Přednáška 11 – Testy pro diskrétní data

Z testů hypotéz, které jsme zatím probrali, byla naprostá většina určena pro výběry spojité náhodné veličiny, i když u některých jsme si poznamenali, že mohou sloužit i pro diskrétní výběry. Dnes se zaměříme výhradně na testy určené pro diskrétní data.

Testy pro diskrétní data budeme rozlišovat podle toho, s kolika výběry pracujeme. V následující tabulce jsou uvedené testy, se kterými budeme pracovat (je také dostupná na webu na odkazu Jak zvolit test hypotéz).

Jeden výběr	Dva výběry
Test podílu – **prop_test** Předpoklady: $n > 30$, $np \geq 5$, $n(1-p) \geq 5$ H_0: $p = p_0 (>, <)$ levo-, pravo-, oboustranný	Test o shodě dvou podílů – **prop_test_2** Předpoklady: nepárové výběry $n > 30$, $np \geq 5$, $n(1-p) \geq 5$ H_0: $p_1 = p_2 (>, <)$ levo-, pravo-, oboustranný
χ^2 test dobré shody – **chisquare_test** Předpoklady: všechny četnosti > 2, alespoň 80% četností > 5 H_0: výběr má teoretické rozdělení	McNemarův test – **mcnemar_test** Předpoklady: binární data, párové výběry, kontingenční tabulka H_0: četnosti jsou stejné
	χ^2 test nezávislosti – **chisquare_test_i** Předpoklady: všechny četnosti > 2, alespoň 80% četností > 5, kontingenční tabulka H_0: jsou nezávislé
	Fisherův exaktní test Předpoklady: nominální data H_0: jsou nezávislé
	Gamma koeficient (Goodmanovo-Kruskalovo gamma) Předpoklady: ordinální data, kategorické rozdělení H_0: jsou nezávislé
	Yule's Q koeficient Předpoklady: ordinální data, alternativní rozdělení H_0: jsou nezávislé

Probereme každý z testů podrobně.

Test podílu pro jeden výběr

Připomeňme si, že výběrový podíl je charakteristika výběru, která se počítá takto (viz přednáška 5):

$$p = \frac{n^+}{n},$$

kde n^+ je počet úspěchů ve výběru a n je počet dat, což znamená, že podíl p je rovnou i pravděpodobností úspěchu. Například, zajímá nás podíl studentů s modrýma očima na dopravní fakultě. Modré oči jsou v tomto případě úspěch, jakékoliv jiné oči jsou neúspěch. Vezmeme náhodný výběr studentů z fakulty. Spočítáme kolik máme studentů s modrýma očima - toto bude počet úspěchů. Vydělíme ho počtem studentů ve výběru a dostaneme podíl modrookých studentů, což je pravděpodobnost toho, že pokud náhodně potkáme nějakého studenta, bude mít modré oči.

Test podílu (**prop_test**) použijeme v případě, když diskrétní náhodná veličina, kterou pozorujeme, má alternativní rozdělení, čili může nabývat dvou možných hodnot:

$$x \in \{0,1\}, \quad \text{kde 0 je neúspěch, 1 úspěch.}$$

Předpoklady k použití testu podílu:

- $n > 30$,
- $np \geq 5$, $n(1-p) \geq 5$, kde p je pravděpodobnost úspěchu. Například,

$$np = 100 * 0.2 = 20 > 5 \rightarrow \text{můžeme použít test,}$$

$$np = 100 * 0.8 = 80 > 5 \rightarrow \text{můžeme použít test,}$$

$$np = 1000 * 0.0001 = 0.1 < 5 \rightarrow \text{nemůžeme.}$$

Obecně nulová hypotéza testu podílu tvrdí

$$H_0 : p = p_0, \quad \text{tj., podíl se rovná předpokládané hodnotě } p_0.$$

Alternativní hypotéza ji popírá:

$$H_A : p \neq (>, <) p_0, \quad \text{– určuje směr testu.}$$

V případě, že $n > 30$, má podíl aproximativně $N(0,1)$ a pro tento test je statistika

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1).$$

Příklad: Na úseku silnice s maximální povolenou rychlostí 80 km/h kontrolujeme rychlost vozidel. Zaznamenali jsme následující rychlosti (viz tabulka). Testujeme hypotézu, že podíl řidičů, kteří překračují rychlost o více než 3km, je menší než 20%.

78	86	65	93	92	85	76	79	...$\rightarrow n > 30$

Řešení: V tomto případě diskrétní náhodná veličina, která má dvě možné hodnoty, je překročení povolené rychlosti o více než 3 km/h. Její dvě možné realizace jsou: úspěch – rychlost je překročena o více než 3 km/h (tj., jsou to všechny rychlosti větší než 83 km/h) a neúspěch – rychlost není překročena (jsou to všechny rychlosti ≤ 83).

Potřebujeme otestovat, zda je podíl řidičů, kteří rychlost překročili o více než 3 km/h, menší než 20% – použijeme tím pádem test podílu (**prop_test**). Pamatujeme si, že podíl je vlastně pravděpodobnost úspěchu, což znamená, že 20% je pravděpodobnost 0.2. Řekneme si nulovou hypotézu :

$$H_0 : p = 0.2 \quad \text{nebo} \quad p < 0.2 \; – \; \text{podle tvrzení,}$$

slovně: podíl řidičů překračujících rychlost o více než 3 km/h se rovná nebo je menší než 0.2. Alternativní hypotéza je opačné tvrzení:

$$H_A : p > 0.2 \; \text{– podíl je větší než 0.2, což určuje, že to je pravostranný test.}$$

Pro použití testu potřebujeme ještě spočítat výběrový podíl p (což je vidět u statistiky testu). Jelikož úspěch je každé překročení povolené rychlosti o více než 3 km/h, takže to je každá rychlost vyšší než 83 km/h. To znamená, že musíme spočítat kolik takových událostí máme – to je počet úspěchů. Dále počet úspěchů vydělíme počtem dat, tj.,:

$$p = \frac{\text{počet úspěchů}}{\text{počet dat}} = \frac{\text{počet rychlostí} > 83}{\text{počet řidičů}} = \frac{1(86) + 1(93) + 1(92) + 1(85) + ...}{n}.$$

Dále následuje obvyklý postup testování.

χ^2 test dobré shody

χ^2 test dobré shody je test rozdělení, který už jsme používali pro testování normality (viz přednáška 7 a tabulka Jak zvolit test hypotéz). Můžeme tento test použít i pro testování libovolného diskrétního rozdělení výběru, případně pro testování, zda mají dva výběry stejné rozdělení. Připomeňme si, že pro použití tohoto testu by měly být všechny četnosti naměřených hodnot větší než 2 a alespoň 80% četností by mělo být větší než 5. χ^2 test dobré shody používá statistiku

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

kde O_i – pozorované četnosti a E_i – očekávané četnosti. Nulová a alternativní hypotézy jsou:

$$H_0 : \text{výběr pochází z teoretického rozdělení},$$

$$H_A : \text{nepochází z teoretického rozdělení}$$

Test je pouze pravostranný. Ukážeme si, jak použijeme χ^2 test dobré shody, tentokrát pro testování diskrétního rozdělení výběru.

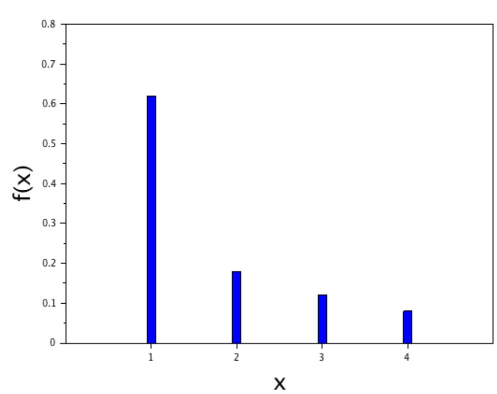
Příklad: Výrobce automobilů nabízí k prodeji v síti svých autorizovaných prodejců vozy s následujícími typy karoserie: hatchback, sedan, kombi a coupé. Vozy se prodávají v následujícím rozdělení:

kombi 62%, hatchback 18%, sedan 12%, coupé 8%.

V rámci rozšíření sítě byla otevřena nová pobočka s autorizovaným prodejcem, který zatím prodal 120 vozů s karoserií kombi, 40 vozů hatchback, 18 sedanů a 22 coupé. Zajímá nás, zda se prodej na nové pobočce neliší od ostatních prodejců.

Řešení: Máme zde diskrétní náhodnou veličinu prodej$\in$ {kombi, hatchback, sedan, coupé}, tj., má 4 možné realizace. Je dáno rozdělení této diskrétní veličiny, tj., pravděpodobnostní funkce f(prodej), kterou můžeme zobrazit jako tabulku nebo graf:

	kombi	hatchback	sedan	coupé
f(prodej)	0.62	0.18	0.12	0.08



Toto je teoretické rozdělení, na shodu s kterým potřebujeme otestovat prodej u nového prodejce. Použijeme k tomu χ^2 test dobré shody (**chisquare_test**). Nulová hypotéza zní:

$$H_0 : \text{prodej na nové pobočce má stejné rozdělení jako u ostatních prodejců v siti},$$

Alternativní hypotéza $\quad H_A$: nemá stejné rozdělení.

Statistika testu používá pozorované četnosti O_i a očekávané četnosti E_i. Pozorované četnosti jsou náš výběr

$$O = [120 \quad 40 \quad 18 \quad 22],$$

tj., data, která jsme naměřili na nové pobočce. Potřebujeme ještě spočítat E_i, tj., četnosti, které bychom očekávali v případě shody rozdělení. Pro jejich výpočet vynásobíme pravděpodobnosti teoretického rozdělení celkovým počtem naměřených dat, který spočteme takto:

$$120 + 40 + 18 + 22 = 200.$$

Dále očekávané četnosti jsou:

$$E = [0.62 * 200 \quad 0.18 * 200 \quad 0.12 * 200 \quad 0.08 * 200] = [124 \quad 36 \quad 24 \quad 16].$$

Četnosti O a E zadáme do testu **chisquare_test**. Výsledná p-hodnota se rovná $0.2286 > 0.05$, takže nezamítáme nulovou hypotézu, že se prodej na nové pobočce neliší od ostatních prodejců.

Test o shodě dvou podílů

Test o shodě dvou podílů (**prop_test_2**) použijeme v případě dvou výběrů diskrétních náhodných veličin, které mohou nabývat dvou možných hodnot – úspěch a neúspěch. Předpoklady použití testu o shodě dvou podílů jsou stejné jako v případě testu podílu pro jeden výběr. Výběry nemusí být párové.

Nulová a alternativní hypotézy jsou také podobné – rozdíl je jenom v tom, že teď pracujeme se dvěma výběrovými podíly. Proto obecně nulová hypotéza testu o shodě dvou podílů tvrdí, že

$$H_0 : p_1 = p_2, \quad \text{tj., podíly jsou stejné.}$$

Alternativní hypotéza ji popírá:

$$H_A : p_1 \neq (>, <) p_2, \quad \text{– určuje směr testu.}$$

Test používá statistiku

$$T = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1).$$

Příklad: Policie ČR tvrdí, že na silnici s maximální povolenou rychlostí 80 km/h směrem z Prahy je menší podíl řidičů překračujících povolenou rychlost než směrem do Prahy. Data jsou v tabulce. Testujeme toto tvrzení na hladině významnosti 0.05.

z Prahy	93	86	85	93	92	85	86	79	$\ldots \to n_1 > 30$
do Prahy	78	96	75	83	97	105	81	79	$\ldots \to n_2 > 30$

Řešení: Podobně jako v příkladě na test podílu, jsou tady dvě diskrétní náhodné veličiny: překročení povolené rychlosti na silnici směrem z Prahy a směrem do Prahy. Za úspěch považujeme překročení, tj., rychlost vyšší než 80 km/h. Použijeme test o shodě dvou podílů (**prop_test_2**). Řekneme si nulovou hypotézu:

$$H_0 : p_z = p_{do} \text{ nebo } p_z < p_{do} \quad \text{– podle tvrzení,}$$

slovně: podíl řidičů překračujících povolenou rychlost na silnici směrem z Prahy je menší než směrem do Prahy. Alternativní hypotéza je opačné tvrzení:

$$H_A : p_z > p_{do} \quad \text{– tj., je větší, takže to je pravostranný test.}$$

Určíme oba výběrové podíly:

$$p_z = \frac{\text{rychlosti z Prahy} > 80}{n_1} = \frac{1(93) + 1(86) + 1(85) + 1(93) + 1(92) + 1(85) + 1(86) + \ldots}{n_1},$$

$$p_{do} = \frac{\text{rychlosti do Prahy} > 80}{n_2} = \frac{1(96) + 1(83) + 1(97) + 1(105) + 1(81) + \dots}{n_2},$$

které využijeme pro funkci **prop_test_2**.

McNemarův test

McNemarův test (**mcnemar_test**) použijeme v případě, když máme dva párové výběry diskrétní náhodné veličiny s binárními hodnotami úspěch a neúspěch, například ano a ne. Nejvíce se test osvědčil při testování efektu nějakého zákroku, léku, aj. – porovnává četnosti před a po zákroku, tj., zda nastala nějaká změna po zákroku a testuje shodu výběrů. K použití testu potřebujeme data ve tvaru kontingenční tabulky, která tady bude velikosti 2×2:

před \ po	ne	ano
ne	a	b
ano	c	d

$$\text{Nulová hypotéza testu říká: } H_0 : b = c, \text{četnosti jsou stejné, tj., není žádná změna po zákroku,}$$

$$\text{Alternativní hypotéza } H_A : b \neq c, \text{četnosti nejsou stejné, tj., je změna po zákroku,}$$

McNemarův test používá statistiku

$$T = \frac{(b-c)^2}{b+c} \sim \chi^2\text{-rozdělení.}$$

Test je pouze pravostranný.

Příklad: Na přechodu pro chodce namontovali nový semafor. Zeptali jsme se 15 lidí na jejich spokojenost s přechodem před instalací semaforu a po instalaci. Odpovědi NE jako 1 a ANO jako 2 máme v tabulce. Testujeme tvrzení, že instalace semaforu nepřispěla ke spokojenosti občanů.

před instalací	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
po instalaci	1	2	1	2	2	2	2	2	2	1	2	2	2	1	2

Řešení: Máme tady diskrétní náhodnou veličinu spokojenost $\in$ {NE=1, ANO=2}, kterou jsme pozorovali před instalací semaforu a po ní, tj., máme dva párové výběry. Vytvoříme kontingenční tabulku, do které napíšeme četnosti pro všechny kombinace hodnot, tj., 1-1, 1-2, 2-1, 2-2 (viz cvičení 10):

před \ po	ne	ano
ne	a=4	b=9
ano	c=0	d=2

Zajímají nás četnosti, které ukazují změnu stavu: před instalací nebyli spokojeni, ale po instalaci jsou – b, nebo před instalací byli spokojeni, ale po instalaci nejsou – c.

$$\text{Nulová hypotéza: } H_0 : b = c \quad \text{– četnosti jsou stejné,}$$

slovně: spokojenost respondentů je stejná před instalací semaforu a po ní, tj., semafor nemá vliv na spokojenost respondentů.

$$\text{Alternativní hypotéza } H_A : b \neq c \quad \text{– četnosti nejsou stejné, tj., semafor má vliv.}$$

Použijeme McNemarův test (**mcnemar_test**). P-hodnota= $0.0026998 < 0.05$, takže zamítáme nulovou hypotézu, že se spokojenost respondentů nezměnila po instalaci semaforu.

Dále k testům pro diskrétní veličiny patří testy nezávislosti, které jsme probírali minulý týden (viz přednáška 10 a tabulka Jak zvolit test hypotéz na webu).

Touto přednáškou končí teoretická část materiálů ze statistiky.