Parametric Statistics

Week 2 - Statistical Modeling and Sampling Distributions

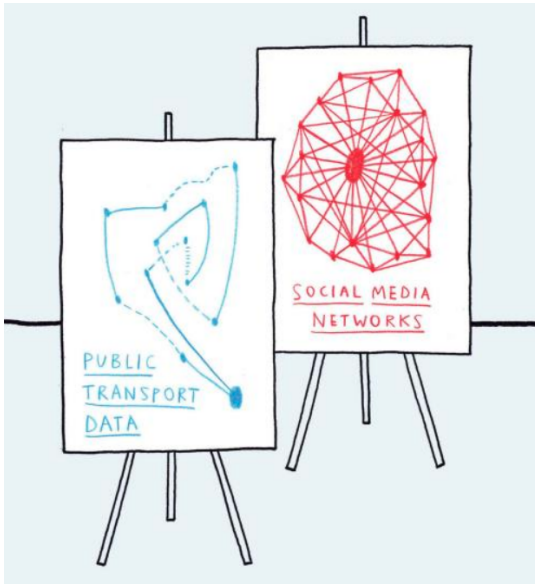
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 23. October 2019

political data science

https://politicaldatascience.blogspot.de



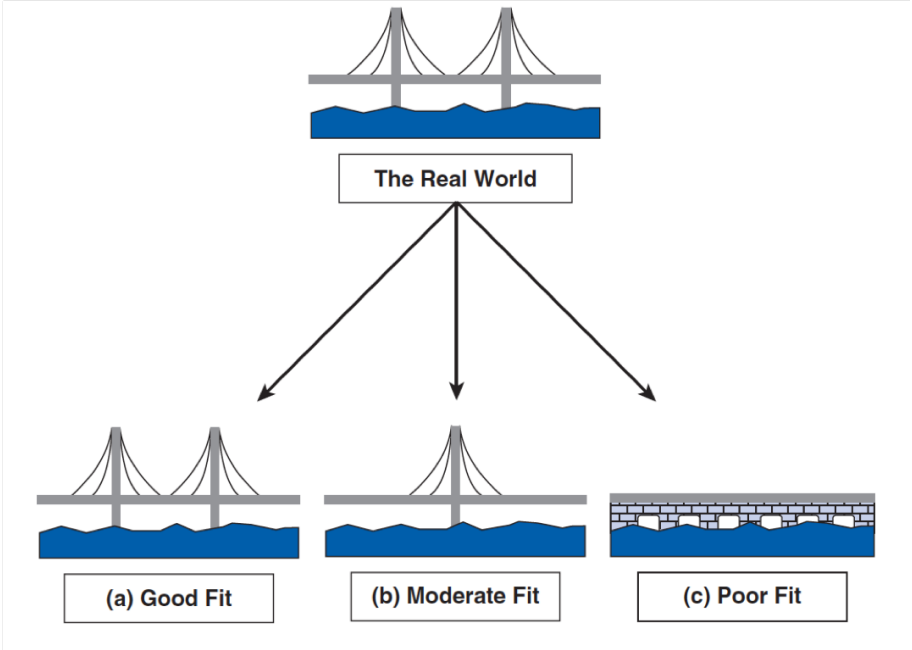
Statistical Models

Models try to approximate real-world phenomena

Fit of the model: The degree to which a statistical model represents the data collected.

A good model can be used to make predictions of the real-world

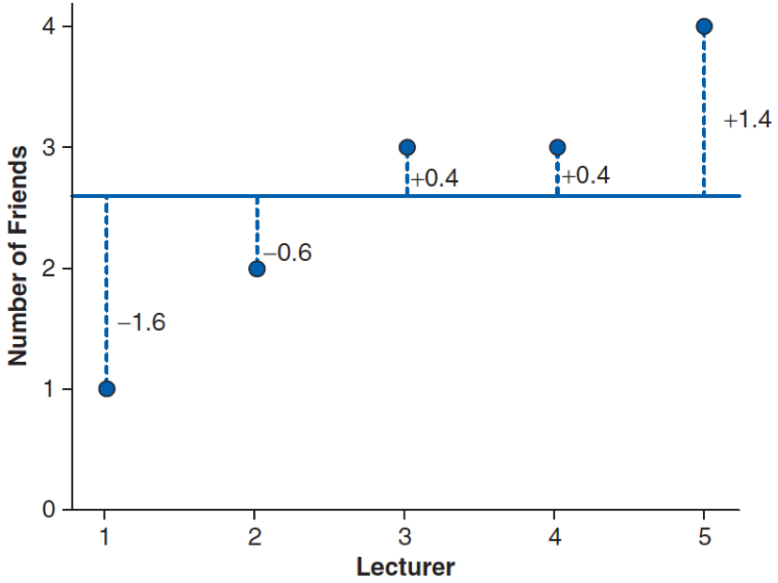
Central part in any research design



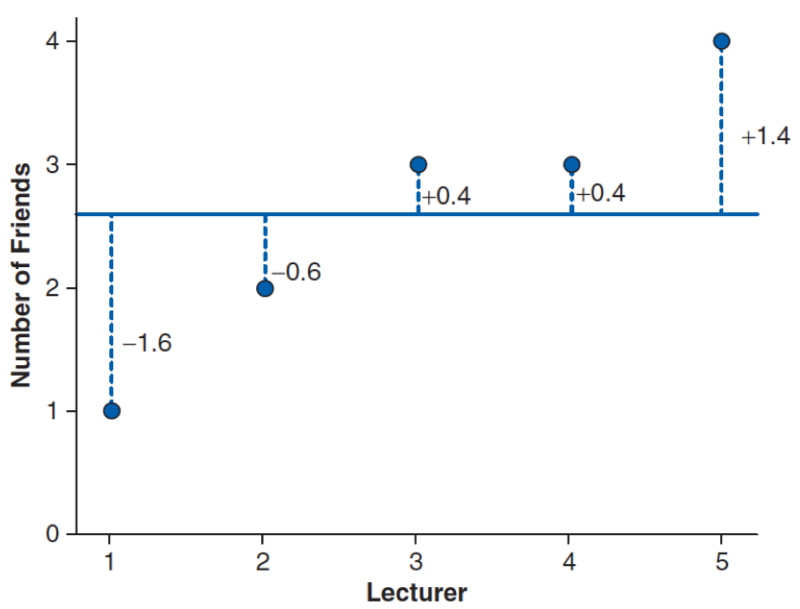
A very simple Statistical Model

Example:

- Research question: Do statistic teachers have friends? How many?
- Hypothesis: ... (Wait for next week)
- Statistical model: Take the mean of sample data
- Data collection: We ask 5 statistic teachers at TUM



A very simple Statistical Model



The vertical lines represent the deviance between the observed data and our model. They can be seen as error in the model.
How to assess the fit? For example, the sum of squared errors (SS):

$$SS = \sum (x_i - \overline{X})^2$$

But it increases with number of datapoints n
Idea: Divide SS by n - 1 ...
Variance! Variance and standard deviation are common measures of model fit.

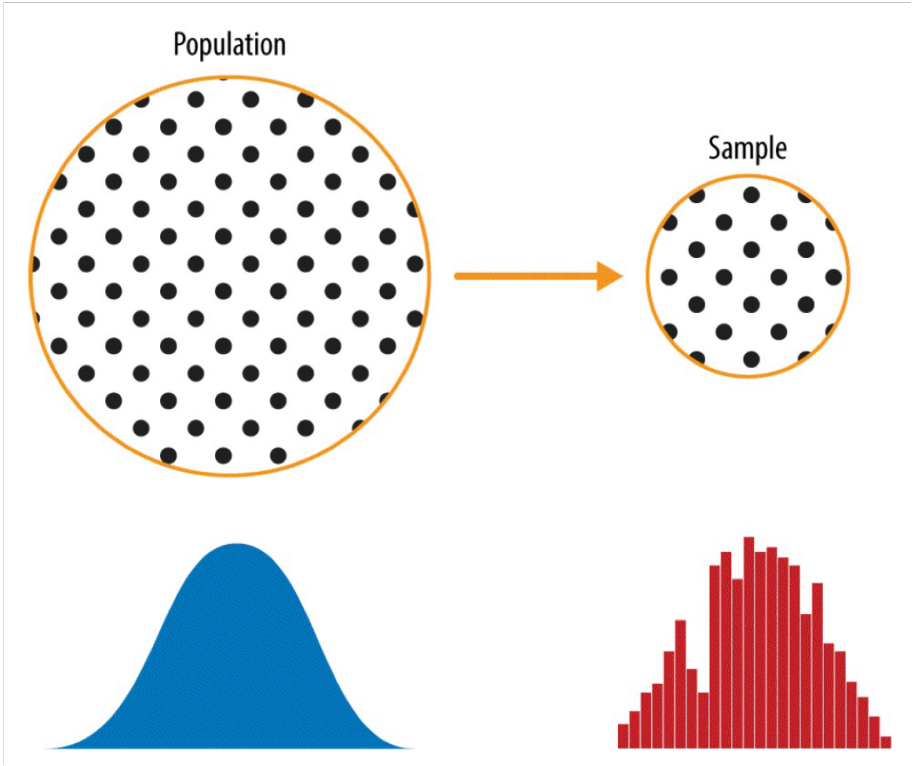
Fitting Statistical Models

$$Outcome_i = Model + error_i$$

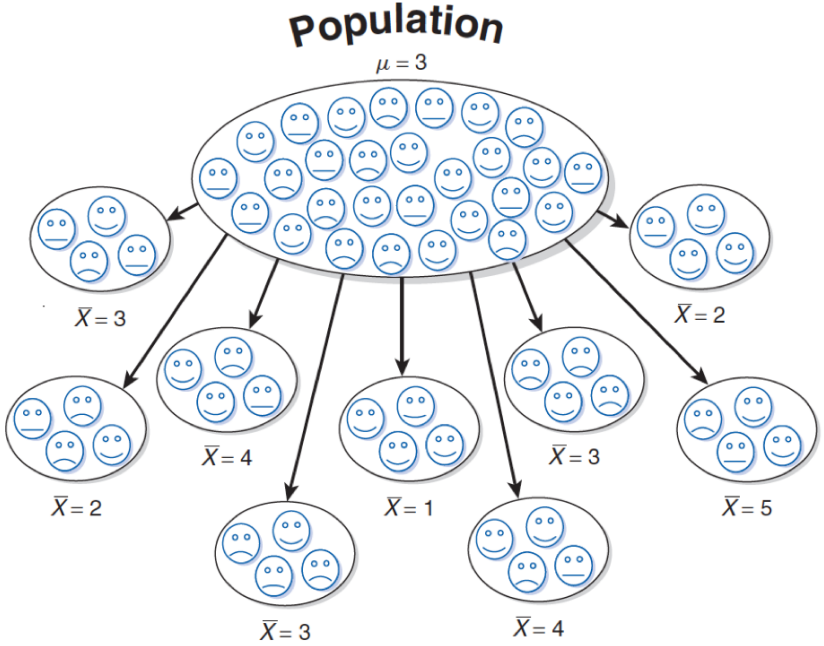
Here we have seen the fit of the model to the sample data, but how about the real world? Is the sample representative of the population?

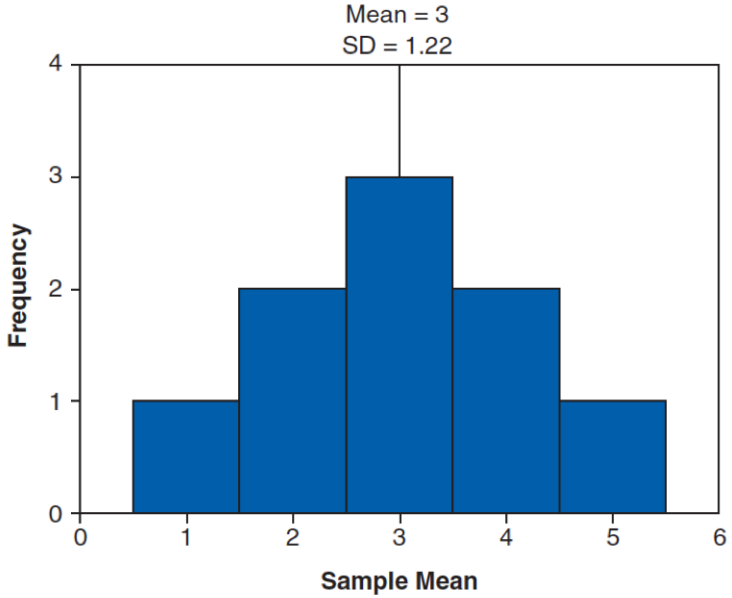
Population and Sample

The real distribution vs. the frequency distribution of a sample:



The Sampling Distribution





Standard Error

Important message 1: Learn to differentiate between probability distribution, frequency distribution and sampling distribution!

Important message 2: The **standard error** is the standard deviation of a sample parameter. In the case of the mean, its called standard error of the mean.

Important message 2: The **standard error** is the standard deviation of a sample parameter. In the case of the mean, its called standard error of the mean.

Important message 2: The **standard error** is the standard deviation of a sample parameter. In the case of the mean, its called standard error of the mean.

Standard Error

Why so important? It tell us how likely it is that one random sample is representative of the population.

If we have sample mean with a big standard error, the mean of many of the samples ($\overline{X}$) is far away from the real mean (μ)

In real life we can not collect hundreds of samples...Sad!
We need to estimate the standard error!
Statistics to the rescue!

Central Limit Theorem

The central limit theorem As samples get large (greater than 30), the sampling distribution has a normal distribution with a mean equal to the population mean, and a standard deviation of:

$$SE = \frac{\sigma}{\sqrt{N}}$$

where σ is the standard deviation of the population, since we normally don't know it, we estimate the standard error:

$$\hat{SE} = \frac{s}{\sqrt{N}}$$

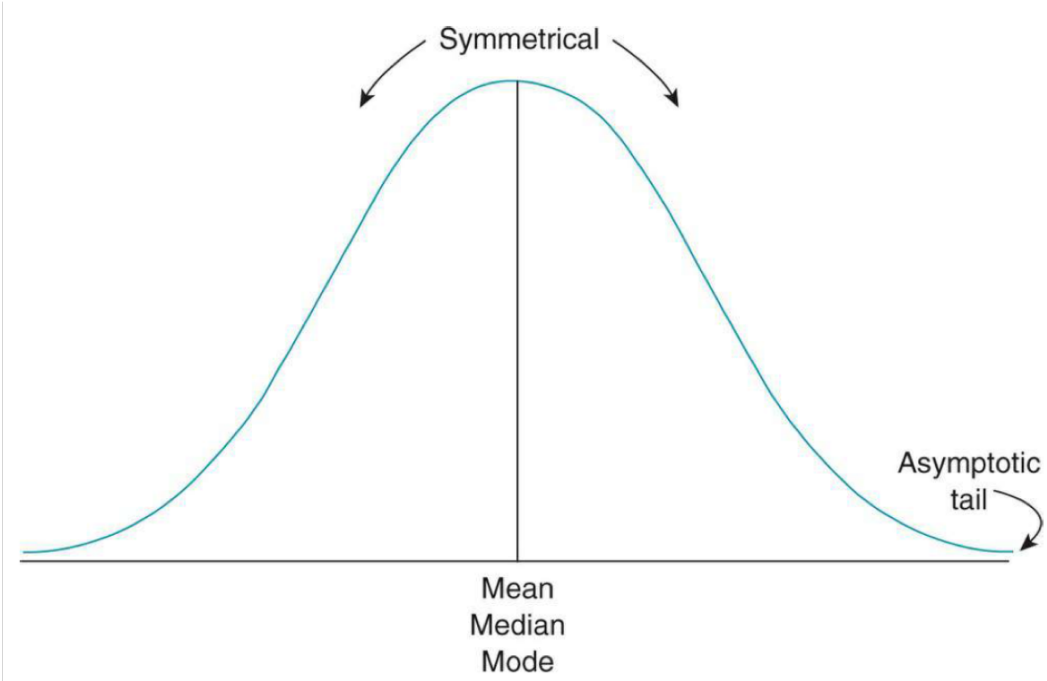
Important consequences:

- As N increases, the standard error decreases.
- Even if the data is not normal (like many real life phenomena), the sample distribution is normal distributed! This is why the normal distribution is so important.
- For samples smaller than 30 the sampling distribution has the shape of a **t-distribution**

Time to explore the normal and t distributions!

The Normal Distribution

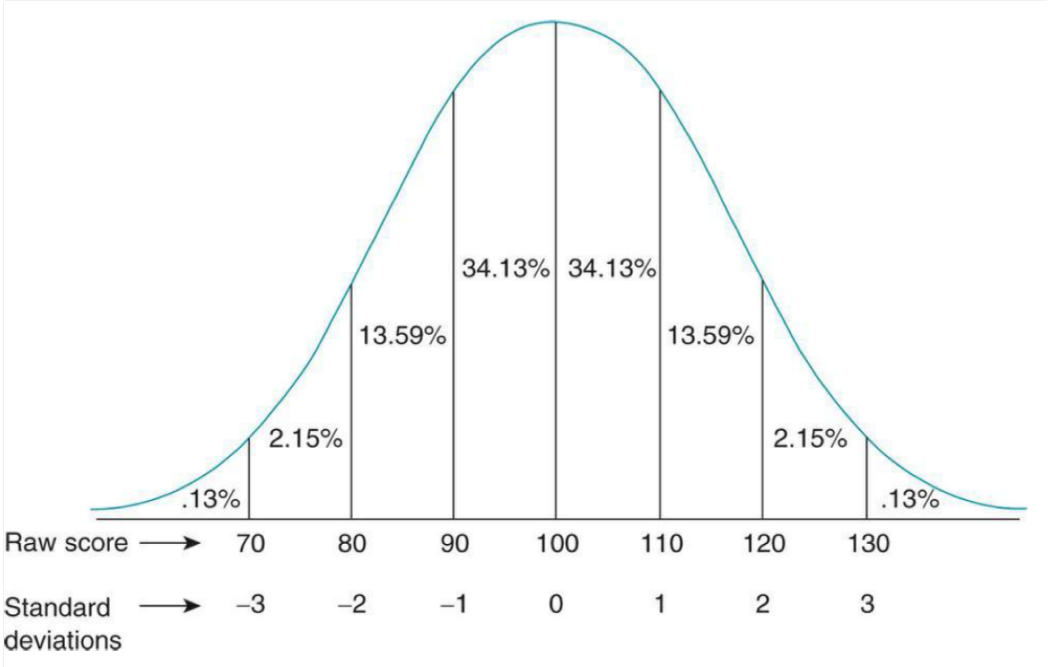
The normal distribution has **TWO** parameters: the mean and the standard deviation



The Normal Distribution

IMPORTANT: Percentage of area under the curve = probability

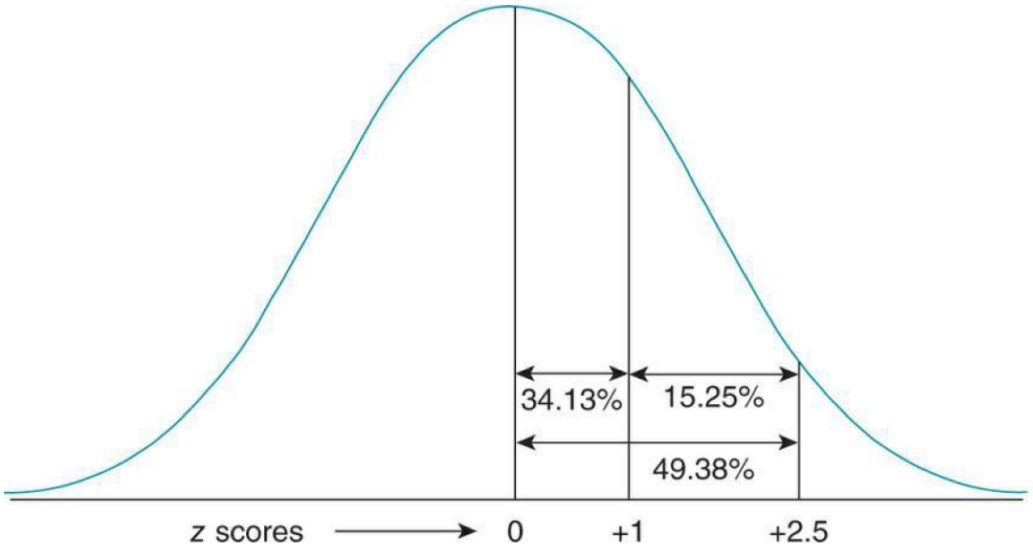
Example Normal distribution with mean $\overline{X} = 100$ and standard deviation $s = 10$



z-scores

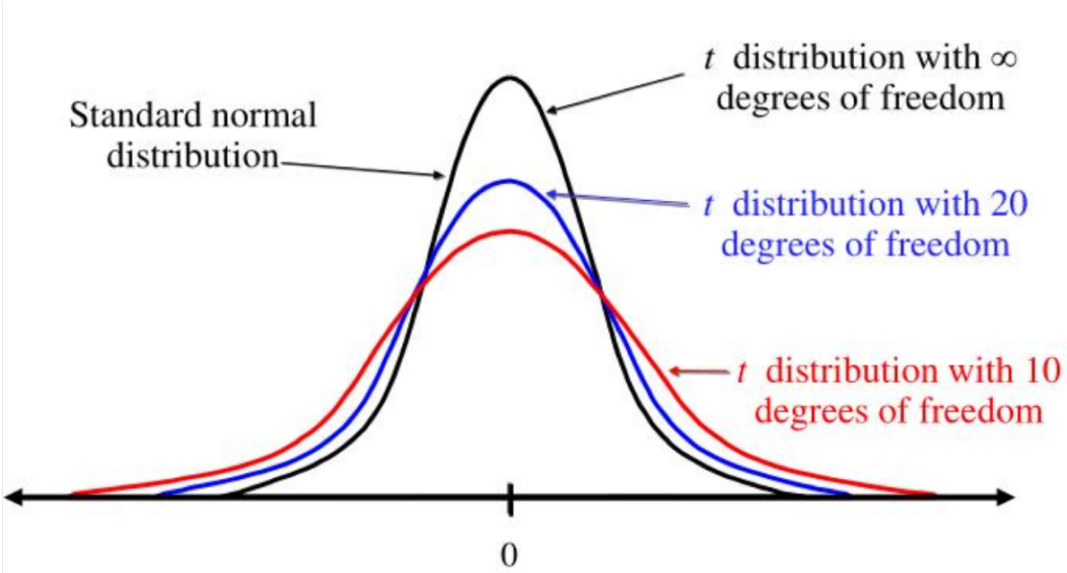
To compare different normal distributions, we need to transform them all to a standard normal distribution.
Standard Normal Distribution: Mean 0, standard deviation 1. To transform the data (X):

$$z = \frac{x - \overline{X}}{s}$$



t-Distribution

Similar to the normal distribution, but with more area on the tails and only **ONE** parameter.
Standard t-distribution: Mean always 0 and standard deviation depends on the degrees of freedom (df) = N - 1



Confidence intervals

Until now, we have...

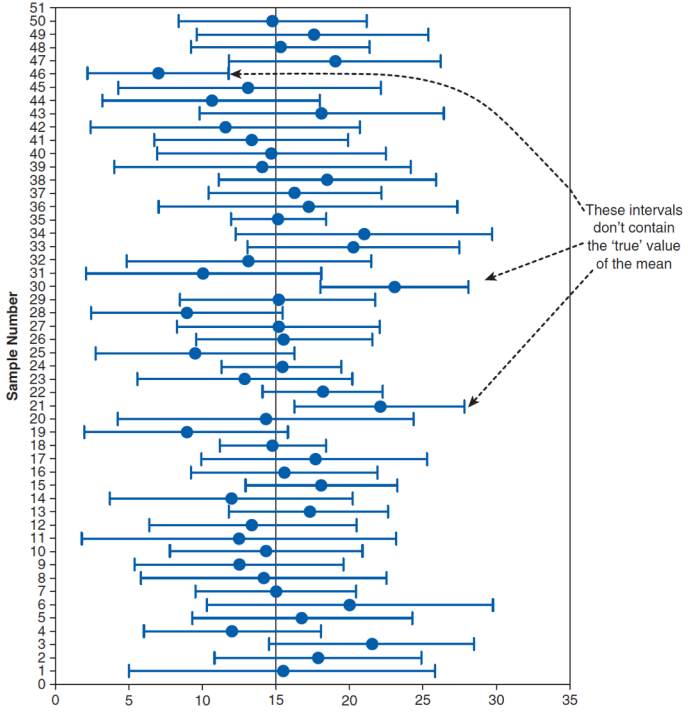
- ... proposed the sample mean as an estimate of the real mean value of the population.
- ... used the standard error to get an idea of the extent of which sample means differ .
- ... looked at the distributions that correspond to sampling distributions of the mean.

One last step...

Confidence intervals: Boundaries within which we believe the true value of the mean will fall.

What does a 95% confidence interval mean? Imagine we collect 100 samples, calculate the mean and for 95 of these samples the confidence intervals will contain the true value of the population

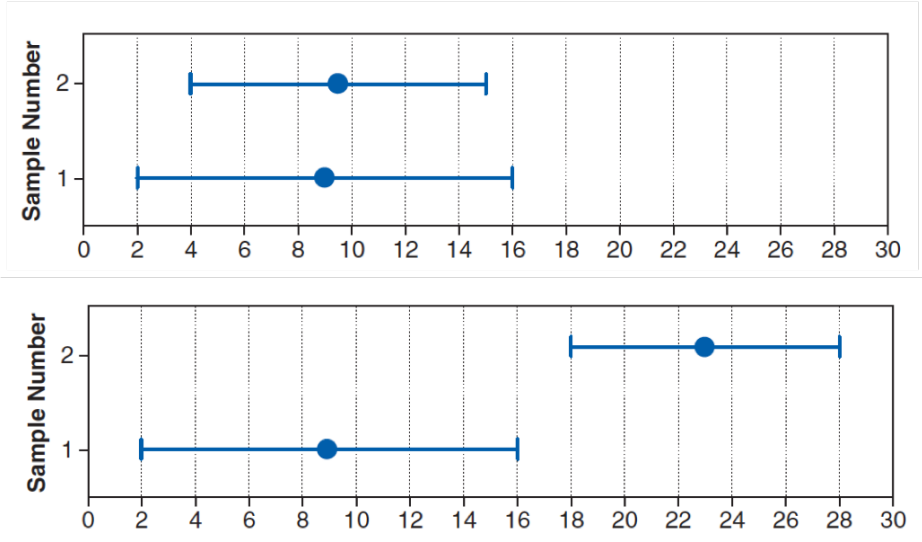
Confidence intervals



Example How many friends do statistical teachers have?

Confidence intervals

Do two samples belong to the same population?



In the second example, the means do not overlap...they are **significantly different** ... More on significance next week!

Confidence intervals

How to calculate them?

95% confidence interval
Observation: In a standard normal distribution, 95% of the z-scores fall between -1.96 and 1.96.
Use the Z-score formula backwards to find the original values:

$$z = \frac{x - \overline{X}}{s}$$

But we are interested in the variability of sample means, not on the variability in the data... Replace s with SE

$$-1.96 = \frac{x - \overline{X}}{SE} \qquad 1.96 = \frac{x - \overline{X}}{SE}$$

Rearrange the equations:

$$Lower \quad boundary = \overline{X} - (1.96 * SE)$$

$$Upper \quad boundary = \overline{X} + (1.96 * SE)$$

If we want a 99% confidence interval we use the values withing which 99% of the z-scores occur (-2.58 and 2.58).

Confidence intervals

How to calculate them?

For small samples:

$$Lower\ boundary = \overline{X} - (t_{n-1} * SE)$$
$$Upper\ boundary = \overline{X} + (t_{n-1} * SE)$$

the t_{n-1} score is equivalent of the z-score but from the t-distribution.

How to get the scores for a specific confidence interval?

- Use the formula of the distribution : Too complicated maths
- Use a table with the distribution values: Too old fashion
- Ask R: Best solution!

We will check this in the tutorial tomorrow

An Alternative Approach

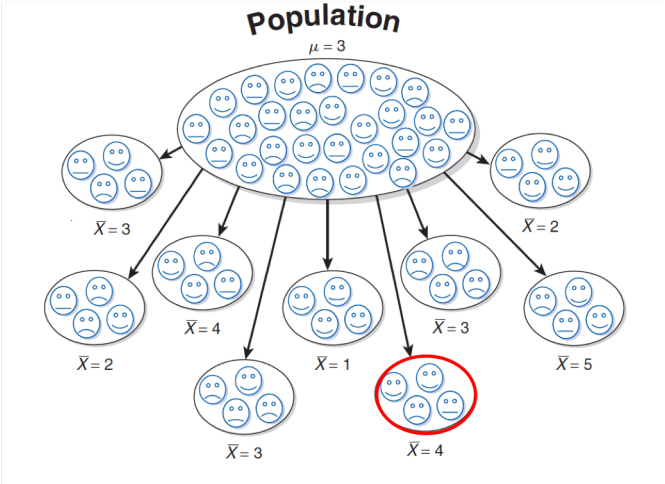
Can we calculate the standard error and confidence intervals without knowing nothing about distributions?

Yes we can! Only with one sample! Thanks to computers

How???

The Bootstrap

Inception idea...
We have one dataset, which is a sample of the population



What if we sample the sample many many times to simulate the other real samples of the population?
=> Treat the sample data as a population from which smaller samples (bootstraping samples) are taken:

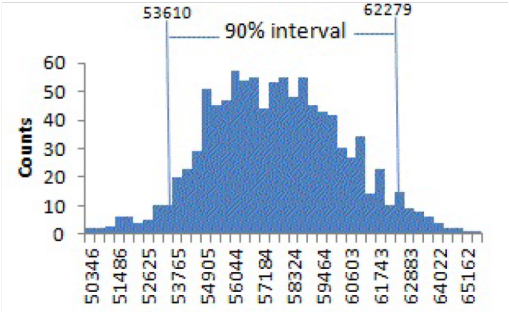
The Bootstrap

Dataset: N observations

Important: Sampling with replacement. (replace each observation after each draw)

Bootstrap Algorithm

1. Draw a sample observation, record it, replace it.
2. Repeat N times.
3. Record the mean of the n resampled values.
4. Repeat steps 1-3 R times (for example 1000 times or more)
5. With the R means, calculate the standard deviation. This estimates the sample mean standard error.
6. With the R means, plot a histogram and find a confidence interval directly.



Approach Selection

Two approaches for finding the reliability of an estimate. When to use which approach?

- Formula approach — Social Scientists for research papers and experiments (You need a theory)
- Bootstraping approach — Data Scientists for business decisions (The data has all the information)

This is just the beginning...

Until now... we calculated the standard error and confidence intervals of the sample mean, but we can do the same for any parameter we want to estimate!

Inferential statistics is based on inferring parameters and saying how confident we are about them

Literature

- Discovering Statistics using R : Chapter 2 (Before 2.6)
- Statistics for people who hate statistics : Chapter 8
- Practical Statistics for Data Science: Chapter 2