χ^2 testy. Test nekorelovanosti.

Petr Pošík

Části dokumentu jsou převzaty (i doslovně)
z *Mirko Navara: Pravděpodobnost a matematická statistika*,
`https://cw.felk.cvut.cz/lib/exe/fetch.php/courses/a6m33ssl/pms_print.pdf`
s laskavým svolením autora.

χ^2 testy 2
χ^2 test dobré shody . . . 4
Př: Test dobré shody . . . 5
Možné problémy a jejich řešení . . . 6
Příklad: Test shody s normálním rozdělením . . . 7
χ^2 test dobré shody 2 diskrétních rozdělení . . . 8
χ^2 test nezávislosti 2 rozdělení . . . 9
Př: χ^2 test nezávislosti . . . 10
Př: χ^2 test nezávislosti . . . 12

Korelace, odhad a testování 13
Korelace . . . 14
Test nekorelovanosti dvou normálních rozdělení . . . 15
Příklad: Test nekorelovanosti . . . 16

Shoda očekávaných a pozorovaných četností

Manažer velké firmy na poradě vedení:

„Mám tu alarmující zprávu z poslední kontroly docházky našich zaměstnanců. Celých 40 % sick-leavů připadá na pondělky a pátky! S tím musíme něco udělat!"

:-)

χ^2 test dobré shody

χ^2 **test dobré shody**:

- Slouží k testování hypotézy, že náhodná veličina má předpokládané rozdělení (umíme hypotézy jen zamítat; nikdy nepotvrdíme, že toto rozdělení skutečně má).
- Testuje shodu s *diskrétním rozdělením*, které ovšem mohlo vzniknout diskretizací spojitého.
- H_0: diskrétní veličina má rozdělení do k tříd s nenulovými pravděpodobnostmi $p_1, \ldots, p_k$.

Testujeme pomocí realizace náhodného výběru rozsahu n. Není důležité pořadí výsledků, pouze jejich **pozorované četnosti** n_i, $i = 1, \ldots, k$, které porovnáváme s **teoretickými (očekávanými) četnostmi** np_i.

Testovací statistikou je

$$T := \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

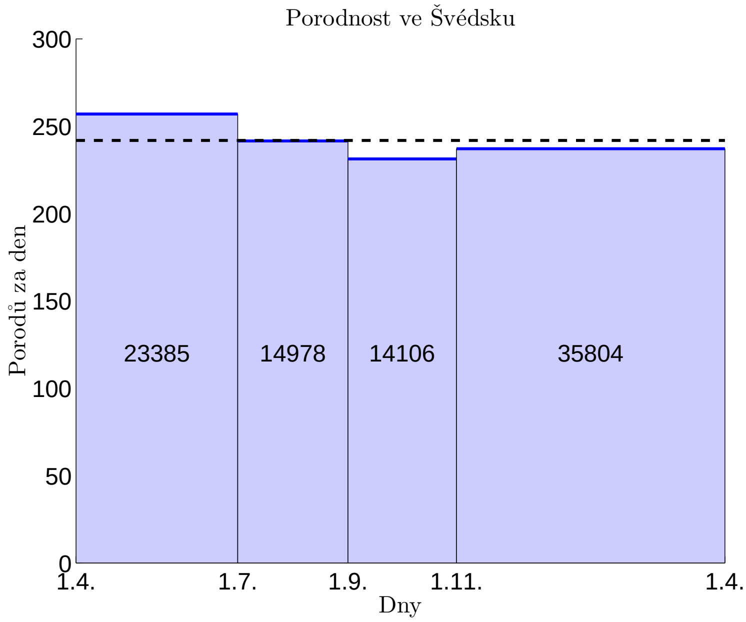
jejíž rozdělení se pro $n \to \infty$ blíží $\chi^2(k-1)$:

- Dosažená hladina významnosti $p = 1 - F_{\chi^2(k-1)}(t)$.
- H_0 zamítáme pro $t > q_{\chi^2(k-1)}(1-\alpha)$, tj. pro $p = 1 - F_{\chi^2(k-1)}(t) < \alpha$.

Příklad: Test dobré shody

Zadání: Ověřte H_0, že porodnost ve Švédsku není ovlivněna ročním obdobím. Na jaře, v létě, na podzim a v zimě se tam narodilo 23385, 14978, 14106 a 36519 dětí.

Řešení: Roční období, jak je znají ve Švédsku, mají odlišný počet dní než v ČR: jaro 91, léto 62, podzim 61 a zima 151.



Použijme χ^2 test dobré shody pozorovaných a očekávaných četností:

i	Období	Měsíce	Dnů	$p_i = \frac{\text{Dnů}}{365}$	n_i	np_i	$(n_i - np_i)$	$\frac{(n_i - np_i)^2}{np_i}$
1	Jaro	duben–červen	91	0.24932	23385	22008	1377	86.16
2	Léto	červenec–srpen	62	0.16986	14978	14944	-16	0.02
3	Podzim	září–říjen	61	0.16712	14106	14752	-646	28.29
4	Zima	listopad–březen	151	0.41370	35804	36519	-715	14.00
$k = 4$	Celkem		365	1	88273	88273	0	128.47

- Realizace testovací statistiky je

$$t = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = 128.47.$$

- Dosažená hladina významnosti $p = 1 - F_{\chi^2(k-1)}(t) = 1 - F_{\chi^2(3)}(128.47) \doteq 0$.
- Zamítáme H_0. Roční období statisticky významně ovlivňuje porodnost ve Švédsku.

Možné problémy a jejich řešení

Problém: Testujeme na rozdělení, kterému se to skutečné jen limitně blíží. Dopouštíme se blíže neurčené dodatečné chyby. Aby byl náš předpoklad oprávněný, *teoretické četnosti tříd nesmí být příliš malé (alespoň 5)*.

- Vychází-li teoretická četnost v nějaké třídě příliš malá, sloučíme ji s jinými třídami, pokud možno „blízkými“, či „podobnými“.

Problém: Zkoumané rozdělení může záviset na neznámých parametrech.

- Parametry odhadneme na základě *jiného* náhodného výběru.
- Parametry odhadneme na základě *stejného* náhodného výběru, který používáme k testu dobré shody. Tím jsme ale snížili počet stupňů volnosti, takže musíme testovat na rozdělení $\chi^2(k-1-q)$, kde q je počet odhadnutých parametrů.

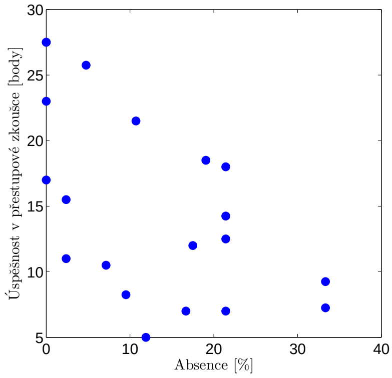
Problém: Chceme testovat shodu se *spojitým* nebo *smíšeným* rozdělením.

- Rozdělení diskretizujeme, tj. všechny možné výsledky rozdělíme do k disjunktních tříd. Prvky v každé třídě si mají být "blízké", jinak snižujeme sílu testu. Všechny teoretické četnosti by měly být dostatečně velké a nejlépe zhruba stejné.

Poznámka: Zásadně musíme pracovat s *jednotkami* (objekty), z nichž každá zvlášť (a nezávisle) je zařazena do nějaké třídy. *Nelze počítat s tisíci, procenty, spojitým množstvím, atd.*

Příklad: Test shody s normálním rozdělením

Zadání: Na jistém gymnáziu v kvartě v předmětu Matematika byla nasbírána následující data týkající se absence při výuce (X) a úspěšnosti ve zkoušce (Y). Výběr obsahuje 21 studentů (2 body v grafu mají četnost 2). Ověřte hypotézy o normalitě obou proměnných.



Řešení: χ^2 test shody s diskretizovaným normálním rozdělením.
Poznámka: *Takto se obvykle normalita netestuje, existují jiné, lepší testy shody s normálním rozdělením (např. Shapiro-Wilk). Zde uvedeno jako demonstrace využití χ^2 testu.*

- Maximálně věrohodné odhady parametrů rozdělení: $X \sim \mathrm{N}(\bar{x}, s_x^2) = \mathrm{N}(12.23, 115.19)$, $Y \sim \mathrm{N}(\bar{y}, s_y^2) = \mathrm{N}(14.73, 49.65)$.
- Diskretizace: Vzhledem k tomu, že rozsah výběru je 21 a že teoretická četnost v každé skupině by měla být alespoň 5, nemůžeme si dovolit diskretizaci do více než 4 skupin. Použijme tedy intervaly s hranicemi $(-\infty$, dolní kvartil $q(\frac{1}{4})$, medián $q(\frac{1}{2})$, horní kvartil $q(\frac{3}{4}), \infty)$.
- Test dobré shody: protože jsme odhadli 2 parametry rozdělení ze stejných dat, musíme testovat na rozdělení $\chi^2(4-1-2) = \chi^2(1)$.

	Četnost v intervalu				t	$p = 1 - F_{\chi^2(1)}(t)$
	$(-\infty, q(\frac{1}{4}))$	$\langle q(\frac{1}{4}), q(\frac{1}{2}))$	$\langle q(\frac{1}{2}), q(\frac{3}{4}))$	$\langle q(\frac{3}{4}), \infty)$		
Očekávané	5.25	5.25	5.25	5.25		
X	8	4	3	6	2.81	0.094
Y	6	6	4	5	0.52	0.469

- Hypotézy o normalitě X a Y nemůžeme zamítnout.

χ^2 test dobré shody 2 diskrétních rozdělení

H_0: Dvě diskrétní náhodné veličiny mají stejné rozdělení p.

- Rozsahy výběrů jsou m a n a pozorované četnosti ve třídách $i = 1, \dots, k$ jsou m_i a n_i takové, že $\sum_i m_i = m$ a $\sum_i n_i = n$.
- Předpokládáme diskrétní rozdělení p s neznámými parametry $p_i, i = 1, \dots, k$.

$$\sum_{i=1}^{k} \frac{(m_i - mp_i)^2}{mp_i} \text{ se blíží } \chi^2(k-1), \quad \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \text{ se blíží } \chi^2(k-1),$$

$$T := \sum_{i=1}^{k} \frac{(m_i - mp_i)^2}{mp_i} + \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \text{ se blíží } \chi^2(2(k-1)).$$

- Neznámé parametry p_i odhadneme metodou maximální věrohodnosti jako

$$p_i = \frac{m_i + n_i}{m + n},$$

  z nichž je ale jen $k-1$ nezávislých, takže výsledný počet stupňů volnosti je $2(k-1) - (k-1) = (k-1)$ a testujeme T na rozdělení $\chi^2(k-1)$.
- H_0 zamítáme pro $t > q_{\chi^2(k-1)}(1-\alpha)$, tj. pro $p = 1 - F_{\chi^2(k-1)}(t) < \alpha$.

Praktičtější ekvivalentní vzorec pro T:

$$T = \left(\frac{1}{m} + \frac{1}{n}\right) \sum_{i=1}^{k} \frac{(m_i - mp_i)^2}{mp_i}.$$

χ^2 test nezávislosti 2 rozdělení

H_0: Dvě diskrétní náhodné veličiny (jejichž rozdělení neznáme) jsou nezávislé.

- X nabývá k hodnot s pravděpodobnostmi $p_1, \ldots, p_k$,
  Y nabývá m hodnot s pravděpodobnostmi $q_1, \ldots, q_m$.
- Realizace dvojrozměrného náhodného výběru $((x_1, y_1), \ldots, (x_n, y_n))$ obsahuje dvojice realizací náhodných veličin X, Y.
- Zajímají nás opět jen *pozorované četnosti* $n_{ij}, i = 1 \ldots, k, j = 1 \ldots, m$, které bývají uspořádány do tzv. **kontingenční tabulky** s km buňkami.
- Za předpokladu nezávislosti proměnných X, Y je pravděpodobnost výsledku ij rovna $p_i q_j$.

$$T := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - np_i q_j)^2}{np_i q_j} \text{ má přibližně rozdělení } \chi^2(km-1).$$

- Neznámé parametry p_i, q_j odhadneme metodou maximální věrohodnosti jako

$$p_i = \frac{\sum_{j=1}^{m} n_{ij}}{n} \quad \text{a} \quad q_j = \frac{\sum_{i=1}^{k} n_{ij}}{n},$$

  z nichž je ale jen $(k-1) + (m-1)$ nezávislých, takže výsledný počet stupňů volnosti je $(km-1) - (k-1) - (m-1) = (k-1)(m-1)$ a testujeme T na rozdělení $\chi^2((k-1)(m-1))$.
- H_0 zamítáme pro $t > q_{\chi^2((k-1)(m-1))}(1-\alpha)$, tj. pro $p = 1 - F_{\chi^2((k-1)(m-1))}(t) < \alpha$.

Př: χ^2 test nezávislosti

Zadání: Zjistěte, zda má streptomycin vliv na léčbu plicní tuberkulózy, z dat pro dva nezávislé výběry (viz níže): první skupina byla léčena streptomycinem, druhá (kontrolní) skupina dostávala placebo.

Řešení: Použijeme χ^2-test nezávislosti:

	i, X (Lék)			
j, Y (Změna stavu)	1, Streptomycin	2, Placebo	$\sum_i n_{ij}$	$q_j = \frac{1}{n}\sum_i n_{ij}$
1, Významné zlepšení	28 16.45	4 15.55	32	0.299
2, Střední / malé zlepšení	10 11.82	13 11.18	23	0.215
3, Beze změn	2 2.57	3 2.43	5	0.047
4, Střední / malé zhoršení	5 8.74	12 8.26	17	0.159
5, Významné zhoršení	6 6.17	6 5.83	12	0.112
6, Smrt	4 9.25	14 8.75	18	0.168
$\sum_j n_{ij}$	55	52	107	1
$p_i = \frac{1}{n}\sum_j n_{ij}$	0.514	0.486	1	

Očekávaná četnost jevu $X =$ Streptomycin $\wedge$ $Y =$ Význ. zlepšení:

$$np_1 q_1 = n \frac{1}{n} \sum_j n_{1j} \frac{1}{n} \sum_i n_{i1} = \frac{55 \cdot 32}{107} = 16.45$$

Očekávaná četnost jevu $X =$ Placebo $\wedge$ $Y =$ Význ. zhoršení:

$$np_2 q_5 = n \frac{1}{n} \sum_j n_{2j} \frac{1}{n} \sum_i n_{i5} = \frac{52 \cdot 12}{107} = 5.83$$

Test nezávislosti (pokr.)

- Realizace testové statistiky:

$$t = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-np_iq_j)^2}{np_iq_j} = 26.96.$$

- Při $k=2$ a $m=6$ se rozdělení T blíží k $\chi^2((k-1)(m-1)) = \chi^2(5)$.
- Dosažaná hladina významnosti:

$$p = 1 - F_{\chi^2(5)}(26.96) = 5.8\cdot 10^{-5}.$$

Závěr: Zamítáme H_0. Způsob léčby a stav pacienta nejsou nezávislé.

V testu jsme se dopustili chyby (porušení předpokladů). Víte kde?

Př: χ^2 test nezávislosti

Problém: Očekávané četnosti v řádku 3 jsou příliš malé ($np_iq_j < 5$) a v řádku 5 jsou blízko hranici.

Řešení: Spojíme řádky tabulky: (3,4) a (5,6), i když řádky 5 a 6 bychom spojovat nemuseli.

	i, X (Lék)			
j, Y (Změna stavu)	1, Streptomycin	2, Placebo	$\sum_i n_{ij}$	$q_j = \frac{1}{n}\sum_i n_{ij}$
1, Významné zlepšení	28 16.45	4 15.55	32	0.299
2, Střední / malé zlepšení	10 11.82	13 11.18	23	0.215
3, Beze změn	2 2.57	3 2.43	5	0.047
4, Střední / malé zhoršení	5 8.74	12 8.26	17	0.159
5, Významné zhoršení	6 6.17	6 5.83	12	0.112
6, Smrt	4 9.25	14 8.75	18	0.168
$\sum_j n_{ij}$	55	52	107	1
$p_i = \frac{1}{n}\sum_j n_{ij}$	0.514	0.486	1	

	i, X (Lék)			
j, Y (Změna stavu)	1, Streptomycin	2, Placebo	$\sum_i n_{ij}$	$q_j = \frac{1}{n}\sum_i n_{ij}$
1, Významné zlepšení	28 16.45	4 15.55	32	0.299
2, Střední / malé zlepšení	10 11.82	13 11.18	23	0.215
3, Beze změn, Střední / malé zhoršení	7 11.31	15 10.69	22	0.206
4, Významné zhoršení, Smrt	10 15.42	20 14.58	30	0.280
$\sum_j n_{ij}$	55	52	107	1
$p_i = \frac{1}{n}\sum_j n_{ij}$	0.514	0.486	1	

- Realizace testové statistiky:

$$t = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-np_iq_j)^2}{np_iq_j} = 24.57.$$

- Při $k=2$ a $m=4$ se rozdělení T blíží k $\chi^2((k-1)(m-1)) = \chi^2(3)$.
- Dosažaná hladina významnosti: $p = 1 - F_{\chi^2(3)}(24.57) = 1.9\cdot 10^{-5}$.

Závěr: Zamítáme H_0. Způsob léčby a stav pacienta nejsou nezávislé.

Korelace

Korelace $\rho(X, Y)$ náhodných veličin X, Y (s nenulovým rozptylem) je střední hodnota součinu odpovídajících normovaných veličin $\frac{X - \mathrm{E}\,X}{\sigma_X}$ a $\frac{Y - \mathrm{E}\,Y}{\sigma_Y}$:

$$\rho(X, Y) = \frac{\mathrm{E}((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y))}{\sigma_X \sigma_Y} \in \langle -1, 1 \rangle$$

- Korelace je nulová pro nezávislé náhodné veličiny, ale i pro některé jiné, tzv. *nekorelované*.
- Krajní hodnoty ± 1 odpovídají lineární závislosti mezi X, Y.

Na základě dvojrozměrného náhodného výběru $((X_1, Y_1), \ldots, (X_n, Y_n))$ můžeme korealaci odhadnout pomocí **výběrového koeficientu korelace**

$$R_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})(Y_i - \bar{\boldsymbol{Y}})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2\right)\left(\sum_{i=1}^{n}(Y_i - \bar{\boldsymbol{Y}})^2\right)}}.$$

Pro výpočet se často používá ekvivalentní jednoprůchodový vzorec

$$R_{X,Y} = \frac{n \sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left(n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right)}}$$

Test nekorelovanosti dvou normálních rozdělení

Předpoklad: Dvojrozměrná náhodná veličina (X, Y) má dvojrozměrné normální rozdělení, $n \geq 3$.

H_0: Veličiny X a Y jsou nekorelované, tj. $\rho(X, Y) = 0$.

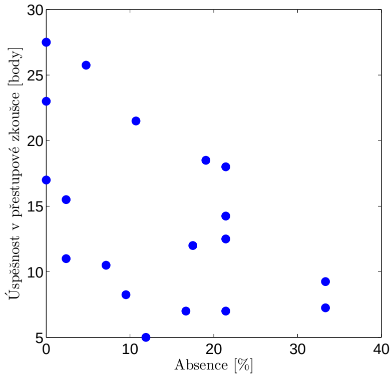
- Testovací statistikou je

$$T = \frac{R_{X,Y}\sqrt{n-2}}{\sqrt{1 - R_{X,Y}^2}}.$$

- Platí-li H_0, T má rozdělení $t(n-2)$.
- Dále postupujeme stejně jako při t-testu.

Příklad: Test nekorelovanosti

Zadání: Na jistém gymnáziu v kvartě v předmětu Matematika byla nasbírána následující data týkající se absence při výuce (X) a úspěšnosti ve zkoušce (Y). Výběr obsahuje 21 studentů (2 body v grafu mají četnost 2), realizace výběrového korelačního koeficientu $r = -0.521$. Ověřte hypotézu, že absence není korelovaná s úspěšností.



Řešení: Test nekorelovanosti.

- Testová statistika

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.521\sqrt{21-2}}{\sqrt{1-0.521^2}} = -2.661.$$

- Dosažená hladina významnosti

$$p = 2(1 - F_{t(n-2)}(|t|)) = 2(1 - F_{t(19)}(2.661)) = 0.015.$$

Závěr: Na hladině významnosti 5 % zamítáme H_0 (ve prospěch H_A, že absence a úspěšnost jsou korelované). Na hladině významnosti 1 % H_0 zamítnout nemůžeme.