JPMC Machine Learning Essentials

LEARNING VOYAGE

Table of Contents

DAY 1

LEARNING VOYAGE

1: Introduction to Data Analysis

Introduction to Data Analysis

The following topics will be covered in this lesson:

- The fundamentals of data analysis
- Statistical foundations
- Setting up a virtual environment

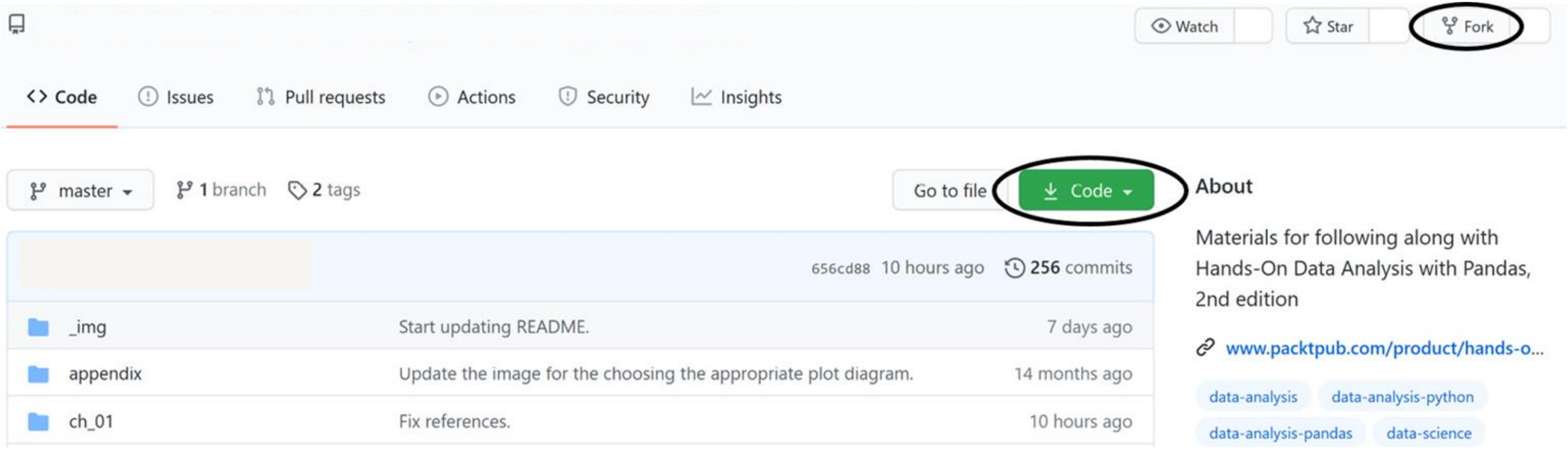
lesson materials

In order to get a local copy of the files, we have a few options (ordered from least useful to most useful):

- Download the ZIP file and extract the files locally.
- Clone the repository without forking it.
- Fork the repository and then clone it.

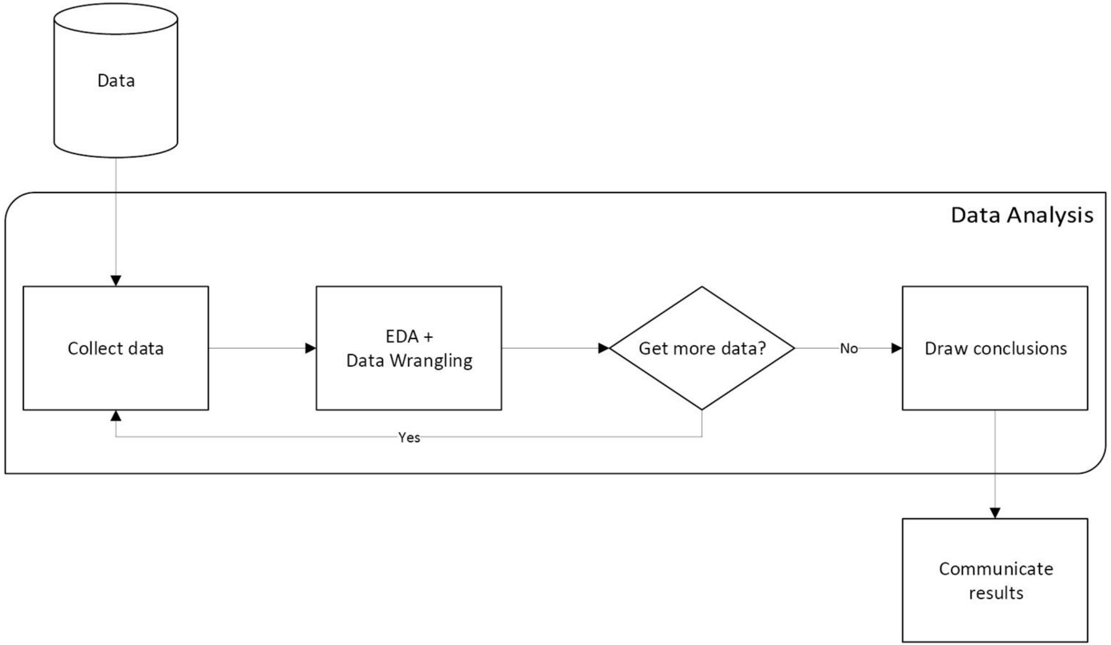
lesson materials

The relevant buttons for initiating this process are circled in the following screenshot:



The fundamentals of data analysis

- The following diagram depicts a generalized workflow:



Data collection

- Data collection is the natural first step for any data analysis—we can't analyze data we don't have.

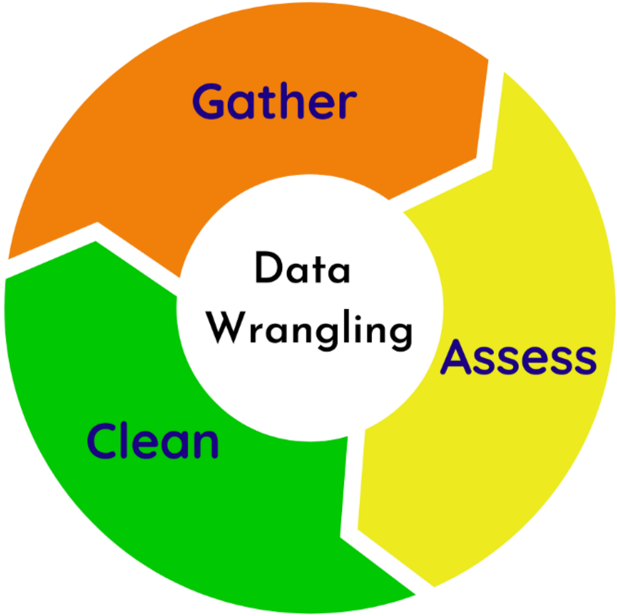
- In reality, our analysis can begin even before we have the data.

- When we decide what we want to investigate or analyze, we have to think about what kind of data we can collect that will be useful for our analysis.

Data wrangling

- Data wrangling is the process of preparing the data and getting it into a format that can be used for analysis.

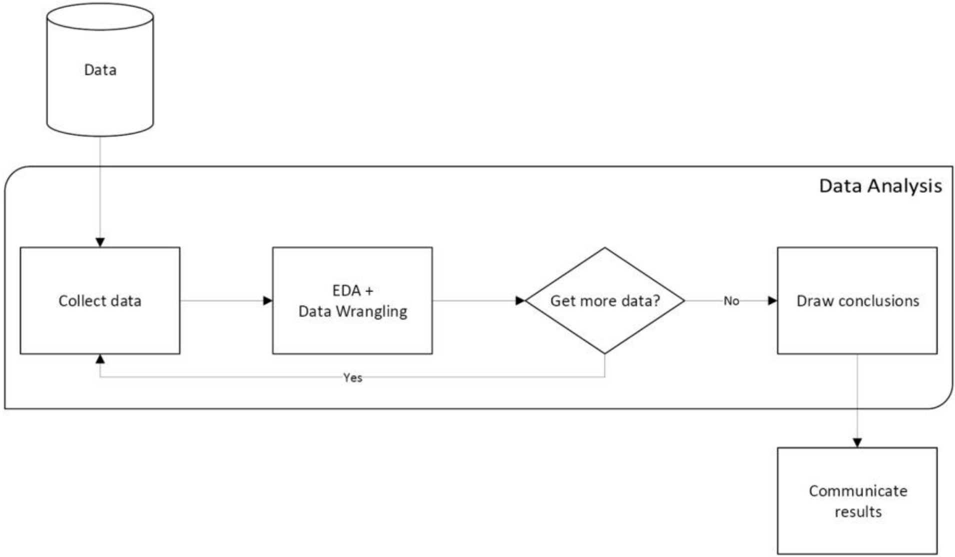
- The unfortunate reality of data is that it is often dirty, meaning that it requires cleaning (preparation) before it can be used.



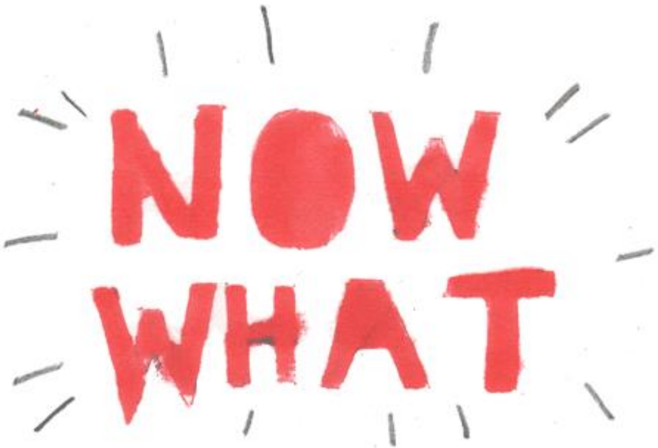
Exploratory data analysis

EDA and data wrangling shared a box.
This is because they are closely tied:

- Data needs to be prepped before EDA.
- Visualizations that are created during EDA may indicate the need for additional data cleaning.
- Data wrangling uses summary statistics to look for potential data issues, while EDA uses them to understand the data.



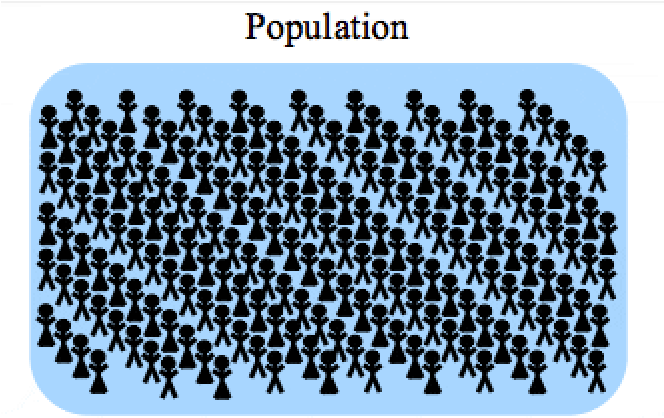
Drawing conclusions



- After we have collected the data for our analysis, cleaned it up, and performed some thorough EDA, it is time to draw conclusions.

- This is where we summarize our findings from EDA and decide the next steps.

Statistical foundations



- When we want to make observations about the data we are analyzing, we often, if not always, turn to statistics in some fashion.
- The data we have is referred to as the sample, which was observed from (and is a subset of) the population.
- Two broad categories of statistics are descriptive and inferential statistics.

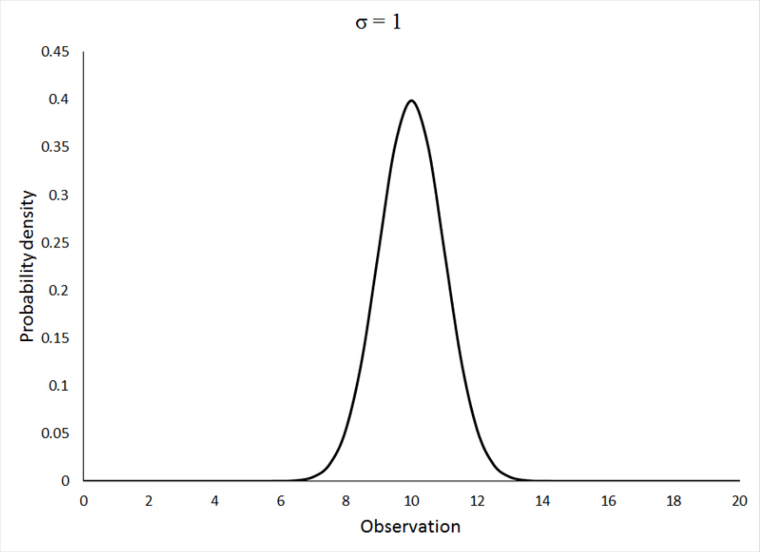
Sampling

- There's an important thing to remember before we attempt any analysis: our sample must be a random sample that is representative of the population.

  - This means that the data must be sampled without bias (for example, if we are asking people whether they like a certain sports team, we can't only ask fans of the team)
  - We should have (ideally) members of all distinct groups from the population in our sample (in the sports team example, we can't just ask men)

Descriptive statistics

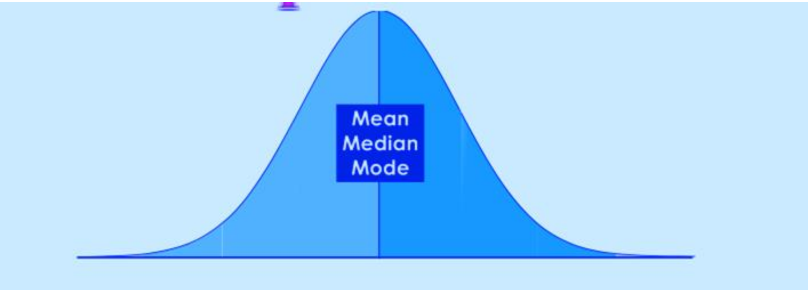
- Descriptive statistics are used to describe and/or summarize the data we are working with.
  - We can start our summarization of the data with a measure of central tendency, which describes where most of the data is centered around, and a measure of spread or dispersion, which indicates how far apart values are.



Descriptive statistics

Measures of central tendency

- Measures of central tendency describe the center of our distribution of data.
- There are three common statistics that are used as measures of center: mean, median, and mode.
- Each has its own strengths, depending on the data we are working with.



Descriptive statistics

Mean

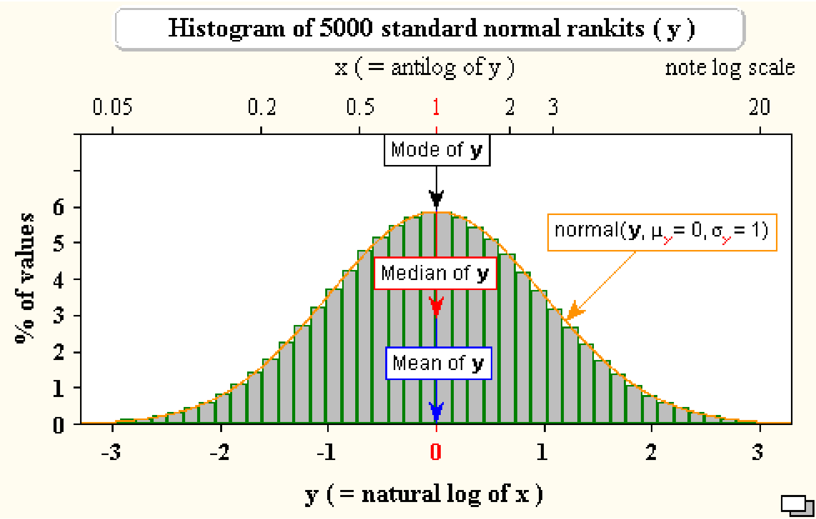
- Perhaps the most common statistic for summarizing data is the average, or mean.
- The population mean is denoted by μ (the Greek letter mu), and the sample mean is written as (pronounced X-bar).
- The sample mean is calculated by summing all the values and dividing by the count of values; for example, the mean of the numbers 0, 1, 1, 2, and 9 is 2.6 ((0 + 1 + 1 + 2 + 9)/5):

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

Descriptive statistics

Median

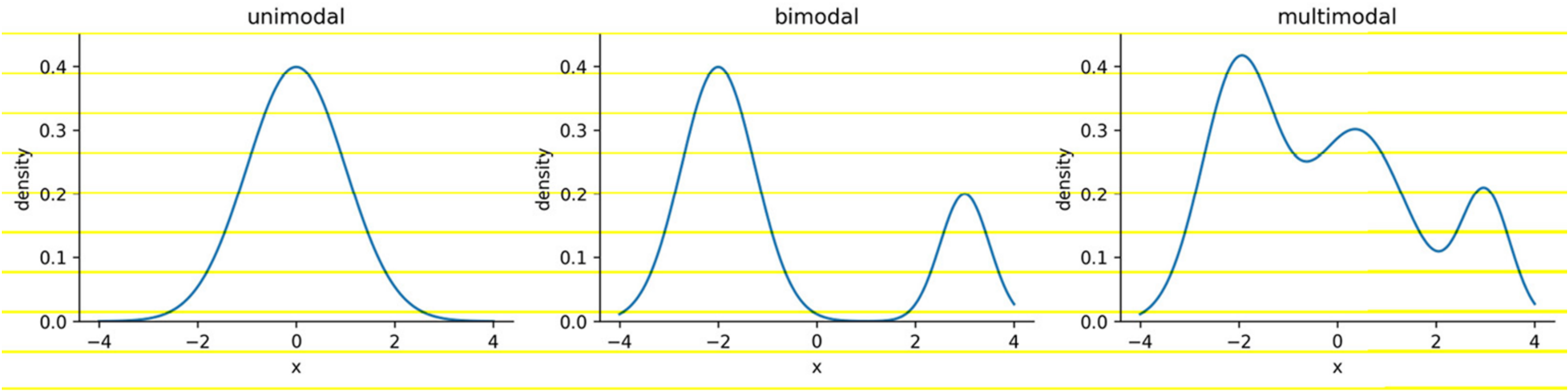
- Unlike the mean, <u>the median is robust to outliers.</u>
- Consider income in the US; the top 1% is much higher than the rest of the population, so this will skew the mean to be higher and distort the perception of the average person's income.
- However, the median will be more representative of the average income because it is the 50th percentile of our data; this means that 50% of the values are greater than the median and 50% are less than the median.



Descriptive statistics

Mode

- The mode is the most common value in the data (if we, once again, have the numbers 0, 1, 1, 2, and 9, then 1 is the mode).



Descriptive statistics

Measures of spread

- Knowing where the center of the distribution is only gets us partially to being able to summarize the distribution of our data—we need to know how values fall around the center and how far apart they are.

Descriptive statistics

Range

- The range is the distance between the smallest value (minimum) and the largest value (maximum).
- The units of the range will be the same units as our data.

$$range = \max(X) - \min(X)$$

Descriptive statistics

Variance

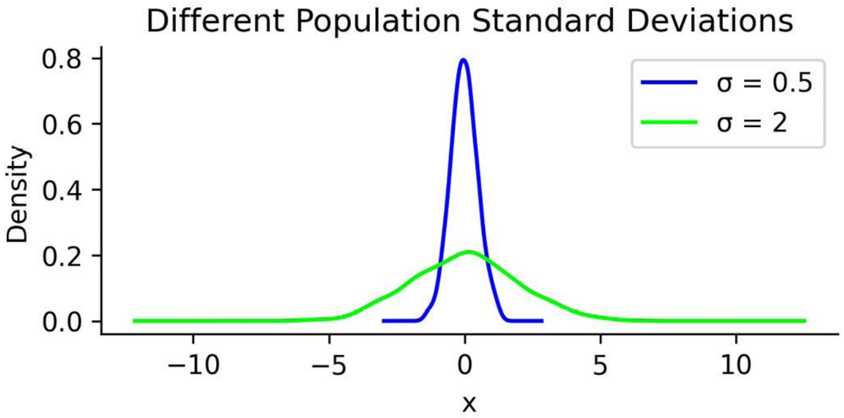
- The variance describes how far apart observations are spread out from their average value (the mean).
- The population variance is denoted as σ2 (pronounced sigma-squared), and the sample variance is written as s2.
- It is calculated as the average squared distance from the mean.

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

Descriptive statistics

Standard deviation

- We can use the standard deviation to see how far from the mean data points are on average.
- A small standard deviation means that values are close to the mean, while a large standard deviation means that values are dispersed more widely.



Descriptive statistics

- The standard deviation is simply the square root of the variance
- By performing this operation, we get a statistic in units that we can make sense of again ($ for our income example):

$$s = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

Descriptive statistics

Coefficient of variation

- When we moved from variance to standard deviation, we were looking to get to units that made sense; however, if we then want to compare the level of dispersion of one dataset to another, we would need to have the same units once again.
- One way around this is to calculate the coefficient of variation (CV), which is unitless.
- The CV is the ratio of the standard deviation to the mean:

$$CV = \frac{s}{\bar{x}}$$

Descriptive statistics

- One common measure for this is the interquartile range (IQR), which is the distance between the 3rd and 1st quartiles:

$$IQR = Q_3 - Q_1$$

Descriptive statistics

Quartile coefficient of dispersion

- This statistic is also unitless, so it can be used to compare datasets.
- It is calculated by dividing the semi-quartile range (half the IQR) by the midhinge (midpoint between the first and third quartiles):

$$QCD = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_1 + Q_3}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Descriptive statistics

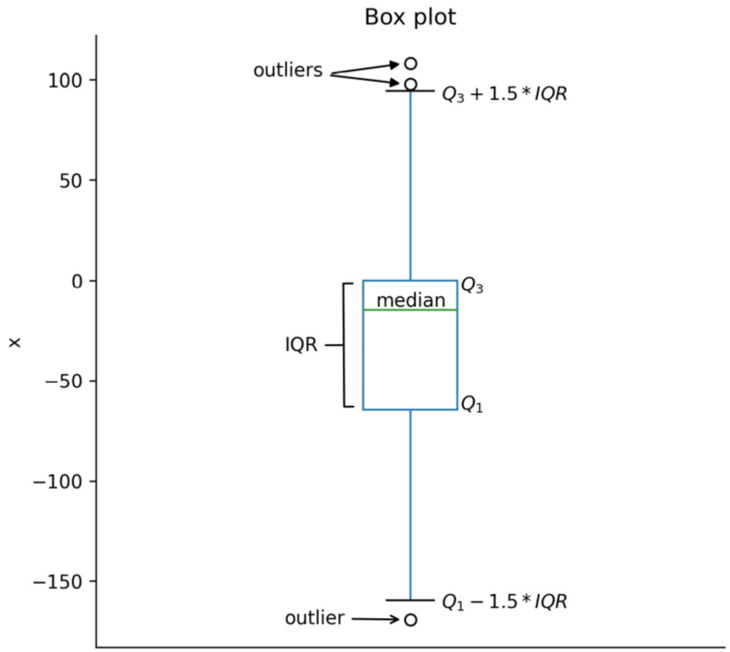
Summarizing data

- We have seen many examples of descriptive statistics that we can use to summarize our data by its center and dispersion; in practice, looking at the 5-number summary and visualizing the distribution prove to be helpful first steps before diving into some of the other aforementioned metrics.
- The 5-number summary, as its name indicates, provides five descriptive statistics that summarize our data:

	Quartile	Statistic	Percentile
1.	Q_0	minimum	0^{th}
2.	Q_1	N/A	25^{th}
3.	Q_2	median	50^{th}
4.	Q_3	N/A	75^{th}
5.	Q_4	maximum	100^{th}

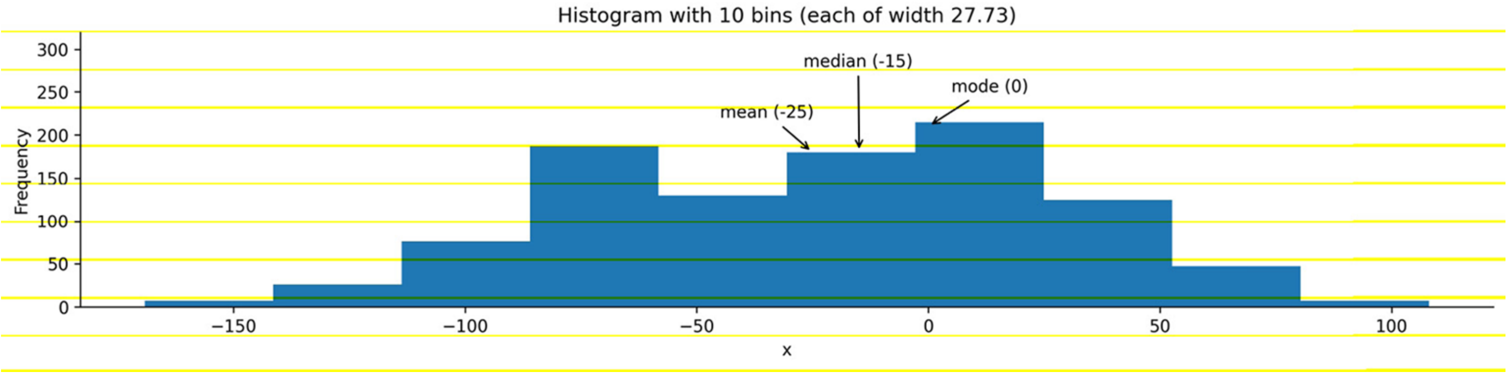
Descriptive statistics

- A box plot (or box and whisker plot) is a visual representation of the 5-number summary.
- The median is denoted by a thick line in the box.
- The top of the box is Q3 and the bottom of the box is Q1.
- Lines (whiskers) extend from both sides of the box boundaries toward the minimum and maximum.



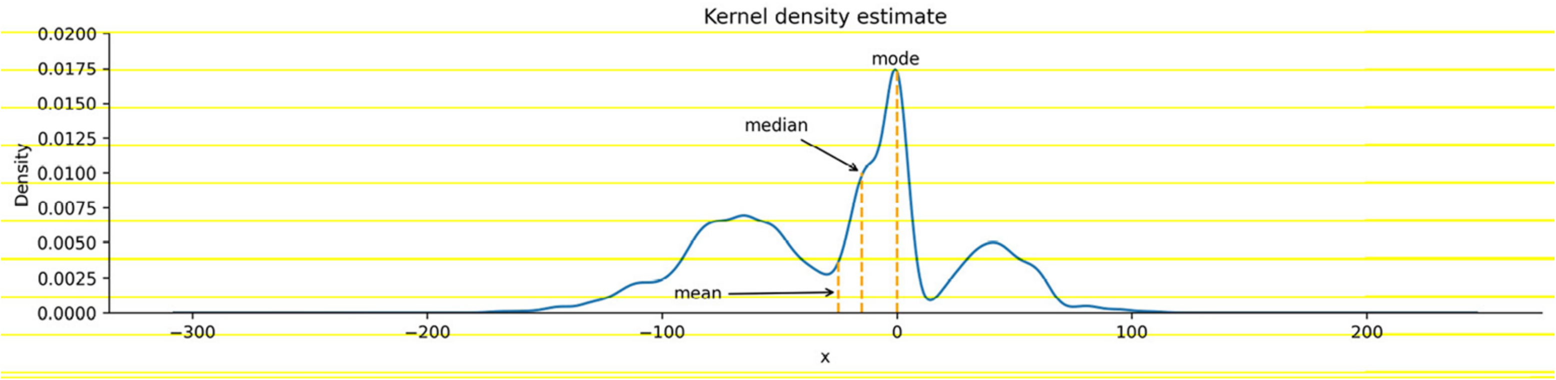
Descriptive statistics

- To make a histogram, a certain number of equal-width bins are created, and then bars with heights for the number of values we have in each bin are added.
- The following plot is a histogram with 10 bins, showing the three measures of central tendency for the same data that was used to generate the box plot in Figure:



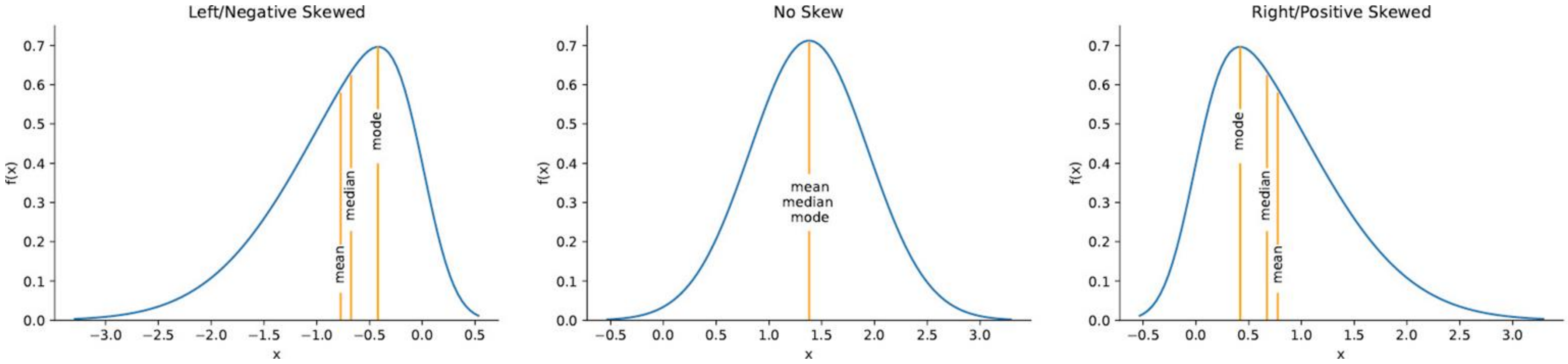
Descriptive statistics

- KDEs are similar to histograms, except rather than creating bins for the data, they draw a smoothed curve, which is an estimate of the distribution's probability density function (PDF).
- The PDF is for continuous variables and tells us how probability is distributed over the values.
- Higher values for the PDF indicate higher likelihoods:



Descriptive statistics

- A left (negative) skewed distribution has a long tail on the left-hand side; a right (positive) skewed distribution has a long tail on the right-hand side.
- In the presence of negative skew, the mean will be less than the median, while the reverse happens with a positive skew.
- When there is no skew, both will be equal:



Descriptive statistics

- When we are interested in the probability of getting a value c ,
we use the cumulative distribution function (CDF), which is the integral (area under the curve) of the PDF:

$$CDF = F(x) = \int_{-\infty}^{x} f(t)dt$$

$$\text{where } f(t) \text{ is the PDF and } \int_{-\infty}^{\infty} f(t)dt = 1$$

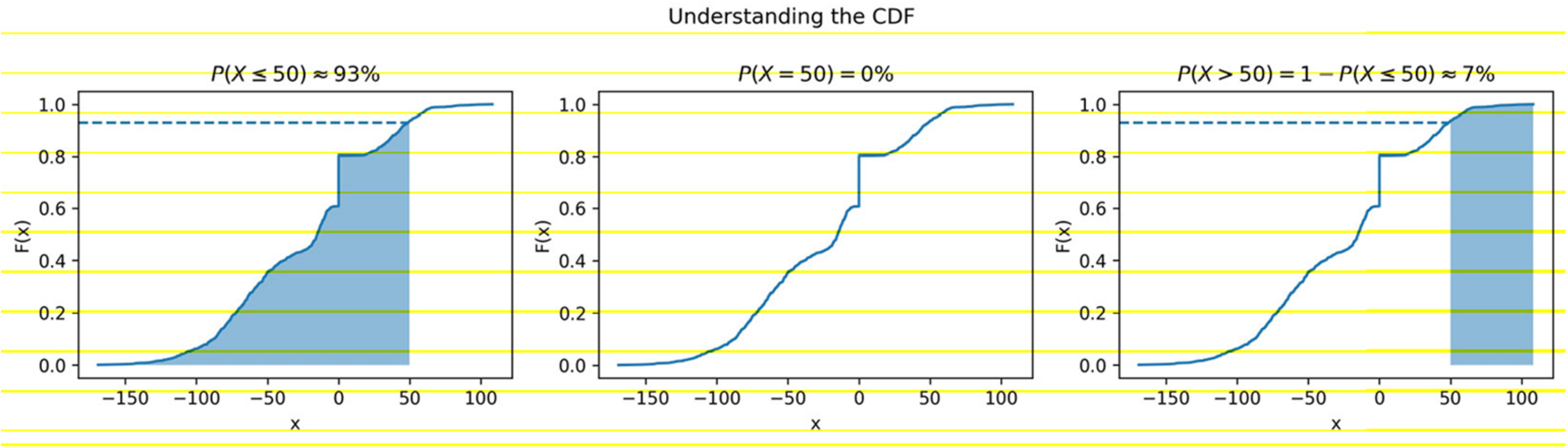
Descriptive statistics

- This is because the probability will be the integral of the PDF from x to x (area under a curve with zero width), which is 0:

$$P(X = x) = \int_{x}^{x} f(t)dt = 0$$

Descriptive statistics

- Let's visualize P(X ≤ 50), P(X = 50), and P(X > 50) as an example:



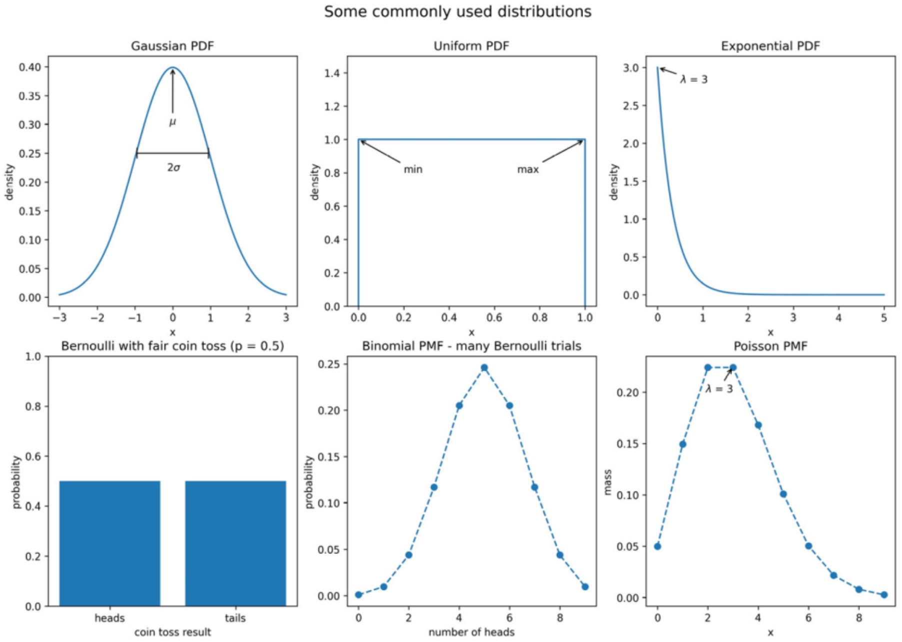
Descriptive statistics

Common distributions

- While there are many probability distributions, each with specific use cases, there are some that we will come across often.
- The Gaussian, or normal, looks like a bell curve and is parameterized by its mean (μ) and standard deviation (σ).
- The standard normal (Z) has a mean of 0 and a standard deviation of 1.
- Many things in nature happen to follow the normal distribution, such as heights.

Descriptive statistics

- We can visualize both discrete and continuous distributions; however, discrete distributions give us a probability mass function (PMF) instead of a PDF:



Descriptive statistics

Scaling data

- In order to compare variables from different distributions, we would have to scale the data, which we could do with the range by using min-max scaling.
- We take each data point, subtract the minimum of the dataset, then divide by the range. This normalizes our data (scales it to the range [0, 1]):

$$x_{scaled} = \frac{x - \min(X)}{range(X)}$$

Descriptive statistics

- This isn't the only way to scale data; we can also use the mean and standard deviation.
- In this case, we would subtract the mean from each observation and then divide by the standard deviation to standardize the data.
- This gives us what is known as a Z-score:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Descriptive statistics

- Quantifying relationships between variables.
- The covariance is a statistic for quantifying the relationship between variables by showing how one variable changes with respect to another (also referred to as their joint variance):

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

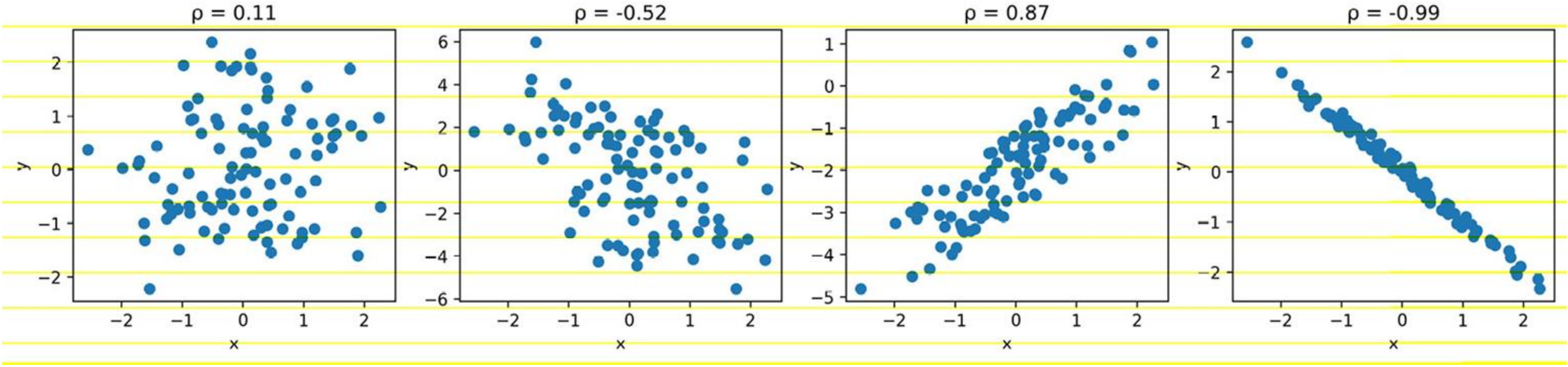
Descriptive statistics

- To find the correlation, we calculate the Pearson correlation coefficient, symbolized by ρ (the Greek letter rho), by dividing the covariance by the product of the standard deviations of the variables:

$$\rho_{X,Y} = \frac{cov(X,Y)}{s_X s_Y}$$

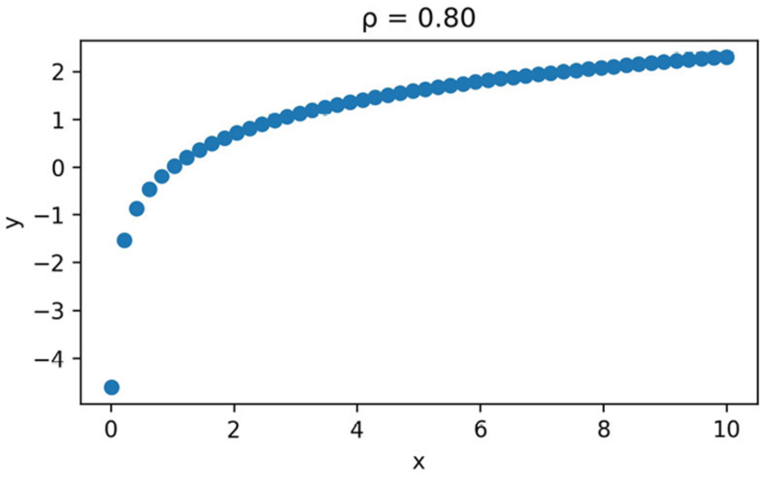
Descriptive statistics

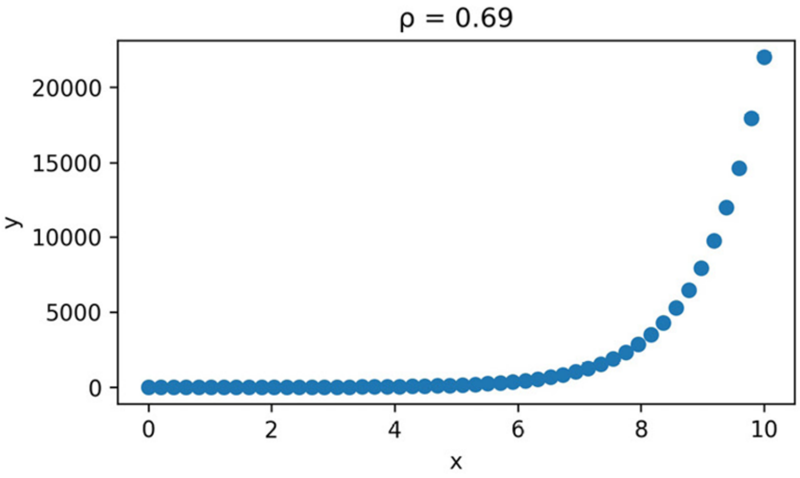
- We can also see how the points form a line:



Descriptive statistics

- Both of the following plots depict data with strong positive correlations, but it's pretty obvious when looking at the scatter plots that these are not linear.
- The one on the left is logarithmic, while the one on the right is exponential:

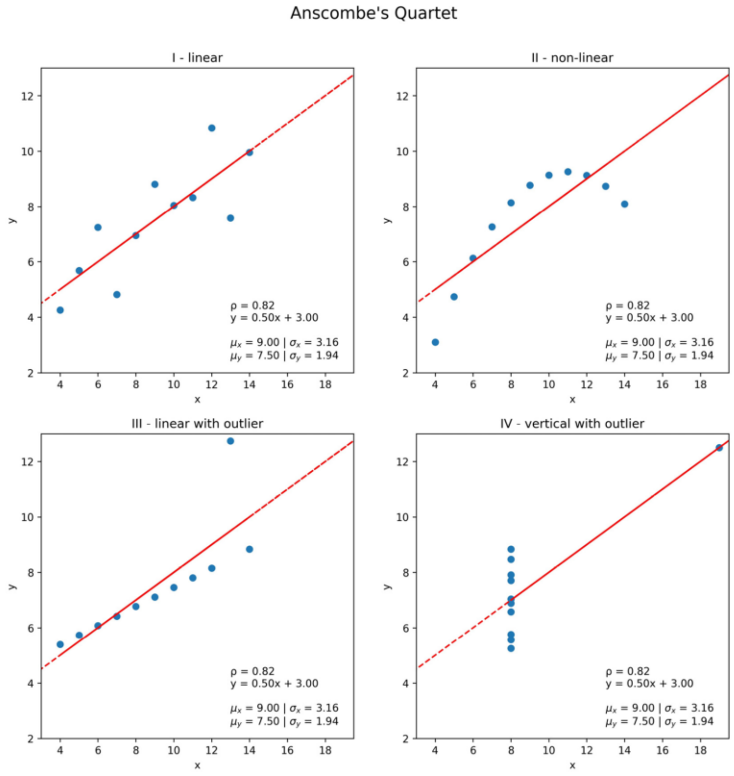




Descriptive statistics

Pitfalls of summary statistics

- Anscombe's quartet is a collection of four different datasets that have identical summary statistics and correlation coefficients, but when plotted, it is obvious they are not similar:



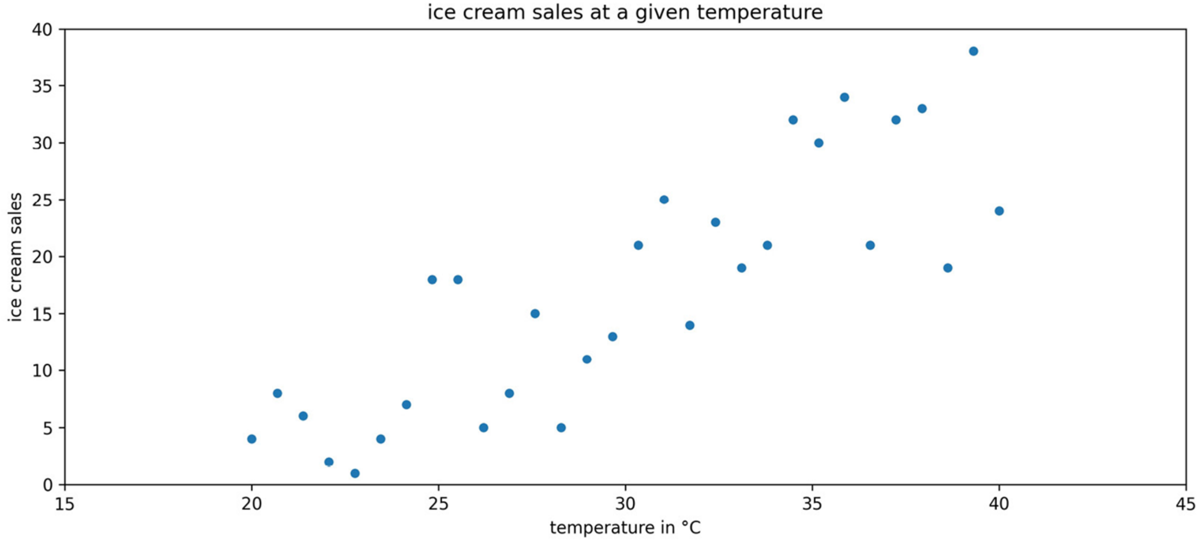
Prediction and forecasting

- Say our favorite ice cream shop has asked us to help predict how many ice creams they can expect to sell on a given day.

- They are convinced that the temperature outside has a strong influence on their sales, so they have collected data on the number of ice creams sold at a given temperature.

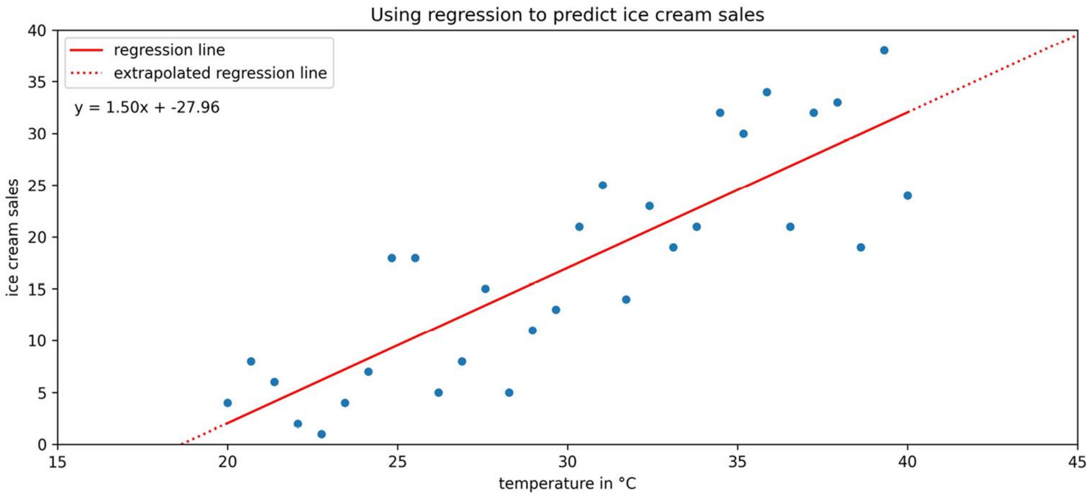
Prediction and forecasting

- We agree to help them, and the first thing we do is make a scatter plot of the data they collected:



Prediction and forecasting

- While we can have many independent variables, our ice cream sales example only has one: temperature.
- Therefore, we will use simple linear regression to model the relationship as a line:



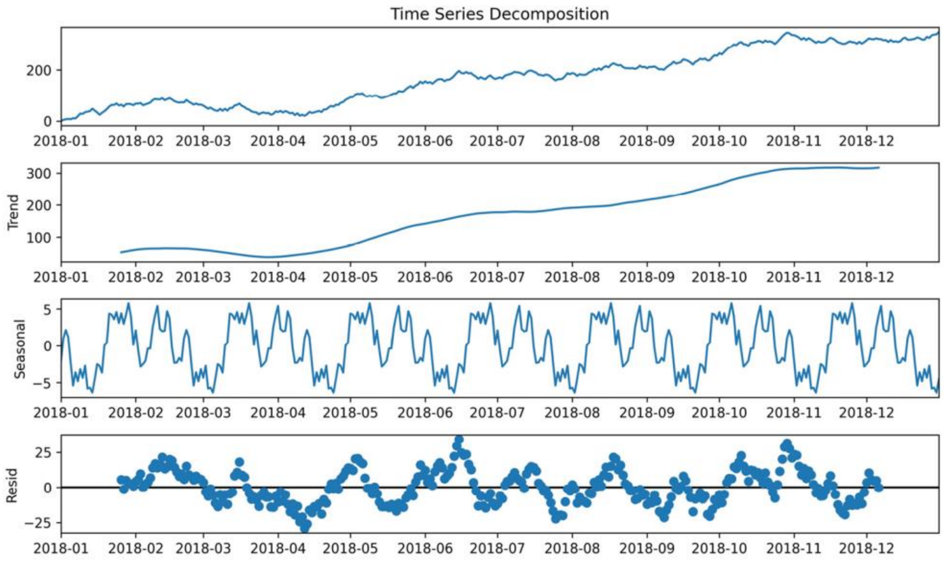
Prediction and forecasting

- The regression line in the previous scatter plot yields the following equation for the relationship:

$$ice\ cream\ sales = 1.50 \times temperature - 27.96$$

Prediction and forecasting

- We can use Python to decompose the time series into trend, seasonality, and noise or residuals.



Inferential statistics

- With an experiment, we are able to directly influence the independent variable and randomly assign subjects to the control and test groups, such as A/B tests (for anything from website redesigns to ad copy).

Setting up a virtual environment

- This course was written using Python 3.7.3, but the code should work for Python 3.7.1+, which is available on all major operating systems.

- In this section, we will go over how to set up the virtual environment in order to follow along with this course.

Virtual environments

- A virtual environment allows us to create separate environments for each of our projects.
- Each of our environments will only have the packages that it needs installed.
- It's good practice to make a dedicated virtual environment for any projects we work on.

Virtual environments

Venv

- Python 3 comes with the venv module, which will create a virtual environment in the location of our choice.

The process of setting up and using a development environment is as follows (after Python is installed):

- Create a folder for the project.
- Use venv to create an environment in this folder.
- Activate the environment.
- Install Python packages in the environment with pip.
- Deactivate the environment when finished.

Virtual environments

- To make a new directory and move to that directory, we can use the following command:

```
$ mkdir my_project && cd my_project
```

Virtual environments

- Before moving on, use cd to navigate to the directory containing this course's repository.

- Note that the path will depend on where it was cloned/downloaded:

```
$ cd path/to/Directory
```

Virtual environments

Windows

- To create our environment for this course, we will use the venv module from the standard library.
- Note that we must provide a name for our environment (course_env).
- Remember, if your Windows setup has python associated with Python 3, then use python instead of python3 in the following command:

```
C:\...> python3 -m venv course_env
```

Virtual environments

- Now, we have a folder for our virtual environment named course_env inside the repository folder that we cloned/downloaded earlier.

- In order to use the environment, we need to activate it:

C:\...> %cd%\course_env\Scripts\activate.bat

Virtual environments

- Note that after we activate the virtual environment, we can see (course_env) in front of our prompt on the command line; this lets us know we are in the environment:

(course_env) C:\...>

- When we are finished using the environment, we simply deactivate it:

(course_env) C:\...> deactivate

Virtual environments

Linux/macOS

- To create our environment for this course, we will use the venv module from the standard library.

- Note that we must provide a name for our environment (course_env):

```
$ python3 -m venv course_env
```

Virtual environments

- Now, we have a folder for our virtual environment named course_env inside of the repository folder we cloned/downloaded earlier.

- In order to use the environment, we need to activate it:

```
$ source course_env/bin/activate
```

- Note that after we activate the virtual environment, we can see (course_env) in front of our prompt on the command line; this lets us know we are in the environment:

```
(course_env) $
```

Virtual environments

- When we are finished using the environment, we simply deactivate it:

```
(course_env) $ deactivate
```

Virtual environments

Conda

- Anaconda provides a way to set up a Python environment specifically for data science.

- It includes some of the packages we will use in this course, along with several others that may be necessary for tasks that aren't covered in this course (and also deals with dependencies outside of Python that might be tricky to install otherwise).

Virtual environments

- To create a new conda environment for this course, called course_env, run the following:

```
(base) $ conda create --name course_env
```

Virtual environments

- Running conda env list will show all the conda environments on the system, which will now include course_env.

- The current active environment will have an asterisk (*) next to it—by default, base will be active until we activate another environment:

```
(base) $ conda env list
# conda environments:
#
base                 *  /miniconda3
course_env              /miniconda3/envs/course_env
```

Virtual environments

- To activate the course_env environment, we run the following command:

(base) $ conda activate course_env

- Note that after we activate the virtual environment, we can see (course_env) in front of our prompt on the command line; this lets us know we are in the environment:

(course_env) $

Virtual environments

- When we are finished using the environment, we deactivate it:

```
(course_env) $ conda deactivate
```

Installing the required Python packages

- Before installing anything, be sure to activate the virtual environment that you created with either venv or conda.

- Be advised that if the environment is not activated before running the following command, the packages will be installed outside the environment:

```
(course_env) $ pip3 install -r requirements.txt
```

Why pandas?

- When it comes to data science in Python, the pandas library is pretty much ubiquitous.
- It is built on top of the NumPy library, which allows us to perform mathematical operations on arrays of single-type data efficiently.

pandas

Jupyter Notebooks

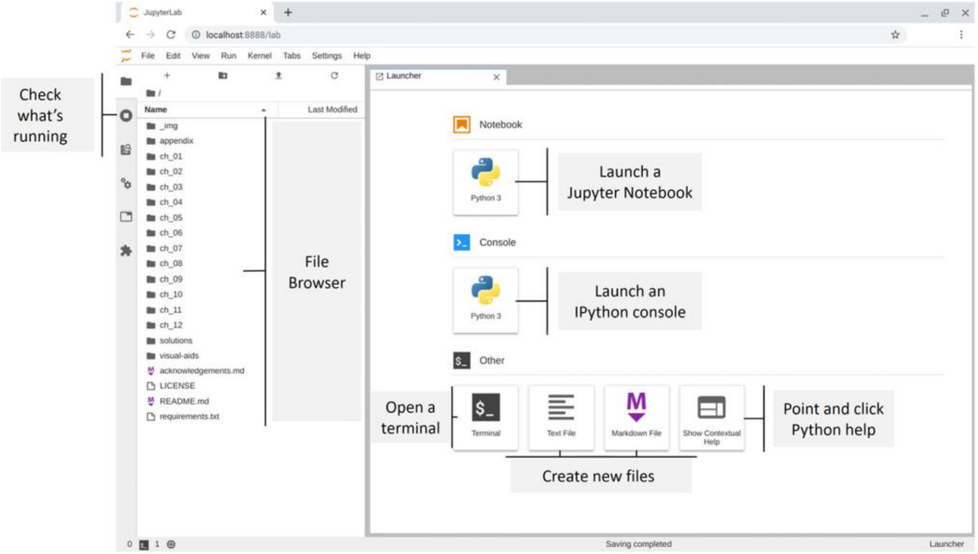
Launching JupyterLab

- JupyterLab is an IDE that allows us to create Jupyter Notebooks and Python scripts, interact with the terminal, create text documents, reference documentation, and much more from a clean web interface on our local machine.
- First, we activate our environment, and then we launch JupyterLab:

```
(course_env) $ jupyter lab
```

Jupyter Notebooks

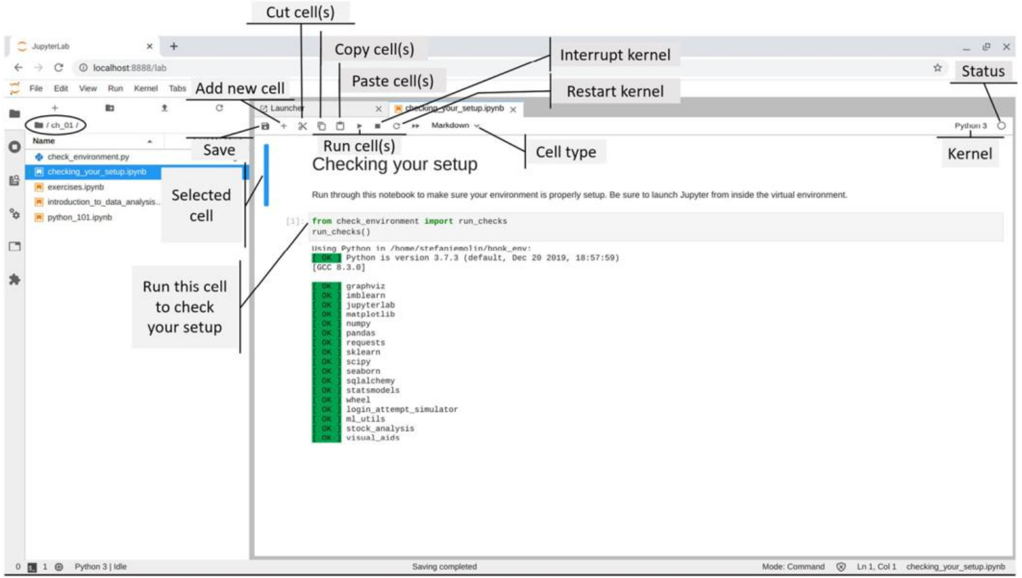
- This will then launch a window in the default browser with JupyterLab.
- We will be greeted with the Launcher tab and the File Browser pane to the left:



Jupyter Notebooks

Validating the virtual environment

- Open the checking_your_setup.ipynb notebook in the ch_01 folder, as shown in the following screenshot:



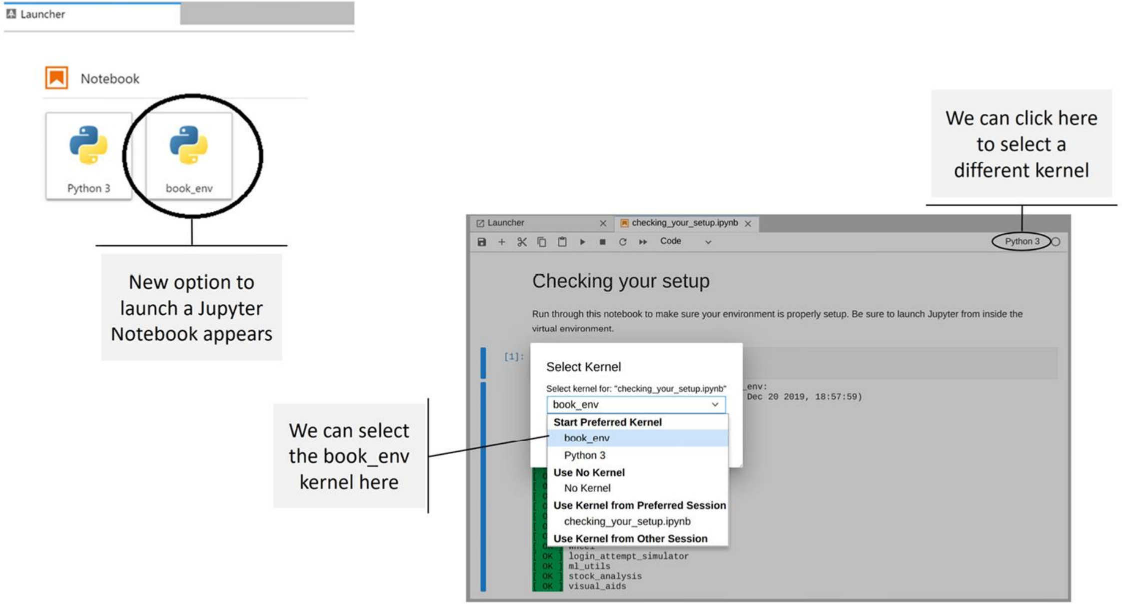
Jupyter Notebooks

- Click on the code cell indicated in the previous screenshot and run it by clicking the play (▶) button.
- If everything shows up in green, the environment is all set up.
- However, if this isn't the case, run the following command from the virtual environment to create a special kernel with the course_env virtual environment for use with Jupyter:

```
(course_env) $ ipython kernel install --user --name=course_env
```

Jupyter Notebooks

- This adds an additional option in the Launcher tab, and we can now switch to the course_env kernel from a Jupyter Notebook as well:



Jupyter Notebooks

Closing JupyterLab

- Closing the browser with JupyterLab in it doesn't stop JupyterLab or the kernels it is running (we also won't get the command-line interface back).

- To shut down JupyterLab entirely, we need to hit Ctrl + C (which is a keyboard interrupt signal that lets JupyterLab know we want to shut it down) a couple of times in the terminal until we get the prompt back:

```
...
[I 17:36:53.166 LabApp] Interrupted...
[I 17:36:53.168 LabApp] Shutting down 1 kernel
[I 17:36:53.770 LabApp] Kernel shutdown: a38e1[...]b44f
(course_env) $
```

Summary

- In this lesson, we learned about the main processes in conducting data analysis: data collection, data wrangling, EDA, and drawing conclusions.
- We followed that up with an overview of descriptive statistics and learned how to describe the central tendency and spread of our data; how to summarize it both numerically and visually using the 5-number summary, box plots, histograms, and kernel density estimates
- How to scale our data; and how to quantify relationships between variables in our dataset.

"Complete Exercises"

"Complete Lab 1"

2: Introduction to Statistics

What is Statistics?

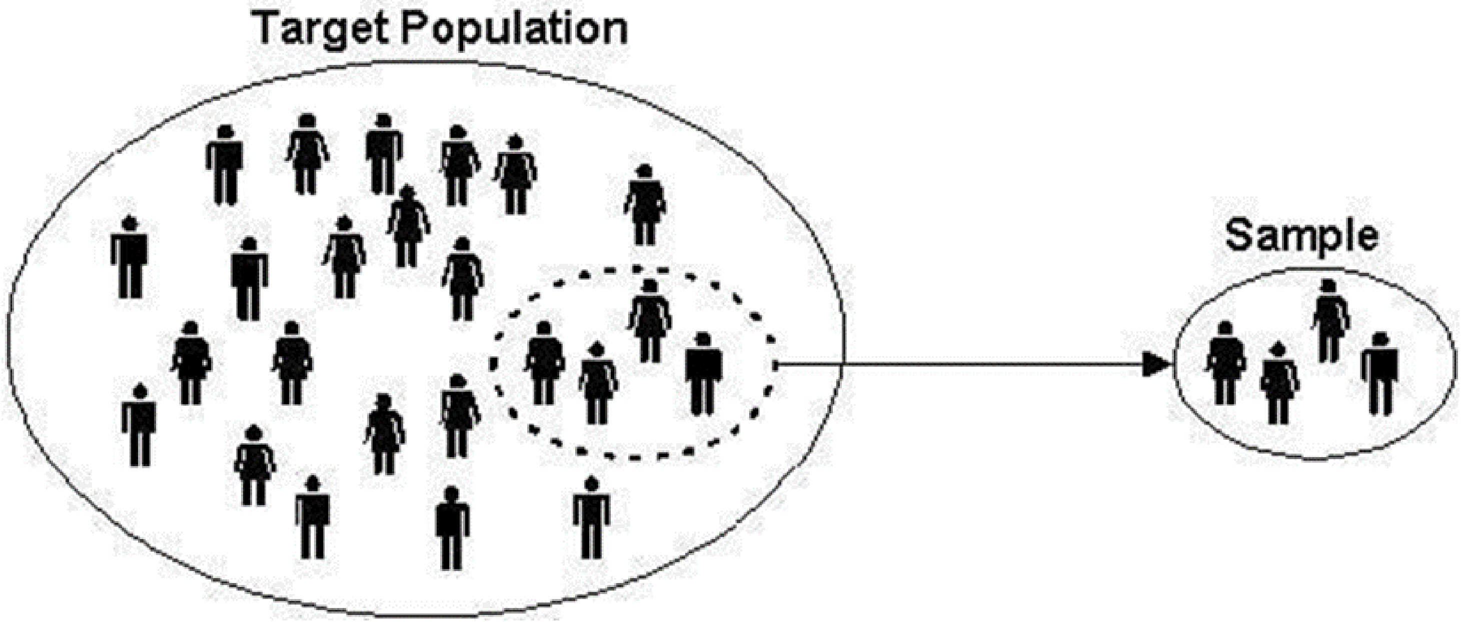
Statistics is “A telescope that allows us to study the large terrain and make it accessible to our unaided vision”

Statistics – Big Picture

Statistics provides a way of organizing data to extract information on a wider and objective basis than relying on personal experience. It is a branch of mathematics working with

- Data Gathering
- Data Understanding
- Data Analysis/Interpretation
- Data Presentation

Basic Statistical Terminology



Parameter and Statistic

Parameter: A descriptive measure of the population. For example,

- population mean - μ
- population variance – σ2
- population standard deviation - σ

Statistic: A descriptive measure of the sample. For example,

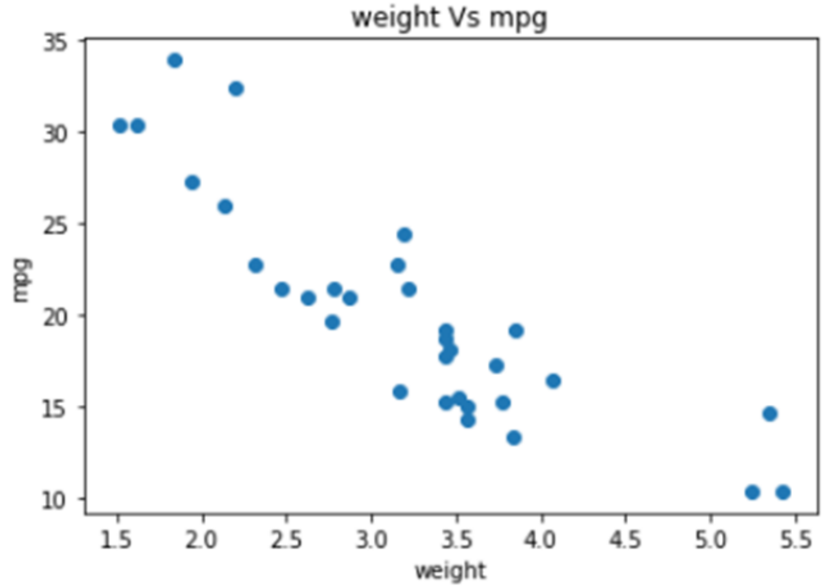
- sample mean - xbar
- sample variance - s2
- sample standard deviation – s

Variables and Data (Example of data)

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4

Variables – Dependent and Independent

- An independent variable (experimental or predictor) is a variable that is being manipulated in an experiment in order to observe the effect on dependent variable (Outcome).



Data

Data is classified into two types Numerical and Categorical

- Categorical Data
- Numerical Data

Levels of Measurement Scales

- **Nominal scale:** The nominal scale could simply be called “labels

Gender	Car Color	Name
Male	Black	Sam
Female	Red	Jack
Male	Blue	John
Female	White	Don

Levels of Measurement Scales

- **Ordinal scale:** The order of the values is what's important and significant, but the difference between each one is not really known. Here are some examples, below

Shirt Size	Feedback
Small	Poor
Medium	Good
Large	Better
Extra Large	Excellent

Descriptive Statistics

- Descriptive statistics involves organizing, summarizing, and presenting data in an informative way.
- Descriptive statistics, unlike inferential statistics, seeks to describe the data, but does not attempt to make inferences from the sample to the whole population.

Different types of Descriptive Statistics

Descriptive statistics are broken down into two categories

- Measure of Central Tendency
- Measure of Variability (Spread)

Mean:

mean
median
mode

- Mean is a central tendency of the data i.e. a number around which a whole data is spread out.

- Formula for sample mean:

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

LEARNING VOYAGE

Mean:

mean
median
mode

- Similarly, for a population data of size N, the population mean is:

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

LEARNING VOYAGE

Median:

mean
median
mode

Median is the value which divides the data into 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either ascending or descending order.

- May not exist as a data point in the set
- Influenced by position of items, but not their values
- Median is not influenced by extreme values

Mode

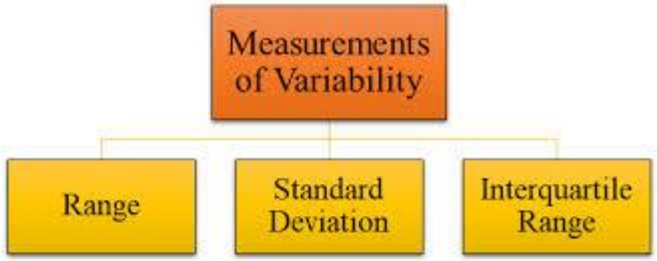
mean
median
mode

Mode: Mode is the most commonly occurring value

- Mode exists as a data point.
- Useful for qualitative data.

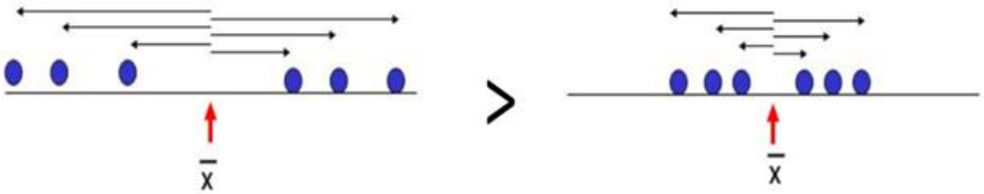
Measure of Variability (Spread / Dispersion)

- The measures that help us to know about the spread of a data set are called measures of dispersion.



Measure of Variability (Spread / Dispersion)

- **Standard deviation:** Standard deviation is the measurement of average distance between each quantity and mean, That is, how data is spread out from mean.
- A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.



Measure of Variability (Spread / Dispersion)

- Sample Standard Deviation is denoted by “S”

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

Measure of Variability (Spread / Dispersion)

- Population Standard Deviation is denoted by “σ” (sigma)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

Variance

- Variance is a square of average distance between each quantity and mean.
- That is, it is a square of standard deviation.

$$\text{Variance} = (S.D.)^2$$

Variance

The variance of Population and Sample

- The variance of a population is:

Population Mean

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

Population Size

- The variance of a sample is:

Sample Mean

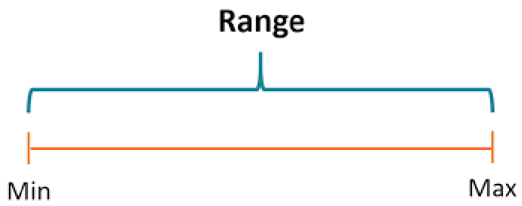
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Note! the denominator is sample size (n) minus one !

Range:

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

- It is easy to calculate.
- It is implemented for both “best” or “worst” case scenarios.
- Too sensitive for extreme values.



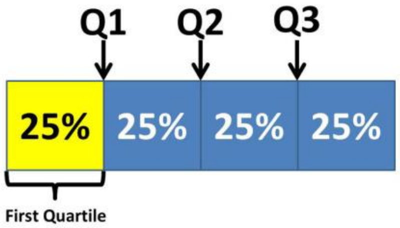
Levels of Measurement Scales

- **Percentile:** Percentile is a way to represent the position of a value in a data set.
- To calculate percentile, values in the data set should always be in ascending order.

Example:
12, 24, 41, 51, 67, 67, 85, 99

Quartile:

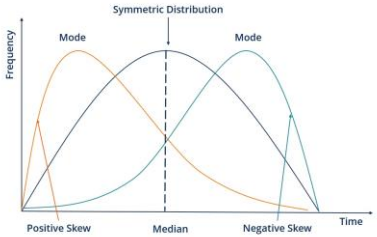
- In statistics and probability, quartile are values that divide your data into quarters provided data is sorted in an ascending order.



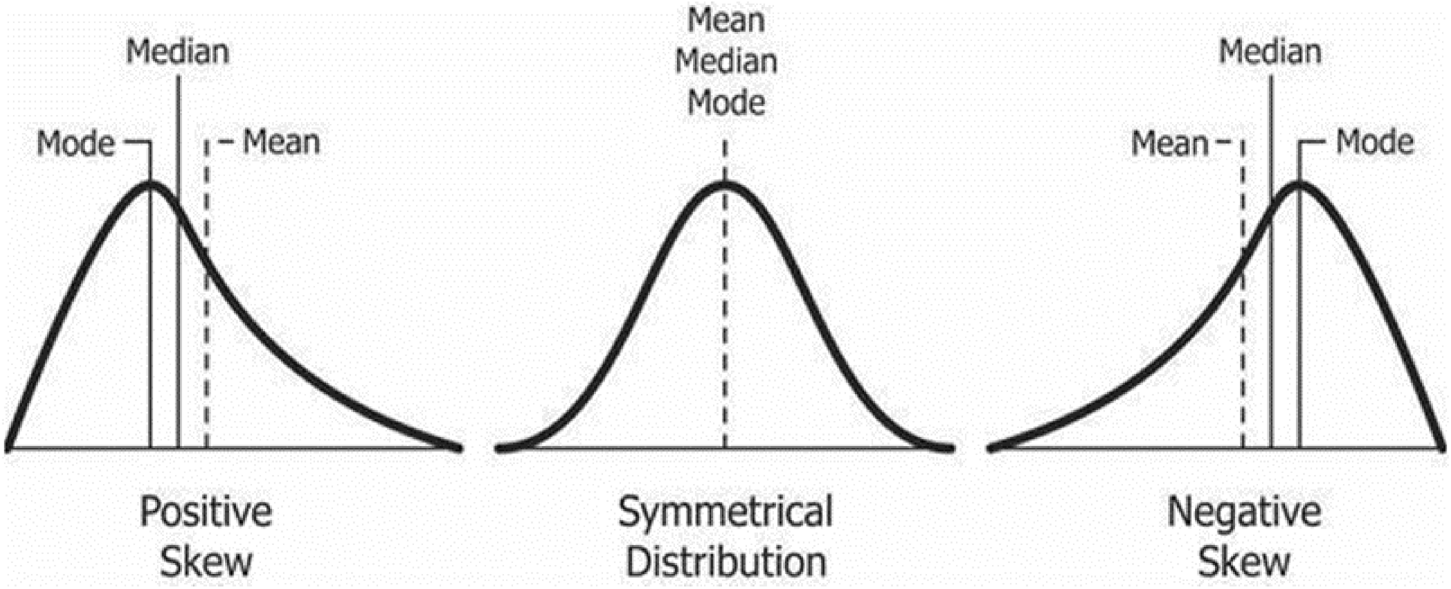
Skewness:

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

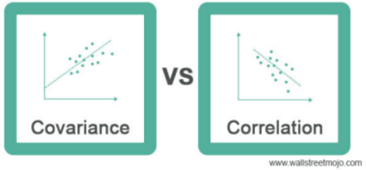
- The skewness value can be positive or negative or undefined.



Skewness:


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Covariance and Correlation



- Covariance studies the direction between two continuous variables and Correlation studies the direction and strength between two continuous variables and helps in understanding how strongly those two continuous variables are associated with each other.

What is Covariance Matrix?

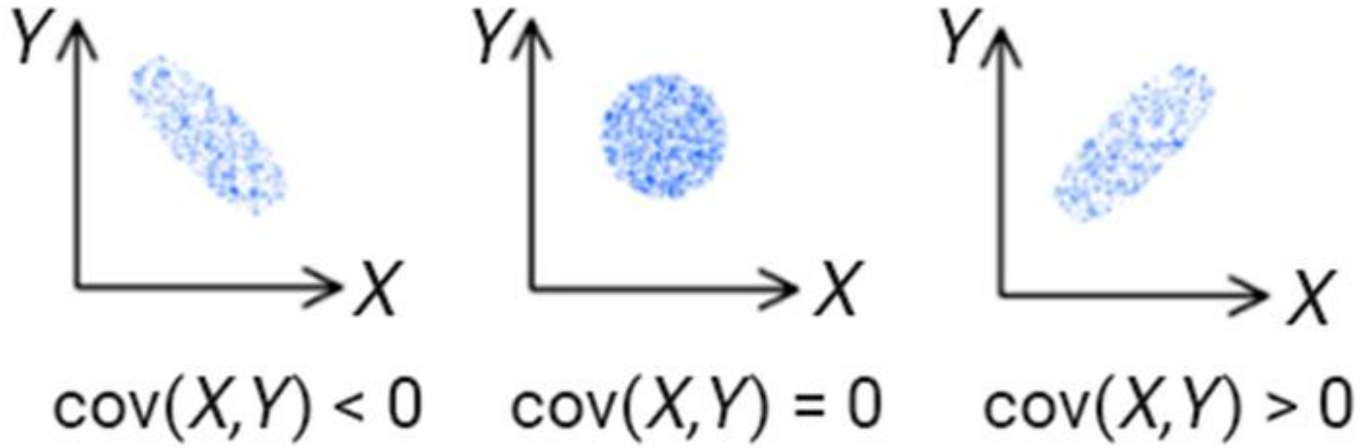
- Suppose we have two variables X and Y, then the covariance between these two variables is represented as Cov (X,Y).

- If ∑(X) and ∑(Y) are the expected values of the variables, the covariance formula can be represented as:

$$\mathbf{COV(X,\ Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

What is Covariance Matrix?

- Here are some plots that highlight how the covariance between two variables could look like in different directions.

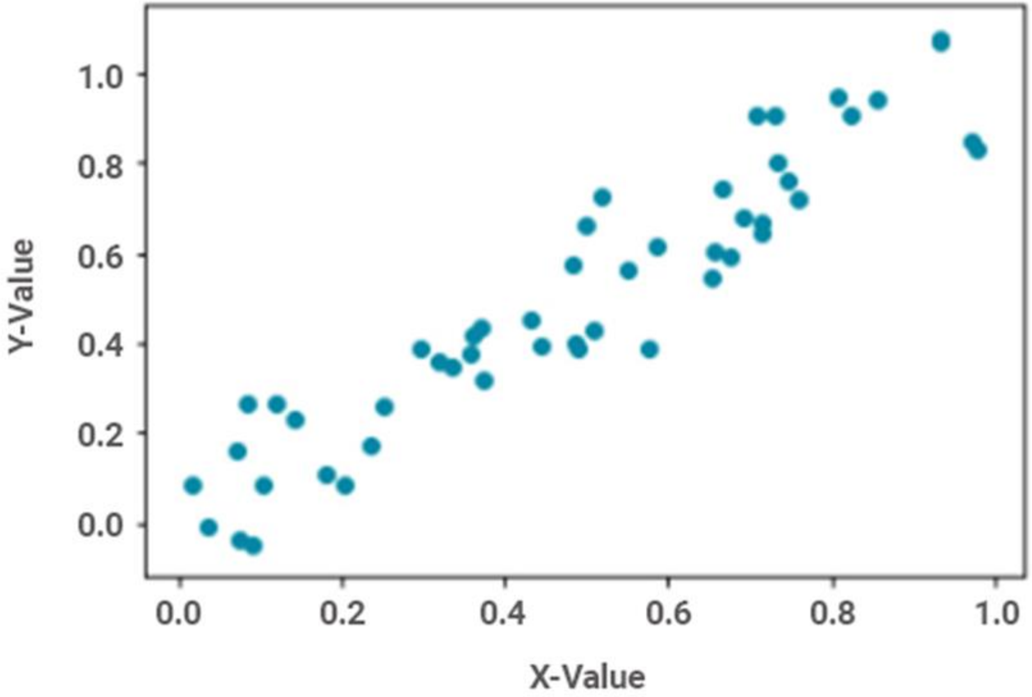


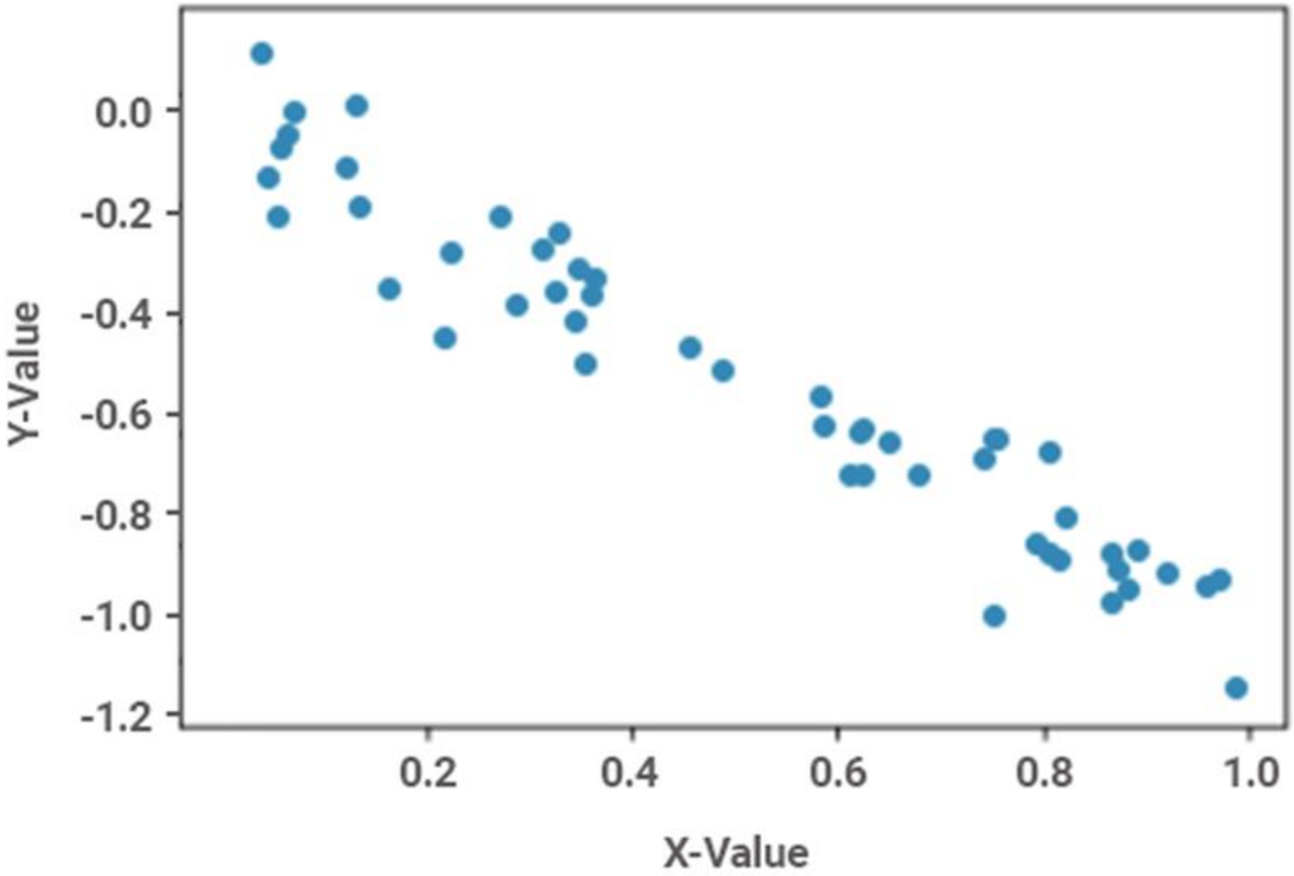
What is a Correlation Matrix?

- A correlation matrix is used to study the strength of a relationship between two variables.
- It not only shows the direction of the relationship, but also shows how strong the relationship is.
- The correlation formula can be represented as:

$$\mathrm{COR}(X, Y) = \frac{\mathrm{COV}(X, Y)}{\sqrt{\mathrm{VAR}(X)\mathrm{VAR}(Y)}}$$

What is Covariance Matrix?





Y-Value
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
0.2
0.4
0.6
0.8
1.0
X-Value

"Complete Lab 2"

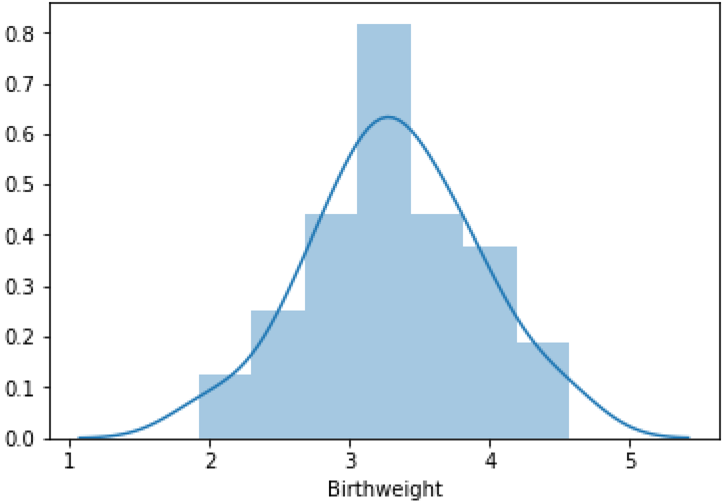
"Complete Case Study"

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
ID	42.0	894.071429	467.616186	27.00	537.25	821.000	1269.5000	1764.00
Length	42.0	51.333333	2.935624	43.00	50.00	52.000	53.0000	58.00
Birthweight	42.0	3.312857	0.603895	1.92	2.94	3.295	3.6475	4.57
Headcirc	42.0	34.595238	2.399792	30.00	33.00	34.000	36.0000	39.00
Gestation	42.0	39.190476	2.643336	33.00	38.00	39.500	41.0000	45.00
smoker	42.0	0.523810	0.505487	0.00	0.00	1.000	1.0000	1.00
mage	42.0	25.547619	5.666342	18.00	20.25	24.000	29.0000	41.00
mnocig	42.0	9.428571	12.511737	0.00	0.00	4.500	15.7500	50.00
mheight	42.0	164.452381	6.504041	149.00	161.00	164.500	169.5000	181.00
mppwt	42.0	57.500000	7.198408	45.00	52.25	57.000	62.0000	78.00
fage	42.0	28.904762	6.863866	19.00	23.00	29.500	32.0000	46.00
fedyrs	42.0	13.666667	2.160247	10.00	12.00	14.000	16.0000	16.00
fnocig	42.0	17.190476	17.308165	0.00	0.00	18.500	25.0000	50.00
fheight	42.0	180.500000	6.978189	169.00	175.25	180.500	184.7500	200.00
lowbwt	42.0	0.142857	0.354169	0.00	0.00	0.000	0.0000	1.00
mage35	42.0	0.095238	0.297102	0.00	0.00	0.000	0.0000	1.00

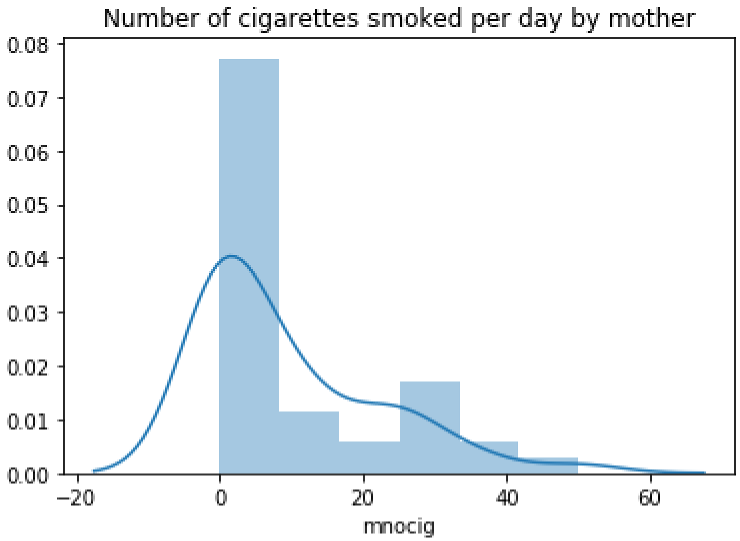
- We can analyze the distribution of the birth weight variable.

```
#plot distibutions of birth weight
sns.distplot(birth_weight['Birthweight'], label="Birth Weight")
```

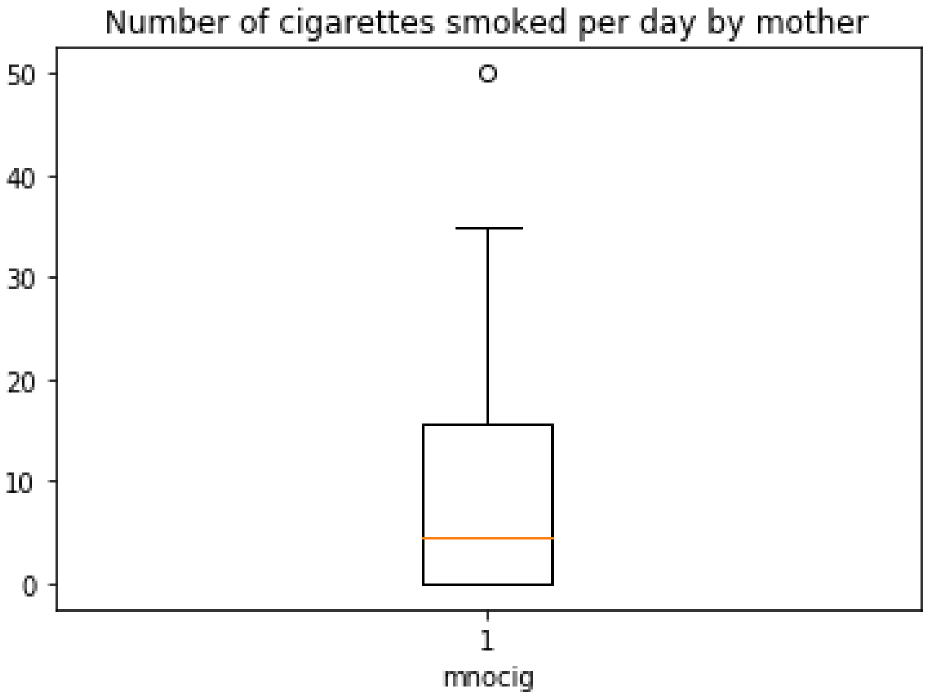


- Let’s analyze the distribution of “mnocig” (Number of cigarettes smoked per day by mother) variable

```
#plot distribution of Number of cigarettes smoked per day by mother
sns.distplot(birth_weight['mnocig'])
plt.title("Number of cigarettes smoked per day by mother")
```



Descriptive Statistics



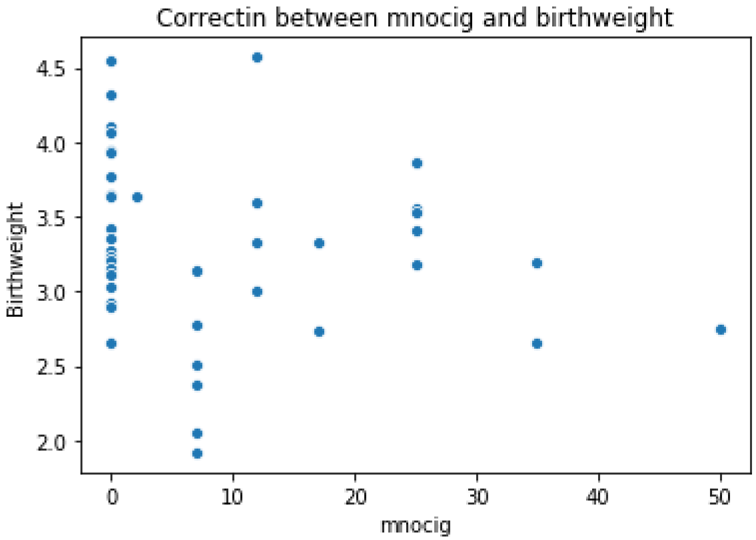
Descriptive Statistics

```
mnocig_46 = np.percentile(birth_weight['mnocig'], 46)
mnocig_75 = np.percentile(birth_weight['mnocig'], 75)
mnocig_90 = np.percentile(birth_weight['mnocig'], 90)

print("46th percentile: ", round(mnocig_46, 0))
print("75th percentile: ", round(mnocig_75, 0))
print("90th percentile: ", round(mnocig_90, 0))
```

```
46th percentile:  0.0
75th percentile:  16.0
90th percentile:  25.0
```

```python
#Correlation between birthweight and mnocig
sns.scatterplot(birth_weight['mnocig'], birth_weight['Birthweight'])
plt.title("Correctin between mnocig and birthweight")
```




```python
#correlation value
birth_weight['Birthweight'].corr(birth_weight['mnocig'])
```

```
-0.1523351844506074
```

SUMMARY

- Statistics deals with collecting, interpreting, and drawing a conclusion from the data.

- Data is measured on different scales like nominal, ordinal, interval and ratio.

- Descriptive statistics aims to summarize a sample data with a single value with the help of mean, median and mode.

"Complete Assessment"

DAY 2

3: Probability Distributions

Probability Distributions

- Probability implies ‘likelihood’ or ‘chance’.
- When an event is certain to happen then the probability of occurrence of that event is 1 and when it is certain that the event cannot happen then the probability of that event is 0.

Assigning Probabilities

- **Classical method** – A prior or Theoretical Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ event\ occurs}{total\ possible\ \#\ of\ outcomes}$$

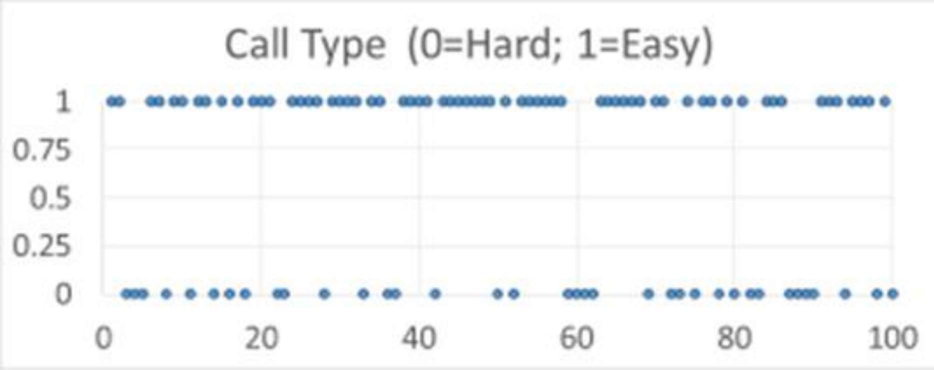
Assigning Probabilities

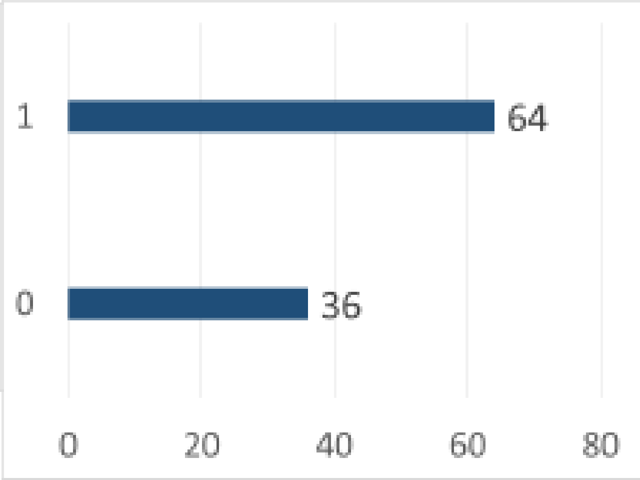
Experiment: Tossing of a fair dice

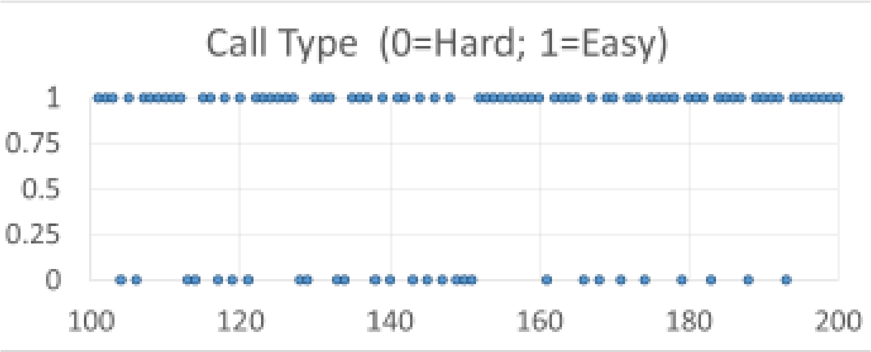
Assigning Probabilities

- Empirical Method – A posteriori or Frequentist
  Probability can be determined post conducting a thought experiment.

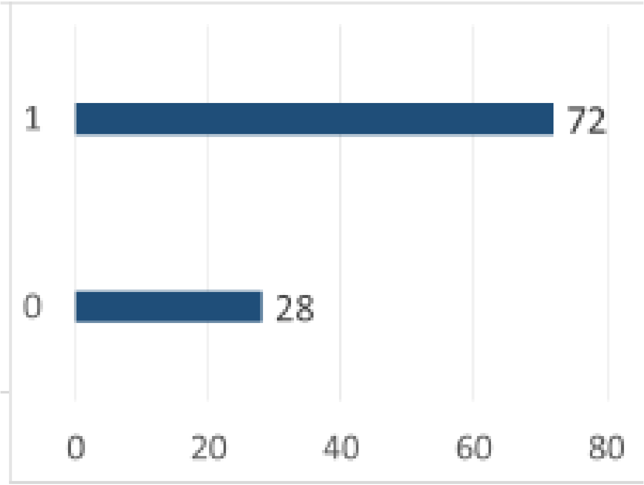
$$P(E) = \frac{\#\ of\ times\ an\ event\ occurred}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$


Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
0
20
40
60
80
100

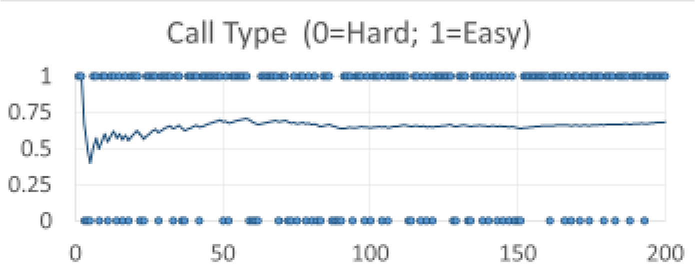

1
64
0
36
0
20
40
60
80



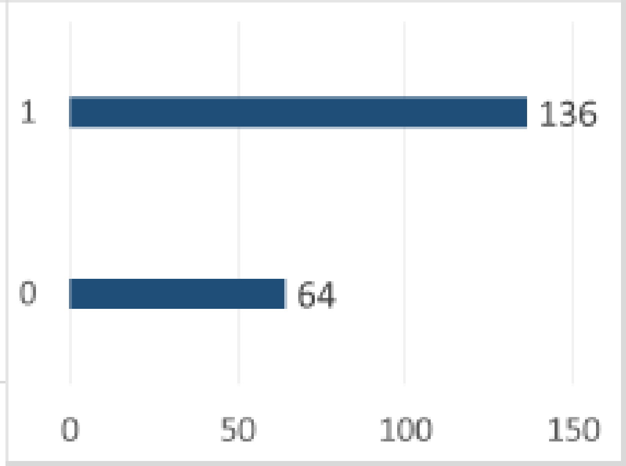
Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
100
120
140
160
180
200



1
72
0
28
0
20
40
60
80



Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
0
50
100
150
200



1
136
0
64
0
50
100
150

P (easy) = 0.7

Probability Terminology

- Sample Space – Set of all possible outcomes, denoted S.

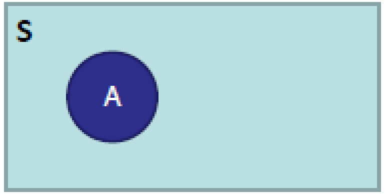
Example: After 2 coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

- Event – A subset of the samples space.

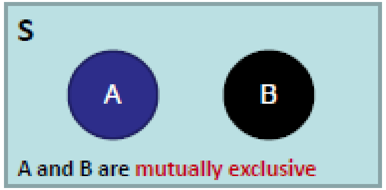
An Event of interest might be – HH

Probability – Rules

$P(S) = 1$



$0 <= P(A) <= 1$



$P(A \text{ or } B) = P(A) + P(B)$

Mutually Exclusive

When two events (call them “A” and “B”) are Mutually Exclusive than it is impossible for them to happen together.

- If A and B are mutually exclusive

$$P\ (A \text{ and } B) = 0$$

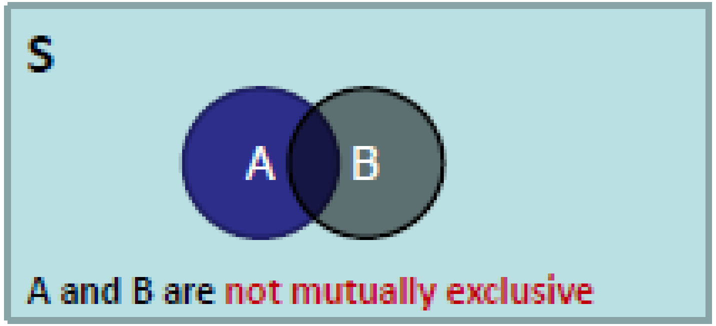
- But the probability of A or B is the sum of the individual probabilities.

$$P\ (A \text{ or } B) = P\ (A) + P\ (B)$$

LEARNING VOYAGE

Mutually Exclusive

- When we combine those two events: P (King or Queen) = (1/13) + (1/13) = 2/13



P (A or B) = P (A) + P (B) – P (A and B)

Mutually Non-Exclusive Events

- Two events A and B are said to be mutually non-exclusive events if both the events A and B have at least one common outcome between them.

Probability – Types

- Contingency table summarizing 2 variables, Loan Default and Age:

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

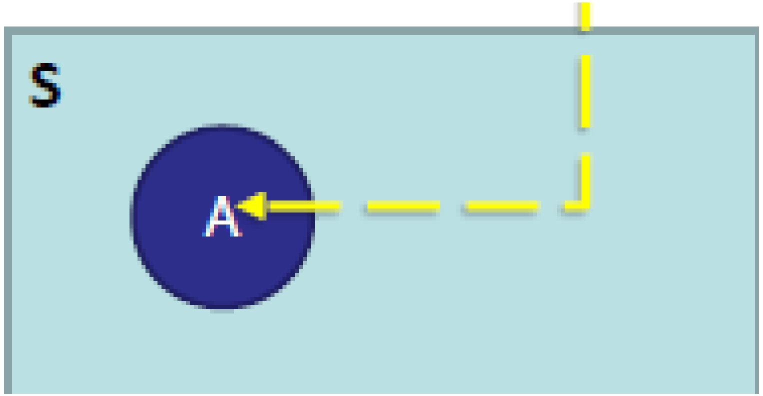
Probability – Types

- Convert it into probabilities:

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.00

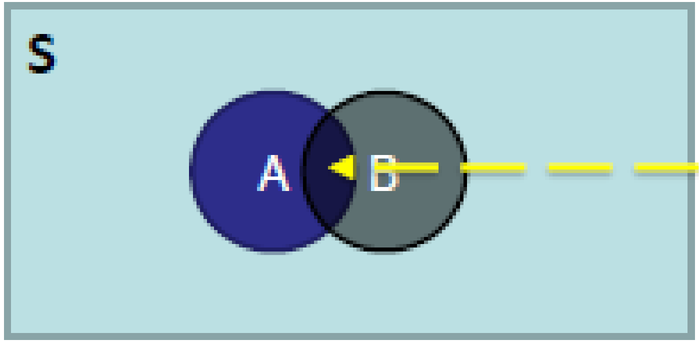
Probability – Types

- Marginal Probability: Probability describing a single attribute



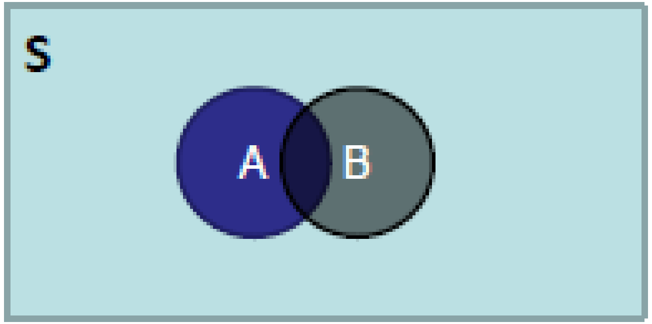
Probability – Types

Joint Probability: Probability describing a combination of attributes



Probability – Types

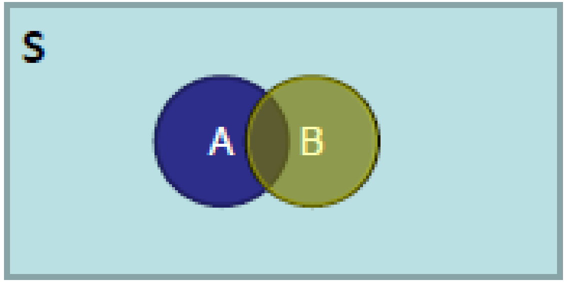
- Union Probability: Probability describing a new set that contains all of the elements that are in at least one of the two sets.



Conditional Probability

The probability of an event (A), given that another event (B) has already occurred.

- The sample space is restricted to a single row or column. This makes the rest of the sample irrelevant.



Example:

What is the probability that a person will not default on the loan payment given he/she is middle-age?

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.69	0.008	1.00

Probability – Types

- Note that this is the ratio of Joint Probability to Marginal Probability, i.e.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Probability – Types

- Equating, we get

P (A/B) * P (B) = P (A) * P (B/A)

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

- Now, given that the probability that someone defaults on a loan is 0.184, find the probability that an older person defaults on the loan.
- Older people make up only 0.8% of the clientele. P (Yes/Old) =?

P (Yes/Old) = (P (Yes) * P (Old/Yes))/P (Old)

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

LEARNING VOYAGE

Histogram:

A series of contiguous rectangles that represent the frequency of data in the given class intervals.

How many class intervals?

- Rule of thumb: 5-15 (not too many and not too few)

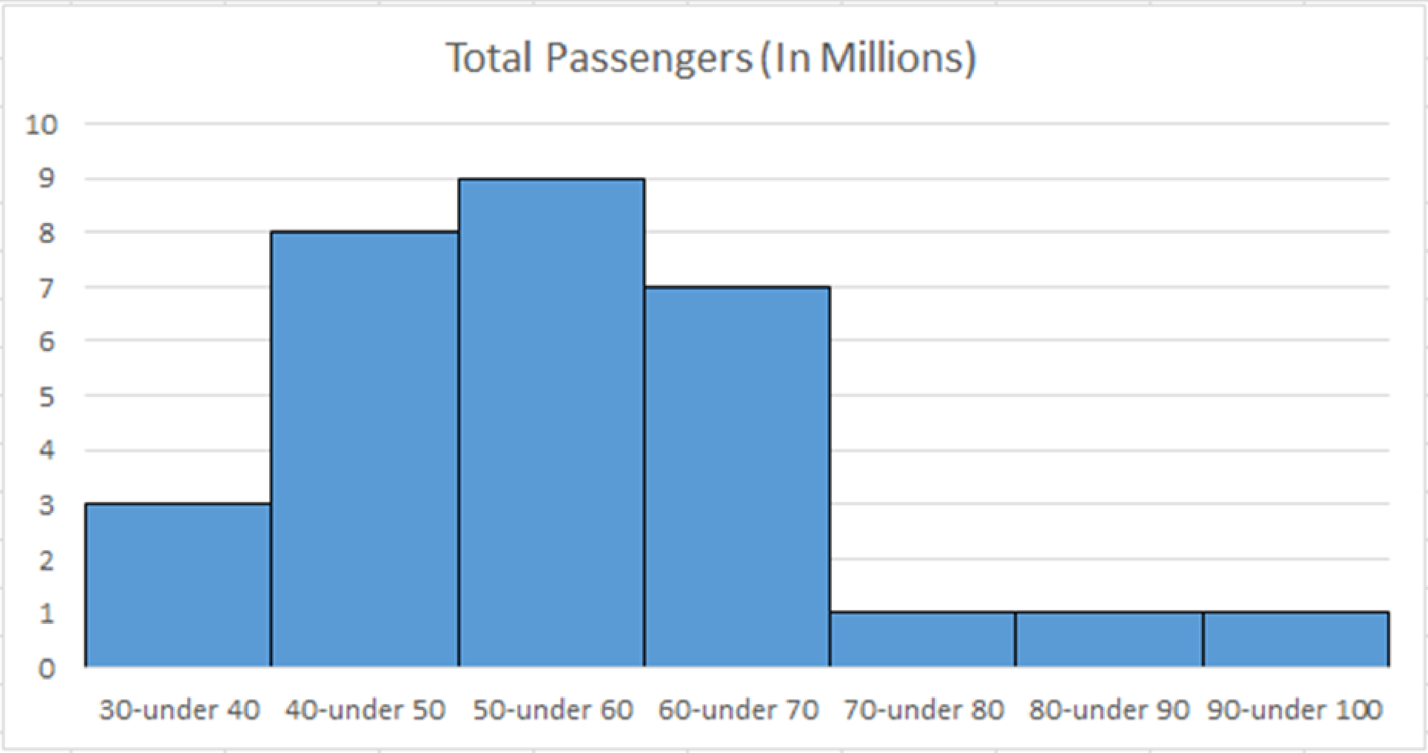
The Freedman-diaconis rule:

$$No.of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

Histogram – Excel

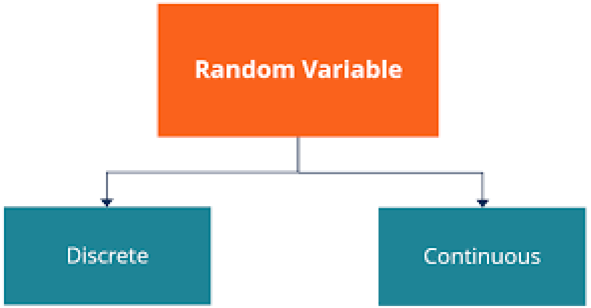
Passenger Traffic 2013 FINAL (Annual)			
Last Update: 22 December 2014			
Passenger Traffic			
Total passengers enplaned and deplaned, passengers in transit counted once			
Rank	**City (Airport)**	**Passengers 2013**	**% Change**
1	ATLANTA GA, US (ATL)	94,431,224	-1.1
2	BEIJING, CN (PEK)	83,712,355	2.2
3	LONDON, GB (LHR)	72,368,061	3.3
4	TOKYO, JP (HND)	68,906,509	3.2
5	CHICAGO IL, US (ORD)	66,777,161	0.2
6	LOS ANGELES CA, US (LAX)	66,667,619	4.7
7	DUBAI, AE (DXB)	66,431,533	15.2
8	PARIS, FR (CDG)	62,052,917	0.7
9	DALLAS/FORT WORTH TX, US (DFW)	60,470,507	3.2
10	JAKARTA, ID (CGK)	60,137,347	4.1
11	HONG KONG, HK (HKG)	59,588,081	6.3
12	FRANKFURT, DE (FRA)	58,036,948	0.9
13	SINGAPORE, SG (SIN)	53,726,087	5
14	AMSTERDAM, NL (AMS)	52,569,200	3
15	DENVER CO, US (DEN)	52,556,359	-1.1
16	GUANGZHOU, CN (CAN)	52,450,262	8.6
17	BANGKOK, TH (BKK)	51,363,451	-3.1
18	ISTANBUL, TR (IST)	51,304,654	13.7
19	NEW YORK NY, US (JFK)	50,423,765	2.3

Histogram – Excel



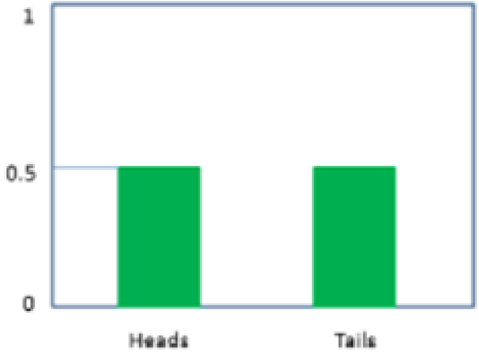
Random Variable

- A Random Variable is a set of possible values from a random experiment.



Discrete Random Variable

- The discrete random variable is a variable that may take on only a countable number of distinct values.



Countable

Probability Distributions

Types of Discrete Probability Distributions

- Bernoulli Distribution.
- Binomial Distribution.
- Poisson Distribution.

Bernoulli distribution

- A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial – a random experiment that has only two outcomes (usually called “Success” or “Failure”.

Binomial Distribution

- A binomial distribution is the probability of the "success" or "failure" outcome of an experiment or survey that is repeated multiple times.
- A binomial distribution is the probability of a "success" or "failure" outcome in an experiment or survey that is repeated multiple times.

Notation: X ~ Bio(n, P)

n: number of times the experiment runs

p: probability of one specific outcome

Probability Mass Function:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n - x}$$

Where:

- b = binomial probability.
- x = total number of “success”.
- P = probability of success on an individual trail.
- n = number of trails.

Mean and Variance of Binomial distribution:

$$E\,(X) = np$$
$$Var\,(X) = npq$$

Criteria – Binominal distribution must meet the following three criteria:

- The number of trials is fixed.
- Each trail is independent of others.
- The probability of “success” (trail, head, fail or pass) is the same from one trail to another.

Poisson distribution

- The Poisson distribution is the discrete probability distribution of the number of events occurring in a given time period, provided that the events occur at a constant mean rate and are independent of the time since last event.

Probability Mass Function:

$$P(X) = \frac{e^{-\mu}\mu^{x}}{x!}$$

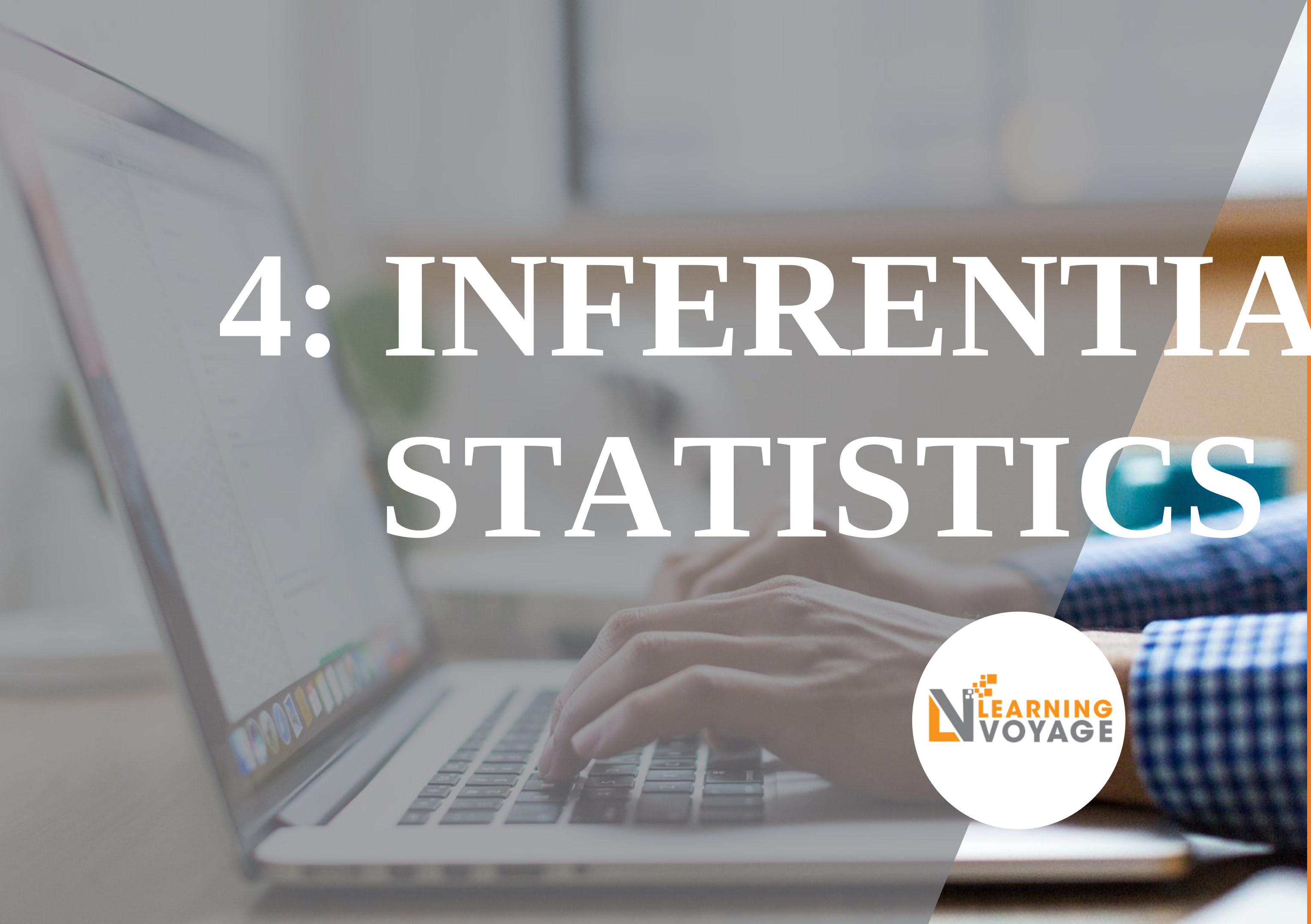
Where:

- The symbol “!” is a factorial.
- M (The expected number of occurrences) is sometimes written as λ. It is sometimes called the event rate or rate parameter.

"Complete Lab 3"

"Complete Case Study"

"Complete Programming Assignment "



4: INFERENTIAL STATISTICS

LEARNING VOYAGE