جامعة الطائف
TAIF UNIVERSITY

# PRINCIPLES OF DATA SCIENCE

2021 - 2022

**Assoc. Prof. Dr. Salha Alzahrani**

**Department of Computer Science**

TU

جامعة الطائف
TAIF UNIVERSITY

## CHAPTER 7: BASIC STATISTICS

- WHAT ARE STATISTICS?
- HOW TO OBTAIN AND SAMPLE DATA
- THE MEASURES OF CENTER, VARIANCE, AND RELATIVE STANDING
- NORMALIZATION OF DATA USING THE Z-SCORE
- THE EMPIRICAL RULE

# WHAT ARE STATISTICS?

# What are statistics?

- Statistics is the field where we try to explain and model the world around us.
- To do that, we have to take a look at the population.
- We can define a population as the entire pool of subjects of an experiment or a model.
- Essentially, your population is who you care about. For example,
  - If you are trying to test if smoking leads to heart disease, your population would be the smokers of the world.
  - If you are trying to study teenage smoking problems, your population would be all teenagers.
- Now, consider that you want to ask a question about your population, for example, if your population is all of your employees (assume that you have over 1,000 employees), perhaps you want to know what percentage of them use illicit drugs. The question is called a parameter.
- We can define a parameter as a numerical measurement describing a characteristic of a population.

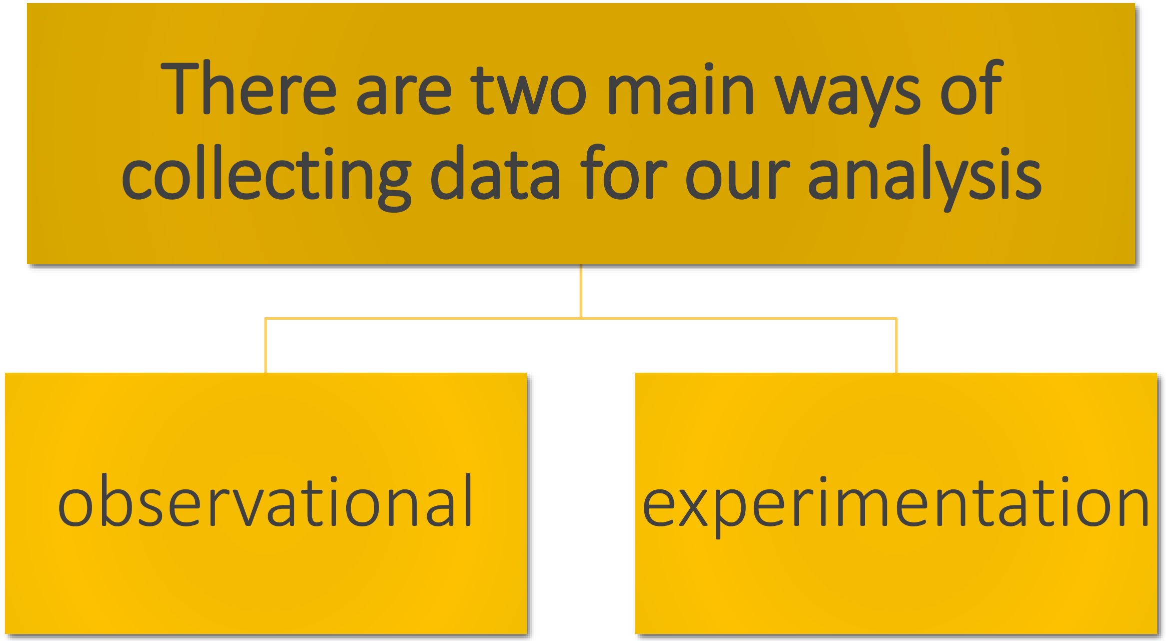
# What are statistics?

- For example, if you ask all 1,000 employees and 100 of them are using drugs, the rate of drug use is 10%. The parameter here is 10%.
- However, let's get real, you probably can't ask every single employee whether they are using drugs. What if you have over 10,000 employees? It would be very difficult to track everyone down in order to get your answer. When this happens, it’s impossible to figure out this parameter. In this case, we can estimate the parameter.
  - First, we will take a sample of the population.
  - We can define a sample of a population as a subset of the population.
  - So, we perhaps ask 200 of the 1,000 employees you have. Of these 200, suppose 26 use drugs, making the drug use rate 13%. Here, 13% is not a parameter because we didn't get a chance to ask everyone. This 13% is an estimate of a parameter. That's a statistic!
- We can define a statistic as a numerical measurement describing a characteristic of a sample of a population.
- This is necessary because you can never hope to give a survey to every single teenager or to every single smoker in the world.

# What are statistics?

That's what the field of statistics is all about taking samples of populations and running tests on these samples.

# HOW DO WE OBTAIN AND SAMPLE DATA?



There are two main ways of collecting data for our analysis

observational

experimentation

- Both these ways have their pros and cons.
- They each produce different types of behavior and, therefore, warrant different types of analysis.

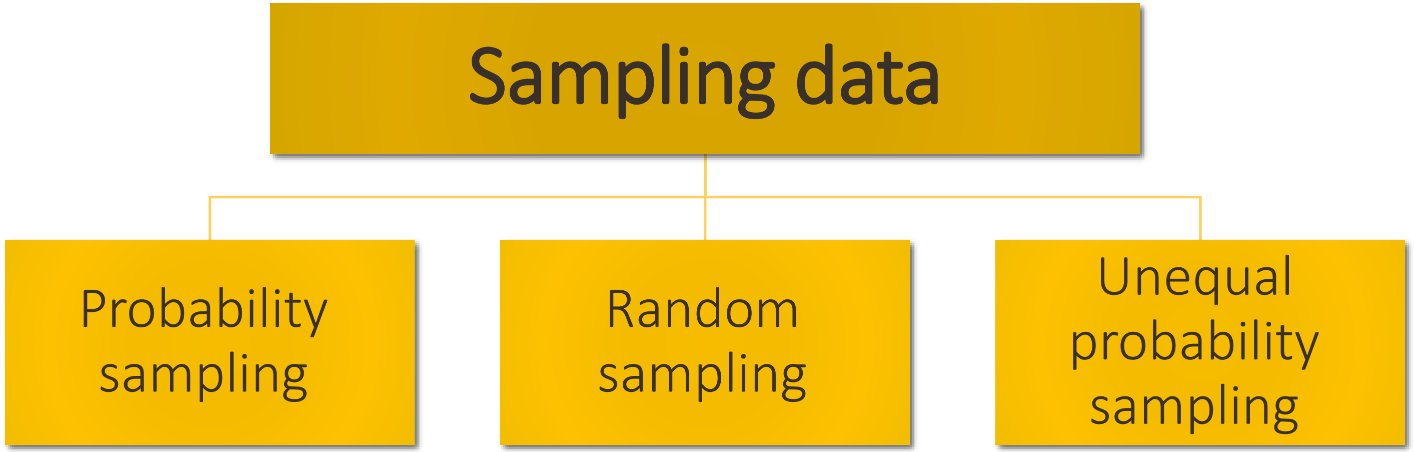
# Obtaining data

## Observational

- An observational study consists of measuring specific characteristics but not attempting to modify the subjects being studied.
- For example, you have a tracking software on your website that observes users' behavior on the website, such as length of time spent on certain pages and the rate of clicking on ads, all the while not affecting the user's experience, then that would be an observational study.

## Experimental

- An experiment consists of a treatment and the observation of its effect on the subjects called experimental units.
- This is usually how most scientific labs collect data. They will put people into two or more groups (usually just two) and call them the control and the experimental group.
  - The control group is exposed to a certain environment and then observed.
  - The experimental group is then exposed to a different environment and then observed.



Sampling data

Probability sampling

Random sampling

Unequal probability sampling

# Probability sampling

- Probability sampling is a way of sampling from a population, in which every person has a known probability of being chosen but that number *might* be a different probability than another user.
- The simplest (and probably the most common) probability sampling method is random sampling.

# Random sampling

- Suppose that we are running an A/B test and we need to figure out who will be in group A and who will be in group B. There are the following three suggestions from your data team:
  - Separate users based on location: Users on the west coast are placed in group A, while users on the east coast are placed in group B
  - Separate users based on the time of day they visit the site: Users who visit between 7 p.m. and 4 a.m. get site A, while the rest are placed in group B
  - Make it completely random: Every new user has a 50/50 chance of being placed in either group

- The first two are valid options for choosing samples and are fairly simple to implement, but they both have one fundamental flaw: they are both at risk of introducing a sampling bias.
- A sampling bias occurs when the way the sample is obtained systemically favors some outcome over the target outcome.

# Random sampling

- Another risk is the impact of a confounding factor into our analysis. A confounding factor is a variable that we are not directly measuring but connects the variables that are being measured.
- Basically, a confounding factor is like the missing element in our analysis that is invisible but affects our results.
- In this case, option 1 is not taking into account the potential confounding factor of geographical taste. For example, if website A is unappealing, in general, to the west coast users, it will affect your results drastically.
- Similarly, option 2 might introduce a temporal (time-based) confounding factor.

# Random sampling

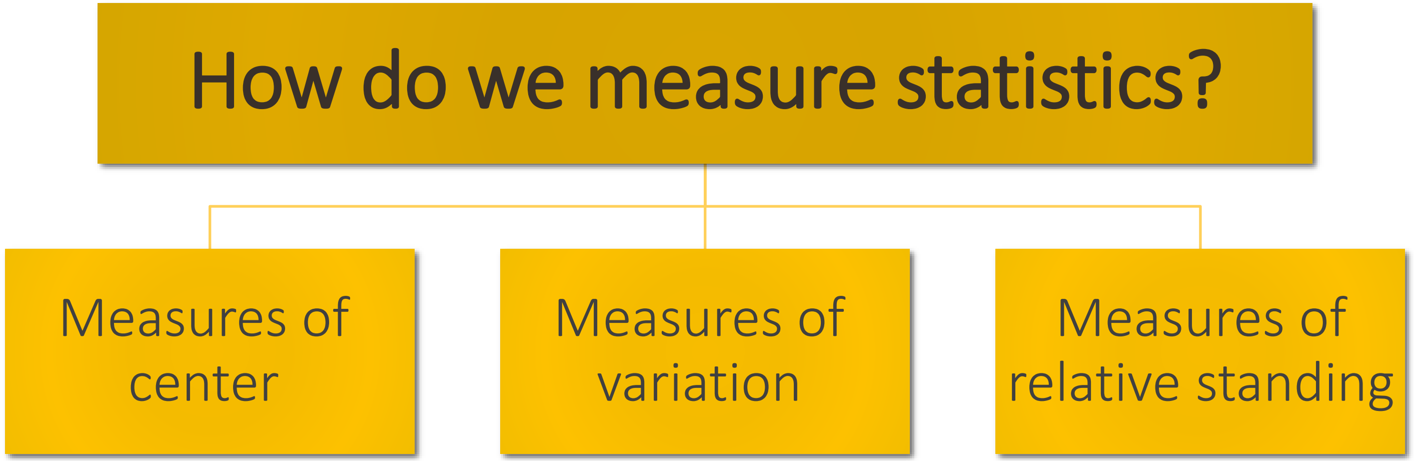
- The best selection!
- What if website B is better viewed in a nighttime environment (which was reserved for A), and users are turned off to the style purely because of what time it is. These are both factors that we want to avoid, so, we should go with option 3, which is a random sample.
- A random sample is chosen such that every single member of a population has an equal chance of being chosen as any other member.
- This is probably one of the easiest and most convenient ways to decide who will be a part of your sample.
- Everyone has the exact same chance of being in any particular group.
- Random sampling is an effective way of reducing the impact of confounding factors.

# Unequal probability sampling

- If we are interested in measuring the happiness level of our employees. We already know that we can't ask every single person on the staff because that would be silly and exhausting. So, we need to take a sample. Does anyone know the percentage of men/women who work here?
- This question is extremely important because sex is likely to be a confounding factor.
- The team found a split of 75% men and 25% women in the company. This means that if we introduce a random sample, our sample will likely favor the results for men and not women.

- To combat this, we can favor including more women than men in our survey in order to make the split of our sample less favored for men.
- We can use unequal sampling to remove systematic bias among gender, race, disability, and so on is much more pertinent.
- Therefore, it can be okay to introduce such a favoring system in your sampling techniques.

# HOW DO WE MEASURE STATISTICS?

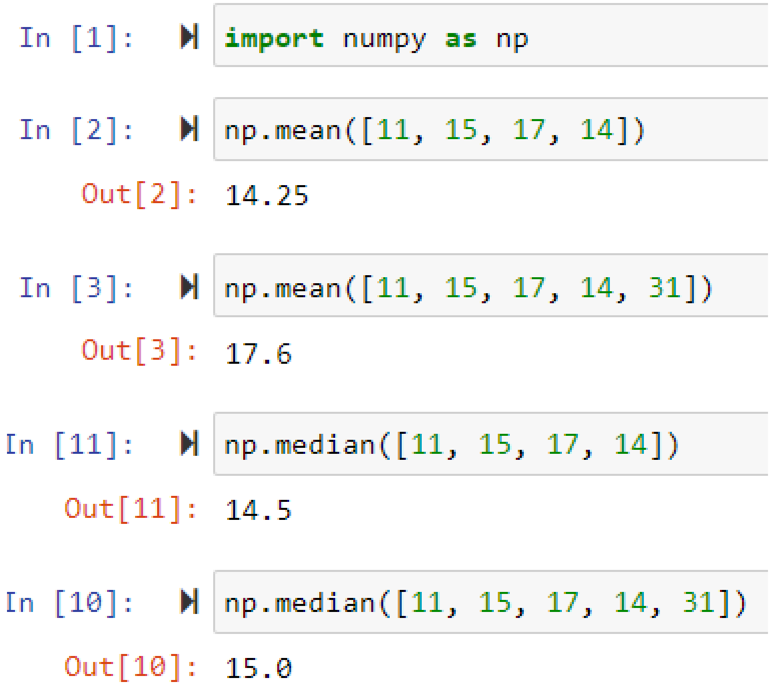


How do we measure statistics?

- Measures of center
- Measures of variation
- Measures of relative standing

# Measures of center

- A measure of center is a value in the "middle" of a dataset.
- It's a way to generalize a large set of data so that it's easier to convey to someone.
- For example, perhaps we're curious about what the average rainfall in Seattle is or what the median height for European males is.
- There are so many different ways of defining the center of data:
  - The arithmetic mean of a dataset is found by adding up all of the values and then dividing it by the number of data values.
  - The median is the number found in the middle of the dataset when it is sorted in order



```
In [1]:  import numpy as np

In [2]:  np.mean([11, 15, 17, 14])
Out[2]:  14.25

In [3]:  np.mean([11, 15, 17, 14, 31])
Out[3]:  17.6

In [11]: np.median([11, 15, 17, 14])
Out[11]: 14.5

In [10]: np.median([11, 15, 17, 14, 31])
Out[10]: 15.0
```

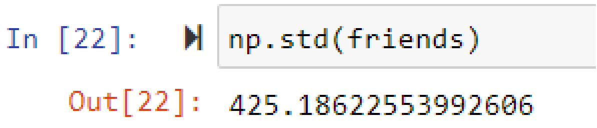
# Measures of variation

- Measures of center are used to quantify the middle of the data, but now we will explore ways of measuring how "spread out" the data we collect is. This is a useful way to identify if our data has many outliers lurking inside.
- The range tells us how far away the two most extreme values are.

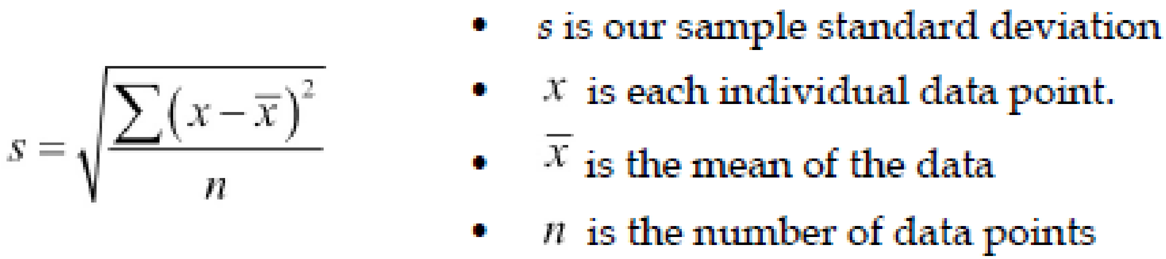
- This is most useful in scientific measurements or safety measurements.
  - Suppose a car company wants to measure how long it takes for an air bag to deploy. Knowing the average of that time is nice, but they also really want to know how spread apart the slowest time is versus the fastest time.
  - This literally could be the difference between life and death.

# Measures of variation

- The most commonly used measure of variation, the standard deviation.
- In essence, standard deviation, denoted by **s** when we are working with a sample of a population, measures how much data values deviate from the arithmetic mean.
- It's basically a way to see how spread out the data is.
- There is a general formula to calculate the standard deviation, which is as follows:



```
In [22]:  np.std(friends)
Out[22]:  425.18622553992606
```



$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

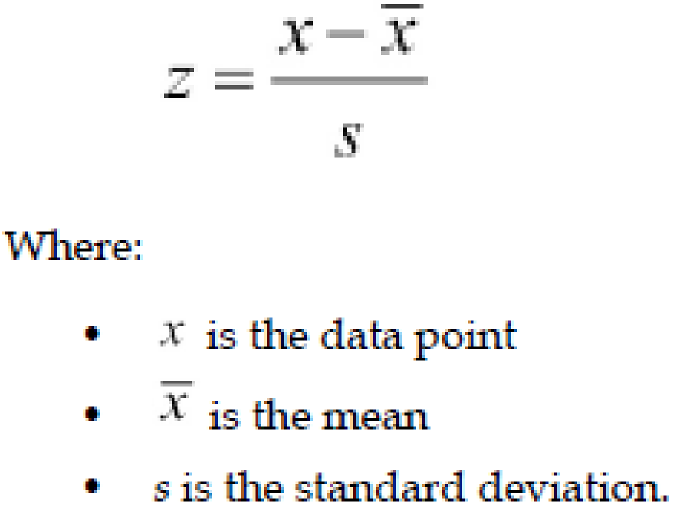
- $s$ is our sample standard deviation
- $x$ is each individual data point.
- $\bar{x}$ is the mean of the data
- $n$ is the number of data points

# Measures of relative standing

- We can combine both the measures of centers and variations to create measures of relative standings.
- Measures of variation measure where particular data values are positioned, relative to the entire dataset.
- Let's begin by learning a very important value in statistics, the z-score.
- The z-score is a way of telling us how far away a single data value is from the mean.

# The z-score

- The z-score of a x data value is as follows:



$$z = \frac{x - \overline{x}}{s}$$

Where:

- $x$ is the data point
- $\overline{x}$ is the mean
- $s$ is the standard deviation.

# The insightful part – correlations in data

- Having data is only one step to a successful data science operation. Being able to obtain, clean, and plot data helps to tell the story that the data has to offer but cannot reveal the moral.
- In subsequent chapters, we will look at a specific machine learning algorithm that attempts to find relationships between quantitative features, called linear regression, but we do not have to wait until then. We have a sample of people, a measure of their online social presence and their reported happiness. The question is—can we find a relationship between the number of friends on Facebook and overall happiness?
- Experiments to answer this question should be conducted in a laboratory setting, but we can begin to form a hypothesis about this question. We have the following three options for a hypothesis:
    - There is a positive association between the number of online friends and happiness (as one goes up, so does the other)
    - There is a negative association between them (as the number of friends goes up, your happiness goes down)
    - There is no association between the variables (as one changes, the other doesn't really change that much)

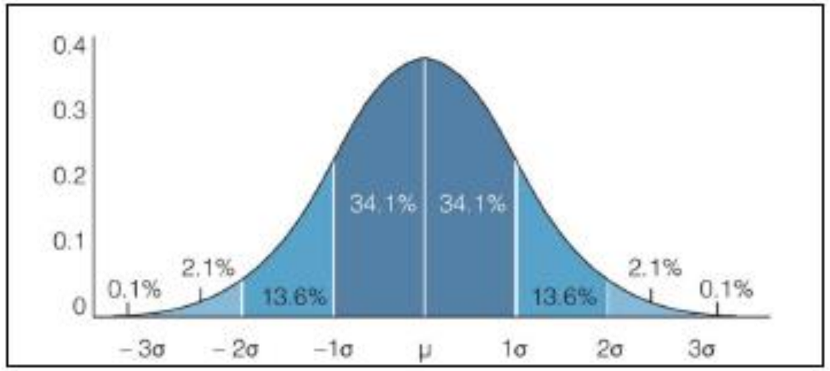
# The insightful part – correlations in data

- Can we use basic statistics to form a hypothesis about this question?
- Correlation coefficients are a quantitative measure that describe the strength of association/ relationship between two variables.
- The correlation between two sets of data tells us about how they move together.
- Would changing one help us predict the other? This concept is not only interesting in this case, but it is one of the core assumptions that many machine learning models make on data. For many prediction algorithms to work, they rely on the fact that there is some sort of relationship between the variables we are looking at. The learning algorithms then exploit this relationship to make accurate predictions.
- A few things to note about a correlation coefficient are as follows:
  - It will lie between -1 and 1
  - The greater the absolute value (closer to -1 or 1), the stronger the relationship between the variables:
    - The strongest correlation is a -1 or a 1
    - The weakest correlation is a 0
  - A positive correlation means that as one variable increases, the other one tends to increase as well
  - A negative correlation means that as one variable increases, the other one tends to decrease

# THE EMPIRICAL RULE

# The Empirical rule

- Recall that a normal distribution is defined as having a specific probability distribution that resembles a bell curve. In statistics, we love it when our data behaves normally. For example, if we have data that resembles a normal distribution, like so:



- The Empirical rule states that we can expect a certain amount of data to live between sets of standard deviations. Specifically, the Empirical rule states for data that is distributed normally:
  - about 68% of the data fall within 1 standard deviation
  - about 95% of the data fall within 2 standard deviations
  - about 99.7% of the data fall within 3 standard deviations

s.zahrani@tu.edu.sa