Дисперсионный анализ, часть 1

Математические методы в зоологии - на R, осень 2015

Марина Варфоломеева

Знакомимся дисперсионным анализом

- Опасности множественных сравнений
- Почему можно использовать дисперсии для сравнения средних
- Модель дисперсионного анализа
- Условия применимости дисперсионного анализа
- Post hoc тесты
- Представление результатов дисперсионного анализа

Вы сможете

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

Пример: сон у млекопитающих

- `TotalSleep` - общая продолжительность сна. В нашем анализе это будет зависимая переменная
- `Danger` - уровень опасности среды для вида, пять градаций (1 - 5)

```
library(readxl)
sleep <- read_excel("sleep.xlsx", sheet = 1)
head(sleep, 2)
```

```
##                  Species BodyWt BrainWt NonDreaming Dreaming
## 1        Africanelephant   6654  5712.0          NA       NA
## 2 Africangiantpouchedrat      1     6.6         6.3        2
##   TotalSleep LifeSpan Gestation Predation Exposure Danger
## 1        3.3     38.6       645         3        5      3
## 2        8.3      4.5        42         3        1      3
```

```
# Сделаем sleep$Danger фактором
sleep$Danger <- factor(sleep$Danger, levels = 1:5, labels = c("очень низкий",
    "низкий", "средний", "высокий", "очень высокий"))
```

Данные: Allison, Cicchetti (1976), электронная версия Statlib database

Задание: Постройте график

Постройте график зависимости общей продолжительности сна от уровня опасности среды. Какой геом лучше подойдет для изображения (`geom_point` или `geom_boxplot`)?

Раскрасьте график в зависимости от уровня опасности среды (используйте эстетики `fill` или `colour`)

Придумайте, каким образом посчитать, в какой группе животных общая продолжительность сна больше?

Дополнительное задание:

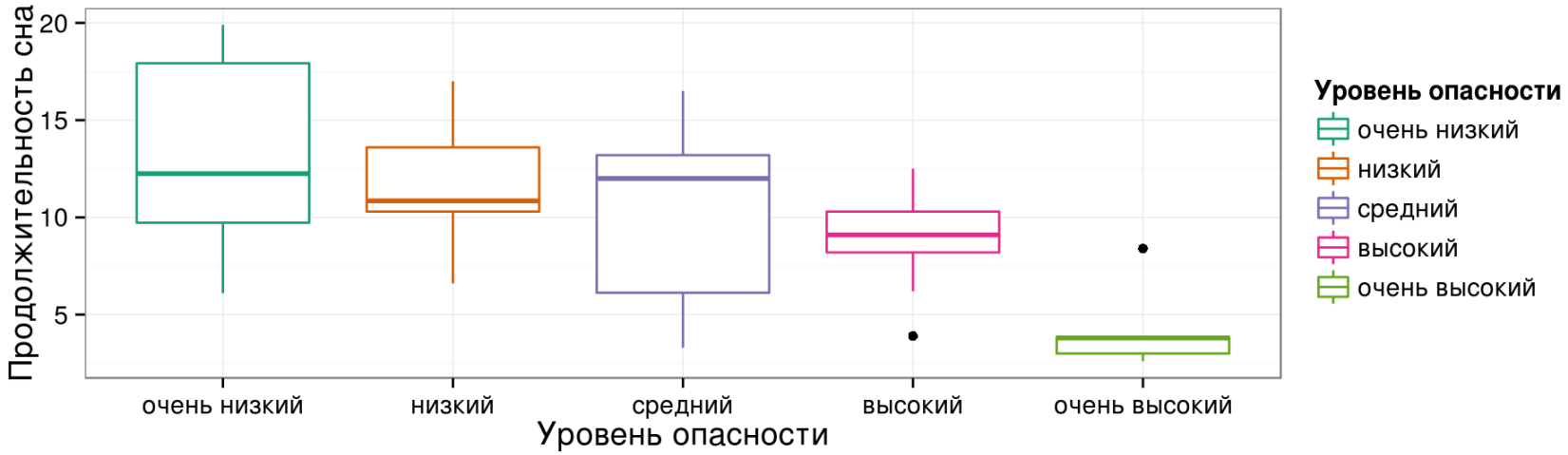
Попробуйте сменить палитру раскраски, используя `scale_colour_brewer` (варианты можно посмотреть в справке в подразделе примеров или в интернете Colors (ggplot2): раздел RColorBrewer palette chart)

Решение

```
library(ggplot2)
theme_set(theme_bw(base_size = 14) + theme(legend.key = element_blank()))

gg_sleep <- ggplot(data = sleep, aes(x = Danger, y = TotalSleep, colour = Danger)) +
    labs(x = "Уровень опасности", y = "Продолжительность сна") +
    scale_colour_brewer("Уровень опасности", palette = "Dark2")
gg_sleep + geom_boxplot()
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



Опасности множественных сравнений

Попарные сравнения средних

2 группы - 1 сравнение

$$\alpha_{\text{для сравнения}} = 0.05$$

4 группы - 6 сравнений

$$\alpha_{\text{для сравнения}} = 0.05$$

А какой будет α для группы из 6 сравнений?

- $\alpha_{\text{для группы сравнений}} = 0.05 \cdot 6 = 0.3$
- Опасно! Случайно найдем различия там, где их нет!

Если нужно много сравнений можно снизить $\alpha_{\text{для сравнения}}$

$$\alpha_{\text{для группы сравнений}} = \alpha_{\text{для сравнения}} \cdot n$$

Хотим зафиксировать $\alpha_{\text{для группы сравнений}} = 0.05$

Поправка Бонферрони:

$$\alpha_{\text{для сравнения}} = \frac{\alpha_{\text{для группы сравнений}}}{n}$$

- для 4 групп, 6 сравнений, $\alpha_{\text{для сравнения}} = 0.008$
  Очень жесткий критерий!

Дисперсионный анализ

Модель дисперсионного анализа

$$y_{ij} = \mu + a_i + \epsilon_{ij}$$

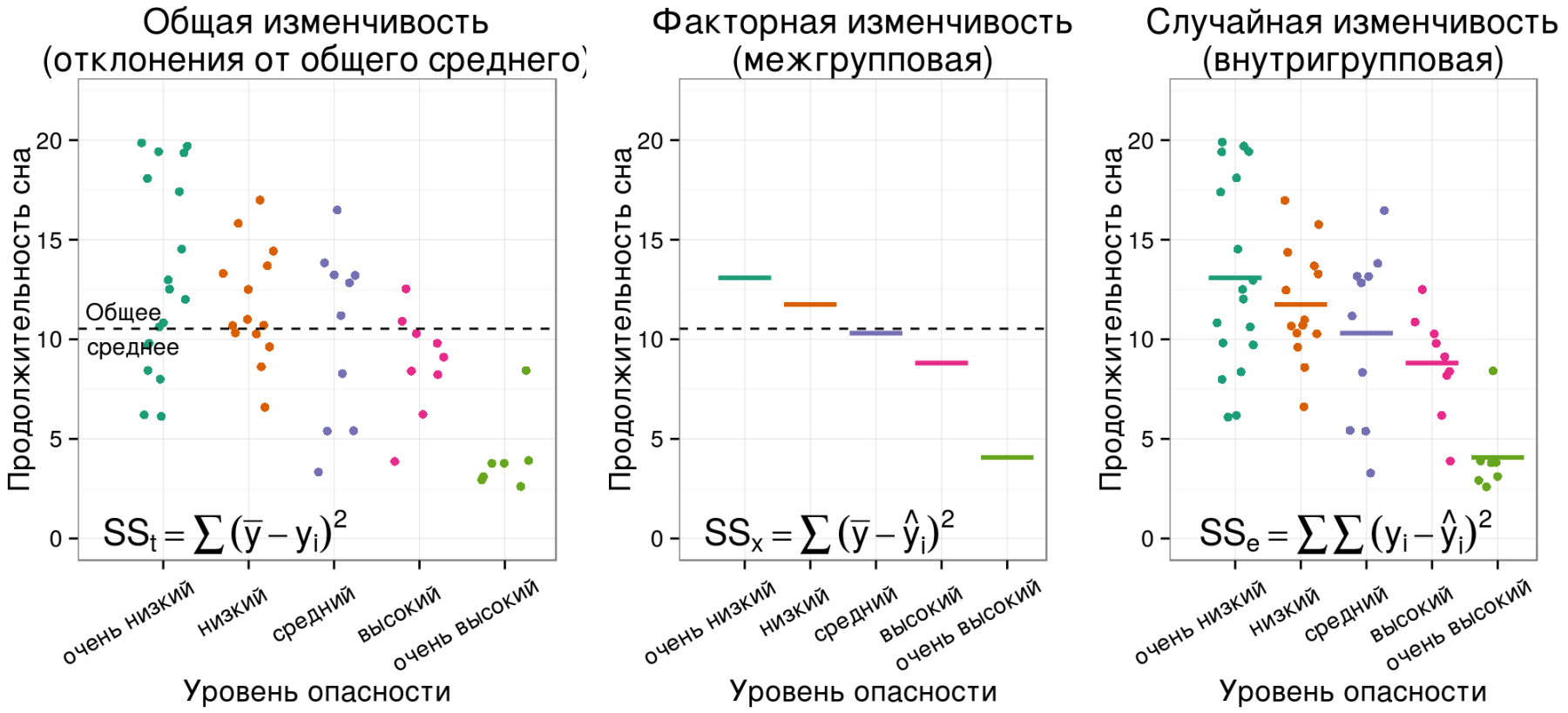
Из чего складываются средние значения в группах по фактору?

Группа	Общее среднее	Эффект	Случайная изменчивость
очень низкий	μ	a_1	ϵ_{1j}
низкий	μ	a_2	ϵ_{2j}
...	...	...	...
очень высокий	μ	a_5	ϵ_{5j}

```
##          Danger TotalSleep
## 1       средний         3.3
## 2  очень низкий        18.3
## 3  очень низкий        12.5
## 4  очень низкий        16.7
## 5  очень низкий        14.5
## 6  очень низкий        15.5
## 7  очень низкий        19.7
## 8  очень низкий        17.4
## 9  очень низкий        13.8
## 10 очень низкий        13.2
## 11 очень низкий        15.6
## 12 очень низкий        14.1
## 13 очень низкий        15.1
## 14 очень низкий        13.2
## 15 очень низкий        12.8
## 16 очень низкий        16.9
## 17 очень низкий        16.6
## 18 очень низкий        15.3
## 19 очень низкий         6.1
## 20 очень низкий        18.1
## 21 очень высокий         NA
## 22 очень высокий        3.8
## 23        низкий        14.4
## 24 очень низкий        12.0
## 25 очень низкий         6.2
## 26 очень низкий        13.0
## 27       средний        13.8
## 28       высокий         8.2
## 29 очень высокий        2.9
## 30 очень низкий        10.8
## 31       высокий         NA
## 32       высокий         9.1
## 33 очень низкий        19.9
## 34 очень низкий         8.0
## 35 очень низкий        10.6
## 36       средний        11.2
## 37       средний        13.2
## 38       средний        12.8
## 39 очень низкий        19.4
## 40 очень низкий        17.4
## 41 очень высокий         NA
## 42        низкий        17.0
## 43       высокий        10.9
## 44        низкий        13.7
## 45       высокий         8.4
## 46 очень высокий         8.4
```

Структура общей изменчивости

Общая изменчивость (SSt) = Факторная (SSx) + Случайная (SSe)



Если выборки из одной совокупности, то Факторная изменчивость = Случайная изменчивость

Таблица дисперсионного анализа

Источник изменчивости	Суммы квадратов отклонений, SS	Число степеней свободы, df	Средний квадрат отклонений (дисперсия), MS	F
Название фактора	$SS_x = \sum (\bar{y} - \hat{y}_i)^2$	$df_x = a - 1$	$MS_x = \frac{SS_x}{df_x}$	$F_{df_r, df_e} = \frac{MS_r}{MS_e}$
Случайная	$SS_e = \sum (y_i - \hat{y}_i)^2$	$df_e = N - a$	$MS_e = \frac{SS_e}{df_e}$	
Общая	$SS_t = \sum (\bar{y} - y_i)^2$	$df_t = N - 1$		

Гипотезы:

$$H_0 : MS_x = MS_e, H_A : MS_x \neq MS_e$$

Дисперсионный анализ в R

Используем Anova из пакета car, хотя есть и другие функции. Зачем? Когда факторов будет больше одного, эта функция сможет правильно оценить достоверность каждого из них независимо от других.

Anova(результат_функции_lm) - дисперсионный анализ

```
library(car)
mod <- lm(TotalSleep ~ Danger, data = sleep)
sleep_anova <- Anova(mod)
sleep_anova
```

```
## Anova Table (Type II tests)
##
## Response: TotalSleep
##           Sum Sq Df F value   Pr(>F)
## Danger       457  4    8.05 0.000038 ***
## Residuals    752 53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Общая продолжительность сна различается у видов животных, которые в разной степени подвержены опасностям в течение жизни ($F_{4,53} = 8.05, p < 0.01$).

Вопрос:

Назовите условия применимости дисперсионного анализа

- Подсказка: дисперсионный анализ - линейная модель, как и регрессия

Ответ:

Условия применененимости дисперсионного анализа:

- Случайность и независимость групп и наблюдений внутри групп
- Нормальное распределение остатков
- Гомогенность дисперсий остатков

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)

Задание: Проверьте условия применимости

Проверьте условия применимости дисперсионного анализа используя графики остатков

Решение

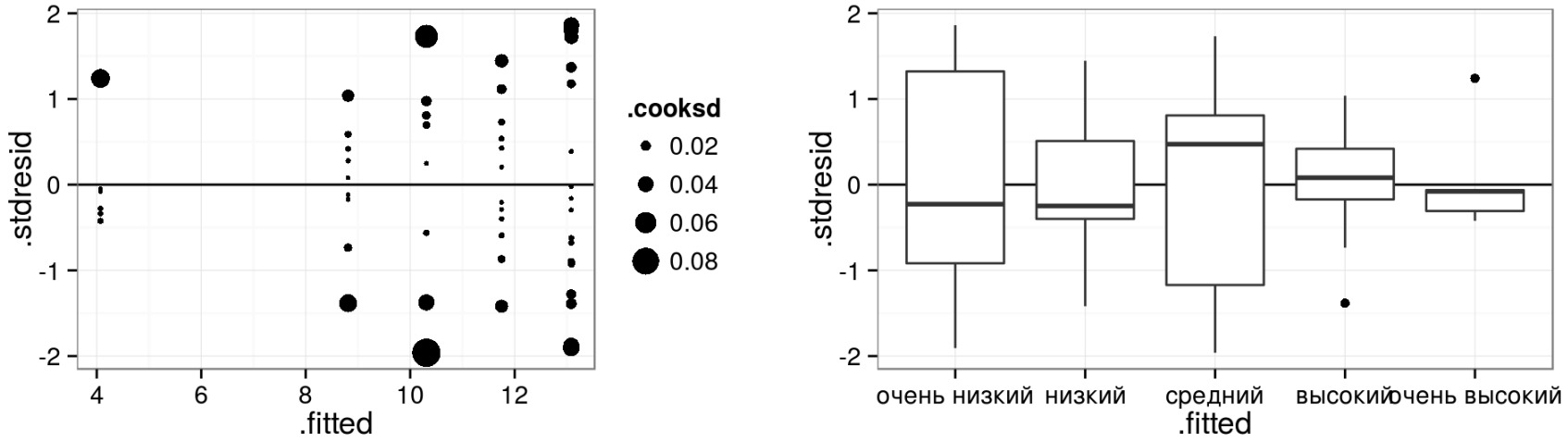
1. Данные для проверки условий применимости на графиках остатков

```
# Данные для анализа остатков
sleep_diag <- fortify(mod)
head(sleep_diag)
```

```
##   TotalSleep       Danger   .hat .sigma  .cooksd .fitted .resid
## 1        3.3      средний 0.1000   3.66 0.085468   10.31 -7.010
## 2        8.3      средний 0.1000   3.79 0.007027   10.31 -2.010
## 3       12.5 очень низкий 0.0556   3.80 0.000299   13.08 -0.583
## 4       16.5      средний 0.1000   3.69 0.066642   10.31  6.190
## 5        3.9      высокий 0.1111   3.73 0.047783    8.81 -4.911
## 6        9.8      высокий 0.1111   3.80 0.001937    8.81  0.989
##   .stdresid
## 1    -1.961
## 2    -0.562
## 3    -0.159
## 4     1.732
## 5    -1.383
## 6     0.278
```

2. Выбросы, гомогенность дисперсий

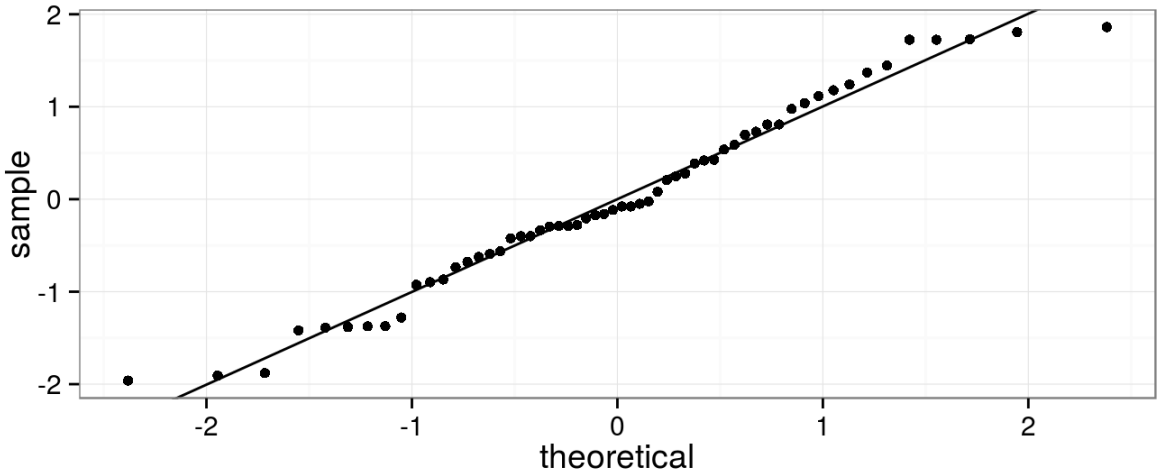
```
gg_res <- ggplot(sleep_diag, aes(x = .fitted, y = .stdresid)) +
  geom_hline(yintercept = 0)
gg_1 <- gg_res + geom_point(aes(size = .cooksd))
gg_2 <- gg_res + geom_boxplot(aes(x = Danger))
library(gridExtra)
grid.arrange(gg_1, gg_2, ncol = 2)
```



- Остатки в пределах двух стандартных отклонений, расстояния Кука маленькие - можно продолжать.
- Подозрительно маленькая дисперсия продолжительности сна в группе с очень высоким уровнем опасности.

3. Нормальность распределения

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

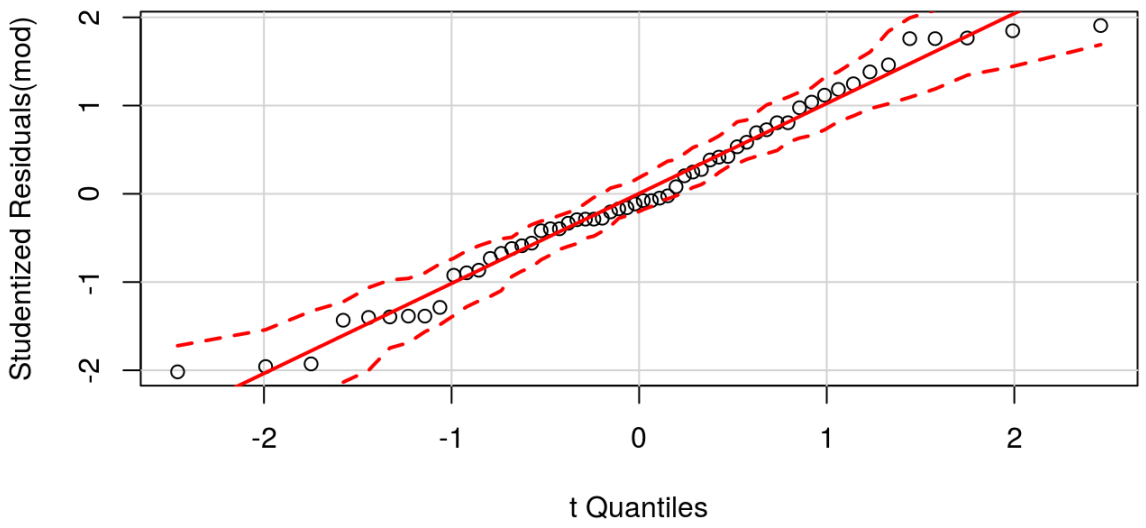


- Распределение практически нормальное

Немного более удобный квантильный график для проверки нормальности распределения

qqPlot() из пакета car

```
qqPlot(mod)
```



- Нет отклонений от нормального распределения

Post hoc тесты

Post-hoc тесты

Дисперсионный анализ показывает, есть ли влияние фактора (= различаются ли средние значения зависимой переменной между группами)

Пост-хок тесты показывают, какие именно из возможных пар средних значений различаются.

Свойства post-hoc тестов для попарных сравнений средних

- Применяются, если влияние фактора значимо
- Делают поправку для снижения вероятности ошибки I рода α, (но не слишком, чтобы не снизилась мощность, чтобы не возросла β)
- Учитывают величину различий между средними
- Учитывают количество сравниваемых пар
- Различаются по степени консервативности (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)
- `Danger = "Tukey"` - тест Тьюки по фактору `Danger`

```
library(multcomp)
sleep_pht <- glht(mod, linfct = mcp(Danger = "Tukey"))
```

Результаты попарных сравнений (тест Тьюки)

```
summary(sleep_pht)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = TotalSleep ~ Danger, data = sleep)
##
## Linear Hypotheses:
##                                    Estimate Std. Error t value Pr(>|t|)
## низкий - очень низкий == 0           -1.33       1.34   -0.99    0.855
## средний - очень низкий == 0          -2.77       1.49   -1.87    0.344
## высокий - очень низкий == 0          -4.27       1.54   -2.78    0.055 .
## очень высокий - очень низкий == 0    -9.01       1.68   -5.37   <0.001 ***
## средний - низкий == 0                -1.44       1.56   -0.92    0.885
## высокий - низкий == 0                -2.94       1.61   -1.83    0.366
## очень высокий - низкий == 0          -7.68       1.74   -4.40   <0.001 ***
## высокий - средний == 0               -1.50       1.73   -0.87    0.907
## очень высокий - средний == 0         -6.24       1.86   -3.36    0.012 *
## очень высокий - высокий == 0         -4.74       1.90   -2.50    0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Описываем результаты пост-хок теста

- Продолжительность сна у видов, подвергающихся очень высокому уровню опасности в течение жизни, значительно меньше, чем у тех, кто живет при среднем, низком и очень низком уровне опасности (тест Тьюки, $p < 0.05$).

Графическое представление результатов пост-хок теста

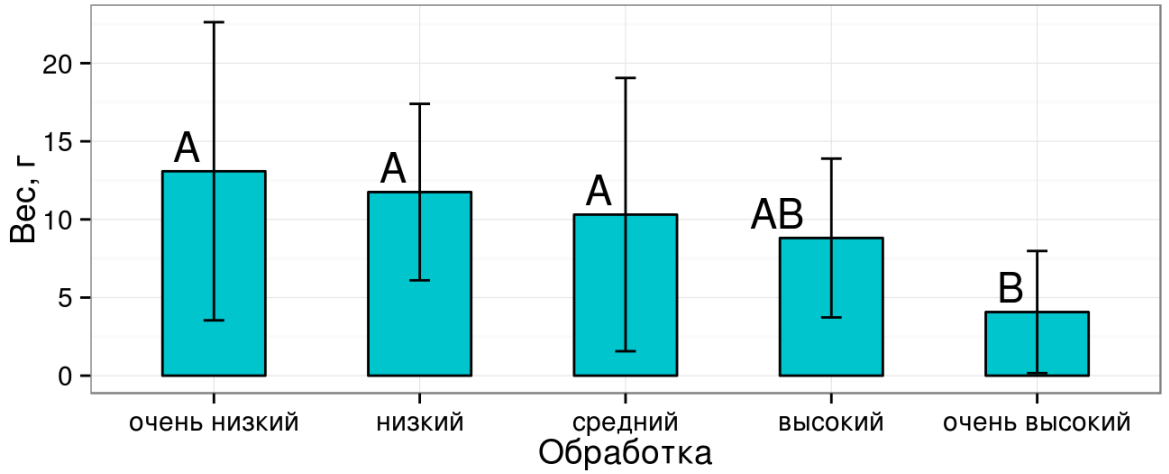
Посчитаем описательную статистику по группам

```
library(dplyr) # есть удобные функции для описания данных
sleep_summary <- sleep %>% # берем датафрейм sleep
  group_by(Danger) %>% # делим на группы по Danger
  # по каждой группе считаем разное
  summarise(
    .n = sum(!is.na(TotalSleep)),
    .mean = mean(TotalSleep, na.rm = TRUE),
    .sd = sd(TotalSleep, na.rm = TRUE),
    .upper_cl = .mean + 1.98*.sd,
    .lower_cl = .mean - 1.98*.sd)
sleep_summary
```

```
## Source: local data frame [5 x 6]
##
##          Danger    .n .mean   .sd .upper_cl .lower_cl
##          (fctr) (int) (dbl) (dbl)     (dbl)     (dbl)
## 1  очень низкий    18 13.08  4.82     22.63     3.540
## 2        низкий    14 11.75  2.85     17.40     6.102
## 3       средний    10 10.31  4.42     19.06     1.564
## 4       высокий     9  8.81  2.57     13.89     3.728
## 5 очень высокий     7  4.07  1.97      7.98     0.162
```

Этот график можно использовать для представления результатов

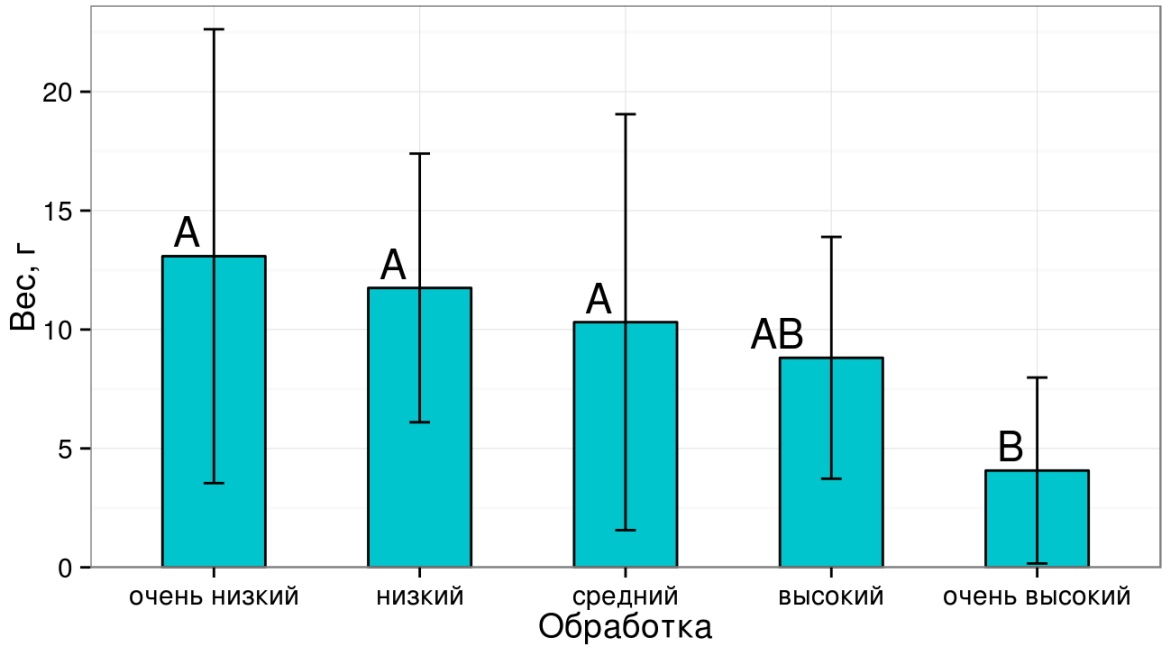
```
gg_means <- ggplot(sleep_summary, aes(x = Danger, y = .mean)) +
  geom_bar(stat = "identity", fill = "turquoise3", colour = "black", width = 0.5) +
  geom_errorbar(aes(ymin = .lower_cl, ymax = .upper_cl), width = 0.1) +
  labs(x = "Обработка", y = "Вес, г") +
  geom_text(aes(label = c("A", "A", "A", "AB", "B"), vjust = -0.3, hjust = 1.5), size = 6)
gg_means
```



- Достоверно различающиеся по пост-хок тесту группы обозначим разными буквами.

Можно "опустить" прямоугольники на ось x

```
upperlimit <- max(sleep_summary$.upper_cl + 1)
gg_means +
  scale_y_continuous(expand = c(0,0),
    limit = c(0, upperlimit))
```



Сохраняем таблицу дисперсионного анализа в файл

```
# 1) в csv
write.csv(sleep_anova, file = "medley_res.csv")

# 2) в xls или xlsx с помощью XLConnect
library(XLConnect)
writeWorksheetToFile(data = sleep_anova, file = "medley_res.xls",
                     sheet = "anova_table")

# или

# 3) отправляем в буфер обмена (только Windows) для вставки в Word-Excel
write.table(file = "clipboard", x = sleep_anova, sep = "\t")
```

Take home messages

- При множественных попарных сравнениях увеличивается вероятность ошибки первого рода. Поправка Бонферрони - способ точно рассчитать, насколько нужно снизить уровень значимости для каждого из сравнений
- При помощи дисперсионного анализа можно проверить гипотезу о равенстве средних значений
- Условия применимости (должны выполняться, чтобы тестировать гипотезы)
- Случайность и независимость групп и наблюдений внутри групп
- Нормальное распределение
- Гомогенность дисперсий
- 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