Introduction to Statistical Literacy

Analyse et traitement de l'information

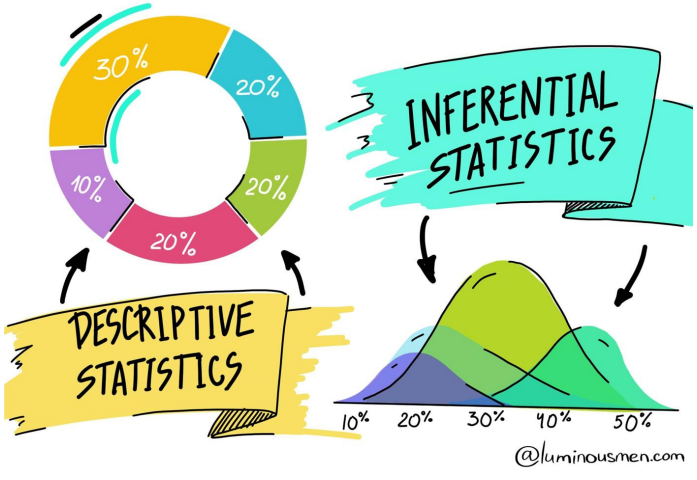
Comprendre le numérique

Stéphane Guerrier

24 Octobre 2023

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 (but how? 🤔). This can lead to contradictory studies and conclusions (as seen during the current COVID-19 pandemic).

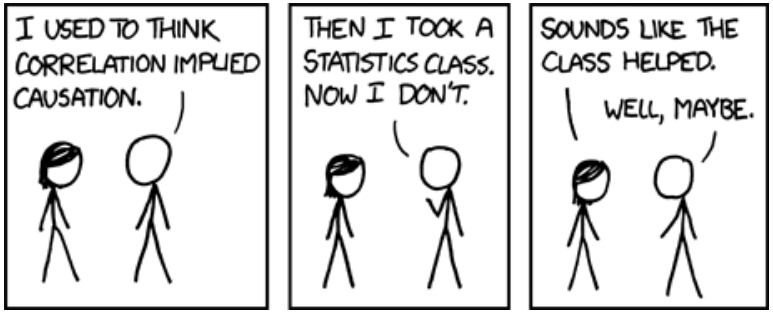
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Source: Atoz Markets

What is statistical literacy?

Statistical Literacy can be defined as follows:

- The ability to understand data, or arguments that use data.
- Being able to apply critical thinking about numbers and about statistics used as evidence in arguments.
- Understanding how statistical associations are used to assess causal connections.

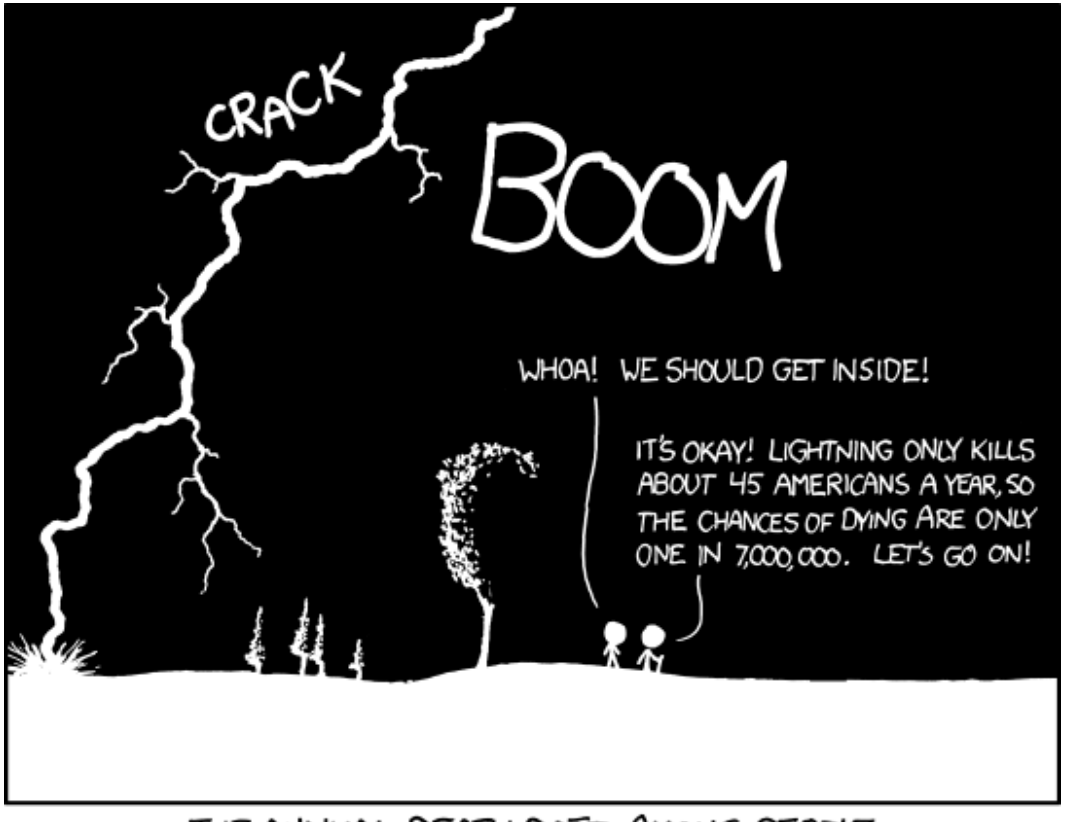


Source: xkcd

👋 If you want to know more about these comic strips take a look here.

Statistical literacy is now becoming an essential skill. It is, for example, needed by anyone who tries to learn new information from data. Scientists also need to develop statistical literacy so that they can produce both rigorous and reproducible research.

(Mis)Understanding statistical results



Source: xkcd

👋 If you want to know more about these comic strips take a look here.

How can statistics be useful?

As mentioned previously, statistics is the science of learning from data. It can be used (among others) to:

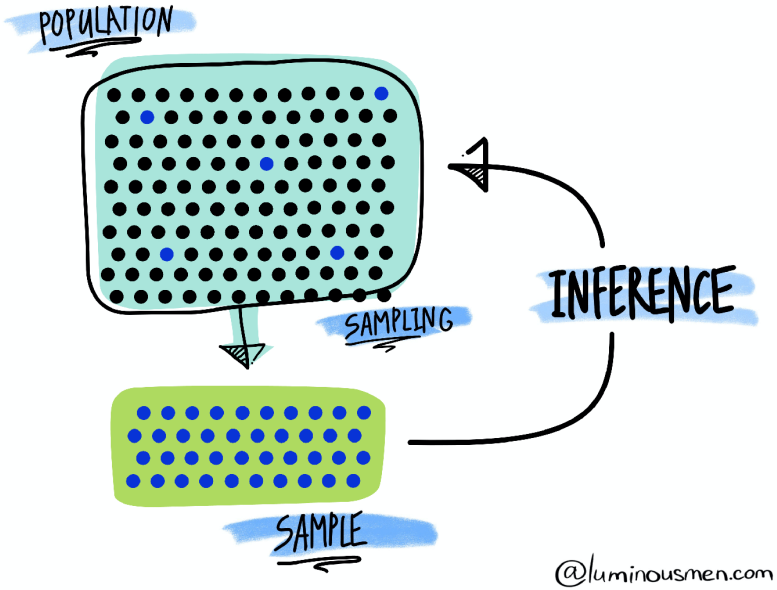
1. Visualize data (e.g. propagation of COVID-19 in different countries).
2. Understand and interpret data (e.g. main causes of cancer).
3. Make predictions (e.g. predicting unemployment or risk indices).
4. **Assess the validity of a hypothesis**.

In this class, we will focus on the last point: "*how to assess the validity of a hypothesis using statistics?*". More precisely, we will discuss how statistical methods can be used to evaluate claims (hypotheses) such as:

1. Coffee consumption increases blood pressure (really 🙄 ☕?).
2. Republican politicians are bad/good for the American Economy.
3. A glass of red wine is as good as an hour at the gym (🍷 🏃 😆).

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**.



Source: luminousmen

How does it work?

- Statistics allows to determine the **decision-making risks** (in the form of probabilities) associated with learning (decisions) based on the analysis of data extracted from samples. This allows to evaluate the validity of a claim (such as the ones we discussed previously).
- Understanding and knowing how to interpret statistical analyses is therefore becoming an increasingly vital skill.
- In some sense, **statistics plays a crucial role in our life** as new discoveries (obtained using statistical methods) directly impact us👋.
- However, understanding statistical results is not intuitive. There exist many pitfalls such as p-hacking or HARKing (e.g. why are there so many contradictory studies?).

👋 This is essentially true in the area of **Big Data** where the discovery of patterns and/or signals hidden in massive amounts of data is increasingly becoming one of the major challenges faced by modern scientific research.

So many studies?

"A glass of red wine is as good as an hour at the gym"? 🙄

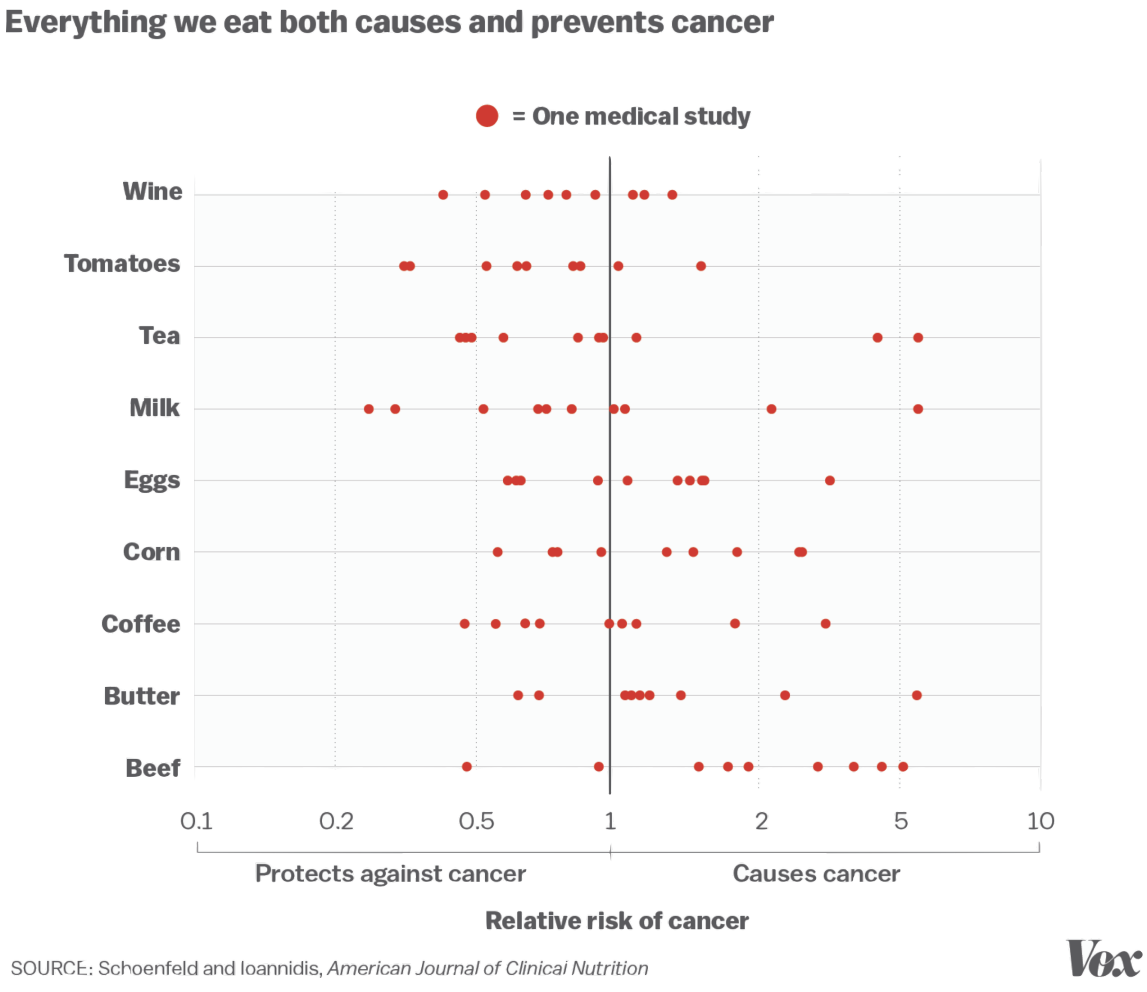
Source: Scientific Studies: Last Week Tonight with John Oliver (HBO)

Contraditory studies?

What about coffee? Is it good?

Source: Scientific Studies: Last Week Tonight with John Oliver (HBO)

So everything causes cancer?



👋 Read the original article: "*This is why you shouldn't believe that exciting new medical study*" here.

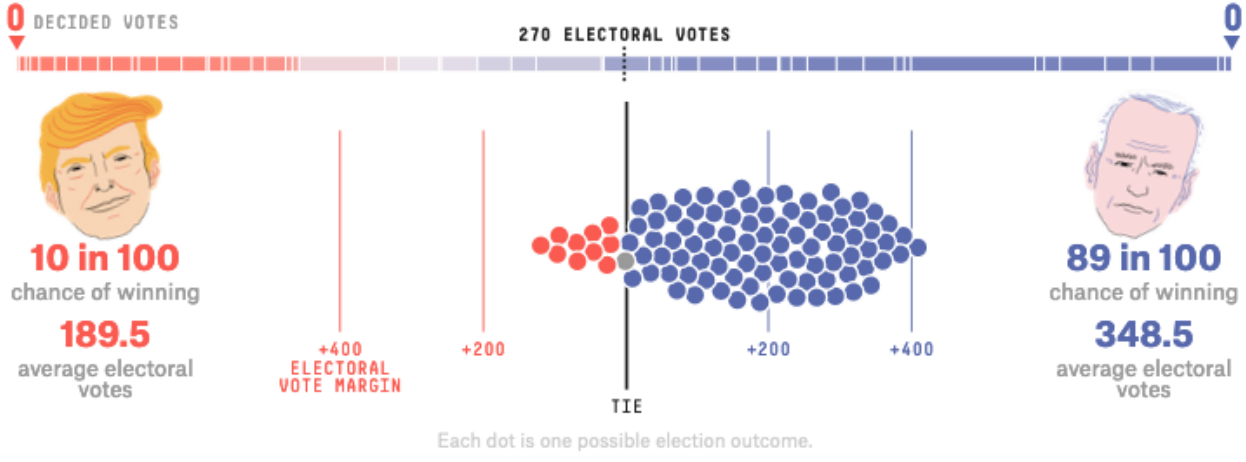
What are the consequences?

Which study should I pick?

Source: Scientific Studies: Last Week Tonight with John Oliver (HBO)

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.

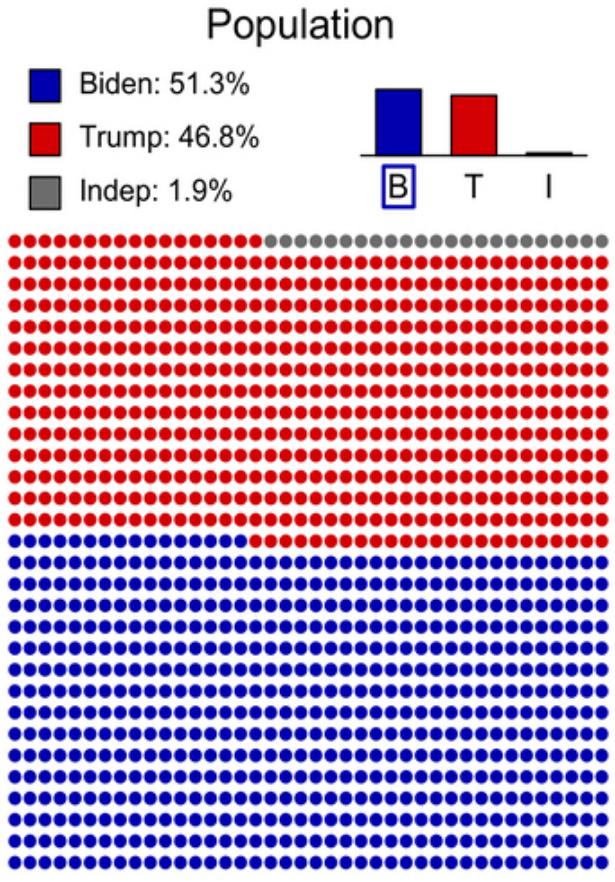


Source: Adapted from fivethirtyeight.
👋 More details on the results can be found here.

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

$$\Pr\Big(X_i = 1\Big) = p,$$

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.).

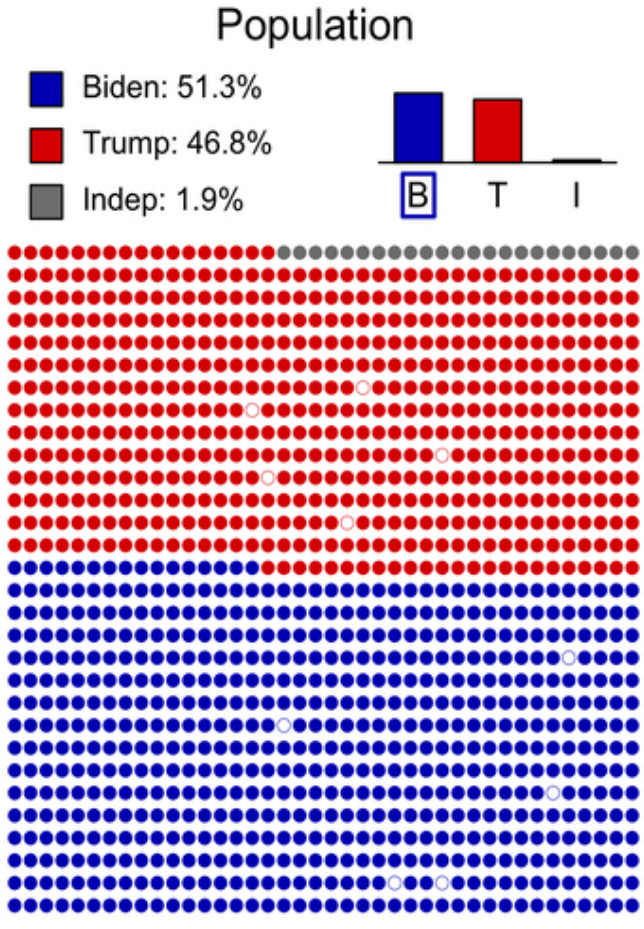
Population and Sample - Example

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

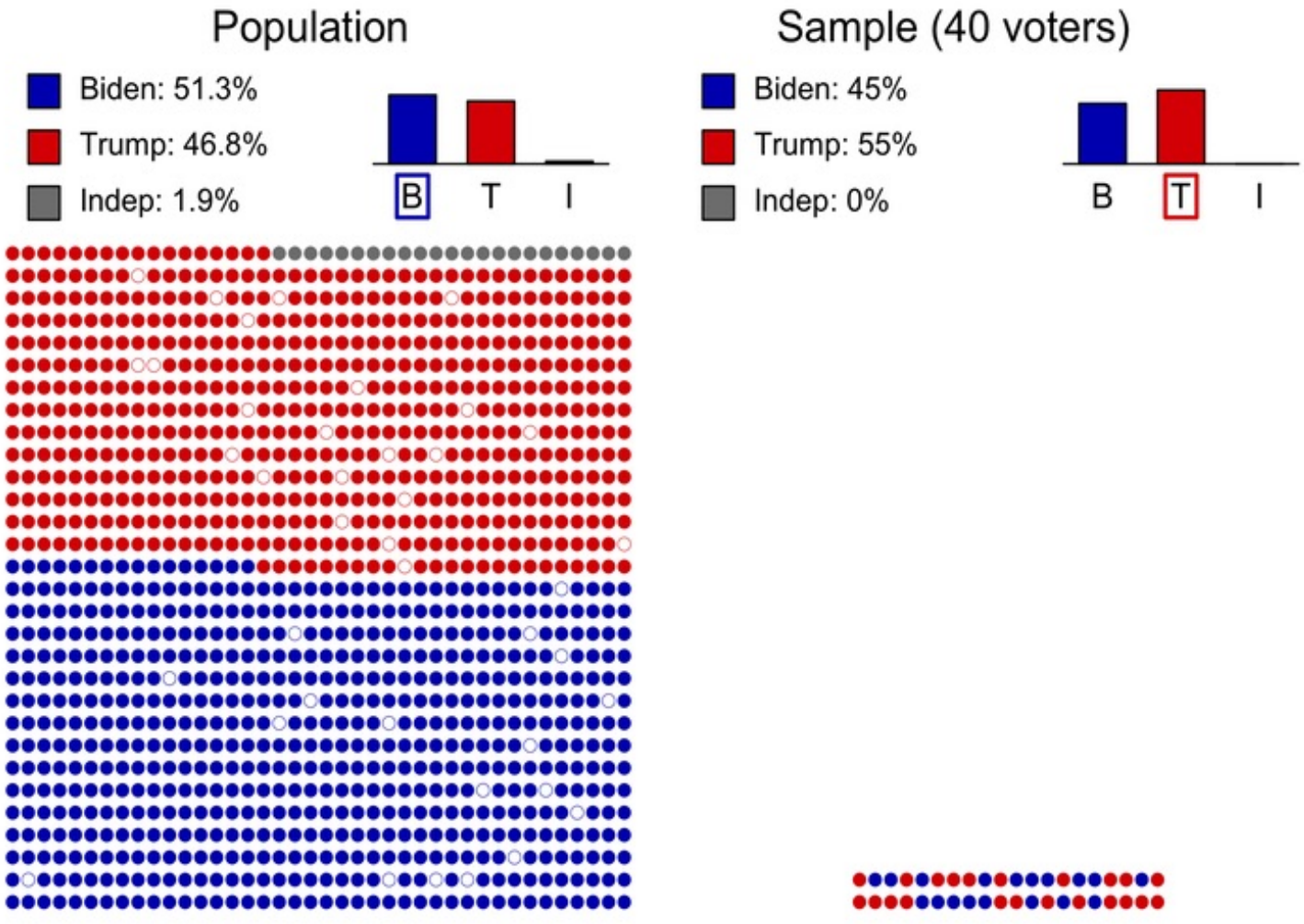
$$x_1 = 1, \ldots, x_5 = 1, x_6 = 0, \ldots, x_{10} = 0$$

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

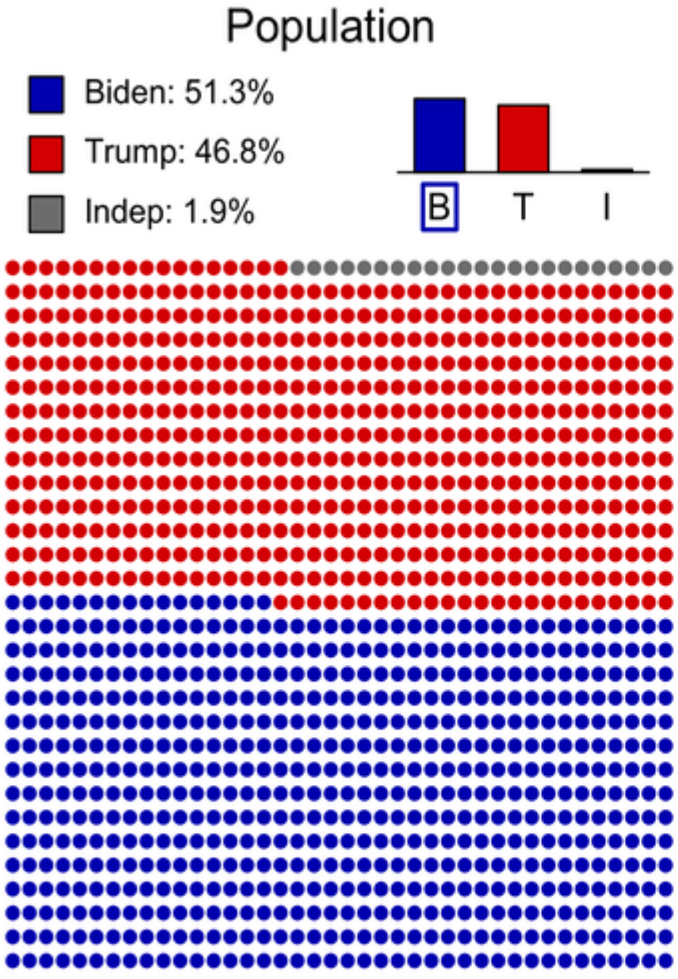
We can now compute our estimator on the observed data (i.e. the realizations) and we obtain $\hat{p} = 0.5$. Therefore, our best guess based on the available data is that 50% of the voters will vote for Biden. Unfortunately, this doesn't really help us. So let's try with a bigger sample size... say $n = 40$.



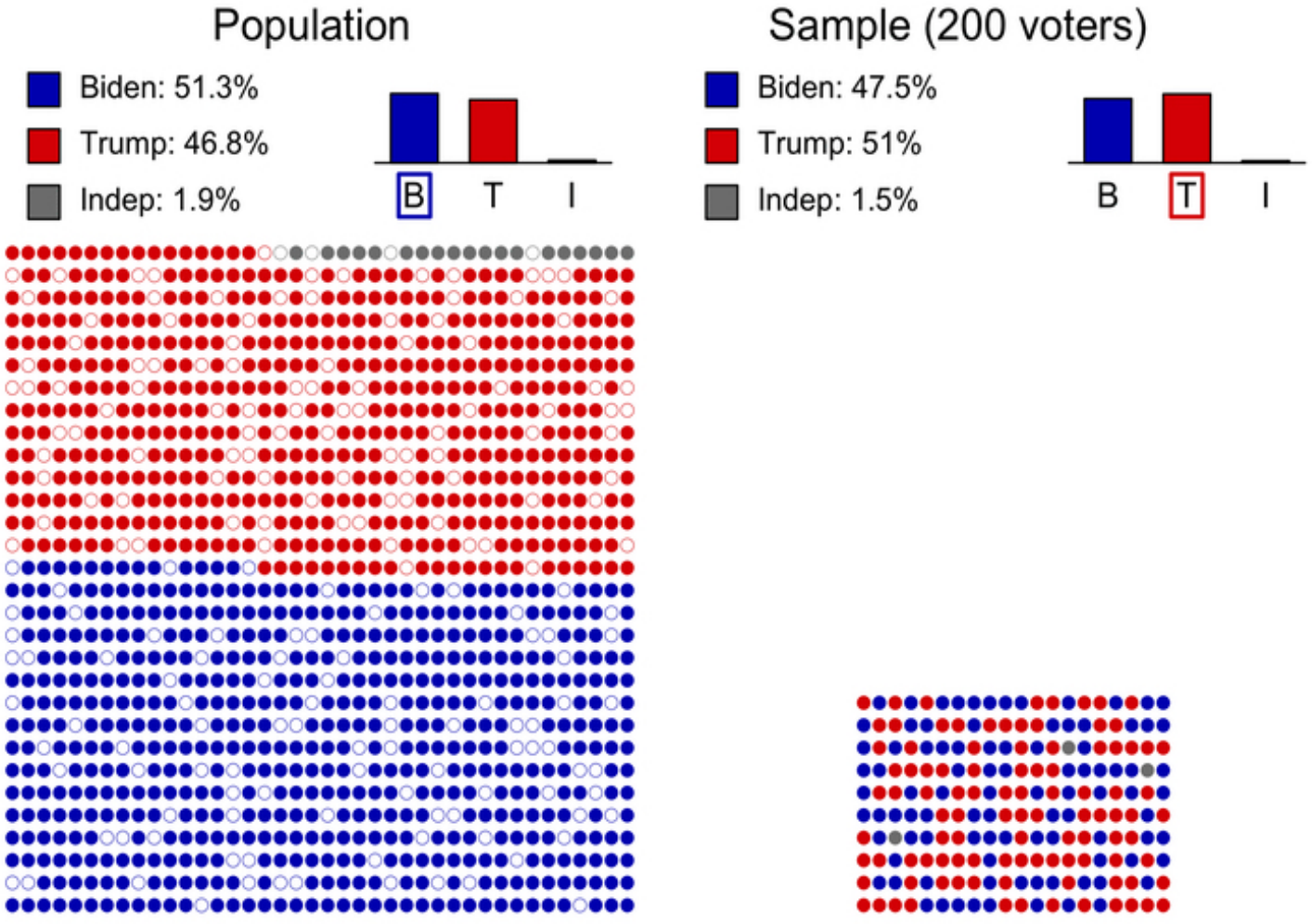
Population and Sample - Example



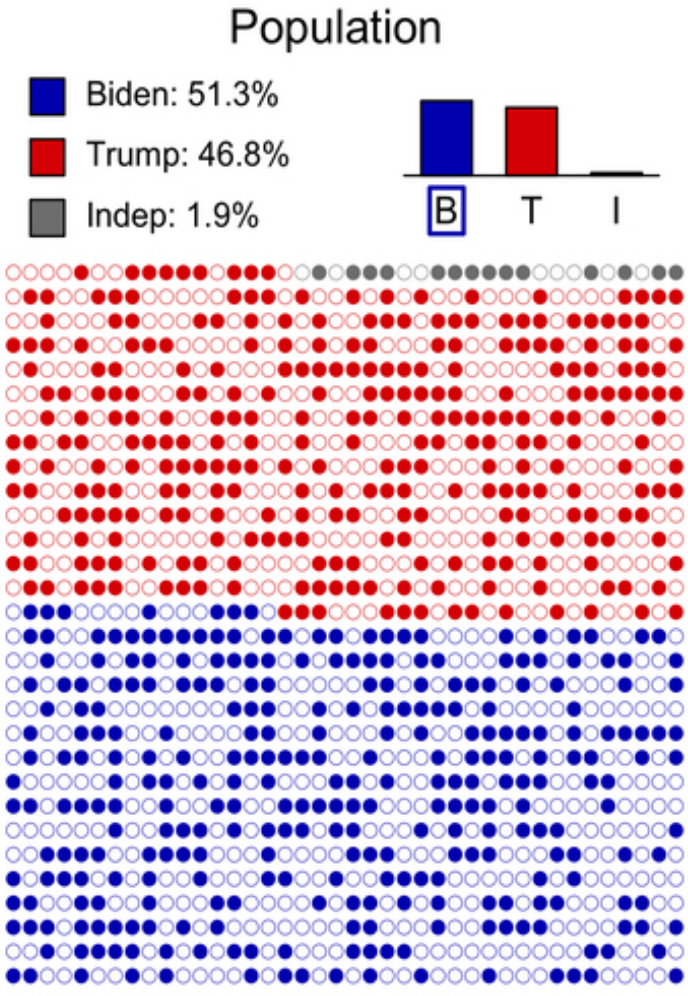
Population and Sample - Example

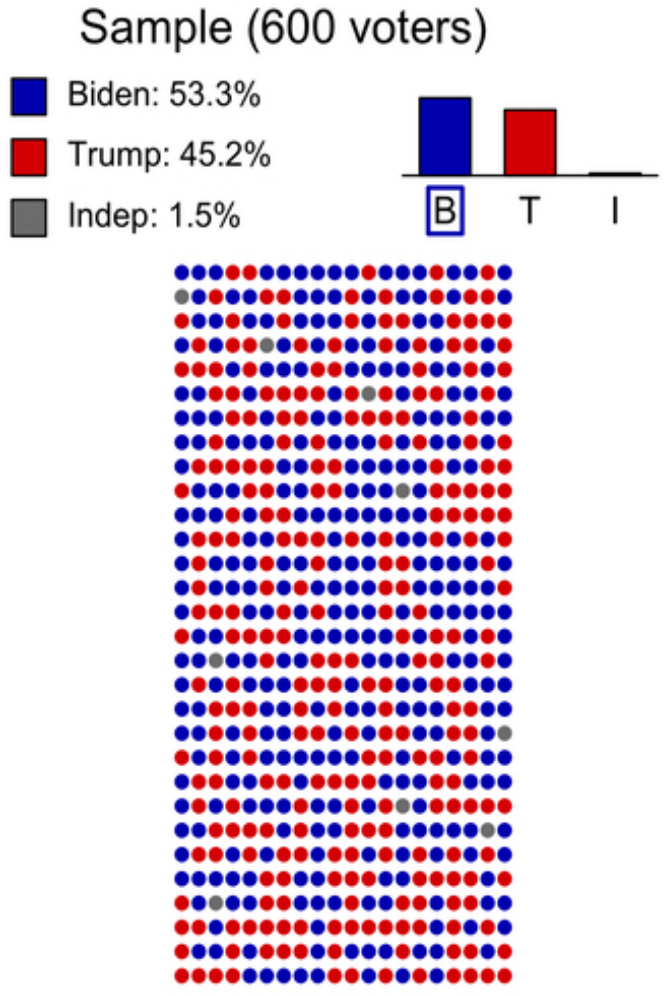


Population and Sample - Example



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

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

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

👋 Check out expected value and variance.

How to measure uncertainty?

In our example, we have

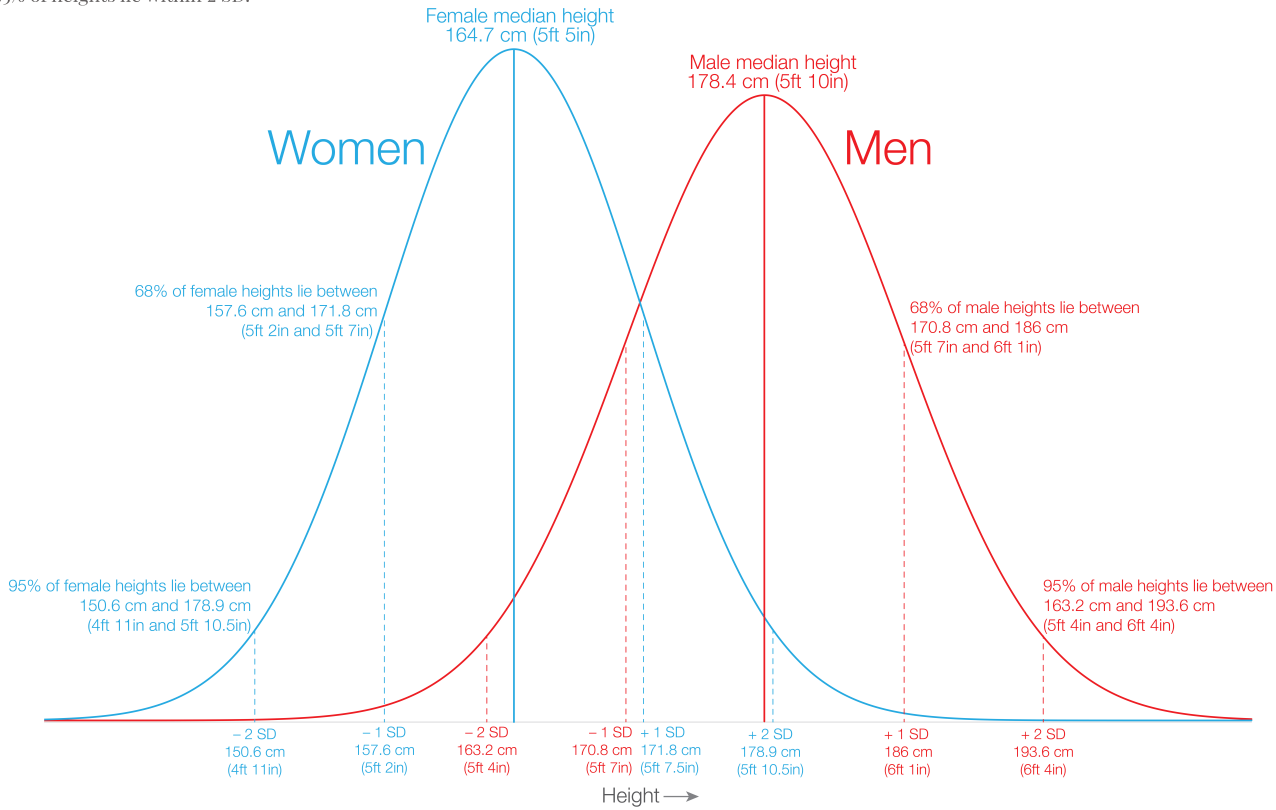
$$\hat{p} \dot{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

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

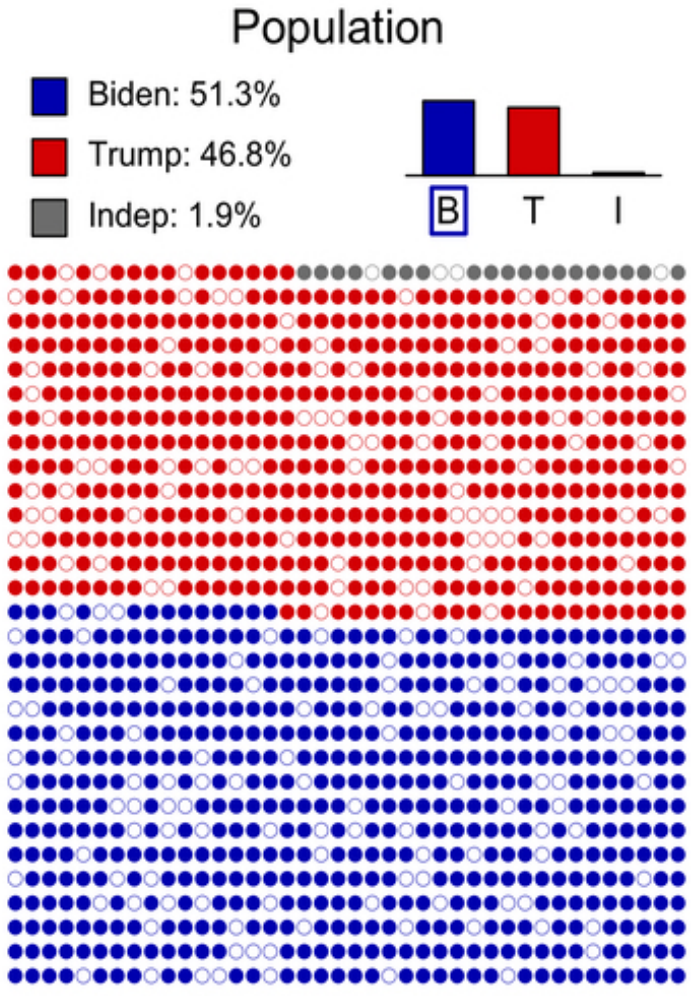
The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human 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.

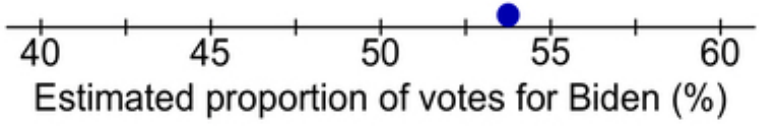


Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.

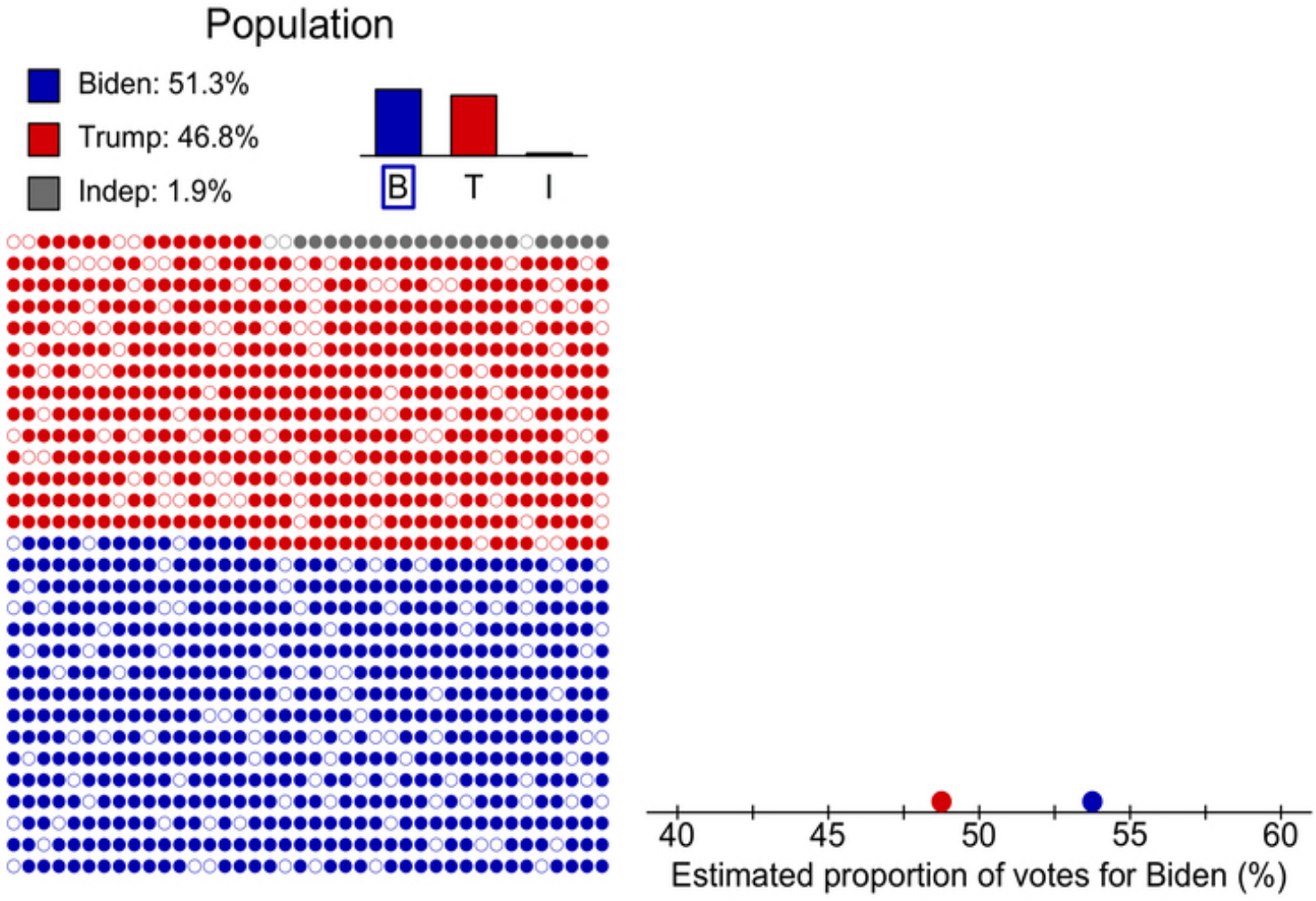
Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.

Central Limit Theorem - Example

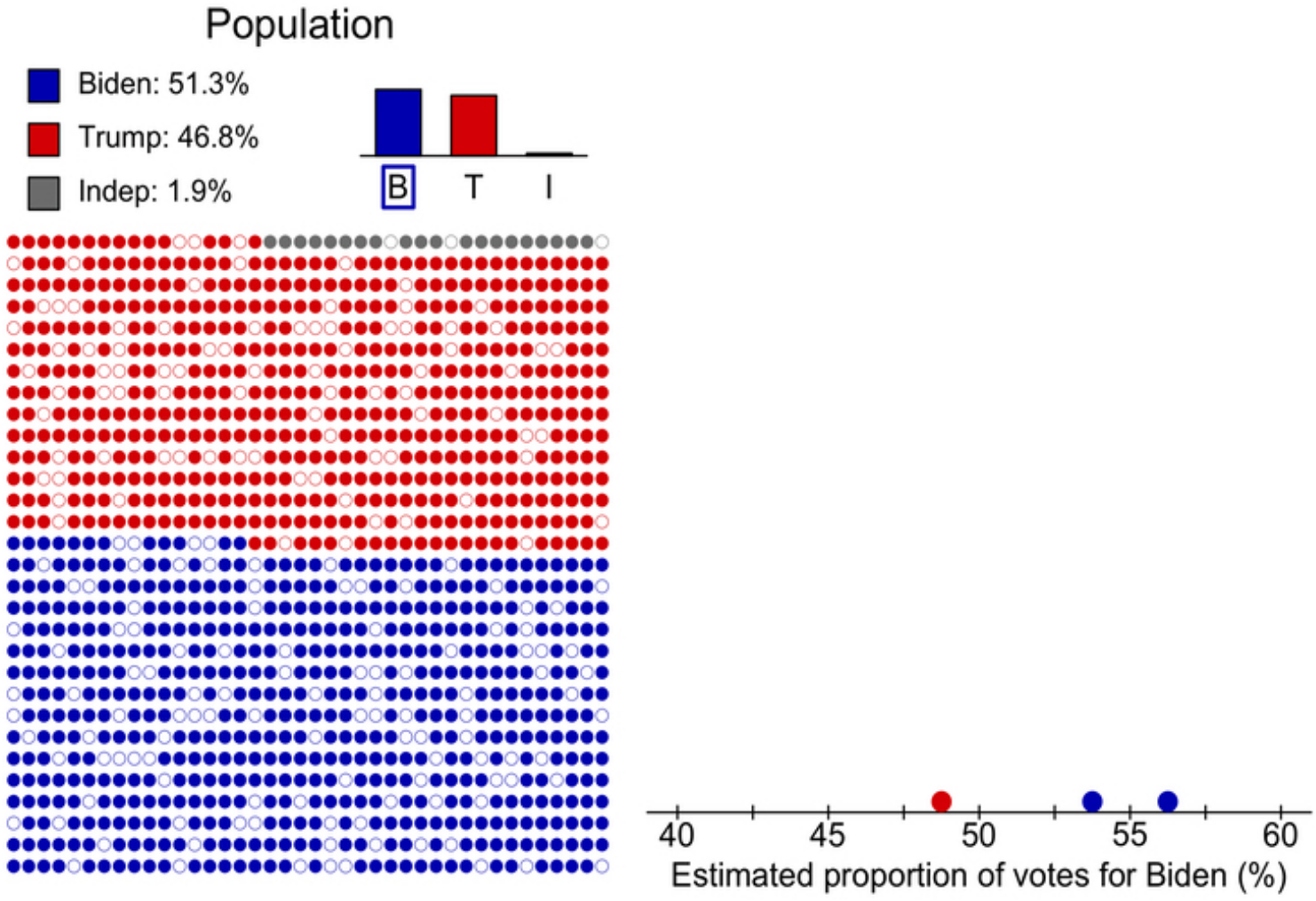




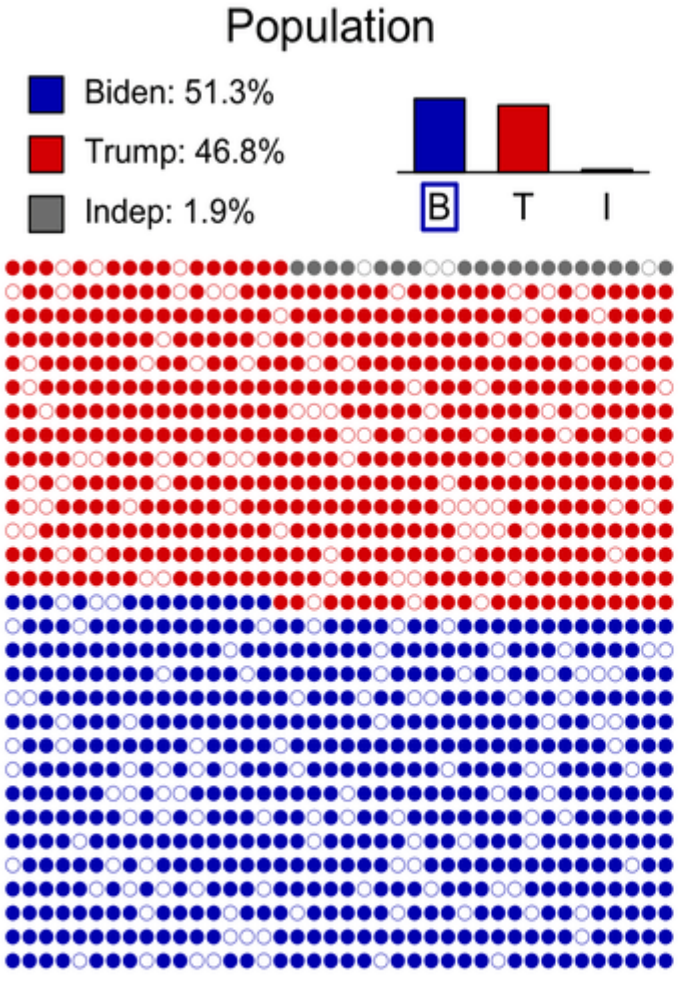
Central Limit Theorem - Example

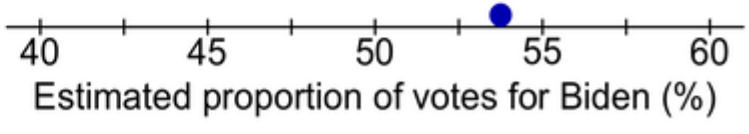


Central Limit Theorem - Example

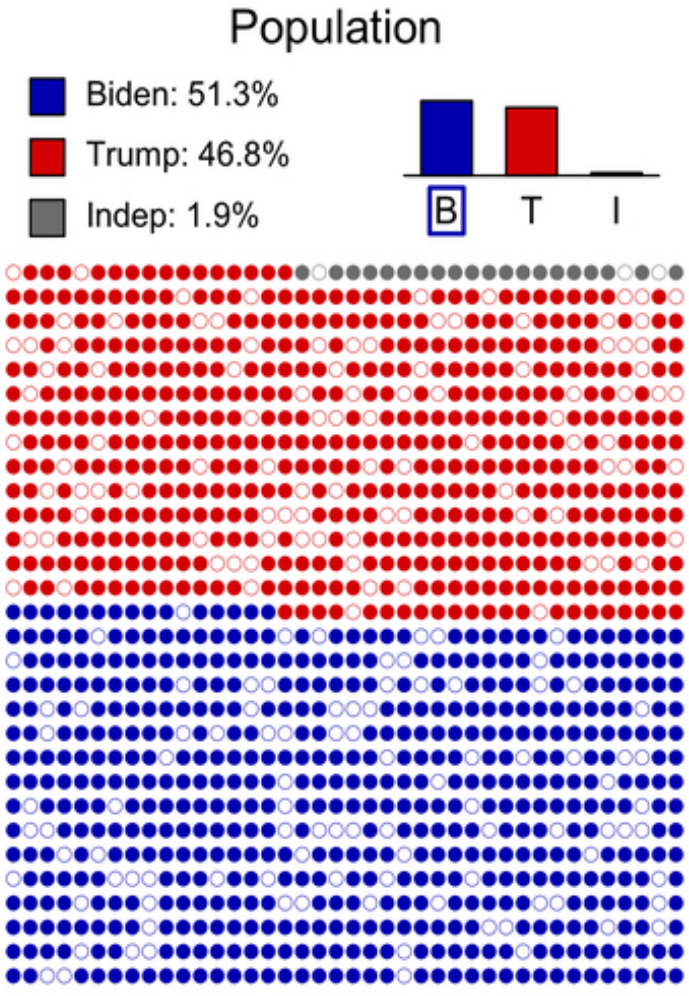


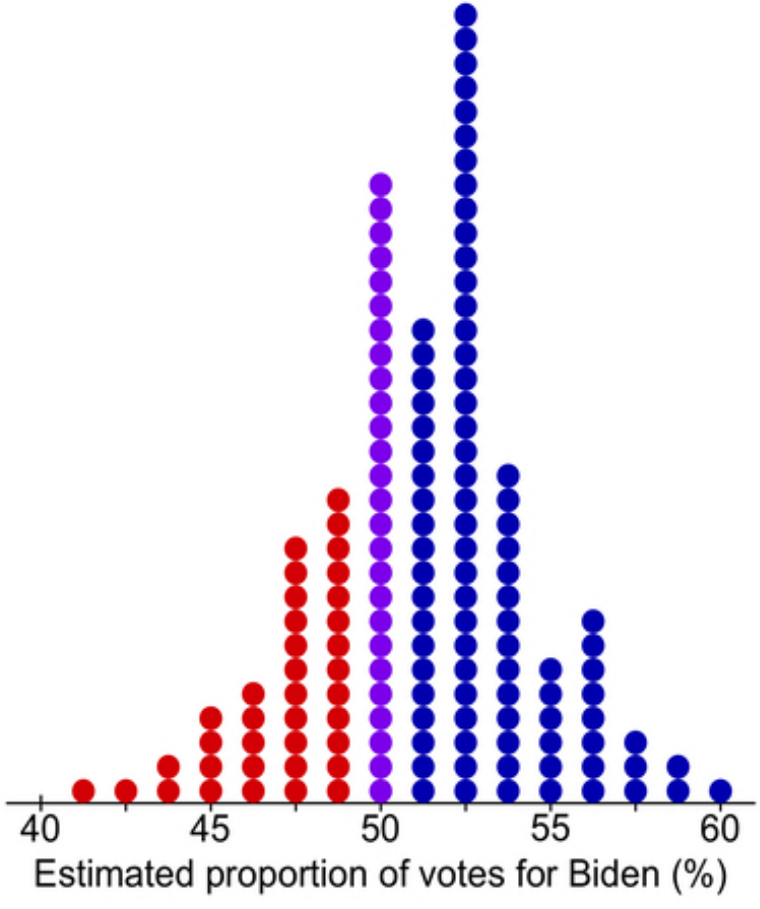
Central Limit Theorem - Example



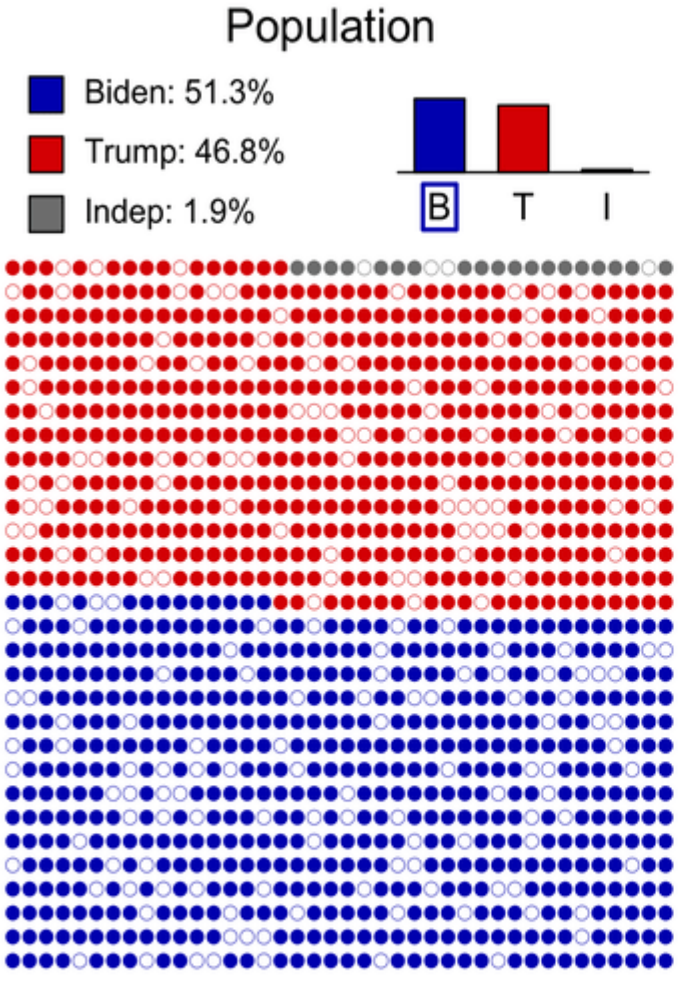


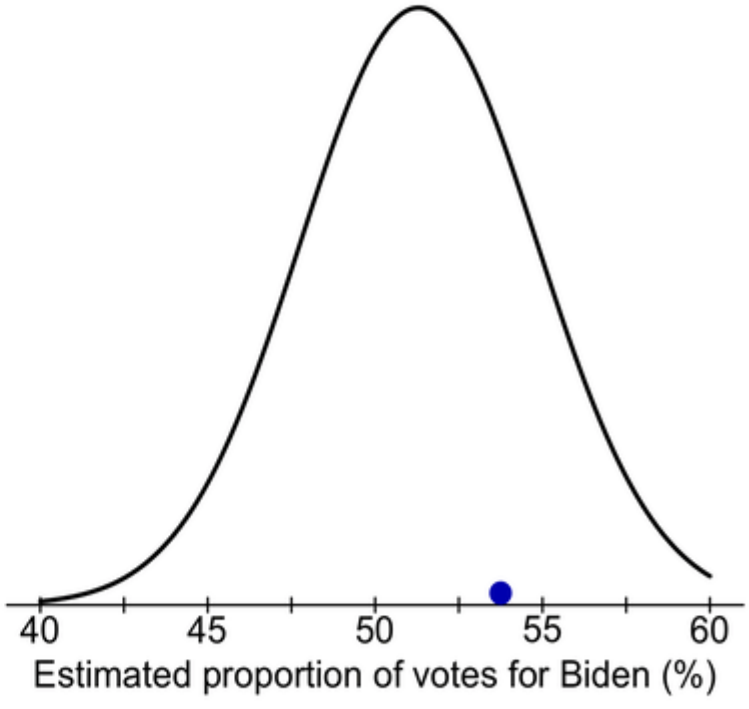
Central Limit Theorem - Example



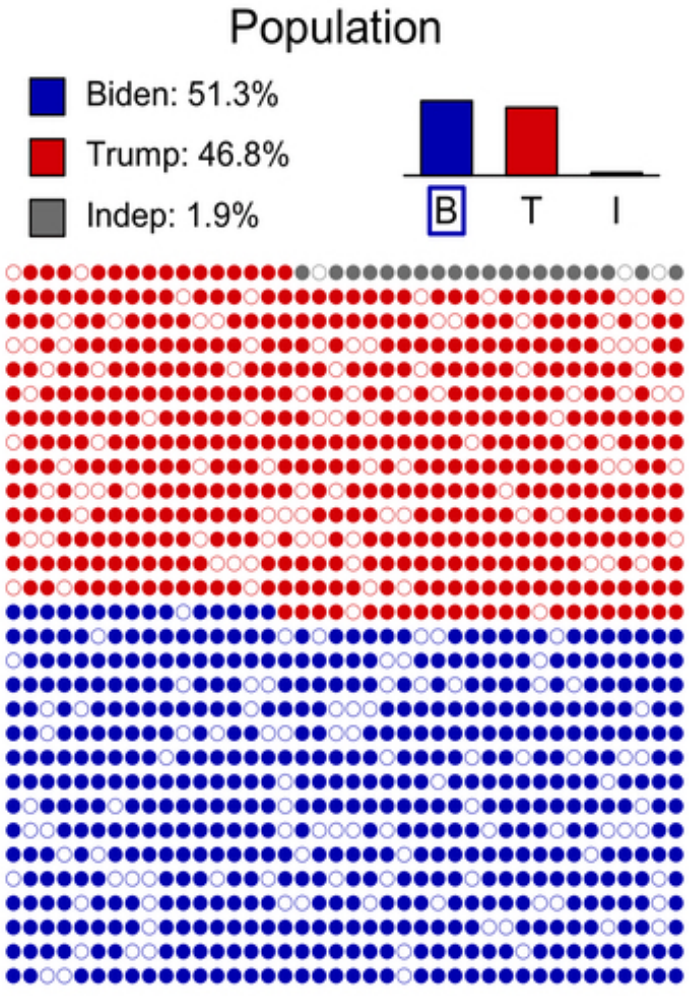


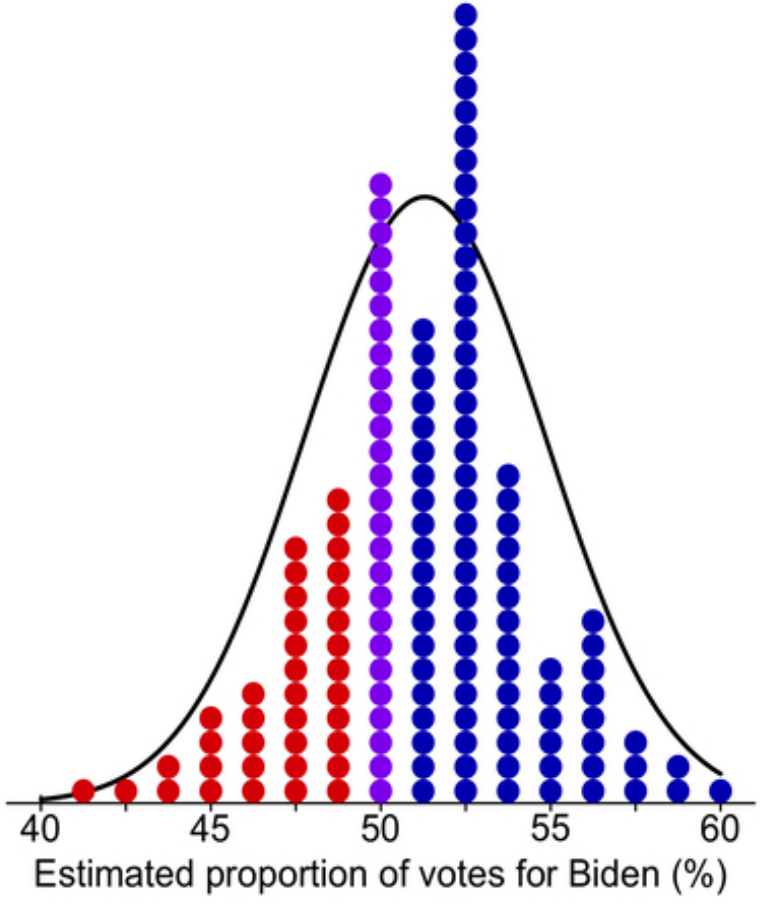
Central Limit Theorem - Example





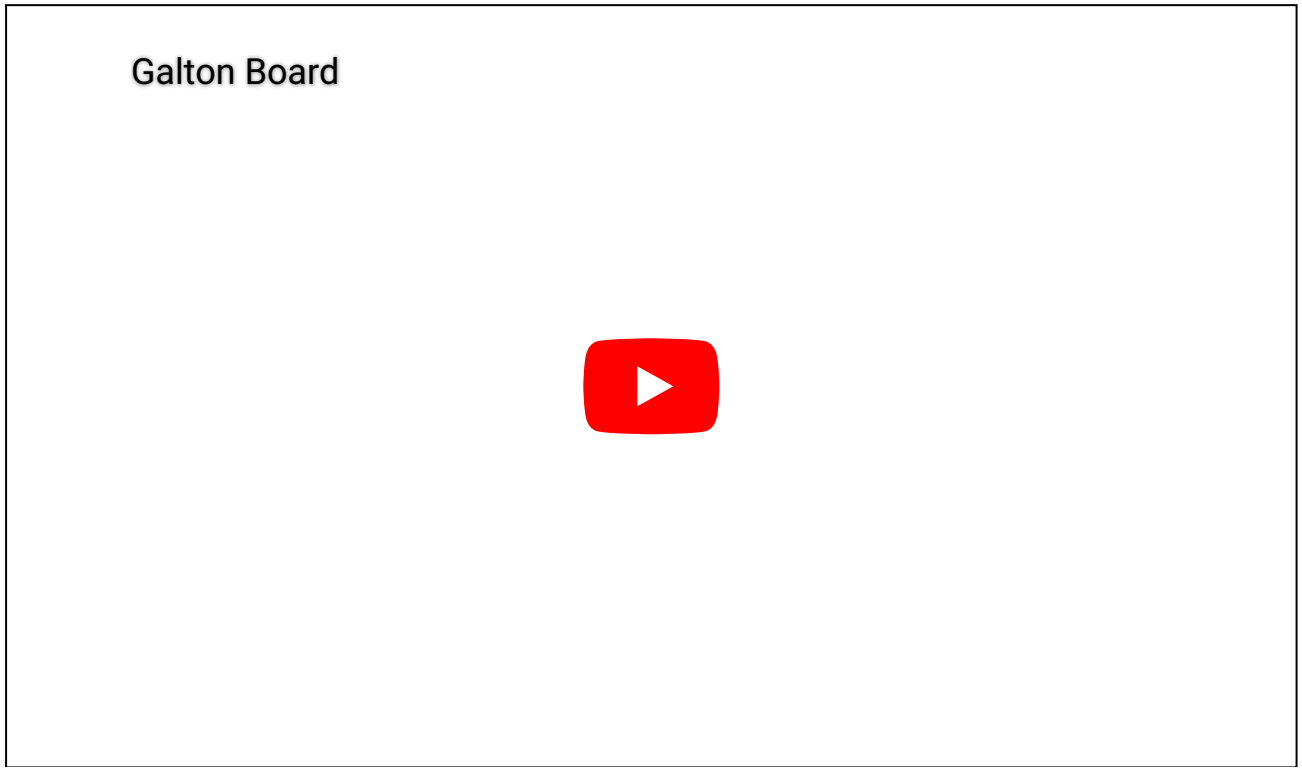
Central Limit Theorem - Example





Central Limit Theorem - Example

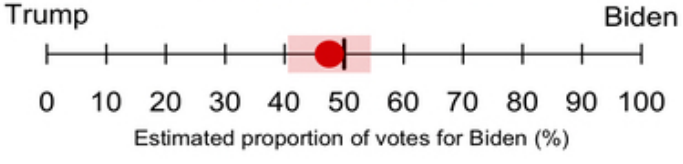
People have used the Galton Board as a practical device to demonstrate the CLT, in particular that with sufficient sample size the binomial distribution approximates a normal distribution.



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? Using the CLT but replacing the unknown p with $\hat{p}$, we can write

$$\hat{p} \dot{\sim} \mathcal{N}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right).$$

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

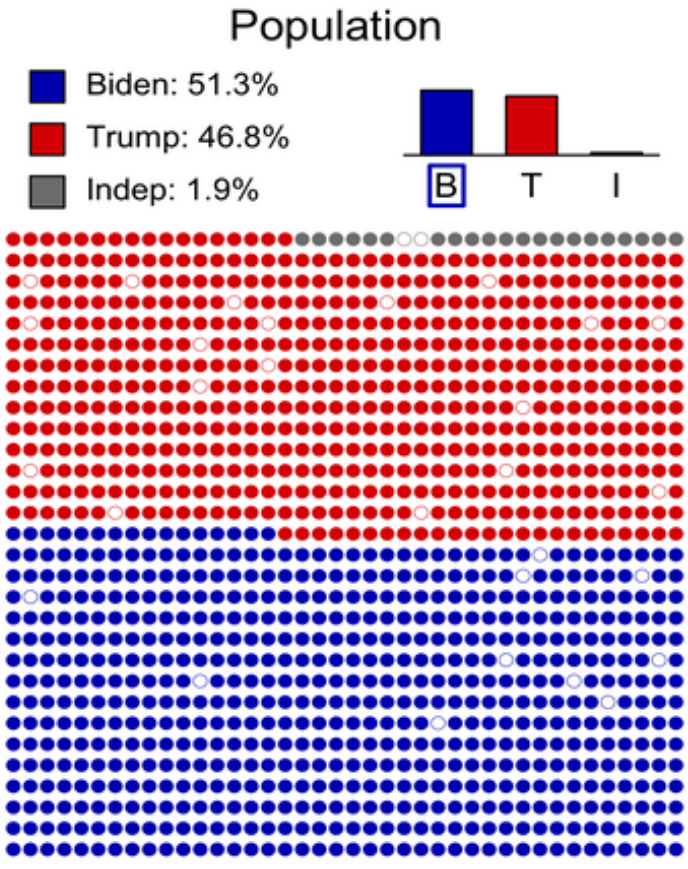
$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

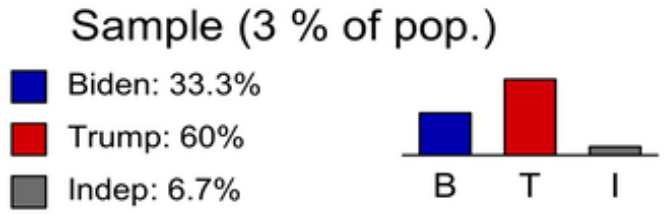
where $Z_{1-\alpha/2}$ corresponds to the $1-\alpha/2$ quantile of a $\mathcal{N}(0,1)$. For example, $Z_{0.975} \approx 1.96$ for 95% confidence interval. Thus, we have

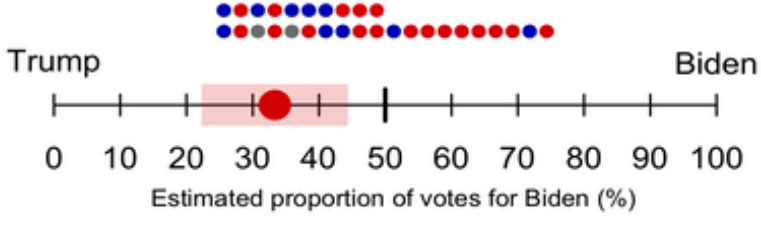
$$\Pr\left(p \in \hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha.$$

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

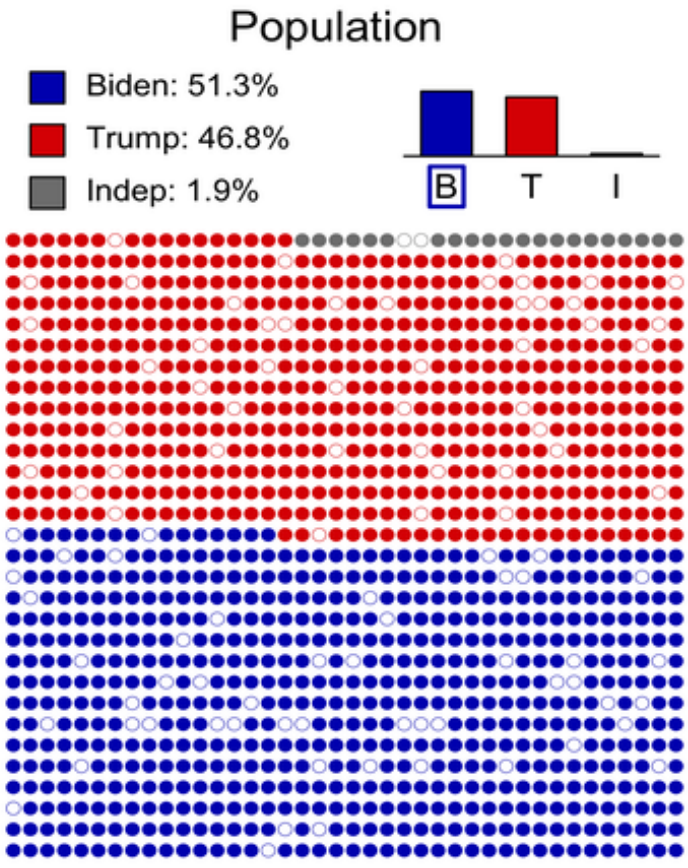
Confidence Interval with $\alpha = 20\%$

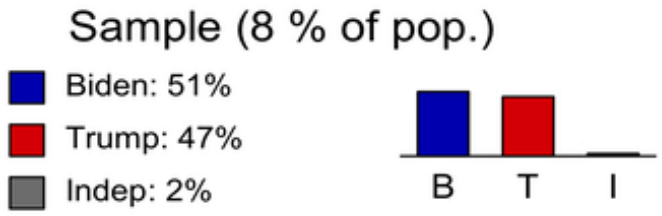


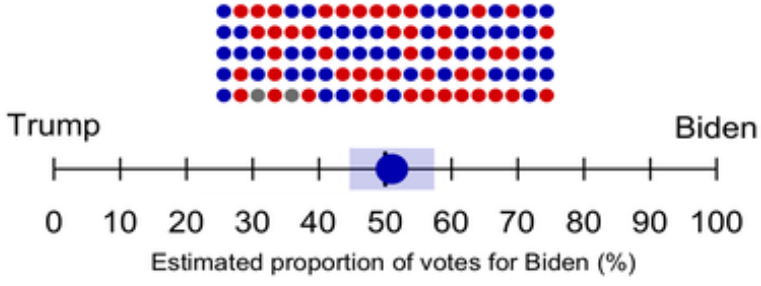




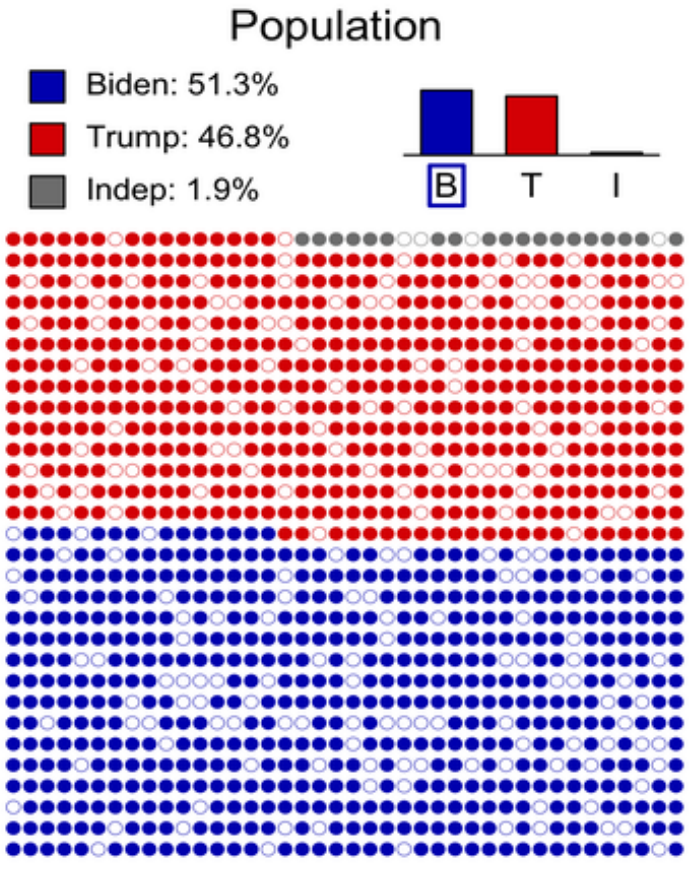
Confidence Interval with $\alpha = 20\%$

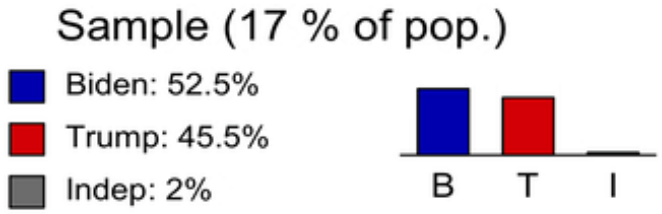


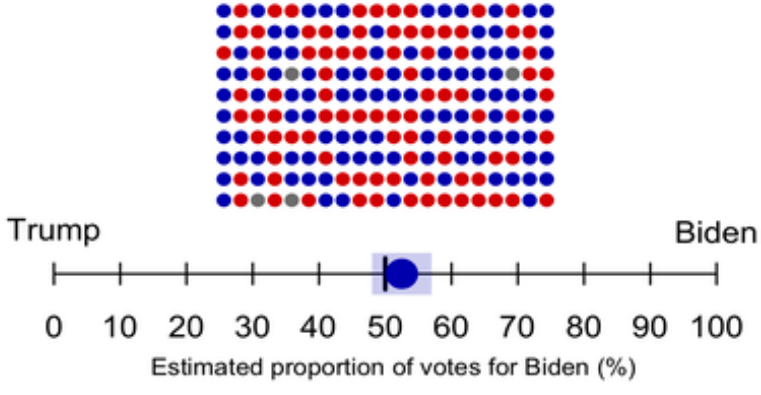




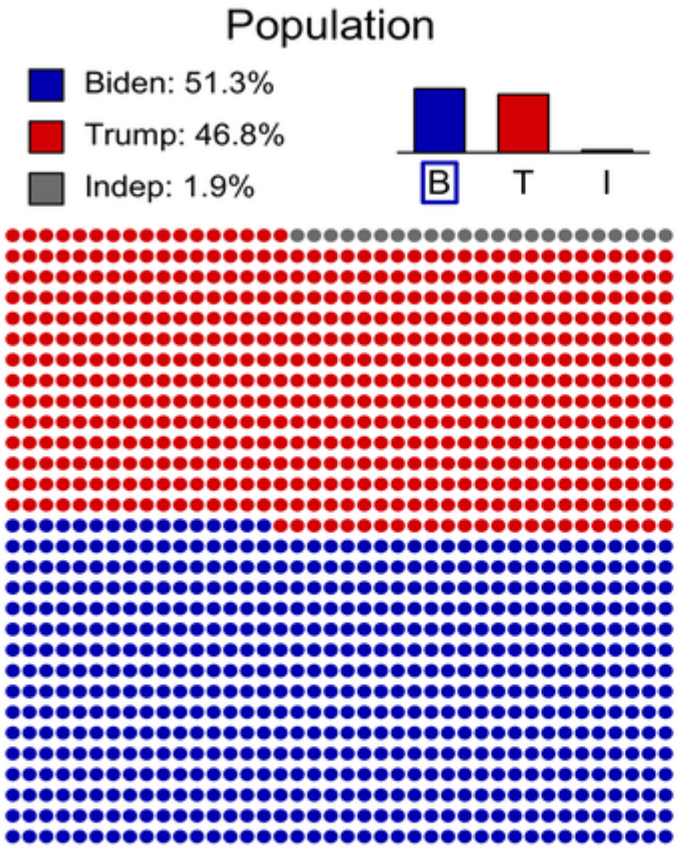
Confidence Interval with $\alpha = 20\%$



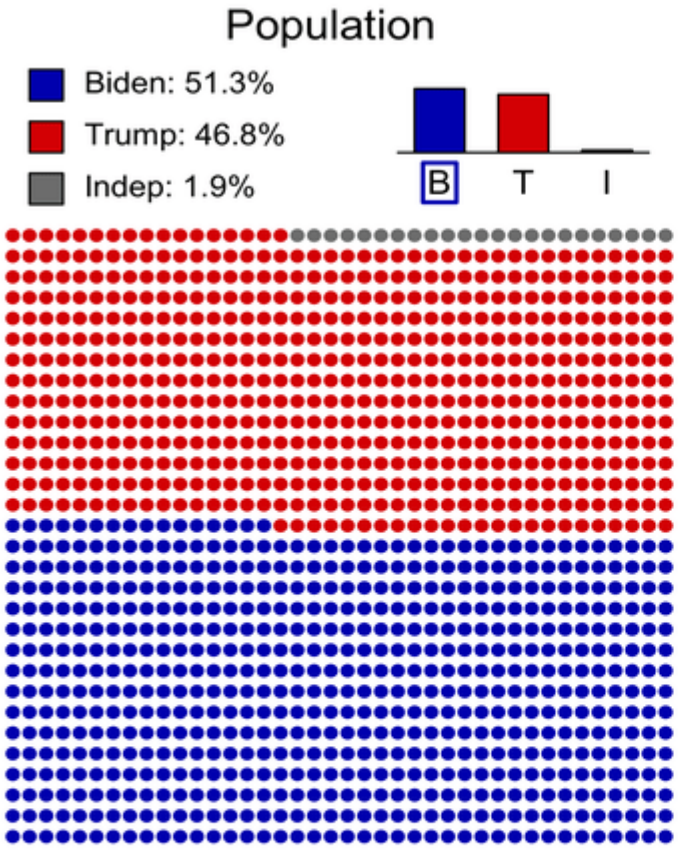




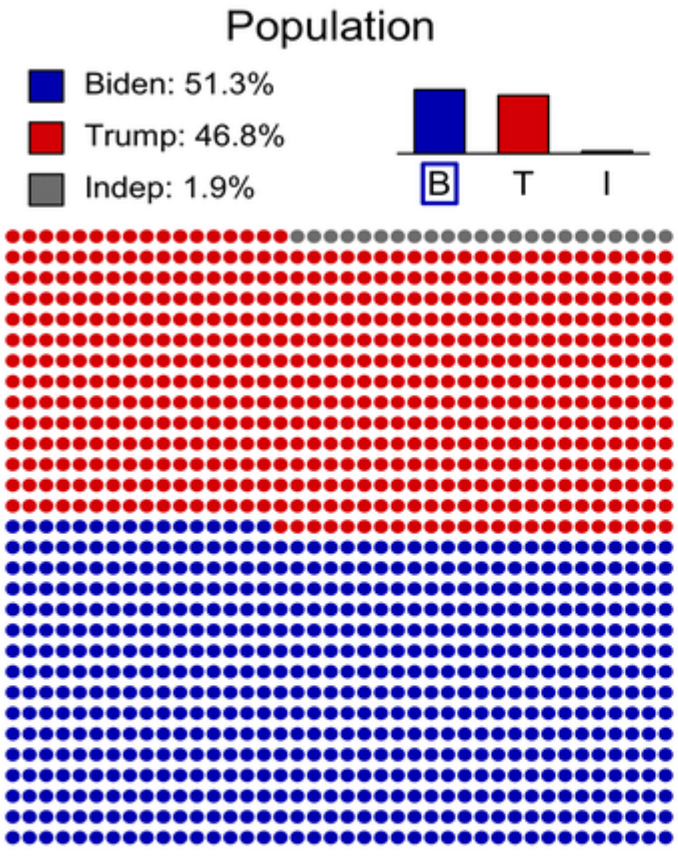
Confidence Intervals with $\alpha = 20\%$



Why is 20%? Why not 5%?



What if I want to be certain? 😳

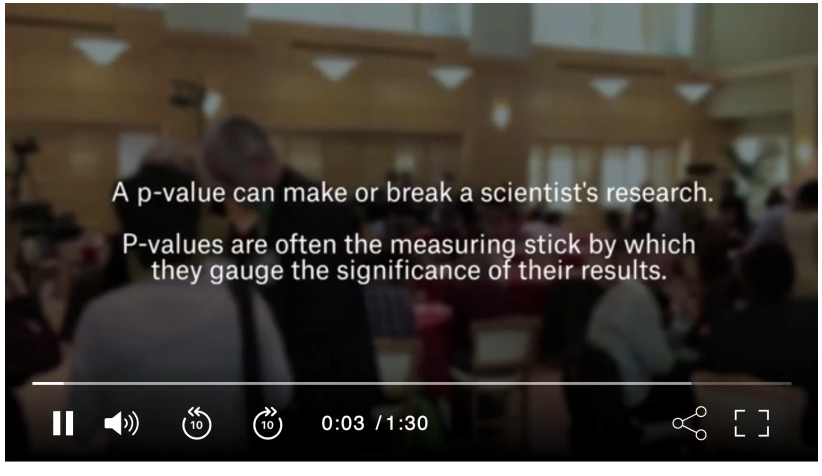


⚠️ Take home message

- 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 - α (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).

How to test a (scientific) hypothesis?

- An alternative summary measure of uncertainty is provided by **p-values** that take values between 0% and 100%. P-values are generally preferred than confidence intervals when we want to assess the validity of a statement.
- 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.

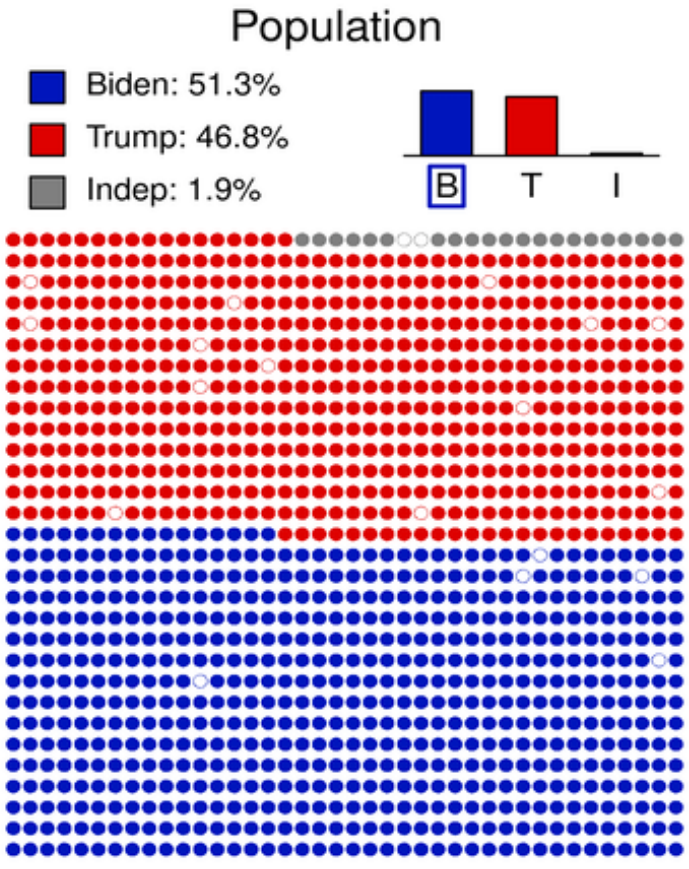
How to test a (scientific) hypothesis?

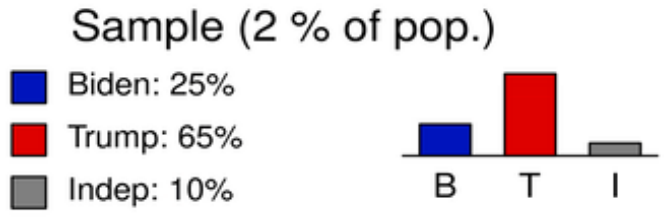
- A p-value is associated to a (couple of) hypotheses about the phenomenon under investigation. For example, for the United States presidential election example, one hypothesis is that the majority votes for Biden's team, and the other one is that there is no majority (the YES and NO are both at 50%)
- 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 (i.e. the most precise hypothes is that the votes are 50%-50%.)
- 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 (e.g. Biden's team receives more than 50% of the votes).
- 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.

Hypothesis testing

- 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 (sample) 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 again this depends on the context.

Hypothesis testing





$\alpha = 5\%$

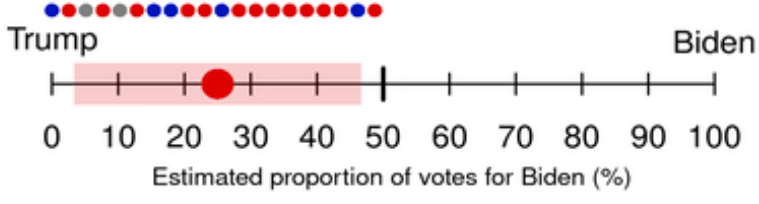
$H_0 : p_{Biden} = 50\%$

$H_1 : p_{Biden} > 50\%$

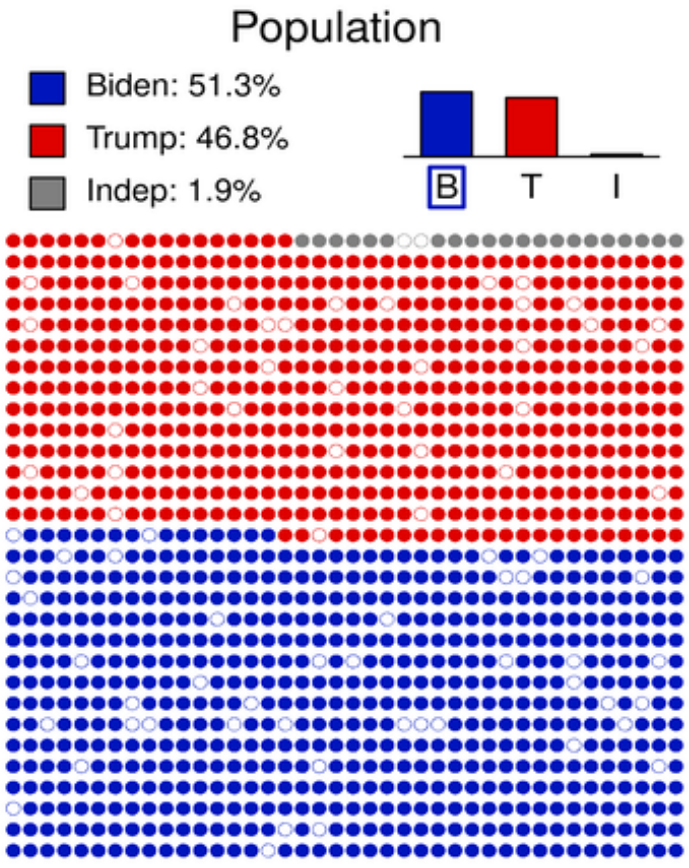
P-value = 99.41%

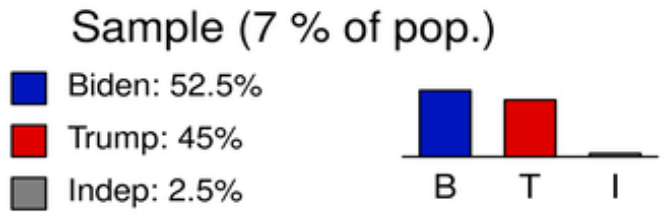
P-value > α

We cannot reject that p_{Biden} is equal to 50%



Hypothesis testing





$\alpha = 5\%$

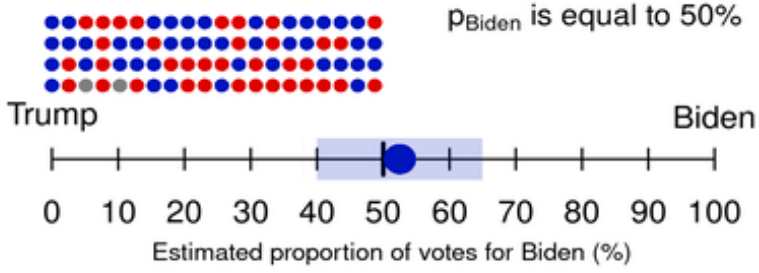
$H_0 : p_{Biden} = 50\%$

$H_1 : p_{Biden} > 50\%$

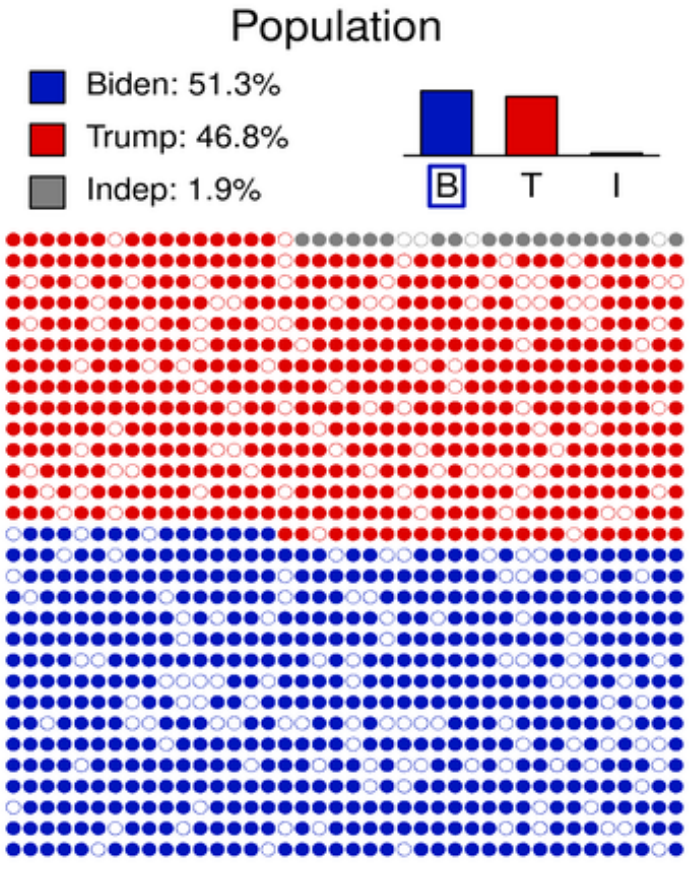
P-value = 36.88%

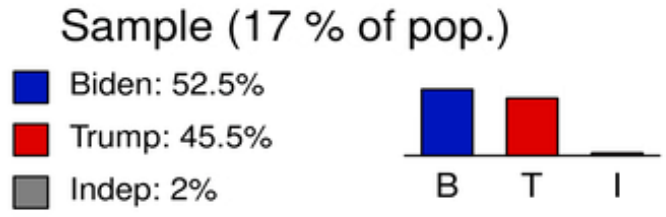
P-value > α

We cannot reject that p_{Biden} is equal to 50%



Hypothesis testing





$\alpha = 5\%$

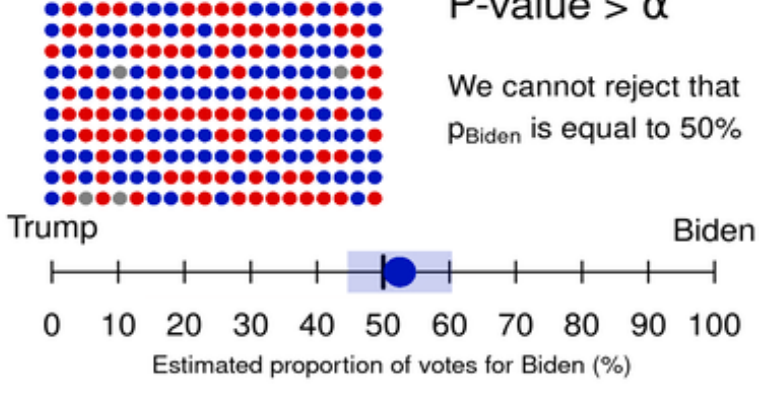
$H_0 : p_{Biden} = 50\%$

$H_1 : p_{Biden} > 50\%$

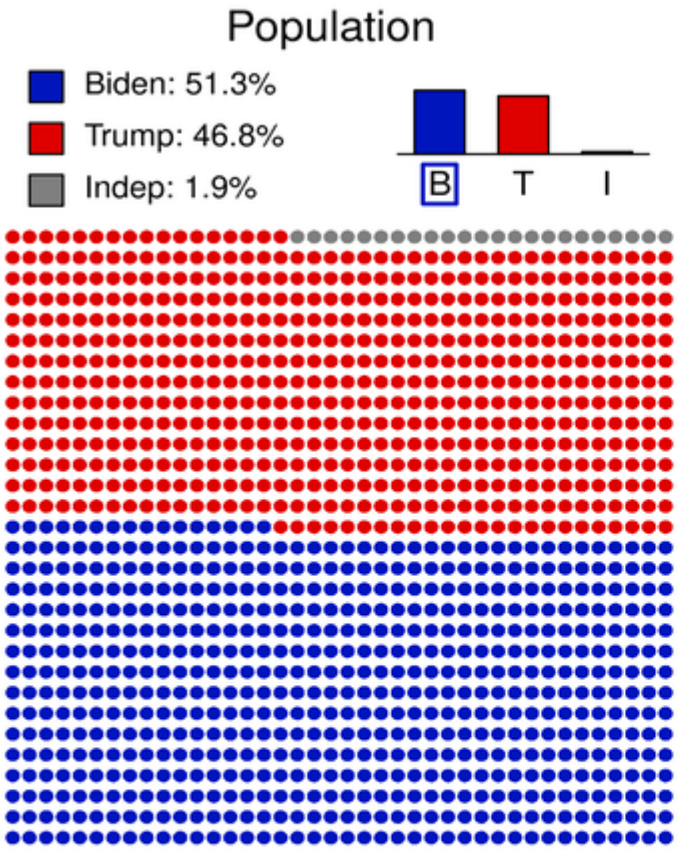
P-value = 26.23%

P-value > α

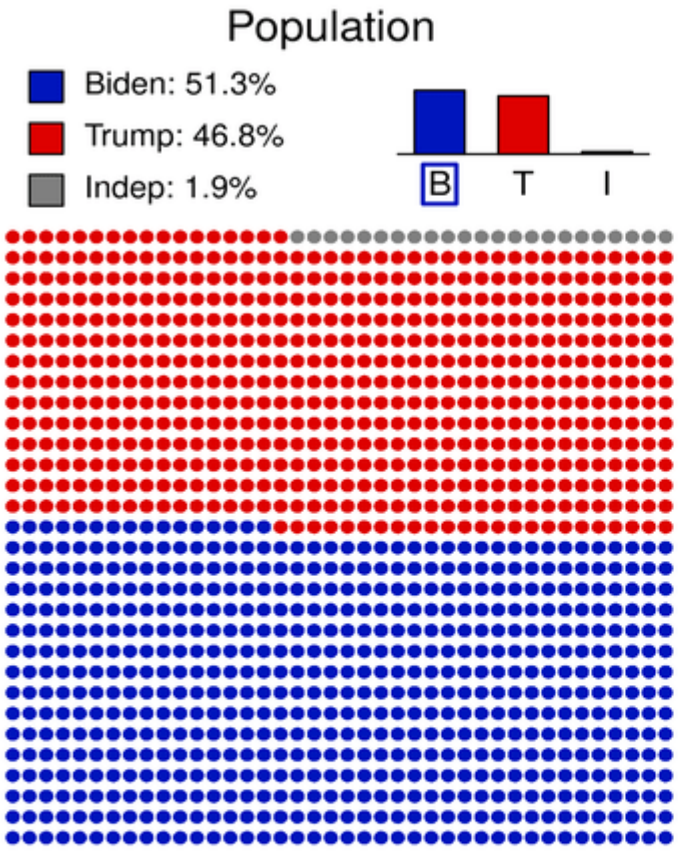
We cannot reject that p_{Biden} is equal to 50%



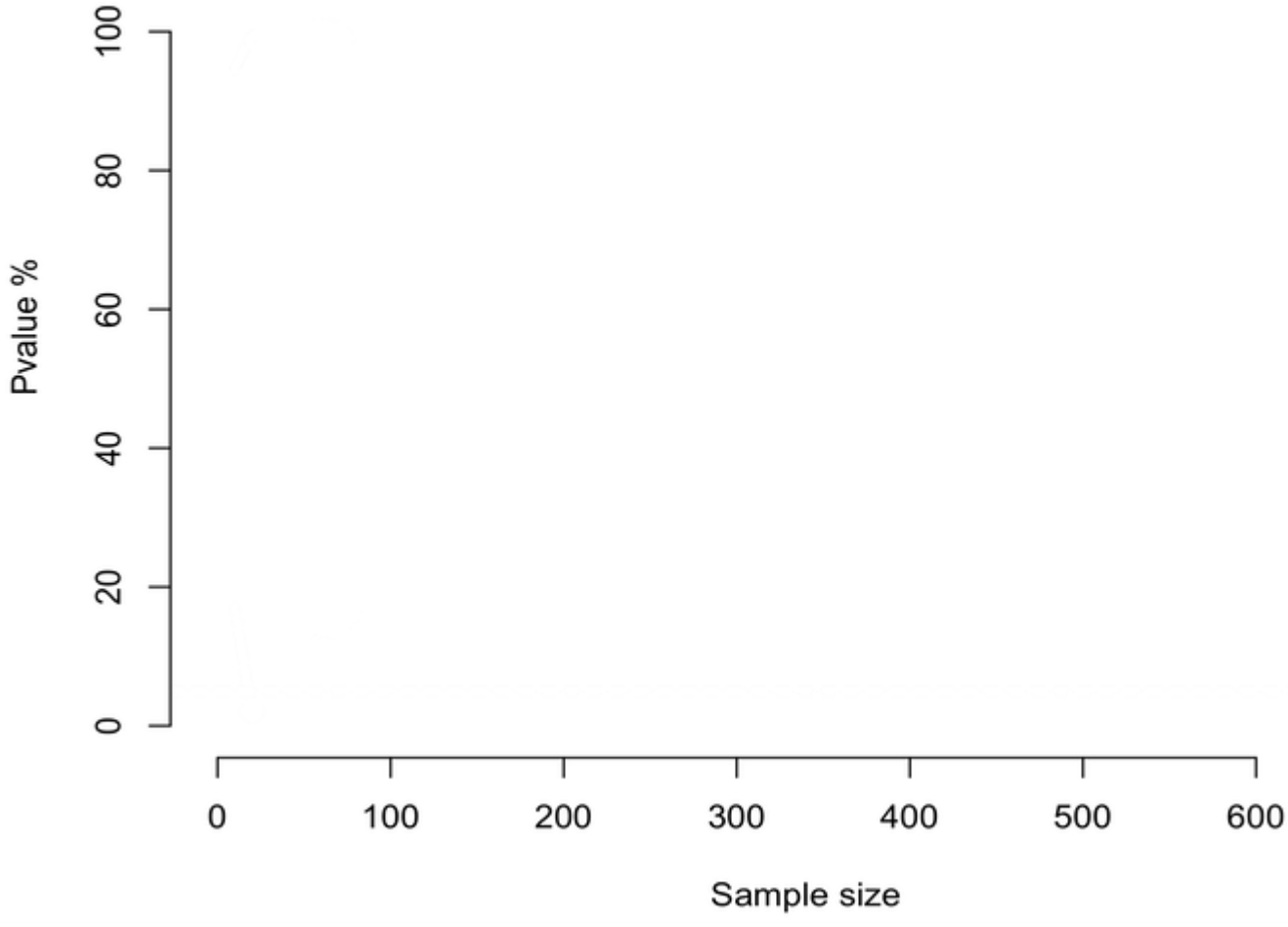
Hypothesis testing



Hypothesis testing



Hypothesis testing

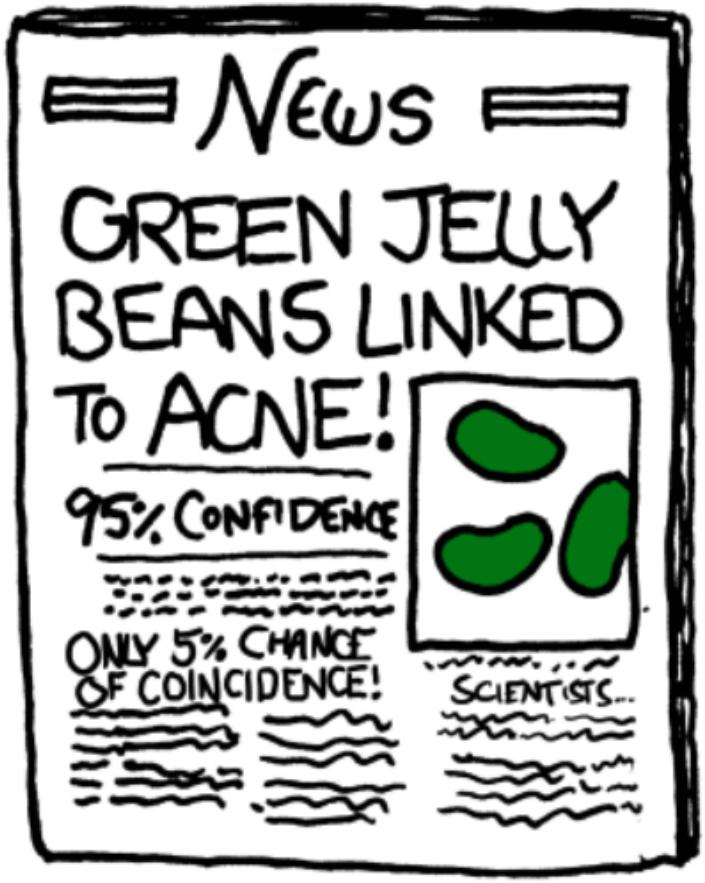


What to remember about p-values

- A p-value is something **random** (i.e. a random variable), since its value depends on the data. Likewise, confidence intervals are random, since their bounds depend on the data. In fact, p-value is highly related to the width of the corresponding confidence interval. The narrower the confidence interval, the more likely that the true value is not included in the interval, and thus the smaller the p-value.
- Hence it might happen that, while the null hypothesis cannot be rejected (supposing it is true), the p-value is smaller than the set threshold. With the latter chosen as 5% then, on average, the (sample) p-value is below 5% one out of twenty times!
- Hence **performing multiple tests, with the same or different data, is dangerous** ⚠️ (but very common! 😟) as it automatically leads to significant results, when actually there are none!
- Confidence intervals and p-values can be used for many types of models, such as regression, correlation, and so on. Therefore, they can be used to assess the validity of most (scientific) hypotheses (as discussed in your statistics classes! 🤓).

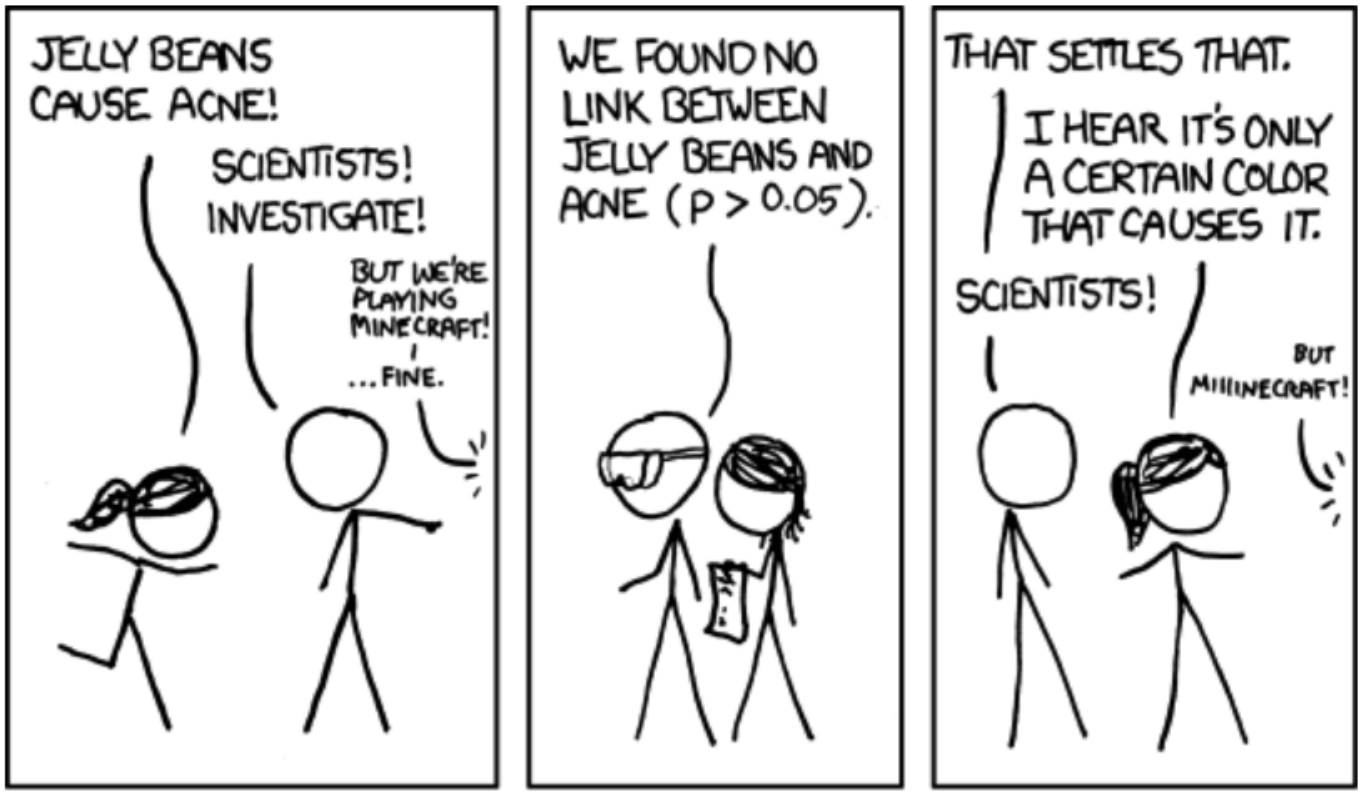
Why are there contradictory studies?

Are jelly beans causing acne? Maybe... but why only green ones? 😐



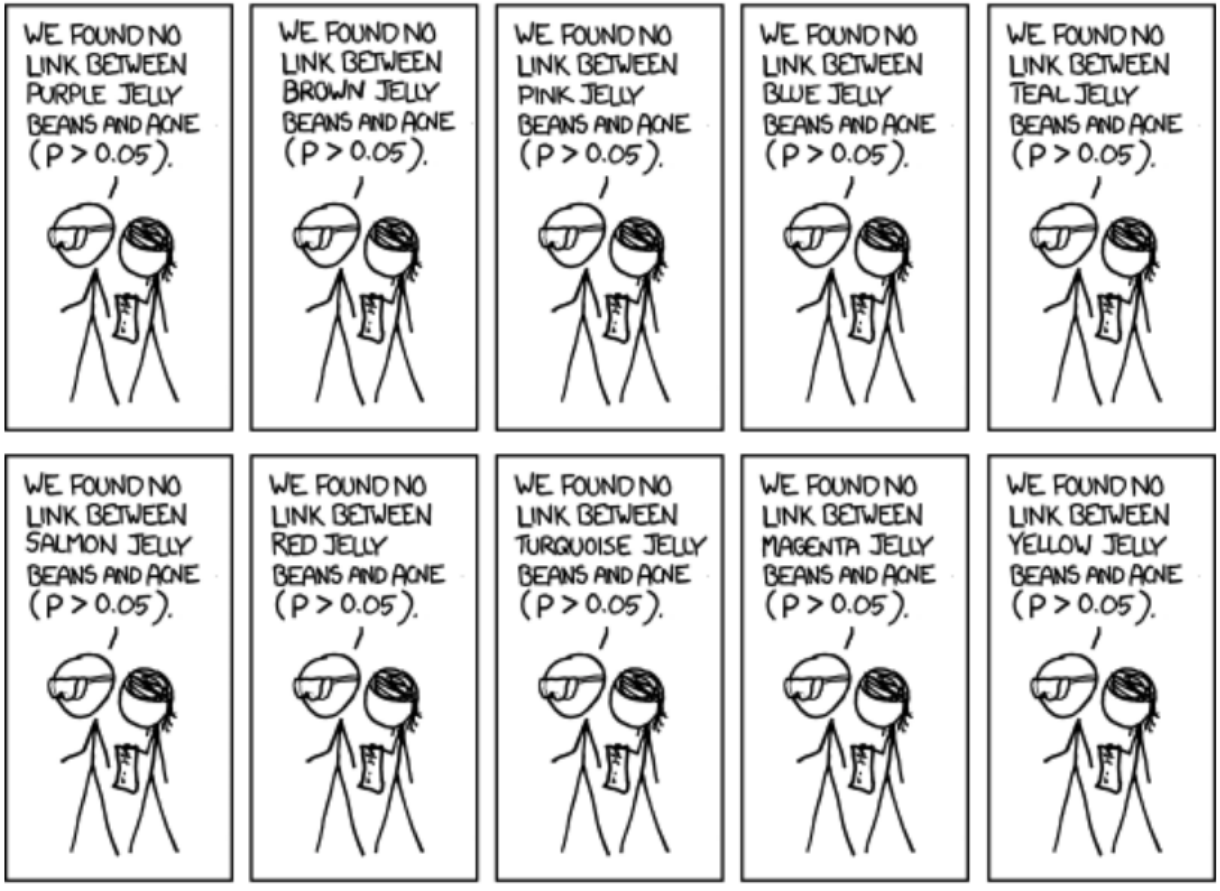
Source: xkcd

Are jelly beans causing acne?



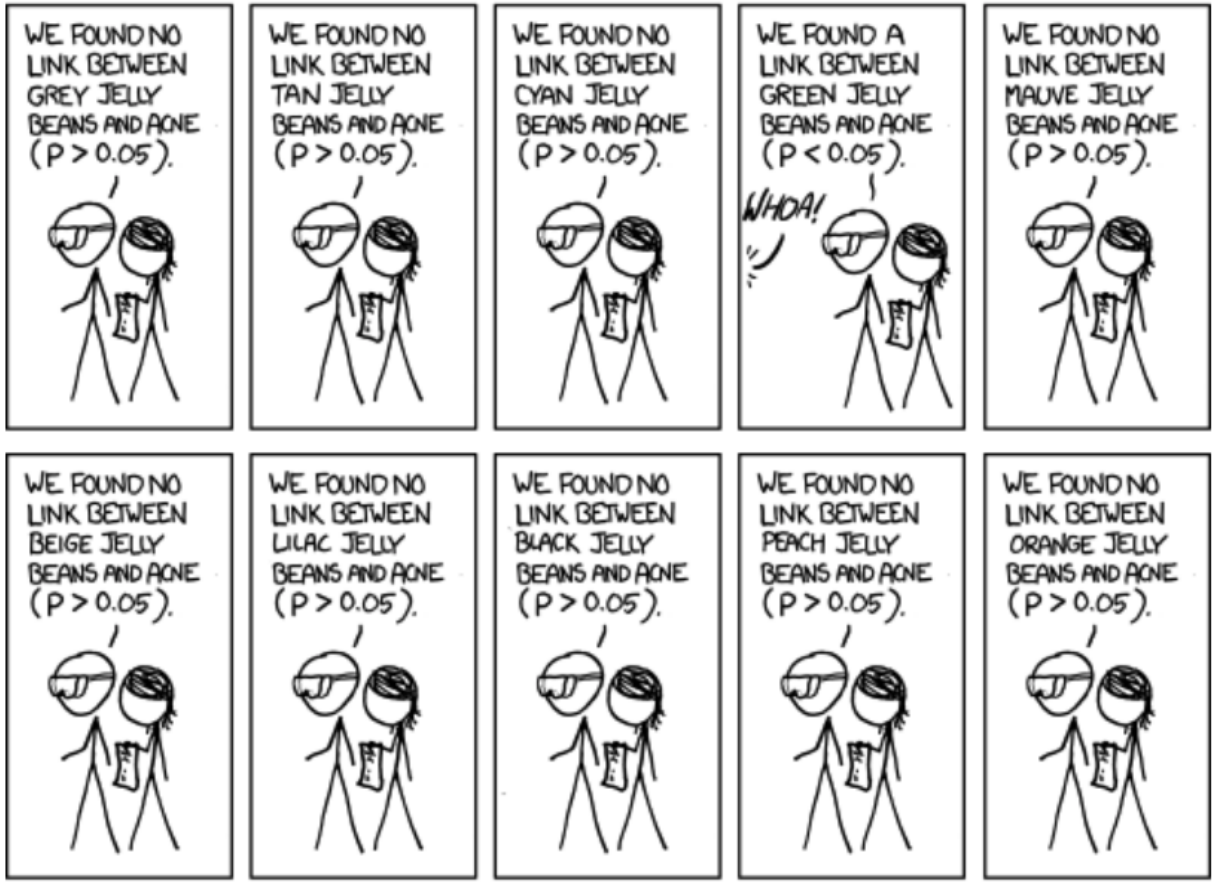
Source: xkcd

Maybe a specific color?



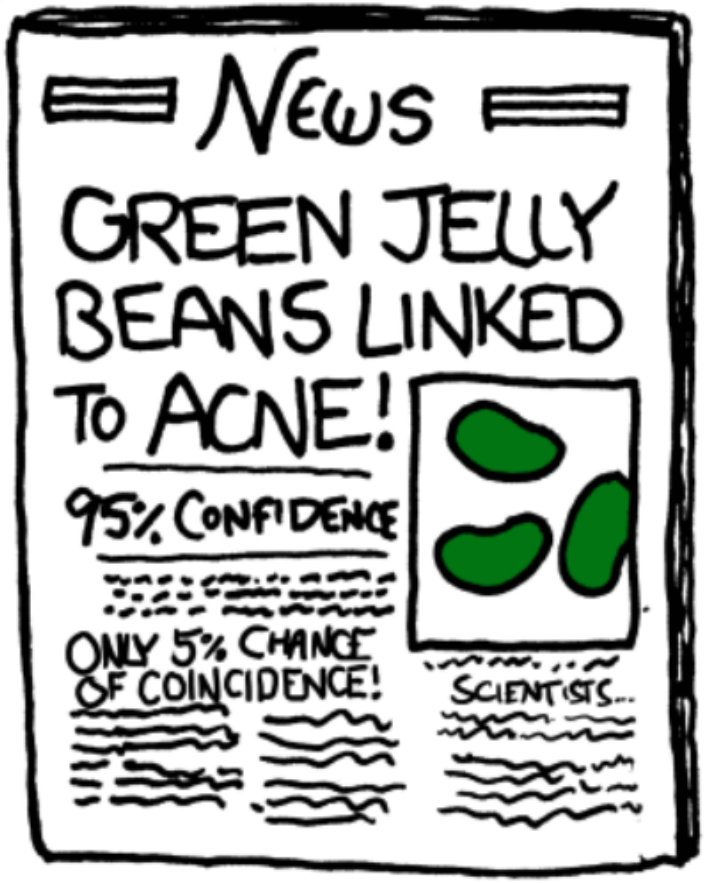
Source: xkcd

Maybe a specific color?



Source: xkcd

And finally...

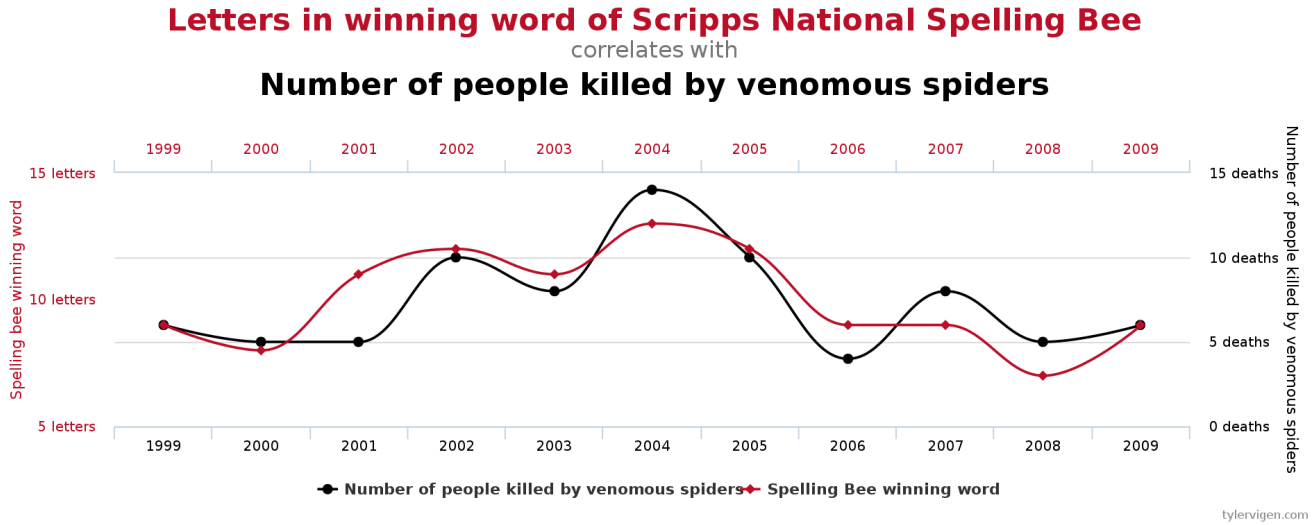


Source: xkcd

👋 If you want to know more about these comic strips take a look here.

This is an example of "P-Hacking"

P-Hacking refers to the misuse of data analysis to find patterns in data that can be presented as statistically significant. This is done by performing many statistical tests on the data and only reporting those that come back with significant results👋. **Example**: a correlation between the number of letters in Scripps National Spelling Bee's winning word and the number of people in the United States killed by venomous spiders.



Source: Wikipedia on Data dredging . 👋 If you want to know more, take a look here.

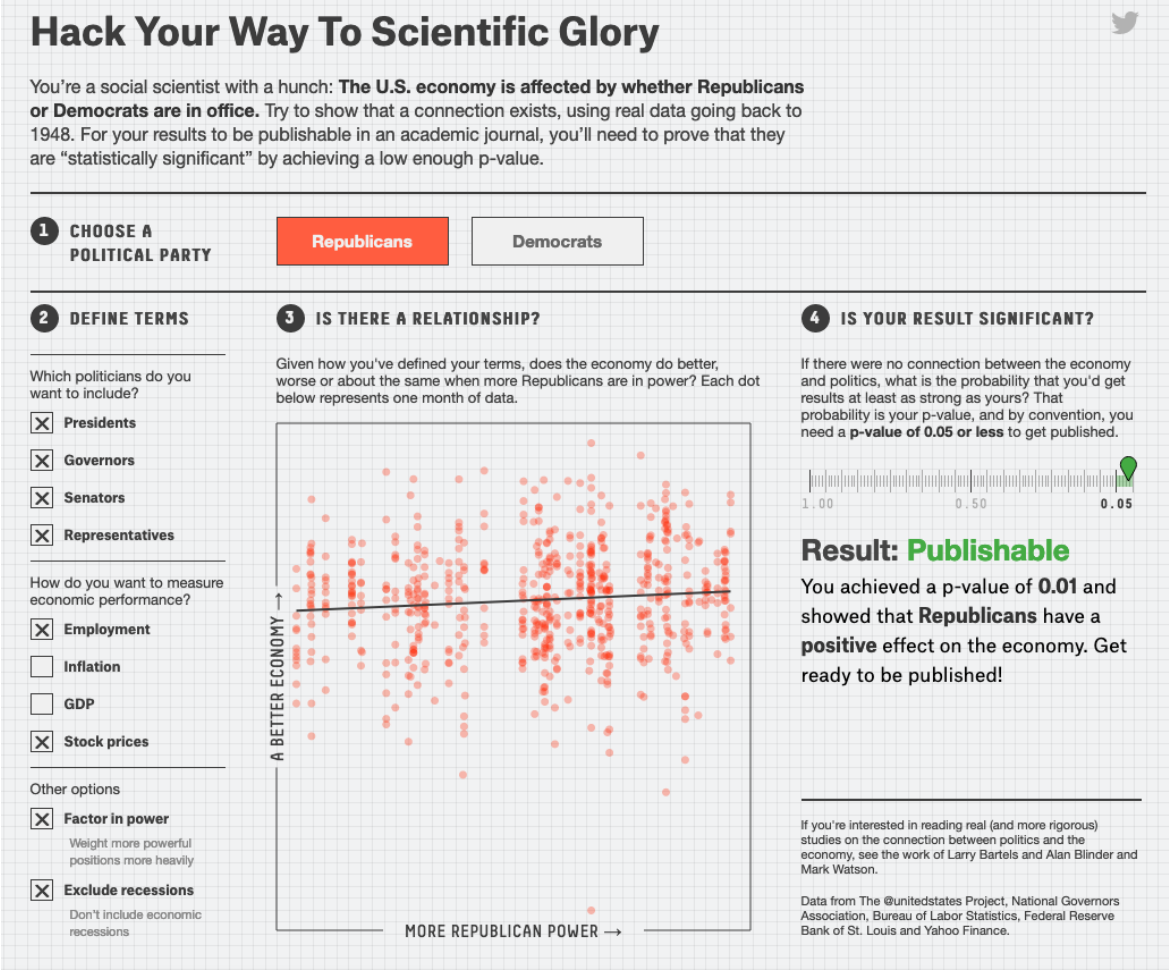
More on P-Hacking

Source: Scientific Studies: Last Week Tonight with John Oliver (HBO)

👋 If you want to know more, take a look here.

How to P-Hack? 🤔

Are republicans good or bad for the American Economy? **Both!!** (see link)



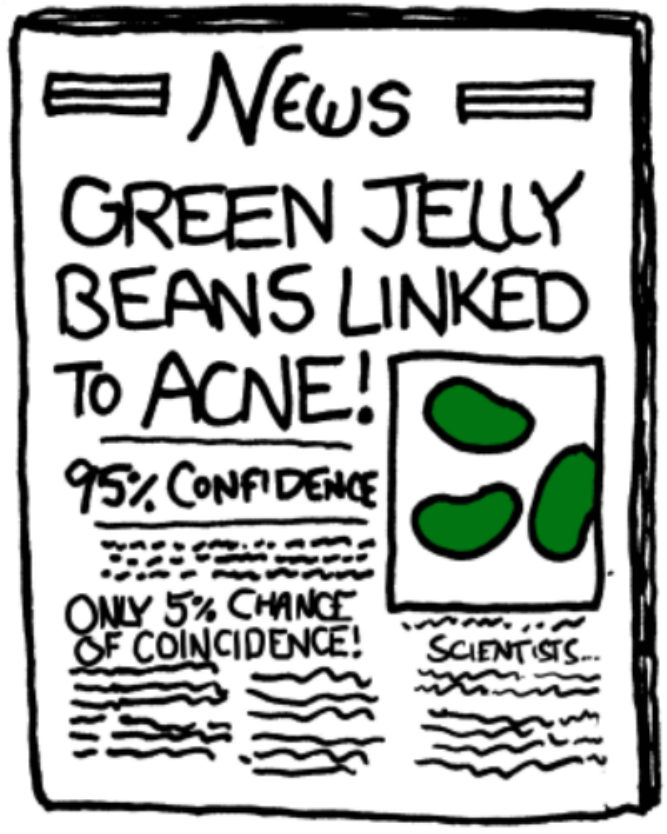
This brings us to "HARKing"

HARKing is an acronym coined by Norbert Kerr for the questionable research practice of "Hypothesizing After the Results are Known". It can be defined as presenting **a post hoc** hypothesis in the introduction of a research report as if it were an **a priori** hypothesis.

How to do it? 🤔

- Test if one jelly bean color is linked to acne.
- Pretend this was our original hypothesis.
- Et voilà!

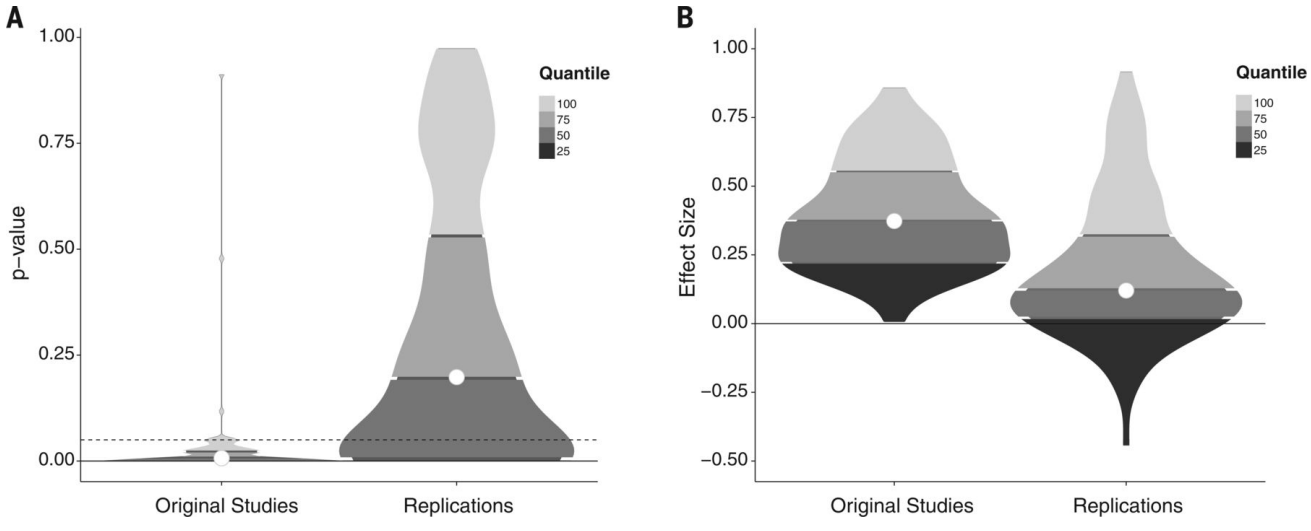
👋 If you want to know more, have a look here and the references therein.



Source: xkcd

The costs of HACKing (and P-Hacking)

HARKing is (arguably) playing an important role in the **replication crisis in science**!



Source: Open Science Collaboration. "Estimating the reproducibility of psychological science." Science 349.6251 (2015).

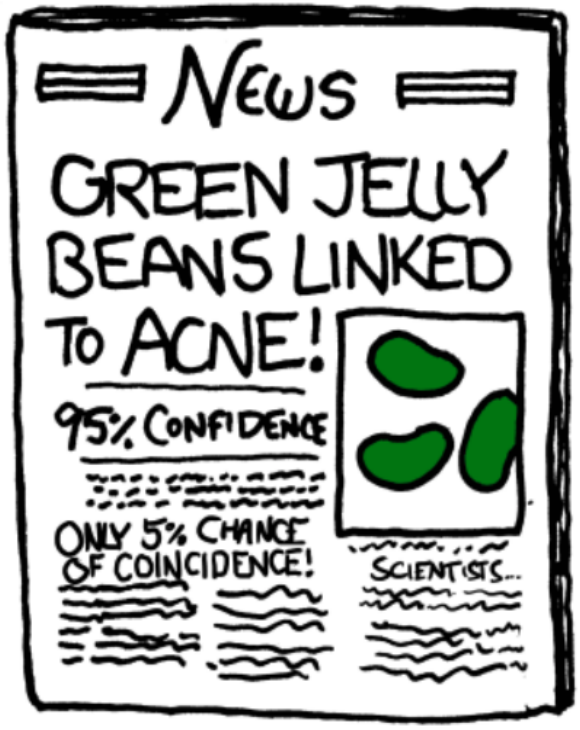
👋 If you want to know more, take a look here and the references therein.

⚠ Take home message

- Any data analysis made on a sample of data is subject to randomness.
- Understanding the random processes underlying any data analysis (including underlying statistics such as confidence intervals and p-values) requires basic knowledge in probability and statistics, or statistical literacy.
- With data, there is never 100% certainty, but rather a conclusion associated to a (estimated) risk that the conclusion might be wrong.
- One has to live with this fact, and any scientific result that claims the validity of a result with 100% certainty on the basis of data, is simply a fallacy.
- However, properly controlling for the statistical risk and properly stating the conclusions that can be drawn from a data analysis can really bring new knowledge, especially in all the sciences that are based on observations.

Thank you very much for your attention

All material is available on GitHub at https://github.com/stephaneguerrier/StatLiteracy



Source: xkcd