# Kontingenční tabulky a analýza rozptylu (ANOVA)

9.1.2020

1. Tabulka 1 shrnuje osudy pasažérů lodě Titanic, která tragicky ztroskotala v roce 1912. Zajímá nás, zda existuje nějaká souvislost mezi třídou, ve které cestující cestoval, a přežitím, nebo zda jsou tyto dva faktory nezávislé.

| | Přežil | |
|---|---|---|
| Třída | Ne | Ano |
| 1 | 122 | 203 |
| 2 | 167 | 118 |
| 3 | 528 | 178 |
| Posádka | 673 | 212 |

Tabulka 1: Data o Titanicu.

Jaký model teď budeme uvažovat a pro co? Co všechno je v modelu náhodné a co naopak není?

2. Nejprve si do R zadáme tabulku 1.

```
titanic=matrix(c(122, 167, 528, 673, 203, 118, 178, 212),ncol=2)
dimnames(titanic)=list(Trida=c("1","2","3","Posadka"),Prezil=c("Ne","Ano"))
titanic
```

Zkontrolujte, že máme tabulku správně zadanou. Kdybychom si chtěli dopočítat marginální četnosti, provedeme to následovně:

```
apply(titanic,1,sum)
apply(titanic,2,sum)
```

3. Opakování: Vhodným obrázkem graficky ilustrujte marginální rozdělení zkoumaných dvou veličin.

4. Vrátíme se zpět ke kontingenční tabulce. Podíváme se na tabulky relativních četností

```
prop.table(titanic)
prop.table(titanic,marg=1)
prop.table(titanic,marg=2)
```

Co nám jednotlivé relativní četnosti odhadují? Jak by měly tabulky přibližně vypadat v případě nezávislosti?
Podíváme se na tutéž věc i graficky:

```
barplot(titanic,beside=T,legend=T)
barplot(prop.table(titanic,mar=2),beside=T,legend=T)
barplot(t(titanic),beside=T,legend=T)
barplot(prop.table(t(titanic),mar=2),beside=T,legend=T)
```

Prozkoumejte jednotlivé obrázky a jak se mezi sebou liší. Co si na základě čísel a grafů myslíte o vztahu zkoumaných dvou veličin? Jsou nezávislé?

5. Provedeme $\chi^2$ test nezávislosti.

```
chisq.test(titanic,correct=FALSE)
```

   Jaký je náš závěr?

   (a) Připomeneme, jak se spočítá testová statistika $\chi^2$ testu:
   ```
   a1=apply(titanic,1,sum)
   a2=apply(titanic,2,sum)
   n=sum(titanic)

   E=a1%o%a2/n
   sum((titanic-E)^2/E )
   ```
   Kolik má stupňů volnosti příslušné asymptotické $\chi^2$ rozdělení? Jak se toto číslo spočítá?

   (b) Ještě si prohlédneme jednotlivé položky, které máme k dispozici po použití funkce chisq.test:
   ```
    CH=chisq.test(titanic,correct=FALSE)
   names(CH)

   CH$residuals
   ```
   Co přesně jsou tato „rezidua“? Které kategorie tabulky nejvíce přispívají k výsledné hodnotě $\chi^2$ statistiky a tím „porušují“ nezávislost?

   Jak bychom shrnuli naše poznatky týkající se přežití pasažérů z jednotlivých tříd?

6. A není to s tím Titanicem celé trochu jinak? Podíváme se na úplně kompletní data, která jsou k dispozici v R :

```
data(Titanic)
Titanic

#nase tabulka 1
apply(Titanic,c(1,4),sum)

#dalsi tabulky:
(t1=apply(Titanic,c(2,4),sum))
prop.table(t1,mar=1)

(t2=apply(Titanic,c(1,2),sum))
prop.table(t2,mar=1)
```

7. Uvažujme $2 \times 2$ tabulku uloženou v `t1`, která shrnuje vztah pohlaví a přežití pasažérů.

   (a) Otestujte nezávislost těchto dvou veličin pomocí $\chi^2$ testu.

   (b) Podívejte se na problém jinak a otestujte shodu pravděpodobností přežití pro muže a pro ženy, pomocí funkce prop.test.

(c) Jak se liší uvažované dva modely v (a) a (b)? V jakém vztahu jsou testové statistiky v (a) a (b)?
(d) Uvažujme model jako v (a). Jaké je rozdělení marginálních řádkových četností $n_{1+}$ a $n_{2+}$? Jaké je rozdělení četností v tabulce, podmíníme-li marginálními četnostmi $n_{1+}$ a $n_{2+}$?

8. Odhadněte poměr šancí na přežití žen a mužů z tabulky `t1`.

```
# nebo rychleji primo z tabulky:
t1[1,1]*t1[2,2]/(t1[1,2]*t1[2,1])
```

Jak budeme interpretovat toto číslo? Jaká hodnota by odpovídala nezávislosti?
Pro poměr šancí můžeme zkonstruovat i interval spolehlivosti:

```
sd=sum(1/t1)
log(odds.ratio)+c(-1,1)*qnorm(0.975)*sqrt(sd)
exp(log(odds.ratio)+c(-1,1)*qnorm(0.975)*sqrt(sd))
```

Měli muži stejnou šanci na přežití jako ženy?

9. Rozhodněte, zda měly ženy více než pětkrát větší šanci na přežití než muži.

Analýza rozptylu (ANOVA). Na pěti různých místech A, B, C, D a E bylo z řeky vyloveno vždy 7 ryb a byla zjišťována koncentrace mědi v jejich játrech. Naměřená data jsou obsažena v datech Med.txt. Otázkou je, zda je znečištění řeky stejné na všech zkoumaných místech nebo zda se nějak významně liší.

10. Stáhněte, načtěte a prohlédněte si data Med.txt. V analýze budeme pracovat s logaritmem koncentrace, tj. s proměnnou lnCu.
    – Porovnáme průměry a směrodatné odchylky na jednotlivých místech. Vše si znázorníme i graficky.

```
attach(Med)
tapply(lnCu,Misto,mean)
tapply(lnCu,Misto,sd)

boxplot(lnCu~Misto,col="orange")
```

11. Na náš problém budeme chtít použít analýzu rozptylu. Připomeňte si, jaké všechny předpoklady tato metoda má. Formulujte $H_0$ a $H_1$.

12. Dále si připomeňte, na jakých principech je analýza rozptylu založena: co je to celkový součet čtverců, součet čtverců skupin a reziduální součet čtverců. Znázorněte pro naše data graficky (viz R kód).

13. Otestujte, zda je znečištění řeky na zkoumaných pěti místech stejné. Test provedeme následovně:

```
model<-aov(lnCu~Misto)
anova(model)
#totez jako
summary(model)
```

Jaký je závěr?

14. Manuální výpočet jednotlivých položek z tabulky analýzy rozptylu:

```
(ni=table(Misto))
(N=sum(ni))
(SSc=sum((lnCu-celk.prumer)^2) )
(SSa=sum(ni*(prumery-celk.prumer)^2))
(SSe=sum((lnCu-fitted(model))^2))
# nebo zde taky takto:
(SSe=sum((lnCu-rep(prumery,7))^2))

p=length(levels(Misto))

SSa/(p-1)
SSe/(N-p)

# testova statistika
(Fa=SSa/(p-1)/(SSe/(N-p)))

# p-hodnota
1-pf(Fa,df1=p-1,df2=N-p)
```

15. Proč jsme nemohli provést test tak, že bychom porovnali (na hladině 5 % pomocí t-testu nebo Welchova testu) všechny dvojice míst a zamítli bychom $H_0$, pokud alespoň jeden z testů odhalí rozdíl?
Jak lze modifikovat výše uvedený postup, abychom mohli provést mnohonásobné porovnání jednotlivých míst na celkové hladině 5 %?

```
lev.mista=levels(Misto)
alpha=0.05
m=5*4/2

#vsechny testy na hladine:
alpha/m
for(i in 1:4) for(j in (i+1):5){
  print(paste(lev.mista[i],"-",lev.mista[j]))
  print(t.test(lnCu[Misto==lev.mista[i]],lnCu[Misto==lev.mista[j]],var.equal=T)$p.val)
}
```

Která místa se významně liší?
Můžeme si vytvořit i přehlednější tabulkový výstup, viz R kód.

16. Podíváme se, jaký je vztah dvouvýběrového t-testu a analýzy rozptylu pro případ $p = 2$. Z našich dat si tedy vybereme pouze místa A a B a ta porovnáme jak t-testem, tak pomocí $F$-testu.

```
detach(Med)
AB=Med[Med$Misto=="A"|Med$Misto=="B",]
AB$Misto=factor(AB$Misto)

(t=t.test(lnCu~Misto,data=AB,var.equal=T))
(a=anova(modelAB<-aov(lnCu~Misto,data=AB)))
```

Je nějaká souvislost mezi uvedenými dvěma testy? Pomocí jakých rozdělení jsou spočtené výše uvedené p-hodnoty?

Samostatná práce.

1. Data dieta.txt porovnávají efekt tří diet na hubnutí. Pro 76 osob máme k dispozici jejich pohlaví, věk, výšku, typ diety (kódováno 1, 2 a 3) a hmotnost před a po 6 týdnech diety. Rozhodněte, zda mají zkoumané tři diety stejný vliv na úbytek hmotnosti, nebo zda je mezi nimi významná odlišnost.

   Návod: Načtěte si data a spočtěte si proměnnou `ubytek`, což je rozdíl mezi váhou před a po dietě. Proměnnou `Diet` si musíte nastavit na typ `factor`.

   ```
   dieta$ubytek=dieta$pre.weight-dieta$weight6weeks
   dieta$Diet=factor(dieta$Diet)
   attach(dieta)
   ```

   Následně proveďte analýzu rozptylu. Problém si také vizualizujte pomocí boxplotů. Připomeňte si uvažovaný model. Pokud zjistíte, že se efekt na hubnutí liší, zjistěte mezi kterými dietami je statisticky významný rozdíl.

2. Závěrečné opakování: Rozhodněte, jaký test (postup) byste použili pro zkoumání následujících problému vztahujících se k datům dieta.txt:
   - Měla dieta 1 efekt na hubnutí? Tj. mají osoby po jejím absolvování nižší hmotnost než před tím?
   - Je pravdivé tvrzení, že díky dietě 3 lidé zhubnou v průměru více než 5 kg?
   - Závisí úbytek hmotnosti na pohlaví?
   - Je pravděpodobnost zhubnutí stejná pro muže a pro ženy?
   - Jsou věkové skupiny $< 30$ let, $30 - 50$ let, $> 50$ zastoupeny v populaci hubnoucích lidí v poměru $1 : 2 : 1$?
   - Je pravděpodobnost zhubnutí stejná pro výše uvedené tři věkové skupiny?