# Тестирование статистических гипотез

## Спецглавы по матобработке данных на R, осень 2014

Вадим Хайтов
Каф. Зоологии беспозвоночных, СПбГУ

## Вы сможете

- Уверенно объяснить, что такое статистический критерий и как он работает
- Применить команды R для проверки наиболее распространенных типов гипотез
- Понять что такое пермутационный метод тестирования гипотез
- Написать R код, позволяющий реализовать пермутационный метод

## ЧАСТЬ 1. Основы основ

- Нормальное распределение величин.
- Параметры распределения.
- Выборочные оценки параметров распределения.

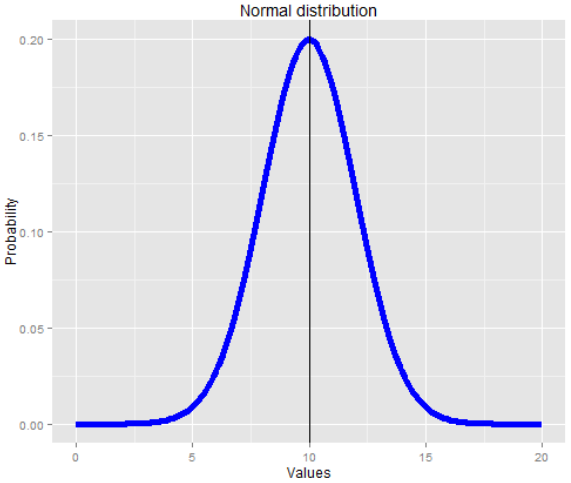
## Нормальное распределение

Распределение - это функция, описывающая связь между значениями величины и вероятностью ее встречи в генеральной совокупности

Нормальное распределение описывается такой формулой

$$p = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$


График этой функции



## Нормальное распределение

Нормальное распределение описывается такой формулой

$$p = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Как и любая функция, функция, описывающая нормальное распределение, имеет параметры

Два параметра нормального распределения

- $\mu$ - Наиболее часто встречающееся в генеральной совокупности знчение.
- $\sigma$ - Характеризует разброс, дисперсию, значений в генеральной совокупности.

## Научимся делать выборки из генеральной совокупности с нормальным распределением величины

Пусть у нас есть величина, для которой в генеральной совокупности $\mu = 50$, а $\sigma = 5$

Возьмем из этой генеральной совокупности выборку в 150 объектов

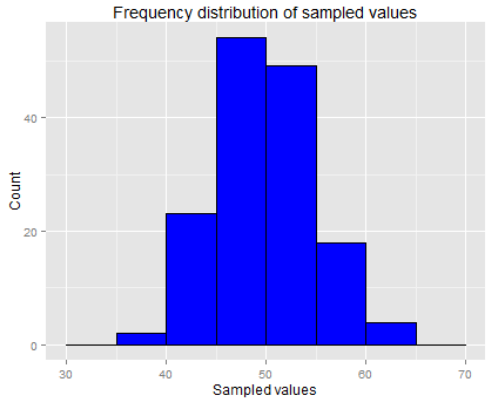
```
set.seed(123)
sample <- rnorm(150, 50, 5)
sample <- data.frame(xi=sample)
head (sample)
```

```
##      xi
## 1 47.20
## 2 48.85
## 3 57.79
## 4 50.35
## 5 50.65
## 6 58.58
```

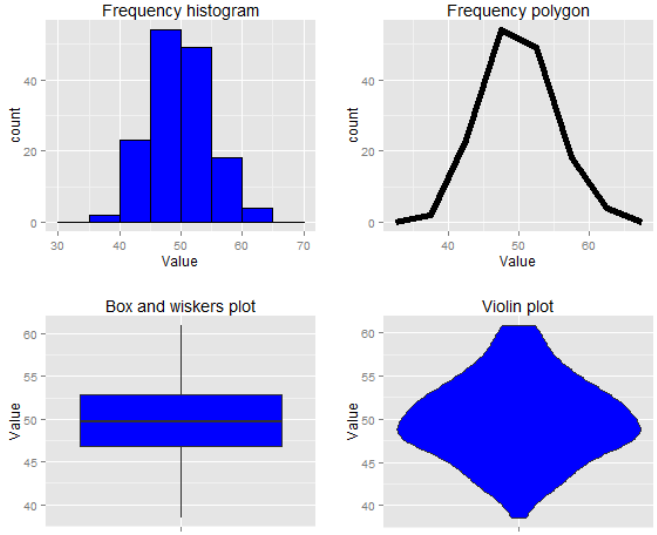
## Научимся строить выборочные частотные распределения

```
library(ggplot2)
pl_sample_distribution <- ggplot(sample, aes(x=xi)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  xlab("Sampled values") +
  ylab("Count") +
  ggtitle("Frequency distribution of sampled values")

pl_sample_distribution
```



## Другие формы отражения частотных распределений



## Выборочные оценки

- Оценкой параметра $\mu$ является среднеее значение в выборке

$$\bar{x} = \frac{\sum x_i}{n}$$

- Оценкой параметра $\sigma$ является среднеквадратичное отклонение

$$sd = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

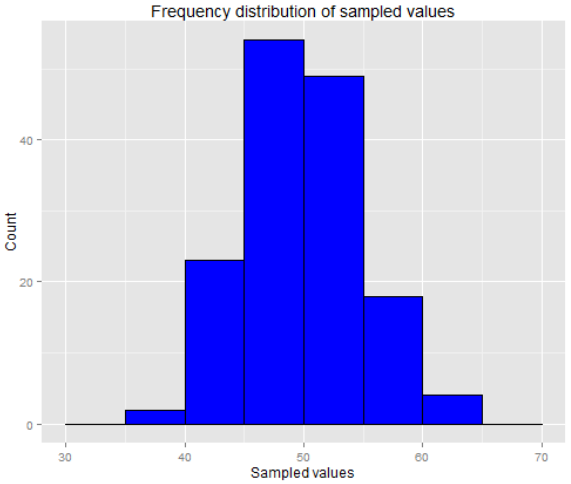
```
mean(sample)
```

```
## [1] 49.88
```

```
sd(sample)
```

```
## [1] 4.749
```

```
pl_sample_distribution
```



## Задача

А если мы возьмем не одну выборку, а много выборок одинакового размера (n) из той же генеральной совокупности, с теми же параметрами, то как будет распеределена такая величина?

## Задание

Пусть у вас имеется озеро, в котором плавают рыбы, и вы знаете, что в этой бесконечно большой популяции (генеральной совокупности) следующие параметры $\mu = 50$ и $\sigma = 5$.

- Напишите R код, который моделирует взятие одной выборки (по 150 особей в каждой): вам понадобится функция `rnorm()`
- Смоделируйте процесс, взятия 1000 аналогичных выборок из той же генеральной совокупности: вам понадобится функция `for(i in 1:1000){}`
- Создайте датафрейм, содержащий средние значения для этих выборок: вам понадобятся функции `for()` и `mean()`.
- Постройте частотую гистограмму, отражающую распределение средних значений: вам понадобится функция `ggplot()` вместе с геомом `geom_histogram()`
- Измените форму представления распределения на частотный полигон: вам понадобится геом `geom_freqpoly()`

# Решение

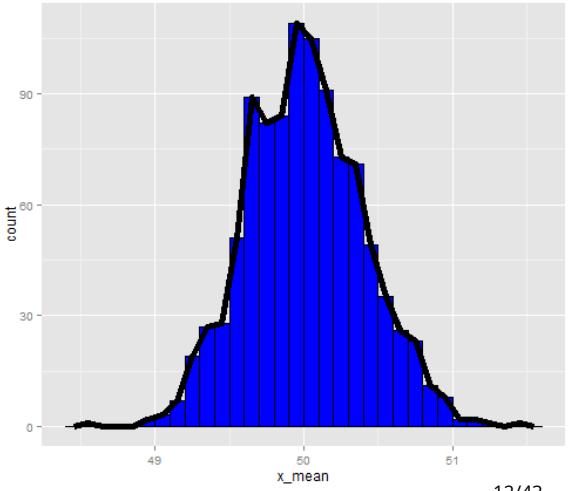
```
means <- data.frame(x_mean = numeric(1000))

for (i in 1:1000) means[i,1] <- mean(rnorm(15(
```

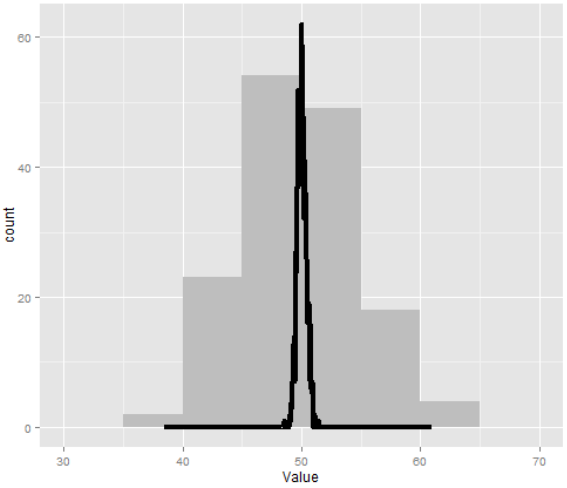
```
head(means)
```

```
##   x_mean
## 1  50.47
## 2  49.77
## 3  50.32
## 4  49.86
## 5  50.31
## 6  50.22
```

```
ggplot(means, aes(x = x_mean)) +
  geom_histogram(binwidth=0.1, fill="blue", col(
  geom_freqpoly(size=2, bin=0.1)
```



# Все то же самое, но совмещенное с распределением исходных величин в одной из выборок



Среднеквадратичное отклонение средних значений в генральной совокупности вычисляется по такой формуле

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Распределение средних значений, взятых из одной генеральной совокупности, имеет два праметра:

- $\mu$
- $SD_{\bar{x}}$

- Выборочная оценка величины $SD_{\bar{x}}$ называется ошибкой среднего

- $$SE_{\bar{x}} = \frac{sd}{\sqrt{n}}$$

## И последнее...

Научимся стандартизировать распределения

Ввведем величину

$$z_i = \frac{x_i - \bar{x}}{sd}$$

```
z <- (sample - mean(sample))/sd(sample)
```

Среднее значение этой величины будет всегда равно 0, а sd = 1

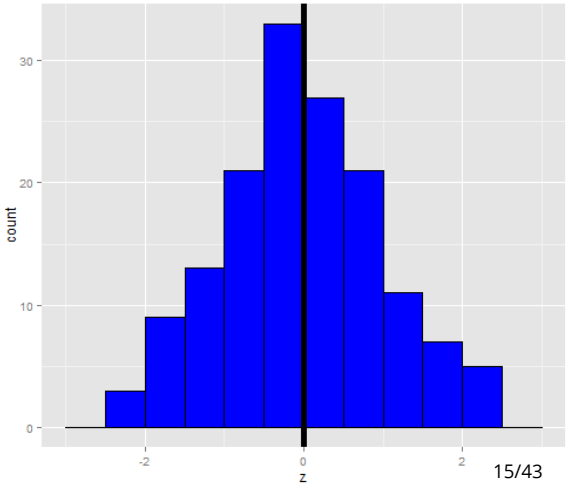
```
mean(z)
```

```
## [1] 2.112e-16
```

```
sd(z)
```

```
## [1] 1
```

```
ggplot(data.frame(z=z), aes(x=z)) +
  geom_histogram(binwidth=0.5, fill="blue", col
  geom_vline(xintercept=0, size=2)
```



**И вот, наконец...**

William Sealy Gosset

t-распеделение Стьюдента (Student, 1908)

$$t = \frac{d}{SE_d}$$

где $d = \bar{x_1} - \bar{x_2}$ - это разность между двумя средними значениями

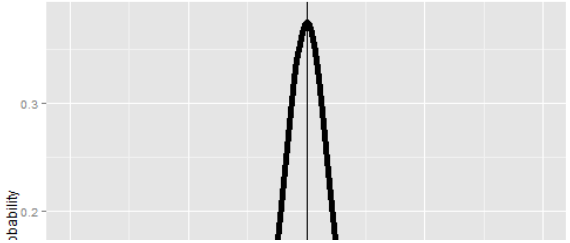
$SE_d$ - Общее среднеквадратичное отклонение разности двух средних

$$SE_d = \sqrt{\frac{sd_1^2(n_1 - 1) + sd_2^2(n_2 - 1)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{}\right.}$$

Если $n_1 = n_2$, то формула существенно упрощается

$$SE_d = \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

**Таким образом t-распределение это всего лишь стандартизованное распределение разностей ДВУХ средних значений, взятых из ОДНОЙ генеральной совокупности!**



# Задача

С какой вероятностью мы можем встретить в ОДНОЙ генеральной совокупности два средних значения (вычисленные по выборкам из 10 объектов, каждая), стандартизированная разность между которыми оказывается больше 2 (или меньше -2)?

$$\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} > 2$$

или

$$\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}} < -2$$

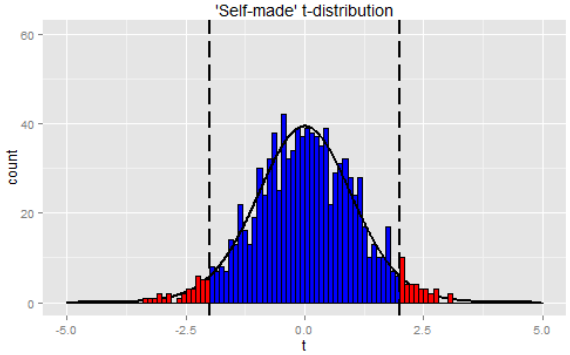
- Эту вероятность можно вычислить строго.
- Но, для того, чтобы "пощупать руками" суть критерия, давайте смоделируем процесс взятия парных выборок из ОДНОЙ генеральной совокупности. То есть представим себе, что мы много раз (например, 1000) взяли повторные парные выборки из одной и той же генеральной совокупности.

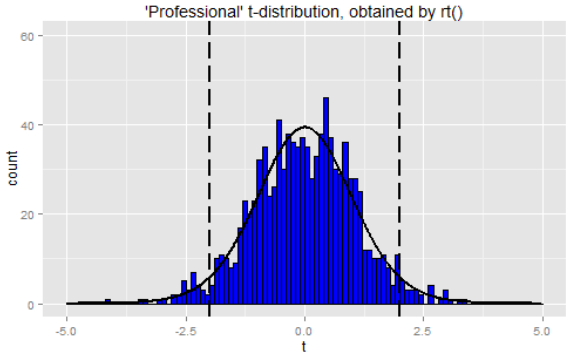
```
t_sample1 <- rep(0,1000)

for (i in 1:1000) {
  # Берем две выборки
  samp1 <- rnorm(10, 50,5)
  samp2 <- rnorm(10, 50, 5)

  # Стандартизируем значения
  t_sample1[i] <- (mean(samp1) - mean(samp2))/sqrt(sd(samp1)^2/length(samp1) + sd(samp2)^2/le
}


#Все то же самое, но с функцией rt()
t_sample2 <- rt(1000, (10 + 10 - 2))
```


'Self-made' t-distribution
count
60
40
20
0
-5.0
-2.5
0.0
2.5
5.0
t


'Professional' t-distribution, obtained by rt()
count
60
40
20
0
-5.0
-2.5
0.0
2.5
5.0
t

# Для оценки интересующей нас вероятности, нам надо понять сколько раз из 1000 мы встретим величину больше 2 (или меньше -2)

Доля значений t > 2 или t < -2

```
p1 <- sum(t_sample1 > 2 | t_sample1 < -2)/length(t_sample1)
p1
```

```
## [1] 0.068
```

```
p2 <- sum(t_sample2 > 2 | t_sample2 < -2)/length(t_sample2)
p2
```

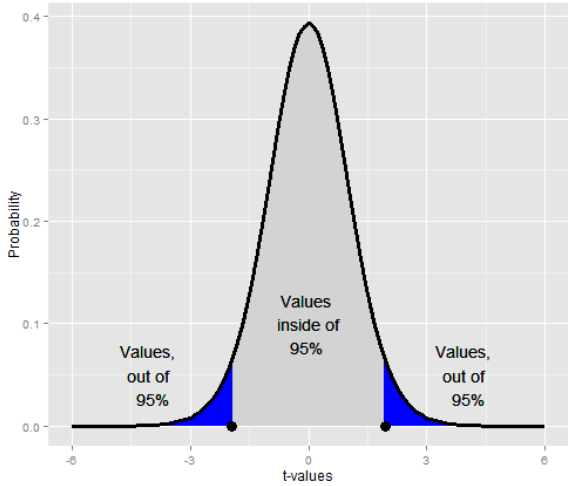
```
## [1] 0.06
```

# Для строгой оценки этой вероятности оценивают долю площади под кривой, описывающей распределение (кривая плотности вероятности).

Для заданной границы t>2 (t<-2) это будет отношение закрашенной площади под кривой к общей площади

# Но! Можно поставить вопрос иначе.

А где находится значение t, которое отделяет область, составляющую, например, 95% площади под кривой?



- Можем договориться, что нас интересуют только те значения t, которые входят в область 95%. Остальные значения мы будем считать не принадлежащими к данной совокупности.

- Соответственно, мы будем считать, что те значения t, которые не попадают в эту, условно ограниченную совокупность, не являются стандартизированными разностями двух средних для выборок, взятых из ОДНОЙ генеральной совокупности

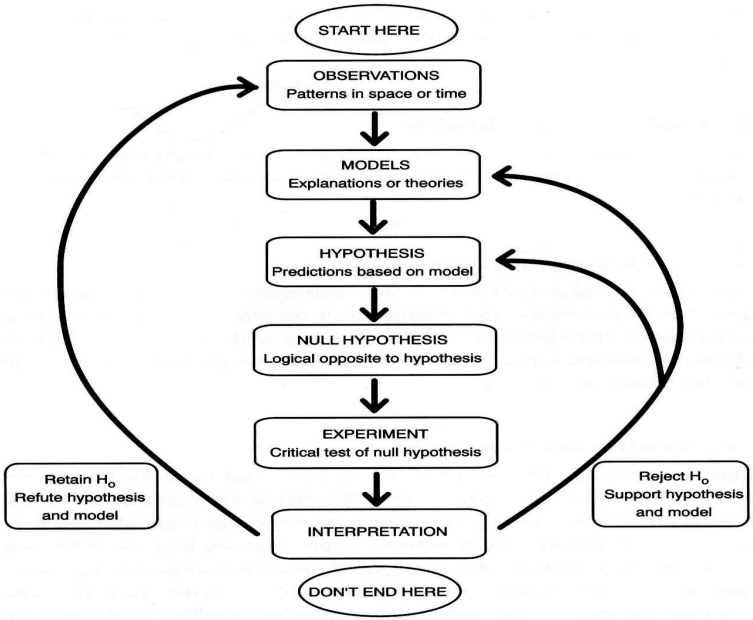
- или иначе

- Если $t > t_{crit}$, то вероятность получить такую стандартизированную разницу средних двух выборок из одной совокупности очень низка ($p<0.05$).

**Теперь у на есть инструмент для проверки статистических гипотез - статистический критерий, или статистический тест**

## ЧАСТЬ 2. Тестирование статистических гипотез

- Формулировка биологической гипотезы
- Численное выражение биологической гипотезы ($H$)
- Формулировка антигипотезы ($H_0$ - нулевой гипотезы)
- Тестирование нулевой гипотезы



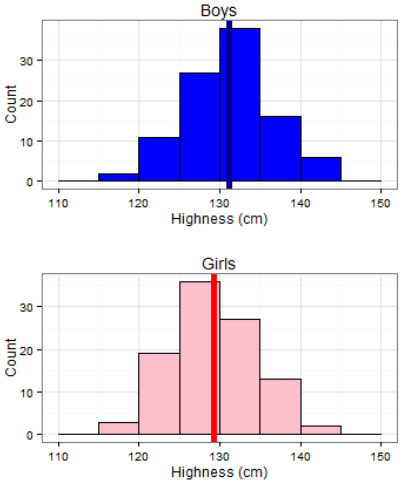

(Underwood, 1997)

## Простейший пример тестирования гипотезы

Создадим две выборки из популяций с нормальным распределением величин и заведомо отличающимися значениями $\mu$

```
set.seed(12345)
male <- rnorm(100, 130, 5)
female <- rnorm(100, 129,5)
```

Частотное распределение этих двух выборок выглядит так



## Сравним две выборки с помощью t-критерия Стьюдента

```
(t<-t.test(male, female))
```

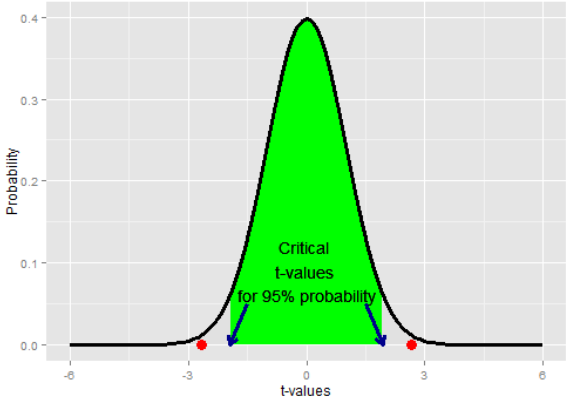
```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = 2.657, df = 196.2, p-value = 0.008522
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5157 3.4839
## sample estimates:
## mean of x mean of y
##     131.2     129.2
```

---

## Вопрос: Вероятность какого события отражает уровень значимости p=0.0085?

Уровень значимости p=0.0085

Это вероятность получения двух выборок, взятых из одной генеральной совокупности ($H_0$ верна), с такими же выборочными оценками (средняя и среднеквадратичное отклонение), которые мы имеем.




Полученное нами эмпирическое значение t = 2.657 не попадает в область, ограниченную критическими значениями!

Это значит, что ошибочный вывод о существовании различий мы будем делать не более, чем в 5% случаев, если подобный эксперимент будет проводиться многократно. И это нас устраивает, поскольку мы приняли $\alpha = 0.05$.

## Допущения (Assumptions) t-критерия

- 1. Нормальное распределение сравниваемых величин
- 2. Равенство дисперсий
- 3. Выборки должны быть сделаны независимо друг от друга

- t-критерий очень чувствителен к нарушению условия 1 и 2 если выборки имеют неравные объемы.
- Лучше сразу планировать сбор материала так, чтобы были сбалансированные выборки.

# Почему в полученных результатах нашего теста df=196.15 дробное число?

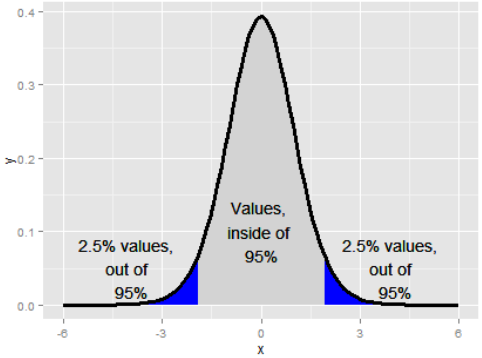
R автоматичски вводит поправку на разность дисперсий (используется модифицированная версия теста - Welch-test)

В этом тесте специально занижается df, что делает китерий более консервативным (то есть он "хуже" отвергает $H_0$)

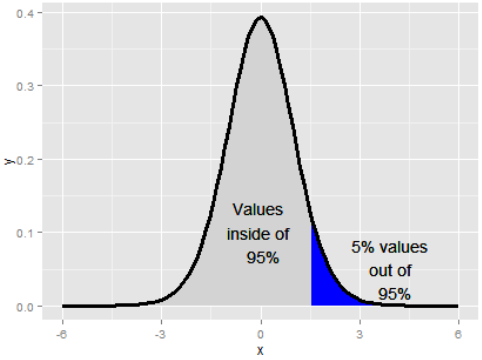
$$df = \frac{(\frac{sd_1}{\sqrt{n_1}} + \frac{sd_2}{\sqrt{n_2}})}{(sd_1/\sqrt{n_1})^2/(n_1+1) + (sd_2/\sqrt{n_2})^2/(n_2+1)} - 2$$

## Двусторонние и односторнние тесты

Двусторонний тест $H_0 : \mu_1 - \mu_2 = 0$ альтернатива $H : \mu_1 \neq \mu_2$ то есть может быть $\mu_1 > \mu_2$ и $\mu_1 < \mu_2$




Односторонний тест $H_0 : \mu_1 - \mu_2 = 0$ альтернатива $H : \mu_2 > \mu_1$




Отвержение $H_0$ происходит при меньшем значении t.

В случае с t-критерием будьте осторожны! Используя односторонние тесты, мы повышаем вероятность неправильного отвержения $H_0$

## Протокол применения t-критерия

- 1.Принимаем априорный пороговый уровень значимости, например $\alpha = 0.05$
- 2.Для двух сравниваемых выборок вычисляем средние значения и значения срднеквадратичного отклонения
- 3.Вычисляем эмпирическое значение t
- 4.Находим число степеней свободы для данного значения t:
- Если дисперсии равны, то $df = n_1 + n2 - 2$
- Если дисперсии не равны, $df = \frac{(\frac{sd_1}{\sqrt{n_1}} + \frac{sd_2}{\sqrt{n_2}})}{(sd_1/\sqrt{n_1})^2/(n_1+1)+(sd_2/\sqrt{n_2})^2/(n_2+1)} - 2$
- 5.Строим референсное t-распределение для даного значения $df$, характеризующее ситуацию для истинной $H_0$
- 6.Вычисляем величину уровня значимсти (p)

  Пункты 3-6 за нас может сделать функция `t.test()`
- 7.Отвергаем $H_0$ если $p < \alpha$, и считаем, что наблюдаются достоверные различия между средними

## ЧАСТЬ 3. Пермутационный метод тестирования гипотез

Пермутации - это перестановки.

Если две сравниваемые выборки взяты из одной совокупности, то обмен элементами между ними ничего не изменит. Степень различия между выборками (значение статистики) останется более или менее тем же самым.

Применим пермутационный метод к нашим двум выборкам, описывающим размеры мальчиков и девочек (`male` и `female`)

```
head (male)
```

```
## [1] 132.9 133.5 129.5 127.7 133.0 120.9
```

```
head (female)
```

```
## [1] 130.1 123.2 131.1 122.4 129.7 126.3
```

Введем статистику

$$d = |\bar{x_1} - \bar{x_2}|$$

```
d_initial <- abs(mean(male) - mean(female))
```

При сравнении векторов `male` и `female` d = 1.9998

При пермутациях мы должны поменять местами, например,

male[10] = 125.4 $\longleftrightarrow$ female[20] = 128.7.

А еще лучше поменять случайное количество элементов одной выборки на такое же количество элементов из другой выборки.

## Получаем распределение статистики $d_{perm}$

Для этого мы много раз случайно перемешиваем выборки и после каждой пермутации вычисляем значение статстики $d_{perm}$

```
Nperm=10000
dperm <- rep(NA, Nperm)

set.seed(12345)
for (i in 1:(Nperm-1))
  {
  BOX <- c(male,female)
  ord <-sample(1:200, 200)
  f <- BOX[ord[1:100]]
  m <- BOX [ord[101:200]]
  dperm[i]=abs(mean(m) - mean(f))  }
head(dperm)
```

```
## [1] 0.3409 0.4027 0.8399 1.4020 0.1568 1.4720
```

## Получаем распределение статистики $d_{perm}$

Посмотрим в конец этого вектора

```
tail(dperm)
```

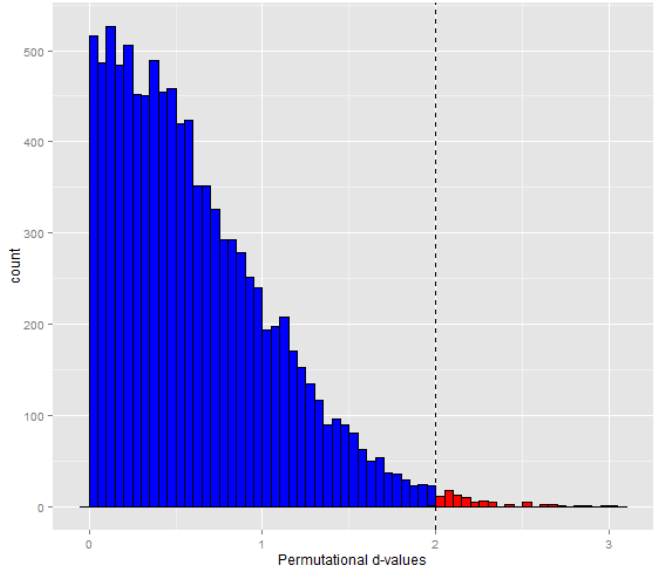
```
## [1] 0.31631 0.83119 0.07813 0.41032 0.09609      NA
```

Последнее 10000-е значение не заполнено!
В него надо вписать исходное, полученное до пермутаций, значение d = 1.9998.

Это необходимо, так как мы тестируем гипотезу о принадлежности этого значения случаному распределению.

```
dperm [Nperm] <- d_initial
```

**Получаем распределение статистики $d_{perm}$**



## Расчитаем величину уровня значимости

$$p_{perm} = \frac{N_{d_{perm}>=d}}{N_{perm}}$$

```
p_perm <- length(dperm[dperm >= d_initial] ) / Nperm
```

Итак, мы получили уровень значимости $p_{perm}$ = 0.0082

Сравним его с уровнем значимости, вычисленным с помощью параметрического t-критерия p=0.0085

Они оба близки и оба выявляют достоверные различия!

## Summary

- 1.Любой статистический критерий работает принципиально так же, как t-критерий: вычисляется значение тестовой статистики, которое сравнивается с референсным распределением, получающимся при истинности $H_0$
- 2.У любого статистического критерия есть свои условия примнимости (assumptions)
- 3. Не надо молиться на уровень значиости p<0.05!

## Что почтать

Гланц С. Медико-биологическая статистика. М: Практика, 1998. 459 с. (Есть в Сети!)