

R - inferenční statistika

Lubor Homolka

12. května 2014

Obsah školení

Teoretická část:

- Domácí cvičení
- Replikace a pravděpodobnost
- Schéma testování hypotéz: vědecká vs. statistická
- Chyby v interpretaci výsledků
- Meta-analýza

Praktická část:

- proporční test prop. test
- t-test a ANOVA
- meta-analýza

Domácí cvičení - dotazy

`https://github.com/luboRprojects`

R-skoleni2 → skript: `plyr_ukazka.R`

Frekvencionalistická statistika

$$\text{Pravděpodobnost} = \frac{\text{Příznivá pozorování}}{\text{Všechna pozorování}}$$

Chyba $\alpha = P(\text{zamítáme } H_0 | H_0 = \text{pravdivá})$

$$\alpha = \frac{\text{Počet článků, které zamítnou}}{\text{Všechny články}}$$

Chyba $\beta = P(\text{nezamítáme } H_0 | H_0 = \text{nepravdivá})$

$$\beta = \frac{\text{Počet článků, které nezamítnou}}{\text{Všechny články}}$$

Hypotézy

- Popperova falzifikace,
- Epistemologický boj: Neymar-Pearson a Fisher
- Vědecká hypotéza: Organizace se správně implementovaným BSC, dosahují vyšší finanční výkonnosti (měřeno ROE).
- Nulová hypotéza: Efekt implementace BSC na ROE neexistuje: $H_0 : \beta = 0$
- Alternativní hypotéza: Efekt implementace BSC na ROE je kladný: $H_1 : \beta > 0$

- nezveřejnění $p\text{-val} > 0.05$.
- statistická významnost vs. věcná významnost
- kontrola splnění předpokladů - alternativní přístupy k asymptotickému odhadu (Monte Carlo)
- nepřesná interpretace intervalů spolehlivosti

Meta-analýza: Analýza síly testu

- Jak ambiciózní je naše analýza?
- Jaké výsledky můžeme očekávat?
- Jak velký má být vzorek?

Analýza síly testu – Power analysis

Analýzovaný problém souvisí s:

- skutečným rozdílem mezi charakteristikami (effect size)
- neurčitost v chování proměnných (variance)
- váha našeho důkazu (observations)
- požadovaná míra chybovosti (α , $1 - \beta$).

Analýza síly testu – Power analysis

Domníváme se, že průměrná hodnota ROE firem, které správně implementují BSC je o 3 % vyšší než u ostatních firem. Očekáváte, že rozptyl ROE je 36%. Protože nevíte, zda opravdu BSC přináší pozitivní efekt, musíte specifikovat toleranci k oběma chybám. Je nepřijatelné, abyste potvrdili, že vliv existuje v případě, kdy neexistuje ($\alpha = 0.05$). Naopak, až tolik nevadí, když nenajdete důkaz pro existenci efektu, pokud existuje ($\beta = 0.2$), neboli síla testu je 0.8.

Jaký je minimální počet respondentů pro náš výzkum?

Analýza síly testu – Power analysis

Test rozdílu středních hodnot pro dvě úrovně je studentův t-test (`t.test`).

effect size pro t-test je stanoven jako:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

```
library(pwr) #install.packages("pwr")
effect.size <- (0.12-0.10)/0.06
pwr.t.test(d = effect.size , sig.level = 0.05,
           power = 0.8, type = c("two.sample"),
           alternative="greater" )
```

Existuje velmi mnoho možností úprav, jako rozdílné velikosti souborů, rozdílné rozptyly apod.

Analýza síly testu – Power analysis

Výsledky:

Two-sample t test power calculation

```
n = 111.9686
d = 0.3333333
sig.level = 0.05
power = 0.8
alternative = greater
```

NOTE: n is number in *each* group

Náš vzorek by musel být více než 220 firem, abychom dokázali detekovat rozdíl 2 %!

Analýza síly testu – Váš úkol

Neprovedli jste power analýzu a nyní Vám nic nevychází! Nasbírali jste totiž jen 80 firem v každé skupině (BSC, non-BSC). Jak velký by musel být skutečný efekt na ROE (rozdíl mezi μ_{BSC} a $\mu_{\text{non-BSC}}$), abyste tento rozdíl byli schopni nalézt?

`github.com/luboRprojects/r_skoleni2/power_analysis.R`

Analýza síly testu – Řešení

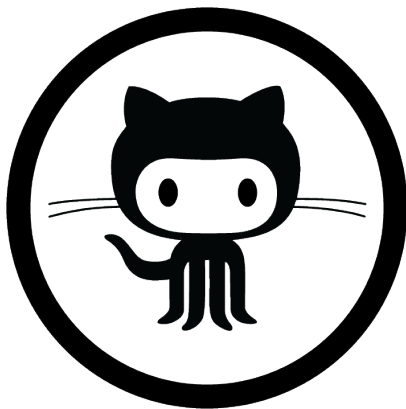
```
> pwr.t.test(n = 80, sig.level = 0.05, power = 0.8,  
  type = c("two.sample"), alternative="greater" )
```

Two-sample t test power calculation

```
      n = 80  
      d = 0.3948399  
sig.level = 0.05  
  power = 0.8  
alternative = greater
```

$$0.3948399 = \frac{\text{efekt}}{0.06}$$

Efekt by musel být minimálně 2.369 %, abychom ho dokázali detekovat.



`github.com/luboRprojects/r_skoleni2/elementary_methods.R`

Proporční test

Testuje hodnotu nebo shodu podílů π . Nulová hypotéza tvrdí, že neexistuje rozdíl $H_0 : \pi_1 = \pi_2$

Zeptali jsme se 100 Čechů a 100 Slováků na preferenci daného výrobku.

- Možné odpovědi: líbí/nelíbí
- 75% Čechů se líbí, Slovákům pouze 70 %
- Je možné tvrdit, že existují různé preference?

```
> prop.test(x=c(75, 70), n=c(100, 100))
```

Jak by se situace změnila, kdyby náš vzorek byl 10× větší? (a proporce se zachovaly)

Analýza kontingenční tabulky

Zjišťujeme, zda struktura kontingenční tabulky získané z našich dat odpovídá té, kterou bychom očekávali nastala, kdyby neexistoval vztah mezi proměnnými.

$$CH = \sum \sum \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

- V případě splnění podmínek má testové kritérium CH rozdělení χ^2
- Porušení podmínek (min. počet pozorování v buňce 5) neumožňuje srovnání CH s χ^2 .
- V případě porušení podmínek buď `fisher.test`, nebo parametr `simulate.p.value=TRUE` ve funkci `chisq.test`.

Kontingenční tabulka - příklad

Je prokazatelné, že se BSC firmy více hlásí k CSR?

```
> data.chi <- as.data.frame(matrix(c(30, 90, 60, 40),  
byrow=TRUE, ncol=2))
```

```
> data.chi
```

	CSR-ne	CSR-ano
BSC	30	90
non-BSC	60	40

```
> chisq.test(data.chi)
```

Pearson's Chi-squared test with Yates'
continuity correction

```
data: data.chi
```

```
X-squared = 26.2121, df = 1, p-value = 3.059e-07
```

Korelační analýza

Míra intenzity lineárního vztahu. Nulová hypotéza tvrdí, že vztah neexistuje, neboli $H_0 : \rho = 0$.

- funkce `cor` vytváří korelační matici, ale netestuje významnost
- funkce `cor.test` provádí párové testy a přináší další možnosti, jako je např. volba typu korelačního koeficientu (Pearson, Spearman).

```
cor(data.in[,c("Fresh", "Milk", "Grocery")])  
cor.test(data.in$Fresh, data.in$Milk)
```

Regresní analýza

Analyzuje kauzální vztah mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými.

Testujeme významnost jak celkového modelu, tak jednotlivých koeficientů.

- hodnoty testů získáme ze `summary(objekt)`
- celkový F-test - je model lepší než průměr? $H_0 : \beta_1 = \dots = \beta_k = 0$
- přispívá koeficient β k vysvětlení? $H_0 : \beta_l = 0$

Prezentace a její obsah je samozřejmě reproducible!



https://github.com/luboRprojects/r_skoleni2/