We’ll be starting shortly!

To help us run the workshop smoothly, please kindly:

- Switch off screen sharing and mute your microphone
- Submit all questions using the Q&A function
- If you have an urgent request, please use the “Raise Hand” function

Thank you!

{<oding:lab}

Shopee CODE LEAGUE 2020

Shopee CODE LEAGUE 2020

{<oding:lab}

Data Science 101

Session 2

About Coding Lab

- Founded by an MIT Graduate who worked in Silicon Valley
- Global Tech advisory team based in New York, Japan and Singapore
- We have Campuses in Japan, Australia and Singapore
- We offer coding classes starting from age 4 to adulthood

{<coding:lab}

Shopee CODE LEAGUE 2020

Features and Partners

PSA
The World's Port of Call

CISCO

NUS
National University of Singapore

Zendesk

UBISOFT

{<oding:lab}

Shopee
CODE LEAGUE
2020

Features and Partners

Shopee
{coding:lab}
1439
I'M THE BEST CODER! CHALLENGE
Training Workshop with Coding Lab (Pre-Tertiary)
15 SEP 2019

Shopee

Ministry of Education
SINGAPORE

THE STRAITS TIMES

zaobao sg
早 晚 全 新

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

企業経営の新潮流を読み解く
Bizコンパス

YOUNG PARENTS

mediacorp

BRAND'S

CGTN

Sassy Mama SG

parentsworld

And many more...

{<coding:lab}

Shopee CODE LEAGUE 2020

Meet our Students

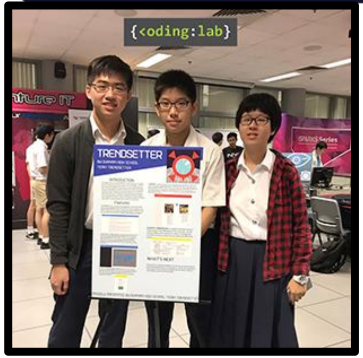
Sarah, 18
Honourable
Mention, NOI 2018

Team ajdisjd
1st Place, iCode
2019

{<coding:lab}
SPECIAL AWARD – YOUNGEST MEDALIST
YONG JUN XIANG, ELIJAH
ANGLO-CHINESE SCHOOL INDEPENDENT

Elijah, 14
Youngest Medalist,
NOI 2019

Surya, 14
Created a stock
rating algorithm



Team Trendsetter
Best Presentation
Award

More schools we have taught at:

eunoia
JUNIOR COLLEGE

EtonHouse
International School • Pre-School

MMI Modern Montessori International (LONDON)

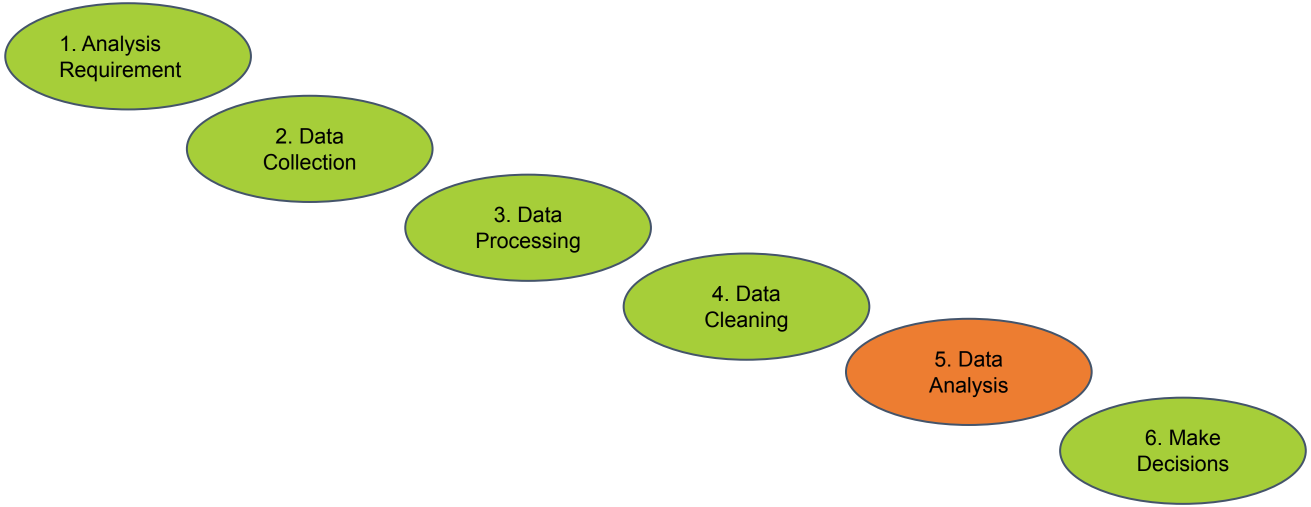
{<coding:lab}

Shopee
CODE
LEAGUE
2020

Let’s get started - Session Overview

- Analytics and Statistics
  - Introduction to Statistics
  - Descriptive Statistics
  - Inferential Statistics
  - Application of Statistics
  - Simple Linear Regression

Data Analytics Process Overview



1. Analysis Requirement
2. Data Collection
3. Data Processing
4. Data Cleaning
5. Data Analysis
6. Make Decisions

{<coding:lab}

Shopee CODE LEAGUE 2020

Introduction to Statistics

{<oding:lab}

Descriptive vs Inferential Statistics

- In descriptive statistics, we use the data that we collect to provide descriptions of the population, either through numerical calculations or graphs or tables
- In Inferential statistics, we make inferences and predictions about an entire population based on a sample of data taken from the population of interest
  - Here, we do not have the full dataset from the entire population

{<oding:lab}

Descriptive Statistics Scenario

- Suppose we want to find out the average number of hours each student in a particular class spends studying at home per week
  - Say there are 20 students in the class
- Since there are only 20 students, it is possible ask them individually and then average out the numbers

Inferential Statistics Scenario

- Suppose we find out that each student spends an average of 8 hours a week studying at home
- What if we are now interested in the entire school?
  - Say there are 3000 students in the school
  - Would it be possible to ask them all 1 by 1?
    - Is it cost effective?
  - We would then have to take sample of data from the school, rather than asking every single person in school

Descriptive Statistics

What Are Descriptive Statistics For?

- Describe what is going on in a data set
    - Allow us to see patterns that exist in our data
- Can only be used to describe the data set that we are studying
    - Simpler interpretation of data

Types of Descriptive Statistics

- Types of Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability

Measures of Central Tendency

- There are 3 Measures of Central Tendency
  - Mean
  - Median
  - Mode

Mean

- The mean is the average of the numbers.
  - It is easy to calculate: add up all the numbers, then divide by how many numbers there are
  - In other words it is the **sum** divided by the **count**
- E.g. Given the numbers 10, 15 and 26, the mean is (10 + 15 + 26) / 3 = 17

{<oding:lab}

Median

- The median is the value in the middle of a sorted list of number
- E.g. Given the numbers 7, 2, 15, 42, 5, 78, 22, 8, 24
  - We first sort the numbers in order: 2, 5, 7, 8, 15, 22, 24, 42, 78
  - The median is the number exactly in the middle
  - In this case the median is 15
- In the case where we have an even number of data, we take the mean of the two values in the middle

Mode

- Mode is the measure of central tendency that identifies the value that occurs most frequently
- E.g. Given the numbers 10, 5, 6, 7, 9, 9, 13, 5, 8, 5
    - We rearrange the numbers in increasing order
    - 5, 5, 5, 6, 7, 8, 9, 9, 10, 13
    - The mode here is 5 as it appears the most number of times

{<oding:lab}

3Ms (Demo/Practice - 1)

- Find the mean, median and mode for the following list of values
  - 13, 18, 13, 14, 13, 16, 14, 21, 13

Checkpoint 1

- Every student must be able to:
    - Know what the 3 central measures of tendency are
    - Calculate mean, median and mode
- For students who are waiting, try the following:
    - Find out the height of your friends and calculate the mean, median and mode!

{<oding:lab}

Measures of Variability

- There are 3 Measures of Variability
  - Range
  - Interquartile Range (IQR)
  - Variance/Standard Deviation

Range

- The range is the difference between the largest and smallest values in a set of values
- Consider the following numbers: 5, 5, 5, 6, 7, 8, 9, 9, 10, 13
  - The range is 13 - 5 = 8

Interquartile Range (IQR)

- The second measure of variability is the Interquartile Range
  - Based on dividing a data set into quartiles
  - Quartiles divide a rank-ordered data set into four equal parts
  - The values that divide each part are called the first, second, and third quartiles
  - Denoted by Q1, Q2, and Q3, respectively

Quartiles

- Q1 is the "middle" value in the first half of the rank-ordered data set
- Q2 is the median value in the set
- Q3 is the "middle" value in the second half of the rank-ordered data set

Finding Interquartile Range

- Consider the following numbers: 5, 5, 5, 6, 7, 8, 9, 9, 10, 13
  - Q1 is the "middle" value in the first half of the set: 5
  - Q2 is the median: (7 + 8) / 2 = 7.5
  - Q3 is the "middle" value in the second half of the set: 9

Variance

- Variance is the average squared deviation from the population mean
  - $\sigma^2 = \Sigma (X_i - \mu)^2 / N$
  - σ^2 is the population variance
  - μ is the population mean
  - X_i is the i-th element from the population
  - N is the number of elements in the population
- Tells you the extent to which your value deviates from the population mean

{<oding:lab}

Calculating Variance

- To calculate Variance:
  - Calculate the mean
  - Subtract the mean from each value then square the result
  - Calculate the average of the squared numbers

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Calculating Variance Example

- Consider the following numbers: 5, 5, 5, 6, 7, 8, 9, 9, 10, 13
    - The mean is 7.7
    - Variance is 6.21
    - $[(5-7.7)^2*3 + (6-7.7)^2 + (7-7.7)^2 + (8-7.7)^2 + (9-7.7)^2*2 + (10-7.7)^2 + (13-7.7)^2] / 10$

Standard Deviation

- Standard deviation is the square root of the variance
  - Denoted by σ

Range, IQR, Variance, SD (Demo/Practice - 2)

- Find the Range, IQR, Variance and SD of the following values
    - 13, 18, 13, 14, 13, 16, 14, 21, 13

{<oding:lab}

Checkpoint 2

- Every student must be able to:
  - Know what the 3 Measures of Variability are
  - Calculate range , IQR, Variance and Standard Deviation
- For students who are waiting, try the following:
  - Explore the statistics package in python https://docs.python.org/3/library/statistics.html

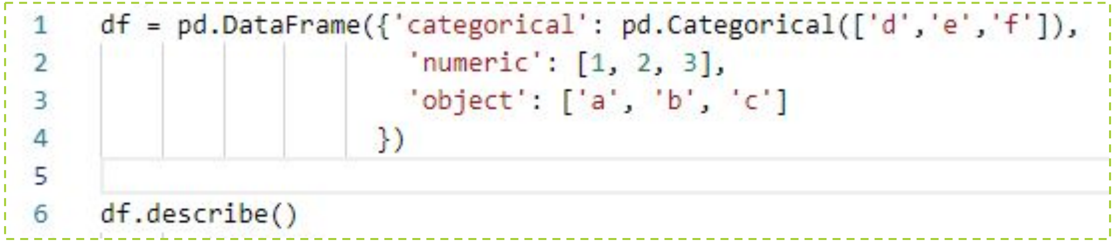
Data Analysis Summary

- Exploratory Data Analysis (EDA) is all about summarising datasets with statistics and visualisation
- However Calculating statistics pertaining to a dataset in plain Python can be a chore.
- In pandas, we simply use one special, extremely helpful method
    - df.describe()

{<oding:lab}

DataFrame.describe()

- With the DataFrame given, this operation returns the descriptive statistics which summarizes the central tendency, dispersion and shape of a dataset's distribution
- Try creating the following DataFrame and see what's the output



```
df = pd.DataFrame({'categorical': pd.Categorical(['d','e','f']),
                   'numeric': [1, 2, 3],
                   'object': ['a', 'b', 'c']
                  })

df.describe()
```

DataFrame.describe() (Demo/Practice - 3)

- Create a DataFrame and perform df.describe() method to your data

Checkpoint 3

- Every student must be able to:
  - Understand and utilise df.describe() method to obtain descriptive statistics of DataFrame
- For students who are waiting, try the following:
  - Explore the statistics package on Pandas
  - What are some useful method to take note of?

{<oding:lab}

Inferential Statistics

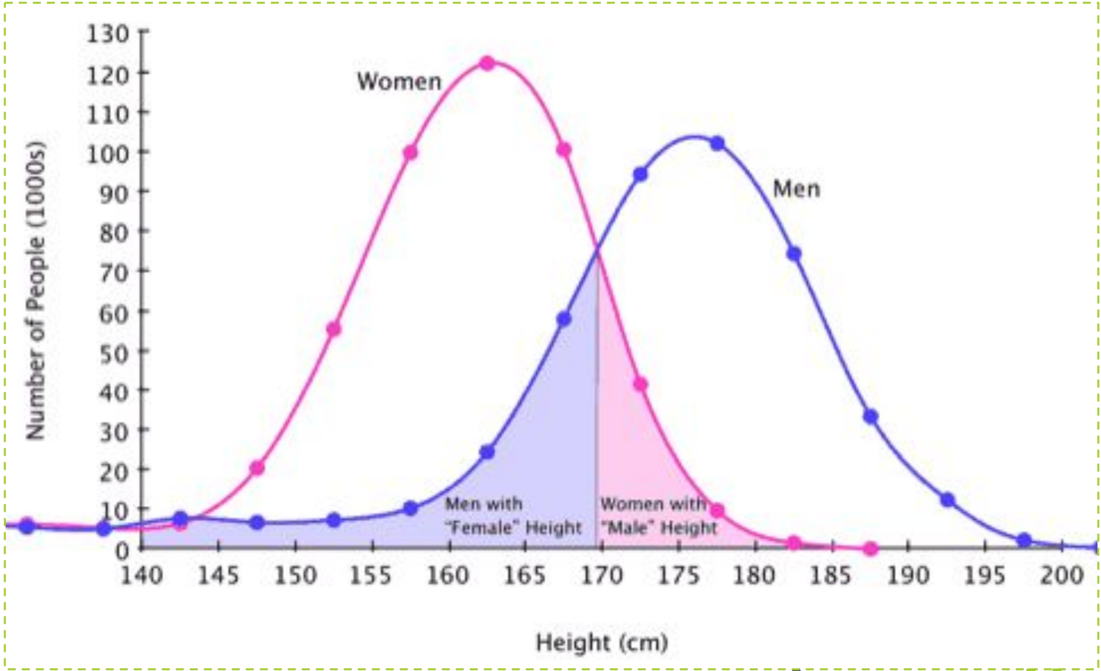
Inferential Statistics

- In inferential statistics, we take data from samples and make generalizations about a population
  - For example, we might stand outside a mall and ask a random sample of 100 people to rate the mall on a scale of 1 to 10
- Two main areas of inferential statistics
  - Estimating parameters
  - Hypothesis Testing

Normal Distribution

- Most important and common probability distribution in statistics
  - Also known as bell curve
- Describes how the values of a variable are distributed
- A symmetric distribution where most values are centralised
- Extreme values on both ends are unlikely to occur

Normal Distribution Graph



lab}

Significance Of The Normal Distribution

- Many natural phenomena follows a normal distribution
  - Height, Blood Pressure
- Easy for statistician to work with
  - Many statistical test can be derived (We'll go deeper into this later)
- Once we obtain the mean and standard deviation of the data, it is easy for us to convert raw data to percentile and vice versa

Normal Distribution (Demo/Practice - 4)

- Randomly generate 100 numbers between 0 to 100 (Do you still remember how from Python I?)
  - Plot them into a histogram as well
  - Does it follow a normal distribution now? Why or why not?

Checkpoint 4

- Every student must be able to:
    - Understand what is a normal distribution
- For those who are waiting, try the following:
    - Generate a random number of datas
        - Is it possible to plot them into a normal distributed graph
        - What built in functions can i use?
    - What other distributions are there?

{<oding:lab}

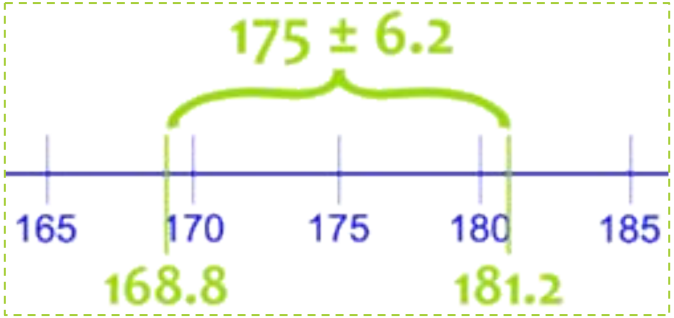
What is A Confidence Interval?

- A type of interval estimate that is computed from the observed data
- In inferential statistics, we want to estimate population parameters using observed sample data
    - E.g. Estimate the average height of male students in a school by conducting random sampling
- Can be used for non-normal distribution but it is easier to understand them in symmetric distributions first

{<coding:lab}

Confidence Interval Example

- We measure the heights of 40 randomly chosen males:
  - mean height 175cm
  - standard deviation 20cm
- The 95% Confidence Interval is as shown below



{<oding:lab}

How To Interpret Confidence Interval

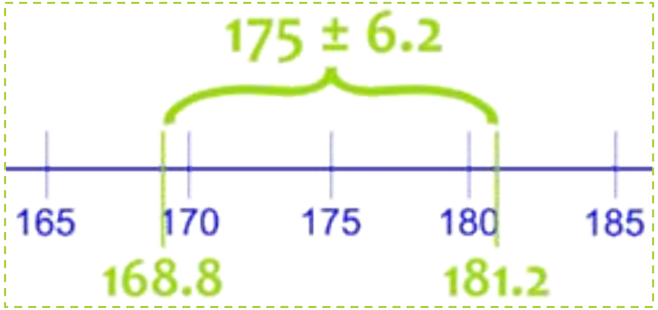
- This confidence interval is saying that the true mean of all the males in this school is likely to be between 168.8cm and 181.2cm
  - Assuming we can measure everyone
- But it might not be true
  - The "95%" confidence interval means that 95% of experiments like the one we just did will include the true mean
    - But 5% will not

True Mean

- E.g. If we run 1000 of the same experiments and create 1000 confidence intervals
  - ~950 of them will contain the true mean
  - ~50 of them will not
  - It's approximate because it's a probability
  - The more experiments we conduct, the closer the ratio will be to 95% and 5%

How Do We Calculate Confidence Interval?

- We measure the heights of 40 randomly chosen males:
    - mean height 1.75m
    - standard deviation 0.2m



{<oding:lab}

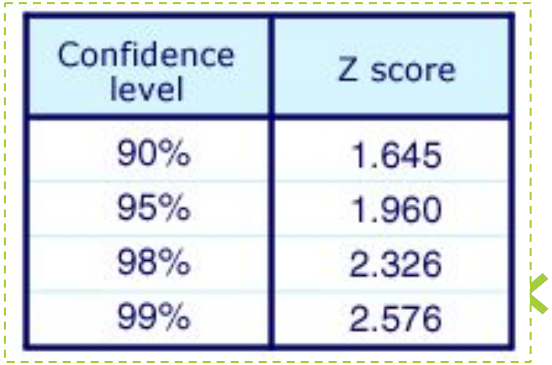
Calculating Confidence Interval (1/4)

- Step 1: find the number of samples n, calculate the mean $\bar{x}$, and the standard deviation s of those samples
  - ○ Using our example
  - ○ Number of samples n = 40
  - ○ Sample Mean $\bar{x}$ = 175
  - ○ Standard Deviation s = 20

{<oding:lab}

Calculating Confidence Interval (2/4)

- Step 2: decide what Confidence Interval we want
  - 90%, 95% and 99% are common choices
- Find the "Z" value for that particular Confidence Interval
  - For 95% the Z value is 1.960



Confidence level	Z score
90%	1.645
95%	1.960
98%	2.326
99%	2.576

Calculating Confidence Interval (3/4)

- Step 3: use that Z in this formula for the Confidence Interval

$$\overline{x} \pm z \frac{s}{\sqrt{n}}$$

  - $\overline{x}$ is the mean
  - Z is the chosen Z-value from the table above
  - s is the standard deviation
  - n is the number of samples

Calculating Confidence Interval (4/4)

- We have 175cm ± 6.20cm

$$175 \pm 1.960 \times 20/\sqrt{40}$$

- Our interval is from 168.8cm to 181.2cm
- The value after the ± is called the margin of error
  - The margin of error in our example is 6.20cm

{<coding:lab}

Confidence Interval (Demo/Practice - 5)

- We measure the math exam scores of 30 randomly selected individuals in a school
  - Mean score 86
  - Standard deviation 5
- Construct the 95% confidence interval
  - Try out different confidence levels
  - What can you conclude?

$$\overline{x} \pm z\frac{s}{\sqrt{n}}$$

Confidence Interval (Demo/Practice - 5) (Answer)

- We know that
  - $\bar{x}$ is the mean = 86
  - Z is the Z-value = 1.960 (from the table above for 95%)
  - s is the standard deviation = 5
  - n is the number of samples = 30
- $86 \pm 1.960 \times 5/\sqrt{30} = 86 \pm 1.79$
- The true mean (of all the math exam scores in the school) is likely to
- be between 84.21 and 87.79

{<oding:lab}

Checkpoint 5

- Every student must be able to:
  - Know what is a confidence interval
  - Calculate confidence interval with given values
- For students who are waiting, try the following:
  - Read up about true mean
  - What is the difference between true mean and sample mean?

Confidence Interval with M&M's

{<oding:lab}

M&M’s Color Breakdown

- https://qz.com/918008/the-color-distribution-of-mms-as-determined-by-a-phd-in-statistics/

M&M's Claim

- We want to investigate the claim that the population of M&M's are:
    - 15% Brown
    - 11% Yellow
    - 14% Red
    - 18% Orange
    - 19% Green
    - 23% Blue

M&M’s Claim: Scenario

- Let’s pick the colour blue:
  - We want to investigate the claim that 23% of all M&M’s are blue
  - Assume we opened a bag of 100 M&M’s and found 30 blue ones

M&M’s Claim: Confidence Interval Formula

- For large random samples a confidence interval for a population proportion is given by

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

M&M’s Claim: Calculating Confidence Interval

- Using the formula:

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

- = 0.3 ± 1.96* √((0.3(1-0.3))/100)
- = 0.3 ± 1.96*(0.04582575694)
- = (0.2101815164 , 0.3898184836)

M&M's Claim: Interpreting The Confidence Interval

- Our confidence interval is (0.2101815164 , 0.3898184836)
  - What can we conclude from this?
  - Is 0.23 within the interval?
    - Our sample mean does not suggest that the proportion of blue M&M's is wrong since 0.23 falls into the interval
    - How can we confirm this?

{<oding:lab}

M&M’s Claim: True Mean

- We can run the test multiple times
- At the 95% confidence level, 95% of all our tests should contain the true mean
    - 95% of all confidence intervals should contain 0.23

{<oding:lab}

Hypothesis Testing - Z Test

{<oding:lab}

What is Hypothesis testing?

- A statistical method used in making statistical decisions using experimental data.
  - An assumption we make about the population parameter
  - The goal is to use sample data to draw conclusions about an entire population
- We use significance levels and p-values to determine whether or not our test results are "statistically significant"

Null Hypothesis vs Alternate Hypothesis

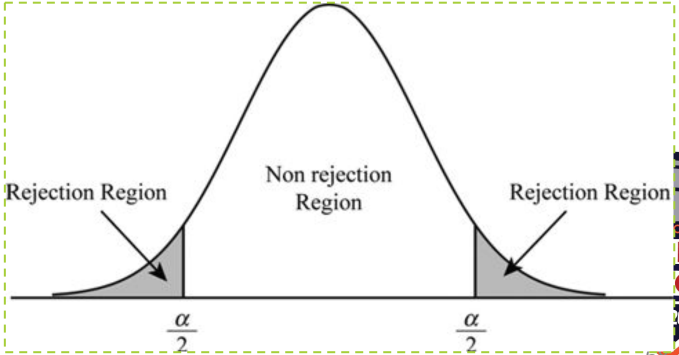
- The hypothesis we claim is true is known as the null hypothesis (H0)
- Contrary to assumption, we would also have an alternate hypothesis (H1) showing results of a real event

Hypothesis Testing Scenario

- For instance we think that the average height of males in a school is equal to 175cm
    - H0: μ = 175
- However, the average height of males in the school is greater than 175 cm
    - Then, H1 : μ > 175

Rejection Region

- The rejection region is a part of the testing process
  - An area of probability that tells us if our theory (hypothesis) is “probably” true
  - leads to rejection of the null hypothesis H0 in a hypothesis test
- Rejection region determined by level of significance



Level of Significance

- Degree of significance in which we accept or reject the null-hypothesis.
- Usually, 100% accuracy is not possible for accepting or rejecting a hypothesis, level of significance is usually selected to help determine the rejection region
  - usually 5%.

{<oding:lab}

Testing A Hypothesis

- There are two ways you can test a hypothesis: with a p-value and with a critical value
  - We will be using critical value
  - You can read up on p-value
- If the value falls in the rejection region, it means we have statistically significant results
  - We can reject the null hypothesis

One Tailed vs Two Tailed Test

- Which type of test is determined by our null hypothesis statement
  - E.g. Is the average height of all males in the school greater than 175cm?
    - This is a one tailed test
    - We are only interested in one direction (greater than 175cm)
    - Also applies to “less than”

Two Tailed Test

- A two tailed test (two rejection regions) would be used when we want to know if there is a difference in both directions (greater than and less than)

How To Do One Tailed Hypothesis Testing (One Tailed z-test)

- Step 1: State the null hypothesis
- Step 2: State the alternate hypothesis
- Step 3: State your level of significance
- Step 4: Find the z-score associated with your alpha level
- Step 5: Find the test statistic using the formula
- Step 6: Compare Step 5 to Step 4 and determine if you can reject the null hypothesis

Hypothesis Testing (One Tailed Example)

- A principal at a certain school claims that the students in his school are of above average intelligence
- A random sample of IQ scores was collected from 30 students
  - Sample mean = 112
- The mean population IQ is 100 with a standard deviation of 15
  - Is there sufficient evidence to support the principal’s claim?

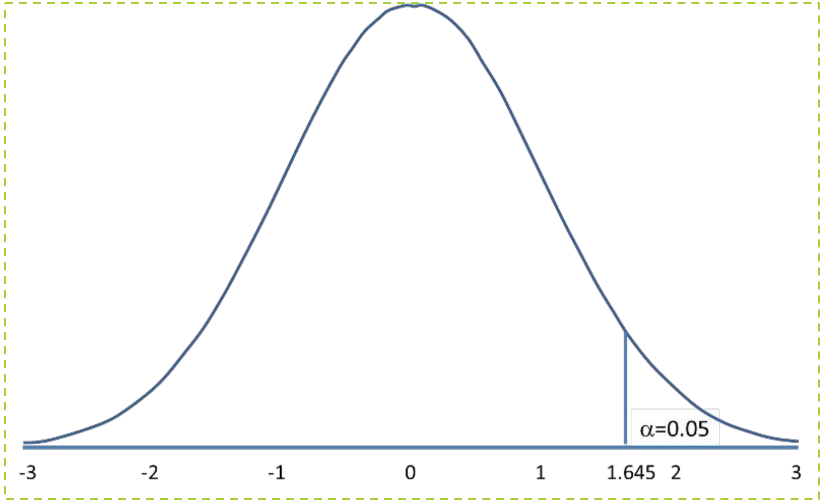
Steps For Hypothesis Testing (One Tailed Example) (1/5)

- Step 1: State the Null hypothesis
  - The accepted fact is that the population mean is 100
    - H0 : μ = 100
- Step 2: State the Alternate Hypothesis
  - The claim is that the students have above average IQ scores
    - H1 : μ > 100
  - We are looking for scores greater than a certain point
    - This is a "one-tailed test"

{<oding:lab}

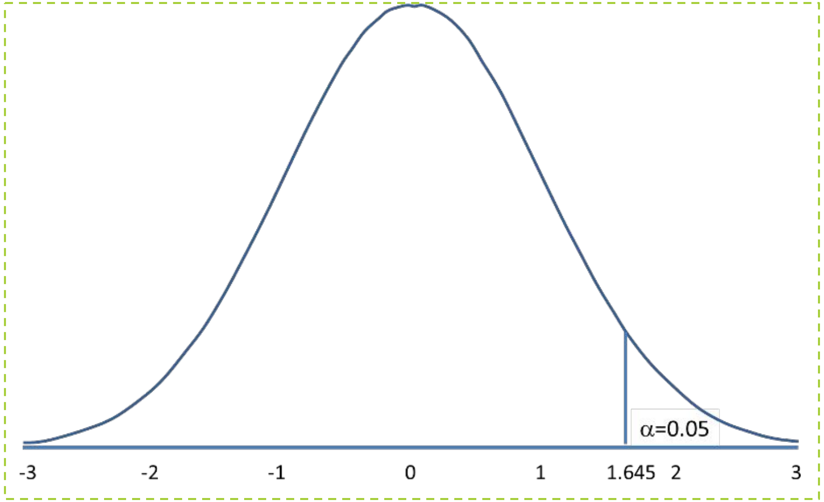
Steps For Hypothesis Testing (One Tailed Example) (2/5)

- Draw a picture to help you visualize the problem
- Step 3: State the level of significance
  - Typically use 5% or 0.05



Steps For Hypothesis Testing (One Tailed Example) (3/5)

- Step 4: Find the rejection region area from the z-table
  - Given by your alpha level
  - An area of 0.05 is equal to a z-score of 1.645



{<coding:lab}

Steps For Hypothesis Testing (One Tailed Example) (4/5)

- Step 5: Find the test statistic:
- We have
  - Sample mean 112.5
  - Population mean 100
  - Population SD 15
  - Sample size 30

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Z = (112.5-100) / (15/√30) = 4.56

Steps For Hypothesis Testing (One Tailed Example) (5/5)

- Step 6: If Step 5 is greater than Step 4, reject the null hypothesis
  - If it is less than Step 4, we cannot reject the null hypothesis
- 4.56 > 1.645, so we can reject the null hypothesis
- We conclude that there is enough evidence to say that the principal’s claim that the students in his school are of above average intelligence is true

{<oding:lab}

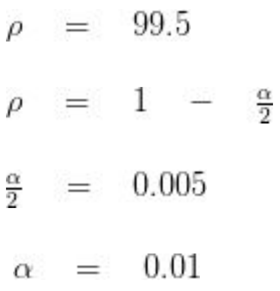
Hypothesis Testing (Demo/Practice - 6)

- A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating. Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down and repaired. This hour's sample of 18 balls has a mean of 1.611 ounces and a standard deviation of 0.065 ounces. Do you shut down the line?

{<oding:lab}

Hypothesis Testing (Demo/Practice - 6) Answer (1/3)

- H0: The population mean (µ) = 1.615
- HA: The population mean ≠ 1.615 (hence a 2-tailed test)
- Since we are doing a two-tailed test, we have to divide the level of significance in half.



$$\rho = 99.5$$

$$\rho = 1 - \frac{\alpha}{2}$$

$$\frac{\alpha}{2} = 0.005$$

$$\alpha = 0.01$$

Hypothesis Testing (Demo/Practice - 6) Answer (2/3)

- Now from the tables we can see that $t_p = 2.898$
- Now the critical regions are > 2.899 and < -2.898
- We calculate t_c now using the formula:

$$t_c = -0.261 \quad t_c = \frac{-\quad 0.004}{0.0153} \quad t_c = \frac{1.611 \quad - \quad 1.615}{0.065/\sqrt{18}}$$

$$t_c = \frac{Y \quad - \quad \mu}{s/\sqrt{n}}$$

{<oding:lab}

Hypothesis Testing (Demo/Practice - 6) Answer (3/3)

- Since t_c is clearly in the “Fail to Reject” region, then we must not reject the null hypothesis. The null hypothesis was “H0: The population mean = 1.615”. Failing to reject that means that the sample was within the bounds of what we would find acceptable if the population mean was 1.615 ounces.

- Therefore, we can say that we do not shut down the line.

Checkpoint 6

- Every student must be able to :
    - Understand the steps for hypothesis testing
    - Know the difference between one tailed test and two tailed test
- For students who are waiting, try the following:
    - Read up on p value
    - Read up on Type 1 and Type 2 error

{<oding:lab}

Other Statistical Testing Methods

{<coding:lab}

Z-test in Hypothesis Testing

- In our previous example, we have been testing out hypotheses with Z-test
  - Assuming the samples are normally distributed
  - Mean and Standard deviation are known and used for evaluations
- What if our samples are not normally distributed?
- What if we do not know the population parameters?
- What tests do we use then?

{<oding:lab}

T-test

- Used to compare the mean of two given sample/hypothesis
- Similar to Z-test, it also assumes the sample follows a normal distribution
- However, the mean and standard deviation are unknown
- 3 main type of T test

{<oding:lab}

Versions of T-test

- One sample t-test which tests the mean of a single group against a known mean
- Independent samples t-test which compares mean for two groups
- Paired sample t-test which compares means from the same group at different times

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

ANOVA Test

- Known as analysis of variance
  - used to compare multiple (three or more) samples with a single test
  - Check if the means of the group are significantly different from each other
  - Similar to T-test but is more reliable if there are more than 2 samples
- 2 Types of Anova Test
  - One-way test - test the mean of 2 or more groups
  - Two-way test - double testing one group

Chi-square Test

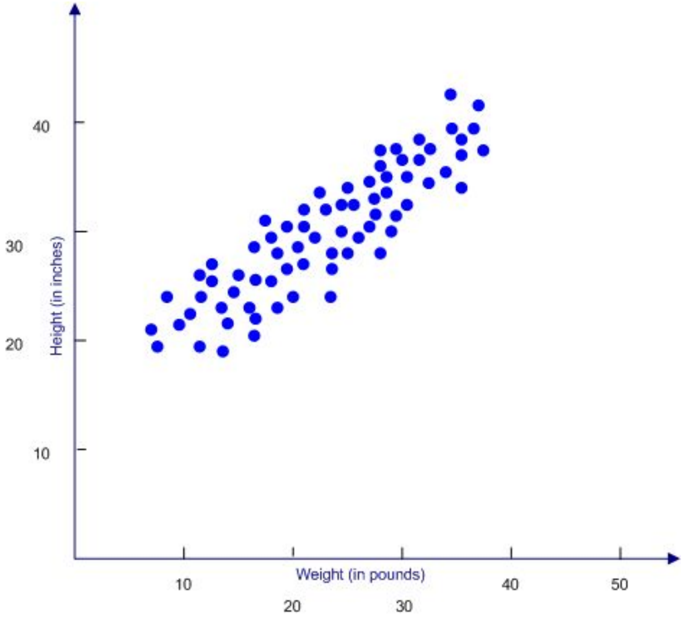
- Used to compare categorical variables
- 2 types of Chi-square Test
  - ○ Determine if a sample matches the population
  - ○ Comparing 2 variables to check if they are related
- A small chi-square value indicates the data fits
- A large chi-square value indicates the other

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

χ^2 = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies

Simple Linear Regression



{<coding:lab}

Simple Linear Regression

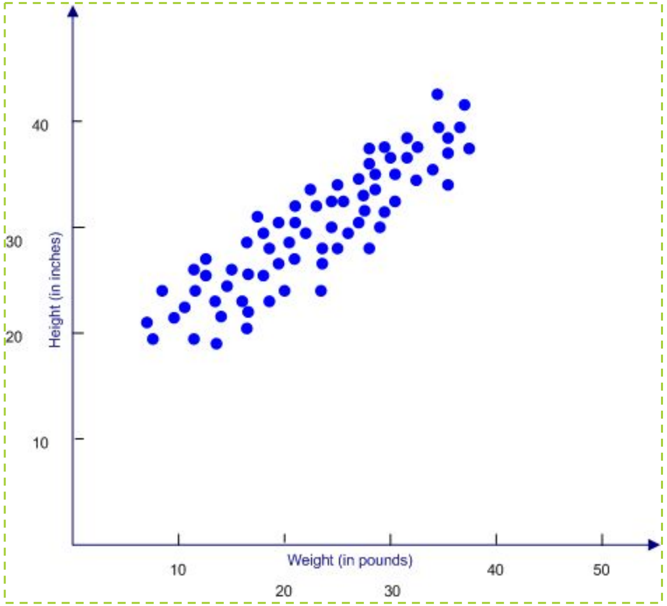
- Linear regression is a linear approach in statistics
  - Model relationship between a scalar dependent variable (y) and one or more explanatory variables (x)
- Allows us to summarize and study relationships between two quantitative variables
  - One variable, denoted x, is regarded as the predictor, explanatory, or independent variable
  - The other variable, denoted y, is regarded as the response, outcome, or dependent variable

Scatterplots

- One of the most commonly used graphs in statistics and data analysis
- displays data that is paired by using a horizontal axis (the x-axis), and a vertical axis (the y-axis)
- Scatterplots help uncover more information regarding the dataset
  - Interprets overall trend among variables

Interpreting scatterplot

- Let's observe the following scatterplot
  - Is there a relationship between the height and the weight of each individual?
- If so, is there a line we can plot to show?



{<oding:lab}

Regression Line

- A regression line is a straight line
  - Describes how a response variable y changes as an explanatory variable x changes
  - We often use a regression line to predict the value of y for a given value of x
  - Regression requires that we have a explanatory variable and a response variable

{<oding:lab}

Linear Patterns in a Regression Line

- When a scatter plot displays a linear pattern, we can describe the overall pattern by drawing a straight line through the points
  - Of course in most cases, no straight line passes exactly through all of the points
- Fitting a line means drawing a line that comes as close as possible to the points

{<oding:lab}

Equation Of A Regression Line

- The equation of a line fitted to the data gives a compact description of how the response variable y is related to the explanatory variable x
- A straight line relating y to x has the form y = mx + c
    - Where m is the slope, the amount by which y changes when x increases by one unit
    - c is the intercept, the value of y when x = 0

Extrapolation

- Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x used to obtain the line
  - Such predictions are often not accurate and should be avoided

Least Square Regression Line

- The least squares regression line of y on x is the line that makes the sum of the squares of the vertical distance of the data points from the line as small as possible
  - Is also referred to as Best Fit Line
- Regression is one of the most common statistical techniques, and least squares is the most common method for fitting a regression line to a dataset

{<coding:lab}

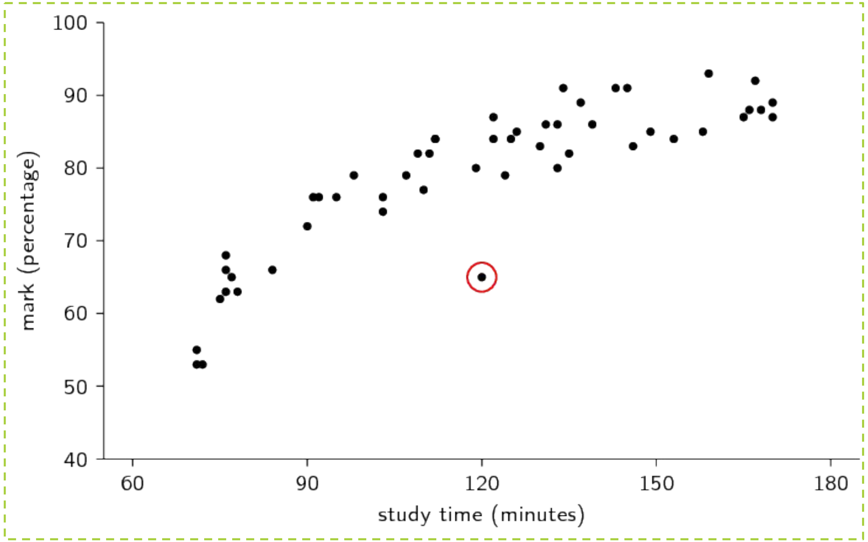
Outliers

- Often, the data presented to us are not pretty and neat
- There would be a few discrepancies in our plot
  - They are referred to as outliers
- An outlier is a data point which differs significantly from the other observations
- Many reasons for the cause of an outlier
  - Variability the the measurement
  - Experimental error

{<oding:lab}

Example of an Outlier

- In the scatterplot below, the dot circled in red is an example of an outlier



Working with Outliers

- Choice of dealing with outliers varies
- Inclusion
  - Understand your data and why there exist an outlier
  - How would the overall result be affected?
- Exclusion
  - If it is evident there is an error, we would drop the data
  - However this should rarely be used because the overall assumption would be affected

Linear Regression Case Study (Demo/Practice - 7)

- Using the dataset Session 04 - Height vs Shoe Size, apply what you have learnt to answer the following questions
  - Which are our explanatory and response variables?
    - Are the variables associated?
    - How are they associated? What is the regression equation?
  - Based on the regression equation, what is the estimated shoe size for someone of height 165cm?
  - From your observation, are there any outliers in the data?
  - Is there any relationship between Height and Shoe Size?

{<coding:lab}

Checkpoint 7

- Every student must be able to:
    - study relationships between two quantitative variables
    - Understand scatterplot and regression line
    - Know that outliers exist in datas

- For students who are waiting, try the following:
    - Read up more on least square regression line

{<oding:lab}

Summary

- Introduction to Statistics
  - Descriptive Statistics vs Inferential Statistics
- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability
- Data Analysis Summary in pandas
  - df.describe()

Summary

- Inferential Statistics
  - Normal Distribution
  - Confidence Interval
  - Hypothesis Testing
    - Z-Test
    - T-Test
    - ANOVA Test
    - Chi-Square Test
- Simple Linear Regression

Your Feedback Matters!

LIKE and follow us for more resources and tips!

@codinglabasia

{<coding:lab}

Shopee CODE LEAGUE 2020

Appendix

{<oding:lab}

Correlation and Causation

{<coding:lab}

Relationship between variables

- Two or more variables are considered related if
  - The value of one variable changes if the other one increases or decreases
  - Eg, the results you get in school is related to number of hours spent studying
    - The more time you spent studying, the better your grades
- We study the relationship between variables via correlation and causation

Correlation

- A statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables
- Correlation has values between -1.0 to 1.0
    - 1 is the perfect correlation
    - 0 indicates no correlation
    - -1 is a negative correlation where when one variable increases, the other decreases and vice versa

{<oding:lab}

Causation

- Indicates that one event is the result of the occurrence of the other event
  - There is a causal relationship between the two events
  - Also referred to as cause and effect

{<oding:lab}

Why are they important?

- One of the objectives of data analysis is to identify the extent to which one variable relates to the other
  - Eg, is there a relationship between a person’s income and his education level?
  - Dis a company’s marketing campaign increase their product sale?
- These questions explore whether there is a correlation between them
  - And if there is, we can research further if one actions has cause the other

Correlation and Causation are different

- Correlation and Causation both includes the study of relationship between variables but they do not mean the same thing
- While correlation is the mutual connection between 2 or more things, Causation is the action of causing something

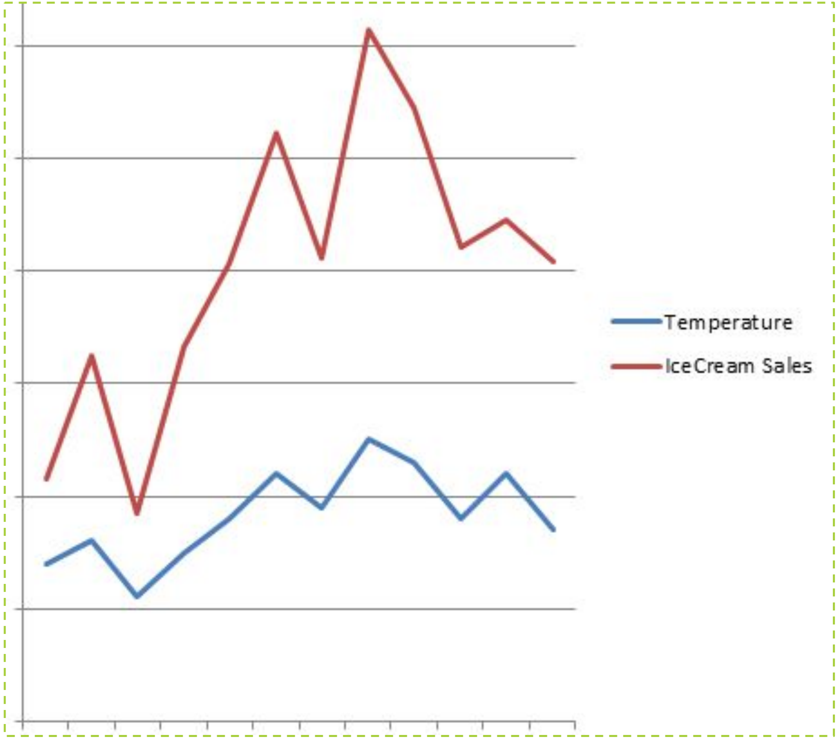
{<oding:lab}

Correlation and Causation (Demo/Practice - A1)

- Observe the graph plotted on the next slide
- Is there a correlation, causation or both between the temperature and ice cream sales?

{<oding:lab}

Correlation and Causation - Graph (Demo/Practice - A1)



{<oding:lab}

Correlation and Causation Examples

- Correlation exists?
  - As the temperature increase, the warmer the weather gets, then the sales of ice cream increase
- Causation exists?
  - The warm weather is the cause and the effect is the increase in ice cream sales
  - Is this an effect? Or simply a relationship
  - The warm weather can cause an increase in heat stroke but does not cause an increase in sales