Дисперсионный анализ, часть 4

Математические методы в зоологии - на R, осень 2013

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Дисперсионный анализ

- Анализ моделей со вложенными факторами (иерархический дисперсионный анализ)
- Компоненты дисперсии для случайных факторов

Вы сможете

- Отличать случаи, когда нужен дисперсионный анализ со вложенными факторами
- Проводить иерархический дисперсионный анализ
- Рассчитывать компоненты дисперсии для случайных факторов

Исходные данные для иерархического дисперсионного анализа

выглядят примерно так

ОБЪЕКТ	ЧАСТЬ ОБЪЕКТА
1	A
1	B
1	C
2	A
2	B
2	C
3	A
3	B
3	C
И т.д.	

В данном случае

Верхний фактор в иерархии - Объект

Вложенный фактор - Часть объекта

Одноименные уровни вложенного фактора несопоставимы между разными объектами!

т.е. А для 1-го объекта не то же самое, что А для второго. Иными словами, ответ на действие вложенного фактора будет разным для разных уровней вышестоящего фактора.

Подберите правильный дизайн дисперсионного анализа

- Какие из этих данных подходят для иерархического дисперсионного анализа?
- Какие из факторов фиксированные, а какие случайные?

ОБЪЕКТ	ЧАСТЬ ОБЪЕКТА
1	A
1	B
1	C
2	A
2	B
2	C
3	A
3	B
3	C
И т.д.	

- Средний размер кукушиных яиц в гнездах одних и тех же 3 видов птиц в 4 лесах
- Число личинок, осевших на 3 вида субстратов в 7 аквариумах (все три субстрата в каждом)
- Уровень экспрессии генов у дрозофил в зависимости от температуры содержания (4 режима содержания по 3 популяции в каждом)

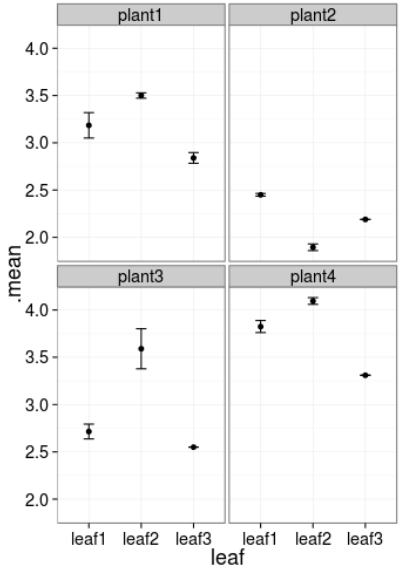
Пример: Кальций в листьях турнепса

Содержание кальция в листьях турнепса

- 4 растения
  - 3 листа с каждого растения (по две пробы с каждого листа)

```
library(XLConnect)
turn <- readWorksheetFromFile(file="./data/turnips.xlsx",
                              sheet = 1)
head(turn)
```

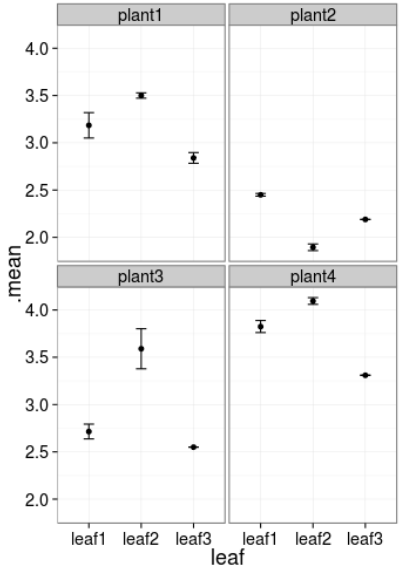
```
##    plant  leaf   ca
## 1 plant1 leaf1 3.28
## 2 plant1 leaf1 3.09
## 3 plant1 leaf2 3.52
## 4 plant1 leaf2 3.48
## 5 plant1 leaf3 2.88
## 6 plant1 leaf3 2.80
```



Снедекор, 1961, с.252

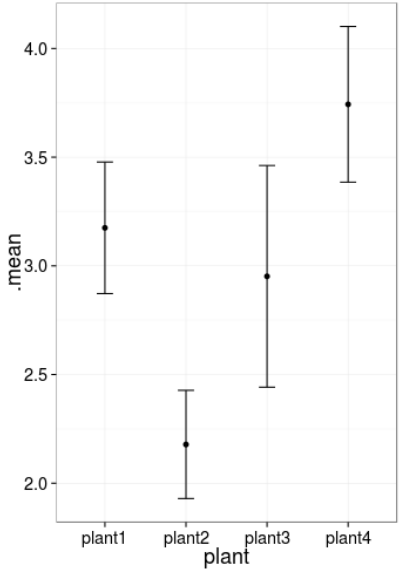
Особенности иерархического дисперсионного анализа

- Минимум два фактора A и B
- Несколько (случайным образом выбранных) градаций фактора B (листья) внутри каждого из уровней фактора A (растения)
- Часто больше одного уровня в иерархии
- Оценка взаимодействия главного фактора и вложенного невозможна



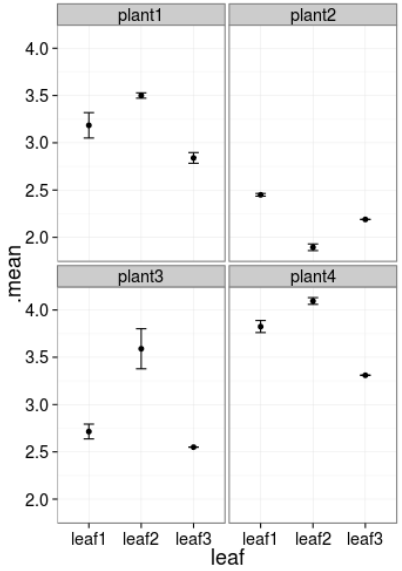
Главный эффект

- Эффект фактора A - изменчивость между средними по фактору A (различия содержания кальция между растениями)



Эффект вложенного фактора

- Эффект фактора B на каждом из уровней фактора A - различия средних по фактору B на каждом из уровней фактора A
  (различия содержания кальция между листьями на одном растении)



Структура изменчивости

Общая = по фактору A + по вложенному фактору + случайная

$$SS_t = SS_A + SS_{B|A} + SS_e$$

- SS_A – различия между средними по фактору A и общим средним
- $SS_{B|A}$ – различия между средними по фактору B и средним на данном уровне A
- SS_e – различия между повторными измерениями в ячейках и общим средним

Как считать F-критерий в иерархическом дисперсионном анализе

ИСТОЧНИК ИЗМЕНЧИВОСТИ	SS	DF	MS	F
A	SS_A	$a-1$	MS_A	$MS_A/MS_{B\|A}$
B|A	$SS_{B\|A}$	$a(b-1)$	$MS_{B\|A}$	$MS_{B\|A}/MS_e$
Случайная	SS_e	$ab(n-1)$	MS_e	
Общая	SS_t			

- Дисперсия каждого фактора оценивается по отношению к дисперсии нижележащего в иерархии
- Вложенный фактор чаще всего случайный, как здесь - смешанная модель

Почему F считается именно так, становится понятно, если посмотреть, что именно оценивают MS

ИСТОЧНИК ИЗМЕНЧИВОСТИ	SS	DF	MS	F	ОЖИДАЕМЫЙ СРЕДНИЙ КВАДРАТ
A	SS_A	$a-1$	MS_A	$MS_A/MS_{B\|A}$	$\sigma^2+n\sigma^2_{B\|A}+nb\sigma^2_A$
$B\|A$	$SS_{B\|A}$	$a(b-1)$	$MS_{B\|A}$	$MS_{B\|A}/MS_e$	$\sigma^2+n\sigma^2_{B\|A}$
Случайная	SS_e	$ab(n-1)$	MS_e		σ^2
Общая	SS_t				

У нас сбалансированный дисперсионный комплекс?

```
table(turn$plant, turn$leaf, useNA = "no")
```

```
##
##         leaf1 leaf2 leaf3
##   plant1     2     2     2
##   plant2     2     2     2
##   plant3     2     2     2
##   plant4     2     2     2
```

Дисперсионный анализ со вложенными факторами для сбалансированных данных

Сначала задаем типы факторов: фиксированные или случайные

```
# install.packages("GAD")
library(GAD) # Дисперсионный анализ по Underwood, 1997
# задаем фиксированные и случайные факторы
turn$plant <- as.fixed(turn$plant)
turn$leaf <- as.random(turn$leaf)
```

Подбираем подель дисперсионного анализа с помощью lm()

Вложенный фактор обозначается так:

`вложенный %in% главный`

```
# модель дисперсионного анализа
model <- lm(ca ~ plant + leaf %in% plant, data = turn)
```

Таблица результатов иерархического дисперсионного анализа

```
model_gad <- gad(model)
options(digits = 3, scipen = 6) # для форматирования чисел в таблице
model_gad
```

```
## Analysis of Variance Table
##
## Response: ca
##            Df Sum Sq Mean Sq F value      Pr(>F)
## plant       3   7.56   2.520    7.67      0.0097 **
## plant:leaf  8   2.63   0.329   49.41 0.000000051 ***
## Residual   12   0.08   0.007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

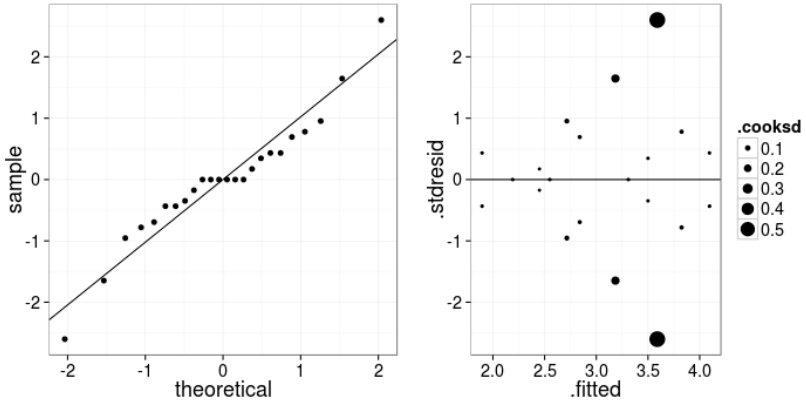
Данные для проверки условий применимости

```
model_diag <- fortify(model) # fortify() из ggplot2
head(model_diag)
```

```
##     ca  plant  leaf .hat .sigma .cooksd .fitted .resid .stdresid
## 1 3.28 plant1 leaf1  0.5 0.0750  0.2260    3.18  0.095     1.647
## 2 3.09 plant1 leaf1  0.5 0.0750  0.2260    3.18 -0.095    -1.647
## 3 3.52 plant1 leaf2  0.5 0.0848  0.0100    3.50  0.020     0.347
## 4 3.48 plant1 leaf2  0.5 0.0848  0.0100    3.50 -0.020    -0.347
## 5 2.88 plant1 leaf3  0.5 0.0835  0.0401    2.84  0.040     0.693
## 6 2.80 plant1 leaf3  0.5 0.0835  0.0401    2.84 -0.040    -0.693
```

Проверим условия применимости

```
# Квантильный график - нормальное распределение остатков
p1 <- ggplot(model_diag) + geom_point(stat = "qq", aes(sample = .stdresid)) +
  geom_abline(yintercept = 0, slope = sd(model_diag$.stdresid))
# График стандартизованных остатков - гомогенность дисперсий остатков
# Расстояние Кука - наличие "выбросов"
p2 <- ggplot(model_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd)) + geom_hline(yintercept = 0)
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)
```



Компоненты дисперсии

- рассчитываются для случайных факторов
- дисперсия между средними во всех возможных группах
- аналоги силы влияния фиксированных факторов

$$s_A^2 = \frac{MS_A - MS_B}{nb}$$

$$s_{B|A}^2 = \frac{MS_B - MS_e}{n}$$

$$s^2 = MS_e$$

Если найти общую, можно будет выразить компоненты дисперсии в процентах

$$s_t^2 = s_A^2 + s_{B|A}^2 + s^2$$

Посчитаем компоненты дисперсии

$$s_A^2 = \frac{MS_A - MS_B}{nb}$$

$$s_{B|A}^2 = \frac{MS_B - MS_e}{n}$$

$$s^2 = MS_e$$

$$s_t^2 = s_A^2 + s_{B|A}^2 + s^2$$

```
table(turn$plant, turn$leaf, useNA = "no")
```

```
##
##          leaf1 leaf2 leaf3
##   plant1     2     2     2
##   plant2     2     2     2
##   plant3     2     2     2
##   plant4     2     2     2
```

```
# Средние квадраты
MSa <-model_gad$'Mean Sq'[1]
MSba <- model_gad$'Mean Sq'[2]
MSe <- model_gad$'Mean Sq'[3]
b <- 3 # число групп по фактору B (листьев на растени
n <- 2 # объем группы (измерений на листе)
VC <- data.frame (VCa = (MSa - MSba)/(n*b),
                  VCba = (MSba - MSe)/n,
                  VCe = MSe)
VC # компоненты дисперсии
```

```
##      VCa  VCba     VCe
## 1 0.365 0.161 0.00665
```

```
VC/sum(VC)*100 # в процентах
```

```
##    VCa VCba  VCe
## 1 68.5 30.2 1.25
```

Осторожно: интерпретация компонент дисперсии для случайных и фиксированных факторов разная!

- Для случайных факторов - дисперсия между средними во всех возможных группах
- Для фиксированных факторов - дисперсия между средними в группах

```
VC[1]/sum(VC)*100 # в процентах
```

```
##    VCa
## 1 68.5
```

Для сравнения доля объясненной изменчивости для фикс. фактора (эта-квадрат и частный эта-квадрат)

```
(etasq_a <- model_gad$'Sum Sq'[1]/sum(model_gad$'Sum Sq'))
```

```
## [1] 0.736
```

```
(p_etasq_a <- model_gad$'Sum Sq'[1]/(model_gad$'Sum Sq'[1] + model_gad$'Sum Sq'[3]))
```

```
## [1] 0.99
```

Пример: Морские ежи и водоросли

Влияет ли плотность морских ежей на обилие нитчаток в сублиторали? (Andrew, Underwood, 1993)

- Обилие ежей - 4 уровня (нет, 33%, 66%, 100%)
- Площадка - 4 штуки (площадь 3-4 $м^2$; по 5 проб на площадке)

```
andr <- readWorksheetFromFile(file = "./data/andrew.xlsx", sheet = 1)
head(andr)
```

```
##   treat patch patchrec quad algae
## 1   con     1       p1    1     0
## 2   con     1       p1    2     0
## 3   con     1       p1    3     0
## 4   con     1       p1    4     6
## 5   con     1       p1    5     2
## 6   con     2       p2    1     0
```

Данные из Quinn, Keough, 2002, p.209

Подготавливаем данные

```
str(andr)
```

```
## 'data.frame':    80 obs. of  5 variables:
##  $ treat   : chr  "con" "con" "con" "con" ...
##  $ patch   : num  1 1 1 1 1 2 2 2 2 2 ...
##  $ patchrec: chr  "p1" "p1" "p1" "p1" ...
##  $ quad    : num  1 2 3 4 5 1 2 3 4 5 ...
##  $ algae   : num  0 0 0 6 2 0 0 0 0 0 ...
```

```
andr$patchrec <- factor(andr$patchrec)
andr$treat <- factor(andr$treat)
str(andr)
```

```
## 'data.frame':    80 obs. of  5 variables:
##  $ treat   : Factor w/ 4 levels "con","rem","t0.33",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ patch   : num  1 1 1 1 1 2 2 2 2 2 ...
##  $ patchrec: Factor w/ 4 levels "p1","p2","p3",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ quad    : num  1 2 3 4 5 1 2 3 4 5 ...
##  $ algae   : num  0 0 0 6 2 0 0 0 0 0 ...
```

Сбалансированный ли у нас дисперсионный комплекс?

Проведите дисперсионный анализ

Проведите диагностику дисперсионного анализа

Проверьте условия применимости дисперсионного анализа

- нормальное распределение остатков
- гомогенность дисперсий остатков

Проверьте наличие "выбросов"

Посчитайте компоненты дисперсии в процентах

$$s_A^2 = \frac{MS_A - MS_B}{nb}$$

$$s_{B|A}^2 = \frac{MS_B - MS_e}{n}$$

$$s^2 = MS_e$$

Постройте график средних значений

А если объемы выборок неравны?

- Лучше использовать оценки максимального правдоподобия (пакеты `nlme`, `lme4`)
- Для тестирования гипотез - G-тест (likelihood-ratio test - сравнение полной и уменьшенной моделей)
- Использование традиционного подхода невозможно - нельзя построить F-распределение для нулевой гипотезы

Take home messages

- Иерархический дисперсионный анализ нужен, когда одноименные уровни вложенного фактора несопоставимы между разными объектами
- Значимость факторов проверяется по отношению к нижележащему в иерархии
- Компоненты дисперсии рассчитываются для случайных факторов (**не только** в иерархическом дисперсионном анализе)
  - дисперсия между средними во всех возможных группах
  - аналоги силы влияния фиксированных факторов

Дополнительные ресурсы

- Quinn, Keough, 2002
- Logan, 2010
- Sokal, Rohlf, 1995
- Zar, 2010