PV252

Tvorba uživatelských rozhraní a hodnocení použitelnosti

12. přednáška

TESTOVÁNÍ HYPOTÉZ II: ZÁKLADNÍ METODY

Martin Dostál
Honeywell International - Aerospace Advanced Technology Europe
Masarykova Univerzita v Brně, Fakulta informatiky

POZOR, NAŠE ZPRACOVÁNÍ TÉTO PROBLEMATIKY JE S OHLEDEM NA PROSTOR ZNAČNĚ POVRCHNÍ. STATISTIKU JE TŘEBA STUDOVAT VE VĚTŠÍM DETAILU.

TŘEBA STUDOVAT VE VĚTŠÍM DETAILU.

Hypotézy a metody

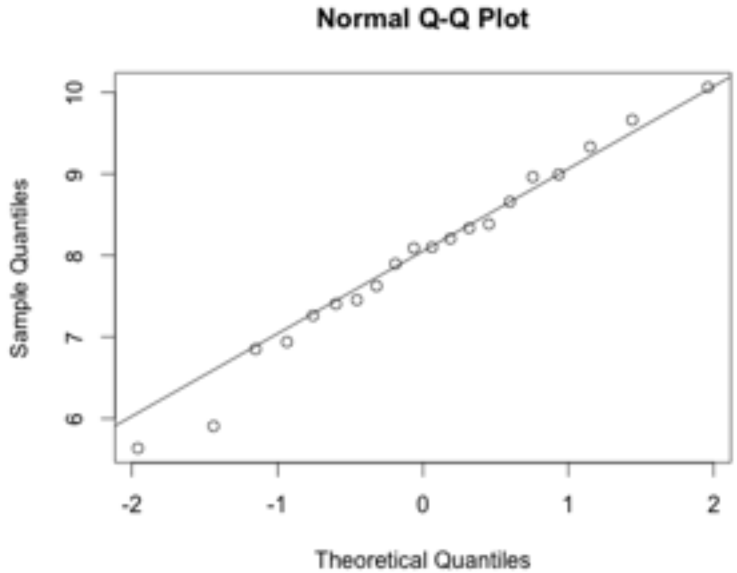
- podle rozdělení pravděpodobnosti
  - parametrické (opírají se předpoklad rozdělení parametru rozdělení, typicky normálního)
  - neparametrické - nemají požadavky na rozdělení, širší aplikovatelnost znamená ale vyšší riziko chyby druhého druhu - parametrické metody jsou slabší
- podle skupin
  - nepárové
  - párové

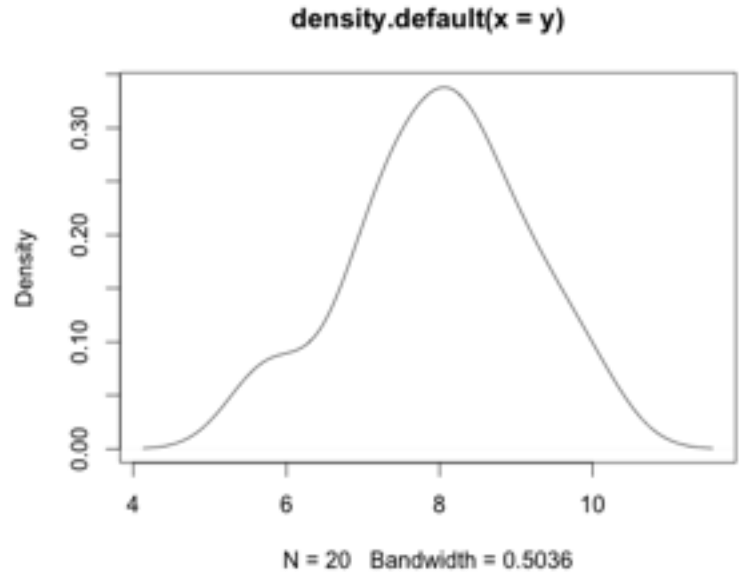
PRO POUŽITÍ PARAMETRICKÝCH METOD, JE TAKÉ MOŽNOST PROVÉST TRANSFORMACI DAT

Přehled základních metod

	Typ závislé proměnné		
	Intervalová/podílová data s normálním rozdělením	**Intervalová/podílová data s jiným než normálním rozdělením, ordinální data**	**Dichotomická**
Porovnání dvou nepárových skupin	Nepárový t-test	Mann-Whitneyho test	Fisherův test
Porovnání dvou párových skupin	Párový t-test	Wilcoxonův test	McNemarův test
Porovnání více než dvou nepárových skupin	Analýza rozptylu (ANOVA)	Kruskal-Wallis test	Chi-kvadrát test
Porovnání více než dvou párových skupin	Analýza rozptylu s opakovanými pozorováními (Repeated-measures ANOVA, RM-ANOVA)	Friedmanův test	Cochranův Q-test
Zjištění korelace mezi dvěma proměnnými	Pearsonův korelační koeficient	Spearmanův korelační koeficent	Cramerovo V

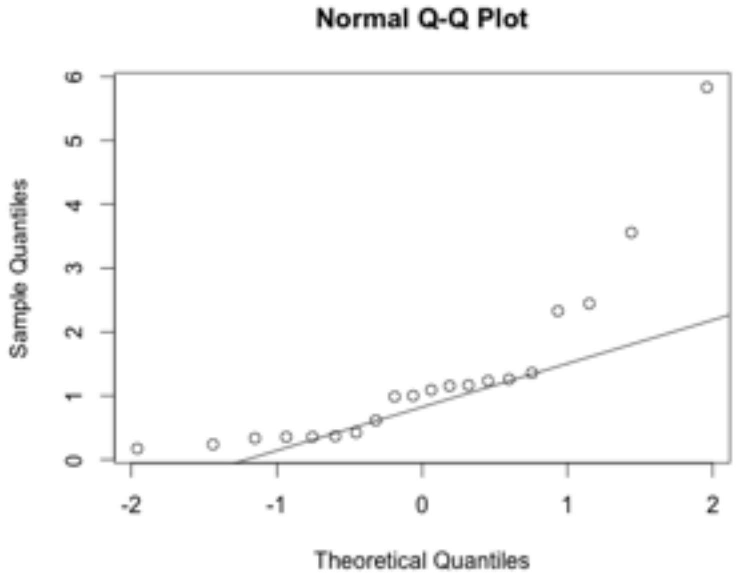
Test normality: QQ-plot

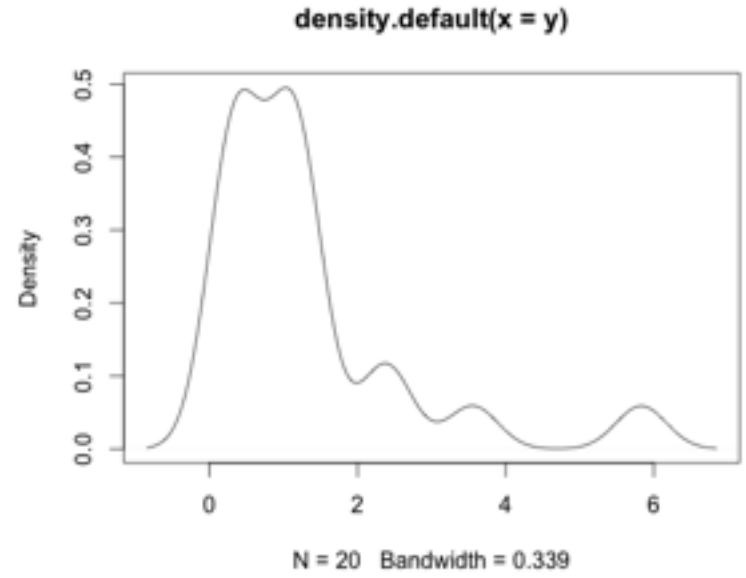





```
> y
 [1] 10.064251  8.101385  7.899375  6.939880  7.406208  9.663936  6.851816  8.995464
9.333238  5.906350
[11]  8.206650  7.262491  7.455530  7.626505  8.091334  8.964504  8.336392  8.386712
5.635199  8.658956
> qqnorm(y)
> qqline(y)
> plot(density(y))
```

Test normality: QQ-plot






```
> y
 [1] 1.2597372 1.1660276 0.3361962 0.3576890 1.3590426 0.3683438 5.8300409 0.2395507
1.1538387 1.2346716
[11] 3.5565921 0.6149820 2.3282096 0.3558715 0.1731463 2.4451087 0.4247259 0.9879564
1.0018932 1.0914047
> qqnorm(y)
> qqline(y)
> plot(density(y))
```

Test normality

- Shapiro-Wilkův test
- nulový hypotéza: data procházejí z normálního rozdělení
- pokud $p < \alpha$, zamítáme nulovou hypotézu

```
> x.normal<-rnorm(20, 5)
> x.normal
 [1] 5.207349 5.646124 4.588337 4.253999 4.155268 4.168005 5.667880 5.751262
5.709069 5.671945 6.373626
[12] 3.318484 6.218892 5.968625 5.843479 3.110242 7.374172 5.834502 4.932141
5.174467
> y.lognormal<-rlnorm(20, meanlog = 0, sdlog = 1)
> shapiro.test(x.normal)

    Shapiro-Wilk normality test

data:  x.normal
W = 0.9506, p-value = 0.3756

> shapiro.test(y.lognormal)

    Shapiro-Wilk normality test

data:  y.lognormal
W = 0.4834, p-value = 2.258e-07
```

Test normality

- Kolmogorovův-Smirnovův test
- porovnává dva rozdělení dvou výběrů
- nulová hypotéza: oba výběry pocházejí ze stejného rozdělení pravděpodobnosti

```
> ks.test(x.normal, "pnorm", mean=mean(x.normal), sd=sd(x.normal))

	One-sample Kolmogorov-Smirnov test

data:  x.normal
D = 0.197, p-value = 0.3703
alternative hypothesis: two-sided

> ks.test(y.lognormal, "pnorm", mean=mean(y.lognormal), sd=sd(y.lognormal))

	One-sample Kolmogorov-Smirnov test

data:  y.lognormal
D = 0.3671, p-value = 0.006339
alternative hypothesis: two-sided
```

t-test

- předpoklad: výběr(y) pocházejí z normálního rozdělení
- varianty: jedno- a dvouvýběrový, nepárový, párový
- u nepárového testu se předpokládá homogenita rozptylů výběrů (použít Bartletův test)

jednovýběrový t-test

- nulová hypotéza: průměr rovná se *mu*
- alternativní hypotéza: nerovnají/menší/větší

```
> x.normal
 [1] 5.207349 5.646124 4.588337 4.253999 4.155268 4.168005 5.667880 5.751262 5.709069
5.671945 6.373626
[12] 3.318484 6.218892 5.968625 5.843479 3.110242 7.374172 5.834502 4.932141 5.174467
> t.test(x, alternative="greater", mu=5)

    One Sample t-test

data:  x
t = -1.877, df = 9, p-value = 0.9534
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:
 3.872459      Inf
sample estimates:
mean of x
 4.429556
```

jednovýběrový t-test

```
> t.test(x.normal, alternative="greater", mu=3)

	One Sample t-test

data:  x.normal
t = 4.7039, df = 9, p-value = 0.000557
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 3.872459      Inf
sample estimates:
mean of x
 4.429556
```

Older age participants reported more time spent with the task (M = 4.42, *SD* = 1.05) than users did in general, $t(9) = 4.70$, $p < .001$

dvouvýběrový t-test

- nulová hypotéza: průměry výběrů se rovnají
- alternativní hypotéza: nerovnají/menší/větší

```
> small.button
 [1] 5.2263125 3.2908618 3.4180683 3.7067978 0.5340702 3.0823482 2.2231250
3.4266038 4.9352018 2.8071005

> large.button
 [1]  1.6873803  1.6217134  1.4451332  2.4695364  0.7222387  3.6875096  2.1387742
-0.4569893  1.6889108
[10]  1.2102160

> t.test(large.button,small.button,alternative="less")

    Welch Two Sample t-test

data:  large.button and small.button
t = -3.0427, df = 17.352, p-value = 0.003612
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.7050047
sample estimates:
mean of x mean of y
 1.621442  3.265049
```

Bartletův test

- nulová hypotéza: rozptyly se rovnají
- alternativní hypotéza: rozptyly se nerovnají

```
> bartlett.test(list(small.button,large.button))

	Bartlett test of homogeneity of variances

data:  list(small.button, large.button)
Bartlett's K-squared = 0.3245, df = 1, p-value = 0.5689
```

párový t-test

```
> red.button = c(19, 20, 21, 22, 23, 22, 27, 25, 27, 28)
> green.button = c(21, 22, 24, 24, 25, 25, 26, 26, 28, 32)
> t.test(green.button,red.button,alternative="greater", paired=TRUE)

    Paired t-test

data:  green.button and red.button
t = 4.3846, df = 9, p-value = 0.0008796
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.105651      Inf
sample estimates:
mean of the differences
                    1.9
```

Mann-Whinteyův/Wilcoxonův znaménkový test

- neparametrický test, pro data s jiným než normálním rozdělením
- určeno pro nepárová data
- dvě různé metody a tedy dvě různé testové statistiky, ale stejná p-hodnota
- nulová hypotéza: oba výběry mají stejné rozdělení pravděpodobnosti (tj. četnosti hodnot v obou souborech jsou stejné)
- alternativní hypotéza: výběry mají různá rozdělení pravděpodobnosti

```
> Question.Baseline <- c(2,4,3,1,2,3,3,2,3,1)
> Question.Test <- c(3,5,4,2,4,3,5,5,3,2)
>
> wilcox.test(Question.Baseline, Question.Test)

	Wilcoxon rank sum test with continuity correction

data:  Question.Baseline and Question.Test
W = 23, p-value = 0.03841
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxonův test - párový

- nulová hypotéza: median rozdílů mezi párovými hodnotami se rovná nule
- alternativní hypotéza: median rozdílů mezi párovými hodnotami je různý od nuly

```
> wilcox.test(Question.Baseline, Question.Test,paired=T)

	Wilcoxon signed rank test with continuity correction

data:  Question.Baseline and Question.Test
V = 0, p-value = 0.01187
alternative hypothesis: true location shift is not equal to 0
```

No jo, ale …

- co když máme více než dva výběry?
- co když mám více nezávislých proměnných?
- co když potřebuju studovat interakce mezi proměnnými?

Nápad

- proč nepoužít sadu t-testů pro dva výběry?
- pravděpodobnost chyby I. druhu:
    - $1 - (1 - \alpha)^n$, pro *n* porovnání potřebných pro srovnání výběrů po dvojicích
    - $n = k*(k-1)/2$, k = počet výběrů
- pravděpodobnost chyby I. druhu rychle roste

ANOVA

- ANOVA = ANalysis Of VAriance, čili analýza rozptylu
- parametrická metoda (existují i neparametrické vairanty)
- existuje mnojo variant této metody podle struktury a podoby dat, viz další slajd
- není pokryto učebním textem „Základy statistické analýzy dat”

Druhy ANOVy

- podle faktorů
  - jednofaktorová
  - vícefaktororová
- podle efektů
  - s pevnými efekty
  - s náhodnými efekty
  - se smíšenými efekty
- podle pozorování
  - bez opakování
  - s opakováním
- podle designu skupin
  - vyvážená
  - nevyvážená

Nulová hypotéza

- nulová hypotéza: průměry výběrů se rovnaji
- alternativní hypotéza: průměry výběrů se nerovnaji
- pozor, alternatiní hypotéza říká že alespoň dva průměry se nerovnají (nikoliv všechny navzájem!)

Princip

- Analysis of variance
- Variance, čili rozptyl
- porovnáváme dva zdroje rozptylu
  - mezi skupinami (výběry)
  - uvnitř skupin (výběrů)
- počítá se tzv. F-statistika

PROMĚNNÉ VYSVĚTLÍME

$$F = \frac{\text{rozptyl "mezi skupinami"}}{\text{rozptyl "uvnitř skupin"}}$$

$MS_A = SS_A/df_a$

$MS_E = SS_E/df_e$

$$F = MS_A/MS_E$$

Předpoklady

- výběry mají normální rozdělení (volnější pojetí: u větších vzorků je možné pracovat s určitou benevolencí)
- rozptyly výběrů se rovnají - vhodný je Bartletův test (volnější pojetí: je-li rozdíl mezi nejmenším a největším rozptylem dvojnásobný, je ještě možné toto kriterium považovat za splněné)

ANOVA a postup

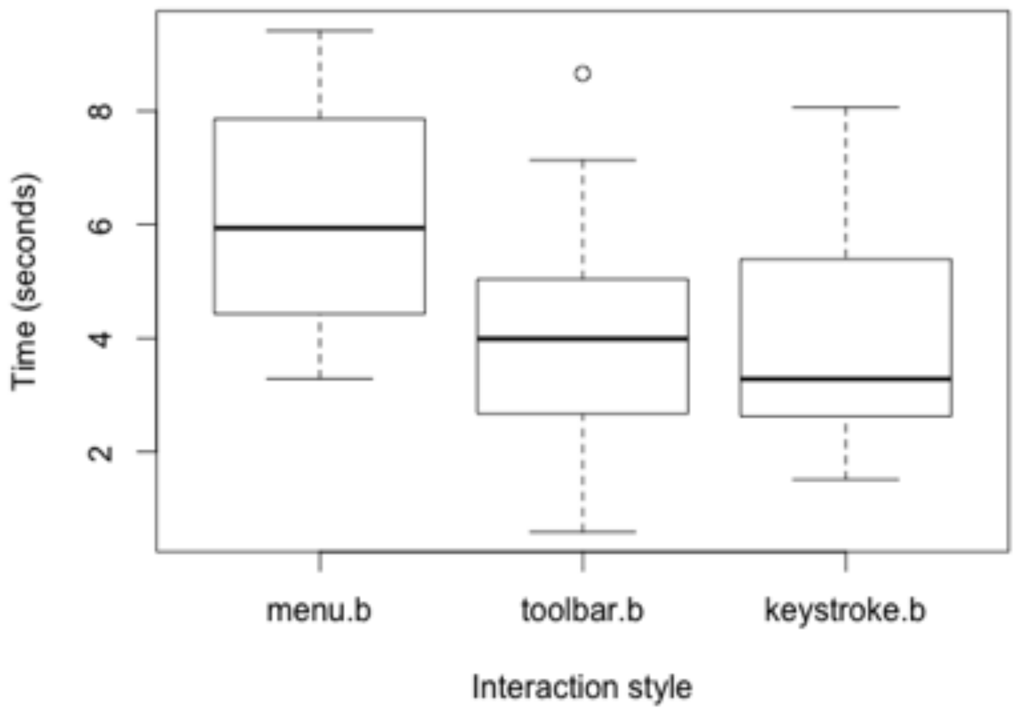
- postup je stejný jako již bylo dříve uvedeno, avšak ...
- výstupem ANOVy je informace zda zamítáme/nezamítáme nulovou hypotézu, my však potřebujeme vědět víc
- existují významné rozdíly mezi průměry výběrů, ale mezi kterými?
- pro tento účel slouží tzv. post-hoc testování

Jednofaktorová ANOVA

- jedna závislá proměnná
- jedna nezávislá proměnná
- nepárová data (design mezi skupinami)

Příklad

- data o interakci
- závislá proměnná: čas
- nezávislá proměnná: interakční styl
  - hladiny: menu, toolbar a klávesové zkratky



Příklad

```
> summary(c(menu.b,toolbar.b,keystroke.b))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5991  3.1750  4.3340  4.6920  6.1340  9.4000

> summary(menu.b)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.296   4.610   5.940   6.160   7.783   9.400

> summary(toolbar.b)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5991  2.8170  3.9920  3.9890  5.0030  8.6530

> summary(keystroke.b)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.519   2.648   3.285   3.925   5.032   8.062

> bartlett.test(list(menu.b,toolbar.b,keystroke.b))

	Bartlett test of homogeneity of variances

data:  list(menu.b, toolbar.b, keystroke.b)
Bartlett's K-squared = 0.1961, df = 2, p-value = 0.9066
```

Příklad: ANOVA

```
> aov.b <- aov(data.full.b$time ~ data.full.b$interaction,
data.full)
> summary(aov.b)
                        Df Sum Sq Mean Sq F value   Pr(>F)
data.full.b$interaction  2  64.75   32.37   8.774 0.000476 ***
Residuals               57 210.31    3.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

VARIABILITA

P-HODNOTA

HODNOTA F-STATISTIKY

Příklad: post-hoc test

- post-hoc testů existuje celá řada
- zde Tukeyho HSD test, porováváme výběry mezi sebou, nezáleží na pořadí

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

Fit: aov(formula = data.full.b$time ~ data.full.b$interaction,
data = data.full)

$`data.full.b$interaction`
                        diff        lwr        upr     p adj
menu-keystroke     2.23506378  0.7733432  3.6967844 0.0014873
toolbar-keystroke  0.06418882 -1.3975318  1.5259094 0.9938631
toolbar-menu      -2.17087496 -3.6325955 -0.7091544 0.0020627
```

Pár poznámek k ANOVA

- nejdůležitější varianty
  - RM-ANOVA
  - vícefaktorová ANOVA
    - můžeme hledat interakce mezi nezávislými proměnnými
- v „klikacích” aplikacích vybereme design experimentu
- v „R” je potřeba sestavit formuli pro testování ANOVA
- pro „R” existují alternativní knihovny pro výpočet ANOVA, zpravidla mají širší funkcionalitu a poskytují bohatší výstup potřebný pro metody následující po ANOVA, či dodatečné výpočty