# Statistics for Data Science Quiz Questions

## Section 1 – Introduction

### 1.2. Population vs sample

1.2.1. The high school principal asked you to conduct a survey on student satisfaction for the entire school. You must contact your classmates about their opinion. Then you present the results to the principal. Would you say this was population or sample data? What is the value you presented called?

a) population, statistic

b) population, parameter

**c) sample, statistic**

d) sample, parameter

Explanation: You interviewed only one class – your own. This is a sample taken from the population of all students studying in the high school. Since it was a sample, the value associated with it was a statistic.

1.2.2. You are trying to estimate the average valuation of start-ups in the USA. Imagine you are able to visit 200 start-ups in Silicon Valley in a random manner. What is a possible problem of your study?

a) The sample is not random.

b) The sample is too small.

**c) The sample is not representative.**

d) The population is unknown.

Explanation: You were estimating a statistic for all start-ups in the USA. However, your sample was taken only from startups in Silicon Valley. While this may be a proxy for California, it does not represent the other 49 states. Therefore, the sample is not representative.

## Section 2 – Descriptive Statistics Fundamentals

### 2.1. Types of data

2.1.1. A variable represents the weight of a person. What type of data does it represent?

a) categorical, discrete

b) categorical, continuous

c) numerical, discrete

**d) numerical, continuous**

Explanation: The data is numerical, as it represents a number. As we said in the lesson, weight is a continuous variable.

2.1.2. A variable represents the gender of a person. What type of data does it represent?

**a) categorical**

b) numerical, discrete

c) numerical, continuous

## 2.2. Levels of Measurement

2.2.1. A variable represents the gender of a person. What is the level of measurement?

**a) nominal**

b) ordinal

c) interval

d) ratio

Explanation: Gender is a nominal variable. The possible categories cannot be put in any order.

2.2.2. A variable represents the weight of a person. What is the level of measurement?

a) nominal

b) ordinal

c) interval

**d) ratio**

Explanation: Weight is a ratio variable. It is quantitative and there is a true zero point.

# Section 3 – Inferential Statistics Fundamentals

## 3.2. What is a distribution

3.2.1. What is a distribution?

a) A graph representing the probability of occurrence of a variable

b) A graph representing the possible values of a variable and the probability of their occurrence

**c) A distribution is a function that shows the possible values for a variable and the probability of their occurrence.**

d) A distribution is a function that shows the probability of a variable.

## 3.3. The Normal distribution

3.3.1. Choose the INCORRECT answer. A Normal distribution is one of the most commonly used distribution types, because:

**a) All variables can be represented by the Normal distribution.**
b) Distributions of sample means with large enough sample sizes could be approximated to Normal.
c) Computable statistics are elegant.
d) Decisions based on Normal distribution insights have a good track record.

Explanation: The Normal distribution approximates a wide number of random variables. However, not all random variables follow a Normal distribution.

## 3.5. The central limit theorem

3.5.1. Choose the INCORRECT answer about the central limit theorem.

a) The means of the samples we extract will be closer to normally distributed if we extract more samples.
b) The distribution of the sample means is expected to have a mean equal to the mean of the original dataset.
c) The distribution of the sample means is expected to have a variance equal to the variance of the original dataset, divided by the sample size.
**d) The means of the samples we extract will be closer to normally distributed, the smaller the samples we extract.**

Explanation: The means of the samples we extract will be closer to normally distributed, the LARGER the samples we extract.

## 3.7. Estimators and estimates

3.7.1. You are given a dataset with a sample mean of 10. In this case, 10 is:

a) a point estimator
**b) a point estimate**
c) an interval estimator
d) an interval estimate

Explanation: A point estimate is a single number given by an estimator. The estimator in this case is a point estimator and is the formula for the mean.

## 3.8. Definition of confidence intervals

3.8.1. How is a confidence interval related to a point estimate?

a) Whenever we can't calculate an interval, the confidence interval is equal to the point estimate.
b) The point estimate is the starting point of the interval.
**c) The point estimate is the midpoint of the interval.**

d) The point estimate is the ending point of the interval.

### 3.10. Student's T distribution

3.10.1. The Student's T distribution:

**a) Approximates the Normal distribution but has fatter tails**

b) Approximates the Normal distribution but has bigger probability

c) Approximates the Normal distribution only for big samples

d) Is a particular case of the Normal distribution

Explanation: The Student's T distribution approximates the Normal distribution but has fatter tails. This means the probability of values being far away from the mean is bigger. For big enough samples, the Student's T distribution coincides with the Normal distribution.

### 3.12. Margin of error

3.12.1. Which of the following would decrease the margin of error?

a) A higher statistic and a higher sample size

b) A higher statistic and a higher standard deviation

c) A lower statistic and a higher standard deviation

**d) A lower statistic and a higher sample size**

Explanation: A higher statistic increases the ME. A higher standard deviation increases the ME. A higher sample size decreases the ME. Therefore: a) and c) are ambiguous. as we don't know which effect is stronger. With b), the ME will definitely increase. In d) both statements decrease the ME.

## Section 4 – Hypothesis Testing

### 4.1. Null vs alternative

4.1.1. You own an ice-cream shop. You suspect you sell as many ice-creams as your competitor, Gelatou. Gelatou sells 100 ice-creams per day. State the null and alternative hypotheses of the test that would answer your question:

a) H0: The average number of ice-creams our shop sells is 100.
H1: The average number of ice-creams our shop sells differs from 100.

**b) H0: The average number of ice-creams our shop sells differs from 100.**
**H1: The average number of ice-creams our shop sells is 100.**

Explanation: Since you are testing if you are selling as much as your competitor, then this is your alternative hypothesis. The null hypothesis is the one we are trying to reject, so it comprises the statement opposite to our sentiment.

4.1.2. You want to check if your height is above average, compared to your classmates. State the null and alternative hypotheses of this test.

a) H0: My height is higher or equal to the average height in the class.
H1: My height is lower than the average height in the class.

b) H0: My height is higher than the average height in the class.

H1: My height is lower or equal to the average height in the class.
**c) H0: My height is lower or equal to the average height in the class.**
**H1: My height is higher than the average height in the class.**
d) H0: My height is lower than the average height in the class.
H1: My height is higher or equal to the average height in the class.

Explanation: Once again, the null hypothesis states the opposite statement. You want to check if your height is above average; therefore, the null hypothesis of the test is 'lower or equal'. You have to include 'equal' in the null hypothesis, as you want to understand if you are ABOVE average.

4.1.3. You want to test if the Obama administration issued fewer executive orders than the Bush administration. State the null and alternative hypotheses of this test.

a) H0: The Obama administration issued more executive orders than the Bush administration.
H1: The Obama administration issued fewer executive orders than the Bush administration.
b) H0: The Obama administration issued fewer executive orders than the Bush administration.
H1: The Obama administration issued more executive orders than the Bush administration.
c) H0: The Obama administration issued at least as many executive orders as the Bush administration.
H1: The Obama administration issued more executive orders than the Bush administration.
**d) H0: The Obama administration issued at least as many executive orders as the Bush administration.**
**H1: The Obama administration issued fewer executive orders than the Bush administration.**

Explanation: We want to test if the Obama administration issued fewer executive orders than the Bush administration. Therefore, the null hypothesis is: 'The Obama administration issued at least as many executive orders as the Bush administration.' The alternative hypothesis encompasses everything else; therefore, the alternative is: The Obama administration issued fewer executive orders than the Bush administration. Answer c) the null hypothesis is right, but the alternative was wrong. The answer is d).

## 4.2. Rejection region and significance level

4.2.1. Which value is NOT typical when assigning a significance level?

a) 0.05
b) 0.01
**c) 0.5**
d) 0.1

Explanation: 0.5 is a very high level of significance. Testing with it will not give much credibility to your test. There is 50% chance you are wrong.

4.2.2. If your test value falls into the rejection region:

a) You will reject the test.
b) You will reject the significance level.
c) You will reject the alternative hypothesis.
**d) You will reject the null hypothesis.**

Explanation: If the test value falls into the rejection region, you will reject the null hypothesis. 'You will reject the test' is a meaningless phrase, as the test is not accepted or rejected; it is executed. The

outcome of the test is rejection or acceptance of the null hypothesis. Similarly, rejecting the significance level has no meaning.

## 4.3. Type I error vs type II error

4.3.1. Type II error is also known as:

a) false error
**b) false negative**
c) false positive
d) it has no other name

4.3.2. Type I error or false positive comprises:

a) accepting a null hypothesis that is true
**b) rejecting a null hypothesis that is true**
c) accepting a null hypothesis that is false
d) rejecting a null hypothesis that is false

4.3.3. What is another way to call the probability of rejecting a null hypothesis that is false?

a) Type I error
b) Type II error
c) Type III error
**d) Power of the test**

4.3.4. You are taking a pregnancy test. The null hypothesis of this test is: I am not pregnant. In reality, you are not pregnant, but the test says you are. Which type of error occurred?

a) No error was made.
b) Both errors were made at the same time.
**c) Type I error**
d) Type II error

Explanation: This is a Type I error. You rejected a null hypothesis that is true. It is also called a false positive, although in this case, it is counter intuitive. While usually no error is made (that's why hypothesis testing is so wide-spread), it is impossible to make both errors at the same time.

## 4.5. p-value

4.5.1. You have a z-score of 2.31. What is the p-value of this test?

a) 0.010
b) 0.021
c) none of those
**d) not enough information**

Explanation: Since we haven't specified if this is a one-tailed or a two-tailed test, there is not enough information to answer this question.

4.5.2. You have a z-score of 2.31 for a one-tailed test. What is the p-value of this test?

**a) 0.010**
b) 0.021
c) none of those
d) not enough information

Explanation: The p-value for a one-tailed test is 0.010. The p-value for a two-tailed test is twice as big, therefore 0.021 (while 0.020 is good enough – differences occur due to rounding).

# Section 5 – Regression Analysis

## 5.1. Introduction

5.1.1. Are you ready for regression analysis?

**a) Yes**
b) No

## 5.3. The linear regression model

5.3.1. You have an ice-cream shop. You noticed a relationship between the number of cones you order and the number of ice-creams you sell. Is this a suitable situation for regression analysis?

a) Yes
**b) No**

Explanation: While it is true that, if you run out of cones, you cannot sell anymore ice-creams, this is not regression analysis material. The two variables go hand-in-hand as (usually) each ice-cream requires a cone.

5.3.2. You are trying to predict the amount of beer consumed in the US, depending on the state. Is this regression material?

**a) Yes**
b) No

Explanation: Yes, logic shows us that, in different states, people drink different amounts of beer. Some states are warmer; other are colder etc. While many more things will be a part of this regression, such as gender, income etc., this is a good basis for regression analysis.

## 5.4. Correlation vs regression

5.4.1. Which statement is false?

a) Correlation does not imply causation.
b) Correlation is symmetrical regarding both variables.
**c) Correlation could be represented as a line.**
d) Correlation does not capture the direction of the causal relationship.

Explanation: Correlation is a single point. It cannot be represented as a line.

## 5.7. Decomposition

5.7.1. Which of the following is true?

**a) SST = SSR + SSE**
b) SSR = SST + SEE
d) SSE = SST + SSR

## 5.8. R-squared

5.8.1. SST = 1245, SSR = 945, SSE = 300. What is the R-squared of this regression?

a) 0.24
b) 0.52
**c) 0.76**
d) 0.87

Explanation: The R-squared is equal to SSR, divided by SST.

5.8.2. The R-squared is a measure that:

a) measures how well your data fits the regression line
b) measures how well your regression line fits your data
c) measures how well your data fits your model
**d) measures how well your model fits your data**

Explanation: The R-squared shows how much of the total variability of the dataset is explained by your regression model. This may be expressed as: how well your model fits your data. It is incorrect to say your regression line fits the data, as the line is the geometrical representation of the regression equation. It also incorrect to say the data fits the model or the regression line, as you are trying to explain the data with a model, not vice versa.

## 5.12. Adjusted R^2

5.12.1. The adjusted R-squared is a measure that:

a) measures how well your model fits the data
**b) measures how well your model fits the data but penalizes the excessive use of variables**
c) measures how well your model fits the data but penalizes excessive use of p-values
d) measures how well your data fits the model but penalizes the excessive use of variables

Explanation: Like the R-squared, the adjusted R-squared measures how well your model fits the data. However, it penalizes the use of variables that are meaningless for the regression.

5.12.2. The adjusted R-squared is:

a) usually bigger than the R-squared
**b) usually smaller than the R-squared**
c) usually the same as the R-squared
d) incomparable to the R-squared

Explanation: Almost always, the adjusted R-squared is smaller than the R-squared. The statement is not true only in the extreme occasions of small sample sizes and a high number of independent variables.

## 5.14. Assumptions

5.14.1. If a regression assumption is violated:

a) Some things change.
b) You cannot perform a regression.
**c) Performing regression analysis will yield an incorrect result.**
d) It is no big deal.

Explanation: Nothing stops you from performing the regression analysis, but it will yield an incorrect result. It is a big deal.

## 5.15. A1. Linearity

5.15.1. Linearity is easy to relax.

**a) True**
b) False

## 5.16. A2. No endogeneity

5.16.1. The easiest way to detect an omitted variable bias is through:

**a) the error term**
b) the independent variables
c) the dependent variable
d) sophisticated software

Explanation: Omitted variable bias occurs when you forget to include a variable. This is reflected in the error term as the factor you forgot about is included in the error. In this way, the error is not random but includes a systematic part (the omitted variable).

## 5.18. A4. No autocorrelation

5.18.1. Autocorrelation is not likely to be observed in:

a) time series data
b) sample data
c) panel data
**d) cross-sectional data**

Explanation: Autocorrelation is not observed in cross-sectional data. You usually spot it at time series data, which is a subset of panel data. Sample data is not relevant for this question.

## 5.19. A5. No multicollinearity

5.19.1. No multicollinearity is:

**a) easy to spot and easy to fix**
b) easy to spot but hard to fix
c) hard to spot but easy to fix
d) hard to spot and hard to fix