

Statistics worksheet 3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
 - a) Total Variation = Residual Variation – Regression Variation
 - b) Total Variation = Residual Variation + Regression Variation
 - c) Total Variation = Residual Variation * Regression Variation
 - d) All of the mentioned.
 - b) Total Variation = Residual Variation + Regression Variation
2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
 - a) random
 - b) direct
 - c) binomial
 - d) none of the mentioned
 - c) binomial
3. How many outcomes are possible with Bernoulli trial?
 - a) 2
 - b) 3
 - c) 4
 - d) None of the mentioned
 - a) 2
4. If H_0 is true and we reject it is called
 - a) Type-I error
 - b) Type-II error
 - c) Standard error
 - d) Sampling error
 - a) Type-I error
5. Level of significance is also called:
 - a) Power of the test
 - b) Size of the test
 - c) Level of confidence
 - d) Confidence coefficient
 - b) Size of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
- a) Decrease
 - b) Increase
 - c) Both of them
 - d) None
- b) Increased
7. Which of the following testing is concerned with making decisions using data?
- a) Probability
 - b) Hypothesis
 - c) Causal
 - d) None of the mentioned
- b) Hypothesis
8. What is the purpose of multiple testing in statistical inference?
- a) Minimize errors
 - b) Minimize false positives
 - c) Minimize false negatives
 - d) All of the mentioned
- d) All of the mentioned
9. Normalized data are centred at and have units equal to standard deviations of the original data
- a) 0
 - b) 5
 - c) 1
 - d) 10
- a) 0

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

The theorem is named after English statistician, Thomas Bayes, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference.

In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance. Some of the applications include but are not limited to, modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

The diagram illustrates Bayes' Theorem with the following components:

- LIKELIHOOD**: The probability of "B" being True, given "A" is True. An arrow points from this label to the numerator term $P(B|A)$.
- PRIOR**: The probability "A" being True. This is the knowledge. An arrow points from this label to the numerator term $P(A)$.
- POSTERIOR**: The probability of "A" being True, given "B" is True. An arrow points from this label to the left side of the equation, $P(A|B)$.
- MARGINALIZATION**: The probability "B" being True. An arrow points from this label to the denominator term $P(B)$.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Note that events A and B are [independent events](#) (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

A special case of the Bayes' theorem is when event A is a [binary variable](#). In such a case, the theorem is expressed in the following way:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A^-)P(A^-) + P(B|A^+)P(A^+)}$$

Where:

- $P(B|A^-)$ – the probability of event B occurring given that event A^- has occurred
- $P(B|A^+)$ – the probability of event B occurring given that event A^+ has occurred

In the special case above, events A^- and A^+ are mutually exclusive outcomes of event A.

11. What is z-score?

A z-score gives us an idea of how far from the mean a data point is. It is an important topic in statistics. Z-scores are a method to compare results to a "normal" population. For

example, we know someone's weight is 70 kg, but if you want to compare it to the "average" person's weight, looking at a vast table of data can be overwhelming. A z-score gives us an idea of where that person's weight is compared to the average population's mean weight. In this article, we will learn what is z score.

A measure of how many standard deviations below or above the population mean a raw score is called z score. It will be positive if the value lies above the mean and negative if it lies below the mean. It is also known as standard score. It indicates how many standard deviations an entity is, from the mean. In order to use a z-score, the mean μ and also the population standard deviation σ should be known. A z score helps to calculate the probability of a score occurring within a standard normal distribution. It also enables us to compare two scores that are from different samples. A table for the values of ϕ , indicating the values of the cumulative distribution function of the normal distribution is termed as a z score table.

Formula

The equation is given by $z = (x - \mu) / \sigma$.

μ = mean

σ = standard deviation

x = test value

When we have multiple samples and want to describe the standard deviation of those sample means, we use the following formula:

$$z = (x - \mu) / (\sigma / \sqrt{n})$$

Interpretation

1. If a z-score is equal to -1, then it denotes an element, which is 1 standard deviation less than the mean.
2. If a z score is less than 0, then it denotes an element less than the mean.
3. If a z score is greater than 0, then it denotes an element greater than the mean.
4. If the z score is equal to 0, then it denotes an element equal to the mean.
5. If the z score is equal to 1, it denotes an element, which is 1 standard deviation greater than the mean; a z score equal to 2 signifies 2 standard deviations greater than the mean; etc.

12. What is t-test?

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

The t-score is a ratio between the difference between two groups and the difference within the groups.

- Larger t scores = more difference between groups.
- Smaller t score = more similarity between groups.

A t score of 3 tells you that the groups are three times as different *from* each other as they are within each other. So when you run a t test, bigger t-values equal a greater probability that the results are repeatable.

13. What is percentile?

Mostly we can define percentile as a number where a certain percentage of scores fall below that given number. Percentile and percentage are often confused, but both are different concepts. The percentage is used to express fractions of a whole, while percentiles are the values below which a certain percentage of the data in a data set is found. If you want to know where you stand compared to the rest of the crowd, you need a statistic that reports relative standing, and that statistic is called a percentile.

For example, you are the fourth tallest person in a group of 20.80% of people who are shorter than you. That means you are at the 80th percentile.

If your height is 5.4inch then "5.4 inch" is the 80th percentile height in that group.

Percentile Formula is given as -

$$\text{Percentile} = n/N \times 100$$

Where,

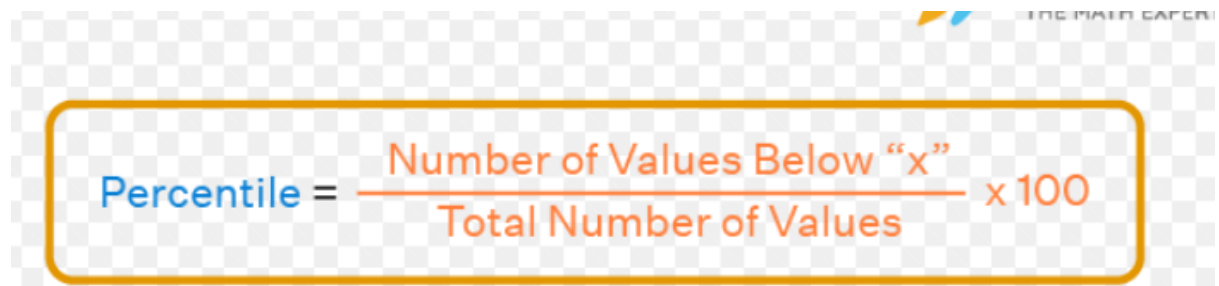
n =ordinal rank of a given value

N = number of values in the data set,

P = Percentile

Percentile is calculated by the ratio of the number of values below 'x' to the total number of values.

The Percentile Formula is given as,


$$\text{Percentile} = \frac{\text{Number of Values Below "x"}}{\text{Total Number of Values}} \times 100$$

14. What is ANOVA?

An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

Another Key part of ANOVA is that it splits the independent variable into 2 or more groups. For example, one or more groups might be expected to influence the dependent variable while the other group is used as a control group, and is not expected to influence the dependent variable.

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

1. An ANOVA can only be conducted if there is no relationship between the subjects in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).
2. The different groups/levels must have equal sample sizes.

3. An ANOVA can only be conducted if the dependent variable is normally distributed, so that the middle scores are most frequent and extreme scores are least frequent.
4. Population variances must be equal (i.e. homoscedastic). Homogeneity of variance means that the deviation of scores (measured by the range or standard deviation for example) is similar between populations.

There are different types of ANOVA tests. The two most common are a “One-Way” and a “Two-Way.”

The difference between these two types depends on the number of independent variables in your test.

One-way ANOVA :

A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.

The one-way ANOVA test for differences in the means of the dependent variable is broken down by the levels of the independent variable.

Two-way ANOVA :

A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

15. How can ANOVA help?

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests . However, it results in fewer type I error and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.