Линейная регрессия. Нормальное распределение. Критерий Стьюдента

Корпусные методы исследований языковых процессов

Даша Попова

30.11.2022

Importing data from a csv file

- > data <-- read.csv(“path/filename.csv”, header = TRUE, sep=“;”)
- > data
- > print(is.data.frame(data))
- > print(ncol(data))
- > print(nrow(data))
- > d1 = subset(data, height>155)

Paired vectors: Functional relations: Linear regression

- Regression analysis is a very widely used statistical tool to establish a relationship model between two variables.
- One of these variable is called predictor variable whose value is gathered through experiments.
- The other variable is called response variable whose value is derived from the predictor variable.
- In Linear Regression these two variables are related through an equation.
- Mathematically a linear relationship represents a straight line when plotted as a graph.
- A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.
- The general mathematical equation for a linear regression is $y = ax + b$.
- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

What is linear regression?

A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

You make this kind of relationships in your head all the time, for example when you calculate the age of a child based on her height, you are assuming the older she is, the taller she will be.

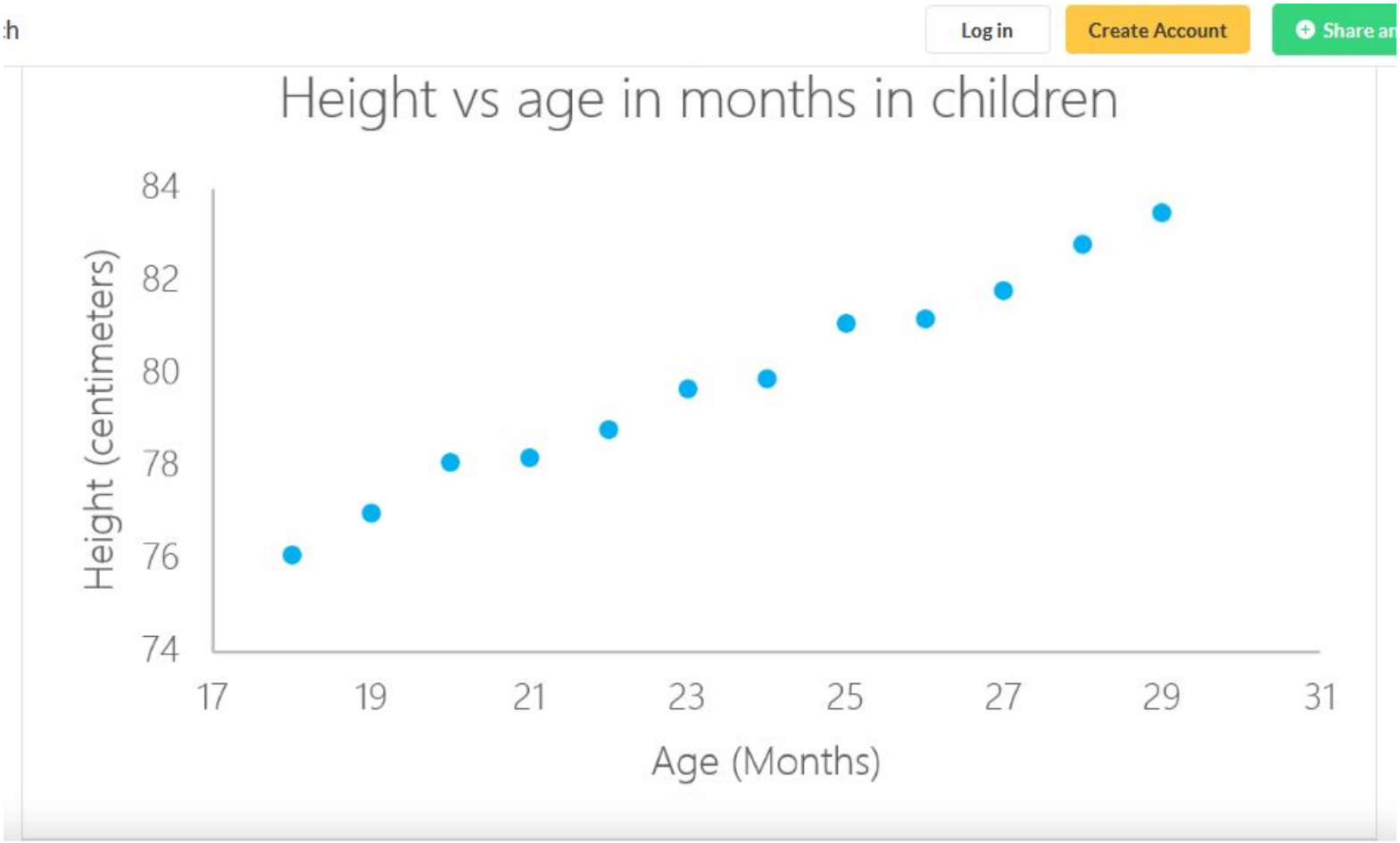
Linear regression is one of the most basic statistical models out there, its results can be interpreted by almost everyone, and it has been around since the 19th century. This is precisely what makes linear regression so popular. It's simple, and it has survived for hundreds of years.

https://www.datacamp.com/community/tutorials/linear-regression-R

Linear regression

- Not every problem can be solved with the same algorithm.
- Linear regression assumes that there exists a linear relationship between the response variable and the explanatory variables.
- This means that you can fit a line between the two (or more variables).

https://www.datacamp.com/community/tutorials/linear-regression-R



https://www.datacamp.com/community/tutorials/linear-regression-R

Linear regression

$$\text{Height} = a + \text{Age} * b$$

- In this case, “a” and “b” are called the intercept and the slope respectively. With the same example, “a” or the intercept, is the value from where you start measuring. Newborn babies with zero months are not zero centimeters necessarily; this is the function of the intercept. The slope measures the change of height with respect to the age in months. In general, for every month older the child is, his or her height will increase with “b”.

- A linear regression can be calculated in R with the command **lm**.

https://www.datacamp.com/community/tutorials/linear-regression-R

Linear regression: coefficients

```
Call:
lm(formula = height ~ age, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

In the red square, you can see the values of the intercept ("a" value) and the slope ("b" value) for the age. These "a" and "b" values plot a line between all the points of the data. So in this case, if there is a child that is 20.5 months old, a is 64.92 and b is 0.635, the model predicts (on average) that its height in centimeters is around 64.92 + (0.635 * 20.5) = 77.93 cm.

https://www.datacamp.com/community/tutorials/linear-regression-R

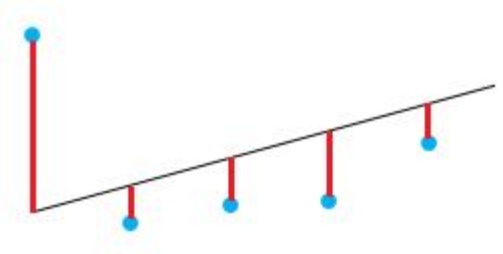
Linear regression: coefficients

- Another aspect to pay attention to in your linear models is the p-value of the coefficients.
- The smaller the p-value the better the predictor

https://www.datacamp.com/community/tutorials/linear-regression-R

Linear regression: residuals

A good way to test the quality of the fit of the model is to look at the residuals or the differences between the real values and the predicted values. The straight line in the image above represents the predicted values. The red vertical line from the straight line to the observed data value is the residual.



https://www.datacamp.com/community/tutorials/linear-regression-R

Linear regression: residuals

The idea in here is that the sum of the residuals is approximately zero or as low as possible. In real life, most cases will not follow a perfectly straight line, so residuals are expected. In the R summary of the lm function, you can see descriptive statistics about the residuals of the model, following the same example, the red square shows how the residuals are approximately zero.

```
Call:
lm(formula = height ~ age, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

https://www.datacamp.com/community/tutorials/linear-regression-R

How to test if your linear model has a good fit?

One measure very used to test how good is your model is the coefficient of determination or R^2. This measure is defined by the proportion of the total variability explained by the regression model.

$$R^2 = \frac{\text{Explained Variation of the model}}{\text{Total variation of the model}}$$

This can seem a little bit complicated, but in general, for models that fit the data well, R^2 is near 1. Models that poorly fit the data have R^2 near 0.

In some fields, an R^2 of 0.5 is considered good.

https://www.datacamp.com/community/tutorials/linear-regression-R

Paired vectors: Functional relations: Linear regression

- file *lr.csv* – read it into R
- relation <- lm(data$height~data$weight)
- > plot(data$weight,data$height,col = "blue",main = "Height & Weight Regression",abline(lm(data$height~data$weight)),cex = 1.3,pch = 16,xlab = "Weight in Kg",ylab = "Height in cm")
- > summary(relation)
- The closer Multiple R-squared value is to 1, the stronger the linear regression.
- plot(relation$residuals, pch = 16, col = "red") #Ideally, when you plot the residuals, they should look random. Otherwise means that maybe there is a hidden pattern that the linear model is not considering.

- Is there a linear dependency between *sleep* and *grade*?

point shapes is pch

- pch = 0,square
- pch = 1,circle
- pch = 2,triangle point up
- pch = 3,plus
- pch = 4,cross
- pch = 5,diamond
- pch = 6,triangle point down
- pch = 7,square cross
- pch = 8,star
- pch = 9,diamond plus
- pch = 10,circle plus
- pch = 11,triangles up and down
- pch = 12,square plus
- pch = 13,circle cross
- pch = 14,square and triangle down
- pch = 15, filled square
- pch = 16, filled circle
- pch = 17, filled triangle point-up
- pch = 18, filled diamond
- pch = 19, solid circle
- pch = 20,bullet (smaller circle)
- pch = 21, filled circle blue
- pch = 22, filled square blue
- pch = 23, filled diamond blue
- pch = 24, filled triangle point-up blue
- pch = 25, filled triangle point down blue

The following arguments can be used to change the **color** and the **size** of the points :

- **col** : color (code or name) to use for the points
- **bg** : the **background** (or fill) color for the open **plot symbols**. It can be used only when **pch = 21:25.**
- **cex** : the size of **pch symbols**
- **lwd** : the **line width** for the **plotting symbols**

```
plot(relation$residuals, pch = 23, col= "red", bg = "yellow", lwd= 2)
```

http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r

Коэффициент корреляции Пирсона

- Коэффициент корреляции Пирсона: cor(data$height,data$weight)

- базовый коэффициент ассоциации переменных, однако стоит помнить, что он дает неправильную оценку, если связь между переменными нелинейна

- если коэффициент положительный — связь между переменными положительная (чем **больше** x, тем **больше** y),
- если коэффициент отрицательный — связь между переменными отрицательная (чем **больше** x, тем **меньше** y);
  - если модуль коэффициента близок к 1 или ей равен — связь между переменными сильная,
  - если модуль коэффициента близок к 0 или ему равен — связь между переменными слабая.

- Поиграем: http://guessthecorrelation.com/

Случайная величина или переменная

- ***Случайная переменная*** — это величина, которая может принимать любое из набора взаимоисключающих значений с определенной вероятностью.

- Распределение вероятности показывает вероятности всех возможных значений случайной переменной. Это теоретическое распределение, которое выражено математически и имеет *среднее* и *дисперсию* — аналоги среднего и дисперсии в эмпирическом распределении.

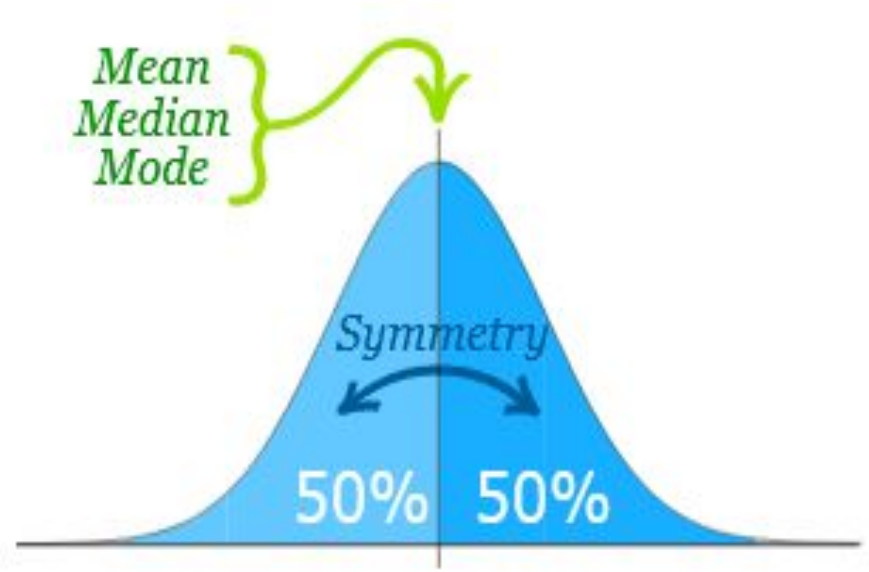
http://statistica.ru/theory/raspredeleniya-veroyatnostey/

Нормальное распределение

- In a random collection of data from independent sources, it is generally observed that the distribution of data is normal.
- Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shape curve.
- The center of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph.
- This is referred as ***normal distribution*** in statistics.

https://www.tutorialspoint.com/r/r_normal_distribution.htm

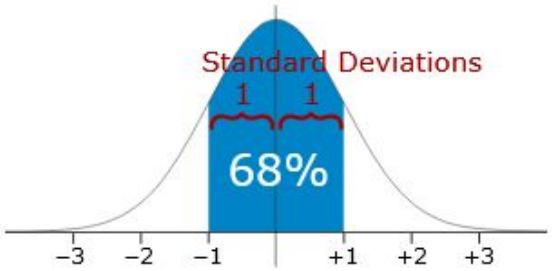
Normal distribution



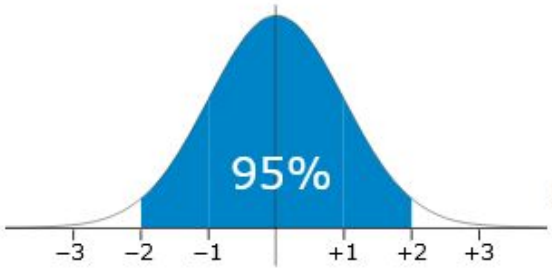
- "Bell curve"
- e.g., heights of people
- e.g., marks on a test
- mean = median = mode
- symmetry about the center
- 50% of values less than the mean
  and 50% greater than the mean

https://www.mathsisfun.com/data/standard-normal-distribution.html

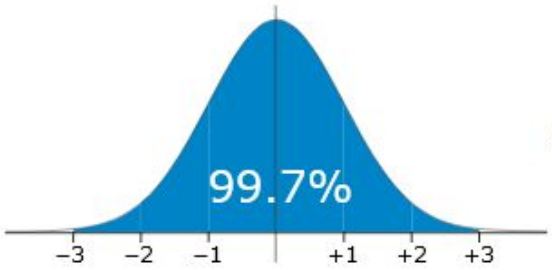
The standard deviation is a measure of how spread out numbers are



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99,7% of values are within
3 standard deviations of the mean

https://www.mathsisfun.com/data/standard-deviation.html

Standardizing

https://www.mathsisfun.com/data/standard-normal-distribution.html

Нормальное распределение

- dnorm(x, mean, sd)
- rnorm(n, mean, sd)
- **x** is a vector of numbers;
- **n** is number of observations(sample size);
- **mean** is the mean value of the sample data. It's default value is zero;
- **sd** is the standard deviation. It's default value is 1.

https://www.tutorialspoint.com/r/r_normal_distribution.htm

Нормальное распределение

- *# create a sequence of numbers between -10 and 10 incrementing by 0.1*
- > x <- seq(-10, 10, by = .1)
- *# choose the mean as 2.5 and standard deviation as 0.5*
- > y <- dnorm(x, mean = 2.5, sd = 0.5)
- > plot(x,y)

- rnorm is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.
- # Create a sample of 50 numbers which are normally distributed
- > y <- rnorm(100, mean =170, sd = 10)
- # Plot the histogram for this sample.
- >hist(y, main = "Normal DIstribution")

- #to test if the variable is normally distributed
- >shapiro.test(y) #Shapiro-Wilk Test for Normality in R
- The data is normal if the p-value is above 0.05.

https://www.tutorialspoint.com/r/r_normal_distribution.htm

T-test, критерий Стьюдента

- t-test is used to compare the means of two groups under the assumption that both samples are random, independent, and come from normally distributed population with unknown but equal variances
- Любопытно, что создал этот метод Уильямом Госсет - химик, приглашенный работать на фабрику Guinness. Разработанный им тест служил изначально для оценки качества пива. Однако, химикам фабрики запрещалось независимо публиковать научные работы под своим именем. Поэтому в 1908 году Уильям опубликовал свою статью в журнале "Biometrika" под псевдонимом "Стьюдент". Позже, выдающийся математик и статистик Рональд Фишер доработал метод, который затем получил массовое распространение под названием Student's t-test
- **Критерий Стьюдента (t-тест)** - это статистический метод, который позволяет сравнивать средние значения двух выборок и на основе результатов теста делать заключение о том, различаются ли они друг от друга статистически или нет. Предполагает, что данные выборок имеют **нормальное распределение**.

https://samoedd.com/soft/r-t-test

T-test, критерий Стьюдента

- **Одновыборочный критерий Стьюдента (one-sample t-test)**
- Одновыборочный t-тест следует выбирать, если Вы сравниваете выборку с общеизвестным средним.
- Например, отличается ли средний возраст жителей Северо-Кавказского Федерального округа от общего по России.
- Нулевая гипотеза: различий между средним ожидаемым уровнем продолжительности по России и республикам Северного Кавказа нет.
- Если различия существуют, то для того, чтобы считать их статистически значимыми *p-value* должно быть менее 0.05.
- >rosstat <-c(79.42, 75.83, 74.16, 73.91, 73.82, 73.06, 72.01) #продолжительность жизни в республиках Кавказа
- >shapiro.test(rosstat) #тестируем на нормальность
- >t.test(rosstat, mu = 70.93) #mu = средняя продолжительность жизни в России
- Результаты t-теста говорят о том, что средняя ожидаемая продолжительность жизни у жителей Северного Кавказа (74.6 лет) действительно выше, чем в среднем по России (70.93 лет), а результаты теста являются статистически значимыми ($p < 0.05$).

https://samoedd.com/soft/r-t-test

T-test, критерий Стьюдента

- **Двувыборочный для независимых выборок (independent two-sample t-test)**
- Двувыборочный t-тест используется, когда Вы сравниваете две независимые выборки.
- Допустим, мы хотим узнать, отличается ли урожайность картофеля на севере и на юге какого-либо региона. Для этого, мы собрали данные с 40 фермерских хозяйств: 20 из которых располагались на севере и сформировали выборку "North", а остальные 20 - на юге, сформировав выборку "South".
- >North <- c(122, 150, 136, 129, 169, 158, 132, 162, 143, 179, 139, 193, 155, 160, 165, 149, 173, 173, 141, 166)
- >shapiro.test(North)
- >South <- c(170, 163, 178, 150, 166, 142, 157, 149, 151, 164, 163, 161, 159, 139, 180, 155, 144, 139, 151, 160)
- >shapiro.test(South)
- >t.test(North, South)
- Результаты теста говорят о том, что средняя урожайность картофеля на севере статистически не отличается от урожайности на юге (p = 0.6339).

- Какой тест используется вместо критерия Стьюдента в случае, если данные распределены не нормально?

https://samoedd.com/soft/r-t-test

T-test, критерий Стьюдента

- **Двувыборочный для зависимых выборок (dependent two-sample *t*-test)**
- Третий вид t-теста используется в том случае, если элементы выборок зависят друг от друга.
- Он идеально подходит для **проверки повторяемости результатов** эксперимента: если данные повтора статистически не отличаются от оригинала, то повторяемость данных высокая.
- Также двувыборочный критерий Стьюдента для зависимых выборок широко применяется **в медицинских исследованиях** при изучении эффекта лекарства на организм до и после приема.
- Для того, чтобы запустить его в R, следует ввести все ту же функцию t.test. Однако, в скобках, после таблиц данных, следует ввести дополнительный аргумент paired = TRUE. Этот аргумент говорит о том, что Ваши данные зависят друг от друга. Например:
- >t.test(experiment, povtor.experimenta, paired = TRUE)
- >t.test(davlenie.do.priema, davlenie.posle.priema, paired = TRUE)
- Также в функции t.test существует два дополнительных аргумента, которые могут улучшить качество результатов теста: var.equal и alternative.
- Если вы знаете, что вариация между выборками равна, вставьте аргумент var.equal = TRUE.
- Если же вы хотите проверить гипотезу о том, что разница между средними в выборках значительно меньше или больше 0, то введите аргумент alternative="less" или alternative="greater" (по умолчанию альтернативная гипотеза говорит о том, что выборки просто отличаются друг от друга: alternative="two.sided").

https://samoedd.com/soft/r-t-test

Рисуем дерево критериев

Категориальные данные?

Да:

критерий хи-квадрат

chisq.test

нулевая гипотеза: переменные независимы

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: переменные зависимы

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Тест Шапиро-Уилка(Вилка)

shapiro.test()

Нулевая гипотеза: распределена нормально

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: распределена не нормально

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Нет:

Критерий Вилкоксона

wilcox.test(x, y)

Нулевая гипотеза H_0: медиана разницы в популяции равна нулю/рейтинги статистически не различаются

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: медиана разницы в популяции не равна нулю

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Да:

Критерий Стьюдента

t.test()

Нулевая гипотеза H_0: различий нет

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: сравниваемые выборки/выборка и величина различаются

Научное знание

- Системность

А.А. Зализняк и С.П. Капица: мин. 5 – 7 и 19 – 24.30: https://www.youtube.com/watch?v=2OmVPytZbGg

- Воспроизводимость
- Верифицируемость – возможность подтвердить утверждение
- Фальсифицируемость – принципиальная возможность опровержения утверждения, опровергаемость, критерий Поппера, который предложил этот критерий в 1935 году.