Neparametrické testy

Mgr. Jiří Mazurek, PhD.

Co přináší neparametrické testování hypotéz

V případě ordinálních (pořadových) nebo nominálních dat odpovídá na specifické otázky:

1. Existuje významný soulad dané charakteristiky vzorku se zadanou charakteristikou?
2. Existuje významný rozdíl dané charakteristiky mezi 2 (nebo více) vzorky?

Charakteristika - např. medián, zadané pořadí, rozdělení pravděpodobnosti (četnosti) aj.

Neparametrické hypotézy

- Neparametrické hypotézy se netýkají parametrů rozdělení náhodné veličiny, nýbrž jiných statistických vlastností, např. tvaru rozdělení, nezávislosti náhodných veličin a podobně.
- O neparametrických testech se také hovoří obecněji v případech, kdy nejsou splněny některé standardně vyžadované předpoklady pro provedení daného testu. (např. u t-testů jsme požadovali splnění jistých podmínek, aby mohl být daný statistických test realizován – požadovali jsme, aby výběr pocházel z normálního rozdělení.)
- Jsou situace, kdy takový předpoklad splněn není, a pak je otázkou jak postupovat.

Možnost testů

- Existují testy „robustnějšího“ charakteru, kterými lze testovat vlastnosti populace, ze které náhodný výběr pochází, a přitom je třeba splnit pouze podmínky velmi obecného charakteru pro využití těchto testů.
- V takových případech hovoříme rovněž o neparametrických testech, byť jimi můžeme testovat konkrétní podobu parametrů daného rozdělení.
- Pod pojmem neparametrický test budeme zahrnovat statistický test, jenž zkoumá jiné vlastnosti neznámé populace či základního souboru než ty vlastnosti, které se týkají přímo parametrů této populace.

Neparametrické testy hypotéz

- Ad 1) Jednovýběrové testy:

  - Má medián populace s neznámým rozdělením stanovenou hodnoru? (mediánový test)

  - Pochází výběr z populace se zadaným (známým) rozdělením pravděpodobnosti? (Chi-kvadrát test, Kolmogorov-Smirnovův test)

- Ad 2) Dvouvýběrové testy:

  - Mají výběry stejný medián? (mediánový test)

  - Pochází výběry ze stejné populace? (Chi-kvadrát test, Mann-Whitneyův test, Wilcoxonův párový test)

Mediánový test

- hodnoty mediánu (prostřední hodnoty v populaci).
- Pokud jde o populaci, která má tu vlastnost, že její populační průměr se shoduje s mediánem, lze mediánový test využít také jako jednovýběrový t-test.
- Jedinou podmínkou pro použití mediánového testu je předpoklad, že rozdělení četností v populaci je možno popsat distribuční funkcí spojitého typu. Nepožaduje se tedy v tomto případě normální rozdělení jako v případě jednovýběrového t-testu.

Mediánový test - předpoklady

- Označme neznámý medián v populaci symbolem $\tilde{\mu}_0$, rozsah vzorku dat, který je k dispozici, je *n*. Předpokládáme větší rozsah výběru, neboť platnost dále popsaného testu se zpřesňuje s růstem rozsahu *n*.

Mediánový test

1. Nulová hypotéza: H_0: $\tilde{\mu} = \tilde{\mu}_0$

   Alternativní hypotéza: H_1: $\tilde{\mu} \neq \tilde{\mu}_0$

2. Testové kritérium: $T = \frac{|2m - n|}{\sqrt{n}}$,
   - kde m je počet pozorování, která jsou menší než $\tilde{\mu}_0$

3. Kritická hodnota je $K = z_{1-\alpha/2}$
   - $z_{1-\alpha/2}$ je kritická hodnota normovaného normálního rozdělení pro zadanou hladinu významnosti α.

4. Jestliže platí $T \geq K$, potom se H_0 zamítá, jinak se H_0 přijímá.

Mediánový test - poznámky

- $z_{1-\alpha/2}$ je kritická hodnota normovaného normálního rozdělení pro zadanou hladinu významnosti α.
- Je to tedy reálné číslo $z_{1-\alpha/2}$ takové, že pravděpodobnost jeho překročení (nebo dorovnání) je rovna hodnotě $1-\alpha/2$.
- Tuto hodnotu nalezneme buď ve statistických tabulkách normovaného normálního rozdělení N(0,1) nebo pomocí Excelu použitím funkce NORMSINV $(1-\alpha/2)$

Příklad 1

- Budeme testovat hypotézu (na hladině významnosti 0,05), že průměrný plat v jistém podniku je 35 000 Kč. Z 50 zaměstnanců podniku mělo 30 zaměstnanců plat nižší než 35 000 Kč.
- Vypočteme testové kritérium $T = \left|\frac{2m-n}{\sqrt{n}}\right| = \left|\frac{2*20-50}{\sqrt{50}}\right| = 0{,}14.$
- Krit. hodnota (z tabulek): K = 1,96.
- H0 přijímáme.

Testy dobré shody

- Další kategorií testů, které probereme, jsou tzv. testy dobré shody.
- Do této skupiny statistických metod patří řada testů, my se budeme zabývat dvěma z nich, které lze považovat za základní a často využívané při marketingových či sociologických výzkumech.
- První test je zaměřen na testování podoby pravděpodobnostního rozdělení, z něhož pochází náhodný výběr, který je k dispozici.
- Druhý test zkoumá statistickou nezávislost dvou znaků. Protože se v obou případech pracuje s rozdělením chí-kvadrát, pokud jde o rozdělení testového kritéria, hovoří se také o chí-kvadrát testech.

Chi-kvadrát test
(X^2 - test pro 1 výběr)

- Data mohou být nominální (nejslabší požadavek)!
- Testuje se (nulová) hypotéza: výběr pochází z populace se zadaným rozdělením
- Zadané rozdělení je obvykle:

  - diskrétní rozdělení s rozdílnými pravdě- podobnostmi (tzv. test dobré shody)

  - diskrétní rozdělení se stejnými pravdě- podobnostmi (tzv. test nezávislosti)

Test dobré shody

- Pro (Pearsonův) test dobré shody předpokládáme, že výsledky náhodného výběru lze uspořádat do *J* nepřekrývajících se tříd.
- Četnosti v jednotlivých třídách značíme $n_1, n_2 \dots\ n_J$, celkový rozsah náhodného výběru je *n*.
- Testovaná hypotéza spočívá v předpokladu určitého modelu pravděpodobnostního rozdělení, tedy předpokladu pravděpodobností pro každou třídu $p_1, p_2 \dots\ p_J$, součet všech pravděpodobností dává hodnotu 1.
- Test dobré shody spočívá v porovnání naměřených (empirických) četností s četnostmi teoretickými.
- Teoretické četnosti $\psi_1, \psi_2 \dots\ \psi_J$ získáte jako součin odpovídající pravděpodobnosti a rozsahu náhodného výběru: $p_i . n_i$
- Podmínkou použitelnosti testu jsou teoretické četnosti větší než 5.

Postup testu

1. Stanovení hypotézy: $H_0 : p_1 = \pi_1, p_2 = \pi_2, \ldots, p_j = \pi_j$ (slovy: dobrá shoda),
   $H_1 : \exists i; p_i \neq \pi_i$ (negace H_0).
2. Testové kritérium: $G = \sum_{i=1}^{J} \frac{(n_i - \psi_i)^2}{\psi_i}$.
3. Obor přijetí: $\langle 0 , \chi^2_{J-1}(\alpha) \rangle$, kritický obor: $(\chi^2_{J-1}(\alpha) , +\infty)$, kde $\chi^2_{J-1}(\alpha)$ je kritická hodnota χ^2-rozdělení s $df = J-1$.
4. Závěr.

Excel

- V Excelu dostanete kritickou hodnotu pomocí funkce CHIINV.
- Další funkce programu Excel, funkce CHITEST(*Aktuální;Očekávané*) vám umožní spočítat *p* hodnotu testu.
- Argumenty funkce CHITEST jsou naměřené – aktuální hodnoty n_i a pak teoretické – očekávané hodnoty ψ_i. Testové kritérium získáte z *p*-hodnoty pomocí funkce CHIINV.

Příklad 2

- Dodavatel slíbil, že dodávka bude obsahovat 70% výrobků 1. jakosti, 20% druhé jakosti a 10% jakosti třetí.
- Při kontrole dodávky kontroloři náhodně vybrali 100 výrobků a zjistili, že 75 kusů je 1. jakosti, 10 kusů je 2. jakosti a 15 kusů je jakosti třetí.
- Na hladině významnosti 0,05 zjistěte, zda dodavatel dodržel smlouvu.

Příklad 2 – řešení

- V následující tabulce je přehled zadání a výpočet teoretických hodnot. Celkový počet pozorování je n = 100.

Četnost výskytu n_i	**Pravděpodobnost p_i**	**Teoretická četnost $\psi_i = p_i n_i$**
75	0,7	70
10	0,2	20
15	0,1	10

Příklad 2 – dosazení do vzorce

1. Stanovení hypotézy: $H_0 : p_1 = 0{,}7,\ p_2 = 0{,}2,\ p_3 = 0{,}1$ (shoda – smlouva dodržena)
   H_1 : negace H_0 (neshoda – dodavatel šidí).
2. Testové kritérium:
   $$G = \sum_{i=1}^{J} \frac{(n_i - \psi_i)^2}{\psi_i} = \frac{(75-70)^2}{70} + \frac{(10-20)^2}{20} + \frac{(15-10)^2}{10} =$$
   $$= \frac{25}{70} + \frac{100}{20} + \frac{25}{10} = 7{,}857.$$
3. Kritickou hodnotu naleznete takto: CHIINV(0,05;2)=5,991476. Potom obor přijetí: $\langle 0;5{,}99 \rangle$, kritický obor: $(5{,}99;+\infty)$.
4. Závěr: Testové kritérium leží v kritickém oboru. Zamítáme hypotézu dobré shody vzorku s předpokladem. Dodavatel nedodržel smlouvu.

Příklad 2 – výpočet pomocí aplikace EXCEL

- Použijete-li k testování funkci CHITEST, naleznete po dosazení naměřených a teoretických hodnot výsledek p = 0,01967 .
- Toto číslo je menší než zadaná hladina významnosti α=0,05, a tedy zamítáme nulovou hypotézu, dodavatel nedodržel smlouvu.
- Testové kritérium získáte z pravděpodobnosti p pomocí funkce CHIINV, jejíž argumenty budou pravděpodobnost a počet stupňů volnosti.
- Zkontrolujte si, že CHIINV(0,019671;2)=7,857.

Test nezávislosti kvalitativních znaků

- Jednou z aplikací testu dobré shody je testování nezávislosti kvalitativních znaků v kontingenční tabulce.
- Jedná se o *n* náhodných pokusů, které nemají přesné výsledky, ale výsledky určují rozdělení do kategorií.
- Příkladem může být kvalitativní znak úspěch s kategoriemi uspěl/neuspěl nebo znak barva s kategoriemi červená/modrá/zelená.
- Sleduje se více znaků, pro dva znaky *A* a *B* by výsledná tabulka četností (kontingenční tabulka) vypadala takto:

Příklad kontingenční tabulky

Kategorie znaku A/B	B_1	B_2	B_3	…	B_s	Marginální součty
A_1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	…	$n_{1,s}$	$n_{1,\cdot}$
A_2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	…	$n_{2,s}$	$n_{2,\cdot}$
A_3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	…	$n_{3,s}$	$n_{3,\cdot}$
…	…	…	…	…	…	…
A_r	$n_{r,1}$	$n_{r,2}$	$n_{r,3}$	…	$n_{r,s}$	$n_{r,\cdot}$
Marginální součty	$n_{\cdot,1}$	$n_{\cdot,2}$	$n_{\cdot,3}$		$n_{\cdot,s}$	Celkový součet n

- Počet kategorií znaku *A* označme *r* a toto číslo současně označuje počet řádků tabulky.
- Počet kategorií znaku *B* označme *s* a tento počet je v tabulce vyjádřen počtem sloupců.
- Celkový počet pozorování je *n*.
- Test nezávislosti se může provádět, jen když je každá z četností $n_{i,j}$ větší než 4.

Teoretické hodnoty

- Chceme-li použít k testování nezávislosti znaků *A* a *B* test dobré shody, potřebujeme mít k dispozici teoretické hodnoty, které pak následně porovnáme s hodnotami naměřenými.
- Teoretické četnosti jsou hodnoty, které by byly v tabulce, kdyby oba znaky byly nezávislé a současně by marginální četnosti zůstaly stejné jak u empirických hodnot.
- Teoretické hodnoty lze vypočítat ze vztahu:

$$\psi_{i,j} = \frac{n_{i,\cdot} \cdot n_{\cdot,j}}{n}$$

Tabulka teoretických četností

Kategorie znaku A/B	B_1	…	B_s	**Marginální součty**
A_1	$\psi_{1,1} = \frac{n_{1,\cdot} \cdot n_{\cdot,1}}{n}$	…	$\psi_{1,s} = \frac{n_{1,\cdot} \cdot n_{\cdot,s}}{n}$	$n_{1,\cdot}$
A_2	$\psi_{2,1} = \frac{n_{2,\cdot} \cdot n_{\cdot,1}}{n}$	…	$\psi_{2,s} = \frac{n_{2,\cdot} \cdot n_{\cdot,s}}{n}$	$n_{2,\cdot}$
…	…	…	…	…
A_r	$\psi_{r,1} = \frac{n_{r,\cdot} \cdot n_{\cdot,1}}{n}$	…	$\psi_{r,s} = \frac{n_{r,\cdot} \cdot n_{\cdot,s}}{n}$	$n_{r,\cdot}$
Marginální součty	$n_{\cdot,1}$	…	$n_{\cdot,s}$	**Celkový součet n**

Postup testování

1. Stanovení hypotézy: $H_0 : p_{i,j} = p_i \cdot p_j$, $i = 1, \dots, r$, $j = 1, \dots, s$ (nezávislost znaků),
   H_1 : negace H_0 .
2. Testové kritérium: $G = \sum_{i=1}^{J} \frac{(n_i - \psi_i)^2}{\psi_i}$.
3. Obor přijetí: $\left\langle 0 , \chi^2_{(r-1)(s-1)}(\alpha) \right\rangle$, kritický obor: $(\chi^2_{(r-1)(s-1)}(\alpha) , +\infty)$, kde $\chi^2_{(r-1)(s-1)}(\alpha)$ je kritická hodnota chi-rozdělení s $df = (r-1)(s-1)$.
4. Závěr.

Příklad 3

- Bylo zkoumáno nákupní chování mužů a žen, které se týkalo návštěv obchodního domu Karolína Ostrava. V Tabulce níže je uveden počet žen a mužů, kteří v Karolíně pravidelně nakupují.

	ANO	**NE**	
Muži	**12**	**34**	46
Ženy	**25**	**16**	41
	37	50	87

Zjistěte na hladině významnosti alfa = 0,05, zda se nákupní zvyklosti mužů a žen liší.

Příklad 3 - pokračovaní

- **V našem případě se kontingenční tabulka nazývá 4polní tabulka.**

	Kat. 1	Kat. 2	
Kat. 1	A	B	
Kat. 2	C	D	

Testové kritérium se vypočte takto: $G = \frac{n(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)}$

Kritickou hodnotu najdeme v tabulkách pro chí-kvadrát rozdělení s 1 stupněm volnosti.

Příklad 3 - pokračování

- **Bylo zkoumáno nákupní chování mužů a žen, které se týkalo návštěv obchodního domu Karolína Ostrava. V Tabulce níže je uveden počet žen a mužů, kteří v Karolíně pravidelně nakupují. Na hladině významnosti alfa = 0.05 otestujte hypotézu, že se nákupní chování mužů a žen neliší.**

	ANO	NE	
Muži	12	34	46
Ženy	25	16	41
	37	50	87

H0: Nákupní chování mužů a žen se neliší.

H1: liší se.

$$G = \frac{87(12*16-25*34)^2}{(12+34)(25+16)(12+25)(34+16)} = 10.79;\ \mathrm{K}\ (df = 1,\ alfa = 0.05) = 3.84.$$

H0 zamítáme.

Příklad 4

- Vysoká škola zjišťovala, jestli existuje závislost mezi známkami z matematiky a mikroekonomie.
- Do výzkumu zahrnula 100 studentů druhých ročníků, kteří měli obě zkoušky za sebou. Výsledky jsou uspořádány v následující kontingenční tabulce.
- Na hladině významnosti 0,05 určete, zda lze pozorovat závislost mezi těmito dvěma předměty.

	Mikroekonomie			
Známka	**1**	**2**	**3**	
	7	5	8	*20*
	5	11	12	*28*
Matematika	14	19	19	*52*
	26	*35*	*39*	*100*

Příklad 4 - řešení

- H0: Mezi známkami z matematiky a mikroekonomie není závislost.

  H1: je závislost.
- Testové kritérium G:
- $G = 100\left(\frac{7^2}{20*26} + \frac{5^2}{20*35} + \frac{8^2}{20*39} + \frac{5^2}{26*28} + \frac{11^2}{35*28} + \frac{12^2}{39*28} + \frac{14^2}{26*52} + \frac{19^2}{35*52} + \frac{19^2}{39*52} - 1\right) = 2.3$
- K (df = 4, alfa = 0.05) = 9.49.
- H0 přijímám.

	Mikroekonomie			
Známka	**1**	**2**	**3**	
	7	5	8	*20*
	5	11	12	*28*
Matematika	14	19	19	*52*
	26	*35*	*39*	*100*

Děkuji za pozornost