Základy biostatistiky

(MD710P09)
ak. rok 2008/2009

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara
katedra pravděpodobnosti a matematické statistiky MFF UK

(naposledy upraveno 5. května 2009)

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání zde přehlížíme
- postupy pro ordinální znaky existují, ale zde není na ně místo
- **příklady**
  - počty osob s krevními skupinami A, B, AB, 0
  - počty dětí narozených v jednotlivých měsících v Praze
  - počty matek se základním, středním, vysokoškolským vzděláním
- statistické jednotky třídíme do k neslučitelných kategorií
- výsledkem je k-tice (vektor) četností
- modelem pro tento vektor je multinomické rozdělení

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

multinomické rozdělení

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$ neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

$$\mathrm{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

souvislost s binomickým rozdělením

- pro $k = 2$ jsou v dílčím pokusu jen dva možné výsledky, binomické rozdělení je speciálním případem multinomického

$$\mathrm{P}\left(N_1 = n_1, N_2 = n_2\right) = \frac{n!}{n_1! n_2!} \pi_1^{n_1} \pi_2^{n_2}$$

je totéž jako (platí přece $n_1 + n_2 = n$)

$$\mathrm{P}\left(N_1 = n_1\right) = \binom{n}{n_1} \pi_1^{n_1} \pi_2^{n-n_1}$$

- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tedy

$$N_j \sim \mathrm{bi}(n, \pi_j), \quad \mathrm{E}\, N_j = n\pi_j$$

souvislost s binomickým rozdělením

- pro $k = 2$ jsou v dílčím pokusu jen dva možné výsledky, binomické rozdělení je speciálním případem multinomického

$$\mathrm{P}\left(N_1 = n_1, N_2 = n_2\right) = \frac{n!}{n_1! n_2!} \pi_1^{n_1} \pi_2^{n_2}$$

je totéž jako (platí přece $n_1 + n_2 = n$)

$$\mathrm{P}\left(N_1 = n_1\right) = \binom{n}{n_1} \pi_1^{n_1} \pi_2^{n-n_1}$$

- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tedy

$$\boxed{N_j \sim \mathrm{bi}(n, \pi_j), \quad \mathrm{E}\, N_j = n\pi_j}$$

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti
  (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $\mathrm{H}_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $\mathrm{E}\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti
  (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti
  (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti
  (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)
(X^2 – velké χ^2)

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$ má přibližně rozdělení χ^2_{k-1}

- **test shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$
  (pravděpodobnosti hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\, N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(\alpha)$, $$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$
- N_j – **experimentální** četnosti,
  $n\pi_j^0$ – **očekávané** (teoretické) četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti
  (měří jejich neshodu)

počty studentů biologie narozených v jednotlivých měsících

nulová hypotéza: děti se rodí během roku **rovnoměrně**

[chisq.test(nn,p=c(31,28,31,30,31,30,31,31,30,31,30,31)/365)]

měsíc	n_j	$n\pi_j^0$	přínos k chí-kvadrát
1	11	9,43	0,2623
2	9	8,52	0,0276
3	13	9,43	1,3539
4	11	9,12	0,3861
5	8	9,43	0,2161
6	5	9,12	1,8635
7	10	9,43	0,0348
8	6	9,43	1,2461
9	13	9,12	1,6473
10	8	9,43	0,2161
11	8	9,12	0,1383
12	9	9,43	0,0194
celkem	111	111,00	7,4115

$$X^2 = 7{,}4115 < \chi^2_{12-1}(0{,}05) = 19{,}675 \qquad p = 76{,}5\ \%$$

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při poloviních četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při poloviních četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při poloviční četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při polovičních četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: reprezentativnost výběru
(porovnat procenta v populaci a výběru **nestačí**)

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určí H_0)
- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní
- při polovičních četnostech ve výběru (28, 36, 27, 9) by vyšlo $\chi^2 = 3{,}98$, $p = 26{,}4\ \%$ (**lze** považovat za reprezentativní)

příklad: barva květů a tvar pylových zrnek

segregace dvou typů genů (C. R. Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1 (dáno)
- tvar pylu – oválný : kulatý v poměru 3 : 1 (dáno)
- platí-li nulová hypotéza (H_0 : jde o **nezávislou** segregaci), pak čtyři možné kombinace musí být v poměru 9 : 3 : 3 : 1

barva tvar	pupurová oválný	červená oválný	purpurová kulatý	červená kulatý	celkem
n_j	296	27	19	85	427
o_j	3843/16	1281/16	1281/16	427/16	427
$\frac{(n_j-o_j)^2}{o_j}$	12,97	35,17	46,57	127,41	222,12

$$\chi^2 = 222{,}12 > \chi^2_3(0{,}05) = 7{,}81$$

- nezávislost jsme **zamítli**

příklad: barva květů a tvar pylových zrnek

segregace dvou typů genů (C. R. Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1 (dáno)
- tvar pylu – oválný : kulatý v poměru 3 : 1 (dáno)
- platí-li nulová hypotéza (H_0 : jde o **nezávislou** segregaci), pak čtyři možné kombinace musí být v poměru 9 : 3 : 3 : 1

barva tvar	pupurová oválný	červená oválný	purpurová kulatý	červená kulatý	celkem
n_j	296	27	19	85	427
o_j	3843/16	1281/16	1281/16	427/16	427
$\frac{(n_j-o_j)^2}{o_j}$	12,97	35,17	46,57	127,41	222,12

$$\chi^2 = 222{,}12 > \chi^2_3(0{,}05) = 7{,}81$$

- nezávislost jsme **zamítli**

příklad: barva květů a tvar pylových zrnek

segregace dvou typů genů (C. R. Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1 (dáno)
- tvar pylu – oválný : kulatý v poměru 3 : 1 (dáno)
- platí-li nulová hypotéza (H_0 : jde o **nezávislou** segregaci), pak čtyři možné kombinace musí být v poměru 9 : 3 : 3 : 1

barva tvar	pupurová oválný	červená oválný	purpurová kulatý	červená kulatý	celkem
n_j	296	27	19	85	427
o_j	3843/16	1281/16	1281/16	427/16	427
$\frac{(n_j-o_j)^2}{o_j}$	12,97	35,17	46,57	127,41	222,12

$$\chi^2 = 222{,}12 > \chi^2_3(0{,}05) = 7{,}81$$

- nezávislost jsme **zamítli**

příklad: barva květů a tvar pylových zrnek

segregace dvou typů genů (C. R. Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1 (dáno)
- tvar pylu – oválný : kulatý v poměru 3 : 1 (dáno)
- platí-li nulová hypotéza (H_0 : jde o **nezávislou** segregaci), pak čtyři možné kombinace musí být v poměru 9 : 3 : 3 : 1

barva tvar	pupurová oválný	červená oválný	purpurová kulatý	červená kulatý	celkem
n_j	296	27	19	85	427
o_j	3843/16	1281/16	1281/16	427/16	427
$\frac{(n_j-o_j)^2}{o_j}$	12,97	35,17	46,57	127,41	222,12

$$\chi^2 = 222{,}12 > \chi^2_3(0{,}05) = 7{,}81$$

- nezávislost jsme **zamítli**

příklad: barva květů a tvar pylových zrnek

- co způsobilo zamítnutí hypotézy?

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- jsou barvy v očekávaném poměru 3 : 1?
  [chisq.test(c(315,112),p=c(3/4,1/4))]

$$\chi^2 = 0{,}3443 \qquad p = 55{,}7\ \%$$

- jsou tvary v očekávaném poměru 3 : 1?

$$\chi^2 = 0{,}0945 \qquad p = 75{,}9\ \%$$

- důvodem zamítnutí určitě závislost

příklad: barva květů a tvar pylových zrnek

- co způsobilo zamítnutí hypotézy?

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- jsou barvy v očekávaném poměru 3 : 1?
  [chisq.test(c(315,112),p=c(3/4,1/4))]

$$\chi^2 = 0{,}3443 \qquad p = 55{,}7\ \%$$

- jsou tvary v očekávaném poměru 3 : 1?

$$\chi^2 = 0{,}0945 \qquad p = 75{,}9\ \%$$

- důvodem zamítnutí určitě závislost

příklad: barva květů a tvar pylových zrnek

- co způsobilo zamítnutí hypotézy?

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- jsou barvy v očekávaném poměru 3 : 1?
  [chisq.test(c(315,112),p=c(3/4,1/4))]

$$\chi^2 = 0{,}3443 \qquad p = 55{,}7\ \%$$

- jsou tvary v očekávaném poměru 3 : 1?

$$\chi^2 = 0{,}0945 \qquad p = 75{,}9\ \%$$

- důvodem zamítnutí určitě závislost

příklad: barva květů a tvar pylových zrnek

- co způsobilo zamítnutí hypotézy?

barva	purpurová	červená	celkem
oválný tvar	296	27	323
kulatý tvar	19	85	104
celkem	315	112	427

- jsou barvy v očekávaném poměru 3 : 1?
  [chisq.test(c(315,112),p=c(3/4,1/4))]

$$\chi^2 = 0{,}3443 \qquad p = 55{,}7\ \%$$

- jsou tvary v očekávaném poměru 3 : 1?

$$\chi^2 = 0{,}0945 \qquad p = 75{,}9\ \%$$

- důvodem zamítnutí určitě závislost

složená nulová hypotéza (hypotéza o struktuře)

- hypotéza určuje vztahy mezi pravděpodobnostmi $\pi_1, \ldots, \pi_k$ některé parametry zůstávají volné, je třeba je odhadnout
- příklad antigen: (Hardy-Weinberg equilibrium) model pro fenotypy AA, Aa, aa

$$\begin{aligned} \mathrm{P}(AA) &\equiv \pi_1(\theta) = \theta^2 \\ \mathrm{P}(Aa) &\equiv \pi_2(\theta) = 2\theta(1-\theta) \\ \mathrm{P}(aa) &\equiv \pi_3(\theta) = (1-\theta)^2 \end{aligned}$$

- neurčený parametr θ – pravděpodobnost alely A
- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem, tj. s H-W rovnováhou?

složená nulová hypotéza (hypotéza o struktuře)

- hypotéza určuje vztahy mezi pravděpodobnostmi $\pi_1, \ldots, \pi_k$ některé parametry zůstávají volné, je třeba je odhadnout
- příklad antigen: (Hardy-Weinberg equilibrium) model pro fenotypy AA, Aa, aa

$$\begin{aligned} \mathrm{P}(AA) &\equiv \pi_1(\theta) = \theta^2 \\ \mathrm{P}(Aa) &\equiv \pi_2(\theta) = 2\theta(1-\theta) \\ \mathrm{P}(aa) &\equiv \pi_3(\theta) = (1-\theta)^2 \end{aligned}$$

- neurčený parametr θ – pravděpodobnost alely A
- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem, tj. s H-W rovnováhou?

složená nulová hypotéza (hypotéza o struktuře)

- hypotéza určuje vztahy mezi pravděpodobnostmi $\pi_1, \ldots, \pi_k$ některé parametry zůstávají volné, je třeba je odhadnout
- příklad antigen: (Hardy-Weinberg equilibrium) model pro fenotypy AA, Aa, aa

$$\begin{aligned} \mathrm{P}(AA) &\equiv \pi_1(\theta) = \theta^2 \\ \mathrm{P}(Aa) &\equiv \pi_2(\theta) = 2\theta(1-\theta) \\ \mathrm{P}(aa) &\equiv \pi_3(\theta) = (1-\theta)^2 \end{aligned}$$

- neurčený parametr θ – pravděpodobnost alely A
- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem, tj. s H-W rovnováhou?

složená nulová hypotéza (hypotéza o struktuře)

- hypotéza určuje vztahy mezi pravděpodobnostmi $\pi_1, \ldots, \pi_k$ některé parametry zůstávají volné, je třeba je odhadnout
- příklad antigen: (Hardy-Weinberg equilibrium) model pro fenotypy AA, Aa, aa

$$\begin{aligned} \mathrm{P}(AA) &\equiv \pi_1(\theta) = \theta^2 \\ \mathrm{P}(Aa) &\equiv \pi_2(\theta) = 2\theta(1-\theta) \\ \mathrm{P}(aa) &\equiv \pi_3(\theta) = (1-\theta)^2 \end{aligned}$$

- neurčený parametr θ – pravděpodobnost alely A
- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem, tj. s H-W rovnováhou?

- odhad θ maximalizací *logaritmické věrohodnostní funkce*

$$\begin{aligned}\ell(\theta) &= \ln(\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3)) \\ &= \ln\left(c_1 \left(\theta^2\right)^{n_1} \left(2\theta(1-\theta)\right)^{n_2} \left((1-\theta)^2\right)^{n_3}\right) \\ &= c_2 + (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta)\end{aligned}$$

$$\hat{\theta} = \frac{2 \cdot N_1 + N_2}{2n} \quad \left(= \frac{2 \cdot 18 + 17}{82} = 0{,}646\right)$$

- obecně se H_0 zamítá, pokud (θ má q nezávislých složek)

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j(\hat{\theta}))^2}{n\pi_j(\hat{\theta})} \geq \chi^2_{k-1-q}(\alpha)$$

- příklad antigen: $\chi^2 = 0{,}355 < \chi^2_{3-1-1}(0{,}05) = 3{,}84$
  $p = 55{,}1\ \%$ hypotézu na 5% hladině nezamítáme

- odhad θ maximalizací *logaritmické věrohodnostní funkce*

$$\begin{aligned}\ell(\theta) &= \ln(\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3)) \\ &= \ln\left(c_1 \left(\theta^2\right)^{n_1} \left(2\theta(1-\theta)\right)^{n_2} \left((1-\theta)^2\right)^{n_3}\right) \\ &= c_2 + (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta)\end{aligned}$$

$$\hat{\theta} = \frac{2 \cdot N_1 + N_2}{2n} \quad \left(= \frac{2 \cdot 18 + 17}{82} = 0{,}646\right)$$

- obecně se H_0 zamítá, pokud (θ má q nezávislých složek)

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j(\hat{\theta}))^2}{n\pi_j(\hat{\theta})} \geq \chi^2_{k-1-q}(\alpha)$$

- příklad antigen: $\chi^2 = 0{,}355 < \chi^2_{3-1-1}(0{,}05) = 3{,}84$
  $p = 55{,}1\ \%$ hypotézu na 5% hladině nezamítáme

- odhad θ maximalizací *logaritmické věrohodnostní funkce*

$$\begin{aligned}\ell(\theta) &= \ln(\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3)) \\ &= \ln\left(c_1 \left(\theta^2\right)^{n_1} (2\theta(1-\theta))^{n_2} \left((1-\theta)^2\right)^{n_3}\right) \\ &= c_2 + (2n_1 + n_2)\ln\theta + (n_2 + 2n_3)\ln(1-\theta)\end{aligned}$$

$$\hat{\theta} = \frac{2 \cdot N_1 + N_2}{2n} \quad \left(= \frac{2 \cdot 18 + 17}{82} = 0{,}646\right)$$

- obecně se H_0 zamítá, pokud (θ má q nezávislých složek)

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j(\hat{\theta}))^2}{n\pi_j(\hat{\theta})} \geq \chi^2_{k-1-q}(\alpha)$$

- příklad antigen: $\chi^2 = 0{,}355 < \chi^2_{3-1-1}(0{,}05) = 3{,}84$
  $p = 55{,}1\ \%$ hypotézu na 5% hladině nezamítáme