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

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

3. How many outcomes are possible with Bernoulli trial?
   a) 2
   b) 3
   c) 4
   d) None of the mentioned

4. If Ho is true and we reject it is called
   a) Type-I error
   b) Type-II error
   c) Standard error
   d) Sampling 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

6. The chance of rejecting a true hypothesis decreases when sample size is:
   a) Decrease
   b) Increase
   c) Both of them
   d) None

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

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

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?
11. What is z-score?
12. What is t-test?
13. What is percentile?
14. What is ANOVA?
15. How can ANOVA help?

10. Bayes' Theorem:

Bayes’ theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of “Causes”.

For example, Bayes’ theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes’ theorem relies on incorporating prior probability. Distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Formula For Bayes’ Theorem:

$P(A|B)=P(B)P(A\cap B)=P(B)P(A)\cdot P(B|A)$
where:
$P(A)$= The probability of A occurring
$P(B)$= The probability of B occurring
$P(A|B)$=The probability of A given B$P(B|A)$= The probability of B given A
$P(A\cap B)$)= The probability of both A and B occurring

What Does Bayes' Theorem State?

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

What Is Calculated in Bayes' Theorem?

Bayes' Theorem calculates the conditional probability of an event, based on the values of specific related known probabilities.
What Is a Bayes' Theorem Calculator?

A Bayes’ Theorem Calculator figures the probability of an event A conditional on another event B, given the prior probabilities of A and B, and the probability of B conditional on A. It calculates conditional probabilities based on known probabilities.

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. Z-Score, also known as the standard score, indicates how many standard deviations an entity is, from the mean.

Z-Score Formula:

It is a way to compare the results from a test to a “normal” population.

If X is a random variable from a normal distribution with mean (μ) and standard deviation (σ), its Z-score may be calculated by subtracting mean from X and dividing the whole by standard deviation.

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.

Example:

The test score is 190. The test has a mean of 130 and a standard deviation of 30. Find the z score. (Assume it is a normal distribution)
Solution:
Given test score x = 190
Mean, μ = 130
Standard deviation, σ = 30
So $z = (x – \mu)/ \sigma$
$= (190 – 130)/ 30$
$= 60/30$
$= 2$
Hence, the required z score is 2.

How to Interpret z-Score:

- Here is how to interpret z-scores:
- A z-score of less than 0 represents an element less than the mean.
- A z-score greater than 0 represents an element greater than the mean.
- A z-score equal to 0 represents an element equal to the mean.
- A z-score equal to 1 represents 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.
- A z-score equal to -1 represents an element, which is 1 standard deviation less than the mean; a z-score equal to -2 signifies 2 standard deviations less than the mean; etc.
- If the number of elements in the set is large, about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2 and about 99% have a z-score between -3 and 3.

What are the Types of Z Score Table?

There are two z-score tables which are:

- Positive Z Score Table: It means that the observed value is above the mean of total values.
- Negative Z Score Table: It means that the observed value is below the mean of total values.

16. What is t-test?

The t-test is a test that is mainly used to compare the mean of two groups of samples. It is meant for evaluating whether the means of the two sets of data are statistically significantly different from each other.

There are many types of t-test. Some of these are:

- The one-sample t-test, which is used to compare the mean of a population with a theoretical value.
- The unpaired two-sample t-test, which is used to compare the mean of two independent given samples.
- The paired t-test, which is used to compare the means between two groups of samples that are related.

T-test Formula:

The T-test formula is given below:

$t = x1^{-} - x2^{-}/(\sqrt{s21n1+s22n2})$

t t-test value
$x1^{-}$: Mean of first set of values
$x2^{-}$: Mean of second set of values
s1: Standard deviation of first set of values
s2: Standard deviation of second set of values
n1: Total number of values in first set
n2: Total number of values in second set.

The formula for standard deviation is given below:

$\mathbf{s=(\sqrt{\sum(x-x^{-})\^{}2/n-1)}}$

s: The standard deviation for a data set
x: Values given in data set
x^{-}: Mean value of data set
n: Total number of values in the data set

13.What is percentile?

In statistics, a percentile is a term that describes how a score compares to other scores from the same set. While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data scores that fall below a given value.

“Percentile” is in everyday use, but there is no universal definition for it. The most common definition of a percentile is a number where a certain percentage of scores fall below that number. You might know that you scored 67 out of 90 on a test. But that figure has no real meaning unless you know what percentile you fall into. If you know that your score is in the

90th percentile, that means you scored better than 90% of people who took the test.

Percentiles are commonly used to report scores in tests, like the SAT, GRE and LSAT. for example, the 70th percentile on the 2013 GRE was 156. That means if you scored 156 on the exam, your score was better than 70 percent of test takers.

The 25th percentile is also called the first quartile.
The 50th percentile is generally the median.
The 75th percentile is also called the third quartile.
The difference between the third and first quartiles is the interquartile range.

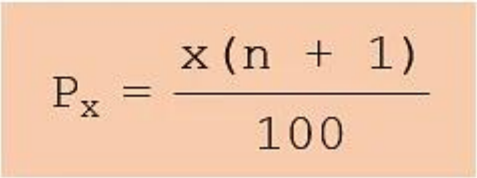
2. Percentile Rank:

The percentile usually indicates that a certain percentage falls below that percentile. For example, if you score in the 25th percentile, then 25% of test takers are below your score. The "25" is called the percentile rank. In statistics, it can get a little more complicated as there are actually three definitions of "percentile." Here are the first two based on an arbitrary "25th percentile":

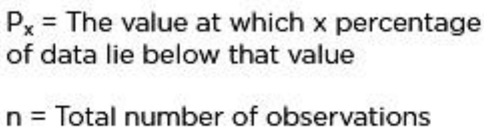
Definition 1: The nth percentile is the lowest score that is greater than a certain percentage ("n") of the scores. In this example, our n is 25, so we're looking for the lowest score that is greater than 25%.

Definition 2: The nth percentile is the smallest score that is greater than or equal to a certain percentage of the scores. To rephrase this, it's the percentage of data that falls at or below a certain observation. This is the definition used in AP statistics. In this example, the 25th percentile is the score that's greater or equal to 25% of the scores.

3. How to Calculate percentile:



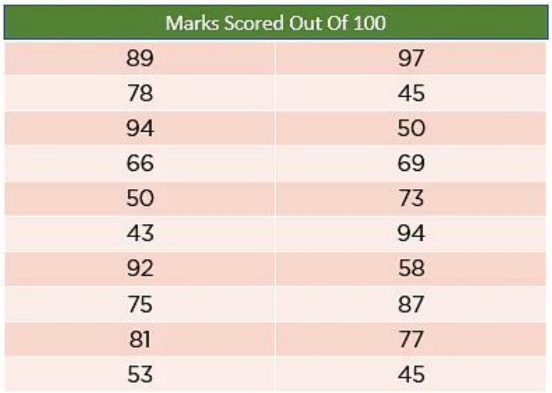
$$P_x = \frac{x(n + 1)}{100}$$



P_x = The value at which x percentage of data lie below that value

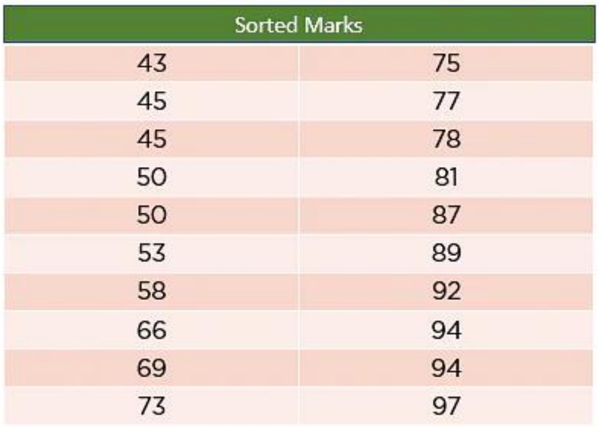
n = Total number of observations

Imagine you have the marks of 20 students. Now, try to calculate the 90th percentile.



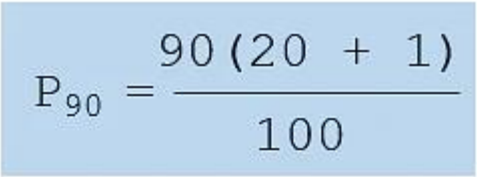
Marks Scored Out Of 100	
89	97
78	45
94	50
66	69
50	73
43	94
92	58
75	87
81	77
53	45

Step 1: Arrange the score in ascending order.

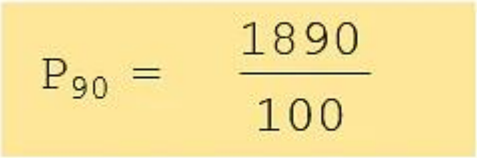


Sorted Marks	
43	75
45	77
45	78
50	81
50	87
53	89
58	92
66	94
69	94
73	97

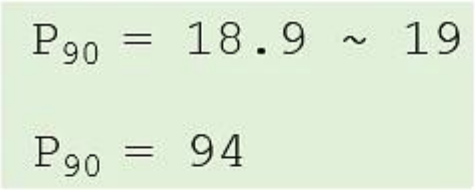
Step 2: Plug the values in the formula to find n.



$$P_{90} = \frac{90(20 + 1)}{100}$$



$$P_{90} = \frac{1890}{100}$$

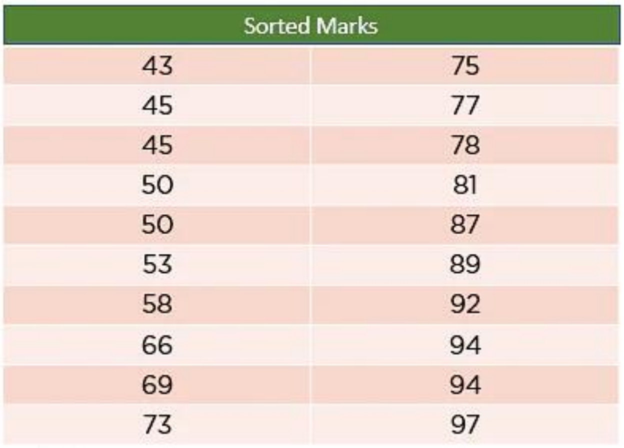


$$P_{90} = 18.9 \sim 19$$

$$P_{90} = 94$$

P90 = 94 means that 90% of students got less than 94 and 10% of students got more than 94
Let’s look at another way how you can find the percentile in statistics.
Suppose you want to find the percentile mark of 78 marks in the data set.

Step 1: Sort the marks in ascending order.



Sorted Marks	
43	75
45	77
45	78
50	81
50	87
53	89
58	92
66	94
69	94
73	97

Step 2: Substitute the value in the formula.

P=n/N*100

n= Ordinal rank of values
N = Total Values in the dataset

P= 12*100/20

P= 60

P = 60 means that 78 marks point to the 60th percentile in the dataset.

14. What is ANOVA?

Data collection, organization, analysis, interpretation, and presentation are all part of statistics, a branch of mathematics. As an interdisciplinary field, statistics have several concepts that have found practical applications. Analysis of Variance, also known as ANOVA, is one such concept that will be discussed in this article.

What Is Analysis of Variance (ANOVA)?

ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples. Analyzing variance tests the hypothesis that the means of two or more populations are equal. If the difference between the two populations is statistically significant, then the two populations are unequal.

ANOVA basics:

An Analysis of Variance (ANOVA) is an inferential statistical tool that we use to find statistically significant differences among the means of two or more populations.

We calculate variance but the goal is still to compare population mean differences. The test statistic for the ANOVA is called F. It is a ratio of two estimates of the population variance based on the sample data.

An ANOVA conducted on a design in which there is only one factor is called a one-way ANOVA. If an experiment has two factors, then the ANOVA is called a two-way ANOVA.

Important Terms Related to ANOVA

Means (Grand and Sample):

A sample mean is the average value for a group, whereas the grand mean is the average of sample means from various groups or the mean of all observations combined.

F-Statistics:

F-statistic or F-ratio is a statistical measure that tells us about the extent of difference between the means of different samples. Lower the F-ratio, closer are the sample means.

Sum of Squares:

The sum of squares is a technique used in regression analysis to determine the dispersion of data points. It is used in the ANOVA test to compute the value of F.

Mean Squared Error (MSE):

The Mean Squared Error gives us the average error in the data set.

Hypothesis:

In ANOVA, we have Null Hypothesis and an Alternative Hypothesis. The Null hypothesis is valid when all the sample means are equal, or they don’t have any major difference. The Alternate Hypothesis is valid when at least one of the sample means is different from the other.

Group Variability:

In ANOVA, a group is a set of samples within the independent variable.

- Between-group variability occurs when there is a significant variation in the sample distributions of individual groups.
- Within-group variability occurs when there are variations in the sample distribution within a single group.

15. How can ANOVA help?

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income). When you understand how each independent variable’s mean is different from the others, you can begin to understand which of them has a connection to your dependent variable (landing page clicks), and begin to learn what is driving that behavior.

What is the difference between one-way and two-way ANOVA tests?

This is defined by how many independent variables are included in the ANOVA test. One-way means the analysis of variance has one independent variable. Two-way means the test has two independent variables. An example of this may be the independent variable being a brand of drink (one-way), or independent variables of brand of drink and how many calories it has or whether it’s original or diet.

How does ANOVA work?

ANOVA compares the means of different groups and shows you if there are any statistical differences between the means. ANOVA is classified as an omnibus test statistic. This means that it can’t tell you which specific groups were statistically significantly different from each other, only that at least two of the groups were.

It’s important to remember that the main ANOVA research question is whether the sample means are from different populations. There are two assumptions upon which ANOVA rests:

First: Whatever the technique of data collection, the observations within each sampled population are normally distributed.

Second: The sampled population has a common variance of s2.

How to conduct an ANOVA test:

Stats iQ and ANOVA:

Stats iQ from Qualtrics can help you run an ANOVA test. When users select one categorical variable with three or more groups and one continuous or discrete variable, Stats iQ runs a one-way ANOVA (Welch’s F test) and a series of pairwise “post hoc” tests (Games-Howell tests). The one-way ANOVA tests for an overall relationship between the two variables, and the pairwise tests test each possible pair of groups to see if one group tends to have higher values than the other.

What does an ANOVA test reveal?

A one way ANOVA will allow you to distinguish that at least two groups were different from each other. Once you begin to understand the difference between the independent variables you will then be able to see how each behaves with your dependent variable.

What are the limitations of ANOVA?

Whilst ANOVA will help you to analyse the difference in means between two independent variables, it won’t tell you which statistical groups were different from each other. If your test returns a significant f-statistic (this is the value you get when you run an ANOVA test), you may need to run an ad hoc test (like the Least Significant Difference test) to tell you exactly which groups had a difference in means.