Statistics

...

Statistical methods

Statistical methods

Statistics is the mathematical discipline that studies how collected data can be analyzed rigorously, to prove or disprove a hypothesis.

The use of statistical methods to ensure accuracy of scientific results is present in almost every publication, and certainly in every field. However, especially with new technologies, such as machine learning, it is often hard to judge the results of a statistical test in order to accept or reject a scientific hypothesis.

Statistical methods

Important points to keep in mind when analyzing a scientific dataset:

1. Data are noisy
2. Sample size is important
3. Metadata, units and normalization of data are important
4. The appropriate method of analysis depends on the dataset

Statistical methods

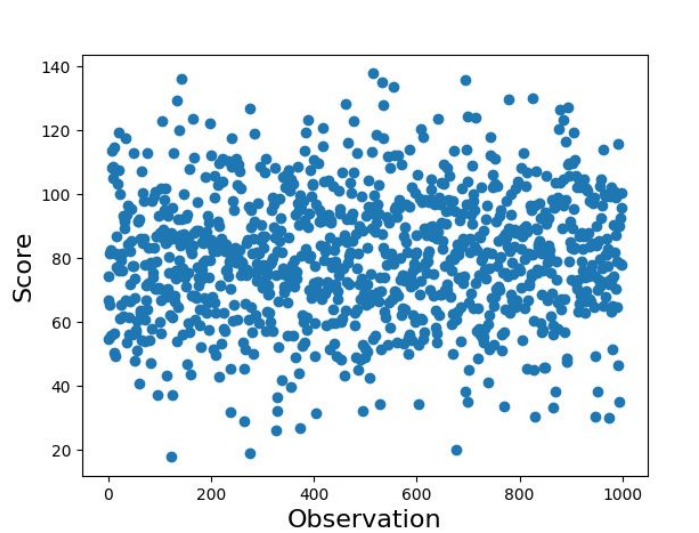
Some data analysis methods that can be useful:

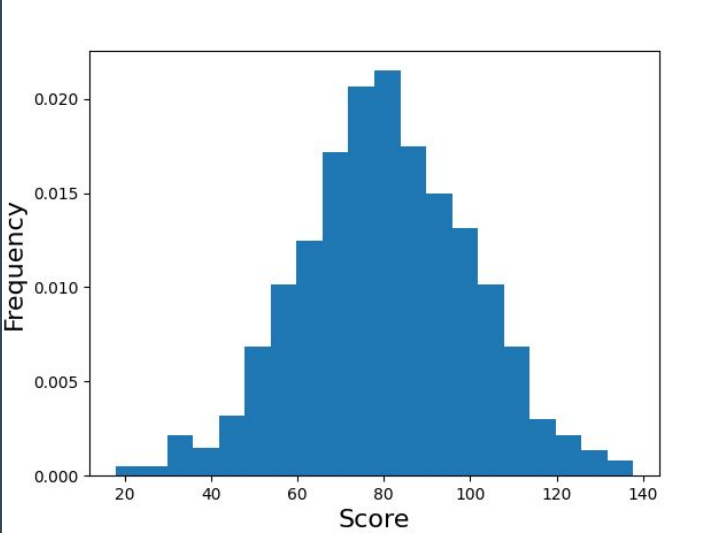
I. Approximation of statistical distributions
II. Comparison of multiple populations
III. Identify correlations

I. Approximate a distribution

Approximate a distribution

Example: Scores in a computational task (i.e. IQ score) from a group of people. We want to test the hypothesis that they follow a normal distribution.





Approximate a distribution

Example: First let's examine the properties of the normal distribution

1. Probability distribution function (PDF)

$$f(x) = P(X = x)$$

PDF of a real valued random variable X is the mathematical formulation of the probability that X will take the value x. This equation describes the probabilities of all possible outcomes to occur during a random experiment.

Approximate a distribution

Example: First let's examine the properties of the normal distribution

1. For a normal distribution the probability distribution function (PDF) for an observation x is given by:

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

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. Cumulative distribution function (CDF)

$$F(x) = P(X \leq x)$$

The CDF of a real-valued random variable X, evaluated at x, is the probability that X will take a value less than or equal to x.

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. CDF and PDF

$$f(x) = \frac{dF(x)}{dx}$$

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

The CDF of a continuous random variable X can be expressed as the integral of its probability density function f (x)

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. For a normal distribution the probability CDF for an observation x is given by:

$$\Phi(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. For a normal distribution the probability CDF for a parametrization of

$$t = \frac{x - \mu}{\sigma}$$

$$dt = \frac{dx}{\sigma}$$

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. For a normal distribution the probability CDF for a parametrization of

$$t = \frac{x - \mu}{\sigma}$$

$$dt = \frac{dx}{\sigma}$$

$$\Phi(x) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

Approximate a distribution

Example: First let's examine the properties of the normal distribution

2. For a normal distribution the probability CDF

$$\Phi(x) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

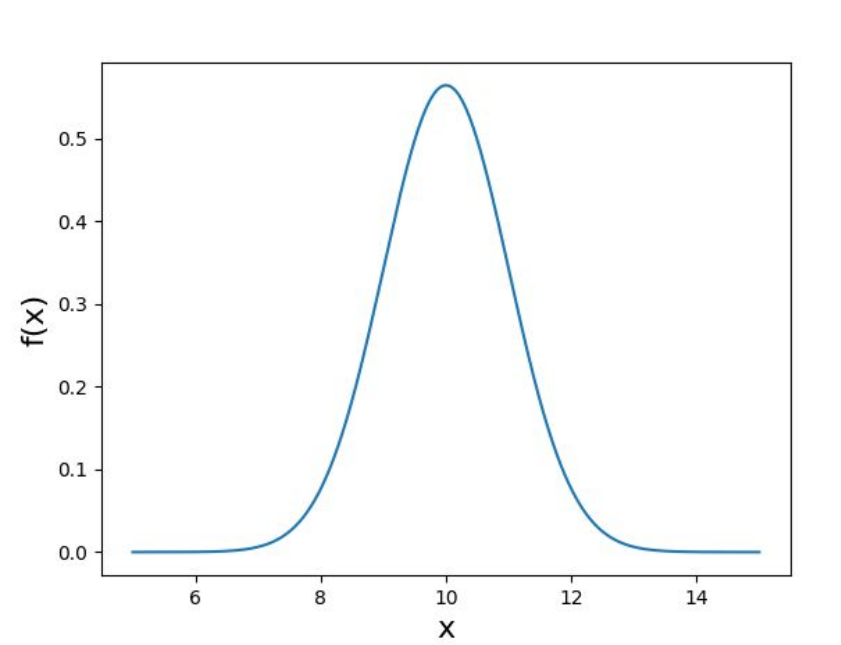
Approximate a distribution

Let's take an example of a normal distribution:

	Value
Mean	10
Standard deviation	1
Number of sample points	1000

Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$



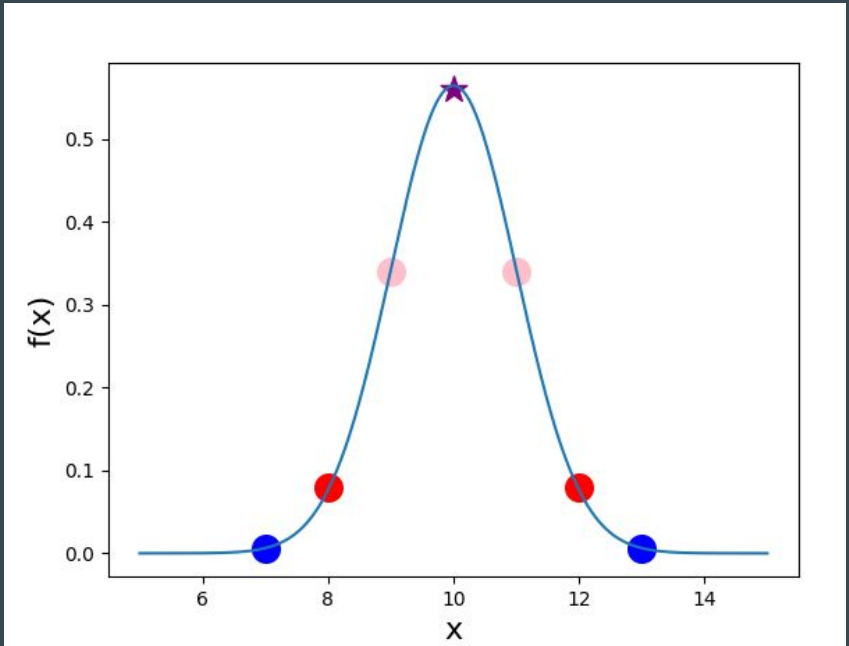
Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$

X	x	f(x): value
10	μ	0.56
9 / 11	σ	0.3
8 / 12	2σ	0.08
7 / 13	3σ	0.006

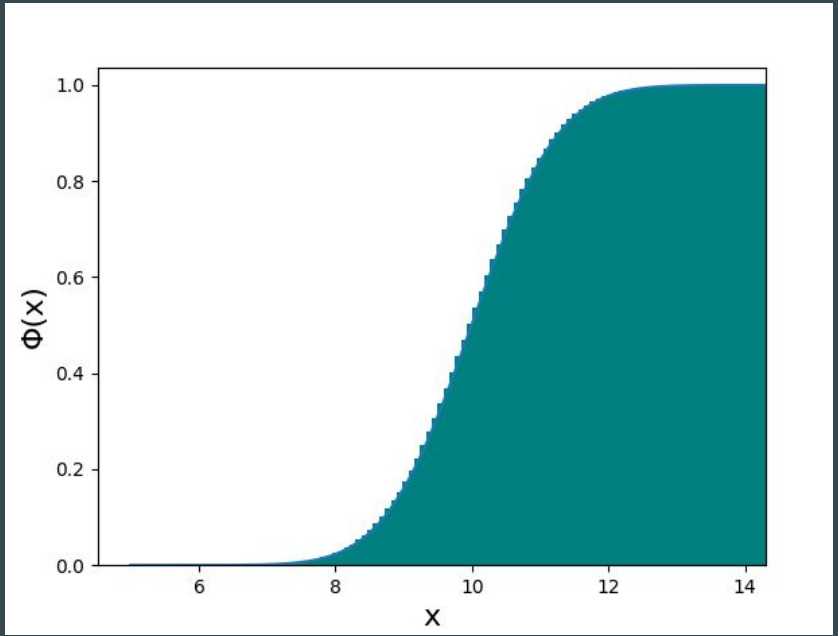
Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$



Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$



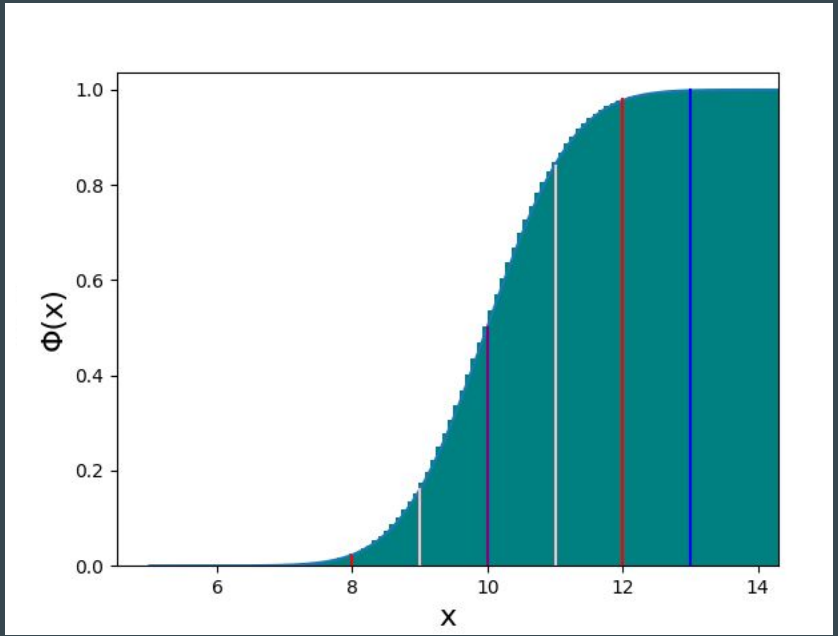
Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$

X	x	Φ(x): value
10	μ	0.50
9 / 11	σ	0.16 / 0.84
8 / 12	2σ	0.02 / 0.98
7 / 13	3σ	0.001 / 0.998

Approximate a distribution

For the normal distribution $N(\mu, \sigma^2) : \mu = 10, \sigma = 1$



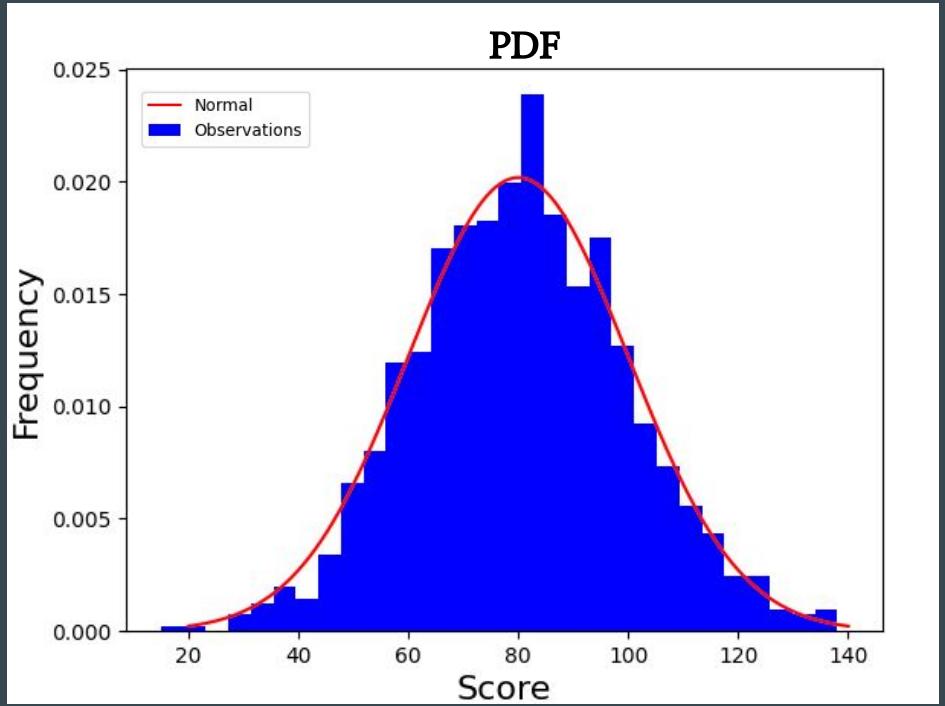
Approximate a distribution

So if we go back to the random experiment of scores, we can test our normal distribution hypothesis by the following process:

1. Identify the mean and standard deviation
2. Generate a normal distribution with these properties
3. Compute the probability of a score to appear (we need to bin the data)
4. Compute the cumulative probability of scores
5. Compare the PDF and CDF of the observations to the normal distribution

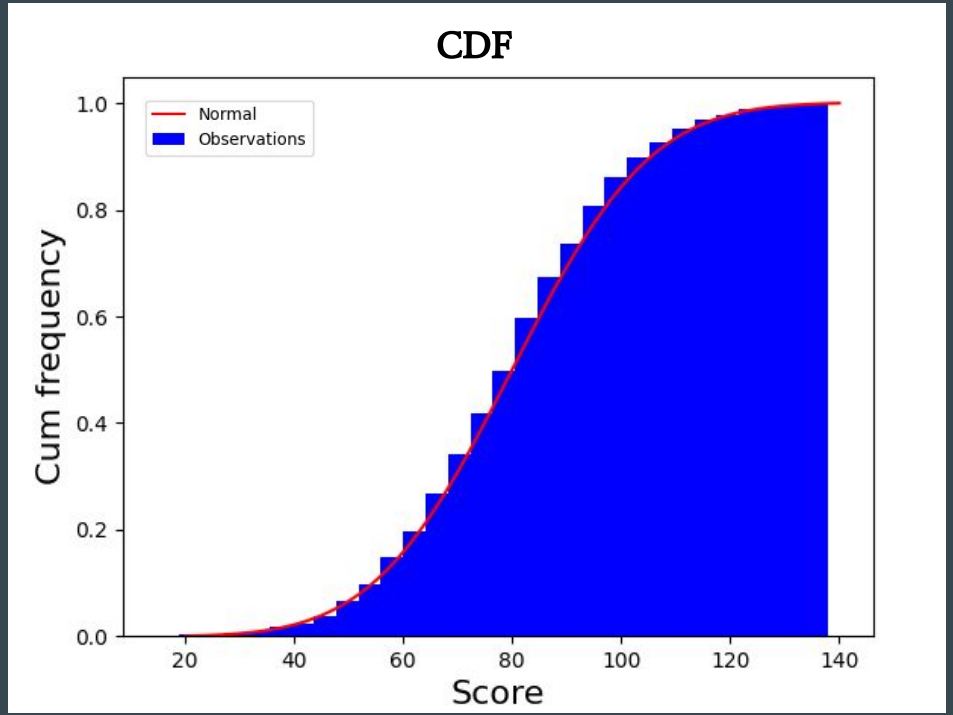
Approximate a distribution

The pdf of the normal distribution:



Approximate a distribution

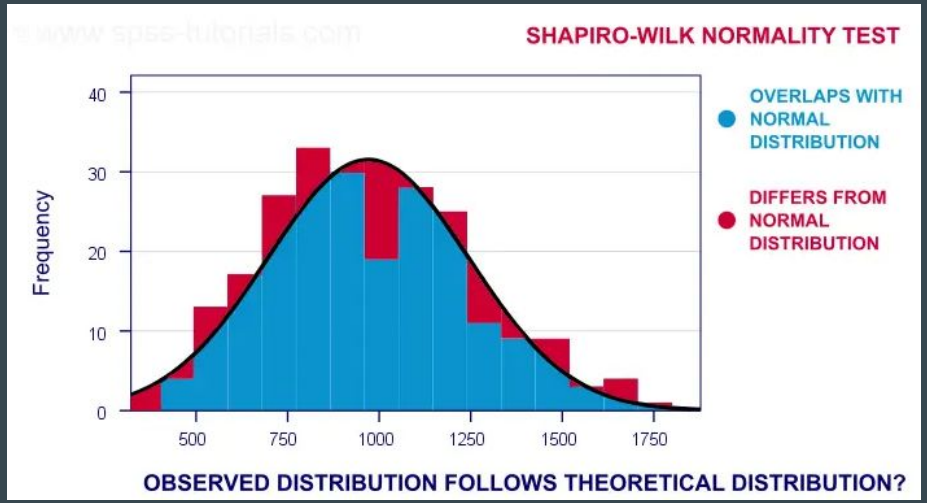
The cdf of the normal distribution:



Approximate a distribution

We can test normality of the dataset with the Shapiro - Wilk test

This test computes the difference between the PDF of the expected and the observed distributions

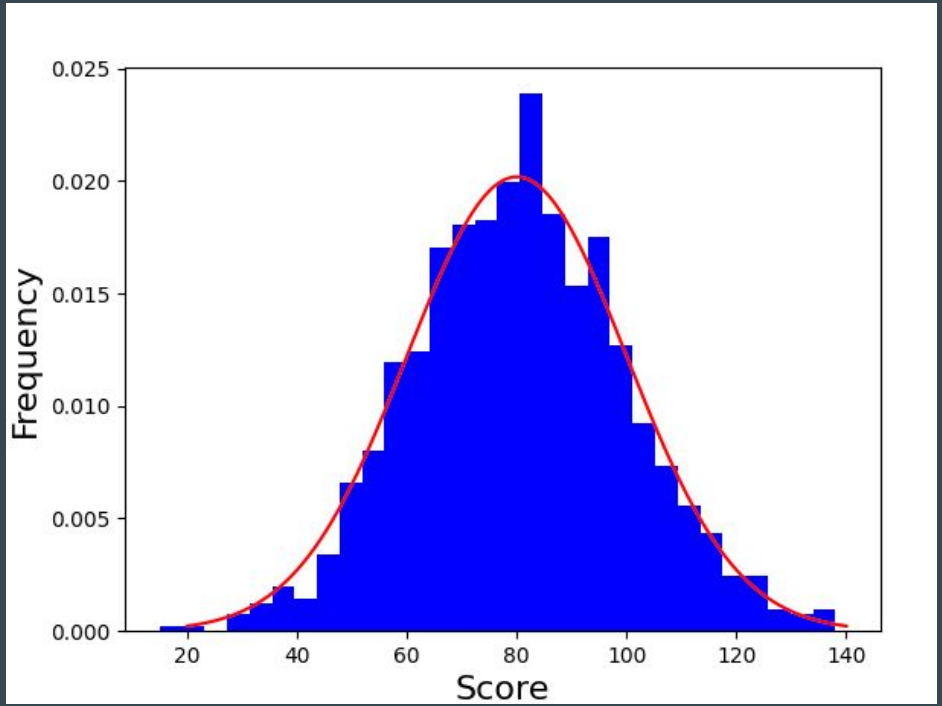


Source: https://www.spss-tutorials.com/spss-shapiro-wilk-test-for-normality/

Approximate a distribution

We can test normality of the dataset with the Shapiro - Wilk test

In our case, that corresponds to the plot we generated for PDF distributions



Approximate a distribution

We can test normality of the dataset with the Shapiro - Wilk test:

Shapiro test

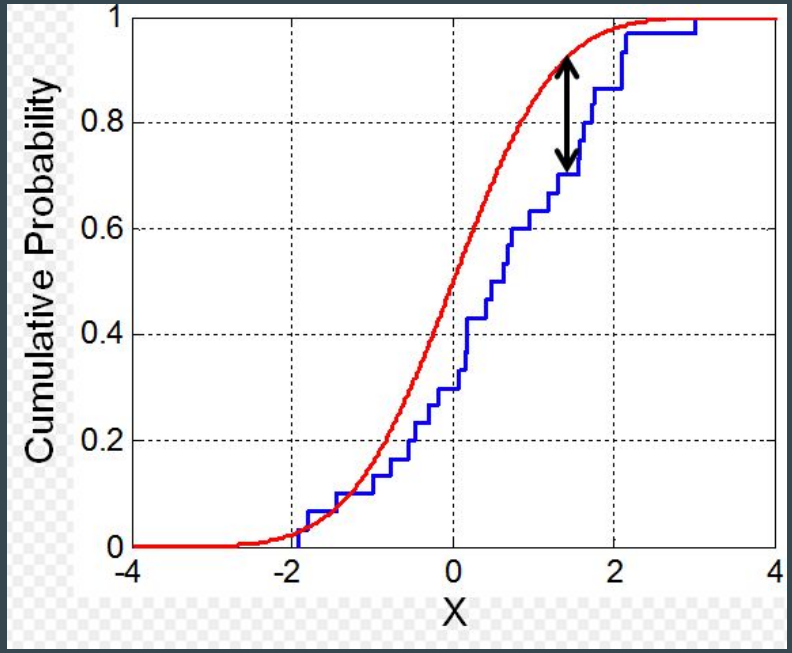
- statistic=0.99
- pvalue=0.63

If p > 0.05 -> supports the hypothesis that data are normally distributed. The data appear to follow a normal distribution.

Approximate a distribution

We can also test normality of the dataset with the Kolmogorov-Smirnov test

KS-test finds the maximum distance between the CDF of the expected and the observed distributions

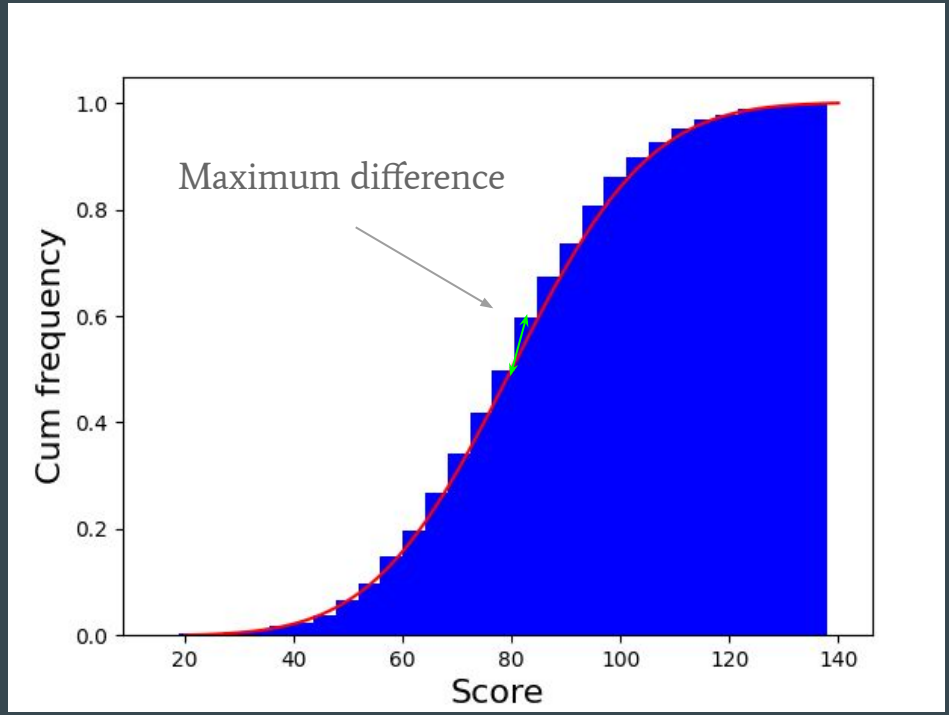


Source: wikipedia

Approximate a distribution

We can also test normality of the dataset with the Kolmogorov-Smirnov test

So in our case, this can be computed from the plot that we generated before



Approximate a distribution

We can also test normality of the dataset with the Kolmogorov-Smirnov test:

KS test not normalized

- statistic=0.99
- pvalue=0.0

KS test normalized

- statistic=0.0055
- pvalue=0.92

If p > 0.05 -> supports the hypothesis that data are normally distributed. The data appear to follow a normal distribution.

Approximate a distribution - Summary

1. Identify the properties of the distribution
2. Generate a normal distribution with these properties
3. Plot the data (PDF, CDF) and compare to normal distribution
4. Compute the statistical scores

Important points:

1. Sample size must be large enough for a statistical test to be accurate
2. Data need to be normalized for more accurate results

II. Comparison of multiple populations

Comparison of multiple populations

Example: Scores in a computational task (i.e. IQ score) from two groups of people. We want to test the hypothesis that two groups of people perform differently at this computational task.

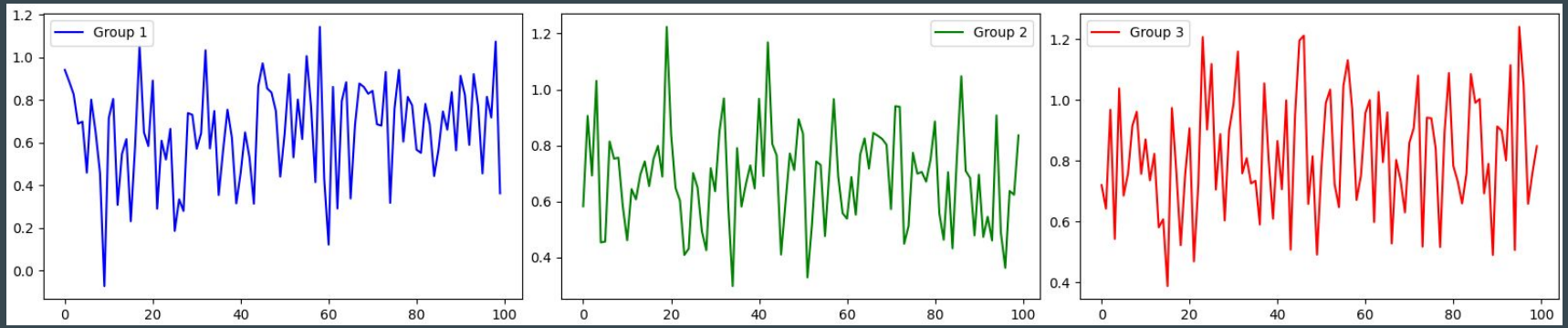
- Plotting the results
- Statistical tests
- Basic machine learning tools

Which one do you think would be more useful?

II. Comparison of multiple populations
(A) plot results

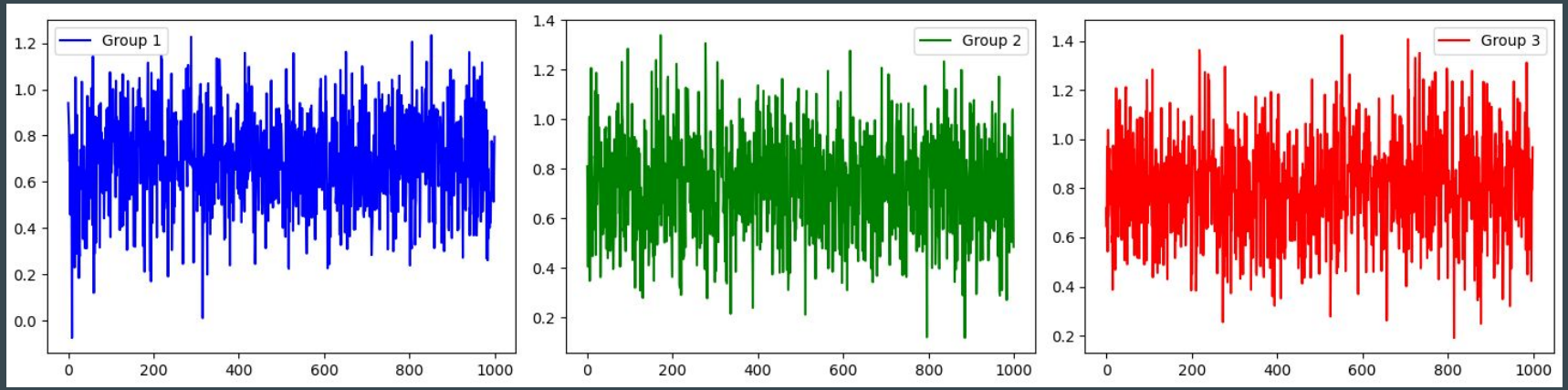
Comparison of multiple populations

Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. We want to test the hypothesis that two groups of people perform differently at this computational task. (sample size: 100)



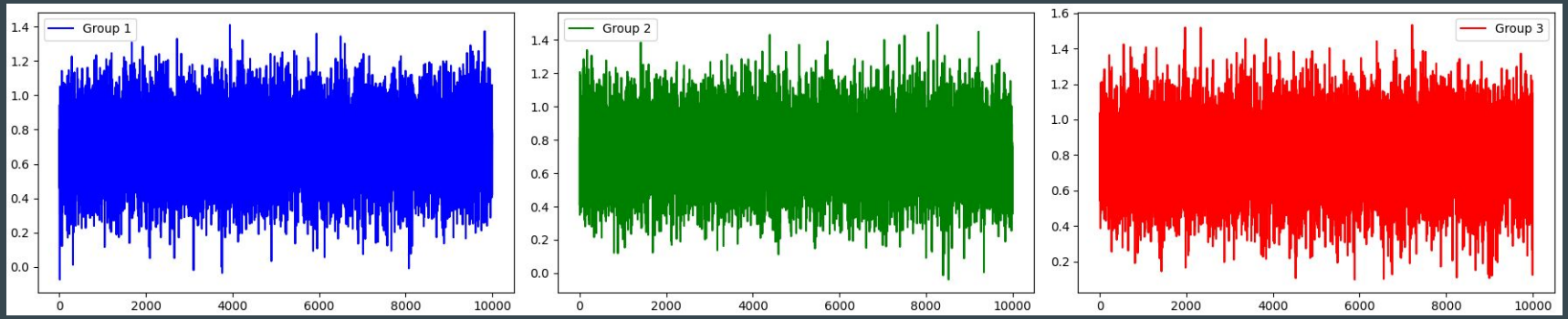
Comparison of multiple populations

Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. We want to test the hypothesis that two groups of people perform differently at this computational task. (sample size: 1000)



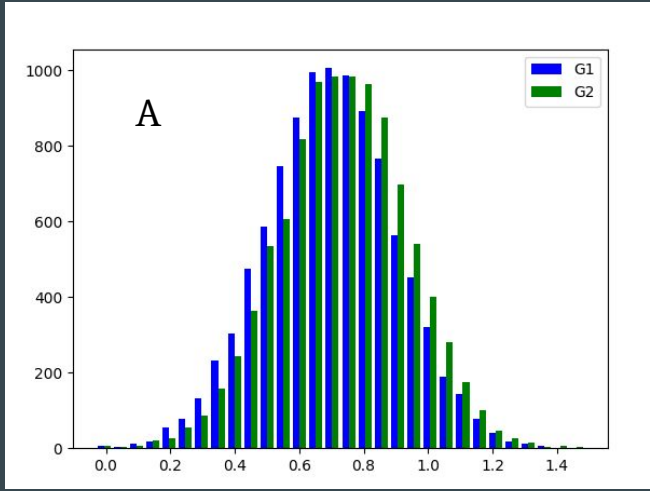
Comparison of multiple populations

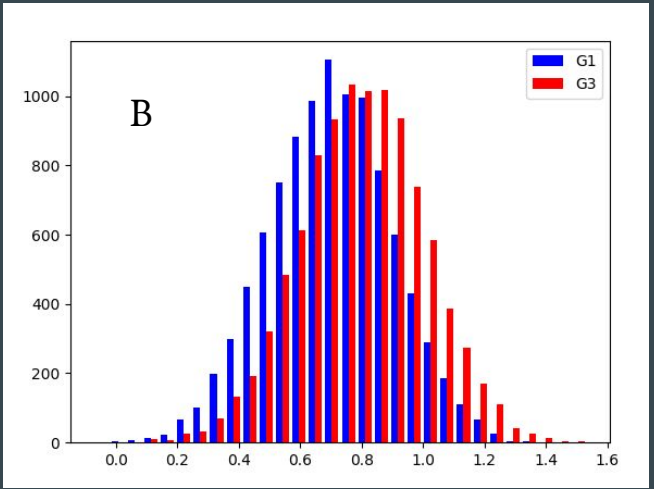
Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. We want to test the hypothesis that two groups of people perform differently at this computational task. (sample size: 10000)



Comparison of multiple populations

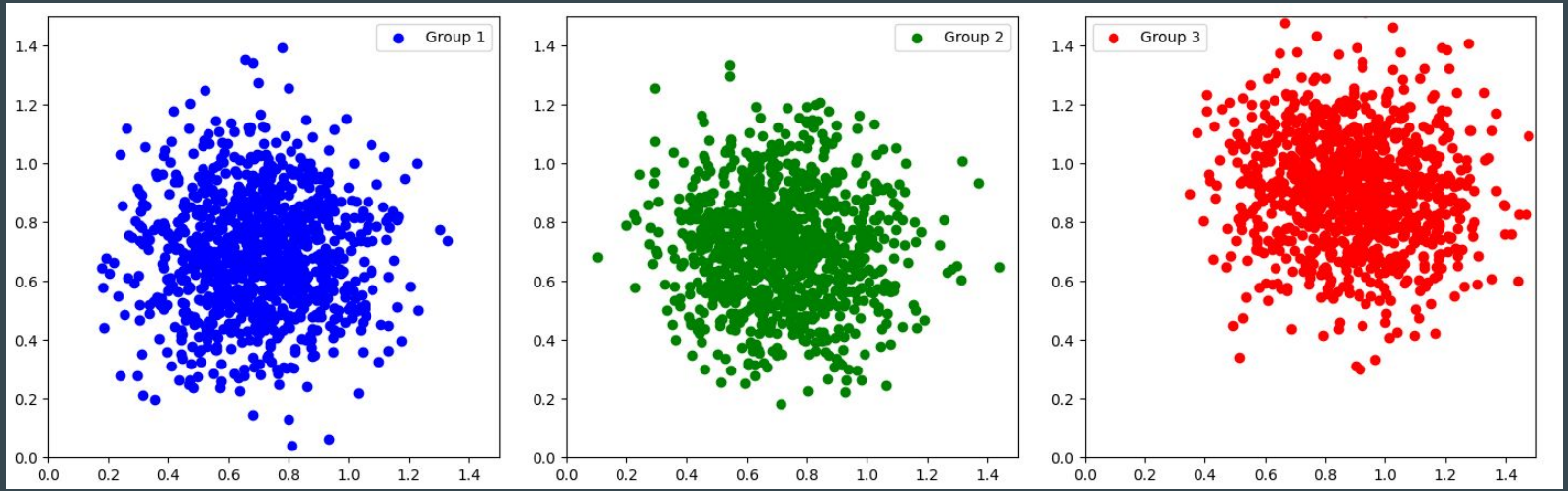
Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. We want to test the hypothesis that two groups of people perform differently at this computational task (G1-G2) and (G1-G3).





Comparison of multiple populations

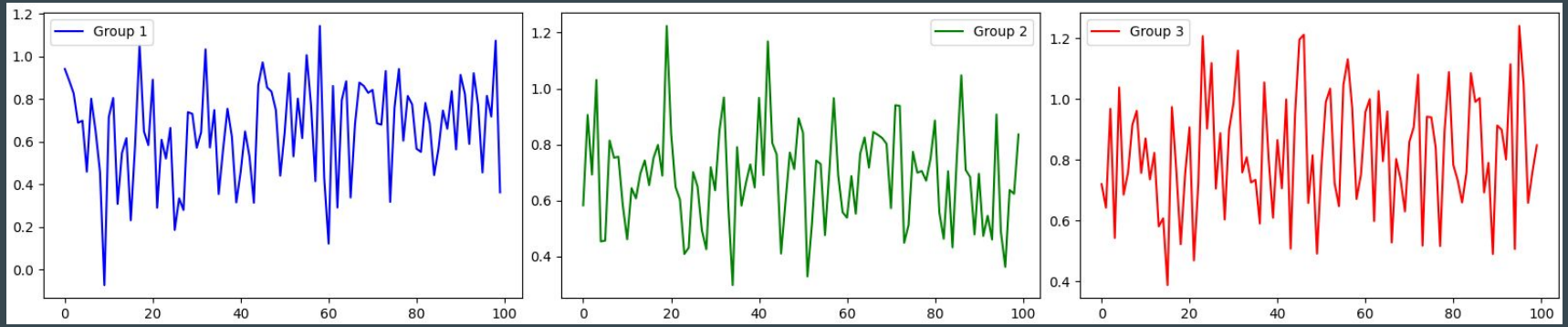
Example A2: Similarly to the previous task we can evaluate a set of 2d data: list of scores in two computational tasks (i.e. score I and score II) from the same two groups of people. We want to test the hypothesis that two groups of people perform differently at these computational tasks.



II. Comparison of multiple populations
(B) statistical analysis

Comparison of multiple populations

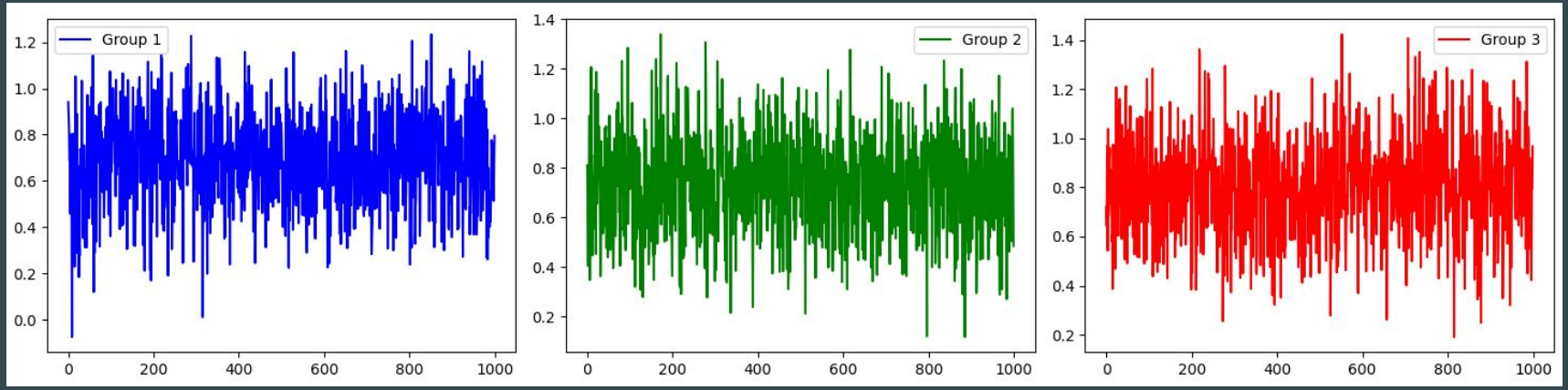
Example A



G1: Mean: 0.65 Std: 0.23 Sample size: 100
G2: Mean: 0.73 Std: 0.19 Sample size: 100
G3: Mean: 0.82 Std: 0.19 Sample size: 100

Comparison of multiple populations

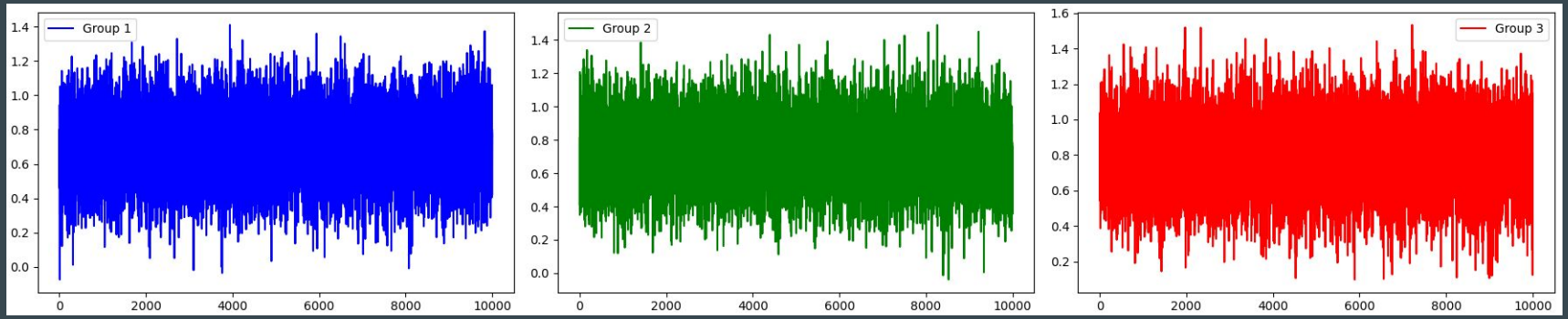
Example A



G1: Mean: 0.70 Std: 0.20 Sample size: 1000
G2: Mean: 0.73 Std: 0.19 Sample size: 1000
G3: Mean: 0.81 Std: 0.20 Sample size: 1000

Comparison of multiple populations

Example A:



G1: Mean: 0.70 Std: 0.20 Sample size: 10000
G2: Mean: 0.73 Std: 0.20 Sample size: 10000
G3: Mean: 0.80 Std: 0.20 Sample size: 10000

Comparison of multiple populations

Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. Let's use a statistical test, for example Kolmogorov-Smirnov test

A: KS test

- distance=0.07
- pvalue=1.1e-19

B: KS test

- distance=0.20
- pvalue=5e-173

In both cases the distributions show statistically significant differences!

Comparison of multiple populations

Example A: List of scores in a computational task (i.e. IQ score) from two groups of people. However... if we test a smaller sample size (#200)

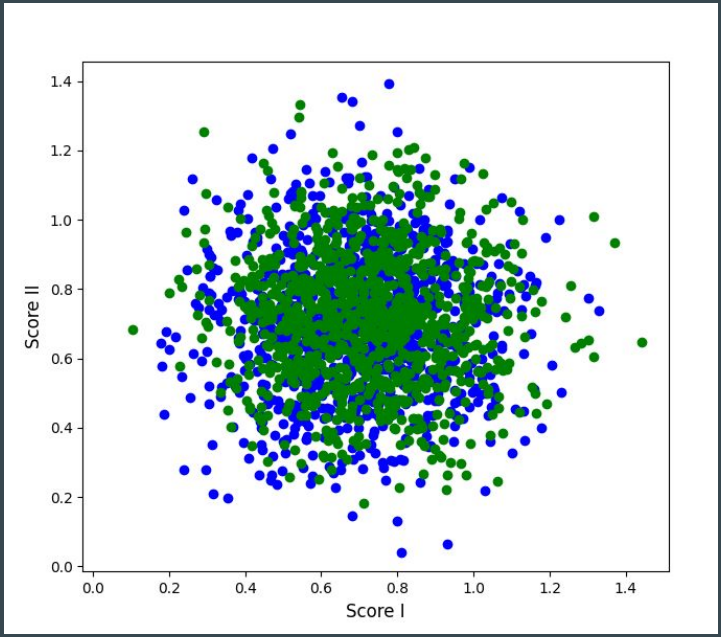
A: KS test
- distance=0.08
- pvalue=0.55

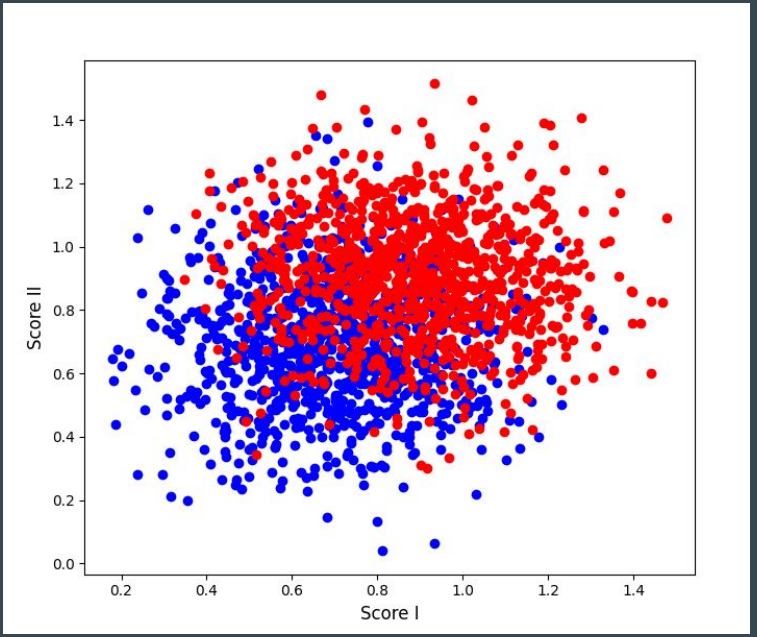
B: KS test
- distance=0.22
- pvalue=0.0001

The p-value is highly sensitive on sample size. So it's important to always report sample size on data! The distances between the two datasets are more consistent.

Comparison of multiple populations: plotting

Example A2: scores for two independent tasks for two groups of people.





Comparison of multiple populations: statistical analysis

Example A2: scores for two independent tasks for two groups of people.

A: KS

- distance=0.059
- pvalue=1.0e-15

B: KS

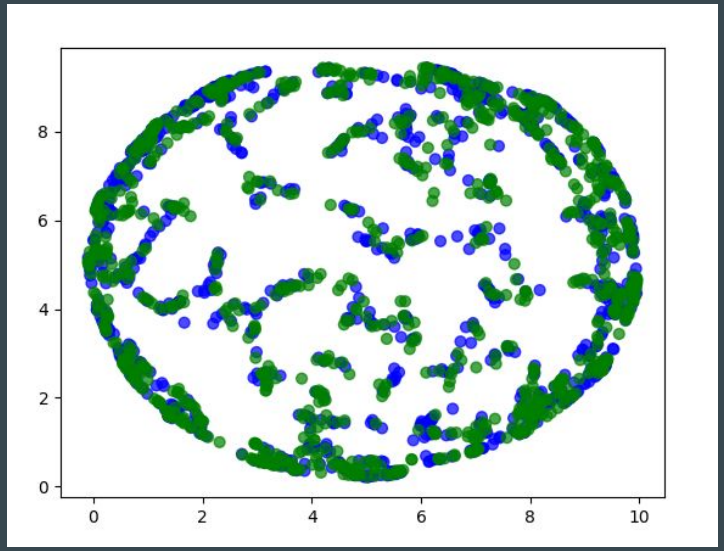
- distance=0.52,
- pvalue=0.0

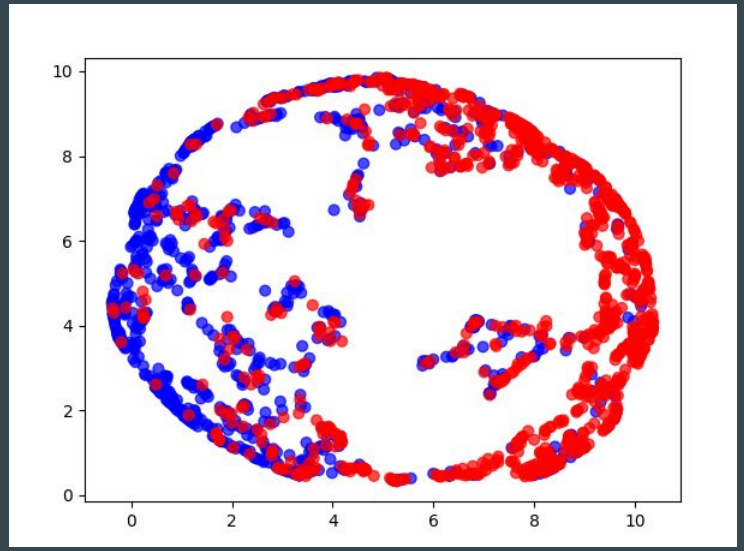
Statistical scores indicate a clear difference between G1 and G2. Even more evidently in G1 and G3

II. Comparison of multiple populations
(C) machine learning

Comparison of multiple populations: machine learning

Example A2: scores for two independent tasks for two groups of people. Plotting data using umap





Comparison of multiple populations: machine learning

Example A2: scores for two independent tasks for two groups of people.

We can train basic classifiers (for example DecisionTrees for supervised, Kmeans for unsupervised learning) to see if they can recognize the individual observations from different groups.

We define accuracy as the number of correct predictions divided by the sample size.

$$acc = \frac{1}{N} \sum_{0}^{N} 1(y_{pred} = y_{real})$$

Comparison of multiple populations: machine learning

Example A2: scores for two independent tasks for two groups of people.

DecisionTrees:

We train the classifier on 50% of the data and test on the remaining 50%. The accuracy of the classifier is:

A: 0.501

B: 0.675

Comparison of multiple populations: machine learning

Similarly, let's collect the scores for two independent tasks for two groups of people.

Kmeans:

We cluster the dataset in two groups. The accuracy of kmeans clustering is:

A: 0.502

B: 0.638

Comparison of multiple populations: machine learning

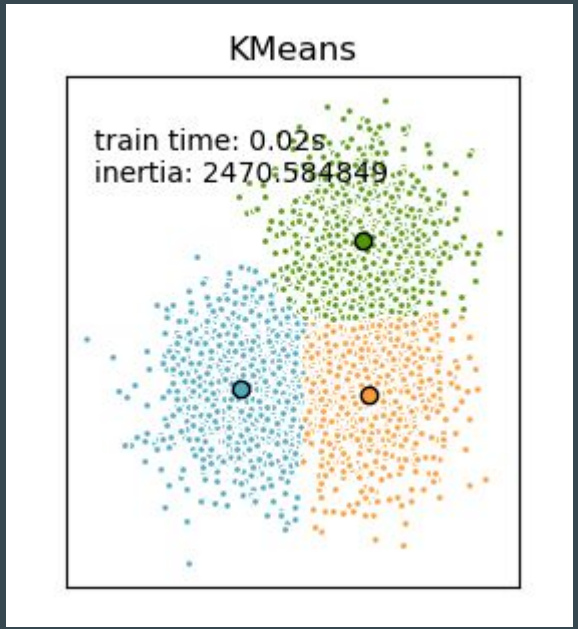
Using machine learning indicates that the first two groups are similar while the first and third groups differ significantly. This is closer to our intuition but mathematically this result is less accurate since the data originally came from different distributions.

Why do we get a less accurate result with ML, i.e. with a more advanced technique?

Comparison of multiple populations: machine learning

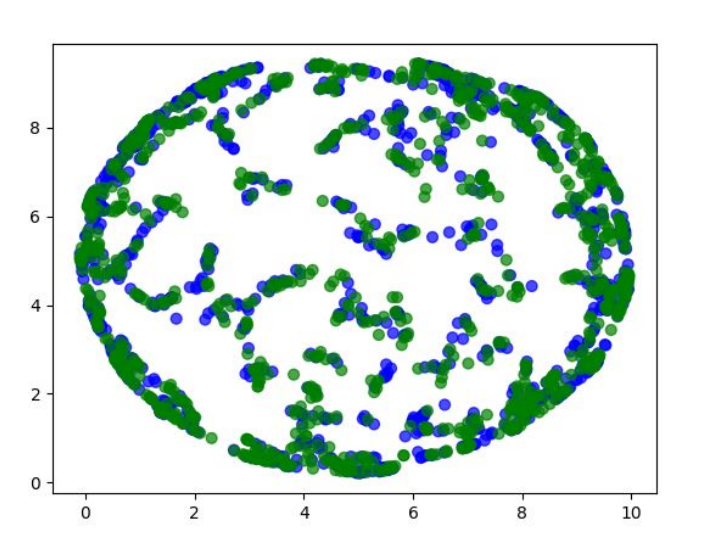
Why do we get a less accurate result with ML?

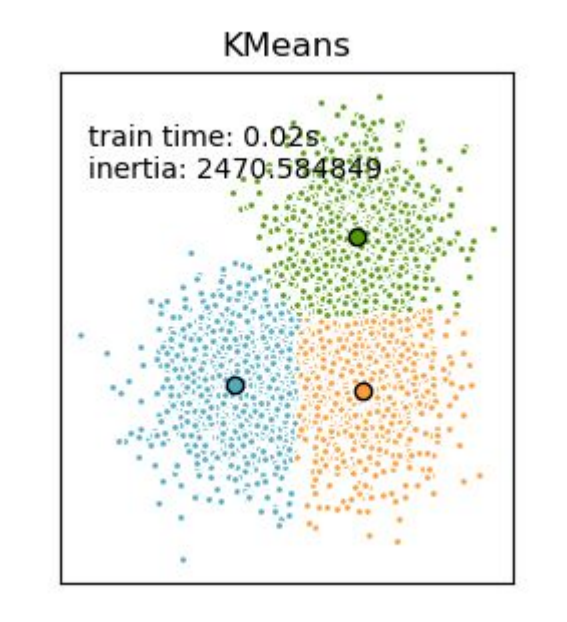
To understand this, we need to delve deeper into the algorithm of kmeans, which depends on the spatial separation of data



Comparison of multiple populations: machine learning

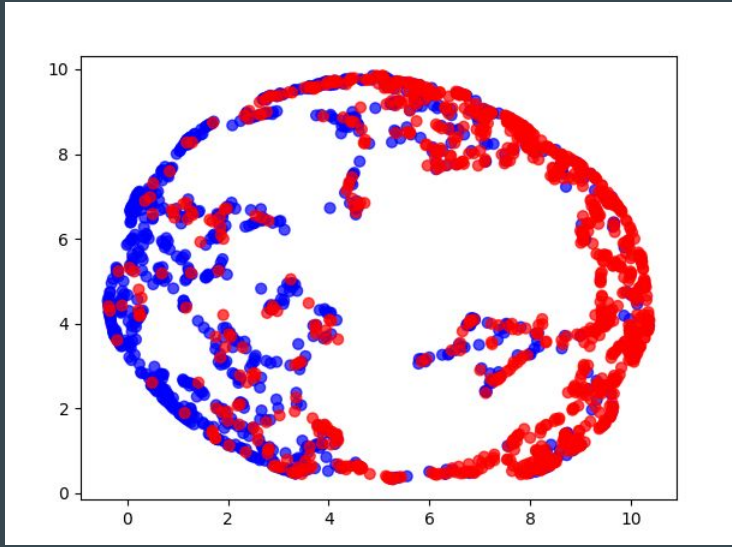
Why do we get a less accurate result with ML? The problem is not the technique but the application to the specific problem:

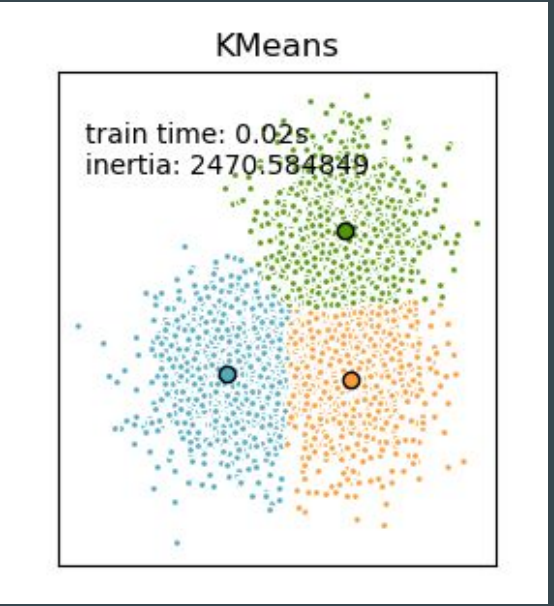




Comparison of multiple populations: machine learning

ML typically separates a dataset of collected observables into groups. However, in our use-case we have only two data points for each individual (score I, score II) and we have seen before that each observable is normally distributed.





Comparison of multiple populations: machine learning

The problem here arises from our attempt to approach a statistical problem with machine learning. How can we evaluate the nature of a problem:

Number of observables: Typically more dimensions make statistical analysis harder and machine learning can be more effective, while smaller dimensions allow the use of statistics.

Sample size: Small sample size is not suitable for machine learning techniques.

Number of groups: Statistical scores can be used for pairs of groups, so problems that need separation into i.e. 10 populations cannot be studied statistically

II. Comparison of multiple populations
Conclusions

Comparison of multiple populations

1. Plotting can be deceiving
2. Statistical tests are more reliable, however they depend on the sample size
3. More advanced techniques (Machine learning) are not always appropriate

III. Correlations in data

Find correlations in measurements

We have a dataset of measurements collected from a set of neurons. For example we measure anatomical properties for a list of neuronal morphologies:

	A	B	C	D	E
1	property|name	axon|mean_number_of_bifurcatio	axon|mean_number_of_leav	axon|mean_number_of_sectio	axon|mean_remote_bifurcation_angl
2	C050800E2_cor.h5	59	60	119	1.2029861559104
3	C120398A-P2.h5	25	26	51	1.31538621816681
4	C120398A-P3.h5	40	41	81	1.42040152246122
5	C271097A-P1.h5	37	38	75	1.17331024837382
6	C271097A-P2.h5	21	22	43	1.18553678311865
7	C271097A-P3.h5	34	35	69	1.30503361454563
8	rp110111_L5-2_idH.h5	61	62	123	1.39761281141949
9	rp120608_P_3_idC.h5	44	47	92	1.18236947545108
10	rp120608_P_3_idD.h5	69	70	139	1.08486835209442
11	sm100429a1-5_INT_idE.h	38	39	77	1.4799778645049
12	sm100429a1-5_INT_idG.h	28	29	57	1.73913560953665
13	vd101102b_INT_idA.h5	43	44	87	1.45461104222372
14	C060998B-P4.h5	10	11	21	1.47125481505973
15	C120398A-P1.h5	10	11	21	1.6064295953642
16	C140600C-P3.h5	12	13	25	1.11694986036925
17	C150897B-P2.h5	32	33	65	1.21858139632941
18	C200897C-P2.h5	19	20	39	1.70497037908832
19	C200897C-P4.h5	41	42	83	1.37332234508014
20	C220498B-P3_cor.h5	25	26	51	1.51601242237315
21	C231296A-P4B2.h5	16	17	33	1.5199786142752
22	C260199A-P2.h5	4	5	9	1.47696729610387
23	C300797C-P2.h5	17	18	35	1.52795803970122
24	C300797C-P4.h5	10	11	21	1.5769699580797
25	C310897A-P4.h5	69	70	139	1.24321043670234
26	C310897B-P3.h5	29	30	59	1.31227750969831
27	C310897B-P4.h5	8	9	17	1.25173531651878
28	Fluo9_left.h5	28	29	57	1.30945300169143

Find correlations in measurements

We have a dataset of measurements collected from a set of neurons.

Two variables can be:

- Positively Correlated: variables change in the same direction.
- Negatively Correlated: variables change in opposite directions.
- Not Correlated: no relationship in the change of the variables.

Find correlations in measurements

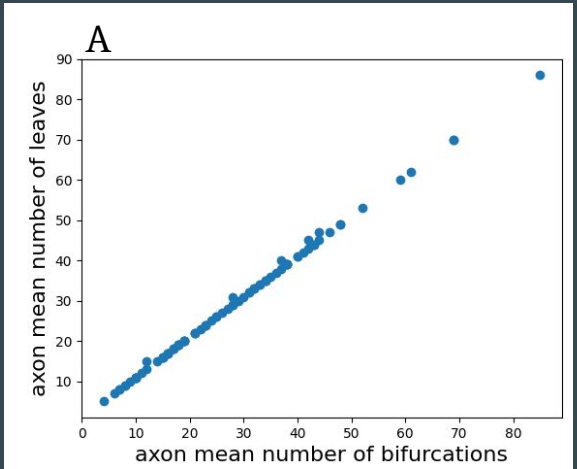
We have a dataset of measurements collected from a set of neurons.

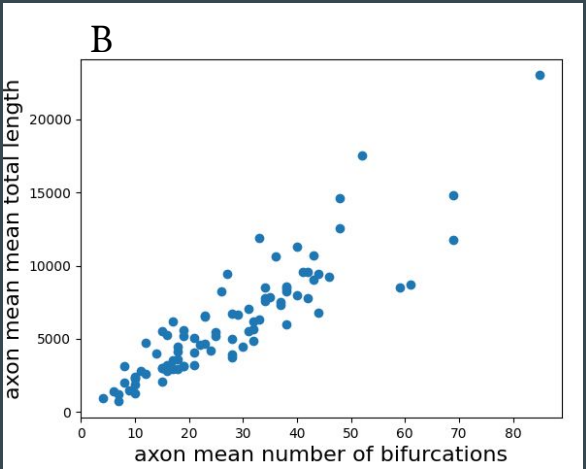
Examples of possible correlations include:

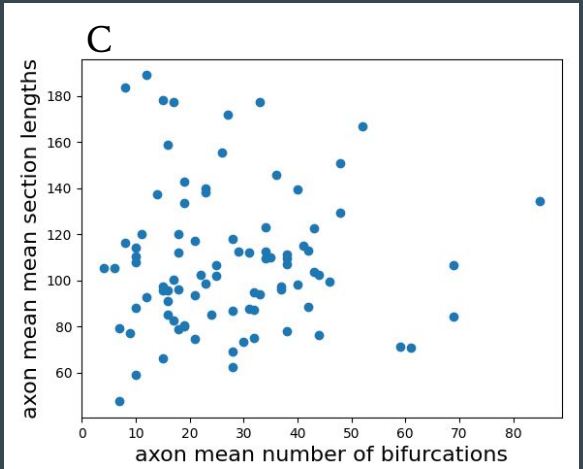
- One variable could be strongly correlated on the values of another variable: one variable is a function of another.
- One variable could be slightly associated with another variable.
- Two variables could depend on a third unknown variable.

Find correlations in measurements

We have a dataset of measurements collected from a set of neurons. How can we check if two measurements are correlated or independent?







Find correlations in measurements

Pearson correlation test between datasets X and Y is computing the covariance divided by the standard deviation of the two datasets.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The P-correlation can be computed as follows from the two datasets:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

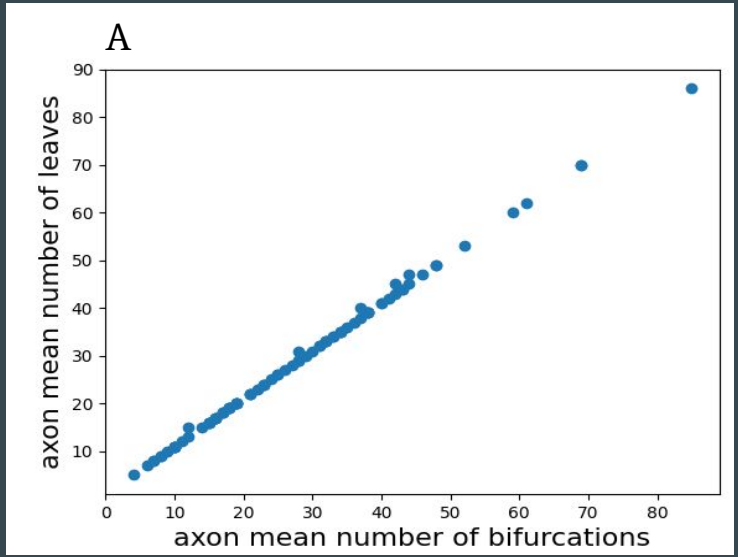
Find correlations in measurements

Pearson correlation test

stats.pearsonr(X_A, Y_A)

Pc: 0.999

pvalue: 6.6e-135



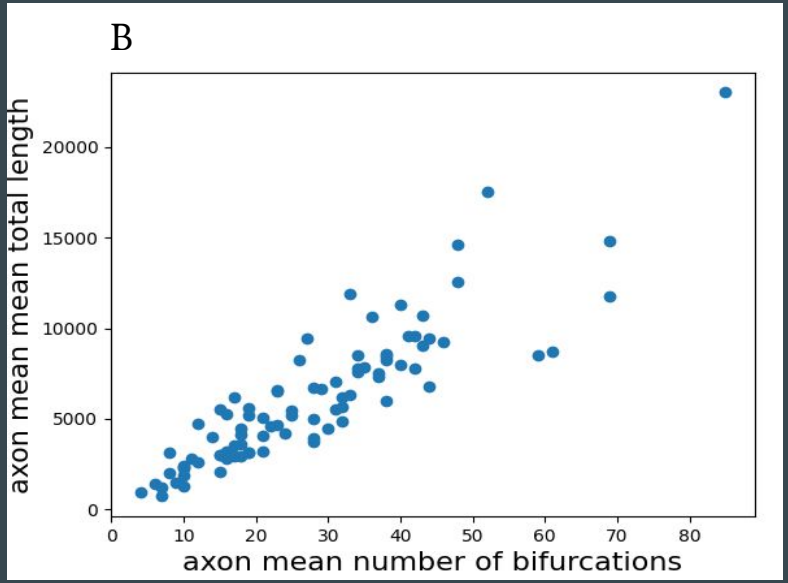
Find correlations in measurements

Pearson correlation test

stats.pearsonr(X_B, Y_B)

Pc: 0.887

pvalue: 5.5e-31



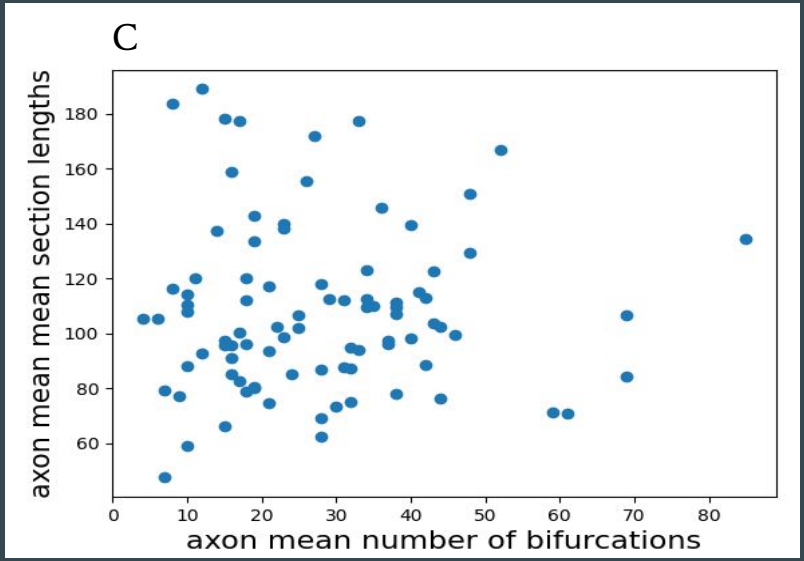
Find correlations in measurements

Pearson correlation test

stats.pearsonr(X_C, Y_C)

Pc: 0.0079

pvalue: 0.94



Find correlations in measurements

Chi square is used to test whether there is a relationship between two categorical variables. Practically it can be used to test whether or not a number of outcomes is occurring in the expected frequency:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O: observed frequency
E: expected frequency

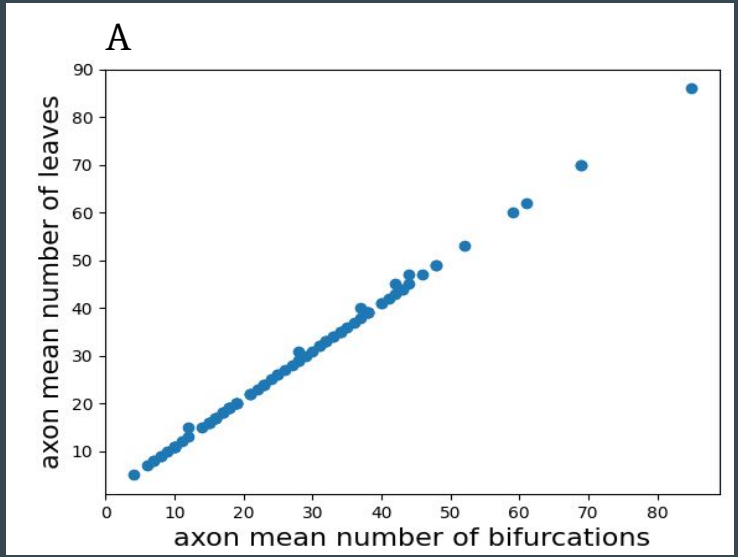
Find correlations in measurements

Chi square test

stats.chi2_contingency(X_A, Y_A)

x2: 0.0

pvalue: 1.0



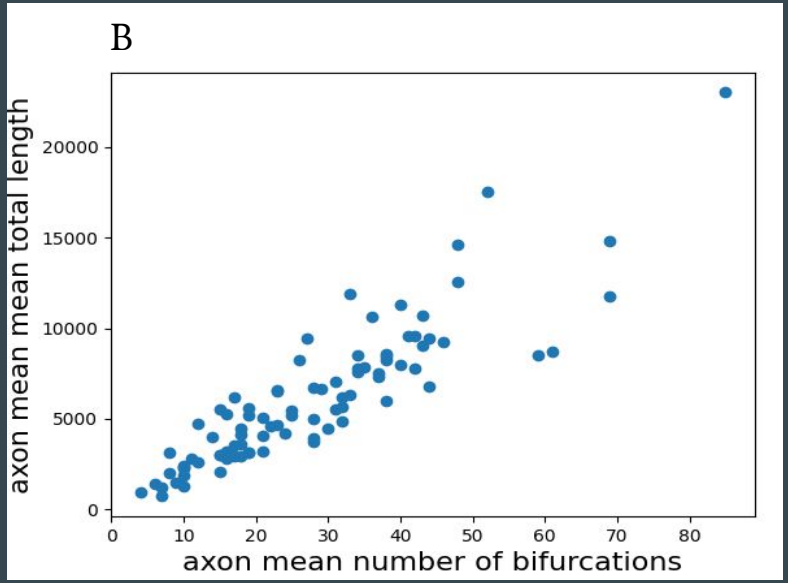
Find correlations in measurements

Chi square test

stats.chi2_contingency(X_B, Y_B)

x2: 0.0

pvalue: 1.0



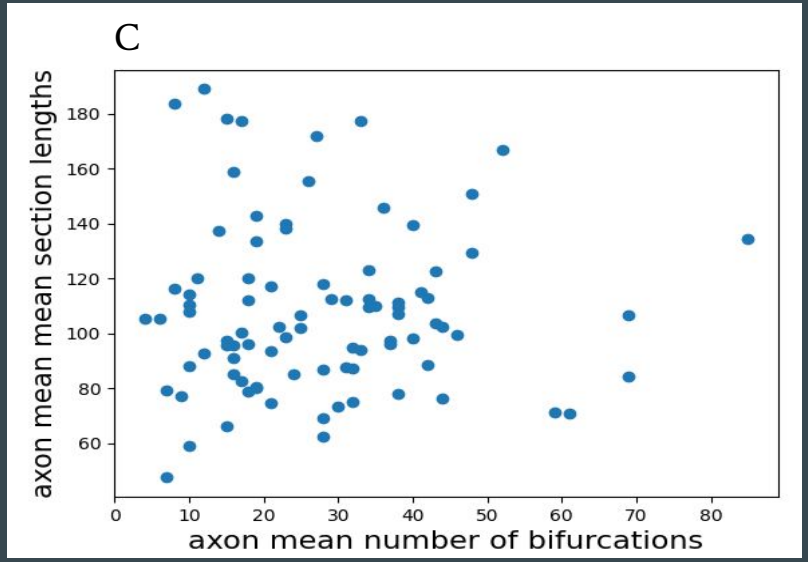
Find correlations in measurements

Chi square test

stats.chi2_contingency(X_C, Y_C)

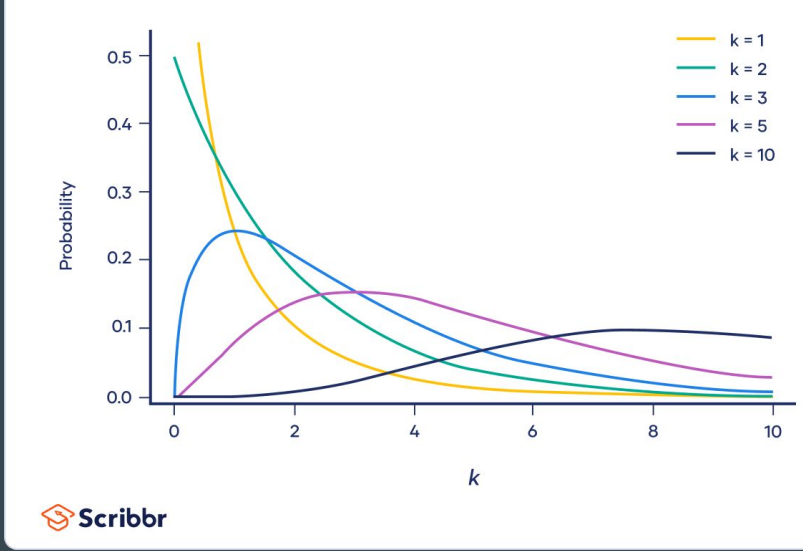
x2: 0.0

pvalue: 1.0



Find correlations in measurements

Chi square is assuming that the distribution of observations is following the chi2 distribution. If this is not the case, it is not an appropriate test to use.



Find correlations in measurements

T-test is used to test the difference between two groups on some continuous variable. For example for two datasets X, Y the t-test can detect if they are correlated or not:

$$t - test = \frac{\overline{X} - \overline{Y}}{SS\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

X, Y: is the mean of each dataset

n_X, n_Y: sample size

SS: sum of squares

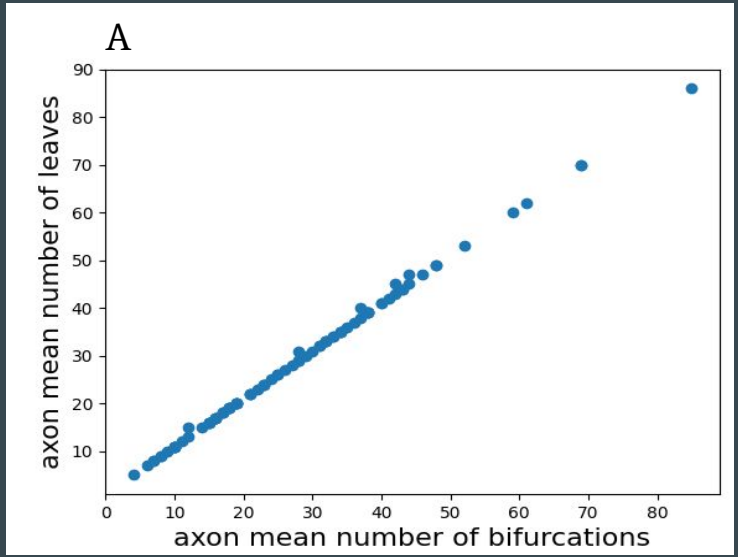
Find correlations in measurements

T-test test

stats.ttest_ind(X_A, Y_A)

ttest: -0.47

pvalue: 0.635



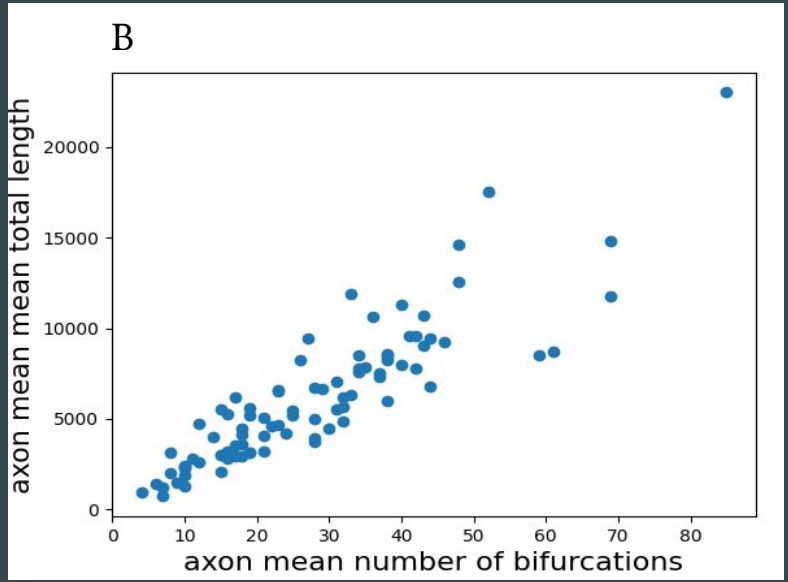
Find correlations in measurements

T-test test

stats.ttest_ind(X_B, Y_B)

ttest: -14.97

pvalue: 3.2e-33



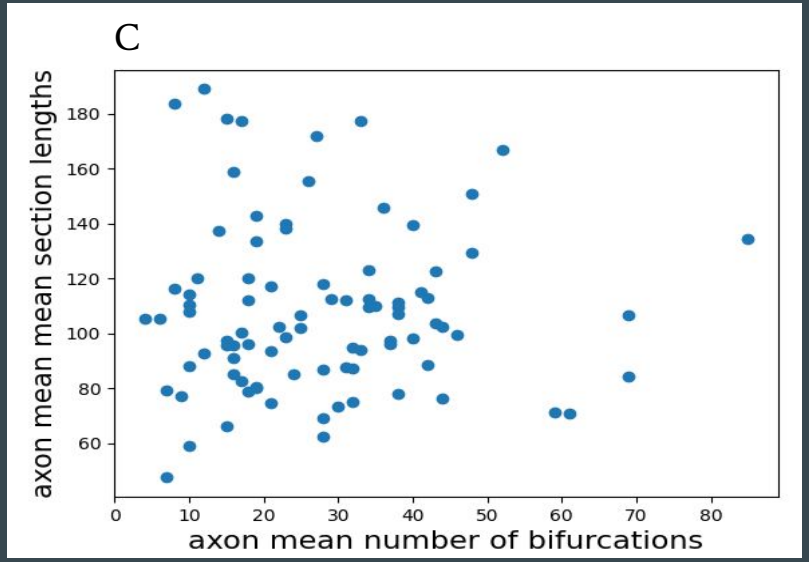
Find correlations in measurements

T-test test

stats.ttest_ind(X_C, Y_C)

ttest: -22.2

pvalue: 4.4e-53



Find correlations in measurements

T-test doesn't seem to work well as an independence test. However... an important point is that we need to normalize the datasets. If we divide each measurement with the maximum element we transform the data:

(0, max) -> (0, 1)

This normalization is important when values of the two variables are in different units and therefore they are not directly comparable.

What would be a better way to compare them using t-test?

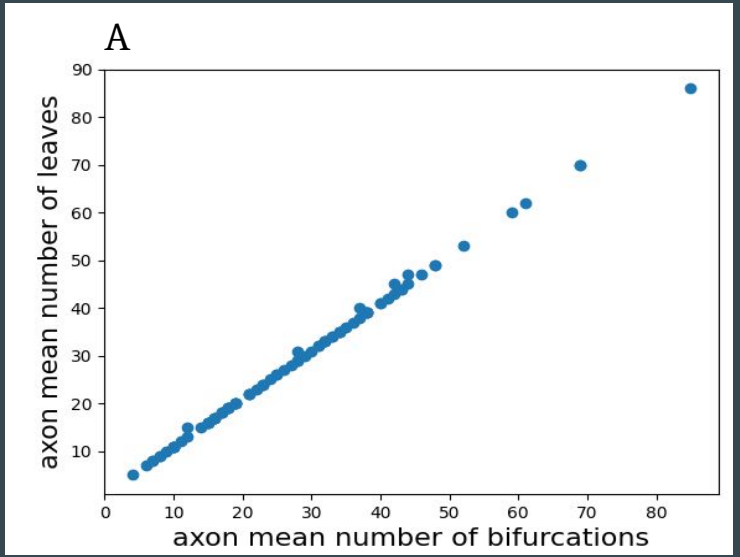
Find correlations in measurements

T-test test

stats.ttest_ind(X_A/max(X_A), Y_A/max(Y_A))

ttest: -0.33

pvalue: 0.739 (was 0.635)



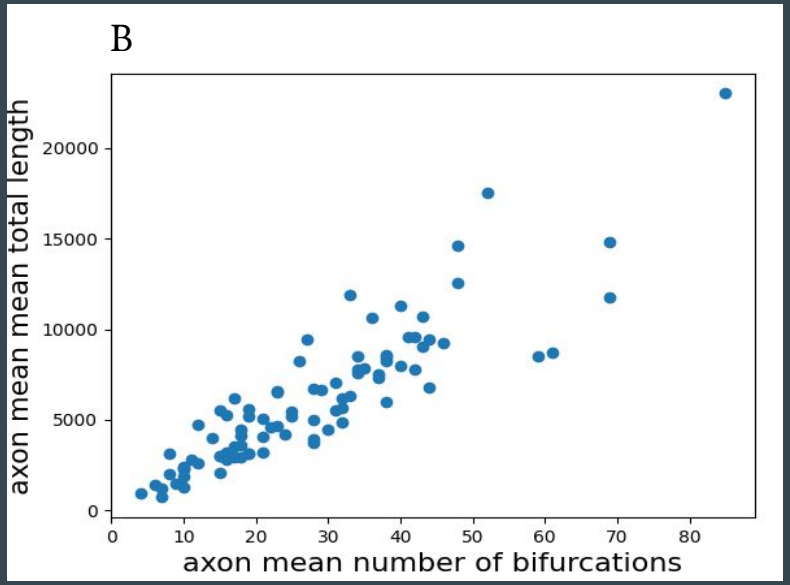
Find correlations in measurements

T-test test

stats.ttest_ind(X_B/max(X_B), Y_B/max(Y_B))

ttest: 2.28

pvalue: 0.024 (was 3.2e-33)



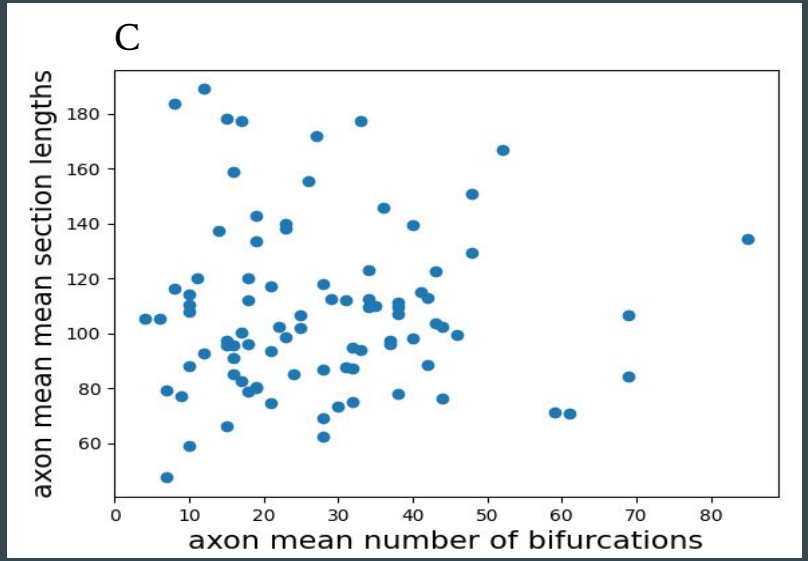
Find correlations in measurements

T-test test

stats.ttest_ind(X_C/max(X_C), Y_C/max(Y_C))

ttest: -9.53

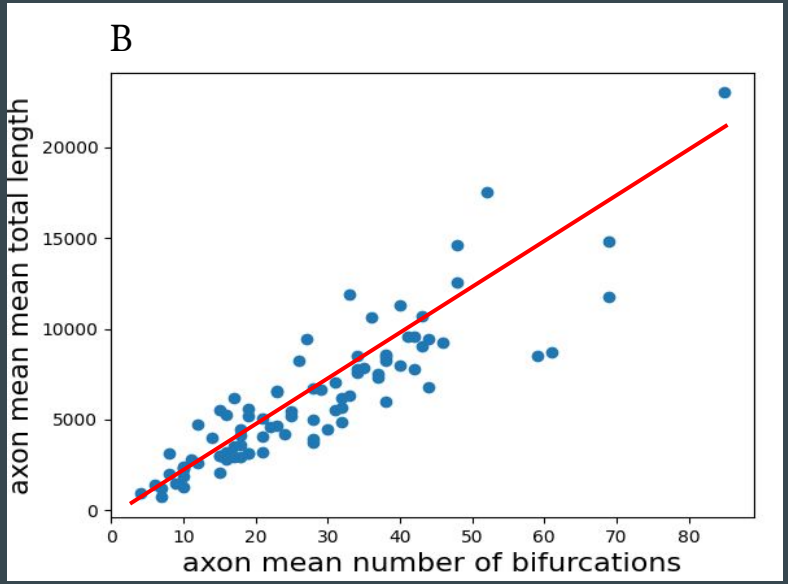
pvalue: 1.2e-17 (was 4.4e-53)



Find correlations in measurements

Once a correlation is identified, the precise relation between the two variables can be explored. One possible method to explore this relation is regression (we will talk more about this in machine learning)

For example linear regression can be used to finds the line that minimizes the distance of points.



Find correlations in measurements

In summary to find correlations between two datasets we can use a variety of methods:

- Pearson, (similar methods: Spearman, Kendall)
- Chi2
- T-test
- Anova

However, we need to be careful about:

- The sample size
- The distribution of the data
- Normalization

Data Detective Methods for Revealing Questionable Research Practices

Gregory Francis and Evelina Thunell

Data Detective Methods for Revealing Questionable Research Practices

There are many types of Questionable Research Practices (QRPs) that all tend to generate statistical information that misrepresents reality. This chapter discusses some methods for detecting the presence of QRPs, mostly by looking for conflicts in different sources of information. These methods typically cannot identify precisely which QRPs were used, and sometimes the conflicts are due to typos or simple mistakes, but either way readers should be skeptical about the validity of studies with inconsistent statistical information. An appropriate mindset for identifying inconsistencies is that of a “data detective” who looks for patterns that do not make sense. We start by describing mathematical inconsistencies between sample sizes and the degrees of freedom in hypothesis tests, which are easy to detect and indicate either a QRP, unreported outlier removal, or sloppiness in reporting.

Data Detective Methods for Revealing Questionable Research Practices

Two additional tests explore inconsistencies across experiments. First, the Test for Excess Success compares the frequency of reported successful outcomes to the expected frequency if the tests were run properly, fully reported, and analyzed without QRPs. Too much success indicates a problem with the reported results. Second, the p-curve analysis examines the distribution of reported p-values for properties that indicate invalid data sets (that are perhaps the result of QRPs).

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

A simple test to start from entails the comparison between the degrees of freedom and the reported p-values.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

Degrees of freedom (df): indicate the number of independent values that can vary in an analysis without breaking any constraints. For a sample size N and P parameters:

$df = N - P$

For example, given a set of observations: x1, x2, ..., x10 and a reported mean (m=0.65) and sample size (N=10), if we know the values of x1, ..., x9 we can compute x10.

In this case $df = 10 - 1 = 9$

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

Degrees of freedom (df): indicate the number of independent values that can vary in an analysis without breaking any constraints.

For t-test (which tests statistical significance of the *mean*)

df = n − 1, where n is the sample size

df = n1 + n2 − 2, where n1 and n2 are the sizes of the two samples

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

Usually scientific papers report number of df and sample size. For example

"As predicted, with n1 = 35 and n2 = 27, we found a significant difference between the control and experimental means t(58) = 2.1, p = 0.04."

For a t-test of n1 = 35, n2 = 27, df = n1 + n2 - 2 = 60 which differs from the reported df=58.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

Degrees of freedom (df): indicate the number of independent values that can vary in an analysis without breaking any constraints.

For one-way ANOVA F-test we have two df terms

$df_{numerator} = K - 1$

$df_{denominator} = N - K.$

Here, K is the number of conditions and N is the sum of sample sizes across all conditions.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK

When testing whether a drug is effective at reducing the duration of a cold, the null hypothesis H0 might look like:

H0: m1 = m2

where m1 and m2 denote the duration of the cold with and without the drug, respectively.

The goal of the hypothesis test is to decide whether to reject the null hypothesis. This decision is based on “statistical significance,” which is determined by a test statistic that is derived from the experimental data.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK:

the statistical test t is:

$$t - test = \frac{\overline{X} - \overline{Y}}{SS\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

where X and Y are the sample means and SS is the standard deviation of the sampling distribution of the difference of means, which is a function of the standard deviation s, and the sample sizes n_X and n_Y.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK:

$$t - test = \frac{\overline{X} - \overline{Y}}{SS\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

If the null hypothesis is true, the t-value is close to 0 (null hypothesis not rejected). If the alternative hypothesis is true and the sample sizes are large enough, the t-value will typically deviate substantially from 0

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK:

Given the degrees of freedom, there is a relationship between t-test and p-value that can be computed. For example, you can use this calculator:

https://www.statology.org/t-score-p-value-calculator/*

To test if a t-test and a p-value are consistent.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK:

Due to random sampling, the t-test value will sometimes deviate from 0 even if the null hypothesis is true, and the researcher will reject the null hypothesis by error. The decision about whether to reject the null hypothesis is often based on the p-value. If $p < \alpha$ (typically $\alpha = 0.05$ or 5%), then the observed t-test value deviates more from 0 than what should be common if the null hypothesis is true. Therefore, there seems to be an effect: the null hypothesis is rejected and the observed difference of means is statistically significant.

Data Detective Methods for RQP: Mathematical Inconsistencies and Data Gleaning

STATCHECK:

For example the statement “As predicted we found a significant difference between the control and experimental conditions, $t(22) = 2.00$, $p < 0.05$.” can be easily tested*

For df = 22, t=2.00 p-value is $p = 0.058 > 0.05$

So due to common mistakes (i.e. incorrect computations, mistyping etc) these errors are possible to occur, but they can (and should) be easily tested.

Summary

Statistical errors are easy to make. However, even small errors can change the results of a study significantly. It is therefore important to pay attention to the details and double check the final results.

Interesting to read:

Data Detective Methods for Revealing Questionable Research Practices

Publication bias and the failure of replication in experimental psychology

Questions?

Publication bias and the failure of replication in experimental psychology

Gregory Francis

Publication bias and the failure of replication in experimental psychology

Replication of empirical findings plays a fundamental role in science:

- successful replication enhances belief in a finding
- failure to replicate is often interpreted to mean that one of the experiments is flawed.

This view is not necessarily accurate.

Because experimental psychology uses statistics, empirical findings should appear with predictable probabilities.

The problem is that in a misguided effort to demonstrate successful replication of empirical findings and avoid failures to replicate, experimental psychologists sometimes report too many positive results.

Publication bias and the failure of replication in experimental psychology

Rather than strengthen confidence in an effect, *too much successful replication* actually indicates publication bias, which invalidates entire sets of experimental findings.

This article shows how an investigation of the effect sizes from reported experiments can test for publication bias by looking for too much successful replication. Simulated experiments demonstrate that the publication bias test is able to discriminate biased experiment sets from unbiased experiment sets, but it is conservative about reporting bias.

Publication bias and the failure of replication in experimental psychology

Ioannidis and Trikalinos (2007) described a test for whether a set of experimental findings contains an excess of statistically significant results. This publication bias test, is central to the present discussion:

The ability of repeated experiments to provide compelling evidence for the validity of an effect must consider the statistical power of the experiments.

If all of the experiments have high power (the probability of rejecting the null hypothesis when it is false), multiple experiments that reject the null hypothesis will indeed be *strong* evidence for the validity of an effect.

Publication bias and the failure of replication in experimental psychology

Even populations with strong effects should have some experiments that do not reject the null hypothesis. Such null findings should not be interpreted as failures to replicate, because if the experiments are run properly and reported fully, such nonsignificant findings are an expected outcome of random sampling.

Some researchers in experimental psychology appear to misunderstand this fundamental characteristic, and they engage in a misguided effort to publish more successful replications than are believable.

If there are *not enough null findings* in a set of moderately powered experiments, the experiments were either *not run properly or not fully reported*, hence there is no reason to believe the reported effect is real.

Publication bias and the failure of replication in experimental psychology

The difference between the observed (O) and the expected (E) number of studies that reject the null hypothesis for a set of (N) reported experiments can be analyzed by a x^2 test:

$$\chi^2(1) = \frac{(O - E)^2}{E} + \frac{(O - E)^2}{(N - E)},$$

The observed number of rejections is the number of reported experiments that reject the null hypothesis. The expected number of rejections is found by first estimating the effect size of a phenomenon across a set of experiments.

Publication bias and the failure of replication in experimental psychology

The difference between the observed (O) and the expected (E) number of studies that reject the null hypothesis for a set of (N) reported experiments can be analyzed by a x^2 test:

$$\chi^2(1) = \frac{(O - E)^2}{E} + \frac{(O - E)^2}{(N - E)},$$

The observed number of rejections is the number of reported experiments that reject the null hypothesis. The expected number of rejections is found by first estimating the effect size of a phenomenon across a set of experiments.