Popisné statistiky
26.2.2013

Úvodní nastavení.
- Otevřete si program R (např. pomocí ikonky "R"na ploše nebo přes nabídku Start→...).
- Změňte si pracovní adresář pomocí File→Change working directory na Váš právě založený adresář statistika.

1. **Popis kvantitativních veličin.**
   (a) Budeme zkoumat veličinu udávající výšku studentů (bez rozdílu pohlaví). Pomocí Statistics →Summaries →Numerical summaries si znovu vypište popisné statistiky této veličiny. Které charakterizují polohy a které variabilitu?
   (b) Vykreslete si histogram výšky pomocí Graphs →Histogram a uvědomte si, co znázorňuje. Co z něj umíte vyčíst?
   (c) Podívejte se, jak se liší jednotlivé formy histogramů, které nám R nabízí.
   (d) Volbou Number of bins můžeme volit počet intervalů, které jsou v histogramu uvažovány. Podívejte se, jak se histogram mění, zvyšujeme-li a snižujeme-li jejich počet.
   (e) Vykreslete si krabicový graf (boxplot) veličiny výška pomocí Graphs →Boxplot. Porovnejte obrázek s popisnými statistikami. Jsou v datech nějaká odlehlá pozorování? Zjistěte, která to jsou.
2. Stejným způsobem si prohlédněte hmotnost studentů a velikost bot. Zaměřte se mimo jiné na odhalení možných „chyb“ v datech a podezřelých záznamů.
3. Ověřete si data v Excelu a opravte chyby v datech a případně odstraňte podezřelá pozorování, na která jste narazili. Data uložte (jako formát csv) a načtěte je znovu do R.
4. **Počítání nových proměnných.**
   (a) Zaveďte novou veličinu, která bude udávat výšku v metrech. Pojmenujte ji např. vyska.m a vytvořte ji pomocí Data →Manage variables in active data set →Compute new variable. Zde zadejte `vyska/100`.
   (b) Nechte si vypsat základní popisné statistiky pro tuto novou veličinu. Porovnejte je s charakteristikami výšky v cm. Jak se změnil průměr a jak směrodatná odchylka?
   (c) Zaveďte si veličinu, vyska.m2, která je dána jako výška v cm minus 100, tj. jde o výšku v cm nad 1 metr. Jak se liší popisné statistiky této veličiny a veličiny vyska.
5. Spočítejte věk studentů v letech a zjistěte, jaký byl nejmladší a nejstarší student na přednášce.
6. Spočtěte rozdíl věku rodičů studentů.
   - Jaký je průměrný rozdíl věků rodičů? Jaká je minimální a maximální hodnota v datech?
   - Identifikujte „extrémní“ případy a prohlídněte si jejich záznam.
7. Zaveďte si novou veličinu udávající BMI (body mass index) studentů.

- Zadejte Data→Manage variables in active data set →Compute new variable... . Zde uveďte jméno nové proměnné BMI a výraz, jakým se má spočítat `vaha/(vyska/100)^2`.
- Prohlédněte si základní popisné statistiky BMI studentů.
- Nechte si vykreslit vhodné popisné grafy.
- Identifikujte „odlehlá“ pozorování a prohlédněte si jejich záznamy.

8. Standardně se uvádí, že „normální“ váha odpovídá BMI v rozmezí 18.5 až 24.9. Zaveďte novou veličinu určující pro každého studenta, zda má podváhu, normální váhu nebo nadváhu.
   - Vyberte z nabídky `Data→Manage variables in active data set →Recode variables`. Zde uveďte jméno nové veličiny (např. nadvaha). Do volného okénka pak napište:
     ```
     0:18.4="podvaha"
     18.5:24.9="normalni"
     else="nadvaha"
     ```
     (jednotlivé kategorie si můžete nazvat i jinak).
   - Zjistěte, kolik studentů má podváhu a kolik nadváhu.
   - Nechte si vykreslit vhodné grafy, které by předchozí počty graficky ilustrovaly.

9. Zaveďte si novou veličinu, pomocí které zjistíte, kolik procent studentů se narodilo na jaře, v létě, na podzim a v zimě. Berte jaro jako duben, květen, červen, léto jako červenec, srpen, září atd.
   - Vypište si procentuální zastoupení jednotlivých ročních období.
   - Namalujte si vhodný obrázek.

10. **Výběr podmnožiny dat.** Nyní si prohlédneme popisné statistiky jednotlivých veličin odpovídajících pouze loňským studentům. Z původních dat si proto vybereme pouze podmnožinu odpovídající roku 2012:
    - Vyberte Data →Active data set →Subset active data set. Zde do políčku Subset expression uveďte Rok==2012 a jako název nového datového souboru napište např. studenti2012.
    - Podívejte se na základní popisné statistiky jednotlivých veličin. Porovnejte je s údaji pro celá data. Je loňský rok v něčem jiný?
    - Na konci se přepněte zpět do celých dat (za všechny roky 2006 až 2012).

11. Zaměřte se jen na studenty s podváhou. Zjistěte
    - kolik je mezi nimi procent žen,
    - kolik mají průměrně sourozenců.

12. **Uložení práce:.** Uložte si File →Save R Workspace pracovní prostředí z R, příště s ním budeme pokračovat.