Petr Šmilauer: Statistika : cvičení s programem R –

Máme následující data z pokusu, kde byla ve čtvercích odhadnuta pokryvnost opadu, spočítány semenáče a změřena teplota: čtverec, pokryvnost opadu (Litt) a počet semenáčů (Seedl), zda byla plocha kosena (KOSEN: 1 - koseno, 0 - nekoseno) a teplotu (Temp).

Čtverec	Litt	Seedl	KOSEN	Temp
1	2	144	1	7.2
2	30	18	0	4.5
3	0	14	1	7.4
4	25	10	0	5.4
5	30	168	0	5.0
6	3	70	1	6.0
7	20	29	0	6.2
8	2	40	1	7.6
9	2	31	1	8.4
10	30	34	0	4.7
11	2	91	1	6.8
12	50	30	0	4.6
13	40	6	0	4.8
14	0	32	1	7.0
15	40	8	0	5.8
16	0	103	1	7.6
17	35	61	0	4.8
18	0	9	1	9.0
19	50	33	0	4.8
20	2	93	1	6.5
21	0	94	1	6.4
22	45	19	0	4.2
23	3	22	1	6.0
24	35	27	0	4.4

Zadejte si data v Excelu. Importujte data do programu R následujícím způsobem:
Tabulku v Excelu vybereme a zkopírujeme do schránky (*Clipboard*). Provedeme příkaz:

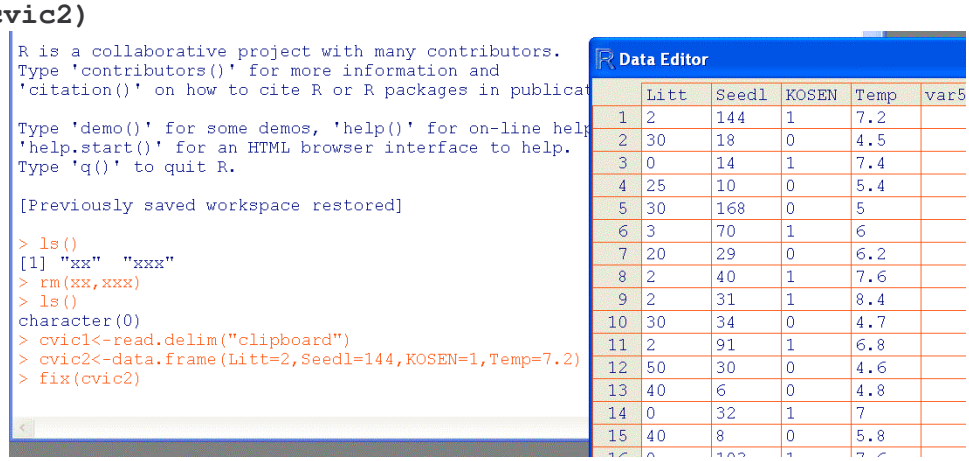
```
cvic1<-read.delim("clipboard")
```

Pokud máme nastavené české prostředí (desetinná čísla u proměnné *Temp* oddělují čárky, ne tečky), musíme použít funkci *read.delim2* místo *read.delim*. Tyto funkce vytvoří proměnnou typu datový rámec (*data frame*), se kterou dále pracujeme v našich analýzách a která obsahuje jednotlivé sloupečky jako proměnné.

R rozlišuje v názvech proměnných malá a velká písmenka. Dobrá rada – všeobecně nepoužívejte v programu R češtinu, diakritika dělá paseku.

Alternativně můžeme data zadat do tabulky (mnohe prostější než v Excelu, ale funkční) přímo v programu R. Pro pojmenování proměnných, případně určení jejich typu (v našem příkladu máme jen číselné proměnné, s faktory se seznámíme později), je dobré definovat první řádek tabulky a dále pokračovat přímo ve spreadsheetu. Tedy:

```
cvic2<-data.frame(Litt=2,Seedl=144,KOSEN=1,Temp=7.2)
fix(cvic2)
```



Jak spočítáme v programu R

Základní popisné statistiky, histogramy a pod. v programu R

Pro každou proměnnou spočítejte základní charakteristiky souboru. Nakreslete Box and whisker plot. Poté pro kosené a nekosené plochy:

```
summary(cvic1)
      Litt            Seedl             KOSEN          Temp
 Min.   : 0.00   Min.   :  6.00   Min.   :0.0   Min.   :4.200
 1st Qu.: 2.00   1st Qu.: 18.75   1st Qu.:0.0   1st Qu.:4.800
 Median :11.50   Median : 31.50   Median :0.5   Median :6.000
 Mean   :18.58   Mean   : 49.42   Mean   :0.5   Mean   :6.046
 3rd Qu.:35.00   3rd Qu.: 75.25   3rd Qu.:1.0   3rd Qu.:7.050
 Max.   :50.00   Max.   :168.00   Max.   :1.0   Max.   :9.000
```

Další statistiky lze pro jednotlivé proměnné v datovém rámci spočítat pomocí funkce *sapply*, např. varianci:

```
sapply(cvic1,var)
        Litt         Seedl         KOSEN          Temp            LS
 354.3405797 1969.2971014     0.2608696     1.8582428     0.8677656
```

nebo směrodatnou odchylku:

```
sapply(cvic1,function(x)sqrt(var(x)))
...
```

Vyneste body (X-opad, Y-počet semenáčů)

```
plot(Seedl~Litt,data=cvic1)
```

Změňte nadpisy, popisy os, fonty, vyhoďte z grafu mříž, vyhoďte přímku závislosti, změňte znak pro body. Pokud bychom chtěli doplnit hlavní popisku nebo změnit popis os, můžeme použít parametry *main*, *xlab* a *ylab*, například:

```
plot(Seedl~Litt,data=cvic1,main="Change of seedling counts with litter",
xlab="Plant litter cover", ylab="Number of seedlings")
```

Typ bodů můžeme změnit pomocí parametru pch (např. *pch=16* zvolí vyplněná kolečka). Přímku závislosti R automaticky nevloží, pokud bychom ji ale chtěli, lze pokračovat těmito příkazy:

```
lm.1<-lm(Seedl~Litt,data=cvic1)
abline(lm.1,col="blue")
```

Vytvořte další proměnnou, log(počtu semenáčů).

```
cvic1$LS<-log(cvic1$Seedl)
```

Přeneste graf do Wordu.

Když je okno s grafem aktivní, vyvoláme pravým tlačítkem myši menu a zvolíme "Copy as metafile". Následně vložíme ve Wordu pomocí *Edit / Paste*. Alternativně můžeme uložit ve vícero formátech pomocí *File / Save as …*

Zkuste vynést trojrozměrný graf (x-opad, y-teplota, z-log(počet semenáčů)).

Místo 3D bodového diagramu nabízím alternativní zobrazení – scatterplot matrix.

```
pairs(cvic1[,c(1,4,5)])
```

Exportujte data do Excelu.

```
write.table(cvic2,"clipboard",sep="\t",row.names=F)
```

Základní box-and-whisker plot:

```
library(lattice)
bwplot(~Seedl,data=cvic1)
```

Box-and-whisker rozdělený podle kategoriální proměnné:

```
bwplot(KOSEN~Seedl,data=cvic1)
```

Frekvenční histogram:

```
histogram(~Seedl,data=cvic1)
```

Pro rozdělení na kosené a nekosené plochy:

```
histogram(~Seedl|KOSEN,data=cvic1)
```

Co s chybějícími pozorováními

Chybějící hodnoty zadáváme předdefinovanou hodnotou *NA*. Některé metody umožňují volbu, jak s takovými případy zacházet, nebo můžeme všechny případy s chybějící hodnotou pro jednu nebo více proměnných vyloučit příkazem:

```
cvic1<-na.omit(cvic1)
```

Test dobré shody

Příklad ze skript: v F_1 generaci byly počty jedinců dominantního a recesivního fenotypu 70 a 10. Liší se od očekávaného mendelovského poměru 3:1? Očekávané frekvence jsou tedy 60:20.
V datovém souboru mám očekávané frekvence v jedné proměnné a napozorované frekvence v druhé (jedno v jednom sloupci, druhé v druhém).

```
chisq.test(c(70,10),p=c(0.75,0.25))
        Chi-squared test for given probabilities
data:  c(70, 10)
X-squared = 6.6667, df = 1, p-value = 0.009823
```

Protože p =0.0098, vidíme, že výsledek je průkazný nejenom na 5%, ale i na 1% hladině významnosti.

Občas může být problém, že funkce *chisq.test* automaticky předpokládá, že počet stupňů volnosti je počet kategorií - 1. Pokud tomu tak není (když třeba odhadujeme očekávané rozdělení na základě dat - např. porovnání s Hardy-Weinbergovskou rovnováhou), potom je správně spočten chi-kvadrát, ale dosaženou hladinu významnosti musíme spočíst sami, třeba takto:

```
pchisq(6.6667,1,lower.tail=F)
[1] 0.00982309
```

Kritickou hodnotu pro danou pravděpodobnost můžeme spočíst takto:

```
qchisq(0.05,3,lower.tail=F)
[1] 7.814728
```

Obdobné funkce pro F, t a normální distribuci se jmenují *pf* a *qf*, *pt* a *qt*, a *pnorm* a *qnorm*.

Cvičení - příklady:

1. Ve druhé filiální generaci (AA x aa) byl očekáván Mendelovský poměr dominantního a recesivního fenotypu 3 : 1. Z 200 zkoumaných individuí bylo 60% dominantních a 40% recesivních jedinců. Liší se výsledek od očekávaného poměru?

2. V první filiální generaci (AA x aa) bylo očekáváno, že všechna individua budou mít dominantní fenotyp. Z 2000 individuí se vyskytla 3 individua s recesivním fenotypem. Liší se výsledek od očekávaného?

```
chisq.test(c(1997,3),p=c(1,0))
        Chi-squared test for given probabilities
data:  c(1997, 3)
X-squared = Inf, df = 1, p-value < 2.2e-16
```

3. Ve druhé filiální generaci (AABB x aabb) je očekávaný poměr fenotypů (AB, Ab, aB, ab) 9:3:3:1. Skutečné počty byly 125, 60, 50, 12. Liší se poměry od očekávaných?

4. Byly zjišťovány preference včel pro určitou barvu. Včely byly po jedné vpouštěny do nádoby, kde byly umístěny terče čtyř barev, a bylo sledováno, na který včela poprvé usedne. Výsledné počty: na červený 10, na žlutý 25, na modrý 18, na zelený 6. Preferují včely některou barvu?

```
chisq.test(c(10,25,18,6))
        Chi-squared test for given probabilities
data:  c(10, 25, 18, 6)
X-squared = 14.5593, df = 3, p-value = 0.002235
```

Pokud nezadáme pravděpodobnosti (parametr *p*), jsou rovnoměrně rozděleny mezi kategorie.

5. V území se vyskytují dva typy luk: jeden lze charakterizovat jako Molinion, druhý jako Violion caninae. Z 59 hnízd puškvorečníka kostkovaného bylo 10 nalezeno v Molinionu a 49 ve Violionu caninae. Co můžeme na základě uvedených dat říci o prefenci puškvorečníka pro jednotlivé biotopy? Které údaje by bylo zajímavé znát, aby byl výsledek biologicky interpretovatelný? Čím ještě může být výsledek ovlivněn?

6. Severní Korea byla obviněna, že pomocí pilulek podávaných těhotným ovlivňuje pohlaví rodících se dětí, aby měla dost vojáků. Její ministerstvo veřejného zdraví a blahobytu vydalo prohlášení, ve kterém se tvrdí, že ze 124 815 dětí, narozených v posledním roce, bylo 62 407 chlapců, což dostatečně přesně odpovídá očekávanému poměru pohlaví 1:1. K jakému názoru dojdeme jako statistici?

7. Najděte kritickou hodnotu pro Chi-kvadrát na 5%-ní hladině významnosti při 3 stupních volnosti.

8. V populaci bylo nalezeno 15 jedinců genotypu AA, 20 jedinců genotypu Aa a 77 jedinců genotypu aa. Testujte shodu s H-W rovnováhou.

Kontingenční tabulky

Kontingenční tabulky se dají v R počítat alespoň dvěma způsoby. Vícerozměrné tabulky s odpovídající loglineární analýzou se dají počítat pomocí funkce *glm*, s volbou Poissonovské distribuce a s počty v buňkách kontingenční tabulky coby vysvětlovanou (závislou) proměnnou**.** Nejjednodušší je ale spočtení testu pro nezávislost řádků a sloupců dvourozměrné kontingenční tabulky pomocí funkce *chisq.test***.** Pokud máme hrubá data, mohu vytvořit kontingenční tabulku pomocí funkce. Příklad (skripta Obr. 3-1 str. 33) - mám data ve tvaru:

Jmeno Pohlavi BarvaVlasu
Novak muz cerna
Ticha zena blond
atd.
uložená v datovém rámci *vlpo* – mohu vypsat např. první tři řádky takto:

```
vlpo[1:3,]
   Jmeno Pohlavi BarvaVlasu
1  Novak     muz       cerna
2  Ticha    zena       blond
3 Hlucna    zena       blond
```

Kontingenční tabulku (založenou na druhé a třetí proměnné v datovém rámci *vlpo*) pak vytvořím pomocí funkce *table*:

```
vlpo.tab<-table(vlpo[,2:3])
```

a výslednou tabulku *vlpo.tab* mohu pak předat funkci *chisq.test*.

Jednodušší ale je, pokud již má spočtené četnosti pro jednotlivé buňky tabulky. Moje tabulka pak může vypadat takto:

Sex barva vlasů:	**černá**	**blond**	**hnědá**	**zrzavá**
muž	32	43	16	9
žena	55	65	64	16

Pro výpočet χ^2 testu pak nejprve vytvořím tabulku:

```
povl.tab<-rbind(c(32,43,16,9),c(55,65,64,16))
```

a spočítám test:

```
chisq.test(povl.tab)
        Pearson's Chi-squared test
data:  povl.tab
X-squared = 8.9872, df = 3, p-value = 0.02946
```

nebo provedu obojí najednou (výsledek jsem tentokrát uložil do proměnné, kterou musím zobrazit, abych dostal výsledek testu):

```
povl.chi<-chisq.test(rbind(c(32,43,16,9),c(55,65,64,16)))

povl.chi
        Pearson's Chi-squared test
data:  rbind(c(32, 43, 16, 9), c(55, 65, 64, 16))
X-squared = 8.9872, df = 3, p-value = 0.02946
```

Proměnná *povl.chi* obsahuje ale více informací:

```
names(povl.chi)
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals"
```

Můžeme tedy například zobrazit očekávané četnosti vypočtené pro jednotlivé vstupy tabulky:

```
povl.chi$expected
     [,1] [,2]     [,3]      [,4]
[1,]   29   36 26.66667  8.333333
[2,]   58   72 53.33333 16.666667
```

Zobrazit lze také residuály, nicméně jde o jiné residuály, než které zobrazuje např. program Statistica:

```
povl.chi$residuals
           [,1]       [,2]      [,3]       [,4]
[1,]  0.5570860  1.1666667 -2.065591  0.2309401
[2,] -0.3939193 -0.8249579  1.460593 -0.1632993
```

Jsou to tzv. Pearsonovy residuály tj. hodnoty (O-E)/√E. Součet jejich druhých mocnin je tedy roven X^2 statistice, tedy 8.9872 v našem příkladě.

Na orientaci tabulky (tj. který faktor definuje řádky a který sloupce) nezáleží, takže stejnou hypotézu (nezávislost faktorů) mohu testovat i takto:

```
chisq.test(rbind(c(32,55),c(43,65),c(16,64),c(9,16)))
        Pearson's Chi-squared test
data:  rbind(c(32, 55), c(43, 65), c(16, 64), c(9, 16))
X-squared = 8.9872, df = 3, p-value = 0.02946
```

Pokud zadáme čtyřpolní (2x2) tabulku, je automaticky prováděna korekce na spojitost (continuity correction), pokud si to ale nepřejeme, můžeme použít parametr *correct=FALSE*.

V mnoha případech (nejčastěji se užívá opět u čtyřpolních tabulek) můžeme vyhodnotit pravděpodobnost chyby prvního druhu nikoliv porovnání testové statistiky, ale pomocí randomizačního testu. Tento typ testu zvolíme pomocí parametru *simulate.p.value=T*. Pokud zvolíme dostatečně velký počet simulací (pomocí parametru *B*, implicitní hodnota jsou 2000), blíží se odhad *p* výsledku tzv. exaktního testu. Vzhledem ke způsobu odhadu (pomocí pseudo-náhodných čísel) je jasné, že při opakovaném volání dostáváme vždy trochu odlišné hodnoty:

```
chisq.test(rbind(c(32,43,16,9),c(55,65,64,16)),simulate.p.value=T)
        Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
data:  rbind(c(32, 43, 16, 9), c(55, 65, 64, 16))
X-squared = 8.9872, df = NA, p-value = 0.02749

chisq.test(rbind(c(32,43,16,9),c(55,65,64,16)),simulate.p.value=T)
        Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)
data:  rbind(c(32, 43, 16, 9), c(55, 65, 64, 16))
X-squared = 8.9872, df = NA, p-value = 0.02699
```

Příklady (pozor, ne všechny musí být na kontingenční tabulky):

1. V pokusu bylo sledováno, jaký vliv má na klíčivost semen hlohu skutečnost, že semeno projde zažívacím traktem kohouta. 50 plodů hlohu bylo dáno sežrat kohoutovi a semena byla po projití zažívacím traktem sebrána a dána klíčit. Dalších 50 semen bylo dáno klíčit přímo. Ze semen traktem prošlých vyklíčilo 56%, ze semen traktem neprošlých 22%. Má projití traktem vliv na klíčivost?

```
chisq.test(rbind(c(28,22),c(11,39)))
        Pearson's Chi-squared test with Yates' continuity correction
data:  rbind(c(28, 22), c(11, 39))
X-squared = 10.7608, df = 1, p-value = 0.001037
```

a totéž bez korekce:

```
chisq.test(rbind(c(28,22),c(11,39)),correct=F)
        Pearson's Chi-squared test
data:  rbind(c(28, 22), c(11, 39))
X-squared = 12.148, df = 1, p-value = 0.0004914
```

2. V Sierra Leone bylo do nemocnice přivezeno s cholerou během určitého časového okamžiku 120 osob. Přitom bylo zjišťováno, zda osoby byly očkovány proti tetanu. Z 65 osob očkovaných proti tetanu přežilo choleru 55. Z 55 osob neočkovaných přežilo 15. Můžeme na základě uvedených dat usoudit, že očkování proti tetanu chrání i před cholerou?

3. Z populace bylo náhodně vybráno 100 rostlin. Z nich bylo (zjištěno izozymovou analýzou) 12 dominantních homozygotů, 20 heterozygotů a zbytek recesivních homozygotů. Je populace v Hardy-Weinbergovské rovnováze?

4. Ze 120 studentek pedagogické fakulty otěhotnělo během sledovaného období 10%, ze 150 studentek biologie 20% a z 160 studentek sociálních věd 5%. Lišila se pravděpodobnost otěhotnění mezi studentkami jednotlivých fakult?

```
chisq.test(rbind(c(12,108),c(30,120),c(8,152)))
        Pearson's Chi-squared test
data:  rbind(c(12, 108), c(30, 120), c(8, 152))
X-squared = 17.3811, df = 2, p-value = 0.0001682
```

5. Během chřipkové epidemie z celkového počtu 170 studentů pedagogické fakulty onemocnělo 80, z 220 biologické fakulty 190 a z 290 studentů sociálních věd 22. Liší se odolnost studentů podle toho, kterou fakultu studují?

6. Byl studován vztah suchopýru úzkolistého a smilky tuhé v lučním společenstvu. Bylo náhodně umístěno 150 čtverců. Z nich 15 obsahovalo oba druhy, 65 pouze smilku, 45 pouze suchopýr a 25 neobsahovalo žádný z druhů. Co můžeme o vztahu těchto druhů říci?

```
chisq.test(rbind(c(15,65),c(45,25)))
        Pearson's Chi-squared test with Yates' continuity correction
data:  rbind(c(15, 65), c(45, 25))
X-squared = 30.385, df = 1, p-value = 3.542e-08
```

7. Předpokládejme, že výška studentů (v cm) má normální rozdělení se střední hodnotou 179 a variancí 121. Jaká je pravděpodobnost, že náhodně vybraný student bude pohodlně sedět v lavici, která je konstruována na výšku postavy 170 až 190 cm? Kolik bude mezi 550 studenty košíkářských postav (tj. s výškou 200 cm a vyšší)? Na jakou výšku musí být konstruovány lavice, aby vyhovovaly 99% posluchačů tak, aby stejnému počtu posluchačů byly příliš malé a stejnému počtu posluchačů příliš velké. Na jakou výšku musí být konstruovány lavice, aby 95% posluchačů nebyly malé?

```
pnorm(190,179,sqrt(121))-pnorm(170,179,sqrt(121))
[1] 0.634718
550*(pnorm(200,179,sqrt(121),lower.tail=F))
[1] 15.46885
```

nebo

```
550*(1-pnorm(200,179,sqrt(121)))
[1] 15.46885
qnorm(0.01/2,179,sqrt(121))
[1] 150.6659
```

- dolní hranice a pro horní hranici:

```
qnorm(0.01/2,179,sqrt(121),lower.tail=F)
[1] 207.3341
qnorm(0.95,179,sqrt(121))
[1] 197.0934
```

1. Test shody dat s rozdělením:

V programu R lze použít pro testování, zda hodnoty dané proměnné pochází z určité distribuce buď obecný Kolmogorov-Smirnovův test.

```
x<-rlnorm(15,0.5,0.2)  # generuje 15 hodnot z lognormální distribuce
x
 [1] 2.436958 1.715585 1.028551 1.650465 1.709717 1.732345 1.514716
 [8] 1.474961 1.563070 2.243551 1.487947 2.268291 2.103358 2.054014
[15] 2.166823
ks.test(x,"pnorm",2,0.5)
        One-sample Kolmogorov-Smirnov test
data:  x
D = 0.3038, p-value = 0.1004
alternative hypothesis: two.sided
```

Pro správnost tohoto testu je ale nutné, aby parametry distribuce, se kterou srovnáváme (v tomto případě normální) byly určeny a priorně, tedy nebyly odhadnuty z testované proměnné. Nemáme-li apriorní hypotézu o těchto parametrech, měli bychom (ale jen pro porovnání s normální distribucí) použít Shapiro-Wilkův test:

```
shapiro.test(x)
        Shapiro-Wilk normality test
data:  x
W = 0.9492, p-value = 0.5116
```

Graficky můžeme naše data porovnat s normální distribucí pomocí funkce qqnorm *(s vynesením referenční přímky, na které by body ležely pro normální distribuci, pomocí funkce* qqline*):*

```
qqnorm(x)
qqline(x,col="blue")
```

2. Jednovýběrový, dvouvýběrový a párový t-test:

Oboustranný jednovýběrový t-test spočítáme pomocí funkce t.test *takto:*

```
t.test(x,mu=2)
        One Sample t-test
data:  x
t = -1.9046, df = 14, p-value = 0.07758
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 1.596095 2.023952
sample estimates:
mean of x
 1.810024
```

Funkce nám zobrazí nejen testovou statistiku, stupně volnosti a signifikanci, ale i odhad průměru a jeho 95%-ní konfidenční interval. Pokud bychom chtěli testovat jednostranou hypotézu, museli bychom použít parametr alternative *s hodnotou* "greater" *nebo* "less" *(místo implicitní hodnoty* "two.sided"*).*

Dvouvýběrový t-test spočítáme pomocí stejné funkce, které ale zadáme dvě proměnné. Na rozdíl od programu Statistica používá funkce t.test *implicitně samostatný odhad variance pro obě srovnávané proměnné, pokud chceme, aby test předpokládal shodu variancí, použijeme parametr* var.equal=T*. V případě párového T-testu musí mít obě zadané proměnné stejnou délku a navíc musíme použít parametr* paired=T.

Jednostranné testy opět zvolíme pro dvouvýběrový či párový test pomocí nastavení správné hodnoty parametru alternative.

Příklady:

1. Byly měřeny délky plůdku ryb. Byla získána následující data:

Délka	kolikrát naměřeno:
3.5mm	3-krát
4mm	15
4.5 mm	26
5mm	18
5.5mm	7
6mm	3
6.5mm	2
7 mm	1
9mm	1

Co můžeme říci o datech? Pocházejí ze základního souboru s normálním rozdělením?
Testujte nulovou hypotézu, že data pocházejí ze základního souboru N(5,2).

```
pludek<-rep(c(seq(3.5,7,by=0.5),9), c(3,15,26,18,7,3,2,1,1))
shapiro.test(pludek)
        Shapiro-Wilk normality test
data:  pludek
W = 0.8231, p-value = 4.083e-08

ks.test(pludek,"pnorm",5,sqrt(2))
        One-sample Kolmogorov-Smirnov test
data:  pludek
D = 0.3158, p-value = 5.224e-07
alternative hypothesis: two.sided
```

2. Délka levé a pravé ruky u tenistů hrajících pravou rukou:

Levá	Pravá
55	57
61	63
65	66
55	57
51	50
52	56

Liší se délka levé a pravé ruky?

```
tenisti<-data.frame(L=c(55,61,65,55,51,52),P=c(57,63,66,57,50,56))
with(tenisti,{
  t.test(L,P,paired=T)})
        Paired t-test
data:  L and P
t = -2.5, df = 5, p-value = 0.05449
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.38038789  0.04705456
sample estimates:
mean of the differences
             -1.666667
```

3. 15 občanů bylo požádáno, aby vypili 4 piva a přitom byla zjišťována změna jejich hmotnosti. Cílem bylo zjistit, zda osoba přibude o předpokládané dva kilogramy. Zjištěné rozdíly byly (v kilogramech):
2.1, 2.2, 1.9, 1.8, 2.5, 2.6, 2.1, 1.7, 1.6, 1.9, 2.3, 2.1, 1.9, 1.8, 2.2.
Liší se v průměru změna váha od předpokladu?

```
t.test(c(2.1,2.2,1.9,1.8,2.5,2.6,2.1,1.7,1.6,1.9,2.3,2.1,1.9,1.8,2.2),
  mu=2)
        One Sample t-test
data:  c(2.1, 2.2, 1.9, 1.8, 2.5, 2.6, 2.1, 1.7, 1.6, ...,1.9, 1.8, 2.2)
t = 0.6341, df = 14, p-value = 0.5362
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 1.888826 2.204507
sample estimates:
mean of x
 2.046667
```

4. Krevní tlak u osob před a po podání léku, který má krevní tlak snížit: Před - Po:
(115 - 110); (125-121); (135-125); (150-152); (120-105); (105-99);(120-110); (108-102);
(99-90); (95-90).
Působí lék skutečně snížení tlaku? (Jaké by bylo správnější pokusné uspořádání?)

```
tlak<-data.frame(pred=c(115,125,135,150,120,105,120,108,99,95),
  po = c(110,121,125,152,105,99,110,102,90,90))
with(tlak,{
  t.test(pred,po,alternative="greater",paired=T)})
        Paired t-test
data:  pred and po
t = 4.7352, df = 9, p-value = 0.000533
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.167569      Inf
sample estimates:
mean of the differences
                   6.8
```

5. Na deseti trvalých plochách byly provedeny rozbory vegetace v roce 1980 a poté v roce 2000. Počty druhů (rok 1980, rok 2000) byly (20, 18), (15, 13), (25, 24), (25, 26), (30, 28), (32, 30), (20, 11), (22, 19), (11, 11), (13, 12).
Dochází v průměru ke změně druhové bohatosti společenstev?

```
bohatost<-data.frame(r1980=c(20,15,25,25,30,32,20,22,11,13),
   r2000=c(18,13,24,26,28,30,11,19,11,12))
with(bohatost,{
  t.test(r1980,r2000,paired=T)})
        Paired t-test
data:  r1980 and r2000
t = 2.473, df = 9, p-value = 0.0354
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1790155 4.0209845
sample estimates:
mean of the differences
                   2.1
```

Porovnání dvou vzorků: t-test a F-test

T-test: H0: průměry se mezi dvěma skupinami neliší.
F-test: H0: variance se mezi dvěma skupinami neliší.

Příklad: porovnáváme délky prašníků dvou pryskyřníků (*Ranunculus acer* a *R. nemorosus*). Od každého měření máme pět pozorování (měli bychom mít asi víc). Potom můžeme zadat data dvojím způsobem:

A. Každý vzorek má zvláštní proměnnou:

Acer	Nemor
5	7
6	8
4	9
6	6
5	8

nebo

B. Všechny hodnoty jsou uloženy v jedné proměnné, a druhá proměnná je klasifikační (říká nám, ke kterému druhu se daný údaj vztahuje):

Druh	delka
Ac	5
Ac	6
Ac	4
Ac	6
Ac	5
Ne	7
Ne	8
Ne	9
Ne	6
Ne	8

V prvém případě (A) vypočteme dvouvýběrový T test takto:

```
ran1<-list(Acer=c(5,6,4,6,5),Nemor=c(7,8,9,6,8))
with(ran1,{
  t.test(Acer,Nemor) })
        Welch Two Sample t-test
data:  Acer and Nemor
t = -3.7947, df = 7.339, p-value = 0.00619
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.8816117 -0.9183883
sample estimates:
mean of x mean of y
      5.2       7.6
```

a otestujeme shodu variancí takto:

```
with(ran1,{var.test(Acer,Nemor)})
        F test to compare two variances
data:  Acer and Nemor
F = 0.5385, num df = 4, denom df = 4, p-value = 0.5635
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.05606329 5.17166994
sample estimates:
ratio of variances
         0.5384615
```

V druhém případě (B) vypočteme T test takto:

```
ran2<-data.frame(Druh=factor(c(rep("Ac",5),rep("Ne",5))),
  delka = c(5,6,4,6,5,7,8,9,6,8))
t.test(delka~Druh,data=ran2)
        Welch Two Sample t-test
data:  delka by Druh
t = -3.7947, df = 7.339, p-value = 0.00619
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.8816117 -0.9183883
sample estimates:
mean in group Ac mean in group Ne
             5.2              7.6
```

a otestujeme shodu variancí pomocí Bartlettova testu:

```
bartlett.test(delka~Druh,data=ran2)
        Bartlett test of homogeneity of variances
data:  delka by Druh
Bartlett's K-squared = 0.3353, df = 1, p-value = 0.5625
```

Jednostranný test shody variancí je možný jen v případě funkce var.test*, a to opět pomocí parametru* alternative.

Příklady

1. Teplota jedné lázně byla měřena teploměry dvou firem (Termom a Celsimet), z nichž každá dodala 10 teploměrů. Cílem bylo zjistit, zda existuje systematická odchylka (tj. teploměry jedné firmy dávají v průměru vyšší hodnoty než teploměry druhé firmy) a zda jsou měření stejně přesná (tj. variabilita mezi teploměry je větší u některé z firem). Byla získána následující data:

Termom: 18, 19, 18, 17, 16, 19, 18, 17, 19, 18 Celsimet: 17, 15, 21, 20, 19, 22, 15, 16 ,18, 17

Otestujte jak vychýlení, tak přesnost měření. (Pozn. Kdyby to byla reálná data, tak bych si při téhle přesnosti nekoupil od žádné z firem už nikdy nic.)

```
teplomery<-list(Termom=c(18,19,18,17,16,19,18,17,19,18),
  Celsimet=c(17,15,21,20,19,22,15,16,18,17))
with(teplomery,{
  t.test(Termom,Celsimet)})
        Welch Two Sample t-test
data:  Termom and Celsimet
t = -0.1196, df = 11.888, p-value = 0.9068
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.923380  1.723380
sample estimates:
mean of x mean of y
     17.9      18.0

with(teplomery,{
   var.test(Termom,Celsimet)})
        F test to compare two variances
data:  Termom and Celsimet
F = 0.1648, num df = 9, denom df = 9, p-value = 0.01302
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.04093767 0.66354348
sample estimates:
ratio of variances
         0.1648148
```

2. Byla srovnávána váha semen dvou druhů, od každého druhu bylo užito deset semen (náhodně vybraných).

Váhy byly

1. druh: 15, 16, 17, 15, 16, 14, 15, 16, 19 , 19 2. druh: 14, 13, 15, 13, 16, 14, 12, 11, 13, 15

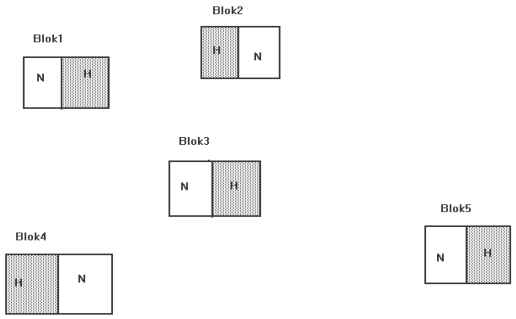
Liší se váhy semen uvedených druhů? Liší se variabilita semen u uvedených dvou druhů?

Spočtěte pro každý druh průměr, a charakteristiku variability a přesnosti odhadu.

Způsob řešení ako předchozí příklad, navíc ale výpočet průměru a charakteristik variability a přesnosti odhadu:

```
mean(druhy$druh1);sqrt(var(druhy$druh1))
[1] 16.2
[1] 1.686548
mean(druhy$druh2);sqrt(var(druhy$druh2))
[1] 13.6
[1] 1.505545
t.test(druhy$druh1)$conf.int
[1] 14.99352 17.40648
attr(,"conf.level")
[1] 0.95
t.test(druhy$druh2)$conf.int
[1] 12.52300 14.67700
attr(,"conf.level")
[1] 0.95
```

3. Pět bloků bylo rozděleno vždy na dvě plochy a polovina z nich byla pohnojena a polovina ne. Pokus tedy vypadal v terénu takto:



Biomasy v plochách byly:

Blok	1	2	3	4	5
Hnojeno	23	25	36	19	22
Nehnojeno	20	24	33	18	21

Má hnojení vliv?

```
hnojeni<-data.frame(hnojeno=c(23,25,36,19,22),
   nehnojeno=c(20,24,33,18,21))
with(hnojeni,{t.test(hnojeno,nehnojeno,paired=T)})
        Paired t-test
data:  hnojeno and nehnojeno
t = 3.6742, df = 4, p-value = 0.02131
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4398252 3.1601748
sample estimates:
mean of the differences
                    1.8
```

4. Deset krys bylo od narození krmeno stravou obohacenou hořčíkem, a deset bylo kontrolních (tatáž strava, ale bez obohacení hořčíkem). Předpokládalo se, že obohacení hořčíkem zvýší počet červených krvinek. Výsledky krevního obrazu (počet červených krvinek na objem) byly:

S hořčíkem: 85, 89, 79, 80, 91, 95, 79, 88, 89, 90.

Kontrola: 79, 80, 75, 79, 80, 71, 75, 80, 76, 80.

Je předpoklad o pozitivním vlivu hořčíku na krevní obraz správný?

```
strava<-list(Mg=c(85,89,79,80,91,95,79,88,89,90),
  Ctrl=c(79,80,75,79,80,71,75,80,76,80))
with(strava,{t.test(Mg,Ctrl,alternative="greater")})
        Welch Two Sample t-test
data:  Mg and Ctrl
t = 4.4814, df = 14.129, p-value = 0.0002531
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.465015      Inf
sample estimates:
mean of x mean of y
     86.5      77.5
```

Jak spočítáme v programu R

1. Mann-Whitney test

H0: mediány se mezi dvěma skupinami neliší [pak předpokládáme shodnost tvaru rozdělení; pokud tento předpoklad nemáme, je nulová hypotéza výběry pocházejí z totožných rozdělení]. Jedná se o neparametrickou obdobu dvouvýběrového t-testu. Jde-li odchylka od nulové hypotézy správným směrem, získáme dosaženou hladinu významnosti pro jednostranný test dělením hodnot hladiny významnosti pro dvoustranný test dvěma.

Příklad ze skript:

Porovnání výšek studentů a studentek

Studenti: 193, 188, 185, 183, 180, 178, 170; studentky: 175, 173, 168, 165, 163:

Data pro dvě porovnávané skupiny můžeme opět zadat buď jako dvě samostatné proměnné nebo pomocí datového rámce, ve kterém jedna proměnná kóduje příslušnost do skupiny, zatímco druhá reprezentuje vlastní hodnoty. Zde si ukážeme jen tuto druhou možnost:

```
vysky<-data.frame(osoba=factor(c(rep("student",7),rep("studentka",5))),
  vyska = c(193,188,185,183,180,178,170,175,173,168,165,163))
wilcox.test(vyska~osoba,data=vysky,paired=F)
        Wilcoxon rank sum test
data:  vyska by osoba
W = 33, p-value = 0.01010
alternative hypothesis: true mu is not equal to 0
```

2. Wilcoxonův test

Jedná se o neparametrickou obdobu párového t-testu, srovnává tedy přes mediány a pořadí pozorování, nikoliv přes průměry.

Příklad ze skript, srnčí nohy (porovnání délek předních a zadních noh u srnčí):

	ZADNI	PREDNI
1	142.000	138.000
2	140.000	136.000
3	144.000	147.000
4	144.000	139.000
5	142.000	143.000
6	146.000	141.000
7	149.000	143.000
8	150.000	145.000
9	142.000	136.000
10	148.000	146.000

Opět možno data zapsat oběma způsoby jako u Mann-Whitneyova testu, zde si pro změnu ukážeme zápis pomocí samostatných proměnných:

```
srnci<-data.frame(zadni=c(142,140,144,144,142,146,149,150,142,148),
  predni=c(138,136,147,139,143,141,143,145,136,146))
with(srnci,{wilcox.test(zadni,predni,paired=T)})
        Wilcoxon signed rank test with continuity correction
data:  zadni and predni
V = 51, p-value = 0.01859
alternative hypothesis: true mu is not equal to 0
```

Pokud bychom chtěli jednostranný test, opět můžeme použít parametr alternative.

Jednoduchý box-and-whisker diagram nakreslíme pomocí funkce boxplot. *Například:*

```
boxplot(vyska~osoba,data=vysky)
```

Příklady:

1. Byla porovnáváne reakce sedmi psů na obrázek černé kočky a bílé kočky. Reakce byla hodnocena na stupnici 1 - 6 (1 - nezájem, 2 - zavrčení .....6 - přímý útok). Výsledky:

Pes	Reakce na černou kočku	R. na bílou kočku
Alík	3	2
Vořech	6	4
Avar	2	2
Míša	5	1
Kačenka	6	6
Ferda	2	1
Alan	4	1

Liší se reakce podle barvy kočky?

```
na.kocku<-data.frame(cerna=c(3,6,2,5,6,2,4),bila=c(2,4,2,1,6,1,1))
with(na.kocku,{wilcox.test(cerna,bila,paired=T)})
        Wilcoxon signed rank test with continuity correction
data:  cerna and bila
V = 15, p-value = 0.05791
alternative hypothesis: true mu is not equal to 0
```

2. Byla porovnávána reakce dvou kultivarů smrku na zakouřené městské prostředí. Po deseti stromech každého kultivaru bylo umístěno poblíž rušné křižovatky po celou vegetačné sezonu a na jejím konci byl zdravotní stav hodnocen na stupnici 1 (zdravý strom, nejlepší zdravotní stav) - 5 (mrtvý strom).

Kultivar A: 2, 2, 1, 2, 3, 4, 2, 3, 1, 5

Kultivar B: 4, 5, 1, 1, 4, 3, 5, 1, 1, 2

Liší se uvedené kultivary svoji odolností?

```
smrcky<-data.frame(kultivar=factor(c(rep("A",10),rep("B",10))),
  stav=c(2,2,1,2,3,4,2,3,1,5,4,5,1,1,4,3,5,1,1,2))
wilcox.test(stav~kultivar,data=smrcky,paired=F)
        Wilcoxon rank sum test with continuity correction
data:  stav by kultivar
W = 49, p-value = 0.969
alternative hypothesis: true mu is not equal to 0
```

3. Na deseti osobách byla testována bolestivost dvou druhů očkování proti tetanu. Vždy do jednoho ramene byla použita vakcína od firmy A, do druhého ramene vakcína od firmy B, přičemž osoba nevěděla, od které firmy je které vakcína. Poté byla osoba požádána, aby ohodnotila bolestivost očkování na stupnici 0 až 10. Liší se vakcíny různých firem bolestivostí reakce?

Osoba	firma A	firma B
1	1	3
2	2	5
3	5	7
4	2	2
5	1	2
6	7	9
7	1	3
8	2	1
9	1	9
10	5	9

Jednocestná ANOVA

Příklad 8.1 ze skript (3 plemena králíků):

Data zadáváme do dvou sloupců obdobně jako pro dvouvýběrový t-test, jen skupin může být více než dvě. Pokud jsou skupiny dvě, výsledky t-testu a jednocestné ANOVy jsou shodné (pokud jich je víc, t-test nesmíme počítat). Jedna proměnná je klasifikační, tedy udává příslušnost objektu ke skupině (v tomto případě plemeno). Této proměnné se také říká kategoriální, neboť se jedná o kategoriální data. Druhá je odpověď, tedy naměřená hodnota (vaha) a říká se jí rovněž závislá proměnná.

plemeno	vaha
1	3
1	3
1	4
1	5
1	5
2	4
2	4
2	5
2	6
2	6
3	5
3	5
3	6
3	7
3	7

V programu R počítáme ANOVu pomocí funkce *aov*. Vysvětlující (klasifikační, kategoriální) proměnná musí být typu factor, proměnnou číselnou nebo znakovou převedeme pomocí funkce *factor*.

Předpoklady ANOVy jsou dva - normalita dat a homogenita variancí. Pokud dojde k výraznému narušení některého z předpokladů, můžeme podle situace: (1) provést transformaci dat, (2) užít zobecněné lineární modely nebo (3) použít neparametrickou obdobu daného testu. ANOVA je relativně robustní proti narušení svých předpokladů a to tím víc, čím víc máme pozorování.

*Pokud se rozhodneme testovat předpoklad o normalitě, musíme provést test pro každou skupinu zvlášť (a nebo lépe testovat normalitu reziduálů). Je třeba si ale uvědomit, že "chceme", aby test vyšel neprůkazně. Síla testu ovšem roste s počtem pozorování – a narušení normality je problémem především při malém počtu pozorování, kdy test je velmi slabý. Homogenitu variancí ověřujeme Bartlettovým testem (*bartlett.test*).*

```
kralici<-data.frame(plemeno=factor(rep(c(1,2,3),c(5,5,5))),
     vaha=c(3,3,4,5,5,4,4,5,6,6,5,5,6,7,7))
summary(kralici)
 plemeno      vaha
 1:5     Min.   :3
 2:5     1st Qu.:4
 3:5     Median :5
         Mean   :5
         3rd Qu.:6
         Max.   :7
```

Nejprve provedeme Bartlettův test:

```
bartlett.test(vaha~plemeno,data=kralici)
        Bartlett test of homogeneity of variances
data:  vaha by plemeno
Bartlett's K-squared = 0, df = 2, p-value = 1
```

(Dosažená hladina významnosti p=1, protože jsem si data vymýšlel tak, aby se mi to dobře počítalo, a všechny variance jsou si rovny. V realitě se to prakticky stát nemůže.)

Vlastní analýzu variance spočteme takto:

```
aov.1<-aov(vaha~plemeno,data=kralici)
```

Výsledek se uložil do objektu aov.1. *Tabulku analýzy variance pak zobrazíme takto:*

```
summary(aov.1)
            Df Sum Sq Mean Sq F value  Pr(>F)
plemeno      2     10       5       5 0.02634 *
Residuals   12     12       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum Sq *znamená sumu čtverců,* ***Df*** *počet stupňů volnosti (DF),* ***Mean Sq*** *průměrnou velikost sumy čtverců,* ***F value*** *je hodnota testovacího kritériea a* ***Pr(>F)*** *je dosažená hladina významnosti* p. *Význam jednotlivých položek je vysvětlen ve skriptech. Do výsledků se píše DF, F a p z řádku příslušícího klasifikační proměnné, v tomto případě plemeno.*

My teď sice víme, že se od sebe ta tři plemena liší ve váze, ale nevíme, jestli se liší každé od každého nebo jenom některá. V případě, že nám tedy hlavní test ANOVy vyšel průkazně, můžeme přistoupit k mnohonásobným porovnáním. Tyto testy se doporučuje provádět pouze pokud vyjde ANOVA průkazně. Budeme používat Tukey-ho test, za pomocí funkce *TukeyHSD*:

```
TukeyHSD(aov.1)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = vaha ~ plemeno, data = kralici)

$plemeno
    diff        lwr      upr     p adj
2-1    1 -0.6873051 2.687305 0.2907518
3-1    2  0.3126949 3.687305 0.0207238
3-2    1 -0.6873051 2.687305 0.2907518
```

Vidíme, že jediná průkazná odlišnost je mezi plemeny 1 a 3 (p=0.0207). Funkce nám zobrazuje 95% konfidenční interval (pokrytí lze změnit pomocí parametru conf.level*, který má implicitní hodnotu* 0.95*) pro hodnotu rozdílu průměrů srovnávaných dvou skupin. Signifikance (*p adj*) odpovídá testu hypotézy, že hodnota tohoto rozdílu je v základní populaci nulová.*

Příklady

1. 15 rostlin bylo rozděleno náhodně do tří skupin po pěti. Rostliny první skupiny byly pěstovány (každá ve zvláštním květináči) v písčité půdě, rostliny druhé skupiny v hlinité půdě a třetí skupiny v rašelině. Výšky rostlin na vrcholu sezóny:
písčitá: 15, 16, 18, 15, 21
hlinitá: 21, 20, 18, 25, 26
rašelina: 22, 26, 27, 30, 29
Má typ půdy vliv na výšku rostlin? Které skupiny se navzájem liší? Nakreslete si příslušný graf. Zkontrolujte homogenitu variancí.

```
aov.2<-aov(vyska~substrat,data=kytky)
summary(aov.2)
            Df  Sum Sq Mean Sq F value    Pr(>F)
substrat     2 240.133 120.067  13.004 0.0009906 ***
Residuals   12 110.800   9.233
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD(aov.2)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = vyska ~ substrat, data = kytky)
$substrat
               diff         lwr        upr     p adj
pisek-hlina    -5.0 -10.1271129  0.1271129 0.0561479
raselina-hlina  4.8  -0.3271129  9.9271129 0.0672929
raselina-pisek  9.8   4.6728871 14.9271129 0.0007081
```

```
plot(TukeyHSD(aov.2))
bartlett.test(vyska~substrat,data=kytky)
        Bartlett test of homogeneity of variances
data:  vyska by substrat
Bartlett's K-squared = 0.2959, df = 2, p-value = 0.8625
```

2. Ze tří náhodně vybraných rostlin všivce lesního byla sebrána semena. 5 semen z každé rostliny bylo za standardních podmínek necháno vyklíčit. Po deseti dnech byly změřeny délky klíčků jednotlivých rostlin (mm):
rostlina 1: 5, 7, 9, 8, 8
rostlina 2: 8, 7, 7, 9, 11
rostlina 3: 6, 8, 9, 11, 6
Liší se délky klíčků v závislosti na mateřské rostlině?

```
vsivce<-data.frame(delka=c(5,7,9,8,8, 8,7,7,9,11, 6,8,9,11,6),
   rostlina=factor(rep(c(1,2,3),rep(5,3))))
summary(aov(delka~rostlina,data=vsivce))
            Df Sum Sq Mean Sq F value Pr(>F)
rostlina     2  2.533   1.267  0.3958 0.6816
Residuals   12 38.400   3.200
```

3. Byla srovnávána váha semen dvou druhů, od každého druhu bylo užito 10 semen (náhodně vybraných). Váhy byly:
1. druh: 15, 16, 17, 15, 16, 14, 15, 16, 19, 19
2. druh: 14, 13, 15, 13, 16, 14, 12, 11, 13, 15
Liší se váhy semen uvedených druhů? Spočtěte pomocí parametrického testu, neparametrického testu a pomocí ANOVy.

```
semeno<-data.frame(druh=factor(rep(c("1","2"), c(10,10))),
  vaha=c(15,16,17,15,16,14,15,16,19,19,14,13,15,13,16,14,12,11,13,15))
summary(aov(vaha~druh,data=semeno))
            Df Sum Sq Mean Sq F value   Pr(>F)
druh         1 33.800  33.800  13.226 0.001886 **
Residuals   18 46.000   2.556
t.test(vaha~druh,data=semeno)
        Welch Two Sample t-test
data:  vaha by druh
t = 3.6368, df = 17.773, p-value = 0.001919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.096631 4.103369
sample estimates:
mean in group 1 mean in group 2
           16.2            13.6
```

Dvouvýběrový test v R a priori nepředpokládá shodu variancí, proto se *p* liší od ANOVA. Ale:

```
t.test(vaha~druh,data=semeno,var.equal=T)
        Two Sample t-test
data:  vaha by druh
t = 3.6368, df = 18, p-value = 0.001886
...

wilcox.test(vaha~druh,data=semeno,paired=F)
        Wilcoxon rank sum test with continuity correction
data:  vaha by druh
W = 88.5, p-value = 0.00355
alternative hypothesis: true mu is not equal to 0
```

Kruskal-Wallis(ův) test

Jedná se o neparametrickou obdobu jednocestné ANOVy.

Data zadáme stejně jako pro jednocestnou analýzu variance (tedy jedna proměnná je klasifikační, druhá proměnná obsahuje zjištěné hodnoty, opět data s králíky). Používáme funkci kruskal.test:

```
kralici<-data.frame(plemeno=factor(rep(c(1,2,3),c(5,5,5))),
  vaha=c(3,3,4,5,5,4,4,5,6,6,5,5,6,7,7))

kruskal.test(vaha~plemeno,data=kralici)
        Kruskal-Wallis rank sum test
data:  vaha by plemeno
Kruskal-Wallis chi-squared = 6.1072, df = 2, p-value = 0.04719
```

Srovnáme-li to s výsledkem jednocestné ANOVy pro stejná data, je vidět, že neparametrický Kruskal-Wallis je slabší test než ANOVA (jako všechny neparametrické testy v porovnání s parametrickými testy).

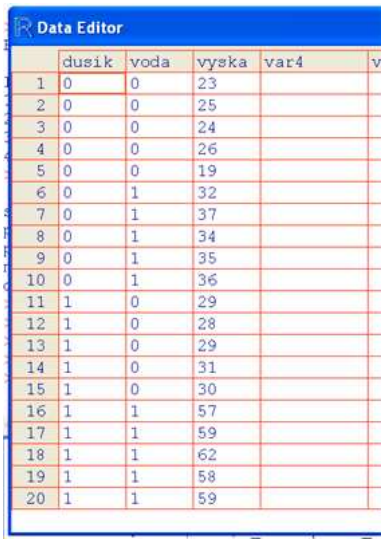
Dvoucestná ANOVA

V případě, že máme dvě nezávislé klasifikační proměnné (dusík byl-nebyl, voda byla-nebyla) a zkoumáme jejich vliv na jednu závislou měřitelnou proměnnou, jedná se o dvoucestnou ANOVu (kdyby byly tři, byla by trojcestná, atd.). Každou klasifikační proměnnou (faktor) zadáme do jedné proměnné (slouce). V příkladě to tedy budou proměnné DUSIK a VODA, měřené hodnoty zadáme do další proměnné (VYSKA). Data tedy budou vypadat:

	DUSIK	VODA	VYSKA
1	0	0	23
2	0	0	25
3	0	0	24
4	0	0	26
5	0	0	19
6	0	1	32
7	0	1	37
8	0	1	34
9	0	1	35
10	0	1	36
11	1	0	29
12	1	0	28
13	1	0	29
14	1	0	31
15	1	0	30
16	1	1	57
17	1	1	59
18	1	1	62
19	1	1	58
20	1	1	59

Data zadáme v R pomocí jednoduché "spreadsheetové tabulky" – nejprve vytvoříme hodnoty prvého řádku a pak voláme funkci *fix*:

```
voda.dus<-data.frame(dusik=0,voda=0,vyska=23)
fix(voda.dus)
```




```
voda.dus$dusik<-factor(voda.dus$dusik)
voda.dus$voda<-factor(voda.dus$voda)
```

Pokud nám jde pouze o efekty jednotlivých klasifikačních proměnných bez jejich interakce, vytvoříme ANOVA model s použitím funkce *aov* takto:

```
aov.3x<-aov(vyska~dusik+voda,data=voda.dus)
```

Pokud nás ale zajímá i jejich interakce (neaditivní spolupůsobení), zadáme model ANOVA následovně:

```
aov.3<-aov(vyska~dusik*voda,data=voda.dus)
```

K tomu však musíme mít faktoriální uspořádání pokusu (tady dusík je-voda není, dusík není-voda je, dusík je-voda je, dusík není-voda není). Stejný model by šlo popsat alternativně i oddělením hlavních efektů od interakce:

```
aov.3<-aov(vyska~dusik+voda+dusik:voda,data=voda.dus)
```

Tabulku analýzy variance, ve které jsou obsaženy testy pro každý z faktorů a pro jejic interakci, zobrazíme následovně:

```
summary(aov.3)
            Df  Sum Sq Mean Sq F value    Pr(>F)
dusik        1 1140.05 1140.05  288.62 1.165e-11 ***
voda         1 2101.25 2101.25  531.96 1.051e-13 ***
dusik:voda   1  414.05  414.05  104.82 1.977e-08 ***
Residuals   16   63.20    3.95
---
```

Další ostupy jsou obdobné jako u jednocestné ANOVy - prvně zkontrolujeme homogenitu variance, pomocí Bartlettova testu:

```
bartlett.test(vyska~voda:dusik,data=voda.dus)
        Bartlett test of homogeneity of variances
data:  vyska by voda by dusik
Bartlett's K-squared = 10.7325, df = 1, p-value = 0.001053
```

Post hoc dostaneme mnohonásobná porovnání provádíme funkcí TukeyHSD *(kde můžeme vybrat faktor, jehož hladiny chceme porovnávat, a to pomocí parametru* which, *např.* which="dusik"; *zde to ale nemá smysl, každý z faktorů má jen dvě hladiny).*

Interaction plot (tj. průměry pro kombinace hladin dvou faktorů, hladiny téhož faktoru spojeny čarou, takže rovnoběžnost čar odpovídá aditivitě efektů) dostaneme pomocí funkce interaction.plot *(ale bez konfidenčních intervalů, na rozdíl od programu Statistica):*

```
with(voda.dus,{
  interaction.plot(voda,dusik,vyska)})
```

Při tomto typu zadání se zadané faktory považují za faktory s pevným efektem. V případě, že potřebujeme zadat efekt s náhodným efektem, je to složitější. Faktor(y) s náhodným efektem, které neinteragují s faktory s pevným efektem (např. když máme úplně znáhodněné bloky, kdy je nutné zadat blok *jako náhodnou proměnnou, viz dále), můžeme zadat ve vzorci funkce* aov *například takto:* `vysvProm~zasah+Error(blok)`

Pokud by ale byla interakce mezi zasah-*em a* blok-*em, museli bychom zadat takto:*

`vysvProm~zasah*blok+Error(blok)`

Co je to náhodný a pevný efekt, záleží na naplánování pokusu a přístupu. Pokud máme pevně danou hladinu hnojiva, kterou přidáváme do květináče, je to pevný efekt. Pokud srovnáváme 3 louky, a jsme zemědělec, kterému ty louky patří, a tedy nás zajímají pouze tyto louky a všechny ostatní louky světa nám jsou ukradené, je louka taky faktor s pevným efektem. Pokud jsme však biologové a tyto 3 louky máme jenom jako výběr a chceme naše výsledky generalizovat na všechny louky, z nichž naše tři jsou náhodným výběrem, pak je to faktor s náhodným efektem. Více zajímavých informací ale pro faktory s náhodným efektem dostaneme až při použitím lineárních modelů se smíšenými efekty (linear mixed effect models, LME, v package nlme, *funkce* lme*).*

Příklady

1. Bylo vybráno 30 osob, a každé z nich byla sdělena skutečnost, která měla vyvola radostnou reakci. Intenzita radostné reakce byla hodnocena na stupnici od 0 (odpověď typu „a co já s tím mám dělat“), přes 1 (odpověď typu „to mě opravdu těší“), atd. až k 10 (radostí nepříčetný). Deseti osobám bylo sděleno, že zcela mimořádně dostávají v restituci pivovar, dalším deseti, že budou jmenováni ministrem školství, a dalším deseti, že vyhráli rekreační zájezd do Ostravy: reakce na jednotlivá sdělení:

Pivovar: 5, 7, 8, 6, 8, 7, 9, 10, 6, 9

Min. škol: 1, 0, 2, 3, 8, 1, 3, 5, 5, 2

Ostrava: 4, 3, 2, 0, 2, 3, 2, 5, 2, 3

Liší se intenzita radostné reakce podle typu sdělení?

```
radost<-
   data.frame(pricina=factor(rep(c("pivo","MSMT","Ostrava"),c(10,10,10))),
   vysledek=c(5,7,8,6,8,7,9,10,6,9,1,0,2,3,8,1,3,5,5,2,4,3,2,0,2,3,2,5,2,3))

kruskal.test(vysledek~pricina,data=radost)
        Kruskal-Wallis rank sum test
data:  vysledek by pricina
Kruskal-Wallis chi-squared = 16.8746, df = 2, p-value = 0.0002166
```

2. V pokusu jsme měli 10 samců a 10 samic krysy. Pětici samic a pětici samců byl od narození podáván hormon ledvinkin. Po dvou měsících byly krysy pitvány a určena váha ledvin. Má hormon ledvinkin vliv na váhu ledvin, liší se váha ledvin u samců a u samic? Je vliv ledvinkinu různý u samců a u samic?

	Samci	Samice
bez ledvinkinu	5, 8, 7, 6, 7	8, 9, 7, 8, 9
s ledvinkinem	11, 15, 14, 12, 15	23, 19, 18, 21, 23

```
pokus<-data.frame(sex=factor(rep(c("mal","fem","mal","fem"),c(5,5,5,5))),
  hormone=factor(rep(c("no","yes"),c(10,10))),
  weight=c(5,8,7,6,7,8,9,7,8,9,11,15,14,12,15,23,19,18,21,23))

summary(aov(weight~sex*hormone,data=pokus))
            Df Sum Sq Mean Sq F value    Pr(>F)
sex          1 101.25  101.25  38.571 1.248e-05 ***
hormone      1 470.45  470.45 179.219 4.153e-10 ***
sex:hormone  1  42.05   42.05  16.019  0.001027 **
Residuals   16  42.00    2.63
```

3. Na deseti rostlinách byla porovnávána hustota průduchů na listech, na korunních plátcích a na řapíku. Byly zjištěny následující hodnoty:

Rostlina	Listy	Korunní plátky	Řapík
1.	9	6	7
2.	15	9	10
3.	7	3	4
4.	15	10	12
5.	11	7	9
6.	20	15	17
7.	19	18	18
8.	4	3	3
9.	16	11	13
10.	14	10	11

Liší se hustoty průduchů na jednotlivých částech rostliny?

```
pruduchy<-data.frame(rostlina=factor(rep(1:10,rep(3,10))),
  misto = factor(rep(c("listy","platky","rapik"),10)),
  hustota=c(9,6,7, 15,9,10, 7,3,4, 15,10,12, 11,7, 9,
            20,15,17, 19,18,18, 4,3,3, 16,11,13, 14,10,11))

summary(aov(hustota~misto+Error(rostlina),data=pruduchy))

Error: rostlina
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9 653.47   72.61

Error: Within
          Df Sum Sq Mean Sq F value    Pr(>F)
misto      2 75.467  37.733  46.734 7.466e-08 ***
Residuals 18 14.533   0.807
---
```

Tímto způsobem zadáváme také opakované měření na jednom jedinci (na stejné ploše), v rámci tzv. *repeated measures ANOVA*. Identita jedince (plochy) vystupuje ve stejné roli jako *rostlina* vyse – tj. jako aditivní faktor s náhodným efektem.

Friedmanův test

Jedná se o neparametrickou obdobu ANOVy pro úplné znáhodněné bloky. Data je nutné zadat stejně, jako když počítáme ANOVu pro úplné znáhodněné bloky (viz předchozí strana):

```
pruduchy<-data.frame(rostlina=factor(rep(1:10,rep(3,10))),
  misto = factor(rep(c("listy","platky","rapik"),10)),
  hustota=c(9,6,7, 15,9,10, 7,3,4, 15,10,12, 11,7, 9,
            20,15,17, 19,18,18, 4,3,3, 16,11,13, 14,10,11))

friedman.test( hustota ~ misto | rostlina, data=pruduchy)
        Friedman rank sum test
data:  hustota and misto and rostlina
Friedman chi-squared = 19.1579, df = 2, p-value = 6.917e-05
```

Hierarchické uspořádání ANOVy

Jedná se o typ ANOVy, ve kterém máme ve skupině podskupiny, např. tři druhy rostlin pěstujeme každý v 5 květináčích, v každém po 5 kytkách a měříme jejich výšku - květináč je hierarchicky nižší kategorie než druh. Není to faktoriální uspořádání, protože první květináč druhu A není srovnatelný s prvním květináčem druhu B či C.

Příklad: rostliny byly pěstovány ve dvou typech substrátu (vždy 2 a 2; je to strašně málo, ale pro ilustraci to stačí, ve skutečnosti by to nestačilo). Z každé rostliny byly odebrány tři okolíky, a v každém z nich byl spočten počet větví. Data vypadají následujícím způsobem, a tak jsou také v programu zadána:

	SUBSTRAT	KYTKA	POCET
1	A	1	6
2	A	1	9
3	A	1	7
4	A	2	9
5	A	2	10
6	A	2	11
7	B	3	5
8	B	3	6
9	B	3	4
10	B	4	7
11	B	4	7
12	B	4	9

Potřebujeme tedy spočítat hierarchickou ANOVu, v níž máme faktory s pevným i náhodným efektem. Postup s použitím funkce aov *nám umožňuje správně otestovat efekt substrátu, nikoliv však signifikanci náhodného efektu identity kytka (ten nás ale často nezajímá):*

```
summary(aov(pocet~substrat+Error(kytka),data=okolik))
```

```
Error: kytka
          Df Sum Sq Mean Sq F value Pr(>F)
substrat   1 16.333  16.333  1.5313 0.3415
Residuals  2 21.333  10.667

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  8 11.3333  1.4167
```

Residuální mean square na hladině variability mezi kytkami (hodnota 10.667, v prvním z řádků začínajících slovem Residuals*, je současně velikostí náhodného efektu* kytka*, ale není zde takto testován. Mohli bychom jej ale testovat sami (v následujícím příkazu je vypočtená hodnota* F *statistiky rovnou porovnávána s F distribucí):*

```
1-pf(10.667/1.4167,2,8)
[1] 0.01448781
```

*Výsledná hodnota (*0.01449*) tedy představuje signifikanci testu náhodného efektu* kytka*. Oba efekty jsou vidět i z přímého hierarchického zadání, ve kterém je vidět vnořenost faktoru* kytka *do faktoru* substrat*, zde ovšem F testy nejsou vůbec počítány, museli bychom spočíst oba sami:*

```
summary(aov(pocet~substrat/kytka+Error(kytka),data=okolik))
Error: kytka
               Df Sum Sq Mean Sq
substrat        1 16.333  16.333
substrat:kytka  2 21.333  10.667

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  8 11.3333  1.4167
1-pf(16.333/10.667,1,2)
[1] 0.341505
```

```
1-pf(10.667/1.4167,2,8)
[1] 0.01448781
```

Alternativně bychom mohli použít i funkce z knihovny nlme*, určené pro fitování lineárních a nelineárních modelů se smíšenými efekty (mixed-effect linear and non-linear models), ty se však obvykle používají ve složitějších situacích:*

```
library(nlme)
anova(lme(pocet~substrat, random=~1|kytka,data=okolik))
            numDF denDF  F-value p-value
(Intercept)     1     8 63.28077  <.0001
substrat        1     2  1.53124  0.3415
```

Test pro (Intercept) *není pro nás zajímavý (testuje hypotézu, že průměrný počet větviček v okolíku je nulový).*

Příklady:

1. Byl testován vliv výluhu dvou druhů (Artemisia vulgaris a Cirsium arvense) na klíčivost semen vlčího máku. Ve čtyřech skupinách (každá po pěti Petriho miskách) bylo zjišťováno procento vyklíčených semen. Uvedená procenta byla:

zalévané destil. vodou: 98, 96, 92, 90, 94

výluh Artemisia : 88, 86, 82, 80, 86

výluh Cirsium : 78, 72, 68, 70, 72

výluh obou : 74, 76, 76, 70, 72

Co můžeme říci o alelopatickém působení výluhů na klíčivost?

```
alelop<-data.frame(AV=factor(rep(c("ne","ano","ne","ano"),rep(5,4))),
                   CV=factor(rep(c("ne","ano"),rep(10,2))),
  klic=c(98,96,92,90,94,88,86,82,80,86,78,72,68,70,72,74,76,76,70,72))

summary(aov(klic~AV*CV,data=alelop))
            Df Sum Sq Mean Sq  F value    Pr(>F)
AV           1   80.0    80.0   7.6923  0.013562 *
CV           1 1344.8  1344.8 129.3077 4.468e-09 ***
AV:CV        1  156.8   156.8  15.0769  0.001320 **
Residuals   16  166.4    10.4
interaction.plot(alelop$CV,alelop$AV,alelop$klic)
```

2. Byl testován vliv tří typů stravy na koncentraci cukru v krvi. Bylo užito 12 krys, tři skupiny po čtyřech. Rozbor krve byl u každé krysy dělán dvakrát. Výsledky z téže krysy jsou spojeny znaménkem &. (Tj. 12 & 14 znamená, že na téže kryse byla dvěma paralelními odběry zjištěna hodnota 12 a 14)

typ stravy

A 12&14, 15&16, 14&15, 11&12

B 19&17, 19&21, 22&21, 19&20

C 12&12, 15&14, 11&12, 10&11

Vyhodnoťte pokus.

```
strav<-data.frame( strava=factor(rep(c("A","B","C"),c(8,8,8))),
                   mys=factor(rep(1:12,rep(2,12))),
  konc=c(12,14,15,16,14,15,11,12, 19,17,19,21,22,21,19,20,
  12,12,15,14,11,12,10,11))
strav
  strava mys konc
1      A   1   12
2      A   1   14
3      A   2   15
...
```

```
summary(aov(konc~strava+Error(mys),data=strav))
Error: mys
          Df  Sum Sq Mean Sq F value    Pr(>F)
strava     2 261.083 130.542  24.350 0.0002338 ***
Residuals  9  48.250   5.361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals 12 10.0000  0.8333
```

3. Je známo, že směrodatná odchylka počtu individuí roupic v půdní sondě je přibližně lineárně závislá na hustotě individuí (tj. na průměrném počtu roupic v sondě). Cílem sledování bylo porovnat populační hustotu v lokalitách Budějovice, Lišov a Třeboň. V každé lokalitě bylo osm sond a byly získány následující počty individuí v sondách:

Budějovice: 12, 8, 15, 22, 25, 0, 10, 12

Lišov: 51, 121, 214, 10, 10, 195, 29, 16

Třeboň: 2, 15, 22, 0, 17, 33, 31, 0

Vyhodnoťte pokus.

```
roupice<-data.frame(lokalita=factor(rep(c("CB","LI","TR"),rep(8,3))),
  pocty=c(12,8,15,22,25,0,10,12, 51,121,214,10,10,195,29,16,
          2,15,22,0,17,33,31,0))
summary(aov(log(pocty+1)~lokalita,data=roupice))
            Df Sum Sq Mean Sq F value  Pr(>F)
lokalita     2 13.363   6.682  4.1466 0.03036 *
Residuals   21 33.839   1.611
---
```

Alternativně lze varianci ve skupinách stabilizovat i odmocninnou transformací:

```
summary(aov(sqrt(pocty)~lokalita,data=roupice))
            Df  Sum Sq Mean Sq F value  Pr(>F)
lokalita     2 111.071  55.535  5.4447 0.01245 *
Residuals   21 214.199  10.200
---
```

A další možností je použití zobecněného lineárního modelu s distribucí zvolenou jako kvasi-Poissonovskou (variance zde roste rychleji než u Poissonovské):

```
> anova(glm(pocty~lokalita,data=roupice,family=quasipoisson(link=log)),
+ test="F")
Analysis of Deviance Table
Model: quasipoisson, link: log
Response: pocty

Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev      F   Pr(>F)
NULL                        23    1380.32
lokalita  2   609.71        21     770.60 8.6693 0.001799 **
---
```

4. Existuje teorie, že výstražné zbarvení hmyzu s žihadlem je tím nápadnější, čím bolestivější je jeho bodnutí. Prvním krokem k ověření teorie byla určitá kvantifikace bolestivosti bodnutí; v rámci pokusu byly užity čtyři druhy vos a sršňů a 10 pokusných osob. Každá osoba byla jednou pokusně bodnuta každým druhem hmyzu a měla kvantifikovat svoje

pocity číslem od jedné (nic moc) do deseti (zatraceně bolí). Testujte hypotézu, že se uvedené druhy hmyzu neliší v bolestivosti bodnutí.

OSOBA	DRUH1	DRUH2	DRUH3	DRUH4
1	4	5	9	2
2	5	6	8	5
3	2	6	10	3
4	5	9	9	7
5	2	3	4	3
6	8	9	10	7
7	5	6	9	7
8	3	7	9	6
9	1	7	9	5
10	3	4	6	6

```
bodani<-data.frame( osoba=factor(rep(1:10,rep(4,10))),
                    druh=factor(rep(1:4,10)),
  pocit=c(4,5,9,2, 5,6,8,5, 2,6,10,3, 5,9,9,7, 2,3,4,3,
          8,8,10,7, 5,6,9,7, 3,7,9,6,  1,7,9,5, 3,4,6,6))

friedman.test(pocit~druh|osoba,data=bodani)
        Friedman rank sum test
data:  pocit and druh and osoba
Friedman chi-squared = 23.1474, df = 3, p-value = 3.763e-05
```

Jednoduchá lineární regrese

Pomocí regrese či korelace popisujeme vztah dvou kontinuálních proměnných. V případě regrese jsme schopni rozhodnout, která proměnná je závislá a která nezávislá. O nezávislé proměnné předpokládáme, že je změřena přesně (v praxi stačí, když chyba měření je u nezávislé proměnné mnohem menší než u závislé). Výsledkem je regresní rovnice, pomocí níž můžeme predikovat hodnotu závislé proměnné při určité hodnotě nezávislé proměnné, a koeficient determinace, R^2, který nám říká míru vysvětlené variability. V případě korelace není jasné, která proměnná je závislá a která nezávislá. Korelační koeficient nám vyjadřuje těsnost vazby. Je zde ovšem těsná souvislost, koeficient determinace je druhou mocninou korelačního koeficientu.

Příklad: Byla studována závislost rychlosti transpirace (TRANSPI) na rychlosti větru (VITR). Byla získána následující data:

	VITR	TRANSPI
1	2	12
2	9	16
3	5	14
4	6	15
5	7	18
6	3	11
7	4	12
8	1	10
9	0	8

Je logické, že transpirace je závislá na rychlosti větru (a nikoliv naopak). Musíme věřit, že rychlost větru je proměnná nezatížená chybou, zatímco měření transpirace chybou zatíženo jistě je.

Pro lineární regresi používáme funkci lm*, která nafituje model, který můžeme dále zkoumat pomocí funkcí* summary, anova, plot *a dalších:*

```
vetry<-data.frame(vitr=c(2,9,5,6,7,3,4,1,0),
                  transpi=c(12,16,14,15,18,11,12,10,8))
lm.vetry<-lm(transpi~vitr,data=vetry)
```

Základní výsledky poskytne funkce summary:

```
summary(lm.vetry)
Call:
lm(formula = transpi ~ vitr, data = vetry)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7226 -0.7903  0.1871  0.2435  2.2548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.8242     0.7668  11.508 8.42e-06 ***
vitr          0.9887     0.1547   6.389 0.000371 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.284 on 7 degrees of freedom
Multiple R-Squared: 0.8536,     Adjusted R-squared: 0.8327
F-statistic: 40.82 on 1 and 7 DF,  p-value: 0.000371
```

*Nejdůležitější je asi tabulka regresních koeficientů (*Coefficients*), ve které je uveden jednak konstantní člen regresní rovnice* (Intercept)*, jednak sklon regresní přímky, označený názvem nezávislé proměnné (*vitr*). Pro oba regresní koeficienty jsou uvedeny i chyby odhadů (*Std. Error*), odpovídající t-test významnosti (*t value *a signifikance* Pr(>|t|)*. Pod tabulkou je také odhad koeficientu determinace R^2, včetně jeho korigované (adjusted) verze. Nakonec jsou uvedeny i výsledky z rozkladu variance pro regresní model, ale ty nám přehledněji zobrazí funkce* anova:

```
anova(lm.vetry)
Analysis of Variance Table

Response: transpi
```

```
          Df Sum Sq Mean Sq F value   Pr(>F)
vitr       1 67.342  67.342  40.825 0.000371 ***
Residuals  7 11.547   1.650
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Celková suma čtverců je rozdělena do modelové sumy čtverců (v řádku vitr*, s hodnotou 67.342) a residuální sumy čtverců (*Residuals*, s hodnotou 11.547). F test, který zde testuje signifikanci celého modelu (v tomto případě, kdy máme jen jednu nezávislou proměnnou, je test ekvivalentní T testu signifikance sklonu přímky, který jsme viděli výše), je založen na F statistice (*F value*), kterou jsme vypočetli vydělením modelového průměrného čtverce residuálním průměrným čtvercem, tj:*

```
67.342/1.650
[1] 40.81333
```

neshoda je dána nepřesností zobrazených MS.i

Pro parametry modelu také můžeme snadno spočítat jejich konfidenční intervaly:

```
confint(lm.vetry,level=0.99)
                0.5 %    99.5 %
(Intercept) 6.1407928 11.507594
vitr        0.4471945  1.530225
```

Pokud si chceme nafitovanou přímku vynést, provedeme to nejjednoduššeji takto:

```
plot(transpi~vitr,data=vetry)
abline(lm.vetry)
```

Na co je třeba se ještě podívat, je reziduální analýza pro nafitovaný regresní model. Doporučuji nejprve vynést residuály modelu proti predikovaným hodnotám, reziduály by měly být rovnoměrně rozloženy kolem osy x:

```
plot(lm.vetry,which=1)
```

Scale-location plot může upozornit na změnu variability residuálů s predikovanou hodnotou, tj. nejčastější případ nehomogenity variance:

```
plot(lm.vetry,which=3)
```

Porovnání residuálů s normální distribucí nám umožňuje normální kvantilový diagram, který vytvoříme nejrychleji takto:

```
plot(lm.vetry,which=2)
```

V případě regresí se musíme rozhodnout, zda při nulové hodnotě nezávislé proměnné bude hodnota závislé proměnné 0 nebo ne. Ve výše uvedeném případě by to byl nesmysl - znamenalo by to, že předem víme, že při nulové rychlosti větru rostliny netranspirují. Ovšem v některých případech regrese procházející počátkem smysl má. Pokud by to pro náš model mělo smysl, odstraníme absolutní člen následujícím zadáním fitovaného modelu:

```
lm.vetry.wrong<-lm(transpi~vitr-1,data=vetry)
summary(lm.vetry.wrong)
...
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
vitr   2.4661     0.3607   6.837 0.000133 ***
---
...
```

Transformace dat

Transformace závislé a/nebo nezávislých proměnných provedeme zadáním transformační funkce přímo ve funkci lm*. Například závislost logaritmu transpirace na logaritmu rychlosti větru bychom fitovali takto:*

```
lm.vetry.loglog <- lm(log(transpi)~log(vitr+1),data=vetry)
```

Jedničku jsme přičítali k hodnotam proměnné vitr *proto, že naše data obsahují i nulovou hodnotu.*

Příklady

1. Byla zjišťována závislost délky trvání vegetační sezony na nadmořské výšce plochy. Byly zjištěny tyto hodnoty:

nadm. výška [m]	délka vegetační sezony [dny]
600	150
650	144
665	145
750	140
850	110
880	105
950	110
1000	99
1005	103
1100	89
1150	92
1200	88

Závisí délka vegetační sezóny na nadmořské výšce? Jak je tato závislost těsná? (Zkontrolujte, zda se reziduály chovají rozumně!).

```
sezona<-data.frame(nv=c(600,650,665,750,850,880,950,1000,1005,1100,
  1150,1200), delka=c(150,144,145,140,110,105,110,99,103,89,92,88))
lm.1<-lm(delka~nv,data=sezona)
summary(lm.1)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 215.293564   9.202105   23.40 4.61e-10 ***
nv           -0.111900   0.009996  -11.19 5.60e-07 ***
...
Multiple R-Squared: 0.9261,     Adjusted R-squared: 0.9187
...
```

Sezóna se zkracuje s každými 100 metry o asi 11.2 dne, těsnost závislosti vyjadřuje R^2=0.926.

2. Byla zjišťována závislost počtu druhů na velikosti plochy: při každé velikosti plochy byla 4 nezávislá stanovení. Byly získány následující výsledky:

velikost plochy (m^2)	počet druhů
0.01	4, 6, 5, 8
0.25	9, 5, 8, 11
1.00	12, 14, 18, 11
4.00	20, 12, 25, 28
9.00	22, 28, 31, 18
16.00	25, 34, 19, 30
64.00	36, 39, 43, 22

Spočtěte regresi, proveďte vhodnou transformaci. Předpokládáme, že platí následující závislost počtu druhů (*S*) na ploše (*A*): $S = c\,A^z$, kde *c* a *z* jsou regresní analýzou odhadnuté koeficienty.

```
spc.area<-data.frame(plocha=rep(c(0.01,0.25,1,4,9,16,64),rep(4,7)),
  pocet=c(4,6,5,8, 9,5,8,11, 12,14,18,11, 20,12,25,28,
          22,28,31,18, 25,34,19,30, 36,39,43,22))
lm.sa<-lm(log(pocet)~log(plocha),data=spc.area)
coef(lm.sa)
(Intercept) log(plocha)
  2.6246644   0.2233922    # z estimate is 0.2233922
exp(coef(lm.sa)[1]) # c estimate
(Intercept)
   13.79994
```

3. Předpokládáme exponenciálně rostoucí populaci. Velikost populace byla zjišťována v jednotlivých časech. Odhadněte růstovou rychlost populace.

čas	velikost populace
0	5
1	7
2	10
3	16
4	19
5	28
6	35
7	49
8	59
9	71
10	101

Můžeme rozumně předpokládat, že variabilita (vyjádřená jako směrodatná odchylka velikosti populace) roste s velikostí populace zhruba lineárně.

$N_t = N_0.e^{rt}$ => $\log(N_t) = \log(N_0) + r*t$

```
rust<-data.frame(time=0:10,size=c(5,7,10,16,19,28,35,49,59,71,101))
lm.rust<-lm(log(size)~time,data=rust)
```

Odhad růstové rychlosti:

```
coef(lm.rust)[2]
     time
0.2951793
```

A 95%ní konfidenční interval pro tuto rychlost:

```
confint(lm.rust,"time")
         2.5 %    97.5 %
time 0.2730446 0.3173140
```

Nafitovaný model odhaduje tuto počáteční velikost (N_0) – srovnej se skutečnou hodnotou 5:

```
exp(coef(lm.rust)[1])
(Intercept)
   5.663679
```

1. Mnohonásobná regrese

Vezměme a rozšiřme příklad z jednorozměrné regrese: byla zjišťována závislost transpirace nejenom na rychlosti větru, ale také na teplotě (TEMP), vlhkosti vzduchu (VLHKOST) a na tom, zda právě svítí slunce (SLUNCE: kvalitativní proměnná, svítí-1, nesvítí-2).

Pozn.: Uvedená data jsou vymyšlená. Ale jsou realistická v tom, že ne všechny proměnné budou vzájemně nekorelované (jeden z teoretických předpokladů mnohonásobné regrese), a dokonce se pravděpodobně vzájemně ovlivňují. Tomu se u reálných dat často nevyhneme. Je vhodné poznamenat, že pro takové množství vysvětlujících proměnných je devět pozorování zoufale málo.

	VITR	TRANSPI	TEMP	VLHKOST	SLUNCE
1	2	12	10	50	1
2	9	16	12	80	0
3	5	14	8	62	0
4	6	15	16	95	1
5	7	18	22	45	0
6	3	11	7	32	0
7	4	12	11	92	1
8	1	10	15	46	1
9	0	8	5	58	0

```
transpi<-data.frame(vitr=c(2,9,5,6,7,3,4,1,0),
  transpi=c(12,16,14,15,18,11,12,10,8),
  temp=c(10,12,8,16,22,7,11,15,5),
  vlhkost=c(50,80,62,95,45,32,92,46,58),
  slunce=as.factor(c("y","n","n","y","n","n","y","y","n")))
```

Nejprve si předvedeme mnohonásobnou regresi pro závislost transpirace na rychlosti větru a teplotě. Model nafitujeme obdobně jako v případě jednoduché přímkové regrese:

```
lm.tr.1<-lm(transpi~vitr+temp,data=transpi)
```

Shrnutí výsledků nám opět poskytne funkce summary*:*

```
summary(lm.tr.1)
Call:
lm(formula = transpi ~ vitr + temp, data = transpi)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9672 -0.6494 -0.2378  0.7981  1.1737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.28104    0.83662   8.703 0.000127 ***
vitr         0.81588    0.13453   6.065 0.000912 ***
temp         0.19135    0.07535   2.539 0.044121 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9631 on 6 degrees of freedom
Multiple R-Squared: 0.9295,     Adjusted R-squared: 0.9059
F-statistic: 39.52 on 2 and 6 DF,  p-value: 0.0003511
```

Tabulka uvedená slovem Coefficients *obsahuje odhady hodnot jednotlivých parametrů - regresních koeficientů (sloupec* Estimate*).*

Na konci výstupu z funkce summary *je i shrnutí analýzy variance regresního modelu a F-testu pro celý statistický model. Podrobnější popis tabulky analýzy variance ale získáme pro fitovaný model funkcí* anova*:*

Do regresní rovnice dosadíme koeficienty, tedy

TRANSPI= 7.28 + 0.82 VITR + 0.19 TEMP

Z t-hodnot (sloupec t value*) a odpovídajících dosažených hladin významnosti (*Pr(>|t|)*) vidíme, že Intercept je významně odlišný od nuly (logické, i při nulové rychlosti větru a nulové teplotě rostliny transpirují, nulová hypotéza byla nesmyslná). Teoreticky bychom mohli říci, že při nulové teplotě je transpirace velmi blízká nule, takže hypotéza zas tak nesmyslná není. Je zde ale další problém – extrapolujeme mimo oblast dat, kde jsme měřili. Z hlediska lineární regrese to nevadí. Nicméně, lineární proložení, které jsme použili pro rozsah teplot, při kterých jsme měřili asi nebude platit mimo tento rozsah. Od nuly jsou odlišné i regresní koeficienty pro obě vysvětlující proměnné (tzn., že obě mají statisticky průkaznou vysvětlující sílu, i když pro teplotu je to na hranici průkaznosti).*

Spočteme všechny regresní diagnostiky, co jsme zkoušeli při jednoduché regresi (s použitím funkce plot*). Také je třeba se podívat na korelace regresních odhadů pro koeficienty jednotlivých nezávislých proměnných – k tomu lze použít parametru* corr=T *při volání funkce* summary*: to nám přidá k výstupu tuto matici:*

```
Correlation of Coefficients:
     (Intercept) vitr
vitr -0.12
temp -0.73       -0.51
```

*Korelací mezi odhadem absolutního členu (*Intercept*) a parametry pro nezávislé proměnné* vitr *a* temp *se můžeme zbavit centrováním obou těchto proměnných (tak aby měly nulový průměr), korelace mezi parametry pro nezávislé proměnné je ale dána vzájemnou korelací mezi těmito proměnnými.*

```
transpi$vitr.cen<-scale(transpi$vitr,scale=F)
transpi$temp.cen<-scale(transpi$temp,scale=F)
summary(lm(transpi~vitr.cen+temp.cen,data=transpi),corr=T)
     ...
Correlation of Coefficients:
         (Intercept) vitr.cen
vitr.cen  0.00
temp.cen  0.00       -0.51
cor(transpi$vitr,transpi$temp)
[1] 0.5059396
```

2. Polynomiální regrese

V případě, že potřebujete spočítat polynomiální regresi, lze to nejprimitivněji (a nejpracněji ☺) udělat tak, že v mnohonásobné regresi zadáme za jednotlivé vysvětlující proměnné x, x^2, x^3 ... atd. (podle požadovaného stupně polynomu). Pozor! Ve vzorci modelu je potřeba zadávat mocniny uzavřené ve funkci I*, protože operátor ^ má ve vzorcích speciální význam (plné interakce mezi proměnnými), tedy např.* ~ x + I(x^2)+I(x^3).

Méně pracně se to dá udělat s použitím funkce poly*. Ta navíc řeší problém, že vyšší mocniny proměnné jsou korelovány s mocninami nižšími (včetně původní proměnné* x*), a to tzv. ortogonalizací jednotlivých členů polynomů. Zadání pak vypadá třeba takto:* ~poly(x, 3).

3. Postupná (stepwise) regrese

*Vybírá soubor nejlepších vysvětlujících proměnných. Pozor na interpretaci, když máme hodně vysvětlujících proměnných, skoro vždycky nám nějaká vyjde průkazná (‘statistical fishing’). V programu R je výběr založen na kritériu úspornosti (*parsimony*), konkrétně AIC statistice, kterou při výběru používá funkce* step*. Nejprve je třeba zadat výchozí (nejčastěji nulový) model. Možné nezávislé proměnné zadáme pomocí parametru* scope*.*

Zkuste v příkladu s transpirací zadat všechny kvantitativní proměnné (VITR, TEMP, VLHKOST) a vybrat z nich ty nejlepší:

```
lm.0<-lm(transpi~+1,data=transpi)

lm.result<-step(lm.0,scope=~vitr+temp+vlhkost)

Start:  AIC= 21.54
 transpi ~ +1
```

```
          Df Sum of Sq    RSS    AIC
+ vitr     1    67.342 11.547  6.243
+ temp     1    39.205 39.684 17.353
<none>                 78.889 21.537
+ vlhkost  1     5.383 73.506 22.901
Step:  AIC= 6.24
 transpi ~ vitr

          Df Sum of Sq    RSS    AIC
+ temp     1     5.981  5.566  1.674
<none>                11.547  6.243
+ vlhkost  1     1.761  9.785  6.753
- vitr     1    67.342 78.889 21.537
Step:  AIC= 1.67
 transpi ~ vitr + temp

          Df Sum of Sq    RSS    AIC
<none>                 5.566  1.674
+ vlhkost  1     0.904  4.662  2.079
- temp     1     5.981 11.547  6.243
- vitr     1    34.118 39.684 17.353
```

Výsledný model je zvýrazněn – další testované kroky (přidání proměnné vlhkost *nebo odebrání proměnné* temp *nebo proměnné* vitr*) již nevedly k snížení hodnoty AIC kritéria, tedy ke zvýšení úspornosti modelu. Výběr sice probíhá na základě AIC statistiky, ale pokud bychom chtěli pro každou zvažovanou změnu modelu vidět hodnotu F statistiky a odpovídající hladinu signifikance, můžeme použít parametr* test="F"*).*
Může se i stát, že se do modelu dostane proměnná, která sama o sobě průkazná není, ale ve spojení s ostatními vysvětluje v modelu signifikantní množství informace.

Analýza kovariance

Analýzu kovariance (ANCOVA) používáme tehdy, když chceme posoudit vliv nějakého faktoru, ale zároveň víme, že na odpověď má vliv i nějaká další SPOJITÁ proměnná.

Opět se vrátíme k příkladu s transpirací. Chceme porovnat, zda se liší transpirace, když svítí a nesvítí slunce. Ale protože víme, že transpirace je závislá i na rychlosti větru, užijeme ji jako kovariátu - covariable. R mezi regresí a analýzou kovariance (a analýzou variance) nerozlišuje tak striktně jako třeba program Statistica. Můžeme tedy opět použít funkce lm *a výsledek shrnout pomocí funkcí* anova *(nebo summary):*

```
lm.ancova<-lm(transpi~vitr+slunce,data=transpi)
anova(lm.ancova)
Analysis of Variance Table

Response: transpi
          Df Sum Sq Mean Sq F value   Pr(>F)
vitr       1 67.342  67.342 36.0944 0.000958 ***
slunce     1  0.352   0.352  0.1889 0.679028
Residuals  6 11.194   1.866
```

Vyčteme zde, že vliv SLUNCE je neprůkazný, průkazný však je vliv větru.

Otázkou však je, zda je závislost transpirace na větru totožná při slunečním svitu nebo bez něj. Tuto hypotézu můžeme testovat následujícím způsobem (zadání odpovídá postupu uvedenému pro program Statistica):

```
lm.ancova.2<-lm(transpi~vitr*slunce,data=transpi)
anova(lm.ancova.2)
Analysis of Variance Table
Response: transpi
            Df Sum Sq Mean Sq F value   Pr(>F)
vitr         1 67.342  67.342 31.2050 0.002536 **
slunce       1  0.352   0.352  0.1633 0.702830
vitr:slunce  1  0.404   0.404  0.1872 0.683263
Residuals    5 10.790   2.158
```

nebo spíše (protože jsme již zjistili, že hlavní efekt faktoru slunce *není průkazný) takto:*

```
lm.ancova.3<-lm(transpi~vitr/slunce,data=transpi)
anova(lm.ancova.3)
Analysis of Variance Table
Response: transpi
            Df Sum Sq Mean Sq F value   Pr(>F)
vitr         1 67.342  67.342 35.0616 0.001034 **
vitr:slunce  1  0.023   0.023  0.0118 0.916992
Residuals    6 11.524   1.921
```

Efekt slunečního svitu je zde "vnořen" do efektu větru, takže v modelu vystupuje jen jako interakce mezi oběma proměnnými.

Oba postupy shodně ukazují, že nemůžeme zamítnou hypotézu, že závislost transpirace na větru je totožná při slunečním svitu i bez něj.

V případě, že máme faktory s náhodným efektem zadáváme je uzavřené do volání funkce Error *a místo funkce* lm *musíme pro fitování modelu použít funkci* aov*.*

Příklady:

1. V kultivačním pokusu byla zjišťována závislost výšky rostliny na (v pokusu řízené) hladině podzemní vody (lze udělat pomocí zvláštního zařízení) a na množství přidaného dusíku. Byly zjištěny následující výsledky.

hladina podz. vody [cm pod povrchem]	přídavek dusíku [nás. zákl. dávky]	výška rostliny [cm]
5	1	15
5	2	17
5	3	20
5	4	21
5	5	26
10	1	13
10	2	16
10	3	17
10	4	19
10	5	21
15	1	10
15	2	12
15	3	15
15	4	15
15	5	18
20	1	10
20	2	12
20	3	13
20	4	14
20	5	17

Vyhodnoťte pokus. Spočtěte mnohonásobnou regresi. Spočtěte matici korelačních koeficientů všech tří proměnných. Nakreslete závislost trojrozměrným grafem.

```
vyska<-data.frame( voda=rep(c(5,10,15,20),rep(5,4)),
        dusik=rep(1:5,4), vyska=c(15,17,20,21,26, 13,16,17,19,21,
                  10,12,15,15,18, 10,12,13,14,17))
lm.vyska<-lm(vyska~voda+dusik,data=vyska)
summary(lm.vyska)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.80000    0.75993   20.79 1.59e-13 ***
voda        -0.46000    0.04195  -10.97 3.95e-09 ***
dusik        2.00000    0.16583   12.06 9.32e-10 ***
```

```
...
Residual standard error: 1.049 on 17 degrees of freedom
Multiple R-Squared: 0.9399,     Adjusted R-squared: 0.9328
F-statistic: 132.8 on 2 and 17 DF,  p-value: 4.196e-11
anova(lm.vyska)
Analysis of Variance Table
Response: vyska
          Df Sum Sq Mean Sq F value    Pr(>F)
voda       1 132.25  132.25  120.23 3.949e-09 ***
dusik      1 160.00  160.00  145.45 9.317e-10 ***
Residuals 17  18.70    1.10
cor(vyska)
             voda     dusik      vyska
voda   1.0000000 0.0000000 -0.6521576
dusik  0.0000000 1.0000000  0.7173229
vyska -0.6521576 0.7173229  1.0000000
summary(vyska)
      voda           dusik       vyska
 Min.   : 5.00   Min.   :1   Min.   :10.00
 1st Qu.: 8.75   1st Qu.:2   1st Qu.:13.00
 Median :12.50   Median :3   Median :15.50
 Mean   :12.50   Mean   :3   Mean   :16.05
 3rd Qu.:16.25   3rd Qu.:4   3rd Qu.:18.25
 Max.   :20.00   Max.   :5   Max.   :26.00
vyska.marg<-list(voda=seq(5,20,by=1),dusik=seq(1,5,by=0.2))
vyska.fit<-predict(lm.vyska,expand.grid(vyska.marg))
res<-persp(vyska.marg$voda,vyska.marg$dusik,matrix(vyska.fit,16),
          xlab="voda",ylab="dusik",zlab="vyska",theta=60)
points(trans3d(vyska$voda,vyska$dusik,vyska$vyska,res))
```

2. Bylo zjišťováno, zda nadměrné pití piva (ANO, NE) má vliv na váhu osoby. Přitom se zjišťovala i výška osoby.

Váha	výška	pije
80	180	0
60	170	0
70	165	1
90	185	0
95	182	1
105	185	1
90	195	0
111	190	1
70	180	0
100	205	0

Má nadměrné pití piva vliv na váhu osoby? Je sklon závislosti váhy na výšce stejný v obou skupinách?

```
pivo<-data.frame(vaha=c(80,60,70,90,95,105,90,111,70,100),
    vyska=c(180,170,165,185,182,185,195,190,180,205),
    pije=factor(c(0,0,1,0,1,1,0,1,0,0)))
anova(lm(vaha~vyska+pije,data=pivo))
Analysis of Variance Table
Response: vaha
          Df  Sum Sq Mean Sq F value    Pr(>F)
vyska      1 1315.54 1315.54  36.219 0.0005325 ***
pije       1  937.11  937.11  25.801 0.0014320 **
Residuals  7  254.25   36.32
```

Porovnání dat s Poissonovou ditribucí

Porovnání dat s jakoukoliv distribucí můžeme provést pomocí Kolmogorov-Smirnovova testu, tak jak jsem si jej ukázali již při porovnávání s normální distribucí. Pro srovnáná s Poissonovou distribucí ale musíme nejprve odhadnou její parametr – tj. střední hodnotu. Testujeme-li tedy například hodnoty uložené v proměnné x*, může zadání testu vypadat takto:*

```
ks.test(x,"ppois",mean(x))
```

Pokud máme data tabelovaná, tj. jako seznam hodnot val *a jejich četnosti* nval*, lze výslednou proměnnou vytvořit pomocí funkce* rep *(jako* rep(val,nval)*).*

Při porovnání s Poissonovou distribucí se ale nejprve podíváme na spočtené hodnoty průměru a variance (v Poissonově distribuci by měly být stejné). Poté můžeme provést test pomocí kriteria:

$$\chi^2 = \frac{s^2(n-1)}{\overline{X}} \qquad (1)$$

a porovnáme s kritickou hodnotou při n-1 stupních volnosti. V R bychom tedy zadali takto:

```
poiss.crit<-var(x)*(length(x)-1)/mean(x)
```

a otestovali takto:

```
1-pchisq(poiss.crit,length(x)-1)
```

Obdobně můžeme spočítat také Lloydův index shlukovitosti. POZOR, Lloydův index je správně

$$L = \frac{\frac{s^2}{\overline{X}} - 1}{\overline{X}} + 1 \qquad (2)$$

(ve skriptech starších versí chybí +1). Takto hodnoty větší než jedna značí shlukovitost, hodnoty menší než 1 pravidelnost. Tak jak je ve skriptech lze užít též, ale pak pozitivní hodnoty značí shlukovitost, negativní pravidelnost. Definici Lloydova indexu tedy odpovídá tato funkce:

```
lloyd.index<-function(x)
  { ((var(x)/mean(x))-1)/mean(x)+1 }
```

Konfidenční interval pro parametr p binomického rozdělení:

Konfidenční interval (implicitně s pokrytím 0.95, ale lze změnit pomocí parametru conf.level*) spočítá funkce* binom.test*, kterou lze užít i pro test shody s teoretickou pravděpodobností. Prvním parametrem funkce je počet případů (*x*), kdy nějaký jev nastal, druhým je celkový počet "pokusů" (*n*), třetím pak teoretická pravděpodobnost, se kterou srovnáváme (*p*). Pomocí parametru* alternative *(s implicitní hodnotou "two.sided") lze provést i jednostranné testy (při užití hodnoty* "less" *nebo* "greater"*).*

Střední chybu odhadu pravděpodobnosti pro binomické rozdělení můžeme spočítat pomocí vzorce:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}$$

kde p *je odhad pravděpodobnosti,* q *je jeho doplněk do jedničky a* n *je celkový počet nezávislých pozorování.*

Příklady:

1. Ve výlovu bylo odebráno 86 kaprů a na každém spočten počet ektoparazitů *Caprozhroutus magnus.* Co můžeme o jejich distribuci říci? (Porovnejte s Poissonovým rozdělením, otestujte pomocí vzorce (1), spočtěte Lloydův index.

Počet kaprů, kteří měli x parazitů

25	0
15	1
11	2
13	3
10	4
6	5
2	6
1	8
1	12
1	15
1	18

```
ncarps<-c(25,15,11,13,10,6,2,1,1,1,1)
npar<-c(0:6,8,12,15,18)
x<-rep(npar,ncarps)
ks.test(x,"ppois",mean(x))
        One-sample Kolmogorov-Smirnov test
data:  x
D = 0.2047, p-value = 0.001482
alternative hypothesis: two.sided

poiss.crit<-var(x)*(length(x)-1)/mean(x)
poiss.crit
[1] 324.9716
1-pchisq(poiss.crit,length(x)-1)
[1] 0
lloyd.index(x)
[1] 2.150686
```

2. Ze 120 náhodně vybraných jablek bylo 56 červivých. Odhadněte procento červivých v populaci, s konf. limitem..

```
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
```

3. Počty sasanek ve čtvercích byly: 0, 2, 5, 10, 2, 0, 5, 6, 1, 6, 2, 4, 5, 6, 1, 2, 3, 2, 3, 2, 0, 0, 0, 2, 3. Otestujte náhodnost rozmístění, různými způsoby.

4. Z 500 lidí, náhodně vybraných, mělo 160 protilátky na boreliozu. Odhadněte (s příslušným konfidenčním intervalem) procento v celé populaci.

5. Byla zjišťována pokryvnost populace metodou point quadrat (tj. jako procento jehel, které zasáhne daný druh). Ze sto jehel zasáhlo 40. Odhadněte pokryvnost a střední chybu odhadu. Kolik jehel bychom potřebovali, aby byla střední chyba odhadu 2%?

```
sqrt(0.4*0.6/99)
[1] 0.0492366
0.4*0.6/(0.02^2)+1
[1] 601
```