Cvičení ze statistiky - 9

Filip Děchtěrenko

Minule bylo..

- Dobrali jsme normální rozdělení
- Tyhle termíny by měly být známé:
  - Inferenční statistika
  - Konfidenční intervaly
  - Z-test

Postup při testování hypotéz

1. Stanovíme si známé parametry
2. Stanovíme si nulovou a alternativní hypotézu
3. Stanovíme hladinu významnosti α a na jejím základě určíme kritickou hodnotu
4. Vypočítáme testovou statistiku (tohle je závislé na testu
5. Srovnáme testovou statistiku s kritickou hodnotou
6. Určíme p-hodnotu testu (=„Jestliže H_0 platí, jaká je pst, že získáme vypočítanou hodnotu nebo ještě neobvyklejší?“)

Druhy testů

- Celkem nás může potkat několik případů
- Budeme uvažovat, že data pocházejí z normálního rozdělení (jinak by se musely použít jiné metody)
- Zkoumáme μ základního souboru
  - Známe σ základního souboru
    - z-test
  - Neznáme σ základního souboru
    - t-test
- Porovnáváme μ_1 a μ_2 dvou základních souborů
  - Výběry na sobě závisí
    - Párový t-test
  - Výběry na sobě nezávisí
    - Dvouvýběrový t-test
- Zkoumáme-li σ základního souboru, používáme speciální testy pracující s χ^2rozdělením (nebudeme testovat) či F-testy (nebudeme testovat)

Z-test

- Výpočet testovací statistiky:

$$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

- Pro alternativní znak

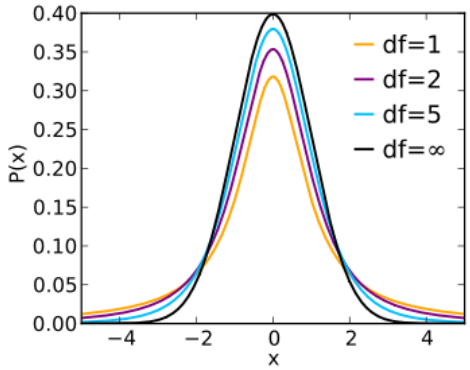
$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Příklad 2

- Na minci nám padlo 22 orlů ze 40 hodů. Preferuje tato mince orly? (řešte pro α=0.01)?
  - p=22/40=0.55, π_0=0.5, n=40
  - $H_0: \pi = \pi_0$
    $H_A: \pi > \pi_0$ (jednostranný test „>“)
  - α=0.01-> $u_{0.99}$=2.33
  - Z=$\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$=$\frac{0.55-0.5}{\sqrt{\frac{0.5(1-0.5)}{40}}}$=0.63
  - Abychom zamítli, musí platit Z>2.33, což neplatí (0.63 < 2.33), Hypotézu H_0 tedy nezamítáme
  - P(Z>0.63)=1-P(Z≤0.63)=1-0.74=0.26 (a to je více než α=0.01)
- Kolik potřebujeme hodů, abychom tato nevyváženost (55%) byla významná?
  - Abychom mohli nulovou hypotézu zamítnout, musí platit, že Z>2.33, tedy
  - $\frac{0.55-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}}$>2.33
    0.05>2.33$\frac{0.5}{\sqrt{n}}$
    $\sqrt{n}$>23.3
    n>542.89
  - Potřebujeme tedy alespoň 543 hodů s 55% výběrovým poměrem, aby to bylo významné

t-rozdělení

- Co když ale neznáme σ základního souboru?
  -> tak to bývá ve většině případů
- Nahradíme-li sigma výběrovou směrodatnou odchylkou, nedává Z-transformace normální rozdělení, ale dostaneme jiné rozdělení
- Nazývá studentovo rozdělení (t-rozdělení)
- $t_{n-1} = \frac{\bar{X}-\mu}{s_{\bar{X}}}$, kde $s_{\bar{X}} = \sqrt{\frac{s^2}{n}}$
- Tvar tohoto rozdělení závisí na parametru označovaným jako stupně volnosti (*df*)
- df=n-1



- Pro df→∞ se rozdělení blíží normálnímu rozdělení

Příklad

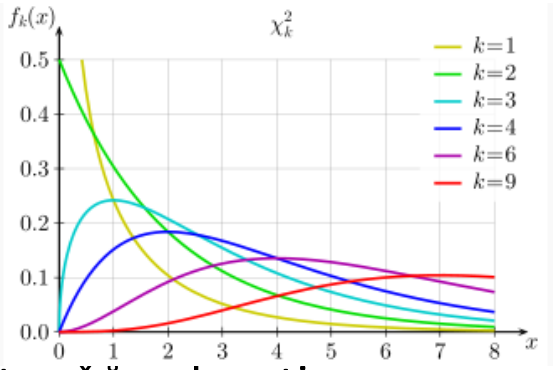
- Spotřeba téhož auta byla testována u 11 řidičů s výsledky 8.8,8.9, 9.0, 8.7, 9.3, 9.0, 8.7, 8.8, 9.4, 8.6, 8.9 (l/100 km). Je pravdivá výrobcem udávaná spotřeba 8,8 l/100 km? Předpokládejte normalitu dat
- n=11 -> df=10, $\bar{x} = 8.918$
  $s^2 = 0.061636$->$s_{\bar{X}}$ =0.0749
- α=0.05→$t_{0.995,10} = 2.228$
- t=$\frac{\bar{X}-\mu}{s_{\bar{X}}}$=$\frac{8.918-8.8}{0.0749}$=1.576
- t< $t_{0.995,10}$, tudíž nezamítáme na hladině významnosti α=0.05 (p-hodnota je těžší na spočítání)

Dva výběry

- Máme-li dva výběry, můžeme testovat, zda se jejich průměry statisticky významně liší (tedy zda nejde o náhodu, že to nevyšlo podobně)
- Pokud jsou na sobě výběry nezávislé (tj. každý pochází z vlastní populace), použijeme *dvouvýběrový t-test*
- Pokud jsou na sobě závislé (např. měříme před a po experimentu, používáme *párový t-test*)
- Můžete je počítat v Excelu TTEST(pole1;pole2;strany;typ)

χ^2-test dobré shody

- Hodili jsme 100 krát mincí, padla nám 60krát panna, 40 orel. Je mince falešná?
- Na tyto otázky se nám hodí χ^2-test dobré shody
- Obecně ho používáme v případech, že chceme porovnat rozdělení vzorku se známým rozdělením základního souboru
- Používáme k tomu χ^2 rozdělení
  - Opět má parametr stupeň volnosti
- Testovou charakteristiku spočítáme jako: $\chi^2_{k-1}=\sum \frac{(PC-OC)^2}{OC}$ kde
  - k je počet kategorií (u mince máme dvě)
  - PC je pozorovaná četnost
  - OC je očekávaná četnost



- Kritickou hodnotu opět hledám v tabulkách podle stupňů volnosti a hladiny významnosti
- Nepoužíváme oboustranný test, protože rozdělení je nesymetrické!

Příklad

- Řetezec cukráren, který nabízí 4 druhy zmrzliny otevřel provozovnu v nové lokalitě. Vyjádřete se pomocí statistického testu ke shodě či odlišnosti struktury prodeje v nové lokalitě oproti dosavadnímu řetězci.
- Prodej řetězce
  - Vanilková 62%
  - Čokoládová 18%
  - Jahodová 12%
  - Pistáciová 8%
- Prodej nové prodejny
  - Vanilková 120 ks
  - Čokoládová 40 ks
  - Jahodová 18 ks
  - Pistáciová 22 ks
- α=0.05, df=3 -> $\chi^2_{3,0.975} = 7.815$
- $\chi^2_3 = \sum \frac{(120-124)^2}{124} + \frac{(36-40)^2}{40} + \frac{(24-18)^2}{18} + \frac{(16-22)^2}{22} = 4.32$
- 4.32<7.815, tedy nulovou hypotézu nezamítáme na hladině významnosti α=0.05

Příklad ze života

- Zkoumali jsme množství násilných činů ve státech USA

```
> head(d)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

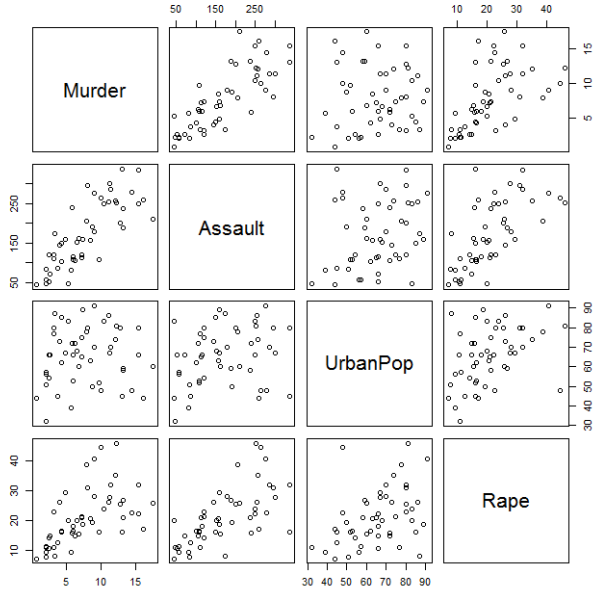
```
> summary(d)
     Murder          Assault         UrbanPop          Rape      
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00  
```

Deskriptivní statistika

- Popisné charakteristiky

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Murder	1	50	7.79	4.36	7.25	7.53	5.41	0.8	17.4	16.6	0.37	-0.83	0.62
Assault	2	50	170.76	83.34	159.00	168.47	110.45	45.0	337.0	292.0	0.22	-1.05	11.79
UrbanPop	3	50	65.54	14.47	66.00	65.88	17.79	32.0	91.0	59.0	-0.21	-0.74	2.05
Rape	4	50	21.23	9.37	20.10	20.36	8.60	7.3	46.0	38.7	0.75	0.35	1.32

- Bodové grafy



Korelace

- Pearsonovy koeficienty korelace

	Murder	Assault	UrbanPop	Rape
Murder	1.00000000	0.8018733	0.06957262	0.5635788
Assault	0.80187331	1.0000000	0.25887170	0.6652412
UrbanPop	0.06957262	0.2588717	1.00000000	0.4113412
Rape	0.56357883	0.6652412	0.41134124	1.0000000

- Hladiny významnosti

	Murder	Assault	UrbanPop	Rape
Murder		0.0000	0.6312	0.0000
Assault	0.0000		0.0695	0.0000
UrbanPop	0.6312	0.0695		0.0030
Rape	0.0000	0.0000	0.0030	

Regresní funkce

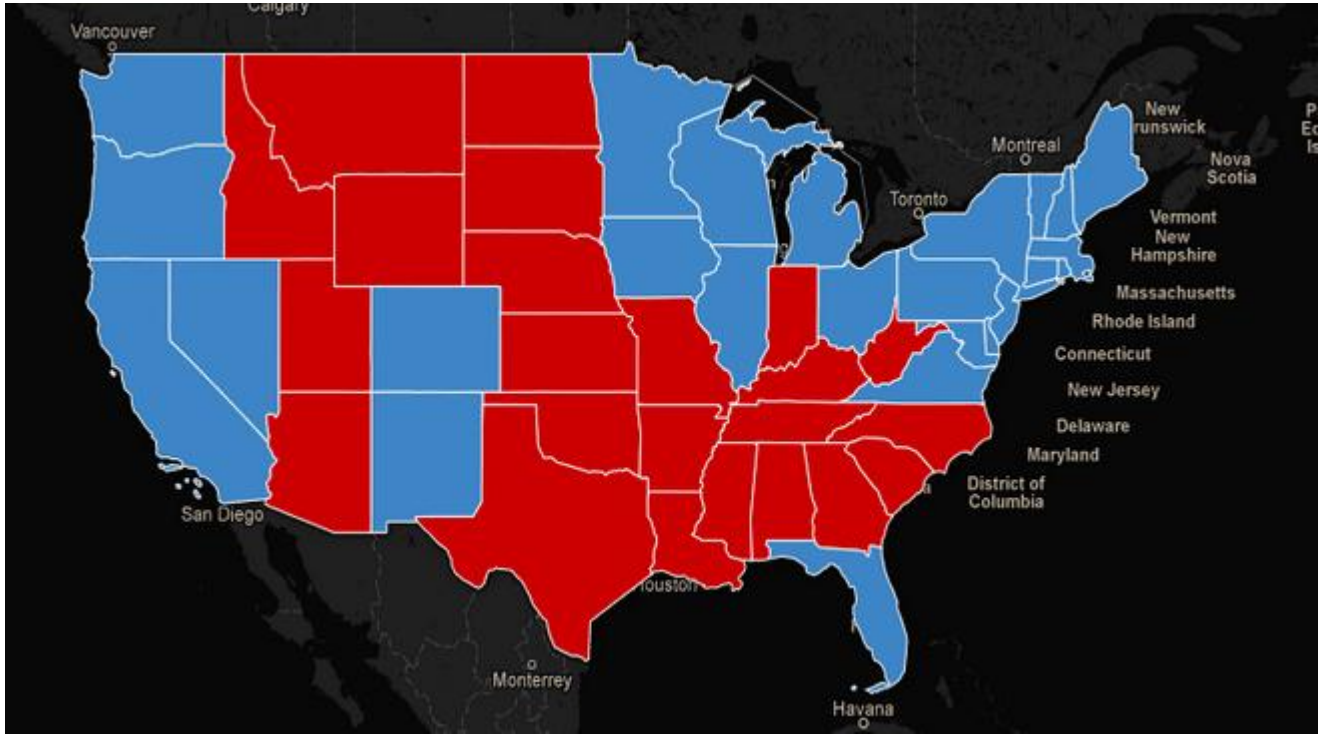
- Vypadá to, že mezi počet vražd a napadení je lineární závislost

```
> lm1.model<-lm(Assault~Murder,data=d)
> summary(lm1.model)

Call:
lm(formula = Assault ~ Murder, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-107.24  -36.35   -3.67   32.15  118.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.27      14.69   3.490  0.00105 **
Murder         15.34       1.65   9.298  2.6e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.31 on 48 degrees of freedom
Multiple R-squared: 0.643,      Adjusted R-squared: 0.6356
F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```

- A vyšlo to významně

Hypotéza – vraždy souvisí s volbou prezidenta

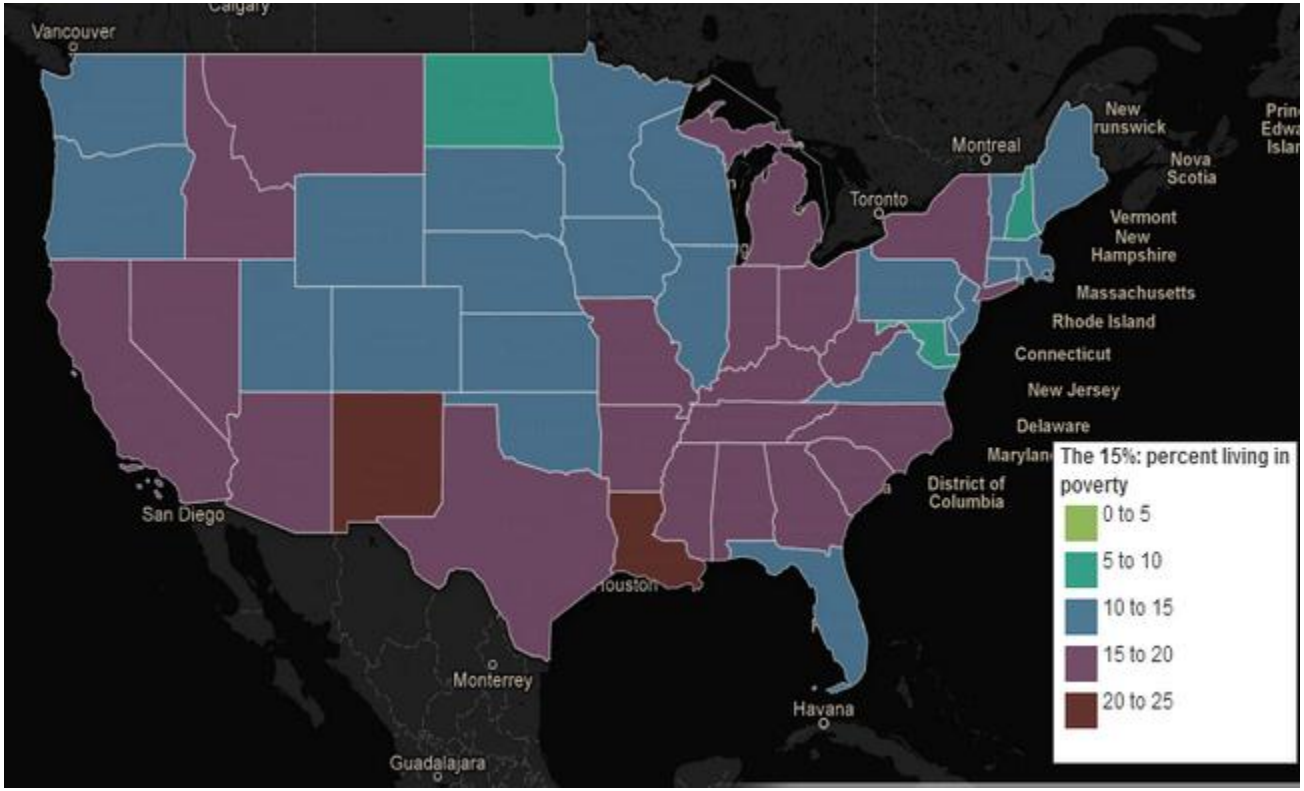
- Výsledky voleb



Zločiny a prezident

```
> t.test(d2[d2$Vote=="Obama",]$Murder,d2[d2$Vote=="Romney",]$Murder,var.equal = TRUE)

        Two Sample t-test

data:  d2[d2$Vote == "Obama", ]$Murder and d2[d2$Vote == "Romney", ]$Murder
t = -1.8003, df = 48, p-value = 0.0781
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.5946322  0.2536065
sample estimates:
mean of x mean of y
 6.746154  8.916667

> t.test(d2[d2$Vote=="Obama",]$Assault,d2[d2$Vote=="Romney",]$Assault,var.equal = TRUE)

        Two Sample t-test

data:  d2[d2$Vote == "Obama", ]$Assault and d2[d2$Vote == "Romney", ]$Assault
t = -0.3324, df = 48, p-value = 0.741
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -55.78147  39.95455
sample estimates:
mean of x mean of y
 166.9615  174.8750

> t.test(d2[d2$Vote=="Obama",]$Rape,d2[d2$Vote=="Romney",]$Rape,var.equal = TRUE)

        Two Sample t-test

data:  d2[d2$Vote == "Obama", ]$Rape and d2[d2$Vote == "Romney", ]$Rape
t = 0.5632, df = 48, p-value = 0.5759
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.864554  6.872246
sample estimates:
mean of x mean of y
 21.95385  20.45000

>
```

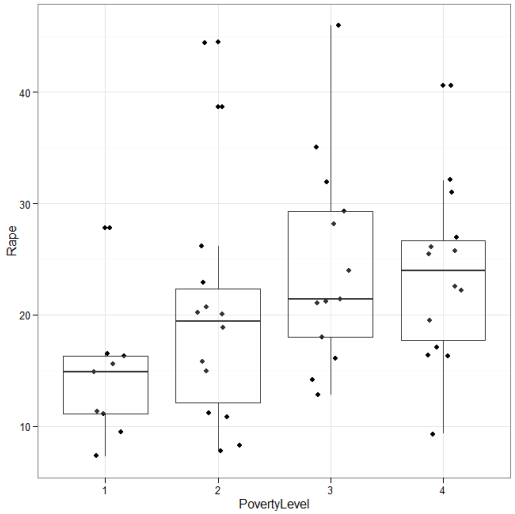
Hypotéza – vraždy souvisí s bohatstvím



Chudoba vs. zločiny

```
> aov1.model<-aov(Rape~PovertyLevel,data=d2)
> aov2.model<-aov(Murder~PovertyLevel,data=d2)
> aov3.model<-aov(Assault~PovertyLevel,data=d2)
>
> summary(aov1.model)
             Df Sum Sq Mean Sq F value Pr(>F)
PovertyLevel  3    656  218.72   2.762 0.0526 .
Residuals    46   3643   79.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov2.model)
             Df Sum Sq Mean Sq F value   Pr(>F)
PovertyLevel  3  393.6  131.21   11.26 1.16e-05 ***
Residuals    46  535.9   11.65
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov3.model)
             Df Sum Sq Mean Sq F value  Pr(>F)
PovertyLevel  3  74113   24704   4.269 0.00967 **
Residuals    46 266200    5787
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> TukeyHSD(aov2.model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Murder ~ PovertyLevel, data = d2)

$PovertyLevel
        diff        lwr       upr     p adj
2-1 1.174603 -2.7125500  5.061756 0.8515304
3-1 4.796581  0.8513594  8.741803 0.0115040
4-1 7.146032  3.2588785 11.033185 0.0000708
3-2 3.621978  0.1176948  7.126261 0.0403423
4-2 5.971429  2.5326517  9.410205 0.0001726
4-3 2.349451 -1.1548327  5.853734 0.2925160
```



A toť je vše

- Hodně štěstí u zkoušky