

VÍCEVÝBĚROVÉ TESTY A ANALÝZA KONTINGENČNÍCH TABULEK

Martina Litschmannová, Adéla Vrtková

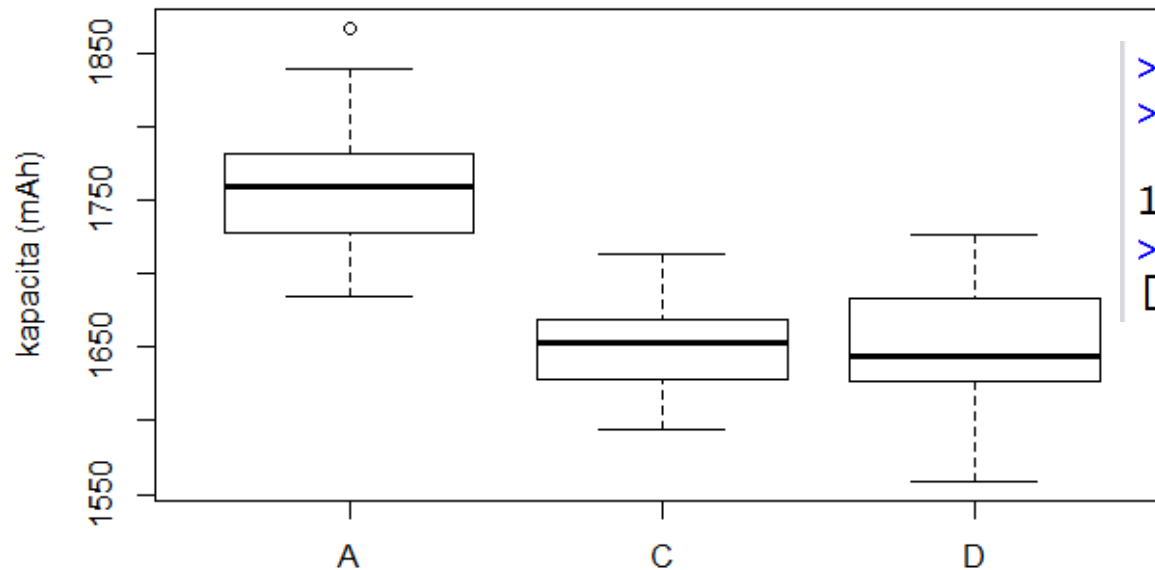


Část 1 – Vícevýběrové testy

- Vícevýběrové testy o shodě rozptylů
- ANOVA (analýza rozptylu = vícevýběrový test o shodě středních hodnot)
- Kruskalův – Wallisův test (vícevýběrový test o shodě mediánů)

Ověření shody rozptylů (homoskedasticity) ve více než dvou výběrech

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?



```
> s2 = tapply(data.acd$kap, data.acd$vyrobce, var)
> s2
```

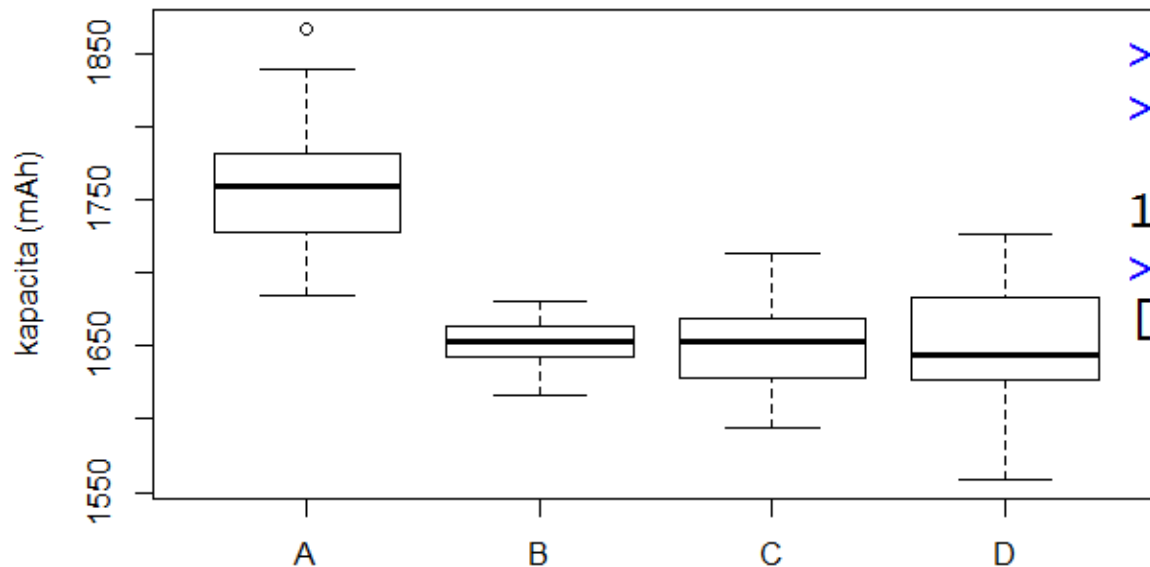
	A	C	D
	1327.2193	860.8913	1481.5059

```
> max(s2)/min(s2)
[1] 1.720898
```

Tato data pravděpodobně **splňují** předpoklad homoskedasticity.

Ověření shody rozptylů (homoskedasticity) ve více než dvou výběrech

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?



```
> s2 = tapply(data$kap, data$vyrobce, var)
> s2
```

A	B	C	D
1327.2193	231.0182	860.8913	1481.5059

```
> max(s2)/min(s2)
```

```
[1] 6.412941
```

$\frac{s_{max}^2}{s_{min}^2} > 2 \Rightarrow$ podezření na heteroskedasticitu
(různé rozptyly)

Tato data pravděpodobně **nesplňují** předpoklad homoskedasticity.

Ověření shody rozptylů (homoskedasticity) ve více než dvou výběrech

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu exaktně pomocí testování hypotéz?

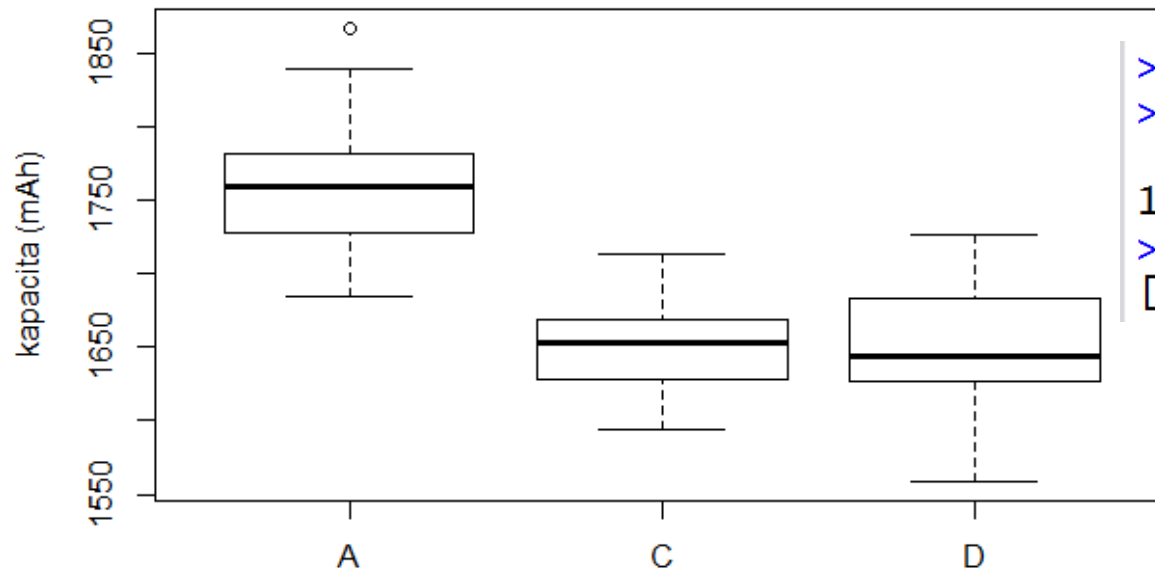
H_0 : Rozptyly měřené veličiny jsou v jednotlivých skupinách shodné.

H_A : $\neg H_0$ (Alespoň ve dvou skupinách se rozptyly měřené veličiny liší.)

Požadovaný typ analýzy	Předpoklady		Testy
Ověření shody rozptylů (<u>homoskedasticity</u>)	Normalita	Vyvážené třídění	<u>Cochranův test</u>
			<u>Hartlevův test</u>
		Nevyvážené třídění	<u>Bartlettův test</u>
	---		<u>Leveneuův test</u>

Ověření shody rozptylů (homoskedasticity) ve více než dvou výběrech

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?



```
> s2 = tapply(data.acd$kap, data.acd$vyrobce, var)
> s2
```

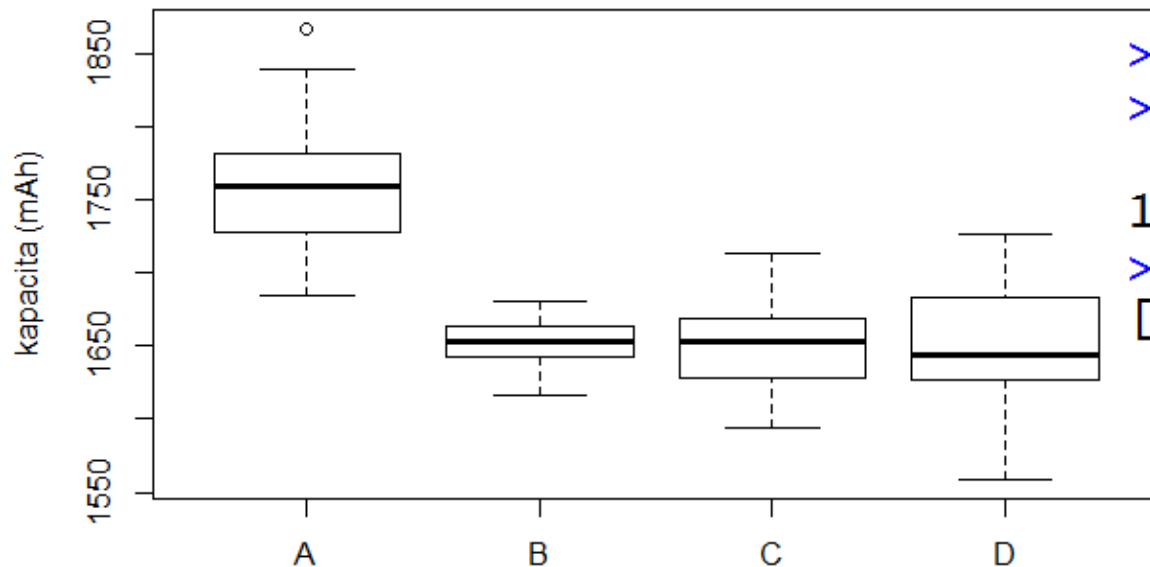
A	C	D
1327.2193	860.8913	1481.5059

```
> max(s2)/min(s2)
[1] 1.720898
```

Na hladině významnosti 5 % nelze zamítnout předpoklad o shodě rozptylů
($p - \text{hodnota} = 0,061$, Bartlettův test).

Ověření shody rozptylů (homoskedasticity) ve více než dvou výběrech

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?



```
> s2 = tapply(data$kap, data$vyrobce, var)
> s2
```

A	B	C	D
1327.2193	231.0182	860.8913	1481.5059

```
> max(s2)/min(s2)
```

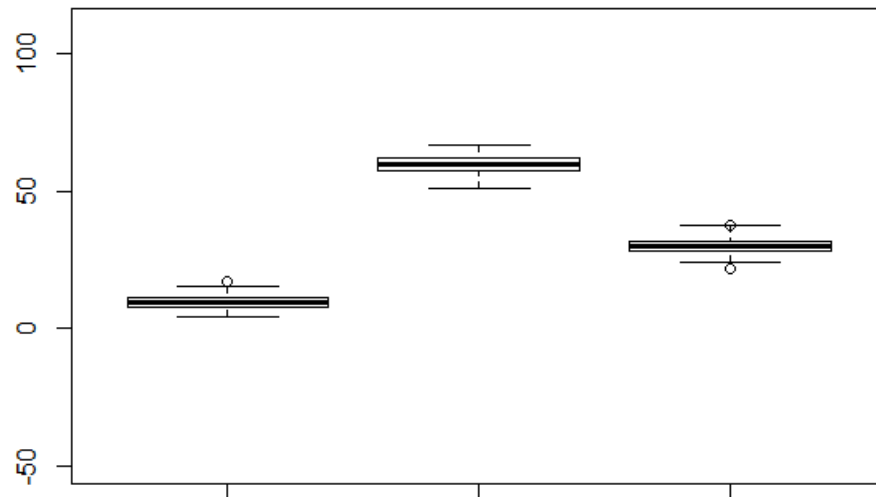
```
[1] 6.412941
```

$\frac{s_{\max}^2}{s_{\min}^2} > 2 \Rightarrow$ podezření na heteroskedasticitu
(různé rozptyly)

Na hladině významnosti 5 % zamítáme předpoklad o shodě rozptylů
($p - \text{hodnota} < 0,001$, Bartlettův test).

Ověření shody středních hodnot ve více než dvou výběrech

Skupina	1	2	3
Průměr	10	60	30

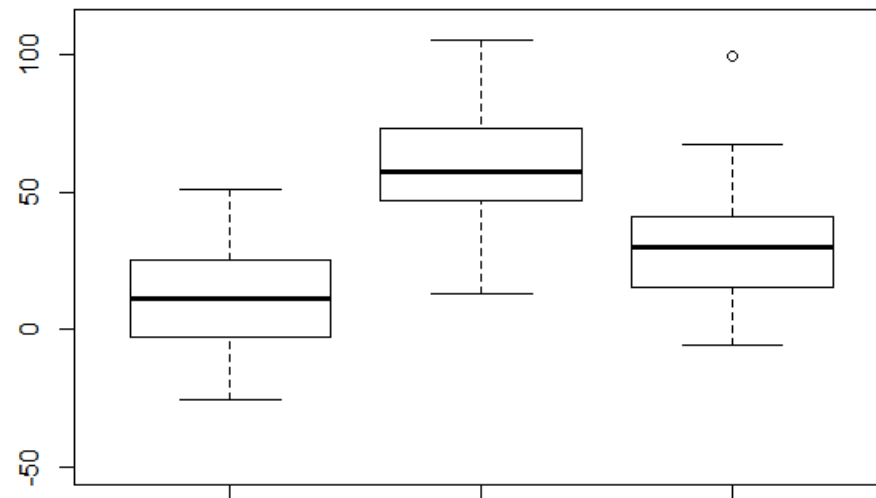
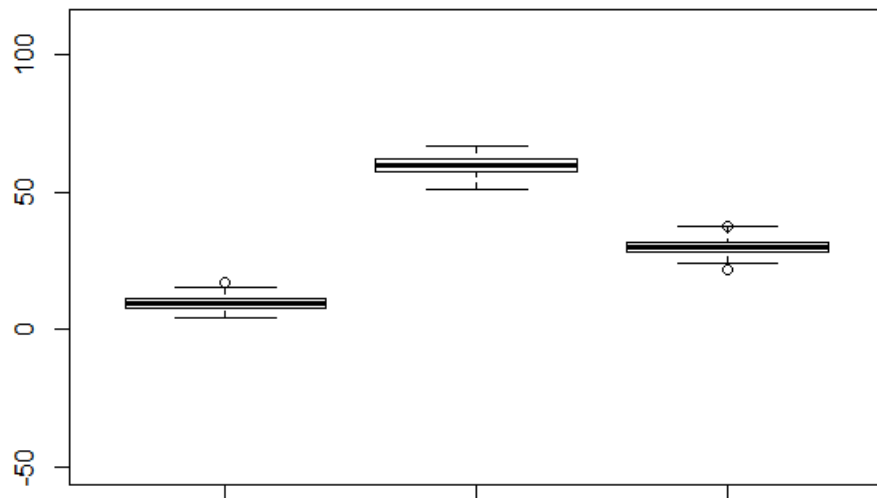


Ověření shody středních hodnot ve více než dvou výběrech

To, zda lze očekávat, že jsou střední hodnoty stejné, záleží nejen na průměrech tříd, ale i na rozptylech uvnitř tříd a rozptylech mezi třídami.

Skupina	1	2	3
Průměr	10	60	30

Skupina	1	2	3
Průměr	10	60	30



Ověření shody středních hodnot ve více než dvou výběrech

ANOVA – analýza rozptylu, tj. test o shodě středních hodnot

Předpoklady testu ANOVA:

- nezávislost výběrů,
- normalita rozdělení (POZOR - nutno ověřit pro každý výběr zvlášť!!!),
- homoskedasticita (tj. identické rozptyly, pro ověření lze použít např. Bartlettův nebo Leveneův test)

Poznámka: ANOVA byla původně navržena pro vyvážené třídění ($n_1 = n_2 = \dots = n_k$). To není předpokladem testu, ale čím těsněji je toto splněno, tím věrohodnější jsou výsledky testu.

Ověření shody středních hodnot ve více než dvou výběrech

ANOVA – analýza rozptylu, tj. test o shodě středních hodnot

- dílčí i celkové výsledky testu ANOVA se zapisují do tabulky,

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Rozptyl (prům. součet čtverců)	F – poměr	p – hodnota
Model	$SS_B = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2$	$df_B = k - 1$	$MS_B = \frac{SS_B}{df_B}$	$\frac{MS_B}{MS_e}$	$1 - F_0(x_{OBS})$
Reziduální	$SS_e = \sum_{i=1}^k (n_i - 1) s_i^2$	$df_e = n - k$	$MS_e = \frac{SS_e}{df_e}$	---	---
Celkový	$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$	$df_T = n - 1$	---	---	---

kde $F_0(x)$ je distribuční funkce Fisherova-Snedecorova rozdělení s df_B stupni volnosti v čitateli a df_e stupni volnosti ve jmenovateli.

Ověření shody středních hodnot ve více než dvou výběrech

Post hoc analýza aneb vícenásobné porovnávání

Zamítneme-li u testu ANOVA nulovou hypotézu, měli bychom zjistit, co bylo příčinou tohoto zamítnutí, tj. najít případné homogenní podskupiny mezi srovnávanými třídami.

Tj. chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti. Použijeme např. **Tukeyovu metodu**.

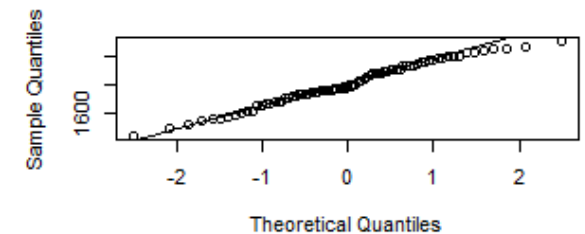
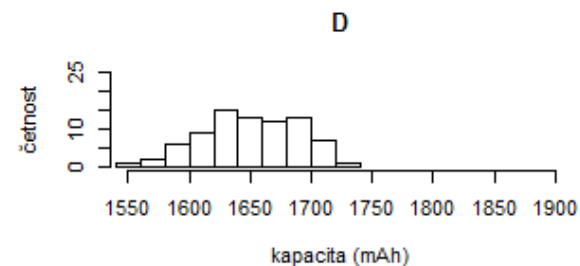
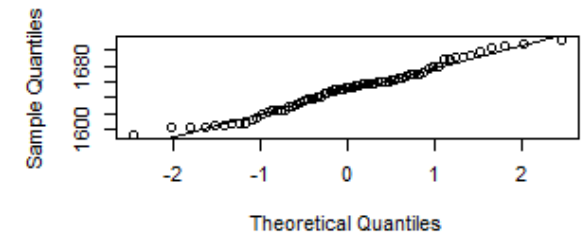
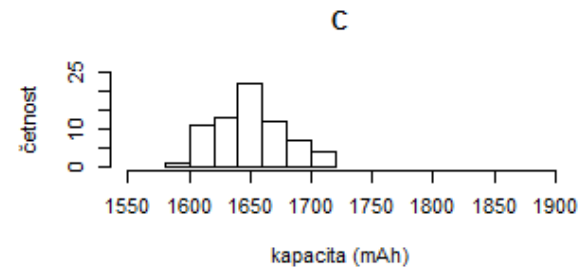
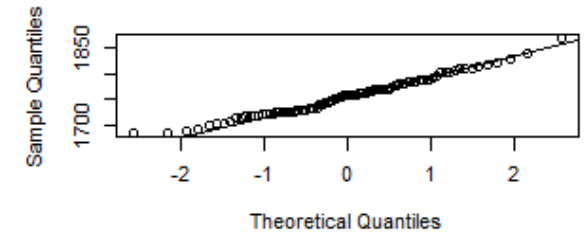
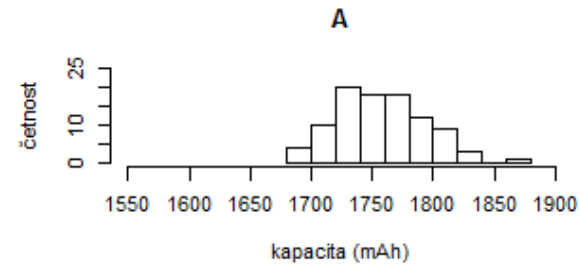
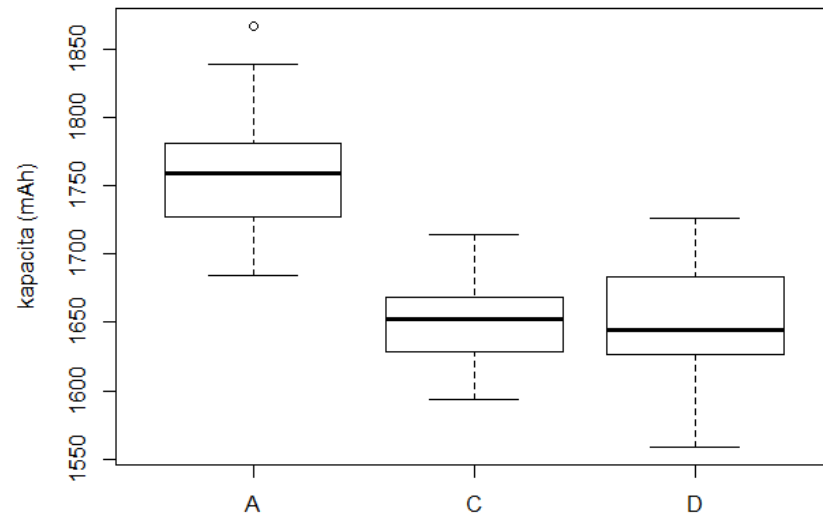
V souboru aku-new.csv jsou uvedeny naměřené kapacity baterií čtyř výrobců po 100 nabíjecích cyklech. Na 5% hladině významnosti ověřte, zda se kapacity akumulátorů výrobců A, C a D statisticky významně liší.

Řešení:

- **Preprocessing dat**: identifikace odlehlých pozorování a jejich zpracování
- **Explorační analýza**: vizualizace dat (vícenásobný boxplot, histogramy), výpočet výběrových charakteristik, posouzení normality (QQ grafy), posouzení shody rozptylů
- **Statistická indukce** (nezapomínat na volbu hladiny významnosti a ověření předpokladů):
 - Ověření hypotézy, zda se průměrné poklesy kapacit baterií výrobců A, C a D statisticky významně liší.

V souboru aku-new.csv jsou uvedeny naměřené kapacity baterií čtyř výrobců po 100 nabíjecích cyklech. Na 5% hladině významnosti ověřte, zda se kapacity akumulátorů výrobců A, C a D statisticky významně liší.

Explorační analýza:



- **Data jsou nezávislá** (každý údaj byl změřen na jiné statistické jednotce).
- **Ověření normality** (exaktní):

H_0 : Data jsou výběrem z normálního rozdělení.

H_A : Data nejsou výběrem z normálního rozdělení. ($\neg H_0$)

výrobce	p-hodnota (Shapirův – Wilkův test)
A	0,330
C	0,197
D	0,271

Na hladině významnosti 5% nezamítáme předpoklad normality (viz tabulka).

- Ověření shody rozptylů

$$H_0: \sigma_A^2 = \sigma_C^2 = \sigma_D^2$$

$$H_A: \neg H_0$$

Na hladině významnosti 5% nelze zamítnout hypotézu o shodě rozptylů ($p - hodnota = 0,061$, Bartlettův test).

Předpoklady pro použití ANOVy nelze zamítnout.

- ANOVA

$$H_0: \mu_A = \mu_C = \mu_D$$

$$H_A: \neg H_0$$

```

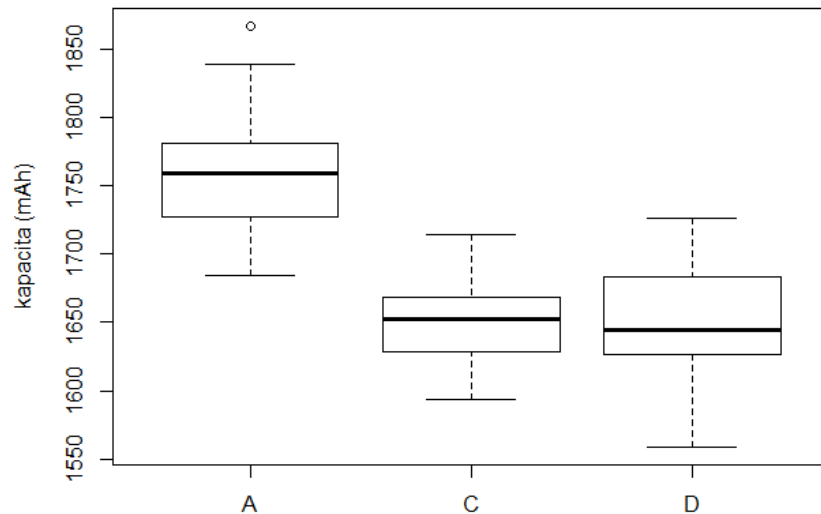
              Df Sum Sq Mean Sq F value Pr(>F)
data.acd$vyrobce    2  653570   326785    262.8 <2e-16 ***
Residuals          241  299718     1244
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Na hladině významnosti 5% zamítáme hypotézu o shodě průměrných kapacit akumulátorů výrobců A, C a D ($p - hodnota \ll 0,001$, ANOVA).

Post hoc analýza aneb vícenásobné porovnávání

Zamítneme-li u testu ANOVA nulovou hypotézu, měli bychom zjistit, co bylo příčinou tohoto zamítnutí, tj. najít případné homogenní podskupiny mezi srovnávanými třídami.



```
$`data.acd$vyrobce`  
      diff      lwr      upr      p adj  
C-A -106.2463158 -119.34644 -93.14619 0.0000000  
D-A -106.0488474 -118.71205 -93.38565 0.0000000  
D-C   0.1974684  -13.45385  13.84879 0.9993586
```

Post hoc analýza Tukeyho metodou

Průměrné kapacity akumulátorů výrobce A se statisticky významně liší od průměrných kapacit akumulátorů výrobců C i D. Průměrné kapacity akumulátorů výrobců C a D se statisticky významně neliší.

Ověření shody mediánů ve více než dvou výběrech

tj. **Kruskalův – Wallisův test**

Předpoklady Kruskalova – Wallisova testu:

- nezávislost výběrů,
- stejný typ rozdělení měřené veličiny pro všechny třídy

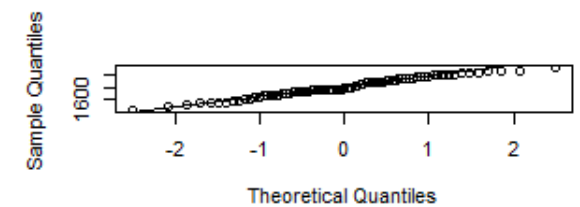
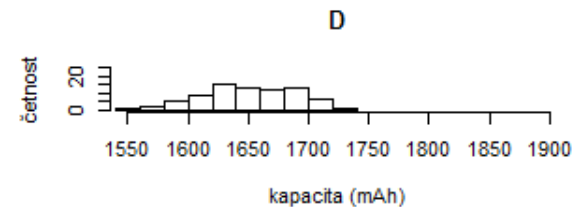
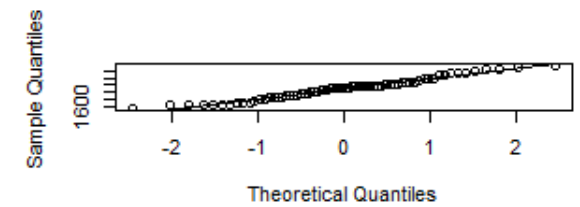
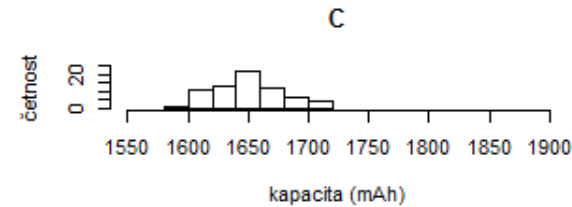
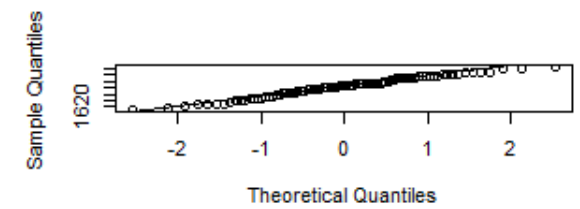
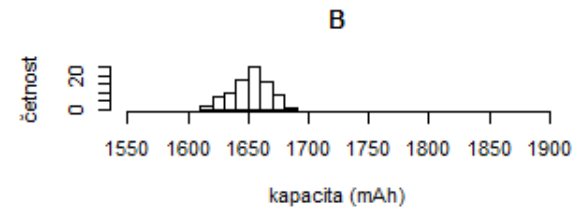
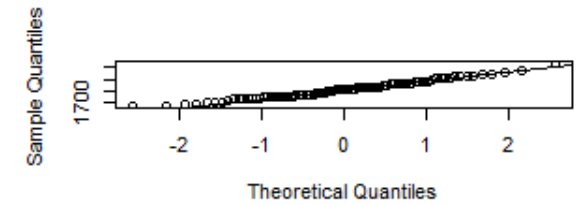
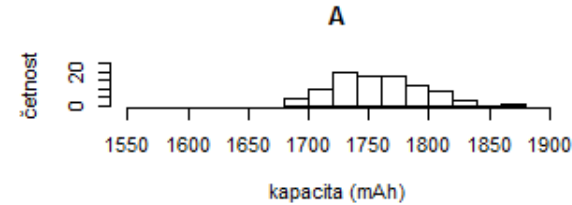
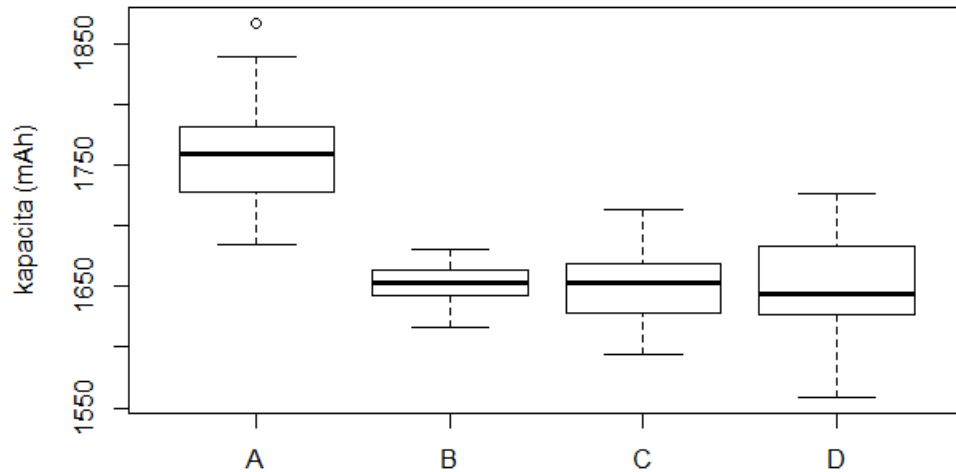
V souboru aku-new.csv jsou uvedeny naměřené kapacity baterií čtyř výrobců po 100 nabíjecích cyklech. Na 5% hladině významnosti ověřte, zda se kapacity akumulátorů výrobců A, B, C a D statisticky významně liší.

Řešení:

- **Preprocessing dat:** identifikace odlehlých pozorování a jejich zpracování
- **Explorační analýza:** vizualizace dat (vícenásobný boxplot, histogramy), výpočet výběrových charakteristik, posouzení normality (QQ grafy), posouzení shody rozptylů
- **Statistická indukce** (nezapomínat na volbu hladiny významnosti a ověření předpokladů):
 - Ověření hypotézy, zda se průměrné poklesy kapacit baterií výrobců A, B, C a D statisticky významně liší.

V souboru aku-new.csv jsou uvedeny naměřené kapacity baterií čtyř výrobců po 100 nabíjecích cyklech. Na 5% hladině významnosti ověřte, zda se kapacity akumulátorů výrobců A, B, C a D statisticky významně liší.

Explorační analýza:



- **Data jsou nezávislá** (každý údaj byl změřen na jiné statistické jednotce).
- **Ověření normality** (exaktní):

H_0 : Data jsou výběrem z normálního rozdělení.

H_A : Data nejsou výběrem z normálního rozdělení. ($\neg H_0$)

výrobce	p-hodnota (Shapirův – Wilkův test)
A	0,330
B	0,198
C	0,197
D	0,271

Na hladině významnosti 5% nezamítáme předpoklad normality (viz tabulka).

- **Ověření shody rozptylů**

$$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$$

$$H_A: \neg H_0$$

Na hladině významnosti 5 % zamítáme hypotézu o shodě rozptylů ($p - \text{hodnota} < 0,001$, Bartlettův test).

Předpoklady pro použití ANOVy nejsou splněny (předpoklad shody rozptylů), proto pro testování volíme Kruskalův – Wallisův test (jeho předpoklady splněny jsou).

- **Kruskalův – Wallisův test**

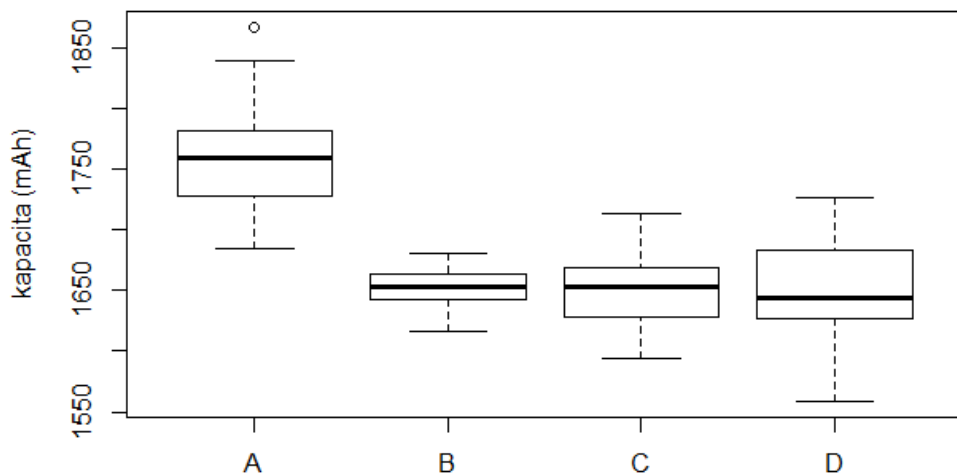
$$H_0: x_{0,5}^A = x_{0,5}^B = x_{0,5}^C = x_{0,5}^D$$

$$H_A: \neg H_0$$

Na hladině významnosti 5 % zamítáme hypotézu o shodě mediánů kapacit akumulátorů výrobců A, B, C a D ($p - \text{hodnota} \ll 0,001$, Kruskalův – Wallisův test).

Post hoc analýza aneb vícenásobné porovnávání

Zamítneme-li u Kruskalova – Wallisova testu nulovou hypotézu, měli bychom zjistit, co bylo příčinou tohoto zamítnutí, tj. najít případné homogenní podskupiny mezi srovnávanými třídami.



Col	Mean-			
Row	Mean	A	B	C
B		11.51268 0.0000		
C		10.93701 0.0000	0.152889 0.4392	
D		11.17325 0.0000	0.019004 0.4924	-0.130902 0.4479

Post hoc analýza Dunnové metodou

Mediány kapacit akumulátorů výrobce A se statisticky významně liší od mediánů kapacit akumulátorů výrobců B, C i D. Mediány kapacit akumulátorů výrobců B, C a D se statisticky významně neliší.

Analýza dat v kontingenčních tabulkách

Část 2 – Analýza závislosti v kontingenčních tabulkách

- Analýza závislosti v kontingenčních tabulkách
- Analýza závislosti v asociačních tabulkách

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnejte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

kap	vyrobce	ind
1654.2	B	ANO
1663.1	B	ANO
1633.3	B	NE
1642.2	B	NE
1656.7	B	ANO
1664.4	B	ANO
1649.8	B	NE

standardní datový formát

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE
B	51	38
D	36	43

tabulka sdružených četností,
tj. kontingenční tabulka

Jak zapsat výsledky měření?

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51	38	89
D	36	43	79
celkem	87	81	168

rozšířená kontingenční tabulka (kont. tabulka rozšířená o **marginální četnosti**)

Jak zapsat výsledky měření?

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (30 %)	38 (23 %)	89
D	36 (21 %)	43 (26 %)	79
celkem	87	81	168

rozšířená kontingenční tabulka (v závorce jsou uvedeny **sdružené rel. četnosti**)

Jak zapsat výsledky měření?

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (59 %)	38 (47 %)	89 (53 %)
D	36 (41 %)	43 (53 %)	79 (47 %)
celkem	87	81	168

rozšířená kontingenční tabulka (v závorce jsou uvedeny **sloupcové rel. četnosti**)

Jak zapsat výsledky měření?

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (57 %)	38 (43 %)	89
D	36 (46 %)	43 (54 %)	79
celkem	87 (52 %)	81 (48 %)	168

rozšířená kontingenční tabulka (v závorce jsou uvedeny **řádkové rel. četnosti**)

Jak zapsat výsledky měření?

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

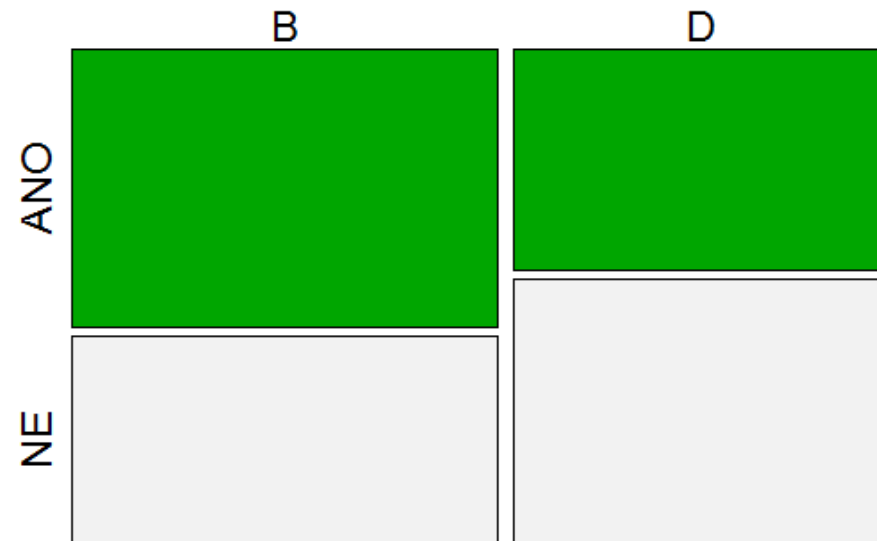
Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (57 %)	38 (43 %)	89
D	36 (46 %)	43 (54 %)	79
celkem	87 (52 %)	81 (48 %)	168

rozšířená kontingenční tabulka (v závorce jsou uvedeny **řádkové rel. četnosti**)

Jak graficky prezentovat sledovanou závislost?

Motivační příklad:

U každého z akumulátorů zjistíte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

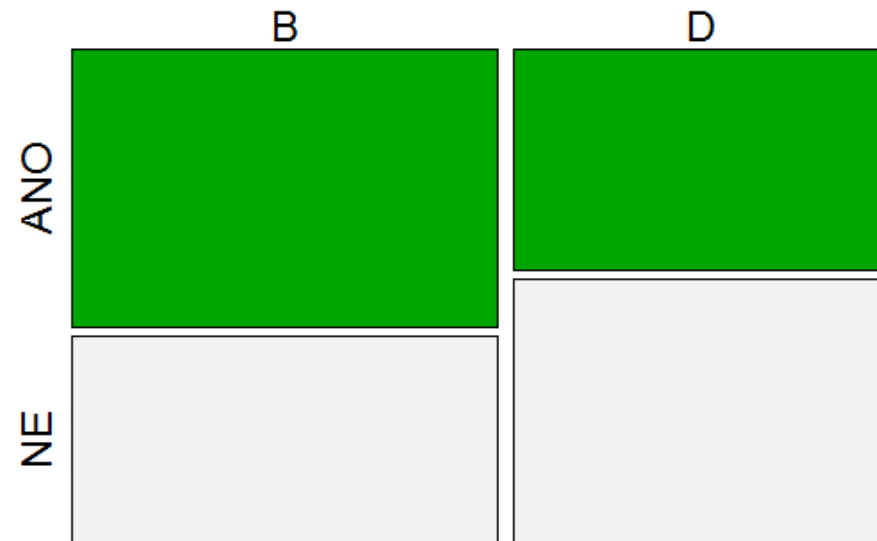


Mozaikový graf

Jak graficky prezentovat sledovanou závislost?

Motivační příklad:

U každého z akumulátorů zjistíte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)



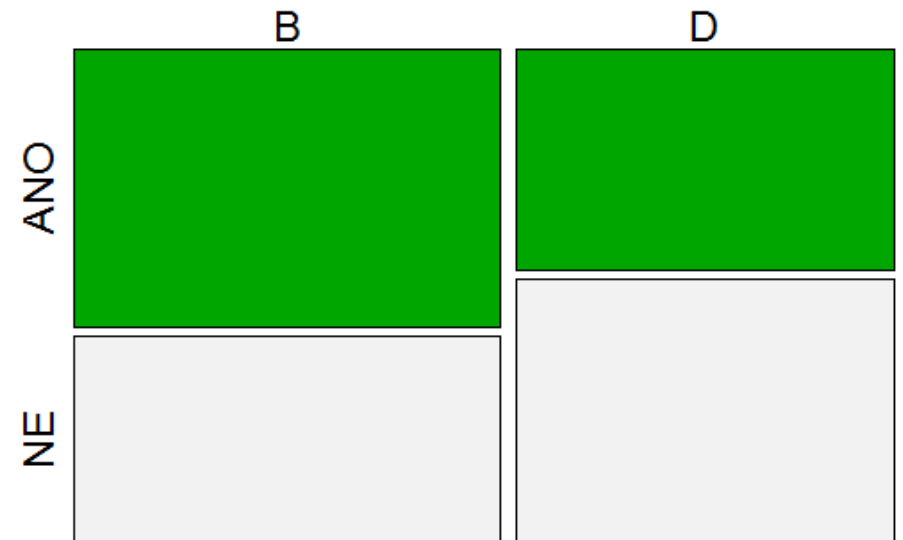
Mozaikový graf

Čím je mozaikový graf členitější, tím silnější závislost mezi proměnnými pozorujeme.

Motivační příklad:

U každého z akumulátorů zjistíte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (57 %)	38 (43 %)	89
D	36 (46 %)	43 (54 %)	79
celkem	87 (52 %)	81 (48 %)	168



Míry kontingence, např. Cramerovo V

Cramerovo V nabývá hodnoty mezi 0 (nezávislé proměnné) a 1 (úplná kontingence).

Cramerovo V = 0,105 (vzorec bude uveden později)

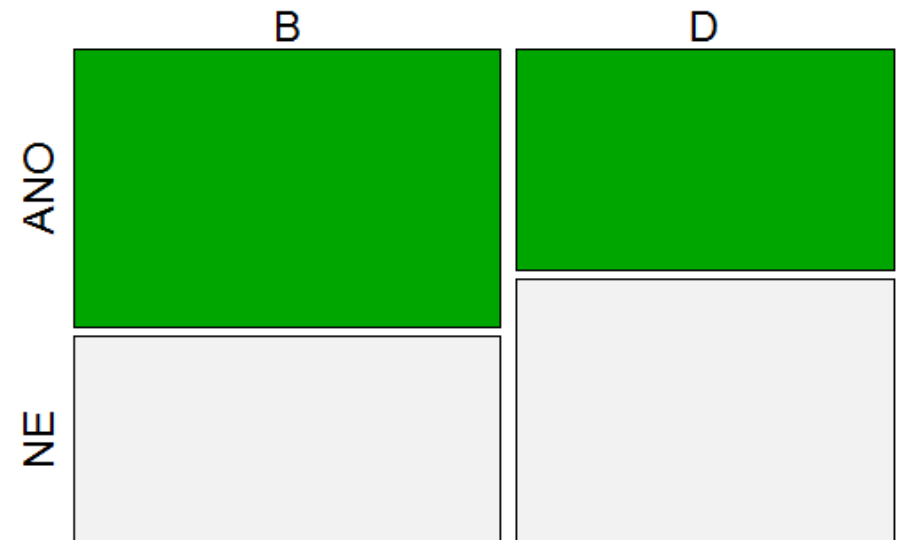
Jak číselně ohodnotit míru kontingence („intenzitu závislosti“)?

Motivační příklad:

U každého z akumulátorů zjistíte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (57 %)	38 (43 %)	89
D	36 (46 %)	43 (54 %)	79
celkem	87 (52 %)	81 (48 %)	168

$$\text{Cramerovo } V = 0,105$$



Chí – kvadrát test nezávislosti v kontingenční tabulce

Jak exaktně ověřit, zda existuje stat. významná souvislost mezi proměnnými?

Chí – kvadrát test nezávislosti

H_0 : Znaky X a Y v kontingenční tabulce **jsou** statisticky **nezávislé**.

H_A : Znaky X a Y v kontingenční tabulce nejsou statisticky nezávislé.

Předpoklady testu:

- všechny očekávané četnosti E_{ij} musí být větší nebo rovny 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

Co jsou to očekávané četnosti?

Očekávané četnosti

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	Celkem
B	51	38	89
D	36	43	79
Celkem	87	81	168

Pozorované (angl. observed) četnosti

O_{ij}

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE
B	$\frac{87 \cdot 89}{168}$	$\frac{81 \cdot 89}{168}$
D	$\frac{87 \cdot 79}{168}$	$\frac{81 \cdot 79}{168}$

Očekávané (angl. expected) četnosti

E_{ij}

Očekávané četnosti E_{ij} ... sdružené četnosti očekávané za předpokladu platnosti H_0

$$E_{ij} = \frac{n_{.j} \cdot n_{i.}}{n},$$

kde $n_{.j}$, $n_{i.}$ jsou příslušné marginální četnosti a n je rozsah výběru

Co jsou to očekávané četnosti?

Očekávané četnosti

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	Celkem
B	51	38	89
D	36	43	79
Celkem	87	81	168

Pozorované (angl. observed) četnosti

O_{ij}

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE
B	46,1	42,9
D	40,9	38,1

Očekávané (angl. expected) četnosti

E_{ij}

Očekávané četnosti E_{ij} ... sdružené četnosti očekávané za předpokladu platnosti H_0

$$E_{ij} = \frac{n_{.j} \cdot n_{i.}}{n},$$

kde $n_{.j}$, $n_{i.}$ jsou příslušné marginální četnosti a n je rozsah výběru

Co jsou to očekávané četnosti?

Chí – kvadrát test nezávislosti

H_0 : Znaky X a Y v kontingenční tabulce **jsou** statisticky **nezávislé**.

H_A : Znaky X a Y v kontingenční tabulce nejsou statisticky nezávislé.

Předpoklady testu:

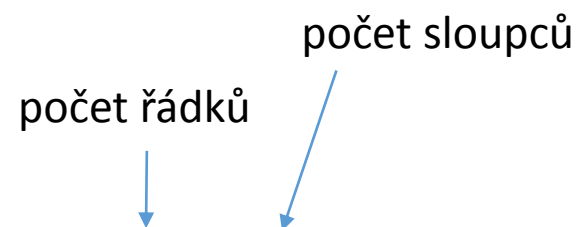
- všechny očekávané četnosti E_{ij} musí být větší nebo rovny 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

Testové kritérium: $x_{OBS} = K = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

p – hodnota: $p\text{-hodnota} = 1 - F_0(x_{OBS}) = P(X > x_{OBS})$,

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $(r - 1)(s - 1)$ stupni volnosti.

počet řádků
počet sloupců

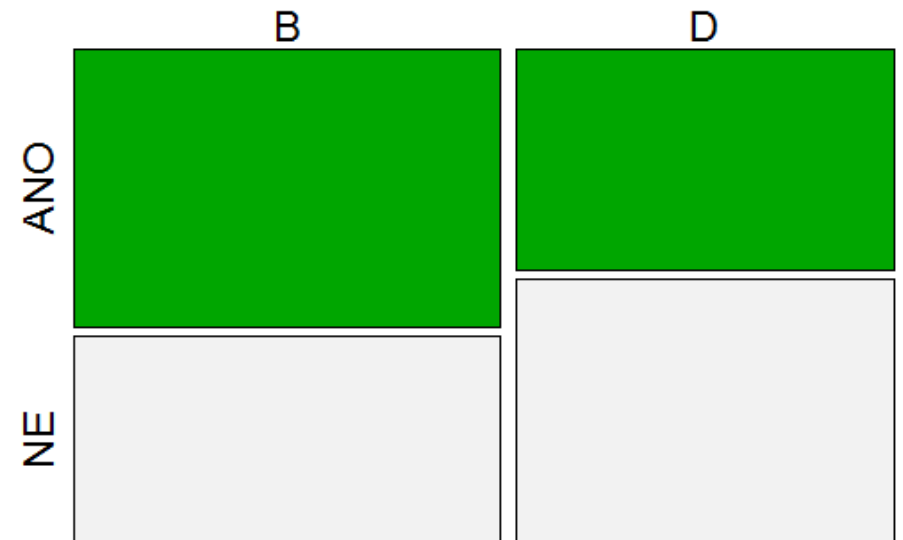


Motivační příklad:

U každého z akumulátorů zjistíte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE	celkem
B	51 (57 %)	38 (43 %)	89
D	36 (46 %)	43 (54 %)	79
celkem	87 (52 %)	81 (48 %)	168

$$\text{Cramerovo } V = 0,105$$



Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

H_0 : Mezi kvalitou akumulátorů (danou daným kritériem) a výrobcem (B, D) **není** statisticky významná **souvislost**.

H_A : Mezi kvalitou akumulátorů (danou daným kritériem) a výrobcem (B, D) je statisticky významná souvislost.

Předpoklady testu:

- všechny očekávané četnosti E_{ij} musí být větší nebo rovny 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

Výrobce \ Kapacita vyšší než 1650 mAh	ANO	NE
B	46,1	42,9
D	40,9	38,1

Předpoklady testu jsou splněny.

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

H_0 : Mezi kvalitou akumulátorů (danou daným kritériem) a výrobcem (B, D) **není** statisticky významná **souvislost**.

H_A : Mezi kvalitou akumulátorů (danou daným kritériem) a výrobcem (B, D) je statisticky významná souvislost.

Předpoklady testu:

- všechny očekávané četnosti E_{ij} musí být větší nebo rovny 2,
- alespoň 80% očekávaných četností E_{ij} musí být větších než 5.

Předpoklady testu jsou splněny.

Rozhodnutí:

Na hladině významnosti 0,05 nelze zamítnout H_0 ($p - hodnota = 0,172$, $df = 1$, Chí – kvadrát test nezávislosti), tj. mezi kvalitou akumulátoru (vzhledem k danému kritériu a výrobcem (A, B) není statisticky významná souvislost).

Cramerovo V

Označme:

- r ... počet variant proměnné X ,
- s ... počet variant proměnné Y ,
- $K = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, kde O_{ij} jsou pozorované sdružené četnosti zapsané v kontingenční tabulce a E_{ij} jsou očekávané četnosti odpovídající součinu příslušných marginálních relativních četností.

Cramerovo V

$$V = \sqrt{\frac{K}{n(\min(r;s)-1)}}$$

Motivační příklad:

U každého z akumulátorů zjistěte, zda je jeho kapacita větší než 1650 mAh. (Definujte si novou dichotomickou proměnnou, která bude nabývat hodnot {ANO, NE}.) Následně ověřte, zda kvalita akumulátoru (daná kapacitou vyšší než 1650 mAh) statisticky významně souvisí s výrobcem akumulátorů. Srovnajte pouze akumulátory výrobce B a D. (viz http://am-nas.vsb.cz/lit40/DATA/aku_new.csv)

Na hladině významnosti 0,05 nelze zamítnout H_0 ($p - hodnota = 0,172$, $df = 1$, Chí – kvadrát test nezávislosti), tj. mezi kvalitou akumulátoru (vzhledem k danému kritériu a výrobcem (A, B) není statisticky významná souvislost.

Pozorovanou závislost lze označit jako slabou (*Cramerovo* $V = 0,105$).

Analýza závislosti v asociační tabulce

Asociační tabulky

- speciální typ kontingenčních tabulek, které používáme k sledování závislosti dvou dichotomických znaků, tj. kategoriálních znaků nabývajících pouze dvou variant. (*asociace = vztah dvou dichotomických znaků*)

$X \text{ (okolnosti)} \backslash Y \text{ (výskyt události)}$	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a + b$
$x_{[2]}$ (II.)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky

Asociační tabulky

- speciální typ kontingenčních tabulek, které používáme k sledování závislosti dvou dichotomických znaků, tj. kategoriálních znaků nabývajících pouze dvou variant. (*asociace = vztah dvou dichotomických znaků*)

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Asociační tabulky

Na asociační tabulku lze sice nahlížet jako na speciální případ kontingenčních tabulek a při analýze používat jejich aparát, nicméně vhodnější je využít specifické metody a charakteristiky asociace.

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Míry asociace

Poměr šancí (angl. odds ratio)

Pozorovaný poměr počtu úspěchů k počtu neúspěchů (tzv. pozorovaná **šance**) za okolností I. je $\frac{a}{b}$, za okolností II. $\frac{c}{d}$. Odhad poměru šancí je pak

$$\widehat{OR} = \frac{ad}{bc}.$$

X (okolnosti) \ Y (výskyt události)	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a + b$
$x_{[2]}$ (II.)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky

Míry asociace

Poměr šancí (angl. odds ratio)

Pozorovaný poměr počtu nemocných k počtu „zdravých“ (tzv. pozorovaná **šance**) u exponované populace je $\frac{a}{b}$, u neexponované populace $\frac{c}{d}$. Odhad poměru šancí je pak

$$\widehat{OR} = \frac{ad}{bc}.$$

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

- Odhad šance novorozeneckého úmrtí u dětí s nízkou porodní váhou je

$$\frac{a}{b} = \frac{618}{4\,597} = 0,134,$$

což odpovídá přibližně 134 novorozeneckým úmrtím na 1 000 přeživších novorozenců s nízkou porodní váhou.

- Obdobně odhadneme šanci novorozeneckého úmrtí u dětí s normální porodní váhou.

$$\frac{c}{d} = \frac{422}{67\,093} = 0,006$$

Lze očekávat přibližně 6 novorozeneckých úmrtí na 1 000 přeživších novorozenců s normální porodní hmotností.

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

- Odhad šance novorozeneckého úmrtí u dětí s nízkou porodní váhou je

$$\frac{a}{b} = \frac{618}{4\,597} = 0,134.$$

- Odhad šance novorozeneckého úmrtí u dětí s normální porodní váhou je

$$\frac{c}{d} = \frac{422}{67\,093} = 0,006$$

$\widehat{OR} = \frac{ad}{bc} = \frac{618 \cdot 67\,093}{4\,597 \cdot 422} \cong 21,4 \Rightarrow$ šance novorozeneckého úmrtí je 21,4 krát vyšší u novorozenců s nízkou porodní váhou než u novorozenců s normální porodní váhou.

Míry asociace

Poměr šancí (angl. odds ratio)

$$\widehat{OR} = \frac{ad}{bc}.$$

$OR < 1$	U exponované populace (populace vystavené sledovanému faktoru) je nižší šance výskytu nemoci.
$OR = 1$	Šance výskytu onemocnění u exponované a neexponované populace jsou shodné.
$OR > 1$	U exponované populace je vyšší šance výskytu nemoci.

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Míry asociace

Poměr šancí (angl. odds ratio)

$$\widehat{OR} = \frac{ad}{bc}.$$

$OR < 1$	U exponované populace (populace vystavené sledovanému faktoru) je nižší šance výskytu nemoci.
$OR = 1$	Šance výskytu onemocnění u exponované a neexponované populace jsou shodné.
$OR > 1$	U exponované populace je vyšší šance výskytu nemoci.

Je-li $OR \neq 1$, potřebujeme zpravidla ještě rozhodnout, zda je indikována asociace statisticky významná.

Woolfova metoda:

100(1 - α)% intervalový odhad OR : $\langle \widehat{OR} \cdot e^{-\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \cdot Z_{1-\frac{\alpha}{2}}}; \widehat{OR} \cdot e^{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \cdot Z_{1-\frac{\alpha}{2}}} \rangle$.

Jestliže 100(1 - α)% intervalový odhad OR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y .

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

- $\widehat{OR} = \frac{ad}{bc} = \frac{618 \cdot 67\,093}{4\,597 \cdot 422} \cong 21,4 \Rightarrow$ šance novorozeneckého úmrtí je 21,4 krát vyšší u novorozenců s nízkou porodní váhou než u novorozenců s normální porodní váhou.

95% intervalový odhad OR je $\langle 19,2; 23,8 \rangle$. Je zcela zřejmé, že šance novorozeneckého úmrtí závisí na porodní váze ($1 \notin \langle 19,2; 23,8 \rangle$).

Míry asociace

Absolutní riziko (angl. absolute risk) výskytu události (onemocnění, úmrtí, ...) v závislosti na okolnostech (přítomnosti sledovaného faktoru)

- odhad absolutního rizika onemocnění u exponovaných respondentů je $\frac{a}{a+b}$,
- odhad absolutního rizika onemocnění u neexponovaných respondentů je $\frac{c}{c+d}$.

Absolutní rizika mohou nabývat hodnot z intervalu (0; 1).

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Míry asociace

Relativní riziko (angl. relative risk)

- poměr odhadů absolutních rizik vzniku onemocnění u exponovaných a neexponovaných osob, tj.

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)}.$$

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\nnovorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

- Odhad absolutního rizika novorozeneckého úmrtí u dětí s nízkou porodní hmotností je $\frac{a}{a+b} = \frac{618}{5\,215} = 0,119$,

tj. novorozenecké úmrtí lze očekávat u cca 119 z 1 000 novorozenců s nízkou porodní váhou),

- u dětí s normální porodní hmotností je absolutní riziko: $\frac{c}{c+d} = \frac{422}{67\,515} = 0,006$,

tj. novorozenecké úmrtí lze očekávat u cca 6 z 1 000 novorozenců s normální porodní váhou.

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrť	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

- Odhad absolutního rizika novorozeneckého úmrtí u dětí s nízkou porodní hmotností je $\frac{a}{a+b} = \frac{618}{5\,215} = 0,119$,
- u dětí s normální porodní hmotností je absolutní riziko: $\frac{c}{c+d} = \frac{422}{67\,515} = 0,006$,
- Odhad relativního rizika novorozeneckého úmrtí

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)} = \frac{0,119}{0,006} = 19,0.$$

Ve sledovaném období bylo u dětí s nízkou porodní váhou 19 krát vyšší riziko novorozeneckého úmrtí než u dětí s normální porodní váhou.

Míry asociace

Relativní riziko (angl. relative risk)

- poměr odhadů absolutních rizik vzniku onemocnění u exponovaných a neexponovaných osob, tj.

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)}.$$

$RR < 1$	Expozice snižuje riziko onemocnění.
$RR = 1$	Mezi expozicí a onemocněním neexistuje žádná asociace.
$RR > 1$	Expozice zvyšuje riziko onemocnění.

X (sledovaný faktor) \ Y (výskyt onemocnění)	D (ANO)	$\bar{D}$ (NE)	Celkem
E (přítomnost faktoru)	a	b	$a + b$
$\bar{E}$ (nepřítomnost faktoru)	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

Schéma rozšířené asociační tabulky (biomedicínská aplikace)

Míry asociace

Relativní riziko (angl. relative risk)

- poměr odhadů absolutních rizik vzniku onemocnění u exponovaných a neexponovaných osob, tj.

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)}.$$

$RR < 1$	Expozice snižuje riziko onemocnění.
$RR = 1$	Mezi expozicí a onemocněním neexistuje žádná asociace.
$RR > 1$	Expozice zvyšuje riziko onemocnění.

Je-li $RR \neq 1$, musíme rozhodnout, zda je indikována asociace statisticky významná.

Katzova metoda:

100(1 - α)% intervalový odhad RR : $\langle \widehat{RR} \cdot e^{-\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot Z_{1-\frac{\alpha}{2}}}; \widehat{RR} \cdot e^{\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot Z_{1-\frac{\alpha}{2}}} \rangle$.

Jestliže 100(1 - α)% intervalový odhad RR nezahrnuje 1, pak zamítáme hypotézu o nezávislosti znaků X a Y .

Závisí novorozenecká úmrtnost (do 7 dnů po porodu) na porodní váze? Data odpovídající situaci v New Yorku v roce 1974 jsou uvedena v následující tabulce.

porodní váha\novorozenecká úmrtí	ANO	NE	Celkem
nízká	618	4 597	5 215
normální	422	67 093	67 515
Celkem	1 040	71 690	72 730

Odhad relativního rizika novorozeneckého úmrtí $\widehat{RR} = 19,0$

⇒ ve sledovaném období bylo u dětí s nízkou porodní váhou 19 krát vyšší riziko novorozeneckého úmrtí než u dětí s normální porodní váhou.

95% intervalový odhad RR je $\langle 17,1; 21,0 \rangle$. Je zcela zřejmé, že riziko novorozeneckého úmrtí závisí na porodní váze ($1 \notin \langle 17,1; 21,0 \rangle$).

DĚKUJEME ZA
POZORNOST!

