Дисперсионный анализ, часть 2

Математические методы в зоологии - на R, осень 2013

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Многофакторный дисперсионный анализ

- Линейная модель многофактроного дисперсионного анализа
- Фиксированные и случайные факторы (I и II модель)
- Дисперсионный анализ сбалансированных данных с фиксированными факторами
- Анализ несбалансированных данных. Типы сумм квадратов (I, II, III).

Вы сможете

- Проводить многофакторный дисперсионный анализ с учетом взаимодействия факторов
- Отличать фиксированные и случайные факторы и выбирать подходящую модель дисперсионного анализа
- Выяснять, сбалансированы ли данные и выбирать подходящий тип сумм квадратов
- Интерпретировать результаты дисперсионного анализа с учетом взаимодействия факторов

Линейные модели для факторных дисперсионных анализов

- Два фактора A и B, двухфакторное взаимодействие

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- Три фактора A, B и C, двухфакторные взаимодействия, трехфакторное взаимодействия

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

Взаимодействие факторов

Эффект фактора B разный в зависимости от уровней фактора A и наоборот.

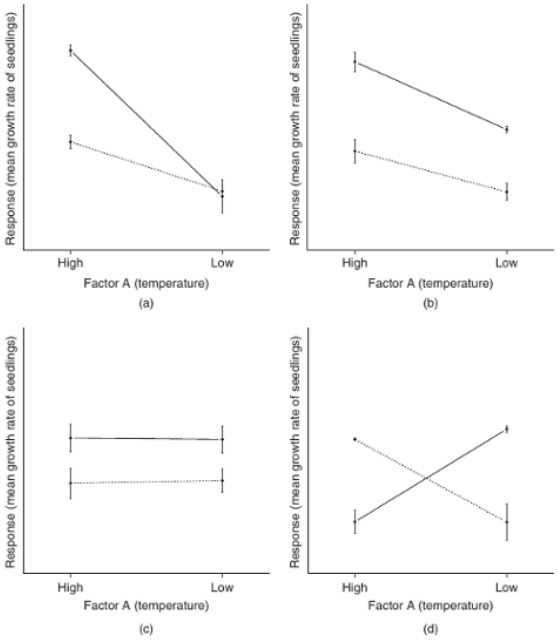


Рисунок из Logan, 2010, fig.12.2

На каких рисунках есть взаимодействие факторов?

Эффект фактора B разный в зависимости от уровней фактора A и наоборот.

- b, c - нет взаимодействия
- a, d - есть взаимодействие

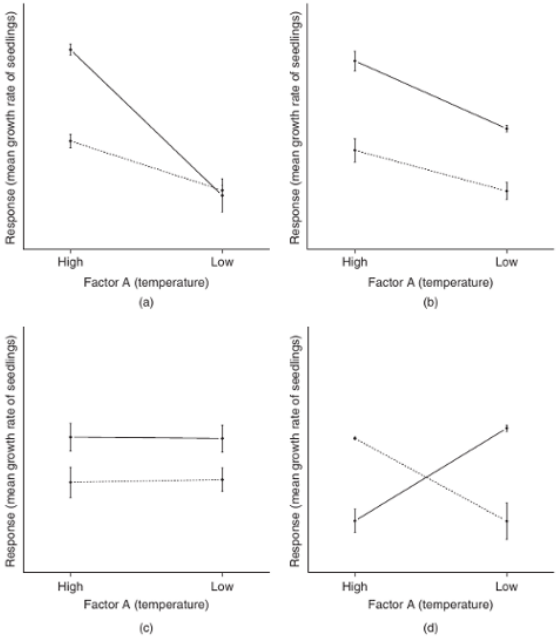


Рисунок из Logan, 2010, fig.12.2

Взаимодействие факторов может маскировать главные эффекты

- Если есть значимое взаимодействие
  - пост хок тесты только по нему.
  - главные эффекты обсуждать не имеет смысла (они могут быть замаскированы взаимодействием)

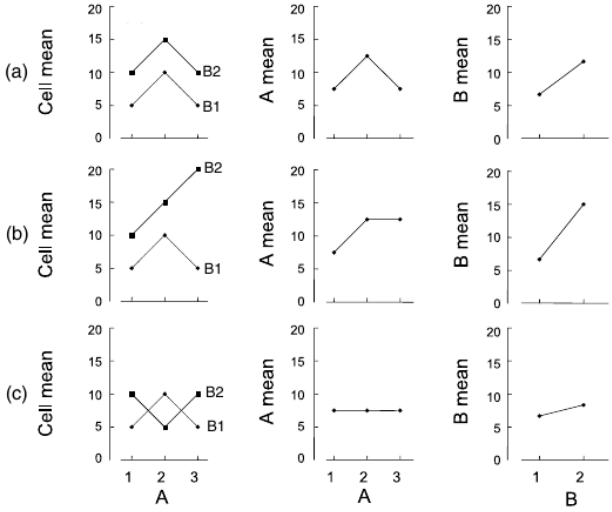


Рисунок из Quinn, Keough, 2002, fig.9.3

Фиксированные и случайные факторы

Две модели дисперсионного анализа

Вспомните, что такое фиксированные и случайные факторы

Какого типа эти факторы?

- Несколько произвольно выбранных градаций плотности моллюсков в полевом эксперименте, где плотностью манипулировали.
- Фактор размер червяка (маленький, средний, большой) в выборке червей.
- Деление губы Чупа на зоны с разной степенью распреснения.

- Может ли один и тот же фактор рассматриваться как случайный или фиксированный?
- Приведите примеры, как тип фактора будет зависеть от проверяемых гипотез

Гипотезы в разных моделях многофакторного дисперсионного анализа

ТИП ФАКТОРА	ФИКСИРОВАННЫЕ ФАКТОРЫ	СЛУЧАЙНЫЕ ФАКТОРЫ
Модель дисп.анализа	I-модель	II-модель
Гипотезы	средние равны	нет увеличения дисперсии связанного с фактором
Для A	$H_{0(A)}: \mu_1 = \mu_2 = \cdots = \mu_i = \mu$	$H_{0(A)}: \sigma^2_\alpha = 0$
Для B	$H_{0(B)}: \mu_1 = \mu_2 = \cdots = \mu_i = \mu$	$H_{0(B)}: \sigma^2_\beta = 0$
Для AB	$H_{0(AB)}: \mu_{ij} = \mu_i + \mu_j - \mu$	$H_{0(AB)}: \sigma^2_{\alpha\beta} = 0$

Рассчет F-критерия для I и II моделей дисперсионного анализа

ФАКТОРЫ	А И В ФИКСИРОВАННЫЕ	А И В СЛУЧАЙНЫЕ	А ФИКСИРОВАННЫЙ, В СЛУЧАЙНЫЙ
A	$\frac{F = MS_a}{MS_e}$	$\frac{F = MS_a}{MS_{ab}}$	$\frac{F = MS_a}{MS_e}$
B	$\frac{F = MS_b}{MS_e}$	$\frac{F = MS_b}{MS_{ab}}$	$\frac{F = MS_b}{MS_a b}$
AB	$\frac{F = MS_{ab}}{MS_e}$	$\frac{F = MS_{ab}}{MS_e}$	$\frac{F = MS_{ab}}{MS_e}$

Внимание: сегодня - только про фиксированные факторы

Дисперсионный анализ для фиксированных факторов

Пример: Возраст и память

Почему пожилые не так хорошо запоминают? Может быть не так тщательно перерабатывают информацию? (Eysenck, 1974)

Факторы:

- `Age` - Возраст:
  - `Younger` - 50 молодых
  - `Older` - 50 пожилых (55-65 лет)
- `Process` - тип активности:
  - `Counting` - посчитать число букв
  - `Rhyming` - придумать рифму к слову
  - `Adjective` - придумать прилагательное
  - `Imagery` - представить образ
  - `Intentional` - запомнить слово

Зависимая переменная - `Words` - сколько вспомнили слов

```
library(ggplot2)
theme_set(theme_bw(base_size = 18))
update_geom_defaults("point", list(shape = 19))
```

```
memory <- read.delim(file="./data/eysenck.csv")
head(memory, 10)
```

```
##        Age  Process Words
## 1  Younger Counting     8
## 2  Younger Counting     6
## 3  Younger Counting     4
## 4  Younger Counting     6
## 5  Younger Counting     7
## 6  Younger Counting     6
## 7  Younger Counting     5
## 8  Younger Counting     7
## 9  Younger Counting     9
## 10 Younger Counting     7
```

http://www.statsci.org/data/general/eysenck.html

Меняем порядок уровней для красоты

```
str(memory)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Age    : Factor w/ 2 levels "Older","Younger": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Process: Factor w/ 5 levels "Adjective","Counting",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Words  : num  8 6 4 6 7 6 5 7 9 7 ...
```

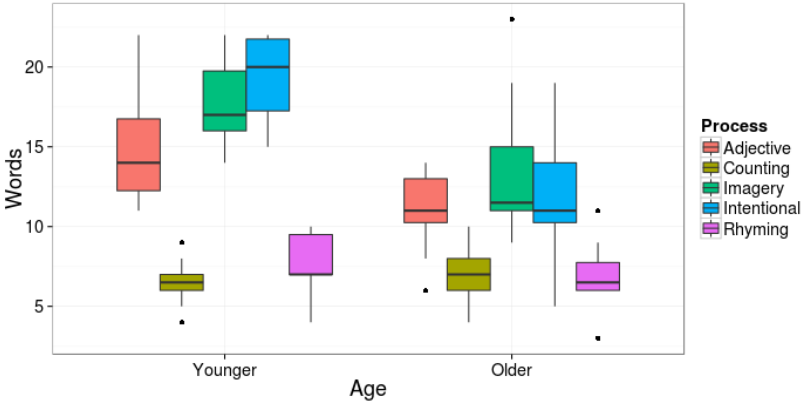
```
levels(memory$Age)
```

```
## [1] "Older"   "Younger"
```

```
# Хотим, чтобы молодые шли первыми - меняем порядок уровней
memory$Age <- relevel(memory$Age, ref="Younger")
```

Посмотрим на боксплот

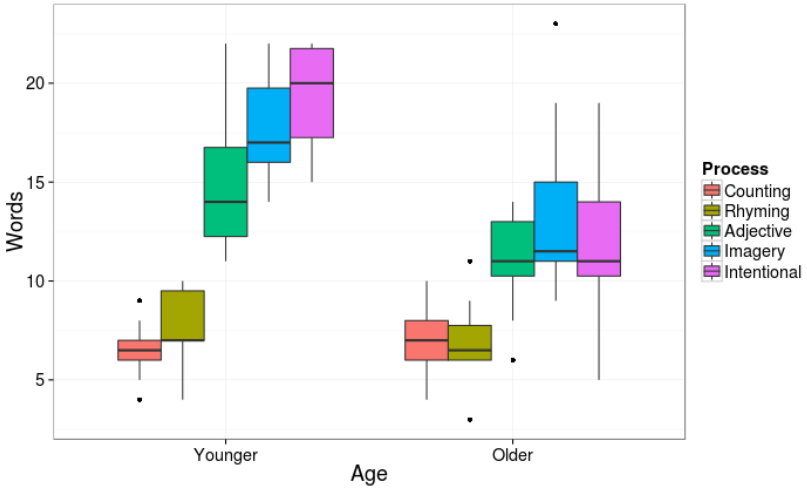
```
#   Этот график нам пригодится для представления результатов
ggplot(data = memory, aes(x = Age, y = Words)) + geom_boxplot(aes(fill = Process))
```




```
# некрасивый порядок уровней memory$Process
# переставляем в порядке следования средних значений memory$Words
memory$Process <- reorder(memory$Process, memory$Words, FUN=mean)
```

Боксплот с правильным порядком уровней

```
mem_p <- ggplot(data = memory, aes(x = Age, y = Words)) +
  geom_boxplot(aes(fill = Process))
mem_p
```



Описательная статистика по группам

- Какого типа здесь факторы?
- Сбалансированный ли дизайн?

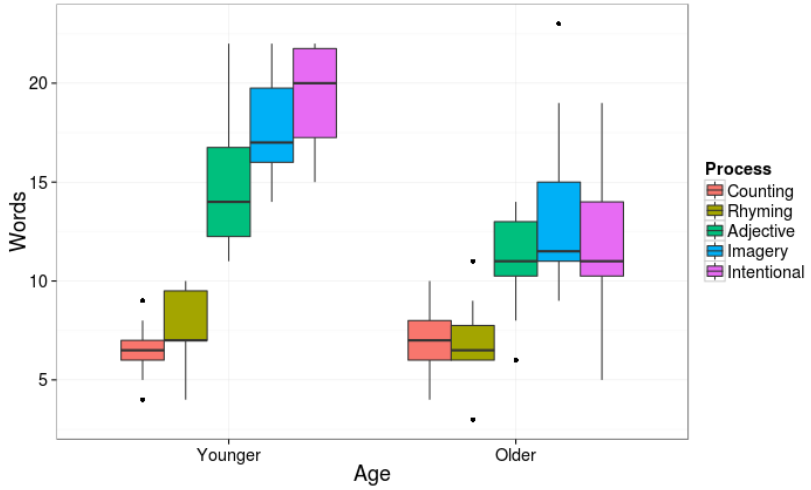
```
library(reshape)
# __Статистика по столбцам и по группам__ одновременно (n, средние, дисперсии, стандартные отклонения)
memory_summary <- ddply(memory, .variables = c("Age", "Process"),
                        summarise,
                        .n = sum(!is.na(Words)),
                        .mean = mean(Words),
                        .var = var(Words),
                        .sd = sd(Words))
memory_summary # краткое описание данных
```

```
##        Age     Process .n .mean  .var  .sd
## 1  Younger    Counting 10   6.5  2.06 1.43
## 2  Younger     Rhyming 10   7.6  3.82 1.96
## 3  Younger   Adjective 10  14.8 12.18 3.49
## 4  Younger     Imagery 10  17.6  6.71 2.59
## 5  Younger Intentional 10  19.3  7.12 2.67
## 6    Older    Counting 10   7.0  3.33 1.83
## 7    Older     Rhyming 10   6.9  4.54 2.13
## 8    Older   Adjective 10  11.0  6.22 2.49
## 9    Older     Imagery 10  13.4 20.27 4.50
## 10   Older Intentional 10  12.0 14.00 3.74
```

Проверяем условия применимости дисперсионного анализа

- Нормальное ли распределение?
- Есть ли гомогенность дисперсий?

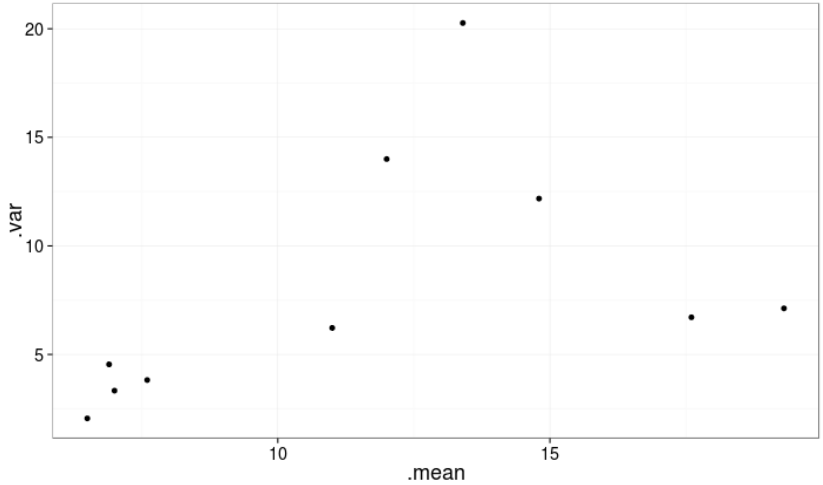
```
mem_p
```



Связь дисперсий и средних

- Есть ли гомогенность дисперсий?

```
# Данные взяли в кратком описании
ggplot(memory_summary, aes(x = .mean, y = .var)) + geom_point()
```



Задаем модель со взаимодействием

`Age:Process` - взаимодействие обозначается `:`

```
memory_aov <- aov(Words ~ Age + Process + Age:Process, data = memory)
```

- То же самое - `Age*Process` - вместо всех факторов

```
memory_aov <- aov(Words ~ Age*Process, data = memory)
```

Данные для графиков остатков

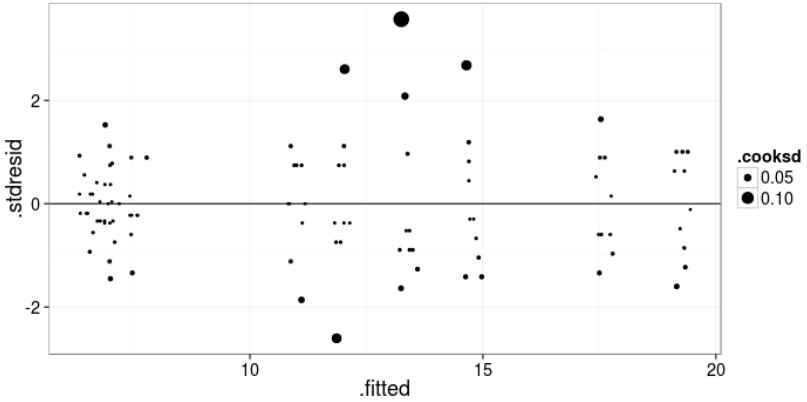
```
memory_diag <- fortify(memory_aov)
head(memory_diag, 3)
```

```
##   Words     Age  Process .hat .sigma  .cooksd .fitted .resid .stdresid
## 1     8 Younger Counting  0.1   2.84 0.003461     6.5    1.5     0.558
## 2     6 Younger Counting  0.1   2.85 0.000385     6.5   -0.5    -0.186
## 3     4 Younger Counting  0.1   2.84 0.009614     6.5   -2.5    -0.930
```

Графики остатков

- Есть ли гомогенность дисперсий?
- Не видно ли трендов в остатках?

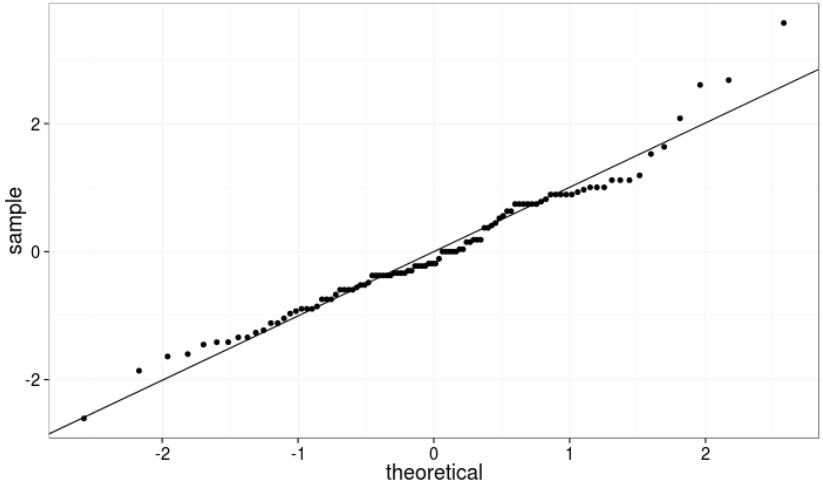
```
ggplot(memory_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd), position = position_jitter(width = .2)) +
  geom_hline(yintercept = 0)
```



Квантильный график

- Нормальное ли у остатков распределение?

```
ggplot(memory_diag) + geom_point(stat = "qq", aes(sample = .stdresid)) +
  geom_abline(yintercept = 0, slope = sd(memory_diag$.stdresid))
```



Результаты дисперсионного анализа

```
anova(memory_aov)
```

```
## Analysis of Variance Table
##
## Response: Words
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Age          1    240     240   29.94 0.0000004 ***
## Process      4   1515     379   47.19   < 2e-16 ***
## Age:Process  4    190      48    5.93   0.00028 ***
## Residuals   90    722       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Пост хок тест

Взаимодействие достоверно, можно другое не тестировать

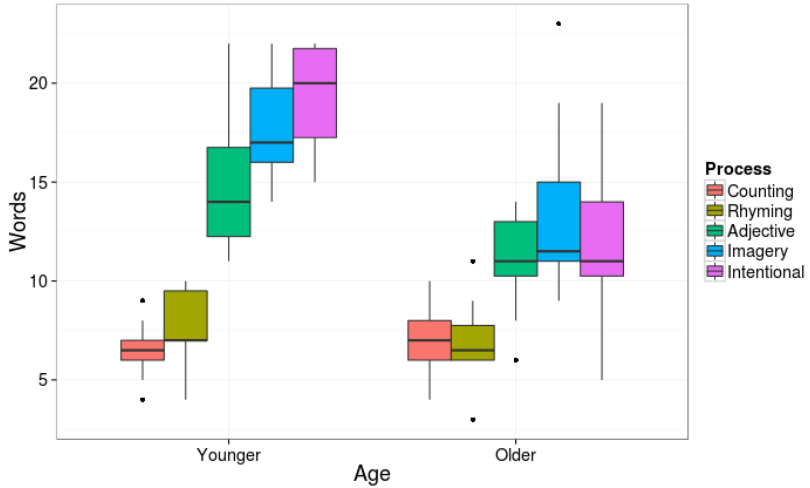
```
TukeyHSD(memory_aov, which=c("Age:Process"))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Words ~ Age * Process, data = memory)
##
## $`Age:Process`
##                                        diff      lwr     upr p adj
## Older:Counting-Younger:Counting         0.5  -3.6105  4.6105 1.000
## Younger:Rhyming-Younger:Counting        1.1  -3.0105  5.2105 0.997
## Older:Rhyming-Younger:Counting          0.4  -3.7105  4.5105 1.000
## Younger:Adjective-Younger:Counting      8.3   4.1895 12.4105 0.000
## Older:Adjective-Younger:Counting        4.5   0.3895  8.6105 0.021
## Younger:Imagery-Younger:Counting       11.1   6.9895 15.2105 0.000
## Older:Imagery-Younger:Counting          6.9   2.7895 11.0105 0.000
## Younger:Intentional-Younger:Counting   12.8   8.6895 16.9105 0.000
## Older:Intentional-Younger:Counting      5.5   1.3895  9.6105 0.001
## Younger:Rhyming-Older:Counting          0.6  -3.5105  4.7105 1.000
## Older:Rhyming-Older:Counting           -0.1  -4.2105  4.0105 1.000
## Younger:Adjective-Older:Counting        7.8   3.6895 11.9105 0.000
## Older:Adjective-Older:Counting          4.0  -0.1105  8.1105 0.063
## Younger:Imagery-Older:Counting         10.6   6.4895 14.7105 0.000
## Older:Imagery-Older:Counting            6.4   2.2895 10.5105 0.000
```

Графики для результатов

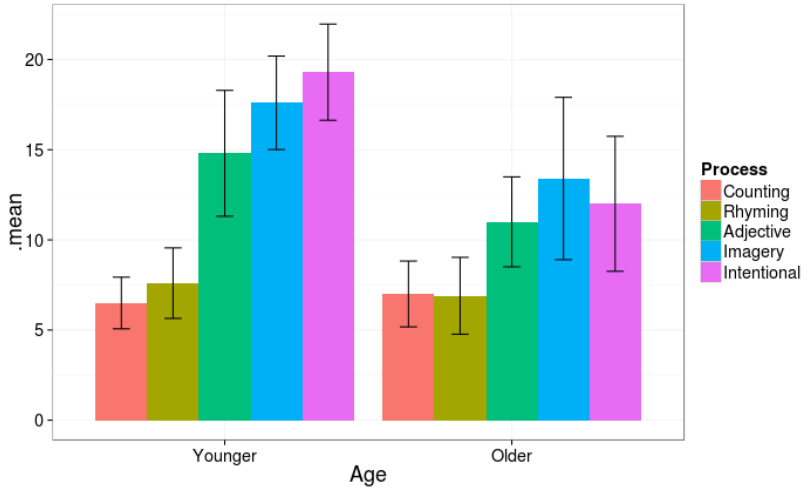
Боксплот

```
mem_p # боксплот у нас уже есть
```



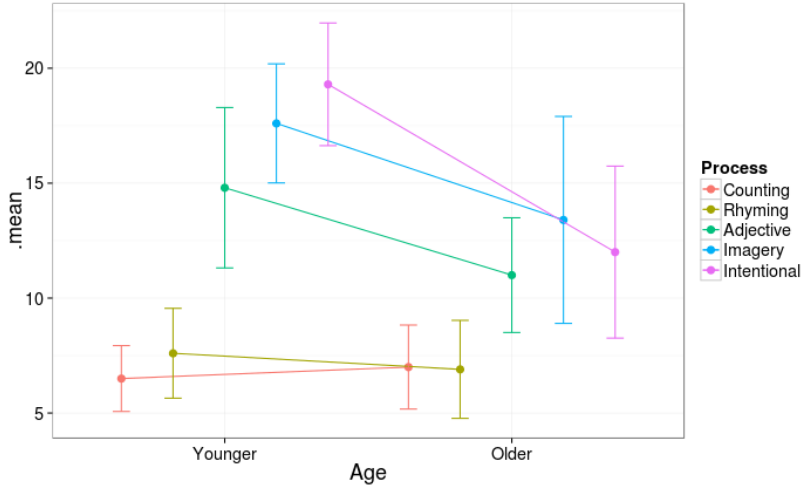
Столбчатый график

```
mem_barp <- ggplot(data = memory_summary, aes(x = Age, y = .mean, ymin = .mean - .sd, ymax = .mean + .sd,
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(width = 0.3, position = position_dodge(width = 0.9))
mem_barp
```



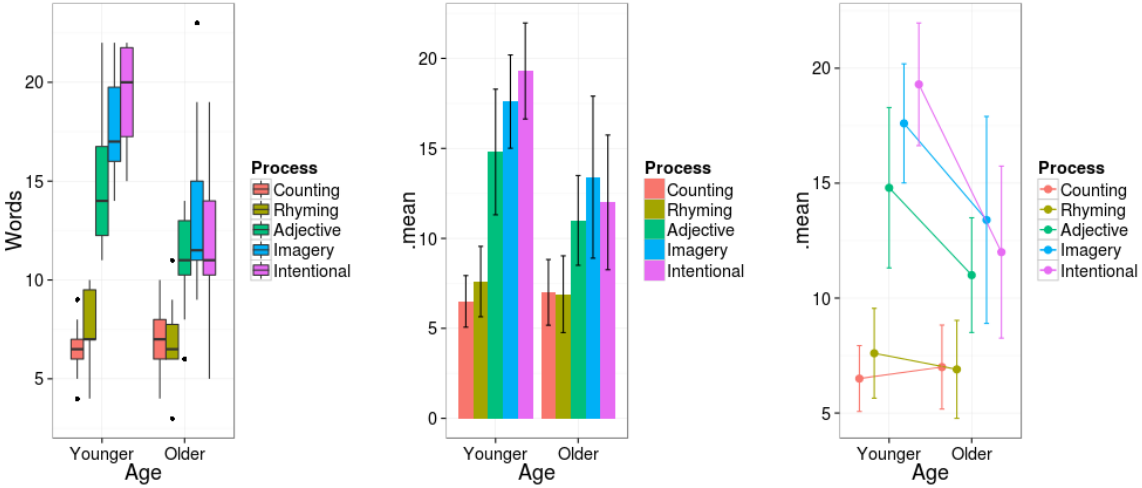
Линии с точками

```
mem_linep <- ggplot(data = memory_summary, aes(x = Age, y = .mean, ymin = .mean - .sd, ymax = .mean + .sd
  geom_point(size = 3, position = position_dodge(width = 0.9)) +
  geom_line(position = position_dodge(width = 0.9)) +
  geom_errorbar(width = 0.3, position = position_dodge(width = 0.9))
mem_linep
```



Какой график лучше выбрать?

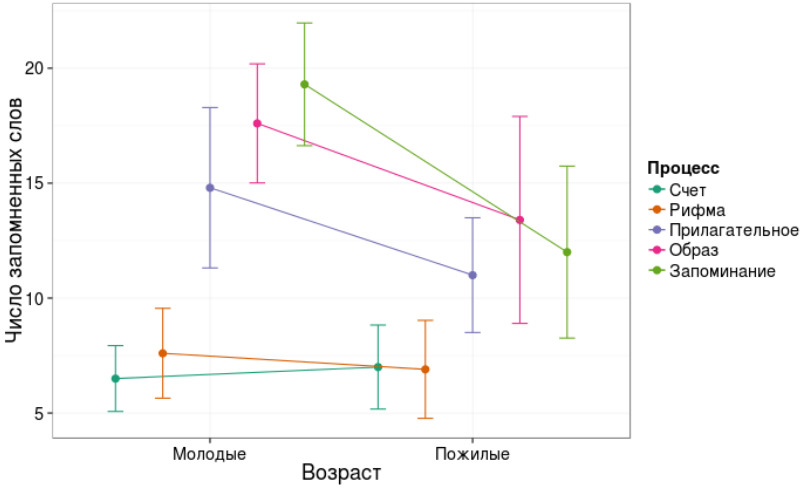
```
library(gridExtra)
grid.arrange(mem_p, mem_barp, mem_linep, ncol = 3)
```



- Должен быть максимум данных в минимуме чернил

Максимум данных в минимуме чернил (Tufte, 1983)

```
mem_linep <- mem_linep + labs(x = "Возраст", y = "Число запомненных слов") +
    scale_x_discrete(labels = c("Молодые", "Пожилые")) + scale_colour_brewer(name = "Процесс",
    palette = "Dark2", labels = c("Счет", "Рифма", "Прилагательное",
        "Образ", "Запоминание")) + theme(legend.key = element_blank())
mem_linep
```



Несбалансированные данные

Сложности с разной численностью групп

Проблемы несбалансированных дизайнов

- Оценки средних в разных группах с разным уровнем точности (Underwood 1997)
- ANOVA менее устойчив к отклонениям от условий применимости (особенно от гомогенности дисперсий) при разных размерах групп (Quinn Keough 2002, section 8.3)
- Сложно рассчитывать компоненты дисперсии (Quinn Keough 2002, section 8.2)
- Проблемы с рассчетом мощности. Если $\sigma_\epsilon^2 > 0$ и размеры выборок разные, то $\frac{MS_{groups}}{MS_{residuals}}$ не следует F-распределению (Searle et al. 1992).

- Для фикс. эффектов неравные размеры не проблема - только если значения p близкие к α
- Мораль: старайтесь *планировать* группы равной численности!

Суммы квадратов в несбалансированных дизайнах

- SSe и SSab также как в сбалансированных
- SSa, SSb - по-разному, суммы квадратов:
  - I тип (Type I SS)
  - II тип (Type II SS)
  - III тип (Type III SS)

Типы сумм квадратов

ТИПЫ СУММ КВАДРАТОВ	I ТИП	II ТИП	III ТИП
Название	Последовательная	Без учета взаимодействий высоких порядков	Иерархическая
Величина эффекта зависит от выборки в группе	Да	Да	Нет
Результат зависит от порядка включения факторов в модель	Да	Да	Нет
Команда R	`aov()`	`Anova()` (пакет `car`)	`Anova()` (пакет `car`)

- Для сбалансированных дизайнов - результаты одинаковы
- Для несбалансированных дизайнов рекомендуют **суммы квадратов III типа** (Maxwell & Delaney 1990, Milliken, Johnson 1984, Searle 1993, Yandell 1997)

Дисперсионный анализ для несбалансированных данных

Данные для демонстрации

```
umemory <- memory
# Случайные целые числа
sample.int(10, 3) # 3 случайных из 10
```

```
## [1] 10  8  6
```

```


# Заменим 5 случайных NA
set.seed(2590) # чтобы на разных системах совп. случайные числа
umemory$Words[sample.int(100, 5)] <- NA
```

Сделайте краткое описание данных

- В каких группах численность меньше 10?

```
# создайте датафрейм umemory_summary
ddply()
summarise()
sum(!is.na())
mean()
var()
sd()

umemory_summary <-
```

Описательная статистика

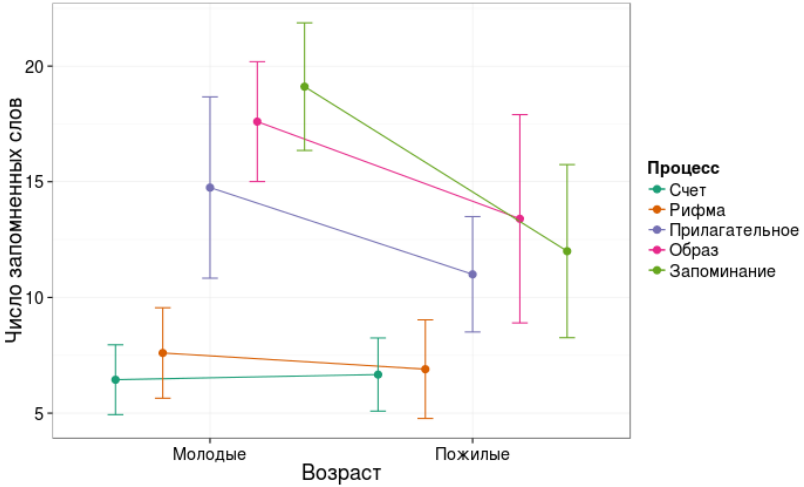
- Внимание! У нас есть NA, нужно добавить `na.rm = TRUE`

```
##        Age     Process .n .mean  .var  .sd
## 1  Younger    Counting  9  6.44  2.28 1.51
## 2  Younger     Rhyming 10  7.60  3.82 1.96
## 3  Younger   Adjective  8 14.75 15.36 3.92
## 4  Younger     Imagery 10 17.60  6.71 2.59
## 5  Younger Intentional  9 19.11  7.61 2.76
## 6    Older    Counting  9  6.67  2.50 1.58
## 7    Older     Rhyming 10  6.90  4.54 2.13
## 8    Older   Adjective 10 11.00  6.22 2.49
## 9    Older     Imagery 10 13.40 20.27 4.50
## 10   Older Intentional 10 12.00 14.00 3.74
```

Красивый график из прошлого примера с другим датафреймом

`%+%` - заменяет датафрейм в `ggplot()`

```
mem_linep %+% umemory_summary
```



Сравните результаты с использованием SS II и SS III

```
library(car)
umem_aov <- aov(Words ~ Age + Process + Age*Process, data = umemory)
```

```
Anova(umem_aov, type=2)
```

```
## Anova Table (Type II tests)
##
## Response: Words
##             Sum Sq Df F value   Pr(>F)
## Age            230  1   27.70 0.000001 ***
## Process       1449  4   43.58  < 2e-16 ***
## Age:Process    163  4    4.89   0.0013 **
## Residuals      707 85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*
```

```
Anova(umem_aov, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: Words
##             Sum Sq Df F value      Pr(>F)
## (Intercept)    374  1   44.96 0.0000000021 *
## Age              0  1    0.03      0.8705
## Process       1251  4   37.60      < 2e-16 *
## Age:Process    163  4    4.89      0.0013 *
## Residuals      707 85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
```

Take home messages

- В зависимости от типа факторов (фиксированные или случайные) по разному формулируются гипотезы и рассчитывается F-критерий.
- Если значимо взаимодействие факторов, то лучше воздержаться от интерпретации их индивидуальных эффектов
- Если численности групп равны - получаются одинаковые результаты с использованием I, II, III типы сумм квадратов
- В случае, если численности групп неравны (несбалансированные данные) по разному тестируется значимость факторов (I, II, III типы сумм квадратов)

Дополнительные ресурсы

- Quinn, Keough, 2002, pp. 221-250
- Logan, 2010, pp. 313-359
- Sokal, Rohlf, 1995, pp. 321-362
- Zar, 2010, pp. 246-266