Definitions

Population	The entire collection of elements (e.g. persons, items) that we want to study or draw conclusions about
Sample	A subset of the population from which we collect data
Random sample	A sample where… ● Each element has an equal chance of being selected ● The selection of each element is independent of the selection of every other element
Unbiased estimate	An estimate such that the average of such estimates obtained from all possible samples of the same size gives the true value of the population parameter
Significance level	The probability of rejecting H_0 when H_0 is in fact true
Critical region	The set of values of the sample mean for which H_0 is rejected
Critical value	The value(s) that defines the critical region, i.e. the value(s) such that if the sample mean is more extreme than that value, H_0 is rejected
P-value	[If H_1: $\mu<\mu_0$ OR H_1: $\mu>\mu_0$] The probability of getting a sample mean less / greater than the observed sample mean [If H_1: $\mu\neq\mu_0$] The probability of getting a sample mean more extreme than the observed sample mean
Independent variable	A variable whose change will have an effect on the dependent variable
Dependent variable	A variable which is affected by changes of the independent variable(s)
Residual	The difference between the actual value of y and the predicted value of y by a model for a given value of x

Remember that only events, not probabilities, can be independent!

Explanations

Hypothesis Testing	
"State your conclusion in context"	[If p-value < α%] Since p-value < α, we reject H_0. There is sufficient evidence at the α% significance level that… [If p-value > α%] Since p-value > α, we do not reject H_0. There is insufficient evidence at the α% significance level that…
Why is the data summarised in the form (x-μ)?	Since the mean of the population parameter is μ, summarising the data in the form (x-μ) gives the difference between the parameter and the mean, allowing the values to be smaller and easier to calculate.
How should the sample be conducted? / Why should the sample be random?	A random sample should be conducted such that the sample is unbiased, i.e. representative of the population. ● For instance, by assigning each element a natural number, generating a list of N random numbers with a calculator, and selecting the corresponding N elements
Why must a sample of size N>30 be taken?	This is because the distribution of the population parameter is not known / the population parameter is not normally distributed, so a large sample of N>30 must be taken such that the Central Limit Theorem applies and the sample mean will be approximately normally distributed.
Should a one-tail or two-tail test be conducted?	[If H_1: $\mu<\mu_0$ OR H_1: $\mu>\mu_0$] A one-tail test should be conducted, as we are only interested in testing if the population mean is less / greater than μ_0, and not whether it is greater / less than μ_0. [If H_1: $\mu\neq\mu_0$] A two-tail test should be conducted, as we are interested in testing if the population mean differs from μ_0. To differ from μ_0, the population mean could either be greater than μ_0 or less than μ_0, making a two-tail test appropriate.
What assumptions must be made for the hypothesis test?	[If N<30] It must be assumed that the population parameter is normally distributed. [If a new variance is not given] It must be assumed that the population variance remains the same.
Upon changing the significance level / population variance, will his conclusion change?	[If H_0 is rejected, and α% increases or the variance decreases] No, the conclusion will not change. This is because… ● H_0 is already rejected at the α% significance level, and α% is less than the new significance level ● The p-value is already less than α%, and the p-value will further decrease if the variance decreases … so H_0 will still be rejected and the conclusion will remain the

	same. [If H_0 is not rejected, and α% increases or the variance decreases] Yes, the conclusion <u>may</u> change. This is because… ● H_0 is not rejected at the α% significance level, but the new significance level is greater than α% — and possibly the p-value ● The p-value is greater than α%, but the p-value will decrease — possibly below α% — if the variance decreases … so H_0 <u>may</u> be rejected and the conclusion <u>may</u> change. *Permute for the other possibilities…*
Random Variables	
Why is a binomial distribution suitable as a model?	This is because… ● There are only two mutually exclusive outcomes of 'success' and 'failure' ● The probability of 'success' remains constant ● Each trial is independent of every other trial
Why is a normal distribution unsuitable as a model?	This could be because… ● The distribution is not symmetrical about the mean ● The median is not equivalent to the mean ● The values are not distributed such that 68% / 95% / 99.7% of the values are within ±σ / ±2σ / ±3σ of the mean ● The probability of getting an impossible / negative value is ___, which is fairly high
What assumptions have to be made for your calculations?	It must be assumed that X_1 and X_2 are independent / X and Y are independent.
Why is the probability of Event A smaller than the probability of Event B?	This is because Event A is a proper subset of Event B — if Event A occurs Event B will definitely occur, but Event B could occur even if Event A does not occur. ● For example…
Correlation and Regression	
Based on the scatter plot, why is a linear model unsuitable?	A linear model is unsuitable because as y increases, x increases / decreases at an increasing / decreasing rate, so the points lie close to a curve and the relationship between x and y is non-linear.
Based on the value of r, why is a linear model unsuitable?	A linear model is unsuitable as |r| is not close to 1 / close to 0, so there is no linear relationship between the x and y.
Is Model A or Model B more suitable?	Under Model A, as x increases, y increases / decreases at an increasing / decreasing / constant rate. Conversely, under Model B, y increases / decreases at an increasing / decreasing / constant rate.

	From the scatter plot, since as x increases, y increases / decreases at an increasing / decreasing / constant rate, and the value of |r| under Model A is greater than the value of |r| under Model B, Model A is more suitable.
Why is the linear model unsuitable for large values of x?	The linear model is unsuitable as for large values of x, the linear model will give negative values of y, which are clearly impossible.
What is the significance of m in y=mx+c?	The value of m represents the increase in y for every unit increase of x.
Why should the regression line of y on x be used?	[If y depends on x] This is because y is the independent variable and x is the dependent variable. [If y does not depend on x] This is because there is no clear independent or dependent variable, and we are estimating the value of y based on a given value of x.
Why is the estimate reliable / unreliable?	[If it is reliable] The estimate is reliable as… ● The data is within the given data range, so we are interpolating ● The value of |r| is close to 1 [If it is unreliable] The estimate is unreliable as… ● The data is outside the given data range, so we are extrapolating ● The value of |r| is not close to 1
Why is it called the “least squares regression line”?	This is because the least squares regression line is the line for which the sum of the squares of the residuals is minimised.
Why are the residuals squared?	This is done to ensure the residuals will not cancel each other out when summed and the sum of the residuals will not be negative, as residuals can be either positive (when they are above the regression line) or negative (when they are below the regression line). *In addition, squaring makes the larger residuals even larger, increasing their impact on the sum of the residuals.*
Will the scaling of units affect the value of r?	No, the scaling of units will not affect the value of r, as the relationship between the variables remains unchanged.
Why should we remove a data point?	This is because the data point is an outlier which may skew or distort the relationship between the two variables, negatively affecting the accuracy of predictions.

	[If applicable] This is because the data point was the result of an inaccurate measurement, or it is not part of the population we are studying.