Základy biostatistiky

(MD710P09)
ak. rok 2007/2008

Karel Zvára

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

(naposledy upraveno 6. května 2008)

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

mnohonásobná lineární regrese

- závislost na dvou (nebo více) nezávisle proměnných
- pozorování $(x_1, v_1, Y_1), \ldots, (x_n, v_n, Y_n)$
- představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{} + E_i$$

- střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- $E_1, \ldots, E_n$ (také $Y_1, \ldots, Y_n$) jsou **nezávislé** náhodné veličiny
- $E_i \sim \mathrm{N}(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

interpretace

- b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- **koeficient determinace** R^2
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $H_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

- **koeficient determinace** R^2
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

- **koeficient determinace** R^2
  podíl celkové variability, který se podařilo vysvětlit závislostí Y na x, v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(\alpha)$$

- p-hodnota tohoto testu bývá uváděna spolu s R^2

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

testy o přínosu jednotlivých regresorů

- model $y = \beta_0 + \beta_1 x + \beta_2 v$
- $H_0 : \beta_2 = 0$
  k vysvětlení chování Y stačí x, tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \qquad \text{zamítat pro } |T_2| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_1 = 0$
  k vysvětlení chování Y stačí v, tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \qquad \text{zamítat pro } |T_1| \geq t_{n-3}(\alpha)$$

- $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	$<$0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	$<$0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	$<$0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

příklad: závislost procenta tuku na výšce a váze

data: Policie

regresor	b_j	S.E.(b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	–0,262	0,110	–2,376	0,0216
weight	0,624	0,0690	9,050	$<$0,0001

- [summary(lm(fat∼height+weight))]
- při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

tabulka analýzy rozptylu

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- $R^2 = 1833{,}11/2676{,}95 = 1 - 843{,}85/2676{,}95 = 0{,}685$
- závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- $s^2 = 17{,}95$
- na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

regresní diagnostika
zda byly splněny předpoklady

a) zvolili jsme správně **tvar závislosti**?

b) je **rozptyl** všude **stejný**?

c) je přiměřeně splněn předpoklad o **normálním rozdělení**?

d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas

- často pomůže transformace (a), b), c)), např. logaritmování závisle proměnné
- [plot(lm(fat∼height+weight))]

hodnocení kvalitativních znaků

- znaky v **nominálním** měřítku
- někdy i v ordinálním měřítku, ale uspořádání 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 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 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 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 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 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 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 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 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 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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \mathrm{bi}(n, \pi_j)}$
- **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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \text{bi}(n, \pi_j)}$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

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

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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \mathrm{bi}(n, \pi_j)}$
- **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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \text{bi}(n, \pi_j)}$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

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

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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \mathrm{bi}(n, \pi_j)}$
- **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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \text{bi}(n, \pi_j)}$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

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

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 $n, \pi_1, \ldots, \pi_k$
- každé N_j (samotné, proti ostatním četnostem) má binomické rozdělení, tj. $\boxed{N_j \sim \text{bi}(n, \pi_j)}$
- **pravděpodobnost** toho, že $N_1 = n_1, \ldots, N_k = n_k$

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

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

vlastnost χ^2 (chí-kvadrát)

- 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, $n\pi_j^0$ – **teoretické** četnosti
- statistika X^2 porovnává experimentální a teoretické četnosti (měří jejich neshodu)

počty studentů biologie narozených v jednotlivých měsících

hypotéza: dětí 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
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

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 %
- ve vzorku pacientů byly počty osob s krevními skupinami po řadě 28, 36, 27, 9
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?
- $$\chi^2 = \frac{(28-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(27-20)^2}{20} + \frac{(9-10)^2}{10}$$
  $$= 3{,}98 \qquad p = 26{,}4\ \%$$
- výběr **lze** považovat za reprezentativní

příklad: reprezentativnost výběru

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 %
- ve vzorku pacientů byly počty osob s krevními skupinami po řadě 28, 36, 27, 9
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?
- $$\chi^2 = \frac{(28-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(27-20)^2}{20} + \frac{(9-10)^2}{10}$$
  $$= 3{,}98 \qquad p = 26{,}4\ \%$$
- výběr **lze** považovat za reprezentativní

příklad: reprezentativnost výběru

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 %
- ve vzorku pacientů byly počty osob s krevními skupinami po řadě 28, 36, 27, 9
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?
- $$\chi^2 = \frac{(28-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(27-20)^2}{20} + \frac{(9-10)^2}{10}$$
  $$= 3{,}98 \qquad p = 26{,}4\ \%$$
- výběr **lze** považovat za reprezentativní

příklad: reprezentativnost výběru

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 %
- ve vzorku pacientů byly počty osob s krevními skupinami po řadě 28, 36, 27, 9
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?
- $$\chi^2 = \frac{(28-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(27-20)^2}{20} + \frac{(9-10)^2}{10}$$
  $$= 3{,}98 \qquad p = 26{,}4\ \%$$
- výběr **lze** považovat za reprezentativní

příklad: reprezentativnost výběru

- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 %
- ve vzorku pacientů byly počty osob s krevními skupinami po řadě 28, 36, 27, 9
- lze považovat tento výběr za reprezentativní vzhledem k výskytu krevních skupin?
- $$\chi^2 = \frac{(28-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(27-20)^2}{20} + \frac{(9-10)^2}{10}$$
  $$= 3{,}98 \qquad p = 26{,}4\ \%$$
- výběr **lze** považovat za reprezentativní

příklad: barva květů a tvar pylových zrnek

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

- barva květů – purpurová : červená v poměru 3 : 1
- tvar pylu – oválný : kulatý v poměru 3 : 1
- jde-li o nezávislou segregaci, pak čtyři možné kombinace 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	$\frac{3843}{16}$	$\frac{1281}{16}$	$\frac{1281}{16}$	$\frac{427}{16}$	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 (Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1
- tvar pylu – oválný : kulatý v poměru 3 : 1
- jde-li o nezávislou segregaci, pak čtyři možné kombinace 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	$\frac{3843}{16}$	$\frac{1281}{16}$	$\frac{1281}{16}$	$\frac{427}{16}$	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 (Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1
- tvar pylu – oválný : kulatý v poměru 3 : 1
- jde-li o nezávislou segregaci, pak čtyři možné kombinace 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	$\frac{3843}{16}$	$\frac{1281}{16}$	$\frac{1281}{16}$	$\frac{427}{16}$	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 (Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1
- tvar pylu – oválný : kulatý v poměru 3 : 1
- jde-li o nezávislou segregaci, pak čtyři možné kombinace 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	$\frac{3843}{16}$	$\frac{1281}{16}$	$\frac{1281}{16}$	$\frac{427}{16}$	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 (Rao: Lineární metody statistické indukce ..., str. 439)

- barva květů – purpurová : červená v poměru 3 : 1
- tvar pylu – oválný : kulatý v poměru 3 : 1
- jde-li o nezávislou segregaci, pak čtyři možné kombinace 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	$\frac{3843}{16}$	$\frac{1281}{16}$	$\frac{1281}{16}$	$\frac{427}{16}$	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

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

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

- je tvar v očekávaném poměru?

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

- důvodem zamítnutí určitě závislost (další přednáška)

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

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

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

- je tvar v očekávaném poměru?

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

- důvodem zamítnutí určitě závislost (další přednáška)

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

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

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

- je tvar v očekávaném poměru?

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

- důvodem zamítnutí určitě závislost (další přednáška)

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

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

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

- je tvar v očekávaném poměru?

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

- důvodem zamítnutí určitě závislost (další přednáška)

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 (neurčený parametr θ)

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

- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem?

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 (neurčený parametr θ)

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

- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem?

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 (neurčený parametr θ)

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

- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$ v souladu s modelem?

- 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

- 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