+

День 2

Outline

- Базовые понятия статистики
- Статистические тесты и условия их применения
- Тесты ассоциации, if, apply, merge.
- Проект по статистической обработке данных
- Знакомство и практика с BioconductoR

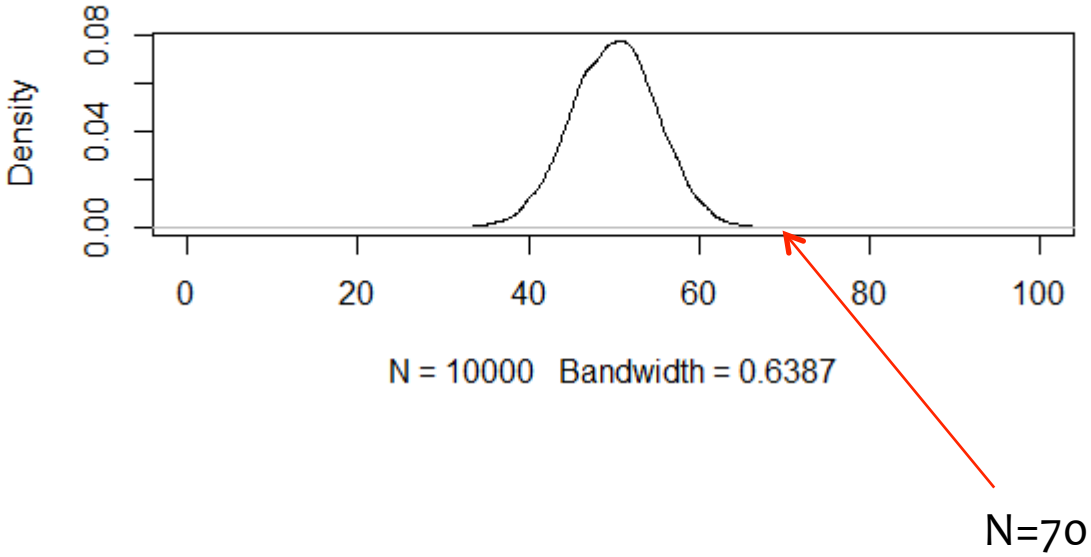
Эксперимент:
как отличить «честную» монетку от «нечестной»?

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?

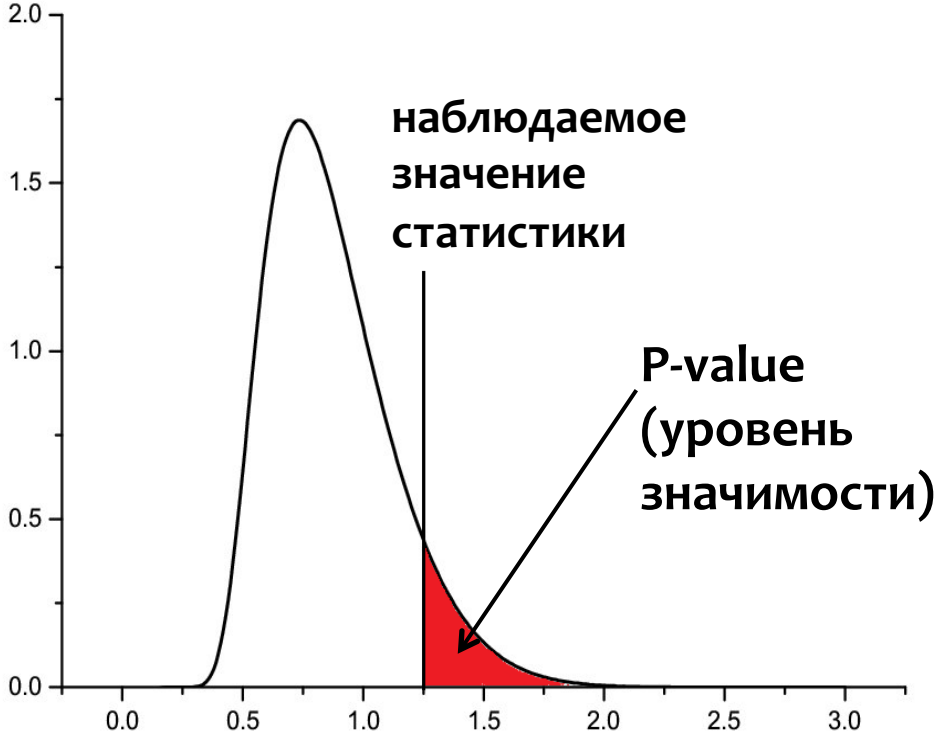
Распределение частот выпадения решки у честной монеты (биномиальное распределение):



H0 – нулевая гипотеза: мы кидали честную монету

H1 – 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!

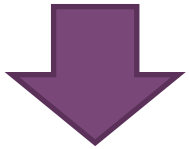
Данные: вес цыплят в зависимости от рациона питания

```
> 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

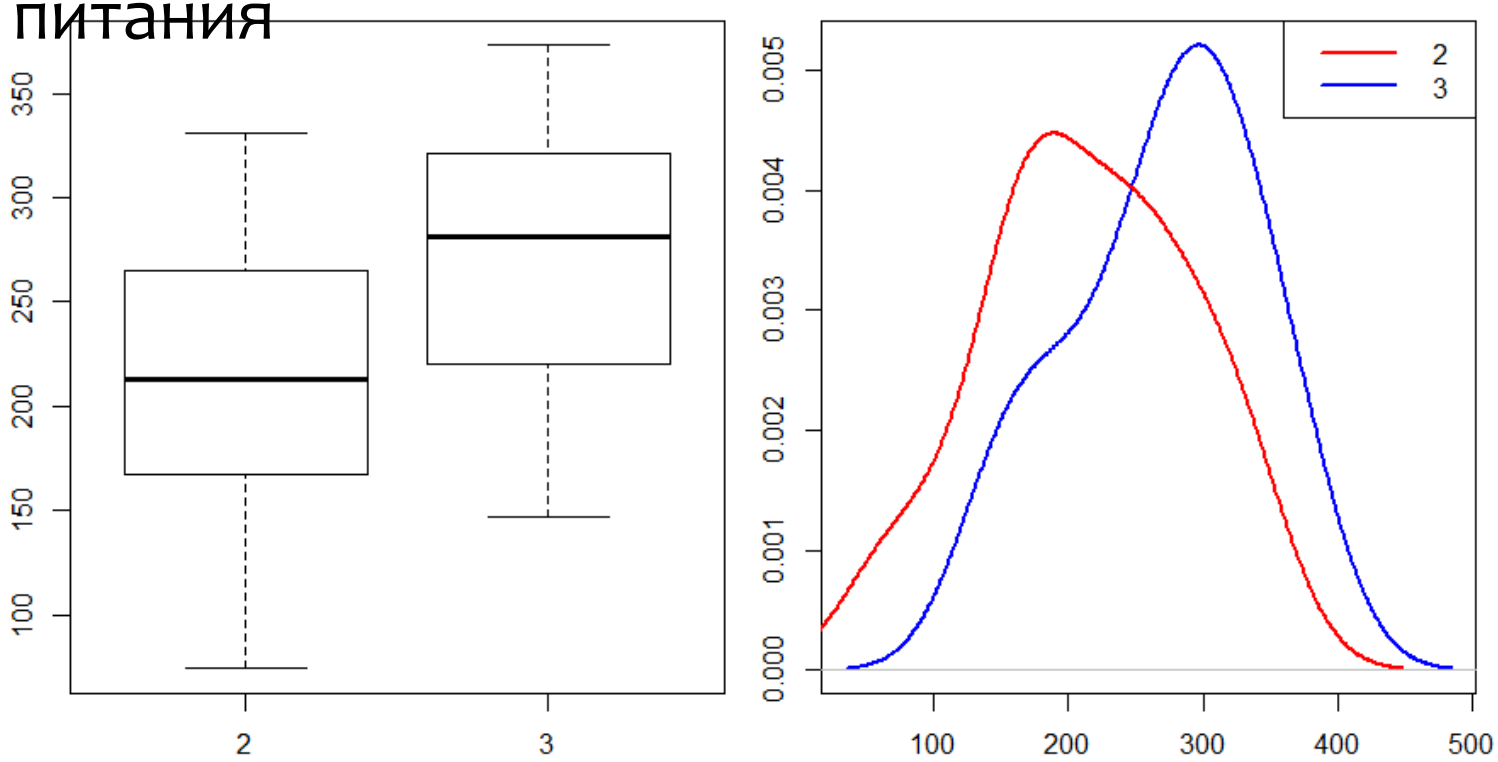
Как распределена каждая выборка?



Сравнение распределения выборки с заданным теоретическим распределением

1. Графический анализ выборок

Вес цыплят в зависимости от рациона питания



Являются ли выборки нормальными?

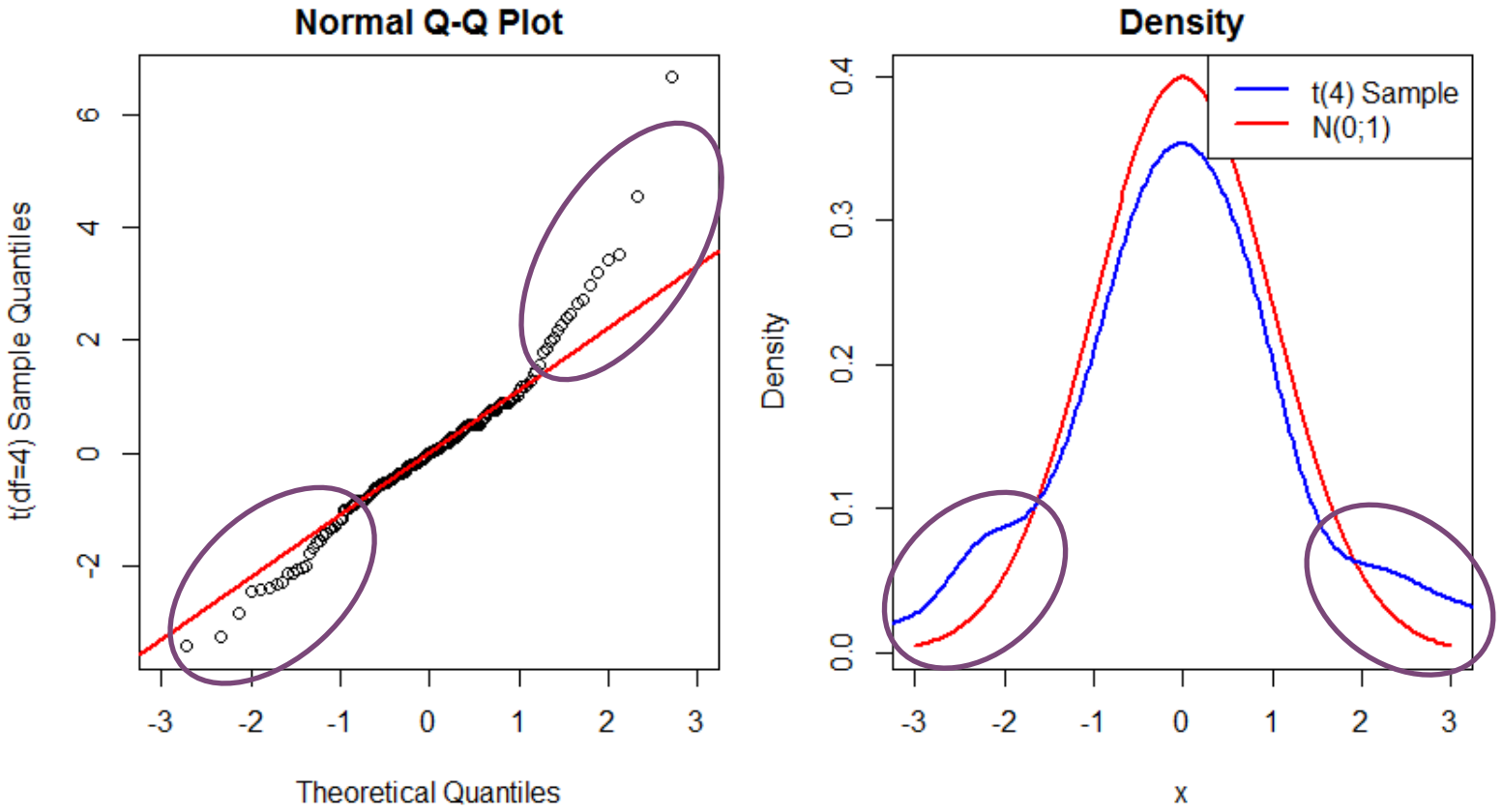
Из одного ли они распределения?

Сравнение формы распределений графически

qqplot – рисует квантили одной выборки напротив другой

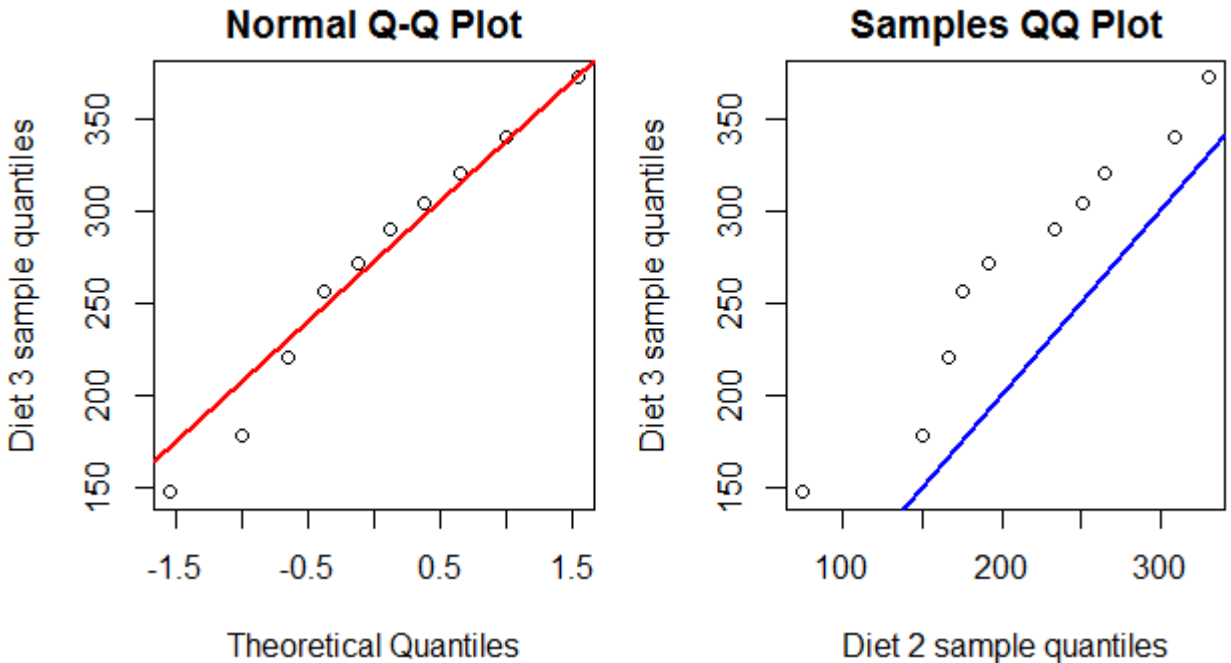
qqnorm – рисует квантили выборки против квантилей нормального распределения

qqline – рисует линию, проходящую через 1 и 3 квартили 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)
```

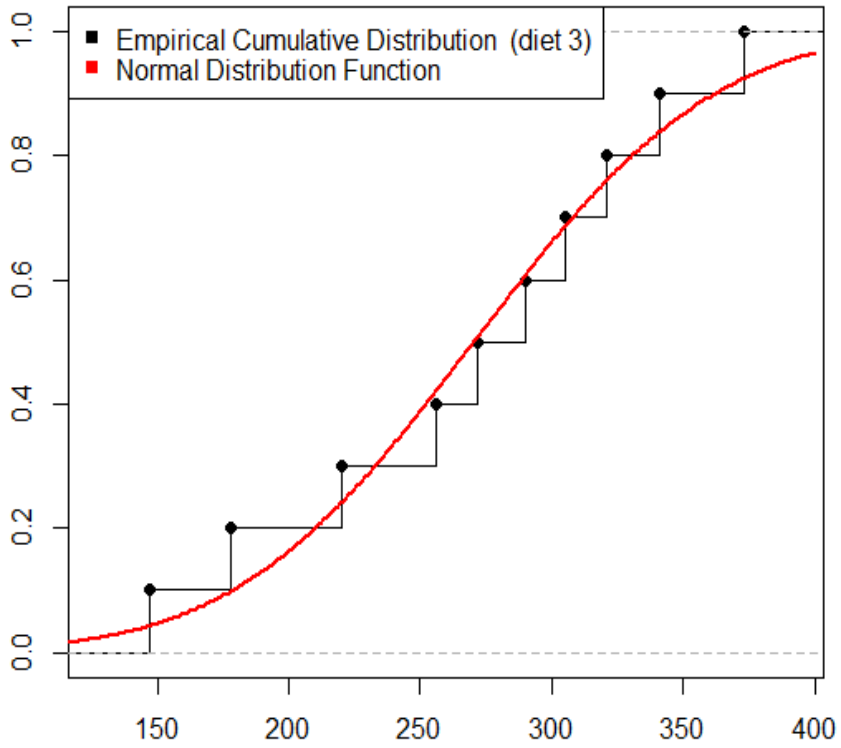


Статистические тесты для сравнения распределений

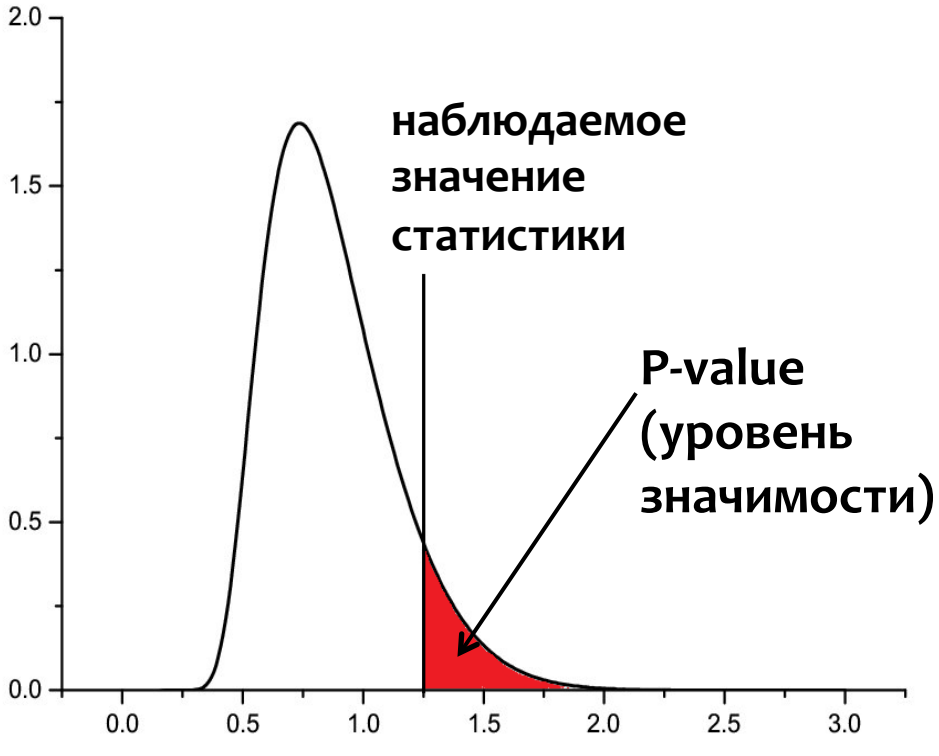
Тест Колмогорова-Смирнова:

- 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



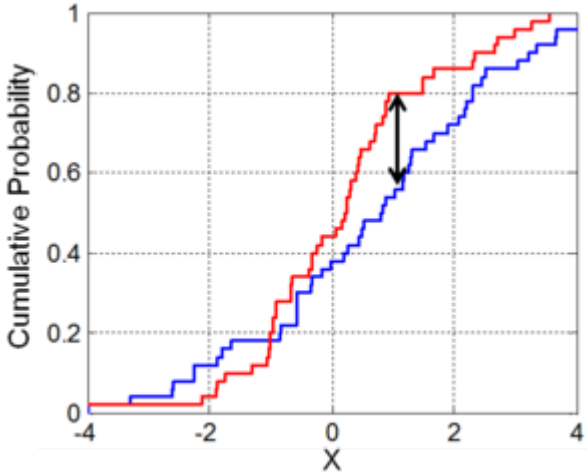
Статистические тесты для сравнения распределений

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	• **t-тест** – проверяет равенство средних	• **U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)** • **Kolmogorov-Smirnov test**
Парные выборки	• **Парный t-test** – проверяет равенство разности случайных величин нулю	• **Т-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

t-test для независимых выборок

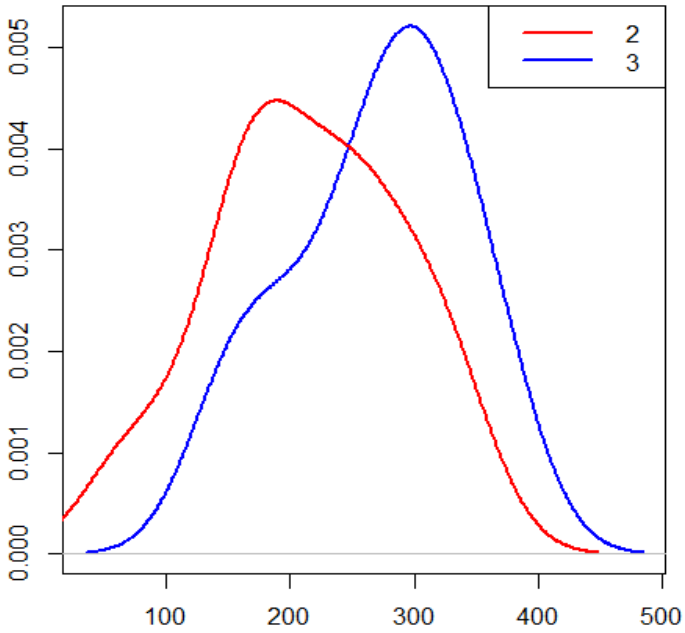
Способ №1:

```
> 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
```

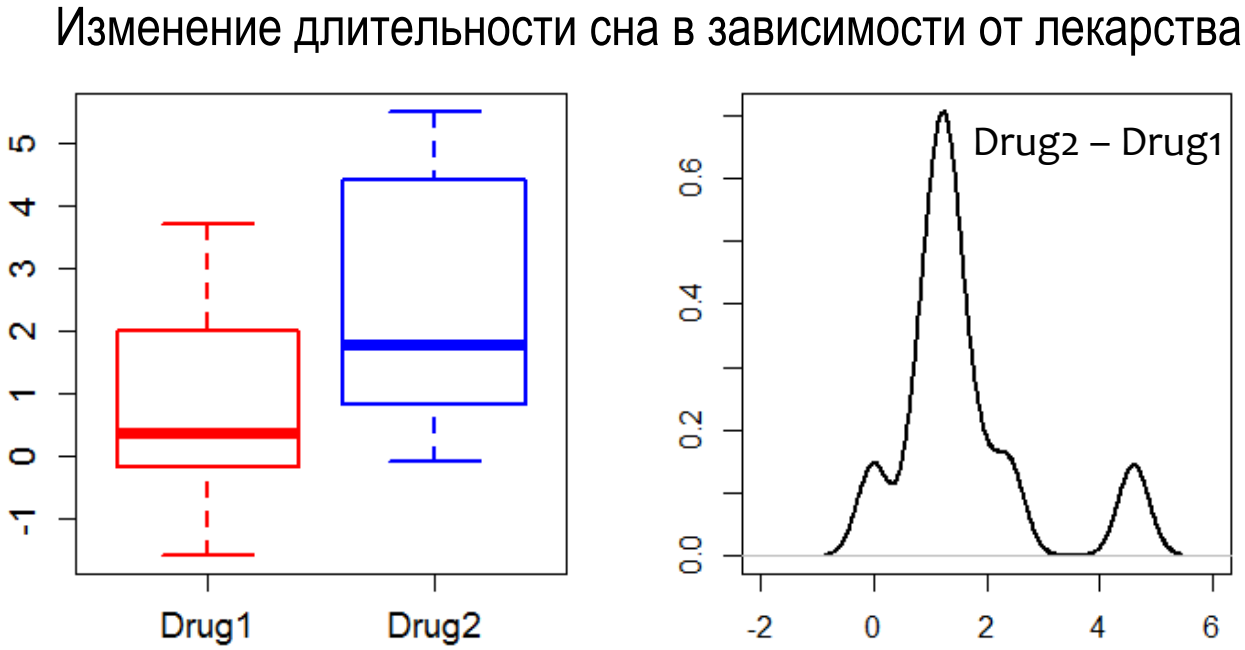


Данные: изменение длительности сна пациентов в зависимости от принимаемого лекарства

```
> 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



Вопрос №2

Сдвинуты ли выборки друг относительно друга?

	Выборки (почти) нормальные	**Выборки совсем не нормальные**
Выборки независимые	• **t-тест** – проверяет равенство средних	• **U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)** • **Kolmogorov-Smirnov test**
Парные выборки	• **Парный t-test** – проверяет равенство разности случайных величин нулю	• **T-Критерий Вилкоксона**

Парный t-тест

Помогло ли лекарство - стали ли пациенты дольше спать?

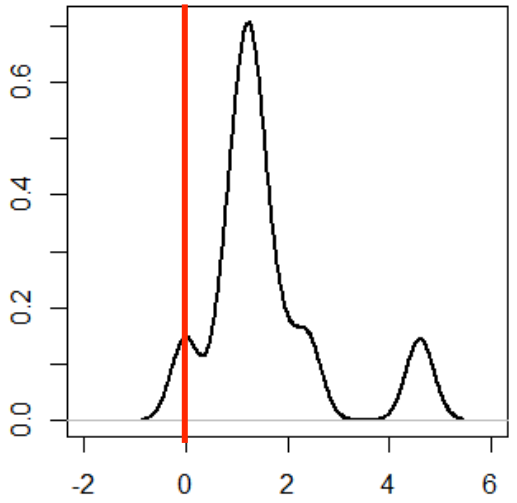
Способ №1:

```
> 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 0

sample estimates:

mean of x

     1.58
```



Если забыть указать, что тест парный:

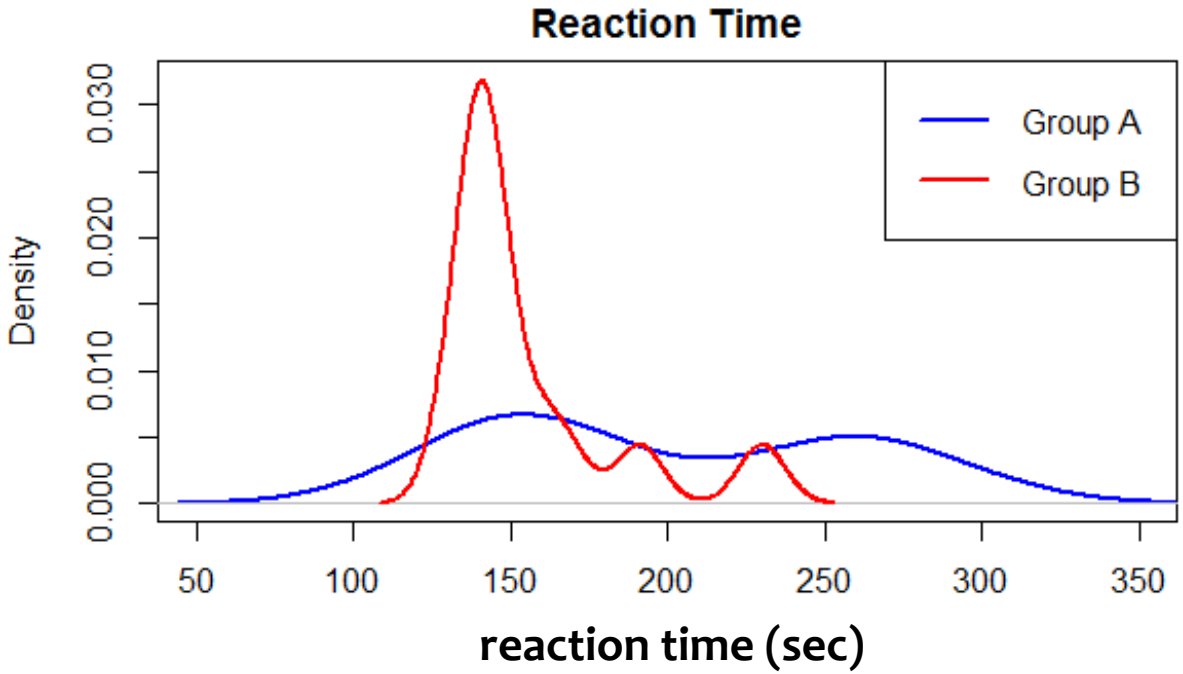
```
> t.test(sleep.paired$Drug1, sleep.paired$Drug2, alternative="less")$p.value

[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	• **t-тест** – проверяет равенство средних	• **U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)** • **Kolmogorov-Smirnov test**
Парные выборки	• **Парный t-test** – проверяет равенство разности случайных величин нулю	• **Т-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-сстатистики приближается к нормальному

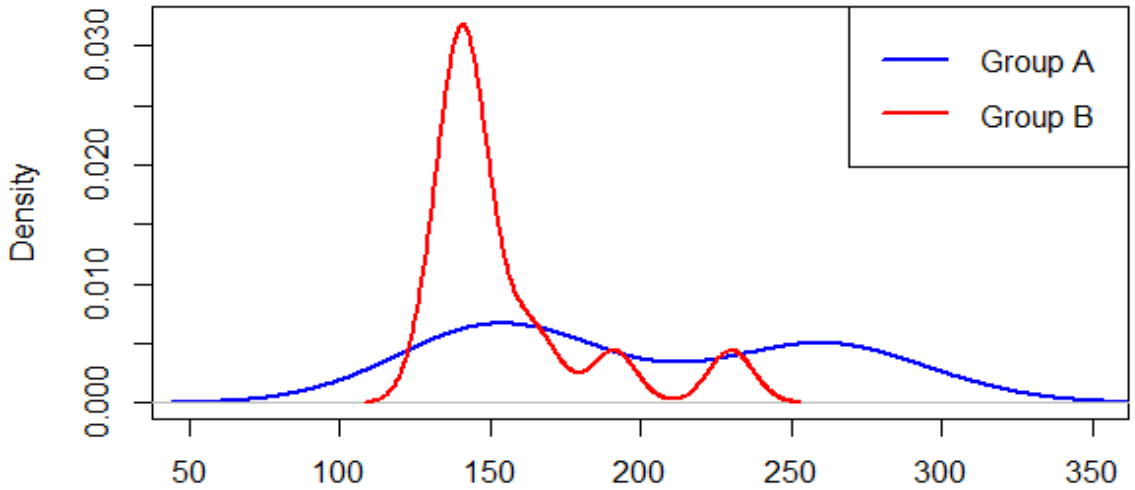
U-критерий Манна-Уитни

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")
```

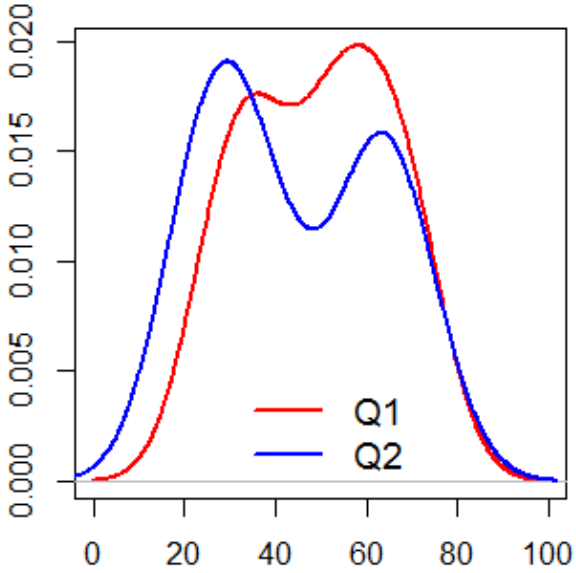


Данные: ответы студентов на вопросы теста

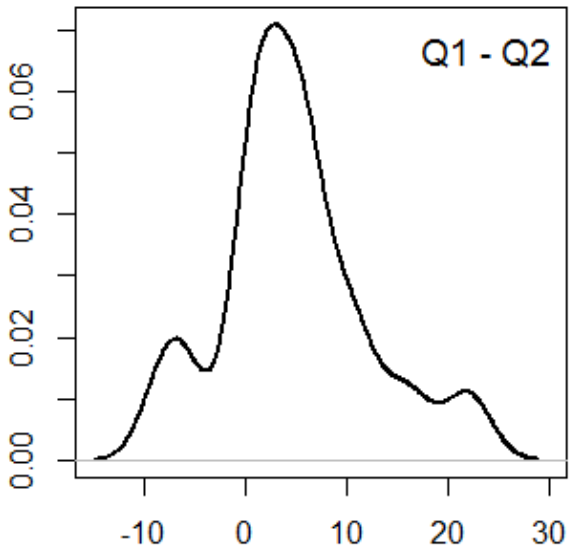
```
> test <- read.csv("StudentTest.csv")
# Student – студент, отвечающий на вопрос
# Q1, Q2 – баллы (от 0 до 100) за 1 и 2 вопросы
```

```
> head(test)
  Student Q1 Q2
1       1 78 67
2       2 24 24
3       3 64 62
4       4 55 58
5       5 74 28
6       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	• **t-тест** – проверяет равенство средних	• **U-критерий Манна-Уитни (Критерий суммы рангов Уилкоксона)** • **Kolmogorov-Smirnov test**
Парные выборки	• **Парный t-test** – проверяет равенство разности случайных величин нулю	• **T-Критерий Вилкоксона**

Т-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-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 студены 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)

QUIZ TIME

?

Скачайте файл students.txt, содержащий список студентов, участвующих в неком межфакультетском курсе.

Какие два факультета (faculty) наиболее представлены на курсе?

bio
econom
psych
soc
vmk

Создайте случайню выборку размера N из нормального распределения с mean=0, sd=1.4 (перед этим установите seed = 1).

При каком минимальном N можно утверждать, что выборка отличается от нормального стандартного распределения в соответствии с тестом Колмогорова-Смирнова при уровне значимости 1%?

25
100
400
1000

Друзья предлагают подарить вам на день рождения сертификат на 1000 р либо в кафе-мороженное, либо в кондитерскую с вкусными булочками. Они предлагают вам выбрать, какой из них вы больше хотите. Вы одинаково сильно любите мороженное и пирожные, поэтому вас больше интересует, в каком из этих мест вы накупите больше вкусняшек на 1000 р. Вы скачали прайс-листы этих заведений и решили сравнить цены.

Какой тест из предложенных наилучшим образом подходит для решения этой задачи?

парный t-тест
U-критерий Манна-Уитни
T-Критерий Вилкоксона
невозможно выбрать из предложенных выше без проверки выборок на нормальность

Брокер, управляющий портфелем из 16 ценных бумаг, следит за изменением стоимости этих ценных бумаг на бирже. Каждый день он сравнивает текущие цены с ценами предыдущего дня и пытается понять, значимо ли изменились цены. Скачайте файл SecurityPriceTest.csv, содержащий данные по стоимости бумаг для двух соседних дней (Day1 идет раньше, чем Day2).

Какое из следующих утверждений верно (на уровне значимости 1%)?

стоимость значимо упала
стоимость значимо выросла
невозможно подтвердить ни увеличение, ни уменьшение стоимости бумаг при требуемом уровне значимости