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

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

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

| Course | Thing | Explanation | Date | Important | Index |
|---|---|---|---|---|---|
| **GEA1000** | **Control Group** | Usage<br>> Provide a baseline for comparing data | 03/02/2022 | | |
| **GEA1000** | **Controlled Experiment** | Types<br>?> Single_Blinded_Experiment<br>?> Double_Blinded_Experiment<br><br>Topics<br>?> Treatment_Group<br>?> Control_Group<br><br>Definition<br>> Intentionally manipulates one variable to cause an effect on another variable<br><br>Usage<br>> Try to aim for randomisation, and or blinding<br>> Choose participants / assessors<br><br>Disadvantages<br>- Unethical for a lot of studies | 03/02/2022 | | 501 |
| **GEA1000** | **Single Blinded Experiment** | Definition<br>> Participants or evaluators don't know which group they are in | 03/02/2022 | | |
| **GEA1000** | **Double Blinded Experiment** | Definition<br>> Participants and evaluators don't know which group they are in<br>> Best method of carrying out comparison studies | 03/02/2022 | | |
| **GEA1000** | **Observational Study** | Disadvantages<br>- Leave so much to the participant that there can be no study to be made<br>- Allocation not random<br>- Creates confounders<br>- Cannot prove causation | 03/02/2022 | | 502 |
| **GEA1000** | **Radiant** | Link<br>> https://vnijs.shinyapps.io/radiant/?SSUID=f4572720b1 | 03/02/2022 | | 503 |
| **GEA1000** | **Statistical Rate** | Definition<br>> Can be treated as probability if there is a probability sampling involved<br><br>Types<br>?> Marginal_Rate<br>?> Conditional_Rate<br>?> Joint_Rate<br><br>Topics<br>?> Association<br>?> Simpson_s_Paradox<br>?> Confounder<br>?> Rules_For_Rates | 17/02/2022 | | |
| **GEA1000** | **Rules For Rates** | Definition<br>> R(A\|B)<=>rate(A\|B') ⟹ R(B\|A)<=>rate(B\|A')<br>> R(A\|B)<>R(A)<>R(A\|B') | 24/03/2022 | | |
| **GEA1000** | **Association** | Types<br>?> Positive_Association<br>?> Negative_Association<br>?> Linear_Association<br><br>Usage<br>> As long as not exactly equal, then they are associated<br><br>Proof<br>> P(A\|B)≠P(A\|~B) or P(B\|A)≠P(B\|~A)<br><br>Disadvantages<br>- Weaker than /Causation<br>- Many controls are needed to establish /Causation | 17/02/2022 | Important | 530 |
| **GEA1000** | **Marginal Rate** | Definition<br>> Only interested in a single column or row in the data<br><br>Usage<br>> P(A), P(B)<br><br>Examples<br>> What proportion of the total did A | 17/02/2022 | | 531 |
| **GEA1000** | **Conditional Rate** | Definition<br>> Interested in a probability given an event<br><br>Usage<br>> P(A\|B), P(B\|A)<br><br>Examples<br>> What proportion of the ones who did B also did A | 17/02/2022 | | 532 |
| **GEA1000** | **Joint Rate** | Definition<br>> Interested in the a single cell in the data<br><br>Usage<br>> P(A∩B)<br><br>Examples<br>> What proportion of the total did A and B | 17/02/2022 | | |
| **GEA1000** | **Positive Association** | Definition<br>> A and B are positively associated ⟺ A increase → B increase<br><br>Proof<br>> P(A\|B) > rate(A\|~B) | 17/02/2022 | | 535 |
| **GEA1000** | **Negative Association** | Definition<br>> A and B are negatively associated ⟺ A increase → B decrease<br><br>Proof<br>> P(A\|B) < rate(A\|~B) | 17/02/2022 | | 536 |

| Course | Thing | Explanation | Date | Important | Index |
|---|---|---|---|---|---|
| **GEA1000** | **Confounder** | Definition<br>> Tests whether a variable affects others<br><br>Usage<br>> Testing A~B, with C as confounder<br><br>Proof<br>> P(A\|C)≠P(A\|~C) ∧ P(B\|C)≠P(B\|~C)<br>> Correlation coefficient between A and C and B and C are not 0 | 17/02/2022 | | 537 |
| **GEA1000** | **Simpson's Paradox** | Definition<br>> Relationship between rates in subgroups are reversed/disappears when subgroups are combined<br><br>Usage<br>> /Confounder is what makes the difference in rates<br><br>Proof<br>> $C_1$: P(A\|B)>P(A\|~B) and $C_2$: P(A\|B)<P(A\|~B)<br>> Simpson's Paradox occurs with C as /Confounder | 17/02/2022 | | 538 |
| **GEA1000** | **Random Assignment** | Usage<br>> To ensure an equal representation of confounders<br><br>Disadvantages<br>- Unethical | 17/02/2022 | | |
| **GEA1000** | **Slicing Data** | Usage<br>> To stratify the data to eliminate /Simpson_s_Paradox | 17/02/2022 | | |
| **GEA1000** | **Excel Convert To Percentage** | Usage<br>> Convert pivot table to percentages<br><br>Process<br>> Right click on value header<br>> Value field settings<br>> Show data as | 17/02/2022 | | 539 |
| **GEA1000** | **Excel Change Chart Type** | Usage<br>> Change type of chart<br><br>Process<br>> Right click on chart<br>> Change chart type | 17/02/2022 | | 540 |
| **GEA1000** | **Excel Format Axis** | Usage<br>> Change axis zero<br><br>Process<br>> Right click on axis<br>> Format axis | 17/02/2022 | | 541 |
| **GEA1000** | **Bivariate Relationship** | Types<br>?> Deterministic_Relationship<br>?> Statistical_Relationship | 23/02/2022 | | |
| **GEA1000** | **Deterministic Relationship** | Definition<br>> Fixed relation between two variables<br><br>Examples<br>> Physics conversions (℃ → Fahrenheit)<br><br>Topics<br>?> True_Value | 23/02/2022 | | |
| **GEA1000** | **True Value** | Definition<br>> Function representing the /Deterministic_Relationship is well-defined<br><br>Types<br>?> Unique_True_Value | 10/03/2022 | | |
| **GEA1000** | **Unique True Value** | Definition<br>> Function is /Injective | 10/03/2022 | | |
| **GEA1000** | **Statistical Relationship** | Definition<br>> Natural variability in relation between two variables<br><br>Topics<br>?> Scatter_Plot<br>?> Regression | 23/02/2022 | | |
| **GEA1000** | **Scatter Plot** | Usage<br>> Get idea of the pattern between two variables | 23/02/2022 | | |
| **GEA1000** | **Regression** | Types<br>?> Linear_Regression<br>?> Non_Linear_Regression<br><br>Topics<br>?> Regression_Analysis | 10/03/2022 | | |
| **GEA1000** | **Regression Analysis** | Definition<br>> You can't extrapolate outside of the range of the data<br><br>Types<br>?> Regression_Direction<br>?> Regression_Form<br>?> Regression_Strength | 23/02/2022 | | |
| **GEA1000** | **Linear Regression** | Definition<br>> Represents a /Statistical_Relationship with a linear equation Y=mX+b<br><br>Properties<br>><MA m = r(SDy/SDx) MA><br><br>Process<br>> Obtained by minimising the squares of differences<br><br>Topics<br>?> Linear_Association<br>?> Excel_Regression_Line | 23/02/2022 | | |

| Course | Thing | Explanation | Date | Important | Index |
|---|---|---|---|---|---|
| **GEA1000** | **Excel Regression Line** | Process<br>::Create regression line<br>> Highlight two columns<br>> Insert /Scatter_Plot<br>> Go to Chart Design and add chart element<br>> Trendline linear<br>> Right-click on trendline<br>> Format trendline and display equation and R value<br>::Create matrix of regression values<br>> Go to Data Analysis (after going tools > excel add ins)<br>> Choose regression and the columns you want | 10/03/2022 | | |
| **GEA1000** | **Linear Association** | Definition<br>> Whether A and B are linearly associated<br><br>Topics<br>?> Correlation_Coefficient | 10/03/2022 | | |
| **GEA1000** | **Non Linear Regression** | Definition<br>> Can use linear law to convert non-linear into linear | 10/03/2022 | | |
| **GEA1000** | **Regression Direction** | Types<br>> Positive<br>> Negative<br>> Neither | 23/02/2022 | | |
| **GEA1000** | **Regression Form** | Types<br>> Linear<br>> Non-linear | 23/02/2022 | | |
| **GEA1000** | **Regression Strength** | Types<br>> Strong relationship<br>> Weak relationship | 23/02/2022 | | |
| **GEA1000** | **Correlation Coefficient** | Definition<br>> Only measures linear association<br>><MA r = m(s□/s□) MA><br><br>Properties<br>> r=0 does not mean there is no relationship, so must look at /Scatter_Plot<br>> Unaffected by /Linear_Transformation of the x and y axes<br><br>Process<br>::Calculate Correlation Coefficient<br>> Calculate SU(x) for all x and all y<br>> Sum up all values and divide by n-1<br><br>Topics<br>?> Standard_Unit | 23/02/2022 | | |
| **GEA1000** | **Standard Unit** | Definition<br>> SU(x) = $(x-\bar{x})/s$□<br>> Z = (X-μ)/s | 10/03/2022 | Important | |
| **GEA1000** | **Statistical Inference** | Definition<br>> Using data to answer questions on data<br><br>Topics<br>?> Basics_Of_Probability<br>?> Proportion<br>?> Conditional_Probability<br>?> Random_Variable<br><br>Types<br>?> Confidence_Interval<br>?> Hypothesis_Testing | 08/03/2022 | | |
| **GEA1000** | **Proportion** | Definition<br>> An estimate for the true probability of the experiment | 08/03/2022 | | |
| **GEA1000** | **Conditional Probability** | Types<br>?> Sensitivity<br>?> Specificity | 08/03/2022 | | |
| **GEA1000** | **Sensitivity** | Definition<br>> True positive rate<br>> P(CVD\|+) | 08/03/2022 | | |
| **GEA1000** | **Specificity** | Definition<br>> True negative rate<br>> P(~CVD\|-) | 08/03/2022 | | |
| **GEA1000** | **Normal Distribution** | Definition<br>> Defined by the mean and variance of the distribution<br>> $P((a-\bar{x})/sd < Z < (b-\bar{x})/sd)$<br>> $N(50, 10^2)$ for P(X<45) ⟹ NORM.DIST(45, 50, 10)<br>> $N(50, 10^2)$ for P(X<a) = 0.05 ⟹ NORM.INV(0.05, 50, 10) | 08/03/2022 | | |
| **GEA1000** | **Confidence Interval** | Definition<br>> Use probability to determine how accurate the estimate is of the population parameters<br>> We are k% confident that the population proportion lies within the k% confidence interval<br>> NOT k% CHANCE SINCE THERE IS NO RANDOM PROCESS<br>> There is a k% chance of selecting a correct confidence interval<br>> When you use the entire population, there is no random error, so confidence interval is width 0<br><br>Types<br>?> Proportion_Confidence_Interval<br>?> Mean_Confidence_Interval<br>?> Difference_Confidence_Interval<br>?> Variance_Confidence_Interval | 08/03/2022 | | |
| **GEA1000** | **Proportion Confidence Interval** | Process<br>::Determine confidence interval using sample population p*, sample size n, value from normal distribution z*<br>> $p^* \pm z^* \sqrt{(p^*(1-p^*)/n)}$ | 08/03/2022 | | |
| **GEA1000** | **Mean Confidence Interval** | Types<br>?> Known_Variance_Mean_Confidence_Interval<br>?> Unknown_Variance_Mean_Confidence_Interval | 08/03/2022 | | |

| Course | Thing | Explanation | Date | Important | Index |
|---|---|---|---|---|---|
| **GEA1000** | **Hypothesis Testing** | Usage<br>> Don't use it when you have access to the full population<br><br>Process<br>> Identify question and state null & alternative hypotheses<br>> Collect relevant data based on test statistic<br>> Determine level of significance and compute p-value<br>> Making conclusion about null hypothesis<br>> Find P(observation & everything rarer than it)<br><br>Properties<br>> Person :- Ronald Fischer<br><br>Types<br>?> T_Test<br>?> Chi_Square_Test<br><br>Topics<br>?> Null_Hypothesis<br>?> Alternative_Hypothesis | 08/03/2022 | | |
| **GEA1000** | **Null Hypothesis** | Definition<br>> Should be the default hypothesis, either no association or sample statistic equals to a certain value<br>> Assumes everything happens due to chance | 08/03/2022 | | |
| **GEA1000** | **Alternative Hypothesis** | Definition<br>> Should be the exceptional case, mutually exclusive to the /Null_Hypothesis<br>> Everything happens for a reason | 08/03/2022 | | |