Kapitola 1

Porovnání výběrů ze spojitého rozdělení

1.1 Volba rozsahu výběru

$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, nezávislé (náhodný výběr z $\mathsf{N}(\mu, \sigma^2)$), $\sigma > 0$ známo

1.1.1 Požadovaná šířka intervalu spolehlivosti

$$\left(\bar{X} - \frac{\sigma}{\sigma} u(\alpha/2), \bar{X} + \frac{\sigma}{\sqrt{n}} u(\alpha/2)\right) = \left(\bar{X} - c, \bar{X} + c\right)$$

$$n \leq \left(\frac{u(\alpha/2)}{c}\right)^2 \sigma^2$$

1.1.2 Požadovaná síla

$\mathsf{H}_0 : \mu_0 = \mu_0$ proti $\mathsf{H}_1 : \mu \neq \mu_0$, při $\mu = \mu_1$ má být síla aspoň $1 - \beta$:

$$n \geq \left(\frac{u(\alpha/2) + u(\beta)}{\mu_1 - \mu_0}\right)^2 \sigma^2$$

Jak to dopadne při jednostranné alternativě?

1.2 Dva nezávislé výběry

nezávislé výběry ze spojitého rozdělení:
$X_1, \ldots, X_m$ s distribuční funkcí $F_X(x)$
$Y_1, \ldots, Y_n$ s distribuční funkcí $F_Y(x)$
$\mathsf{H}_0 : F_X \equiv F_Y$ proti $\mathsf{H}_1 : F_X \not\equiv F_Y$

1.2.1 Dvouvýběrový t test

(Janko, 1958, str. 33–34), (Anděl, 1993, str. 88)
Nechť $X_i \sim \mathsf{N}(\mu_1, \sigma_1^2)$, $Y_i \sim \mathsf{N}(\mu_2, \sigma_2^2)$

$\sigma_1^2 = \sigma_2^2$ (klasický test)

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{mn}{m+n}} = \frac{\bar{X} - \bar{Y}}{\mathsf{S.E.}(\bar{X} - \bar{Y})}$$

zamítat, je-li $|T| \geq t_{m+n-2}(\alpha)$ ve prospěch alternativy $\mu_1 \neq \mu_2$, kde

$$\begin{aligned} S^2 &= \frac{1}{m+n-2} \left(\sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right) \\ &= \frac{1}{m+n-2} \left((m-1) S_X^2 + (n-1) S_Y^2 \right) \end{aligned}$$

$\sigma_1^2 \neq \sigma_2^2$ (přibližný test)

$$T = \frac{\bar{X} - \bar{Y}}{S_{\bar{X} - \bar{Y}}} = \frac{\bar{X} - \bar{Y}}{\mathsf{S.E.}(\bar{X} - \bar{Y})}$$

kde je $S^2_{\bar{X}-\bar{Y}} = S_X^2/m + S_Y^2/n$. Označíme-li $v_X = S_X^2/m$, $v_Y = S_Y^2/n$, pak je $S^2_{\bar{X}-\bar{Y}} = v_X + v_Y$.

a) zamítat, je-li $|T| \geq \frac{v_X}{v_X+v_Y} t_{m-1}(\alpha) + \frac{v_Y}{v_X+x_Y} t_{n-1}(\alpha)$,
b) (Welch, R, NCSS) zamítat, je-li $|T| \geq t_f(\alpha)$, kde

$$f = \frac{(v_X + v_Y)^2}{\frac{1}{m-1} v_X^2 + \frac{1}{n-1} v_Y^2},$$

c) (Satterthwaite) zamítat, je-li $|T| \geq t_f(\alpha)$, kde

$$f = \frac{(v_X + v_Y)^2}{\frac{1}{m+1} v_X^2 + \frac{1}{n+1} v_Y^2} - 2.$$

```
> attach(vredy)
> x<-ambul; y<-hospital; d<-c(x,y)
> t.test(x,y,var.equal=T)

        Two Sample t-test
```

```
data:  x and y
t = 0.5069, df = 62, p-value = 0.614
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -85.91362 144.28862
sample estimates:
mean of x mean of y
 115.3125   86.1250

> t.test(x,y,var.equal=F)

        Welch Two Sample t-test

data:  x and y
t = 0.5069, df = 59.496, p-value = 0.6141
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -86.01006 144.38506
sample estimates:
mean of x mean of y
 115.3125   86.1250
```

1.2.2 Mannův-Whitneyův test

(Anděl, 1993, str. 91), (Anděl, 1978, 229–230), (Antoch, Vorlíčková, 1992, str. 88–89), (Jurečková, 1981, str. 45–51)

Lze vytvořit celkem mn dvojic (X_i, Y_j). Označme symbolem U_1 počet dvojic, kde je $x_i < y_j$, symbolem U_2 počet dvojic, kde je $x_i > y_j$. V případě shody zvětšíme každou ze statistik U_1, U_2 o jednu polovinu. Kritický obor by měl obsahovat ty výběry, kde jsou obě statistiky U_1, U_2 od sebe příliš různé.

Za nulové hypotézy je zřejmě $\mathsf{E}\, U_1 = \mathsf{E}\, U_2$. Vzhledem k tomu, že zřejmě platí $U_1 + U_2 = mn$, je za nulové hypotézy $\mathsf{E}\, U_1 = mn/2$.

Testová statistika Wilcoxonova dvouvýběrového testu je založena na pořadích, jaká jednotlivá pozorování dostanou, když přestaneme rozlišovat ke kterému ze dvou výběrů patří. Za platnosti nulové hypotézy by průměrná pořadí měla být podobná. Jsou-li T_1 součet pořadí pozorování z prvního výběru a T_2 součet pořadí z druhého výběru, nutně platí $T_1 + T_2 = (m+n)(m+n+1)/2$. Tento vztah platí i v případě, že v případě shodných pozorování použijeme vždy průměrné pořadí.

Navíc platí $U_1 + T_1 = mn + m(m+1)/2$, podobně $U_2 + T_2 = mn + n(n+1)/2$. Rozhodování podle U_1 resp. podle T_1 je tedy ekvivalentní, navíc zřejmě platí $\mathsf{var}\, U_1 = \mathsf{var}\, T_1$.

Rozptyl T_1 je odvozen od součtu čtverců *všech* pořadí od 1 do $m+n$, ať už jsou tato pořadí původní, nebo se mezi nimi vyskytují průměrná pořadí. Platí

$$\mathsf{var}\, T_1 = \frac{mn}{12}\left(m+n+1-\frac{\sum t_j(t_j^2-1)}{(m+n)(m+n+1)},\right)$$

kde $t_1, t_2 \dots$ jsou počty opakování shodných hodnot v původním spojeném souboru, které způsobily potřebu průměrných pořadí.

Tabelovány bývají kritické hodnoty pro $\min(U_1, U_2)$, přičemž se předpokládá, že je $m \le n$. Pro větší m, n se používá normální aproximace

$$Z = \frac{U_1 - \mathsf{E}\, U_1}{\sqrt{\mathsf{var}\, U_1}},$$

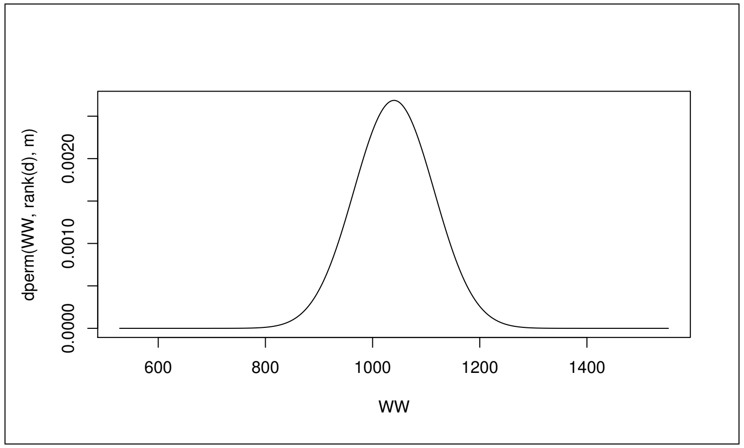
případně s opravou na spojitost, kdy se čitatel přibližuje o jednu polovinu k nule.

Poznámka: Přesnou dosaženou hladinu testu (při oboustranné alternativě) bychom mohli určit tak, že bychom postupně našli všechny m-tice (pořadí prvního výběru) z hodnot $1, \dots, m+n$ (případně včetně shod) a určili, jaký díl z nich dá hodnotu testové statistiky aspoň tolik vzdálenou od její střední hodnoty, jako dal aktuální výběr. Pro velkou hodnotu $\binom{m+n}{m}$ je tento postup výpočetně velmi náročný. Aproximaci normálním rozdělením odpovídá hodnota $2(1-\Phi(|Z|))$, kde $\Phi(x)$ je distribuční funkce normovaného normálního rozdělení.

Test je citlivý vůči alternativě posunutí, je málo citlivý vůči alternativě nestejných rozptylů. Pokud mají oba výběry normální rozdělení se stejnými rozptyly, pak je asymptotická relativní eficience (ARE) Mannova-Whitneyova testu v porovnání s t testem rovna 0,955, obecně není při alternativě posunutí horší než 0,864. (Eficience zde ukazuje, kolikrát méně stačí u t testu pozorování k dosažení stejné síly.)

```
> attach(vredy)
> x<-ambul; y<-hospital
> t.test(x,y,var.equal=T)

         Two Sample t-test

data:  x and y
t = 0.5069, df = 62, p-value = 0.614
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -85.91362 144.28862
sample estimates:
mean of x mean of y
 115.3125   86.1250
```

```
> wilcox.test(x,y)

         Wilcoxon rank sum test with continuity correction

data:  x and y
W = 694, p-value = 0.01388
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in: wilcox.test(x, y)
> wilcox.exact(x,y)

         Exact Wilcoxon rank sum test

data:  x and y
W = 694, point prob = 0, p-value = 0.01310
alternative hypothesis: true mu is not equal to 0

> W.min<-m*(m+1)/2; W.max<-(m+n)*(m+n+1)/2
> WW<-W.min:W.max
> plot(WW,dperm(WW,rank(d),m),type="l")
```



1.2.3 Kolmogorovův-Smirnovův test

(Anděl, 1993, str. 94–97), (Antoch, Vorlíčková, 1992, str. 95–98), (Jurečková, 1981, str. 60–63) (Likeš, Laga, 1978, tab. 30)

Porovnávají se empirické distribuční funkce

$$F_m(x) = \frac{\#(X_i < x)}{m}, \qquad F_n(x) = \frac{\#(Y_i < x)}{n},$$

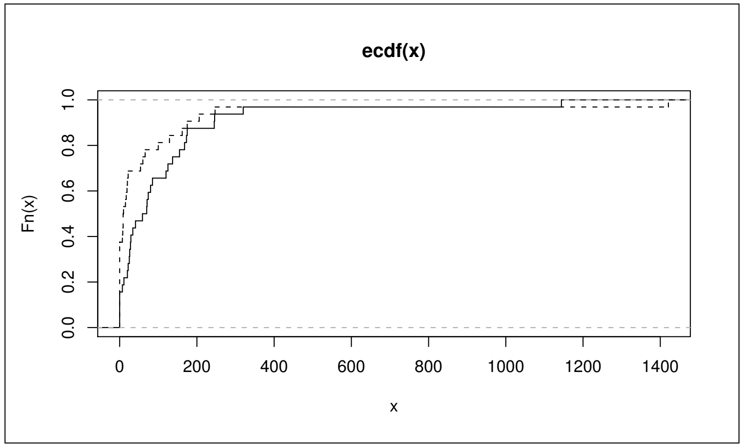
které pro zvolené x ukazují, jaký je díl pozorování (relativní četnost) *menších*, než zvolené x. Míra neshody (vzdálenost) empirických distribučních funkcí je rovna $D_{m,n} = \sup_x |F_m(x) - F_n(x)|$. Je to tedy maximální zjištěný rozdíl mezi zmíněnými relativními četnostmi.

Tabelovány jsou kritické hodnoty $D^*_{m,n}(\alpha)$. Jako aproximace (zejména pro výpočet dosažené hladiny testu) se používá (viz (Anděl, 1993, str. 96)):

$$D^*_{m,n} \doteq \sqrt{\frac{1}{2}\left(\frac{1}{m} + \frac{1}{n}\right) \ln \frac{2}{\alpha}}.$$

Na rozdíl od Mannova-Whitneyova testu je K-S test citlivý na všechna porušení nulové hypotézy.

```
> attach(vredy)
> x<-ambul; y<-hospital
> ks.test(x,y)

         Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.4062, p-value = 0.01017
alternative hypothesis: two.sided

Warning message:
cannot compute correct p-values with ties in: ks.test(x, y)

> plot(ecdf(x),verticals=T,do.points=F,xlim=range(d))
> plot(ecdf(y),verticals=T,do.points=F,add=T,lty=2)
```



1.2.4 Mediánový test

(Antoch, Vorlíčková, 1992, str. 89–90), (Jurečková, 1981, str. 52–53)

Opět spojme oba výběry do jednoho společného soubory a spočítejme jeho výběrový medián. Pokud platí testovaná hypotéza, pak přibližně polovina pozorování z prvního výběru by měla být menších, než tento společný medián. Stejná úvaha platí pro počet pozorování z prvního výběru, která jsou větší, než společný medián.

Označme symbolem U počet prvků prvního výběru, které jsou *větší*, než společný medián. Potom platí

$$\mathsf{E}\,U = \frac{mt}{m+n}, \qquad \mathsf{var}\,U = \frac{mnt(m+n-t)}{(m+n)^2(m+n-1)},$$

kde je $t = [(m+n)/2]$ (celá část).

Když ve výrazu pro rozptyl jmenovatele nepatrně zvětšíme na $(m+n)^3$, dostaneme vyjádření

$$\mathsf{var}\,U \doteq \frac{mnt(m+n-t)}{(m+n)^3}.$$

Spočítejme čtverec normovaného výrazu pro U:

$$\begin{aligned} Z^2 &= \left(\frac{U - \mathsf{E}\,U}{\sqrt{\mathsf{var}\,U}}\right)^2 \\ &= (m+n)\frac{(U(m+n) - mt)^2}{mnt(m+n)t} \\ &= (m+n)\frac{((m-U)(t-u) - U(n+U-t))^2}{mnt(m+n-t)}, \end{aligned}$$

což je známá statistika X^2 ve čtyřpolní tabulce

	výběr		
	první	druhý	
$\leq$ medián	$m-U$	$n+U-t$	$m+n-t$
$>$ medián	U	$t-U$	t
	m	n	$m+n$

V tomto tvaru (kontingenční tabulka) se zpravidla mediánový test používá. Je zřejmé i rozšíření na několik nezávislých výběrů.

```
> attach(vredy)
> x<-ambul; y<-hospital; d<-c(x,y)
> median(d)
[1] 25.5
> table(d<=median(d),(1:64)<33)

        FALSE TRUE
  FALSE    10   22
  TRUE     22   10
> chisq.test(table(d<=median(d),(1:64)<33))

         Pearson's Chi-squared test with Yates' continuity correction

data:  table(d <= median(d), (1:64) < 33)
X-squared = 7.5625, df = 1, p-value = 0.00596

> pperm2(sum(y>median(d)),d>median(d),32)
[1] 0.005534567
```

1.2.5 Permutační test

(Jurečková, 1981, str. 95)

Navažme na úvahu uvedenou v poznámce v oddílu 1.2.2 o Mannovu-Whitneyovu testu. Přesnou dosaženou hladinu jsme počítali tak, že jsme mezi všemi možnými m-ticemi vybranými z napozorovaných $m+n$ hodnot zjistili, jaký jejich díl dá hodnotu testové statistiky aspoň tak špatnou pro nulovou hypotézu, jako dala ona skutečná m-tice hodnot $X_1, \dots, X_m$.

Proč bychom měli pracovat právě s testovou statistikou T_1 (nebo U_1)? K zhodnocení toho, nakolik dané rozdělení $m+n$ pozorování do dvou skupin nesvědčí pro nulovou hypotézu bychom mohli použít některý z výrazů

$$\begin{aligned} K_1 &= |\bar{X} - \bar{Y}|, \\ K_2 &= (\bar{X} - \bar{Y})^2, \\ K_3 &= \sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \end{aligned}$$

nebo dokonce statistiku T klasického dvouvýběrového t testu.

Dá se ukázat, že platí

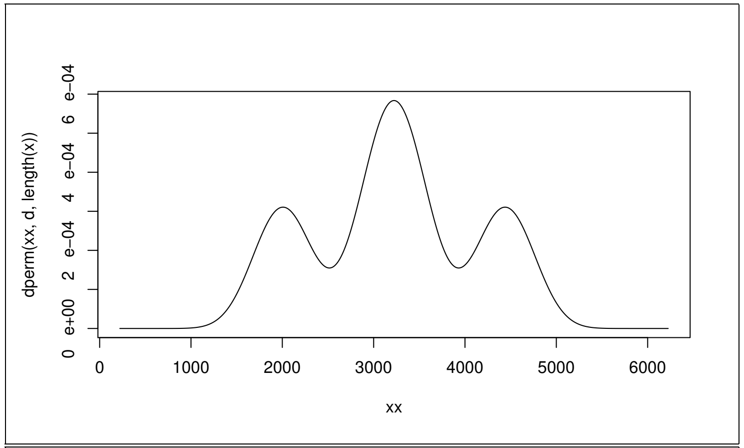
$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{(m+n)\left(\sum X_i^2 + \sum Y_i^2\right) - \left(\sum X_i + \sum Y_i\right)^2 - mn\left(\bar{X} - \bar{Y}\right)^2}{mn(m+n-2)}}},$$

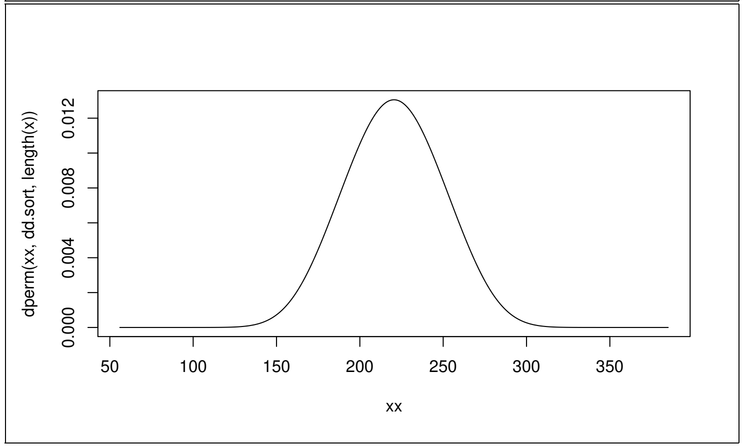
odkud je zřejmé, že absolutní hodnota t statistiky je monotonní funkcí rozdílu $|\bar{X} - \bar{Y}|$. Podobně se zjistí, že všechna uvedená kritéria vedou ke stejnému srovnání možných rozdělení $m+n$ pozorování na soubory o m a n prvcích. Dokonce stačí pracovat se součtem hodnot menšího z obou souborů, např. s výrazem $\sum X_i$.

```
> attach(vredy)
> x<-ambul; y<-hospital; d<-c(x,y)
> pperm2(sum(x),d,length(x))
[1] 0.5771305

> d.sort(d)
> x.min<-sum(d.sort[1:32])
> x.max<-sum(d.sort[33:64])
> xx<-seq(x.min,x.max,by=1)
> plot(xx,dperm(xx,d.sort,length(x)),type="l")

> dd<-round(sqrt(d),0)
> pperm2(sum(dd[seq(along=x)]),dd,length(x))
[1] 0.1212443
> dd.sort<-sort(dd)
```



```
> x.min<-sum(dd.sort[1:32])
> x.max<-sum(dd.sort[33:64])
> xx<-seq(x.min,x.max,by=1)
> plot(xx,dperm(xx,dd.sort,length(x)),type="l")
```





1.2.6 Waldův-Wolfowitzův test (runs test)

(Rao, 1978, str. 545), (Likeš, Laga, 1978, Tab. 36)

Waldův-Wolfowitzův test hodnotí počet iterací Q prvků dvou druhů. Mějme posloupnost prvků dvou druhů, například x, x, y, x, y, y, x. *Iterací* nazveme skupinu prvků jednoho druhu ohraničenou prvky druhého druhu. V našem malém příkladě je celkem 5 iterací. Test se používá k testování náhodnosti uspořádání prvků dvojího druhu. Máme-li m prvků prvního druhu a n prvků druhého druhu a má-li každá m-tice pozic prvků prvního druhu stejnou pravděpodobnost (rovnou kombinačnímu číslu $\binom{m+n}{m}$), pak platí

$$\begin{aligned} \mathsf{E}\, Q &= 1 + \frac{2mn}{m+n}, \\ \mathsf{var}\, Q &= \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}. \end{aligned}$$

Chceme-li testovat stejnou hypotézu, jako u Mannova-Whitneyova testu, pak nejprve spojené výběry srovnáme podle velikosti. Platí-li nulová hypotéza, bude takto vzniklá posloupnost prvků dvou druhů dobře promíchána. Proti platnosti nulové hypotézy svědčí *malý* počet iterací. Použijeme buď tabulky nebo asymptotické rozdělení normovaného výrazu pro Q.

Test je velice jednoduchý, stačí umět porovnat každé dva prvky. Problém nastává, když dochází ke shodám mezi prvky obou druhů. Pak totiž počet iterací není dán jednoznačně.

1.3 Několik nezávislých výběrů

nezávislé výběry ze spojitého rozdělení:
$Y_{11}, \ldots, Y_{1n_1}$ s distribuční funkcí $F_1(x)$
$\ldots$
$Y_{I1}, \ldots, Y_{In_I}$ s distribuční funkcí $F_I(x)$
$\mathsf{H}_0 : F_1 \equiv \ldots \equiv F_I$

1.3.1 Analýza rozptylu jednoduchého třídění

Lze-li předpokládat $Y_{ij} \sim \mathsf{N}(\mu_i, \sigma^2)$, lze použít klasickou analýzu rozptylu jednoduchého třídění. Testová statistika F porovnává variabilitu jednotlivých průměrů $\bar{Y}_{i\bullet} = (1/n_i)\sum_{j=1}^{n_i} Y_{ij}$ s odhadem σ^2 pomocí reziduálního rozptylu

$$\begin{aligned} S^2 &= \frac{1}{n-I}\sum_{i=1}^{I}(n_i-1)S_i^2 \\ &= \frac{1}{n-I}\sum_{i=1}^{I}\sum_{i=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2. \end{aligned}$$

Pokud je $I = 2$, testová statistika analýzy rozptylu těsně souvisí s testovou statistikou t testu. Platí $F = T^2$ a protože kritické hodnoty jsou vázány vztahem $F_{1,n_1+n_2-2}(\alpha) = t^2_{n_1+n_2-2}(\alpha)$, je při oboustranné alternativní hypotéze rozhodování oběma metodami totožné.

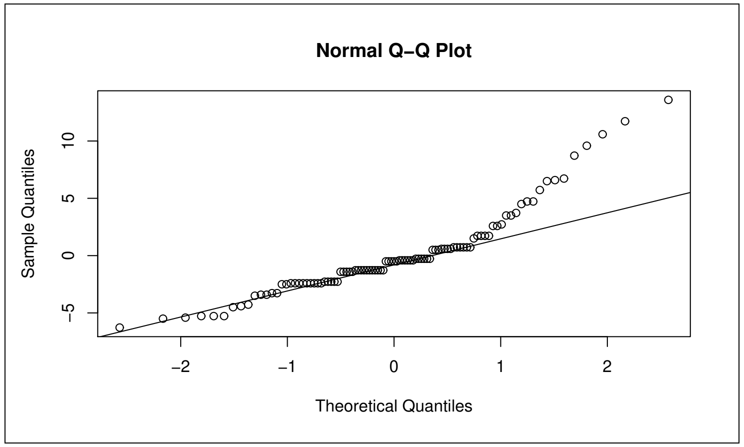
Předpoklad normálního rozdělení se ověřuje až pomocí reziduí $Y_{ij} - \bar{Y}_{i\bullet}$. K ověření předpokladu, že rozptyl je ve všech I výběrech shodný, lze použít vedle Bartlettova testu také přibližný test Levenův. Ten počítá analýzu rozptylu jednoduchého třídění s veličinami $Y^*_{ij} = |Y_{ij} - \bar{Y}_{i\bullet}|$ nebo modifikovaný Levenův test s veličinami $Y^{**}_{ij} = |Y_{ij} - \tilde{Y}_{i\bullet}|$, kde $\tilde{Y}_{i\bullet}$ je medián i-tého výběru. Za tímto testem je vlastnost normálního rozdělení, podle které platí

$$\mathsf{E}\,|Y_{ij} - \bar{Y}_{i\bullet}| = \sqrt{\frac{2}{\pi}}\sqrt{\frac{n_i-1}{n_i}}\sigma_i,$$

když jsme připustili, že v i-tém výběru je rozptyl σ_i^2. Variantu s Y^*_{ij} používá program STATISTICA, variantu s Y^{**}_{ij} program NCSS.

Často se vyskytuje situace, kdy s průměrem roste také směrodatná odchylka S_i spočítaná z i-tého výběru. Bývá to známkou toho, že závisle proměnná má logaritmicko-normální rozdělení. V takovém případě pomůže logaritmická transformace závisle proměnné.

```
> anova(lm(vek.m~vzdelani))
Analysis of Variance Table

Response: vek.m
          Df  Sum Sq Mean Sq F value    Pr(>F)
vzdelani   2  334.77  167.38  12.099 2.061e-05 ***
Residuals 96 1328.14   13.83
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
> qqnorm(resid(lm(vek.m~vzdelani)))
> qqline(resid(lm(vek.m~vzdelani)))
> DAgostino.test(resid(lm(vek.m~vzdelani)))
               skewness    kurtosis      omnibus
statistics 4.734166e+00 3.128867874 3.220214e+01
p-value    2.199579e-06 0.001754812 1.017171e-07
```



1.3.2 Kruskalův-Wallisův test

Když nelze předpokládat normální rozdělení, mohli jsme použít v případě dvou výběrů dvouvýběrový Wilcoxonův test.

Přímým rozšířením uvedeného testu je test Kruskalův-Wallisův. Jsou-li R_{ij} pořadí náhodných veličin Y_{ij} ve sdruženém výběru $n = \sum_{i=1}^{I} n_i$ hodnot, pak součty pořadí v jednotlivých výběrech označíme $T_1, \ldots, T_I$. Za nulové hypotézy očekáváme, že průměrná pořadí v I původních výběrech budou přibližně stejná. Podobně, jako v analýze rozptylu jednoduchého třídění vydělíme variabilitu průměrných pořadí odhadem rozptylu pořadí spočítaným za platnosti hypotézy. To vede ke statistice

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{T_i^2}{n_i} - 3(n+1),$$

která má za platnosti nulové hypotézy asymptoticky rozdělení $\chi^2(I-1)$. (Připomeňte si, že v analýze rozptylu jednoduchého třídění rozdělení za platnosti nulové hypotézy platí $S_A/\sigma^2 \sim \chi^2(I-1)$.)

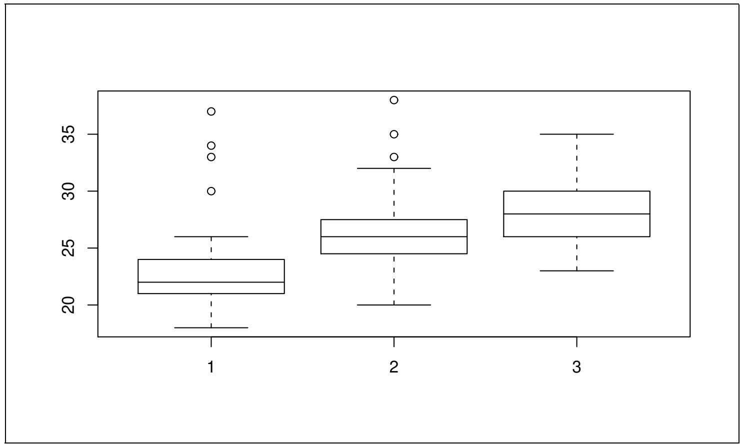
Pokud se v datech vyskytují shody, používají se průměrná pořadí a testovou statistiku lze upravit tím, že se vydělí koeficientem

$$1 - \frac{\sum (t_j^3 - t_j)}{n^3 - n}.$$

Jistě nepřekvapí, že pro $I = 2$ můžeme použít dvouvýběrový Wilcoxonův test (Mannův-Whitneyův) se statistikou Z, přičemž platí $Q = Z^2$.

Podobně jako ve dvouvýběrové úloze je Kruskalův-Wallisův test citlivý zejména na porušení nulové hypotézy posunutím.

```
> kruskal.test(vek.m,vzdelani)

        Kruskal-Wallis rank sum test

data:  vek.m and vzdelani
Kruskal-Wallis chi-squared = 29.4765, df = 2, p-value = 3.974e-07
```



1.3.3 Mediánový test

Opět určíme medián $\tilde{Y}$ všech pozorování a spočítáme kontingenční tabulku, která má v prvním sloupci četnosti $\#(Y_{1j} <= \tilde{Y}), \ldots, \#(Y_{Ij} <= \tilde{Y})$, ve druhém sloupci pak četnosti $\#(Y_{1j} > \tilde{Y}), \ldots, \#(Y_{Ij} > \tilde{Y})$. Testujeme nulovou hypotézu homogenity, že pro všechny výběry je pravděpodobnost pozorování „patřit do prvního sloupce“ stejná. Jak víme, testová statistika X^2 má za platnosti nulové hypotézy asymptoticky rozdělení $\chi^2(I-1)$. Někdy se používá modifikovaný postup, kdy se testová statistika X^2 vynásobí zlomkem $(n-1)/n$.

```
> attach(kojeni)
> table(vzdelani,vek.m>median(vek.m))
```

```
vzdelani FALSE TRUE
       1    28    6
       2    23   24
       3     3   15
> chisq.test(table(vzdelani,vek.m>median(vek.m)))

        Pearson's Chi-squared test

data:  table(vzdelani, vek.m > median(vek.m))
X-squared = 21.617, df = 2, p-value = 2.023e-05
```

1.4 Párové testy

Mějme náhodný výběr $(Y_1, Z_1)', \dots, (Y_n, Z_n)'$ z nějakého dvourozměrného rozdělení. Složky těchto dvourozměrných vektorů mohou být korelované. Pro praxi je zpravidla výhodná co největší pozitivní korelace.

1.4.1 Normální rozdělení rozdílů

Zde stačí předpokládat, aby *rozdíly* $X_i = Y_i - Z_i$ měly normální rozdělení $\mathsf{N}(\mu, \sigma^2)$. Hypotéza $\mu_1 = \mu_2$ je pak ekvivalentní hypotéze $\mu = 0$ zhodnocené pro náhodný výběr $X_1, \dots, X_n$ z rozdělení $\mathsf{N}(\mu, \sigma^2)$.

Testová statistika

$$T = \frac{\bar{X}}{\mathsf{S.E.}(\bar{X})} = \frac{\bar{X}}{S}\sqrt{n},$$

závisí na odhadu S^2 rozptylu σ^2. Jak souvisí parametr σ^2 s parametry výchozích veličin Y_i, Z_i? Je-li ρ korelační koeficient Y_i, Z_i, pak je

$$\sigma^2 = \mathsf{var}\, X_i = \mathsf{var}\,(Y_i - Z_i) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2.$$

Čím je korelační koeficient ρ blíže jedničce, tím je rozptyl σ^2 menší a lze očekávat větší sílu testu (při pevném rozdílu $\mu = \mu_1 - \mu_2$).

1.4.2 Dvourozměrné normální rozdělení

Předpokládejme, že platí

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Test hypotézy $\mu_1 = \mu_2$ jsme uvedli v předchozím odstavci.

Abychom vysvětlili, jak testovat hypotézu $\sigma_1 = \sigma_2$ (proč nemůžeme použít F test?), zaveďme $H_i = Y_i + Z_i$. Snadno zjistíme, že platí

$$\mathsf{cov}(X_i, H_i) = \mathsf{cov}(Y_i - Z_i, Y_i + Z_i) = \sigma_1^2 - \sigma_2^2.$$

Hypotéza o shodě rozptylů je tedy totožná s hypotézou o nezávislosti náhodných veličin X_i, H_i, které mají sdružené normální rozdělení (proč?). Stačí použít známý test nulové hypotézy $\mathsf{H}: \rho_{XH} = 0$ ((Anděl, 1978, str. 117), (Anděl, 1993, str. 217)), který lze upravit na (test Morganův-Pitmanův, viz (Anděl, 1993, str. 85))

$$\begin{aligned} T &= \frac{r_{XH}}{\sqrt{1 - r_{XH}^2}}\sqrt{n-2} \\ &= \frac{S_Y^2 - S_Z^2}{2 S_Y S_Z}\sqrt{\frac{n-2}{1 - r_{YZ}^2}} \end{aligned}$$

s rozdělením $\mathtt{t}(n-2)$ za platnosti testované hypotézy.

```
> t.test(vyska.o,vyska.m,paired=T)

        Paired t-test

data:  vyska.o and vyska.m
t = 15.018, df = 98, p-value = < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.66855 13.91731
sample estimates:
mean of the differences
               12.29293

> X<-vyska.o-vyska.m
> H<-vyska.o+vyska.m
> cor.test(X,H)

        Pearson's product-moment correlation

data:  X and H
t = 1.0555, df = 97, p-value = 0.2938
alternative hypothesis: true correlation is not equal to 0
sample estimates:
      cor
0.1065639
```

1.4.3 Jednovýběrový Wilcoxonův test

(Anděl, 1993, str. 78), (Antoch, Vorlíčková, 1992, str. 101), (Jurečková, 1981, str. 75)

Jde vlastně o test symetrie. Předpokládáme, že náhodná veličina X s hustotou $f(x)$ má rozdělení symetrické kolem parametru a (platí $f(a-x) = f(a+x)$ pro všechna x), takže medián (v případě její existence i střední hodnota) je roven a. Testujeme hypotézu, že je $a = 0$. Označme symbolem R_i^+ pořadí veličiny $|X_i|$. Platí-li nulová hypotéza, pak očekáváme, že součty pořadí

$$S^+ = \sum_{i:X_i>0} R_i^+, \qquad S^- = \sum_{i:X_i<0} R_i^+$$

budou podobné. Protože zřejmě platí $S^+ + S^- = n(n+1)/2$, stačí pracovat jen se statistikou $\min(S^+, S^-)$, pro kterou jsou konstruovány tabulky. Pro velké hodnoty n lze použít toho, že statistika

$$Z = \frac{S^+ - \mathsf{E}\, S^+}{\sqrt{S^+}},$$

kde je

$$\mathsf{E}\, S^+ = \frac{n(n+1)}{4}, \qquad \mathsf{var}\, S^+ = \frac{n(n+1)(2n+1)}{24}.$$

Všude se předpokládá, že žádná z hodnot X_i není rovna nule, jinak se taková hodnota ze zpracování vynechá (a příslušně se sníží n). Test je citlivý vůči alternativě posunutí, ale také vůči na nesymetrickému rozdělení.

```
> wilcox.test(vyska.o,vyska.m,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  vyska.o and vyska.m
V~= 4428, p-value = 2.220e-16
alternative hypothesis: true mu is not equal to 0

> wilcox.test(vyska.o-10,vyska.m,paired=T)

        Wilcoxon signed rank test with continuity correction

data:  vyska.o - 10 and vyska.m
V~= 2678.5, p-value = 0.01110
alternative hypothesis: true mu is not equal to 0
```


1.4.4 Znaménkový test

Je-li $X_1, \dots, X_n$ ze spojitého rozdělení s mediánem $\tilde{x}$, pak pro každí i platí $\mathsf{P}(X_i < \tilde{x}) = \mathsf{P}(X_i > \tilde{x}) = 1/2$. Testujme hypotézu $\tilde{x} = x_0$. Stejně jako v předchozím oddílu nejprve vynecháme pozorování, která jsou totožná s x_0. Nechť Y značí počet pozorování, která jsou menší, než x_0. Za platnosti nulové hypotézy je $Y \sim \mathsf{bi}(n, 0{,}5)$, čehož použijeme při rozhodování. Pro dostatečně velká n použijeme centrální limitní větu, podle které má statistika

$$Z = \frac{Y - n/2}{\sqrt{n/4}} = \frac{2Y - n}{\sqrt{n}}$$

asymptoticky rozdělení $\mathsf{N}(0,1)$.

```
> binom.test(sum(vyska.o-10>vyska.m),length(vyska.m))

        Exact binomial test

data:  sum(vyska.o - 10 > vyska.m) and length(vyska.m)
number of successes = 59, number of trials = 99, p-value = 0.0699
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4926181 0.6934221
sample estimates:
probability of success
             0.5959596
```

1.5 Náhodné bloky

Když chceme vyšetřit vliv ošetření na nějakou vlastnost, používáme při tom často *náhodné bloky*. Jsou to co možná homogenní skupiny objektů, na nichž se budou provádět měření. Máme-li porovnat J ošetření, měla by každá taková skupina mít právě J prvků (případně stejný celočíselný násobek J prvků). Při přípravě pokusu pak jednotlivá ošetření náhodně přidělíme náhodně vybraným prvkům skupiny, při větším počtu pozorování v bloku přiřadíme každému ošetření stejný počet opakování.. Nejznámějším příkladem takovéto experimentální situace je model, který vede k párovému t testu, kdy jsou náhodné bloky dvouprvkové.

Nechť Y_{ij} je hodnota měřeného znaku z i-tého bloku a s j-tým ošetřením, která má spojité rozdělení s distribuční funkcí $F_{ij}(x)$.

1.5.1 Analýza rozptylu

Nejjednodušším modelem je

$$Y_{ij} = \mu + A_i + \beta_j + e_{ij},$$

kde $A_1, \dots, A_I, e_{11}, \dots, e_{IJ}$ jsou nezávislé náhodné veličiny s rozdělením $A_i \sim \mathsf{N}(0, \sigma_A^2)$ a $e_{ij} \sim \mathsf{N}(0, \sigma^2)$. Náhodné veličiny A_i tedy vyjadřují individuální přínos jednotlivých bloků (náhodný efekt). Parametry β_j vyjadřují vliv jednotlivých ošetření (pevný efekt).

Zpravidla nás zajímá nulová hypotéza, že ošetření vliv nemají:

$$\mathsf{H}_B : \beta_1 = \ldots = \beta_J,$$

někdy je zajímavá také nulová hypotéza o tom, že náhodné bloky se mezi sebou neliší (je mezi nimi nulová variabilita) $\mathsf{H}_A : \sigma_A^2 = 0$. Rozhodování o těchto hypotézách je založeno na rozkladu součtu čtverců, na zhodnocení variability mezi náhodnými veličinami Y_{ij}:

$$\begin{aligned} S_T &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\bullet\bullet}\right)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{i\bullet} - \bar{Y}_{\bullet\bullet}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2 \\ &\qquad + \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}\right)^2 \\ &= S_A + S_B + S_e. \end{aligned}$$

O hypotéze H_B (ošetření nemají vliv) rozhodujeme stejně, jako o nulových sloupcových efektech v modelu analýzy rozptylu dvojného třídění bez interakcí s jedním pozorováním v podskupině (bez opakování). Tuto hypotézu zamítáme, je-li

$$F_B = \frac{S_B/(J-1)}{S^2} \geq F_{J-1,(I-1)(J-1)}.$$

kde jsme použili označení pro reziduální rozptyl $S^2 = S_e/((I-1)(J-1))$. Podobně hypotézu H_A (o nulové variabilitě mezi bloky) zamítneme, když vyjde

$$F_A = \frac{S_A/(I-1)}{S^2} \geq F_{I-1,(I-1)(J-1)},$$

Je dobré si uvědomit, že v modelu náhodných bloků obecně již nejsou náhodné veličiny Y_{ij}, Y_{ik} se stejného bloku nezávislé:

$$\mathsf{cov}(Y_{ij}, Y_{ik}) = \mathsf{cov}(A_i + e_{ij}, A_i + e_{ik}) = \mathsf{var}\, A_i = \sigma_A^2.$$

K jejich nezávislosti je třeba, aby platila hypotéza H_A.


```
> attach(maky)
> summary(maky)
   pocet.maku        osetreni      pole
 Min.   : 16.00   Min.   :1   Min.   :1.00
 1st Qu.: 29.75   1st Qu.:2   1st Qu.:1.75
 Median : 82.00   Median :3   Median :2.50
 Mean   :193.50   Mean   :3   Mean   :2.50
 3rd Qu.:377.80   3rd Qu.:4   3rd Qu.:3.25
 Max.   :538.00   Max.   :5   Max.   :4.00
> anova(lm(sqrt(pocet.maku+3/8)~factor(pole)+factor(osetreni)))
Analysis of Variance Table

Response: sqrt(pocet.maku + 3/8)
                 Df Sum Sq Mean Sq F value    Pr(>F)
factor(pole)      3  22.44    7.48  1.8485    0.1922
factor(osetreni)  4 863.93  215.98 53.3634 1.519e-07 ***
Residuals        12  48.57    4.05
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
```

1.5.2 Friedmanův test

Nemají-li náhodné veličiny Y_{ij} normální rozdělení, máme k disposici následující pořadový test. V každém bloku (pro pevné i) zjistíme nejprve pořadí R_{ij} jednotlivých veličin $Y_{i1}, \dots, Y_{iJ}$. Platí-li nulová hypotéza, při které distribuční funkce F_{ij} nezávisí na indexu j (jsou pro všechna ošetření stejné), potom bude pro každé i nabývat vektor $(R_{i1}, \dots, R_{iJ})'$ všech svých možných $J!$ permutací čísel $1, \dots, J$) se stejnou pravděpodobností. Za platnosti nulové hypotézy by měly sloupcové součty pořadí $\sum_{i=1}^{I} R_{ij}$ být přibližně stejné, měly by mít malou variabilitu.

Jako testová statistika se používá výraz

$$\begin{aligned} Q &= \frac{12}{IJ(J+1)} \sum_{j=1}^{J}\left(\sum_{i=1}^{I} R_{ij} - \frac{I(J+1)}{2}\right)^2 \\ &= \frac{12}{IJ(J+1)} \sum_{j=1}^{J}\left(\sum_{i=1}^{I} R_{ij}\right)^2 - 3I(J+1), \end{aligned}$$

která má za platnosti nulové hypotézy (pro každý blok mají všechny permutace pořadí stejnou pravděpodobnost) asymptoticky (s rostoucím počtem bloků) rozdělení $\chi^2(J-1)$.

Motivace testové statistiky Q je prostá. V analýze rozptylu dvojného třídění rozhodujeme o hypotéze H_B na základě toho, že za platnosti této hypotézy platí

$$\frac{1}{\sigma^2}S_B = \frac{1}{\sigma^2}\sum_{j=1}^{j}\sum_{i=1}^{I}\left(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(J-1). \tag{1.1}$$

Místo hodnot Y_{ij} použijeme jejich pořadí R_{ij}:

$$\begin{aligned}\sum_{j=1}^{j}\sum_{i=1}^{I}\left(\bar{R}_{\bullet j} - \bar{R}_{\bullet\bullet}\right)^2 &= I\left(\sum_{j=1}^{J}\left(\frac{1}{I}\sum_{i=1}^{I}R_{ij}\right)^2 - J\left(\frac{J+1}{2}\right)^2\right)\\ &= \frac{1}{I}\sum_{j=1}^{J}\left(\sum_{i=1}^{I}R_{ij}\right)^2 - \frac{IJ(J+1)^2}{4}.\end{aligned}$$

Rozptyl σ^2 charakterizuje variabilitu *uvnitř* jednoho (kteréhokoliv) bloku, je tedy roven

$$\begin{aligned}\frac{1}{J-1}\sum_{j=1}^{J}\left(j - \frac{J+1}{2}\right)^2 &= \frac{1}{J-1}\left(\sum_{j=1}^{J}j^2 - J\left(\frac{J+1}{2}\right)^2\right)\\ &= \frac{1}{J-1}\left(\frac{J(J+1)(2J+1)}{6} - \frac{j(J+1)^2}{4}\right)\\ &= \frac{J(J+1)}{12}.\end{aligned}$$

Místo (1.1) tedy dostaneme právě vztah pro Q.

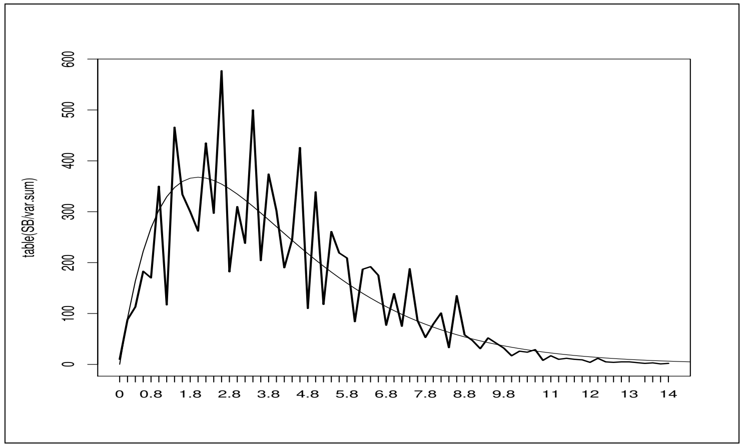
Je-li $J = 2$, pak jde o stejný model, jako u znaménkového testu. Platí pak $Q = Z^2$, kde Z je normovaný výraz znaménkového testu.

Jako příklad použijeme data o vlčích mácích z klasické knihy Snedecor, Cochran (1967), tab. 11.15.1.

```
> friedman.test(pocet.maku,osetreni,pole)

        Friedman rank sum test

data:  pocet.maku, osetreni and pole
Friedman chi-squared = 14.4, df = 4, p-value = 0.006122
```

Ukažme si ještě porovnání skutečného rozdělení statistiky Q s aproximací pomocí $\chi^2(J-1)$. Protože pro každé z $I = 4$ políček máme $J = 5$ ošetření, potřebovali bychom spočítat statistiku celkem pro $J!^{I-1} = 120^4 \cdot 2 \cdot 10^8$ tabulek, což je příliš. (Jeden řádek pořadí můžeme nechat pevný, protože srovnávací statistika na pořadí sloupců nezávisí.) Proto statistiku spočítáme jen $N = 10000$ náhodně vybraných tabulek.


```
> attach(maky)
> I<-max(pole); J<-max(osetreni)
> m<-matrix(pocet.maku,nrow=I,ncol=J,byrow=T)
> m
     [,1] [,2] [,3] [,4] [,5]
[1,]  438  538   77   17   18
[2,]  442  422   61   31   26
[3,]  319  377  157   87   77
[4,]  380  315   52   16   20
> r<-t(apply(m,1,rank))
> r
     [,1] [,2] [,3] [,4] [,5]
[1,]    4    5    3    1    2
[2,]    5    4    3    2    1
[3,]    4    5    3    2    1
[4,]    5    4    3    1    2
> var.sum<-I*J*(J+1)/12
> SB.0<-(J-1)*var(apply(r,2,sum))
> SB.0/var.sum
[1] 14.4
> N<-10000
> SB<-1:N
> for (i in 1:N) SB[i]<-(J-1)*var(r[1,]+r[2,sample(J)]+
+             r[3,sample(5)]+r[4,sample(J)])
> c(mean(SB/var.sum),var(SB/var.sum))
[1] 4.015120 5.979865
> p<-sum(SB>=SB.0)/N
> p
[1] 0
> plot(table(SB/var.sum),type="l")
> xx<-seq(0,15,by=0.2)
> lines(xx,N*dchisq(xx,4)*0.2)
```



Kapitola 2

Spolehlivost měření

V následujícím textu nejde vůbec o obsah znalostí, text se zabývá matematickými a statistickými modely měření znalostí (testováním). Když pomineme obsah otázky a možnosti technicky přípustných odpovědí, můžeme vycházet z představy, že odpovědí je hodnota nějakého znaku v nominálním nebo ordinálním měřítku, případně hodnota spojitého znaku. K prvním dvěma možnostem patří také odpovědi typu ano/ne (nula-jedničkové).

2.1 Model

Následující model je vhodný pro spojité znaky (s hodnotami zapisovanými pomocí libovolného reálného čísla z nějakého intervalu, ovšem pojmy odtud odvozené se používají i v případě dvouhodnotových znaků (nula-jedničkových veličin).

Rozlišujme hodnoty měřené náhodné veličiny A a skutečně změřené hodnoty $Y = A + e$, kde členem e zachycujeme náhodnou nepřesnost měření. Předpoklad nepřítomnosti systematických chyb zapíšeme jako předpoklad $\mathsf{E}\, e = 0$. Důležitý je předpoklad *nezávislosti* obou náhodných veličin na pravé straně, tedy nezávislosti chyby měření na měřené hodnotě. Označme ještě střední hodnotu měřené veličiny jako μ a její rozptyl jako σ_A^2. Variabilitu chyb měření popíšeme rozptylem σ_e^2. Odtud dostáváme, že platí $\mathsf{E}\, Y = \mu$, $\mathsf{var}\, Y = \mathsf{var}\,(A + e) = \sigma_A^2 + \sigma_e^2$.

Spolehlivost měření bývá charakterizována **reliabilitou** (též spolehlivost, angl. reliability)

$$R_A = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}, \tag{2.1}$$

která je totožná s **koeficientem vnitrotřídní korelace** (intraclass correlation) z analýzy rozptylu. Koeficient vnitrotřídní korelace je korelačním koeficientem mezi dvěma nezávislými měřeními téže hodnoty měřené náhodné veličiny A:

$$\mathsf{cov}(Y_1, Y_2) = \mathsf{cov}(A + e_1, A + e_2) = \mathsf{var}\, A = \sigma_A^2.$$

V netriviálním případě, kdy je $\sigma_A > 0$ a $\sigma_e > 0$, platí $R_A \in (0,1)$, přičemž velké hodnoty R_A svědčí o velké spolehlivosti měření (chybová složka málo ovlivňuje výsledek měření).

Ukážeme, že jsme schopni parametr R odhadnout, Určíme tak také přesnost samotných měření, protože korelační koeficient mezi měřenou hodnotou $\mu + A$ a měřením $\mu + A + e$ je roven

$$\rho_{\mu+A,\mu+A+e} = \frac{\sigma_A^2}{\sqrt{\sigma_A^2(\sigma_A^2+\sigma_e^2)}}.$$

2.2 Důsledky nespolehlivosti měření

Nyní se pokusíme kvantifikovat následky, které má skutečnost, že neměříme veličinu A přímo, ale prostřednictvím veličiny Y, která v sobě obsahuje chybovou složku e.

Korelační koeficient

Uvažujme náhodné veličiny A, B s korelačním koeficientem ρ_{AB}, nepřímo měřené pomocí X, Y. Snadno dostaneme

$$\begin{aligned}\rho_{X,Y} &= \frac{\mathsf{cov}(A+e, B+f)}{\sqrt{\mathsf{var}\,(A+e)\mathsf{var}\,(B+f)}}\\ &= \frac{\mathsf{cov}(A,B)}{\sigma_A\sigma_B}\sqrt{\frac{\sigma_A^2}{\sigma_A^2+\sigma_e^2}\frac{\sigma_B^2}{\sigma_B^2+\sigma_f^2}}\\ &= \rho_{AB}\sqrt{R_A R_B}\end{aligned}$$

Protože jsou obě reliability R_A, R_B v intervalu (0,1), platí nutně

$$|\rho_{XY}| < |\rho_{AB}|.$$

Chybové členy e, f tedy zdánlivě zeslabují (zřeďují) sledovanou závislost.

Síla testu

Mějme dva nezávislé výběry veličin X, Y a testujme hypotézu, že obojí střední hodnoty jsou stejné. Pro jednoduchost budeme předpokládat, že oba rozsahy jsou stejné, že také rozptyly σ_A^2, σ_B^2 a reliability R_A, R_B jsou stejné. Kdybychom dokázali měřit veličiny A, B přímo, potom bychom pro $\delta = \mu_A - \mu_B$ dosáhli požadované síly $1-\beta$ dvouvýběrového t testu, pokud by platilo

$$n \geq n^* = 2\sigma_A^2\left(\frac{u(\alpha/2)+u(\beta)}{\delta}\right)^2.$$

Použijeme-li místo veličin A, B s rozptylem σ_A^2 „náhradní“ veličiny X, Y s rozptylem $\sigma_A^2 + \sigma_e^2$ (který je $1/R_A$ krát větší), budeme potřebovat k dosažení stejné síly aspoň n^*/R_A pozorování.

2.3 Odhad reliability

Nejjednodušší odhad reliability dostaneme, když vyjdeme z toho, že je rovna korelačnímu koeficientu dvou nezávislých měření stejného objektu. Předpokládáme-li, že každý objekt byl změřen právě J-krát, můžeme jako odhad reliability použít průměrný korelační koeficient

$$\bar{r} = \frac{2}{J(J-1)}\sum_{j<k}\sum r_{jk}, \tag{2.2}$$

kde r_{jk} je korelační koeficient mezi j-tým a k-tým měřením stejného objektu.

```
> attach(brusle)
> options(digits=3)
> brusle
  body rozhodci zavodnik
1   5.9        1        1
2   5.8        1        2
3   5.5        1        3
4   5.6        1        4
     . . . . . . . . .
51  5.3        9        3
52  5.7        9        4
53  5.5        9        5
54  5.3        9        6
> b<-matrix(body,nrow=nlevels(zavodnik),ncol=nlevels(rozhodci))
> b
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]  5.9  5.9  5.9  5.9  5.8  5.8  5.9  5.9  5.9
[2,]  5.8  5.7  5.8  5.7  5.7  5.7  5.8  5.8  5.8
[3,]  5.5  5.7  5.7  5.8  5.3  5.6  5.5  5.7  5.3
[4,]  5.6  5.8  5.6  5.7  5.4  5.6  5.3  5.6  5.7
[5,]  5.4  5.5  5.4  5.4  5.4  5.4  5.4  5.4  5.5
[6,]  5.5  5.2  5.2  5.3  5.1  5.4  5.2  5.2  5.3
> print(r <- cor(b))
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
 [1,] 1.000 0.645 0.790 0.691 0.856 0.911 0.844 0.790 0.851
 [2,] 0.645 1.000 0.919 0.934 0.772 0.864 0.678 0.919 0.727
 [3,] 0.790 0.919 1.000 0.951 0.859 0.957 0.881 1.000 0.718
```

```
 [4,] 0.691 0.934 0.951 1.000 0.694 0.925 0.726 0.951 0.579
 [5,] 0.856 0.772 0.859 0.694 1.000 0.844 0.929 0.859 0.920
 [6,] 0.911 0.864 0.957 0.925 0.844 1.000 0.859 0.957 0.771
 [7,] 0.844 0.678 0.881 0.726 0.929 0.859 1.000 0.881 0.733
 [8,] 0.790 0.919 1.000 0.951 0.859 0.957 0.881 1.000 0.718
 [9,] 0.851 0.727 0.718 0.579 0.920 0.771 0.733 0.718 1.000

> rr <- NULL
> for (i in 2:9) for (j in 1:(i-1)) rr<-c(rr,r[i,j])
> mean(rr)
[1] 0.83
```

Nezávislá měření objektů

K odhadu spolehlivosti z opakovaných pozorování na stejném objektu můžeme použít vyvážený model analýzy rozptylu jednoduchého třídění s náhodnými efekty

$$Y_{ij} = A_i + e_{ij}, \qquad 1 \le j \le J, \quad 1 \le i \le I.$$

Navíc budeme předpokládat normální rozdělení všech náhodných veličin modelu, tedy $A_i \sim \mathsf{N}(\mu, \sigma_A^2)$, $e_{ij} \sim \mathsf{N}(0, \sigma_e^2)$, přičemž všechny náhodné veličiny jsou nezávislé. Standardní rozklad součtu čtverců je dán vztahy

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2,$$

tedy

$$SS_T = SS_A + SS_e.$$

Odpovídající stupně volnosti jsou pak $f_A = I-1$, $f_e = I(J-1)$. Průměrné čtverce z tabulky analýzy rozptylu označíme jako $MS_A = SS_A/f_A$, $MS_e = SS_e/f_e$.

Zřejmě platí

$$\bar{Y}_{i\bullet} = \mu + A_i + \bar{e}_{i\bullet} \sim \mathsf{N}\left(\mu, \sigma_A^2 + \frac{\sigma_e^2}{J}\right)$$

pro $i = 1, \dots, I$, takže

$$SS_A = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(A_i + \bar{e}_{i\bullet} - \bar{A}_\bullet - \bar{e}_{\bullet\bullet}\right)^2 \sim \left(J\sigma_A^2 + \sigma_e^2\right)\chi^2(I-1).$$

Protože pro pevné i je

$$Y_{ij} - \bar{Y}_{i\bullet} = e_{ij} - e_{i\bullet},$$

platí nutně

$$\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2 \sim \sigma_e^2\chi^2(J-1),$$

takže nakonec po sečtení přes I nezávislých výběrů dostaneme

$$SS_e \sim \sigma_e^2\chi^2(I(J-1)).$$

Použijeme-li známé vlastnosti χ^2 rozdělení, budou střední hodnoty rovny

$$\mathsf{E}\,MS_A = \sigma_e^2 + J\sigma_A^2, \tag{2.3}$$
$$\mathsf{E}\,MS_e = \sigma_e^2. \tag{2.4}$$

Ztotožníme-li uvedené střední hodnoty s průměrnými čtverci, dostaneme odhady

$$s_e^2 = MS_e, \tag{2.5}$$
$$s_A^2 = (MS_A - MS_e)/J. \tag{2.6}$$

Dosadíme-li tyto odhady do definice reliability, dostaneme odhad

$$\hat{R}_A = \frac{s_A^2}{s_A^2 + s_e^2} = \frac{MS_A - MS_e}{MS_A + (J-1)MS_e}. \tag{2.7}$$

Měření stejným expertem

Až dosud jsem předpokládali, že opakovaná měření na i-tém objektu jsou nezávislá. Mnohdy ale každý z J expertů provede vlastní měření téhož i-tého objektu. Potom použijeme jeden z modelů

$$Y_{ij} = \mu + A_i + \beta_j + e_{ij}, \tag{2.8}$$
$$Y_{ij} = \mu + A_i + B_j + e_{ij}, \tag{2.9}$$

podle toho, zda považujeme efekt expertů za pevný (výpověď se týká *daných* expertů) nebo náhodný (danou J-tici expertů považujeme za vzorek možných expertů). Podobně jako dříve předpokládáme $\sum \beta_j = 0$ resp. $B_j \sim \mathsf{N}(0, \sigma_B^2)$ a nezávislost všech náhodných veličin.

Tentokrát se celková variabilita SS_T rozkládá na tři sčítance

$$SS_T = SS_A + SS_B + SS_e,$$

přičemž je tentokrát

$$SS_B = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2,$$
$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2.$$

V obou případech (v modelech (2.8) a (2.9)) dostaneme

$$\mathsf{E}\, MS_A = \mathsf{E}\, \frac{S_A}{I-1} = J\sigma_A^2 + \sigma_e$$
$$\mathsf{E}\, MS_e = \mathsf{E}\, \frac{S_e}{(I-1)(J-1)} = \sigma_e^2,$$

takže (2.3) a (2.4) zůstávají v platnosti. Můžeme tedy použít odhad $\hat{R}_A$ reliability podle (2.7), jen průměrné čtverce MS_e se počítají trochu jinak.

Je však otázka, nakolik lze definovat spolehlivost R vztahem (2.1) také v případě, že (sloupcové) efekty expertů (testových položek) musíme považovat za náhodné. V takovém případě je korelační koeficient mezi dvěma nezávislými opakovanými měření stejné hodnoty roven ($i \neq k$)

$$\rho_{Y_{ij},Y_{ik}} = \frac{\mathsf{cov}(\mu + A_i + B_j + e_{ij}, \mu + A_i + B_k + e_{ik})}{\sqrt{\mathsf{var}\,(\mu + A_i + B_j + e_{ij})\mathsf{var}\,(\mu + A_i + B_k + e_{ik})}}$$
$$= \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_e^2} \leq \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

2.4 Cronbachovo alfa

V psychologických či pedagogických vědách je běžné použít několik měření (testových položek) Y_{ij} a měřený i-tý objekt charakterizovat celkovým součtem $Y_{i\bullet} = \sum_{j=1}^{J} Y_{ij}$. Potom platí

$$\begin{aligned}\mathsf{var}\, Y_{i\bullet} &= \mathsf{var}\left(\sum_{j=1}^{J} Y_{ij}\right)\\ &= \sum_{j=1}^{J} \mathsf{var}\, Y_{ij} + \sum\sum_{j\neq j'} \mathsf{cov}(Y_{ij}, Y_{ij'})\\ &= J(\sigma_A^2 + \sigma_e^2) + J(J-1)\sigma_A^2\\ &= J^2\sigma_A^2 + J\sigma_e^2. \end{aligned} \tag{2.10}$$

Cronbachovo alfa porovnává rozptyl tohoto součtu se součtem rozptylů jednotlivých položek. Rozdíl rozptylů je roven

$$\mathsf{var}\, Y_{i\bullet} - \sum_{j=1}^{J} \mathsf{var}\, Y_{ij} = J(J-1)\sigma_A^2.$$

Když tento rozdíl vztáhneme k rozptylu součtu $Y_{i\bullet}$, dostaneme zlomek

$$\frac{(J-1)\sigma_A^2}{J\sigma_A^2 + \sigma_e^2}.$$

Má-li tato charakteristika ukazovat reliabilitu součtu, potom v případě zcela přesných pozorování, kdy je $\sigma_e^2 = 0$, by měla mít hodnotu rovnou jedničce. Proto se *Cronbachovo* α zavádí vztahem

$$\alpha = \frac{J}{J-1} \frac{\mathsf{var}\, Y_{i\bullet} - \sum_{j=1}^{J} \mathsf{var}\, Y_{ij}}{\mathsf{var}\, Y_{i\bullet}} \tag{2.11}$$
$$= \frac{J}{J-1} \frac{(J-1)\sigma_A^2}{J\sigma_A^2 + \sigma_e^2} = \frac{J\sigma_A^2}{J\sigma_A^2 + \sigma_e^2} \tag{2.12}$$
$$= \frac{\sigma_A^2}{\sigma_A^2 + \frac{1}{J}\sigma_e^2} = \frac{JR_A}{1 + (J-1)R_A}. \tag{2.13}$$

Cronbachova alfa je tedy rostoucí funkcí reliability R_A a také rostoucí funkcí počtu měření J. Uvážíme-li, že rozptyl průměru $\bar{Y}_{i\bullet}$ je roven $\sigma_A^2 + (1/J)\sigma_e^2$, a že měřená veličina $\mu + A_i$ má stále rozptyl σ_A^2, je vidět, že **Cronbachovo α je rovno reliabilitě průměru**. Protože se součet a průměr liší pouze konstantním násobkem a reliabilita nezávisí na měřítku, jde také o reliabilitu součtu $Y_{i\bullet}$. Protože reliabilita je, jak jsme viděli, korelačním koeficientem mezi dvěma nezávislými měřeními téže hodnoty měřené veličiny, lze Cronbachovo α interpretovat jako **korelační koeficient mezi dvěma nezávislými měřeními pomocí součtu $Y_{i\bullet}$ či průměru $\bar{Y}_{i\bullet}$**.

Poznámka Vztah mezi reliabilitou R_A a Cronbachovým α je do jisté míry podobný, jako je vztah mezi směrodatnou odchylkou σ a střední chybou (průměru) $\sigma/\sqrt{n}$, kde n je počet pozorování, z nichž je průměr počítán.

2.4.1 Přímý odhad Cronbachova alfa

První odhad Cronbachova alfa dostaneme z průměrného korelačního koeficientu. Dosadíme-li do definičního vztahu (2.13) průměr korelačních koeficientů, dostaneme *normované Cronbachovo alfa*

$$\bar{\alpha} = \frac{J\bar{r}}{1 + (J-1)\bar{r}}. \tag{2.14}$$

Abychom navrhli další odhad pro α, nechme se inspirovat úlohou o testováním studentů. Protože všichni studenti odpovídají na *stejné* otázky (položky), musíme použít některý z modelů *dvojného třídění* (2.8) nebo (2.9). Zvolíme druhý z těchto modelů, i když výsledné vztahy na této volbě nezávisí. Předpokládáme tedy závislost

$$Y_{ij} = \mu + A_i + \beta_j + e_{ij}, \qquad 1 \leq i \leq J, 1 \leq j \leq I, \tag{2.15}$$

kde symbolem β_j značíme zatím pevný efekt j-té otázky, o němž předpokládáme, že je v průměru přes všechny otázky nulový.

Tabulka 2.1: Explicitní vyjádření součtů čtverců v modelu analýzy rozptylu (2.15) pro Cronbachovo α

variabilita	součet čtverců S	st. volnosti f	stř. hodnota MS
studenti	$SS_A = J\sum_{i=1}^{I}\left(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}\right)^2$	$I-1$	$J\sigma_A^2 + \sigma_e^2$
otázky	$SS_B = I\sum_{j=1}^{J}\left(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2$	$J-1$	$\frac{I}{J-1}\sum_{j=1}^{J}\beta_j^2 + \sigma_e^2$
reziduální	$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet}\right)^2$	$(I-1)(J-1)$	σ_e^2

Běžná analýza rozptylu vede k rozkladu celkového součtu čtverců $SS_T = SS_A + SS_B + SS_e$. Explicitní vyjádření jednotlivých součtů čtverců, stupně volnosti a střední hodnoty průměrných čtverců jsou uvedeny v tabulce 2.1. Když použijeme momentovou metodu a dosadíme do (2.12) členy s odpovídající střední hodnotou, dostaneme odhad Cronbachova α

$$\hat{\alpha} = \frac{MS_A - MS_e}{MS_A} \tag{2.16}$$

$$= 1 - \frac{1}{F_A}. \tag{2.17}$$

Je zajímavé, že tento odhad nijak nezávisí na tom, zda sloupcové efekty (otázky) mají nulový rozptyl či nikoliv. Ve tvaru (2.17) je uvedena souvislost s testem hypotézy o řádkovém efektu v tabulce analýzy rozptylu dvojného třídění. **Cronbachovo alfa je tím větší, čím lépe použitá sada položek rozlišuje testované jedince** (řádkové efekty).

V literatuře se ovšem odhad (2.16) udává v poněkud jiném tvaru. Označme jako s_{jt} výběrovou kovarianci j-tého a t-tého měření. Výrazy pro MS_A a MS_e lze vyjádřit jako

$$MS_A = \frac{1}{J}\sum_{j=1}^{J}\sum_{j=1}^{J} s_{jt},$$

$$MS_e = \frac{1}{J(J-1)}\left(J\sum_{j=1}^{J} s_{jj} - \sum_{j=1}^{J}\sum_{t=1}^{J} s_{jt}\right),$$

takže lze psát

$$\hat{\alpha} = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} s_{jj}}{\sum_{j=1}^{J}\sum_{t=1}^{J} s_{jt}}\right), \tag{2.18}$$

což jsme mohli dostat také tak, že do definice (2.11) dosadíme výběrové rozptyly a kovariance. Je přitom užitečné si uvědomit, že výběrový rozptyl součtů $Y_{i\bullet}$ lze vyjádřit jako $\sum_{j=1}^{J}\sum_{t=1}^{J} s_{jt}$.

Z tvaru vzorečku plyne, že nezáleží na tom, jaký odhad varianční matice použijeme, zda dělíme počtem objektů I nebo dělíme $I-1$, abychom dostali nestranné odhady.

Pokud jsme všechna měření předem normovali tak, že měření každého experta mají stejný (bez újmy na obecnosti jednotkový) rozptyl, pak po nepatrné úpravě dostaneme právě normované Cronbachovo alfa. Je tedy možné, že jsou odhady $\hat{\alpha}$ a $\bar{\alpha}$ velmi blízké. Pokud jsou výrazně odlišné, mělo by to pro uživatele být varováním, že měření jednotlivých expertů mají nestejnou variabilitu. V praxi psychologických nebo pedagogických testů se to může stát zejména tehdy, když maximální počet bodů, které lze získat za jedno měření (položku), není konstantní.

```
> print(s <- var(b))
        [,1]   [,2]  [,3]   [,4]  [,5]   [,6]   [,7]  [,8]   [,9]
 [1,] 0.0377 0.0313 0.040 0.0313 0.043 0.0283 0.0457 0.040 0.0423
 [2,] 0.0313 0.0627 0.060 0.0547 0.050 0.0347 0.0473 0.060 0.0467
 [3,] 0.0400 0.0600 0.068 0.0580 0.058 0.0400 0.0640 0.068 0.0480
 [4,] 0.0313 0.0547 0.058 0.0547 0.042 0.0347 0.0473 0.058 0.0347
 [5,] 0.0430 0.0500 0.058 0.0420 0.067 0.0350 0.0670 0.058 0.0610
 [6,] 0.0283 0.0347 0.040 0.0347 0.035 0.0257 0.0383 0.040 0.0317
 [7,] 0.0457 0.0473 0.064 0.0473 0.067 0.0383 0.0777 0.064 0.0523
 [8,] 0.0400 0.0600 0.068 0.0580 0.058 0.0400 0.0640 0.068 0.0480
 [9,] 0.0423 0.0467 0.048 0.0347 0.061 0.0317 0.0523 0.048 0.0657
> J<-nlevels(rozhodci)
> print(Cronbach.klas <- (J/(J-1))*(1-sum(diag(s))/sum(s)))
[1] 0.974
> print(Cronbach.klas.r <- (J/(J-1))*(1-sum(diag(r))/sum(r)))
[1] 0.978
```

Je tedy $\hat{\alpha} = 0{,}974$ resp. $\bar{\alpha} = 0{,}978$.

Uveďme ještě jednu **poznámku**. Vztah (2.16) umožňuje snadno zjistit, za jakých okolností platí $\hat{\alpha} = 1$. Je to právě tehdy, když je $MS_e = 0$, tedy právě tehdy, když jsou všechny sčítance v reziduálním součtu čtverců nulové. To je možné, právě když existují čísla $a_1, \ldots, a_I, b_1, \ldots, b_J, m$ splňující $\sum_i a_i = \sum_j b_j = 0$ a taková, že pro všechny možné dvojice indexů i, j platí

$$Y_{ij} = m + a_i + b_j.$$

Pak je ovšem výběrový průměr j-tých měření roven

$$\bar{Y}_{\bullet j} = \frac{1}{J}\sum_{i=1}^{I}(m + a_i + b_j) = m + b_j,$$

takže výběrová kovariance j-tého a t-tého měření rovna

$$s_{jt} = \frac{1}{J}\sum_{i=1}^{I}\left(Y_{ij} - \bar{Y}_{\bullet j}\right)\left(Y_{it} - \bar{Y}_{\bullet t}\right) = \frac{1}{J}\sum_{i=1}^{I} a_i^2.$$

Zjišťujeme tedy, že odhad Cronbachova alfa $\hat{\alpha}$ je roven jedné právě tehdy, když všechny kovariance tedy i **všechny rozptyly jsou stejné** a **všechny korelační koeficienty jsou rovny jedné**.

Odhad Cronbachova alfa rozdělením expertů do dvou skupin

Další dva odhady reliability jsou založeny na rozdělení expertů do dvou disjunktních skupin (např. program Statistica).

Rozdělme experty náhodně do dvou skupin. Jejich rozhodování je tím spolehlivější, čím jsou odpovědi těchto dvou skupin podobnější. Označme jako r_{skup} korelační koeficient součtů (průměrů?) hodnocení v těchto dvou skupinách. Použijeme-li stejnou ideu, jako je v definici (2.14) normovaného Cronbachova alfa, dostaneme předpis pro *Spearmanův-Brownův koeficient spolehlivosti* (Spearman-Brown split-half reliability)

$$\hat{\alpha}_{SB} = \frac{2r}{1+r}. \tag{2.19}$$

```
> set.1 <- diag(1:J)%*%all.subsets(J)
> set.2 <- diag(1:J)%*%(1-all.subsets(J))
> r.12 <- NULL
> print(n.subsets.2 <- dim(set.1)[2]-2)
[1] 510
> for (i in 1:n.subsets.2)
+ r.12 <- c(r.12,cor(b%*%set.1[,i+1],b%*%set.2[,i+1]))
> Sp.Brown <- 2*r.12/(1+r.12)
> options(digits=6)
> c(mean(Sp.Brown),range(Sp.Brown))
[1] 0.965011 0.881873 0.999520
```

Hodnoty Spearmanova-Brownova koeficientu spolehlivosti kolísají v mezích od 0,8819 do 0,9995 s průměrnou hodnotou 0,9650.

Podobně když definici (2.11) použijeme na dva částečné součty s výběrovými rozptyly s_1^2 a s_2^2. Označíme-li jako s^2 výběrový rozptyl součtu odpovědí všech expertů, pak můžeme zavést *Guttmanův koeficient spolehlivosti* (Guttman split-half reliability) vztahem

$$\hat{\alpha}_G = \frac{2(s^2 - (s_1^2 + s_2^2))}{s^2}. \tag{2.20}$$

Spearmanův-Brownův odhad je vhodnější tam, kde lze předpokládat, že v obou dílčích součtech jsou rozptyly stejné. Pokud tento předpoklad není udržitelný, je vhodnější odhad Guttmanův.

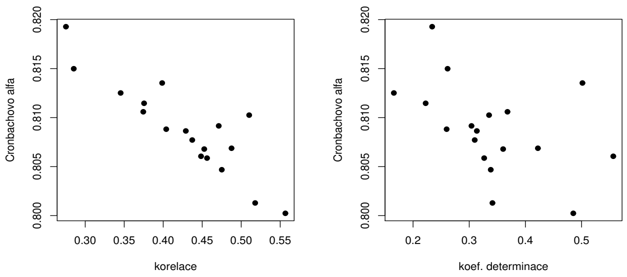
Příklad Ve své diplomové práci studentka PřF UK hodnotila, jak studenti druhých ročníků gymnázií znají jistou partii biologie. V testu bylo mimo jiné 18 otázek s otevřenými odpověďmi (například uvést název zobrazeného objektu) a počítal se počet správných odpovědí. U jednotlivých položek tedy bylo lze získat jeden až šest bodů.

Tabulka 2.2: Položková analýza NCSS dat testu z biologie

	položka			součet položek kromě jedné				koef.
otázka	max	prům.	sm. odch.	prům.	sm. odch.	alfa	korelace	det.
o2	3	1,8961	1,2811	37,290	9,9033	0,8125	0,3454	0,1658
o5	6	3,9826	1,4445	35,203	9,7867	0,8115	0,3755	0,2224
o9A	5	3,1385	1,1336	36,047	9,8494	0,8059	0,4562	0,3266
o9B	5	3,4199	1,4146	35,766	9,6123	0,8013	0,5179	0,3413
o9C	5	2,6103	1,5700	36,575	9,6329	0,8077	0,4372	0,3098
o10A	5	2,9696	1,2489	36,216	9,8482	0,8088	0,4039	0,2598
o10B	4	3,0909	1,1630	36,095	9,8125	0,8047	0,4751	0,3383
o14A	4	1,3463	0,7920	37,839	10,0063	0,8069	0,4876	0,4222
o14B	3	2,4935	0,9549	36,692	10,0202	0,8106	0,3744	0,3681
o14C	3	1,2727	1,1263	37,913	9,7464	0,8002	0,5566	0,4852
o16A	5	2,3722	1,5179	36,813	9,8948	0,8193	0,2753	0,2340
o16B	4	2,0216	1,0650	37,164	10,0615	0,8150	0,2853	0,2615
o17	6	3,5757	1,6103	35,610	9,6226	0,8086	0,4290	0,3135
o19A	4	2,2813	1,2865	36,904	9,7748	0,8061	0,4485	0,5565
o19B	1	0,7229	0,4485	38,463	10,2285	0,8135	0,3987	0,5015
o20A	2	0,5714	0,6479	38,614	10,0944	0,8092	0,4712	0,3040
o20B	1	0,4588	0,4993	38,727	10,1518	0,8103	0,5102	0,3353
o20C	2	0,9610	0,9794	38,225	9,9355	0,8068	0,4526	0,3603

Program NCSS dal odhady pro Cronbachovo alfa: $\hat{\alpha} = 0{,}8176$ a $\bar{\alpha} = 0{,}8396$. Poměrně dobrá shoda obou koeficientů naznačuje, že kdybychom u každé otázky brali jako odezvu procento maximálního možného počtu bodů, dostali bychom jen nepatrně lepší odhad měřené vlastnosti, než dá prostý součet. Velký koeficient determinace u otázky o19A ukazuje, že odpovědi na tuto otázku dokážeme poměrně dobře predikovat z ostatních odpovědí. Nepřekvapí proto, že po vyloučení této otázky bychom dostali téměř stejné Cronbachovo alfa (0,806). Existují však jiné otázky, po jejichž vyloučení klesne tato statistika ještě méně (u položky o19B je to 0,8135), po vynechání položky o16A bychom dostali Cronbachovo alfa dokonce nepatrně větší než na základě všech položek (0,8193).

Když rozdělíme data do dvou skupin (v pořadí sudé a liché otázky), dá Statistica $\hat{\alpha}_{SB} = 0{,}8573$ a $\hat{\alpha}_G = 0{,}8571$, když je korelační koeficient mezi dvěma



Obrázek 2.1: Závislost Cronbachova alfa na korelačním koeficientu resp. koeficientu mnohonásobné korelace spočítaných po vynechání jedné položky

částečnými součty roven $r = 0{,}7503$.

Dalším (zatím) empirickým zjištěním je skutečnost, že s rostoucím korelačním koeficientem r_{iT} mezi i-tou položkou a součtem ostatních položek klesá Cronbachovo alfa spočítané pomocí součtu všech položek s výjimkou i-té. Položka s velkým korelačním koeficientem by tedy v testu velmi chyběla. Svědčí o tom i obrázek 2.1.

Dolní diagram na obrázku 2.1 naznačuje, že souvislost Cronbachova alfa s koeficientem determinace (závislosti dané položky na ostatních) již není tak jednoduchá. Nicméně i tady je patrná podobná souvislost: vynechání málo předpověditelné položky znamená spíše snížení Cronbachova alfa.

Odhad pro nula-jedničkové položky

Velmi často se používají uzavřené testové položky s předem nabídnutými odpověďmi, z nich jediná je správná. Nejde tedy o spojité veličiny (dokonce s normálním rozdělením), pro které byla až dosud uvedená teorie odvozena. Každou odpověď můžeme vyjádřit pomocí jedničky nebo nuly (je nebo není správná). Označíme-li p_j průměr hodnot j-té položky (relativní četnost úspěšných), pak lze odhad Cronbachova alfa z (2.18) vyjádřit speciálně jako

$$\hat{\alpha} = \frac{J}{J-1}\frac{s^2 - \sum_{j=1}^{J} p_j(1-p_j)}{s^2}, \qquad (2.21)$$

když jsme znovu použili označení s^2 pro rozptyl součtu, tentokrát ovšem musíme dělit I, nikoliv $I-1$, protože tak počítáme odhady rozptylu.. Vztah (2.21) je v literatuře uváděn jako *Kuderův-Richardsonův vzorec 20* (Kuder-Richardson-20 formula).

U nula-jedničkových položek nejsou samozřejmě splněny předpoklady modelu analýzy rozptylu. Odpovědi nemají normální rozdělení a zejména rozptyl závisí na střední hodnotě.

Při přípravě takového testu je třeba velmi obezřetně volit nabízené nesprávné odpovědi. Stejně jako u všech testů je třeba před jejich cílovým použitím ověřit skutečné vlastnosti, což nelze jinak, než na podobné populaci, jako je ta cílová.

2.5 Tetrachorický korelační koeficient

Jak jsme viděli, můžeme odhad Cronbachova alfa založit na výpočtu korelačních koeficientů mezi položkami. V případě nula-jedničkových veličin se někdy místo běžného (Pearsonova) korelačního koeficientu používá korelační koeficient *tetrachorický*. Ten je založen na představě, že dvojice nula-jedničkových veličin je jen neúplným pozorováním dvojice veličin s dvourozměrným normálním rozdělením s korelačním koeficientem ρ. Pro každou z veličin jednička znamená, že příslušná realizace překročila jakousi (neznámou) pevnou mez. Tetrachorický korelační koeficient je odhadem parametru ρ. K výpočtu odhadu je třeba uplatnit iterační postup. Jako hrubou aproximaci lze použít výraz

$$r_{tet} = \cos\frac{2\pi}{1+\sqrt{n_{01}n_{10}/n_{00}n_{11}}},$$

kde $n_{00}, n_{01}, n_{10}, n_{11}$ jsou četnosti běžné čtyřpolní tabulky.

Příklad Věnujme se nyní jedenácti položkám z testu z biologie s uzavřenými odpověďmi. Výsledná hodnota Cronbachova alfa ($\hat{\alpha} = 0{,}4029$) je velmi malá, po-

Tabulka 2.3: Položková analýza NCSS nula-jedničkových dat testu z biologie

	položka		součet položek kromě jedné				koef.
položka	prům.	sm. odch.	průměr	sm. odch.	alfa	korelace	det.
t1	0,4372	0,4971	5,2511	1,9240	0,4380	-0,0107	0,0463
t3	0,6797	0,4676	5,0087	1,9019	0,4135	0,0520	0,0309
t4	0,5455	0,4990	5,1429	1,8560	0,3884	0,1268	0,0508
t6	0,3463	0,4768	5,3420	1,8509	0,3774	0,1559	0,0728
t7	0,7576	0,4295	4,9307	1,8938	0,3969	0,0969	0,0525
t8	0,2987	0,4587	5,3896	1,8497	0,3709	0,1748	0,0442
t14	0,7229	0,4485	4,9654	1,8079	0,3331	0,2830	0,1473
t15	0,4805	0,5007	5,2078	1,8697	0,3995	0,0972	0,0552
t16	0,3203	0,4676	5,3680	1,8318	0,3591	0,2069	0,1328
t20	0,6364	0,4821	5,0519	1,8074	0,3434	0,2463	0,1125
t24	0,4632	0,4997	5,2251	1,8061	0,3482	0,2308	0,1290

dobná je i normované alfa ($\bar{\alpha} = 0{,}4060$). Kdybychom nepoužili testovou položku t1, dostali bychom poněkud spolehlivější test ($\hat{\alpha} = 0{,}4380$). Tuto skutečnost dokumentuje také záporný korelační koeficient mezi položkou t1 a součtem ostatních položek. Nejsnazší položkou je t6, nejobtížnější položka t7. Spolehlivost by nejvíce oslabilo vyloučení položky t14.

Kdybychom použili místo Pearsonova korelačního koeficientu koeficient tetrachorický, dostali bychom větší hodnotu normovaného alfa $\bar{\alpha}_{tet} = 0{,}5432$. Právě z toho důvodu, že tetrachorický korelační koeficient vychází zpravidla větší, než Pearsonův a protože je jeho interpretace omezena na předpokládané mnohorozměrné rozdělení, někteří statistici jeho použití nedoporučují.

2.6 Proč právě součet?

Matematik si položí otázku: proč výsledný odhad měřeného znaku založit právě na součtu výsledků jednotlivých testů (měření na stejném objektu), když například otázky nejsou stejně obtížné nebo z objektivních důvodů některé testové položky umožňují získat více bodů než jiné. Například v testu z biologie se vyskytují položky s nestejným maximálním počtem bodů. Nebylo by vhodnější počítat nějaký vážený průměr nebo nějakou lineární funkci? Nebylo by vhodnější jednotlivé položky nějak vážit?

Normované Cronbachovo alfa je vlastně běžné Cronbachova alfa, v němž jsme každou položku nejprve vyjádřili jako z skór (odečti průměr, vyděl směrodatnou odchylkou). Protože na posunutí (odečítání průměru) nezávisí, znamená to prakticky, že pracujeme místo se součtem položek s jejich lineární funkcí, přičemž u j-té položky je koeficient $1/\sqrt{s_{jj}}$, kde s_{jj} je výběrový rozptyl j-té položky.

Uvažujme nyní lineární funkci $Y = \mathbf{c}'\mathbf{Y}$ náhodného vektoru $\mathbf{Y}$ s varianční maticí $\mathbf{V}$. Hledáme takový vektor konstant $\mathbf{c}$, aby Cronbachovo alfa pro Y bylo co možná největší. Znamená to maximalizovat funkci

$$\alpha(\mathbf{c}) = \frac{J}{J-1}\frac{\mathbf{c}'\mathbf{V}\mathbf{c} - \mathbf{c}'\mathrm{diag}(\mathbf{V})\mathbf{c}}{\mathbf{c}'\mathbf{V}\mathbf{c}},$$

kde $\mathrm{diag}(\mathbf{V})$ je diagonální matice obsahující právě diagonálu matice $\mathbf{V}$.

Platí zobecnění známé extremální vlastnosti vlastních čísel a vektorů symetrické matice (např. (Rao, 1978, str. 101)): Jsou-li $\mathbf{A}, \mathbf{B}$ symetrické matice, přičemž $\mathbf{B}$ je pozitivně definitní, potom je

$$\max_{\mathbf{c}\neq\mathbf{0}} \frac{\mathbf{c}'\mathbf{A}\mathbf{c}}{\mathbf{c}'\mathbf{B}\mathbf{c}} = \lambda_1,$$

kde λ_1 je největší z kořenů rovnice $\det(\mathbf{A}-\lambda\mathbf{B}) = 0$. Maximum nastává pro nenulové řešení soustavy rovnic $(\mathbf{A} - \lambda_1\mathbf{B})\,\mathbf{c} = \mathbf{0}$.

Na místě matice $\mathbf{B}$ máme varianční matici (při odhadu Cronbachova alfa výběrovou varianční matici), na místě matice $\mathbf{A}$ máme varianční matici, v níž jsme předem všechny diagonální prvky nahradili nulami.

Příklad Navažme na příklad o testu z biologie. Použijeme stejných 18 položek s otevřenými odpověďmi. Např. pomocí programu MATLAB dostaneme $\lambda_1 = 0{,}7982$ (pro zajímavost, nejmenší kořen je $\lambda_{18} = -2{,}7417$). Optimální vektor $\mathbf{c}$ je uveden v tabulce 2.4. Sloupec nadepsaný optimální norm. obsahuje takový násobek optimálního vektoru vah, který má součet složek rovný počtu složek. To umožňuje snadno zjistit, které položky mají při optimálních vahách větší a které menší koeficient, než je standardní jednička.

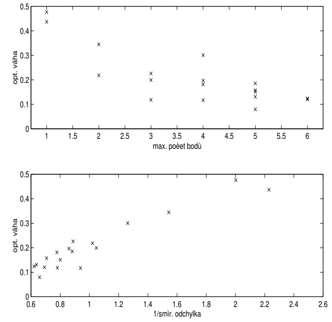
V předposledním sloupci jsou uvedeny podobně normované optimální váhy, když jsme zvlášť hledali nejlepší lineární kombinaci ve dvou skupinách po devíti položkách, které jsme náhodně vybrali. Jak je patrno, optimální váha každé položky závisí na tom, se kterými položkami má současně odhadovat Cronbachovo alfa.

Na obrázku 2.2 jsou tyto koeficienty porovnány s maximálními počty bodů položek a převrácenými hodnotami jejich směrodatných odchylek.

Tabulka 2.4: Optimální váhy testu z biologie

název	max.	směr.	optimální		půlení	
položky	bodů	odch.	váha	norm.	norm.	skup.
o2	3	0,7806	0,1184	0,5662	0,6177	B
o5	6	0,6923	0,1205	0,5762	0,5096	B
o9A	5	0,8821	0,1855	0,8870	0,8418	B
o9B	5	0,7069	0,1578	0,7543	0,6153	A
o9C	5	0,6369	0,1309	0,6257	0,6144	A
o10A	5	0,8007	0,1512	0,7231	0,4478	B
o10B	4	0,8598	0,1968	0,9408	0,8764	A
o14A	4	1,2625	0,3005	1,4368	0,7281	B
o14B	3	1,0472	0,1991	0,9520	0,6318	B
o14C	3	0,8878	0,2261	1,0812	0,8471	B
o16A	5	0,6588	0,0798	0,3813	0,7167	A
o16B	4	0,9389	0,1169	0,5591	1,4287	A
o17	6	0,6210	0,1239	0,5923	1,0318	A
o19A	4	0,7773	0,1812	0,8664	1,1794	A
o19B	1	2,2296	0,4369	2,0891	1,5486	A
o20A	2	1,5433	0,3445	1,6472	2,2616	B
o20B	1	2,0025	0,4762	2,2770	0,9887	A
o20C	2	1,0210	0,2184	1,0444	2,0944	B

Z dolního diagramu obrázku 2.2 je zřetelná úměrnost optimálních vah převráceným hodnotám směrodatných odchylek a převráceným hodnotám (odmocnin) maximálních počtů bodů za jednotlivé položky. Běžné Cronbachovo alfa je, jak víme, rovno $\hat{\alpha} = 0{,}8176$, normované alfa $\bar{\alpha} = 0{,}8396$. Použijeme-li optimální váhy, dostaneme $\hat{\alpha}_{opt} = 0{,}8451$, tedy hodnotu nepříliš jinou, než bylo $\hat{\alpha}$, přestože v posledním sloupci tabulky 2.4 jsou hodnoty velmi vzdálené od jedniček.



Obrázek 2.2: Optimální váhy v závislosti na maximálním počtu bodů a směrodatné odchylce položky

Při zmíněném rozdělení položek do dvou skupin jsme dostali v jednotlivých skupinách odhady Cronbachova alfa 0,6790 a 0,7192. Protože korelační koeficient součtů ve dvou skupinách je roven 0,6703, vychází Spearmanův-Brownův odhad podle (2.19) $\hat{\alpha}_{SB} = 0{,}8026$, kdežto odhad Guttmanův $\hat{\alpha}_G = 0{,}8009$.

Optimální váhy dají po řadě odhady $\hat{\alpha} = 0{,}7274$ a $\hat{\alpha} = 0{,}7418$. Protože korelační koeficient optimálních kombinací položek je roven 0,7051, vychází $\hat{\alpha}_{SBopt} = 0{,}8271$. Podobně z rozptylů optimálních lineárních kombinací položek dostaneme Guttmanův odhad $\hat{\alpha}_{Gopt} = 0{,}8152$.

Kapitola 3

Závislost nominálních a ordinálních znaků

Rozhodováni o nezávislosti nominálních znaků je jednoduché: nejčastěji se používají kontingenční tabulky. Nechť první znak nabývá r různých hodnot a druhý znak c různých hodnot. Symbolem N_{ij} označíme počet případů, kdy první znak nabyl své i-té hodnoty a současně druhý znak své j-té hodnoty. Marginální četnosti

$$N_{i\bullet} = \sum_{j=1}^{c} N_{ij}, \quad N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$$

jsou pak náhodné, na rozdíl od celkového počtu pozorování

$$n = \sum_{i=1}^{r} N_{i\bullet} = \sum_{j=1}^{c} N_{\bullet j}.$$

Testová statistika má tvar

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - N_{i\bullet}N_{\bullet j}/n)^2}{N_{i\bullet}N_{\bullet j}/n} \tag{3.1}$$

$$= n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}^2}{N_{i\bullet}N_{\bullet j}} - n \tag{3.2}$$

a kritický obor asymptotického testu je dán nerovností $X^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$.

3.1 Nominální znaky

Pomocí statistiky X^2 můžeme sice rozhodovat o nezávislosti vyšetřovaných znaků v nominálním měřítku, sotva ji však můžeme použít jako (empirickou) míru síly závislosti (velmi často též míru asociace), která by měla ležet v intervalu $\langle 0, 1\rangle$. Pro nezávislé veličiny by měla být míra asociace nulová (nejlépe jenom pro nezávislé), podobně jednička by se měla objevit (právě) v případě funkční závislosti.

Aspoň pro čtyřpolní tabulku v tomto intervalu však již leží koeficient

$$\Phi = \sqrt{\frac{X^2}{n}}. \tag{3.3}$$

Že můžeme vypadnout z požadovaného intervalu nahlédneme snadno, zvolíme-li za matici $\mathbf{N}$ jednotkovou matici řádu 3, která dá $\Phi = \sqrt{2}$. Stačí však ještě málo: *Cramerovo* V, definované vztahem

$$V = \sqrt{\frac{X^2}{n(m-1)}} = \sqrt{\frac{\Phi^2}{m-1}}, \tag{3.4}$$

kde je $m = \min(r, c)$, již v intervalu $\langle 0, 1\rangle$ nutně leží. Pokud v každém řádku (nebo v každém sloupci) je jediná nenulová četnost, takže jde o jednoznačnou závislost sloupcového (řádkového) indexu na indexu řádkovém (sloupcovém), což lze snadno dokázat pomocí vztahu (3.2).

V požadovaném intervalu nepochybně leží také *(Pearsonův) koeficient kontingence*

$$C = \sqrt{\frac{X^2}{n + X^2}}. \tag{3.5}$$

Bohužel, horní meze tohoto intervalu dosahuje pouze v limitním případě. Je-li tabulka například čtvercová a ukazuje na vzájemně (proč vzájemně?) jednoznačnou závislost, pak je nutně $X^2 = n(r-1)$. Odtud hned vychází, že nejvyšší možná hodnota $C = \sqrt{(r-1)/r}$ závisí na počtu rozlišovaných kategorií.

Obdobou Cramerova V je *Čuprovovo* T

$$T = \sqrt{\frac{X^2}{n\sqrt{(r-1)(c-1)}}}, \tag{3.6}$$

které je pro čtvercové matice shodné s V.

Nesymetrické míry asociace

Uvažujme nyní, nakolik zlepší známá hodnota řádkového znaku předpověď hodnoty znaku sloupcového. Když hodnotu řádkového znaku neznáme, budeme předpovídat, že nastane ta hodnota sloupcového znaku, která je nejpravděpodobnější. Odhadem pravděpodobnosti chybného rozhodnutí bude tedy

$$q_0 = 1 - \frac{\max_j N_{\bullet j}}{n}.$$

Pokud však řádkový index známe, budeme hledat největší pravděpodobnost jen v tomto řádku. Pravděpodobnost chybného rozhodnutí dostaneme vážením pomocí pravděpodobností hodnot jednotlivých řádků, takže odhad pravděpodobnosti chybného rozhodnutí bude

$$q_1 = \sum_{i=1}^{r} \frac{N_{i\bullet}}{n}\left(1 - \frac{\max_j N_{ij}}{N_{i\bullet}}\right).$$

Nyní použijeme stejnou myšlenku, jaká je aplikována například v pojmu koeficientu determinace. Zjistíme, nakolik se zmenšila neurčitost predikce tím, že známe řádkový index. Neurčitost budeme měřit pravděpodobností chybného rozhodnutí a vyjádříme ji (stejně, jako v případě koeficientu determinace) relativně vzhledem k výchozí hodnotě:

$$\begin{aligned} \lambda_c &= \frac{q_0 - q_1}{q_0} \\ &= \frac{\sum_{i=1}^r \max_j N_{ij} - \max_j N_{\bullet j}}{n - \max_j N_{\bullet j}}. \end{aligned} \tag{3.7}$$

Koeficient λ_c ukazuje, nakolik závisí hodnoty sloupcového indexu na indexu řádkovém. Po záměně indexů dostaneme podobnou nesymetrickou míru, která vyjadřuje sílu závislosti řádkového indexu na sloupcovém:

$$\lambda_r = \frac{\sum_{j=1}^c \max_i N_{ij} - \max_i N_{i\bullet}}{n - \max_i N_{i\bullet}}.$$

Symetrickou míru dostaneme po sečtení čitatelů a jmenovatelů definičních vztahů pro λ_r a λ_c:

$$\lambda = \frac{\sum_{i=1}^r \max_j N_{ij} + \sum_{j=1}^c \max_i N_{ij} - \max_j N_{\bullet j} - \max_i N_{i\bullet}}{2n - \max_j N_{\bullet j} - \max_i N_{i\bullet}}.$$

Postup, který jsme použili při zavedení λ_c, spočíval v porovnání střední hodnoty (přes podmínku) míry neurčitosti podmíněného rozdělení (sloupcové) veličiny Y při daném x s mírou neurčitosti marginálního rozdělení samotného Y. Když použijeme jinou míru nejistoty, dostaneme jinou nesymetrickou míru závislosti.

Použijme *entropii*, která je pro multinomické rozdělení s pravděpodobnostmi $\pi_1, \ldots, \pi_k$ definována jako

$$H = -\sum_{i=1}^{k} \pi_i \log \pi_i.$$

Neznámé pravděpodobnosti sdruženého i marginálních rozdělení nahradíme odpovídajícími relativními četnostmi (maximálně věrohodnými odhady)

$$\hat\pi_{ij} = \frac{N_{ij}}{n}, \quad \hat\pi_{i\bullet} = \frac{N_{i\bullet}}{n}, \quad \hat\pi_{\bullet j} = \frac{N_{\bullet j}}{n}.$$

Entropie marginálního rozdělení sloupcové veličiny je odhadnuta výrazem

$$H(Y) = -\sum_{j=1}^{c} \hat{\pi}_{\bullet j} \log \hat{\pi}_{\bullet j}.$$

Podobně odhad střední hodnoty entropie podmíněného rozdělení je

$$\begin{aligned} \hat{\mathsf{E}}_X H(Y|X=x) &= \sum_{i=1}^{r} \hat{\pi}_{i\bullet} \left(-\sum_{j=1}^{c} \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}} \right) \\ &= -\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}} \\ &= -\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{N_{ij}}{n} \log \frac{N_{ij}}{N_{i\bullet}}. \end{aligned}$$

Porovnáním dostaneme *koeficienty nejistoty* pro závislost sloupcového indexu na řádkovém

$$U_c = -\frac{\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij} \log \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j}}}{\sum_{j=1}^{c} \hat{\pi}_{\bullet j} \log \hat{\pi}_{\bullet j}}. \tag{3.8}$$

Záměnou jmenovatele bychom dostali podobný koeficient neurčitosti U_r vyjadřující sílu závislosti řádkového indexu na sloupcovém indexu. Opět sečtením čitatelů a jmenovatelů definičních vztahů dostaneme také symetrický koeficient nejistoty.

Entropie patří v biologii mezi míry diverzity (bohatosti druhů). Jinou používanou mírou diverzity je *Simpsonův-Giniho* index

$$1 - \sum_{i=1}^{k} \pi_i = \sum_{i=1}^{k} \pi_i (1 - \pi_i),$$

který lze interpretovat jako pravděpodobnost mezidruhového setkání. Když jej použijeme místo entropie, dostaneme *Goodmanův-Kruskalův koeficient kontingence*

$$\tau_c = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{\pi}_{ij}^2 / \hat{\pi}_{i\bullet} - \sum_{j=1}^{c} \hat{\pi}_{\bullet j}}{1 - \sum_{j=1}^{c} \hat{\pi}_{\bullet j}}. \tag{3.9}$$

```
> attach(kojeni)
> print(N <- table(vzdelani,plan))
        plan
vzdelani  0  1
       1 20 14
       2 16 31
       3  5 13
> N.radky<-apply(N,1,sum)
> N.sloupce<-apply(N,2,sum)
> chisq.test(N)
```



```
         Pearson's Chi-squared test

data:  N
X-squared = 6.6794, df = 2, p-value = 0.03545

> fisher.test(N)

         Fisher's Exact Test for Count Data

data:  N
p-value = 0.03804
alternative hypothesis: two.sided
> n <- sum(N)
> print(H.Y <- -sum((N.sloupce/n)*log(N.sloupce/n)))
[1] 0.6783305
> print(H.Yx <- -sum((N/n)*log(N/N.radky)))
[1] 0.6445645
> print(U <- (H.Y-H.Yx)/H.Y)
[1] 0.04977806
```

3.2 Ordinální znaky

Podstatnou odlišností ordinálního měřítka je uspořádání jeho hodnot. Pro každá dvě pozorování x_i, x_j takového znaku můžeme rozhodnout, která z možností $x_i \prec x_j$, $x_i \approx x_j$ a $x_i \succ x_j$ nastala. Sledujme nyní dva znaky. Dvojici pozorování (x_i, y_i) a (x_j, y_j) nazveme *konkordantní*, když v obou znacích je lepší stejný prvek

$$\begin{aligned} x_i \prec x_j \quad &\text{a současně} \quad y_i \prec y_j \\ x_i \succ x_j \quad &\text{a současně} \quad y_i \succ y_j. \end{aligned}$$

V případě, že v každém znaku je lepší jiný prvek, tj. platí některá z možností

$$\begin{aligned} x_i \prec x_j \quad &\text{a současně} \quad y_i \succ y_j \\ x_i \succ x_j \quad &\text{a současně} \quad y_i \prec y_j, \end{aligned}$$

jde o dvojici *diskordantní*. Mohou existovat také dvojice pozorování, které nejsou ani konkordantní ani diskordantní.

Počet konkordantních dvojic mezi pozorováními $(x_1, y_1), \ldots, (x_n, y_n)$ se zpravidla značí symbolem P, počet diskordantních symbolem Q. Při čtení vzorců je třeba se ujistit, zda autor má na mysli uspořádané dvojice (náš případ) nebo dvojice neuspořádané. V prvním případě jsou čísla P, Q nutně sudá a platí $P + Q \leq n(n-1)$.

Jednoduchá míra asociace mezi ordinálními veličinami je dána vztahem

$$\gamma = \frac{P-Q}{P+Q}. \tag{3.10}$$

Porovnává odhady pravděpodobností, že náhodně zvolená dvojice je konkordantní a diskordantní.

Poměrně málo se používá *Kendallovo* τ, které rozdíl $P-Q$ porovnává s celkovým počtem všech dvojic

$$\tau = \frac{P-Q}{n(n-1)}. \tag{3.11}$$

Statistika τ má pěknou interpretaci. Nahradíme-li každou dvojici (x_i, x_j) znaménkem rozdílu $x_i - x_j$, podobně pro dvojice (y_i, y_j), pak v případě, že mezi pozorováními nejsou netriviální shody, je τ korelačním koeficientem těchto nových hodnot.

Jakmile dojde mezi pozorováními ke shodám, nemůže statistika τ nabývat krajních hodnot. Data můžeme shrnout do kontingenční tabulky rozměru $r \times c$, kde r je počet různých hodnot mezi x_i a c je počet různých hodnot mezi y_i. Korelační koeficient je pak roven statistice *Kendallova* τ_b

$$\tau_b = \frac{P-Q}{\sqrt{n^2 - \sum_i N_{i\bullet}^2}\sqrt{n^2 - \sum_j N_{\bullet j}^2}}. \tag{3.12}$$

Jeho zjednodušenou verzí je *Stuartovo (Kendallovo)* τ_c

$$\tau_c = \frac{P-Q}{n^2}\frac{m}{m-1}, \tag{3.13}$$

kde je opět $m = \min(r, c)$.

```
> print(P0 <- sum(sgn.plan * sgn.vzdelani>0))
[1] 2176
> print(Q0 <- sum(sgn.plan * sgn.vzdelani<0))
[1] 898
> (P0-Q0)/(P0+Q0)
[1] 0.4157450
> print(P <- sum(sgn.plan * sgn.vzdelani>0))
[1] 2176
> print(Q <- sum(sgn.plan * sgn.vzdelani<0))
[1] 898
> print(gamma <- (P-Q)/(P+Q))
[1] 0.4157450
> print(tau.B <- (P-Q)/sqrt((n**2-sum(N.radky**2))*(n**2-sum(N.sloupce**2))))
[1] 0.2370382
> print(m <- min(dim(N)))
[1] 2
> print(tau.C <- (m/(m-1))*(P-Q)/(n**2))
[1] 0.2607897
```


Nesymetrickou variantou těchto statistik při předpovídání sloupcového indexu ze známého řádkového indexu je *Sommersovo* $d_{Y|x}$

$$d_{Y|x} = \frac{P-Q}{n^2 - \sum_i N_{i\bullet}^2}. \tag{3.14}$$

Podobně lze zavést $d_{X|y}$ malou úpravou jmenovatele. Všimněte si, že platí

$$d_{Y|x} d_{X|y} = \tau_b^2,$$

což je obdoba známého vztahu mezi směrnicemi regresních přímek závislostí Y na x a X na y a korelačním koeficientem mezi X, Y.

```
> print(d.Yx <- (P-Q)/(n**2-sum(N.radky**2)))
[1] 0.2090969
> print(d.Xy <- (P-Q)/(n**2-sum(N.sloupce**2)))
[1] 0.2687132
> d.Xy*d.Yx
[1] 0.05618709
> tau.B**2
[1] 0.05618709
```

Poznámka Zvolili jsme označení, které používá program Statistica, tedy Y znamená index sloupce a x index řádku.

3.3 Přesný test

Spíše jako příklad ukažme použití myšlenky Fisherova exaktního (faktoriálového) testu na situaci s ordinálními veličinami (viz (Agresti, 1990, str. 64)).

	cigarety/den			
	0	$1-24$	≥ 25	
kontroly	25	25	12	62
infarkty	0	1	3	4
	25	26	15	66

Budeme se zabývat možností prokázat souvislost mezi počtem kouřených cigaret a výskytem infarktu myokardu. Pokud jsou dva znaky nezávislé, pak pro dané marginální četnosti $N_{i\bullet}, N_{\bullet j}$, je podmíněná pravděpodobnost dané tabulky rovna (jak se to dostane?)

$$\frac{\prod_{i=1}^r N_{i\bullet}! \prod_{j=1}^c N_{\bullet j}!}{n! \prod_{i=1}^r \prod_{j=1}^c N_{ij}!}. \tag{3.15}$$

Když testujeme hypotézu, že dva znaky jsou nezávislé proti alternativní hypotéze, že *kouření zvyšuje* riziko infarktu, sečteme pravděpodbnost naší tabulky a tabulek ještě více svědčících pro alternativní hypotézu. K porovnání použijeme statistiku γ.

Pro naši tabulku platí

$$\begin{aligned}P &= 25*(1+3)+25*3+1*25+3*(25+25)=350,\\ Q &= 12*1+1*12=24,\\ \gamma &= \frac{350-24}{350+24}=0{,}871658.\end{aligned}$$

Jedinou tabulku, u níž je tento poměr horší, dostaneme, když budou všichni postižení infarktem patřit mezi nejsilnější kuřáky. Pak vyjde

$$\begin{aligned}P &= (25+26)*4+4*(25+26)=408,\\ Q &= 0,\\ \gamma &= 1.\end{aligned}$$

Dosažená hladina (p-hodnota) je tedy rovna

$$\begin{aligned}p &= \frac{62!4!25!26!15!}{66!}\left(\frac{1}{25!25!12!0!1!3!}+\frac{1}{25!26!11!0!0!4!}\right)\\ &\doteq 0{,}018308.\end{aligned}$$

Na běžné 5 % hladině jsme závislost prokázali. Vlastní výpočet v R:

```
> c<-lgamma(63)+lgamma(5)+lgamma(26)+lgamma(27)+lgamma(16)-lgamma(67)
> p1<-exp(c-(lgamma(26)+lgamma(26)+lgamma(13)+lgamma(1)+lgamma(2)+lgamma(4)))
> p1
[1] 0.01641414
> p2<-exp(c-(lgamma(26)+lgamma(27)+lgamma(12)+lgamma(1)+lgamma(1)+lgamma(5)))
> p2
[1] 0.001893939
> p1+p2
[1] 0.01830808
```

3.4 Střední chyby

Máme-li porovnávat míry asociace nezávisle odhadnuté ve dvou populacích, potřebovali bychom znát jejich výběrovou variabilitu. Existují asymptotické vzorce, jsou to však komplikované aproximace (viz například encyklopedickou knihu Kendall, Stuart (1969)) nebo manuály dobře dokumentovaného programového vybavení. I zde je třeba být obezřetný, zhusta se uvádějí dva vzorce, jeden předpokládá nezávislost, druhý je vhodnější při určování intervalu spolehlivosti pro populační protějšek.

3.5 Test trendu v tabulce $r \times 2$

Uvažujme o závislosti dvou znaků. Řádky kontingenční tabulky jsou dány r hodnotami ordinálního znaku, sloupce jsou dány právě dvěma hodnotami nominálního (či ordinálního) znaku. Může jít například o situaci, kdy máme r nezávislých výběrů z binomického rozdělení

$$N_{i1} \sim \mathsf{bi}(n_{i\bullet}, \pi_{i1}), \qquad i = 1, \ldots, r.$$

Test nulové hypotézy $\mathsf{H}_0 : \pi_{11} = \ldots = \pi_{r1}$ vede zřejmě ke klasické testové statistice X^2 z (3.1). Kritický obor je zvolen tak, aby byl citlivý na všechna možná porušení H_0, předpokládá tedy alternativní hypotézu H_1, která je negací H_0.

Budeme uvažovat alternativní hypotézu s monotonní posloupností pravděpodobností $\mathsf{H}_< : \pi_{11} < \ldots < \pi_{r1}$. Do kritického oboru by měly padnout takové výsledky pokusu, kdy relativní četnosti $N_{i1}/n_{i\bullet}$ spíše rostou s indexem i. V dalším popíšeme postup s oboustranným protějškem takové alternativní hypotézy, kdy pravděpodobnosti buď monotonně rostou nebo klesají. Řešení pro případ jednostranné alternativní hypotézy se dostane snadnou úpravou.

Připomeňme tvrzení o vážené lineární regresi. Mějme náhodný vektor $\mathbf{V} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{W}^{-1})$, kde známá matice $\mathbf{W}$ je pozitivně definitní. Odhadem $\boldsymbol{\beta}$ je pak

$$\mathbf{b}_W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}.$$

Je-li navíc matice $\mathbf{W}$ diagonální, s prvky w_i na diagonále, lze reziduální součet čtverců napsat pomocí $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}_W$ jako $RSS = \sum_i w_i(Y_i - \hat{Y}_i)^2$.

Vraťme se k úloze s klasickou alternativní hypotézou H_1 a testovou statistiku odvoďme poněkud jinak. V použitém modelu pro dostatečně velké marginální četnosti $n_{i\bullet}$ mají relativní četnosti $\hat{\pi}_{i\bullet} = N_{i1}/n_{i\bullet}$ asymptoticky normální rozdělení

$$\hat{\pi}_{i1} \sim \mathsf{N}\left(\pi_{i1}, \frac{\pi_{i1}(1-\pi_{i1})}{n_{i\bullet}}\right). \tag{3.16}$$

Uvažujme nyní speciální případ, kdy je $\pi_{i1} = \pi$ pro všechna i. Označíme-li $\sigma^2 = \pi(1-\pi)$ a použijeme-li místo w_i četnosti $n_{i\bullet}$ a místo Y_i podíly $N_{i1}/n_{i\bullet}$, máme (asymptoticky) model vážené lineární regrese s maticí $\mathbf{X} = \mathbf{1}$ a jednoprvkovým vektorem $\boldsymbol{\beta}$. Protože je pak $\mathbf{X}'\mathbf{W}\mathbf{X} = \sum_i n_{i\bullet} = n$ a $\mathbf{X}\mathbf{W}\mathbf{Y} = \sum_i N_{i1} = N_{\bullet 1}$, dostaneme odhad $b_W = N_{\bullet 1}/n = \hat{\pi}$. Konzistentním odhadem σ^2 bude výraz $\hat{\pi}(1-\hat{\pi})$, takže asymptoticky bude mít výraz

$$X^2 = \frac{RSS}{\hat{\pi}(1-\hat{\pi})} = \frac{\sum_i n_{i\bullet}(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}(1-\hat{\pi})} \tag{3.17}$$

rozdělení $\chi^2(r-1)$. Nepřekvapí proto tvrzení, že poslední výraz je jen jiným vyjádřením pro klasickou testovou statistiku X^2.

Vraťme se k alternativní hypotéze $\mathsf{H}_<$. Předpokládejme, že pravděpodobnosti π_i jsou lineární funkcí nějakých *daných* čísel x_i. Místo modelu (3.16) budeme nyní uvažovat podmodel

$$\hat\pi_{i1} \sim \mathsf{N}\left(\beta_0 + \beta_1 x_i, \frac{\pi_{i1}(1-\pi_{i1})}{n_{i\bullet}}\right). \tag{3.18}$$

Hypotéza H_0 znamená $\beta_1 = 0$ a také $\pi_{i1} = \pi$. Při odhadu β_1 a výpočtu reziduálního součtu čtverců předpokládáme $\pi_{i1} = \pi$. Matice $\mathbf{X}$ má tentokrát (jako u regresní přímky) dva sloupce (jedničky, hodnoty x_i). Odhady se zjednoduší na

$$b_{W1} = \frac{n\sum_i N_{i1}x_i - N_{\bullet 1}\sum_i n_{i\bullet}x_i}{n\sum_i n_{i\bullet}x_i^2 - (\sum_i n_{i\bullet}x_i)^2}, \tag{3.19}$$

$$\mathsf{var}\, b_{W1} = \frac{n\pi(1-\pi)}{n\sum_i n_{i\bullet}x_i^2 - (\sum_i n_{i\bullet}x_i)^2}. \tag{3.20}$$

Použijeme-li stejný konzistentní odhad rozptylu $\pi(1-\pi)$, můžeme rozhodovat pomocí statistiky (Waldův test)

$$W = \frac{b_1}{\sqrt{\mathsf{var}\, b_1}} = \frac{n}{\sqrt{N_{\bullet 1}N_{\bullet 2}}}\frac{\sum_i N_{i1}x_i - (N_{\bullet 1}/n)\sum_i n_{i\bullet}x_i}{\sqrt{\sum_i n_{i\bullet}x_i^2 - (\sum_i n_{i\bullet}x_i)^2/n}},$$

která má za nulové hypotézy asymptoticky rozdělení $\mathsf{N}(0,1)$.

Pokud se zajímáme o oboustrannou alternativní hypotézu, kdy pravděpodobnosti $\pi_{11}, \ldots \pi_{r1}$ tvoří monotonní posloupnost, stačí použít W^2, která má za platnosti H_0 asymptoticky rozdělení $\chi^2(1)$. Velikost W^2 vypovídá o nulovosti či nenulovosti regresního koeficientu β_1, rozdíl $X^2 - W^2$ vypovídá o linearitě závislosti pravděpodobností π_i na x_i, a tedy vlastně také o vhodnosti volby skórů x_i.

Ještě k samotnému výpočtu v prostředí R. Je možno použít běžný lineární model `lm()`. Když si představíme, že závisle proměnná nabývá hodnot 1 (v i-tém souboru v N_{i1} případech) a hodnoty 0 (v N_{i2} případech. Potom vztah (3.19) udává běžnou směrnici přímky a následující vzorec (3.20) podobně rozptyl uvedeného odhadu. Prakticky jediný rozdíl je v tom, že v našem specielním modelu souvisí parametr σ^2 s odhadovanými pravděpodobnostmi. Proto statistiku W lze dopočítat z t statistiky T o odhadu b_1 pomocí

$$W = T\sqrt{\frac{s^2}{\hat\pi(1-\hat\pi)}} = T\frac{sn}{\sqrt{N_{\bullet 1}N_{\bullet 2}}}$$

```
> attach(kojeni)
> print(N<-table(vzdelani,plan))
        plan
vzdelani  0  1
       1 20 14
```



```
       2 16 31
       3  5 13
> N.c<-apply(N,2,sum)
> N.n<-sum(N)
> print(a<-summary(lm(as.numeric(plan)~as.numeric(vzdelani))))

Call:
lm(formula = as.numeric(plan) ~ as.numeric(vzdelani))

Residuals:
    Min      1Q  Median      3Q     Max
-0.7827 -0.4438  0.2173  0.3868  0.5562

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.27433    0.13524   9.423 2.44e-15 ***
as.numeric(vzdelani)  0.16946    0.06867   2.468   0.0153 *
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

Residual standard error: 0.4827 on 97 degrees of freedom
Multiple R-Squared: 0.05908,    Adjusted R-squared: 0.04938
F-statistic:  6.09 on 1 and 97 degrees of freedom,      p-value: 0.01535

> W<-a$coefficients[2,3]*a$sigma*N.n/sqrt(N.c[1]*N.c[2])
> W
       0
2.418371
> 1-pchisq(W**2,1)
         0
0.01559016
> chisq.test(N)

         Pearson's Chi-squared test

data:  N
X-squared = 6.6794, df = 2, p-value = 0.03545

> dif<-chisq.test(N)$statistic-W**2
> print(dif<-chisq.test(N)$statistic-W**2)
X-squared
0.8308548
> 1-pchisq(dif,1)
X-squared
0.3620255
```

Závislost podílu plánovaných těhotenství na dosaženém vzdělání jsme dokázali jednat „univerzálně“, při běžné alternativní hypotéze běžným χ^2-testem s dosaženou hladinou $p = 0{,}03545$. Statistiku $\chi^2 = 6{,}6794$ lze rozložit na dvě složky, z nichž první $W^2 = 2{,}4184^2 = 5{,}8485$ vypovídá o nenulovosti směrnice ($p = 0{,}0156$) a druhá o případné nelinearitě: $\chi^2 - w^2 = 0{,}8309$, což linearitu nevylučuje ($p = 0{,}3620$).

Je tu zajímavá souvislost s počtem konkordantních a diskordantních dvojic. Označme rozdíl jejich počtů symbolem $S = P - Q$. Lze ukázat, že v případě nezávislosti (zde v případě platnosti H_0) je (viz Armitage (1955))

$$\mathsf{var}\, S = \frac{N_{\bullet 1} N_{\bullet 2}}{3n(n-1)} \left(n^3 - \sum_i n_{i\bullet}^3 \right).$$

Zvolíme-li speciálně $x_i = i$, pak se dá ukázat, že platí

$$\frac{S^2}{\mathsf{var}\, S} = \frac{n-1}{n} \frac{b_1^2}{\mathsf{var}\, b_1}.$$

V uvedeném speciálním případě tedy máme dva asymptoticky ekvivalentní testy.

Pro úplnost, uveďme ještě další řešení naší úlohy, pomocí zobecněného lineárního modelu (logistické regrese).

```
> summary(a.lin<-glm(plan~as.numeric(vzdelani),family=binomial))

Call:
glm(formula = plan ~ as.numeric(vzdelani), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7232  -1.0750   0.7168   0.9772   1.2834

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.9825     0.5893  -1.667   0.0954 .
as.numeric(vzdelani)   0.7368     0.3111   2.368   0.0179 *
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.31  on 98  degrees of freedom
Residual deviance: 128.29  on 97  degrees of freedom
AIC: 132.29

Number of Fisher Scoring iterations: 3
```



```
> summary(a.anova1<-glm(plan~vzdelani,family=binomial))

Call:
glm(formula = plan ~ vzdelani, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6006  -1.0302   0.8068   0.9123   1.3321

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4201     0.2339   1.796   0.0726 .
vzdelani.L    0.9279     0.4460   2.080   0.0375 *
vzdelani.Q   -0.2956     0.3598  -0.822   0.4114
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.31  on 98  degrees of freedom
Residual deviance: 127.62  on 96  degrees of freedom
AIC: 133.62

Number of Fisher Scoring iterations: 3

> anova(a.lin,a.anova1,test="Chisq")
Analysis of Deviance Table

Model 1: plan ~ as.numeric(vzdelani)
Model 2: plan ~ vzdelani
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        97    128.294
2        96    127.624  1    0.670     0.413
```

Došli jsme k analogickým závěrům: závislost na vzdělání lze považivat za lineární a je průkazná. S rostoucím vzděláním matky roste pravděpodobnost, že těhotenství je považováno za plánované.

Poznámka Aby to nebylo tak jednoduché. Podívejme se na možnou závislost podílu plánovaných těhotenství na věku matky, podívejme se také na souvislost věku matky a jejího vzdělání.

```
> anova(glm(plan~vek.m,family=binomial),test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit
```

```
Response: plan

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       98    134.309
vek.m  1    4.041        97    130.268     0.044
> kruskal.test(vek.m,vzdelani)

         Kruskal-Wallis rank sum test

data:  vek.m and vzdelani
Kruskal-Wallis chi-squared = 29.4765, df = 2, p-value = 3.974e-07
```

Věk a vzdělání spolu (pochopitelně) prokazatelně souvisejí. Co tedy ovlivňuje podíl plánovaných těhotenství, vzdělání nebo věk?

Literatura

A. Agresti (1990). *Categorical Data Analysis.* Wiley, New York.

J. Anděl (1978). *Matematická statistika.* SNTL, Praha.

J. Anděl (1993). *Statistické metody.* MATFYZPRESS, Praha.

J. Antoch, D. Vorlíčková (1992). *Vybrané metody statistické analýzy dat.* Academia, Praha.

P. Armitage (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, pages 375–386.

J. Janko (1958). *Statistické tabulky.* Academia, Praha.

J. Jurečková (1981). Pořadové testy.

M. G. Kendall, A. Stuart (1969). *The Advanced Theory of Statistics*, volume II. Griffin.

J. Likeš, J. Laga (1978). *Základní statistcké tabulky.* SNTL, Praha.

C. R. Rao (1978). *Lineární metody statistické indukce a jejich aplikace.* Academia, Praha.

G. W. Snedecor, W. G. Cochran (1967). *Statistical Methods.*