Analýza rozptylu— pokračování

9.11.2012

Úvodní nastavení.

– Otevřete si R Studio. Z internetu si stáhněte data vodivost.csv.

Popis dat: Data jsou součástí studie, která zjišťovala vliv obsahu vybraných chemických látek na vlastnosti vody. Konkrétně nás bude zajímat, zda přítomnost těchto látek ovlivňuje vodivost vody.

conductivity	vodivost vzorku při 25 stupních Celsia,
NO3	obsah dusičnanů,
Fe	obsah železa,
Cl2	obsah chlóru.

1. Načtěte si do R data vodivost.csv a podívejte se na základní popisné statistiky všech proměnných v souboru.

2. Pomocí vhodných obrázků se podívejte na závislost vodivosti na obsahu dusičnanů, na obsahu železa a obsahu chlóru (vždy zvlášť).

3. Pro zjednodušení následné analýzy nebudeme uvažovány přímo obsahy chemických látek (číselné hodnoty), ale omezíme se pouze na faktorové veličiny udávající, zda byl obsah příslušné látky vysoký (nad mediánem) nebo nízký (pod mediánem). Vytvoříme si proto tři nové proměnné, které budou nabývat jen dvou kategorií (vysoká a nízka hladina dané látky). Pro chlór:

   ```
   Cl2_level=ifelse(Cl2>median(Cl2),1,0)
   Cl2_level=factor(Cl2_level,labels=c("nizka","vysoka"))
   ```

   a analogicky pro další dvě látky.

4. Nejprve nás bude zajímat vliv dusičnanů společně s vlivem železa na vodivost. Podíváme se proto na závislost vodivosti na těchto faktorech:

   ```
   tapply(conductivity,Fe_level,mean)
   tapply(conductivity,Cl2_level,mean)
   ```

5. Podívejte se na grafické znázornění závislosti vodivosti na úrovních daných dvou chemických látek.

6. Jelikož boxploty v předchozím bodě nebyly velmi „čitelné“, můžeme se podívat na graf průměrů s intervaly spolehlivosti:

   ```
   library(gplots)
   plotmeans(conductivity~Fe_level)
   plotmeans(conductivity~NO3_level)
   ```

Jak závisí vodivost na obsahu železa? A jak na obsahu dusičnanů?

7. Otestujeme, jak je to se závislostí vodivosti na obsahu železa a na obsahu dusičnanů. Jelikož uvažujeme dva faktory, použijeme analýzu rozptylu dvojného třídění (nejprve bez interakcí):

   ```
   model=aov(conductivity~Fe_level+NO3_level)
   ```

   Tento model předpokládá, že vlivy obou faktorů jsou aditivní, tj. úroveň jednoho faktoru neovlivňuje efekt druhého. Test významnosti je opět založen na rozkladu celkového součtu čtverců. V případě dvou a více faktorů se však na „významnost“ můžeme dívat různě:

   ```
   anova(model)

   # pro porovnani vysledek, jsou-li faktory zadany v opacnem poradi
   model2=aov(conductivity~NO3_level+Fe_level)
   anova(model2)
   ```

   Funkce anova dává tzv. sekvenční rozklad čtverců. Naproti tomu použitím

   ```
   library(car)
   Anova(model,type="III")
   ```

   dostaneme tzv. marginální rozklad čtverců. V případě, kdy jsou do modelu zahrnuty interakce, dostaneme ještě trochu jiný výsledek, použitím

   ```
   Anova(model)
   ```

8. I v případě dvojného třídění bez interakcí lze provádět mnohonásobná porovnání, např.

   ```
   TukeyHSD(model,"Fe_level")
   ```

   (zde máme jen dvě úrovně faktoru, takže toto nebylo nutné.)

9. Nakonec ještě musíme zjistit, jak je to s předpoklady testu: ověřit normalitu a shodu rozptylů. (Opakování z minula).

10. Jelikož máme podezření, že by koncentrace železa mohla ovlivňovat vliv dusičnanů na vodivost, podíváme se na příslušné průměry ve skupinách:

    ```
    tapply(conductivity,list(Fe_level,Cl2_level),mean)
    ```

    Graficky si vše lze znázornit pomocé grafu interakcí:

    ```
    interaction.plot(Fe_level,NO3_level,conductivity,col=2:3)

    interaction.plot(NO3_level,Fe_level,conductivity,col=2:3)
    ```

11. Na základě předchozích grafů je zřejmé, že v tomto příkladě je nutné uvažovat model s interakcemi:

```
model=aov(conductivity~Fe_level*NO3_level)
```

Opět otestujeme významnost jednotlivých členů. Jak již bylo uvedeno výše, lze uvažovat tři různé pohledy:

```
anova(model)
```

```
Anova(model,type="III")
```

```
Anova(model)
```

Vyjdou-li interakce významné, nedoporučuje se provádět mnohonásobná porovnání na hlavní efekty. Lze provést mnohonásobné porovnání na skupiny dle obou faktorů:

```
TukeyHSD(model)
```

v dolní části výstupu (viz srovnání na obrázku interakcí).

12. Nakonec se opět musíme podívat na to, zda jsou splněny předpoklady testu a podle toho výsledek řádně interpretovat.

Vícenásobné modely ANOVA

13. Do výše uvažovaného modelu přidáme ještě vliv chlóru na vodivost. Opět se můžeme podívat na průměry v jednotlivých skupinách

```
tapply(conductivity,list(Fe_level,NO3_level,Cl2_level),mean)
tapply(conductivity,list(Cl2_level,NO3_level,Fe_level),mean)
tapply(conductivity,list(Cl2_level,Fe_level,NO3_level),mean)
```

nebo graficky

```
interaction.plot(NO3_level,Cl2_level,conductivity,col=1:2)
interaction.plot(Fe_level,Cl2_level,conductivity,col=1:2)
```

14. Budeme tak uvažovat model tzv. trojného třídění

```
model=aov(conductivity~Fe_level*NO3_level*Cl2_level)
```

Testy významnosti jednotlivých probíhá stejně jako u dvojného třídění. Lze uvažovat tři různé pohledy:

```
anova(model)
```

```
Anova(model,type="III")
```

```
Anova(model)
```

Samostatná práce

1. Načtěte si do R data Platy.txt. Datový soubor obsahuje informace o platech 242 osob ČR. K dispozici máme i údaj o nejvyšším dosaženém vzdělání a pohlaví respondenta.

Pohlavi	pohlaví zaměstnance (1 - žena, 0 - muž)
Plat	měsíční hrubý plat (v Kč)
Vzdelani	stupeň vzdělání (1 - ZŠ, 2 - SŠ, 3 - VŠ)

2. Prohlédněte si základní popisné statistiky pro všechny ze zahrnutých proměnných (funkce summary).

3. Pro kategoriální věličiny nám příkaz summary dává „nesmyslné“ výstupy. Musíme proto R-ku říci, že hodnoty těchto veličin nemá chápat jako čísla, ale jen jako označení úrovní. Např. pro pohlaví:

   ```
   pohlavi=factor(pohlavi,labels=c("Muz","Zena"))
   ```

   Stejně pro ostatní kategoriální proměnné v datech.

4. Pomocí vhodných obrázků a popisných statistik si udělejte představu o tom, zda
   - závisí plat na vzdělání,
   - závisí plat na pohlaví,
   - jaké je procentuální zastoupení jednotlivých kategorií vzdělání pro muže a pro ženy.
     `barplot(table(vzdelani,pohlavi),beside=T,legend=T)`

5. Otestujte, zda mají muži v průměru vyšší platy než ženy. Jestliže prokážete, že ano, tak odhadněte, o kolik procent je plat mužů vyšší.

6. Otestujte, zda plat závisí na vzdělání. Které kategorie se od sebe statisticky významně liší?

7. Uvažujte společný model, ve kterém má vzdělání a pohlaví vliv na měsíční plat. Na základě vhodných průměrů a obrázků se rozhodněte, zda je vhodné uvažovat model s interakcemi nebo bez nich.

8. Proveďte analýzu rozptylu.

9. Jaké jsou Vaše závěry o platových podmínkách v ČR?