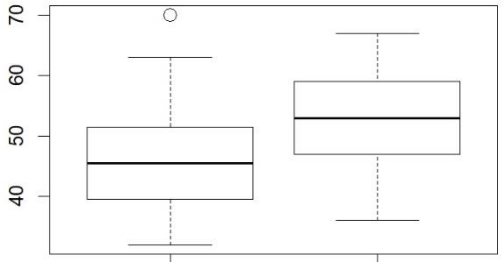
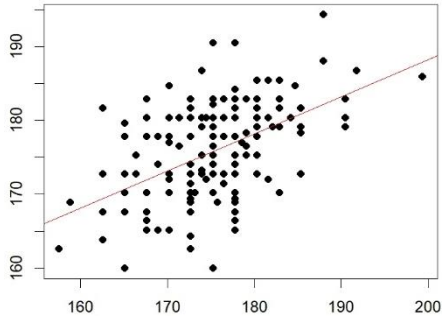
Přehled základních metod:

<table>
<tr><td></td><td colspan="2">Závislá proměnná</td></tr>
<tr><td>Nezávislá proměnná</td><td>Spojitá (kvantita)
(délka, hmotnost, množství)</td><td>Nominální (kvalita)
(kategorie, vlastnosti)</td></tr>
<tr><td>Nominální
(skupiny, faktory, kategorie)</td><td> 

t-test, Wilcoxon, ANOVA</td><td><table>
<tr><th rowspan="2">Porodnice</th><th colspan="3">Vzdělání matky</th></tr>
<tr><th>ZŠ</th><th>SŠ</th><th>VŠ</th></tr>
<tr><td>pražská</td><td>23</td><td>30</td><td>17</td></tr>
<tr><td>okresní</td><td>11</td><td>17</td><td>1</td></tr>
</table>
Kontingenční tabulky</td></tr>
<tr><td>Spojitá
(délka, hmotnost, množství, …)</td><td> 

Korelační a regresní analýza</td><td>(Logistická regrese
výskyt rakoviny plic [ano/ne] v závislosti na počtu vykouřených cigaret)</td></tr>
</table>

Příkladová data

Je plánování těhotenství závislé na vzdělání matky?

Plán?	Vzdělání matky		
	ZŠ	SŠ	VŠ
Plánované	14	31	13
Neplánované	20	16	5

Je rozdíl ve struktuře vzdělání matek rodících v Praze a v okresní porodnici?

Porodnice	Vzdělání matky		
	ZŠ	SŠ	VŠ
pražská	23	30	17
okresní	11	17	1

Hodnocení kvalitativních dat

- Pozorování jsou na nominální škále (krevní skupiny: A – B – AB – 0).
- Lze zahrnout i ordinální škálu, ale nepracujeme s vlastností uspořádání hodnot. (Postupy pro ordinální škálu také existují.)
- Nominální proměnná se chová jako faktor: má ***k*** úrovní, pozorování padne do právě jedné úrovně (kategorie).
- Výsledkem třídění do úrovní je potom ***k***-tice četností.
  50 zkoumaných osob charakterizujeme těmito četnostmi: $(24, 10, 3, 13)$, tj. 24 osob s krevní skupinou ‚*A*', 10 osob ‚*B*', 3 osoby ‚*AB*' a 13 osob ‚*0*'
- Výsledné četnosti považujeme za konkrétní realizace náhodných proměnných, které dohromady tvoří náhodný vektor.
- Pravděpodobnostním modelem pro takový vektor je multinomické rozdělení.

Multinomické rozdělení

- Řada nezávislých pokusů (celkem $\boldsymbol{n}$), v každém pokusu získám právě jednu hodnotu z $\boldsymbol{k}$ možných (např. krevní skupina: *A* nebo *B* nebo *AB* nebo *0*).
- Jednotlivé hodnoty nastávají s pravděpodobnostmi $\boldsymbol{\pi_1, \pi_2, \pi_3, \dots, \pi_k}$ a tyto pravděpodobnosti jsou pro všechny pokusy stejné.
- Samozřejmě platí $\boldsymbol{\pi_1 + \pi_2 + \pi_3 + \cdots + \pi_k = 1}$.
- $\boldsymbol{n_j}$ je počet pokusů (**četnost**), ve kterých jsem získala $\boldsymbol{j}$-tou hodnotu (např. krevní skupinu *AB*). V „jazyku“ náhodných veličin zapíšeme $\boldsymbol{Y_j = n_j}$.
- Zde platí $\boldsymbol{n_1 + n_2 + \cdots + n_k = n}$.
- Hodnoty $\boldsymbol{n_j}$ (či Y_j) jsou vzájemně závislé, musí splňovat podmínku $\sum_{j=1}^{k} \boldsymbol{n_j = n}$.
- Náhodný vektor $(\boldsymbol{Y_1, Y_2, \dots, Y_k})$ má multinomické rozdělení s parametry $(\boldsymbol{n, \pi_1, \pi_2, \pi_3, \dots, \pi_k})$.

Multinomické rozdělení $(Y_1, Y_2, \dots, Y_k) \sim Multi(n, \pi_1, \pi_2, \dots, \pi_k)$

Pravděpodobnost, že náhodný vektor nabyde hodnot právě $n_1, n_2, n_3, \dots, n_k$ je

$$P(Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k) = \frac{n!}{n_1!\, n_2! \dots n_k!} \cdot \pi_1^{n_1} \cdot \pi_2^{n_2} \cdot \dots \cdot \pi_k^{n_k}$$

Faktoriálový člen uvádí počet možných kombinací, jak můžeme ***n*** pokusů rozdělit tak, aby platilo $Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k$.

Vlastnost: ***k***-tice náhodných veličin $(Y_1, Y_2, \dots, Y_k)$ má ***k − 1*** stupňů volnosti, protože při daném celkovém počtu pokusů ***n*** mohu „volit“ ***k − 1*** hodnot Y_j, ale tu poslední hodnotu musím dopočítat tak, aby byl součet roven ***n***, tj. $Y_k = n - Y_1 - \dots - Y_{k-1}$

Příklad krevní skupiny: podíly čtyř krevních skupin v populaci jsou v tabulce.
Budu-li zkoumat skupinu 50 osob, **očekávám střední četnosti** v jednotlivých skupinách $n \cdot \pi_j$
Skutečné počty se budou pohybovat kolem této střední hodnoty, jsou to náhodné veličiny.

Krevní skupina	***A***	***B***	***AB***	***0***
Podíl v populaci:	43 %	19 %	10 %	28 %
Pravděpodobnost π_j	0.43	0.19	0.1	0.28
Střední četnosti $n\pi_j$	21.5	9.5	5	14

Multinomické rozdělení – souvislost s binomickým rozdělením

- $(\boldsymbol{Y_1, Y_2, \dots, Y_k}) \sim \boldsymbol{Multi(n, \pi_1, \pi_2, \pi_3, \dots, \pi_k)}$
- Pro $\boldsymbol{k}$ = 2 máme $\boldsymbol{P(Y_1 = n_1, Y_2 = n_2) = \frac{n!}{n_1! n_2!} \cdot \pi_1^{n_1} \cdot \pi_2^{n_2}}$

  a protože $n_1 + n_2 = n \rightarrow n_2 = n - n_1$

  a $\pi_1 + \pi_2 = 1 \rightarrow \pi_2 = 1 - \pi_1$

  Odtud: $\boldsymbol{P(Y_1 = n_1, Y_2 = n_2) = \frac{n!}{n_1!(1-n_1)!} \cdot \pi_1^{n_1} \cdot (1 - \pi_1)^{n-n_1}}$

- Obecně každé $\boldsymbol{Y_j}$ z vektoru $(\boldsymbol{Y_1, Y_2, \dots, Y_k})$ má samostatně binomické rozdělení $\boldsymbol{Y_j \sim Bi(n, \pi_j)}$, kde $1 - \pi_1 = \pi_2 + \pi_3 + \cdots + \pi_k$

  a $n - n_1 = n_2 + n_3 + \cdots + n_k$

  Potom také $\boldsymbol{EY_j} = n \cdot \pi_j$ a $\boldsymbol{var\ Y_j} = n\pi_j(1 - \pi_j)$

Test dobré shody, χ^2-test [goodness of fit, chi-squared test]

Určitou obdobou centrální limitní věty je tvrzení, že pro dostatečně velké ***n*** má statistika $X^2 = \sum_{j=1}^{k} \frac{(Y_j - n\pi_j)^2}{n\pi_j}$ asymptoticky rozdělení $\chi^2_{(k-1)}$.

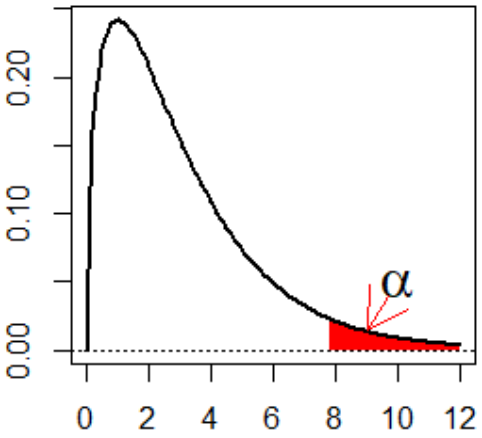
X^2 je velké χ , čti [chí kvadrát.]

Dostatečně velké ***n*** většinou znamená, že pro všechna ***j*** jsou $n\pi_j \geq 5$.

Pomocí této statistiky testuji hypotézu H_0 , že četnosti znaků v mém výběru odpovídají očekávaným četnostem pro dané populační pravděpodobnosti.

Jiné formulace: – že výběr je reprezentativní
– že výběr pochází z dané populace

Hypotézu zamítám, když $X^2 > \chi^2_{(k-1)}(1-\alpha)$, tedy celou α vložím na pravý chvost rozdělení. To proto, že nulové hypotéze odporují velké odchylky od očekávaných četností, tedy velký součet všech ***k*** zlomků.



Test dobré shody - příklad

Krevní skupiny.
podíly čtyř krevních skupin v populaci jsou v tabulce.
Budu-li zkoumat skupinu 50 osob,
očekávám střední četnosti
v jednotlivých skupinách $n \cdot \pi_j$

Krevní skupina	***A***	***B***	***AB***	***0***
Podíl v populaci:	43 %	19 %	10 %	28 %
Pravděpodobnost π_j	0.43	0.19	0.1	0.28
Střední četnosti $n\pi_j$	21.5	9.5	5	14

V souboru 50 pokusných osob jsme zaznamenali tyto četnosti:

Pozorované četnosti:	25	7	1	17

Odpovídají tyto četnosti frekvencím v celé populaci?

Test dobré shody - příklad

Krevní skupiny. Odpovídají pozorované četnosti podílům v celé populaci?

R: **`chisq.test(x=c(25,7,1,17),p=c(.43, .19, .1, .28))`**

Zadávám vektor pozorovaných četností a „očekávaných“ pravděpodobností.

```
Chi-squared test for given probabilities
data:  c(25, 7, 1, 17)
X-squared = 5.0705, df = 3, p-value = 0.1667
```

Krevní skupina	***A***	***B***	***AB***	***0***
Podíl v populaci:	43 %	19 %	10 %	28 %
Pravděpodobnost π_j	0.43	0.19	0.1	0.28
Střední četnosti $n\pi_j$	21.5	9.5	5	14
Pozorované četnosti:	25	7	1	17

Test dobré shody - poznámka

Statistika $\boldsymbol{X^2 = \sum_{j=1}^{k} \frac{(Y_j - n\pi_j)^2}{n\pi_j}}$ se někdy zjednodušeně zapisuje jako

$$X^2 = \sum_{j=1}^{k} \frac{(\text{empirické} - \text{očekávané})^2}{\text{očekávané}} .$$

Empirické = pozorované, naměřené četnosti.

Toto značení v sobě skrývá past. Jako překlad do angličtiny se používají termíny **o**bserved a **e**xpected, která mají počáteční písmena přesně opačná než české pojmy.

Vzorec tvaru $X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$ odpovídá anglickému názvosloví.

Zde budu používat $\boldsymbol{\sum_{j=1}^{k} \frac{(n_j - o_j)^2}{o_j}}$, tj. „**n**aměřené mínus **o**čekávané“.

Kontingenční tabulky [contingency tables, crosstabulations tables]

Dvě různé otázky:

(a) Hypotéza o nezávislosti proměnných

(b) Hypotéza o pravděpodobnostní struktuře výběrů

Příklad: Kojení – vzdělání matky a plánované těhotenství

→ ***1** výběr, třídění podle **2** nominálních proměnných (faktorů)*

Plán?	**Vzdělání** základní	**Vzdělání** střední	**Vzdělání** VŠ
Plánované	14	31	13
Neplánované	20	16	5

Otázka: je plánování těhotenství závislé na vzdělání matky?

Kontingenční tabulky

Nové pojmy: **sdružené četnosti** a **marginální četnosti**

Plán?	**Vzdělání** základní	**Vzdělání** střední	**Vzdělání** VŠ	Řádkové součty
Plánované	14	31	13	58
Neplánované	20	16	5	41
Sloupcové součty	34	47	18	99

Proměnná R	**Proměnná C**			Celkem
	1	…	c	
1	n_{11}	…	n_{1c}	$n_{1\bullet}$
r	n_{r1}	…	n_{rc}	$n_{r\bullet}$
Celkem	$n_{\bullet 1}$	…	$n_{\bullet c}$	n

Kontingenční tabulky

K četnostem patří také pravděpodobnosti:

- π_{ij} – teoretická pravděpodobnost pro výskyt kombinace úrovní ***i + j***
- p_{ij} – odhad pravděpodobnosti π_{ij}
- $\pi_{i\bullet}$ – teoretická pravděpodobnost úrovně ***i*** nezávisle na sloupcových úrovních ***j***
- $\pi_{\bullet j}$ – teoretická pravděpodobnost úrovně ***j*** nezávisle na řádkových úrovních ***i***

Proměnná R	**Proměnná C**			Celkem
	1	…	c	
1	π_{11}	…	π_{1c}	$\pi_{1\bullet}$
r	π_{r1}	…	π_{rc}	$\pi_{r\bullet}$
Celkem	$\pi_{\bullet 1}$	…	$\pi_{\bullet c}$	1

Kontingenční tabulky – test nezávislosti

- Nezávislost mezi nominálními proměnnými znamená, že četnosti v úrovních proměnné (faktoru) R nejsou závislé na konkrétní úrovni proměnné C.
- **Definice nezávislosti:** $\boldsymbol{P\left(X_i = n_i,\ X_j = n_j\right) = P(X_i = n_i) \cdot P\left(X_j = n_j\right)}$
- Pro pravděpodobnosti v kont. tabulce: $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$
- Pravděpodobnosti neznáme, musíme je tedy odhadnout z dat.
- Pravděpodobnosti pozorované odhadujeme jako relativní četnosti ze sdružených četností: $p_{ij} = \frac{n_{ij}}{n}$.
- Pravděpodobnosti očekávané za platnosti hypotézy o nezávislosti proměnných odhadujeme jako $p_{i\bullet} = \frac{n_{i\bullet}}{n}$ a $p_{\bullet j} = \frac{n_{\bullet j}}{n}$.
- Nyní mohu spočítat četnosti očekávané za předpokladu nezávislosti:

$$\boldsymbol{o_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j}} = n \cdot \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

Kontingenční tabulky – test nezávislosti

→ Nulová hypotéza H_0: úrovně (znaky) obou nominálních proměnných se vyskytují na sobě nezávisle.

Testová statistika chí-kvadrát:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-o_{ij})^2}{o_{ij}} \dot{\sim}_{H_0} \chi^2_{(r-1)(c-1)}$$

Když nulová hypotéza platí, vyjde χ^2 blízko nuly, protože naměřené četnosti budou odpovídat těm očekávaným.

- Stupně volnosti $df = (r-1)(c-1)$
- ! Podmínka pro užití χ^2 testu: všechny očekávané četnosti $o_{ij} \geq 5$.
- H_0 zamítám, když $X^2 > \chi^2_{(r-1)(c-1)}(1-\alpha)$
- Poznámka: v tomto případě jsou četnosti sdružené i marginální považovány za náhodné veličiny. Četnosti na všech místech tabulky mohou dopadnout pro každý výběr trochu jinak, náhodně v rámci rozsahu výběru.

Test nezávislosti - příklad

Data Kojení: Je plánování těhotenství závislé na vzdělání matky?

Plán?	**Vzdělání** základní	**Vzdělání** střední	**Vzdělání** VŠ	Řádkové součty
Plánované	14	31	13	58
Neplánované	20	16	5	41
Sloupcové součty	34	47	18	99

Jiný formát zapsání dat do tabulky:

R: ze sloupcových proměnných vytvořím tabulku:

```
plan.tab <- xtabs(pocet~plan+vzdelani,
      data=planvzd)
```

	plan	vzdelani	pocet
1	ano	ZS	14
2	ano	SS	31
3	ano	VS	13
4	ne	ZS	20
5	ne	SS	16
6	ne	VS	5

Test nezávislosti - příklad

Data „planovani“: Je plánování těhotenství závislé na vzdělání matky?

```
R:> ktplan <- xtabs(~plan+vzdelani, data=planovani)

  > chisq.test(ktplan)
→  Pearson's Chi-squared test
    data:  ktplan
    X-squared = 6.6794, df = 2, p-value = 0.03545

  > chisq.test(ktplan)$expected
→         vzdelani
    plan   SS   VS   ZS
      ano 27.5 10.5 19.9
      ne  19.5  7.5 14.1
```

Zde kontrola nejmenších četností.
Nicméně Rko kontrolu provádí automaticky a v případě malých četností píše upozornění:
`Chi-squared approximation may be incorrect.`

Kontingenční tabulky – test homogenity

(b) Hypotéza o pravděpodobnostní struktuře (homogenity) výběrů

Příklad: planovani – počet porodů podle vzdělání matky v porodnici pražské a v porodnici okresní

→ ***2*** *(i více) nezávislých výběrů,* ***1*** *nominální proměnná*

Porodnice	Vzdělání matky			Celkem
	základní	střední	VŠ	
pražská	23	30	17	70
okresní	11	17	1	29
Celkem	34	47	18	99

Otázka: je rozdíl ve struktuře vzdělání maminek rodících v Praze a v okresní porodnici?

Kontingenční tabulky – test homogenity

- Jestliže obě „populace“ mají stejnou strukturu/pravděpodobnosti pro jednotlivé skupiny matek podle vzdělání (π_j), můžeme tyto prsti odhadnout takto:

ZŠ: $p_1 = \frac{n_{\bullet 1}}{n}$

SŠ: $p_2 = \frac{n_{\bullet 2}}{n}$

VŠ: $p_3 = \frac{n_{\bullet 3}}{n}$

Porodnice	Vzdělání matky: základní	střední	VŠ	Celkem
pražská	23	30	17	70
okresní	11	17	1	29
Celkem	34	47	18	99
Hypotéza:	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	

- Očekávané četnosti za předpokladu nulové hypotézy potom spočítáme vzhledem k (marginálnímu) počtu maminek v příslušné porodnici:

ZŠ: $o_{11} = n_{1\bullet} \cdot p_1 = \frac{n_{1\bullet} \cdot n_{\bullet 1}}{n}$ SŠ: $o_{12} = n_{1\bullet} \cdot p_2 = \frac{n_{1\bullet} \cdot n_{\bullet 2}}{n}$ atd.

- Obecně: $o_{ij} = n_{i\bullet} \cdot p_j = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$

tedy opět stejný vzorec pro o_{ij}, ale došli jsme k němu z jiných předpokladů!

Kontingenční tabulky – test homogenity

→ Nulová hypotéza H_0: pravděpodobnosti jednotlivých úrovní nominální proměnné jsou pro všechny výběry stejné.

Testová statistika chí-kvadrát (stejná jako v předchozím):

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - o_{ij})^2}{o_{ij}} \dot{\sim}_{H_0} \chi^2_{(r-1)(c-1)}$$

- Stupně volnosti $df = (r-1)(c-1)$
- Kontrola podmínky pro užití χ^2 testu: všechny očekávané četnosti $o_{ij} \geq 5$?
- H_0 zamítám, když $X^2 > \chi^2_{(r-1)(c-1)}(1-\alpha)$
- Poznámka: v tomto případě jsou řádkové marginální četnosti (tj. rozsahy výběrů) považovány za pevné, nenáhodné hodnoty. Protože pro opakování pokusu bychom měli použít opět stejné rozsahy výběrů.

Test homogenity – příklad

Data planovani: Je rozdíl ve struktuře vzdělání maminek rodících v Praze a v okresní porodnici?

Porodnice	Vzdělání matky			Celkem
	základní	střední	VŠ	
pražská	23	30	17	70
okresní	11	17	1	29
Celkem	34	47	18	99

Nulová hypotéza:
Pravděpodobnost, že rodící matka má ZŠ vzdělání je stejná v Praze i mimo Prahu. Totéž platí zároveň pro SŠ i VŠ vzdělání.

```
> ktporod <-xtabs(~porodnice+vzdelani, data=planovani)

> chisq.test(ktporod)
        Pearson's Chi-squared test
data:  ktporod
X-squared = 6.1238, df = 2, p-value = 0.0468
```

$X^2 = 6.12,\ p = 0.047$

Na hladině α = 0.05 těsně zamítám hypotézu o stejné pravděpodobnostní struktuře vzdělání matek v pražské a okresní porodnici.

Čtyřpolní kontingenční tabulky

- Situace: ***1*** *výběr,* ***2*** *nominální proměnné, každá má jen 2 úrovně.*
- Obecně užívané speciální značení, objevuje se např. u koeficientů podobnosti:

a	b	***a+b***
c	d	***c+d***
a+c	***b+d***	***n***

$$\rightarrow X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \dot{\sim}_{H_0} \chi^2_1$$

- Hypotézy nezávislosti nebo homogenity formuluji stejně.
- Kontrola podmínky pro užití χ^2 testu: všechny očekávané četnosti $o_{ij} \geq 5$?
- Pro malé četnosti o_{ij} se používá **Yatesova oprava na spojitost**:

$$X^2_{corr} = \frac{n\left(|ad-bc| - \frac{1}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Potom $X^2_{corr} < X^2$ => $p_{corr} > p$, test je tedy konzervativnější.

Čtyřpolní tabulka – příklad hraboš

Hraboš: 515 hrabošů bylo vyšetřováno na výskyt některého ze dvou druhů parazitů. Jen u 4 jedinců byl zjištěn současný výskyt obou druhů. Existuje nějaká souvislost mezi jejich výskytem?

	Sarcocystis +	Sarcocystis -	celkem
Frenkelia +	4	27	31
Frenkelia -	11	473	484
Celkem	15	500	515

R:
```
> kthrabos = xtabs(pocty~ Frenkelia+Sarcocystis, data=hrabos)
> chisq.test(kthrabos)
        Pearson's Chi-squared test with Yates' continuity correction
data:  kthrabos
X-squared = 8.187, df = 1, p-value = 0.004219
Warning message:    #upozorňuje na malé očekávané četnosti
In chisq.test(hrabos.tab): Chi-squared approximation may be incorrect
> chisq.test(kthrabos)$expected
          Sarcocystis
Frenkelia       Sano       Sne
     Fano  0.9029126  30.09709
     Fne  14.0970874 469.90291
```

Použití Yatesovy korekce je tedy na místě. Doporučit lze i Fisherův přesný test (dále).

Míry těsnosti vazby, asociace (2x2 tabulka) [coefficients of association]

– např. pro studium mezidruhové vazby

Φ – koeficient [phi coefficient]

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

Nedostatek: koeficient ϕ ztratí při výpočtu znaménko +/-, tedy indikátor pozitivní nebo negativní asociace.

Cramérovo V

$$V = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(existuje i složitější varianta pro větší tabulky)

- Interpretace jako pro korelační koeficient, $\in \langle -1 , 1 \rangle$
- koeficienty, jejichž hodnoty nezávisí na počtu pozorování
- POZOR na uspořádání tabulky! Pro rozumnou interpretaci se musí na diagonále potkat ANO-ANO a NE-NE. Výpočetní vzoreček nezohledňuje význam úrovní…

	ANO	*NE*
ANO	a	*b*
NE	*c*	*d*

Míry těsnosti vazby, asociace (2x2 tabulka) [coefficients of association]

R: např. balík „vcd“

```
install.package(„vdc“)  # pozor na uvozovky, R má svoje speciální!
library(vcd)
assocstats(kthrabos)
                     X^2   df     P(> X^2)
Likelihood Ratio  6.8004    1   0.00911382   (G-test poměrem věroh.)
Pearson          11.6429    1   0.00064449   (bez Yatesovy korekce)

Phi-Coefficient   : 0.15    (pro 2x2 tabulky stejné jako Cramerovo V)
Contingency Coeff.: 0.149
Cramer's V        : 0.15    (bez +/- znamének)
```

Fisherův exaktní (faktoriálový) test (2x2 tabulky)

- Jiný princip než χ^2-test, p-hodnotu počítá přímo, patří mezi tzv. podmíněné testy.
- Podmínka: marginální četnosti jsou dány (tedy je nepovažujeme za náhodné veličiny, jako u testu nezávislosti pomocí χ^2 rozdělení).
- Sčítá pravděpodobnost skutečně realizované tabulky a tabulek s danými (stejnými) marginálními četnostmi, které ještě více odporují nulové hypotéze.
- Test je to spíše konzervativní, tzn. že skutečná pravděpodobnost chyby 1. druhu může být výrazně menší než zvolená hladina α.
- Pro tabulky s velkými četnostmi je to výpočetně (časově i paměťově) náročný test.

Fisherův test – příklad hraboš

STAT: nepočítá Fisherův test pro větší četnosti…

R:
```
> fisher.test(kthrabos)

        Fisher's Exact Test for Count Data

data:  kthrabos
p-value = 0.009226
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.377865 23.215915
sample estimates:
odds ratio
  6.322939
```

Fisherův test v Rkovém provedení testuje hypotézu o poměru šancí [odds ratio]. Jako bonus máme odhad tohoto poměru (pro zájemce viz Zvára str. 218) a konfidenční interval tohoto odhadu.

McNemarův test – test symetrie

- Situace: ***1*** *výběr a* ***1*** *nominální proměnná, opakované (párové) měření*
  - na skupině subjektů zjišťujeme četnosti všech úrovní nominální proměnné dvakrát, např. před ošetřením (zásahem) a po ošetření, nebo v jednom roce a v následujícím roce.
- Kontingenční tabulka je pak čtvercová:

Příklad stromy: počty stromů podle míry houbové nákazy ve dvou sezónách. Zajímá nás, jestli došlo mezi sezónami ke změně v četnosti nákazy.

	1995			
1994	**Žádná**	**Mírná**	**Silná**	celkem
Žádná	35	15	1	51
Mírná	11	21	7	39
Silná	3	3	4	10
Celkem	49	39	12	100

McNemarův test – test symetrie

K této úloze je možné vyslovit dvě hypotézy, které se mírně liší:

(1) H_0: marginální pravděpodobnosti jsou shodné, tedy pravděpodobnosti jednotlivých stupňů nákazy jsou v obou sezónách stejné. (**Stuartův test**)

(2) H_0: matice pravděpodobností je symetrická, tedy platí $\pi_{ij} = \pi_{ji}$. (**Bowkerův test symetrie**)
Je-li matice pravděpodobností symetrická, pak jsou také marginální pravděpodobnosti shodné. Proto se test symetrie někdy používá i k testování shody marginálních pravděpodobností.

McNemarův test je původně konstruovaný pro tabulky 2x2, ale název se používá i pro Bowkerův test (např. v Rku). STATistica počítá tento test jen pro čtyřpolní tabulky.

	1995			
1994	**Žádná**	**Mírná**	**Silná**	celkem
Žádná	35	15	1	51
Mírná	11	21	7	39
Silná	3	3	4	10
Celkem	49	39	12	100

McNemarův test – test symetrie

Nulová hypotéza H_0: matice pravděpodobností je symetrická.

Testová statistika:

$$X^2 = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \dot{\sim}_{H_0} \chi^2_{c(c-1)/2}$$

Promyslete význam posunutého indexování ve sčítacích sumách…

1994	1995		
	Žádná	Mírná	Silná
Žádná	35	15	1
Mírná	11	21	7
Silná	3	3	4

Tvar testové statistiky pro tabulku 2x2:

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \dot{\sim}_{H_0} \chi^2_1$$

McNemarův test – příklad

Stromy: Je rozdíl v rozložení houbové nákazy mezi sezónami? Neboli jsou marginální pravděpodobnosti (potažmo četnosti) srovnatelné?

1994	1995			celkem
	Žádná	Mírná	Silná	
Žádná	35	15	1	51
Mírná	11	21	7	39
Silná	3	3	4	10
Celkem	49	39	12	100

R:

```
> stromy=matrix(data=c(35,15,1,11,21,7,3,3,4), nrow=3, ncol=3, byrow=T)
> mcnemar.test(stromy)
        McNemar's Chi-squared test
data:  stromy
McNemar's chi-squared = 3.2154, df = 3, p-value = 0.3596
```

Nezamítám hypotézu o symetrii matice, tedy ani hypotézu o shodnosti marginálních pravděpodobností houbové nákazy v jedné a druhé sezóně. „Houbová“ situace v lese se nezměnila.

Čtyřpolní tabulka – příklad hraboš

	Sarcocystis +	Sarcocystis -	celkem
Frenkelia +	4	27	31
Frenkelia -	11	473	484
Celkem	15	500	515

Hraboš: Jsou pravděpodobnosti výskytu stejné pro oba parazity?
Pozor, pozorování jsou párová, nemůžeme tedy použít test homogenity pro výběr Frenkelia a pro výběr Sarcocystis, každého hraboše bychom pak měli v testu dvakrát!

R:
```
> mcnemar.test(kthrabos)
        McNemar's Chi-squared test with continuity correction
data:  kthrabos
McNemar's chi-squared = 5.9211, df = 1, p-value = 0.01496
```

McNemarův test má přednastavené použití Yatesovy opravy na spojistost, což je v tomto případě správné.
Hypotézu o stejné pravděpodobnosti výskytu zamítáme. Znamená to, že výskyt parazitů je sice korelován (tj. existuje nějaká závislost), ale frekvence výskytu není podobná (symetrická). Pozor, test neříká nic o kauzální závislosti, je to čistě statistická korelace.

Kontingenční tabulky – další témata

- Test poměrem věrohodností [(log-)likelyhood ratio test, G-test]
  R-packages: **`vcd`**, funkce `assocstats`
  `RVAideMemoire`, fce `G.test, pairewise.G.test` (pro mnohonásobné porovnávání)
- Odhad poměru šancí [odds ratio] a jeho konfidenční interval.
- Nabídka eRkové funkce chisq.test(…, simulate.p.value = FALSE, B = 2000), která po přepnutí na TRUE simuluje odhad p-hodnoty …