PALAWAN STATE UNIVERSITY
College of Sciences

MATHEMATICS IN THE MODERN WORLD

MATHEMATICS AS A TOOL

MODULE 3

DATA MANAGEMENT

Table of Contents

Learning Objectives

After going through in this module, you should be able to:

- ✓ **Use varied and appropriate statistical tools to process and manage numerical data.**
- ✓ **Apply the methods of linear regression and correlation to predict the value of a variable given certain conditions.**
- ✓ **Use statistical data in drawing conclusions and making important decisions.**

Overview

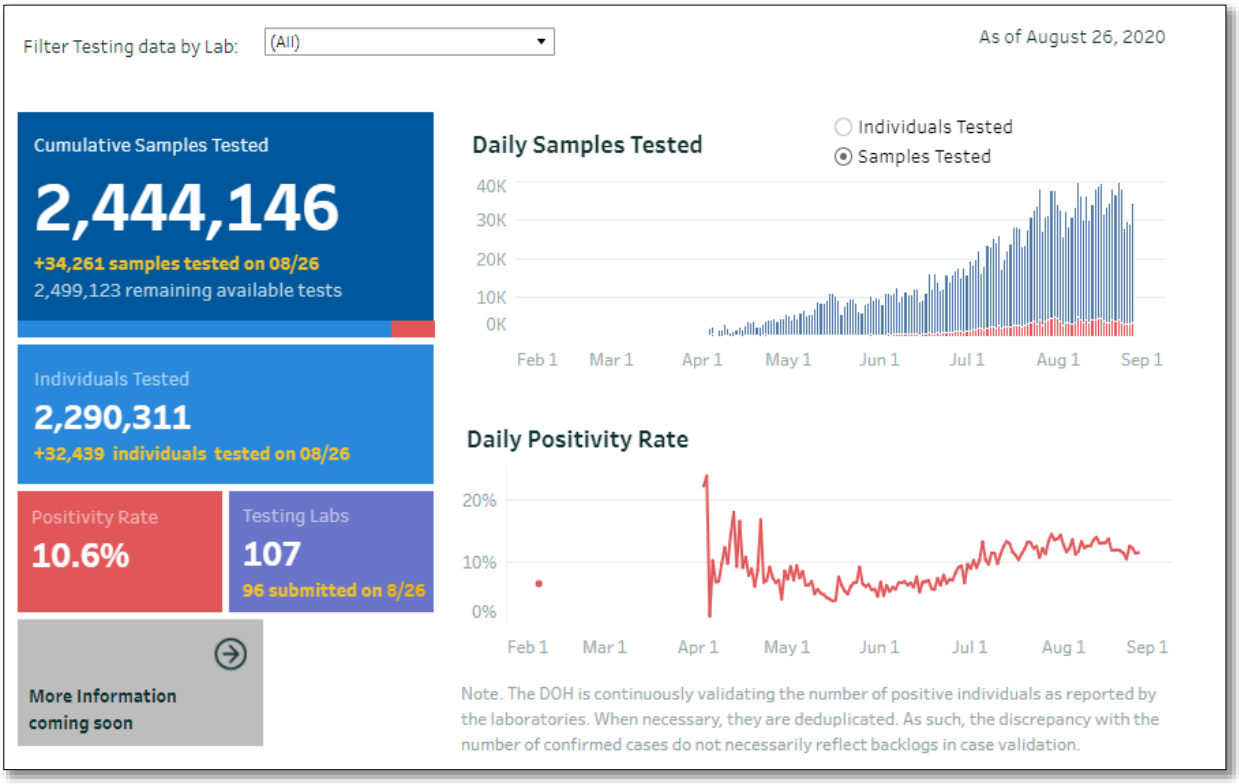
Hello students! How are you keeping up with our modular classes? I hope that you are all doing fine. If not, please feel free to ask for our help In this module, we are going to review some concepts that you have learned from the Statistics courses you took in your basic education. Specifically, this will cover the following topics: Descriptive Statistics, Inferential Statistics and Planning or Conducting an Experiment or Study. We will briefly discuss the measures of central tendency, measures of variation, normal distributions, and linear regression and correlation.

This module is designed for you to finish in three weeks. Time management and self-motivation is the key to accomplishing all designated tasks/activities of this module, to achieve every objectives and outcomes of this course.

I hope you will be able to appreciate the importance of managing data as it is relevant to the pandemic that we are struggling with right now. It can help us fight through this battle and all other disasters which could strike us in the future. Stay safe everyone!

Initial Activity

I want you to access the DOH COVID-19 Tracker (doh.gov.ph, 2020) through the link www.doh.gov.ph/covid19tracker and browse the webpage. If you can't access this page, refer to a screenshot below of a portion the webpage. Of all the data being shown in this platform, can you identify which data show average values? Can you interpret these? Do the graphs make sense to you? Do they display a trend which could help us predict the behavior of the data in the next 7-14 days or more?



Discussion

A. Review of Descriptive Statistics

Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users (Galleto, 2016). Data are individual pieces of factual information recorded and used for the purpose of analysis. It is the raw information from which statistics are created (Macalester.edu, n.d.). Statistics are the results of data analysis - its interpretation and presentation. Statistics has two branches. The one that involves collection, organization, summarization and presentation is called **descriptive statistics**. The branch of statistics which involves interpretation and drawing conclusions is called **inferential statistics**.

Typically, there are two general types of statistic that are used to describe data: **measures of central tendency** and **measures of variance**. A measure of central tendency, which is sometimes referred to as "averages", describes a set of data by identifying its central position. It provides us the value that is typical or representative of the whole data set. In this module, will consider three types of averages: mean, median and mode. In the first section of discussion, we will take a look at these measures specifically on how to calculate them.

Measures of Central Tendency

Mean

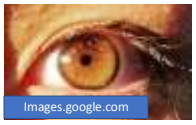
The mean, also called the arithmetic mean, is the most frequently used measure of central tendency. It is obtained by getting the sum of all values and divide it by the number of values in the data set. So if we have n values in the data set and they have values $x_1, x_2, x_3, \dots, x_n$, the mean usually denoted by $\bar{x}$ (read as x bar) is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

$$\bar{x} = \frac{\sum x}{n}$$

The summation notation "Σ" denotes the sum of all values in a given data set.

x is the variable usually used to represent the individual data values
n represents the number of values in a sample



Did you know?

Seaweed can grow up to 12 inches per day averagely.

An average person eats almost 1500 pounds of food a year.

The human eye blinks an average of 4,200,000 times a year.

The most children born to one woman is 69, she was a peasant who lived a 40 year life, in which she had 16 twins, 7 triplets, and 4 quadruplets.

*Source: statisticbrain.com

To be precise, $\bar{x}$ is known as *sample mean*. In some situations, data are collected from small portions of a large group in order to represent the determined information about

Discussion

the group. The whole group under consideration is called a **population**, while any subset of the population is called a **sample**.

So if we want to compute for the value of the *population* mean, denoted by μ, we use the formula:

$$\mu = \frac{\sum x}{N}$$

where N is the size of the population.

Example

1.) In a grocery store, price (₱) of bath soap products of different brands are the following:

35 46 40 31 39

Find the mean of the prices of these five brands of bath soap.

Solution

The five brands are all of the bath soap brands in the grocery store (population size N = 5). We use μ to represent the mean.

$$\mu = \frac{\sum x}{N} = \frac{35 + 46 + 40 + 31 + 39}{5}$$

Thus, the mean of the prices of these five brands of bath soap is ₱*38.20*.

2.) The following are the ages (in years) of eight of the 67 employees of a small company:

25 32 61 42 39 48 56 29

Find the mean age of the employees.

Solution

Because the given data set only includes eight of the 67 employees of the company, it represents the population. Hence, *N* = 8. The population mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{25 + 32 + 61 + 42 + 39 + 48 + 56 + 29 = 332}{8} = \frac{332}{8} = 41.5 \approx 42$$

Thus, the mean age of all eight employees of this company is *42 years*.

Discussion

Sometimes a data set may contain a few very small or a few very large values; such values are called **outliers** or extreme values. Outliers may contain valuable information or be meaningless aberrations caused by measurement and recording errors. A major shortcoming of the mean as a measure of central tendency is that it is very sensitive to outliers.

Example

3.) The following are the savings (₱) of five siblings of a certain family (arranged from the youngest to the eldest sibling):

2500 3000 2800 2900 35900

Notice that the eldest sibling's savings (35900) is very large compared to the others. Hence, this is an outlier. Show how this affects the value of the mean.

Solution

If we do not include the data of the eldest sibling, then the mean is

$$Mean = \frac{2500 + 3000 + 2800 + 2900}{4} = ₱2800$$

Now, if we include it to our data, then the mean is

$$Mean = \frac{2500 + 3000 + 2800 + 2900 + 35900}{5} = ₱9420$$

Thus, including the savings of the eldest sibling causes more than three times increase in the value of the mean, which changes from ₱2800 to ₱9420.

Note:

In the last example, where we included the outlier in the computation of the mean, it seemed that the measure that we got does not provide us the value that is typical or representative of the whole data set. The first computation seemed to be the more sensible value. However, it is NOT acceptable to drop an observation *just* because it is an outlier. Outliers can be legitimate observations and are sometimes the most interesting ones. It is important to investigate the nature of the outlier before deciding whether or not to drop an outlier in the analysis of data. Please check out the following link https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/ for guidelines in dealing with outliers.

Learning Check

Activity 1

Compute the mean of the following set of data values

a.) Anika went to the supermarket to buy some packed potatoes, with her uncertainty of estimating the difference between the sizes of the packed potatoes, she looks at the price tags and finds the following prices (₱)

125 142 132.5 201.25 160 172.75

b.) Archie collects stamps and glues them to his notebook. The following shows the tally of his monthly collection from June to November

16 25 18 30 29 23

c.) The following shows the grades of a student in Physics, Math, English, Biology, Chemistry and Filipino, respectively.

86 89 91 93 88 90

Discussion

Median

Another important measure of central tendency is the median. The median of the set of scores is the **middle value when the scores are arranged in order** of increasing (or decreasing) magnitude. Aside from being a measure of central tendency, it is also a positional score as it divides the data set into two equal parts: one- half of the observations above it, and the other half below or equal to it. To find the median of a data set, we first rank the data values (arrange them in increasing or decreasing order), then get the median in one of the following ways.

- ✓ If the number of scores is odd, the median is the number that is exactly in the middle of the list.
- ✓ If the number of scores is even, the median is found by computing the mean of the two middle numbers. That is, we add the two middlemost values and divide the sum by 2.

Discussion

Example

1.) The following data give the age of students in a ballet class. Find the median age.

12 15 17 11 13 16 19

Solution

First, we rank the given data in an increasing order. Since there are seven (odd) values in this data set as it was arranged accordingly, the fourth term is the middle term. Hence,

11 12 13 **15** 16 17 19

Median

The median age of the students in a ballet class is 15.

2.) The following are the number of smart phone users in 10 households in a certain barangay:

7 9 4 10 5 6 5 8 12 5

Solution

Since there are ten values in this data set as it was arranged accordingly, the fifth term and the sixth term are the middle term:

4 5 5 5 **6 7** 8 9 10 12

Median

To get the median, we get the average of the two middle terms

$$Median = \frac{6+7}{2} = 6.5$$

The median number of smart phone users among the 10 households in the barangay is 6.5.

Learning Check

Activity 2

Compute the median of each of the following sets of data

a.) 3, 4, 7, 11, 12, 12, 15, 16

b.) -8, -5, -12, -1, 4, 7, 11

c.) 6, 4, 8.5, 9, 11, 8.25, 6.5, 8.75

d.) The following are the ages of the 5 siblings of the Reyes family. Find the age of the middle child.

25 21 23 18 27

e.) In a supermarket, the following table shows the number of tissue roll of different brands sold in a certain week (Brand 1 to Brand 8, respectively)

25 12 5 17 20 13 9 21

Discussion

Mode

Another measure of central tendency is the **mode**, which is the value that occurs most frequently in a data set. There is no formula in finding the mode of an ungrouped data, it is just found by inspection. When two scores occur with the same greatest frequency, each one is a mode and the data set is bimodal. When more than two scores occur with the same greatest frequency, each is a mode and the data set is said to be multimodal. It is also possible that a data set has no mode; that is when no score is repeated more than the others. In this case, we stipulate that there is no mode.

Example

1.) The following gives the general weighted average of the top 10 students with the highest grades in a certain class.

95 93.5 91 89 92 94 91 93 90 91.5

Solve for the mode.

Solution

In this data set, all values appeared only once except for 91 which appeared twice. Because 91 has the highest frequency, hence

$$\boldsymbol{Mode} = 91$$

2.) The following are the number of ball pens that each of the eight students owned in a certain group of friends: 1 3 2 0 2 1 3 5

Solve for the mode.

Solution

There are three data values with the highest frequency in this data set which are 1, 2, and 3. Therefore, the data set is multimodal and the modes are:

$$\boldsymbol{Mode} = 1, 2, 3$$

Discussion

3.) A statistician conducted a survey in a certain barangay to obtain the profile of households wherein the statistician gathered information about the appliances owned by each household. The table below shows the summary of the number of appliances of by the households

Name of Appliance	***No. of Households***
Television	35
Refrigerator	26
Microwave oven	10
Kitchen Stove	36
Washing Machine	15
Clothing Iron	30

What is the most popular appliance owned in that barangay?

Solution

Since the kitchen stove is the appliance that has the highest frequency, then the most popular appliance owned in that barangay is kitchen stove.

Example

4.) The following are the number of tourist arrivals in Honda bay for a sample of 1 week, beginning from Monday to Sunday:

87, 45, 63, 49, 75, 80, 100.

What is the busiest day of the week in Honda Bay?

Solution

Since the number of tourists is the highest on Sunday, then the busiest day of the week in Honda Bay is on Sunday.

5.) The following are the scores of 10 students in Math.

82, 78, 95, 83, 89, 75, 90, 80, 88, 92

What is the modal grade?

Solution

Since there is no grade that occurred more than the others, then this data set has **no mode**.

Discussion

Note:

- ✓ Among the different measures of central tendency, the mode is the only one that can be used with the data at the nominal level of measurement. The median can only be used if the variable is at least at the ordinal level of measurement. While the median can only be used if the variable is at least at the ordinal level of measurement. For a review of the levels of measurement, check out this link https://conjointly.com/kb/levels-of-measurement/.

- ✓ Unfortunately, the term average is sometimes used for any measure of central tendency and is sometimes used for the mean. Because of this ambiguity, we should not use the term average when referring to a specific measure of central tendency. Instead we should use the specific term, such as mean, median, or mode.

Learning Check

Activity 3

Solve for the mode of each of the given data set.

a.) 13, 15, 16, 12, 11, 17, 13, 15, 10, 14

b.) 7.5, 8.5, 6.5, 3.5, 5, 5.5, 5, 9.5, 10.5

c.) -11, -4, -5, |-1|, 0, 1, 5, 4, 11

d.) The following shows the number of hops in a skipping rope that 10 students can complete in one minute

60 35 55 78 56 55 69 48 58 57

e.) The following table shows an estimate of the number of strawberries that a farmer was able to harvest per day in a certain week

Day	***No. of Strawberries Harvested***
Sunday	560
Monday	980
Tuesday	760
Wednesday	670
Thursday	950
Friday	980
Saturday	1000

Discussion

Mean for Grouped Data

If the data are given in the form of a frequency table, we no longer know the values of individual observations. In such cases, we cannot obtain the sum of individual values. We find an approximation for the sum of these values using the formulas shown below

$$\mu = \frac{\sum mf}{n} \qquad\qquad \bar{x} = \frac{\sum mf}{N}$$

where *m* is the midpoint; *f* is the frequency of a class interval; mf is the product of the midpoint and frequency in a class interval, and $\sum mf$ is the summation of the product of the midpoint and frequency in all class intervals.

For a quick review of some important terms about a frequency distribution table, we have the following sidetrip lesson:

The Frequency Distribution (fd). When the set of data contains a large number of elements or observations, grouping a data set using a frequency distribution (fd) table can give us a better picture of the behavior of the data. A frequency distribution table is an arrangement of data into mutually exclusive classes along with the corresponding frequency falling in each class. Below is an example of a simple frequency distribution and some terms described.

Height of Men (in inches) (X)	Number of Men (frequency)
50-54	1
55-59	2
60-64	3
65-69	49
70-74	46
75-79	1

- ✓ Class intervals are the mutually exclusive classes or categories. In our sample frequency the class intervals are 50-54, 55-59, 60-64, 65- 69, 70-74, and 75- 79.
- ✓ The class size or class width is the distance from each lower limit to its corresponding upper limit, and it is uniform in all classes. For example, in the interval 50- 54, the lower limit is 50 and the upper limit is 54, and the distance between them is 5 (same with the distance between 55 and 59, 60 and 64, and so on.)
- ✓ The midpoint of an interval is found by adding the lower limit and upper limit of the class and dividing the sum by 2.

Discussion

Example

The table below gives the frequency distribution of the daily commuting times (in minutes) from home to work for all 40 employees of a company.

Daily Commuting Time	***Number of Employees***
1-9	4
10-18	12
19-27	9
28-36	7
37-45	5
46-54	3

Let *x* denote the daily commuting times (in minutes) from home to work of the 40 employees of a company, and *f* denote the frequency. The values of *m* and *mf* are calculated in the table below

x	f	m	mf
1-9	4	5	20
10-18	12	14	168
19-27	9	23	207
28-36	7	32	224
37-45	5	41	205
46-54	3	50	150
	N = 40		$\sum \boldsymbol{mf = 974}$

$$\mu = \frac{974}{40} = \mathbf{24.35}$$

Hence, the mean daily commuting times of the 40 employees of the company is 24.35 minutes.

Learning Check

Activity 4

The table below shows the frequency distribution of the number of customers received in a cafe each day during the past 24 days.

Number of Orders	Number of Days
10-12	12
13-15	20
16-18	14
19-21	25
22-24	15

Learning Check

Median of Grouped Data

To compute for the median of a grouped set of data, the median class shall be identified by locating the $\frac{n}{2}th$ (or half of the) data at the >*cf* column. Then, we will use the formula below:

$$Median = L + \frac{\frac{n}{2} - cf_B}{f_m} \times w$$

where *L* is the lower class boundary of the median class, n is the number of observations in the data set, cf_B is the cumulative frequency of the class before the median class, f_m is the frequency of the median class, and *w* is the class width.

Note:

- ✓ The median class is found by dividing n by 2 and looking it up through the cf< column. Pick the row that has a value that is equal or nearest greater than the quotient $\frac{n}{2}$.
- ✓ The cumulative frequency less than (cf<) column is found by starting at the frequency of the lowest class interval and adding the frequency of the next class each time.
- ✓ The lower boundary L is found by subtracting one- half of the distance between an upper limit and the succeeding lower limit from the lower limit. Sounds vague? Here's what it means. Suppose the class intervals are 50-54, 55-59, 60-64, 65- 69, 70-74, and 75- 79, the distance between an upper limit and the succeeding lower limit is 1. (See 54 and 55; 59 and 60, and so on.) One- half of 1 is 0.5. So the lower boundary of the class 50-54 is 49.5 (that's 50- 0.5), the lower boundary of the class 55-59 is 54.5, and so on. Now, suppose the class intervals are 1.0- 1.4, 1.5- 1.9, 2.0- 2.4, 2.5- 2.9. The distance between an upper limit and the succeeding lower limit is 0.1, and one- half of it is 0.05. So, the lower boundary of the class 1.0- 1.4 is 0.95 (that's 1.0- 0.05), the lower boundary of the class 1.5- 1.9 is 1.45, and so on. (This second example is seldom used.)

Learning Check

Example

The table below gives the frequency distribution of the number of hours spent in studying by 50 students before a quarter exam.

Hours	Number of Students
0-3	9
4-7	13
8-11	10
12-15	11
16-19	7

Solution

Let *x* denote the number of hours spent in studying by 50 students before a quarter exam, and f denote the frequency. The values of $m - \bar{x}$ and $(m - \bar{x})^2$, and $f(m - \bar{x})^2$ are calculated in the table below.

X	***f***	***cf<***	***CB***
0-3	9	9	-0.5-3.5
4-7	13	22	3.5-7.5
8-11	10	32	7.5-11.5
12-15	11	43	11.5-15.5
16-19	7	50	15.5-19.5

Median class

First, we have to identify the median class. In this example, $\frac{n}{2} = \frac{50}{2} = 25$. **Hence, the 25th data can be found at the third class.** (Since the second class contains only 22 data, and the third class already contains 32 data which includes the $25th$ one.) Remember to pick the class that has a value that is equal or nearest greater than the quotient $\frac{n}{2}$.

With the third class as the median class, $L = 7.5$, $cf_B = 22$, $f_m = 10$, and $w = 4$

$$Median = 7.5 + \frac{25 - 22}{10} \times 4 = \mathbf{8.7}$$

Therefore, the median number of hours spent in studying by 50 students before a quarter exam is 8.7 hours.

Learning Check

Activity 5

Calculate the median score in a Math quiz as shown in the table below:

Score	Number of Students
12-15	4
16-19	6
20-23	4
24-27	5
28-31	7
32-35	4

Discussion

Mode of Grouped Data

To compute for the mode of a grouped set of data, the modal class shall be identified by locating the modal class- the class with the highest frequency at the *f* column. If in case there is more than 1 class with the highest frequency, we will have to compute for more than 1 mode. Then, use the formula below:

$$Mode = L + \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} \times w$$

Where L is the lower boundary of the modal class, f_{mo} is the frequency of the modal class, $f_{mo-1}$is the frequency of the class before the modal class, and f_{mo+1} is the frequency of the class afer the modal class and w is the class width.

The formula above can be written as

$$Mode = L + \frac{d_1}{d_1 + d_2} \times w$$

where
d_1= the difference between the frequency of the modal class and the frequency of the class before the modal class;
d_2= the difference between the frequency of the modal class and the frequency of the class after the modal class

Discussion

Example

Calculate the mode of the number of hours spent in studying by 50 students before a quarter exam.

Hours	Number of Students
0-3	9
4-7	13
8-11	10
12-15	11
16-19	7

Solution

Let *x* denote the number of hours spent in studying by 50 students before a quarter exam, and *f* denote the frequency. The values of *f* and mf are calculated in the table below.

x	***f***	***>cf***	***CB***
0-3	9	9	-0.5-3.5
4-7	13	22	3.5-7.5
8-11	10	32	7.5-11.5
12-15	11	43	11.5-15.5
16-19	7	50	15.5-19.5

← Modal class

In this frequency distribution table, the modal class is the second class because it has the highest frequency which is 13. So, L = 3.5, $f_{mo} = 13$, $f_{mo-1} = 9$, $f_{mo+1} = 10$, $w = 4$

$$Mode = 3.5 + \frac{13 - 9}{(13 - 9) + (13 - 10)} \times 4$$

$$= 3.5 + \frac{4}{4 + 3} \times 4$$

$$= 3.5 + \frac{4}{7} \times 4$$

$$= 5.79$$

Hence, the mode of the number of hours spent in studying by 50 students before a quarter exam is 5.79 hours.

Learning Check

Activity 6

Calculate the modal score in a Math quiz as shown in the table below:

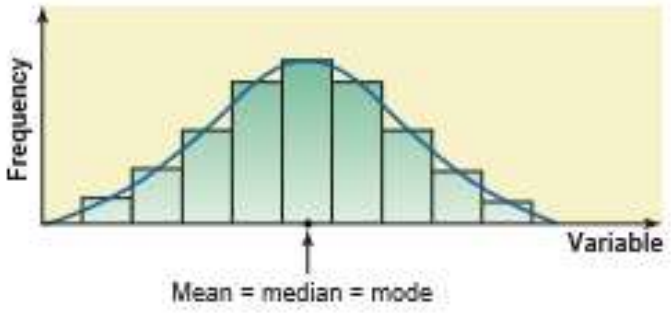
Score	Number of Students
12-15	4
16-19	6
20-23	4
24-27	5
28-31	7
32-35	4

Discussion

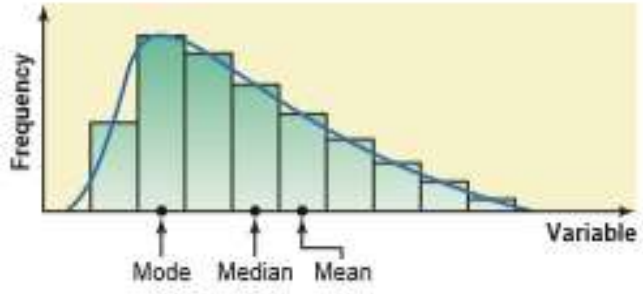
Relationships among the Mean, Median, and Mode

A histogram or a frequency distribution curve can assume shapes which are symmetric and skewed. The shape of a frequency distribution curve can be identified or described using the knowledge of the values of the mean, median, and mode of a certain set of data.

1. If the values of the mean, median, and mode are identical, and they lie at the center of the distribution, then, a symmetric histogram and frequency distribution curve has one peak. This graph is bell- shaped and is called a normal curve.



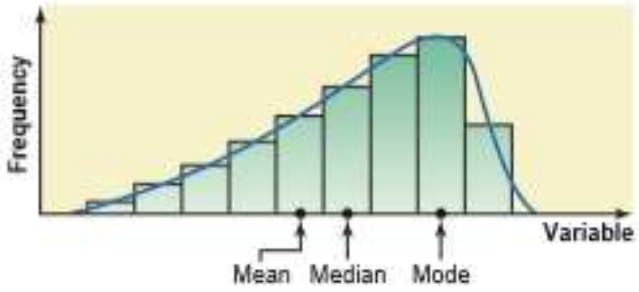
2. If the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two, then a histogram and a frequency distribution curve is skewed to the right (Notice that the mode always occurs at the peak point.) This graph is called positively skewed.



The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail. These outliers pull the mean to the right. (Note: The horizontal line in the graph is the x- axis, so values further to the right are greater than those on the left. Hence, we say that the mean here is the largest among the three averages.)

Discussion

3.) If the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two, then, histogram and a frequency distribution curve are skewed to the left.



In this case, the outliers in the left tail pull the mean to the left. This graph is called negatively skewed.

Following is a matrix comparing the three averages.

COMPARISON OF MEAN, MEDIAN, AND MODE

Average	Definition	How Common	Existence	Takes every score into account?	Affected by Extreme Scores?	Advantages and Disadvantages
Mean	$\bar{X} = \frac{\sum X}{n}$	Most familiar"average"	Always exists	Yes	Yes	Works well with many statistical methods
Median	middle score	Commonly used	Always exists	No	No	Often a good choice if there are some extreme scores
Mode	most frequent score	Sometimes used	Might not exist; may be more than one mode	No	No	Appropriate for data at the nominal level

Discussion

Here are some more takeaways in the comparison of the three averages.

- ✓ For a data collection that is approximately symmetric with one mode, the mean, median, and mode tend to be about the same.
- ✓ For a data collection that is obviously symmetric, it would need to report both the mean and median.
- ✓ The mean is relatively reliable. That is, when samples are drawn from the same population, the sample means tend to be more consistent than the other averages (consistent in a sense that the means of samples drawn from the same population don't vary as much as the other averages).
- ✓ A comparison of the mean, median and mode can reveal information about the characteristic of **skewness**. A distribution of data is skewed if it is not symmetric and **extends more to one side than the other**.
- ✓ With a **symmetric** distribution, if the data is graphed using a histogram, we will see that **the left half of the histogram is roughly a mirror image of its right half**. The graph is roughly **bell-shaped**.
- ✓ Data skewed to the left are said to be **negatively skewed**; the mean and median are to the left of the mode. Although not always predictable, negatively skewed data generally have the mean to the left of the median.
- ✓ Data skewed to the right are said to be **positively skewed**; the mean and median are to the right of the mode. Again, although not always predictable, negatively skewed data generally have the mean to the right.

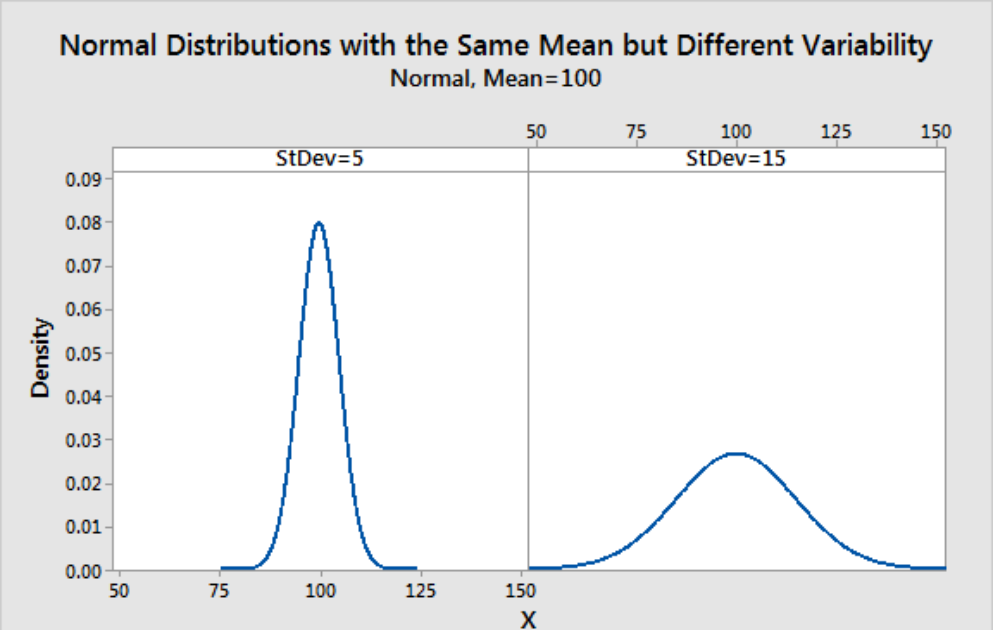
Discussion

Measures of Variation

In examining averages, some characteristics of a set of data may not be evident. For instance, the example from the discussion of mean about the savings of siblings in a family, the given shows that the mean of the savings are equal despite of having an inconsistent value which is quite far from the others. This example shows that the average does not reflect the dispersion, spread or scatter of data.

In this section, we will be introducing the measures of variability. In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. A measure of **variability** is a summary statistic that represents the amount of dispersion in a dataset. It tells us how spread out are the values. While a measure of central tendency describes the typical value in a data set, measures of variability define how far away the data points tend to fall from the center. We talk about variability in the context of a distribution of values. A low dispersion indicates that the data points tend to be clustered tightly around the center. High dispersion signifies that they tend to fall further away.

The two plots below show the difference graphically for distributions with the same mean but more and less dispersion. The panel on the left shows a distribution that is tightly clustered around the average, while the distribution in the right panel is more spread out.



https://i1.wp.com/statisticsbyjim.com/wp

Discussion

When a distribution has lower variability, the values in a dataset are more consistent. However, when the variability is higher, the data points are more dissimilar and extreme values become more likely. Understanding variability helps us grasp the likelihood of unusual events as it provides critical information.

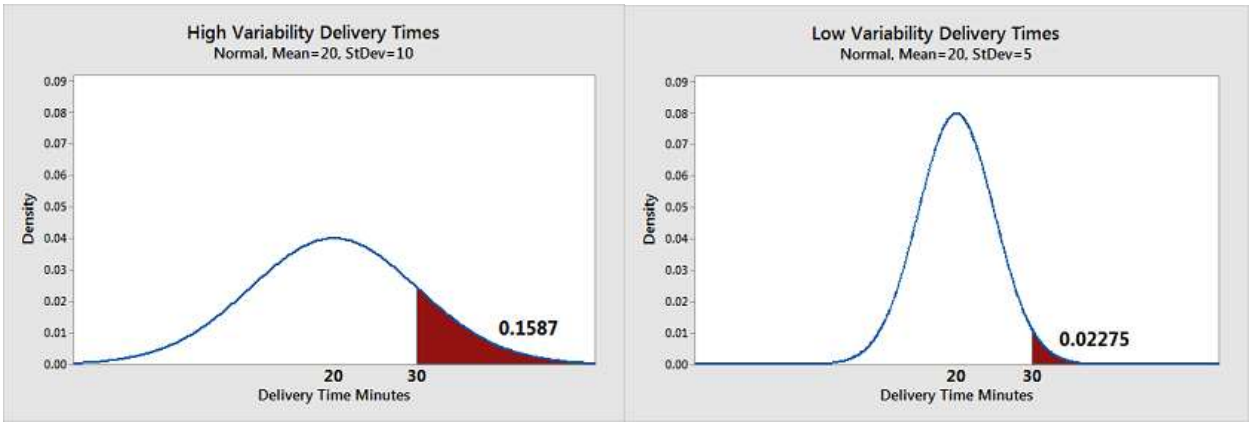
Variability is everywhere. Your travel time from the city to your municipality varies a bit every time. When you order a favorite snack at a restaurant repeatedly, it isn't exactly the same each time. The effectiveness of a medicine may vary from one patient to another.

These are all examples of real-life variability. While some degree of variation is unavoidable, too much inconsistency can cause problems. If your travel time takes much longer than the mean travel time, it would be difficult to schedule a fetch at the terminal. If the restaurant snack is much different than how it usually is, you might not like it at all. And, if a medicine's effect is too inconsistent, it would be dangerous to take.

Example: Suppose we want to order food online and we are considering two restaurants as options. They have the same menu and both restaurants sound equally good! Also, they both advertise a mean delivery time of 20 minutes. Which restaurant do we choose?

Since the mean does not differentiate the two restaurants, we need to we need to analyze the restaurant's variability. Suppose we studied their delivery times, calculated the variability for each place, and learned that their variabilities are different. Suppose again that the graphs below display the distribution of delivery times for these two restaurants. The restaurant with more variable delivery times (first chart) has the broader distribution curve.

Discussion



https://i0.wp.com/statisticsbyjim.com/wp

In these graphs, we consider a 30-minute wait or longer to be unacceptable. The shaded area in each chart represents the proportion of delivery times that surpass 30 minutes. Nearly 0.1587 or 16% of the deliveries for the high variability restaurant exceed 30 minutes. On the other hand, only 2% of the deliveries take too long with the low variability restaurant. They both have an average delivery time of 20 minutes, but where would you place your order when you're hungry? Of course, you should pick the restaurant with more consistent delivery time.

You see, the central tendency doesn't always provide complete information. We also need to understand the variability around the middle of the distribution to get the full picture.

Just as there are multiple measures of central tendency, there are also several measures of variability. In this module, you'll learn the four measures of variability, namely, the range, average deviation, variance and standard deviation.

Range

The range of a set of data values is obtained by getting the difference between the largest data value and the lowest data value.

$$Range = Largest\ Data - Smallest\ Data$$

Discussion

Example

1.) In a cafeteria, there are three *vendo* machines which are set to dispense 10 ounces of coffee to a cup. For the first five trials, the given table below shows the volume (in oz.) of coffee dispensed by the machines to each cup. Find the range for each machine.

Machine 1	Machine 2	Machine 3
11.52	10.01	8.02
8.41	9.99	12.03
12.07	9.95	7.94
7.85	10.03	11.99
10.15	10.02	10

Machine 1: Range = 12.07 – 7.85 = 4.22 oz.
Machine 2: Range = 10.03 – 9.95 = 0.08 oz.
Machine 3: Range = 12.03 – 7.94 = 4.09 oz.

This indicates that Machine 1 varies the most in volume of dispensing soft drinks, which shows its inconsistency unlike for Machine 2, the volume dispensed is quite consistent.

2.) Calculate the range of the following set of data values:

a.) 10, 2, 5, 6, 7, 3, 4 — ***Range= 10 – 2 = 8***
b.) 99, 45, 23, 67, 45, 91, 82, 78, 62, 51 — ***Range = 99 – 23 =76***
c.) 15, 14, 17, 13, 11, 19, 25, 22, 21 — ***Range = 25 –11 = 14***

Discussion

Activity 7

Compute the range of the following data values

a.) 1, 2, 5, 7, 8, 19, 22

b.) 3, 4, 7, 11, 12, 12, 15, 16

c.) -8, -5, -12, -1, 4, 7, 11

d.) 2, 4, 6, 7, 9, -5, -7

e.) 12, 16, 12, 16, 14, 12, 11

Discussion

Average Deviation

The range is the easiest measure of variation that you can get. But it only considers the extreme values; it does not tell us anything about the values between these extreme values. A measure of variation that takes into consideration the deviations of the individual data scores from an average is therefore generally considered more accurate and reliable than those determined by only the single value like the range. One such measure is the ***average deviation or mean deviation*****.**

In calculating the average deviation of a data set of values, we must obtain the mean of the data set first. The average deviation of an **ungrouped data set** can then be obtained by using the equation:

$$Average\ deviation = \frac{\sum |x-\mu|}{N} \qquad \text{for population data}$$

$$Average\ deviation = \frac{\sum |x-\bar{x}|}{n} \qquad \text{for sample data}$$

In calculating the average deviation for a set of **grouped data**, we have to use the formula:

$$Average\ deviation = \frac{\sum f|m-\mu|}{N} \qquad \text{for population data}$$

$$Average\ deviation = \frac{\sum f|m-\bar{x}|}{n} \qquad \text{for sample data}$$

where *m* is the midpoint of a class interval and f is the frequency of the class.

Note:
In both formulas we take the absolute value of each difference from the mean because if we don't, we will always get a 0. That's actually one property of the mean: the sum of the differences of each score from the mean is zero, which reinforces that the mean is at the center of distribution.

Discussion

Example

1.) Solve the average deviation of the scores: 17, 15, 18, 19, 15, 13, 14, 16

Score	$\|x - \mu\|$
17	1.125
15	0.875
18	2.125
19	3.125
15	0.875
13	2.875
14	1.875
16	0.125
Mean = 15.875	Sum = 13

$$Average\ deviation = \frac{13}{8} = 1.625$$

Hence, the average deviation of the scores is 1.625 or 1.63.

Example: Compare the two groups below using mean deviation and interpret your answer.

Test scores of a sample of Math 2 students in the morning and afternoon sessions.

Morning Session	Afternoon Session
15	10
18	12
16	12
20	24
22	24
23	26
23	26
25	28
28	28
30	30

Discussion

Solution: We first compute for the mean of each group.

Morning session: $\bar{x} = \frac{15+18+16+20+22+23+23+25+28+30}{10} = \frac{220}{10} = 22$

Afternoon session: $\bar{x} = \frac{10+12+12+24+24+26+26+28+28+30}{10} = \frac{220}{10} = 22$

Here, we once again see that a measure of central tendency is not enough to differentiate two data sets. These data sets have entirely different values but they have the same mean. The next thing that we will do is subtract each data point from the mean and take the absolute value of the difference. Then we add each set of differences. We get the following table.

Morning Session	$\lvert x - \bar{x} \rvert$	Afternoon Session	$\lvert x - \bar{x} \rvert$
15	7	10	12
18	4	12	10
16	6	12	10
20	2	24	2
22	0	24	2
23	1	26	4
23	1	26	4
25	3	28	6
28	6	28	6
30	8	30	8
Sum	38	Sum	64

We now plug in the sum of the differences to the mean deviation formula for ungrouped data.

Morning Session: $Average\ deviation = \frac{\sum |x-\bar{x}|}{n}$

$= \frac{38}{10}$

$= 3.8$

Afternoon Session: $Average\ deviation = \frac{\sum |x-\bar{x}|}{n}$

$= \frac{64}{10}$

$= 6.4$

This means that the performance of the students in the morning session is more uniform than that of the afternoon session.

Discussion

Example

2.) Solve the average deviation of the scores of students in a Math Quarter exam with the given data from the frequency distribution below

Score	Number of Students
50-59	5
60-69	3
70-79	5
80-89	10
90-99	12

Solution

Let *x* denote the scores of the scores of students in a Math Quarter exam. The values of m and , $|m-\mu|$ and $f|m-\mu|$ are calculated in the table below.

x	f	m	mf	$\|m-\mu\|$	$f\|m-\mu\|$
50-59	5	54.5	272.5	26	130
60-69	3	64.5	193.5	16	48
70-79	5	74.5	372.5	6	30
80-89	10	84.5	845	4	40
90-99	12	94.5	1134	14	168
	$N = 35$		$\mu = 80.5$		416

$$Average\ deviation = \frac{416}{35} = 11.89$$

Hence, the average deviation of the scores of students in a Math Quarter exam is 11.89.

Learning Check

Activity 8

Compute the average deviation of the following data values

a.) 1, 2, 5, 7, 8, 19, 22
b.) 3, 4, 7, 11, 12, 12, 15, 16
c.) -8, -5, -12, -1, 4, 7, 11
d.)

x	*f*
16-18	3
19-21	2
22-24	5
25-27	8
28-30	7

Discussion

Variance and Standard Deviation for Ungrouped Data

The standard deviation is the most commonly used measure of variability. The value of the standard deviation indicates how close the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. Otherwise, it indicates that data set are spread over a relatively larger range around the mean.

The standard deviation is obtained using the basic formulas which are used to calculate the variance:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \qquad s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

where σ^2 is the population variance and $\boldsymbol{s^2}$ is the sample variance?

Consequently, to obtain the formula for calculating the standard deviation is:

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

where σ is the population standard deviation and $\boldsymbol{s}$ is the sample standard deviation.

Discussion

Example

The following table gives the estimation of 2008 market values of five international companies (rounded to trillion of pesos). Compute for the population variance and standard deviation.

Company	Market Value
PepsiCo	3.75
Google	5.35
PetroChina	13.55
Johnson and Johnson	6.90
Intel	3.55

Solution

Let *x* denote the 2008 market value (in billions of dollars) of a company. The values of *x*, $x - \mu$, and $(x - \mu)^2$ are calculated in the table below

x	$x - \mu$	$(x - \mu)^2$
3.75	-2.87	8.2369
5.35	-1.27	1.6129
13.55	6.93	48.0249
6.90	0.28	0.0784
3.55	-3.07	9.4249
$\boldsymbol{\mu = 6.62}$		$\boldsymbol{\sum (x - \mu)^2 = 67.378}$

$\sigma^2 = \frac{67.378}{5} = \mathbf{13.48}$ **variance**

$\sigma = \sqrt{13.48} = \mathbf{3.67}$ **standard deviation**

Hence, the population variance of the 2008 market value of the international companies is 13.47, while the standard deviation is 3.67.

Discussion

Example: Compare the two groups below using variance and standard deviation and interpret your answer.

Test scores of a sample of Math 2 students in the morning and afternoon sessions.

Morning Session	Afternoon Session
15	10
18	12
16	12
20	24
22	24
23	26
23	26
25	28
28	28
30	30

Solution: From our previous solution, we computed the mean as both 22. The next thing that we will do is subtract each data point from the mean and take the square of each difference. Then we add each set of differences. We get the following table.

Morning Session	$x - \bar{x}$	$(x - \bar{x})^2$	Afternoon Session	$x - \bar{x}$	$(x - \bar{x})^2$
15	7	49	10	12	144
18	4	16	12	10	100
16	6	36	12	10	100
20	2	4	24	2	4
22	0	0	24	2	4
23	1	1	26	4	16
23	1	1	26	4	16
25	3	9	28	6	36
28	6	36	28	6	36
30	8	64	30	8	64
		$(x - \bar{x})^2 = 216$	Sum		$(x - \bar{x})^2 = 520$

Discussion

We now plug in the sum of the squared differences to the variance and standard deviation formulas for sample ungrouped data.

Morning Session:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \qquad s = \sqrt{s^2}$$

$$= \frac{216}{9} \qquad = \sqrt{24}$$

$$= 24 \quad \text{variance} \qquad = 4.90 \quad \text{standard deviation}$$

Afternoon Session:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \qquad s = \sqrt{s^2}$$

$$= \frac{520}{9} \qquad = \sqrt{57.78}$$

$$= 57.78 \quad \text{variance} \qquad = 7.60 \quad \text{standard deviation}$$

These values further verifies that the performance of the students in the morning session is more uniform than that of the afternoon session.

Learning Check

Activity 9

The following table shows the grades of a Grade 7 student in the previous quarter. Calculate for its population variance and standard deviation.

Subject	Grade
Math	91
Filipino	94
English	89
Araling Panlipunan	90
PE	95

Discussion

Variance and Standard Deviation of Grouped Data

Following are the *basic formulas* used to calculate the population and sample variances for a set of grouped data respectively:

$$\sigma^2 = \frac{\sum f(m-\mu)^2}{N} \qquad s^2 = \frac{\sum f(x-\bar{x})^2}{n-1}$$

where *m* is the midpoint of a class and and $\bar{x}$ or μ are the means. Also, the standard deviation is obtained by taking the positive square root of the variance.

$$\sigma^2 = \sqrt{\sigma} \qquad s = \sqrt{s}$$

Example

The table below gives the frequency distribution of the daily commuting times (in minutes) from home to work for a sample of 25 employees of a company. Compute for the sample variance and sample standard deviation.

Daily Commuting Time of Employees (minutes)	Number of Employees
1-10	4
11-20	9
21-30	6
31-40	4
41-50	2

Solution

(1) Let *x* denote the daily commuting time of employees of a company. The values of $m - \bar{x}$, $(m-\bar{x})^2$, and $f(m-\bar{x})^2$ are calculated in the table below.

x	*f*	*m*	*mf*	$m-\bar{x}$	$(m-\bar{x})^2$	$f(m-\bar{x})^2$
1-10	4	5.5	22	-16.4	268.96	1075.84
11-20	9	15.5	139.5	-6.4	40.96	368.64
21-30	6	25.5	153	3.6	12.96	77.76
31-40	4	35.5	142	13.6	184.96	739.84
41-50	2	45.5	91	23.6	556.96	1113.92
			$\bar{x} = 21.9$			$\sum = 3376$

$$s^2 = \frac{3376}{25-1} = 140.67 \qquad s = \sqrt{140.67} = 11.86$$

Hence, the sample variance of the daily commuting times of the employees of a company is 140.67, and the sample standard deviation is 11.86.

Learning Check

Activity 10

Calculate the sample variance and sample standard deviation of the scores of students in a Chemistry quiz as shown in the table below:

Score	Number of Students
16-18	9
19-21	6
22-24	7
25-27	5
28-30	3

Discussion

B. Review of Inferential Statistics

- Descriptive statistics applied to populations, and the properties of populations, like the mean or standard deviation are called **parameters.** A **parameter** is a numerical measurement describing some characteristics of a population.
- Properties of samples, such as the mean or standard deviation, are not called parameters, but **statistics** (not to be confused with statistics as a science, this one has a singular term, statistic). A **statistic** is a numerical measurement describing some characteristic of a sample (McGuckian, n.d.).

Inferential statistics are techniques that allow us to use samples to make generalizations about the populations from which the samples were drawn (Nazir, 2015). The process of accurately achieving this is called sampling. Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population.

The methods of inferential statistics are (1) the estimation of parameter(s) and (2) testing of statistical hypotheses (Laerd Statistics, 2018).

In this module we are only going to focus on testing of statistical hypothesis.

Testing of Statistical Hypothesis

The main purpose of statistics is to test a hypothesis. For example, you might run an experiment and find that a certain drug is effective at treating mild symptoms of COVID-19. But if you can't repeat that experiment, no one will take your results seriously.

A **hypothesis** is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation (American Public University System, 2020). For example:

- A new medicine you think might work.
- A new mode of learning you think might be better.
- A possible location of new species.
- A fairer way to administer standardized tests.

It can really be *anything at all* as long as you can put it to the test.

Discussion

Statistical Hypothesis

A statistical hypothesis is an assertion or conjecture concerning one or more populations (Paiva, 2010).

Examples:

1. A medical researcher claims that:
   "The mean body temperature of healthy adults is not equal to 98.6°F (37°C)."

2. The president of a TV company claims that:
   "The majority of all adults are not annoyed by violence on television."

3. The president of a bank claims that:
   "With a single line, customers have more consistent waiting times, with less than 6.2-min standard deviation."

In the next page is the general sense of the thinking used in the tests of such statements, we call this **hypothesis tests** or **tests of significance.**

Example:

Suppose that you take a coin from your pocket and claim that it favors heads when it is flipped. The claim is a hypothesis, and we can test it by flipping the coin 100 times. We would expect to get around 50 heads with a fair coin. If heads occur 94 times out of 100 tosses, most people would agree that the coin favors heads. If heads occur 51 times out of 100 tosses, we should not conclude that the coin favors heads because we could easily get 51 heads with a fair and unbiased coin. But what about 60 heads out of 100 tosses? Or 75? **Here is the key point:** We should conclude that the coin favors heads only if we get ***significantly*** more heads than we would expect with an unbiased coin. A result of 51 heads out of 100 tosses could easily happen by chance- it is not significant. But an outcome of 94 heads out of 100 tosses is not likely to happen by chance and is therefore significant.

This brief example illustrates the basic approach used in testing hypotheses. The formal method involves a variety of standard terms and conditions incorporated in an organized procedure.

Discussion

Components of A Formal Hypothesis Test

Two types of Statistical Hypotheses:

1. The **null hypothesis** (denoted by H_0) is a statement about the value of a population parameter (such as the mean (μ), and it must contain the condition of equality (that is, it must be written with the symbol $=, \leq, \geq$).

   For the mean, the null hypothesis will be stated in only one of three possible forms:

   H_0: $\mu =$ some value or
   H_0: $\mu \leq$ some value or
   H_0: $\mu \geq$ some value

   In short, the null hypothesis is a hypothesis of no difference, and it is ordinarily formulated with the purpose of being accepted or rejected.

For example, the null hypothesis corresponding to the common belief that the mean body temperature is 98.6°F is expressed as H_0: $\mu = 98.6$. We test the hypothesis directly in the sense that the conclusion will be either a rejection of H_0 or a failure to reject H_0.

2. The **alternative hypothesis** (denoted by H_1) is a statement that must be true if the null hypothesis is false.

   For the mean, the alternative hypothesis will be stated in only one of three possible forms:

   H_1: $\mu \neq$ some value or
   H_1: $\mu <$ some value or
   H_1: $\mu >$ some value

For example, if H_0: $\mu = 98.6$, then it follows that the alternative hypothesis is given by

$$H_1: \mu \neq 98.6.$$

The alternative Hypothesis is classified as directional and nondirectional.

a.) ***Nondirectional*** – asserts that there is a significant difference between two measures (uses = or ≠)
b.) ***Directional*** – asserts that one measure is greater (or less) than another measure of similar nature (uses > or <)

Discussion

Very Important notes:

- Depending on the original wording of the problem, the original claim will sometimes be the null hypothesis H_0, and other times it will be the alternative hypothesis H_1. Regardless of whether the original claim corresponds to H_0 or H_1, the null hypothesis H_0 must always contain equality (with the symbolic form of $=, \leq, or \ \geq$).

- Even though we sometimes express H_0 with the symbol $\leq or \geq$ like in H_0: $\mu \leq 98.6$ or H_0: $\mu \geq 98.6$ we conduct the test by assuming that H_0: 98.6 is true. We must have a fixed and specific value for μ so that we can work with a single distribution having a specific mean.

- If we are making our own claims, we should arrange the null and alternative hypothesis so that the most serious error would be the rejection of a true null hypothesis. Ideally, all claims would be made so that they would all be null hypotheses. Unfortunately, our real world is not ideal- there are some problems that involve claims that are null hypothesis, whereas others involve claims that are alternative hypothesis. For example, a salesperson might claim that Diehard batteries last more than 5 years. The original claim of $\mu > 5$ does not contain equality, so it is an alternative hypothesis and the null hypothesis becomes $\mu \leq 5$. If the salesperson claimed that the mean was at least 5 years ($\mu \geq 5$), than the original claim would become the null hypothesis because it would contain equality.

Discussion

Example

The following are examples of statistical hypothesis:

1. H_0: The average monthly salary of nurses in a government hospital is ₱12,000 (μ = ₱12,000).
   H_1: The average monthly salary of nurses in a government hospital is ₱12,000 ($\mu \neq$ ₱12,000).

2. H_0: There is no significant difference between the average battery life of smartphone brand X and that of smartphone brand Y ($\mu_x = \mu_y$).
   H_1: There is a significant difference between the average battery life of smartphone brand X and that of smartphone brand Y ($\mu_x \neq \mu_y$).

3. H_0: The percentage of Puerto Princesa City college students who prefer modular learning is 60% (p = 60%).
   H_1: The percentage of Puerto Princesa City college students who prefer modular learning is greater than 60% (p > 60%).

4. H_0: The percentage of Twitter users posting tweets from 6:00 to 7:00 in the evening is the same on Fridays and Saturdays ($p_1 = p_2$).
   H_1: The percentage of Twitter users posting tweets from 6:00 to 7:00 in the evening is less on Fridays than Saturdays ($p_1 < p_2$).

*We can observe that the first two examples are directional while the last two examples are nondirectional.

Learning Check

Activity 11

Write a pair of null and alternative statistical hypothesis (can be about any topic).

Criteria for Rating

Score	Criterion Description
6-10	Student was able to create a null and alternative hypothesis about a certain topic correctly.
1-5	The null and alternative hypothesis provided by the student was original and/or doesn't fit the definition of a hypothesis.
0	Student has no output for this activity.

Discussion

Two types of errors

When testing a null hypothesis, we arrive at a conclusion of rejecting it or failing to reject it. Such conclusions are sometimes correct and sometimes wrong, but there are two different types of error that can be made.

This is a table showing the four possible consequences of decisions in testing hypotheses and it is summarized in the table below:

DECISION	H_0 IS TRUE	H_0 IS FALSE
ACCEPT H_0	Correct decision	*Type II error*
REJECT H_0	*Type I error*	Correct decision

- **Type I error.** The mistake of rejecting the null hypothesis when it is true. The type I error is not a miscalculation or procedural misstep; it is an actual error that can occur when a rare event happens by chance. The probability of rejecting the null hypothesis when it is true is called the **significance level**; that is, the significance level is the probability of a type I error. The symbol ***α* (alpha)** is used to represent the significance level. The values of $\alpha = 0.05$ and $\alpha = 0.01$ are commonly used.
- **Type II error.** The mistake of failing to reject the null hypothesis when it is false. The symbol ***β* (beta)** is used to represent the probability of a type II error.

In short, if we accept a true null hypothesis or reject a false null hypothesis, we arrive at a perfect decision. However, if we reject H_0 when, in fact, it is true, we commit a *type I error;* and if we accept H_0 when, in fact, it is false, we commit a *type II error.*

The probability of committing a type I error is denoted by α and the probability for committing type II error is denoted by β. Both errors commonly have a value of .01 or .05. Generally, the lower the value of α is, the lesser the probability of rejecting a true null hypothesis and higher probability of accepting a false null hypothesis.

Level of Significance

One step in our procedure for testing hypothesis involves the selection of α. However, we don't select β. It would be great if we could always have $\alpha = 0$ and $\beta = 0$, but in reality, that is not possible and we must attempt to manage the α and β error probabilities.

In specifying the probability of committing a type I error α, which is commonly known as the level of significance, we assume some degree of confidence with our decision.

Discussion

The following terms are associated with the key components in hypothesis testing procedure.

- Test statistic: A sample statistic or a value based on the sample data. A test statistic is used in making the decision about the rejection of the null hypothesis.

- Critical region: The set of all values of the test statistic that would cause us to reject the null hypothesis.

- Critical value: The value or values that separate the critical region from the values of the test statistic that would not lead to rejection of the null hypothesis. The critical values depend on the nature of the of the null hypothesis, the relevant sampling distribution, and the level of significance α.

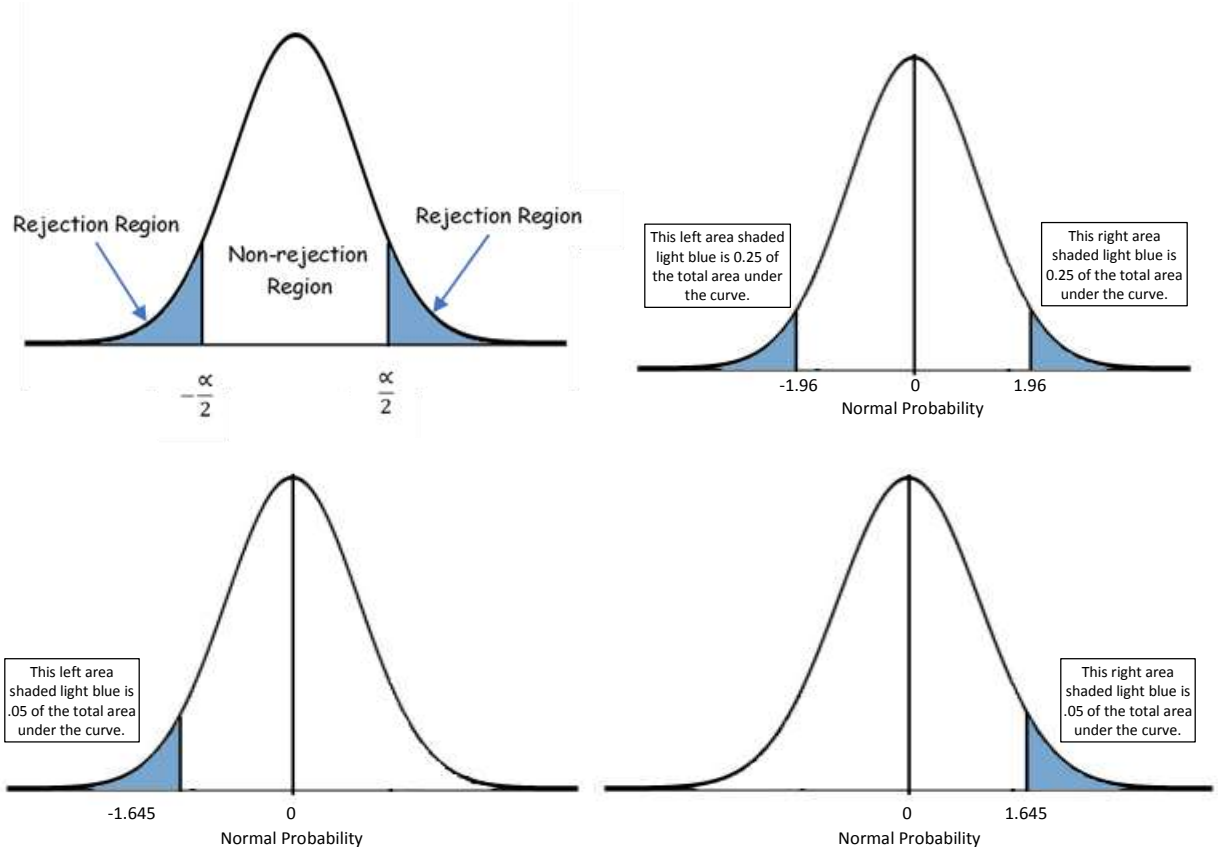
In short, the use of level of significance can determine a critical value which defines a *rejection region* (or *critical region*) and *acceptance region*, which serves as basis for either accepting or rejecting a hypothesis. The level of significance also gives the area of the rejection region. For example, when α= 0.05, the area of the rejection region is 0.05 and the area of the acceptance region is 0.95.

One-tailed and Two-tailed Tests

The **tails** in a distribution are the extreme regions bounded by critical values. Our informal example of hypothesis testing involved a **two- tailed** test in the sense that the critical region is in the two extreme regions under the curve. We reject the null hypothesis H_0 if our test statistic is in the critical region because that indicates a significant discrepancy between the null hypothesis and the sample data. Some tests will be **left- tailed**, with the critical region located in the extreme left region under the curve. Other tests may be **right-tailed**, with the critical region in the extreme right region under the curve.

In two- tailed tests, the level of significance α is divided equally between the two tails that constitute the critical region. For example, in a two- tailed test with a significance level of $\alpha = 0.05$, there is an area of 0.025 in each of the two tails. In tests that are right- or left-tailed, the area of the critical region is α.

Discussion



In short, a *one-tailed test* is a test of any statistical hypothesis when H_1 is directional. A *two-tailed test* is when H_1 is nondirectional. In the first test, the rejection region lies entirely at one side of the distribution, while in the second test, the rejection region is split into two equal areas placed at the tails of the distribution.

In determining the critical value, the type of test and level of significance is considered in connection with the rejection or acceptance of H_0.

Level of Significance	One-sided	Two-sided
α = .05	z > 1.645 (or z < -1.645)	z > 1.96 or z < -1.96
α = .01	z > 2.33 (or z < -2.33)	z > 2.575 or z < -2.575

Discussion

For a one-tailed test the critical value is z = ±1.645 when α = .05, and z = ±2.33 when α = .01. Meanwhile, the critical value for a two- tailed test which requires the rejection region to lie on the right/left tail of the distribution is z = ±1.96 when α = .05 and z = ±2.575, respectively, when α = .01.

Example: A claim about the population mean weight of all aspirin tablets is tested with a significance level of $\alpha = 0.05$. The conditions are such that the standard normal distribution can be used. Find the critical value(s) of z if the test is

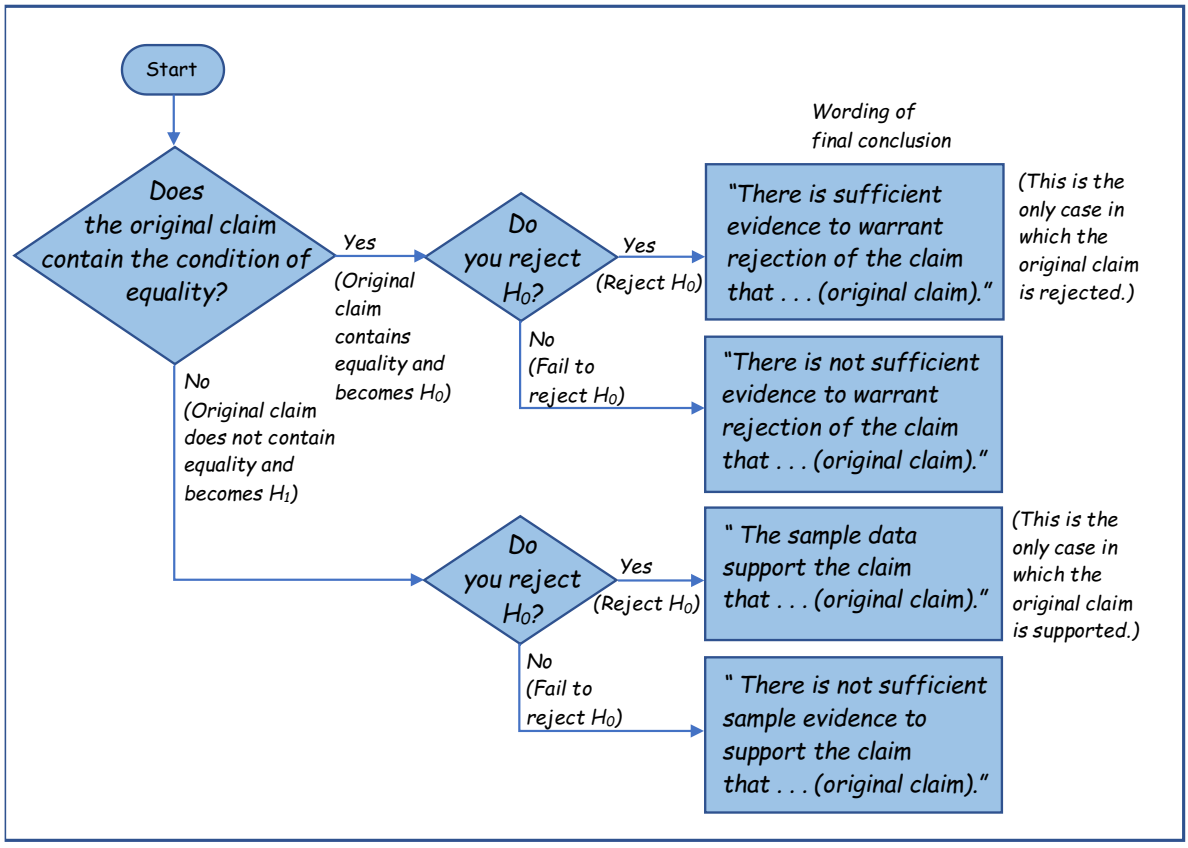
(a) two- tailed
(b) left- tailed
(c) right- tailed.

Solutions:

(a) two- tailed: z >1.96 or z < -1.96
(b) left- tailed: z < -1.645
(c) right- tailed: z > 1.645

Writing Conclusions in Hypothesis Testing

Use the chart below to guide you in writing the conclusion in hypothesis testing. Remember that it is important to be precise in the language used.



Discussion

Example: After analyzing 106 body temperatures of healthy adults, a medical researcher makes a claim that the mean body temperature is *less than* $98.6°\text{F}$ *(or* $37°\text{C}$*)*

a. Express this claim in symbolic form.
b. Identify the null hypothesis H_0.
c. Identify the alternative hypothesis H_1.
d. Identify this test as being two- tailed, left- tailed, or right- tailed.
e. Identify the Type I error for this test.
f. Identify the Type II error for this test.
g. Assume that the conclusion is to reject the null hypothesis. State the conclusion in nontechnical terms; be sure to address the original claim.
h. Assume that the conclusion is failure to reject the null hypothesis. State the conclusion in nontechnical terms; be sure to address the original claim.

Solutions:

a. The claim in symbolic form: The mean body temperature is $\mu < 98.6°\text{F}$.

b. Null hypothesis H_0: The mean body temperature is $\mu \geq 98.6°\text{F}$.[1]

c. Alternative hypothesis H_1: The mean body temperature is $\mu < 98.6°\text{F}$.

d. This test is left- tailed.

e. We commit Type I error if we reject the claim that the mean body temperature is less than $98.6°\text{F}$ and find out it is true.

f. We commit Type II error if we accept the claim that the mean body temperature is less than $98.6°\text{F}$ and find out it is not true.

g. If we reject the null hypothesis, the conclusion would be: There is sufficient evidence that the mean body temperature is less than $98.6°\text{F}$.

h. If we fail to reject the null hypothesis, the conclusion would be: There is no sufficient evidence that the mean body temperature is less than $98.6°\text{F}$.

Note: We use the more technically correct term “fail to reject” instead of “accept” the null hypothesis because NOT REJECTING a claim does not necessarily mean that we are ACCEPTING it.

[1] Since the opposite of < is > or =

Discussion

Fundamentals of Hypothesis Testing

The following are the basic steps in testing a hypothesis:

1. Formulate the null and alternative hypothesis (H_0).
2. Specify the level of significance α.
3. Choose the appropriate test statistic.
4. Establish the critical region.
5. Compute for the value of the statistical test[2].
6. Make a decision and draw a conclusion, if possible.

Testing a Hypothesized Value of the Mean

Testing a Claim About a Mean (Large Samples n≥30)

The test statistic that is applicable in testing the hypothesis of a mean is the z statistic (or z- test) if:

1. The parameter σ is given.
2. The parameter σ is not given, but the sample size *n* ≥ 30

The formula for z- test is $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

In some situations, σ is not given, but if n ≥ 30, we can we use the sample standard deviation *s* to estimate σ. So, to compute the z statistic when σ is not given, we have:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Example: It is currently a common belief that the mean body temperature of healthy adults is 98.6°F. But when a study was conducted by University of Maryland researchers, they found out that for a sample of 106 temperatures, the sample mean is $\bar{x} = 98.2$, the sample standard deviation is $s = 0.62$, and the distribution is approximately bell shaped. Use a 0.05 significance level to test the claim that the mean body temperature of healthy adults is equal to $98.6°F$.

[2] The statistical test will depend on the hypothesis and size of sample.

Discussion

Solution:

Step 1: Null hypothesis: The mean body temperature of healthy adults is equal to $98.6°\text{F}$.

(H_0: $\mu = 98.6$)

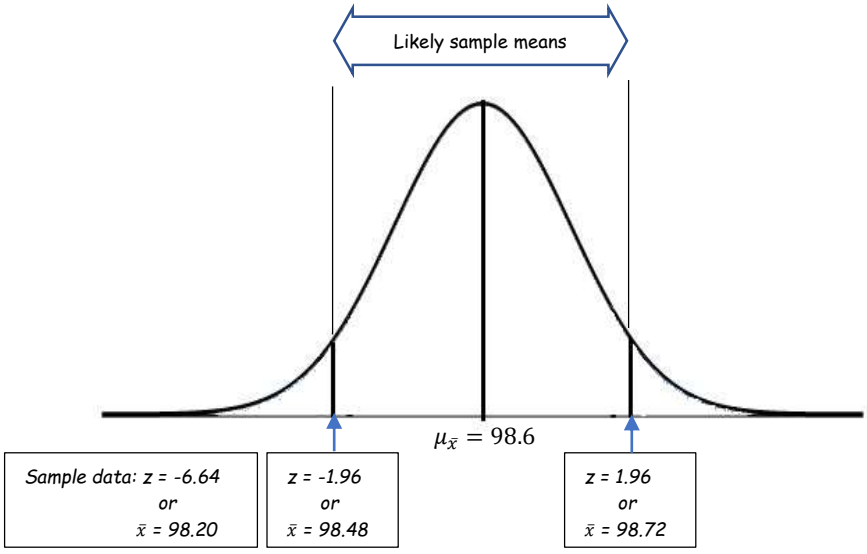
Alternative hypothesis: The mean body temperature of healthy adults is not equal to $98.6°\text{F}$.

(H_0: $\mu \neq 98.6$)

Step 2: The significance level is $\alpha = 0.05$

Step 3: Test statistic: Because the claim is made about the population mean μ, and the sample is large, the test statistic will be z- test, and since the symbol used in the alternative hypothesis is $\neq$ (nondirectional), the test will be two- tailed.

Step 4: The critical region is z<-1.96 or z>1.96 as shown in the picture below.



Step 5. Compute the z- statistic. From the given problem we have the following information: $\mu_0 = 98.6$; $n = 106$; $\bar{x} = 98.2$; and since the sample is large, we can use s=0.62 to estimate σ.

$$z = \frac{\bar{x} - \mu_0}{\sigma/n} = \frac{98.2 - 98.6}{\left(\frac{0.62}{\sqrt{106}}\right)} = \frac{-0.4}{0.0602} = -6.64$$

Step 6: Decision and conclusion: Reject H_0 (because -6.64<-1.96, it is within the critical region). There is sufficient evidence to warrant rejection of the claim that the mean body temperature of healthy adults is $98.6°\text{F}$.

Discussion

Example: The Mill Valley Brewery distributes beer in bottles labeled 32 oz. The Bureau of Weights and Measures randomly selects 50 of these bottles, measures their contents, and obtains a sample mean of 31.8 oz and a sample standard deviation of 0.75 oz. Using a 0.01 significance level, test the Bureau's claim that the brewery is cheating consumers (that is, the mean content of beer bottles is less than 32 oz).

Solution:

Step 1: Null hypothesis: The mean content of beer bottles is greater than or equal to 32 oz.

(H_0: $\mu \geq 32$)

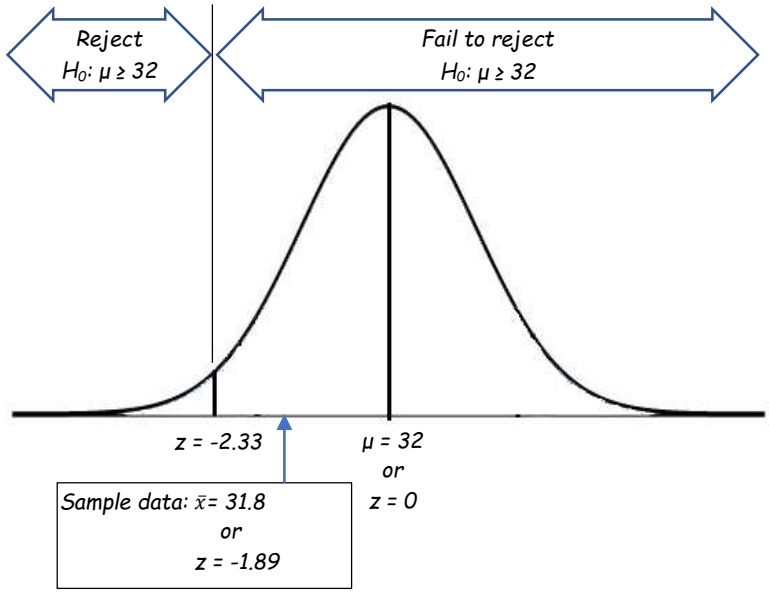
Alternative hypothesis: The mean content of beer bottles is less than 32 oz.

(H_0: $\mu < 32$)

Step 2: The significance level is $\alpha = 0.01$

Step 3: Test statistic: Because the claim is made about the population mean μ, and the sample is large, the test statistic will be z- test, and since the symbol used in the alternative hypothesis is $<$ (nondirectional), the test will be one- tailed, specifically left- tailed.

Step 4: The critical region is z<-2.33.



Step 5. Compute the z- statistic. From the given problem, we have the following information: $\mu = 32; n = 50, \bar{x} = 31.8$, and since the sample is large, we can use s=0.75 to estimate σ.

Discussion

$$z = \frac{\bar{x} - \mu_0}{\sigma/n} = \frac{31.8 - 32}{\left(\frac{0.75}{\sqrt{50}}\right)} = \frac{-0.2}{0.1061} = -1.89$$

Step 6: Decision and conclusion: Fail to reject H_0 (because -1.89 is not less than -1.96, it is not within the critical region). There is no sufficient evidence to support the claim that the brewery is cheating consumers.

Note: A similar conclusion is: There is no sufficient evidence to warrant rejection of the claim that the mean content of beer bottles is greater than or equal to 32 oz.

Example

An instructor gives his class a midterm exam, in which from all the midterm exams he's given to his previous students, his students gets an average score of 81. His current class of 40 students gets a mean of 88 and a standard deviation mean of 8.5. Can he claim that his current class is superior to his previous classes? (Use α= .01)

Solution

H_0 : μ = 81

H_1 : μ > 81

Significance level: α= .01, one-tailed test
Test statistic: z statistic
Critical region: z > 2.33 (superior implies greater than)
Computations:

n = 40
$\bar{x}$ = 81
μ_0 = 88
σ = 8.5

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \left(\frac{8.5}{\sqrt{40}}\right) = 1.3439$$

$$z = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{(88 - 81)}{1.3439} = 5.21$$

The computed z value is greater than 1.96 so we reject H_0 and conclude that there is sufficient evidence to believe that the instructor's current class is superior.

Discussion

Testing a Claim About a Mean (Small Samples n<30)

If the parameter σ is unknown and the sample size *n* is less than 30, the applicable test statistic in hypothesis testing is t statistic.

The formula for t value is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

with *df = n – 1*, where $s_{\bar{x}}$ is the standard deviation of the sample mean.

Note: df means degree of freedom, together with the level of significance, it gives us the critical region for the hypothesis testing. A table for degrees of freedom for a t-test is provided in the appendix.

Discussion

Example

A doctor at a certain hospital claims that the average blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect. (Use α= .05)

Solution

H_0 : The average blood glucose level of obese patients is 100. (**$\mu = 100$**)

H_1 : The average blood glucose level of obese patients is not 100. (**$\mu \neq 100$**)

Significance level: α= .05
Test statistic: t statistic, two- tailed
Critical region: t > 2.04523

(because df=n-1=30-1=29 and α= $\frac{.05}{2}$ =0.025 because the test is two- tailed. Use t-table in the appendix of the module, select the intersection of the row for 29 and 0.025)

Computations:

$n = 30$

$\bar{x} = 140$

$\mu_0 = 100$

$\sigma = 15$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \left(\frac{15}{\sqrt{30}}\right) = 2.7386$$

$$t = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{(140 - 100)}{2.7386} = 14.61$$

The computed z value is greater than 2.04523 so we reject H_0 .and we can say that there is enough evidence to believe that the average blood glucose level of obese patients is not 100, and the raw cornstarch diet had an effect.

Discussion

Example

Maharlika Corporation replaced their old machines to new ones (latest model) which produces parts of computer monitors. The following data are the numbers of computer monitors produced at the company for a sample of 10 days.

28 31 40 23 28 27 23 35 33 29

If the average number of computer monitors produced per day using the old machines is 30, is the management justified in stating that the number of monitors produced per day can be increased with the new machines? (Use α= .01)

Solution

H_0 : The new machines do not change the average number of monitors produced per day. (**$\mu \leq 30$**)

H_1 : The new machines increases the average number of monitors produced per day. (**$\mu > 30$**)

Significance level: α= .01
Test statistic: t statistic, one-tailed test with *df* = 10 – 1 = 9
Critical region: t > 2.82144
Computations: We first need to compute for the sample mean and standard deviation.

x	$(x-\bar{x})^2$
28	2.89
31	1.69
40	106.09
23	44.89
28	2.89
27	7.29
23	44.89
35	28.09
33	10.89
29	0.49
Σ = 297	**Σ = 250.1**

$$\bar{x} = \frac{\sum x}{n} = \frac{297}{10} = 29.7$$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{250.1}{9}} = \sqrt{27.79} = 5.27$$

Discussion

Solution (continuation)

Computations: Then we plug in these computed values to the t- test formula.

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{(29.7 - 30)}{5.27/\sqrt{10}} = -0.1800$$

Since the computed t-value does not lie in the critical region, we fail to reject the H_0 and conclude that the new machines do not change the average number of monitors produced per day.

Learning Check

Activity 12

1.) A pharmaceutical firm claims that the average time for a drug to take effect is 15 minutes with a standard deviation of 1.5 minutes. In a sample of 40 trials, the average time was 23 minutes. Test thee claim that the alternative that the average time is not equal to 15 minutes, using a .01level of significance.

2.) Past experience has shown that the average length of time for students to enroll in a university is 6.4 hours. A new online enrollment procedure is being tested. If a random sample of 7 students spent 3.5, 5.3, 4.8, 6, 5, 4.7, and 2 hours for the new online enrollment, can it be concluded that average number of hours is less than the old procedure?

Discussion

Testing the Difference Between Means

Sometimes, we need to determine whether or not the mean of one population (μ_1) is equal to the mean of another population (μ_2). We can formulate statistical hypothesis in a two-sample case in three ways:

1. $H_o : \mu_1 = \mu_2$
   $H_1 : \mu_1 \neq \mu_2$ (two-tailed)

2. $H_o : \mu_1 = \mu_2$
   $H_1 : \mu_1 < \mu_2$ (one-tailed)

3. $H_o : \mu_1 = \mu_2$
   $H_1 : \mu_1 > \mu_2$ (one-tailed)

Where μ_1 = mean of population 1 and μ_2 = mean of population 2

Just like in the previous section, we use z-statistic when

	Standard Error of the difference between means	z-statistic
1.) σ_1 and σ_2 are known	$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$z = \frac{\bar{x}_1-\bar{x}_2}{\sigma_{\bar{x}_1-\bar{x}_2}} = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}}}$
2.) σ_1 and σ_2 are unknown n ≥ 30	$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$z = \frac{\bar{x}_1-\bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1}{n_1}+\frac{s_2}{n_2}}}$
3.) σ_1 and σ_2 are unknown n < 30	$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_1-1)s_2^2}{n_1+n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ *or* $s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n_1+n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$t = \frac{\bar{x}_1-\bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1}{n_1}+\frac{s_2}{n_2}}}$ $df = n_1+n_2-2$

Discussion

Example

An instructor wishes to determine which of two methods of teaching A or B is more effective in teaching a certain concept in Biology. In a class of 40 students, he used method A and in another class where there are 37 students, he used method B. He gave the two classes the same examination and obtained the following results:

Method A	Method B
$x_1 = 87$	$x_2 = 85$
$s_1 = 6.1$	$s_2 = 5.2$

Is he correct in assuming that method A is more effective than method B? Use α=.05

Solution

H_0 : There is no significant difference between methods A and B. (**$\mu_1 = \mu_2$**)

H_1 : Method A is more effective than method B. (**$\mu_1 = \mu_2$**)

Significance level: α= .05, one-tailed test
Test statistic: z statistic
Critical region: z > 1.645
Computations:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{6.1^2}{40} + \frac{5.2^2}{37}} = \sqrt{0.93025 + 0.730811} = \mathbf{1.28882}$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} = \frac{87 - 85}{1.5518} = \mathbf{1.5518}$$

We may reject H_0 and accept H_1 since the computed value of z is not in the critical region.

Discussion

Example

For a sample of 16 PinoyPhone smartphones, the mean average battery life is 42 hours with a standard deviation of 5.6 hours. For a sample of 12 Doon smartphones, the mean average battery life is 45 hours with a standard deviation of 6.7 hours. (Use α= .01)

Solution

H_0 : There is a significant difference between the average battery lives of two smartphone brands. ($\mu_1 = \mu_2$)

H_1 : There is a significant difference between the average battery lives of two smartphone brands. ($\mu_1 \neq \mu_2$)

Significance level: α= .01, two-tailed test
Test statistic: t statistic with *df* = 16+12 – 2 = 26
Critical region: t > 2.998 or t < -2.998
Computations:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_1-1)s_2^2}{n_1+n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{(16-1)5.6+(12-1)6.7}{16+12}\left(\frac{1}{16}+\frac{1}{12}\right)} = 0.906286$$

$$t = \frac{42-45}{0.906286} = \mathbf{-3.31}$$

We accept the null hypothesis, since the computed value of t is in the region of acceptance.

Discussion

Learning Check

Activity 13

A History instructor wants to test if online learning mode is more effective than modular learning mode. In the final exam, the average score of the 15 students (normal population) in online mode is 87 with a standard deviation σ_1=9.2. The average score of the 18 students (normal population) in modular mode is 75 with a standard deviation σ_2=7.6. Can he claim that her students perform online learning mode do a better performance compared to those who are on modular learning mode?

Discussion

Correlation

- Two variables are correlated if they share a statistical dependence / relationship. For example, measurements of temperature at noon and 1pm every day are correlated, because they both lie consistently above the mean daily temperature
- Correlations between variables are important because they indicate some underlying physical relationship between those variables

Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

- Positive Correlation – when the value of one variable increases with respect to another.
- Negative Correlation – when the value of one variable decreases with respect to another.
- No Correlation – when there is no linear dependence or no relation between the two variables.

Correlation coefficient (r) Formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (y)^2]}}$$

Where:

n = number of pairs of scores
$\sum xy$ = sum of the products of paired scores
$\sum x$ = sum of x scores
$\sum y$ = sum of y scores
$\sum x^2$ = sum of squared x scores
$\sum y^2$ = sum of squared y scores

Discussion

Here's the correlation coefficient interpretation:

To interpret its value, see which of the following values your correlation *r* is closest to:

R-value	Interpretation
(±) Exactly 1	A perfect uphill/downhill linear relationship
(±)0.70	A strong uphill/downhill linear relationship
(±)0.50	A moderate uphill/downhill relationship
(±)0.30	A weak uphill/downhill linear relationship
0	No linear relationship

Correlation Example

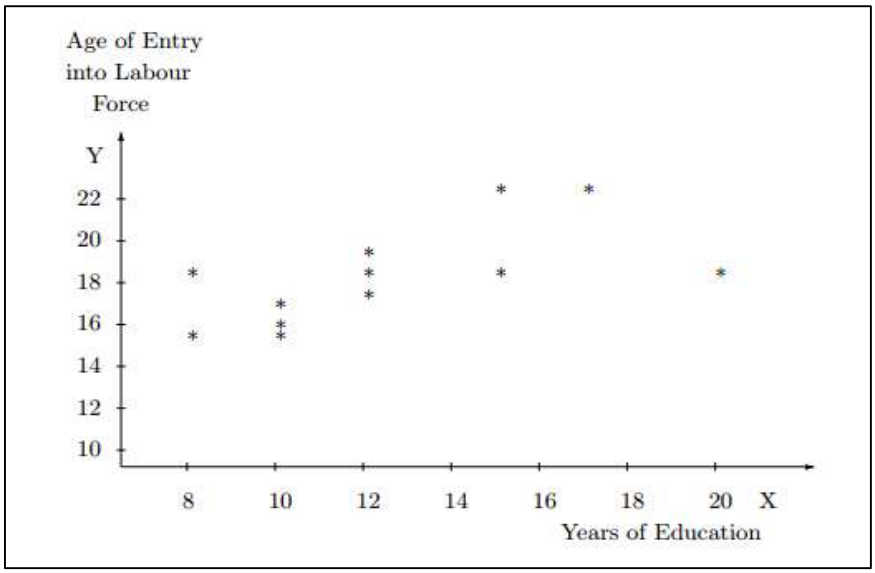
The table below gives the number of years of formal education (X) and the age of entry into the labor force (Y), for 12 males from the Regina Labour Force Survey. Both variables are measured in years, a ratio level of measurement and the highest level of measurement. All of the males are aged 30 or over, so that most of these males are likely to have completed their formal education.

Respondent Number	X	Y
1	10	16
2	12	17
3	15	18
4	8	15
5	20	18
6	17	22
7	12	19
8	15	22
9	12	18
10	10	15
11	8	18
12	10	16

Table 1. **Years of Education and Age of Entry into Labor Force for 12 Regina Males**

Discussion

Since most males enter the labour force soon after they leave formal schooling, a close relationship between these two variables is expected. By looking through the table, it can be seen that those respondents who obtained more years of schooling generally entered the labour force at an older age. The mean years of schooling is $\bar{X}$ = 12.4 years and the mean age of entry into the labour force is $\bar{Y}$ = 17.8, a difference of 5.4 years.



This difference roughly reflects the age of entry into formal schooling, that is, age five or six. It can be seen through that the relationship between years of schooling and age of entry into the labor force is not perfect. Respondent 11, for example, has only 8 years of schooling but did not enter the labor force until age 18. In contrast, respondent 5 has 20 years of schooling but entered the labor force at age 18. The scatter diagram provides a quick way of examining the relationship between X and Y.

Discussion

Example

Based on the given data from Table 1, determine the correlation between the years of education and age of entry into labor force for 12 Regina males.

Solution

We have to compute for $\sum xy$,$\sum x$,$\sum y$,$\sum x^2$, and $\sum y^2$

Respondent Number	x	y	xy	x^2	y^2
1	10	16	160	100	256
2	12	17	204	144	289
3	15	18	270	225	324
4	8	15	120	64	225
5	20	18	360	400	324
6	17	22	374	289	484
7	12	19	228	144	361
8	15	22	330	225	484
9	12	18	216	144	324
10	10	15	150	100	225
11	8	18	144	64	324
12	10	16	160	100	256
$\sum$	$\sum x$= 149	$\sum y$=214	$\sum xy$=2716	$\sum x^2$ =1999	$\sum y^2$ =3876

Substituting this value to the r formula:

$$r = \frac{12(2716) - (149)(214)}{\sqrt{[12(1999) - (149)^2][12(3876) - (214)^2]}}$$

$$r = \frac{706}{\sqrt{1788(716)}} = \frac{706}{\sqrt{111279492}} \approx 0.624$$

With the r-value of 0.624, which is closest to +0.70, there is a strong uphill linear relationship between the years of education and age of entry into labor force for 12 Regina males.

Learning Check

Activity 13

In a Chemistry quiz, the table below shows the scores of 10 male and female students.

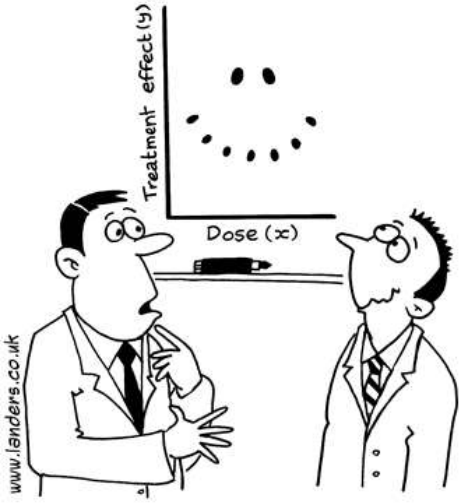
Male (X)	Female (Y)
10	12
15	14
8	10
13	11
7	8
9	9
11	15
12	10
10	9
5	11

Compute the correlation coefficient and make an interpretation.

Discussion

Linear Regression

- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

For instance, a student wants to know whether there is a relationship between the number of hours of sleep and his/her grade in Math. The table below gives the data.

Number of hrs of sleep (x)	8.5	6.8	7.2	8.6	5.9	6.1	5.2	6.3	6.5	8
Grade in Math (y)	2.0	2.25	1.75	2.50	2.25	1.75	3.0	2.75	2.0	1.0

Using these data, a **scatter plot** can be drawn, as shown below.

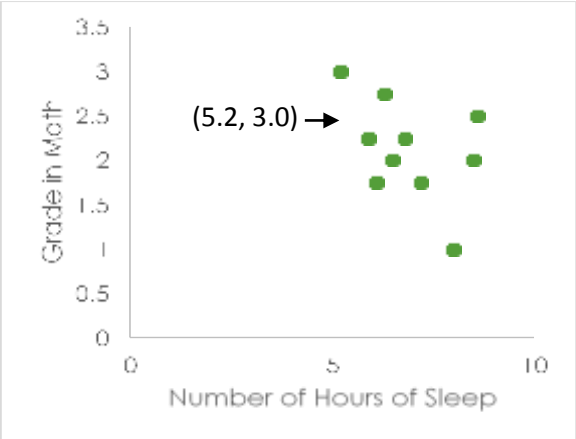


Figure 1

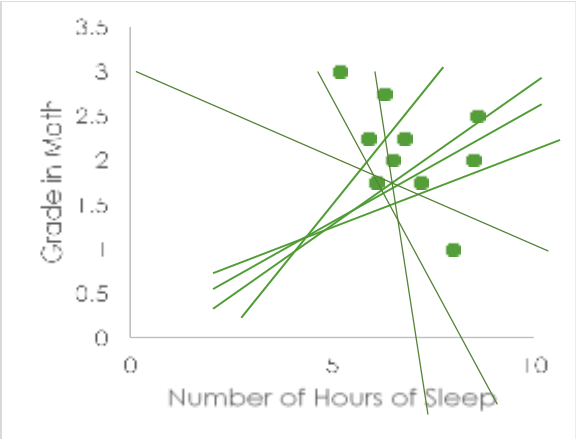


Figure 2

Discussion

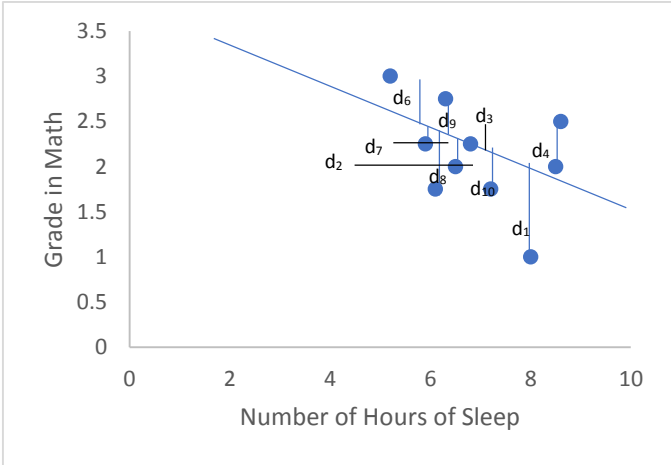
- One way to create a model of relationship between the two variables is to find a line that approximates the data points plotted in the scattered plot.

- Out of all these possible lines, on that is closest to the behavior of the data is called the best-fit line or **least-squares regression line**. This line is best fits the data better than any other line that can be drawn.

- It can be defined as a set of data that minimizes the sum of squares of the vertical deviations from each data point to the line. From this definition we get that the linear equation minimizes the sum

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 + d_7^2 + d_8^2 + d_9^2 + d_{10}^2$$

  is the equation of the best-fit line, where d_n represents the distance from the data point n to the line.



The Formula for the Least-Squares Line

The equation of the least-squares line for the *n* ordered pairs is

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ is $y = ax + b, where\ a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$ and $b = \bar{y}$

Discussion

Example

Apply the Least-Squares Line formula to the given data of the students' number of hours of sleep and grade in Math. Predict the student's grade in Math if he/she sleeps for 10 hours.

Solution

First we find the value of each summation.

$$\sum x = 69.1 \qquad \sum y = 21.25 \qquad \sum x^2 = 489.29 \qquad \sum xy = 144.275$$

We use the values to find the value of *a*.

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{10(144.275) - (69.1)(21.25)}{10(489.29) - (69.1)^2} = -0.216995512$$

When then find the values of $\bar{x}$ and $\bar{y}$,

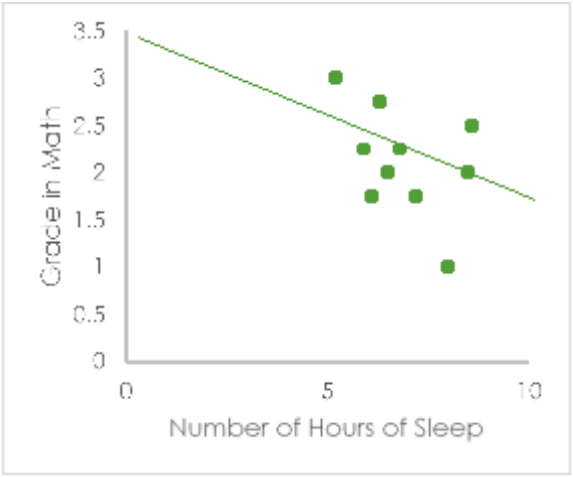
$$\bar{x} = \frac{\sum x}{n} = \frac{69.1}{10} = 6.91 \qquad \text{and} \qquad \bar{y} = \frac{\sum y}{n} = \frac{21.25}{10} = 2.125$$

We use them to find the x-intercept *b*

$$b = \bar{y} - a\bar{x}$$
$$b = 2.125 - (-0.216995516)(6.91) = 3.624439016$$

The regression equation is

$$\boldsymbol{y = -0.216995512x + 3.624439016}$$



Using this equation, to predict the student's grade we substitute 10 hrs to x

$$y = -0.216995512x + 3.624439016 = -0.216995512(10) + 3.624439016 = \mathbf{1.45}$$

We can say that a student who sleeps for 10 hours is likely to get a grade of 1.45 in Math.

Learning Check

Activity 14

A researcher wants to find out the relationship between the speed of an adult man and a cow's pace (stride/walk) and his length of pace. The following gives the data of the two variables for eight men.

a. Adult man

Pace length (meters) x	2.6	2.9	3.2	3.6	3.7	3.9	4.3	4.4
Speed (meters/seconds) y	3.3	4.8	5.6	6.7	7.1	7.6	8.2	8.6

b. Cow

Pace length (meters) x	2.7	3.2	3.4	3.6	3.7	4.0	4.2	4.4
Speed (meters/seconds) y	2.5	3.9	4.6	5.2	5.7	6.4	7.3	7.8

Predict the speed of the adult man and the cow if their pace lengths are 3.0 m and 5.0 m.

Reflection

Express briefly your learning, points of clarification, insights, and feelings towards the given learning material.

Rubrics	**%**
Substance *(depth and validity of the content)*	40%
Relevance *(connection to the topic)*	30%
Comprehensiveness *(extensiveness of the content)*	20%
Clarity *(organization of though)*	10%
Total	100%

Evaluation

I. Solve the puzzle below by encircling the words identified by the descriptions in the following items.

X	A	B	N	O	I	T	A	L	E	R	R	O	C
S	Q	E	X	U	U	N	I	M	O	D	A	L	R
Z	V	O	Y	A	H	I	B	D	D	V	N	E	E
H	J	K	U	W	F	B	M	R	E	L	N	H	G
Y	D	E	W	T	F	H	C	D	B	A	R	N	R
P	Y	J	G	D	L	E	Z	E	I	D	F	D	E
O	M	E	N	E	R	I	X	D	T	O	H	Q	S
T	O	M	R	A	H	Y	E	D	O	M	L	E	S
H	A	K	A	I	Q	M	W	R	E	I	R	T	I
E	B	U	V	E	C	X	Z	P	O	T	N	I	O
S	I	R	O	C	E	S	Y	R	R	L	A	E	N
I	R	M	O	D	E	U	B	N	A	U	E	R	C
S	N	I	N	A	N	Y	O	E	O	M	M	A	I
L	U	T	O	P	R	S	E	N	N	A	E	M	R

1.) This measure of central tendency is defined as the data with the highest frequency
2.) Statistical dependence / relationship between two variables.
3.) This measure of central tendency can be identified by dividing the sum of the data values by the total frequency of the sample or population.
4.) This data is comparatively large or small to the other values among the data set.
5.) An educated guess.

II. Find the mean, median, mode, and average deviation of the given set of data.

1.) -5, 44, 66, 44, 33, 55, 77
2.) 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5
3.) -2, -11, -31, -5, -4, -18, -17
4.) 1, 4, 7, 11, 1, -5, -8, -9, -11
5.) 12.7, 13.7, 11.7, 14.7, 13.7

Evaluation

III. Compute for the mean, median, mode, and variance of the given set of sample data. Round off to the nearest tenth.

1.) The following data shows the number of players there are in each university basketball teams in a certain state.

18	12	16	13	14	12	13	10
11	15	17	8	9	10	21	10

2.) A survey of 10 households noted the average number of visitors they receive in a week. The results are given below.

12	5	8	7	13	20	18	3	5	1

3.) The box below shows a random sample of the number of candies 7-year old children eat in a single day.

6	10	7	6	5	12	9	8

IV. Solve the mean, median, mode, variance, and standard deviation of the given data set in the following problems. Write all necessary solutions and final answer(s).

1. The following is a distribution for the number of runners in a 21-kilometer category of a marathon, with the respective distance which they have run in the marathon.

Distance Ran	Number of Runners
7 – 9	8
10 – 12	6
13 – 15	5
16 – 18	10
19 – 21	6

2. The following is a distribution for the number of quizzes given by a Statistics teacher in the entire school year

Number of Weeks	Number of Companies
16 – 20	5
21 – 25	4
26 – 30	6
31 – 35	7
36 – 40	3

Evaluation

IV. Compute for the correlation coefficient of the given data and make an interpretation.

The following table represents the survey results from the 7 online stores.

Online Store	*Monthly E-commerce Sales (x) (in 1000 s)*	*Online Advertising Sales (y) (1000 s)*
1	9200	42.5
2	8500	37.5
3	13625	70
4	13850	125
5	8275	32.5
6	11900	55
7	9400	32.5

V. Using the same data in the previous part, find the equation least-squares line. Predict the online advertising sales of an online store if its monthly-ecommerce sales is 10,500 and 7,250.

VI. Solve the following word problems.

1.) The average amount of cellular data usage of teenagers daily is normally distributed with a mean of 625 mb. What can you say about this claim if a random sample of 1,254 16-year olds spends 743 mb with a standard deviation of 60 mb? Use a .05 level of significance.

2.) A certain brand of canned sardines is advertised as having a net weight of 15 oz. If the net weights of a random sample of 10 cans are 13.9, 15.2, 14.5, 14.7, 14.3, 15.5, 14.9, 16, 15.2, and 13.5 oz, can it be concluded that the average weight of cans is less than the advertised amount? Use a .01 level of significance.

3.) In a certain university in Iloilo, a study was conducted to determine whether the IQ scores of students who came from provincial high schools differ significantly from those students who came from city high schools. An IQ test was given to 300 (100 from each group) college freshmen and the results are as follows:

Students from provincial high schools: $\bar{x}_1 = 99, s_1^2 = 5$

Students from city high schools: $\bar{x}_2 = 102, s_2^2 = 8$

Evaluation

4.) In order to determine the effect of music on the speed of students in answering exams, an instructor administered a comprehensive test while listening to music to his students and recorded the time that they completed and submitted the test. He divided the class of 32 students into 2 groups. 17 were in the exam room with music playing, and 15 were in the exam room without music. The 17 students who listened to music had an average score of 59 with a standard deviation of 10. The other had an average of 65 with a standard deviation of 15. Can the instructor conclude that music distracts students from taking exams?

Design

In a group of 5 members, conduct a study and create a min-research paper wherein you have to collect data from at least 30 respondents. Apply either Pearson correlation or Linear regression for data analysis. The format of the paper will be provided to you by your teacher.

Answer Key

Learning Check (Activity 1: Mean - Ungrouped Data)

a. ₱155.58
b. 23.5
c. 89.5

Learning Check (Activity 2: Median - Ungrouped Data)

a. 11.5
b. -1
c. 8.38
d. 23
e. 15

Learning Check (Activity 3: Mode - Ungrouped Data)

a. 13
b. 5
c. 1
d. 55
e. 980

Learning Check (Activity 4: Mean – Grouped Data)

Mean = 17.38 = **17**

Learning Check (Activity 5: Median – Grouped Data)

Median = 24.3 = **24**

Learning Check (Activity 6: Mode – Grouped Data)

Median = 29.1= **29**

Answer Key

Learning Check (Activity 7: Range – Ungrouped Data)

a. 21
b. 13
c. 23
d. 16
e. 5

Learning Check (Activity 8: Range)

a. 6.49
b. 4
c. 6.77
d. 0.79

Learning Check (Activity 9: Variance and Standard Deviation-Ungrouped Data)

$\sigma^2 = 5.36; \quad \sigma = 2.31$

Score	Number of Students
16-18	9
19-21	6
22-24	7
25-27	5
28-30	3

Learning Check (Activity 9: Variance and Standard Deviation-Ungrouped Data)

$\sigma^2 = 5.36; \quad \sigma = 2.31$

Learning Check (Activity10: Variance and Standard Deviation-Grouped Data)

$\sigma^2 = 3.39; \quad \sigma = 1.84$

Answer Key

Learning Check (Activity 11: Null and Alternative Hypothesis)

Criteria for Rating

Score	Criterion Description
6-10	Student was able to create a null and alternative hypothesis about a certain topic correctly.
1-5	The null and alternative hypothesis provided by the student was original and/or doesn't fit the definition of a hypothesis.
0	Student has no output for this activity.

Learning Check (Activity 12: z-test and t-test)

H_0 : μ = 15

H_1 : μ ≠ 15

Significance level: α= .01, one-tailed test
Test statistic: z statistic
Critical region: z > 2.575 or z <-2.575
Computations:

n = 40
$\bar{x}$ = 23
μ_0 = 15
σ = 1.5
$s_{\bar{x}} = 0.237171$

Computed value: $z = 33.73$

The computed z value is greater than 2.575 so we reject H_0

Answer Key

Learning Check (Activity 13)

$H_0 : \mu_1 \leq \mu_2$

$H_1 : \mu_1 > \mu_2$

Significance level: α= .05, one-tailed test
Test statistic: t statistic
Critical region: z > 1.645
Computations:

n_1 = 15
n_2 = 18
μ_1 = 87
μ_2 = 75
σ_1= 9.2
σ_2= 7.6

Compute value: $z = 4.2359$

The computed z value is greater than 1.645 so we reject H_0

Learning Check (Activity 14: Pearson Correlation)

r = 0.506933

interpretation: A moderate uphill relationship

Learning Check (Activity 15: Linear Regression)

a. $y = -3.46965x + 2.785217$
b. $y = -6.15023x + 3.171296$

References

1. American Public University System. (2020). Is a hypothesis the same as a theory? [20 June 2020]. Retrieved from: https://apus.libanswers.com/writing/faq/189986
2. Bluman, A. G. (2003). Elementary Statistics: A Step by Step Approach. 5th Ed. McGraw Hill, Inc.
3. CENGAGE (2018). Mathematics in the Modern World.
4. Febre, F. A. (2003). Introduction to Statistics. Quezon City: Phoenix Publishing House, Inc.
5. Galleto, M. (2016). What is Data Management? [10 Oct 2019]. Retrieved from: https://www.ngdata.com/what-is-data-management/
6. Kiernan, D. (2020). The Fundamentals of Hypothesis Testing. [20 June 2020]. Retrieved from: https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Natural_Resources_Biometrics_(Kiernan)/03%3A_Hypothesis_Testing/3.01%3A_The_Fundamentals_of_Hypothesis_Testing
7. Laerd Statistics. (2018). Descriptive and Inferential Statistics. [15 May 2020]. Retrieved from: https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php
8. Macallestter.edu. (n.d.). Data Module #1: What is Research Data? [10 Oct 2019]. Retrieved from: https://libguides.macalester.edu/c.php?g=527786&p=3608664
9. McGuckian, D. (n.d.). Chapter One Notes. [15 May 2020]. Retrieved from: http://faculty.fiu.edu › STA2122_Notes_ChapterOne
10. Nazir, H. (2015). Inferential Statistics. [15 May 2020]. Retrieved from: https://quizlet.com/78042053/inferential-statistics-flash-cards/
11. Paiva, A. (2010). Hypothesis Testing. [20 June 2020]. Retrieved from: https://www.sci.utah.edu/~arpaiva/classes/UT_ece3530/hypothesis_testing.pdf
12. Walpole, M. and M. (2002). Probability and Statistics for Engineers and Scientists. 7th Ed. Prentice Hall Int'l. Inc.
13. wac.colostate.edu http://wac.colostate.edu/docs/llad/v4n1/jamison.pdf
14. https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture9.pdf
15. https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/