Bertelsmann Tech Scholarship Challenge Course - Data Track Nanodegree Program

Introduction to Data Analysis & Programming - 36 hours 17 minutes

LESSON 1

Welcome to the Scholarship Challenge Course!

LESSON 2

Intro to Research Methods

LESSON 3

PS 1a: Intro to Research Methods

LESSON 4

PS 1b: Additional Practice (Optional)

LESSON 5

Visualizing Data

LESSON 6

PS 2a: Visualizing Data

LESSON 7

PS 2b: Additional Practice (Optional)

LESSON 8

Google Spreadsheet Tutorial

LESSON 9

Central Tendency

To get started, we'll discuss some of the ways you can summarize the central tendencies of data, like mean, median and mode. This lesson is optional and if you are familiar with these concepts you can skip them.

LESSON 10

PS 3a: Central Tendency

LESSON 11

PS 3b: Additional Practice (Optional)

LESSON 12

Variability

LESSON 13

PS 4: Variability

LESSON 14

Standardizing

LESSON 15

PS 5a: Standardizing

LESSON 16

PS 5b: Additional Practice (Optional)

LESSON 17

Normal Distribution

LESSON 18

PS 6: Normal Distribution

LESSON 19

Sampling Distributions

Bertelsmann Tech Scholarship Challenge Course - Data Track Nanodegree Program

LESSON 20

PS 7: Sampling Distributions

LESSON 21

Why Python Programming

Welcome to Introduction to Python! Here's an overview of the course.

LESSON 22

Data Types and Operators

Familiarize yourself with the building blocks of Python! Learn about data types and operators, compound data structures, type conversion, built-in functions, and style guidelines.

LESSON 23

Control Flow

Build logic into your code with control flow tools! Learn about conditional statements, repeating code with loops and useful built-in functions, and list comprehensions.

LESSON 24

Functions

Learn how to use functions to improve and reuse your code! Learn about functions, variable scope, documentation, lambda expressions, iterators, and generators.

LESSON 25

Scripting

Setup your own programming environment to write and run Python scripts locally! Learn good scripting practices, interact with different inputs, and discover awesome tools.

LESSON 26

Basic SQL

In this section, you will gain knowledge about SQL basics for working with a single table. You will learn the key commands to filter a table in many different ways.

LESSON 27

SQL Joins

In this lesson, you will learn how to combine data from multiple tables together.

Bertelsmann Tech Scholarship Challenge Course - Data Track Nanodegree Program

LESSON 28

SQL Aggregations

In this lesson, you will learn how to aggregate data using SQL functions like SUM, AVG, and COUNT. Additionally, CASE, HAVING, and DATE functions provide you an incredible problem solving toolkit.

LESSON 29

SQL Subqueries & Temporary Tables

In this lesson, you will be learning to answer much more complex business questions using nested querying methods - also known as subqueries.

LESSON 30

SQL Data Cleaning

Cleaning data is an important part of the data analysis process. You will be learning how to perform data cleaning using SQL in this lesson.

LESSON 31

Challenge Course Wrap-Up

Bertelsmann Tech Scholarship Challenge Course - Data Track Nanodegree Program

LESSON 2

Intro to Research Methods

2.20 Quiz: Visualize Relationship ---------->
Visualize data to make it easier to draw conclusions

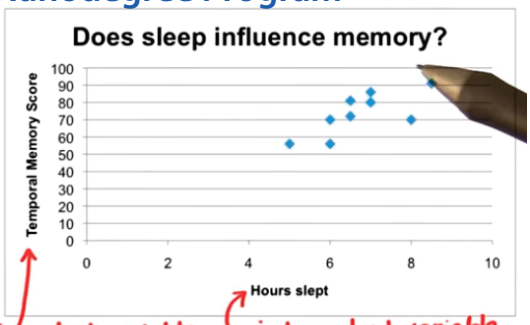
2.22 Golden Arches Theory
Correlation does not mean/ imply causation.
2 vars are related, this doesn't mean that 1 causes other to occur.

Show relationships ⇒ Observational studies Surveys

2.25 Causal Inference -->

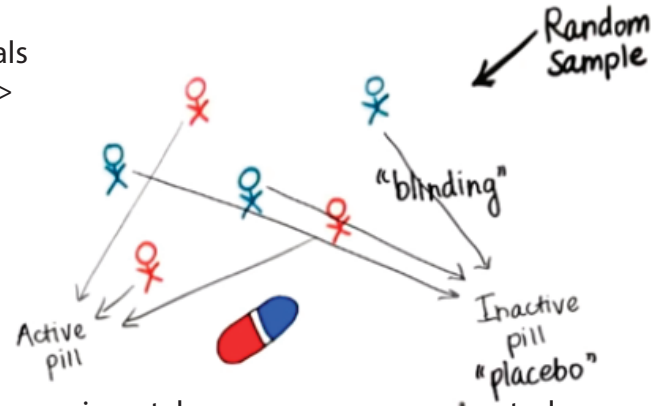
Show causation ⇒ Controlled experiment

Hours Slept	Temporal Memory
7	86
8	70
6	56
5	56
6	70
7	80
6.5	72
8.5	91
6.5	81
7	86



Raw data — Visualize data = Summarize data

2.27 Quiz: Downsides of Surveys
Surveys used to analyze constructs - Surveys control experimentals
2.30 Double Blind ---------->
Researchers + participants donot know.
2.35 Correlation does not mean causation.



experimentals group ---------- control group

LESSON 3 PS = Problem Set

PS 1a: Intro to Research Methods

3.2 Quiz 1: All California Residents
- A population is the entire group of everyone (or everything) you're interested in, while
- A Sample is a smaller group selected from the population. populations and samples are groups of people or things you want to study, while
- Statistics and populations are characteristics of those groups.
More info --> https://www.cliffsnotes.com/

Dr. Snycerski used an anonymous survey to investigate the alcohol use of all California residents. The entire group of California residents is an example of a(n):
- sample
- statistic
- ● population
- parameter

3.3 Quiz 2: Sleepy College Students
A research on American college students. A group of 125 students are an example of a sample .

- ○ sample
- ○ statistic
- ○ population
- ○ parameter

3.4 Quiz 3: Not Enough Sleep...?
A group of 125 students sleep average of 6.2 hour/day. This average is an example of a statistic .
- parameter is a characteristic of a population, while
- a statistic is a characteristic of a sample.

- ○ sample
- ● statistic
- ○ population
- ○ parameter

3.5 Quiz 6: Characteristic of a Population
- Statistic is a number that describes a sample
- A var is a characteristic that describes individual data points, student's gender, age.
- A constant = your experiment that does not change = everyone test at 9am.

A number that describes a population is called a:
- ● parameter
- ○ statistic
- ○ variable
- ○ constant

3.6 Quiz 7. Sample Approximates Population

population. The difference between the sample and population averages is known as:
- ● sampling error

3.8 Quiz 10. Which Are Constructs?
- A construct is a var that is not directly observable or measurable. Would there be any difficulty in measuring someone's annual income in USD?

What is an operational definition? Choose all that apply.
- ☐ An abstract concept that we are interested in studying
- ☑ A way of turning constructs into variables we can measure
- ☐ The difference between a sample statistic and population parameter
- ☑ A way of describing a variable in terms of the way we measure it
- ☐ A group of individuals of interest in a research study
- ☐ Estimates of population parameters

3.9 Quiz 12. Define Operational Definition

Bertelsmann Tech Scholarship - Data Track

3.10. Quiz 13: Research Studies

- Too expensive to collect data entire population, so to learn about population using a sample. By estimate population parameters using sample statistics. We can't expect our estimates to be exactly accurate when we do this.

3.12 Quiz: 16. Which Are Hypotheses?

- scientific fact >< hypothesis

3.14 Quiz 18. Symbols = Fun Fun Fun!

3.17 Quiz 23. Sample vs. Population

- sampling error = difference betw population parameter and sample statistic.

$\bar{x}$ = sample statistic = average in the sample

μ = population par. = average value for population.

3.19 Ouiz 33. All Kinds of Variables

In an experiment, researcher manipulates the __ var measure changes in the __ variable, and seeks to control __ variables.

- lurking; dependent; independent
- continuous; discrete; dependent
- dependent; independent; lurking
- independent; dependent; lurking

3.21 Quiz 37. Insomnia

- Observational study, you look at existing data, so which participants are in the experiment group vs. the control group is outside of your control.

In most research studies (choose all that apply):

- [ ] Every individual in the population is included in the study
- [ ] Data from all individuals in the population are used to learn about a sample
- [x] Data from individuals in a sample are used to learn about a population
- [ ] We expect our best guesses (estimates) of the population parameters to be exactly equal to the actual values
- [x] We expect our best guesses (estimates) of the population parameters to differ from the actual values
- [x] We expect our sample statistics will not be exactly equal to the population parameters they are estimating

$\bar{x}$ is the statistical symbol for the __, whereas μ is the statistical symbol for the__:

- sample average; population average

Using a random sample (n = 100), researchers found that the average US resident spends 32 hours/week online. Imagine the true value for the entire population of US residents is 25 hours/week spent online. Given this scenario, which of the following is true? Choose all that apply.

- [x] The sample statistic is 32 and the population parameter is 25.
- [x] The difference between 32 and 25 is called "sampling error".
- [ ] The sample value is wrong because the sample was random.
- [x] We should not be surprised that the sample average is different from the population average.
- [ ] The sample statistic is 25 and the population parameter is 32.
- [x] The amount of sampling error in this example is 7 hours/week (32 – 25).
- [x] $\bar{x} = 32$; $\mu = 25$
- [ ] $\mu = 32$; $\bar{x} = 25$
- [x] The sample had 100 US residents in it.
- [x] We might have found a sample average closer to the population average if we used a larger sample (for example, n = 1000).
- [ ] We should be surprised that our sample average is different than our population average because random samples guarantee 100% accurate estimates.

- [ ] This is an example of an observational study, so we cannot make causal conclusions about the effectiveness of zolpidem.
- [x] This study is an experimental study, so we can make causal conclusions about the effectiveness of zolpidem.

LESSON 4

PS 1b: Additional Practice (Optional)

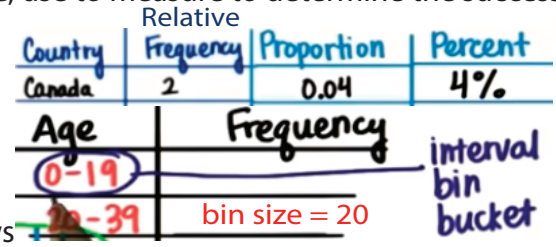
4.28 Quiz 49. Depression

- independent variable = var, that is different between the control and experiment group
- dependent variable, use to measure to determine the success of the experiment.

LESSON 5

Visualizing Data

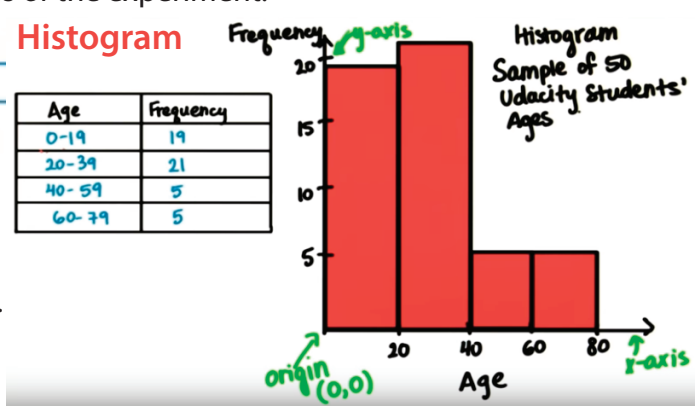
5.8 Convert to Percentage



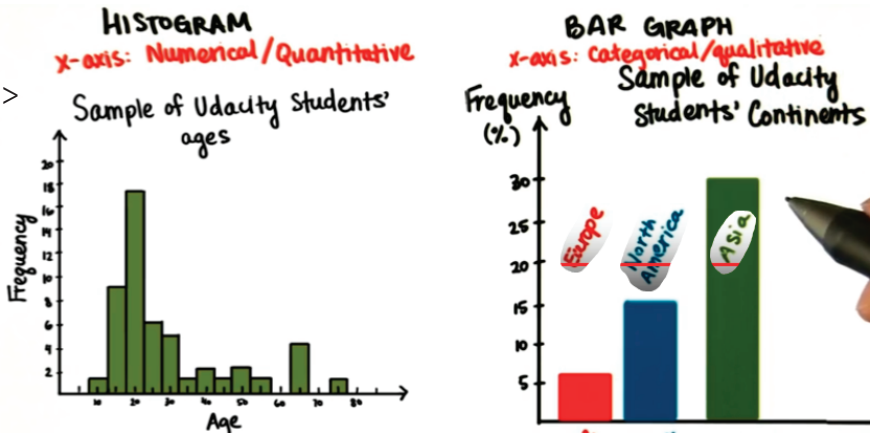
5.11 Number of Rows

5.14 Histogram

- Frequency is always on Y- axis = Dependent var, called the outcome
- On X- axis = Independent var = predictor var



5.23 Difference Between Graphs



Bertelsmann Tech Scholarship -

LESSON 6

PS 2a: Visualizing Data

6.1 Quiz 1. Blood Types
- Σ symbol = total sum = capital sigma.
- *f* stands for frequency (count)
- *p* stands for proportion.

6.15 Quiz 29. Frequency Axis
In general for histograms, x-axis has var you're interested in, broken down into bins. x-axis should be numerical.

6.16 Quiz 30. X-Axis Represent!
- Histograms should have a numerical x-axis.

If x-axis is categorical, the graph is called a bar graph. Space between each bar to indicate x-axis is not numerical.
- Frequency is on y-axis; countries and colors are categorical data. y-axis has the frequency of values in that bin.

LESSON 7

PS 2b: Additional Practice (Optional)

LESSON 8

Google Spreadsheet Tutorial

LESSON 9

Central Tendency

To get started, we'll discuss some of the ways you can summarize central tendencies of data, like mean, median and mode. This lesson is optional and if you are familiar with these concepts you can skip them.

9.3 Quiz Which Number to Choose?
- The value at highest frequency = mode = highest bin = most common value in dataset. Mode of a histogram should be the bin w/ the highest frequency. No mode = uniform distribution, but some distributions has multiple modes. mode occurs on the X-axis.
- The value in the middle distribution = median
- Average = statistic, rest at a specific spot in the middle distribution.

The mode, median, average > all help describe the distribution.
Each has strengths & weaknesses.

9.7 Quiz Mode - Uniform Distribution

9.8 More than One Mode?
- Distribution has multi modals when it has more clear trends.

Data tells a story about male, female foot sizes.

9.9 Quiz Mode of Categorical Data
mode occurs on X-axis, so look for what value has highest frequency.

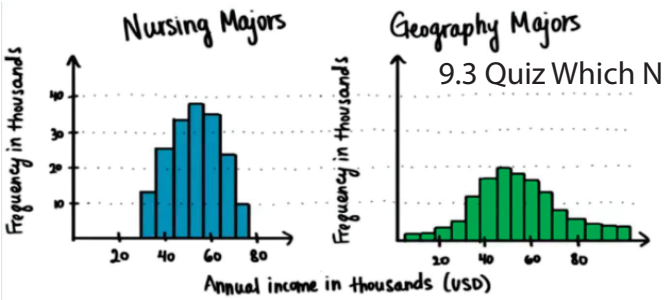
9.10 Quiz More o' Mode!
- Cann't describe mode w/ equation, so we use the mean or average.

9.11 Find the Mean
- mean = average of all values

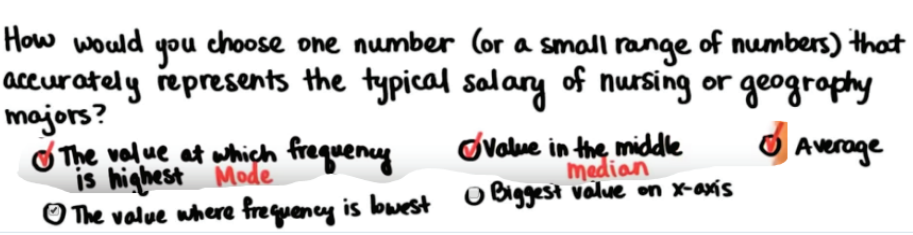
9.12 Procedure for Finding Mean
- Procedure = steps to solve a problem.

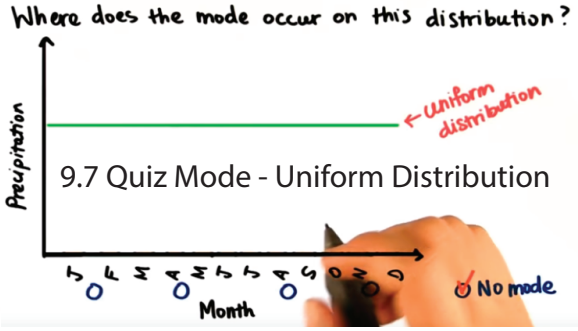
9.14 Helpful Symbols
- Math = universal lg = symbols. Mathematics is a way of thinking, we symbolize these thoughts.

This statistical notation is shorthand instructions that tell us what to do.

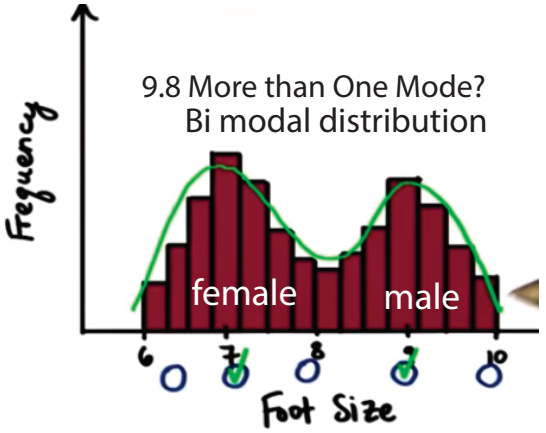


9.3 Quiz Which Number to Choose?





9.7 Quiz Mode - Uniform Distribution



9.8 More than One Mode?
Bi modal distribution