Cvičení z NSTP129

# Popisná statistika

15. 12. 2011

**Úvodní nastavení.** Na disku `H` si založte speciální adresář na toto cvičení (např. `NSTP129` nebo `Statistika` apod.) Ze stránky si stáhněte soubory `vysledky.dat` a `spokojenost.dat`.

Otevřete si program `R` (Start→Všechny programy →Statistika→R apod). V `R` si spusťte `R Commander` pomocí Packages→Load Package, zde vyberte package `Rcmdr`. Mělo by se Vám otevřít nové okno s názvem `R Commander`. Zde si změňte pracovní adresář přes `File→Change working directory` na Váš právě založený na disku `H`. V rámci tohoto cvičení se omezíme na práci v `R Commander` .

## 1. Společná práce: Analýza výsledků písemky

Nejprve se společně seznámíme s tím, jak provádět v `R Commander` některé základní operace.

1. Načtení dat: Pomocí `Data→Import Data→from text file, clipboard or URL` si načteme data `vysledky.dat`. Do prvního políčka vyplňte název, pod kterých chcete mít tato data uložená (např. `vysledky` nebo `data` apod.) Ostatní položky neměňte.
   Pomocí `View data set` se podívejte na načtená data; prohlédněte si jednotlivé proměnné. Společně si okomentujeme, co jednotlivé sloupce znamenají.
2. Přes `Statistics →Summaries →Active data set` si nechte vypsat základní popisné statistiky všech veličin zahrnutých v datech. Ujistěte se, že víte, co jednotlivé položky znamenají.
   Všimněte si dále rozdílu mezi tím, co `R` vypsalo pro počet bodů a pro variantu písemky.
3. Na základě výše uvedeného výstupu zjistěte, jaký je průměrný počet bodů a medián počtu bodů. Liší se tyto dvě hodnoty? Co odhadují?
4. Kolik byste museli mít bodů, abyste patřili mezi 5 % (resp. 10 %) nejlepších studentů?
   (Uvědomte si, že chcete spočítat určitý výběrový kvantil. Toho lze dosáhnout přes `Statistics →Summaries →Numerical summaries`.)
5. Vykreslete si histogram počtu bodů (`Graphs →Histogram`). Interpretujte tento obrázek.
6. Na základě histogramu si rozmystele, zda je možné předpokládat, že počet bodů náhodně vybraného studenta je náhodná veličina s normálním rozdělením. Jak by měl vypadat histogram náhodného výběru z normálního rozdělení?
   Vykreslete si dále tzv. Q-Q plot (`Graphs →Quantile-comparison plot`). Jak by musel tento graf vypadat v případě normálního rozdělení?
7. Uveďte, kolik studentů psalo kterou variantu, a to jak číselně, tak procentuelně. Ilustrujte tato čísla také graficky pomocí `Graphs →Bar graph`.
8. Zajímají nás bodové zisky v jednotlivých variantách písemky.
   (a) Popište rozdělení počtu bodů v jednotlivých skupinách (průměr, max, min, směrodatná odchylka atd).
   (`Statistics→Summaries →Numerical summaries`, zde vybrat `by groups`).

(b) Vykreslete si boxploty počtu bodů v jednotlivých skupinách (`Graphs`→`Boxplot` a vyberte `Plot by groups`). Interpretujte, co boxplot znázorňuje a jak ho máme chápat.
Co usuzujete na základě tohoto obrázku o obtížnosti jednotlivých variant?

9. Podobně jako v předchozím bodě proveďte porovnání bodových zisků z pondělního a čtvrtečního cvičení (pomocí popisných statistik a obrázku).
K tomuto účelu si zaveďte novou veličinu `den`: `Data`→`Manage variables in active data set` →`Recode variables`. Zde vyberte v sloupci `Variables to recode` veličinu `varianta`, novou veličinu nazvěte `den`, do `Enter code directives` zapište

```
"A"="pondeli"
 "B"="pondeli"
 else="ctvrtek"
```

10. K úspěšnému napsání písemky bylo potřeba 12 bodů. Zajímá nás nyní úspěšnost.
    (a) Zaveďte novou veličinu `zapocet`, která pro každého studenta udává, zda byl úspěšný nebo nikoliv. Provedete to následujícím způsobem: `Data`→`Manage variables in active data set` →`Recode variables`. Zde vyberte v sloupci `Variables to recode` veličinu `body`, novou veličinu nazvěte `zapocet`, do `Enter code directives` zapište

    ```
    12:20="ano"
     else="ne"
    ```

    (b) Zjistěte úspěšnost na písemce.
    (c) Vypište si tabulku úspěšnost vs. den (`Statistics` →`Contingency tables` →`Two-way table`). Jaká je úspěšnost v pondělním cvičení?
    (d) Podobným způsobem zjistěte, jaká je úspěšnost ve variantě, kterou jste psali Vy.

## 2. Samostatná práce

**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) |

Řešte následující úkoly:

1. Načtěte si do `R Commander` data `spokojenost.dat`. Prohlédněte si data.
2. Které proměnné v datovém souboru odpovídají spojité náhodné veličině?
Jsou-li některé veličiny kategoriální a jsou kódovány pomocí čísel, musíte `R`-ku sdělit, že je má chápat jako tzv. faktory: `Data`→`Manage variables in active data set` →`Convert numeric variables to factors`.
3. Co je možné říci o složení zaměstnanců dané firmy z hlediska vzdělání? Nechte si vypsat vhodnou tabulku a vykreslit vhodný obrázek.

4. Jak byste popsali spokojenost zaměstnanců v dané firmě? (Použijte opět vhodnou tabulku a vhodný obrázek.)
5. Popište platové podmínky v dané firmě (zatím bez rozlišení vzdělání či pohlaví).
6. Pomocí vhodného obrázku ilustrujte (empirické) rozdělení platu v dané firmě. Rozmyslete si, zda je možné předpokládat, že má veličina plat normální rozdělení.
7. Zajímá nás, zda se liší plat v závislosti na vzdělání zaměstnanců.
   (a) Pomocí vhodných charakteristik popište plat v závislosti na vzdělání.
   (b) Nakreslete vhodný obrázek, který porovná plat v uvedených třech skupinách dle vzdělání.
   (c) Na základě výše spočtených popisných statistik a obrázku si udělejte názor na závislost resp. nezávislost platu na vzdělání.
8. Přes `Graphs` →`Scatterplot` si vykreslete obrázek, který ukáže, jak spolu souvisí plat a doba strávená v dané firmě. Myslíte si, že jsou tyto dvě veličiny nějakým způsobem závislé? Jak? Spočtěte dále výběrovou korelaci mezi dobou strávenou ve firmě a platem (`Statistics` →`Summaries` →`Correlation matrix`). Interpretujte toto číslo.
9. Uvažujme 2 kategorie zaměstnanců podle toho, jak dlouho již pracují v dané firmě: ne více než 15 let (ti, co pracují krátce), a více než 15 let (pracují dlouho).
   (a) Vytvořte si novou veličinu přes `Data`→`Manage variables in active data set` →`Recode variables`. Zde vyberte v sloupci `Variables to recode` veličinu `Doba`, novou veličinu nazvěte např. `Doba2`, do `Enter code directives` zapište
   ```
   0:15="kratce"
   else="dlouho"
   ```
   (b) Popište rozdělení této kategoriální veličiny.
   (c) Pomocí vhodných charakteristik a obrázků rozhodněte, zda plat souvisí s dobou strávenou ve firmě (kde uvažujete zmíněné dvě skupiny podle délky strávené ve firmě).

Na závěr si nezapomeňte **uložit svou práci**. Uložte si zejména skript (příkazy do R, které se objevovali v horní části `R Commander` ) a output (výsledky v dolní části `R Commander` ). `R` workspace neukládejte.