

Занятие 3

Факторы – Файлы - Статистика

23 сентября 2015

Виноградова Светлана

План

- Факторы
- Работа с файлами
- Элементарная статистика

Факторы

Факторы

Используются для представления категориальных данных (да/нет, низкий/средний/высокий, мужчина/женщина...)

```
> f <- factor(c("yes", "yes", "no", "yes", "no"))  
> f
```

```
[1] yes yes no yes no  
Levels: no yes
```

```
> levels(f) # возможные значения в факторе  
[1] "no" "yes"
```

```
> levels(f) <- c(levels(f), "maybe")
```

```
> table(f)
```

```
f  
no  yes  maybe  
 2    3     0
```

Факторы

Уровни можно упорядочивать при создании фактора (может быть важно в линейной регрессии):

```
> f <- factor(c("yes", "yes", "no", "yes",  
"no"), levels = c("yes", "no"))
```

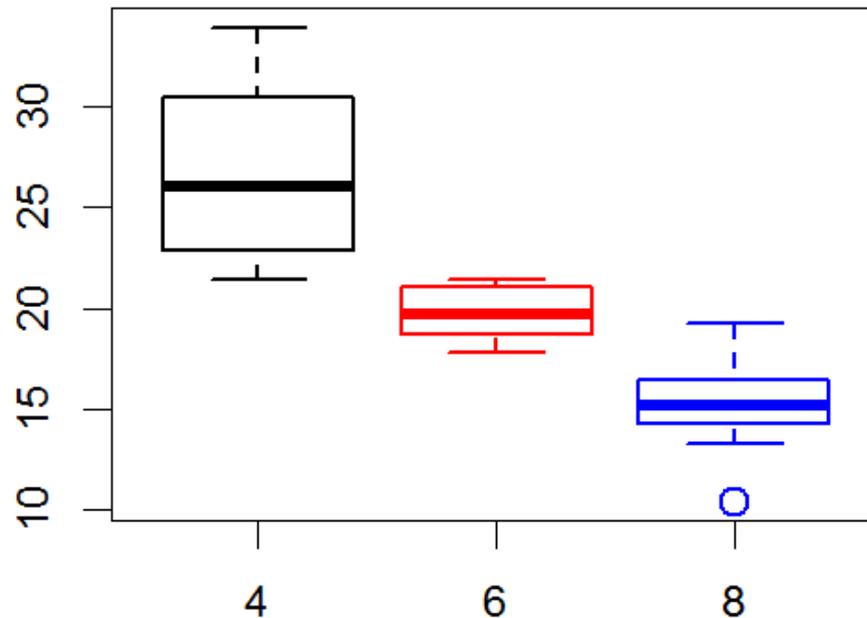
```
> f  
[1] yes yes no yes no  
Levels: yes no
```

(по умолчанию, уровни в факторе упорядочиваются в лексикографическом порядке)

Факторы

Разбиение вектора по фактору:

```
> boxplot(mtcars$mpg ~ mtcars$cyl)
```



```
mpg: 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 ...  
cyl: 6 6 4 6 8 6 8 4 ...
```

Работа с файлами

Работа с файлами: основные функции

Чтение	Запись	Применение
<i>read.table</i>	<i>write.table</i>	Чтение/запись табулированных текстовых файлов
<i>read.csv</i>	<i>write.csv</i>	Чтение/запись файлов в формате CSV
<i>readLines</i>	<i>writeLines</i>	Чтение/запись текстовых файлов по строкам
<i>load</i>	<i>save</i>	Загрузка/сохранение объектов R из/в бинарные файлы (.RData)

Работа с файлами: рабочая директория

Узнать рабочую директорию:

```
> getwd()
```

```
[1] "C:/Users/anna/FBB/R"
```

Поменять рабочую директорию:

```
> setwd("week3") # путь указан относительно рабочей директории!
```

```
> getwd()
```

```
[1] "C:/Users/anna/FBB/R/week3"
```

Узнать список файлов в рабочей директории

```
> dir()
```

Узнать список файлов в указанной директории

```
> dir("C:/Users/anna/FBB/R/")
```

В RStudio:

закладка Files (справа внизу) -> выбрать нужную директорию -> More -> Set As Working Directory

Работа с файлами: *read.table*

- Читает файл с разделителями
- Возвращает *data.frame*

```
> students <- read.table("FBBRStudents.tab", sep="\t",  
header=T)
```

```
> students[101:102,]
```

	Name	Faculty	Level	Year
101	Широкий В. Р.	химический	специалитет	4
102	Базылев С. С.	биологический	бакалавриат	1

Работа с файлами: *read.table*

Основные аргументы:

- **file** – имя файла или соединение (connection)
- **header** – есть ли в файле заголовок? (по умолчанию, FALSE)
- **sep** – разделитель полей (колонок) (по умолчанию, пробел)
- **colClasses** – вектор с названиями классов колонок
- **nrows** – количество строчек, которые нужно прочитать
- **skip** – количество строчек, которые нужно пропустить
- **comment.char** – знак комментариев
- **stringsAsFactors** – преобразовывать строковые поля в фактор? (по умолчанию, TRUE)

Работа с файлами: *read.table*

```
> students<-read.table("FBBRStudents.tab", sep="\t", header=T,  
+ colClasses = c("character", "factor", "factor", "integer"))
```

```
> str(students)
```

```
'data.frame': 141 obs. of 4 variables:
```

```
 $ Name : chr "Антонов С. В." "Дмитриев Д. И." "Золотов И.  
А." "Иванова Т. В." ...
```

```
 $ Faculty: Factor w/ 10 levels "биологический",...: 3 3 3 3  
3 3 3 3 3 3 ...
```

```
 $ Level : Factor w/ 3 levels "бакалавриат",...: 3 3 3 3 3 3  
3 3 3 3 ...
```

```
 $ Year : int 3 3 3 3 4 4 4 4 4 4 ...
```

Работа с файлами: *read.csv*, *write.csv*, *readLines*

- ***read.csv*** – то же, что `read.table`, но с другими дефолтными значениями параметров (`header=TRUE`, `sep=","`)

- ***write.csv***:

```
> write.csv(students, "FBBRStudents.csv")
```

- ***readLines***:

```
> lines <- readLines("FBBRStudents.txt", 3)
```

```
> lines
```

```
[1] "Name\tFaculty\tLevel\tYear"
```

```
[2] "Антонов С. В.\tмеханико-математический\tспециалитет\t3"
```

```
[3] "Дмитриев Д. И.\tмеханико-математический\tспециалитет\t3"
```

Работа с файлами: *save, load*

Сохраняем объекты *students* и *lines* в файл:

```
> save(students, lines, file="Students.RData")
```

Удаляем все объекты из рабочего пространства:

```
> rm(list=ls())
```

```
> ls()
```

```
character(0)
```

Загружаем объекты из файла:

```
> load("Students.RData")
```

```
> ls()
```

```
[1] "lines" "students" # объекты появляются в  
# рабочем пространстве
```

Соединения

- **file** – открывает соединение с файлом
- **gzfile, bzfile** – открывает соединение с архивированным файлом
- **url** – открывает соединение с веб-страницей

```
> con <- file("FBBRStudents.txt", "r")
```

```
> readLines(con, 1)
```

```
[1] "Name\tFaculty\tLevel\tYear"
```

```
> readLines(con, 1)
```

```
[1] "Антонов С. В.\tмеханико-математический\tспециалитет\t3"
```

```
> close(con)
```

```
> con <- gzfile("FBBRStudents.gz")
```

```
> read.csv(con, nrow=2)
```

```
X Name Faculty Level Year
```

```
1 1 Антонов С. В. механико-математический специалитет 3
```

```
2 2 Дмитриев Д. И. механико-математический специалитет 3
```

```
> close(con)
```

Элементарная статистика

Эксперимент:

как отличить «честную» монетку от «нечестной»?

Честная монетка: вероятность орла **0.5**, вероятность решки **0.5**

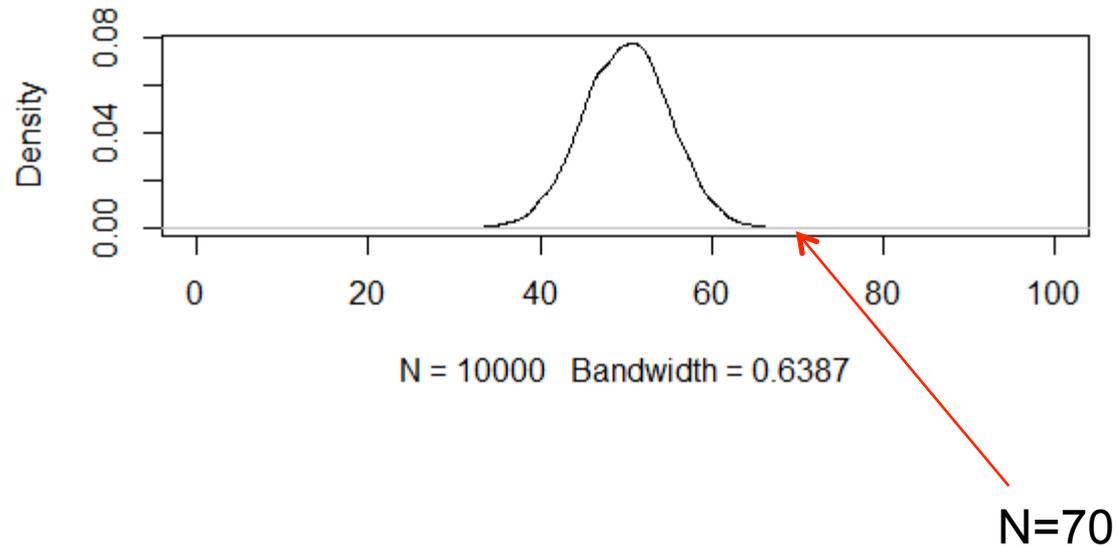
Нечестная монетка: вероятность орла **0.2**, вероятность решки **0.8**

Подбросим монетку 100 раз.

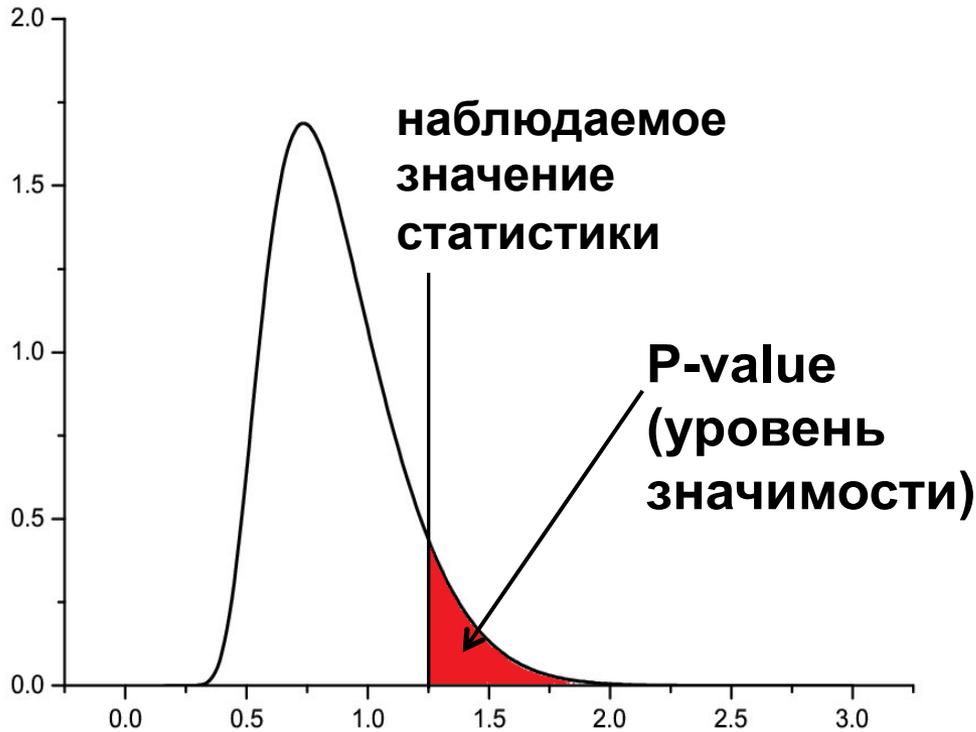
Решка выпала **70** раз. Какая у нас монетка?

Насколько можно быть уверенным в этом?

Распределение частот выпадения решки у честной монеты
(биномиальное распределение):



H_0 – нулевая гипотеза: мы кидали честную монету
 H_1 – альтернативная гипотеза: монета кривая



Что такое P-value?

- ✓ Вероятность наблюдаемого при нулевой гипотезе
- ✓ Вероятность ошибочно отвергнуть нулевую гипотезу (когда она верна)

Не строгие математические определения, главное – понять смысл!

Данные: вес цыплят в зависимости от рациона питания

```
> chick.w <- read.table("ChickWeight.tab", header=T)
```

```
> dim(chick.w)
[1] 20 2
```

```
> head(chick.w)      # weight - вес цыпленка (в граммах)
  weight Diet      # Diet - тип рациона (2 или 3)
```

```
232    331    2
244    167    2
256    175    2
268     74    2
280    265    2
292    251    2
```

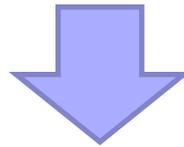
```
> tail(chick.w,3)
  weight Diet
436    290    3
448    272    3
460    321    3
```

Задача:

понять, влияет ли рацион на вес

Вопрос №1

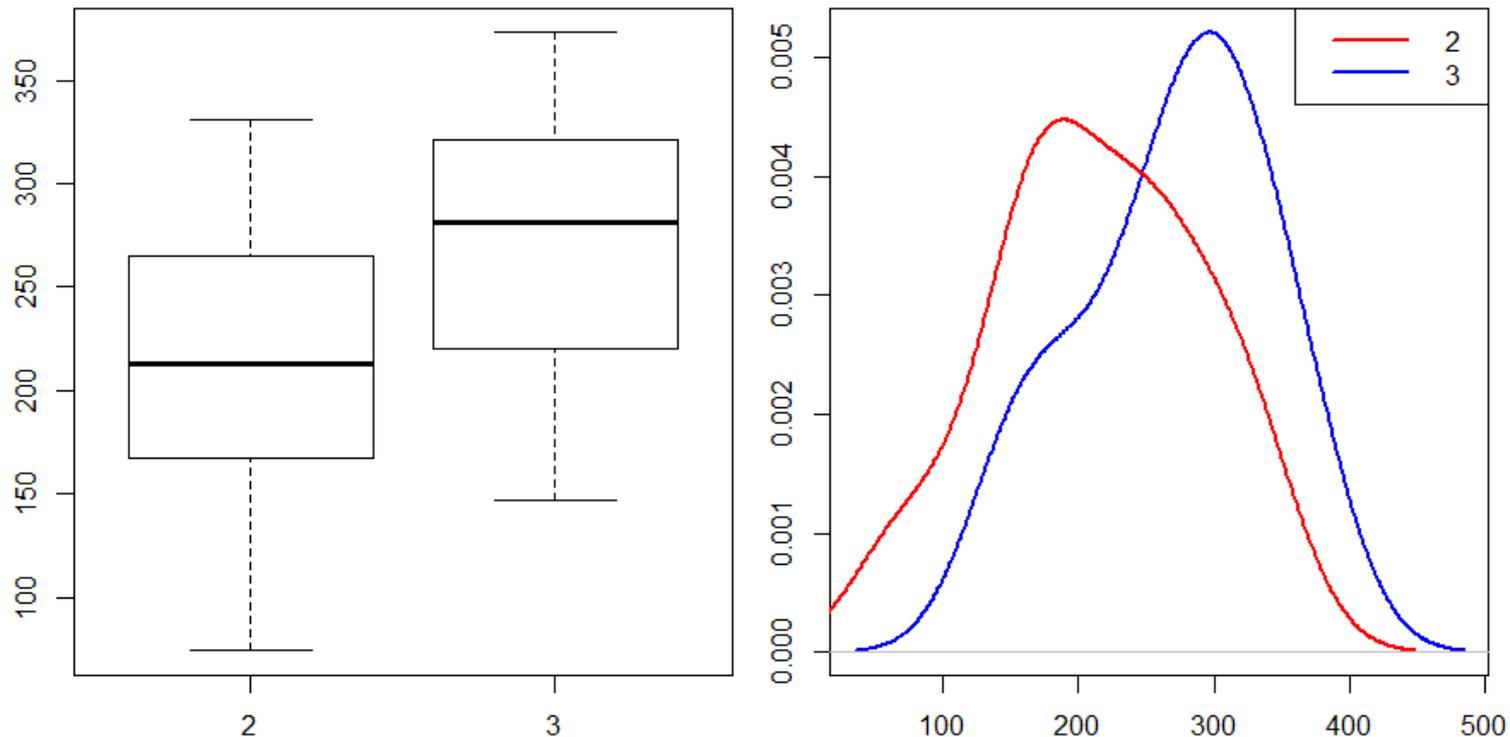
Как распределена каждая выборка?



Сравнение распределения выборки с заданным теоретическим распределением

1. Графический анализ выборок

Вес цыплят в зависимости от рациона питания



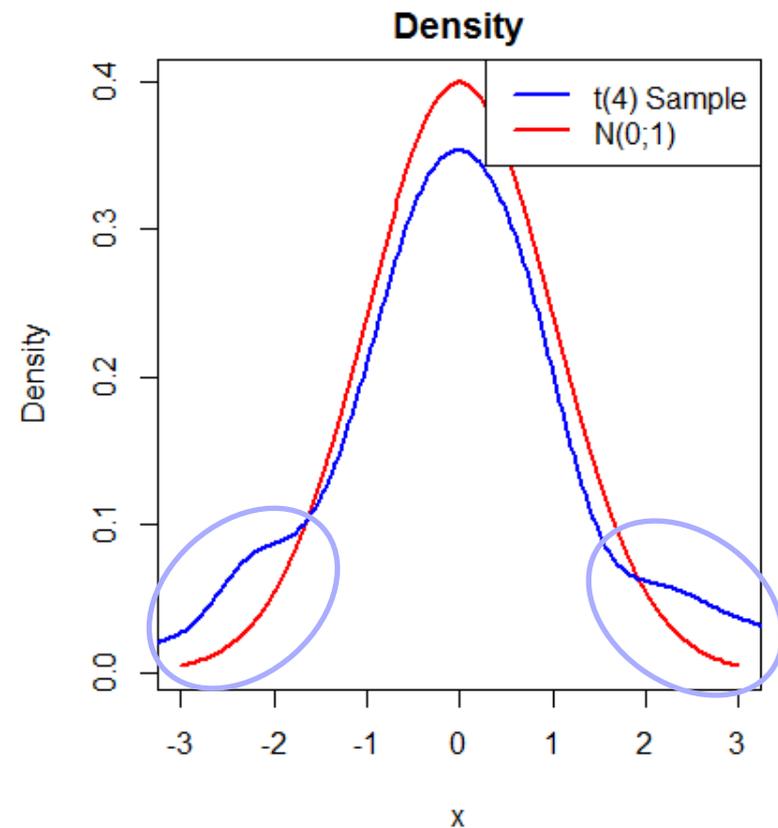
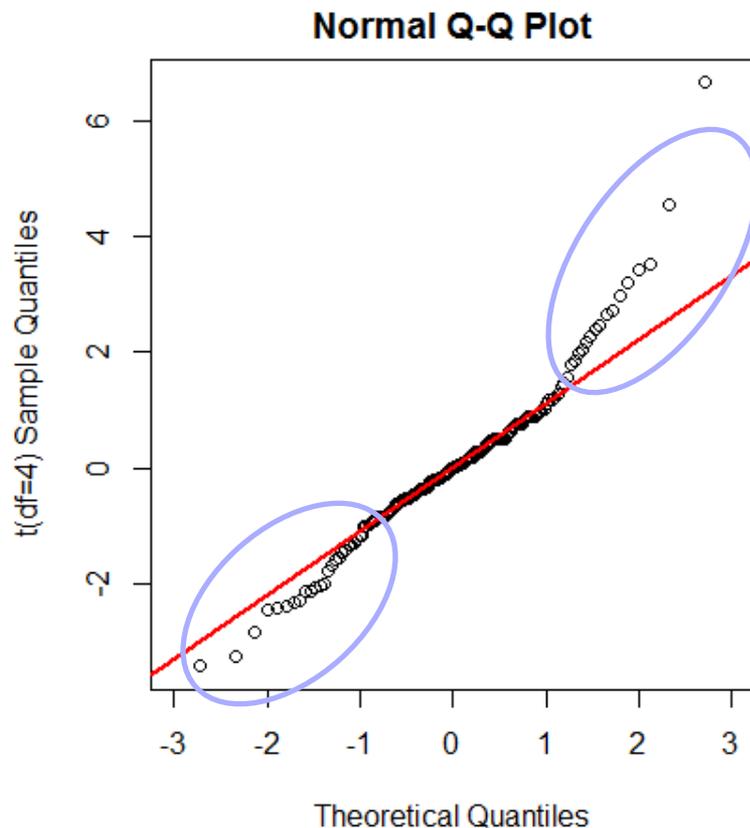
Являются ли выборки нормальными?
Из одного ли они распределения?

Сравнение формы распределений графически

qqplot – рисует квантили одной выборки напротив другой

qqnorm – рисует квантили выборки против квантилей нормального распределения

qqline – рисует линию, проходящую через 1 и 3 квантили теоретического (нормального) распределения

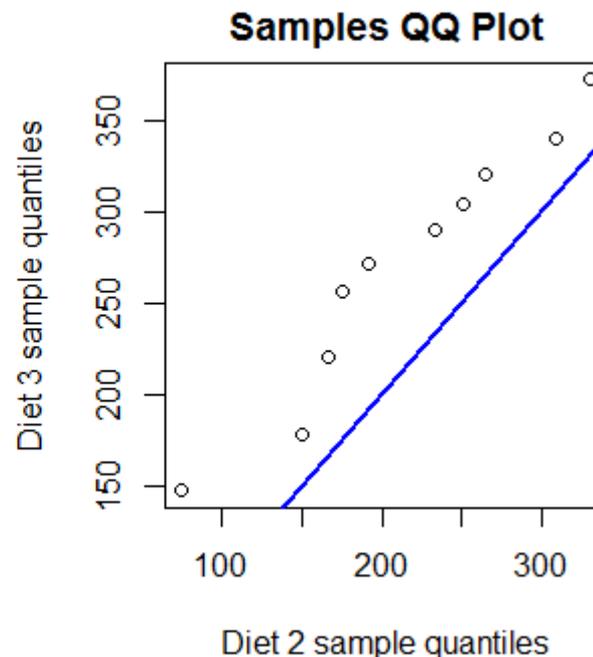
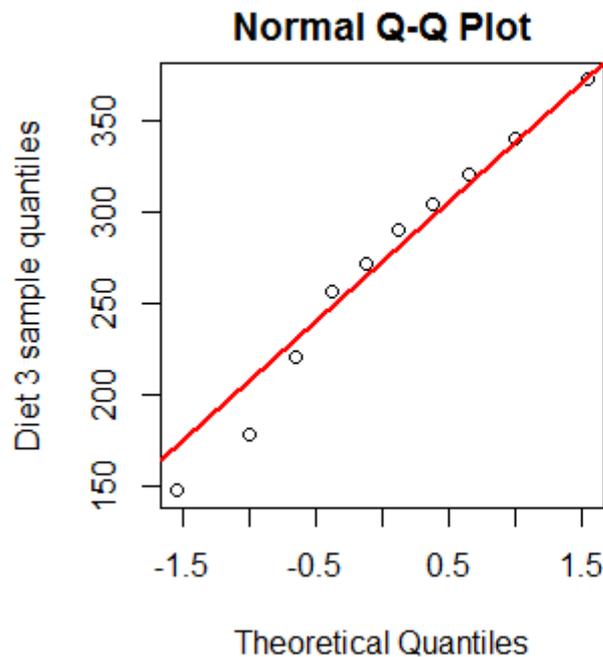


QQ Plot для веса цыплят

```

> par(mar=c(4,4,2,1),mfrow=c(1,2))
> w.diet.2 <- chick.w[chick.w$Diet==2,"weight"]
> w.diet.3 <- chick.w[chick.w$Diet==3,"weight"]
> qqnorm(w.diet.3, ylab="Diet 3 sample quantiles")
> qqline(w.diet.3,col="red",lwd=2)
> qqplot(w.diet.2,w.diet.3,xlab="Diet 2 sample quantiles",
+ ylab="Diet 3 sample quantiles", main="Samples QQ Plot")
> abline(0,1,col="blue",lwd=2)

```

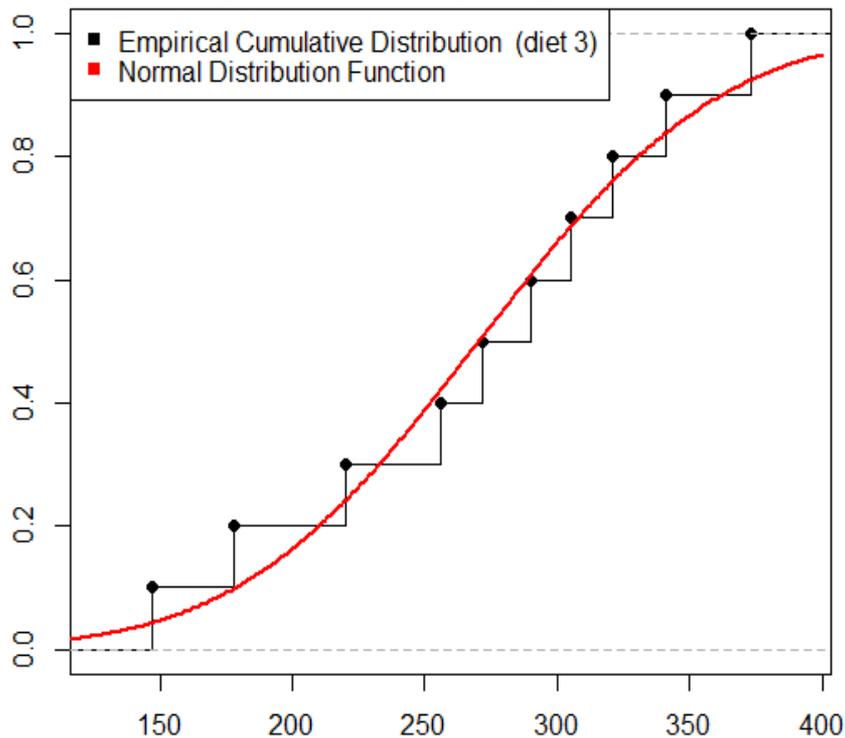


Статистические тесты для сравнения распределений

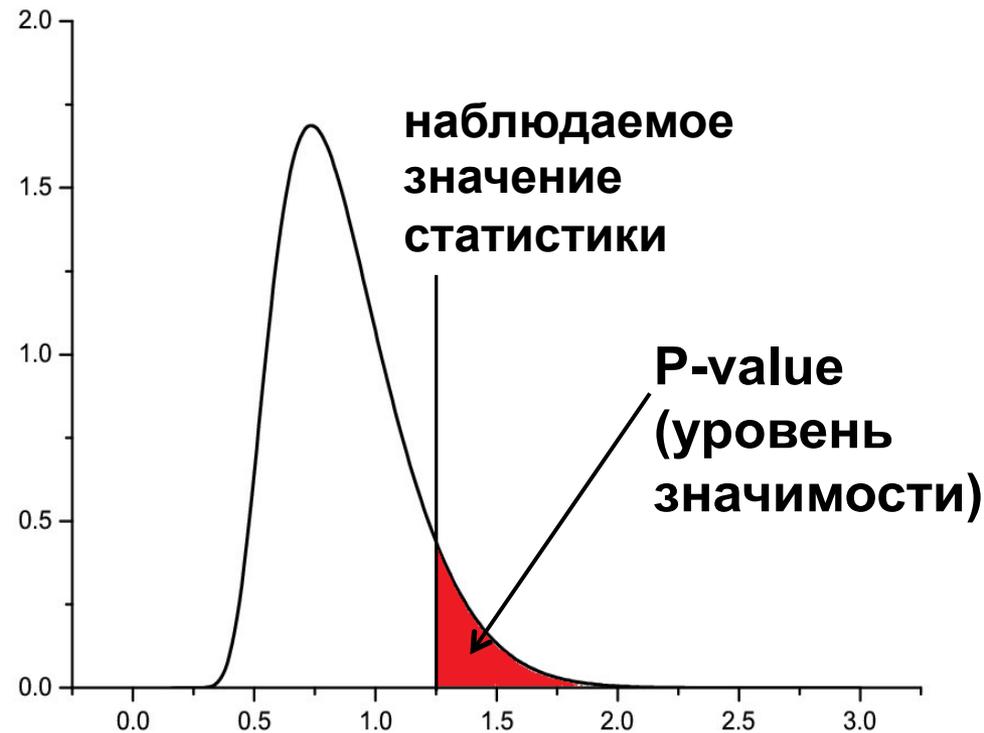
Тест Колмогорова-Смирнова:

- чувствителен к отличиям в форме распределений и их сдвигу относительно друг друга
- H_0 : распределения совпадают
- **плохо работает на маленьких выборках**
- применим только для непрерывных распределений

Сравнение эмпирического и теоретического распределений



Распределение статистики при нулевой гипотезе



Статистические тесты для сравнения распределений

Сравнение эмпирического распределения с теоретическим:

тест на нормальность

```
> ks.test(w.diet.3, "pnorm", mean(w.diet.3), sd(w.diet.3))
```

One-sample Kolmogorov-Smirnov test

```
data: w.diet.3
```

```
D = 0.1209, p-value = 0.9944
```

```
alternative hypothesis: two-sided
```

Сравнение распределений двух выборок:

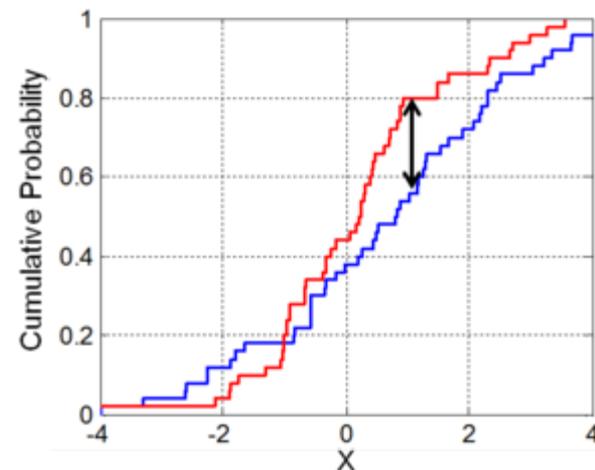
```
> ks.test(w.diet.2, w.diet.3)
```

Two-sample Kolmogorov-Smirnov test

```
data: w.diet.2 and w.diet.3
```

```
D = 0.4, p-value = 0.4175
```

```
alternative hypothesis: two-sided
```



Объект класса *htest*

Многие статистические тесты в R возвращают объект класса *htest*:

```
> diet3.ks <- ks.test(w.diet.2,w.diet.3)
```

```
> diet3.ks
```

```
Two-sample Kolmogorov-Smirnov test
```

```
data: w.diet.2 and w.diet.3
```

```
D = 0.4, p-value = 0.4175
```

```
alternative hypothesis: two-sided
```

```
> class(diet3.ks)
```

```
[1] "htest"
```

```
> names(diet3.ks)
```

```
[1] "statistic" "p.value" "alternative" "method"
```

```
[5] "data.name"
```

```
> diet3.ks$statistic
```

```
D
```

```
0.4
```

```
> diet3.ks$p.value
```

```
[1] 0.4175
```

Статистические тесты для сравнения распределений

Тест Shapiro-Wilk:

- проверяет гипотезу, что выборка пришла из **нормального распределения**
- H_0 : выборка является нормальной
- мощнее, чем тест Колмогорова-Смирнова (то есть с меньшей вероятностью ошибочно принимает H_0)
- размер выборки от 3 до 5000

```
> shapiro.test(w.diet.3) # возвращает объект htest
```

```
Shapiro-wilk normality test
```

```
data: w.diet.3
```

```
w = 0.9705, p-value = 0.895
```

```
> shapiro.test(w.diet.2)$p.value
```

```
[1] 0.948785
```

Вопрос №2

Сдвинуты ли выборки друг относительно друга?

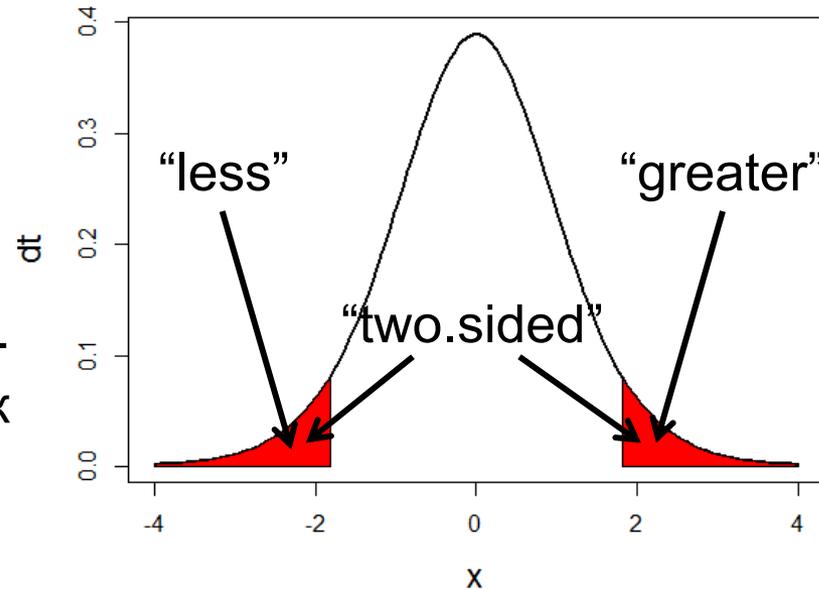
	Выборки (почти) нормальные	Выборки совсем не нормальные
Выборки независимые	<ul style="list-style-type: none"> • t-тест – проверяет равенство средних 	<ul style="list-style-type: none"> • U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона) • Kolmogorov-Smirnov test
Парные выборки	<ul style="list-style-type: none"> • Парный t-test – проверяет равенство разности случайных величин нулю 	<ul style="list-style-type: none"> • T-Критерий Вилкоксона

Student's (Gosset's) t-тест

- Введен Вильямом Госсетом в 1908 для оценки качества пива на пивоварне Guinness
- Используется для:
 - проверки равенства выборочного среднего заданному значению
 - проверки равенства средних значений двух серий измерений, сделанных для тех же объектов в разных условиях (например, состояние пациентов до и после лечения) – **paired t-test**
 - проверки равенства средних двух независимых выборок

- Предполагается, что случайные величины распределены **примерно нормально**
- При больших размерах выборок, распределение t-статистики приближается к нормальному

Распределение t-статистики



t-test для независимых выборок

Способ №1:

```
> chick.test <- t.test(w.diet.2, w.diet.3,  
alternative="less")
```

Способ №2:

```
> chick.test <- t.test(chick.w$weight ~ chick.w$Diet,  
alternative="less")
```

```
> chick.test$p.value
```

welch Two Sample t-test

data: chick.w\$weight by chick.w\$Diet

t = -1.6588, df = 17.865, p-value = 0.05731

alternative hypothesis: true difference in
means is less than 0

95 percent confidence interval:

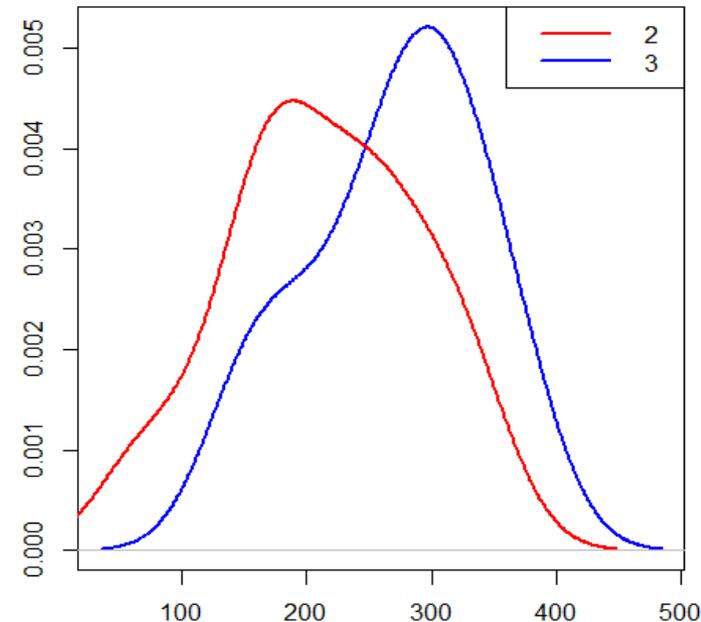
-Inf 2.548154

sample estimates:

mean in group 2 mean in group 3

214.7

270.3

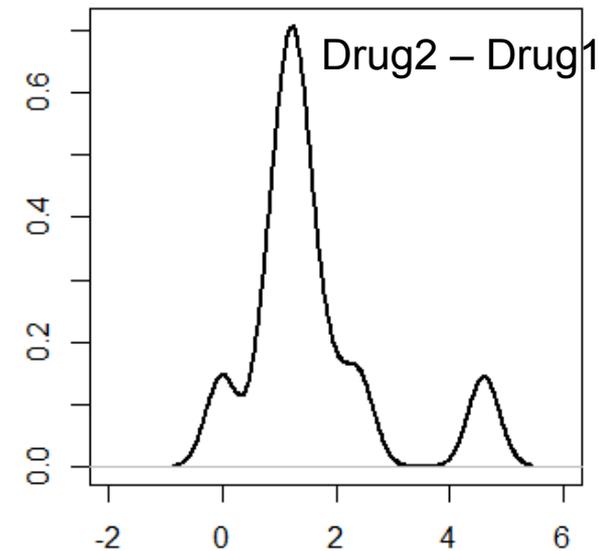
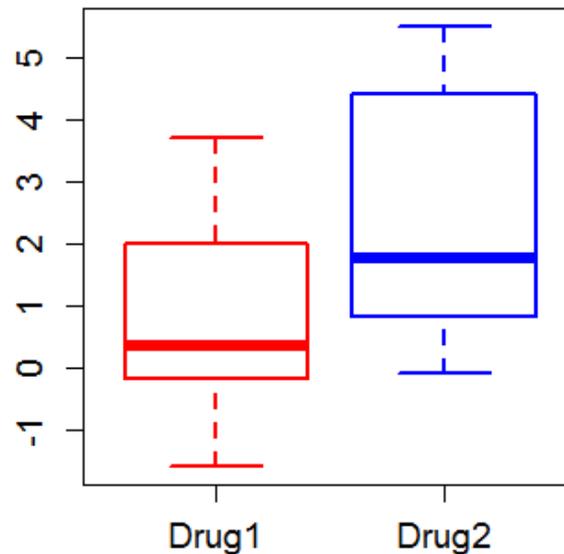


Данные: изменение длительности сна пациентов в зависимости от принимаемого лекарства

```
> sleep.paired <- read.table("sleep.paired.tab",header=T)
# ID – идентификатор пациента
# Drug1 и Drug2 – изменение длительности сна (в часах) при
# приеме лекарств 1 и 2
> sleep.paired
```

	ID	Drug1	Drug2
1	1	0.7	1.9
2	2	-1.6	0.8
3	3	-0.2	1.1
4	4	-1.2	0.1
5	5	-0.1	-0.1
6	6	3.4	4.4
7	7	3.7	5.5
8	8	0.8	1.6
9	9	0.0	4.6
10	10	2.0	3.4

Изменение длительности сна в зависимости от лекарства



Вопрос №2

Сдвинуты ли выборки друг относительно друга?

	Выборки (почти) нормальные	Выборки совсем не нормальные
Выборки независимые	<ul style="list-style-type: none"> • t-тест – проверяет равенство средних 	<ul style="list-style-type: none"> • U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона) • Kolmogorov-Smirnov test
Парные выборки	<ul style="list-style-type: none"> • Парный t-test – проверяет равенство разности случайных величин нулю 	<ul style="list-style-type: none"> • T-Критерий Вилкоксона

Парный t-тест

Помогло ли лекарство - стали ли пациенты дольше спать?

Способ №1:

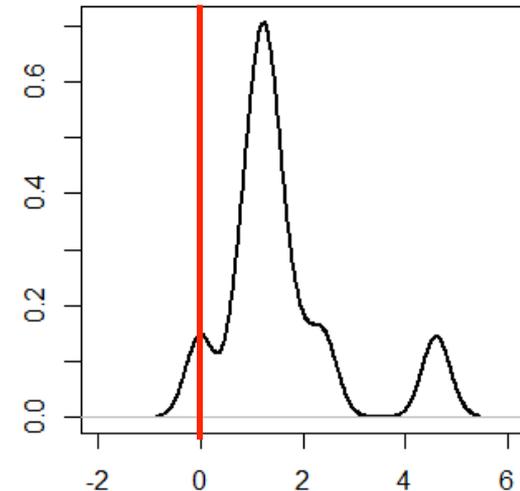
```
> sleep.test <- t.test(sleep.paired$Drug1,
+ sleep.paired$Drug2, paired=T, alternative="less")
```

Способ №2:

```
> diff <- sleep.paired$after - sleep.paired$before
> t.test(diff) # объект htest
```

One Sample t-test

```
data: diff
t = 4.0621, df = 9, p-value = 0.001416
alternative hypothesis: true mean is greater than
95 percent confidence interval:
 0.8669947      Inf
sample estimates:
mean of x
 1.58
```



Если забыть указать, что тест парный:

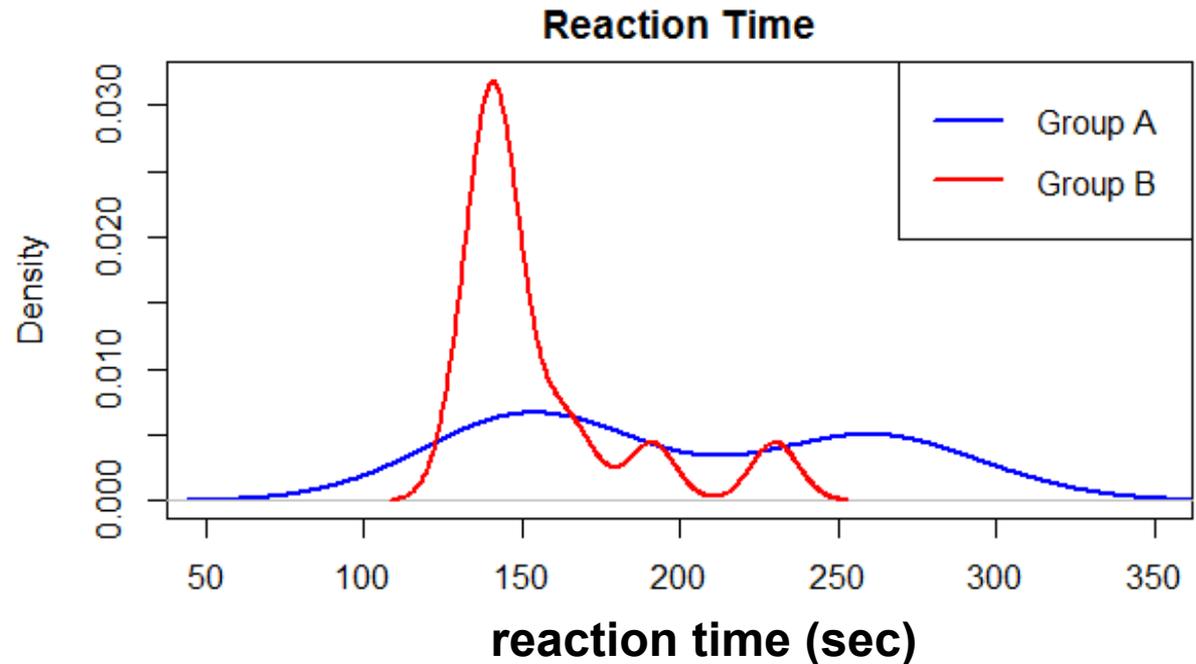
```
> t.test(sleep.paired$Drug1, sleep.paired$Drug2, alternative="less")
$p.value
[1] 0.03969707
```

Данные: влияние анестетика на время реакции пациентов на световой раздражитель

```
rt <- read.table("anaesthetic.reaction.time.tab",
  sep="\t", header=T)
```

Mean.RT – среднее время реакции; Group: A/B – with/without anesthetic

```
> head(rt)
  Mean.RT Group
1     131     B
2     135     A
3     138     B
4     138     B
5     139     A
6     141     B
```



Больше ли время реакции у пациентов под воздействием анестетика?

Вопрос №2

Сдвинуты ли выборки друг относительно друга?

	Выборки (почти) нормальные	Выборки совсем не нормальные
Выборки независимые	<ul style="list-style-type: none">• t-тест – проверяет равенство средних	<ul style="list-style-type: none">• U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)• Kolmogorov-Smirnov test
Парные выборки	<ul style="list-style-type: none">• Парный t-test – проверяет равенство разности случайных величин нулю	<ul style="list-style-type: none">• T-Критерий Вилкоксона

U-критерий Манна-Уитни

- Используется для тестирования гипотезы, что значения в одной из выборок в среднем (стохастически) больше, чем в другой
- H_0 : выборки не отличаются
- Позволяет выявлять различия в значении параметра между малыми выборками
- При больших размерах выборок, распределение U-статистики приближается к нормальному

U-критерий Манна-Уитни

Способ №1:

```
> wilcox.test(rt$Mean.RT~rt$Group, alternative="greater")
```

wilcoxon rank sum test with continuity correction

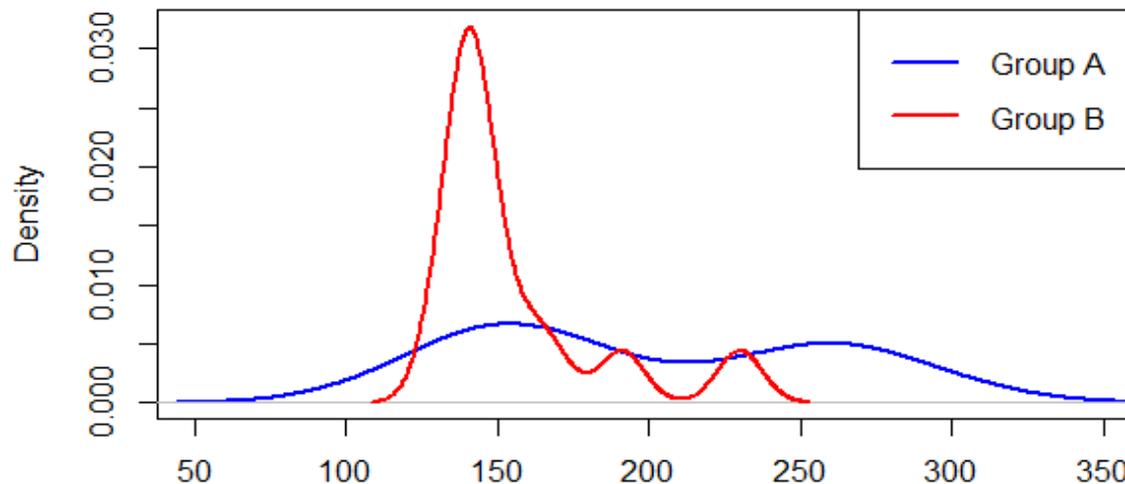
data: rt\$Mean.RT by rt\$Group

W = 126, p-value = 0.01633

alternative hypothesis: true location shift is greater than 0

Способ №2:

```
> wilcox.test(rt[rt$Group=="A", "Mean.RT."], rt[rt$Group=="B", "Mean.RT."], alternative="greater")
```



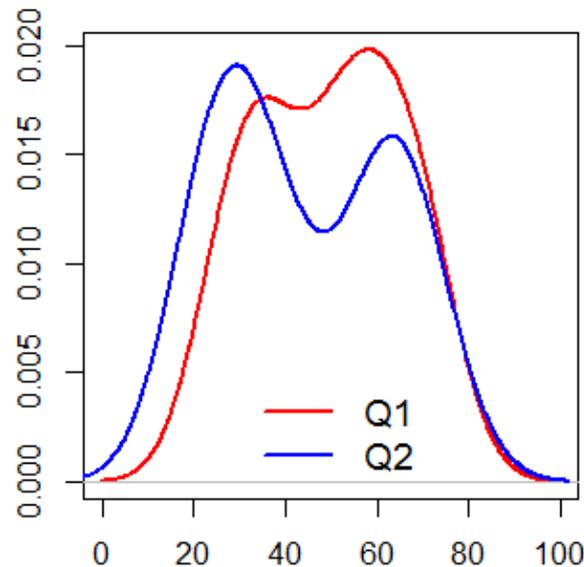
Данные: ответы студентов на вопросы теста

```
> test <- read.csv("StudentTest.csv")
# Student – студент, отвечающий на вопрос
# Q1, Q2 – баллы (от 0 до 100) за 1 и 2 вопросы
```

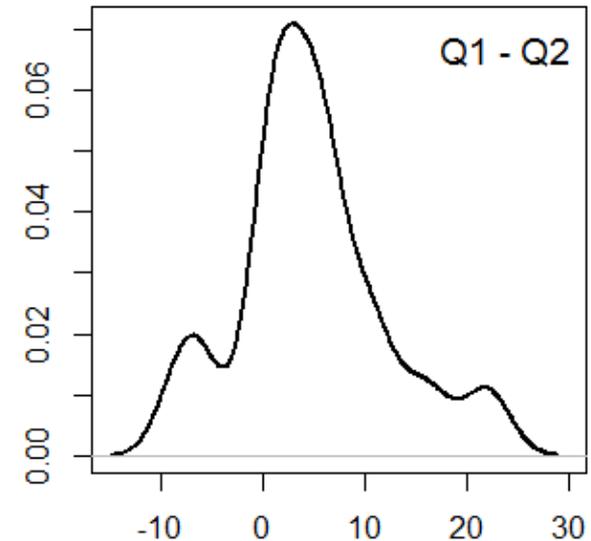
```
> head(test)
```

Student	Q1	Q2
1	78	67
2	24	24
3	64	62
4	55	58
5	74	28
6	52	36

Распределения баллов
за вопросы



Распределения разницы
баллов за вопросы



Предположение: студенты лучше отвечали на первый вопрос.

Вопрос №2

Сдвинуты ли выборки друг относительно друга?

	Выборки (почти) нормальные	Выборки совсем не нормальные
Выборки независимые	<ul style="list-style-type: none">• t-тест – проверяет равенство средних	<ul style="list-style-type: none">• U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)• Kolmogorov-Smirnov test
Парные выборки	<ul style="list-style-type: none">• Парный t-test – проверяет равенство разности случайных величин нулю	<ul style="list-style-type: none">• T-Критерий Вилкоксона

T-Критерий Вилкоксона

Способ №1:

```
> wilcox.test(test$Q1, test$Q2, paired=T,  
alternative="greater", exact=F)
```

wilcoxon signed rank test with continuity correction

data: test\$Q1 and test\$Q2

$V = 114.5$, $p\text{-value} = 0.008503$

alternative hypothesis: true location shift is greater than 0

Способ №2:

```
> wilcox.test(test$Q1-test$Q2, alternative="greater",  
exact=F)
```

Предположение, что студенты лучше отвечали на первый вопрос, подтвердилось на 1% уровне значимости.

Элементарная статистика: типичные вопросы

- Как распределены наблюдения в выборке?
 - является ли выборка нормальной? (`ks.test`, `shapiro.test`)

- Сравнение двух выборок:
 - из одного ли они распределения? (`ks.test`)
 - сдвину-ты ли они друг относительно друга? (`t.test`, `wilcox.test`, `ks.test`)