Cvičení ze statistiky

Filip Děchtěrenko

ZS 2012/2013

Cvičení ze statistiky

- Pondělí 16:40, C328
- http://www.ms.mff.cuni.cz/~dechf7am
- Praktické zaměření

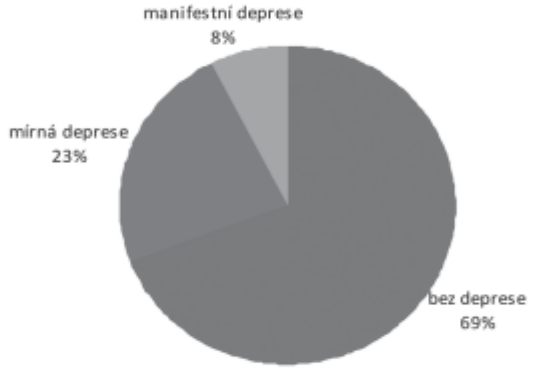
Proč potřebuji statistiku, když chci dělat …(doplň)?

Porozumění článků

- Články jsou důležité, protože publikace jsou zastaralé už v době vydání

votním stavu. Významně vyšší celkové skóry Logo-testu (nižší míru smysluplnosti) měli senioři hospitalizovaní v LDN, NNP a na ošetřovatelských odděleních DS v porovnání s těmi, kteří žili doma s někým ($p < 0,001$) a na běžných odděleních domovů pro seniory ($p < 0,001$) (výsledky uvádíme po Bonferroniho korekci). Skóry těch, kdo žili doma osaměle se od institucionalizovaných v LDN, NNP a na ošetřovatelských odděleních DS významně neodlišovaly ($p = 0,085$).

V souboru bylo 69 % osob bez deprese, u 23,4 % respondentů jsme zjistili mírnou depresi a u 7,5 % zkoumaných seniorů manifestní depresi (graf 2).



Graf 2 Výskyt deprese u respondentů

Výsledky měření deprese ve vztahu k demografickým charakteristikám souboru jsou shrnuty v tab. 2. Celkový GDS skór byl 4,19 ± 3,51 (škála 0–15). Nenalezli jsme statisticky významný vztah GDS skóru k věku, pohlaví ani k dosaženému vzdělání respondentů. Významné rozdíly jsme zjistili ve vztahu k původnímu povolání zkoumaných osob, přičemž nejnižších skórů (bez deprese) dosáhli ti, kteří v minulosti zastávali odbornou profesi (zdravotníci, pedagogičtí pracovníci, technici). Co se týče rezidence respondentů, významně vyšší skóry GDS mají senioři, kteří potřebují zvýšenou míru ošetřovatelské péče, a tedy žijí v LDN, NNP a na ošetřovatelských lůžkách DS, a to v porovnání s těmi, kteří žijí doma sami ($p = 0,022$), doma v partnerství, širší rodině či v DsPS ($p < 0,001$) a na standardních odděleních DS ($p < 0,001$). Výsledky uvádíme po Bonferroniho korekci.

Porozumění článků 2

V zásadě se jedná o jednotlivé významové proměnné seřazené podle klesající četnosti výskytu. Horní kvartil tedy tvoří významové kategorie, které byly respondentem používány nejčastěji, dolní kvartil tvoří zpravidla významové proměnné, které respondent při svých odpovědích nepoužil ani jednou nebo jenom velmi zřídka. Významové proměnné z horního kvartilu se vyznačují tzv. pozitivním výskytem, významové proměnné z dolního kvartilu tzv. negativním výskytem v rámci významového profilu[5].

Statistická analýza

Získaná data v podobě četností (absolutních a relativních) jednotlivých sémantických kategorií byla zpracována jednak pomocí neparametrického Mannova-Whitneyova U-testu a v další etapě metodou kontingenčních tabulek (pozitivní a negativní výskyt proměnných ve významovém profilu, popř. jejich absence, jsme převedli na hodnoty +1, -1, popř. 0).

4. Výsledky

Zpracování dat proběhlo ve třech fázích. Nejprve byly porovnány obě skupiny z hlediska četnosti použití jednotlivých významových proměnných, poté byly porovnány profily obou skupin a nakonec jsme se pokusili interpretovat profil fobických jedinců.

Porovnání skupin

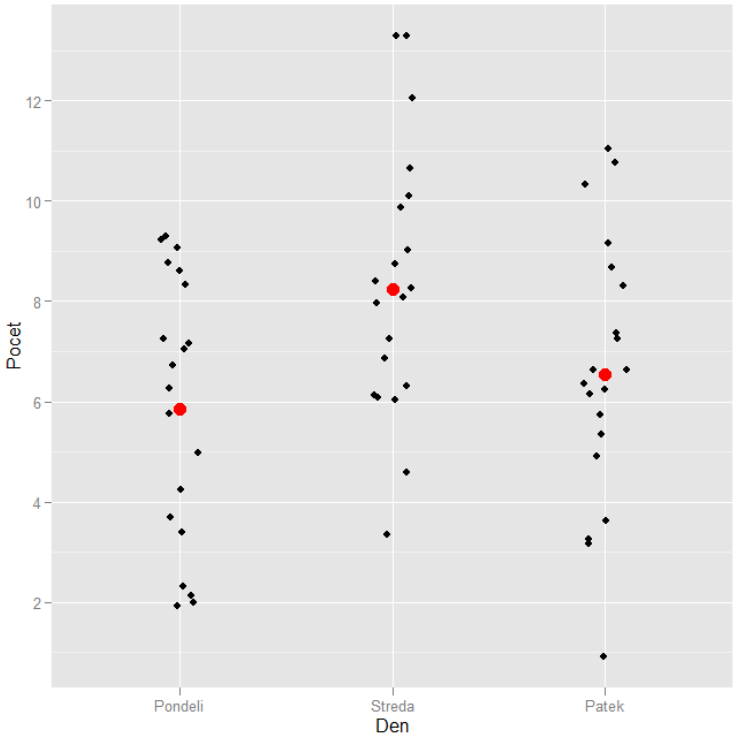
Na základě porovnání absolutních četností jednotlivých významových proměnných (Mannovým-Whitneyovým U-testem) bylo zjištěno, že jedinci z obou skupin se především výrazně lišili v celkovém průměrném počtu významových jednotek (skup. F v průměru 122,2; skup. K 177,1; průměrné pořadí u skup. F 16,5; u skup. K 24,5; U = 120,5; p = 0,03). To znamená, že porovnání obou skupin je třeba provádět na základě relativních četností, aby výsledky nebyly ovlivněny tímto celkovým rozdílem v úrovni produkce.

Tab. 2 obsahuje výsledky tohoto porovnání. Opět jsme použili neparametrický Mannův-Whitneyův U-test, protože obě skupiny jsou tvořeny pouze 20 respondenty a rozložení mnoha proměnných se odchyluje od normálního.

Porozumění zákonitostem kolem nás

- Otevřeli jsme si manželskou poradnu, lidé mohou chodit v pondělí, ve středu a v pátek
- Po dvaceti týdnech fungování musíme kvůli časovým důvodům jednu zrušit, kterou?
- Lidí celkem/průměr:
  - Po: **117/5.85**
  - St: 165/8.29
  - Pa: 129/6.45
- Zrušíme tedy pondělí
- Není to náhoda, že zrovna přišlo tolik lidí?

	pondeli	utery	patek
1	7	8	3
2	2	6	4
3	6	10	9
4	2	8	7
5	3	11	7
6	5	9	6
7	7	13	6
8	7	13	1
9	2	7	10
10	2	12	7
11	9	5	11
12	9	8	6
13	4	3	8
14	9	6	3
15	9	10	5
16	4	6	7
17	6	6	5
18	7	9	9
19	9	8	6
20	8	7	11



Materiály

- Vše, co je v sylabu (třeba Hendl je dobrý)
- Online kurz statistiky https://class.coursera.org/stats1-2012-001/
- Jiný online kurz statistiky http://www.udacity.com/overview/Course/st101/CourseRev/1
- Kurz z JČU http://www2.ef.jcu.cz/~rost/courses/stata/
- Kurz biostatistiky (=statistika v biologii) http://botanika.prf.jcu.cz/suspa/vyuka/statistika.php

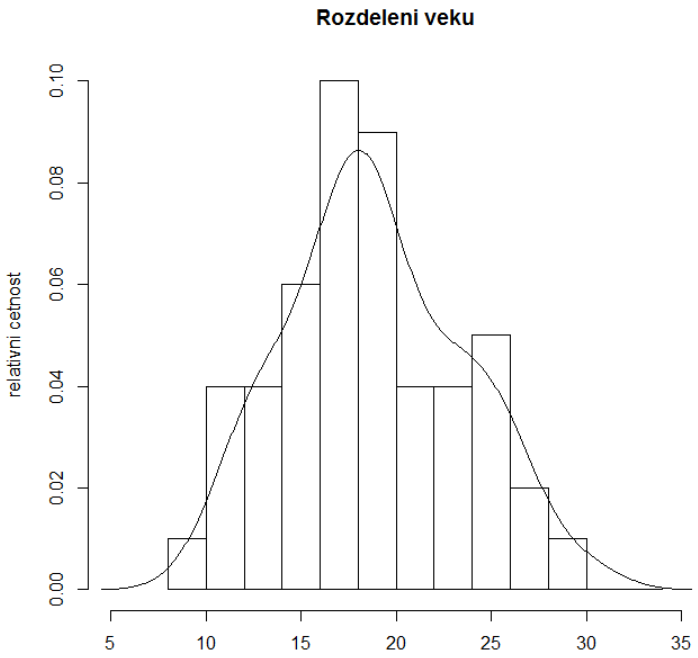
Rozdělní statistiky

- *Deskripční* (popisná) statistika – popisuje vlastnosti naměřených dat, z čistých dat není nic vidět
- *Inferenční* (odvozovací) statistika – odhaduje vlastnosti všech dat na základě naměřených dat
- Terminologie:
  - *Populace* (population)/základní soubor
  - *Vzorek* (sample)/výběrový soubor
- Tedy deskripční statistika dělá závěry jen o vzorku, zatímco inferenční dělá závěry o celé populaci

Deskripční statistika

- Zvyšuje přehlednost dat:
- Věky lidí:
  14 21 12 22 18 17 16 19 27 24 16 18 17 10 15 20 25 25 23 19 18 12 19 30 20 13 22 16 23 20 18 25 18 16 17 13 12 15 26 19 23 19 18 12 22 27 14 19 17 25
- Oproti:
  průměrný věk: 18.92
  nejmladší člověk: 10
  nejstarší člověk: 30

- Ztrácíme tím některé informace, ale za „dobrou cenu“



Základní statistické charakteristiky

- Libovolná data můžeme popsat pomocí četnosti a relativní četnosti
- Data, která můžeme porovnávat (později) můžeme popsat ještě pomocí kumulované četnosti a kumulované relativní četnosti
- Data, která se navíc chovají jako opravdová čísla (např. váha) můžeme popsat ještě lépe pomocí míry polohy a variability

Četnost

- Jednoduchá charakteristika, říká nám, kolik máme daných pozorování
- Př.: V obchodě jsme koupili: cibule, petržel, cibule, mrkev, brambory
  četnost jednotlivých položek:
  - Cibule - **2**
  - Petržel - **1**
  - Mrkev - **1**
  - Brambory - **1**

Relativní četnost

- Někdy je lepší vyjádřit data poměrově vůči celku
- Vyjadřujeme v procentech
- Spočítáme jako četnost/celkový počet prvků
- Relativní četnost jednotlivých položek (celkem 5):
  - Cibule – 2 -> rel.čet=2/5=40%
  - Petržel – 1 -> rel.čet=1/5=20%
  - Mrkev - 1 -> rel.čet=1/5=20%
  - Brambory - 1 -> rel.čet=1/5=20%
- Celkem musí být relativní četnost 100%

Kumulativní četnost

- Pokud můžeme data porovnávat podle velikosti, máme k dispozici i kumulovanou četnost
- Kumulovaná četnost pro prvek ***x***, značí počet prvků **menších nebo rovno** než x (obyčejná četnost vyjadřuje jen počet prvků rovno x)
- Obdobně máme i kumulovanou relativní četnost (akorát pracujeme z relativní četností)
- Dá se spočítat z četností (resp. relativních četností)

Výpočet kumulativní četnosti

- Ve gymnáziu jsou počty studentů v jednotlivých ročnících takto:
  - 1. ročník: 65 studentů
  - 2. ročník: 45 studentů
  - 3. ročník: 48 studentů
  - 4. ročník: 80 studentů
- Celkem tedy 65+45+48+80=238 studentů

Ročník	četnost	Rel.četnost	Kum. Čet.	Kum.rel.čet.
1	65	0.273	65	0.273
2	45	0.189	65+45=110	0.273+0.189=0.462
3	48	0.202	110+48=158	0.462+0.201=0.663
4	80	0.336	158+80=238	0.663+0.336=**1**

- Tedy na otázku, kolik studentů chodí do 1. nebo 2. ročníku odpovíme 46.2%

Příklad

- Děti ve škole psaly test. Jako statistici jste dostali známky jednotlivých dětí, určete četnost, relativní četnost, kumulovanou četnost a kumulovanou relativní četnost jednotlivých známek

Jméno	Známka
Anna	2
Bára	2
Cyril	2
Dominik	4
Eva	3
Filip	2
Gustav	2
Hubert	3
Ilona	2
Jana	3
Klára	2
Lukáš	5
Martin	1
Norbert	3
Otto	3
Petra	3
Richard	3

Míra středu a polohy

- Máme-li za data obyčejná čísla, můžeme použít charakteristiky středu a polohy
- Charakteristiky středu – jak jedním čísel popsat celý vzorek
- Charakteristiky rozptýlenosti – jak moc špatně jsme určili střed
- Tohle už dávno známe! Byť si to možná neuvědomujeme☺

Charakteristiky středu

- Pořádali jsme večírek pro našeho mladšího bratra a přišli nám na něj tito lidé (pro kompaktnost uvedeme jen stáří):
  5, 7, 6, 7, 8, 7
- Jak byste popsali kamarádovi, jak staří tam byli lidé?
- „Byly tam děti kolem 7 let“
- A tomu se matematicky říká ***průměr***

Charakteristiky středu 2

- Pojďme to zkomplikovat..
- Na párty našeho bratra přišel i děda jednoho z bratrových kamarádů, stáří lidí na večírku: 5, 7, 6, 7, 8, 7, 64
- Problém: průměrný věk vychází na 14.86 (a přitom tam žádný teenager není…)
- Řešení: uvedeme prostřední hodnotu (a tomu se matematicky říká *medián*) - 7

Charakteristiky středu 3

- Na párty se dostavil prarodič každého dítěte, stáří lidí na večírku:
  5, 7, 6, 7, 8, 7, 64, 58, 70, 66, 59, 60
- Průměr nám nepomůže (34.5 není dobrý popis)
- Medián taky ne (je nestabilní – přidá se jedno dítě nebo jeden prarodič a medián se změní)
  5, 7, 6, 7, **8**, 7, 64, 58, 70, 66, 59, 60, 6
  5, 7, 6, 7, 8, 7, 64, **58**, 70, 66, 59, 60, 80
- Řešení – uvedeme nejčastější hodnotu (a tomu se matematicky říká *modus*) - 7

A nyní matematicky..

- Vzorek zapíšeme pomocí vektoru, tj. X=(5, 7, 6, 7, 8, 7)
- Jednotlivé prvky označujeme pomocí x_i, kde i značí pořadí prvku ve výběru
- X=($x_1, x_2, x_3, x_4, x_5, x_6$)
- Otázka: kolik je $2 * (x_1 + x_5)$?
- X=(**5**, 7, 6, 7, **8**, 7) -> 2*(5+8)=26

  $(x_1, x_2, x_3, x_4, x_5, x_6)$
- Obecně X=($x_1, x_2, \ldots, x_n$)
  n říkáme *rozsah výběru*

Průměr

- Sečteme a vydělíme počtem prvků

$$\overline{x} = m = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Angl. *Mean*

Medián

- Seřadíme a vybereme prostřední člen, značíme $\tilde{x}$, angl. *Median*
- Musí platit, že 50% dat je větších nebo rovno než medián a 50% dat je menších nebo rovno
  formálně: rel.četnost($\tilde{x}$≤medián) ≤0.5 & rel.četnost($\tilde{x}$≥medián) ≥ 0.5
- Neseřazené: (5, 7, 6, 7, 8, 7, 64)
- Po seřazení: (5, 6, 7, **7**, 7, 8, 64)
- Co když bude sudý počet vzorků?

Medián 2

- Při sudém počtu vzorků spočítáme průměr dvou prostředních čísel
- Neseřazené: (5, 7, 6, 7, 8, 7, 64, 66)
- Po seřazení: (5, 6, 7, **7**, **7**, 8, 64, 66)
- Spočítáme průměr prostředních hodnot: $\frac{7+7}{2} = 7$
- Otázka: pokud $(x_1, x_2, \ldots, x_n)$ značí setříděnou posloupnost, jak zapsat matematicky průměr prostředních dvou?

Modus

- Spočítáme počty výskytů jednotlivých prvků a je to ten nejčastější, značíme $\hat{x}$, angl. *Modus*
- Může jich být i více
- (5, **7**, 6, **7**, 8, **7**, 64, 58, 70, 66, 59, 60)

hodnota	5	6	**7**	8	58	59	60	64	66	70
četnost	1	1	3	1	1	1	1	1	1	1

Kvantily

- Jde o hodnoty, které nám rozdělují setříděná data podobně jako medián
- < ----------/---------->

  p.100% x_p (1-p).100 %
- Tedy $x_{0.33}$ rozděluje data tak, že 33% dat je menších nebo rovno než $x_{0.33}$ a 66% je větších nebo rovno než $x_{0.33}$

Kvantily 2

- Nejčastěji se používají kvartily – $x_{0.25}$, $x_{0.50}$, $x_{0.75}$
- Otázka: jak jinak značíme $x_{0.50}$?
- Výpočet: spočítáme medián dvakrát
- Neseřazené: 5, 7, 6, 7, 8, 7, 64, 66
- Seřazené: 5, 6, 7, **7**, **7**, 8, 64, 66
- Medián spodních 50%: 5,6,7,**7** -> 6.5
- Medián horních 50%: **7**, 8, 64, 66 -> 36
- Dolní kvartil se také značí K_d, horní K_h

Příklad

- Děti ve škole psaly test. Jako statistici jste dostali počty bodů jednotlivých dětí, určete charakteristiky středu (střední hodnotu, medián, modus, horní a dolní kvartil) pro následující data

Jméno	Počet bodů
Anna	11
Bára	13
Cyril	12
Dominik	7
Eva	8
Filip	13
Gustav	12
Hubert	8
Ilona	11
Jana	9
Klára	12
Lukáš	4
Martin	16
Norbert	10
Otto	9
Petra	9
Richard	8

Charakteristiky variability

- Charakteristiky středu mohou vycházet stejně pro různé vzorky
- (7,7,7,7,7) a (5,6,7,8,9) mají stejné průměry (i mediány), ale evidentně u první vypovídá o vzorku lépe
- Data jsou kolem středu různě *rozptýlená*
- Používáme v životě běžně: „V kolik přijdeš?“ „Ve 4, **+- 20 minut**“

Rozpětí

- Nejjednodušší míra variabilty
- Stačí odečíst minimum a maximum
- Na data bez extrémů(*outliers*) to stačí, ale co třeba na (1,2,3,4,500)
- Rozpětí vychází 500-1=499, přitom většina dat je z rozsahu 1-4

Mezikvartilové rozpětí a odchylka

- Řeší problémy s extrémy
- Budeme pracovat s rozdíly kvartilů na rozdíl od klasického rozpětí
- $K_h - K_d$
- Pokud budeme chtít popsat, jak se data odchylují od mediánu, vydělíme dvěma
- $(K_h - K_d)/2$

Rozptyl

- Nejpoužívanější míra rozptýlenosti
- Sečteme druhé mocniny všech odchylek od průměru

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Jde o celkovou míru rozptýlenosti (ale moc nám neříká, jak jsou data průměrně rozptýlená)

Rozptyl příklad

- X=(4,5,7,10,14)
- $\bar{x}$=8, n=5

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
4	-4	16
5	-3	9
7	-1	1
10	2	4
14	6	36

- Součet všech $(x_i - \bar{x})^2$: 16+9+1+4+36=66
- Rozptyl tedy je 66/4= 16.5

Jiný vzorec pro rozptyl

- Výraz $(x_i - \bar{x})$ nám může dát ošklivá čísla (pokud $\bar{x}$ nebude přirozené) -> upravíme si vzorec, abychom odčítali jen pěkná čísla

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Jiný vzorec příklad

- X=(4,5,7,10,14)
- n=5

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

x_i	x_i^2
4	16
5	25
7	49
10	100
14	196

- Součet x_i je 4+5+7+10+14=40
- Součet x_i^2 je 16+25+49+100+196=386
- Rozptyl tedy je (386-1600/5)/4=16.5

Směrodatná odchylka

- Určuje průměrnou odchylku od středu
- Stačí odmocnit rozptyl, tedy

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Pro náš příklad tedy
  X=(4,5,7,10,14)
  s=4.06

Variační koeficient

- Slouží ke studiu, zda není s daty něco podivného
- Vydělíme směrodatnou odchylku průměrem
- Počítáme v procentech

$$V = \frac{s}{\overline{x}} \cdot 100$$

- Hodnoty větší než 15%-30% (záleží, jakého charakteru mám data) svědčí o nějakém problému (třeba kdybychom porovnávali stáří babičky a vesmíru)