Přednáška X.
Testování hypotéz o kvalitativních proměnných

- Testování hypotéz o podílech
- Kontingenční tabulka, čtyřpolní tabulka
- Testy nezávislosti, Fisherův exaktní test, McNemarův test
- Testy dobré shody pro ověření rozdělení pravděpodobnosti

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Opakování – analýza rozptylu

- Proč je výhodnější provést srovnání průměrů spojité veličiny u více než dvou skupin pomocí analýzy rozptylu než pomocí testů pro všechny dostupné dvojice sledovaných skupin?

- Jak lze řešit situaci, kdy chceme provést více testů zároveň?

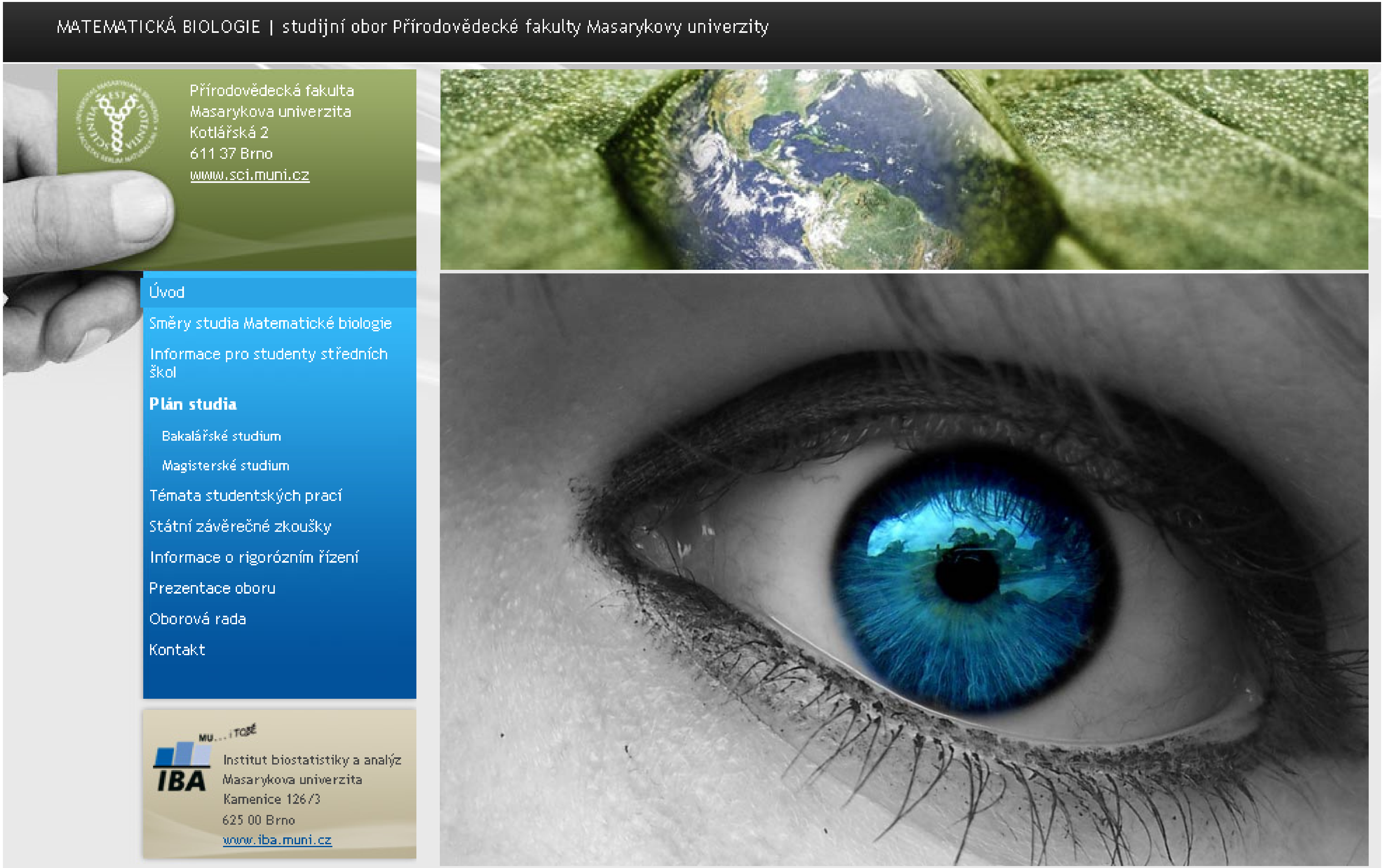
Opakování – princip analýzy rozptylu

- Jaký je princip analýzy rozptylu?

- Jaké jsou předpoklady analýzy rozptylu?

1. Motivace

Matematická biologie × modré oči



Studenti matematické biologie s modrýma očima

- **Budeme sledovat podíl studentů matematické biologie** (současných i bývalých), **kteří mají modré oči**.

- Náhodná veličina *A* = modrá barva očí – alternativní náhodná veličina.

$$A = \begin{cases} 1 & \text{když student má modré oči} \\ 0 & \text{když student nemá modré oči} \end{cases} \qquad \begin{aligned} P(A=1) &= \pi \\ P(A=0) &= 1-\pi \end{aligned}$$

- Náhodná veličina *X* = počet studentů matematické biologie s modrýma očima – binomická náhodná veličina. Je to součet *n* alternativních veličin.

$$X = \sum_{i=1}^{n} A_i \qquad X \sim Bi(n, \pi)$$

- Odhad parametru π: $\hat{\pi} = p = X / n$

Studenti matematické biologie s modrýma očima

- **Budeme sledovat podíl studentů matematické biologie, kteří mají modré oči**.
- Výsledky v tabulce:

	Modrá barva očí	Jiná barva očí	Celkem
Studenti matematické biologie (současní i bývalí)	17	43	60

- Odhad parametru π:

$$\hat{\pi} = p = X / n = 17 / 60 = 0{,}283$$

Studenti matematické biologie s modrýma očima

- **Budeme se zajímat o to, jestli podíl studentů matematické biologie, kteří mají modré oči, souvisí s obdobím studia.**
- Výsledky v tabulce:

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	11	31	42
Bývalí	6	12	18
Celkem	17	43	60

2. Testování hypotéz o podílech

Co nás bude zajímat?

- Binární data jsou v medicíně i biologii častá – výskyt ano/výskyt ne, úspěch/neúspěch, …

- Kromě bodového odhadu nás může zajímat
  - Interval spolehlivosti pro parametr π
  - Test o parametru π proti konstantě π_0
  - Test o parametru π ve dvou souborech

Aproximace na normální rozdělení

- Pravděpodobnost, že náhodná veličina *X* bude při své realizaci rovna hodnotě *k* lze přesně stanovit pomocí vzorce:

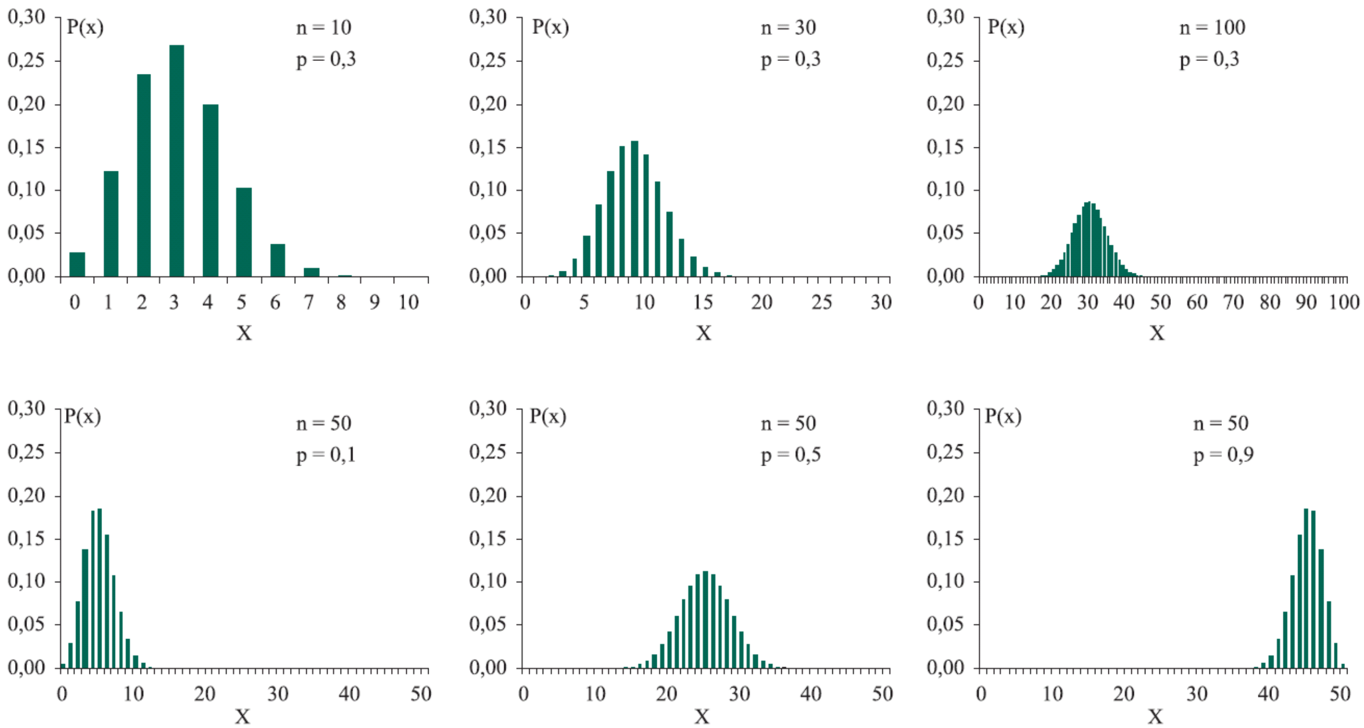
$$P(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

- Pro větší *n* (a tedy větší rozsah možných hodnot *k*) je jednodušší použít aproximaci normálním rozdělením.
- Vychází z CLV – součty se pro dostatečné *n* chovají normálně.
- Předpokladem aproximace na normální rozdělení je součin *np*(1-*p*) větší než 5, nebo ještě lépe součin *np*(1-*p*) větší než 10.
- Pak platí:

$$Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)$$

Proč $np(1-p)$ větší než 5?

- Souvisí s množstvím informace nutné pro dosažení „tvaru normálního rozdělení“ → nutné pro vhodnost, respektive přesnost aproximace.
- Pro π = 0,5 je jednodušší dosáhnout „tvar normálního rozdělení“ než pro π = 0,1 nebo π = 0,9. Pro π hodně blízká 0 nebo 1 není aproximace vhodná.



Interval spolehlivosti pro podíl

- Máme *n* studentů Matematické biologie a mezi nimi *x* s modrýma očima.
- Rozdělení pravděpodobnosti odhadu parametru π: $\hat{\pi} = p = x/n$

$$E(p) = E(x/n) = E(x)/n = n\pi/n = \pi$$

$$D(p) = D(x/n) = D(x)/n^2 = n\pi(1-\pi)/n^2 = \pi(1-\pi)/n$$

- Při konstrukci intervalu spolehlivosti neznáme hodnotu π, proto je logické ji v odhadu rozptylu (a *SE*) nahradit odhadem *p*:

$$SE(p) = \sqrt{D(p)} = \sqrt{p(1-p)/n}$$

- Při splnění podmínek pro aproximaci normálním rozdělením má 100(1-α)% IS tvar:

$$p \pm z_{1-\alpha/2} SE(p) = p \pm z_{1-\alpha/2} \sqrt{p(1-p)/n}$$

Příklad s modrýma očima

- Máme 60 studentů Matematické biologie a mezi nimi 17 s modrýma očima.

	Modrá barva očí	Jiná barva očí	Celkem
Studenti matematické biologie (současní i bývalí)	17	43	60

- Odhad parametru π: $\hat{\pi} = p = X / n = 17 / 60 = 0{,}283$
- Chceme sestrojit 95% IS pro parametr π.
- Splnění podmínky pro aproximaci normálním rozdělením:

$$np(1-p) = 60 * 0{,}283 * (1-0{,}283) = 12{,}2$$

- Pak $SE(p) = \sqrt{D(p)} = \sqrt{p(1-p)/n} = \sqrt{0{,}283(1-0{,}283)/60} = 0{,}058$

$$95\%\ \mathrm{IS}:\ p \pm z_{1-\alpha/2} SE(p) = 0{,}283 \pm 1{,}96 * 0{,}058 = (0{,}169; 0{,}397)$$

Test pro podíl u jednoho výběru

- Chceme testovat rovnost odhadu parametru π získaného na náhodném výběru *n* jedinců předem dané hodnotě π_0: $H_0 : \pi = \pi_0$
- Při splnění podmínek pro aproximaci normálním rozdělením víme, že platí:

$$Z = \frac{p - \pi}{SE(p)} = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$$

- To za platnosti H_0 znamená:

$$Z = \frac{p - \pi_0}{SE(p)} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,1)$$

- Vypočteme hodnotu testové statistiky a nulovou hypotézu zamítáme podle toho, jakou máme alternativu a hladinu významnosti α.
- Pro alternativu $H_1 : \pi \neq \pi_0$ zamítáme H_0 když $|Z| > z_{1-\alpha/2}$

Příklad s modrýma očima

- Chceme testovat na hladině významnosti α=0,05 rovnost odhadu parametru π získaného na výběru 60 matematických biologů předem dané hodnotě π_0=0,40:

$$H_0 : \pi = 0{,}4$$

- Splnění podmínky pro aproximaci normálním rozdělením máme ověřeno.
- Testová statistika:

$$Z = \frac{p - \pi_0}{SE(p)} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0{,}283 - 0{,}400}{\sqrt{0{,}4(1-0{,}4)/60}} = -1{,}85$$

- Srovnání s kvantilem:

$$|Z| = 1{,}85 < z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$$

⇒ Nezamítáme H_0: π = 0,40.

Je rozdíl mezi IS a testem?

- Pokud ano, v čem?

Je rozdíl mezi IS a testem?

- Ano je…

- Konstrukce IS: $SE(p) = \sqrt{p(1-p)/n}$
- Test H_0: $SE(p) = \sqrt{\pi_0(1-\pi_0)/n}$

- Binomické rozdělení má různou variabilitu pro různé hodnoty π – největší je pro π = 0,5, směrem k 0 a 1 variabilita klesá.

- **Neplatí ekvivalence mezi intervalem spolehlivosti a testem proti π_0 jako tomu bylo v případě průměru jako odhadu střední hodnoty.**

IS pro podíl ve dvou souborech

- Máme *n* studentů Matematické biologie a mezi nimi *x* s modrýma očima, x_1 je současných a x_2 je již vystudovaných. Zajímá nás interval spolehlivosti pro rozdíl podílů studentů s modrýma očima ve skupině současných a již vystudovaných studentů: $\pi_1 - \pi_2$.
- Podmínka pro aproximaci normálním rozdělením musí být splněna v obou výběrech.
- Rozdělení pravděpodobnosti odhadu parametru π v jednotlivých souborech:

$$\hat{\pi}_1 = p_1 = \frac{x_1}{n_1} \qquad \hat{\pi}_2 = p_2 = \frac{x_2}{n_2}$$

$$SE(p_1 - p_2) = \sqrt{D(p_1) + D(p_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Při splnění podmínek pro aproximaci normálním rozdělením má 100(1-α)% IS tvar:

$$p_1 - p_2 \pm z_{1-\alpha/2} SE(p_1 - p_2) = p_1 - p_2 \pm z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Příklad s modrýma očima

- Máme 60 studentů Matematické biologie a mezi nimi 17 s modrýma očima, 11 je současných a 6 je již vystudovaných. Chceme 95% IS pro $\pi_1 - \pi_2$.

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	11	31	42
Bývalí	6	12	18
Celkem	17	43	60

- Splnění podmínek pro aproximaci – zde je to pouze pro ilustraci.
- Odhady: $\hat{\pi}_1 = p_1 = x_1 / n_1 = 11/42 = 0{,}262 \qquad \hat{\pi}_2 = p_2 = x_2 / n_2 = 6/18 = 0{,}333$

$$SE(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0{,}262(1-0{,}262)}{42} + \frac{0{,}333(1-0{,}333)}{18}} = 0{,}130$$

- 95% IS pro $\pi_1 - \pi_2$:

$$p_1 - p_2 \pm z_{1-\alpha/2} SE(p_1 - p_2) = -0{,}071 \pm 1{,}96 * 0{,}130 = (-0{,}326; 0{,}184)$$

Test pro podíl ve dvou výběrech

- Chceme testovat rovnost odhadu parametru π získaného na dvou náhodných výběrech n_1 a n_2 jedinců: $H_0: \pi_1 = \pi_2 = \pi$

- Nejlepším odhadem parametru π je za platnosti H_0: $\hat{\pi} = p = \dfrac{x_1 + x_2}{n_1 + n_2}$

- Odhady pro jednotlivé výběry: $\hat{\pi}_1 = p_1 = x_1 / n_1$ $\qquad \hat{\pi}_2 = p_2 = x_2 / n_2$
- Při splnění podmínky pro aproximaci normálním rozdělením (musí být splněna v obou souborech zároveň) víme, že platí:

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} \sim N(0,1)$$

kde $SE(p_1 - p_2) = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$

- Pro alternativu $H_1: \pi_1 \neq \pi_2$ zamítáme H_0 když $|Z| > z_{1-\alpha/2}$

Příklad s modrýma očima

- Máme 60 studentů Matematické biologie a mezi nimi 17 s modrýma očima, 11 je současných a 6 je již vystudovaných. Testujeme $H_0 : \pi_1 = \pi_2 = \pi$

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	11	31	42
Bývalí	6	12	18
Celkem	17	43	60

- Odhady: $\hat{\pi} = p = 0{,}283$ $\hat{\pi}_1 = p_1 = 0{,}262$ $\hat{\pi}_2 = p_2 = 0{,}333$

$$SE(p_1 - p_2) = \sqrt{p(1-p)(\tfrac{1}{n_1} + \tfrac{1}{n_2})} = \sqrt{0{,}283(1-0{,}283)(\tfrac{1}{42} + \tfrac{1}{18})} = 0{,}127$$

- Testová statistika:

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{0{,}262 - 0{,}333}{0{,}127} = -0{,}56$$

$$|Z| = 0{,}56 < z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$$ ⇒ Nezamítáme H_0.

3. Analýza kontingenčních tabulek

Kontingenční tabulka

- Frekvenční sumarizace dvou nominálních nebo ordinálních veličin pomocí tabulky.
- Proměnné reprezentujeme diskrétními náhodnými veličinami *X* a *Y*.
- Speciální případ: **2 × 2 tabulka** = čtyřpolní tabulka.
- **Př.**: Sumarizace pacientů diagnostikovaných s melanomem dle lokalizace onemocnění a roku diagnózy.

Období	**Lokalizace**				**Celkem**
	Horní končetina	Dolní končetina	Trup	Hlava a krk	
1994-2000	50	103	116	7	276
2001-2005	106	157	310	54	627
2006-2009	115	142	316	52	625
Celkem	271	402	742	113	**1528**

Kontingenční tabulka - hypotézy

- Kontingenční tabulky umožňují testování různých hypotéz:

Nezávislost (Pearsonův chí-kvadrát test)

- Jeden výběr, dvě charakteristiky – obdoba nepárového uspořádání
- Př.: studenti matematické biologie – modré oči × období studia

Shoda struktury (Pearsonův chí-kvadrát test)

- Více výběrů, jedna charakteristika – obdoba nepárového uspořádání
- Př.: pacienti s IM v několika nemocnicích × věková struktura

Symetrie (McNemarův test)

- Jeden výběr, opakovaně jedna charakteristika – obdoba párového uspořádání
- Př.: stromy – posouzení jejich stavu ve dvou sezónách

Značení

- Proměnné reprezentujeme diskrétními náhodnými veličinami *X* a *Y*.

- Označme n_{ij} počet subjektů, pro které platí, že *X=i* a *Y=j* (*i* = 1, ..., *r; j* = 1, ..., *c*).
- Marginální četnosti: $n_{i.} = \sum_{j=1}^{c} n_{ij}$ $\quad n_{.j} = \sum_{i=1}^{r} n_{ij}$
- Celkový počet subjektů: $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$

- Relativní četnosti lze vztahovat:
  - Vzhledem k celkovému *n* $\quad p_{ij} = n_{ij} / n$
  - Vzhledem k řádkovým součtům $n_{i.}$ $\quad p_{ij}^{r} = n_{ij} / n_{i.}$
  - Vzhledem k sloupcovým součtům $n_{.j}$ $\quad p_{ij}^{c} = n_{ij} / n_{.j}$

Pointa testu pro kontingenční tabulku

- Celkem 17 studentů s modrýma očima = 28,3 %. Pokud modré oči nesouvisí s obdobím studia, mělo by stejné zastoupení modrookých platit i v rámci skupin → očekávaná četnost za platnosti H_0 o nezávislosti: $e_{ij} = n_{i.}n_{.j} / n$

- Ekvivalentně lze nezávislost vyjádřit následovně: $p_{ij} = p_{i.}p_{.j}$
- Z toho plyne:

$$e_{ij} = np_{i.}p_{.j} = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

- Očekávané četnosti v příkladu s modrýma očima:

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	11,9	30,1	42
Bývalí	5,1	12,9	18
Celkem	17	43	60

Příklad – melanomy

Období = veličina X	Lokalizace = veličina Y Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	Celkem
1994-2000 $X = 1$	$50 = n_{11}$	$103 = n_{12}$	$116 = n_{13}$	$7 = n_{14}$	$276 = n_{1.}$
2001-2005 $X = 2$	$106 = n_{21}$	$157 = n_{22}$	$310 = n_{23}$	$54 = n_{24}$	$627 = n_{2.}$
2006-2009 $X = 3$	$115 = n_{31}$	$142 = n_{32}$	$316 = n_{33}$	$52 = n_{34}$	$625 = n_{3.}$
Celkem	$271 = n_{.1}$	$402 = n_{.2}$	$742 = n_{.3}$	$113 = n_{.4}$	**1528** $= n$

Období = veličina X	Lokalizace = veličina Y Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	Celkem
1994-2000 $X = 1$	18.12 %	37.32 %	42.03 %	2.54 %	100 %
2001-2005 $X = 2$	16.91 %	25.04 %	49.44 %	8.61 %	100 %
2006-2009 $X = 3$	18.40 %	22.72 %	50.56 %	8.32 %	100 %
Celkem	17.74 %	26.31 %	48.56 %	7.40 %	100 %

Pearsonův chí-kvadrát test nezávislosti

- **Založen na myšlence srovnání pozorovaných a očekávaných četností jednotlivých hodnot, kterých nabývá náhodná veličina *X*.**
- Pozorované četnosti jednotlivých variant *X*=*i* a *Y*=*j* nám vyjadřují n_{ij}.
- Za platnosti nulové hypotézy lze očekávané četnosti jednotlivých variant *X*=*i* a *Y*=*j* vypočítat pomocí:

$$e_{ij} = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

- Karl Pearson odvodil, že statistika

$$\mathrm{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

má za platnosti H_0 chí-kvadrát rozdělení s (*r*-1)(*c*-1) stupni volnosti: $\mathrm{X}^2 \sim \chi^2_{(r-1)(c-1)}$

- Nulovou hypotézu o nezávislosti *X* a *Y* zamítáme na hladině významnosti α, když $\mathrm{X}^2 \geq \chi^2_{(1-\alpha)}(r-1)(c-1)$

Předpoklady Pearsonova chí-kvadrát testu

- Nezávislost jednotlivých pozorování
- Alespoň 80 % buněk musí mít očekávanou četnost (e_{ij}) větší než 5
- 100 % buněk musí mít očekávanou četnost (e_{ij}) větší než 2

Příklad – melanomy

Období = veličina *X*	**Lokalizace = veličina *Y*** Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	**Celkem**
1994-2000 $X = 1$	$50 = n_{11}$	$103 = n_{12}$	$116 = n_{13}$	$7 = n_{14}$	$276 = n_{1.}$
2001-2005 $X = 2$	$106 = n_{21}$	$157 = n_{22}$	$310 = n_{23}$	$54 = n_{24}$	$627 = n_{2.}$
2006-2009 $X = 3$	$115 = n_{31}$	$142 = n_{32}$	$316 = n_{33}$	$52 = n_{34}$	$625 = n_{3.}$
Celkem	$271 = n_{.1}$	$402 = n_{.2}$	$742 = n_{.3}$	$113 = n_{.4}$	$\mathbf{1528} = n$

Období = veličina *X*	**Lokalizace = veličina *Y*** Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	**Celkem**
1994-2000 $X = 1$	$e_{11} = 48.95$	$e_{12} = 72.61$	$e_{13} = 134.03$	$e_{14} = 20.41$	276
2001-2005 $X = 2$	$e_{21} = 111.20$	$e_{22} = 164.96$	$e_{23} = 304.47$	$e_{24} = 46.37$	627
2006-2009 $X = 3$	$e_{31} = 110.85$	$e_{32} = 164.43$	$e_{33} = 303.50$	$e_{34} = 46.22$	625
Celkem	271	402	742	113	1528

Příklad – melanomy

- **Př.**: Sumarizace pacientů diagnostikovaných s melanomem dle lokalizace onemocnění a roku diagnózy.
- Testová statistika:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Výpočet:

$$X^2 = \frac{(50-48{,}95)^2}{48{,}95} + \frac{(103-72{,}61)^2}{72{,}61} + \frac{(116-134{,}03)^2}{134{,}03} + \frac{(7-20{,}41)^2}{20{,}41} + \frac{(106-111{,}20)^2}{111{,}20} + \frac{(157-164{,}96)^2}{164{,}96} +$$

$$+ \frac{(310-304{,}47)^2}{304{,}47} + \frac{(54-46{,}37)^2}{46{,}37} + \frac{(115-110{,}85)^2}{110{,}85} + \frac{(142-164{,}43)^2}{164{,}43} + \frac{(316-303{,}50)^2}{303{,}50} + \frac{(52-46{,}22)^2}{46{,}22} = 30{,}41$$

- Kritická hodnota: $\chi^2_{(1-\alpha)}(r-1)(c-1) = \chi^2_{(0{,}95)}(6) = 12{,}59$

$X^2 \geq \chi^2_{(0{,}95)}(6)$ ⇒ Zamítáme H_0 o nezávislosti.

Příklad s modrýma očima

- Máme 60 studentů Matematické biologie a mezi nimi 17 s modrýma očima, 11 je současných a 6 je již vystudovaných. Testujeme nezávislost.
- Testová statistika:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Výpočet:

$$X^2 = \frac{(11-11{,}9)^2}{11{,}9} + \frac{(31-30{,}1)^2}{30{,}1} + \frac{(6-5{,}1)^2}{5{,}1} + \frac{(12-12{,}9)^2}{12{,}9} = 0{,}32$$

- Kritická hodnota: $\chi^2_{(1-\alpha)}(r-1)(c-1) = \chi^2_{(0{,}95)}(1) = 3{,}84$

$X^2 < \chi^2_{(0{,}95)}(1)$ ⟹ Nezamítáme H_0 o nezávislosti.

4. Čtyřpolní tabulky

Co je čtyřpolní tabulka

- Nejjednodušší možná kontingenčí tabulka, kdy obě sledované veličiny mají pouze dvě kategorie.
- **Příklad z 2. přednášky**: Zajímá nás přesnost vyšetření jater ultrazvukem, tedy schopnost vyšetření UTZ identifikovat maligní ložisko v pacientových játrech. Přesnost je vztažena k histologickému ověření odebrané tkáně.

Vyšetření UTZ	Histologické ověření		
	Maligní	Benigní	Celkem
Maligní	32	2	34
Benigní	3	24	27
Celkem	35	26	61

- Zde jsme závislost neověřovali, ale dokonce předpokládali!

Asociace ve čtyřpolní tabulce

- Můžeme rozhodovat o závislosti/nezávislosti dvou sledovaných veličin – nyní.
- Můžeme rozhodovat i o míře (těsnosti) této závislosti – příští přednáška.

Veličina *X*	Veličina *Y*		
	Y = 1	*Y* = 2	Celkem
X = 1	*a*	*b*	*a* + *b*
X = 2	*c*	*d*	*c* + *d*
Celkem	*a* + *c*	*b* + *d*	*n*

- Při rozhodování o nezávislosti můžeme použít Pearsonův chí-kvadrát test, ale pro malá *n* je standardem v klinických analýzách tzv. **Fisherův exaktní test** („Fisher exact test“).

Fisherův exaktní test

- Určen zejména pro čtyřpolní tabulky, **je vhodný i pro tabulku s malými četnostmi – pro ty, které nesplňují předpoklad Pearsonova testu**.
- Založen na výpočtu „přesné“ *p*-hodnoty, která zde hraje roli testové statistiky.
- **Pointa je ve výpočtu pravděpodobnosti, se kterou bychom získali čtyřpolní tabulky stejně nebo více „odchýlené“ od nulové hypotézy při zachování marginálních četností.**
- Pravděpodobnost konkrétní tabulky (s pevně zvolenou hodnotou *a* při zachování marginálních četností) lze získat:

$$p_a = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!}$$

- Pointa = spočítáme p_a všech možných tabulek při zachování marginálních četností a výsledná *p*-hodnota je součtem p_a menších nebo stejných jako p_a, která přísluší pozorované tabulce.

Příklad s modrýma očima

- Sledujeme vztah modrých očí a období studia matematické biologie.
- Pomocí Fisherova exaktního testu chceme testovat H_0 o nezávislosti.

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	11	31	42
Bývalí	6	12	18
Celkem	17	43	60

- Pravděpodobnost pozorované tabulky:

$$p_a = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!} = \frac{42!17!18!43!}{60!11!31!6!12!} = 0{,}205$$

- Tento výsledek sám o sobě znamená, že nezamítáme H_0, protože $p_a > 0{,}05$.

Příklad s modrýma očima

- Vypočítejme pravděpodobnosti pro jednotlivé možnosti kontingenční tabulky:

Studenti BIMAT	Modrá barva očí	Jiná barva očí	Celkem
Současní	a	b	42
Bývalí	c	d	18
Celkem	17	43	60

$$p_a = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!}$$

Příklad s modrýma očima

Možnosti	*a*	*b*	*c*	*d*	p_a
1.	0	42	17	1	$4{,}6 \times 10^{-14}$
2.	1	41	16	2	$1{,}7 \times 10^{-11}$
3.	2	40	15	3	$1{,}8 \times 10^{-9}$
4.	3	39	14	4	$9{,}1 \times 10^{-8}$
5.	4	38	13	5	$2{,}5 \times 10^{-6}$
6.	5	37	12	6	$4{,}1 \times 10^{-5}$
7.	6	36	11	7	$4{,}3 \times 10^{-4}$
8.	7	35	10	8	0,003
9.	8	34	9	9	0,015
10.	9	33	8	10	0,050
11.	10	32	7	11	0,121
12.	11	31	6	12	0,205
13.	12	30	5	13	0,245
14.	13	29	4	14	0,202
15.	14	28	3	15	0,111
16.	15	27	2	16	0,039
17.	16	26	1	17	0,008
18.	17	25	0	18	$6{,}6 \times 10^{-4}$

$p = 1 – 0{,}245 = 0{,}755$

⇓

Nezamítáme H_0

Fisherův × Pearsonův test

- Pearsonův chí-kvadrát test lze použít na jakoukoliv kontingenční tabulku, ALE je nutné hlídat předpoklady: 80 % e_{ij} větších než 5 – u čtyřpolní tabulky to znamená 100 %.
- Nedodržení předpokladů pro Pearsonův chí-kvadrát test může stejně jako u *t*-testu a analýzy rozptylu vést k nesmyslným závěrům!
- Situace s malými n_{ij} a tedy i e_{ij} jsou ale v medicíně i biologii velmi časté – Fisherův exaktní test je klíčový pro hodnocení čtyřpolních tabulek.

Test hypotézy o symetrii – McNemarův test

- Mám 20 pacientů, u každého opakovaně sleduji výskyt otoků před podáním a po podání léku.
- Která tabulka je správně?

	Před podáním léku	Po podání léku	Celkem
Bez otoku (úspěch)	7	12	19
S otokem (neúspěch)	13	8	21
Celkem	20	20	40

	Po podání bez otoku	Po podání s otokem	Celkem
Před podáním bez otoku	5	2	7
Před podáním s otokem	7	6	13
Celkem	12	8	20

McNemarův test

- Je to obdoba párového testu (test symetrie pro čtyřpolní tabulku).
- **Zaměřuje se pouze na pozorování, u kterých jsme při opakovaném měření zaznamenali rozdílné výsledky – za platnosti H_0 by jejich četnosti (označeny *b* a *c*) měly být stejné.**
- Testová statistika pro čtyřpolní tabulku:

$$X^2 = \frac{(b-c)^2}{b+c}$$

- Za platnosti H_0 má statistika chí-kvadrát rozdělení s 1 stupněm volnosti.
- Nulovou hypotézu o nezávislosti *X* a *Y* zamítáme na hladině významnosti α, když $X^2 \geq \chi^2_{(1-\alpha)}(1)$

- Testová statistika pro obecnou kontingenční tabulku: $X^2 = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$

Příklad – McNemarův test

- Mám 20 pacientů, u každého opakovaně sleduji ústup otoků po podání léku A a léku B. Zajímá mě rozdíl v četnosti otoků.

	Po podání B bez otoku	Po podání B s otokem	Celkem
Po podání A bez otoku	5	2	7
Po podání A s otokem	7	6	13
Celkem	12	8	20

- Testová statistika pro čtyřpolní tabulku:

$$\mathrm{X}^2 = \frac{(b-c)^2}{b+c} = \frac{(2-7)^2}{2+7} = 2{,}78$$

- Kritická hodnota: $\chi^2_{(1-\alpha)}(1) = \chi^2_{(0{,}95)}(1) = 3{,}84$

$\mathrm{X}^2 < \chi^2_{(0{,}95)}(1)$ ⟹ Nezamítáme H_0 o tom, že není rozdíl ve výskytu otoků před a po podání léku.

5. Testy o rozdělení náhodné veličiny

Testy o rozdělení náhodné veličiny

- **Kolmogorovův-Smirnovovův test** – založen na srovnání výběrové distribuční funkce s teoretickou distribuční funkcí odpovídající rozdělení, které chceme testovat. K-S test hodnotí maximální vzdálenost mezi těmito dvěma distribučními funkcemi.

- **Pearsonův chí-kvadrát test = chí-kvadrát test dobré shody** – i pro testování shody s teoretickým rozdělením je založen na myšlence srovnání pozorovaných a očekávaných četností jednotlivých hodnot, kterých nabývá náhodná veličina *X*.

- **Q-Q plot** – zobrazuje proti sobě kvantily pozorovaných hodnot a kvantily teoretického rozdělení pravděpodobnosti.

Chí-kvadrát test dobré shody

- Předpokládejme, že náhodná veličina *X* může nabývat *r* různých hodnot B_1, B_2, … ,B_r, každé s pravděpodobností p_1, p_2, … , p_r – s tím, že $\sum_{i=1}^{r} p_i = 1$

- Uvažujme *n* pozorování náhodné veličiny *X*: pokud je pravděpodobnostní model správný, měl by se počet pozorování jednotlivých variant, v_i, blížit hodnotě np_i – s tím, že $\sum_{i=1}^{r} v_i = n$

Chí-kvadrát test dobré shody

- Označme pozorovanou četnost *i*té varianty náhodné veličiny o_i („observed“) a očekávanou četnost *i*té varianty náhodné veličiny e_i („expected“).
- Opět platí, že statistika

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(o_i - e_i)^2}{e_i}$$

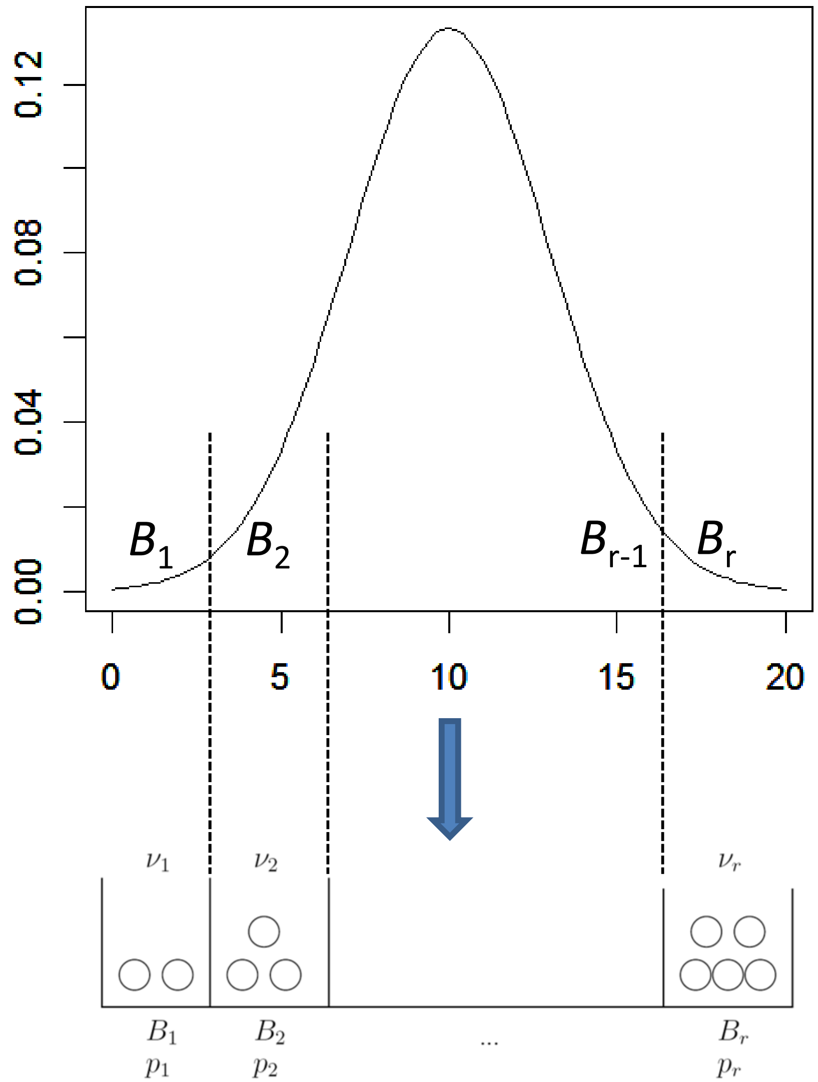
  má za platnosti H_0 chí-kvadrát rozdělení s *r*-1 stupni volnosti: $\mathrm{X}^2 \sim \chi^2_{(r-1)}$
- Nulovou hypotézu o shodě rozdělení veličiny *X* s předpokládaným rozdělením zamítáme na hladině významnosti α, když $\mathrm{X}^2 \geq \chi^2_{(1-\alpha)}(r-1)$

- Když H_0 specifikuje pouze typ rozdělení, ale ne jeho parametry, pak musí být tyto parametry odhadnuty z pozorovaných hodnot. Za každý takto odhadnutý parametr se počet stupňů volnosti testové statistiky snižuje o 1.

Chí-kvadrát test pro spojité veličiny

- Spojitá veličina samozřejmě může nabývat nespočetně mnoho hodnot v určitém intervalu.

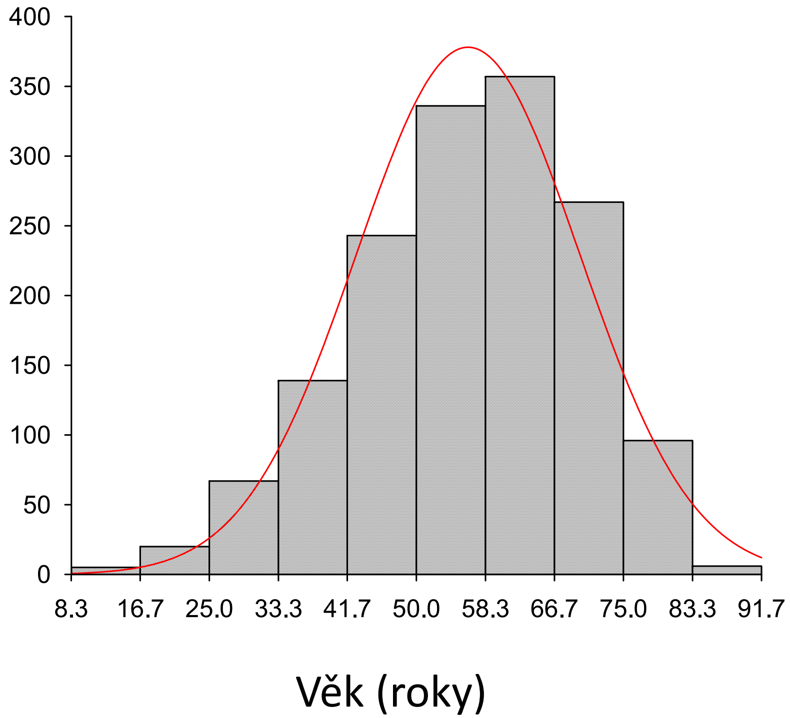
- Chí-kvadrát test dobré shody lze použít i pro spojité veličiny, které však musíme kategorizovat → rozdělit obor možných hodnot do *r* disjunktních intervalů.



Příklad – melanom a normální rozdělení

- Chceme zjistit, jestli věk u pacientů s melanomem vykazuje normální rozdělení.

$$N(\mu = 56{,}2, \sigma^2 = 182{,}4)$$

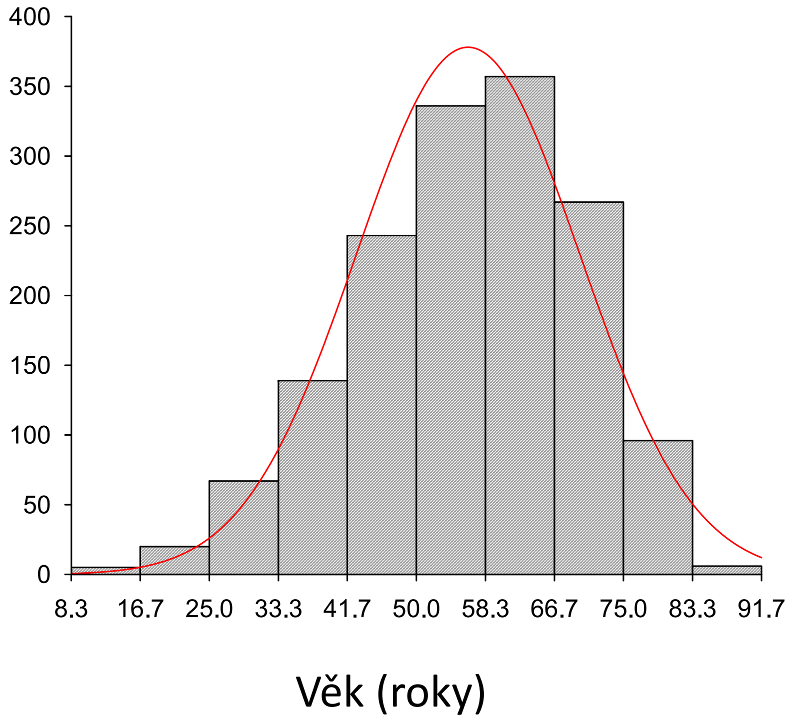


Věk – *i*-tý interval	o_i	e_i	$o_i - e_i$
0,0 – 8,3	0	0.30	-0.30
8,3 – 16,7	5	2.30	2.70
16,7 – 25,0	20	13.30	6.70
25,0 – 33,3	67	53.09	13.91
33,3 – 41,7	139	146.42	-7.42
41,7 – 50,0	243	279.13	-36.13
50,0 – 58,3	336	367.95	-31.95
58,3 – 66,7	357	335.43	21.57
66,7 – 75,0	267	211.46	55.54
75,0 – 83,3	96	92.16	3.84
83,3 – 91,7	6	27.76	-21.76
91,7 – 100,0	0	6.70	-6.70

Příklad – melanom a normální rozdělení

- Chceme zjistit, jestli věk u pacientů s melanomem vykazuje normální rozdělení.

$$N(\mu = 56{,}2, \sigma^2 = 182{,}4)$$



$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(o_i - e_i)^2}{e_i} = 56{,}6$$

$$df = r - 1 - 2 = 12 - 1 - 2 = 9$$

Odhad parametrů μ a σ^2 z dat.

$$\mathrm{X}^2 = 56{,}6 \geq \chi^2_{(1-\alpha)}(r-1-2) = \chi^2_{(0{,}95)}(9) = 16{,}92$$

$$p < 0{,}001$$

Zamítáme H_0 o normalitě rozdělení věku pacientů s melanomem.

Příklad – Poissonovo rozdělení

- **Chceme ověřit, že počet pacientů, kteří přijdou ve všední den na zubní pohotovost se řídí Poissonovým rozdělením**. Jednotkou času bude 30 minut. Celkem byly zaznamenány údaje za 1200 půlhodinových úseků.
- **H_0**: Počet příchodů pacientů během 30 minut má Poissonovo rozdělení.
- **H_1**: Počet příchodů pacientů během 30 minut nemá Poissonovo rozdělení.
- Neznáme parametr λ, je třeba ho odhadnout z dat:

$$\hat{\lambda} = \bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \frac{1}{1200}(79 \cdot 0 + 188 \cdot 1 + \ldots + 0 \cdot 11) = \frac{3364}{1200} = 2{,}80$$

- S odhadem λ lze vypočítat pravděpodobnosti pro jednotlivé hodnoty *X*:

$$p_i = P(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

- Kvůli splnění předpokladu pro aproximaci na normální rozdělení sloučíme kategorie 8, 9, 10 a 11 pacientů.

Příklad – Poissonovo rozdělení

Počet pacientů	Pozorovaná četnost	Očekávaná četnost
x_i	o_i	$e_i = np_i$
0	79	72,97
1	188	204,32
2	282	286,05
3	275	266,98
4	196	186,89
5	114	104,66
6	45	48,84
7	10	19,54
8 a více	11	9,75
Celkem	1200	1200

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(o_i - e_i)^2}{e_i} = 8{,}50$$

$$r = 9$$

$$df = r - 1 - 1 = 7$$

$$\mathrm{X}^2 = 8{,}50 < \chi^2_{(1-\alpha)}(r-1-1) = \chi^2_{(0{,}95)}(7) = 14{,}07$$

Nezamítáme H_0 o tom, že data pochází z výběru s Poissonovým rozdělením pravděpodobnosti.

Poděkování…

Rozvoj studijního oboru „Matematická biologie“ PřF MU Brno je finančně podporován prostředky projektu ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ