Week 1: Introduction to Statistics

Unit 1: Introduction to Statistics

openSAP Course Overview

Weeks 1 through 6

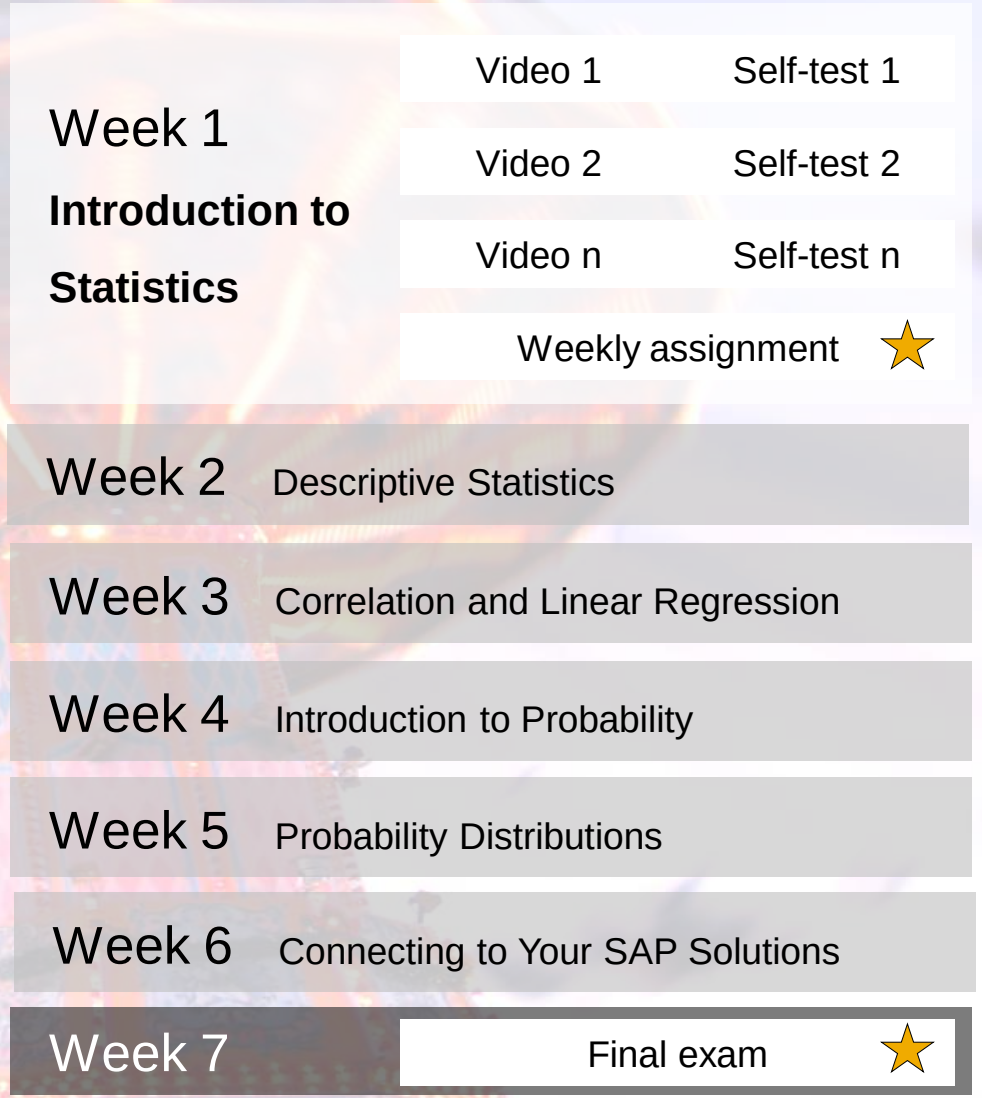
- 4-7 video units and self-tests
- 1 weekly assignment
- Online discussion forum (collaborate, ask questions)
- ~3-4 hours of effort each week

Week 7

- Final exam

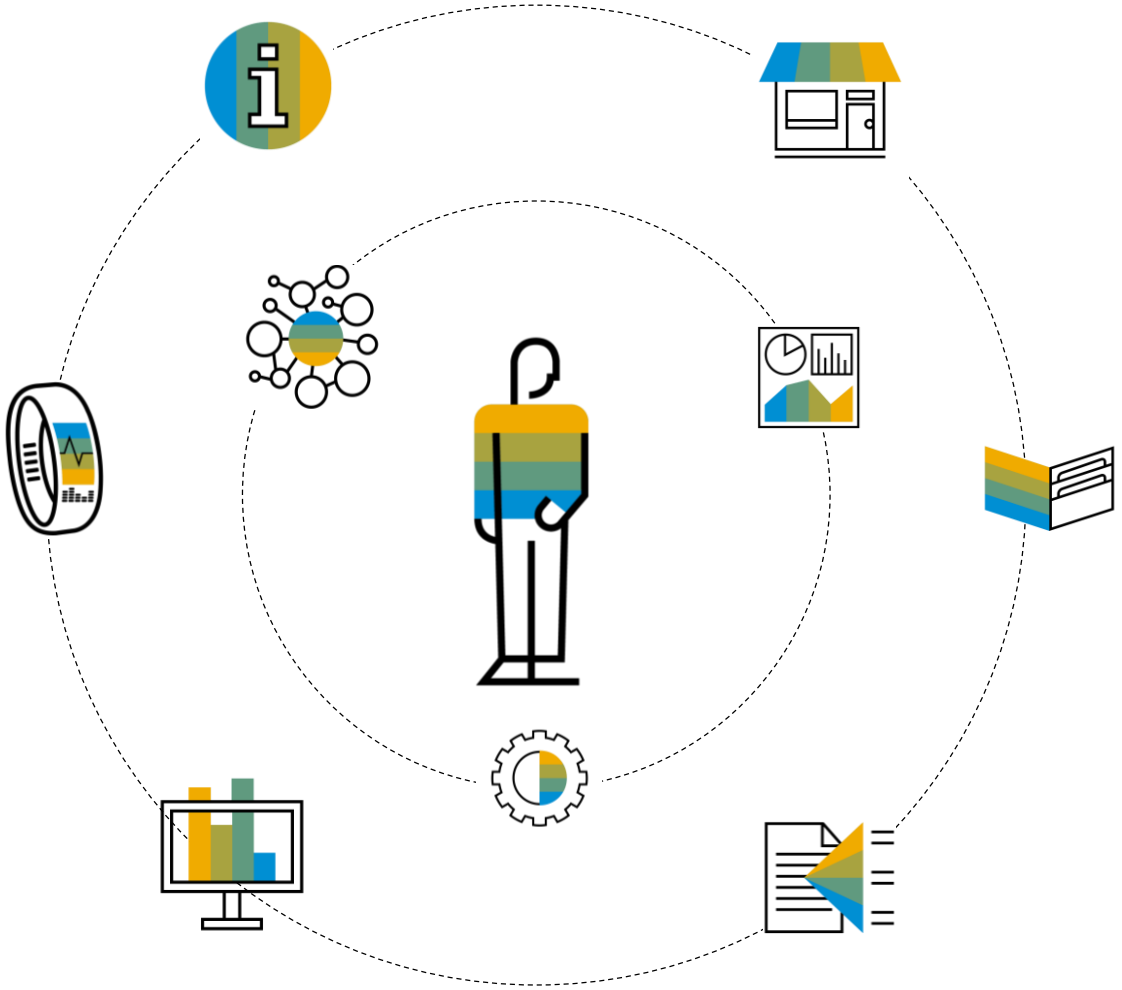
Record of achievement

- Collect at least 50% of the total points available in all online tests during the course



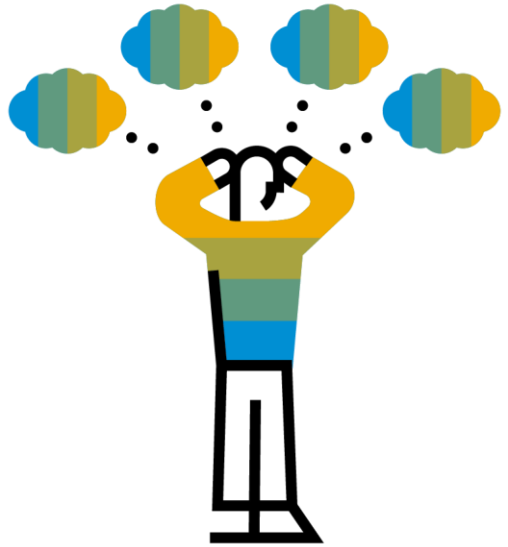
Numbers are everywhere!

- During every hour of every day, we make decisions and judgments based on data. For example, a house purchase …
  - Location
  - Size of town
  - Proximity to services, shops, and sea
  - Crime rates
  - Property prices
  - Size and number of rooms
  - Condition of house



How do I know?

- Crime has gone down/up
- By stopping eating meat you will improve your health and save the environment
- Climate change cannot be true because we had a cold winter!
- Our government has invested 14% more in social services



How do we know? How do we test the claims?!

What is a statistical problem?

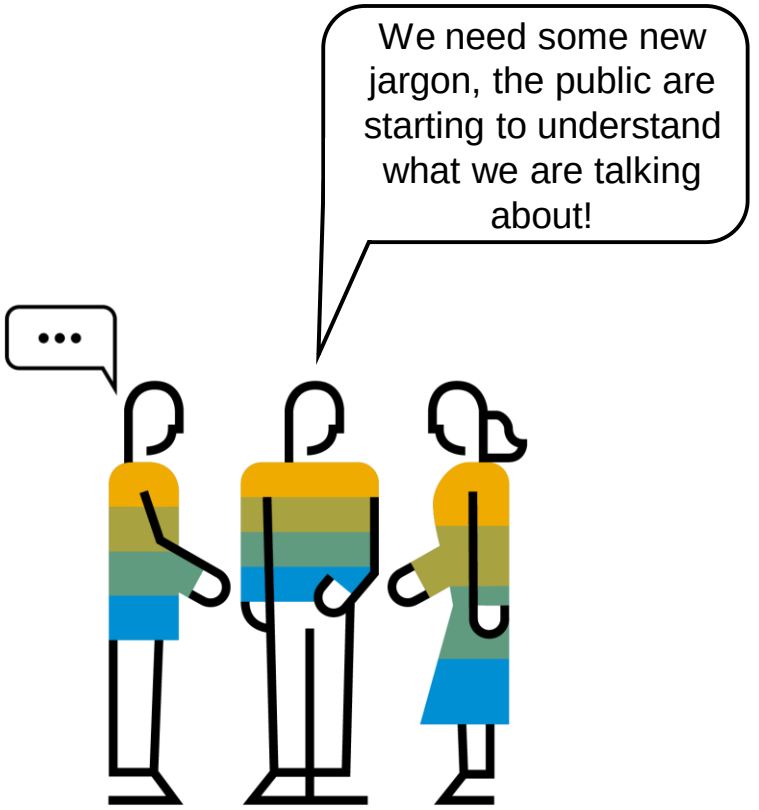
- How much sleep does the average person get?
- Is there a difference by age/gender/ethnicity etc?
- What lifestyle characteristics influence sleep quantity and quality?
- How do I test whether a sleep intervention is effective or not?

What is a statistical problem in business?

- How much profit are we making?
- How many product defects are we discovering? What's the trend?
- Has the process change led to a significant increase or decrease in employee satisfaction?
- What is our customer churn rate?

Some key statistical terms

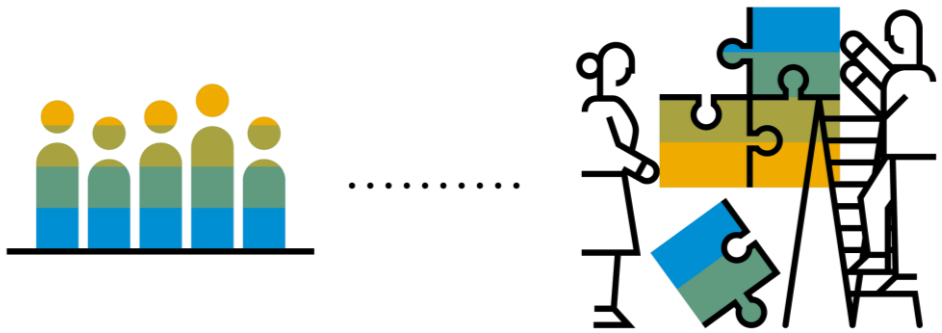
- Population versus sample
- Randomness
- Descriptive statistics
- Distributions
- Inference
- Probability
- Correlation



Population versus sample

“**Population**" data sets and "**sample**" data sets

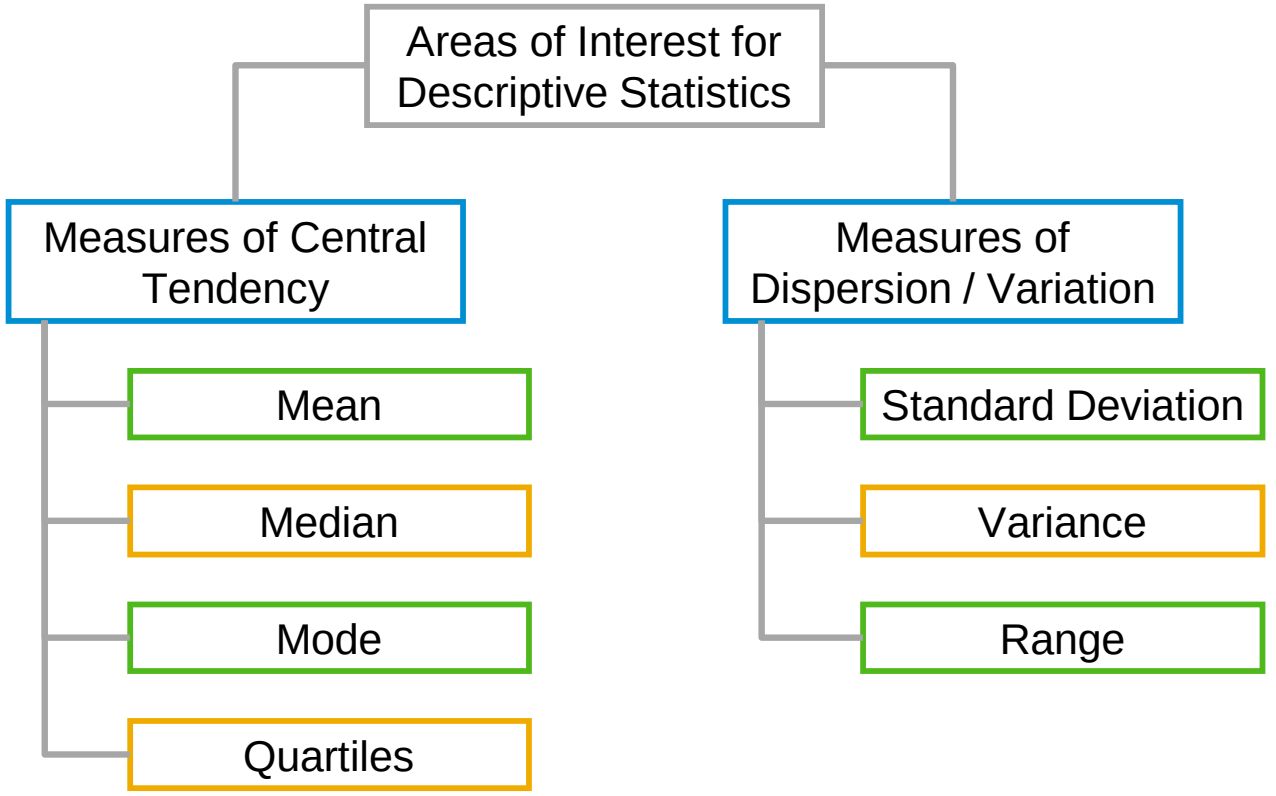
- A **population** data set contains all members of a specified group (the entire list of possible data values).
- A **sample** data set contains a part, or a subset, of a **population**. The size of a **sample** is always less than the size of the **population** from which it is taken.



https://mathbitsnotebook.com/Algebra1/StatisticsData/STPopSample.html

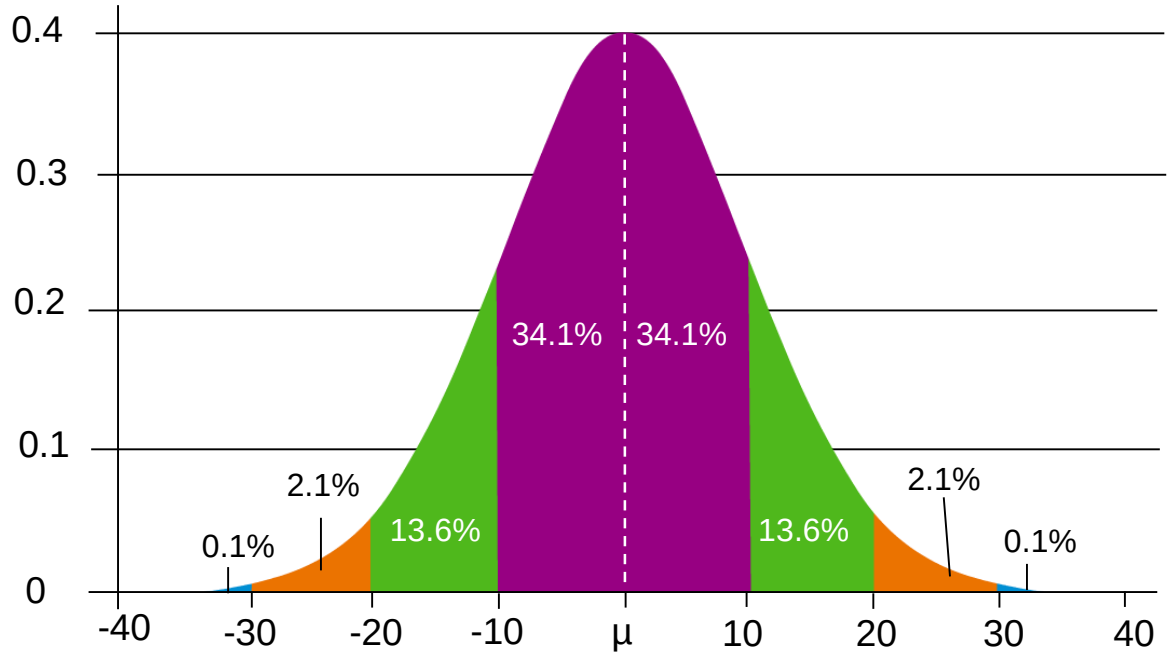
Descriptive statistics

- **Descriptive statistics** attempt to summarize a large body of data so that you can highlight key information.
- This is mainly through measures of central tendency and measures of dispersion.



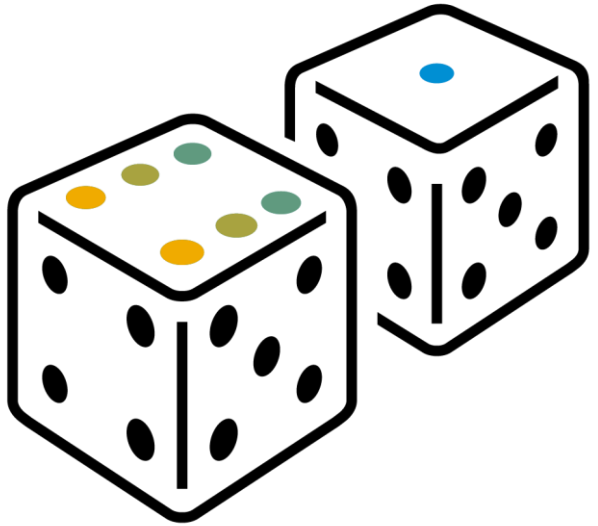
Distributions

- A *frequency* distribution provides a way of viewing all the values of a sample in a table view or a histogram.
- A *probability* distribution is a mathematical function that describes the probability of getting any particular result, such as the outcome when you roll 2 dice.



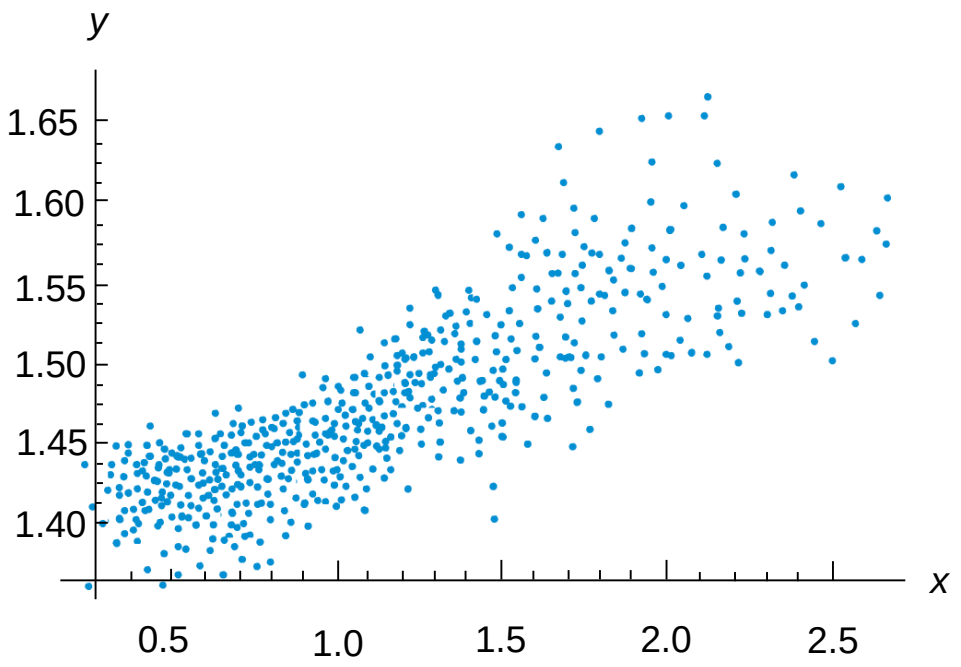
Probability

- Probability is a statistical measure of the likelihood of an event happening.
- Probability is measured between 0 and 1, so that at 0, an event definitely will not happen and at 1 it certainly will.
- Most probabilities fall between the two figures.



Correlation

- In statistics, dependence or association is any statistical relationship, whether causal or not, between two random variables.
- This statistical relationship is described as **correlation**.



Summary

- Statistics is everywhere, and everyday we are expected to make different kinds of statistical estimations.
- As part of the above, we are bombarded with different kinds of statistical claims from parties who want our vote, our time, or our money.
- We will learn how to evaluate those claims.
- There are some key statistical terms, which we need to understand to be able to develop our skills in understanding statistics.
- The first terms we need to understand are population vs. sample, randomness, descriptive statistics, distributions, probability, and correlation.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

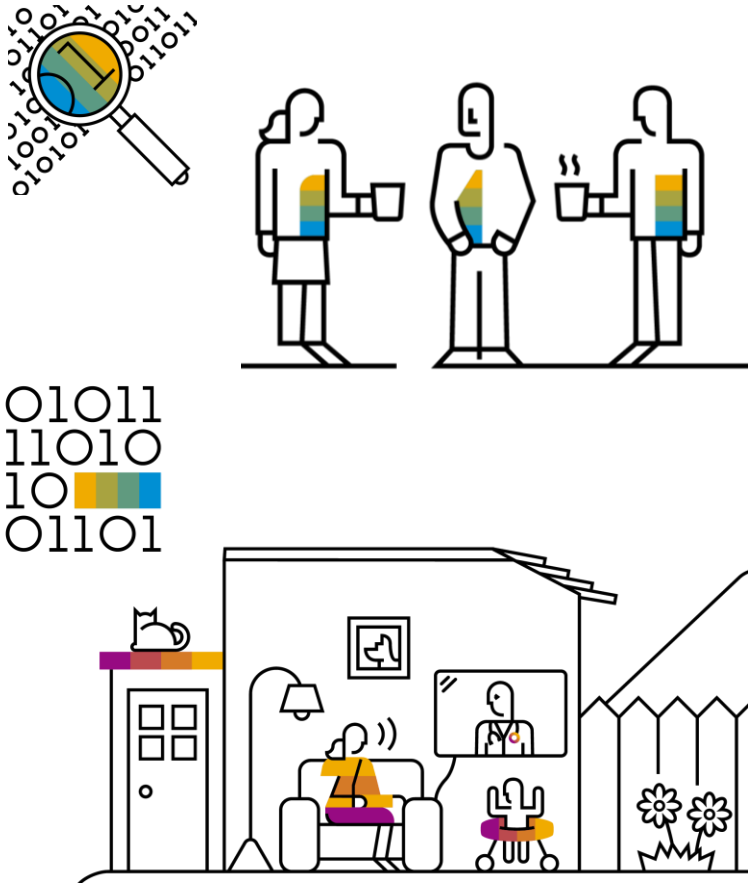
THE BEST RUN SAP®

Week 1: Introduction to Statistics

Unit 2: Numbers in Everyday Life

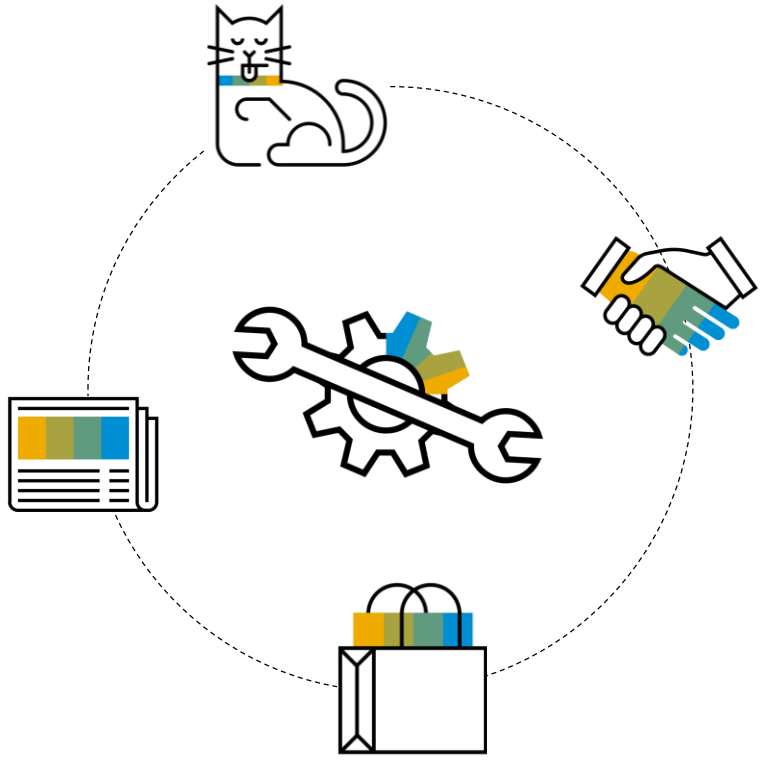
Introduction

- In this unit, we'll look at everyday examples where we have to evaluate and make judgments using numbers and statistics.
- In many cases, we hardly think that we're doing statistics, but we are. There are countless examples from our everyday lives, but I've chosen 3 from the day I was building this unit.
- For each of the claims made, we'd like you to think about how you'd evaluate and test them yourself. After each one, we'll compare notes with a few things I thought about.



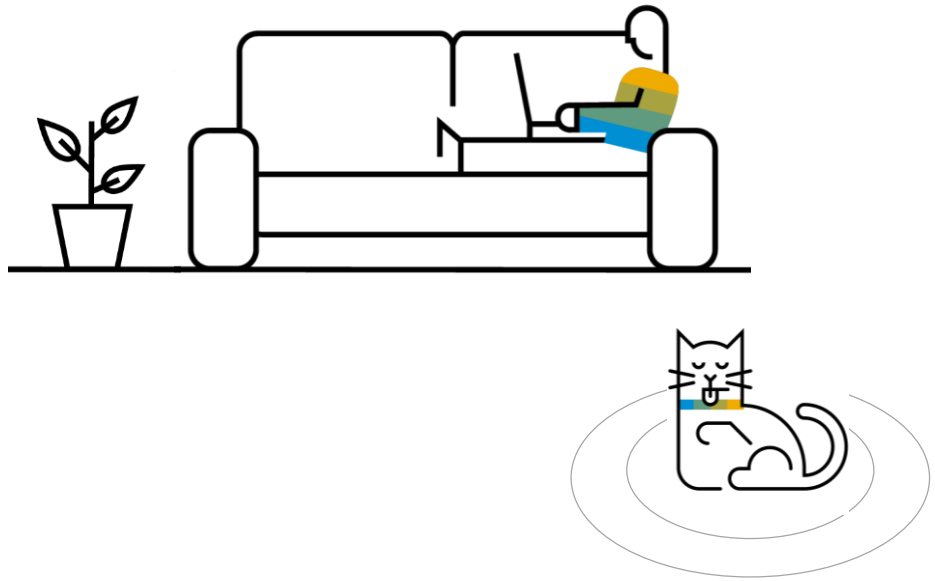
Statistics every day

- Let’s look at some statistical claim examples from my day …
  - Advertising – “9 out of 10 cat owners say their cats prefer x brand of cat food”
  - Newspaper articles – “*“Bill Gates says poverty is decreasing. He couldn’t be more wrong” (The Guardian – 29/01/2019)*
  - Shopping – Does it make sense to buy the BOGOF (Buy-One-Get-One-Free) offer?
- Let’s test these one by one …



Questioning the statistics

- What questions should you be raising about these claims?…
- Advertising – *“9 out of 10 cat owners say their cats prefer x brand of cat food”*
  - What sample size did they use?
  - How could they test that the cats preferred the cat food to other food?
  - Was the testing done against all top brands?
  - How could you tell that the cats didn’t simply enjoy the food because it was new and different? Will they get bored?



Questioning the statistics

- What questions should you be raising about these claims?…
- Newspaper – *"Bill Gates says poverty is decreasing. He couldn't be more wrong" (The Guardian – 29/01/19) Read the article first*
  - How are they defining poverty?
  - What is the baseline?
  - Do those making these claims have some reason to show bias?
  - What extra information do you need to have to legitimately make the above claims?

https://www.theguardian.com/commentisfree/2019/jan/29/bill-gates-davos-global-poverty-infographic-neoliberal

Questioning the statistics

- Which questions should you be considering?...
- Shopping – *Which bananas should I buy? A bag for £1? Individual ones for 25p each? Large ones or small ones? Organic?*
  - How many bananas do you get in a bag? Are the sizes the same as the individual ones?
  - Is the quality/ripeness comparable between them?
  - If there is a BOGOF (Buy-One-Get-One-Free) offer on the bagged bananas, would that make a difference?

Summary

- Numbers, summary statistics, and claims are part of our everyday lives and interactions.
- Those selling goods and services often use tactics which are not necessarily untruthful, but make it difficult to check claims.
- We are constantly battling to evaluate these claims to understand whether they are biased or truthful.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®

Week 1: Introduction to Statistics

Unit 3: Use and Abuse of Numbers

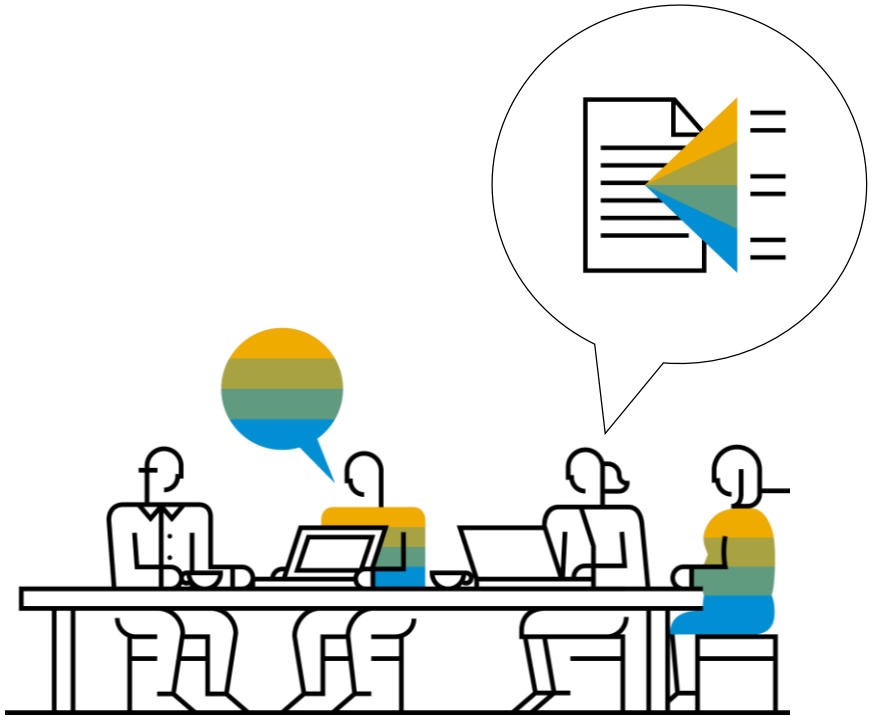
Introduction

- 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.

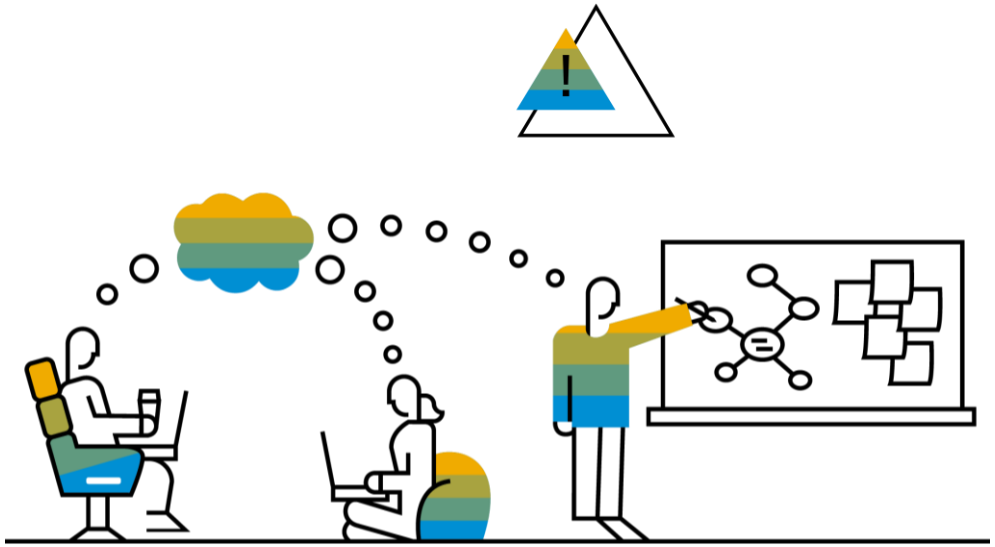


Loaded questions

- 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.

Overgeneralization

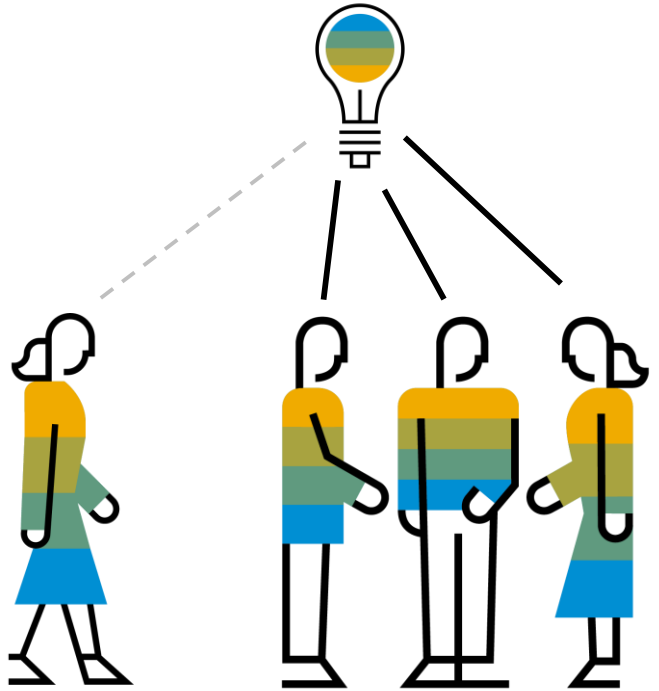
- **Overgeneralization** is a logical fallacy that occurs when a conclusion about a group is drawn from an unrepresentative sample, especially a sample that is too small or too narrow.



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.

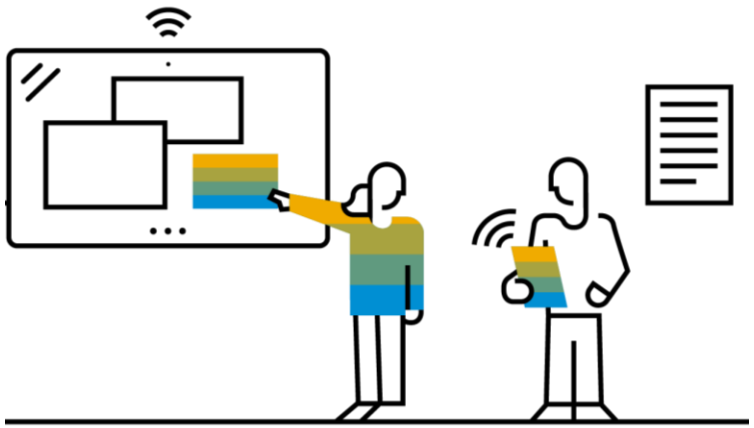


Classic example

- 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.

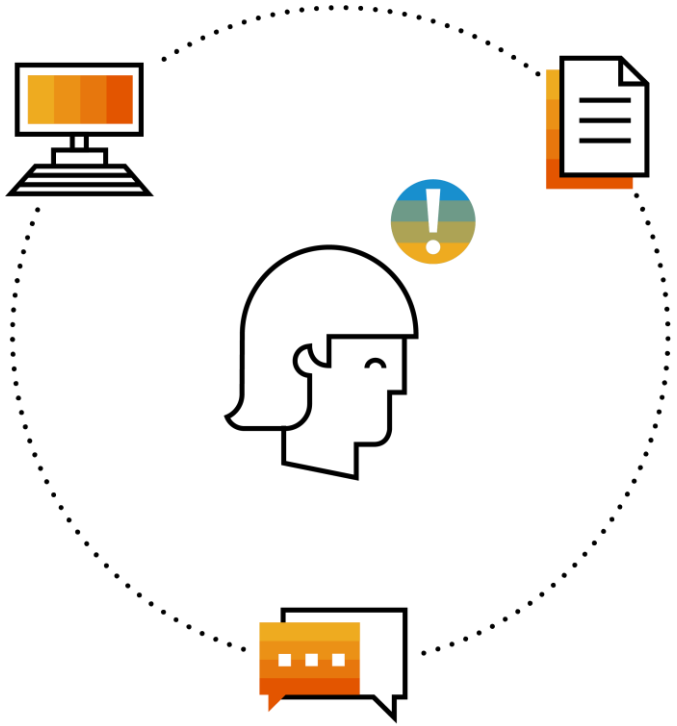
Misreporting estimated error

- 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.



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

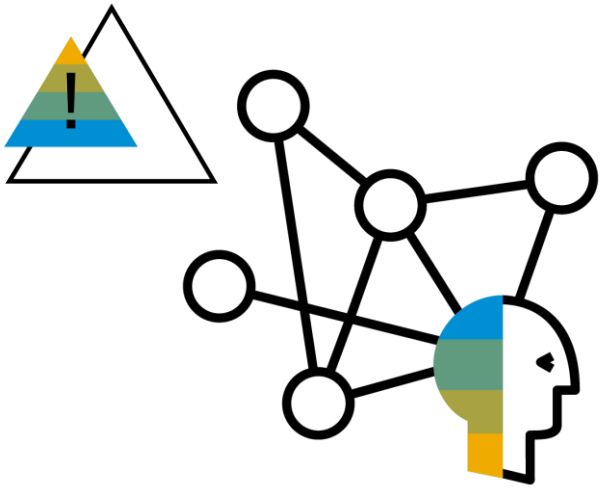
- **Statistical significance** is concerned with whether a research result is due to chance or sampling variability.
- **Practical significance** is concerned with whether the result is large enough to be of value in the real world.

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/

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 BEST RUN SAP®

Week 1: Introduction to Statistics

Unit 4: Bias

Introduction

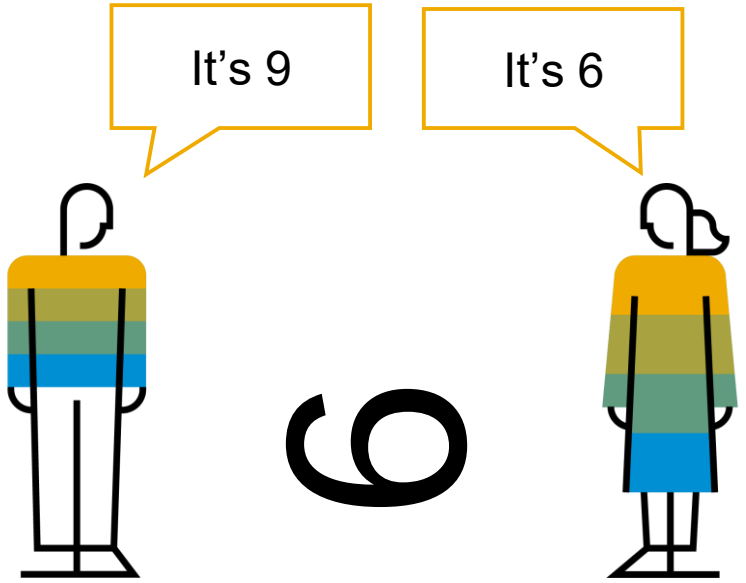
- **Bias** refers to the tendency of a measurement process to over or under-estimate the value of a population parameter.
- There are may different types of bias.

Bias

It is inevitable!

“With careful and prolonged planning, we may reduce or eliminate many potential sources of bias, but seldom will we be able to eliminate all of them. Accept bias as inevitable and then endeavor to recognize and report all exceptions that do slip thought the cracks."

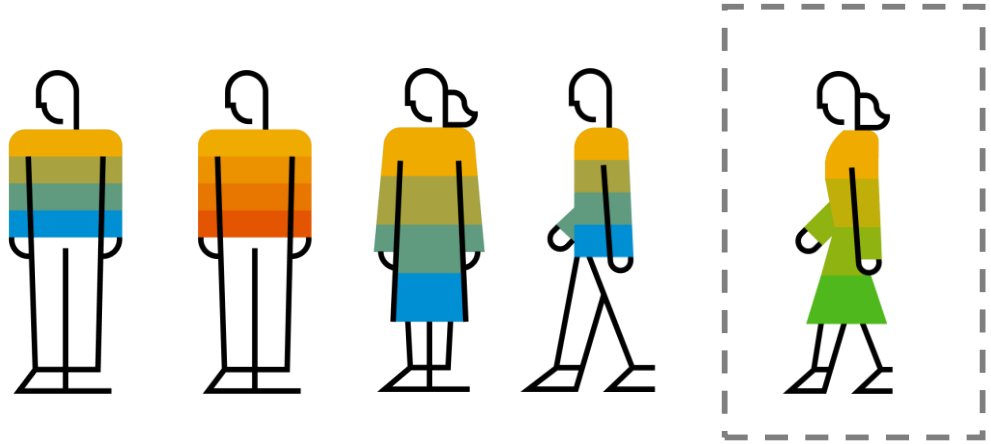
Good and Hardin (2006) Common Errors in Statistics (and How to Avoid Them), p. 113



Different types of bias

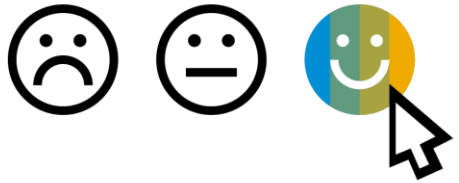
- We will look at some of the different types of bias:
  - Selection/sampling bias
  - Self-selection bias
  - Confirmation bias
  - Overfitting
- “Unlike error related to random variability, bias cannot be assessed without external knowledge of the world”

Herbert I. Weisberg (2010), *Bias and Causation: Models and Judgment for Valid Comparisons*, p. 26

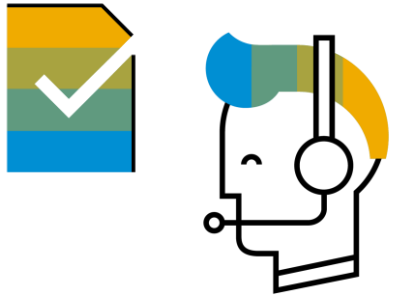


Sampling bias

Surveying of attitudes to Brexit on Facebook
(favors younger audiences)



Face-to-face surveying of attitudes to equality at a football match (self-selection)



Telephone surveying of holiday preferences

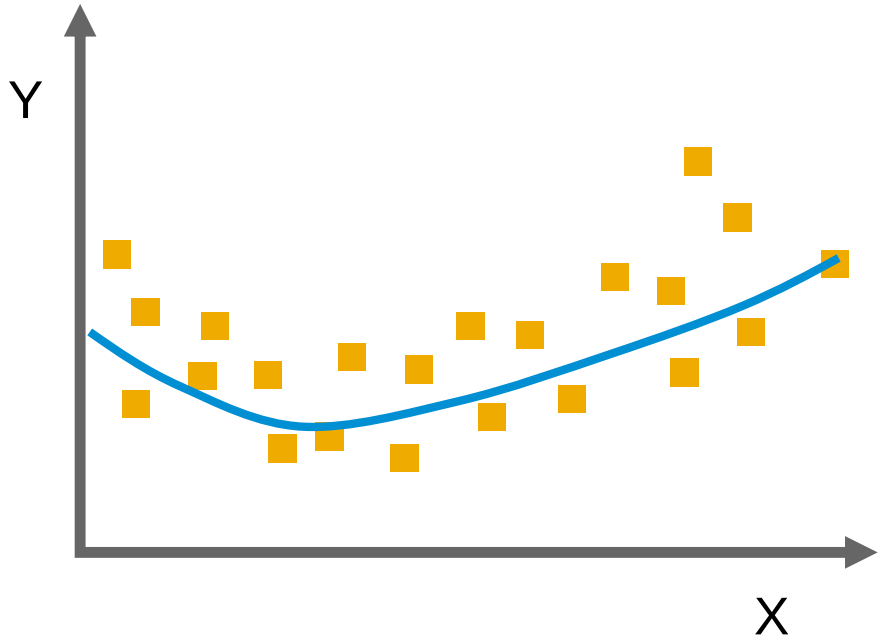
Self-selection bias

- If you use data taken from a *voluntary response sample*, i.e. the participants volunteered to take part, it becomes very difficult to avoid bias.
- The self-selected group will contain more participants with a particular set of beliefs about your study.

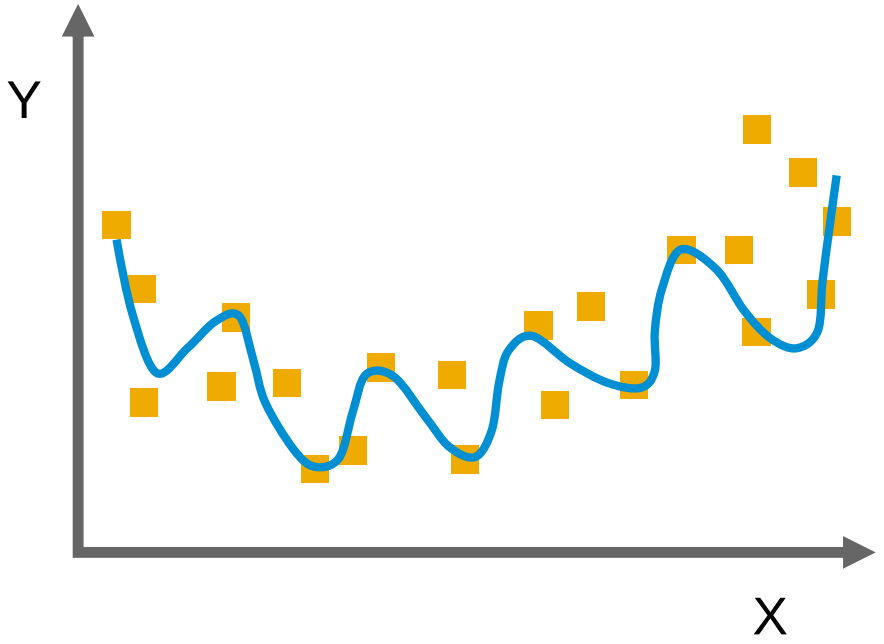
Confirmation bias

- “Confirmation bias” is one of a range of “cognitive biases” that affect how we read and interpret the insights we think we have found.
- “Cognitive bias” means that it is inbuilt into us, as humans, and how we think.
- “Confirmation bias” reflects our tendency to pick out those parts of the data and information in a way to support our previously held beliefs.

Overfitting and extrapolation



Just right!



overfitting

Summary

- It is almost impossible to avoid bias in its various forms, but an awareness of bias can help mitigate its worst effects.
- There are various forms of bias, e.g. technical, cognitive, and others, which impact what data to collect and how it should be interpreted.
- In this unit we have examined 4 key examples of bias – sampling, self-selection, confirmation, and overfitting. There are many other kinds of bias.
  See: https://en.wikipedia.org/wiki/Bias (statistics)

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®

Week 1: Introduction to Statistics

Unit 5: Different Kinds of Analytic Approaches

Qualitative vs. quantitative data

- The type of analytical approach you take depends on the type of data you have collected and the question you are answering.
- There are two types of data: **qualitative** and **quantitative**.

Qualitative Data

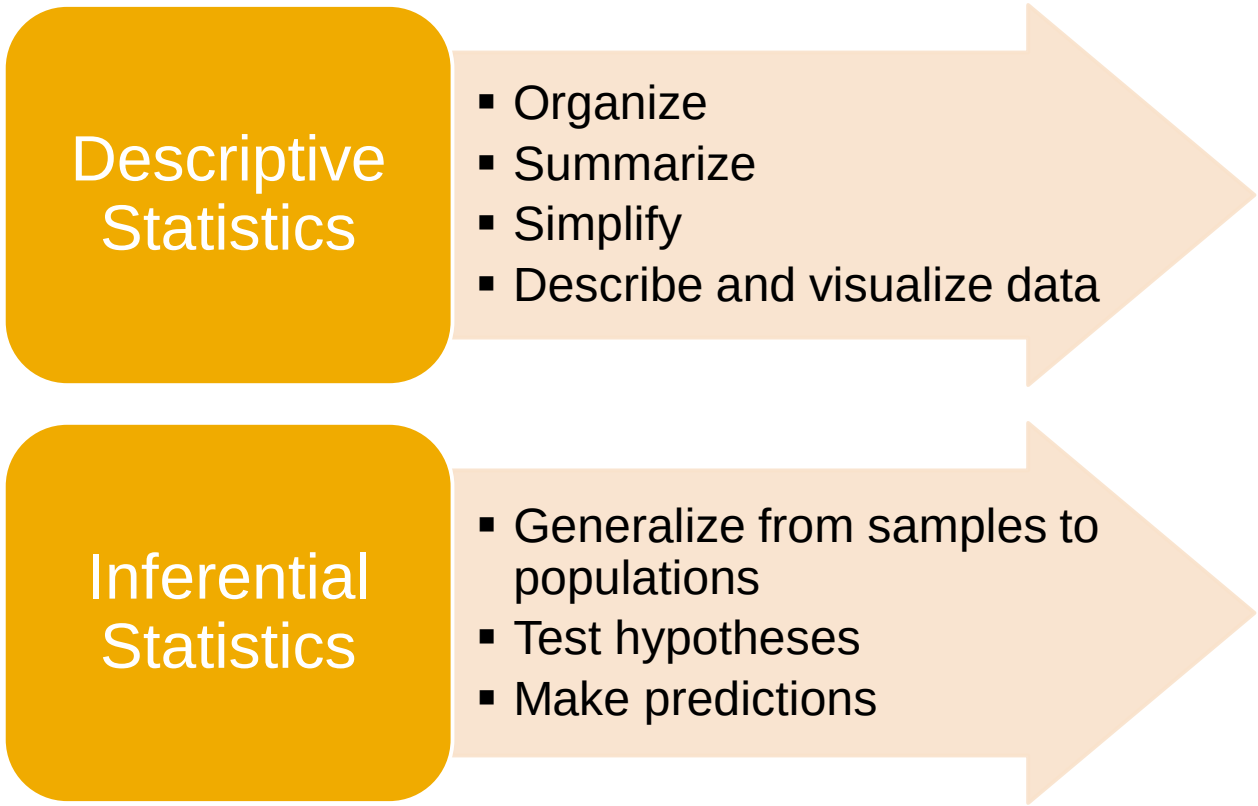
- Qualitative → Quality
- Deals with descriptions
- Data can be observed but not measured
- Colors, textures, smells, tastes, appearance, etc.

Quantitative Data

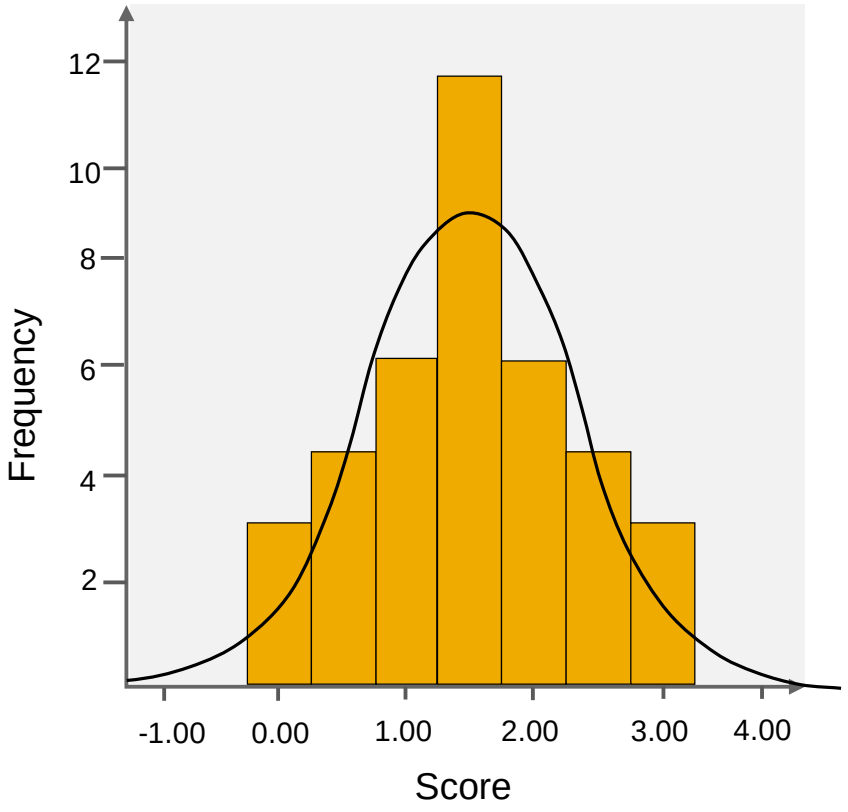
- Quantitative → Quantity
- Data that can be measured
- Length, height, area, volume, weight, speed, time, temperature, cost, etc.

Descriptive vs. inferential analysis

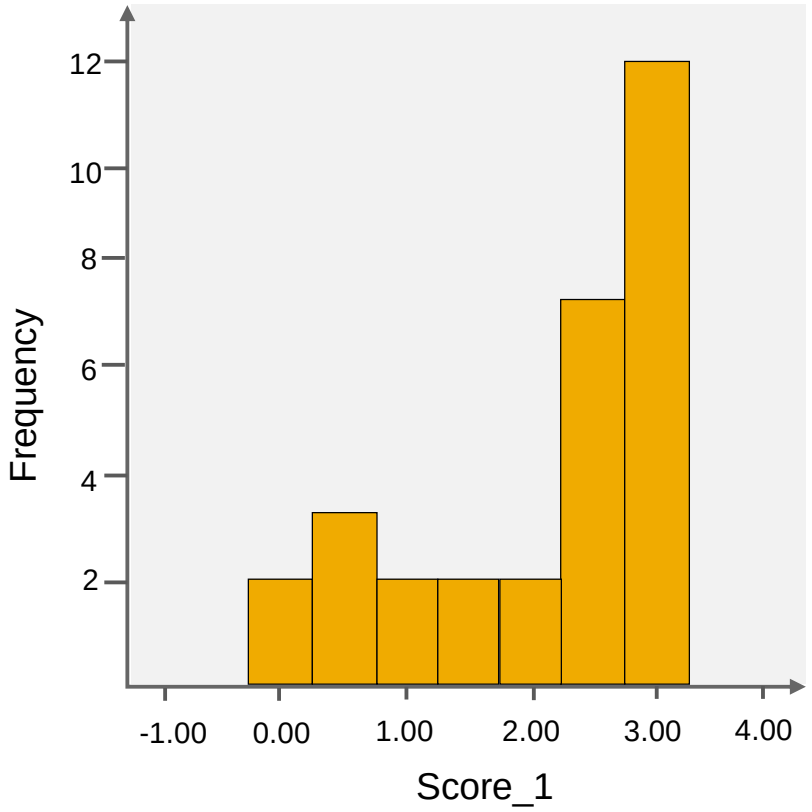
- There are two common types of analysis that are referred to as “descriptive” and “inferential”.



Normal vs. non-normal



A **normal distribution** looks like the bell curve



A **non-normal distribution**

https://cyfar.org/inferential-analysis for more information

Common parametric statistical tests

Correlation: Analyze the association between variables	
Pearson correlation	Tests for the strength of the association between two continous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not assume that data is normally distributed)
Chi-square	Tests for the strength of the association between two categorical variables

Comparison of means: Analyze the difference between the means of variables	
One sample T-test	Compares the mean of a sample to a pre-specified value and tests for a deviation from that value
Paired T-test	Tests for the difference between two related variables
Independent T-test	Tests for the difference between two independent variables
Analysis of Variance (ANOVA)	Tests for the difference between group means in a sample after any other variance in the outcome variable is accounted for

Regression: Analyze how change in one variable predicts change in another variable	
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predicts the level of change in the outcome variable

Some of these tests will be described in more detail later in this course.

https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740

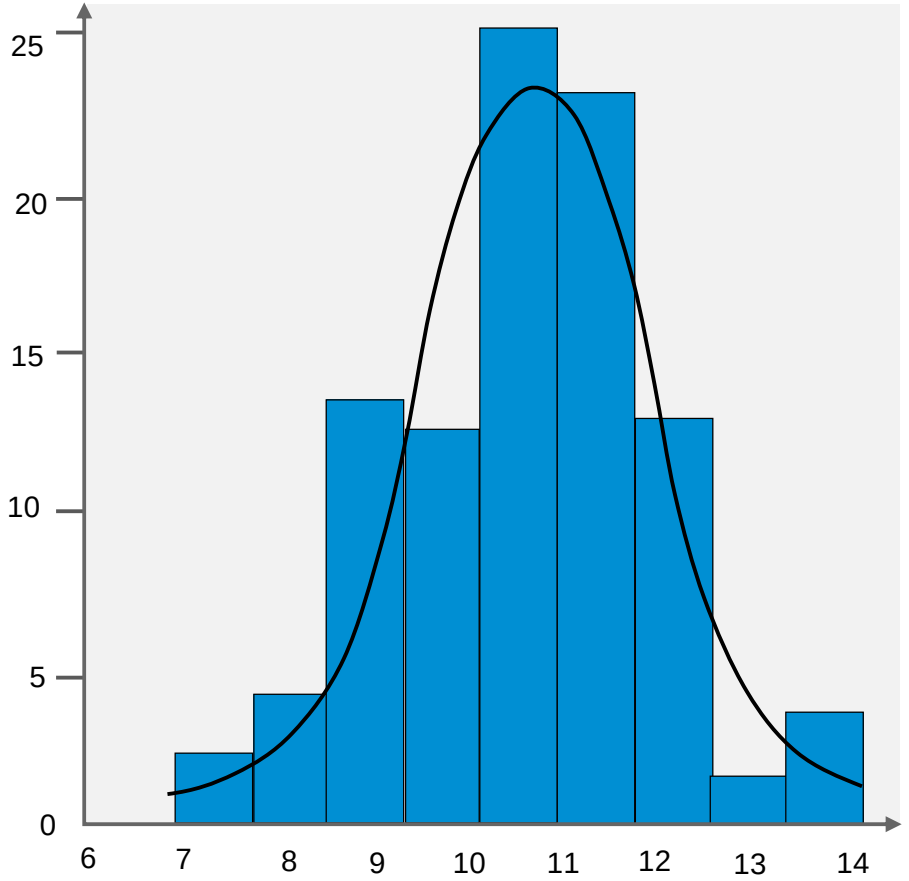
Common nonparametric statistical tests

Nonparametric: used when the data does not meet the assumptions required for parametric tests	
Sign test	Tests if two related variables are different – ignores the magnitude of change, only takes into account direction. The sign is an alternative to one sample T-test or a paired T-test.
Wilcoxon rank-sum test	Tests for the difference between two independent variables – takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for the difference between two related variables – takes into account the magnitude and direction of difference

For more information, see
http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric_print.html

Choosing parametric or nonparametric statistical tests

- It can sometimes be difficult to assess whether a continuous outcome follows a normal distribution and whether a parametric or nonparametric test is appropriate.
- The most practical approach to assessing normality involves analyzing the distribution of the outcome in the sample using a histogram.



https://en.wikipedia.org/wiki/Goodness_of_fit for more information

Summary

- **Descriptive analysis** informs you about the basic qualities of the data.
- **Inferential analysis** uses statistical tests to analyze whether a pattern in the data is due to chance or due to the intervention that is observed, and what the strength of that relationship is.
- In this course, you'll learn about some of these descriptive and inferential statistical techniques, and how these techniques can be misused.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®