Analýza variance (ANOVA), jednocestná, faktor s náhodným efektem; hierarchická ANOVA

1. Jednocestná ANOVA, faktor s náhodným efektem

Zajímá nás, zda se liší produkce semen smrku v rámci dané populace. K zodpovězení této otázky je třeba náhodně vybrat určitý počet stromů a u každého z nich stanovit počet semen v několika náhodně vybraných šiškách. Porovnáváme tedy určitý počet geneticky odlišných jedinců (faktorem je genetický jedinec), ale to, který jedinec bude nakonec vybrán pro srovnání, je dílem náhody. Produkce semen jednotlivými smrky populace bude mimo jiné závicet na genetických dispozicích, které reálně není možné zjistit (ani ovlivnit). Navíc, otázka zní, jestli je signifikantní rozdíl v produkci semen v dané populaci a není podstatné, zda a jak se mezi sebou liší některé konkrétní stromy! Takovýto typ faktoru odpovídá faktoru s náhodným efektem (oproti tomu u faktoru s pevným efektem nás zajímají právě rozdíly mezi jednotlivými hladinami).

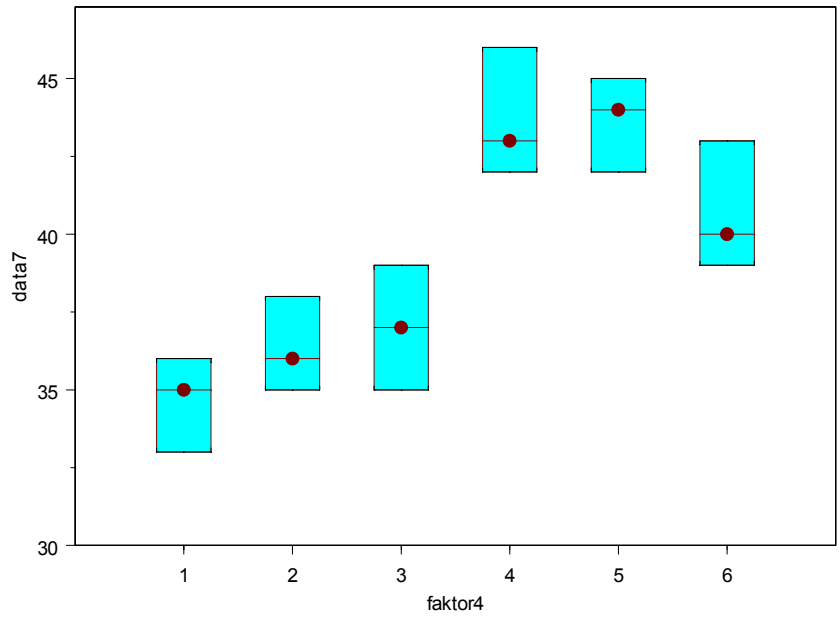
Máme k dispozici údaje o počtu semen v šiškách smrku ze šesti stromů (genetických jedinců).

Počet semen; Data7	33	35	36	36	38	35	39	35	37	42	43	46	44	45	42	39	40	43
Jedinec; Faktor4	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6

Pokud data importujeme či zadáváme přímo do tabulky je třeba nastavit sloupec s faktorem na datový typ *faktor*. Použijeme pravé tlačítko myši či menu Data>Change Data Type...!

Analýzu budeme tedy počítat jako ***jednocestnou ANOVu***, kdy faktor je s ***náhodným efektem*** Jak jsme již zmínili u jednocestné ANOVy s faktory s pevnými efekty, výpočet je u obou typů shodný. Rozdíl je tedy pouze v interpretaci výsledné analýzy a předem neplánování použití mnohonásobných srovnání.
Pro výpočet v S+ tedy můžeme počítat přes menu Statistics>Compare Samples>k samples>One-way ANOVA... či přes Statistics>ANOVA>Random effects... Avšak pokud budeme ANOVU počítat dle druhého způsobu, budeme si muset sami dopočítat hodnotu testovací F statistiky a hodnotu dosažené pravděpodobnosti.

Neměli bychom zapomenout na předběžný pohled na zdrojová data pomocí krabicových diagramů.



Pokud tedy počítáme přes nabídku One-way ANOVA dostáváme tuto tabulku:

```
      ***  One-Way ANOVA for data in data7 by faktor4 ***

Call:
   aov(formula = structure(.Data = data7 ~ faktor4, class =
      "formula"), data = SDF2)

Terms:
                 faktor4 Residuals
 Sum of Squares 226.6667   39.3333
Deg. of Freedom        5        12

Residual standard error: 1.810463
Estimated effects are balanced

           Df Sum of Sq  Mean Sq  F Value         Pr(F)
  faktor4   5  226.6667 45.33333 13.83051 0.0001251073
Residuals 12   39.3333  3.27778
```

Pokud použijeme nabídku ANOVA>Random effects tabulka vypadá takto:

```
*** Analysis of Variance Model ***

Short Output:
Call:
   raov(formula = data7 ~ faktor4, data = SDF2, na.action = na.exclude)

Terms:
                 faktor4 Residuals
 Sum of Squares 226.6667   39.3333
Deg. of Freedom        5        12

Residual standard error: 1.810463
Estimated effects are balanced

           Df Sum of Sq  Mean Sq Est. Var.
  faktor4   5  226.6667 45.33333  14.01852
Residuals 12   39.3333  3.27778   3.27778
```

Teď ještě dopočítáme hodnotu F statistiky a pravděpodobnost. Nejlépe je spočítáme v "Commands" okně

```
> 45.33333/3.27778
[1] 13.8305
```

výsledná F statistika

```
> 1-pf(13.83, 5, 12)
[1] 0.0001251299
```

použitá funkce pf vypočte kumulativní pravděpodobnost F rozdělení a má tři parametry (F hodnota, stupně volnosti čitatele, stupně volnosti jmenovatele).

Výsledek je tedy v obou případech shodný a vysoce průkazný. Lze tedy zamítnout nulovou hypotézu, že produkce semen se v populaci smrku významně neliší. Provedením mnohonásobných srovnání bychom mohli ověřit, které ze šesti smrků se vzájemně liší (pozor, je třeba si uvědomit, že a priori plánovat párové testy v tomto konkrétním případě nemá smysl).

Tabulka Modelů I-III

	I.	II.	III.
	A, B s pevnými efekty	A, B s náhodnými efekty	A s pevnými, B s náhodnými efekty
Faktor A	MS_A/MS_e	MS_A/MS_{AB}	MS_A/MS_{AB}
Faktor B	MS_B/MS_e	MS_B/MS_{AB}	MS_B/MS_e
Interakce A×B	MS_{AB}/MS_e	MS_{AB}/MS_e	MS_{AB}/MS_e

MS_e průměrný čtverec reziduí

ANOVA se třemi faktory, všechny pevné efekty: faktory a interakce se testují proti MS_e.

ANOVA se třemi faktory. A, B - s pevnými efekty, C - náhodné efekty

A MS_A/MS_{AC}
B MS_B/MS_{BC}
C MS_C/MS_e
AB MS_{AB}/MS_{ABC}
AC MS_{AC}/MS_e
BC MS_{BC}/MS_e
ABC MS_{ABC}/MS_e

2. Hierarchická ANOVA

Předchozí příklad můžeme rozvést. Předpokládejme, že jsme (při zachování náhodnosti výběru) získali první tři stromy z předchozího příkladu ze svahu orientovaného na sever a druhou polovinu z jižní expozice. Otázka je, zda se liší produkce semen smrku na svazích jižní a severní orientace. Protože se však může lišit produkce semen i u stromů na jednotlivých svazích (mimo jiné opět v důsledku genetické variability jednotlivých stromů), což by mohlo vést k zastření vlivu rozdílné orientace svahů, je třeba tuto část variability oddělit.

Počet semen; Data7	33	35	36	36	38	35	39	35	37	42	43	46	44	45	42	39	40	43
Expozice; Faktor5	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
Jedinec; Faktor4	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3

Poznámka ke kódování jedinců. Přestože se jedná o odlišné jedince a jedinec č. 1 z jižní orientace nemá žádný vztah k jedinci č. 1 ze severní expozice je potřeba pro S+ kódovat v rámci každé hierarchické skupiny shodně.

Oproti předchozímu příkladu máme ANOVu se dvěma faktory. První faktor je orientace svahu (jasně definovaná, tedy pevný efekt). Faktor genetického jedince je s náhodným efektem a je zároveň **podřazený** (vhnízděný - nested) faktoru orientace. Cílem je tedy oddělit efekt náhodné variability jednotlivých stromů od variability dané vlivem nadřazeného faktoru. Tento typ analýzy řešíme ***hierarchickou*** (nested) ***ANOVou***, u které je podmínka, že podřazený faktor je s náhodným efektem.

V S+ je nejvhodnější spočítat tuto analýzu přes menu Statistics>ANOVA>Random effects... Tím získáme Průměrné čtverce pro jednotlivé faktory a pak si již snadno spočteme hodnoty F statistiky a odpovídající pravděpodobnosti. Podřazené faktory v S+ zapisujeme takto: *Nadřazený/Podřazený* (takto můžeme mít i hlubší strukturu např. Kontinent/Stát/Okres/Město). Druhá možnost zápisu je *Podřazený %in% Nadřazený*. Obecný vzorec pro výpočet tedy bude: `data7 ~ faktor5 + faktor5/faktor4`

Z výsledné tabulky získáme hodnoty průměrných čtverců

```
                        Df     Sum of Sq   Mean Sq     Est. Var.
                faktor5  1     200.0000    200.0000    21.48148
faktor4 %in% faktor5     4     26.6667     6.6667      1.12963
              Residuals  12    39.3333     3.2778      3.27778
```

Hodnoty F statistiky pro *faktor5* získáme podílem $MS_{faktor5}/MS_{faktor4\%in\%faktor5}$, tedy 200/6.6667 což je 30.0 a tedy dosažená hladina pravděpodobnosti je při daných stupních volnosti 0.005408.
Pro podřazený faktor je testovací kritérium podílem odpovídajícího MS ku průměrnému čtverci reziduí (6.6667/3.2778=2.03), a následně hladina pravděpodobnosti 0.153414.
Průměry pro faktor expozice jsou:

```
faktor5
      1      2
 36.000 42.667
```

Výsledkem hierarchické ANOVy je závěr, že nejsou průkazné rozdíly v rámci podřazeného faktoru (jednotlivé stromy se mezi sebou v produkci semen průkazně neliší). Průkazný je však rozdíl nadřazeného faktoru - existuje tedy signifikantní rozdíl v produkci semen mezi svahy rozdílné expozice.