Testování hypotéz o střední hodnotě — pokračování
2.11.2012

Úvodní nastavení.
– Otevřete si R Studio. Z internetu si stáhněte data smokers.txt.

Opakování z minula: Kterou testovací metodu použijeme na problémy, které byly na rozmyšlení? Jaké předpoklady musíme ověřit a jak? Jsou předpoklady nezávislosti pozorování v 2. a 3. reálné?

Analýza rozptylu (ANOVA) — jednoduché třídění (one-way)

1. Data smokers.txt obsahují údaje o procentuálním zastoupení kuřáků v populaci podle věku a měsíčním příjmu rodiny (z let 1978-1980, USA). Jednotlivé proměnné udávají:

Smokers	procento pravidelných kuřáků
Income	roční rodinný příjem: pod 5000, 5001–9999, 10 000-14 999, 15 000-24 999, 25 000 a více
Age	věk: 17-30, 31-64, 65 a více

   Bude nás zajímat, zda se liší procentuální zastoupení kuřáků v jednotlivých věkových skupinách.

2. Načtěte si data do R. Zpřístupněte si jednotlivé proměnné

```
attach(smokers)
```

3. Podívejte se na graf (boxplot) znázorňující zkoumanou závislost. Které skupiny se zřejmě budou lišit?

4. Spočtěte průměry a směrodatné odchylky v jednotlivých skupinách dle věku.

5. Statistické srovnání provedeme pomocí analýzy rozptylu (ANOVA):

```
model=aov(Smokers~factor(Age))
anova(model)
```

   Jaký je výsledek testu?

6. Abychom zjistili, které skupiny se od sebe významně liší, provedeme Tukeyho test mnohonásobného porovnávání:

```
TukeyHSD(model)

#graficke znazorneni
plot(TukeyHSD(model))
```

   K jakému závěru jsme dospěli?

7. Předpoklady analýzy rozptylu jsou: nezávislost, normalita a shoda rozptylů. Ověříme nejprve, jak je to s normalitou. Budeme se dívat na rozdíly „pozorování - průměr v příslušné skupině“. Ty by všechny měly mít stejné normální rozdělení. Tyto rozdíly získáme jako tzv. rezidua z našeho modelu.

```
r=resid(model)
shapiro.test(r)

hist(r)
library(car)
qqPlot(r,dist="norm")
```

Předpoklad shody rozptylů se ověřuje pomocí Levenova nebo Bartlettova testu

```
bartlett.test(Smokers~factor(Age))

# v knihovne library(car)
leveneTest(Smokers~factor(Age))
```

Co lze konstatovat o splnění předpokladů?

Samostatná práce

8. Byl vyšetřován vliv čtyř různých druhů penicilínu na růst bacilu B substilis. Vyšetřete, zda je účinek všech čtyř penicilínů shodný. Naměřená data jsou uvedena v souboru penicilin.csv.

9. Datový soubor srazky.csv obsahuje průměrné měsíční teploty a srážky v letech 2000 až 2011. Zjistěte, zda a jak závisí průměrné denní srážky na ročním období. Kdy prší nejvíce? Kdy naopak nejméně?

10. Pro stejná data proveďte porovnání pro jednotlivé měsíce.

11. Datový soubor football.csv obsahuje údaje o hmotnosti fotbalových hráčů z pěti různých týmu. Zjistěte, zda se hmotnost hráčů v jednotlivých týmech liší, a pokud ano, tak jak se liší.

Analýza rozptylu (ANOVA)

Jednoduché třídění

Situace: Máme k nezávislých výběrů (každý obecně s jiným rozsahem) a chceme testovat, zda se střední hodnota mezi výběry neliší nebo liší (tj. alespoň dvě jsou odlišné).
Rozdělení do výběrů většinou odpovídá nějaké k-úrovňové kategoriální veličině, tzv. faktoru.

Zajímá nás vliv k-úrovňového faktoru na měřenou spojitou veličinu.

Příklady: Vliv tří různých způsobů hnojení na výnos z půdy, závislost teploty na ročním období atd.

Hypotézy: H_0 : střední hodnoty jsou stejné ve všech výběrech
H_1 : alespoň dvě střední hodnoty jsou odlišné

Předpoklady:
- výběry jsou nezávislé,
- normální rozdělení ve všech výběrech,
- rozptyl je stejný ve všech výběrech.

Ověření předpokladů

- nezávislost - z principu sběru dat
- normalita - histogram, QQ grafy, testy normality (Shapirův-Wilkův test)
  vyvážený model (v každé skupině stejný počet pozorování) není velmi citlivý na porušení
- shoda rozptylů - neformální posouzení směrodatných odchylek, testy: Levenův, Bartlettův
  vyvážený model není velmi citlivý na porušení

Post-hoc mnohonásobná porovnávání: Jestliže H_0 zamítneme, zajímá nás, která dvojice se tedy liší. Jednou z možných metod je Tukeyho metoda.

Testy o střední hodnotě

1. jeden výběr srovnáváme s referenční hodnotou (předepsaná hodnota, známá z minulosti, expertní domněnka...)
   - normální data nebo dostatek pozorování →jednovýběrový t-test
   - symetrické rozdělení →Wilcoxonův jednovýběrový test
2. párová data (na jednom subjektu měříme dvakrát – např. před a po, různé laboratorní podmínky...)
   - normální data rozdílu nebo dostatek pozorování →párový t-test
   - symetrické rozdělení rozdílu →párový Wilcoxonův test
3. dva nezávislé výběry (jednu veličinu měříme ve dvou skupinách, za různých podmínek, atd.) sledujeme závislost spojité proměnné na kategoriální proměnné se dvěma kategoriemi
   - normální data nebo dostatek pozorování →dvouvýběrový t-test
   - obecné rozdělení →Wilcoxonův (Mannův-Whitneyův) test
4. více než dva nezávislé výběry (měříme ve více než dvou skupinách)
   - ANOVA
   - neparametrické alternativy ANOVA (např. Kruskalův-Wallisův test)

Závislost dvou proměnných

1. obě proměnné kategoriální →kontingenční tabulky
2. jedna proměnná kategoriální, druhá je spojitá →ANOVA nebo dvouvýběrový t-test
3. obě proměnné spojité →korelace, lineární regrese