Дисперсионный анализ, часть 2

Математические методы в зоологии - на R, осень 2015

Марина Варфоломеева

Многофакторный дисперсионный анализ

- Модель многофакторного дисперсионного анализа
- Взаимодействие факторов
- Несбалансированные данные, типы сумм квадратов
- Многофакторный дисперсионный анализ в R
- Фиксированные и случайные факторы

Вы сможете

- Проводить многофакторный дисперсионный анализ и интерпретироваот его результаты с учетом взаимодействия факторов
- Отличать фиксированные и случайные факторы и выбирать подходящую модель дисперсионного анализа

Модель многофакторного дисперсионного анализа

Линейные модели для факторных дисперсионных анализов

- Два фактора A и B, двухфакторное взаимодействие

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

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

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

Взаимодействие факторов

Взаимодействие факторов

Взаимодействие факторов - когда эффект фактора B разный в зависимости от уровней фактора A и наоборот

На каких рисунках есть взаимодействие факторов?

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

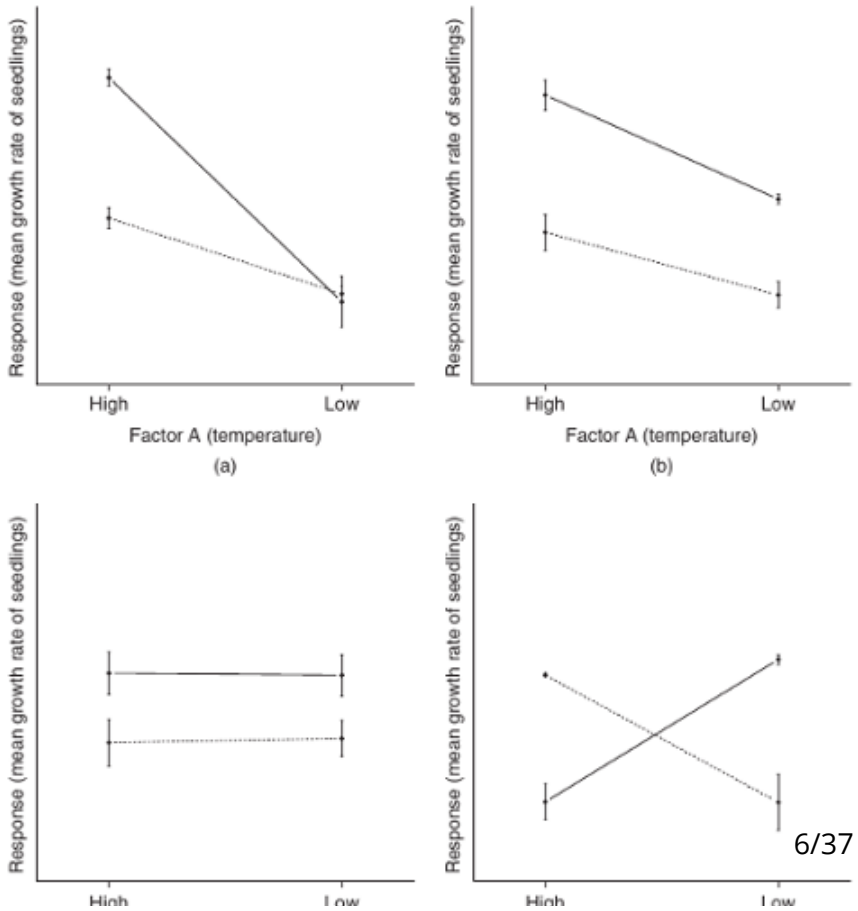


Рисунок из Logan, 2010, fig.12.2

Взаимодействие факторов может маскировать главные эффекты

- Если есть значимое взаимодействие
- пост хок тесты только по нему.
- главные эффекты обсуждать не имеет смысла

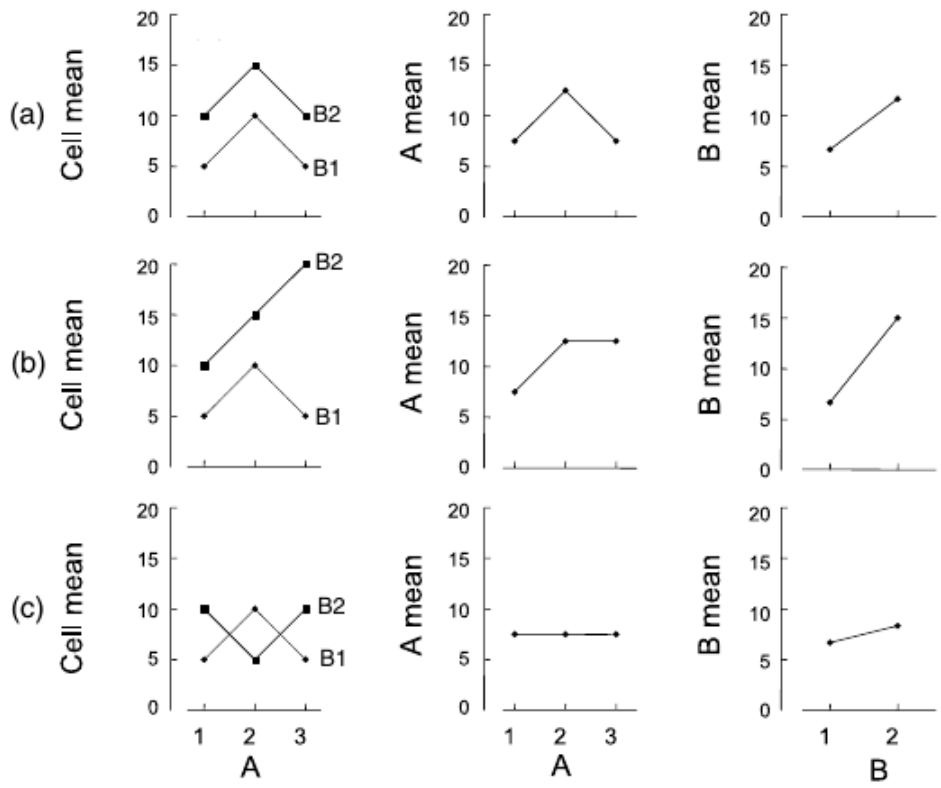


Рисунок из Quinn, Keough, 2002, fig.9.3

Задаем модель со взаимодействием в R

Взаимодействие обозначается `:` - двоеточием

Если есть факторы A и B, то их взаимодействие A:B

Для такой модели $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

Формула модели со взаимодействием:

Y ~ A + B + A:B

Сокращенная запись такой же модели обозначает, что модель включает все главные эффекты и их взаимодействия:

Y ~ A*B

Несбалансированные данные, типы сумм квадратов

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)
- Проблемы с рассчетом мощности. Если $\sigma^2_\epsilon > 0$ и размеры выборок разные, то $\frac{MS_{groups}}{MS_{residuals}}$ не следует F-распределению (Searle et al. 1992).
- Для фикс. эффектов неравные размеры - проблема только если значения p близкие к α
- Мораль: старайтесь *планировать* группы равной численности!

Если несбалансированные данные, выберите правильный тип сумм квадратов

- SSe и SSab также как в сбалансированных
- SSa, SSb - три способа расчета
- Для сбалансированных дизайнов - результаты одинаковы
- Для несбалансированных дизайнов рекомендуют **суммы квадратов III типа** если есть взаимодействие факторов (Maxwell & Delaney 1990, Milliken, Johnson 1984, Searle 1993, Yandell 1997)

Типы сумм квадратов в дисперсионном анализе

Типы сумм квадратов	**I тип**	**II тип**	**III тип**
Название	Последовательная	Без учета взаимодействий высоких порядков	Иерархическая
SS	SS(A), SS(B | A) SS(AB | B, A)	SS(A | B) SS(B | A) SS(AB | B, A)	SS(A | B, AB) SS(B | A, AB) SS(AB | B, A)
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`)

Многофакторный дисперсионный анализ в R

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

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

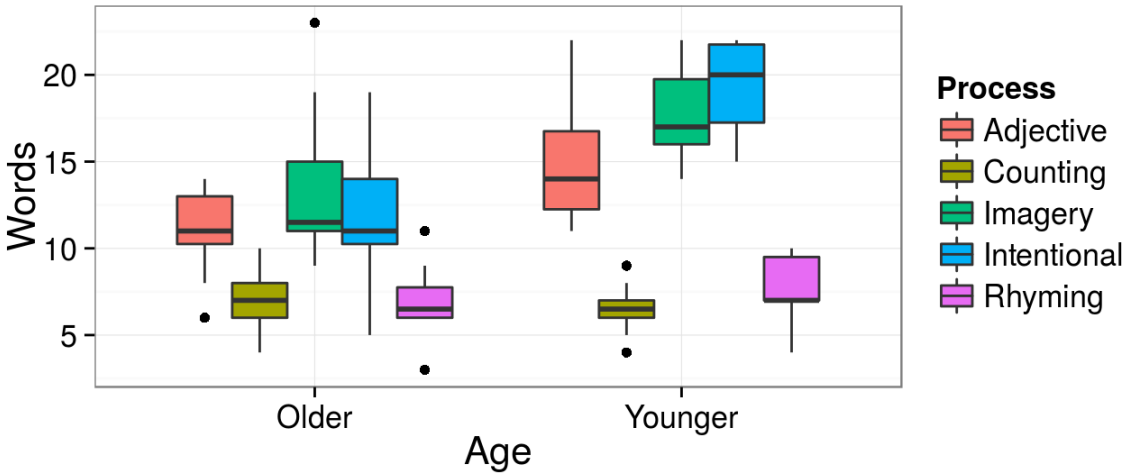
Открываем данные

```
memory <- read.delim(file = "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
```

Посмотрим на боксплот

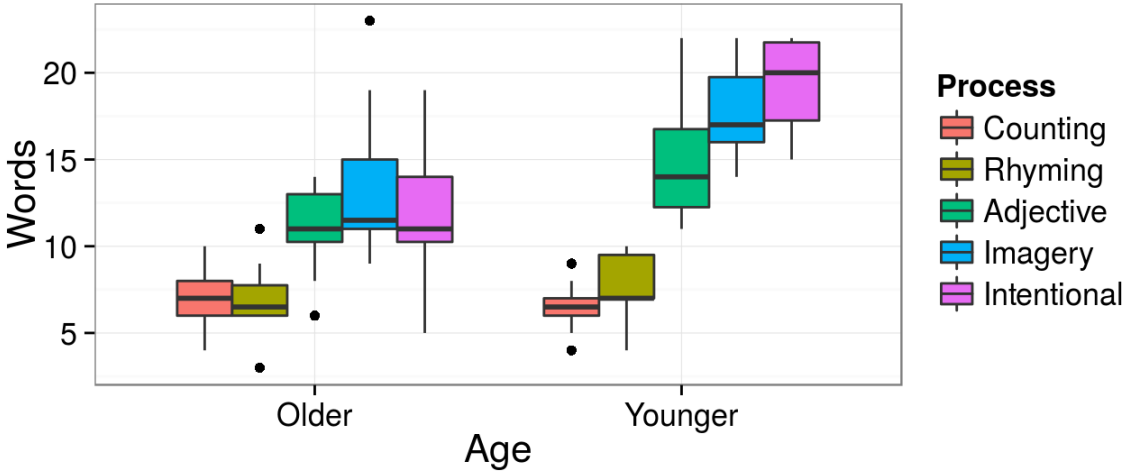
```
library(ggplot2)
theme_set(theme_bw(base_size = 16) + theme(legend.key = element_blank()))
ggplot(data = memory, aes(x = Age, y = Words)) +
  geom_boxplot(aes(fill = Process))
```



Некрасивый порядок уровней `memory$Process`

Боксплот с правильным порядком уровней

```
# переставляем в порядке следования средних значений memory$Words
memory$Process <- reorder(memory$Process, memory$Words, FUN=mean)
mem_p <- ggplot(data = memory, aes(x = Age, y = Words)) +
  geom_boxplot(aes(fill = Process))
mem_p
```



Подбираем линейную модель

Внимание: при использовании III типа сумм квадратов, нужно при подборе линейной модели **обязательно указывать тип контрастов для факторов**. В данном случае - `contrasts=list(Age=contr.sum, Process=contr.sum)`

```
memory_fit <- lm(formula = Words ~ Age * Process, data = memory,
contrasts=list(Age=contr.sum, Process=contr.sum))
```

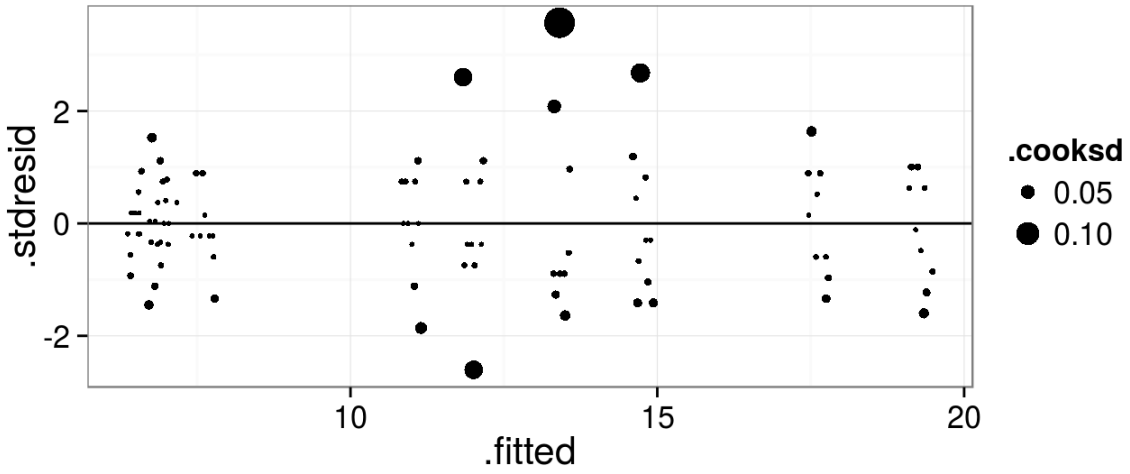
Задание: Проверьте условия применимости дисперсионного анализа

- Есть ли гомогенность дисперсий?
- Не видно ли трендов в остатках?
- Нормальное ли у остатков распределение?

Решение: 1. Проверяем условия применимости

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

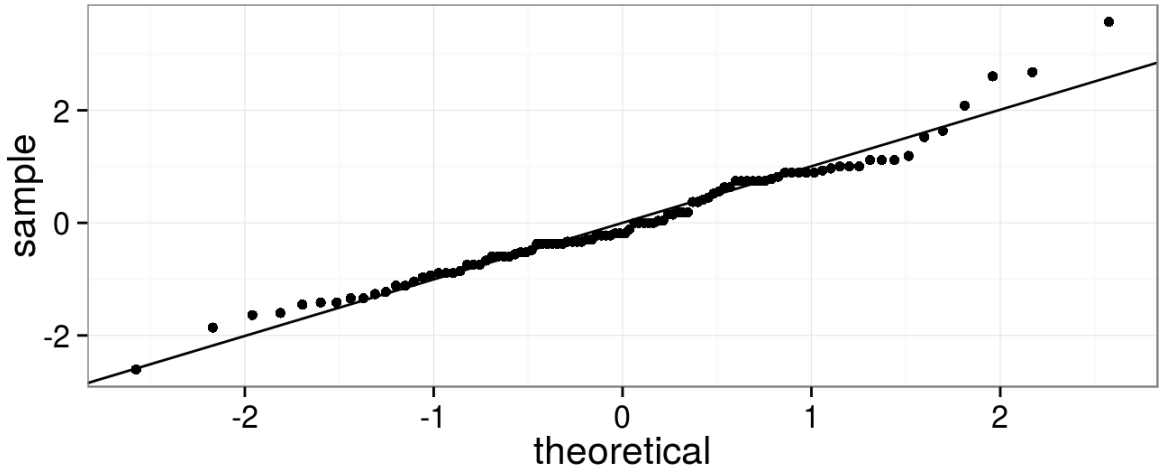
```
memory_diag <- fortify(memory_fit)
ggplot(memory_diag, aes(x = .fitted, y = .stdresid)) + geom_point(aes(size = .c
    position = position_jitter(width = 0.2)) + geom_hline(yintercept = 0)
```



Решение: 2. Проверяем условия применимости

- 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

```
library(car)
Anova(memory_fit, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: Words
##             Sum Sq Df F value    Pr(>F)
## (Intercept)  13479  1 1679.54   < 2e-16 ***
## Age            240  1   29.94 0.0000004 ***
## Process       1515  4   47.19   < 2e-16 ***
## Age:Process    190  4    5.93   0.00028 ***
## Residuals      722 90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Взаимодействие достоверно, факторы отдельно можно не тестировать, тк. взаимодействие может все равно изменять их эффект до неузнаваемости.
- Нужно делать пост хок тест по взаимодействию факторов

Пост хок тест по взаимодействию факторов

```r
# 1. создаем переменную-взаимодействие
memory$AgeProc <- interaction(memory$Age, memory$Process)
# 2. подбираем модель без intercept
cell_means <- lm(Words ~ AgeProc - 1, data = memory)
# 3. делаем пост хок тест
library(multcomp)
memory_tukey <- glht(cell_means, linfct = mcp(AgeProc = "Tukey"))
options(width = 90)
summary(memory_tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Words ~ AgeProc - 1, data = memory)
##
## Linear Hypotheses:
##                                                Estimate Std. Error t value Pr(
## Younger.Counting - Older.Counting == 0            -0.50       1.27   -0.39   1
## Older.Rhyming - Older.Counting == 0               -0.10       1.27   -0.08   1
## Younger.Rhyming - Older.Counting == 0              0.60       1.27    0.47   1
```

Смотрим на результаты пост хок теста

В виде таблицы результаты практически не читаемы. Лучше построить график.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Words ~ AgeProc - 1, data = memory)
##
## Linear Hypotheses:
##                                              Estimate Std. Error t value Pr(>|t|)
## Younger.Counting - Older.Counting == 0          -0.50       1.27   -0.39    1.000
## Older.Rhyming - Older.Counting == 0             -0.10       1.27   -0.08    1.000
## Younger.Rhyming - Older.Counting == 0            0.60       1.27    0.47    1.000
## Older.Adjective - Older.Counting == 0            4.00       1.27    3.16    0.063 .
## Younger.Adjective - Older.Counting == 0          7.80       1.27    6.16    <0.01 ***
## Older.Imagery - Older.Counting == 0              6.40       1.27    5.05    <0.01 ***
## Younger.Imagery - Older.Counting == 0           10.60       1.27    8.37    <0.01 ***
## Older.Intentional - Older.Counting == 0          5.00       1.27    3.95    <0.01 **
## Younger.Intentional - Older.Counting == 0       12.30       1.27    9.71    <0.01 ***
## Older.Rhyming - Younger.Counting == 0            0.40       1.27    0.32    1.000
## Younger.Rhyming - Younger.Counting == 0          1.10       1.27    0.87    0.997
## Older.Adjective - Younger.Counting == 0          4.50       1.27    3.55    0.021 *
## Younger.Adjective - Younger.Counting == 0        8.30       1.27    6.55    <0.01 ***
## Older.Imagery - Younger.Counting == 0            6.90       1.27    5.45    <0.01 ***
## Younger.Imagery - Younger.Counting == 0         11.10       1.27    8.76    <0.01 ***
```

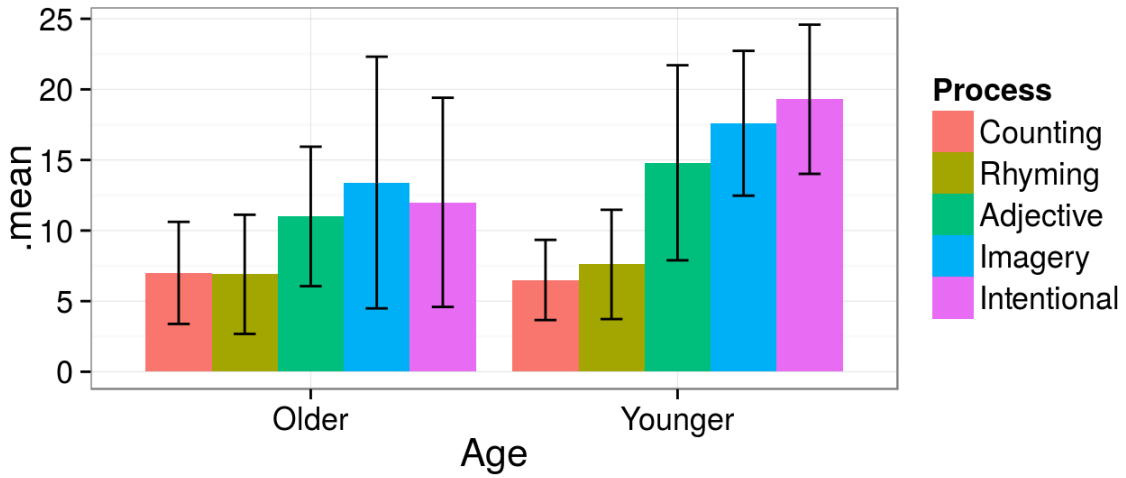
Данные для графиков

```
# __Статистика по столбцам и по группам__ одновременно (n, средние,
# стандартные отклонения)
library(dplyr)
memory_summary <- memory %>%
  group_by(Age, Process) %>%
  summarise(
    .n = sum(!is.na(Words)),
    .mean = mean(Words, na.rm = TRUE),
    .sd = sd(Words, na.rm = TRUE))
memory_summary
```

```
## Source: local data frame [10 x 5]
## Groups: Age [?]
##
##       Age     Process    .n .mean   .sd
##    (fctr)      (fctr) (int) (dbl) (dbl)
## 1   Older    Counting    10   7.0  1.83
## 2   Older     Rhyming    10   6.9  2.13
## 3   Older   Adjective    10  11.0  2.49
## 4   Older     Imagery    10  13.4  4.50
## 5   Older Intentional    10  12.0  3.74
## 6 Younger    Counting    10   6.5  1.43
## 7 Younger     Rhyming    10   7.6  1.96
```

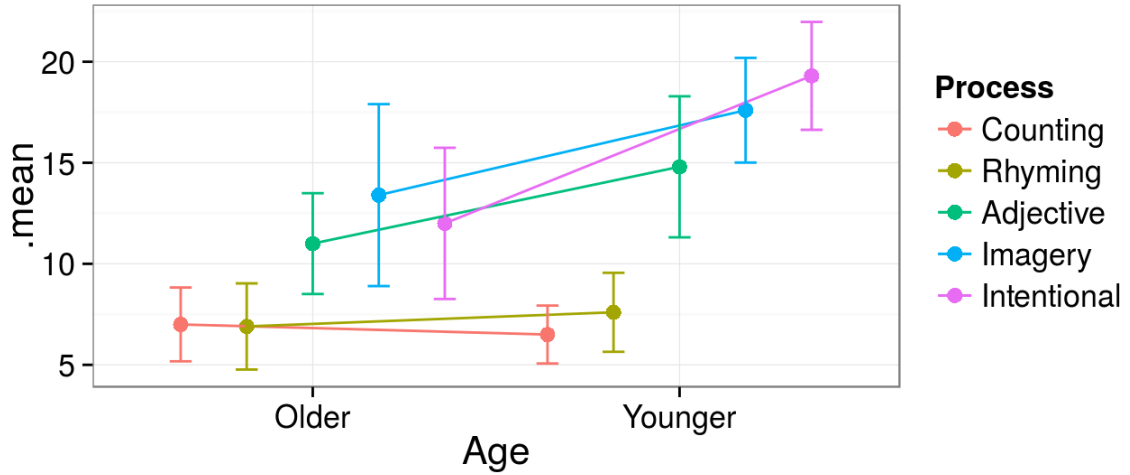
Графики для результатов: Столбчатый график

```
mem_barp <- ggplot(data = memory_summary,
                   aes(x = Age, y = .mean, ymin = .mean - 1.98*.sd,
                       ymax = .mean + 1.98*.sd, fill = Process)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(width = 0.3, position = position_dodge(width = 0.9))
mem_barp
```



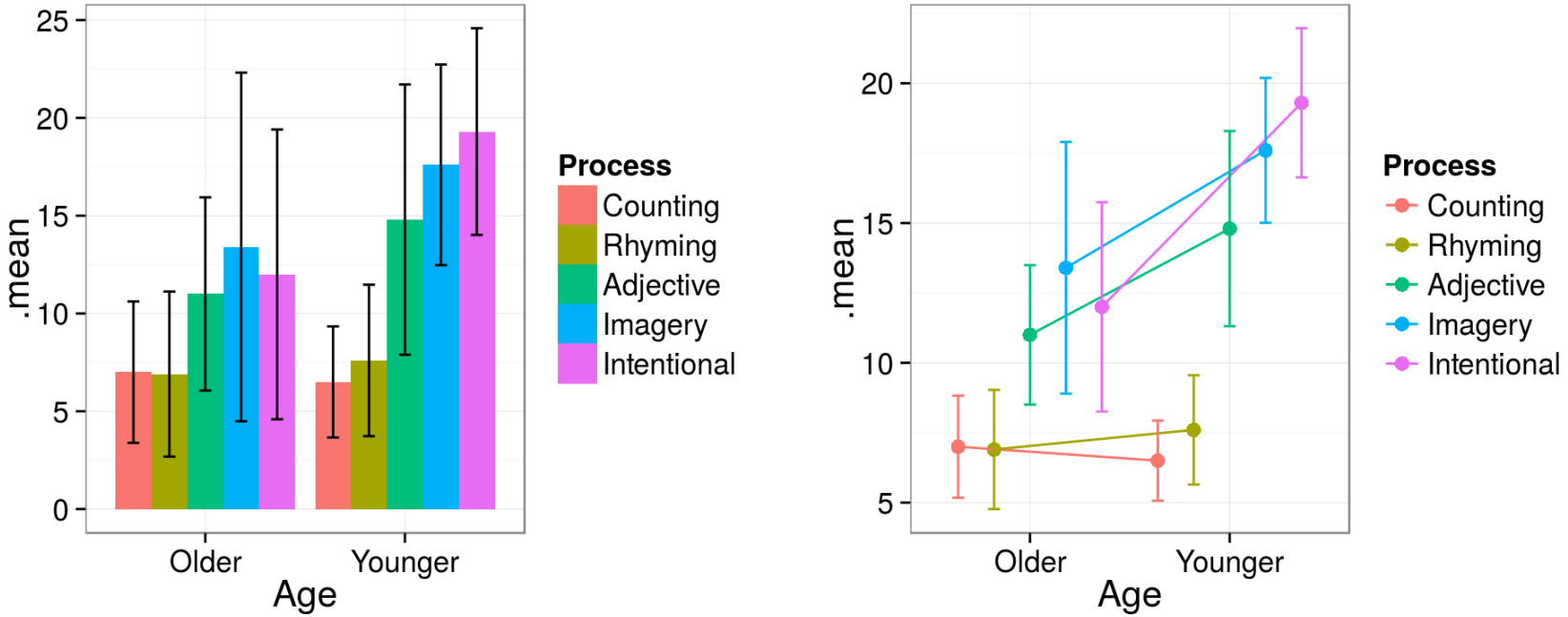
Графики для результатов: Линии с точками

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



Какой график лучше выбрать?

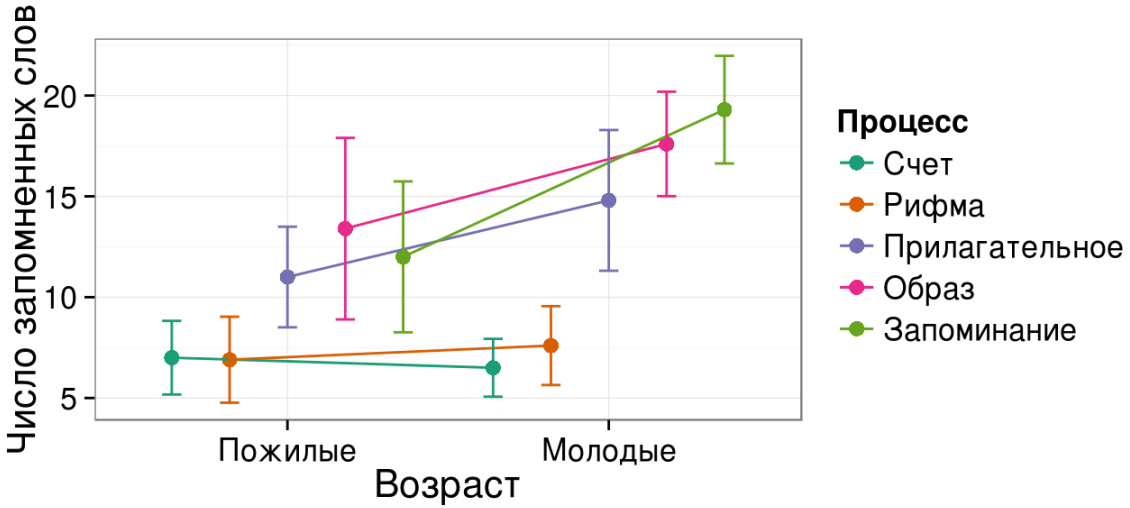
```
library(gridExtra)
grid.arrange(mem_barp, mem_linep, ncol = 2)
```



- Должен быть максимум данных в минимуме чернил (Tufte, 1983)

Приводим понравившийся график в приличный вид

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



Фиксированные и случайные факторы

Фиксированные и случайные факторы

- Фиксированные факторы
  - возможные градации фактора заранее известны, уровни фактора выбраны не случайно из небольшого числа возможных
  - предсказывать можно только для существующих в модели значений факторов
- Случайные факторы
  - возможные градации фактора неизвестны заранее, уровни фактора выбраны случайно из множества возможных
  - предсказывать можно для любых значений факторов

Задание: Примеры фиксированных и случайных факторов

Опишите ситуации, когда эти факторы будут фиксированными, а когда случайными

- 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.

- Приведите другие примеры того, 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	A и B фиксированные	A и B случайные	A фиксированный, B случайный
A	$F = \frac{MS_a}{MS_e}$	$F = \frac{MS_a}{MS_{ab}}$	$F = \frac{MS_a}{MS_e}$
B	$F = \frac{MS_b}{MS_e}$	$F = \frac{MS_b}{MS_{ab}}$	$F = \frac{MS_b}{MS_a b}$
AB	$F = \frac{MS_{ab}}{MS_e}$	$F = \frac{MS_{ab}}{MS_e}$	$F = \frac{MS_{ab}}{MS_e}$

Внимание: сегодня говорили только про фиксированные факторы.

Если есть случайные факторы - смешанные модели. О них в магистратуре.

Пакеты `nlme` и `lme4`

Книги:

- Pinheiro, J., Bates, D., 2000. Mixed-Effects Models in S and S-PLUS. Springer.
- Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology With R. Springer.

Take home messages

- Многофакторный дисперсионный анализ позволяет оценить взаимодействие факторов. Если оно значимо, 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) >- В зависимости от типа факторов (фиксированные или случайные) по разному формулируются гипотезы и рассчитывается F-критерий.

Дополнительные ресурсы

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