

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.  
t-тест  
критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест  
Мани—Уитни,  
Уилкоксон  
критерий  $\chi^2$   
критерий Мак Немара  
multiple testing effect

выбор теста

послесловие

# Описательная и классическая статистика в R

Г. Мороз

# Данные

## данные

### описательные статистики

### разведочный анализ данных

### типология

### одновыборочные

#### доверительный инт.

#### t-тест

#### критерий Уилкоксона

#### биномиальный тест

### двухвыборочные

#### t-тест

#### Манн—Уитни,

#### Уилкоксон

#### критерий $\chi^2$

#### критерий Мак Немара

#### multiple testing effect

### выбор теста

### послесловие

- количественные
  - непрерывные
  - дискретные
- номинативные / категориальные

Бывает, имеет смысл перевести количественные данные в категориальные, т. е. составить группы, в которые будут попадать те или иные значения. Как обосновать те или иные границы — дело исследователя. Зная границы, легко узнать, сколько наблюдений в каждой из групп:

```
a <- sin(1:100)
b <- c(-1, -0.5, 0, 0.5, 1)
table(cut(a, breaks = b))
```

# создадим вектор  
# зададим границы

$(-1, -0.5]$	$(-0.5, 0]$	$(0, 0.5]$	$(0.5, 1]$
35	15	16	34

```
table(cut(a, breaks = b), right = F)
```

# переворачивает границы

$[-1, -0.5)$	$[-0.5, 0)$	$[0, 0.5)$	$[0.5, 1)$
35	15	16	34

# Описательные статистики

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послеловие

## ○ средние

- `mean(x)` # среднее арифметическое
- `mean(x, trim = 0.05)` # среднее усеченное
- `weighted.mean(x)` # среднее арифметическое взвешенное
- `1/mean(1/x)` # среднее гармоническое
- `prod(x)**(1/length(x))` # среднее геометрическое
- подробнее о разнице смотрите [stackexchange](#)

## ○ `median(x)`

# медиана

## ○ `range(x), min(x), max(x)`

## ○ `sd(x)`

# среднеквадратическое отклонение

## ○ `quantile(x, 0.25)`

# квантиль

## ○ `IQR(x)`

# IQR

`library("moments")`

## ○ `skewness(x)`

# коэффициент асимметрии

## ○ `kurtosis(x)`

# коэффициент эксцесса

# Описательная статистика

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,

Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

```
v <- c('m', 'f', 'm', 'm', 'f', 'f', 'f', 'f')
df <- data.frame(sex = c('m', 'f', 'm', 'm', 'f', 'f', 'f', 'f'),
hand = c('lf', 'rh', 'rh', 'rh', 'rh', 'lf', 'am', 'rh'))
```

- `table(v)` # частотное распределение
- `table(df)` # таблицы сопряженности
- `prop.table(table(v))`, `prop.table(table(df))` # доля
- `prop.table(table(v))*100`, `prop.table(table(df))*100` # проценты

# NA в выборке

Все функции описательных статистик болезненно относятся к наличию значений NA, поэтому они содержат аргумент `na.rm`, позволяющий игнорировать NA при значении TRUE.

```
x <- c(NA, 4, 2, 3, 2, 9, NA, 9, 4, 5, 2, 4, 7)
mean(x, na.rm = T)
> 4.636364
```

Достаточно легко проверить, есть ли в нашей выборке значения NA при помощи функции `is.na()`:

```
x <- c(NA, 4, 2, 3, 2, 9, NA, 9, 4, 5, 2, 4, 7)
sum(is.na(x))                                     # почему сумма?
> 2
```

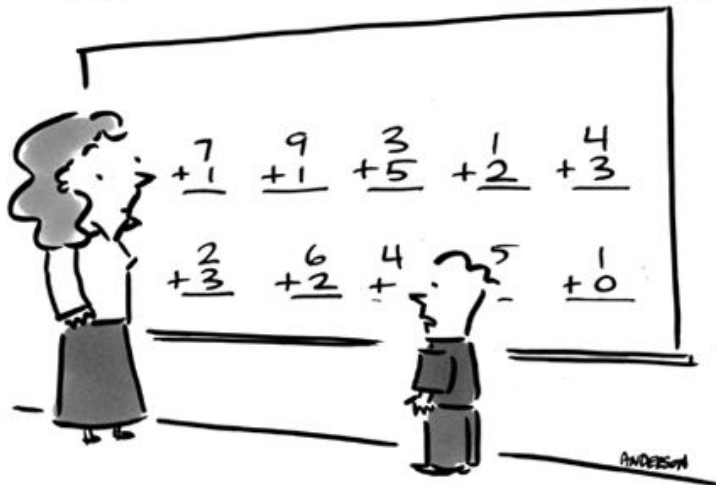
Кроме того, в R есть функция `complete.cases()`, которая позволяет брать только те данные, которые NA не содержат:

```
x <- c(NA, 4, 2, 3, 2, 9, NA, 9, 4, 5, 2, 4, 7)
complete.cases(x)
> [1] FALSE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE
> [13] TRUE
```

# Как будут проходить наши занятия? Увы...

© MARK ANDERSON

WWW.ANDERTOONS.COM

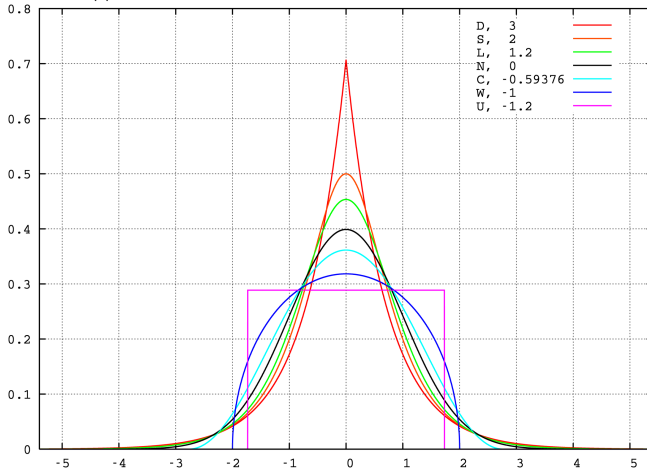


"All I'm saying is we plug these into Excel, let it do its thing, and then we can all play until lunch!"

презентация доступна: <http://goo.gl/kCpxyr>

# Как выглядит распределение?

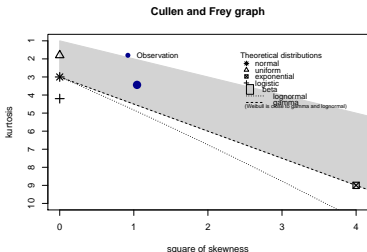
Распределения с разными коэффициентом эксцесса, но одинаковыми стандартным отклонением и средними. Картинка из Википедии:



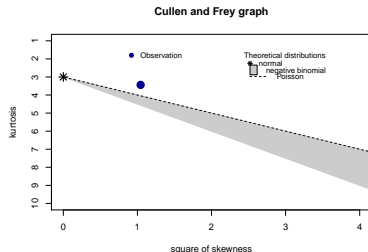
# Как выглядит распределение?

Коэффициенты эксцесса и асимметрии сильно зависят от размера выборки и дают сбой, если выборка не унимодальна. Визуальную оценку типа распределения можно сделать по следующим графикам:

Рис. : library("fitdistrplus")



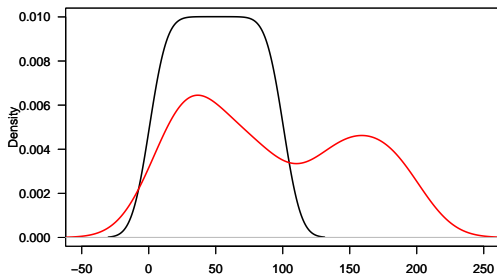
(a) descdist(x)



(b) descdist(x, discrete = T)



# Как сравнивать два распределения?



данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

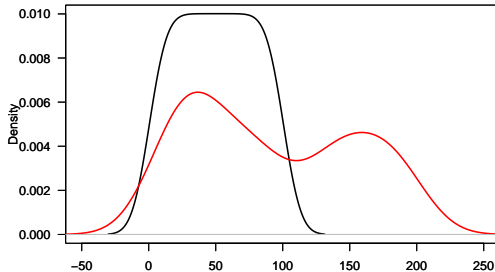
критерий Мак Немара

multiple testing effect

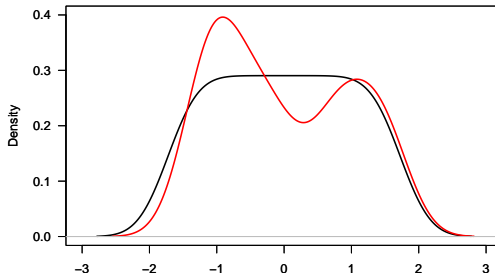
выбор теста

послесловие

# Как сравнивать два распределения?



scale(x)



# z-преобразование

# Garbage in — garbage out

- Данные получить **легко**.
- Скормить полученное компьютеру **тоже легко**.

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

# Garbage in — garbage out

- Данные получить **легко**.

- Скормить полученное компьютеру **тоже легко**.



**Тяжело** помнить, как тот или иной метод работает и какие требования предъявляет к анализируемым данным.

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

# Garbage in — garbage out

- Данные получить **легко**.


- Скормить полученное компьютеру **тоже легко**.



**Тяжело** помнить, как тот или иной метод работает и какие требования предъявляет к анализируемым данным.

⇒ Следует проводить **разведочный анализ данных**.

# Garbage in — garbage out

- Данные получить **легко**.
- Скормить полученное компьютеру **тоже легко**.
-  **Тяжело** помнить, как тот или иной метод работает и какие требования предъявляет к анализируемым данным.

⇒ Следует проводить **разведочный анализ данных**.

- визуализация
- формальные статистические тесты

В некоторых работах по статистике можно встретить предостережения относительно некоторых тестов на нормальность и аргументы в пользу графических методов.

# Garbage in — garbage out

- Данные получить **легко**.
- Скормить полученное компьютеру **тоже легко**.
- 🐱 **Тяжело** помнить, как тот или иной метод работает и какие требования предъявляет к анализируемым данным.

⇒ Следует проводить **разведочный анализ данных**.

- **визуализация**
- **формальные статистические тесты**

В некоторых работах по статистике можно встретить предостережения относительно некоторых тестов на нормальность и аргументы в пользу графических методов.

В работе [Zuur et al. 2010] разработан **протокол разведочного анализа данных**.

# Разведочный анализ данных [Zuur et al. 2010]

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

- **Outliers** *boxplot and Cleveland dotplot*
- **Homogeneity** *conditional boxplot*
- **Normality** *histogram or QQ-plot*
- **Zeros in data** *frequency plot or corrgram*
- **Collinearity** *VIF and scatterplots correlation and PCA*
- **Relationships between variables** *multi-panel scatterplots, conditional boxplots*
- **Interactions** *coplots*
- **Independence** *ACF and varlogram, plot versus time/space*



Statistics are used much like a drunk uses a lamppost: for support, not illumination.

A.E. Housman (commonly attributed to Andrew Lang)

частотная vs. байесовская статистики

A frequentist uses impeccable logic to answer the wrong question, while a Bayesean answers the right question by making assumptions that nobody can fully believe in.

P. G. Hammer

(все так пишут, сам я первоисточника не нашел...)

параметрические vs. непараметрические

одновыборочные vs. двухвыборочные vs. многовыборочные тесты

парные vs. непарные тесты

односторонние vs. двусторонние

презентация доступна: <http://goo.gl/kCpxyr>

# Одновыборочные тесты (one-sample tests)

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

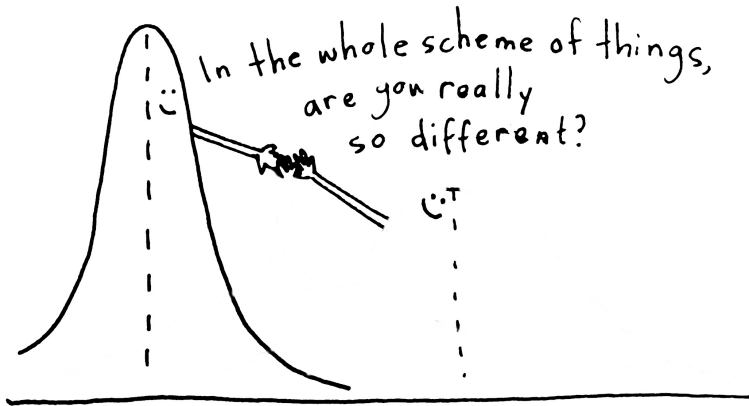
доверительный инт.  
t-тест  
критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест  
Манн—Уитни,  
Уилкоксон  
критерий  $\chi^2$   
критерий Мак Немара  
multiple testing effect

выбор теста

послесловие



## Задача 1: доверительный интервал

У носителей деревни диалектные формы распределены по-разному, у некоторых — много, у некоторых — мало или вообще отсутствуют. Из индивидуальных интервью с  $n$  носителей из середины были взяты по 30 минут и измерялось количество диалектных форм. В среднем в интервью встречается  $\bar{g}$  диалектных черт со стандартным отклонением  $s$ . Что мы можем сказать о средней у всех носителей деревни?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

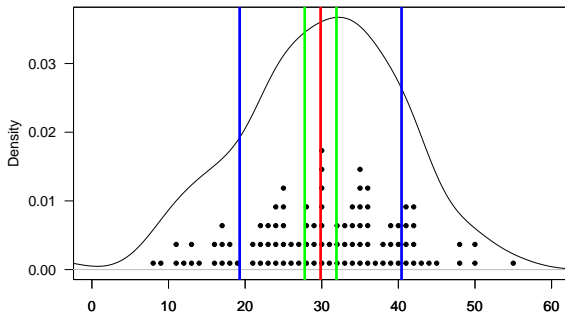
послеловие

# Задача 1: доверительный интервал

У носителей деревни диалектные формы распределены по-разному, у некоторых — много, у некоторых — мало или вообще отсутствуют. Из индивидуальных интервью с  $n$  носителей из середины были взяты по 30 минут и измерялось количество диалектных форм. В среднем в интервью встречается  $\bar{g}$  диалектных черт со стандартным отклонением  $s$ . Что мы можем сказать о средней у всех носителей деревни?

тип данных: количественный  
тип теста: одновыборочный,  
непараметрический,  
непарный

# Доверительный интервал



ДЛЯ  $x > 30$

○ **mean(x)**

**# среднее**

○ **sd(x)**

**# стандартное отклонение**

○ **sd(x)/sqrt(x)**

**# стандартная ошибка среднего**

○ **library("plotrix"); std.error(x)**

**# стандартная ошибка среднего**

○ **mean(x) ± 1.96\*std.error(x)**

**# 95% доверительный интервал**

○ **mean(x) ± 2.58\*std.error(x)**

**# 99% доверительный интервал**

К. Magnusson создал **визуализацию доверительных интервалов**.  
презентация доступна: <http://goo.gl/kCpxyr>

## Задача 2: доверительный интервал

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1.93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Как нам определить, количество информантов  $n$ , которых надо опросить в данной деревне, если мы хотим, чтобы 95% доверительный интервал был меньше 1?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послеловие

## Задача 2: доверительный интервал

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1,93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Как нам определить, количество информантов  $n$ , которых надо опросить в данной деревне, если мы хотим, чтобы 95% доверительный интервал был меньше 1?

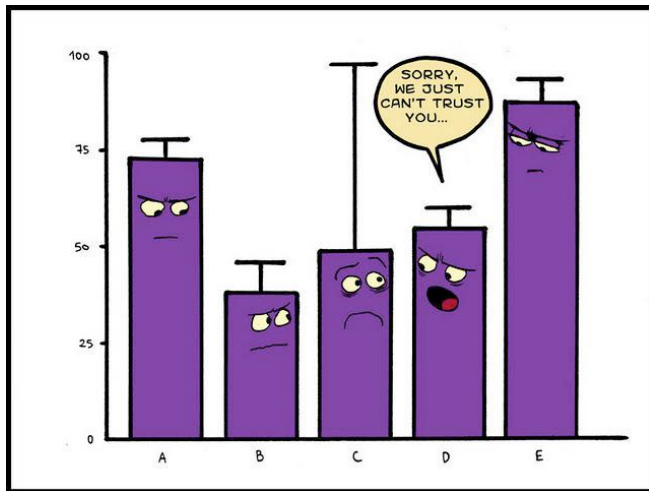
$$CI = \left( 1.96 \times \frac{sd}{\sqrt{n}} \right) \times 2$$

$$n > \left( \left( \frac{1.96 \times sd}{CI} \right) \times 2 \right)^2$$

$$n > 57.2383$$

## Задача 2: доверительный интервал

Чем больше элементов в выборке, тем меньше доверительный интервал.





# Как рисовать доверительные интервалы? R base

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

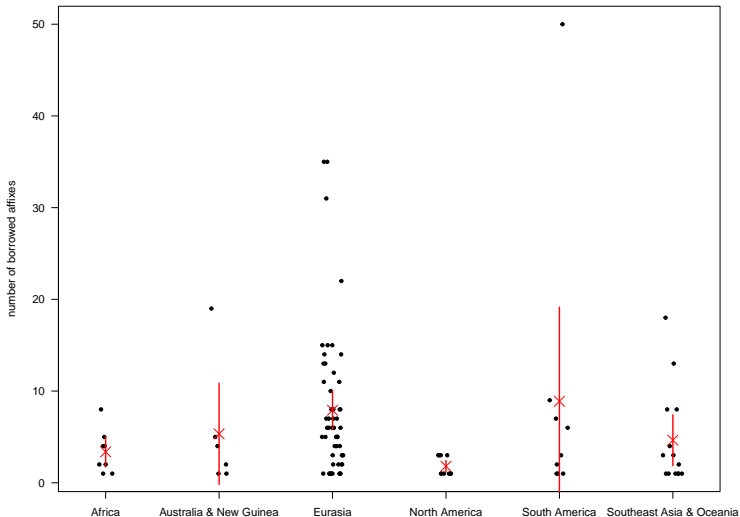
критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие



# Как рисовать доверительные интервалы? R base

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара  
multiple testing effect

выбор теста

послеловие

```
library("plotrix")
a <- read.csv("http://goo.gl/Vlvc5M") # качаем базу данных AfBo
result <- cbind.data.frame(          # создадим дата фрейм
  aggregate(number.of.borrowed.affixes ~ Area, a, mean), # со средним
  aggregate(number.of.borrowed.affixes ~ Area, a, std.error)) # и ст. ошибк.
names(result)[c(2, 4)] <- c("mean", "std.error") # переименуем столбцы

stripchart(a$number.of.borrowed.affixes ~ a$Area, # рисуем данные
  las = 1, pch = 20, method = "jitter", vertical = T,
  ylab = "number of borrowed affixes") # переименуем ось

points(result$mean, pch = 4, cex = 2, col = "red") # рисуем средние
# нарисуем доверительный интервал
segments(x0 = 1:6, x1 = 1:6,
  y0 = result$mean-1.96*result$std.error,
  y1 = result$mean+1.96*result$std.error,
  lwd = 2, col = "red")
```

# Как рисовать доверительные интервалы? ggplot2

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

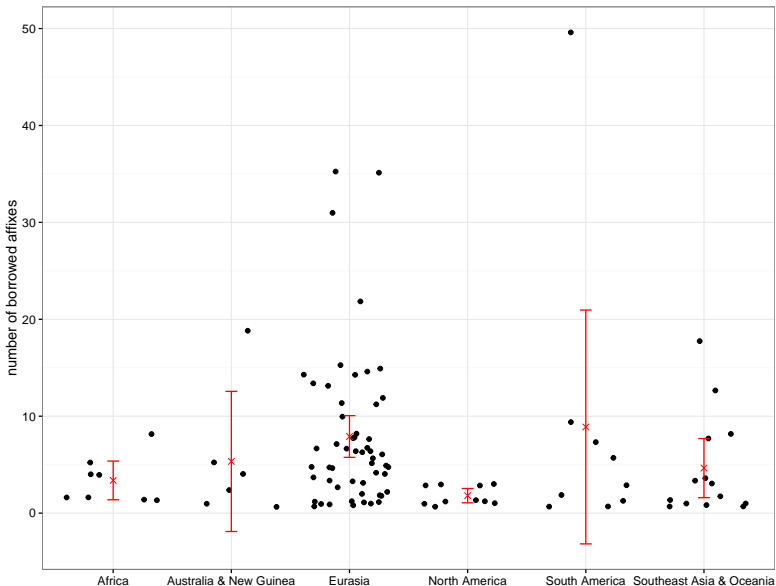
Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара  
multiple testing effect

выбор теста

послесловие



# Как рисовать доверительные интервалы? ggplot2

```
library("plotrix")
library("ggplot2")
a <- read.csv("http://goo.gl/Vlvc5M") # качаем базу данных AfBo
result <- cbind.data.frame(          # создадим дата фрейм
  aggregate(number.of.borrowed.affixes ~ Area, a, mean),      # со средним
  aggregate(number.of.borrowed.affixes ~ Area, a, std.error)) # и ст. ошибк.
names(result)[c(2, 4)] <- c("mean", "std.error") # переименуем столбцы

ggplot(a, aes(a$Area, a$number.of.borrowed.affixes))+
  geom_jitter()+ # нарисуем данные
  xlab() +
  ylab("number of borrowed affixes")+ # переименуем ось
  theme_bw()
stat_summary(fun.y = mean, geom = "point", # рисуем средние
  size = 2, col = "red", shape = 4)+
stat_summary(fun.data = mean_cl_normal,
  geom = "errorbar", # рисуем доверительный интервал
  width = 0.1, col = "red")
```

## Задача 3: одновыборочный t-тест

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1,93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Мы опросили 20 носителей деревни N и выяснили, что средняя равна 4.198775, а стандартное отклонение равно 1.976299. Является ли данная разница статистически значимой?

тип данных: количественный

тип теста: одновыборочный,

требует нормального распределения  
непарный

# Формулировка гипотезы

Классические статистические тесты сравнивают два или более набора данных. Чаще всего:

- строится нулевая гипотеза ( $H_0$ ), о том, что два набора данных могли бы происходить из одной выборки
- строится альтернативная гипотеза ( $H_1$ ), о том, что два набора данных не могли бы происходить из одной выборки
- устанавливается порог статистической значимости (в лингвистике принят порог 5%)
- проводится эксперимент
- а дальше предъявляется p-value — вероятность случайно получить результаты эксперимента (или отличающиеся еще больше), если мы принимаем за истину нулевую гипотезу

# Формулировка гипотезы

"Whatever the approach to formal inference, formalization of the research question as being concerned with aspects of a specified kind of probability model is clearly of critical importance. It translates a subject-matter question into a formal statistical question and that translation must be reasonably faithful and, as far as is feasible, the consistency of the model with the data must be checked. **How this translation from subject-matter problem to statistical model is done is often the most critical part of an analysis.** Furthermore, all formal representations of the process of analysis and its justification are at best idealized models of an often complex chain of argument".

[Cox 2006: 197]

# Как интерпретировать p-value?

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
$\geq 0.1$	

If all else fails, use "significant at a  $p > 0.05$  level" and hope no one notices.

Комикс xkcd p-values. Объяснение.

презентация доступна: <http://goo.gl/kCpxyr>



## Задача 3: одновыборочный t-тест

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1,93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Мы опросили 20 носителей деревни N и выяснили, что средняя равна 4.198775, а стандартное отклонение равно 1.976299. Является ли данная разница статистически значимой?

```
t.test(a, mu = 5.31)      # первое — вектор значений, второе — среднее
```

One Sample t-test

data: a

```
t = -2.5146, df = 19, p-value = 0.02108
```

alternative hypothesis: true mean is not equal to 5.31

95 percent confidence interval:

3.273838 5.123711

sample estimates:

mean of x

4.198775

# Задача 4: нормально ли распределение?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Мани—Уитни,

Уилкоксон

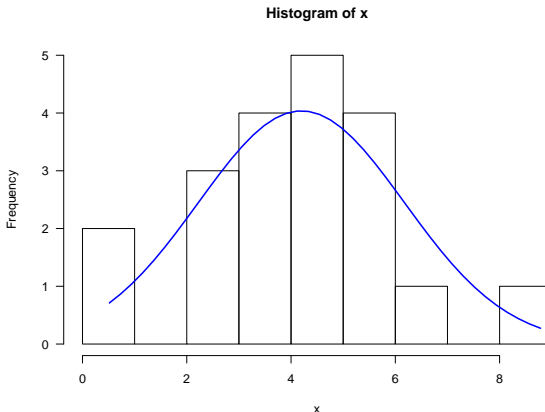
критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

последствие



```
h <- hist(x, las = 1)
xfit <- seq(min(x),max(x),length=40)
yfit <- dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue lwd=2)
```

презентация доступна: <http://goo.gl/kCpxyr>

```
# записывает параметры
# создает выборку
# получает параметры

# рисует результат
```

# Задача 4: нормально ли распределение?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.  
**t-тест**

критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест

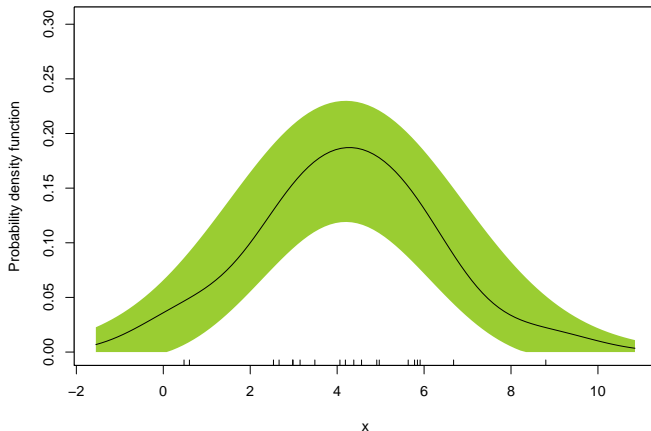
Мани—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара  
multiple testing effect

выбор теста

послесловие



```
library(sm)
sm.density(x, model = "Normal col.band="yellowgreen")
```

презентация доступна: <http://goo.gl/kCpxyr>

# Задача 4: нормально ли распределение?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

**t-тест**

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Мани—Уитни,

Уилкоксон

критерий  $\chi^2$

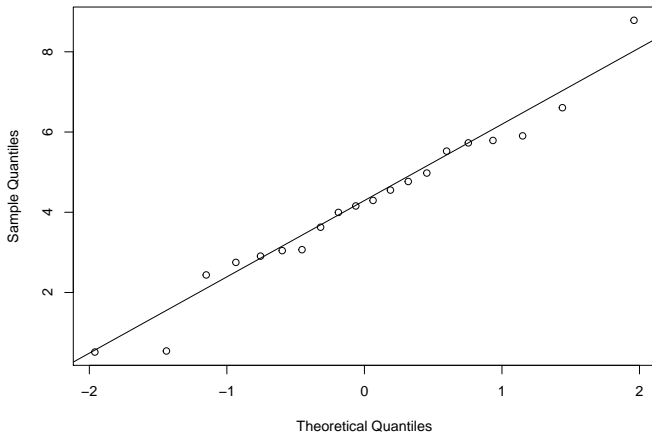
критерий Мак Немара

multiple testing effect

выбор теста

послесловие

Normal Q-Q Plot



`qqplot(x)`

`qqline(x)`

презентация доступна: <http://goo.gl/kCpxyr>

## Задача 4: нормально ли распределение?

Критерий Шапиро-Уилка:

если наблюдений  $< 60$

$H_0$ : данные распределены нормально

$H_1$ : данные не распределены нормально

`shapiro.test(x)`

Shapiro-Wilk normality test

data: x

$W = 0.9718$ ,  $p\text{-value} = 0.7923$

Одновыборочный критерий Колмогорова-Смирнова:

$> 60$

`ks.test(x, "pnorm")`

One-sample Kolmogorov-Smirnov test

data: x

$D = 0.12647$ ,  $p\text{-value} = 0.0816$

alternative hypothesis: two-sided

# Гомоскедастичность (гомогенность) дисперсии

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

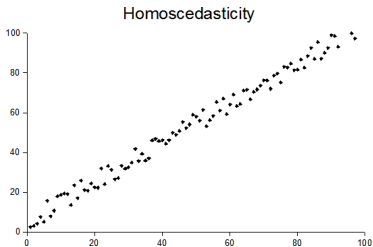
критерий  $\chi^2$

критерий Мак Немара

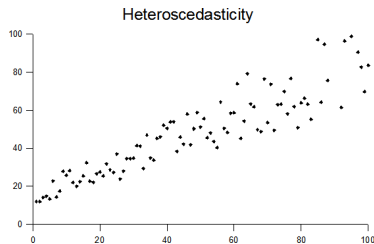
multiple testing effect

выбор теста

послеловие



(с) гомоскедастичное



(d) гетерогедастичное

распределения из Википедии

Гомоскедастичность можно проверить тестом Бартлетта:

`bartlett.test(m, n)`

Bartlett test of homogeneity of variances

data: m, n

Bartlett's K-squared = 2.0949, df = 1, p-value = 0.1478

## Задача 5: сколько нужно наблюдений?

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1.93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Как нам определить, количество информантов  $n$ , которых надо опросить в данной деревне, если мы хотим, чтобы мы могли наблюдать разницу в 1 слог с вероятностью совершить ошибку первого рода  $\alpha 0.05$  и мощностью теста 0.8?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,

Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послеловие

## Задача 5: сколько нужно наблюдений?

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1.93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Как нам определить, количество информантов  $n$ , которых надо опросить в данной деревне, если мы хотим, чтобы мы могли наблюдать разницу в 1 слог с вероятностью совершить ошибку первого рода  $\alpha$  0.05 и мощностью теста 0.8?

```
power.t.test(sig.level = 0.05,  
power = 0.8,  
delta = 1,  
sd = 1.93,  
type = "one.sample",  
alternative = "one.sided")
```

```
One-sample t test power calculation  
n = 24.44055
```

...

```
#  $\alpha$   
# мощность теста  
# наблюдаемая разница  
# стандартное отклонение
```



# Задача 6: выборка не распределена нормально?

```
wilcox.test(x, mu = 5,31)
```

Wilcoxon rank sum test

data: x and 31

$W = 0$ ,  $p\text{-value} = 0.04878$

alternative hypothesis: true location shift is not equal to 5

**тип данных: количественный**

**тип теста: одновыборочный,  
непараметрический,  
непарный**

## Задача 7: биномиальный тест

В частотном словаре [Ляшевская, Шаров 2009], созданном на корпусе объемом 92 млн. словоупотреблений, существительное *кенгуру* имеет абсолютную частотность 0.0000021, а предлог *к* — 0.005389 (его алломорф *ко* в расчет не берется). В некотором тексте, имеющем 61981 слов существительное *кенгуру* встречается 58 раз, а предлог *к* — 254. Каковы вероятности получить такие результаты?

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

**биномиальный тест**

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

## Задача 7: биномиальный тест

В частотном словаре [Ляшевская, Шаров 2009], созданном на корпусе объемом 92 млн. словоупотреблений, существительное *кенгуру* имеет абсолютную частотность 0.0000021, а предлог *к* — 0.005389 (его алломорф *ко* в расчет не берется). В некотором тексте, имеющем 61981 слов существительное *кенгуру* встречается 58 раз, а предлог *к* — 254. Каковы вероятности получить такие результаты?

```
binom.test(58, size = 61981, p = 0.0000021)
binom.test(58, 61981, 0.0000021)
binom.test(254, 61981, 0.005389)
```

```
# про кенгуру
# про кенгуру
# про к
```

тип данных: категориальный  
тип теста: одновыборочный,  
непараметрический,  
непарный

# Двухвыборочные тесты (two-sample tests)

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

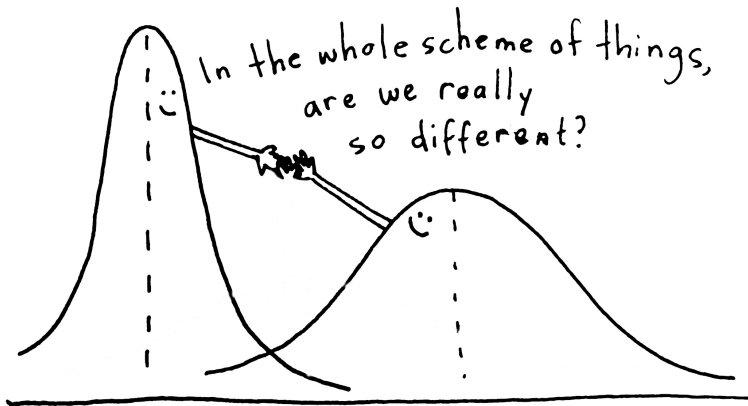
доверительный инт.  
t-тест  
критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест  
Манн—Уитни,  
Уилкоксон  
критерий  $\chi^2$   
критерий Мак Немара  
multiple testing effect

выбор теста

послесловие



## Задача 8: двухвыборочный t-тест

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1,93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Является ли данная разница между мужчинами и женщинами статистически значимой?

тип данных: количественный

тип теста: двухвыборочный,

требует нормального распределения и гомоскедастичности  
непарный

## Задача 8: двухвыборочный t-тест

Из статьи С. Степановой мы знаем, что носители русского языка в среднем говорят 5.31 слога в секунду со стандартным отклонением 1,93 (мужчины 5.46 слога в секунду со средним отклонением 2.02, женщины 5.23 слога в секунду со средним отклонением 1.84, дети 3.86 слога в секунду со средним отклонением 1.67). Является ли данная разница между мужчинами и женщинами статистически значимой?

`t.test(a, b)`

# первое и второе — векторы значений

Welch Two Sample t-test

data: a and b

`t = 0.38408, df = 37.919, p-value = 0.7031`

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.9984148 1.4659358

# CI для величины эффекта

sample estimates:

mean of x mean of y

5.465256 5.231496

## Задача 9: парный t-тест

Во время экспедиции от 20 информантов (10 мужчин, 10 женщин) были записаны список слов, а потом были измерены длительности ударных гласных. Получилось, что у мужчин средняя длительность гласного 93 миллисекунды, а у женщин — 107 миллисекунд. Является ли данная разница между мужчинами и женщинами статистически значимой?

тип данных: количественный

тип теста: двухвыборочный,

требуется нормального распределения и гомоскедастичности  
парный

## Задача 9: парный t-тест

Во время экспедиции от 20 информантов (10 мужчин, 10 женщин) были записаны список слов, а потом были измерены длительности ударных гласных. Получилось, что у мужчин средняя длительность гласного 95 миллисекунды ( $sd = 32$ ), а у женщин — 104 миллисекунды ( $sd = 45$ ). Является ли данная разница между мужчинами и женщинами статистически значимой?

```
t.test(m, f, paired = T)
```

Paired t-test

data: m and f

$t = -2.3743$ ,  $df = 199$ ,  $p\text{-value} = 0.01853$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-16.208597 -1.500364

# CI для величины эффекта

sample estimates:

mean of the differences

-8.85448



# Парные тесты

Парные тесты применяют, если исследуют:

- наблюдения до/после (измерения скорости речи при первом рассказе и при пересказе)
- наблюдения одного и того же объекта, полученные разными методами (например, кроссвалидация разметки)

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послесловие

# Задача 9: выборки не распределена нормально?

`wilcox.test(a, b)`

# критерий Манна—Уитни

Wilcoxon rank sum test

data: a and b

W = 243, p-value = 0.2534 alternative hypothesis: true location shift is not equal to 0

`wilcox.test(c,d, paired = T)`

# критерий Уилкоксона

Wilcoxon signed rank test

data: c and d

V = 80, p-value = 0.3683

alternative hypothesis: true location shift is not equal to 0

Если добавить аргумент `conf.int = T`, то отобразится еще и 95% доверительный интервал, в котором находится величина эффекта.

тип данных: количественный

тип теста: двухвыборочный,

непараметрический,

непарный, парный

<http://goo.gl/kCpxyr>

## Задача 10: $\chi^2$ (с поправкой Йейтса)

Из интервью с носителями одной деревни произвольным образом выбрали по пол часа и посчитали кол-во реализаций диалектных форм vs. недиалектных. В результате получилось что у женщин было 107 диалектных форм vs. 93 недиалектные, а у мужчин — 74 vs. 45.

тип данных: категориальный  
тип теста: двухвыборочный,  
непараметрический,  
непарный

## Задача 10: $\chi^2$ (с поправкой Йейтса)

Из интервью с носителями одной деревни произвольным образом выбрали по пол часа и посчитали кол-во реализаций диалектных форм vs. недиалектных. В результате получилось что у женщин было 107 диалектных форм vs. 93 недиалектные, а у мужчин — 74 vs. 45. Значима ли зафиксированная разница?

Сначала следует составить таблицу сопряженности:

`table(dialect)`

# таблица сопряженности

sex	feature	
	-d	+d
f	107	93
m	74	45

А дальше используем тест:

`chisq.test(table(dialect))`

Pearson's Chi-squared test with Yates' continuity correction

data: table(dialect)

X-squared = 1.9525, df = 1, p-value = 0.1623

# Критерий Фишера, критерий Крамера

Иногда R может сказать:

Chi-squared approximation may be incorrect

Критерий  $\chi^2$  плохо использовать:

- если хотя бы в одной из клеток есть значения меньше 5  
→ тест Фишера (`fisher.test()`)
- если между числами есть большой разрыв<sup>1</sup>  
→ проверяем величину эффекта критерием Крамера (`cramersV()` в пакете `lsr`)
- ...вообще таблицы сопряженности бывают разные, да и тестов куда больше см. [Lydersen et al. 2009]

---

<sup>1</sup> "All differences are significant with a large enough sample size"

## Задача 11: критерий Мак Немара

Во время диалектологической экспедиции от 20 информантов (10 мужчин, 10 женщин) были записаны списки слов. Получилось, что мужчины произносят велярный фрикативный  $\gamma$  в 13 случаях, а велярный стоп  $g$  в 43. У женщин получилось другое распределение: 19  $\gamma$  против 37  $g$ . Является ли данная разница между мужчинами и женщинами статистически значимой?

тип данных: категориальный

тип теста: двухвыборочный,  
непараметрический,  
парный

## Задача 11: критерий Мак Немара

Во время диалектологической экспедиции от 20 информантов (10 мужчин, 10 женщин) были записаны списки слов. Получилось, что мужчины произносят велярный фрикативный  $\gamma$  в 13 случаях, а велярный стоп  $g$  в 43. У женщин получилось другое распределение: 19  $\gamma$  против 37  $g$ . Является ли данная разница между мужчинами и женщинами статистически значимой?

Сначала следует составить таблицу сопряженности:

`table(stopfricg)`

# таблица сопряженности

	feature	
sex	fric	stop
f	21	35
m	13	43

А дальше используем тест:

`mcnemar.test(table(stopfricg))`

McNemar's Chi-squared test with continuity correction

data: table(stopfricg)

McNemar's chi-squared = 9.1875, df = 1, p-value = 0.002437

презентация доступна: <http://goo.gl/kCpxyr>

# Об эффекте множественных сравнений

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона

биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

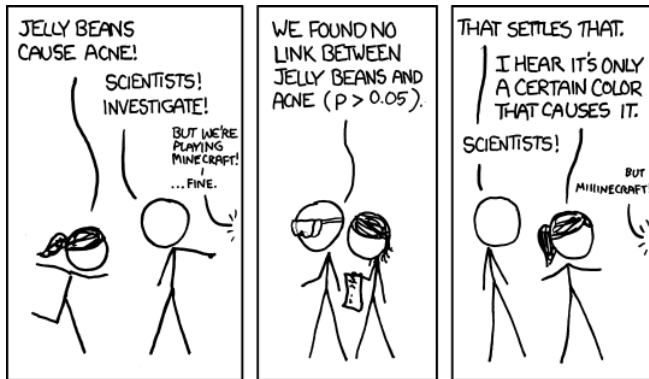
критерий  $\chi^2$

критерий Мак Немара

multiple testing effect

выбор теста

послеловие





# Об эффекте множественных сравнений

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

доверительный инт.

t-тест

критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест

Манн—Уитни,  
Уилкоксон

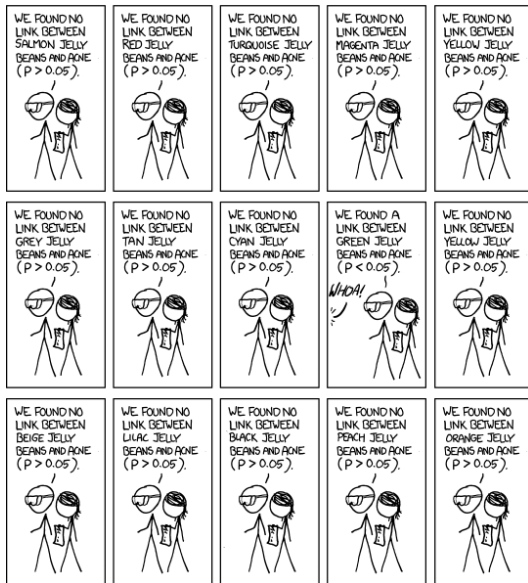
критерий  $\chi^2$

критерий Мак Немара

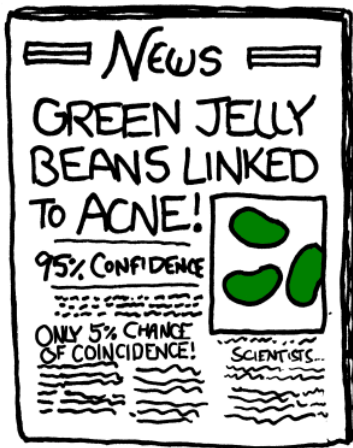
multiple testing effect

выбор теста

послеловие



# Об эффекте множественных сравнений



Комикс xkcd Significant. Объяснение.

Это называют: data dredging, data fishing, data snooping, equation fitting, p-hacking...

презентация доступна: <http://goo.gl/kCpxyr>

# Об эффекте множественных сравнений

При проверке каждой статистической гипотезы закладывается возможность ошибки первого рода (т. е. отклонение верной нулевой гипотезы). Чем больше гипотез мы проверяем на одних и тех же данных, тем больше будет вероятность допустить как минимум одну такую ошибку. Вероятность того, что из 21 теста (включая первый тест, без исключения цвета) не будет допущена ошибка первого рода равна

$$P = (1 - \alpha)^m = (1 - 0.05)^{21} = 0.34$$

# Многовыборочные тесты (multiple-sample tests)

данные

описательные  
статистики

разведочный  
анализ данных

типология

одновыборочные

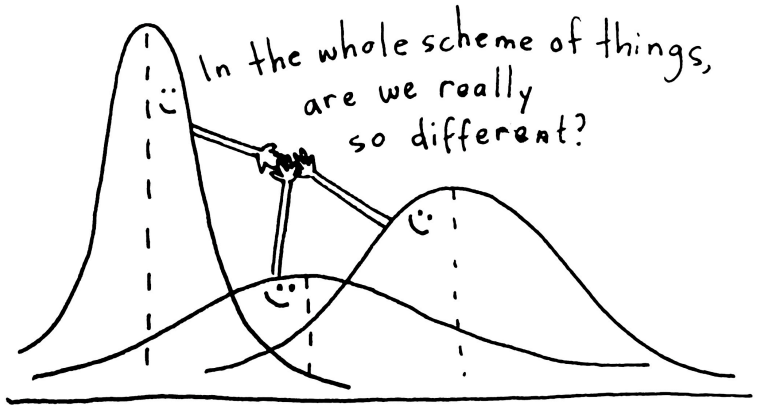
доверительный инт.  
t-тест  
критерий Уилкоксона  
биномиальный тест

двухвыборочные

t-тест  
Мани—Уитни,  
Уилкоксон  
критерий  $\chi^2$   
критерий Мак Немара  
**multiple testing effect**

выбор теста

послесловие



# Выбор теста

тип данных и распределение  
тип группы

количество групп

тест

норм.	с заданным значением	1	одновыборочный t-test
	независимые	2	t-test для независимых выборок
	зависимые	2	парный t-test
не норм.	с заданным значением	1	критерий Уилкоксона
	независимые	2	критерий Манна-Уитни
	зависимые	2	критерий Уилкоксона
категор.	с заданным значением	1	биномиальный тест
	независимые	2	$\chi^2$ с поправкой Йейтса, Фишер, Крамер
	зависимые	2	критерий Мак-Нимора

Если количество групп превышает 2, то с используют многовыборочные тесты: ANOVA (и всякие варианты ANCOVA, MANOVA, MANCOVA), критерии Краскела-Уоллиса, критерий Фридмана, Q-критерий Кокрена и  $\chi^2$  с поправкой на правдоподобие.

# Величина эффекта (effect size)

В статье [Sullivan, Feinn 2012] приводится ряд аргументов в пользу того, что следует приводить не только  $p$ -value, но и величину эффекта:

- величина эффекта — основной результат количественного исследования,  $p$ -value лишь говорит о том, что эффект с некоторой вероятностью есть
- при работе со значительными выборками статистические тесты всегда будут давать статистическую значимость, даже если величина эффекта незначительна

# p-value очень много ругают

- за то, что его очень часто понимают неправильно [Gigerenzer 2004], [Goodman 2008]
- за то, что само по себе  $p\text{-value} < 0.05$  слабый довод [Sterne, Smith 2001], [Nuzzo et al. 2014], [Wasserstein, Lazar 2016]

*Q: Why do so many colleges and grad schools teach  $p = 0.05$ ?*

*A: Because that's still what the scientific community and journal editors use.*

*Q: Why do so many people still use  $p = 0.05$ ?*

*A: Because that's what they were taught in college or grad school*

[Wasserstein, Lazar 2016]

В связи с этим, сейчас можно наблюдать

- большое обсуждение  $p\text{-value}$  vs. доверительные интервалы
- все нарастающую популярность Байесовской статистики

"Есть жизнь" и вне Пирсоновской и Байесовской статистик.

презентация доступна: <http://goo.gl/kCpxyr>

# Спасибо за внимание

Пишите письма  
[agricolamz@gmail.com](mailto:agricolamz@gmail.com)



# Список литературы

- Cox, David Roxbee (2006). *Principles of statistical inference*. Cambridge University Press.
- Gigerenzer, Gerd (2004). Mindless statistics. *The Journal of Socio-Economics* 33(5), 587--606.
- Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. 45(3), 135--140.
- Lydersen, Stian, Morten W Fagerland, Petter Laake (2009). Recommended tests for association in  $2 \times 2$  tables. *Statistics in medicine* 28(7), 1159--1175.
- Nuzzo, Regina et al. (2014). Statistical errors. *Nature* 506(7487), 150--152.
- Sterne, J. A. C., G. D. Smith (2001). Sifting the evidence—what's wrong with significance tests? *Physical Therapy* 81(8), 1464--1469.
- Sullivan, G. M., R. Feinn (2012). Using effect size-or why the p value is not enough. *Journal of graduate medical education* 4(3), 279--282.
- Wasserstein, R., L., N. A. Lazar (2016). The asa's statement on p-values: context, process, and purpose. *The American Statistician* 70, ???--??
- Zuur, Alain F, Elena N Ieno, Chris S Elphick (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1), 3--14.
- Ляшевская, О. Н., С. А. Шаров (2009). *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. Азбуковник.