Statistics

1 The central limit theorem is a statistical theorem that states that the sampling distribution of the mean of a large number of independent, identically distributed random variables will be approximately normally distributed, regardless of the distribution of the individual variables.

This theorem is important because it allows us to make statistical inferences about a population based on a sample drawn from that population. It allows us to estimate the mean and standard deviation of a population based on a sample, and to test hypotheses about the population parameters.

The central limit theorem is widely used in statistical analysis and has many practical applications. It is used to estimate population parameters, to test hypotheses about population parameters, and to construct confidence intervals for population parameters. It is also used in the analysis of statistical data, in the design of experiments, and in the interpretation of statistical results.

2 Sampling is the process of selecting a subset of individuals from a larger population to study in order to draw conclusions about the population as a whole. Sampling is an important part of statistical analysis because it allows researchers to study a portion of a population rather than the entire population, which can be time-consuming, costly, and sometimes impractical.

There are several different methods of sampling, including:

Simple random sampling: This method involves selecting a sample of individuals from the population at random, with each individual having an equal chance of being selected.

Stratified sampling: This method involves dividing the population into subgroups or strata based on certain characteristics, and then selecting a random sample from each stratum.

Cluster sampling: This method involves dividing the population into groups or clusters, and then selecting a random sample of clusters to study.

Systematic sampling: This method involves selecting every nth individual from the population, where n is determined by the size of the sample and the size of the population.

Convenience sampling: This method involves selecting individuals from the population who are readily available or convenient to study.

Quota sampling: This method involves selecting a sample that is representative of certain characteristics of the population, such as age, gender, or income level.

Snowball sampling: This method involves selecting a small number of individuals from the population and then asking them to refer others who meet the sampling criteria.

3 In statistical hypothesis testing, a type I error is a error that occurs when the null hypothesis is rejected when it is actually true. This error is also known as a "false positive" or "false alarm." It is represented by the Greek letter alpha (α) and is typically set at a small value, such as 0.05 or 0.01, to minimize the probability of a type I error.

A type II error, on the other hand, is a error that occurs when the null hypothesis is not rejected when it is actually false. This error is also known as a "false negative" or "miss." It is represented by the Greek letter beta (β) and is typically set at a small value, such as 0.10 or 0.20, to minimize the probability of a type II error.

In hypothesis testing, it is important to consider both type I and type II errors, as making either type of error can have serious consequences. For example, in the medical field, a type I error could lead to unnecessary treatment or surgery, while a type II error could result in a failure to diagnose or treat a disease.

4 Normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is defined by its mean and standard deviation. It is a symmetrical distribution, with the majority of the observations clustered around the mean and fewer observations as you move away from the mean in either direction.

The normal distribution is also important because of the central limit theorem, which states that the sampling distribution of the mean of a large number of independent, identically distributed random variables will be approximately normally distributed, regardless of the distribution of the individual variables.

5 Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. It is expressed as a value between -1 and 1, where:

- A value of -1 indicates a strong negative relationship, meaning that as one variable increases, the other variable decreases.
- A value of 0 indicates no relationship between the two variables.
- A value of 1 indicates a strong positive relationship, meaning that as one variable increases, the other variable also increases.

Covariance is a measure of the degree to which two variables vary together. It is calculated as the product of the standard deviations of the two variables and the correlation between them. Covariance can be positive, negative, or zero. A positive covariance indicates that the two variables tend to vary together, while a negative covariance indicates that the two variables tend to vary in opposite directions. A zero covariance indicates that there is no relationship between the two variables.

Both correlation and covariance are used to analyze the relationship between two variables. Correlation is generally easier to interpret and is more commonly used, but covariance is more general and can be used in more advanced statistical analyses.

6 Univariate analysis is the analysis of a single variable. It involves exploring the characteristics of the variable, such as its mean, median, mode, range, and standard deviation, and may include graphical representations of the data, such as histograms or box plots.

Bivariate analysis is the analysis of two variables. It involves exploring the relationship between the two variables, such as whether there is a correlation or causal relationship between them. Bivariate analysis may include techniques such as scatterplots, correlation analysis, and regression analysis.

Multivariate analysis is the analysis of three or more variables. It involves exploring the relationships between the variables and may include techniques such as principal component analysis, factor analysis, and multivariate regression analysis. Multivariate analysis is often used in situations where

there are many variables that may be related to each other and it is necessary to identify the most important variables and understand how they are related.

7 Sensitivity is a measure of the ability of a test or model to correctly identify individuals who have a certain condition or characteristic. It is calculated as the number of true positive results divided by the total number of individuals with the condition.

For example, suppose a medical test is used to diagnose a certain disease and it is applied to 100 individuals known to have the disease and 100 individuals known not to have the disease. The test produces the following results:

- True positive: The test correctly identifies 95 individuals with the disease.
- False negative: The test incorrectly identifies 5 individuals without the disease as having the disease.
- False positive: The test incorrectly identifies 10 individuals without the disease as having the disease.
- True negative: The test correctly identifies 90 individuals without the disease.

To calculate the sensitivity of the test, we would use the following formula:

Sensitivity = True positive / (True positive + False negative) = 95 / (95 + 5) = 0.95

This means that the test correctly identified 95 out of 100 individuals with the disease, or 95% of the individuals with the disease. Sensitivity is often expressed as a decimal or as a percentage, and is typically represented by the symbol "Se" or "Sen."

Sensitivity is an important characteristic of a diagnostic test because it tells us how well the test can identify individuals who have a certain condition. A test with high sensitivity is more likely to correctly identify individuals with the condition, while a test with low sensitivity is more likely to miss individuals with the condition. Sensitivity is often used in conjunction with specificity, which is a measure of the ability of a test to correctly identify individuals who do not have a certain condition or characteristic.

8 Hypothesis testing is a statistical procedure that is used to determine whether a hypothesis about a population parameter is supported by the data. It involves formulating a null hypothesis, which is a statement about the population that is assumed to be true unless there is sufficient evidence to reject it, and an alternative hypothesis, which is a statement about the population that is tested against the null hypothesis.

The null hypothesis, denoted as H0, represents the status quo or the default assumption, and is usually a statement of no difference or no relationship between the variables being studied. For example, in a study to determine whether a new drug is effective in treating a certain medical condition, the null hypothesis might be that the drug has no effect on the condition.

The alternative hypothesis, denoted as H1, represents the opposite of the null hypothesis and is the hypothesis that is tested against the null hypothesis. In the example above, the alternative hypothesis might be that the drug is effective in treating the medical condition.

In a two-tail test, the null and alternative hypotheses are formulated in such a way that the test can reject the null hypothesis if the data are either too high or too low. For example, in the study of the effectiveness of the new drug, the null hypothesis might be that the drug has no effect on the

medical condition, and the alternative hypothesis might be that the drug either increases or decreases the effectiveness of the treatment. In this case, the two-tail test would reject the null hypothesis if the data showed that the drug was either more or less effective than the current treatment

9 Quantitative data is numerical data that can be measured and quantified. It can be continuous, meaning that it can take on any value within a given range, or discrete, meaning that it can only take on specific values. Examples of quantitative data include height, weight, age, and income.

Qualitative data is non-numerical data that cannot be measured or quantified. It is used to describe characteristics or attributes of a phenomenon or a group. Qualitative data can be categorical, meaning that it falls into one of several categories, or ordinal, meaning that it can be ranked or ordered. Examples of qualitative data include race, gender, marital status, and occupation.

10 To calculate the range of a set of data, you need to find the difference between the highest and lowest values in the data set. The formula for calculating the range is:

Range = Maximum value - Minimum value

For example, suppose you have a data set containing the following values: 2, 4, 6, 8, 10

To find the range of this data set, you would first find the maximum value, which is 10, and the minimum value, which is 2. Then you would subtract the minimum value from the maximum value to get the range:

Range = 10 - 2 = 8

The interquartile range (IQR) is a measure of the dispersion of a data set, similar to the standard deviation. It is defined as the difference between the upper quartile (Q3) and the lower quartile (Q1). The upper quartile is the value that separates the top 25% of the data from the bottom 75%, and the lower quartile is the value that separates the bottom 25% of the data from the top 75%.

To calculate the interquartile range, you first need to find the upper and lower quartiles of the data set. To do this, you need to:

1. Sort the data in ascending order.
2. Find the median of the data set (Q2).
3. Divide the data set into two halves, with the lower half containing the values below the median and the upper half containing the values above the median.
4. Find the median of the lower half of the data (Q1).
5. Find the median of the upper half of the data (Q3).

Then you can use the following formula to calculate the interquartile range:

IQR = Q3 - Q1

For example, suppose you have the following data set: 2, 4, 5, 6, 8, 9, 10

To find the interquartile range, you would first need to sort the data in ascending order: 2, 4, 5, 6, 8, 9, 10

Then you would find the median of the data set, which is 6. The lower half of the data is 2, 4, 5, and the upper half is 8, 9, 10. The medians of the lower and upper halves are 4 and 9, respectively.

Finally, you would use the formula to calculate the interquartile range:

IQR = Q3 - Q1 = 9 - 4 = 5

The interquartile range is a useful measure of dispersion because it is less sensitive to outliers than the range or the standard deviation. It is often used in statistical analysis to describe the spread of a data set.

11 The bell curve, also known as the normal distribution or Gaussian distribution, is a continuous probability distribution that is defined by its mean and standard deviation. It is a symmetrical distribution, with the majority of the observations clustered around the mean and fewer observations as you move away from the mean in either direction.

The bell curve is a useful model for many real-world phenomena, such as IQ scores, height, weight, and blood pressure, because it can accurately describe the distribution of many continuous variables. The normal distribution is also important because of the central limit theorem, which states that the sampling distribution of the mean of a large number of independent, identically distributed random variables will be approximately normally distributed, regardless of the distribution of the individual variables.

The bell curve is defined by the following probability density function:

f(x) = 1/(σ * √(2π)) * e^(-(x - μ)^2 / (2σ^2))

where:

- x is the value of the random variable
- μ is the mean of the distribution
- σ is the standard deviation of the distribution
- π is the constant approximately equal to 3.14159
- e is the base of the natural logarithm, approximately equal to 2.71828

The bell curve is often represented graphically as a symmetrical curve with a "bell" shape, as the name suggests. The x-axis represents the values of the variable, and the y-axis represents the probability of each value occurring. The mean is located at the peak of the curve, and the standard deviation determines the width of the curve.

12 One method for finding outliers in a data set is to use the interquartile range (IQR) method. This method involves calculating the difference between the upper quartile (Q3) and the lower quartile (Q1) of the data set, and then identifying values that are outside of the range defined by the following formula:

Lower bound = Q1 - 1.5 * IQR Upper bound = Q3 + 1.5 * IQR

Values that fall outside of this range are considered to be outliers.

To use the IQR method, you first need to calculate the upper and lower quartiles of the data set, as well as the interquartile range. To do this, you need to:

1. Sort the data in ascending order.
2. Find the median of the data set (Q2).
3. Divide the data set into two halves, with the lower half containing the values below the median and the upper half containing the values above the median.
4. Find the median of the lower half of the data (Q1).
5. Find the median of the upper half of the data (Q3).
6. Calculate the interquartile range as Q3 - Q1.

Then you can use the lower bound and upper bound formulas to identify values that are outside of the range defined by the interquartile.

13 In hypothesis testing, the p-value is a measure of the strength of the evidence against the null hypothesis. It is the probability of obtaining a result as extreme or more extreme than the one observed, given that the null hypothesis is true.

The p-value is calculated based on the test statistic and the sampling distribution of the test statistic under the null hypothesis. If the p-value is small, it means that the observed result is unlikely to have occurred by chance if the null hypothesis is true, and the null hypothesis can be rejected. The smaller the p-value, the stronger the evidence against the null hypothesis.

The p-value is often used as a threshold for determining whether to reject or fail to reject the null hypothesis. It is generally accepted that a p-value of less than 0.05 is considered to be statistically significant, meaning that there is a less than 5% chance of the observed result occurring by chance if the null hypothesis is true. This means that if the p-value is less than 0.05, the null hypothesis can be rejected and the alternative hypothesis can be accepted.

It is important to note that the p-value is not a measure of the truth of the null or alternative hypotheses, but rather a measure of the strength of the evidence against the null hypothesis. A small p-value does not necessarily mean that the alternative hypothesis is true, but rather that the observed result is unlikely to have occurred by chance if the null hypothesis is true.

14 The binomial probability formula is a statistical formula used to calculate the probability of a specific outcome in a binomial experiment, which is a type of experiment that has only two possible outcomes. A binomial experiment is defined by the following characteristics:

- The experiment consists of a fixed number of trials, n.
- Each trial has only two possible outcomes, which are usually referred to as "success" and "failure."
- The probability of success, p, is the same for each trial.
- The trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

The binomial probability formula is used to calculate the probability of a specific number of successes, x, in n trials. It is given by the following formula:

15 ANOVA, or analysis of variance, is a statistical procedure used to test whether there are significant differences between the means of two or more groups. It is a hypothesis testing technique that compares the variance between groups to the variance within groups.

ANOVA is used to determine whether there is a significant difference between the means of two or more groups, or whether the differences observed between the groups are simply due to chance. It is a useful tool for comparing the means of groups and determining whether the groups are significantly different from one another.

ANOVA can be used in a variety of applications, including:

- Comparing the means of two or more groups to determine whether there is a significant difference between the groups.
- Testing the effectiveness of a new treatment or intervention by comparing the mean outcomes of a treatment group to a control group.
- Analyzing data from experiments with more than one independent variable, in order to determine which variables have a significant effect on the dependent variable.

ANOVA is a widely used statistical technique in many fields, including psychology, economics, and biology. It is a useful tool for comparing group means and determining whether the differences between the groups are statistically significant.

Machine Learning

1 c

2 c

3 c

4 a

5 c

6 b

7 c

8 b,c

9 a,b,d

10 a,b,d

11 Outliers are observations that are significantly different from the rest of the data in a data set. They may be extreme values that fall outside of the range of most of the other observations, or they may be values that are substantially different from the mean or median of the data set. Outliers can have a significant impact on the statistical properties of a data set, such as the mean and standard deviation, and can potentially affect the results of statistical analyses.

The Inter Quartile Range (IQR) method is a statistical method for identifying outliers in a data set. It involves calculating the difference between the upper quartile (Q3) and the lower quartile (Q1) of the data set, and then identifying values that are outside of the range defined by the following formula:

Lower bound = Q1 - 1.5 * IQR Upper bound = Q3 + 1.5 * IQR

Values that fall outside of this range are considered to be outliers.

To use the IQR method, you first need to calculate the upper and lower quartiles of the data set, as well as the interquartile range. To do this, you need to:

1. Sort the data in ascending order.
2. Find the median of the data set (Q2).
3. Divide the data set into two halves, with the lower half containing the values below the median and the upper half containing the values above the median.
4. Find the median of the lower half of the data (Q1).
5. Find the median of the upper half of the data (Q3).
6. Calculate the interquartile range as Q3 - Q1.

Then you can use the lower bound and upper bound formulas to identify values that are outside of the range defined by the interquartile range. The IQR method is a useful tool for identifying outliers because it is less sensitive to the presence of extreme values than other methods, such as the range or standard deviation.

12 Bagging and boosting are two ensemble learning techniques that are used to improve the performance of machine learning models. Ensemble learning involves training multiple models and combining their predictions to make a more accurate overall prediction.

The primary difference between bagging and boosting algorithms is the way in which the individual models are trained and combined.

Bagging algorithms, or bootstrap aggregating algorithms, train multiple models independently on different subsets of the training data, and then combine their predictions by averaging or voting. The goal of bagging is to reduce the variance of the ensemble model, by training models that are slightly different from one another and averaging their predictions.

Boosting algorithms, on the other hand, train multiple models sequentially, with each model attempting to correct the errors of the previous model. The goal of boosting is to reduce the bias of the ensemble model, by training models that focus on the errors of the previous model and gradually improving the overall accuracy of the ensemble.

Overall, the primary difference between bagging and boosting algorithms is the way in which the individual models are trained and combined. Bagging algorithms train multiple models independently and combine their predictions by averaging or voting, while boosting algorithms train multiple models sequentially and focus on correcting the errors of the previous model.

13 In linear regression, the R2 value, or coefficient of determination, is a measure of the strength of the relationship between the predictor variables and the response variable. It is a measure of how well the model fits the data, and is calculated as the proportion of the variance in the response variable that is explained by the predictor variables.

The adjusted R2 value, also known as the adjusted coefficient of determination, is a modified version of the R2 value that takes into account the number of predictor variables in the model. It is a more conservative measure of the model fit because it penalizes models with more predictor variables, even if those variables improve the fit of the model.

The adjusted R2 value is calculated as follows:

Adjusted R2 = 1 - (1 - R2) * (n - 1) / (n - p - 1)

where:

- R2 is the coefficient of determination.
- n is the number of observations in the data set.
- p is the number of predictor variables in the model.

The adjusted R2 value is a useful tool for comparing the fit of different linear regression models, because it adjusts for the number of predictor variables in the model and provides a more accurate measure of the model fit. It is generally considered to be a more reliable measure of the model fit than the R2 value, particularly when comparing models with different numbers of predictor variables.

14 Standardization and normalization are two techniques that are used to scale variables in a data set. Both techniques are used to transform variables to a common scale, which can be useful for comparison purposes and for certain types of statistical analyses.

Standardization, also known as z-score normalization, involves transforming the variables to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the variable from each value and dividing the result by the standard deviation. The formula for standardization is:

z = (x - mean) / std

where:

- z is the standardized value of x.
- x is the original value of the variable.
- mean is the mean of the variable.
- std is the standard deviation of the variable.

Standardization is useful when the variables have different scales and you want to compare their relative magnitudes. It is also useful when the variables are normally distributed, because the transformed values will also be normally distributed.

Normalization, on the other hand, involves transforming the variables to have a range of 0 to 1. This is done by subtracting the minimum value of the variable from each value and dividing the result by the range of the variable. The formula for normalization is:

x' = (x - min) / (max - min)

where:

- x' is the normalized value of x.
- x is the original value of the variable.
- min is the minimum value of the variable.
- max is the maximum value of the variable.

Normalization is useful when the variables have different scales and you want to compare their relative proportions. It is also useful when you want to ensure that the transformed values have a common range, which can be important for certain types of machine learning algorithms.

Overall, the main difference between standardization and normalization is the way in which the variables are transformed. Standardization transforms the variables to have a mean of 0 and a standard deviation of 1, while normalization transforms the variables to have a range of 0 to 1.

15 Cross-validation is a statistical technique that is used to evaluate the performance of machine learning models and to select the best model. It involves dividing the data into a number of folds or subsets, training the model on one or more of the folds, and evaluating the model on the remaining fold or folds. This process is repeated a number of times, with each fold serving as the test set at least once. The performance of the model is then averaged across all of the iterations, providing a more reliable estimate of the model's performance.

One advantage of using cross-validation is that it allows you to evaluate the performance of a machine learning model on multiple different subsets of the data, providing a more reliable estimate of the model's generalization ability. This is particularly useful when working with small data sets, as it allows you to make better use of the available data for model training and evaluation.

One disadvantage of using cross-validation is that it can be computationally intensive, particularly for large data sets or models with many parameters. It can also be time-consuming to implement, as it requires training and evaluating the model multiple times on different subsets of the data. However, in most cases, the benefits of using cross-validation outweigh the costs, making it a useful technique for evaluating the performance of machine learning models.

SQL

```
1  SELECT shippedDate, AVG(orderNumber)
FROM orders GROUPBY shippedDate;

2  SELECT orderDate, AVG(orderNumber)
FROM orders GROUPBY orderDate;

3 SELECT productname, MIN(MSRP)
FROM products;

4 SELECT productname, MAX(quantityinstock)
FROM products;

5 SELECT p.productname, MAX(o.quantityordered)
FROM orderdetails o
INNER JOIN products p
ON o.productcode = p.productcode
GROUP BY p.productname;

6 SELECT c.customername, MAX(p.amount)
FROM payments p
INNER JOIN customers c
ON p.customerNumber = c.customerNumber
GROUP BY c.customername;

7 SELECT customerNumber, customerName
FROM customers
WHERE city = 'Melbourne city';
```

```
8 SELECT customerName
FROM customers
WHERE customerName LIKE 'N%';

9 SELECT customerName
FROM customers
WHERE phone LIKE '7%' AND city = 'LasVegas';

10 SELECT customerName
FROM customers
WHERE creditLimit < 1000 AND city IN ('Las Vegas', 'Nantes', 'Stavern');

11 SELECT orderNumber
FROM orderdetails
WHERE quantityOrdered < 10;

12 SELECT o.orderNumber
FROM orders o
WHERE c.customerName LIKE 'N%'
INNER JOIN customers c
ON o.customerNumber = c.customerNumber;

13 SELECT c.customerName
FROM customers c
WHERE o.status = 'Disputed'
INNER JOIN orders o
ON c.customerNumber = o.customerNumber;
```

```
14 SELECT c.customerName
FROM customers c
WHERE p.checkNumber LIKE 'H%' AND p.paymentDate = '2004-10-19'
INNER JOIN payments p
ON c.customerNumber = p.customerNumber;

15 SELECT checkNumber
FROM payments
WHERE amount > 1000;
```