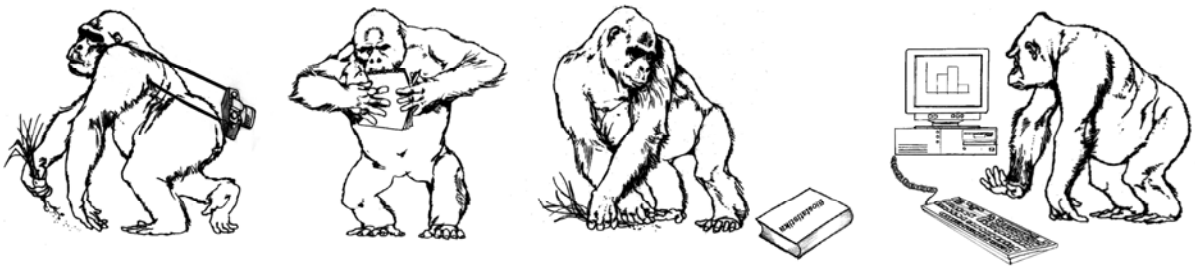
Biostatistika

Jan Šuspa Lepš & Petr Šmilauer



Přírodovědecká fakulta

Jihočeská Univerzita v Českých Budějovicích

České Budějovice, květen 2014

Obsah

Úvod

Moderní biologie je kvantitativní vědou. Biolog váží, měří, počítá, ať už se jedná o krvinky, individua, či nukleární DNA. Každé číslo, které takto získáme, je ale zatíženo náhodnou variabilitou. Počty krvinek v krvi krysy, počítané nezávisle několikrát z jednoho odběru se budou lišit. Více se budou lišit počty krvinek mezi odběry různých krys, patřících do stejné experimentální skupiny. A možná se ještě více budou lišit počty krvinek u krys, patřících do různých experimentálních skupin. Stejně tak se bude lišit obsah nukleární DNA u rostlin patřících do téže populace, obsah dusíku v půdních vzorcích z téže lokality nebo hustota populace buchanky v různých odběrech z téhož rybníka. Říkáme, že data obsahují náhodnou složku; čísla, která získáme, jsou náhodné veličiny. Co je to vlastně náhoda? Touto otázkou se dostáváme až do oblasti filozofie nebo k základům teorie pravděpodobnosti: co je to pravděpodobnost? Biolog se většinou spokojí s pragmatickou představou: náhodné je to, pro co nemáme vysvětlení, a s touto představou si vystačíme i zde.

Statistika je obor, který nám dává návod, jak pracovat s daty obsahujícími náhodnou složku a jak odlišit zákonitosti od náhodné variability (lidová moudrost praví, že statistika je nauka o přesném počítání s nepřesnými čísly). Termín statistika má více významů. V laickém chápání je to uspořádaný soubor dat (statistika střel na branku, statistika hlasování poslanců v parlamentu, statistika počtu aut projíždějících hlavní třídou). Statistika jako věda (někdy též nazývaná matematická statistika) je nástrojem, jak z těchto souborů dat získat užitečnou informaci. Je samostatnou vědní disciplinou, do určité míry ji lze považovat za aplikaci teorie pravděpodobnosti. Termín statistika se používá ještě v dalším významu: je to veličina spočítaná z dat. Např. všeobecně známý průměr můžeme nazvat statistikou, charakterizující daný soubor.

Ve vědeckém uvažování rozlišujeme dva přístupy: deduktivní a induktivní. Deduktivní je takový přístup, kdy ze známých pravd logickou cestou docházíme k důsledkům těchto pravd. Sherlock Holmes na základě faktu, že místnost je uzamčena, nemá okna a nikdo není vevnitř, hbitě *vydedukuje* závěr, že místnost musela být uzamčena zvenčí. Typickým příkladem deduktivního systému je matematika: na základě axiomů můžeme čistě logickou (deduktivní) cestou odvozovat další a další věty, které jsou vždy pravdivé, pokud jsou pravdivé axiomy, z nichž vycházíme (a pokud jsme v odvozování a důkazech vět neudělali chybu). Při deduktivním přístupu postupujeme tedy přísně logicky a nepotřebujeme žádná srovnání s realitou.

Opakem je induktivní myšlení: z mnoha pozorování se snažíme dojít k obecným zákonitostem. Jestliže stokrát stoupneme na led 1cm silný a stokrát se probořím, dojdeme k názoru, že 1cm silný led neunese váhu dospělého člověka. K tomuto závěru jsme došli induktivním myšlením. Ke stejnému závěru jsme mohli ovšem dojít i deduktivní cestou, na základě znalostí fyzikálních zákonů, charakteristik pevnosti pro led a váhy dospělého člověka. Když stoupeme na tenký led, většinou nevíme přesně, jak je silný, a někdy se probořím, někdy ne. Obvykle zjistíme, že jsme se probořili, pokud byl led velmi tenký. Někdy se ale proboří i relativně silný led - kdy, to je ovlivněno řadou okolností, které nejsme schopni kvantifikovat (struktura ledu, opatrnost našeho našlápnutí, atd.) a které tudíž považujeme za náhodné. Z mnoha pozorování potom můžeme odhadnout závislost pravděpodobnosti proboření na tloušťce ledu - k tomu použijeme metod matematické statistiky. Statistika je tedy nástrojem induktivního myšlení v těch případech, kdy výsledek pokusu nebo pozorování je zatížen náhodnou variabilitou.

Díky užití výpočetní techniky se statistika stala dostupnou pro většinu biologů. Podmínkou přijetí článku k publikaci ve většině uznávaných biologických časopisů je, aby kvantitativní data byla statisticky vyhodnocena (některé časopisy dokonce vyžadují, aby byla data **správně** statisticky vyhodnocena). V současné době nelze úplně porozumět většině článků v takových časopisech, jako např. Ecology, American Naturalist (tedy v časopisech čistě biologických) bez znalosti základů statistiky. Všichni biologové plánují pozorování a pokusy, které budou provádět. Přitom jenom správným způsobem sebraná data lze statisticky vyhodnotit. A sebrat data správným způsobem vyžaduje alespoň základní přehled o statistice.

Znalost základů statistiky se tedy stává podmínkou úspěšné práce prakticky ve všech biologických oborech. Statistiky se ovšem také často zneužívá: tradice praví, že známe tři typy lži: úmyslnou, neúmyslnou a statistiku. Jednak lze velmi špatná data "vyšperkovat" užitím složité statistiky tak, že vypadají jako podstatný příspěvek k vědění (a prosadí se tím i do velmi solidních časopisů). Druhým, neméně častým případem je vydávání statistické ("korelační") závislosti za závislost příčinnou. Tímto způsobem lze "dokázat" téměř cokoliv. Oblíbenými, zvlášte v politické a veřejné sféře, jsou různé triky s daty, kdy užitím procent bez udání základu, ze kterého jsou počítány, nebo vhodným posunutím na stupnici při grafickém vynesení vyvoláme ve čtenáři žádoucí efekt. Např. před lety v jakémsi článku v denním tisku, který dokazoval, že naše normy pro znečištění ovzduší jsou příliš přísné, se konstatovalo, že zatímco v Anglii snížili za posledních pět let určitou míru znečištění o 5%, u nás byla snížena o celých 10%. To zní dobře. Nevýhoda byla, že nebyly udány základy, ze kterých jsou procenta počítána. Může to totiž také znamenat, že zatímce před pěti lety byla u nás daná míra znečištění pětkrát vyšší, dnes je vyšší „jenom“ 4.6-krát. Tím by se asi nikdo nechlubil. Znalost statistiky tedy slouží biologovi také k tomu, aby odlišil, která ze sdělení používajících statistiku přinášejí cenné informace, kde je statistika pouze zástěrkou pro myšlenkovou nebo biologickou prázdnotu, kde byla statistika zneužita k důkazu nesprávného tvrzení a kde manipulace s daty mají vyvolat ve čtenáři falešnou představu.

Dostupnost statistického softwaru výrazně změnila přístup k použití statistiky. Dnes si prakticky každý může na svém osobním počítači vyhodnotit svá data - často stačí pár kliknutí myší. Přitom počítač (téměř) vždy dá nějaký výsledek, často ve formě efektního grafu. Je to sice pohodlné, ale skrývá to v sobě nebezpečí. Mnozí lidé prezentují výsledky takových statistických programů, aniž by chápali, co vlastně daný program spočetl. Cílem těchto skript je tedy nejen naučit posluchače vyhodnotit data, ale i chápat, co výsledky statistického zpracování znamenají.

A co je to biostatistika? Nemyslíme si, že jde o samostatnou vědní disciplinu. Spíše tímto termínem naznačujeme, že se jedná o aplikaci statistiky na biologické problémy. V podobném významu se někdy užívá také termín biometrika. Důraz tedy klademe na pochopení principu metod a zásad jejich použití, nikoliv na odvození jednotlivých postupů. Zde je vhodné zdůraznit, že metody, u kterých uvádíme jen princip a příslušné výpočetní postupy, lze většinou analyticky odvodit a klasickým matematickým způsobem (věta, důkaz) dokázat jejich vlastnosti. Protože je toto skriptum určeno biologům, postupujeme opačně: ukazujeme problémy, které daná metoda řeší, a potom uvádíme princip a předpoklady. Při výkladu předpokládáme, že student absolvoval základní kurs matematiky včetně základů počtu pravděpodobnosti. Přesto se snažíme, kde to jde, vyhnout se užití složitějších matematických metod.

Toto skriptum poskytuje pouze základní informace. Každému čtenáři doporučujeme přečíst si o uvedených metodách více. Nejužívanějšími učebnicemi v zahraničí jsou Sokal

a Rohlf (2012), Zar (2007), a Quinn a Keough (2002). Poslední dvě učebnice mají k biologům nejblíže, protože jejich autoři se sami účastní výzkumu v oboru ekologie. Z učebnice Zarovy pochází některé nápady použité v těchto skriptech. Za každou kapitolou uvádíme strany, na kterých může čtenář získat další poučení o probíraných metodách právě v těchto učebnicích. Upozorňuji, že toto skriptum je pouze pomůckou pro základní kurs rozsahu 3/2, a nevešly se sem tedy některé metody, které jsou biologovi velmi užitečné. Jde především o metody mnohorozměrné analýzy, které se tradičně probírají v samostatných učebnicích a kurzech.

Většina odborné biologické literatury je v angličtině. V této literatuře je nutné porozumnět i výrazům, popisujícím analýzu dat. Proto za důležitými termíny – včetně názvů některých podkapitol - uvádíme v závorce jejich anglické ekvivalenty, s použitím *kurzivy*. Klasickým problémem, který řeší český uživatel statistického software, je zda oddělovat desetinou část čísel čárkou (jak velí lokální zvyklosti) nebo tečkou, jak je tomu v anglosaské literatuře a v části statistického software (např. program R, ale třeba program Statistica používá v aktivním českém prostředí desetinou čárku). V učebnici používáme desetinnou tečku, ale čtenář musí pamatovat na to, že ta může být v programech Microsoft Excel nebo Statistica – v závislosti na nastavení operačního systému – nahrazena čárkou.

Předpokládáme, že biolog bude svoje data hodnotit na počítači a ilustrujeme proto postup zpracování a podobu výsledků ve dvou odlišných typech software, často užívaných v Čechách a na Moravě. Program Statistica reprezentuje typ programu méně náročného na uživatele, s příjemným systémem nabídek v menu a dialog boxech a snadno dostupnou a upravitelnou grafickou prezentací výsledků. Program R představuje náročnější alternativu z pohledu uživatelské přívětivosti, ale obsahuje prakticky všechny známé statistické postupy, včetně těch nejmodernějších a navíc je program dostupný zcela zadarmo (viz webové stránky *www.cran.r-project.org*). Předpokládáme, že uživatelé těchto programů znají základní postupy práce s daným statistickým programem, jakými jsou jejich uživatelský interface, import dat nebo export získaných výsledků. Tyto postupy jsou shrnuty v příloze 1 těchto skript, umístěné za poslední kapitolou.

Ve skriptech také ukazujeme, jak mohou být výsledky získané ve statistických programech prezentovány v odborných publikacích a jak použitou statistickou metodu popsat v metodice. V obou případech používáme anglický jazyk, ve kterém je publikován většina odborných sdělení.

Ve skriptech budeme nejčastěji odkazovat na následující učebnice statistiky:

* Zar J. H. (2007): Biostatistical analysis. 5th Edition. Pearson, 960 pp.
* Quinn G. P. & Keough M. J. (2002): Experimental design and data analysis for biologists. Cambridge University Press, 556 pp.

Další užitečné učebnice jsou:

* Sokal R. R. & Rohlf F. J. (2012): Biometry. 4th Edition, W.H. Freeman, San Francisco, 960 pp.
* Green R. H. (1979): Sampling design and statistical methods for environmental biologists. - J. Wiley, New York.
* Jongman R.H.G., ter Braak C.J.F. & van Tongeren, O.F.R. (1995): Data analysis in community and landscape ecology. - Cambridge University Press, 324 pp.
* Šmilauer P. & Lepš J. (2014): Multivariate analysis of ecological data using Canoco 5. Cambridge University Press. 375 pp.

Pro pokročilejší, ale velmi užitečná je učebnice:

* Mead R. (1990): The design of experiments. Statistical principles for practical application. - Cambridge University Press, Cambridge. 636 pp.

Tam, kde je to třeba, citujeme další literaturu na konci příslušné kapitoly.

Velkou pomocí při psaní těchto skript byly reakce studentů na předcházející verze.

1 Základní statistické pojmy, charakteristiky souboru

Typy biologických dat

Při svých výzkumech, ať již v laboratoři nebo v terénu, sledujeme objekty (případy, *cases*), které nás zajímají, a získáváme o nich informace. Všechny údaje, které o sledovaných objektech získáme, budeme nazývat daty. Za data charakterizující rostlinu můžeme považovat její barvu květů, počet jejích listů, výšku jejího stonku, nebo její biomasu. Každá taková charakteristika měřená či odhadovaná pro určitý objekt se nazývá proměnná (*variable*). Rozeznáváme několik typů dat, které se liší svými vlastnostmi a tudíž i způsobem, jakým s nimi při statistickém hodnocení zacházíme.

Data na poměrové stupnici (*data on a ratio scale*), např. výška rostliny, počet listů rostliny, váha krysy atd. Jedná se o kvantitativní data, většinou znázorňující měřitelné množství - hmoty, délky, energie. Pro tato data je typické, že je konstantní rozdíl mezi přilehlými jednotkami (mezi 5 a 6 cm je stejný rozdíl jako mezi 8 a 9 cm) a smysluplná nula. Pro tato data má smysl uvažovat i o jejich poměrech (odtud dostal tento typ dat jméno), například 8 cm je dvakrát více než 4 cm.

Data na intervalové stupnici (*data on an interval scale*) jsou např. stupně Celsia. Opět jde o kvantitativní data, kde rozdíl mezi přilehlými jednotkami je konstantní, není zde ale smysluplná nula. Například stupně Celsia a Fahrenheita mají každý nulu jinde, a každá z nich je stanovena arbitrárně. Nemá také smysl hovořit o poměrech: nelze říci, že 8°C je dvakrát více než 4°C. Zvláštním případem jsou pak data na cirkulární stupnici (hodiny dne, azimut, dny roku, kdy největší možná hodnota je buď identická s nebo přilehlá k nejmenší hodnotě (např. 0° a 360°).

Data na ordinální stupnici (*data on an ordinal scale*), jako např. klasifikační stupně: výborně, velmi dobře, dobře, neprospěl; klasifikace zdravotního stavu: zcela zdráv, lehce nemocen, těžce nemocen, mrtev. Tato data jsou charakteristická tím, že není konstantní rozdíl mezi přilehlými jednotkami, nelze například říci, že rozdíl mezi výborně a velmi dobře je týž jako mezi dobře a neprospěl. Hodnoty ale lze seřadit, lze určit vztah mezi každou dvojicí (je větší, je menší). V biologii se často tato data užívají jako náhražka, kdy nejsme schopni danou charakteristiku měřit lépe (kvantitativně, na poměrové nebo intervalové stupnici), například odhad relativní významnosti jednotlivých rostlinných druhů při popisu vegetace.

Data na nominální stupnici, nominální data, někdy též kategoriální data (*data on a nominal scale*, *categorial variables*, *categorical variables*, *factors*). Příklady mohou být barva, příslušnost ke druhu,typ horniny. Tato data značí příslušnost sledovaného objektu k určité třídě objektů, jeho určitou kvalitativní charakteristiku. Nejsou zde ani konstantní rozdíly mezi kategoriemi, ani nelze jednotlivá pozorování seřadit. Kategoriální data, která mohou nabývat pouze dvou hodnot se nazývají data binární. Většinou se jedná o přítomnost - či nepřítomnost znaku, např. listy lysé nebo chlupaté, samci nebo samice, bakterie gramnegativní nebo grampozitivní, buňky obsahují nebo neobsahují alkoholdehydrogenázu, organismus byl/nebyl očkován, organismus je/není živý apod..

Ordinální i kategoriální data jsou často v programech kódována jako přirozená čísla. Můžeme například kódovat červenou barvu jako jedničku, zelenou jako dvojku a modrou jako trojku. Program nepozná, že se jedná o kategoriální data (pokud mu to nějak nesdělíme) a spočítá nám průměr z červené, zelené a modré - získáme tak zcela nesmyslné číslo. Proto pozor, určité operace lze provádět pouze s určitými typy dat.

Kvantitativní data (na poměrové či intervalové stupnici) se ještě dělí na diskrétní a spojitá:

Diskrétní a spojitá data (*discrete and continuous data*). Pro spojitá data (např. váha) je typické, že mezi kterýmikoliv dvěma hodnotami měření může ležet další. Opakem jsou diskrétní data (např. počet listů). Nejčastěji se jedná o počty, tedy celá čísla, ale ne nezbytně. V biologii toto rozlišení není vždy respektováno. Například pro většinu účelů lze počet krevních destiček v 1 ml krve považovat za spojitou proměnnou (vysoké číslo, přesnost měření je obvykle stejně menší než jedna destička). Na druhou stranu, v některých případech proměnné, které jsou typicky spojité, jsme schopni měřit pouze s určitou přesností (např. výšku stromu hypsometrem obvykle měříme s přesností na půl metru, nebo i na jeden metr). Potom i když je měřená proměnná spojitá, mají naměřené hodnoty diskrétní charakter. Tato diskrétnost je ale artefaktem způsobu měření, nikoliv vlastností měřené proměnné: u zaznamenané výšky stromů se nám budou hodnoty opakovat, zatímco pravděpodobnost, že dva stromy v lese mají stejnou výšku, je prakticky nulová.

Základní soubor a náhodný výběr

Náš výzkum většinou charakterizuje větší (až potenciálně nekonečnou) skupinu případů, **základní soubor** (*statistical population*)**,** na základě zkoumání menší skupiny případů (*cases*, *observations*), která je její součástí. Tato menší skupina pozorování se označuje jako **náhodný výběr** (*sample, random sample;* i když přívlastek *random* nepoužijeme, předpokládáme, že výběr je náhodný). Termín (*statistical*) *population* pro základní soubor velmi často neodkazuje na biologickou populaci jedinců, slovo *population* je zde užíváno v obecnějším smyslu. Procesu, kdy získáváme výběr, se říká v angličtině *sampling.*

Abychom získali **náhodný výběr** (jiný výběr nelze statisticky hodnotit), musíme při výběru pozorování dodržet určitá pravidla: každý člen (jedinec) základního souboru má stejnou a nezávislou šanci, že bude vybrán. Náhodnost by měla být zaručena užitím náhodných čísel. V nejjednodušší variantě všechna individua v souboru očíslujeme od jedné do *N* a poté získáme příslušný počet náhodných přirozených čísel z intervalu (1, N) tak, aby každé číslo mělo stejnou pravděpodobnost, že bude vybráno a žádné se neopakovalo. Poté vybereme a změříme individua příslušející k vybraným číslům. V terénních studiích odhadujeme např. biomasu na ploše tak, že vybereme z celé plochy několik vzorkových plošek, ze kterých odebíráme biomasu. Tyto plošky vybíráme tak, že pro plochu zvolíme systém pravoúhlých souřadnic, a generujeme náhodné souřadnice středu zkusných ploch. Předpokládáme, že základní plocha má tvar pravoúhelníku a že základní plocha je dostatečně velká, abychom mohli zanedbat možnost, že se vzorkovací plošky překrývají.

Podstatně obtížnější je náhodný výběr individuí z populace divoce žijících organismů, kdy není možné všechna individua očíslovat. Zde většinou provádíme výběr, o kterém doufáme, že je náhodnému blízký, a dál s ním pracujeme jako s výběrem náhodným, aniž si často uvědomíme nebezpečí ovlivnění výsledků. Budeme např. studovat hraboše na obilném poli. Získáme jich určitý počet odchytem do pastí, přičemž velikost základního souboru neznáme. Hraboše, kteří se chytili do pastí, považujeme za náhodný výběr, ačkoliv jím pravděpodobně není: starší a zkušenější individua se spíše pasti vyhnou a budou proto ve výběru zastoupena méně. Pro posouzení možných následků této nenáhodnosti, případně pro vypracování odchytového programu, který bude bližší náhodnému výběru, je bezpodmínečně nutná znalost biologie sledovaných druhů. Problémem je ovšem i výběr individuí sedentárních organismů. Očíslovat všechny jedince osívky jarní na pěti arech pískovny a potom z nich vybrat podle pravidel náhodný výběr je sice principiálně možné, ale prakticky

neproveditelné. Je třeba použít metodu, odpovídající studovanému objektu a jeho rozmístění v prostoru. Zde je třeba upozornit, že často užívaný postup, kdy zvolíme náhodný bod v ploše (tj. náhodně generujeme souřadnice) a vybereme individuum nejbližší tomuto náhodnému bodu, není náhodným výběrem: soliterní individua mají větší šanci, že budou do výběru zařazena, než individua, která se vyskytují ve shluku. Pokud jsou individua ve shluku menší (což se díky kompetici stává), budou všechny odhady parametrů založené na tomto výběru vychýlené.

Existují metody výběru, kdy si celý soubor rozdělím na několik homogenních podsouborů a teprve v nich provádím náhodný výběr. Například při průzkumech veřejného mínění jsou respondenti vybíráni náhodně, ale zvlášť v rámci určitých podsouborů, o kterých předpokládáme, že jsou homogennější než základní soubor; zvlášť se vybírají obyvatelé Prahy, zvlášť obyvatelé velkých měst, zvlášť venkované. V každé dílčí skupině je ale nutné provést náhodný výběr. Pro takto strukturovaný výběr musíme níže uvedené odhady parametrů a jejich přesnosti modifikovat podle struktury výběru.

Subjektivní výběr individuí, ať už typických nebo zdánlivě náhodných (např. jdu podél řádky na poli a občas vyberu rostlinu) není náhodným výběrem.

Základní soubor může být hypotetický. Např. 5 králíků v pokusu reprezentuje potenciální (imaginární) množinu všech možných králíků stejného druhu stejně živených atd.

Charakteristiky souboru

Předpokládejme, že chceme popsat výšku souboru padesáti studentů. Padesát hodnot výšky podává sice úplnou, ale značně nepřehlednou informaci. Proto se budeme snažit tuto informaci zjednodušit a zpřehlednit, a to tak, aby došlo k minimální ztrátě informace. To můžeme provést dvojím způsobem; buď si informaci převedeme do grafické podoby nebo se budeme snažit soubor popsat pomocí několika charakteristik, nazývaných popisné statistiky (*descriptive statistics*), které vystihnou nejdůležitější vlastnosti celého souboru. Z grafických shrnutí se často užívá histogram četností. Zkonstruujeme jej tak, že rozsah hodnot proměnné rozdělíme do několika tříd stejné šíře a do histogramu vynášíme počet případů v každé třídě. Někdy se místo počtu případů vynášejí relativní četnosti, jako procento případů z celého souboru (tvar histogramu ani vypovídací schopnost se tím nemění, mění se pouze stupnice na svislé ose). Jestliže máme dostatečně velký počet pozorování a dostatečně úzké třídy, tvar histogramu četnosti odpovídá charakteristice rozdělení, kterou nazýváme hustota pravděpodobnosti (viz kapitola 2). Další možnosti grafického znázornění jsou uvedeny v podkapitole o grafickém shrnutí dat.

Druhou možností je použití popisných statistik. Budou nás zajímat především dvě věci: jak jsou studenti „v průměru“ vysocí a jak se liší výšky studentů v rámci souboru. První typ informace nám podávají charakteristiky polohy (též centrální tendence), druhou charakteristiky variability. Charakteristiky konečného souboru (např. náhodného výběru, ale i konečného základního souboru) můžeme zjistit přesně, charakteristiky nekonečného základního souboru (nebo základního souboru, kde jsme nezměřili všechna individua) odhadujeme právě na základě náhodného výběru. Platí pravidlo, že charakteristiky základního souboru (a parametry rozdělení) se obvykle píší řeckými písmeny, parametry výběru písmeny latinky. Výjimkou je počet prvků v souboru: **N** - počet prvků v základním souboru; **n** - počet prvků ve výběru.

Charakteristiky polohy

Otázky: Jaká je výška studentů prvního ročníku?; jak velký je obsah PCB v mléce prodávaném v Českých Budějovicích? (Výšku všech studentů prvního ročníku můžeme teoreticky zjistit přesně, tj. mohu všechny studenty nastupující do prvního ročníku změřit, nebo se mohu spokojit s odhadem na základě náhodného výběru; v případě mléka prodávaného v Českých Budějovicích musím vystačit s odhadem na základě náhodně odebraných vzorků.) Zajímá nás, jaké jsou „v průměru" hodnoty veličin, jaká je poloha dat na zvolené škále. Tuto intuitivně chápanou „průměrnou hodnotu" může charakterizovat několik parametrů:

Aritmetický průměr (*arithmetic mean*)

průměr základního souboru je

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Vz. 1-1

průměr výběru je

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Vz. 1-2

Výběrový průměr je odhadem průměru základního souboru. Průměr je definován pro data na poměrové a intervalové stupnici.

Příklad: Výšky studentů v cm (soubor pěti studentů) byly 151, 155, 161, 180, 205; průměr = (151 + 155 + 161 + 180 + 205) / 5 = 170.4

Pozor, aritmetický průměr ani jiné charakteristiky polohy nelze užít bez modifikace pro data na cirkulární škále. Například zjišťujeme, jaká je průměrná hodnota orientace kmene stromu, na kterém byl nalezen určitý druh lišejníku. Dostali jsme hodnoty (ve stupních, kde 0 i 360 značí sever): 5, 10, 355, 350, 15, 145. Podle Vz. 1-2 dostáváme hodnotu 180, indikující, že průměrná orientace kmene byla na jih (přičemž všechny byly obráceny k severu). Pro data na cirkulární škále je třeba užít zvláštní metody, a to pro všechny charakteristiky. Způsob práce s těmito daty popisuje např. Zar (1984 pp. 422-469).

Medián a kvantily (*median and other quantiles*)

Pro medián není obecně uznávaný symbol. Medián je definován tak, že stejný počet pozorování leží pod mediánem jako nad mediánem. Nebo přesněji je pravděpodobnost, že hodnota sledované proměnné bude u náhodně vybraného individua větší než medián, stejná, jako že bude menší než medián. V teoretických rozděleních je to hodnota náhodné proměnné, kdy je distribuční funkce rovna 0.5. Medián je definován pro data na poměrové, intervalové i ordinální stupnici. Kromě mediánu se užívají i jiné **kvantily**. Nejčastěji užívané jsou **kvartily** - horní kvartil je definován jako hodnota, nad kterou se nachází čtvrtina pozorování, dolní obdobně: čtvrtina pozorování leží pod ním. Obdobně můžeme definovat další **kvantily**. Ke kvantilům rozdělení se vrátíme při popisu charakteristik rozdělení.

Pro výšky studentů uvedené v příkladu pro výpočet průměru je medián 161. Medián vypočítáme tak, že pozorování nejprve seřadíme podle velikosti. Když je *n* liché, medián je roven $X_{(n+1)/2}$, tj. prostřední ze všech pozorování . Když je *n* sudé, medián je střed intervalu dvou prostředních pozorování, tj. $(X_{n/2}+X_{n/2+1}) / 2$. Pro sudý počet pozorování, např. váhy studentů 50, 52, 60, 63, 70, 94: medián je 61.5 Zvláštním způsobem se medián někdy počítá, když padne mezi pozorování stejné hodnoty (*tied observations*), viz Zar (1984, p. 22).

Jak uvidíme dále, medián se shoduje s průměrem, pokud mají data symetrické rozdělení. Jak se průměr a medián liší v rozděleních asymetrických, ukazuje následující příklad: máme dvě skupiny živočichů o jedenácti individuích, každá získává potravu jiným způsobem. Množství potravy (přepočtené na gramy organického uhlíku za den) získané každým individuem je následující:

Skupina 1: 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 21

Skupina 2: 5, 5, 6, 6, 7, 8, 9, 15, 35, 80, 120.

V první skupině je průměr množství zkonzumované stravy 17.8, ve druhé je 26.9. Průměrná konzumace stravy, charakterizovaná aritmetickým průměrem, je tedy ve druhé skupině vyšší. Naproti tomu medián v první skupině je 18, zatímce ve druhé pouze 8. Průměrné individuum (charakterizované tím, že polovina individuí se nají víc než ono a polovina míň) tedy hladoví podstatně více ve druhé skupině.

Modus (*mode*)

Modus je definován jako nejčastěji se vyskytující pozorování. Pro spojitá rozdělení je to hodnota proměnné odpovídající lokálnímu maximu (nebo lokálním maximům) hustoty pravděpodobnosti. Modus nemusí být nutně jeden, rozdělení totiž mohou být i bimodální (se dvěma mody), popř. polymodální. Modus definován pro všechny typy dat. Pro spojitá data odhadujeme modus obvykle jako střed intervalu nejvyššího sloupce v histogramu četností, v případě polymodálních dat jako polohu sloupců, které převyšují sousední sloupce. Zde je třeba upozornit, že tento odhad závisí na naší volbě šíře třídy a že skutečnost, že jsme na základě výběru a při daných intervalech dostali histogram s více „mody“ nemusí nutně znamenat, že základní soubor je polymodální.

Geometrický průměr (*geometric mean*)

Je to n-tá odmocnina součinu *n* hodnot:

$$\mathrm{GM} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

Vz. 1-3

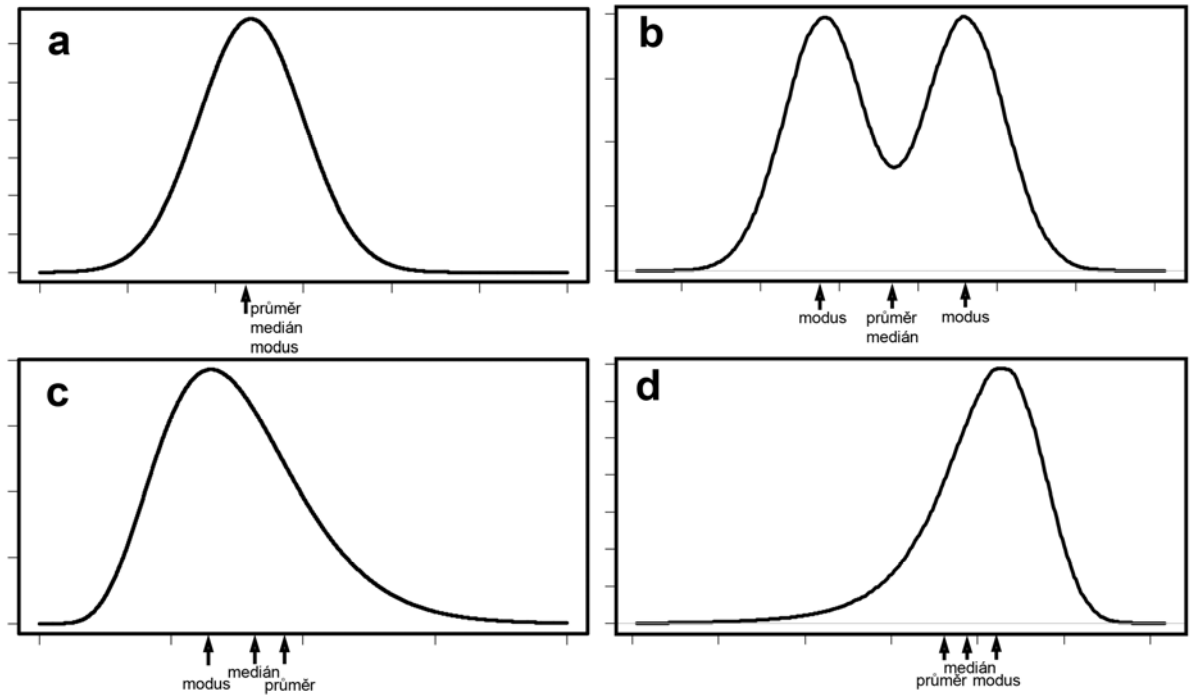
Harmonický průměr (*harmonic mean*)

Je to převrácená hodnota průměru převrácených hodnot:

$$\mathrm{HM} = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{1}{X_i}}$$

Vz. 1-4

Geometrický a harmonický průměr se užívají pro data na poměrové stupnici, pokud neobsahují nuly.



Obr. 1-1 Idealizované frekvenční histogramy (hustoty pravděpodobnosti) s označenými charakteristikami polohy. Hodnoty proměnné jsou vynášeny podél horizontální osy (*abscissa*) a frekvence na vertikální ose (*ordinate*). Rozdělení **a** a **b** jsou symetrická, **c** je pozitivně šikmé, **d** je negativně šikmé. Rozdělení **a**, **c** a **d** jsou unimodální a rozdělení **b** je bimodální.

Charakteristiky variability (rozptylu)

Kromě toho, jaká je „v průměru" hodnota sledované proměnné, nás také zajímá, jak se hodnoty v rámci sledovaného souboru mezi sebou liší, jak jsou variabilní. Na to dávají odpověď charakteristiky variability.

Otázka: Jak variabilní je výška studentů?

Rozsah (*range*)

Rozsah je rozdíl mezi největším a nejmenším pozorováním. Pro data o výškách studentů uvedená výše to je 44 cm. Pozor, se zvětšováním výběru rozsah většinou roste a rozsah výběru není proto dobrým odhadem rozsahu základního souboru.

Variance, rozptyl (*variance*)

Variance a hodnoty z ní odvozené jsou nejužívanějšími charakteristikami variability. Variance definována jako průměrná hodnota druhé mocniny (čtverce) odchylky od průměru.

Variance základního souboru je definována takto:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Vz. 1-5

Variance výběru (jako odhad variance základního souboru):

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Vz. 1-6

Namísto s^2 se někdy též užívá zkratky *var* nebo *VAR*. Variance výběru je odhadem variance základního souboru.

Variance výšky studentů je $((151-170.4)^2+(155-170.4)^2+(161-170.4)^2+(180-170.4)^2+(205-170.4)^2)/5 = 398.24$, pokud uvedených pět studentů pokládáme za základní soubor a $((151-170.4)^2+(155-170.4)^2+(161-170.4)^2+(180-170.4)^2+(205-170.4)^2)/4 = 497.8$, pokud je uvedených pět studentů pokládáno za výběr z širšího základního souboru.

Směrodatná odchylka (*standard deviation*)

Směrodatná odchylka je odmocnina z variance (pro výběr i pro základní soubor). Kromě *s* se také často značí jako *S.D.*, *s.d.* nebo *SD*, zvláště v anglických textech.

Směrodatná odchylka základního souboru je

$$\sigma = \sqrt{\sigma^2}$$

Vz. 1-7

Směrodatná odchylka výběru

$$s = \sqrt{s^2}$$

Vz. 1-8

Považujeme-li pět studentů za základní soubor, potom je směrodatná odchylka $\sqrt{398.24} = 19.96$; jako výběr je $s = \sqrt{497.8} = 22.31$

Variační koeficient (*coefficient of variation*)

Jde o podíl směrodatné odchylky a průměru:

$$CV = \frac{s}{\overline{X}}$$

Vz. 1-9

Variační koeficient je smysluplný pro data na poměrové stupnici. Užívá se tam, kde chceme porovnat variabilitu nestejně velkých druhů objektů. Můžeme se například ptát, zda se je větší variabilita výšek v populaci netýkavky žláznaté nebo netýkavky nedůtklivé (netýkavka žláznatá dosahuje výšek přes 2 m, netýkavka nedůtklivá bývá vysoká do 30 cm). Potom je třeba míru variability vztáhnout na průměrnou výšku porovnávaných objektů, a tedy užít variační koeficient. Naproti tomu, porovnáváme-li variabilitu teplot (data na intervalové stupnici), nemá smysl směrodatnou odchylku teploty vztahovat k průměrným teplotám; nemáme smysluplnou nulu, a variační koeficient vychází jinak pro Celsiovy stupně, jinak pro Fahrenheitovy.

Za míru disperse lze také považovat **mezikvartilové rozpětí** *(interquartile range)*, tj. rozdíl mezi horním a dolním kvartilem (ten není, na rozdíl od rozsahu hodnot, systematicky ovlivněn velikostí výběru).

Přesnost odhadu průměru, střední chyba průměru

Výběrový průměr je také náhodná veličina (zatímco průměr základního souboru **není** náhodná veličina). Průměr jako náhodná veličina má tedy také svoji varianci. Provedeme-li několik výběrů z téhož základního souboru, jejich průměry se budou lišit. Tuto varianci můžeme odhadnout pomocí variance základního souboru (nebo jejího odhadu). Variance průměru je

$$s^2_{\bar{x}} = s^2_x / n$$

Vz. 1-10

Její odmocnina je tedy směrodatná odchylka výběrového průměru jako náhodné veličiny, a je také nazývána střední chyba průměru (*the standard error of mean*). Značí se $s_{\bar{x}}$, *s.e.*, *s.e.m.*, *SEM* a je nejčastěji užívanou charakteristikou přesnosti odhadu průměru (jinou charakteristikou je konfidenční interval, který bude probírán později a který se počítá na jejím základě). Ze Vz. 1-10 tedy dostáváme vzorec pro výpočet střední chyby průměru

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

Vz. 1-11

Poznámka: *jak získáme vzorec Vz. 1-10. Pokud jsou dvě proměnné (x, y) nezávislé, potom platí*

$$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$$

Vz. 1-12

Dále platí, je-li k konstanta, potom

$$\sigma^2_{kx} = \sigma^2_x \cdot k^2$$

Vz. 1-13

Ze Vz. 1-12 můžeme ukázat, že variance součtu n nezávislých pozorování (tj. n náhodných proměnných, z nichž každá pochází ze souboru s variancí s_x^2) je $n.s^2$. Protože průměr je součet dělený počtem pozorování, je podle Vz. 1-13 variance průměru n^2-krát menší, než variance součtu a tím dostáváme vzorec Vz. 1-10.

Nezaměňujte: směrodatná odchylka popisuje variabilitu dat, se kterými pracujeme; její odhadovaná hodnota není závislá na velikosti výběru. Střední chyba průměru popisuje přesnost našeho odhadu; její hodnota klesá s rostoucí velikostí výběru – čím větší výběr, tím přesnější odhad.

Grafická shrnutí dat

Většina článků přináší obvykle údaj o průměru a směrodatné odchylce, případně o střední chybě průměru. Tím ovšem ztrácíme nemalou část informace o datech, například o typu rozdělení. Obecně platí, že dobře zvolený graf, který data shrnuje, o nich řekne více než jedno nebo dvě čísla, představující sumární statistiky. Představu o tvaru rozdělení dat získáme nejsnáze tak, že si vyneseme tzv. histogram četností (*frequency histogram*, viz Obr. 1-2). Jiným typem zobrazení je tzv. *box-and-whisker plot* (česky snad krabice s fousy, užívá se někdy krabicový nebo obdélníkový graf). Význam jednotlivých symbolů vysvětluje Obr. 1-3. Některé programy, včetně programu *Statistica*, užívají box-and-whisker plot

k vynášení aritmetického průměru a směrodatných odchylek (a charakteristik od nich odvozených). To je přístup vhodný, pokud můžeme předpokládat, že základní statistická populace má normální rozdělení (viz dále v této kapitole), ale obecně je více informativní vynášet tento typ diagramu založený na mediánu a kvartilách, protože takto lépe vynikne případná nesymetričnost distribuce hodnot a případně lze identifikovat i přítomnost neobvyklých hodnot (podle voleb pro kreslení diagramu).

Příkladová data

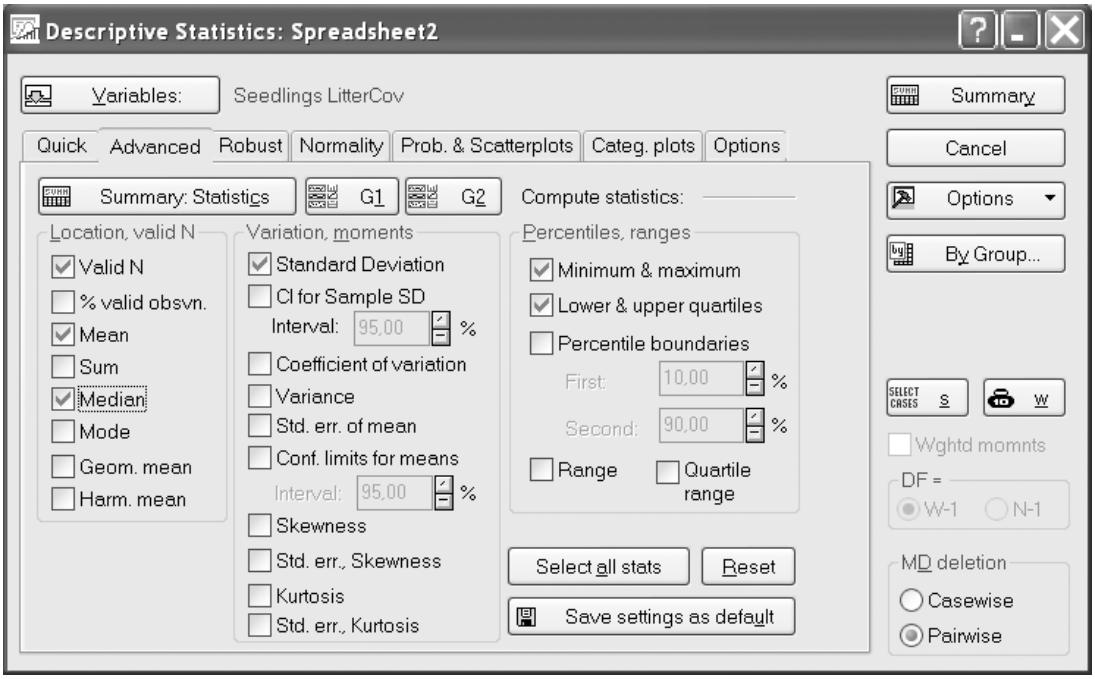
Data v listu *Chap1* souboru *biostat-data.xls* obsahují pozorování z 24 experimentálních ploch terénního pokusu, ve kterém byl studován vliv kosení luční (proměnná *Mown* označuje, zda daná plocha byla kosena či ne) na počet semenáčků rostlin, které vyklíčily během roku v této ploše (proměnná *Seedlings*). Protože badatel předpokládal, že vliv kosení se může projevovat přes změnu množství rostlinného opadu na povrchu půdy, byla také zaznamenávána procentická pokryvnost tohoto opadu (proměnná *LitterCov*).

Cílem analýzy těchto dat je (v kontextu této kapitoly) je spočtení základních výběrových statistik proměnných *Seedlings* a *LitterCov*, jednak pro celý soubor, jednak zvlášť pro kosené a nekosené plochy a také grafické shrnutí těchto dat.

Jak postupovat v programu Statistica

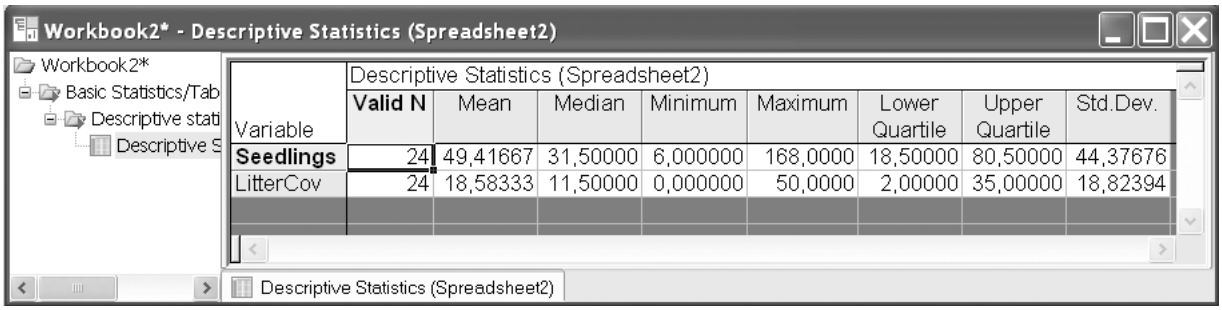
Statistiky pro celý výběr a jeho části

Z menu zvolíme *Statistics* | *Basic Statistics/Tables*, v zobrazeném seznamu vybereme položku *Descriptive statistics* a zvolíme tlačítko *OK*; v zobrazeném dialog boxu vybereme záložku *Advanced*.



Nejprve vybereme pomocí dialog boxu zobrazeného tlačítkem *Variables:* proměnné *Seedlings* a *LitterCov*, pro které budeme statistiky počítat. Výběr statistik, které nám Statistica nabízí jako předvolbu, doplníme pro náš příklad o *Median* a *Lower & upper quartiles* (viz zobrazení boxu výše) a nakonec zvolíme tlačítko *Summary*. Statistica zobrazí

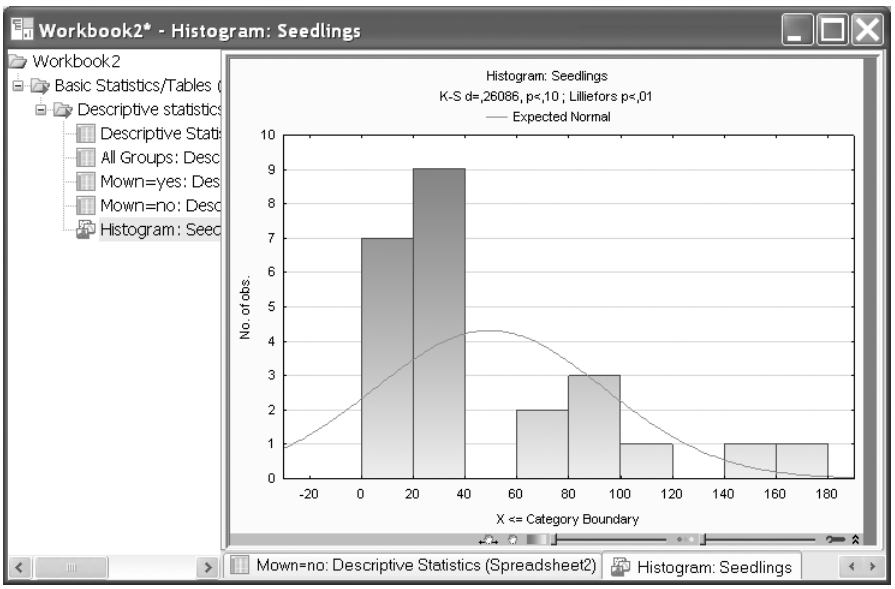
nové okno (*Workbook*), ve kterém jsou zobrazeny zvolené statistiky, které můžeme kopírovat do jiných aplikací.



Do stále aktivního dialogu (*Descriptive Statistics*) se můžeme vrátit volbou jeho tlačítka u dolní hrany pracovního prostoru programu Statistica. Pokud bychom chtěli spočíst tyto statistiky pro skupiny pozorování, definované jedním nebo více kategoriálními proměnnými, můžeme aktivovat režim *By Group* pomocí tlačítka na pravém okraji boxu. V zobrazeném dialog boxu vybereme proměnnou (či proměnné) definující skupiny (v našich datech proměnná *Mown*) a případně zaškrtneme volbu *Output to single folder* pro jednodušší zobrazení výsledků. Po návratu do hlavního boxu *Descriptive Statistics* spočteme statistiky opět pomocí tlačítka *Summary*. Tentokrát se vytvoří ne jedna, ale tři tabulky výsledků: první odpovídá původním výsledkům (pro všechna pozorování), další dvě představují statistiky pro pozorování, u kterých má proměnná *Mown* hodnotu *yes* nebo *no*, tedy odděleně pro kosené a nekosené plochy. Režim *By Group* zůstává v platnosti, dokud box *Descriptive Statistics* neuzavřeme. Podmnožiny pozorování můžeme definovat také složitějším, ale více flexibilním způsobem pomocí tlačítka *SELECT CASES*.

Grafické shrnutí kvantitativních proměnných

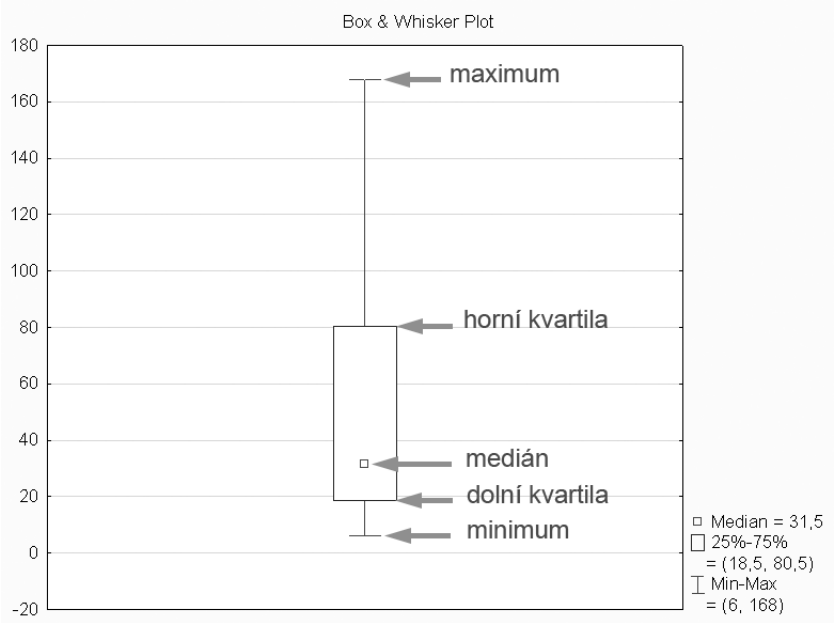
Rozsáhlejší nabídku pro tvorbu grafů je možné najít v menu *Graphs*, ale základní grafická shrnutí lze provádět z dialog boxu *Descriptive Statistics*. Pro jejich ilustraci vybereme nejprve tlačítko *Variables* a tentokrát vybereme pouze proměnnou *Seedlings*. V záložce *Quick* dialog boxu *Descriptive Statistics* můžeme zvolit tlačítko *Histograms*, které zobrazí nejen frekvenční histogram pro zvolenou proměnnou, ale také porovná distribuci jejích hodnot s normální distribucí (blíže viz kapitola 4).



Obr. 1-2

Počet intervalů, do kterých je proměnná rozdělena na horizontální ose, lze nastavit na stránce *Normality* dialog boxu *Descriptive Statistics*, odkud lze histogram také vytvářet. Počtem intervalů také určujeme jejich šířku. Tu musíme volit i s ohledem na přesnost měření. Pokud jsme například měřili výšku stromů s přesností na jeden metr, musí být šířka intervalu celistrvým násobkem jednoho metru. Pokud bychom např. zvolili šířku menší, dostaneme v histogramu intervaly, které nebudou obsahovat žádná měření.

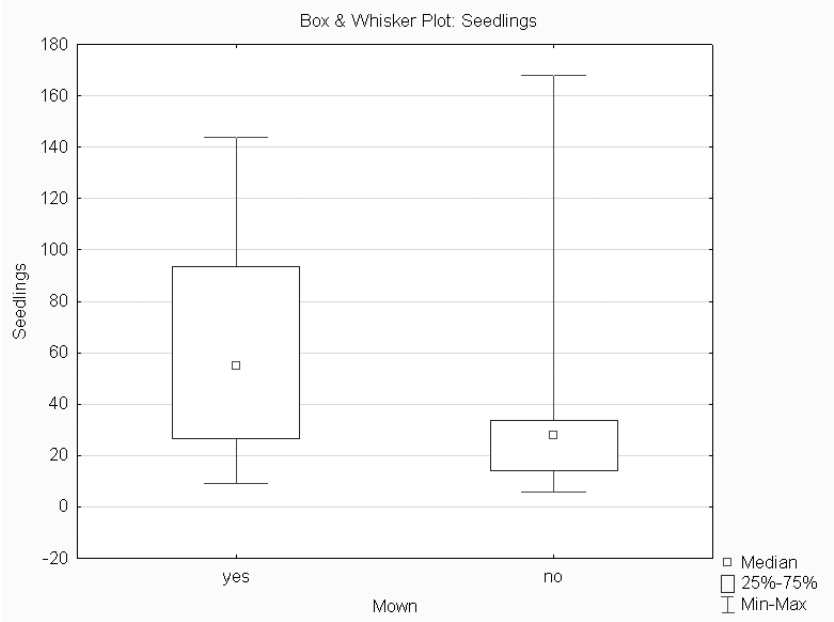
Před vytvořením box-and-whiskers grafu v klasické podobě je třeba na stránce *Options* stejného dialog boxu zvolit v oblasti *Options for Box-Whisker plots* variantu *Median/Quartiles/Range*. Graf lze pak vytvořit pomocí tlačítka *Box & whisker plot for all variables* na záložce *Quick*.



Obr. 1-3

Výsledný graf (doplněný informací o významu jeho jednotlivých prvků) naznačuje (podobně jako histogram a také skutečnost, že medián má výrazně nižší hodnotu než aritmetický průměr) nesymetrickou distribuci počtu semenáčků, s několika málo výrazně většími hodnotami a většinou pozorování v dolní části celkového rozsahu hodnot.

Pokud bychom chtěli vynášet histogramy či box-and-whisker grafy samostatně pro skupiny pozorování definované jednou či více kategoriálními proměnnými, lze to provést opět pomocí tlačítka *By Group*. Tento postup ale vytvoří oddělené grafy pro každou skupinu, takže je pak obtížné je navzájem porovnat. Je proto lepší použít příkazy ze záložky *Categ. plots*, ze kterých zde představujeme použití tlačítka *Categorized box & whisker plots*. Po jeho volbě se zobrazí nový dialog box, ve kterém je třeba vybrat klasifikující proměnnou (či proměnné), v našem příkladu proměnnou *Mown*. Před zobrazením diagramu se ještě objeví box, ve kterém je možné vybrat jen část kategorií přítomných ve vybrané kategoriální proměnné, v našem případě (*Mown* obsahuje jen dvě kategorie) to ale nedává smysl. Box-and-whisker diagram v programu Statistica zobrazuje jako "whiskers" celý rozsah hodnot, ale je také schopen omezit tento rozsah na přilehlé hodnoty, tj. vyloučit z něj hodnoty odlehlé (*outlying observations* či *outliers*). Toho můžeme dosáhnout změnou voleb již vytvořeného diagramu (lze je otevřím dvojitým poklepáním na graf), v záložce *Box/Whisker*, kde změníme hodnotu *Whisker value* z *Min-Max* na *Non-Outlier Range*. Následující ilustrace ale ukazuje diagram bez této modifikace.



Jak postupovat v programu R

Data v listu *Chap1* importujeme do datového rámce *chap1*. Základní popisné statistiky získáme pro numerické proměnné v datovém rámci pomocí funkce *summary*:

```
> summary(chap1)
   Seedlings       Mown      LitterCov
 Min.   :  6.00   no :12   Min.   : 0.00
 1st Qu.: 18.75   yes:12   1st Qu.: 2.00
 Median : 31.50            Median :11.50
 Mean   : 49.42            Mean   :18.58
 3rd Qu.: 75.25            3rd Qu.:35.00
 Max.   :168.00            Max.   :50.00
```

1st Qu. a *3rd Qu.* jsou dolní a horní kvartila, význam ostatních statistik je asi jasný. Další statistky lze pro proměnné v datovém rámci spočítat za pomocí funkce *sapply*, například varianci proměnných:

```
> sapply(chap1,var)
   Seedlings         Mown    LitterCov
1969.2971014    0.2608696  354.3405797
```

nebo pro směrodatnou odchylku:

```
> sapply(chap1,function(x)sqrt(var(x)))
 Seedlings       Mown  LitterCov
44.3767631  0.5107539 18.8239364
```

Statistiky udávané pro kategoriální proměnnou (faktor) *Mown* jsou založeny na převodu jejích hodnot na nuly (*no*) a jedničky (*yes*).

Pokud chceme spočíst pro jednotlivé proměnné výběrové statistiky pro jednotlivé skupiny pozorování, pomohou nám funkce *split* (rozdělující hodnoty proměnné, která je prvním parametrem na skupiny definované druhým parametrem) a opět *sapply*, například:

```
> x <- with( chap1, split( Seedlings, Mown))
> sapply(x, mean)
      no      yes
36.91667 61.91667
```

```
> sapply(x, var)
      no      yes
1926.447 1850.265
```

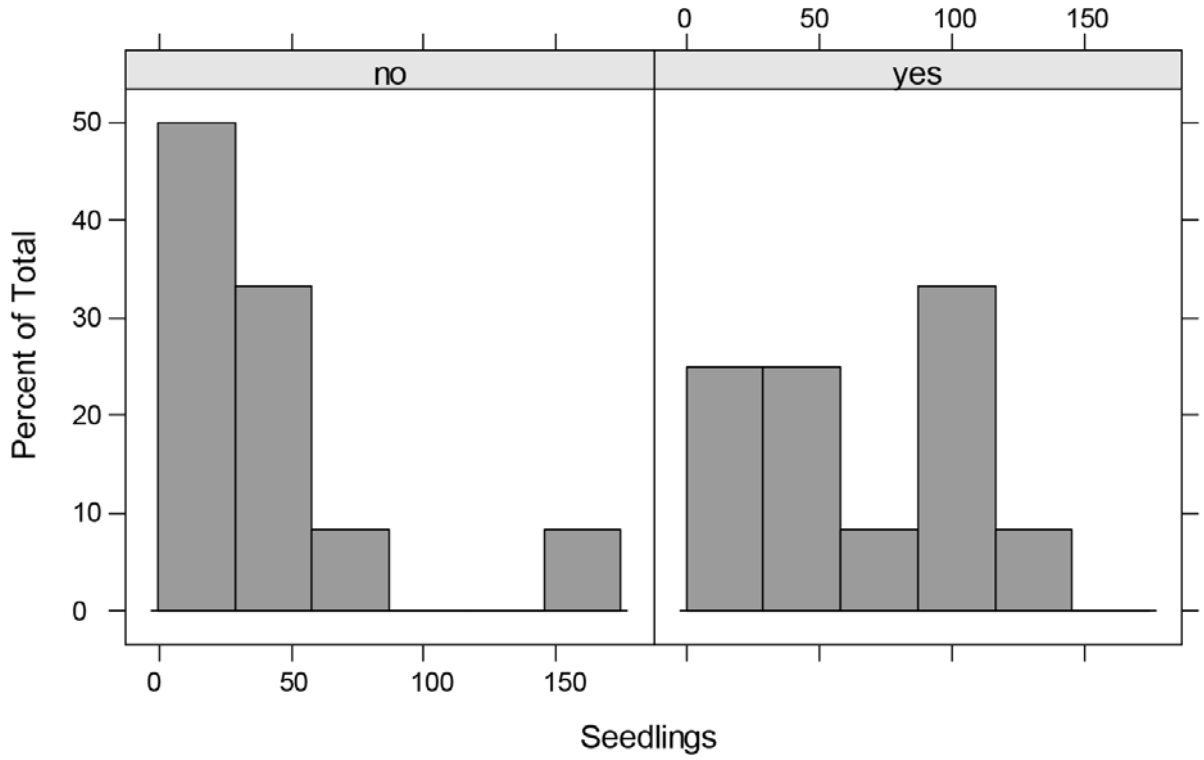
Tvorbu obrázků graficky shrnujících data si ukážeme s použitím knihovny *lattice*, i když základní grafická knihovna také podporuje tvorbu jednoduchých histogramů a box-and-whisker diagramů.

```
> library(lattice)
> histogram(~Seedlings,data=chap1)
```

Pokud chceme tyto diagramy porovnat mezi skupinami pozorování (např. kosené a nekosené plochy v našem příkladě), můžeme to provést takto:

```
> bwplot(Mown~Seedlings,data=chap1)
> histogram(~Seedlings|Mown,data=chap1)
```

Příkaz pro histogram vytvoří následující graf:



Popis analýz v článku

Shrnutí hodnot měřených proměnných se neobjevuje v článcích běžně (obvykle jen na vyžádání recenzenta) a počet zobrazených charakteristik je v takovém případě omezený (typicky na jednu charakteristiku polohy a jednu charakteristiku rozptylu, případně rozsah pozorovaných hodnot). Pro náš příklad jsme k tradičním parametrickým charakteristikám (průměr a směrodatná odchylka) přidali ještě medián.

Methods

Measured quantitative variables were summarized using mean, median, and sample standard deviation.

Results

We have recorded seedling counts and estimated percentage cover of plant litter on all 24 experimental plots (Table 1).

Table 1: Summary statistics of measured variables.

	Mean	Median	Standard Deviation
Seedling count	49.4	31.5	44.4
Cover of litter (%)	18.6	11.5	18.8

Náhodné veličiny, rozdělení, distribuční funkce, hustota pravděpodobnosti

Všechny dosud uvedené vzorce lze užít pouze pro soubory nebo výběry konečné velikosti. Abychom mohli spočítat průměr pro soubor, musíme změřit všechna individua v daném souboru, a to lze pouze pro soubor konečné velikosti. Představme si nyní, že máme nekonečný základní soubor, z něhož vybíráme individua, nebo máme náhodný proces, který můžeme libovolněkrát opakovat a kterého výsledkem je určitá hodnota - určitá náhodná veličina. Například při studiu rozšiřování rostlin pouštíme semeno trubicí z určité výšky a měříme jeho rychlost na konci trubice (tzv. *terminal velocity* - ta se považuje za dobrou charakteristiku schopnosti šířit se větrem). Proces můžeme teoreticky opakovat nekonečněkrát. (Není to zase tak jednoduché, většinou chceme charakterizovat druh, a v tom případě bych měli pokaždé pouštět jiné semeno daného druhu, a jednotlivá semena by měla být náhodným výběrem ze všech semen daného druhu.) Rychlost, kterou semeno dopadá, považujeme za náhodnou proměnnou a časy, které naměříme, jsou realizacemi této náhodné proměnné. Realizace náhodné proměnné jsou vlastně náhodným výběrem z potenciálně nekonečné množiny všech možných rychlostí, kterými semeno mohlo dopadat.

Náhodnou proměnnou potom charakterizujeme pomocí pravděpodobností, se kterou může nabývat dané hodnoty, tedy pomocí **rozdělení pravděpodobností** (*probability distribution*, často ale užíváme jenom **rozdělení**, *distribution*). Následující teorie se vztahuje na kvantitativní data, tj. na data na poměrové nebo intervalové stupnici. V teorii přísně rozlišujeme diskrétní a spojité veličiny. V praxi mnohdy užíváme metody určené pro spojité veličiny i pro veličiny diskrétní, zvláště pokud mohou diskrétní veličiny nabývat velmi mnoha hodnot (viz dřívější příklad s počtem krvinek).

Rozdělení pravděpodobností, a distribuční funkce diskrétních náhodných veličin

Pro **diskrétní** náhodnou veličinu lze její jednotlivé (možné) hodnoty očíslovat (může nabývat nejvýše spočetně mnoha hodnot). Můžeme ji popsat buď rozdělením pravděpodobností nebo distribuční funkcí.

Rozdělením pravděpodobností nazýváme výčet všech možných hodnot x_i a jim příslušejících pravděpodobností, že náhodná veličina dané hodnoty nabude, tj. $p_i = P(X = x_i)$. Rozdělení pravděpodobností může být zadáno tabulkou (Tab. 1-1) nebo vzorcem.

$\mathbf{x_i}$	x_1	x_2	...	x_n
$\mathbf{p_i}$	p_1	p_2	...	p_n

Tab. 1-1

Je logické, že součet všech pravděpodobností p_i se musí rovnat jedné:

$$\sum_{i=1}^{n} p_i = 1,$$

Vz. 1-14

kde počet možných hodnot (n) může být konečný nebo nekonečný.

Funkce $F(x)$, která se rovná pravděpodobnosti $\mathbf{P}(X < x)$ toho, že náhodná veličina bude menší než zvolené číslo x, se nazývá **distribuční funkce** *(distribution function*, někdy též *cumulative distribution function*, podle způsobu výpočtu) náhodné veličiny X. Platí, že

$$F(x) = \sum_{x_i<x} p_i,$$

Vz. 1-15

kde se sčítání provádí přes všechny hodnoty i, pro které je $x_i < x$. Např. pravděpodobnost, že x je menší než 5, se rovná součtu pravděpodobností příslušejících všem hodnotám x menším než 5.

Distribuční funkce a hustota pravděpodobnosti spojité náhodné veličiny

Náhodná veličina X, která může nabývat libovolných číselných hodnot z daného intervalu a pro kterou existuje pro libovolné x z tohoto intervalu limita

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}$$

Vz. 1-16

se nazývá spojitá náhodná veličina. Funkce $f(x)$ se nazývá **funkce hustoty pravděpodobnosti** (*probability density function*). O čem nás informuje graf hustoty pravděpodobnosti? Vzpomeňme si na konstrukci histogramu četností v úvodní kapitole. Pokud by se Δx ve Vz. 1-16 rovnalo jedné, a vynášeli bychom s krokem 1 hodnotu $P(x<X<x+\Delta x)/\Delta x$ proti x, dostaneme histogram relativních četností, kde každá hodnota bude odpovídat pravděpodobnosti, že v daném intervalu (x, x+1) leží hodnota náhodné veličiny. Pokud budeme zužovat tento interval, bude se nám snižovat pravděpodobnost, že náhodná proměnná dané hodnoty nabude. Proto je v čitateli Δx, které nám závislost této pravděpodobnosti na šíři intervalu koriguje. V limitním přechodu, kdy se Δx blíží nule, dostáváme funkci hustoty pravděpodobnosti. Hustotu pravděpodobnosti můžeme tedy chápat jako idealizovaný histogram relativních četností pro nekonečně velký základní soubor.

Spojitá náhodná veličina může být určena také (kumulativní) distribuční funkcí $F(x) = P(X < x)$, kde x je libovolné reálné číslo. Je to tedy pravděpodobnost toho, že náhodná veličina je menší než x.

Distribuční funkce $F(x)$ má tyto základní vlastnosti:

1. $P(a \le X < b) = F(b) - F(a)$;
2. $F(x_1) \le F(x_2)$ pro $x_1 < x_2$;
3. $\lim_{x \to +\infty} F(x) = 1$;

4. $\lim_{x \to -\infty} F(x) = 0$;

Vz. 1-17

To znamená, že:

1. Pravděpodobnost, že X nabude hodnoty z intervalu (a, b), je rovna rozdílu hodnot příslušných distribučních funkcí. Například pravděpodobnost, že X bude ležet mezi hodnotami 3 a 5 je rovna rozdílu pravděpodobností, že X bude menší než 5 a že X bude menší než 3.

2. Distribuční funkce je neklesající. Přirozeně, pravděpodobnost, že $X<3$ nemůže být větší, než pravděpodobnost, že $X<5$.

3. a 4. X je jistě menší než $+\infty$ a jistě není menší než $-\infty$.

Hustota pravděpodobnosti $f(x)$ má tyto základní vlastnosti:

1. $f(x) \geq 0$;

2. $f(x) = \mathrm{d}F(x) / \mathrm{d}x$;

3. $P(a \leq X < b) = \int_a^b f(x)dx$;

4. $\int_{-\infty}^{+\infty} f(x)dx = 1$

Vz. 1-18

To znamená, že:

1. Hustota pravděpodobnosti je vždy nezáporná.

2. Tato funkce je derivací distribuční funkce.

3. Určitý integrál od a do b z hustoty pravděpodobnosti nám udává pravděpodobnost, s jakou náhodná veličina nabude hodnoty v intervalu od a do b.

4. Hodnota X jistě leží mezi $-\infty$ a $+\infty$.

Pravděpodobnost, že spojitá náhodná proměnná nabude přesně libovolné hodnoty a, se limitně blíží nule. Proto se pro praktické účely nemusíme příliš starat o to, kde jsou ve Vz. 1-17 a Vz. 1-18 ostré či neostré nerovnosti. Z praktického hlediska je pravděpodobnost, že $X>5$ pro spojitou proměnnou stejná, jako pravděpodobnost, že $X\geq5$.

Pomocí distribuční funkce můžeme definovat kvantily. Veličina x_p, definovaná rovností $F(x_p) = p$, se nazývá **kvantil**. Často se vyjadřuje v procentech. Například $x_{0.95}$ se nazývá 95%-ní kvantil; je to taková hodnota X, pro kterou je distribuční funkce rovná 0.95. To znamená, že ji náhodná proměnná překročí s pravděpodobností p=0.05 (s 5%-ní pravděpodobností). Kvantil $x_{0.5}$ se nazývá **medián,** kvantily $x_{0.25}$ a $x_{0.75}$ jsou dolní a horní **kvartil**. Jestliže má hustota maximum, pak ta hodnota x, při které platí $f(x) = \max$, se nazývá **modus.**

Podobně jako konečný soubor můžeme i rozdělení charakterizovat pomocí průměru (ten se pro rozdělení nazývá střední hodnota rozdělení) a variance. Pro diskrétní proměnnou je střední hodnota definována

$$\mu = \sum_{i=1}^{n} x_i p_i ,$$

Vz. 1-19

a variance je dána vzorcem

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p_i ,$$

Vz. 1-20

Směrodatná odchylka σ je druhá odmocnina z variance.

Střední hodnota rozdělení spojité náhodné proměnné je dána vzorcem

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx,$$

Vz. 1-21

a variance vzorcem

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Vz. 1-22

Některá rozdělení lze teoreticky odvodit, lze pro ně analyticky vyjádřit funkci hustoty pravděpodobnosti, popř. distribuční funkci. Takováto rozdělení mívají svá jména (např. normální, binomické, χ^2), a hodnoty jejich distribuční funkce lze spočítat ve statistických programech (postup je popsán v kapitole 4).

Náhodným procesem vedoucím k realizaci náhodné proměnné může být také provedení náhodného výběru, ze kterého spočteme určitou charakteristiku. Například výběrový průměr je náhodnou proměnnou. Dokonce jsme schopni odvodit vztah charakteristik tohoto rozdělení k charakteristikám rozdělení základního souboru.

Doporučená četba

Zar (2007), pp. 1 - 26.

Quinn & Keough (2002), pp. 7 – 17; pp. 58 – 61 pro grafická shrnutí.

2 Testování hypotéz, testy dobré shody

Dva základní statistické postupy jsou odhad parametrů a testování hypotéz. V minulé kapitole jsme si ukázali, jak odhadujeme charakteristiky základního souboru, v této kapitole probereme základy testování hypotéz (*hypothesis testing*). Mezi základní poučky metodologie vědy patří, že shoda dat s hypotézou ještě neznamená, že hypotéza je pravdivá; naproti tomu data odporující hypotéze ukazují na to, že hypotéza pravdivá není. Hypotézu nelze na základě dat dokázat; hypotézu však lze na základě dat vyvrátit. Z toho vychází i statistické testování hypotéz. Ukážeme si jej na příkladě vyhodnocení nominálních dat; ne proto, že by se pro jiné typy dat neužívalo, ale proto, že je na nominálních datech nejsnáze pochopitelné.

Postup je následující: Formulujeme nulovou hypotézu. Nulová hypotéza je formulována tak, aby ji mohla data vyvrátit v případě, že není pravdivá. Většinou to tedy bývá opak toho, co chceme dokázat. Nulová hypotéza (*null hypothesis*, H_0) je většinou formulována jako: něco se neliší; není rozdíl; není závislost; platí zákon atd. Poté se snažíme dokázat, že určitá data nejsou slučitelná (jsou v rozporu) s touto nulovou hypotézou. Pokud to dokážeme, zamítám nulovou hypotézu a přijímáme alternativní hypotézu H_A, někdy též H_1, která je negací nulové hypotézy.

Příklad: Studujeme dědičnost rostliny a ptáme se, zda pro barvu květů platí jednoduchá mendelovská dědičnost. Předpokládáme, že v F_2 generaci bude poměr počtu červenokvětých k bělokvětým 3:1. Mám 80 potomků. Potom předpokládáme, že v potomstvu bude 60 červenokvětých a 20 bělokvětých jedinců. My ale máme 70 červenokvětých a 10 bělokvětých.. Jsou naše výsledky v rozporu s poměrem 3:1, tzn. s předpokladem, že každé individuum má pravděpodobnost 0.75 být červenokvěté a pravděpodobnost 0.25 být bělokvěté?

I v případě, že pravděpodobnosti jsou 0.75 a 0.25, můžeme s určitou pravděpodobností dostat výsledný poměr 70:10. Dokonce můžeme dostat všech 80 červenokvětých (s pravděpodobností 0.75^{80}, což je řádově 10^{-10}). V takovém případě ovšem nebudeme ochotni věřit, že šlo pouze o náhodu, a dojdeme k názoru, že nulová hypotéza neplatí (zamítneme ji).

Nulovou hypotézu zamítáme, pokud dostaneme uspořádání dat, které je za předpokladu platnosti nulové hypotézy velmi nepravděpodobné.

Co to ale je, "velmi nepravděpodobné"? Statistika nám k tomu poskytuje následující návod: Zvolíme si, jak nepravděpodobný výsledek za předpokladu platnosti nulové hypotézy musíme dostat, abychom se rozhodli pro závěr, že nulová hypotéza neplatí. Většinou se rozhodujeme pro 5% nebo 1%. Této hodnotě říkáme hladina významnosti (*significance level*) testu a značíme ji α; bývá zvykem ji vyjadřovat desetinným číslem, např. α=0.05. Potom spočteme testové kriterium (někdy tuto hodnotu nazýváme testová statistika; zde je další význam termínu statistika). Pro toto kriterium je známo, jaké má rozdělení v případě platnosti nulové hypotézy. Je tedy známo, kterou hodnotu překročí s pravděpodobností 5%, kterou hodnotu překročí s pravděpodobností 1% atd. Těmto hodnotám říkáme kritické hodnoty (*critical values*).

Jestliže hodnota testového kriteria překročí kritickou hodnotu pro zvolenou hladinu významnosti, zamítneme nulovou hypotézu na dané hladině významnosti. Říkáme potom, že výsledek (nesouhlas s nulovou hypotézou) je průkazný (signifikantní) na dané hladině významnosti. Pro kategoriální data používáme testy dobré shody (*goodness of fit*) a používáme kritéria (statistiky) χ^2, čti „chí-kvadrát" (angl. *chi-square*, čti „kaj skvér").

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

Vz. 2-1

k je celkový počet kategorií, které sleduji (v našem příkladě 2), $\hat{f}_i$ je očekávaná četnost v i-té kategorii (často se také značí *E,* z anglického *expected*), f_i je četnost skutečná (pozorovaná, někdy též *O, observed*). V našem příkladě tedy formulujeme nulovou hypotézu: v F_1 generaci je poměr pravděpodobností výskytu červenokvětých k bělokvětým 3:1. V 80-tičlenném potomstvu tedy předpokládám 60 a 20 individuí.

$$\chi^2 = \frac{(70-60)^2}{60} + \frac{(10-20)^2}{20} = 6.66$$

Hodnotu 6.66 porovnávám s kritickou hodnotou χ^2 distribuce pro danou hladinu významnosti α (většinou volíme 0.05 nebo 0.01) a daný počet stupňů volnosti. Stupně volnosti, (*degrees of freedom*), značíme je většinou df, DF nebo ν (řecké písmeno „ný“). Pro testy tohoto typu je počet stupňů volnosti roven počtu kategorií minus jedna (*k*-1). Je to počet četností ve skupinách, které potřebujeme znát, abychom znali celý výsledek. Počet případů v poslední kategorii můžeme dopočítat ze znalosti předcházejících *k*-1 kategorií a celkového počtu pozorování (ten je v testech považován za fixní). Víme-li, že z osmdesáti jedinců bylo sedmdesát červenokvětých, znám výsledek celého pokusu. Získaná hodnota 6.66 je větší než kritická hodnota $\chi^2_{0.05,1}$, (tj. při 5% hladině významnosti a jednom stupni volnosti), jejíž hodnota je 3.84. Zamítáme tedy nulovou hypotézu na pětiprocentní hladině významnosti. Závěr by tedy zněl: Pozorovaná data se významně (signifikantně) na 5%-ní hladině významnosti liší od četností, předpokládaných jednoduchou mendelovskou dědičností. V našem případě by odlišnost byla průkazná i na 1%-ní hladině významnosti.

Pozor: v testu užíváme přímo napozorované četnosti. Nelze převést nejprve údaje na procenta a potom počítat s procenty!!!

Příklad: V jeskyni je velké množství netopýrů (pro nás jich je nekonečně mnoho), samci a samice. Chceme zjistit, zda je poměr samců a samic 1:1. Nejsme ale schopni prohlédnout všechny netopýry v jeskyni. Chytíme jich tedy 100 a podle nich se snažíme rozhodnout. Oněch 100 individuí musí být náhodným výběrem! Nulová hypotéza zní: V jeskyni je stejně samců jako samic (což je totéž jako: pravděpodobnost, že náhodně vybrané individuum je samec, je stejná, jako že náhodně vybrané individuum je samice).

Existují dvě možnosti, jak je tomu ve skutečnosti:

1. V jeskyni je stejně samců jako samic, obě pravděpodobnosti jsou tedy 0.5. To znamená, že nulová hypotéza platí (je pravdivá). Výsledek pokusu ale může být dvojího typu:

1a) např. 55 samců; 45 samic. Potom $\chi^2=(55-50)^2/50+(45-50)^2/50 = 1.1 < 3.84$. Nemůžeme zamítnout nulovou hypotézu. Správné rozhodnutí.

1b) např. 60 samců; 40 samic. Potom $\chi^2=(60-50)^2/50+(40-50)^2/50 = 4.4 > 3.84$. Zamítáme nulovou hypotézu na 5%-ní hladině významnosti. Udělali jsme **chybu prvního druhu (***Type I error*). Pravděpodobnost této chyby známe: je to α. Hladina významnosti α je tedy podmíněná pravděpodobnost zamítnutí nulové hypotézy za předpokladu, že nulová hypotéza platí.

2. Samci tvoří 60% individuí, náhodně vybrané individuum bude samec s pravděpodobností 0.6; samice s pravděpodobností 0.4. Nulová hypotéza tedy neplatí - je nepravdivá. Výsledek pokusu může být opět dvojí:

2a) např. 55 samců; 45 samic. Potom $\chi^2=(55-50)^2/50+(45-50)^2/50 = 1.1 < 3.84$. Nemůžeme zamítnout nulovou hypotézu. Dopustili jsem se tak **chyby druhého druhu** (*Type II error*). Její pravděpodobnost označujeme jako β a většinou ji neznáme. 1 - β je **síla testu** (*power of the test*). Obecně platí, že síla testu roste s odchylkou od nulové hypotézy a s počtem pozorování. Dále platí, že čím menší je α, tím větší je β. Protože β neznáme, je správná formulace výsledku: **Na základě dat nemůžeme zamítnout nulovou hypotézu**. Formulace: *Dokázali jsme nulovou hypotézu* je **nesprávná!**

2b) např. 60 samců; 40 samic. Potom $\chi^2=(60-50)^2/50+(40-50)^2/50 = 4.4 > 3.84$. Zamítáme nulovou hypotézu na 5%-ní hladině významnosti. Správné rozhodnutí.

Máme tedy dvě možnosti, jaká je realita (nulová hypotéza buď platí nebo neplatí), a naše rozhodnutí může být také dvojí (nulovou hypotézu zamítám, nebo nezamítám). Celý proces je zvykem ilustrovat tabulkou:

		Skutečnost	
		Je-li H_0 správná	Je-li H_0 špatná
Naše rozhodnutí	H_0 jsme zamítli	Chyba 1. druhu	Správné rozhodnutí
	H_0 jsme nezamítli	Správné rozhodnutí	Chyba 2. druhu

Tab. 2-1 Chyba 1. a 2. druhu při statistickém rozhodování

Chyby prvního i druhého druhu jsou vlastní statistickému rozhodování a vyplývají ze stochastického (náhodného) charakteru studovaných procesů; nelze je tedy žádným způsobem z našeho rozhodování zcela eliminovat. Čím menší pravděpodobnost chyby prvního druhu jsme ochotni připustit, tím větší máme pravděpodobnost chyby druhého druhu. Představme si v příkladu netopýrů, že jsme ochotni přijmout pravděpodobnost chyby prvního druhu pouze 0.01. Kritická hodnota testu je 6.63. Co je toho důsledkem? V případě 1b jsme se díky přísnosti kriteria nedopustili chyby prvního druhu; naproti tomu jsme se v případě 2b dopustili chyby druhého druhu. Za lepší ochranu před chybou prvního druhu platíme větší pravděpodobností, že uděláme chybu druhého druhu.

Na tomto příkladě však lze demonstrovat nebezpečí jiných chyb, které sice nejsou vlastní statistice, ale při aplikaci statistických metod na biologické problémy se jim většinou také nevyhneme. Oněch 100 individuí pokládáme za náhodný výběr. Nicméně, abychom opravdu mohli provést náhodný výběr, museli bychom všechna individua očíslovat a potom podle tabulky náhodných čísel vybrat 100 individuí - to logicky není možné. Za náhodný výběr obvykle považujeme ta individua, která se nám podaří získat. Předpokládejme, že sbíráme netopýry v zimě, když visí ze stropu jeskyně. Pokud si např. samci vybírají pro přezimování pro člověka obtížněji dostupná místa než samice, nebo se samci rychleji probudí a dřív nás pokoušou, takže nám jich víc uletí, je pravděpodobné, že v našem výběru bude (statisticky významně) více samic. Pokud by netopýři viseli ze stropu ve skupinách, které jsou složeny převážně z jednoho pohlaví, a my, když už se nám podaří ke tropu vylézt, tak vyšetříme všechny jedince ve skupině, bude narušen předpoklad, že výběr jedince je nezávislý na tom, zda bude vybrán jedinec jiný (a ve výsledku to může znamenat větší odchylku od nulové hypotézy. Ve shora popsaných případech se nám nepodařilo provést náhodný výběr, a ve všech případech může být výsledkem velká odchylka od stavu předpokládaného při platnosti nulové hypotézy. V praxi tedy může být zamítnutí nulové hypotézy důsledkem tří skutečností:

1. Nulová hypotéza neplatí.

2. Nulová hypotéza platí, ale dopustili jsme se chyby 1. druhu.

3. Nulová hypotéza platí, ale my jsme nesplnili všechny předpoklady pro užití testu.

Test dobré shody a základní vzorec

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

můžeme použít pro libovolný počet kategorií.

Následující příklad představuje výsledky sledování dvou znaků na semenech (semena zelená, žlutá; svraskalá a hladká). Žlutá a hladká se považují za dominantní projevy znaku. Očekávaný poměr je potom 9:3:3:1. V tomto případě je počet kategorií k = 4, DF (ν) = 3. Bylo sledováno 250 semen. Pozorované četnosti fenotypů byly 126, 55, 60, 9.

Nulová hypotéza (H_0): Sledovaný výběr semen pochází ze základního souboru charakterizovaného poměrem fenotypů žlutých hladkých, žlutých svraskalých, zelených hladkých a zelených svraskalých 9:3:3:1 (to můžeme též formulovat jako "pravděpodobnosti výskytu daných fenotypů jsou v poměru 9:3:3:1"). Alternativní hypotéza (H_A): Semena pocházejí ze základního souboru, který nemá poměr shora uvedených fenotypů 9:3:3:1.

Očekávané četnosti spočteme trojčlenkou. Např. očekávaný počet žlutých hladkých semen je 250 . (9/16) =140.625

	žlutá hladká	žlutá svraskalá	zelená hladká	zelená svraskalá	n
pozorované	126	55	60	9	250
očekávané	140.625	46.875	46.875	15.625	

Tab. 2-2 Příklad užití χ^2 testu

$$\chi^2 = \frac{14.625^2}{140.625} + \frac{8.125^2}{46.875} + \frac{13.125^2}{46.875} + \frac{6.625^2}{15.625} = 1.521+1.4083+3.675+2.809=\mathbf{9.413}$$

Protože 9.431 je větší než kritická hodnota pro α=0.05 (ta je 7.815), zamítáme nulovou hypotézu na 5%-ní hladině významnosti. Můžeme tedy zamítnout hypotézu, že data odpovídají modelu jednoduché mendelovské dědičnosti s nezávislými znaky. Z hodnot jednotlivých sčítanců vidíme, že nejvýraznější příspěvek k vysoké hodnotě testovacího kritéria dává třetí kategorie (semena zelená a hladká).

Uveďme další příklady užití tohoto testu:

(**1**) Včely jsou postupně vpouštěny do pokusného prostoru se žlutými, červenými a modrými terči. Sledujeme barvu terče, na který každá včela poprvé usedne. Nulová hypotéza je, že pravděpodobnost usednutí nezávisí na barvě terče. Tímto způsobem zjišťujeme, zda se včely vizuálně orientují a zda při této orientaci hrají nějakou úlohu barvy. Data: bylo vpuštěno 100 včel; četnosti barev, na které poprvé usedly: žlutá 47, červená 38, modrá 15. Lze z těchto dat usoudit, že včely některou barvu preferovaly? Nulová hypotéza bude znít: Pravděpodobnost usednutí včely na terč nezávisí na barvě terče, a očekávané četnosti tedy budou 33.3, 33.3 a 33.3 pro jednotlivé barvy. Na tomto pokusu můžeme demonstrovat další podmínky použití tohoto testu:

(a) **Pozorované četnosti pocházejí z nezávislých pokusů**. Proto vpouštíme včely do pokusného prostoru po jedné, a zaznamenáváme chování každé včely. Kdybychom vpustili všechny včely do prostoru najednou a spočetli počet, který se usadil na každém terči, může být (průkazná) odchylka od nulové hypotézy dána tím, že včely poletí společně jako roj a společně usednou na terč, který náhodně vybere jedna z nich, jakási „vedoucí roje“. Při provádění pokusu si musíme být jisti, že usednuvší včela nenechá na terči nějakou značku (např. pachovou), která by umožnila dalším včelám se orientovat. Pokud si tím nejsme jisti, musíme terče vyměňovat před vpuštěním každé další včely. Dále je třeba zajistit, aby včely nemohly preferovat určitý terč nikoliv kvůli barvě, ale kvůli pozici v pokusném prostoru. Proto bychom ve správně prováděném pokusu pozice barevných terčů náhodně střídali po každé jednotlivé včele.

(b) **Před pokusem jsme měli pevně daný celkový rozsah výběru.** Nesprávný (ale občas užívaný) postup je takový, kdy po prvních 100 včelách zjistíme, zda je výsledek testu průkazný; pokud není, „zvětšíme velikost výběru“, přidáme dalších 30 včel a proceduru opakujeme, a tak to zkoušíme několikrát a sledujeme, zda dostaneme kýžený průkazný výsledek, který nám umožní publikovat zásadní práci o tom, jak se včely orientují podle barvy. Takovýto postup mnohonásobně zvyšuje pravděpodobnost chyby prvního druhu!

(**2**) Porovnání poměru pohlaví (*sex ratio*) ve skupině s očekávaným poměrem 1:1. Data: Za poslední měsíc se v porodnici města X narodilo 89 chlapců a 99 děvčat. Byl poměr průkazně odlišný od očekávaného poměru 1:1? Nulová hypotéza tedy zní: Pravděpodobnosti narození chlapce a děvčete byly stejná. Očekávané četnosti jsou tedy 94:94.

Ze statistického hlediska nám nic nebrání testovat nulovou hypotézu, že pravděpodobnost narození chlapce byla dvakrát větší než pravděpodobnost narození děvčete. Potom by ovšem očekávané četnosti byly 125.34 a 62.66. Tuto hypotézu bychom jistě zamítli. Ovšem zamítnutí takové hypotézy je málo zajímavé, neboť není žádný důvod předpokládat, proč by měla platit. Naproti tomu, pokud bychom zamítli nulovou hypotézu o poměru 1:1, můžeme hledat smysluplná vysvětlení, proč tomu tak je. Obdobně, při testování štěpných poměrů v genetice nám ze statistického hlediska nic nebrání testovat nulovou hypotézu, že štěpný poměr je např. 1:17. Její zamítnutí nám ovšem potvrdí to, co jsme předem věděli, že štěpný poměr 1:17 je zřejmý nesmysl. Poměr 3:1 očekáváme, protože známe zákony mendelovské dědičnosti, a zároveň víme, že existují mechanismy, které tento poměr narušují. Ty budeme hledat v systémech, kde jsme dokázali odchylku od očekávaného štěpného poměru. Formulace nulové hypotézy je tedy, stejně jako uspořádání pokusu či plán sledování, věcí nejen statistickou, ale především věcí znalosti problému a jeho biologické podstaty.

Nulovou hypotézu formulujeme v matematických termínech, v uvedených příkladech používáme pravděpodobnosti nebo očekávané četnosti jevů. Vše ostatní je mimostatistické uvažování. V příkladu s barevnými terči jistě můžeme usoudit, že pokud zamítneme nulovou hypotézu, že pravděpodobnost usednutí na terč nezávisí na barvě terče, potom včely musí mít schopnost barvy rozlišit. Nicméně hypotéza “včely nejsou schopny rozlišit barvy”, není nulovou hypotézou statistického testu.

Testování Hardy-Weinbergovy rovnováhy

Testování shody s Hardy-Weinbergovou rovnováhou je příkladem užití testu dobré shody v případě, že použijeme výsledná data pro stanovení očekávaných frekvencí za předpokladu platnosti nulové hypotézy. Hardy-Weinbergova rovnováha předpokládá, že genotypy dominantní homozygoti, heterozygoti, a recesivní homozygoti jsou v populaci zastoupeni

v poměru p^2, 2pq a q^2, kde p je frekvence dominantní alely. Tu odhadneme **z dat** jako (2 x počet dominantních homozygotů + počet heterozygotů)/(2 x počet všech individuí). A protože jeden parametr nulového modelu odhadujeme z dat, musíme při závěrečném testování odečíst jeden stupeň volnosti.

Velikost výběru

Test dobré shody je pouze přibližný. Přiblížení je velmi dobré, pokud je velikost výběru velká; doporučuje se, aby žádná očekávaná četnost nebyla menší než 1 a aby méně než 20% četností bylo menších než 5. Pokud tomu tak není, můžeme některé kategorie s malými četnostmi spojit.

Kritické hodnoty; dosažená hladina významnosti

Základem mnoha statistických testů je následující postup. Spočteme **testovou statistiku**, např. χ^2, o které víme, jaké má rozdělení v případě platnosti nulové hypotézy. Např. víme, že testová statistika, která vznikne sečtením

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} = \sum_{i=1}^{k} \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

Vz. 2-2 O je pozorovaná četnost (*observed*), E je očekávaná četnost (*expected*).

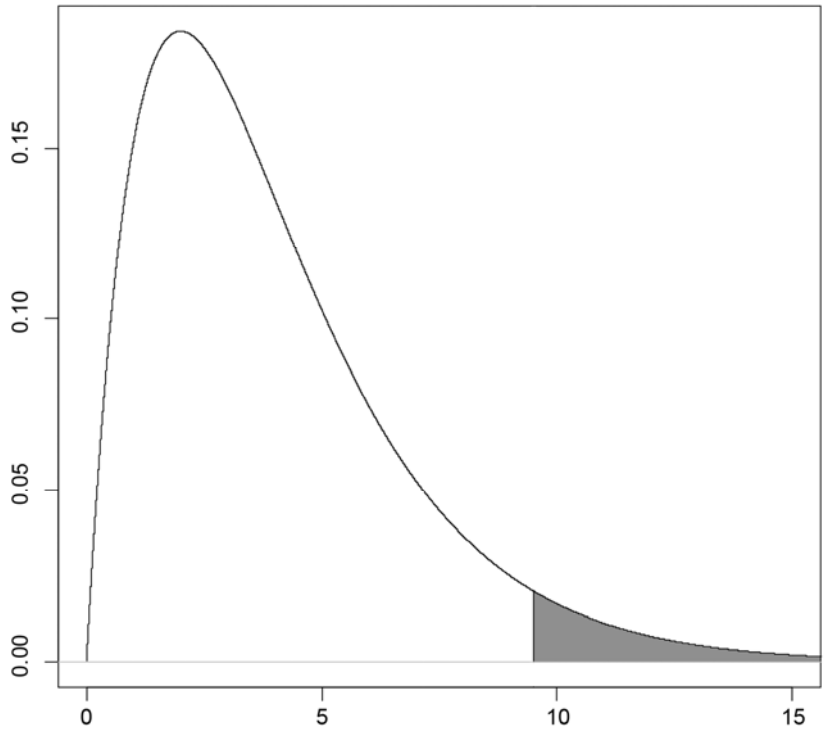
má za předpokladu platnosti nulové hypotézy (na jejímž základě jsou očekávané četnosti počítány) rozdělení, které jsme schopni charakterizovat distribuční funkcí a tuto distribuční funkci vyčíslit. Toto rozdělení se nazývá, stejně jako testová statistika, χ^2. Toto rozdělení je spojité; testové kriterium počítáme z počtu případů, tedy nutně z dat diskrétních, a proto může testové kriterium nabývat pouze určitých hodnot. Proto musí být velikost výběru dostatečně velká, aby „nespojitost“ příliš nevadila. Toto rozdělení patří mezi tzv. výběrová rozdělení a tvar jeho distribuční funkce závisí na počtu stupňů volnosti. Protože toto rozdělení je známé, lze spočítat jeho 95%-ní kvantil.

Víme, že podle definice je 95%-ní kvantil hodnota, kterou náhodná proměnná překročí s pravděpodobností 0.05. Pro shora uvedený test je tedy 95%-ní kvantil rozdělení kritickou hodnotou na 5%-ní hladině významnosti (tj. při $\alpha = 0.05$). Víme, že hodnota kriteria je tím větší, čím je větší odchylka od nulové hypotézy. Pokud tedy hodnota testového kriteria překročí kritickou hladinu na 5%-ní hladině významnosti, můžeme říci, že pokud nulová hypotéza platí, potom pravděpodobnost, že dostaneme výsledek takto nebo více odlišný od nulové hypotézy je menší než 5%.

Většina statistických programů přímo s hodnotou testové statistiky poskytuje také odpovídající hodnotu pravděpodobnosti, kterou nejčastěji nazývá *Probability,* případně jenom *P* či *p*, ale také někdy *Significance level* (tj. hladina signifikance). Je to 1 - hodnota distribuční funkce pro spočtenou hodnotu testového kriteria, což je totéž jako hodnota určitého integrálu z hustoty pravděpodobnosti od spočtené hodnoty do $+\infty$. Na grafu hustoty pravděpodobnosti (Obr. 2-1) je to plocha, kterou pokrývá „ocas“ rozdělení od dané hodnoty do nekonečna. Tato hodnota nám udává přímo pravděpodobnost, s jakou dostaneme takto nebo více od nulové hypotézy odlišný výsledek za předpokladu, že nulová hypotéza platí. Této hodnotě se říká dosažená hladina významnosti. Pokud je dosažená hladina významnosti menší než 0.05, znamená to, že test je průkazný při $\alpha = 0.05$. V biologických článcích se nyní nejčastěji referuje o výsledcích testů následujícím způsobem (výsledek testu z Tab. 2-2):

V pokusu získaný štěpný poměr 126:55:60:9 se statisticky významně lišil od poměru předpokládaného jednoduchou mendelovskou dědičností ($\chi^2 = 9.41$, df=3, $p<0.05$).

Pro prezentaci výsledků statistických testů je vhodné dodržovat následující: α používáme pro předem stanovenou hladinu významnosti, takže píšeme např. „test je průkazný při $\alpha=0.05$“; *P* (nebo *p*) používáme pro dosaženou hladinu významnosti, takže píšeme $p<0.05$. Pokud nám program napíše, že $p=0$ (nebo často $p=0.0000$, znamená to, že hodnota dosažené hladiny významnosti je menší než přesnost zobrazení v tomto programu. Nepište do článků $p=0$, ale např. $p<10^{-6}$. Častěji ale udáváme přímo dosaženou hladinu významnosti. Sdělením $p=0.049$ naznačujeme, že výsledek testu byl sice průkazný na 5% hladině významnosti, ale „s odřenýma ušima“. Podobně sdělení $p=0.052$ naznačuje, že jsme nulovou hypotézu sice nezamítli, ale mnoho nechybělo, aby k tomu došlo. Dosažená hladina významnosti je velmi cennou informací v publikacích a proto tento způsob prezentace doporučujeme.



Obr. 2-1 Hustota pravděpodobnosti rozdělení χ^2 se čtyřmi stupni volnosti. Celá plocha vymezená křivkou a osou x je rovna jedné, velikost šedé plochy odpovídá pravděpodobnosti, že proměnná nabude hodnotu větší než 9.5. Jestliže jsme dostali hodnotu testového kriteria 9.5, potom velikost šedé plochy odpovídá dosažené hladině významnosti testu (zde p=0.05).

Klasická statistika doporučovala striktně postup, kdy nejprve pevně stanovíme hladinu významnosti a poté dostaneme jednoznačnou odpověď: zamítáme nebo nezamítáme hypotézu. Dnes se užívá víceméně jen ten přístup, kdy prezentujeme dosaženou hladinu významnosti a podle ní posuzujeme i důvěryhodnost výsledku: pokud posuzujeme poměr pohlaví v populaci netopýrů, s určitou nejistotou se smíříme, není třeba zcela jednoznačně rozhodnout ano nebo ne. Naproti tomu prvý způsob musíme nutně použít tam, kde na základě testu činíme rozhodnutí typu ano/ne. Například se rozhodujeme, zda zavést výrobu preparátu, jehož účinnost má test prokázat. Zde si musíme předem stanovit míru rizika, kterou jsme jako výrobce ochotni nést (bude pravděpodobně velmi malá, například 0.001), a pokud výsledek pokusu nebude průkazný, výrobu nezavedeme. Obdobný postup se užívá u klinických testů při zavádění nových léků.

Příliš dobré, aby to byla pravda (*too good to be true*)

Příklad: Společnost, vyrábějící nový druh žvýkačky, byla obviněna, že pravidelné žvýkání jejích produktů vede u mužů ke zvýšené mortalitě spermií nesoucích chromosom X (a pak se jim budou rodit převážně synové). Společnost najala pokusnou osobu, která dva roky intenzivně žvýkala její produkty; poté bylo provedeno vyšetření jeho spermatu. Ve zveřejněné zprávě společnost uvádí, že provedla test na přítomnost chromosomu X u 10000 spermií pokusné osoby, a zjistila přítomnost chromosomu X v 5001 případě, tzn. nepřítomnost v 4999 případech. Společnost na základě toho konstatuje, že shoda s očekávaným poměrem 1:1 je dostatečně jasná a že tedy její produkty jsou z tohoto hlediska zcela neškodné. Co k tomu můžeme říci jako statistici?

Odhlédněme nyní od toho, že pozorování nebylo nejlépe naplánováno, chybí kontrola, jedná o vliv na jediného člověka, a pokusme se vyhodnotit porovnání poměru spermií s chromosomem X k počtu spermií s chromosomem Y s očekávaným poměrem 1:1. Použijme χ^2- test dobré shody. Dostáváme $\chi^2 = 4.10^{-4}$, P=0.984. Výsledek testu je tedy neprůkazný, ale dosažená hladina významnosti se blíží jedné. To je velmi podezřelé. Co nám to říká? Předpokládejme, že poměr 1:1 je v základním souboru opravdu zachován. Potom s pravděpodobností více než 98% dostaneme v náhodném výběru spermií větší odchylku od poměru 1:1, než v našich datech. Nebo řečeno jinak: šance, že dostaneme takto dobrou shodu s poměrem 1:1 byla menší než 2%. Buď tedy měla společnost z pekla štěstí, nebo spíše výsledky zfalšovala tak, jak jí vyhovovaly. Jsou příliš dobré na to, aby to mohla být pravda.

Uvedený příklad je jistě vymyšlený. Nicméně ukazuje na to, jak se dá statistikou objevit falšování dat. Statisticky obdobný příklad je ovšem ve světové vědě znám: vyhodnotíme-li uvedeným postupem výsledky originálních Mendelových pokusů, zjistíme, že jsou z uvedeného hlediska „příliš dobré“, shoda se štěpnými poměry je nepravděpodobně dobrá (upozornil na to slavný statistik R.A. Fisher v roce 1936). Mendel sám ovšem o statistice netušil a nikde netvrdí, že se řídil pravidly pro náhodný výběr; naopak, konstatuje, že tam, kde byl výběr malý, přidával další individua. Závěr, hojně citovaný v době, kdy u nás byla genetika nazývána buržoasní pavědou, že „prelát Mendel falšoval data“, je tedy nesmyslný - pro zájemce doporučujeme článek T. Havránka (1986). Nicméně stačí zadat do vyhledávače „Mendel, Fisher, a to-good-to-be-true“, a s překvapením zjistíte, kolik lidí se tímto tématem po celém světě zabývá. Poučení pro nás je ovšem dvojí: když přinášíme zprávu o výsledku pokusu, popišme detailně, jak jsme k datům přišli a při použití historických dat nepředpokládejme, že data byla sebrána způsobem odpovídajícím statistickým zásadám.

Příkladová data

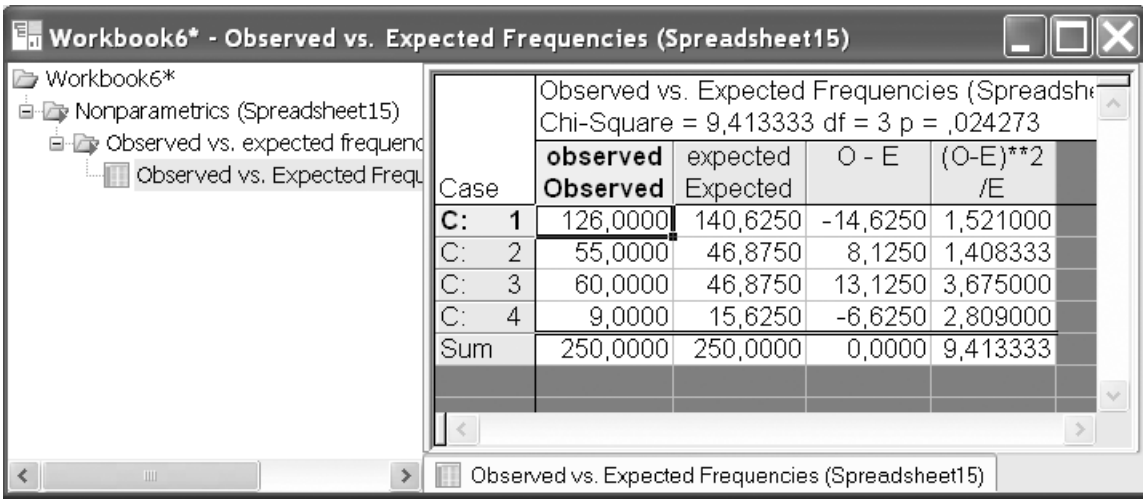
Všechna data pro tuto kapitolu jsou v listu *Chap2* souboru *biostat-data.xls*. Standardní test dobré shody budeme ilustrovat na dvou příkladech. První z nich byl již popsán spolu se vzorečkem pro výpočet χ^2 statistiky tohoto testu (semena hrachu žlutá vs. zelená a hladká vs. svraskalá). Údaje o počtu pozorovaných semen pro každou ze čtyř kategorií jsou v proměnné *Observed*, očekávané počty jsou vypočteny ve sloupci označeném *Expected*. Jako druhý příklad uvádíme následující výsledky: v první filiální generaci (AA x aa) bylo očekáváno, že všechna individua budou mít dominantní fenotyp. Mezi 2000 pozorovanými individui se vyskytla tři s recesivním fenotypem. Liší se výsledek od očekávaného? Pozorované četnosti jsou v proměnné *Obs_ind* (hodnoty 1997 a 3), očekávané četnosti (2000 a 0) jsou v proměnné *Exp_ind*.

Pro ilustraci způsobu testování Hardy-Weinbergovy rovnováhy (a nezávislého výpočtu pravděpodobnosti chyby prvního typu) použijeme následující příklad: v populaci diploidních savců bylo nalezeno 15 jedinců genotypu *AA*, 20 jedinců genotypu *Aa*, a 77 jedinců s genotypem *aa*. Testujte shodu s Hardy-Weinbergovou rovnováhu. Pozorované počty jsou uloženy v proměnné *Individuals*, očekávané počty (které jsou vypočteny na základě Hardy-Weinbergova modelu, viz samostatný list *Chap2-HW* zobrazující postup výpočtu) jsou v proměnné *Exp_indiv*.

Jak postupovat v programu Statistica

Test dobré shody

V menu zvolíme příkaz *Statistics* | *Nonparametrics* a z nabídnutého seznamu vybereme položku *Observed versus expected* X^2. V dialogovém okně nejprve zvolíme tlačítko *Variables* a v následně zobrazeném okně vybereme proměnnou *Observed* v levém seznamu a proměnnou *Expected* v pravém. Po návratu do prvního dialogového okna zvolíme tlačítko *Summary* a program zobrazí výsledky v oknu *Workbook*.



Vlastní tabulka představuje detaily výpočtu χ^2 statistiky, shrnutí je v nejhořejší části nad tabulkou. Hodnota testové statistiky je (po zaokrouhlení) 9.41 a porovnáním s χ^2 distribucí se 3 stupni volnosti odhadneme signifikanci testu p = 0.0243. Nulovou hypotézu o shodě s poměrem 9:3:3:1 tedy zamítáme na hladině $\alpha = 0.05$.

Tento test si ještě procvičíme s druhým příkladem, ve kterém nulová hypotéza tvrdí, že všichni jedinci musí mít dominantním fenotyp. To znamená, že pravděpodobnost nalezení recesivního fenotypu je nulová. Z toho vyplývá, že nalezení byť jediného jedince s recesivním fenotypem naši nulovou hypotézu jednoznačně vyvrací. Podívejme se ale, co se stane, pokud proměnné *Obs_ind* a *Exp_ind* použijeme stejným způsobem jako v předchozím příkladě. Zde jsou výsledky zobrazené programem Statistica.

Observed vs. Expected Frequencies (biostat-data)
Chi-Square = ,0045000 df = 1 p = ,946516

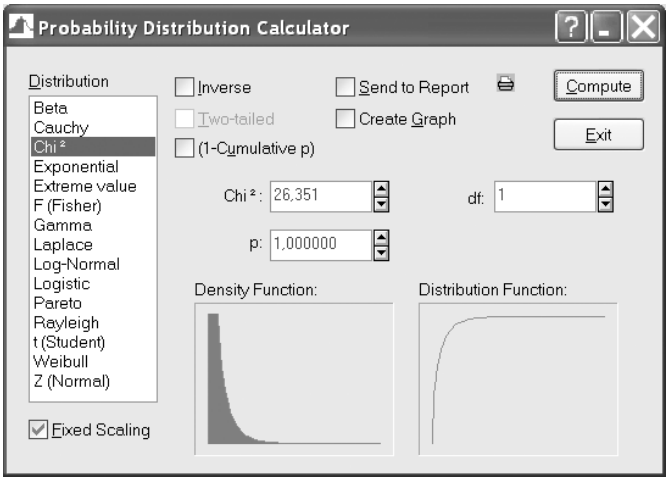
Case	observed Obs_ind	expected Exp_ind	O - E	(O-E)**2 /E
C: 1	1997,000	2000,000	-3,00000	0,004500
C: 2	3,000	0,000	3,00000	0,000000
Sum	2000,000	2000,000	0,00000	0,004500

Tyto výsledky naznačují, že pravděpodobnost chyby, které bychom se mohli dopustit zamítnutím nulové hypotézy, je velmi vysoká a tedy že naše data jsou ve shodě s touto hypotézou. To ale odporuje logické úvaze, tak jak jsme ji popsali výše. Důvodem rozporu je to, že výpočet χ^2 statistiky je v tomto případě nesprávný. Pokud se v zobrazené tabulce podíváme do řádky označené *C: 2*, tvrdí nám poslední sloupec, že když vydělíme hodnotu 3.00 nulou (tj. hodnotou E), dostaneme nulu. Ve skutečnosti je ale výsledkem dělení nulou hodnota kladného nekonečna a pravděpodobnost, že takovou hodnotu testové statistiky si náhodou „vytáhneme“ z χ^2 distribuce je nulová. Toto je prakticky jediný případ, kdy můžeme opravněně napsat p=0.

Způsob zadání pozorovaných a očekávaných četností jedinců v příkladu na Hardy-Weinbergovu rovnováhu je shodný s předchozími příklady, liší se ale způsob, jakým k očekávaným četnostem dospějeme, a v důsledku i tím, jak odhadneme významnost testu. Očekávané počty nejsou jen důsledkem námi *a priori* formulované hypotézy, ale jde o výpočet založený na relativních četnostech dvou alel daného genu, přičemž tyto četnosti jsou odvozeny z pozorovaných dat. Připravíme se tak o další stupeň volnosti a musíme proto, se třemi kategoriemi genotypu, porovnávat získanou hodnotu testové statistiky (26.351) ne s χ^2 distribucí se dvěma stupni volnosti (jak to automaticky provádí Statistica), ale jen s jedním stupněm. To lze provést postupem popsaným v následující sekci.

Výpočet kritických hodnot a průkaznosti testové statistiky

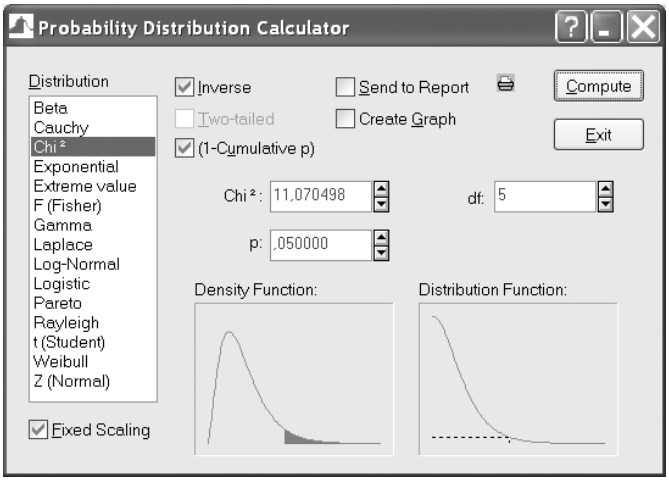
Zvolíme v menu příkaz *Statistics* | *Probability Calculator* | *Distributions* a v dialogovém oknu nejprve zvolíme na levé straně typ distribuce, pro náš příklad *Chi*^{2}. Pokud chceme – jako v příkladu na Hardy-Weinbergovu rovnováhu, započatém v předchozí sekci – spočíst dosaženou hladinu významnosti, zadáme spočtenou hodnotu statistiky do políčka uvedeného slovem *Chi*^{2}*:* a v následujícím políčku změníme počet stupňů volnosti (*df:*) (pro nás na hodnotu 1). V případě práce v českém prostředí je třeba statistiku zadat s desetinnou čárkou (ne tečkou). Již při vyplňování hodnoty statistiky se vyprázdní níže položené políčko označené *p:* a po výběru tlačítka *Compute* se v něm zobrazí spočtená hodnota 1.000000.



Hodnota 1.000 ale odpovídá pravděpodobnosti, že si z referenční distribuce náhodně “vytáhneme” zadanou hodnotu nebo hodnotu menší. My ale potřebujeme pravděpodobnost komplementární, tj. že dostaneme takovou nebo větší hodnotu. Její výpočet není v tomto případě obtížný, obecně si ale dopočet do jedničky můžeme usnadnit zaškrtnutím volby *(1-Cumulative p)*. Tento příklad také ilustruje jinou věc, na kterou si při prezentaci výsledků z programu Statistica musíme dávat pozor. Získaná hodnota je sice zobrazena jako přesná nula, ale to jen proto, že Statistica zobrazuje výsledky s přesností šesti desetinných míst a na

tuto přesnost získaný výsledek zaokrouhluje. Tedy například hodnota 0.0000004 bude zobrazena jako 0.000000. Výslednou hodnotu *p* tedy musíme prezentovat nikoliv jako nulu, ale *p<0.000001*, alternativně *p<1.0e-6*, nebo spíše $p<10^{-6}$. Až budeme podávat o výsledcích zprávu, určitě se nespokojíme s konstatováním, že populace není v H-W rovnováze, ale uvedeme i která kategorie je v nadbytku/ která chybí - to nám umožní uvažovat o mechanismech, které H-W rovnováhu narušují.

Stejné dialogové okno můžeme použít také pro výpočet kritické hodnoty, tedy např. pokud bychom chtěli spočíst kritickou hodnotu χ^2 distribuce s 5 stupni volnosti na pětiprocentní hladině významnosti, zadáme hodnotu *5* do políčka *df*, hodnotu *0.05* (nebo *0,05* v případě aktivního českého prostředí) do políčka *p* (v případě, kdy je zaškrtnuta volba *(1-Cumulative p)*, jinak zadáme *0.95*) a po zmáčknutí tlačítka *Compute* je kritická hodnota zobrazena v políčku *Chi*^{2}.



Jak postupovat v programu R

Při importu příkladových dat do programu R z listu *Chap2* doporučujeme naimportovat každou dvojici proměnných do samostatného datového rámce (tj. *chap2.a*, *chap2.b* a *chap2.c*), protože se tyto proměnné liší počtem řádků. Ve skutečnosti ale druhou z proměnných v každé dvojici nebudeme potřebovat, protože v programu R se naše očekávání (nulová hypotéza) nezadává jako očekávané frekvence, ale jako pravděpodobnosti pro jednotlivé kategorie.

Prvý příklad vyřešíme tedy následovně (první příkaz jen ukazuje obsah používané proměnné *Observed*):

```
> chap2.a$Observed
[1] 126  55  60   9
> with( chap2.a, chisq.test( Observed, p=c(9/16,3/16,3/16,1/16)))
        Chi-squared test for given probabilities
data:  Observed
X-squared = 9.4133, df = 3, p-value = 0.02427
```

Druhý příklad spočteme obdobně a všimneme si, že program R se nedopouští chyby, kterou dělá program Statistica, hodnotu testové statistiky správně odhadne jako +nekonečno:

```
> chap2.b$Obs_ind
[1] 1997    3
> with( chap2.b, chisq.test( Obs_ind, p=c(1,0)))
        Chi-squared test for given probabilities
data:  Obs_ind
X-squared = Inf, df = 1, p-value < 2.2e-16
Warning message:
...
```

A nakonec spočtem příklad na ověřování toho, zda populace je v Hardy-Weinbergově rovnováze. Tentokrát si pomůžeme k získání pravděpodobností očekávaných podle nulové hypotézy již vypočtenými očekávanými četnostmi:

```
> chap2.c
  Individuals Exp_indiv
1          15      5.58
2          20     38.84
3          77     67.58
> hw <- with( chap2.c, chisq.test(Individuals,p=Exp_indiv/sum(Exp_indiv)))
> pchisq(hw$statistic, 1, lower.tail=F)
   X-squared
2.841857e-07
```

V programu R získáváme přesnější odhad pravděpodobnosti prvního druhu ($p=2.8.10^{-7}$) Ve výše uvedeném kódu jsme si výsledek vrácený funkci *chisq.test* uložili do proměnné *hw* a to nám umožňuje nejen předání přesnější hodnoty χ^2 statistiky funkci *pchisq*, ale třeba i pro zobrazení residuálních frekvencí:

```
> hw$residuals
[1]  3.987804 -3.023022  1.145887
```

Pokud bychom naopak chtěli pro danou pravděpodobnost spočíst kritickou hodnotu, může použít funkci *qchisq*, například takto:

```
> qchisq(0.95,5)
[1] 11.0705
```

Obdobné dvojice funkcí existují i pro další typy distribucí, jako je například normální distribuce (*pnorm* a *qnorm*), t-distribuce (*pt* a *qt*) nebo F-distribuce (*pf* a *qf*), potřebné parametry se ale liší, podle toho, jakými parametry jsou dané distribuce určeny (například průměr a směrodatná odchylka u normální distribuce).

Popis analýz v článku

Methods

The difference of the observed frequency of the four phenotypes from the expected 9:3:3:1 ratio was tested using a goodness-of-fit test based on χ^2 statistic.

Results

Frequencies of the four distinguished phenotypes differed significantly ($\chi^2_3=9.41, p=0.0243$) from the expected ratio. The largest difference was found in the frequency of the fourth category.

Doporučená četba

Zar (2007), pp. 40-60, Sokal-Rohlf (1981), pp. 692-730, Quinn & Keough (2002), pp. 32-57 (testování hypotéz), p. 381 (test dobré shody)

Havránek T. (1986): Gregor Mendel a experimentální data. - Vesmír 65: 331-333.

3 Kontingenční tabulky

Pro sledování závislosti dvou nebo více kategoriálních proměnných (proměnných na nominální stupnici) používáme kontingenční tabulky (*contingency tables*). V praxi nejčastěji studujeme závislost dvou znaků, kterou hodnotíme pomocí dvourozměrných tabulek.

Dvourozměrné tabulky

Dvourozměrné tabulky nám umožňují popsat závislost dvou kategoriálních proměnných.

Příklady:

1. Sledujeme závislost výskytu určitého druhu na typu obhospodařování. Máme 45 pokusných ploch: z nich je 15 koseno v létě, 15 na podzim a 15 jsou nekosené kontroly. Druh se vyskytl ve 12 plochách kosených v létě, ve 13 plochách kosených na podzim a v 6 plochách nekosených. Je průkazný rozdíl ve výskytu druhu v plochách různě obhospodařovaných? Přesněji: můžeme zamítnout nulovou hypotézu, že pravděpodobnost výskytu druhu nezávisí na typu obhospodařování?

2. Ve 100 plochách náhodně rozmístěných v porostu byl sledován výskyt druhů. Dva druhy se vyskytly společně 45-krát, 15-krát se vyskytl prvý druh samostatně, 5-krát druhý druh samostatně a 35 ploch neobsahovalo žádný druh. Můžeme zamítnout nulovou hypotézu, že se druhy vyskytují na sobě nezávisle? Obdobná je úloha, kdy u **n** individuí savců (náhodný odchyt) sledujeme výskyt dvou druhů parazitů a ptáme se, zda se paraziti vyskytují na sobě nezávisle nebo zda u nich můžeme pozorovat vzájemnou závislost výskytu.

3. Ze 100 semen hlohu bylo 50 kontrolních a 50 bylo dáno sežrat kohoutovi a sebráno z výkalů po projití zažívacím traktem. Z kontrolních semen vyklíčilo 8 a ze semen, která prošla zažívacím traktem kohouta, vyklíčilo 28. Má projití traktem vliv na klíčivost?

4. V etologickém pozorování vybraných primátů bylo sledováno, zda se liší přístup samců a samic k získání pochoutky zavěšené na stromě. Každý jedinec byl na základě minutového pokusu zařazen do tří kategorií: (1) použil k získání pochoutky klacek, který byl k dispozici, (2) vylezl na strom a tím pochoutku získal a (3) pochoutku nezískal. Pokus byl proveden na 100 samcích a 100 samicích.

Výsledky a zpracování tohoto příkladu jsou v Obr. 3-1.

Obecně vypadá tabulka 2 x 3 takto:

			FAKTOR 2		
		Kategorie 1	**Kategorie 2**	**Kategorie 3**	**Součty**
FAKTOR 1	**Kategorie 1**	f_{11}	f_{12}	f_{13}	R_1
	Kategorie 2	f_{21}	f_{22}	f_{23}	R_2
	Součty	C_1	C_2	C_3	**n**

Tab. 3-1 Kontingenční tabulka 2x3

Je zcela lhostejné, která proměnná tvoří sloupce (*columns*) a která řádky (*rows*).

Počet sloupců označujeme *c*, počet řádků *r* (z angl. *column* a *row*). f_{ij} je četnost v i-tém řádku a j-tém sloupci, tj. počet případů, kdy sledované individuum bylo klasifikováno do kategorie *i* v prvé klasifikaci (v prvém faktoru) a zároveň do kategorie *j* v druhém faktoru.

H_0: Úspěšnost a způsob získání pochoutky ve studované populaci nezávisí na pohlaví.
H_A Úspěšnost a způsob získání pochoutky ve studované populaci závisí na pohlaví.
Pozorované frekvence jsou v tabulce uváděny spolu s frekvencemi očekávanými při platnosti H_0 (v závorkách).

	Řešení			
Pohlaví	Klacek	Strom	nezískal	**Celkem**
Samec	32	43	25	100
	(23.5)	(54.0)	(22.5)	(=R1)
Samice	15	65	20	100
	(23.5)	(54.0)	(22.5)	(=R2)
Celkem	47	108	45	200
	(=C1)	(=C2)	(=C3)	(=n)

$$\chi^2 = \sum\sum \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} = \frac{(32-23.5)^2}{23.5} + \frac{(43-54.00)^2}{54.00} + \frac{(25-22.5)^2}{22.5} +$$

$$\frac{(15-23.5)^2}{23.5} + \frac{(65-54.0)^2}{54.0} + \frac{(20-22.5)^2}{22.5}$$

=3.074+2.241+0.278+3.074+2.241+0.278 = 11.186
ν=(r-1)(c-1)=(2-1)(3-1)=2; p=0.0037 , zamítáme tedy H_0.

Obr. 3-1 Vyhodnocení kontingenční tabulky popsané v příkladu 4.

$$R_i = \sum_{j=1}^{c} f_{ij}$$

Vz. 3-1

jsou součty v řádcích, obdobně C_j součty ve sloupcích. Součty v řádcích a ve sloupcích se nazývají marginální četnosti; *n* je celkový počet pozorování.

Pro zpracování kontingenčních tabulek používáme opět χ^2 test. V tomto případě testové kritérium vychází z obecného vzorce:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

Vz. 3-2

kde f a $\hat{f}$ jsou pozorované a očekávané četnosti. Používáme přitom pozorované a očekávané četnosti ve všech polích kontingenční tabulky. Vzorec má tedy podobu:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$$

Vz. 3-3

Očekávané četnosti odhadujeme na základě marginálních četností. Pravděpodobnost, že individuum bude klasifikováno do kategorie *i* v prvém faktoru, odhadneme takto: $P_i = R_i / n$; obdobně v druhém faktoru $Pj = C_j / n$. Pokud jsou oba faktory nezávislé, platí, že pravděpodobnost společného výskytu dvou **nezávislých** jevů je součinem jejich pravděpodobností, tj. $P_{ij} = P_i P_j$. Očekávaná četnost je pak součinem pravděpodobnosti a celkového počtu sledovaných případů:

$$\hat{f}_{ij} = P_{ij}n = \frac{R_i C_j}{n}$$

Vz. 3-4

Takto odhadnuté četnosti dosadíme do vzorce Vz. 3-3. Celý postup je zřejmý z příkladu v Obr. 3-1. Musíme také určit počet stupňů volnosti: *DF = (c - 1)(r - 1)*; z marginálních součtů jsme vždy schopni dopočítat četnost v posledním políčku řádky nebo sloupce na základě hodnot předcházejících.

Yatesova korekce: Někteří autoři doporučují užívat tzv. Yatesovu korekci na spojitost (kontinuitu), pokud jsou očekávané frekvence nízké (platí obecně pro χ^2, lze aplikovat i pro kontingenční tabulky). Pokud tuto korekci použijeme, má vzorec podobu

$$\chi^2 = \sum_{i=1}^{k} \frac{(\left|f_i - \hat{f}_i\right| - 0.5)^2}{\hat{f}_i}$$

Vz. 3-5

Takto dostáváme velmi konzervativní test: to znamená, že pravděpodobnost chyby prvního druhu je nižší než stanovené α, ale tím je pravděpodobnost chyby druhého druhu vyšší. O této korekci se uvažuje, pokud některá očekávaná četnost klesne pod hodnotu 5, názory autorů se ale různí. Rozdělení testové statistiky má jen *přibližně* χ^2-rozdělení. χ^2 je totiž rozdělení spojité náhodné proměnné, zatímco statistika χ^2 je zde počítána z frekvencí, tj. počtu případů - tedy z proměnné, která je diskrétní (nespojitá). Pokud se jedná o vysoká čísla nebo sčítáme hodně členů, přiblížení je velmi dobré; pokud sčítáme málo členů a každý sčítanec je počítán na základě malých frekvencí, je přiblížení horší.

Likelihood-ratio test: Místo klasického χ^2 lze užít ke stejným účelům též G-test, nazývaný též test poměrem věrohodností (*the log-likelihood ratio*). Vychází z jiných předpokladů, užívá jiný vzorec, ale výsledná statistika má také přibližně χ^2 rozdělení. Vzorec je

$$G = 2 \cdot \left[\sum_i \sum_j f_{ij} \ln f_{ij} - \sum_i R_i \ln R_i - \sum_j C_j \ln C_j + n \ln n \right]$$

Vz. 3-6

nebo

$$G = 4.60517 \left[\sum_i \sum_j f_{ij} \log f_{ij} - \sum_i R_i \log R_i - \sum_j C_j \log C_j + n \log n \right]$$

Vz. 3-7

Příklad v Obr. 3-2 vychází z týchž dat jako příklad v Obr. 3-1 (ukazuje vyhodnocení příkladu 4). Vidíme, že jsme pomocí zcela odlišného vzorce dostali přibližně stejnou hodnotu testové statistiky jako pomocí klasického vzorce.

H0: Úspěšnost a způsob získání pochoutky ve studované populaci nezávisí na pohlaví.
HA Úspěšnost a způsob získání pochoutky ve studované populaci závisí na pohlaví.
Pozorované frekvence jsou uvedeny v tabulce.

		Řešení		
Pohlaví	Klacek	Strom	nezískal	***Celkem***
Samec	32	43	25	*100*
Samice	15	65	20	*100*
Celkem	*47*	*108*	*45*	*200*

$G = 4.60517\ (\sum\sum f_{ij} \log f_{ij} - \sum R_i \log R_i - \sum C_j \log C_j + n \log n) =$
$4.60517[(32)(1.50515)+(43)(1.63347)+(25)(1.39794) +...+(20)(1.30103)-(100)(2.000)-(100)(2.000) -$
$(47)(1.6721)-(108)(2.0334)-(45)(1.65321) +(200)(2.30103)]= 4.60517(2.46685) = 11.360$ s DF=2, $p=0.0034$, zamítáme proto H_0.

Obr. 3-2 Vyhodnocení příkladu 4 pomocí G-statistiky.

Čtyřpolní tabulky

Čtyřpolní tabulky (2x2 *tables*) jsou nejjednodušším příkladem dvourozměrných kontingenčních tabulek, použili bychom je například pro řešení příkladů 2 a 3 na začátku této kapitoly. Základní vzorec Vz. 3-3 lze zjednodušit dosazením za $\hat{f}_{ij}$ ve Vz.3-3 podle rovnice Vz.3-4. Dostáváme výpočetní tvar, který se používá

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{C_1 C_2 R_1 R_2}$$

V mnoha učebnicích bývají pro čtyřpolní tabulky tradičně užívány symboly *a, b, c, d* namísto $f_{11}, f_{12}, f_{21}, f_{22}$, a symboly *m, n, r, s* pro marginální součty C_1, C_2, R_1, R_2.

Statistika χ^2 pro čtyřpolní tabulku se porovnává s χ^2 distribucí s jedním stupněm volnosti. Je tedy nejvíce ovlivněna nespojitým charakterem dat, především pokud jsou očekávané frekvence nízké. Pro čtyřpolní tabulku je ale možné spočítat **Fisherův exaktní test**, který (na základě kombinatoriky) spočte přímo pravděpodobnost, s jakou dostaneme za předpokladu platnosti nulové hypotézy tabulku stejně nebo více odlišnou od nulové hypotézy.

Míry těsnosti vazby

Doposud jsme se ptali, zda jsou sledované kategoriální proměnné nezávislé; odpověď na tuto otázku dává statistický test. Můžeme se ale také ptát, jak silná je závislost mezi sledovanými proměnnými. Rozdíl mezi oběma přístupy si ukážeme na příkladu použití čtyřpolních tabulek pro zjišťování mezidruhových vazeb. Uspořádání výzkumu odpovídá příkladu 2 na načátku této kapitoly – liší se jenom tím, že jsme do téhož porostu v prvním příkladě náhodně umístili 1500 ploch, ve kterých jsme zaznamenávali přítomnost dvou vybraných druhů. Ve druhém případě bylo ploch pouze 150. Získali jsme následující výsledky:

		Druh 1		
		přítomen	nepřítomen	součet
Druh 2	přítomen	100	200	300
	nepřítomen	200	800	1000
	součet	300	1000	1300

Tab. 3-2

Pro tuto tabulku je $\chi^2 = 23.11$ a $p<0.0001$; zamítáme tedy jasně nulovou hypotézu, očekávaná četnost společných výskytů je 300 x 300 / 1500 = 60. Druhy se spolu tedy vyskytují častěji, než bychom očekávali při absenci jejich vztahu, a mluvíme proto o kladné vazbě, v opačném případě bychom mluvili o záporné vazbě. Připomeňme, že kladná vazba ještě neznamená aktivní pozitivní působení jednoho druhu na druhý, jak ukážeme v závěru této kapitoly.

Srovnejme nyní shora uvedenou tabulku s jinou, podobnou, založenou na pouhých 150 plochách:

		Druh 1		
		přítomen	nepřítomen	součet
Druh 2	přítomen	10	20	30
	nepřítomen	20	80	100
	součet	30	100	130

Tab. 3-3

Tato tabulka má stejné relativní četnosti, ale desetkrát nižší počet pozorování. Hodnota $\chi^2 = 2.311$ je tedy desetkrát nižší, $p = 0.128$ a nulovou hypotézu nemůžeme zamítnout. Přitom je byly obě tabulky získány výzkumem ve stejné porostu, a liší se jen počtem čtverců. Hodnota χ^2 (a tím i dosažená hladina významnosti) se tedy i při stejné intenzitě závislosti mění s celkovým počtem pozorování. To je v pořádku: čím více máme pozorování, tím menší je pravděpodobnost, že ke stejné relativní odchylce od stavu očekávaného při nezávislosti může dojít náhodou. A čím víc máme pozorování, tím mám v ruce více důkazů, které umožňují zamítnout nulovou hypotézu. Vidíme tedy opět, že síla testu roste s počtem pozorování. Hodnota χ^2 ale nemůže být mírou intenzity vazby. K tomu účelu se používají statistiky, jejichž hodnoty při stejných relativních odchylkách od náhodnosti nezávisí na počtu pozorování. První skupinu těchto statistik tvoří ty, které jsou založeny na poměru $f_{11}f_{22}$: $f_{21}f_{12}$. Tento poměr se někdy nazývá Y. Je-li $f_{11}f_{22} > f_{21}f_{12}$, pak je vazba kladná, je-li $f_{11}f_{22} < f_{21}f_{12}$, pak je vazba záporná. Nejčastější je užíván koeficient

$$Q = \frac{f_{11}f_{22} - f_{21}f_{12}}{f_{11}f_{22} + f_{21}f_{12}},$$

Vz. 3-8

který je funkcí **Y**. Může nabývat hodnot od -1 do +1. Hodnota koeficientu Q je 0,333 , shodně pro obě tabulky. V témže rozmezí se pohybuje koeficient

$$V = \frac{(f_{11}f_{22} - f_{12}f_{21})}{\sqrt{C_1C_2R_1R_2}}$$

Vz. 3-9

Jak je vidět z porovnání s výpočetním vzorcem pro χ^2, tato hodnota je vlastně

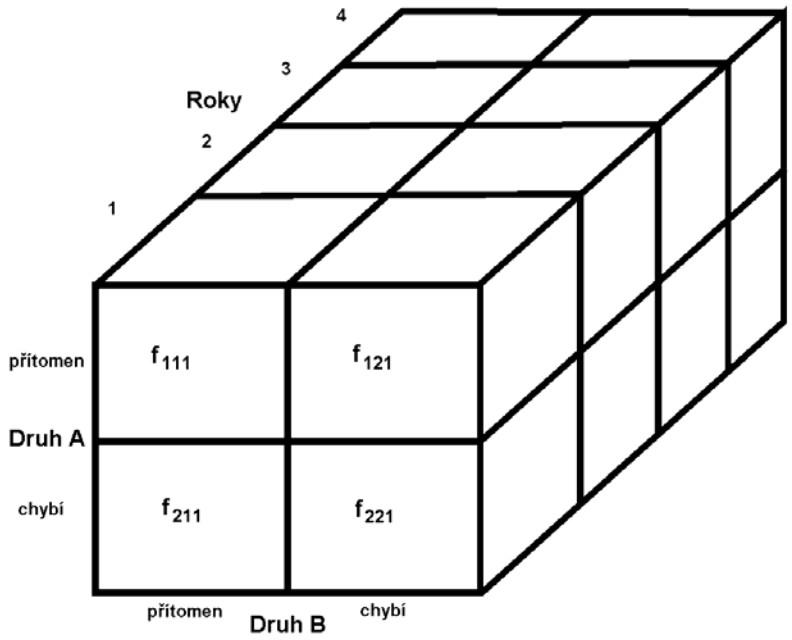
$$\sqrt{\frac{\chi^2}{n}}$$

Vz. 3-10

s příslušným znaménkem, tj. kladným pro kladnou vazbu, záporným pro zápornou vazbu. Také pro tento koeficient je hodnota shodná pro obě tabulky, a to 0.1333.

Vícerozměrné tabulky

Někdy potřebujeme studovat závislost více kategoriálních proměnných. Například jsme studovali vzájemnou závislost dvou druhů ve více letech. Dostáváme potom tabulku (či spíše kostku), jaká je na Obr. 3-3.



Obr. 3-3. Příklad trojrozměrné kontingenční tabulky

Zde je nulová hypotéza: tři sledované proměnné byly na sobě nezávislé. To znamená, že frekvence žádného druhu se nemění s lety a druhy jsou navzájem nezávislé. Očekávané hodnoty (za předpokladu platnosti nulové hypotézy) vypočteme jako *n*-násobek pravděpodobnosti současného výskytu tří nezávislých jevů, tedy

$$\hat{f}_{ijl} = P_i P_j P_l n = R_i C_j T_l / n^2$$

Vz. 3-11

R_i je celkový počet jednotek nabývajících hodnoty *i* v první proměnné, C_j je počet jednotek nabývajících hodnoty *j* v druhé proměnné, obdobně T_l v třetí proměnné; v našem případě je f_{111} počet ploch snímkovaných v prvém roce sledování, které obsahovaly druh 1 i druh 2. Očekávaná hodnota je součinem počtu ploch snímkovaných v prvém roce, celkového počtu ploch, kde se vyskytl druh 1 a celkového počtu ploch, kde se vyskytl druh 2, lomená druhou mocninou celkového počtu ploch, sledovaného ve všech letech. Můžeme ovšem testovat i jiné hypotézy, např. předpokládáme, že frekvence druhů se mění s časem, ale druhy navzájem jsou nezávislé.

Kromě interakcí prvního řádu (dvojice proměnných) můžeme očekávat i interakce vyššího řádu. Příkladem takové interakce by v našem případě bylo, kdyby dva sledované druhy byly na sobě statisticky závislé, ale intenzita závislosti by se měnila s časem (případně by na začátku vykazovaly kladnou vazbu a na konci zápornou). K testování těchto hypotéz slouží tzv. log-lineární modely, které jsou speciálním případem zobecněných lineárních modelů. Modelem zde rozumíme strukturu závislostí; hledáme model, který je nejjednodušší a přitom dostatečně vysvětluje pozorovaná data. Název pochází z užívané logaritmické transformace: závislost

$$\hat{f}_{ijl} = R_i C_j T_l / n^2$$

můžeme vyjádřit jako

$$\log \hat{f}_{ijl} = \log R_i + \log C_j + \log T_l - 2 \log n \, .$$

Vz. 3-12

V praxi postupujeme tak, že testujeme data proti nulové hypotéze dané nejjednodušším modelem; v případě, že tento model nemůžeme zamítnout, znamená to, že nejsme schopni prokázat jakoukoliv závislost mezi proměnnými. Pokud tento model můžeme zamítnout, testujeme postupně složitější a složitější modely; za základ pro interpretaci potom bereme nejjednodušší model, který nemůžeme zamítnout. V případě, že porovnáváme více proměnných, existují různé strategie výběru onoho nejjednoduššího modelu.

Statistická a kauzální závislost

Jak vidíme z příkladů na začátku této kapitoly, dvourozměrné kontingenční tabulky (ale obdobně to platí i pro vícerozměrné tabulky) se užívají pro studium závislostí kategoriálních proměnných jak v případě, že jedna z proměnných je manipulovaná, a chceme prokázat její vliv na druhou kategoriální proměnnou, tak v případě observačních studií, kdy žádná z proměnných není manipulovaná, ale předpokládáme, že jedna může být příčinou a druhá následkem, a konečně v případě, kdy studujeme dvě stejnocenné proměnné (příklad mezidruhových vazeb). Zatímco vypočetní postup a výsledky statistického testu jsou ve všech případech stejné, interpreatce výsledků se může výrazně lišit.

Začněme příkladem: porovnejme příklad 3 ze začátku této kapitoly (100 semen hlohu, 50 kontrolních, 50 prošlo zažívacím traktem kohouta a sledujeme klíčivost) s jiným uspořádáním. V terénu bylo sebráno 50 semen hlohu, která ležela pod keřem a zřetelně nebyla sežrána, a 50 semen, která byla nalezena ve výkalech bažantů. Co nám říká průkazný výsledek prvého a co průkazný výsledek druhého testu? V prvním případě šlo o manipulativní experiment - ze 100 semen bylo náhodně vybráno 50, která byla dána sežrat kohoutovi. Pokud se liší klíčivost mezi sežranými a nesežranými, může být **příčinou** pouze projití zažívacím traktem (nakolik projití zažívacím traktem kohouta v chovu dobře simuluje žraní bažantů v přírodě je ovšem otázkou další interpretace).

Naproti tomu ve druhém případě 50 semen, která bažanti sežrali, zdaleka nemusí být náhodným výběrem semen z většího souboru. Právě naopak - můžeme předpokládat, že dobře zralé plody budou bažanty více lákat k sežrání, zatímco neduživé plody spadnou na zem nepovšimnuty. Je zde tedy velká pravděpodobnost, že se dva srovnávané soubory lišily již předtím, než se mohl projevit vliv bažantů. Můžeme tedy v případě statisticky významného výsledku testu konstatovat, že se klíčivost semen nalezených ve výkalech lišila od klíčivosti semen na zem spadlých; hypotéza, že semena zvýšila svou klíčivost projitím zažívacího traktu bažantů, je ale jen jedním z možných vysvětlení.

Uvedený příklad ilustruje důležitost manipulativních experimentů, kdy experimentátor uměle manipuluje hodnotu jedné proměnné a sleduje změny druhé proměnné. Jen tak lze dokázat kauzální závislost. Naproti tomu společný výskyt dvou znaků na jednom objektu (v tomto případě klíčivost a průchod trávicím traktem bažanta) je pouze závislostí statistickou, za kterou může a nemusí být kauzální vazba. Velmi často jsou dvě sledované proměnné ovlivněny jinou, třetí, kterou nesledujeme (*confounding variable*). Protože manipulativní experimenty jsou, zvláště v některých terénních oborech, dosti obtížně

proveditelné, je otázka důkazu kauzálních závislostí a jejich odlišení od pouhé statistické závislosti ožehavým problémem mnoha výzkumů.

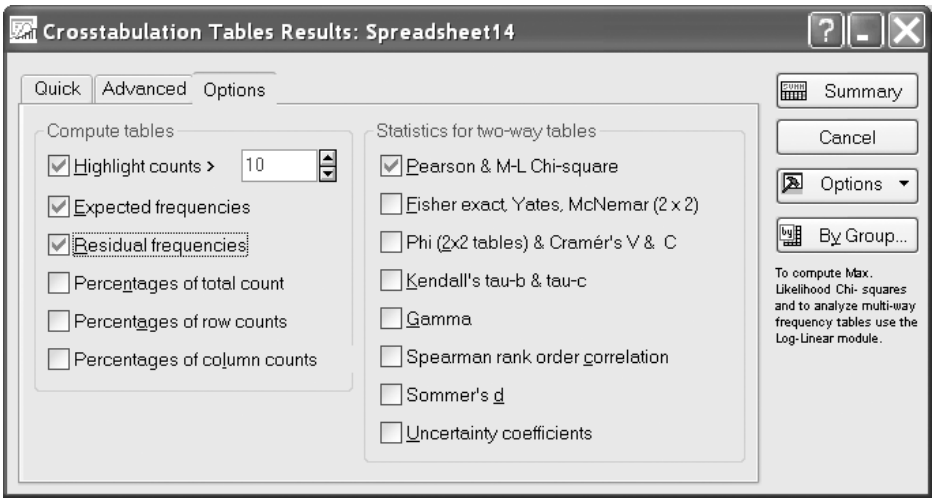
Příkladová data

Analýzu kontingenčních tabulek ukážeme na příkladu č. 4 popsaném na začátku této kapitoly a také v Obr. 3-1 a 3-2. Tato data jsou uložena v listu *Chap3* souboru *biostat-data.xls*. Data nejsou uspořádána do dvourozměrné tabulky, ale jsou uspořádána lineárně, s příslušností do řádků a sloupců určenou dvěma kategoriálními proměnnými *Gender* a *Solution*. Počet výskytů jednotlivých kombinací pohlaví a řešení problému je uveden ve třetí proměnné *Count*. Pro ilustraci výpočtů u čtyřpolních tabulek užíváme data z Tab. 3-2.

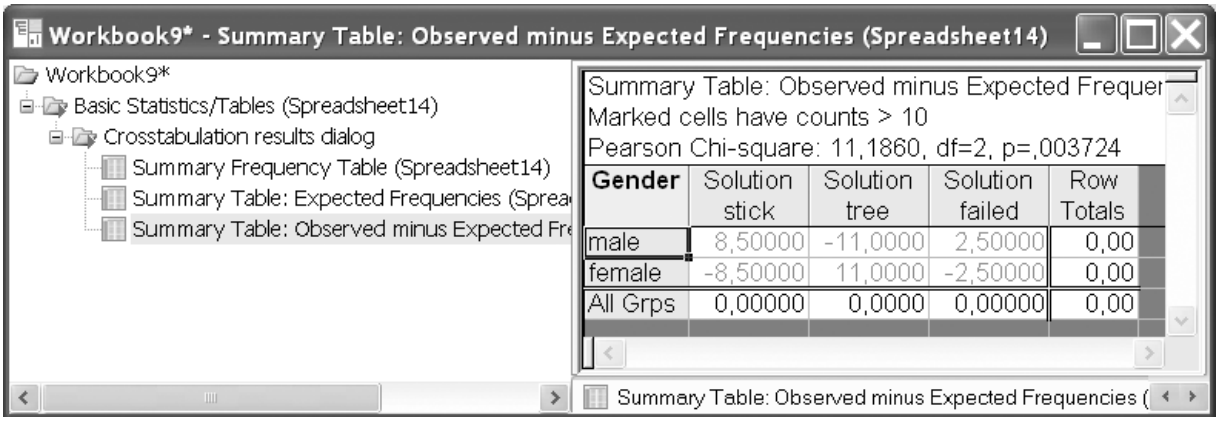
Jak postupovat v programu Statistica

Chceme-li ve Statistice analyzovat kontingenční tabulky, tak s výjimkou čtyřpolních tabulek (2 x 2) je musíme zadat v expandované podobě, kdy každé políčko v tabulce je představováno jedním řádkem ve spreadsheetu a dvě (či více, pro vícerozměrné tabulky) kategoriální proměnné popisují, do jakého řádku / sloupce daná buňka patří. V této podobě jsou příkladová data již přítomna v listu *Chap3*. Obecné kontigenční tabulky zadáme a analyzujeme následujícím způsobem.

V menu *Statistics* vybereme příkaz *Basic Statistics/Tables*, z nabídky vybereme *Tables and banners* a zvolíme tlačítko *OK*. V dialogovém okně vybereme tlačítko *Specify tables (select variables)*, každý faktor popisující hrany tabulky (zde *Gender* a *Solution*) zvolíme v samostatném seznamu (pro náš příklad tedy jen v prvním a druhém seznamu) a zmáčkneme tlačítko *OK*. Počty případů zadáme pomocí tlačítka váhy (blízko pravého dolního okraje dialogu, se závažíčkem a písmenem *w*). V novém zobrazeném dialogovém okně doporučujeme zvolit *Use weights for this Analysis/Graphs only*, pak zvolíme v políčku *Weight variable* proměnnou s počty (*Count*) a musíme ještě tuto proměnnou aktivovat změnou volby v oblasti *Status* z *Off* na *On*. Po uzavření dialogového okna tlačítkem *OK* nás Statistica ještě upozorní na aktivace vah. Po návratu to dialogového okna *Crosstabulation Tables* pokračujeme tlačítkem *OK*, pak se zobrazí okno *Crosstabulation Tables Results*. Je vhodné nejprve přepnout na záložku *Options*. Tam je nutné – pro výpočet a zobrazení χ^2 testu zadat zobrazení očekávaných frekvencí, případně i residuálních frekvencí a zaškrtnout výpočet *Pearson & M-L Chi-square*, jak je ilustrováno níže.

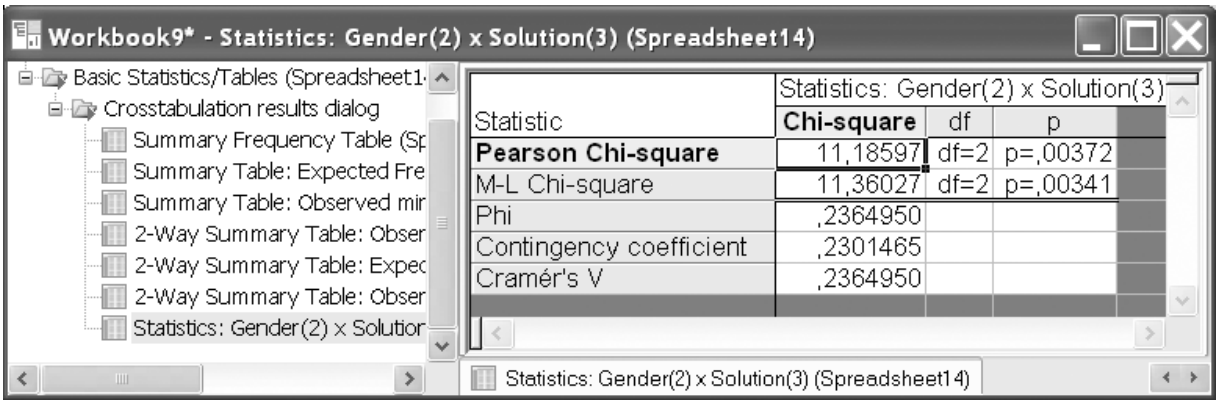


Po volbě tlačítka *Summary* se nám zobrazí výsledky následujícím způsobem:



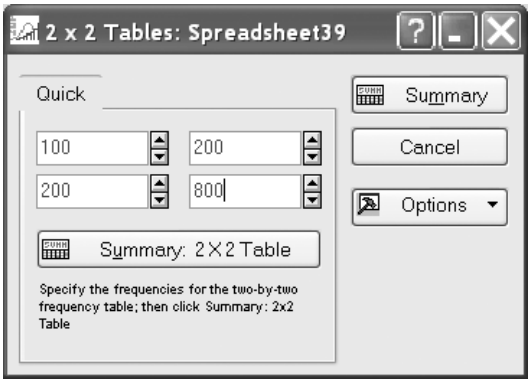
Levá část workbooku nám ukazuje, že byly vytvořeny tři tabulky, obsahující (v tomto pořadí) pozorované četnosti, očekávané četnosti a rozdíly mezi pozorovanými a očekávanými. Pouze poslední dvě tabulky obsahují v pravé horní části výpočet testu. Hodnota testové statistiky je 11.186 a po porovnání s χ^2 distribucí se 2 stupni volnosti je odhadem chyby 1. druhu (průkaznosti testu) p=0.003724, tedy ve shodě s našimi výpočty v Obr. 3-1 výše.

Pokud v záložce *Options* zaškrtneme druhou volbu (*Fisher exact, Yates*, *McNemar (2 x 2)*) a třetí volbu (*Phi (2x2 tables) & Cramér's V & C*) a pak na záložce *Advanced* zvolíme tlačítko *Detailed two-way tables*, zobrazí se nám další statistiky, některé z nich diskutované výše v této kapitole.



Kromě již dříve zobrazeného klasického testu je zobrazen i výsledek G (maximum-likelihood) testu a statistika v řádku *Phi* představuje hodnotu V koeficientu. Yatesova korekce je ale počítána jen pro čtyřpolní tabulky, pro náš příklad s 2x3 tabulkou není proto zobrazena.

Pro čtyřpolní tabulky se ale k výsledkům testů dostaneme snadněji, když v menu zvolíme *Statistics* | *Nonparametrics* a v zobrazeném seznamu zvolíme *2 x 2 Tables (...)*. Čtyřpolní tabulku zde zadáme velmi lehce v původním tvaru (viz také *Tab. 3-2*).



Po volbě tlačítka *Summary* se rovnou zobrazí všechny statistiky:

	2 x 2 Table (Spreadsheet39)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	100	200	300
Percent of total	7,692%	15,385%	23,077%
Frequencies, row 2	200	800	1000
Percent of total	15,385%	61,538%	76,923%
Column totals	300	1000	1300
Percent of total	23,077%	76,923%	
Chi-square (df=1)	23,11	p= ,0000	
V-square (df=1)	23,09	p= ,0000	
Yates corrected Chi-square	22,37	p= ,0000	
Phi-square	,01778		
Fisher exact p, one-tailed		----	
two-tailed		----	
McNemar Chi-square (A/D)	542,89	p=0,0000	
Chi-square (B/C)	,00	p= ,9601	

Je mezi nim jak klasická χ^2 statistika, tak stejný test s Yatesovou korekcí. Hodnota V koeficientu se zobrazuje jako druhá mocnina (*Phi-square*), takže je bohužel vždy kladná a nerozlišíme tedy kladnou a zápornou vazbu při porovnávání výskytu druhů. Q statistiku nepočítá Statistica ani tady, ani v proceduře *Tables and banners*.

Jak postupovat v programu R

Pro výpočet klasického χ^2 testu je nejprve nutné převést naimportovaný datový rámec (*chap3*) do podoby kontingenční tabulky pomocí funkce *xtabs*:

```
> chap3.tab <- xtabs(Count~Gender+Solution,data=chap3)
> chap3.tab
        Solution
Gender   failed stick tree
  female     20    15   65
  male       25    32   43
```

Vlastní test provedeme pak takto:

```
> sol.chisq <- chisq.test(chap3.tab)
> sol.chisq
        Pearson's Chi-squared test
data:  chap3.tab
X-squared = 11.186, df = 2, p-value = 0.003724
```

Objekt vrácený funkcí *chisq.test* obsahuje i další údaje, lze například zobrazit residuální hodnoty:

```
> sol.chisq$residuals
        Solution
Gender       failed      stick       tree
  female -0.5270463 -1.7534161  1.4969104
  male    0.5270463  1.7534161 -1.4969104
```

Tyto residuály ale neodpovídají těm, které zobrazuje Statistica. Jde o takzvané Pearsonovy residuály, tedy hodnoty $(O_i-E_i)/\sqrt{E_i}$. Součet jejich druhých mocnin je tedy roven chi-square statistice, tedy 11.186 v našem příkladě. Funkce *chisq.test* používá Yatesovu korekci jen pro čtyřpolní (2x2) tabulky a v takovém případě to provádí automaticky, pokud nepoužijeme při jejím volání parametr *correct=FALSE*.

Funkce *chisq.test* také umí testovat hypotézu o nezávislosti faktorů pomocí Monte Carlo simulace. Ta je do značné míry podobná Fisherovu exaktnímu testu, ale neodhaduje pravděpodobnost chyby 1. typu na základě systematického výčtu možných četností v jednotlivých buňkách tabulky, ale provádí náhodný výběr možných kombinací. S rostoucím počtem náhodných kombinací se výsledek stává přesnější:

```
> chisq.test(chap3.tab,simulate.p.value=T,B=1000)
        Pearson's Chi-squared test with simulated p-value (based on 1000
        replicates)
data:  chap3.tab
X-squared = 11.186, df = NA, p-value = 0.002997

> chisq.test(chap3.tab,simulate.p.value=T,B=1000000)
        Pearson's Chi-squared test with simulated p-value (based on 1e+06
        replicates)
data:  chap3.tab
X-squared = 11.186, df = NA, p-value = 0.003759
```

Pro výpočet G-testu je nejlepší použít log-lineární model, tj. zobecněný lineární model s předpokládanou Poissonovou distribucí (viz kapitola XX pro bližší vysvětlení):

```
> chap3.glm0 <- glm(Count~Gender+Solution,data=chap3,family=poisson)
> chap3.glm  <- update(chap3.glm0,.~.+Gender:Solution)

> anova(chap3.glm0,chap3.glm,test="Chisq")
Analysis of Deviance Table
Model 1: Count ~ Gender + Solution
Model 2: Count ~ Gender + Solution + Gender:Solution
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2      11.36
2         0       0.00  2    11.36 0.003413 **
```

Kontingenční tabulky můžeme vytvořit i přímo, pomocí funkce *rbind*. Následné použití ve funkcích *chisq.test* ilustruje výpočet pro 2x2 tabulku bez a s Yatesovou korekcí.

```
> tab3.2 <- rbind(c(100,200),c(200,800))
> tab3.2
     [,1] [,2]
[1,]  100  200
[2,]  200  800

> chisq.test(tab3.2,correct=F)
        Pearson's Chi-squared test
data:  tab3.2
X-squared = 23.1111, df = 1, p-value = 1.529e-06

> chisq.test(tab3.2,correct=T)
        Pearson's Chi-squared test with Yates' continuity correction
data:  tab3.2
X-squared = 22.3661, df = 1, p-value = 2.253e-06
```

Fisherův exaktní test je k dispozici ve funkci *fisher.test*:

```
> fisher.test(tab3.2)
```

```
        Fisher's Exact Test for Count Data
data:  tab3.2
p-value = 3.514e-06
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.484557 2.684159
sample estimates:
odds ratio
  1.998839
```

Koeficient V je pro kontingenční tabulky k dispozici v knihovně *vcd*, funkci *assocstats*, opět pod názvem Phi:

```
> library( vcd)
> assocstats(tab3.2)
                    X^2 df   P(> X^2)
Likelihood Ratio 21.817  1 2.9986e-06
Pearson          23.111  1 1.5290e-06
Phi-Coefficient   : 0.133
Contingency Coeff.: 0.132
Cramer's V        : 0.133
```

Popis analýz v článku

Methods

The dependency between the problem solution and ape gender was tested using Pearson's contingency table chi-squared test [with Yates correction for discontinuity].

nebo

The dependency between the two factors was tested using likelihood-ratio test [using a log-linear model].

Results

The chosen problem solution and failure rates differ between males and females (χ^2=11.186, df=2, p=0.0037). Contingency table residuals suggest there is small differences between both genders in the failure rate, but males prefer the use of stick, while females more often climb up the tree.

Doporučená četba

Zar (2007), p. 61-78; Quinn & Keough (2002), pp. 381-393 (kontingenční tabulky) a pp. 393-400 (log-lineární modely).

4 Normální rozdělení

Ve statistice nejčastěji používaným rozdělením je rozdělení normální (*normal distribution*). Má ve statistice do jisté míry „výsadní“ postavení. Velké množství statistických metod předpokládá, že data, se kterými pracujeme, mají normální rozdělení. Teoreticky je normální rozdělení rozdělením spojité proměnné na intervalové škále. V praxi se jím ale běžně charakterizují i spojitá data na poměrové stupnici, pokud je průměr alespoň o několik směrodatných odchylek větší než nula a v některých případech i diskrétní data (pokud mohou nabývat dostatečného počtu diskrétních hodnot). Mnohá biologická data lze skutečně úspěšně normálním rozdělením charakterizovat. Normální rozdělení se někdy nazývá Gaussovo rozdělení (*Gaussian distribution*). Jeho hustota pravděpodobnosti je symetrická, zvonovitá a je dána

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Vz. 4-1

Uvedená funkce obsahuje dvě konstanty (π a e, tj. Ludolfovo číslo a základ přirozených logaritmů) a dva parametry: μ a σ. Pozor, zde je jistá nedůslednost: v tomto textu používáme μ a σ ve dvou významech: jednak obecně jako symbol pro střední hodnotu a směrodatnou odchylku a jednak jako parametry normálního rozdělení. Lze ukázat, že μ je střední hodnotou: platí, že pro funkci f(x) ze Vz. 4-1 je

$$\int_{-\infty}^{+\infty} x f(x) dx = \mu$$

Vz. 4-2

a σ^2 je variance, tzn. že

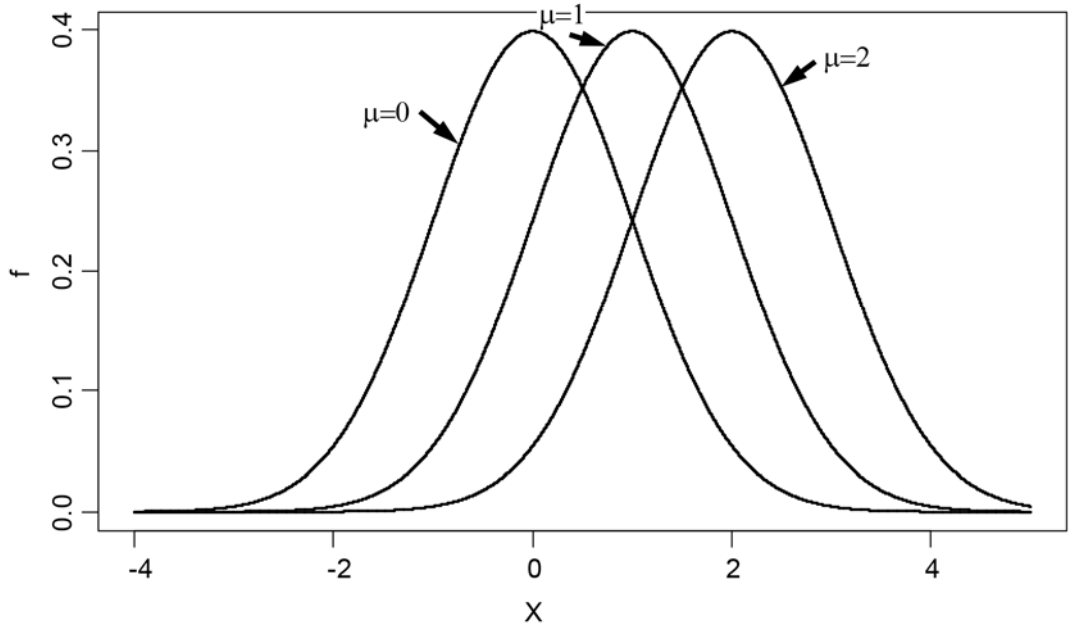
$$\int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = \sigma^2$$

Vz. 4-3

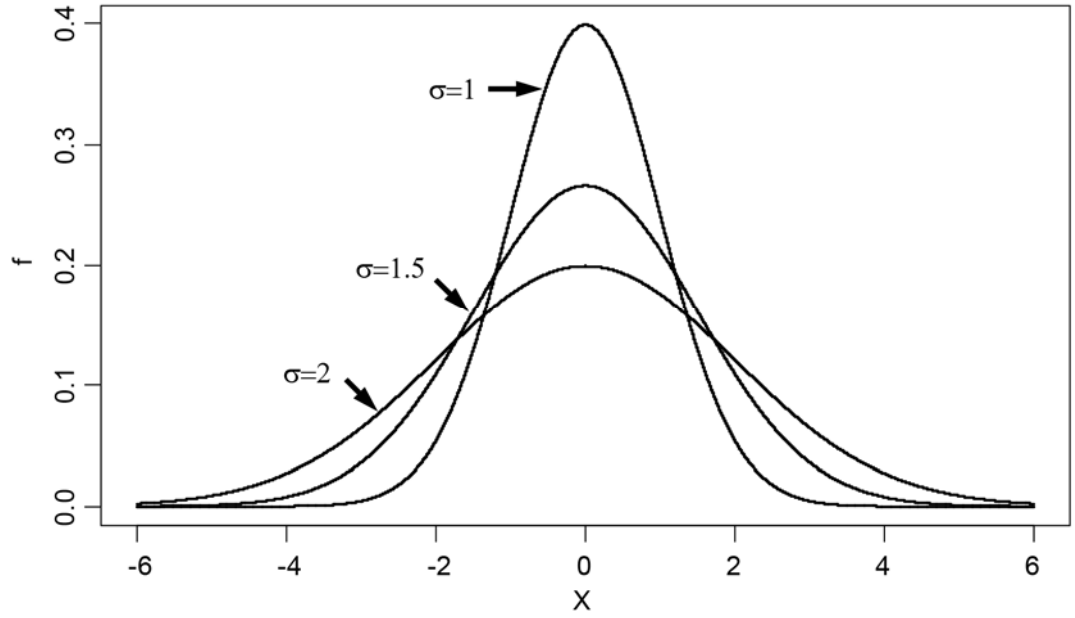
Význam obou parametrů ukazují Obr. 4-1 a Obr. 4-2. Při konstantním σ zůstává tvar křivky konstantní, μ určuje její polohu. Naopak σ určuje šíři „zvonu“.

Výjimečné postavení normálního rozdělení je do jisté míry důsledkem platnosti *centrální limitní věty*. Vyplývá z ní, že průměr „velmi velkého“ náhodného výběru je náhodnou veličinou s přibližně normálním rozdělením, i když má základní soubor rozdělení jiné než normální[*].

[*] Z toho také vyplývá, že narušení předpokladu normality vadí více při práci s malými výběry, než při práci s velkými výběry.



Obr. 4-1 Normální distribuce s hodnotami parametru μ 0, 1 nebo 2. Parametr σ je vždy 1.



Obr. 4-2 Normální distribuce s μ=0 a měnícím se rozptylem (*spread*) určeným hodnotou parametru σ.

Šikmost a špičatost

Šikmost - Jak již bylo uvedeno (kapitola 1), průměrná hodnota druhé mocniny odchylky od průměru, variance, je mírou variability souboru.[x] Průměrná hodnota třetí mocniny odchylky od průměru (tj. třetí centrální moment, κ_3) je mírou šikmosti. Součet odchylek od průměru je (z definice) roven nule. Pokud je jedna velká kladná odchylka kompenzována několika malými zápornými odchylkami, potom třetí mocnina velké kladné odchylky je velké kladné

[x] Obecně: *i*-tá mocnina odchylky od průměru se nazývá *i*-tý centrální moment, často se značí κ_i; variance je tedy κ_2 - druhý centrální moment

číslo, zatímco třetí mocniny malých záporných odchylek jsou sice záporné, co do absolutní velikosti ale zanedbatelné ve srovnání s velkou kladnou.

Takové rozdělení je pozitivně šikmé (*positively skewed*) rozdělení, opakem je záporně šikmé (*negatively skewed*) rozdělení; viz Obr. 1-1 *c* a *d*. κ_3 je udáváno ve třetích mocninách jednotek (je-li měření v cm, κ_3 je v cm^3). Jeho hodnota závisí na celkové variabilitě i na užitých jednotkách (užijeme-li místo metrů centimetry, bude milionkrát větší). Proto se užívá parametr, kde je v čitateli třetí mocnina směrodatné odchylky: $\gamma_1 = \frac{\kappa_3}{\sigma^3}$. Ten je bezrozměrný a vztahuje se pouze ke tvaru funkce hustoty pravděpodobnosti. Pro normální rozdělení je $\gamma_1 = \kappa_3 = 0$.

Špičatost - čtvrtý centrální moment charakterizuje špičatost. Špičatá rozdělení se nazývají leptokurtická, plochá jsou platykurtická. Normální rozdělení je mesokurtické. Pro něj platí: $\frac{\kappa_4}{\sigma^4} = 3$ za míru špičatosti se potom považuje $\gamma_2 = \frac{\kappa_4}{\sigma^4} - 3$. Špičatá (leptokurtická) rozdělení mají $\gamma_2 > 0$; opak je pravdou pro platykurtická rozdělení.

Standardizované normální rozdělení

Pokud má proměnná X normální rozdělení s parametry μ, σ^2, potom po transformaci

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Vz. 4-4

má proměnná Z normální rozdělení se střední hodnotou 0 a variancí 1 (směrodatná odchylka je tedy také 1). Takovému rozdělení se říká standardizované normální rozdělení (angl. *standard score*, *normal deviate*). Standardizace obecného normálního rozdělení se dříve užívala pro určení pravděpodobností, že hodnota pozorování bude ležet v určitém intervalu, protože hodnoty distribuční funkce F(x) standardizovaného normálního rozdělení byly tabelovány. Dnes můžeme tyto pravděpodobnosti (nebo naopak kvantily odpovídající zvoleným pravděpodobnostem) spočíst již přímo za pomocí statistických programů.

Některé kvantily standardizovaného normálního rozdělení jsou dobře známé a běžně používané. Hodnoty -1.96 a 1.96 jsou 2.5% a 97.5% kvantily standardizovaného normálního rozdělení. To znamená, že mezi těmito hodnotami leží 95% studované populace, má-li proměnná normální rozdělení. To je také důvod, proč se v grafech někdy vynáší průměr ± dvě směrodatné odchylky. Pokud by se jednalo o skutečné parametry populace, mezi těmito hodnotami leží přibližně 95% všech hodnot. Může být užitečné si pamatovat, že -1 a 1 jsou 15.9% a 84.1% kvantily normovaného normálního rozdělení, takže v intervalu střední hodnota ± směrodatná odchylka leží přibližně 68.2% pozorování.

Ověřování normality rozdělení

V některých případech potřebujeme zjistit, zda data v našem výběru pocházejí ze souboru s normálním rozdělením. Existuje řadu způsobů, jak v takovém případě postupovat, od grafických popisných metod až po formální statistické testy.

V následujících odstavcích uvádíme dvě grafické metody a tři způsoby testu nulové hypotézy, že data jsou náhodným výběrem ze základního souboru, který je charakterizován

náhodným rozdělením (statistikové říkají přesněji, že data jsou realizací náhodné proměnné s normálním rozdělením).

1. Nejjednodušším a často používaných způsobem je posouzení vzhledu histogramu četností, který si vyneseme z vlastních dat. Na vytvořeném histogramu hlavně sledujeme, zda je symetrický. Histogram můžeme případně porovnat s teoretickou křivkou hustoty pravděpodobnosti normálního rozdělení, které má stejnou střední hodnotu a směrodatnou odchylku.
2. Vynesení grafu užívajícího „normální pravděpodobnostní stupnici“ (*normal probability plot*). Na vodorovnou osu vynášíme hodnotu proměnné a na svislou osu procento (jako desetinné číslo) pozorování, menších než daná hodnota. Je to tedy vlastně graf kumulativních relativních četností, odhad distribuční funkce. Distribuční funkce normálního rozdělení má sigmoidní tvar. Proto se procenta (odhady pravděpodobností) vynášejí na tzv. pravděpodobnostní stupnici (*probability* scale). Pravděpodobnostní stupnice je zhuštěná kolem hodnoty 0.5. Body vynesené na pravděpodobnostní stupnici leží na přímce, pokud mají data normální rozdělení. Odchylky od normality se projeví v tomto vynesení nelinearitou. Variantou tohoto diagramu, která se v dnešní době používá více, je takzvaný QQ diagram (*quantile-quantile* diagram). Ten se liší vpodstatě jen v jednotkách užívaných na svislé ose, kde jsou pro naše pozorování (seřazené od nejmenšího do největšího) vynášeny odpovídající kvantily normálního rozdělení, se kterým data porovnáváme (diagram ale bývá často vynášen s prohozenými osami, tj. pozorované hodnoty na svislé a očekávané na vodorovné ose). Tvorba tohoto typu diagramu je ukázána níže v částech “Jak postupovat”.
3. χ^2 test - Spočtu průměr a varianci z dat. Ty pokládám za charakteristiky normálního rozdělení. Rozdělím rozsah pozorování na určitý počet tříd (podle celkového počtu pozorování). Spočtu očekávané četnosti a χ^2 testem je porovnám se skutečnými. Počet stupňů volnosti je *počet tříd - 1 - počet odhadovaných parametrů rozdělení*, pro normální rozdělení tedy *počet tříd - 3*. Mohu ale také testovat shodu s rozdělením, jehož parametry předem znám a neodhaduji je z dat. Potom je počet stupňů volnosti *počet tříd - 1*.
4. Kolmogorov-Smirnovův test hledá maximální rozdíl mezi očekávanou a napozorovanou relativní kumulativní frekvencí.
5. Lze spočíst šikmost nebo špičatost a poté testovat, zda se spočtené hodnoty významně liší od nuly (např. pomocí tzv. standardizované šikmosti a špičatosti).

Zde je nutné připomenout, že neprůkazný výsledek testu nulové hypotézy není důkazem její platnosti. Neprůkazný výsledek testu shody s normálním rozdělením tedy není důkazem toho, že data mají normální rozdělení: zvláště pokud je náš výběr malý, jsou testy velmi slabé. Při ověřování normality před užitím některých metod, které předpokládají normalitu (např. t-test, analýza variance, viz následující kapitoly) může dojít k paradoxní situaci: při malých výběrech, kde narušení normality vadí, nejsme schopni nulovou hypotézu zamítnout, což nás neoprávněně uklidní, zatímco při velkých souborech dat, kdy jsou t-test a analýza variance poměrně robustní k narušení normality (robustní - angl. *robust* - znamená nepříliš citlivé k narušení předpokladů testu), nulovou hypotézu zamítáme, často jen při nepatrné odchylce od normálního rozdělení. Přesto editoři nebo recenzenti časopisů takový test někdy vyžadují.

Níže popisujeme, jak testovat shodu s normální distribucí v programech Statistica a R a také jak takový test popsat v odborném článku. To ale neznamená, že se osobně s takovým postupem ztotožňujeme – jednak z důvodů uváděných v předchozím odstavci, jednak proto, že testovat normalitu například pro závislou (vysvětlovanou) proměnnou v analýze variance či v regresi postrádá smysl. To, co by mělo mít rozdělení blízké normálnímu rozdělení, jsou residuály těchto modelů, tj. variabilita zůstavající po odečtení systematických vlivů, a také bychom se měli soustředit na neměnnou variabilitu těchto residuálu, spíše než na (přísnější) požadavek shody s konkrétním tvarem rozdělení.

χ^2 test a Kolmogorov-Smirnovův test můžeme použít i pro testování shody dat s jinými teoretickými rozděleními.

Příkladová data

Předpokládáme, že výška studentů má normální rozdělení se střední hodnotou 175 cm a směrodatnou odchylkou 14 cm. (a) Jak velká část studentů bude mít výšku větší než 190 cm? (b) Jaká část studentů bude mít výšku mezi 160 a 180 cm? (c) Kolik studentů ze souboru 380 osob bude mít výšku větší než 200 cm? (d) V jakém rozsahu budou výšky 10% studentů s nejmenší výškou?

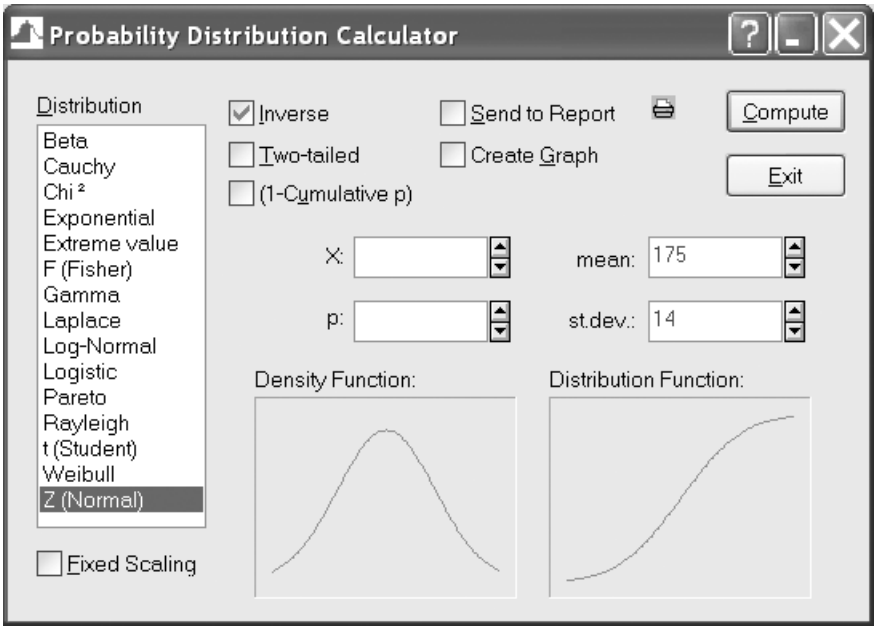
V listu *Chap4* příkladových dat (*biostat-data.xlsx*) jsou data o počtu semenáčků lučních rostlin a pokryvnosti opadu (viz kapitola 1 pro bližší popis). Srovnejte rozdělení hodnot proměnné *Seedlings* s normální distribucí. Pokud najdete velké odchylky, porovnejte s normální distribucí logaritmované hodnoty této proměnné.

Jak postupovat v programu Statistica

Hledání hodnoty distribuční funkce a kvantilů

Se základy výpočtu pravděpodobnosti, že si ze zvoleného rozdělení náhodně vyberu hodnotu menší nebo rovnou zvolené konstantě, a opačného postupu, tj. určení limitní (kritické) hodnoty pro zvolenou pravděpodobnost, jsme se seznámili již v kapitole 2 na příkladu χ^2 rozdělení a čtenáři doporučujeme si nejprve prostudovat návod tam uvedený. Jediný rozdíl proti kapitole 2, že zvolená distribuce nepředstavuje očekávanou distribuci testové statistiky, ale předpokládanou distribuci pro naše data.

Zvolíme příkaz *Statistics* | *Probability Calculator* | *Distributions* a v zobrazeném dialogovém oknu nejprve zvolíme typ distribuce: *Z (Normal)* a zadáme její parametry (v políčkách *mean* a *st.dev.*). Je užitečné přitom **nemít** zaškrtnutou volbu *fixed scaling* v levém dolním rohu – v grafech pak hezky vidíme, které části distribuční křivky námi spočtené pravděpodobnosti odpovídají.



Nyní můžeme poměrně snadno odpovídat na výše uvedené otázky:

(a) Zaškrtneme volbu *(1-Cumulative p*) a do políčka *X* zadáme hodnotu 190. Po volbě tlačítka *Compute* se v políčku *p* objeví číslo *0.14199*. Větší výšku než 190 cm bude tedy mít zhruba 14% studentů. To také znamená, že zhruba 86% bude mít výšku menší než 190 cm.

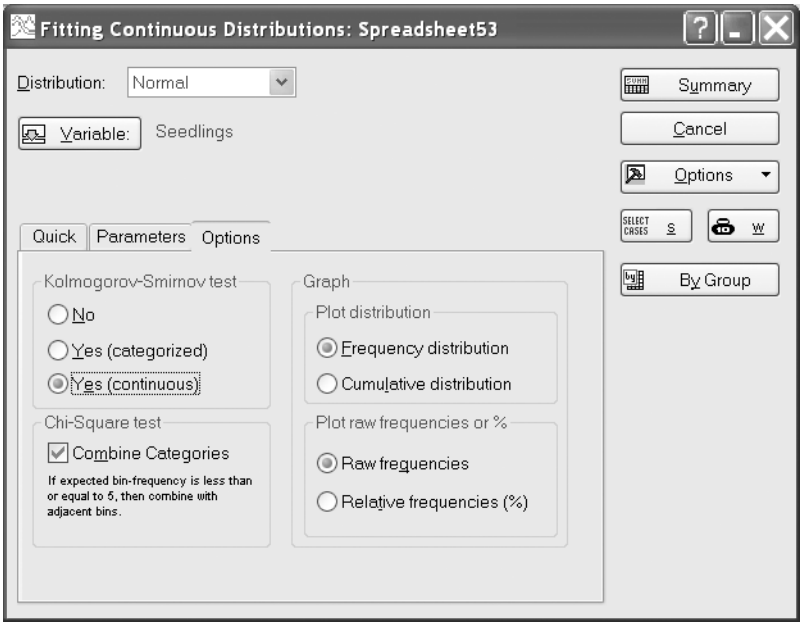
(b) K odpovědi na tuto otázku musíme spočíst podíl studentů s výškou do 180 cm a od něj odečíst podíl studentů s výškou do 160 cm. Zrušíme tedy volbu *(1-Cumulative p)* a nejprve zadáme do *X* hodnotu 180 a zmáčkneme *Compute*: v *p* je zobrazeno číslo (po zaokrouhlení) 0.640. Pak změníme *X* na 160 a opět zmáčkneme *Compute*: v *p* je teď 0.142. Podíl studentů s výškou mezi 160 a 180 cm je tedy 0.64-0.142, tedy 49.8%.

(c) Tato úloha je podobná otázce (a), pouze odhadnutou pravděpodobnost násobíme velikosti uvažované skupiny (výběru), čímž dostáváme očekávaný počet jedinců. Hodnota *p* pro *X=200* (při zaškrtnuté volbě *(1-Cumulative p)*) je 0.037 a po vynásobení 380 dostáváme očekávaný počet asi 14 studentů s výškou nad 200 cm.

(d) 10% nejmenších studentů bude mít výšku do hodnoty kvantilu 0.10 pro normální distribuci se zvolenými parametry. Zrušíme opět volbu *(1-Cumulative p)*, do políčka *p* zadáme hodnotu 0.10 a zmáčkneme tlačítko *Compute*. Hodnota v políčku *X* nám říká, že výška 10% nejmenších studentů bude 157 cm a méně.

Testování shody s teoretickou distribucí

Po importu dat z listu *Chap4* zvolíme v menu *Statistics* | *Distribution Fitting* a v dialogovém okně zvolíme v levém seznamu (*Continuous Distributions*) *Normal* a zmáčkneme *OK*. V dalším dialogovém okně zvolíme záložku *Options*, zaškrtneme v oblasti *Kolmogorov-Smirnov test* volbu *Yes (continuous*) a pomocí tlačítka *Variable* vybereme proměnnou *Seedlings*.



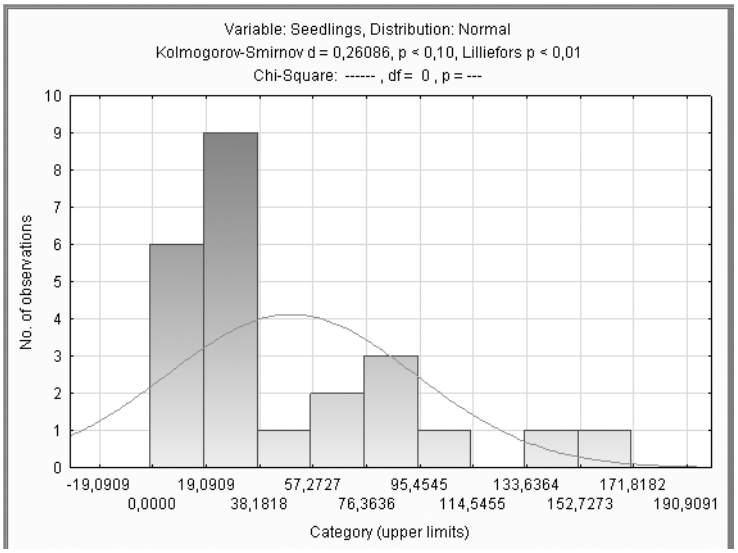
Tlačítkem *Summary* zobrazíme výsledky χ^2 a Kolmogorov-Smirnovova testu.

Variable: Seedlings, Distribution: Normal (Spreadsheet53)
Kolmogorov-Smirnov d = 0,26086, p < 0,10, Lilliefors p < 0,01
Chi-Square = 4,24657, df = 1 (adjusted) , p = 0,03933

Upper Boundary	**Observed Frequency**	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected
<= -10,00000	0	0	0,00000	0,0000	2,167184	2,16718
0,00000	0	0	0,00000	0,0000	1,018376	3,18556
10,00000	4	4	16,66667	16,6667	1,307459	4,49302
20,00000	3	7	12,50000	29,1667	1,595835	6,08885
30,00000	4	11	16,66667	45,8333	1,851771	7,94063
40,00000	5	16	20,83333	66,6667	2,042804	9,98343
50,00000	0	16	0,00000	66,6667	2,142425	12,12585

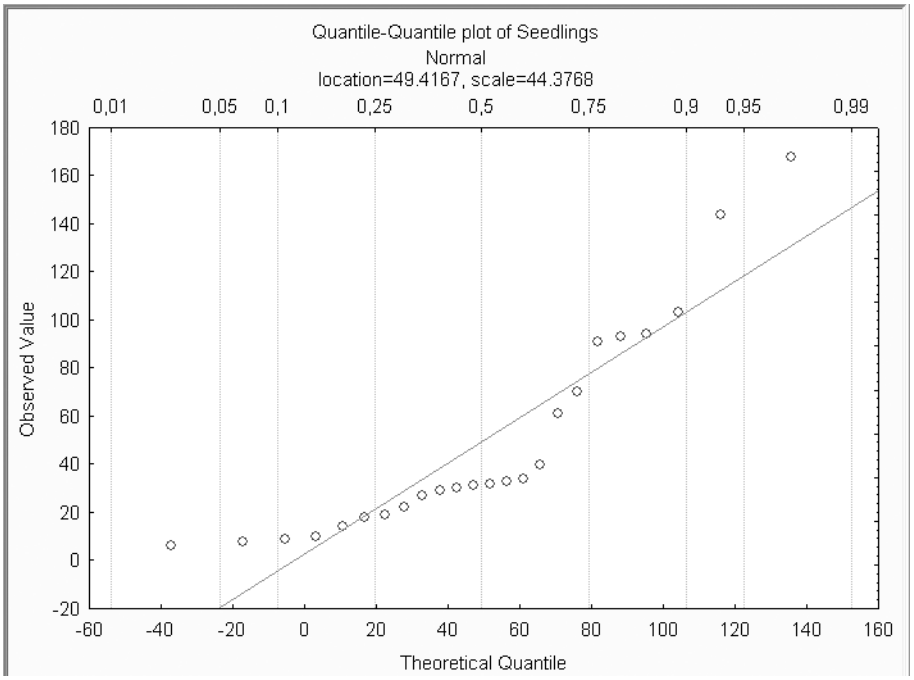
Uvedené testy naznačují, že můžeme zamítnout nulovou hypotézu o shodě s normální distribucí. Všimněte si, že χ^2 test má jen jeden stupeň volnosti, přestože tabulka zobrazuje daleko větší počet intervalů, ve kterých byly pozorované a očekávané počty pozorování porovnávány. Je to tím, že Statistica sdružuje sousedící intervaly až do dosažení minimální hodnoty 5 ve sloupci *Expected Frequency* (viz volba *Combine Categories* ve výše zobrazeném dialogovém okně) a také proto, že další dva stupně volnosti jsou odečteny, protože průměr a variance normální distribuce, se kterou srovnáváme, byly odhadnuty z našich dat. Ve výše uvedené tabulce je užitečné zkontrolovat, jak program vytvořil kategorie (intervaly), a to zvláště tehdy, pokud měříme data s určitou přesností. Je třeba se vyvarovat situace, kdy šířka intervalu není celistvým násobkem nejmenšího rozdílu, který jsme schopni detekovat (jestliže např. měříme délku s přesností na metry, musí být šířka intervalu jeden metr nebo jeho celistvým násobkem).

Jednoduché grafické porovnání rozdělení pro naše data s normálním rozdělením získáme po návratu do dialogového okna *Fitting Continuous Distributions* tak, že na záložce *Parameters* nejprve snížíme hodnotu pro *Number of categories* na 11 (pro získání kompaktnějšího histogramu) a pak na záložce *Quick* zvolíme *Plot of observed and expected distribution*.



Vidíme, že shoda se symetrickou, zvonovitou křivkou normální distribuce opravdu není příliš dobrá. To, že hodnoty *p* v testech nejsou více přesvědčivé je dáno malou sílou testu pro náš malý výběr s 24 pozorováními.

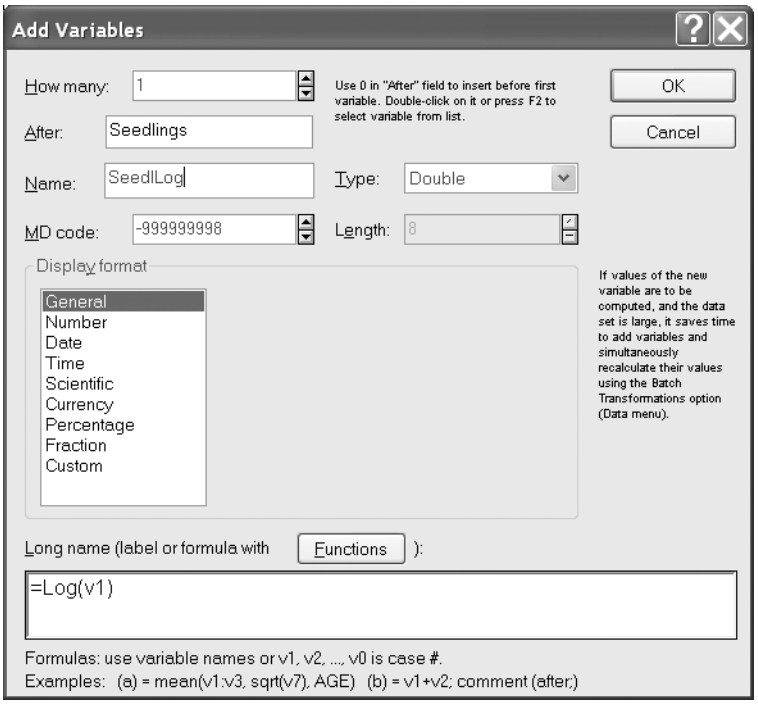
QQ diagram vytvoříme po zvolení příkazu *Statistics | Distributions & Simulation*. V prvním dialogovém okně zvolíme *Fit Distribution* a tlačítko *OK*, v dalším okně zadáme za pomoci tlačítka *Variables* proměnnou *Seedlings* v seznamu *Continuous variables* (a opět zvolíme *OK*) a na záložce *Continuous variables* pak zrušíme všechny zaškrtnuté volby kromě první (*Normal*). Po volbě tlačítka *OK* se objeví nové okno, kde můžeme zvolit *Q-Q plot* a získáme tak následující obrázek.



Je příjemné, že Statistica vynáší pro horizontální osu nejen očekávané hodnoty kvantil (na dolním okraji) ale paralelně i odpovídající pravděpodobnosti (na horním okraji). Odchylky bodů od referenční přímky naznačují jednak to, že na rozdíl od normální distribuce jsou pozorované hodnoty zdola omezeny nulou (počet semenáčků nemůže být záporný), jednak pomalejší nárůst hodnot (část podlézající referenční přímku) odpovídající pozitivně šikmé distribuci.

Pro testování shody log-transformovaných počtů semenáčků s normálním rozdělením musíme nejprve vytvořit novou proměnnou s logaritmy původních hodnot (budeme užívat přirozené logaritmy, ale volba základu nemá vliv na tvar rozdělení). V datovém spreadsheetu přidáme novou proměnnou například pomocí příkazu *Data* | *Variables* | *Add* a v zobrazeném dialogovém okně zvolíme nejprve jméno proměnné (například *SeedlLog* místo *NewVar*) a pak ve větším bílém políčku v spodní části okna zadáme vzoreček definující hodnoty této proměnné, například jako *=Log(v1)*

Číslo *1* v názvu *v1* odkazuje na skutečnost, že chceme logaritmovat první proměnnou (první sloupeček) v datech. Alternativně bychom také mohli zadat *=Log(Seedlings)*



Po zmáčknutí *OK* je (při správném zadání vzorečku) nová proměnná vytvořena a vyplněna hodnotami. Dále již postupujeme stejně jako v případě původní proměnné. Při testování shody s normální distribucí nejsme pak schopni tuto hypotézu zamítnout a zobrazený histogram má alespoň náznak symetričnosti distribuční křivky.

Jak postupovat v programu R

Hledání hodnoty distribuční funkce a kvantilů

Podobně jako v příkladech pro kapitolu 2, i zde budeme používat dvě funkce pro spočtení kvantilů a/nebo kumulativní pravděpodobnosti. Jejich jména jsou *qnorm* a *pnorm* a jejich první parametr odpovídá hodnotě pravděpodobnosti nebo hodnotě proměnné, zatímco další dva parametry udávají střední hodnotu a směrodatnou odchylku referenční normální distribuce.

```
> 1 - pnorm( 190, 175, 14)
[1] 0.1419884
> pnorm( 180, 175, 14) - pnorm( 160, 175, 14)
[1] 0.4975192
> 380 * (1 - pnorm( 200, 175, 14))
[1] 14.08765
> qnorm( 0.10, 175, 14)
[1] 157.0583
```

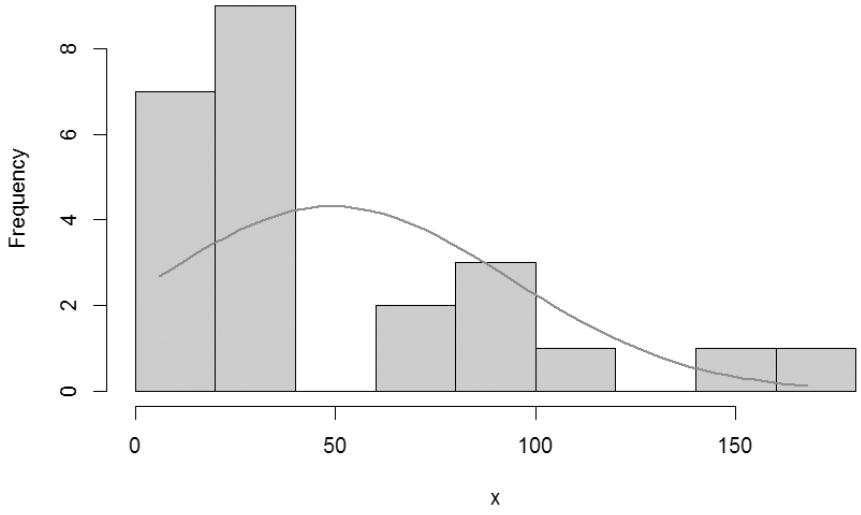
Testování shody s teoretickou distribucí

Pro grafické porovnání histogramu s normálním rozdělením si můžeme definovat a použít následující pomocnou funkci:

```
> hist.norm <- function(x,nbins=10)
{
  hist.x <- hist( x, breaks=nbins, col="light blue",main="")
  x.val  <- seq( min(x), max(x), length=50)
  x.fit  <- dnorm( x.val, mean=mean(x), sd=sqrt(var(x)))
  x.fit  <- x.fit * diff(hist.x$mids[1:2]) * length(x)
  lines( x.val, x.fit, col="red", lwd=2)
}
```

Tato funkce umožňuje změnit počet intervalů pomocí parametru *nbins*, ostatní volby (hladkost křivky pro normální rozdělení a barva histogramu a křivky) jsou zvoleny pevně. Graf pak vytvoříme voláním této funkce.

```
> hist.norm(chap4$Seedlings)
```



QQ diagram vytvoříme následovně (výsledný obrázek vynechán):

```
> qqnorm(chap4$Seedlings)
> qqline(chap4$Seedlings,distribution=qnorm,lwd=2)
```

Pearson χ^2 test je v programu R k dispozici v jednoduché podobě ve volitelné knihovně *nortest* (typicky ji musíte doinstalovat), ve funkci *pearson.test*:

```
> pearson.test(chap4$Seedlings)
        Pearson chi-square normality test
data:  chap4$Seedlings
P = 20.6667, p-value = 0.0009363
```

Tato funkce ale neprovádí automatické slučování sousedících intervalů s nízkými očekávanými četnostmi, lze toho ale přibližně dosáhnout zadáním výchozího počtu intervalů:

```
> pearson.test(chap4$Seedlings,n.classes=4)
        Pearson chi-square normality test
data:  chap4$Seedlings
P = 4.3333, p-value = 0.03737
```

Kolmogorov-Smirnovův test lze spočíst takto:

```
> with(chap4,ks.test(Seedlings,"pnorm",mean(Seedlings),sqrt(var(Seedlings))))
        One-sample Kolmogorov-Smirnov test
data:  Seedlings
D = 0.2609, p-value = 0.06268
```

```
alternative hypothesis: two-sided
```

Tento způsob použití Kolmogorov-Smirnovova testu (kdy jsou parametry normální distribuce odhadnuty z dat) ale nápověda funkce nedoporučuje, místo něj je doporučován Shapiro-Wilk test:

```
> shapiro.test(chap4$Seedlings)
        Shapiro-Wilk normality test
data:  chap4$Seedlings
W = 0.8305, p-value = 0.0009702
```

Obdobné grafy a testy můžeme provést pro logaritmované počty semenáčků tak, že všude, kde je užívána proměnná *Seedlings* (nebo *chap4$Seedlings*) použijeme *log(Seedlings)* (nebo *log(chap4$Seedlings)*).

Popis analýz v článku

Methods

We have tested the consistency of the distribution of the observed number of seedlings with Normal distribution using the Pearson's χ^2 test with adjusted degrees of freedom [using the Kolmogorov-Smirnov test] [using the Shapiro-Wilk test of normality].

Results

The number of seedlings was found to be significantly different from Normal distribution (χ^2=4.247, df=1, p=0.03933) and this discrepancy was fixed by log-transforming the variable values.

V případě, že se odchylka od normality po logaritmické transformaci zmenšila, ale odchylka od normality je stále průkazná (což se může stát, když máme velký soubor dat, u kterého tato odchylka při velkém počtu pozorování nemusí vadit), můžeme užít alespoň tuto formulaci: data were log-transformed to improve the normality.

Doporučená četba

Sokal & Rohlf (1981), pp. 98-127; Zar (2007), pp. 79 – 98, Quinn & Keough (2002) pp. 17-18.

5 Studentovo t-rozdělení a jeho použití

Příklady řešených problemů:

1. Chodci jsou se zavázanýma očima vypouštěni směrem na cíl. Na linii kolmé ke spojnici start - cíl je měřena jejich odchylka od cíle v metrech (nalevo záporné, napravo kladné hodnoty). Otázka zní, zda existuje systematická odchylka od přímého směru, tedy od nuly.
2. Známe koncentraci stabilního izotopu ^{13}C ve vzduchu a považujeme ji za pevnou hodnotu. Poté změříme koncentraci izotopu v deseti pokusných rostlinách (a ta se bude mezi rostlinami lišit). Ptáme se, jestli je koncentrace izotopu v rostlinách (její střední hodnota) shodná s koncentrací ve vzduchu. Pokud by nebyla, je to možné považovat za důkaz, že rostliny při fotosyntéze diskriminují mezi izotopy uhlíku.
3. U vzorku populace měříme koncentraci cholesterolu v periferní krvi a v krvi odebrané ze žíly. Ptáme se, jestli je mezi těmito dvěma hodnotami v celé populace systematický rozdíl (tj. jedna hodnota je systematicky vyšší než druhá). Tuto otázku můžeme formulovat také tak, zda je rozdíl těchto dvou hodnot systematicky odlišný od nuly.
4. Měříme koncentraci Pb v mase náhodně vybraných kaprů z určitého rybníka. Chceme znát interval, který nám s určitou pravděpodobností (nejčastěji 95%) pokryje neznámou střední hodnotu této koncentrace u všech kaprů v daném rybníce.

V minulé kapitole jsem si ukázali, že má-li proměnná X normální rozdělení se střední hodnotou μ a variancí (rozptylem) σ^2_X, potom proměnná

$$Z = \frac{\overline{X} - \mu}{\sigma_X}$$

Vz. 5-1

má normální rozdělení se střední hodnotou nula a jednotkovou variancí (standardizované normální rozdělení). Již dříve bylo ukázáno, že pokud má základní soubor střední hodnotu μ a rozptyl σ^2_X, potom je výběrový průměr $\overline{X}$ náhodnou proměnnou se střední hodnotou μ a směrodatnou odchylkou*

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \ .$$

Vz. 5-2

Směrodatná odchylka výběrového průměru se nazývá střední chyba průměru. Platí také, že odpovídá-li základní soubor normálnímu rozdělení, potom i rozdělení průměrů je normální. Můžeme vyslovit dokonce silnější tvrzení: rozdělení průměrů je bližší normálnímu než rozdělení základního souboru. Z toho vyplývá, že i veličina

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

Vz. 5-3

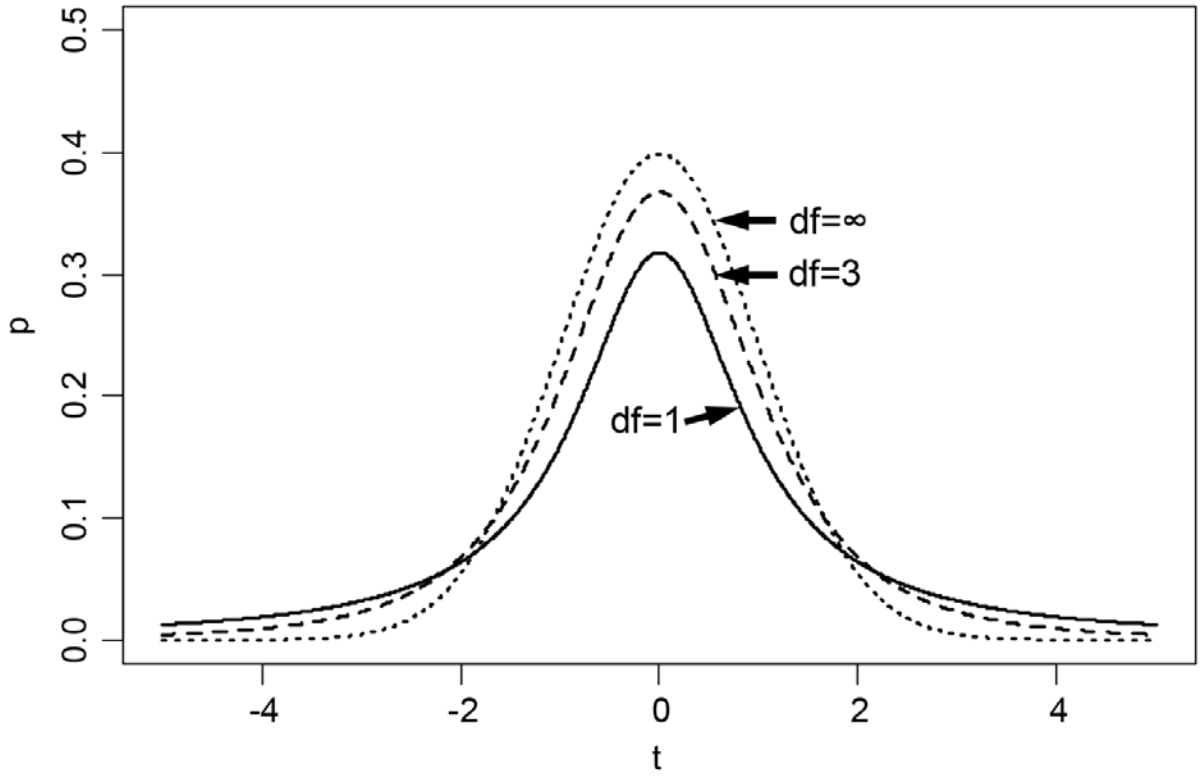
* **Náhodný výběr z konečně velkého základního souboru.** V biologii většinou provádíme náhodný výběr z (potenciálně) nekonečného základního souboru nebo je základní soubor mnohem větší než průměr. Pokud tvoří výběr větší část základního souboru (alespoň 5%), je odhad průměru přesnější, než jak ukazují shora uvedené výpočty, a je možné při odhadu střední chyby průměru užít korekci na konečnost souboru.

má standardizované normální rozdělení. Této skutečnosti by bylo možné využít pro testování různých hypotéz o průměru, například zda se liší námi získaný výběrový průměr významně od určité teoreticky předpokládané hodnoty. Pro rozdělení Z můžeme říci, s jakou pravděpodobností se daná hodnota Z vyskytuje. Problém je ale v tom, že v drtivé většině praktických aplikací neznáme $\sigma_{\bar{x}}$, ale pouze jeho odhad $s_{\bar{x}}$ a tento odhad je zatížen určitou chybou (je tedy také náhodnou proměnnou). Obdobně získaná veličina

$$t = \frac{\overline{X} - \mu}{s_{\bar{X}}}$$

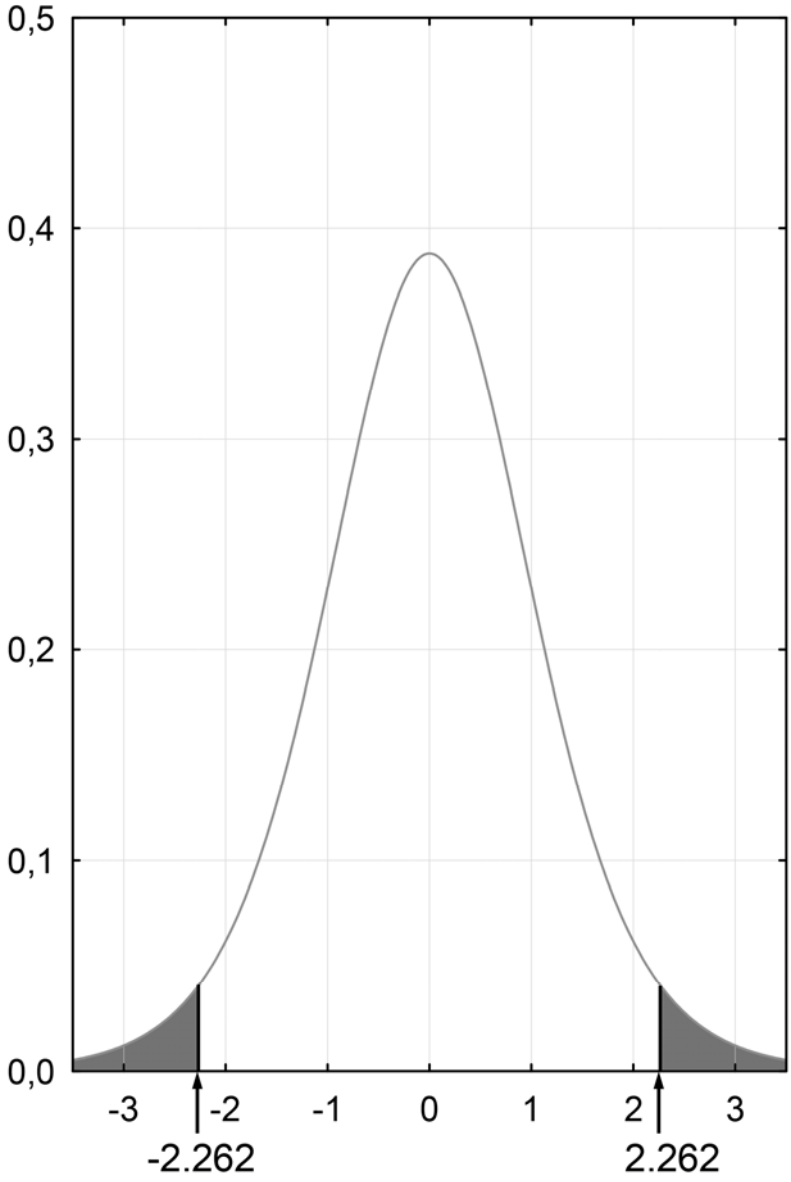
Vz. 5-4

má *t*-rozdělení - tzv. Studentovo t (*Student's t distribution,* Student byl vědecký pseudonym chemika Guinnessova pivovaru v Dublinu, který test vymyslel), které se mírně liší od rozdělení normálního. Je to rozdělení výběrové, závisející na počtu stupňů volnosti (*df* = *n*-1, kde *n* je rozsah výběru). Čím je rozsah výběru větší, tím více se blíží *t*-rozdělení normálnímu (viz Obr. 5-1).



Obr. 5-1 *t*-distribuce pro různé stupně volnosti, v. Pro df=∞, *t* distribuce je identická s normální distribucí.

t-rozdělení je jedním z nejpoužívanějších v praktické statistice. Nejjednodušší je jeho použití pro testování hypotéz o jednom výběru: můžeme se ptát, zda se zjištěný průměr významně liší od určité teoretické hodnoty. Protože známe charakteristiky *t*-rozdělení, jsme schopni spočítat jeho kvantily. *t*-rozdělení je symetrické, střední hodnota je nula. Za předpokladu platnosti nulové hypotézy je pravděpodobnost, že *t* bude buď menší než 2.5%-ní kvantil nebo větší než 97.5%-ní kvantil rovna 5%, jak ukazuje Obr. 5-2.



Obr. 5-2 *t* distribuce pro df =9, se znázorněnou kritickou oblastí (stínovaná plocha) pro dvoustranný test při užití α=0.05 (kritická hodnota *t* je 2.262).

Jak ukazuje obrázek, 2.5%-ní kvantil je záporné číslo a 97.5%-ní kvantil je kladné číslo, jejich absolutní hodnota je ale stejná. Nulovou hypotézu zamítáme na 5%-ní hladině významnosti, je-li spočtené *t* menší než 2.5%-ní kvantil nebo větší než 97.5%-ní kvantil. Jinak to lze formulovat takto: je-li $|t| >$ 97.5%-ní kvantil. 97.5%-ní kvantil je tedy kritickou hodnotou pro **dvoustranný test** na 5%-ní hladině významnosti. Obecně platí, že kritická hodnota pro dvoustranný test na hladině významnosti α je rovna (1 - α/2) x 100%-nímu kvantilu rozdělení. Značí se tedy $t_{\alpha(2),\nu}$. Dvojka v závorce značí, že se jedná o dvoustranný test, ν je počet stupňů volnosti (d.f.). Dvoustranným testem (*two-tailed test* - tedy doslova dvouocasový) míníme, že nulovou hypotézu zamítáme, pokud spočtená hodnota padne do jednoho ze dvou „ocasů“ křivky rozdělení.

Příklad: Koncentrace izotopu uhlíku ^{13}C se standardně vyjadřuje ja $\delta^{13}C$ – jeho hodnotu ve vzduchu budeme považovat za konstantní, a rovnou -8. Odebrali jsme deset rostlin (které považujeme za náhodný výběr z vhodně zvoleného základního souboru), a ptáme se, jestli bude tato hodnota v biomase rostlin průkazně odlišná od hodnoty ve vzduchu. Nulová hypotéza tedy může být formulována takto: střední hodnota $\delta^{13}C$ (μ) v biomase rostlin je

rovna hodnotě ve vzduchu, tj. -8. Tedy H_0: $\mu=\mu_0$ (v našem případě μ_0 =-8). Alternativní hypotéza zní: H_A: $\mu \neq \mu_0$. Hodnotu *t* vypočteme podle vzorce

$$t = \frac{\overline{X} - \mu_0}{s_{\overline{X}}}.$$

Vz. 5-5

Postup při hodnocení výše uvedeného příkladu ukazuje následující Obr. 5-3.

Hodnoty $\delta^{13}C$ v deseti zkoumaných rostlinách byly (vzali jsme C_4 rostliny): -10, -12, -13, -11, -15, -13, -16, -19, -11, -14.

H_0: μ=-8 $\quad\overline{X}$ = -13.4

H_A: $\mu\neq$-8 $\quad$ s = 2.716

α=0.05 $\quad s_{\overline{X}} = \frac{2.716}{\sqrt{10}} = 0.8589$

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{-13.4 - (-8)}{0.8589} = -6.287$$

df = n - 1 = 10 - 1 = 9

Dosažená hladina signifikance je p=0.000143 (viz komentář níže a Obr. 5-4), zamítáme H_0 a tvrdíme, že výběr 10 rostlin pochází ze základního souboru, jehož průměr je jiný než -8.

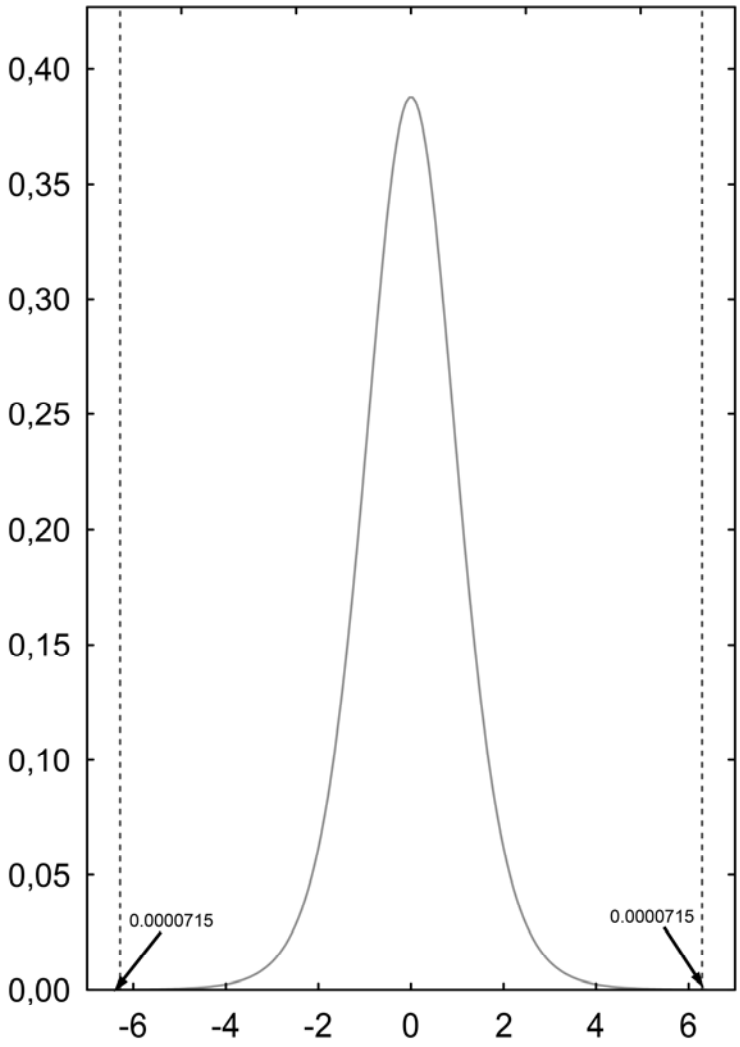
$t_{0,975\,,\,9}$ (tj. 97.5%ní kvantil t-rozdělení s df=9) = 2.262 (viz Obr. 5-2).

Obr. 5-3 Dvoustranný *t*-test na shodu (signifikantní odlišnost) výběrového průměru s hypotetizovaným průměrem základního souboru μ_0=-8.

Při prezentaci t-testu ale dnes obvykle nesrovnáváme s kritickou hodnou, nýbrž přímo určíme, jak pravděpodobné je, že dostaneme odchylku od nulové hypotézy tak velkou nebo větší, než je (absolutní) hodnota vypočtené testové statistky (tj. 6.287 v našem příkladě). Takto spočtená pravděpodobnost (p=0.000143 v našem příkladu, viz Obr. 5-3) je pak významností (signifikancí) dvoustranného t-testu. Způsob výpočtu je také ilustrován v následujícím Obr. 5-4.

Jedním z nejčastějších použití tohoto testu je testování, zda se dvě veličiny, měřené na sledovaných objektech liší. Např. vybereme 10 jedinců, a u každého zjistíme koncentraci cholesterolu v krvi odebrané z prstu a odebrané ze žíly a spočteme jejich rozdíl (vždy stejným způsobem, tedy např. žíla – prst, tj. kladné hodnoty znamenají, že koncentrace je vyšší v žíle). Testovanou proměnnou je rozdíl koncentrací. Máme tedy 10 rozdílů koncentrací a ptáme se, zda je průměrná hodnota rozdílů významně odlišná od nuly. Nulová hypotéza tedy zní: Střední hodnota rozdílu koncentrací je rovna 0. Toto je princip **párového t-testu**. Obdobně lze např. testovat, zda je u lidí rozdíl mezi obvodem levého a pravého zápěstí: testovanou proměnnou je rozdíl obvodu levého a pravého zápěstí u každé testované osoby.

Tento postup se často užívá při terénních pokusech. Např. se můžeme ptát, zda přidáním fosforu zvýšíme výnosy. Protože existuje značná prostorová variabilita, je výhodné mít vždy dvojice ploch (kontrolní a s přidáním fosforu) co nejblíže u sebe a za testovanou proměnnou považovat právě rozdíl výnosů každé dvojice. Pokud bychom neměli dvojice ploch spárovány, nelze párový test použít a je třeba užít dvouvýběrový test.



Obr. 5-4 Dvoustranný jednovýběrový *t*-test shody průměru s očekávanou hodnotou (-8.0). Vypočtená testová statistika (viz Obr. 5-2) je -6.287. Pravděpodobnost, že dostanu takto extrémní hodnotu (<-6.287 nebo >+6.287) z t-distribuce s df=9, je rovna součtu pravděpodobností (ploch pod distribuční křivkou), tj. 2*0.0000715=0.000143.

Jednostranné testy

Doposud jsme předpokládali nulovou hypotézu H_0: $\mu=\mu_0$, kterou zamítáme, je-li odchylka od nulové hypotézy nepravděpodobně velká buď na kladnou nebo na zápornou stranu. Alternativní hypotézou je tedy H_A: $\mu \neq \mu_0$. V některých případech nás ale zajímá odchylka od nulové hypotézy jen na jednu stranu. Např. podáváme lék na snížení tlaku, a proto nás zajímá pouze snížení. Nulová hypotéza je potom: Podání léku nemá žádný vliv nebo způsobí zvýšení tlaku; alternativou je: lék způsobí snížení tlaku. Formálně: H_0: $\mu \geq \mu_0$; H_A: $\mu < \mu_0$, v našem případě $\mu_0 = 0$. Za průkazný potom označíme pouze takový výsledek testu, kdy t je menší než 5%-ní kvantil. Nebo naopak pokud nás zajímá pouze kladná odchylka, když t je větší než 95%-ní kvantil. Připomeňme, že 5%-ní a 95%-ní kvantil se liší pouze znaménkem. Obecně platí, že kritická hodnota pro jednostranný test na hladině významnosti α je rovna (1 - α) x 100%-nímu kvantilu rozdělení. Příklad výpočtu viz Obr. 5-5 a Obr. 5-6.

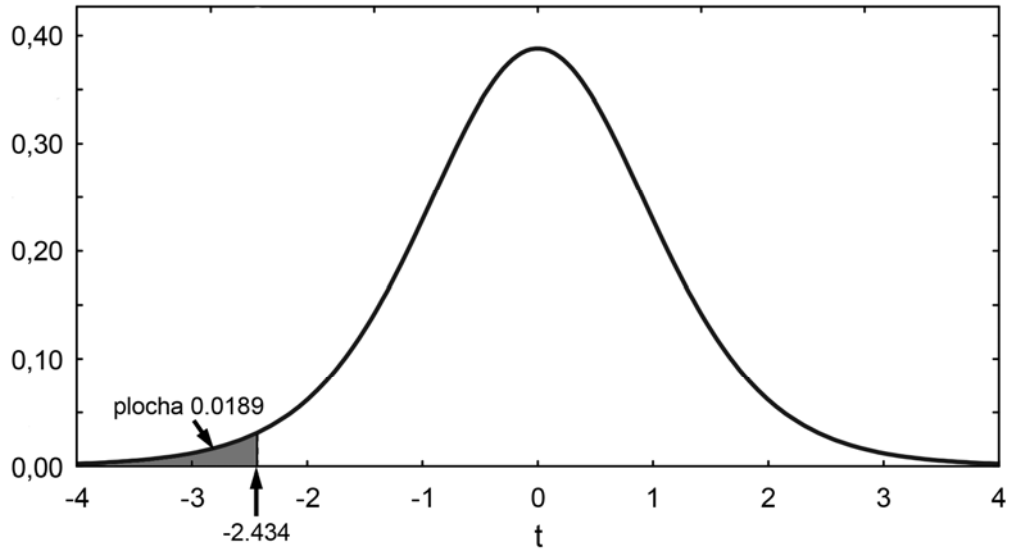
Data představují změny krevního tlaku u 10 osob po podání léku, který by měl tlak snižovat. Každá hodnota je tlak po podání mínus tlak před podáním léku: -5, -4, -3, +2, +5, -6, -1, -6, -9 a -5.
n=10, $\overline{X}$ =-3.2, s=4.158,

$s_{\overline{X}} = \frac{4.158}{\sqrt{10}} = 1.315$

$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{-3.2}{1.315} = -2.434$

df=n-1 = 9, p=0.0189
Zamítáme tedy H0 na hladině významnosti 0.0189.

Obr. 5-5 Příklad jednostranného *t* - testu.



Obr. 5-6 Křivka hustoty pravděpodobnosti t rozdělení s vyznačenou pozicí testové statistiky jednostranného testu z příkladu v Obr. 5-5

Je ale třeba poznamenat, že uvedené experimentální uspořádání nebylo správné. Pokusem jsme dokázali, že celá procedura podání léku způsobila, že v průměru tlak pacienta klesl. Ale nemůžeme na základě toho říci, zda to bylo účinnou látkou v léku podanou, nebo zda to bylo čistě díky tomu, že jsme procedurou pacienta uklidnili, a tím mu následně klesl tlak. Pokud by měl být uspořádán pokus správně, museli bychom mít ještě kontrolní skupinu, které bychom podávali placebo, a její výsledky bychom porovnávali se skupinou, které podáváme testovanou látku.

Náš příklad také hezky ilustruje potřebu rozlišovat mezi průkazností efektu a jeho velikostí. Ačkoliv jsme demonstrovali, že se tlak u zkoumaných osob průkazně snížil, průměrné snížení o 3 mm rtuťového sloupce by asi žádného lékaře nenadchlo.

Je vidět, jednostranné testy jsou silnější než odpovídající dvoustranné testy. Proto je výhodné je užít všude tam, kde to odpovídá logice věci. Užití jednostranného testu je dáno otázkou, kterou si položíme, a je často závislé na naší apriorní znalosti problému. V prvním příkladu na začátku kapitoly nemáme žádný apriorní důvod předpokládat, že by pokusné osoby měly častěji zahýbat jedním směrem. Proto je zcela namístě použít dvoustranný test. Naproti tomu, pokud je naším cílem prokázat, že určitý lék snižuje tlak, je užití dvoustranného testu zbytečné: lék nebudeme používat ani když bude výsledek testu neprůkazný, ale ani pokud by lék tlak průkazně zvyšoval.

Užití jednostranných testů někdy vede k určité podezíravosti recenzentů – často je to skutečně tak, že o jednostranném testu začneme uvažovat teprve poté, co nám

v oboustranném testu vyjde hodnota *P* mezi 0,05 a 0,1. Proto je třeba v pracech věnovat příslušnou pozornost zdůvodnění, proč jsme jej užili a neužívat jej jen proto, abychom "stlačili" *p* pod vytouženou hodnotu 0,05.

Konfidenční interval pro průměr

Z definice ve Vz. 5-4 vyplývá, že hodnota *t*, spočtená pro náhodný výběr z populace s průměrem μ, bude s pravděpodobností 95% ležet v intervalu ($-t_{0,975,\,df}$, $t_{0.975,\,df}$), kde $t_{0.975,\,df}$ je 97.5%ní kvantil t-rozdělení s příslušnými stupni volnosti. To můžeme také vyjádřit jako nerovnost

$$P\left(-t_{0,975,\,df} < \frac{\overline{X}-\mu}{s_{\overline{X}}} < t_{0.975,\,df}\right) = 0.95$$

Vz. 5-6

Po úpravách (násobíme $s_{\overline{X}}$, odečteme $\overline{X}$, pak násobíme -1 se změnou směru nerovnosti) dostáváme

$$P\left(\overline{X} - t_{0..975,\,df}.s_{\overline{X}} < \mu < \overline{X} + t_{0.975,\,df}.s_{\overline{X}}\right) = 0.95$$

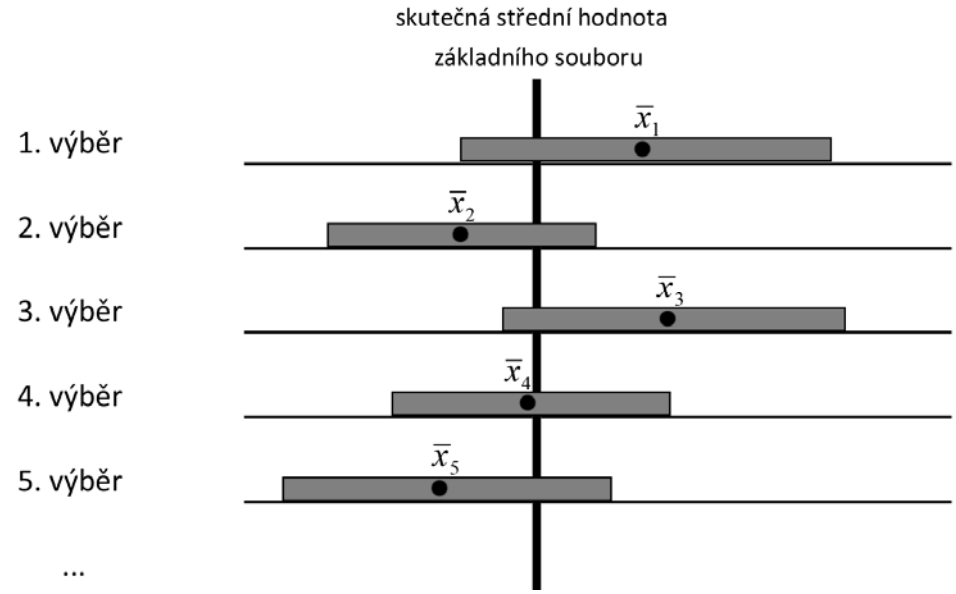
Vz. 5-7

Dostáváme tedy interval, ve kterém s 95% pravděpodobností leží skutečné střední hodnota. Tento interval se nazývá (95%-ní) **konfidenční interval** nebo též interval spolehlivosti (*confidence interval*, meze intervalu jsou *confidence limits*). Zcela analogicky můžeme sestrojit i konfidenční interval pro jinou hodnotu než 95%, nejčastěji se z dalších užívá 99%. Ve statistice se hovoří o bodových (odhadujeme jednu hodnotu) a intervalových odhadech. Biologové tuto terminologii většinou neužívají, mluví o průměru a jeho konfidenčním intervalu. Je důležité si uvědomit, že **skutečná střední hodnota základního souboru není náhodná proměnná, ta je pevná; náhodné proměnné jsou naše odhady: výběrový průměr a meze konfidenčního intervalu**. Ukazuje to Obr. 5-7.

Lze ukázat, že platí následující: pokud konfidenční interval pro průměr neobsahuje určitou hodnotu, potom je průměr na odpovídající hladině významnosti průkazně odlišný od dané hodnoty. Například jestliže 95%ní konfidenční interval pro průměr velikosti změny váhy studenta neobsahuje nulu, znamená to, že v oboustranném testu na 5%ní hladině významnosti můžeme zamítnout nulovou hypotézu, že průměrná velikost změny je rovna nule.

Předpoklady užití metod

Při odvozování Vz. 5-4 se předpokládá, že data jsou náhodným výběrem ze základního souboru s normálním rozdělením. Proto jak t-test, tak stanovení konfidenčního intervalu předpokládají normalitu dat (odhad střední chyby průměru naproti tomu předpoklad normality nemá). Metody jsou tím robustnější k narušení tohoto předpokladu, čím větší výběry máme. Je to dáno tím, že do vzorců pro t-test nevstupují původní data, ale už tam vstupují jen průměr a výběrová směrodatná odchylka. Potřebujeme tedy hlavně, aby měl výběrový průměr jako náhodná veličina normální rozdělení. A pro průměr spočtený z velkého množství dat to obvykle platí i v případě, že původní data se od normálního rozdělení odchylovala.



Obr. 5-7 Příklad, kdy z téhož základního souboru vybíráme větší počet nezávislých výběrů. Pro každý výběr dostaneme jiný výběrový průměr (černá kolečka) a spočteme jiné meze 95%-ního konfidenčního intervalu (šedé obdélníky). Pro všechny intervaly platí, že s pravděpodobností 95% pokrývají neznámou (ale pevnou) střední hodnotu základního souboru. Kdybychom na obrázku měli dvacet výběrů, můžeme očekávatat, že se jeden z nich „netrefí“ a bude ležet celý mimo skutečnou střední hodnotu.

Podáváme zprávu o variabilitě a o přesnosti odhadu

Nejběžnějším statistickým úkolem, se kterým se každý biolog setká, je podat zprávu o variabilitě nebo o přesnosti odhadu střední hodnoty. V následujícím je několik ukázek jak to můžeme udělat. O variabilitě informuje směrodatná odchylka, případně rozsah, o přesnosti střední chyba průměru, nebo konfidenční interval. **V každém případě musíme vždy uvést velikost výběru.** Známe-li velikost výběru, lze na základě znalosti jedné veličiny z trojice směrodatná odchylka, středná chyba průměru, meze konfidenčního intervalu dopočítat zbylé dvě charakteristiky. Rozsah hodnot také závisí na velikosti výběru: čím větší máme výběr, tím větší je šance, že se do něj dostanou extrémní hodnoty.

V textové nebo tabelární formě často užíváme zápis typu $\bar{x} \pm d$. Pozor, tato forma se užívá pro průměr v kombinaci se směrodatnou odchylku, tak i s jeho střední chybu ale i pro meze konfidenčního intervalu. Pokud nenapíšeme, co danou formou myslíme, může si takový zápis vykládat každý jinak. Aritmetický průměr je vhodná charakteristika polohy pro data se symetrickou distribucí, pro sešikmené distribuce (často se vyskytující v kontextu biologických dat) ale tolik informativní není, lepší informaci zde poskytuje medián: viz též Obr. 5-9 a 5-10 a také diskusi pod nimi. Ten se často doplňuje hodnotami dolního a horního kvartilu, jeho kombinace se směrodatnou odchylkou či standardní chybou aritmetického průměru nedává smysl. Medián v kombinaci s kvartily a případně i dalšími neparametrickými odhady je základem klasického *box-and-whiskers* diagramu, ale stejné grafické zobrazení můžeme použít i pro parametrické odhady (například průměr a směrodatná odchylka. V každém případě je tedy třeba takový graf popsat a vysvětlit smysl jednotlivých symbolů.

Tab. 5-1 Počet pozorování a výběrové statistiky pro koncentraci dusičnanových iontů v půdě experimentálních ploch – C – kontrola, N – přidání dusíku. *n* je počet pozorování, *SD* je výběrová směrodatná odchylka a *SE* je standardní chyba průměru.

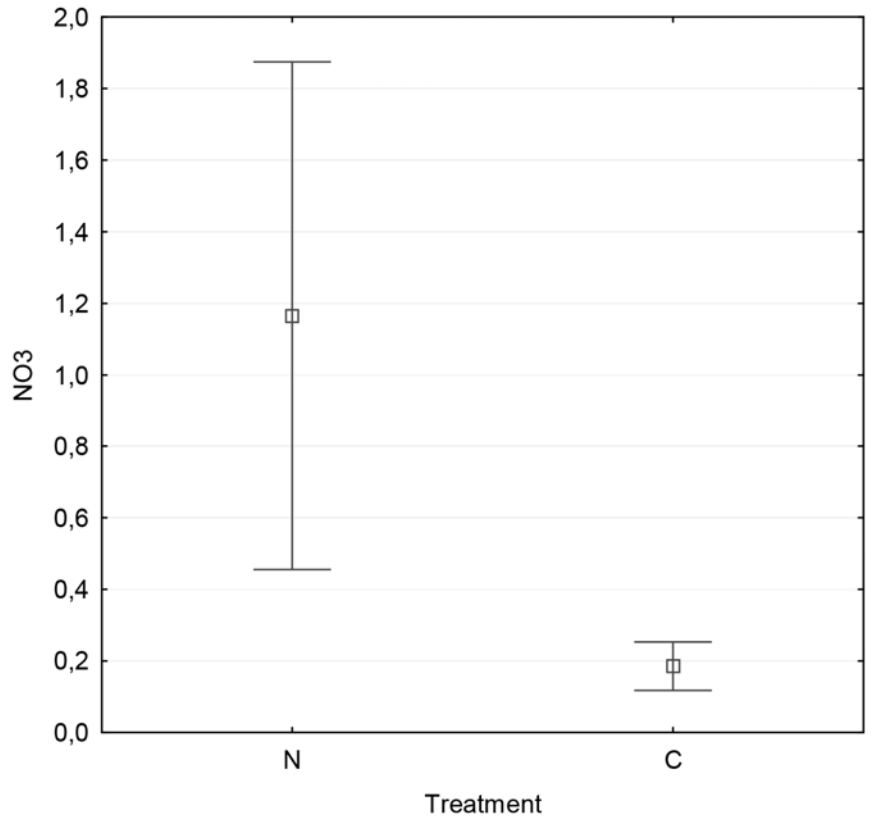
Zásah	*n*	*průměr*	*medián*	*SD*	*SE*
C	25	0.185	0.12	0.1642	0.0328
N	21	1.165	0.43	1.5593	0.3403

Tab. 5-2 Shrnutí parametrů půdní chemie v experiment (C- kontrola, N – přidání dusíku). Údaje představují průměr ± SE (střední chyba průměru), v závorce je počet pozorování. Použité jednotky: NO_3^- $mg.kg^{-1}$ suché půdy, NH_4^+ **...**

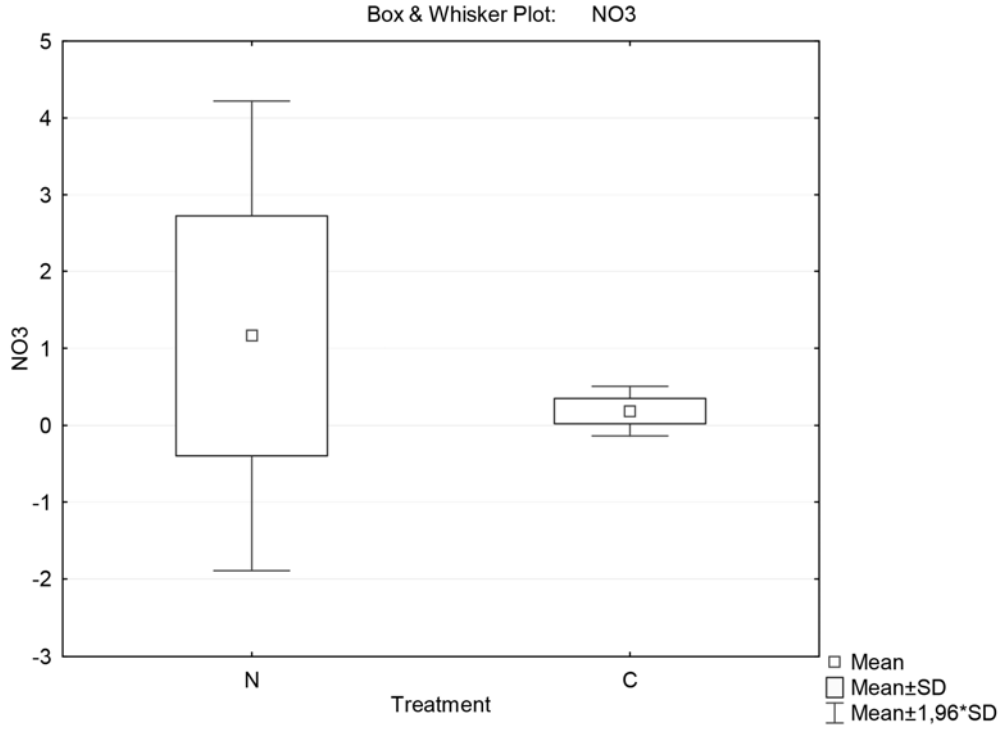
	Zásah	
	C	**N**
NO3-	0.185 ± 0.033 (25)	1.165 ± 0.340 (21)
NH4+	1.312 ± 0.636 (25)	3.863 ± 1.724 (20)
PO4-	14.278 ± 1.752 (25)	16.354 ± 1.925 (21)
celkový N	15.21 ± 2.17 (25)	83.91 ± 5.31 (25)
celkový P	2.372 ± 0.872 (25)	9.231 ± 1.382 (25)
pH	7.15 ± 0.22 (20)	7.04 ± 0.36 (20)

Výše uvedené tabulky představují dva z vícero možných způsobů prezentace sumárních statistik pro data analyzovaná pro potřeby článku či závěrečné zprávy. Obr. 5-8 až Obr. 5-10 přinášejí ukázky, jak lze čtenáře o přesnosti nebo variabilitě informovat v grafické podobě. Existují také modifikace *Box and Whisker Plot* (např. *notched BWP*), které informují zároveň o variabilitě a přesnosti. Přes tyto možnosti se ale často omezujeme na jednoduché vynesení průměru a pomocí svislé čáry přidáme jednu z charakteristik: SD, S.E. nebo meze konfidenčního intervalu (viz Obr. 5-8; vždy napište, co svislá čára znamená!).

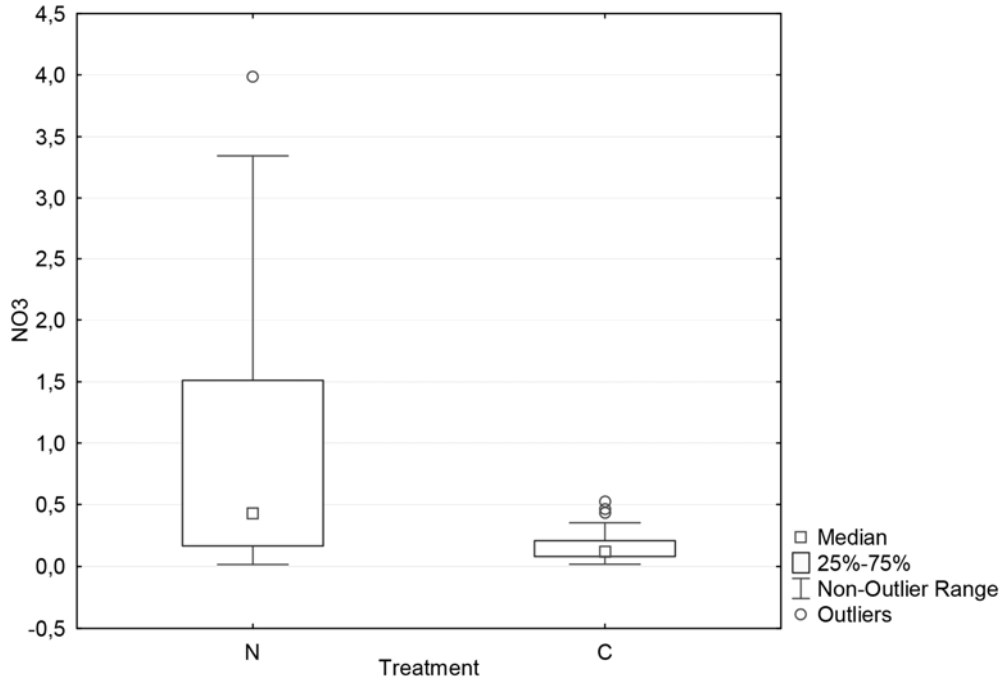
Při udávání variability se někdy vynáší průměr ±2 SD (nebo se užívá 1.96-násobek místo dvojnásobku). Důvodem je, že pokud by byl průměr skutečnou střední hodnotou a hodnota výběrové směrodatné odchylky byla směrodatnou odchylkou základního souboru a data měla normální rozdělení, potom by v uvedeném intervalu leželo přibližně 95% všech pozorování: 2.5%ní a 97.5%ní kvantily normálního rozdělení jsou -1.96, +1.96.



Obr. 5-8 Podáváme zprávu o přesnosti formou diagramu, viz též Tab. 5-1. V obrázku jsou znázorněny průměry (symboly) a 95% konfidenční interval pro půdní koncentraci NO_3^- v experimentálních plochách typu N a C. Počet pozorování byl n=21 pro N plochy a n=25 pro C plochy.



Obr. 5-9 Podáváme zprávu o variabilitě dat formou diagramu, viz též Obr. 5-10. V obrázku jsou znázorněny průměry (symboly), rozsah ±1 směrodatná odchylka od průměru (krabičky) a rozsah ±1.96 směrodatné odchylky od průměru (intervalové čáry), pro půdní koncentraci NO_3^- v experimentálních plochách typu N a C. Počet pozorování byl n=21 pro N plochy a n=25 pro C plochy.



Obr. 5-10 Podáváme zprávu o variabilitě dat formou diagramu, viz též Obr. 5-9. V obrázku jsou znázorněny mediány (symboly), rozsah od dolní kvartily (25% percentil) do horní kvartily (75% percentil) (krabičky) a rozsah přilehlých hodnot (intervalové čáry) a odlehlé hodnoty (kolečka), pro půdní koncentraci NO_3^- v experimentálních plochách typu N a C. Počet pozorování byl n=21 pro N plochy a n=25 pro C plochy.

Porovnání box-and-whisker diagramů v Obr. 5-9 a 5-10 pěkně ilustruje neadekvátnost parametrických shrnutí dat, jejichž distribuce není symetrická a tedy málo podobná normální distribuci. Rozsah hodnot, ve kterých by mělo ležet 95% pozorování pro skupinu N v Obr. 5-9 implikuje relativně častý výskyt negativních hodnot, ty ale pro koncentraci iontů nedávají smysl. Naproti tomu Obr. 5-10 ilustruje jak skutečný rozsah hodnot, tak nesymetričnost (sešikmenost) distribuce, jejímž důsledkem je mimo jiné i to, že odhady mediánu jsou mnohem nižší než odhady aritmetického průměru. Takováto data (obecněji různé typy koncentrací či hustot výskytu, ale také třeba biomasy či rozměry) by měla daleko více symetrickou distribuci a také podobnost k normální distribuci po logaritmické transformaci. Ta by také umožnila více korektní porovnání středních hodnot mezi skupinami pomocí parametrických testů, které předpokládají podobnou variabilitu hodnot ve srovnávaných skupinách (viz kapitola 6). Porovnání Obr. 5-8 (tedy zprávy o přesnosti výběru) se zbývajícími dvěma obrázky také ukazuje, že graficky prezentovaná zpráva o přesnosti spíš vizuálně přesvědčí čtenáře o rozdílech mezi skupinami, než zpráva o variabilitě - zvláště čtenáře, který si dostatečně neuvědomuje rozdíl mezi zprávou o přesnosti a o variabilitě (všechny tři obrázky jsou založeny na týchž datech).

Jak velký výběr potřebujeme?

Když se chystáme provést pokus nebo rozsáhlejší pozorování, bývá nejčastější otázkou: Jak velký má být náhodný výběr, abych dostal dostatečně přesný výsledek? Pro první přiblížení můžeme formulovat tuto úlohu takto: ze známé variance základního souboru určete velikost výběru tak, aby očekávaná střední chyba průměru byla menší než hodnota q. Víme, že $\sigma_{\bar{x}} = \sigma_X / \sqrt{n}$. Požadujme, aby $\sigma_X / \sqrt{n} < q$. Z toho plyne, že

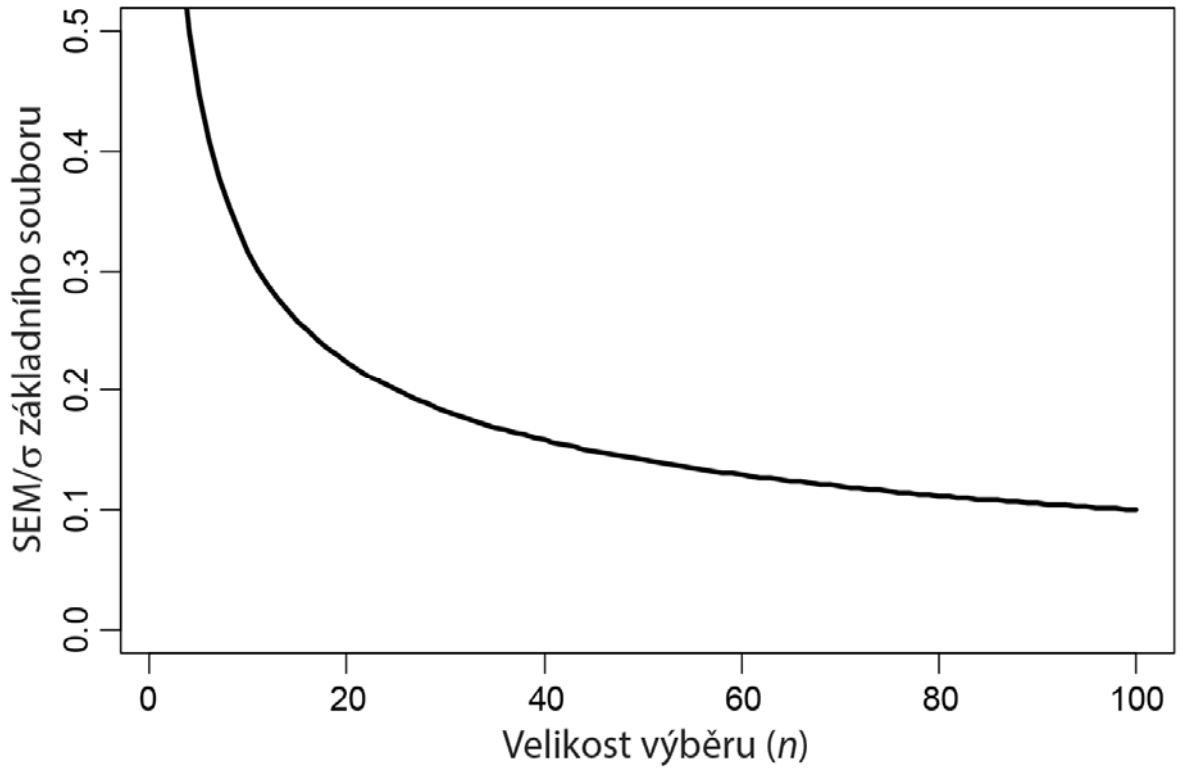
$$n > (\sigma_X / q)^2$$

Vz. 5-8

Problém je v tom, že charakteristiky základního souboru, tedy ani jeho varianci, neznáme. Nicméně máme často představu o variabilitě materiálu, se kterým pracujeme; pokud ji nemáme, je vždy třeba se před zahájením rozsáhlejšího výzkumu nebo před naplánováním pokusu o rozsahu této variability přesvědčit předběžným sledováním. Pokud tak neučiníme, můžeme se dočkat nemilých překvapení - např. výsledku nákladného pokusu, kde velká šíře konfidenčních intervalů znemožní jakoukoliv interpretaci výsledků.

Příklad: Odhadujeme (z předchozích pokusů), že směrodatná odchylka váhy pětiměsíčních krys, chovaných za standardních podmínek, je 100 g. Pro určitý pokus potřebuji, aby střední chyba průměru byla menší než 20 g. Kolik krys mám mít v pokusné skupině? Podle Vz. 5-8 dostáváme $n > (\sigma_X/q)^2$, $n > (100/20)^2$, $n > 25$. Potřebujeme tedy nejméně 25 krys.

Na tomto příkladě je také zřejmé, jak nejisté bývají údaje, které do těchto výpočtů vstupují; je proto třeba tyto výpočty považovat za orientační. Existují statisticky přesné metody, jak spočíst rozsah výběru na základě předběžného vzorku tak, aby konfidenční interval byl menší než zadaná hodnota. Na podobném základě lze spočíst doporučenou velikost výběru pro *t*-test, při daném α a dané síle testu. Víme totiž, že neprůkazný výsledek testu je nejčastěji buď důsledkem toho, že rozdíly neexistují, nebo toho, že máme málo dat. Rozumné plánování experimentů shora naznačeným způsobem nám umožní snížit pravděpodobnost druhé alternativy. Jak klesá střední chyba průměru se zvětšujícím se rozsahem výběru ukazuje Obr. 5-11.



Obr. 5-11 Střední chyba výběrového průměru v závislosti na velikosti výběru. Na svislé ose je vynesen poměr střední chyby průměru a směrodatné odchylky základního souboru.

Příkladová data

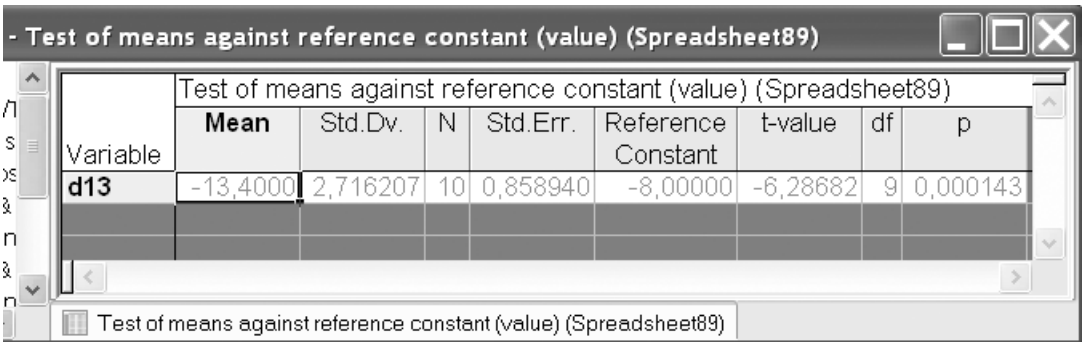
Pro ilustraci jednovýběrového t-testu užíváme příklad s $\delta^{13}C$ hodnotami deseti rostlin (viz Obr. 5-3) s daty v proměnné *d13* v listu *Chap5*, pro klasický párový t-test pak příklad z Obr. 5-5, ale s daty v podobě, v jaké by byla během experimentu zaznamenáváná (proměnné *PressBefore* a *PressAfter* představující tlak u 10 zkoumaných osob před a po podání léku).

Pro ilustraci textové a grafické prezentace středních hodnot, variability dat a přesnosti odhadu průměru používáme výsledky stanovení obsahu dusičnanových iontů (NO_3^-) v půdě, při kterém byly především srovnávány důsledky experimentálního zásahu (proměnná *Treatment*: *C* – kontrolní plochy, *N* – plochy hnojené v předchozím roce).

Jak postupovat v programu Statistica

Jednovýběrový a párový t-test

Jednovýběrový t-test spočteme v programu Statistica volbou příkazu *Statistics* | *Basic Statistics/Tables* a pak položky *t-test, single sample*. Pomocí tlačítka *Variables* vybereme proměnnou s testovanými hodnotami (*d13* v našem příkladu) a zvolíme hodnotu, se kterou chceme průměr našeho výběru srovnávat, v políčku následujícím volbu *Test all means against*. Velmi často srovnáváme proti nule, ale v našem příkladě je třeba změnit hodnotu na -8. Po zvolení tlačítka *Summary* se zobrazí tabulka s výsledky.



- Test of means against reference constant (value) (Spreadsheet89)

Test of means against reference constant (value) (Spreadsheet89)

Variable	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
d13	-13,4000	2,716207	10	0,858940	-8,00000	-6,28682	9	0,000143

Test of means against reference constant (value) (Spreadsheet89)

Výsledky testu v řádku **d13** jsou zobrazeny červeně, Statistica tak upozorňuje na zamítnutí H_0 na hladině α=0.05.

Pro výpočet párového t-testu buď můžeme zadat přímo rozdíly hodnot v párech a použít jednovýběrový t-test (viz Obr. 5-5 ilustrující postup na našich příkladových datech) nebo můžeme zadat výsledky jako dvě proměnné tak, že každý porovnávaný pár představuje jeden řádek se dvěma hodnotami těchto proměnných. V této podobě máme data v proměnných *PressBefore* a *PressAfter*. Z menu zvolíme příkaz *Statistics* | *Basic Statistics/ Tables* a v seznamu pak zvolíme *t-test, dependent samples*. Obě proměnné zadáme pomocí tlačítka *Variables* (jednu do levého seznamu, druhou do pravého) a zvolíme tlačítko *Summary*.

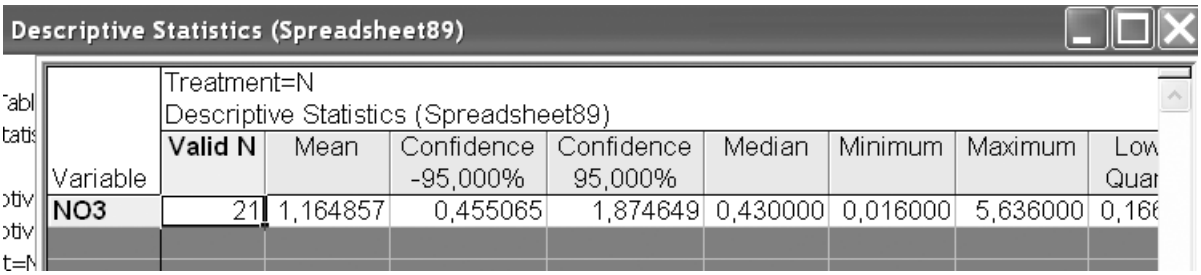
T-test for Dependent Samples (Spreadsheet89)
Marked differences are significant at p < ,05000

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p	Confidence -95,000%	Confidence +95,000%
PressBefore	88,60000	6,963396								
PressAfter	85,40000	7,515909	10	3,200000	4,157991	2,433697	9	0,037753	0,225552	6,174448

Průměrné hodnoty pro obě skupiny ve sloupci *Mean* a jejich rozdíl ve sloupci *Diff* ukazují, že při srovnání přes všechny osoby došlo ke snížení tlaku, výsledky výpočtů jsou shodné s údaji v Obr. 5-5, s výjimkou hodnoty průkaznosti *p*, kde je zobrazeno číslo 0.0378, místo očekávaného 0.0189. Je to proto, že program Statistica provádí jen dvoustranné testy. Pokud tedy testujeme jednostrannou hypotézu (H_0 je zde, že se tlak nezměnil nebo se zvýšil) a ověřili jsme si, že pozorovaná odchylka mezi skupinami je směrem, který předpokládá alternativní hypotéza (což je případ našeho příkladu), správnou hodnotu *p* získáme vydělením zobrazené hodnoty dvěma, tj. p=0.0378/2 = 0.0189. Tabulka zobrazuje také 95% interval spolehlivosti pro průměr párových rozdílů, ten ale nelze použít přímo k alternativnímu způsobu testování, s ohledem na nesymetrickou povahu testované hypotézy.

Shrnutí variability a popis přesnosti odhadu střední hodnoty

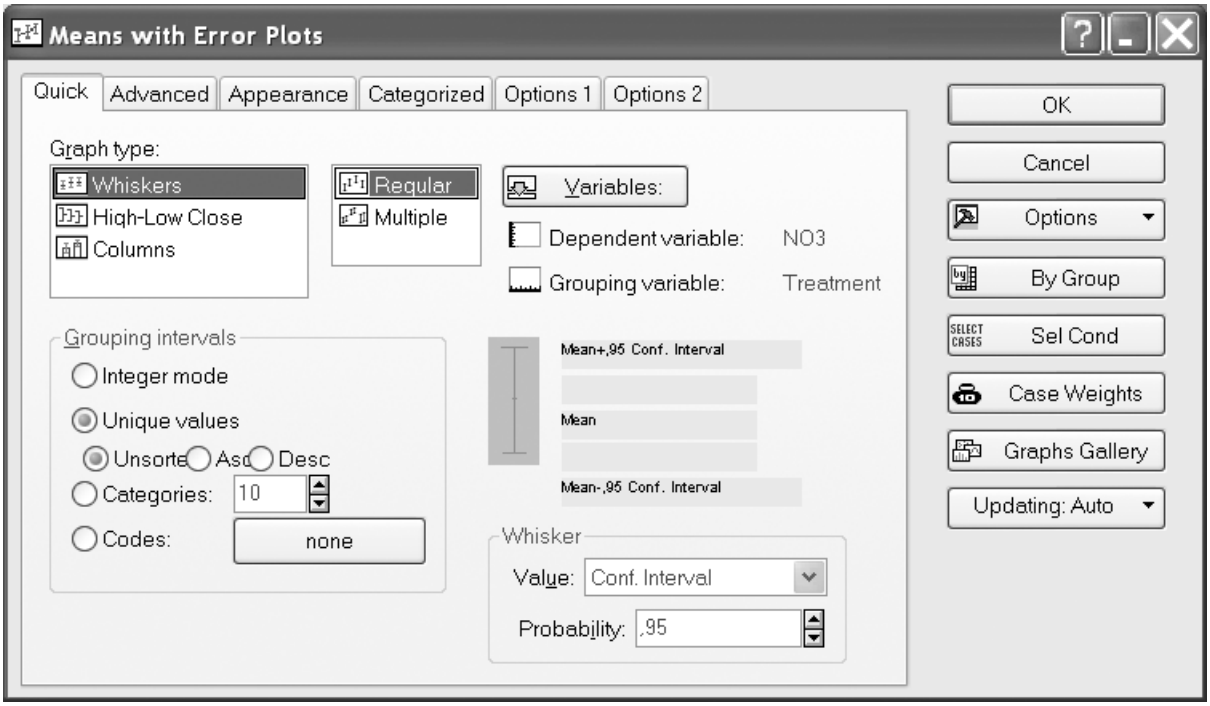
Budeme vytvářet číselná i grafická shrnutí středních hodnot, variability a také přesnosti odhadu střední hodnoty koncentrací dusičnanů (proměnná *NO3*), a to odděleně pro dvě skupiny pozorování, odpovídající dvěma typům experimentálního zásahu (proměnná *Treatment*, zásahy *C* a *N*). Číselné shrnutí získáme pomocí příkazu *Statistics | Basic Statistics/ Tables | Descriptive statistics*. V dialogovém okně zvolíme nejprve proměnnou *NO3* pomocí tlačítka *Variables*, pak na záložce *Advanced* zvolíme příslušné statistiky (pro výpočet všech diskutovaných typů statistik je třeba zvolit *Valid N*, *Mean*, *Median*, *Standard Deviation*, *Std. err. of mean*, *Conf. Limits for means*, *Minimum & maximum* a *Lower & upper quartiles*), a nakonec - chceme-li vypočítat tyto statistiky odděleně pro skupiny pozorování, jako v našem příkladě – zvolíme tlačítko *By Group* a v novém dialogovém okně zvolíme tlačítkem *Grouping Variable* proměnnou *Treatment*. V dialogovém okně *By Group* je taktéž předvolena možnost *Output "All Groups" results*, která zvolené statistiky spočte také pro spojená pozorování ze všech skupin. Nakonec v dialogu *Descriptive Statistics* zvolíme tlačítko *Summary*. Příklad výstupu (neúplného) pro skupinu pozorování se zásahem *N* ukazuje následující obrázek.


Descriptive Statistics (Spreadsheet89)

Treatment=N
Descriptive Statistics (Spreadsheet89)

Variable	Valid N	Mean	Confidence -95,000%	Confidence 95,000%	Median	Minimum	Maximum	Low Quar
NO3	21	1,164857	0,455065	1,874649	0,430000	0,016000	5,636000	0,16

Jednoduché zobrazení průměrných hodnot s konfidenčními intervaly (případně jinými charakteristikami) získáme pomocí příkazu *Graphs | Means w/Error Plots*. V záložce *Quick* zobrazeného dialogového okna nejprve zvolíme proměnnou, kterou sumarizujeme, a případně i proměnnou definující skupiny (v našem případě je to *Treatment*).



Jak je vidět z obrázku, Statistica implicitně volí zobrazení průměru s konfidenčním intervalem a klasické intervalové zobrazení (*Whiskers*), význam intervalu lze ale změnit v oblasti nazvané *Whisker* vpravo dole a vynášet například směrodatnou odchylku, střední chybu nebo celý rozsah hodnot. Na záložce *Advanced* lze zvolit ještě větší počet variant, včetně nahrazení průměru mediánem. Důležitou volbou na této záložce je také *Connect middle points*: pro většinu grafů by tato volba neměla být zaškrtnutá, protože jednotlivé skupiny představují nezávislé kategorie a spojování průměru čárou nedává smysl. V tomto dialogovém okně je také možné přidávat konfidenční intervaly (či jiné rozsahy) ke sloupečkovém diagramu. Obr. 5-8 (uvedený dříve) ale ilustruje vynášení průměrů s konfidenčními intervaly klasickou metodou.

Pro grafické znázornění variability vynášíme směrodatné odchylky (SD), případně doplněné násobkem SD (viz komentář v části o vynášení variability a přesnosti odhadu). Takový graf lze vytvořit také v rámci dialogového okna příkazu *Statistics* | *Basic Statistics/ Tables* | *Descriptive statistics*, ze záložky *Categ. plots* (pokud tedy chceme tuto informaci zobrazit pro samostatné skupiny, jako v našem příkladě). Po volbě tlačítka *Categorized box & whiker plots* se nejprve zobrazí dialogové okno, kde zvolíme (pouze) v levém seznamu proměnnou *Treatment* definující dvě kategorie (ale můžeme mít zjevně i dvojitou či trojitou klasifikaci), další dialogové okno nám umožní zvolit jen podmnožinu těchto kategorií (pro náš příklad můžeme rovnou pokračovat tlačítkem *OK*). Výsledný graf je reprodukován v Obr. 5-9.

Vynášené statistiky můžeme změnit tak, že v době, kdy je okno s grafem aktivní (v popředí) zvolíme z menu *Format* | *Graph Options* nebo můžeme dvojitě poklepat (*double-click*) na okno grafu. V zobrazeném oknu *Graph Options* zvolíme vlevo položku *Box/Whisker* a na zobrazené stránce můžeme měnit identitu středového bodu, krabičkového rozsahu a intervalového rozsahu. Jednoduchou změnu všech parametrů zobrazovaného diagramu dosáhneme změnou volby *Middle point* z hodnoty *Mean* na *Median*. Tím se změní i definice rozsahu krabiček, který teď představuje dolní a horní kvartilu, a také intervalového rozsahu (*whisker value*), představující *Non-Outlier Range* (kterému se ve statistické literatuře spíše říká *adjacent value range*). Abychom viděli v grafu plný rozsah hodnot, můžeme doplnit vynášení odlehlých hodnot (*outliers*) tak, že zmáčkneme tlačítko *More* a v dialogové

okně zvolíme v oblasti *Outliers* hodnotu *Outliers* místo *Off* (a pak zvolíme *Close* a v dalším okně *OK*). Výsledný graf je v Obr. 5-10. Za *Outliers* jsou v programu Statistica považovány takové hodnoty, které jsou o víc než 1.5 mezikvartilového rozpětí větší než horní kvartil, nebo o víc než 1.5 mezikvartilového rozpětí menší než spodní kvartil.

Jak postupovat v programu R

V příkladovém skriptu pro R jsme naimportovali – s ohledem na odlišný počet pozorování – zvlášť první tři proměnné (datový rámec *chap5A*) a zbylé dvě proměnné (datový rámec *chap5B*).

Jednovýběrový a párový t-test

Jednovýběrový i párový t-test (ale také dvouvýběrový t-test, který je popsán v následující kapitole) můžeme v programu R spočíst pomocí funkce *t.test*. Způsob zadání se pro jednotlivé typy t-testu samozřejmě liší. Program R také umožňuje, na rozdíl od programu Statistica, přímo zadat jednostranný test, včetně směru odchylky předpokládaného alternativní hypotézou. Nejprve spočteme jednovýběrový t-test, který je pro náš příklad dvoustranný.

```
> with(chap5A, t.test(d13, mu=-8))
        One Sample t-test
data:  d13
t = -6.2868, df = 9, p-value = 0.0001432
alternative hypothesis: true mean is not equal to -8
95 percent confidence interval:
 -15.34306 -11.45694
sample estimates:
mean of x
    -13.4
```

Klíčové pro zadání testu pro náš příklad je zadání hodnoty, se kterou průměr srovnáváme, pomocí parametru *mu*. Pokud bychom testovali jednostrannou hypotézu, museli bychom ještě použít parametr *alternative* (viz následující příklad).

Párový test pro náš druhý příklad můžeme spočíst následovně:

```
> with(chap5A, t.test(PressBefore,PressAfter, paired=TRUE,alternative="greater"))
        Paired t-test
data:  PressBefore and PressAfter
t = 2.4337, df = 9, p-value = 0.01888
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.7896908       Inf
sample estimates:
mean of the differences
                   3.2
```

Se dvěma porovnávanými proměnnými je nezbytné pro párový test zadat parametr *paired* s hodnotou *TRUE* (implicitní hodnotou je *FALSE* a tedy dvouvýběrový test). Náš test je navíc jednostranný a to jsme zadali pomocí parametru *alternative*, jehož hodnota udává, že alternativní hypotéza předpokládá, že hodnoty první zadané proměnné (tlak před podáním léku) jsou v párech větší než hodnoty proměnné druhé. Výstup funkce také (na rozdíl od programu Statistica) zobrazuje odhad intervalu spolehlivosti, který odpovídá jednostrannosti naší hypotézy.

Shrnutí variability a popis přesnosti odhadu střední hodnoty

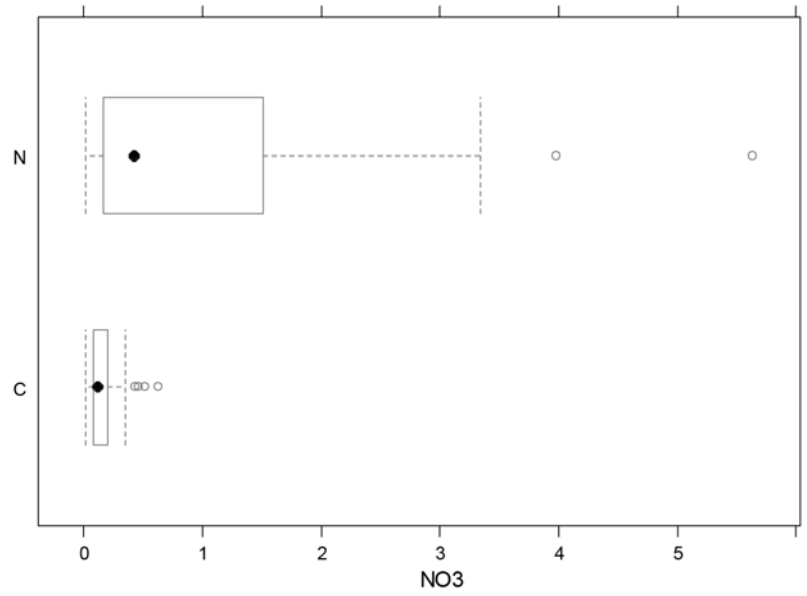
Program R obsahuje více verzí funkce zobrazující klasický box-and-whisker plot, nejsnáze dostupná je funkce *boxplot*, kterou lze užít takto:

```
> boxplot(NO3~Treatment, data=chap5B)
```

Pokročilejší zobrazení lze vytvořit pomocí funkce *bwplot* v knihovně *lattice*:

```
> library(lattice)
> bwplot(NO3~Treatment, data=chap5B)
> bwplot(Treatment~NO3, data=chap5B)
```

Odlišné pořadí proměnných *NO3* a *Treatment* ve vzorci, který je prvním parametrem funkce *bwplot*, umožňuje zvolit vertikální nebo horizontální orientaci sumarizované proměnné v grafu. Druhé volání například vytvoří takovýto diagram:



Diagramy vynášející parametrické statistiky (průměry s intervaly spolehlivosti či standardními chybami) lze vytvořit například pomocí funkce *plotmeans* v knihovně *gplots*:

```
> plotmeans(NO3~Treatment, data=chap5B,connect=F)
```

Při výše uvedeném způsobu zadání obsahuje graf také informaci o počtu pozorování použitých pro výpočet statistik.

Popis analýz v článku

Methods

Stable isotope $\delta^{13}C$ values obtained for experimental plants were compared with isotope concentration in the atmosphere using a one-sample t-test.

Decrease of systolic blood pressure after drug administration to 10 patients was tested using a one-sided paired t-test.

Shrnující charakteristiky typu průměru, mediánu, konfidenčních intervalů atd. jsou považovány za obecně známe, takže se jejich výpočet v metodice obvykle nepopisuje; v řadě časopisů by totéž platilo i o t-testech.

Results

The average of $\delta^{13}C$ values in plants was -13.4, significantly lower then -8, the air concentration value (t_9=-6.287, p=0.000143).

Blood pressure decreased significantly (t_9=-2.434, p=0.0189) due to drug administration, but the average size of the effect (decrease by 3.2 mm Hg) was quite small. *(v diskusi bychom pak museli napsat, že to nemuselo být jen účinnou látkou, ale mohl to být placebo efekt)*.

Doporučená četba

Sokal & Rohlf (1981), pp. 128-178, Zar (2007), pp. 97-121, Quinn & Keough (2002), pp. 17-22, 35-37.

6 Porovnání dvou výběrů

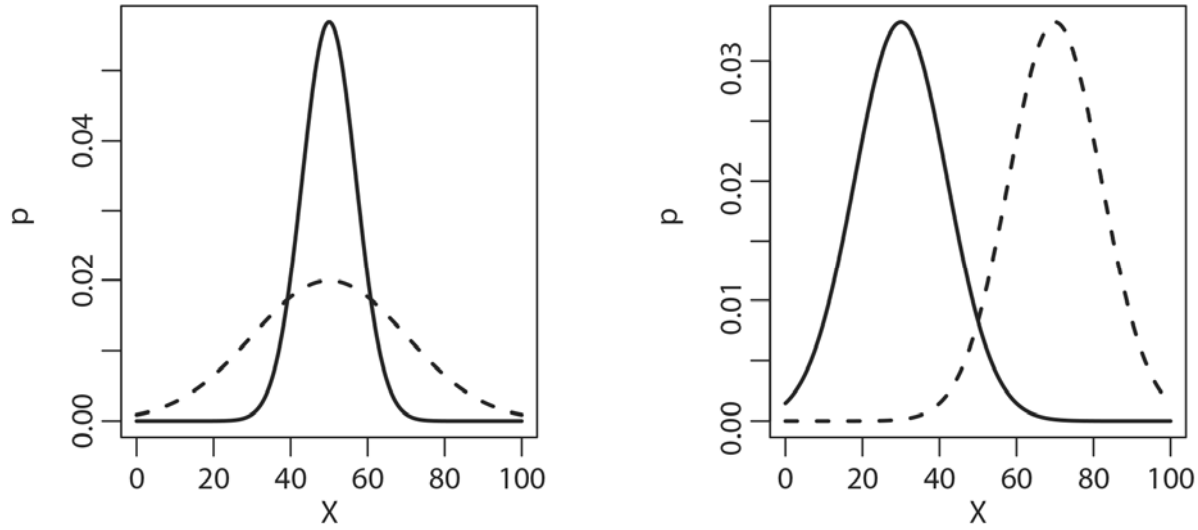
Příklady problémů:

1. Dvacet rostlin bylo pěstováno v běžné půdě a dvacet rostlin v půdě obohacené fosforem, každá rostlina v jednom květináči. Po dvou měsících byly tyto rostliny sklizeny, vysušeny do konstantní váhy a zváženy. Má obohacení půdy fosforem vliv na biomasu rostlin, tj. liší se od sebe průkazně tyto dva soubory?

2. V populaci rostlin se vyskytují diploidní a tetraploidní jedinci. U náhodně vybraných jedinců jsme změřili délku prašníků a zjistili jejich ploidii. Naším cílem je zjistit, zda se liší délka prašníků mezi diploidy a tetraploidy.

3. Určitý druh můr byl chytán do lapačů dvou různých konstrukcí. Jedenáct lapačů prvního typu a osm lapačů druhého typu bylo náhodně rozmístěno do porostu. Po noci fungování byl spočten počet ulovených můr v každém lapači. Liší se tyto dva typy lapačů v účinnosti?

Předpokládejme, že oba srovnávané vzorky jsou výběry ze základních souborů charakterizovaných normálním rozdělením. Tyto základní soubory se mohou lišit průměrem a/nebo variancí (viz Obr. 6-1). Obojí je možné testovat.



Obr. 6-1 Hustota pravděpodobnosti dvou normálních rozdělení, lišících se variancí (vlevo) a střední hodnotou (vpravo).

Testování rozdílů ve varianci

Máme dva výběry ze základních souborů a ptáme se, zda se variance základních souborů liší. Nulová hypotéza tedy zní: H_0: $\sigma_1^2=\sigma_2^2$, alternativní hypotéza je H_A: $\sigma_1^2 \neq \sigma_2^2$. Test provádíme následujícím způsobem: spočteme výběrové variance s_1^2 a s_2^2 a ptáme se, jaká je pravděpodobnost, že dva výběry se budou lišit ve svých variancích tolik jako nebo více než naše výběry za předpokladu, že pocházejí ze základních souborů se stejnými variancemi. Pokud je tato pravděpodobnost malá (obvykle 0.05 nebo 0.01, jako v dříve popsaných testech), zamítáme nulovou hypotézu. Pokud je pravděpodobnost větší, nemáme dostatek důkazů, abychom nulovou hypotézu zamítli. K tomu nám slouží F test. Hodnotu F spočteme jako poměr větší výběrové variance k menší. Tedy, je-li $s_1^2 > s_2^2$, potom

$$F = \frac{s_1^2}{s_2^2}$$

Vz. 6-1

Testová statistika (hodnota F) má v případě platnosti nulové hypotézy F rozdělení (jméno rozdělení je podle Sira R.A. Fishera, jednoho ze zakladatelů statistiky). Toto rozdělení závisí (podobně jako např. χ^2 - rozdělení) na počtu stupňů volnosti. Protože se ale jedná o charakteristiku závislou na dvou výběrech, jsou zde dva počty stupňů volnosti: prvý pro výběr, jehož variance je v čitateli (*numerator df*) a druhý pro výběr, jehož variance je ve jmenovateli (*denominator df*). Počet stupňů volnosti je pro každý výběr jeho velikost minus 1, tedy *n-1*. Je-li spočtená hodnota F statistiky vyšší než kritická hodnota F rozdělení pro danou hladinu významnosti a pro dané počty stupňů volnosti, zamítáme nulovou hypotézu na dané hladině významnosti. Pro oboustranný test (testujeme nulovou hypotézu $\sigma_1^2 = \sigma_2^2$) je kritickou hodnotou pro α-procetní hladinu významnosti (např. 0.05) (1 - α/2) kvantil (např. tedy 0.975) – α dělíme dvěma, protože jsme se arbitrárně rozhodli, že větší hodnota variance bude v čitateli. Statistický software ale udává ke každé spočtené hodnotě *F* přímo odpovídající pravděpodobnost *p* (dosaženou hladinu významnosti). Postup výpočtu ukazuje příklad na Obr. 6-2

Dvoustranný test na poměr variancí pro hypotézy H_0: $\sigma_1^2=\sigma_2^2$ a H_A: $\sigma_1^2\neq\sigma_2^2$ Data jsou délky prašníků [mm] u diploidních a tetraploidních jedinců v populaci určitého druhu pryskyřníku.
H_0: $\sigma_1^2=\sigma_2^2$
H_A: $\sigma_1^2\neq\sigma_2^2$

Diplodní jedinci: 2.9, 3.1, 3.2, 2.8, 2.9, 3.3, 3.4, 2.8, 2.7, 3.0, 3.1
Tetraploidní jedinci: 3.5, 3.8, 3.7, 3.8, 3.7, 3.5, 3.6, 3.9
$n_1 = 11$, $df_1 = 10$
$n_2 = 8$, $df_2 = 7$
$s_1^2 = 0.0496$ mm^2 $s_2^2 = 0.0213$ mm^2
$F = \frac{s_1^2}{s_2^2} = \frac{0.0496}{0,0213} = 2.336$
$p = 0.273$ a proto nezamítáme H_0 (kritická hodnota je 4.76).
Protože se variance neliší, můžeme odhadnout společnou varianci obou souborů jako
$s_p^2=(0.496 + 0.1491) / (10 + 7) = 0.0379$ mm^2.

Obr. 6-2 Příklad na oboustranný test poměru variancí.

Pokud nemůžeme zamítnout nulovou hypotézu a rozhodneme se odhadnout společnou (*pooled*) varianci pro oba výběry (s_p^2), použijeme vzorce

$$s_p^2 = \frac{df_1.s_1^2 + df_2.s_2^2}{df_1 + df_2} = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Vz. 6-2

Zde je třeba upozornit, že test je dosti slabý, zvláště při malých výběrech. Pokud bychom porovnávali výběry o velikosti 10 (v biologii často větší výběry nemáme), potom kritická hodnota pro oboustranný F-test při $\alpha = 0.05$ je 4.026: to znamená, že jedna hodnota výběrové variance musí být 4-krát větší než druhá, abychom mohli nulovou hypotézu o shodě variancí zamítnout. Proto musíme být velmi opatrní se závěrem, že oba výběry pocházejí z rozdělení se stejnou variancí, pravděpodobnost chyby druhého druhu je vysoká.

Popsaný test je oboustranný, lze ale užívat i jednostranné testy: např. chceme dokázat, že variabilita (charakterizovaná variancí) velikosti snůšky vajec je větší v přírodě než u ptáků chovaných v zajetí. Potom je nulová hypotéza formulována takto: variance velikosti snůšky

v přírodě je stejná nebo menší než u ptáků chovaných v zajetí. Pro test na hladině významnosti α je pak kritickou hodnotou kvantil $(1-\alpha)$. Po té, co nám statistický program odhadne hodnotu *p* pro oboustranný test, přesvědčíme se nejprve, že výběrové variance odpovídají svoji relativní velikostí naší alternativní hypotéze, a pokud ano, získáme správný odhad významnosti pro jednostranný test vydělením hodnoty *p* dvěma.

Lze spočítat i konfidenční interval pro poměr variancí. $1-\alpha$ konfidenční interval obsahuje hodnotu 1 právě tehdy, když nulovou hypotézu o rovnosti variancí nelze při dvoustranném testu zamítnout na hladině významnosti α. Je-li tedy 95%-ní konfidenční interval (interval spolehlivosti) pro poměr variancí (0.49, 9.23) jako v našem příkladě (Obr. 6-2), potom to znamená, že nulovou hypotézu o rovnosti variancí nemůžeme zamítnout na 5%-ní hladině významnosti. Naopak, pokud by např. 99%-ní interval spolehlivosti byl (pro jiná data) (1.45, 6.54), můžeme nulovou hypotézu zamítnout na 1%-ní hladině významnosti ($p < 0.01$).

Porovnání průměrů

K porovnání průměrů dvou výběrů slouží *t*-test pro dva výběry. Jedná se o pravděpodobně nejčastěji používaný statistický test vůbec. Jeho předpokladem je, že oba výběry pocházejí ze základních souborů, ve kterých má sledovaná proměnná normální rozdělení. V základní verzi test předpokládá, že oba výběry pocházejí ze základních souborů se stejnou variancí, ale existuje několik řešení problému, když se porovnávané soubory liší variancí. Testujeme nulovou hypotézu H_0: $\mu_1 = \mu_2$ proti alternativní H_A: $\mu_1 \neq \mu_2$. V základní verzi použijeme vzorec

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}1-\overline{X}2}}$$

Vz. 6-3

V tomto případě je $s_{\overline{X}1-\overline{X}2}$ střední chyba rozdílu průměrů. Protože předpokládáme, že variance je stejná pro oba základní soubory, použijeme pro odhad společné variance s_p^2 vzorec Vz. 6-2 a střední chybu rozdílu průměrů odhadneme jako

$$s_{\overline{X}1-\overline{X}2} = \sqrt{\frac{s^2{}_p}{n_1} + \frac{s^2{}_p}{n_2}}$$

Vz. 6-4

Vz. 6-3 má potom tvar

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2{}_p}{n_1} + \frac{s^2{}_p}{n_2}}}$$

Vz. 6-5

Počet stupňů volnosti je součtem počtu stupňů volnosti pro oba výběry, tedy $(n_1-1) + (n_2-1) = n_1 + n_2 - 2$. Nulovou hypotézu zamítáme, pokud spočtená významnost (hodnota p) je menší než zvolené α (tedy pokud vypočtená statistika přesáhne kritickou hodnotu na dané hladině významnosti). Použití základního oboustranného *t*-testu ukazuje příklad na Obr. 6-3, kdy jsou porovnávány délky prašníků dvou ploidních úrovní (z Obr. 6-2):

Dvouvýběrový *t*-test pro oboustranné hypotézy H_0: $\mu_1 = \mu_2$ a H_A: $\mu_1 \neq \mu_2$. Tyto hypotézy lze také vyjádřit jako H_0: $\mu_1 - \mu_2 = 0$ a H_A: $\mu_1 - \mu_2 \neq 0$).
Data jsou opět délky prašníků (v mm) u diploidů a tetraploidů určitého druhu pryskyřníku.
Diplodní jedinci: 2.9, 3.1, 3.2, 2.8, 2.9, 3.3, 3.4, 2.8, 2.7, 3.0, 3.1
Tetraploidní jedinci: 3.5, 3.8, 3.7, 3.8, 3.7, 3.5, 3.6, 3.9

$n_1 = 11$	$n_2 = 8$
$df_1 = 10$	$df_2 = 7$
$\overline{X}_1 = 3.02$ mm	$\overline{X}_2 = 3.69$ mm
$SS_1 = 0.4964$ mm	$SS_2 = 0.1488$ mm

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{0.4964 + 0.1488}{10 + 7} = \frac{0.6451}{17} = 0.0379 \text{ mm}^2$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{0.0379}{11} + \frac{0.0379}{8}} = \sqrt{0.00345 + 0.004744} = \sqrt{0.008193} = 0.0905 mm$$

$$t = \frac{\overline{X}_1 + \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{3.02 - 3.69}{0.0905} = \frac{-0.67}{0.0905} = -7.394$$

$p<0.001$, zamítáme proto H_0 (kritická hodnota je $t_{0.05(2),17} = 2.11$).

Obr. 6-3 Dvouvýběrový *t*-test pro oboustrannou hypotézu

Zcela analogicky jako při jednovýběrovém testu lze použít i jednostranný test místo testu dvoustranného. Potom testujeme nulovou hypotézu H_0: $\mu_1 \leq \mu_2$ proti alternativě H_0: $\mu_1 > \mu_2$ (nebo naopak).

Narušení předpokladů: Při tomto testu předpokládáme, že oba výběry pocházejí ze základních souborů s normálním rozdělením a se stejnou variancí. Naštěstí je známo, že i dosti velké narušení těchto předpokladů výsledky testů příliš neovlivní (říkáme, že test je vůči narušení těchto předpokladů **robustní**), zvláště pokud je velikost výběrů dostatečná a pokud jsou oba výběry přibližně stejně velké. Pokud se variance porovnávaných výběrů výrazně liší, bývá zvykem užít výpočetní vzorec známý jako „Welchovo přibližné *t*“:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}}$$

Vz. 6-6

s přibližným počtem stupňů volnosti

$$df = \frac{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^2}{\frac{\left(\frac{s1^2}{n1}\right)^2}{n1-1} + \frac{\left(\frac{s2^2}{n2}\right)^2}{n2-1}}$$

Vz. 6-7

Hodnota spočtená podle Vz. 6-7 nemusí být celé číslo, například pro příklad v Obr. 6-2 je to 16.882 (a hodnota *t* statistiky je -7.905, $p<0.0001$). Užívají se i další vzorce pro přibližné *t*, pokud se variance nerovnají.

Varování: Dvouvýběrový *t*-test používáme, pokud chceme srovnat **pouze** dva výběry. Pokud máme výběrů více než dva, musíme užít analýzu variance s následným mnohonásobným porovnáním. Nelze tedy testovat *t*-testem postupně všechny páry výběrů, protože pravděpodobnost chyby 1. druhu je α v každém testu a testy na sobě nejsou nezávislé. I v případě, že platí nulová hypotéza, je při porovnávání mnoha výběrů pravděpodobnost, že nalezneme průkazný rozdíl alespoň v jednom páru, velmi vysoká a roste s počtem prováděných testů.

Příkladová data

Jde o příklad u kterého je srovnávána délka prašníku u 11 diploidních a 8 tetraploidních jedinců jednoho druhu. Stejná data jsou použita pro srovnání variancí i pro srovnání průměrů, ostatně takto (a v tomto pořadí) obvykle při aplikaci dvouvýběrového testu postupujeme. Data jsou uložena v listu *Chap6* souboru *biostat-data.xlsx*. Proměnná *Ploidy* určuje, ke které skupině jedinec v daném řádku patří, proměnná *Anther* pak pro daného jedince udává délku prašníku. Pro potřeby F-testu i dvouvýběrového t-testu doporučujeme používat toto uspořádání dat, protože to pak přirozeně přechází do testování rozdílů mezi více skupinami, například pomocí analýzy variance.

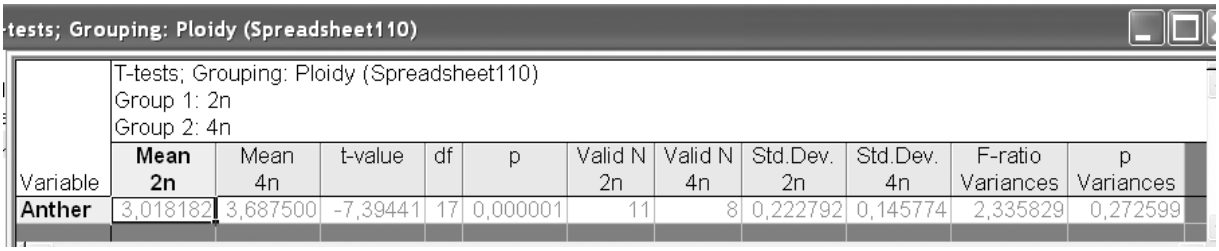
Jak postupovat v programu Statistica

F test rovnosti variancí

Ověření shody variancí se v programu Statistica provádí společně s dvouvýběrovým t-testem, viz následující sekce.

Dvouvýběrový t test rovnosti průměrů

Zvolíme z menu příkaz *Statistics* | *Basic Statistics/Tables* | *t-test, independent, by groups*. V dialogovém okně nejprve zadáme pomocí tlačítka *Variables* proměnnou s hodnotami (*Anther*, v levém seznamu) a proměnnou definující skupiny (*Ploidy*, v pravém seznamu). Po návratu do hlavního dialogového okna zvolíme tlačítko *Summary*. Výsledky jsou zobrazeny v jednom řádku tabulky, nejprve výsledky dvouvýběrového testu srovnávajícího střední hodnoty, následované výsledky F-testu srovnávajícího variance.


-tests; Grouping: Ploidy (Spreadsheet110)

T-tests; Grouping: Ploidy (Spreadsheet110)
Group 1: 2n
Group 2: 4n

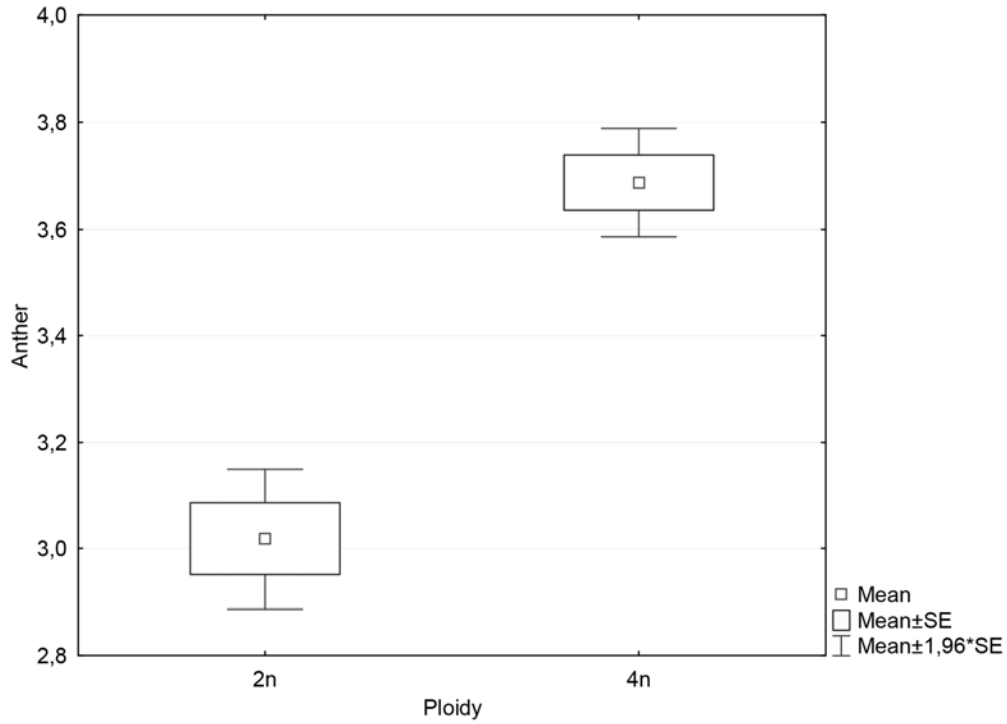
Variable	Mean 2n	Mean 4n	t-value	df	p	Valid N 2n	Valid N 4n	Std.Dev. 2n	Std.Dev. 4n	F-ratio Variances	p Variances
Anther	3,018182	3,687500	-7,39441	17	0,000001	11	8	0,222792	0,145774	2,335829	0,272599

Podíváme se nejprve na F-test na pravé straně okna. Hodnota testové statistiky je ve sloupečku *F-ratio variance* a pravděpodobnost chyby 1. typu vycházející z porovnání této statistiky s F distribucí s 10 a 7 stupni volnosti (pro náš příklad) je v posledním sloupci *p Variances* (viz též Obr. 6-2).

Výsledky dvoustranného dvouvýběrového t-testu jsou pak reprezentovány (krom dílčích informací o průměrech každé skupiny a počtech pozorování) hodnotou testové statistiky ve sloupci *t-value*, počty stupňů volnosti ve sloupci *df* a významností testu (chybou 1. typu) ve sloupci *p*. Výsledky lze porovnat s těmi v Obr. 6-3. Pokud bychom nemohli předpokládat shodu variancí ve srovnávaných dvouch skupinách, lze naši hypotézu přibližně otestovat za užití korekce: v dialogovém okně pro t-test musíme na záložce *Options* zaškrtnout volbu *Test w/ separate variance estimates*.

V rámci záložek *Quick* a *Advanced* můžeme také použít tlačítko *Box & whisker plot*, které nám dává možnost vytvořit grafickou prezentaci dvouvýběrového t-testu. V následující nabídce *Box-Whisker Type* je pak asi nejlepší možností volbou *Mean/SE/1.96*SE*, která nám pro dostatečně velká data (větší, než pro náš příklad, aproximace je slušná pro n>60)

představuje pomocí „whiskers“ 95% interval spolehlivosti pro střední hodnoty porovnávaných skupin.



Obr. 6-3 Rozdíly **v délce prašníků** u diploidů a tetraploidů (t_{17}=-7.394, p<0.001)

Toto vynesení ovšem není příliš informativní – informuje jen o čtyřech číslech – dvou průměrech a dvou středních chybách průměru (dvvojnásobek střední chyby si člověk lehce představí). Pokud se počty pozorování hodně liší, může být užitečné vynést průměr, střední chybu, a směrodatnou odchylku: při různých počtech pozorování je poměr směrodatné odchylky a střední chyby průměru různý.

Jak postupovat v programu R

Data opět doporučujeme importovat ve stejném uspořádání, v jakém jsou v listu *Chap5* příkladových dat, tj. s jednou číselnou proměnnou představující hodnoty obou skupin a s jedním faktorem představujícím přiřazení pozorování do skupin. Zde je příklad naimportovaného datového rámce *chap6*:

```
> summary(chap6)
 Ploidy      Anther
 2n:11   Min.   :2.70
 4n: 8   1st Qu.:2.95
         Median :3.30
         Mean   :3.30
         3rd Qu.:3.65
         Max.   :3.90
```

F test rovnosti variancí

Shodu variancí můžeme otestovat pomocí funkce *var.test* takto:

```
> var.test(Anther~Ploidy, data=chap6)
        F test to compare two variances
data:  Anther by Ploidy
F = 2.3358, num df = 10, denom df = 7, p-value = 0.2726
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
 0.4906053 9.2261131
sample estimates:
ratio of variances
          2.335829
```

V případě potřeby může testovat i jednostrannou hypotézou správnou volbou hodnoty parametru *alternative* (*"less"* nebo "*greater*").

Dvouvýběrový t test rovnosti průměrů

Klasický dvouvýběrový t-test spočteme následujícím způsobem:

```
> t.test(Anther~Ploidy,data=chap6,var.equal=T)
        Two Sample t-test
data:  Anther by Ploidy
t = -7.3944, df = 17, p-value = 1.048e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8602919 -0.4783445
sample estimates:
mean in group 2n mean in group 4n
        3.018182         3.687500
```

Jde o test oboustranné (symetrické) hypotézy; pokud bychom chtěli testovat jednostrannou hypotézu, opět musíme zvolit neimplicitní hodnotu parametru *alternative*. Pokud bychom nemohli předpokládat shodu variancí, můžeme spočíst (přibližný) Welch-ův test vynechání parametru *var.equal* (takže bude mít svou implicitní hodnotu *FALSE*):

```
> t.test(Anther~Ploidy,data=chap6,var.equal=F)

        Welch Two Sample t-test

data:  Anther by Ploidy
t = -7.9052, df = 16.882, p-value = 4.5e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8480473 -0.4905891
sample estimates:
mean in group 2n mean in group 4n
        3.018182         3.687500
```

Popis analýz v článku

Methods

The homogeneity of variances was tested for the two compared groups of diploid and tetraploid specimens using F-ratio test.

The difference in the mean length of anthers for diploid and tetraploid specimens was tested using two-sample t-value test [using Welch correction].

Pro řadu časopisů je ale t-test natolik triviální, že stačí uvést v Results *výsledek, a předpokládáme, že čtenář z toho pochopí, že jsme prováděli t-test.*

Results

We have found no differences in the variance of diploid and tetraploid plants ($F_{10,7}$=2.336, n.s.).

The diploid and tetraploid specimens differed significantly in the length of their anthers (t_{17}=-7.394, $p<0.001$) with average anther length, respectively for diploids and tetraploids, 3.0 and 3.7 mm.

Doporučená četba

Sokal & Rohlf (1981), pp. 170 - 179, Zar (2007), pp. 122-149, Quinn & Keough (2002), pp. 37-42.

7 Neparametrické metody

Jak jednovýběrový (případně párový) *t*-test a dvouvýběrový *t*-test, tak i *F*-test vycházejí z předpokladu normálního rozdělení základního souboru. Ve všech těchto testech formulujeme a testujeme hypotézy o **parametrech** rozdělení. Proto se tyto testy nazývají parametrické. Tyto testy jsou naštěstí robustní, což znamená, že malé odchylky od předpokladů mají zanedbatelný vliv. Co však dělat, když jsou odchylky opravdu velké?

Zde máme v zásadě tři možnosti: (1) jsou známy určité transformace dat, které (za určitých předpokladů) změní rozdělení dat na normální nebo normálnímu blízké; těmito transformacemi se budeme zabývat později; nebo (2) rozdělení dat, se kterými pracujeme, je blízké nějakému jinému rozdělení, které je k dispozici v rámci tzv. zobecněných lineárních modelů (*generalized linear models*), které jsou blíže popsány v kapitole XX; nebo (3) můžeme užít neparametrické metody (*nonparametric methods*).

Většina parametrických metod má své neparametrické protějšky. Velká část neparametrických metod je založena na pořadí. K neparametrickým metodám můžeme zařadit i tzv. permutační metody. Tyto metody také nezávisí na předpokladech o rozdělení dat. Bude o nich zmínka na závěr této kapitoly.

Mann-Whitney(ův) test

Tento test je neparametrickou obdobou dvouvýběrového *t*-testu. Testujeme nulovou hypotézu: oba výběry pocházejí ze stejného rozdělení (už nikoliv hypotézu, že střední hodnoty jsou stejné). Test provádíme tak, že všechny hodnoty z obou výběrů seřadíme; nejvyšší hodnota má pořadí 1, nejnižší pořadí *N*. $N = n_1 + n_2$, tj. součet počtu pozorování v prvním a ve druhém výběru. Pokud mají některá pozorování stejné hodnoty, přiřadí se všem stejným hodnotám průměr jejich pořadí. Např. sledujeme počet vyklíčených semenáčů v ošetřených a neošetřených plochách. Počty z obou ploch seřadíme společně. Pozorované počty mohou být např. 17, 15, 12, 12, 12, 10, ...; tyto hodnoty pak dostanou pořadí 1, 2, 4, 4, 4, 6, ...(4 je průměr z hodnot 3, 4, a 5). Poté spočteme statistiku

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Vz. 7-1

kde R_1 je součet pořadí v prvním výběru. Zcela analogicky můžeme spočíst statistiku

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Vz. 7-2

Platí ovšem vztah

$$U + U' = n_1 n_2,$$

Vz. 7-3

takže v praxi počítáme jen jednu z nich podle Vz. 7-1 nebo Vz. 7-2, a zbylou dopočteme ze Vz. 7-3. Při oboustranném testu potom vybereme menší z hodnot *U, U'* a tu porovnáme s distribucí, kterou mají U statistiky za platnosti nulové hypotézy, čímž stanovíme pravděpodobnost, že testová statistika z této distribuce pochází (*p*). Distribuci hodnot U lze odhadnout v případě relativně malých souborů dat exaktně tak, že za pomoci počítače

vytvoříme všechna možná rozmístění pozorovaných hodnot do dvou skupin shodné velikosti jakou mají ty naše a spočteme pro každou takovou kombinaci hodnotu U. Postup ilustruje příklad porovnání zdravotního stavu dvou kultivarů smrku pod vlivem znečištěného ovzduší v Obr. 7-1.

Zdravotní stav stromků byl odhadnut pomocí přibližné stupnice s hodnotami od 1 (zcela zdravý strom) do 5 (zcela mrtvý strom).
H_0: Zdravotní stav jedinců se neliší mezi dvěma kultivary A a B.
H_A: Zdravotní stav jedinců se liší mezi kultivary A a B.

Zdravotní stav jedinců kultivaru A: 2, 2, 1, 2, 3, 4, 2, 3, 1, 5 $n_1=10$
Zdravotní stav jedinců kultivaru B: 4, 5, 3, 1, 4, 3, 5, 2, 1, 2 $n_2=10$

Po převodu na pořadí (s průměrným pořadím pro jedince se shodnou hodnotou zdravotního stavu) dostáváme:
Kultivar A: 13.5, 13.5, 18.5, 13.5, 8.5, 5.0, 13.5, 8.5, 18.5, 2.0 $R_1=115.0$
Kultivar B: 5.0, 2.0, 8.5, 18.5, 5.0, 8.5, 2.0, 13.5, 18.5, 13.5 $R_2 = 95.0$

$U = n_1n_2 + n_1(n_1+1)/2 - R_1 = (10)(10) + (10)(11)/2 - 115 = 100 + 55 - 115 = 40$
$U'=n_1n_2 - U = (10)(10) - 40 = 60$
$p=0.481$ (exaktní odhad), H_0 proto nezamítáme: zdravotní stav stromků nezávisí na kultivaru.

Obr. 7-1 Mann-Whitney-ův test pro neparametrické testování oboustranné hypotézy, že není rozdíl ve zdravotním stavu stromků lišících se svým kultivarem.

Pokud užíváme jednostranný test (opět platí, že kritická hodnota pro oboustranný test na hladině významnosti α je rovna kritické hodnotě pro jednostranný test na hladině $\alpha/2$), užijeme U v případě, že chceme dokázat, že skupina 1 je větší (tzn. že nulová hypotéza zní: velikost ve skupině 2 je stejná nebo větší než ve skupině 1) a U', pokud chceme dokázat opak.

Poznámka a varování: shora řečené platí pouze, pokud řadíme hodnoty od největší k nejmenší; je jistě možné řadit hodnoty opačně, tj. od nejmenší k největší; potom je v jednostranných testech nutno zaměnit U a U'. Spíše než mechanické určení, kdy užít U a kdy U', doporučuji logickou úvahu: jsou-li hodnoty v daném výběru velké, je součet jejich pořadí při řazení odzhora malý a odpovídající hodnota statistiky U, popř. U' je vysoká.

Pokud jsou výběry dostatečně velké (řekněme přes 50 pozorování v obou výběrech), potom má za platnosti nulové hypotézy U přibližně normální rozdělení se střední hodnotou

$$\mu_U = \frac{n_1 n_2}{2}$$

Vz. 7-4

a směrodatnou odchylkou

$$\sigma_U = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

Vz. 7-5

($N=n_1+n_2$ je celkový počet pozorování). Hodnota $Z = (U-\mu_U)/\sigma_U$ má tedy přibližně normované normální rozdělení. Někdy se také užívá tzv. korekce na kontinuitu, tzn. že od absolutní hodnoty rozdílu $U-\mu_U$ odečítáme 0.5.

Na rozdíl od *t*-testu můžeme užít Mann-Whitney-ův test i pro data na ordinální škále (stupnici), tak jak to ukazuje náš příklad.

Zde je třeba upozornit, že Mann-Whitney test testuje nulovou hypotézu o shodě rozdělení, ze kterého pocházejí porovnávané výběry. Pokud testujeme nulovou hypotézu o shodě polohy (tedy o mediánu nebo průměru), musíme předpokládat, že se distribuce neliší tvarem.

Poznámka: Protože Wilcoxon položil základy těmto testům, objevuje se často jeho jméno i u Mann-Whitney testu jako Wilcoxon-Mann-Whitney, někdy i Wilcoxonův test pro dva výběry, apod. Někdy se k totožným výsledkům dochází různými výpočetními postupy.

Mediánový test: Ještě jednodušší možností je spočíst společný medián a pomocí čtyřpolní tabulky porovnat počet pozorování nad společným mediánem a pod společným mediánem v prvním a druhém výběru. Tento test testuje přímo hypotézu o rovnosti mediánů, je však velmi slabý.

Wilcoxonův test pro párová pozorování

Jako je Mann-Whitney(ův) test neparametrickou obdobou dvouvýběrového t-testu, Wilcoxonův test je obdobou párového *t*-testu. Test provádíme tak, že spočteme nejprve rozdíly mezi hodnotami pozorování v párech pozorování s nulovým rozdílem vyloučíme a potom tyto rozdíly seřadíme podle velikosti jejich absolutní hodnoty od nejmenšího k největšímu. Všimněme si, že prvním krokem je zde počítání rozdílů a teprve potom tyto rozdíly převedeme na pořadí. Znamená to, že při užití testu předpokládáme, že můžeme původní hodnoty odečítat – proto hovoříme v příkladu v Obr. 7-2 o počtu přidělených bodů. Pokud bychom použili tento test pro data na ordinální škále, tedy např. pro zdravotní stav (viz příklad v Obr. 7-1), naznačujeme tím, že jsou konstantní rozdíly mezi kategoriemi, tj. považujeme rozdíl mezi zdravotním stavem 1 a 2 (tj. zcela zdravý strom versus mírně nemocný strom) shodný s rozdílem stupňů 4 a 5 (tj. kriticky nemocný strom a mrtvý strom). Tím vlastně nejprve převedeme ordinální stupnici, mírně násilným způsobem, na stupnici intervalovou. Nicméně totéž děláme, když počítáme "průměrnou známku za studium".

Pro rozdíly stejné absolutní velikosti užijeme průměrné pořadí, jako v předcházejícím testu. Poté spočteme součet pořadí kladných a součet pořadí záporných rozdílů (označujeme je T_+ a T_-). Protože součet řady čísel 1 až n je $n(n+1)/2$, lze snadno dopočítat $T_+=n(n+1)/2-T_-$. Menší z hodnot lze opět buď porovnat se známým rozdělením této statistiky nebo (pro větší vzorky) užít aproximaci normálním rozdělením, kde

$$\mu_T = \frac{n(n+1)}{4}$$

Vz. 7-6

a

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Vz. 7-7

a tedy spočítat hodnotu Z a tu porovnat s normovaným normálním rozdělením N(0,1).

Použití testu ukazuje příklad v Obr. 7-2, kde byla porovnávána hodnocení dvou expertních skupin.

Dvě skupiny expertů měly hodnotit úspěšnost agro-environmentálních opatření na jednotlivých farmách. Každá skupina měla možnost přidělit podle standardizovaného protokolu každé farmě 0 (nejhorší) až 10 bodů (nejlepsí). Aby došlo ke standardizaci, v pilotním pokusu hodnotily obě skupiny, na sobě nezávisle, deset vybraných farem. Zatímco předpokládáme, že určitá variabilita hodnocení je nevyhnutelná, chceme zjistit, zda mezi skupinami nejsou systematické rozdíly, tj. zda jedna skupina nehodnotí systematicky lépe než druhá.
H_0: Hodnocení oběma skupinami se systematicky neliší.
H_A:V hodnocení je systematický posun mezi oběma expertními skupinami.

Farma	Exp. skupina 1	Exp. skupina 2	Rozdíl	Pořadí	Pořadí se znaménkem
1	5	7	-2	6.5	-6.5
2	4	5	-1	3	-3.0
3	1	1	0	xxx	xxx
4	8	9	-1	3	-3.0
5	7	6	+1	3	+3.0
6	3	5	-2	6.5	-6.5
7	1	4	-3	8	-8.0
8	0	4	-4	9	-9.0
9	9	10	-1	3	-3.0
10	2	3	-1	3	-3.0

n=10,
$T_+ = 3.0$
$T_- = 6.5 + 3.0 + 3.0 + 6.5 + 8.0 + 9.0 + 3.0 + 3.0 = 42.0$
p=0.021, H_0 tedy zamítáme: hodnocení dvěma skupinami expertů se systematicky liší.

Obr. 7-2 Wilcoxonův párový test aplikovaný na data srovnávající dvě expertní skupiny.

Tento test nepředpokládá normální rozdělení, ale na rozdíl od předešlého testu předpokládá, že rozdělení rozdílů je symetrické okolo mediánu (což bývá často splněno). Pokud tomu tak není, je možné užít tzv. **znaménkový test**, který porovnává počet kladných a záporných rozdílů: Pokud se soubory neliší, předpokládáme stejný počet kladných i záporných diferencí. Tento test je ovšem velmi slabý (např. v našem příkladě dostáváme p=0.0455), je ho ale možné užít v případech, kdy nelze užít ani párový *t*-test, ani Wilcoxonův test.

Užívání testů založených na pořadí

Názory na užití neparametrických testů se liší. Někteří autoři volají po častějším využití těchto testu při analýze biologických dat (Potvin & Roff 1993), jiní autoři (např. Johnson 1995) ale upozorňují na jejich často nesprávnou interpretaci (například Mann-Whitneyův test nelze interpretovat jako test rozdílu ve středních hodnotách, testuje rozdíly v celém rozdělení) a také na nesprávnou interpretaci předpokladů nezbytných pro použití parametrických testů (jen průměry musí mít normální distribuci, ne nutně analyzovaná data).

Mann-Whitney(ův) a Wilkoxonův test jsou jedny z nejužívanějších a nejsilnějších neparametrických testů. Pokud jsou splněny předpoklady parametrického testu (tj. *t*-testu), je parametrický test o něco silnější. Pokud jsou výrazně narušeny, je silnější a spolehlivější neparametrický test.

Obecně lze říci, že pokud si jsme rozumně jisti, že nejsou příliš narušeny předpoklady parametrického testu, je lepší užít parametrický test. Dále také platí, že vliv narušení předpokladů parametrického testu je větší u malých výběrů.

Permutační testy

Klasické testy jsou založeny na předpokladu, že známe rozdělení testové statistiky v případě platnosti nulové hypotézy - např. víme, že testová statistika má t-rozdělení s určitým počtem stupňů volnosti. V případě t-testu, ale i mnoha jiných testů, je pak ale většinou nezbytné, aby srovnávané základní soubory, ze kterých výběry (jejich skupiny v případě porovnávání skupin) pocházejí, měly normální rozdělení. Testy v rámci zobecněných lineárních modelů umožňují předpokládat, že výběry pocházejí ze základních souborů s jiným typem rozdělení (např. Poissonovo či binomické). Pokud nechceme spoléhat na předpoklady o typu rozdělení, můžeme si rozdělení testové statistiky za předpokladu nulové hypotézy „náhodně nasimulovat".

Jak bychom to provedli v případě porovnání dvou výběrů? Použijeme náš příklad z kapitoly 6 s délkami prašníků pro 10 diploidních a 10 tetraploidních jedinců. Máme tedy dvacet hodnot délky. Spočteme klasickou hodnotu testového kriteria *t*. Nyní každé z dvaceti čísel napíšeme na malé lístečky a vhodíme do klobouku. Se zavřenýma očima vytáhneme deset lístečků a budeme je považovat za délky prašníků u diploidních jedinců, zbylých deset za délky u tetraploidů. Spočteme pro ně hodnotu testového kriteria. Lístečky vrátíme, zamícháme a znova taháme deset diploidů a znova spočteme hodnotu *t* pro pozorování náhodně rozdělená do dvou skupin. Tahání lístečků z klobouku za nás může obstarat počítač a generátor náhodných čísel. Jestliže náhodné rozdělení souboru provedeme tisíckrát, získáme slušnou představu o tom, jaké je rozdělení testové statistiky za předpokladu, že se výšky ve skupinách neliší.

Podíváme se, v kolika procentech vyšla absolutní hodnota testového kriteria rovná nebo vyšší než je absolutní hodnota *t* statistiky spočtená z reálných dat. Tato procenta nám dávají odhad pravděpodobnosti, s jakou dostaneme z dat spočtenou hodnotu testového kriteria za předpokladu, že platí nulová hypotéza, a můžeme je tedy považovat za dosaženou hladinu významnosti. Přesněji se dosažená hladina významnosti odhaduje jako $\frac{x+1}{n+1}$, kde x je počet případů, kdy absolutní hodnota testového kriteria pro náhodnou simulaci vyšla vyšší nebo rovná absolutní hodnotě v datech a n je počet náhodných permutací. Jestliže by nám vyšla absolutní hodnota *t* vyšší než v datech ve 14 případech z 999 náhodných permutací, odhadujeme, že dosažená hladina významnosti p=0.015.

Protože pracujeme s absolutními hodnotami, dostáváme dvoustranný test (odhadujeme velikost obou „ocasů" rozdělení). Jednostranný test bychom provedli tak, že nebudeme pracovat s absolutními hodnotami a budeme odhadovat velikost toho „ocasu" rozdělení, který nás zajímá. Uvedený příklad se týkal porovnání dvou výběrů. Pro každý problém musíme zkonstruovat jiný typ náhodných permutací, a to tak, aby odpovídal nulové hypotéze. V praxi se používají permutační testy spíše pro složitější problémy, než je problém porovnání jednorozměrné proměnné u dvou výběrů.

Příkladová data

Data pro příklady této kapitoly jsou v listu *Chap7*. Pro ilustraci neparametrického dvouvýběrového testu použijeme data ze studie srovnávající zdravotní stav stromků dvou kultivarů, popsaná v Obr. 7-1. Ta jsou pro oba programy zadána tak, že všechny údaje o zdravotním stavu jsou v jedném proměnné (*Health*), zatímco proměnná *Cultivar* kóduje pomocí textu A a B příslušnost každého pozorování k jednomu z kultivarů.

Neparametrický párový test používá příklad z Obr. 7-2, srovnávající hodnocení dvěma skupinami expertů. Pro tento typ testů je vpodstatě nezbytné (u obou statistických programů) zadávat data jako dvě proměnné, ve kterých jsou uvedené párované údaje ve stejném řádku.

Jak postupovat v programu Statistica

Mann-Whitneyův test

Z menu zvolíme příkaz *Statistics* | *Nonparametrics* a v seznamu vybereme *Comparing two independent samples (groups)*. Po kliknutí tlačítka *OK* zadáme v zobrazeném dialogovém okně (*Comparing Two Groups*) proměnné (*Variables*): proměnnou *Health* zadáme v levém seznamu, proměnnou *Cultivar* v pravém. Pak zvolíme tlačítko *Mann-Whitney U test* (nebo vpravo nahoře tlačítko *M-W U test*). Statistica zobrazí výsledky tímto způsobem:

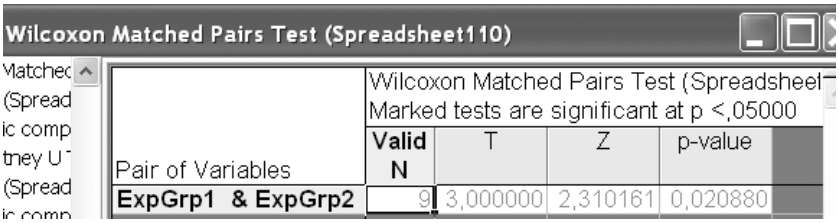
Mann-Whitney U Test (Spreadsheet110)
By variable Cultivar
Marked tests are significant at p <,05000

variable	Rank Sum A	Rank Sum B	U	Z	p-value	Z adjusted	p-value	Valid N A	**Valid N B**	2*1sided exact p
Health	95,00000	115,0000	40,00000	-0,718132	0,472676	-0,735770	0,461871	10	10	0,481251

Spočtené součty pořadí jsou v prvních dvou sloupcích, nižší z hodnot U statistiky je ve třetím, aproximace normálním rozdělením a odpovídající průkaznost jsou ve sloupcích *Z* a *p-value*, ale asi přesnější jsou údaje z následujících dvou sloupců (*Z adjusted* a za ním následující *p-value*, kde je použita korekce na diskontinuitu). Pro menší velikosti souborů (jako je ten náš) Statistica počítá i přesné *p* založené na výčtu (v posledním sloupci *2*1sided exact p*) a pokud je přítomno, představuje asi nejlepší odhad pravděpodobnosti chyby 1. typu a proto jej doporučujeme užít.

Wilcoxonův párový test

Z menu zvolíme příkaz *Statistics* | *Nonparametrics* a v seznamu vybereme položku *Comparing two dependent samples (variables*). Po zobrazení dialogového okna zadáme pomocí tlačítka *Variables* proměnnou *ExpGrp1* v jednom seznamu a *ExpGrp2* v druhém a po návratu do původního okna zvolíme tlačítko *Wilcoxon matched pair test* (znaménkový test bychom spočetli pomocí tlačítka *Sign test*).



Statistica zobrazila hodnotu T_+ a její transformaci do Z statistiky, pro kterou pak spočetla průkaznost.

Jak postupovat v programu R

Protože mají oba příklady odlišné uspořádání dat a tedy odlišné délky proměnných, importovali jsme první dva sloupce do datového rámce *chap7a* a další dva do datového rámce *chap7b*.

Mann-Whitney-ův test

Vyhodnocení provádíme stejnou funkcí jako pro Wilcoxonův párový test, jen nezadáme parameter *paired*:

```
> wilcox.test(Health~Cultivar,data=chap7a)
        Wilcoxon rank sum test with continuity correction
data:  Health by Cultivar
W = 40, p-value = 0.4619
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(x = c(2L, 2L, 1L, 2L, 3L, 4L, 2L, 3L, 1L,  :
  cannot compute exact p-value with ties
```

Funkce automaticky provádí korekci na spojitost, pokud bychom si to tak nepřáli, lze přidat parametr *correct=FALSE*. Zajímavou možností je výpočet intervalu spolehlivosti pro medián rozdílů mezi srovnávanými vzorky:

```
> wilcox.test(Health~Cultivar,data=chap7a,conf.int=T)
        Wilcoxon rank sum test with continuity correction
...
95 percent confidence interval:
 -2.0000486  0.9999876
sample estimates:
difference in location
           -0.03741339
```

Wilcoxonův párový test

Tento test provedeme obdobně jako Mann-Whitneyův test, ale dvě proměnné se srovnávanými výběry zadáváme jako první dva parametry a musíme také přidat parametr *paired=TRUE*. V této podobě (bez vzorečku modelu) funkce nerozpoznává parametr *data*, používáme proto vnořené volání s funkcí *with*:

```
> with(chap7b, wilcox.test(ExpGrp1,ExpGrp2,paired=TRUE))
        Wilcoxon signed rank test with continuity correction
data:  ExpGrp1 and ExpGrp2
V = 3, p-value = 0.02182
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(ExpGrp1, ExpGrp2, paired = TRUE) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(ExpGrp1, ExpGrp2, paired = TRUE) :
  cannot compute exact p-value with zeroes
```

Permutační testy

Princip konstrukce permutačních testů v programu R ukážeme na příkladu dvouvýběrového testu shody medianů. Nejprve definujeme příslušnou funkci (smysl jednotlivých příkazů v jazyce S je naznačen v komentářích následujících znak #):

```
> permtest.two.groups <- function(x,y,N=9999)
{
  xy <- c(x,y)         # spolecny vektor
  len1  <- length(x)   # velikost prvni skupiny
  len2  <- length(y)   # velikost druhe skupiny
  len12 <- len1 + len2 # celkova velikost
  diffs <- numeric(N+1)# rozdily mezi skupinami
  diffs[1] <- median(x) - median(y) # pozorovana hodnota
  for( i in 2:(N+1))
  { # vyber podmnozinu odpovidajici prvni skupine
    idx <- sample( 1:len12, size=len1, replace=F)
    xx  <- xy[idx]  # simulace prvni skupiny
    yy  <- xy[-idx] # simulace druhe skupiny
    # rozdil v medianech
```

```
    diffs[i] <- median(xx) - median(yy)
  }
  # odhad p: kolik rozdilu je vetsich nez pozorovany?
  mean(abs(diffs) >= abs(diffs[1]))
}
```

Funkci pak můžeme použít například takto:

```
> with(chap7a, permtest.two.groups(Health[1:10],Health[11:20]))
[1] 0.6611
```

Jde o Monte Carlo odhad průkaznosti (pro náš není rozdíl v mediánech průkazně odlišný od nuly), tj. vybírá se jen určitá podmnožina z možných přiřazení jednotlivých pozorování do dvou skupin. Při opakovaném volání se stejnými daty proto funkce vrací různé odhady (0.6489, 0.6514, 0.6521,...). Test jednostranné hypotézy bychom dosáhli změnou posledního příkazu, kdy bychom zadali příslušný typ nerovnosti bez použití funkce *abs*.

Popis analýz v článku

Methods

We have used Mann-Whitney U test, with *p*-value estimated by normal approximation with a continuity correction to test the hypothesis about the identity of the health response of trees of the two compared cultivars.

nebo, pokud máme k dispozici exaktní hodnotu p:

We have used Mann-Whitney U test with an exact Type I error estimate (based on enumerating all possible value assignments into the groups) to test the hypothesis ...

We have used the Wilcoxon matched-pair test with p-value estimated by normal approximation to test for judgement differences between the two compared expert groups

Results

The distribution of health status values does not differ significantly between the two cultivars (Z=-0.7357, n.s.).

We found a significant bias between the two expert groups in their scoring of the success of agri-environmental measures (Z=2.311, p=0.0209).

Doporučená četba

Zar (2007): pp. 138-146, 153-156, Sokal & Rohlf (1969): pp. 387-395, Quinn & Keough (2002), pp. 45-48.

D. H. Johnson (1995): Statistical sirens: the allure of nonparametrics. *Ecology* **76**: 1998-2000.

C. Potvin & D. A. Roff (1993): Distribution-free and robust statistical methods: viable alternatives to parametric statistics? *Ecology* **74**: 1617-1628.

8 Analýza variance (ANOVA): jednoduché třídění

Příklady problémů:

1. Pokusní králíci byli náhodně rozděleni do tří skupin. Prvá skupina byla krmena standardní potravou, druhá skupina potravou obohacenou vápníkem a třetí skupina potravou obohacenou železem. Po měsíci byl u každého králíka proveden rozbor krve. Otázkou je, zda se liší počty červených krvinek u králíků živených různými typy stravy.

2. Semena kostřavy červené byla získána z pěti různých populací. Poté byly ze semen vzešlé rostliny pěstovány (každé individuum ve zvláštním květináči) za standardních podmínek a u každého individua byla zjišťována určitá bionomická nebo morfologická charakteristika (např. počet výběžků). Liší se sledovaná charakteristika mezi populacemi rostlin, pocházejícími z různých populací?

3. Byl zjišťován obsah cukru v plodech tří různých kultivarů jabloní, pěstovaných v témže sadu. Ze sklizně bylo náhodně vybráno 10 plodů z každé odrůdy (ne více než jeden plod z každého stromu) a v nich provedena analýza obsahu cukru. Liší se obsah cukru v plodech uvedených třech kultivarů?

Jak jsme si ukázali, pro porovnání dvou průměrů používáme *t*-test. Pokud porovnáváme více než dva průměry, nelze *t*-test (každý s každým) použít. Je tomu tak proto, že pravděpodobnost chyby prvního druhu je α v každém páru, a pravděpodobnost, že v případě platnosti nulové hypotézy nalezneme průkazný rozdíl alespoň v jednom páru, je potom podstatně vyšší než α. Jednotlivé testy nejsou nezávislé, proto je poněkud obtížnější pravděpodobnost chyby prvního druhu alespoň v jednom testu odhadnout. Udává ji Tab. 8-1.

	Hladina signifikance užívaná v *t* testech			
Počet průměrů (k)	0.05	0.02	0.01	0.001
2	0.05	0.02	0.01	0.001
3	0.13	0.05	0.03	0.003
4	0.21	0.09	0.05	0.006
5	0.23	0.13	0.07	0.009
10	0.63	0.37	0.23	0.034
20	0.92	0.71	0.52	0.109
∞	1.00	1.00	1.00	1.00

Tab. 8-1 Pravděpodobnost, že se dopustíme chyby I. druhu, budeme-li užívat více *t* testů při hledání rozdílů mezi všemi páry ve skupině *k* průměrů.

Chceme-li testovat hypotézu H_0: $\mu_1=\mu_2=\mu_3= ...= \mu_k$, kde k je počet porovnávaných skupin, použijeme analýzu variance (angl. *analysis of variance*, ANOVA; výraz ANOVA se běžně používá i v češtině, i když spíše nespisovně; někteří jej dokonce i skloňují). Obecně je analýza variance vlastně celým odvětvím statistiky, které je schopno testovat i složitější hypotézy, a zároveň je součástí většího komplexu metod, kterým se říká obecné lineární modely, *general linear models*. Ve své nejjednodušší podobě ANOVA testuje shora naznačenou hypotézu H_0: $\mu_1=\mu_2=\mu_3= ...= \mu_k$. To znamená, že máme objekty klasifikované (zařazené do skupin) podle hodnot jednoho faktoru. Proto se této analýze říká analýza variance jednoduchého třídění, častěji v češtině jednocestná analýza variance (angl. *single factor ANOVA, one-way ANOVA*).

V příkladu 1 je klasifikačním faktorem typ stravy. O každém typu stravy říkáme, že je to *hladina* faktoru. V nejjednodušší podobě jsou experimentální skupiny náhodnými, vzájemně nezávislými výběry: mluvíme o zcela náhodném experimentálním uspořádání (*completely randomized experimental design*). Rozdíly mezi skupinami můžeme demonstrovat pouze pokud známe variabilitu uvnitř skupin. Princip analýzy variance můžeme přibližně vysvětlit takto: testujeme nulovou hypotézu, že se střední hodnoty mezi sebou neliší. Protože předpokladem ANOVy je rovnost variancí, můžeme si představit, že za platnosti nulové hypotézy se jedná o několik výběrů z téhož souboru. Varianci základního souboru odhadneme na základě variancí uvnitř jednotlivých skupin. Na základě této variance jsme schopni předpovědět, jaká je variabilita mezi skupinami. Tuto předpověď potom porovnáme se skutečnou variabilitou mezi skupinami. Pokud je variabilita mezi skupinami nepravděpodobně velká (otestujeme *F*-testem), potom zamítáme nulovou hypotézu o rovnosti průměrů.

Předpoklady: V analýze variance předpokládáme, že všechny výběry pocházejí ze základního souboru s normálním rozdělením a jednotlivé výběry pocházejí ze základních souborů se stejnou variancí. Tyto předpoklady jsou důležité pro *F*-test.

Výpočet

Předpokládejme, že máme pokusné objekty rozděleny do k pokusných skupin (někdy nazývaných třídami). $X_{i,j}$ je hodnota sledované proměnné na j-tém objektu v i-té skupině ($i=1, .., k$). V pokusu 1 je tedy $X_{2,4}$ počet krvinek u čtvrtého králíka ve druhé pokusné skupině. Počet objektů v i-té skupině je n_i, celkový počet pozorování je $N = \sum_{i=1}^{k} n_i$. Průměrná hodnota ve skupině i se značí $\overline{X}_i$, průměr všech hodnot ve všech skupinách jako $\overline{X}$. Princip analýzy variance spočívá v porovnání variability uvnitř skupin s variabilitou mezi skupinami. Hodnotu uvnitř skupin charakterizuje „průměrný čtverec odchylky od průměru uvnitř skupin“, nazývaný též reziduální (*within group mean square, error mean square* MS_E). Získáme jej tak, že nejprve spočteme tzv. součet čtverců (uvnitř skupin, reziduální, *within group sum of squares, error sum of squares* nebo *residual sum of squares*) odchylek od průměrů skupin (budeme jej značit SS_E)

$$SS_E = \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i \right)^2 \right)$$

Vz. 8-1

Odpovídající počet stupňů volnosti DF_E je součtem počtu stupňů volnosti v jednotlivých skupinách:

$$DF_E = \sum_{i=1}^{k} (n_i - 1) = N - k$$

Vz. 8-2

Průměrný čtverec

$$MS_E = \frac{SS_E}{DF_E}$$

Vz. 8-3

je také odhadem společné variance σ^2 ve všech skupinách. Připomeňme, že stejným způsobem jsme odhadovali společnou varianci v *t*-testu a že podobně jako v *t*-testu

předpokládáme rovnost variancí ve skupinách. Obdobně variabilitu mezi skupinami charakterizují součet čtverců a průměrný čtverec mezi skupinami (budeme je označovat s indexem G, *among group sum of squares, among group degrees of freedom*, atd., často jen *group sum of squares*). Součet čtverců získáme součtem čtverců odchylek průměrů skupin od celkového průměru, násobených počtem objektů ve skupině:

$$SS_G = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$$

Vz. 8-4

$$DF_G = k - 1$$

Vz. 8-5

$$MS_G = \frac{SS_G}{DF_G}$$

Vz. 8-6

Lze ukázat, že pokud platí nulová hypotéza o rovnosti průměrů, potom i MS_G je odhadem společné variance σ^2. Pokud neplatí, je variabilita mezi průměry skupin (MS_G) podstatně větší než uvnitř skupin (MS_E). K jejich porovnání můžeme použít *F*-testu (připomeňme, že *F*-test jsme používali pro porovnání variancí dvou výběrů). Pokud tedy nulová hypotéza platí, má podíl $F = MS_G / MS_E$ *F*-rozdělení s odpovídajícími počty stupňů volnosti DF_G, DF_E. Porovnáním s tímto rozdělením můžeme tedy určit pravděpodobnost, s jakou by bylo dosaženo stejné nebo vyšší hodnoty *F* za předpokladu, že nulová hypotéza platí a spočítat tak hodnotu *p*. V případě, že je toto *p* menší než zvolené α (např. 0.05), zamítáme nulovou hypotézu.

V analýze variance bývá zvykem uvádět ještě celkovou sumu čtverců a odpovídající stupně volnosti (*total sum of squares*, SS_{TOT}). Tu spočteme jako výběrovou varianci celého souboru, tedy

$$SS_{TOT} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X} \right)^2$$

Vz. 8-7

$$DF_{TOT} = N - 1$$

Vz. 8-8

V analýze variance se tyto hodnoty uvádějí tradičně, i když se přímo k testu nepoužívají. Používaly se při výpočtech, platí totiž

$$SS_{TOT} = SS_G + SS_e$$

Vz. 8-9

$$DF_{TOT} = DF_G + DF_e$$

Vz. 8-10

Patnáct rostlin (pěstovaných v samostatných květináčích) bylo rozděleno náhodně do tří skupin po pěti. Rostliny první skupiny byly pěstovány v písčité půdě, rostliny druhé skupiny v hlinité půdě a třetí skupiny v rašelině. Cílem studie bylo zjistit, zda typ substrátu ovlivní velikost rostlin. Byly změřeny výšky rostlin na vrcholu sezóny (v cm):
Písek: 15, 16, 18, 15, 21
Hlína: 21, 20, 18, 25, 26
Rašelina: 22, 26, 27, 30, 29
Z dat nejprve spočteme průměry: $\overline{X}_1 = 17$, $\overline{X}_2 = 22$, $\overline{X}_3 = 26.8$, celkový průměr je $\overline{X} = 21.9333$
Počet stupňů volnosti v každé skupině je čtyři, tedy $DF_E=4+4+4=12$, $DF_G=3-1=2$.
$SS_E=(15-17)^2 +(16-17)^2 + (18-17)^2 + ... + (21-22)^2 + ... + (30-26.8)^2 = 110.8$
$SS_G= 5x(17-21.9333)^2 + 5x(22-21.9333-5)^2 + 5x(26.8-21.9333)^2= 240.13$.
$MS_G= 240.13/2 = 120.07$, $MS_E = 110.8/12 =9.23$. $F= MS_G / MS_E = 120.07/9.23 = 13.00$.
Pravděpodobnost, že získáme z F distribuce s parametry (stupni volnosti) 2 a 12 získáme hodnotu 13.00 nebo větší je rovna 0.00099. Nulovou hypotézu o rovnosti středních hodnot proto zamítáme.

Obr. 8-1 Příklad analýzy variance jednoduchého třídění.

Případným zamítnutím nulové hypotézy ovšem zjistíme pouze, že se alespoň jedna střední hodnota liší, tj. že všechny nejsou stejné. Které se liší nám ale výsledek F-testu neříká. K tomu musíme použít další metody, nejčastěji se užívají tzv. mnohonásobná porovnání (*multiple comparisons*).

ANOVA pro k=2 a t-test

Analýzu variance můžeme použít i pro případ, kdy $k=2$, tedy pro porovnání dvou průměrů. V tomto případě je výsledek (tj.dosažená hladina významnosti) zcela shodný s výsledkem dvoustranného *t*-testu. Při použití metody ANOVA ale nelze testovat jednostrannou hypotézu a nemůžeme ani testovat nulovou hypotézu $H0: \mu_1 - \mu_2 = \mu_0$, pokud se μ_0 nerovná nule (při *t*-testu tyto možnosti máme). Při užití t-testu také můžeme užít modifikaci, která zohledňuje narušení homogenity variancí.

Dva modely analýzy variance

V analýze variance rozlišujeme dva modely: tzv. model I nebo také model s pevnými efekty (*fixed effect model*) a model II neboli model s náhodnými efekty (*random effect model*). Příkladem úloh na model I (tj. s pevnými efekty) jsou příklady 1 a 3 na začátku kapitoly. Ptáme se, zda se liší uvedené typy stravy ve svém vlivu na krevní obraz: zajímají nás právě ty typy stravy, které byly užity v pokusu. Podobně v pokusu z příkladu 3 nás zajímají právě tři uvedené odrůdy jabloní.

Naproti tomu v pokusu 2 se ptáme, zda se mohou charakteristiky druhu lišit podle lokalit (každá populace obývá jednu oddělenou lokalitu) - lokality jsou náhodným výběrem z mnoha možných různých lokalit: cílem není říci, že se liší kostřava od Budějovic od kostřavy od Třeboně, ale ukázat, že různé populace téhož druhu se mohou v určitých charakteristikách lišit. V jednocestné analýze variance je výpočetní postup stejný pro oba modely. Pokud budeme pracovat s více faktory (dvoucestná analýza variance, složitější modely), potom se výpočty budou lišit.

V posledních desítiletích se významně rozvinulo užívání modelů s náhodnými efekty, které zahrnují nejen různé varianty analýzy variance, ale rozšiřují celou skupinu obecných lineárních modelů (s vysvětlujícími proměnnými jak kategoriálními, tak kvantitativními), ale i skupinu zobecněných linearních modelů. Modelům, které zahrnují jak náhodné, tak pevné efekty se dnes obvykle říká modely se smíšenými efekty (*linear mixed-effect models*,

případně *generalized mixed-effect models*). Náš příklad analýzy variance s jedním náhodným efektem je pak nejjednodušší formou takového modelu.

Autoři metodiky modelů se smíšenými efekty ale také zvolili odlišné pojetí počtu stupňů volnosti pro náhodné efekty, z čehož vyplývá odlišný způsob testování těchto efektů. V našem příkladě 2 (odlišnost morfologických parametrů kostřavy mezi stanovišti) považujeme náhodný efekt stanoviště za parametr, který je určen jen jedním číslem, konkrétně variabilitou odchylek stanovištních průměrů od celkového průměru (předpokládáme, že tyto odchylky pocházejí z normální distribuce $N(0,\sigma^2_A)$. a proto při testu používáme jen jeden stupeň volnosti, nikoliv 4, které bychom používali při testování pevného efektu pro faktor s pěti hladinami. Tento přístup k analýze náhodných efektů není ale podporován programem Statistica a budeme jej ilustrovat v praktické sekci věnované programu R.

Síla testu

Jako ve všech předchozích případech různých testů si musíme uvědomit, že neprůkazný výsledek testu znamená buď, že se střední hodnoty neliší, nebo je výsledkem chyby druhého druhu. Zatímco pravděpodobnost chyby prvního druhu známe (je jím nominálně stanovená hladina významnosti), pravděpodobnost chyby druhého druhu neznáme. Za určitých předpokladů je ale možné tuto pravděpodobnost odhadnout. Tento odhad by měl být proveden vždy před založením rozsáhlejšího nebo náročnějšího pokusu. Postup odhadu je relativně složitý, je hezky popsán v učebnicích Zar (1984, p.171) a Sokal a Rohlf (1981, p. 262). Jako základní informaci je vhodné si zapamatovat, že síla testu roste s velikostí výběru, s rozdíly mezi skupinami, a klesá s variabilitou materiálu uvnitř skupin a s počtem skupin. Síla testu klesá s nevyvážeností počtů pozorování ve skupinách. Vezměme v úvahu pokus 1. Máme k dispozici celkem 50 králíků (více jich např. z finančních důvodů nemůžeme použít) a chtěli bychom testovat vliv více typů stravy. Pokud máme představu o možné velikosti efektu typu stravy a o variabilitě v krevním obrazu, můžeme spočítat, kolik typů stravy můžeme v jednom pokuse testovat, aby síla testu byla alespoň 95%.

Narušení předpokladů

Jak bylo uvedeno, analýza variance (ANOVA) předpokládá normalitu (uvnitř skupin!) a rovnost (homogenitu) variancí. Obojí lze testovat, pro normalitu bychom testovali normalitu reziduálů, pro homogenitu variancí se nejčastěji užívá Bartlettův test. Naštěstí je ANOVA relativně robustní vůči narušení obou předpokladů. Navíc, jak je u testování předpokladů obvyklé, „potřebujeme“ aby nám testy vyšly neprůkazně. Tak se nám stane, že při malém počtu pozorování (kdy je např. Bartletův test velmi slabý) nás uklidní neprůkazný výsledek testu, zatím ce při velkém počtu pozorování i malé narušení předpokladů, vůči kterému je ANOVA dostatečně robustní, vede k průkaznému výsledku testu (a my budeme přemýšlet, co s tím). Podstatně užitečnější než formální testování je podívat se, jak se chovají reziduály.

ANOVA nevyžaduje stejné počty pozorování ve skupinách. Nicméně, pokud se počty ve skupinách výrazně liší, ztrácí ANOVA svou robustnost vůči narušení předpokladu o rovnosti variancí. Proto když plánujeme pokus, snažíme se užívat stejně velké skupiny. Pokud jsou všechny skupiny stejně velké, říkáme, že třídění je *vyvážené* (*balanced design*). Robustnost vůči narušení normality je tím větší, čím více je pozorování v jednotlivých skupinách. Pokud je narušení normality příliš velké, je možné užít transformace dat, užít

neparametrickou obdobu analýzy variance (Kruskal-Wallisův test) nebo užít zobecněné lineární modely. Jeden z těchto přístupů by byl pravděpodobně nutný v příkladu 2; pokud by průměrné počty výběžků u kostřavy byly nízké, můžeme je pouze těžko považovat za spojitou proměnnou s normálním rozdělením (aproximace Poissonovým rozdělením v zobecněném lineárním modelu by byla vhodným řešením). Při vyšších průměrných počtech by asi narušení předpokladů bylo snesitelné.

Mnohonásobná porovnání

Průkazný výsledek analýzy variance (tj. zamítnutí nulové hypotézy) nás informuje o tom, že se alespoň jeden z porovnávaných průměrů liší od zbývajících. Rozdílů mezi průměry může být ale více. Pokud se jedná o model s faktorem s pevným efektem, tj. pokud je každá skupina jednoznačně definována a není náhodným výběrem z většího množství skupin, potom nás zajímá, které průměry se mezi sebou liší. K zodpovězení slouží mnohonásobná porovnání (*multiple comparisons*). Běžný postup je takový, že nejprve spočteme analýzu variance a pouze pokud zamítneme nulovou hypotézu o rovnosti průměrů, pokračujeme v mnohonásobných porovnáních. Protože porovnání provádíme **po** vlastní ANOVě, nazývají se *a posteriori*. Kromě těchto metod se někdy užívají tzv. plánovaná (*a priori*) porovnání. To v případě, že na základě „vnější“ informace rozhodneme již před provedením pokusu, že nás zajímají pouze vybraná srovnání. Tento postup je v některých případech užitečný (síla testu je vyšší než u *a posteriori* porovnaní), v praxi je se užívá méně často. Bližší informace uvádí Sokal a Rohlf (1981, p. 232).

Při užití mnohonásobných porovnání vlastně řešíme problém, který je při praktickém použití statistiky běžný. Výsledky většiny experimentů nám umožní odpovídat na více než jednu otázku, a tudíž budeme s velkou pravděpodobností provádět více než jeden test. Pokud ale budeme mít v každém pravděpodobnost chyby prvního druhu omezenou hladinou α, tak bude pravděpodobnost, že se dopustíme alespoň jedné chyby prvního druhu v celém experimentu (tj. něco nám vyjde průkazně, přestože příslušná nulová hypotéza platí), mnohem vyšší (jak také ukazuje tabulka 9-1). Můžeme se pak rozhodnout, zda budeme kontrolovat četnost chyby prvního druhu v jednotlivém testu (*comparison-wise Type I error rate*) nebo v rámci celého experimentu (*experiment-wise Type I error rate*). První přístup ale povede při větším množství testů k tomu, že budeme mít mnoho „falešně pozitivních“ výsledků. Obvykle se kontrola frekvence chyb prvního druhu provádí pro skupinu logicky souvisejících testů (např. porovnávání všech hladin jednoho z testovaných faktorů, řešená v této kapitole, nebo výběr vysvětlujících proměnných pro jeden regresní model, viz kapitola 14). Výsledek takovéto kontroly se pak nazývá *family-wise Type I error rate*.

Metod mnohonásobných porovnání existuje řada, což samo o sobě indikuje, že žádná z nich není zcela ideální. Zde si ukážeme jednu, která je všeobecně přijímána a která je k dispozici ve většině statistických programů: Tukey(ho) test. Z dalších (Duncanův test, SNK-test, Scheffé) se doporučuje především SNK-test (Student-Newman-Keuls test) a Scheffé-ho test. Všechny tři mají tu vlastnost, že jsou konstruovány tak, aby pravděpodobnost výskytu chyby prvního druhu v alespoň jednom z dílčích srovnání celého experimentu byla rovna stanovené hladině významnosti. Tuto vlastnost nemá Duncanův test, kde je pravděpodobnost chyby prvního druhu vztažena na jednotlivá porovnání (bere ake v úvahu pořadový rozdíl průměru porovnávaného výběru). Proto pomocí Duncanova testu získáme nejvíce průkazných rozdílů (ale pravděpodobnost chyby prvního druhu je vyšší).

Kromě mnohonásobných pozorování existuje ještě další úloha: neporovnáváme všechny dvojice mezi sebou, ale několik různých zásahů, každý pouze proti jedné společné

kontrole: k tomuto účelu slouží Dunnettův test. Mnohonásobná porovnání předpokládají, že se jedná o ANOVA model I (tj. model s pevnými efekty). Pokud porovnáváme tři vybrané odrůdy, zajímá nás, které se od sebe liší. Naproti tomu pokud chceme dokázat, že se budou lišit populace trávy z pěti náhodně vybraných lokalit (tj. model II), rozdíly mezi dvojicemi lokalit už netestujeme. Při průkazném výsledku hlavního testu se v ANOVA modelu II někdy odhadují podíly vlivů na varianci (tj. poměrný vliv variability uvnitř tříd a mezi třídami), např. poměr variability v rámci lokality a mezi lokalitami, případně se vyjadřují koeficientem vnitrotřídní korelace: 1 znamená, že uvnitř tříd jsou všechna pozorování totožná, rozdíly jsou mezi třídami, 0 - není žádný rozdíl mezi třídami (bližší vysvětlení Sokal a Rohlf, 1981, p. 215-216).

V praxi je někdy obtížné rozhodnout, zda se jedná o model I nebo II. Vraťme se k příkladu s kostřavou. Za určitých podmínek můžeme považovat čtyři stanoviště za náhodný výběr z většího množství stanovišť: pak nás nezajímají konkrétní rozdíly, ale míra variability mezi stanovišti a uvnitř stanovišť. Naproti tomu, mnohdy jsme o stanovištích schopni říci něco dalšího a zajímavou informací může být, která stanoviště se mezi sebou liší; potom má smysl považovat vliv stanoviště za vliv s pevným efektem a provést mnohonásobné porovnání.

Rozhodnutí, o který ANOVA model se jedná, může tedy za určitých okolností záviset na kontextu a na otázkách, které si klademe. Porovnáváme-li populace určitého druhu, může být pro některé čtenáře zajímavá informace, že populace ze Třeboně a z Budějovic se neliší, zatímco populace od Plzně se od nich významně liší (byť by naším primárním cílem bylo najít a kvantifikovat mezipopulační vnitrodruhovou variabilitu). Naproti tomu, uvedeme-li do článku, kde porovnáváme potomstvo osmi náhodně vybraných rostlin, o kterých nevíme nic dalšího, že potomstvo náhodně vybraných individuí číslo 1, 3 a 8 se vzájemně nelišilo, ale tato skupina se lišila od potomstva náhodně vybraných individuí 2, 4, 5, 6 a 7, má tato informace pro čtenáře velmi malou cenu: při náhodném výběru mohla individua 1, 3, a 8 být stejně tak dobře označena jako 4, 5, 7. Určitou informaci ovšem získáme, pokud uvidíme, zda se jedná o dvě dobře odlišené skupiny nebo zda je potomstvo jednotlivých individuí relativně stálé, ale průměry pokrývají více méně rovnoměrně celou šíři variability. Protože nás může svádět interpretovat v tomto smyslu překrývající a nepřekrývající se skupiny ve výsledcích mnohonásobných porovnání, upozorněme, že překrývající se skupiny jsou důsledkem chyby druhého druhu, a jako takové je třeba je interpretovat.

Tukeyho test

Tukeyho test se provádí analogicky *t*-testu. Obdobou kriteria *t* je zde kriterium *q*, které spočteme pro libovolnou dvojici porovnávaných průměrů skupin *A* a *B*

$$q = \frac{\overline{X}_A - \overline{X}_B}{SE}$$

Vz. 8-11

kde *SE* je střední chyba odhadu rozdílu průměrů skupin *A* a *B*. Ta se spočte jako

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Vz. 8-12

s^2 je odhad společné variance - reziduální (uvnitř skupin) průměrný čtverec MS_E z analýzy variance, n_A a n_B jsou počty pozorování ve srovnávaných skupinách. Připomeňme, že předpokládáme homogenitu variance a že doporučení stejné velikosti skupin zde platí naléhavěji než pro hlavní test ANOVA modelu.

SNK test se počítá zcela shodně; v tomto testu se seřadí průměry skupin podle velikosti a kritická hodnota závisí na rozdílu pořadí porovnávaných průměrů, nikoliv na celkovém počtu porovnávaných průměrů. SNKtest je silnější než Tukeyho test, ale má větší pravděpodobnost chyby prvního druhu. Říká se, že skutečná pravděpodobnost chyby prvního druhu je u SNK vyšší než hladina významnosti a u Tukeyho nižší, ale je velmi obtížné to přesně vyjádřit.

Výsledky mnohonásobných porovnání můžeme prezentovat v textové podobě nebo grafiky. V prvém případě můžeme do tabulky zobrazit seřazené průměry každý do jednoho řádku a neprůkaznost rozdílů vyznačit stejným písmenem nebo hvězdičkou či křížkem v témže sloupci, nebo můžeme vytvořit tabulku porovnávající každou hladinu se všemi ostatními, nebo můžeme spočíst a zobrazit intervaly spolehlivosti pro odhad rozdílů mezi skupinovými průměry. V případě grafického zobrazení doplníme graf s průměry (body či sloupečky histogramu) opět písmenky, která jsou shodná pro ty průměry, které se od sebe navzájem neliší. Všechny tyto možnosti jsou ilustrovány v části "Jak postupovat v programu Statistica".

Poměrně často se nám stane například při porovnáváná tří různých skupin, že se skupiny 1 a 3 se průkazně liší ale skupinu 2 nelze odlišit průkazně ani od skupiny 1, ani od skupiny 3. Není ale možné, aby se skupiny 1 a 3 průměry lišily a skupina 2 byla totožná s oběma. Došlo zde tedy s největší pravděpodobností k chybě druhého druhu. Takovýto výsledek bývá častý zvláště pokud porovnáváme více průměrů a uvedenou prezentací vyjadřujeme i nejistotu, která je dána nedostatkem dat. Občas se také stane, že mnohonásobná porovnání neukáží žádný průkazný rozdíl, i když je hlavní *F*-test ANOVA modelu průkazný. To je také důsledkem toho, že síla mnohonásobných porovnání je menší než síla *F*-testu. V tom případě můžeme očekávat, že experiment s větším množstvím dat by pravděpodobně přinesl průkazné výsledky i v mnohonásobných porovnáních.

Dunnettův test

Pokud chceme porovnat jednu kontrolu s více druhy pokusného zásahu, použijeme Dunnetův test. Jeho použití by bylo vhodné pro řešení příkladu 1, kdybychom chtěli testovat, přídavky kterých látek do stravy vliv na krevní obraz (v porovnání se standardní neobohacenou stravou), ale nezajímalo by nás, zda se případný vliv liší mezi přídavkem vápníku a železa.

Předpokládáme, že porovnáváme výběr *A* proti kontrole. Hodnotu *q* vypočteme obdobně jako v Tukey-testu (Vz. 8-11):

$$q = \frac{\overline{X}_A - \overline{X}_{kontrola}}{SE},$$

Vz. 8-13

kde *SE* je střední chyba odhadu rozdílu průměrů skupin *A* a kontroly. Ta se spočte

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_{kontrola}}\right)}$$

Vz. 8-14

Význam symbolů je analogický Vz. 8-11 a Vz. 8-12. Podobně jako v SNK-testu, i zde závisí výsledek na rozdílu pořadí mezi kontrolou a porovnávaným vzorkem (obdoba *p* v SNK-testu). Pokud víme, kterým směrem by zásahy měly vychýlit výsledek vůči kontrole, je výhodné použít jednostranný test: pokud například přidávám do stravy prvky a měřím krevní obraz jako odpověď, pak je to zřejmě proto, že chci, aby tyto prvky krevní obraz zlepšily. Protože provádíme méně porovnání, je síla Dunnetova testu vyšší než u mnohonásobných porovnání, kde provádíme test pro každou dvojici. Protože kontrola vstupuje do testu vícekrát, je potřeba, aby byla nejpřesněji odhadnuta - doporučuje se proto, aby počet pozorování v kontrole byl o něco méně než $\sqrt{k-1}$ krát větší a ostatní skupiny se nelišily ve své velikosti. *k* je počet všech skupin včetně kontroly.

Neparametrická analýza variance

Data, která jsme testovali jednocestnou analýzou variance (tzn. data ze zcela náhodného experimentálního uspořádání - completely randomized experimental design) můžeme testovat také pomocí neparametrického testu, založeného na pořadí. Jedná se o Kruskal-Wallisův test (někdy také nazývaný analysis of variance by ranks). Podobně, jako je Mann-Whitneyův test neparametrickou obdobou dvouvýběrového t-testu, je Kruskal-Wallisův test obdobou modelu ANOVA jednoduchého třídění; pokud porovnáváme dvě skupiny, doporučuje se užít Mann-Whitneyův test. Zatímco ANOVA pro dvě skupiny a dvouvýběrový t-test dávají totožné výsledky, výsledky pro Kruskal-Wallisův test a Mann-Whitneyův test jsou různé.

Pokud jsou splněny podmínky pro ANOVA model (normalita a homogenita variance), je síla Kruskal-Wallisova testu asi 95% parametrického ANOVA modelu. Tento test lze ale užít i v mnoha případech, kdy podmínky pro parametrický test splněny nejsou. Je také možné jej použít v případě hodnocení dat na ordinální stupnici V testu postupujeme následujícím způsobem. Nejprve přiřadíme každému pozorování jeho pořadí mezi všemi pozorováními bez ohledu na zařazení do skupiny. Potom spočteme statistiku

$$H = (N-1)\frac{\sum_{i=1}^{k} n_i (r_i - r)^2}{\sum_{i=1}^{k}\sum_{j=}^{n_i} (r_{ij} - r)^2}$$

Vz. 8-15

kde n_i je počet pozorování ve skupině *i,* N=$\sum_{i=1}^{k} n_i$, tedy celkový počet pozorování ve všech skupinách, *k* je počet skupin, r_{ij} je (celkové) pořadí *j*-tého pozorování z *i*-té skupiny, r_i je průměrná hodnota pořadí pro pozorování ve skupině *i*, a *r* je průměrné pořadí v celém souboru dat (to je rovno 0.5.(N+1)). Statistika H má za platnosti nulové hypotézy přibližně χ^2-rozdělení s počtem stupňů volnosti *k-1*. Postup při výpočtu ukazuje následující příklad v Obr. 8-2. I u Kruskall-Wallisova testu můžeme po průkazném výsledku použít proceduru mnohonásobného porovnání, viz Zar (1984, p.199).

V dotazníkové akci, zjišťující vztah různých sociálních skupin k ochraně přírody, odpovídali respondenti na otázku: Jak velkým problémem je pro vás vymírání druhů na planetě Zemi? Odpovědi byly na ordinální stupnici: 0 – žádný problém v tom nevidím, 1 – drobný problém, bez většího významu pro lidstvo, až po 10 – zásadní problém, který ohrožuje lidskou existenci. Respondenti byli náhodně vybráni z kontrastních typů sídel: velká průmyslová města a průmyslové aglomerace; malá města, bez velkého průmyslového podniku; sídla do 10 000 obyvatel. Liší se přístup obyvatel k problému v závislosti na tom, kde bydlí? Při interpretaci výsledků si ale musíme uvědomit, že rozdíly mezi skupinami nelze jednoznačně interpretovat jako závislost názoru na ochranu přírody na tom, kde dotyčná osoba bydlí. Kauzalita může být zcela opačná – lidé, kteří mají zájem o přírodu, se možná tolik nestěhují do průmyslových aglomerací.

H_0: Názory na vymírání se neliší mezi třemi srovnávanými skupinami.
H_A: Názory na vymírání se mezi skupinami liší.
Data jsou následující (s pořadím - ranky - hodnot v závorkách):
Průmyslová města: 1 (2.5), 0 (1.0), 2 (4.0), 1 (2.5) a 4 (5.5)
Malá města: 5 (8.5), 7 (13.5) 9 (16.5), 5 (8.5), 6 (11.5) a 4 (5.5)
Vesnice: 7 (13.5), 9 (16.5), 6 (11.5), 8 (15.0), 5 (8.5) a 5 (8.5)
$N = 5 + 6 + 6 = 17$
$H = 16.(254.09/400.5) = 10.15098$, $p=0.00625$ (při srovnání s χ^2_2). Zamítáme tedy H_0.

Obr. 8-2 Kruskal-Wallisův test (analýza variance pomocí pořadí).

Příkladová data

Pro ilustraci klasického ANOVA modelu jednoduchého třídění použijeme data o výškách rostlin pěstovaných ve třech typech substrátu, představená v Obr. 8-1. V souboru příkladových dat jsou na listu *Chap8*: proměnná *Height* udává výšku rostliny v centimetrech, proměnná *Substrate* identifikuje substrát, ve kterém byla daná rostlina pěstována.

Pro ilustraci testování náhodného efektu v programu R použijeme následující data (opět v listu *Chap8*), odpovídající zčásti příkladu 2 na začátku této kapitoly. Byla studována geneticky podmíněná variabilita ve velikosti rostlin kostřavy mezi jednotlivými populacemi: semena z dané populace byla pěstována ve srovnatelných podmínkách a po 2 měsících byla změřena délka nejdelšího výběžku (tilleru) v centimetrech. Délky pro jednotlivé jedince jsou v proměnné *Till.Len*, příslušnost k jedné z pěti vzorkovaných populací je uvedena v proměnné *Population*. Pokud bychom se drželi kontextu klasické ANOVy, můžeme tato data vyhodnotit i v programu Statistica – výsledek pak bude stejný, považujeme-li lokalitu za faktor s pevným nebo náhodným efektem.

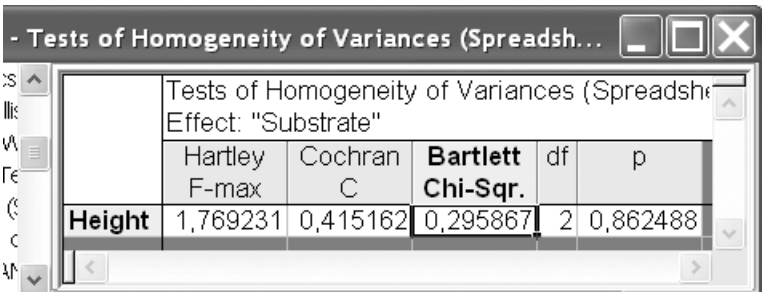
Kruskal-Wallisův test je ilustrován příkladem z Obr. 8-2. Typ sídla je popsán faktorem *Settlement*, názor respondenta proměnnou *Importance*.

Jak postupovat v programu Statistica

Jednocestná ANOVA

Zvolíme z menu příkaz *Statistics* | *ANOVA* a v zobrazeném seznamu vybereme položku *One-way ANOVA*. Pomocí tlačítka *Variables* zvolíme proměnnou *Height* v levém seznamu (*Dependent variable list*) a proměnnou *Substrate* v pravém (*Categorical predictor (factor)*). Pokud bychom chtěli porovnat jen některé z hladin vybraného faktoru (například jen rostliny pěstované v písku a v rašelině), můžeme tyto hladiny zvolit za pomocí tlačítka *Factor codes*, jinak můžeme nechat implicitní stav, který použije všechny. Po volbě tlačítka *OK* v dialogovém okně *ANOVA/MANOVA One-Way ANOVA* se objeví nové okno s titulkem *ANOVA Results*.

Začneme ověřením předpokladu homogenity variancí. K tomu musíme nabídku v tomto okně rozšířit, pomocí tlačítka *More results* v levém dolní rohu okna. Po jeho volbě vybereme v rozšířeném okně záložku *Assumptions* a zvolíme tlačítko *Cochran C, Hartley, Bartlett*.


- Tests of Homogeneity of Variances (Spreadsh...

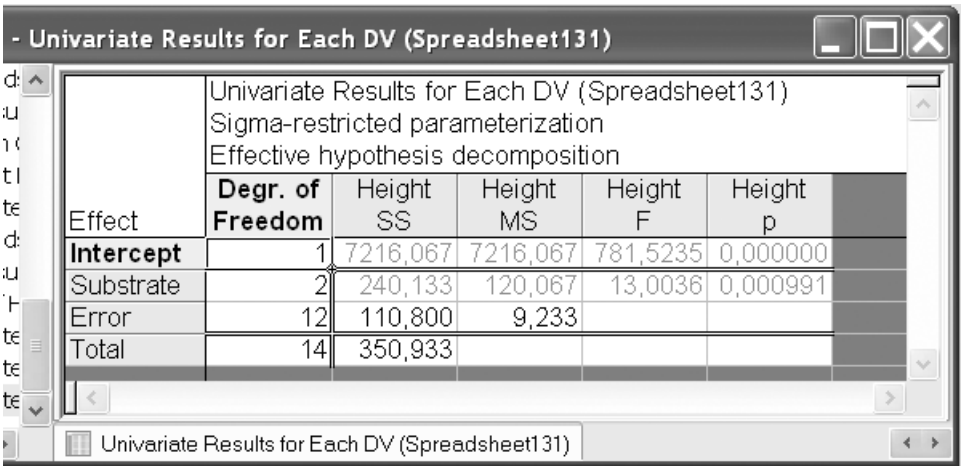
Tests of Homogeneity of Variances (Spreadsh
Effect: "Substrate"

	Hartley F-max	Cochran C	**Bartlett Chi-Sqr.**	df	p
Height	1,769231	0,415162	0,295867	2	0,862488

Výsledky Bartlettova testu jsou v posledních třech sloupcích, hodnotu testové statistiky srovnáváme s χ^2 distribucí, v našem příkladě se třemi skupina používáme dva stupně volnosti. Pravděpodobnost chyby 1. druhu je příliš vysoká na to, abychom zamítli nulovou hypotézu o shodě variancí, v tomto kontextu je to ale uspokojivý výsledek, protože znamená, že předpoklad homogenity variancí nebyl pro naše data zpochybněn.

Na záložce *Assumptions* je užitečné také vizuálně zkontrolovat, jak vypadají reziduály (vynést jejich histogram četností). Protože nejčastějším případem narušení homogenity variance je situace, kdy je směrodatná odchylka pozitivně závislá na průměru skupiny, je dobré se také podívat na graf závislost (tlačítko *Plot means vs. Std. Deviations*).

Nyní můžeme provést základní F-test z dialogového okna *ANOVA Results*. Na záložce *Summary* vybereme tlačítko *Univariate results* (podobný výsledek ale dává i *Test all effects*) a zobrazí se nám následující tabulka:


- Univariate Results for Each DV (Spreadsheet131)

Univariate Results for Each DV (Spreadsheet131)
Sigma-restricted parameterization
Effective hypothesis decomposition

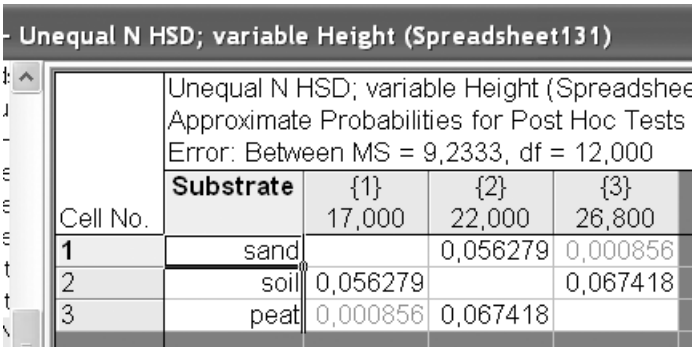
Effect	**Degr. of Freedom**	Height SS	Height MS	Height F	Height p
Intercept	1	7216,067	7216,067	781,5235	0,000000
Substrate	2	240,133	120,067	13,0036	0,000991
Error	12	110,800	9,233		
Total	14	350,933			

Univariate Results for Each DV (Spreadsheet131)

Řádek označený *Intercept* testuje odlišnost celkového průměru výšky rostlin od nuly, což je vpravdě nesmyslný test pro tento typ dat a jiné statistické programy tento test nezobrazují (pokud tedy tuto ANOVA tabulku někde prezentujete, tak tento řádek vymažte). Další tři řádky ukazují ve sloupci *Height SS* rozklad celkové sumy čtverců (SS_{TOT}, v řádku *Total*) na meziskupinovou sumu čtverců (SS_G, v řádku *Substrate*) a vnitroskupinovou (reziduální) sumu čtverců (SS_E, v řádku *Error*). Stupně volnosti jsou ve sloupci *Degr. Of Freedom* a podíl sum čtverců a stupňů volnosti (*mean squares*) jsou ve sloupci *Height MS*. Testová statistika je pak ve sloupci *Height F* a odpovídající průkaznost ve sloupci *Height p*. Nulovou hypotézu o shodě průměrů tedy zamítáme.

Mnohonásobná porovnání

Abychom zjistili, které typy substrátů se od sebe skutečně liší, provedeme mnohonásobné porovnání. V dialogovém okně *ANOVA Results* vybereme záložku *Post-hoc* a zvolíme tlačítko *Unequal N HSD*. Statistica zobrazí následující tabulku.



Unequal N HSD; variable Height (Spreadsheet131)

Unequal N HSD; variable Height (Spreadshee
Approximate Probabilities for Post Hoc Tests
Error: Between MS = 9,2333, df = 12,000

Cell No.	Substrate	{1} 17,000	{2} 22,000	{3} 26,800
1	sand		0,056279	0,000856
2	soil	0,056279		0,067418
3	peat	0,000856	0,067418	

Jde o symetrickou matici porovnání průměrné výšky každé skupiny definované typem substrátu se zbývajícími dvěma, zobrazeny jsou průkaznosti Tukeyho testu, v červené barvě pokud jsou menší než 0.05. Vidíme, že se objevila situace popsaná v posledním odstavci sekce *Tukeyho test* v této kapitole, tj. výsledky nám tvrdí, že se průkazně liší výška mezi pískem a rašelinou, ale ani jedna z těchto dvou skupin se průkazně neliší od hlíny.

Tlačítko *Tukey HSD* dává stejné výsledky porovnání v případě stejného počtu pozorování ve skupinách, ale *Unequal N HSD* je obecnější procedura použitelná i pro nevyvážené uspořádání, proto doporučujeme její používání. Další tlačítka v záložce *Post-hoc* umožňují provést i jiné, dříve diskutované postupy mnohonásobných porovnání (SNK test se skrývá pod tlačítkem *Newman-Keuls*), níže si ukážeme Dunnettův test.

Pokud v oblasti nazvané *Display* na záložce *Post-hoc* zvolíme místo *Significant differences* volbu *Homogeneous groups* (můžeme pak upravit hladinu α v políčku vpravo od této položky) a opět zvolíme tlačítko *Unequal N HSD*, Statistica zobrazí výsledky tímto způsobem:

Unequal N HSD; variable Height (Spreadsheet131)
Homogenous Groups, alpha = ,05000
Error: Between MS = 9,2333, df = 12,000

Cell No.	Substrate	Height Mean	1	2
1	sand	17,00000	****	
2	soil	22,00000	****	****
3	peat	26,80000		****

Sloupce 1 a 2 indikují hvězdičkami skupiny (typy substrátu), které se svými průměry od sebe průkazně neliší. Sloupce 1 a 2 lze také převést na písmenka (např. *a* a *b*), která pak lze prezentovat v grafu (viz obrázek níže). Nekonzistence v případě hlíny (*soil*) je odražena v tom, že tato hladina má hvězdičky v obou sloupcích. V grafu proto musí u této skupiny být jak písmenko *a*, tak písmenko *b*.

Základní grafické porovnání mezi hladinami testovaného faktoru získáme ze záložky *Summary* pomocí tlačítka *All effects/Graphs*. Po jeho volbě se objeví dialogové okno *Table of All Effects*, kde ponecháme zvolený faktor (v jednocestném ANOVA modelu jediný) a také další volby a zvolíme tlačítko *OK*. Statistica vytvoří následující graf (text nad grafem byl odstraněn):

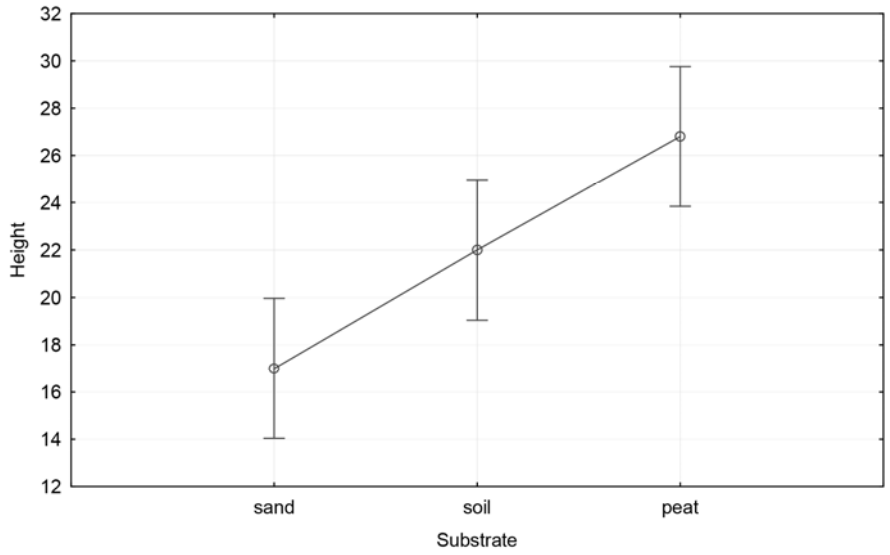
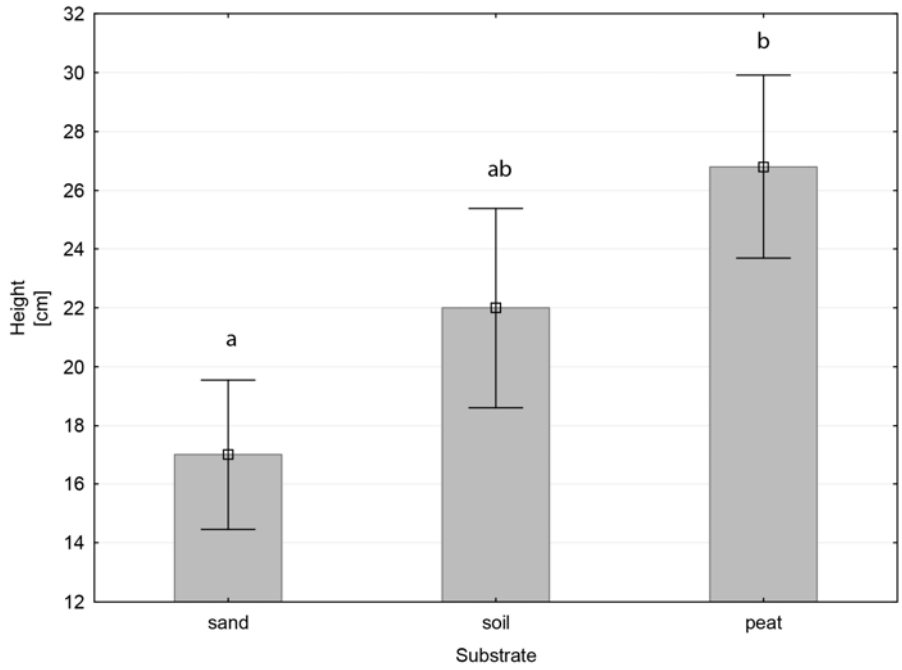


Diagram pěkně ukazuje odlišnost rostlin z písku a rašeliny, jejichž 95% konfidenční intervaly se nepřekrývají a také to, že pro daný rozsah dat nejsme schopni odlišit rostliny pěstované v písku (či v rašelině) od rostlin pěstovaných v hlíně.

Čáry spojující průměry jednotlivých skupin doporučujeme odstranit. Ty jsou užitečné v případě ANOVA modelu s více faktory pro pochopení jejich interakce, ale ve výše uvedeném grafu navozují nesprávný dojem spojitosti proměnné *Substrate*. Odstranění lze provést rychlým dvojitým poklepnutím na již vybranou čáru a zrušením zaškrtnutí položky *Line* v sekci *Plot* | *General* zobrazeného dialogového okna *Graph Options*. Alternativní způsob zobrazení dat i výsledků představuje následující graf.



Graf byl vytvořen volbou příkazu *Graphs* | *Means w/Error Plots*. V záložce *Quick* bylo pro *Graph type* zvoleno *Columns* a v oblasti *Whisker* zvoleno *Std dev* s hodnotou *Coefficient* 1. Sloupečky tedy zobrazují průměrné výšky ve skupinách a rozsahy zobrazují +- jednu směrodatnou odchylku (v popisce grafu musíte tuto informaci uvést – výběrem směrodatné odchylky informujete čtenáře o variabilitě dat). Graf byl následně modifikován změnou barev, odstraněním čáry spojující průměry. Písmena informující o výsledcích

mnohonásobných porovnání na hladině 0.05 lze do grafu přidat pomocí příkazu *Insert* a následné volby *text.* Pak zadáme vhodná písmena a text přesuneme na požadované místo.

Dunnettův test obvykle používáme při zodpovídání odlišných badatelských otázek, než které vedly k experimentu s typy substrátu, ale můžeme si alespoň představit situaci, kdy bychom pěstovali nějaký zemědělsky významný druh plodiny v hlíně a uvažovali bychom o změně substrátu z hlíny buď na rašelinu nebo na písek, tak abychom zvýšili produkci (zde odhadovanou pomocí výšky). Můžeme se pak ptát, zda rostliny pěstované v rašelině nebo v písku jsou průkazně vyšší než ty pěstovaných dosavadním způsobem, a použít Dunnettův test, srovnávající jednu referenční hladinu (zde hlínu) se zbylými dvěma hladinami. Na záložce *Post-hoc* v dialogovém okně *ANOVA Results* zvolíme pro *Display* hodnotu *Significant differences* a v dolní části okna (*Comparisons with a Control Group*) zvolíme *> CG* (testujeme tedy jednostrannou H0, že výška v dané skupině není větší než ve skupině kontrolní) a změníme hodnotu *CG cell #* z 1 na 2, tak abychom testovali proti hlíně. Pak zvolíme tlačítko *Dunnett*.

Dunnett test; variable Height (Spreadsheet131
Probabilities for Post Hoc Tests (M>Control)
Error: Between MS = 9,2333, df = 12,000

Cell No.	Substrate	{2} 22,000
1	sand	0,997765
2	soil	
3	peat	0,025245

Vzhledem k naší volbě jednostranného testu a k tomu, že průměrná výška rostlin pěstovaných na písku je menší než těch z hlíny, je hodnota *p* v řádku *sand* očekávatelná, je ale zajímavé vidět, jak to, že naše hypotéza je více konkrétní a provádíme menší počet porovnání, ovlivnilo průkaznost srovnání mezi hlínou a rašelinou (p=0.025) – připomeňme, že v Tukeyho testu toto srovnání bylo neprůkazné..

Síla testu

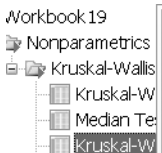
Pokud bychom chtěli zjistit, kolik rostlin bychom potřebovali v každé skupině, abychom byli schopni odlišit všechny tři substráty od sebe v ANOVA modelu jednoduchého třídění, můžeme postupovat takto: nejprve musíme odhadnout průměrné hodnoty v jednotlivých skupinách (viz například diagramy vytvořené výše) a také odhadnout variabilitu uvnitř skupin pomocí směrodatné odchylky (viz kapitola 1 – procedura *Descriptive statistics* s použitím tlačítka *By Group*): ta je kolem 3 (od 2.5 pro písek po 3.4 pro hlínu, ale variance se mezi skupinami průkazně neliší – viz Bartlettův test výše). To odpovídá tomu, jak by měl podobný výzkum probíhat. Nejprve provedeme tzv. pilotní pokus (za něj můžeme považovat pokus právě vyhodnocený) s relativně malým množství opakování, a na jeho základě pak naplánujeme velký pokus.

Zvolíme *Statistics* | *Power analysis* a pak vybereme *Sample Size Calculation* v levém seznamu a *Several Means, ANOVA, 1-Way* v pravém seznamu. V dalším dialogovém okně můžeme upravit nabízené volby, například nastavit hodnotu *Alpha* na 0.01 pro chybu 1. druhu, a hodnotu 0.95 pro zamýšlenou sílu testu (to je 1 – pravděpodobnost chyby 2. druhu). Hodnotu RMSSE odhadneme pomocí tlačítka *Calculate Effects*. Po jeho zvolení zadáme v novém okně v levé dolní části skupinové průměry (17, 22 a 26.8) a do políčka *Sigma* její odhad 3.0. Po zavření tohoto okna pomocí *OK* vybereme tlačítko *OK* i v okně *1-Way ANOVA: Sample Size Parameters* a v novém okně zvolímě tlačítko *Calculate N*. Objeví se výsledky, mezi kterými je asi nejzajímavější poslední řádek (*Required Sample Size*

(N)), který nám říká, že k dosažení zvolených parametrů testu bychom potřebovali alespoň 6 rostlin v každé skupině. Stejné dialogové okno nabízí i možnost vynesení diagramů, které ukazuje, kolik opakování je třeba v každé skupině pro dosažení určitého α nebo určité síly testu.

Kruskal-Wallisův test

Zvolíme z menu příkaz *Statistics* | *Nonparametrics* a v seznamu vybereme položku *Comparing multiple indep. Samples (groups)*. Pomocí tlačítka *Variables* zvolíme proměnnou *Importance* v levém seznamu a proměnnou *Settlement* v pravém seznamu a pak zvolíme tlačítko *Summary*. **Pozor! Statistica provede po Kruskall-Wallisově testu ještě mediánový test a to je také ten, který vidíme nejprve po zobrazení výsledků. Abychom se podívali na výsledky Kruskall-Wallisova testu, musíme v pravé části okna s výsledky (Workbook) zvolit předposlední zobrazenou položku!**



Kruskal-Wallis ANOVA by Ranks; Importance (Spreadsl
Independent (grouping) variable: Settlement
Kruskal-Wallis test: H (2, N= 17) =10,15098 p =,0062

Depend.: Importance	Code	Valid N	Sum of Ranks	Mean Rank
industrial	101	5	15,50000	3,10000
town	102	6	64,00000	10,66667
village	103	6	73,50000	12,25000

Po zjištění průkazného rozdílu mezi skupinami (jako v našem příkladě p=0.0062) můžeme dále zjistit, které páry skupiny se liší. V dialogovém okně *Kruskal-Wallis ANOVA and Median Test* zvolíme tlačítko *Multiple comparisons of mean ranks for all groups*.

Multiple Comparisons p values (2-tailed); Imp
Independent (grouping) variable: Settlement
Kruskal-Wallis test: H (2, N= 17) =10,15098

Depend.: Importance	**industrial R:3,1000**	town R:10,667	village R:12,250
industrial		0,040020	0,008305
town	0,040020		1,000000
village	0,008305	1,000000	

Pro naše příkladová data se od zbylých dvou odlišuje jen skupina *industrial*, výrazněji (p=0.0083) od *village*, ale stále průkazně i od *town* (p=0.0400).

Jak postupovat v programu R

Vzhledem k odlišnému počtu pozorování byly poslední dvě proměnné z listu *Chap8* importovány do samostatného datového rámce (*chap8b*), první čtyři proměnné jsou v datovém rámci *chap8a*.

Jednocestná ANOVA

Ověření homogenity variancí provedeme samostatnou funkcí *bartlett.test*:

```
> bartlett.test(Height~Substrate,data=chap8a)
        Bartlett test of homogeneity of variances
data:  Height by Substrate
Bartlett's K-squared = 0.2959, df = 2, p-value = 0.8625
```

Většinu variant klasického ANOVA modelu (včetně těch probíraných v pozdějších kapitolách) lze zpracovat pomocí funkce *aov*. Objekt vrácený touto funkcí lze pak předávat

dalším funkcím pro získání klasické ANOVA tabulky nebo pro provedení mnohonásobných porovnání.

```
> aov.1 <- aov( Height~Substrate, data=chap8a)
> summary( aov.1)
            Df Sum Sq Mean Sq F value   Pr(>F)
Substrate    2  240.1  120.07      13 0.000991 ***
Residuals   12  110.8    9.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1°
```

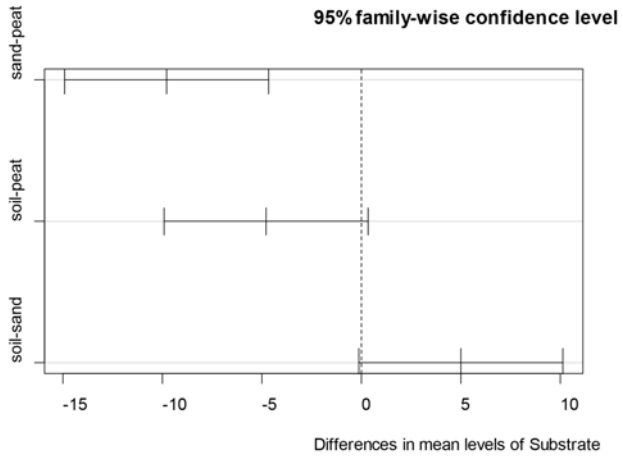
Mnohonásobná porovnání

Tukeyho test můžeme provést jednodušše pomocí funkce *TukeyHSD*:

```
> TukeyHSD(aov.1)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = Height ~ Substrate, data = chap8a)
$Substrate
          diff         lwr        upr     p adj
sand-peat -9.8 -14.9271129 -4.6728871 0.0007081
soil-peat -4.8  -9.9271129  0.3271129 0.0672929
soil-sand  5.0  -0.1271129 10.1271129 0.0561479
```

Výsledky jsou zobrazeny ve formě rozdílů mezi průměry jednotlivých skupin, jejich konfidenčních intervalů (pokrytí intervalu lze změnit pomocí parametru *conf.level* s implicitní hodnotou 0.95) a odhadu *p* pro test nulové hypotézy, že se daný rozdíl rovná nule. Grafické zobrazení téhož získáme tak, že hodnotu vrácenou touto funkcí předáme funkci *plot*:

```
> plot(TukeyHSD(aov.1))
```



Další typy mnohonásobných porovnání lze provádět pomocí specializované knihovny *multcomp*. Nejprve příklad provedení stejného testu jaký provádí funkce *TukeyHSD* (v této knihovně jej ale lze provést i s jinými typy regresních modelů než jen s ANOVA modelem):

```
> summary(glht(aov.1,linfct=mcp(Substrate="Tukey")))
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = Height ~ Substrate, data = chap8a)
Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
sand - peat == 0   -9.800      1.922  -5.099   <0.001 ***
soil - peat == 0   -4.800      1.922  -2.498   0.0673 .
soil - sand == 0    5.000      1.922   2.602   0.0561 .
```

Dunnetův test lze provést v rámci knihovny *multcomp* různými postupy, zde si ukážeme poněkud složitější, ale současně ilustrující výpočet tzv. plánovaných porovnání. Ta

popíšeme pomocí jednoduché matice kontrastů. Nejprve ale musíme zjistit, v jakém pořadí jsou jednotlivé typy substrátu seřazeny ve faktoru *Substrate*:

```
> levels(chap8a$Substrate)
[1] "peat" "sand" "soil"
```

Pro porovnání rašeliny s půdou a písku s půdou tedy musíme odečítat třetí hladinu faktoru od první a pak třetí hladinu od druhé. Tomu odpovídá následující matice (funkce *rbind* spojuje zadané vektory jako řádky matice, jejich jednotlivé hodnoty vytvoří sloupce):

```
> CompSoil <- rbind("sand vs. soil" = c(  0, +1, -1),
+                   "peat vs. soil" = c( +1,  0, -1))
> CompSoil
              [,1] [,2] [,3]
sand vs. soil    0    1   -1
peat vs. soil    1    0   -1
```

Plánované kontrasty pak spočteme takto:

```
> summary(glht(aov.1,linfct=mcp(Substrate=CompSoil)))
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: User-defined Contrasts
Fit: aov(formula = Height ~ Substrate, data = chap8a)
Linear Hypotheses:
                   Estimate Std. Error t value Pr(>|t|)
sand vs. soil == 0   -5.000      1.922  -2.602   0.0418 *
peat vs. soil == 0    4.800      1.922   2.498   0.0504 .
```

Pokud chceme reprodukovat výpočet Dunnetových kontrastů ilustrovaný výše pro program Statistica, musíme nejprve zjistit, zda rozdíl má správné znaménko a jen v tom případě považovat výsledek za potenciálně průkazný a vydělit získanou pravděpodobnost dvěma. Rozdíl mezi pískem a hlínou tedy není průkazný (rozdíl je záporný, viz sloupec *Estimate*) a pro rozdíl mezi rašelinou a hlínou je p=0.0504/2 = 0.0252.

Test náhodného efektu lokality

Přestože zde nebudeme testovat pomocí klasického ANOVA modelu, je i v případě práce s modely s náhodnými efekty vhodné nejprve otestovat homogenitu variancí ve srovnávaných skupinách.

```
> bartlett.test(Till.Len~Population,data=chap8a)
        Bartlett test of homogeneity of variances
data:  Till.Len by Population
Bartlett's K-squared = 12.0723, df = 4, p-value = 0.01682
```

Vidíme, že populace se ve variabilitě délek odnoží dosti liší mezi populacemi. Pro údaje typu rozměrů, koncentrací či vah se často užívá logaritmická transformace, která obvykle nejen variabilitu stabilizuje mezi skupinami, ale také přiblíží rozdělení hodnot ve skupinách normálnímu:

```
> bartlett.test(log(Till.Len)~Population,data=chap8a)
        Bartlett test of homogeneity of variances
data:  log(Till.Len) by Population
Bartlett's K-squared = 5.9544, df = 4, p-value = 0.2026
```

Budeme tedy používat logaritmované délky odnoží. Test náhodného efektu populace provedeme pomocí knihovny *nlme*, která umožňuje odhadovat (fitovat) jak lineární tak nelineární modely se smíšeným efekty. Nejprve ale nafitujeme model bez náhodného efektu populace a ten pak porovnáme s modelem, do kterého jsme přidaly tento náhodný efekt:

```
> library(nlme)
> lm.0 <- lm(log(Till.Len)~1,data=chap8a)
> lme.1 <- lme(log(Till.Len)~1,random=~1|Population,data=chap8a)
> anova(lme.1,lm.0)
```

```
      Model df      AIC      BIC     logLik   Test  L.Ratio p-value
lme.1     1  3 25.09871 27.01588  -9.549356
lm.0      2  2 27.55328 28.83139 -11.776639 1 vs 2 4.454568  0.0348
```

Výsledek ve sloupci *p-value* ukazuje (spolu se změnami hodnot statistik parsimonie – AIC a BIC), že přítomnost náhodného efektu populace má v modelu své oprávnění. Testová statistika *L.Ratio* a její srovnání s χ^2 distribucí s jedním stupněm volnosti představují tzv. test poměru věrohodností (*likelihood ratio test*). Pro doplnění ještě uvádíme část výstupu z funkce *summary*, ze které lze vyčíst velikost náhodného efektu:

```
> summary(lme.1)
Linear mixed-effects model fit by REML
...
Random effects:
 Formula: ~1 | Population
        (Intercept)  Residual
StdDev:   0.4066099 0.3432269
```

Tato informace ukazuje, že variabilita mezi populacemi (vyjádřená směrodatnou odchylkou populačních průměrů) je o trochu větší než variabilita uvnitř populací, i když se tyto dvě hodnoty (představující jednu z jednodušších verzí tzv. *variance components*) asi vzájemně neliší, jak je vidět, když si spočteme přibližné odhady spolehlivosti: (0.171, 0.965) pro mezipopulační variabilitu a (0.221, 0.532) pro vnitropopulační variabilitu.

```
> intervals(lme.1)
Approximate 95% confidence intervals
 Fixed effects:
               lower     est.    upper
(Intercept) 2.136281 2.587004 3.037727
...
 Random Effects:
  Level: Population
                    lower      est.     upper
sd((Intercept)) 0.1713998 0.4066099 0.9645963

 Within-group standard error:
    lower      est.     upper
0.2214352 0.3432269 0.5320052
```

Kruskal-Wallisův test

Kruskal-Wallisův test spočteme takto:

```
> kruskal.test(Importance~Settlement,data=chap8b)
        Kruskal-Wallis rank sum test
data:  Importance by Settlement
Kruskal-Wallis chi-squared = 10.151, df = 2, p-value = 0.006248
```

Mnohonásobné porovnání neparametrickou metodou je k dispozici v knihovně *pgirmess*:

```
> kruskalmc(Importance~Settlement,data=chap8b,p=0.05)
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                   obs.dif critical.dif difference
industrial-town    7.566667     7.320256       TRUE
industrial-village 9.150000     7.320256       TRUE
town-village       1.583333     6.979591      FALSE
```

Hodnoty *TRUE* a *FALSE* v posledním sloupci tabulky udávají, které páry skupin se průkazně liší na hladině nižší než zadané *p*.

Popis analýz v článku

Methods

Differences in plant height among the three compared substrate types were tested using one-way ANOVA with post-hoc comparisons using Tukey HSD method. Homogeneity of variances was tested using Bartlett test.

*Výsledky Bartlettova testu nebývají obvykle popisovány ve výsledcích, ale pokud se v některém z nich ukáže průkazný rozdíl mezi skupinami a např. logaritmická transformace tento rozdíl odstraní, v metodice může být napsáno: Tiller length values were log-transformed to achieve homogeneity of variances required by the F-test in one-way ANOVA. Když ani zvolená transformace úplně nepomůže (může se to stát například pokud máme hodně pozorování), můžeme v popisu alespoň použít „*to improve*" místo „*to achieve*".*

We have tested for differences in the attitude towards species extinction among the three sampled social groups using Kruskal-Wallis test with following non-parametric post-hoc comparisons.

Based on the pilot study results (group means and data variation estimates), we have performed power analysis of the single factor design to determine the required number of replicates needed in each group to achieve the target test power $\beta=0.95$ with $\alpha<=0.01$.

The random effect of population was tested using a likelihood-ratio test in the *nlme* package of R software [R citation here].

Results

There were significant differences in plant height among substrate types ($F_{2,12}=13.0$, $p=0.001$) and follow-up tests have demonstrated significant difference between the plants grown in sand and in peat ($p=0.00086$). The differences of plants grown in soil from the plants in other substrates were nearly significant ($p=0.056$ for sand-soil difference and $p=0.067$ for peat-soil difference) and the lack of stronger evidence is likely due to a limited size of our sample. The average values together with the group standard deviations and indications of significant differences from one-way ANOVA are shown in Figure XX (viz sloupečkový diagram výše).

There were significant differences in the attitude towards species extinction among the respondents from different types of living places ($H=10.15$, $p=0.0062$). The post-hoc tests revealed that this is due to a strong difference between the respondents from industrial centers and either small towns ($p=0.040$) or villages ($p=0.008$). There was, however, no difference between the respondents from small towns and those from villages.

We have found a significant variation in tiller length among the five sampled populations ($\chi^2_1=4.45$, $p=0.035$).

Doporučená četba

Sokal a Rohlf (1981, p. 208 - 270 ANOVA a 429-444 neparametrická ANOVA), Zar (2007), p. 162-205, Quinn & Keough (2002), pp. 173-207.

9 Dvoucestná analýza variance

Metoda dvoucestné analýzy variance (*two-way ANOVA*) je někdy též nazývána dvoufaktorová ANOVA (*two-factor ANOVA*), nebo ANOVA dvojného třídění.

Pokud testujeme současně více faktorů než jeden, užíváme obvykle tzv. faktoriální uspořádání pokusu (*factorial design*). Znamená to, že (pokud možno se stejným počtem opakování) testujeme všechny možné kombinace hladin všech faktorů. Například můžeme sledovat vliv dvou faktorů: obohacení dusíkem a zálivky na růst rostliny. U každého faktoru budou dvě hladiny: normální a zvýšená. Budeme mít tedy čtyři skupiny (bez obohacení, obohacenou dusíkem, bez obohacení dusíkem s vyšší zálivkou, a obohacenou dusíkem s vyšší zálivkou), v každé kombinaci např. pět opakování.

Stejně tak by bylo možné uspořádat faktoriálním způsobem i pokus 1 z kapitoly 8 (králíci krmení normální stravou, stravou obohacenou vápníkem a stravou obohacenou železem) tím, že bychom přidali čtvrtou skupinu, kde by králíci měli stravu obohacenou jak železem tak vápníkem. Faktoriálně můžeme uspořádat i pokusy s více než dvěma faktory a každý faktor může mít dvě i více hladin. Faktoriální uspořádání je považováno za jedno z nejefektivnějších - šetří čas i peníze, vypovídá i o interakci studovaných faktorů. V přírodních vědách používáme experimenty se třemi či čtyřmi factory méně často, s více než čtyřmi velmi zřídka: pro čtyři faktory, každý jen se dvěma hladinami, potřebujeme 16 pokusných skupin - 2^4, pro 5 faktorů 32 pokusných skupin.

Dva faktory (nebo i více) mohou být uspořádány i jiným způsobem než faktoriálním: mohou být uspořádány hierarchicky. Například v pokusu č. 2 z kapitoly 8 (rozdíly mezi kostřav z různých populací) bychom sebrali na každé lokalitě semena z pěti rostlin. Poté bychom sledovali nejen rozdíly mezi lokalitami, ale i rozdíly mezi rostlinami pocházejícími z různých rodičů téže lokality. Pro zpracování bychom užili hierarchickou analýzu variance (*hierarchical ANOVA*, někdy též *nested design* nebo *nested ANOVA*). Ta bude probírána v kapitole 11.

V dvoucestné analýze variance je výrazný rozdíl mezi vyváženým a nevyváženým modelem. Vyvážený model je takový, kde je pro každou kombinaci faktorů stejný počet opakování. Nejjednodušší výpočty jsou pro vyvážený model, kterému odpovídá největší síla testu při daném celkovém počtu pozorování. Pokud z nějakých důvodů nemůžeme vyvážený model dodržet, je výhodné mít uspořádání proporční. V tom případě je počet pozorování v 'buňce' (*cell*) *i,j,* tj. při *i*-té hladině prvního faktoru a při *j*-té hladině druhého faktoru, roven

$$n_{ij} = \frac{(\text{počet pozorová ní s hladinou 1.faktoru i})(\text{počet pozorová ní s hladinou 2.faktoru j})}{N}$$

Vz. 9-1

kde N je celkový počet všech pozorování. Vzpomeneme-li si na kontingenční tabulky, zjistíme, že kontingenční tabulka počtu pozorování při proporčním uspořádání pokusu je tabulkou nulové hypotézy, tzn. že pro ni je $\chi^2=0$. To tedy znamená, že sledované experimentální faktory jsou vzájemně nezávislé. Pokud nejsou faktory nezávislé, nebude možné zcela odlišit vliv jednotlivých experimentálních faktorů. Dvoucestná analýza variance je (za určitých předpokladů) schopna testovat i data, kde je jen jeden údaj (jedno pozorování) pro každou kombinaci faktorů.

Modelem jednocestné analýzy variance může být rovnice[*]:

[*] Jedná se spíše o symbolické znázornění, rovnice by byla zapsána přesněji.

pozorování = celkový průměr + vliv faktoru + náhodná variabilita

Vz. 9-2

s tím, že testujeme nulovou hypotézu H_0: vliv faktoru=0 pro všechny skupiny. Obdobně, dvoucestnou analýzu variance, se dvěma faktory A a B, můžeme znázornit takto:

pozorování = celkový průměr + vliv A + vliv B + vliv interakce A a B + náhodná variabilita

Vz. 9-3

Zde můžeme testovat tři nulové hypotézy: (1) vliv faktoru A=0 pro všechny hladiny, (2) vliv faktoru B=0 pro všechny hladiny, 3) vliv interakce je 0 pro všechny kombinace faktorů.

Co to znamená, že vliv interakce je roven nule (také říkáme, že interakce je rovna nule, nebo že není interakce mezi faktory)? Pokud je interakce rovna nule, je vliv faktorů čistě aditivní (sčitatelný). Znamená to, že rozdíl průměrů ve skupiných definovaných hladinami faktoru A je konstantní a nezávisí na hladině faktoru B. Obvykle označujeme $\overline{X}_{i\bullet}$ průměr pro všechna pozorování s hladinou prvního (A) faktoru *i*, obdobně pro druhý faktor $\overline{X}_{\bullet j}$, $\overline{X}_{ij}$ je průměr v příslušné buňce a $\overline{X}$ je celkový průměr. Pokud je interakce nulová a zanedbáme vliv náhodné variability, potom platí

$$\overline{X}_{ij} = \overline{X}_{i\bullet} + \overline{X}_{\bullet j} - \overline{X}$$

Vz. 9-4

Jak mohou v případě aditivity vypadat průměry v jednotlivých buňkách (když zanedbáme náhodnou variabilitu) ukazuje Obr. 9-1. Graficky si to můžeme znázornit pomocí tzv. Interakčních diagramů (*interaction plots*). Jejich použití je ilutrováno na konci této kapitoly, kde ukazujeme, jak je sestrojit v programech Statistica a R.

		faktor A	
		a1	*a2*
faktor B	*b1*	15	25
	b2	18	28

		faktor A			
		a1	*a2*	*a3*	*a4*
faktor B	*b1*	21	29	14	25
	b2	18	26	11	22

Obr. 9-1 Možné hodnoty průměrů ve faktoriálním pokusu, pokud je vliv faktorů A a B čistě aditivní, s nulovou interakcí (zanedbáváme náhodnou variabilitu). V prvním případě mají oba faktory jen dvě hladiny. V druhém případě má faktor A čtyři hladiny, faktor B dvě hladiny. Pro každou hladinu faktoru A platí, že při jedné hladině faktoru B se průměr liší o 3 jednotky od průměru druhé hladin**y**.

Výpočet

Předvedeme zde část výpočetního postupu pro vyvážený model dvoucestné analýzy variance. Označme počet hladin faktoru *A* jako *a*, počet hladin faktoru *B* jako *b* a protože předpokládáme vyvážený model, počet pozorování v každé kombinaci hladin faktorů je *n*. Potom platí

$$SS_{TOT} = SS_A + SS_B + SS_{AB} + SS_e$$

Vz. 9-5

kde SS_{TOT} je součet čtverců odchylek všech pozorování od celkového průměru (odpovídající počet stupňů volnosti je $DF_{TOT} = N-1$, N je celkový počet pozorování, tedy $N=abn$). SS_A je součet čtverců odchylek průměrů skupin definovaných faktorem A od celkového průměru, násobených pro každou skupinu odpovídajícím počtem pozorování, spočtený jako bychom počítali jednocestnou analýzu variance. Odpovídající počet stupňů volnosti je $DF_A=a-1$. Zcela obdobně lze spočíst pro druhý faktor SS_B. Reziduální součet čtverců je součet odchylek všech pozorování v buňkách od příslušného průměru buňky; odpovídající počet stupňů volnosti je $DF_e=ab(n-1)$. Interakční součet čtverců SS_{AB} je (počtem případů vážený) součet odchylek průměrů v jednotlivých buňkách od průměrů očekávaných podle Vz. 9-4. Odpovídající počet stupňů volnosti je

$$DF_{AB}=(a-1)(b-1).$$

Vz. 9-6

Vydělením součtu čtverců (*sum of squares*, *SS*) příslušnými stupni volnosti dostáváme hodnotu průměrného čtverce (*mean square*, *MS*).

Podobně jako v jednocestné ANOVě platí, že residuální součet čtverců (tj. uvnitř buněk) je odhadem společné variance. Lze ukázat, že v případě, že vliv faktoru A je nulový, je odhadem společné variance i MS_A, podobně pro faktor B, a pokud je interakce nulová, potom je odhadem společné variance MS_{AB}. Pokud bychom počítali s každou buňkou jako se samostatnou skupinou v jednocestné analýze variance, dostáváme součet čtverců mezi buňkami (*cell SS*), pro který platí $SS_{cell}=SS_A + SS_B+ SS_{AB}$. Pokud žádný z faktorů nemá vliv na hodnotu sledované proměnné, je příslušný *MS* odhadem společné variance. Tuto skutečnost používáme v testech příslušných hypotéz.

Testy hypotéz se v dvoucestné ANOVě liší podle toho, jedná-li se o model I (model s pevnými efekty), nebo model II (s náhodnými efekty). Ve dvoucestné ANOVě máme ještě model III (tzv. smíšený, jeden faktor má pevné a druhý náhodné efekty.) Pokud se jedná o model s pevnými efekty (pravděpodobně nejčastější případ), porovnáváme vždy příslušný průměrný čtverec s reziduálním (MS_E). To znamená, že hodnota testového kriteria pro příslušný faktor (nebo interakci) je

$$F = MS_{faktor} / MS_E.$$

Vz. 9-7

Většina počítačových programů má test pro model I jako předvolbu (*default*), ale některé programy umožňují zvolit, co bude v čitateli F-testu. Výpočetní postup je zdlouhavý, nebudeme ho zde proto předvádět a čtenáře odkazujeme na výsledky získané s příkladovými daty (popsanými v Tab. 9-1; jde o výsledky pokusu popsaného na začátku kapitoly: výšky rostlin ovlivněné dvěma faktory, přídavkem dusíku a vody), tak jak jsou prezentovány v sekcích „Jak postupovat v programu“...

	dusík NE voda NE	dusík NE voda ANO	dusík ANO voda NE	dusík ANO voda ANO
	23	32	29	57
	25	37	28	59
	24	34	29	62
	26	35	31	58
	19	36	30	59
průměr:	**23.4**	**34.8**	**29.4**	**59.0**

Tab. 9-1 Příkladová data.

Ve výsledcích zobrazených ke konci této kapitoly vidíme test pro oba hlavní efekty a také pro jejich interakci: všechny tyto testy vysoce průkazné. Průkaznost interakce značí (v případě našich dat), že výsledný efekt obou faktorů je více než jejich součet (synergismus). Důvodem by zřejmě bylo, že při nedostatku vláhy rostlina není schopna dostatečně využít zvýšenou nabídku živin.

Pokud se jedná o jiný model než s pevnými efekty, musíme pro výpočet testové charakteristiky F použít ve jmenovateli v některých případech MS_{AB} místo MS_e. Kdy a jak doporučuje Tab. 9-2, převzatá z učebnice Zar (1984).

Testovaný efekt	Model I (faktory A i B jsou fixované)	Model II (faktory A i B jsou náhodné)	Model III (faktor A fixovaný ; faktor B je náhodný)
Faktor A	MS_A/MS_e	MS_A/MS_{AB}	MS_A/MS_{AB}
Faktor B	MS_B/MS_e	MS_B/MS_{AB}	MS_B/MS_e
A x B interakce	MS_{AB}/MS_e	MS_{AB}/MS_e	MS_{AB}/MS_e

Tab. 9-2 Výpočet F statistiky pro testy signifikance u dvoufaktorové ANOVA s opakováními.

Sokal a Rohlf (1981) doporučují ještě o něco komplikovanější postup, odlišně zase k testu efektů přistupují moderní implementace lineárních modelů se smíšenými efekty.

Proč se musí výpočty lišit, si předvedeme na pokusu, který vyhodnotíme jednou pomocí modelu I a jednou pomocí modelu III. Máme pokus na vliv hnojení a provádíme jej na třech lokalitách s porostem kostřavy. Za odpověď budeme považovat délku nejdelšího výběžku. Na každé lokalitě vybereme deset rostlin, které pohnojíme a deset rostlin nehnojených bude sloužit jako kontrola. Máme tedy faktoriální uspořádání s faktory hnojení a lokalita. Faktor hnojení je jasně faktor s pevnými efekty. Faktor lokalita mohu chápat jako faktor s pevnými efekty (zajímá mě porovnání konkrétních tří lokalit, a hlavně jenom efekt hnojení na těchto třech lokalitách), ale také jako faktor s náhodnými efekty (zajímá mě, zda se obecně lokality mezi sebou liší, uvedené tři lokality považuji za náhodný výběr ze všech lokalit daného typu v daném regionu, a hlavně, efekt hnojení chci testovat pro všechny tyto lokality).

V prvním případě (tj. lokalita je faktor s pevnými efektem) užiji pro výpočet F statistiky pro faktor hnojení v čitateli reziduální MS, ale pokud bude lokalita faktor s náhodným efektem, použiji interakční MS. V prvním případě dostanu skoro jistě silnější test. Ale to je v pořádku. Pokud jsem považoval faktor lokalita za faktor s pevným efektem, znamená to, že svoje výsledky zobecňuji na všechny rostliny nacházející se na třech zkoumaných lokalitách. Naopak, pokud jsem považoval faktor lokalita za faktor s náhodným efektem, potom svoje výsledky zobecňuji na základní soubor všech rostlin na všech

lokalitách, ze kterých jsou moje uvedené tři lokality náhodným výběrem. Proto zjištěný efekt testuji proti variabilitě odpovědi na hnojení na jednotlivých lokalitách, tj. proti interakčnímu MS. Vyplývá z toho jeden důležitý závěr pro sílu testu. Jestliže je A faktor s pevným efektem a B faktor s náhodným efektem, potom síla testu pro faktor A roste s počtem hladin faktoru B: v našem případě to znamená, že pokud chci svoje výsledky zobecnit na všechny možné lokality, je test tím silnější, čím více lokalit zkoumám.

ANOVA s interakcemi a bez interakcí

V dvoucestné analýze variance jsme dosud uvažovali možnost, že vliv faktorů není čistě aditivní, že mezi nimi je interakce. V některých případech se můžeme předem rozhodnout, že existenci interakce a priori zamítáme. Testujeme pak jenom vliv hlavních efektů, anglicky se taková analýza někdy nazývá *main effects ANOVA*. Potom srovnáváme *MS* pro jednotlivé faktory s *MS*, který vypočteme takto: $(SS_e + SS_{AB})/(DF_e + DF_{AB})$. Veškeré odchylky od aditivity tedy v takovém případě považujeme za náhodnou variabilitu a používáme je k odhadu společné variance. Zar (1984) užívá pro součet v čitateli označení *remainder SS*. Většina statistických programů dovoluje rozhodnout se, zda chceme počítat analýzu variance s interakcemi nebo bez nich.

V dvoucestné (dvoufaktorové) analýze tedy existuje řada voleb, které nám umožňují analýzu přizpůsobit co nejvíce charakteru studovaného problému. Při těchto volbách rozhodujeme vlastně, co je variabilita, která má být naším modelem vysvětlená, a co je variabilita náhodná. Ve vícefaktorové analýze máme voleb ještě více (můžeme uvažovat interakce pouze mezi některými faktory, můžeme uvažovat i o interakcích vyšších řádů, atd.) Výběr modelu by měl být založen na naší apriorní znalosti studovaného problému. Nesprávný je (často užívaný) přístup, kdy spočteme všechny modely a přijmeme ten, jehož výsledky se nejlépe shodují s naší představou, co mělo vyjít. Naproti tomu existují metody, které nám pomohou vybrat model, který nejlépe odpovídá našim datům.

Dvoucestná ANOVA bez opakování

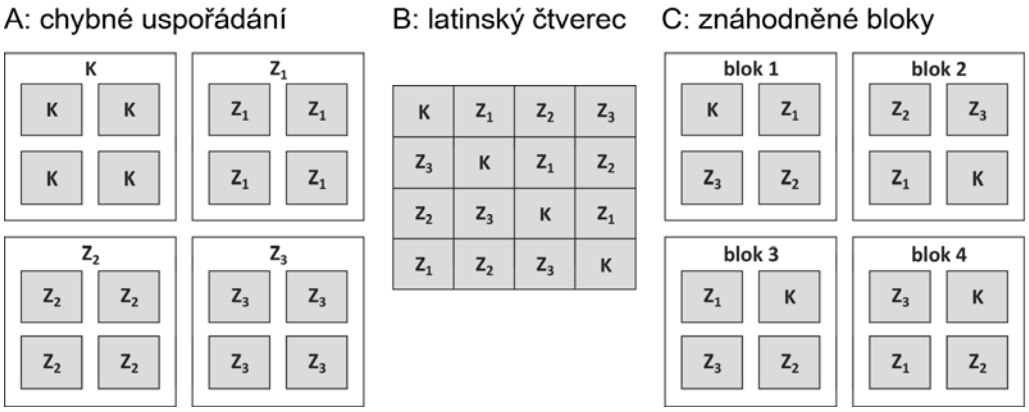
V dvoucestné analýze variance můžeme vyhodnotit i pokus, ve kterém je vždy jen jedno opakování v každé buňce, tj. každá kombinace faktorů je zastoupena právě jedním údajem. V tomto případě musíme použít dvoucestnou analýzu variance bez interakce: tu nemůžeme testovat, protože není žádná variabilita uvnitř buněk a pro odhad celkové variance musíme použít pouze odchylky od aditivity.

Uspořádání pokusů

V této části budou stručně probrány obecné zásady uspořádání pokusů (zejména rozmístění ploch). Budou užity příklady z uspořádání ploch v terénu, ale obdobné postupy se dají užít i v laboratorních podmínkách. Nebudou probírány specifické rysy, které se liší podle typu studovaného materiálu. Sem patří především problémy s vlivem okrajů pokusné plochy (zásah provádíme většinou na větší ploše, hodnotíme její střední část) případně s možným vlivem zásahu na okolí.

Aby bylo možné výsledky terénních pokusů statisticky vyhodnotit, musí se uspořádání experimentu řídit určitými pravidly (viz např. Mead 1988). Většina biotických i abiotických charakteristik se v ploše kontinuálně mění - blízké plochy jsou si proto podobnější než plochy vzdálené. Pokusy v terénu mají velkou tradici v zemědělském

výzkumu, kde byly vypracovány standardní postupy pro uspořádání polních pokusů. Cílem takového uspořádání je co nejvíce omezit vliv heterogenity pokusné plochy při zachování pravidel nutných pro statistické zpracování (nezávislost opakování). Jak ukázal Hurlbert (1984), v ekologii často nejsou tato pravidla dodržována; výsledkem je snížení důvěryhodnosti výsledků a často i chybné závěry. Přestože jsou Hurlbertovy závěry založeny na analýze ekologických článků, nedodržování pravidel a z nich vyplývající nedůvěryhodnost výsledků je zřejmě obecnějším jevem v celé biologii.



Obr. 9-2 Některé možnosti experimentálního uspořádání pokusu.

Jak tedy má a jak nemá uspořádání experimentu v prostoru vypadat? Předpokládejme, že máme jednu kontrolu (K) a tři různé zásahy (Z1, Z2, Z3). Potom možná uspořádání pokusu ukazuje Obr. 9-2. Chybou je (viz část A), pokud je každý typ zásahu proveden na jediné souvislé ploše a v rámci plochy potom odebíráme jednotlivé „vzorky“ (opakování). Taková opakování nejsou nezávislá a statistické zpracování z nich vycházející může vést ke zcela nesprávným výsledkům. Nejběžnější správná uspořádání ukazují části B a C. Část B představuje latinský čtverec - každý řádek a každý sloupec obsahuje právě jednu plochu od každého typu zásahu. V obrázku chybí úplně náhodné uspořádání, kdy máme soubor experimentálních ploch a nějakým náhodným procesem (pomocí generátoru náhodných čísel) rozhodneme, který typ zásahu přiřadíme té které ploše, s tím, že každá plocha má stejnou pravděpodobnost mít kterýkoliv typ zásahu.

Uspořádání v úplných znáhodněných blocích (část C) má některé výhody, zvláště pokud je celková plocha studovaného společenstva dostatečně velká. V tom případě vybereme počet bloků, který se rovná žádanému počtu opakování. Bloky jsou vybírány tak, aby byly vnitřně maximálně homogenní a naopak odlišnosti mezi nimi pokrývaly variabilitu podmínek, přes které chceme naše závěry zobecnňvoat. V každém bloku potom umístíme počet ploch rovnající se počtu typů zásahů a náhodným procesem určíme, který zásah provedeme na které ploše. Toto uspořádání umožňuje dobře odlišit prostorovou variabilitu od vlivu experimentálního zásahu.

V metodách analýzy variance a lineárních modelů můžeme ale obvykle dobře vyhodnotit i další typy uspořádání, např. neúplné znáhodněné bloky. Obecně se prosazuje tendence, že je třeba přizpůsobit experimentální design použitému biologickému materiálu: např. vhodným blokem pro experimenty jsou mláďata pocházející z jednoho vrhu; nelze však donutit samici, aby měla v každém vrhu přesně tolik mláďat, kolik my potřebujeme srovnávat typů zásahů. Proto je nutné užít neúplné znáhodněné bloky. Výpočet je ovšem potom výrazně komplikovanější. Je však třeba zdůraznit, že určitá uspořádání (jako např. Část A v Obr. 9-2) nejdou spolehlivě vyhodnotit žádnou statistickou metodou. Pokud se rozhodnete užívat

nestandardní uspořádání, raději předem konzultujte statistika. Velmi podrobný přehled problematiky podává Mead (1988).

Vyhodnocení pokusů ve znáhodněných blocích a v latinském čtverci

Uspořádání ve znáhodněných blocích je vlastně zobecněním párového uspořádání. ANOVA znáhodněných bloků hodnotí pokusy podobně jako párový t-test, ale srovnáváme zde více než dvě skupiny. Jde vlastně o smíšený model analýzy variance bez opakování. Faktor s pevnými efekty je zkoumaný faktor, faktor s náhodným efektem je blok, interakci nemůžeme stanovit kvůli absenci opakování uvnitř bloků. Podobný ANOVA model se také používá, pokud na témže objektu provádíme opakovaná měření. Např. deseti králíkům nabídneme postupně tři typy stravy (různé druhy trav) v definovaném množství a zjišťujeme, kolik za hodinu sežerou. U každého jedince zkoušíme každý typ jen jednou. Před každým nabídnutím necháme králíka standardní dobu vytrávit. Zjišťujeme preferenci pro typ stravy. „Blokem“ (faktorem s náhodným efektem) je zde každý králík, testovaný faktor je faktor s pevným efektem - typ stravy. Tento přístup nelze ale použít, pokud je nebezpečí, že první zásah může nějakým způsobem ovlivnit reakci individua na druhý zásah, i když tento problém můžeme redukovat větším počtem pokuzných zvířat v kombinaci s pořadím nabízených typů stravy znáhodněným nezávisle pro každé zvíře.

Data z latinského čtverce hodnotíme zvláštním modelem trojcestné analýzy variance. Třemi vysvětlujícími proměnnými jsou řádek, sloupec a typ zásahu. Interakce pochopiteně nemůžeme stanovit.

Mnohonásobná porovnání

Pokud prokážeme průkazný vliv některého faktoru s pevným efektem, který má více než dvě hladiny, můžeme užít kteroukoliv metodu zmíněnou v předešlé kapitole (doporučuji Tukeyho test, pro porovnání všech hladin s kontrolou Dunnettův test). Namísto k (počet skupin) užijeme v rovnicích a nebo b, tedy počet hladin daného faktoru, za n dosazujeme počet všech pozorování při dané hladině faktoru. Pro model I pro s^2 použijeme MS_E, tedy průměrný čtverec uvnitř buňky, s odpovídajícím počtem stupňů volnosti. Pokud provádíme analýzu variance bez opakování (např. znáhodněné úplné bloky), musíme pro s^2 užít 'remainder mean square' s odpovídajícím počtem stupňů volnosti (tj. tu hodnotu DF, která patřila k hodnotě ve jmenovateli při výpočtu hodnoty F statistiky). V některých případech (zvláště pokud je odhadujeme interakci a ta je průkazná) nás bude zajímat ne srovnání hladin faktoru, ale srovnání průměrů v jednotlivých buňkách (tj. porovnání kombinací hodnot faktorů, které jsou součástí dané interakce). V tom případě postupujeme jako po jednocestné ANOVě, kde by každá buňka byla nezávislá skupina; (k - celkový počet skupin je potom $a * b$). Naopak, při průkazné interakci se mnohonásobná porovnání obvykle neprovádějí – pokud efekty nejsou aditivní, porovnání hladin jednoho faktoru závisí na hladině faktoru druhého.

Neparametrické metody

Analýza variance je poměrně robustní k narušení předpokladů. Pokud je ale narušení příliš velké, je jednou z možností užít neparametrické metody. Kruskal-Wallisův test je neparametrickou obdobou jednocestné analýzy variance. Existují také zobecnění Kruskal-Wallisova testu pro dvoucestnou analýzu variance, ale ani Statistica ani R tento postup

nenabízejí, popisuje jej např. Zar (1984, p. 219). Někdy se používá přístup tzv. transformace pořadí (*rank transformation*, RT), při které původní hodnoty převedeme na jejich pořadí a s novými hodnotami pracujeme ve standardním ANOVA modelu. Tento přístup lze ale užít jen pro testování hlavních efektů, ne s ANOVA modely zahrnujícími interakce.

Pro neparametrickou analýzu dat ze znáhodněných bloků je možné použít Friedmanův test. Ten je založen na pořadí v rámci bloků: jsou-li tedy čtyři typy zásahů, každé hodnotě v bloku přiřadíme číslo 1 až 4 podle pořadí v bloku. Vypočítáme testovou statistiku

$$\chi^2_r = \frac{12}{ba(a+1)} \sum_{i=1}^{a} R^2_i - 3b(a+1)$$

Vz. 9-8

kde a je počet hladin studovaného faktoru, b je počet bloků a R_i je součet pořadí pro i-tou hladinu studovaného faktoru. Jestliže např. máme 5 bloků a pro první hladinu studovaného faktoru byla měřená hodnota čtyřikrát nejmenší a jednou druhá nejmenší, potom $R_1 = 6$. Za předpokladu platnosti nulové hypotézy má tato statistika rozdělení blízké χ^2 s $a-1$ stupni volnosti. Shoda je tím lepší, čím jsou a a b větší. Při průkazném výsledku můžeme pokračovat v mnohonásobných porovnáních; existuje jak možnost mnohonásobných porovnání každý s každým (modifikace Tukeyho metody), tak pro porovnání s kontrolou (modifikace Dunnettovy metody), viz Zar (1984, p. 229). Mnohonásobná porovnání ale nejsou k dispozici ani v programu Statistica ani v programu R.

Příkladová data

Jako příklad pro analýzu variance používáme data uvedená v Tab. 9-1. Faktory představující experimentální zásahy (dusík přidaný do substrátu a navýšení zálivky) jsou v proměnných *Nitrogen* a *Water*, výsledná výška rostliny je v proměnné *Height*.

Pro ilustraci analýzy dat ze znáhodněných bloků použijeme výsledky pokusu publikovaného v článku Špačková et al. (1998), ve kterém byl studován vliv modifikace vegetačního pokryvu na obnovu lučních rostlin ze semen. Modifikace spočívala v odstranění rostlinného opadu (*rem_litt*), odstranění (vypletí) dominantního druhu (smilka, *Nardus stricta*; hladina *rem_NS*), odstranění opadu i mechů (*rem_litt_moss*) a navíc zde byly kontrolní plochy bez modifikace (*ctrl*). Příslušnost k jednomu z těchto čtyř typů je zaznamenána v proměnné *Treatment*, jedna plocha od každého typu byla založena v každém ze čtyř experimentálních bloků (faktor *Block*). Celkový počet semenáčků (všech druhů) zaznamenaný v plochách během sezóny je v proměnné *SeedlSum*. V ukázkových analýzách byly pro jednoduchost počty semenáčků použity bez transformace, ale bylo by vhodnější je transformovat (logaritmovat či odmocnit), ještě lepší by bylo použít zobecněný lineární model (GLM) s předpokládanou Poissonovou distribucí (viz kapitola XX).

Příklad ilustrující data sebraná z uspořádání latinského čtverce pochází ze studie Šmilauer & Šmilauerová (2000), ve které byl studován vliv potlačení arbuskulární mykorrhizní symbiózy (pomocí fungicidu bavistinu, viz proměnná *Bav)* v kombinaci s případným zvýšením dostupnosti fosforu v plochách (viz proměnná *P*) na složení lučního rostlinného společenstva. V našem příkladě používáme nadzemní biomasu dominantní trávy *Poa pratensis ssp. angustifolia* (proměnná *PoaAngus*). Umístění plochy v rámci latinského čtverce je určeno faktory *Row* a *Column*, udávajícími příslušnost k řádku a sloupci.

Posledními příkladovými daty jsou údaje pro Friedmanův test, představujících výsledky testování citlivosti osob na různé alergeny. Do kůže se vpraví malé množství

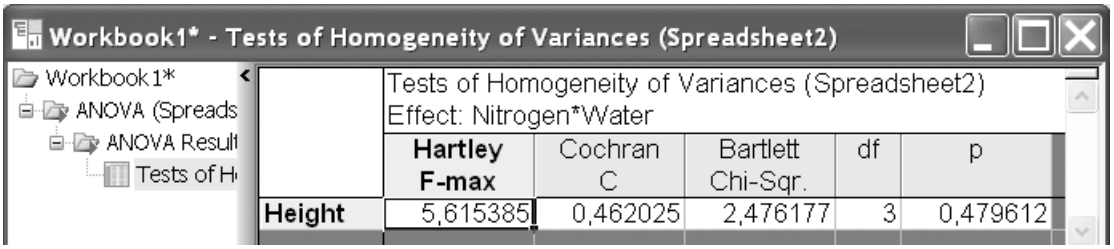
určitého alergenu a sleduje se reakce na stupnici 0 (žádná reakce) až 4 (největší reakce). Každá osoba dostává současně více alergenu, v našem případě byly tři: pelyněk, ambrosie a bříza. Friedmanův test nám umožňuje se vyjádřit k tomu, zda je v populaci na některý z alergenů větší citlivost. Data jsou pro tento typ testu uspořádána tak, že jednotlivé typy zásahů (zde typy alergenů) jsou reprezentovány samostatnými proměnnými (sloupci, *R-Artemisia*, *R-Ambrosia*, *R-Betula*), přičemž pozorování na určité osobě jsou v jednom řádku.

Jak postupovat v programu Statistica

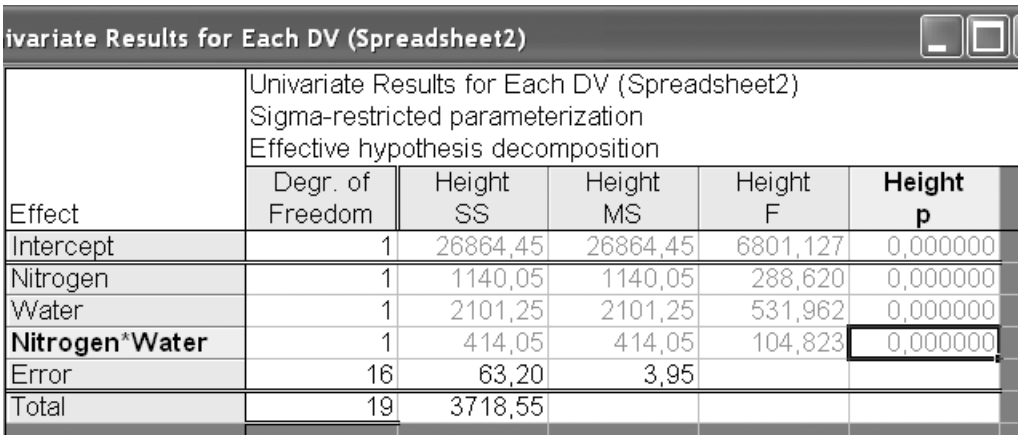
Faktoriální ANOVA se dvěma faktory

Zvolíme příkaz *Statistics* | *ANOVA* a v dialogovém okně vybereme *Factorial ANOVA*. Po volbě tlačítka *OK* se objeví nové dialogové okno, kde nejprve zvolíme za pomocí tlačítka *Variables* proměnnou *Height* v levém seznamu (*Dependent variable list*) a proměnné *Nitrogen* a *Water* v pravém seznamu (*Categorical predictors (factors)*). Pokud bychom chtěli v analýze porovnávat jen některé z kategorií definovaných jedním či druhým faktorem, můžeme je vybrat pomocí tlačítka *Factor codes*, ale jinak jeho použití není nutné a můžeme pokračovat tlačítkem *OK*.

Následně se zobrazí nové dialogové okno *ANOVA Results*. Podobně jako u ANOVA modelu jednoduchého třídění začneme nejprve ověřením předpokladu homogenity variance, a pro to musíme zobrazit toto okno ve větším rozsahu, za pomocí tlačítka *More results* v levém dolním rohu. Pak vybereme záložku *Assumptions* a jako testované skupiny (položka *Effect*) ponecháme nabízenou kombinaci faktorů *Nitrogen*Water*. Díky této volbě budeme porovnávat mezi sebou variance čtyř skupin, definovaných kombinací obou faktorů. Po zvolení tlaíčka *Cochran C, Harley, Bartlett* se objeví výsledek ukazující, že se variance mezi skupinami neliší:



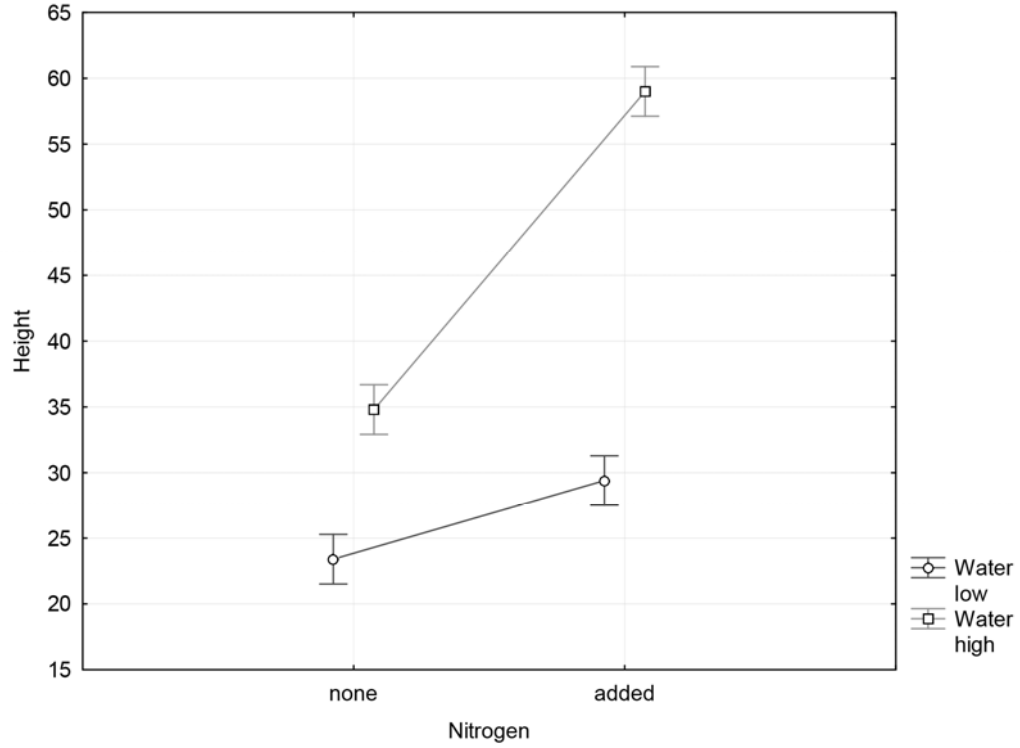
Vrátíme se tedy zpět do dialogového okna *ANOVA Results* a na záložce *Summary* zvolíme *Univariate results*.



Effect	Degr. of Freedom	Height SS	Height MS	Height F	**Height p**
Intercept	1	26864,45	26864,45	6801,127	0,000000
Nitrogen	1	1140,05	1140,05	288,620	0,000000
Water	1	2101,25	2101,25	531,962	0,000000
Nitrogen*Water	1	414,05	414,05	104,823	0,000000
Error	16	63,20	3,95		
Total	19	3718,55			

Výsledná tabulka (opět doporučujeme ignorovat řádek *Intercept*, který jednak testuje zde nesmyslnou nulovou hypotézu, že průměrná výška rostlin je rovna nule, jednak není součástí rozkladu celkové sumy čtverců z řádku *Total*) ukazuje hlavní efekty obou faktorů a také efekt představující interakci mezi těmito faktory. Všechny tyto efekty jsou vysoce průkazné. Průkaznost hlavního efektu proměnné *Nitrogen* nám říká, že se od sebe liší průměry skupiny nehnojených a skupiny hnojených rostlin a podobně u proměnné *Water* a rostlin s nižší a s vyšší zálivkou.

Důležité ale také je, že průkazná interakce *Nitrogen*Water* ($F_{1,16}$=104.8, p<0.000001) nám říká, že míra odlišnost mezi hnojenými a nehnojenými rostlinami závisí na tom, zda měly vyšší či nižší zálivku. Interakce mezi faktory je symetrická, takže ji můžeme také popsat tak, že míra odlišnosti výšky mezi více a méně zalévanými rostlinami závisí na tom, zda byly rostliny přihnojovány či nikoliv. Povahu interakce lze nejlépe určit z interakčního diagramu, který vytvoříme následujícím způsobem. V dialogovém okně *ANOVA Results* zvolíme na záložce *Summary* tlačítko *All effects/Graphs* a v zobrazeném seznamu vybereme interakci (v našem příkladě *Nitrogen*Water*). Ujistíme se, že v oblasti *Display* je zvoleno *Graph* a zmáčkneme tlačítko *OK*. V dalším dialogovém okně zvolíme (v případě interakce dvou faktorů), který faktor je představován symboly s odlišným umístěním podél horizotnální osy (*x-axis, upper*) a který je představován odlišnými čárami (*Line pattern*). Můžeme například zvolit faktor *Nitrogen* v prvním (levém) sloupci a faktor *Water* ve sloupci druhém.



Výsledný graf (text nad grafem byl odstraněn) jasně ukazuje, že efekt přidání dusíku na výšku rostlin byl výraznější v případě rostlin s větší zálivkou. Takto ale nemusí interakce mezi dvěma faktory vypadat vždy: místo jejich "synergického" efektu se mohou jejich efekty vzájemně rušit (a čáry v interakčním diagramu se pak kříží) nebo by mohl nastat méně extrémní a docela častý příklad, kdy se hladiny jednoho z faktorů liší pouze pro určitou hladinu (určité hladiny) faktoru druhého. V našem případě by to mohlo vypadat tak, že pro nízkou zálivku by efekt dusíku nebyl průkazný a pro vysokou ano. V takovýchto případech pak někdy nastává situace, že jeden či oba hlavní efekty nejsou průkazné, interakce faktorů ale průkazná je.

Ještě musíme zdůraznit, že v případě ANOVA modelu s více faktory s pevným efektem nemá smysl provádět mnohonásobná porovnání samostatně pro faktory, které vystupují v průkazném interakčním efektu. V našem případě to ani nepřichází v úvahu, protože oba faktory mají jen dvě hladiny, ale přesto zde můžeme ilustrovat, jak mnohonásobná porovnání v případě průkazných interakcí provádět. Na záložce *Post-hoc* v dialogovém okně *ANOVA Results* musíme v poli *Effect* zvolit interakci, nikoliv hlavní efekt, tedy například *Nitrogen* Water* pro náš příklad a pak zvolit typ srovnání (např. *Tukey HSD*).

Tukey HSD test; variable Height (Spreadsheet2)
Approximate Probabilities for Post Hoc Tests
Error: Between MS = 3,9500, df = 16,000

Cell No.	Nitrogen	Water	{1} 23,400	{2} 34,800	{3} 29,400	{4} 59,000
1	none	low		0,000186	0,001206	0,000185
2	none	high	0,000186		0,002923	0,000185
3	added	low	0,001206	0,002923		0,000185
4	added	high	0,000185	0,000185	0,000185	

Vidíme, že jsou mezi sebou srovnávány možné kombinace faktorů vstupujících do zvolené interakce. S ohledem na výsledky zobrazené v interakčním diagramu výše nás nepřekvapí, že se každá kombinace liší od všech zbývajících.

Nakonec ještě zmíníme skutečnost, že současné přístupy k statistické analýze zdůrazňují, že výsledky by měly obsahovat nejen výsledky testu nulové hypotézy (reprezentované testovou statistikou a průkazností *p*), ale také představu o relativní významu daného faktoru (vysvětlující proměnné) – jak moc jsou hodnoty proměnné závislé (vysvětlované) skutečně ovlivněny (a samozřejmě také jakým směrem, viz například interakční diagram výše). Taková informace se obvykle označuje velikost efektu (*effect size*). Velikost efektu můžeme mezi jednotlivými faktory porovnávat pomocí specializovaných statistik (viz tlačítko *Effect sizes* na záložce *Summary*), ale poměrně dobrou statistikou velikosti efektu je i hodnota F statistiky zobrazená v tabulce analýzy variance. Pro naše příkladová data srovnání F statistik ukazuje, že výška rostlin byla více ovlivněna změnou objemu zálivky, méně pak přihnojením, a odlišnosti výšky oproti předpokládanému aditivnímu efektu zálivky a hnojení byly nejmenší.

Analýza znáhodněných bloků a latinských čtverců

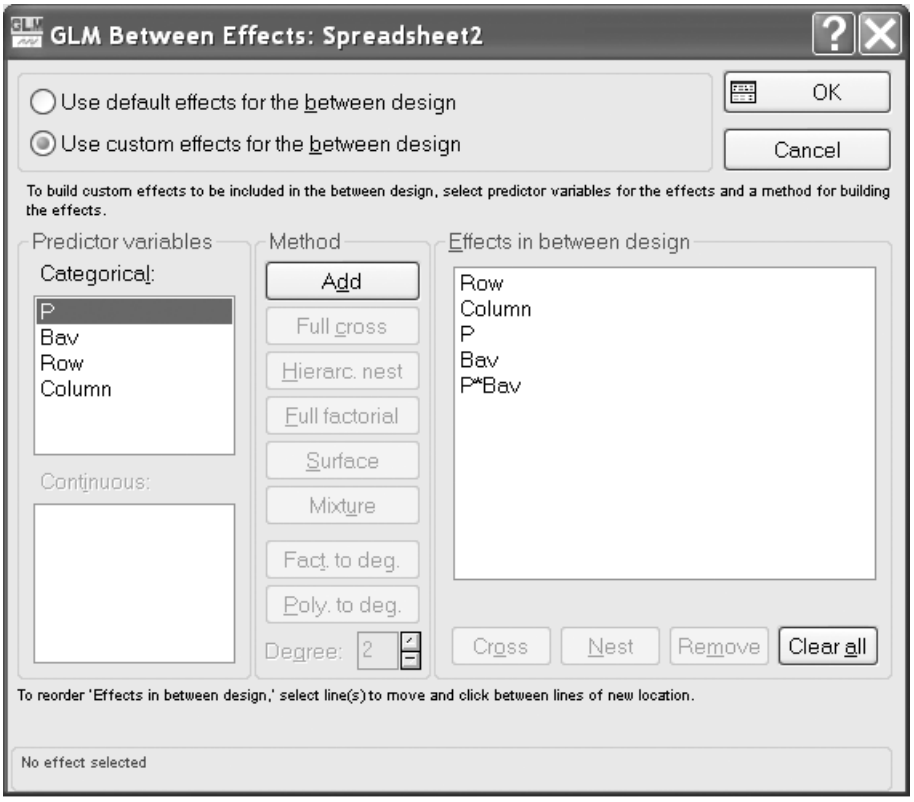
V případě jednoduchých modelů, ve kterých je jen jeden faktor představující bloky na jedné úrovni a jen jeden faktor s pevným efektem, s hladinami zopakovanými v každém bloku se stejným počtem opakování, můžeme použít analýzu variance s hlavními efekty následujícím způsobem. Zvolíme příkaz *Statistics* | *ANOVA* a v seznamu vybereme položku *Main effects ANOVA* a po volbě tlačítka *OK* zadáme proměnné tlačítkem *Variables*: *SeedlSum* jako *Dependent variable* a proměnné *Block* a *Treatment* jako *Categorical predictors*. Při ověřování homogenity variance nemůžeme porovnávat skupiny pozorování pro kombinace hladin obou faktorů, protože pro každou takovou kombinaci máme jen jedno pozorování. Je proto nutné otestovat homogenitu variancí zvlášť pro faktor *Block* a zvlášť pro faktor *Treatment*. Zatímco pro bloky nenachází Bartlettův test žádnou odlišnost, v případě experimentálních faktorů se variance mezi skupinami liší. Zde budeme tuto skutečnost ignorovat, ale správný způsob, jak v takovém případě postupovat, probereme v příští kapitole. Na záložce *Summary* zvolíme tlačítko *Univariate results*.

Effect	Univariate Results for Each DV (Spreadsheet2) Sigma-restricted parameterization Effective hypothesis decomposition Degr. of Freedom	SeedlSum SS	SeedlSum MS	SeedlSum F	SeedlSum p
Intercept	1	122675,1	122675,1	114,7740	0,000002
Block	3	646,7	215,6	0,2017	0,892645
Treatment	3	13539,7	4513,2	4,2225	0,040278
Error	9	9619,6	1068,8		
Total	15	23805,9			

Tento model analýzy lze také zadat pomocí procedury *Repeated measures ANOVA* (nabízené po volbě příkazu *Statistics* | *ANOVA*), ale data pak musí být zadána odlišně (každý typ zásahu jako samostatná proměnná, jednotlivé řádky odpovídají blokům). *Repeated measures ANOVA* se ale užívá především pro zpracování dat, ve kterých jsou stejné objekty měřeny opakovaně v čase, a její použití probereme v kapitole 11.

Pro složitější modely analýzy variance, včetně těch s náhodným efektem bloků, lze užít proceduru *General Linear Models*. Tu teď ukážeme na příkladu analýzy dat z experimentu s uspořádáním ploch v latinském čtverci. Z menu vybereme příkaz *Statistics* | *Advanced Linear/Nonlinear Models* | *General Linear Models* a v zobrazeném seznamu vybereme (poslední) položku *General Linear Models* a pak tlačítko *OK*. V dialogovém okně *GLM General linear models*[1] zvolíme nejprve tlačítko *Variables* a zadáme proměnnou *PoaAngus* v prvním seznamu (*Dependent*) a faktory *P*, *Bav*, *Row* a *Column* v seznamu druhém (*Categorical pred.*). V třetím seznamu (*Continuous pred.*) pro tento model nevybereme nic. Po návratu do původního okna zvolíme tlačítko *Between effects* a zadáme definici vysvětlujících proměnných v modelu (tj. jaké efekty budou v modelu odhadovány). Statistica implicitně navrhuje model pouze s hlavními efekty všech čtyř faktorů, ale nás by zajímala také možná interakce mezi aplikacemi fosforu (*P*) a fungicidu (*Bav*). Interakce s faktory *Row* a *Column* není naproti tomu možné studovat, protože (a) pro každou kombinaci řádku a sloupce máme jen jedno pozorování (jednu plochu) a (b) každá kombinace faktorů *P* a *Bav* se také vyskytuje jen jednou v kterémkoliv řádku či sloupci. V dialogovém okně *GLM Between Effects* vybereme v horní části volbu *Use custom effects for the between design*. Pak v seznamu *Categorical* vybereme faktory *Row* a *Column* a zvolíme tlačítko *Add*. Potom zvolíme v témže seznamu faktory *P* a *Bav* a zvolíme tlačítko *Full factorial*. Výsledná podoba okna před jeho zavřením tlačítkem *OK* bude tedy následující:

[1] Všimněme si, že program Statistica používá zkratku GLM odlišně než mnohé jiné programy, kde znamená *general**ized** linear model* a odpovídá tedy obecnějším typům modelů



Po návratu do okna *GLM General linear models* ještě vybereme záložku *Options* a pomocí tlačítka *Random factors* zvolíme faktory *Row* a *Column* jako proměnné s náhodným efektem (vybereme je v seznamu a zvolíme tlačítko *OK*). Dále postoupíme opět tlačítkem *OK* a v záložce *Quick* zvolíme tlačítko *All effects*.

Univariate Tests of Significance for PoaAngus (Spreadsheet2)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: 2,100804

Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixed	1140,582	1	1140,582	4,780536	53,02746	21,50926	0,006313
Row	Random	64,410	3	21,470	6,000000	4,41338	4,86476	0,047791
Column	Random	107,912	3	35,971	6,000000	4,41338	8,15040	0,015444
P	Fixed	127,408	1	127,408	6,000000	4,41338	28,86849	0,001707
Bav	Fixed	127,069	1	127,069	6,000000	4,41338	28,79181	0,001719
P*Bav	Fixed	43,527	1	43,527	6,000000	4,41338	9,86251	0,020056
Error		26,480	6	4,413				

Vidíme, že biomasa lipnice se liší mezi sloupci i řádky experimentálního uspořádání (to odráží prostorovou variabilitu vlhkosti a dalších půdních vlastností), ale tato variabilita je menší než ta, kterou lze vysvětlit přidáním fosforu a fungicidu nebo také jejich interakcí). Interakční diagram (který zde nezobrazujeme, ale čtenář si jej může zobrazit ze záložky *Quick* volbou tlačítka *All effects/Graphs* a následnou volbou efektu *P*Bav*) pak ukazuje, že lipnice víceméně nereaguje na přidání fungicidu na plochách bez aplikace fosforu, ale po jeho přidání se její biomasa zvýší především tam, kde byl aplikován i fungicid.

Friedmanův test

K výpočtu Friedmanova testu zvolíme příkaz *Statistics* | *Nonparametrics* a v nabídnutém seznamu vybereme položku *Comparing multiple dep. samples (variables)*. Pomocí tlačítka *Variables* nejprve zvolíme proměnné *R-Artemisia*, *R-Ambrosia* a *R-Betula* a po návratu do

okna *Friedman ANOVA by Ranks* zvolíme tlačítko *Summary*. Tabulka s výsledky podává přehled o dílčích výsledcích výpočtu, vlastní výsledek testu je v horní bílé oblasti.

iedman ANOVA and Kendall Coeff. of Concordance (Spreadsheet2)

Friedman ANOVA and Kendall Coeff. of Concordance (S
ANOVA Chi Sqr. (N = 10, df = 2) = 15,76471 p = ,00038
Coeff. of Concordance = ,78824 Aver. rank r = ,76471

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
R-Artemisia	1,300000	13,00000	1,100000	0,994429
R-Ambrosia	2,900000	29,00000	2,900000	0,994429
R-Betula	1,800000	18,00000	1,700000	1,337494

Je vidět, že nulovou hypotézu o shodě distribucí skóre kožní reakce mezi třemi typy alergenů můžeme zamítnout s p=0.00038. Procedura sice nenabízí mnohonásobná porovnání, ale jak průměrné pořadí (*Average Rank* v tabulce výsledků), tak box-and-whisker diagram (který můžeme vytvořit pomocí tlačítka *Box & whisker plot for all variables* v dialogovém okně *Friedman ANOVA by Ranks*) ukazují, že za odlišnost mezi alergeny je zodpovědná výraznější odezva na pyl z ambrosie.

Jak postupovat v programu R

V příkladech příkazů v prostředí R předpokládáme, že data z listu *Chap9* byla naimportována do více datových rámců, lišících se počtem případů: *chap9a* s daty pro faktoriální analýzu variance, *chap9b* s daty pro příklad úplných znáhodněných bloků a také příklad latinského čtverce, a nakonec *chap9c* s daty pro Friedmanův test.

Faktoriální ANOVA se dvěma faktory

Pokud chceme ověřit homogenitu variancí pro ANOVA model se dvěma či více faktory, je asi nejlepším řešením srovnat variabilitu mezi skupinami definovanými kombinací všech použitých faktorů, zde tedy čtyř kombinací (ne)přidání dusíku a standardní/zvýšené zálivky:

```
> bartlett.test(Height~interaction(Nitrogen,Water),data=chap9a)
        Bartlett test of homogeneity of variances
data:  Height by interaction(Nitrogen, Water)
Bartlett's K-squared = 2.4762, df = 3, p-value = 0.4796
```

Mezi skupinami tedy není průkazný rozdíl ve variabilitě. Nafitujeme proto model dvoufaktorové analýzy s interakcí:

```
> summary(aov(Height~Nitrogen*Water,data=chap9a))
              Df Sum Sq Mean Sq F value   Pr(>F)
Nitrogen       1 1140.0  1140.0   288.6 1.17e-11 ***
Water          1 2101.3  2101.3   532.0 1.05e-13 ***
Nitrogen:Water  1  414.0   414.0   104.8 1.98e-08 ***
Residuals      16   63.2     3.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

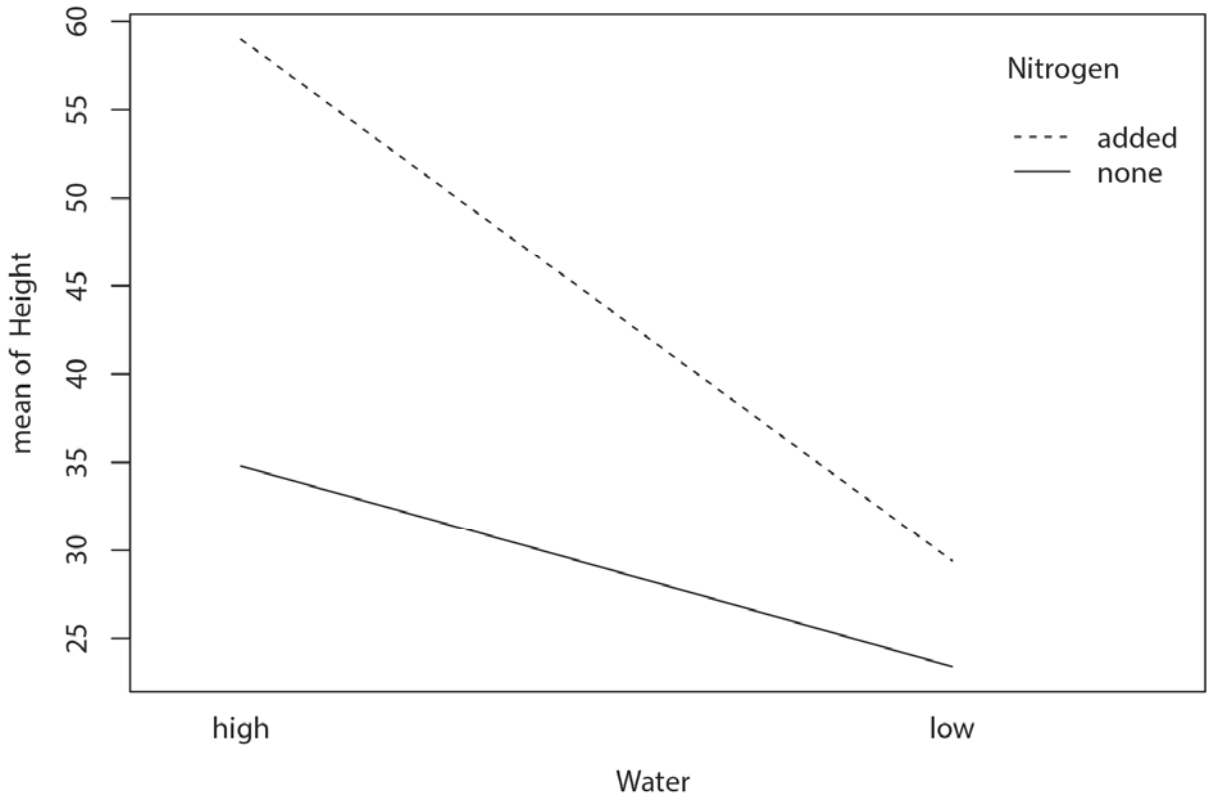
Oba hlavní efekty dusíku a zálivky mají výrazný a průkazný efekt, významná je ale také interakce mezi oběma faktory. Pro podrobnější interpretaci výsledků odkazujeme na komentáře v části *Jak postupovat v programu Statistica*. Zápis *Nitrogen*Water* použitý v zadání modelu funkci *aov* odpovídá něčemu jinému než v případě programu Statistica, kde se takto značí samotná interakce. V programu R se vlastní interakce zapisuje jako *Nitrogen:Water* a zápis *Nitrogen*Water* představuje jak hlavní efekty, tak jejich interakci.

Stejný model bychom tedy zadali i zápisem *Nitrogen+Water+Nitrogen:Water* (Statistica této kombinaci hlavních efektů a interakce říká "full cross").

Interakční diagram vytvoříme v programu R takto:

```
> with(chap9a, interaction.plot(Water, Nitrogen, Height))
```

a výsledný diagram vypadá zhruba takto:



Pokud bychom chtěli provést mnohonásobná porovnání na kombinacích obou faktorů (viz sekce pro program Statistica pro detailnější diskusi), můžeme to provést následovně:

```
> library(multcomp)
> NW <- with(chap9a,interaction(Nitrogen,Water))
> summary(glht(aov(Height~NW,data=chap9a),linfct=mcp(NW="Tukey")))
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = Height ~ NW, data = chap9a)
Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
none.high - added.high == 0  -24.200      1.257 -19.252  < 0.001 ***
added.low - added.high == 0  -29.600      1.257 -23.548  < 0.001 ***
none.low - added.high == 0   -35.600      1.257 -28.322  < 0.001 ***
added.low - none.high == 0    -5.400      1.257  -4.296  0.00288 **
none.low - none.high == 0    -11.400      1.257  -9.069  < 0.001 ***
none.low - added.low == 0     -6.000      1.257  -4.773  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Analýza znáhodněných bloků a latinských čtverců

Analýzu příkladových dat pro znáhodněné bloky můžeme pro tento nejjednodušší design provést s použitím bloku jako pevného faktoru ve standardním modelu analýzy variance bez interakce:

```
> summary(aov(SeedlSum~Block+Treatment,data=chap9b))
            Df Sum Sq Mean Sq F value Pr(>F)
Block        3    647     216   0.202 0.8926
Treatment    3  13540    4513   4.223 0.0403 *
Residuals    9   9620    1069
```

Pro složitější modely je ale třeba postupovat odlišně, s použitím výrazu *Error*, jak ukazujeme níže. Pro naše data jsou výsledky totožné, s výjimkou skutečnosti, že není zobrazován test odlišnosti mezi bloky.

```
> summary(aov(SeedlSum~Treatment+Error(Block),data=chap9b))
Error: Block
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3  646.7   215.6

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Treatment  3  13540    4513   4.223 0.0403 *
Residuals  9   9620    1069
```

Model analýzy variance pro příkladová data z latinského čtverce můžeme fitovat v programu R následujícím způsobem:

```
> summary(aov(PoaAngus~Row+Column+P*Bav,data=chap9b))
            Df Sum Sq Mean Sq F value  Pr(>F)
Row          3  64.41   21.47   4.865 0.04779 *
Column       3 107.91   35.97   8.150 0.01544 *
P            1 127.41  127.41  28.868 0.00171 **
Bav          1 127.07  127.07  28.792 0.00172 **
P:Bav        1  43.53   43.53   9.863 0.02006 *
Residuals    6  26.48    4.41
```

Vidíme zde výrazné (a průkazné) efekty fosforu a fungicidu a také jejich interakce, k interpretaci směru a velikosti efektů můžeme použít buď interakční diagram nebo odhady parametry ANOVA modelu. Ten lze totiž považovat za speciální případ regresního modelu, ve kterém jsou vlivy jednotlivých proměnných vyjádřeny regresními koeficienty (viz kapitola XX). Ty získáme pro náš model následujícím příkazem:

```
> coef(aov(PoaAngus~Row+Column+P*Bav,data=chap9b))
(Intercept)       Rowr2       Rowr3       Rowr4    Columnc2    Columnc3
    4.23625     2.76000    -0.75250    -2.84500    -3.09500     0.58000
   Columnc4        Pyes      Bavyes Pyes:Bavyes
    4.21750     2.34500     2.33750     6.59750
```

Hodnoty koeficientů nám ukazují (budeme zde ignorovat rozdíly mezi řádky a sloupci), že plochy s přidaným fosforem (ať již byl přidán fungicid nebo ne) měly v průměru o 2.345 g vyšší (*Pyes*) biomasu lipnice a podobně plochy s přidaným fungicidem (průměrováno přes plochy bez i s fosforem) měly v průměru biomasu o 2.3375 g vyšší (*Bavyes*). Vzhledem k průkazné interakci ale neměly plochy s aplikací fosforu i Bavistinu biomasu vyšší o (2.345+2.3375), ale o (2.345+2.3375+6.5975), viz regresní koeficient *Pyes:Bavyes*.

Friedmanův test

Vzhledem k uspořádání, ve kterém máme příkladová data pro Friedmanův test (každý zásah, zde typ alergenu, jako samostatná proměnná), jej můžeme v programu R spočítat takto:

```
> with(chap9c, friedman.test(cbind(R.Artemisia,R.Ambrosia,R.Betula)))
        Friedman rank sum test
data:  cbind(R.Artemisia, R.Ambrosia, R.Betula)
Friedman chi-squared = 15.7647, df = 2, p-value = 0.0003773
```

V programu R můžeme ale tento neparametrický test provést i na datech ve stejném uspořádání jako pro klasický test s úplnými znáhodněnými bloky, tj. se všemi hodnotami reakcí v jedné numerické proměnné a s dvěma faktory, kódujícími příslušnost pozorování k osobě a příslušnost pozorování k typu alergenu (číslo 10 odpovídá počtu sledovaných osob, číslo 3 počtu typů zásahů):

```
> Resp <- with(chap9c, c(R.Artemisia,R.Ambrosia,R.Betula))
> Person <- as.factor(rep(1:10,3))
> Allergen <- as.factor(rep(c("Artem","Ambros","Betula"),rep(10,3)))
> friedman.test(Resp~Allergen|Person)
        Friedman rank sum test
data:  Resp and Allergen and Person
Friedman chi-squared = 15.7647, df = 2, p-value = 0.0003773
```

Popis metod v článku

Methods

The effects of nitrogen addition and increased water availability were studied in a factorial experiment (five replications of each treatment levels combination), with completely randomized position of individual plants. Data were analysed using a two-factor ANOVA model, with a post-hoc test of differences among the four combinations of experimental factors using Tukey HSD method.

The effects of plant cover manipulation were evaluated using analysis of variance with an added effect of block.

nebo alternativně

... using general linear model with the main effects of manipulation and block identity.

The differences in skin response to three allergens were tested using Friedman (ANOVA) test.

Results

The results of the ANOVA method are summarized in Table X. The differences in the average values can be seen in Figure Y. The effect of nitrogen addition was less pronounced than the effect of added water, but their joing application had a synergistic effect.

Figure Y neuvádíme, ale byl by to nejspíše sloupečkový diagram pro čtyři kombinace zálivky a hnojení, s přidanými standardními efekty odhadu průměru; v tomtéž diagramu by mohly být písmenky nad sloupci reflektovány i průkazné rozdíly z mnohonásobného porovnání – viz kapitola 8 pro příklad – ale v našem příkladě by stačilo do popisky obrázku uvést, že se všechny kombinace od sebe navzájem lišily s $p<0.01$. Jinou možností je prezentovat interaction plot (například tak, jak jsme ho vytvořili v programu Statistica). Pak ale nezapomeneme v popisku uvést, že chybové úsečky (error bars) jsou 95% konfidenční intervaly, a je vhodné uvést, že spojnice průměrů neznačí interpolaci, ale že nerovnoběžnost spojnic slouží k lepší vizualizaci interakce.

Table X: ANOVA table of nitrogen and water effects upon plant height. The $F_{1,16}$ column represents the values of test statistic for each model term, *p* represents the test significance for this term.

	$F_{1,16}$	p
nitrogen	288.6	<0.001
water	532.0	<0.001
nitrogen*water	104.8	<0.001

We have found significant differences among the plant cover treatments ($F_{3,9}$=4.22, p=0.040). Multiple comparisons suggest that the only significant contrast was between the removal of both the plant litter and of moss layer and the removal of dominant grass (p=0.032), with highest seedling density at the plot with litter and moss removal. (v našem příkladě jsme mnohonásobná porovnání neprováděli, ale pro reálnou studii by to bylo nezbytné).

ANOVA results are summarized in Table W (zde ji neuvádíme, jde o obdobné zjednodušení výsledků zobrazených v programu Statistica či R jako v prvním příkladě). Both the phosphate and fungicide applications significantly affected the aboveground biomass of *Poa* and we have also demonstrated their synergistic effect, with the biomass almost four times higher in plots with joint application of phosphate and fungicide, as compared to control plots (see Figure Z). (opět obrázek vynecháváme).

Skin response differed significantly among the three allergen types (χ^2_2=15.76, p=0.0004), with the strongest response observed for the *Ambrosia* pollen grains.

Doporučená četba

Hurlbert S.H. (1984): Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187-211.

Mead R. (1988): The design of experiments. Statistical principles for practical applications. Cambridge University Press.

Šmilauer P. & Šmilauerová M. (2000): Effect of AM symbiosis exclusion on grassland community composition. Folia Geobotanica 35: 13-25

Špačková I., Kotorová I. & Lepš J. (1998): Sensitivity of seedling recruitment to moss, litter and dominant removal in an oligotrophic wet meadow. Folia Geobotanica 33: 17-30.

Sokal R.R & Rohlf F.J. (1981): pp. 321-371; Zar J. H. (1984): pp. 206-235; Quinn & Keough (2002): pp. 221-259 (faktoriální uspořádání s dvěma a více faktory), pp. 262-300 (analýza znáhodněných bloků a jednoduchých opakovaných měření na objektech)

10 Transformace dat v analýze variance

Jak ve Studentově t-testu, tak v analýze variance předpokládáme, že data splňují jisté předpoklady. Základním předpokladem je, že data představují náhodné nezávislé výběry. Pokud data tento předpoklad nesplňují (třeba tím, že byly vybírány úmyslně jako typické exempláře, nebo že bylo použito nevhodného experimentálního uspořádání - viz předcházející kapitolu), je důvěryhodnost statistického zpracování vážně narušena a žádná statistická procedura tento nedostatek nemůže napravit. Dalšími předpoklady jsou: (1) předpoklad normality: každý vzorek pochází ze základního souboru s normálním rozdělením[x], (2) předpoklad homogenity variance (*homogeneity of variances* nebo *homoscedasticity*): srovnávané základní soubory mají stejné variance, a (3) předpoklad aditivity (sčitatelnosti) efektů hladin jednotlivých faktorů. Předpoklady (1) a (2) jsme již diskutovali, víme, že proti jejich narušení je ANOVA vcelku robustní, pokud toto narušení není příliš velké. Pro t-test také existuje přibližná metoda, jak postupovat při narušení předpokladu rovnosti variancí. Co znamená předpoklad aditivity? Znamená to, že se vlivy sčítají (odchylky od součtu potom považujeme za interakci). Vycházíme přitom z modelu analýzy variance:

pozorování = celkový průměr + vliv fakt. A + náhod. variab.

Vz. 10-1

případně pro dvoucestnou ANOVu

pozorování = celkový průměr + vliv faktoru A + vliv faktoru B + vliv interakce A a B + náhodná variabilita

Vz. 10-2

Aditivní uspořádání bez interakcí v dvoucestné ANOVě ukazuje část A v Tab. 10-1. Mnohé vlivy pozorované v přírodních vědách nemají aditivní účinek, ale multiplikativní (násobný) účinek. Pěstujeme-li v půdě obohacené dusíkem rostlinu, která v nehnojemé půdě vyroste 20 cm, ve hnojené vyroste například 25. Pěstujeme-li rostlinu, která v nehnojené vyroste 40, v hnojené vyroste 50. V obou případech se hnojením výška zvýší ne **o** určitý počet cm, ale 1,25 **krát**. Faktor má tedy multiplikativní účinek (příkladem je část B v Tab. 10-1). Potom je často multiplikativní i vliv náhodné variability. Modelem takového uspořádání (pro jednocestnou ANOVu) je

pozorování = celkový průměr × vliv faktoru A × náhodná variabilita

Vz. 10-3

V tomto uspořádání je jak průměr, tak náhodná variabilita násobena vlivem faktoru A. Z toho plyne, že můžeme očekávat, že bude existovat lineární závislost směrodatné odchylky na průměru. Pokud celou rovnici logaritmujeme, dostáváme

log(pozorování) = log(celkový průměr) + log(vliv faktoru A) + log(náhodná variabilita)

Vz. 10-4

[x] Ale pozor: pokud by čtenář chtěl tento předpoklad testovat (což mu nedoporučujeme), musí si uvědomit, že pozorování nemusí pocházet jen z jedné distribuce. Pokud totiž například v jednocestném ANOVA modelu s faktorem se třemi hladinami neplatí nulová hypotéza, data pro každou ze tří skupiny by měla pocháze ze tří normálních distribucí lišících se průměrem a nelze je proto srovnávat v rámci jednoho testu, ale pro jednotlivé skupiny odděleně. Tím se ale ještě více sníží síla testu pro malá data. Alternativně (a správněji) bychom měli testovat residuály (odchylky hodnot závislé proměnné od skupinových průměrů) s normální distribucí s nulovým průměrem.

Z toho plyne, že má-li faktor multiplikativní účinek na měřenou proměnnou, má aditivní účinek na její logaritmus. Příklad je v Tab. 10-1, část C.

Hypotetické uspořádání pro dvoufaktorovou analýzu variance, ve kterém jsou efekty faktorů aditivní (data v gramech)
Část A:

	Faktor A		
Faktor B	**Hladina 1**	**Hladina 2**	**Hladina 3**
Hladina 1	10	20	25
Hladina 2	20	30	35

Část B:
Hypotetické uspořádání pro dvoufaktorovou analýzu variance, ve kterém jsou efekty faktorů multiplikativní (data v gramech)

	Faktor A		
Faktor B	**Hladina 1**	**Hladina 2**	**Hladina 3**
Hladina 1	10	30	60
Hladina 2	20	60	120

Část C:
Hypotetické uspořádání, vycházející z předcházející tabulky, ukazující logaritmy původních hodnot.

	Faktor A		
Faktor B	**Hladina 1**	**Hladina 2**	**Hladina 3**
Hladina 1	1.00	1.48	1.78
Hladina 2	1.30	1.78	2.08

Tab. 10-1 Aditivní a multiplikativní efekty (vlivy)

Pokud nejsou splněny podmínky analýzy variance, máme několik možností: buď užijeme některou odpovídající neparametrickou metodu, nebo použijeme zobecněné lineární modely (viz kapitola XX) nebo provedeme **transformaci dat**. Transformace dat je taková operace, kdy matematickou operací získáme z jedné proměnné (X) jinou proměnnou (řekněme X'). V mnoha případech se nám podaří jednou operací změnit data tak, že rozumně vyhovují všem třem podmínkám. Nejčastěji užívaná transformace jsou logaritmická, pro některé typy dat se užívají i transformace arcsinová (pro relativní podíly) a odmocninová (pro počty případů).

Logaritmická transformace

Pro data na poměrové stupnici, kde mají faktory multiplikativní účinek a směrodatná odchylka je lineárně závislá na průměru (jak jsme si ukázali na příkladu rovnice ve Vz. 10-3, oba tyto jevy se často vyskytují současně), je doporučeno užít pro vysvětlovanou (závislou) proměnnou logaritmickou transformaci. Logaritmická transformace má tvar

$$X' = \log(X).$$

Vz. 10-5

V případě, že data obsahují nuly (samozřejmě předpokládáme, že data neobsahují záporná čísla)

$$X' = \log(X+c),$$

Vz. 10-6

kde *c* je konstanta, často jedna (nebo jiné číslo zhruba odpovídající minimálním nenulovým hodnotám v datech). . Obě transformace mají podobné účinky. Jak je vidět z Tab. 10-1, logaritmická transformace je schopna změnit multiplikativní účinek na aditivní, pro log(x+c) to ale platí pouze přibližně. Proto v případě, že data neobsahují nuly užíváme log(*X*), a ne log(x+c). Pokud je směrodatná odchylka lineárně závislá na průměru (tzn. pokud je variační koeficient konstantní) mají logaritmované hodnoty přibližně stejné variance.

Logaritmická transformace ovšem působí i na typ rozdělení: dokáže „zesymetričtit" výrazně pozitivně šikmá rozdělení. Pokud má log(*X*) normální rozdělení, potom se rozdělení proměnné *X* nazývá logaritmicko-normální, zkráceně lognormální, a je pozitivně šikmé. Zkušenost i teoretické úvahy naznačují, že pokud mají faktory multiplikativní účinky, můžeme očekávat lineární závislost směrodatné odchylky na průměru a zároveň lognormální rozdělení dat.

Pro mnoho datových souborů skutečně platí, že mají lognormální rozdělení (obecně se říká, že v biologii je lognormální rozdělení častější, než normální) a zároveň spíše koeficient variance, než variance sama, je konstantní. U takových dat je často účinek faktoru multiplikativní: v takovémto případě dokáže logaritmická transformace vyřešit všechny problémy najednou. V mnoha případech je užitečné ji užít např. pro charakteristiky představující hmotnosti či rozměry a také pro počty individuí, zvláště pokud jsou individua rozmístěna výrazně shlukovitě. I když zde jsou teoreticky správnější tzv. kontagiosní rozdělení, např. negativně binomické, dokáže lognormální rozdělení data dobře aproximovat.

Dlužno upozornit na některá nebezpečí. Odlogaritmováním průměru logaritmovaných hodnot získáme nikoliv aritmetický ale geometrický průměr, který je vždy nižší než aritmetický. Paušální užití logaritmické (ale i kterékoliv jiné) transformace bez ohledu na to, jak data vypadají, může vést k nesmyslům, jak ukazuje následující příklad. Srovnáváme počty individuí ve dvou plochách, v každé ploše máme 18 zkusných jednotek. V první ploše jsou individua rozmístěna zcela pravidelně, v druhé shlukovitě, takže dostáváme následující počty:

Plocha 1: 18-krát 20 individuí. (log (20) = 1.30)

Plocha 2: 16-krát 10, dvakrát 100. (log(10)=1; log(100)=2).

V obou případech je průměrný počet individuí 20 na jednotku. Pokud použijeme logaritmovaných hodnot, získáme v prvé lokalitě průměr 1.30 a ve druhé *(16 x 1 + 2 x 2)/18 = 1.11*, statistický test by ukázal průkazný rozdíl. Užili jsme logaritmickou transformaci tam, kde ani neplatilo, že směrodatná odchylka je lineárně závislá na průměru, ani neměla originální data lognormální rozdělení.

Arcsinová transformace

Pokud mají být hodnoceny údaje o poměrech (o procentech), užívá se tzv. angulární nebo arcsinová transformace. Označme procenta (vyjádřená jako číslo od nuly do jedné) *p*. Potom proměnná

$$p' = \arcsin\sqrt{p}$$

Vz. 10-7

má přibližně normální rozdělení. Transformace se doporučuje používat, pokud se v datech vyskytují i hodnoty *p* menší než 0.3 nebo větší než 0.7. Pokud jsou hodnoty *p* mezi 0.3 a 0.7, je rozdělení dostatečně blízké normálnímu.

Odmocninová transformace

Pokud jsou data brána ze souboru charakterizovaného Poissonovou distribucí, můžeme použít odmocninovou transformaci (ovšem logaritmická transformace je zde také dobrou volbou). Poissonova distribuce bude probírána později (kapitola 17): její vlastností je, že se variance rovná průměru. Poissonova distribuce je charakteristická např. pro počty individuí ve zkusných jednotkách, pokud jsou individua rozmístěna náhodně. Používá se buď odmocnina, nebo lépe transformace

$$X' = \sqrt{X + 0.5}$$

Vz. 10-8

doporučovaná zvláště tehdy, když data obsahují nuly, nebo

$$X' = \sqrt{X + \frac{3}{8}}$$

Vz. 10-9

Výsledná proměnná má rozdělení blízké normálnímu a variance je méně závislá na průměru.

Někdy se užívá i Box-Cox(ova) transformace (viz např. Sokal a Rohlf 1981, p. 423). Je to v podstatě forma mocninné transformace, s tím, že hodnota parametru transformace je určována pomocí iterační procedury na základě dat tak, aby transformace vedla k co nejlepší shodě dat s normalitou, případně s homogenitou variance.

Všechny transformace se dají užít i pro výpočet průměru a jeho konfidenčního intervalu, a to tak, že nejprve vypočteme průměr a hranice konfidenčního intervalu transformovaných hodnot, a poté vše „odtransformujeme“. Nicméně průměr je poněkud vychýlený a konfidenční interval je asymetrický (což je rozumný výsledek, protože původní rozdělení je také asymetrické). Není ale smysluplné uvádět odtransformované hodnoty střední chyby průměru. Tyto odhady jsou výrazně robustní (méně citlivé) k extrémním hodnotám, které se mohou, zvláště v rozděleních výrazně pozitivně šikmých, vyskytnout (tzv. *long-tailed distributions* - rozdělení s dlouhým ocasem).

Užití transformací je poněkud kontroversním tématem; někteří autoři proti jejich užívání varují. Naproti tomu je v mnoha případech tradičně užíváme, aniž si to uvědomujeme Měříme-li například kyselost půdy či vody hodnotou pH, užíváme logaritmické transformace koncentrace vodíkových iontů. Tato koncentrace má zřejmě lognormální rozdělení, její variabilita je závislá na průměru a lze rozumně očekávat, že účinek různých faktorů i náhodné variability je multiplikativní. Nikoho nenapadne, že by bylo vhodné pracovat přímo s koncentrací vodíkových iontů. Lze také argumentovat tak, že rozhodnutí ponechat hodnoty na škále našich měření je (do značné míry arbitrární) transformací. Velmi často je zvolená škála dána měřící procedurou a nemáme žádnou záruku, že je optimální pro popis

procesů probíhajících v přírodě. To je případ například koncentrace živin v půdě či ve vodě, nejen koncentrace vodíkých kationtů.

Pro jednoduché statistické modely typu analýzy variance či regresní analýzy je lepším řešením používání zobecněných lineárních modelů (GLM), které mimochodem také zahrnují implicitní změnu škály vysvětlované proměnné (viz kapitola XX). I v případě GLM a podobných typů modelů (GAM, GLMM, atd.) ale stále zůstává problém vhodné škály pro vysvětlující (nezávislé) kvantitativní proměnné. Pro více složité statistické modely (včetně ordinačních metod, viz kapitola XXX) zobecnění typu změny z lineárního modelu či analýzy variance na GLM nejsou příliš dostupná a transformace analyzovaných dat má zde nezastupitelnou roli.

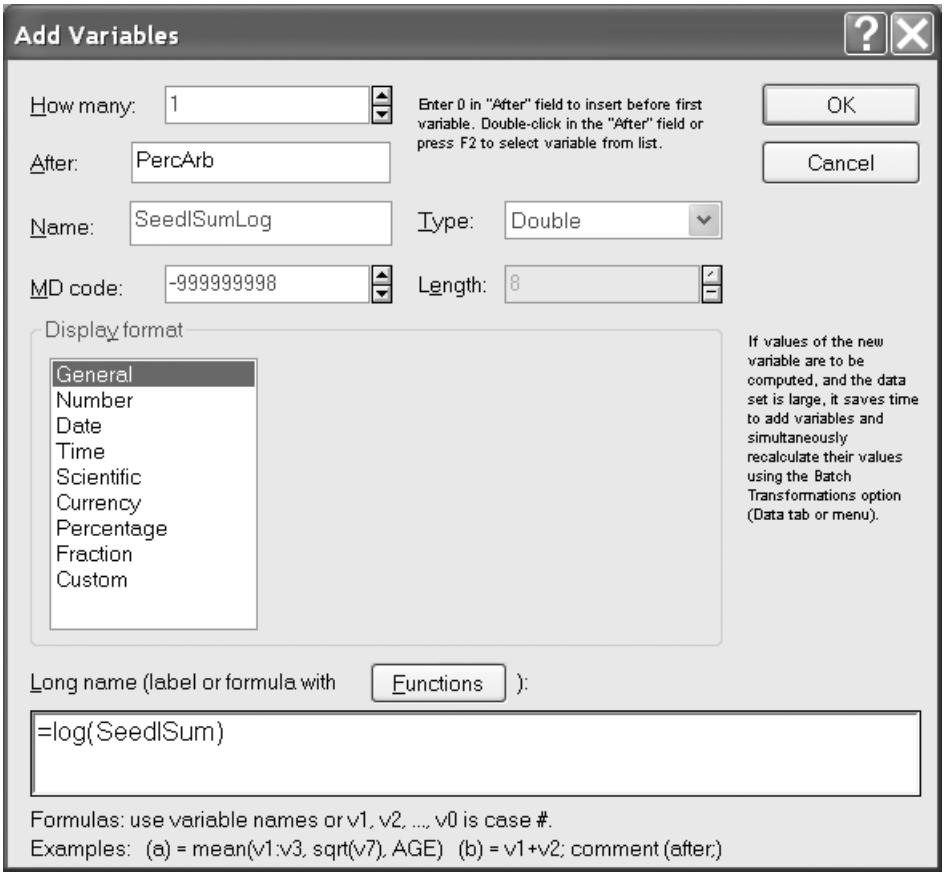
Příkladová data

V prvém příkladě se vrátíme k datům z předchozí kapitoly, popisujícím vztah počtu semenáčků lučních rostlin (*SeedlSum*) pozorovaných na experimentálních plochách k manipulaci s vegetací těchto ploch (*Treatment*). Jde o pokus s uspořádáním úplných znáhodněných bloků, jejichž identita je v proměnné *Block*. V minulé kapitole jsme zjistili, že je porušen předpoklad homogenity variancí a tady si ověříme, zda tento problém odstraní logaritmická transformace počtu semenáčků.

Druhý příklad představuje informaci o procentické frekvenci mykorrhizních hub v kořenech medyňku (*Holcus lanatus*), odebraných z ploch lišících se kosením (proměnná *Mown* s hodnotami *yes* pro kosené a *no* pro nekosené plochy) a zvýšenou nabídkou fosfátů (proměnná *P* s hodnotami *yes* pro přihnojované a *no* pro nepřihnojované plochy). Frekvence hub (*PercArb*) představuje údaj, kolik ze 100 průniků kořenů (pozorovaných pod mikroskopem) s mřížkovým rastrem v okuláru odpovídalo místům s mykorrhizní symbiózou. Každé pozorování (rostlina medyňku) bylo odebráno z odlišné plochy.

Jak postupovat v programu Statistica

Pro první příklad musíme nejprve vytvořit v tabulce dat importovaných do programu Statistica novou proměnnou odpovídající logaritmovaným počtům semenáčků (v proměnné *SeedlSum*). K tomu zvolíme z menu příkaz *Data* | *Variables* | *Add* a v dialogovém okně *Add Variables* provedeme tyto změny: zvolíme název nové proměnné v políčku *Name* (například *SeedlSumLog*) a pak zadáme definici hodnot této proměnné v políčku *Long name* v dolní části okna. Jde o vzoreček podobný těm, které užíváme v programu Excel, jména funkcí ale nejsou vždy stejná a na existující data odkazujeme odlišně. Můžeme zde například zapsat *=log(SeedlSum)*, ale lze použít i zápis *=log(v3)* – pokud je proměnná *SeedlSum* umístěna ve třetím sloupci datové tabulky (na velikosti písmen při odkazu na funkci *log* nezáleží). Použili jsme přirozený logaritmus (dekadickému logaritmu odpovídá funkce *log10*), ale základ nehraje žádnou roli (hodnoty se liší jen násobnou konstantou, tedy škálou). Výsledný stav okna může vypadat takto:



Po zmáčknutí tlačítka *OK* se objeví nová proměnná v datové tabulce, obsahující logaritmy počtu semenáčků. Tu pak použijeme místo proměnné *SeedlSum* při zadání analýzy variance s hlavním efekty, tak jak bylo popsáno v předchozí kapitole. Bartlettův test jasně ukazuje (χ^2_3=3.41, p=0.332), že se variance počtu semenáčků mezi čtyřmi typy zásahů neliší. Efekt zásahu ale zůstává nadále průkazný:

	Univariate Results for Each DV (Spreadsheet22) Sigma-restricted parameterization Effective hypothesis decomposition		
Effect	SeedlSumLog MS	SeedlSumLog F	SeedlSumLog p
Intercept	308,2277	2881,960	0,000000
Block	0,0441	0,413	0,747943
Treatment	0,4735	4,427	0,035778
Error	0,1070		
Total			

Pro data o frekvenci mykorrhizní symbiózy použijeme arcsinovou transformaci, nesmíme ale zapomenout na to, že údaje musí být nejprve převedeny do rozsahu 0 až 1 a pak také odmocněny před aplikací funkce *arcsin*. Použijeme tedy stejný postup jako v případě logaritmické transformace výše, ale v dialogovém okně *Add Variables* zadáme vzoreček *=ArcSin(sqrt(PercArb/100))* .

Tato data lze analyzovat faktoriální analýzou variance, proměnnou *PercArbSin* s transformovanými daty zadáme jako *Dependent variable* a oba faktory (*Mown* a *P*) jako *Categorical predictors*. Po ověření homogenity variancí můžeme zobrazit výslednou tabulku s testy pomocí tlačítka *Univariate results* na záložce *Summary*.

Effect	Univariate Results for Each DV (Spreadsheet22) Sigma-restricted parameterization Effective hypothesis decomposition Degr. of Freedom	PercArbAsin SS	PercArbAsin MS	PercArbAsin F	PercArbAsin p
Intercept	1	3,861037	3,861037	329,3251	0,000000
Mown	1	0,036437	0,036437	3,1079	0,099719
P	1	0,401459	0,401459	34,2422	0,000042
Mown*P	1	0,032183	0,032183	2,7451	0,119787
Error	14	0,164137	0,011724		
Total	17	0,637429			

Pouze aplikace fosfátu má průkazný vliv na mykorrhizní symbiózu, pomocí grafu lze ověřit, že ji snižuje.

Jak postupovat v programu R

Proměnné pro naše dva příklady jsou importovány samostatně do datových rámců *chap10a* a *chap10b*. Logaritmickou transformaci pro prvý příklad dosáhneme použitím funkce *log* v místě, kde bychom jinak zadávali název vysvětlované proměnné.

```
> summary(aov(log(SeedlSum)~Block+Treatment,data=chap10a))
            Df Sum Sq Mean Sq F value Pr(>F)
Block        3 0.1324  0.0441   0.413 0.7479
Treatment    3 1.4204  0.4735   4.427 0.0358 *
Residuals    9 0.9626  0.1070
```

Obdobně pro frekvence mykorrhizních hub v kořenech je zadání jednoduché (arcsinová funkce má v programu R jiný název):

```
> summary(aov(asin(sqrt(PercArb/100))~Mown*P,data=chap10b))
            Df Sum Sq Mean Sq F value   Pr(>F)
Mown         1 0.0569  0.0569   4.851   0.0449 *
P            1 0.3842  0.3842  32.773 5.25e-05 ***
Mown:P       1 0.0322  0.0322   2.745   0.1198
Residuals   14 0.1641  0.0117
```

Za povšimnutí stojí, že ačkoliv jsou výsledky testu pro interakci stejné jako v programu Statistica a stejný je i odhad residuální sumy čtverců, testy pro hlavní efekty se liší. Jde o důsledek skutečnosti, že programy Statistica a R používají odlišné způsoby rozkladu objasněné sumy čtverců na příspěvek jednotlivých efektů a tyto odhady se liší v případě nevyváženého uspořádání: v našich datech nemáme stejný počet opakování pro všechny kombinace faktorů a to vede ke korelaci mezi nimi a v důsledku k odlišnosti testů založených na odlišných metodách rozkladu sumy čtverců.

Postup, který je používán funkcí *summary* v programu R, je tzv. sekvenciální rozklad, kdy příspěvky jednotlivých hlavních efektů jsou počítány v pořadí, ve kterém jsou tyto efekty zadány do modelu (a výsledek proto závisí na pořadí, ve kterém efekty zadáváme). V našem případě vidíme, že hlavní efekt faktoru *Mown* má odhad sumy čtverců 0.0569, tj. zhruba o 0.02 vyšší než v programu Statistica, a naopak příspěvek druhého faktoru *P* je o stejný objem nižší.[x] Tento sekvenciální rozklad ale není pro data s nevyváženým uspořádáním příliš dobrý. Statistica má jako předvolbu rozklad sumy čtverců, který není příliš rozšířený, ale běžně jsou doporučovány rozklady označované jako *Type II Sums of Squares* nebo *Type III Sums of*

[x] Pokud jsou nezávislé proměnné spolu korelovány, nemůžeme pro část vysvětlené variability jednoznačně rozhodnout, které z korelovaných proměnných ji vysvětluje a záleží na metodě rozkladu sumy čtverců, jak jsou tyto „překryvy“ rozdělovány.

Squares (sekvenciální rozklad se označuje jako *Type I Sums of Squares*). Jejich použití závisí na typu testovaných hypotéz, ale pro náš příklad s neprůkaznou interakcí je vhodnější *Type II SS*. Pro jeho výpočet v programu R je nejjednodušší použít funkci *Anova* v knihovně *car*:

```
> library( car)
> Anova(aov(asin(sqrt(PercArb/100))~Mown*P,data=chap10b),type=2)
Anova Table (Type II tests)
Response: asin(sqrt(PercArb/100))
           Sum Sq Df F value    Pr(>F)
Mown      0.03073  1  2.6212    0.1277
P         0.38423  1 32.7730 5.252e-05 ***
Mown:P    0.03218  1  2.7451    0.1198
Residuals 0.16414 14
```

Výsledek není zcela shodný s výsledky v programu Statistica[*], ale závěry jsou obdobné (jen hlavní efekt fosfátu je průkazný).

Uvedené rozdíly ukazují, že i pro relativně základní metody, jako je dvoucestná analýza variance s neproporčním uspořádáním počtu pozorování, můžeme dostat různé výsledky v závislosti na použitém programu. V tomto případě je to proto, že Statistica a R mají různé předvolby (pokud bychom je sjednotili, výsledky by měly být stejné). Pro uživatele z toho vyplývají dvě skutečnosti – (1) i u relativně jednoduchých metod je třeba ve článcích uvádět, který jste použili software a (2) je užitečné vědět, které předvolby ten který program používá.

Popis analýz v článku

Methods

Seedling counts were log transformed [Frequency of mycorrhizal symbionts was arcsin-transformed] to meet the assumptions of ANOVA method [to achieve homogeneity of variances]. *Formulací* „to improve normality and homogeneity of variances/homoscedascity“ *pak můžeme vyjádřit situaci: „snažil jsem se, ale nezaručuji, že se to tak úplně povedlo“*.

Results

Ve výsledcích se provedenou transformací již nezabýváme.

Doporučená četba

Sokal R.R & Rohlf F.J. (1981) pp. 417-428, Zar (1984) pp. 236-243, Quinn & Keough (2002) pp. 64-67.

[*] V programu Statistica lze Type I, Type II a Type III SS také zvolit, na záložce *Options* prvního dialogového okna zobrazeného po výběru metody (tj. okna *ANOVA/MANOVA Factorial ANOVA*).

11 Hierarchická ANOVA, split-plot ANOVA a opakovaná měření

Hierarchická ANOVA

V těchto skriptech definujeme hierarchickou analýzu variance (*hierarchical ANOVA* nebo *nested ANOVA*) jako takový ANOVA model s více hierarchicky uspořádanými zdroji náhodné variability, ve kterém se faktory s pevným efektem buď nevyskytují nebo se vyskytují jen na jedné hierarchické úrovni. Modely, ve kterých jsou přítomny pevné efekty na dvou či více hierarchických úrovních označujeme jako split-plot analýzu variance a probíráme je v další sekci.

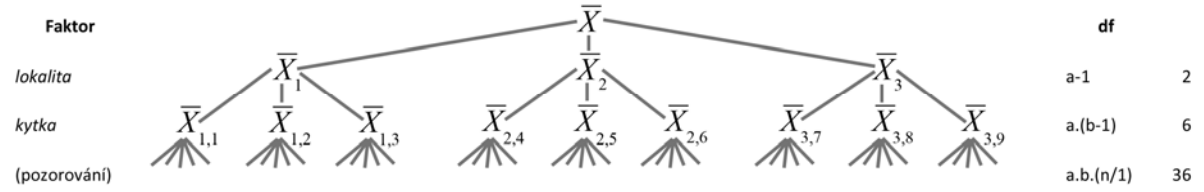
Příklady problémů pro hierarchickou analýzu variance:

1. Srovnáváme vliv typu hnojiva (tři různé typy, tj. tři hladiny faktoru) na složení tkání listů. V každé skupině máme pět rostlin a z každé rostliny odebereme náhodně čtyři terčíky a ty analyzujeme na obsah dusíku. Ptáme se, zda se liší obsah dusíku v listech při různých typech hnojení půdy. Máme tedy dva faktory - typ hnojiva (1 až 3) a identitu rostliny (rostlina 1 až 15). Tyto faktory nemají faktoriální uspořádání: v datech přítomné hladiny faktoru, představujícího identitu rostliny, se mezi typy hnojiva nepřekrývají, rostlina č. 7 může být kombinována jen s jedním typem hnojiva. Faktory jsou uspořádány hierarchicky, faktor rostliny je vnořen (*nested*) ve faktoru hnojiva. Alternativně také můžeme říct, že v tomto uspořádání jsou (na rozdíl od případu, kdy bychom z každé rostliny odebrali jen jeden terčík) dva zdroje náhodné variability – rozdíly mezi rostlinami pěstovanými se stejným typem hnojiva a rozdíly mezi vzorky (terčíky) odebranými na téže rostlině. Pokud jde o test vlivu typu hnojiva, nezávislými opakováními jsou jen jednotlivé rostliny, nikoliv jednotlivé terčíky.

V hierarchické ANOVě může být nejvyšší faktor buď s pevnými efekty (a tento faktor nás obvykle nejvíce zajímá, jako v našem příkladu 1) a nižší faktory jsou faktory s náhodnými efekty, nebo jsou všechny faktory s náhodnými efekty. Je možné mít i víc než dva faktory; v následujícím případě je faktorů více, všechny s náhodnými efekty.

2. Měříme délku trubky květní u hluchavky. Máme tři náhodně vybrané oblasti, v každé máme pět náhodně vybraných lokalit, z každé lokality čtyři náhodně vybrané rostliny a na každé rostlině měříme 7 náhodně vybraných květů. Zajímají nás rozdíly mezi oblastmi, mezi lokalitami v rámci oblasti, a mezi rostlinami z téže lokality. Chceme též porovnat variabilitu mezi skupinami na různých hierarchických úrovních.

Obrázek 11-1 ukazuje zjednodušenou variantu příkladu 2 (tak, aby se vešlo do obrázku). Porovnáváme zde 3 lokality, v každé byly zkoumány 3 rostliny a na každé bylo změřeno pět květů.



Obr. 11-1 Znázornění dat, sloužících jako příklad na hierarchickou analýzu variance.

Pro tento příklad označíme hierarchicky nejvyšší faktor *lokalita*, ten bude mít *a* úrovní, nižší faktor *kytka* bude mít *b* úrovní, *n* je počet opakování v nejnižší hierarchické jednotce (tj. počet květů měřených na každé rostlině). Zde má faktor *lokalita* tři hladiny, faktor *kytka* také tři hladiny, v každé základní skupině je pět opakování. Pro každou hladinu je v Obr. 11-1 udán počet stupňů volnosti. Součet čtverců je součtem čtverců odchylek průměrů na dané hierarchické úrovni od příslušného průměru hierarchicky nejblíže vyššího vážený (násobený) počtem všech pozorování v dané skupině. Např. pro faktor *kytka* je součet čtverců

$$(\bar{X}1 - \bar{X}11)^2 + (\bar{X}1 - \bar{X}12)^2 + (\bar{X}1 - \bar{X}13)^2 + (\bar{X}2 - \bar{X}21)^2 + (\bar{X}2 - \bar{X}22)^2 + (\bar{X}2 - \bar{X}23)^2 + (\bar{X}3 - \bar{X}31)^2 + (\bar{X}3 - \bar{X}32)^2 + (\bar{X}3 - \bar{X}33)^2$$

dále násobený pěti. Průměrný čtverec (MS) spočteme tak, že dělíme součet čtverců příslušným počtem stupňů volnosti.

Testovací kriterium F pro test nulové hypotézy, že hladiny faktoru dané hierarchické úrovně se neliší, vypočteme tak, že dělíme příslušný MS hierarchicky nejbližší (níže postavenou) hodnotou MS.

Zajímavým údajem je i to, jakým dílem přispívají jednotlivé hierarchické úrovně k celkové variabilitě. Tyto hodnoty – složky variance (*variance components*) se spočtou tak, že se od MS dané hierarchické úrovně odečte MS nejbližší nižší a vydělí se počtem pozorování v dané skupině. Pro faktor *kytka* je to 5 (květů na každé rostlině), pro faktor *lokalita* je to 15 (3 rostliny krát 5 květů). Faktor nejnižší hierarchické úrovně zůstává nezměněn. Pokud je MS pro vyšší hierarchický faktor menší než MS pro nejbližší nižší, potom se příspěvek tohoto vyššího hierarchického faktoru považuje za nulový. Hodnota příspěvku dané hierarchické úrovně k celkové variabilitě se často vyjadřuje v procentech.

I pro hierarchickou analýzu variance můžeme spočítat mnohonásobná porovnání pro hladiny hierarchicky nejvyššího faktoru, pokud se jedná o faktor s pevnými efekty. Jako odhad s^2 používáme *MS*, který jsme použili ve jmenovateli při výpočtu hodnoty *F* a jemu odpovídající počet stupňů volnosti.

Výše uvedené příklady ukazují dva časté typy užití hierarchické analýzy variance. V prvním případě (reprezentovaným příkladem 1) nás vpodstatě zajímá jen hierarchicky nejvyšší factor (který bývá faktorem s pevným efektem) a ostatní faktory jsou jen nutností, která odráží experimentální uspořádání. Je pravděpodobné, že rostliny se budou lišit svým obsahem dusíku, takže případný test významnosti rozdílů mezi rostlinami není tak zajímavý. Více terčíků z každé rostliny bereme proto, abychom lépe odhadli průměrný obsah dusíku v listech každé jednotlivé rostliny. Podobné uspořádání bývá běžné v terénních pokusech.

Například budeme sledovat vliv pastvy na diversitu společenstva, charakterizovanou počtem druhů na m^2 (např. porovnání pastvy skotu a ovcí). Máme zde praktická omezení: zvířata budou v experimentálních ohradách, ale aby se zvířata chovala přirozeně, nesmí být velikost ohrad příliš malá (řekněme, že potřebujeme hektarové ohrady). Je jasné, že pro každý typ pastvy musí být ohrady replikované (typ pastvy budeme testovat proti variabilitě jednotlivých ohrad), ale každou ohradu budeme mít reprezentovanou řadou čtverců, náhodně umístěných v rámci každé ohrady. Opět nás rozdíly mezi ohradami zajímají jen z toho hlediska, že proti variabilitě mezi ohradami se stejným způsobem pastvy budeme testovat vliv typů pastvy. Pro sílu testu na nejvyšší hierarchické úrovni (tedy efektu, který nás zajímá) je limitující počet opakování a variabilita na nejbližší nižší hierarchické úrovni – tedy počet ohrad každého typu, variabilita mezi ohradami. Síla testu ale dále poklesne, pokud jsou průměrné bohatosti v rámci každé ohrady odhadnuty s velkou chybou, k čemuž dojde v případě, že máme v každé ohradě málo čtverců na spočtení druhové bohatosti. Při plánování pokusu je třeba brat v úvahu, že udržování pastvy na hektarové ploše je podstatně

nákladnější, než stanovení počtu druhů na vybraných metrových čtvercích. Hierarchický rozklad variability, spolu se zohledněním nákladů na opakování na jednotlivých hierarchických úrovních nám umožní naplánovat experiment s maximální úsporností.

Druhý příklad reprezentuje typ užití hierarchické analýzy, kde nás zajímá více rozklad variability, prakticky přes všechny hierarchické úrovně, a většinou je spojen s popisnými, observačními daty. Takové problémy budeme studovat v taxonomických, případně biogeografických studiích.

Split-plot ANOVA

TODO

ANOVA pro opakovaná měření

TODO

Příkladová data

Příkladová data v prvých třech sloupcích listu *Chap11* představují výsledky následujícího pokusu. Byl studován vliv půdního typu (proměnná *Soil*, s variantami písčitá – *sandy* – a jílovitá půda – *clay*) na hmotnost semen rostliny. V každém květináči (faktor *Pot*) byly pěstovány 3 rostliny a údaj v jednom řádku proměnné *Seedweight* představuje průměrnou hmotnost semen konkrétní rostliny.

Další čtyři proměnné listu *Chap11* (ve sloupcích D až G) použijeme pro ilustraci výpočtu složek variance a pocházejí z následující studie. Standardní metodou byla studována společenstva bentosu v horských potocích. V každém standardním vzorku byl stanoven počet druhů. Byla porovnávána tři pohoří, v každém pohoří tři oblasti, v každé oblasti byly vybrány tři potoky, a z každého potoka byly získány tři vzorky. Vždy se jednalo o horské potoky (ve stejném pásu nadmořských výšek), oblasti, potoky a místa pro získání vzorků v potocích byly vybírány tak, aby je bylo možné považovat za náhodný výběr. Cílem bylo zjistit, jak je rozložena druhová bohatost na různých geografických škálách (z výsledků lze usuzovat např. na limitaci druhové bohatosti šiřitelností organismů).

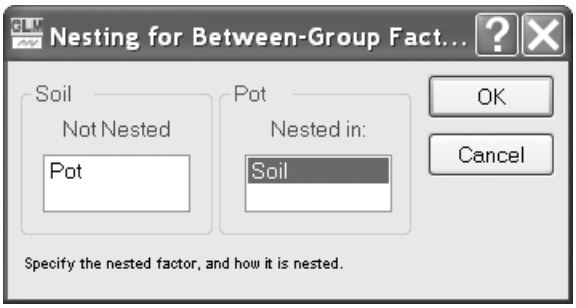
TODO – příkladová data pro split-plot a pro repeated-measures

Jak postupovat v programu Statistica

Hierarchická ANOVA

V programu Statistica musíme zadat model s hierarchickým uspořádáním faktorů pomocí pokročilých lineárních modelů. Z menu zvolíme příkaz *Advanced Linear/Nonlinear Models* | *General Linear Models* a z nabídnutého seznamu vybereme položku *Nested design ANOVA*. V dialogovém okně zadáme proměnné pomocí tlačítka *Variables*: *Seedweight* v levém seznamu (*Dependent variable list*), a proměnné *Soil* a *Pot* v druhém seznamu. Pak musíme kliknout na tlačítko *Between effects* a objeví se nové dialogové okno. Způsob, kterým Statistica nabízí výběr toho, který faktor je vnořen ve kterém, je poněkud matoucí. Při volbě musíme vycházet z toho, že po výběru proměnné v bílém políčku se změní text nad políčkem

z *Not Nested* na *Nested in*. Proto zvolíme proměnnou *Soil* v pravém políčku a dialogové okno pak vypadá následovně:

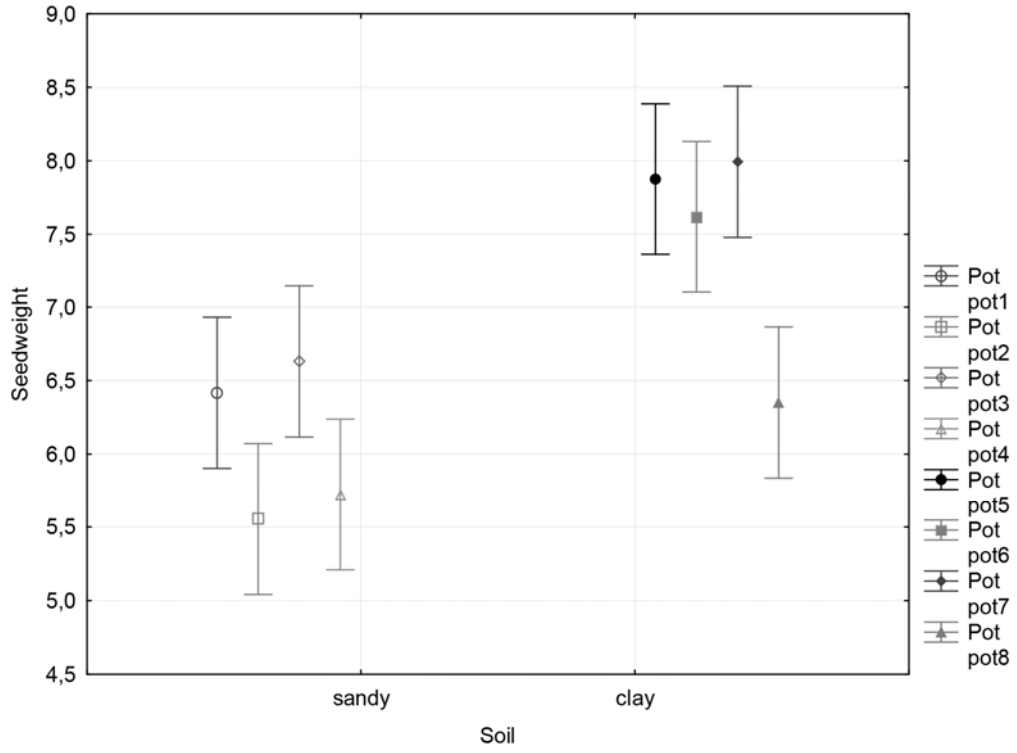


Po návratu do okna *GLM Nested Design ANOVA* přejdeme na záložku *Options* a pomocí tlačítka *Random factors* zadáme *Pot* jako faktor s náhodným efektem. Po volbě tlačítka *OK* se zobrazí výsledky analýzy variance a ANOVA tabulku můžeme zobrazit pomocí tlačítka *All effects*.

Univariate Tests of Significance for Seedweight (Spreadsheet30)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: ,4202826

Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixed	1099,990	1	1099,990	6	1,264422	869,9543	0,000000
Soil	Fixed	11,371	1	11,371	6	1,264422	8,9933	0,024043
Pot(Soil)	Random	7,587	6	1,264	16	0,176638	7,1583	0,000764
Error		2,826	16	0,177				

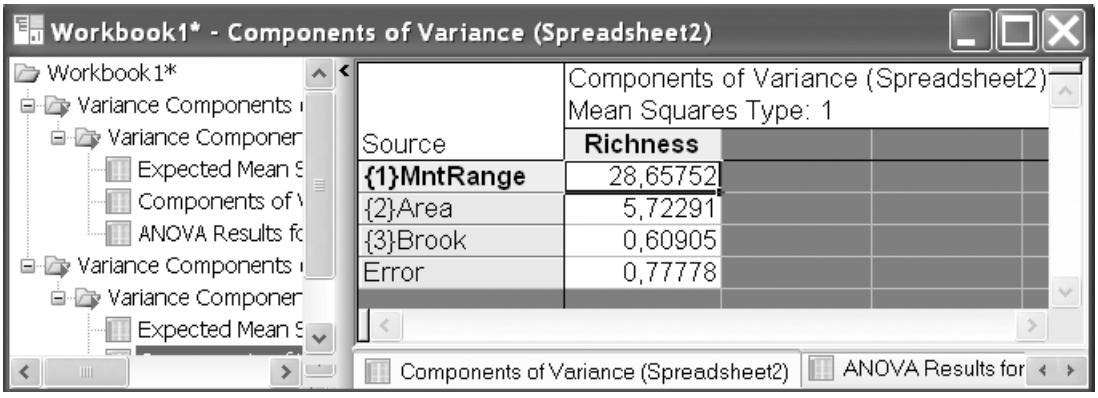
Všimněme si, že obsah tabulky neodpovídá běžné faktoriální analýze variance se dvěma faktory. To, že je efekt květináče vnořen do efektu půdního typu, znamená, že při výpočtu F statistiky používáme ve jmenovateli variabilitu (průměrný čtverec, MS) hmotnosti semen mezi květináči (tj. spočtenou z květináčových průměrů) s hodnotou 1.264, nikoliv variabilitu mezi rostlinami (v rámci květináče), kterou reprezentuje Error MS (0.177). Nezávislým pozorováním zde nejsou jednotlivé rostliny, ale trojice rostlin sdílejících květináč, tomu také odpovídá počet residuálních stupňů volnosti (6, protože máme 8 květináčů – 1 df pro odhad celkového průměru hmotnosti – 1 df pro efekt půdního typu). Alternativně lze tuto situaci popsat tak, že jednotlivé rostliny nemohou být nezávislými pozorováními, protože uspořádání experimentu neumožňovalo každou rostlinu náhodně a nezávisle na ostatních přiřadit jednomu ze dvou půdních typů. Výsledek můžeme doplnit informativním grafem porovnávajícím rozdíly mezi průměry hmotnosti semen pro jednotlivé květináče s rozdíly mezi půdními typy.



Tento graf lze vytvořit ze záložky *Quick* (nebo *Summary*) pomocí tlačítka *All effects/ Graphs* s následnou volbou položky *Pot(Soil)* pro vytvoření grafu.

Složky variance

Z menu zvolíme příkaz *Statistics* | *Advanced Linear/Nonlinear Models* | *Variance Components*. Pomocí tlačítka *Variables* zadáme v dialogovém okně proměnné (*Richness* ve sloupci *Dependent vars* a *MntRange*, *Area* a *Brook* ve sloupci *Random factors*). Na záložce *Model* pak zvolíme *Hierarchically nested design* a pak *Codes identify consecutive overall levels*. Po volbě tlačítka *OK* se objeví nové okno *Variance Components and Mixed Model ANOVA/ANCOVA Results*, kde zvolíme tlačítko *Summary: Components of variance*. Tím vytvoříme ve worksheetu troje výsledky, tabulka s odhady velikostí složek variance je ta prostřední. Z výsledků je jasné, že je zde velká variabilita mezi jednotlivými pohořími, zatímco variabilita mezi potoky v rámci oblasti je zcela zanedbatelná.



Poslední tabulka uvádí test odlišnosti složek variance (v základní populaci) od nuly. Všechny složky jsou průkazné, i když se jejích velikost zásadně liší.

Split-plot ANOVA

TODO

ANOVA s opakovanými měřeními

TODO

Jak postupovat v programu R

Hierarchická ANOVA

Tento jednoduchý typ hierarchické analýzy variance lze spočíst pomocí funkce *aov* za použití speciálního členu *Error* v zadání modelu. Tento člen specifikuje faktor, který představuje skupiny na sobě závislých pozorování (kterými jsou v našich datech trojice rostlin sdílejících květináč).

```
> summary(aov(Seedweight~Soil+Error(Pot),data=chap11a))
Error: Pot
          Df Sum Sq Mean Sq F value Pr(>F)
Soil       1 11.371  11.371   8.993  0.024 *
Residuals  6  7.587   1.264
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 16  2.826  0.1766
```

Řádky *Error: Pot* a *Error: Within* od sebe oddělují rozklad variability na dvou úrovních, oddělených faktorem zadáným pomocí členu *Error(Pot)*: mezikvětináčová variabilita, ze které je vysvětleno 11.371 faktorem *Soil* a 7.587 zůstává nevysvětleno, a variabilita mezi rostlinami v rámci květináčů (tedy variabilita hodnot kolem květináčových průměrů) o velikosti 2.826.

Velikost efektu půdního typu můžeme lehce získát z fitovaného ANOVA modelu pomocí funkce *coef*:

```
> coef(aov(Seedweight~Soil+Error(Pot),data=chap11a))
(Intercept) :
(Intercept)
       6.77

Pot :
Soilsandy
-1.376667

Within :
numeric(0)
```

Tento výsledek ukazuje, že průměrná hmotnost semen je u rostlin pěstovaných v písčité půdě o 1.377 mg nižší než u rostlin pěstovaných v jílovité půdě (kde je průměrná hmotnost 6.77 mg).

V programu R není při tomto zadání počítána průkaznost efektu květináče, ale můžeme ji spočíst ze zobrazených údajů – variabilitu mezi květináči nevysvětlenou půdním typem vyjadřuje řádek *Residuals* v sekci *Error: Pot*, zatímco variabilitu uvnitř květináčů vyjadřuje řádek *Residuals* v sekci *Error: Within*:

```
> 1-pf(1.264/0.1766,6,16)
[1] 0.0007644262
```

Tento test ale představuje testování efektu květináče coby faktoru s pevným efektem, správnější by ale bylo testovat jej jako náhodný efekt v rámci lineárních modelů se smíšenými efekty.

Složky variance

Data pro tento příklad byla importována do datového rámce *chap11b*. Pro odhad složek variance jsou již připraveny knihovny pro lineární modely se smíšenými efekty. Nejprve si postup ukážeme s knihovnou *nlme*:

```
> library(nlme)
> lme.1 <- lme(Richness~1,random=~1|MntRange/Area/Brook,data=chap11b)
> VarCorr(lme.1)
            Variance     StdDev
MntRange =  pdLogChol(1)
(Intercept) 28.6568539   5.3532097
Area =      pdLogChol(1)
(Intercept)  5.7231591   2.3923125
Brook =     pdLogChol(1)
(Intercept)  0.6090636   0.7804253
Residual     0.7777722   0.8819139
```

Hodnoty složek variance jsou čísla uváděná ve sloupci *Variance*: variabilita mezi pohořími (*MntRange*) je tedy 28.6568, mezi oblastmi (*Area)* je 5.7232, mezi potoky 0.6091 a v rámci potoků 0.7778. Pro vyjádření na procentické škále bychom museli jako základ použít součet těchto hodnot. Knihovna *nlme* pracuje primárně s odmocninami těchto variancí (tj. se směrodatnými odchylkami), jak je vidět i z výstupu z funkce *summary* (uvádíme jen jeho část):

```
> summary(lme.1)
Linear mixed-effects model fit by REML
 Data: chap11b
       AIC      BIC    logLik
  282.4709 294.3811 -136.2355

Random effects:
 Formula: ~1 | MntRange
        (Intercept)
StdDev:     5.35321

 Formula: ~1 | Area %in% MntRange
        (Intercept)
StdDev:    2.392313

 Formula: ~1 | Brook %in% Area %in% MntRange
        (Intercept)  Residual
StdDev:   0.7804253 0.8819139

       ...
```

Na škále směrodatných odchylek můžeme získat i intervaly spolehlivosti pro velikost spočtených složek variance, pro převod na škálu variancí bychom tedy museli hodnoty uváděné v části *Random Effects* umocnit:

```
> intervals(lme.1)
Approximate 95% confidence intervals

 Fixed effects:
              lower     est.    upper
(Intercept) 10.2819 16.69136 23.10081
attr(,"label")
[1] "Fixed effects:"
```

```
 Random Effects:
  Level: MntRange
                   lower    est.    upper
sd((Intercept)) 1.874621 5.35321 15.28674
  Level: Area
                   lower     est.    upper
sd((Intercept)) 1.319976 2.392313 4.335805
  Level: Brook
                    lower      est.    upper
sd((Intercept)) 0.4867092 0.7804253 1.251391

 Within-group standard error:
    lower      est.     upper
0.7304182 0.8819139 1.0648314
```

V knihovně *lme4* postupujeme podobně, zadání náhodných komponent modelu je ale odlišné:

```
> library(lme4)
> lmer.2 <- lmer(Richness~1+(1|MntRange)+(1|Area)+(1|Brook),data=chap11b)
> VarCorr(lmer.2)
 Groups   Name        Std.Dev.
 Brook    (Intercept) 0.78042
 Area     (Intercept) 2.39226
 MntRange (Intercept) 5.35328
 Residual             0.88192
```

A pro intervaly spolehlivosti pak (zde jsme již umocnili, takže jsou na škále variancí, poslední položka *Intercept* je ale jediný pevný efekt v modelu, průměr druhové bohatosti, zde umocněný, a tudíž nesmyslný):

```
> confint(lmer.2)^2
Computing profile confidence intervals ...
                  2.5 %     97.5 %
.sig01       0.2010616   1.550909
.sig02       2.0185579  23.887322
.sig03       2.7974516 176.271604
.sigma       0.5454127   1.163184
(Intercept) 88.4300254 574.993520
```

Split-plot ANOVA

TODO

ANOVA s opakovanými měřeními

TODO

Popis metod v článku

Methods

We have evaluted the effect of soil type on average seed weight using ANOVA with a random effect of pot.

nebo ... with the effect of pot nested within the effect of soil type.

TODO

Results

Soil type significantly affected the average weight of seeds ($F_{1,6}=8.99$, $p=0.024$), with the seeds being on average lighter by 1.38 mg on plants grown on sandy soils (see Figure X).

Obrázek X by obsahoval znázornění průměrů ve skupinách, nejspíše s konfidenčními intervaly

TODO

Doporučená četba

Sokal R.R & Rohlf F.J. (1981), pp. 271-320, Zar (1984), pp. 253-260, Quinn & Keough (2002): pp. 208-220 (hierarchická ANOVA), pp. 301-338 (split-plot a opakovaná měření).

12 Závislost dvou kvantitativních proměnných: regrese

Regrese a korelace

Doposud jsme porovnávali rozdíly mezi skupinami objektů. Tuto úlohu můžeme také popsat jako hledání závislosti jedné kvantitativní proměnné na proměnné kvalitativní; kvalitativní proměnnou je zde zařazení objektů do skupin. Nyní se budeme zabývat vztahem dvou kvantitativních proměnných. Nejčastěji půjde o dvě (nebo i více) spojitých proměnných na poměrné nebo intervalové stupnici.

Příklad 1: Měříme na řadě stromů průměr jejich kmene v prsní výšce a jejich výšku a studujeme závislost těchto dvou kvantitativních proměnných. Pravděpodobně zjistíme, že čím je větší průměr kmene, tím je větší výška.

Příklad 2: Studujeme závislost množství plovoucího opadu stromů (větve, klády) v jezerech, v závislosti na hustotě stromů na jeho pobřeží. Očekáváme, že se s densitou stromů bude zvětšovat množství dřevního materiálu a chceme popsat tento vztah..

Příklad 3: Měříme délku křídla u různě starých mláďat vrabce domácího. Vidíme, že délka křídla se stářím ptáka roste. Chceme popsat tuto závislost.

Ve všech uvedených příkladech se jedná o závislost dvou proměnných - a přece se uvedené příklady liší. Ve druhém a třetím příkladu jsme schopni říci, která charakteristika závisí na které: je smysluplné předpokládat, že množství větví a klád v jezeře závisí na denzitě stromů v okolí jezera a nikoliv naopak, a podobně že délka křídla závisí na věku, ale nikoliv, že věk závisí na délce křídla. Pokud chceme vyjádřit závislost jedné proměnné na jiné, metodický postup se nazývá **regrese** (*regression*); předpokládáme při ní, že nezávislá proměnná (*independent variable*) je změřena přesně, zatímco závislá proměnná (*dependent variable*) je zatížena náhodnou variabilitou. Naproti tomu v prvním případě není zřejmé, že by jedna z proměnných byla závislá na druhé, nemůžeme označit jednu proměnnou za nezávislou a druhou za závislou; obě proměnné jsou zřejmě zatíženy určitou náhodnou variabilitou. Pro vyjádření závislosti v těchto případech užíváme **korelaci** (*correlation*).

Jak ale uvidíme v praxi, regrese je často počítána i tehdy, když je „nezávislá“ proměnná také zatížena chybou (tam tomu bude určitě v našich příkladech 1 a 2 a nejspíš také v příkladu 3) a není zřejmý kauzální vztah jedné proměnné k druhé, zvláště díváme-li se na statistiku jako na exploratorní analýzu dat. Např. většina biologů by s čistým svědomím spočetla regresi výšky na průměru kmene, přestože zde není možné pokládat jednu proměnnou za závislou na druhé a náhodná variabilita obou proměnných bude přibližně stejná (tato závislost se často počítá a nazývá se alometrickým vztahem). Často se potom hovoří o **vysvětlující proměnné** (*explanatory variable*) a **vysvětlované proměnné** (*explained variable*) místo o proměnné nezávislé a závislé. Vysvětlovanou proměnnou někdy též nazýváme odpovědí (*response variable*), vysvětlující proměnná prediktorem (*predictor*). V případě, že X je náhodná proměnná, chápeme regresi *Y* na *X* jako studium závislosti odpovědi na zjištěných hodnotách vysvětlující proměnné. Odhady parametrů regresního modelu nejsou příliš dobré, pokud je náhodná variabilita proměnné X podobná nebo větší než náhodná variabilita Y (viz sekci *Nezávislá proměnná s náhodnou variabilitou* ke konci této kapitoly).

K tomu je třeba poznamenat, že korelační koeficient je průkazně odlišný od nuly právě tehdy, když je průkazná lineární regrese jedné z uvažovaných proměnných na druhé.

Jednoduchá lineární regrese

Nejjednodušší typ regrese se označuje jako jednoduchá lineární regrese (*simple linear regression* nebo *bivariate regression*). Jednoduchá zde znamená, že máme jen jednu nezávislou (vysvětlující) proměnnou. Obecně v regresi může být nezávislých proměnných několik. Lineární znamená, že závislost můžeme vyjádřit přímkou. Příklad dat a jejich vynesení podává Tab 12-1 a Obr. 12-1. Nezávislou (vysvětlující) proměnnou značíme obvykle *X* a vynášíme ji na vodorovnou osu (*abscissa*), závislá (vysvětlovaná) proměnná bývá značena *Y* a vynáší se na svislou osu (*ordinate*). Rovnice přímky potom je

$$EY = \beta_0 + \beta_1 X$$

Vz. 12-1

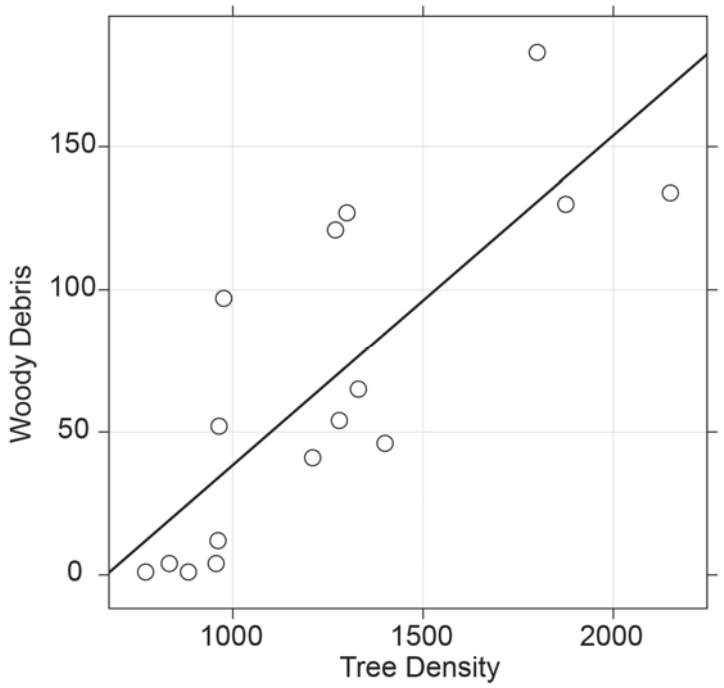
EY je očekávaná (*expected*) hodnota závislé proměnné, β_1 je směrnice této přímky (často zvaná sklon, *slope*, je to tangent úhlu, který svírá regresní přímka s vodorovnou osou) a β_0 je hodnota *Y* v případě kdy *X=0*, tedy souřadnice průsečíku přímky s osou y (*intercept*). Připomeňme, že β_1 udává, o kolik jednotek se změní hodnota *Y* při zvětšení hodnoty *X* o jednu jednotku, a závisí tedy na tom, v jakých jednotkách měřím obě proměnné, je udána v jednotkách (jednotky Y . [jednotky X]$^{-1}$). β_0 je v jednotkách proměnné *Y*.

Tab. 12-1 Závislost množství dřevního opadu (plovoucí větve a klády) na hustotě stromů na pobřeží (do vzdálenosti 50 m od pobřežní čáry).

X: Hustota stromů [počet / km pobřeží]	**Y: Množství dřevní hmoty [m^2 / km pobřeží]**
1270	121
1210	41
1800	183
1875	130
1300	127
2150	134
1330	65
964	52
961	12
1400	46
1280	54
976	97
771	1
833	4
883	1
956	4

Otázkou je, jak vybrat z různých možných přímek, které vynesené body prokládají, tu nejlepší. Možností je několik, ale ve statistice se běžně užívá kriterium nejmenších čtverců

(*least squares*).[*] Označme X_i a Y_i zjištěné hodnoty nezávislé a závislé proměnné, a $\hat{Y}_i$ je hodnota závislé proměnné, predikovaná podle rovnice ve Vz. 12-1. Potom za nejlepší prohlásíme takovou přímku, pro kterou platí, že $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ je minimální (n je celkový počet pozorování; v dalším textu budeme pro jednoduchost vynechávat rozsah sumace: vždy předpokládáme od jedné do n). Tato hodnota se nazývá reziduální součet čtverců (*residual sum of squares* - RSS, také *error sum of squares*). Kritérium nejmenších čtverců by se tedy mělo přesně nazývat kriteriem nejmenšího součtu reziduálních čtverců.



Obr. 12-1 Závislost množství dřevního opadu (plovoucí větve a klády) na hustotě stromů na pobřeží

Určit přesně hodnoty β_0 a β_1 by znamenalo znát všechny hodnoty základního souboru. My z něj ale většinou známe jen výběr: nezajímá nás, jak závisí množství dřevního opadu na hustotě stromů u 16 zkoumaných jezer, ale jak závisí u - potenciálně mnohem většího - souboru všech jezer podobného charakteru). Proto parametry základního souboru odhadujeme na základě výběru. Odhad parametru β_0 značíme b_0, odhad parametru β_1 značíme b_1. Lze ukázat, že regresní přímka vždy prochází bodem $\left[\overline{X},\overline{Y}\right]$. Směrnici regresní přímky vypočteme

$$b_1 = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2}$$

Vz. 12-2

Pro vlastní výpočty se užívá tzv. výpočetní tvar, který dává totožné výsledky, ale je jednodušší pro vlastní užití.[x] Parametr b_0 v praxi odhadujeme tak, že do rovnice ve Vz. 12-1 dosadíme $\left[\overline{X},\overline{Y}\right]$ (víme, že přímka musí procházet tímto bodem) a dostáváme

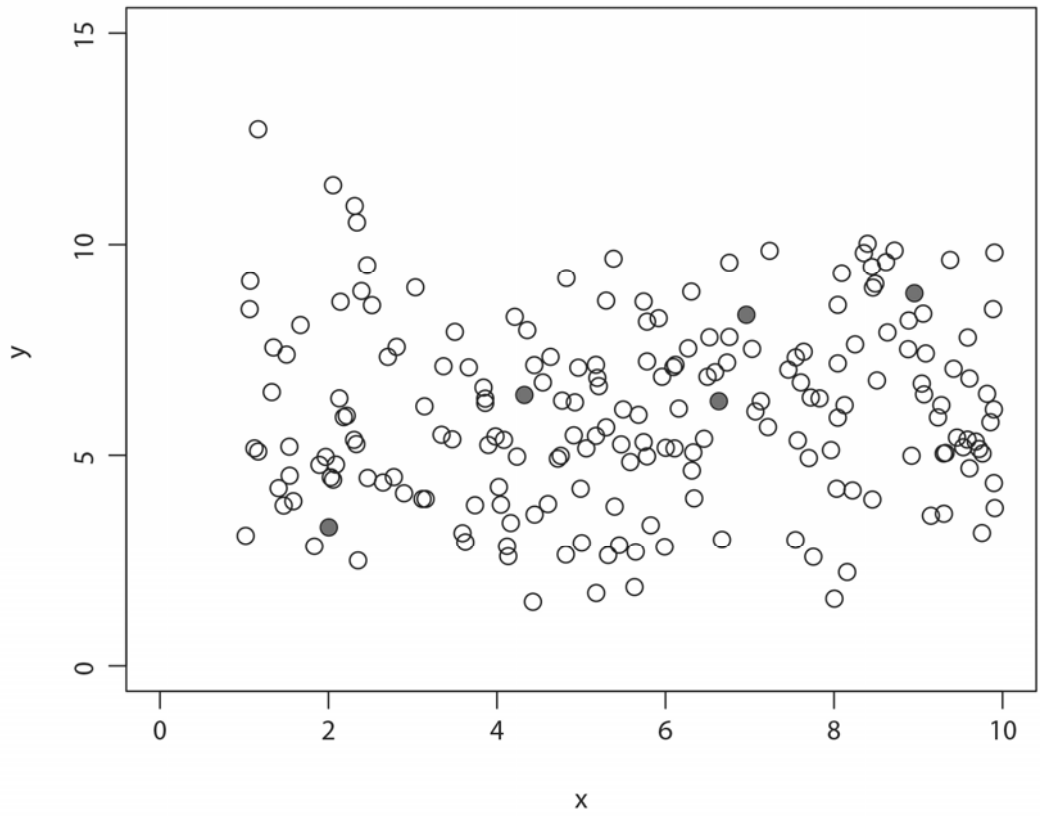
[*] Obecnějším často užívaným kriteriem je kriterium největší věrohodnosti (*maximum likelihood*); pro lineární regresi jsou jeho odhady parametrů totožné s kriteriem nejmenších čtverců.

[x] Odvození odhadu vychází z toho, že do součtu čtverců dosadíme za $\hat{Y}$ z rovnice 12-1. Výsledný výraz derivujeme podle β_1 i podle β_0, oba získané výrazy položíme rovny nule (hledáme lokální minimum, proto musí být derivace rovna nule). Tím dostáváme soustavu dvou rovnic o dvou neznámých, kterou vyřešíme.

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Vz. 12-3

Hodnota b_1 může být kladná i záporná. Hodnota ve jmenovateli Vz. 12-2 je vždy kladná. Pokud jsou kladné odchylky Y od průměru sdruženy s kladnými odchylkami X, je hodnota v čitateli také kladná a b_1 je kladné, v opačném případě je b_1 záporné.



Obr. 12-2 Hypotetický základní soubor dat, s regresním koeficientem β_1 rovným nule. Šedě vyplněné body mohou být možným výběrem pěti pozorování a takovýto výběr povede k odhadu b_1 průkazně odlišnému od nuly.

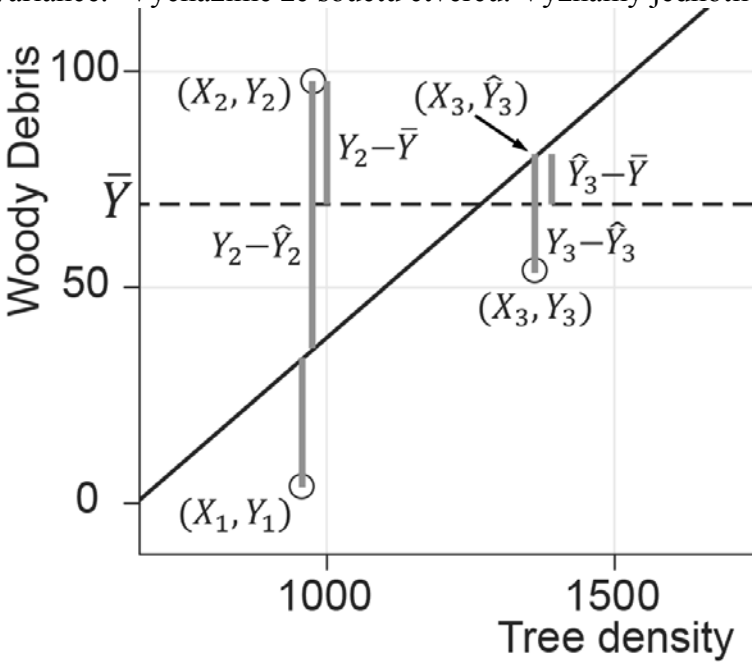
Již jsme konstatovali, že odhady parametrů jsou zatíženy určitou chybou. Jsou to tedy náhodné proměnné. Proto chceme většinou znát, jak velká tato chyba může být: testujeme některé hypotézy o parametrech, případně o celém **regresním modelu** (který nám představuje regresní rovnice například ve Vzorci 12-1). Tento předpoklad je obvykle těžké splnit, obvykle se spokojujeme s konstatováním, že chyba X je podstatně menší než chyba Y - potom je metoda vcelku robustní. Dále předpokládáme, že pro každý bod X má proměnná Y normální rozdělení a že variance tohoto rozdělení není závislá na hodnotě X (to je obdoba požadavku homogenity variancí v metodách analýzy variance). Alternativně můžeme tuto skutečnost vyjádřit tak, že závislost popíšeme modelem $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, kde ε_i je náhodná proměnná s normálním rozdělením a nulovou střední hodnotou, nezávislá na hodnotě X. Samozřejmě by mělo také platit, že jednotlivé měřené objekty byly vybrány náhodně a na sobě nezávisle!

Testy významnosti

Teoreticky se může stát, že sledované veličiny jsou v základním souboru nezávislé, ale do výběru se náhodou dostanou ty, které určitou závislost vykazují: příklad ukazuje Obr. 12-2. Proto se ptáme, jaká je pravděpodobnost, že závislost dané síly najdeme jako důsledek

náhody při výběru. Testujeme tedy nulovou hypotézu: v základním souboru nezávisí hodnota závislé proměnné na hodnotě nezávislé proměnné.

Pro test významnosti regrese používáme nejčastěji rozklad sumy čtverců na část vysvětlenou a část residuální, které porovnáváme pomocí F statistiky, podobně jako v modelech analýzy variance.[x] Vycházíme ze součtů čtverců. Významy jednotlivých rozdílů ukazuje Obr. 12-3.



Obr. 12-3 Zvětšená část Obr. 12-1, ukazující komponenty odchylek hodnot Y

Celkový součet čtverců (*total sum of squares*)

$$SS_{TOT} = \sum \left(Y_i - \bar{Y} \right)^2$$

Vz. 12-4

popisuje celkovou variabilitu závislé proměnné. Regresní součet čtverců (*regression sum of squares,* též modelový součet čtverců, *model sum of squares*)

$$SS_{REG} = \sum \left(\hat{Y}_i - \bar{Y} \right)^2$$

Vz. 12-5

odpovídá variabilitě vysvětlené regresním modelem a reziduální součet čtverců (*residual sum of squares*)

$$SSe = \sum \left(Y_i - \hat{Y}_i \right)^2$$

Vz. 12-6

odpovídá variabilitě modelem nevysvětlené. Platí

[x] Protože se rozklad celkové sumy čtverců používá jak v modelech analýzy variance (kde jsou nezávislými proměnnými faktory), tak v lineární regresi (jde jsou nezávislými proměnnými kvantitativní proměnné), v terminologii statistické analýzy vznikl zmatek, protože termín ANOVA (analýza variance) se používá ve dvou významech (tj, jako model analýzy variance a jako rozklad variability pro téměř jakýkoliv statistický model). V kontextu lineární (či jiné) regresní analýzy doporučujeme používat termín ANOVA jen v jednoznačných kombinacích, jako třeba ANOVA tabulka regresního modelu, odpovídající testu průkaznosti, který zde popisujeme.

$$SS_{TOT} = SS_{REG} + SSe$$

Vz. 12-7

čehož se využívá při výpočtech. Odpovídající stupně volnosti jsou

$$DF_{TOT} = n-1,$$

Vz. 12-8

$$DF_{REGR} = \text{počet odhadovaných parametrů} - 1 = 1$$

Vz. 12-9

a

$$DF_e = DF_{TOT} - DF_{REGR} = n-2.$$

Vz. 12-10

Hodnoty průměrného čtverce (*MS*, *mean square*) získáme opět jako podíl součtu čtverců a příslušného počtu stupňů volnosti. Opět platí aditivita součtu čtverců i aditivita počtu stupňů volnosti. MS_e ($=SS_e/DF_e$)je odhadem společné variance („rozptylu pozorování kolem regresní přímky“). Lze ukázat, že pokud platí nulová hypotéza o nezávislosti, všechny tři hodnoty *MS* jsou odhadem variance proměnné Y. Proto lze poměr

$$F = \frac{MS_{REGR}}{MS_e}$$

Vz. 12-11

považovat za testové kriterium významnosti regrese. Na základě součtu čtverců se počítá i koeficient determinace (*coefficient of determination*)

$$R^2 = \frac{SS_{REGR}}{SS_{TOT}}$$

Vz. 12-12

který udává podíl vysvětlené variability. Jiným způsobem testování je testování parametrů regresního modelu (koeficientů β_i) pomocí Studentova *t*. Obecně platí, že

$$t = \frac{\text{(odhad parametru-hypotetická hodnota parametru)}}{\text{střední chyba odhadu parametru}}$$

Vz. 12-13

To, co automaticky tiskne většina programů, je test nulové hypotézy, že hodnota parametru se rovná nule. V případě b_1 je dosažená hladina významnosti totožná s dosaženou hladinou významnosti u analýzy variance. To je logické: pokud $\beta_1=0$, je regresní přímka vodorovná a hodnoty závislé proměnné nejsou ovlivněny hodnotami nezávislé proměnné. Na rozdíl od F testu pro regresní model můžeme ale t-test použít i pro jednostranný test. V případě parametru b_0 test pro hypotézu $\beta_0=0$ většinou nemá biologický smysl: testujeme nulovou hypotézu, že přímka prochází počátkem, a tedy že hodnota závislé proměnné je nulová pro nulové hodnoty nezávislé proměnné.

Většina biologických závislostí zřejmě není lineární; přesto se lineární regrese hojně užívá: jednak lze při dostatečně malém rozsahu hodnot nezávislé proměnné každou funkci aproximovat přímkou (otázkou zůstává, co budeme považovat za dostatečně malý rozsah),

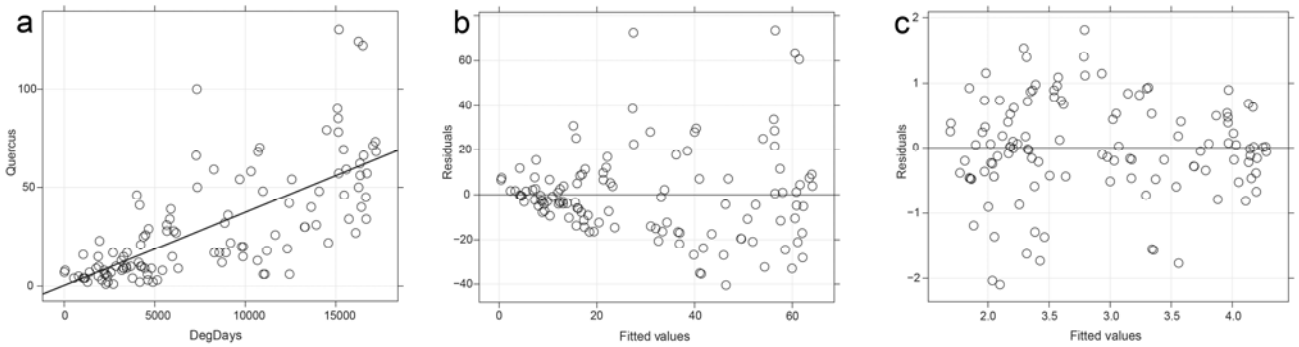
jednak v případě, že nemáme apriorní představu, jak má model funkční závislosti vypadat, snažíme se užít nejjednodušší možný model - a tím je model lineární odpovědi.

Konfidenční a predikční intervaly

Konfidenční interval (interval spolehlivosti) pro střední hodnotu rozdělení proměnné *Y* při dané hodnotě *X*, zvaný obvykle *confidence band* nebo *confidence limits* je definován tak, že střední hodnota leží v intervalu s danou pravděpodobností. Naproti tomu v predikčním intervalu (*prediction interval, prediction limits*) se nachází s udanou pravděpodobností libovolné měření. Predikční interval je tedy (při stejné zadané pravděpodobnosti) výrazně širší než konfidenční. Konfidenční interval se stále zužuje při vzrůstu počtu pozorování, predikční interval se pouze přibližuje určité mezi, za kterou už nemůže klesnout. Oba intervaly jsou nejužší pro hodnoty kolem průměru nezávislé (vysvětlující) proměnné X.

Transformace dat v regresi

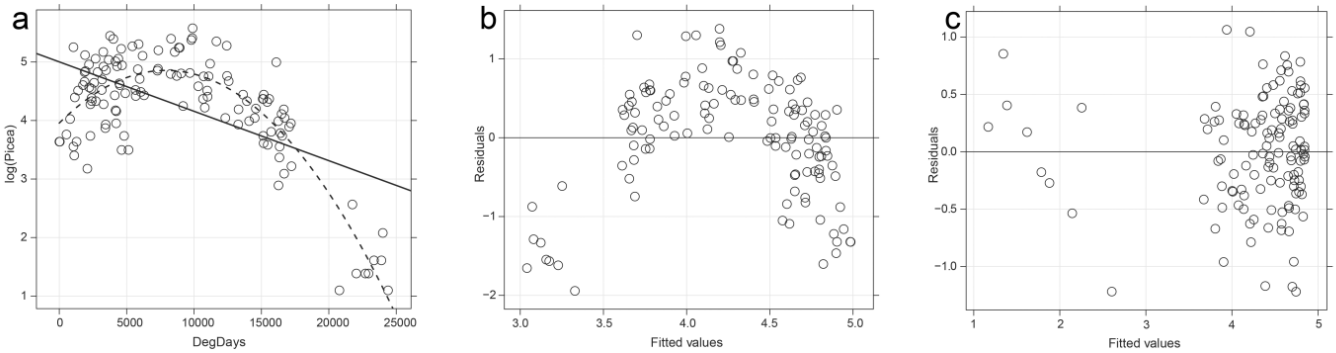
Jak poznáme, zda jsou splněny předpoklady regrese (nejčastěji linearita vztahu a nezávislost variance na hodnotě nezávislé proměnné)? První možnost je z pouhého pohledu na vynesenou závislost. Přesnější a názornější je vynést regresní reziduály (tj. pozorovaná minus predikovaná hodnota: $Y_i - \hat{Y}_i$) proti predikované hodnotě $\hat{Y}_i$, jak ukazují Obr. 12-4 a Obr. 12-5. V ideálním případě by reziduály měly ležet v pásu kolem osy X, jak ukazuje např.Obr. 12-4c.*



Obr. 12-4 Regresní model, u kterého data nesplňují předpoklad nezávislosti variance na fitované hodnotě (část a). Po vynesení reziduálů proti predikované hodnotě (část b) vidíme jejich zvětšující se rozptyl pro vyšší predikované hodnoty Závislá proměnná by měla být transformována. Část c ukazuje stejný typ diagramu jako část b, ale po logaritmické transformaci závislé proměnné.

Dva nejběžnější typy narušení předpokladů ukazují Obr. 12-4 a Obr. 12-5. V případě Obr. 12-5 je závislost nelineární. V případě Obr. 12-4 je funkce sice lineární, reziduály jsou symetricky rozloženy kolem osy X, ale jejich absolutní hodnota stoupá s predikovanou hodnotou; v některých případech naopak zjistíme, že jejich absolutní velikost klesá podél horizontální osy. Vyneseme-li absolutní hodnoty reziduálů proti hodnotám X, případná regrese vyjde průkazná (někdy se podobně vynášejí čtverce reziduálů). Tento případ je poměrně běžný, u mnoha proměnných je stálý spíše variační koeficient než sama variance.

* Připomeňme, že pokud bychom počítali přímku regresní závislosti reziduálu na hodnotě nezávisle proměnné, musíme vždy dostat osu X, tj. $b_0=0$ i $b_1=0$.



Obr. 12-5 Regresní model, který je špatně specifikován (přímka v části a). Reziduály ukazují (část b), že závislost není lineární, spíše je kvadratická (kvadratický model, užívající polynom druhého stupně pro nezávislou proměnnou, je zobrazen v části a čárkovanou čárou. V části c jsou vyneseny regresní reziduály a predikované hodnoty pro tento kvadratický model. Oba zde zobrazené regresní modely používají logaritmované hodnoty závislé proměnné (viz Obr. 12-4)..

Zkontrolovat reziduály je asi nejjednodušším a nejběžnějším způsobem, jak si ověřit vhodnost použitého modelu. V současné době se ovšem užívají i další metody **regresní diagnostiky.** K nejběžnějším patří sledování vlivu jednotlivých bodů v regresi (*leverage*). Největší váhu mají odlehlé body (*outliers*). Takováto inspekce nám může mnohé napovědět, zvláště máme-li o bodech další informace. Sledujeme-li např. závislost počtu druhů na velikosti ostrova a zjistíme-li, že Trinidad má největší vliv (tzn. je „nejodlehlejším" bodem), budeme uvažovat proč. Zjistíme-li, že jsou „vlivné" body nahloučeny, např. pro extrémní hodnoty nezávislé proměnné, je zřejmé, že jsme použili nevhodný model. Někdy takový přístup může ukázat i na chyby v datech. Některé programy umožňují vylučování takových bodů z regrese. Vylučování bodů z regrese považujeme za nebezpečný přístup (zvláště bychom si po něm nedovolili provádět testy významnosti), pokud k němu nemáme ještě jiný a pádný důvod než odlehlost bodu (např. zjistím-li, že se velmi výrazně liší jediné pozorování, a to pozorování, provedené určitou osobou 1. ledna ve dvě hodiny ráno). Naproti tomu výrazně doporučujeme provádět regresní diagnostiku modelů.

Vhodnou transformací dat můžeme některé odchylky od předpokladů napravit. Pro transformace dat v regresi platí obdobná pravidla (i obdobná varování) jako pro transformace v analýze variance. Transformace kterékoliv proměnné ovšem změní tvar závislosti - v některých případech tak, jak potřebujeme, ale v jiných naopak může narušit lineární charakter závislosti. Pokud předpokládáme logaritmický tvar závislosti, tj $Y = b_0 + b_1 \log(X)$, a nezávislost reziduálů na predikované hodnotě, stačí provést logaritmickou transformaci nezávislé proměnné a dostáváme lineární vztah. Užít můžeme přirozené i dekadické logaritmy. Často se užívá logaritmická transformace (nebo $\log(X+1)$) závislé proměnné. Pokud je v původních datech směrodatná odchylka lineárně závislá na průměru a data mají spíše logaritmicko-normální než normální rozdělení a závislost má exponenciální tvar, tj.

$$Y = e^{b_0 + b_1 X} = e^{b_0} e^{b_1 X}$$

Vz. 12-14

potom po logaritmické transformaci dostáváme lineární závislost se stálou variancí a normálním rozdělením dat. Logaritmováním Vz. 12-14 dostáváme

$$\ln Y = b_0 + b_1 X$$

Vz. 12-15

Typicky tak můžeme např. odhadnout růstovou rychlost exponenciálně rostoucí populace. Ta je charakterizována rovnicí $N_t = N_0\ e^{rt}$, kde N značí velikost populace v časech daných indexem, r je růstová rychlost a t je čas. Logaritmickou transformací (musíme užít přirozený logaritmus) dostáváme $\ln(N_t) = \ln(N_0) + rt$. Sklon přímky v regresi přirozeného logaritmu na čase nám tedy dává přímo odhad růstové rychlosti.

Pokud platí lineární závislost směrodatné odchylky od průměru a lognormalita rozdělení a závislost má tvar

$$Y = b_0 x^{b_1}$$

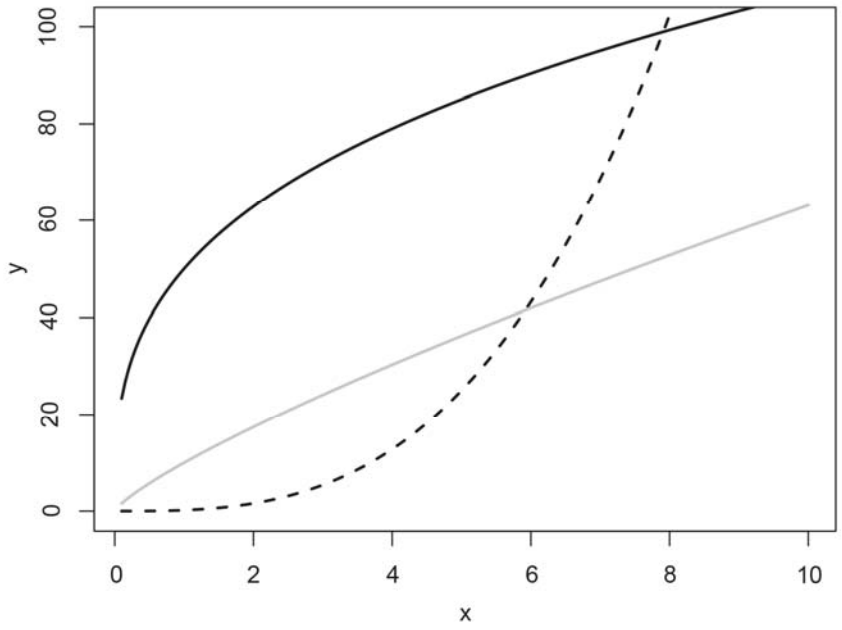
Vz. 12-16

doporučuje se provést logaritmickou transformaci obou proměnných. Logaritmováním Vz. 12-16 dostáváme

$$\ln Y = \ln b_0 + b_1 \ln X$$

Vz. 12-17

Běžně se v tomto případě mluví o log-log transformaci. Ta je schopna „zlinearizovat" většinu závislostí, které jsou monotónní, nemají inflexní bod a procházejí počátkem. Možné tvary křivek ukazuje Obr. 12-6. Je známo, že takový tvar má většinou závislost počtu druhů na ploše (tzv. *species-area curve*), nebo alometrické závislosti (např. závislost objemu dřeva stromu na průměru kmene). V případě *species-area* závislosti se tradičně uvádí tvar závislosti $S = c.A^z$, kde S je počet druhů, A je velikost sledované plochy a c a z jsou parametry odhadované regresní analýzou. V té závislost zlogaritmujeme, a dostáváme $\log(S) = \log(c) + z \log(A)$ – tedy počítáme regresi logaritmu počtu druhů na logaritmu plochy.



Obr. 12-6 Různé křivky tvaru $y=b_0 x^{b1}$

Těmto postupům se někdy říká linearizovaná regrese. Zde je třeba si uvědomit, že transformaci nezávislé proměnné tak, aby výsledná závislost byla lineární, můžeme provádět dle libosti. Nezávislá proměnná není zatížena náhodnou variabilitou (je *error free*). Naproti

tomu transformace závislé proměnné mění jak tvar závislosti, tak typ rozdělení dat a stálost variance. Logaritmická transformace závislé proměnné může zlepšit stálost variance, byla-li směrodatná odchylka lineárně závislá na průměru; pokud ovšem byly předpoklady lineární regrese splněny u netransformovaných dat, nebudou splněny u transformovaných. Parametry odhadujeme metodou nejmenších čtverců. Minimalizujeme tedy reziduální součet a zároveň platí, že součet všech odchylek je roven nule. V linearizované regresi to ovšem platí pro transformovaná data, nikoliv už pro data po „odtransformování". Byla-li např. použita logaritmická transformace, je po odtransformování součet kladných reziduálů vyšší než součet záporných: odhad je tedy do jisté míry vychýlený.

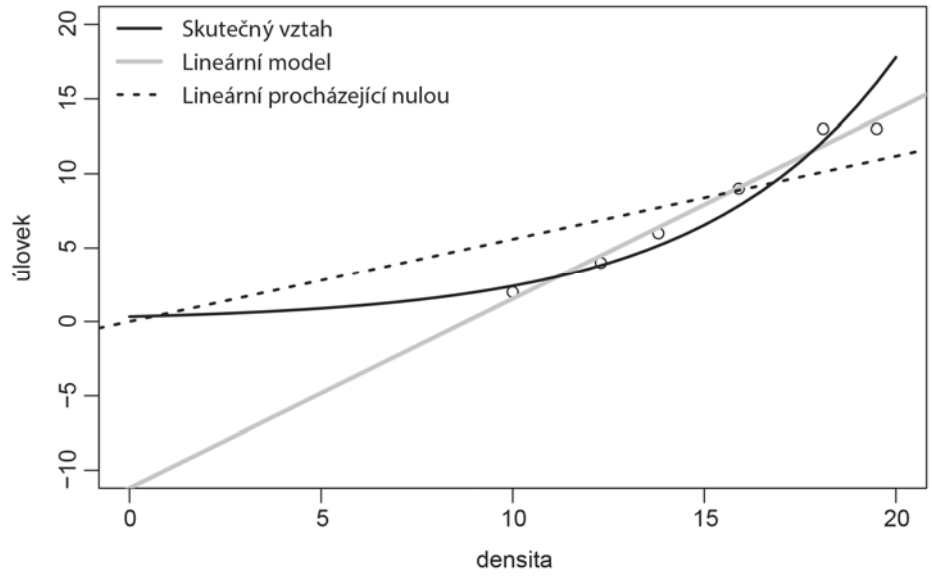
Pokud jsou podmínky kromě linearity splněny u netransformovaných dat a transformace nezávislé proměnné nepomáhá (logaritmická transformace nemůže pomoci např, při nemonotónních závislostech), je možné použít polynomiální nebo nelineární regresi. Lepší alternativou linearizované regrese je užití metod zobecněných lineárních modelů (viz kapitola XX). Ty umožňují velkou škálu funkčních závislostí a odhadují parametry modelu metodou maximální věrohodnosti (tedy nikoliv metodou nejmenších čtverců). Využívají při tom apriorní informaci o podobě statistického rozdělení dat. Tyto metody z větší části již nahradily linearizovanou regresi a časem ji pravděpodobně nahradí zcela.

Regrese procházející počátkem

V některých případech víme předem, že by regresní přímka měla procházet počátkem. Např. studujeme jak závisí počet hrabošů ulovených poštolkou na hustotě populace hraboše (tzv. funkcionální odezva, *functional response*) a předpokládáme odezvu prvního typu, tj. lineární vztah pro nízké hustoty. Přitom víme, že závislost musí procházet počátkem (při nulové hustotě se nic nedá ulovit). Potom můžeme použít regresi procházející počátkem, tj. předpokládáme závislost $Y = b_1 X$. Vzorce uvádí Zar (1984, p. 284).

Použití regrese procházející počátkem by ovšem nemělo být aplikováno v podobných případech automaticky. U poštolky se ale může jednat i o tzv. odezvu třetího typu, kdy při nízkých hodnotách neuloví nic (dravec začně registrovat typ stravy, ažkdyž přesáhne určitou hustotu). Tvar závislosti je na Obr. 12-7. V naší krajině se ovšem území s tak nízkou hustotou hrabošů nevyskytují – studujeme potom jen určitý rozsah hodnot hustoty hrabošů. Závislost může být v daném rozsahu lineární. Ovšem extrapolace pro hodnoty populace hrabošů blízké nule bude předpovídat záporný počet ulovených hrabošů. Pokud bychom v daném případě „donutili" regresní závislost procházet počátkem, dostaneme velmi nerealistickou závislost (viz Obr. 12-7). Je to tím, že závislost je zřejmě lineární ve sledovaném rozsahu, ale není lineární, pokud bychom extrapolovali mimo tento studovaný rozsah.

Pokud je závislost ve studovaném rozsahu lineární, je správní¨ým řešením lineární regrtese, s tím, že jednoznačně víme, že danou závislost můžeme užít jen v rozsahu hodnot prediktoru, se kterým jsme pracovali, a vyvarujeme se jakýchkoliv extrapolací. Zcela obecně je si třeba uvědomit, že lineární závislosti nejsou v přírodě pravidlem, ale lineární závislostí můžeme často v určitém omezeném rozsahu prediktoru závislost aproximovat. Pokud ovšem budeme v takovém případě užívat extrapolaci, můžeme dojít ke zcela nesmyslným výsledkům. Zvlášť u proměnných na poměrové stupnici si musíme být navíc vědomi toho, že v blízkosti nuly se závislost bude chovat jinak, než ve zbytku rozsahu hodnot prediktoru. Z tohoto hlediska je třeba se dívat i na testy koeficientu b_0 . Pokud je prediktor proměnná na poměrové stupnici, a její studovaný rozsah je ale nule vzdálen, potom nemá smysl testovat, za přímka prochází počátkem, protože test je založen na předpokladu, že závislost je lineární v celém rozsahu, a tento předpoklad nemáme jak ověřit.



Obr. 12-7 Závislost mezi počtem ulovených hrabošů a jejich nabídkou (hustotou populace). Pokud aproximujeme skutečnou závislost (plná černá čára) lineárním modelem (plná šedá čára), predikce extrapolované směrem k nízkým hustotám nejsou realistické. Použití lineárního modelu procházejícího nulou (čárkovaná černá čára) ale není řešením, výsledný model nepopisuje dobře sebraná data.

Síla testu

Ani u regrese není možné nezamítnutí nulové hypotézy přímo interpretovat jako důkaz nezávislosti Y na X. Musíme vždy uvažovat o síle testu. Jako u jiných testů, při dané hladině významnosti síla stoupá především s počtem pozorování a samozřejmě s těsností závislosti (vyjádřenou R^2). Přesnější určení síly je stejné jako pro korelaci a bude probráno při korelaci. Zvýšení síly testu dosáhneme také zvětšením rozsahu hodnot nezávislé proměnné. Například v manipulativních pokusech je nezávislá proměnná určována experimentátorem, a proto si můžeme zvolit její rozsah. Musíme ale počítat s tím, že čím větší rozsah, tím bývá u biologických závislostí větší odchylka od linearity.

Nezávislá proměnná s náhodnou variabilitou

Jak jsme již diskutovali výše, lineární regrese má sice předpoklad, že hodnoty nezávislé proměnné jsou pevné, bez náhodné variability, ale většina dat, se kterými se lineární regrese používá, tento předpoklad nesplňuje. V případě, že naším cílem při užití regrese není předpovídat (predikovat) nové hodnoty závislé proměnné, ale kvantifikovat skutečný vztah mezi proměnnými X a Y (tj, správně odhadnout parametr β_1), a přitom je variabilita nezávislé proměnné X podobně velká či dokonce větší než pro závislou proměnnou Y, použití klasické regrese odhadované metodou nejmenších čtverců (občas nazývané též regrese s modelem I) není správné. Odhad koeficientu β_1 je v takovém případě podceněn (posunut směrem k nule).

Existují ale také metody tzv. regrese s modelem II (*model II regression*), které variabilitu proměnné X berou do úvahy. Nejběžněji používanými technikami pro odhad těchto modelů jsou tzv. regrese hlavní osy (*major axis – MA - regression*) případně metoda regrese standardní hlavní osy (*standard major axis – SMA – regression*, někdy též *reduced major axis regression*). Důležitým omezením je, že běžné implementace této metody dovedou pracovat jen s jednou nezávislou proměnnou X. Typickým příkladem užití jsou alometrické vztahy.

Lineární kalibrace

Občas se vyskytne případ, kdy potřebujeme predikovat nějakou hodnotu, kterou jsme potenciálně schopni zjistit přesně (a tato hodnota tedy není zatížená chybou), pomocí metody, která je dostupnější, ale méně přesná. Například nadzemní biomasu bylin můžeme určit velmi přesně tím, že plochu sklidíme, usušíme a zvážíme. Tato metoda je pracná a je také destruktivní - na sklizeném místě nezůstane žádná biomasa. Naproti tomu existuje metoda, která je levná, rychlá (takže můžeme provést velké množství jednotlivých stanovení přes velké území), a navíc vegetaci příliš neponičí: po tyči spouštíme disk a měříme, v jaké výšce nad zemí jej vegetace zastaví. Tato metoda se běžně užívá např. při odhadu množství potravy pro herbivory. Tato metoda je ovšem zatížena relativně velkou chybou a nedává nám absolutní odhad biomasy, ale jen odhad relativní – kde je víc biomasy, tam se disk zastaví výše.

Proto tuto metodu potřebujeme zkalibrovat. To uděláme tak, že na vybraných místech nejprve odhadneme biomasu pomocí disku a poté plochu sklidíme a zjistíme přesnou hodnotu. Funkci pro přepočet výšky získáme pomocí regrese (někdy funguje i jednoduchá lineární regrese), ve které užíváme jako nezávislou (vysvětlující) proměnnou výšku zastavení disku (tedy proměnnou zatíženou chybou) a jako závislou (vysvětlovanouú proměnnou skutečnou biomasu. Děláme to tak proto, že v konečném použití funkce bude výška zastavení disku prediktorem. A také je naším cílem minimalizovat rozdíly mezi predikovanou a skutečnou hodnotou biomasy, nikoliv mezi predikovanou a skutečnou výškou zastavení disku. Podobných případů možného použití kalibrace je v biologii řada, hezký je příklad, kdy odhadujeme množství organismů na mořském dně z lodi (nepřesná rychlá metoda), kterou kalibrujeme tím, že se na mořské dno potopíme a všechna individua sklidíme.

Příkladová data

List *Chap12* obsahuje v prvých dvou sloupcích proměnné použitié pro ilustraci jednoduché lineární (přímkové) regrese. Jde o data z 16 jezer Severní Ameriky (Christensen et al. 1996), popisujících vztah mezi množstvím stromů na pobřeží jezera (proměnná *TreeDens*, km^{-1}) a množstvím hrubého dřevního opadu (*WoodDebris*, v $m^2.km^{-1}$). Naším cílem může být například predikce množství tohoto opadu pro nová jezera se známou densitou stromů na pobřeží.

Proměnné *W_body* a *W_brain* ve sloupcích D a E představují hmotnost těla a mozku u vybrané skupiny savců. Naším cílem je ověřit hypotézu, že alometrický poměr mezi logaritmovanou hmotností mozku a logaritmovanou hmotností těla je v základní populaci roven 2/3, protože hmotnost těla je úměrná jeho objemu, zatímco hmotnost mozku je úměrná povrchu těla, s ohledem na významný podíl inervace tohoto povrchu na kapacitě mozku.

Jak postupovat v programu Statistica

Jednoduchá regrese

Základní model lineární regrese (s jednou ale i více nezávislými proměnnými) lze odhadnout pro naše data volbou příkazu *Statistics* | *Multiple regression*. Pomocí tlačítka *Variables* zadáme proměnné (*WoodDebris* v levém seznamu *Dependenvt var.* a *TreeDens* v pravém seznamu *Independent variable list*). Pokud bychom chtěli zadat model regrese procházející počátkem nebo zvolit postupný výběr nezávislých proměnných (viz další kapitola), museli

bychom ještě na záložce *Advanced* zaškrtnout volbu *Advanced options*, ale pro náš jednoduchý model pokračujeme přímo tlačítkem *OK*.

Objeví se dialogové okno *Multiple Regression Results*, kde jsou základní charakteristiky odhadnutého regresního modelu zobrazeny v horní části (velké bílé pole), my si ale jednotlivé výsledky ukážeme za pomocí tlačítek dostupných v dolní části okna.

Základní shrnutí regresního modelu nám zobrazí tlačítko *Summary: Regression results* na záložce *Quick* (a také *Advanced*).

Regression Summary for Dependent Variable: WoodDebris (Spreadsheet9)

Regression Summary for Dependent Variable: WoodDebris (Spreadsheet9)
R= ,79654886 R²= ,63449008 Adjusted R²= ,60838223
F(1,14)=24,303 p<,00022 Std.Error of estimate: 36,318

N=16	b*	Std.Err. of b*	b	Std.Err. of b	t(14)	p-value
Intercept			-77,0991	30,60801	-2,51892	0,024552
TreeDens	0,796549	0,161579	0,1155	0,02343	4,92977	0,000222

V horní části zobrazené tabulky jsou dva odhady koeficientu determinace. Výsledek R^2 (s hodnotou *0.*6345) představuje koeficient determinace spočítaný podle Vz. 12-12 a lze jej interpretovat jako podíl z celkové variability hodnot závislé proměnné, který se nám modelem podařilo vysvětlit. Je ale známo, že tento koeficient je zkresleným odhadem (*biased estimate*) skutečného podílu vysvětlené variability v základní populaci, a to zejména pro složité modely založené na malém počtu pozorování (tedy když se DF_{REGR} blíží DF_{TOT}). Proto doporučujeme použití upraveného koeficientu determinace (*adjusted* R^2, často též zapisovaného jako R^2_{adj}), jehož hodnota pro naše data je 0.6084. Znamená to, že hustota stromů na pobřeží jezer vysvětluje zhruba 61% z variability hodnot množství dřevního opadu ve vodách jezera. V horní části pak následuje test průkaznosti celého modelu užívající F statistiku, ale ten probereme níže v rámci tabulky analýzy variance regresního modelu.

Ve dvou řádcích vlastní tabulky jsou zobrazeny odhady parametrů regresního modelu (pozor-až ve třetím sloupci označeném *b*, první dva sloupce jsou vysvětleny v kapitole 14. To znamená, že při známé hustotě stromů na pobřeží *TD* (například 1000 stromů na kilometr) můžeme očekávanou hodnotu dřevního opadu předpovídat pomocí rovnice

WD = -77.10 + 0.116*TD, tedy například 116-77.1 = 38.9 $m^2.km^{-1}$.

I když jsme si ještě regresní přímku nevynesli, kladná hodnota odhadu regresního koeficientu b_1 nám ukazuje, že je vztah mezi densitou stromů a dřevním opadem kladný: čím větší hustota, tím více opadu. Z povahy modelu regresní přímky dále vyplývá, že můžeme tento vztah popsat bez ohledu na hodnotu koeficientu průsečíku, například tak, že „se vzrůstem density stromů na pobřeží o 100 stromů/km vzroste průměrná pokryvnost opadu o 11.6 m^2 na km pobřeží. Negativní hodnota průsečíku (-77.1) ukazuje na limitace použitého modelu pro naše data – jezero bez stromů na pobřeží by sice mohlo existovat, ale sotva by mělo pokryvnost dřevního opadu -77.1. Zde je třeba si uvědomit, že naše minimální hodnota prediktoru je blízká 800 stromů na km, takže predikce -77.1 pokryvnosti při nulové denzitě stromů je extrapolací daleko za rozsah studovaných hodnot. Použití modelu přímky procházející počátkem by vedlo k závislosti, která bude procházet mimo zjištěná data. Lepší je použít zobecněný lineární model vhodného typu, kde budeme moci specifikovat, že pokryvnost má rozdělení, které může nabývat pouze nezáporných hodnot, ale ani tak se nevyhneme základnímu omezení, kterým je skutečnost, že nemáme jediný údaj pro situaci, kdy je hustota stromů na pobřeží menší než 770 stromů na km.

Odhady střední chyby pro odhady regresních koeficientů jsou ve sloupci nazvaném *Std.Err. of b* a sloupec *t(14)*[*] představuje hodnotu *t* statistiky pro test daného regresního koeficientu, tak jak popsán ve Vz. 12-13 a v textu pod ním. Zejména tyto dílčí testy založené na *t* statistice jsou citlivé na narušení předpokladu o homogenitě variancí. Výsledky nám ukazují, že oba koeficienty jsou průkazně odlišné od nuly.

Pomocí tlačítka *ANOVA (Overall goodness of fit)* na záložce *Advanced* si zobrazíme ANOVA tabulku pro náš regresní model.

Analysis of Variance; DV: WoodDebris (Spreadsheet9)

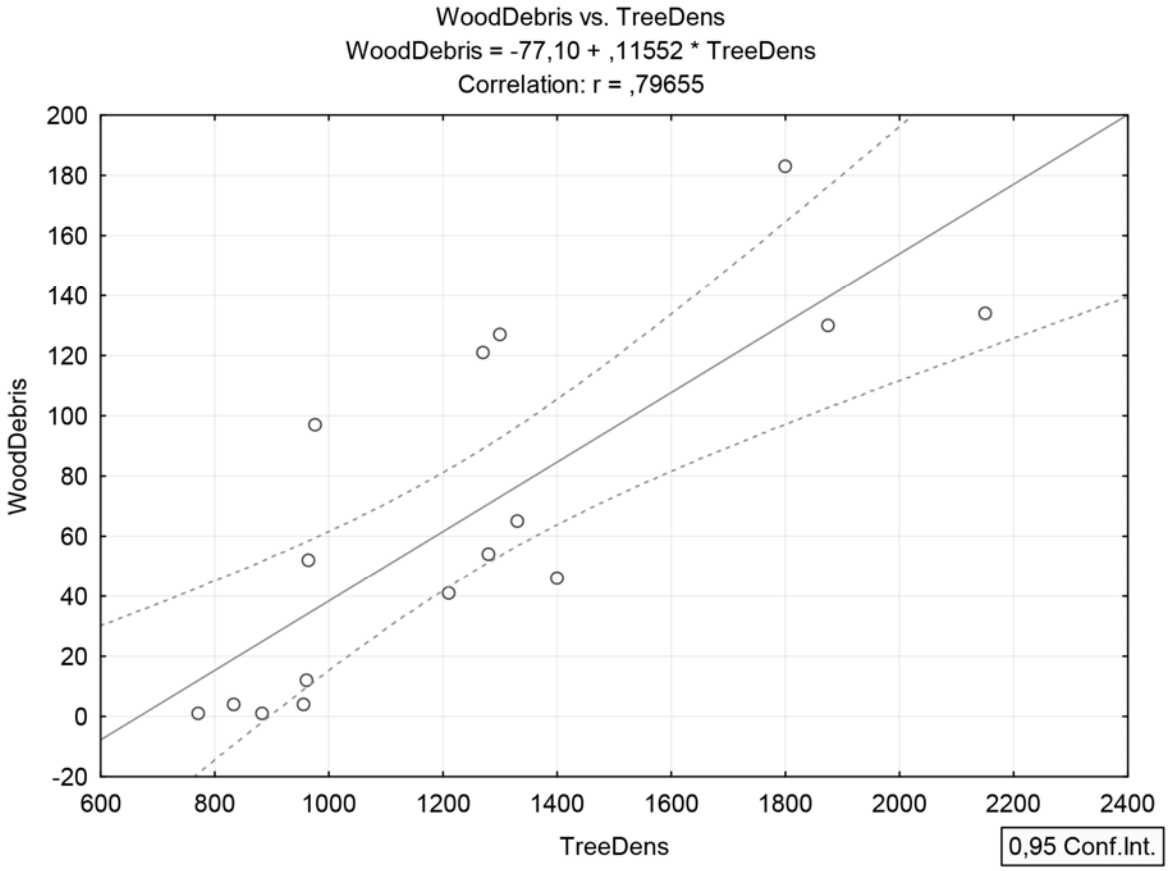
Sums of Squares	df	Mean Squares	F	p-value
32054,44	1	32054,44	24,30265	0,000222
18465,56	14	1318,97		
50520,00				

Z celkové sumy čtverců závislé proměnné (SS_{TOT}, 50520) náš model vysvětlil (SS_{REG}) 32054.4 a 18465.6 zůstalo nevysvětleno (SS_e). Poměr MS_{REG} a MS_e představuje F statistiku (24.3) s parametry 1, 14, která by za platnosti nulové hypotézy o absenci efektu nezávislých proměnných měla pocházet z $F_{1,14}$ distribuce. V případě našich dat tomu ale tak pravděpodobně není ($p<0.001$) a výsledek tak potvrzuje průkazný efekt density stromů na množství dřevního opadu.[x] V případě jednoduché lineární regrese s jednou nezávislou (vysvětlující) proměnnou nám tento F test nedává informaci odlišnou od t testu pro regresní koeficient β_1 – hodnoty *p-value* jsou totožné (a *F* statistika je druhou mocninou hodnoty *t* statistiky), ale při více nezávislých proměnných představuje F test zhodnocení celého modelu, umožňuje ale také porovnávat například dva modely lišícíce přítomností/absencí jedné nezávislé proměnné (viz kapitola 14).

Záložka *Residuals/assumptions/prediction* nám umožňuje jednak vytvářet grafy regresní diagnostiky, jednak vynést nafitovaný model (v případě jednoduché přímkové regrese), a také předpovídat hodnoty závislé proměnné na základě odhadnutého modelu (tlačítko *Predict dependent variable*).

Po volbě tlačítka *Perform residual analysis* se objeví nové dialogové okno *Residual Analysis*, nejdůležitější procedure jsou zde dostupné ze záložky *Scatterplots*. Vlastní regresní model můžeme vynést pomocí tlačítka *Bivariate correlation*, kde v novém dialogovém okně vybereme nezávislou proměnnou (*TreeDens*) v levém seznamu a závislou proměnnou (*WoodDebris*) v pravém seznamu.

[*] Číslo v závorce představuje počet reziduálních stupňů volnost a závisí tedy na velikosti dat a složitosti modelu.
[x] Na tomto místě opět zdůrazníme popisnou povahu regresního modelu: vyjádření „průkazný efekt density“ je statistická formulace a nijak neimplikuje kauzální efekt, i když ten v tomto příkladě nepochybně také existuje.



Model odhadnutá regresní přímka je představována plnou čárou, zatímco prohnuté čárkované čáry představují interval spolehlivosti, pokrývající (při splnění některých předpokladů, včetně vhodnosti zvoleného přímkového modelu) s pravděpodobností 0.95 střední hodnoty množství dřevního odpadu pro danou hustotu stromů. Pokud bychom chtěli vynést predikční interval, můžeme zvolit z menu příkaz *Format* | *Graph Options* a pak zvolit (v seznamu nalevo) *Plot / Regression Bands* a buď změnit typ z *Confidence* na *Prediction* nebo použít tlačítko *Add new pair of bands* a nový typ přidat ke stávajícímu.

V dialogovém okně *Residual Analysis* můžeme také na záložce *Scatterplots* vytvořit diagramy podobné těm v Obr. 12-4b,c a 12-5b,c, a to tlačítkem *Predicted vs. residuals*. Pro detekci měnící se variance residuálů je ale vhodnější varianta *Predicted vs. squared residuals*.

Regrese s modelem II

Tento typ regrese není v programu Statistica k dispozici, viz návod pro program R níže.

Jak postupovat v programu R

Proměnné ve sloupcích A a B byly importovány do datového rámce *chap12a*, zatímco proměnné ve sloupcí D a E do datového támce *chap12b*.

Jednoduchá regrese

Lineární regresní model odhadneme pomocí funkce *lm*:

```
> lm.1 <- lm(WoodDebris~TreeDens,data=chap12a)
```

Základní shrnutí obsahující (po stručné charakteristice regresních residuálů) odhady regresních koeficientů, standardních chyb těchto odhadů a t-testy regresních koeficientů v tabulce, následované informací o koeficientu determinace a také o F testu celého modelu:

```
> summary(lm.1)
Call:
lm(formula = WoodDebris ~ TreeDens, data = chap12a)

Residuals:
   Min     1Q Median     3Q    Max
-38.62 -22.41 -13.33  26.16  61.35

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -77.09908   30.60801  -2.519 0.024552 *
TreeDens      0.11552    0.02343   4.930 0.000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.32 on 14 degrees of freedom
Multiple R-squared:  0.6345,    Adjusted R-squared:  0.6084
F-statistic:  24.3 on 1 and 14 DF,  p-value: 0.0002216
```

Interpretace zobrazených regresních koeficientů a objasnění obou variant koeficientu determinace čtenář nalezne v popisu odhadu regresního modelu v programu Statistica výše. ANOVA tabulku pro regresní model lze zobrazit také funkcí *anova*, ta ale umožňuje i porovnání dvou či více regresních modelů.

```
> anova(lm.1)
Analysis of Variance Table

Response: WoodDebris
          Df Sum Sq Mean Sq F value    Pr(>F)
TreeDens   1  32054   32054  24.303 0.0002216 ***
Residuals 14  18466    1319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

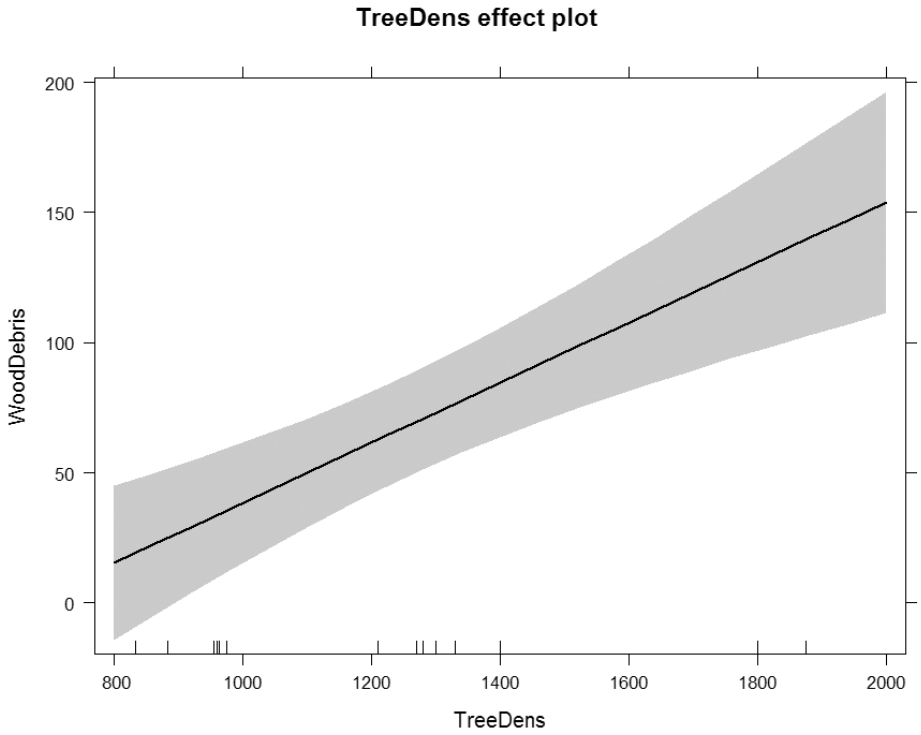
Odhadnutý přímkový model lze zobrazit nejsnadněji takto

```
> plot(WoodDebris~TreeDens,data=chap12a)
> abline(lm.1,lwd=2)
```

... ale přidání intervalu (regionu) spolehlivosti je pak o dost obtížnější. Doporučujeme v takovém případě použít knihovnu *effects*:

```
> library(effects)
> plot(allEffects(lm.1))
```

Tento kód zobrazí následující graf, kde je interval spolehlivosti reprezentován světle šedou oblastí.



Svislé čáry vyčnívající z horizontální osy do plochy diagramu představují hodnoty nezávislé (vysvětlující) proměnné pro jednotlivá pozorování.

Diagramy regresní diagnostiky lze vytvářet buď pomocí funkce *plot*, pokud zadáme objekt s odhadnutým modelem jako první parametr, nebo pomocí extrakčních funkcí typu *residuals* nebo *fitted*:

```
> plot(lm.1,which=1,add.smooth=F)
> plot(lm.1,which=3,add.smooth=F)
> plot(residuals(lm.1)~fitted(lm.1))
```

Regrese s modelem II

Tento typ regresního modelu lze odhadovat například s knihovnou *lmodel2*:

```
> lm2.1 <-lmodel2(log(W_brain)~log(W_body),data=chap12b,"interval","interval",
+ nperm=999)
```

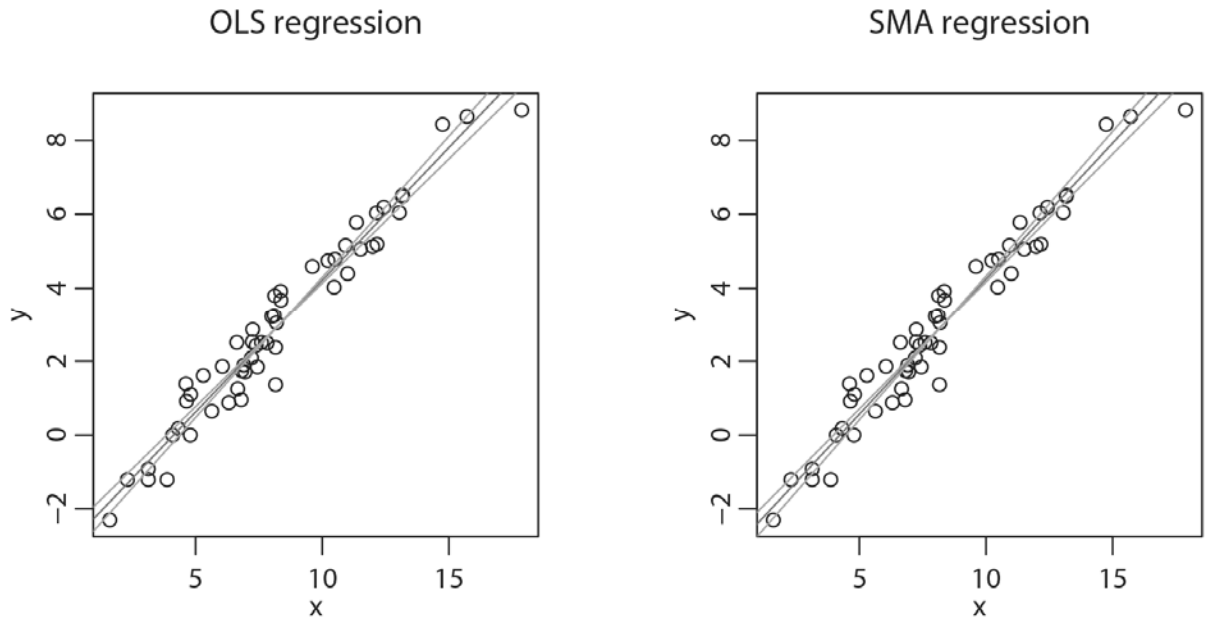
Funkce *lmodel2* odhaduje jak klasický lineární model metodou nejmenších čtverců (ve výsledcích odkazován zkratkou *OLS*, *ordinary least squares*), tak třemi metodami regrese pro model II: *major axis (MA) regression*, *standard major axis (SMA) regression* a také *ranged major axis regression (RMA*). Pomocí parametru *nperm* lze také zvolit provedení permutačního testu. Přehled výsledků lze získat zobrazením výsledného objektu (část výstupu vynechána):

```
> lm2.1
Model II regression
...
n = 54   r = 0.9753605   r-square = 0.951328
Parametric P-values:   2-tailed = 8.347005e-36    1-tailed = 4.173503e-36
Angle between the two OLS regression lines = 1.364954 degrees

Permutation tests of OLS, MA, RMA slopes: 1-tailed, tail corresponding to sign
A permutation test of r is equivalent to a permutation test of the OLS slope
P-perm for SMA = NA because the SMA slope cannot be tested
```

```
Regression results
  Method Intercept     Slope Angle (degrees) P-perm (1-tailed)
1    OLS -2.968594 0.7183359        35.69104             0.001
2     MA -3.072829 0.7309896        36.16642             0.001
3    SMA -3.118077 0.7364825        36.37100                NA
4    RMA -3.129393 0.7378562        36.42199             0.001

Confidence intervals
  Method 2.5%-Intercept 97.5%-Intercept 2.5%-Slope 97.5%-Slope
1    OLS      -3.372758       -2.564430  0.6731221   0.7635497
2     MA      -3.460514       -2.701784  0.6859464   0.7780527
3    SMA      -3.501951       -2.757048  0.6926552   0.7830829
4    RMA      -3.525572       -2.759170  0.6929129   0.7859504
...
```

Odhadnuté regresní koeficienty lze najít (pro jednotlivé metody odhadu parametrů modelu) ve sloupcích *Intercept* a *Slope* tabulky *Regression results*, spolu s průkazností odhadnutou permutačním testem (*P-perm (1-tailed)*), tabulka *Confidence intervals* obsahuje intervaly spolehlivosti a ty (pro všechny tři odhady Type II modelu) ukazují, že hypotetická hodnota 0.6667 není zahrnuta, tj. $p < 0.05$ pro hypotézu H_0: $\beta_1=0.667$. Můžeme si také vytvořit obrázek s nafitovanou přímkou zvoleného typu (níže je zobrazeno srovnání klasické regresní přímky s přímkou odhadnutou metodou *SMA*:

```
> par(mfrow=c(1,2))
> plot(lm2.1,"OLS")
> plot(lm2.1,"SMA")
> par(mfrow=c(1,1))
```



Popis metod v článku

Methods

The relation between shore tree density and the amount of raw woody debris was described using a linear regression model.

We have tested the agreement of allometric relation between mammalian brain and body weights with the hypothesized 2/3 ratio using standard major axis (Model II) regression on log-transformed weight values.

Results

We have found significant relation of woody debris relative area and shore tree density ($F_{1,14}=24.3$, $p=0.00022$) with the debris area increasing by 11.6 m^2 per kilometer of shore with each increase of tree density by 100 trees per km (see Figure X).

Obrázek Figure X by nejspíše obsahoval původní data spolu s průběhem regresní přímky, případně s intervaly spolehlivosti a s rovnicí obsahující odhady parametrů uvedenou v popisce obrázku. Detailnost popisu parametrů modelů závisí na typu časopisu a také relativní důležitosti daného modelu v kontextu článku.

Our estimates of the standard major axis regression model including the 95% confidence interval for the slope (0.693, 0.784) suggest that the allometric coefficient is larger than the hypothesized 2/3 value.

Doporučená četba

Sokal a Rohlf (1981) pp.454-560, Zar (1984) pp. 261-305, Quinn & Keough (2002) pp. 77-105.

D. L. Christensen et al. (1996): Impacts of lakeshore residential development on coarse woody debris in north temperate lakes. *Ecological Applications* **64**: 1143-1149.

13 Závislost dvou kvantitativních proměnných: korelace

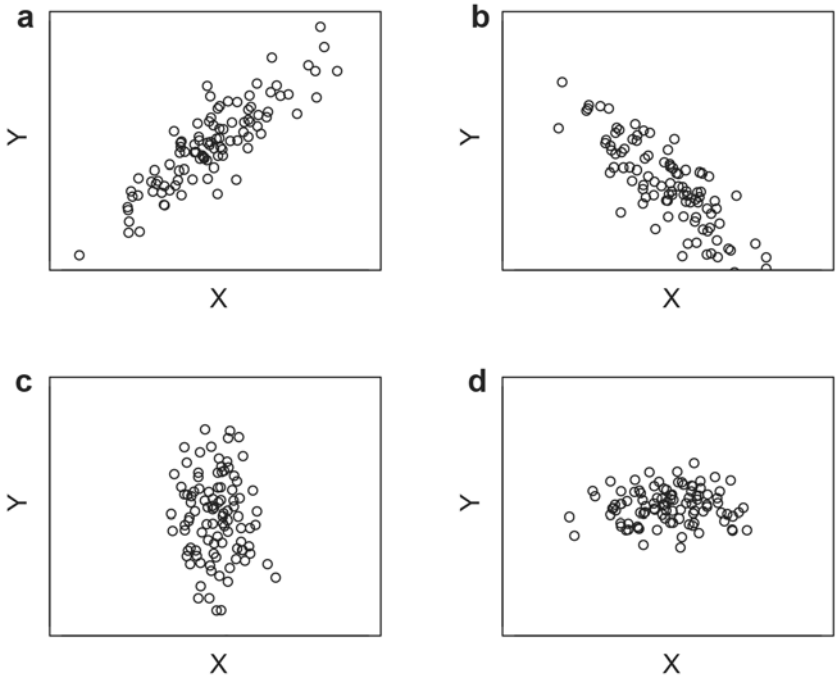
Při regresi jsme vycházeli z předpokladu, že mezi proměnnými existuje funkční závislost a že jsme schopni odlišit závislou a nezávislou proměnnou (nebo alespoň vysvětlující a vysvětlovanou proměnnou). Přitom jsme předpokládali, že nezávislá proměnná není zatížena chybou. Naproti tomu v korelační analýze předpokládáme, že není nutně funkční závislost jedné proměnné na druhé, dvě proměnné jsou pouze *korelovány* a obě proměnné jsou zatíženy náhodnou variabilitou.

Předpokládáme přitom (pokud nepoužíváme neparametrické korelace, viz níže), že proměnné pocházejí z tzv. dvourozměrného normálního rozdělení, tzn. že pro každou hodnotu X má proměnná Y normální rozdělení a pro každou hodnotu Y má proměnná X normální rozdělení. Z tohoto předpokladu plyne linearita vztahu. Mírou těsnosti vztahu je **korelační koeficient** (někdy též zvaný Pearsonův: uvede-li se korelační koeficient bez přívlastku, míní se tím obyčejně tento), který se vypočte takto:

$$r = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2 \sum\left(Y_i - \overline{Y}\right)^2}}$$

Vz. 13-1

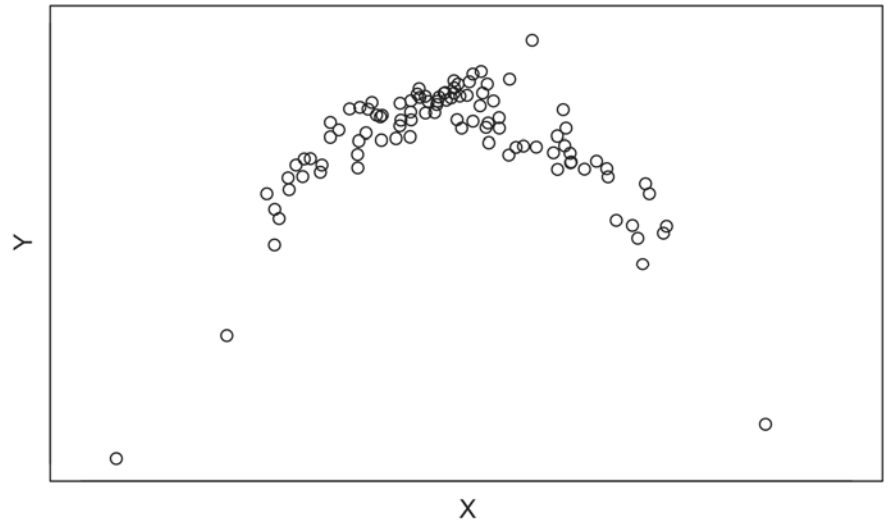
Přestože v tomto případě mají obě proměnné stejný význam (není zde závislá a nezávislá proměnná), používají se tradičně písmena X a Y. Jak vidíme, výraz je symetrický, tzn. že záměna X za Y nehraje žádnou roli. Výraz v čitateli je vždy kladný. Pokud jsou kladné odchylky od průměru v X většinou spojeny s kladnými odchylkami Y od průměru, je většina sčítaných členů v čitateli kladná (to platí zvláště, pokud jsou velké kladné odchylky v jedné proměnné spojeny s velkými kladnými v druhé) a čitatel i celý výraz je kladný. V opačném případě je záporný. Význam kladné a záporné korelace ukazuje Obr. 13-1.



Obr. 13-1 Jednoduchá lineární korelace: pozitivní korelace (a), negativní korelace (b), bez korelace (c a d)

Hodnoty *r* mohou být od *-1* do *1*. Hodnota *-1* značí deterministickou negativní závislost, *+1* deterministickou positivní závislost. Hodnota *r* rovná *0* znamená, že mezi proměnnými

není žádná lineární závislost. Obr. 13-2 ukazuje příklad, kdy dvě proměnné jsou vzájemně závislé, ovšem nikoliv lineárně; jejich korelační koeficient je velmi blízký nule. Všimněme si, že pro určité hodnoty Y má proměnná X bimodální rozdělení - je tedy výrazně narušen předpoklad normality. Korelační koeficient je tedy dobrou mírou těsnosti vztahu pouze pokud data pocházejí z dvourozměrného normálního rozdělení.



Obr. 13-2 Příklad dat, kde dvě proměnné nejsou nezávislé, ale korelační koeficient je (téměř) roven nule (r=-0.013)

Pokud bychom počítali regresi jedné proměnné na druhé, je koeficient determinace numericky roven druhé mocnině korelačního koeficientu (proto jej označujeme R^2).

Všimněme si rozdílu korelačního a regresního koeficientu. Regresní koeficient (b_1) nám říká, o kolik se změní závislá proměnná na jednotku nezávislé proměnné. Proto je uveden v jednotkách závislé proměnné na jednotky nezávislé proměnné; v případě pokryvnosti dřevního opadu (viz předchozí kapitola) by to byly m2/jednoho jedince stromu. Proto se také hodnota b_1 změní, pokud změníme jednotky měření. Koeficient b_1 může nabývat hodnoty od minus nekonečna do plus nekonečna. Naproti tomu korelační koeficient je bezrozměrné číslo, které vyjadřuje těsnost vztahu, a jeho hodnota je nezávislá na použitých jednotkách.

Hodnotu r prakticky vždy počítáme pro výběr pozorování; pokládáme potom r za odhad parametru základního souboru ρ. Chceme tedy většinou testovat hypotézy o ρ, nejčastěji nulovou hypotézu, $H_0: \rho=0$. K tomu můžeme použít výpočet střední chyby odhadu:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Vz. 13-2

Potom spočteme

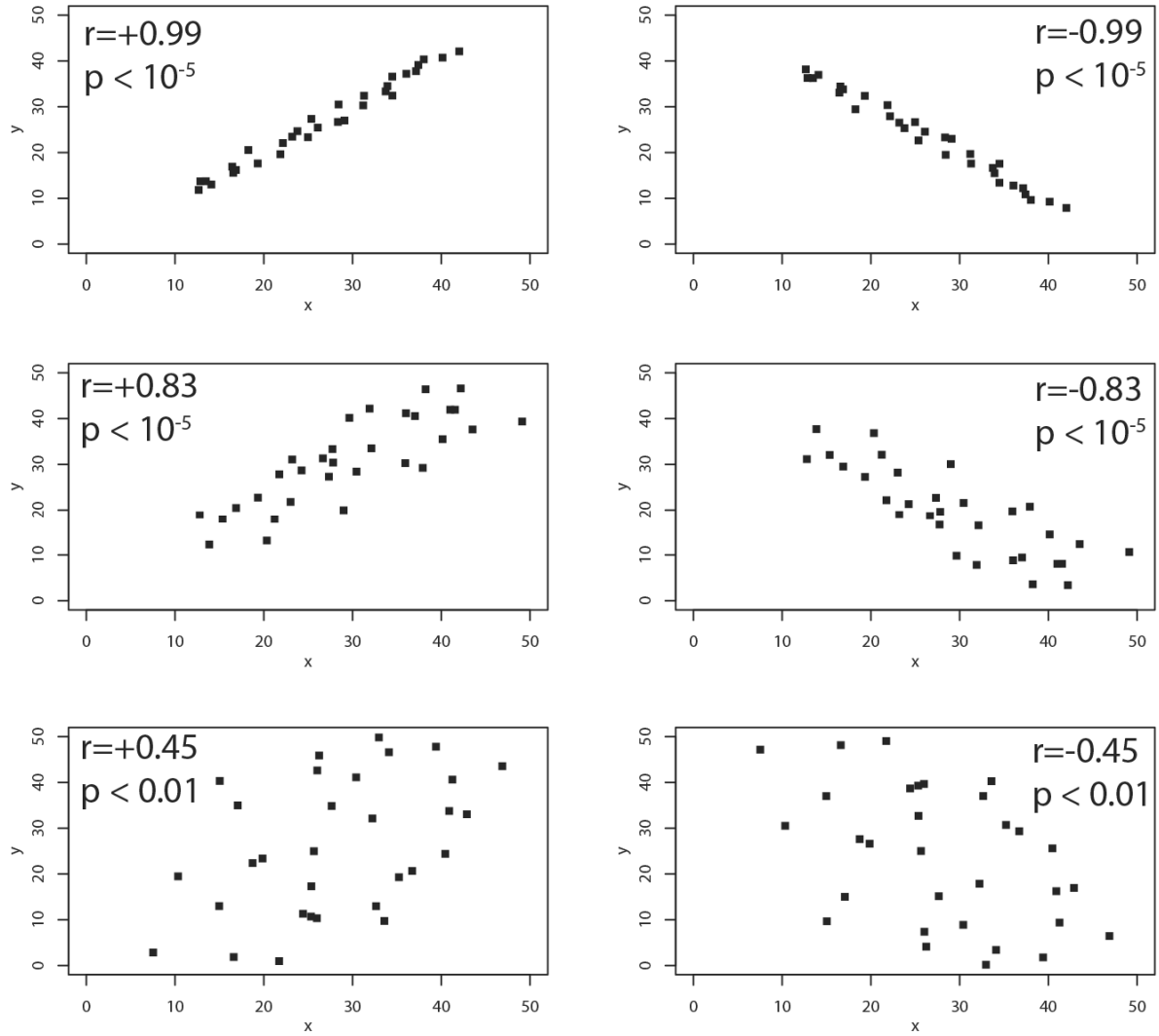
$$t = \frac{r}{s_r}$$

Vz. 13-3

a porovnáme s t rozdělením při $n-2$ stupních volnosti. Pokud platí $\rho = 0$, má r přibližně normální rozdělení. Můžeme použít jednostranný i dvoustranný test. Pro daný počet pozorování můžeme přímo vypočítat kritické hodnoty r (Tab. 13-1). Je třeba poznamenat, že

dosažená hladina významnosti (p) pro test nulové hypotézy $\rho = 0$, je shodná s p pro regresi jedné proměnné na druhé.

Představu o tom, jak těsné jsou závislosti při různých hodnotách korelačního koeficientu může dát Obr. 13-3.



Obr. 13-3 Závislost dvou proměnných při různých hodnotách korelačního koeficientu. Všechny koeficienty jsou spočteny na základě 31 pozorování

Pokud je ρ_0 různé od nuly, potom r **nemá** normální rozdělení. Abychom získali proměnnou s normálním rozdělením, musíme nejprve provést tzv. z-transformaci

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

Vz. 13-4

z má přibližně normální rozdělení, se směrodatnou odchylkou

$$\sigma_z = \sqrt{\frac{1}{n-3}}$$

Vz. 13-5

Síla testu

Protože platí, že regrese Y na X je průkazná právě tehdy, když je průkazný korelační koeficient, platí pro obě metody společné zásady o určování síly testu. Platí tedy, že síla testu stoupá s těsností vztahu (tj. s hodnotou korelačního koeficientu základního souboru) a s velikostí výběru. Protože při korelačním výzkumu provádíme náhodný výběr ze základního souboru a žádná z proměnných není ovlivněna experimentátorem, nelze (jako při regresi) zvětšit rozsah hodnot některé proměnné.

Nejjednodušším způsobem (který je zcela neformální, statisticky nepřesný, a tedy slouží jen k hrubé orientaci) je podívat se do přiložené tabulky kritických hodnot korelačního koeficientu (Tab. 13-1). Se zvětšováním výběru se střední hodnota výběrového korelačního koeficientu nemění. Jestliže např. předpokládáme, že korelační koeficient základního souboru bude přibližně 0.5 a chceme provádět oboustranný test na 5%-ní hladině významnosti, potom pokud budeme mít méně než 16 pozorování, tj. 14 stupňů volnosti, velmi pravděpodobně nedokážeme zamítnout nulovou hypotézu. Uvědomme si, že pokud budeme mít při daném počtu pozorování střední hodnotu korelačního koeficientu rovnou jeho kritické hodnotě (tedy např. při 17 pozorováních budeme předpokládat, že korelační koeficient základního souboru je roven 0.48) máme přibližně jen 50% šanci, že nulovou hypotézu zamítneme. Úvaha není statisticky zcela přesná, ale pro hrubou orientaci postačuje.

Tab. 13-1 Kritické hodnoty korelačního koeficientu pro dvoustranný test nulové hypotézy $\rho = 0$. Počet stupňů volnosti je n-2. Nulovou hypotézu zamítáme, pokud absolutní hodnota výběrového korelačního koeficientu překročí kritickou hodnotu.

df	3	4	5	6	7	10	15	20	30	50	100
$\alpha = 0.05$	0.8783	0.8114	0.7545	0.7067	0.6664	0.5760	0.4821	0.4227	0.3494	0.2732	0.1946
$\alpha = 0.01$	0.9587	0.9172	0.8745	0.8343	0.7977	0.7079	0.6055	0.5368	0.4487	0.3541	0.2540

Přesnější je následující postup: Máme spočten výběrový korelační koeficient. Spočteme *z*-transformaci *r* (*z*) a *z*-transformaci kritické hodnoty při stanovené hladině α: z_{krit}. Potom pravděpodobnost, se kterou bude normovaná normální proměnná ($\mu=0$, $\sigma=1$) menší než hodnota $Z = (z - z_{krit})\sqrt{n-3}$ je rovna síle testu. Např. při spočteném korelačním koeficientu *0.866* (*z=1.31*) a *12* pozorováních chceme odhadnout sílu dvoustranného testu na 5%-ní hladině významnosti, tj. za předpokladu, že je korelační koeficient základního souboru roven spočtenému výběrovému korelačnímu koeficientu. Počet stupňů volnosti je *10*. Kritická hodnota *r* je *0.576* (z tabulky), $z_{krit}=0.656$. Z toho dostáváme $Z=(1.31 - 0.6565) \sqrt{(12-3)} = 1.98$. Pravděpodobnost, že normovaná normální náhodná proměnná je menší než 1.98 je 0.98 (viz *Statistics* | *Probability Calculator* | *Distributions* v programu Statistica nebo funkce *pnorm* v programu R). Je tedy síla testu 0.98, tzn. že v 98% případů budeme schopni zamítnout na výběrech o velikosti 12 z dané populace nulovou hypotézu.

Pokud máme předpokládanou hodnotu korelačního koeficientu v základním souboru (často nejmenší hodnota, která nás ještě zajímá), stanovenu hladinu významnosti α a požadovanou sílu testu $(1-\beta)$, potom potřebná velikost výběru se spočte

$$n = \left(\frac{Z_\beta + Z_\alpha}{z_{\text{předp}}} \right)^2 + 3$$

Vz. 13-6

Z_β a Z_α jsou kritické hodnoty normovaného normálního rozdělení (Z_β vždy jednostranný, Z_α jedno nebo dvoustranný, podle typu testu, získáme z tabulek nebo v programu Statgraphics) a $z_{předp}$ je z-transformovaná hodnota přepodkládané hodnoty korelačního koeficientu.

Příklad: Chceme v oboustranném testu na 5% hladině významnosti zamítnout nulovou hypotézu o nezávislosti dvou veličin (tj. že $\rho=0$) s 99%-ní pravděpodobností, pokud skutečná hodnota korelačního koeficientu je v absolutní hodnotě alespoň 0.5. $\beta=0.01$, kritická hodnota pro jednostranný test je tedy 99%-ní kvantil normovaného normálního rozdělení, tj. 2.3263; $\alpha=0.05$, kritická hodnota pro dvoustranný test je 97.5%-ní kvantil, tj. 1.9600). Když dosadíme za r ve Vz. 13-4 hodnotu 0.5 (z-transformace předpokládané hodnoty korelačního koeficientu) dostáváme $z_{předp}= 0.5493$. Dosazením do Vz. 13-6 získáme

$$n = \left(\frac{2.3263 + 1.9600}{0.5493} \right)^2 + 3 = 63.9$$

Potřebujeme tedy 64 pozorování.

Neparametrické metody

Pokud nemají data dvourozměrné normální rozdělení a odchylka od předpokladu je katastrofálně velká, můžeme použít neparametrickou metodu. Nejčastější neparametrickou metodou je pro zjištění korelace výpočet Spearmanova korelačního koeficientu (angl. *Spearman* nebo *rank correlation coefficient*). Je založen na pořadí. Nahradíme skutečné hodnoty v každé proměnné jejich pořadím a z nich spočteme korelační koeficient. Dá se použít vzorce **Error! Unknown switch argument.** nebo s týmž výsledkem:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Vz. 13-7

kde d_i je diference v pořadí. Je však třeba upozornit, že pro data odpovídající Obr. 13-2 bude r_S také přibližně rovno nule.

Ještě více neparametrický je koeficient Kendallův (*Kendall tau coefficient*), ten pracuje jen s počtem souhlasných a nesouhlasných pořadí hodnot dvou srovnávaných proměnných, nepočítá rozdíly v hodnotách pořadí.

Poznámky k interpretaci

Jak ve spojení s regresí, tak i s korelací, se často objevují otázky: jak velký korelační koeficient (nebo R^2 v regresi) je známkou dostatečně těsné vazby nebo závislosti? Někdy se dokonce za těsnost vazby vydává dosažená hladina významnosti, což je nesmyslné. Odstrašujícím příkladem může být prohlášení typu „závislost je velice těsná, podařilo se nám ji prokázat na 0.01%-ní hladině významnosti“. Ve skutečnosti jsme jen zamítli hypotézu, že vztah není žádný: při velkém počtu pozorování na to stačí relativně slabá závislost.

Těsnost vazby měří jen korelační koeficient, případně R^2. V regresi R^2 udává podíl variability závislé proměnné vysvětlené nezávislou proměnnou. Proto to, co budeme považovat za rozumnou sílu vazby, závisí případ od případu. V některých případech nás bude zajímat jen míra těsnosti a test je prakticky nesmyslný: porovnáváme-li dvě chemické metody stanovení dusíku, potom testovat nulovou hypotézu, že mezi dvěma metodami není žádná

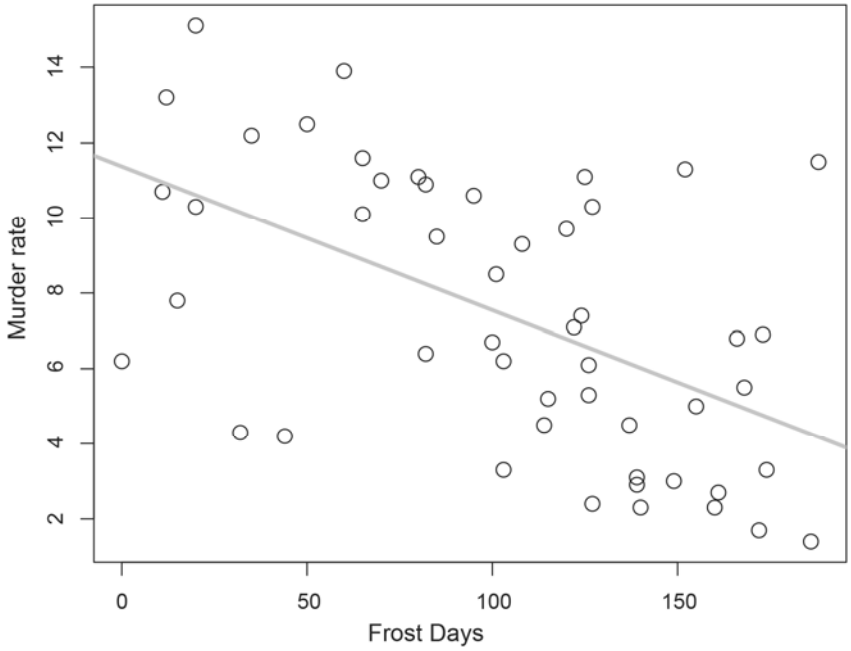
závislost, je nesmysl. Ještě větší nesmysl je prohlásit, že metody jsou ekvivalentní, protože vypočtený korelační koeficient je vysoce průkazný. Hodnotu korelačního koeficientu 0.90 budu pravděpodobně považovat za nízkou a nevyhovující, bez ohledu na to, že se mi bude statisticky vysoce průkazně lišit od nuly.

Naproti tomu, pokud budeme studovat závislost počtu druhů na m^2 na množství půdního humusu, potom je na místě nejprve otestovat nulovou hypotézu a teprve poté uvažovat o těsnosti vztahu. V takovém případě bude i velmi nízký koeficient determinace zajímavý, pokud bude regrese průkazná.

Před výpočtem korelací mezi proměnnými bychom si vždy měli graficky (XY diagram) ověřit povahu vztahu mezi těmito proměnnými. Vztah by měl být lineární v případě Pearsonova *r* a monotónní (rostoucí či klesající závislost, ale ne nutně lineárně) v případě neparametrických koeficientů. Pro vztahy, které monotónní nejsou (např. Obr. 13-2), můžeme těsnost vztahu odhadnout tak, že zvolíme vhodný typ regresního modelu (například polynom druhého stupně pro Obr. 13-2) a spočteme korelaci mezi předpovídanými (fitovanými) a skutečnými hodnotami Y (tedy odmocninu z koeficientu determinace R^2) – pro Obr. 13-2 je například tato hodnota r=0.940.

Statistická závislost a kauzalita

Už při vyhodnocování kontingenčních tabulek jsme upozorňovali, že statistická závislost nemusí vždy znamenat závislost příčinnou. Pokud se jedná o korelace, na obě proměnné pohlížíme stejně, a tedy průkaznou korelaci můžeme těžko považovat za důkaz příčinné závislosti. To nám ale nebrání v tom, abychom hledali příčinu této korelace. Složitějším případem jsou průkazné výsledky regresní analýzy. Zde máme závislou a nezávislou proměnnou a často uvažujeme o kauzálním vztahu. Kdy jsou tyto úvahy oprávněné?



Obr. 13-4 Výsledky regresní analýzy počtu vražd na 100 000 obyvatel v roce 1976 (*Murder rate*) v jednotlivých státech USA v závislosti na průměrném počtu mrazových dní v hlavním městě daného státu v letech 1931-1960 (*Frost Days*).

Vezměme příklad regresní analýzy (Obr. 13-4), kde je vysvětlovanou (závislou) proměnnou počet vražd na 100 000 obyvatel v jednotlivých státech USA a vysvětlující

proměnnou je průměrný počet mrazových dní v hlavním městě daného státu. Předpokládat, že počet vražd ve státě může ovlivnit, kdy mrzne, by bylo velmi odvážné. Naproti tomu se mohou vyskytnout teorie o tom, že velká vedra vzbuzují v lidech násilné choutky, zatímco mrazy tlumí jejich aktivitu i v násilnické oblasti. Překvapivě, Obr. 13-3 je v souhlase s touto teorií. Když se však podíváme na celý statistický soubor dat o USA, ze kterého jsou tyto údaje převzaty, zjistíme, že počet mrazových dní skvěle vysvětluje i průměrný příjem obyvatele, procento maturantů v dospělé populaci a procento negramotných. Není to ovšem kauzální vliv; tradičně jsou v USA tzv. jižanské státy chudší a také divočejší (v kovbojkách se vždycky nejvíce střílelo při mexické hranici). Existuje tedy určitý „kulturní" severojižní gradient, který je pochopitelně korelován s teplotou. Vidíme tedy, že obecně může být vysoce průkazná regrese výsledkem faktu, že vysvětlující i vysvětlovaná proměnná jsou závislé na nějaké třetí proměnné (či více proměnných). Z takových dat pochází nejvíce nesmyslných korelací, vydávaných čas od času za kauzální závislosti. Klasickým případem takového použití regrese je případ, kdy jsou za jednotlivá pozorování brány údaje z jednotlivých let. Tak lze například ukázat, že je vysoce průkazná pozitivní korelace mezi koncentrací oxidu uhličitého a počtem obyvatel Indie (obojí během času vzrůstá), nebo mezi počtem ledniček a počtem rozvodů za posledních 90 let.

Kdy tedy můžeme považovat průkaznou regresi za důkaz kauzální závislosti? Pouze tehdy, pokud se jedná o výsledky manipulativního experimentu, kde je nezávislá proměnná (ve správném experimentálním uspořádání) manipulována experimentátorem. Jestliže vyberu na louce 10 ploch, každé náhodně přiřadím dávku hnojiva od 0 do 9 a výsledná biomasa bude průkazně závislá na dávce hnojiva, potom mohu tvrdit, že hnojivo ovlivnilo výslednou biomasu.

To, co bylo řečeno o regresi (a co známe z vyhodnocování kontingenčních tabulek), platí i v širším rámci studia závislosti dvou proměnných, bez ohledu na to, zda se jedná o proměnné kvalitativní (faktory) nebo kvantitativní. Studuji například, zda početnost populace pavouků na ostrovech závisí na tom, zda na daném ostrově žije či nežije druh hmyzožravé ještěrky. Mohu t-testem porovnat míry početnosti na ostrovech s ještěrkou a bez nich. Pokud dostanu rozdíly, mohou být způsobeny buď opravdu tím, že ještěrka pavouky žere, nebo mně neznámým faktorem, který způsobuje, že na daných ostrovech je populace pavouků nízká a zároveň způsobuje, že zde ještěrky nemohou žít; případně zde může být i opačný kauzální vztah. Tam, kde je z určitého (mně neznámého) důvodu málo pavouků, ještěrky hladoví a nakonec na ostrově vymřou. Pokud na náhodně vybrané polovině ostrovů ještěrky vysadím a na polovině nevysadím, a tam, kde jsem ještěrky vysadil, klesne populace pavouků, zatímco na ostrovech bez ještěrek ke změnám nedojde, mohu mluvit o kauzální závislosti.

Pouze manipulativní experiment je dobrým průkazem kauzality.

V některých reálných situacích jsou ovšem experimentální důkazy obtížné až nemožné, např. v některých případech v ekologii v krajinném měřítku, v některých evolučních studiích, v řadě studií člověka. Ale i shora popsaný pokus s ještěrkami by jistě ochranu přírody nepotěšil. Potom se často musíme spokojit se statistickou závislostí a hledáme pro své tvrzení o kauzalitě další nepřímé podpůrné „důkazy".

Příkladová data

Proměnné *Conduct* a *Ca* v listu *Chap13* představují vodivost a obsah vápenných kationtů ve vzorcích vody z 33 potoků Šumavy. Očekáváme, že tyto dva parametry budou spolu pozitivně korelovány, a zajímá nás, zda tomu tak je a jak je tento vztah těsný.

Jak postupovat v programu Statistica

Z menu zvolíme příkaz *Statistics* | *Basic Statistics/Tables* a ze seznamu vybereme *Correlation matrices*. Tuto proceduru můžeme používat i pro spočtení korelací mezi větším počtem proměnných. Pokud chceme počítat korelaci každé naší proměnné se všemi ostatními, zadáme proměnné pomocí tlačítka *One variable list*. Pokud bychom ale chtěli korelovat všechny proměnné z jedné skupiny s proměnnými skupina další (ale ne korelovat proměnné v rámci skupin), užijeme tlačítko *Two lists*. To je výhodné, pokud chceme například spočíst korelace četností vybraných druhů se změřenými charakteristikami prostředí. V našem případě se dvěma korelovanými proměnnými vedou obě zadání ke stejným výsledkům. Po volbě tlačítka *Summary* na záložce *Quick* získáme výsledek:

	Correlations (Spreadsheet3) Marked correlations are significant at p < ,05000 N=33 (Casewise deletion of missing data)			
Variable	**Means**	Std.Dev.	Conduct	Ca
Conduct	150,8485	85,94552	1,000000	0,537024
Ca	19,4576	11,84145	0,537024	1,000000

Vidíme, že Pearsonův lineární korelační koeficient je vskutku pozitivní (0.537) a průkazně odlišný od nuly. Pro zjištění přesnější odhady průkaznosti tohoto testu bychom v dialogovém okně *Product-Moment and Partial Correlations* museli na záložce *Options* zvolit *Display r, p-values, and N's* před volbou tlačítka *Summary*. Zde je odhadnutá průkaznost zobrazena jako 0.001, naše hypotéza ale nebyla symetrická (očekávali jsme, jako H_A, že korelace bude pozitivní), takže s ohledem na kladné znaménko výsledku můžeme za průkaznost jednostranného testu považovat hodnotu 0.001/2 = 0.0005. V dialogovém okně lze také (na záložce *Quick* nebo *Advanced*) vytvořit XY diagramy. Pro menší počet porovnávaných proměnných doporučujeme tlačítko *Scatterplot matrix*, ukazující všechna porovnání v jednom grafu, spolu s nafitovanými regresními přímkami.

Neparametrické korelační koeficienty lze vypočíst volbou příkazu *Statistics* | *Nonparametrics* a následným výběrem položky *Correlations (Spearman, Kendall tau, gamma)*. Spearmanův korelační koeficient je k dispozici na záložkách *Quick* i *Advanced*.

Jak postupovat v programu R

Klasický (Pearsonův) i neparametrický Spearmanův koeficient korelace lze spočíst pomocí funkce *cor*.

```
> with(chap13,cor(Ca,Conduct))
[1] 0.5370242
> with(chap13,cor(Ca,Conduct,method="spearman"))
[1] 0.5841063
```

Test Pearsonova koeficientu (a také výpočet intervalu spolehlivosti) provádí funkce *cor.test*. Můžeme v ní, pomocí parametru *alternative* specifikovat i případnou nesymetrickou hypotézu, jak je tomu v případě našich dat:

```
> with(chap13,cor.test(Ca,Conduct,alternative="greater"))
        Pearson's product-moment correlation
data:  Ca and Conduct
t = 3.5445, df = 31, p-value = 0.0006358
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.2909983 1.0000000
```

```
sample estimates:
      cor
0.5370242
```

Grafy znázorňující vztahy mezi porovnávanými proměnnými můžeme vytvářet pomocí funkcí *plot* (či *xyplot* v knihovně *lattice*), matice párových XY grafů pak pomocí funkce *pairs* nebo (v knihovně *lattice*) funkce *splom*.

Popis metod v článku

Methods

We have quantified the correlation between conductivity and calcium concentration using Pearson linear correlation and tested their expected positive correlation using one-sided t test.

Due to a non-linear relation between the examined variables, we have used Spearman correlation coefficient to quantify the strength of relation between water conductivity and calcium ion concentrations.

Results

We have confirmed a positive, medium strength correlation between the conductivity and calcium concentration ($r=0.537$, $n=33$, $p<0.001$).

Doporučená četba

Sokal & Rohlf (1981) pp. 561-616; Zar J. H. (1984) pp. 306-327, Quinn & Keough (2002) pp. 72-77

14 Mnohonásobná regrese a obecné lineární modely

Doposud jsme se zabývali případy, kdy jedna vysvětlovaná (závislá) proměnná závisí na jedné vysvětlující (nezávislé) proměnné. Nyní se budeme zabývat případy, kdy jedna vysvětlovaná proměnná závisí na mnoha vysvětlujících proměnných – jde o model mnohonásobné regrese (*multiple regression*). Například budeme studovat závislost nadzemní biomasy lipnice luční na množství N a P v půdě. S ohledem na povahu proměnných zde pracujeme se všemi třemi proměnnými na logaritmické škále. Nezávislé proměnné budeme označovat X_1, X_2 atd., závislá proměnná bude Y. j-té pozorování první nezávislé proměnné bude tedy $X_{1,j}$. Pro jednu vysvětlující proměnnou byla regresní rovnice

$$\hat{Y} = \beta_0 + \beta_1 X$$

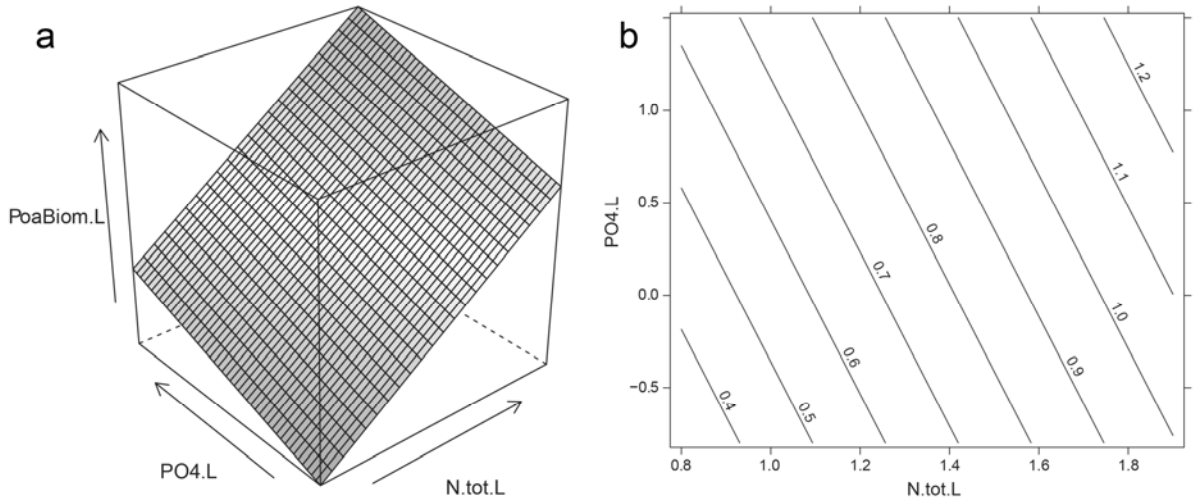
Vz. 14-1

a jejím obrazem byla přímka. Pro dvě vysvětlující proměnné je regresní závislost

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Vz. 14-2

a jejím obrazem je rovina v prostoru (jak ukazuje Obr. 14-1).



Obr. 14-1 Závislost nadzemní biomasy lipnice (*PoaBiom.L*) na množství celkového dusíku (*N.tot.L*) a koncentraci fosfátových iontů (*PO4.L*).

Místo Vz. 14-1 bychom mohli psát

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Vz. 14-3

kde ε je náhodná variabilita - proměnná s normálním rozdělením, střední hodnotou nula a stálou variancí (nezávislou na hodnotách vysvětlujících proměnných, neměnící se s hodnotou Y). Předpoklady jsou tedy stejné jako u jednoduché regrese - aditivita chyby i efektů, normalita rozdělení, stálost variance. Případ pro dvě proměnné můžeme zobecnit pro více proměnných:

$$Y = \beta_0 + \sum_i \beta_i X_i + \varepsilon$$

Vz. 14-4

Koeficienty β_i se nazývají parciální regresní koeficienty. Hodnoty těchto koeficientů neznáme; odhadujeme je jako parametry b_0 a b_i na základě výběru, a to tak, aby ve výběru splňovaly kriterium nejmenších čtverců. Ke každému regresnímu koeficientu lze také odhadnout střední chybu jeho odhadu. Podíl *t = hodnota parametru / střední chyba parametru* slouží jako testové kriterium pro test nulové hypotézy, že hodnota příslušného koeficientu (β_0, β_i) je rovna nule -u β_i jde tedy o test významnosti vlivu dané proměnné. Za předpokladu platnosti nulové hypotézy má tato statistika *t*-rozdělení s počtem stupňů volnosti *df = n-m-1*, kde *m* je počet vysvětlujících proměnných.

Obvykle se test používá hlavně pro koeficienty b_i , kde je testem průkaznosti parciálního efektu jednotlivých proměnných; test pro koeficient b_0 často nemá reálný smysl, testuje nulovou hypotézu, že hodnota závislé proměnné je nula pro nulové hodnoty všech nezávislých proměnných. I když je tento test někdy zdánlivě smysluplný, často bývá založen na extrapolacích (zvláště u prediktorů měřených na poměrové stupnici). V našem případě, vzhledem k tomu, že všechny proměnné máme logaritmicky transformované, testujeme nulovou hypotézu, že logaritmus biomasy druhu je pro nulové hodnoty logaritmu množství dusíku a fosforečnanů roven nule, tedy že pro množství dusíku 1 a fosforu 1 by biomasa měla být také 1 (vše v jednotkách, které jsme použili). Není jediný důvod, proč takovou hypotézu testovat.

Zatímco *t*-testy pro jednotlivé koeficienty b_i slouží k testování dílčích hypotéz o vlivu jednotlivých vysvětlujících (nezávislých) proměnných, slouží F test pro ANOVA tabulky regresního modelu jako celkový test nulové hypotézy, že celý **regresní model** nevysvětluje žádnou část variability vysvětlované proměnné. Odpovídá tedy nulové hypotéze, že všechny koeficienty β_i jsou rovny nule. Setkáváme se zde, podobně jako v jednoduché regresi (viz kapitola 12), opět s celkovým součtem čtverců (*total sum of squares*), se součtem čtverců pro regresi (*regression sum of squares* nebo *model sum of squares*), a s reziduálním součtem čtverců (*residual sum of squares*). Příslušné počty stupňů volnosti jsou *n-1* pro celkový, *m* pro regresní a *n-m-1* pro reziduální součet čverců. Příslušný průměrný čtverec (*mean square, MS)* je roven *MS = SS / DF*. Stejně jako u jednoduché regrese se pro test nulové hypotézy užívá podílu

$$F = \frac{MS_{REGR}}{MS_e}$$

Vz. 14-5

a stejně tak se spočte koeficient determinace

$$R^2 = \frac{SS_{REGR}}{SS_{TOT}}$$

Vz. 14-6

Tento odhad je ale vychýleným odhadem koeficientu determinace základního souboru: čím méně dat a více vysvětlujících proměnných máme, tím vyjde vyšší (viz též kapitola 12). Připomeňme, že lineární kombinací *m* proměnných lze proložit *m*+1 bodů přesně. Pokud bychom měli jen *m+2* bodů v regresi s *m* proměnnými, vyjde koeficient determinace značně vysoký i v případě, že sledované proměnné jsou nezávislé. Proto se používá korekce a počítá se takto:

$$R^2{}_{adj} = R^2 - \frac{m}{n-m-1}\left(1 - R^2\right)$$

Vz. 14-7

(*adj = adjusted*). V počítačových výstupech se uvádějí obvykle obě hodnoty (R^2 i R^2_{adj}). Tato korekce je zanedbatelná, pokud máme velký soubor dat a relativně málo vysvětlujících proměnných.

Upozorněme, že se může stát, že některý z regresních koeficientů vyjde průkazně odlišný od nuly, zatímco celkový test vyjde neprůkazně. To většinou indikuje, že máme v regresním modelu zbytečně velké množství vysvětlujících proměnných, které nic nevysvětlují. Pokud tedy F-test pro celý regresní model nevyjde průkazný, nedoporučuje se již příliš důvěřovat průkaznosti jednotlivých parciálních regresních koeficientů: pravděpodobnost chyby prvního druhu je rovna zvolené hladině významnosti α v každém dílčím testu.

Naopak se může stát, že celková regrese vyjde průkazná, zatímco žádný z parciálních regresních koeficientů průkazně odlišný od nuly není; to většinou znamená, že vysvětlující proměnné jsou vzájemne korelované. Vysvětlující (nezávislé) proměnné by měly být skutečně nezávislé, tzn. korelované by být neměly. V praxi ovšem mnohdy jsou - například stanoviště bohatá na celkový dusík bývají bohatší i na fosforečnanové ionty. V tom případě je ovšem těžké rozhodnout, která z vysvětlujících proměnných má vliv na závislou proměnnou - to vede k velkým hodnotám střední chyby odhadu parametrů.

Ne všechny vysvětlující proměnné mají stejně silný vliv na vysvětlovanou proměnnou, některé je často možné z regresního modelu vyloučit. To lze provést buď ručně (porovnáním výsledků různých regresních modelů) nebo automaticky. Ve většině programů pro to existují procedury, zvané postupný výběr (*stepwise selection*) či postupná regrese (*stepwise regression*), které podle předem stanovených pravidel postupně přidávají nebo ubírají proměnné. Takový postup vybere podsoubor vysvětlujících proměnných tak, aby přidání další proměnné už nevedlo k významnému snížení nevysvětlené variability, ale aby vypuštění kterékoliv proměnné vedlo ke zvýšení nevysvětlené variability.

Pokud měníme soubor vysvětlujících proměnných (někdy jim také říkáme prediktory), mění se hodnoty parciálních regresních koeficientů. Mění se tím více, čím více jsou vysvětlující proměnné korelované - mění se i jejich střední chyby a tím i výsledky testů. Test příslušného parciálního regresního koeficientu tedy testuje, zda testovaná proměnná přináší **v rámci daného souboru proměnných** ještě významné množství informace k vysvětlení variability závisle proměnné, ostatními proměnnými nevysvětlené.

Příklad: Na hoře Pektu v Koreji byl prováděn výzkum místního druhu modřínu a změn jeho charakteristik na gradientu nadmořské výšky. Jednou z otázek bylo, zda se mění tvar stromu s nadmořskou výškou. Za charakteristiku tvaru stromu byla považována tzv. alometrická rovnice:

$$\log(\text{výška stromu}) = b_0 + b_1 \log(\text{dbh})$$

Vz. 14-8

dbh je průměr kmene v prsní výšce (*diameter at breast height*), běžná lesnická charakteristika. Důkazem, že se tato rovnice (a tedy i tvar stromu) mění s nadmořskou výškou je, když v rovnici

$$\log(\text{výška stromu}) = b_0 + b_1 \log(\text{dbh}) + b_2 \text{ nadm.výška}$$

Vz. 14-9

vyjde hodnota parciálního regresního koeficientu b_2 průkazně odlišná od nuly. Protože tato hodnota vyšla průkazně záporná, můžeme konstatovat, že se stoupající nadmořskou výškou se mění tvar stromu; při stejném průměru kmene jsou stromy nižší. Naproti tomu zjištění, že regresní koeficient *b* v rovnici

$$\log(\text{výška stromu}) = b_0 + b_1 \text{ nadm.výška}$$

Vz. 14-10

je průkazně odlišný od nuly (záporný) říká pouze, že výška stromů s nadmořskou výškou klesá. Toto zjištění je triviální, celková velikost stromů s nadmořskou výškou klesá (podobně klesá i dbh s nadmořskou výškou; prediktory ve Vz. 14-9 jsou tedy vzájemně korelovány). Hodnota koeficientu b_2 ve Vz. 14-9 se liší od hodnoty koeficientu b_1 ve Vz. 14-10.

Parciální korelace

Představme si, že studujeme vzájemné závislosti mnoha proměnných, řekněme X_i, X_j, X_k, X_l. První možností je spočítat korelační koeficienty pro každou dvojici proměnných.Uděláme-li to pro všechny dvojice, dostaneme tzv. korelační matici, případně takto spočtené korelace doplnit testy významnosti. Upozorněme, že se jedná o hladiny významnosti pro každý dílčí test a pokud spočteme korelaci pro dostatečně velký soubor proměnných, velmi pravděpodobně najdeme hodnoty některých korelačních koeficientů průkazně odlišné od nuly jako důsledek chyby prvního druhu, to znamená i tehdy, jsou-li proměnné v základním souboru nekorelované (je zde určitá analogie s mnohonásobnými porovnáními v analýze variance). Výpočet korelační matice je také podkladem pro některé složitější statistické metody.

Výpočet korelačních koeficientů pro dvojice ovšem nemůže postihnout interakce vyššího řádu. K tomu nám mohou sloužit parciální korelační koeficienty (*partial correlation coefficients)*. Vyjadřují vzájemnou závislost dvou proměnných za předpokladu, že další proměnná (proměnné) se nemění. Tak např. $r_{ij.k}$ vyjadřuje vzájemnou závislost proměnných X_i a X_j za předpokladu, že X_k se nemění.

Význam parciální korelace si můžeme představit pomocí výpočtu její hodnoty poněkud netradičním způsobem. Spočteme nejprve regresi X_i na X_k, a vypočteme reziduály. Poté uděláme totéž pro závislost X_j na X_k. Normální korelační koeficient mezi reziduály z prvé a ze druhé regrese je parciálním korelačním koeficientem $r_{ij.k}$. Lze tedy říci, že to je korelační koeficient po „odfiltrovaní“ vlivu třetí proměnné. Protože je zde závislost podmíněná jednou proměnnou, mluvíme o parciálním korelačním koeficientu prvního řádu; kdybychom chtěli podmínit korelaci dvěma proměnnými, jednalo by se o koeficient druhého řádu a značili bychom jej $r_{ij.kl}$. Hodnota parciálního korelačního koeficientu $r_{ij.k}$ je průkazně odlišná od nuly právě tehdy, když je od nuly průkazně odlišný parciální regresní koeficient příslušný k proměnné X_j v regresi X_i na X_j a X_k, nebo (což je ekvivalentní) pokud je od nuly průkazně odlišný parciální regresní koeficient příslušný k proměnné X_i v regresi X_j na X_i a X_k.

Obecné lineární modely a analýza kovariance

Obecné lineární modely (*general linear models*) jsou poměrně rozsáhlou oblastí statistických modelů, které použijeme, chceme-li vysvětlit závislost spojité proměnné na kategoriálních proměnných (potom se jedná o analýzu variance), spojitých proměnných (lineární regrese) nebo na směsi kategoriálních a spojitých proměnných. Lineární regrese i analýza variance jsou jen zvláštními případy obecných lineárních modelů. Společným principem těchto metod je rozklad celkové variability závislé proměnné (charakterizované celkovou sumou čtverců) na části, vysvětlitelné jednotlivými vysvětlujícími proměnnými, případně jejich interakcemi, a na variabilitu nevysvětlenou.

Nyní se zmíníme o možnostech vyhodnocení úloh, kde je vysvětlujícími proměnnými směs kvantitativních a kategoriálních proměnných. Například studujeme závislost hrubé primární produkce (GPP) společenstva na nadmořské výšce (kvantitativní spojitá proměnná); je ovšem pravděpodobné, že tato závislost se bude měnit na různých matečných horninách. Musíme tedy zavést další vysvětlující proměnnou, a tou bude typ matečné horniny. Tato proměnná bude kategoriální: například užijeme - podle typu prostředí, ve kterém pracujeme - tři kategorie: bazické sedimenty (vápenec); kyselé horniny (žuly, ruly); vyvřeliny (uvedené členění jistě není vyčerpávající ani přesné). Budeme tedy studovat závislost GPP na dvou vysvětlujících proměnných, z nichž jedna je kategoriální, druhá spojitá kvantitativní.

Zvláštním případem zobecněných lineárních modelů je **analýza kovariance** (*analysis of covariance*, ANCOVA). Používá se tehdy, když chceme primárně dokázat vliv jedné nebo více kategoriálních proměnných, ale víme, že odpověď je ovlivněna i kvantitativní proměnnou, kterou nejsme schopni z pokusu eliminovat. Například studujeme vliv kompetice na plodnost rostliny (10 rostlin pěstujeme v prostředí bez kompetice, 10 v prostředí s kompeticí). Již na začátku pokusu ale rostliny měly různou výšku. V těchto případech je vždy výhodné zaznamenat výšku každé rostliny na začátku pokusu (dříve, než mohla být ovlivněna kompeticí!) a tu potom použít jako pomocnou vysvětlující proměnnou (*covariate* nebo *covariable*) v analýze kovariance - velmi často to výrazně sníží nevysvětlenou variabilitu a tím vede ke zvýšení síly testu hlavního efektu (v našem příkladu efektu kompetice). V analýze kovariance nejprve „odečteme“ vliv spojité proměnné a poté testujeme, jestli se jednotlivé skupiny mezi sebou liší.

Celou proceduru si můžeme představit tak, že předpokládáme, že závislost na spojité proměnné je ve všech skupinách lineární a má stejný sklon (předpoklad stejného sklonu lze i testovat). Potom testujeme, zda parametr b_0 (tj. průsečík s osou) je u všech skupin stejný. Typickým příkladem může být následující úloha: Porovnáváme váhy mužů ve věku 40 - 45 let, kteří nadměrně pijí pivo, s těmi, kteří pivo (víceméně) nepijí. Pokud bychom úlohu počítali t-testem nebo ekvivalentní jednocestnou analýzou variance, dostaneme velmi slabý test, protože variabilita materiálu bude obrovská, budeme zde mít muže od 150 do 210 cm a váha výrazně závisí i na výšce. Proto je vhodné spočítat analýzu kovariance a výšku užít jako *covariable*. Tím se nám nevysvětlená variabilita zmenší a test bude mnohem silnější. Závislost váhy na výšce je přitom triviální a příliš nás nebude zajímat.

Srozumitelný přehled analýzy kovariance podávají Snedecor a Cochran (1967), str. 419 - 446 a Sokal & Rohlf (1981).

Příkladová data

Příkladová data na mnohonásobnou regresi jsou v listu *Chap14* ve sloupcích A až C. Proměnná *PoaBiom.L* představuje nadzemní biomasu lipnice luční (*Poa pratensis*) na vrcholu sezóny, dvě charakteristiky půdní chemie (*N.tot.L* a *PO4.L*) jsou ale doplněny dvěma dalšími (*NH4.L* a *NO3.L*) pro ilustraci metody postupného výběru vysvětlujících proměnných. Všechny tyto charakteristiky (spolu s nadzemní biomasou lipnice) jsou logaritmicky transformovány (tato transformace se ukazuje jako vhodná ve většině případů užití koncentrací, objemů či ploch), původní data jsou ve sloupcích *J* až *N*, ale do programů tyto původní sloupce není třeba importovat.

Sloupce F až H obsahují příkladová pro ilustraci analýzy kovariance a odpovídají příkladu zmíněnému výše. Vysvětlujeme zde hmotnost (*Weight*) mužů středního věku nadměrnou konzumací alkoholu (faktor *Drinks* s hodnotami *yes* a *no*), ale také číselnou

kovariátou *Height* (výška muže v cm). Primárně nás zajímá vliv konzumace, závislost hmotnosti na výšce spíše jen předpokládáme.

Jak postupovat v programu Statistica

Mnohonásobná regrese

Model mnohonásobné regrese odhadneme pro naše data ("nafitujeme") pro apriorně zvolené vysvětlující proměnné takto. Z menu zvolíme příkaz *Statistics* | *Multiple Regression* a v zobrazeném dialogovém okně zvolíme tlačítkem *Variables* v levém sloupci (*Dependent var.*) proměnnou *PoaBiom.L* a v pravém sloupci obě vysvětlující proměnné (*N.tot.L* a *PO4.L*). Protože zde nechceme provádět postupný výběr vysvětlujících proměnných, pokračujeme rovnou tlačítkem *OK* a zobrazí se nám přehled výsledků v novém okně (*Multiple Regression Results*). Většina výsledků a postupů pro jejich zobrazení je shodná s tím, co jsme popsali již pro jednoduchou regresi v kapitole 12 a čtenáře tam proto odkazujeme. Po zobrazení hlavních výsledků pomocí tlačítka *Summary: Regression results* vidíme následujícíc tabulku.

Regression Summary for Dependent Variable: PoaBiom.L (Spreadsheet8)
R= ,69580670 R²= ,48414696 Adjusted R²= ,45626302
F(2,37)=17,363 p<,00000 Std.Error of estimate: ,19707

N=40	b*	Std.Err. of b*	b	Std.Err. of b	t(37)	p-value
Intercept			-0,066501	0,176228	-0,377359	0,708063
N.tot.L	0,585875	0,118439	0,613473	0,124018	4,946651	0,000017
PO4.L	0,332337	0,118439	0,130335	0,046449	2,805981	0,007952

Z toho, že oba odhady regresních koeficientů pro proměnné *N.tot.L* a *PO4.L* jsou kladné a průkazně odlišné od nuly (viz sloupce *p-value* a *t(37)* můžeme usuzovat, že jak dostupnost dusíku, tak anorganického fosforu mají pozitivní vztah s nadzemní biomasou lipnice. Musíme si ale uvědomit, že: (a) tato pozitivní korelace není nutně projevem kauzálního vztahu (jde o pozorování nemanipulovaných ploch) – možná, že vyšší biomasa vegetace (umožněná například větší půdní vlhkostí, tu jsme zde neměřili) vede k většímu množství rostlinného opadu, jehož rozklad pak vede k vyšší koncentraci živin; a (b) jde o parciální regresní koeficienty, které jsou ovlivněny přítomností druhé vysvětlující proměnné v modelu (viz diskuse výše). V některých situacích se stává (ne v tomto případě, kdy je korelace mezi *N.tot.L* a *PO4.L* poměrně malá), že regresní koeficient vysvětlující proměnné z jednoduché regrese má opačné znaménko (a implikuje tedy opačný vztah) než koeficient téže proměnné v modelu mnohonásobné regrese.

Další problém, kterému při interpretaci výsledků mnohonásobné regrese musíme řešit, je ten, že není možné porovnat hodnoty regresních koeficientů (zde například b_1 a b_2) pro posouzení relativní velikosti vlivu obou vysvětlujících proměnných. Je to proto, že tyto regresní koeficienty současně převádějí jednotky vysvětlující proměnné na jednotky proměnné vysvětlované. Pokud bychom například množství fosforečnanů měřili v gramech na gram sušiny půdy (místo stávajících mg na g sušiny), všechny hodnoty by byly tisíckrát menší, a v důsledku by hodnota koeficientu b_2 byla tisíckrát větší. Program Statistica nám proto nabízí odhady koeficientů, které bychom získali v regresi, kde by jak všechny vysvětlující, tak vysvětlovaná proměnná byly nejprve standardizovány na nulový průměr a jednotkovou varianci. Tyto koeficienty jsou zobrazeny ve sloupci b* a umožňují nám konstatovat, že (v rozsahu pozorované variability půdních parametrů) je efekt dusíku skoro dvakrát větší než efekt fosforečnanů.

Podobně jako u jednoduché regrese doporučujeme pro odhadnutý model mnohonásobné regrese zobrazit minimálně jeden graf regresní diagnostiky, s hodnotami reziduálů vynesenými proti predikovaným hodnotám. Ten nám umožní orientačně posoudit jak možné změny variability reziduálů s predikovanou hodnotou (tj. absenci homogenity variance), tak přílišnou jednoduchost zvoleného lineárního modelu: v grafu reziduálů bychom byli schopni najít přetrvávající jasný tvar (*pattern*) závislosti. Tento graf zobrazíme volbou tlačítka *Perform residual analysis* na záložce *Residuals/assumptions/prediction* a následující volbou tlačítka *Predicted vs. residuals* na záložce *Scatterplots* v dialogovém okně *Residual Analysis*. Výsledek ukazuje (nezobrazen), že se variabilita residuálů nijak výrazně nemění a také že specifikace modelu je ve značné shodě s daty. To potvrzuje, že naše rozhodnutí pracovat se všemi proměnnými na logaritmické stupnici bylo správné.

Pokud bychom chtěli model se dvěma vysvětlujícími proměnnými (pro tři a více již máme smůlu) zobrazit, můžeme z menu *Graphs* zvolit *3D XYZ Graphs* a pak buď *Scatterplots*, *Surface Plots* nebo *Contour Plots*; poslední z možností považujeme za nejvíce informativní zobrazení modelu. Po její volbě zadáme proměnné tlačítkem *Variables* (vysvětlovaná tj. závislá proměnná by měla být zadána jako *Z*) a zvolíme volbu *Linear* v položce *Fit*.

Postupný výběr proměnných

Pro postupný výběr proměnných začneme stejně jako v předchozím příkladě (příkaz *Statistics | Multiple Regression*), ale pomocí tlačítka *Variables* zvolíme čtyři vysvětlující proměnné (*N.tot.L* až *NO3.L*) a pak ještě na záložce *Advanced* zaškrtneme volbu *Advanced options*, před zvolením tlačítka *OK*. Zobrazí se nám nové dialogové okno *Model Definition*, kde na záložce *Stepwise* (nebo i *Quick*) zvolíme *Forward stepwise* v políčku *Method*. Alternativní volbou zde je *Backward stepwise*, které začíná z regresního modelu se všemi možnými vysvětlujícími proměnnými a z něj odebírá nejméně významné proměnné. Tento postup ale v případě analýzy exploratorních dat (tj. výsledků pozorování, typicky s velkou nabídkou proměnných) nedoporučujeme. Na záložce *Stepwise* můžeme zvolit také kritéria, která musí kandidátní vysvětlující proměnná splnit, aby byla do modelu v určitém kroku přidána (*F to enter*), případně z něj odebrána (*F to remove*). Nabídka programu Statistica v tomto modulu *Multiple Regression* ale není moc šťastná, kritériem by neměla být absolutní hodnota testové statistiky, spíše hodnota pravděpodobnosti chyby I. typu nebo nějaké kritérium úspornosti modelu (Statistica nabízí obojí při postupném výběru u obecných lineárních modelů). Zde můžeme pro ilustraci zvolit hodnotu *F to enter* rovnou 2.0. Abychom viděli průběh výběru proměnných, je dále třeba změnit volbu *Display results* na *At each step*.

Po zvolení tlačítka *OK* se nám zobrazí okno *Multiple Regression Results*, ale v jeho bílém poli vidíme modrý nápis *Step 0: No variables in the regression equation*, tj. počáteční stav modelu, před výběrem první vysvětlující proměnné. Po té, co v tomto okně zvolíte tlačítko *Next*, procedura odhadne čtyři modely jednoduché regrese, každý s jednou z kandidujících vysvětlujících proměnných, a vybereme ten, který má největší hodnotu F, pokud je tato hodnota větší než námi dříve zvolené *F to enter*. Pro naše data je vybrána proměnná *N.tot.L* a okno *Multiple Regression Results* se znovu zobrazí s tlačítkem *Next*. Nyní bude procedura zkoumat, která ze tří zbývajících vysvětlujících proměnných by vedla k nejlepšímu modelu se dvěma proměnnými a testuje významnost této změny pomocí parciálního F testu (porovnávajícího zvýšení modelové sumy čtverců s hodnotou reziduální sumy čtverců výchozího modelu, ovšem po standardizaci do podoby *mean squares*). Po opětovné volbě tlačítka *Next* se ukáže, že druhá nejlepší vysvětlující proměnná je *PO4.L*, optimální model se dvěma prediktory je tedy tentýž, který jsme si dříve vybrali a priori.

Tlačítko *Next* se ale teď změnilo na *OK*, což nám ukazuje, že ten lepší ze dvou modelů se třemi vysvětlujícími proměnnými již není dostatečně dobrý, alespoň na základě námi zvoleného kriteria *F to enter*, takže současný model je již finální výsledkem výběru. S tímto modelem pak můžeme pracovat stejně, jak jsme si ukázali v předchozí sekci. Musíme si ovšem uvědomit, že výše popsaným způsobem jsme vybrali nejlepší prediktory ze všech nabídnutých. Je tedy pravděpodobné, že vybraný model vysvětlí průkazně více než očekáváme v případě platnosti nulové hypotézy, a to i v případě, že nulová hypotéza platí. Toto nebezpečí je tím větší, čím byl větší soubor proměnných, ze kterých jsme touto procedurou vybírali.

Parciální korelace

K výpočtu parciálních korelací můžeme použít příkaz *Statistics* | *Basic Statistics/Tables* a v zobrazeném seznamu zvolíme *Correlation matrices*. V dialogovém okně *Product-Moment and Partial Correlations* pak zvolíme tlačítko *Two lists* a v levém seznamu zadáme proměnné, jejichž vzájemné korelace chceme počítat (například proměnné *PoaBiom.L* a *N.tot.L*), a v druhém seznamu proměnné, jejichž vliv na korelované proměnné má být odečten (například proměnné *NH4.L* a *NO3.L*). Výsledek, který pak zobrazíme pomocí tlačítko *Partial correlations* na záložce *Advanced*, je pak následující

Partial Correlations (Spreadsheet8)
Controlling for: NH4.L and NO3.L
Marked correlations are significant at p < ,05000
N=40 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	PoaBiom.L	N.tot.L
PoaBiom.L	0,832650	0,267254	1,000000	0,593975
N.tot.L	1,402900	0,255231	0,593975	1,000000

Hodnota 0.594 představuje parciální korelaci mezi biomasou lipnice a celkovým obsahem dusíku v půdě při kontrole (lineárního) vlivu proměnných představujících dostupnost anorganických forem dusíku.

Analýza kovariance

Z menu zvolíme příkaz *Statistics* | *Advanced Linear/Nonlinear Models* | *General Linear Models* a z následně zobrazené nabídky zvolíme *Analysis of covariance*. Pomocí tlačítka *Variables* zadáme proměnnou *PoaBiom.L* jako *Dependent* (variable), *Drinks* jako *Categorical pred*(ictor) a proměnnou *Height* jako *Continous pred*(ictor). Po volbě tlačítka *OK* se objeví okno s výsledky (*GLM Results*) a vlastní výsledky ANCOVA modelu si můžeme nejsnáze zobrazit tlačítkem *All effects* na záložce *Quick*.

Univariate Tests of Significance for Weight (Spreadsheet8)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 6,026

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	692,028	1	692,028	19,05290	0,003294
Height	1809,834	1	1809,834	49,82832	0,000201
Drinks	937,110	1	937,110	25,80051	0,001432
Error	254,250	7	36,321		

Vidíme, že efekt konzumace piva má průkazný efekt na váhu, průkazný (a výraznější) efekt má i výška člověka. Čtenáři doporučujeme zkusit, jak by test vlivu konzumace piva dopadl, pokud bychom údaje o výšce postavy neměli nebo je ignorovali, protože výsledek je

velmi instruktivní. Test můžeme provést buď pomocí jednocestné analýzy variance nebo (protože faktor *Drinks* má jen dvě kategorie) pomocí dvouvýběrového t-testu. Efekt konzumace piva zde není průkazný ($F_{1,8}$=1.72, p=0.227): rozdíly mezi pijany a ostatními jsou zastíněny variabilitou ve velikosti těla, kterou jsme zde ignorovali.

Pokud bychom chtěli zkoumat, zda je změna hmotnosti s výškou odlišná pro konzumenty piva a pro kontrolní skupinu, můžeme po volbě příkazu *Statistics* | *Advanced Linear/Nonlinear Models* | *General Linear Models* zvolit proceduru *Homogeneity-of-slopes model*, zaměřenou na test paralelismu přímek. Tlačítkem *Variables* zvolíme *Weight* jako *Dependent*, *Drinks* jako kategoriální a *Height* jako spojitý prediktor a po volbě tlačítka *OK* můžeme v okně výsledků zvolit tlačítko *All effects*, které nám zobrazí tabulku rozkladu modelové sumy čtverců s řádkem *Drinks*Height*, který představuje test interakce mezi faktorem *Drinks* a numerickou proměnnou *Height*. Interakce není průkazná ($F_{1,6}$=2.61, n.s.), změna hmotnosti s výškou se tedy mezi oběma skupinami neliší.

Jak postupovat v programu R

Dva příklady užité v této kapitole jsou importovány do samostatných datových rámců (*chap14a* pro sloupce A až E a *chap14b* pro sloupce F až H).

Mnohonásobná regrese

Apriorně zvolený model s vysvětlujícími proměnnými *N.tot.L* a *PO4.L* odhadneme pomocí funkce *lm*:

```
> lm.1 <- lm(PoaBiom.L~N.tot.L+PO4.L,data=chap14a)
```

Pro zobrazení hlavních výsledků použijeme funkci *summary*:

```
> summary(lm.1)
Call:
lm(formula = PoaBiom.L ~ N.tot.L + PO4.L, data = chap14a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48315 -0.11565  0.03657  0.14820  0.29407

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06650    0.17623  -0.377  0.70806
N.tot.L      0.61347    0.12402   4.947 1.66e-05 ***
PO4.L        0.13033    0.04645   2.806  0.00795 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1971 on 37 degrees of freedom
Multiple R-squared:  0.4841,    Adjusted R-squared:  0.4563
F-statistic: 17.36 on 2 and 37 DF,  p-value: 4.805e-06
```

Kromě zobrazení odhadnutých regresních koeficientů, odhadů standardní chyby a dílčích t-testů je zobrazen i celkový F test modelu a dvě varianty koeficientu determinace, spolu s jednoduchým shrnutím regresních reziduálů.

Program R nabízí i alternativní F testy pro jednotlivé vysvětlující proměnné, rozklad vysvětlené sumy čtverců mezi jednotlivé proměnné je ale prováděn podle jejich pořadí v modelu (viz diskusi v kapitole 10, sekce *Jak postupovat v programu R*).

```
> anova(lm.1)
Analysis of Variance Table
```

```
Response: PoaBiom.L
          Df  Sum Sq Mean Sq F value   Pr(>F)
N.tot.L    1 1.04284 1.04284 26.8523 8.03e-06 ***
PO4.L      1 0.30578 0.30578  7.8735 0.007952 **
Residuals 37 1.43694 0.03884
```

Základní graf regresní diagnostiky můžeme zobrazit funkcí *plot*

```
> plot(lm.1,which=1)
```

Postupný výběr proměnných

Pro postupný výběr vysvětlujících proměnných je dobré si nejprve definovat výchozí (nulový) model a také popis maximálního rozsahu vysvětlujících proměnných v modelu (tzv. *scope*):

```
> lm.0 <- lm(PoaBiom.L~+1,data=chap14a)
> lm.scope <- ~N.tot.L+PO4.L+NH4.L+NO3.L
```

Pokud bychom chtěli provádět výběr "ručně", můžeme jednotlivé kroky, testující varianty změny aktuálního stavu modelu, provádět pomocí funkce *add1*. Ta normálně porovnáva alternativní modely pomocí kritéria úspornosti modelu (*model parsimony*), tzv. AIC (*Akaike information criterion*). Chceme-li tedy použít parametrický výběr složitosti modelu, musíme přidat parametr *test* do volání funkce:

```
> add1(lm.0,lm.scope,test="F")
Single term additions

Model:
PoaBiom.L ~ +1
        Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              2.7856 -104.58
N.tot.L  1   1.04284 1.7427 -121.34 22.7392 2.728e-05 ***
PO4.L    1   0.39832 2.3872 -108.75  6.3405   0.01613 *
NH4.L    1   0.24334 2.5422 -106.23  3.6374   0.06408 .
NO3.L    1   0.12156 2.6640 -104.36  1.7339   0.19580
```

Výsledek nám ukazuje, že nejlepším kandidátem na vysvětlující proměnnou je proměnná *N.tot.L* a do modelu ji můžeme přidat takto:

```
> lm.2 <- update(lm.0, .~.+N.tot.L)
```

Další rozšíření nového modelu pak testujeme tímto příkazem:

```
> add1(lm.2,lm.scope,test="F")
Single term additions

Model:
PoaBiom.L ~ N.tot.L
       Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>              1.7427 -121.34
PO4.L   1  0.305777 1.4369 -127.06  7.8735 0.007952 **
NH4.L   1  0.054763 1.6880 -120.61  1.2004 0.280319
NO3.L   1  0.067331 1.6754 -120.91  1.4870 0.230403
```

... a protože je parciální efekt *PO4.L* stále průkazný, přidáme i tuto proměnnou do modelu. Nové volání funkce *add1* ale ukazuje, že zbylé dva prediktory již další smysluplný příspěvek do modelu nevnesou:

```
> lm.3 <- update(lm.2, .~.+PO4.L)
> add1(lm.3,lm.scope,test="F")
Single term additions

Model:
```

```
PoaBiom.L ~ N.tot.L + PO4.L
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              1.4369 -127.06
NH4.L   1 0.0017191 1.4352 -125.10  0.0431 0.8367
NO3.L   1 0.0112441 1.4257 -125.37  0.2839 0.5974
```

Pokud bychom vybírali pomocí kriteria AIC, výsledný model by byl (v tomto příkladě, ne obecně!) shodný, jak ukazují i hodnoty AIC zobrazené ve výstupech funkce *add1*. Při porovnání více modelů pomocí kriteria AIC netestujeme žádnou hypotézu, ale vybíráme model s nejmenší hodnotou kriteria (v našem případě tedy s nejnižší zápornou hodnotou). Použití kriteria AIC ilustrujeme na příkladu automatického výběru proměnných do modelu, pomocí funkce *step*. Ta zobrazuje jednotlivé kroky výběru:

```
> step(lm.0,lm.scope)
Start:  AIC=-104.58
PoaBiom.L ~ +1

          Df Sum of Sq    RSS     AIC
+ N.tot.L  1   1.04284 1.7427 -121.34
+ PO4.L    1   0.39832 2.3872 -108.75
+ NH4.L    1   0.24334 2.5422 -106.23
<none>                 2.7856 -104.58
+ NO3.L    1   0.12156 2.6640 -104.36

Step:  AIC=-121.34
PoaBiom.L ~ N.tot.L

          Df Sum of Sq    RSS     AIC
+ PO4.L    1   0.30578 1.4369 -127.06
<none>                 1.7427 -121.34
+ NO3.L    1   0.06733 1.6754 -120.91
+ NH4.L    1   0.05476 1.6880 -120.61
- N.tot.L  1   1.04284 2.7856 -104.58

Step:  AIC=-127.05
PoaBiom.L ~ N.tot.L + PO4.L

          Df Sum of Sq    RSS     AIC
<none>                 1.4369 -127.06
+ NO3.L    1   0.01124 1.4257 -125.37
+ NH4.L    1   0.00172 1.4352 -125.10
- PO4.L    1   0.30578 1.7427 -121.34
- N.tot.L  1   0.95029 2.3872 -108.75

Call:
lm(formula = PoaBiom.L ~ N.tot.L + PO4.L, data = chap14a)

Coefficients:
(Intercept)      N.tot.L        PO4.L
    -0.0665       0.6135       0.1303
```

Protože jsme výsledný model neukládali do žádného objektu, je jeho obsah jen stručně zobrazen na konci výstupu.

Parciální korelace

Parciální korelace lze spočítat pomocí funkce *pcor* v knihovně *ppcor*. Například zadání příkazu

```
> pcor(chap14a)
$estimate
            PoaBiom.L   N.tot.L      PO4.L       NH4.L       NO3.L
PoaBiom.L  1.00000000 0.4575827  0.3534303  0.02691278 -0.08576814
N.tot.L    0.45758269 1.0000000  0.2338523  0.58204942  0.47326235
```

```
PO4.L      0.35343029 0.2338523  1.0000000 -0.46915262 -0.24131528
NH4.L      0.02691278 0.5820494 -0.4691526  1.00000000 -0.08649805
NO3.L     -0.08576814 0.4732624 -0.2413153 -0.08649805  1.00000000

$p.value
            PoaBiom.L      N.tot.L       PO4.L        NH4.L       NO3.L
PoaBiom.L 0.000000000 2.330443e-03 0.025405650 8.734516e-01 0.610550501
N.tot.L   0.002330443 0.000000e+00 0.154752280 2.288695e-05 0.001481269
PO4.L     0.025405650 1.547523e-01 0.000000000 1.672859e-03 0.141259552
NH4.L     0.873451566 2.288695e-05 0.001672859 0.000000e+00 0.607493537
NO3.L     0.610550501 1.481269e-03 0.141259552 6.074935e-01 0.000000000

$statistic
           PoaBiom.L  N.tot.L     PO4.L      NH4.L      NO3.L
PoaBiom.L  0.0000000 3.044530  2.235178  0.1592758 -0.5092878
N.tot.L    3.0445299 0.000000  1.422944  4.2346877  3.1783283
PO4.L      2.2351782 1.422944  0.000000 -3.1428950 -1.4711167
NH4.L      0.1592758 4.234688 -3.142895  0.0000000 -0.5136545
NO3.L     -0.5092878 3.178328 -1.471117 -0.5136545  0.0000000

$n
[1] 40

$gp
[1] 3

$method
[1] "pearson"
```

spočetlo a zobrazilo parciální korelace mezi všemi páry proměnných v datovém rámci *chap14a*, s vyloučením vlivu zbývajících tří proměnných. Pokud bychom chtěli např. spočíst parciální korelaci mezi biomasou lipnice a celkovým obsahem dusíku, s vyloučením vlivu jen dvou proměnných *NH4.L* a *NO3.L*, musíme proměnnou *PO4.L* z výpočtu vyloučit.

```
> pcor(chap14a[,-3])
$estimate
           PoaBiom.L   N.tot.L       NH4.L       NO3.L
PoaBiom.L  1.0000000 0.5939745 -0.16813504 -0.18842636
N.tot.L    0.5939745 1.0000000  0.55010528  0.44177347
NH4.L     -0.1681350 0.5501053  1.00000000  0.03117279
NO3.L     -0.1884264 0.4417735  0.03117279  1.00000000
...
```

Výsledná hodnota parciální korelace je tedy 0.594. Funkce *pcor* je schopna spočítat parciální korelace i pro neparametrické regresní koeficienty, typ korelace můžeme zadat parametrem *method*.

Analýza kovariance

Statistické modely v programu R nerozlišují mezi použitím kvantitativních a kategoriálních vysvětlujících proměnných a obecný lineární model lze tedy zadat jak pomocí funkce *aov*, tak ve funkci *lm* (jediný rozdíl mezi nimi je v tom, jakou informaci zobrazuje funkce *summary* a možnost specifikace dalších zdrojů variability pomocí členu *Error* ve funkci *aov*). Náš příklad analýzy kovariance si ilustrujeme za použití funkce *lm*:

```
> lm.4 <- lm(Weight~Height+Drinks,data=chap14b)
> anova(lm.4)
Analysis of Variance Table
Response: Weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
Height     1 1315.54 1315.54  36.219 0.0005325 ***
Drinks     1  937.11  937.11  25.800 0.0014320 **
Residuals  7  254.25   36.32
```

V testu proměnné *Height* opět vidíme výsledek numericky odlišný od výsledku v programu Statistica (viz diskuse v kapitole 10), test pro faktor *Drinks* je ale shodný. Díky použití funkce *lm* získáme snadno i hodnoty koeficientů pomocí funkce *summary*

```
> summary(lm.4)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -154.1595    33.4987  -4.602 0.002478 **
Height         1.2690     0.1798   7.059 0.000201 ***
Drinksyes     20.3514     4.0066   5.079 0.001432 **
```

... která nám ukazuje, že jak rostoucí výška, tak konzumace piva zvyšují očekávanou hmotnost zkoumané osoby.

Pro test paralelismu přímek, představujících změnu hmotnosti osob s jejich výškou pro skupiny konzumentů a nekonzumentů piva, odhadneme model s interakcí mezi oběma prediktory a buď jej srovnáme s modelem *lm.4* nebo prostě užijeme funkci *anova*, která bude interakci testovat jako poslední a poskytne tak ekvivalentní výsledek:

```
> lm.5 <- lm(Weight~Height*Drinks,data=chap14b)

> anova(lm.4,lm.5)
Analysis of Variance Table

Model 1: Weight ~ Height + Drinks
Model 2: Weight ~ Height * Drinks
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      7 254.25
2      6 177.25  1    76.995 2.6062 0.1576

> anova(lm.5)
Analysis of Variance Table

Response: Weight
              Df  Sum Sq Mean Sq F value    Pr(>F)
Height         1 1315.54 1315.54 44.5304 0.0005483 ***
Drinks         1  937.11  937.11 31.7208 0.0013407 **
Height:Drinks  1   76.99   76.99  2.6062 0.1575710
Residuals      6  177.25   29.54
```

Popis metod v článku

Methods

We have estimated the effect of total soil nitrogen and phosphate concentration in soil on the total aboveground biomass of *Poa* using multiple regression. Both the response variable and the explanatory variables were log-transformed to achieve additivity of their effects and to increase homogeneity of variances in the response variable.

Parsimonious model predicting *Poa* biomass was selected using stepwise selection based on Akaike information criterion (Akaike 1978).

nebo

... stepwise selection using partial F-test and acceptance threshold α=0.05 at each step.

The effect of beer drinking on the body weight of examined men was tested using analysis of covariance (ANCOVA) with the total height used as a covariate.

We have checked for a difference in body weight increase with subject height betwen the drinking and nondrinking group using a F test of the interaction between height and drinking status in a general linear model.

Results

Fitted regression model is summarized in Table X. Both the total N and phosphate concentrations had significant positive relations with the aboveground biomass of *Poa*.

Table X: Multiple linear regression model of the relation of *Poa* biomass (log-transformed) to log-transformed total N content and PO_4^{3-} concentration in soil. The partial *t* tests of individual predictors were done with df=37. Absolute term estimate (b0) was -0.0665.

Model term	b_i	$se(b_i)$	t	P
Total N	0.613	0.124	4.95	<0.001
PO_4^{3-}	0.130	0.046	2.81	0.008

We have found a significant positive effect of beer drinking on the body weight ($F_{1,7}$=25.8, p=0.0014) alongside the significant (positive) effect of height ($F_{1,7}$=49.8, p=0.0002).

We have found no evidence for a difference of body weight change with increasing height between the two groups of subjects ($F_{1,6}$=2.6, n.s.).

Doporučená četba

Zar (1984) pp.328-360, Sokal & Rohlf (1981) pp. 509-529 (ANCOVA), pp. 617-641 (mnohonásobná regrese), Quinn & Keough (2002) pp. 111-142 (mnohonásobná regrese, postupný výběr vysvětlujících proměnných), pp. 339-352 (ANCOVA).

Analýza kovariance: Snedecor G.W. & Cochran W.G. (1967) Statistical Methods. Sixth Edition. Iowa State University Press, Ames. pp. 419 - 446.

Akaike H. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**: 9-14.

15 Nelineární závislost

Lineární regrese je bezesporu nejčastěji užívanou metodou studia závislosti vysvětlované proměnné (odpovědi) na vysvětlujících proměnných. Přitom ale můžeme rozumně předpokládat, že mnoho (asi většina) biologických závislostí je nelineární. Jakým způsobem můžeme takovou nelineární odezvu hodnotit? Jedna z možností již byla ukázána: linearizovat vztah pomocí transformace proměnných. Další možností je užití polynomiální regrese, užití klasické nelineární regresi minimalizující součet čtverců odchylek, užití metod zobecněných lineárních modelů (v jistém smyslu jde o rozšíření transformačního přístupu, ale i u nich lze použít polynomy pro vysvětlující proměnné), nebo můžeme data „vyhladit“, tj. použít tzv. *smoothing* (např. metoda loess nebo metody tzv. zobecněných aditivních modelů, *generalized additive models*). V této kapitole probírame pouze metody polynomiální a nelineární regrese.

Pokud při regresní analýze zjistíme, že závislost dvou proměnných je nelineární, ale rozptyl hodnot kolem předpokládané závislosti je přibližně konstantní, máme několik možností, jak pokračovat. Jednou z oblíbených je užití polynomiální regrese. Regresní rovnice má (v případě, kdy mám jen jednu vysvětlující proměnnou X) tvar:

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_m X^m$$

Vz. 15-1

odhad parametru β_0 značíme b_0, odhady parametrů β_i značíme b_i. Prokládáme tedy data polynomem m-tého stupně; v angličtině se tento postup někdy nazývá *polynomial curve fitting*, obecně se často mluví spíše o prokládání křivek (*curve fitting*), než o regresi. Obr. 15-1 ukazuje, jak jsou tatáž data prokládána postupně polynomy vyšších a vyšších stupňů.

Takovéto křivky nám často jsou schopny velmi hezky proložit naše data, interpretovat jednotlivé koeficienty je ovšem často obtížné nebo i zhola nemožné, zvláště u polynomů vyšších stupňů. V biologické praxi se používá regrese kvadratická (proložení polynomem druhého stupně), vzácněji kubická (polynom třetího stupně). V případě modelování vztahu velikosti (či *fitness*) populace k vlastnostem prostředí představuje kvadratická regrese v kombinaci s logaritmováním vysvětlované proměnné známý model jednovrcholové (unimodální) odezvové křivky (*unimodal response curve*) a odhadnuté regresní koeficienty lze pak transformovat do dvou biologicky smysluplnějších parametrů: pozice optima druhu (kde je hodnota křivky nejvyšší) a tolerance druhu (šířka křivky). Optimum spočteme z regresních koeficientů takto:

$$x_{opt} = \frac{-b_1}{2.b_2}$$

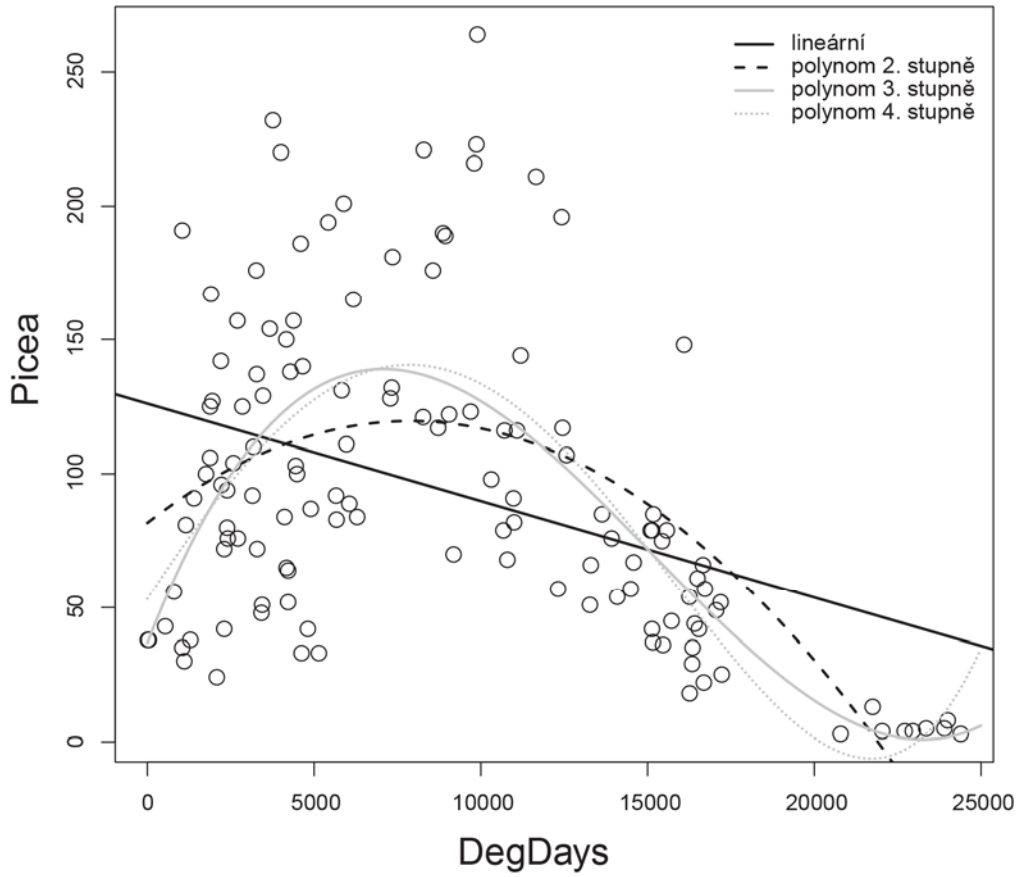
Vz. 15-2

ale musíme být připraveni na to, že to v některých případech může být minimum, spíše než maximum křivky (pokud je $b_2 > 0$), tedy pozice „sedla“ spíše než „vrcholu“ kopce.

Rovnice ve Vz. 15-1 velmi připomíná rovnici pro mnohonásobnou regresi – a skutečně, polynomiální regrese je zvláštním případem mnohonásobné regrese, kde za jednotlivé nezávislé proměnné dosazujeme postupně první, druhou, atd. mocninu jediné nezávisle proměnné. Obdobně jako v mnohonásobné regresi můžeme testovat, zda jednotlivé koeficienty jsou průkazně odlišné od nuly. Postupujeme přitom tak, že začínáme s nejnižším počtem členů - tedy s lineární regresí - a sledujeme, zda přidáním vyššího členu významně zlepšíme proložení. Pokud je kvadratický člen průkazně liší od nuly, znamená to, že závislost není lineární, což může být biologicky zajímavé. Interpretovat biologicky kubický člen může

být obtížné; pokud studujeme odezvu druhu na gradient prostředí, znamená průkazný kubický člen, že křivka odezvy není symetrická kolem optima druhu. Polynom alespoň třetího stupně je třeba použít, pokud předpokládáme, že funkce má inflexní bod. Je take třeba poznamenat, že v případě kvadratické závislosti, kdy je maximum nebo minimum přibližně uprostřed rozsahu nezávislé proměnné, lineární regrese nebude průkazná, přestože budeme mít těsnou kvadratickou závislost. Proto je vždy třeba si závislost vynést do XY diagramu a podle tvaru závislosti naznačeného uspořádáním bodů se rozhodnout o vhodném postupu.

Mnohonásobná polynomiální regrese umožňuje odhadovat závislost jedné proměnné na více polynomech různých nezávislých proměnných, přičemž pro různé proměnné lze použít polynomy různého stupně. To se používá například při popisu trendů v prostorovém rozložení hodnot vysvětlované proměnné, kdy používáme polynomy geografických souřadnic jednotlivých pozorování (tzv. *trend surface analysis*). Pro odhady parametrů opět užíváme kriteria nejmenších čtverců (v obecném případě má vždy jedno minimum). Analýzu variance můžeme používat stejně jako pro mnohonásobnou regresi.



Obr. 15-1 Prokládání dat polynomiálními regresními modely s polynomy různého stupně. Data představují počet pylových zrn ve vzorku, získaného v různých geografických oblastech, které jsou (na horizontální ose) charakterizovány veličinou *degree days* (součet pozitivních denních průměrných teplot za celý rok). Zobrazené regresní modely používají vysvětlovanou proměnnou bez transformace, ale logaritmická transformace by zde jistě prospěla mimo jiné i k omezení předpovídaných hodnot (na původní škále) na pozitivní čísla. Všimněte si také nerealistického vzestupu kubické křivky pro vysoké hodnoty DegDays.

Nelineární regrese

V některých případech známe teoreticky očekávaný tvar závislosti a často dokonce známe významy jednotlivých parametrů. Například pro popis rychlosti fixace uhlíku při fotosyntéze

v závislosti na koncentraci CO_2 se často užívá asymptotická křivka s posunem (*asymptotic curve with an offset*)

$$FR = b_0\{1 - e^{-e^{b_1}(x-b_2)}\}$$

Vz. 15-3

kde *FR* je rychlost fixace C, *x* je koncentrace CO_2 ve vzduchu a parametry b_i představují asymptotickou hodnotu rychlosti fixace (b_0), logaritmus rychlosti nárůstu fixace se vzestupem koncentrace CO2 (b_1) a hodnotu offsetu (b_2), tj. minimální koncentrace CO_2, od které začne fixace C probíhat. Pro odhad parametrů takovéto křivky můžeme použít nelineární regresi. Pokud použijeme kriterium nejmenších čtverců, potom hledáme takové hodnoty parametrů, pro které je reziduální součet čtverců nejmenší. Na rozdíl od lineární regrese toto minimum nejsme schopni nalézt v obecném případě jinak, než numericky. Používají se různé iterační metody, jsou to metody numerické matematiky, které systémem směrovaný pokus - korekce - nový pokus - nová korekce atd. postupně hledají polohu minima součtu čtverců, tj. takovou kombinaci hodnot parametrů, pro kterou je součet čtverců nejmenší.

Metod odhadu je více, žádná není dokonalá. Největší nevýhodou je, že na rozdíl od lineární a polynomiální regrese může mít funkce závislosti velikosti součtu čtverců na hodnotách parametrů mnoho lokálních minim. Všechny procedury jsou schopny nějaké minimum najít, problém je ovšem v tom, že žádná procedura není schopna zjistit, zda se jedná skutečně o globální minimum. Většina programů proto vyžaduje pro obecné zadání nelineárního modelu iniciální hodnoty odhadu parametrů, ale pro konkrétní funkce dokážou některé programy tyto odhady spočíst samy. Ve většině metod se využívají hodnoty derivace funkce, a proto je některé programy také vyžadují (jiné si spočítají derivaci numericky).

Podobně jako pro lineární regresi, i zde je předpokladem užití metody stálá hodnota variance. Tento předpoklad bývá u nelineárních modelů ale často narušen. I ve shora uvedeném příkladu je variance zřejmě nejvyšší v optimu, se vzdáleností od optima klesá. V mnoha případech pomohou transformace. Jinou možností je užít jiného kriteria než nejmenších čtverců (většinou kriteria maximální věrohodnosti). ¨

Příkladová data

Příkladová data v listu *Chap15* (prvé dva sloupce) představují měření závislosti rychlosti fixace uhlíku rostlinou (proměnná *C.uptake*) na koncentrací oxidu uhličitého ve vzduchu (proměnná *CO2*). 21 rostlin bylo pěstováno v sedmi různých ambientních koncentracích CO_2 (v každé koncentraci 3 rostliny) za stejné teploty, vzdušné vlhkosti a světelných podmínek a určena rychlost fixace. Takové pokusy se obvykle odehrávají v omezeném počtu klimaboxů a rostliny sdílející stejný klimabox (např. se shodnou koncetrací CO_2) pak není možné považovat za nezávislá pozorování, ale předpokládajme zde, že tomu tak v našem příkladě není. Tato data použijeme jak na ilustraci polynomiální regrese, tak pro ilustraci odhadu nelineárního modelu metodou nejmenších čtverců (což je bezesporu vhodnější přístup než polynomiální regrese, i s ohledem na biologickou interpretovatelnost parametrů zvoleného nelineárního modelu.

Jak postupovat v programu Statistica

Polynomiální regrese

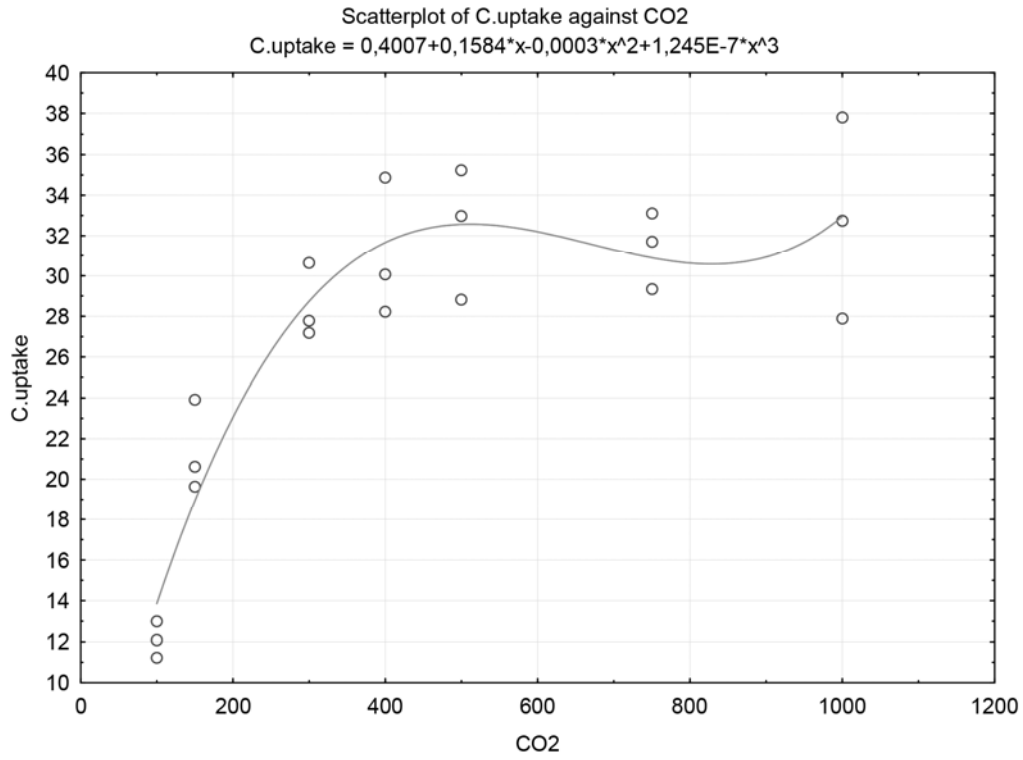
Z menu zvolíme příkaz *Statistics* | *Advanced Linear/Nonlinear Models* | *General Linear Models* a v následné nabídce pak položku *Polynomial regression*. Pomocí tlačítka *Variables* vybereme v levém seznamu (*Dependent variable list*) proměnnou *C.uptake* a v pravém seznamu (*Predictor variables*) *CO2*. Pomocí tlačítka *Between effects* zvolíme složitost polynomu – s ohledem na nesymetričnost závislosti (asymptotický růst k maximální rychlosti fixace C) nebude polynom druhé stupně dostatečný, zvolíme tedy druhou možnost (*Use polynomials to specified degree*) a změním hodnotu ze 2 na 3. Pak pokročíme z původního dialogového okna pomocí tlačítka *OK*. V dialogovém okně *GLM Results* můžeme odhady parametrů modelu zobrazit například tlačítkem *Univariate results*:

Univariate Results for Each DV (Spreadsheet2)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	C.uptake SS	C.uptake MS	C.uptake F	C.uptake p
Intercept	1	0,131	0,1314	0,01500	0,903953
"CO2"	1	297,069	297,0691	33,92392	0,000020
"CO2"^2	1	153,201	153,2009	17,49484	0,000625
"CO2"^3	1	100,282	100,2817	11,45171	0,003528
Error	17	148,868	8,7569		
Total	20	1180,120			

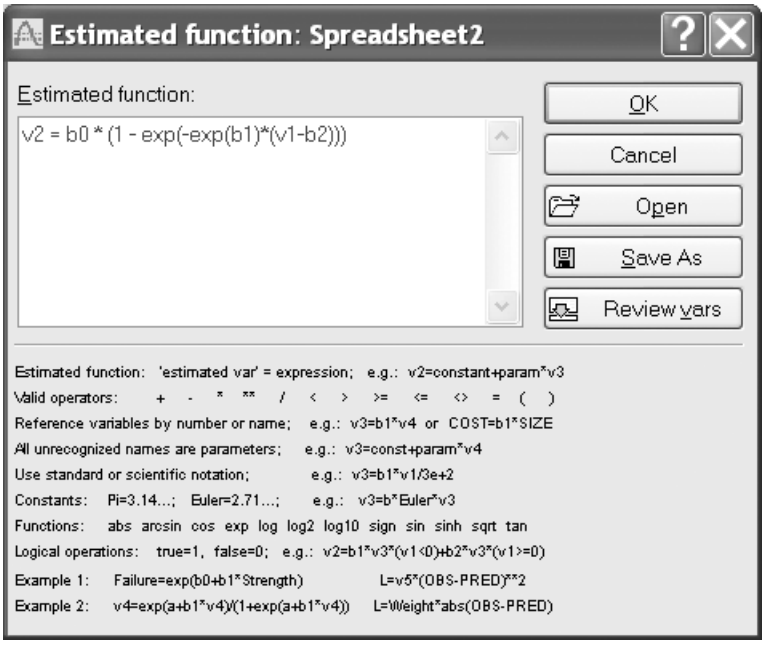
Statistica zobrazuje testy jednotlivých polynomiálních členů modelu, odhadnuté hodnoty regresních koeficientů můžeme zobrazit pomocí tlačítka *Coefficients* (odhad koeficientu pro třetí mocninu CO2 není nulový (jak ukazuje výše uvedená tabulka, je dokonce statisticky vysoce průkazně odlišný od nuly), jen je menší než zobrazené přesnost, protože odpovídající hodnoty jsou třetí mocniny hodnot v řádu stovek, takže odpovídající koeficient je nutně velmi malý). Test celého modelu lze najít ve výstupu zobrazeném tlačítkem *Whole model R*, společně s koeficientem determinace ukazujícím, že model vysvětlil 87% z celkové variability. V této proceduře ale nemůžeme jednoduše zobrazit nafitovaný model.

K tomu musíme zvolit příkaz *Graphs* | *2D Graphs* | *Scatterplots*, pomocí tlačítka *Variables* zvolit *CO2* pro osu X a *C.uptake* pro osu *Y*, a nakonec na záložce *Advanced* zvolíme v políčku *Fit* volbu *Polynomial* a na záložce *Options 2* zvolíme hodnotu *Cubic* pro položku *Polynomial order*. Na výsledném obrázku vidíme fit tohoto polynomu pro naše data, prohnutí křivky směrem dolů pro vyšší koncentrace CO_2 ovšem biologické realitě neodpovídá. Zde take vidíme, že koeficient pro třetí mocninu je $1.24 \cdot 10^{-7}$.



Nelineární regrese

Pro odahd nelineárního modelu (konkrétně toho popsaného ve Vz. 15-3) zvolíme z menu příkaz *Statistics* | *Advanced Linear/Nonlinear Models* | *Nonlinear Estimation* a v seznamu vybereme položku *User-specified regression, least squares*. V dialogovém okně klikneme na tlačítko *Function to be estimated* a zadáme rovnici modelu:



Zavřeme toto okno tlačítkem *OK* a postoupíme dále opět tlačítkem *OK*. V dialogovém okně *Nonlinear Least Squares Model Estimation* ponecháme volbou metody odhadu *Levenberg-Marquardt*, ale na záložce *Advanced* musíme zvolit počáteční odhady modelu, s implicitními hodnotami 0.1 pro všechny tři parametry by odhad modelu selhal. Klikneme na

tlačítko *Start values*. Odhad rozumných počátečních hodnot máme usnadněn naší znalostí jejich významu. Pohledem na rozmístění bodů (pozorování) v XY diagramu výše vidíme, že limitní hodnota, ke které rychlost fixace uhlíku roste je někde kolem 35 (*b0*) a že pro nejnižší pozorovanou koncentraci oxidu uhličitého (100) je hodnota fixace kolem 12, ale směrem k nižším hodnotám rychle padá, takže lze předpokládat, že pozitivní se stává kolem 50 (*b2*). Těžší je odhadnout rychlost růstu, respektive její logaritmus. Necháme tedy hodnotu parametru *b1* na nabízené hodnotě 0.1 a upravíme jen hodnoty *b0 a b2*.

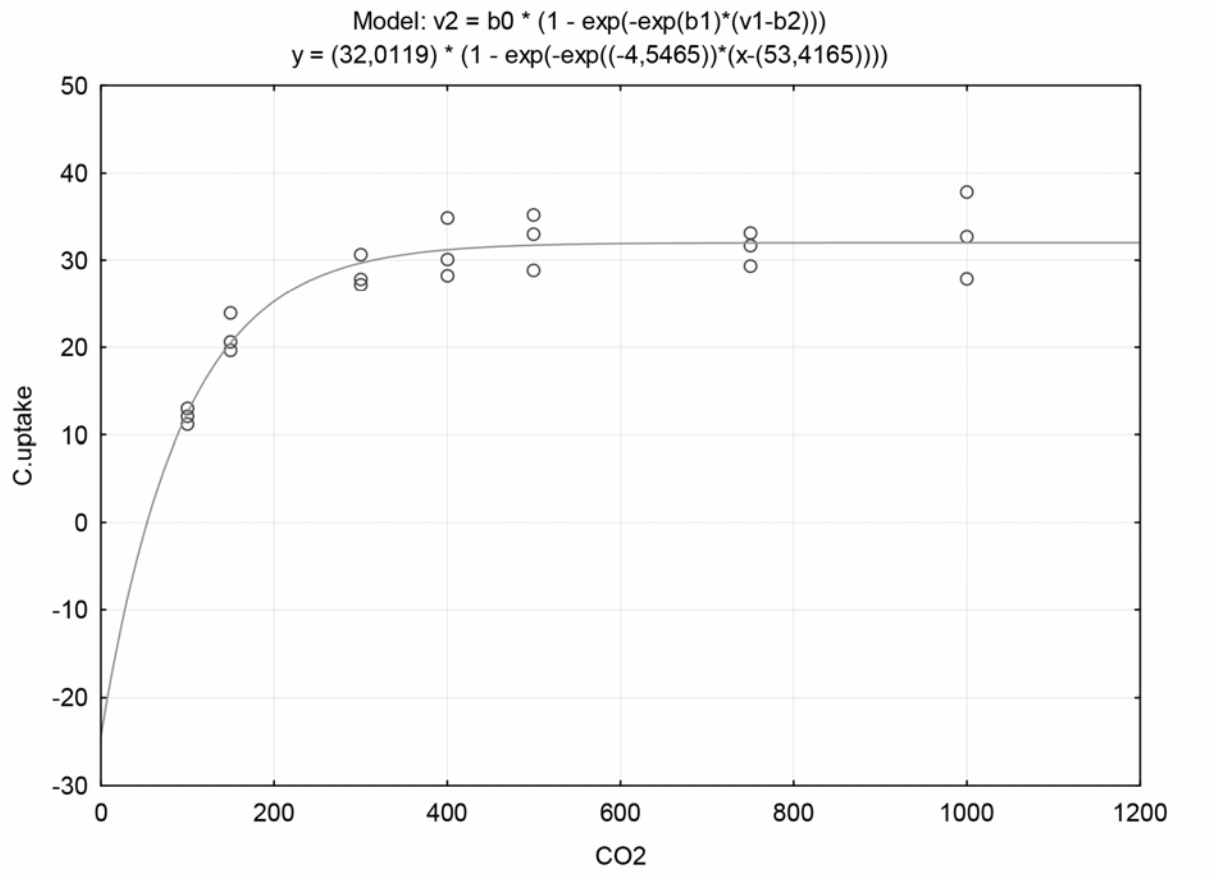
Po uzavření okna pro specifikaci odhadů a volbě tlačítka *OK* se objeví dialogové okno *Results*. Po té, co zvolíme tlaíčtko *Summary* ale vidíme, že odhad neproběhl dobře (viz poznámka *Caution: Degenerate solution ...* v záhlaví tabulky, ale také nulové odhady pro standardní chyby odhadů *b1* a *b2*). Zkusíme tedy lépe odhadnout počáteční hodnotu parametru *b1*. Uzavřeme tedy okno *Results* pomocí *Cancel* a v původním dialogovém okně rovnou zmáčkneme tlačítko *OK* a na záložce *Advanced* opět zvolíme tlačítko *Start values*. Zvolíme tentokrát menší hodnotu -2, implikující, že růstová rychlost (nelogaritmovaná) je menší než jedna. Tentokrát dává tlačítko *Summary* v okně *Results* smysluplnější výsledky:

Model is: v2 = b0 * (1 - exp(-exp(b1)*(v1-b2))) (Spreadsheet2)
Dep. Var. : C.uptake
Level of confidence: 95.0% (alpha=0.050)

	Estimate	Standard error	t-value df = 18	p-value	Lo. Conf Limit	Up. Conf Limit
b0	32,01186	0,87557	36,5611	0,000000	30,17235	33,85137
b1	-4,54646	0,25330	-17,9489	0,000000	-5,07862	-4,01429
b2	53,41646	15,57406	3,4298	0,002987	20,69657	86,13635

Statistica nezobrazuje přímo informaci o množství vysvětlené variability, ale klasický koeficient determinace R^2 lze dopočítat z údajů zobrazených tlačítkem *Analysis of Variance*. Je ale třeba dát pozor na to, jaké hodnoty použijeme. Modelová suma čtverců v řádku *Regression* není obvyklá MSS, stejně jako není celková suma čtverců v řádku *Total* (TSS; obě sumy představují odlišnost od nulového modelu, ve kterém je hodnota vysvětlované proměnné rovna nula, nikoliv průměru této proměnné). Správná TSS je v řádku *Corrected Total* (1180.12) a správnou modelovou sumu čtverců zjistíme jako rozdíl této hodnoty a hodnoty v řádku *Residual*, tj. 1180.12-130.08=1050.04. Koeficient determinace je tedy roven R^2=1050.04/1180.12=0.8897 (tedy objasněno asi 89.0%).

Po volbě tlačítka *Iteration history* nám tabulka názorně ukazuje, jak byly během iterativního hledání optimálního řešení původní odhady parametrů modifikovány a jak se jejich změna během posledních iterací zmenšovala (došlo ke konvergenci řešení). Užitečnější je ale asi tlačítko *Fitted 2D function & observed vals*, které nám umožňuje nafitovaný model snadno vynést.



Jak postupovat v programu R

Polynomiální regrese

Model s polynomem třetího stupně odhadneme a sumarizujeme v programu R takto:

```
> lm.1 <- lm(C.uptake~poly(CO2,3),data=chap15)
> summary(lm.1)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.0857     0.6458  41.944  < 2e-16 ***
poly(CO2, 3)1  24.4695     2.9592   8.269 2.32e-07 ***
poly(CO2, 3)2 -18.2268     2.9592  -6.159 1.05e-05 ***
poly(CO2, 3)3  10.0141     2.9592   3.384  0.00353 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.959 on 17 degrees of freedom
Multiple R-squared:  0.8739,    Adjusted R-squared:  0.8516
F-statistic: 39.25 on 3 and 17 DF,  p-value: 7.362e-08
```

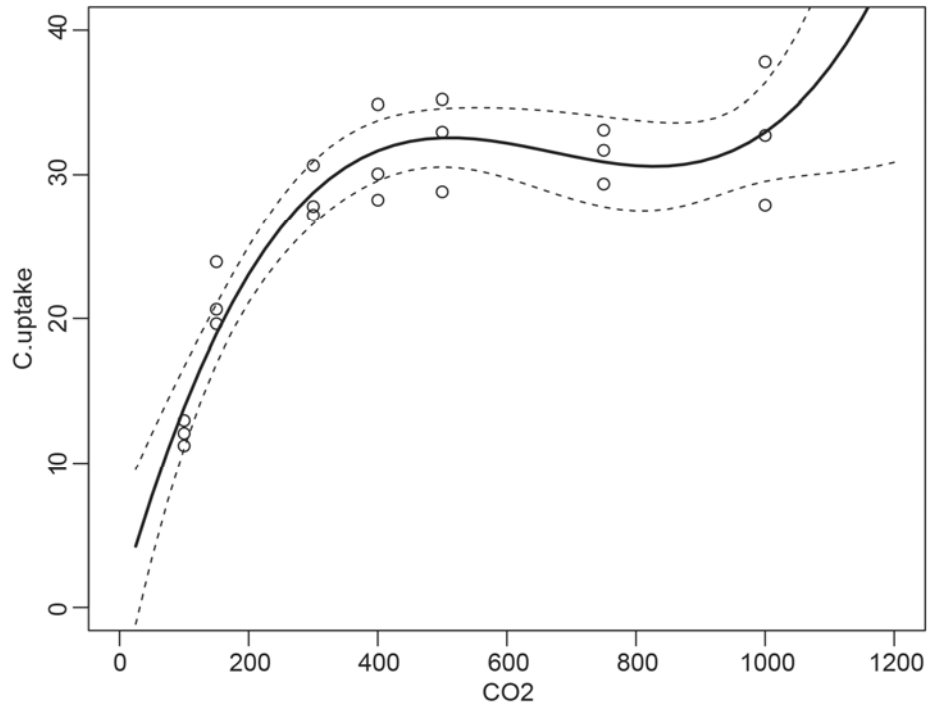
Všimněme si ale, jakým způsobem jsou členy polynomu zapsány. Odpovídající tři vysvětlující proměnné nelze interpretovat jako „původní hodnoty CO2“, „druhé mocniny CO2“ a „třetí mocniny CO2“, protože užití členu *poly* ve vzorci modelu způsobilo, že tato trojice prediktorů byla nahrazena jinou, která vysvětlí stejné množství variability a předpovídá shodné hodnoty *C.uptake*, ale hodnoty těchto tří prediktorů nejsou vzájemně korelovány: jde o tzv. ortogonální polynom (*orthogonal polynomial*). Pokud bychom chtěli získat koeficienty, se kterými lze snadno předpovídat z koncentrace CO_2 rychlost fixace uhlíku, museli bychom model fitovat takto:

```
> lm.2 <- lm(C.uptake~CO2+I(CO2^2)+I(CO2^3),data=chap15)
```

Tato varianta ale není vhodná např. pro testy jednotlivých parametrů modelu.

Nafitovaný model můžeme (spolu s 95% regionem spolehlivosti) zobrazit takto:

```
> plot(C.uptake~CO2,data=chap15,xlim=c(0,1200),ylim=c(0,40))
> lines(xpred$CO2,lm.1fit$fit,lwd=2)
> lines(xpred$CO2,lm.1fit$fit-2*lm.1fit$se.fit,lty=2)
> lines(xpred$CO2,lm.1fit$fit+2*lm.1fit$se.fit,lty=2)
```



Nelineární regrese

Parametry nelineárních regresních modelů můžeme odhadovat poměrně snadno pomocí funkce *nls*, která je součástí knihovny *nlme*. To ale platí pouze v případě, kdy jde o jednu z nelineárních funkcí, které mají v knihovně přímou podporu tzv. *self-starting* funkcí (jejich názvy začínají písmeny *SS*). V takovém případě tyto funkce poskytnou dobré počáteční odhady parametrů pro daná data a při odhadu modelu jsou užívány známé derivace modelové funkce. Pro ostatní případy musíme vytvořit vlastní funkci, která jednak vrací fitovanou funkce, jednak její derivaci, a také dodat vlastní počáteční odhady parametrů. Zde budeme ilustrovat ale jen tu jednodušší variantu, protože rovnice ve Vz. 15-3 je podporována funkcí *SSasympOff*:

```
> nls.1 <- nls(C.uptake~SSasympOff(CO2,b0,b1,b2),data=chap15)
> summary(nls.1)

Formula: C.uptake ~ SSasympOff(CO2, b0, b1, b2)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0  32.0119     0.8756   36.56  < 2e-16 ***
b1  -4.5465     0.2533  -17.95 6.18e-13 ***
b2  53.4158    15.5749    3.43  0.00299 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

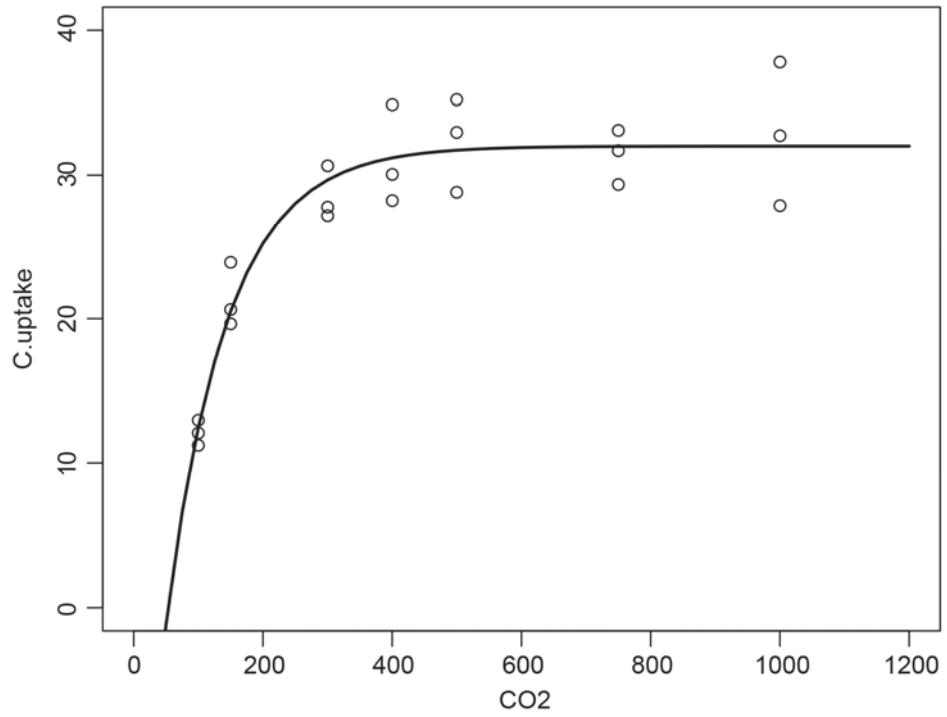
```
Residual standard error: 2.688 on 18 degrees of freedom

Number of iterations to convergence: 0
Achieved convergence tolerance: 6.864e-07
```

Koeficient determinace není pro nelineární model počítán, ale zjistíme ho poměrně snadno jako druhou mocninu korelace mezi předpovídanými a skutečnými hodnotami vysvětlované proměnné:

```
> cor(predict(nls.1),chap15$C.uptake)^2
[1] 0.8897759
```

Graf s původními hodnotami a fitovanou funkcí vytvoříme takto:

```
> xpred <- data.frame(CO2=seq(25,1200,by=25))
> nls.fit <- predict(nls.1,xpred)
> plot(C.uptake~CO2,data=chap15,xlim=c(0,1200),ylim=c(0,40))
> lines(xpred$CO2,nls.fit,lwd=2)
```



Popis metod v článku

Methods

The nonlinear nature of the change of C fixation rate with ambient CO_2 concentration was summarized using a third order polynomial model.

We have used a nonlinear model of asymptotic growth curve with offset (see Formula 15-3) to describe the change of C fixation rate with increasing ambiend CO_2 concentration, estimated with least-squares minimizing algorithm of Levenberg and Marquardt (Moré 1977).

... algorithm of Gauss-Newton algorithm (Bates & Chambers 1992).

Results

Uvádíme popis pro nelineární model, prezentace polynomiálního modelu by byla obdobná, odkazované tabulky a grafy zde neopakujeme:

Estimated non-linear model of asymptotic growth (see Figure X) explained about 89% of the variation in C uptake rate and its parameters are summarized in Table Y.

Doporučená četba

Polynomiální regrese: Zar (1984): pp. 361-368; Sokal & Rohlf (1981): pp. 671-683, Quinn & Keough (2002), pp. 133-135.

Nelineární regrese: Quinn & Keough (2002), pp. 150-152.

D.M. Bates & J. M Chambers (1992): Chapter 10 - Nonlinear Models. In: J. M. Chambers & T. J. Hastie (1992): Statistical Models in S, Wadsworth Press, Pacific Grove, p. 421-454.

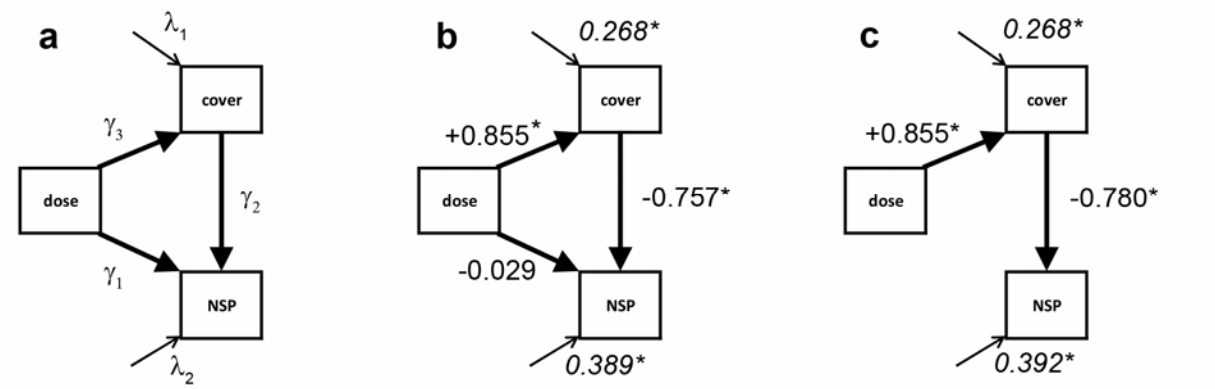
J. J. Moré (1977): The Levenberg-Marquardt Algorithm: Implementation and Theory. In G.A. Watson (ed.), Lecture Notes in Mathematics 630. Berlin: Springer-Verlag, p. 105-116.

16 Modely strukturních rovnic

V této kapitole popíšeme skupinu metod zvanou modely strukturních rovnic (*structural equation modeling* nebo *structural equation models*, *SEM*). Tyto modely nám umožňují popsat systém tří nebo více charakteristik (proměnných, často představujících nějaké biologické procesy), definovat hypotetický model kauzálních vztahů mezi nimi (která charakteristika / který proces ovlivňuje které další) a tento hypotetický model dále testovat a odhadovat jeho parametry na základě jeho porovnání s pozorovanými hodnotami proměnných a zejména s jejich vzájemnými korelacemi, včetně korelací parciálních (viz kapitola 14).

SEM ale umožňují popsat i situace, kdy pro některé procesy či jevy máme jen obecnou představu jejich významu, ale žádná z měřených proměnných tuto představu nevyjadřuje přesně. Jde o tzv. latentní proměnné (*latent variables*), které pak v těchto modelech spojíme s větším počtem měřených proměnných, které je do určité míry charakterizují (každá jinak) a umožňují nám odhadnout vztah latentní proměnné k ostatním, ať již reálně měřeným nebo latentním. Latentní proměnné se vyskytují často v aplikaci SEM v sociologii, etologii či psychologii (charakteristiky jako "schopnost řešit problémy" nebo "míra empatie"), v biologii se ale nejčastěji používá jednodušší forma SEM, ve které jsou všechny proměnné konkrétní a nezastupují žádné obecné koncepty. Tato jednodušší forma, která se nazývá *path analysis* (české jméno pokud víme neexistuje), tedy vychází z apriorního modelu příčinných závislostí mezi našimi proměnnými a odhaduje jeho parametry a snaží se určit, jak pravděpodobný s ohledem na naše data je.

Následující příklad představuje nejjednodušší možný případ SEM, se třemi vzájemně souvisejícími pozorovanými proměnnými a bez jakýchkoliv latentních proměnných. V pokuse (viz Pyšek & Lepš 1991) byl sledován vliv hnojení na druhovou bohatost společenstva plevelů v porostech ječmene. Mimo množství aplikovaného hnojiva a počtu druhů v dané experimentální ploše byla také zaznamenávána pokryvnost ječmene, protože ta (díky zastínění) výrazně ovlivňuje rozvoj plevelů. To je jeden z nejdůležitějších rozdílů této metody oproti mnohonásobné regresi – jedna proměnná (zde pokryvnost ječmene) může vystupovat zároveň v roli ovlivňované i ovlivňující proměnné.



Obr. 16-1 Path analýza vlivu dávky hnojiva a pokryvnosti ječmene na druhovou bohatost plevelového společenstva; **a**: schématické znázorněné kauzálního modelu, **b**: odhad parametrů tohoto modelu (standardizované koeficienty, čísla následovaná hvězdičkou představují průkazné vlivy), **c**: odhad parametrů pro zjednodušený model, ve kterém byl neprůkazný přímý efekt dávky hnojiva na počet druhů vynechán.

Příčinné závislosti mezi proměnnými jsou v Obr. 16-1 značeny šipkou, šipky jdoucí „odnikud“ značí vliv neznámých faktorů (a tedy nevysvětlenou variabilitu příslušných proměnných). Za základní (nezávislou) proměnnou považujeme dávku hnojiva (*dose*). Tato dávka ovlivňuje přímo pokryvnost ječmene (*cover*), ale také druhovou bohatost (*NSP*). Pokryvnost ječmene pak také přímo ovlivňuje druhovou bohatost. Řeckými písmeny u šipek jsou značeny tzv. *path coefficients*; jsou to vlastně standardizované parciální regresní koeficienty a vyjadřují sílu vlivu jedné proměnné na druhou. Koeficienty u „odnikud“ jdoucích šipek vyjadřují sílu vlivu neznámého faktoru (tedy odpovídají nevysvětlené variabilitě). V diagramech na Obr. 16-1 jsou tyto koeficienty zobrazeny ve standardizované podobě (tj. jako by byly spočteny z proměnných normalizovaných na nulový průměr a jednotkovou varianci).

Na Obr. 16-1 b jsou tyto koeficienty odhadnuty (statisticky významné hodnoty jsou označeny hvězdičkou). Je vidět, že nepřímý vliv hnojení na druhovou bohatost přes pokryvnost ječmene je výrazný (a také průkazný), zatímco přímý vliv hnojení na druhovou bohatost průkazný není (a jeho relativní velikost je malá).

Metoda *path analysis* stále není v oblasti biologických věd dostatečně doceněna. Metody popisované dříve v těchto skriptech (zejména analýzy variance a regresní modely) umožňují testovat naše hypotézy o vztazích a také, v případě dat získaných z vhodně definovaných manipulativních experimentů, výsledky těchto testů interpretovat jako podporující či nepodporující naše domněnky o kauzálních vztazích, nicméně jsou v této oblasti omezeny na relativně jednoduché systémy vztahů.

Přestože při užití této metody uvažujeme o příčinných (kauzálních) závislostech, zamítnutí nulové hypotézy nemůžeme považovat za důkaz příčinné závislosti, ale pouze za závislost statistickou (Petraitis et al. 1996). Pro důkaz kauzální závislosti by bylo třeba provést manipulativní experimenty: ve výše uvedeném příkladě bychom experimentálně odstranili biomasu ječmene ve všech variantách dávky hnojení, abychom prokázali její přímý vliv na druhovou bohatost. *Path analysis* je tedy velmi užitečná metoda pro zpracování observačních dat; z nich získáme hypotézy, které jsou testovány experimentálně. Tato metoda se ovšem často užívá i tam, kde není možné experimenty provést, a podle okolností se potom může uvažovat o kauzalitě, i když se o její přímý důkaz nejedná (je relativně populární při různých evolučních úvahách). Pro bližší informace o SEM v kontextu biologických oborů doporučujeme knihu Shipley (2000).

Příkladová data

Příkladová data pocházejí z experimentu, ve kterém byla manipulována dávka hnojiva a byl sledován její vliv na pokryvnost plodiny (ječmene) a také na druhovou bohatost plevelového společenstva. Experiment a apriorní hypotézy jsou blíže popsány na začátku této kapitoly a v Obr. 16-1a. Proměnné v listu *Chap16* souboru s příkladovými daty představují dávku (*dose*), pokryvnost ječmene (*cover*) a počet druhů plevelů (*NSP*).

Jak postupovat v programu Statistica

Ačkoliv program Statistica obsahuje samostatný a poměrně propracovaný modul pro SEM (dostupný z příkazu *Statistics* | *Advanced Linear/Nonlinear Models* | *Structural Equation Modeling*), námi zvolený model jsme v něm nebyli schopni odhadnout s dostatečnou spolehlivostí parametrů (vyjádřenou velikostí standardních chyb odhadů). Proto praktické užití SEM demonstrujeme jen pro program R v následující sekci.

Jak postupovat v programu R

V programu R existuje několik knihoven implementujících nezávisle SEM analýzy, zde si ukážeme práci s knihovnou *sem*.

Začneme s modelem popsaným v Obr. 16-1a. Popis tohoto diagramu (označovaného často jako *path diagram*) zadáme následovně (krom příkazu v první řádce musíme také zadat text uvedený za jednotlivými kombinacemi čísla a dvojtečky, za poslední s číslem 6 rovnou zmáčkneme *Enter*, čímž zadávání ukončíme):

```
> sem.mod1 <- specifyEquations()
1: NSP = gam1 * dose + gam2 * cover
2: cover = gam3 * dose
3: V(cover) = lam1
4: V(NSP)   = lam2
5: V(dose)  = 1
6:
Read 5 items
```

Model strukturních rovnic pak odhadneme pomocí funkce *sem*:

```
> sem.1 <- sem( sem.mod1,data = chap16)
```

Výsledný model můžeme stručně charakterizovat pomocí funkce *summary*:

```
> summary(sem.1)

 Model Chisquare =  24.8828   Df =  1 Pr(>Chisq) = 6.092372e-07
 AIC =  34.8828
 BIC =  20.07877

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.9840 -3.6620 -1.8240 -0.5105  2.5690  3.1570

 R-square for Endogenous Variables
   NSP  cover
0.6115 0.7316

 Parameter Estimates
     Estimate     Std Error    z value   Pr(>|z|)
gam1  -0.08501071  0.316893104 -0.268263 7.884968e-01 NSP <--- dose
gam2  -0.06174312  0.008924294 -6.918544 4.563088e-12 NSP <--- cover
gam3  30.37217053  1.672389178 18.160947 1.051952e-73 cover <--- dose
lam1 338.42315289 43.509328422  7.778175 7.357848e-15 cover <--> cover
lam2   3.26131810  0.419290936  7.778175 7.357848e-15 NSP <--> NSP

 Iterations =  0
```

Z testů nulových hypotéz rovnosti jednotlivých koeficientů nule je vidět, že jediný koeficient *gam1* (reprezentující přímý vliv hnojení na druhovou bohatost) není průkazný. Revidujeme proto původní model a tuto přímou vazbu vyloučíme:

```
> sem.mod2 <- specifyEquations()
1: NSP = gam2 * cover
2: cover = gam3 * dose
3: V(cover) = lam1
4: V(NSP)   = lam2
5: V(dose)  = 1
6:
Read 5 items


> sem.2 <- sem(sem.mod2,data=chap16)
```

Oba modely můžeme ještě porovnat pomocí LRT (*likelihood-ratio test*):

```
> anova(sem.1,sem.2)
LR Test for Difference Between Models

      Model Df Model Chisq Df LR Chisq Pr(>Chisq)
sem.1        1      24.883
sem.2        2      24.939  1  0.05613     0.8127
```

Mezi oběma modely není průkazný rozdíl v nevysvětlené variabilitě, dáme proto přednost jednoduššímu *sem.2*. Na tomto místě je asi vhodné zmínit, že naše modely jsou testovatelné jen díky tomu, že i ten složitější má nenulový počet residuálních stupňů volnosti (přesněji řečeno df=1). V modelu jsme odhadovali 5 parametrů, celkový počet DF ale není roven počtu pozorování, nýbrž počtu (odlišných) hodnot v matici variance a kovariancí mezi třemi proměnnými užívanými v modelu. Těchto hodnot je 6 (tři různé kovariance mezi páry proměnných a tři variance těchto proměnných), takže pokud bychom odhadovali v našem modelu i variabilitu proměnné *dose*, žádný stupeň volnosti by nám nezbyl pro residuální variabilitu. My jsme ale využili skutečnosti, že *dose* je experimentátory nastavený faktor a variabilitu jsme zadali jako fixní (s hodnotou jedna), není tedy odhadována z dat.

Ještě si výsledný model *sem.2* shrneme:

```
> summary(sem.2)

 Model Chisquare =  24.93893   Df =  2 Pr(>Chisq) = 3.842211e-06
 AIC =  32.93893
 BIC =  15.33088

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.9840 -3.6620 -1.7720 -0.5319  2.5340  3.0700

 R-square for Endogenous Variables
 cover    NSP
0.7316 0.6079

 Parameter Estimates
       Estimate     Std Error   z value    Pr(>|z|)
gam2    -0.06334064  0.00462451 -13.696725 1.062083e-42 NSP <--- cover
gam3    30.37217053  1.67238918  18.160947 1.051952e-73 cover <--- dose
lam1   338.42315289 43.50932842   7.778175 7.357848e-15 cover <--> cover
lam2     3.26283133  0.41948548   7.778175 7.357848e-15 NSP <--> NSP

 Iterations =  0
```

Hodnoty koeficientů a odhady variancí jsou na původní škále těchto proměnných, v diagramech pro *path analysis* (viz např. Obr. 16-1c) jsou ale uváděny v hodnotách, které bychom získali použitím proměnných standardizovaných na nulový průměr a jednotkovou varianci, takže lze koeficienty lépe porovnávat. Tyto standardizované koeficienty získáme takto:

```
> coef(sem.2,stand=T)
      gam2       gam3       lam1       lam2
-0.7796839  0.8553363  0.2683998  0.3920930
```

Konfidenční intervaly pro koeficienty na původní škále můžeme získat pomocí funkce *bootSem*:

```
> boot.sem2 <- bootSem(sem.2,R=1000)
> summary(boot.sem2)
Call: bootSem(model = sem.2, R = 1000)

Lower and upper limits are for the 95 percent perc confidence interval
```

```
         Estimate          Bias    Std.Error        Lower        Upper
gam2  -0.06334064  7.231297e-05  0.006046914  -0.07526459  -0.05088769
gam3  30.37217053 -1.302913e-01  2.067418679  26.03369971  34.15688207
lam1 338.42315289 -4.518392e+00 42.271458828 255.92859890 425.09314372
lam2   3.26283133 -7.397739e-02  0.454660991   2.34366019   4.11920179
```

Popis metod v článku

Methods

To model direct and indirect effects of fertilizer dose on weed species richness, we have used structural equation models (SEM, Shipley 2000) estimated in the package *sem* of R software. The model was simplified based on Z-approximation test of path coefficients, eliminating non-significant paths from the model.

Results

Resulting structural equation model is shown in Figure 16-1c. This model represents an adjustment of our original hypothesis, because the direct effect of fertilizer dose was not supported by our data (γ_1=-0.029, Z=-0.268, n.s.).

Doporučená četba a citovaná literatura

Petraitis P.S., Dunham A.E. & Niewiarowski P.H. (1996): Inferring multiple causality: the limitations of path analysis. *Functional Ecology*, **10**: 421-431.

Pyšek P. & Lepš J. (1991): Response of a weed community to nitrogen fertilization: a multivariate analysis. *Journal of Vegetation Science*, **2**: 237-244.

Shipley B. (2000): Cause and correlation in biology. A user's guide to path analysis, structural equations and causal inference. Cambridge University Press, 317 pp.

Sokal R.R. & Rohlf F.J. (1981) pp. 642-656. Quinn & Keough (2002), pp. 145-150.

17 Diskrétní rozdělení a jejich užití; charakteristiky rozmístění v prostoru

V této kapitole probereme dva nejznámější typy diskrétních rozdělení (*discrete distributions*) a jejich užití v praxi. Charakteristika diskrétních rozdělení byla podána v kapitole 1; většinou biolog používá těch diskrétních rozdělení, která mohou nabývat pouze celočíselných hodnot. K nim patří i Poissonovo rozdělení a binomické rozdělení, o nichž bude řeč v této kapitole. Protože se srovnání s Poissonovým rozdělením používá často k testování náhodnosti rozmístění v prostoru, budou spolu s tímto rozdělením probrány i jiné metody popisu prostorového rozmístění.

Poissonovo rozdělení

Příklady:

1. Bylo odchyceno 50 myší a na každé byla provedena analýza ektoparazitů. V jejím rámci byl zjištěn počet klíšťat na každém individuu. Ptáme se, zda jsou klíšťata rozmístěna na myších náhodně; pokud jsou rozmístěna náhodně, znamená to, že všechna individua myší mají stejnou pravděpodobnost, že se jich klíště chytne a že přítomnost jednoho klíštěte nezvýší ani nesníží pravděpodobnost výskytu jiného klíštěte.

2. Na ploše bylo rozmístěno 100 pokusných ploch standardní velikosti. V každé ploše byl zjištěn počet jedinců kruštíku bahenního. Ptáme se, zda jsou jedinci tohoto druhu rozmístěni v ploše náhodně, na sobě nezávisle.

3. Byl zjišťován počet rekombinačních nodů na chromosomových bivalentech. Ptáme se, zda jsou rekombinační nody rozmístěny na bivalentech náhodně (tzn. že existence nodu na bivalentu ani nezvýší, ani nesníží pravděpodobnost vzniku dalšího nodu).

Jedním z nejčastěji užívaných diskrétních rozdělení je Poissonovo rozdělení. Popisuje počet náhodných vzájemně nezávislých jevů v jednotce času nebo prostoru. Poissonovo rozdělení bude mít např. počet bakterií v jednotce objemu vodní suspense. Poissonovu distribuci použijeme pro data, která představují malý objem nebo pokud bude suspenze velmi řídká. Při vysokých hodnotách průměru se Poissonovo rozdělení blíží normálnímu rozdělení.

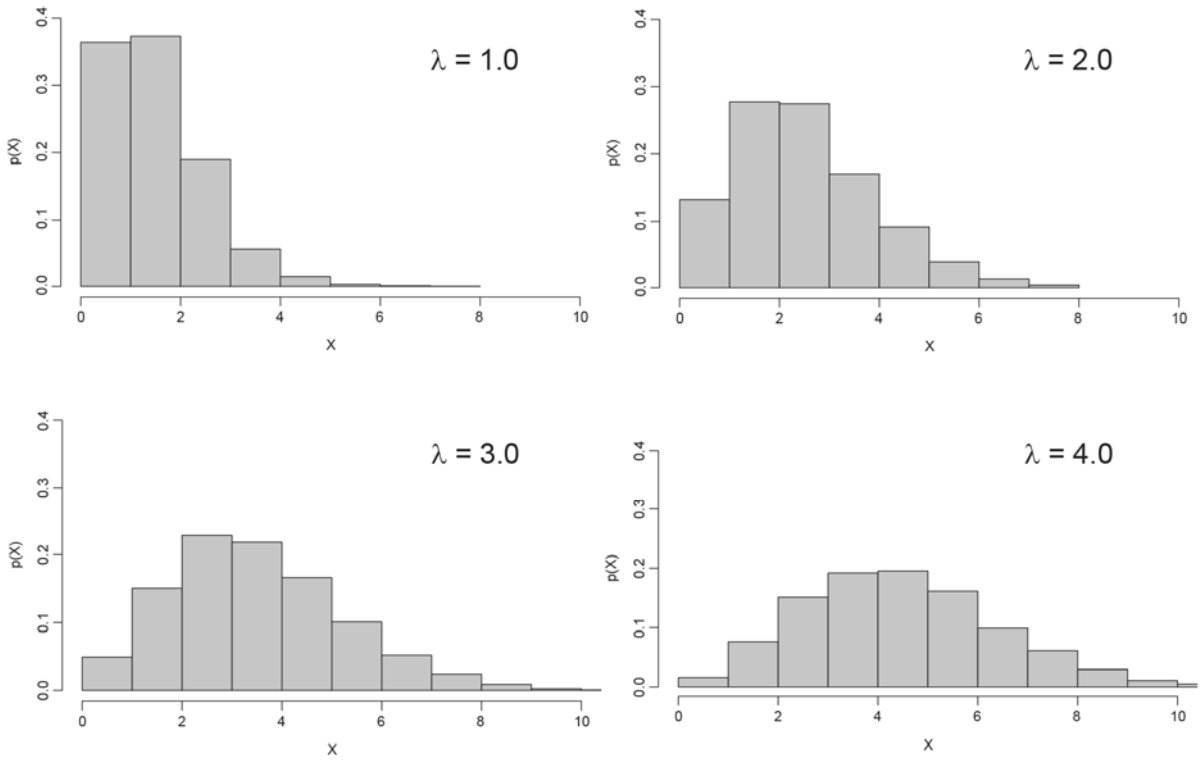
Jiným příkladem může být počet nezávislých kolonizací ostrova za časovou jednotku: předpokládejme, že se určitý druh vyskytuje na pevnině a nevyskytuje se na vzdáleném ostrově. Druh má konstantní pravděpodobnost, že jeho diaspora bude přenesena z pevniny na ostrov (to ještě neznamená úspěšnou kolonizaci). Pokud předpokládáme, že jednotlivé přenosy budou nezávislé, potom bude mít počet přenosů za desetiletí Poissonovo rozdělení. Pokud předpokládáme, že jednotlivá velká zemětřesení jsou nezávislá, potom bude mít počet velkých zemětřesení, která postihnou kontinent za desetiletí také Poissonovu distribuci.

Proměnná, která má Poissonovo rozdělení, může nabývat hodnoty celých nezáporných čísel. Pravděpodobnost, že proměnná x nabude hodnoty X, je dána funkcí

$$P(x = X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$$

Vz. 17-1

X! je faktoriál (např. *4!=4.3.2*, *6!=6.5.4.3.2,* atd.), λ je jediný parametr tohoto rozdělení. Platí, že jak střední hodnota, tak variance tohoto rozdělení jsou rovny λ. Dále platí, že součet dvou navzájem nezávislých proměnných, které mají Poissonovo rozdělení, má také Poissonovo rozdělení. Jestliže tedy má počet velkých zemětřesení za desetiletí Poissonovo rozdělení, potom i počet zemětřesení za půl století má Poissonovo rozdělení (pokud není zemětřesná aktivita v po sobě následujících desetiletích korelována). Obr. 17-1 ukazuje Poissonovo rozdělení pro různé hodnoty parametru λ. Vidíme, že rozdělení je výrazně pozitivně šikmé, zvláště pro nízké hodnoty parametru λ. Pro vyšší hodnoty se začíná blížit normálnímu; uvádí se, že při hodnotách parametu λ vyšších než 10 lze toto rozdělení úspěšně aproximovat rozdělením normálním.



Obr. 17-1 Poissonova distribuce pro různé hodnoty λ.

V praxi se s Poissonovým rozdělením setkáme nejčastěji ve dvou případech:

(1) Pokud víme, že má určitá proměnná Poissonovo nebo Poissonovu blízké rozdělení, víme také, že rozdělení je pozitivně šikmé a že variance není nezávislá na průměru. Šikmost je tím větší, čím nižší je parametr λ. To vadí jak v regresi, tak v analýze variance. V tomto případě se doporučuje odmocninová transformace závislé proměnné (viz kapitola 10).

(2) Test shody s Poissonovým rozdělením se provádí, pokud chceme zjistit, zda určité jevy nastávají náhodně, navzájem nezávisle. Největší tradici má tento postup v ekologii a v parazitologii (viz příklady 1 a 2 výše). Zjišťujeme počet individuí ve zkusných jednotkách a sledujeme, zda tyto počty mají nebo nemají Poissonovo rozdělení. Shoda odpovídá náhodnému rozmístění individuí do jednotek.

Test provádíme nejčastěji pomocí testu dobré shody χ^2: nejprve odhadneme hodnotu parametru λ pomocí výběrového průměru ($\overline{X}$) a dosadíme do Vz. 17-1. Tak získáme

očekávané pravděpodobnosti a vynásobením počtem pozorování i očekávané frekvence. Ty porovnáme pomocí testu dobré shody (χ^2) s frekvencemi skutečnými. Kategorie s nízkými očekávanými frekvencemi obvykle spojujeme. Počet stupňů volnosti je počet kategorií minus dva.

Porovnání variance a průměru

Jinou možností je porovnání hodnot variance a průměru. Poissonovo rozdělení je charakterizováno tím, že variance je rovna střední hodnotě. Pokud se variance průkazně liší od průměru, znamená to, že můžeme zamítnou nulovou hypotézu, že data pocházejí z Poissonova rozdělení. Pokud test používáme proto, abychom charakterizovali rozmístění objektů (individuí) v prostoru, znamená variance větší než průměr shlukovitost a variance menší než průměr rovnoměrnost (až pravidelnost) rozmístění. K testování používáne statistiku

$$\frac{s^2}{\overline{X}}(n-1)$$

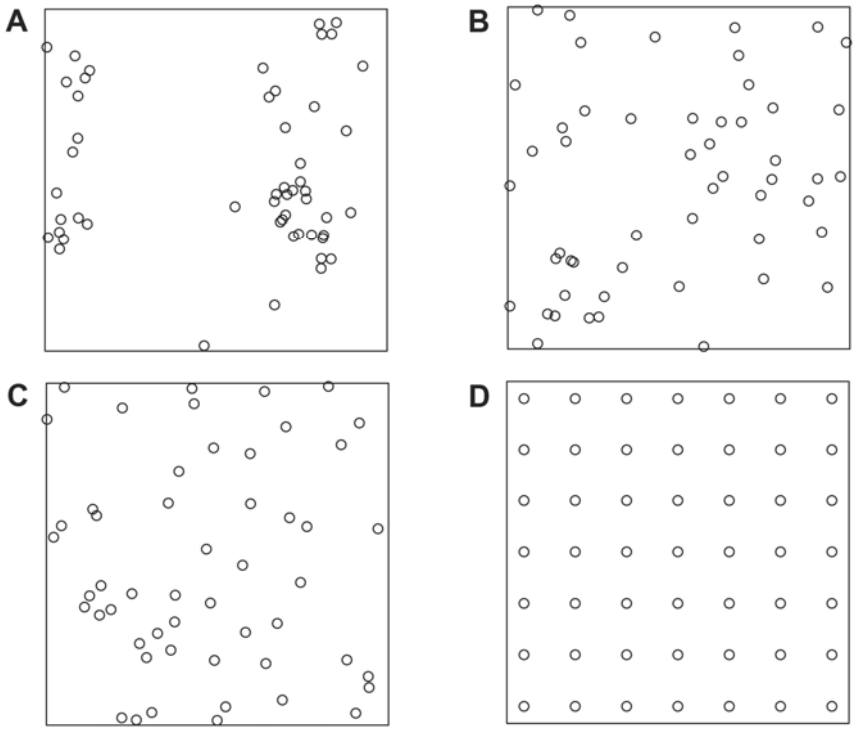
Vz. 17-2

Pokud jsou pozorované hodnoty výběrem ze základního souboru s Poissonovým rozdělením, potom má hodnota statistiky ve Vz. 17-2 přibližně χ^2 rozdělení s *n-1* stupni volnosti (*n* je velikost výběru, tj. počet studovaných zkusných jednotek). Nulovou hypotézu zamítáme, pokud je hodnota spočtená ve Vz. 17-2 menší než $\alpha/2*100$%-ní kvantil (to znamená, že variance je průkazně menší než průměr a individua jsou tedy zřejmě rozmístěna náhodně), nebo pokud je hodnota větší než $(1-\alpha/2)*100$%-ní kvantil (variance větší než průměr, odpovídající shlukovitému rozmístění individuí). Můžeme případně provést i jednostranný test, např. zda je variance průkazně vyšší než průměr (nebo menší než průměr). Zde testujme proti nulové hypotéze, že je variance rovna nebo nižší než průměr (nebo rovna nebo vyšší), kritickou hodnotou pak je $(1-\alpha)*100$%-ní kvantilem (případně $\alpha*100$%-ní kvantil).

Každý z testů (tj. test dobré shody s rozdělením a test založený na porovnání variance a průměru) má své výhody i nevýhody: variance může být rovna průměru i v případě, že se rozdělení výrazně liší od Poissonova – v tom případě se tato odchylka odrazí pouze v testu dobré shody, a nikoliv při porovnání variance a průměru; na druhou stranu porovnání variance a průměru naznačí směr odchylky od náhodnosti (tj. interpretaci výsledků jako shlukovitost či pravidelnost rozmístění studovaných objektů) a umožní užití jednostranného testu.

Rozmístění populace	Odpovídající distribuce počtu individuí ve zkusné jednotce	Poměr variance a střední hodnoty (σ^2/μ) rozdělení	Vzájemný poměr výskytu individuí	Nejčastější ekologické příčiny odchylky od náhodnosti
pravidelné (rovnoměrné)	např. binomické	< 1	Výskyt jednoho individua v jednotce snižuje pravděpodobnost výskytu jiného individua	Vnitrodruhová konkurence, teritoriální chování
náhodné	Poissonovo	1	Výskyt individuí je navzájem nezávislý	
shlukovité	kontagiózní (např. negativně binomické, Neymanovo)	> 1	Výskyt jednoho individua v jednotce zvyšuje pravděpodobnost výskytu jiného individua	způsob rozmnožování, heterogenita prostředí

Tab. 17-1 Typy rozmístění populace a odpovídající statistické a ekologické charakteristiky



Obr. 17-2 Typy rozmístění individuí ve spojitém prostoru: shlukovité (A), náhodné (B), rovnoměrnější (C) a zcela pravidelné ve čtvercové síti (D).

Prokážeme-li, že variance je větší než průměr, obvykle konstatujeme, že jevy (individua) se vyskytují shlukovitě. Naproti tomu, pokud je variance menší než průměr, mluvíme o rovnoměrném nebo pravidelném rozmístění jevů (individuí).* Shlukovité rozmístění jevů může být způsobeno buď tím, že jsou jevy na sobě závislé, nebo tím, že různé jednotky mají různou pravděpodobnost výskytu jevu: např. shlukovité rozmístění parazitů (klíšťat na myších) může být buď důsledkem toho, že některé myši jsou pro klíšťata atraktivnější než jiné, nebo důsledkem toho, že existují místa se zvýšenou intenzitou výskytu

* Rozlišujte rozdělení jako statistický termín a rozmístění jako uspořádání v prostoru. Jedinci v prostoru mají určité rozmístění, počet jedinců ve zkusné jednotce je náhodná proměnná s určitým rozdělením. Vztah těchto ukazuje Tab. 17-1, typy rozmístění Obr. 17-2.

klíšťat, a když myš takovým místem proběhne, obvykle „chytí“ víc než jedno klíště. Obdobně, rostliny mohou být rozmístěny shlukovitě díky způsobu šíření nebo díky tomu, že některá místa v porostu jsou pro ně výhodnější.

V ekologii platí, že většina populací je v prostoru rozmístěna shlukovitě. Počet individuí v jednotce potom odpovídá tzv. kontagiosním rozdělením (nejznámější z nich jsou negativně binomické a Neymannovo). Pokud je rozmístění rovnoměrné až pravidelné, obvykle v ekologických případech uvažujeme o vlivu kompetice mezi sousedícími individui. O vlivu kompetice uvažujeme i tehdy, když v průběhu času intenzita shlukovitosti poklesne. Za míru intenzity shlukovitosti se v těchto případech doporučuje použít tzv. Lloydův index shlukovitosti

$$\frac{\frac{s^2}{\overline{X}} - 1}{\overline{X}} + 1$$

Vz. 17-3

Má tu vlastnost, že pokud individua v jednotlivých zkusných jednotkách vymírají nezávisle na tom, kolik je ve zkusné jednotce individuí, pak se jeho hodnota při vymírání jedinců v populaci nemění. Naproti tomu, pokud individua z jednotek obsazených více individui vymírají rychleji, jeho hodnota klesá. Proto se užívá k zjištění přítomnosti procesů zřeďování závislých na hustotě (*density-dependent processes*). Jednoznačným důkazem by ovšem byl pouze experiment s manipulovanou hustotou populace.

Způsob rozmístění individuí v prostoru se v angličtině označuje *spatial pattern*. Pokud studujeme rozmístění individuí v jednotkách přirozeně definovaných (klíšťata na myších, roztoči v jednotlivých rourkách choroše), nepotřebujeme rozhodovat o velikosti jednotky. Pokud studujeme rozmístění individuí ve spojitém prostoru (např. rostliny v porostu), musíme se rozhodnout pro velikost zkusné jednotky. Je zřejmé, že výsledek naší analýzy bude záviset na tom, jak velkou jednotku použijeme. Pokud individua nejsou rozmístěna náhodně a tvoří shluky určité velikosti, potom je výsledek závislý na velikosti užité zkusné plochy. Existuje celý soubor metod (užívaný hlavně v ekologii) zvaný *spatial pattern analysis*, který hledá velikosti shluků a odhaduje intenzitu shlukovitosti - používá čtverce uspořádané do sítí nebo do transektů.

Míry shlukovitosti založené na vzdálenosti

Jinou možnost, jak charakterizovat rozložení individuí v ploše, nám dává užití vzdálenosti mezi sousedy a vzdálenosti náhodného bodu k nejbližšímu sousedu. Tyto metody jsou založeny na skutečnosti, že pokud jsou individua v ploše rozmístěna náhodně, potom je průměrná vzdálenost od náhodného bodu k nejbližšímu individuu stejná, jako průměrná vzdálenost od náhodně vybraného individua k jeho nejbližšímu sousedu. Toho využívá index *A* Hopkinse a Skellama, definovaný jako

$$A = \frac{\sum r_1^2}{\sum r_2^2}$$

Vz. 17-4

kde r_1 je vzdálenost od náhodného bodu k nejbližšímu individuu a r_2 je vzdálenost od náhodně vybraného individua k jeho nejbližšímu sousedu (připomeňme, že individuum nejbližší k náhodnému bodu není náhodně vybrané individuum). Předpokládáme stejný počet

měření r_1 i r_2. Pokud jsou individua rozmístěna náhodně, je očekávána hodnota $A=1$, pokud shlukovitě, je $A>1$, v případě rovnoměrného rozdělení je $A<1$.

Pokud chceme odchylku od náhodnosti testovat, spočteme nejprve hodnotu $x=A/(A+1)$ a potom proměnnou $Z = 2(x - 0.5)\sqrt{2n+1}$, kde n je počet měřených vzdáleností každého typu. V případě, že jsou individua rozmístěna náhodně a n je dostatečně velké (doporučuje se alespoň 50), má Z normované normální rozdělení (tj. střední hodnota 0, variance 1) a tuto hodnotu lze použít jako testové kriterium.

Jinou užívanou mírou nenáhodnosti je index Clarka a Evanse. Porovnává skutečnou vzdálenost rostliny k nejbližšímu sousedu se vzdáleností očekávanou v případě náhodného rozmístění individuí. Označme hustotu individuí ρ. Potom střední hodnota vzdálenosti k sousedu za předpokladu náhodnosti je

$$\frac{1}{2\sqrt{\rho}}$$

Vz. 17-5

Index

$$R = 2\bar{r}\sqrt{\rho}$$

Vz. 17-6

je tedy poměrem zjištěné ($\bar{r}$) a očekávané vzdálenosti k nejbližšímu sousedu. $R<1$ indikuje shlukovitost, $R>1$ pravidelnost. I tuto hodnotu lze testovat (viz Pielou 1977, p. 155).

Další informace o popisu rozmístění individuí v prostoru lze najít v češtině v práci Lepše (1989). Tyto metody byly určeny pro rychlé stanovení typu rozmístění v terénu, jejich problémem ovšem bylo, že není jednoduché vybrat v terénu náhodné individuum. Lze to provést počítačově v případě, že máme mapu všech individuí ve studované ploše (nebo spíše pro výpočet souřadnice všech individuí). V tom případě můžeme ale o typu rozmístění zjistit mnohem víc, než nám může vypovědět jediná hodnota indexu (proto mají výše uvedené dva indexy dnes již spíše historický význam). Zde doporučujeme užívat analýzu K-funkcí (*K-function analysis*). Tato metoda na základě porovnání počtu „sousedů“ individua očekávaného za předpokladu náhodného rozmístění a počtu skutečného charakterizuje shlukovitost populace. Za „sousedy“ považujeme objekty (jedince) nacházející se do určité vzdálenosti. Tuto vzdálenost měníme a tím dostáváme funkci, která nám popisuje velmi dobře rozmístění individuí v prostoru a umožňuje odhad případné velikosti shluků.

Hodnota K funkce roste pro náhodné rozmístění s druhou mocninou vzdálenosti, ale běžněji se používá linearizovaná podoba nazývaná L-funkce. Změna počtu sousedů se vzdáleností od jednoho pozorování, který vyjadřuje K-funkce (či L-funkce) je ale jen jedním možným aspektem prostorového rozmístění (*spatial pattern*) bodů. Používají se proto také F-funkce (kumulativní funkce vzdáleností od náhodně vybrané souřadnice k nejbližšímu pozorování) a G-funkce (kumulativní funkce vzdáleností od jednoho pozorování k jeho nejbližšímu sousedu). Tyto dvě funkce tedy odpovídají parametrům r_1 a r_2 ve Vz. 17-4, ale hodnoty vzdáleností se zde nekombinují do jednoho čísla. Občas se používá také J-funkce, spočtená z předchozích dvou vzorcem (1-G)/(1-F), která je v jistém smyslu obdobou koeficientu A ze Vz. 17-4. Hodnota 1 odpovídá náhodnému rozmístění bodů (na dané prostorové škále), hodnota <1 nahloučenosti bodů a hodnota >1 více pravidelnému rozmístění bodů.

Novější verse těchto metod umožňují také brát v úvahu nejen umístění, ale i vlatnosti jednotlivých objektů, tzv. *marked point pattern analysis*. Jedná-li se například o vymapované stromy v ploše, pak je třeba každý strom charakterizován svojí druhovou identitou, velikostí, zdravotním stavem a pod.). To umožňuje testování velmi specifických hypotéz o vzájemném prostorovém vztahu individuí různých druhů, či individuí různých velikostí.

Popis podává např. Diggle (2013) a elektronické materialy prof. A. Baddeleye (http://www.csiro.au/resources/pf16h.html).

Binomické rozdělení

Předpokládejme, že provádíme *n* nezávislých pokusů, jejichž výsledek je možné hodnotit zařazením do jedné ze dvou kategorií (často se uvádí úspěch - neúspěch). V biologii např. pozorujeme *n* individuí (náhodně nezávisle vybraných) a sledujeme, zda jsou to samci či samice. Nebo nakazíme *n* pokusných krys určitým virem a sledujeme, zda přežijí nebo nepřežijí.

Počet úspěchů (označme jej *x*) je potom náhodná proměnná s **binomickým rozdělením**. Binomické rozdělení (*binomial distribution*) je charakterizováno dvěma parametry: *p* - pravděpodobnost úspěchu v jednotlivém pokusu, a *n* - počet pokusů. Pravděpodobnost, že proměnná *x* nabude přesně hodnoty *X*, je potom

$$P(x = X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

Vz. 17-7

q=1-p je pravděpodobnost neúspěchu (odvození vzorce je klasickým cvičením kombinatoriky a teorie pravděpodobnosti). Střední hodnota rozdělení

$$\mu_x = np$$

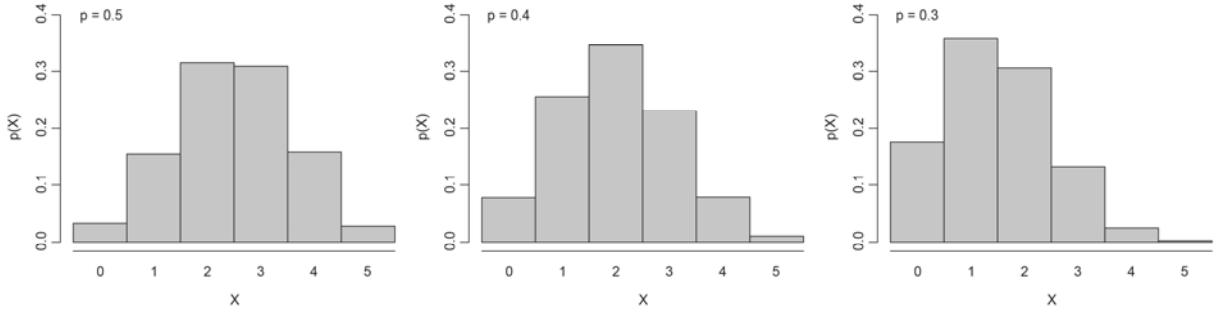
Vz. 17-8

a variance

$$\sigma_x^2 = npq$$

Vz. 17-9

Pokud je *p=q=0.5*, potom je binomické rozdělení symetrické, jinak je asymetrické (viz Obr. 17-3). Pokud je *p* velmi malé a *n* relativně velké, blíží se binomické rozdělení Poissonovu, pokud je *n* rozumně velké a *p* není blízko nuly nebo jedničky, blíží se binomické rozdělení normálnímu (toho se často užívá při aproximaci normálním rozdělením).



Obr. 17-3 Binomické rozdělení s *n=5* a různými hodnotami parametru *p* (parametr *q* je vždy *1-p*).

V praxi se binomického rozdělení užívá především při odhadu a porovnání relativních četností jevů. Např. odhadujeme zastoupení samic v populaci - odchytem získáme *n* individuí daného druhu, z nich je *X* samic. Ptáme se, jaké je procento (relativní zastoupení) samic v populaci (tedy odhadujeme parametr *p*) a k odhadu chceme znát i konfidenční interval. Přirozeně, odhadem *p* bude

$$\hat{p} = \frac{X}{n}$$

Vz. 17-10

$\hat{q} = 1 - \hat{p}$. $\hat{p}$ je náhodná proměnná, která má varianci ze Vz. 17-9.

$$\sigma_{\hat{p}}^2 = \frac{pq}{n}$$

Vz. 17-11

Protože *p* ani *q* neznáme, ale známe pouze jejich odhady, odhadujeme varianci pomocí nich:

$$s_{\hat{p}}^2 = \frac{\hat{p}\hat{q}}{n-1}$$

Vz. 17-12

Odmocnina z tohoto výrazu je střední chyba odhadu. Pokud můžeme užít normální aproximaci, je *(1-α)* konfidenční interval dán vzorcem

$$\hat{p} \pm \left(Z_{(1-\alpha/2)} s_{\hat{p}} + \frac{1}{2n} \right)$$

Vz. 17-13

$Z_{(1-\alpha/2)}$ je *(1-α/2)×100*-procentní kvantil normovaného normálního rozdělení. Vodítko, kdy je rozumné použít normální aproximaci, dává Tab. 17-3 - čím víc se *p* liší od 0.5, tím větší musí být *n*:

Tab. 17-3 Vhodnot normální aproximace pro binomické rozdělení.

$\hat{p}$	n
0.5	≥ 30
0.4 nebo 0.6	≥ 50
0.3 nebo 0.7	≥ 80
0.2 nebo 0.8	≥ 200
0.1 nebo 0.9	≥ 600

Pokud nemůžeme užít normální aproximaci, postupujeme podle následujících vzorců (v těch případech je rozdělení nesymetrické, a také konfidenční interval je nesymetrický).

$$\text{Dolní mez} = \frac{X}{X + (n - X + 1)F_{(1-\alpha/2),\nu 1,\nu 2}}$$

Vz. 17-14

$F_{(1-\alpha/2),\nu 1,\nu 2}$ je *(1-α /2)*×100%-ní kvantil, s příslušnými stupni volnosti: $\nu_1 = 2(n-X+1)$ a $\nu_2 = 2X$.

$$\text{Horní mez} = \frac{(X+1)F_{(1-\alpha/2),\nu'_1,\nu'_2}}{n-X+(X+1)F_{(1-\alpha/2),\nu'_1,\nu'_2}}$$

Vz. 17-15

Počty stupňů volnosti jsou $\nu_1'=2(X+1)$ a $\nu_2' = 2(n-X)$.

Na podobném základě, jako jsme určovali konfidenční intervaly, lze také porovnávat proporce: např. na lokalitě 1 bylo 15 samic ze 60-ti individuí, na lokalitě 2 bylo 10 samic z 50-ti individuí, a my se ptáme, zda se liší relativní podíl samic v základních populací těchto dvou lokalit? Ve většině případů je ale jednodušší a výhodnější použít kontingenční tabulky a χ^2-test, v případech, kdy jsou frekvence velmi nízké, Fisherův exaktní test.

Často se ptáme, jak velký potřebujeme výběr k odhadu p s požadovanou přesností. Požadujeme-li, aby střední chyba odhadu $\hat{p}$ byla přibližně rovna w, potom potřebná velikost výběru je

$$n = \frac{pq}{w^2}$$

Vz. 17-16

Střední hodnota střední chyby průměru bude potom rovna w, tzn. že s přibližně 50%-ní pravděpodobností dostaneme střední chybu větší a s 50%-ní pravděpodobností střední chybu menší. Předpokládáme-li, že v populaci je přibližně 20% jedinců s mutací určitého typu a chceme-li jejich zastoupení určit se střední chybou 1% (95%-ní konfidenční interval bude potom přibližně odhad ±2%), potřebujeme vyšetřit $n=(0.2 \times 0.8)/0.01^2=1600$ jedinců.

Příkladová data

Ve výlovu bylo odebráno 86 kaprů a na každém spočten počet ektoparazitů druhu *Caprozhroutus magnus*. Proměnná *NumParas* udává počet nalezených parazitů na jednom jedinci, zatímco odpovídající hodnota v proměnné *NumCarps* udává, na kolika kaprech byl tento počet parazitů nalezen. Co můžeme říci o distribuci tohoto parazita mezi jedince kaprů?

Bylo mapováno rozmístění jedinců koniklece velkokvětého na ploše 100 x 100 metrů. Proměnné *xPos* a *yPos* udávají souřadnice jednotlivých rostlin. Testujte hypotézu, že jejich rozmístění je shloučené (agregované).

Ze 120 náhodně vybraných jablek bylo 56 červivých. Odhadněte procento červivých jablek v populaci, spolu s 95% konfidenčním intervalem.

Jak postupovat v programu Statistica

Hodnoty z prvého příkladu (proměnné *NumCarps* a *NumParas*) můžeme porovnat s Poissonovou distribucí následujícími způsoby.

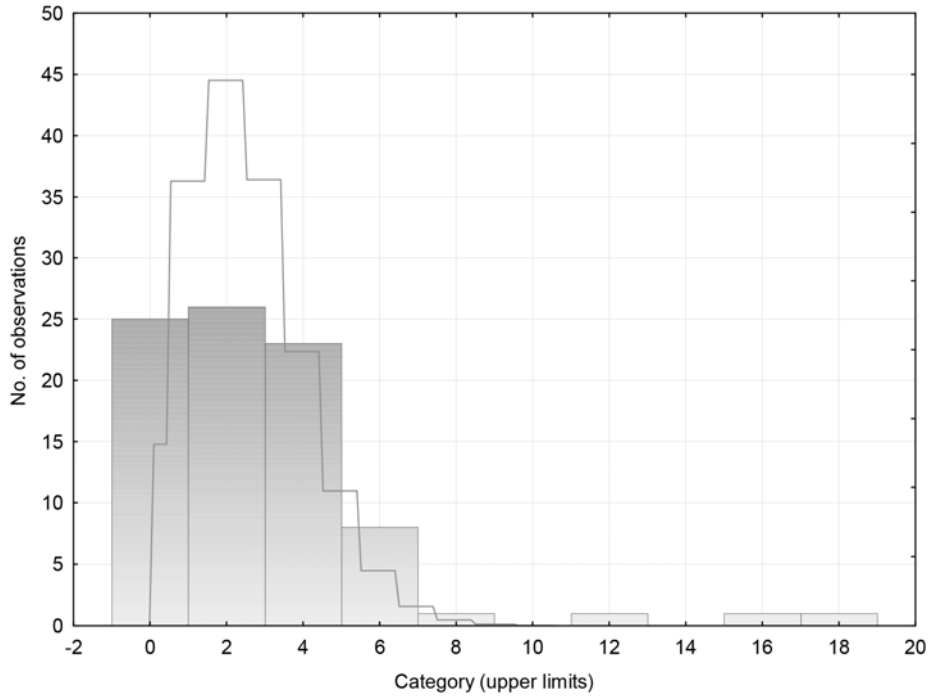
Z menu zvolíme příkaz *Statistics* | *Distribution Fitting*, v dialogovém okně *Distribution Fitting* zvolíme variantu *Discrete Distributions* a v seznamu *Poisson*, a pak zvolíme tlačítko *w* (vpravo dole) pro zadání vah pozorovaných četností parazita (*NumParas*). Těmito váhami budou počty kaprů v proměnné *NumCarps*. V dialogovém okně *Analysis/Graph Case Weights* zvolíme *Use weights for this Analysis/Graph only*, zadáme jméno proměnné do políčka *Weight variable* (lze na něj dvakrát kliknout a vybrat jméno proměnné, *NumCarps*, ze seznamu) a ještě zvolíme hodnotu *On* v rámečku *Status*. Okno

zavřeme tlačítkem *OK* (potvrdíme informační zprávu) a zvolíme *OK* i v původním dialogovém okně. V dalším okně (*Fitting Discrete Distributions*) zadáme proměnnou *NumParas* pomocí tlačítka *Variable* a na záložce *Options* ještě zvolíme v rámečku *Kolmogorov-Smirnov test* variantu *Yes (categorized)*. Po volbě tlačítka *Summary* se objeví tyto výsledky:

Variable: NumParas, Distribution: Poisson, Lambda = 2,45 (Spreadsheet44)
Kolmogorov-Smirnov d = 0,20470, p < 0,01
Chi-Square = 49,55984, df = 2 (adjusted) , p = 0,00000

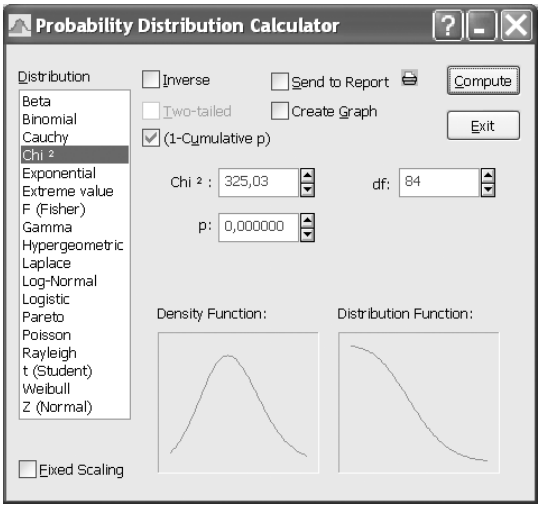
Category	**Observed Frequency**	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected
<= 0,00000	25	25	29,06977	29,0698	7,39541	7,39541	8,59931
2,00000	26	51	30,23256	59,3023	40,40326	47,79866	46,98053
4,00000	23	74	26,74419	86,0465	29,36955	77,16821	34,15063
6,00000	8	82	9,30233	95,3488	7,71943	84,88764	8,97609
8,00000	1	83	1,16279	96,5116	1,02610	85,91374	1,19314
10,00000	0	83	0,00000	96,5116	0,08176	85,99551	0,09507
12,00000	1	84	1,16279	97,6744	0,00433	85,99983	0,00503
14,00000	0	84	0,00000	97,6744	0,00016	86,00000	0,00019
16,00000	1	85	1,16279	98,8372	0,00000	86,00000	0,00001
< Infinity	1	86	1,16279	100,0000	0,00000	86,00000	0,00000

V záhlaví tabulky je zobrazen výsledek Kolmogorov-Smirnovova testu, který porovnává distribuci počtu parazitů s Poissonovou distribucí s parametrem λ odhadnutým z našich dat. Podobně postupuje test dobré shody, jehož výsledky jsou prezentovány v posledním řádku záhlaví (a výpočetní postup ilustruje vlastní tabulka). Počet intervalů, do kterých byly hodnoty proměnné *NumParas* rozděleny je sice 10, ale počet stupňů volnosti je uveden jako *df=2*, protože (a) Statistica spojuje intervaly tak, aby očekávaný počet případů (viz sloupec *Expected Frequency*) nebyl nikdy nižší než 5 (takže zbudou jen 4 kombinované intervaly: 0, <=2, 2 až 4, a >4) a (b) další stupeň volnosti je odečten s ohledem na odhad parametru distribuce z našich dat. Výsledky obou testů ukazují průkaznou odlišnost od Poissonovy distribuce, zatím ale nevím, kterým směrem. K pochopení může napomoci graf, který vytvoříme z okna *Fitting Discrete Distributions* pomocí tlačítka *Plot of observed and expected distribution*.



Vidíme, že počet nulových hodnot je vyšší, než by pro Poissonovu distribuci s daným průměrem měl být, a také že vyšší počty parazitů jsou více časté. To ukazuje, že odchylka od náhodné distribuce parazitů je směrem k agregaci parazitů na určitých jedincích kapra.

Pomocí příkazu *Statistics* | *Basic Statistics/Tables* a následné volby *Descriptive statistics* spočtene průměr a varianci proměnné *NumParas*. Pozor, i zde musíme zadat hodnoty proměnné *NumCarps* jako váhy, stejným způsobem jako výše. Výsledné hodnoty jsou *2.453* pro průměr a *9.380* pro varianci. Podle Vz. 17-2 pak vypočteme hodnotu testové statistiky jako *9.38*85/2.453 = 325.03* (pozor, počet pozorování je roven počtu zkoumaných kaprů, nikoliv počtu řádků v našich datech) a tu pak porovnáme s χ^2 rozdělením s 84 stupni volnosti pomocí příkazu *Statistics* | *Probability Calculator* | *Distributions*.



Hypotézu o shodě s Poissonovou distribucí tedy zamítáme s $p < 0.000001$ a to, že je variance výrazně vyšší než průměr opět ukazuje na agregaci parazitů. Ze známých hodnot variance a průměru také můžeme spočítat Lloydův index například v programu Excel pomocí vzorečku (9.38/2.453 - 1)/2.453 + 1, s výsledkem *2.151*, který opět ukazuje na shlukovitost rozmístění parazitů.

Data, ve kterých je rozmístění jedinců v ploše reprezentováno souřadnicemi (proměnné *xPos* a *yPos*), nelze v programu Statistica jednoduše zpracovat, je třeba přes souřadnice položit čtvercovou síť (ať již graficky či v podobě výpočetního algoritmu) a spočíst počty jedinců v každém čtverci. Tato data lze pak analyzovat obdobně jako v předchozím příkladu.

Podíl červivých jablek v třetím příkladu spočtem snadno jako *56/120 = 0.467*, což je tedy zároveň odhad parametru *p* binomické distribuce. S ohledem na *n=120* a hodnotu *p* můžeme pro odhad konfidenčního intervalu použít aproximaci normální distribucí podle Vz. 17-13 a Vz. 17-12. Variance odhadu *p* je tedy *0.467*0.533/(120-1) = 0.002092*, ve Vz. 17-13 ale používáme směrodatnou odchylku, tj. odmocninu spočtené hodnoty: *0.0457*. Kvantil normované normální distribuce (spočteme jej v okně *Probability Calculator*) je pro 95% konfidenční interval (a tedy *p=0.975*) roven *1.96*, a tedy konfideční interval je (i s korekcí členem *1/(2*n)*) roven *(0.373, 0.561)*.

Konfidenční interval binomického rozdělení lze spočíst v programu Statistica i přímo. Zvolíme z menu příkaz *Statistics* | *Power Analysis*, v zobrazeném okně zvolíme v levém seznamu *Interval Estimation* a vpravo zvolíme *One Proportion, Z, Chi-Square Test* a pak tlačítko *OK*. Zadáme odhadnuté *p* (0.467) v políčku *Observed Proportion p* a počet pozorování (120) v *Sample Size*, ponecháme *Conf. Level* rovné *0.95* a tlačítkem *Compute*

zobrazíme konfidenční intervaly spočtené různými výpočetními variantami. Přesný (*exact*) konfidenční interval má hodnotu (0.375, 0.560).

Jak postupovat v programu R

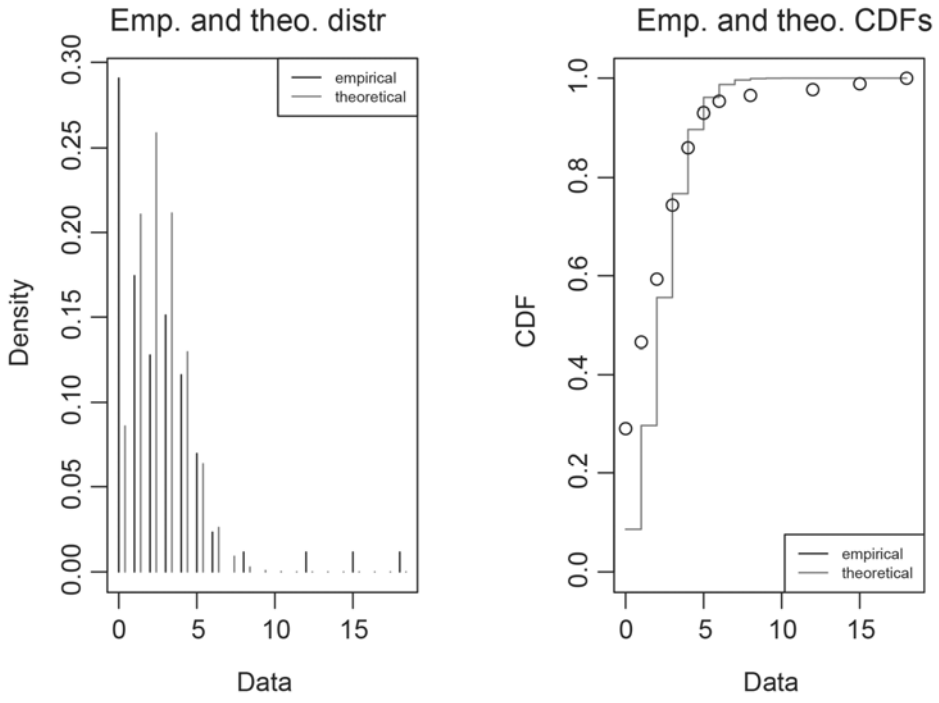
Údaje o počtu parazitů na jednotlivých kaprech, obsažené v kondezované podobě v proměnných *NumCarps* a *NumParas* expandujeme do proměnné obsahující počty parazitů na jedincích tímto příkazem:

```
> nPar <- with(chap17a, rep(NumParas,NumCarps))
```

Odhadnout parametry zvolené distribuce z dat a srovnat data s touto distribucí umožňuje knihovna *fitdistrplus*:

```
> library(fitdistrplus)
> fit.pois <- fitdist(nPar,"pois")
> summary(fit.pois)
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
       estimate Std. Error
lambda 2.453488  0.1689051
Loglikelihood:  -221.0932   AIC:  444.1865   BIC:  446.6408
> plot(fit.pois)
```

Na výsledných diagramech je pěkně vidět vyšší počet nul a vysokých počtů proti očekávání, založenému na předpokladu Poissonovy distribuce.



Srovnání mezi fitovanou a pozorovanou distribucí hodnot pomocí χ^2 testu lze provést funkcí *gofstat*:

```
> gofstat(fit.pois)
Chi-squared statistic:  50.89291
Degree of freedom of the Chi-squared distribution:  4
Chi-squared p-value:  2.350246e-10
Chi-squared table:
     obscounts theocounts
<= 0 25.000000   7.395405
```

```
<= 1 15.000000  18.144541
<= 2 11.000000  22.258710
<= 3 13.000000  18.203829
<= 4 10.000000  11.165721
> 4  12.000000   8.831793

Goodness-of-fit criteria
                                1-mle-pois
Aikake's Information Criterion   444.1865
Bayesian Information Criterion   446.6408
```

Výsledky (ale nikoliv závěry) jsou odlišné od testu provedeného v programu Statistica, protože byly zvoleny jiné intervaly hodnot. Volby programu Statistica ale můžeme snadno reprodukovat takto:

```
> gofstat(fit.pois,chisqbreaks=c(0,2,4))
Chi-squared statistic:  49.55984
Degree of freedom of the Chi-squared distribution:  2
Chi-squared p-value:  1.73068e-11
Chi-squared table:
     obscounts theocounts
<= 0 25.000000   7.395405
<= 2 26.000000  40.403251
<= 4 23.000000  29.369550
> 4  12.000000   8.831793
...
```

S Poissonovou distribucí můžeme naše data srovnávat i pomocí Kolmogorov-Smirnovova testu:

```
> ks.test(nPar,"ppois",mean(nPar))
        One-sample Kolmogorov-Smirnov test
data:  nPar
D = 0.2047, p-value = 0.001482
alternative hypothesis: two-sided

Warning message:
In ks.test(nPar, "ppois", mean(nPar)) :
  ties should not be present for the Kolmogorov-Smirnov test
```

Jak funkce upozorňuje, opakované hodnoty jsou při tomto testu problémem, výsledky jsou tedy jen přibližné.

Můžeme také spočíst χ^2 statistiku srovnávající varianci s průměrem, podle Vz. 17-2 a porovnat ji s χ^2 distribucí (poslední výsledek je odpovídající hodnota *p*, i když ne příliš přesná):

```
> x <- var(nPar)*(length(nPar)-1)/mean(nPar)
> x
[1] 324.9716
> (1-pchisq(x,length(nPar)-1))
[1] 0
```

Pro výpočet Lloydova indexu si vytvoříme samostatnou funkci:

```
> lloyd.index <- function(x){ ((var(x)/mean(x))-1)/mean(x)+1}
> lloyd.index(nPar)
[1] 2.150686
```

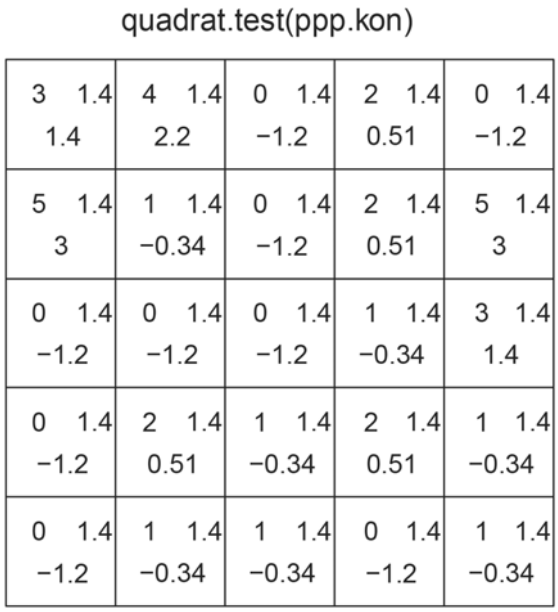
Data o prostorových souřadnicích jedinců koniklece můžeme poměrně rozsáhle analyzovat pomocí knihovny *spatstat*. Zde ukážeme jen test shody s Poissonovou distribucí, ale knihovna obsahuje rozsáhlý soubor funkcí pro analýzu bodových uspořádání, včetně výpočtu K funkce (a souvisejících F, G a J funkcí) a vytváření modelů, ve kterých odchylky

od náhodnosti (jak směrem k pravidelnosti, tak směrem k agregaci) můžeme vysvětlovat pomocí změřených proměnných, například vlastností prostředí.

```
> library( spatstat)
> ppp.kon <- with(chap17b, ppp(x=xPos,y=yPos,c(0,100),c(0,100)))
> quadrat.test(ppp.kon)
        Chi-squared test of CSR using quadrat counts
data:  ppp.kon
X-squared = 41.4286, df = 24, p-value = 0.02988
alternative hypothesis: two.sided

Quadrats: 5 by 5 grid of tiles
Warning message:
Some expected counts are small; chi^2 approximation may be inaccurate
```

Funkce *ppp* vytvořila ze souřadnic v datovém rámci *chap17b* datový objekt představující bodové uspořádání. Nezbytnou součástí jeho tvorby je definice polygonu, představující zkoumané území. Zde jsme tento polygon zadali v nejjednodušší možné podobě, tj. jako obdélník v uvedeném rozsahu souřadnic. Funkce *quadrat.test* tento obdélník rozdělila na čtverce (5 x 5 čtverců, jak ukazuje výstup funkce) a srovnala počet pozorování v jednodlivých čtvercích s očekávaným počtem. Výstup z této funkce ale varuje, že očekávané počty jsou příliš nízké a to je vidět i z detailnějšího výstupu, který zobrazí funkce *plot*:

```
> plot(quadrat.test(ppp.kon))
```



V každém čtverci je uveden pozorovaný počet konikleců (*O*, vlevo nahoře), očekávaný počet (*E*, vpravo nahoře, zde – s ohledem na přesné zaplnění obdélníku těmito čtverci – stejná hodnota pro všechny, tj. 1.4) a také (v dolní části každého čtverečku) příspěvek k celkové hodnotě χ^2 statistiky (odmocnina z $(O\text{-}E)^2/E)$).

Pro odhad intervalu spolehlivosti pro podíl červivých jablek můžeme použít funkci *binom.test*, které zadáme počet případů a celkový počet sledovaných objektů.

```
> binom.test(56,120)
        Exact binomial test
data:  56 and 120
number of successes = 56, number of trials = 120, p-value = 0.523
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
```

```
 0.3750729 0.5599445
sample estimates:
probability of success
             0.4666667
```

Tato funkce také provádí test shody pozorované relativní frekvence s apriorně zvolenou hodnotou *p* (implicitní hodnota je 0.5, vhodná například pro testy podílu pohlaví ve vrhu), pro náš příklad takový test ale není smysluplný.

Popis metod v článku

Methods

Randomness of parasite distribution across fish bodies was tested by comparing the observed counts with Poisson distribution using χ^2 goodness-of-fit test. The nature of observed point pattern was summarized using Lloyd's index of patchiness (Lloyd 1967).

We have estimated the confidence interval for the frequency of worm infestation of apples using normal approximation of the estimated *p* and estimated standard error of *p* calculated using Vz. 17-12 (Zar 20xx).

Results

We have found significant deviation from random distribution of parasite individuals (χ^2_2=49.6, p < 0.00001) towards aggregated presence (Lloyd index value L=2.151)

The average rate of worm infestation was 0.467, with 95% confidence interval (0.373, 0.561).

Doporučená četba

Poissonovo a binomické rozdělení:

Zar J. H. (1984) pp. 369-420, Sokal & Rohlf (1981) pp. 62-97.

Studium rozmístění v prostoru:

Diggle P.J. (2013): Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition. Chapman & Hall / CRC, 300 pp.

Lepš J. (1989): Metody studia populací. In: Dykyjová D. [ed.]: Metody studia ekosystémů. - Academia, Praha. p. 230 - 302.

M. Lloyd (1967): Mean crowding. *J. of Animal Ecology*, **36**: 1-30.

Pielou E.C. (1977): Mathematical Ecology. Wiley, New York.

18 Shluková analýza

Zatímco výklad v předcházejících kapitolách byl veden tak, aby čtenář byl schopen popisované metody používat, počínaje touto kapitolou je naší snahou jen to, aby čtenář věděl, že metody řešící dané typy úloh existují, a aby zjistil, kde se o nich dočte více (případně který program použít pro spočtení), a také aby byl zhruba schopen chápat smysl prací z vlastního oboru, které uvedené statistické metody použijí, případně aby si byl vědom omezenosti příslušných metod a byl příslušně „imunní" proti neopodstatněným dalekosáhlým interpretacím. I když bude čtenář díky přístupnému programovému vybavení schopen některé zde popisované metody přímo spočíst, měl by si před jejich použitím vyhledat příslušnou literaturu a seznámit se s ní.

Mějme řadu objektů (často odpovídajích nezávislým pozorováním). Na každém objektu měříme určitý soubor proměnných, můžeme ale také říci, že měříme jednu nebo i více mnohorozměrných proměnných. Souboru metod, zabývajícím se analýzou takovýchto dat, říkáme metody mnohorozměrné analýzy, často zkráceně mnohorozměrné metody (*multivariate methods*). Zatímco v běžné statistice je kladen většinou důraz buď na odhad parametrů nebo na testování hypotéz, při hodnocení mnohorozměrných proměnných máme občas i skromnější cíle: vyznat se v datech. Ptáme se, zda se určité typy objektů pravidelně opakují: hledáme *repeatable patterns*. Jejich nalezení nám potom umožní hypotézy navrhnout. Jednou z pomůcek nám v tom může být shluková analýza. Je ovšem třeba připomenout, že existují i mnohé metody mnohorozměrné analýzy, umožňující testování hypotéz.

Cílem shlukové analýzy (*cluster analysis*) je nalézt v celém souboru takové skupiny objektů, které jsou si navzájem podobné, ale které se liší od objektů ostatních skupin, nebo najít takové skupiny proměnných, které jsou navzájem korelovány. Tato metoda má v biologii velmi blízko k taxonomii – a také se v úzkém spojení s taxonomií vyvíjela, jako její zvláštní větev, tzv. numerická taxonomie.[*] Další velice časté použití je v ekologii, zvláště v ekologii společenstev. Obzvláštní oblibě se tato metoda těší ve fytocenologii, kde byla (a někdy dosud i je) klasifikaci společenstev věnována neobvyklá péče.

Typickým příkladem je užití v taxonomii: studujeme například vnitrodruhovou variabilitu určitého druhu. Na padesáti lokálních populacích změříme biometrické charakteristiky (např. délku nejvyššího listu, délku korunní trubky, šířku pysku, počet květů atd.) a ptáme se, zda jsou si určité skupiny populací podobnější, zda tvoří shluky (*clusters*). Jiným příkladem použití by bylo, kdybychom měli charakterizovány všechny druhy určitého rodu, a ptali bychom se, zda jsou si některé druhy v rámci rodu podobnější (zda existují shluky druhů v rámci rodu).

Typickým použitím v ekologii je klasifikace zápisů složení vegetace (snímků): hledáme skupiny snímků navzájem podobných. Můžeme ale také hledat skupiny druhů, které se spolu častěji vyskytují; o těch můžeme předpokládat, že tvoří ekologicky definované skupiny. **Upozorněme, že shluková analýza je především metodou prvního stupně analýzy dat, která má shrnout variabilitu ve studovaných datech, případně navrhnout určité hypotézy.** Neměla by být konečným cílem žádné práce, ale spíše prvním vodítkem.

Metody hledání shluků můžeme rozdělit na dvě velké skupiny: hierarchické a nehierarchické. Nehierarchické metody dělí soubor na několik shluků stejné úrovně.

[*] Numerická taxonomie vznikla původně jako numerická fenetika; dnes existuje i numerická kladistika, která se považuje za součást numerické taxonomie. Numerická taxonomie využívá metody shlukové analýzy, ale vyvíjí i některé své speciální metody; zvláště to platí o metodách kladistiky.

Výsledný počet shluků může být buď zadán předem nebo je součástí procedury určit podle nějakého kriteria optimální počet shluků. Hierarchické klasifikace vytvářejí shluky, které mají různou hierarchickou úroveň - shluky nejvyšší hierarchické úrovně obsahují shluky nižší úrovně, ty ještě nižší atd. Výsledky hierarchických procedur bývá zvykem znázornit dendrogramem. Hierarchické metody jsou buď aglomerativní (postupujeme „odspoda“, vytváříme jádra shluků z nejpodobnějších dvojic a na ně postupně „nabalujeme“ další a další objekty), nebo divizivní (postupujeme „odshora“, tj. dělíme nejprve celý soubor na většinou dvě části a s každou z těchto částí potom pracujeme jako se samostatným souborem a dělíme jej znovu na dvě části atd). V biologii jsou tradičně užívány hlavně metody hierarchické aglomerativní.

Data

Je zřejmé, že objekty mohou být charakterizovány jak kvalitativními, tak kvantitativními daty: např. v taxonomii může být znak jak délka semena, tak barva květu. Kvantitativní data přitom mohou být udávána v různých jednotkách (délky, váhy) a i data délková se pohybují v různých řádech. Proto bývá zvykem proměnné různě standardizovat: většinou se standardizací kvantitativních proměnných (faktory transformovat nelze) míní tzv. *Z*-transformace: pro každou proměnnou spočteme její průměr a směrodatnou odchylku v sledovaném souboru a poté spočteme transformaci

$$z = \frac{x - \overline{X}}{s_x}$$

Vz. 18-1

Tak budou všechny standardizované proměnné bezrozměrné a budou mít nulový průměr a jednotkovou varianci. Tato standardizace je vhodná, pokud lze předpokládat, že jednotlivé proměnné mají alespoň přibližně normální rozdělení. Kromě toho se podle potřeby užívají různé transformace dat (např. logaritmická), případně různé jiné standardizace, často specifické pro jednotlivé vědní obory.

Podobnost

Jak bylo řečeno výše, chceme nalézt skupiny objektů, které jsou si navzájem podobné. Musíme proto definovat, co je to podobnost. Dnes existuje mnoho měr podobnosti nebo nepodobnosti mezi objekty, často specifických pro jednotlivé vědní disciplíny (nebo shodné pro různé disciplíny, ale v každé nazývané jinak). Zde si předvedeme jen nejjednodušší z nich. Pro kvalitativní data je asi nejjednodušší míra podobnosti:

$$\frac{\text{počet shodných znaků}}{\text{počet vsech sledovaných znaků}}$$

Vz. 18-2

Jak to u měr podobnosti často bývá, její hodnota je 1 pro objekty zcela shodné a 0 pro objekty, které se liší ve všech znacích. Doplněk této hodnoty do jedničky by mohl být považován za míru nepodobnosti.

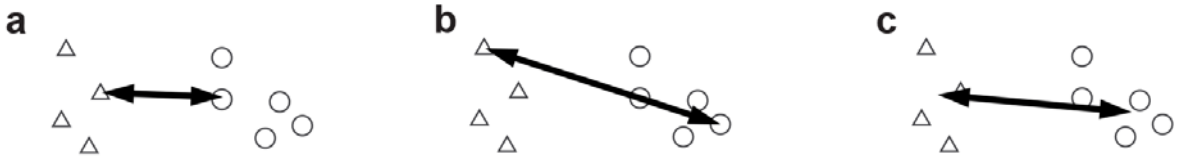
Pro kvantitativní znaky označme $x_{1,i}$ hodnotu *i*-té proměnné na prvém objektu, $x_{2,i}$ hodnotu *i*-té proměnné na druhé objektu (v taxonomii to může být např. délka pysku, při studiu společenstev pokryvnost či četnost určitého druhu). Za míru nepodobnosti pro kvantitativní data se často užívá tzv. Euklidovská vzdálenost (*Euclidean distance*),

vycházející z představy vzdálenosti dvou bodů v *n*-rozměrném prostoru, kde *n* je počet sledovaných proměnných:

$$ED = \sqrt{\sum_i (x_{1,i} - x_{2,i})^2}$$

Vz. 18-3

Při shodě dvou objektů je její hodnota 0, horní mez je dána daty. Často se užívá s daty po standardizaci. Pokud chceme najít shluky proměnných, potřebujeme míru „podobnosti“ mezi proměnnými. Zde se často užívá korelační koeficient nebo míry jemu podobné. Připomeňme, že korelační koeficient je dobrou měrou vztahu dvou proměnných, které mají dvourozměrné normální rozdělení.



Obr. 18-1 Schématické znázornění způsobu měření vzdálenosti (nepodobnosti) dvou skupin objektů při užití metody: **a** - *single linkage* (jednospojné), **b** – *complete linkage* (všespojné), **c** – *average linkage* (středospojné).

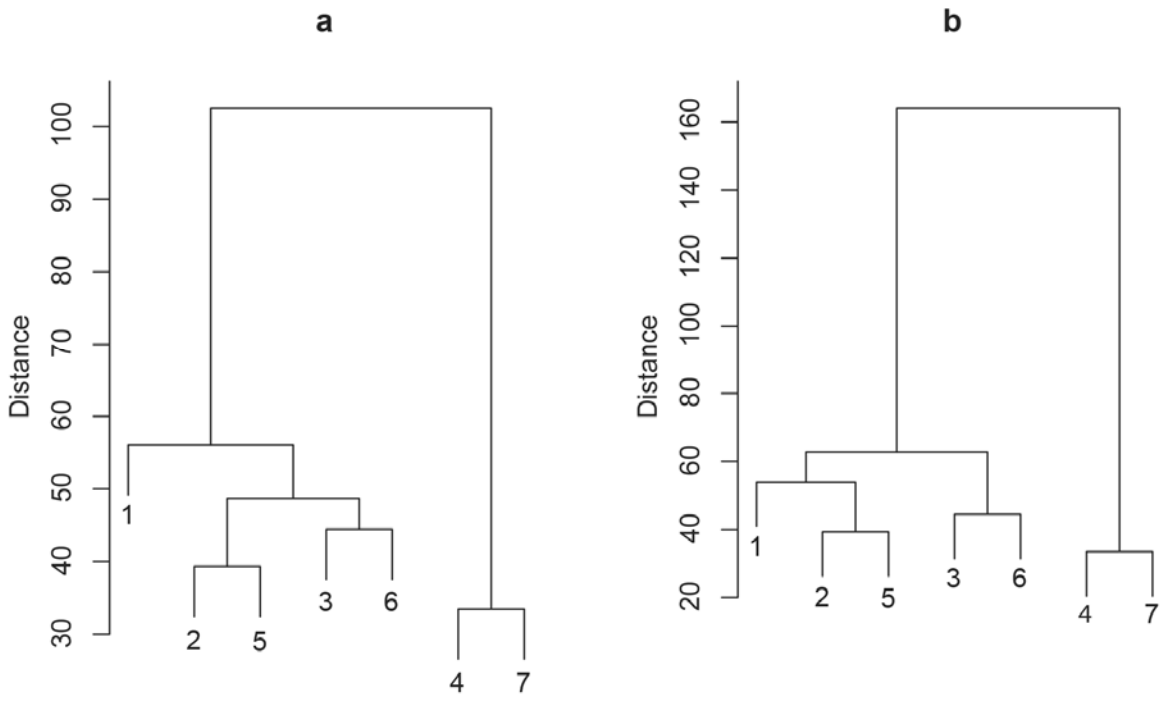
Shlukovací algoritmy

Budou stručně probrány jen aglomerativní algoritmy, které jsou častěji užívány. Pokud máme definovanou podobnost mezi objekty, můžeme učinit prvý krok shlukové analýzy - spočíst matici podobností všech dvojic objektů. Na tuto matici potom aplikujeme některou shlukovací proceduru: najdeme mezi všemi objekty nejpodobnější dvojici a tu spojíme v jeden shluk. Poté přepočítáme podobnost (vzdálenost) všech objektů k tomuto shluku. Jednotlivé procedury se liší podle toho, jak je definována podobnost (nebo vzdálenost) dvou skupin objektů.

Tři různé možnosti ukazuje Obr. 18-1. Poznamenejme, že v češtině existují nejméně tři terminologie, jak jednotlivé metody nazývat, a tak, chce-li se člověk domluvit, musí užít terminologii anglickou, i když i zde existuje určitá různorodost užívaných terminologií.

Znázornění výsledku

Výsledek bývá zvykem znázornit dendrogramem (Obr. 18-2). Na svislé ose je dána hladina nepodobnosti, na které jsou dva objekty (skupiny objektů) spojovány. Na obou částech obrázku vidíme, že soubor je tvořem dvěma výraznými skupinami, do první skupiny patří objekty 1, 2, 3, 5 a 6, do druhé skupiny objekty 4, a 7. Číslováni objektů je dáno předem, objekty by mohly být charakterizovány i svými názvy. Nejpodobnější dvojicí jsou objekty 4 a 7, hierarchické spojování je podobné, ale ne shodné (v části **a** jsou si více podobné dvojice (2,5) a (3,6), v části **b** je dvojice (2,5) nejvíce podobná objektu 1).



Obr. 18-2 Příklad zpracování téhož souboru dat dvěma různými shlukovacími algoritmy. Všechny postupy jsou shodné, pouze v části **a** byla užita metoda *average linkage* a v části **b** metoda Wardova.

Divizivní metody

V současné době se v ekologii užívá z divizivních metod jen jedna, a to *Two Way Indicator Species Analysis* - program a zkratka *TWINSPAN*. Je konstruována pro klasifikaci souboru vegetačních snímků, ale v principu lze užít i pro jiné typy dat. Dělí soubor vždy na dvě skupiny, pro každé dělení vybere dvě skupiny druhů, charakteristických pro jednu a druhou stranu dichotomie a pak dělení do dvou skupin opakovaně provádí na obou podskupinách, až do stanoveného limitu velikosti skupin.

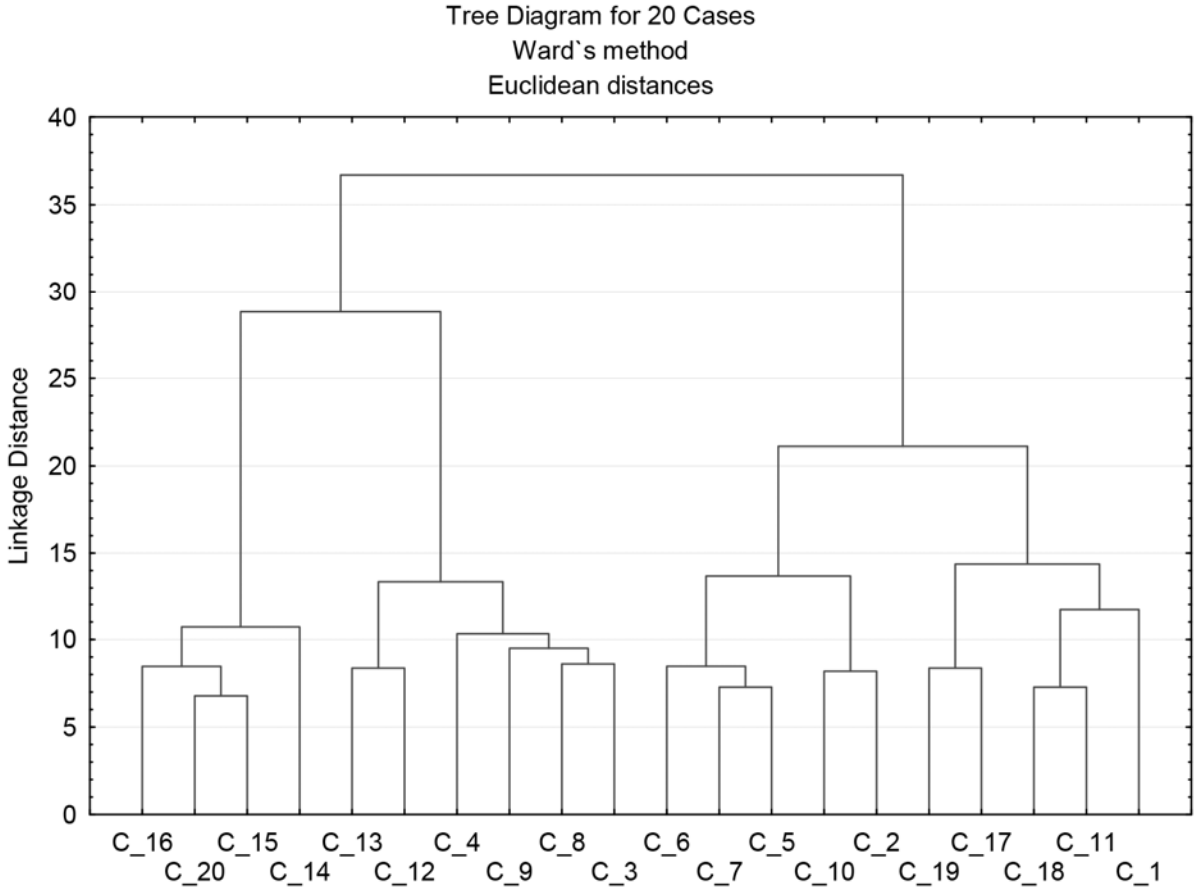
Příkladová data

Příkladová data ve sloupcích A až AD listu *Chap18* souboru *biostat-data.xlsx* představují 20 záznamů (vegetačních snímků) o složení travinné vegetace na ostrově Terschelling při pobřeží Holandska (Batterink & Wijffels 1983). Sloupce představují 30 druhů rostlin, zaznamenaných v plochách a čísla reprezentují významnost jednotlivých druhů v dané ploše (jde o převod subjektivní odhadové stupnice na pořadová čísla, ale budeme zde s nimi zacházet jako by to byly hodnoty na poměrové škále). Naším cílem je najít skupiny záznamů, které jsou si složením vegetace podobné, případně najít skupiny druhů, které se často vyskytují společně.

Jak postupovat v programu Statistica

Z menu vybereme příkaz *Statistics* | *Multivariate Exploratory Techniques* | *Cluster Analysis* a z následně nabízeného seznamu vybereme položku *Joining (tree clustering*). Zvolíme tlačítko *Variables* a v dialogovém okně *Select variables for the analysis* vybereme všechny proměnné pomocí tlačítka *Select All*. Na záložce *Advanced* dialogového okna *Cluster Analysis* zvolíme v položce *Cluster*, že chceme klasifikovat vzorky: *Cases (rows)*. Jako

shlukovací metodu (*Amalgamation (linkage) rule*) zvolíme *Ward's method*. Program Statistica nenabízí většinu z měr nepodobnosti (vzdálenosti), které jsou běžně používány v ekologii, ale pro tato data (s relativně nízkou heterogenitou) lze užít volbu *Euclidean distances*, kterou vybereme v položce *Distance measure*. Dialogové okno s výsledky (*Joining Results*), které se zobrazí po volbě tlačítka *OK* nabízí zobrazení výsledného dendrogramu ve dvou možných orientacích, obvyklejší bývá ta, kterou získáme tlačítkem *Vertical icicle plot*.



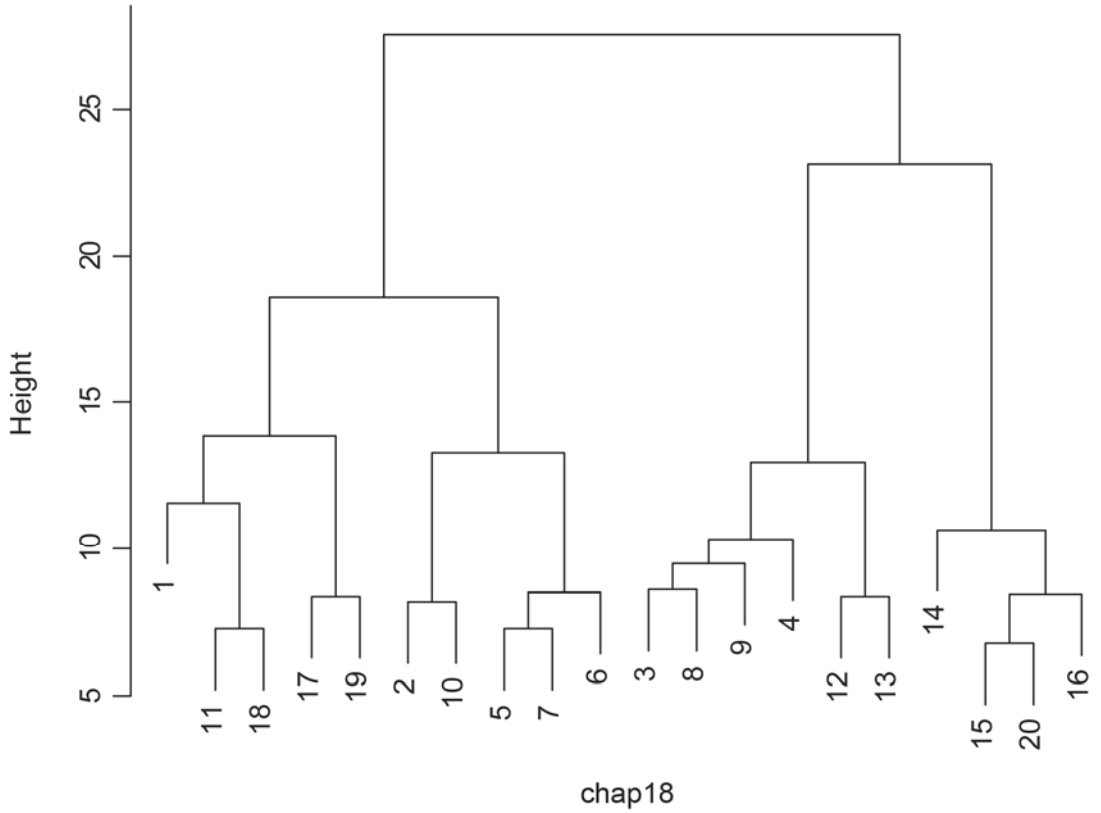
V dendrogramu můžeme rozpoznat 2 – 4 shluky, podle toho, na jaké hodnotě svislé osy (*Linkage Distance*) „povedeme řez". Tyto shluky bychom dále mohli klasifikovat typickými druhy, které se ve snímcích z daného shluku vyskytují. Při hledání takových druhů nebo jejich skupin nám může pomoci klasifikace rostlinných druhů.

Pro vytvoření dendrogramu klasifikujícího rostlinné druhy. zvolíme na záložce *Advanced* dialogového okna *Cluster Analysis* v položce *Cluster* možnost *Variables (columns*) a jako *Distance measure* zvolíme *1-Pearson r*.

Jak postupovat v programu R

Rozsáhlý soubor algoritmů pro shlukovou analýzu nabízí knihovna *cluster*, aglomerativní hierarchické metody jsou v ní k dispozici ve funkci *agnes*. Následujícím příkazem provedem klasifikaci vzorků se zvoleným algoritmem (Wardova metoda), Euklidovská distance je zde implicitní volbou:

```
> library( cluster)
> ag.1 <- agnes( chap18, method="ward")
> plot(ag.1, which.plots=2, main="",sub="")
```



Výsledný diagram je ekvivalentní tomu, který jsme vytvořili v programu Statistica (viz výše), ale jejich srovnání pěkně ilustruje skutečnost, že pořadí vzorků (a jejich větviček) může být pro dva různé dendrogramy představující obsahově shodnou klasifikaci hodně odlišné. V tomto případě jsou dendrogramy prakticky shodné, jen zrcadlově převrácené, ale kdybychom změnili pořadí, ve kterém objekty zadáváme, měnilo by se i pořadí snímků v dendrogramu. Důležitá je tedy informace, které objekty ten který shluk tvoří (a na jaké hodnotě podobnosti jsou vytvořeny), nikoliv to, jak jsou objekty seřazeny.

Vytvoření dendrogramu pro druhy (proměnné) je trochu složitější, protože klasifikaci proměnných funkce *agnes* přímo nepodporuje a také nenabízí výpočet nepodobnosti založený na korelaci:

```
> ag.2 <- agnes( as.dist(cor(chap18)), method="complete")
> plot(ag.2, which.plots=2, main="",sub="",xlab="")
```

Jiné programy

V ekologii je pro hierarchickou klasifikaci často používán program PC-Ord (viz webové stránky *www.pcord.com*). Jiným specializovaným programem pro klasifikaci dat je program CLUSTAN (*www.clustan.com*). Divisivní klasifikační metodu TWINSPAN je implementována v programu Twinspan for Windows, který si lze zdarma stáhnout na adrese: *http://www.ceh.ac.uk/products/software/wintwins.html*

Popis metod v článku

Methods

Classification of vegetation records was performed using hierarchical agglomerative clustering using Euclidean distance among records and Ward's method of joining the clusters.

Results

Clustering results are summarized by a dendrogram in Figure X. We interpret the dendrogram as suggesting three distinct clusters of observations. The cluster with observations 14-16 and 20 corresponds to most wet meadows, characterized by presence of species like ...

Doporučená četba a citovaná literatura

Jongman R.H., ter Braak C.J.F. & van Tongeren O.F.R. (1987): Data analysis in community and landscape ecology. - Pudoc, Wageningen.

Legendre P. & Legendre L. (2012): Numerical Ecology. Third English Edition. Elsevier, Amsterdam, p.337-424 pro shlukovou analýzu, ale i p. 265-335 pro míry nepodobnosti

Batterink M. & Wijffels G. (1983): Een vergelijkend vegetatiekundig onderzoek naar de typologie en invloeden van het beheer van 1973 tot 1982 in de Duinweilanden op Terschelling. Report Agricultural University, Department of Vegetation Science, Plant Ecology and Weed Science. Wageningen, 101 pp.

19 Další mnohorozměrné metody

V této kapitole se budeme zabývat třemi skupinami mnohorozměrných metod. **Metody klasické (neomezené) ordinace** charakterizuje následující úloha: máme na řadě objektů měřen větší počet proměnných a chceme popsat a většinou následně i zobrazit v grafu strukturu podobnosti mezi objekty a závislosti mezi proměnnými. Dosáhneme toho tím, že původní proměnné „nahradíme“ určitým počtem hypotetických proměnných tak, aby došlo k co nejmenší ztrátě informace. Jedná se většinou o metody explorační analýzy dat, jejich cílem je především umožnit orientaci v datech. Pro tuto skupinu metod se hledá obtížně společný název. V ekologii se kromě označení (neomezená) ordinace (*unconstrained ordination*) užívá také název nepřímá gradientová analýza (*indirect gradient analysis)*. V jiných oborech (například sociologie) bývají někdy shrnovány pod pojem faktorová analýza v nejširším slova smyslu.

Diskriminační analýza a příbuzné metody se užívájí tam, kde máme řadu objektů popsanou mnoha proměnnými (jako v předchozím případě) a navíc máme tyto objekty rozdělené do skupin. Cílem této metody je nalézt klasifikační pravidlo, které umožní na základě měřených proměnných určit příslušnost objektů k jednotlivým skupinám.

Metody omezené ordinace (*constrained ordination*) či kanonické analýzy (*canonical analysis*) užíváme tam, kde máme na každém objektu měřeny dvě skupiny proměnných: jednu skupinu považujeme za proměnné vysvětlující, druhou skupinu za vysvětlované. Jedná se tedy o určitou obdobu regrese, kde je ale jak vysvětlující, tak vysvětlovaná proměnná mnohorozměrná.

Metody neomezené ordinace

Výchozí struktura dat je stejná jako v případě shlukové analýzy: máme soubor objektů, každý charakterizovaný řadou proměnných (obvykle spolu v různé míře korelovaných). Opět hledáme v datech určitou strukturu, určité zákonitosti. Cílem těchto metod je nahradit původní proměnné menším množstvím složených proměnných, které budou vzájemně nekorelované, ale dostatečně vysvětlí strukturu sledovaného souboru. Tyto nové proměnné označujeme jako ordinační osy a ve výsledných grafech skutečně představují osy diagramu.

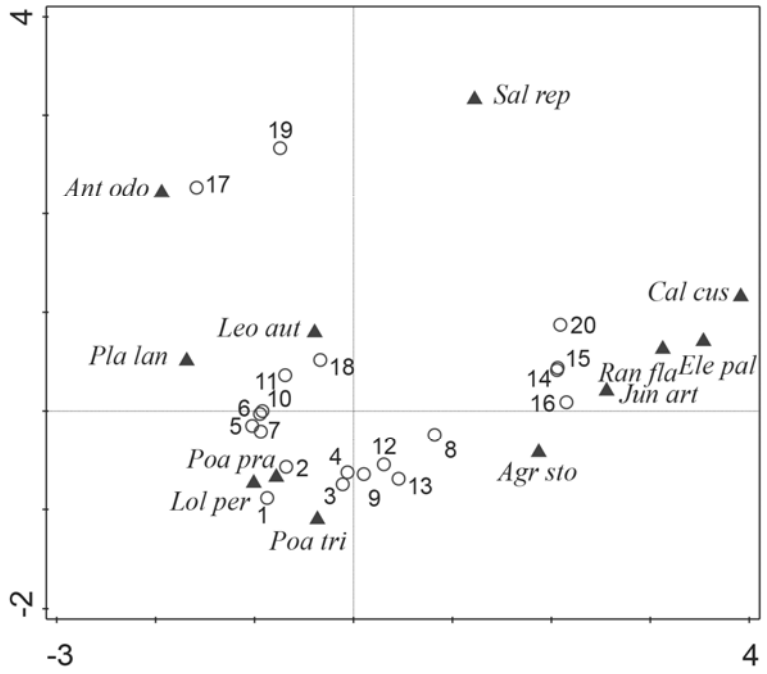
Například při studiu žákovských vysvědčení zjistíme, že celý soubor známek by bylo možno s relativně malým zkreslením nahradit třemi proměnnými; první z těchto proměnných by bylo možné interpretovat jako celkovou nadanost žáka, druhé jako relativní schopnosti v přírodovědných předmětech a třetí jako relativní schopnosti v humanitních předmětech. Lze ukázat, že tyto tři proměnné vysvětlují velkou část celkové variability souboru všech dvanácti proměnných, charakterizujících jednotlivé předměty. Metod, které užíváme pro konstrukci složených proměnných, je celá řada a mají různé předpoklady, při jejichž splnění se dají použít. Termín faktorová analýza *Factor Analysis* se někdy používá souhrně pro celý soubor příbuzných metod, častěji pro omezený soubor metod.

V biologii se nejčastěji užívají následující metody: analýza hlavních komponent (*principal components analysis*, *PCA*), analýza hlavních koordinát (*principal coordinates analysis*, *PCO* či *PCoA*), korespondenční analýza (*correspondence analysis CA*, *reciprocal averaging*), nemetrické mnohorozměrné škálování (*non-metric multidimensional scaling*, *MDS* či *NMDS*). Faktorová analýza v úzkém slova smyslu se dnes v biologii víceméně neužívá.

Úlohu ordinačních metod můžeme formulovat několika způsoby. Nejjednodušší je představa, že studované objekty jsou vlastně body v *n*-rozměrném prostoru, kde *n* je počet proměnných a umístění bodu na dané ose odpovídá hodnota dané proměnné. Úkolem těchto metod je pak promítnout uspořádání bodů v *n* rozměrném prostoru do méně rozměrného prostoru (v praxi obvykle dvou- nebo třírozměrného) tak, aby došlo k minimálnímu zkreslení. Výsledkem je potom nejen promítnutí jednotlivých bodů, představujících jednotlivé objekty, ale často také promítnutí původních os (to neumí například metoda MDS). Obecnějším vyjádřením předchozího zadání je: najít takové uspořádání bodů v dvou- nebo třírozměrném (zřídka vícerozměrném) prostoru, aby vzdálenosti bodů v tomto novém prostoru co nejlépe odpovídaly nepodobnosti studovaných objektů. Jiným zadáním může být: nahradit stávající, tj. skutečné, měřené, proměnné, které jsou často mezi sebou korelované souborem méně proměnných, které vzájemně korelované nejsou, ale jsou buď lineární kombinací nebo váženým průměrem měřených proměnných. Lze ukázat, že za určitých předpokladů dostáváme ze obou výše definovaných zadání ekvivalentní výsledky, tj. že určitá ordinační metoda odpovídá všem výše uvedeným definicím.

Podobně jako shluková analýza, i tyto metody jsou oblíbeny v taxonomii a při hledání opakujících se typů společenstev. V případě studia společenstev předpokládáme, že vegetační snímky (zápisy, ale také záznamy ze zemních pastí, vzorky zooplanktonu, standardizované úlovky ryb) jsou objekty a charakteristikami je zastoupení druhů. Dále předpokládáme, že zastoupení druhů je určeno několika málo výraznými gradienty prostředí. Proto doufáme, že složené proměnné vytvořené ordinačními metodami budou těmto gradientům prostředí odpovídat. Užití těchto metod má určité předpoklady, které není vhodné ignorovat - např. v metodě PCA jsou nové osy (hlavní komponenty) lineární kombinací měřených proměnných a proto předpokládáme lineární vztahy mezi proměnnými.

Příklad použití v ekologii je na **Error! Reference source not found.**. V tomto případě se jedná o ordinaci snímků (označených kolečky) a druhů (trojúhelníčky se zkrácenými názvy druhů) metodou korespondenční analýzy. Tato metoda předpokládá, že existují gradienty prostředí určující druhové složení a také že druhy mají na každém gradientu své optimum. Vidíme, že jsou zde shluky podobných snímků (např. snímky 5, 6, 7 a 10 si jsou velmi podobné). Druhy jsou charakterizovány svými optimy (na rozdíl od metody PCA, kde se předpokládá, že se zastoupení druhu s osou lineárně mění). První ordinační osa tedy bude odpovídat asi vlhkostnímu gradientu: vlhkomilné druhy *Ranunculus flammula*, *Juncus articulatus* na pravé straně oproti relativně suchomilným *Anthoxanthum odoratum* nebo *Plantago lanceolata* na levé straně.



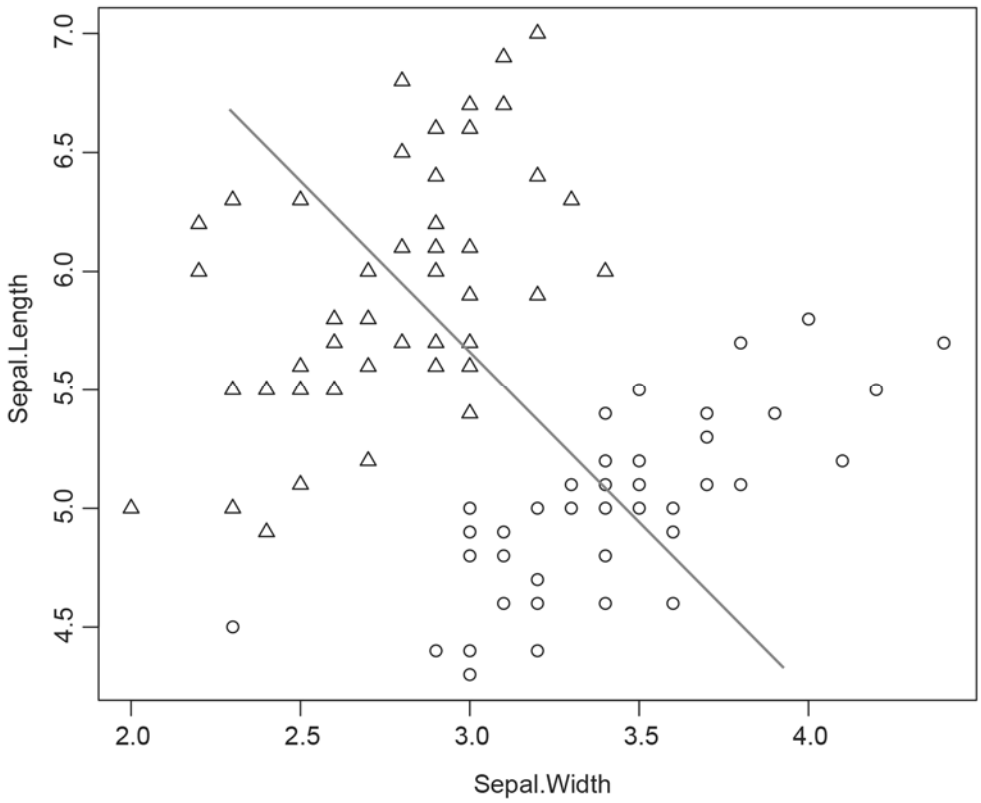
Obr. 19-1 Ordinační diagram korespondenční analýzy (*CA*) dat o složení travinné vegetace na ostrově Terschelling (viz sekci Příkladová data). První ordinační osa je vynesena horizotálně, druhá vertikálně. Plochy (snímky) jsou označeny kolečky, druhy vyplněnými trojúhelníčky.

Diskriminační analýza

V případě užití diskriminační analýzy (*discriminant analysis*) jsou objekty charakterizovány řadou proměnných, jako v předchozích případech, ale navíc jsou klasifikovány podle **nezávislého** kriteria. Hledáme pak klasifikační pravidlo (vyjádřené obvykle jako lineární kombinace hodnot měřených proměnných) nebo více takových pravidel, které predikují klasifikaci objektů do skupin. Příklady užití:

1. Byla sledována velká skupina zaměstnanců jednoho velkého podniku. Každému z nich byl proveden rozbor krve, rozbor moče, změřen krevní tlak atd. Poté byli všichni sledováni po dobu 5-ti let. Někteří z nich dostali během těchto 5-ti let infarkt (nemoc byla ve skutečnosti diagnostikována přesněji). Znamená to, že každý zaměstnanec je charakterizován jednak výsledky vyšetření (sem počítáme i věk) a jednak je klasifikován do dvou skupin - dostal / nedostal infarkt. Ptáme se, zda je možné s předstihem předpovědět na základě vyšetření (a podle jakého pravidla), zda člověk dostane infarkt.

2. Existují dva příbuzné, morfologicky velmi si podobné druhy, které se liší počtem chromozomů. Protože je obtížné počítat při každém určování chromosomy, ptáme se, zda je možné nalézt pravidlo, které by pomocí kombinace morfometrických údajů umožnilo druh spolehlivě určit. Zde je důležité, že příbuzné druhy byly odlišeny pomocí nějakého nezávislého kriteria (tady počet chromozómů) a při jejich určení nebyly užívány znaky použité v analýze.



Obr. 19-2 Ilustrace diskriminační funkce rozlišující dva druhy kosatce podle dvou znaků měřených na květech.

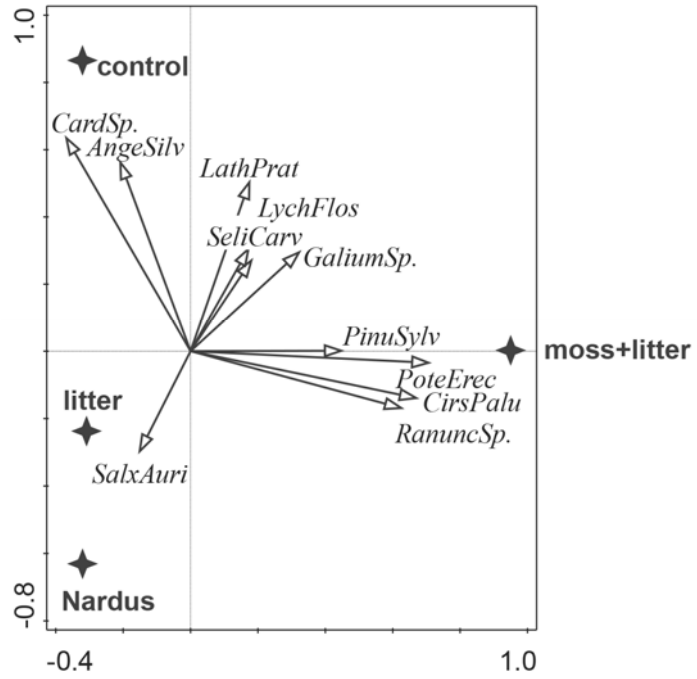
Princip diskriminační analýzy ilustruje Obr. 19-2 odpovídající druhém příkladu. Jde o zjednodušenou variantu prastarých dat, publikovaných R.A. Fisherem pro ilustraci lineární diskriminační analýzy (Fisher 1936). Naší snahou je odlišit dva druhy kosatců na základě dvou morfometrických údajů o jejich květech. Ani jedna z charakteristik není schopna sama tyto dva druhy odlišit. Můžeme ale zkonstruovat lineární funkci $Z = \lambda_1 X_1 + \lambda_2 X_2$ tak, že příslušníci jednoho druhu budou mít nízké hodnoty Z a příslušníci druhého druhu vysoké hodnoty, a tuto funkci použít pro jejich rozlišení. Diskriminační analýza má mnoho společného s mnohonásobnou regresí. Podobně jako v mnohonásobné regresi existuje i zde možnost automatického výběru prediktorů - tzn. že můžeme zjistit, že k rozlišení dvou druhů nám postačí pouze omezený počet proměnných z mnoha nabídnutých. Někteří statistici považují automatický výběr vhodných prediktorů za poněkud nebezpečný (pro diskriminační analýzu, ale i v jiných souvislostech) a dávají přednost výběru ručnímu. Objekty mohou být klasifikovány do více než dvou skupin. Pro rozlišení k skupin typicky potřebujeme $k-1$ diskriminačních funkcí.

Metody omezené ordinace (kanonické analýzy)

Metody omezené ordinace (*constrained ordination*) představují soubor metod, které hodnotí vzájemný vztah dvou skupin proměnných. Patří sem zejména metody CCA (*canonical correspondence analysis*) a RDA (*redundancy analysis*), ale existují i jiné, v ekologii méně používané (například *canonical correlation analysis*). Metody CCA a RDA jsou oblíbené zejména v ekologii společenstev. Při jejich typickém využití vycházíme ze soubor snímků (ploch, lokalit), charakterizovaných jednak zastoupením jednotlivých druhů a jednak změřenými charakteristikami prostředí. Druhů je obvykle mnoho, charakteristik prostředí

obyčejně jen několik (a mohou být jak kvantitativní, tak kvalitativní). Omezené ordinace umožní hodnotit vztah těchto dvou skupin proměnných. .

Zatímco metody neomezených (tradičních) ordinací (jako jsou PCA, CA nebo NMS) slouží především k explorační analýze dat a k navrhování hypotéz, tyto metody jsou velmi silným nástrojem pro testování hypotéz a jsou proto velmi vhodné k hodnocení experimentů na úrovni společenstva. Při testování používáme tzv. Monte Carlo permutační test. V těchto testech můžeme zohlednit různá experimentální uspořádání, např. úplné znáhodněné bloky. Příkladem mohou být výsledky na Obr. 19-3. V pokusu (Špačková et al. 1998) provedeném v úplných znáhodněných blocích byla v jednotlivých plochách odstraňována dominanta (*Nardus stricta*), stařina nebo stařina spolu s mechovým patrem. Bylo sledováno množství a druhové složení semenáčků na plochách. Pro vyhodnocení byla použita analýza RDA (použití tzv. kovariát dovolilo zohlednit blokovou strukturu pokusu). Monte Carlo permutační test prokázal významné rozdíly mezi jednotlivými zásahy. Diagram ukazuje vztah jednotlivých druhů k zásahům. Např. semenáčky druhů *Cirsium palustre, Potentilla erecta* a *Ranunculus* sp. se nacházely nejvíce na plochách, kde byla odstraněna stařina i mech.



Obr. 19-3 Ordinační diagram metody RDA (první dvě osy), shrnující výsledky pokusu, který zkoumal vliv odstranění dominanty (*Nardus*), opadu (*litter*) a kombinace opadu a mechového patru (*moss+litter*) na abundanci semenáčků lučních rostlin.

Metody mnohorozměrné analýzy jsou v poslední době velmi populární v ekologii. Dobré vysvětlení všech těchto metod je v publikacích Šmilauer & Lepš (2014), Jongman et al. (1987) a v mnohem detailnějším rozsahu pak v Legendre & Legendre (2012). Šíře použití a množství metod zdaleka přesahují rámec těchto skript a čtenář proto odkazujeme na uvedenou literaturu.

Příkladová data

Příkladová data pro ordinační metody jsou v listu *Chap19* rozdělena do dvou tabulek: sloupce A až AD představují údaje o přítomnosti a významnosti jednotlivých druhů rostlin (sloupce) v jednotlivých plochách (řádky) a sloupce AF až AI představují vysvětlující proměnné:

A1Horiz je mocnost horního půdního horizontu, *Moisture* je přibližný odhad půdní vlhkosti, *Mngmnt* je faktor charakterizující typ vlastníka (*SF*: standardní zemědělec, *BF*: biozemědělec, *HF*: rekreační, víkendový zemědělec a *NM*: obhospodařuje ochrana přírody), a *Manure* je množství aplikovaného hnoje. Pro metody neomezené ordinace budeme používat jen první z těchto tabulek, pro metody omezené ordinace obě.

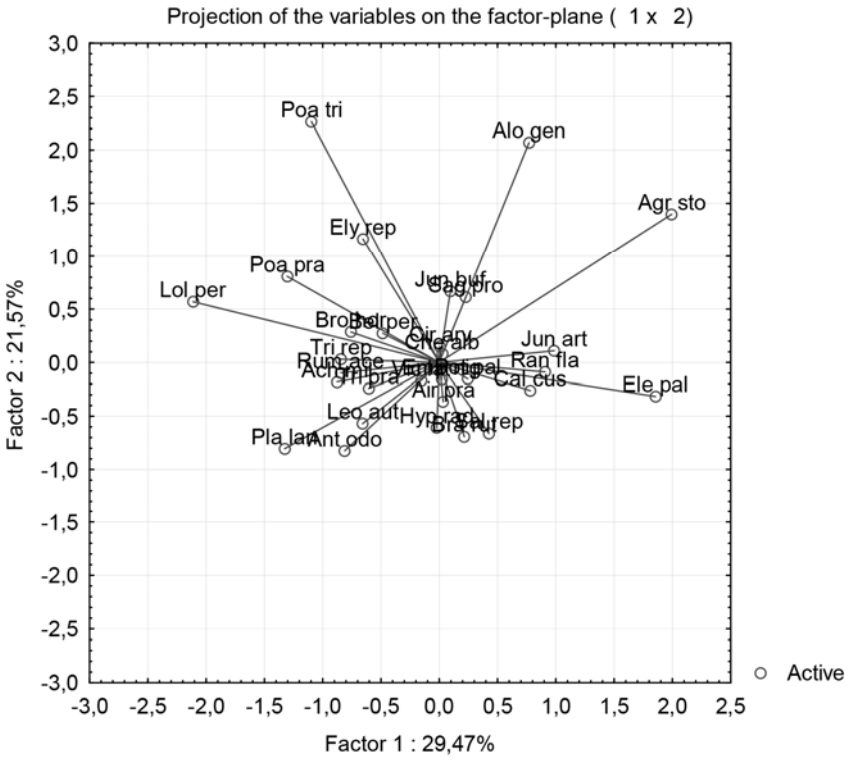
Pro ilustraci diskriminační analýzy použijeme taxonomická data, prvně publikovaná A.R. Fisherem (Fisher 1936). Ve sloupcích AK až AN jsou čtyři změřené charakteristiky květů tří druhů kosatců, příslušný druh je identifikován pro dané pozorování (řádek) ve sloupci AO (proměnná *Species*).

Jak postupovat v programu Statistica

Neomezené ordinační metody

Program Statistica nabízí analýzu hlavních komponent (PCA), korespondenční analýzu (CA) i nemetrické mnohorozměrné škálování (NMS), přičemž pro metodu NMS musíme jako vstupní data použít matici vzdáleností (nepodobností) mezi vzorky. Zde si ukážeme jen výpočet PCA.

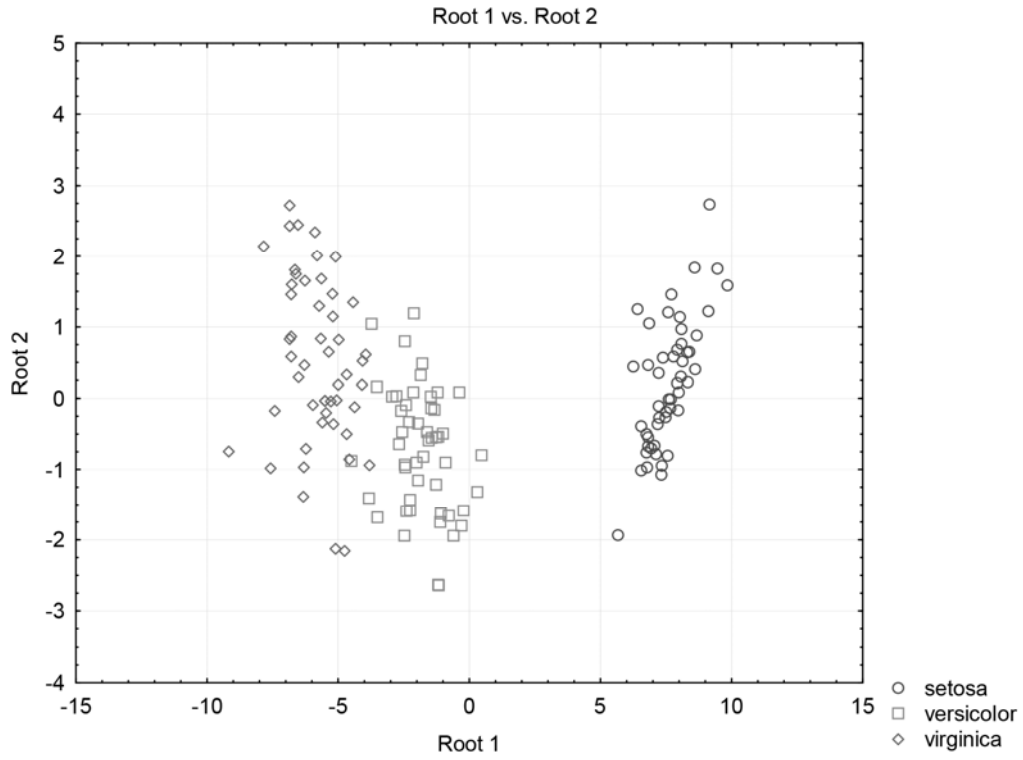
Z menu zvolíme příkaz *Statistics* | *Multivariate Exploratory Techniques* | *Principal Components & Classification Analysis*. Pomocí tlačítka *Variables* zadáme všechny proměnné reprezentující rostlinné druhy jako *Variables for analysis*. Pokud bychom chtěli v ordinačním diagramu odlišit jednotlivá pozorování do skupin (například definovaných proměnnou *Mngmnt*), můžeme příslušný faktor zadat jako *Grouping variable*. Na záložce *Advanced* si můžeme zvolit, zda je analýza založená na matici korelací či kovariancí. Vpodstatě tato volba určuje, zda jsou proměnné před analýzou standardizovány na jednotkovou varianci (*Correlations*) nebo ne (*Covariances*). Standardizace je třeba u proměnných s odlišnými jednotkami měření, v našem případě ale doporučujeme volbu *Covariances*. Po volbě tlačítka *OK* se objeví okno s výsledky (*Principal Components and Classification Analysis Results*). Ze záložky *Quick* můžeme vybrat buď diagram vynášející pozice proměnných (ilustrován níže) tlačítkem *Plot var. factor coordinates, 2D* nebo diagram s pozicemi vzorků – tlačítkem *Plot case factor coordinates, 2D*. V obou případech musíme nejprve vybrat, které ordinační osy (zde nazývané faktory) chceme vynášet. Nejvíce variability původních dat zobrazí diagram s prvou a druhou osou (tj. *Factor 1* pro *x-axis* a *Factor 2* pro *y-axis*).



Ordinační diagram je vhodné v publikaci doplnit informací, kolik procent z variability v původních datech vysvětlují zobrazené osy a tuto informaci získáme z tabulky (ve sloupci *% Total variance*) zobrazené tlačítkem *Eigenvalues*. Z grafu můžeme např. vyčíst, že druhy *Lolium perenne* a *Poa pratensis* se často vyskytují spolu (jejich pokryvnosti jsou pozitivně korelovány, a můžeme tedy předpokládat, že podobně odpovídají na charakteristiky prostředí), zatímco *Eleocharis palustris* je s nimi korelována negativně. S použitím našich znalostí ekologie těchto druhů se pak můžeme pokusit interpretovat první osu - asi bude souviset s vlhkostí, druhy na pravé straně jsou výrazně vlhkomilné.

Diskriminační analýza

Pro diskriminační analýzu jsme samostatně naimportovali z listu *Chap19* sloupce AK až AO. Z menu vybereme příkaz *Statistics* | *Multivariate Exploratory Techniques* | *Discriminant Analysis*. Pomocí tlačítka *Variables* zadáme proměnnou *Species* jako *Grouping variable* a zbylé čtyři proměnné (*Sepal.Length*,...) v *Independet variable list*. Dialogové okno s výsledky se zobrazí po zvolení tlačítka *OK*. Tlačítko *Summary* zobrazí informaci o významnosti (včetně testu průkaznosti) jednotlivých vysvětlujících (nezávislých) proměnných. Koeficienty diskriminačních funkcí lze získat pomocí tlačítka *Classification functions*. Pro zobrazení grafu, který umožní posoudit míru separace mezi třemi druhy, vybereme nejprve tlačítko *Perform canonical analysis* na záložce *Advanced* a pak v nově zobrazeném dialogovém okně zvolíme tlačítko *Scatterplot of canonical scores* na záložce *Canonical scores*.



Osy *Root 1* a *Root 2* představují (poněkud nešťastně pojmenované) dvě osy diskriminační analýzy (více jich není s ohledem na to, že se snažíme rozlišit tři typy objektů).

Omezené ordinační metody

Z této skupiny metod nabízí Statistica pouze kanonickou korelační analýzu (pomocí příkazu *Statistics* | *Multivariate Exploratory Techniques* | *Canonical Analysis*), která se s ohledem na její omezení v ekologii prakticky téměř nepoužívá. Pro provádění těchto metod (ale i metod neomezené ordinace) doporučujeme užití programu Canoco (ter Braak & Šmilauer 2012).

Jak postupovat v programu R

Každá ze tří tabulek v listu *Chap19* byla importována do samostatného datového rámce (*chap19a*, *chap19b* a *chap19c*).

Neomezené ordinační metody

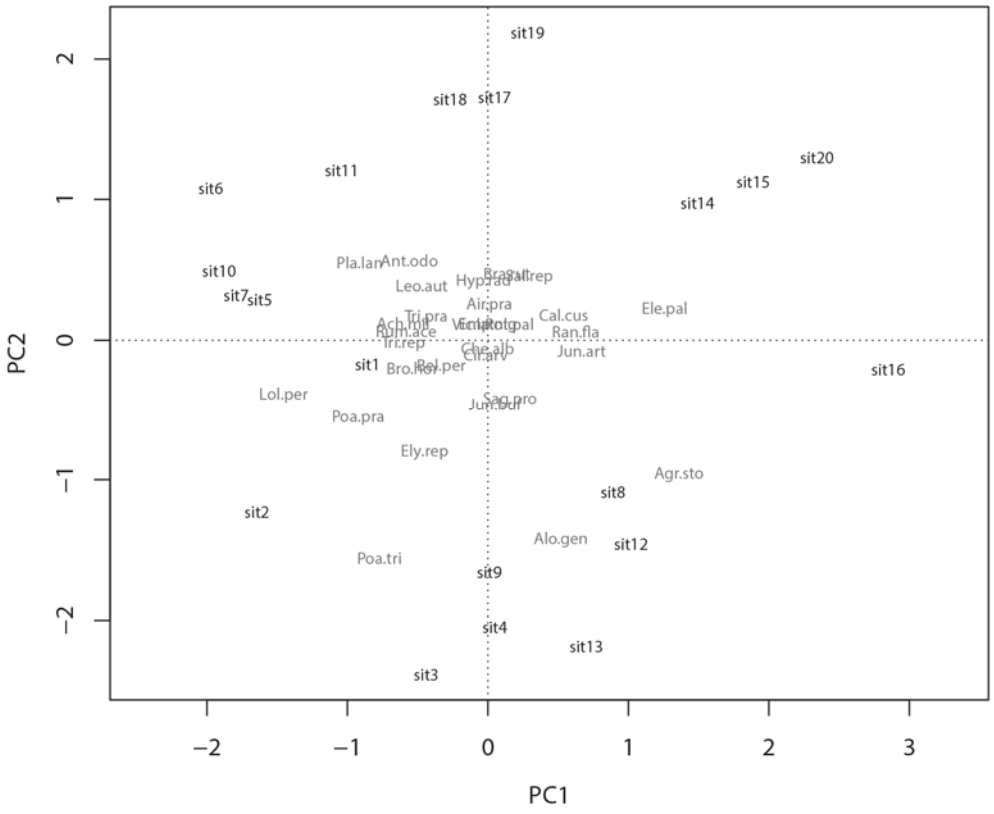
Pro metody neomezené ordinace doporučujeme užívat knihovnu *vegan*. Zde je příklad výpočtu analýzy hlavních komponent pro druhová data v datovém rámci *chap19a*:

```
> library(vegan)
> pca.1 <- rda(chap19a,scale=FALSE)
> summary(pca.1)

Call:
rda(X = chap19a, scale = FALSE)

Partitioning of variance:
              Inertia Proportion
Total           84.12          1
Unconstrained   84.12          1
```

```
Eigenvalues, and their contribution to the variance

Importance of components:
                          PC1     PC2     PC3     PC4    PC5     PC6     PC7     PC8    PC9
Eigenvalue           24.7953 18.1466 7.62913 7.15277 5.6950 4.33331 3.19936 2.78186 2.4820
Proportion Explained  0.2948  0.2157 0.09069 0.08503 0.0677 0.05151 0.03803 0.03307 0.0295
Cumulative Proportion 0.2948  0.5105 0.60115 0.68618 0.7539 0.80539 0.84342 0.87649 0.9060
...

Species scores
              PC1       PC2       PC3       PC4       PC5       PC6
Ach.mil -0.603786  0.12392  0.008464  0.159574  0.40871  0.127857
Agr.sto  1.373953 -0.96401  0.166905  0.266466 -0.08765  0.047368
Air.pra  0.023415  0.25078 -0.194768 -0.326043  0.05574 -0.079619
...

Site scores (weighted sums of species scores)
           PC1     PC2     PC3     PC4      PC5      PC6
sit1  -0.85678 -0.1724  2.6079 -1.1296  0.45074 -2.49113
sit2  -1.64477 -1.2299  0.8867 -0.9859  2.03463  1.81057
sit3  -0.44010 -2.3827  0.9297 -0.4601 -1.02783 -0.05183
...
```

Pro výpočet PCA používáme (poněkud nelogicky) funkci zvanou *rda*, pro kterou nebyly specifikovány vysvětlující proměnné. Volba *scale=FALSE* způsobí, že je PCA počítána z variančně-kovariančí matice, nikoliv z matice korelací, jednotlivé proměnné tudíž nejsou standardizovány na jednotkovou varianci. Řádek *Proportion explained* udává procento variability objasněné jednotlivými osami (ale na škále 0-1, ne 0-100). Ordinační diagram můžeme zobrazit pomocí funkce *plot*. Zde vynášíme souřadnice jednotlivých objektů a také souřadnice proměnných (druhů rostlin). Ale jak tady, tak v programu Statistica si můžeme vybrat, co do grafu vynášíme.

```
> plot(pca.1)
```



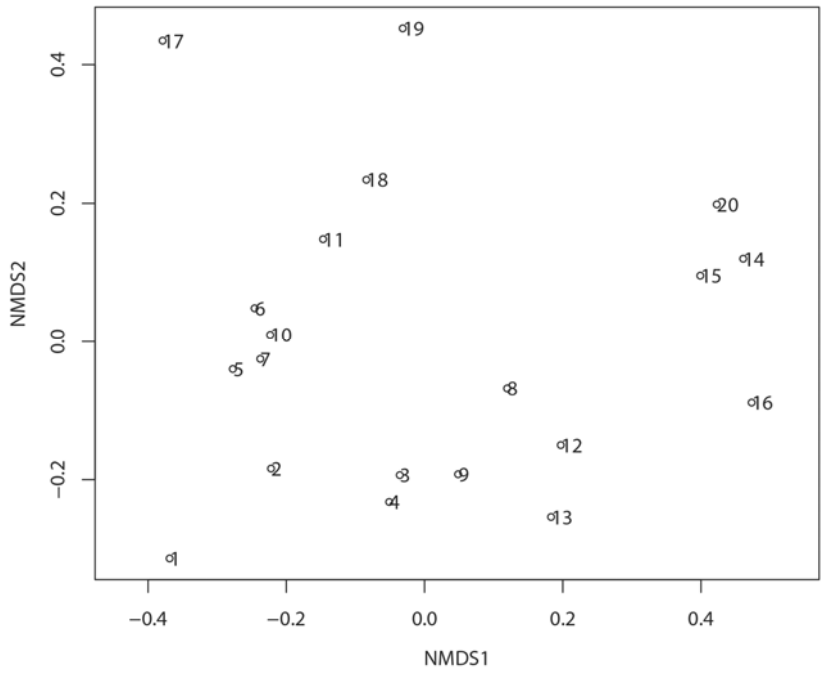
Z diagramu např. vyčteme, že stanoviště 5, 7 a 10 jsou si velmi podobná svým druhovým složením, naopak, stanoviště 16 je velmi odlišné. Porovnáním ordinačního diagramu druhů (proměnných) a stanovišť (objektů) můžeme usoudit, které druhy jsou za diferenciaci typů zodpovědné.

Ještě si ukážem výpočet nemetrického mnohorozměrného škálování (NMS) na základě předtím spočtené matice nepodobností. Knihovna *vegan* nabízí relevantní, v ekologii užívané míry nepodobnosti. V kombinaci s metodou NMS se často užívá Bray-Curtisova distance a matici s těmito distancemi spočteme pro naše data takto:

```
> dist.1 <- vegdist(chap19a,method="bray")
```

NMDS spočteme (s apriorně zvolenými dvěma osami) a výsledek zobrazíme takto:

```
> mds.1 <- metaMDS(dist.1, k=2)
Run 0 stress 0.1192678
Run 1 stress 0.1808916
Run 2 stress 0.1192678
... New best solution
... procrustes: rmse 5.437267e-05  max resid 0.0001662021
*** Solution reached
> plot(mds.1)
> text(mds.1,adj=0)
```



Diskriminační analýza

Diskriminační analýzu můžeme spočíst pomocí funkce *lda* v knihovně *MASS*:

```
> library( MASS)
> lda.1 <- lda(Species~.,data=chap19c)
> lda.1
Call:
lda(Species ~ ., data = chap19c)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
```

```
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Omezené ordinační metody

Pro omezené ordinační metody opět použijeme knihovnu *vegan* a funkce *cca* nebo *rda*, zde je příklad RDA metody pro data obsažená v datových rámcích *chap19a* (vysvětlované proměnné) a *chap19b* (vysvětlující proměnné).

```
> rda.1 <- rda(chap19a~A1Horiz+Moisture+Mngmnt+Manure,data=chap19b)
> summary(rda.1)

Call:
rda(formula = chap19a ~ A1Horiz + Moisture + Mngmnt + Manure, data = chap19b)

Partitioning of variance:
              Inertia Proportion
Total           84.12     1.0000
Constrained     47.11     0.5601
Unconstrained   37.01     0.4399
...
```

V uvedené počáteční části výstupu z funkce *summary* můžeme vyčíst, že námi zvolené čtyři proměnné vysvětlily 56% z variability v druhových datech (řádek *Constrained*, sloupec *Proportion*), vynechaná část obsahuje informace obdobné výstupu z PCA výše, jsou zde ale i souřadnice v ordinačním prostoru pro vysvětlující proměnné.

Vztah mezi složením lučního společenstva a vysvětlujícími proměnnými můžeme testovat následovně:

```
> anova(rda.1,step=1000)
Permutation test for rda under reduced model

Model: rda(formula = chap19a ~ A1Horiz + Moisture + Mngmnt + Manure, data =
chap19b)
         Df    Var      F N.Perm Pr(>F)
Model     6 47.114 2.7581    999  0.001 ***
Residual 13 37.010
```

Vidíme tedy, že mezi složením společenstva a hodnotami vysvětlujících proměnných je průkazný vztah (to neznamená nutně, že složení společenstva má průkazný vztah s každou z proměnných, obdobně jako v mnohonásobné regresi). Parametr *step* udává (zjednodušeně řečeno) počet permutací během testu.

Jiné programy

Pro analýzu ekologických dat ordinačními metodami doporučujeme program Canoco 5 (*http://www.canoco5.com*), který je uživatelsky přívětivější než programy popisované výše

a vytváří také kvalitnější ordinační diagramy (viz Obr. 19-1 nebo Obr. 19-3). Canoco 5 nabízí jak neomezené ordinace (PCA, CA, DCA, PCoA, NMDS), tak ordinace omezené (CCA, RDA) a další doplňující metody. Pro popis práce s tímto programem doporučujeme naši knížku Šmilauer & Lepš (2014).

Většinu z uváděných mnohorozměrných metod implementuje také program PC-Ord: *http://www.pcord.com*

Popis metod v článku

Methods

We have summarized the compositional community variation in studied grasslands using principal component analysis (PCA) calculated from centered (but not standardized) data.

The relation between grassland community composition and explanatory variables representing management regime and environmental characteristics was summarized using redundancy analysis (RDA) using centered response data. The significance of the relation was tested with Monte Carlo permutation test, using 999 permutations.

Results

The results of PCA are summarized in the ordination diagram in Figure X. The first two axes (shown in the diagram) explain 51% of the total variation in community composition.

We have found significant relation between plant community composition and selected explanatory variables (pseudo-F=2.76, p=0.001), which have explained 56% of the total variation in plant community composition. The results of RDA are summarized in the ordination diagram in Figure Y.

Citovaná literatura

Fisher R.A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**: 179-188.

Jongman R.H., ter Braak C.J.F. & van Tongeren O.F.R. (1987): Data analysis in community and landscape ecology. - Cambridge University Press.

Legendre P. & Legendre L. (2012): Numerical Ecology. Third English Edition. Elsevier, Amsterdam. p. 425-520 pro metody neomezené (klasické) ordinace a p.625-710 pro metody omezené ordinace (včetně diskriminační analýzy).

Sokal R.R. & Rohlf F.J. (1981): Biometry. 2nd ed. Freeman and comp., San Francisco. (discriminant analysis, pp. 683-687).

Šmilauer P. & Lepš J. (2014): Multivariate analysis of ecological data using Canoco 5. Cambridge University Press, Cambridge.362 pp.

20 Index