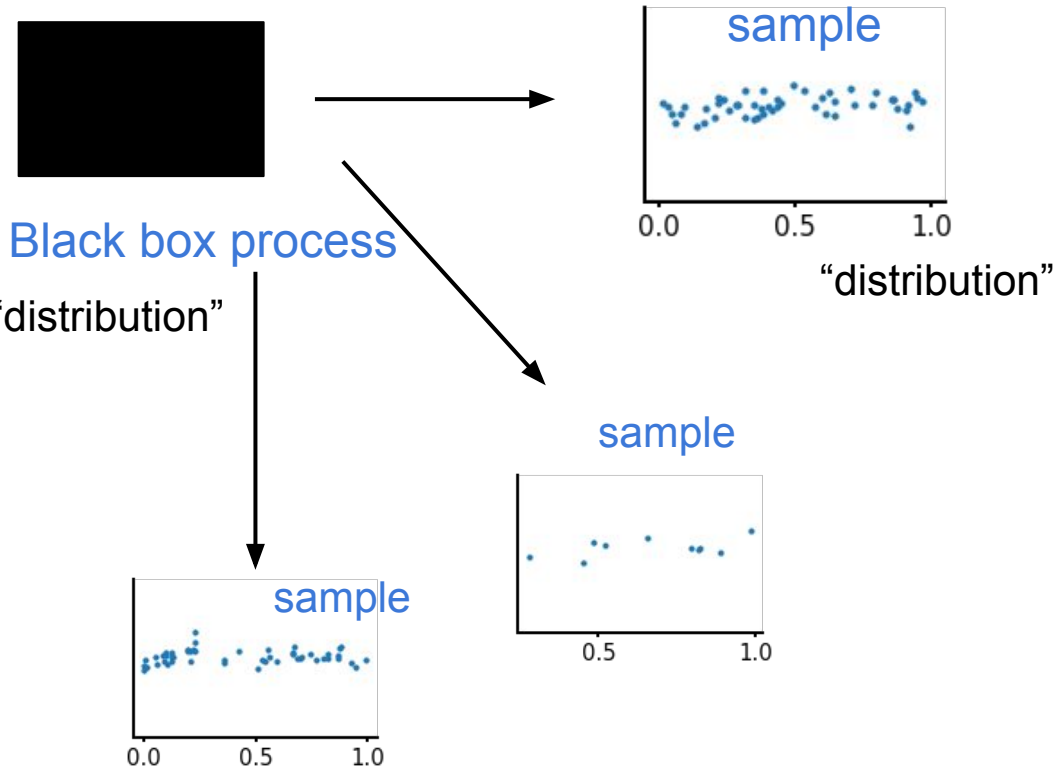
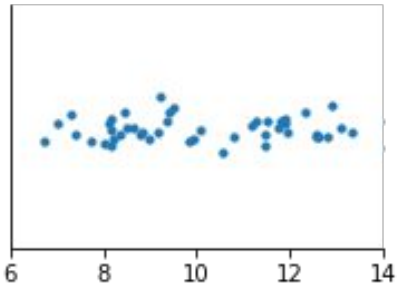
Statistical Data Analysis Lab

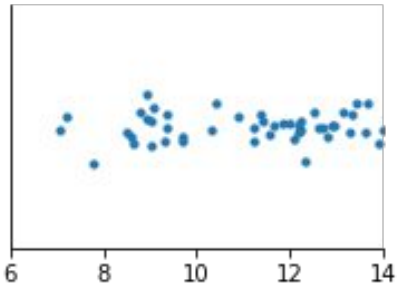
Statistical tests and assumptions 1

Anna Tkachev


sample
0.0
0.5
1.0
Black box process
“distribution”
“distribution”
sample
0.5
1.0
sample
0.0
0.5
1.0

Statistical testing

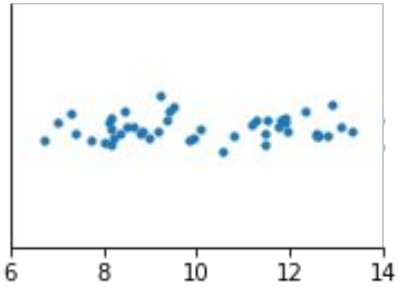




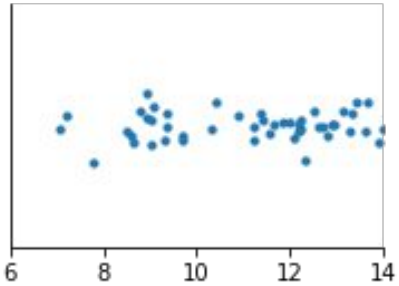
Two random samples

Statistical testing

Black box process



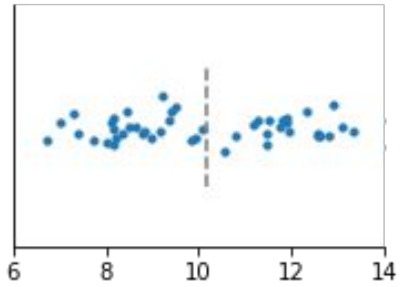
Black box process

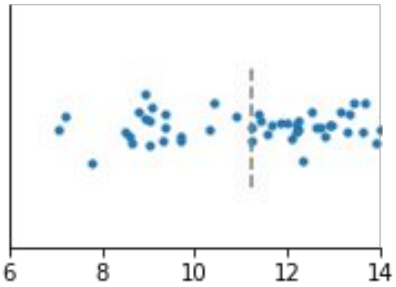


Two random samples

Q: Is the underlying distribution the same?

Statistical testing

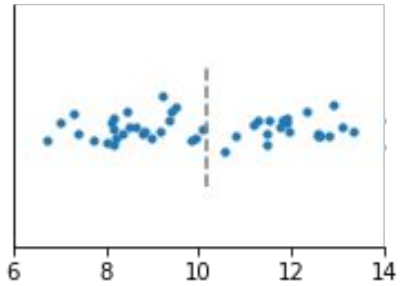


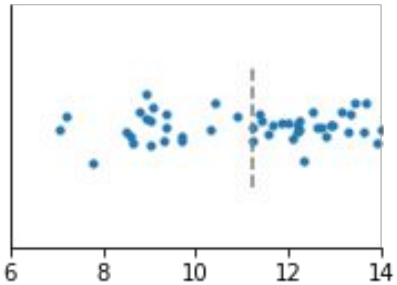


Q: Is the underlying distribution the same?

Or more specifically: is the mean value of the underlying distribution the same?

Statistical testing





Q: Is the underlying distribution the same?

Or more specifically: is the mean value of the underlying distribution the same?

One such test:

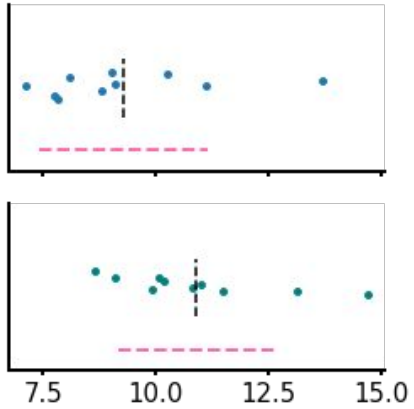
T-test

(Welch-test)

How likely is it for a normal distribution to generate two sample means that are so far apart?

Statistical testing

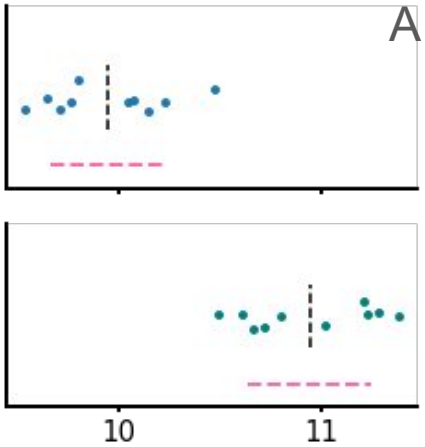
How likely is it for a normal distribution to generate two sample means that are so far apart?

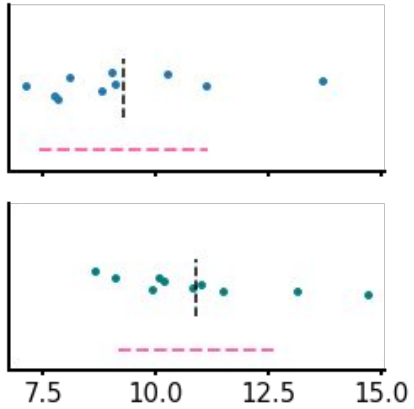


Statistical testing

How likely is it for a normal distribution to generate two sample means that are so far apart?

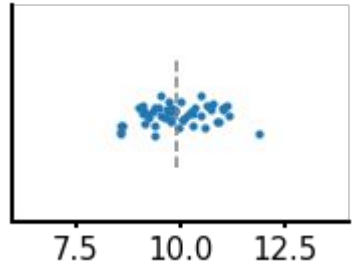
Which case is more unlikely?

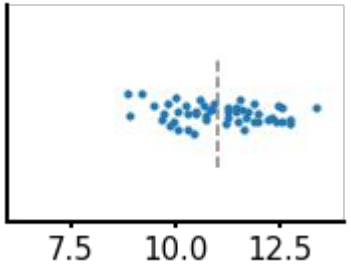




Statistical testing

How likely is it for a normal distribution to generate two sample means that are so far apart?

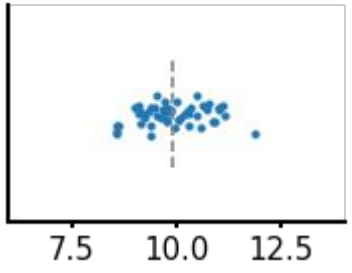




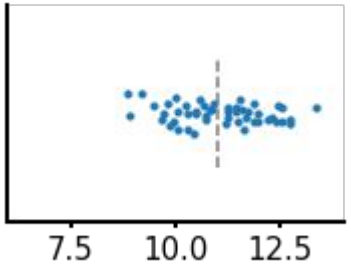
Statistical testing

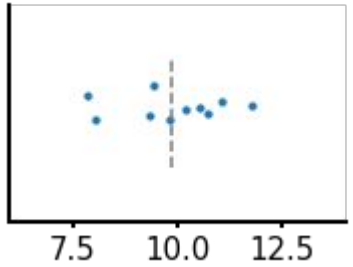
How likely is it for a normal distribution to generate two sample means that are so far apart?

Which case is more unlikely?

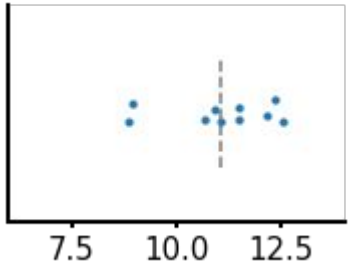


A

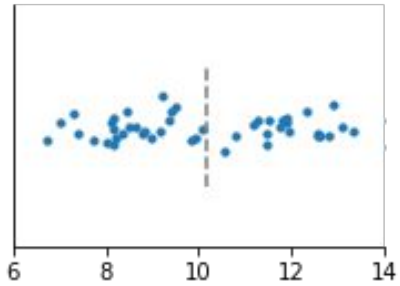


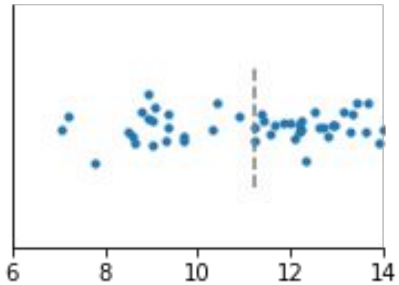


B



Statistical testing





Q: Is the underlying distribution the same?

Or more specifically: is the mean value of the underlying distribution the same?

One such test:

T-test

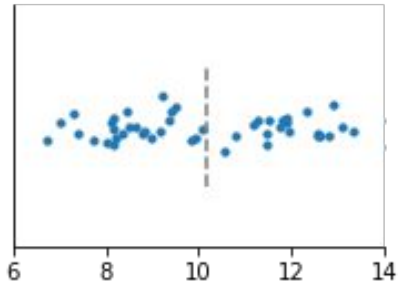
(Welch-test)

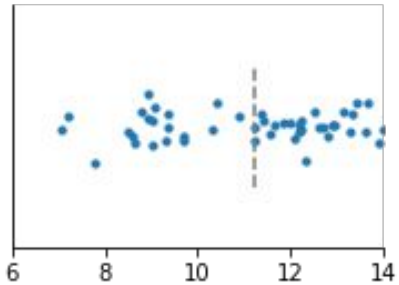
$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

T-statistics is calculated from the data:
-The sample standard deviation is used to normalize the difference in sample means
-Depends on sample size n

How likely is it for a normal distribution to generate two sample means that are so far apart?

Statistical testing





Q: Is the underlying distribution the same?

Or more specifically: is the mean value of the underlying distribution the same?

One such test:

T-test

(Welch-test)

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

T-statistics is calculated from the data:
-The sample standard deviation is used to normalize the difference in sample means
-Depends on sample size n

How likely is it for a normal distribution to generate two sample means that are so far apart?

The larger the t-statistics, the more unlikely it is

Statistical testing

In statistical testing in general, we calculate:

Pvalues :
(not an exact definition)

"The probability of the observed distribution of sample data points, if the underlying "black box" distribution of the two samples is the same"

Statistical testing

In statistical testing in general, we calculate:

Pvalues :
(not an exact definition)

"The probability of the observed distribution of sample data points, if the underlying "black box" distribution of the two samples is the same"

In two-sample statistical testing, we calculate:

"The probability observing such extreme differences in the distribution of sample data points, if the underlying "black box" distribution of the two samples is the same"

Statistical testing

In statistical testing in general, we calculate:

Pvalues :
(not an exact definition)

"The probability of the observed distribution of sample data points, if the underlying "black box" distribution of the two samples is the same"

In two-sample statistical testing, we calculate:

"The probability observing such extreme differences in the distribution of sample data points, if the underlying "black box" distribution of the two samples is the same"

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In the case of t-test:

The probability of the observing such t-statistics, or one even further from 0, if the underlying "black box" distributions are normal and the mean value is the same.

Statistical testing

-P value is a probability value, so taking values from 0 to 1

-It is actually a random variable, because it is calculated from the data, which is a sample from a random variable.

Statistical testing

-P value is a probability value, so taking values from 0 to 1

-It is actually a random variable, because it is calculated from the data, which is a sample from a random variable.

-p value is small => the probability of observing such difference in the data is low, if the distributions of the two samples were the same. For this reason this is evidence that the distributions are actually different.

“Hypothesis testing”, the hypothesis is that the distribution of the samples is the same.
In the case of t-test, the hypothesis is that the mean value is the same.
Normality is just assumed

Statistical testing

-P value is a probability value, so taking values from 0 to 1

-It is actually a random variable, because it is calculated from the data, which is a sample from a random variable.

-p value is small => the probability of observing such difference in the data is low, if the distributions of the two samples were the same. For this reason this is evidence that the distributions are actually different.

But because it's a probability, this means sometimes the p value is small even when the distributions are actually the same.

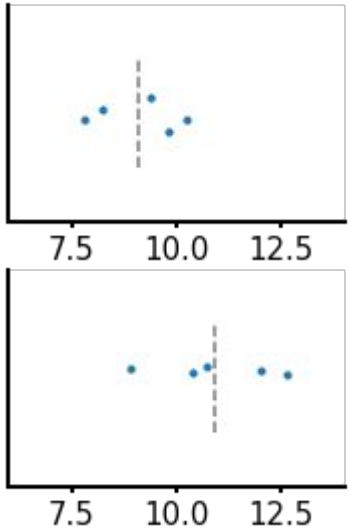
Statistical significance

"Hypothesis testing", the hypothesis is that the distribution of the samples is the same.
In the case of t-test, the hypothesis is that the mean value is the same.
Normality is just assumed

Statistical testing

- **Statistical significance**
- **Power**

The underlying distributions actually have different means. But our t-test (for this example) will have p-value <0.05 in only 50% of the cases. This value of the test is called **power,** and in this case can be considered to be quite low (ex. because of small sample size)

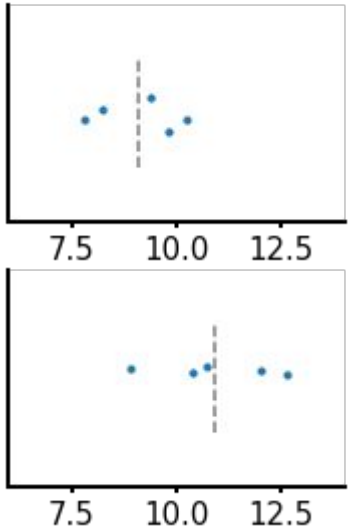


Statistical testing

- **Statistical significance**
- **Power**

Statistical significance is calculated during statistical testing.
Power is only calculated if you want to estimate how many data points you want to collect.

The underlying distributions actually have different means. But our t-test (for this example) will have p-value <0.05 in only 50% of the cases. This value of the test is called **power,** and in this case can be considered to be quite low (ex. because of small sample size)



Statistical testing

- **Statistical significance**
- **Power**

Collected data and performed test involving two samples

"Is the underlying distribution of the two collected samples is the same?"

Result of statistical testing is significant
=> there is evidence that your two samples come from different distributions

Result of statistical testing is not significant
=> there is no evidence that your two samples come from different distributions

But we make no conclusion about how likely it is that the distributions are the same.

Assumptions of t-test (Welch test)

All statistical tests have assumptions

-The observations in each of the two samples are independent

-Each of the two samples come from normal distributions

-The variances of both distributions is the same
(not required in the case of Welch test)

Mann Whitney U test

All statistical tests have assumptions

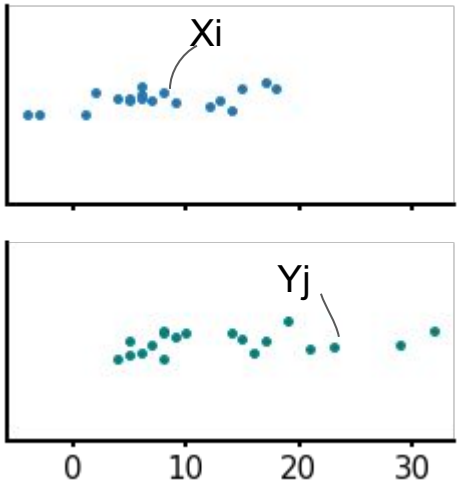
-The observations in each of the two samples are independent

~~-Each of the two samples come from normal distributions~~

Mann Whitney U test

In “Hypothesis testing”, the hypothesis is that the distribution of the samples is the same.

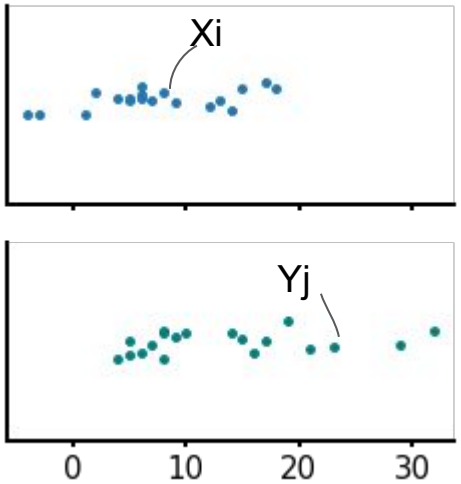
Hypothesis being tested:



Mann Whitney U test

In “Hypothesis testing”, the hypothesis is that the distribution of the samples is the same.

Hypothesis being tested:



Ranks instead of data values, for each of the two samples.

value	rank
-1	1
-0.1	2
1	3
2	4

…

19	rank
22	35
23	36
129	37
32	38

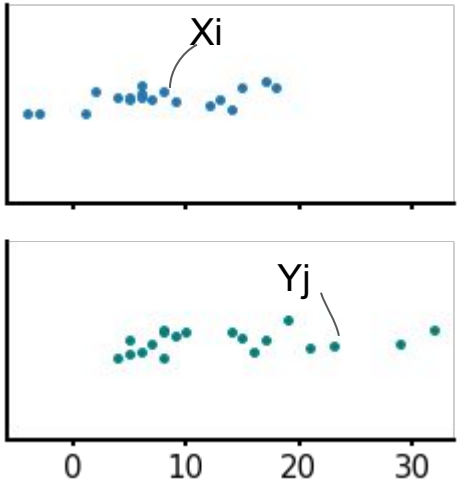
Mann Whitney U test

In “Hypothesis testing”, the hypothesis is that the distribution of the samples is the same.

Hypothesis being tested:

Is the mean rank the same for the two distributions?

(for two given points from the two distributions, will one be larger than the other, on average? ex. Yi>Xi)

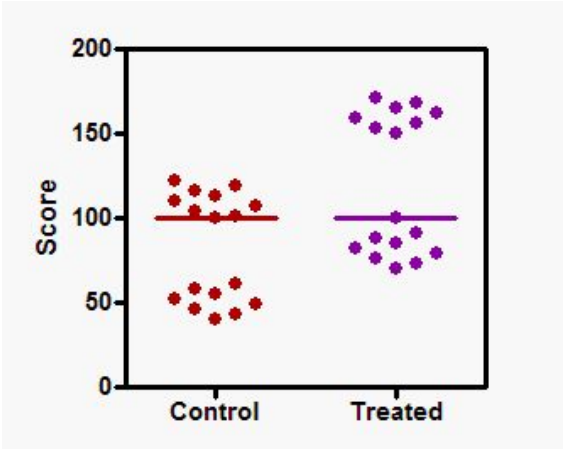


Ranks instead of data values, for each of the two samples.

value	rank
-1	1
-0.1	2
1	3
2	4

...

19	rank
22	35
23	36
129	37
32	38


200
150
100
50
0
Score
Control
Treated

Examples

Other types of tests

var1	var2
X1	Y1
X2	Y2
X3	Y3
X4	Y4

…

Xi - data points in first variable
Yi - data points in second variable

Are two variables related?

Pearson correlation test:
is there a linear relationship between two variables?

Other types of tests

var1	var2
X1	Y1
X2	Y2
X3	Y3
X4	Y4

...

Xi - data points in first variable
Yi - data points in second variable

Are two variables related?

Pearson correlation test:
is there a linear relationship between two variables?

In the process we calculate the *correlation coefficient:*

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson correlation test

In the process we calculate the *correlation coefficient:*

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

-r^2 represents proportion of variance from one variable that is explained by the other

-pvalue of the tests represent how likely it is that the r is in fact, not equal to 0

Pearson correlation test

In the process we calculate the *correlation coefficient:*

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

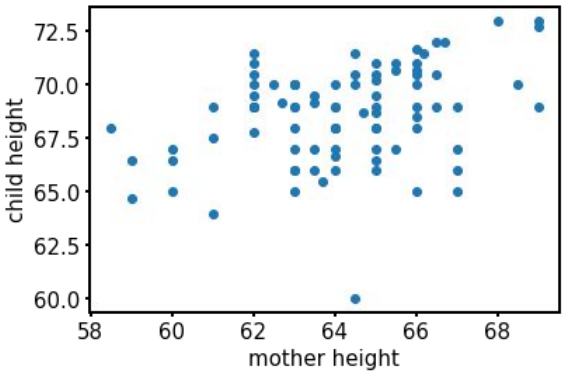
-r^2 represents proportion of variance from one variable that is explained by the other

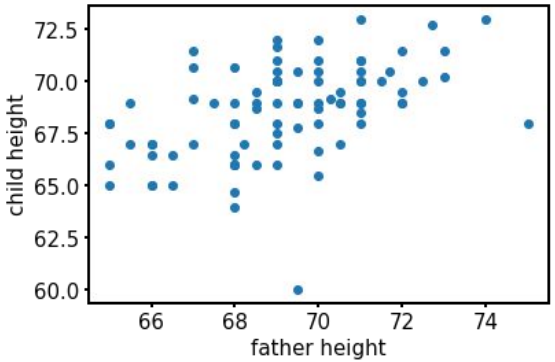
-pvalue of the tests represent how likely it is that the r is in fact, not equal to 0

Higher r = more linear relationship

pvalue = the probability that the (theoretical) r is in fact ≠ 0

Pearson correlation vs scatter plot

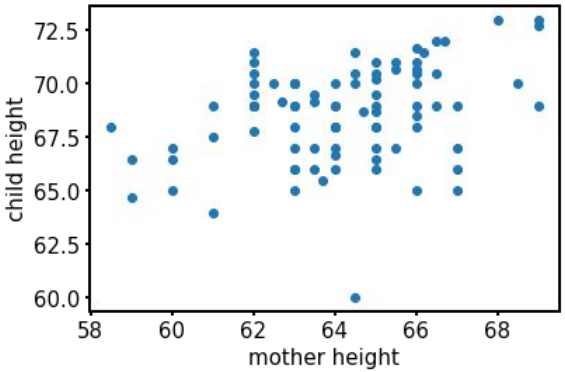


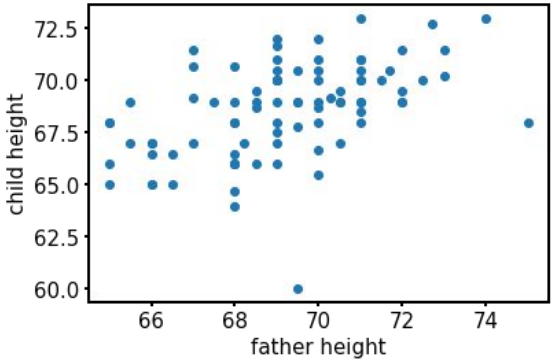


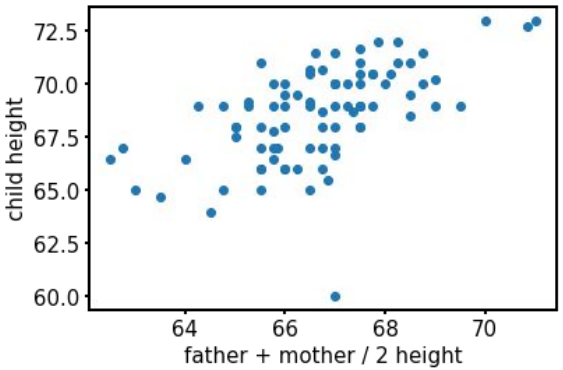
family	father	mother	children	childNum	gender	childHeight
005	75.0	58.5	6	3	male	68.0
007	74.0	68.0	6	3	male	73.0
016	73.0	65.0	9	3	male	70.2
017	73.0	64.5	6	3	male	71.5
020	72.7	69.0	8	3	male	72.7
...	...	...	...	...	...	...
189	65.0	66.0	5	3	male	65.0
190	65.0	65.0	9	3	male	68.0

Male children, Galton dataset (n = 87):

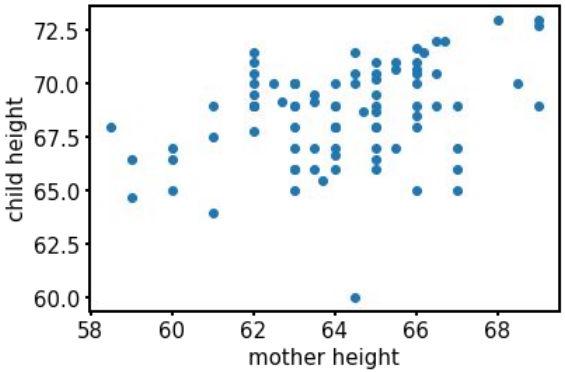
Pearson correlation vs scatter plot

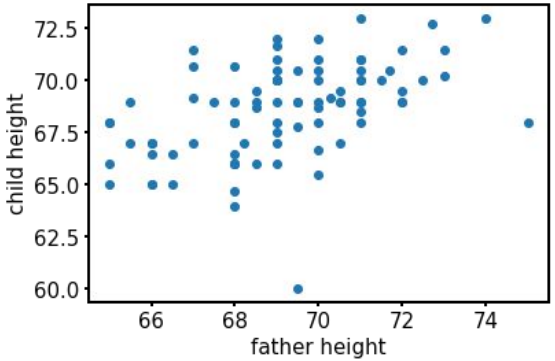


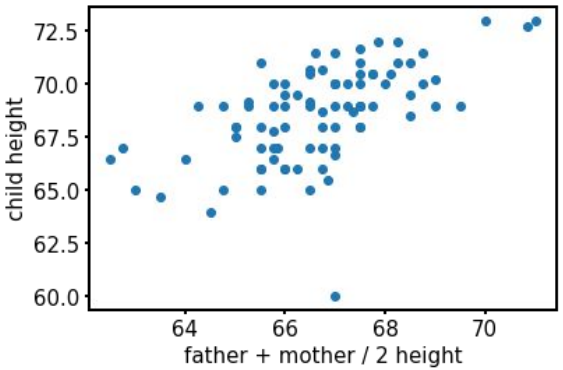




Male children, Galton dataset (n = 87):

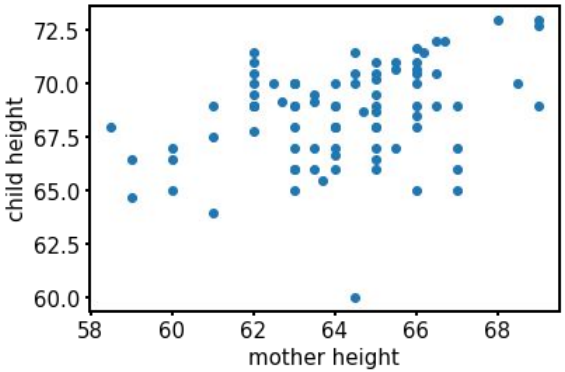


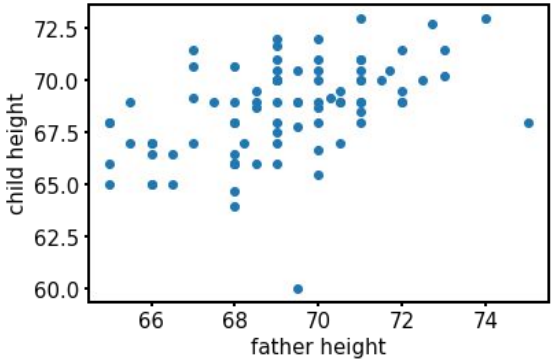


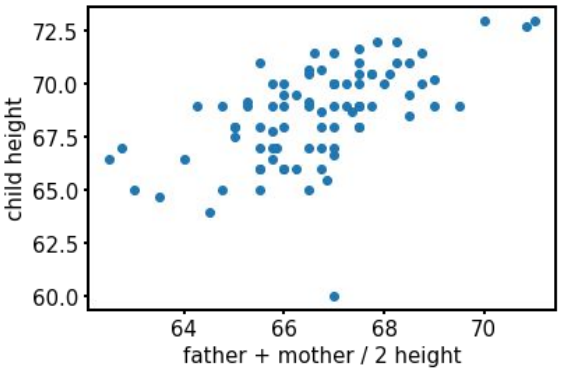


Correlation coefficient = 0.35 = 0.48 = 0.58

Male children, Galton dataset (n = 87):







Correlation coefficient = 0.35 = 0.48 = 0.58

P value = 0.00079 = 2e-06 = 3e-09

Pearson correlation

Requires some assumptions for the validity of the calculated **p-values** (but not necesseirly to calculate pearson correlation coefficients)

-The observations in each of the two samples are independent

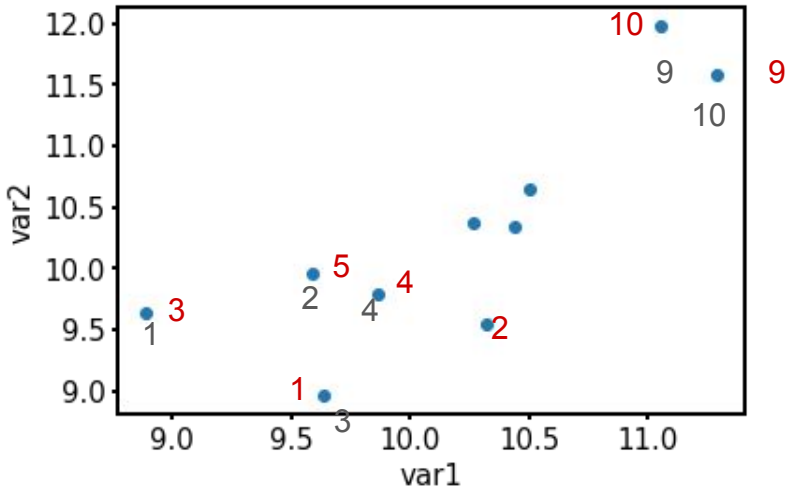
- Each of the two samples come from normal distributions

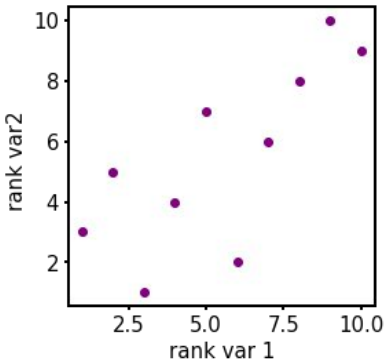
Spearman correlation

Are two variables related?

Spearman correlation test:
is there a monotonic relationship between two variables?

Similarly to the Mann Whitney U test, values are first converted to ranks, and the relationship between the ranks for the two variables is assessed





Examples

Other types of tests

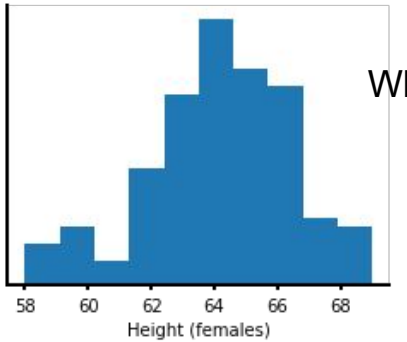
Testing for similarity of disrtibutions

or

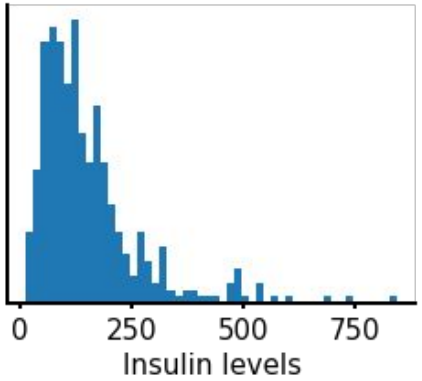
Testing if sample has an underlying normal distribution

As all statistical tests, depends on sample size

Shapiro test for normality and sample size

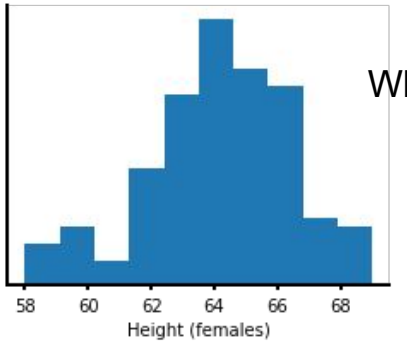


Whole dataset

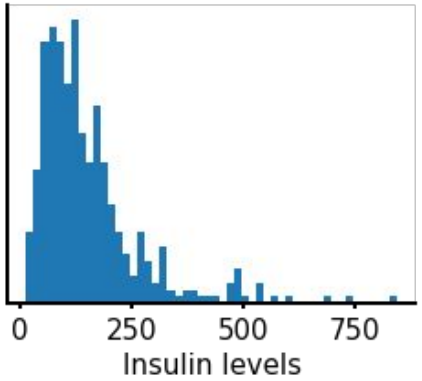


Whole dataset

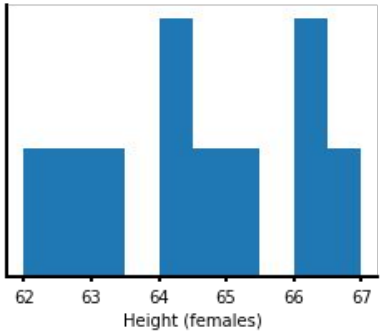
Shapiro test for normality and sample size



Whole dataset

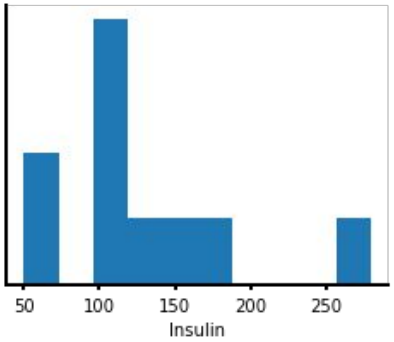


Whole dataset



Subset of 10 samples

pv= 0.85



Subset of 10 samples

pv= 0.11