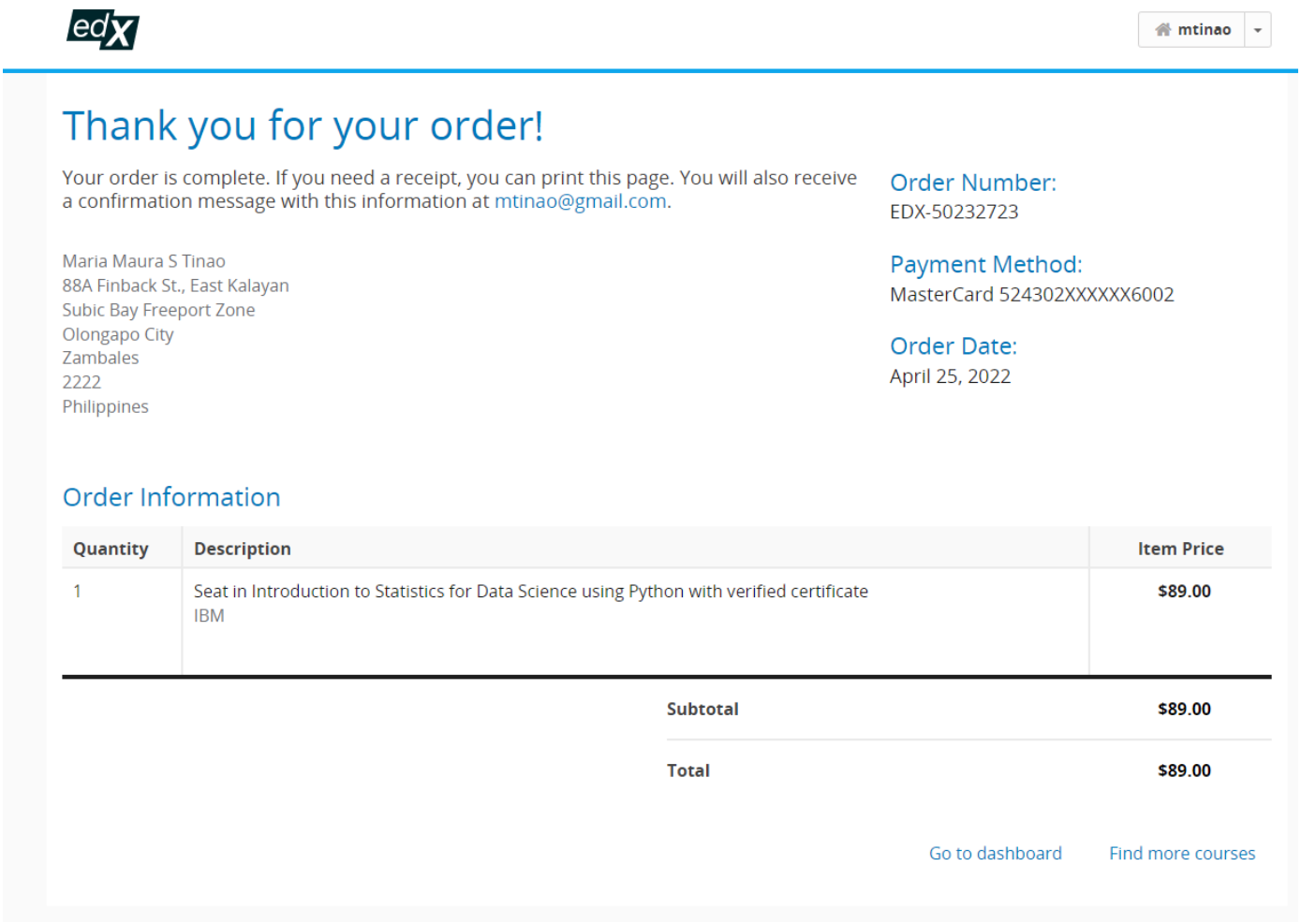


https://github.com/suneelpatel/Statistics-for-Data-Science-using-Python/tree/master/Descriptive_Statstics_using_Python

Introduction to Statistics for Data Science using Python

Begin your course today Start course

Expand all

Welcome to the Course

- Welcome from your instructors! (2:56)
- Course Overview
- Python Packages for Data Science (2:37)
- (Optional) Basics of Jupyter Notebooks
- (Optional) Lab: Python Review

Descriptive Statistics

- Welcome to Statistics (4:26)
- Types of Data (5:33)
- Measures of Central Tendency (5:04)
- Measure of Dispersion (4:05)
- Lab: Descriptive Statistics
- Quiz: Introduction to Descriptive Statistics (1 Question)
  Graded Quiz due May 2, 2022, 8:46 AM GMT+8

Visualizing Data with graphs and charts

- Visualization Fundamentals (3:07)
- Statistics by Groups (6:47)
- Statistical Charts (4:05)
- Introducing the Teacher's Rating Data (4:32)
- Lab: Visualizing Data
- Quiz: Data Visualization (1 Question)
  Graded Quiz due May 5, 2022, 8:46 PM GMT+8

Hypothesis testing and Probability Distribution

- Random Numbers and Probability Distributions (4:46)
- State Your Hypothesis (3:35)
- Normal Distribution (3:59)
- T-Distribution (4:50)
- Probability of Getting a High or Low Teaching Evaluation (4:18)
- Lab: Introduction to Probability Distributions
- Quiz: Introduction to Probability Distributions (1 Question)
  Graded Quiz due May 9, 2022, 8:46 AM GMT+8

Testing for mean differences and relationships

- Z-Test or T-Test (4:04)
- Dealing with Rejections and Tails (4:33)
- Equal vs unequal variances (2:42)
- ANOVA (4:36)
- Correlation tests (7:01)
- Lab: Hypothesis Testing
- Quiz: Hypothesis Testing (1 Question)
  Graded Quiz due May 12, 2022, 8:46 PM GMT+8

Regression analysis as an alternative to Hypothesis Testing

- Regression: The Workhorse of Statistical Analysis (4:19)
- Regression in place of T-Test (2:14)
- Regression in place of ANOVA (3:07)
- Regression in place of Correlation (2:00)
- Python Packages for Data Science (2:41)
- Lab: Regression Analysis
- Quiz: Regression Analysis (1 Question)
  Graded Quiz due May 16, 2022, 8:46 AM GMT+8

Final Exam

- Final Exam (1 Question)
  Final Exam due May 19, 2022, 8:46 PM GMT+8

Badge

Welcome from your instructors!

Hello and welcome to this course on statistics. In this course, our goal is to make learning statistics fun and enable you to apply statistical methods for data analysis and Data Science.

My name is Murtaza Haider, I'm your instructor for this course. I'm also an associate professor at the Ted Rogers School of Management at Ryerson University in Toronto. The author of Getting Started with Data Science, Making Sense of Data with Analytics.

My research interests are in urban economics as they relate to housing markets and transportation. I blog regularly, and you can find my blogs on Huffington Post.

By way of training, I have a Master's in Transportation Engineering and Planning and a PhD in Civil Engineering with a focus on Urban Systems Analysis.

My name is Aije Egwaikhide. I'm the co-instructor for this course. I'm a Senior Data Scientist and Statisticians with the IBM developers skills networks team. I have field experience working on supervised and unsupervised Machine Learning algorithms for oil and gas clients.

During my high school in Nigeria, it was easy to put off mathematics and statistics and focus on the seemingly easier courses. I always love a good challenge. My interest in statistics and mathematics spiked as a result of people around me saying it was hard. So I made my parents invest in textbooks and I always made sure to be ahead of the class. When I got to the University of Manitoba for my undergrad, picking statistics alongside economics was easy. When I had to do a postgrad, picking Data Science and Business Analytics was a no-brainer.

On my off days, I'm a fashion and carrier blogger on Instagram. This course consists of five modules: Introduction and Descriptive Statistics, Data Visualization, Introduction to Probability Distribution, Hypothesis Testing, and Regression Analysis. Each module will comprise of four to six videos, and we'll include exercises for you to practice on.

The hands-on lab will utilize Jupiter Notebooks using the Python programming language, which is one of the easiest programming languages to learn.

Let's get started and happy learning.

Course Outline

- This course contains five Modules
  1. Introduction and Descriptive Statistics
  2. Data Visualization
  3. Introduction to Probability Distribution
  4. Hypothesis Testing
  5. Regression Analysis
- Each Module will include videos and practical exercises
- Hands-on labs with Jupyter notebooks using Python

Course Overview

This Statistics for Data Science course is designed to introduce you to the basic principles of statistical methods and procedures used for data analysis. After completing this course you will have practical knowledge of crucial topics in statistics including – data gathering, summarizing data using descriptive statistics, displaying and visualizing data, examining relationships between variables, probability distributions, expected values, hypothesis testing, introduction to ANOVA (analysis of variance), regression and correlation analysis.

You will take a hands-on approach to statistical analysis using Python programming language and Jupyter Notebooks – the tools of choice for modern Data Scientists and Data Analysts. At the end of the course, you will complete a project to apply various concepts in the course to a Data Science problem involving a real-life inspired scenario and demonstrate an understanding of foundational statistical thinking and reasoning. The focus is on developing a clear understanding of the different approaches for different data types, developing an intuitive understanding, making appropriate assessments of the proposed methods, using Python to analyze our data, and interpreting the output accurately.

This course is suitable for a variety of professionals and students intending to start their journey in data and statistics-driven roles such as Data Scientists, Data Analysts, Business Analysts, Statisticians, and Researchers. We strongly recommend taking the Python for Data Science course before starting this course to get familiar with the Python programming language, Jupyter notebooks, and libraries. Optional refreshers for using Jupyter notebook and popular Python libraries are also provided. They are intended to help review some of the usages of Jupyter and Python in this course and are not a replacement for becoming proficient in Python.

Important

Every module of the course has:

- a **hands-on assignment** in Python,
- a **practice quiz** to test your understanding of the material, and
- a **graded quiz** that will contribute to 17% of the total grade for the course.

In the final week, you will be given a **project** that you will need to complete and submit for peer review. This project will be worth 15% of the final grade for the course.

Python Packages for Data Science

Welcome. In order to do data analysis in Python, we should first tell you a little bit about the main packages relevant to analysis in Python.

A Python Library

- is a collection of functions and methods that allows you to perform lots of actions without writing your code.
- The libraries usually contain built-in modules providing different functionalities which you can use directly, and their extensive libraries offering a broad range of facilities.

We divided the Python data analysis libraries into three groups. The first group is called scientific computing libraries.

Pandas,

- offers data structure and tools, for affective data manipulation and analysis.
- It provides fast access to structured data.
- The primary instrument of Pandas is a two dimensional table, consisting of columns and rows labels, which is called a data frame. It is designed to provide an easy indexing function.

NUM PY Library

- uses arrays as their inputs and outputs.
- It can be extended to objects for matrices, and with a little change of coding, developers perform faster a processing.

SciPy

- includes functions for some advanced math problems as listed in the slide, as well as data visualization.
- Using data visualization methods, are the best way to communicate with others and show the meaningful results of analysis. These libraries enable you to create graphs, charts, and maps.

The Matplotlib package

- is the most well known library for data visualization. T
- his package is great for making graphs and plots. The graphs are also highly customizable.

Another high level visualization library is

Seaborn.

- It is based on Matplotlib. It's very easy to generate some sort of plots like heat maps, time series, and violin plots.

With machine learning algorithms, were able to develop a model using our data set, and obtain predictions. The algorithmic libraries tackle some machine learning tasks from basic to complex.

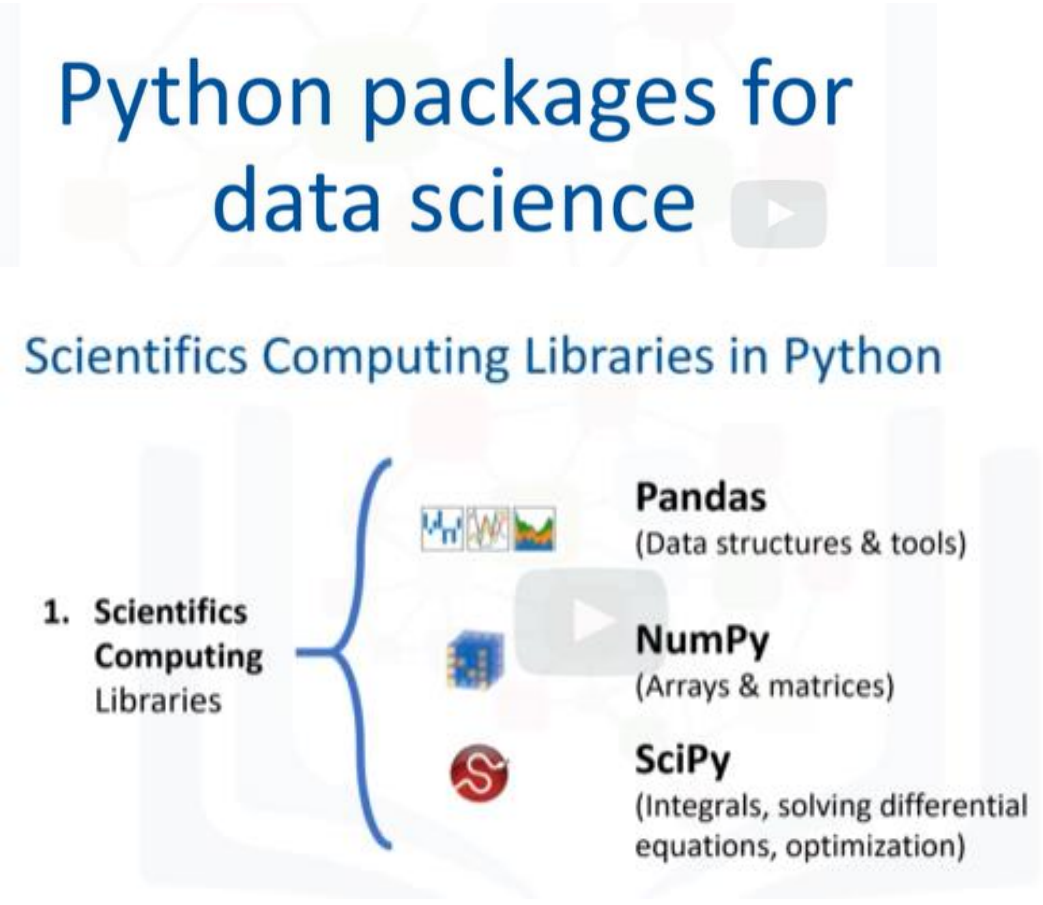
We introduce two packages.

The Scikit-learn Library

- contains tools for statistical modeling including regression, classification, clustering, and so on. I
- It is built on NUM PY, SciPY, and Matplotlib.

Statsmodels

- is also a Python module that allows users to explore data, estimate statistical models, and perform statistical tests.



Pandas,

- offers data structure and tools, for affective data manipulation and analysis.
- It provides fast access to structured data.

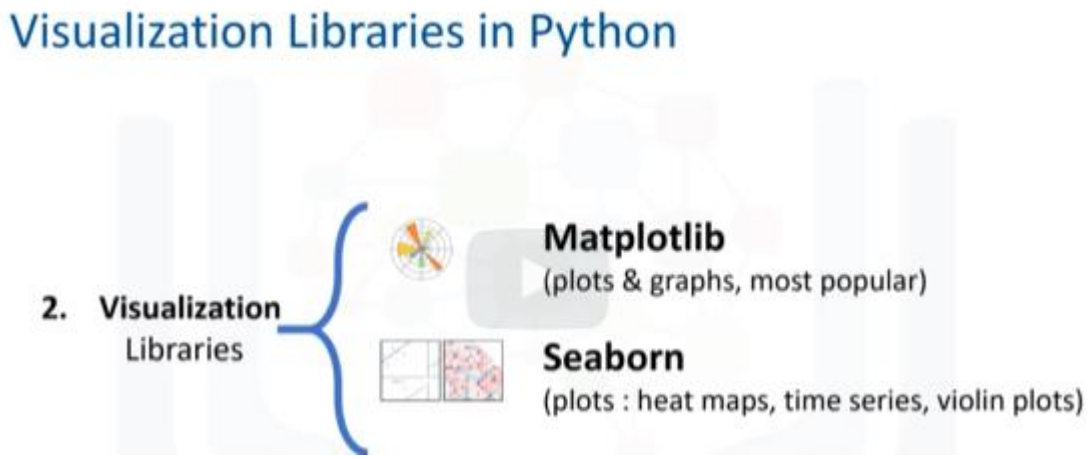
- The primary instrument of Pandas is a two dimensional table, consisting of columns and rows labels, which is called a data frame. It is designed to provide an easy indexing function.

NUM PY Library

- uses arrays as their inputs and outputs.
- It can be extended to objects for matrices, and with a little change of coding, developers perform faster a processing.

SciPy

- includes functions for some advanced math problems as listed in the slide, as well as data visualization.
- Using data visualization methods, are the best way to communicate with others and show the meaningful results of analysis. These libraries enable you to create graphs, charts, and maps.



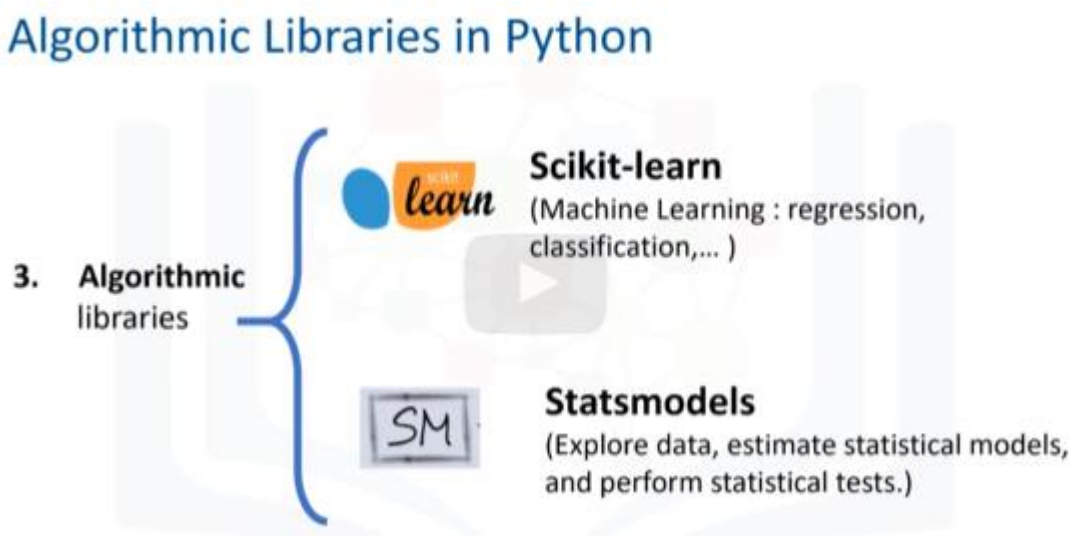
The Matplotlib package

- is the most well known library for data visualization. T
- his package is great for making graphs and plots. The graphs are also highly customizable.

Another high level visualization library is

Seaborn.

- It is based on Matplotlib. It's very easy to generate some sort of plots like heat maps, time series, and violin plots.



We introduce two packages.

The Scikit-learn Library

- contains tools for statistical modeling including regression, classification, clustering, and so on. I
- It is built on NUM PY, SciPY, and Matplotlib.

Statsmodels

- is also a Python module that allows users to explore data, estimate statistical models, and perform statistical tests.

Basics of Jupyter Notebooks

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and explanatory text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, machine learning and, much more."

– description from Project Jupyter

Jupyter Notebook has become the "tool of choice" for the modern Data Scientist. Jupyter took the concept of a paper notebook that scientists use to record experiments and made it digital and interactive. It is also a fantastic learning tool that allows researchers, teachers, and learners to combine computations with descriptive text, multimedia, and graphs. It helps to document experiments by combining computation (code and data) and a narrative explanation, share it with others, and to have it come alive clearly explaining your findings and the path you took to arrive at your conclusions. In short, it helps you tell a story about your experiment. You can also say that Jupyter Notebook and its more modern version JupyterLab is an Integrated Development Environment (IDE). But we prefer to think of it as a storytelling tool that makes data come alive.

To make it easy for you to get started, hands-on labs and assignments in this course will use a virtual JupyterLab Notebook environment provided by the Skills Network Labs. You will find it very convenient as you will not have to download, install, or configure any software. You will simply click on the "**Open Tool**" button. It's like magic!

Note: you can get your own Jupyter Notebooks is by installing Anaconda or you can use Jupyter Notebooks in IBM Watson Studio (we will go in-depth about this later in the course). We don't recommend doing it at this time at the risk of getting distracted. Like a real Data Scientist, you will use Jupyter to:

- Write and run interactive code (we use Python in this course),
- Document and explain what you were doing using markdown text,
- Share your notebooks with other learners, colleagues, and even the world.

Here are just a few very useful tips for using a Jupyter Notebook. Jupyter Notebooks are composed of *cells*. Each cell can be either:

- a *code cell* containing computer code (Python in our course),
- or a *markdown cell* containing text, pictures, videos, etc. formatted with markdown (a very simple formatting language).

To get an output from a code cell, you have to run the cell. To run the cell, place your cursor in the cell and click on the **Cell** menu on the top and click **Run Cells**. To use a shortcut, press **Shift + Enter**.

2. To add a new cell: Click on **Insert** and insert cell above or below depending on your preference.

3. To delete a cell. Click on **Edit** and choose **Delete Cells**

4. For a full list of the shortcuts, click on the **Help** button.

The next lesson will be an optional Python notebook to get you familiar with the pandas and numpy libraries.

Introduction Notebook

Estaimted time needed: **10** minutes

Objectives

After completing this lab you will be able to:

- Acquire data in various ways
- Obtain insights from Data with Pandas library

Table of Contents

Estimated Time Needed: **10 min**

Data Acquisition

There are various formats for a dataset, .csv, .json, .xlsx etc. The dataset can be stored in different places, on your local machine or sometimes online.
In this section, you will learn how to load a dataset into our Jupyter Notebook.
In our case, the Automobile Dataset is an online source, and it is in CSV (comma separated value) format. Let's use this dataset as an example to practice data reading.

- data source: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data
- data type: csv

The Pandas Library is a useful tool that enables us to read various datasets into a data frame; our Jupyter notebook platforms have a built-in **Pandas Library** so that all we need to do is import Pandas without installing. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install Pandas or Numpy.

```
[1]: #install specific version of libraries used in  lab
     #! mamba install pandas==1.3.3  -y
     #! mamba install numpy=1.21.2 -y
```

```
[2]: # import library
     import pandas as pd
     import numpy as np
```

Read Data

We use pandas.read_csv() function to read the csv file. In the bracket, we put the file path along with a quotation mark, so that pandas will read the file into a data frame from that address. The file path can be either an URL or your local file address.
Because the data does not include headers, we can add an argument headers = None inside the read_csv() method, so that pandas will not automatically set the first row as a header.
You can also assign the dataset to any variable you create.
This dataset was hosted on IBM Cloud object click HERE for free storage.

```
[3]: # Read the online file by the URL provides above, and assign it to variable "df"
     other_path = "https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/auto.csv"
     df = pd.read_csv(other_path, header=None)
```

After reading the dataset, we can use the dataframe.head(n) method to check the top n rows of the dataframe; where n is an integer. Contrary to dataframe.head(n), dataframe.tail(n) will show you the bottom n rows of the dataframe.

```
[4]: # show the first 5 rows using dataframe.head() method
     print("The first 5 rows of the dataframe")
     df.head(5)
```

```
The first 5 rows of the dataframe
```

[4]:

	0	1	2	3	4	5	6	7	8	9	...	16	17	18	19	20	21	22	23	24	25
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	13495
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	16500
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154	5000	19	26	16500
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	102	5500	24	30	13950
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	115	5500	18	22	17450

5 rows × 26 columns

Question #1:

check the bottom 10 rows of data frame "df".

[5]:
```
# Write your code below and press Shift+Enter to execute
print("The last 10 rows of the dataframe\n")
df.tail(10)
```

```
The last 10 rows of the dataframe
```

[5]:

	0	1	2	3	4	5	6	7	8	9	...	16	17	18	19	20	21	22	23	24	25
195	-1	74	volvo	gas	std	four	wagon	rwd	front	104.3	...	141	mpfi	3.78	3.15	9.5	114	5400	23	28	13415
196	-2	103	volvo	gas	std	four	sedan	rwd	front	104.3	...	141	mpfi	3.78	3.15	9.5	114	5400	24	28	15985
197	-1	74	volvo	gas	std	four	wagon	rwd	front	104.3	...	141	mpfi	3.78	3.15	9.5	114	5400	24	28	16515
198	-2	103	volvo	gas	turbo	four	sedan	rwd	front	104.3	...	130	mpfi	3.62	3.15	7.5	162	5100	17	22	18420
199	-1	74	volvo	gas	turbo	four	wagon	rwd	front	104.3	...	130	mpfi	3.62	3.15	7.5	162	5100	17	22	18950
200	-1	95	volvo	gas	std	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	114	5400	23	28	16845
201	-1	95	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	8.7	160	5300	19	25	19045
202	-1	95	volvo	gas	std	four	sedan	rwd	front	109.1	...	173	mpfi	3.58	2.87	8.8	134	5500	18	23	21485
203	-1	95	volvo	diesel	turbo	four	sedan	rwd	front	109.1	...	145	idi	3.01	3.40	23.0	106	4800	26	27	22470
204	-1	95	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	114	5400	19	25	22625

10 rows × 26 columns

Question #1 Answer:

Run the code below for the solution!

```
Double-click <b>here</b> for the solution.

<!-- The answer is below:

print("The last 10 rows of the dataframe\n")
df.tail(10)

-->
```

Add Headers

Take a look at our dataset; pandas automatically set the header by an integer from 0.

To better describe our data we can introduce a header, this information is available at: https://archive.ics.uci.edu/ml/datasets/Automobile
Thus, we have to add headers manually.

Firstly, we create a list "headers" that include all column names in order. Then, we use dataframe.columns = headers to replace the headers by the list we created.

[6]:
```
# create headers list
headers = ["symboling","normalized-losses","make","fuel-type","aspiration", "num-of-doors","body-style",
           "drive-wheels","engine-location","wheel-base", "length","width","height","curb-weight","engine-type",
           "num-of-cylinders", "engine-size","fuel-system","bore","stroke","compression-ratio","horsepower",
           "peak-rpm","city-mpg","highway-mpg","price"]
print("headers\n", headers)
```

headers
['symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiration', 'num-of-doors', 'body-style', 'drive-wheels', 'engine-location', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-type', 'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', 'highway-mpg', 'price']

We replace headers and recheck our data frame

[7]:
```
df.columns = headers
df.head(10)
```

[7]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131

fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
mpfi	3.47	2.68	9.0	111	5000	21	27	13495
mpfi	3.47	2.68	9.0	111	5000	21	27	16500
mpfi	2.68	3.47	9.0	154	5000	19	26	16500
mpfi	3.19	3.40	10.0	102	5500	24	30	13950
mpfi	3.19	3.40	8.0	115	5500	18	22	17450
mpfi	3.19	3.40	8.5	110	5500	19	25	15250
mpfi	3.19	3.40	8.5	110	5500	19	25	17710
mpfi	3.19	3.40	8.5	110	5500	19	25	18920
mpfi	3.13	3.40	8.3	140	5500	17	20	23875
mpfi	3.13	3.40	7.0	160	5500	16	22	?

10 rows × 26 columns

we need to replace the "?" symbol with NaN so the dropna() can remove the missing values

[8]:
```
df1=df.replace('?',np.NaN)
```

we can drop missing values along the column "price" as follows

[9]:
```
df=df1.dropna(subset=["price"], axis=0)
df.head(20)
```

[9]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...
5	2	NaN	audi	gas	std	two	sedan	fwd	front	99.8	...
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...
7	1	NaN	audi	gas	std	four	wagon	fwd	front	105.8	...
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...
10	2	192	bmw	gas	std	two	sedan	rwd	front	101.2	...
11	0	192	bmw	gas	std	four	sedan	rwd	front	101.2	...
12	0	188	bmw	gas	std	two	sedan	rwd	front	101.2	...
13	0	188	bmw	gas	std	four	sedan	rwd	front	101.2	...
14	1	NaN	bmw	gas	std	four	sedan	rwd	front	103.5	...
15	0	NaN	bmw	gas	std	four	sedan	rwd	front	103.5	...
16	0	NaN	bmw	gas	std	two	sedan	rwd	front	103.5	...
17	0	NaN	bmw	gas	std	four	sedan	rwd	front	110.0	...
18	2	121	chevrolet	gas	std	two	hatchback	fwd	front	88.4	...
19	1	98	chevrolet	gas	std	two	hatchback	fwd	front	94.5	...
20	0	81	chevrolet	gas	std	four	sedan	fwd	front	94.5	...

engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
130	mpfi	3.47	2.68	9.0	111	5000	21	27	13495
130	mpfi	3.47	2.68	9.0	111	5000	21	27	16500
152	mpfi	2.68	3.47	9.0	154	5000	19	26	16500
109	mpfi	3.19	3.40	10.0	102	5500	24	30	13950
136	mpfi	3.19	3.40	8.0	115	5500	18	22	17450
136	mpfi	3.19	3.40	8.5	110	5500	19	25	15250
136	mpfi	3.19	3.40	8.5	110	5500	19	25	17710
136	mpfi	3.19	3.40	8.5	110	5500	19	25	18920
131	mpfi	3.13	3.40	8.3	140	5500	17	20	23875
108	mpfi	3.50	2.80	8.8	101	5800	23	29	16430
108	mpfi	3.50	2.80	8.8	101	5800	23	29	16925
164	mpfi	3.31	3.19	9.0	121	4250	21	28	20970
164	mpfi	3.31	3.19	9.0	121	4250	21	28	21105
164	mpfi	3.31	3.19	9.0	121	4250	20	25	24565
209	mpfi	3.62	3.39	8.0	182	5400	16	22	30760
209	mpfi	3.62	3.39	8.0	182	5400	16	22	41315
209	mpfi	3.62	3.39	8.0	182	5400	15	20	36880
61	2bbl	2.91	3.03	9.5	48	5100	47	53	5151
90	2bbl	3.03	3.11	9.6	70	5400	38	43	6295
90	2bbl	3.03	3.11	9.6	70	5400	38	43	6575

20 rows × 26 columns

Now, we have successfully read the raw dataset and add the correct headers into the data frame.

Question #2:

Find the name of the columns of the dataframe

[10]:
```
# Write your code below and press Shift+Enter to execute
print(df.columns)
```

```
Index(['symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiration',
       'num-of-doors', 'body-style', 'drive-wheels', 'engine-location',
       'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-type',
       'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'stroke',
       'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg',
       'highway-mpg', 'price'],
      dtype='object')
```

```
Double-click <b>here</b> for the solution.

<!-- The answer is below:

print(df.columns)

-->
```

Save Dataset

Correspondingly, Pandas enables us to save the dataset to csv by using the dataframe.to_csv() method, you can add the file path and name along with quotation marks in the brackets.
For example, if you would save the dataframe **df** as **automobile.csv** to your local machine, you may use the syntax below:

```
df.to_csv("automobile.csv", index=False)
```

We can also read and save other file formats, we can use similar functions to **pd.read_csv()** and **df.to_csv()** for other data formats, the functions are listed in the following table:

Read/Save Other Data Formats

Data Formate	Read	Save
csv	`pd.read_csv()`	`df.to_csv()`
json	`pd.read_json()`	`df.to_json()`
excel	`pd.read_excel()`	`df.to_excel()`
hdf	`pd.read_hdf()`	`df.to_hdf()`
sql	`pd.read_sql()`	`df.to_sql()`
...	...	...

Basic Insight of Dataset

After reading data into Pandas dataframe, it is time for us to explore the dataset.
There are several ways to obtain essential insights of the data to help us better understand our dataset.

Data Types

Data has a variety of types.
The main types stored in Pandas dataframes are **object**, **float**, **int**, **bool** and **datetime64**. In order to better learn about each attribute, it is always good for us to know the data type of each column. In Pandas: dtypes returns a Series with the data type of each column.

```
[12]: # check the data type of data frame "df" by .dtypes
      print(df.dtypes)
```

```
symboling              int64
normalized-losses     object
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                  object
stroke                object
compression-ratio    float64
horsepower            object
peak-rpm              object
city-mpg               int64
highway-mpg            int64
price                 object
dtype: object
```

As a result, as shown above, it is clear to see that the data type of "symboling" and "curb-weight" are int64, "normalized-losses" is object, and "wheel-base" is float64, etc.
These data types can be changed; If you want to learn more about data types or need more preparation before starting this course please visit:

- Coursera: https://www.coursera.org/learn/data-analysis-with-python#syllabus
- CC.ai: https://cognitiveclass.ai/courses/data-analysis-python

Describe

If we would like to get a statistical summary of each column, such as count, column mean value, column standard deviation, etc. We use the describe method:

```
dataframe.describe()
```

[13]:

	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000
mean	0.840796	98.797015	174.200995	65.889055	53.766667	2555.666667	126.875622	10.164279	25.179104	30.686567
std	1.254802	6.066366	12.322175	2.101471	2.447822	517.296727	41.546834	4.004965	6.423220	6.815150
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.800000	64.100000	52.000000	2169.000000	98.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.500000	66.600000	55.500000	2926.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.000000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

This shows the statistical summary of all numeric-typed (int, float) columns.
For example, the attribute "symboling" has 205 counts, the mean value of this column is 0.83, the standard deviation is 1.25, the minimum value is -2, 25th percentile is 0, 50th percentile is 1, 75th percentile is 2, and the maximum value is 3.
However, what if we would also like to check all the columns including those that are of type object.

You can add an argument include = "all" inside the bracket. Let's try it again.

[14]:
```
# describe all the columns in "df"
df.describe(include = "all")
```

[14]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size
count	201.000000	164	201	201	201	199	201	201	201	201.000000	...	201.000000
unique	NaN	51	22	2	2	2	5	3	2	NaN	...	NaN
top	NaN	161	toyota	gas	std	four	sedan	fwd	front	NaN	...	NaN
freq	NaN	11	32	181	165	113	94	118	198	NaN	...	NaN
mean	0.840796	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	98.797015	...	126.875622
std	1.254802	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6.066366	...	41.546834
min	-2.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	86.600000	...	61.000000
25%	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	94.500000	...	98.000000
50%	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	97.000000	...	120.000000
75%	2.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	102.400000	...	141.000000
max	3.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	120.900000	...	326.000000

engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
201.000000	201	197	197	201.000000	199	199	201.000000	201.000000	201
NaN	8	38	36	NaN	58	22	NaN	NaN	186
NaN	mpfi	3.62	3.40	NaN	68	5500	NaN	NaN	8921
NaN	92	23	19	NaN	19	36	NaN	NaN	2
126.875622	NaN	NaN	NaN	10.164279	NaN	NaN	25.179104	30.686567	NaN
41.546834	NaN	NaN	NaN	4.004965	NaN	NaN	6.423220	6.815150	NaN
61.000000	NaN	NaN	NaN	7.000000	NaN	NaN	13.000000	16.000000	NaN
98.000000	NaN	NaN	NaN	8.600000	NaN	NaN	19.000000	25.000000	NaN
120.000000	NaN	NaN	NaN	9.000000	NaN	NaN	24.000000	30.000000	NaN
141.000000	NaN	NaN	NaN	9.400000	NaN	NaN	30.000000	34.000000	NaN
326.000000	NaN	NaN	NaN	23.000000	NaN	NaN	49.000000	54.000000	NaN

11 rows × 26 columns

Now, it provides the statistical summary of all the columns, including object-typed attributes.
We can now see how many unique values, which is the top value and the frequency of top value in the object-typed columns.
Some values in the table above show as "NaN", this is because those numbers are not available regarding a particular column type.

Question #3:

You can select the columns of a data frame by indicating the name of each column, for example, you can select the three columns as follows:

```
dataframe[['column 1', 'column 2', 'column 3']]
```

Where "column" is the name of the column, you can apply the method ".describe()" to get the statistics of those columns as follows:

```
dataframe[['column 1', 'column 2', 'column 3']].describe()
```

Apply the method to ".describe()" to the columns 'length' and 'compression-ratio'.

```
[ ]: # Write your code below and press Shift+Enter to execute
df[['length', 'compression-ratio']].describe()
```

```
Double-click <b>here</b> for the solution.

<!-- The answer is below:

df[['length', 'compression-ratio']].describe()

-->
```

Info

Another method you can use to check your dataset is:

```
dataframe.info()
```

This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.

[15]:
```
# Look at the info of "df"
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 201 entries, 0 to 204
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   symboling          201 non-null    int64
 1   normalized-losses  164 non-null    object
 2   make               201 non-null    object
 3   fuel-type          201 non-null    object
 4   aspiration         201 non-null    object
 5   num-of-doors       199 non-null    object
 6   body-style         201 non-null    object
 7   drive-wheels       201 non-null    object
 8   engine-location    201 non-null    object
 9   wheel-base         201 non-null    float64
 10  length             201 non-null    float64
 11  width              201 non-null    float64
 12  height             201 non-null    float64
 13  curb-weight        201 non-null    int64
 14  engine-type        201 non-null    object
 15  num-of-cylinders   201 non-null    object
 16  engine-size        201 non-null    int64
 17  fuel-system        201 non-null    object
 18  bore               197 non-null    object
 19  stroke             197 non-null    object
 20  compression-ratio  201 non-null    float64
 21  horsepower         199 non-null    object
 22  peak-rpm           199 non-null    object
 23  city-mpg           201 non-null    int64
 24  highway-mpg        201 non-null    int64
 25  price              201 non-null    object
dtypes: float64(5), int64(5), object(16)
memory usage: 42.4+ KB
```

Excellent! You have just completed the Introduction Notebook!

Thank you for completing this lab!

Author

Joseph Santarcangelo

https://github.com/suneelpatel/Statistics-for-Data-Science-using-Python

Welcome to Statistics

Welcome to Statistics! In this module, we will explain how statistics surround our daily lives. All we have to do is to think of the conversations we have on a regular basis. A day starts with this concern about rain or snow. We turn to the weather channel to see whether it will rain or snow today or tomorrow. When the weather channel informs you that the chance of rain is 35% or 60% you are essentially relying on statistical tools and technologies to come up with those forecasts so that you may be better prepared for either rain or snow.

If you happen to live in a large city in North America or Europe or in East Asia, housing affordability is likely to be a concern. And when you hear in the news media that housing is becoming more expensive over time, this analysis is coming out of statistical analysis. At the same time, you will hear if the unemployment rate has fallen or has risen over time, or how millennials are looking for jobs that may not be full-time. And when we track those numbers over time we realize we are using statistics. Statistics are not just confined to economics, we appreciate players based on their performance and we judge their performance using statistics.

Millennials and the sharing economy is perhaps redefining the way we understand economics and business these days. The way millennials have defined their preferences, different from previous

generations, is something of interest. Many say that they would like to rent than to own a house. That didn't used to be the case in the past. But the norms are changing. Who gets paid how much, and not just at your work, but in Hollywood is again coming out of statistics.

Similarly, if you're thinking of pursuing business analytics as a career, you may be interested in knowing what is the average salary of a starting business analyst and, again, this comes out of statistics. Any comparison of salary between two professions such as an engineer or an economist would require statistics. And if you happen to be in Chicago, you probably have not missed that crime has spiked in recent years. Similarly, we compare crime, especially violent crime, over years, and this comparison requires us to use statistics. So, when we say average income, average age, average height, we're relying on average, which is a statistical parameter.

Highest paid athlete: we're looking at the maximum salary; fastest sprinter: we're looking at the maximum speed; lowest unemployment rate of all the OECD countries; we're looking at a minimum value; % of females who study engineering: requires us to compute percentages; the chance for rain tomorrow: is in fact likelihood; and how consistent is a stock performance over the past three months: we're concerned about variance, which again is a statistical parameter. And then this question of an average, "Do men spend more on clothes than women?" we probably would use a t-test to determine this difference, again, relying on statistics.

If you were to recall your conversations in a given day, you probably realize now that you have been using the language of statistics on a daily basis. At the same time the news media use statistics all the time to

demonstrate how trends are changing. 2016 was the year American presidential elections were held. Big surprises there between what the polls forecasted and what the outcome was. But again, you see these numbers portrayed in the newspapers. At the same time, you have other publications that show you how housing prices or other development-related statistics vary over countries. In a nutshell, the information we consume and the conversations we have, every day, involve a lot of statistics. So it pays one to learn some statistics.

The Language of Statistics

- We use Statistics everyday without really being mindful of it
  - Average income, age, height …
  - Highest paid (Maximum) athlete
  - Fastest (Maximum) sprinter
  - Lowest (Minimum) unemployment rate of all OECD countries
  - Percentage of females studying engineering
  - The chance (likelihood) for rain tomorrow
  - How consistent (variance) is a stock performance over the past three months?
  - On average, do men spend more (t-test) on clothes than women?

We see statistics in news media

Early-Stage (January-June) Favorability Ratings
New York Times / CBS News Polls, Since 1976
Incumbents Shaded in Gray

Year	Candidate	Favorable	Unfavorable	Net	Result
1984	Reagan	54	29	+25	WON
1976	Carter	41	21	+20	WON
1988	Dukakis	34	16	+18	LOST
1976	Ford	52	35	+17	LOST
2008	Obama	43	28	+15	WON
1996	Clinton	48	36	+12	WON
1980	Reagan	42	30	+12	WON
2000	Bush	37	31	+6	WON
2008	McCain	35	30	+5	LOST
1984	Mondale	38	34	+4	LOST
2004	Kerry	28	27	+1	LOST
1988	Bush	34	35	-1	WON
2012	Obama	41	42	-1	???
2004	Bush	40	41	-1	WON
1992	Bush	38	41	-3	LOST
2000	Gore	32	36	-4	LOST
1996	Dole	27	37	-10	LOST
1992	Clinton	19	30	-11	WON
2012	Romney	26	37	-11	???
1980	Carter	33	58	-25	LOST
	AVERAGE	37	34	+3	

***The Economist*'s house-price indices**
% change on a year earlier

	Q3 2004*	Q3 2003	1997-2004
South Africa	35.1	20.9	227
Hong Kong	31.2	-13.6	-49
Spain	17.2	16.5	149
New Zealand	16.4	21.2	56
France	14.7	11.5	76
Britain	13.8	11.0	139
United States	13.0	6.0	65
Ireland	10.8	14.8	187
China	9.9	4.1	*na*
Sweden	9.8	5.5	81
Italy	9.7	10.6	69
Belgium	9.3	5.5	50
Australia	8.2	17.6	112
Denmark	7.3	3.4	50
Canada	6.7	6.5	43
Netherlands	3.3	1.9	76
Switzerland	2.2	2.4	12
Singapore	*nil*	-2.3	*na*
Germany	-1.7†	-4.5	-3
Japan	-6.4	-4.8	-24

*Or 2004 latest †Second half 2003

Sources: ABSA; Bulwien; ESRI; Japan Real Estate Institute; Nomisma; NVM; ODPM; OFHEO; Quotable Value; Stadim; Swiss National Bank; government offices

Types of Data

The first step in analytics or statistics is to have a good look at your data and before you begin try to understand what kind of variable you're working with, and based on the type of variable you will decide what kind of analytics could be performed with it. Let's have a look at various different types of data that we encounter and is commonly used in our daily lives. T

The most common one would be a cross-sectional data, which is basically looking at a measurement taken at one point in time. Census in a given year is a cross-section of the society.

As students evaluate course and instructor, that's a cross-section at any given point. Compared to the cross-sectional data, we can have panel or cross-sectional panel data, which is essentially asking the same group of individuals the same questions repeatedly over time. So you may pick a group of people, constitute it as a panel, and then ask the same questions once every year over a given period of time.

The time series data is rather different; you're looking at a particular phenomenon, such as unemployment rate, and then you measure it every month and then display that data or analyze that data, which is repeated measurements on the same phenomena over time -- so you may have monthly data going back to 1940s, or climate data going back to hundreds of years. So, based on the type of data cross-sectional, panel time series, we will pick appropriate tools, statistical tools, to deal with that.

If your dataset has only one variable, it's called the univariate dataset, and if you have multiple variables in your dataset, then it's a multivariate dataset.

Let us now look at variable types and start with categorical or nominal variables.

Let's consider home ownership, for instance, one can either own a home or rent a home. And knowing that there are only two categories here, owning and renting, that is a categorical variable. The tenure status of an individual is essentially a categorical variable. In this particular case, because you only have two choices, own or rent, it's a binomial variable.

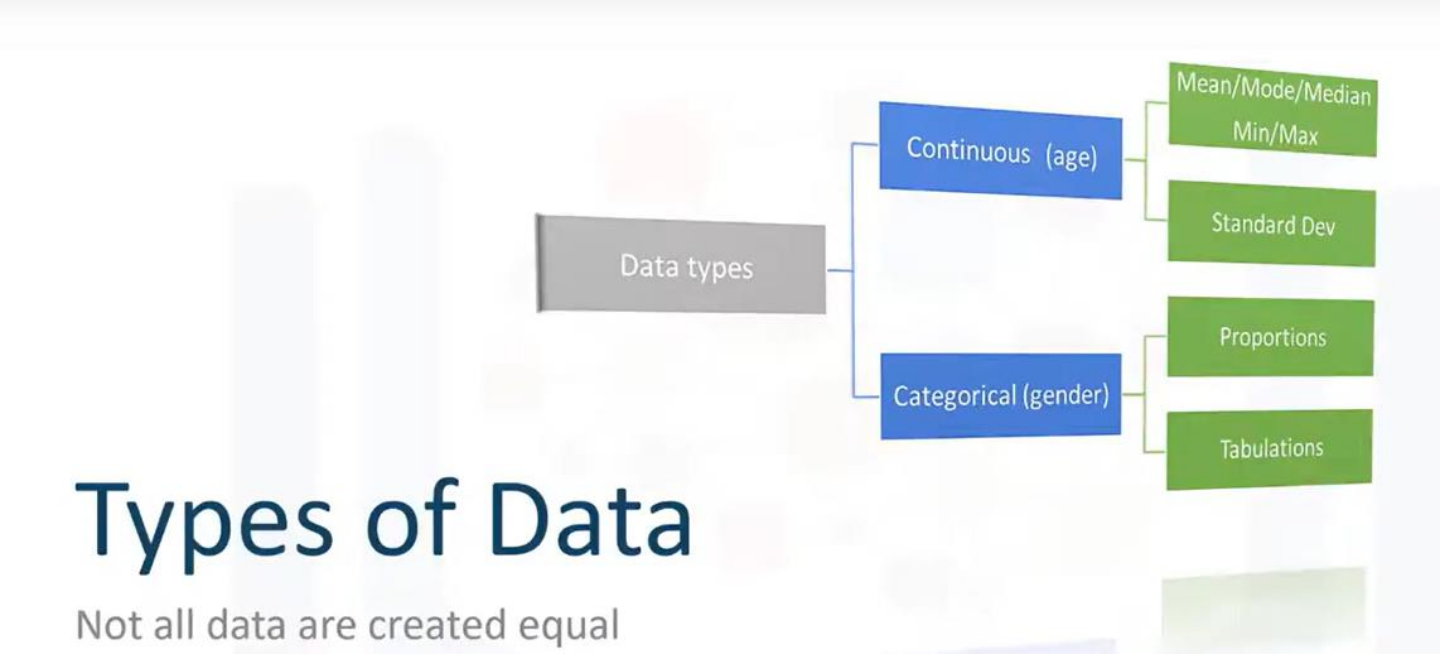
Consider travel choices: you can go to work by driving, or by someone driving you there, so you're a passenger, you can take public transit, or you can walk or bike. So, in this particular case, you have four choices, more than two, so we call it multinomial. So, both binomial and multinomial variables are part of categorical variables.

You cannot have any quantitative relationships among categories, and for these types of variables, averages are usually meaningless. So, if you have a mode of travel, and you have four categories, an average category would mean absolutely nothing of use.

A particular type of categorical variable is ordinal data, where data are ranked or ordered in some particular fashion. So, for instance, number of cars owned by household. A household may have zero car, one car, two cars, three or more cars, and that essentially is an ordinal data, where 0 represents 0, and 0 cannot be coded as 1 and 1 cannot be coded as 0. So the order in which a variable has been recorded matters. Categories can be compared with one another, and you still cannot use regular statistics. The differences are also meaningless, in this particular case.

Another type of data is called ratio data, which is the dataset that have a natural 0, for example, sales dollars, length of a distance, or weight of an object. These are all examples of ratio data, and I often would use the term continuous data, or continuous variable. So, a variable such as distance from point a to b could be 8 kilometers, 8.5 kilometers, 6.2 miles, and the variable is continuous. And, if 0 makes some logical sense, in this particular variable, so for instance, you say I have 0 dollars and 0 means something here. It's the strongest form of measurement; and you can compute ratios and differences.

And another type of variable is interval data or interval variables that are ordered and characterized by a specific measure of distance between observations. And it may not have a natural 0. So, temperature is a good example; and when you say that it's 0 degrees Celsius, it does not mean that there is no temperature; it's freezing, but it is measuring something that exists. And so, ratios are also meaningless. So, for example, if someone said, "Well, you know the temperature in some African countries, 50 degrees compared to somewhere in tropical where it was 25 degrees ,it doesn't mean that the temperatures in the African desert is two times or twice as hot as it is in the tropics, but we can say that there's a difference of 25 degrees between the two places.



Data Types - 1

- Type of Data
  - Cross-Sectional – measurements taken at one time period
    - e.g., students course evaluations in a course
    - Cross sectional Panel
      - Same student's evaluation of different courses in a particular year or in subsequent years
  - Time series – data collected over time
    - E.g., unemployment rate, monthly retail sales
    - Time Series Panel
      - Students' annual satisfaction rating of Ryerson University over 4 years

Data Types - 2

- Number of Variables
  - Univariate– data consisting of a single variable to measure some entity
  - Multivariate– data consisting of two or more variables to measure some entity

Variable Types - 1

- Categorical (nominal) – data sorted into mutually exclusive (an observation cannot belong to more than one category) categories
  - Geographical region, type of employee, gender, state of birth, type of automobile owned
  - Discrete choices
    - Mode of travel (multinomial)
      - Auto drive
      - Auto passenger
      - Public transit
      - Walk/bike
    - Home ownership (binomial)
      - Own
      - Rent
- Properties
  - No quantitative relationships among categories
  - Statistics such as averages are usually meaningless

Variable Types - 2

- Ordinal data – data ordered or ranked according to some relationship to one another
  - Number of cars owned by a household
- Properties
  - Categories can be compared with one another
  - Statistics usually meaningless because of no fixed units of measurement; i.e., differences are meaningless

Variable Types - 3

- Ratio data – data that have a natural zero
  - Sales dollars, length, weight, time from start of a process, most business and economic data
- Properties
  - Strongest form of measurement; both ratios and differences are meaningful

Variable Types - 4

- Interval data – data that are ordered and characterized by a specified measure of distance between observations, but with no natural zero.
  - Temperature scales, time, survey scales that are assumed to be interval
- Properties
  - Ratios are meaningless (50 degrees is not twice as hot as 25 degrees)
  - Differences are meaningful, so statistics such as averages may be compared

Measures of Central Tendency

The **measures of central tendency** are the most commonly used in statistical analysis. We know them as mean, median, and mode, and their use is ubiquitous in statistical analysis.

So let's see how it works. Before we begin, let's take a quick look at our dataset in this course. We have been using the teaching evaluation data from the University of Texas.

The data set comprises of 463 courses, in which we have information about the teaching evaluation score received by the instructor, and we have information about the attributes of the instructor, as well as the characteristics of the course.

Once you have imported a csv file with a Pandas Python library, the first step in getting to know your data is to discover the different data types it contains. You can display all columns and their **data types** with dataframe.info. In this case, we have named our **dataframe ratings_df**. It tells you how many rows you have. For the teaching rating data, we have 463 entries from 0 to 462 because Python starts counting from 0, and then it also gives you information about the data types: object represents strings, int 64 represents integer or whole numbers, and float represents real numbers, which could take on decimal points.

Before we begin, let's have a conversation about population and samples. Essentially, if you have all the information of interest for a particular decision, about every individual that is supposed to be involved in that decision, that is called **a population**. So, if you're interested in looking at some attribute of driving, and we have information about all possible automobile drivers in the U.S. and then we call this the population.

The **sample**, on the other hand, is a subset of population. So, for example, if we have data on all married drivers over the age of 25, then that's a subset. And, within that subset, if we were to randomly select 5% of those married drivers over the age of 25, that would be our sample. We use samples especially in cases where we do not want to incur the cost of collecting data for the entire population. Now, let's consider that there are 230 million individuals in the country. A sample size of say, 330 to 500 individuals, randomly selected, would suffice. This reduces the cost, especially in cases where you cannot collect information for the entire population. Therefore, using samples, it's really helpful and cost effective.

Here you see some Greek symbols on the screen, but don't be afraid; they mostly show the formula. We will then proceed from here. While they may differ in notation, essentially, the **mean** for a population and sample are the same. It is the sum of all the observations, then divided by the number of observations, to get the mean, which we call **averages**.

There are several properties of the mean, and they're meaningful, but one of the characteristics of a mean is that if you take the difference between the average value for a variable and subtract from all the observations and sum them up, that sum would be equal to 0.

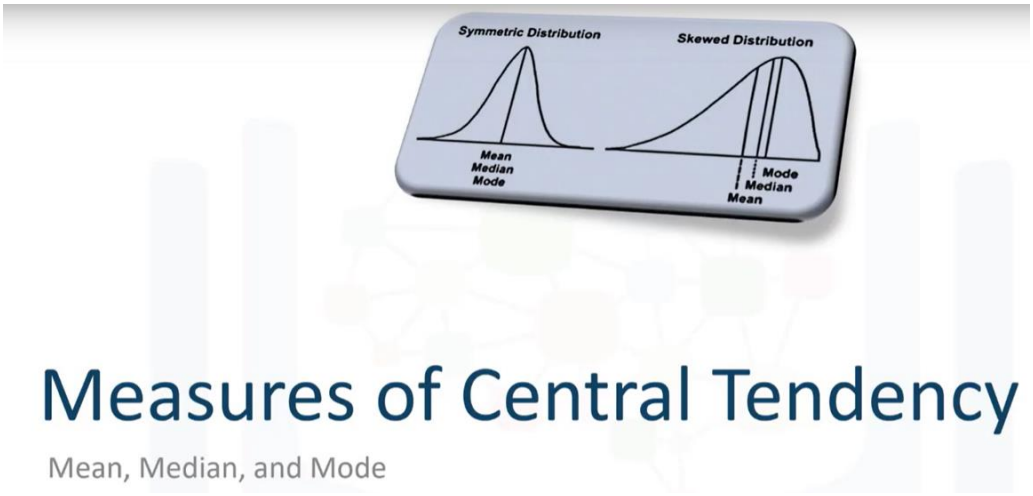
The **median** is different from the mean. When you order the data from the smallest value to the largest value, the result is in the middle, that is, the value in the middle indicating that there are an equal number of observations that are above, and the equal number of observations are below that family. That value is called the median. So, if the median salary in some city is $45,000,

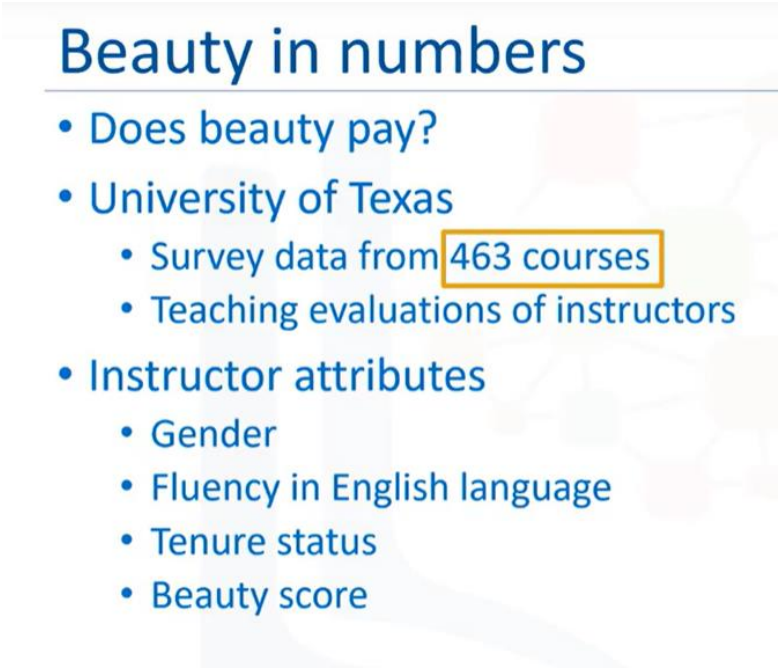
it means that 50% of the people make more than $45,000, and the other 50% make less than $45,000.

Mode is essentially the value that occurs most frequently. Therefore, if the most common age in a class of students is 16, then that's the mode.

We will now turn to Python for our hands-on training to estimate the summary statistics values for beauty score, teaching you about evaluation and age. We will use the **dataframe.describe** function to find the summary statistics.

This prints out the number of rows, mean, standard deviation, minimum value, 25th, 50th, and 75th percentile, and the maximum value. To find the summary statistics for a subset of the variables, you will have to state the column names, as we can see here. Otherwise, for the full population, we will call the .describe function on the dataframe.





Teaching Evaluation Data

```
## get information about each variable
ratings_df.info()
```

```
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
minority          463 non-null object
age               463 non-null int64
gender            463 non-null object
credits           463 non-null object
beauty            463 non-null float64
eval              463 non-null float64
division          463 non-null object
native            463 non-null object
tenure            463 non-null object
students          463 non-null int64
allstudents       463 non-null int64
prof              463 non-null int64
PrimaryLast       463 non-null int64
vismin            463 non-null int64
female            463 non-null int64
single_credit     463 non-null int64
upper_division    463 non-null int64
English_speaker   463 non-null int64
tenured_prof      463 non-null int64
dtypes: float64(2), int64(11), object(6)
```

Teaching Evaluation Data

```
1 ## get information about each variable
2 ratings_df.info()
```

```
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
minority            463 non-null object
age                 463 non-null int64
gender              463 non-null object
credits             463 non-null object
beauty              463 non-null float64
eval                463 non-null float64
division            463 non-null object
native              463 non-null object
tenure              463 non-null object
students            463 non-null int64
allstudents         463 non-null int64
prof                463 non-null int64
PrimaryLast         463 non-null int64
vismin              463 non-null int64
female              463 non-null int64
single_credit       463 non-null int64
upper_division      463 non-null int64
English_speaker     463 non-null int64
tenured_prof        463 non-null int64
dtypes: float64(2), int64(11), object(6)
```

Populations and Samples

- Population – all items of interest for a particular decision or investigation
  - All drivers in the U.S.
  - All individuals who do not own a cell phone
- Sample – a subset of a population
  - All married drivers in the U.S. over age 25
- Why samples are used?
  - To reduce costs of data collection
  - When a full census cannot be taken

Terminology and Notation

- x_i represents the i^{th} observation
- $\sum$ indicates the operation of addition
- N is the size of the population; n is the size of the sample
- f_i is the number of observations in cell i of a frequency distribution

Mean or Average

- Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

It is the sum of all the observations, then divided by the number of observations, to get the mean, which we call **averages**.

- Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Properties of the Mean

- Meaningful for interval and ratio data (continuous variables)
- Affected by unusually large or small observations (outliers)
  - Hence median is also useful
- The only measure of central tendency where the sum of the deviations of each value from the measure is zero; i.e.,

$$\sum(x_i - \bar{x}) = 0$$

Median

- Middle value when data are ordered from smallest to largest. This results in an equal number of observations above the median as below it
  - Unique for each set of data
  - Not affected by extremes
  - Meaningful for ratio, interval, and ordinal data

When you order the data from the smallest value to the largest value, the result is in the middle, that is, the value in the middle indicating that there are an equal number of observations that are above, and the equal number of observations are below that family. That value is called the median. So, if the median salary in some city is $45,000, it means that 50% of the people make more than $45,000, and the other 50% make less than $45,000.

Mode

- Observation that occurs most frequently; for grouped data, the midpoint of the cell with the largest frequency (approximate value)
  - Useful when data consist of a small number of unique values

Let's calculate the average beauty!

	beauty	eval	age
count	4.630000e+02	463.000000	463.000000
mean	6.271140e-08	3.998272	48.365011
std	7.886477e-01	0.554866	9.802742
min	-1.450494e+00	2.100000	29.000000
25%	-6.562689e-01	3.600000	42.000000
50%	-6.801430e-02	4.000000	48.000000
75%	5.456024e-01	4.400000	57.000000
max	1.970023e+00	5.000000	73.000000

Let's calculate the average beauty!

```
1  ratings_df[['beauty', 'eval', 'age']].describe()
```


	beauty	eval	age
count	4.630000e+02	463.000000	463.000000
mean	6.271140e-08	3.998272	48.365011
std	7.886477e-01	0.554866	9.802742
min	-1.450494e+00	2.100000	29.000000
25%	-6.562689e-01	3.600000	42.000000
50%	-6.801430e-02	4.000000	48.000000
75%	5.456024e-01	4.400000	57.000000
max	1.970023e+00	5.000000	73.000000

Measures of Dispersion

Dispersion, which is also called variability, scatter, or spread, is the extent to which the data distribution is stretched or squeezed. The common measures of dispersion are standard deviation and variance, and if you are at a university or college, you may have heard about the bell curve, which looks like this. And you will

often hear, "This is within one standard deviation of the mean," or, "Within two standard deviations of the mean."

So, let's see what that means. Let's look at the age of an instructor. Let's say the average age is 52. This means that the individual ages may differ, some may be 48, or maybe 55, or 75. So, the average age is an estimate. But what we also need is an estimate for the dispersion in the dataset. The other thing to note is the range in our data set, for example, the difference of the range is from a minimum of 29 years of age to a maximum of 73 years, and this to you refers to a distance, or the difference between the minimum and the maximum.

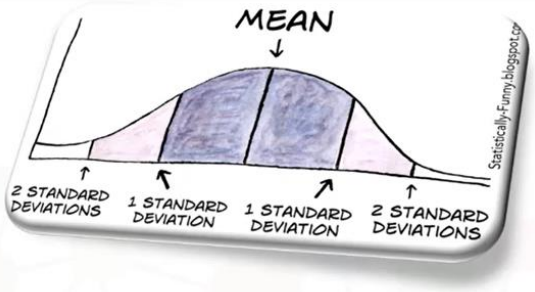
Unlike the difference between population and sample mean, the difference between a variance for the population and a sample, is that when you compute the population variance, denoted as sigma squared, you divided by the total number of observations, these are the deviations between observation and the mean squared, then added, and then divided by the total number of observations. For sample variance, which is denoted as s squared, you divided by n minus 1. The purpose of using n minus 1 is so that our estimate is unbiased in the long run. That means that if we take a second sample, we'll get a different value of s squared. If we take a third sample, we'll get a third value of s squared, and so on.

We use n minus 1 so that the average of all these values of s squared is equal to sigma squared. We usually talk about squares called standard deviation rather than the variance. **Standard deviation** is essentially the square root of the variance, or the variances in square units, so it's good to use the standard deviation because it's exactly the same units as the variable.

The standard deviation of age will also be measured in years, rather than years squared. Here you see that we just took the square root of variance, and this becomes standard deviation. We will return to our dataset and we'll look at the variables that we computed before. You can see that the standard deviation for beauty, evaluation scores, and age were computed with the descriptive statistics using the describe function in Python Pandas.

Let me explain why mean and standard deviation have to go hand in hand. I will use the example from basketball about the two giants, Michael Jordan and Wilt Chamberlain, who preceded Michael Jordan. Now, if you consider their average score per game, you would notice they didn't differ much. Their average was around 30 points for both Jordan and Chamberlain, however, when you look at the standard deviation of their performance, Jordan was around 4.76, compared to Chamberlain who was at around 10.59. If you were to plot this distribution to see Michael Jordan's scores using the mean and standard deviation, assuming that their scores are normally distributed, you would notice, even though both players had around the same mean, the tighter distribution for Jordan suggests that he was more consistent in his performance than Chamberlain.

The main takeaway is that average will only paint a partial picture. If you really want to understand the complete picture about a variable or dataset, it is important to compute both the average and the standard deviation to get insights on what the data is telling us. So, a mean, with a standard deviation, means something more useful than the mean by itself.



Measures of Dispersion

The Standard Deviants

Measures of Dispersion

- Dispersion is the degree of variation in the data
  - E.g., the age of instructors {48, 49, 50, 51, 52}
- Range is the difference between the maximum and minimum observations
  - The minimum age of an instructor was 29 and maximum age was 73

Variance

- Population $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- Sample $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Standard Deviation

- The standard deviation is the square root of the variance
- The variance is in "square units" so the standard deviation is in the same units as x

- Population $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Sample $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard deviation is essentially the square root of the variance, or the variances in square units, so it's good to use the standard deviation because it's exactly the same units as the variable. The standard deviation of age will also be measured in years, rather than years squared. Here you see that we just took the square root of variance, and this becomes standard deviation.

The aging, beautiful standard deviants

```
1  ratings_df[['beauty', 'eval', 'age']].describe()
```

	beauty	eval	age
count	4.630000e+02	463.000000	463.000000
mean	6.271140e-08	3.998272	48.365011
std	7.886477e-01	0.554866	9.802742
min	-1.450494e+00	2.100000	29.000000
25%	-6.562689e-01	3.600000	42.000000
50%	-6.801430e-02	4.000000	48.000000
75%	5.456024e-01	4.400000	57.000000
max	1.970023e+00	5.000000	73.000000

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players
- On average, they both scored almost the same points per game
  - Mean – 30.12 for Jordan
  - Mean – 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD – Jordan: 4.76
  - SD – Chamberlain: 10.59

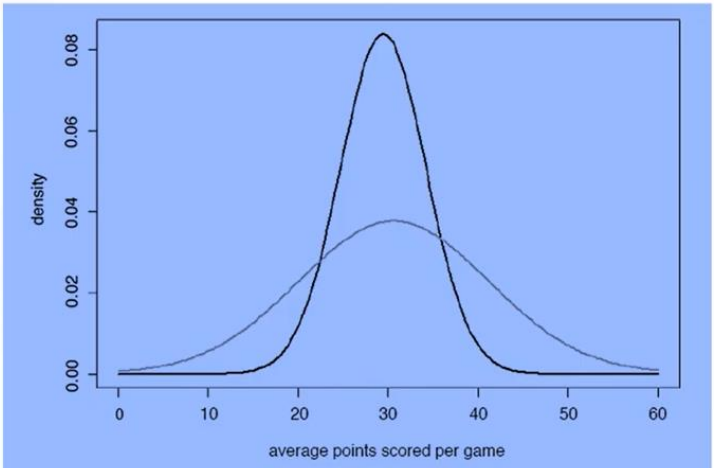


Figure 6.15 Normal distribution curves for Michael Jordan and Wilt Chamberlain

Source: Chapter 6: *Getting Started with Data Science: Making Sense of Data with Analytics*

Let me explain why mean and standard deviation have to go hand in hand. I will use the example from basketball about the two giants, Michael Jordan and Wilt Chamberlain, who preceded Michael Jordan. Now, if you consider their average score per game, you would notice they didn't differ much. Their average was around 30 points for both Jordan and Chamberlain, however, when you look at the standard deviation of their performance, Jordan was around 4.76, compared to Chamberlain who was at around 10.59. If you were to plot this distribution to see Michael Jordan's scores using the mean and standard deviation, assuming that their scores are normally distributed, you would notice, even though both players had around the same mean, the tighter distribution for Jordan suggests that he was more consistent in his performance than Chamberlain.

Reliability

- Average paints a partial picture
- Average statistics are incomplete without standard deviation/variance
- Risk metrics are all about variance

Measure of Center

In [0]:
```
import numpy as np
x=[2,4,6,7,20,10,22]
y=np.array(x)
```

Mean:

In [3]:
```
print("Mean is : ",y.mean())
```

```
Mean is :  10.142857142857142
```

Median:

In [4]:
```
print("Median is : ",np.median(y))
```

```
Median is :  7.0
```

Mode:

In [5]:
```
from statistics import mode
print("Mode is:",mode([1, 1, 2, 3, 3, 3, 3, 4]))
```

```
Mode is: 3
```

15 lines (10 sloc) | 268 Bytes

```
1  import numpy as np
2
3  x=[2,4,6,7,20,10,22]
4  y=np.array(x)
5
6  print("Mean is : ",y.mean())
7
8  print("Median is : ",np.median(y))
9  print("\n")
10 print("Mean is : ",y.mean())
11
12 print("\n")
13
14 from statistics import mode
15 print("Mode is:",mode([1, 1, 2, 3, 3, 3, 3, 4]))
```

Range :

Range = X(lagest) - X (lowest)

In [4]:
```
import numpy as np
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5], [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
A

B=A.T
B

a=np.ptp(B, axis=0)
b=np.ptp(B,axis=1)

print("Range in Array A:",a)
print("Range in Array B:",b)
```

```
Range in Array A: [12.  10.   8.   6.5]
Range in Array B: [ 7.   6.5  8.   7.5  4.5 11.  11.5]
```

Quartile

In [5]:
```
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5], [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])

B=A.T

a=np.percentile(B,27,axis=0, interpolation='lower')
b=np.percentile(B,25,axis=1, interpolation='lower')
c=np.percentile(B,75,axis=0, interpolation='lower')
d=np.percentile(B,50,axis=0, interpolation='lower')

print(a)

print(b)

print(c)

print(d)
```

```
[9.5 9.  7.5 9. ]
[8.  7.5 9.  7.  9.5 7.  7.5]
[14.  15.5 11.  12. ]
[11.  14.5 10.5 11.5]
```

inter-qurtile range

In [7]:

```
import numpy as np
from scipy.stats import iqr
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5], [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])

B=A.T

a=iqr(B, axis=0 , rng=(25, 75), interpolation='lower')
b=iqr(B, axis=1 , rng=(25, 75), interpolation='lower')

print(a,b)
```

```
[4.5 6.5 3.5 3. ] [3.5 3.5 2.5 5.  2.5 8.  8. ]
```

Variance

In [8]:
```
import numpy as np
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5],
  [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])

B=A.T

a = np.var(B,axis=0)
b = np.var(B,axis=1)

print(a)

print(b)
```

```
[13.98979592 12.8877551   6.12244898  3.92857143]
[ 6.546875  5.921875  8.796875  7.546875  2.875    16.5      19.0625  ]
```

Standard deviation

In [9]:
```
import numpy as np
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5],
  [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
B=A.T
a = np.std(B,axis=0)
b = np.std(B,axis=1)
print(a)

print(b)
```

```
[3.74029356 3.58995196 2.4743583  1.98206242]
[2.55868619 2.43349029 2.96595263 2.74715762 1.6955825  4.0620192
 4.3660623 ]
```

```
#Range
import numpy as np
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5], [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
A

B=A.T
B

a=np.ptp(B, axis=0)
b=np.ptp(B,axis=1)

print(a)
print(b)


#Quartile
A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5], [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])

B=A.T

a=np.percentile(B,27,axis=0, interpolation='lower')
b=np.percentile(B,25,axis=1, interpolation='lower')
c=np.percentile(B,75,axis=0, interpolation='lower')
d=np.percentile(B,50,axis=0, interpolation='lower')

print(a)

print(b)

print(c)

print(d)


```

```
35
36  #inter-qurtile range
37  import numpy as np
38  from scipy.stats import iqr
39  A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5] [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
40
41  B=A.T
42
43  a=iqr(B, axis=0 , rng=(25, 75), interpolation='lower')
44  b=iqr(B, axis=1 , rng=(25, 75), interpolation='lower')
45
46  print(a,b)
47
48
49  #Variance
50
51  import numpy as np
52  A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5],
53    [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
54
55  B=A.T
56
57  a = np.var(B,axis=0)
58  b = np.var(B,axis=1)
59
60  print(a)
61
62  print(b)
63
64
65  #Standard deviation
66  import numpy as np
67  A=np.array([[10,14,11,7,9.5,15,19],[8,9,17,14.5,12,18,15.5],
68    [15,7.5,11.5,10,10.5,7,11],[11.5,11,9,12,14,12,7.5]])
69  B=A.T
70  a = np.std(B,axis=0)
71  b = np.std(B,axis=1)
72  print(a)
73
74  print(b)
```

https://github.com/suneelpatel/Statistics-for-Data-Science-using-Python/blob/master/Descriptive_Statstics_using_Python/Measure_of_Spread.py

Descriptive Statistics

Estimated time needed: **30** minutes

In this lab, you'll go over some hands-on exercises using Python.

Objectives

- Import Libraries
- Read in Data
- Lab exercises and questions

Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[1]:
```
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install matplotlib=3.4.3-y
```

Import the libraries we need for the lab

[4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as pyplot
```

Read in the csv file from the URL using the request library

[3]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.csv'
ratings_df=pd.read_csv(ratings_url)
```

Data Description

Variable Description

minority	Does the instructor belong to a minority (non-Caucasian) group?
age	The professor's age

Variable	Description
gender	Indicating whether the instructor was male or female.
credits	Is the course a single-credit elective?
beauty	Rating of the instructor's physical appearance by a panel of six students averaged across the six panelists and standardized to have a mean of zero.
eval	Course overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent).
division	Is the course an upper or lower division course?
native	Is the instructor a native English speaker?
tenure	Is the instructor on a tenure track?
students	Number of students that participated in the evaluation.
allstudents	Number of students enrolled in the course.
prof	Indicating instructor identifier.

Display information about the dataset

1. Structure of the dataframe
2. Describe the dataset
3. Number of rows and columns

print out the first five rows of the data

```
[5]: ratings_df.head()
```

[5]:

	minority	age	gender	credits	beauty	eval	division	native	tenure	students	allstudents
0	yes	36	female	more	0.289916	4.3	upper	yes	yes	24	43
1	yes	36	female	more	0.289916	3.7	upper	yes	yes	86	125
2	yes	36	female	more	0.289916	3.6	upper	yes	yes	76	125
3	yes	36	female	more	0.289916	4.4	upper	yes	yes	77	123
4	no	59	male	more	-0.737732	4.5	upper	yes	yes	17	20

prof	PrimaryLast	vismin	female	single_credit	upper_division	English_speaker	tenured_prof
1	0	1	1	0	1	1	1
1	0	1	1	0	1	1	1
1	0	1	1	0	1	1	1
1	1	1	1	0	1	1	1
2	0	0	0	0	1	1	1

get information about each variable

[6]:
```
ratings_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   minority         463 non-null    object
 1   age              463 non-null    int64
 2   gender           463 non-null    object
 3   credits          463 non-null    object
 4   beauty           463 non-null    float64
 5   eval             463 non-null    float64
 6   division         463 non-null    object
 7   native           463 non-null    object
 8   tenure           463 non-null    object
 9   students         463 non-null    int64
 10  allstudents      463 non-null    int64
 11  prof             463 non-null    int64
 12  PrimaryLast      463 non-null    int64
 13  vismin           463 non-null    int64
 14  female           463 non-null    int64
 15  single_credit    463 non-null    int64
 16  upper_division   463 non-null    int64
 17  English_speaker  463 non-null    int64
 18  tenured_prof     463 non-null    int64
dtypes: float64(2), int64(11), object(6)
memory usage: 68.9+ KB
```

get information about each variable

[6]:
```
ratings_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   minority         463 non-null    object
 1   age              463 non-null    int64
 2   gender           463 non-null    object
 3   credits          463 non-null    object
 4   beauty           463 non-null    float64
 5   eval             463 non-null    float64
 6   division         463 non-null    object
 7   native           463 non-null    object
 8   tenure           463 non-null    object
 9   students         463 non-null    int64
 10  allstudents      463 non-null    int64
 11  prof             463 non-null    int64
 12  PrimaryLast      463 non-null    int64
 13  vismin           463 non-null    int64
 14  female           463 non-null    int64
 15  single_credit    463 non-null    int64
 16  upper_division   463 non-null    int64
 17  English_speaker  463 non-null    int64
 18  tenured_prof     463 non-null    int64
dtypes: float64(2), int64(11), object(6)
memory usage: 68.9+ KB
```

[7]:
```
ratings_df.shape
```

[7]:
```
(463, 19)
```

Lab Exercises

Can you identify whether the teachers' Rating data is a time series or cross-sectional?

Print out the first ten rows of the data

1. Does it have a date or time variable? - No - it is not a time series dataset
2. Does it observe more than one teacher being rated? - Yes - it is cross-sectional dataset

The dataset is a Cross-sectional

[8]:
```
ratings_df.head(10)
```

[8]:

	minority	age	gender	credits	beauty	eval	division	native	tenure	students
0	yes	36	female	more	0.289916	4.3	upper	yes	yes	24
1	yes	36	female	more	0.289916	3.7	upper	yes	yes	86
2	yes	36	female	more	0.289916	3.6	upper	yes	yes	76
3	yes	36	female	more	0.289916	4.4	upper	yes	yes	77
4	no	59	male	more	-0.737732	4.5	upper	yes	yes	17
5	no	59	male	more	-0.737732	4.0	upper	yes	yes	35
6	no	59	male	more	-0.737732	2.1	upper	yes	yes	39
7	no	51	male	more	-0.571984	3.7	upper	yes	yes	55
8	no	51	male	more	-0.571984	3.2	upper	yes	yes	111
9	no	40	female	more	-0.677963	4.3	upper	yes	yes	40

allstudents	prof	PrimaryLast	vismin	female	single_credit	upper_division	English_speaker	tenured_prof
43	1	0	1	1	0	1	1	1
125	1	0	1	1	0	1	1	1
125	1	0	1	1	0	1	1	1
123	1	1	1	1	0	1	1	1
20	2	0	0	0	0	1	1	1
40	2	0	0	0	0	1	1	1
44	2	1	0	0	0	1	1	1
55	3	0	0	0	0	1	1	1
195	3	1	0	0	0	1	1	1
46	4	0	0	1	0	1	1	1

Find the mean, median, minimum, and maximum values for students

Find Mean value for students

[9]:
```
ratings_df['students'].mean()
```

[9]:
```
36.62419006479482
```

Find the Median value for students

[11]:
```
ratings_df['students'].median()
```

[11]:
```
23.0
```

Find the Minimum value for students

[12]:
```
ratings_df['students'].min()
```

[12]:
```
5
```

Find the Maximum value for students

[13]:
```
ratings_df['students'].max()
```

[13]:
```
380
```

Produce a descriptive statistics table

[14]:
```
ratings_df.describe()
```

[14]:

	age	beauty	eval	students	allstudents	prof	PrimaryLast	4
count	463.000000	4.630000e+02	463.000000	463.000000	463.000000	463.000000	463.000000	
mean	48.365011	6.271140e-08	3.998272	36.624190	55.177106	45.434125	0.203024	
std	9.802742	7.886477e-01	0.554866	45.018481	75.072800	27.508902	0.402685	
min	29.000000	-1.450494e+00	2.100000	5.000000	8.000000	1.000000	0.000000	
25%	42.000000	-6.562689e-01	3.600000	15.000000	19.000000	20.000000	0.000000	
50%	48.000000	-6.801430e-02	4.000000	23.000000	29.000000	44.000000	0.000000	
75%	57.000000	5.456024e-01	4.400000	40.000000	60.000000	70.500000	0.000000	

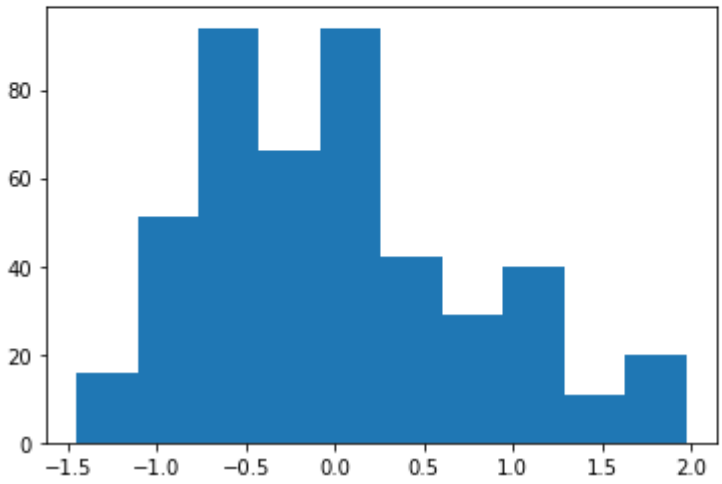
vismin	female	single_credit	upper_division	English_speaker	tenured_prof	1.000000
463.000000	463.000000	463.000000	463.000000	463.000000	463.000000	
0.138229	0.421166	0.058315	0.660907	0.939525	0.779698	
0.345513	0.494280	0.234592	0.473913	0.238623	0.414899	
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	
0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	
0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

Create a histogram of the beauty variable and briefly comment on the distribution of data

using the matplotlib library, create a histogram

```
[15]: pyplot.hist(ratings_df['beauty'])
```

```
[15]: (array([16., 51., 94., 66., 94., 42., 29., 40., 11., 20.]),
       array([-1.45049405, -1.10844234, -0.76639063, -0.42433892, -0.08228722,
              0.25976449,  0.6018162 ,  0.94386791,  1.28591962,  1.62797133,
              1.97002304]),
       <BarContainer object of 10 artists>)
```



here are few conclusions from the histogram most of the data for beauty is around the -0.5 and 0 the distribution is skewed to the right therefore looking at the data we can say the mean is close to 0

Does average beauty score differ by gender? Produce the means and standard deviations for both male and female instructors.

Use a group by gender to view the mean scores of the beauty we can say that beauty scores differ by gender as the mean beauty score for women is higher than men

[16]:
```
ratings_df.groupby('gender').agg({'beauty':['mean', 'std', 'var']}).reset_index()
```

[16]:

gender		beauty		
		mean	std	var
0	female	0.116109	0.81781	0.668813
1	male	-0.084482	0.75713	0.573246

Calculate the percentage of males and females that are tenured professors. Will you say that tenure status differ by gender?

First groupby to get the total sum

[18]:
```
tenure_count = ratings_df[ratings_df.tenure == 'yes'].groupby('gender').agg({'tenure': 'count'}).reset_index()
```

Find the percentage

[19]:
```
tenure_count['percentage'] = 100 * tenure_count.tenure/tenure_count.tenure.sum()
tenure_count
```

[19]:

	gender	tenure	percentage
0	female	145	40.166205
1	male	216	59.833795

Practice Questions

Question 1: Calculate the percentage of visible minorities are tenure professors. Will you say that tenure status differed if teacher was a visible minority?

[23]:
```
### we can use a groupby function for this
## first groupby to get the total sum
tenure_count = ratings_df.groupby('minority').agg({'tenure': 'count'}).reset_index()
# Find the percentage
tenure_count['percentage'] = 100 * tenure_count.tenure/tenure_count.tenure.sum()
##print to see
tenure_count
```

[23]:

	minority	tenure	percentage
0	no	399	86.177106
1	yes	64	13.822894

Double-click **here** for the solution.

Question 2: Does average age differ by tenure? Produce the means and standard deviations for both tenured and untenured professors

[21]:
```
## group by tenureship and find the mean and standard deviation for each group
ratings_df.groupby('tenure').agg({'age':['mean', 'std']}).reset_index()
```

[21]:

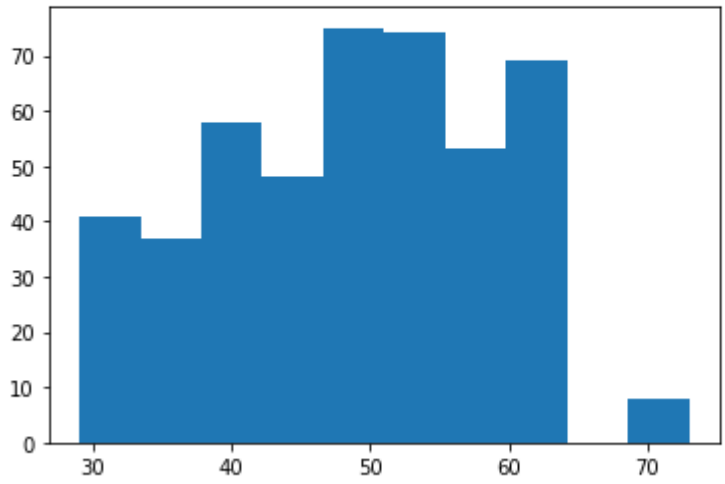
	tenure	age	
		mean	std
0	no	50.186275	6.946372
1	yes	47.850416	10.420056

Double-click **here** for the solution.

Question 3: Create a histogram for the age variable

[22]:
```
pyplot.hist(ratings_df['age'])
```

[22]:
```
(array([41., 37., 58., 48., 75., 74., 53., 69.,  0.,  8.]),
 array([29. , 33.4, 37.8, 42.2, 46.6, 51. , 55.4, 59.8, 64.2, 68.6, 73. ])
 <BarContainer object of 10 artists>)
```



Double-click **here** for the solution.

Question 4: What is the Median evaluation score for tenured Professors?

[24]:
```
## insert code here
## you can index just tenured professors and find their median evaluation scores
ratings_df[ratings_df['tenure'] == 'yes']['eval'].median()
```

[24]:
```
4.0
```

Quiz Introduction to Descriptive Statistics

Bookmarked

Graded Quiz due May 2, 2022 08:46 +08 Completed

Multiple Choice

11/11 points (graded)

What is the difference between the maximum and minimum data entries in the set?

- (●) Range
- ( ) Mode
- ( ) Mean
- ( ) Variance

✔

Answer
Correct: Correct!

Find the median of the data set. 3 ,8 ,9 ,11, 12, 15

- ○ 9
- ◉ 10
- ○ 11
- ○ 12

✔

Answer
Correct: Correct!

The measurements of spread or scatter of the individual values about the central point is called:

- ◉ Measures of dispersion
- ○ Measures of central tendency
- ○ Measure of skewness
- ○ Measures of central tendency and measures of dispersion

✔

Answer
Correct: Correct!

Which of the following is an example of time series data?

- ○ Number of dolphins in the Pacific Ocean
- ○ Number of trees in Jardin du Luxemburg in Paris
- ○ Batting average of a baseball player
- ◉ Annual average housing price in New York City

✔

Answer
Correct:
Correct! A time series is a sequence taken at successive, equally spaced points in time. This is a time series with data taken in equally spaced year-long intervals.

What is the 25th percentile of the following data set?

- ◉ 3
- ○ 1
- ○ 5.5
- ○ 3.5

✔

Answer
Correct: Correct!

Which of the following is a measure of variability?

- ○ Mean
- ○ Median
- ◉ Variance
- ○ Mode

✔

Answer
Correct: Correct!

Which of the following measures of central tendency will always change if a single value in the data changes?

- ◉ Mean
- ○ Median
- ○ Mode
- ○ All of the above

✔

Answer
Correct: Correct!

Which of the following datasets has a mean of 10 and a standard deviation of 0?

- ○ 0, 10, 20
- ◉ 10, 10, 10
- ○ 0, 0, 0
- ○ 15, 15, 15

✔

Answer
Correct:
Correct! Many data sets can have a mean of 10. However, if you force the standard deviation to be 0, you have only one choice: 10, 10, 10. A standard deviation of 0 means the average distance from the data values to the mean is 0. In other words, the data values don't deviate from the mean at all, and hence they have to be the same value.

What is meta data?

- ○ The metabolism data in a clinical trial
- ○ The data about metamorphism
- ◉ It's the data about data
- ○ Data about metal fatigue

✔

Answer
Correct: Correct!

Which of the following is an example of categorical data?

- ○ Number of children at a kindergarten
- ○ Length of the river Nile
- ○ Number of fire hydrants in Toronto
- ◉ Mode of travel of work

✔

Answer
Correct: Correct!

If the variance of a dataset is correctly computed with the formula using (n-1) in the denominator, which of the following options is true?

- ◉ The dataset is a sample
- ○ The dataset is a population
- ○ The data contains other variables with categorical data
- ○ The data is from an unknown source

✔

Answer
Correct: Correct!

Submit You have used 2 of 2 attempts Show answer

Visualization Fundamentals

Let us begin with the fundamentals of visualization.

What you see on the right side is a pictograph or a bar chart prepared by students in the kindergarten class, four year-olds, at the St. Luke's Catholic School in Mississauga. This bar chart presents the frequency for transportation modes the kids have used to come to school, some have been dropped to school by car, others by bus, and two students walk to school.

In the brave big world of big data, we can see that children are being trained at the earliest possible age with data and data science. The most important thing to realize is that the type of visualization you will use depends upon the type of variables you're trying to analyze.

For instance, if you're working with categorical variables, such as, gender, you have to rely on a certain type of charting tools, than if you were to be working with continuous variables, such as, age and income.

In one case, you may be using bar charts, in another case, you will be required to use scatter plots.

I would like to draw your attention to the extreme presentation method developed by Dr. Andrew Abela.

Essentially, it lays out the possible ways of depicting data, based on what kind of variables you have at your disposal. This visualization, this graphic, is developed by Dr. Abela that shows that if you are interested in comparing variables, or demonstrating their distribution, or composition, or the relationship between two variables or more, you have to rely on specific type of graph.

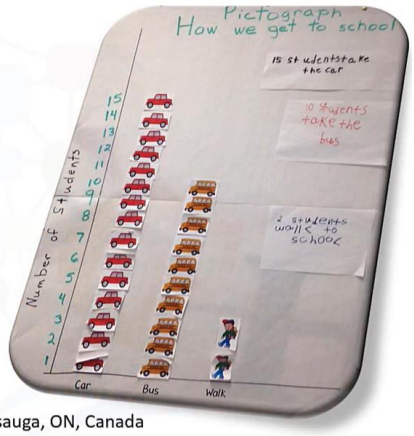
If you're, you can use

1. bar charts or column charts - comparing items with few categories
2. line chart - comparing behaviors over time, and if you have the time periods running into months, several months, and, then you can use
3. columns and other approaches - if the time periods are not that many
4. scatter plot - depict relationships between two continuous variables
5. bubble chart - depict two variables on x and y-axis the third variable will be depicted by the size of the circle, so you can essentially have 3 variables
6. histogram (could be a bar chart type of histogram, or a line histogram/ scatterplots - depicting the distribution of the dataset
7. pie charts - show the composition, and if it's static data
8. stacked columns - showing the composition that changes over time, for few periods
9. stack area charts - showing the composition that changes over time, for several periods

For hands-on in Python, we will be using the:

seaborn library and the matplotlib library to create visualizations in the labs. We will learn how to use different functions within the library to create different kinds of charts.

Visualization Fundamentals



Kindergarten students at
St. Luke Catholic School, Mississauga, ON, Canada

Variable type Matters

Chart types should be based on the type of variables being depicted

Examples:

- Categorical variables:
  - Counts, e.g., How many instructors are male and how many are female
- Continuous variables:
  - Scatter plots
  - Statistical properties (averages)

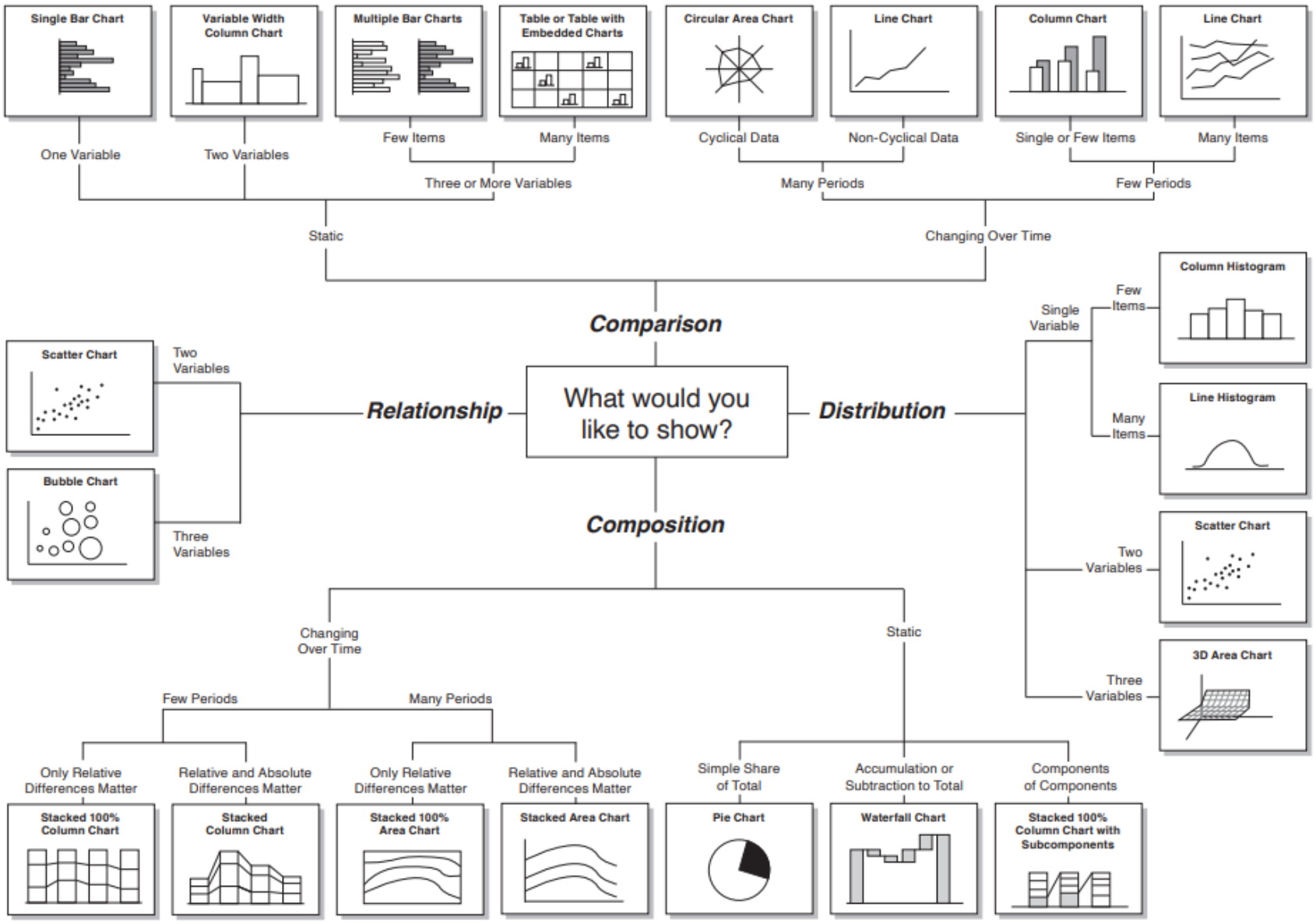
The Extreme Presentation Method

Dr. Andrew Abela:

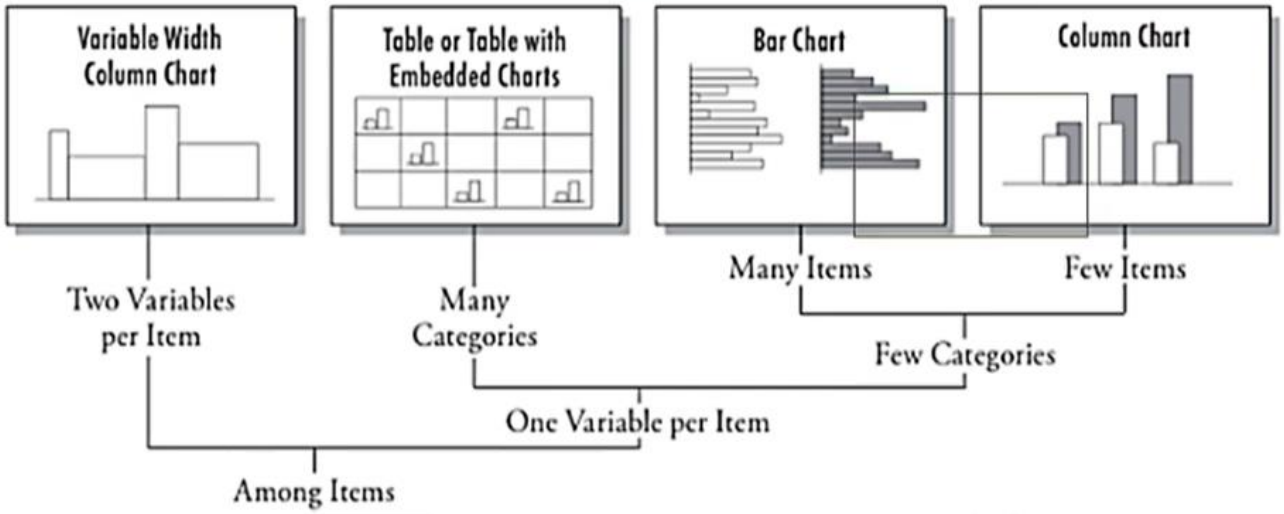
- The Extreme Presentation method is a step-by-step approach for designing presentations of complex or controversial information in ways that drive people to action.
- https://extremepresentation.com/

Chart suggestions

Chart Chooser

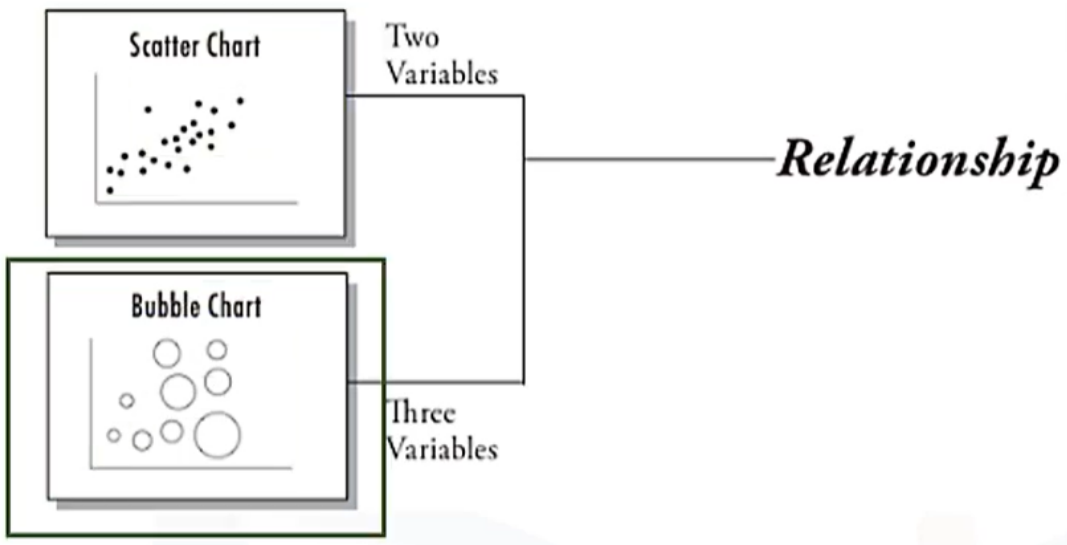


Comparison – among items

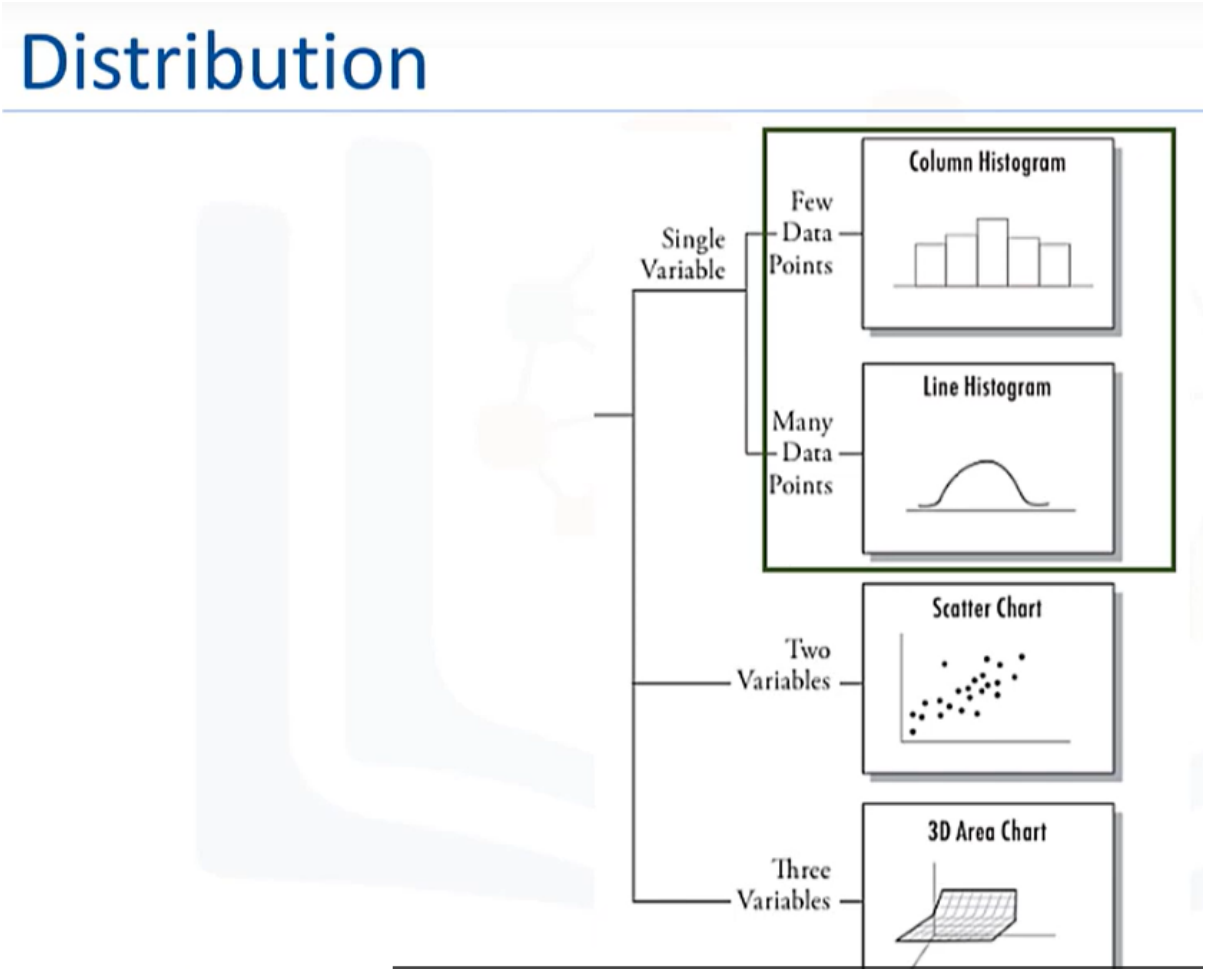


1. bar charts or column charts - comparing items with few categories
2. line chart - comparing behaviors over time, and if you have the time periods running into months, several months, and, then you can use
3. columns and other approaches - if the time periods are not that many

Relationship

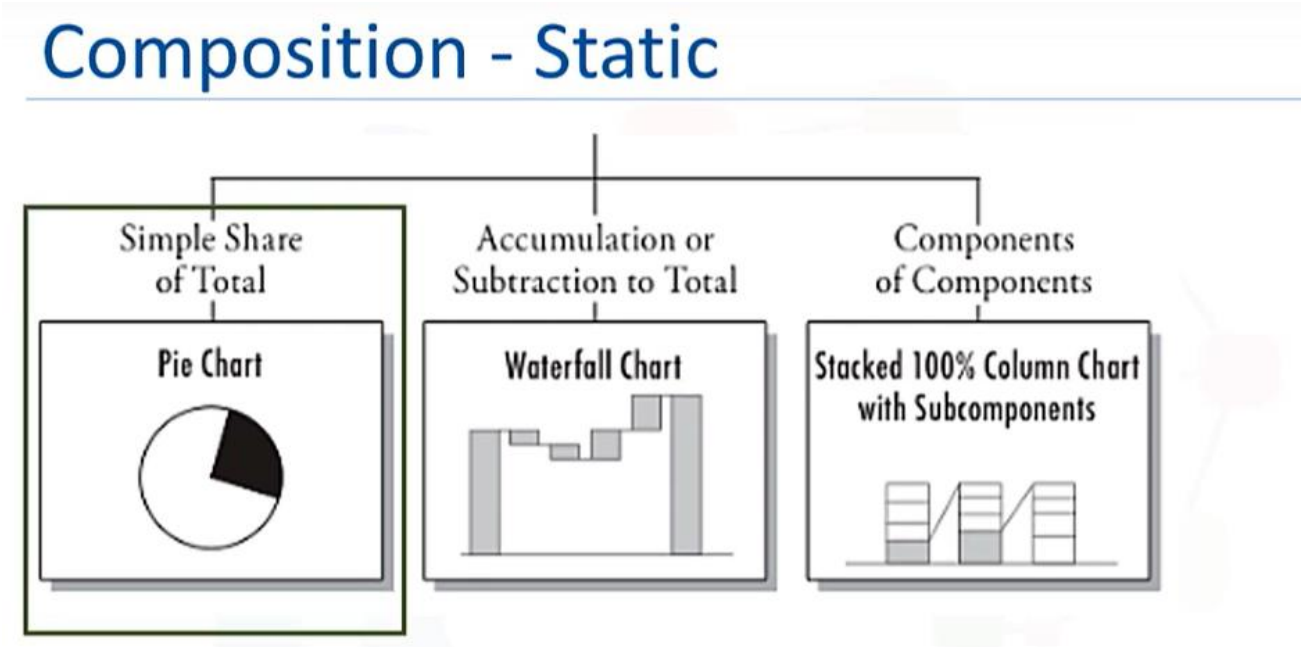


1. scatter plot - depict relationships between two continuous variables
2. bubble chart - depict two variables on x and y-axis the third variable will be depicted by the size of the circle, so you can essentially have 3 variables

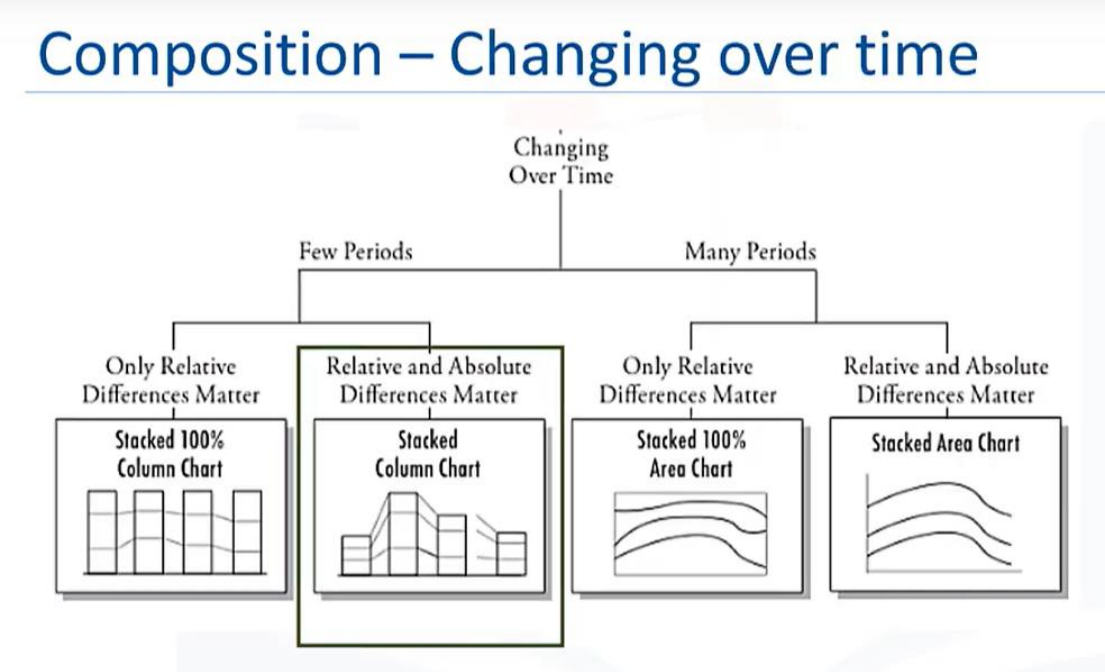


6.

histogram (could be a bar chart type of histogram, or a line histogram/ scatterplots - depicting the distribution of the dataset

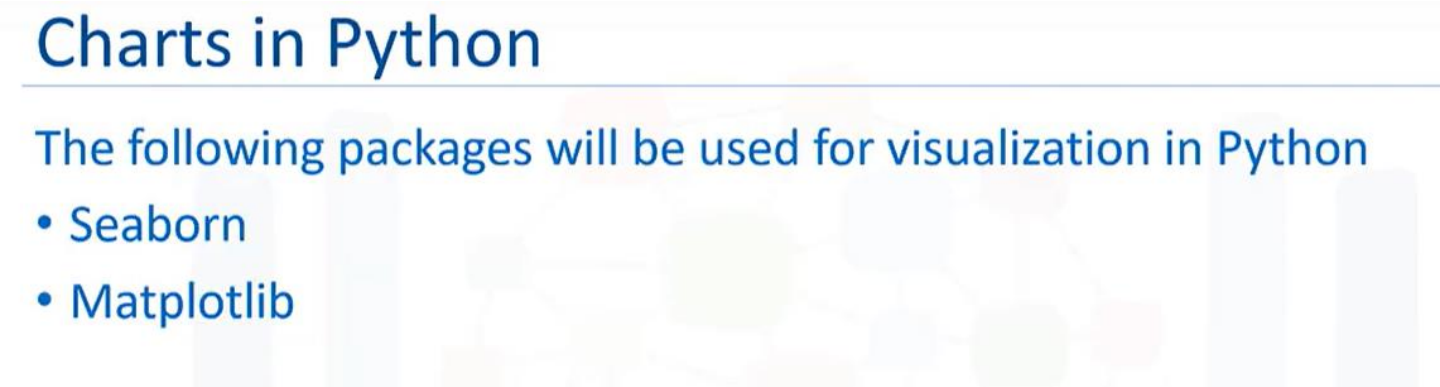


7. pie charts - show the composition, and if it's static data



8. stacked columns - showing the composition that changes over time, for few periods

9. stack area charts - showing the composition that changes over time, for several periods



Statistics by Groups

We have learned to compute averages and standard deviations, but now we will use the same information of same knowledge to make comparisons between groups. So we will use the same dataset that we have used so far, that is, the teaching evaluation data from University of Texas comprising 463 courses.

We are looking at the teaching evaluation beauty and age and we're comparing the averages for these three variables for female instructor variables. So those who are females, their average teaching evaluation was 3.9, compared to those of men, 4.06. Here, we're looking at the average teaching evaluation for tenured professors, 3.96, versus untenured, 4.13.

The average age of untenured professors was 50.2 years, and that for tenured professors, 47.85 years. One thing that is very important in statistical analysis is to think about the question and to think about the population of sample that you are working with. We are computing averages across 463 courses and we find the average age or beauty, but these are the attributes of instructors. We know from our data that there are 94 instructors who have collectively taught 463 courses. And we know that there are duplicates, that is, the same instructor that has, who has, taught multiple courses. So when I compute the average age using 463 courses,

it's not necessarily the average age of the instructors because it could be true that older aged individuals may have taught more courses than younger individuals, resulting in a higher average; that is, not

necessarily the average age of the instructors. So to avoid this problem, we have to subset the data so that we remove the duplicates and have only one observation per individual instructor in the data set. So instead of 463 observations, you should have just 94 observations. Now, let's look at the comparison when we use 94 observations, where no instructor is repeated in the dataset, the average age, or average beauty score, is 0.25. When we look at the 463 courses, the average value is 0.11.

Or, let's compare the age: the average age using 94 observations for males is 49.4 and for females it's 44.9. And, you see here that as for age, we don't see much difference, whether we use 463 observations or 94. But we certainly see much difference in the beauty scores if we were to use the wrong data set, that is, the dataset where individuals are repeated multiple times.

Data visualization is a critical piece of modern day statistical analysis. Their staples are helpful, so you don't have to eyeball the output to figure out what the trends are. The visual displays are much easier to understand. We will use the same datasets of teaching evaluations and ask this question, "Do instructors teaching single credit courses get higher evaluations?" We see that, yes, they do. By mean evaluation, when plotted as a chart, you see that instructors who teach single credit courses have a slightly higher average teaching evaluation.

Let us start by determining how many courses were taught by male instructors and how many by female instructors. For this, we can use a bar chart. Notice that the information is complete from a statistical point of view, in that we know how many courses were taught by males versus females, but we do not have some critical information from this chart as it relates to communication. Therefore, we can say this chart serves a statistical purpose, but it doesn't serve a communication purpose. Let me illustrate this with an example.

Here you are looking at a street map. You can see the streets and the buildings and the highways, but you don't see the street names, and without street names it is hard to determine where you are and in which direction you should be heading. Even though it is according to scale, it may be accurate in its depiction of the streets in the neighborhood, but it still lacks the ability to communicate information to you. To add communication value to this map, you can simply add the street names. Let us apply the same philosophy to our graphic.

But once we add information about this info-graphic, for example, adding just a title makes this chart more informative. To do this in Python, we'll use the count.plot function in the cborn library and set the title label.

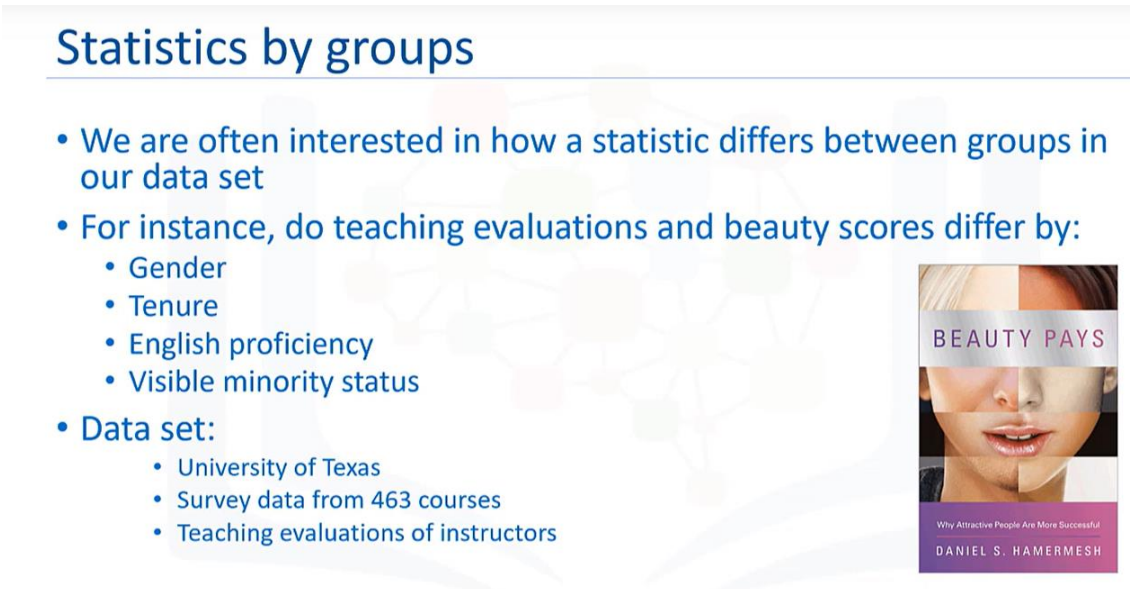
This helps your graph to be more informative. We can also add more dimensions to the data. In addition to the gender of the instructors, we can add the tenure status of the instructors, as well to the graphic. To do that in Python, you add the hue argument to the count plot.

We can add another dimension to the data regenerating the same graphic with the same information, that is, the number of courses taught by gender and tenure, and then adding the dimension of courses: being upper division and lower division, and presenting them in two rows or columns. To do this in Python, we can specify the rows argument using the cat.plot function.

Now let's look at the situation where our primary variables of interest are continuous variables. We would like to explore the relationship between the two, while adding further categorical variables as an additional dimension. Using the teaching evaluation data, we asked this question, "Does age affect teaching evaluations?" We then add two additional dimensions, which are gender and tenure. So our dataset consists

of age and teaching evaluation, which are the two primary variables of interest, and are continuous. And then we add two other dimensions, i.e., gender and tenure. These are categorical variables, age is on the x-axis and the teaching evaluation score is on the y-axis.

The orange-colored circles represent males and the blue-colored circles represent females. The top panel is for tenured professors and the bottom panel is for the untenured instructors. To do this in Python, we use the facet.grid option, which works for multi-plot gridding and allows tweaking the plot. You create the row and hue for the categorical variables. In our case, tenure and gender. And then we use the map to apply a plotting function to each subset of the data.



Summary statistics for groups

Case Summaries

	gender	eval			beauty			age		
		mean	std	var	mean	std	var	mean	std	var
0	female	3.901026	0.538803	0.290308	0.116109	0.81781	0.668813	45.092308	8.532031	72.795559
1	male	4.069030	0.556652	0.309861	-0.084482	0.75713	0.573246	50.746269	9.993396	99.867964

	tenure	eval			beauty			age		
		mean	std	var	mean	std	var	mean	std	var
0	no	4.133333	0.556747	0.309967	0.028359	0.876656	0.768525	50.186275	6.946372	48.252087
1	yes	3.960111	0.549104	0.301516	-0.008013	0.763074	0.582282	47.850416	10.420056	108.577562

To Eliminate Duplicates

Let's think for a second

- Average teaching evaluation is an attribute of the course
  - There are 463 courses
- Average age or beauty is an attribute of the instructor
  - There are 94 instructors in the data set who taught 463 courses
- Eliminate 'duplicates' to avoid repeated measure for the age and beauty variable
- Identify duplicates

```
1 no_duplicates_ratings_df = ratings_df.drop_duplicates(subset =['prof'])
```

Age and Beauty corrected

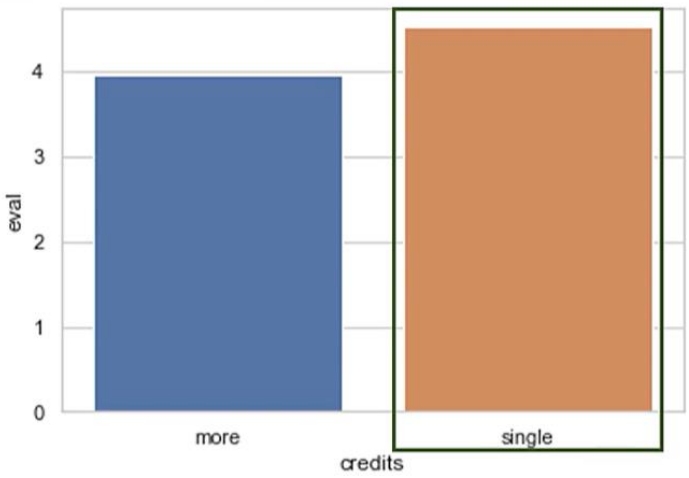
With 94 observations

	gender	beauty			age		
		mean	std	var	mean	std	var
0	female	0.252303	0.843667	0.711774	44.950000	8.935524	79.843590
1	male	-0.033098	0.801559	0.642497	49.481481	10.813585	116.933613

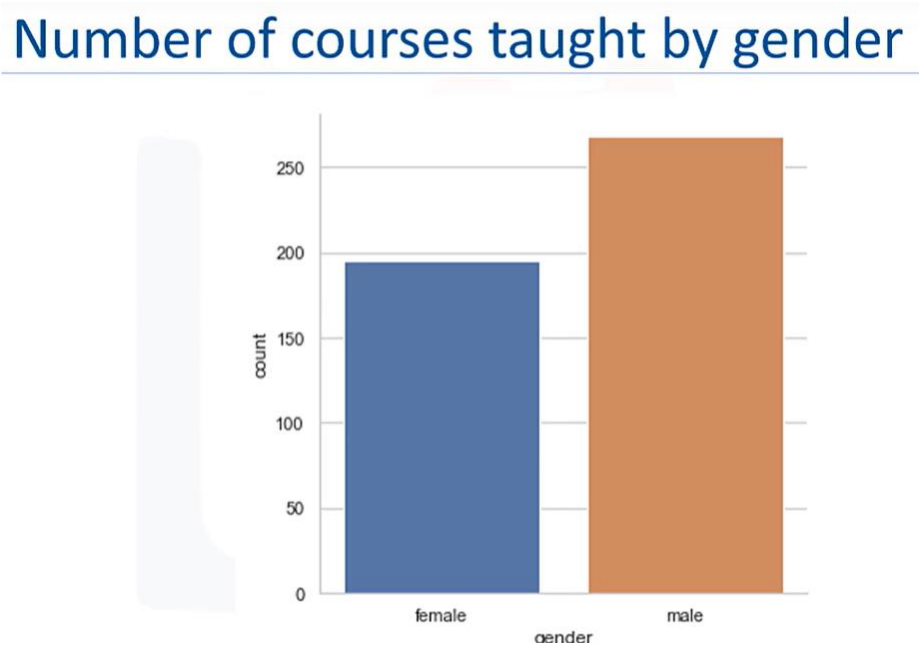
	gender	eval			beauty			age		
		mean	std	var	mean	std	var	mean	std	var
0	female	3.901026	0.538803	0.290308	0.116109	0.81781	0.668813	45.092308	8.532031	72.795559
1	male	4.069030	0.556652	0.309861	-0.084482	0.75713	0.573246	50.746269	9.993396	99.867964

Course Evaluation by number of credits

- Do instructors teaching single credit courses get higher evaluations?

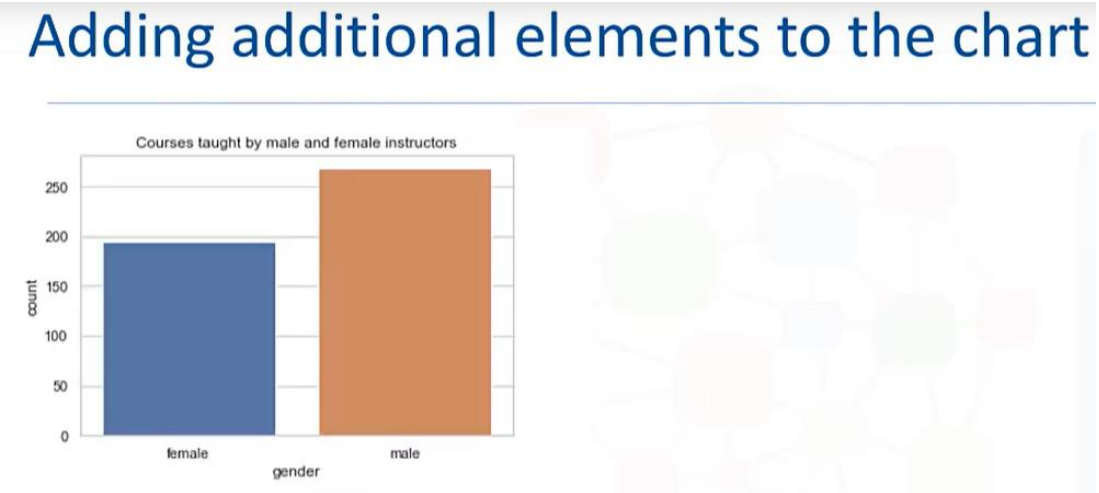


The visual displays are much easier to understand. We will use the same datasets of teaching evaluations and ask this question, "Do instructors teaching single credit courses get higher evaluations?" We see that, yes, they do. By mean evaluation, when plotted as a chart, you see that instructors who teach single credit courses have a slightly higher average teaching evaluation.



Let us start by determining how many courses were taught by male instructors and how many by female instructors. For this, we can use a bar chart. Notice that the information is complete from a statistical point of view, in that we know how many courses were taught by males versus females, but we do not have some critical information from this chart as it relates to communication. Therefore, we can say this chart serves a statistical purpose, but it doesn't serve a communication purpose. Let me illustrate this with an example.

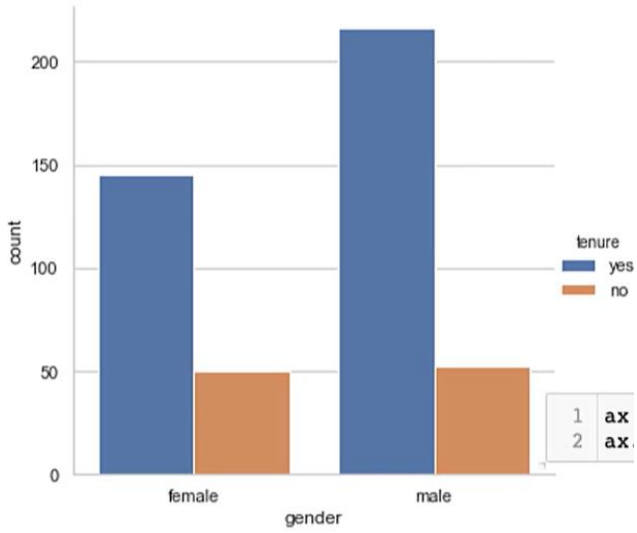
Adding a title




```
1  ax = sns.countplot(x='gender', data=ratings_df)
2  ax.set_title("Courses taught by male and female instructors")
```

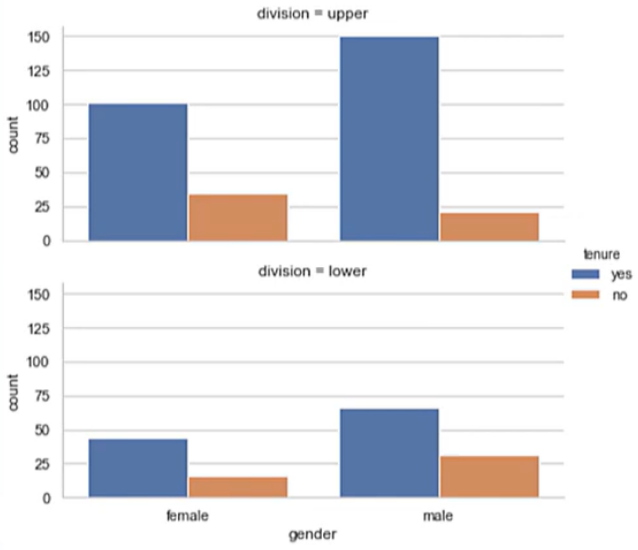
Adding tenure

Courses taught by male and female instructors

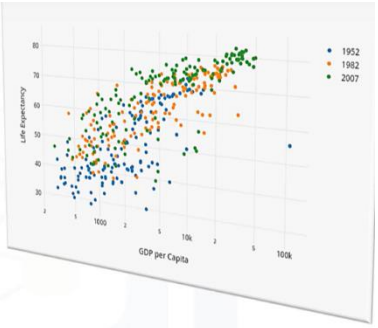



```
1 ax = sns.countplot(x='gender', data=ratings_df, hue = 'tenure')
2 ax.set_title("Courses taught by male and female instructors")
```

Adding tenure + division




```
1 ax = sns.catplot(x='gender', data=ratings_df, kind ='count', hue = 'tenure', row = 'division')
```

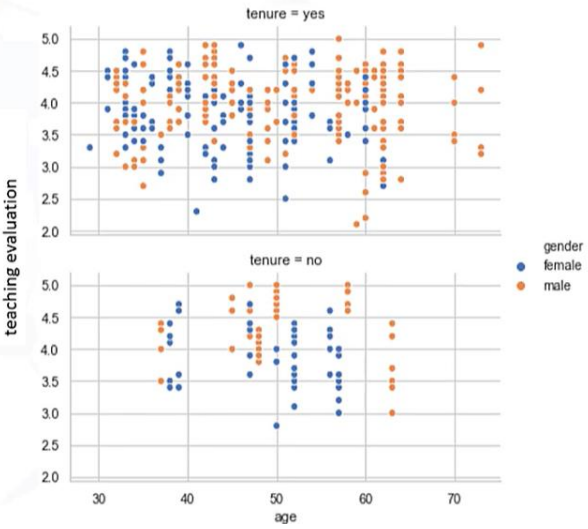


Scattered plots

Continuous variables

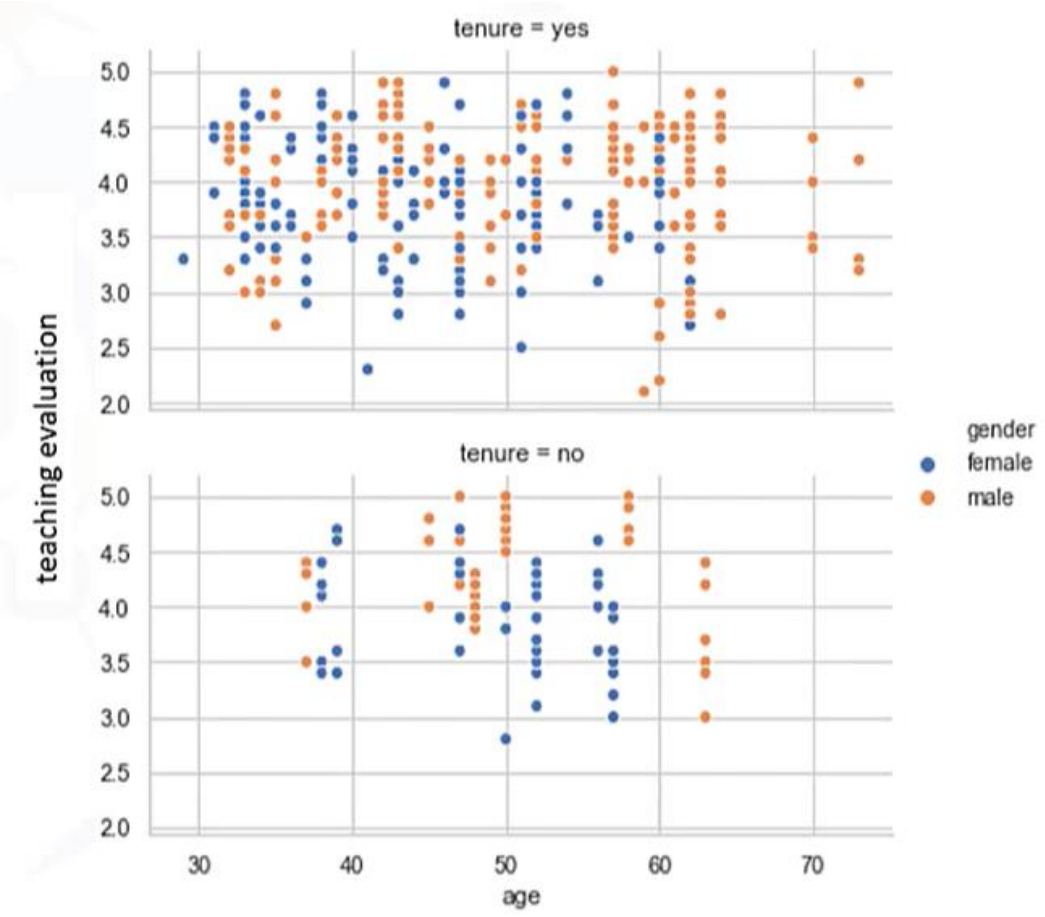
Does age affect teaching evaluations?

- Continuous variables:
  - Age
  - Tenure
- Categorical variables
  - Gender
  - Tenure



"Does age affect teaching evaluations?" We then add two additional dimensions, which are gender and tenure. So our dataset consists of age and teaching evaluation, which are the two primary variables of interest, and are continuous. And then we add two other dimensions, i.e., gender and tenure. These are categorical variables, age is on the x-axis an d the teaching evaluation score is on the y-axis. The orange-colored circles represent males and the blue-colored circles represent females. The top panel is for tenured professors and the bottom panel is for the untenured instructors.

Does age affect teaching evaluations?



```
g = sns.FacetGrid(ratings_df, row="tenure",  hue="gender")
g = (g.map(plt.scatter, "age", "eval")
       .add_legend())
```

Statistical Charts

So far, we have displayed data as averages and counts. Now let's look at some other statistical parameters that we will illustrate as graphics. We have not yet shown anything about variance and I think the first thing that one should look into beyond averages is to look at variance or how the data are distributed, and a good way of looking at the distribution of data, especially if it were to be a continuous variable, is to look at histograms. And if you are interested in displaying something more than the average, maybe the median, and the quartiles, then perhaps box plots should be our choice.

Using the teaching evaluation data, we have plotted a histogram of teaching evaluation scores. You could see that the mean score is around 4, but then you could see very low teaching evaluation scores, not many frequently, but most frequently it's the around the average, and then you'll see that some have lower teaching evaluation scores, and some have fairly high teaching evaluation scores.

The histogram approximates the normal distribution curve. Essentially, you have 3.99 or 4 as the mean. The standard deviation of 0.55, looking at 463 records. This gives you a good idea of how your data are distributed. You can, in fact, plot multiple histograms, such that you can see the difference between the subgroups. So here, you have the histograms overlaid for males and females. These frequent lower teaching evaluations, for females, is likely to influence the average teaching evaluation score for females versus for males.

A box plot essentially looks like this. The thick line in the box represents the median. The top part of the box is the third quartile. The bottom part of the box is the first quartile. The line at the bottom is the minimum value, and, the line at the top is the maximum value. And the range between the first quartile and the third quartile is called the interquartile range.

In this graphic, we have created the box plots for the age variable. We can see that the median age of males is higher than the median age of females. Also, the maximum age of the males is higher than the maximum

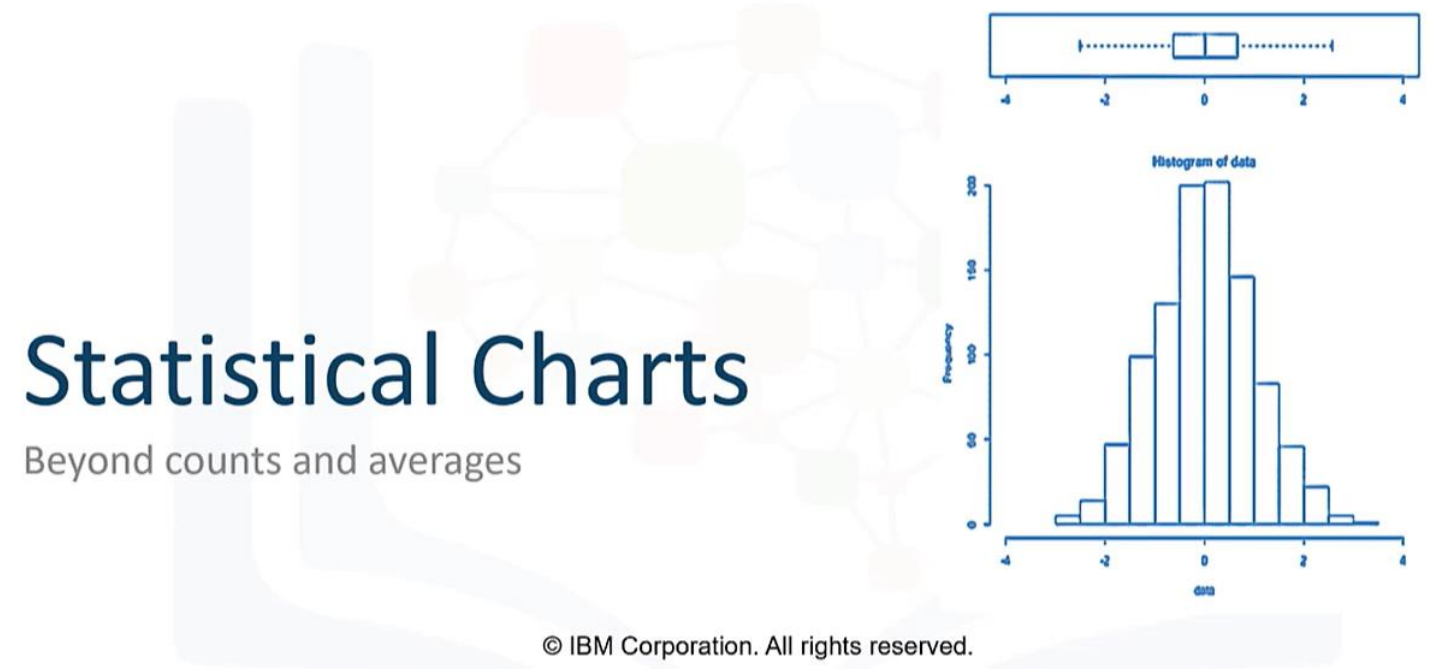
age of females. To do this in Python, we use the box plot function in the seaborne library. We will put the gender on the y-axis and the age of the instructor on the x-axis. You can play around with the x and y-axis.

If you wanted a horizontal style box plot, for readability, I like to use vertical box plots. We can also add another dimension. Here we will add tenure: so those who are tenured are plotted on the right and those who are not tenured are plotted on the left; and the blue color represents the female instructors; and the orange color represents the male instructors.

We can see the differences between male and female. Instructors, male tenured instructors, are older than male untenured instructors; whereas female tenured instructors are younger than female untenured instructors.

To do this in Python, add the hue argument to the box plot function. A pie chart is another way of looking at your data. You can see here in this graphic that the number of courses taught by male instructors is larger than the number of courses taught by female instructors. To do this in Python, we will use the matplotlib library.

First we specify the labels, get the number of courses taught both by male and females, and assign it to a sizes variable. Create a subplot, insert the sizes, labels, and percentage to one decimal place in the pi function, and print out the pie chart with the show function.

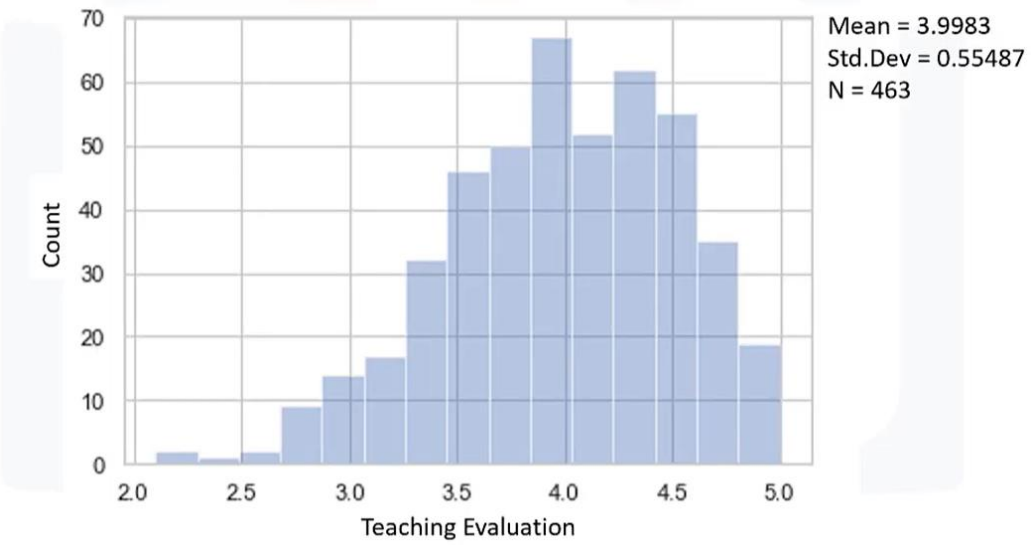


Better than the average

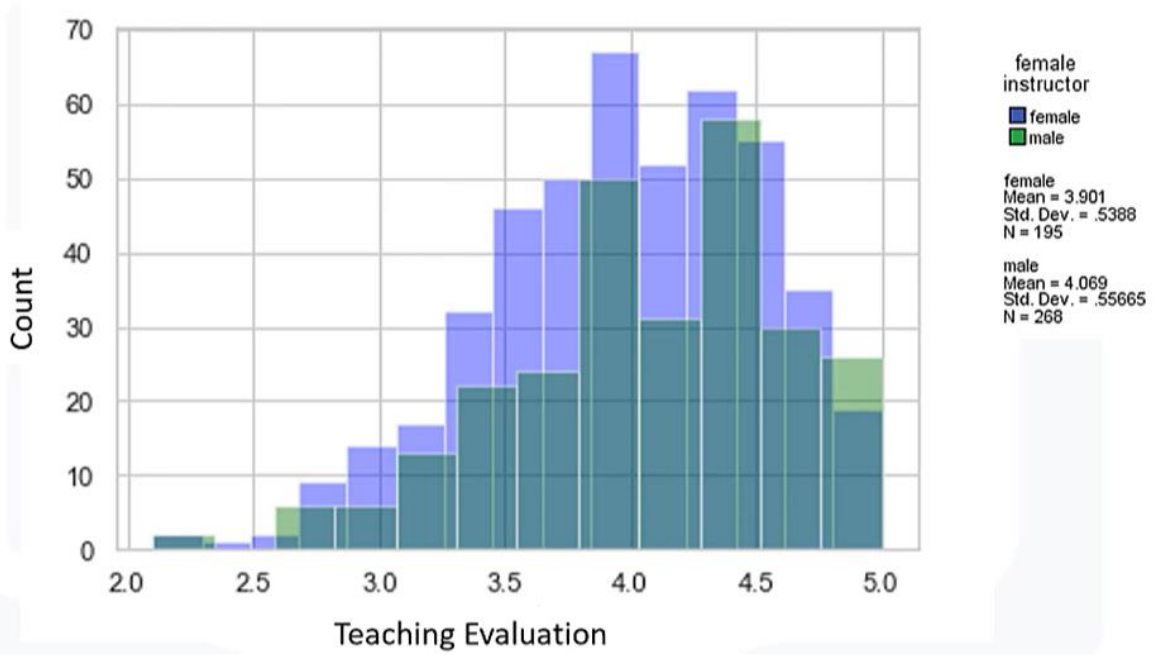
Statistical parameters illustrated in charts:

- Distribution and variance: Histogram
- Box plot: Displaying mean, median quartile and outliers

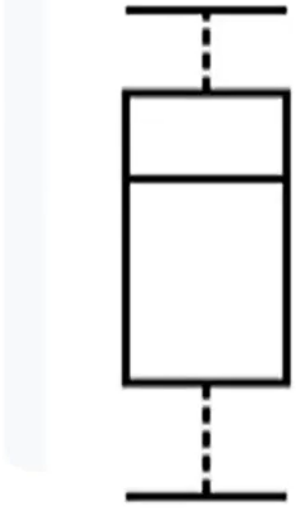
Simple Histogram – teaching evaluation score

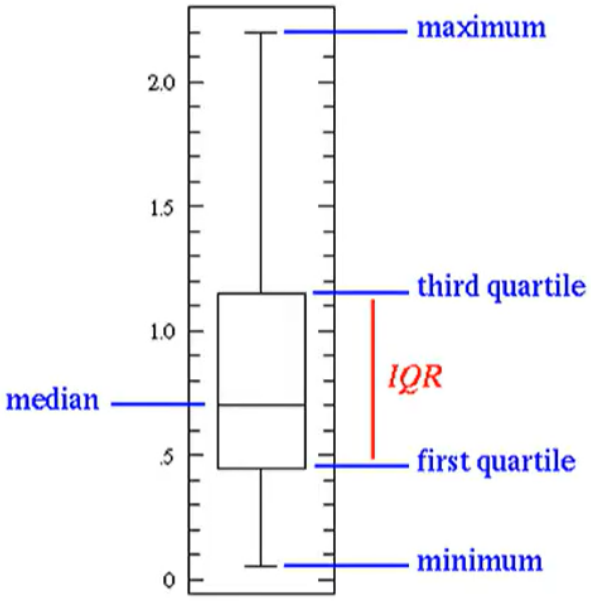


Evaluation by gender-Histogram



Box plots

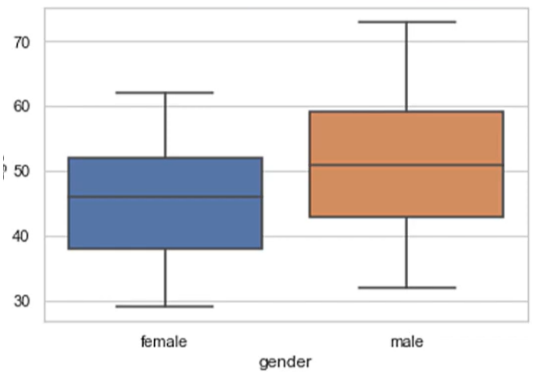


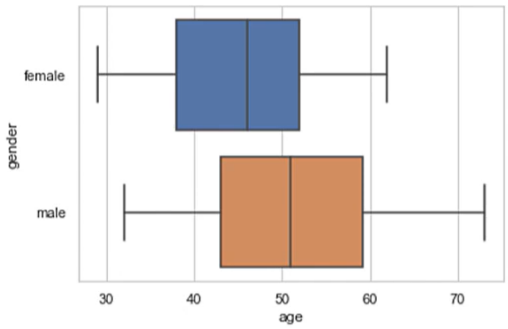


For normally distributed data:
*IQR = 1.35 * standard deviation*

A box plot essentially looks like this. The thick line in the box represents the median. The top part of the box is the third quartile. The bottom part of the box is the first quartile. The line at the bottom is the minimum value, and, the line at the top is the maximum value. And the range between the first quartile and the third quartile is called the interquartile range.

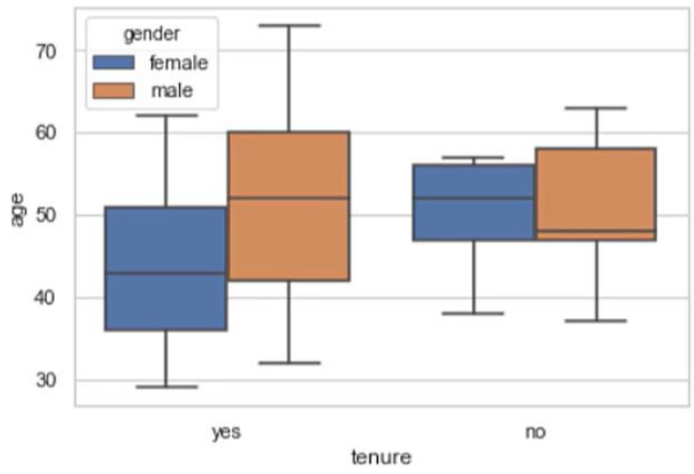
Box plots – age of the instructor by gender






```
1  ax = sns.boxplot(x="gender", y="age", data=ratings_df)
```

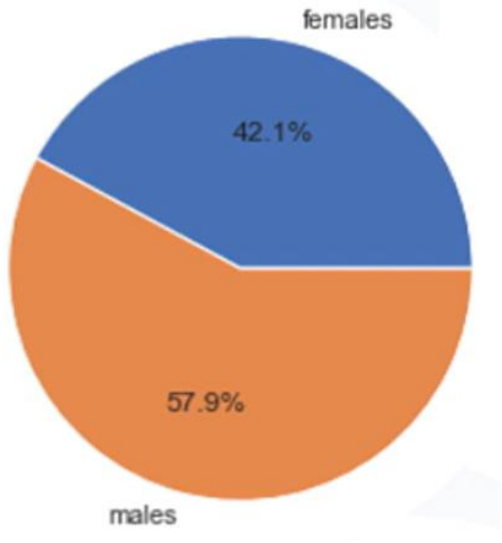
Comparing age along tenure and gender



Here we will add tenure: so those who are tenured are plotted on the right and those who are not tenured are plotted on the left; and the blue color represents the female instructors; and the orange color represents the male instructors. We can see the differences between male and female. Instructors, male tenured instructors, are older than male untenured instructors; whereas female tenured instructors are younger than female untenured instructors.

```
1  ax = sns.boxplot(x="tenure", y="age", hue="gender",
2                   data=ratings_df)
```

Pie in the sky




```
1  import matplotlib.pyplot as plt
```

A pie chart is another way of looking at your data. You can see here in this graphic that the number of courses taught by male instructors is larger than the number of courses taught by female instructors.

To do this in Python, we will use the matplotlib library. First we specify the labels, get the number of courses taught both by male and females, and assign it to a sizes variable. Create a subplot, insert the sizes, labels, and percentage to one decimal place in the pi function, and print out the pie chart with the show function.

Introducing Teaching Ratings Data

In this video, I will introduce the teaching ratings data. We have been working with the teaching ratings data from University of Texas, and the underlying question is, "If students teaching evaluations are influenced by the looks of individual instructors?" Or, you can ask, "If their teaching evaluations differ by gender?" Or, "If good-looking instructors get higher teaching evaluations?" I obtained this data from Professor Daniel Hamermesh who has written a paper about how beauty may impact an instructor's teaching evaluation.

In fact, he has written an amazing book called, "Beauty Pays," in which he answers these questions, such as, "Do do you think good looking employees get higher pay or faster promotions? Do you think good looking instructors get higher teaching evaluations?" The data comes from University of Texas. It's a survey and data from, obtained from, 463 courses.

The data is first referenced in the book, "Getting Started with Data Science: Making sense of data with analytics." In Chapter 4, there are variables that essentially define the attributes of instructors, and characteristics of the courses. Some variables are continuous, others are dichotomous or categorical variables. So, the primary two variables of our interest are "beauty score," which is basically the physical appearance of an instructor, which was ranked by a panel of 6 students and I think that they have normalized the beauty score, such that it had a mean of 0 and variance of 1 or a standard deviation of 1. It's the same as z transformation. The dependent variable, the variable of interest, is evaluation, which is basically the teaching evaluation ranging between a scale of 1 to 5, 1 being very unsatisfactory, and if the student found the course to be excellent, then it's 5. And then there are other dichotomous or binary or categorical variables, such as "minority," if the instructor was a non-Caucasian. "Age" is a continuous variable into the professor's age or instructor's age. "Gender," being male or female; "native" stands for native English speaker; if the instructor was a native speaker of English language, 1, 0 otherwise. If the professor was tenured, 1, 0 otherwise. I have produced some descriptive statistics for your reference.

For instance, for the continuous variables, such as age and beauty, and teaching evaluation, and the number of students who were enrolled in the course, or the number of students who performed the teaching evaluations, I produced the descriptive statistics, such as the minimum, the maximum, mean, and standard deviation. For categorical variables, such as gender, female "yes" or "no," visible minority, "yes" or "no," person being a tenured professor or otherwise. I produced the frequency distributions and percentage of individuals falling in one category or otherwise. Notice that the teaching evaluation score is 3.99, with a standard deviation of 0.55. And, let's see if I were to produce a histogram of this variable, teaching evaluation, how will it look like with raw data; and if I were to use the normal distribution and feed the two parameters, that is, the mean and the standard deviation, how will the same distribution look like, using a normal distribution?

I am presenting here the distribution of the raw data on left side and the presentation of the same data with the same parameters, the mean, and standard deviation using normal distribution. You could see that the data are not exactly following a bell curve; this is the raw data, and it seldom does, but then the theoretical distribution looks like this. Essentially, the same dataset with a mean of 3.998 and standard deviation is

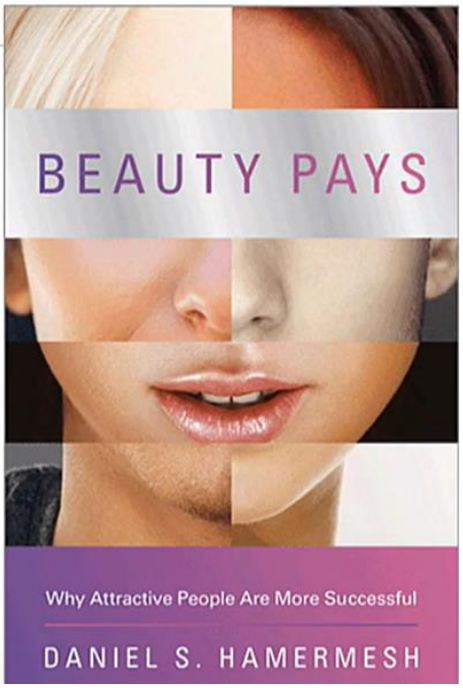
presented here; and then a normal distribution drawn from these two parameters appears here. So it's much smoother; theoretical distributions are much smoother than the raw data.

Introducing Teaching Ratings Data

Does instructor evaluation score differ by gender?

Beauty in numbers

- Does beauty pay?
- Do you think good-looking employees get higher pay or more promotions?
- Do you think good looking professors get better teaching evaluations?
- Let's look at data from the University of Texas
  - Survey data from 463 courses



Meta data: The data about data

- Details are available in Chapter 4 of *Getting Started with Data Science*
- Variable type:
  - Real number/Integer
  - Continuous/ categorical variables
  - Dichotomous variable

Variable	Description
beauty	Rating of the instructor's physical appearance by a panel of six students, averaged across the six panelists, transformed to have a mean of zero.
eval	Course overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent).
minority	Factor variable. Does the instructor belong to a minority (non-Caucasian)?
age	The professor's age.
gender	Factor indicating instructor's gender (male/female).
native	Factor variable. Is the instructor a native English speaker?
tenure	Factor variable. Is the instructor on tenure track?
credits	Factor variable. Is the course a single-credit elective (for example, yoga, aerobics, dance)?
division	Factor variable. Is the course an upper or lower division course? (Lower division courses are mainly large freshman and sophomore courses.)
students	Number of students who participated in the evaluation.
allstudents	Number of students enrolled in the course.
prof	Factor variable indicating instructor identifier.

Descriptive Statistics

	age	beauty	eval	students	allstudents
count	463.000000	4.630000e+02	463.000000	463.000000	463.000000
mean	48.365011	6.271140e-08	3.998272	36.624190	55.177106
std	9.802742	7.886477e-01	0.554866	45.018481	75.072800
min	29.000000	-1.450494e+00	2.100000	5.000000	8.000000
25%	42.000000	-6.562689e-01	3.600000	15.000000	19.000000
50%	48.000000	-6.801430e-02	4.000000	23.000000	29.000000
75%	57.000000	5.456024e-01	4.400000	40.000000	60.000000
max	73.000000	1.970023e+00	5.000000	380.000000	581.000000

native

		count
0	no	28
1	yes	435

minority

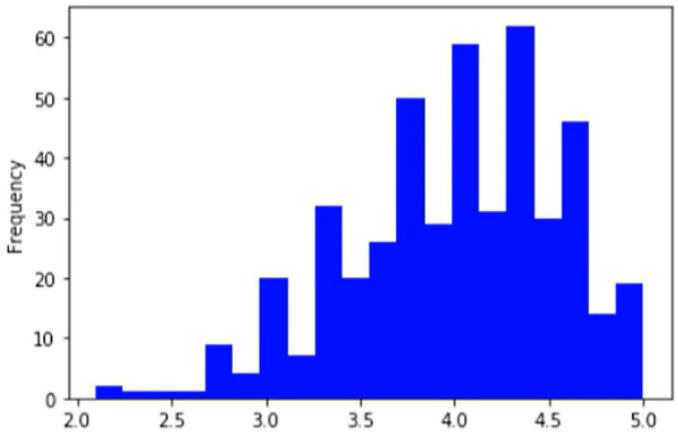
		count
0	no	399
1	yes	64

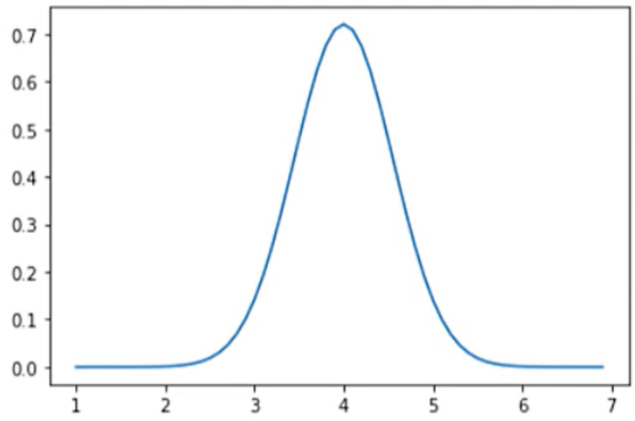
gender

		count
0	female	195
1	male	268

tenure

		count
0	no	102
1	yes	361





Teaching evaluation score with normal distribution curve

I were to produce a histogram of this variable, teaching evaluation, how will it look like with raw data; and if I were to use the normal distribution and feed the two parameters, that is, the mean and the standard deviation, how will the same distribution look like, using a normal distribution? I am presenting here the distribution of the raw data on left side and the presentation of the same data with the same parameters, the mean, and standard deviation using normal distribution. You could see that the data are not exactly following a bell curve; this is the raw data, and it seldom does, but then the theoretical distribution looks like this. Essentially, the same dataset with a mean of 3.998 and standard deviation is presented here; and then a normal distribution drawn from these two parameters appears here. So it's much smoother; theoretical distributions are much smoother than the raw data.

Data Visualization

Estimated time needed: **30** minutes

In this lab, you will learn how to visualize and interpret data

Objectives

- Import Libraries
- Lab Exercises
    - Identifying duplicates
    - Plotting Scatterplots
    - Plotting Boxplots

Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[1]:
```
#install specific version of libraries used in lab
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install scipy=1.7.1-y
#!  mamba install seaborn=0.9.0-y
#!  mamba install matplotlib=3.4.3-y
```

Import the libraries we need for the lab

[2]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Read in the csv file from the url using the request library

[4]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.csv'
ratings_df = pd.read_csv(ratings_url)
```

Lab Exercises

Identify all duplicate cases using prof. Using all observations, find the average and standard deviation for age. Repeat the analysis by first filtering the data set to include one observation for each instructor with a total number of observations restricted to 94.

Identify all duplicate cases using prof variable - find the unique values of the prof variables

```
[5]: ratings_df.prof.unique()
```

```
[5]: array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
        18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36,
        37, 38, 39, 41, 42, 43, 44, 45, 46, 48, 49, 50, 51, 52, 53, 54, 55,
        56, 57, 58, 59, 60, 63, 64, 65, 66, 67, 68, 70, 71, 72, 73, 74, 75,
        76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92,
        93, 94, 22, 30, 40, 47, 61, 62, 69])
```

Print out the number of unique values in the prof variable

```
[6]: ratings_df.prof.nunique()
```

```
[6]: 94
```

Using all observations, Find the average and standard deviation for age

```
[7]: ratings_df['age'].mean()
```

```
[7]: 48.365010799136066
```

```
[8]: ratings_df['age'].std()
```

```
[8]: 9.802742037864821
```

Repeat the analysis by first filtering the data set to include one observation for each instructor with a total number of observations restricted to 94.

first we drop duplicates using prof as a subset and assign it a new dataframe name called no_duplicates_ratings_df

```
[9]: no_duplicates_ratings_df = ratings_df.drop_duplicates(subset =['prof'])
     no_duplicates_ratings_df.head()
```

[9]:

	minority	age	gender	credits	beauty	eval	division	native	tenure	students	allstudents
0	yes	36	female	more	0.289916	4.3	upper	yes	yes	24	43
4	no	59	male	more	-0.737732	4.5	upper	yes	yes	17	20
7	no	51	male	more	-0.571984	3.7	upper	yes	yes	55	55
9	no	40	female	more	-0.677963	4.3	upper	yes	yes	40	46
17	no	31	female	more	1.509794	4.4	upper	yes	yes	42	48

prof	PrimaryLast	vismin	female	single_credit	upper_division	English_speaker	tenured_prof
1	0	1	1	0	1	1	1
2	0	0	0	0	1	1	1
3	0	0	0	0	1	1	1
4	0	0	1	0	1	1	1
5	0	0	1	0	1	1	1

> Use the new dataset to get the mean of age

```
[10]: no_duplicates_ratings_df['age'].mean()
```

```
[10]: 47.5531914893617
```

```
[11]: no_duplicates_ratings_df['age'].std()
```

```
[11]: 10.25651329515495
```

Using a bar chart, demonstrate if instructors teaching lower-division courses receive higher average teaching evaluations

```
[12]: ratings_df.head()
```

[12]:

	minority	age	gender	credits	beauty	eval	division	native	tenure	students	allstudents	prof
0	yes	36	female	more	0.289916	4.3	upper	yes	yes	24	43	1
1	yes	36	female	more	0.289916	3.7	upper	yes	yes	86	125	1
2	yes	36	female	more	0.289916	3.6	upper	yes	yes	76	125	1
3	yes	36	female	more	0.289916	4.4	upper	yes	yes	77	123	1
4	no	59	male	more	-0.737732	4.5	upper	yes	yes	17	20	2

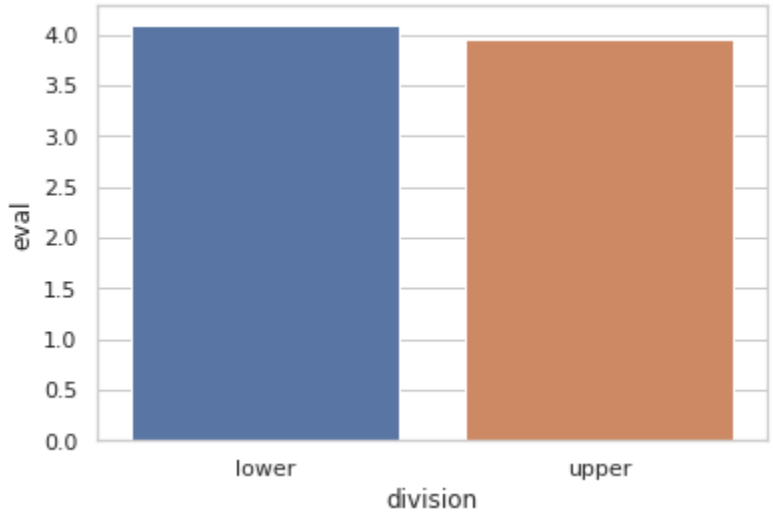
PrimaryLast	vismin	female	single_credit	upper_division	English_speaker	tenured_prof
0	1	1	0	1	1	1
0	1	1	0	1	1	1
0	1	1	0	1	1	1
1	1	1	0	1	1	1
0	0	0	0	1	1	1

Find the average teaching evaluation in both groups of upper and lower-division

[13]:
```
division_eval = ratings_df.groupby('division')[['eval']].mean().reset_index()
```

Plot the barplot using the seaborn library

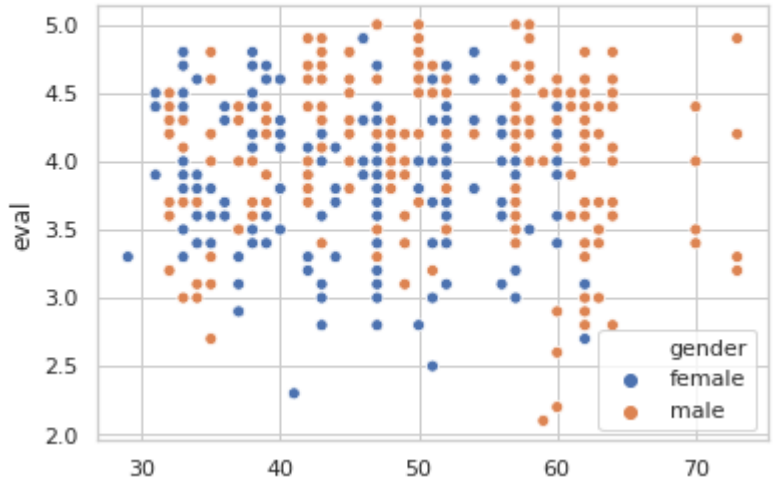
[14]:
```
sns.set(style="whitegrid")
ax = sns.barplot(x="division", y="eval", data=division_eval)
```



Using gender-differentiated scatter plots, plot the relationship between age and teaching evaluation scores.

Create a scatterplot with the scatterplot function in the seaborn library this time add the hue argument

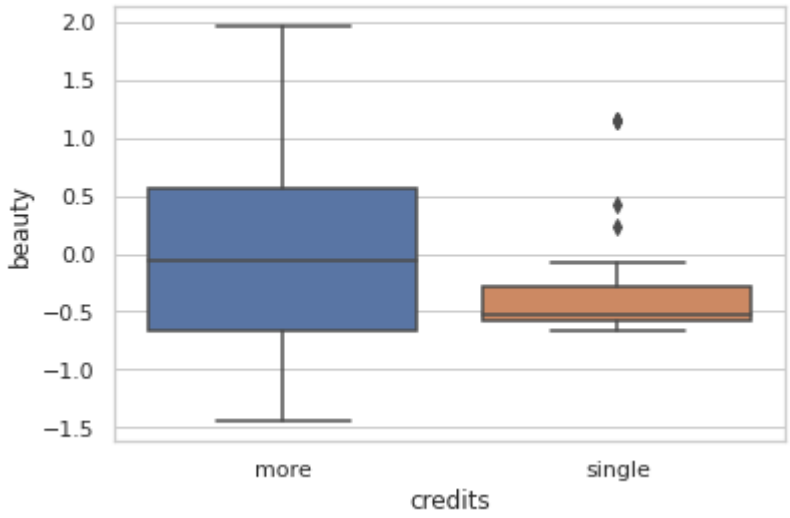
```
[16]: ax = sns.scatterplot(x='age', y='eval', hue='gender',
                           data=ratings_df)
```



Create a box plot for beauty scores differentiated by credits.

We use the boxplot() function from the seaborn library

```
[17]: ax = sns.boxplot(x='credits', y='beauty', data=ratings_df)
```

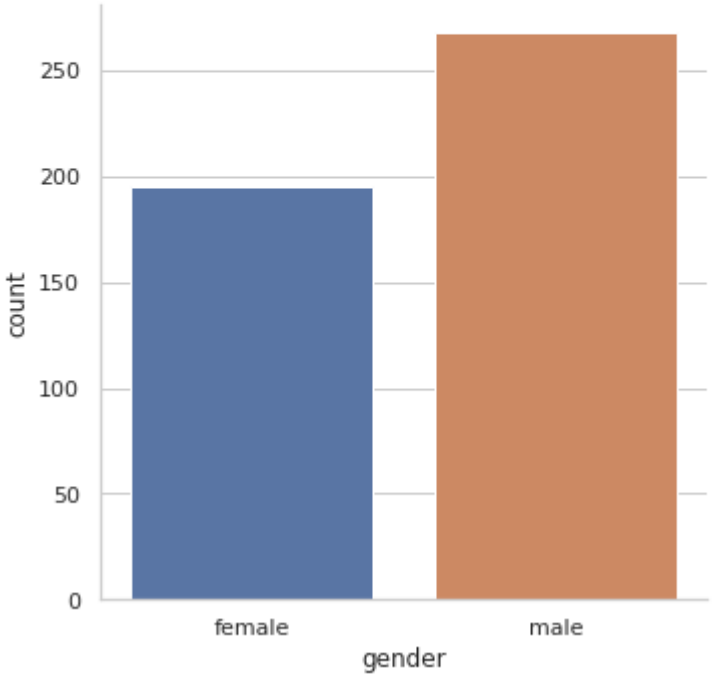


What is the number of courses taught by gender?

We use the `catplot()` function from the seaborn library

[18]:
```
sns.catplot(x='gender', kind='count', data=ratings_df)
```

[18]:
```
<seaborn.axisgrid.FacetGrid at 0x7f8c6c919a90>
```

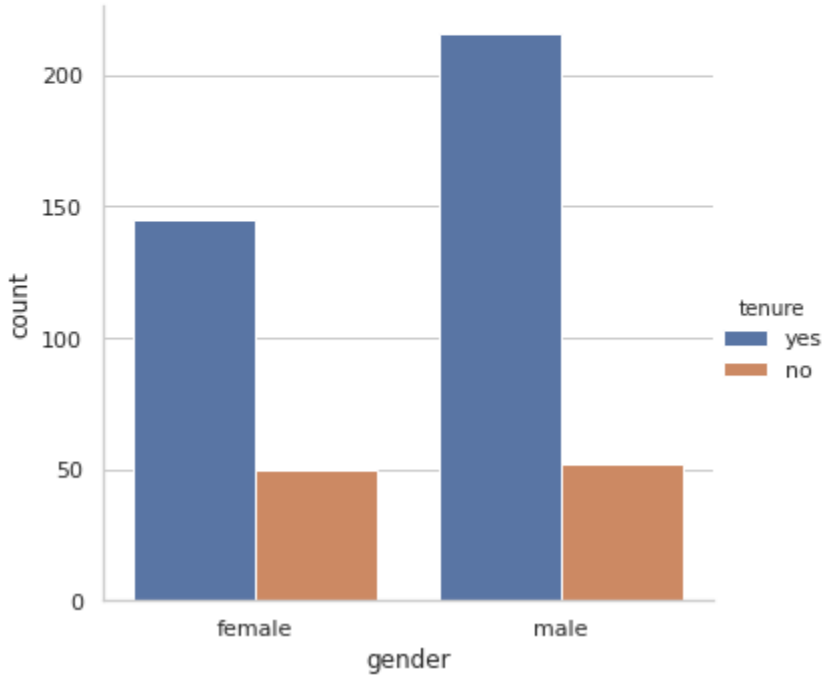


Create a group histogram of taught by gender and tenure

We will add the `hue = Tenure` argument

[19]:
```
sns.catplot(x='gender', hue = 'tenure', kind='count', data=ratings_df)
```

[19]:
```
<seaborn.axisgrid.FacetGrid at 0x7f8c67fca5d0>
```

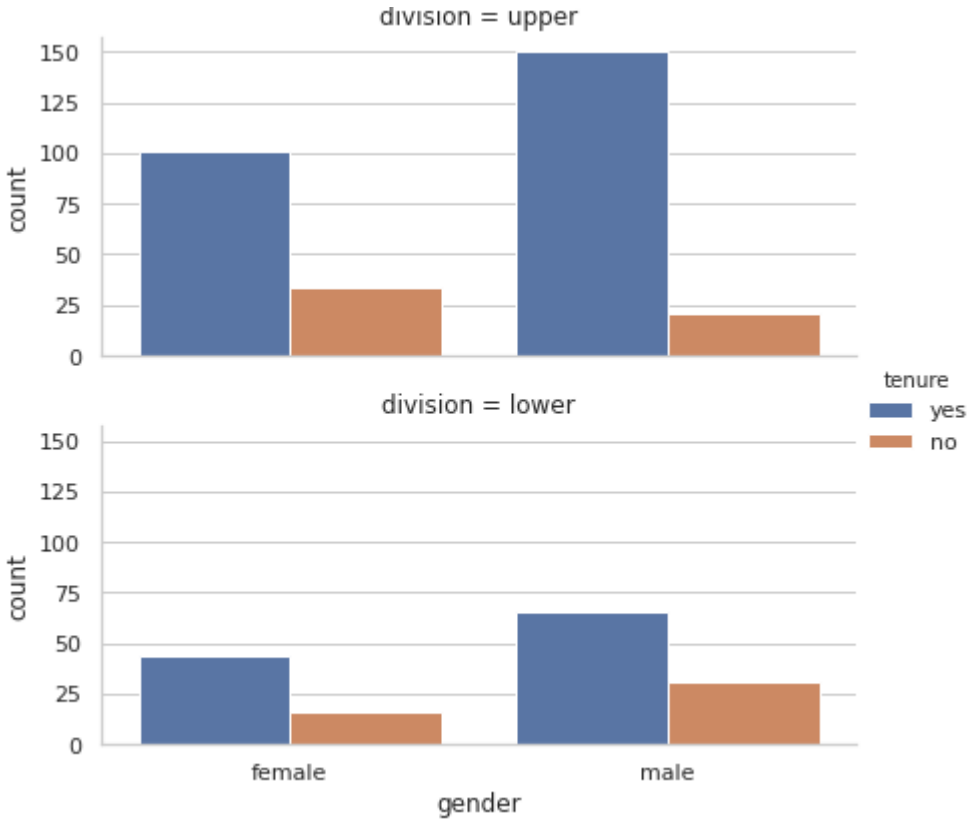


Add division as another factor to the above histogram

We add another argument named `row` and use the division variable as the row

[20]:
```
sns.catplot(x='gender', hue = 'tenure', row = 'division',
            kind='count', data=ratings_df,
            height = 3, aspect = 2)
```

[20]:
```
<seaborn.axisgrid.FacetGrid at 0x7f8c67d7e210>
```

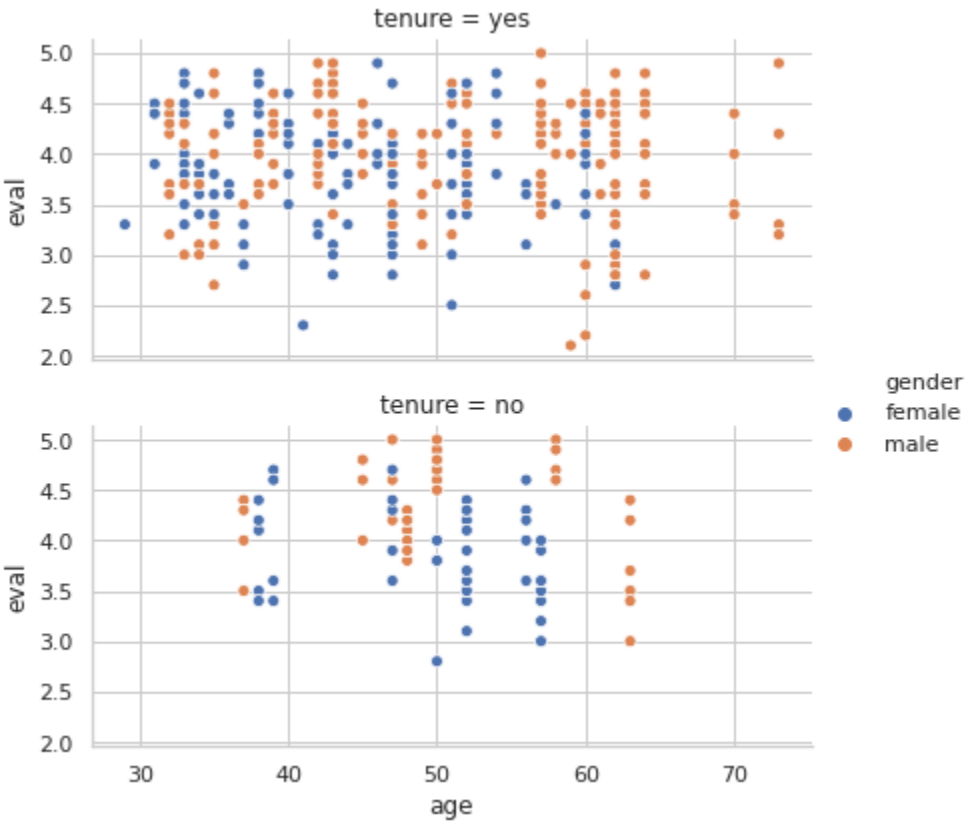


Create a scatterplot of age and evaluation scores, differentiated by gender and tenure

Use the `relplot()` function for complex scatter plots

[21]:
```
sns.relplot(x="age", y="eval", hue="gender",
            row="tenure",
            data=ratings_df, height = 3, aspect = 2)
```

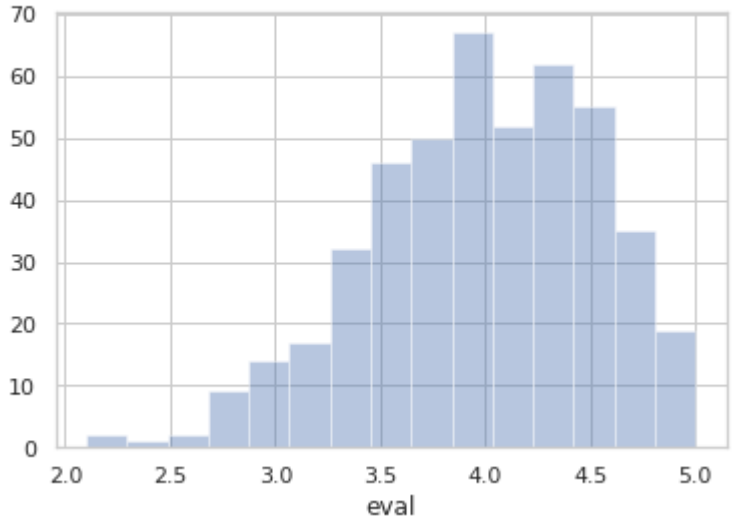
[21]:
```
<seaborn.axisgrid.FacetGrid at 0x7f8c67b95690>
```



Create a distribution plot of teaching evaluation scores

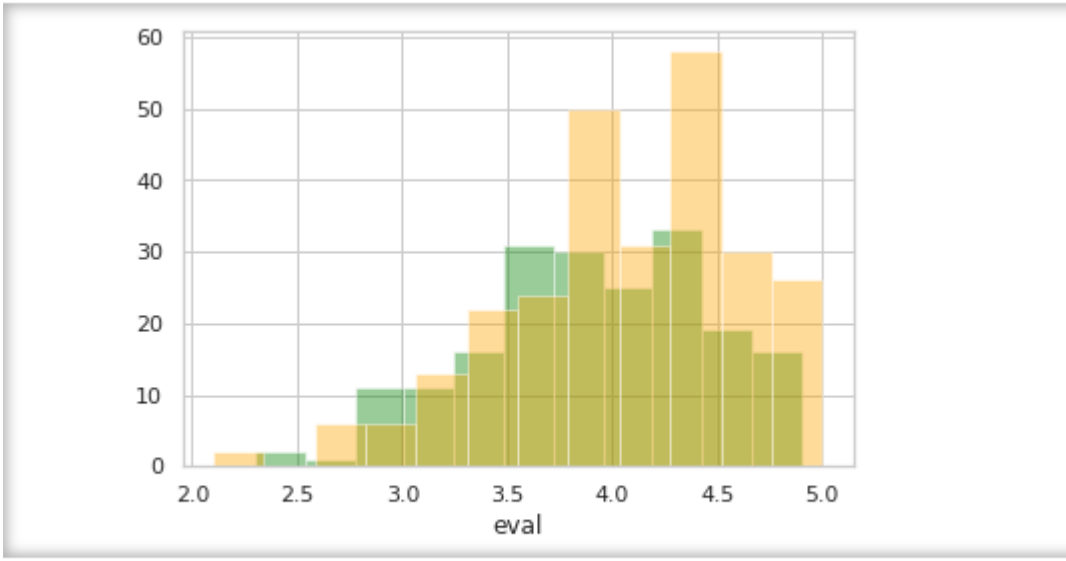
We use the distplot() function from the seaborn library, set kde = false because we don'e need the curve

[22]:
```
ax = sns.distplot(ratings_df['eval'], kde = False)
```



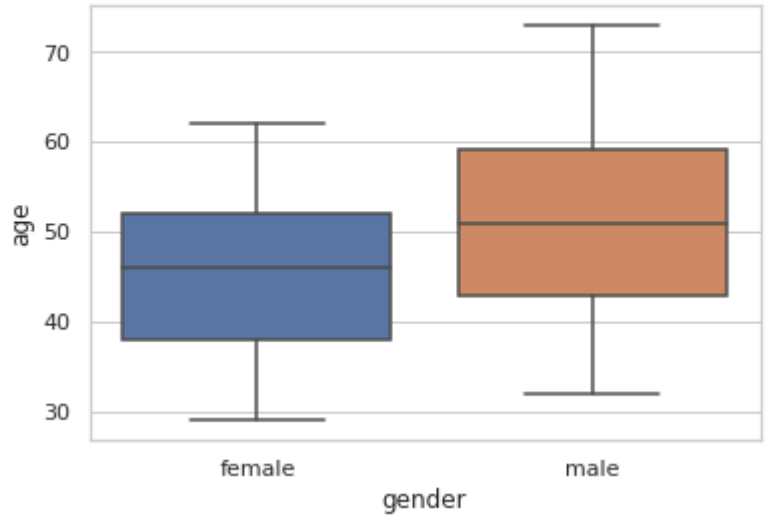
Create a distribution plot of teaching evaluation score with gender as a factor

[23]:
```
## use the distplot function from the seaborn library
sns.distplot(ratings_df[ratings_df['gender'] == 'female']['eval'], color='green', kde=False)
sns.distplot(ratings_df[ratings_df['gender'] == 'male']['eval'], color="orange", kde=False)
plt.show()
```



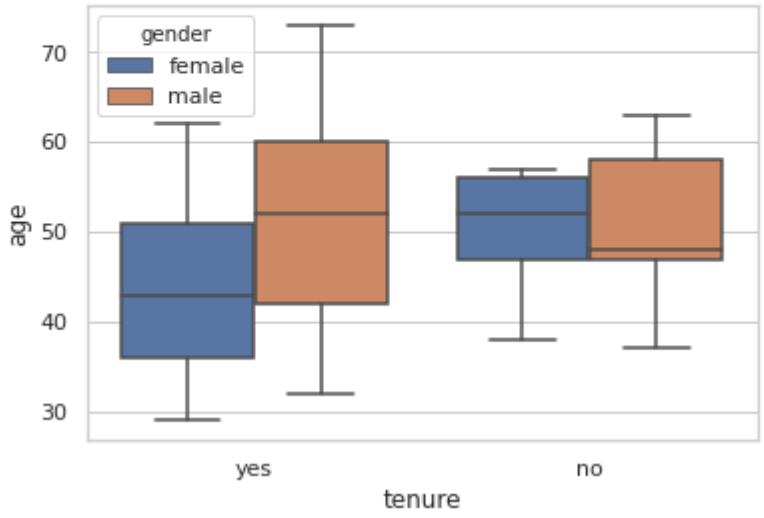
Create a box plot - age of the instructor by gender

[24]:
```
ax = sns.boxplot(x="gender", y="age", data=ratings_df)
```



Compare age along with tenure and gender

[25]:
```
ax = sns.boxplot(x="tenure", y="age", hue="gender",
                 data=ratings_df)
```

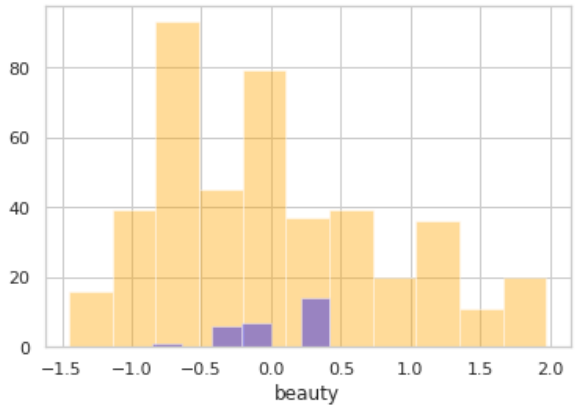


Practice Questions

Question 1: Create a distribution plot of beauty scores with Native English speaker as a factor

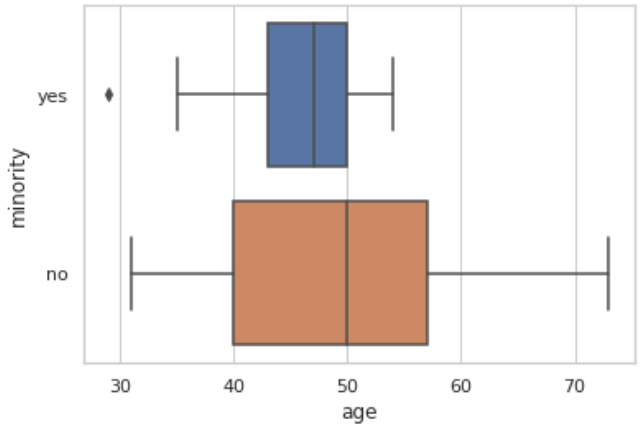
- Make the color of the native English speakers plot - orange and non - native English speakers - blue

[26]:
```
## insert code
sns.distplot(ratings_df[ratings_df['native'] == 'yes']['beauty'], color="orange", kde=False)
sns.distplot(ratings_df[ratings_df['native'] == 'no']['beauty'], color="blue", kde=False)
plt.show()
```



Question 2: Create a Horizontal box plot of the age of the instructors by visible minority

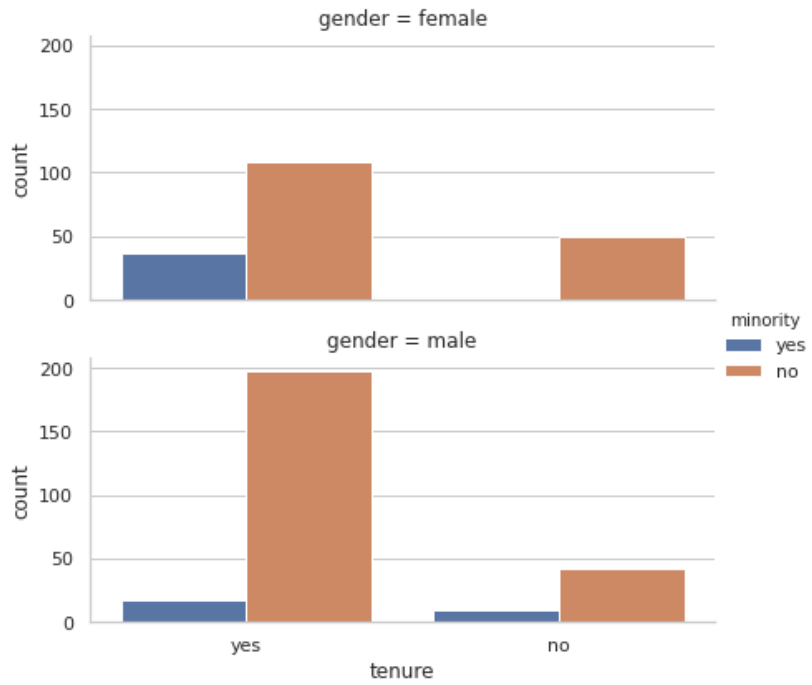
[27]:
```
## insert code
ax = sns.boxplot(x="age", y="minority", data=ratings_df)
```



Question 3: Create a group histogram of tenure by minority and add the gender factor

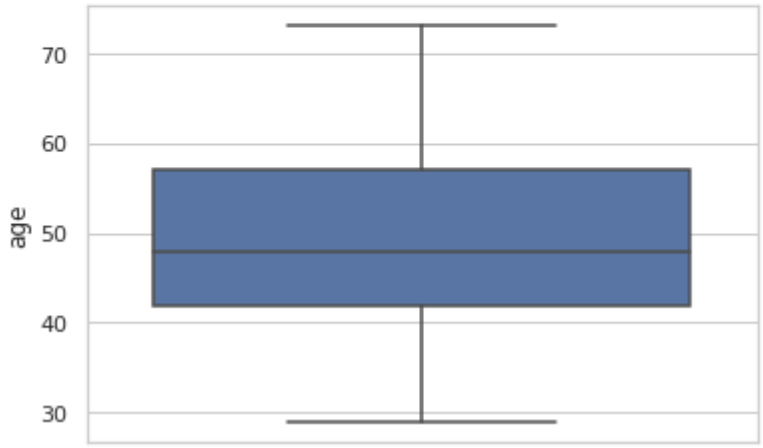
[28]:
```
## insert code
sns.catplot(x='tenure', hue = 'minority', row = 'gender',
            kind='count', data=ratings_df,
            height = 3, aspect = 2)
```

[28]:
```
<seaborn.axisgrid.FacetGrid at 0x7f8c667c66d0>
```



Question 4: Create a boxplot of the age variable

[29]:
```
## insert code
ax = sns.boxplot(y="age", data=ratings_df)
```



Authors

Aije Egwaikhide is a Data Scientist at IBM who holds a degree in Economics and Statistics from the University of Manitoba and a Post-grad in Business Analytics from St. Lawrence College, Kingston. She is a current employee of IBM where she started as a Junior Data Scientist at the Global Business Services (GBS) in 2018. Her main role was making meaning out of data for their Oil and Gas clients through basic statistics and advanced Machine Learning algorithms. The highlight of her time in GBS was creating a customized end-to-end Machine learning and Statistics solution on optimizing operations in the Oil and Gas wells. She moved to the Cognitive Systems Group as a Senior Data Scientist where she will be providing the team with actionable insights using Data Science techniques and further improve processes through building machine learning solutions. She recently joined the IBM Developer Skills Network group where she brings her real-world experience to the courses she creates.

Notes:

1. When the sum of two or more ranges is equal to 100, a line chart type is not ideally suited to display the data. Rather a pie-chart would be appropriate.

Explanation:

A pie chart classically represents numbers in percentages, used to visualize a whole relationship or a portion of a constitution.

Pie charts are not meant to compare individual sections or show exact values.

A pie chart displays a fixed number and how the groups represent part of a whole.

A pie chart represents numbers in percentages, and the sum of all squares must equal 100%.

2.

1.

Question 1

Which of the following is the suitable way to display the average income earned by men and women in a city?

1 / 1 point

- ○ A histogram
- ◉ A bar chart
- ○ A pie chart
- ○ A scatter plot

Correct! A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent

2.

Question 2

What is a suitable way to display relationship between two continuous variables?

1 / 1 point

○ A pie chart

○ A histogram

◉ A scatter plot

○ A bar chart

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

5.

Question 5

When multiple observations are reported for each respondent in the data set, to compute statistics for variables about the respondents, one must:

1 / 1 point

○ Ignore the presence of duplicates and compute statistics as usual

○ Weight data by duplicates

◉ Remove duplicates before running analysis

○ None of the above

Quiz: Data Visualization

Bookmark this page

Graded Quiz due May 20, 2022 06:40 +08

Multiple Choice

4/4 points (graded)

Which of the following is the suitable way to display the average income earned by men and women in a city?

- ○ A pie chart
- ◉ A bar chart
- ○ A histogram
- ○ A scatter plot

✔

Answer
Correct:
Correct! A bar chart of bar graph is a chart or graph that presents the categorical data with rectangular bars with lengths proportional to the values they represent - higher values equate to longer bars.

What is a suitable way to display the relationship between two continuous variables?

- ○ A bar chart
- ○ A pie chart
- ○ A histogram
- ◉ A scatter plot

✔

Answer
Correct:
Correct! Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

When the sum of two or more categories is equal to 100, what chart type is ideally suited for displaying this data?

- ○ A histogram
- ◉ a pie chart
- ○ A box plot
- ○ A line chart

✔

Answer
Correct: Correct! A pie chart is well suited to display the relative proportions of a dataset.

When multiple observations are reported for each respondent in the data set, to compute statistics for variables about the respondents, one must:

- ○ Ignore the presence of duplicates and compute statistics as usual
- ○ Add weights to the data by duplicates
- ◉ Remove the duplicates before performing the analysis
- ○ None of the above

✔

Answer
Correct: Correct!

Submit You have used 1 of 2 attempts Save Show answ

✔ Correct (4/4 points)

Hypothesis testing and Probability Distribution

- Random Numbers and Probability Distributions (4:46)
- State Your Hypothesis (3:35)
- Normal Distribution (3:59)
- T-Distribution (4:50)
- Probability of Getting a High or Low Teaching Evaluation (4:18)
- Lab: Introduction to Probability Distributions
- Quiz: Introduction to Probability Distributions (1 Question)
  Graded Quiz due May 9, 2022, 8:46 AM GMT+8

Random Numbers and Probability Distribution

Now, let us visit some basic definitions about probability as it relates to the most commonly used concepts in statistics. Essentially, probability is a measure between 0 and 1 for the likelihood that something or some event might occur, for instance, you may hear that the stock markets, the chance of stock markets, rising above some point, or falling below some point, is x percent. Or you may hear that the chance for rain is 45% tonight. These are all coming from this very concept of probability.

Essentially, probabilities are measured between 0 and 1, so 45 would be 0.45. The discussion about probability is not complete without a discussion about random variables. Essentially, a random variable is a quantity whose possible values depend in some clearly defined way, on a set of some random events. It's a function that maps out outcomes, that is, points in a probability space.

So probability space essentially is all possible outcomes. If you roll a die, it can have one out of six outcomes, so that's the probability space there. If you roll two dice, you can have one out of 36 outcomes, where each outcome could be considered a random outcome. And a probability distribution is a theoretical model that depicts the possible values any random variable may assume, along with the probability of its occurrence. We'll define this more with examples using two dice.

So consider two dice, a die has six faces. If you roll two dice, it can assume 1 out of 36 discrete outcomes. So, if you were to roll 2 dice, the probability that both die would have 1 as the outcome, would be 1 plus 1 = 2 and there's only 1 possibility of getting that, and that's 1 out of 36. So here we have two die, one black and one white, and if you were to look at the possibility of getting 1 on black and 2 on white, that's one outcome, so that's 1 plus 2 is 3, or you can have 2 on black and 1 on white, that's 2 plus 1 = 3 again.

So you have two ways of getting 3 by rolling 2 dice. So the outcome, or the probability, is 2 out of 36 possible outcomes that are mapped out here. So if you think about the sum of 2 dice being 2, there's only 1 possibility out of 36; getting a 3, you have 2 possibilities; getting a 4, you have 3 possibilities; getting a 5, you have 4 possibilities, and so on and so forth. The maximum most frequently possible number to have as a sum of two dice is 7, and the probability is 6 out of 36, which is 0.167. And if you were to sum these probabilities up, and that is 0.02 plus 0.056 you get 0.083. So, if you sum up all these, they all sum up to 1, and the probability of getting 6 or less than or greater than 6, is 0.58 or getting less than or equal to 6 is 0.417. You sum these up, it's 1.028 plus 0.972 is 1. This plus this is 1; this plus this is 1, and obviously 1 plus 0 is 1. The probability of getting 12, that is, both die show 6, there's 1 out of 36 possible outcomes, which is 0.028.

So the probability sums up to 1, and the probability of getting more than 12 is obviously 0, because the two dice can maximum produce this number. So, nice way of looking at the way a probability distribution space is created by rolling two dice. And if I were to look at the probability of, say, getting some number or less than some number, that's called the cumulative distribution function. If you were to simply chart this probability outcomes in a chart, you get this graph here. So here we have age as our variable, and I created a histogram of age; and then using the mean value of 48.37, which is the minimum mean average age, and a standard deviation of 9.8 years, I can fit a theoretical normal distribution with these two parameters.

Probability – the frequentist approach

- Probability is a measure between zero and one of the likelihood that an event might occur.
  - An event could be the likelihood of a stock market falling below or rising above a certain threshold.
- You are familiar with the weather forecast that often describes the likelihood of rainfall in terms of probability or chance.
  - You often hear the meteorologists explain that the likelihood of rainfall is, for instance, 45%. Thus, 0.45 is the probability that the event, rainfall, might occur.
- The probability associated with any outcome or event must fall in the zero and one (0–1) interval.
  - The probability of all possible outcomes must equate to one.

Random variables

- A random variable is a "quantity whose possible values depend, in some clearly-defined way, on a set of random events."
  - It "is a function that maps outcomes (that is, points in a probability space)."
  - Rolling two dice can have one of 36 outcomes where each outcome could be considered a random outcome.
- A probability distribution is essentially a theoretical model depicting the possible values a random variable may assume along with the probability of occurrence.
  - We can define probability distributions for both discrete and continuous random variables.

Casino Royale: Roll the Dice

A die has six faces, so rolling two dice can assume one of the 36 discrete outcomes:

- Each die can assume one of the six outcomes in a roll. Hence rolling two dice together will return one out of 36 outcomes.
- If one (1) comes up on each die, the outcome will be 1 + 1 = 2, and the probability associated with this outcome is one out of thirty-six (1/36) because no other combination of the two dice will return two (2).
- On the other hand, I can obtain three (3) with the roll of two dice by having either of the two dice assume one and the other assuming two and vice versa.
- Thus, the probability of an outcome of three with a roll of two dice is 2 out of 36 (2/36).

36 ways to feel lucky

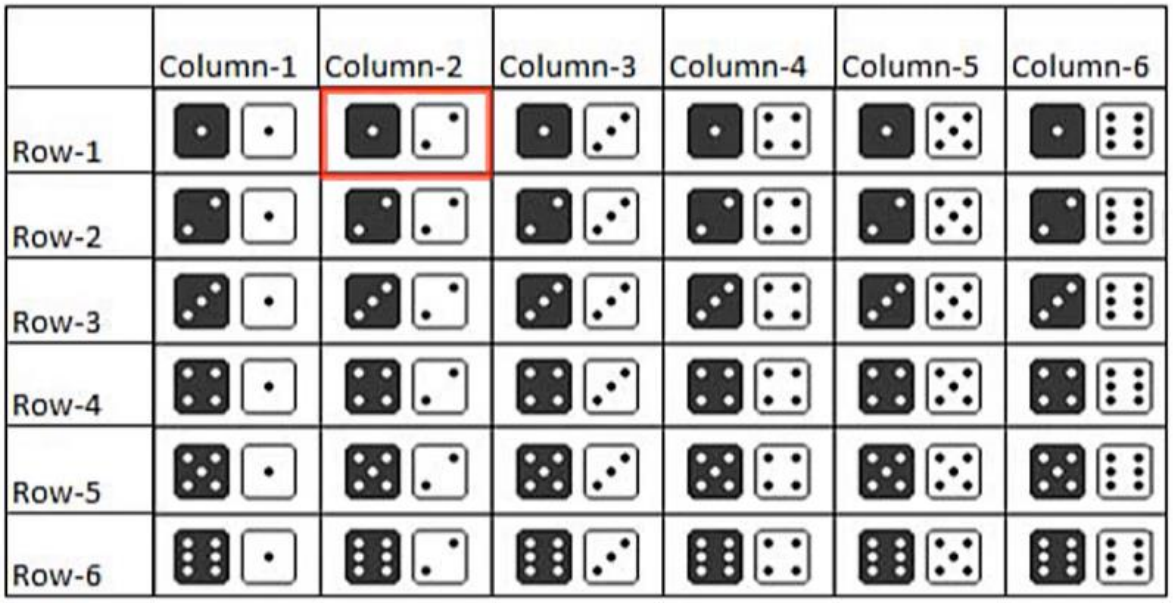


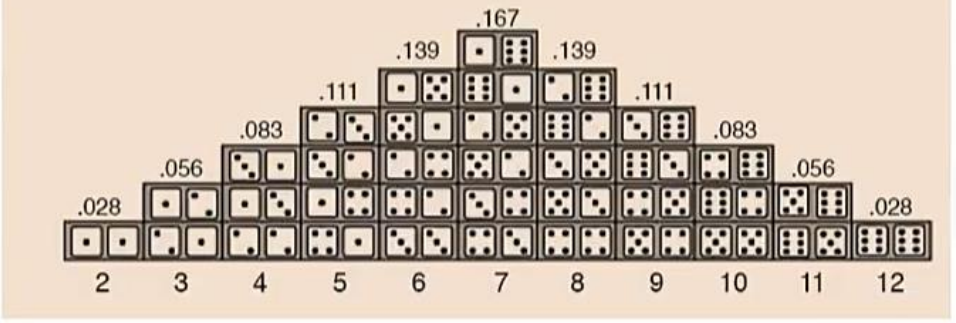
Figure 6.2 All possible outcomes of rolling two dice

Source: http://www.edcollins.com/backgammon/diceprob.htm

So here we have two die, one black and one white, and if you were to look at the possibility of getting 1 on black and 2 on white, that's one outcome, so that's 1 plus 2 is 3, or you can have 2 on black and 1 on white, that's 2 plus 1 = 3 again. So you have two ways of getting 3 by rolling 2 dice. So the outcome, or the probability, is 2 out of 36 possible outcomes that are mapped out here.

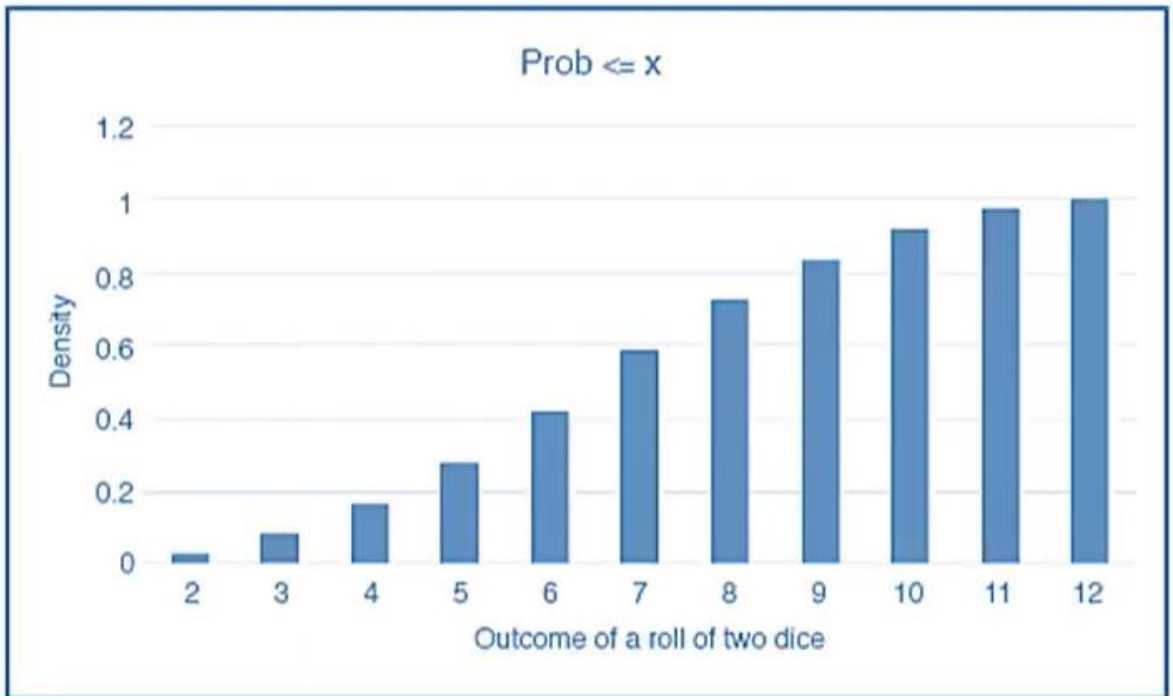
Lucky seven

Sum of Two Dice, x	f(x)	Probability	F(x)	Prob <=x	Prob > x
2	1/36	0.028	1/36	0.028	0.972
3	2/36	0.056	3/36	0.083	0.917
4	3/36	0.083	6/36	0.167	0.833
5	4/36	0.111	10/36	0.278	0.722
6	5/36	0.139	15/36	0.417	0.583
7	6/36	0.167	21/36	0.583	0.417
8	5/36	0.139	26/36	0.722	0.278
9	4/36	0.111	30/36	0.833	0.167
10	3/36	0.083	33/36	0.917	0.083
11	2/36	0.056	35/36	0.972	0.028
12	1/36	0.028	1	1	0



So if you think about the sum of 2 dice being 2, there's only 1 possibility out of 36; getting a 3, you have 2 possibilities; getting a 4, you have 3 possibilities; getting a 5, you have 4 possibilities, and so on and so forth. The maximum most frequently possible number to have as a sum of two dice is 7, and the probability is 6 out of 36, which is 0.167. And if you were to sum these probabilities up, and that is 0.02 plus 0.056 you get 0.083. So, if you sum up all these, they all sum up to 1, and the probability of getting 6 or less than or greater than 6, is 0.58 or getting less than or equal to 6 is 0.417. You sum these up, it's 1.028 plus 0.972 is 1. This plus this is 1; this plus this is 1, and obviously 1 plus 0 is 1. The probability of getting 12, that is, both die show 6, there's 1 out of 36 possible outcomes, which is 0.028. So the probability sums up to 1, and the probability of getting more than 12 is obviously 0, because the two dice can maximum produce this number. So, nice way of looking at the way a probability distribution space is created by rolling two dice.

Probability of less than equal to



And if I were to look at the probability of, say, getting some number or less than some number, that's called the cumulative distribution function. If you were to simply chart this probability outcomes in a chart, you get this graph here. So here we have age as our variable, and I created a histogram of age; and then using the mean value of 48.37, which is the minimum mean average age, and a standard deviation of 9.8 years, I can fit a theoretical normal distribution with these two parameters.

State Your Hypothesis

In this video, I will illustrate how to state your hypothesis when you're comparing the averages between two entities. We will use basketball as an example, using Michael Jordan and Wilt Chamberlain, the two highest scorers in the history of basketball, as examples. If you were to recall, you would know that Michael Jordan, on average, scored 30.12 in each game, and Chamberlain averaged around 0.36 points per game. If you were to compare the two averages between Michael Jordan and Chamberlain, and even though they are very similar looking numbers, we need to set up a statistical hypothesis testing.

We are interested in comparing the average points scored by the two basketball players, and the comparison of means or averages is available in three flavors. First, we can assume that the average points per game scored by the two players, Jordan and Chamberlain, are the same, that is, the difference between their mean scores is 0. If their averages are the same, so average of one minus average of other should be 0, and this becomes our null hypothesis. Let's say if mu j, mu subscript j, represents the average points per game scored by Michael Jordan, and mu subscript c represents the average points per game scored by Wilt Chamberlain, we can state the null hypothesis to be: mu j equal mu c, that is, the average scored by Jordan and average scored by Chamberlain are the same.

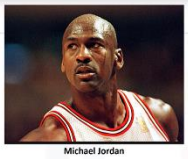
And the alternative hypothesis would be that, no, the averages are not the same; they're different. So the alternative hypothesis h subscript a compared to null hypothesis, h subscript 0, or o, the alternative is that the averages are not the same; their average scores are different.

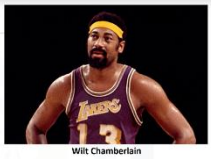
Now, the other option, the second option, is to assume that Jordan scored better or higher and, in that case, our null hypothesis is the average score by Michael Jordan is greater than or equal to the average score by Wilt Chamberlain. And, in this particular case, the alternative hypothesis would be different. It wouldn't be not equal to, but it would be less than. So, the alternative would be that the average scored by Jordan is less than that by Wilt Chamberlain. And, by the same account, our third option would be that the null hypothesis is that, in fact, Jordan scored less than Wilt Chamberlain. And the alternative hypothesis will be the reverse of it, saying that, no, Michael Jordan scored higher than Wilt Chamberlain.

So, in a nutshell, we have three options. **1.** We can say the averages are the same and the null would be, no, they are not the same, not equal. **2.** We can say the average is less than the null is that Jordan's average is higher than Chamberlain's and the alternative would be the reverse of it. **3.** And the third option is to say that the average scored by Jordan is less than the average scored by Chamberlain, and the reverse of it will be the alternative hypothesis. So, three ways of defining a hypothesis.

State Your Hypothesis

Murtaza Haider and Aije Egwaikhide


Michael Jordan


Wilt Chamberlain

The Basketball Giants

Jordan versus Chamberlain

Jordan's 30.12 points per game

Chamberlain's 30.06 points per game

Comparing basketball giants

- The comparison of means (averages) comes in three flavors.
- First, you can assume that the mean points per game scored by both Jordan and Chamberlain are the same.
- That is, the difference between the mean scores of the two basketball legends is zero.
  - This becomes our null hypothesis.
- Let μ_j represent the average points per game scored by Jordan
- Let μ_c represent the average points per game scored by Chamberlain.

The Null and Alternative hypotheses

- Null Hypothesis: H_o
- Alternative Hypothesis: H_a

H_0:$\mu_j = \mu_c$

The alternative hypothesis, denoted as H_a, is as follows:

H_a:$\mu_j \neq \mu_c$; their average scores are different.

we can state the null hypothesis to be: mu j equal mu c, that is, the average scored by Jordan and average scored by Chamberlain are the same. And the alternative hypothesis would be that, no, the averages are not the same; they're different. So the alternative hypothesis h subscript a compared to null hypothesis, h subscript 0, or o, the alternative is that the averages are not the same; their average scores are different

Option 2: Jordan is better

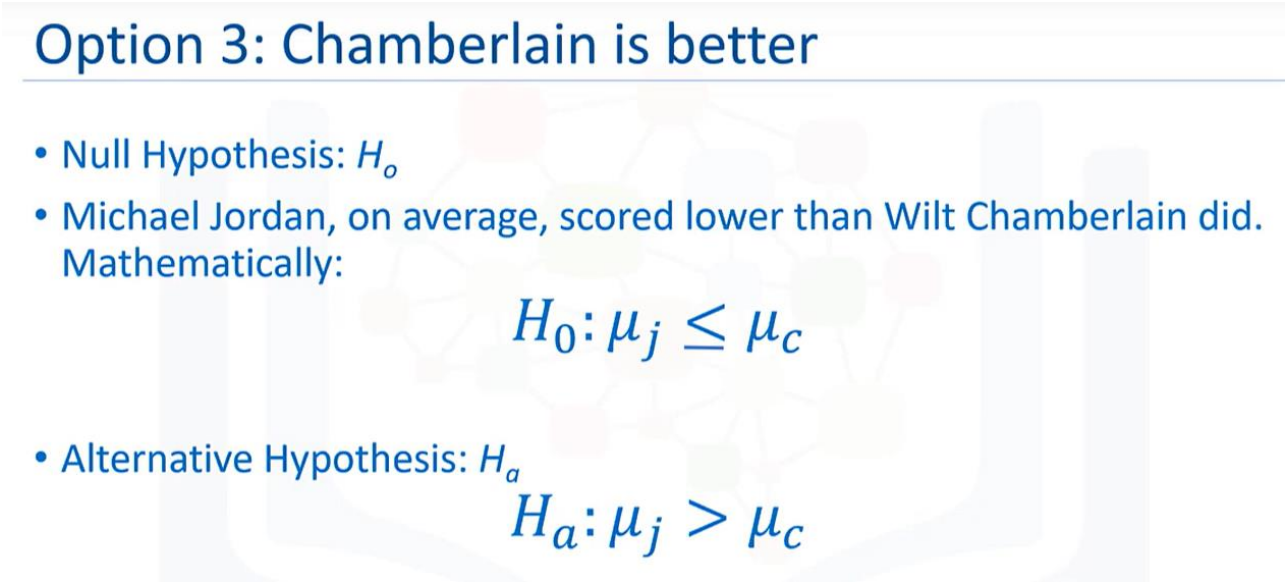
- Null Hypothesis: H_o
- Now let us work with a different null hypothesis and assume that Michael Jordan, on average, scored higher than Wilt Chamberlain did. Mathematically:

$$H_0: \mu_j \geq \mu_c$$

- Alternative Hypothesis: H_a

$$H_a: \mu_j < \mu_c$$

our null hypothesis is the average score by Michael Jordan is greater than or equal to the average score by Wilt Chamberlain. And, in this particular case, the alternative hypothesis would be different. It wouldn't be not equal to, but it would be less than. So, the alternative would be that the average scored by Jordan is less than that by Wilt Chamberlain.



Our third option would be that the null hypothesis is that, in fact, Jordan scored less than Wilt Chamberlain. And the alternative hypothesis will be the reverse of it, saying that, no, Michael Jordan scored higher than Wilt Chamberlain.

Three options.

1. We can say the averages are the same and the null would be, no, they are not the same, not equal.

2. We can say the average is less than the null is that Jordan's average is higher than Chamberlain's and the alternative would be the reverse of it.

3. And the third option is to say that the average scored by Jordan is less than the average scored by Chamberlain, and the reverse of it will be the alternative hypothesis. So, three ways of defining a hypothesis.

Normal Distribution

Let me introduce you to normal distribution, which is one of the most commonly used distributions in statistical analysis, and even in everyday conversations. A large body of academic, scholarly, and professional work rests on the assumption that the underlying data follows a normal distribution.

The defining characteristics of normal distribution is this bell-shaped curve, which you're familiar with, from your textbooks. Mathematically, normal distribution is presented by this equation. We can say that the normal distribution relies on three inputs and f stands for functions, a function of x mu and sigma. X is your data.

x is a random variable and it can attain any very reasonable value. Mu stands for the mean, and sigma is standard deviation; and the mathematical formulation is right here, which is 1 divided by sigma, times, and then the square root of 2 times pi; pi is the 3.14, or 22 divided by 7. And then you have the exponential here; do not forget this minus sign!

So, exponential of this entity, which is minus, and then the numerator x minus mean, or x minus mu 2 whole square, divided by 2 times sigma squared. So let me explain. 1 divided by the standard deviation, and then the square root of 2 times pi; 2 is known, and pi is known, so is the value for exponential.

And what is not known is the sigma, which you can obtain from the data, that is, standard deviation; and the mean, which is also coming from data. So, you have the mean and the standard deviation, and x is the random variable whose mean and standard deviation you're using. You put this all together and then you get the normal distribution, the bell-shaped curve that you saw earlier.

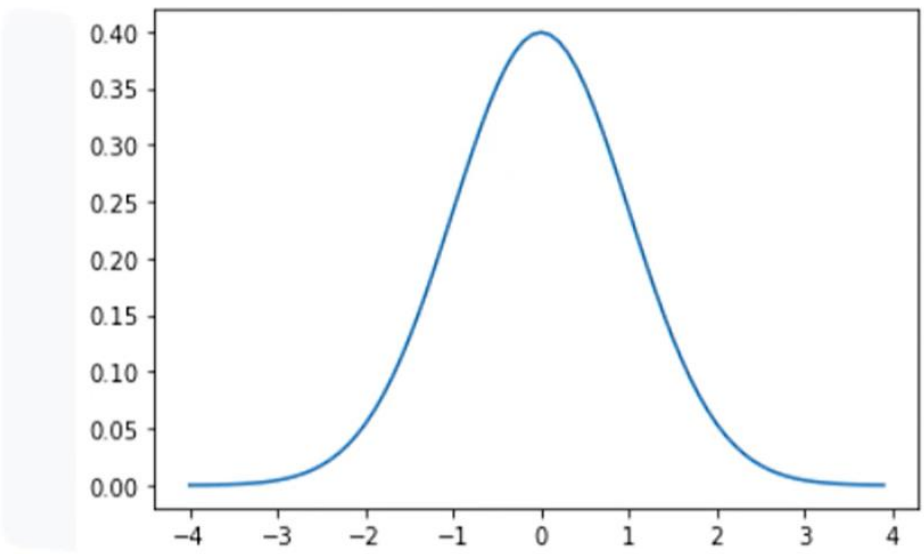
I also introduce you to the standard normal. A standard normal is, and when we say that there's x is a variable, that has a mean zero, and standard deviation of 1. So what's mean 0 and standard deviation 1 look like? If you replace mu with zero, and standard deviation or sigma with 1, the equation is reduced to this entity, which is 1 divided by 2 times pi. Notice the sigma here (which I have grayed out a bit so that it doesn't interfere), sigma is 1 so 1 times anything would be the same, so I've removed this. And then e to the power, minus x, minus mu, but because mu is 0, so x minus 0 is x, so x squared divided by 2, times sigma square, sigma remember is 1, the square of 1 is 1, so 2 times 1.

So removing sigma, because it's 1, and anything multiplied with 1 is the same entity. So how do i generate this normal density or a bell curve? Let's assume that the underlying variable has a mean 0, and the standard deviation of 1 and x varies between -4 and 4. So the mean is 0, and the minimum value is -4, the maximum value is 4, and I would substitute these -4 to 4 values

in this equation; this is the only thing that's changing; the x here is the only entity that is changing. And let's see if this could generate the standard normal curve. Let us do this in Python. We'll use the matplotlib function, which you are already familiar with for the graphics, numpy library, as well as the norm.pdf function in the scipy.stats library. In this example, I have used increments of 0.1. This will generate the standard normal curve that you hear about in statistics.

Normal Distribution

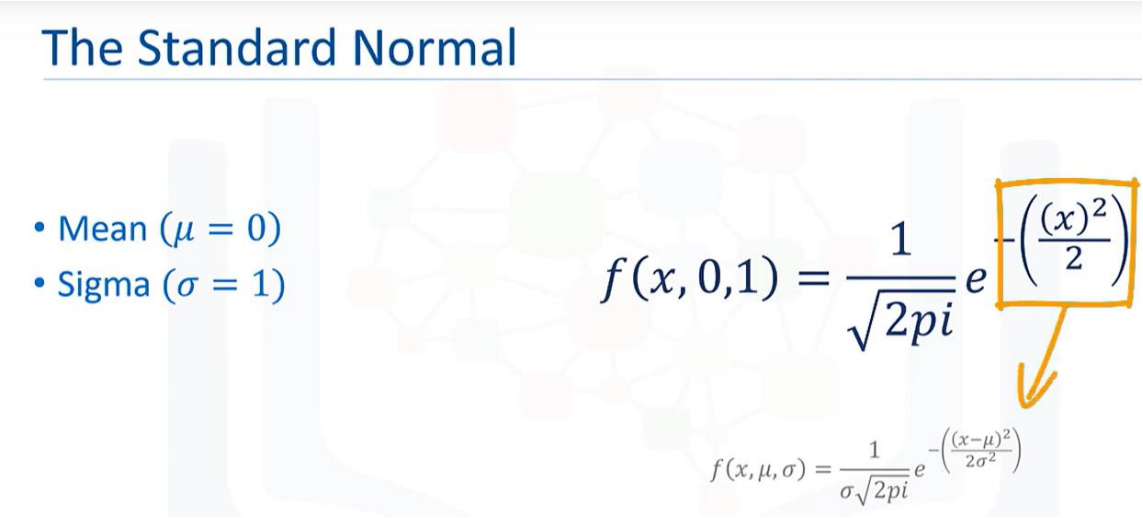
Normal Distribution



The Math behind Normality

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Explanation. 1 divided by the standard deviation, and then the square root of 2 times pi; 2 is known, and pi is known, so is the value for exponential. And what is not known is the sigma, which you can obtain from the data, that is, standard deviation; and the mean, which is also coming from data. So, you have the mean and the standard deviation, and x is the random variable whose mean and standard deviation you're using. You put this all together and then you get the normal distribution, the bell-shaped curve that you saw earlier.



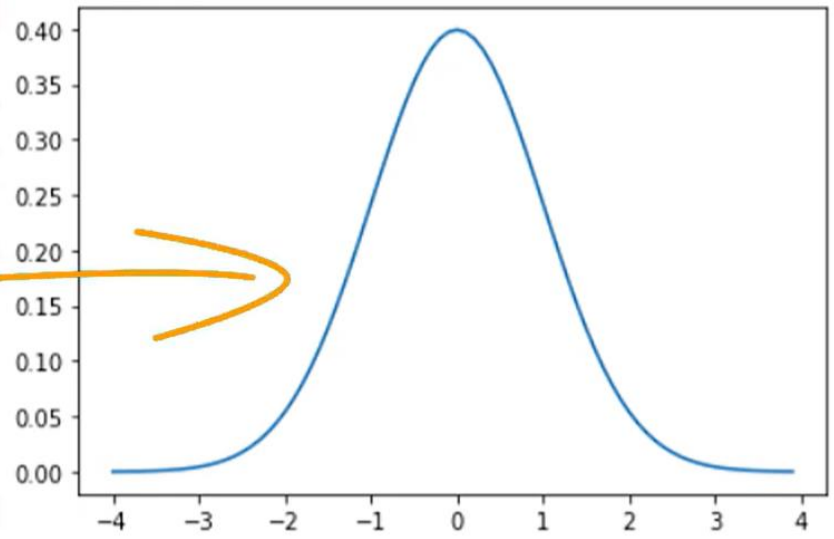
Explanation: 1 divided by the standard deviation, and then the square root of 2 times pi; 2 is known, and pi is known, so is the value for exponential. And what is not known is the sigma, which you can obtain from the data, that is, standard deviation; and the mean, which is also coming from data. So, you have the mean and the standard deviation, and x is the random variable whose mean and standard deviation you're using. You put this all together and then you get the normal distribution, the bell-shaped curve that you saw earlier.

A standard normal is, and when we say that there's x is a variable, that has a mean zero, and standard deviation of 1. So what's mean 0 and standard deviation 1 look like? If you replace mu with zero, and standard deviation or sigma with 1, the equation is reduced to this entity, which is 1 divided by 2 times pi. Notice the sigma here (which I have grayed out a bit so that it doesn't interfere), sigma is 1 so 1 times anything would be the same, so I've removed this. And then e to the power, minus x, minus mu, but because mu is 0, so x minus 0 is x, so x squared divided by 2, times sigma square, sigma remember is 1, the square of 1 is 1, so 2 times 1. So removing sigma, because it's 1, and anything multiplied with 1 is the same entity.

The Bell Curve

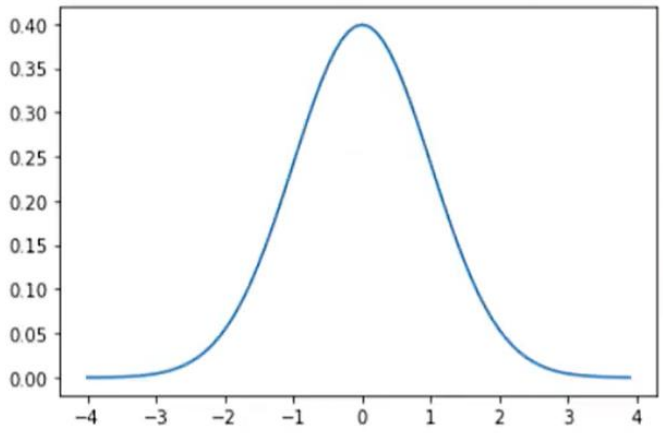
$$\mu = 0, \sigma = 1$$
$$-4 < x < 4$$

$$f(x, 0,1) = \frac{1}{\sqrt{2 * pi}} e^{\frac{(x)^2}{2}}$$



To generate normal curve in Python:

The Bell Curve – in Python




```
1  import numpy as np
2  import matplotlib.pyplot as plt
3  from scipy.stats import norm
4
5  # Plot between -4 and 4 with 0.1 steps.
6  x_axis = np.arange(-4, 4, 0.1)
7  # Mean = 0, SD = 1.
8  plt.plot(x_axis, norm.pdf(x_axis,0,1))
9  plt.show()
```

T-Distribution / T-test

The other commonly used statistical distribution is known as the student's t-distribution. William C. Lee Gossett specified the t-distribution. In fact, he published a paper in Biometrica in 1908, and he published it under the pseudonym student. He worked for the Guinness Brewery in Durban Ireland where he worked with small samples of barley. Mr. Gossett is the unsung hero of statistics. He published his work under a pseudonym because of the restrictions from his employer.

Apart from his published work, his other contributions to statistical analysis are equally significant. The Cult of Statistical Significance, a must read book for anyone interested in data science, chronicles Mr. Gosset's work and how other influential statisticians of the time, namely Ronald Fisher and Egon Pearson, by way of their academic bonafides, ended up being more influential than the equally deserving Mr. Gosset.

The normal distribution describes the mean for a population, whereas the t-distribution describes the mean of samples drawn from a population. The t-distribution for each sample could be different, and the t-distribution resembles the normal distribution for large sample sizes. Here, I present normal distribution, which is drawn in blue, and the t-distribution with a degree of freedom of 1. As the degrees of freedom increase, the t-distribution curve becomes more similar to the normal distribution.

In statistical analysis, several statistical tests rely on t-distribution. For instance, a comparison of means, use the t-distribution and it's also known as the t-test. We have been working with a dataset comprising teaching evaluations of instructors from University of Texas. And I will illustrate the use of t-test or t-distribution with the question of, "Does instructor evaluation score differ by gender; do males and females get different teaching evaluations from students?" Now, if I were to take the same dataset and compute the means and standard deviations, I can test this statistically.

I have computed the visual representation of the average teaching evaluation score for male and female instructors. The blue bar represents the average teaching evaluation value for female; the orange bar represents the average teaching evaluation value for male.

By eyeballing it, it is around 4, and slightly less for female. Now it's a small difference between males and females. The question is: "Is this difference statistically significant?" To use a t-test, you have to make sure some assumptions are met.

The first assumption for a t-test is that:

- the scale of measurement applied to the data collected follows a continuous or ordinal scale.
- The second assumption is that the data is collected from a representative, randomly selected portion of the total population.
- The third assumption is the data, when plotted, will follow a normal distribution.
- And the final assumption, is homogeneity of variance to avoid the test statistics to be biased towards larger sample sizes.

There is a test for this, which will be discussed later.

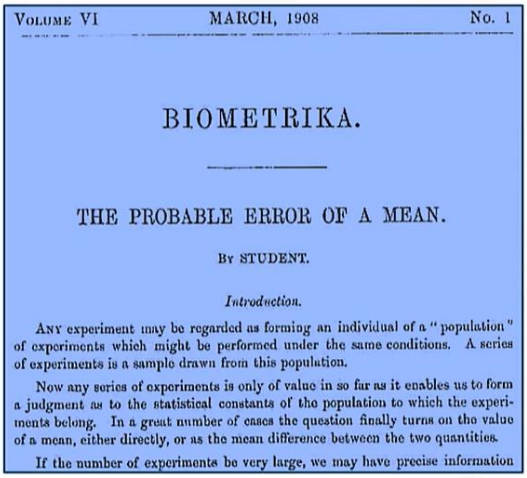
Before we go perform the test in Python, first, we will state our hypothesis. The null hypothesis is as follows: "There is no difference in evaluation scores for males and females."

The alternate hypothesis is: "There is a difference in evaluation scores between males and females."

Then, set alpha level to 0.05. To do this in Python, we will use the t-test independent sample in the scipy.stats function. The function takes in the two samples it is trying to test; the statistical difference of means, 4, in our example, is the female evaluation scores versus all the male evaluation scores. It will return a t-statistic and a p-value. Since the p-value is less than 0.05, the alpha level, we reject the null hypothesis, as there is enough evidence that there is a statistical difference in teaching evaluations based on gender.

T Distribution

William Sealy Gossett



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

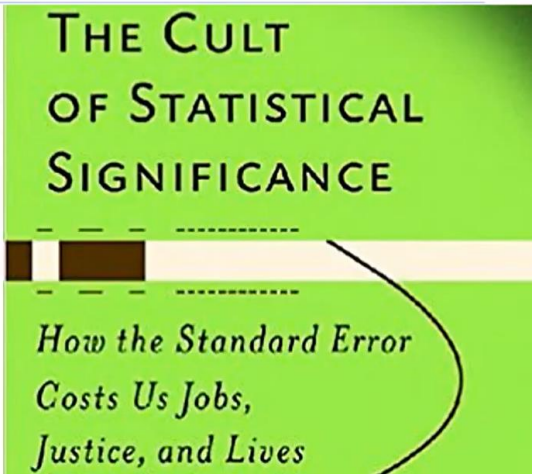
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

William C. Lee Gossett specified the t-distribution. In fact, he published a paper in Biometrica in 1908, and he published it under the pseudonym student. He worked for the Guinness Brewery in Durban Ireland where he worked with small samples of barley. Mr. Gossett is the unsung hero of statistics. He published his work under a pseudonym because of the restrictions from his employer.

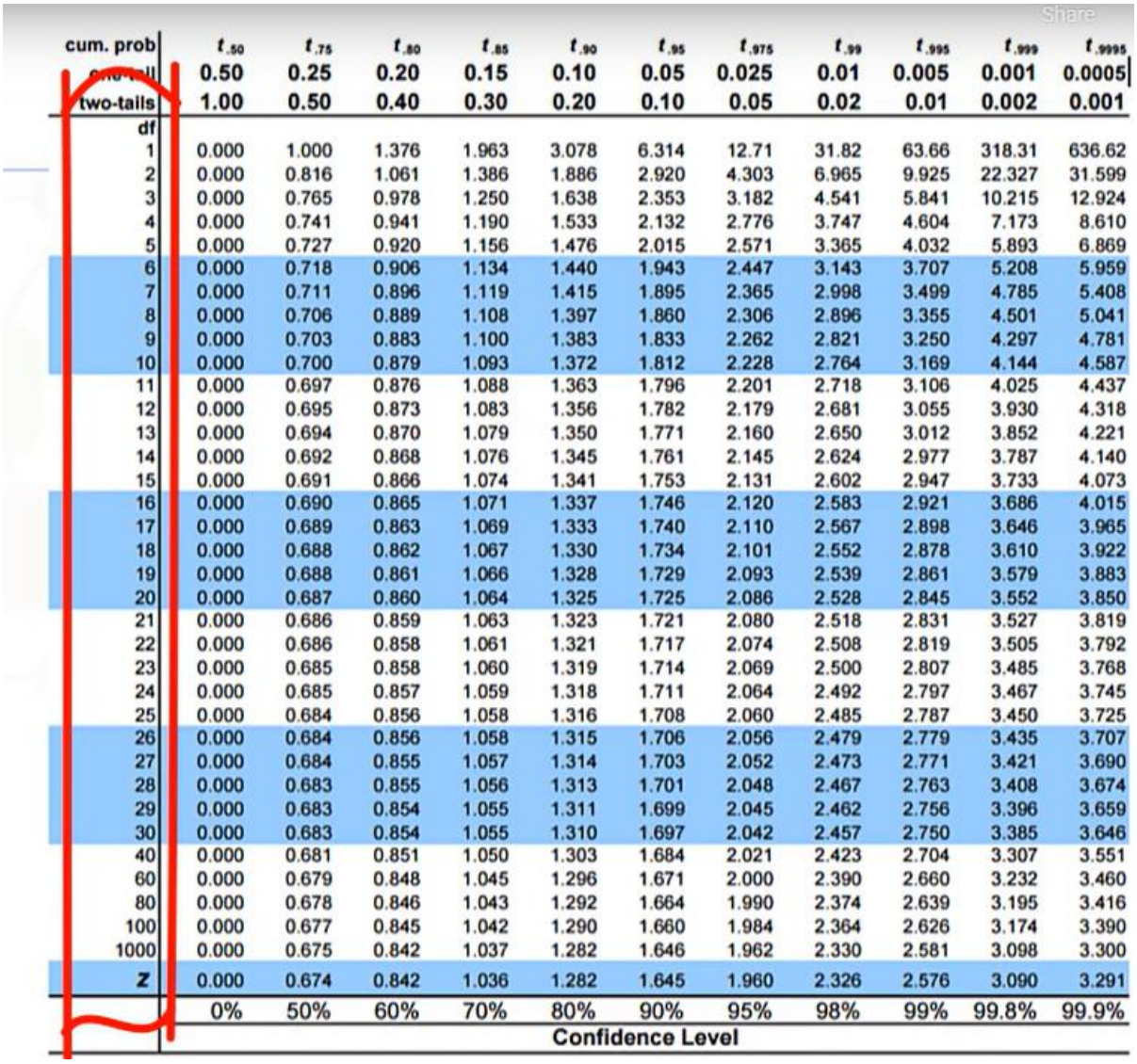
Ziliak and McCloskey



The Cult of Statistical Significance, a must read book for anyone interested in data

science, chronicles Mr. Gosset's work and how other influential statisticians of the time, namely Ronald Fisher and Egon Pearson, by way of their academic bonafides, ended up being more influential than the equally deserving Mr. Gossett.

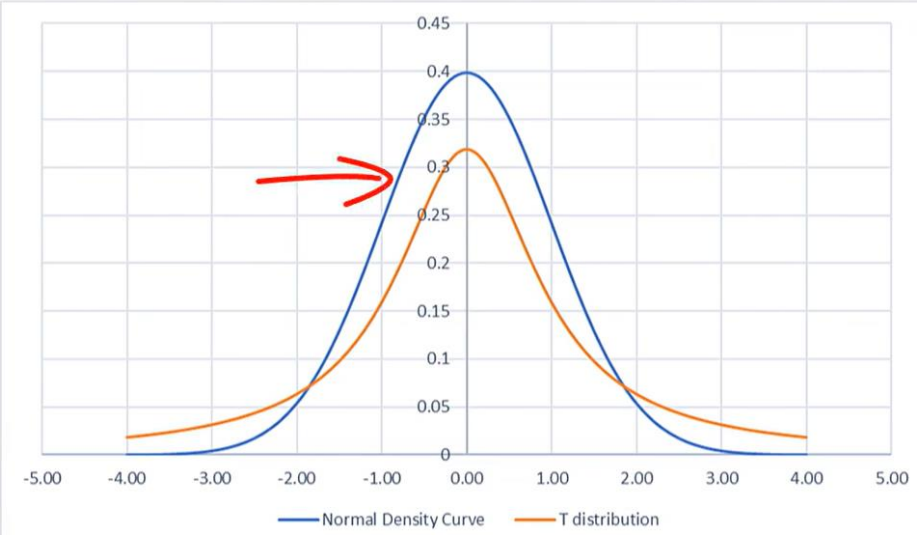
T-Distribution



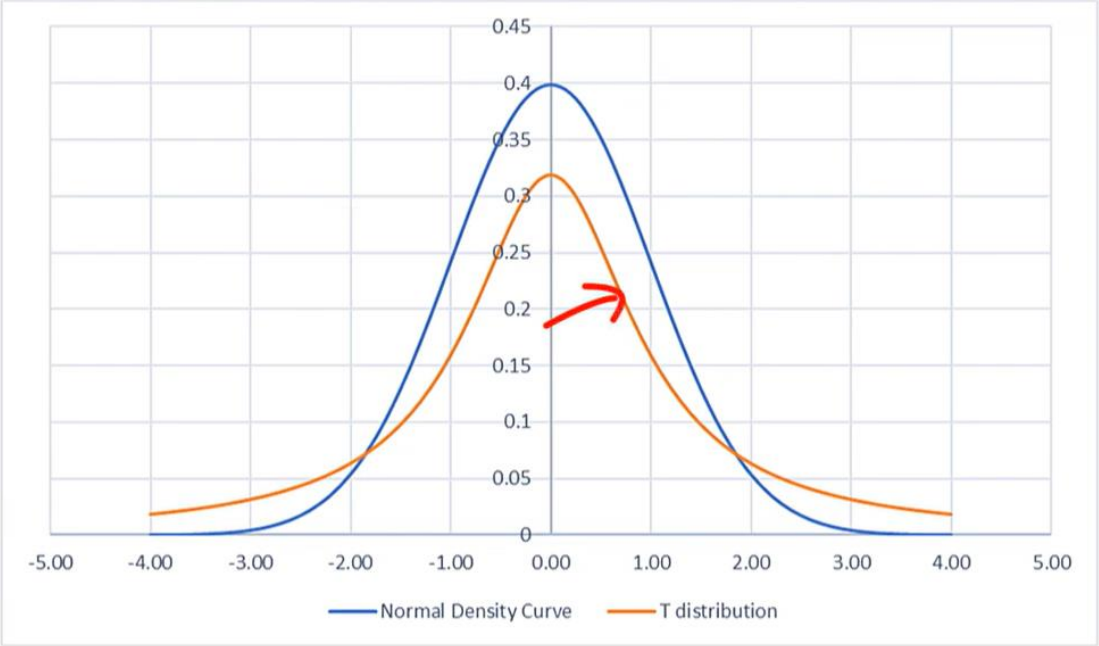
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

The normal distribution describes the mean for a population, whereas the t-distribution describes the mean of samples drawn from a population. The t-distribution for each sample could be different, and the t-distribution resembles the normal distribution for large sample sizes.

Comparing Normal and T-Distribution

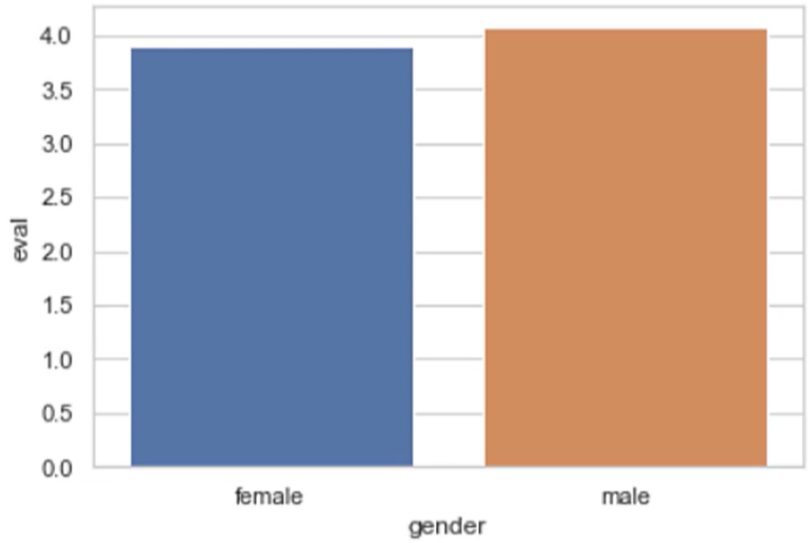


Comparing Normal and T-Distribution



Here, I present normal distribution, which is drawn in blue, and the t-distribution with a degree of freedom of 1. As the degrees of freedom increase, the t-distribution curve becomes more similar to the normal distribution. In statistical analysis, several statistical tests rely on t-distribution. For instance, a comparison of means, use the t-distribution and it's also known as the t-test.

Is the Difference Statistically Significant?



Does instructor evaluation score differ by gender; do males and females get different teaching evaluations from students?" Now, if I were to take the same dataset and compute the means and standard deviations, I can test this statistically. I have computed the visual representation of the average teaching evaluation score for male and female instructors. The blue bar represents the average teaching evaluation value for female; the orange bar represents the average teaching evaluation value for male. By eyeballing it, it is around 4, and slightly less for female. Now it's a small difference between males and females. The question is: "Is this difference statistically significant?"

T-test

Testing for statistical significance

- First the Assumptions
  - Scale of measurement
  - Simple Random Sample
  - Bell-shaped distribution
  - Homogeneity of variance

T-test

Testing for statistical significance

- First the Assumptions
  - Scale of measurement
  - Simple Random Sample
  - Bell-shaped distribution
  - Homogeneity of variance
- State your hypothesis
  - Null hypothesis: μ1 = μ2 ("there is no difference in evaluation scores for male and females")
  - Alternative hypothesis: μ1 ≠ μ2 ("there is a difference in evaluation scores between male and females")
  - alpha (α) level = 0.05

T-test in python

```
1 import scipy.stats
```

```
1 scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],
2                       ratings_df[ratings_df['gender'] == 'male']['eval'])
```

```
Ttest_indResult(statistic=-3.249937943510772, pvalue=0.0012387609449522217)
```

Since the p-value is less than 0.05, the alpha level, we reject the null hypothesis, as there is enough evidence that there is a statistical difference in teaching evaluations based on gender.

Hypothesis testing and Probability Distribution/ Probability of Getting a High or Low Teaching Evaluation

Let me illustrate how to obtain the probability of getting a high or low teaching evaluation score from our dataset. First, an important concept is **the standardization of a variable, such that it returns a dataset with a mean of 0, and a standard deviation of 1.** I use the formula in equations shown here, where the standardization is taking a variable x and subtracting from it the average value mu. Then dividing it by the standard deviation, so that if the teaching evaluation score of an instructor, on a scale of 1 to 5 is 4.5, we subtract the average teaching evaluation of 3.998 from it, and divide it by the standard deviation, which is 0.554, resulting in a z score of 0.906.

If we were to just display the data as a histogram, you would see that it has a mean around 0 and the spread is shown on a scale, where the x axis varies from -3 to 2. In a case where you do not have access to a computer with statistical software, you can still compute probabilities from a probability table, using a simple and standard normal table found in statistics textbooks, or downloaded online. A copy of such a table is on the right. Notice that the normal distribution graph to the left is grayed out in some parts. That grayed out area represents the probability of getting some value z, in this case z. This value of z, or less than, we will need to first standardize the variable to determine the probability of a teaching evaluation score higher than 4.5 or less than 4.5.

So let's say we have a dataset where the average teaching evaluation is 3.998, and the standard deviation is 0.554 and we are interested in determining the probability of getting a teaching evaluation score of 4.5, or less. So, from the table that I showed in the last slide, we can determine this. If we were to standardize the data, it becomes 0.906 because the accuracy of this table is only good for two decimal places. So 0.906 effectively becomes 0.91. So we get a 0.8186 value here, hence the probability of obtaining a teaching evaluation score of 4.5, or less, is 0.8186.

If you were to look at this graphic, you will see that I have plotted the area under the curve by shading it gray. That's the area that depicts the probability of an instructor receiving a teaching evaluation of less than or equal to 4.5, and that probability is 0.8176 or 81.76 percent.

Now what will be the probability of receiving a teaching evaluation score of greater than 4.5? In fact, you can see from the next graphic, that the probability is: the reverse of one that we saw earlier, and hence the probability of obtaining a teaching evaluation score of greater than 4.5 is 18.24 percent, which is the area shaded in gray. The reason for this is because the area under the normal distribution curve is equal to 1. So 1 minus 0.8176 will give you the area for evaluation scores greater than 4.5. Let me illustrate the example of getting a teaching evaluation score of greater than 4.5 in Python.

We will use the norm.cdf function in the scipy.stats package. After finding the mean and standard deviation, we plug it into the function with the x value of 4.5, and you'll get the area to the left, which is the less than 4.5 area. Because we want the area to the right of 4.5, that is, the probability of greater than 4.5, we will remove the value from 1, as indicated here.

Probability of Getting a High or Low Teaching Evaluation

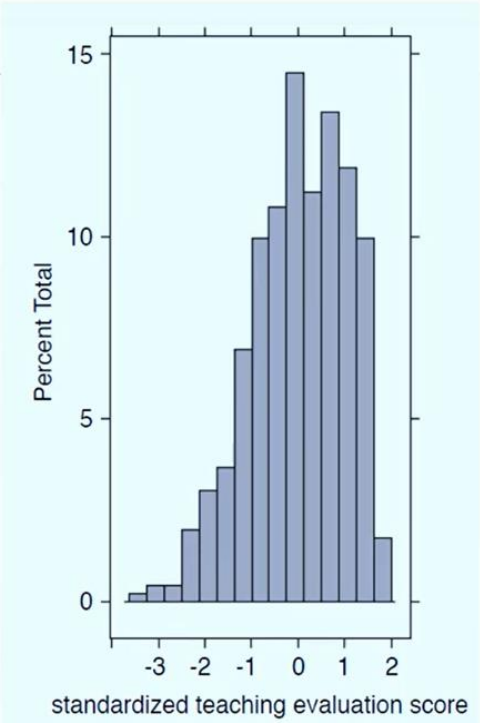
*data taken from pervious example)

an important concept is **the standardization of a variable, such that it returns a dataset with a mean of 0, and a standard deviation of 1.**

Standardization

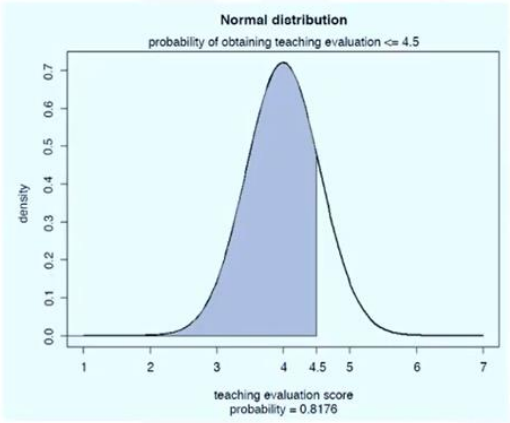
$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{4.5 - 3.998}{0.554} = 0.906$$



the formula in equations shown here, where the standardization is taking a variable x and subtracting from it the average value mu. Then dividing it by the standard deviation, so that if the teaching evaluation score of an instructor, on a scale of 1 to 5 is 4.5, we subtract the average teaching evaluation of 3.998 from it, and divide it by the standard deviation, which is 0.554, resulting in a z score of 0.906. If we were to just display the data as a histogram, you would see that it has a mean around 0 and the spread is shown on a scale, where the x axis varies from -3 to 2.

Normal Distribution Table



Probability Content from -oo to Z

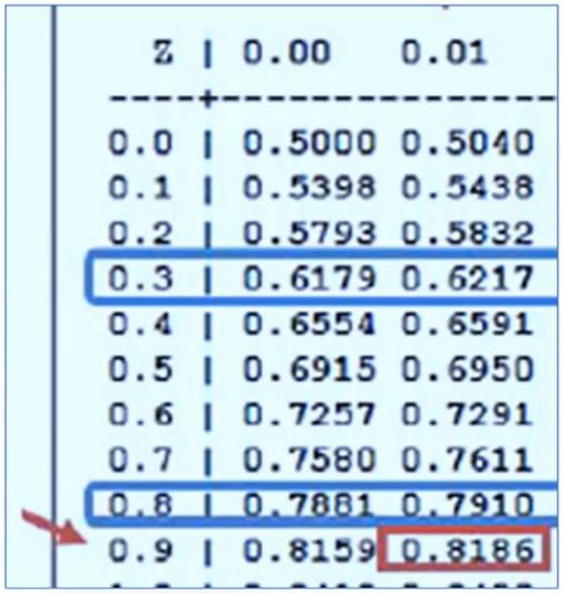
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

In a case where you do not have access to a computer with statistical software, you can still compute probabilities from a probability table, using a simple and standard normal table found in statistics textbooks, or downloaded online. A copy of such a table is on the right. Notice that the normal distribution graph to the left is grayed out in some parts. That grayed out area represents the probability of getting some value z, in this case z. This value of z, or less than, we will **need to first standardize the variable to determine the probability of a teaching evaluation score higher than 4.5 or less than 4.5**

Standardization

$$z = \frac{x - \mu}{\sigma}$$

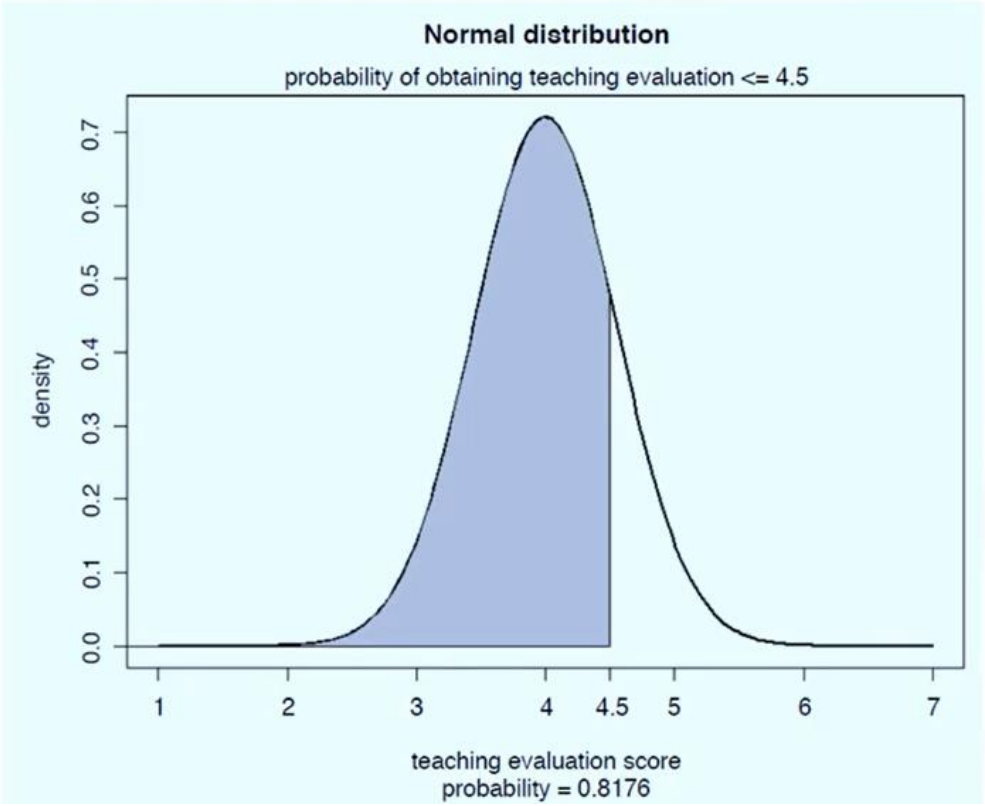
$$z = \frac{4.5 - 3.998}{0.554} = 0.906$$



z	0.00	0.01
0.0	0.5000	0.5040
0.1	0.5398	0.5438
0.2	0.5793	0.5832
0.3	0.6179	0.6217
0.4	0.6554	0.6591
0.5	0.6915	0.6950
0.6	0.7257	0.7291
0.7	0.7580	0.7611
0.8	0.7881	0.7910
0.9	0.8159	0.8186

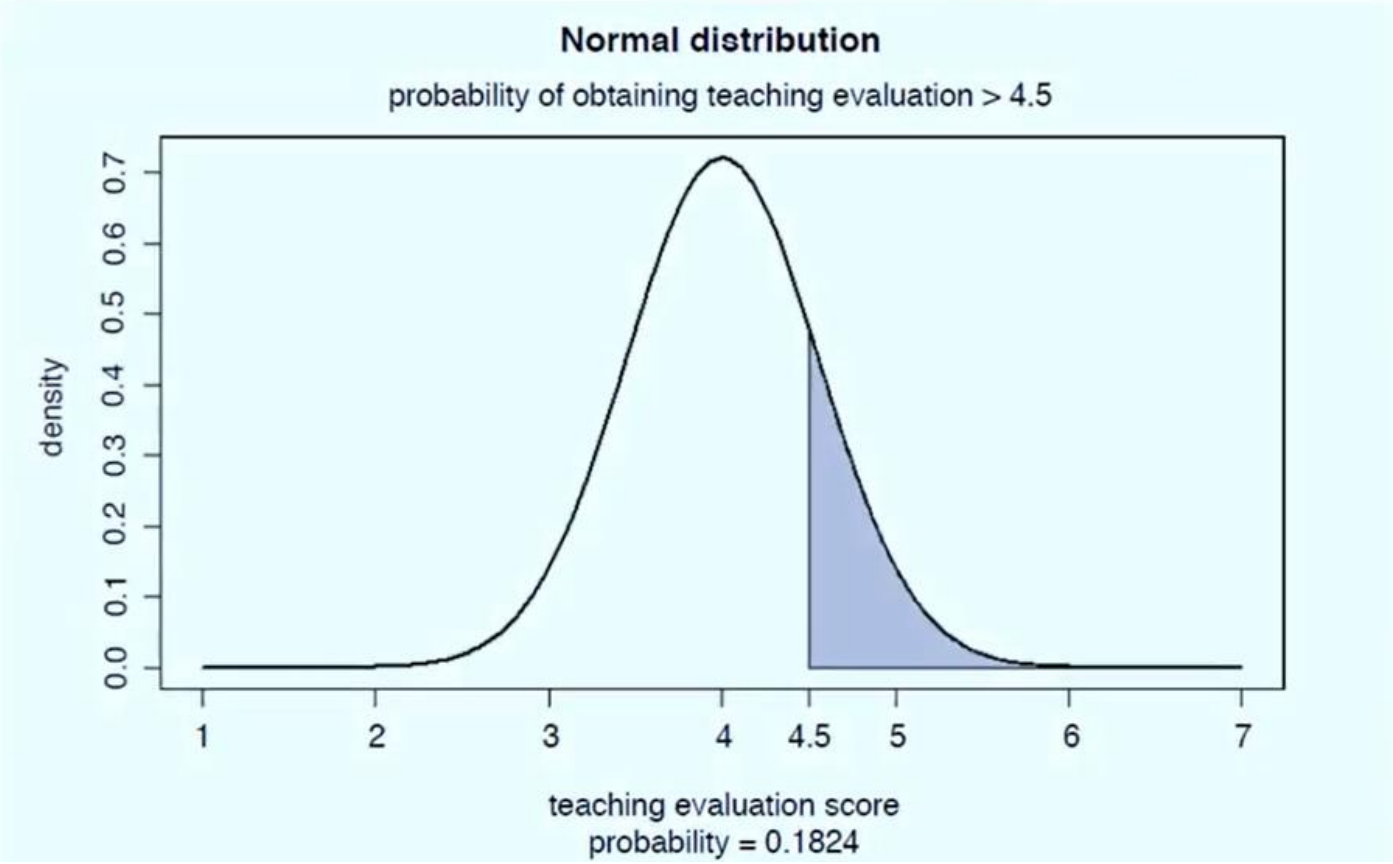
So let's say we have a dataset where the average teaching evaluation is 3.998, and the standard deviation is 0.554 and we are interested in determining the probability of getting a teaching evaluation score of 4.5, or less. So, from the table that I showed in the last slide, we can determine this. If we were to standardize the data, it becomes 0.906 because the accuracy of this table is only good for two decimal places. So 0.906 effectively becomes 0.91. So we get a 0.8186 value here, hence the probability of obtaining a teaching evaluation score of 4.5, or less, is 0.8186.

Probability <= 4.5



f you were to look at this graphic, you will see that I have plotted the area under the curve by shading it gray. That's the area that depicts the probability of an instructor receiving a teaching evaluation of less than or equal to 4.5, and that probability is 0.8176 or 81.76 percent.

Probability > 4.5



Now what will be the probability of receiving a teaching evaluation score of greater than 4.5? In fact, you can see from the next graphic, that the probability is: the reverse of one that we saw earlier, and hence the probability of obtaining a teaching evaluation score of greater than 4.5 is 18.24 percent, which is the area shaded in gray. The reason for this is because the area under the normal distribution curve is equal to 1. So 1 minus 0.8176 will give you the area for evaluation scores greater than 4.5.

Python Syntax

Find the mean and Standard deviation

```
1 eval_mean = round(ratings_df['eval'].mean(), 3)
2 eval_sd = round(ratings_df['eval'].std(), 3)
3 print(eval_mean, eval_sd)
```

```
3.998 0.555
```

Using the norm.cdf package in scipy.stats, find the probability value.

```
1 import scipy.stats
```

```
1 prob0 = scipy.stats.norm.cdf((4.5 - eval_mean)/eval_sd)
2 print(1 - prob0)
```

```
0.1828639734596742
```

Z- scores are always to the left of the curve so we remove from one to get the opposite side

Python Syntax

Find the mean and Standard deviation

```
1 eval_mean = round(ratings_df['eval'].mean(), 3)
2 eval_sd = round(ratings_df['eval'].std(), 3)
3 print(eval_mean, eval_sd)
```

```
3.998 0.555
```

Using the norm.cdf package in scipy.stats, find the probability value.

```
1 import scipy.stats
```

```
1 prob0 = scipy.stats.norm.cdf((4.5 - eval_mean)/eval_sd)
2 print(1 - prob0)
```

```
0.1828639734596742
```

Z- scores are always to the left of the curve so we remove from one to get the opposite side

Let me illustrate the example of getting a teaching evaluation score of greater than 4.5 in Python. We will use the norm.cdf function in the scipy.stats package. After finding the mean and standard deviation, we plug it into the function with the x value of 4.5, and you'll get the area to the left, which is the less than 4.5 area. Because we want the area to the right of 4.5, that is, the probability of greater than 4.5, we will remove the value from 1, as indicated here.

Introduction to Probability Distribution

Estimated time needed: **30** minutes

In this lab, you will familiarize yourself with the normal probability distributions and work on some exercises

Objectives

- Import Libraries
- Introduction to Probability Distributions
    - Normal Distributions
- Lab Exercises

Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[1]:
```
#install specific version of libraries used in lab
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install scipy=1.7.1-y
#!  mamba install matplotlib=3.4.3-y
#!  mamba install statsmodels=0.12.0-y
```

Import the libraries we need for the lab

[2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats
from math import sqrt
```

Read in the csv file from the url using the request library

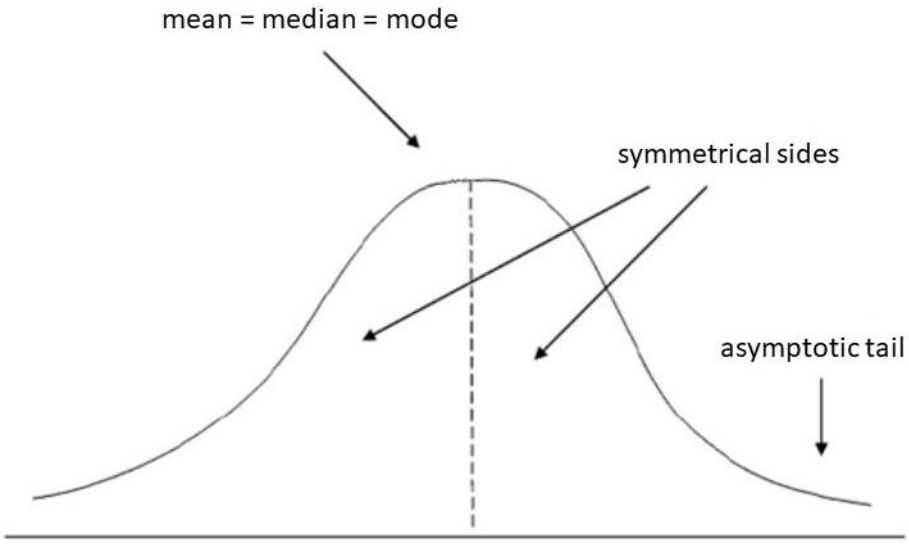
[3]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.csv'
ratings_df = pd.read_csv(ratings_url)
```

Introduction to Probability Distribution

In this section, you will learn how to create the plot distributions using the scipy library in python

Normal Distribution

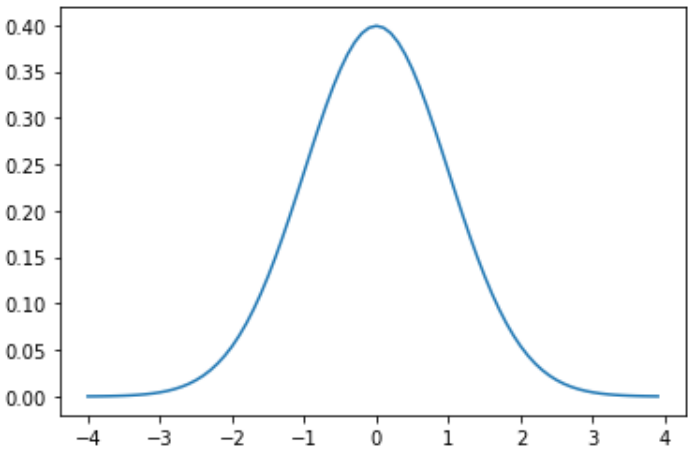
A normal distribution is a bell-shaped density curve described by its mean μ and standard deviation σ. The curve is symmetrical and centered around it's mean. A normal distribution curve looks like this:



We can visualize the curve. Import norm from scipy.stat and plot graph with matplotlib

[4]:
```
from scipy.stats import norm

# Plot between -4 and 4 with 0.1 steps.
x_axis = np.arange(-4, 4, 0.1)
# Mean = 0, SD = 1.
plt.plot(x_axis, norm.pdf(x_axis, 0, 1))
plt.show()
```



Lab Exercises

Using the teachers' rating dataset, what is the probability of receiving an evaluation score of greater than 4.5

Find the mean and standard deviation of teachers' evaluation scores

```
[5]: eval_mean = round(ratings_df['eval'].mean(), 3)
     eval_sd = round(ratings_df['eval'].std(), 3)
     print(eval_mean, eval_sd)

     3.998 0.555
```

Use the scipy.stats module. Because python only looks to the left i.e. less than, we do remove the probability from 1 to get the other side of the tail

Using the teachers' rating dataset, what is the probability of receiving an evaluation score greater than 3.5 and less than 4.2
First we find the probability of getting evaluation scores less than 3.5 using the norm.cdf function

```
[7]: x1 = 3.5
     prob1 = scipy.stats.norm.cdf((x1 - eval_mean)/eval_sd)
     print(prob1)

     0.1847801491443654
```

Then for less than 4.2

```
[8]: x2 = 4.2
     prob2 = scipy.stats.norm.cdf((x2 - eval_mean)/eval_sd)
     print(prob2)

     0.642057540461896
```

The probability of a teacher receiving an evaluation score that is between 3.5 and 4.2 is:

```
[9]: round((prob2 - prob1)*100, 1)

[9]: 45.7
```

Using the two-tailed test from a normal distribution:

A professional basketball team wants to compare its performance with that of players in a regional league.

The pros are known to have a historic mean of 12 points per game with a standard deviation of 5.5.

A group of 36 regional players recorded on average 10.7 points per game.

The pro coach would like to know whether his professional team scores on average are different from that of the regional players.

State the null hypothesis

$H\_0: x=\mu\_1$ ("The mean point of the regional players is not different from the historic mean")

$H\_1: x\neq\mu\_1$ ("The mean point of the regional players is different from the historic mean")

When the population standard deviation is given and we are asked to deal with a sub-sample, the size (n) of the sub-sample is used in the formula:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

```
[10]: ## because it is a two-tailed test we multiply by 2
      2*round(scipy.stats.norm.cdf((10.7 - 12)/(5.5/sqrt(36))), 3)

[10]: 0.156
```

Conclusion: Because the p-value is greater than 0.05, we fail to reject the null hypothesis as there is no sufficient evidence to prove that the mean point of the regional players is different from the historic mean

Practice Questions

Question 1: Using the teachers' rating dataset, what is the probability of receiving an evaluation score greater than 3.3?

[11]:
```
## insert code here
##calculate the probability less than 3.3
prob_less_than = scipy.stats.norm.cdf((3.3 - eval_mean)/eval_sd)
##then remove the probability from 1 to get the area to the right of 3.3
print(1 - prob_less_than)
```

```
0.8957422041794154
```

```
Double-click **here** for the solution.

<!-- The answer is below:
##calculate the probability less than 3.3
prob_less_than = scipy.stats.norm.cdf((3.3 - eval_mean)/eval_sd)
##then remove the probability from 1 to get the area to the right of 3.3
print(1 - prob_less_than)
-->
```

Question 2: Using the teachers' rating dataset, what is the probability of receiving an evaluation score between 2 and 3?

[12]:
```
## insert code here
## find the probablity of reciving a score of less than 2
prob_less_than_2 = scipy.stats.norm.cdf((x1 - eval_mean)/eval_sd)
print(prob_less_than_2)

## find the probablity of reciving a score of less than 3
prob_less_than_3 = scipy.stats.norm.cdf((x2 - eval_mean)/eval_sd)
print(prob_less_than_3)

## remove both probabilities from each other
round((prob_less_than_3 - prob_less_than_2)*100, 1)
```

```
0.1847801491443654
0.642057540461896
```

[12]:
```
45.7
```

```
Double-click **here** for the solution.

<!-- The answer is below:
## find the probablity of reciving a score of less than 2
prob_less_than_2 = scipy.stats.norm.cdf((x1 - eval_mean)/eval_sd)
print(prob_less_than_2)

## find the probablity of reciving a score of less than 3
prob_less_than_3 = scipy.stats.norm.cdf((x2 - eval_mean)/eval_sd)
print(prob_less_than_3)

## remove both probabilities from each other
round((prob_less_than_3 - prob_less_than_2)*100, 1)
```

Question 3: To test the hypothesis that sleeping for at least 8 hours makes one smarter, 12 people who have slept for at least 8 hours every day for the past one year have their IQ tested.

- Here are the results: 116, 111, 101, 120, 99, 94, 106, 115, 107, 101, 110, 92
- Test using the following hypotheses: H0: μ = 100 or Ha: μ > 100

[13]:
```
## insert code here
### remember to remove from 1 because we want the value for when IQs are greater than 100
iqs = [116, 111, 101, 120, 99, 94, 106, 115, 107, 101, 110, 92]
sample_size = len(iqs)
degree_freedom = sample_size - 1
iq_mean = sum(iqs) / sample_size
mean_diff = [(iq - iq_mean) ** 2 for iq in iqs]
iq_std = sqrt(sum(mean_diff) / degree_freedom)
variance = iq_std ** 2
print(f"IQ mean is {iq_mean}, sd is {iq_std}, variance is {variance}")
round(1-scipy.stats.norm.cdf((iq_mean - 100)/(iq_std/sqrt(12))), 3)
```

```
IQ mean is 106.0, sd is 8.831760866327848, variance is 78.00000000000001
```

[13]:
```
0.009
```

Double-click **here** for a hint.

<!-- The hint is below:
```
### find the mean and standard deviation of the 12 IQs
iqs = [116, 111, 101, 120, 99, 94, 106, 115, 107, 101, 110, 92]
sample_size = len(iqs)
degree_freedom = sample_size - 1
iq_mean = sum(iqs) / sample_size
mean_diff = [(iq - iq_mean) ** 2 for iq in iqs]
iq_std = sqrt(sum(mean_diff) / degree_freedom)
variance = iq_std ** 2
```
-->

Double-click **here** for the solution.

<!-- The answer is below:
```
### remember to remove from 1 because we want the value for when IQs are greater than 100
iqs = [116, 111, 101, 120, 99, 94, 106, 115, 107, 101, 110, 92]
sample_size = len(iqs)
degree_freedom = sample_size - 1
iq_mean = sum(iqs) / sample_size
mean_diff = [(iq - iq_mean) ** 2 for iq in iqs]
iq_std = sqrt(sum(mean_diff) / degree_freedom)
variance = iq_std ** 2
print(f"IQ mean is {iq_mean}, sd is {iq_std}, variance is {variance}")
round(1-scipy.stats.norm.cdf((iq_mean - 100)/(iq_std/sqrt(12))), 3)
```
-->

Authors

Aije Egwaikhide is a Data Scientist at IBM who holds a degree in Economics and Statistics from the University of Manitoba and a Post-grad in Business Analytics from St. Lawrence College, Kingston. She is a current employee of IBM where she started as a Junior Data Scientist at the Global Business Services (GBS) in 2018. Her main role was making meaning out of data for their Oil and Gas clients through basic statistics and advanced Machine Learning algorithms. The highlight of her time in GBS was creating a customized end-to-end Machine learning and Statistics solution on optimizing operations in the Oil and Gas wells. She moved to the Cognitive Systems Group as a Senior Data Scientist where she will be providing the team with actionable insights using Data Science techniques and further improve processes through building machine learning solutions. She recently joined the IBM Developer Skills Network group where she brings her real-world experience to the courses she creates.

1.) https://brainly.in/question/15438909

Ravi's score on the test was 180

Given that, the mean score in the test (μ) = 150

Standard deviation (σ) = 20

z-score (Z) = 1.50

We know that,

$X = \mu + Z\sigma$

where Z is Ravi's score on the test

Replacing the values, we get,

$X = 150 + (1.50*20)$

$= 150 + 30$

$= 180$

180 is the required answer.

2.

https://frequentlyaskedquestions.info/if-a-negatively-skewed-distribution-i-e-skewed-to-the-left-has-a-median-of-50-which-of-the-following-statements-are-true-select-all-that-apply/

3.

Question 3

If a negatively skewed distribution (i.e. skewed to the left) has a median of 50, which of the following statements are true? (Select all that apply)

1 / 1 point

☐ None of the above

☑ Mode is greater than 50

Correct! Mean tends to move towards the tail of the data and mode does the opposite

4.

If two fair coins are tossed, what is the probability of getting two heads?

Statistics › Probability › Basic Probability Concepts

$$P(H, H) = \frac{1}{4}$$

Explanation:
There are several possibilities:

tail, tail
tail, head
head, tail
head, head

Each of these four outcomes is equally probable, so each has a 1 in 4 chance. So the probability of getting two heads is:

$$1 \text{ in } 4 = 0.25 = 25\,\% = \frac{1}{4}$$

Probabilities are usually given as fractions.

(Now, had the question been "What is the probability of getting one head and one tail?" - the answer would be $2 \text{ in } 4 = 0.50 = 50\,\%$ or $\frac{2}{4} = \frac{1}{2}$ because there are two ways for the two coins to yield the mixed results.)

6.

https://frequentlyaskedquestions.info/what-is-the-area-under-a-conditional-cumulative-density-function/

6.

Question 6

What is the area under a conditional Cumulative Density Function?

1 / 1 point

○ 0.5

◉ 1

○ 0

○ 2

7. https://frequentlyaskedquestions.info/which-of-the-following-is-a-possible-alternative-hypothesis-h1-for-a-two-tailed-test/

7.

Question 7

Which of the following is a possible alternative hypothesis H1 for a two-tailed test.

1 point

○ **μ is not equal to 85**

○ μ is greater than 85

○ μ is equal to 85

○ μ is less than 85

Quiz: Introduction to Probability Distributions

Bookmark this page

Graded Quiz due May 23, 2022 18:40 +08

Multiple Choice

8/9 points (graded)

A test is administered annually. The test has a mean score of 150 and a standard deviation of 20. If Chioma's z-score is 1.50, what was her score on the test?

- (●) 180
- ( ) 130
- ( ) 30
- ( ) 150

✔

Answer
Correct: Correct!

If a negatively skewed distribution (i.e. skewed to the left) has a median of 50, which of the following statements are true? (Select all that apply)

- [ ] Mean is greater than 50
- [x] Mean is less than 50
- [x] Mode is greater than 50
- [ ] None of the above

✔

Answer
Correct:
Correct!
Correct!

What is the probability of getting two heads when two coins are flipped?

- ○ 1/8
- ◉ 1/4
- ○ 1/2
- ○ 1

✔

Answer
Correct: Correct! Tbe probability of getting two heads is 1/2 × 1/2 (or 0.5 × 0.5) which is 1/4 (or 0.25)

What is the probability of getting two of the same result by rolling TWO six-sided dice (with sides labeled as 1,2,3,4,5,6)?

- ○ 1/36
- ○ 1/6 ✔
- ◉ 2/36
- ○ 1

✖

Answer
Incorrect:
Incorrect! Calculate each individual probability of getting a double. e.g. for 1, 1 - the probability is 1/36. Now calculate the probability of getting a double.

What is the area under a conditional Cumulative Density Function?

- (●) 1
- ( ) 0
- ( ) 2
- ( ) 0.5

✔

Answer
Correct:
Correct! The area under a Cumulative Density Function is calculated by adding the indiviudal probabilities. This sum must always equal to 1.

Which of the following is a possible alternative hypothesis, H1, for a two-tailed test?

- (●) μ is not equal to 85
- ( ) μ is greater than 85
- ( ) μ is less than 85
- ( ) μ is equal to 85

✔

Answer
Correct: Correct!

A normal distribution can best be described by which of the following attributes? (Select all that apply)

- [x] Symmetric
- [ ] Uniform
- [x] Bell-shaped
- [ ] Skewed

Answer
Correct:
Correct!
Correct!

In its standardized form, the normal distribution:

- ( ) Has a mean of 1 and a variance of 0
- (•) Has a mean of 0 and standard deviation of 1
- ( ) Has an area equal to 0.5
- ( ) Cannot be used to approximate discrete probability distributions

Answer
Correct: Correct!

Submit You have used 2 of 2 attempts

Z-Test or T-Test

In traditional hypothesis testing, one has the option to perform a z-test or a t-test and the question is: "Under what circumstances should one perform a z-test or a t-test?" Well, the answer is rather simple. If one is aware of the population's standard deviation or variance, we use the z-test.

And that is when we are comparing the sample mean to a hypothetical or a population mean. And if the sample, the population standard deviation is not known, and we're comparing the sample mean against the population mean with an unknown standard deviation, then we use the t-test. Now there are four scenarios in which we perform these tests.

- First scenario is where we are comparing a sample mean to a population mean and the population standard deviation is known. In that particular case, we use a z-test.
- And in cases where we are comparing a sample mean to a population mean, with an unknown standard deviation, we use a t-test. Now, this I covered earlier

in the last slide.

- The new thing here is that when we compare the means of two independent samples, that is, comparing the means of two independent samples with unequal variances. If we are using, if we are facing with this kind of a question, we use a t-test.
- Again if we are comparing the means of two independent samples, with equal variances, we still use a t-test.

The underlying theory is that when you're using a z-test, you're basing your results on normal distribution. And when you are deploying t-test, you're basing your results on t-distribution. And this could be made - the process of hypothesis testing could be made rather simple - by looking at these thresholds. If you are comparing the means, and,

in this particular case, you are looking at the null being that the two averages are the same, against two averages not being the same, then you're using a two-tailed test. And in that particular case, you're looking for a t-statistic or a z-statistic of 1.96, the absolute value of 1.96. If that were to be the case, you reject the null hypothesis, that is, you're comparing, you're conducting a two-tailed test, you can be using normal distribution or a t-distribution, and you get the calculated z or t-statistics of greater than absolute value of 1.96 and the expected p-value, the probability of that happening would be less than 0.05, and you reject the null [hypothesis]; the null being that the two means are equal. In the case of one-tailed test, where you're testing whether the mean or average of one entity is greater or less than the other, here, the absolute value for z or t-statistic is 1.64. And the probability would still be less than 0.05. If that were to be the case, you reject the null [hypothesis].

If the calculated value for z or t-statistic is less than 1.96, you fail to reject the null hypothesis and the null being the two averages are the same. In case of a one-tailed test and the value calculated value for z or t statistic is less than 1.64, you fail to reject the null [hypothesis] and the null could be that one value, one average, is less than equal to the other, or is greater than the other.

Z Test or T Test

Z test or T test

If the population's standard deviation is known, use z test

Otherwise, use T-test

Comparing means – 4 cases

Comparing sample mean to a population mean when the population standard deviation is known	• Use Z test
Comparing sample mean to a population mean when the population standard deviation is not known	• Use T Test
Comparing the means of two independent samples with unequal variances	• Always use T Test
Comparing the means of two independent samples with equal variances	• Always use T Test

First scenario is where we are comparing a sample mean to a population mean and the population standard deviation is known. In that particular case, we use a z-test. And in cases where we are comparing a sample mean to a population mean, with an unknown standard deviation, we use a t-test. Now, this I covered earlier

in the last slide. The new thing here is that when we compare the means of two independent samples, that is, comparing the means of two independent samples with unequal variances. If we are using, if we are facing with this kind of a question, we use a t-test. Again if we are comparing the means of two independent samples, with equal variances, we still use a t-test. The underlying theory is that when you're using a z-test, you're basing your results on normal distribution. And when you are deploying t-test, you're basing your results on t-distribution.

Rules of Thumb

Type of Test	z or t Statistics*	Expected p-value	Decision
Two-tailed test	The absolute value of the calculated z or t statistics is greater than 1.96	Less than 0.05	Reject the null hypothesis
One-tailed test	The absolute value of the calculated z or t statistics is greater than 1.64	Less than 0.05	Reject the null hypothesis

* In large samples this rule of thumb holds true for the t-test because in large sample sizes, the t-distribution is approximate to a normal distribution

And this could be made - the process of hypothesis testing could be made rather simple - by llooking at these thresholds. If you are comparing the means, and, in this particular case, you are looking at the null being that the two averages are the same, against two averages not being the same, then you're using a two-tailed test. And in that particular case, you're looking for a t-statistic or a z-statistic of 1.96, the absolute value of 1.96. If that were to be the case, you reject the null hypothesis, that is, you're comparing, you're conducting a two-tailed test, you can be using normal distribution or a t-distribution, and you get the calculated z or t-statistics of greater than absolute value of 1.96 and the expected p-value, the probability of that happening would be less than 0.05, and you reject the null [hypothesis]; the null being that the two means are equal. In the case of one-tailed test, where you're testing whether the mean or average of one entity is greater or less than the other, here, the absolute value for z or t-statistic is 1.64. And the probability would still be less than 0.05. If that were to be the case, you reject the null [hypothesis]. If the calculated value for z or t-statistic is less than 1.96, you fail to reject the null hypothesis and the null being the two averages are the same. In case of a one-tailed test and the value calculated value for z or t-statistic is less than 1.64, you fail to reject the null [hypothesis] and the null could be that one value, one average, is less than equal to the other, or is greater than the other.

Dealing with Rejections and Tails

One needs to understand the theory behind the hypothesis testing and how do you reject a null hypothesis or otherwise.

There are rules of thumb, that is, in case of a two-tailed test, one can use 1.96 as the calculated threshold for either z or t statistics to reject the null hypothesis, or for a one-tailed test, the absolute value of 1.64 to reject a null hypothesis.

But what does it mean? How do you get to 1.64 or 1.96?

There is some theory to it. It involves statistical distribution and now perhaps is a good time to learn about those. Now, imagine if the mean values of two variables are the same, that is, we are assuming that the difference between the two means is essentially zero.

Let's say the mean of variable “a” and the mean of “b” we assume that they are equal, that is, mu "a" equals mu "b", and if that were to be the case, the difference between the two should be equal to 0.

So, the alternative hypothesis could be that the difference is mean is not equal to 0, so you would say that mu “a” is not equal to mu “b”, or the difference is greater than 0, that is, mu "a" is greater than mu "b" or the difference is less than 0, where mu “a” is less than MU “b”.

And in these three circumstances, the rejection region, or how do you reject the null hypothesis means three different things, and for this we were to revert to the normal distribution curve.

Imagine that you are conducting a z-test using a normal distribution, and the shape of the curve will be very similar, this image would be very similar, if we were to do a two-tailed test for T distribution.

But let's assume that we're working with normal distribution, and you have a rejection region that is to the left and to the right of the curve. As you saw the normal distribution curve, let’s say our alternative hypothesis is that the mean difference is not equal to 0, it could be greater than or less than 0, but it is not equal to 0. So if we will call this a two-tailed test because we are not making the assumption that the difference is greater than or less than 0.

And then we have to define the rejection region in both tails, that is, the left tail and the right tail of the normal distribution. And remember, we only consider 5% of the area under the normal curve to define the rejection region and for a two-tailed test, that 5% gets divided into half of it goes into the left region, the left tail, and the other half goes into the right tail. So, 2.5% under the curve in each tail. And, graphically you could see this again, the same as we saw earlier that this is a normal distribution curve and 2.5% is in the left tail, and the other 2.5% is in the right tail. And if the test statistic is 1.96; if the absolute value of the test statistics is greater than 1.96 or less than 1.96, it falls in the rejection region and you can safely reject the null [hypothesis], and the null would be the difference of mean is equal to 0 or in common balance; what we are saying, is that the two means are not the same.

Now let us work with the assumption or the situation where we are testing if the difference of mean is less than zero. We are only interested in the left tail. Our alternative hypothesis is that the difference of mean is less than 0. In this case, the entire rejection region, that is, 5% of the rejection region is to the left, and in any situation for a one-tailed test, if we were to get the t-stat of 1.64 or less, we would reject the null [hypothesis]

that the mean is greater than 0, in favor of the alternative that the difference is less than 0. And the exact opposite to this, would be the right-tailed test, where the alternative hypothesis is that the mean is greater than 0 and if you get the t or test statistics of greater than 1.64, or a right-tailed test, you reject the null [hypothesis] in favor of the alternative, that the mean difference is greater than 0.

Dealing With Rejections & Tails

Rules of Thumb

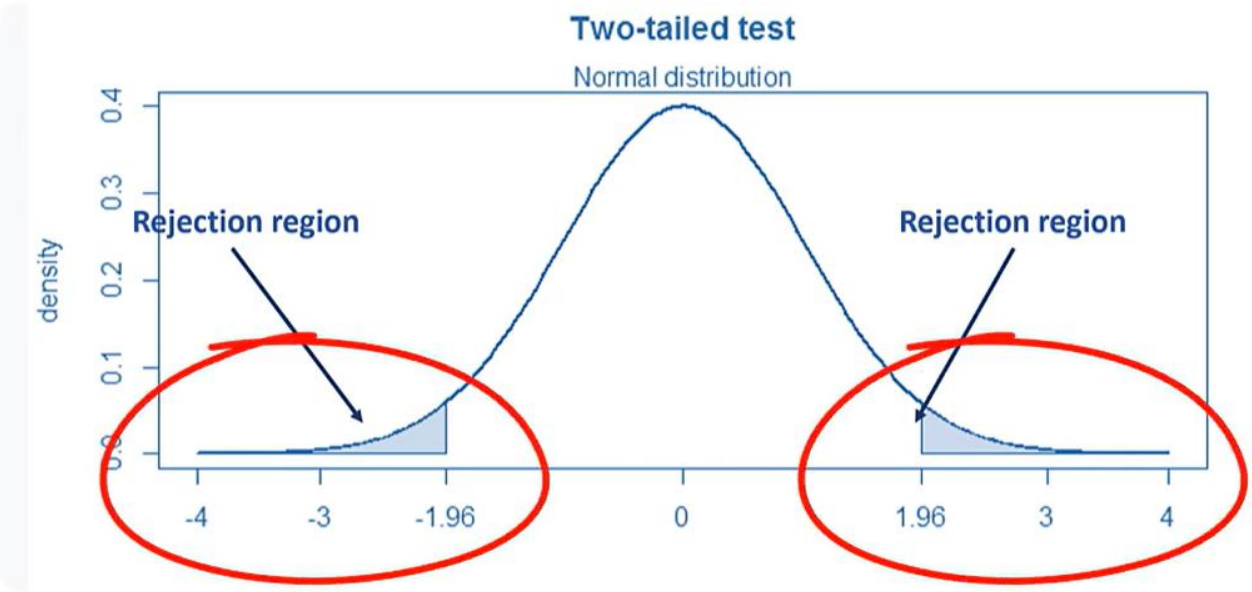
Type of Test	z or t Statistics*	Expected p-value	Decision
Two-tailed test	The absolute value of the calculated z or t statistics is greater than 1.96	Less than 0.05	Reject the null hypothesis
One-tailed test	The absolute value of the calculated z or t statistics is greater than 1.64	Less than 0.05	Reject the null hypothesis

* In large samples this rule of thumb holds true for the t-test because in large sample sizes, the t-distribution is approximate to a normal distribution

The Difference in means equals 0

- If the mean values of two variables is the same, the difference between the two means is essentially 0.
- Mathematically:
  - If $\mu_a = \mu_b$,
  - then $\mu_a - \mu_b = 0$
- Therefore, the alternative hypothesis comes in three flavors:
  - The difference in means is not equal to 0
  - The difference is greater than 0
  - The difference is less than 0.

Normal distribution and rejection regions

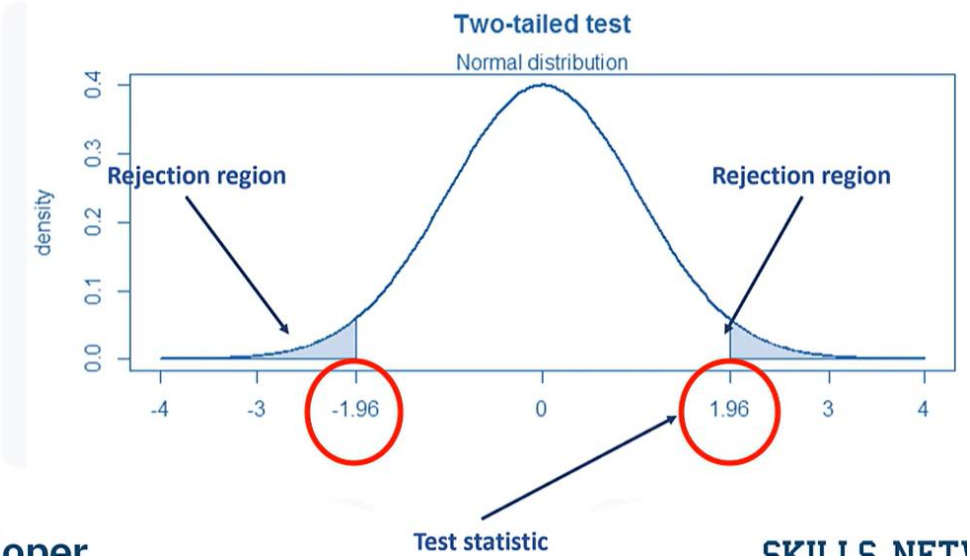


Imagine that you are conducting a z-test using a normal distribution, and the shape of the curve will be very similar, this image would be very similar, if we were to do a two-tailed test for T distribution. But let's assume that we're working with normal distribution, and you have a rejection region that is to the left and to the right of the curve.

Rejection Regions and Rules of Thumb

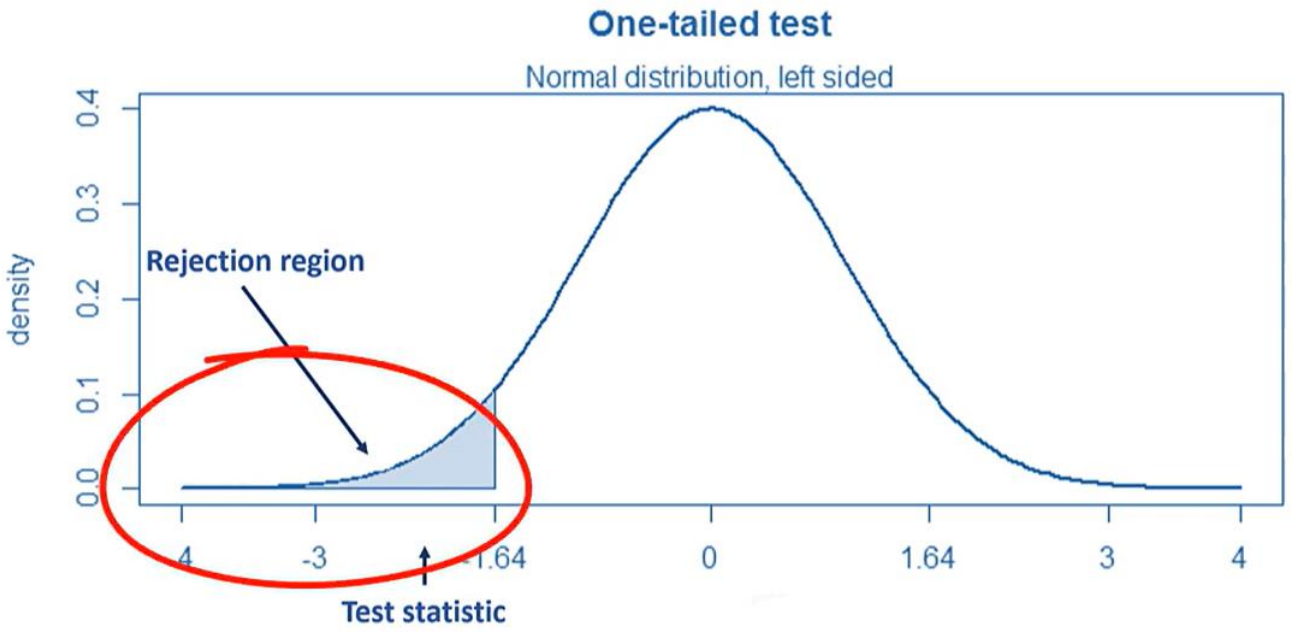
- Alternative hypothesis: mean difference is not equal to 0
  - Difference could be greater or less than zero
  - We call this the two-tailed test
- We will define a rejection region in both tails (left and right) of the Normal distribution
- Remember, we only consider 5% of the area under the Normal curve to define the rejection region
- For a two-tailed test, we divide 5% into two halves and define rejection regions covering 2.5% under the curve in each tail

Normal distribution and rejection regions



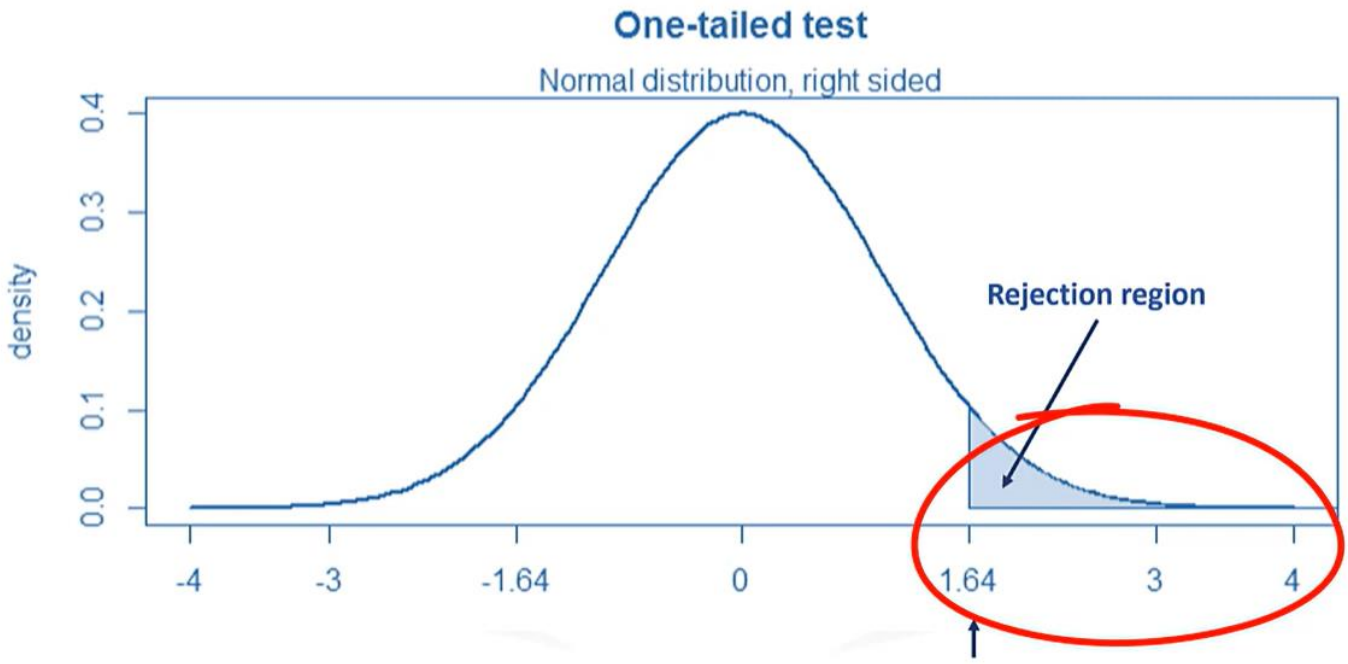
And if the test statistic is 1.96; if the absolute value of the test statistics is greater than 1.96 or less than 1.96, it falls in the rejection region and you can safely reject the null [hypothesis], and the null would be the difference of mean is equal to 0 or in common balance; what we are saying, is that the two means are not the same.

Left tailed test



Now let us work with the assumption or the situation where we are testing if the difference of mean is less than zero. We are only interested in the left tail. Our alternative hypothesis is that the difference of mean is less than 0. In this case, the entire rejection region, that is, 5% of the rejection region is to the left, and in any situation for a one-tailed test, if we were to get the t-stat of 1.64 or less, we would reject the null [hypothesis] that the mean is greater than 0, in favor of the alternative that the difference is less than 0.

Right tailed test



And the exact opposite to this, would be the right-tailed test, where the alternative hypothesis is that the mean is greater than 0 and if you get the t or test statistics of greater than 1.64, or a right-tailed test, you reject the null [hypothesis] in favor of the alternative, that the mean difference is greater than 0.

Equal Versus Unequal Variances

A t-test is the comparison of average values between two groups. For example, you could be comparing whether teaching evaluations of male instructors is the same for female instructors. You can either assume that the variance or standard deviation is equal or unequal.

How do we determine this? We have the teaching evaluation data. We calculated the average teaching evaluation for female instructors to be 3.9, with a standard deviation of 0.53. The average teaching evaluations for male instructors, on the other hand, was calculated as 4.06, with a standard deviation of 0.55. When we conduct a t-test, we are faced with whether to assume equal or unequal variances.

We have a t-test called Levene's test to determine the equality of variances. The null hypothesis of the Levene's test is that population variances are equal. If the p-value of the test is less 0.05, reject the null hypothesis of equal variances and assume that the variances are unequal.

Let us look at the example for that of teaching evaluation scores for male and female instructors. We will use the Levene's function in the scipy.stats package. We will run it against both samples and we will specify the center argument, as means, since our t-test we test for mean differences. We will get a p-value of 0.66, which is greater than 0.05. That means we will fail to reject the null hypothesis and we will assume equal variances when conducting our t-test.

So when you run your t-test, you set the equal_var option to true and if you got a p-value less than 0.05, you set that option to false. If you were to do a t-test by hand, the formula would be different for calculating equal versus unequal variances. You must calculate the pooled variance, as shown here. Then calculate the standard deviation that you will use in the t-test.

If the variances were unequal, calculate the t-test with each of the individual standard deviations and sample size. You will need to find the degree of freedom to check the t-test table. With this formula. The rule of thumb for assuming equal variance when calculating by hand is defined by the ratio of the larger group's variance, to the smaller group's variance to be less than 1.5.

Equal versus Unequal Variances

Variances – equally unequal!

Case Summaries

teaching evaluation

female instructor	N	Mean	Std. Deviation
female	195	3.9010	.53880
male	268	4.0690	.55665
Total	463	3.9983	.55487

Levene's Test

Levene's Test is an inferential statistic to assess the equality of variances.

Null hypothesis:

- Population variances are equal
- If the p-value < 0.05, reject the Null Hypothesis of Equal Variances

Conclusion: Variance does not differ

```
scipy.stats.levene(ratings_df[ratings_df['gender'] == 'female']['eval'],
                   ratings_df[ratings_df['gender'] == 'male']['eval'], center='mean')
```

```
LeveneResult(statistic=0.1903292243529225, pvalue=0.6628469836244741)
```

Variance is equal

```
scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],
                      ratings_df[ratings_df['gender'] == 'male']['eval'], equal_var = True)
```

We will use the Levene's function in the scipy.stats package. We will run it against both samples and we will specify the center argument, as means, since our t-test we test for mean differences. We will get a p-value of 0.66, which is greater than 0.05. That means we will fail to reject the null hypothesis and we will assume equal variances when conducting our t-test. So when you run your t-test, you set the equal_var option to true and if you got a p-value less than 0.05, you set that option to false.

Equal and Unequal Variances

Equal Variances

$$sdev = \sqrt{\frac{vpool * (n_1 + n_2)}{n_1 * n_2}}$$

$$vpool = \frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}$$

$$t = \frac{x_1 - x_2}{sdev}$$

Unequal Variances

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$dof = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

If you were to do a t-test by hand, the formula would be different for calculating equal versus unequal variances. You must calculate the pooled variance, as shown here. Then calculate the standard deviation that you will use in the t-test. If the variances were unequal, calculate the t-test with each of the individual standard deviations and sample size. You will need to find the degree of freedom to check the t-test table. With this formula. The rule of thumb for assuming equal variance when calculating by hand is defined by the ratio of the larger group's variance, to the smaller group's variance to be less than 1.5.

ANOVA

Most of us are familiar with comparing the average values between two groups, for example, the average teaching evaluation score for male instructors compared with that of female instructors and the groups are too, and we know that such comparisons are made using the t-test.

But what if you're dealing with more than two groups; what if there are three, four, or more groups? In that particular case, we would use ANOVA, or analysis of variance, where our intent or goal is to compare the means of more than two groups.

So, in order to do that, in order to accomplish this, we return back to our teaching evaluation data, and in that particular case, we have a variable called "age" where the age of the instructor is recorded in number of years, but we will discretize this age variable, that is, we will create three groups: so instructors who are 40 years and younger, we put them in one group; those between 40 and 56.5 years of age, they are in another group; and those who are 57 years and older, we put them in the 3rd group.

So, you have a set of younger instructors, middle age instructors, and rather slightly older instructors, and the number of observations, of course, is taught by each group is reported under n. And what we also have here, is the the teaching evaluation score for each group, which is not differing much; it's pretty much 4 for each group, and for the older professors, it's slightly less at 3.9

in their respective standard deviations. So we have three groups. Let's say what we are interested in is to determine if their average, of these 3 averages for the 3 respective age categories, if these averages are statistically the same, or they are different. So, we use the one-way analysis of variance, or ANOVA, and using the ANOVA, the F distribution to compare the mean values for more than two groups.

Our null hypothesis is that samples in all groups are drawn from the same population with the same mean values, and we fail to reject the null hypothesis if the the p-value, or the significance for the F test, is greater than .05 and we then infer equal means.

Let's say we are interested in determining if the beauty score for instructors differs by age. We have three groups: younger, middle-aged, and older professors. We have the summary statistics for the standardized beauty scores. We see that there is a difference; as the age goes up, the average value for Beauty score goes down.

So let's run an ANOVA to see if the differences are statistically significant. Our null hypothesis will be: mean beauty scores for instructors do not differ by age, and the alternative hypothesis will be that at least one of the means is different. First, this variable does not exist in out data. We will need to group or bin the continuous age data using the .lock function in Pandas. Then use the f_oneway function in the scipy.stats library to perform the ANOVA test.

We will then print out the F statistics and the p-value. What we can see, is that the p-value is 4.32, times 10, raised to the power of -8, and that is less than 0.05. We will reject the null hypothesis, as there is significant evidence that at least one of the means differ.

If I do the same test for teaching evaluation scores that we observed for the three groups, and we run ANOVA on these three mean values, we find out that the p-value is 0.295, which is greater than 0.05. We will

fail to reject the null [hypothesis] and infer equal means, that is, that the three means are not statistically different. Here we have the analysis of variance performed on two samples: one is the Beauty score. We notice that the difference in means for Beauty scores between the three groups is based on the significance value. This leads us to conclude that at least one mean is different and we reject the null hypothesis that states equal means. And here, because the p-value for teaching evaluation scores between three groups is greater than 0.05, we fail to reject the null hypothesis. We believe that these three means are statistically equal.

ANOVA– Comparing means of more than two groups

use ANOVA, or analysis of variance, where our intent or goal is to compare the means of more than two groups.

Groups of three or more

Let's discretize age:

Evaluation score

	age_group	count	mean	std
0	40 years and younger	113	4.002655	0.505763
1	between 40 and 57 years	228	4.030702	0.537923
2	57 years and older	122	3.933607	0.624250

ANOVA

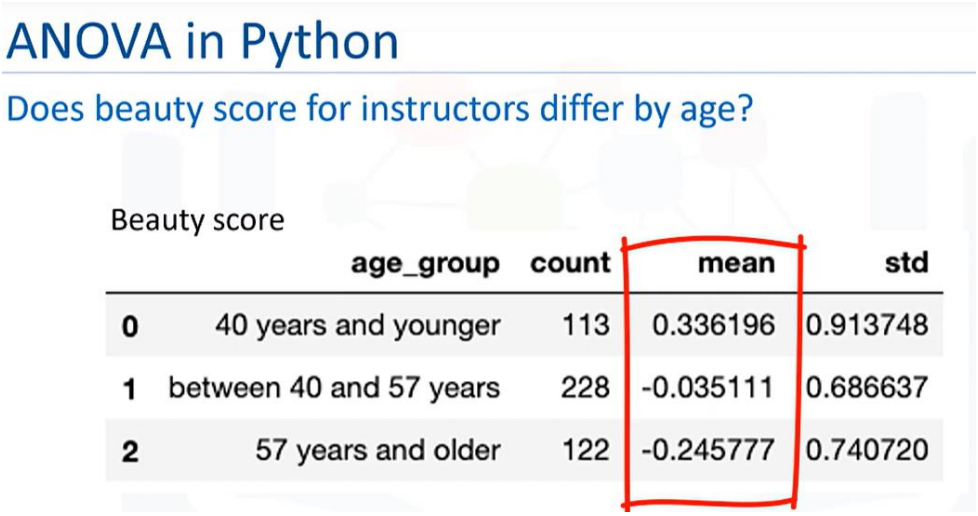
One-way analysis of variance is used to compare means of more than two groups using the F distribution.

Null hypothesis:

- "Samples in all groups are drawn from populations with the same mean values."
- We fail to reject the Null if the p-value of the F-Test > 0.05 and infer equal means.

So, we use the one-way analysis of variance, or ANOVA, and using the ANOVA, the F distribution to compare the mean values for more than two groups. Our null hypothesis is that samples in all groups are drawn from the same population with the same mean

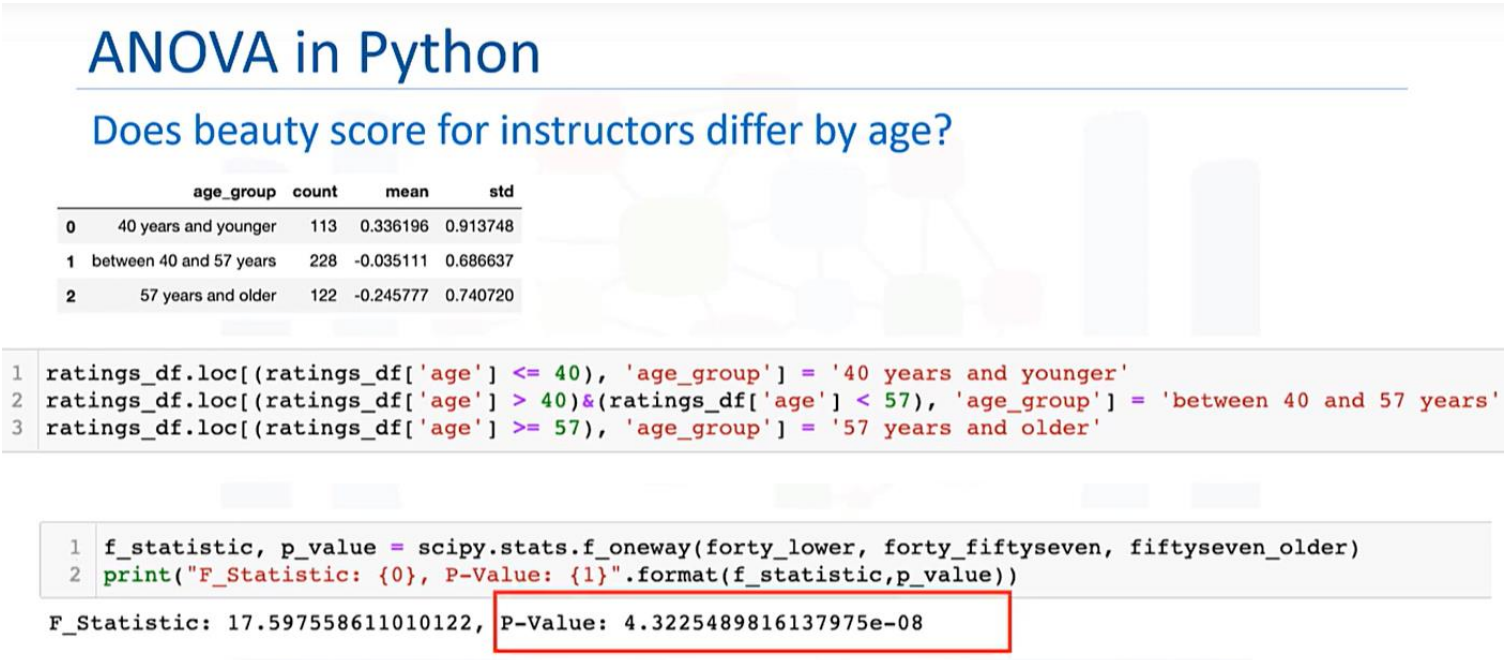
values, and we fail to reject the null hypothesis if the the p-value, or the significance for the F test, is greater than .05 and we then infer equal means.



	age_group	count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	between 40 and 57 years	228	-0.035111	0.686637
2	57 years and older	122	-0.245777	0.740720

Let's say we are interested in determining if the beauty score for instructors differs by age. We have three groups: younger, middle-aged, and older professors. We have the summary statistics for the standardized beauty scores. We see that there is a difference; as the age goes up, the average value for Beauty score goes down.

So let's run an ANOVA to see if the differences are statistically significant. Our null hypothesis will be: mean beauty scores for instructors do not differ by age, and the alternative hypothesis will be that at least one of the means is different.



	age_group	count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	between 40 and 57 years	228	-0.035111	0.686637
2	57 years and older	122	-0.245777	0.740720

```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

First, this variable does not exist in out data. We will need to group or bin the continuous age data using the .lock function in Pandas. Then use the f_oneway function in the scipy.stats library to perform the ANOVA test. We will then print out the F statistics and the p-value. What we can see, is that the p-value is 4.32, times 10, raised to the power of -8, and that is less than 0.05. We will reject the null hypothesis, as there is significant evidence that at least one of the means differ.

ANOVA

Does teaching evaluation score for instructors differ by age?

	age_group	count	mean	std
0	40 years and younger	113	4.002655	0.505763
1	between 40 and 57 years	228	4.030702	0.537923
2	57 years and older	122	3.933607	0.624250

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower_eval, forty_fiftyseven_eval, fiftyseven_older_eval)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 1.2226327996572204, P-Value: 0.29540894225417536
```

If I do the same test for teaching evaluation scores that we observed for the three groups, and we run ANOVA on these three mean values, we find out that the p-value is 0.295, which is greater than 0.05. We will fail to reject the null [hypothesis] and infer equal means, that is, that the three means are not statistically different.

ANOVA in Python

Does beauty score for instructors differ by age?

	age_group	count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	between 40 and 57 years	228	-0.035111	0.686637
2	57 years and older	122	-0.245777	0.740720

```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

Here we have the analysis of variance performed on two samples: one is the Beauty score. We notice that the difference in means for Beauty scores between the three groups is based on the significance value. This leads us to conclude that at least one mean is different and we reject the null hypothesis that states equal means.

ANOVA

Does teaching evaluation score for instructors differ by age?

	age_group	count	mean	std
0	40 years and younger	113	4.002655	0.505763
1	between 40 and 57 years	228	4.030702	0.537923
2	57 years and older	122	3.933607	0.624250

```
1 f_statistic, p_value = scipy.stats.f_oneway(forty_lower_eval, forty_fiftyseven_eval, fiftyseven_older_eval)
2 print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 1.2226327996572204, P-Value: 0.29540894225417536
```

And here, because the p-value for teaching evaluation scores between three groups is greater than 0.05, we fail to reject the null hypothesis. We believe that these three means are statistically equal.

Correlation Tests

Now moving ahead from comparing the average values between two or more groups, we are looking at two variables. We want to know if there is a statistically significant correlation between these two variables and what is needed for this to happen.

We would need to look back to the earlier definition of types of variables. We generally define the variables in two groups: the categorical variables and continuous variables. So if we were to go back to our teaching ratings data, we have instructors who are male and female. Some instructors are visible minorities and some are Caucasian.

So we have two variables: male and female and a visible minority status. These two variables are examples of categorical variables. And if we are comparing or trying to determine the correlation between two categorical variables, we would use the chi-square test.

We would begin with a cross-tabulation between the two values. If we have two continuous variables, for example, the teaching evaluation score and the beauty score of an instructor, then these are two continuous variables, and they can assume any reasonable value within the range.

In this case, we use a Pearson correlation test. We usually begin with a scatter-plot to see what's the nature of the relationship between the two variables. Let us start with categorical variables. We will use the chi-square test for association.

- First we state our hypothesis. We will test the null hypothesis that gender and tenure-ship are

independent against the alternative hypothesis that they are associated.

- Let's begin with a cross tabulation between gender male and female and tenure, that is, tenured profs, then followed by a chi-square test. So, we do the tabulations.
- In the rows we have tenured "no" versus tenured "yes" and female instructors and males. We would like to eyeball these numbers before we turn them into percentages. Looking at instructors who are non-tenured, we notice that 50 of the instructors are female, versus 52 who are male.

But for the instructors who are tenured, 145 of them are female and 216 of them are male. So within the tenured group, we see greater probability for males to be tenured, but in the untenured group, the distribution between males and females looks similar.

Before we go to Python, let's do this by hand to understand the concept.

The formula for chi-square is given as follows: the summation of the observed value, i.e., the counts in each group minus the expected value all squared, divided by the expected value. Expected values are based on the given totals. What would we say each individual value would be if we did not know the observed values? So, to calculate the expected value of untended female instructors, we take the row total, which is 102, multiplied by the column total, 195, divided by the grand total of 463. This will give you 42.96. If we do the same thing for tenured male instructors, we will take the row total, 361, multiplied by the column total, 268, divided by 463, we get 208.96.

If we repeat the same procedure for all of them, we get these values. If we take the row totals, column totals, and grand total, we will get the same values as the totals as the observed values.

Now going back to this formula, if we take a summation of all the observed minus the expected values, all squared, divided by the expected value, we will get a chi-square value of 2.557, and the degree of freedom will be 1. On the chi-square table, we check the degree of freedom equals row 1 and find the value closest to 2.557. Here we can see that 2.557 will most likely fall in between a p-value of 0.1 and 0.25, therefore, we can say that the p-value is greater than 0.1. Since the p-value is greater than .05, we fail to reject the null hypothesis that the two variables are independent and therefore we will conclude that the alternative hypothesis that there is an association between gender and tenure-ship does not exist.

To do this in Python, we will use the chi-square contingency function in the SciPy statistics package, that is, a chi-square test value of 2.557 and the second value is the p-value of about 0.11 and a degree of freedom of 1. If you remember the chi-square table did not give an exact p-value, but a range in which it falls. Python will give the exact p-value.

We can see the same results as on the previous slide. It also prints out the expected values, which we also calculated by hand. Since the p-value is 0.11, which is greater than .05, we fail to reject the null hypothesis that the two variables are independent and therefore we will conclude the alternative hypothesis that there is an association between gender and tenure-ship does not exist. This was an example of testing independence between two categorical variables.

Now to continuous variables. Using a Pearson correlation test from the teaching ratings data, we will test the null hypothesis that there is no correlation between an instructor's beauty score and their teaching evaluation score, against the alternative hypothesis that there is a correlation between both variables. We have the normalized beauty score on the x-axis and the teaching evaluation score on the y-axis. You can eyeball a positive upward sloping curve, but let's run a Pearson correlation test to find out.

We will use the Pearson R package in the scipy.stats package and check for the correlation. We will get a coefficient value of how strong the relationship is and in what direction. Correlation coefficient values lie between -1 and 1, where -1 means a strong negative correlation and visually represented by a downward sloping curve, and 1 means a strong positive relationship and visually represented by an upward sloping curve. In our case, we have a Pearson coefficient of 0.18 and a p-value of 4.25, times 10, raised to power minus 5. Since the p-value is less than the 0.05 we reject the null hypothesis and conclude that there exists a relationship between an instructor's beauty score and teaching evaluation score.

Correlation Tests

Correlations

Types of variables:

- Categorical variables
  - Chi-square test
    - But start with a cross-tab
- Continuous variables
  - Pearson correlation test
    - But start with a scatter plot

We would need to look back to the earlier definition of types of variables. We generally define the variables in two groups: the categorical variables and continuous variables. So if we were to go back to our teaching ratings data, we have instructors who are male and female. Some instructors are visible minorities and some are Caucasian. So we have two variables: male and female and a visible minority status. These two variables are examples of categorical variables. And if we are comparing or trying to determine the correlation between two categorical variables, we would use the chi-square test. We would begin with a cross-tabulation between the two values. If we have two continuous variables, for example, the teaching evaluation score and the beauty score of an instructor, then these are two continuous variables, and they can assume any reasonable value within the range. In this case, we use a Pearson correlation test. We usually begin with a scatter-plot to see what's the nature of the relationship between the two variables.

Categorical variables

To test for relationships between categorical variables:

- We use the Chi-square Test for Association
- State your hypothesis
  - H_0: There is no association between gender and being tenured
  - H_a: There is an association between gender and being tenured

Categorical variables

Is there an association between gender and being tenured?

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Observed value

gender / tenure	female	male	All
no	50	52	102
yes	145	216	361
All	195	268	463

Row total * Column total / Grand total

The formula for chi-square is given as follows: the summation of the observed value, i.e., the counts in each group minus the expected value all squared, divided by the expected value. Expected values are based on the given totals. What would we say each individual value would be if we did not know the observed values? So, to calculate the expected value of untended female instructors, we take the row total, which is 102, multiplied by the column total, 195, divided by the grand total of 463.

Categorical variables

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

gender / tenure	female	male	All
no	50	52	102
yes	145	216	361
All	195	268	463

Observed value

Expected value

tenure / gender	female	male	
no	42.96	59.04	102
yes	152.04	208.96	361
	195	268	

Row total * Column total / Grand total

So, to calculate the expected value of untended female instructors, we take the row total, which is 102, multiplied by the column total, 195, divided by the grand total of 463. This will give you 42.96. If we do the same thing for tenured male instructors, we will take the row total, 361, multiplied by the column total, 268, divided by 463, we get 208.96. If we repeat the same procedure for all of them, we get these values. If we take the row totals, column totals, and grand total, we will get the same values as the totals as the observed values. This will give you 42.96. If we do the same thing for tenured male instructors, we will take the row

total, 361, multiplied by the column total, 268, divided by 463, we get 208.96. If we repeat the same procedure for all of them, we get these values. If we take the row totals, column totals, and grand total, we will get the same values as the totals as the observed values.

Categorical variables

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Degree of freedom = (row-1)*(column-1)

$$\chi_c^2 = 2.557$$

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92

P-value > 0.05, we fail to reject the null hypothesis that the two variables are independent and conclude that a systematic association *does not* exist between gender and tenure.

Now going back to this formula, if we take a summation of all the observed minus the expected values, all squared, divided by the expected value, we will get a chi-square value of 2.557, and the degree of freedom will be 1. On the chi-square table, we check the degree of freedom equals row 1 and find the value closest to 2.557. Here we can see that 2.557 will most likely fall in between a p-value of 0.1 and 0.25, therefore, we can say that the p-value is greater than 0.1. Since the p-value is greater than .05, we fail to reject the null hypothesis that the two variables are independent and therefore we will conclude that the alternative hypothesis that there is an association between gender and tenure-ship does not exist.

Categorical variables

gender	female	male
tenure		
no	50	52
yes	145	216

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

```
1  scipy.stats.chi2_contingency(cont_table, correction = False)
```

```
(2.557051129789522,
 0.10980322511302845,
 1,
 array([[ 42.95896328,  59.04103672],
        [152.04103672, 208.95896328]]))
```

P-value of 0.109 > 0.05, we fail to reject the null hypothesis that the two variables are independent and conclude that a systematic association *does not* exist between gender and tenure.

To do this in Python, we will use the chi-square contingency function in the SciPy statistics package, that is, a chi-square test value of 2.557 and the second value is the p-value of about 0.11 and a degree of freedom of 1. If you remember the chi-square table did not give an exact p-value, but a range in which it falls. Python will give the exact p-value.

We can see the same results as on the previous slide. It also prints out the expected values, which we also calculated by hand. Since the p-value is 0.11, which is greater than .05, we fail to reject the null hypothesis that the two variables are independent and therefore we will conclude the alternative hypothesis that there is an association between gender and tenure-ship does not exist. This was an example of testing independence between two categorical variables.

Continuous variables

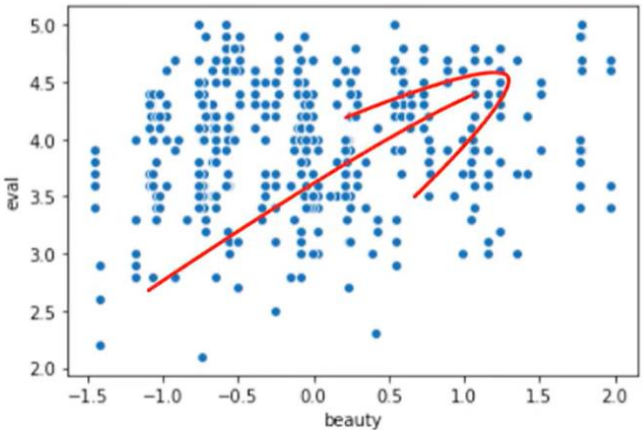
To test for relationships between continuous variables:

- We use the Pearson Correlation Test
- State your hypothesis
  - H_0: There is no correlation between an instructor's beauty score and their teaching evaluation score.
  - H_a: There is a correlation between an instructor's beauty score and their teaching evaluation score.

Now to continuous variables. Using a Pearson correlation test from the teaching ratings data, we will test the null hypothesis that there is no correlation between an instructor's beauty score and their teaching evaluation score, against the alternative hypothesis that there is a correlation between both variables.

Continuous variables

Is teaching evaluation score correlated with beauty score?



We have the normalized beauty score on the x-axis and the teaching evaluation score on the y-axis. You can eyeball a positive upward sloping curve, but let's run a Pearson correlation test to find out.

Pearson Correlation test

```
1 scipy.stats.pearsonr(ratings_df['beauty'], ratings_df['eval'])
```

```
(0.1890390908404521, 4.247115419812614e-05)
```

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)} \cdot \sqrt{\operatorname{var}(y)}}$$

Null Hypothesis: There is no association between an instructor's looks and teaching evaluation score.

Since the p-value (Sig. (2-tailed) < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between beauty and teaching evaluation score.

Correlation coefficient varies between -1 and 1.

We will use the Pearson R package in the scipy.stats package and check for the correlation. We will get a coefficient value of how strong the relationship is and in what direction. Correlation coefficient values lie between -1 and 1, where -1 means a strong negative correlation and visually represented by a downward sloping curve, and 1 means a strong positive relationship and visually represented by an upward sloping curve. In our case, we have a Pearson coefficient of 0.18 and a p-value of 4.25, times 10, raised to power minus 5. Since the p-value is less than the 0.05 we reject the null hypothesis and conclude that there exists a relationship between an instructor's beauty score and teaching evaluation score.

Hypothesis Testing

Estimated time needed: **30** minutes

The goal of hypothesis testing is to answer the question, “Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?”

1. The first step is to quantify the size of the apparent effect by choosing a test statistic (t-test, ANOVA, etc).
2. The next step is to define a null hypothesis, which is a model of the system based on the assumption that the apparent effect is not real.
3. Then compute the p-value, which is the probability of the null hypothesis being true, and finally interpret the result of the p-value, if the value is low, the effect is said to be statistically significant, which means that the null hypothesis may not be accurate.

Objectives

- Import Libraries
- Lab exercises
    - Stating the hypothesis
    - Levene's Test for equality
    - Preparing your data for hypothesis testing
- Quiz

Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[1]:
```
#install specific version of libraries used in lab
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install scipy=1.7.1-y
#!  mamba install seaborn=0.9.0-y
#!  mamba install matplotlib=3.4.3-y
#!  mamba install statsmodels=0.12.0-y
```

Import the libraries we need for the lab

[2]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats
```

Read in the csv file from the URL using the request library

[3]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.csv'
ratings_df = pd.read_csv(ratings_url)
```

Lab Exercises

T-Test: Using the teachers' rating data set, does gender affect teaching evaluation rates?

We will be using the t-test for independent samples. For the independent t-test, the following assumptions must be met.

- One independent, categorical variable with two levels or group
- One dependent continuous variable
- Independence of the observations. Each subject should belong to only one group. There is no relationship between the observations in each group.
- The dependent variable must follow a normal distribution
- Assumption of homogeneity of variance

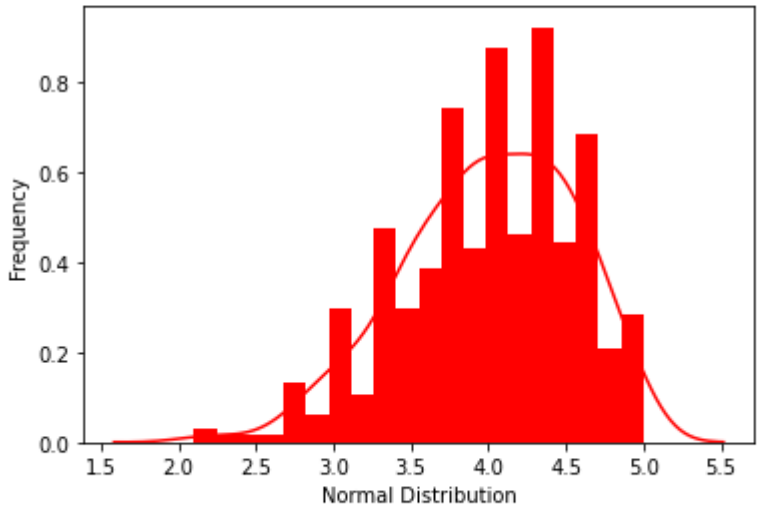
State the hypothesis

- H_0:μ_1=μ_2H_0:μ_1=μ_2 ("there is no difference in evaluation scores between male and females")
- H_1:μ_1≠μ_2H_1:μ_1≠μ_2 ("there is a difference in evaluation scores between male and females")

We can plot the dependent variable with a historgram

```
[4]: ax = sns.distplot(ratings_df['eval'],
                       bins=20,
                       kde=True,
                       color='red',
                       hist_kws={"linewidth": 15,'alpha':1})
     ax.set(xlabel='Normal Distribution', ylabel='Frequency')
     ## we can assume it is normal
```

```
[4]: [Text(0.5, 0, 'Normal Distribution'), Text(0, 0.5, 'Frequency')]
```



We can use the Levene's Test in Python to check test significance

```
[5]: scipy.stats.levene(ratings_df[ratings_df['gender'] == 'female']['eval'],
                       ratings_df[ratings_df['gender'] == 'male']['eval'], center='mean')

     # since the p-value is greater than 0.05 we can assume equality of variance
```

```
[5]: LeveneResult(statistic=0.19032922435292574, pvalue=0.6628469836244741)
```

Use the `ttest_ind` from the `scipy_stats` library

```
[6]: scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],
                           ratings_df[ratings_df['gender'] == 'male']['eval'], equal_var = True)
```

```
[6]: Ttest_indResult(statistic=-3.249937943510772, pvalue=0.0012387609449522217)
```

Conclusion: Since the p-value is less than alpha value 0.05, we reject the null hypothesis as there is enough proof that there is a statistical difference in teaching evaluations based on gender

ANOVA: Using the teachers' rating data set, does beauty score for instructors differ by age?

First, we group the data into cateries as the one-way ANOVA can't work with continuous variable - using the example from the video, we will create a new column for this newly assigned group our categories will be teachers that are:

- 40 years and younger
- between 40 and 57 years
- 57 years and older

[7]:
```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

State the hypothesis

- $H\_0 : \mu\_1 = \mu\_2 = \mu\_3$ (the three population means are equal)
- H_1 : At least one of the means differ

Test for equality of variance

[8]:
```
scipy.stats.levene(ratings_df[ratings_df['age_group'] == '40 years and younger']['beauty'],
                   ratings_df[ratings_df['age_group'] == 'between 40 and 57 years']['beauty'],
                   ratings_df[ratings_df['age_group'] == '57 years and older']['beauty'],
                   center='mean')
# since the p-value is less than 0.05, the variance are not equal, for the purposes of this exercise, we will move along
```

[8]:
```
LeveneResult(statistic=8.60005668392584, pvalue=0.000215366180993476)
```

First, separate the three samples (one for each job category) into a variable each.

[9]:
```
forty_lower = ratings_df[ratings_df['age_group'] == '40 years and younger']['beauty']
forty_fiftyseven = ratings_df[ratings_df['age_group'] == 'between 40 and 57 years']['beauty']
fiftyseven_older = ratings_df[ratings_df['age_group'] == '57 years and older']['beauty']
```

Now, run a one-way ANOVA.

[10]:
```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

Conclusion: Since the p-value is less than 0.05, we will reject the null hypothesis as there is significant evidence that at least one of the means differ.

ANOVA: Using the teachers' rating data set, does teaching evaluation score for instructors differ by age?

Test for equality of variance

```
[11]: scipy.stats.levene(ratings_df[ratings_df['age_group'] == '40 years and younger']['eval'],
                         ratings_df[ratings_df['age_group'] == 'between 40 and 57 years']['eval'],
                         ratings_df[ratings_df['age_group'] == '57 years and older']['eval'],
                         center='mean')
```

```
[11]: LeveneResult(statistic=3.820237661494229, pvalue=0.022621418520219 39)
```

```
[12]: forty_lower_eval = ratings_df[ratings_df['age_group'] == '40 years and younger']['eval']
      forty_fiftyseven_eval = ratings_df[ratings_df['age_group'] == 'between 40 and 57 years']['eval']
      fiftyseven_older_eval = ratings_df[ratings_df['age_group'] == '57 years and older']['eval']
```

```
[13]: f_statistic, p_value = scipy.stats.f_oneway(forty_lower_eval, forty_fiftyseven_eval, fiftyseven_older_eval)
      print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 1.2226327996572206, P-Value: 0.29540894225417536
```

Conclusion: Since the p-value is greater than 0.05, we will fail to reject the null hypothesis as there is no significant evidence that at least one of the means differ.

Chi-square: Using the teachers' rating data set, is there an association between tenure and gender?

State the hypothesis:

- H_0:H_0: The proportion of teachers who are tenured is independent of gender
- H_1:H_1: The proportion of teachers who are tenured is associated with gender

Create a Cross-tab table

```
[14]: cont_table = pd.crosstab(ratings_df['tenure'], ratings_df['gender'])
      cont_table
```

[14]:

gender	female	male
tenure		
no	50	52
yes	145	216

Use the scipy.stats library and set correction equals False as that will be the same answer when done by hand, it returns: χ2 value, p-value, degree of freedom, and expected values.

```
[15]: scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
[15]: (2.20678166999886,
       0.1374050603563787,
       1,
       array([[ 42.95896328,  59.04103672],
              [152.04103672, 208.95896328]]))
```

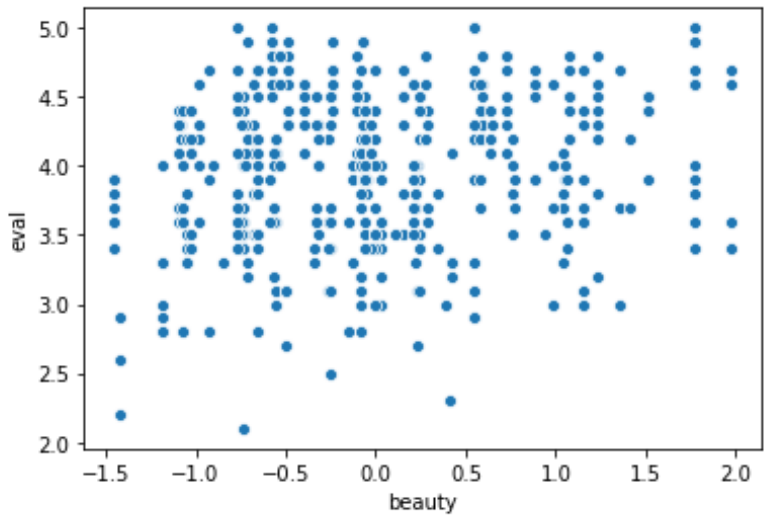
Correlation: Using the teachers rating dataset, Is teaching evaluation score correlated with beauty score?

State the hypothesis:

- H_0:H_0: Teaching evaluation score is not correlated with beauty score
- H_1:H_1: Teaching evaluation score is correlated with beauty score

Since they are both continuous variables we can use a Pearson correlation test and draw a scatter plot

```
[16]: ax = sns.scatterplot(x="beauty", y="eval", data=ratings_df)
```




```
[ ]: scipy.stats.pearsonr(ratings_df['beauty'], ratings_df['eval'])
```

Conclusion: Since the p-value (Sig. (2-tailed) < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between beauty and teaching evaluation score.

Practice Questions

Question 1: Using the teachers rating data set, does tenure affect teaching evaluation scores?

- Use α = 0.05

```
[18]: ## insert code here
      scipy.stats.ttest_ind(ratings_df[ratings_df['tenure'] == 'yes']['eval'],
                            ratings_df[ratings_df['tenure'] == 'no']['eval'], equal_var = True)
```

```
[18]: Ttest_indResult(statistic=-2.8046798258451777, pvalue=0.005249471210198792)
```

```
Double-click **here** for the solution.

<!-- The answer is below:
scipy.stats.ttest_ind(ratings_df[ratings_df['tenure'] == 'yes']['eval'],
                      ratings_df[ratings_df['tenure'] == 'no']['eval'], equal_var = True)
```

The p-value is less than 0.05 that means that - we will reject the null hypothesis as there evidence that being tenured affects teaching evaluation scores.

Question 2: Using the teachers rating data set, is there an association between age and tenure?

- Discretize the age into three groups 40 years and youngers, between 40 and 57 years, 57 years and older (This has already been done for you above.)
- What is your conclusion at α = 0.01 and α = 0.05?

```
[19]: ## insert code here
      scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
[19]: (2.20678166999886,
       0.1374050603563787,
       1,
       array([[ 42.95896328,  59.04103672],
              [152.04103672, 208.95896328]]))
```

```
Double-click **here** for a hint.

<!-- The hint is below:
## state your hypothesis
Null Hypothesis: There is no association between age and tenure
Alternative Hypothesis: There is an association between age and tenure

## don't forget to create a cross tab of the data
cont_table  = pd.crosstab(ratings_df['tenure'], ratings_df['age_group'])
```

```
Double-click **here** for the solution.

<!-- The answer is below:
## use the chi-square function
scipy.stats.chi2_contingency(cont_table, correction = True)
```

At the α = 0.01, p-value is greater, we fail to reject null hypothesis as there is no evidence of an association between age and tenure

At the α = 0.05, p-value is less, we reject null hypoothesis as there is evidence of an association between age and tenure

Question 3: Test for equality of variance for beauty scores between tenured and non-tenured instructors

- Use α = 0.05

```
[20]: ## insert code here
scipy.stats.levene(ratings_df[ratings_df['tenure'] == 'yes']['beauty'],
                   ratings_df[ratings_df['tenure'] == 'no']['beauty'],
                   center='mean')
```

```
[20]: LeveneResult(statistic=0.48842416527504556, pvalue=0.4849835158609811)
```

```
Double-click **here** for the solution.

<!-- The answer is below:
### use the levene function to find the p-value and conclusion
scipy.stats.levene(ratings_df[ratings_df['tenure'] == 'yes']['beauty'],
                   ratings_df[ratings_df['tenure'] == 'no']['beauty'],
                   center='mean')
```

Since the p-value is greater than 0.05, we will assume equality of variance of both groups

Question 4: Using the teachers rating data set, is there an association between visible minorities and tenure?

- Use α = 0.05

```
[21]: ## insert code here
      scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
[21]: (2.20678166999886,
       0.1374050603563787,
       1,
       array([[ 42.95896328,  59.04103672],
              [152.04103672, 208.95896328]]))
```

```
Double-click **here** for a hint.

<!-- The hint is below:
##State you hypothesis and Create a cross-tab:
Null Hypothesis: There is no association between visible minorities and tenure
Alternative Hypothesis: There is an association between visible minorities and tenure

cont_table  = pd.crosstab(ratings_df['vismin'], ratings_df['tenure'])
```

```
Double-click **here** for the solution.

<!-- The answer is below:
## run the chi2_contingency() on the contigency table
scipy.stats.chi2_contingency(cont_table, correction = True)
```

Since the p-value is greater than 0.05, we fail to reject null hypothesis as there is evidence of an association between visible minorities and tenure

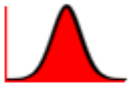
Consider a normally distributed data set with mean μ = 63.18 inches and standard deviation σ= 13.27 inches. What is the z-score when x = 91.54 inches? (To 3 decimal places)

Result

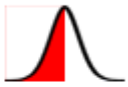
Z-score = -4,742.00904

Probability of x<91.54: 0

Probability of x>91.54: 1



Probability of 91.54<x<63018: 0.5



Steps:

$$\text{Z score} = \frac{x - \mu}{\sigma}$$

$$= \frac{91.54 - 63018}{13.27}$$

$$= -4{,}742.00904$$

P-value from Z-Table:

P(x<91.54) = 0

P(x>91.54) = 1 - P(x<91.54) = 1

P(91.54<x<63018) = 0.5 - P(x<91.54) = 0.5

A room in a laboratory is only considered safe if the mean radiation level is 400 or less. When a sample of 10 radiation measurements were taken, the mean value of the radiation was 414 with a standard deviation of 17. There are concerns that mean radiation is above 414. Radiation levels in the lab are known to follow a normal distribution with standard deviation 22. We will like to conduct a hypothesis test at the 5% level of significance to determine whether there is evidence that the laboratory is unsafe.

What will be the appropriate test?

1 / 1 point

- (●) z-test
- ( ) t-test
- ( ) ANOVA
- ( ) Chi-square

Correct! We use a z-test when the population standard deviation is known

The mineral content of a particular brand of supplement pills is normally distributed with mean 490 mg and variance of 400. What is the probability that a randomly selected pill contains at least 500 mg of minerals?

Question 6

The mineral content of a particular brand of supplement pills is normally distributed with mean 490 mg and variance of 400. What is the probability that a randomly selected pill contains at least 500 mg of minerals?

1 / 1 point

- (●) 0.3085
- ( ) 0.2023
- ( ) 0.0525
- ( ) 0.7967

Question 7

The P-value for a normally distributed right-tailed test is P=0.042. Which of the following is **INCORRECT**?

1 / 1 point

- ( ) The P-value for a two-tailed test based on the same sample would be P=0.084
- (●) The P-value for a left-tailed test based on the same sample would be P= -0.042
- ( ) The z-score test statistic is approximately z=1.73
- ( ) We will reject H0 at α=0.05, but not at α=0.01

Question 8

The time X taken by a cashier in a grocery store express lane to complete a transaction follows a normal distribution with mean 90 seconds and standard deviation 20 seconds.What is the first quartile of the distribution of (in seconds)?

1 / 1 point

- 88.0
- 76.6
- 73.8
- 81.2

9.

Question 9

A man accused of committing a crime is taking a polygraph (lie detector) test. The polygraph is essentially testing the hypotheses

H0: The man is telling the truth vs. Ha: The man is not telling the truth.

Suppose we use a 5% level of significance. Based on the man's responses to the questions asked, the polygraph determines a P-value of 0.08. We conclude that:

1 / 1 point

○ The probability that the man is telling the truth is 0.08.

◉ We fail to reject the null hypothesis as there is insufficient evidence that the man is not telling the truth.

○ The probability that the man is not telling the truth is 0.08.

○ We reject the null hypothesis as there is sufficient evidence that the man is telling the truth.

Question 10

The average hourly wage at a fast-food restaurant is $5.85 with a standard deviation of $0.35. Assume that the wages are normally distributed. The probability that a selected worker earns more than $6.90 is

1 / 1 point

○ 0

○ 0.4987

◉ 0.0013

○ 0.9987

Quiz: Hypothesis Testing

Bookmarked

Graded Quiz due Jun 13, 2022 19:31 +08

Quiz: Hypothesis Testing

9/9 points (graded)

Using the teacher's rating data, is there an association between native (native English speakers) and the number of credits taught? What test will you use?

- (●) Chi-Square test for Association
- ( ) ANOVA
- ( ) T-test
- ( ) Z-test

✔

Answer
Correct: Correct!

If I wanted to use the Chi-Square test to determine where there is an association between gender (Male or Female) and tenure-ship (tenured or not tenured), what will be my degree of freedom?

1 ✔

1

Answer
Correct: Formula for degree of freedom for Chi-Squre is (r-1)*(c-1)

Consider a normally distributed data set with mean μ = 63.18 inches and standard deviation σ= 13.27 inches. What is the z-score when x = 91.54 inches? (To 3 decimal places)

2.137 ✔

2.137

A room in a laboratory is only considered safe if the mean radiation level is 400 or less. When a sample of 10 radiation measurements were taken, the mean value of the radiation was 414 with a standard deviation of 17. There are concerns that mean radiation is above 414. Radiation levels in the lab are known to follow a normal distribution with standard deviation 22. We will like to conduct a hypothesis test at the 5% level of significance to determine whether there is evidence that the laboratory is unsafe. What will be the appropriate test?

- ○ t-test
- ○ ANOVA
- ○ Chi-square
- ◉ z-test

✔

Answer
Correct: Correct! We use a z-test when the population standard deviation is known.

The mineral content of a particular brand of supplement pills is normally distributed with mean 490 mg and variance of 400. What is the probability that a randomly selected pill contains at least 500 mg of minerals?

- ◉ 0.3085
- ○ 0.2023
- ○ 0.0525
- ○ 0.7967

✔

Answer
Correct: Correct!

The P-value for a normally distributed right-tailed test is P=0.042. Which of the following is INCORRECT?

- ○ The P-value for a two-tailed test based on the same sample would be P=0.084
- ◉ The P-value for a left-tailed test based on the same sample would be P= -0.042
- ○ The z-score test statistic is approximately z=1.73
- ○ We will reject H0 at α=0.05, but not at α=0.01

✔

Answer
Correct: Correct! P-values are proportion and range from 0 to 1. The left-tail test for this will also be 0.042

The time X taken by a cashier in a grocery store express lane to complete a transaction follows a normal distribution with mean 90 seconds and standard deviation 20 seconds.What is the first quartile of the distribution of X (in seconds)?

- (●) 76.6
- ( ) 81.2
- ( ) 88.0
- ( ) 73.8

✔

Answer
Correct: Correct!

A man accused of committing a crime is taking a polygraph (lie detector) test. The polygraph is essentially testing the hypotheses H0: The man is telling the truth vs. Ha: The man is not telling the truth. Suppose we use a 5% level of significance. Based on the man's responses to the questions asked, the polygraph determines a P-value of 0.08. We conclude that:

- ( ) The probability that the man is telling the truth is 0.08
- ( ) The probability that the man is not telling the truth is 0.08.
- ( ) We reject the null hypothesis as there is sufficient evidence that the man is telling the truth.
- (●) We fail to reject the null hypothesis as there is insufficient evidence that the man is not telling the truth.

✔

Answer
Correct: Correct! p-value is greater than 0.05

The average hourly wage at a fast-food restaurant is $5.85 with a standard deviation of $0.35. Assume that the wages are normally distributed. The probability that a selected worker earns more than $6.90 is:

- (●) 0.0013
- ( ) 0.4987
- ( ) 0.9987
- ( ) 0

✔

Answer
Correct: Correct!

Regression: The Workhorse of Statistical Analysis

In this video, we will introduce the fundamentals of **regression analysis**, which we believe is the workhorse of statistical analysis. Now, in terms of hypothesis testing, these tests measure the strength of relationship between two or more variables; and you have to run them independently, but if you know how to run regression, we say that as a practical data scientist you can forego these tests and go straight to regression, which is available in most spreadsheets and, also, in all statistical software.

So, here are the fundamentals, the basics, of regression model.

1. First of all, you need a question to answer using the regression model; for instance, do male instructors get higher teaching evaluations than female instructors?
2. Or, does the beauty score decrease with the age of the individual instructor?
3. Or, is there an association between an instructor's looks and the teaching evaluation score they receive? Do good looking professors get higher teaching evaluation scores?

So, with these questions in mind, we focus now on the terminology of regression model. So, there are **two types of regression variables** that we use: **1.** one is a dependent variable, that is, the variable that we are really interested in, for example, the teaching evaluation score of an individual instructor; and, the explanatory variables that explain the variance or differences or values of the dependent variable.

So, for example, teaching evaluation score could be explained by the looks or the gender or the English language proficiency of an individual instructor. So, you have two types of variables: dependent and explanatory.

Now, let's look at the notation for a regression model. The dependent variable is donated as y, so this y, would be the teaching evaluation score; and the explanatory variables are denoted as x's, so beauty, the gender, and the English language proficiency would be an x, and the underlying assumption is that y is explained by x, that is, teaching evaluation score; y is explained by x, that is, the beauty score, or y is a function of x, which we write as y equal to function of x, that is, the teaching evaluation score is some function of beauty.

Statistically, if you run an estimate, a regression model, y is equal to some constant, and then a weighting factor for the variable x, if it's a beauty score, then the weighting factor for the building score, and the error term. An error term is whatever we cannot explain by the model; that goes into error term; and I will explain this a little more in a minute. So y is equal to, let's say the constant is beta naught, plus some factor or weight, which is beta1 for x, plus the error term, which we represent as epsilon.

And then, if you are familiar with your basic statistics text, if there are more than one variables, then y is equal to beta naught, that's the constant, plus beta1, x1, beta1 is the factor for one variable that could be beauty, plus beta2, the other weight explaining another x2, another variable, such as, English language proficiency. So the weight for English language proficiency would be beta2 and plus the epsilon, which is the error term explaining or capturing whatever the model couldn't capture. If I had estimated a regression model, using the dataset, teaching evaluations dataset, it would look like the following.

It will be the teaching evaluation score of an individual instructor, is equal to some constant, plus the weight for the beauty variable, and then times the beauty score, plus the error. So here, teaching evaluation score is equal to, according to the model, 3.998, that's the constant, plus the weight for beauty score, which is 0.133, plus the error. And the error is epsilon, which is essentially the difference between the actual teaching evaluation score, that we have recorded in the dataset, and the one that we have forecasted using this model. So the difference between the actual values and the forecasted values is the error term.

Hypothesis Testing

Regression: The ultimate tool for hypothesis testing

It can, arguably, replace

- T-test
- ANOVA
- Pearson Correlation tests

Regression is available in most spreadsheets and all stats software

Fundamentals of Regression Model

Introduction to Regression Models

The Basics:

- We need a question. For example:
  - Do male instructors get higher teaching evaluations than female instructors?
  - Does beauty score decrease with age?
  - Is there an association between an instructor's looks and teaching evaluation score?

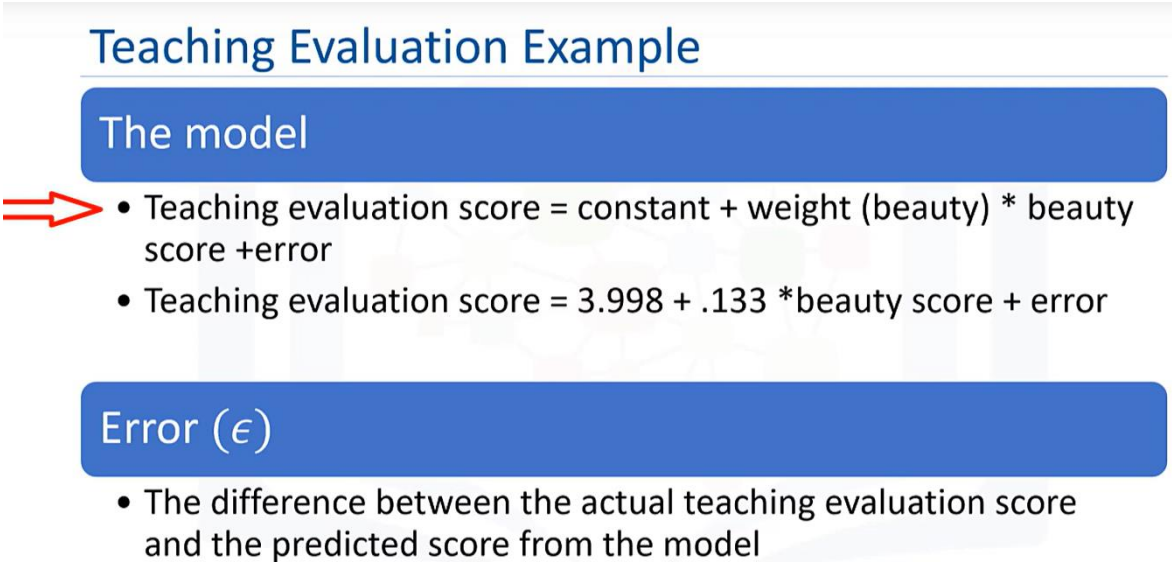
Terminology

Dependent variable	Explanatory variables

• The variable we are primarily interested in
  • Teaching Evaluation Score
	• Variables that influence the dependent variable
	  • Beauty
	  • Gender
	  • English proficiency

Notation

- Dependent variable is denoted as y
- Explanatory variables are denoted as x
- y is explained by x or y is a function of x
- Mathematically:
  - $y = f(x)$
- Statistically:
  - $y = constant + weight_x(x) + error$
  - $y = \beta_0 + \beta_1 x + \epsilon$
  - If there are more than one explanatory variables:
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$



If I had estimated a regression model, using the dataset, teaching evaluations dataset, it would look like the following. It will be the teaching evaluation score of an individual instructor, is equal to some constant, plus the weight for the beauty variable, and then times the beauty score, plus the error. So here, teaching evaluation score is equal to, according to the model, 3.998, that's the constant, plus the weight for beauty score, which is 0.133, plus the error. And the error is epsilon, which is essentially the difference between the actual teaching evaluation score, that we have recorded in the dataset, and the one that we have forecasted using this model. So the difference between the actual values and the forecasted values is the error term.

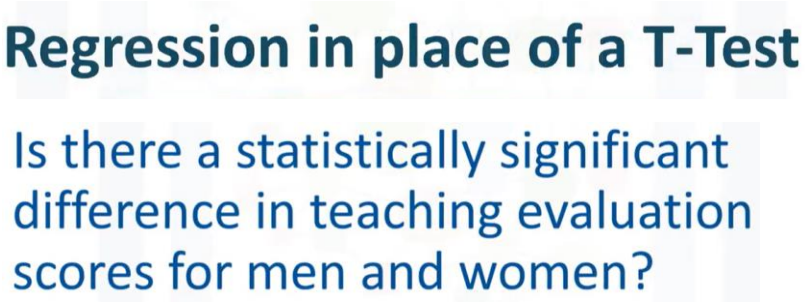
Regression in Place of T-Test

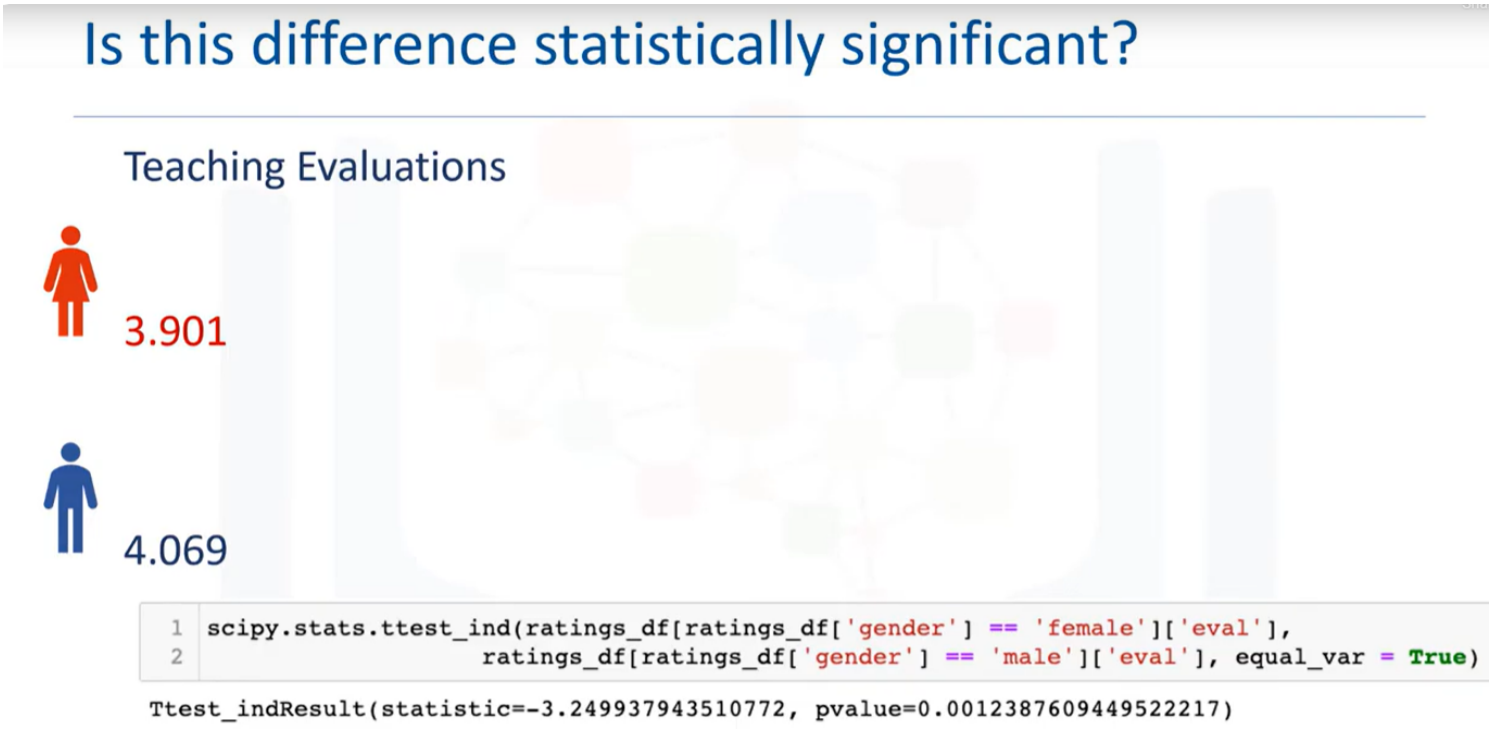
In this video, we will illustrate how to use regression analysis in place of a t-test. We will begin with a question, and the question is:

"Is there a statistically significant difference in teaching evaluation scores for men and women?" When we compute the averages while using the teaching evaluation data set, we find that the teaching evaluation score for women is around 3.9, and for men it's around 4.06.

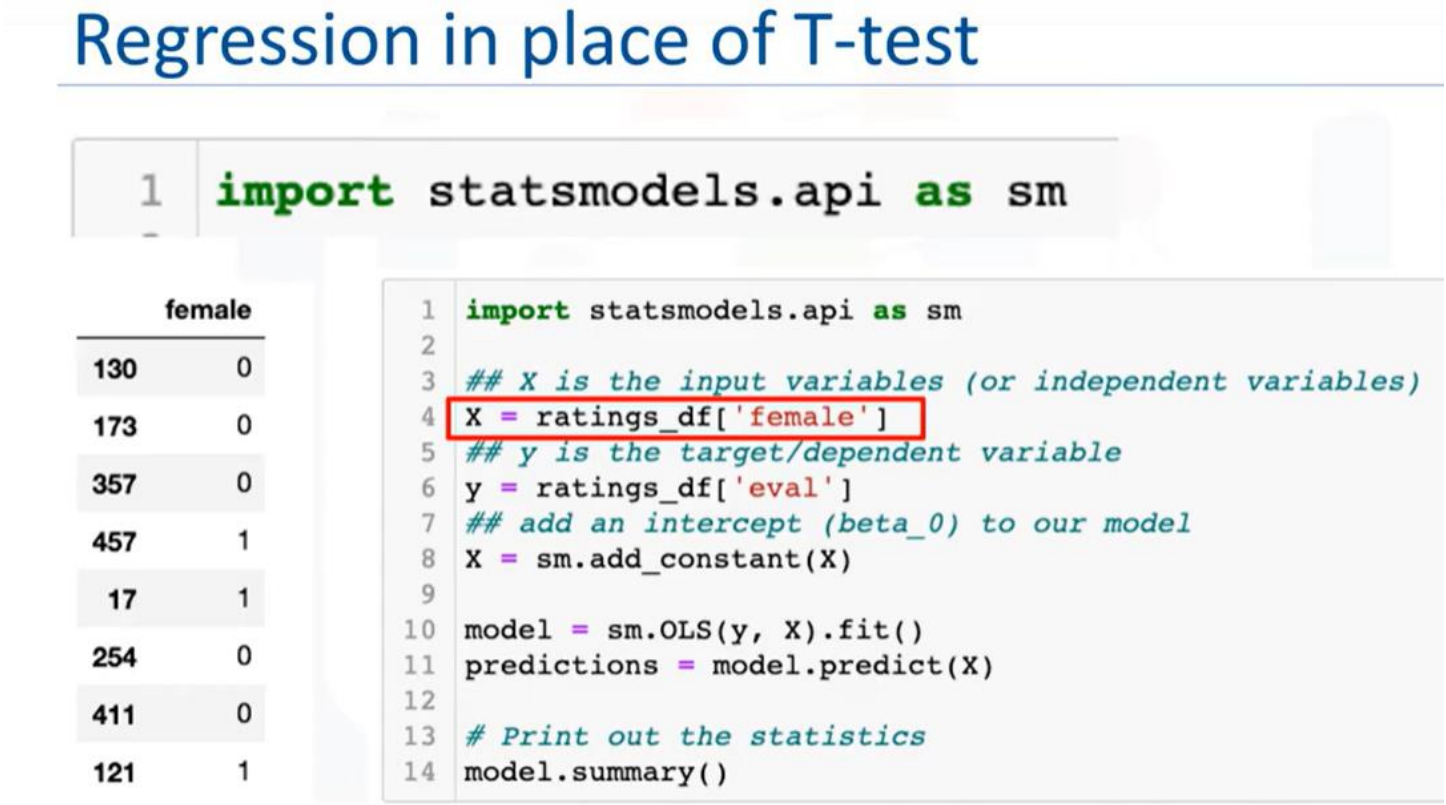
The question is: "Is this difference, even though it's small, statistically significant?" We can run a t-test using Python and compute the statistical significance for the t-test. Here our conclusion is that the teaching evaluation scores difference between men and women is statistically significant. What if we were to do the same thing with the regression model? We will do the linear regression in Python. We will be using the stats model library. We will create a list for the independent variable, that is, the female variable, which has been turned into a binary variable, where one equals female and 0 is male. We will also create a list for the dependent variable "teaching evaluation score." We will manually add the constant, beta0, then we will fit and make predictions; and print out the model summary.

The model summary will print out a table like this. But we are only interested in this part of this table for the t-test. It prints out the coefficient error: t-statistics and p-value. We can see the t-statistics for the female variable is negative 3.25, and the p-value is less than 0.05. That means that there is a statistical difference in mean values for male and female instructors. The coefficient means that you are most likely to lose about 0.17 marks for being a female. We can see that the results from using a regression model and the conclusion is identical, if we run a t-test.





The question is: "Is this difference, even though it's small, statistically significant?" We can run a t-test using Python and compute the statistical significance for the t-test. Here our conclusion is that the teaching evaluation scores difference between men and women is statistically significant.



We can run a t-test using Python and compute the statistical significance for the t-test.

Here our conclusion is that the teaching evaluation scores difference between men and women is statistically significant. What if we were to do the same thing with the regression model? We will do the linear regression in Python. We will be using the stats model library. We will create a list for the independent variable, that is, the female variable, which has been turned into a binary variable, where one equals female and 0 is male. We will also create a list for the dependent variable "teaching evaluation score." We will manually add the constant, beta0, then we will fit and make predictions; and print out the model summary.

Regression in place of T-test

Dep. Variable:	eval	**R-squared:**	0.022
Model:	OLS	**Adj. R-squared:**	0.020
Method:	Least Squares	**F-statistic:**	10.56
Date:	Thu, 03 Sep 2020	**Prob (F-statistic):**	0.00124
Time:	14:50:47	**Log-Likelihood:**	-378.50
No. Observations:	463	**AIC:**	761.0
Df Residuals:	461	**BIC:**	769.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0690	0.034	121.288	0.000	4.003	4.135
female	-0.1680	0.052	-3.250	0.001	-0.270	-0.066

Omnibus:	17.625	**Durbin-Watson:**	1.209
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18.970
Skew:	-0.496	**Prob(JB):**	7.60e-05
Kurtosis:	2.981	**Cond. No.**	2.47

Regression in place of T-test

Dep. Variable:	eval	R-squared:	0.022
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	10.56
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	0.00124
Time:	14:50:47	Log-Likelihood:	-378.50
No. Observations:	463	AIC:	761.0
Df Residuals:	461	BIC:	769.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0690	0.034	121.288	0.000	4.003	4.135
female	-0.1680	0.052	-3.250	0.001	-0.270	-0.066

Omnibus:	17.625	Durbin-Watson:	1.209
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.970
Skew:	-0.496	Prob(JB):	7.60e-05
Kurtosis:	2.981	Cond. No.	2.47

But we are only interested in this part of this table for the t-test. It prints out the coefficient error: t-statistics and p-value. We can see the t-statistics for the female variable is negative 3.25, and the p-value is less than 0.05. That means that there is a statistical difference in mean values for male and female instructors. The coefficient means that you are most likely to lose about 0.17 marks for being a female. We can see that the results from using a regression model and the conclusion is identical, if we run a t-test.

Regression in place of ANOVA

When we are comparing the difference in means, or when we are comparing the averages between groups that are more than two, we will use ANOVA, or analysis of variance. We know that if there are only two groups we can use the T-test, but when we are comparing averages for more than two groups, we use analysis of variance.

Working with our teaching evaluation data set, we took the teaching evaluation scores and then we wanted to see what would happen if we took the instructors, and divided them into three groups: 40 years and younger, those between 40 and 57 years of age, and those that are 57 years or older. We computed the average value for teaching evaluation score for the three groups. We wanted to determine if the three mean values were statistically different.

To recap, we ran the analysis of variance test, which uses F distribution. The p-value is less than 0.05, so we reject the null hypothesis: that averages of the group are equal and concluded that the differences are statistically significant.

Now, let us do this with a regression model. We will use the stats model library and also import the ols function. We will create or initiate a linear model of the beauty score, which is our y-variable. Please note that when dealing with a linear regression model, the y-variable has to be a continuous variable, otherwise results will not be accurate.

Now, create the linear model and fit it using the fit function. Use the ANOVA_IM function to create a table that prints out the results of the test statistics. The results will look like this. It will print out the degree of freedom, the sum of square, f-statistics, and the p-value. Like ANOVA from the SciPy package we get the same results, which is that we will reject the null hypothesis: the averages of the group are equal, and conclude that the differences are statistically significant.

You can also turn the age group values into dummy values and run it like you run the regression for t-test. To do that, you will need to create dummy variables for the age groups using the get_dummies function in Pandas. It will look like this: where 1 means they belong to that group, and 0 means otherwise. Just like a binary variable values can only belong to one group. Run the same as you did for the t-test by fitting the variables into an ols function. Predict and print out the model summary. We will get results like this.

Taking a closer look we can see the same results for the f-statistics and the p-value.

Regression in place of ANOVA

Groups of three or more

Let's discretize age

	age_group	eval		
		count	mean	std
0	40 years and younger	113	4.002655	0.505763
1	57 years and older	122	3.933607	0.624250
2	between 40 and 57 years	228	4.030702	0.537923

We wanted to determine if the three mean values were statistically different. To recap, we ran the analysis of variance test, which uses F distribution. The p-value is less than 0.05, so we reject the null hypothesis: that averages of the group are equal and concluded that the differences are statistically significant.

To recap:

ANOVA in Python

Does beauty score for instructors differ by age?

	age_group	beauty		
		count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	57 years and older	122	-0.245777	0.740720
2	between 40 and 57 years	228	-0.035111	0.686637

```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

Ran the analysis of variance test, which uses F distribution. The p-value is less than 0.05, so we reject the null hypothesis: that averages of the group are equal and concluded that differences are statistically significant.

Regression for ANOVA

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```

Now, create the linear model and fit it using the fit function. Use the ANOVA_IM function to create a table that prints out the results of the test statistics.

Regression for ANOVA

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```


	df	sum_sq	mean_sq	F	PR(>F)
age_group	2.0	20.422744	10.211372	17.597559	4.322549e-08
Residual	460.0	266.925153	0.580272	NaN	NaN

Now, create the linear model and fit it using the fit function. Use the ANOVA_IM function to create a table that prints out the results of the test statistics. The results will look like this. It will print out the degree of freedom, the sum of square, f-statistics, and the p-value. Like ANOVA from the SciPy package we get the same results, which is that we will reject the null hypothesis: the averages of the group are equal, and conclude that the differences are statistically significant.

Regression for ANOVA

```
1  X = pd.get_dummies(ratings_df[['age_group']])
```

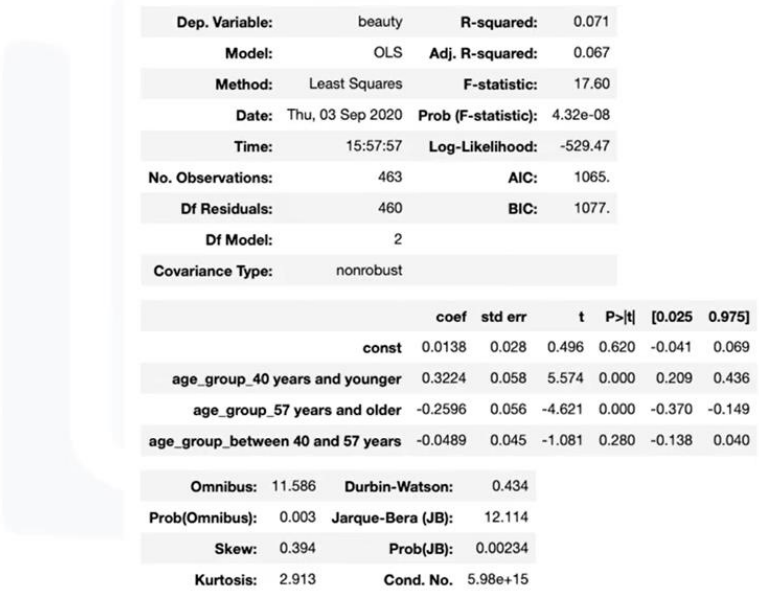
	age_group_40 years and younger	age_group_57 years and older	age_group_between 40 and 57 years
359	0	0	1
107	0	0	1
356	0	0	1
52	0	0	1
440	0	1	0
287	1	0	0

You can also turn the age group values into dummy values and run it like you run the regression for t-test. To do that, you will need to create dummy variables for the age groups using the get_dummies function in Pandas. It will look like this: where 1 means they belong to that group, and 0 means otherwise. Just like a binary variable values can only belong to one group.

Regression for ANOVA

```
1  y = ratings_df2['beauty']
2  ## add an intercept (beta_0) to our model
3  X = sm.add_constant(X)
4
5  model = sm.OLS(y, X).fit()
6  predictions = model.predict(X)
7
8  # Print out the statistics
9  model.summary()
```

Regression for ANOVA



Dep. Variable:	beauty	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.067
Method:	Least Squares	F-statistic:	17.60
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	4.32e-08
Time:	15:57:57	Log-Likelihood:	-529.47
No. Observations:	463	AIC:	1065.
Df Residuals:	460	BIC:	1077.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0138	0.028	0.496	0.620	-0.041	0.069
age_group_40 years and younger	0.3224	0.058	5.574	0.000	0.209	0.436
age_group_57 years and older	-0.2596	0.056	-4.621	0.000	-0.370	-0.149
age_group_between 40 and 57 years	-0.0489	0.045	-1.081	0.280	-0.138	0.040

Omnibus:	11.586	Durbin-Watson:	0.434
Prob(Omnibus):	0.003	Jarque-Bera (JB):	12.114
Skew:	0.394	Prob(JB):	0.00234
Kurtosis:	2.913	Cond. No.	5.98e+15

Run the same as you did for the t-test by fitting the variables into an ols function. Predict and print out the model summary.

We will get results like this.

Regression for ANOVA

Dep. Variable:	beauty	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.067
Method:	Least Squares	F-statistic:	17.60
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	4.32e-08
Time:	15:57:57	Log-Likelihood:	-529.47
No. Observations:	463	AIC:	1065.
Df Residuals:	460	BIC:	1077.
Df Model:	2		
Covariance Type:	nonrobust		

Taking a closer look we can see the same results for the f-statistics and the p-value.

Regression in Place of Correlation

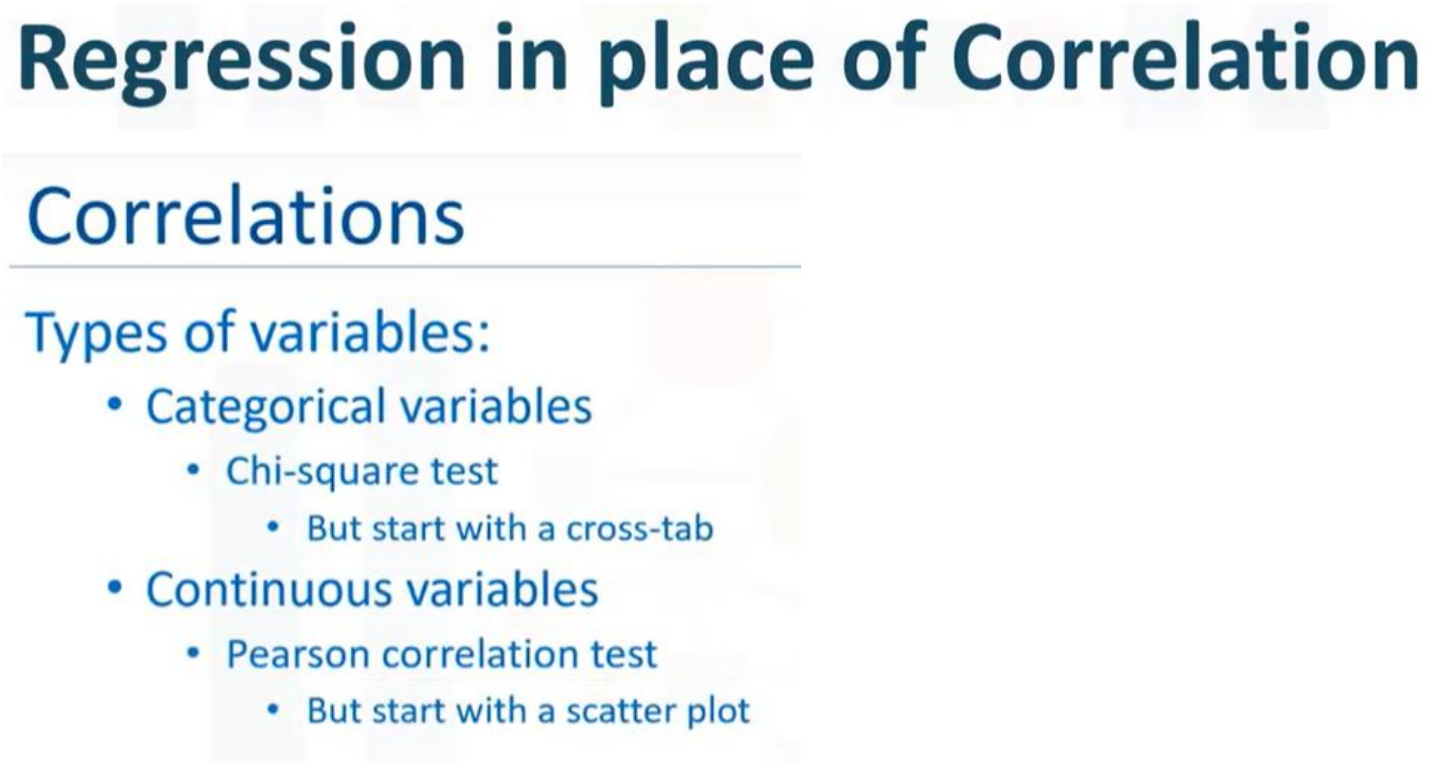
In this video, we will illustrate how one can use regression models in place of tests conducted for correlation analysis. We will return to the basics.

There are two types, or mostly two types, of variables. First are the categorical variables for which we use chi-square tests to determine if there is an association between the two; and second, are the categorical variables; or, we could have continuous variables where we use the Pearson correlation test. We will focus on just the continuous variables.

We can plot two continuous variables in a scatter plot. The teaching evaluation scores are on the y-axis, and the normalized beauty scores are on the x-axis. You could sort of see a relationship between the two variables.

It's an upward sloping type of a relationship. We see that as the beauty score increases, so does the teaching evaluation score. Remember, we use the Pearson correlation test to determine the relationship and its significance level.

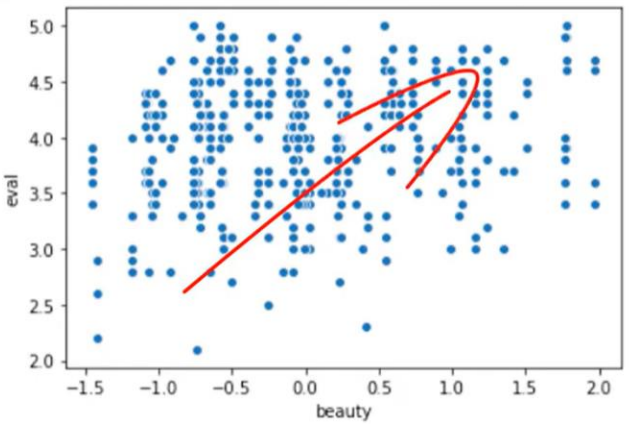
Now let's do the same in regression. Just like we did with the T-test and the F-test, we will fit a linear model for both the beauty and evaluation score values; and print out the model summary. Taking a closer look it prints out a p-value of 4.25 times 10 raised to power negative 5, which is less than 0.05. That is very similar to when we run the Pearson R function. It will also give us the R-squared value, that is, if we took the square root of 0.036 it will give us 0.189, which is the same value as the correlation coefficient from computing the Pearson R.



There are two types, or mostly two types, of variables. First are the categorical variables for which we use chi-square tests to determine if there is an association between the two; and second, are the categorical variables; or, we could have continuous variables where we use the Pearson correlation test. We will focus on just the continuous variables.

Continuous variables

Is teaching evaluation score correlated with beauty score?



We can plot two continuous variables in a scatter plot. The teaching evaluation scores are on the y-axis, and the normalized beauty scores are on the x-axis. You could sort of see a relationship between the two variables. It's an upward sloping type of a relationship. We see that as the beauty score increases, so does the teaching evaluation score. Remember, we use the Pearson correlation test to determine the relationship and its significance level.

Pearson Correlation Test

```
1 scipy.stats.pearsonr(ratings_df['beauty'], ratings_df['eval'])
```

```
(0.1890390908404521, 4.247115419812614e-05)
```

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)} \cdot \sqrt{\operatorname{var}(y)}}$$

Null Hypothesis: There is no association between an instructor's looks and teaching evaluation score.

Since the p-value (Sig. (2-tailed) < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between beauty and teaching evaluation score.

Correlation coefficient varies between -1 and 1.

Remember, we use the Pearson correlation test to determine the relationship and its significance level. Now let's do the same in regression.

Association between beauty and teaching scores

```
## X is the input variables (or independent variables)
X = ratings_df['beauty']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

Just like we did with the T-test and the F-test, we will fit a linear model for both the beauty and evaluation score values; and print out the model summary. Just like we did with the T-test and the F-test, we will fit a linear model for both the beauty and evaluation score values;

Association between beauty and teaching scores

Dep. Variable:	eval	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	17.08
Date:	Thu, 03 Sep 2020	**Prob (F-statistic):**	4.25e-05
Time:	16:36:25	**Log-Likelihood:**	-375.32
No. Observations:	463	**AIC:**	754.6
Df Residuals:	461	**BIC:**	762.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.9983	0.025	157.727	0.000	3.948	4.048
beauty	0.1330	0.032	4.133	0.000	0.070	0.196

Omnibus:	15.399	**Durbin-Watson:**	1.238
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16.405
Skew:	-0.453	**Prob(JB):**	0.000274
Kurtosis:	2.831	**Cond. No.**	1.27

Association between beauty and teaching scores

Dep. Variable:	eval	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	17.08
Date:	Thu, 03 Sep 2020	**Prob (F-statistic):**	4.25e-05
Time:	16:36:25	**Log-Likelihood:**	-375.32
No. Observations:	463	**AIC:**	754.6
Df Residuals:	461	**BIC:**	762.9
Df Model:	1		
Covariance Type:	nonrobust		

Pearson R – P value

4.247115419812614e-05)

and print out the model summary. Taking a closer look it prints out a p-value of 4.25 times 10 raised to power negative 5, which is less than 0.05. That is very similar to when we run the Pearson R function. It will also give us the R-squared value, that is, if we took the square root of 0.036 it will give us 0.189, which is the same value as the correlation coefficient from computing the Pearson R.

Python Packages for Data Science

In order to do data analysis in Python, we should first tell you a little bit about the main packages relevant to analysis in Python. A Python library is a collection of functions and methods that allows you to perform lots of actions without writing your code. The libraries usually contain built-in modules, providing different functionalities, which you can use directly. And there are extensive libraries, offering a broad range of facilities. We divided the Python data analysis libraries into three groups.

The first group is called

1. Scientific Computing Libraries.

- Pandas offers data structure and tools for effective data manipulation and analysis. It provides fast access to structured data. The primary instrument of Pandas is a two-dimensional table consisting of columns and rows' labels, which is called a data frame. It is designed to provide an easy indexing function.
- Numpy library uses arrays as their inputs and outputs. It can be extended to objects for matrices. And with a little change of coding developers perform fast array processing.
- SciPy includes functions for some advanced math problems, as listed in the slide, as well as data visualization.

2.Using data visualization methods are the best way to communicate with others and show the meaningful results of analysis. These libraries enable you to create graphs, charts, and maps.

- The Matplotlib package is the most well-known library for data visualization. This package is great for making graphs and plots. The graphs are also highly customizable.
- Another high level visualization library is Seaborn. It is based on Matplotlib. It's very easy to generate some sort of plots like heat-maps, time series, and violin plots.

With machine learning algorithms, we're able to develop a model using our data set, and obtain predictions.

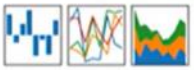
3.The algorithmic libraries tackle some machine learning tasks from basic to complex. We introduced two packages.

- The Scikit-learn library contains tools for statistical modeling including regression, classification, clustering, and so on. It is built on NumPy, SciPy, and Matplotlib. Statsmodels is also a Python module that allows users to explore data, estimate statistical models, and perform statistical tests. Now let's get into Statistics!

Python packages for data science

Scientifics Computing Libraries in Python

1. **Scientifics Computing** Libraries



Pandas
(Data structures & tools)

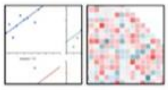
NumPy
(Arrays & matrices)

SciPy
(Integrals, solving differential equations, optimization)

Visualization Libraries in Python

2. **Visualization** Libraries

Matplotlib
(plots & graphs, most popular)



Seaborn
(plots : heat maps, time series, violin plots)

Algorithmic Libraries in Python

3. **Algorithmic** libraries

Scikit-learn
(Machine Learning : regression, classification,...)

Statsmodels
(Explore data, estimate statistical models, and perform statistical tests.)

Regression Analysis

Estimated time needed: **30** minutes

The goal of regression analysis is to describe the relationship between one set of variables called the dependent variables, and another set of variables, called independent or explanatory variables. When there is only one explanatory variable, it is called simple regression.

Objectives

After completing this lab you will be able to:

- Import Libraries
- Regression analysis in place of the t-test
- Regression analysis in place of ANOVA
- Regression analysis in place of correlation

Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[1]:
```
#install specific version of libraries used in lab
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#! mamba install scipy=1.7.1-y
#!  mamba install seaborn=0.9.0-y
#!  mamba install matplotlib=3.4.3-y
#!  mamba install statsmodels=0.12.0-y
```

Import the libraries we need for the lab

[2]:
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
```

Read in the csv file from the URL using the request library

```
[3]: ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.csv'
     ratings_df = pd.read_csv(ratings_url)
```

In this section, you will learn how to run regression analysis in place of the t-test, ANOVA, and correlation

Regression with T-test: Using the teachers rating data set, does gender affect teaching evaluation rates?

Initially, we had used the t-test to test if there was a statistical difference in evaluations for males and females, we are now going to use regression. We will state the null hypothesis:

- H_0:β1H_0:β1 = 0 (Gender has no effect on teaching evaluation scores)
- H_1:β1H_1:β1 is not equal to 0 (Gender has an effect on teaching evaluation scores)

We will use the female variable. female = 1 and male = 0

```
[4]: ## X is the input variables (or independent variables)
     X = ratings_df['female']
     ## y is the target/dependent variable
     y = ratings_df['eval']
     ## add an intercept (beta_0) to our model
     X = sm.add_constant(X)

     model = sm.OLS(y, X).fit()
     predictions = model.predict(X)

     # Print out the statistics
     model.summary()
```

[4]:

OLS Regression Results

Dep. Variable:	eval	**R-squared:**	0.022
Model:	OLS	**Adj. R-squared:**	0.020
Method:	Least Squares	**F-statistic:**	10.56
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	0.00124
Time:	03:20:51	**Log-Likelihood:**	-378.50
No. Observations:	463	**AIC:**	761.0
Df Residuals:	461	**BIC:**	769.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0690	0.034	121.288	0.000	4.003	4.135
female	-0.1680	0.052	-3.250	0.001	-0.270	-0.066

Omnibus:	17.625	**Durbin-Watson:**	1.209
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18.970
Skew:	-0.496	**Prob(JB):**	7.60e-05
Kurtosis:	2.981	**Cond. No.**	2.47

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Conclusion: Like the t-test, the p-value is less than the alpha (α) level = 0.05, so we reject the null hypothesis as there is evidence that there is a difference in mean evaluation scores based on gender. The coefficient -0.1680 means that females get 0.168 scores less than men.

Regression with ANOVA: Using the teachers' rating data set, does beauty score for instructors differ by age?

State the Hypothesis:

- H_0:μ1=μ2=μ3H_0:μ1=μ2=μ3 (the three population means are equal)
- H_1:H_1: At least one of the means differ

Then we group the data like we did with ANOVA

[5]:
```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

Use OLS function from the statsmodel library

[6]:
```
from statsmodels.formula.api import ols
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```

```
                df      sum_sq    mean_sq          F        PR(>F)
age_group      2.0   20.422744  10.211372  17.597559  4.322549e-08
Residual     460.0  266.925153   0.580272        NaN           NaN
```

Conclusion: We can also see the same values for ANOVA like before and we will reject the null hypothesis since the p-value is less than 0.05 there is significant evidence that at least one of the means differ.

Regression with ANOVA option 2

Create dummy variables - A dummy variable is a numeric variable that represents categorical data, such as gender, race, etc. Dummy variables are dichotomous, i.e they can take on only two quantitative values.

[7]:
```
X = pd.get_dummies(ratings_df[['age_group']])
```

[8]:
```
y = ratings_df['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

[8]:

OLS Regression Results

Dep. Variable:	beauty	**R-squared:**	0.071
Model:	OLS	**Adj. R-squared:**	0.067
Method:	Least Squares	**F-statistic:**	17.60
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	4.32e-08
Time:	03:22:23	**Log-Likelihood:**	-529.47
No. Observations:	463	**AIC:**	1065.
Df Residuals:	460	**BIC:**	1077.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0138	0.028	0.496	0.620	-0.041	0.069
age_group_40 years and younger	0.3224	0.058	5.574	0.000	0.209	0.436
age_group_57 years and older	-0.2596	0.056	-4.621	0.000	-0.370	-0.149
age_group_between 40 and 57 years	-0.0489	0.045	-1.081	0.280	-0.138	0.040

Omnibus:	11.586	**Durbin-Watson:**	0.434
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	12.114
Skew:	0.394	**Prob(JB):**	0.00234
Kurtosis:	2.913	**Cond. No.**	6.90e+15

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 1.35e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

You will get the same results and conclusion

Correlation: Using the teachers' rating dataset, Is teaching evaluation score correlated with beauty score?

[9]:
```
## X is the input variables (or independent variables)
X = ratings_df['beauty']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

[9]:

OLS Regression Results

Dep. Variable:	eval	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	17.08
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	4.25e-05
Time:	03:22:52	**Log-Likelihood:**	-375.32
No. Observations:	463	**AIC:**	754.6
Df Residuals:	461	**BIC:**	762.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.9983	0.025	157.727	0.000	3.948	4.048
beauty	0.1330	0.032	4.133	0.000	0.070	0.196

Omnibus:	15.399	**Durbin-Watson:**	1.238
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16.405
Skew:	-0.453	**Prob(JB):**	0.000274
Kurtosis:	2.831	**Cond. No.**	1.27

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Conclusion: $p < 0.05$ there is evidence of correlation between beauty and evaluation scores

Practice Questions

Question 1: Using the teachers' rating data set, does tenure affect beauty scores?

- Use $\alpha = 0.05$

[10]:
```
### insert code here
## put beauty scores in a list
y = ratings_df['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

[10]:

OLS Regression Results

Dep. Variable:	beauty	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	1.010e+34
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	0.00
Time:	03:24:01	**Log-Likelihood:**	16160.
No. Observations:	463	**AIC:**	-3.232e+04
Df Residuals:	461	**BIC:**	-3.231e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.674e-18	7.84e-18	1.107	0.269	-6.73e-18	2.41e-17
beauty	1.0000	9.95e-18	1.01e+17	0.000	1.000	1.000

Omnibus:	35.667	**Durbin-Watson:**	0.445
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	41.953
Skew:	0.729	**Prob(JB):**	7.76e-10
Kurtosis:	3.221	**Cond. No.**	1.27

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Double-click **here** for a hint.

```
Double-click **here** for a hint.

<!-- The hint is below:
## State Hypothesis
Null Hypothesis: Mean beauty scores for tenured and non-tenured instructors are equal
Alternative Hypothesis: There is a difference in mean beauty scores for tenured and non-tenured instructors

## use the dummy variable for tenure - the OLS library doesn't recognize texts
X = ratings_df['tenured_prof']
-->
```

```
Double-click **here** for the solution.

<!-- The answer is below:
## put beauty scores in a list
y = ratings_df['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

p-value is greater than 0.05, so we fail to reject the null hypothesis as there is no evidence that the mean difference of tenured and untenured instructors are different

Question 2: Using the teachers' rating data set, does being an English speaker affect the number of students assigned to professors?

Use "allstudents"

Use α = 0.05 and α = 0.1

[11]:
```
## insert code here
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

[11]:

OLS Regression Results

Dep. Variable:	beauty	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	1.010e+34
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	0.00
Time:	03:24:44	**Log-Likelihood:**	16160.
No. Observations:	463	**AIC:**	-3.232e+04
Df Residuals:	461	**BIC:**	-3.231e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.674e-18	7.84e-18	1.107	0.269	-6.73e-18	2.41e-17
beauty	1.0000	9.95e-18	1.01e+17	0.000	1.000	1.000

Omnibus:	35.667	**Durbin-Watson:**	0.445
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	41.953
Skew:	0.729	**Prob(JB):**	7.76e-10
Kurtosis:	3.221	**Cond. No.**	1.27

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
Double-click **here** for a hint.

<!-- The hint is below:
## State Hypothesis
Null Hypothesis: Mean number of students assigned to native English speakers vs non-native English speakers are equal
Alternative Hypothesis: There is a difference in mean number of students assigned to native English speakers vs non-native English speakers

## Is the instructor a native English speaker - make sure to use the binary variable "English speaker"
X = ratings_df['English_speaker']
## You can use the students or all students variable
y = ratings_df['allstudents']
-->
```

```
Double-click **here** for the solution.

<!-- The answer is below:
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

At α = 0.05, p-value is greater, we fail to reject the null hypothesis as there is no evidence that being a native English speaker or a non-native English speaker affects the number of students assigned to an instructor.

At α = 0.1, p-value is less, we reject the null hypothesis as there is evidence that there is a significant difference of mean number of students assigned to native English speakers vs non-native English speakers.

Question 3: Using the teachers' rating data set, what is the correlation between the number of students who participated in the evaluation survey and evaluation scores?

Use "students" variable

[12]:
```
## insert code here
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

[12]:

OLS Regression Results

Dep. Variable:	beauty	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	1.010e+34
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	0.00
Time:	03:25:26	**Log-Likelihood:**	16160.
No. Observations:	463	**AIC:**	-3.232e+04
Df Residuals:	461	**BIC:**	-3.231e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.674e-18	7.84e-18	1.107	0.269	-6.73e-18	2.41e-17
beauty	1.0000	9.95e-18	1.01e+17	0.000	1.000	1.000

Omnibus:	35.667	**Durbin-Watson:**	0.445
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	41.953
Skew:	0.729	**Prob(JB):**	7.76e-10
Kurtosis:	3.221	**Cond. No.**	1.27

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
Double-click **here** for a hint.

<!-- The hint is below:
## create a list of students and evaluation socres
X = ratings_df['students']
y = ratings_df['eval']
-->
```

```
Double-click **here** for the solution.

<!-- The rest of the answer is below:
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

R-square is 0.001, R will be √0.001, correlation coefficient is 0.03 (close to 0). There is a very weak correlation between the number of students who participated in the evaluation survey and evaluation scores

1.

Question 1

Does running an ANOVA give the same p-value results as running a regression analysis when testing the difference in group means?

1 / 1 point

- (●) True
- ( ) False

Correct! We can run the regression in place of ANOVA

2.

Question 2

Give the results of the regression analysis below, what is the correlation coefficient?

1 / 1 point

- ( ) 0.036
- (●) 0.19
- ( ) 17.08
- ( ) 0.034

3.

Question 3

Given the results for tenure-ship vs teaching evaluation, if our null hypothesis is that there is no difference in mean evaluation scores for professors who are tenured vs professors who are not tenured. What will be the conclusion of the t-test statistics?

1 / 1 point

- (●) P-value is less than 0.05, that means that there is a difference in mean values for professors who are tenures versus professors who are not tenured.
- ( ) P-value is less than 0.05, we will fail to reject the null hypothesis.
- ( ) There is no conclusive evidence in the results above.

4.

Question 4

We run a regression analysis in place of a t-test to test if there is a difference in number of students enrolled in classes with professors who are native english speakers (English_speakers = 1) vs professors who are not (English_speakers = 0). The table is shown below. What doe the coefficient for English_speakers mean?

1 / 1 point

- ( ) Professors who are English speakers get about 30 more students enrolled on average
- ( ) Professors who are English speakers get about 27 less students enrolled on average
- ( ) We can't conclude because the error is too large and if factored in could change the conclusion of the results
- (●) Professors who are English speakers get about 27 more students enrolled on average

5.

Question 5

Which of these are correct about correlation coefficient? (Select all that apply)

1 / 1 point

☑ A correlation coefficient of -0.9 indicates a strong linear relationship?

Correct! The negative sign means they are strongly negatively correlated

☐ A correlation coefficient of -0.9 indicates a weak linear relationship?

☐ The correlation coefficient (r) ranges from 0 to 1

☑ The correlation coefficient (r) ranges from -1 to 1

Correct! Values can be positively and negatively related

6.

Question 6

Which of these options is most likely to be the null hypothesis for testing correlation between two variables?

1 / 1 point

◉ There is no association between an instructor's looks and teaching evaluation score.

○ There is an association between an instructor's looks and teaching evaluation score.

○ There is a partial association between an instructor's looks and teaching evaluation score.

7.

Question 7

If we ran a regression analysis between two continuous variables amount of time spent running on a treadmill vs the amount of calories burnt. If I get a coefficient of 0.33 for the amount of time running on the treadmill and an R-square value of 0.81. What is the correlation coefficient?

1 / 1 point

○ 0.81

○ 0.77

○ 0.66

◉ 0.9

8.

Question 8

Which of the following best explains a scatter plot?

1 / 1 point

◉ A two-dimensional graph of data values.

○ A one-dimensional graph of randomly scattered data.

○ A two-dimensional graph of a curved line.

○ A two-dimensional graph of a straight line.

Correct! A scatter plot represents the relationship between two continuous data

Quiz: Regression Analysis

Bookmark this page

Graded Quiz due Jun 17, 2022 07:31 +08

Quiz: Regression Analysis

8/8 points (graded)

Running an ANOVA gives the same p-value results as running a regression analysis when testing the difference in group means.

- (●) True
- ( ) False

✔

Answer
Correct: Correct! We can run the regression in place of ANOVA.

Give the results of the regression analysis below, what is the correlation coefficient?

Dep. Variable:	eval	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	17.08
Date:	Thu, 03 Sep 2020	**Prob (F-statistic):**	4.25e-05
Time:	16:36:25	**Log-Likelihood:**	-375.32
No. Observations:	463	**AIC:**	754.6
Df Residuals:	461	**BIC:**	762.9
Df Model:	1		
Covariance Type:	nonrobust		

- (●) 0.19
- ( ) 0.036
- ( ) 17.08
- ( ) 0.034

✔

Answer
Correct: Correct!

Given the results for tenure-ship vs teaching evaluation, if our null hypothesis is that there is no difference in mean evaluation scores for professors who are tenured vs professors who are not tenured. What will be the conclusion of the t-test statistics?

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.1333	0.055	75.791	0.000	4.026	4.241
tenured_prof	-0.1732	0.062	-2.805	0.005	-0.295	-0.052

- ○ P-value is less than 0.05, we will fail to reject the null hypothesis.
- ◉ P-value is less than 0.05, that means that there is a difference in mean values for professors who are tenures versus professors who are not tenured.
- ○ There is no conclusive evidence in the results above.

✔

Answer
Correct: Correct!

We run a regression analysis in place of a t-test to test if there is a difference in number of students enrolled in classes with professors who are visible minority(vismin = 1) vs professors who are not (vismin = 0). The table is shown below. What does the coefficient for vismin mean?

	coef	std err	t	P>|t|	[0.025	0.975]
const	58.0902	3.745	15.513	0.000	50.731	65.449
vismin	-21.0746	10.072	-2.092	0.037	-40.867	-1.282

- ◉ Professors who are visible minority get about 21 students less on average that professors who aren't visible minority.
- ○ Professors who are visible minority get about 21 students more on average that professors who aren't visible minority.
- ○ Professors who are visible minority get about 58 students less on average that professors who aren't visible minority.
- ○ We can't conclude because the error is too large and if factored could change the conclusion of the tests.

✔

Answer
Correct: Correct!

Which of these are correct about correlation coefficient? (Select all that apply)

- [x] The correlation coefficient (r) ranges from -1 to 1
- [ ] The correlation coefficient (r) ranges from 0 to 1
- [ ] A correlation coefficient of -0.9 indicates a weak linear relationship
- [x] A correlation coefficient of -0.9 indicates a strong linear relationship

Which of these options is most likely to be the null hypothesis for testing correlation between two variables?

- (●) There is no association between an instructor's looks and teaching evaluation score.
- ( ) There is an association between an instructor's looks and teaching evaluation score.
- ( ) There is a partial association between an instructor's looks and teaching evaluation score.

✔

Answer
Correct: Correct!

If we ran a regression analysis between two continuous variables amount of time spent running on a treadmill vs the amount of calories burnt. If I get a coefficient of 0.33 for the amount of time running on the treadmill and an R-square value of 0.81. What is the correlation coefficient?

- ( ) 0.77
- ( ) 0.81
- ( ) 0.66
- (●) 0.9

✔

Answer
Correct: Correct!

Which of the following best explains a scatter plot?

- ( ) A one-dimensional graph of randomly scattered data.
- ( ) A two-dimensional graph of a straight line.
- (●) A two-dimensional graph of data values.
- ( ) A two-dimensional graph of a curved line.

✔

Answer
Correct: Correct! A scatter plot represents the relationship between two continuous data.

Final Exam

Bookmark this page

Final Exam due Jun 20, 2022 19:31 +08 Completed

Final Exam

14/14 points (graded)

What is the 75th percentile of the following data set: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

- (●) 7
- ( ) 5.5
- ( ) 3
- ( ) 8

✔

Answer
Correct: Correct!

The median represents a value in the data set where:

- ( ) Most observations are negative
- (●) Half of the observations are above the median and the other half below it
- ( ) Half of the observations are known and the other half not known
- ( ) Most observations are positive

✔

Answer
Correct: Correct!

Which of the following is NOT a descriptive statistic?

- ○ Mean
- ◉ T-test
- ○ Standard deviation
- ○ Median

✔

Answer
Correct: Correct!

What's the best way to display median and outliers?

- ○ A scatter plot
- ○ A time series plot
- ◉ A box plot
- ○ A bubble plot

✔

Answer
Correct:
Correct! Boxplots are a way of displaying the distribution of data based on a five number summary ("minimum", first quartile, median, third quartile, and "maximum"). It also displays the outliers of the dataset.

What is a suitable way to display the average basketball scores between two teams?

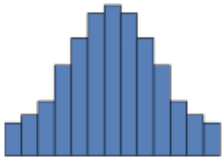
- ◉ A bar chart
- ○ A pie chart
- ○ A histogram
- ○ A scatter plot

✔

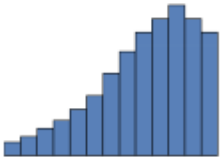
Answer
Correct:
Correct! A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

Given the histograms below, which histogram most closely depicts a normal distribution?

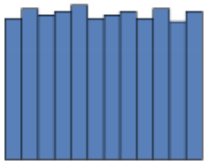
Histogram A



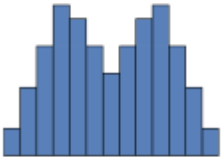
Histogram B



Histogram C



Histogram D



- (●) A
- ( ) B
- ( ) C
- ( ) D

✔

Answer

Correct: Correct! A normal distribution is symmetrical and Bell-shaped.

If you got a 75 on a test in a class with a mean score of 85 and a standard deviation of 5, the z-score of your test score would be:

- (●) -2
- ( ) -3
- ( ) 2
- ( ) 3

✔

Answer

Correct: Correct!

The spread of the normal curve depends upon the value of:

- ○ Median
- ○ Mean
- ○ 1st quartile
- ◉ Standard deviation

✔

Answer
Correct: Correct! The Standard Deviation measures if how spread out the data are.

The weekly earnings of bus drivers are normally distributed with a mean of $395. If only 0.84% of the bus drivers have a weekly income of more than $429.35, the standard deviation of the weekly earnings of the bus drivers is approximately:

- ◉ 14.37
- ○ 17
- ○ 34.83
- ○ 2.39

✔

Answer
Correct: Correct!

For the following samples assume they follow a normal distribution and we assume equal variance, we will like to know if there is a difference between both sample means. If we perform a two-sample t-test for independent samples. What is the p-value for the test Statistics? Sample1 = 9, 11, 10,11,10,12, 9,11,12, 9, 10 Sample2 = 10, 13, 10, 13, 12, 9, 11, 12, 12, 12, 13

- (●) 0.0384
- ( ) 2.21
- ( ) 0.0885
- ( ) 0.975

✔

Answer
Correct: Correct!

Which test is used to test the equality of variance?

- ( ) t-test
- ( ) ANOVA
- (●) Levene's test
- ( ) z-test

✔

Answer
Correct: Correct!

We run a regression analysis between two continuous variables amount of food eaten vs the amount of calories burnt. If I get a coefficient of -0.33 for the amount of food eaten and an R-square value of 0.81. What is the correlation coefficient?

- (●) -0.9
- ( ) 0.9
- ( ) 0.66
- ( ) -0.66

✔

Answer
Correct: Correct!

In the simple linear regression equation, the term B0 represents the:

- ( ) Estimated or predicted response
- (●) Estimated intercept
- ( ) Estimated slope
- ( ) Explanatory variable

✔

Answer
Correct: Correct!

The Pearson correlation is concerned with:

- ( ) The relationship between two categorical variables
- (●) The relationship between two quantitative variables
- ( ) The relationship between a quantitative explanatory variable and a categorical response variable
- ( ) The relationship between a categorical explanatory variable and a quantitative response variable.

✔

Answer
Correct: Correct!