

# Analýza rozptylu, kontingenční tabulky a $\chi^2$ testy

—  
Supplementum ke cvičení 4ST201 Statistika

Lubomír Štěpánek<sup>1, 2</sup>



<sup>1</sup>Oddělení biomedicínské statistiky  
Ústav biofyziky a informatiky  
1. lékařská fakulta  
Univerzita Karlova, Praha



<sup>2</sup>Katedra biomedicínské informatiky  
Fakulta biomedicínského inženýrství  
České vysoké učení technické v Praze

(2019) Lubomír Štěpánek, CC BY-NC-ND 3.0 (CZ)



Dílo lze dále svobodně šířit, ovšem s uvedením původního autora a s uvedením původní licence. Dílo není možné šířit komerčně ani s ním jakkoliv jinak nakládat pro účely komerčního zisku. Dílo nesmí být jakkoliv upravováno. Autor neručí za správnost informací uvedených kdekoli v předložené práci, přesto vynaložil nezanedbatelné úsilí, aby byla uvedená fakta správná a aktuální, a práci sepsal podle svého nejlepšího vědomí a svých „nejlepších“ znalostí problematiky.

# Obsah

- 1 Opakování
- 2 Jednofaktorová analýza rozptylu
- 3 Kontingenční tabulka
- 4  $\chi^2$  testy
- 5 Literatura

# Příklad

- Balící linka má zhotovovat kilogramové balíčky cukru. Z dvaceti náhodně vybraných balíčků cukru byla zjištěna jejich průměrná hmotnost 0,997 kg a výběrová směrodatná odchylka 0,015 kg. Připravuje balící linka balíčky cukru lehčí než jeden kilogram? Testujme na hladině významnosti 0,05.

# Příklad

- Ze 100 hodů jednou mincí padla hlava 60-krát. Je pravděpodobnost padnutí hlavy a orla shodná na hladině významnosti 0,05? Kolikrát by během 100 hodů musela padnout hlava, abychom zamítli hypotézu o shodné pravděpodobnosti padnutí hlavy a orla na hladině významnosti 0,01?

# Jednofaktorová analýza rozptylu

- též zvaná jednofaktorová ANOVA (Analysis of Variance)
- hodnotí závislost kvantitativní proměnné  $Y$  na nějaké kvalitativní proměnné, která má  $k \geq 3$  hodnot<sup>1</sup>
  - např. průměrná výše mzdy podle dosaženého vzdělání (základní, střední, vysoké)
- kvalitativní proměnná tak rozdělí hodnoty kvantitativní proměnné na skupiny hodnot  $Y_1, Y_2, \dots, Y_k$  pro  $k \geq 3$  (ale i  $k = 2$ )

---

<sup>1</sup>pro  $k = 2$  hodnoty kvalitativní proměnné lze užít dvouvýběrový  $t$ -test

# Jednofaktorová analýza rozptylu

- předpokládáme, že jednotlivé skupiny hodnot  $Y_j$  pro  $\forall j \in \{1, 2, \dots, k\}$  rozdělené podle  $k$  hodnot kvalitativní proměnné jsou navzájem nezávislé a že  $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$  tak že  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$ , tedy že všechny skupiny hodnot mají stejný rozptyl  $\sigma^2$  (tzv. *homoskedasticita*)
  - shodnost rozptylů v jednotlivých skupinách lze formálně testovat Bartlettovým nebo Leveneovým testem
- jednofaktorová analýza rozptylu tak zkoumá nulovou hypotézu  $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$  o shodnosti středních hodnot v  $k$  skupinách

# Jednofaktorová analýza rozptylu

- za předpokladu platnosti nulové hypotézy tak očekáváme, že průměry všech  $k$  skupin rovnají; tento společný celkový průměr  $\bar{X}$  lze odhadnout z hodnot výběru
- je-li  $\forall j \in \{1, 2, \dots, k\}$  v  $j$ -té skupině právě  $n_j$  pozorování, pak celkový součet čtverců odchylek je

$$S_x = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{i,j} - \bar{X})^2$$

a lze rozdělit na meziskupinový součet čtverců  $S_{x.m}$   
a vnitroskupinový součet čtverců  $S_{x.v}$

$$S_{x.m} = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2 n_j \quad S_{x.v} = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{i,j} - \bar{X}_j)^2$$



# Jednofaktorová analýza rozptylu

- protože platí  $S_x = S_{x.m} + S_{x.v}$  a hodnota  $S_x$  je jednoznačně dána hodnotami ve výběru, svědčí pro alternativní hypotézu vysoká meziskupinová variabilita
- testovým kritériem je statistika  $F$

$$F = \frac{\frac{S_{x.m}}{k-1}}{\frac{S_{x.v}}{n-k}} \sim F(k-1, n-k)$$

sledující Fisherovo-Snedecorovo rozdělení o počtech stupňů volnosti  $k-1$  a  $n-k$

- protože proti nulové hypotéze (tedy ve prospěch alternativní) svědčí velké hodnoty  $S_{x.m}$  a tedy i  $F$ , zamítáme nulovou hypotézu pro taková  $F$ , že

$$W_\alpha = \{F : F \geq F_{1-\alpha}\}$$

# Příklad

- Každou ze tří variant testu (A, B a C) napsali vždy čtyři náhodní studenti. Je třeba rozhodnout, zda jsou na základě výsledků v tabulce varianty testu obdobně náročné, předpokládáme-li, že apriorní znalosti všech dvanácti studentů jsou srovnatelné.

		výsledek testu			
		1. student	2. student	3. student	4. student
varianta	A	91	81	74	57
	B	83	72	63	47
	C	71	69	58	40

# Kontingenční tabulka

- jde o tabulku o rozměrech  $r \times s$ , kde buňka v  $i$ -tém řádku  $j$ -tého sloupce tak, že  $i \in \{1, 2, \dots, r\}$  a  $j \in \{1, 2, \dots, s\}$ , uvádí, kolik případů ( $n_{i,j}$ ) bylo pozorováno pro kombinaci  $i$ -té hodnoty řádkové kvalitativní proměnné a  $j$ -té hodnoty sloupcové kvalitativní proměnné

		hodnoty sloupcové proměnné			
		1	2	...	$s$
hodnoty řádkové proměnné	1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,s}$
	2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,s}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$r$	$n_{r,1}$	$n_{r,2}$	...	$n_{r,s}$

# $\chi^2$ test nezávislosti

- předpokládáme kontingenční tabulku o rozměrech  $r \times s$  tak, že v  $i$ -tém řádku  $j$ -tého sloupce je hodnota  $n_{i,j}$
- pak zřejmě je

$$\sum_{i=1}^r \sum_{j=1}^s n_{i,j} = \sum_{i=1}^r n_{i,\bullet} = \sum_{j=1}^s n_{\bullet,j} = n$$

kde  $n_{i,\bullet}$  je součet  $i$ -tého řádku kontingenční tabulky a  $n_{\bullet,j}$  je součet  $j$ -tého sloupce kontingenční tabulky

- $\chi^2$  test nezávislosti testuje nulovou hypotézu  $H_0$  o tom, že řádková a sloupcová kvalitativní proměnná jsou nezávislé
- testovou statistikou je

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet} n_{\bullet,j}}{n}} \sim \chi^2_{1-\alpha}((r-1)(s-1))$$

# $\chi^2$ test nezávislosti

- pro alternativní hypotézu tedy svědčí velké rozdíly mezi pozorovanými četnostmi  $n_{i,j}$  a očekávanými četnostmi  $\frac{n_{i,\bullet}n_{\bullet,j}}{n}$ , tedy obecně velké hodnoty statistiky  $\chi^2$
- proto nulovou hypotézu zamítáme pro takové  $\chi^2$ , že

$$W_\alpha = \{\chi^2 : \chi^2 \geq \chi^2_{1-\alpha}((r-1)(s-1))\}$$

# Příklad

- Z průzkumu trhu byla sestavena kontingenční tabulka preference budoucího bydliště dle pohlaví. Výsledky jsou uvedeny níže. Závisejí preference budoucího bydliště na pohlaví? Zkoumejme na hladině významnosti 0,05.

		preference budoucího bydliště	
		město	venkov
pohlaví	muž	71	91
	žena	82	56

# $\chi^2$ test dobré shody

- $\chi^2$  test dobré shody testuje nulovou hypotézu  $H_0 : \pi_j = \pi_{0,j}$  pro  $\forall j \in \{1, 2, \dots, k\}$  o tom, zda je pravděpodobnost nastání jevu našeho zájmu  $\pi_j$  v  $j$ -té kategorii rovna očekávané pravděpodobnosti  $\pi_{0,j}$
- je-li očekávaná pravděpodobnost  $j$ -té kategorie rovna  $\pi_{0,j}$ , pak zřejmě očekávaná četnost je  $n\pi_{0,j}$ , kde  $n$  je celkový počet pozorování,
- naopak pozorovaná četnost  $j$ -té kategorie ať je  $n_j$
- $\chi^2$  test dobré shody lze tak aplikovat i na kontingenční tabulku
- testovou statistikou je

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - n\pi_{0,j})^2}{n\pi_{0,j}} \sim \chi^2_{1-\alpha}(k-1)$$

# $\chi^2$ test dobré shody

- pro alternativní hypotézu tedy svědčí velké rozdíly mezi pozorovanými četnostmi  $n_j$  a očekávanými četnostmi  $n\pi_{0,j}$ , tedy obecně velké hodnoty statistiky  $\chi^2$
- proto nulovou hypotézu zamítáme pro takové  $\chi^2$ , že

$$W_\alpha = \{\chi^2 : \chi^2 \geq \chi_{1-\alpha}^2(k-1)\}$$





# Příklad

- Při 600 hodech hrací kostkou byly zjištěny následující četnosti jednotlivých šesti stran

85, 99, 91, 108, 119, 98.

Lze na hladině významnosti 0,05 považovat takovou kostku za spravedlivou?

# Literatura

-  Hindls, Richard, Stanislava Hronová, Jan Seger a Jakub Fischer. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. ISBN: 978-80-86946-43-6.
-  Marek, Luboš. *Statistika v příkladech*. Praha: Professional Publishing, 2015. ISBN: 978-80-7431-153-6.

Děkuji za pozornost!

lubomir.stepanek@vse.cz

lubomir.stepanek@lf1.cuni.cz

lubomir.stepanek@fbmi.cvut.cz

[https://github.com/LStepanek/4ST201\\_Statistika](https://github.com/LStepanek/4ST201_Statistika)