Week 1: Introduction to Statistics

Unit 3: Use and Abuse of Numbers

openSAP
open.sap.com

THE BEST RUN SAP

Use and Abuse of Numbers

Introduction

- Although numbers don’t lie, they can be used to mislead with half-truths. This is known as the “abuse (or misuse) of statistics”.
- To be able to interpret data, it is important that you are familiar with the basics of statistical misuse. In this presentation, you’ll review some of the most common forms.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2685008/

Cherry picking

- Often, when a company promotes a product, they will undertake studies to “prove” the product’s effectiveness.
- So the company could be very selective and cherry-pick the results.

Use and Abuse of Numbers

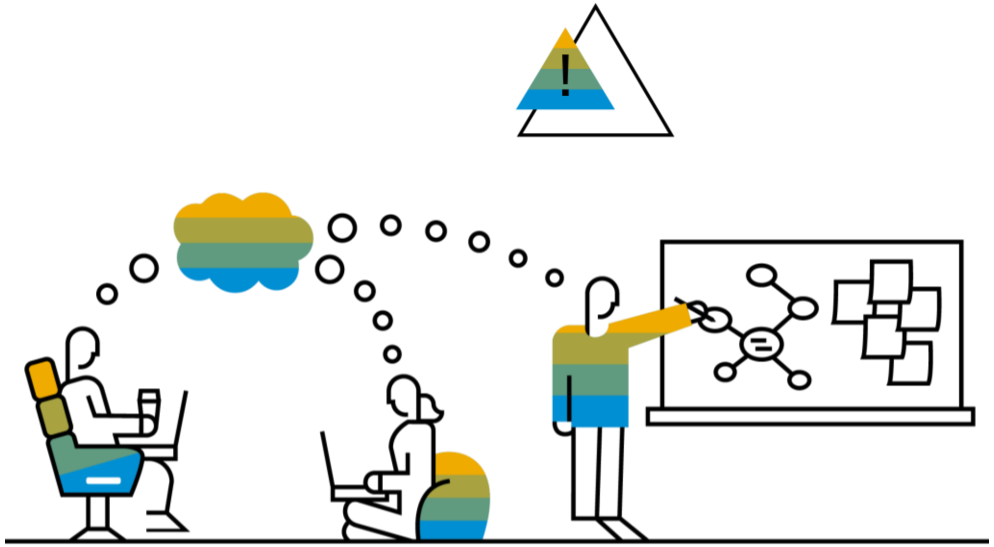
Loaded questions

- The manner in which questions are phrased can have a massive impact on the way an audience answers them.
- Specific wording patterns have a persuasive effect, and influence respondents to answer in a predictable manner.

Use and Abuse of Numbers

Overgeneralization

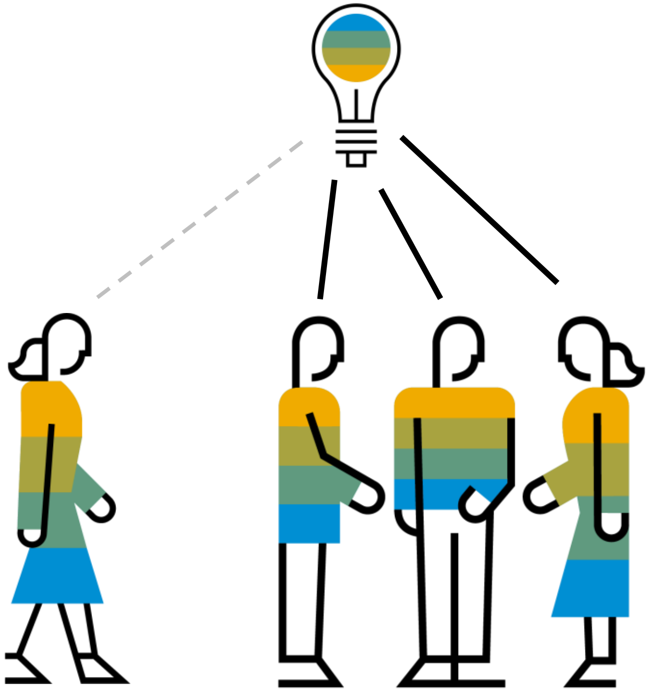
- **Overgeneralization** is a logical fallacy that occurs when a conclusion about a group is drawn from an unrepresentative sample, especially a sample that is too small or too narrow.



https://rampages.us/noelta/tag/overgeneralization/

Biased samples

- Sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others.



Use and Abuse of Numbers

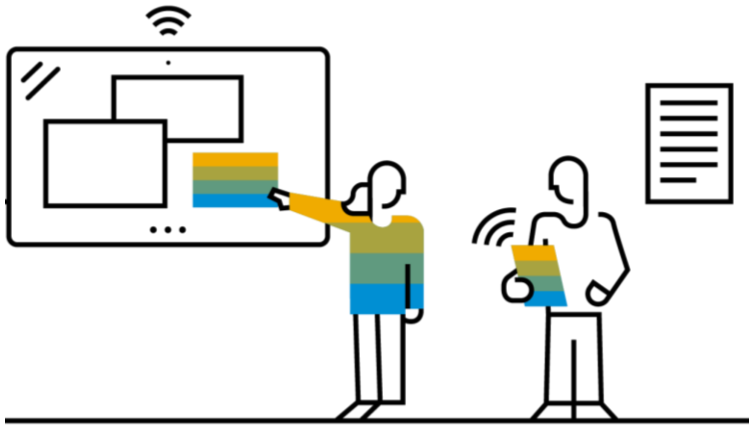
Classic example

- On election night of the 1948 presidential election, the Chicago Tribune printed the headline DEWEY DEFEATS TRUMAN. Truman won!
- In the morning, the grinning president-elect, Harry S. Truman, was photographed holding a newspaper bearing this headline.
- The reason the Tribune was mistaken was due to the results of a biased phone survey.

Use and Abuse of Numbers

Misreporting estimated error

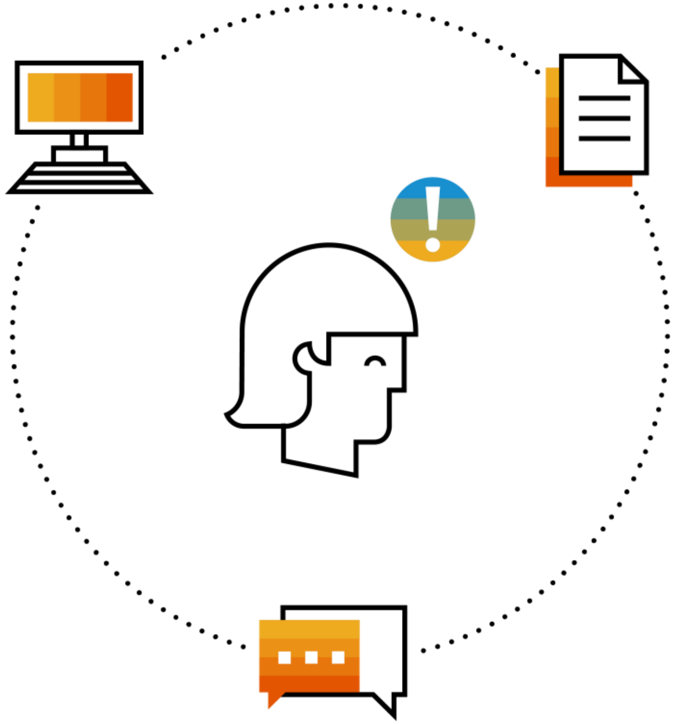
- If you want to know how 1 million people feel about a topic, it is impractical to ask all of them. Therefore, you choose a random sample.
- The confidence is the "plus or minus" figure often quoted for statistical surveys.
  - For example, a survey might have an estimated error of ±5% at 95% confidence.
- The smaller the estimated error, the larger the required sample, at a given confidence level.
- Many people might assume, that if the confidence figure is omitted, then there is a 100% certainty that the true result is within the estimated error. Of course, this is not mathematically correct.



Use and Abuse of Numbers

Correlation and causation

- In statistics, many statistical tests calculate the correlation between variables, and when two variables are found to be correlated, it is tempting to assume that this shows that one variable causes the other.
- However, correlation does not imply causation!!



https://www.quackwatch.org/01QuackeryRelatedTopics/emf.html

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Statistical vs. practical significance

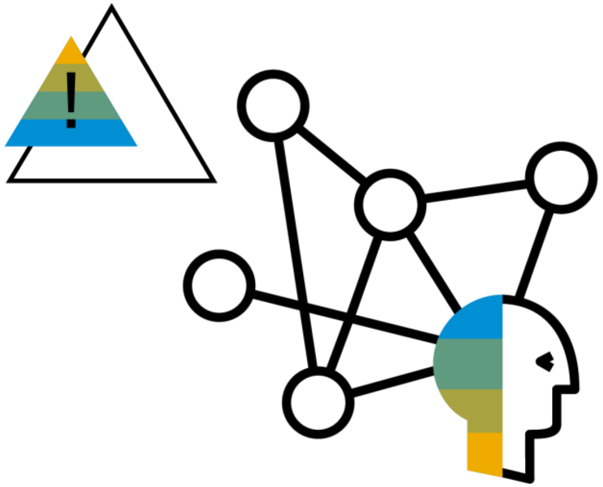
- **Statistical significance** is concerned with whether a research result is due to chance or sampling variability.
- **Practical significance** is concerned with whether the result is large enough to be of value in the real world.

Panel from a 2011 xkcd cartoon explaining p-hacking, in which scientists look for relationships between many colors of jelly beans and acne, and find a p value <0.05 only for green ones.

https://www.explainxkcd.com/wiki/index.php/882:_Significant

Data dredging

- Data dredging (sometimes called “data fishing”, “data snooping”, and “p-hacking”) is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect.
- “p-hacking” is when a data scientist analyses and presents the data in a way that supports pre-conceived answers.
  - They know that by selectively munging, binning, constraining, cleansing, and sub-segmenting data, they can get it to tell almost any story or validate almost any “fact”.



https://en.wikipedia.org/wiki/Data_dredging

https://infocus.dellemc.com/william_schmarzo/management-challenge-p-hacking/

https://www.nngroup.com/articles/understanding-statistical-significance/

https://www.nngroup.com/articles/probability-theory-and-fishing-significance/

Use and Abuse of Numbers

Summary

- We expect statistics should make data easier for us to understand.
- Unfortunately, it’s easy for statistics to be used in a misleading way to trick the casual observer into believing something other than what the data shows.
- This misuse of statistics occurs when a statistical argument asserts a falsehood.
- In some cases, the misuse may be accidental.
- In others, it is purposeful and is designed to trick us into believing a lie.
- This presentation will hopefully help you recognize the common forms of misuse.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE’s or its affiliated companies’ strategy and possible future developments, products, and/or platforms, directions, and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

THE BEST RUN SAP