Линейные модели с дискретными предикторами

Линейные модели...

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Линейные модели с дискретными предикторами (дисперсионный анализ)

Вы сможете

- Объяснить, в чем опасность множественных сравнений, и как с ними можно бороться
- Рассказать, как в дисперсионном анализе моделируются значения зависимой переменной
- Интерпретировать и описать результаты, записанные в таблице дисперсионного анализа
- Перечислить и проверить условия применимости дисперсионного анализа
- Провести множественные попарные сравнения при помощи post hoc теста Тьюки, представить и описать их результаты
- Построить график результатов дисперсионного анализа

Множественные сравнения

Пример: яйца кукушек

- `species` — вид птиц-хозяев (фактор)
- `length` — длина яиц кукушек в гнездах хозяев (зависимая переменная)

```
library(DAAG)
data("cuckoos")
# Положим данные в переменную с коротким названием, чтобы меньше печатать
cu <- cuckoos
head(cu, 3)
```

```
#   length breadth      species id
# 1   21.7    16.1 meadow.pipit 21
# 2   22.6    17.0 meadow.pipit 22
# 3   20.9    16.2 meadow.pipit 23
```

Данные: Latter, 1902; источник: Tippett, 1931

Исследуем данные

```
# Пропущенных значений нет
colSums(is.na(cu))
```

```
#  length breadth species      id
#       0       0       0       0
```

```
# Данные не сбалансированы (размеры групп разные)
table(cu$species)
```

```
#
# hedge.sparrow  meadow.pipit  pied.wagtail         robin    tree.pipit
#            14            45            15            16            15
#          wren
#            15
```

Изменим названия уровней фактора, чтобы было легче понять о каких птицах речь

```
levels(cu$species)
```

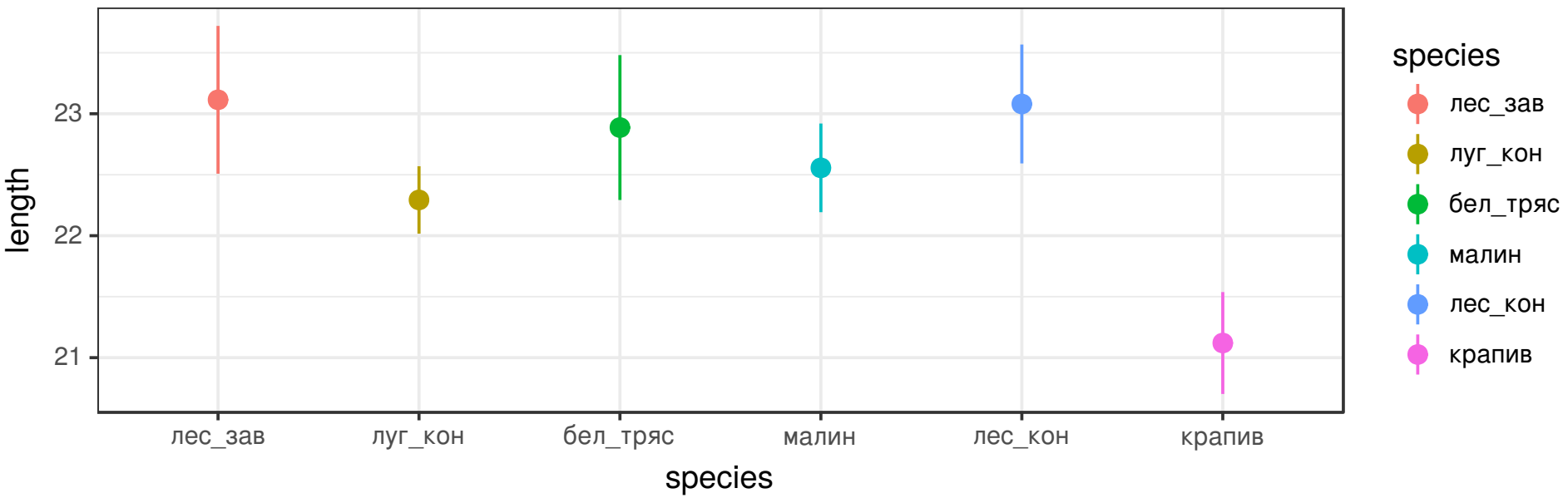
```
# [1] ”hedge.sparrow” ”meadow.pipit”  ”pied.wagtail”  ”robin”
# [5] ”tree.pipit”    ”wren”
```

```
levels(cu$species) <- c(”лес_зав”, ”луг_кон”, ”бел_тряс”,
                        ”малин”, ”лес_кон”, ”крапив”)
```

Задание 1

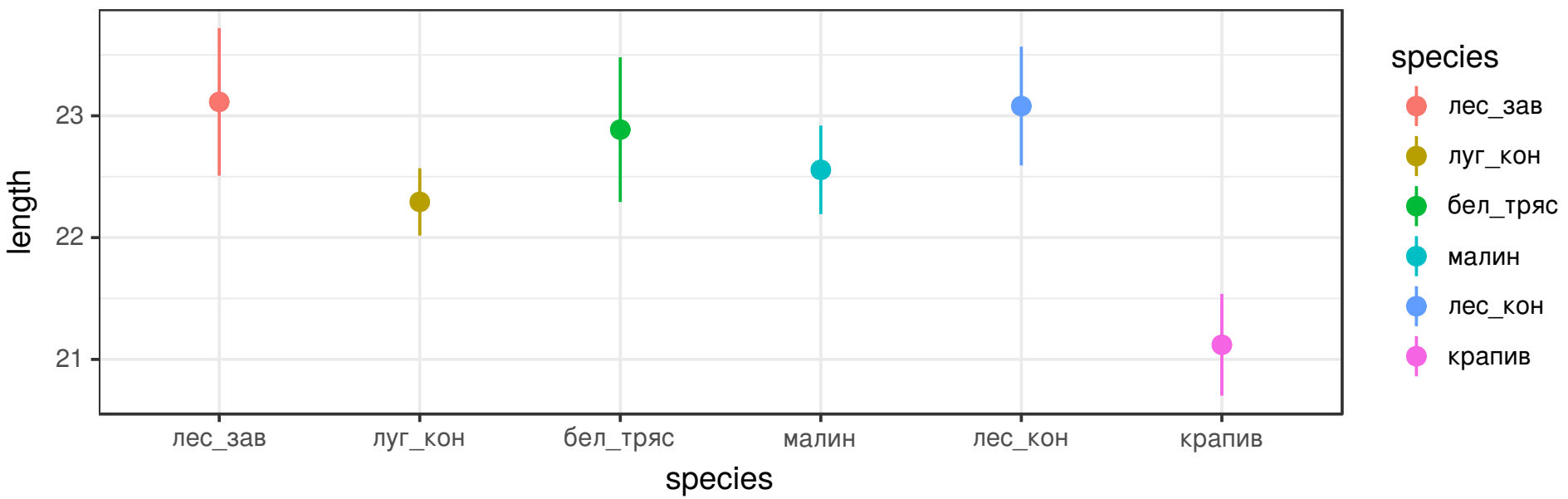
Дополните код, чтобы построить график зависимости размера яиц кукушек (`length`) от вида птиц-хозяев (`species`), в гнездах которых были обнаружены яйца. На графике должны быть изображены средние значения и их 95% доверительные интервалы, а цвет должен соответствовать виду птиц-хозяев.

```
theme_set( )
ggplot(data = , aes()) +
  stat_summary(geom = "pointrange", fun.data = mean_cl_normal)
```



Решение

```
library(ggplot2)
theme_set(theme_bw())
ggplot(data = cu, aes(x = species, y = length, colour = species)) +
  stat_summary(geom = "pointrange", fun.data = mean_cl_normal)
```



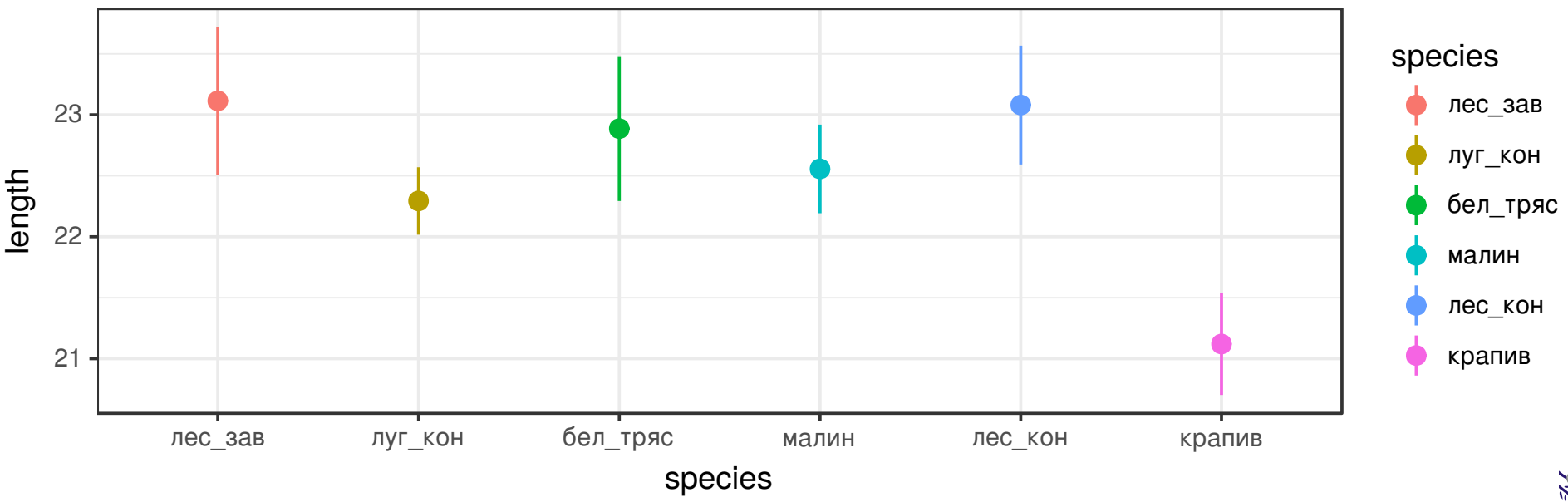
“Некрасивый” порядок уровней на графике

На этом графике некрасивый порядок уровней: средние для разных способов запоминания `cu$species` расположены, как кажется, хаотично.

Порядок групп на графике определяется порядком уровней фактора

```
# ”старый” порядок уровней
levels(cu$species)
```

```
# [1] ”лес_зав”  ”луг_кон”  ”бел_тряс” ”малин”    ”лес_кон”  ”крапив”
```



Меняем порядок уровней

Давайте изменим порядок уровней в факторе `cu$species` так, чтобы он соответствовал возрастанию средних значений длины яиц `cu$length`.

```
# "старый" порядок уровней
levels(cu$species)
```

```
# [1] "лес_зав"  "луг_кон"  "бел_тряс" "малин"    "лес_кон"  "крапив"
```

```
# переставляем уровни в порядке следования средних значений
cu$species <- reorder(cu$species, cu$length, FUN = mean)
# "новый" порядок уровней стал таким
levels(cu$species)
```

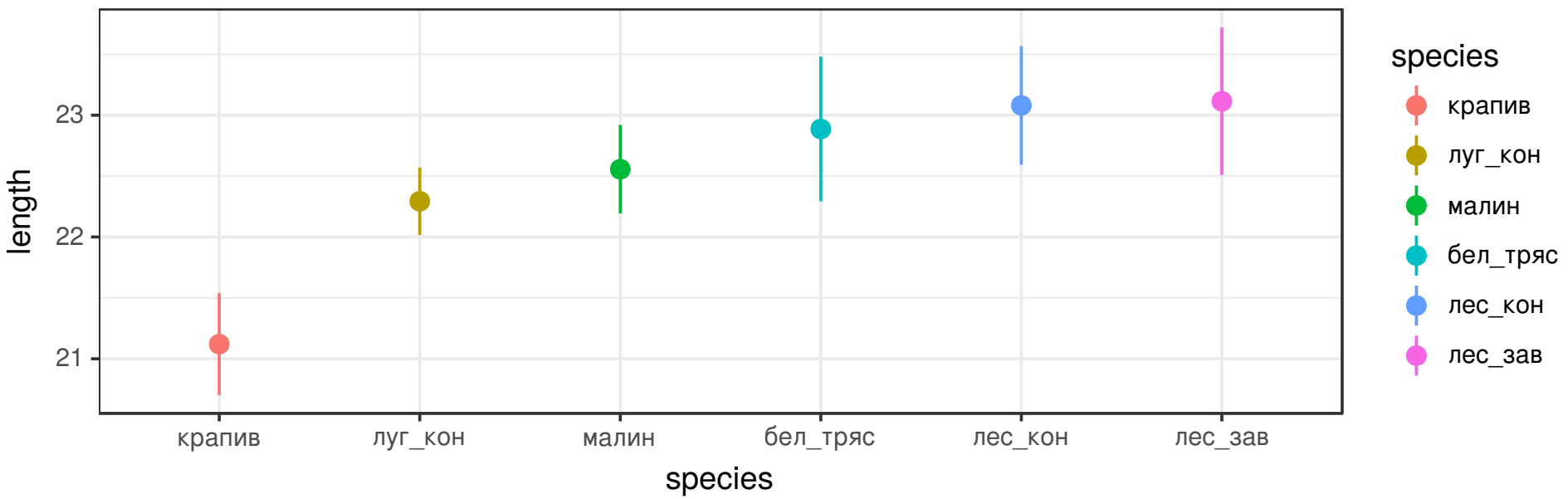
```
# [1] "крапив"   "луг_кон"  "малин"    "бел_тряс" "лес_кон"  "лес_зав"
```

График с новым порядком уровней

С новым порядком уровней нам легче визуально сравнивать друг с другом число запомненных слов при разных способах запоминания.

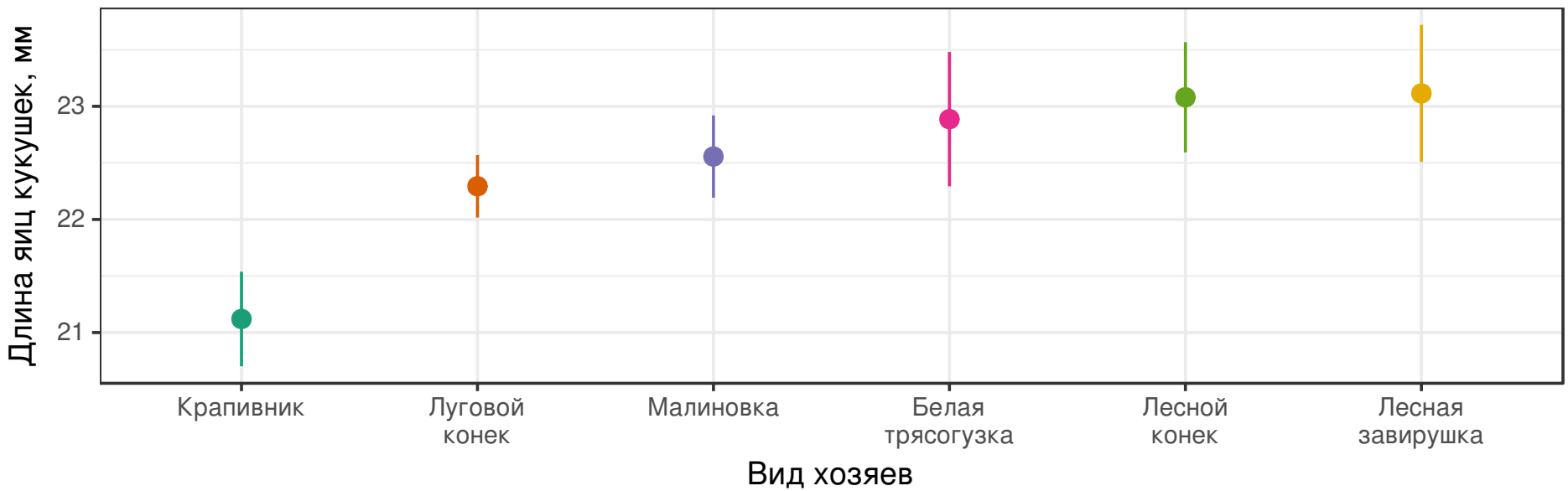
Поскольку, изменив порядок уровней, мы внесли изменения в исходные данные, придется полностью обновить график (т.к.`ggplot()` хранит данные внутри графика).

```
ggplot(data = cu, aes(x = species, y = length, colour = species)) +
  stat_summary(geom = "pointrange", fun.data = mean_cl_normal)
```



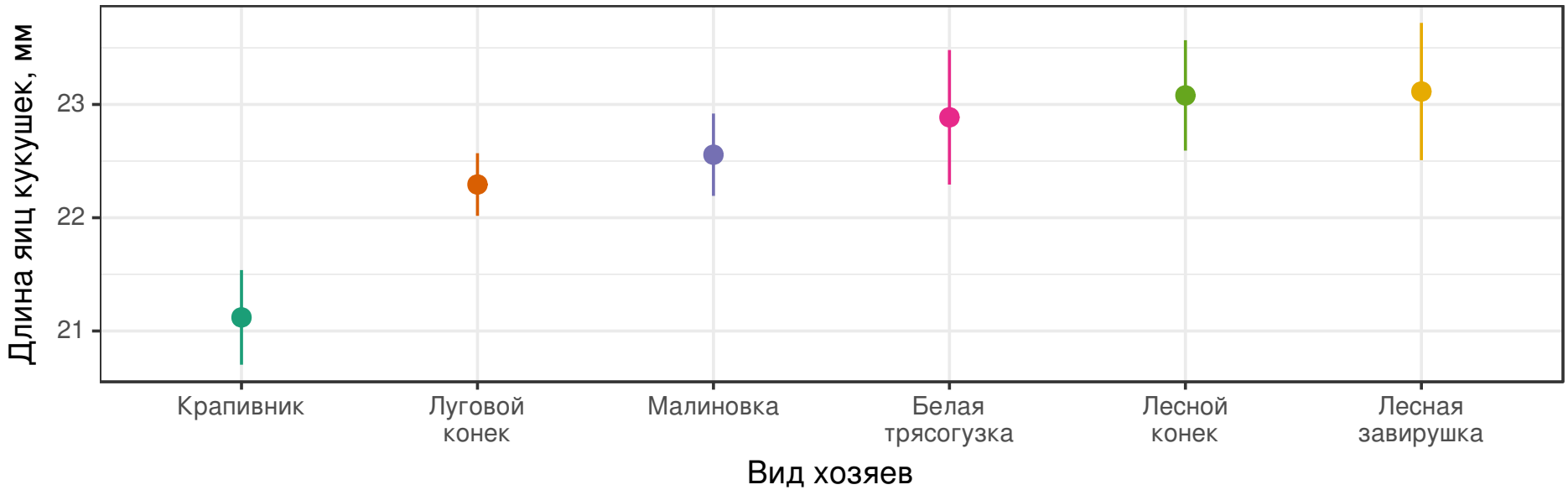
Понравившийся график, если понадобится, можно в любой момент довести до ума, а остальные удалить

```
ggplot(data = cu, aes(x = species, y = length, colour = species)) +
  stat_summary(geom = "pointrange", fun.data = mean_cl_normal) +
  labs(x = "Вид хозяев", y = "Длина яиц кукушек, мм") +
  scale_colour_brewer(name = "Вид \nхозяев", palette = "Dark2") +
  scale_x_discrete(labels = c("Крапивник", "Луговой\nконек", "Малиновка",
"Белая\nтрясогузка", "Лесной\nконек", "Лесная\nзавирушка")) +
  theme(legend.position = "none")
```



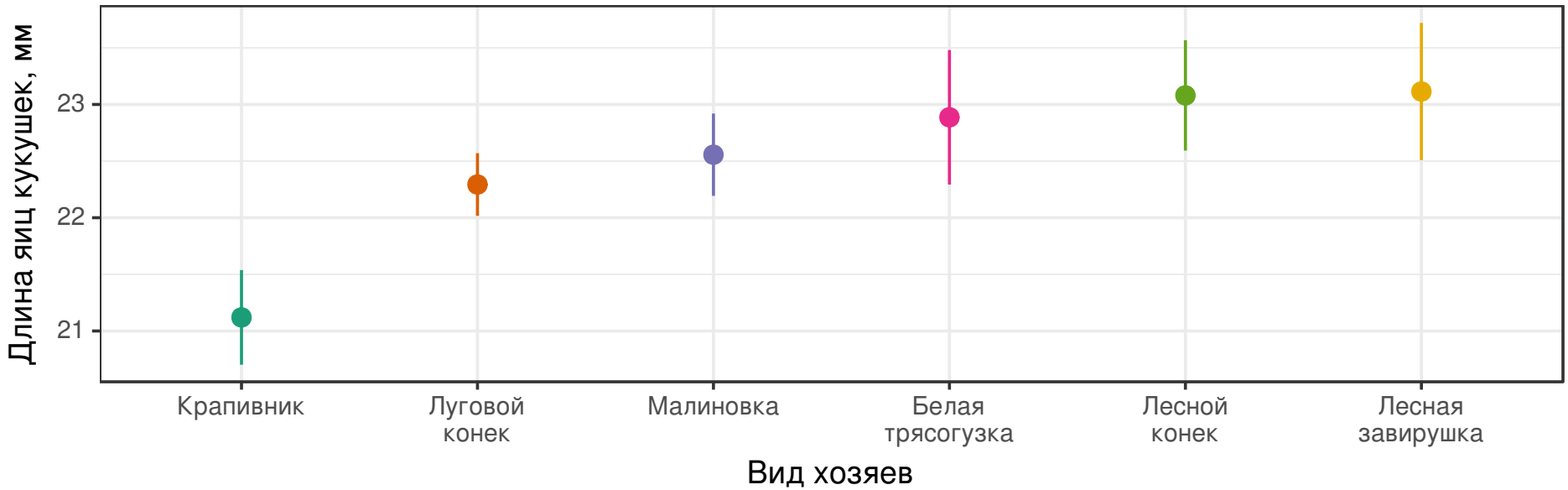
Множественные сравнения

Мы могли бы сравнить длину яиц в гнездах резных хозяев при помощи t-критерия. У нас всего 6 групп. Сколько возможно между ними попарных сравнений?



Множественные сравнения

Мы могли бы сравнить длину яиц в гнездах резных хозяев при помощи t-критерия. У нас всего 6 групп. Сколько возможно между ними попарных сравнений?

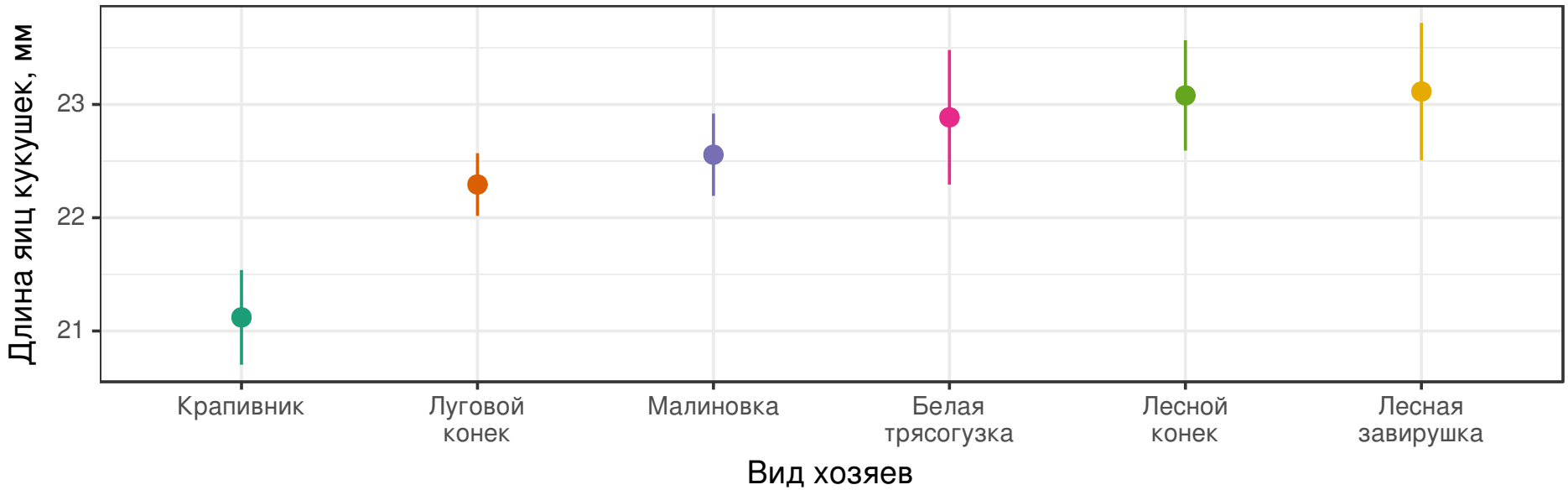


Всего возможно 15 сравнений.

Если для каждого сравнения вероятность ошибки первого рода будет $\alpha_{per\ comparison} = 0.05$, то для группы из 15 сравнений — ?

Множественные сравнения

Мы могли бы сравнить длину яиц в гнездах резных хозяев при помощи t-критерия. У нас всего 6 групп. Сколько возможно между ними попарных сравнений?



Всего возможно 15 сравнений.

Если для каждого сравнения вероятность ошибки первого рода будет $\alpha_{per\ comparison} = 0.05$, то для группы из 15 сравнений — ?

$\alpha_{family\ wise} = 0.05 * 15 = 0.75$

Мы рискуем найти различия там где их нет с 75% вероятностью!!!

Поправка Бонферрони

Если нужно много сравнений, можно снизить $\alpha_{per\ comparison}$ до общепринятого уровня

$$\alpha_{per\ comparison} = \frac{\alpha_{family\ wise}}{n}$$

Поправка Бонферрони

Если нужно много сравнений, можно снизить $\alpha_{per\ comparison}$ до общепринятого уровня

$$\alpha_{per\ comparison} = \frac{\alpha_{family\ wise}}{n}$$

Например, если хотим зафиксировать $\alpha_{family\ wise} = 0.05$

С поправкой Бонферрони $\alpha_{per\ comparison} = 0.05/15 = 0.003$

Это очень жесткая поправка! Мы рискуем не найти достоверных различий, даже там, где они есть…

Но есть выход. Вместо множества попарных сравнений можно использовать один тест — дисперсионный анализ (analysis of variation, ANOVA).

Линейная модель с одним дискретным предиктором

Линейная модель с одним дискретным предиктором

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$$

Коэффициенты линейной модели обозначают отклонения от базового уровня. Такой способ кодирования называется **параметризация фиктивных переменных** или **параметризация тритментов** (dummy (indicator) coding = treatment parametrization = reference cell model), и он вам уже знаком по предыдущим двум лекциям. В R этот способ используется по-умолчанию и обозначается **contr.treatment**.

Переменные-индикаторы (= переменные-болванки):

Если у дискретного фактора k уровней, уравнение модели будет включать $m = k - 1$ переменных, кодирующих принадлежность к этим уровням ($x_{1i}, \ldots, x_{mi}$, i - номер наблюдения). Первый уровень фактора считается базовым и для его кодирования не нужна отдельная переменная.

Коэффициенты:

- β_0 — значение свободного члена для базового уровня дискретного фактора (это среднее значение для базового уровня).
- $\beta_1, \ldots, \beta_m$ — коррекция свободного члена для других уровней (это отклонения средних для других уровней фактора от базового уровня).

Задание 2

- Сколько переменных нужно, чтобы записать модель зависимости длины яиц кукушек от вида птиц-хозяев, если использовать параметризацию тритментов?

Решение:

- Сколько переменных нужно, чтобы записать модель зависимости длины яиц кукушек от вида птиц-хозяев?
- Понадобится 5 переменных, т.к. 6 уровней у фактора `species`

```
levels(cu$species)
```

```
# [1] "крапив"   "луг_кон"  "малин"    "бел_тряс" "лес_кон"  "лес_зав"
```

Уровень крапив будет базовым, и для его кодирования не нужна отдельная переменная).

Модель с одним дискретным предиктором в матричном виде

Уравнение линейной модели для этого примера (в параметризации фиктивных переменных, contr.treatment):

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

- Здесь $i = 1, \ldots, n$, т.е. порядковый номер наблюдения,
- $x_{1i}, \ldots, x_{5i}$ — переменные-индикаторы для фактора `species`

Если расписать эту формулу, получится по отдельному уравнению для каждого из наблюдений:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_5 x_{51} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{12} + \cdots + \beta_5 x_{52} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{1n} + \cdots + \beta_5 x_{5n} + \varepsilon_n
\end{aligned}$$

Эту систему уравнений

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_5 x_{51} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{12} + \cdots + \beta_5 x_{52} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{1n} + \cdots + \beta_5 x_{5n} + \varepsilon_n
\end{aligned}
$$

можно переписать в виде матриц:

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{51} \\ 1 & x_{12} & \cdots & x_{52} \\ \vdots & & & \\ 1 & x_{1n} & \cdots & x_{5n} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

Для такой длинной формы записи матриц

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{51} \\ 1 & x_{12} & \cdots & x_{52} \\ \vdots & & & \\ 1 & x_{1n} & \cdots & x_{5n} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

есть сокращенная форма записи:

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

- **Y** - матрица значений зависимой переменной, один столбец в n строк (n - число наблюдений)
- **X** - матрица независимых переменных с n строк, ее первый столбец содержит единицы, далее по столбцу для каждой из переменных в модели
- β - матрица коэффициентов линейной модели, столбец
- ε - матрица остатков, один столбец в n строк

Модельная матрица **X** в дисперсионном анализе

Посмотреть своими глазами на эти переменные-индикаторы можно так:

```
X <- model.matrix(~ species, data = cu)
head(X)
```

```
#   (Intercept) speciesлуг_кон speciesмалин speciesбел_тряс
# 1           1              1             0                0
# 2           1              1             0                0
# 3           1              1             0                0
# 4           1              1             0                0
# 5           1              1             0                0
# 6           1              1             0                0
#   speciesлес_кон speciesлес_зав
# 1              0              0
# 2              0              0
# 3              0              0
# 4              0              0
# 5              0              0
# 6              0              0
```

Задание 3

- Подберите линейную модель зависимости длины яиц кукушек в гнездах от вида птиц-хозяев
- Вспомните, что означают коэффициенты этой линейной модели, и вычислите, чему равен ожидаемый размер яиц в гнездах каждого из видов-хозяев.

Решение:

```
cmod <- lm(length ~ species, data = cu)
coef(cmod)
```

```
#     (Intercept)  speciesлуг_кон    speciesмалин speciesбел_тряс
#           21.12            1.17            1.44            1.77
#  speciesлес_кон  speciesлес_зав
#            1.96            1.99
```

Первый коэффициент — средний размер яиц кукушек в гнездах крапивников (на базовом уровне):

- Крапивник — $length = b_0 = 21.12$

Другие коэффициенты — разницу размеров яиц кукушек в гнездах других хозяев и в гнездах крапивников (отклонения от базового уровня):

- Луговой конек — $length = b_0 + b_1 = 22.29$
- Малиновка — $length = b_0 + b_2 = 22.56$
- Белая трясогузка — $length = b_0 + b_3 = 22.89$
- Лесной конек — $length = b_0 + b_4 = 23.08$
- Лесная завирушка — $length = b_0 + b_5 = 23.11$

t-тесты значимости коэффициентов линейной модели с одним дискретным предиктором

- t-тест для первого коэффициента показывает отличается ли от нуля среднее на базовом уровне
- t-тесты значимости других коэффициентов показывают достоверность отличий средних значений в группах от среднего на базовом уровне.

По значениям коэффициентов нельзя сказать влияет ли дискретный фактор целиком (исключение — фактор с двумя градациями).

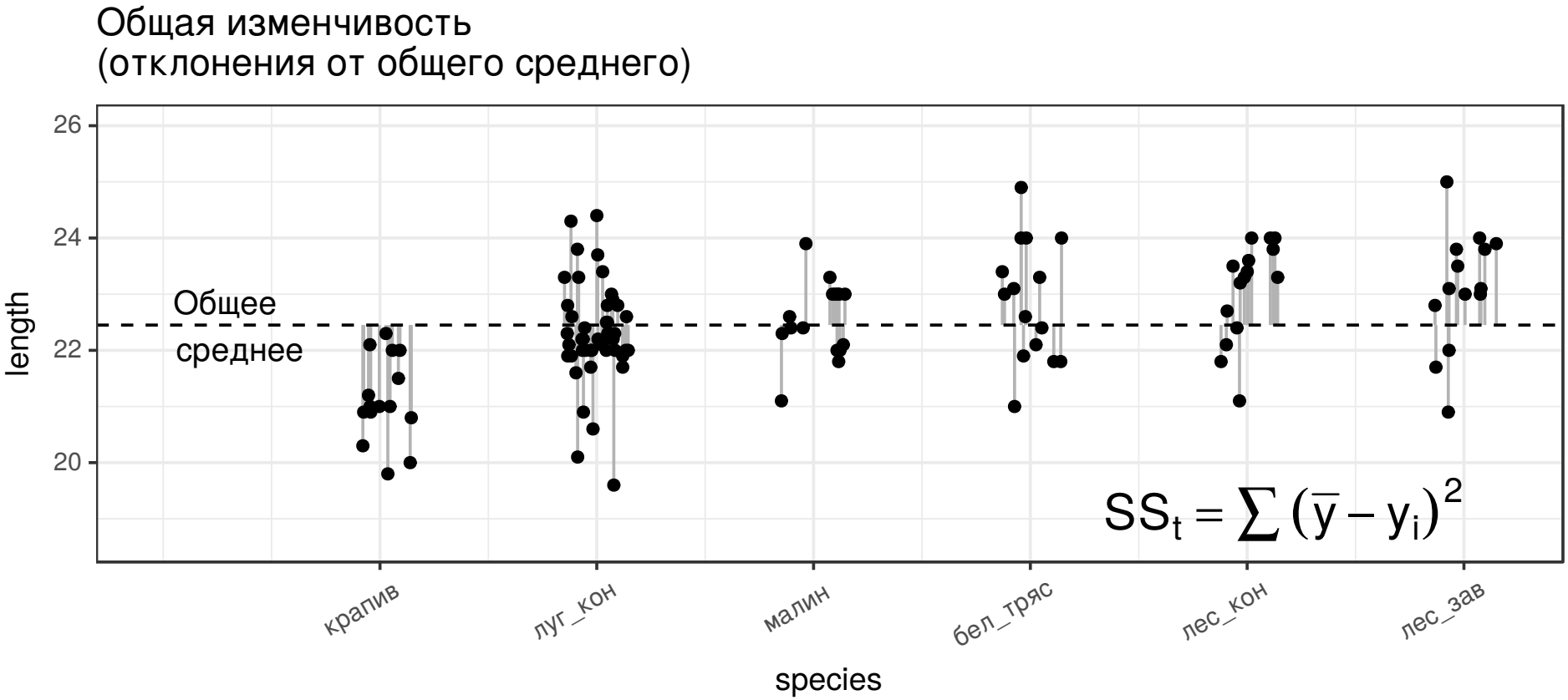
```
coef(summary(cmod))
```

```
#                  Estimate Std. Error t value  Pr(>|t|)
# (Intercept)         21.12      0.234   90.36 6.20e-108
# speciesлуг_кон       1.17      0.270    4.35  3.01e-05
# speciesмалин         1.44      0.325    4.41  2.31e-05
# speciesбел_тряс      1.77      0.331    5.34  4.70e-07
# speciesлес_кон       1.96      0.331    5.93  3.31e-08
# speciesлес_зав       1.99      0.336    5.93  3.33e-08
```

Дисперсионный анализ

Общая изменчивость

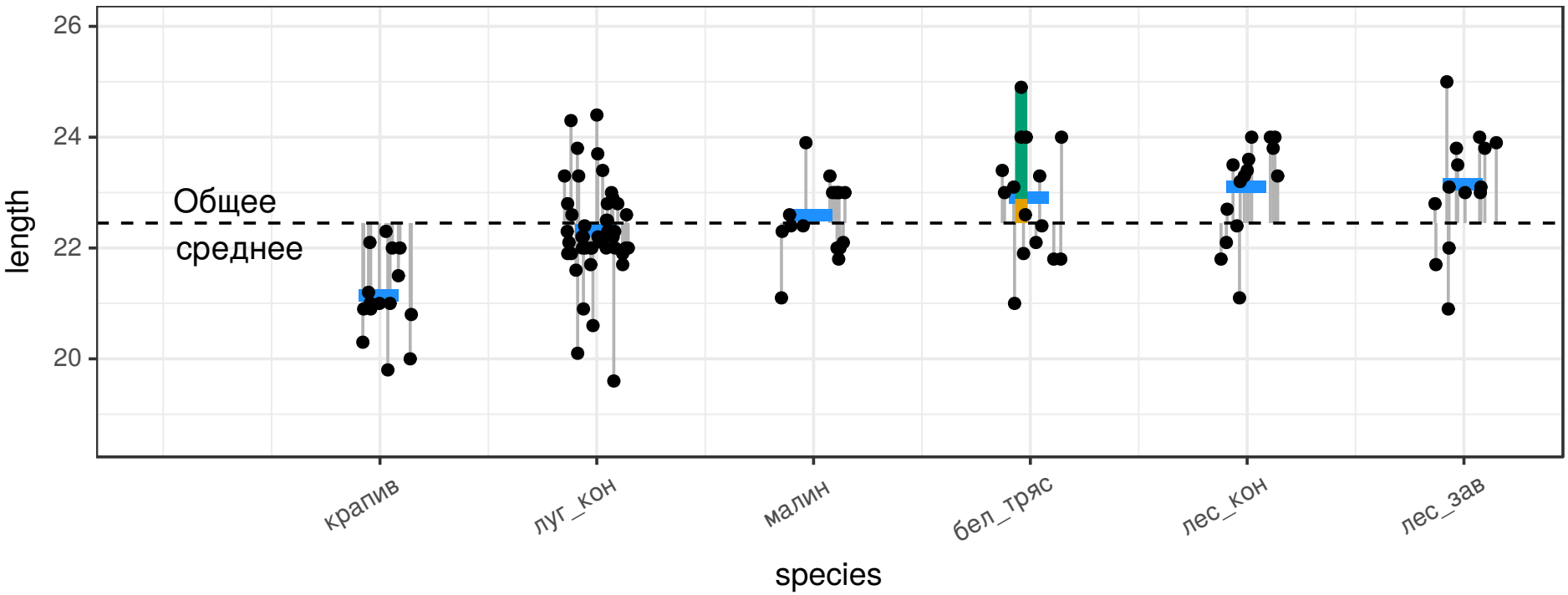
Общая изменчивость SS_t — это сумма квадратов отклонений наблюдаемых значений y_i от общего среднего $\bar{y}$



Отклонения от общего среднего

Отклонения от общего среднего складываются из двух составляющих:

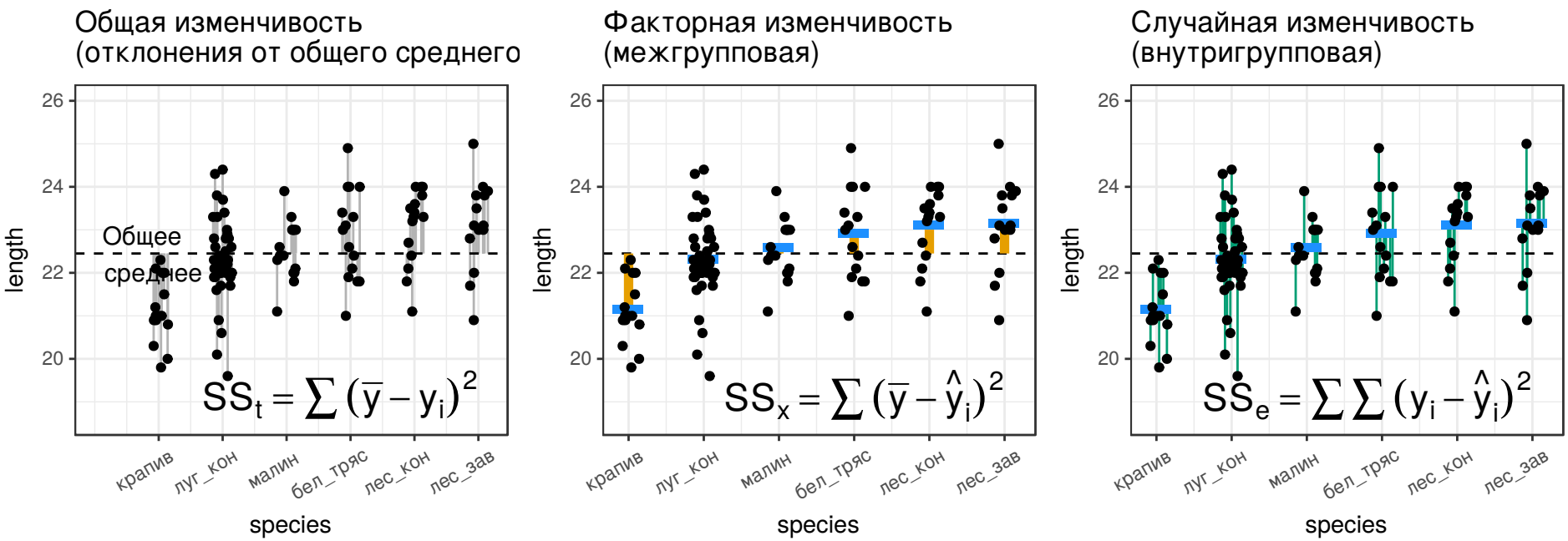
- Внутригрупповые отклонения — отклонения наблюдаемых значений от внутригрупповых средних
- Межгрупповые отклонения — отклонения внутригрупповых средних от общего среднего (“эффекты” групп)



Структура общей изменчивости

Общая изменчивость SS_t складывается из изменчивости связанной с фактором SS_x и случайной изменчивости SS_e

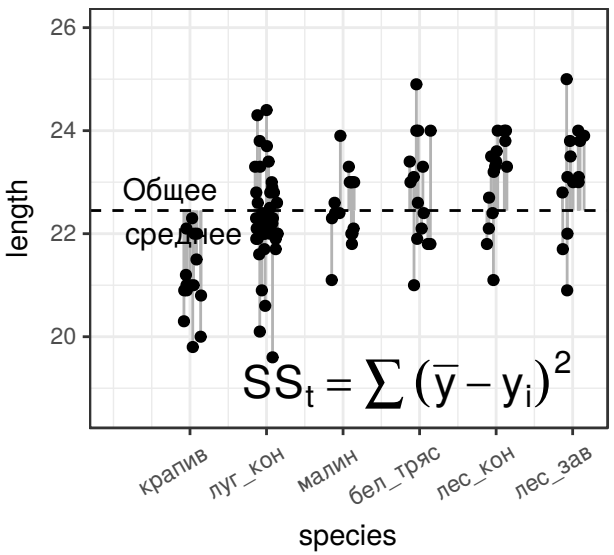
$$SS_t = SS_x + SS_e$$



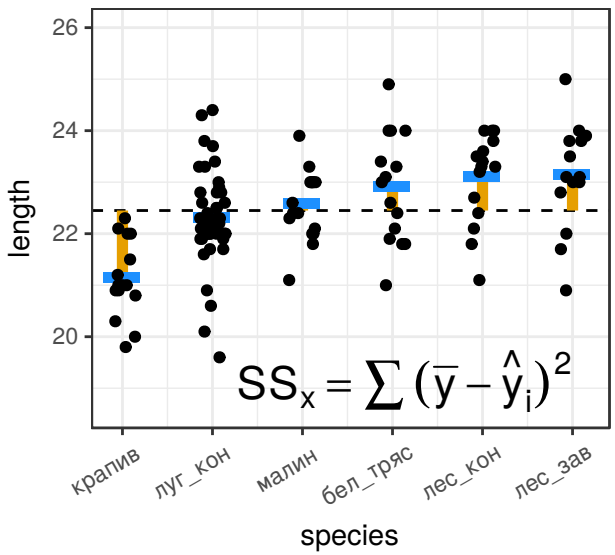
Средние квадраты отклонений

Если поделить суммы квадратов отклонений (*SS*) на их число степеней свободы, то получатся средние квадраты отклонений (*MS*) — дисперсии

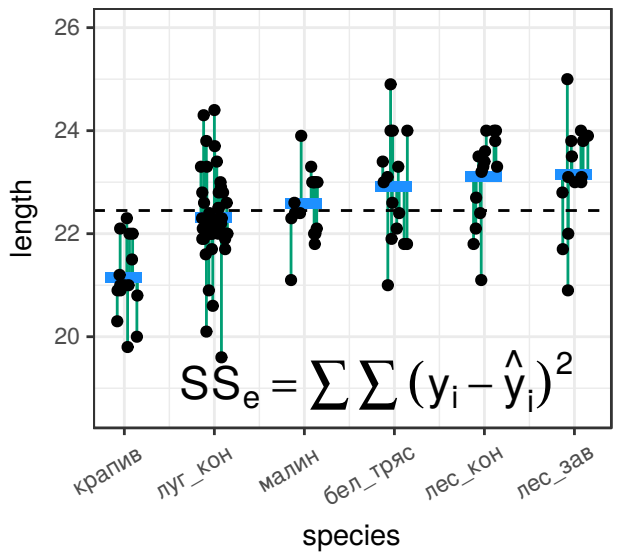
Общая изменчивость
(отклонения от общего среднего



Факторная изменчивость
(межгрупповая)



Случайная изменчивость
(внутригрупповая)



MS_t полная дисперсия | MS_x факторная дисперсия | MS_e остаточная дисперсия

$$MS_t = \frac{SS_t}{df_t} \qquad MS_x = \frac{SS_x}{df_x} \qquad MS_e = \frac{SS_e}{df_e}$$

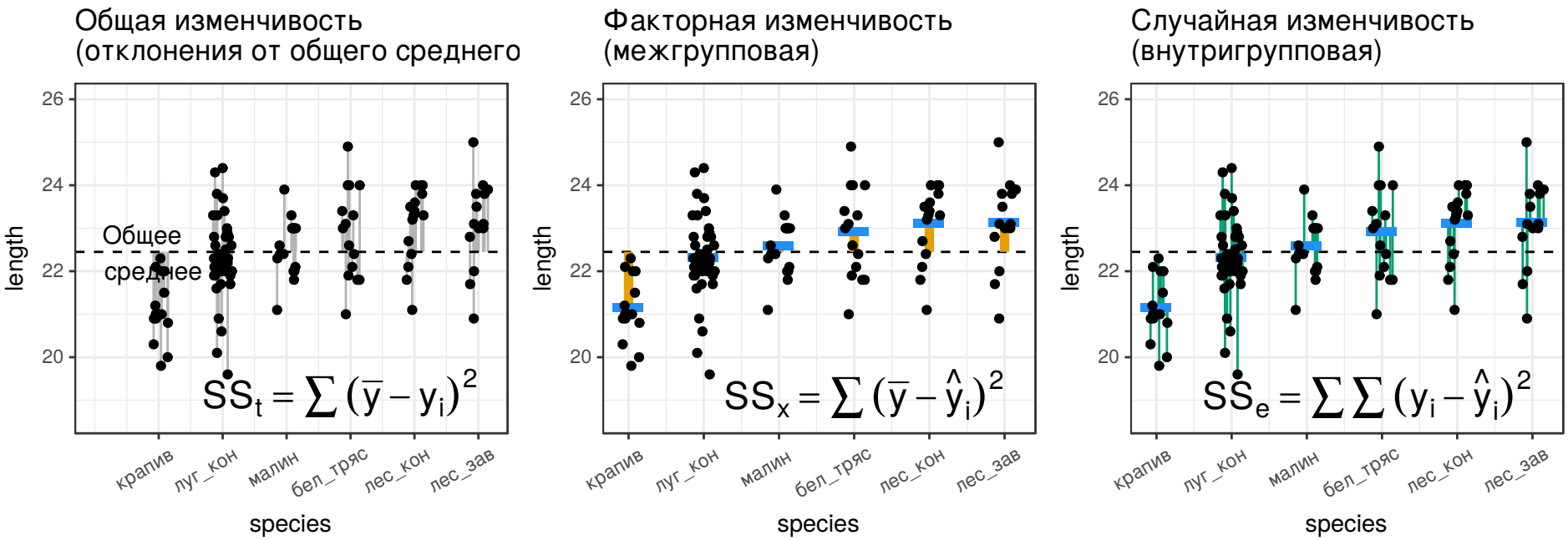
$$SS_t = \sum (\bar{y} - y_i)^2 \qquad SS_x = \sum (\hat{y} - \bar{y})^2 \qquad SS_e = \sum (\hat{y} - y_i)^2$$

$$df_t = N - 1 \qquad df_x = a - 1 \qquad df_e = N - a$$

Если выборки из одной совокупности, то

Если выборки из одной совокупности, то наблюдения из разных групп будут отличаться друг от друга не больше, чем наблюдения из одной группы, т.е. факторная дисперсия будет близка к случайной дисперсии $MS_x \sim MS_e$. Их равенство можно проверить при помощи F-критерия

$$F_{df_x, df_e} = \frac{MS_x}{MS_e}$$



F-критерий

$$F_{df_x, df_e} = \frac{MS_x}{MS_e}$$

Гипотезы:

H_0: все выборки взяты из одной совокупности — $\bar{X}_1 = \bar{X}_2 = \cdots = \bar{X}_a$. Тогда $MS_x = MS_e$

H_A: какая-то из выборок из другой совокупности, т.е. какое-то среднее значение $\bar{X}_i$ отличается от других. Тогда $MS_x \neq MS_e$

F-статистика подчиняется F-распределению. Форма F-распределения зависит от двух параметров: $df_x = a - 1$ и $df_e = N - a$

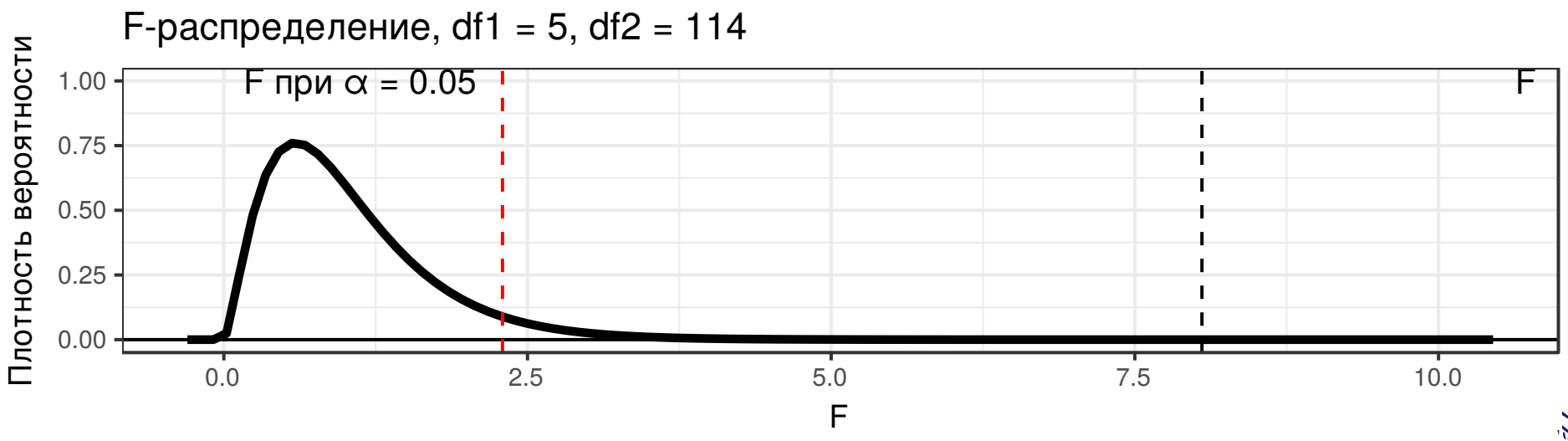


Таблица дисперсионного анализа

Источник изменчивости	SS	df	MS	F
Название фактора	$SS_x = \sum (\bar{y} - \hat{y}_i)^2$	$df_x = a - 1$	$MS_x = \frac{SS_x}{df_x}$	$F_{df_x df_e} = \frac{MS_x}{MS_e}$
Случайная	$SS_e = \sum (y_i - \hat{y}_i)^2$	$df_e = N - a$	$MS_e = \frac{SS_e}{df_e}$	
Общая	$SS_t = \sum (\bar{y} - y_i)^2$	$df_t = N - 1$		

Минимальное упоминание результатов в тексте должно содержать F_{df_x, df_e} и p.

Делаем дисперсионный анализ в R

В R есть много функций для дисперсионного анализа. Мы рекомендуем `Anova()` (**с большой буквы**) из пакета `car`. Зачем? Эта функция умеет тестировать влияние факторов в определенном порядке. Когда факторов будет больше одного, это станет важно для результатов.

```
library(car)
cu_anova <- Anova(cmod)
cu_anova
```

```
# Anova Table (Type II tests)
#
# Response: length
#           Sum Sq  Df F value      Pr(>F)
# species     42.8   5    10.4 0.000000029 ***
# Residuals   93.4 114
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Результаты дисперсионного анализа

Результаты дисперсионного анализа можно описать в тексте:

- Длина яиц кукушек в гнездах разных птиц-хозяев достоверно различается ($F_{5,114} = 10.45$, $p < 0.01$).

Результаты дисперсионного анализа

Результаты дисперсионного анализа можно представить в виде таблицы

- Длина яиц кукушек достоверно различалась в гнездах разных птиц-хозяев (Табл. 1).

Table 1: Результаты дисперсионного анализа длины яиц кукушек в гнездах разных птиц-хозяев. SS — суммы квадратов отклонений, df — число степеней свободы, F — значение F-критерия, P — доверительная вероятность.

	SS	df	F	P
Хозяин	42.8	5	10.4	<0.01
Остаточная	93.4	114		

Условия примененимости дисперсионного анализа

Результатам тестов можно верить, если выполняются условия применимости

Условия применимости дисперсионного анализа:

- Случайность и независимость наблюдений внутри групп
- Нормальное распределение остатков
- Гомогенность дисперсий остатков
- Отсутствие колинеарности факторов (независимость групп)

Другие ограничения:

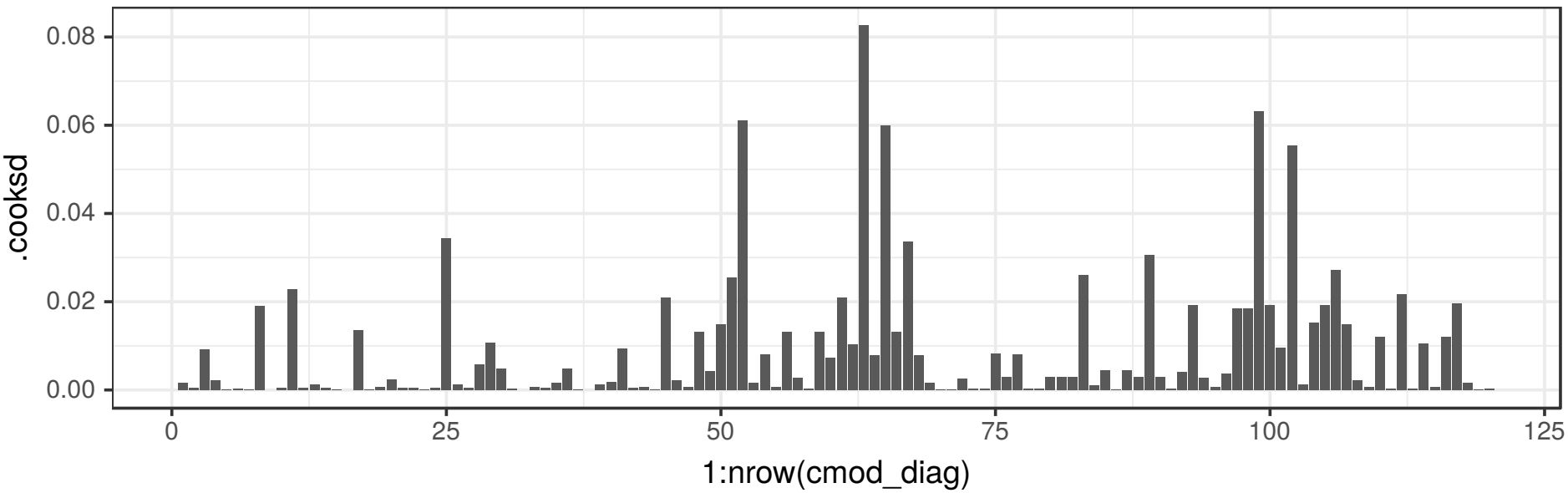
- Лучше работает, если размеры групп примерно одинаковы (т.наз. сбалансированный дисперсионный комплекс)
- Устойчив к отклонениям от нормального распределения (при равных объемах групп или при больших выборках)

Проверяем выполнение условий применимости

```
# Данные для графиков остатков
cmod_diag <- fortify(cmod)
```

1) График расстояния Кука

```
ggplot(cmod_diag, aes(x = 1:nrow(cmod_diag), y = .cooksd)) +
  geom_bar(stat = "identity")
```



- Выбросов нет

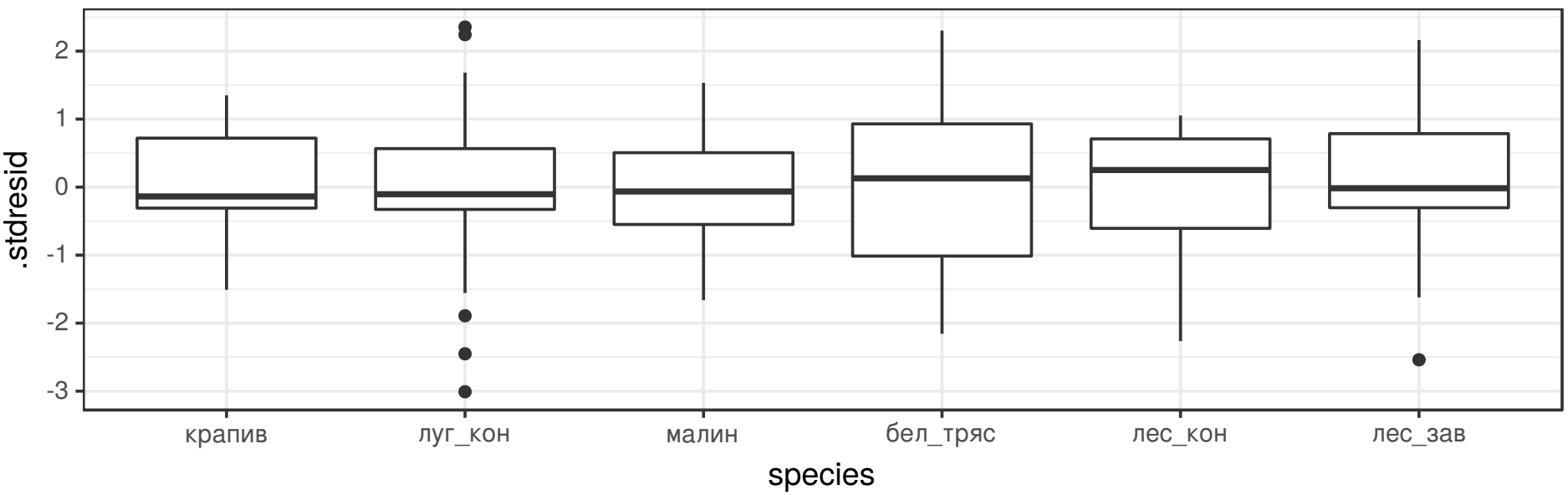
2) График остатков от предсказанных значений

```
ggplot(cmod_diag, aes(x = .fitted, y = .stdresid)) + geom_jitter()
```

У нас один единственный дискретный предиктор, поэтому удобнее сразу боксплот

3) Графики остатков от предикторов в модели и не в модели

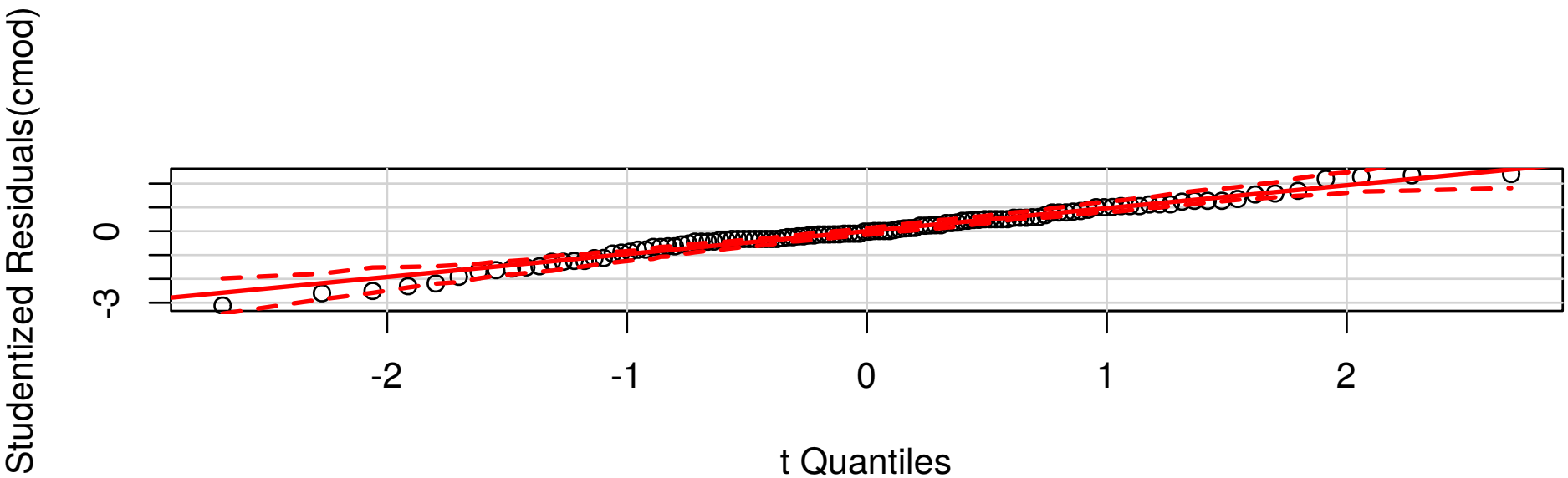
```
ggplot(cmod_diag, aes(x = species, y = .stdresid)) + geom_boxplot()
```



Дисперсии почти одинаковые. Может быть, в одной из групп чуть больше

4) Квантильный график остатков

```
library(car)
qqPlot(cmod)
```



- Распределение остатков отличается от нормального

Пост хок тесты

Как понять, какие именно группы различаются

Дисперсионный анализ говорит нам только, есть ли влияние фактора, но не говорит, какие именно группы различаются.

Коэффициенты линейной модели в `summary(cmod)` содержат лишь часть ответа — сравнение средних значених всех групп со средним на базовом уровне.

Если нас интересуют другие возможные попарные сравнения, нужно сделать пост хок тест.

Post hoc тесты

Пост хок тесты — попарные сравнения средних **после того, как дисперсионный анализ показал, что влияние фактора достоверно**

Свойства post hoc тестов:

- **Применяются, только если влияние фактора значимо**
- Делают поправку для снижения вероятности ошибки I рода α, (но не слишком большую, чтобы не снизилась мощность, и чтобы не возросла вероятность ошибки II рода β)
- Учитывают величину различий между средними
- Учитывают количество сравниваемых пар
- Различаются по степени консервативности (тест Тьюки — разумный компромисс)
- Работают лучше при равных объемах групп, при гомогенности дисперсий

Пост хок тест Тьюки в R

- `glht()` — “general linear hypotheses testing”
- `linfct` — аргумент, задающий гипотезу для тестирования
- `mcp()` — функция, чтобы задавать множественные сравнения (обычные пост хоки)
- `species` = “Tukey” — тест Тьюки по фактору `species`

```
library(multcomp)
cu_ph <- glht(cmod, linfct = mcp(species = "Tukey"))
```

Результаты попарных сравнений (тест Тьюки)

Таблица результатов пост хок теста практически нечитабельна.

```
summary(cu_ph)
```

```
#
#    Simultaneous Tests for General Linear Hypotheses
#
# Multiple Comparisons of Means: Tukey Contrasts
#
#
# Fit: lm(formula = length ~ species, data = cu)
#
# Linear Hypotheses:
#                          Estimate Std. Error t value Pr(>|t|)
# луг_кон - крапив == 0     1.1733     0.2699    4.35   <0.001 ***
# малин - крапив == 0       1.4362     0.3253    4.41   <0.001 ***
# бел_тряс - крапив == 0    1.7667     0.3305    5.34   <0.001 ***
# лес_кон - крапив == 0     1.9600     0.3305    5.93   <0.001 ***
# лес_зав - крапив == 0     1.9943     0.3364    5.93   <0.001 ***
# малин - луг_кон == 0      0.2629     0.2635    1.00    0.915
# бел_тряс - луг_кон == 0   0.5933     0.2699    2.20    0.241
# лес_кон - луг_кон == 0    0.7867     0.2699    2.91    0.047 *
# лес_зав - луг_кон == 0    0.8210     0.2770    2.96    0.041 *
# бел_тряс - малин == 0     0.3304     0.3253    1.02    0.909
# лес_кон - малин == 0      0.5238     0.3253    1.61    0.587
# лес_зав - малин == 0      0.5580     0.3313    1.68    0.538
# лес_кон - бел_тряс == 0   0.1933     0.3305    0.58    0.992
```

Результаты пост хок теста

Результаты пост хок теста можно привести в виде текста…

- Размер яиц кукушек в гнездах крапивника достоверно меньше, чем в гнездах лугового конька (тест Тьюки, $p < 0.01$). Размер яиц кукушек в гнездах лесной завирушки, белой трясогузки, малиновки и лесного конька не различается, но яйца в гнездах этих видов крупнее, чем у лугового конька или крапивника (тест Тьюки, от $p < 0.01$ до 0.05).

…или построить график

Данные для графика при помощи predict()

```
MyData <- data.frame(
  species = factor(levels(cu$species),
                   levels = levels(cu$species)))

MyData <- data.frame(
  MyData,
  predict(cmod, newdata = MyData, interval = "confidence")
  )

MyData
```

```
#    species  fit  lwr  upr
# 1   крапив 21.1 20.7 21.6
# 2  луг_кон 22.3 22.0 22.6
# 3    малин 22.6 22.1 23.0
# 4 бел_тряс 22.9 22.4 23.3
# 5  лес_кон 23.1 22.6 23.5
# 6  лес_зав 23.1 22.6 23.6
```

Задание 4

Создайте MyData вручную:

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

```
MyData <- data.frame(
  species = factor(levels(cu$species),
                   levels = levels(cu$species)))

X <- model.matrix()
betas <-
MyData$fit <-  %*%
MyData$se <- sqrt(diag(X %*% vcov(cmod) %*% t(X)))
MyData$lwr <- MyData$ - 1.96 * MyData$
MyData$upr <- MyData$ + 1.96 * MyData$
```

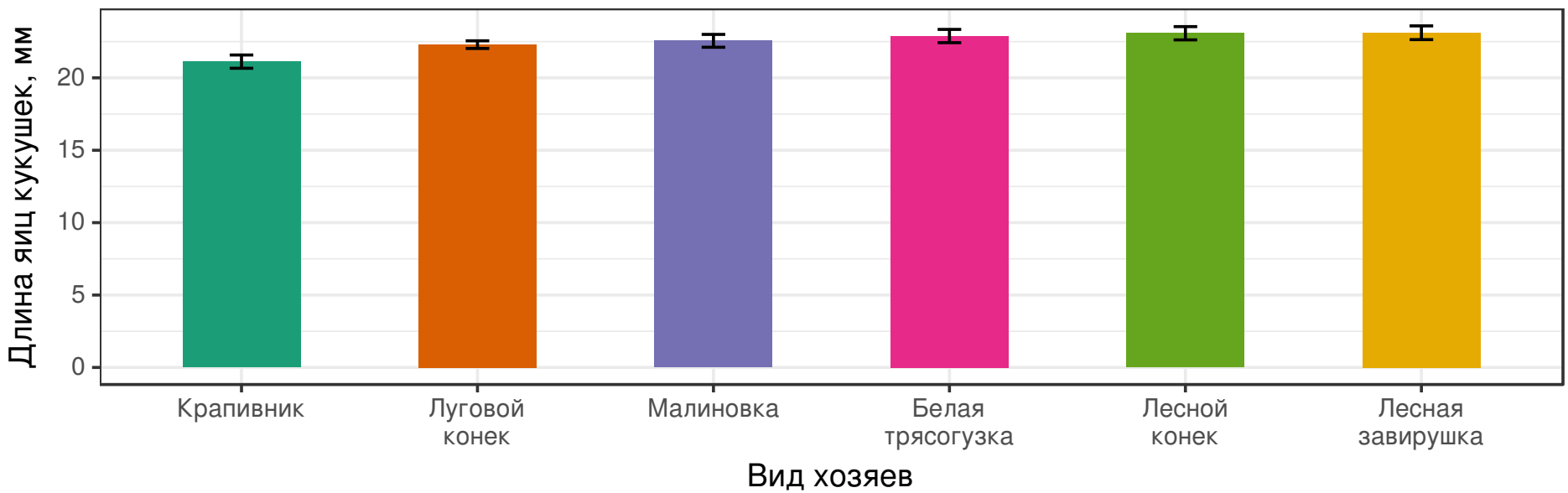
Решение:

```
MyData <- data.frame(
  species = factor(levels(cu$species),
                   levels = levels(cu$species)))
X <- model.matrix(~species, data = MyData)
betas <- coef(cmod)
MyData$fit <- X %*% betas
MyData$se <- sqrt(diag(X %*% vcov(cmod) %*% t(X)))
MyData$lwr <- MyData$fit - 1.96 * MyData$se
MyData$upr <- MyData$fit + 1.96 * MyData$se
MyData
```

```
#    species  fit    se  lwr  upr
# 1   крапив 21.1 0.234 20.7 21.6
# 2  луг_кон 22.3 0.135 22.0 22.6
# 3    малин 22.6 0.226 22.1 23.0
# 4 бел_тряс 22.9 0.234 22.4 23.3
# 5  лес_кон 23.1 0.234 22.6 23.5
# 6  лес_зав 23.1 0.242 22.6 23.6
```

Столбчатый график

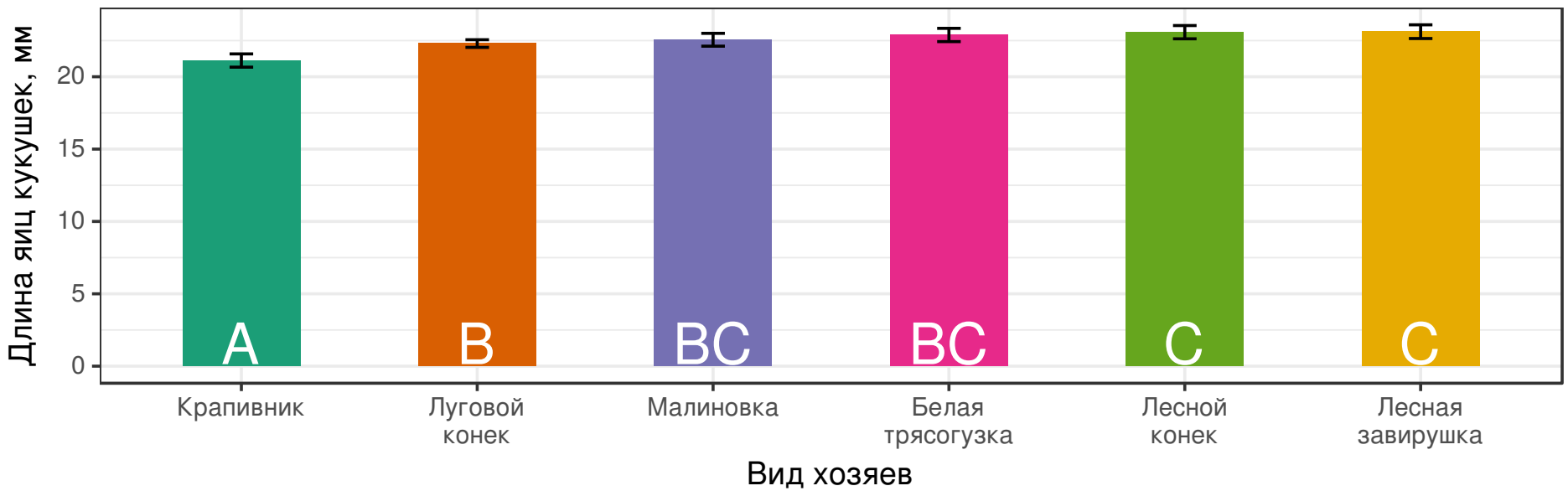
```
gg_bars <- ggplot(data = MyData, aes(x = species, y = fit)) +
  geom_bar(stat = "identity", aes(fill = species), width = 0.5) +
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.1) +
  labs(x = "Вид хозяев", y = "Длина яиц кукушек, мм") +
  scale_fill_brewer(name = "Вид \nхозяев", palette = "Dark2") +
  scale_x_discrete(labels = c("Крапивник", "Луговой\nконек", "Малиновка",
                              "Белая\nтрясогузка", "Лесной\nконек", "Лесная\nза
  theme(legend.position = "none")
gg_bars
```



Можно привести результаты пост хок теста на столбчатом графике

Достоверно различающиеся группы обозначим разными буквами

```
gg_bars_coded <- gg_bars +
  geom_text(aes(y = 1.6,  label = c("A", "B", "BC", "BC", "C", "C")),
            colour = "white", size = 7)
gg_bars_coded
```



Take home messages

- Дисперсионный анализ — линейная модель с дискретными предикторами, существует в нескольких параметризациях, которые отличаются трактовками коэффициентов
- При помощи дисперсионного анализа можно проверить гипотезу о равенстве средних значений в группах
- Условия применимости дисперсионного анализа
  - Случайность и независимость групп и наблюдений внутри групп
  - Нормальное распределение в группах
  - Гомогенность дисперсий в группах
- При множественных попарных сравнениях увеличивается вероятность ошибки первого рода, поэтому нужно вносить поправку для уровня значимости
- Post hoc тесты — это попарные сравнения после дисперсионного анализа, которые позволяют сказать, какие именно средние различаются

Дополнительные ресурсы

- Quinn, Keough, 2002, pp. 173–207
- Logan, 2010, pp. 254–282
- Open Intro to Statistics, pp.236–246
- Sokal, Rohlf, 1995, pp. 179–260
- Zar, 2010, pp. 189-207