Дисперсионный анализ, часть 2

Математические методы в зоологии с использованием R

Марина Варфоломеева

1. **Модель многофакторного дисперсионного анализа**
2. **Взаимодействие факторов**
3. **Несбалансированные данные, типы сумм квадратов**
4. **Многофакторный дисперсионный анализ в R**
5. **Фиксированные и случайные факторы**

Многофакторный дисперсионный анализ

Вы сможете

- Проводить многофакторный дисперсионный анализ и интерпретироватm 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

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

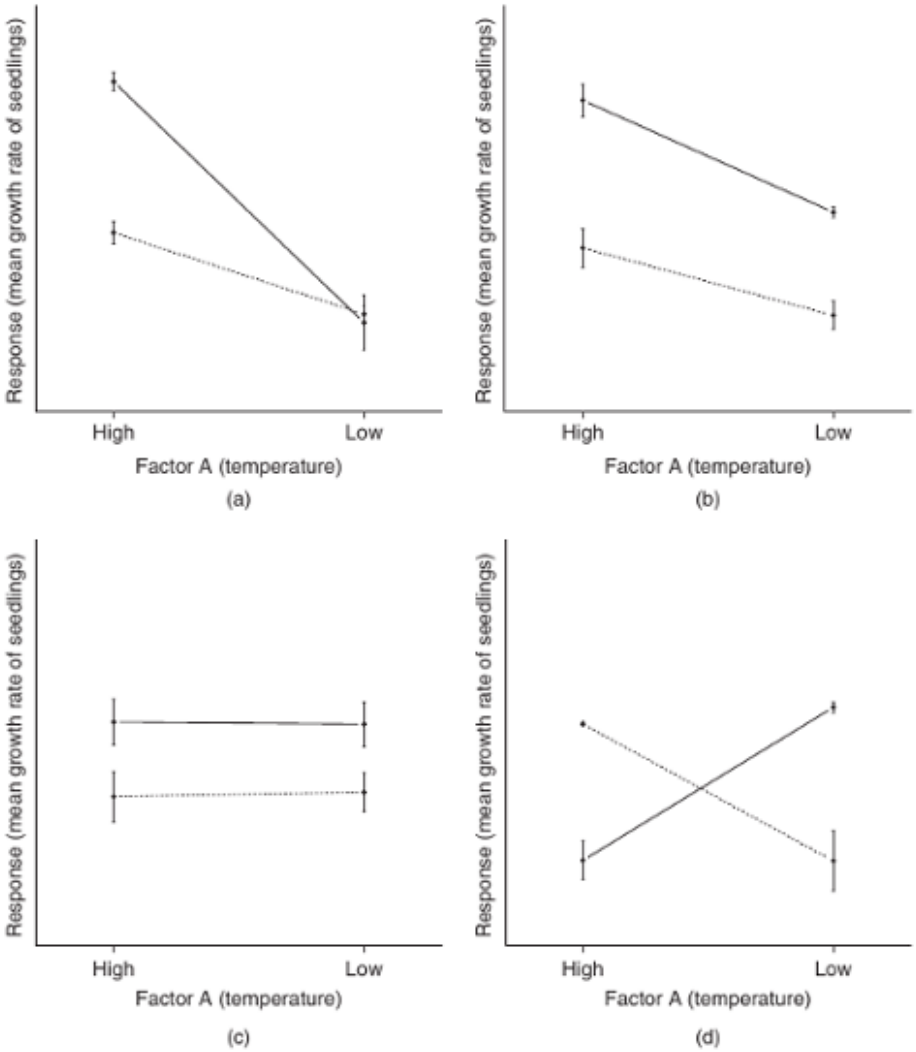
- Три фактора A, B и C, двухфакторные взаимодействия, трехфакторное взаимодействия

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

Взаимодействие факторов

Взаимодействие факторов

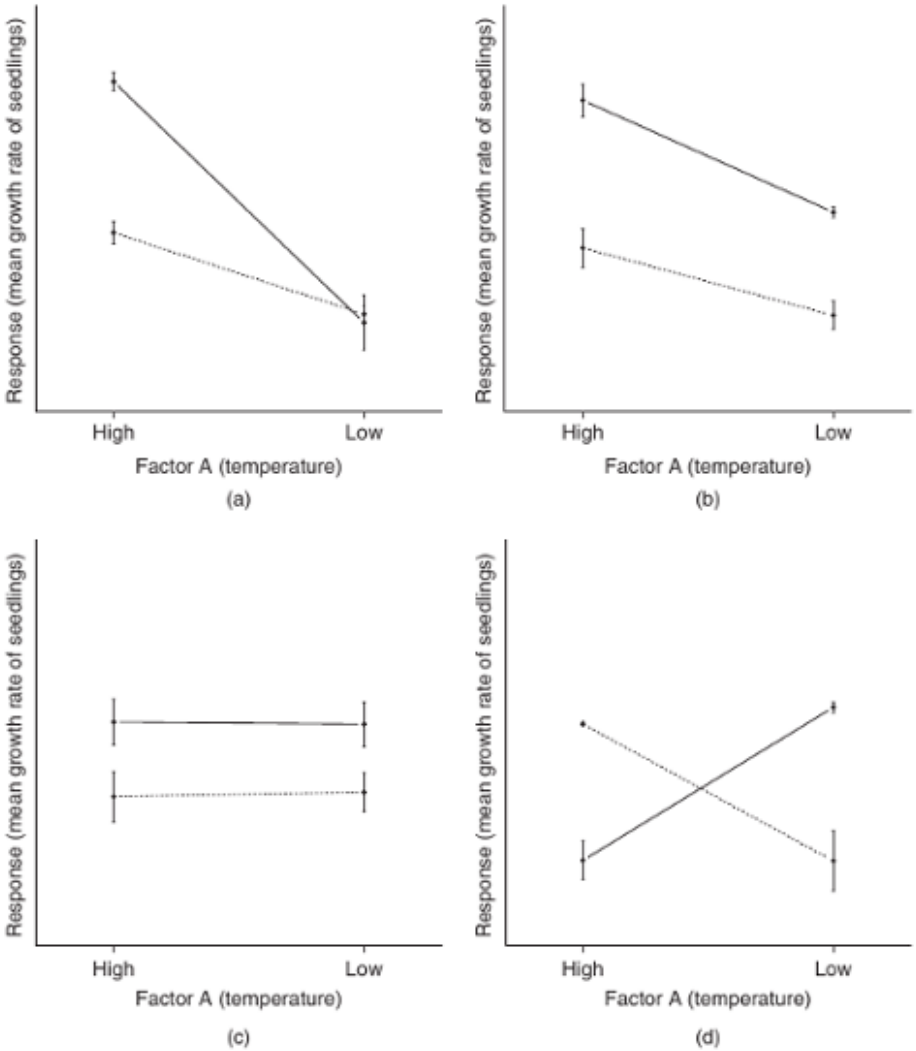
Взаимодействие факторов — когда эффект фактора B разный в зависимости от уровней фактора A и наоборот



На каких рисунках есть взаимодействие факторов?

Взаимодействие факторов

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

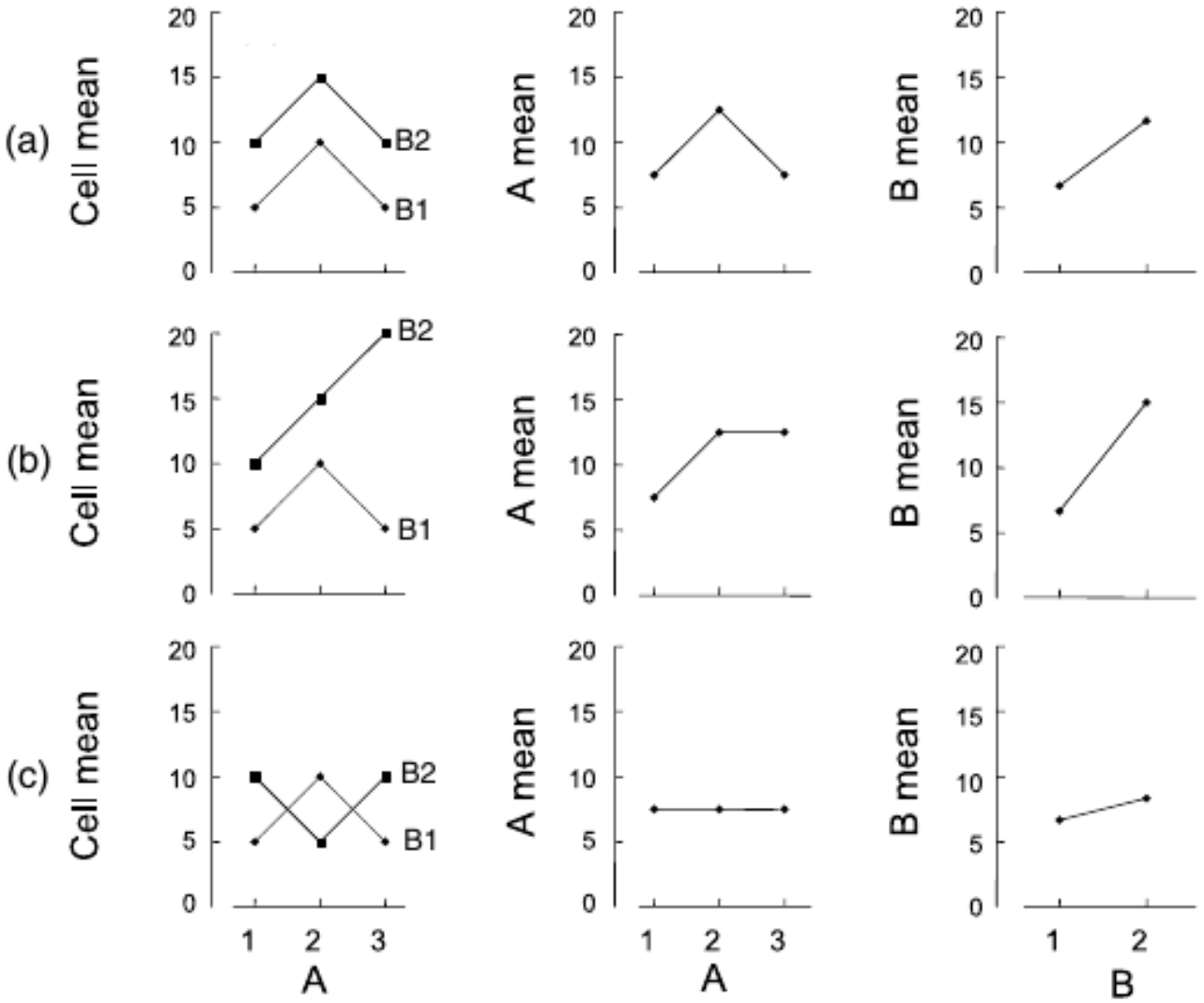


На каких рисунках есть взаимодействие факторов?

- b, c - нет взаимодействия (эффект фактора B одинаковый для групп по фактору A, линии для разных групп по фактору B на графиках расположены параллельно)
- a, d - есть взаимодействие (эффект фактора B разный для групп по фактору A, на графиках линии для разных групп по фактору B расположены под наклоном).

Рисунок из Logan, 2010, fig.12.2

Взаимодействие факторов может маскировать главные эффекты



Если есть значимое взаимодействие

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).

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).

- Старайтесь *планировать* группы равной численности!
- Но если не получилось - не страшно:
  - Для фикс. эффектов неравные размеры - проблема только если значения доверительной вероятности *p* близки к выбранному критическому уровню значимости α

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). (Правда, этот способ не самый правильный с точки зрения статистики, т.к. основные эффекты факторов тестируются так, как если бы взаимодействие было включено в модель).

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	Нет	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

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

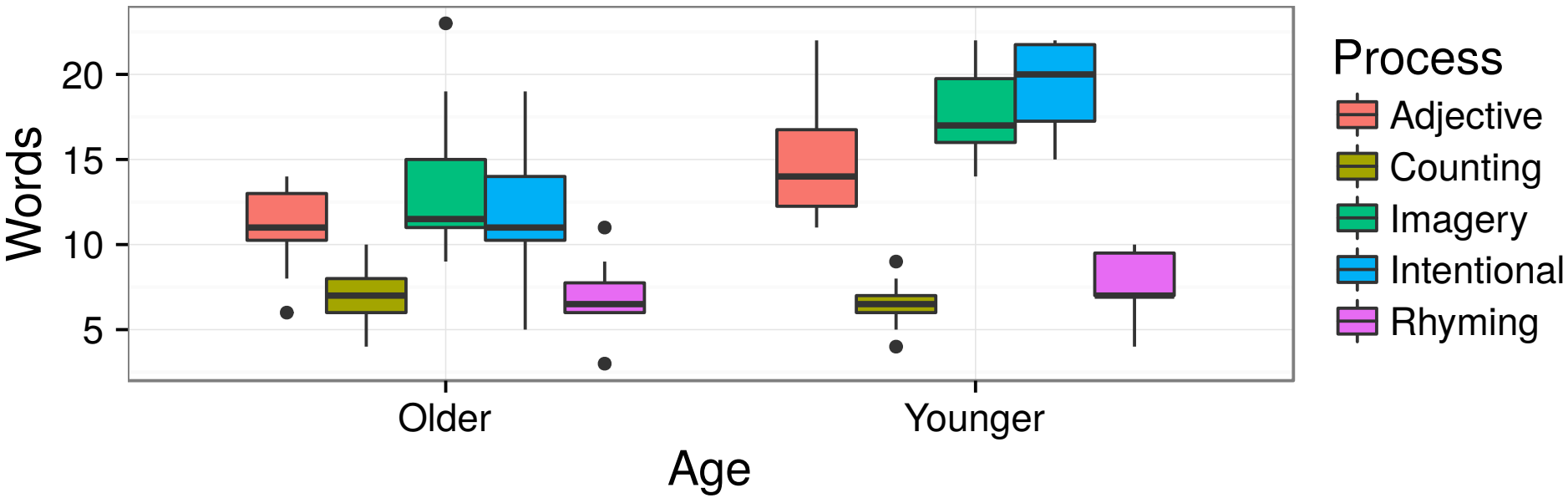
Открываем данные

```
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
```

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))
```

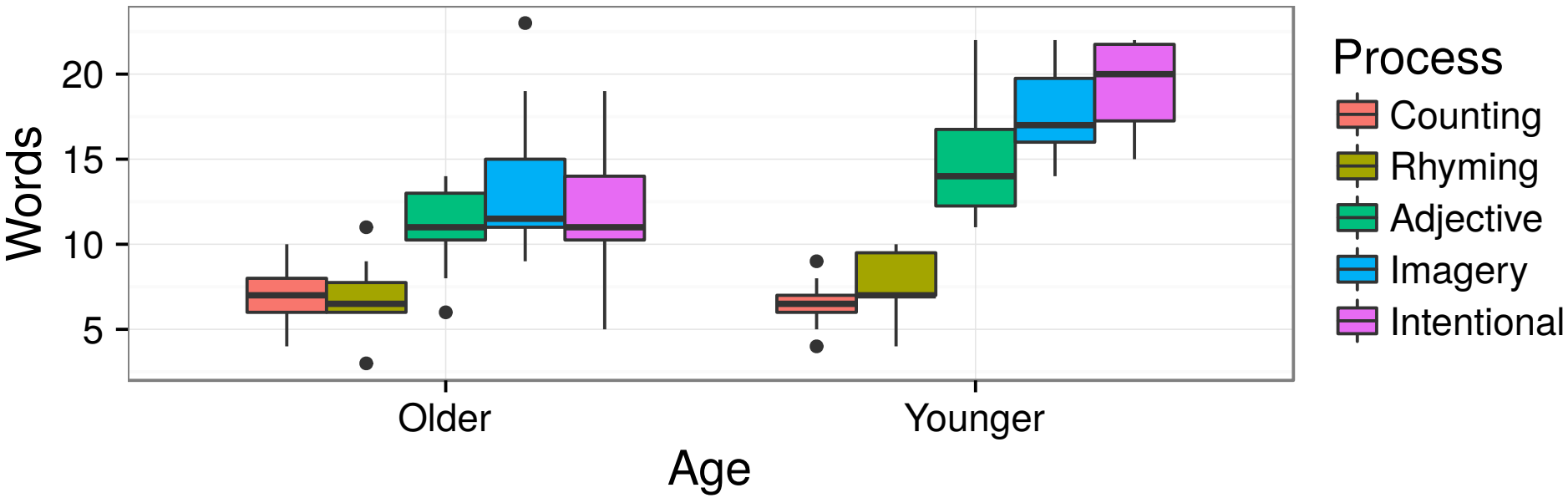


Некрасивый порядок уровней `memory$Process`

Боксплот с правильным порядком уровней

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

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



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

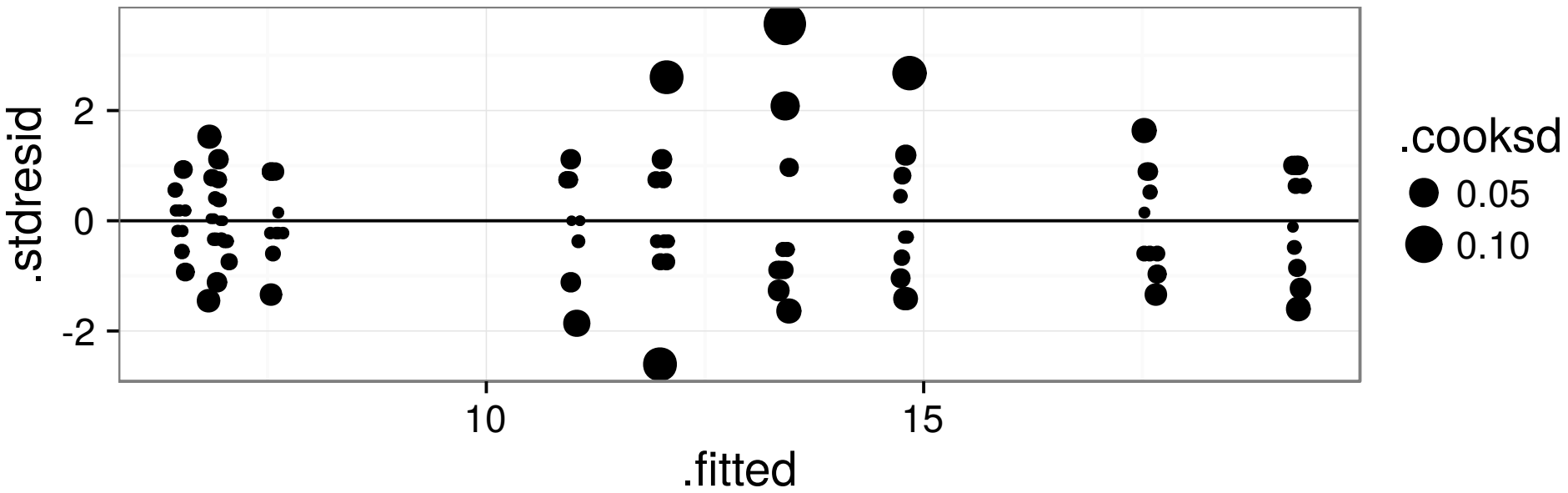
Проверьте условия применимости дисперсионного анализа

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

Решение: 1. Проверяем условия применимости

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

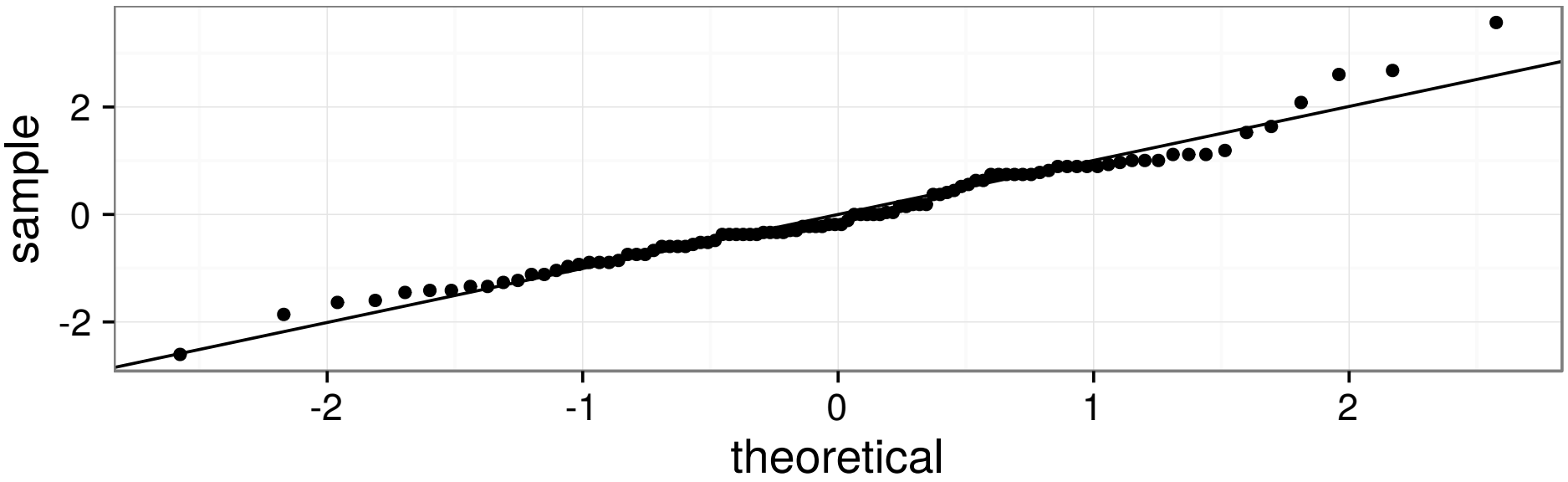
```
memory_diag <- fortify(memory_fit)
ggplot(memory_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point(aes(size = .cooksd), position = position_jitter(width = .2)) +
  geom_hline(yintercept = 0)
```



Решение: 2. Проверяем условия применимости

- Нормальное ли у остатков распределение?

```
ggplot(memory_diag) +
  geom_point(stat = "qq", aes(sample = .stdresid)) +
  geom_abline(aes(intercept = 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) 13479.2  1 1679.5361    < 2.2e-16 ***
# Age           240.3  1   29.9356 0.0000003981 ***
# Process      1514.9  4   47.1911    < 2.2e-16 ***
# Age:Process   190.3  4    5.9279    0.0002793 ***
# Residuals     722.3 90
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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) 13479.2  1 1679.5361    < 2.2e-16 ***
# Age           240.3  1   29.9356 0.0000003981 ***
# Process      1514.9  4   47.1911    < 2.2e-16 ***
# Age:Process   190.3  4    5.9279    0.0002793 ***
# Residuals     722.3 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

Пост хок тест для взаимодействия факторов делается легче всего “обходным путем”

1. Создаем переменную-взаимодействие
2. Подбираем модель без свободного члена
3. Делаем пост хок тест для этой модели

```
memory$AgeProc <- interaction(memory$Age, memory$Process)
cell_means <- lm(Words ~ AgeProc - 1, data = memory)
library(multcomp)
memory_tukey <- glht(cell_means, linfct = mcp(AgeProc = "Tukey"))
summary(memory_tukey)
```

Смотрим на результаты пост хок теста

В виде таблицы результаты нечитаемы. Лучше построить график.

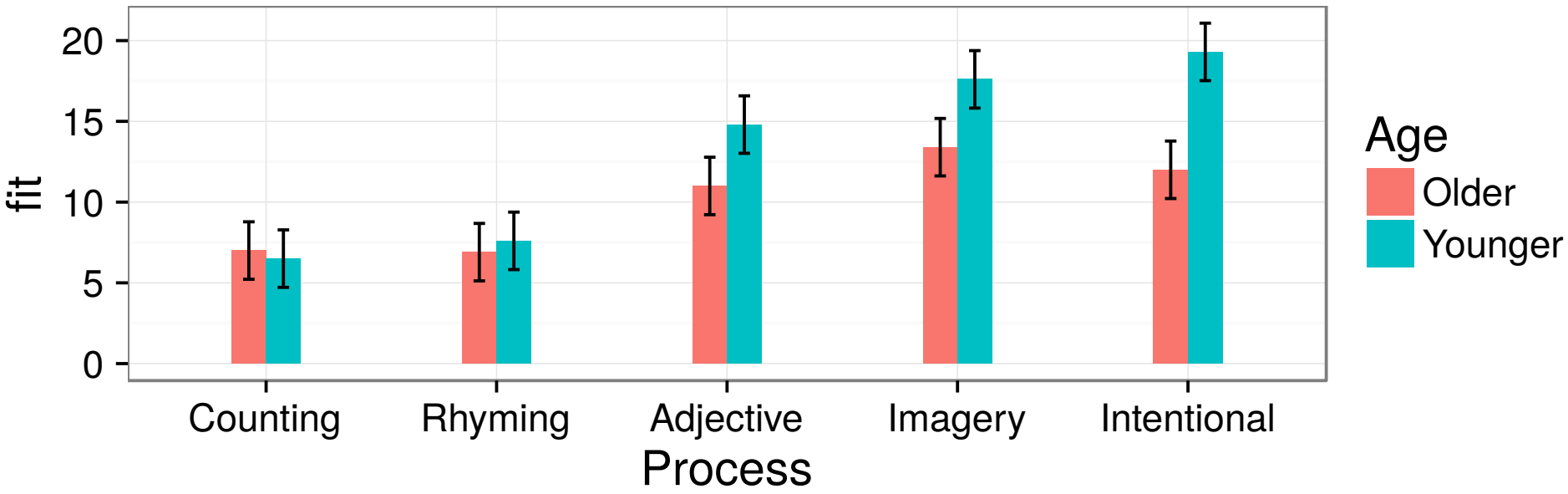
```
#
#    Simultaneous Tests for General Linear Hypotheses
#
# Multiple Comparisons of Means: Tukey Contrasts
#
#
# Fit: lm(formula = Words ~ AgeProc - 1, data = memory)
#
# Linear Hypotheses:
#                                                Estimate Std. Error t value Pr(>|t|)
# Younger.Counting - Older.Counting == 0           -0.500      1.267  -0.395   1.0000
# Older.Rhyming - Older.Counting == 0              -0.100      1.267  -0.079   1.0000
# Younger.Rhyming - Older.Counting == 0             0.600      1.267   0.474   1.0000
# Older.Adjective - Older.Counting == 0             4.000      1.267   3.157   0.0630 .
# Younger.Adjective - Older.Counting == 0           7.800      1.267   6.157    <0.01 *
# Older.Imagery - Older.Counting == 0               6.400      1.267   5.052    <0.01 *
# Younger.Imagery - Older.Counting == 0            10.600      1.267   8.367    <0.01 *
# Older.Intentional - Older.Counting == 0           5.000      1.267   3.947    <0.01 *
# Younger.Intentional - Older.Counting == 0        12.300      1.267   9.709    <0.01 *
# Older.Rhyming - Younger.Counting == 0             0.400      1.267   0.316   1.0000
# Younger.Rhyming - Younger.Counting == 0           1.100      1.267   0.868   0.9970
# Older.Adjective - Younger.Counting == 0           4.500      1.267   3.552   0.0208 *
# Younger.Adjective - Younger.Counting == 0         8.300      1.267   6.551    <0.01 *
```

Данные для графиков

```
process <- levels(memory$Process)
fprocess <- factor(process, levels = process)
MyData <- expand.grid(Age = levels(memory$Age),
                      Process = fprocess)
MyData <- data.frame(MyData,
  predict(memory_fit, newdata = MyData,
          interval = "confidence"))
```

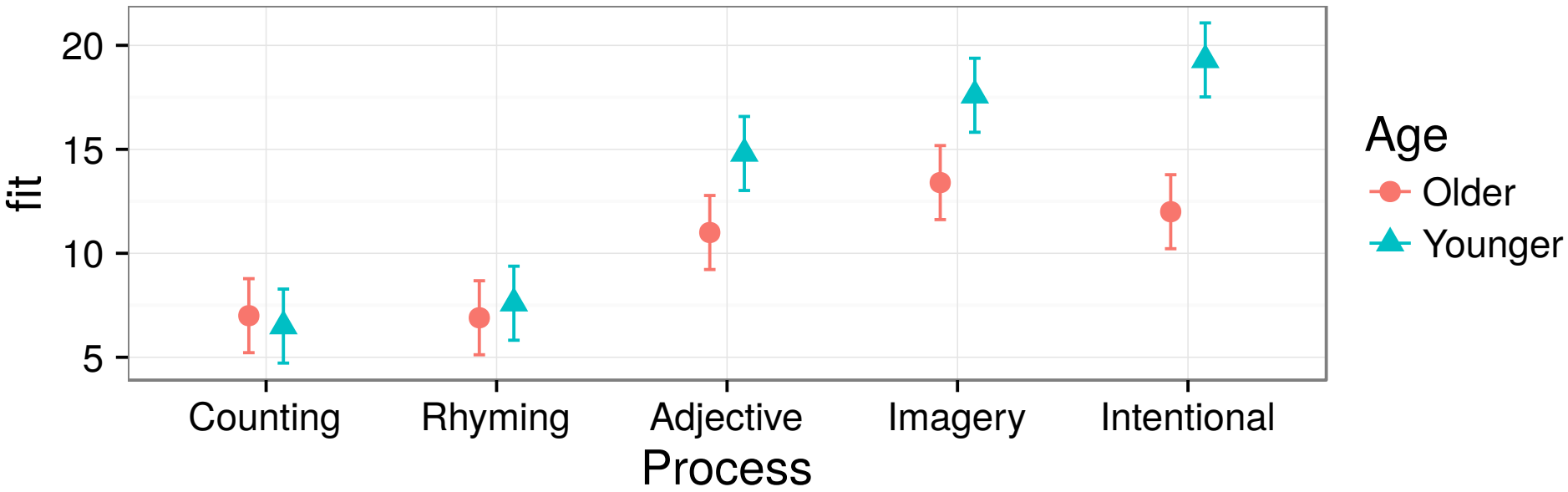
Графики для результатов: Столбчатый график

```
pos <- position_dodge(width = 0.3)
gg_barp <- ggplot(data = MyData, aes(x = Process, y = fit,
             ymin = lwr,  ymax = upr, fill = Age)) +
  geom_bar(stat = "identity", position = pos, width = 0.3) +
  geom_errorbar(width = 0.1, position = pos)
gg_barp
```



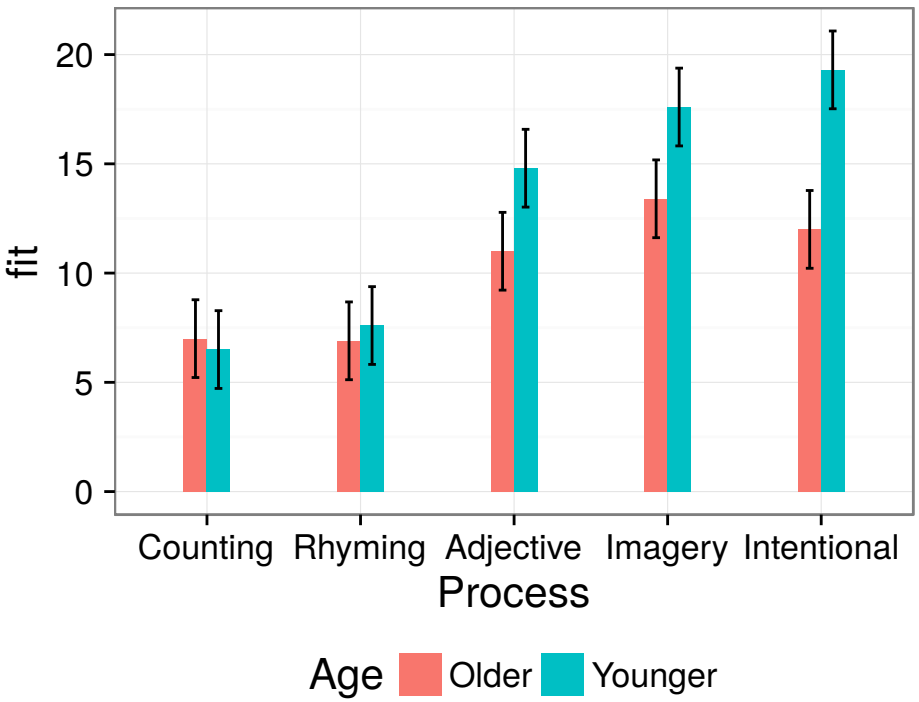
Графики для результатов: Точки

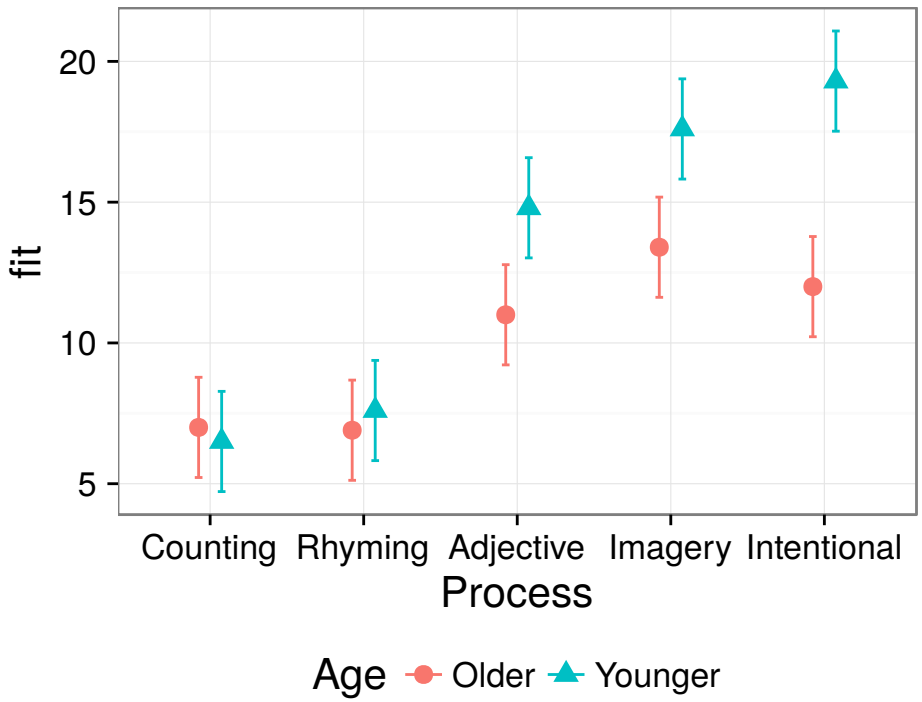
```
gg_pointp <- ggplot(data = MyData, aes(x = Process, y = fit,
                ymin = lwr,  ymax = upr, colour = Age)) +
  geom_point(aes(shape = Age), size = 3, position = pos) +
  # geom_line(aes(group = Age), position = pos) +
  geom_errorbar(width = 0.1, position = pos)
gg_pointp
```



Какой график лучше выбрать?

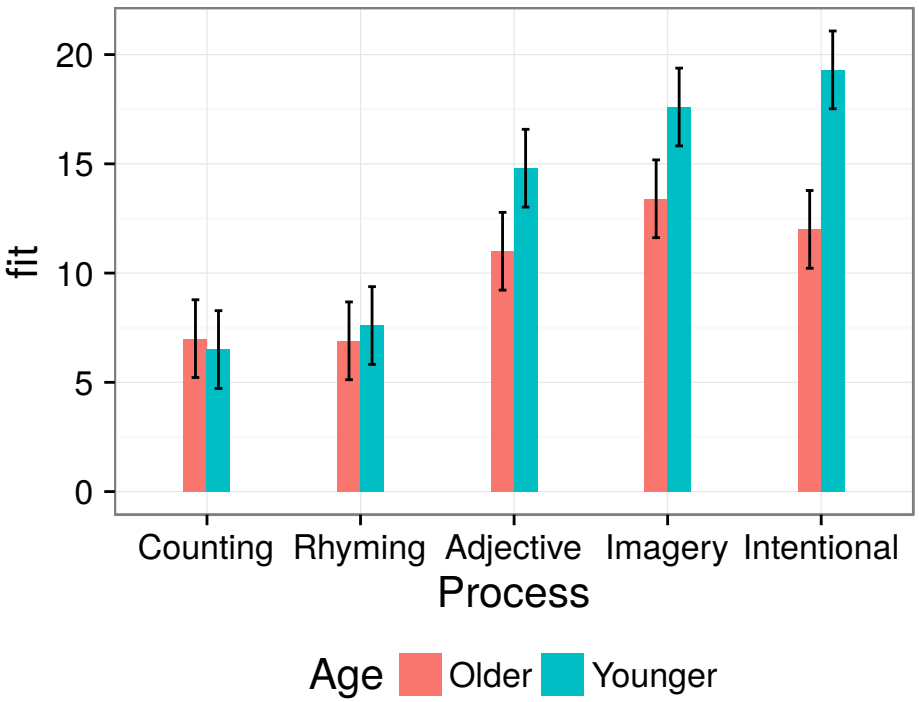
```
library(gridExtra)
grid.arrange(gg_barp + theme(legend.position = "bottom"),
             gg_pointp + theme(legend.position = "bottom"),
             ncol = 2)
```

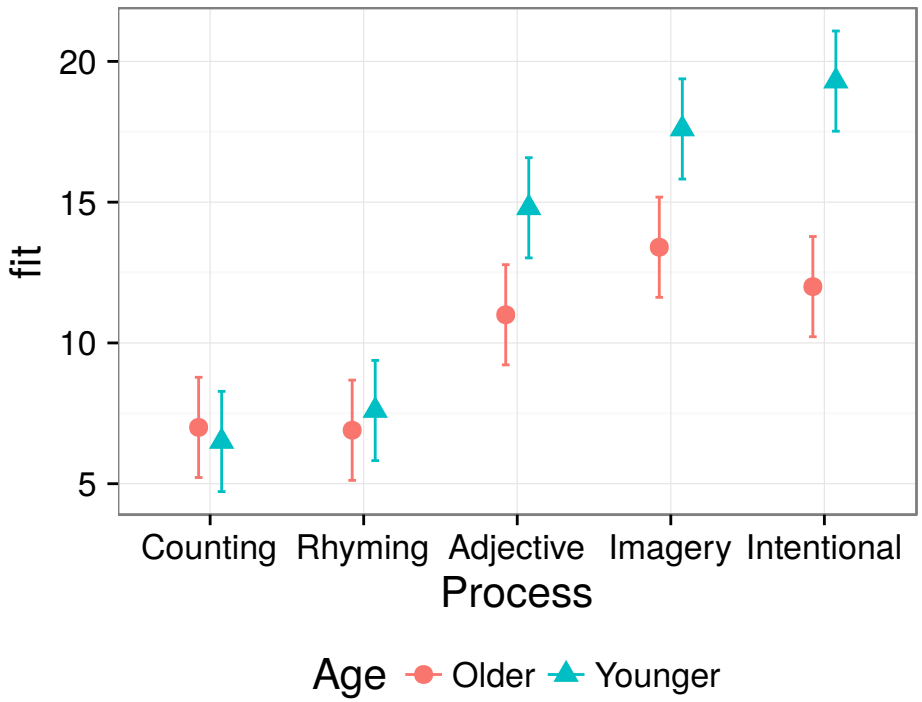




Какой график лучше выбрать?

```
library(gridExtra)
grid.arrange(gg_barp + theme(legend.position = "bottom"),
             gg_pointp + theme(legend.position = "bottom"),
             ncol = 2)
```



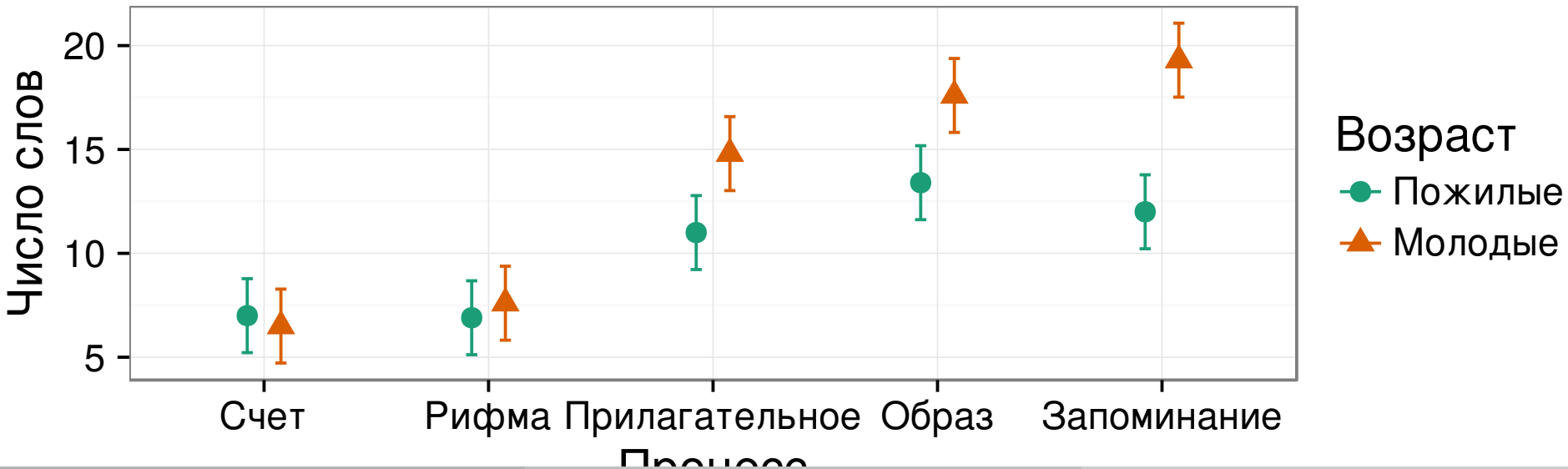


- Максимум данных в минимуме чернил (Tufte, 1983)

Приводим понравившийся график в приличный вид

```
gg_final <- gg_pointp +
  labs(y = "Число слов") +
   scale_colour_brewer(name = "Возраст", palette = "Dark2",
                       labels = c("Пожилые", "Молодые")) +
  scale_shape_discrete(name = "Возраст",
                       labels = c("Пожилые", "Молодые")) +
  scale_x_discrete(name = "Процесс", palette = "Dark2",
                       labels = c("Счет", "Рифма",
                                  "Прилагательное", "Образ", "Запоминание"))

gg_final
```



Фиксированные и случайные факторы

Фиксированные и случайные факторы

Свойства	Фиксированные факторы	Случайные факторы
Уровни фактора	фиксированные, заранее определенные и потенциально воспроизводимые уровни	случайная выборка из всех возможных уровней
Используются для тестирования гипотез	о средних значениях отклика между уровнями фактора $H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu$	о дисперсии отклика между уровнями фактора $H_0 : \sigma^2_{rand.fact.} = 0$
Выводы можно экстраполировать	только на уровни из анализа	на все возможные уровни
Число уровней фактора	Осторожно! Если уровней фактора слишком много, то нужно подбирать слишком много коэффициентов — должно быть много данных	Важно! Для точной оценки σ нужно нужно много уровней фактора — не менее 5

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.

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

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