Odhady a testování hypotéz
16.4.2013

Popis dat:. Management jedné velké nadnárodní firmy potřebuje zhodnotit mzdovou politiku a spokojenost svých zaměstnanců, aby mohl podniknout případné kroky v personální politice. Proto si nechal udělat průzkum mezi 100 náhodně vybranými zaměstnanci. Byl zaznamenán plat dotazovaného zaměstnance, doba strávená ve firmě, dosažené vzdělání, pohlaví a spokojenost ve firmě. Vše je zaznamenáno v souboru `spokojenost.dat`.

Proměnné v datech:

Id	identifikační číslo zaměstnance,
Pohlavi	pohlaví zaměstnance (0 - žena, 1 - muž)
Plat	měsíční plat (v Kč)
Doba	doba strávená ve firmě (v letech)
Vzdelani	stupeň vzdělání zaměstnance (1 - ZŠ, 2 - SŠ, 3 - VŠ)
Spokojenost	spokojenost zaměstnance ve firmě (1 - velmi spokojen, 2 - spíše spokojen, 3 - spíše nespokojen, 4 - velmi nespokojen)

1. Načtěte si do R Commanderu soubor `spokojenost.csv`.

2. Prohlédněte si data a základní popisné statistiky všech veličin. Zamyslete se nad tím, které veličiny jsou kategoriální, a případě změňte jejich formát na `factor`.

3. Nejprve se budeme zajímat o plat ve firmě:

   (a) Odhadněte střední hodnotu, směrodatnou odchylku, medián, 10% a 90% kvantily.

   (b) Dále nás bude zajímat, zda lze předpokládat, že má plat ve firmě normální rozdělení:
   - Nechte si vykreslit histogram. Z minulé hodnoty už víte, jak by měl vypadat histogram normálního rozdělení. Jak bychom tedy rozhodli o platu?
   - Pro lepší představu si do obrázku přidáme hustotu normálního rozdělení s příslušnými parametry:
     ```
     curve(dnorm(x,mean=mean(spokojenost$Plat),sd=sd(spokojenost$Plat)),10000,40000,add=T)
     ```
   - Na posouzení normality slouží také tzv. QQ-graf. Nechte si jej vykreslit pro veličinu Plat pomocí nabídky Graphs →Quantile-comparison plot. Společně si řekneme, co tento graf zobrazuje a jak by měl vypadat pro výběr z normálního rozdělení.

   (c) Necháme si spočítat odhad distribuční funkce platu ve firmě:

   ```
   e=ecdf(spokojenost$Plat)
   plot(e)
   ```

   Pro srovnání si do stejného obrázku přidáme distribuční funkci normálního rozdělení s příslušnými parametry:

   ```
   curve(pnorm(x,mean(spokojenost$Plat),sd(spokojenost$Plat)),add=T,col="red")
   ```

4. Podobným způsobem se zajíme o rozdělení doby strávené ve firmě a jeho charakteristiky.
   - Odhadněte střední dobu strávenou ve firmě, směrodatnou odchylku, medián a 10% a 90% kvantily.

– Zjistěte, zda je možné předpokládat, že se doba strávená ve firmě řídí normálním rozdělením. Vyzkoušejte všechny způsoby, které zatím znáte.

5. Zajímá nás, zda a jakým způsobem spolu souvisí plat a doba strávená ve firmě.

   (a) Podívejte se na vztah těchto dvou veličin graficky. Jaký vztah je z obrázku patrný?

   (b) Pomocí Statistics →Summaries →Correlation matrix si nechte vypsat tzv. (výběrový) korelační koeficient. Jak interpretujeme jeho hodnotu?

6. Ověřte domněnku, že je typický plat ve firmě (bez ohledu na pohlaví) roven 20 000 Kč měsíčně.

   (a) Co myslíme typickým platem? Jaký model pro data uvažujeme? Formulujte nulovou a alternativní hypotézu.

   (b) Proveďte vhodný test, tj. **jednovýběrový t-test**. Připomeňte si předpoklady tohoto testu. Zvolte hladinu významnosti $\alpha = 5\%$.

   Jednovýběrový t-test provedeme pomocí `Statistics→Means →Single-sample t-test`. Zde vhodně doplňte jednotlivá políčka.

   (c) Prohlédněte si pečlivě výstup testu. Společně si vysvětlíme, co znamenají jednotlivé položky.

   (d) Pro zodpovězení otázky nás zajímá zejména tzv. `p-value`, tj. p-hodnota testu (dosažená hladina testu). Vzpomeňte si, co udává a jak pomocí ní rozhodneme o výsledku testu (znáte z přednášky).

   (e) Jaký je náš závěr? Vyvracíme nebo nevyvracíme uvedenou domněnku o typickém platu?

   (f) Jak by se změnil náš závěr, pokud bychom testovali na hladině významnosti 1 %?

   (g) Vzpomeňte si na vzoreček pro testovou statistiku t-testu a ověřte, že ji R spočítalo správně.

7. Otestujte, zda byl náš předpoklad normality rozdělení platu v předchozím bodě vhodný. Použijte k tomu tzv. Shapirův-Wilkův test pomocí `Statistics→Summaries →Shapiro-Wilk test of normality`.

   (a) Co je zde nulová hypotéza a co alternativní hypotéza?

   (b) Jaký je náš závěr?

8. Vedení firmy tvrdí, že je typický plat ve firmě vyšší než 20 000 Kč měsíčně. Ověřte toto tvrzení.

   (a) Formulujte nulovou a alternativní hypotézu.

   (b) Proveďte jednovýběrový t-test (změňte vhodně `Alternative Hypothesis`).

   (c) Prohlédněte si výstup a porovnejte ho s výstupem z úkolu 3. V čem se liší? V čem se neliší?

   (d) Jaký je Váš závěr? Říká vedení firmy pravdu či nikoliv?

9. Jeden naštvaný zaměstnanec tvrdí, že je typický plat ve firmě určitě nižší než 21 000 Kč měsíčně. Ověřte jeho tvrzení.

   (a) Formulujte nulovou a alternativní hypotézu.

(b) Proveďte test a interpretujte jeho výsledek.

10. Management firmy tvrdí, že je střední doba strávená ve firmě vyšší než 10 let. Ověřte toto tvrzení.

    (a) Formulujte nulovou a alternativní hypotézu.

    (b) Jak je to s předpoklady t-testu? Lze ho použít?

    (c) Jaký je náš závěr?

11. Samostatně pracujte pouze s platem žen resp. mužů (dle Vašeho výběru). Ověřte následující tvrzení:
    A: Typický plat žen (mužů) je nižší než 22 000 Kč měsíčně.
    B: Typický plat žen (mužů) je roven 25 000 Kč měsíčně.
    C: Typický plat žen (mužů) je vyšší než 19 000 Kč měsíčně.
    D: Střední doba strávená ve firmě mezi ženami (muži) je nižší než 15 let.
    E: Střední doba strávená ve firmě mezi ženami (muži) je vyšší než 13 let.
    Vždy zformulujte nulovou a alternativní hypotézu a výsledek řádně interpretujte. Předpoklad normality ověřte jak graficky, tak také Shapirovým-Wilkovým testem.