Дисперсионный анализ, часть 2

Линейные модели…

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Многофакторный дисперсионный анализ

- Модель многофакторного дисперсионного анализа
- Взаимодействие факторов
- Несбалансированные данные, типы сумм квадратов
- Многофакторный дисперсионный анализ в R
- Дисперсионный анализ в матричном виде

Вы сможете

- Проводить многофакторный дисперсионный анализ и интерпретировать его результаты с учетом взаимодействия факторов

Данные

Пример: Удобрение и беспозвоночные

Влияет ли добавление азотных и фосфорных удобрений на беспозвоночных?

Небольшие искуственные субстраты экспонировали в течение разного времени в верхней части сублиторали (Hall et al., 2000).

Зависимая переменная:

- `richness` — Число видов

Факторы:

- `time` — срок экспозиции (2, 4 и 6 месяцев)
- `treat` — удобрения (добавляли или нет)

Планировали сделать 5 повторностей для каждого сочетания факторов

Данные из Quinn, Keough, 2002

Знакомимся с данными

```
fert <- read.csv(file="data/hall.csv")
str(fert)
```

```
# 'data.frame': 29 obs. of  3 variables:
#  $ TREAT   : Factor w/ 2 levels "control","nutrient": 1 1 1 1 1 1 1 1 1 1 .
#  $ TIME    : int  2 2 2 2 2 4 4 4 4 4 ...
#  $ RICHNESS: int  5 7 5 7 5 20 18 20 18 17 ...
```

```
# Для удобства названия переменных маленькими буквами
colnames(fert) <- tolower(colnames(fert))
# Время делаем фактором
fert$time <- factor(fert$time)
levels(fert$time)
```

```
# [1] "2" "4" "6"
```

Пропущенные значения

```
sum(is.na(fert))
```

```
# [1] 0
```

```
sapply(fert, function(x)sum(is.na(x)))
```

```
#    treat     time richness
#        0        0        0
```

- Нет пропущенных значений

Объемы выборок в группах

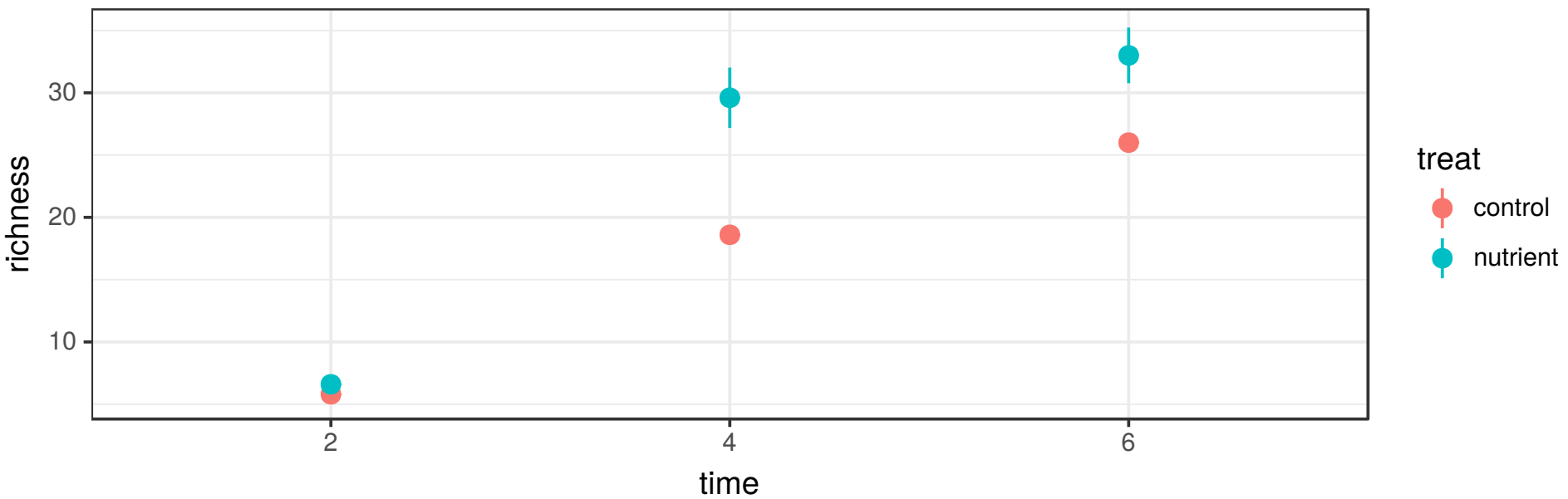
```
table(fert$time, fert$treat)
```

```
#
#      control nutrient
#   2        5         5
#   4        5         5
#   6        4         5
```

- Группы разного размера

Посмотрим на боксплот

```
library(ggplot2)
theme_set(theme_bw())
gg_rich <- ggplot(data = fert, aes(x = time, y = richness, colour = treat)) +
  stat_summary(geom = "pointrange", fun.data = mean_se)
gg_rich
```



- Вполне возможно, здесь есть гетерогенность дисперсий.

Преобразовываем данные

Зависимая переменная `richness` – это счетная величина. Она подчиняется распределению Пуассона (и чем больше ее среднее значение, тем больше дисперсия).

Правильно было бы воспользоваться обобщенными линейными моделями с Пуассоновским распределением ошибок вместо нормального.

Но пока что мы попробуем преобразовать зависимую переменную, чтобы ее распределение стало больше походить на нормальное. Это может помочь, а может и нет.

```
fert$log_rich <- log10(fert$richness + 1)
```

Линейные модели с разным числом дискретных предикторов и взаимодействиями

Линейные модели с разным числом дискретных предикторов

- 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

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

Линейные модели с разным числом дискретных предикторов

- Два фактора A и B, двухфакторное взаимодействие

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

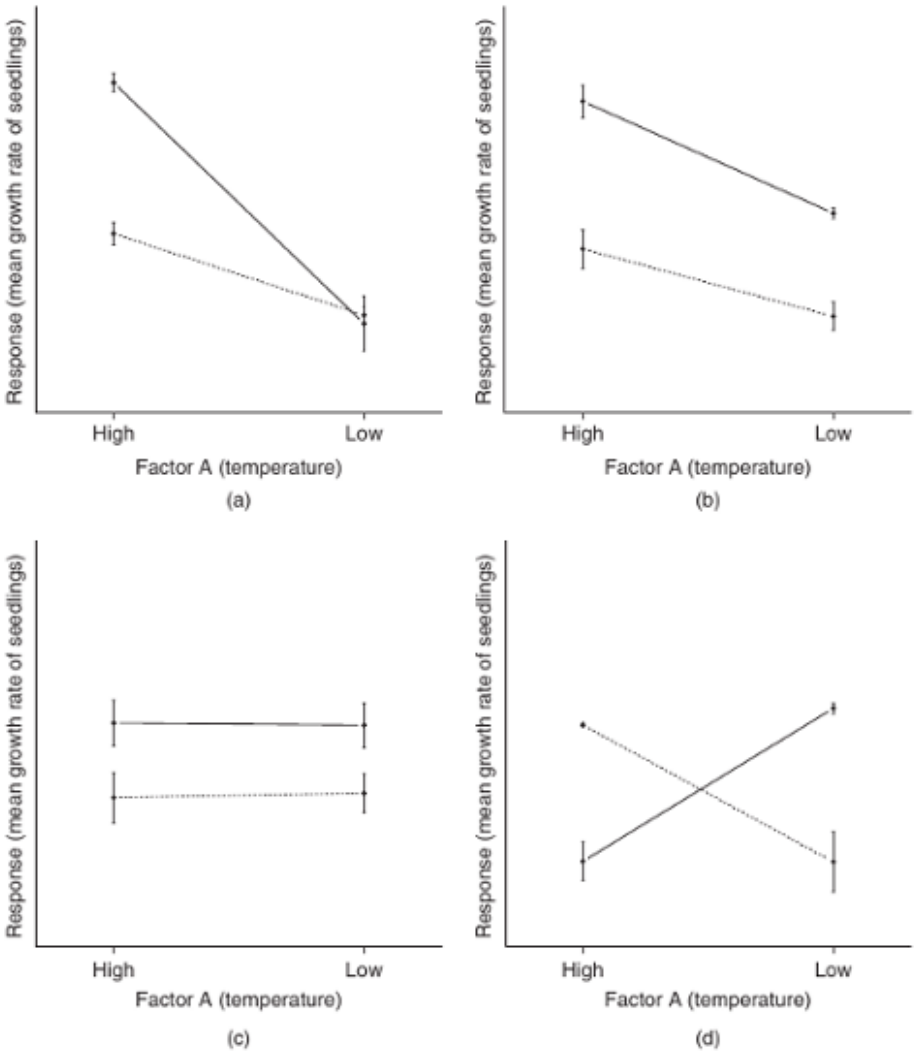
- Три фактора A, B и C, двухфакторные взаимодействия, трехфакторное взаимодействия

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

α, β и γ — это группы коэффициентов, кодирующие соответствующие факторы.

Что такое взаимодействие дискретных предикторов

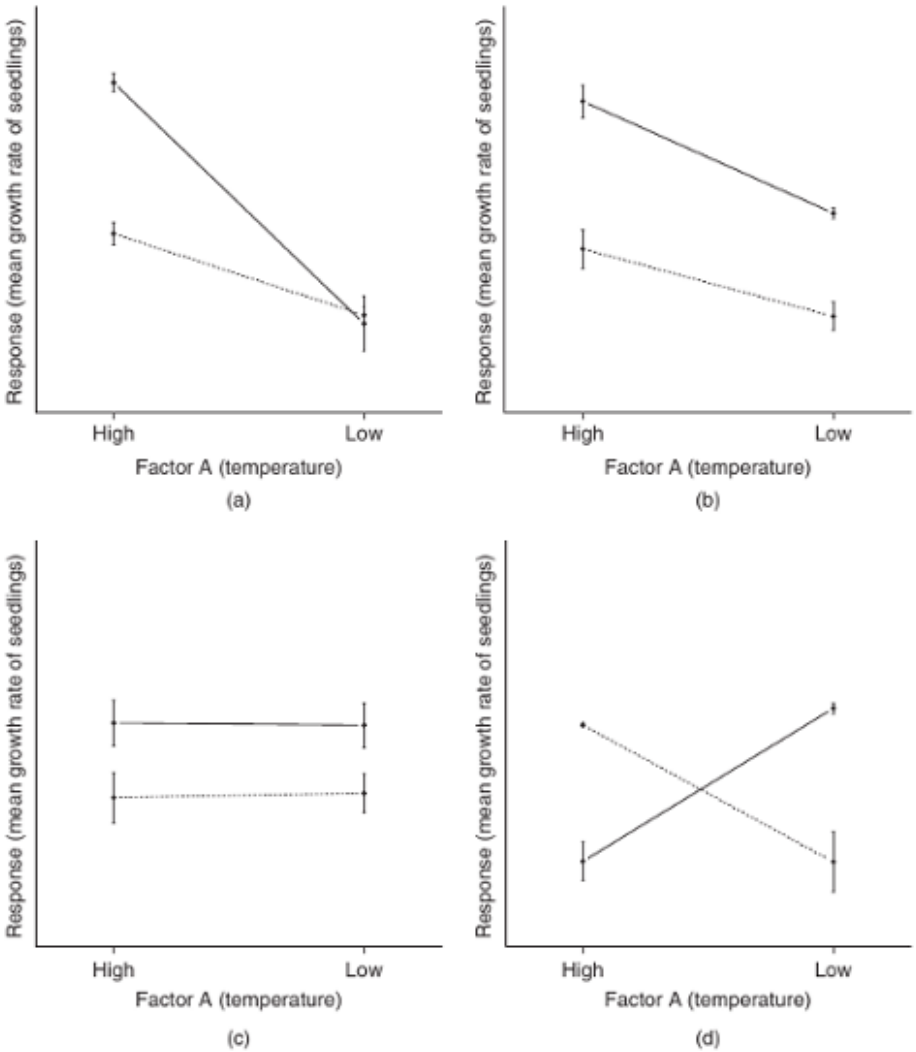
Взаимодействие факторов - когда эффект фактора B разный в зависимости от уровней фактора A и наоборот



На каких рисунках есть взаимодействие факторов?

Что такое взаимодействие дискретных предикторов

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



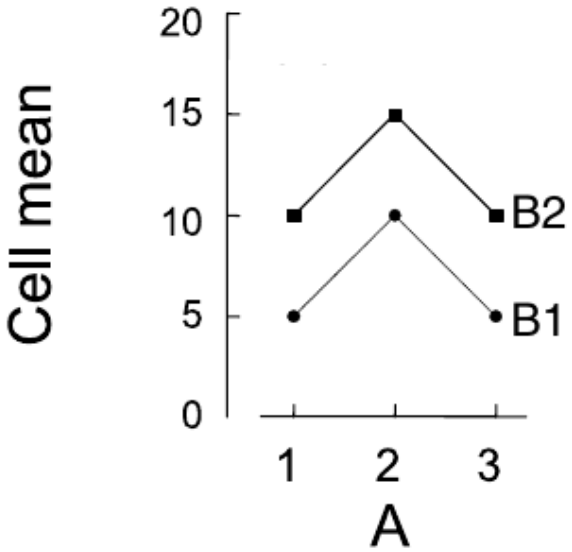
На каких рисунках есть взаимодействие факторов?

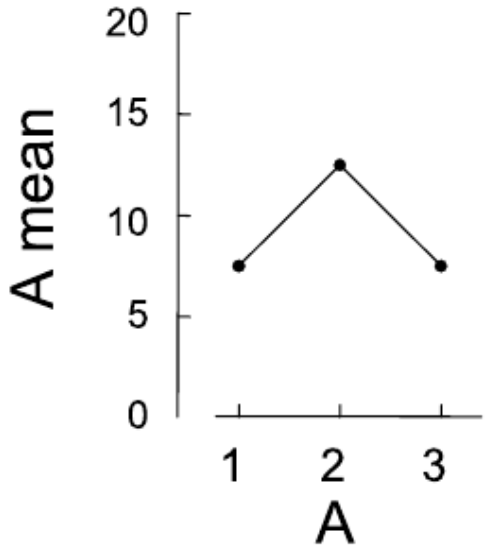
- b, c - нет взаимодействия (эффект фактора B одинаковый для групп по фактору A, линии для разных групп по фактору B на графиках расположены параллельно)
- a, d - есть взаимодействие (эффект фактора B разный для групп по фактору A, на графиках линии для разных групп по фактору B расположены под наклоном).

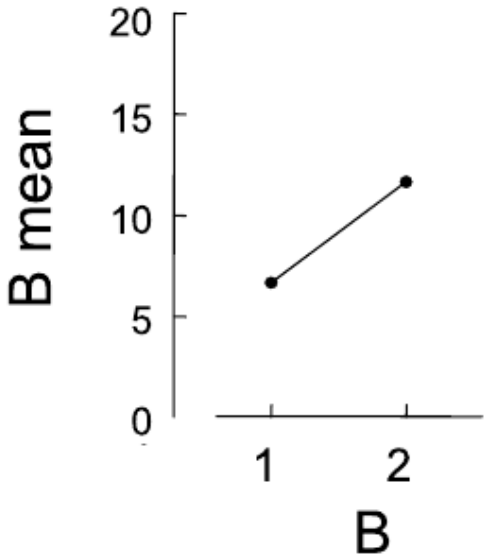
Рисунок из Logan, 2010, fig.12.2

Влияют ли главные эффекты и взаимодействие?

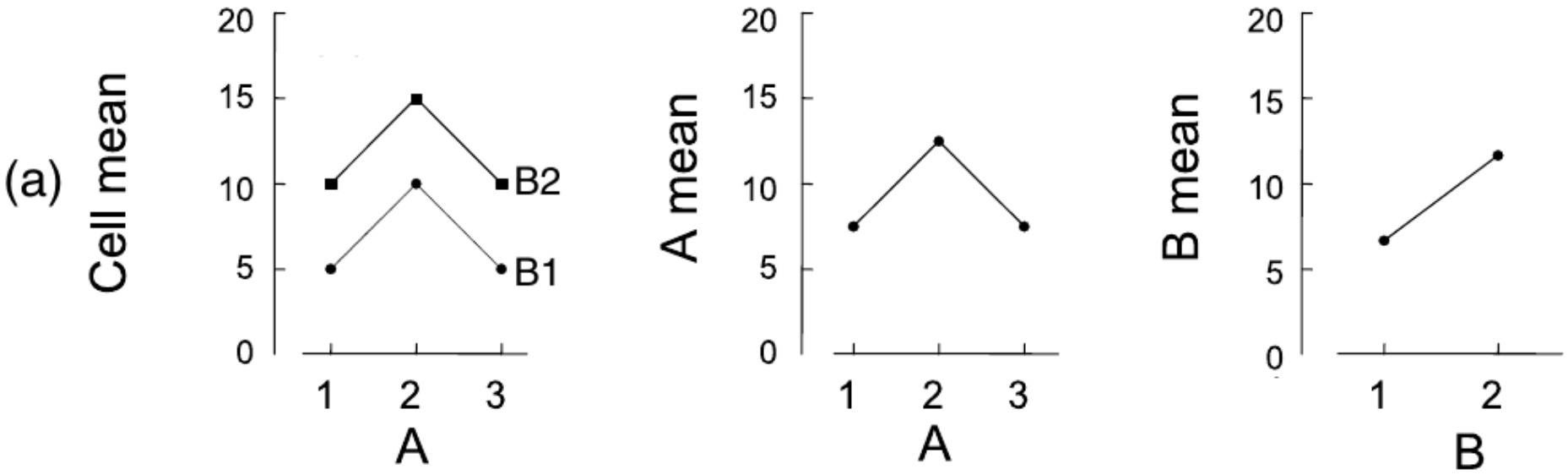
(a)







Влияют ли главные эффекты и взаимодействие?

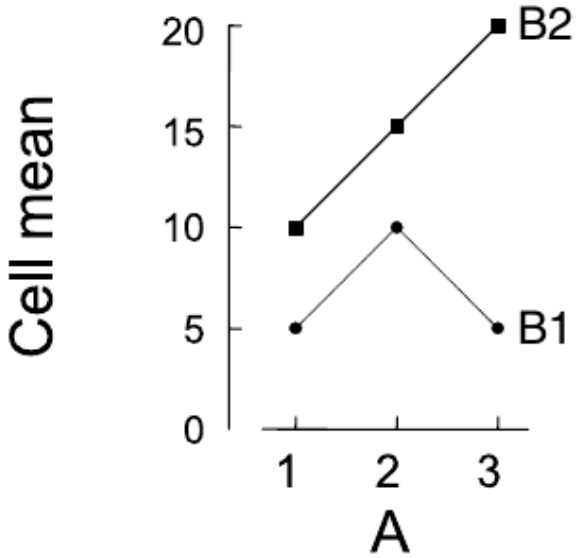


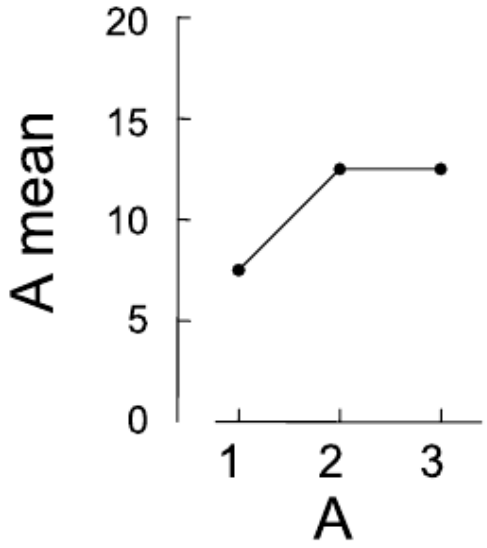
- взаимодействие недостоверно
- фактор A влияет
- фактор B влияет

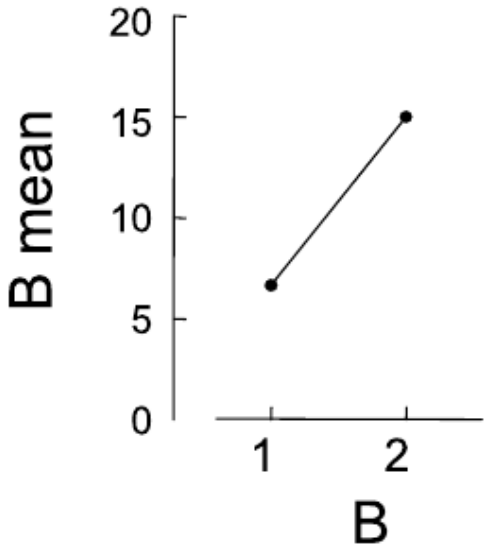
Рисунок из Quinn, Keough, 2002, fig.9.3

Влияют ли главные эффекты и взаимодействие?

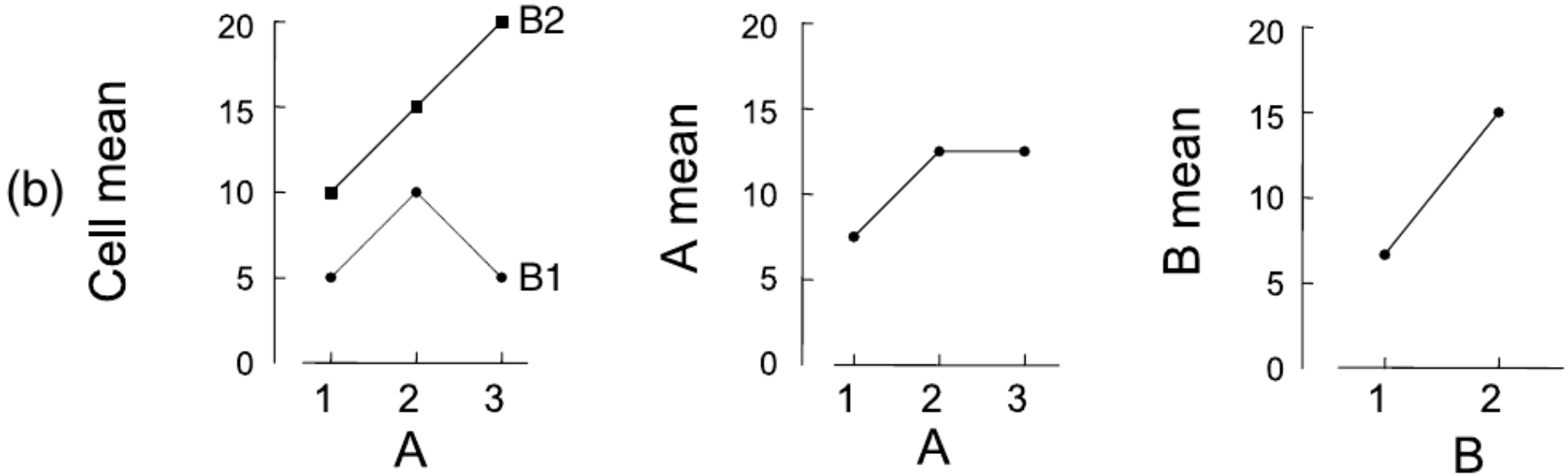
(b)







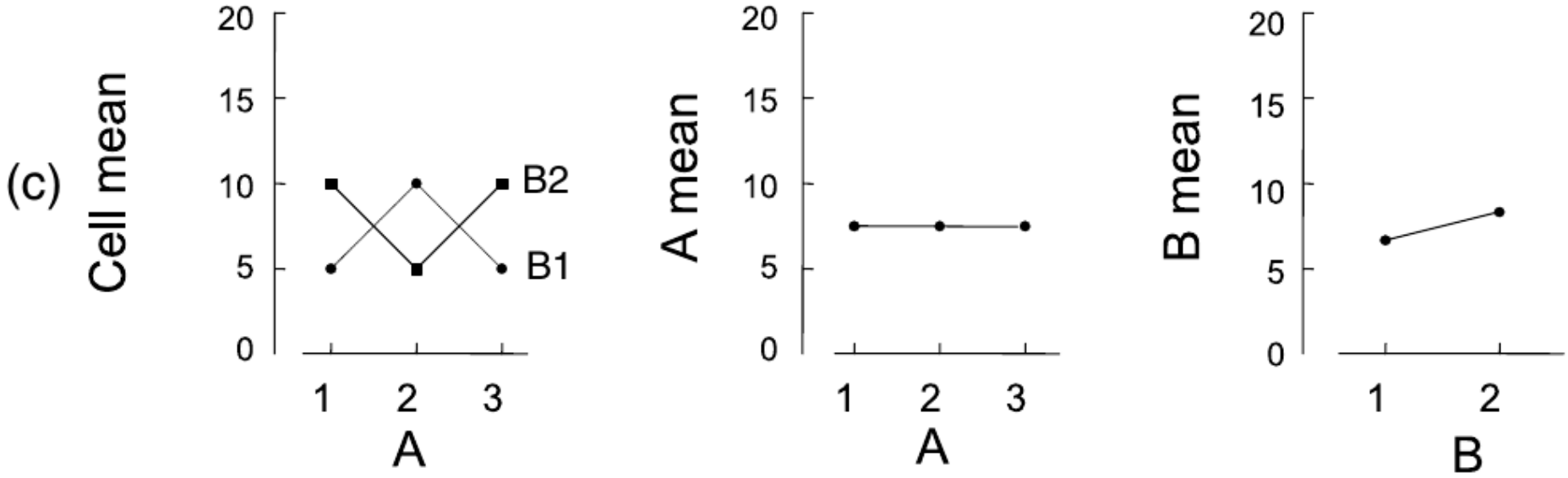
Влияют ли главные эффекты и взаимодействие?



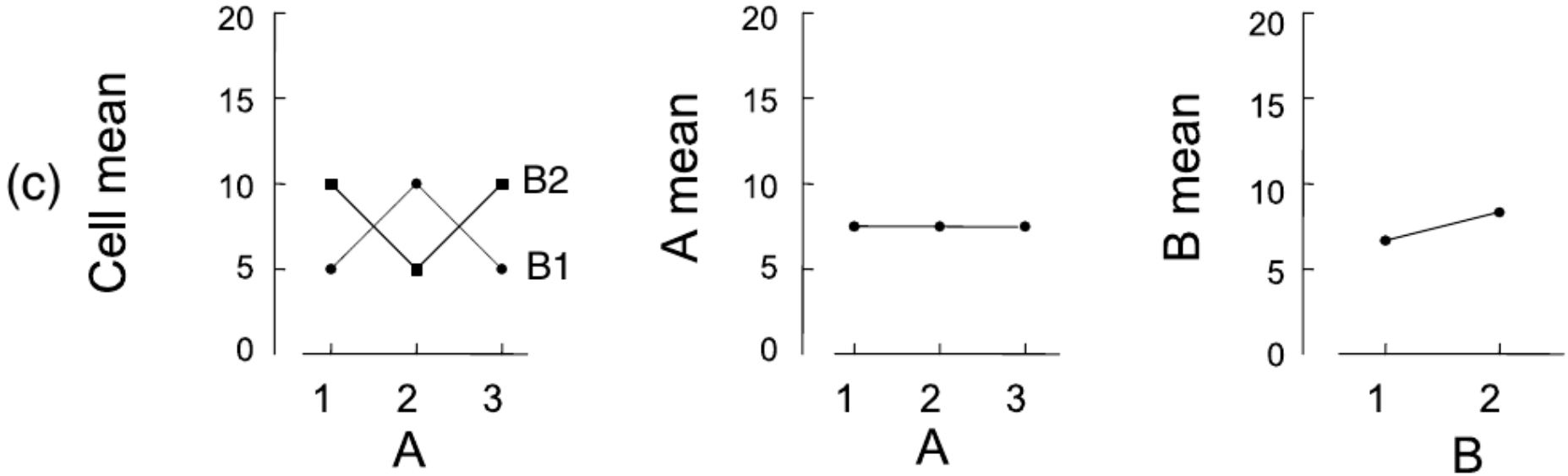
- взаимодействие достоверно и мешает интерпретировать влияние факторов отдельно:
  - для B2 зависимая переменная возрастает с изменением уровня A
  - для B1 зависимая переменная возрастает только на A2, но не различается на A1 и A3
- если смотреть на главные эффекты, можно сделать неправильные выводы:
  - фактор A влияет, группы A2 и A3 не отличаются
  - фактор B влияет, в группе B2 зависимая переменная больше, чем в B1

Рисунок из Quinn, Keough, 2002, fig.9.3

Влияют ли главные эффекты и взаимодействие?



Влияют ли главные эффекты и взаимодействие?



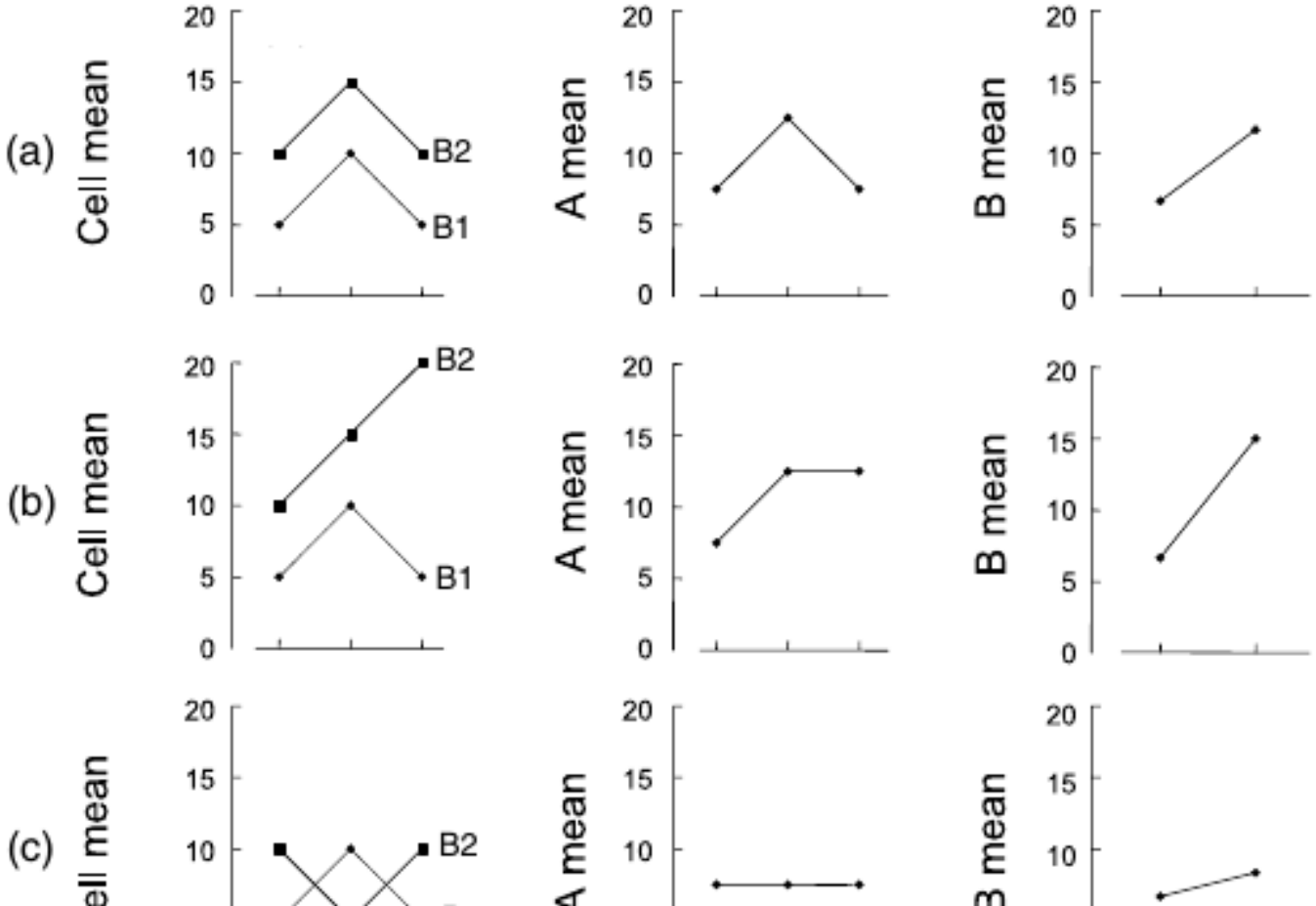
- взаимодействие достоверно и мешает интерпретировать влияние факторов отдельно:
  - на уровне A2 меняется порядок различий уровней фактора B
- если смотреть на главные эффекты, можно сделать неправильные выводы:
  - факторы A и B не влияют

Рисунок из Quinn, Keough, 2002, fig.9.3

Взаимодействие факторов может маскировать главные эффекты

Если есть значимое взаимодействие

- главные эффекты обсуждать не имеет смысла
- пост хок тесты проводятся только для ваимодействия



Двухфакторный дисперсионный анализ в матричном виде

Уравнение линейной модели в параметризации фиктивных переменных (contr.treatment)

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_5 x_{5j} + \varepsilon_j$$

В этом примере отклик — видовое богатство, и два дискретных фактора: `treat` (2 уровня, базовый `control`), `time` (3 уровня, базовый 2). Для их кодирования потребуются переменные-индикаторы: $x_{1j}, \ldots, x_{5j}$ ($j = 1, \ldots, n$ — порядковый номер наблюдения).

Коэффициенты:

- β_0 — значение отклика в контроле при экспозиции 2 (на базовом уровне обоих факторов)
- β_1 — изменение отклика при добавлении удобрений при экспозиции 2 (эффект фактора `treat`)
- β_2 и β_3 — изменение отклика в контроле при экспозициях 4 и 6 соответственно (эффект фактора `time`)
- β_4 и β_5 — изменение отклика при добавлении удобрений при экспозициях 4 и 6 соответственно (эффект взаимодействия факторов)

Остальное все так же как в предыдущем примере с однофакторным анализом.

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

В параметризации фиктивных переменных

Уровни фактора “удобрения” закодированы с помощью одной переменной-индикатора:

```
contr.treatment(levels(fert$treat))
```

```
#          nutrient
# control         0
# nutrient        1
```

Уровни фактора “время экспозиции” — с помощью двух переменных-индикаторов:

```
contr.treatment(levels(fert$time))
```

```
#   4 6
# 2 0 0
# 4 1 0
# 6 0 1
```

Модельная матрица целиком

```
X_trt <- model.matrix(~ treat*time, data = fert)
X_trt
```

Для параметризации эффектов (contr.sum) дискретные предикторы кодируются иначе

В модели эффектов (contr.sum) переменные-болванки будут закодированы при помощи -1, 0 и 1, так, чтобы сумма кодов для возможных состояний одной переменной была равна нулю.

Уровни фактора “удобрения” закодированы с помощью одной переменной:

```
contr.sum(levels(fert$treat))
```

```
#          [,1]
# control     1
# nutrient   -1
```

Уровни фактора “время экспозиции” — с помощью двух переменных:

```
contr.sum(levels(fert$time))
```

```
#   [,1] [,2]
# 2    1    0
# 4    0    1
# 6   -1   -1
```

Модельная матрица целиком

```
X_sum <- model.matrix(~ treat*time, data = fert,
  contrasts = list(treat = "contr.sum", time = "contr.sum"))
```

Двухфакторный дисперсионный анализ в матричном виде (contr.sum)

Уравнение линейной модели для этого примера (в параметризации эффектов, contr.sum):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i} + \varepsilon_i$$

$x_{1i}, \ldots, x_{5i}$ — переменные-индикаторы ($i = 1, \ldots, n$ — порядковый номер наблюдения)

- β_0 - среднее значение отклика по всем данным
- β_1 - отклонение значений отклика для тритментов от общего среднего (фактор `treat`)
- β_2, β_3 - отклонение отклика при разной экспозиции от общего среднего (фактор `time`)
- β_4, β_5 - отклонение отклика в тритментах при разных экспозициях от общего среднего (взаимодействие)

Несбалансированные данные, типы сумм квадратов

Несбалансированные данные - когда численности в группах по факторам различаются

Например так,

	A1	A2	A3
B1	5	5	5
B2	5	4	5

или так,

	A1	A2	A3
B1	3	8	4
B2	4	7	4

Проблемы из-за несбалансированности данных

- 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 расчетом мощности. Если $\sigma^2_\epsilon > 0$ и размеры выборок разные, то $\frac{MS_x}{MS_e}$ не следует F-распределению (Searle et al. 1992).

Проблемы из-за несбалансированности данных

- 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 расчетом мощности. Если $\sigma^2_\epsilon > 0$ и размеры выборок разные, то $\frac{MS_x}{MS_e}$ 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 при нарушении условий применимости только, если значения доверительной вероятности *p* близки к выбранному критическому уровню значимости α

Суммы квадратов в многофакторном дисперсионном анализе со взаимодействием

Если данные сбалансированы, то …

взаимодействие и эффекты факторов независимы, их суммы квадратов и соответствующие тесты можно посчитать в одном анализе и его результат не будет зависеть от того, в каком порядке мы рассматриваем факторы.

Если данные несбалансированы, то …

взаимодействие и эффекты факторов уже не являются полностью независимыми, суммы квадратов для факторов не равны общей сумме квадратов. Если делать все как обычно, результат анализа будет зависеть от порядка включения факторов в модель. (Для вычислений используется регрессионный подход к дисперсионному анализу)

Если несбалансированные данные, выберите подходящий порядок тестирования гипотез

- SSe и SSab всегда рассчитываются одинаково, вне зависимости от порядка тестирования гипотез и от сбалансированности данных
- SSa, SSb — есть три способа расчета (суммы квадратов I, II и III типа, терминология пришла из SAS) в зависимости от порядка тестирования значимости факторов

- 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, Glantz, Slinker 2000)

Порядок тестирования гипотез в дисперсионном анализе

”Типы сумм квадратов”	I тип	II тип	III тип
Название	Последовательный	Без учета взаимодействий высоких порядков	Иерархический
Порядок расчета 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()	Anova() (пакет car)	Anova() (пакет car)

Многофакторный дисперсионный анализ в R

Дисперсионный анализ со II типом сумм квадратов

При таком способе, сначала тестируется взаимодействие, затем отдельные факторы в модели без взаимодействия

```
fmod2 <- lm(log_rich ~ treat * time, data = fert)
library(car)
Anova(fmod2, type = "II")
```

```
# Anova Table (Type II tests)
#
# Response: log_rich
#            Sum Sq Df F value   Pr(>F)
# treat       0.091  1   28.42 0.000021 ***
# time        2.216  2  344.88  < 2e-16 ***
# treat:time  0.025  2    3.84    0.036 *
# Residuals   0.074 23
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Дисперсионный анализ с III типом сумм квадратов

Опишем процедуру на тот случай, если вдруг вам понадобится воспроизвести в R дисперсионный анализ с III типом сумм квадратов.

При этом способе вначале тестируют взаимодействие, когда все другие факторы есть в модели. Затем тестируют факторы, когда все другие факторы и взаимодействие есть в модели.

Внимание: при использовании III типа сумм квадратов, нужно обязательно указывать тип контрастов для факторов (`contrasts=list(фактор_1 = contr.sum, фактор_2=contr.sum)`).

```
fmod3 <- lm(log_rich ~ treat * time, data = fert,
            contrasts = list(treat = contr.sum, time = contr.sum))
Anova(fmod3, type = 3)
```

```
# Anova Table (Type III tests)
#
# Response: log_rich
#             Sum Sq Df  F value   Pr(>F)
# (Intercept)   44.2  1 13776.11  < 2e-16 ***
# treat          0.1  1    28.04 0.000022 ***
# time           2.2  2   344.75  < 2e-16 ***
# treat:time     0.0  2     3.84    0.036 *
# Residuals      0.1 23
# ---
```

Пост хок тест для взаимодействия факторов

Пост хок тесты в многофакторном дисперсионном анализе

- Поскольку взаимодействие достоверно, факторы отдельно можно не тестировать. Проведем пост хок тест по взаимодействию, чтобы выяснить, какие именно группы различаются
- Если бы взаимодействие было недостоверно, мы бы провели пост хок тест по тем факторам, влияние которых было бы достоверно. Как? См. предыдущую презентацию.

Пост хок тест для взаимодействия 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

Задание 1

Дополните этот код, чтобы посчитать пост хок тест Тьюки по взаимодействию факторов

```
# Создаем переменную-взаимодействие
fert$treat_time <- interaction(fert$treat, fert$time)
# Подбираем линейную модель от этой переменной без свободного члена
fit_inter <- lm()
# Делаем пост хок тест для этой модели
library(multcomp)
dat_tukey <- glht(, linfct = mcp( = "Tukey"))
summary()
```

Решение

```
# Создаем переменную-взаимодействие
fert$treat_time <- interaction(fert$treat, fert$time)
# Подбираем линейную модель без свободного члена
fit_inter <- lm(log_rich ~ treat_time - 1, data = fert)
# Делаем пост хок тест для этой модели
library(multcomp)
dat_tukey <- glht(fit_inter, linfct = mcp(treat_time = "Tukey"))
summary(dat_tukey)
```

Результаты пост хок теста

В виде таблицы результаты нечитабельны Лучше построить график.

```
#
#    Simultaneous Tests for General Linear Hypotheses
#
# Multiple Comparisons of Means: Tukey Contrasts
#
#
# Fit: lm(formula = log_rich ~ treat_time - 1, data = fert)
#
# Linear Hypotheses:
#                             Estimate Std. Error t value Pr(>|t|)
# nutrient.2 - control.2 == 0    0.0504     0.0358    1.41    0.723
# control.4 - control.2 == 0     0.4633     0.0358   12.93   <0.001 ***
# nutrient.4 - control.2 == 0    0.6519     0.0358   18.19   <0.001 ***
# control.6 - control.2 == 0     0.6031     0.0380   15.86   <0.001 ***
# nutrient.6 - control.2 == 0    0.6997     0.0358   19.52   <0.001 ***
# control.4 - nutrient.2 == 0    0.4129     0.0358   11.52   <0.001 ***
# nutrient.4 - nutrient.2 == 0   0.6015     0.0358   16.78   <0.001 ***
# control.6 - nutrient.2 == 0    0.5527     0.0380   14.54   <0.001 ***
# nutrient.6 - nutrient.2 == 0   0.6493     0.0358   18.11   <0.001 ***
# nutrient.4 - control.4 == 0    0.1885     0.0358    5.26   <0.001 ***
# control.6 - control.4 == 0     0.1398     0.0380    3.68    0.014 *
# nutrient.6 - control.4 == 0    0.2364     0.0358    6.59   <0.001 ***
# control.6 - nutrient.4 == 0   -0.0488     0.0380   -1.28    0.791
# nutrient.6 - nutrient.4 == 0   0.0478     0.0358    1.33    0.763
# nutrient.6 - control.6 == 0    0.0966     0.0380    2.54    0.153
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
# (Adjusted p values reported -- single-step method)
```

Данные для графика при помощи predict()

У нас два дискретных фактора, поэтому вначале используем expand.grid()

```
MyData <- expand.grid(treat = levels(fert$treat),
                      time = levels(fert$time))
MyData <- data.frame(
  MyData,
  predict(fmod2, newdata = MyData, interval = "confidence")
  )
# Обратная трансформация (не забываем про единичку, которую прибавляли)
MyData$richness <- 10^MyData$fit - 1
MyData$LWR <- 10^MyData$lwr - 1
MyData$UPR <- 10^MyData$upr - 1
MyData
```

```
#      treat time   fit   lwr   upr richness   LWR   UPR
# 1  control    2 0.828 0.776 0.881     5.73  4.97  6.60
# 2 nutrient    2 0.879 0.826 0.931     6.56  5.70  7.53
# 3  control    4 1.291 1.239 1.344    18.56 16.34 21.07
# 4 nutrient    4 1.480 1.428 1.532    29.20 25.76 33.07
# 5  control    6 1.431 1.373 1.490    25.99 22.58 29.89
# 6 nutrient    6 1.528 1.475 1.580    32.71 28.88 37.04
```

Задание 2

Создайте MyData вручную для модели в обычной параметризации:

- предсказанные значения
- стандартные ошибки
- верхнюю и нижнюю границы доверительных интервалов

```
MyData <- expand.grid(treat = levels(fert$treat),
                      time = levels())
X <- model.matrix(~ , data = )
betas <- coef()
MyData$fit <-
MyData$se <-   (X %*% vcov(fmod2) %*% t(X))
MyData$lwr <- MyData$ - 1.96 *
MyData$upr <- MyData$ + 1.96 *

# Обратная трансформация
MyData$richness <-
MyData$LWR <-
MyData$UPR <-
MyData
```

```
#       treat time   fit     se   lwr   upr richness   LWR   UPR
# 1   control    2 0.828 0.0253 0.778 0.878     5.73  5.00  6.55
# 2 nutrient    2 0.879 0.0253 0.829 0.928     6.56  5.74  7.48
```

Решение:

```
MyData <- expand.grid(treat = levels(fert$treat),
                      time = levels(fert$time))
X <- model.matrix(~ treat * time, data = MyData)
betas <- coef(fmod2)
MyData$fit <- X %*% betas
MyData$se <- sqrt(diag(X %*% vcov(fmod2) %*% t(X)))
MyData$lwr <- MyData$fit - 1.96 * MyData$se
MyData$upr <- MyData$fit + 1.96 * MyData$se
# Обратная трансформация
MyData$richness <- 10^MyData$fit - 1
MyData$LWR <- 10^MyData$lwr - 1
MyData$UPR <- 10^MyData$upr - 1
MyData
```

```
#       treat time   fit     se   lwr   upr richness   LWR   UPR
# 1   control    2 0.828 0.0253 0.778 0.878     5.73  5.00  6.55
# 2 nutrient    2 0.879 0.0253 0.829 0.928     6.56  5.74  7.48
# 3   control    4 1.291 0.0253 1.242 1.341    18.56 16.45 20.93
# 4 nutrient    4 1.480 0.0253 1.430 1.530    29.20 25.93 32.86
# 5   control    6 1.431 0.0283 1.376 1.487    25.99 22.75 29.67
# 6 nutrient    6 1.528 0.0253 1.478 1.577    32.71 29.07 36.80
```

Задание 3

Постройте график результатов, на котором будут изображены предсказанные средние значения видового богатства в зависимости от тритмента и времени экспозиции.

```
pos <- position_dodge(width = 0.2)
gg_linep <- ggplot(data = , aes()) +
  geom_  (position = pos) +
  geom_  (aes(group = ), position = pos) +
  geom_  (position = pos, width = 0.1)
gg_linep
```

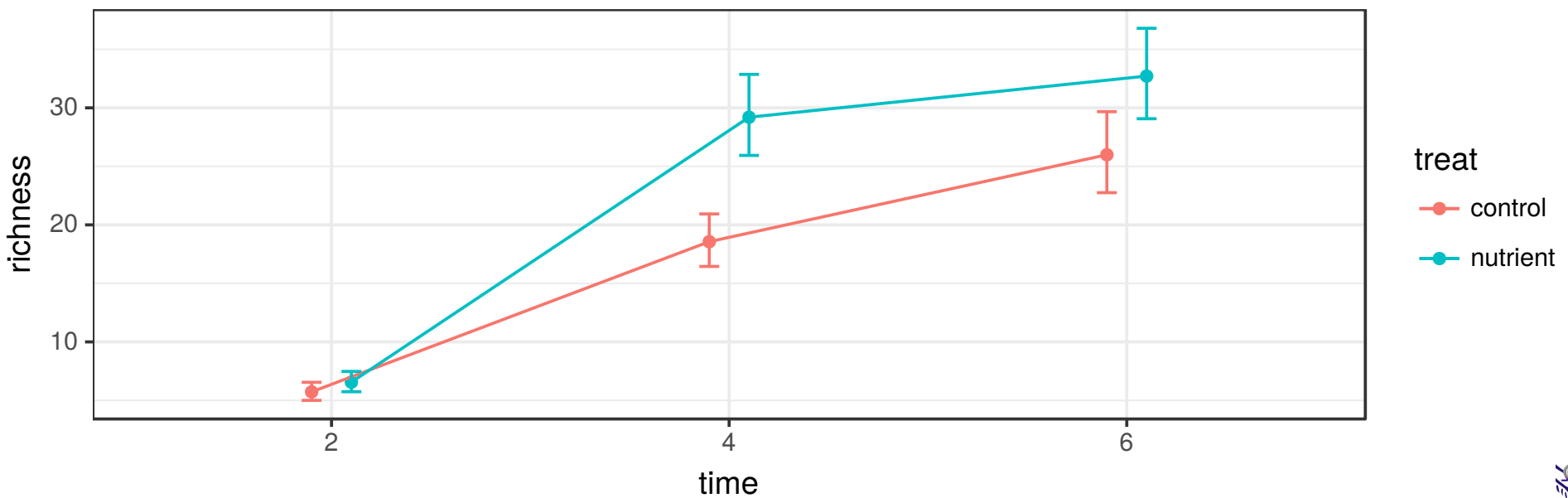
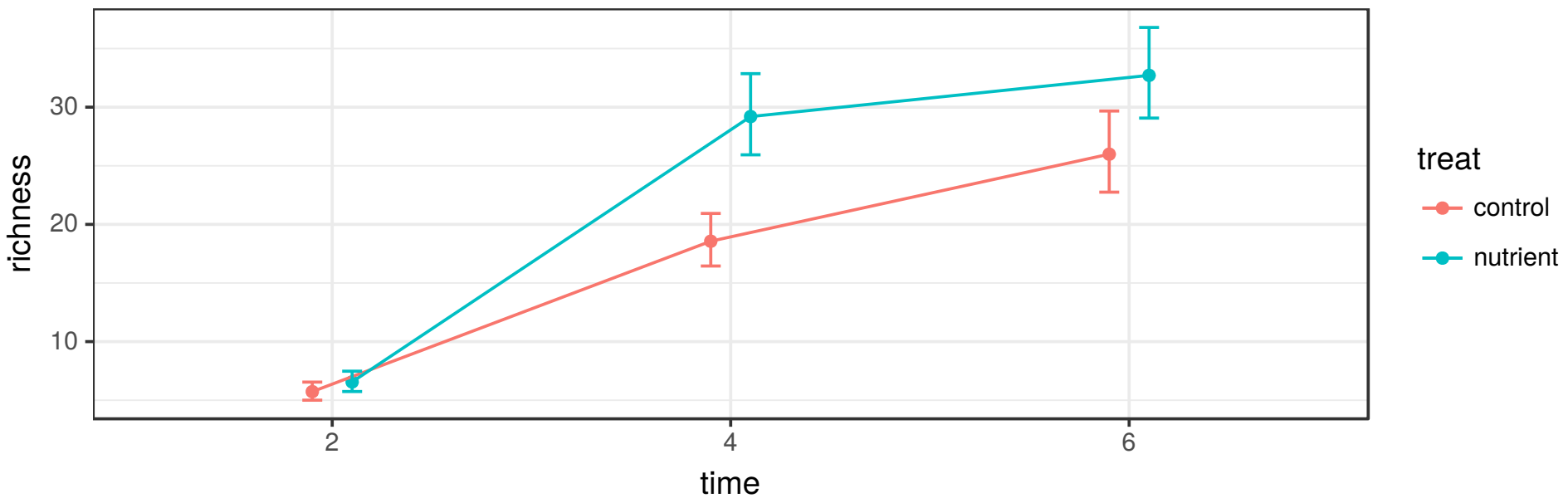


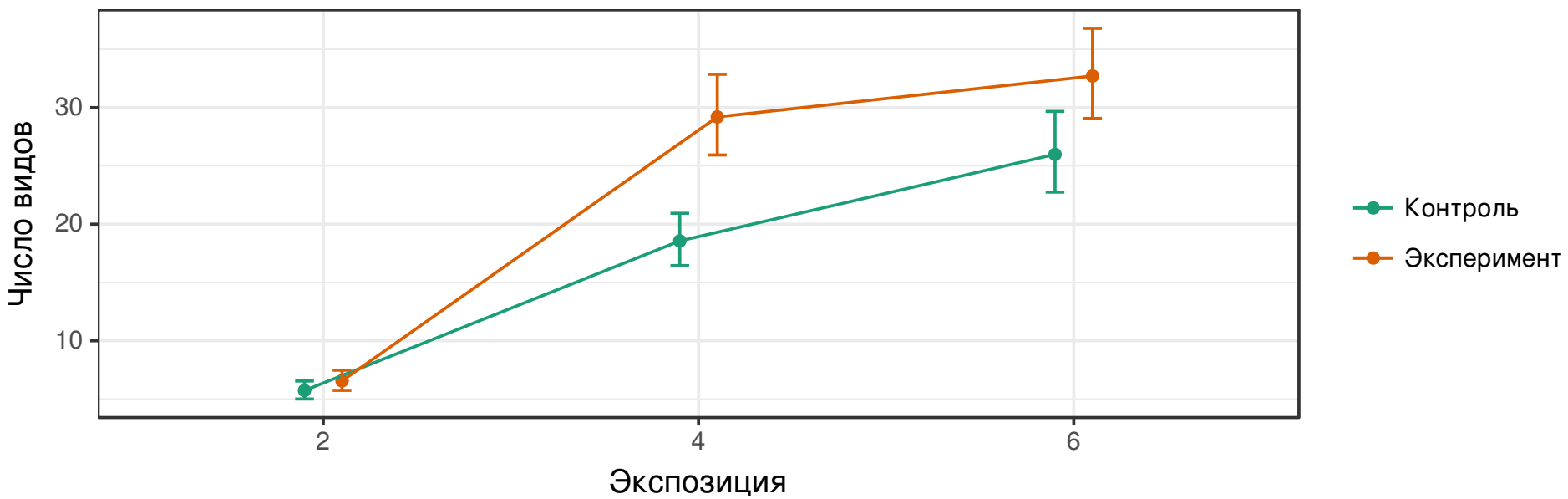
График результатов: Линии с точками

```
pos <- position_dodge(width = 0.2)
gg_linep <- ggplot(data = MyData, aes(x = time, y = richness,
                ymin = LWR,  ymax = UPR, colour = treat)) +
  geom_point(position = pos) +
  geom_line(aes(group = treat), position = pos) +
  geom_errorbar(position = pos, width = 0.1)
gg_linep
```



Приводим график в приличный вид

```
gg_final <- gg_linep + labs(x = "Экспозиция",  y = "Число видов") +
  scale_colour_brewer(name = "", palette = "Dark2",
    labels = c("Контроль", "Эксперимент"))
gg_final
```



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 вне зависимости от порядка тестирования значимости факторов
- В случае, если численности групп неравны (несбалансированные данные), есть несколько способов тестирования значимости 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