Introduction to Statistics

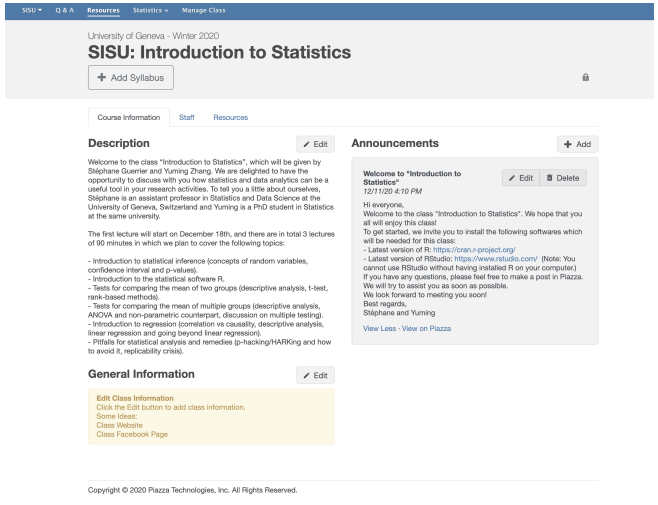
Part I: Introduction to Statistical Inference

Stéphane Guerrier & Yuming Zhang
18 December 2020

Introduction

Welcome to the class "**Introduction to Statistics**"!

Throughout this class, we will use a platform called **Piazza** for the share of all the teaching materials and Q&A.



- Signup Link: https://piazza.com/configure-classes/winter2020/sisu
- Access Code: statisfun

R and RStudio

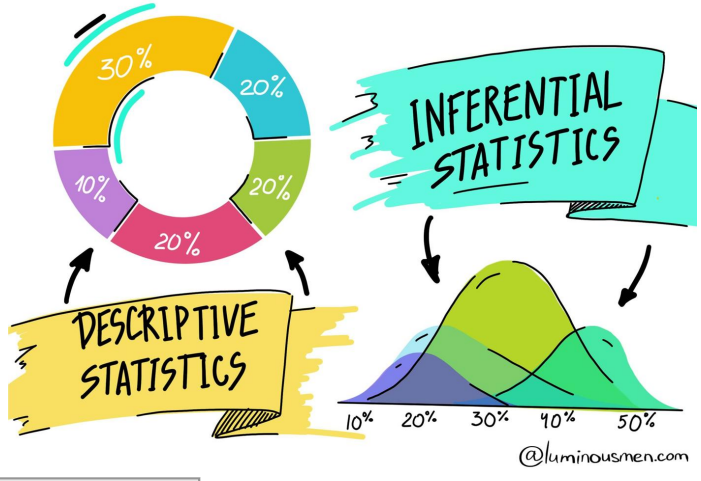
In this class, we will use the statistical software **R** together with the integrated development environment **R Studio**, which can be downloaded with the following:

- Latest version of R: https://cran.r-project.org/
- Latest version of R Studio: https://www.rstudio.com/

Note: You cannot use RStudio without having installed R on your computer.

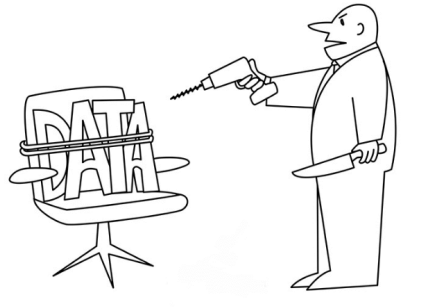
What is statistics?

Statistics is a science that uses mathematics and computer science to deal with the collection, analysis, interpretation, and presentation of masses of numerical data. Informally, it is the science of learning from data.



Source: luminousmen

Statistics is a crucial part of our life. However, statistical methods are often consciously (or not) misused. This can lead to contradictory studies and conclusions (as seen during the current COVID-19 pandemic).



"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Source: Atoz Markets

How can statistics be useful?

Statistics can be used (among others) to

1. Visualize data (e.g. propagation of COVID-19 in different countries).
2. Understand and interpret data (e.g. main causes of cancer).
3. Assess the validity of a hypothesis (e.g. is a drug working?).
4. Make predictions (e.g. predicting unemployment or risk indices).

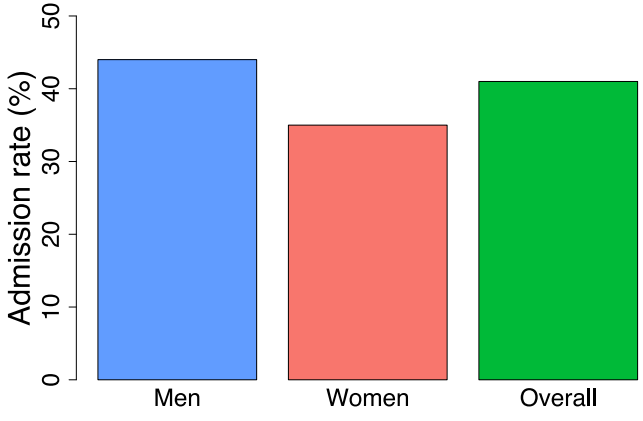
Learning more about statistics allows to

1. Better understand arguments based on data.
2. Be able to apply critical thinking about statistics used as evidence.
3. Understand how statistical associations are used to evaluate claims (hypotheses) and assess causal connections.

Understanding and knowing how to interpret statistical analyses is therefore becoming an increasingly vital skill.

Simpson's Paradox

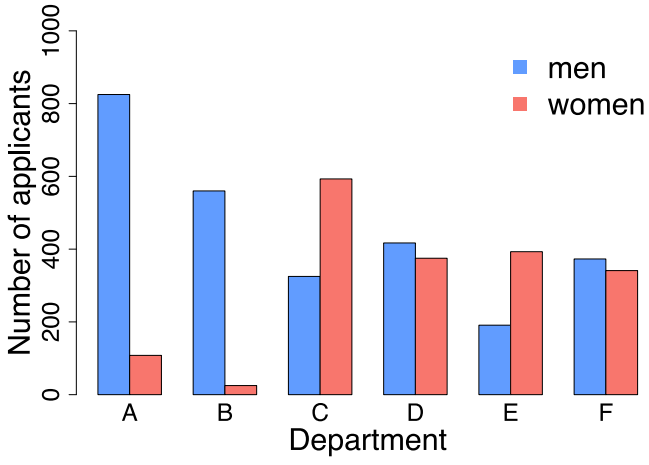
Statistical analysis can be tricky. Here we give an example of a study of gender bias among graduate school admissions to University of California, Berkeley, for the fall of 1973.

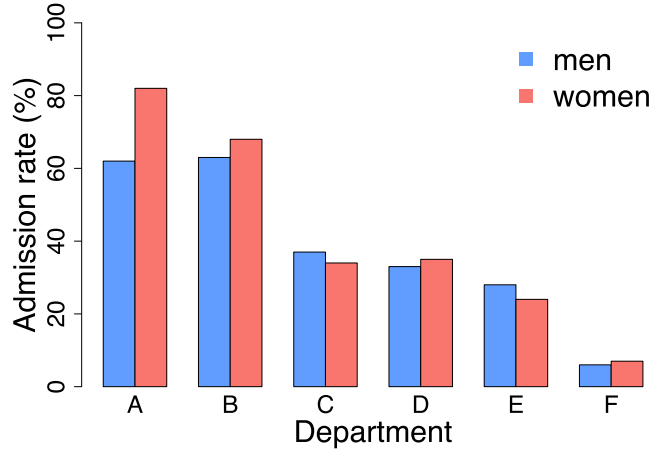


The data show that among 8,442 men applicants 44% were admitted while among 4,321 women applicants only 35% were admitted. The overall admission rate was around 41%. The difference is quite large (9%) and it is a large sample with 12,763 applicants, so it is unlikely that this is due to chance. Therefore, the data suggest that men applying were more likely to be admitted than women.

However, when people looked more into the data, they found that this conclusion is actually completely incorrect. In fact, a correct analysis showed that "small but statisticall significant bias in favor of women". But why? 😮

Simpson's Paradox



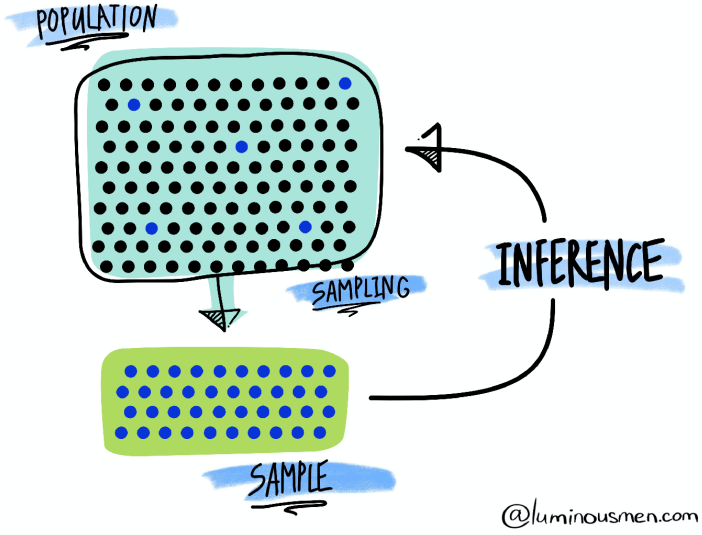


- Men applicants tended to apply for "easy" departments, i.e. departments that had high admission rates.
- Women applicants tended to apply to "hard" departments, i.e. departments that had low admission rates.
- So it turns out that most of the departments actually had a slightly higher success rate for women.

This phenomenon is actually very common and known as **Simpson's Paradox**.

How does it work?

- Statistical methods are based on several fundamental concepts, the most central of which is to consider the information available (in the form of data) resulting from a random process.
- As such, the data represent a **random sample** of a totally or conceptually accessible **population**.
- Then, statistical inference allows to infer the properties of a population based on the observed sample. This includes deriving estimates and testing hypotheses.



Source: luminousmen

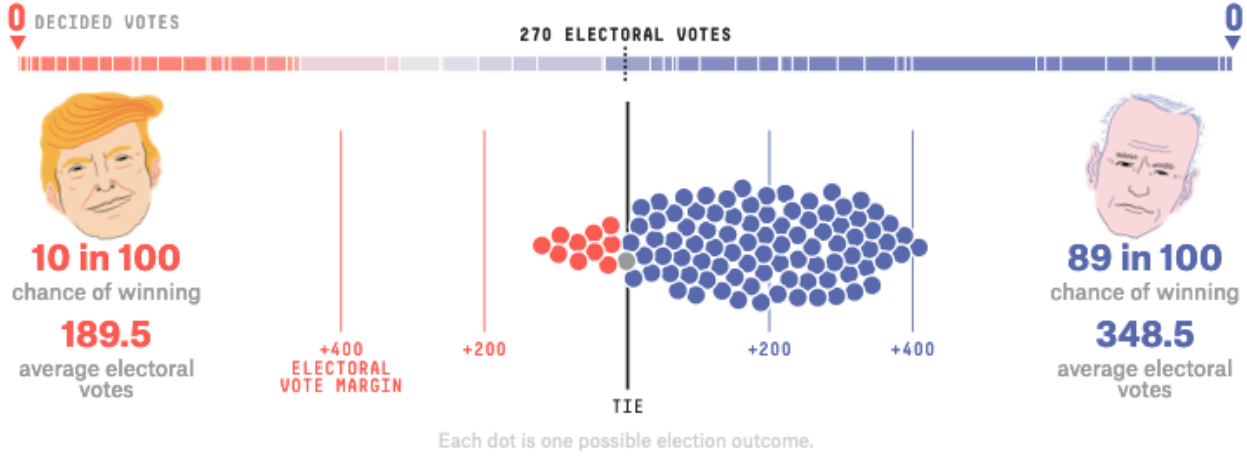
Outline

In this class, we plan to cover the following topics:

- Introduction to statistical inference (concepts of random variables, confidence interval and p-values).
- Introduction to the statistical software R.
- Tests for comparing the mean of two groups (descriptive analysis, t-test, rank-based methods).
- Tests for comparing the mean of multiple groups (descriptive analysis, ANOVA and non-parametric counterparts, discussion on multiple testing).
- Introduction to regression (correlation vs causality, descriptive analysis, linear regression and going beyond linear regression).
- Pitfalls for statistical analysis and remedies (p-hacking/HARKing and how to avoid it, replicability crisis).

Population and Sample - Example

To fix ideas we will consider a simple example. The 2020 United States presidential election was the 59th quadrennial presidential election, held on November 3, 2020. According to the latest estimates, **Biden's team received 51.3% of the votes while Trump's received 46.8%** 👋. Naturally, the result of American elections is not determined by the popular vote but suppose that we were interested in collecting data **before the vote** to assess if Biden's team will receive more than 50% of the votes.



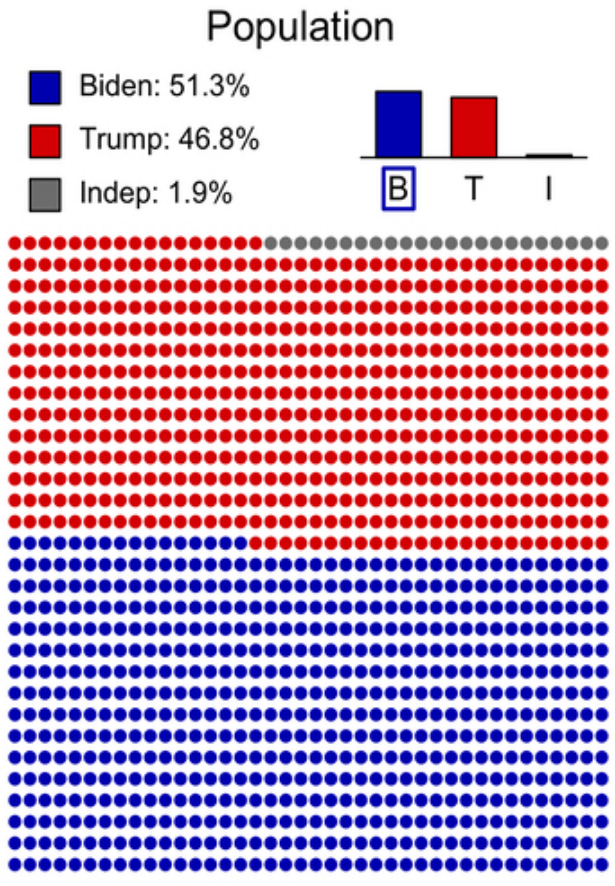
Source: Adapted from fivethirtyeight.

👋 More detailed analysis of the results can be found here.

Population and Sample - Example

In this example, we will make the following assumptions for simplicity:

- The American population of voters is composed of 1200 individuals (616 for Biden, 561 for Trump and 23 independents).
- We can perfectly sample the population (everyone is available, no double sampling, and the sampling is random).
- People don't change their mind and they don't lie.



Proportion Estimation

Suppose we ask n voters for which candidate (Biden, Trump or independent) they intend to vote in order to estimate the proportion of voters for Biden. For that we define the **random variables** $X_1, \ldots, X_n$ where X_i is defined as:

$$X_i = \begin{cases} 1 & \text{if voter } i \text{ intends to vote for Biden} \\ 0 & \text{otherwise.} \end{cases}$$

The random variables $X_1, \ldots, X_n$ are called a (random) **sample** and we refer to n as the **sample size**. Let p denote the (true) proportion of voters for Biden (which in this case is 51.3%), we then write

$$\Pr\left(X_i = 1\right) = p,$$

where $\Pr(A)$ denotes the probability of the event A.

Proportion Estimation

Using the random variables $X_1, \ldots, X_n$ we can define an **estimator** of p, which we often write as $\hat{p}$ and is given by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n},$$

where m denotes the number of voters in our sample in favor of Biden, and n is the sample size (as described previously).

An estimator is defined as a function of the data (i.e. $X_1, \ldots, X_n$), and therefore, theoretically any function of $X_1, \ldots, X_n$ can be an estimator.

However, in this case $\hat{p}$ is the best possible estimator of p 👋 and therefore it is not useful (in this case) to search for better estimators.

👋 More precisely, this estimator is unbiased (more info.) and has the smallest possible variance (more info.) as it attains the Cramér–Rao bound (more info.).

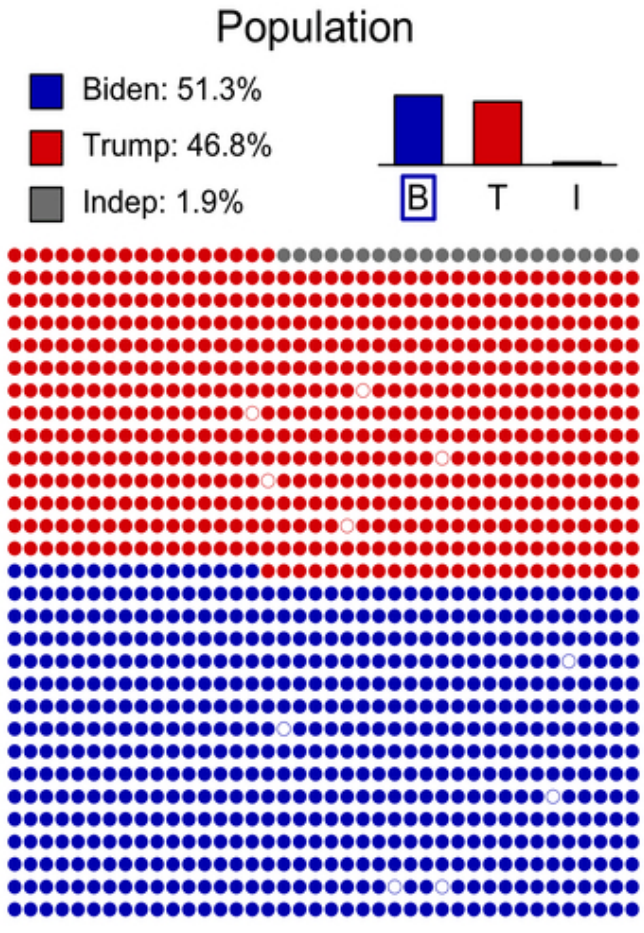
Population and Sample - Example

Consider a sample of $n = 10$ voters (randomly chosen from the population) as shown here 👉. In this case, we have 5 voters for Biden and 5 for Trump. We say that

$$x_1 = 1, \ldots, x_5 = 1, x_6 = 0, \ldots, x_{10} = 0$$

are **realizations** of the random variables $X_1, \ldots, X_{10}$.

We can now compute our estimator on the observed data (i.e. the realizations) and we obtain $\hat{p} = 0.5$. Therefore, our best guess based on the available data is that 50% of the voters will vote for Biden. Unfortunately, this doesn't really help us. So let's try with a bigger sample size... say



Population and Sample - Example

Population and Sample - Example

Population and Sample - Example

Population and Sample - Example

Population and Sample - Example

- In our example, we are interested in knowing if p, a population-level quantity, is bigger or smaller than 50%. Unfortunately, p is generally unknown as we cannot access the whole population (otherwise let's not bother with Statistics! 🤪). Therefore, we use $\hat{p}$ instead, a sample-dependent quantity.
- However, as we can see, $\hat{p}$ is random in the sense that it can change depending on the collected sample (e.g. we get different answers when $n = 200$, where Trump is leading, and when $n = 600$, where Biden is leading).
- To address this issue, we need to assess the **uncertainty** of $\hat{p}$ (i.e. assess how different $\hat{p}$ and p can be).
- Statistics can provides us many tools allowing to determine uncertainty as well as the associated decision-making risks.

How to measure uncertainty?

Uncertainty can be measured in many different ways. A common approach (in statistics) is to use **confidence intervals**, which rely on the **Central Limit Theorem (CLT)** that states:

"The sampling distribution of the sample mean approaches to a normal distribution as the sample size gets larger."

Loosely speaking, we can translate the CLT as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\cdot}{\sim} \mathrm{N}(\mu, \sigma^2),$$

where $\mathrm{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 (typically computed using the data)👋. Here $\bar{X}$ denotes the sample mean and $\overset{\cdot}{\sim}$ represents "approximately distributed as".

👋 Check out expected value and variance.

How to measure uncertainty?

In our example, we have

$$\hat{p} \overset{\cdot}{\sim} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right).$$

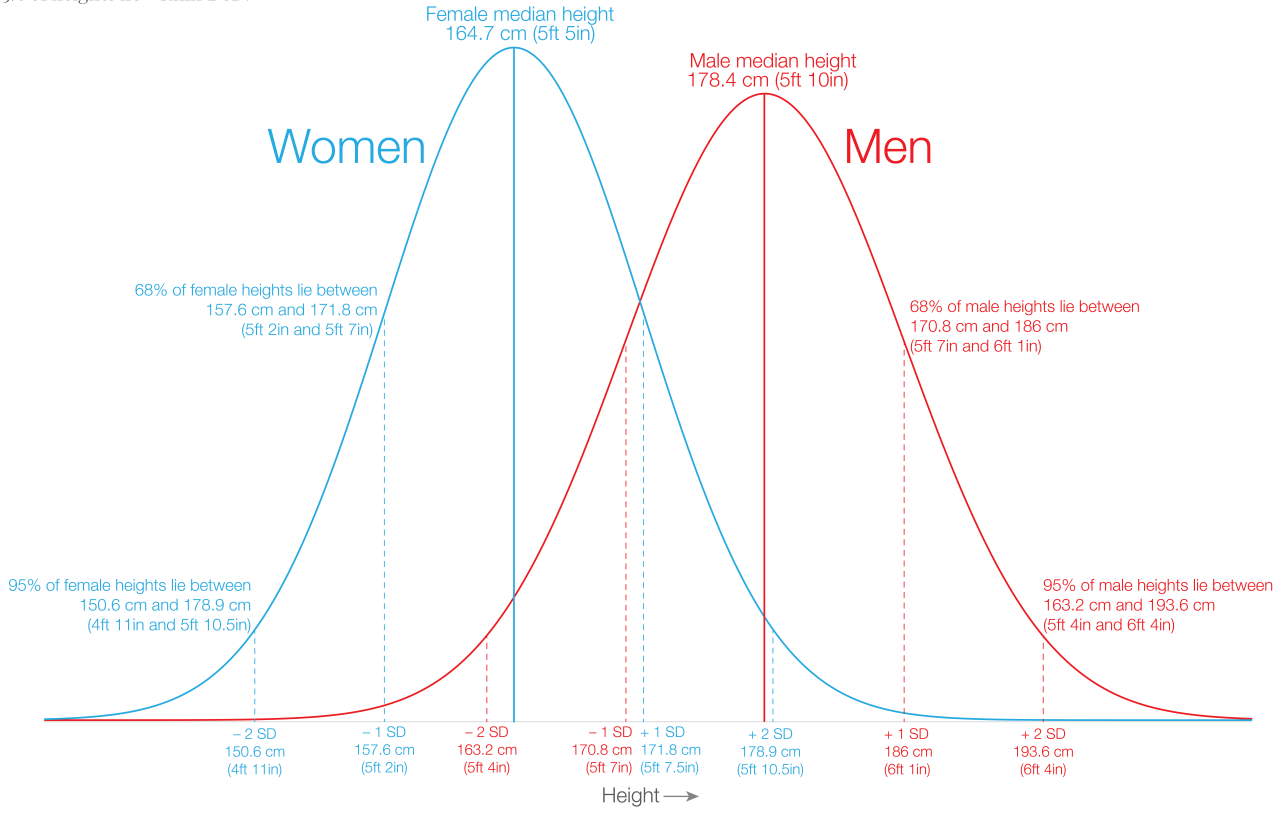
How to understand the practical implications of the CLT? Informally, it means that when a measurement can be thought of as the sum (or the average) of **numerous** factors, its distribution tends to go to a normal distribution. For example, the height of adults can be thought of as the sum of their genetic information, diet, life style, ...

Distribution of heights

The distribution of male and female heights

The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.

Our World in Data



Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.

Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.

Central Limit Theorem - Example

Central Limit Theorem - Example

Central Limit Theorem - Example

Central Limit Theorem - Example

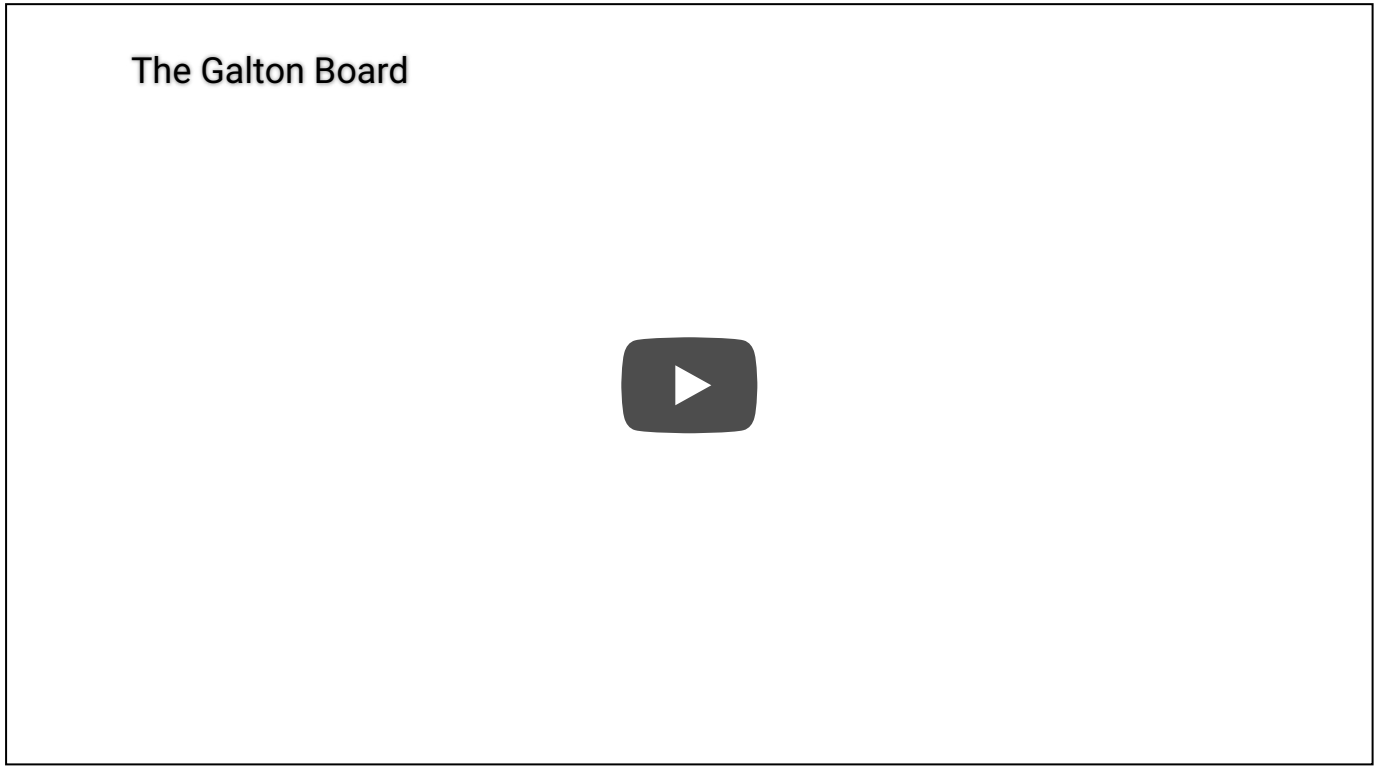
Central Limit Theorem - Example

Central Limit Theorem - Example

Central Limit Theorem - Example

Central Limit Theorem - Example

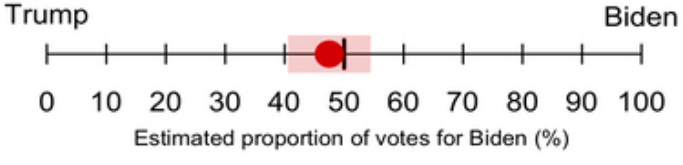
People have used the Galton Board as a practical device to demonstrate the CLT, in particular that with sufficient sample size the binomial distribution approximates a normal distribution.



Confidence Intervals

Now that we learn about CLT, how can we use it to construct **confidence intervals**?

- Confidence intervals correspond to a range of values that are likely to include the population value with a certain level of confidence. The level of confidence is a probability expressed as a percentage (%).
- In our example, we are interested in the true (population) proportion of voters in favor of Biden (51.3%). Using our sample of $n = 200$, we had $m = 95$ and we can construct the following 95% confidence interval:



So what does it mean? 🤔 It means that with a **probability of 95%**, the true proportion of voters for Biden (51.3% in this case) is between 40.58% and 54.42%.

Confidence Intervals

How is this confidence interval computed? We recall that by CLT,

$$\hat{p} \overset{\cdot}{\sim} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right).$$

However, as we **do not know** the true value of p in practice, we replace it with the **best guess** we have, which is $\hat{p}$. We then can write

$$\hat{p} \overset{\cdot}{\sim} \mathrm{N}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right).$$

The $1-\alpha$ confidence interval for p is then given by

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

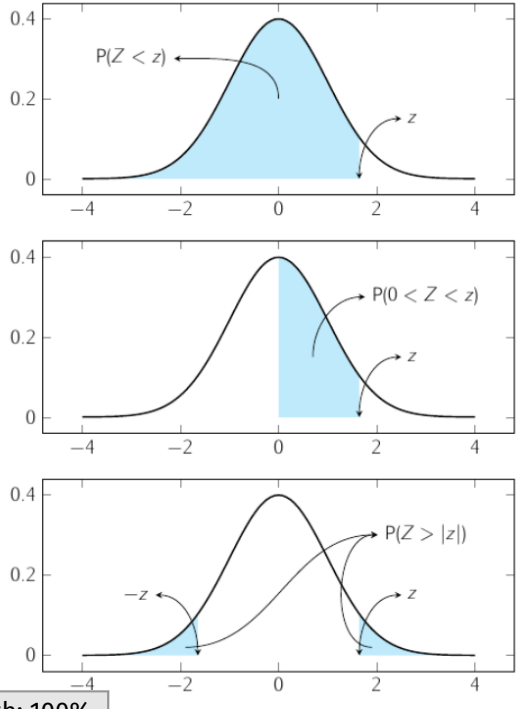
Processing math: 100%

$Z_{1-\alpha/2}$? 🤓

Confidence Intervals

$Z_{1-\alpha/2}$ corresponds to the $1-\alpha/2$ quantile of a standard normal distribution $\mathrm{N}(0, 1)$, that is,

$$Z_{1-\alpha/2} \text{ is such that } \Pr\left(Z \leq Z_{1-\alpha/2}\right) = 1-\alpha/2 \text{ where } Z \sim \mathrm{N}(0, 1).$$



- For 90% confidence interval, $Z_{0.95} \approx 1.64$.
- For 95% confidence interval, $Z_{0.975} \approx 1.96$.
- For 99% confidence interval, $Z_{0.995} \approx 2.58$.

Processing math: 100%

Confidence Intervals

Therefore, a confidence interval corresponds to a range of values that contains the true unknown population-level quantity we are considering with a probability of approximately $1 - \alpha$ (typically 95%).

Basically, we have

$$\Pr\left(p \in \hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha.$$

⚠ This means that a fraction of $100 \times \alpha\%$ of confidence intervals **don't include** p.

Confidence Intervals

Confidence Intervals

How to compute confidence intervals

Example R Code - 💪

Previously, we said that in a sample of $n = 200$, we had $m = 95$. In this case, we mentioned that (40.58%, 54.42%) was a 95% confidence interval. But how are these numbers computed? We have

$$\hat{p} = \frac{m}{n} = \frac{95}{200} = 47.5\%.$$

Then, we have

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.0353.$$

To obtain a 95% confidence interval we used $Z_{1-\alpha/2} = Z_{0.975} \approx 1.96$ and we get

Confidence Interval with $\alpha = 20\%$

Confidence Interval with $\alpha = 20\%$

Confidence Interval with $\alpha = 20\%$

Confidence Intervals with $\alpha = 20\%$

Why is 20%? Why not 5%?

What if I want to be certain? 😳

⚠ Take home message

- Since the data is available through sampling, it is **random**. Therefore, a decision or prediction can never be made with certainty!
- The only certainty one can have is that, for example, a proportion will always be included in the interval from 0% to 100%. **However, this is neither informative nor useful** and it does not even depend on the data.
- There exists a trade-off between **risk** as measured by $1 - \alpha$ (typically 95%) the confidence level, and the **precision of the conclusion** as measured, for example, by the confidence interval length.
- Moreover, the larger the sample size, the more precise the conclusion, for the same confidence level.
- Therefore, every decision based on statistical methods has a risk and how much risk is acceptable depends on the context (e.g. safety in airplanes vs which soft drink tastes better).

Example: Are there more 👦 than 👧?

Problem | R Code - 💪 | R Code - 🤖

An American found 13,173 boys were born among 25,468 newborn children. Is this sample an evidence that the birth of boys may be more common than the birth of girls in the entire population? So, we have $n = 25,468$ and $m = 13,173$. Therefore, we have

$$\hat{p} = \frac{m}{n} = \frac{13173}{25468} \approx 51.728\%.$$

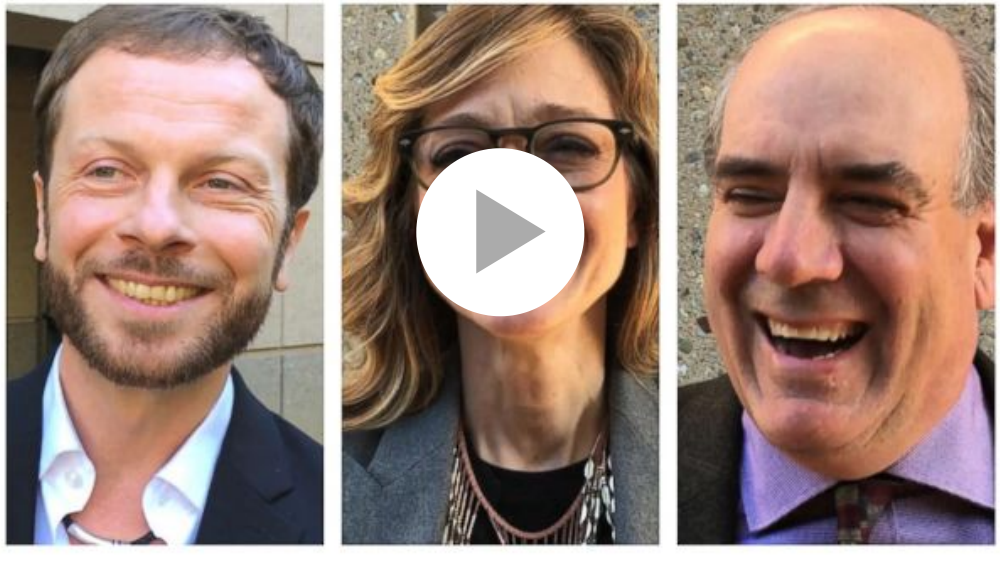
Then, we have

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.003131.$$

To obtain a 95\% confidence interval we used $Z_{1-\alpha/2} = Z_{0.975} \approx 1.96$ and we get

How to test a (scientific) hypothesis?

- An alternative summary measure of uncertainty is provided by **p-values** that take values between 0% and 100%.
- However, **p-values have been misused** many times because understanding what they mean is not intuitive.



👋 If you want to know more have a look here.

How to test a (scientific) hypothesis?

- A p-value is associated to a (couple of) hypotheses about the phenomenon under investigation. A hypothesis testing is designed to assess the strength of evidence against a baseline hypothesis that is called null hypothesis H_0 and in favor of another hypothesis that is called alternative hypothesis H_a.

- In the Biden-Trump example, we can write

$$H_0 : p = 0.5 \text{ and } H_a : p > 0.5.$$

- Each hypothesis excludes the other, so that one can exclude one in favor of the other using the data.
- The null hypothesis is the one that one will never be able to prove because the data is random.
- The alternative hypothesis is the one that offers more choice of values and hence has a chance to be favored with respect to the null hypothesis.

Testing

- Informally, **a p-value can be understood as a measure of plausibility of the null hypothesis given the data**. The smaller the p-value the greater the incompatibility of the null hypothesis with the data.
- When the p-value is small enough (typically smaller than 5%), one says that the test based on the null and alternative hypotheses is **significant** or that the null hypothesis is rejected in favor of the alternative. This is generally what we want because it "verifies" our (research) hypothesis.
- When the p-value is not small enough (typically larger than 5%), with the available data, we cannot reject the null hypothesis and then **nothing** can be concluded. 🤔
- With a sample of data, the obtained p-value (associated to a couple of hypotheses) summarizes somehow the **incompatibility between the data and the model** (random process) constructed under the set of assumptions.
- The p-value is usually compared to a threshold value that sets the (subjective) risk level of decision in favor of the incompatibility. The risk level is called the significance level and is a small value, usually 5%, but this depends on the context.

Example: Biden-Trump

Problem R Code

Returning to our Biden-Trump example, suppose we believe (or hope to show that) Biden will have more than 50% of the votes. We collect data with $n = 600$ and $m = 322$. We will consider the following steps to set up the test:

1. Define hypotheses: $H_0 : p = 0.5$ and $H_a : p > 0.5$.
2. Define α: We consider $\alpha = 5\%$.
3. Compute p-value: p-value = 3.959% (see computation tab for details).
4. Conclusion: We have p-value < α so we can reject the null hypothesis and conclude that the proportion of voters for Biden is greater than 50%.

Example: Biden-Trump

Problem R Code

What if we want to check if Trump will have more than 50% of the votes. Using the same data, we set up the test as follows:

1. Define hypotheses: $H_0: p = 0.5$ and $H_a: p < 0.5$.
2. Define α: We consider $\alpha = 5\%$.
3. Compute p-value: p-value = 96.04% (see computation tab for details).
4. Conclusion: We have p-value > α and we cannot reject the null hypothesis. We don't have enough evidence to conclude that Trump will have more than 50% of the votes.

Example: Breastfeeding and income

Problem R Code

A researcher who is studying the effects of income levels on breastfeeding of infants hypothesizes that countries where the income level is lower have a higher rate of infant breastfeeding than higher income countries. It is **known** that in Germany (high-income country), 22% of all babies are breastfeed. In Tajikistan (low-income country) researchers found that in a random sample of 500 new mothers that 125 were breastfeeding their infants. What can we conclude?

1. Define hypotheses: $H_0 : p = 0.22$ and $H_a : p > 0.22$.
2. Define α: We consider $\alpha = 5\%$.
3. Compute p-value: p-value = 5.874% (see computation tab for details).
4. Conclusion: We have p-value > α and we cannot reject the null hypothesis. We don't have enough evidence to conclude that countries where the income level is lower have a higher rate of infant breastfeeding than higher income countries.