Cvičenie 1

Generovanie dát

Na cvičeniach si budeme často generovať náhodné dáta s požadovanými vlastnosťami. Pri generovaní je vhodné si najprv nastaviť inicializáciu generátora náhodných čísel (inak by sa vygenerovali pri každom spustení rozdielne dáta). Nasledujúci príklad vygeneruje náhodný vektor 100 hodnôt podľa normálneho rozdelenia s 0 strednou hodnotou a štandardnou odchýlkou 0.5:

```
set.seed(1234)
x <- rnorm(100, mean=0, sd=0.5)
```

Hodnoty môžeme vypočítať aj zadaním funkcie. Nasledujúce príkazy vygenerujú postupnosť čísel x od 0 do 10 s krokom 0.5 pre ktorú vypočítajú hodnoty lineárnej funkcie y = 3 x + 2.

```
x <- seq(from=0, to=10, by=0.5)
f <- function(x) { 3*x + 2 }
y <- sapply(x,f)
plot(y ~ x)
```

K vygenerovaným dátam môžeme pridať náhodný šum s požadovaným rozdelením pravdepodobnosti. Nasledujúci príklad vygeneruje náhodný vektor s uniformným rozdelením s intervalom [-1,1], ktorý pripočíta k hodnotám y:

```
noise <- runif(length(y), min=-1, max=1)
y <- y + noise
plot(y ~ x)
```

Inštalácia príkladov

Na cvičení budete pracovať s dátami, ktoré sú súčasťou balíka „AER“. Balík „AER“ obsahuje dátové súbory pripravené pre publikáciu „Applied Econometrics with R”. Nasledujúce príkazy nainštalujú balík do prostredia R a načítajú dátovú množinu CPS1985. Dátovú množinu si premenujeme na cps.

```
install.packages(“AER”)
data(“CPS1985”, package=”AER”)
cps <- CPS1985
```

Popis dátovej množiny môžete získať príkazom str. Príkaz head vypíše niekoľko prvých záznamov. Dáta obsahujú demografické údaje z prieskumu obyvateľstva z mája roku 1985 a obsahujú základné atribúty ako napr. príjem (wage), stupeň vzdelania, vek, atď.

```
str(cps)
head(cps)
```

Príkazom attach si vytvoríme odkazy na stĺpce dátovej množiny, aby sme sa na nich mohli priamo odkazovať.

```
attach(cps)
```

Základné štatistiky pre číselné atribúty

Prehľadové základné štatistiky pre číselné atribúty je možné vypísať príkazom summary. Jednotlivé štatistiky môžeme vypočítať aj samostatne funkciami mean (priemer), sd (štandardná odchýlka), var (variancia), atď. Nasledujúce príkazy vypočítajú základné štatistiky pre príjem.

```
summary(wage)
mean(wage)
sd(wage)
```

Graficky si môžeme zobraziť histogram hodnôt. Štandardne sa na histograme zobrazí početnosť príkladov, ktoré spadajú do daného rozsahu histogramu. Ak chceme zobraziť iba odhad pravdepodobnosti (počet príkladov s hodnotami v danom rozsahu / celkový počet príkladov), parameter freq nastavíme na FALSE. Hodnoty pravdepodobnosti môžeme aproximovať funkciou hustoty density. Nasledujúce príkazy vykreslia histogram s aproximovanou funkciou hustoty.

```
hist(wage)
hist(wage, freq=FALSE)
lines(density(wage))
```

Aproximovaná funkcia hustoty nám slúži napr. na lepšie vizuálne porovnanie či majú hodnoty normálne rozdelenie, ktoré má nasledujúci priebeh:

```
curve(dnorm, from=-5, to=5)
```

Dáta za príjem sú vychýlené k nižším hodnotám (tzn. väčší počet ľudí má menší plat). Aby sme dáta normalizovali, skúsime ich transformovať logaritmickou funkciou. Nasledujúce príkazy vykreslia histogram a aproximovanú funkciu hustoty pre logaritmicky transformované dáta.

```
hist(log(wage), freq=FALSE)
lines(density(log(wage)))
```

Základné štatistiky pre faktory (nominálne resp. ordinálne atribúty)

Pre faktory prehľad zobrazí početnosť jednotlivých hodnôt, ktoré je možné graficky zobraziť stĺpcovým grafom.

```
summary(occupation)
barplot(table(occupation))
```

Závislosti medzi dvoma číselnými atribútmi

Základnou štatistikou je korelačný koeficient, ktorý meria lineárnu závislosť medzi dvoma číselnými atribútmi. Nasledujúce príkazy vypočítajú korelačný koeficient medzi príjmom a vzdelaním. Závislosť si môžeme zobraziť aj graficky kde môžeme potvrdiť linearitu.

```
cor(wage, education)
cor(log(wage), education)
plot(log(wage) ~ education)
```

Závislosti medzi dvoma faktormi

Závislosť medzi dvoma faktormi môžeme zistiť z kontingenčnej tabuľky, ktorá určuje početnosť spoločného výskytu hodnôt dvoch faktorov. Početnosť je možné prehľadne zobraziť graficky vykreslením, napr. mozaikovým grafom. Podobne ako sa pre číselné atribúty používa korelačný koeficient, tak sa pre faktory používa chi-kvadrát test. Ak je kritická hodnota $p < 0.05$ atribúty sa považujú za závislé.

```
table(gender, occupation)
plot(gender ~ occupation)
chisq.test(table(gender, occupation))
```

Závislosti medzi jedným číselným atribútom a jedným faktorom

Pomocou funkcie tapply vieme vypočítať štatistiky číselného atribútu pre jednotlivé hodnoty faktora, napr. nasledujúci príkaz vypočíta priemerný plat samostatne pre mužov a ženy (hodnoty faktora gender). Súhrne je možné závislosti medzi číselným atribútom a faktorom zobraziť pomocou krabicového grafu (box-plot).

```
tapply(log(wage), gender, mean)
boxplot(log(wage) ~ gender)
```

Podrobnejšie je možné zobraziť graf kvantilov (q-q plot). Najprv si odfiltrujeme dáta podľa pohlavia do dvoch podmnožín ktorých kvantily zobrazíme funkciou qqplot. Do grafu pridáme aj 45° referenčnú priamku predeľujúcu zobrazenie. Keďže väčšina kvantilov je pod referenčnou čiarou, muži zarábajú viac.

```
mwage <- subset(cps, gender == "male")$wage
fwage <- subset(cps, gender == "female")$wage
qqplot(mwage, fwage, xlim=range(wage), ylim=range(wage))
abline(0,1)
```

Úlohy na cvičení

1. Zobrazte histogramy pre všetky číselné atribúty z dátovej množiny CPS, zistite ktorý z nich má približne normálne rozdelenie. Transformujte atribúty použitím logaritmickej funkcie a zistite, či majú transformované dáta približne normálne rozdelenie.
2. Zobrazte početnosti hodnôt pre všetky faktory (nominálne resp. ordinálne atribúty). Zistite, ktoré z týchto rozdelení je približne uniformné. Zobrazte stĺpcové grafy s relatívnymi početnosťami (počet hodnôt/celkový počet príkladov).
3. Vypočítajte korelačnú maticu medzi všetkými číselnými atribútmi pomocou funkcie cor (číselné atribúty si môžete vybrať z množiny x použitím funkcie x[sapply(x, is.numeric)]). Pre dvojicu atribútov s najväčšou koreláciou nakreslite X-Y graf ich závislosti.
4. Vypočítajte chi-kvadrát test pre všetky dvojice faktorov. Pre faktory, ktoré majú najväčšiu závislosť nakreslite mozaikový graf.
5. Vypočítajte a zobrazte priemernú mzdu pre všetky hodnoty a všetky faktory samostatne. Pre vypočítané hodnoty zobrazte krabicové grafy. Zistite kde sú najväčšie rozdiely v platoch a pre danú závislosť nakreslite q-q graf.
6. Vygenerujte náhodné vektory x a y o veľkosti 100 hodnôt z normálneho rozdelenia so strednou hodnotou 1 a štandardnou odchýlkou 2. Zobrazte ich závislosť na X-Y grafe a vypočítajte korelačný koeficient. Pridajte do oboch vektorov jednu extrémnu hodnotu (napr. 10, 20) pomocou funkcie c(). Vypočítajte korelačný koeficient na zmenených dátach a zobrazte X-Y závislosť.
7. Vygenerujte náhodné vektory x a y o veľkosti 100 hodnôt z normálneho rozdelenia so strednou hodnotou 1 a štandardnou odchýlkou 2. Vypočítajte korelačný koeficient a zobrazte ich závislosť. Postupne vygenerujte nové hodnoty vo vektore y s rovnakým stredom 1 ale zmenšujúcou sa štandardnou odchýlkou s hodnotami 2, 1.5, 1 a 0.5. Pre zmenené hodnoty y vypočítajte korelačný koeficient a zobrazte X-Y priebeh.
8. Vygenerujete si sekvenciu dát x od 0 do 10 s krokom 0.5. Pre vygenerované hodnoty x vypočítajte hodnotu funkcie y = 2.5x + 2 a pridajte k nej šum s normálnym rozdelením s 0 strednou hodnotou a štandardnou odchýlkou 0.5. Zobrazte priebeh funkcie na X-Y grafe. Postupne pridávajte viac šumu s väčšou štandardnou odchýlkou 1, 1.5, 2.0, 3.0. Po každej zmene zobrazte dáta a vypočítajte korelačný koeficient medzi x a y.
9. Vygenerujete si vektor x so 100 hodnotami z uniformného rozdelenia s min 0 a max 10. Pre vygenerované hodnoty x vypočítajte hodnotu funkcie y = 2*sin(x) + x a pridajte k nej šum s normálnym rozdelením s 0 strednou hodnotou a štandardnou odchýlkou 0.1. Zobrazte priebeh funkcie na X-Y grafe a vypočítajte korelačný koeficient medzi x a y.