CIS 635 Knowledge Discovery & Data Mining

Basics Statistics

Basic Statistics - Data Science

- Let’s say you are working (as a DS) for a **Drug Discovering** company
- Your company has a plan (**long and short term**) to develop a drug (as they are the experts).
- Your **competitor** is another Giant, and they already have a very popular drug in the market (say for last **10 years**) and the **efficacy** rate (as many people has reported) to be **~90%**.

- **What are your thoughts on the problem?**
- **How would you approach?**

Basic Statistics - Data Science

- Let’s say you are working (as a DS) for a **Drug Discovering** company
- Your company has a plan (**long and short term**) to develop a drug (as they are the experts).
- Your **competitor** is another Giant, and they already have a very popular drug in the market (say for last **10 years**) and the **efficacy** rate (as many people has reported) to be **~90%**.

- **What are your thoughts on the problem?**
- **How would you approach?**

> In today's world, the buzzword is *data*, as in, “Do you have any data to support your claim?” “What data do you have on this?” “The data supported the original hypothesis that ...,” “Statistical data show that ...,” and “The data bear this out ....” But the field of statistics is not just about data.

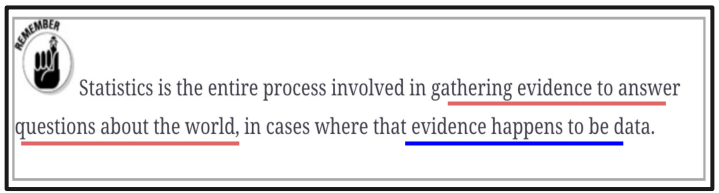
Basic Statistics - Data Science

- When we talk about data, the first thing comes to our mind is “Statistics”; right?
- What are the associations?

In today's world, the buzzword is *data*, as in, “Do you have any data to support your claim?” “What data do you have on this?” “The data supported the original hypothesis that ...,” “Statistical data show that ...,” and “The data bear this out ....” But the field of statistics is not just about data.

Basic Statistics - Data Science

Statistics is involved in every aspect of the scientific method



In today's world, the buzzword is *data*, as in, “Do you have any data to support your claim?” “What data do you have on this?” “The data supported the original hypothesis that ...,” “Statistical data show that ...,” and “The data bear this out ....” But the field of statistics is not just about data.

Basic Statistics - Data Science

Chapter 4 (Statistics for Dummies - Rumsey)

- **Tools of the Trade**
  - **Learning some basic (but important concepts)**

Basic Statistics - Data Science

Data: Most data, at lower level, fall into two groups:

- Numerical, or
- Categorical

Basic Statistics - Data Science

Data: Most data, at lower level, fall into two groups:

- **Numerical**, or
- Categorical

Numerical: Also known as "**Quantitative data**"

- **Measurements:** person's height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Basic Statistics - Data Science

Data: Most data, at lower level, fall into two groups:

- **Numerical**, or
- Categorical

Numerical data can be further broken into two types: discrete and continuous

- **Discrete:** items that can be counted; they can take on possible values that can be listed out.
- **Continuous**: Usually represents measurements; their possible values cannot be counted such as: a person's height, weight, IQ, or blood pressure

Numerical: Also known as "**Quantitative data**"

- **Measurements:** person's height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Basic Statistics - Data Science

Data: Most data, at lower level, fall into two groups:

- Numerical, or
- **Categorical**

Numerical data can be further broken into two types: discrete and continuous

- **Discrete:** items that can be counted; they can take on possible values that can be listed out.
- **Continuous**: Usually represents measurements; their possible values cannot be counted such as: a person's height, weight, IQ, or blood pressure

Numerical: Also known as "**Quantitative data**"

- **Measurements:** person's height, weight, or blood pressure; or
- **Counts**: such as number of stock shares a person wons, number of teeth a dog has, or the number of pages your favourite book contains

Categorical data: Categorical data represent characteristics such as a person's gender, marital status, country of birth, or the types of movies they like.

- Can be ordinal (say, student grades A, B, C; days of week, moths of week)
- Non ordinal data (person's gender, marital status, country of birth)

Basic Statistics - Data Science

Data set: The collection of **data** taken for a study is called a **sample**; for example

- If you measured the **weights of packages**, and those weights were 12, 22, 22, 68, and 3 pounds those five numbers **(12, 22, 22, 68, 3)** constitute your data sample

- If you just record their **corresponding sizes**, say, (**medium, medium, medium, large, small)** then this can be another representation of the same data (you must have your **size definition** somewhere)

Basic Statistics - Data Science

Data set: The collection of **data** taken for a study is called a **sample**; for example

- If you measured the **weights of packages**, and those weights were 12, 22, 22, 68, and 3 pounds those five numbers **(12, 22, 22, 68, 3)** constitute your data sample

- If you just record their **corresponding sizes**, say, (**medium, medium, medium, large, small)** then this can be another representation of the same data (you must have your **size definition** somewhere)

Variable: A variable is any characteristic or numerical value that varies from individual to individual.

- A variable can represent a count (for example, the number of siblings you have); or a measurement (say, your weight, **weights of packages**). Or
- It can be categorical where an individual is placed into a group based on certain criteria (say, your educational level), **size of packages**
- Actual pieces of information recorded on individuals regarding a variable are the **data**

Basic Statistics - Data Science

Population: Your **focused group of individuals/entities** related to your **query/research** (say, a group of people, cities, animals, and so on). Here are some sample questions:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Many times researchers want to study and make conclusions about a broad population, but in the end — to save time, money, or just because they don't know any better — they study only a narrowly defined population. That shortcut can lead to big trouble when conclusions are drawn. For example, suppose a college professor wants to study how TV ads persuade consumers to buy products. Her study is based on a group of her own students who participated to get five points extra credit. This test group may be convenient, but her results can't be generalized to any population beyond her own students, because no other population was represented in her study.

Basic Statistics - Data Science

Population: Your **focused group of individuals/entities** related to your **query/research** (say, a group of people, cities, animals, and so on). Here are some sample questions:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Many times researchers want to study and make conclusions about a broad population, but in the end — to save time, money, or just because they don't know any better — they study only a narrowly defined population. That shortcut can lead to big trouble when conclusions are drawn. For example, suppose a college professor wants to study how TV ads persuade consumers to buy products. Her study is based on a group of her own students who participated to get five points extra credit. This test group may be convenient, but her results can't be generalized to any population beyond her own students, because no other population was represented in her study.

Basic Statistics - Data Science

Population: Your **focused group of individuals/entities** related to your **query/research** (say, a group of people, cities, animals, and so on). Here are some sample questions:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Many times researchers want to study and make conclusions about a broad population, but in the end — to save time, money, or just because they don't know any better — they study only a narrowly defined population. That shortcut can lead to big trouble when conclusions are drawn. For example, suppose a college professor wants to study how TV ads persuade consumers to buy products. Her study is based on a group of her own students who participated to get five points extra credit. This test group may be convenient, but her results can't be generalized to any population beyond her own students, because no other population was represented in her study.

Basic Statistics - Data Science

Population: Your **focused group of individuals/entities** related to your **query/research** (say, a group of people, cities, animals, and so on). Here are some sample questions:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Many times researchers want to study and make conclusions about a broad population, but in the end — to save time, money, or just because they don't know any better — they study only a narrowly defined population. That shortcut can lead to big trouble when conclusions are drawn. For example, suppose a college professor wants to study how TV ads persuade consumers to buy products. Her study is based on a group of her own students who participated to get five points extra credit. This test group may be convenient, but her results can't be generalized to any population beyond her own students, because no other population was represented in her study.

Basic Statistics - Data Science

Statistic: A statistic is a number that summarizes the data collected as a sample. For example,

- Data can be summarized as a **percentage** (**60%** of the US households sampled won more than 2 cars).
- An **average** (the **average** household income in West Michigan is **..**)
- A **median** (the **median** number of family member in US is **..**), or
- A **percentile** (your height is is at the **90th** percentile of the Americans based on the data provided by Stats USA).

Parameter: If data is collected is from entire population, it is called **census**.

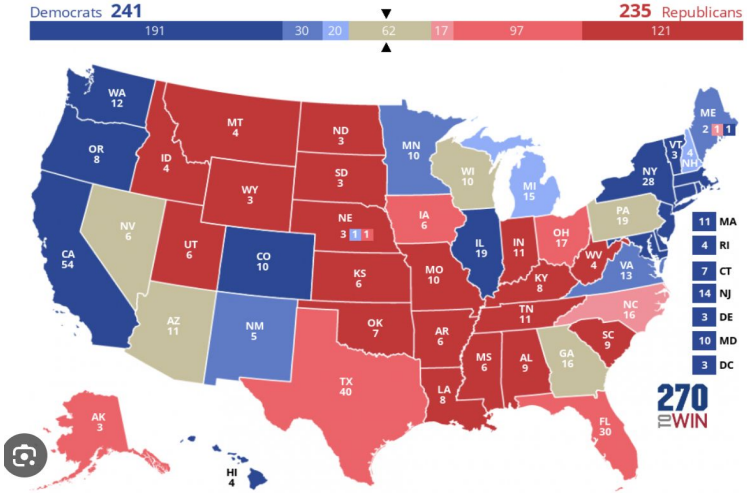
- If we summarize the census information from one variable to a single number, the number is called a **parameter**.

Basic Statistics - Data Science

Bias: Bias is systematic favoritism that may present in the data collection process, resulting in lopsided, misleading results. Bias can occur in many ways:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Mismatch to your research query!



Basic Statistics - Data Science

Bias: Bias is systematic favoritism that may present in the data collection process, resulting in lopsided, misleading results. Bias can occur in many ways:

- What do Americans think about the president's foreign policy?
- What percentage of planted crops in Wisconsin did deer destroy last year?
- What's the prognosis for breast cancer patients taking a new experimental drug?
- What percentage of all cereal boxes get filled according to specification?

Mismatch to your research query!

- **In the way the sample is selected:** For example, if you want to estimate how much holiday shopping people in the United States plan to do this year, and you take your clipboard and head out to a shopping mall on the day after Thanksgiving to ask customers about their shopping plans, you have bias in your sampling process. Your sample tends to favor those die-hard shoppers at that particular mall who were braving the massive crowds on that day known to retailers and shoppers as “Black Friday.”
- **In the way data are collected:** Poll questions are a major source of bias. Because researchers are often looking for a particular result, the questions they ask can often reflect and lead to that expected result. For example, the issue of a tax levy to help support local schools is something every voter faces at one time or another. A poll question asking, “Don't you think it would be a great investment in our future to support the local schools?” has a bit of bias. On the other hand, so does “Aren't you tired of paying money out of your pocket to educate other people's children?” Question wording can have a huge impact on results.

Basic Statistics - Data Science

Measures of Central Tendency

Central, or typical value of a Probability distribution; help you find the middle, or the average, of a dataset

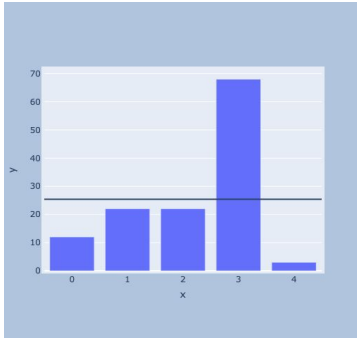
- Mean (Arithmetic)
- Median
- Mode

Basic Statistics - Data Science

Mean: Mean, also referred to as the average, is the most common **static** used to measure the **center**, or middle of a **numerical** data set.

- The mean is the sum of all numbers divided by the total number of observations.
- The mean of the entire population is called the **population mean**
- The mean of a sample is called the **sample mean**.

The mean may not be a fair representation of the data, because the average is easily is influenced by outliers

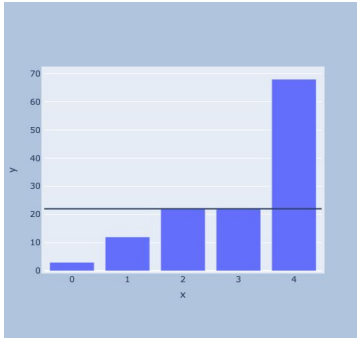


Weights of five packages: **(12, 22, 22, 68, 3)** constitute our data

Mean = $\frac{(12+22+22+68+3)}{5}$ = **25.4**

Basic Statistics - Data Science

Median: The median is another way to measure the center of a numerical data set. A threshold where 50% of data points are below and above the point.



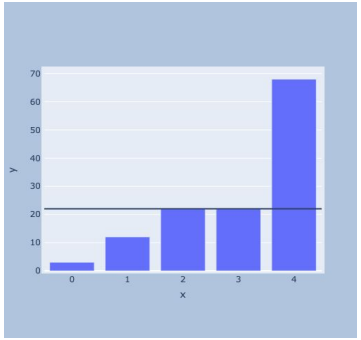
Weights of five packages: **(12, 22, 22, 68, 3)** constitute our data

Median = (3, 12**, 22**, 22, 68)

The mean may not be a fair representation of the data, because the average is easily is influenced by outliers

Basic Statistics - Data Science

Mode: The mode represents the highest concentration of points.



Weights of five packages: **(12, 22, 22, 68, 3)** constitute our data

Mode = (3, 12**, 22**, **22**, 68)

The mean may not be a fair representation of the data, because the average is easily is influenced by outliers

Basic Statistics - Data Science

Variance: The average of the squared differences from the Mean. This measures of how spread out numbers are.

Standard Deviation: sqrt (variance); Another static to measure how spread out numbers are.

$$\sqrt{\bar{}(variance)}$$

- On average how far away from the **average/mean**
- Comparte students groups
  - G1: with only A and B grade, **vs**
  - G2: with A, B, and C grade

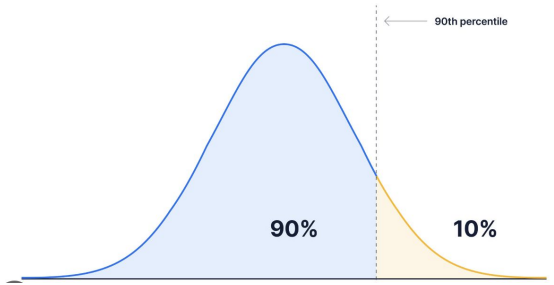
Basic Statistics - Data Science

	Population	Sample
# of subjects	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
	Note: S^2 is the formula for unbiased sample variance, since we're dividing by $n-1$.	
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
	Note: Finding S by taking $\sqrt{S^2}$ reintroduces bias.	

Basic Statistics - Data Science

Percentile: The **percentile** reported for a given variable (score) is the percentage of values in the data set that **falls below** that threshold. For example,

- If **your score** was reported to be at the **90th percentile**, that means 90% of the other people who took the test scored lower than you did.



QA