EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Cvičné příklady ze statistických testů

Cvičné příklady

6BZST1

Základy statistiky

doc. RNDr. Lenka Komárková, Ph.D.

2018

Rozvoj vzdělávací a dalších činností a podpora kvality na VŠE v Praze, r. č. CZ.02.2.69/0.0/0.0/16_015/0002342

Cvičné datové soubory

Cvičné datové soubory Salaries.xlsx, Domy1987.xlsx a Nemoci.xlsx jsou k dispozici ke stažení v univerzitním studijním informačním systému InSIS, a to v dokumentové složce k předmětu.

První dva datové soubory Salaries.xlsx a Domy1987.xlsx byly podrobně představeny v rámci cvičných příkladů č. 1. Proměnné v datovém souboru Nemoci.xlsx jsou popsány níže.

Data Nemoci.xlsx

Zdroj dat: nasimulovaná data

Management nejmenované firmy potřeboval zhodnotit nemocnost svých zaměstnanců a zjistit jejich názor na novelu zákona ohledně výplaty nemocenského.

- **id** – identifikační číslo zaměstnance;
- **nazor** – názor na novelu zákona *(1 – líbí/2 – nelíbí/3 – je mi to jedno);*
- **vek** – věk zaměstnance (v letech);
- **nemoc2009** – doba strávená na nemocenské v r. 2009 (ve dnech);
- **nemoc2010** – doba strávená na nemocenské v r. 2010 (ve dnech);
- **pohlavi** – pohlaví zaměstnance *(0 - žena/1 - muž)*.

Zadání

Příklad 1 (Salaries)

Stáhněte si data „Salaries.xlsx“ z InSIS a načtěte je pomocí MS Excel.

a) Určete 95% interval spolehlivosti pro populační proporci akademických pracovníků na amerických vysokých školách bez profesorského titulu.

b) Otestujte na 5% hladině významnosti, zda tato proporce může být **30 %**.

- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
- Zvolte vhodný test a ověřte jeho předpoklady.
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Interpretujte výsledek provedeného testu a porovnejte ho s výše sestaveným intervalem spolehlivosti.

Příklad 2 (Salaries)

a) Ověřte graficky, zda **plat docentů** na amerických univerzitách se řídí normálním rozdělením.

b) Stanovte bodový odhad pro průměrný plat docentů pracujících na amerických univerzitách. Odhadněte i směrodatnou chybu odhadu. Jaký má vztah k výběrové směrodatné odchylce?

c) Sestavte intervalový odhad, který s 95% pravděpodobností pokryje průměrný plat docentů na amerických univerzitách.

d) Sestavte intervalový odhad, který pokryje výši platů pro 95% docentů pracujících na amerických univerzitách.

Příklad 3 (Salaries)

a) Rozhodněte pomocí vhodného testu, zda výše průměrného platu docentů na amerických univerzitách je **90 tis. USD**. Pro rozhodnutí použijte hladinu významnosti 5%; resp. 1 %.

- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
- Zvolte vhodný test a ověřte jeho předpoklady.
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Interpretujte výsledek provedeného testu.

b) Na 5% hladině významnosti rozhodněte, zda výše průměrného platu docentů na amerických univerzitách přesahuje 90 tis. USD. Sestavte i odpovídající interval spolehlivosti.

Příklad 4

Na začátku roku byl u 200 respondentů proveden průzkum zájmu o typ letošní letní dovolené. Výsledky průzkumu lze nalézt v níže uvedené kontingenční tabulce. Na základě provedeného průzkumu otestujte, zda v letošním roce podíl zájemců o tuzemskou a zahraniční letní dovolenou bude v poměru 1:2. Rozhodnutí učiňte na 5% hladině významnosti.

Pohlaví	tuzemská	zahraniční
muž	30	50
žena	30	90

Příklad 5 (Salaries)

a) Pomocí vhodného grafu porovnejte rozložení akademických hodností mezi muži a ženami.

b) Pomocí vhodného testu (na 5% hladině významnosti) rozhodněte, zda akademická hodnost závisí na pohlaví.

- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
- Zvolte vhodný test a ověřte jeho předpoklady.
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Interpretujte výsledek provedeného testu (při interpretaci využijte i vytvořený graf).

Příklad 6 (Salaries)

a) Pomocí vhodného testu rozhodněte, zda platová situace žen a mužů pracujících na amerických univerzitách se liší. Pro rozhodnutí použijte hladinu významnosti 5 %. Porovnejte platovou situaci žen a mužů.

- Vyberte vhodný test a ověřte jeho předpoklady.
- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
  Jakou populační charakteristiku jste vybrali pro srovnání?
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Sestavte odpovídající interval spolehlivosti, pokud to lze.
- Interpretujte výsledek provedeného testu (případně společně s intervalem spolehlivosti).

b) Jak se Vaše výsledky změní, pokud byste chtěli testovat, že platová situace žen oproti mužům je v akademické sféře horší.

Příklad 7 (Salaries)

a) Otestuje, zda poměrné zastoupení žen je na katedrách s teoretickým a aplikačním zaměřením stejné. Použijte α = 5 %.

b) Sestavte i odpovídající interval spolehlivosti a interpretujte ho.

Příklad 8

Podnik uspořádal v rámci dalšího vzdělávání svých zaměstnanců školení výpočetní techniky pro technickohospodářské pracovníky (THP) zaměřené na MS Excel. Provedl i průzkum na vzorku 70 THP o jeho pravidelném využívání. Výsledky tohoto průzkumu lze nalézt v níže uvedené tabulce:

Před školením	**Po školení**	
	Používá	Nepoužívá
Používá	28	4
Nepoužívá	23	15

a) Odhadněte podíl THP používajících Excel pravidelně ve své práci před školením, resp. po školení.

b) Lze prokázat, že školení ovlivnilo podíl THP používajících tento software pravidelně ve své práci? Pro rozhodnutí použijte hladinu významnosti 5 %.

- Vyberte vhodný test a ověřte jeho předpoklady.
- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Interpretujte výsledek provedeného testu.

Příklad 9 (Salaries)

Liší se platová situace mezi odbornými asistenty, docenty a profesory? Pro rozhodnutí použijte 5% hladinu významnosti.

- Použijte vhodné tabulky a grafy z popisné statistiky k dané analýze.
- Vyberte vhodný test a ověřte jeho předpoklady.
- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně. Jakou populační charakteristiku jste vybrali pro srovnání?
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Interpretujte výsledek provedeného testu (případně společně s popisnou statistikou a výsledky z dvouvýběrových testů).

Příklad 10 (Domy1987)

Stáhněte si data „Domy1987.xlsx“ z InSIS a načtěte je pomocí MS Excel.

a) Zjistěte, zda lze na 5% hladině významnosti prokázat, že přítomnost klimatizace má vliv na průměrnou cenu domu.

b) Zjistěte, zda lze na 5% hladině významnosti prokázat, že přítomnost klimatizace vede k nárůstu průměrné ceny domu **o více jak 24 tis. CAD**.

Příklad 11 (Domy1987)

Rozhodněte, zda lze na 5% hladině významnosti prokázat, že průměrná cena nemovitostí ve Windsoru je **nižší než 70 tis. CAD.** Proveďte test a sestavte i odpovídající interval spolehlivosti.

Příklad 12 (Nemoci)

Ověřte, zda průměrná nemocnost zaměstnanců se ve sledovaných letech 2009 a 2010 statisticky významně lišila (α=0,05).

- Vyberte vhodný test a ověřte jeho předpoklady.
- Zformulujte nulovou a alternativní hypotézu symbolicky i slovně.
- Určete hodnotu testové statistiky a dosaženou hladinu testu.
- Sestavte odpovídající interval spolehlivosti.
- Interpretujte výsledek provedeného testu (případně společně s intervalem spolehlivosti).

Řešení

Příklad 1

Jednovýběrový problém o proporci

a) **Interval spolehlivosti**
funkce **CONFIDENCE.NORM,** nebo doplněk **Analýza dat**, *kdy je potřeba kódováním (0 –profesoři, 1- ostatní) převést kategoriální proměnnou na číselnou, pro kterou pak lze vypočítat 95% interval spolehlivosti pro populační průměr*
Výsledek: (0,286; 0,376)

b) **Test (Chí-kvadrát test dobré shody** dává stejné výsledky jako **jednovýběrový test o proporci)**
funkce **CHISQ.TEST(pozorované, očekávané)**
předem je potřeba si vytvořit sloupeček s pozorovanými a očekávanými četnostmi

pozorovane	ocekavane	**p-hodnota**
131	119,1	**0,1924748**
266	277,9	

Příklad 2

Rozdíl mezi směrodatnou odchylkou (SD) a směrodatnou chybou (SE)

a) krabicový diagram a histogram – zkontrolujeme, zda se jedná o symetrické rozdělení bez odlehlých hodnot
b) **Průměr**: 93876,4; **Směrodatná odchylka:** 13831,7; **Směrodatná chyba:** 1729,0
c) 95% interval spolehlivosti: **průměr +- 2 směrodatné chyby … (**90418,5; 97334,4)
d) Interval pro rozmezí 95% hodnot: **průměr +- 2 směrodatné odchylky … (**66213,0; 121539,8**)**

Příklad 3

Jednovýběrový Studentův t-test

Testovou statistiku spočítáme podle vzorečku:

T = (průměr – hyp. hodnota) / směrodatná chyba = 2,24206

a) **p-hodnota (oboustr.)** se pak spočítá pomocí funkce **T.DIST.2T(testová statistika; počet dat - 1): T.DIST.2T(2,24206;63) … p=0,028**
b) **p-hodnota (jednostr. typu >)** se spočítá pomocí **T.DIST.RT(testová statistika; počet dat - 1): T.DIST.RT(2,24206;63) … p=0,014**
Odpovídající 95% jednostranný interval je (90990,1; ∞)

Příklad 4

Chí-kvadrát test dobré shody

pozorovane	ocekavane	p-hodnota
60	66,6666667	**0,3173105**
140	133,3333333	

Příklad 5

Chí-kvadrát test nezávislosti

a) (skládaný) sloupcový graf
b) **Pearsonův chí-kvadrát test nezávislosti**
pro účely testu je nutné spočítat očekávané četnosti, které se počítají pomocí okrajových četností: *„odpovídající řádek · odpovídající sloupec / počet dat“*

Očekávané	AssocProf	AsstProf	Prof	**Celkem**
Female	6,29	6,58	26,13	**39**
Male	57,71	60,42	239,87	**358**
Celkem	**64**	**67**	**266**	**397**

funkce **CHISQ.TEST(pozorované, očekávané)**

p-hodnota: 0,014

Příklad 6

Dvouvýběrový t-test

a) Oboustranná verze
   I. funkce **T.TEST(1.skupina, 2. skupina, 2, typ)**
   a na předpoklad homoskedasticity **F.TEST(1. skupina, 2.skupina)**
   typ: 2 – dvouvýběrový homoskedastický (Studentův), 3 – dvouvýběrový heteroskedastický (Welchův)
   Fisherův-Snedecorův F-test: p = 0,231**, Studentův test: p = 0,006** (Welchův test: p = 0,003)
   II. doplněk **Analýza dat**
   - **„dvouvýběrový F-test pro rozptyl“** (pozor dělá jednostranný) p = 0,115
   - **„dvouvýběrový t-test s rovností rozptylů“** k dispozici ve výstupu jak oboustranná, tak jednostranná verze testu
   - **„dvouvýběrový t-test s nerovností rozptylů“**
b) Jednostranná verze
   T.TEST(1.skupina, 2. skupina, 1, 2) … p = 0,003

Příklad 7

Dvouvýběrový problém o proporcích

Pearsonův chí-kvadrát test nezávislosti

Očekávané	Female	Male	Celkem
A	17,78086	163,2191	181
B	21,21914	194,7809	216
Celkem	39	358	397

funkce **CHISQ.TEST(pozorované, očekávané)**

výsledná p-hodnota: 0,941

Příklad 8

McNemarův test

Testová statistika dle vzorce: (4-23)^2 / (4+23) =**13,37**

p-hodnota: CHISQ.DIST.RT(testová statistika; 1) ... **p=0,000256**

Příklad 9

ANOVA test

Provedeme přes doplněk *Analýza dat „Anova: jeden faktor"*

Výstup v Excelu:

Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
OA	67	5411991	80775,99	66816117
DOC	64	6008092	93876,44	1,91E+08
PROF	266	33721381	126772,1	7,68E+08

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	***F***	***Hodnota P***	*F krit*
Mezi výběry	1,43E+11	2	7,16E+10	**128,2174**	**1,29305E-43**	3,018626
Všechny výběry	2,2E+11	394	5,59E+08			
Celkem	3,63E+11	396				

Příklad 10

Dvouvýběrový Welchův t-test: různé typy alternativ

1. skupina: domy s klimatizací, 2. skupina: domy bez klimatizace

a) $H_0: \mu_1 - \mu_2 = 0$ vs. $H_1: \mu_1 - \mu_2 \neq 0$, výsledek: **$p = 1{,}93 \cdot 10^{-22}$**

b) $H_0: \mu_1 - \mu_2 = 24000$ vs. $H_1: \mu_1 - \mu_2 > 24000$, **výsledek: p=0,206**

Příklad 11

Jednovýběrový t-test s jednostrannou alternativou

Testová statistika: T = (68121,5971 – 70000) / 1142,7688 = **-1,64373**

P-hodnota v Excelu přes funkci T.DIST(-1,64373;545;PRAVDA): **p = 0,050404**

Příklad 12

Párový t-test

přes doplněk Analýza dat „douvýběrový párový t-test na střední hodnotu“

Hypotetický rozdíl středních hodnot: **0**

Číst výsledky pro oboustrannou alternativu: **T = -13,55 a p < 0,00001**

Odpovídající interval: **(-17,6; -13,1)**

je potřeba udělat rozdíl těch dvou sloupců *nemoc2009 - nemoc2010*, a pak sestavit 95% oboustranný interval spolehlivosti v doplňku Analýza dat – Popisná statistika