Course	Thing	Explanation	Date	Important	Index
GEA1000	**GEA1000**	Topics ?> PPDAC	03/02/2022		
GEA1000	**PPDAC**	Definition > Problem > Plan > Data > Analysis > Conclusion Topics ?> Research_Question ?> Sampling ?> Exploratory_Data_Analysis ?> Variables ?> Study_Design ?> Statistical_Inference ?> Univariate_Analysis ?> Bivariate_Relationship	03/02/2022		
GEA1000	**Exploratory Data Analysis**	Definition > Explore the data to come up with answers to questions	03/02/2022		
GEA1000	**Sampling**	Topics ?> Sampling_Frame ?> Sampling_Bias ?> Estimate ?> Population_Parameter ?> Random_Error Types ?> Probability_Sampling ?> Non_Probability_Sampling	03/02/2022		
GEA1000	**Population**	Definition > The group of people you want to know about	03/02/2022		
GEA1000	**Research Question**	Topics ?> Census ?> Population Types ?> Estimation_Question ?> Test_Claim_Question ?> Comparison_Question	03/02/2022		
GEA1000	**Estimation Question**	Examples > What is the average number	03/02/2022		
GEA1000	**Test Claim Question**	Examples > Is the average number	03/02/2022		
GEA1000	**Comparison Question**	Examples > Is A bigger than B	03/02/2022		
GEA1000	**Census**	Definition > 100% accuracy, 100% response rate studies	03/02/2022		494
GEA1000	**Sampling Frame**	Definition > Sampling frame will decide how generalisable the study is to the target population	03/02/2022		495
GEA1000	**Population Parameter**	Definition > The statistic about the population that you want to know about, like average age etc	03/02/2022		
GEA1000	**Random Error**	Definition > Despite having a perfect sample, sometimes random deviances happen and are out of control	03/02/2022		
GEA1000	**Sampling Bias**	Types ?> Selection_Bias ?> Non_Response_Bias	03/02/2022		
GEA1000	**Selection Bias**	Definition > Researcher's problem Examples > Imperfect /Sampling_Frame > Improper /Probability_Sampling	03/02/2022		496
GEA1000	**Non Response Bias**	Definition > Participants' problem Examples > Disinterest > Inconvenient > Unwilling	03/02/2022		
GEA1000	**Probability Sampling**	Definition > Everyone has a chance of participating Benefits + Mitigates /Selection_Bias Types ?> Simple_Random_Sampling ?> Systematic_Sampling ?> Stratified_Sampling ?> Cluster_Sampling	03/02/2022		497
GEA1000	**Non Probability Sampling**	Definition > Humans choose the participants Types ?> Convenience_Sampling ?> Volunteer_Sampling	03/02/2022		

Course	Thing	Explanation	Date	Important	Index
GEA1000	Convenience Sampling	Definition > Study the most convenient people Benefits + Very easy Disadvantages - Subject to /Selection_Bias and /Non_Response_Bias Examples > Mall survey	03/02/2022		
GEA1000	Volunteer Sampling	Definition > Participants choose themselves Benefits + Easy and mitigates /Non_Response_Bias Disadvantages - Attracts an unrepresentative group of people Examples > Optional surveys	03/02/2022		
GEA1000	Simple Random Sampling	Definition > Random number generator ?> Uniform_Probability Benefits + Good representation of population Disadvantages - Subject to /Non_Response_Bias	03/02/2022		
GEA1000	Uniform Probability	Definition > Every outcome has the same probability 1/n	03/02/2022		
GEA1000	Systematic Sampling	Definition > Apply a pattern for selecting Benefits + Simpler selection process Disadvantages - Subject to /Selection_Bias if wrong pattern	03/02/2022		
GEA1000	Stratified Sampling	Definition > Break down population into strata > Conduct /Simple_Random_Sampling on each strata > Do weighted calculations to find population Benefits + Can get representative sample from each stratum Disadvantages - Need information about sampling frame and stratum	03/02/2022		
GEA1000	Cluster Sampling	Definition > Break down population into clusters > Randomly sample a fixed number of clusters > Include all obsetvations Benefits + Less tedious Disadvantages - High variability due to dissimilar clusters Examples > Mental wellbeing study in separate schools	03/02/2022		
GEA1000	Estimate	Definition > Estimate = /Population_Parameter + /Sampling_Bias + /Random_Error > AKA Sample Statistic Types ?> Good_Estimate	03/02/2022	Important	498
GEA1000	Good Estimate	Properties > Sampling frame > Probability sampling > Large enough > High response rate	03/02/2022		499
GEA1000	Variables	Types ?> Categorical_Data ?> Numerical_Data ?> Independent_Variable ?> Dependent_Variable	03/02/2022		500
GEA1000	Univariate Analysis	Topics ?> Shape_Of_Data ?> Outlier ?> Summary_Statistics	10/03/2022		
GEA1000	Summary Statistics	Types ?> Standard_Deviation ?> Coefficient_Of_Variation ?> Interquartile_Range ?> Five_Number_Summary	03/02/2022		
GEA1000	Shape Of Data	Definition > Used to get a rough idea of the distribution within the group Topics ?> Statistical_Skew ?> Data_Peak ?> Histogram	10/03/2022		
GEA1000	Histogram	Usage > Give a better understanding of the distribution of the data	10/03/2022		
GEA1000	Data Peak	Definition >	10/03/2022		

Course	Thing	Explanation	Date	Important	Index
GEA1000	Statistical Skew	Types ?> Left_Skewed_Data ?> Right_Skewed_Data ?> Symmetric_Data Definition > A high degree of skew will cause the mean to shoot up > As such, it is more common to use median to represent the data if the data is very skewed > Skewness ≠ outliers	03/02/2022	Important	
GEA1000	Symmetric Data	Definition > No skew	10/03/2022		
GEA1000	Left Skewed Data	Definition > Most of the bulk is on the right Properties > Mode < Median < Mean	03/02/2022		
GEA1000	Right Skewed Data	Definition > Most of the bulk is on the left Properties > Mode > Median > Mean	03/02/2022		
GEA1000	Box Plot	Definition > Used to represent the /Five_Number_Summary	10/03/2022		
GEA1000	Five Number Summary	Types ?> Minimum_Data ?> Median_Data ?> Maximum_Data ?> Q1_Data ?> Q3_Data Topics ?> Box_Plot ?> Robust_Statistics	03/02/2022		
GEA1000	Robust Statistics	Definition > Statistics that are unaffected by outliers			
GEA1000	Standard Deviation	Properties ><MA SD ≈ Range/4 ≈ IQR*0.75 MA> > SD changes when data is multiplied	03/02/2022		
GEA1000	Interquartile Range	Definition ><MA Q3-Q1 MA>	03/02/2022		
GEA1000	Minimum Data	Definition > Minimum	03/02/2022		
GEA1000	Maximum Data	Definition > Maximum	03/02/2022		
GEA1000	Median Data	Definition > Median Properties > Unaffected by outliers	03/02/2022		
GEA1000	Q1 Data	Definition > 25 Percentile	03/02/2022		
GEA1000	Q3 Data	Definition > 75 Percentile	03/02/2022		
GEA1000	Outlier	Definition ><MA x < Q1-1.5IQR MA> ><MA x > Q3+1.5IQR MA> Properties > Can mess up mean & standard deviation	03/02/2022	Important	
GEA1000	Coefficient Of Variation	Definition ><MA SD / Mean MA>	03/02/2022		
GEA1000	Independent Variable	Definition > Researcher control	03/02/2022		
GEA1000	Dependent Variable	Definition > Researcher wants to know	03/02/2022		
GEA1000	Categorical Data	Types ?> Nominal_Data ?> Ordinal_Data	03/02/2022		
GEA1000	Numerical Data	Definition > Arithmetic operations make sense Types ?> Discrete_Data ?> Continuous_Data	03/02/2022		
GEA1000	Nominal Data	Definition > Data is in unordered groups Examples > Country	03/02/2022		
GEA1000	Ordinal Data	Definition > Data is in ordered groups Examples > Education level	03/02/2022		
GEA1000	Discrete Data	Definition > Finite possibilities for data	03/02/2022		
GEA1000	Continuous Data	Definition > Infinite possibilities for data	03/02/2022		
GEA1000	Study Design	Topics ?> Controlled_Experiment ?> Observational_Study	03/02/2022		
GEA1000	Treatment Group	Usage > Changes the independent variable to attempt to give change to the dependent variable	03/02/2022		
GEA1000	Control Group	Usage > Provide a baseline for comparing data	03/02/2022		

Course	Thing	Explanation	Date	Important	Index
GEA1000	**Controlled Experiment**	Types ?> Single_Blinded_Experiment ?> Double_Blinded_Experiment Topics ?> Treatment_Group ?> Control_Group Definition > Intentionally manipulates one variable to cause an effect on another variable Usage > Try to aim for randomisation, and or blinding > Choose participants / assessors Disadvantages - Unethical for a lot of studies	03/02/2022		501
GEA1000	**Single Blinded Experiment**	Definition > Participants or evaluators don't know which group they are in	03/02/2022		
GEA1000	**Double Blinded Experiment**	Definition > Participants and evaluators don't know which group they are in > Best method of carrying out comparison studies	03/02/2022		
GEA1000	**Observational Study**	Disadvantages - Leave so much to the participant that there can be no study to be made - Allocation not random - Creates confounders - Cannot prove causation	03/02/2022		502
GEA1000	**Radiant**	Link > https://vnijs.shinyapps.io/radiant/?SSUID=f4572720b1	03/02/2022		503
GEA1000	**Statistical Rate**	Types ?> Marginal_Rate ?> Conditional_Rate ?> Joint_Rate Topics ?> Association ?> Simpson_s_Paradox ?> Confounder	17/02/2022		
GEA1000	**Association**	Types ?> Positive_Association ?> Negative_Association ?> Linear_Association Usage > As long as not exactly equal, then they are associated Proof > P(A|B)≠P(A|~B) or P(B|A)≠P(B|~A) Disadvantages - Weaker than /Causation - Many controls are needed to establish /Causation	17/02/2022	Important	530
GEA1000	**Marginal Rate**	Definition > Only interested in a single column or row in the data Usage > P(A), P(B) Examples > What proportion of the total did A	17/02/2022		531
GEA1000	**Conditional Rate**	Definition > Interested in a probability given an event Usage > P(A|B), P(B|A) Examples > What proportion of the ones who did B also did A	17/02/2022		532
GEA1000	**Joint Rate**	Definition > Interested in the a single cell in the data Usage > P(A∩B) Examples > What proportion of the total did A and B	17/02/2022		
GEA1000	**Positive Association**	Definition > A and B are positively associated ⟺ A increase → B increase Proof > P(A|B) > rate(A|~B)	17/02/2022		535
GEA1000	**Negative Association**	Definition > A and B are negatively associated ⟺ A increase → B decrease Proof > P(A|B) < rate(A|~B)	17/02/2022		536
GEA1000	**Confounder**	Definition > Tests whether a variable affects others Usage > Testing A~B, with C as confounder Proof > P(A|C)≠P(A|~C) ∧ P(B|C)≠P(B|~C) > Correlation coefficient between A and C and B and C are not 0	17/02/2022		537

Course	Thing	Explanation	Date	Important	Index
GEA1000	Simpson's Paradox	Definition > Relationship between rates in subgroups are reversed/disappears when subgroups are combined Usage > /Confounder is what makes the difference in rates Proof > C_1: $P(A \mid B)>P(A \mid \sim B)$ and C_2: $P(A \mid B)<P(A \mid \sim B)$ > Simpson's Paradox occurs with C as /Confounder	17/02/2022		538
GEA1000	Random Assignment	Usage > To ensure an equal representation of confounders Disadvantages - Unethical	17/02/2022		
GEA1000	Slicing Data	Usage > To stratify the data to eliminate /Simpson_s_Paradox	17/02/2022		
GEA1000	Excel Convert To Percentage	Usage > Convert pivot table to percentages Process > Right click on value header > Value field settings > Show data as	17/02/2022		539
GEA1000	Excel Change Chart Type	Usage > Change type of chart Process > Right click on chart > Change chart type	17/02/2022		540
GEA1000	Excel Format Axis	Usage > Change axis zero Process > Right click on axis > Format axis	17/02/2022		541
GEA1000	Bivariate Relationship	Types ?> Deterministic_Relationship ?> Statistical_Relationship	23/02/2022		
GEA1000	Deterministic Relationship	Definition > Fixed relation between two variables Examples > Physics conversions (℃ → Fahrenheit) Topics ?> True_Value	23/02/2022		
GEA1000	True Value	Definition > Function representing the /Deterministic_Relationship is well-defined Types ?> Unique_True_Value	10/03/2022		
GEA1000	Unique True Value	Definition > Function is /Injective	10/03/2022		
GEA1000	Statistical Relationship	Definition > Natural variability in relation between two variables Topics ?> Scatter_Plot ?> Regression	23/02/2022		
GEA1000	Scatter Plot	Usage > Get idea of the pattern between two variables	23/02/2022		
GEA1000	Regression	Types ?> Linear_Regression ?> Non_Linear_Regression Topics ?> Regression_Analysis	10/03/2022		
GEA1000	Regression Analysis	Definition > You can't extrapolate outside of the range of the data Types ?> Regression_Direction ?> Regression_Form ?> Regression_Strength	23/02/2022		
GEA1000	Linear Regression	Definition > Represents a /Statistical_Relationship with a linear equation Y=mX+b Properties ><MA m = r(SDy/SDx) MA> f(x) cannot be used to predict x Process > Obtained by minimising the squares of differences Topics ?> Linear_Association ?> Excel_Regression_Line	23/02/2022		
GEA1000	Excel Regression Line	Process ::Create regression line > Highlight two columns > Insert /Scatter_Plot > Go to Chart Design and add chart element > Trendline linear > Right-click on trendline > Format trendline and display equation and R value ::Create matrix of regression values > Go to Data Analysis (after going tools > excel add ins) > Choose regression and the columns you want	10/03/2022		

Course	Thing	Explanation	Date	Important	Index
GEA1000	Linear Association	Definition > Whether A and B are linearly associated Topics ?> Correlation_Coefficient	10/03/2022		
GEA1000	Non Linear Regression	Definition > Can use linear law to convert non-linear into linear	10/03/2022		
GEA1000	Regression Direction	Types > Positive > Negative > Neither	23/02/2022		
GEA1000	Regression Form	Types > Linear > Non-linear	23/02/2022		
GEA1000	Regression Strength	Types > Strong relationship > Weak relationship	23/02/2022		
GEA1000	Correlation Coefficient	Definition > Only measures linear association ><MA r = m(s□/s□) MA> Properties > r=0 does not mean there is no relationship, so must look at /Scatter_Plot > Unaffected by /Linear_Transformation of the x and y axes Process ::Calculate Correlation Coefficient > Calculate SU(x) for all x and all y > Sum up all values and divide by n-1 Topics ?> Standard_Unit	23/02/2022		
GEA1000	Standard Unit	Definition > SU(x) = (x-X̄)/s□	10/03/2022		
GEA1000	Statistical Inference	Definition > Using data to answer questions on data Topics ?> Basics_Of_Probability ?> Proportion ?> Conditional_Probability ?> Random_Variable Types ?> Confidence_Interval ?> Hypothesis_Testing	08/03/2022		
GEA1000	Proportion	Definition > An estimate for the true probability of the experiment	08/03/2022		
GEA1000	Conditional Probability	Types ?> Sensitivity ?> Specificity	08/03/2022		
GEA1000	Sensitivity	Definition > True positive rate > P(CVD|+)	08/03/2022		
GEA1000	Specificity	Definition > True negative rate > P(~CVD|-)	08/03/2022		
GEA1000	Normal Distribution	Definition > Defined by the mean and variance of the distribution	08/03/2022		
GEA1000	Confidence Interval	Definition > Use probability to determine how accurate the estimate is of the population parameters > We are k% confident that the population proportion lies within the k% confidence interval Types ?> Proportion_Confidence_Interval ?> Mean_Confidence_Interval	08/03/2022		
GEA1000	Proportion Confidence Interval	Process ::Determine confidence interval using sample population p*, sample size n, value from normal distribution z* > p*±z*√(p*(1-p*)/n)	08/03/2022		
GEA1000	Mean Confidence Interval	Process ::Determine confidence interval using sample mean X̄, sample size n, sample deviation s, value from t-distribution t* > X̄±t*s/√n	08/03/2022		
GEA1000	Hypothesis Testing	Process > Identify question and state null & alternative hypotheses > Collect relevant data based on test statistic > Determine level of significance and compute p-value > Making conclusion about null hypothesis Types ?> T_Test ?> Chi_Square_Test Topics ?> Null_Hypothesis ?> Alternative_Hypothesis	08/03/2022		
GEA1000	Null Hypothesis	Definition > Should be the default hypothesis, either no association or sample statistic equals to a certain value	08/03/2022		
GEA1000	Alternative Hypothesis	Definition > Should be the exceptional case, mutually exclusive to the /Null_Hypothesis	08/03/2022		