WORKSHEET

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

answer: 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

answer: a) random

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

answer: **a) 2**

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

answer: **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

answer: c) Level of confidence

6. The chance of rejecting a true hypothesis decreases when sample size is:

a) Decrease

b) Increase

c) Both of them

d) None

answer: b) Increase

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

answer: 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

answer: a) Minimize errors

WORKSHEET

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

answer: a) 0

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

10. What Is Bayes' Theorem?

Answer: The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics. In other words, it's used to figure out how likely an event is based on its proximity to another. Bayes law or Bayes rule are other names for the theorem.

e results, the sample space will be: {1, 2, 3, 4, 5, 6}

Event

An event is the outcome of a random experiment. Getting heads when you toss a coin is an event. Getting a 4 when you roll a fair die is an event.

Random Variable

A random variable is a variable with an unknown value or a function that assigns values to each of the outcomes of an experiment. A random variable can be discrete (meaning it has specific values) or continuous (meaning it has no specific values).

Exhaustive Events

Two or more events associated with a random experiment are exhaustive if their union is the sample space.

Let's say A is the event of a red card being drawn from a pack, and B is the event of a black card being drawn. Because the sample space S = {red, black}, A and B are exhaustive.

Independent Events

When the occurrence of one event has no bearing on the occurrence of the other, the two events are said to be independent. Two events A and B, are said to be independent in mathematics if:

$P(A \cap B) = P(AB) = P(A)*P(B)$

For example, if A gets a 3 on a die roll and B gets a jack of hearts from a well-shuffled deck of cards, then A and B are independent events.

Conditional Probability

Let A and B be the two events associated with a random experiment. Then, the probability of A's occurrence under the condition that B has already occurred and P(B) ≠ 0 is called the Conditional Probability. It is denoted by P (A/B). Thus, you have:

conditional-probability-1

What Is Bayes Theorem?

The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics. In other words, it's used to figure out how likely an event is based on its proximity to another. Bayes law or Bayes rule are other names for the theorem.

The Bayes theorem is the foundation of Naive Bayes, one of the most widely used classification algorithms in data science.

Bayes Theorem Formula

The formula for the Bayes theorem can be written in a variety of ways. The following is the most common version:

$P(A \mid B) = P(B \mid A)P(A) / P(B)$

$P(A \mid B)$ is the conditional probability of event A occurring, given that B is true.

$P(B \mid A)$ is the conditional probability of event B occurring, given that A is true.

P(A) and P(B) are the probabilities of A and B occurring independently of one another.

Example of Bayes Theorem

Now, try to solve a problem using the Bayes theorem.

Problem 1: Three urns contain 6 red, 4 black; 4 red, 6 black, and 5 red, 5 black balls respectively. One of the urns is selected at random and a ball is drawn from it. If the ball drawn is red, find the probability that it is drawn from the first urn.

Solution: Let E1, E2, E3, and A be the events defined as follows:

E1 = urn first is chosen

E2 = urn second is chosen

E3 = urn third is chosen

A = ball drawn is red

Since there are three urns and one of the three urns is chosen at random, therefore:

$P(E1) = P(E2) = P(E3) = ⅓$

If E1 has already occurred, then urn first has been chosen, containing 6 red and 4 black balls. The probability of drawing a red ball from it is 6/10.

So, P(A/E1) = 6/10

Similarly, you have P(A/E2) = 4/10 and P(A/E3) = 5/10

You are required to find the P(E1/A) i.e., given that the ball drawn is red, what is the probability that it is drawn from the first urn.

By Bayes theorem, you have

P(E1/A) = P(E1) P(A/E1)P(E1) P(A/E1) + P(E2) P(A/E2) + P(E3) P(A/E3)

= 1/3 * 6/10(1/3 * 6/10) + (1/3 * 4/10) + (1/3 * 5/10)

= ⅖

11. What is z-score?

Answer: A z score is simply defined as the number of standard deviation from the mean. The z-score can be calculated by subtracting mean by test value and dividing it by standard value. Where x is the test value, µ is the mean and σ is the standard value.

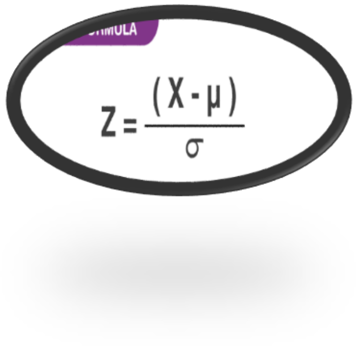
standard normal table (also called the unit normal table or z-score table) is a mathematical table for the values of ϕ, indicating the values of the cumulative distribution function of the normal distribution. Z-Score, also known as the standard score, indicates how many standard deviations an entity is, from the mean.

Since probability tables cannot be printed for every normal distribution, as there is an infinite variety of normal distribution, it is common practice to convert a normal to a standard normal and then use the z-score table to find probabilities.

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.



Where, x = test value

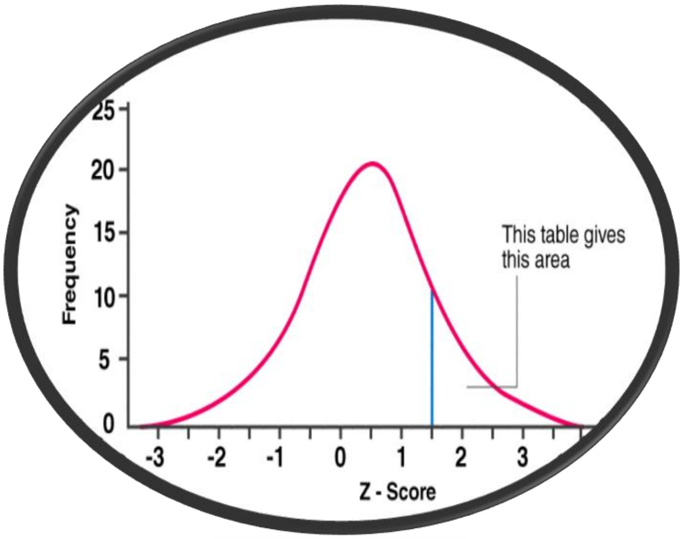
µ is mean and

σ is SD (Standard Deviation)

For the average of a sample from a population 'n', the mean is µ and the standard deviation is σ.

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.



Let us understand the concept with the help of a solved example:

Example: The test scores of students in a class test has a mean of 70 and with a standard deviation of 12. What is the probable percentage of students scored more than 85?

Solution: The z score for the given data is,

z= (85-70)/12=1.25

From the z score table, the fraction of the data within this score is 0.8944.

This means 89.44 % of the students are within the test scores of 85 and hence the percentage of students who are above the test scores of 85 = (100-89.44)% = 10.56

12. What is t-test?

Answer: 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.

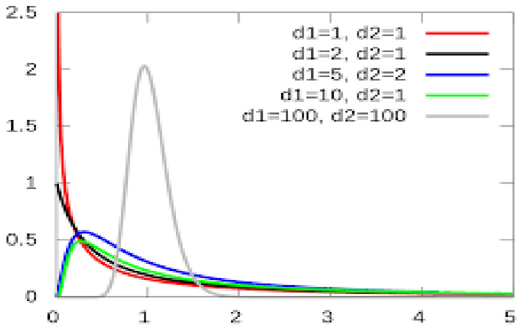
- The one-sample t-test, which is used to compare the mean of a population with a theoretical value.
- The unpaired two-sample t-test, which is used to compare the mean of two independent given samples.

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



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	

Also,

The formula for standard deviation is given below:

$s=(\surd\sum(x-x^{-})2n-1)$

Where,

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

Solved Examples

Q.1: Find the t-test value for the following given two sets of values:

7, 2, 9, 8 and

1, 2, 3, 4?

Solution: For first data set:

Number of terms in first set i.e. n_1 = 4

Calculate mean value for first data set using formula:

x¯1=∑x1n1

i.e. x¯1=7+2+9+84

i.e. x¯1=6.5

Construct the following table for standard deviation:

x1	x1−x¯1	(x1−x¯1)2
7	0.5	0.25
2	-4.5	20.25
9	2.5	6.25
8	1.5	2.25

Thus , ∑((x1−x¯1)2)=29

Now, compute the standard deviation usng formula as,

s1=(√∑(x1−x¯1)2n1−1)

i.e. s1=(√294−1)

i.e. s1=(√9.66)

s1=3.11

Therefore, standard deviation for the first set of data: s_1 = 3.11

For second data set:

Number of terms in second set i.e.

. n2=4

Calculate mean value for second data set using formula:

x¯2=∑x2n2
i.e. x¯2=1+2+3+44
i.e. x¯2=2.5

Construct the following table for standard deviation:

x2	x2−x¯2	(x2−x¯2)2
1	-1.5	2.25
2	-0.5	0.25
3	0.5	0.25
4	1.5	2.25

Thus, ∑((x2−x¯2)2)=5

Now, compute the standard deviation using formula as,

s2=(√∑(x2−x¯2)2n2−1)
i.e. s2=(√54−1)
i.e. s1=(√1.66)
s1=1.29
Therefore, standard deviation for the second set of data: s2=1.29

Now, apply the formula for t-test value:

t=x1¯−x2¯(√s21n1+s22n2)
t=6.5–2.5(√3.1124+1.2924)
=4(√9.36674+1.6674)

t = 2.38

Hence t-test value for the two data sets is = 2.38

13. What is percentile?

Answer: Percentile (also referred to as Centile) is the percentage of scores that range between 0 and 100 which is less than or equal to the given set of distribution. Percentiles divide any distribution into 100 equal parts.

“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 (if you’re using the third definition—see below).

The 75th percentile is also called the third quartile.

The difference between the third and first quartiles is the interquartile range.

There are actually three definitions of “percentile.” Here are the first two (see below for definition 3), 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.

They may seem very similar, but they can lead to big differences in results, although they are both the 25th percentile rank.

Example question: Find out where the 25th percentile is in the above list.

Score	Rank
30	1
33	2
43	3
53	4
56	5
67	6
68	7
72	8

* to Find a Percentile

Step 1: Calculate what rank is at the 25th percentile. Use the following formula:

Rank = Percentile / 100 * (number of items + 1)

Rank = 25 / 100 * (8 + 1) = 0.25 * 9 = 2.25.

A rank of 2.25 is at the 25th percentile. However, there isn't a rank of 2.25 (ever heard of a high school rank of 2.25? I haven't!), so you must either round up, or round down. As 2.25 is closer to 2 than 3, I'm going to round down to a rank of 2.

Step 2: Choose either definition 1 or 2:

Definition 1: The lowest score that is greater than 25% of the scores. That equals a score of 43 on this list (a rank of 3).

Definition 2: The smallest score that is greater than or equal to 25% of the scores. That equals a score of 33 on this list (a rank of 2).

Depending on which definition you use, the 25th percentile could be reported at 33 or 43! A third definition attempts to correct this possible misinterpretation:

Definition 3: A weighted mean of the percentiles from the first two definitions.

In the above example, here's how the percentile would be worked out using the weighted mean:

Multiply the difference between the scores by 0.25 (the fraction of the rank we calculated above). The scores were 43 and 33, giving us a difference of 10:

(0.25)(43 – 33) = 2.5

Add the result to the lower score. 2.5 + 33 = 35.5

In this case, the 25th percentile score is 35.5, which makes more sense as it's in the middle of 43 and 33.

Percentile Range

A percentile range is the difference between two specified percentiles. these could theoretically be any two percentiles, but the 10-90 percentile range is the most common. To find the 10-90 percentile range:

Calculate the 10th percentile using the above steps.

Calculate the 90th percentile using the above steps.

Subtract Step 1 (the 10th percentile) from Step 2 (the 90th percentile).

14. What is ANOVA?

Answer: Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher.

A factorial ANOVA is an Analysis of Variance test with more than one independent variable, or “factor“. It can also refer to more than one Level of Independent Variable. For example, an experiment with a treatment group and a control group has one factor (the treatment) but two levels (the treatment and the control).

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Basically, you’re testing groups to see if there’s a difference between them. Examples of when you might want to test different groups:

A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.

A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.

Students from different colleges take the same exam. You want to see if one college outperforms the other.

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

One-way ANOVA between groups: used when you want to test two groups to see if there’s a difference between them.

Two way ANOVA without replication: used when you have one group and you’re double-testing that same group. For example, you’re testing one set of individuals before and after they take a medication to see if it works or not.

Two way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

15. How can ANOVA help?

Answer: Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

One-Way Analysis of Variance (ANOVA) tells you if there are any statistical differences between the means of three or more independent groups.

You might use Analysis of Variance (ANOVA) as a marketer, when you want to test a particular hypothesis. You would use ANOVA to help you understand how your different groups respond, with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).

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.

You may want to use ANOVA to help you answer questions like this:

Do age, sex, or income have an effect on whether someone clicks on a landing page?

Do location, employment status, or education have an effect on NPS score?

One-way ANOVA can help you know whether or not there are significant differences between the groups of your independent variables (such as USA vs Canada vs Mexico when testing a Location variable). You may want to test multiple independent variables (such as Location, employment status or education). When you understand how the groups within the independent variable differ (such as USA vs Canada vs Mexico, not location, employment status, or education), you can begin to understand which of them has a connection to your dependent variable (NPS score).

"Do all your locations have the same average NPS score?"

Although, you should note that ANOVA will only tell you that the average NPS scores across all locations are the same or are not the same, it does not tell you which location has a significantly higher or lower average NPS score.

How does ANOVA work?

Like other types of statistical tests, 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.

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.

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.