Introduction to Statistical Literacy

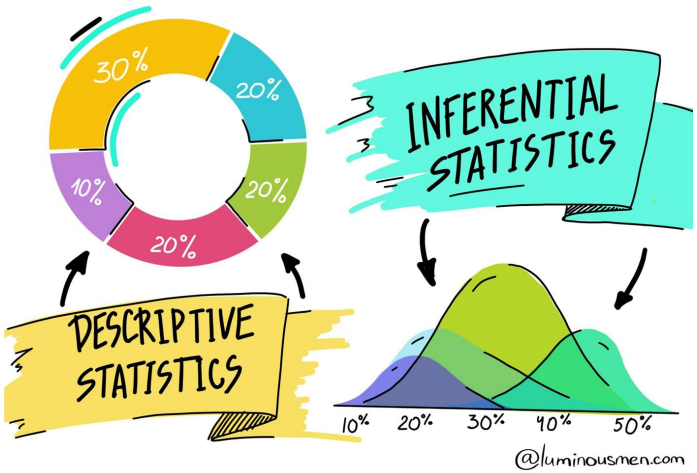
Analyse et traitement de l'information

Comprendre le numérique

Stéphane Guerrier
22 October 2020

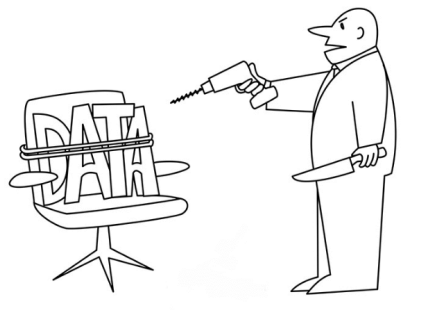
What is statistics?

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).



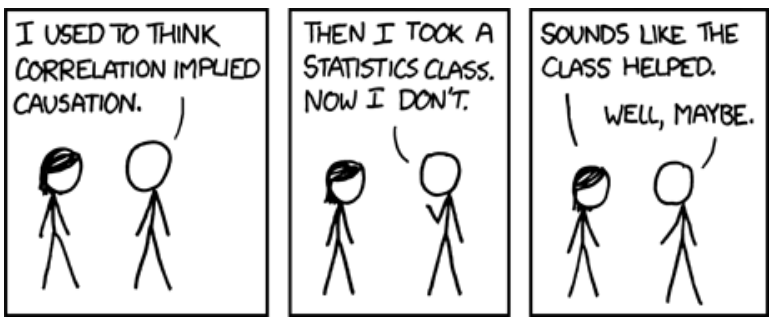
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

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.

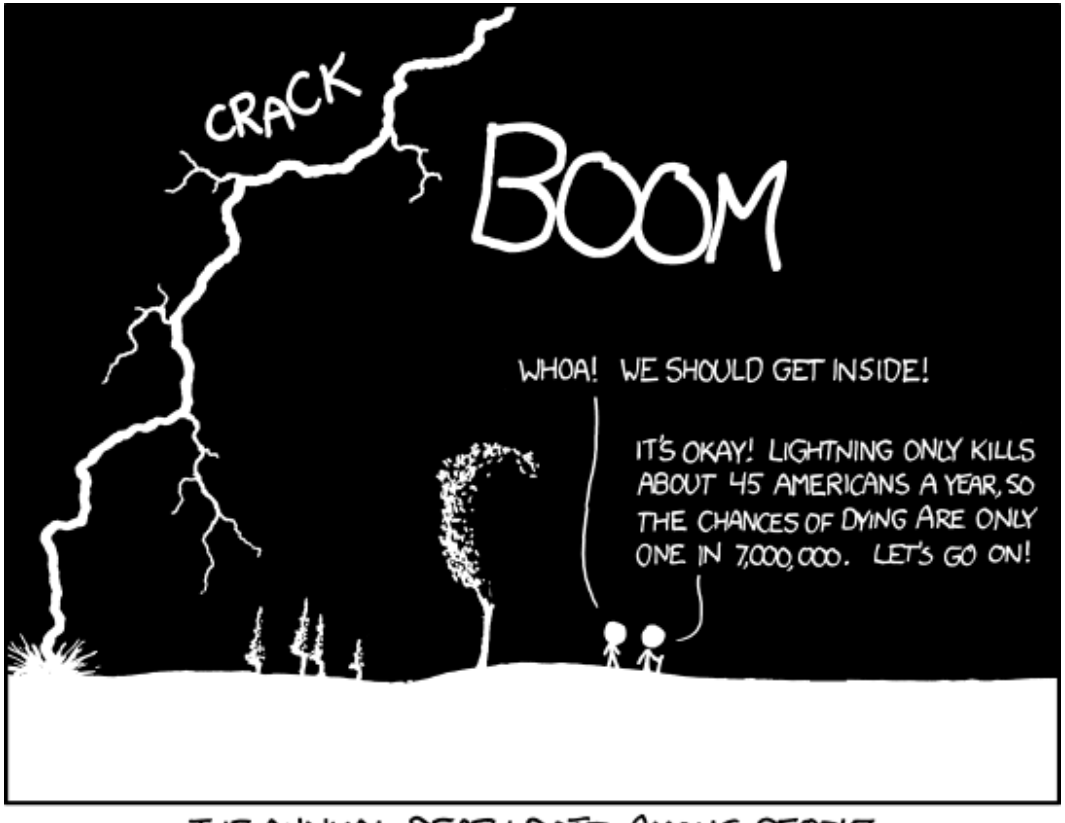


Source: xkcd

👋 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:

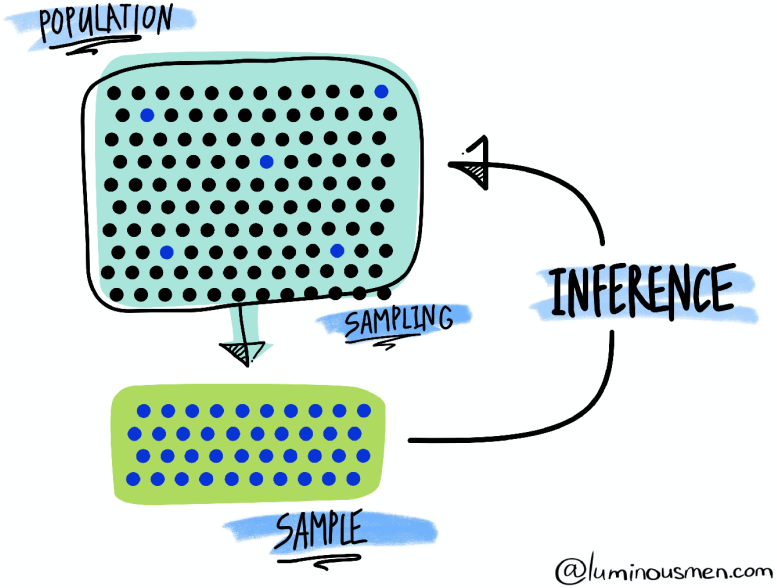
1. Visualize data (e.g. propagation of COVID-19 in different countries).
2. Understand and interpret data (e.g. main causes of cancer).
3. Make predictions (e.g. predicting unemployment or risk indices).
4. **Assess the validity of a hypothesis**.

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

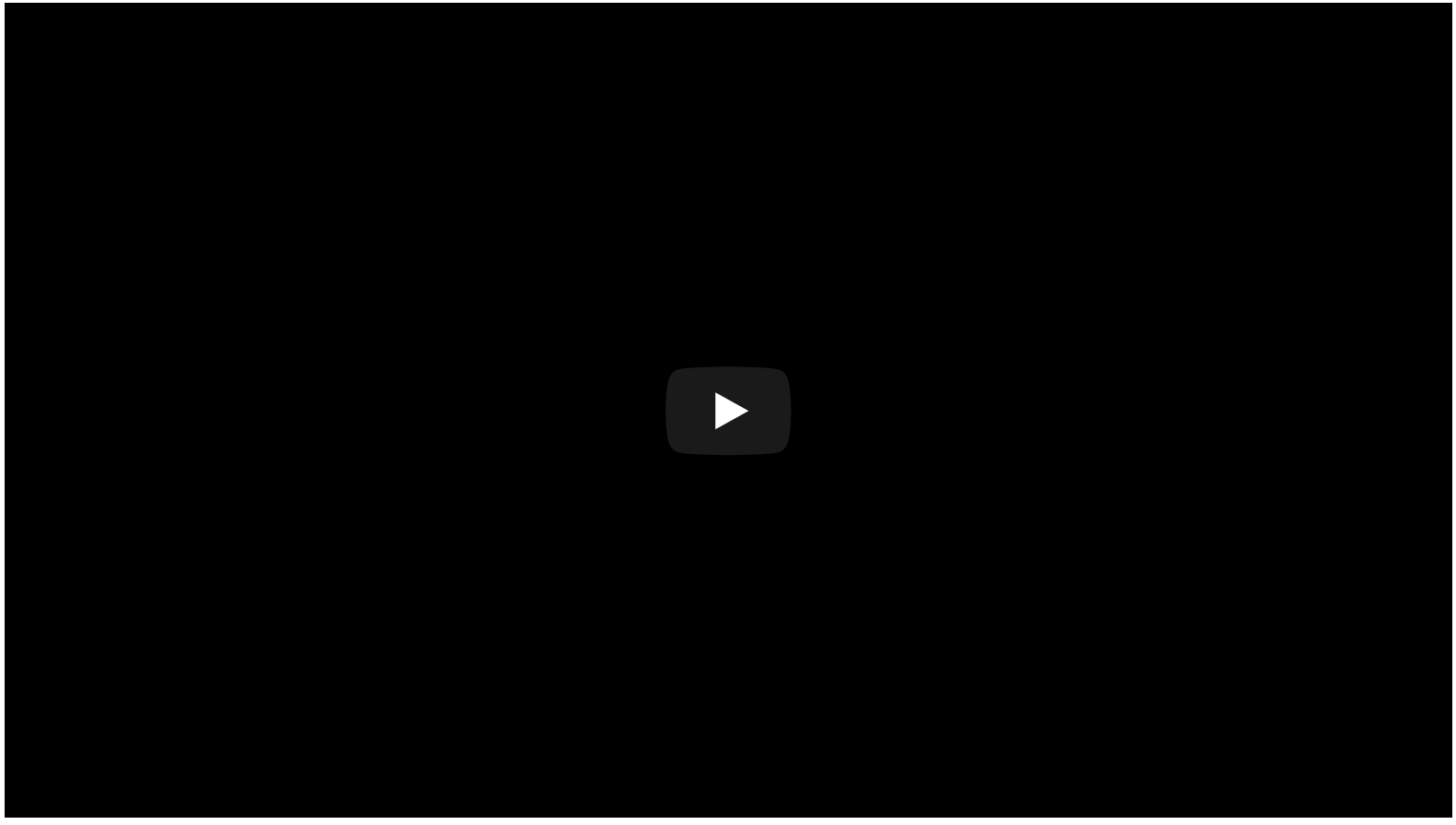
How does it work?

- 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? Is Hydroxychloroquine working or not?).

👋 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?

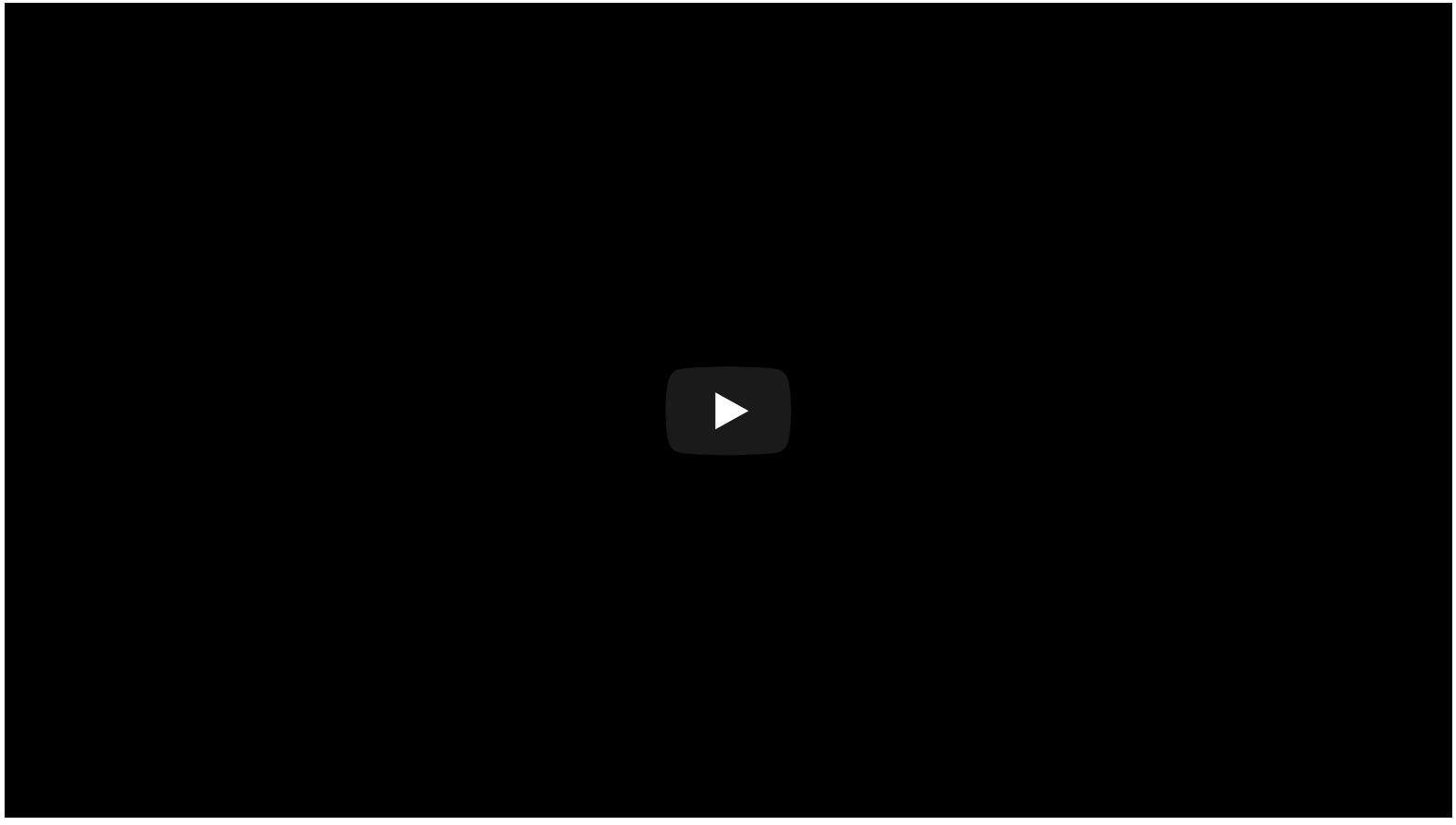
"A glass of red wine is as good as an hour at the gym"? 🙄



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

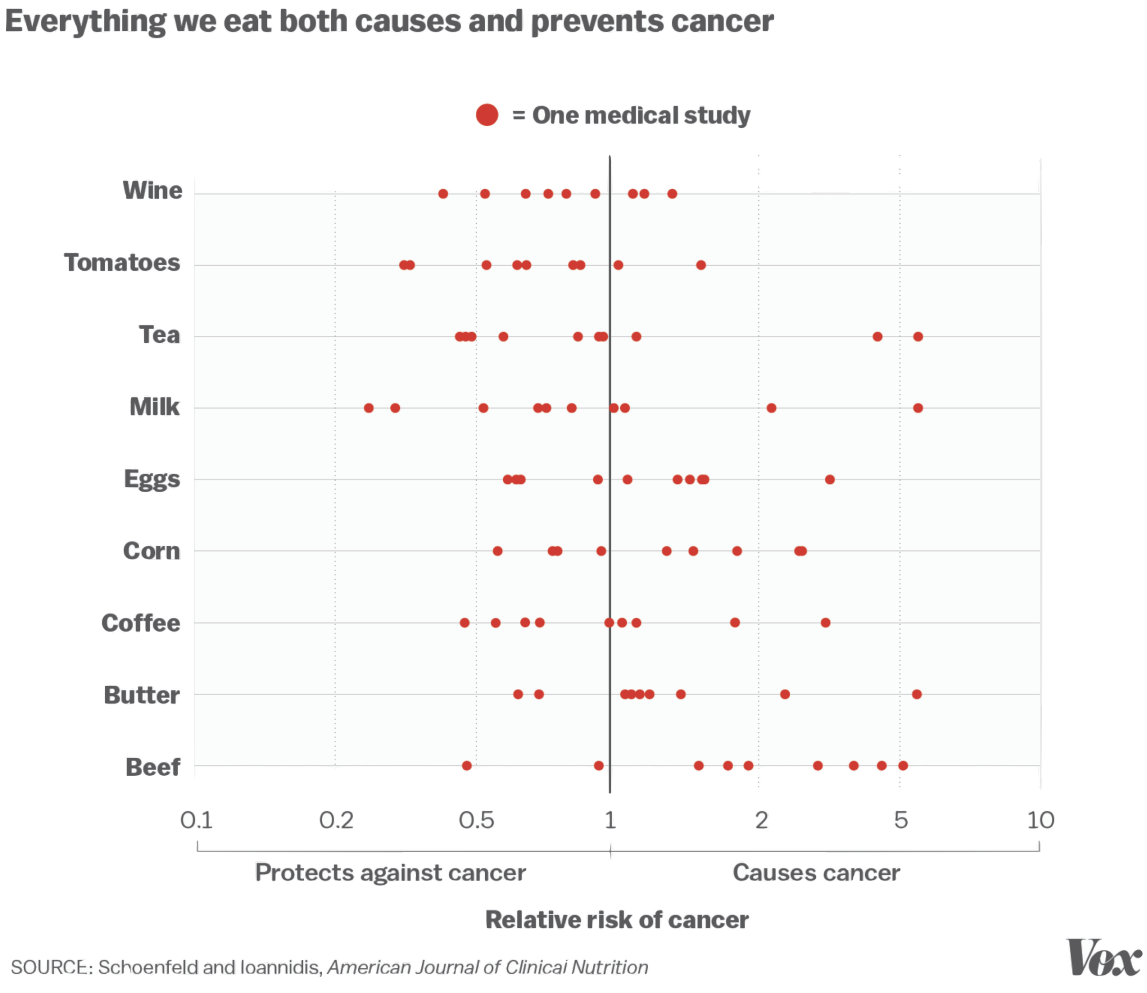
Contraditory studies?

What about coffee? Is it good?



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

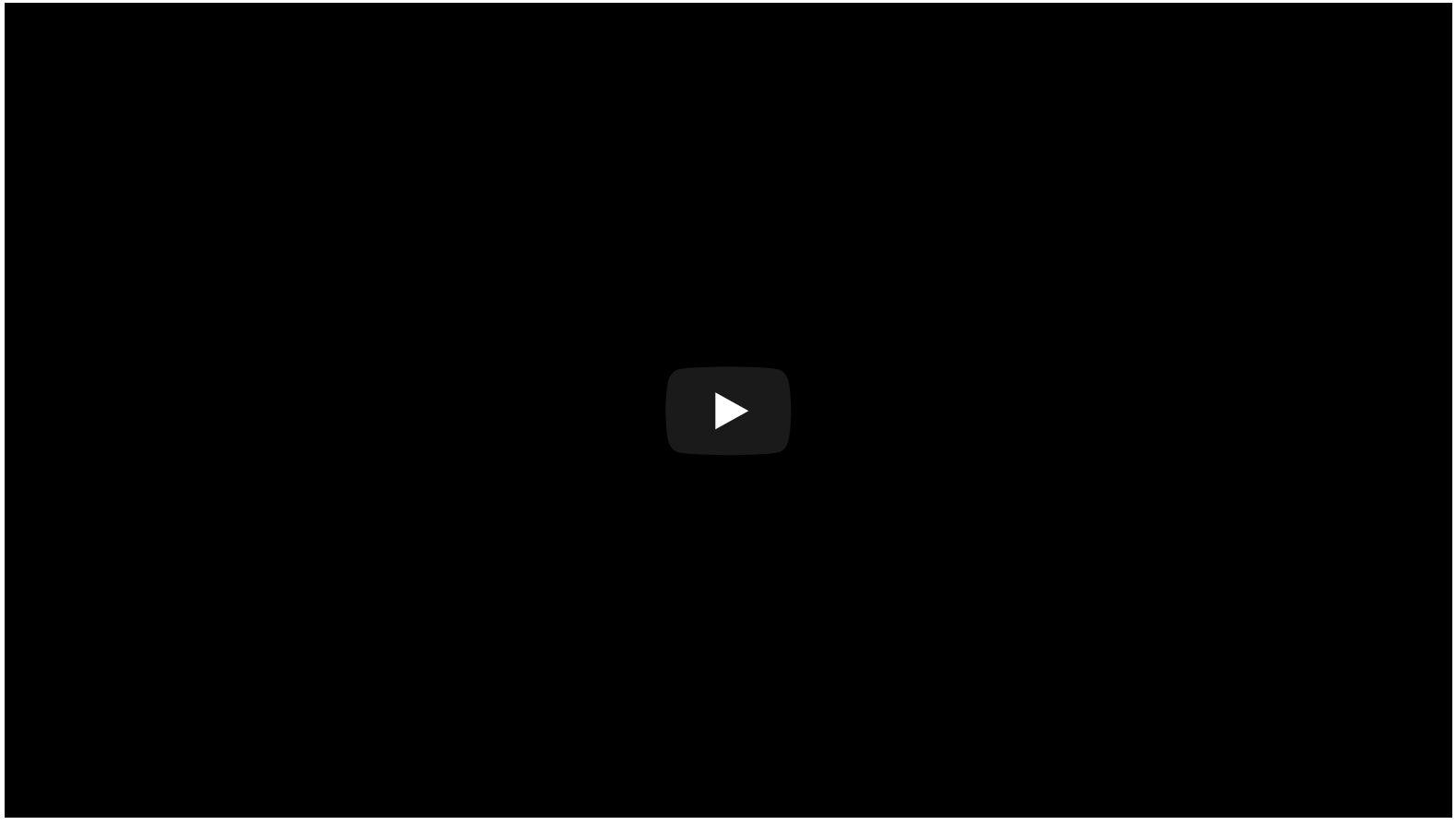
So everything causes cancer?



👋 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)

The Hydroxychloroquine Controversy 🦠

Since the beginning of 2020, various research teams have suggested that (hydroxy)chloroquine combined with other drugs (e.g. azithromycin) is an **effective treatment** for the coronavirus disease 2019. Since then there have been **many many many contradictory publications** (and some retractions!) and the efficacy of these drugs remains unclear.

As summarized in the Wikipedia entry dedicated to hydroxychloroquine (and citing various scientific sources): "*There is no strong scientific evidence to support the use of hydroxychloroquine for preventing or treating coronavirus disease 2019 (COVID-19)*".

So what does this mean? 🤔

Does it mean that hydroxychloroquine is not effective in preventing or treating COVID-19? That it is effective? Actually, neither... 🤔

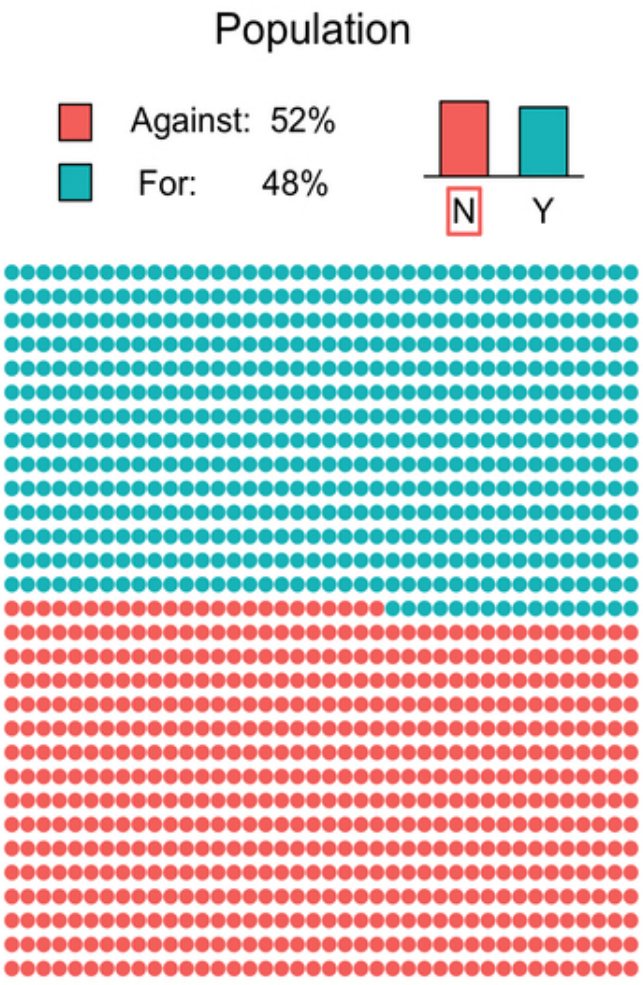
Population and Sample - Example

To fix ideas we will consider a simple example. On September 27 2020, the Swiss population voted on the *Loi fédérale sur la chasse et la protection des mammifères et oiseaux sauvages (Loi sur la chasse, LChP)*. The proposal was not accepted as it only received 48% of supporting votes👋. Suppose that we were interested in collecting data **before the vote** to assess if this proposal was likely to be accepted. In this example, we will make the following assumptions for simplicity:

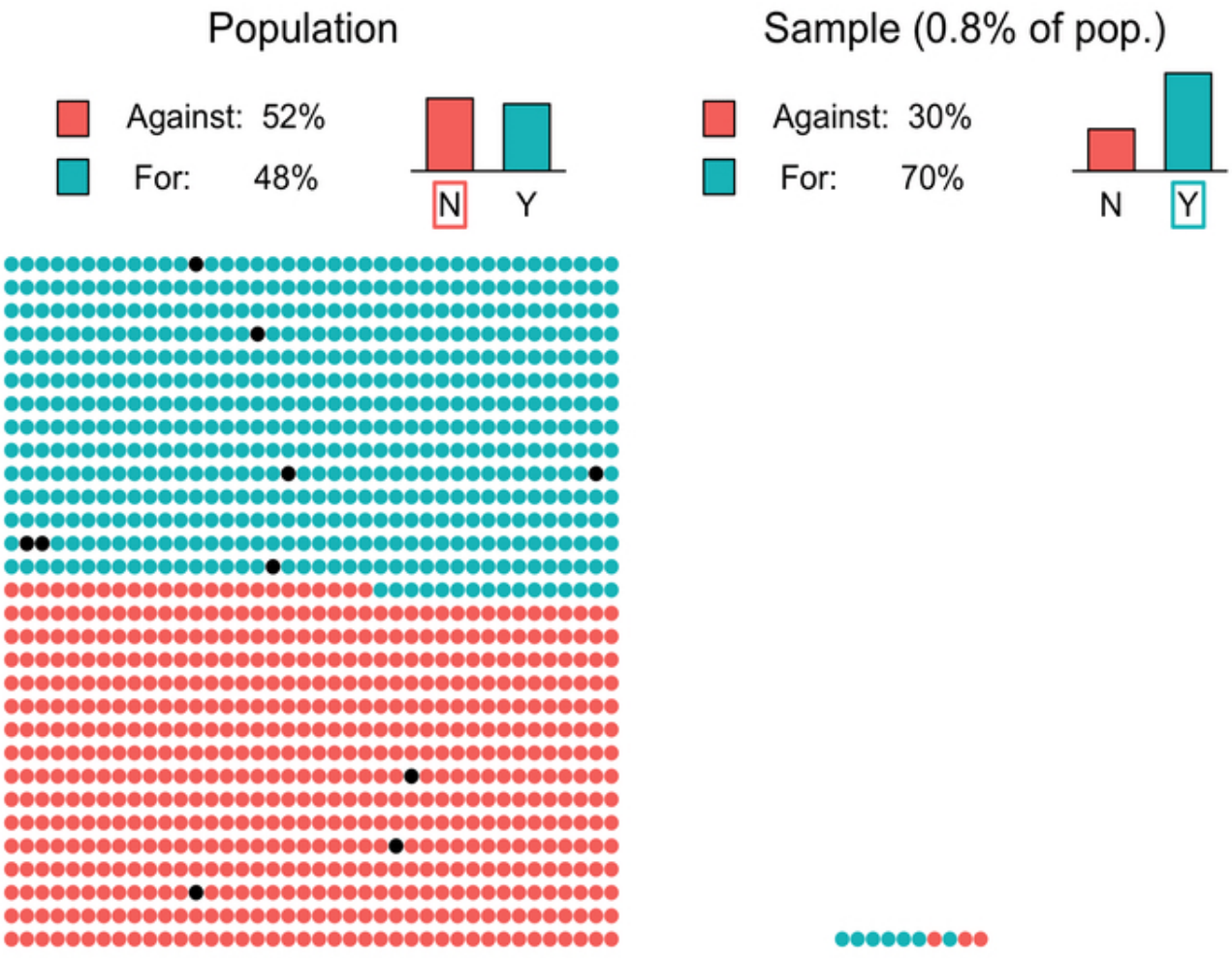
- The Swiss population of voters is composed of 1200 individuals (576 for the project, 624 against).
- We can perfectly sample the population (everyone is available, no double sampling, and the sampling is random).
- People don't change their mind and they don't lie.

👋 More details on the results can be found here.

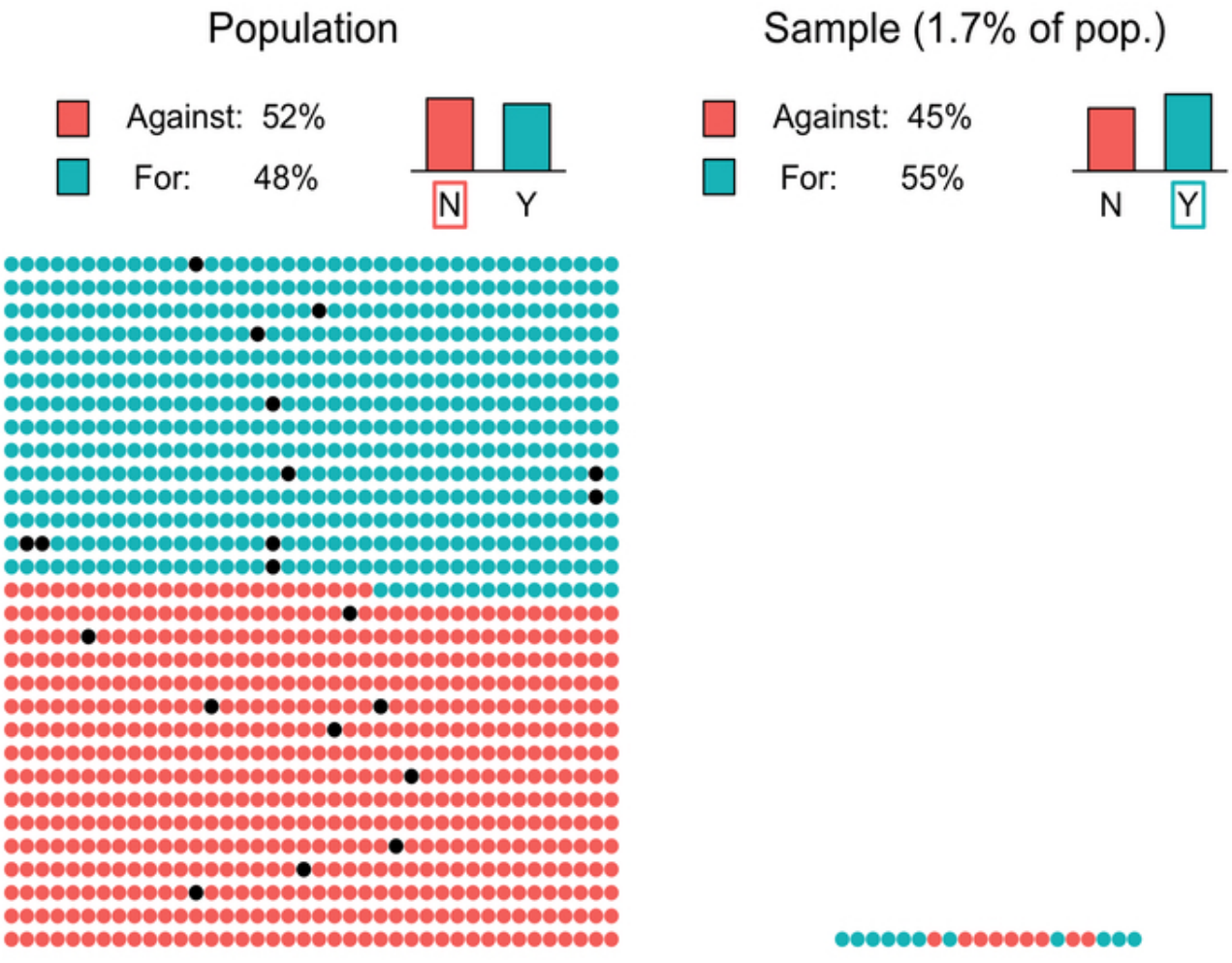
Population and Sample - Example



Population and Sample - Example

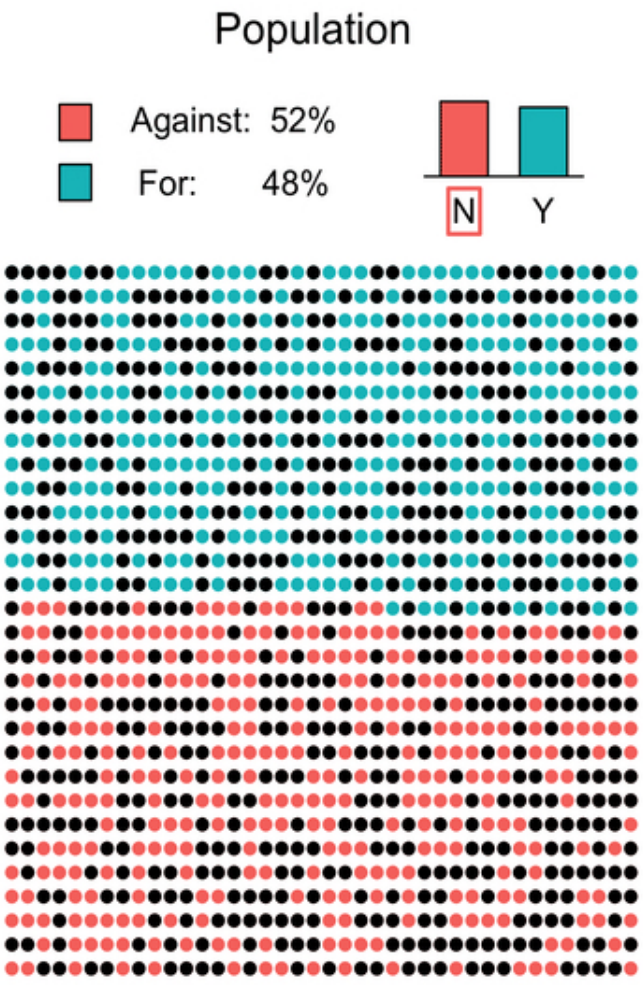


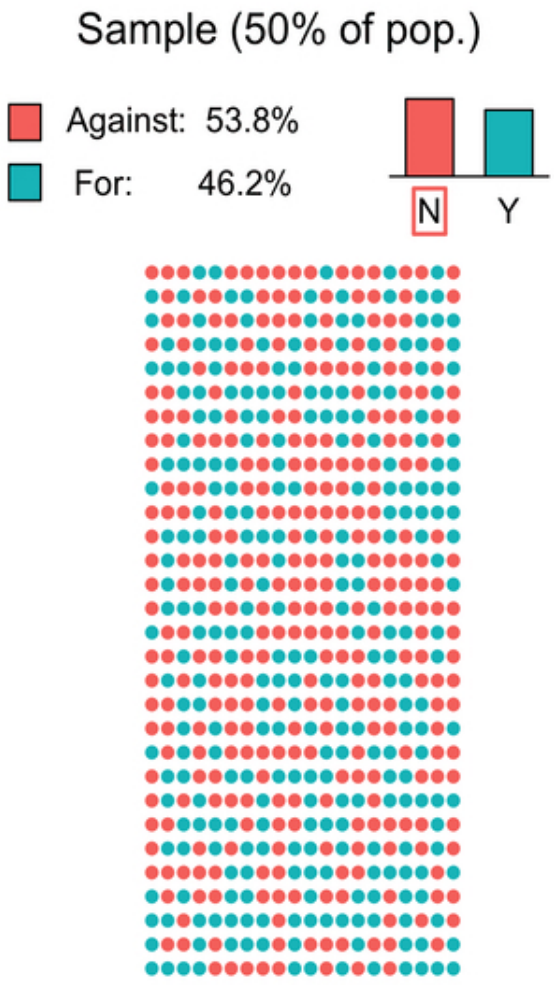
Population and Sample - Example



Population and Sample - Example

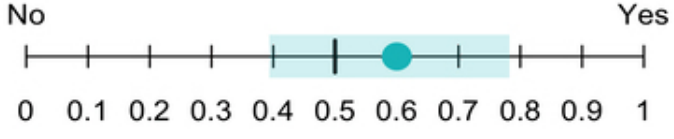
Population and Sample - Example





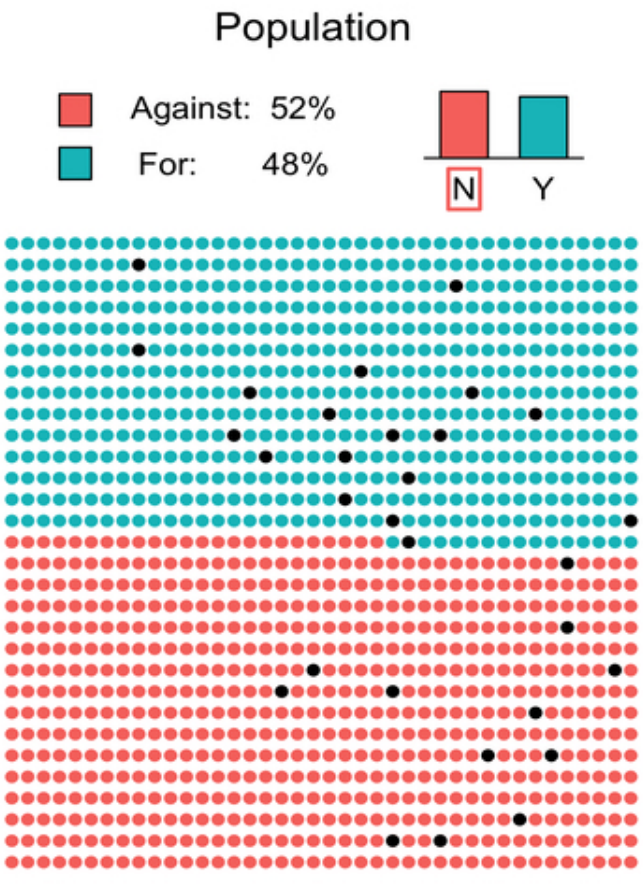
How to measure uncertainty?

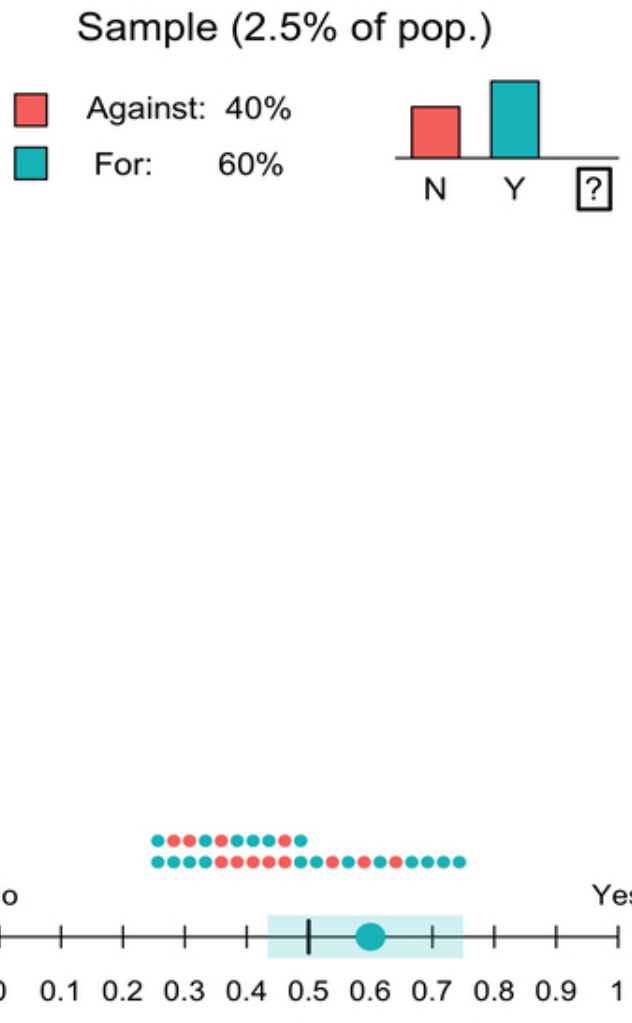
- Uncertainty can be measured in many different ways.
- A common approach (in statistics) is to use **confidence intervals**.
- These intervals correspond to a range of values that are likely to include the population value with a certain level of confidence. The level of confidence is a probability expressed as a percentage (%).
- In our example, we are interested in the true (population) proportion of voters in favor of the proposal (48%). Using our second sample (8 against and 12 for) we could construct the following 90% confidence interval:



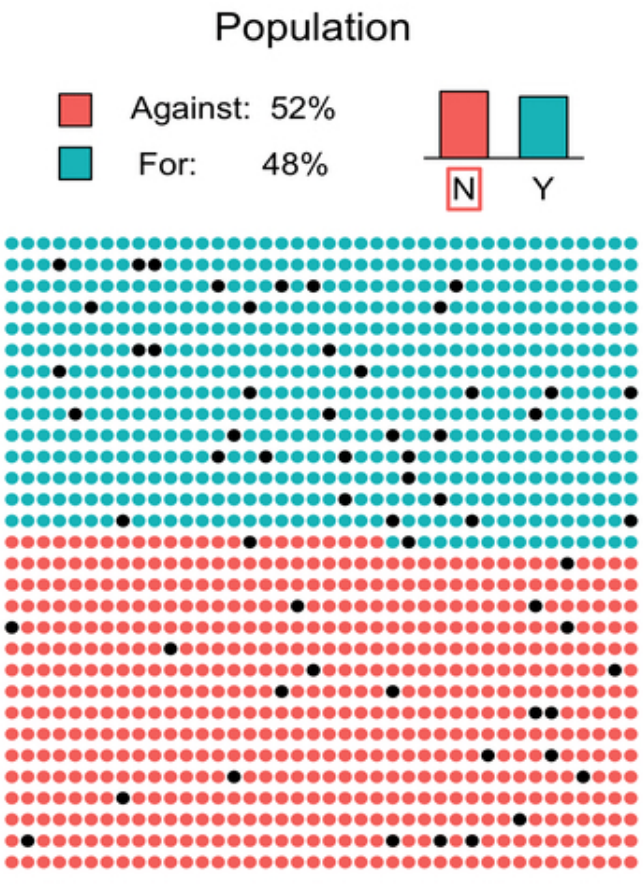
So what does it mean? 🤔 It means that with a **probability of 90%** the true proportion of voters for the project (48% in this case) is between say 40% and 80%.

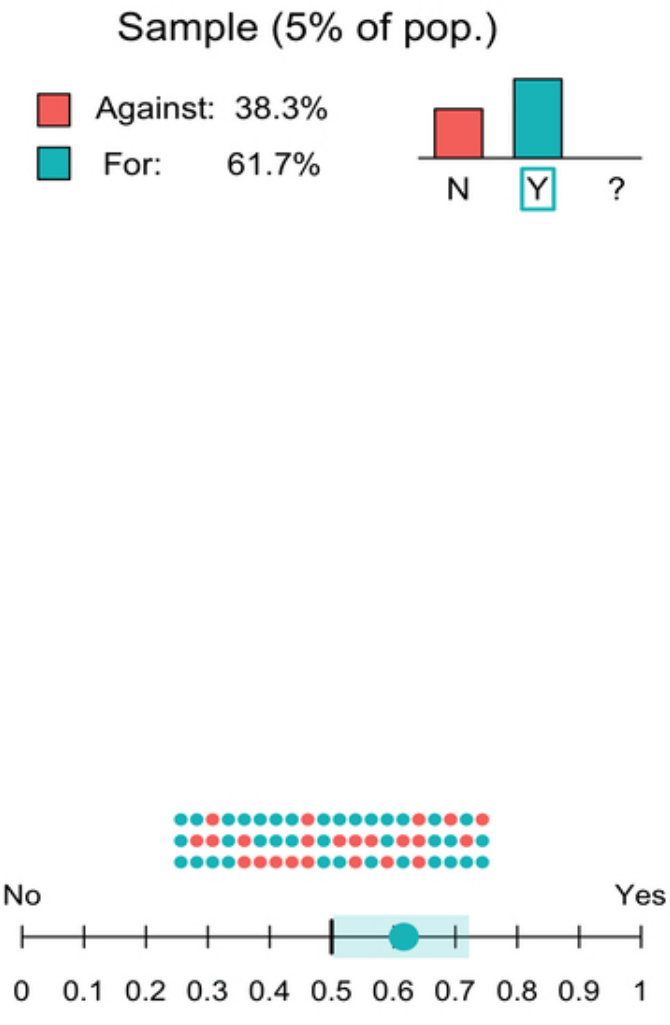
How to measure uncertainty?



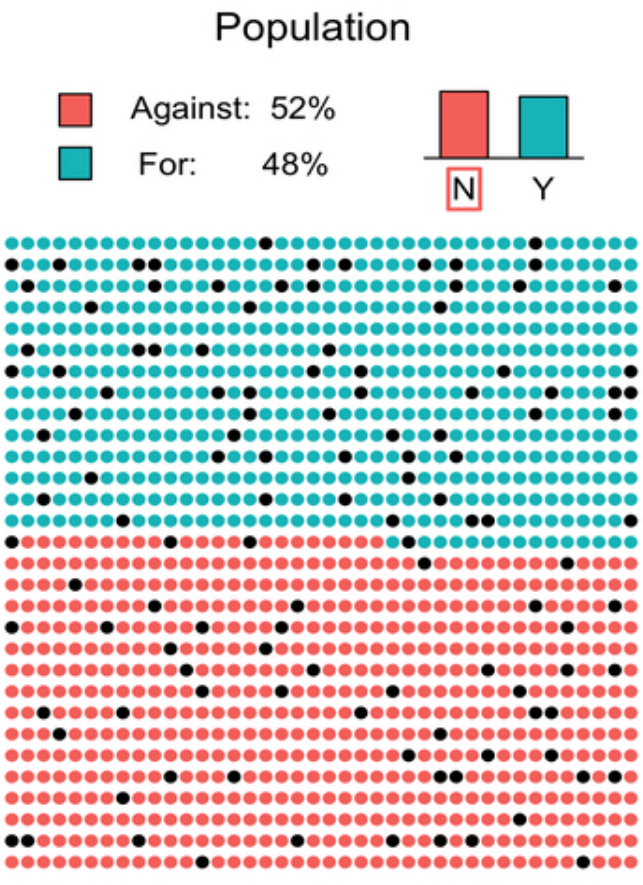


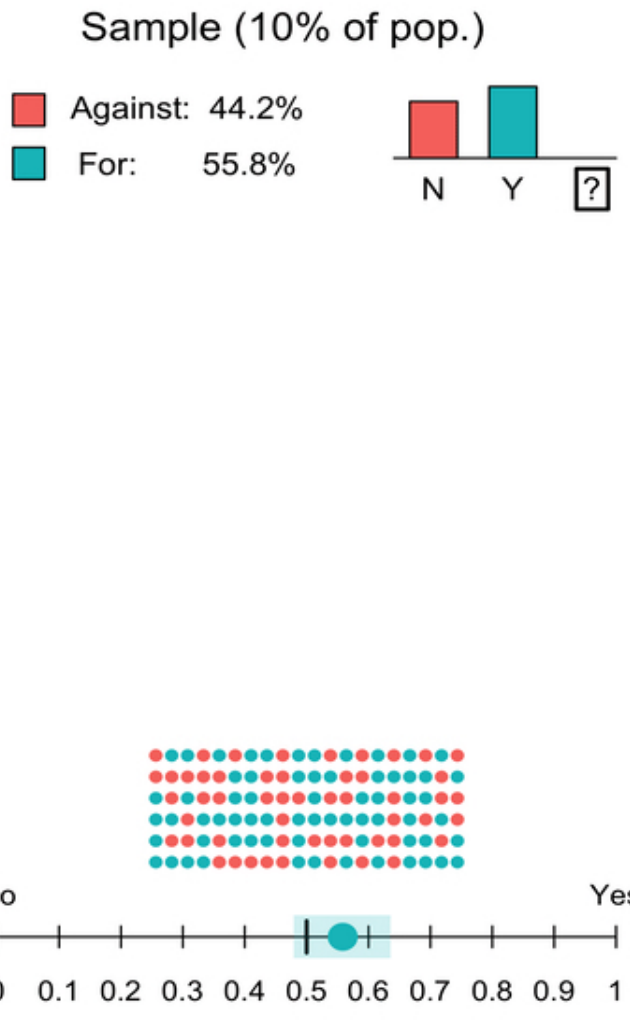
If we had more data? 😳





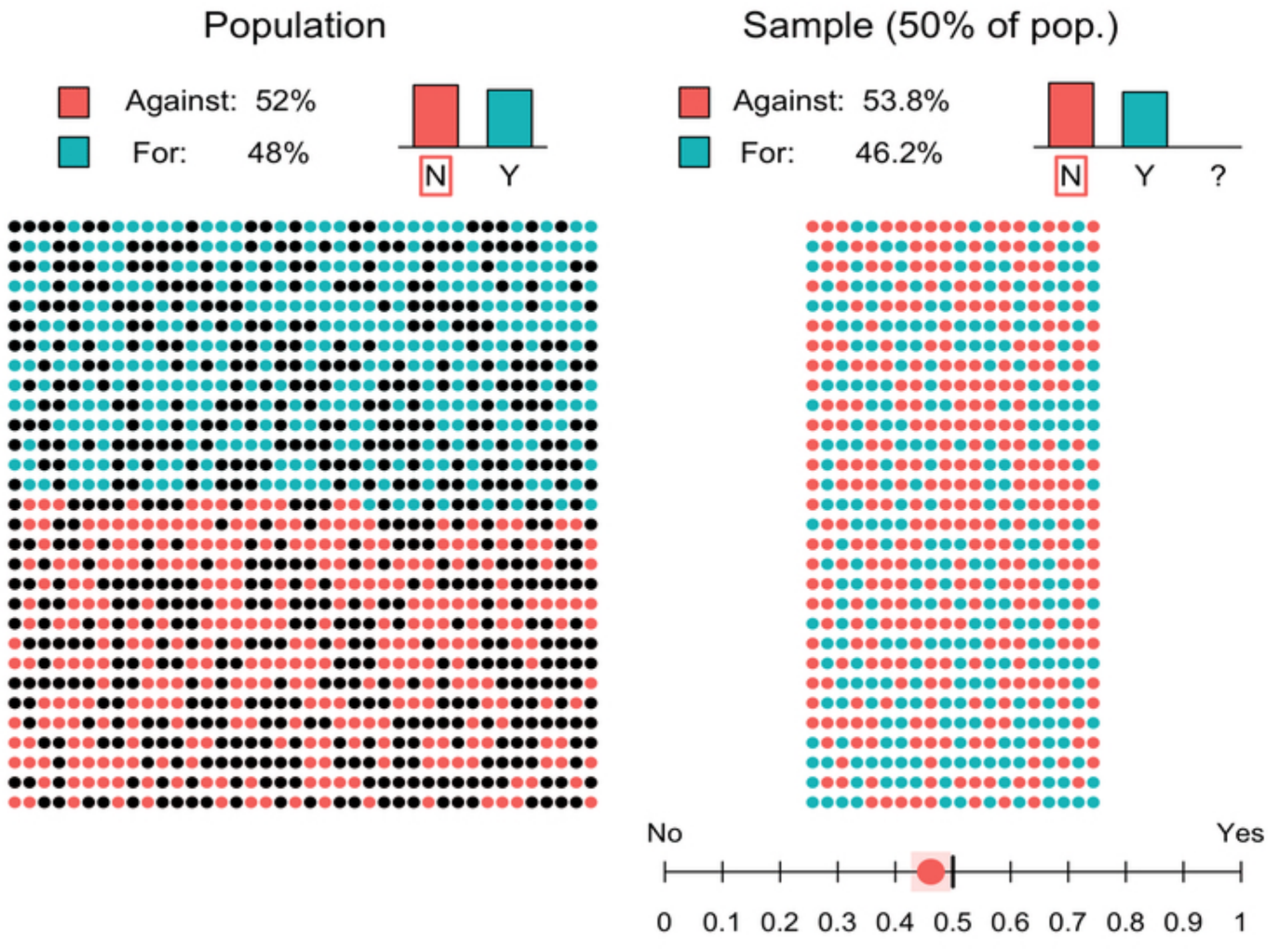
If we had more data? 🤔





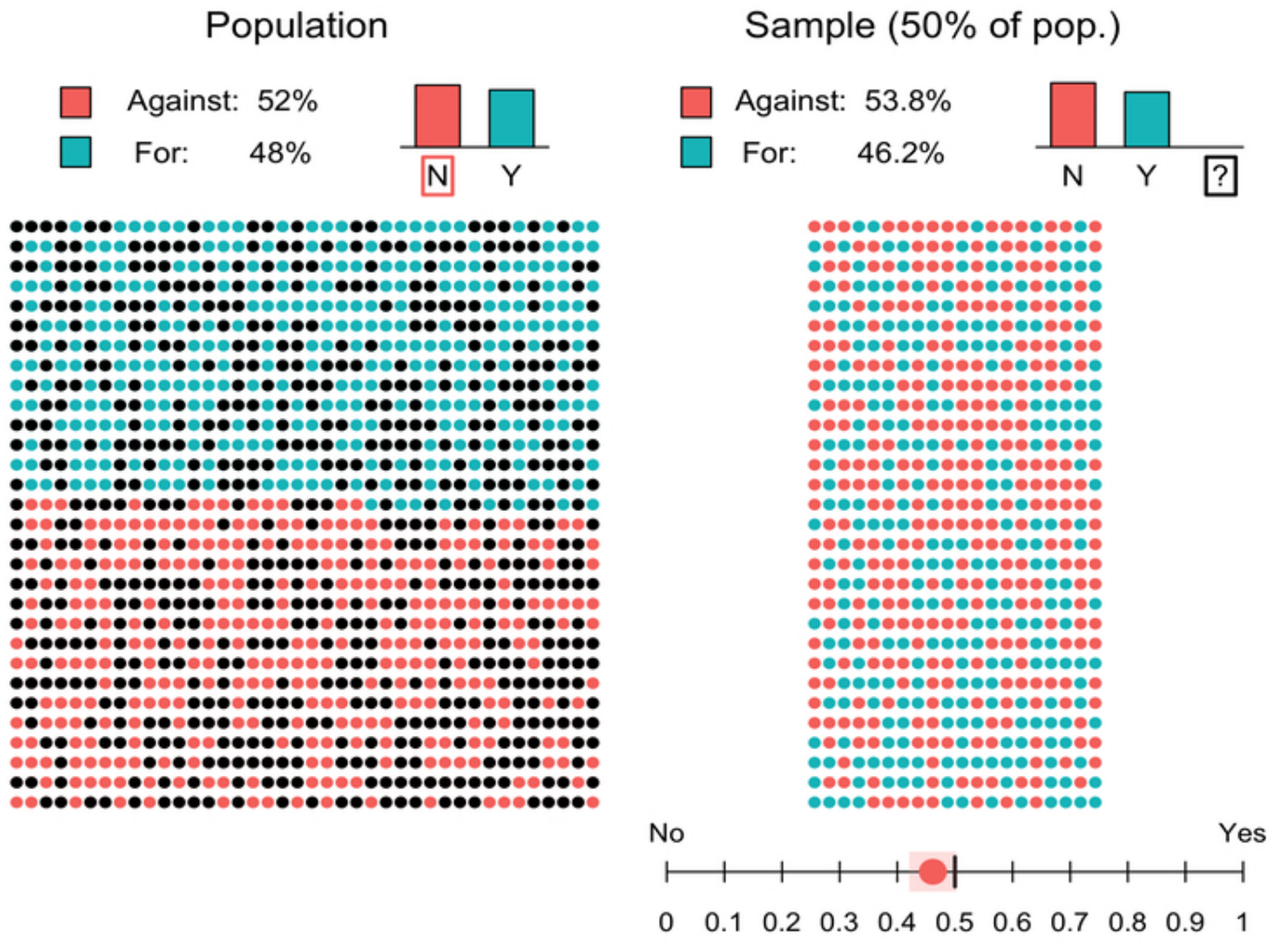
How to measure uncertainty?

How to measure uncertainty?



Why 90%? Why not 95%?

Why 90%? Why not 95%?



Is it too much? Maybe 60%?

What if I want to be really sure? 99%?

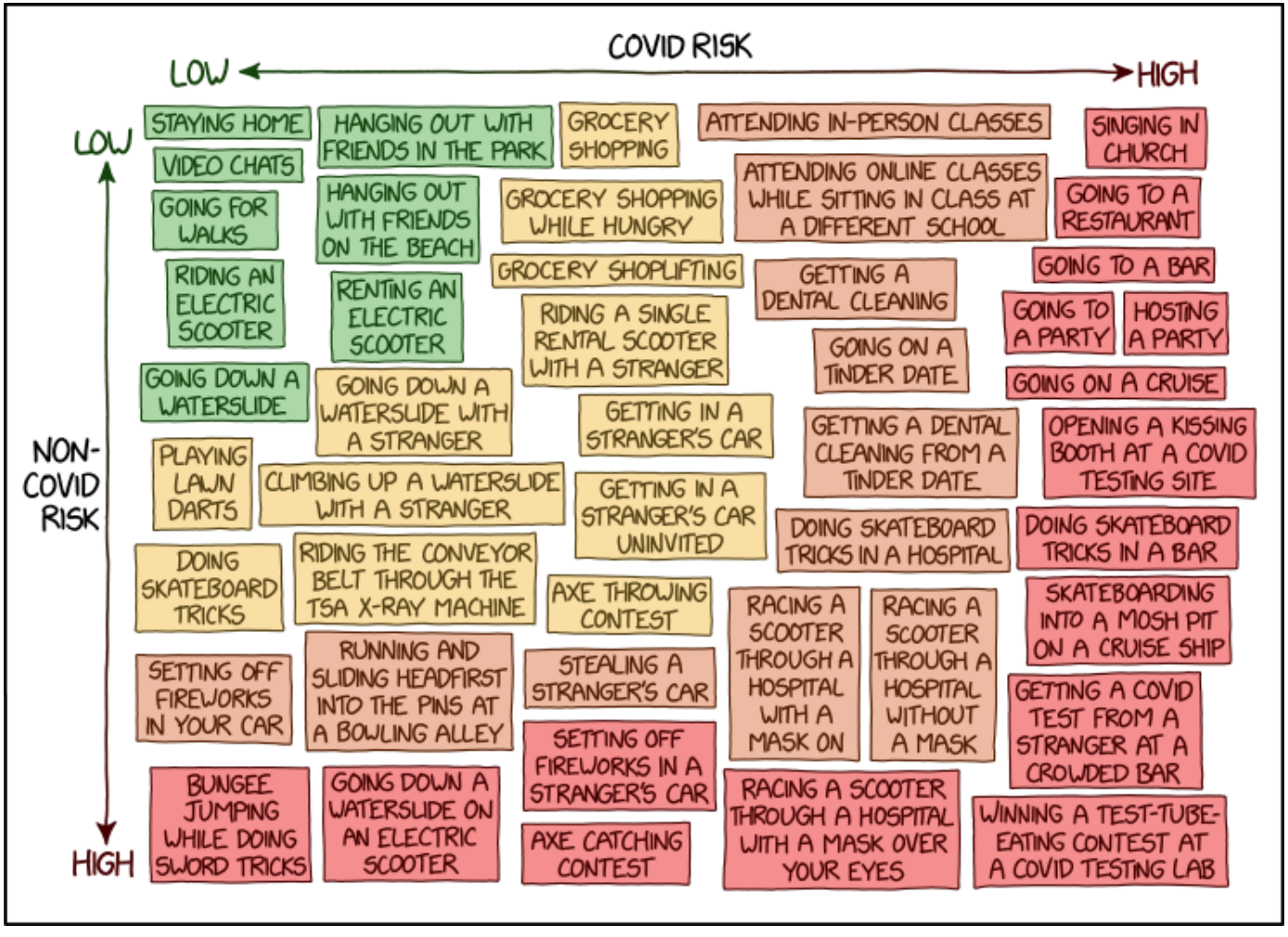
A bit more? How about 99.99%?

What if I want to be certain? 100%? 😳

⚠️ 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).

⚠ At the end it's a question of risk



Source: xkcd

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.

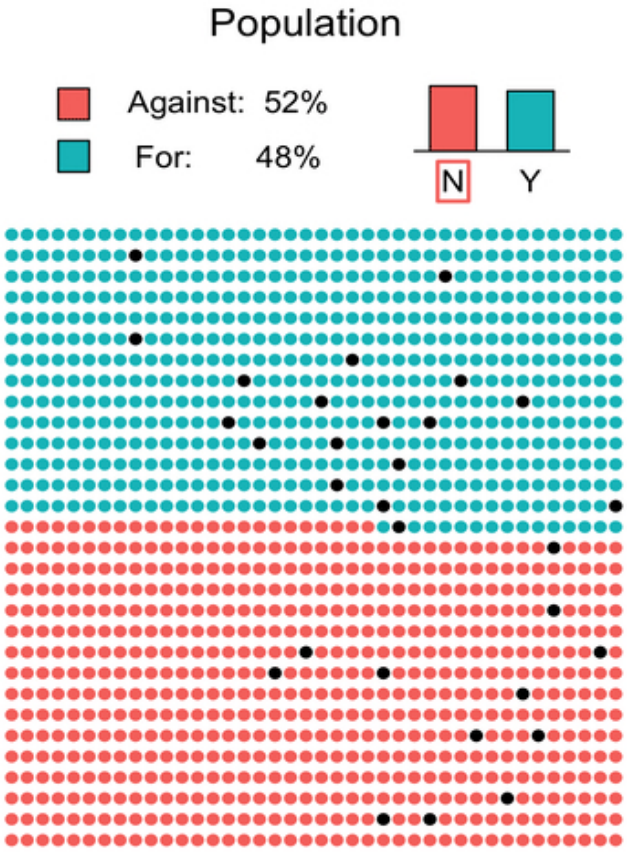
How to test a (scientific) hypothesis?

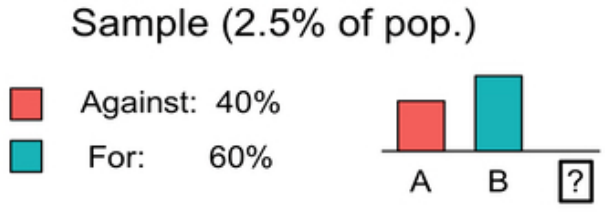
- A p-value is associated to a (couple of) hypotheses about the phenomenon under investigation. For example, for the voting proposal, one hypothesis is that the majority is against the proposal, and the other one is that there is no majority (the YES and NO are both at 50%)
- Each hypothesis excludes 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 hypothesis 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. proportion of NO is smaller than 50%).
- 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.

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.

Testing





$\alpha = 5\%$

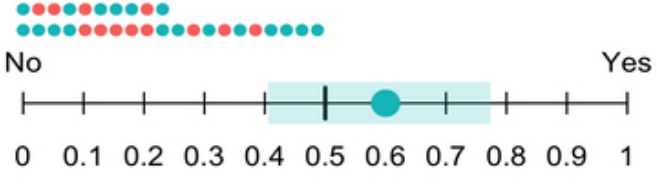
$H_0 : p_{yes} = 50\%$

$H_1 : p_{yes} < 50\%$

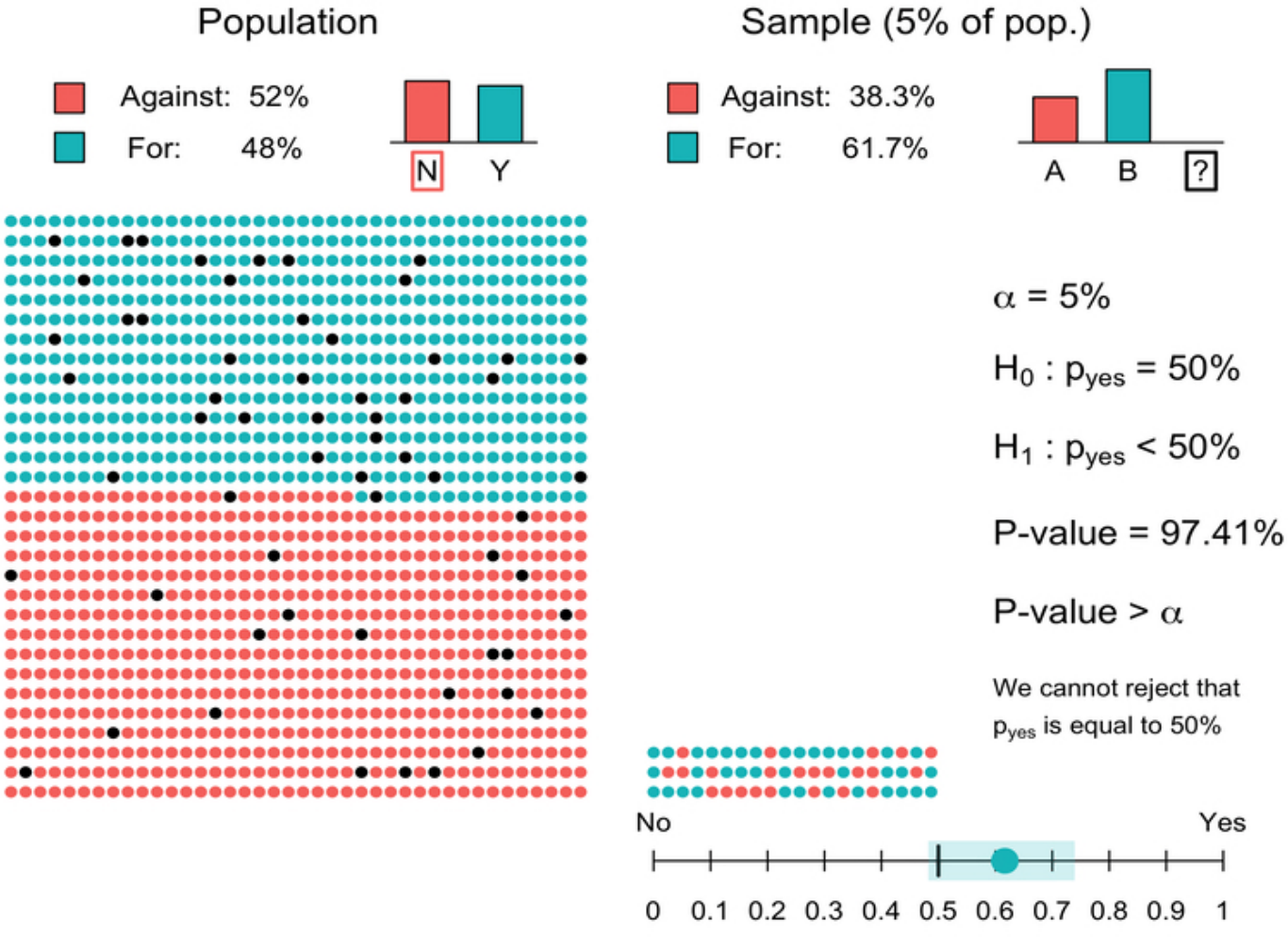
P-value = 89.98%

P-value > α

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



Testing

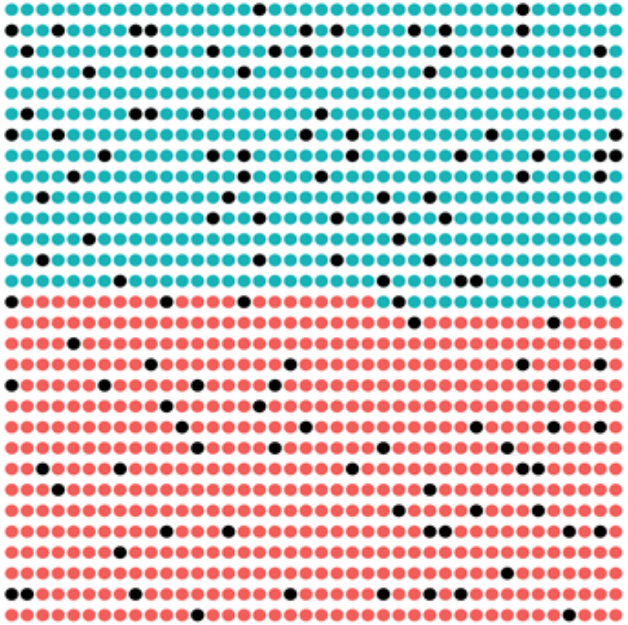

Population
Against: 52%
For: 48%
N
Y
Sample (5% of pop.)
Against: 38.3%
For: 61.7%
A
B
?
α = 5%
H0 : pyes = 50%
H1 : pyes < 50%
P-value = 97.41%
P-value > α
We cannot reject that
pyes is equal to 50%
No
Yes
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Testing

Population

Against: 52%

For: 48%



Sample (10% of pop.)

Against: 44.2%

For: 55.8%

$\alpha = 5\%$

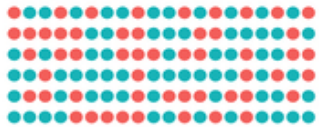
$H_0 : p_{yes} = 50\%$

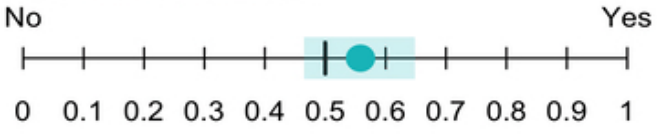
$H_1 : p_{yes} < 50\%$

P-value = 91.47%

P-value > α

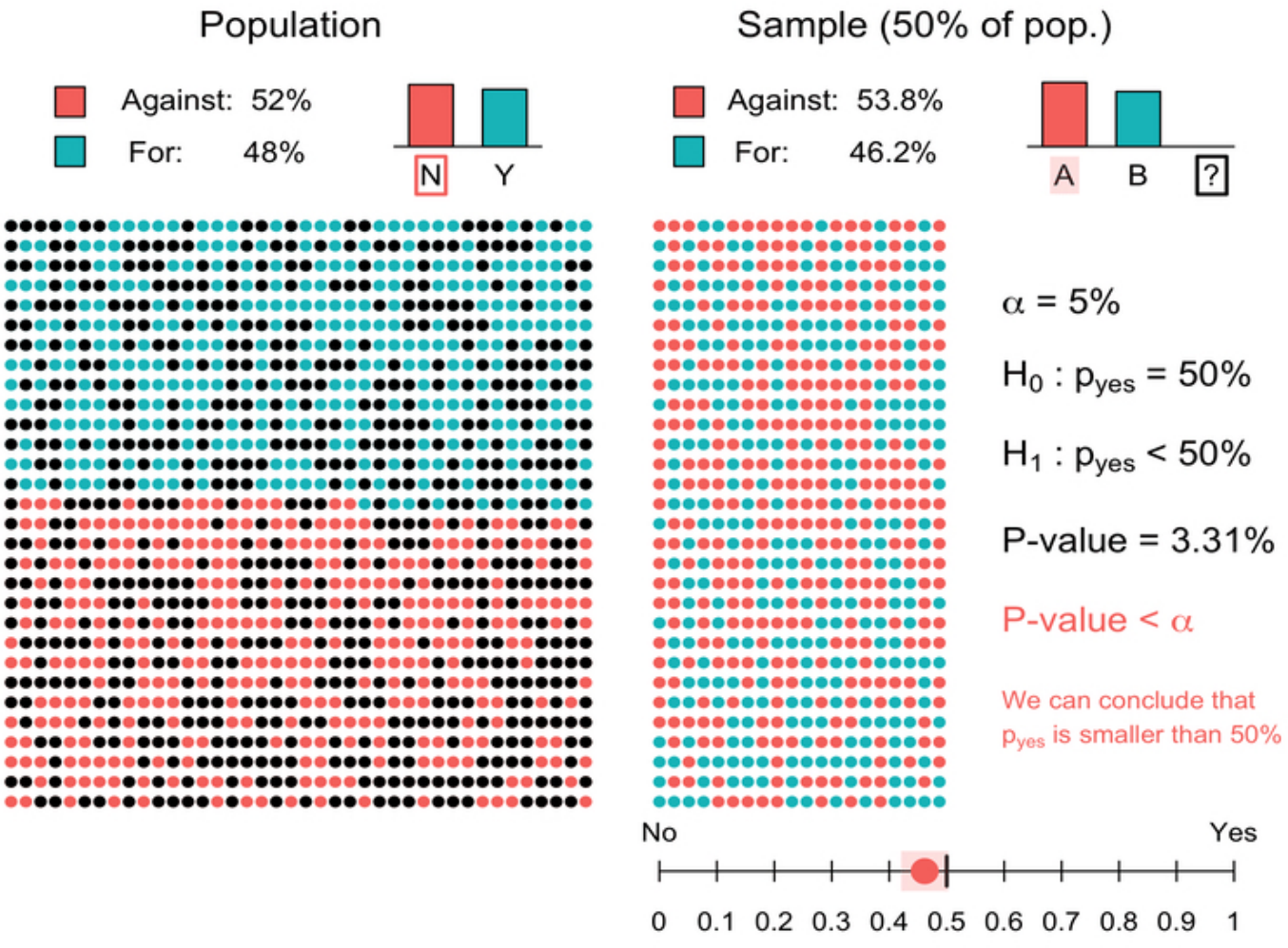
We cannot reject that p_{yes} is equal to 50%





Testing

Testing

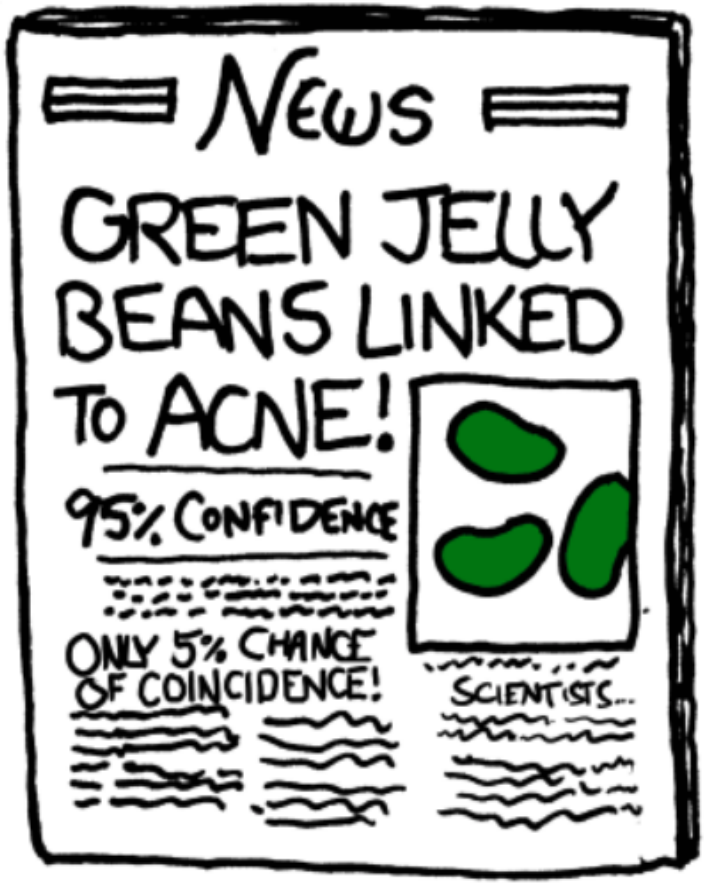

Population
Against: 52%
For: 48%
N
Y
Sample (50% of pop.)
Against: 53.8%
For: 46.2%
A
B
?
α = 5%
H0 : pyes = 50%
H1 : pyes < 50%
P-value = 3.31%
P-value < α
We can conclude that
pyes is smaller than 50%
No
Yes
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

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.
- 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! 🤓).

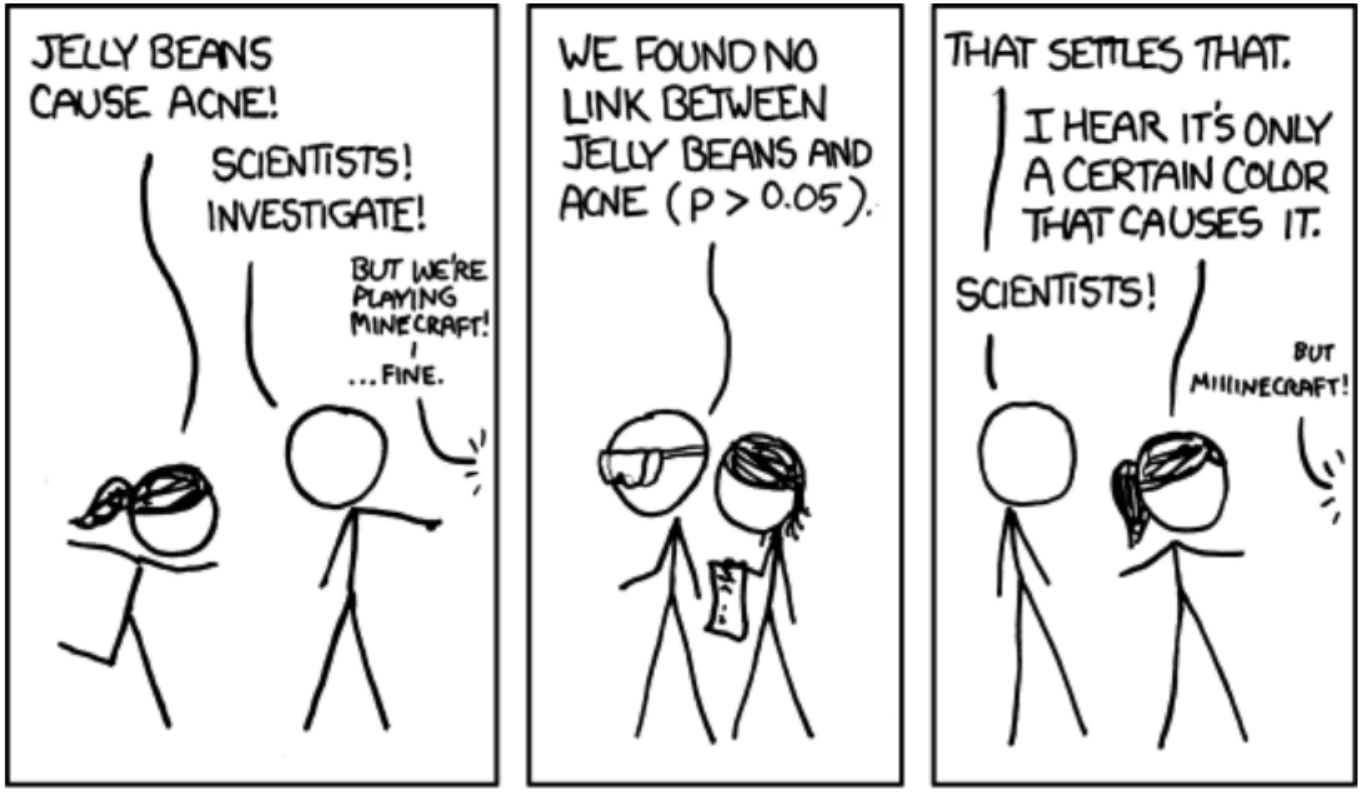
Why are there contradictory studies?

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



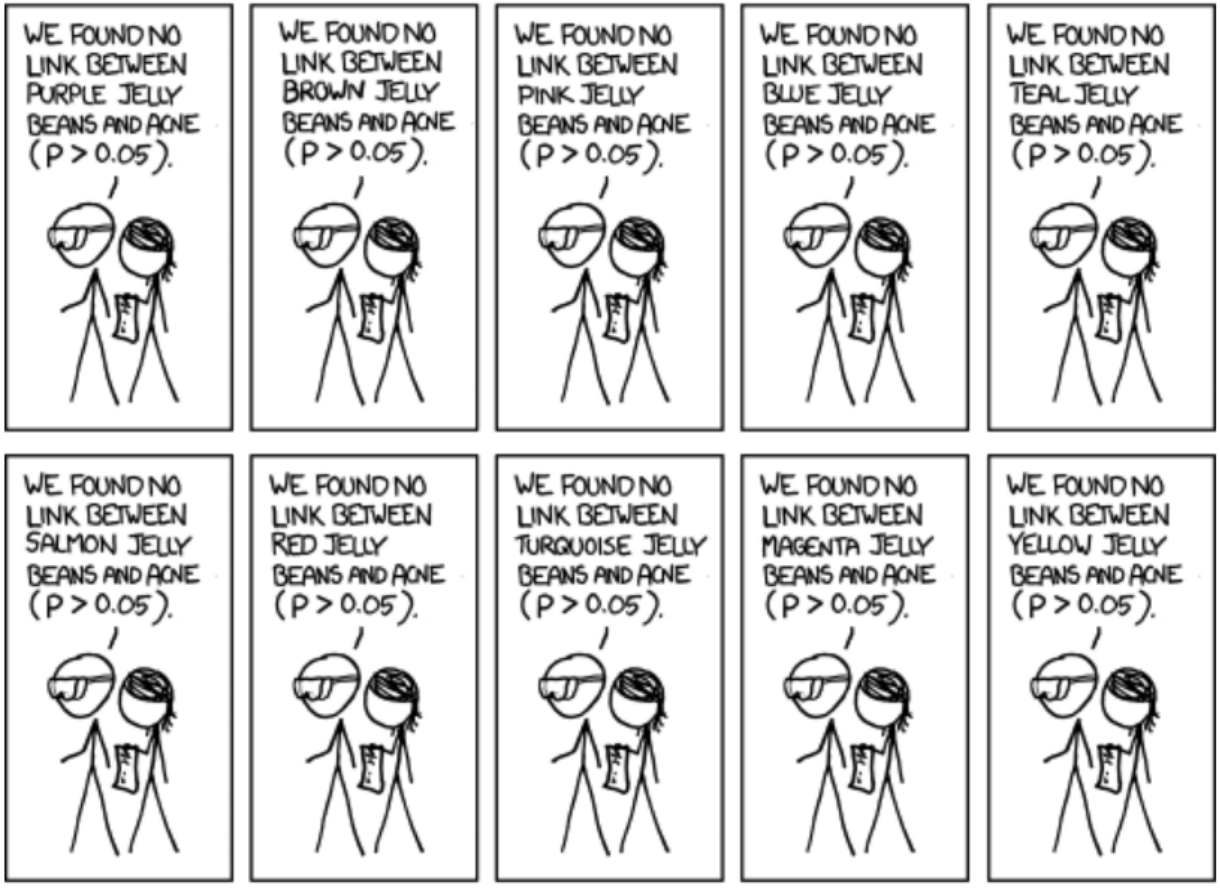
Source: xkcd

Are jelly beans causing acne?



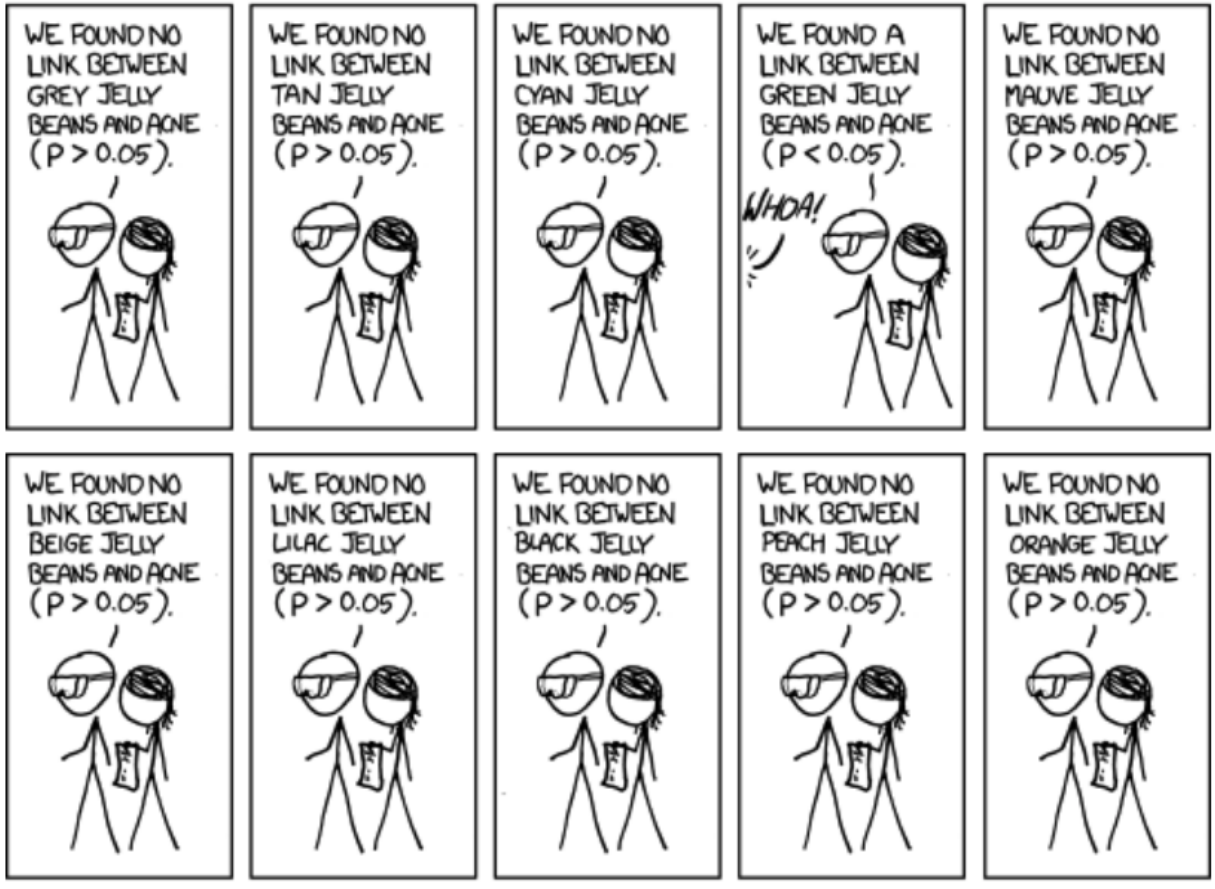
Source: xkcd

Maybe a specific color?



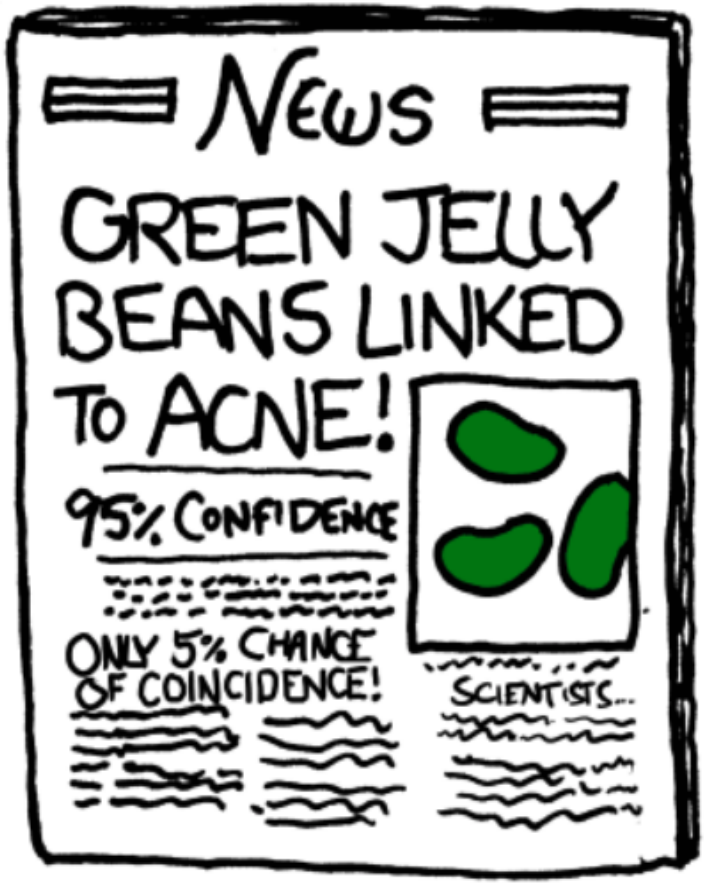
Source: xkcd

Maybe a specific color?



Source: xkcd

And finally...

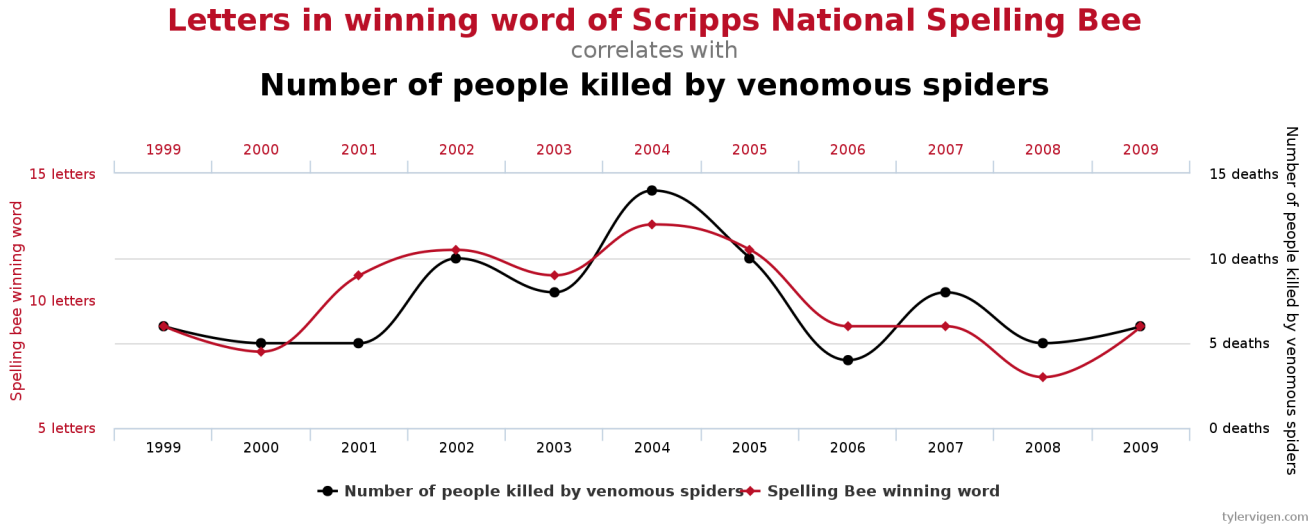


Source: xkcd

👋 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.

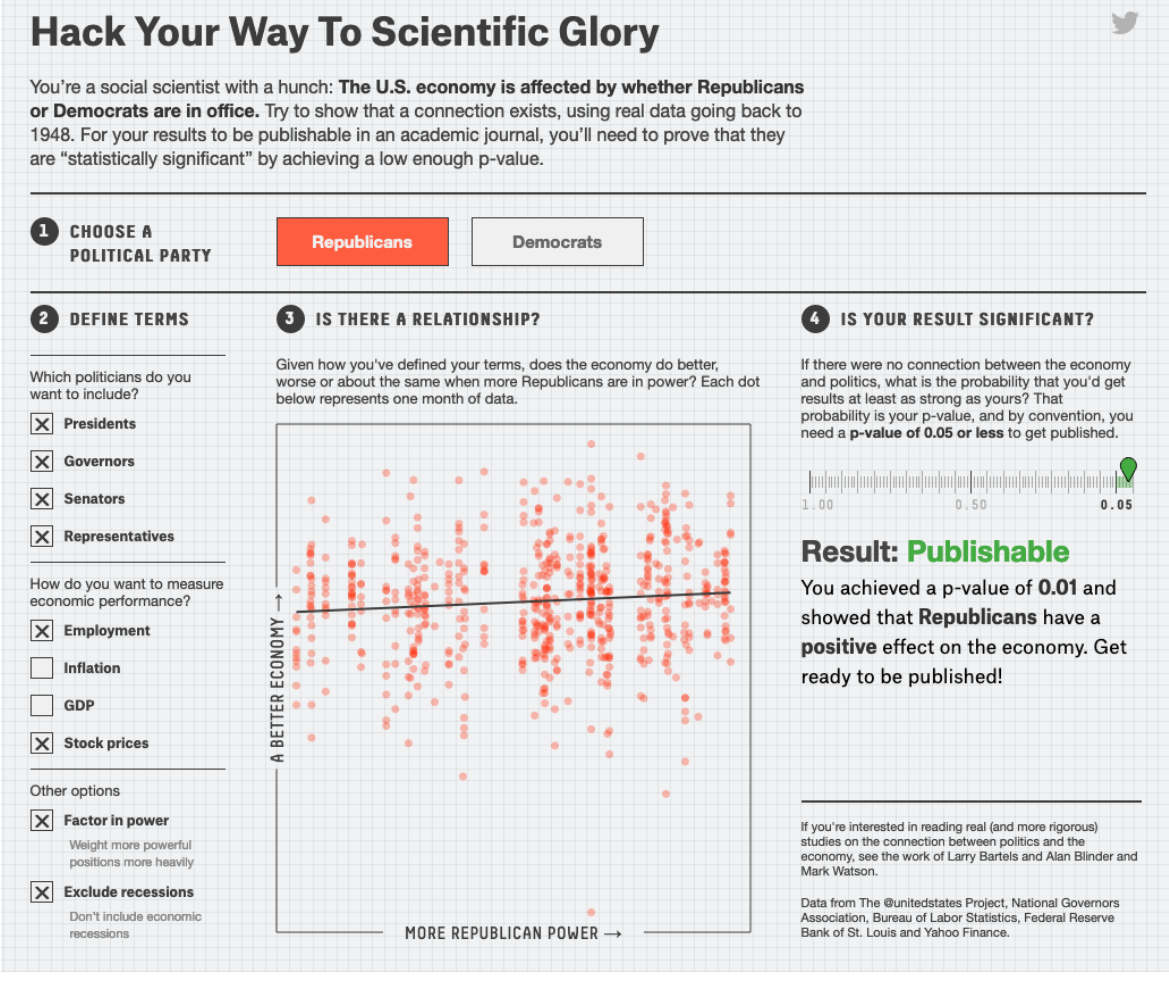
More on P-Hacking

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 Ecy? **Both!!** (see link)



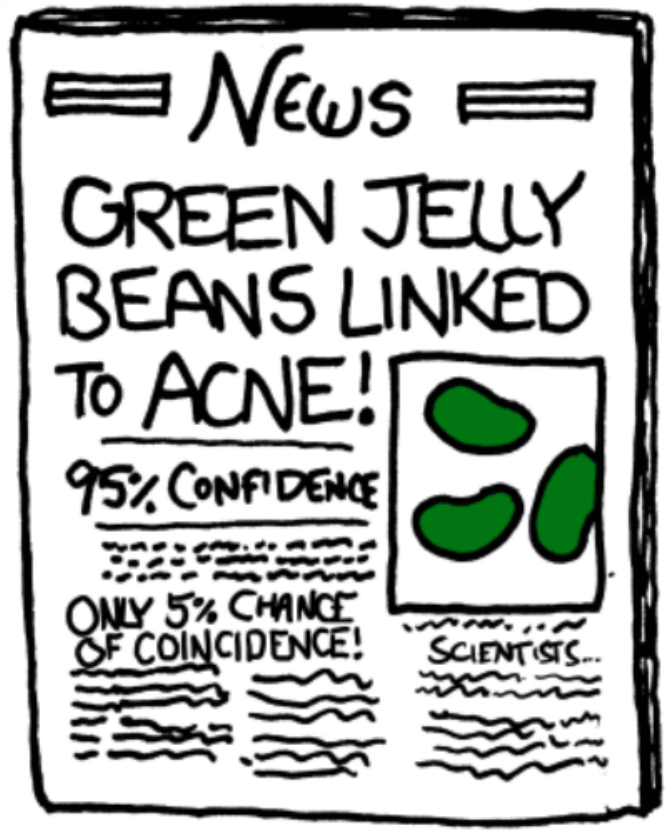
This bring us to "HARKing"

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!

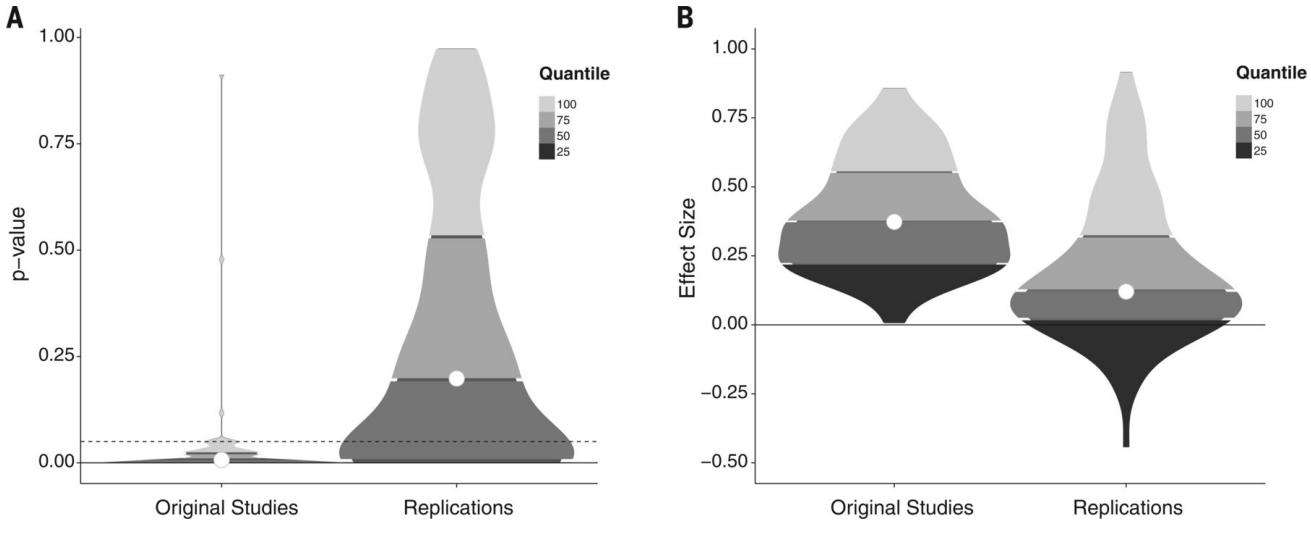
👋 If you want to know more, have a look here and the references therein.



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).

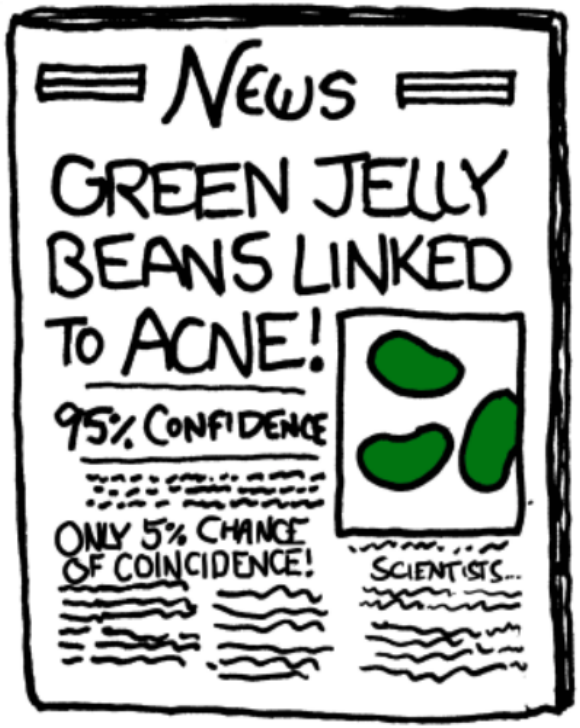
👋 If you want to know more, take a look here and the references therein.

⚠️ 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