Popisné statistiky, testování hypotéz
19.10.2012

Úvodní nastavení.

– Otevřete si `R Studio` a pomocí `Workspace` →`Load Workspace` si nahrajte soubor s příponou .RData z minulé hodiny. Pokud by se Vám to nedařilo, načtěte si data stejným způsobem jako minule ze souboru csv.

Popisné statistiky II

1. Zajímá nás, zda se nějak liší obsah alkoholu v jednotlivých třech odrůdách vína. Případnou závislost budeme zkoumat pomocí obrázku a vhodných číselných charakteristik.

```
plot(alcohol~vintage)
# je to totez jako plot(vintage,alcohol) nebo boxplot(vintage,alcohol)

library(lattice)
histogram(~alcohol|vintage)

tapply(alcohol,vintage,summary)
tapply(alcohol,vintage,sd)
```

2. Stejným způsobem se podívejte na obsah flavonoidů (antioxidantů) v závislosti na odrůdě.

3. Podíváme se na vztah intenzity barvy vína a obsahu alkoholu.

```
plot(col_int~alcohol)
abline(lm(col_int~alcohol),col="blue")
lines(lowess(col_int~alcohol),col="red")

library(car)
scatterplot(col_int~alcohol)

plot(col_int~alcohol,col=as.numeric(vintage),pch=as.numeric(vintage))

scatterplot(col_int~alcohol|vintage)
```

   Pro číselný popis použijeme korelační koeficient:

```
cor(col_int,alcohol)
cor(col_int,alcohol,method="spearman")
```

4. Nyní se podíváme stejným způsobem na vztah mezi alkoholem a flavonoidy.

5. Vyzkoušejte také příkazy

```
plot(wine1)
cor(wine1[,1:10])

pairs(~alcohol+malic_acid+magnesium+flavonoids+col_int)

scatterplotMatrix(~alcohol+magnesium+flavonoids+col_int)
```

6. A konečně se podíváme i na popis dvou kvalitativních veličin. Protože máme v datech jen jednu takovou proměnnou (vintage), musíme si nějakou další vytvořit:

```
silne= cut(alcohol,breaks=c(0,13,100),labels=c("slabe","silne"))

table(vintage,silne)
prop.table(table(vintage,silne),margin=1)

barplot(table(silne,vintage),beside=T,legend=T)
```

Ověření normality

7. Předpokladem velkého množství testů je normalita rozdělení dat. Zejména v menších souborech je potřeba normalitu dat ověřit.

```
hist(alcohol,prob=T)
lines(density(alcohol))
curve(dnorm(x,mean(alcohol),sd(alcohol)),min(alcohol),max(alcohol),add=T,col="red")

qqnorm(alcohol)
qqline(alcohol)

# nebo po zavolani knihovny library(car)
qqPlot(alcohol, dist="norm")

shapiro.test(alcohol)
```

Jednovýběrové testy

8. Otestujte, zda je střední hodnota koncentrace alkoholu rovna 13%.

```
t.test(alcohol,mu=13)
```

9. Je možné prokázat, že je obsah popelovin vyšší než 2 g/L?

```
t.test(ash,mu=2,alternative="greater")
```

10. Neparametrický Wilcoxonův test pro 8. a 9.

```
wilcox.test(alcohol,mu=13)
wilcox.test(ash,mu=2,alternative="greater")
```

Statistická indukce

Statistika
– popisná $\leftrightsquigarrow$ popis konkrétních dat
– induktivní $\leftrightsquigarrow$ zobecnitelné závěry

Předpoklad: Naše data tvoří reprezentativní **náhodný výběr** = nezávislé realizace náhodné veličiny s neznámým rozdělením

1. Teorie odhadu: Na základě dat chceme odhadnout rozdělení a jeho charakteristiky (střední hodnota atd)
– bodové odhady $\leftrightsquigarrow$ popisné statistiky
– intervalové odhady (intervaly spolehlivosti, konfidenční intervaly)

2. Testování hypotéz
– ověřování platnosti nějakého výroku
– rozhodujeme na základě statistického testu

Hypotéza je výrok, o jehož platnosti chceme rozhodnout na základě nasbíraných dat:
– nulová hypotéza H_0
– alternativní hypotéza H_1.

Statistický test = rozhodovací pravidlo (postup), na jehož základě se rozhodneme:
– **zamítáme** H_0 ve prospěch H_1 (*naše data svědčí proti* H_0, **prokazujeme** *platnost* H_1)
– **nezamítáme** H_0 (*na základě našich dat nelze* H_0 *zamítnout, naše data nejsou v rozporu s* H_0)

Můžeme se dopustit **chyby**:

	Skutečnost	
Rozhodnutí	H_0 platí	H_1 platí
zamítáme H_0	chyba 1.druhu	OK
nezamítáme H_0	OK	chyba 2.druhu

Chyba 1.druhu je závažnější (falešně něco prokazujeme) $\rightsquigarrow$ její pravděpodobnost chceme kontrolovat. Volíme α = maximální přípustná pravděpodobnost chyby 1.druhu (většinou $\alpha = 0.05$) $\rightarrow$ test na hladině α.

Síla testu $= 1 - \mathsf{P}(\text{chyba 2.druhu}) = \mathsf{P}(\text{prokázání platné } H_1)$
– obecně o ní moc nevíme, můžeme ovlivnit volbou testu, počtu měření atd.

P-hodnota (angl. p-value) = dosažená hladina testu
– pravděpodobnost, že dostaneme výsledek, který stejně nebo ještě více svědčí proti H_0, jestliže H_0 ve skutečnosti platí
– „stupeň důvěry“ v platnost H_0
–

$$\text{p-hodnota} \leq \alpha \quad \Longrightarrow \quad \text{zamítáme } H_0 \text{ a prokazujeme tak } H_1$$