OSMEÑA COLLEGES

Academic Year 2021-2022,
2nd Semester

ADVANCE PROBABILITY AND STATISTICS

APS

This Module belongs to:

Name of Student

Course, Year & Section

Address/Contact Number

Module Prepared by:

SHIELA MAE B. INOPIA
Instructor

www.oc.edu.ph
College of Teacher Education

Sampling Distributions and Estimations

Module 7

OVERVIEW

Have you ever wondered how the mean, or average, amount of money per person in a population is determined? It would be impossible to contact 100% of the population, so there must be a statistical way to estimate the mean number of peso per person in the population.

LEARNING OUTCOMES

By the end of this module, you should be able to:

- Understand the inferential relationship between a sampling distribution and a population parameter.
- Graph a frequency distribution of sample means using a data set.
- Understand the relationship between sample size and the distribution of sample means.
- Understand the Central Limit Theorem and calculate a sampling distribution using the mean and standard deviation of a normally distributed random variable.
- Understand the relationship between the Central Limit Theorem and the normal approximation of a sampling distribution.
- Calculate the mean of a sample as a point estimate of the population mean.
- Construct a confidence interval for a population mean based on a sample mean.
- Calculate a sample proportion as a point estimate of the population proportion.
- Construct a confidence interval for a population proportion based on a sample proportion.
- Calculate the margin of error for a point estimate as a function of sample mean or proportion and size.
- Understand the logic of confidence intervals, as well as the meaning of confidence level and confidence intervals.

LEARNING EXPERIENCES & SELF-ASSESSMENT ACTIVITIES (SAA)

Activity

If you had 100 pennies and were asked to record the age of each penny, predict the shape of the distribution. (The age of a penny is the current year minus the date on the coin.)

Analysis

1. Construct a histogram of the ages of the pennies.
2. Calculate the mean of the ages of the pennies.

Abstraction

The purpose of sampling is to select a set of units, or elements, from a population that we can use to estimate the parameters of the population. Random sampling is one special type of probability sampling. Random sampling erases the danger of a researcher consciously or unconsciously introducing bias when selecting a sample. In addition, random sampling allows us to use tools from probability theory that provide the basis for estimating the characteristics of the population, as well as for estimating the accuracy of the samples.

Probability theory is the branch of mathematics that provides the tools researchers need to make statistical conclusions about sets of data based on samples. As previously stated, it also helps statisticians estimate the parameters of a population. A parameter is a summary description of a given variable in a population. A population mean is an example of a parameter. When researchers generalize from a sample, they're using sample observations to estimate population parameters. Probability theory enables them to both make these estimates and to judge how likely it is that the estimates accurately represent the actual parameters of the population.

Probability theory accomplishes this by way of the concept of sampling distributions. A single sample selected from a population will give an estimate of the population parameters. Other samples would give the same, or slightly different, estimates. Probability theory helps us understand how to make estimates of the actual population parameters based on such samples.

Estimating Population Parameters from a Small Sample

Suppose you were to randomly select a sample of only one person from the ten. How close will this sample be to the population mean?

The ten possible samples are represented in the diagram below, which shows the peso bills possessed by each sample. Since samples of one are being taken, they also represent the means you would get as estimates of the population.

₱1.00

₱3.00

₱0.00

₱6.00

₱7.00

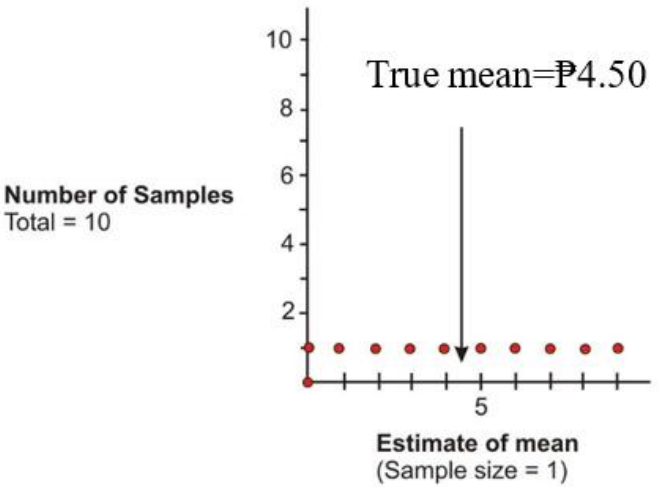
₱5.00

₱9.00

₱8.00

₱2.00

The graph below shows the results:

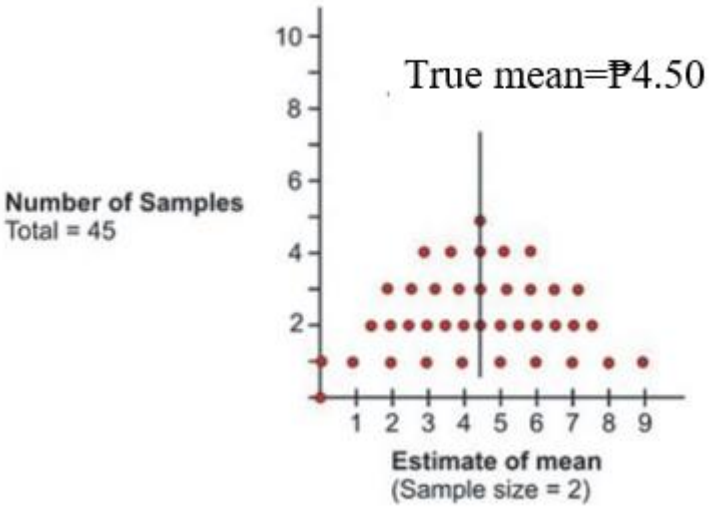


The distribution of the dots on the graph is an example of a sampling distribution. As can be seen, selecting a sample of one is not very good, since the group's mean can be estimated to be anywhere from ₱0.00to ₱9.00, and the true mean of ₱4.50 could be missed by quite a bit.

Estimating Population Parameters from a Larger Sample

What happens if we take samples of two or more?

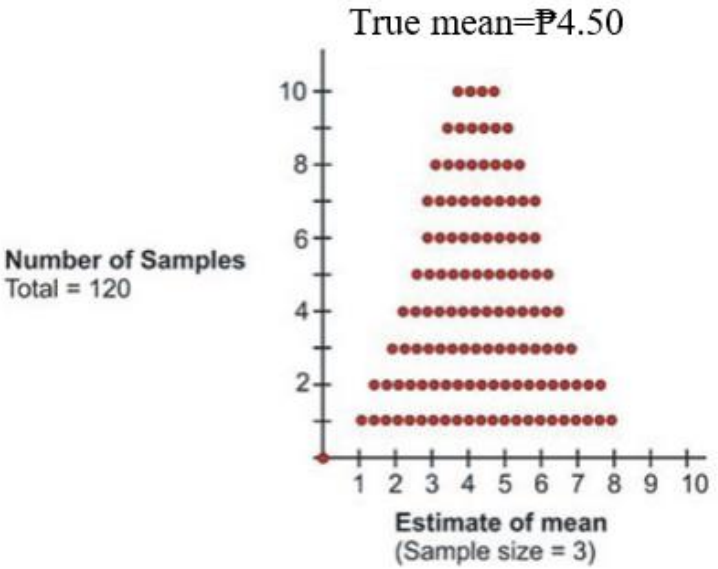
First let's look at samples of size two. From a population of 10, in how many ways can two be selected if the order of the two does not matter? The answer, which is 45, can be found by using a graphing calculator as shown in the figure below. When selecting samples of size two from the population, the sampling distribution is as follows:



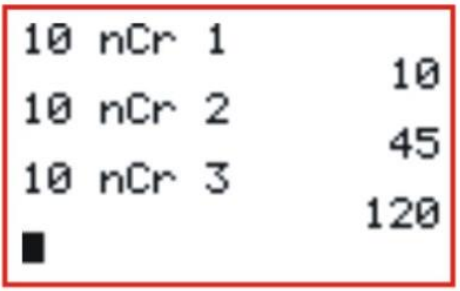
Increasing the sample size has improved your estimates. There are now 45 possible samples, such as (₱0, ₱1), (₱0, ₱2), (₱7, ₱8), (₱8, ₱9), and so on, and some of these samples produce the same means. For example, (₱0, ₱6), (₱1, ₱5), and (₱2, ₱4) all produce means of ₱3. The three dots above the mean of 3 represent these three samples. In addition, the 45 means are not evenly distributed, as they were when the sample size was one. Instead, they are more clustered around the true mean of ₱4.50. (₱0, ₱1) and (₱8, ₱9) are the only two samples whose means deviate by as much as ₱4.00. Also, five of the samples yield the true estimate of ₱4.50, and another eight deviate by only plus or minus 50 cents.

If three people are randomly selected from the population of 10 for each sample, there are 120 possible samples, which can be calculated with a graphing calculator as shown below. The sampling distribution in this case is as follows:

10 Math → PRB ↓ 3. $_nC_r$ 3 **enter** 120

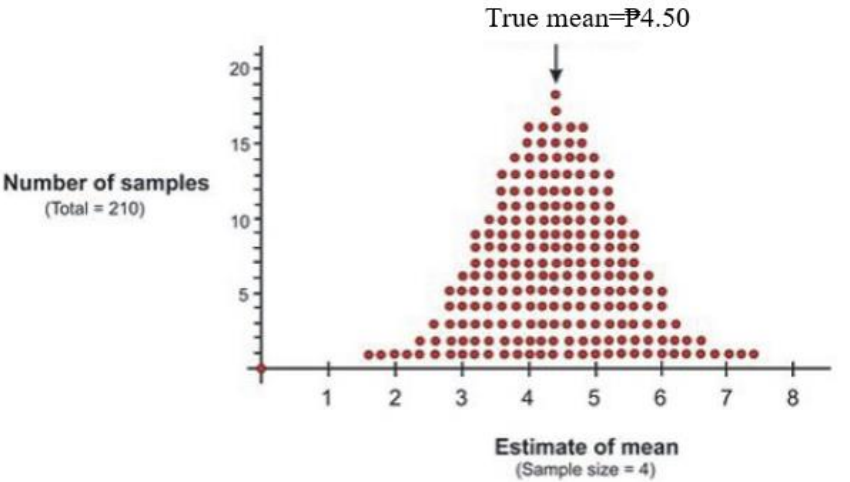


Here are screen shots from a graphing calculator for the results of randomly selecting 1, 2, and 3 people from the population of 10. The 10, 45, and 120 represent the total number of possible samples that are generated by increasing the sample size by 1 each time.

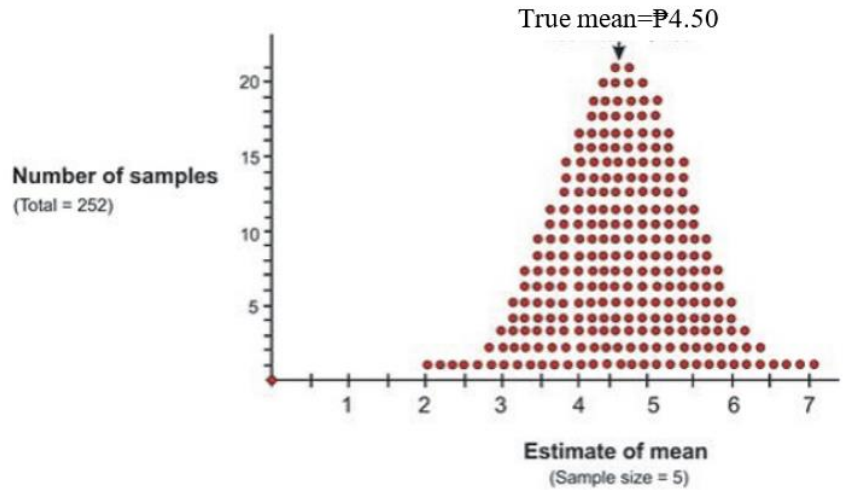


Next, the sampling distributions for sample sizes of 4, 5, and 6 are shown:

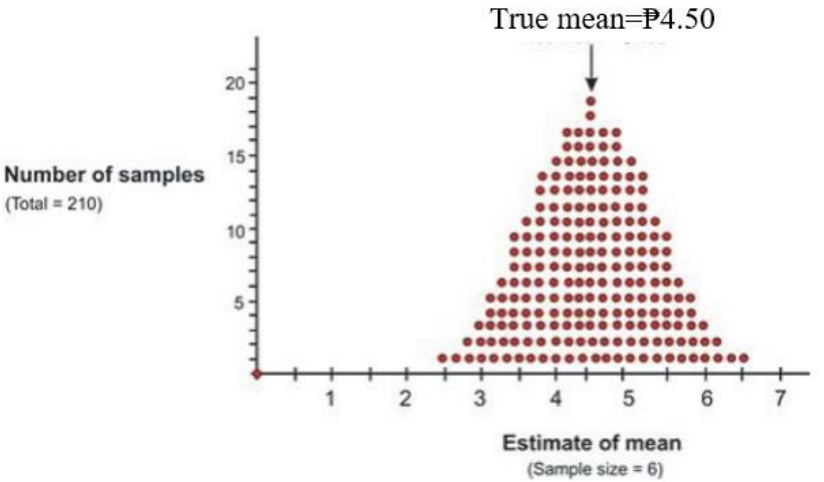
10 Math → PRB ↓ 3. $_nC_r$ 4 **enter** 210



10 Math → PRB ↓ 3. $_nC_r$ 5 **enter** 252



10 Math → PRB ↓ 3. $_nC_r$ 6 **enter** 210



From the graphs above, it is obvious that increasing the size of the samples chosen from the population of size 10 resulted in a distribution of the means that was more closely clustered around the true mean. If a sample of size 10 were selected, there would be only one possible sample, and it would yield the true mean of ₱4.50. Also, the sampling distribution of the sample means is approximately normal, as can be seen by the bell shape in each of the graphs.

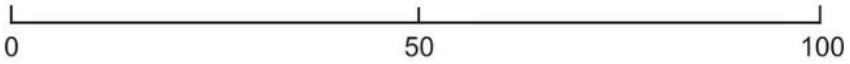
Now that you have been introduced to sampling distributions and how the sample size affects the distribution of the sample means, it is time to investigate a more realistic sampling situation.

Studying a Population through Sampling

Assume you want to study the student population of a university to determine approval or disapproval of a student dress code proposed by the administration. The study's population will be the 18,000 students who attend the school, and the elements will be the individual students. A random sample of 100 students will be selected for the purpose of estimating the opinion of the entire student body, and attitudes toward the dress code will be the variable under consideration. For simplicity's sake, assume that the attitude variable has two variations: approve and disapprove. As you know from the last chapter, a scenario such as this in which a variable has two attributes is called binomial.

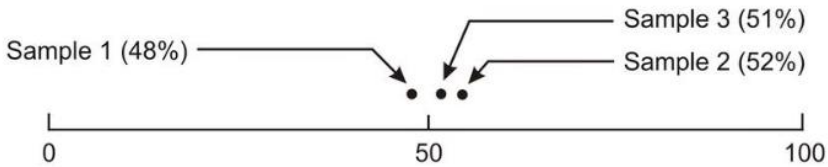
The following figure shows the range of possible sample study results. It presents all possible values of the parameter in question by representing a range of 0 percent to 100 percent of students approving of the dress code. The number 50 represents the midpoint, or 50 percent of the students approving of the dress code and 50 percent disapproving. Since the sample size is 100, at the midpoint,

half of the students would be approving of the dress code, and the other half would be disapproving.



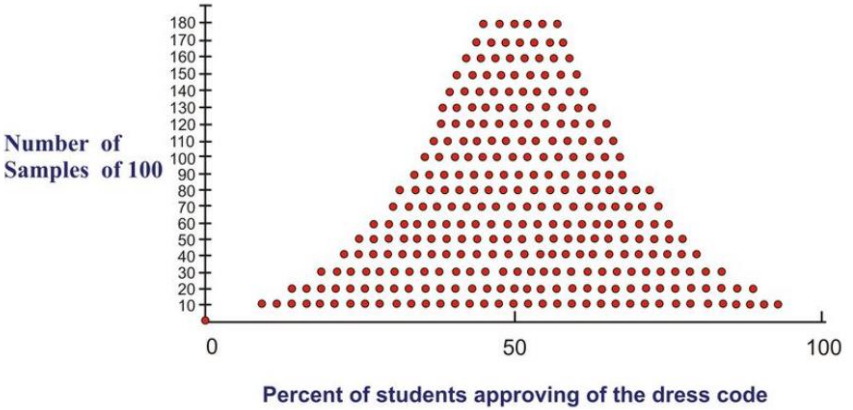
Percent of students approving of the dress code

To randomly select the sample of 100 students, every student is presented with a number from 1 to 18,000, and the sample is randomly chosen from a drum containing all of the numbers. Each member of the sample is then asked whether he or she approves or disapproves of the dress code. If this procedure gives 48 students who approve of the dress code and 52 who disapprove, the result would be recorded on the figure by placing a dot at 48%. This statistic is the sample proportion. Let's assume that the process was repeated, and it resulted in 52 students approving of the dress code. Let's also assume that a third sample of 100 resulted in 51 students approving of the dress code. The results are shown in the figure below.



Percent of students approving of the dress code

In this figure, the three different sample statistics representing the percentages of students who approved of the dress code are shown. The three random samples chosen from the population give estimates of the parameter that exists for the entire population. In particular, each of the random samples gives an estimate of the percentage of students in the total student body of 18,000 who approve of the dress code. Assume for simplicity's sake that the true proportion for the population is 50%. This would mean that the estimates are close to the true proportion. To more precisely estimate the true proportion, it would be necessary to continue choosing samples of 100 students and to record all of the results in a summary graph as shown:



Sampling Error

Notice that the statistics resulting from the samples are distributed around the population parameter. Although there is a wide range of estimates, most of them lie close to the 50% area of the graph. Therefore, the true value is likely to be in the vicinity of 50%. In addition, probability theory gives a formula for estimating how closely the sample statistics are clustered around the true value. In other words, it is possible to estimate the sampling error, or the degree of error expected for a given sample design. The formula $s = \sqrt{\frac{p(1-p)}{n}}$ contains three variables: the parameter,p, the sample size, n, and the **standard error**, s. The symbols p and $1 - p$ in the formula represent the population parameters.

Calculating Standard Error

If 60 percent of the student body approves of the dress code and 40% disapproves, p and $1 - p$ would be 0.6 and 0.4, respectively. The square root of the product of p and $1 - p$ is the population standard deviation. As previously stated, the symbol n represents the number of cases in each sample, and s is the standard error.

If the assumption is made that the true population parameters are 0.50 approving of the dress code and 0.50 disapproving of the dress code, when selecting samples of 100, the standard error obtained from the formula equals 0.05:

$$s = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05$$

This calculation indicates how tightly the sample estimates are distributed around the population parameter. In this case, the standard error is the standard deviation of the sampling distribution.

The **Empirical Rule** states that certain proportions of the sample estimates will fall within defined increments, each increment being one standard error from the population parameter. According to this rule, 34% of the sample estimates will fall within one standard error above the population parameter, and another 34% will fall within one standard error below the population parameter. In the above example, you have calculated the standard error to be 0.05, so you know that 34% of the samples will yield estimates of student approval between 0.50 (the population parameter) and 0.55 (one standard error above the population parameter). Likewise, another 34% of the samples will give estimates between 0.5 and 0.45 (one standard error below the population parameter). Therefore, you know that 68% of the samples will give estimates between 0.45 and 0.55. In addition, probability theory says that 95% of the samples will fall within two standard errors of the true value, and 99.7% will fall within three standard errors. In this example, you can say that only three samples out of one thousand would give an estimate of student approval below 0.35 or above 0.65.

The size of the standard error is a function of the population parameter. By looking at the formula $s = \sqrt{\frac{p(1-p)}{n}}$, it is obvious that the standard error will increase as the quantity $p(1-p)$ increases. Referring back to our example, the maximum for this product occurred when there was an even split in the population. When$p = 0.5$, $p(1-p) = (0.5)(0.5) = 0.25$. If $p = 0.6$, then $p(1-p) = (0.6)(0.4) = 0.24$. Likewise, if p=0.8, then $p(1-p) = (0.8)(0.2) = 0.16$. If p were either 0 or 1 (none or all of the student body approves of the dress code), then the standard error would be 0. This means that there would be no variation, and every sample would give the same estimate. The standard error is also a function of the sample size. In other words, as the sample size increases, the standard error decreases, or the bigger the sample size, the more closely the samples will be clustered around the true value. Therefore, this is an inverse relationship. The last point about that formula that is obvious is emphasized by the square root operation. That is, the standard error will be reduced by one-half as the sample size is quadrupled.

The Central Limit Theorem

The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

Put another way, CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Furthermore, these samples approximate a

normal distribution, with their variances being approximately equal to the variance of the population as the sample size gets larger, according to the law of large numbers.
Before going any further, you should become familiar with (or reacquaint yourself with) the symbols that are commonly used when dealing with properties of the sampling distribution of sample means. These symbols are shown in the table below:

	Population Parameter	Sample Statistic	Sampling Distribution
Mean	μ	$\bar{x}$	$\mu_{\bar{x}}$
Standard Deviation	σ	s	$S_{\bar{x}}$ or $\sigma_{\bar{x}}$
Size	N	n	

As the sample size, n, increases, the resulting sampling distribution would approach a normal distribution with the same mean as the population and with $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The notation $\sigma_{\bar{x}}$ reminds you that this is the standard deviation of the distribution of sample means and not the standard deviation of a single observation.

The Central Limit Theorem states the following:
If samples of size n are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means, $\bar{x}$, approximates a normal distribution as n increases.

The mean of this sampling distribution approximates the population mean, and the standard deviation of this sampling distribution approximates the standard deviation of the population divided by the square root of the sample size: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
These properties of the sampling distribution of sample means can be applied to determining probabilities. If the sample size is sufficiently large (>30), the sampling distribution of sample means can be assumed to be approximately normal, even if the population is not normally distributed.

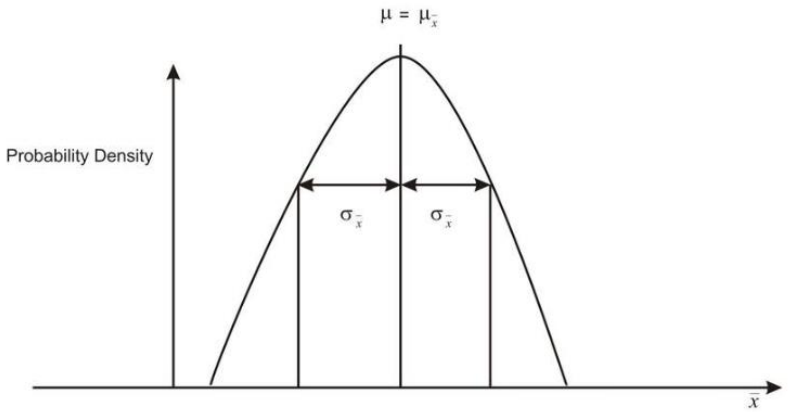
Applying the Central Limit Theorem

1. Suppose you wanted to answer the question, “What is the probability that a random sample of 20 families in Canada will have an average of 1.5 pets or fewer?” where the mean of the population is 0.8 and the standard deviation of the population is 1.2.

For the sampling distribution, $\mu_{\bar{x}} = \mu = 0.8$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{20}} = 0.268$

Therefore, the probability that the sample mean will be below 1.5 is 0.9952. In other words, with a random sample of 20 families, it is almost definite that the average number of pets per family will be less than 1.5.

The properties associated with the Central Limit Theorem are displayed in the diagram below:



The vertical axis now reads probability density, rather than frequency, since frequency can only be used when you are dealing with a finite number of sample means. Sampling distributions, on the other hand, are theoretical depictions of an infinite number of sample means, and probability density is the relative density of the selections from within this set.
The question indicates that multiple samples of size 40 are being collected from a known population, multiple sample means are being calculated, and then the sampling distribution of the sample means is being studied. Therefore, an understanding of the Central Limit Theorem is necessary to answer the question.

Confidence Intervals

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence. Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Your desired confidence level is usually one minus the alpha (a) value you used in your statistical test:

Confidence level = 1 – a

So if you use an alpha value of $p < 0.05$ for statistical significance, then your confidence level would be $1 − 0.05 = 0.95$, or 95%.

When do you use confidence intervals?

You can calculate confidence intervals for many kinds of statistical estimates, including:

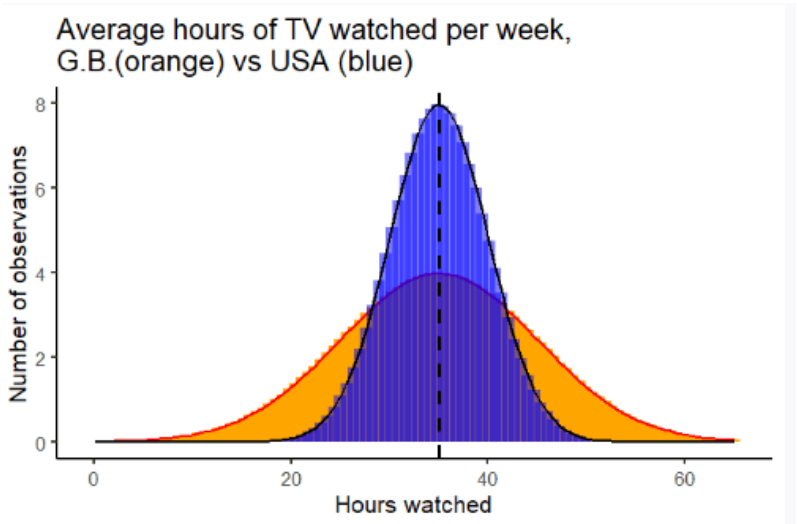
- Proportions
- Population means
- Differences between population means or proportions
- Estimates of variation among groups

These are all point estimates, and don't give any information about the variation around the number. Confidence intervals are useful for communicating the variation around a point estimate.

Example: Variation around an estimate

You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week. However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.



Calculating a confidence interval: what you need to know

Most statistical programs will include the confidence interval of the estimate when you run a statistical test.

If you want to calculate a confidence interval on your own, you need to know:

1. The point estimate you are constructing the confidence interval for
2. The critical values for the test statistic
3. The standard deviation of the sample
4. The sample size

Once you know each of these components, you can calculate the confidence interval for your estimate by plugging them into the confidence interval formula that corresponds to your data.

Point estimate

The point estimate of your confidence interval will be whatever statistical estimate you are making (e.g. population mean, the difference between population means, proportions, variation among groups).

Example: Point estimate

In the TV-watching example, the point estimate is the mean number of hours watched: 35.

Finding the critical value

Critical values tell you how many standard deviations away from the mean you need to go in order to reach the desired confidence level for your confidence interval.

There are three steps to find the critical value.

1. Choose your alpha (a) value.

The alpha value is the probability threshold for statistical significance. The most common alpha value is p = 0.05, but 0.1, 0.01, and even 0.001 are sometimes used. It's best to look at the papers published in your field to decide which alpha value to use.

2. Decide if you need a one-tailed interval or a two-tailed interval.

You will most likely use a two-tailed interval unless you are doing a one-tailed t-test.

For a two-tailed interval, divide your alpha by two to get the alpha value for the upper and lower tails.

3. Look up the critical value that corresponds with the alpha value.

If your data follows a normal distribution, or if you have a large sample size (n > 30) that is approximately normally distributed, you can use the z-distribution to find your critical values.

For a z-statistic, some of the most common values are shown in this table:

Confidence level	90%	95%	99%
alpha for one-tailed CI	0.1	0.05	0.01
alpha for two-tailed CI	0.05	0.025	0.005
z-statistic	1.64	1.96	2.57

If you are using a small dataset ($n \leq 30$) that is approximately normally distributed, use the t-distribution instead.

The t-distribution follows the same shape as the z-distribution, but corrects for small sample sizes. For the t-distribution, you need to know your degrees of freedom (sample size minus 1).

Check out this set of t tables to find your t-statistic. The author has included the confidence level and p-values for both one-tailed and two-tailed tests to help you find the t-value you need.

For normal distributions, like the t-distribution and z-distribution, the critical value is the same on either side of the mean.

Example: Critical value

In the TV-watching survey, there are more than 30 observations and the data follow an approximately normal distribution (bell curve), so we can use the z-distribution for our test statistics.

For a two-tailed 95% confidence interval, the alpha value is 0.025, and the corresponding critical value is 1.96.

This means that to calculate the upper and lower bounds of the confidence interval, we can take the mean ±1.96 standard deviations from the mean.

Finding the standard deviation

Most statistical software will have a built-in function to calculate your standard deviation, but to find it by hand you can first find your sample variance, then take the square root to get the standard deviation.

1. **Find the sample variance**

Sample variance is defined as the sum of squared differences from the mean, also known as the mean-squared-error (MSE):

$$S^2 = \sum_{i=1}^{n} \frac{(Xi - \overline{X})^2}{n-1}$$

To find the MSE, subtract your sample mean from each value in the dataset, square the resulting number, and divide that number by n − 1 (sample size minus 1).

Then add up all of these numbers to get your total sample variance (s2). For larger sample sets, it's easiest to do this in Excel.

2. **Find the standard deviation.**

The standard deviation of your estimate (s) is equal to the square root of the sample variance/sample error (S^2):

$$S = \sqrt{(S^2)}$$

Example: Standard deviation
In the television-watching survey, the variance in the GB estimate is 100, while the variance in the USA estimate is 25. Taking the square root of the variance gives us a sample standard deviation (s) of:

- 10 for the GB estimate.
- 5 for the USA estimate.

Sample size

The sample size is the number of observations in your data set.
Example: Sample size
In our survey of Americans and Brits, the sample size is 100 for each group.

Application

1. Random samples of size 225 are drawn from a population with mean 100 and standard deviation 20 . Find the mean and standard deviation of the sample mean.
2. There are 250 dogs at a dog show who weigh an average of 12 pounds, with a standard deviation of 8 pounds. If 4 dogs are chosen at random, what is the probability they have an average weight of greater than 8 pounds and less than 25 pounds?

KEY POINTS

- ✓ **Central Limit Theorem** states that if samples are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means approximates a normal distribution as the sample size increases beyond 30.
- ✓ **Sample Mean** is the mean only of the members of a sample or subset of a population.
- ✓ **Sampling distribution**. The probability distribution of a test statistic computed for each sample is called a sampling distribution.

- ✓ **Confidence intervals** is the interval within which you expect to capture a specific value. The confidence interval width is dependent on the confidence level.
- ✓ **Confidence level** is the probability value associated with a confidence interval.
- ✓ **Point estimate** of a population parameter is a single value used to estimate the population parameter.

END OF MODULE ASSESSMENT

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. On every module's "End of Module Assessment" (this part), write your answer/s on a one whole sheet of yellow pad paper.

LOOKING AHEAD

Congratulations for making it till the end of this module! The next topic will deal on Hypothesis Testing! Happy learning!

REFERENCES

https://www.ck12.org/book/ck-12-advanced-probability-and-statistics-concepts/

https://www.investopedia.com/terms/c/central_limit_theorem.asp#:~:text=Key%20Takeaways-,The%20central%20limit%20theorem%20(CLT)%20states%20that%20the%20distribution%20of,for%20the%20CLT%20to%20hold.

https://www.scribbr.com/statistics/confidence-interval/#:~:text=A%20confidence%20interval%20is%20the,another%20way%20to%20describe%20probability.

https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/06%3A_Sampling_Distributions/6.E%3A_Sampling_Distributions_(Exercises)

END OF MODULE ASSESSMENT (Answer Sheet)

(Please do not forget to provide information on this part)

Name: _____ Course& Year: _____
Module Number and Title: _____ Date accomplished: _____

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. Write your answers on a one whole sheet of yellow pad paper.

I. Explain the following questions

1. Does the mean of the sampling distribution equal the mean of the population?
2. Are there any restrictions on the size of the sample that is used to estimate the parameters of a population?

II. Problem Solving

1. In a local teaching district, a technology grant is available to teachers in order to install a cluster of four computers in their classrooms. From the 6,250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. Of those selected, 142 teachers felt that computers were an essential teaching tool.
   A. Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential teaching tool.
   B. How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?
2. A random sample of 16 light bulbs has a mean life of 650 hours and a standard deviation of 32 hours. Assume the population has a normal distribution. Construct a 90% confidence interval for the population mean.

Hypothesis Testing

Module 8

OVERVIEW

In this module we will explore hypothesis testing, which involves making conjectures about a population based on a sample drawn from the population. Hypothesis tests are often used in statistics to analyze the likelihood that a population has certain characteristics. For example, we can use hypothesis testing to analyse if a senior class has a particular average SAT score or if a prescription drug has a certain proportion of the active ingredient.

A hypothesis is simply a conjecture about a characteristic or set of facts. When performing statistical analyses, our hypotheses provide the general framework of what we are testing and how to perform the test.

These tests are never certain and we can never prove or disprove hypotheses with statistics, but the outcomes of these tests provide information that either helps support or refute the hypothesis itself.

LEARNING OUTCOMES

By the end of this module, you should be able to:

- Develop null and alternative hypotheses to test for a given situation.
- Understand the critical regions of a graph for one- and two-tailed hypothesis tests.
- Calculate a test statistic to evaluate a hypothesis.
- Test a hypothesis about a population proportion by applying the binomial distribution approximation.
- Test a hypothesis about a mean.
- Understand how the shape of Student's t−distribution corresponds to the sample size (which corresponds to a measure called the "degrees of freedom.")
- Identify situations that contain dependent or independent samples.
- Calculate the pooled standard deviation for two independent samples.
- Calculate the test statistic to test hypotheses about dependent and independent data pairs for both large and small samples.
- Calculate the test statistic to test hypotheses about the difference of proportions between two independent samples.

LEARNING EXPERIENCES & SELF-ASSESSMENT ACTIVITIES (SAA)

Activity

Suppose the present success rate in treating a certain type of lung cancer is .75. A research group hopes to demonstrate that the success rate of a new treatment of this cancer is better. Write the null and alternative hypotheses.

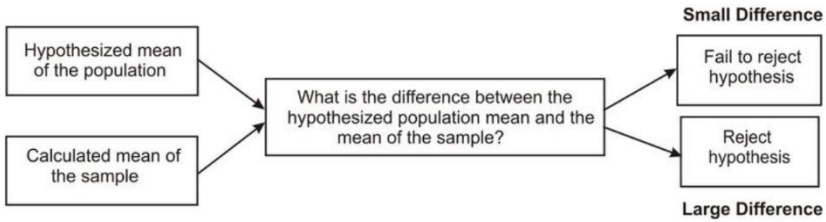
Analysis

1. How did you know the Null Hypothesis of the given statement above?
2. Base on the statement above, how did you write the alternative hypothesis?

Abstraction

Null and Alternative Hypotheses

Hypothesis testing involves testing the difference between a hypothesized value of a population parameter and the estimate of that parameter which is calculated from a sample. If the parameter of interest is the mean of the populations in hypothesis testing, we are essentially determining the magnitude of the difference between the mean of the sample and they hypothesized mean of the population. If the difference is very large, we reject our hypothesis about the population. If the difference is very small, we do not. Below is an overview of this process.



In statistics, the hypothesis to be tested is called the null hypothesis and given the symbol H_o. The alternative hypothesis is given the symbol H_a

The null hypothesis defines a specific value of the population parameter that is of interest. Therefore, the null hypothesis always includes the possibility of equality. Consider

$$H_0: \mu = 3.2$$
$$H_a: \mu \neq 3.2$$

In this situation if our sample mean,$\bar{x}$, is very different from 3.2 we would reject H_o. That is, we would reject H_o if $\bar{x}$ is much larger than 3.2 or much smaller than 3.2. This is called a 2-tailed test. An $\bar{x}$ that is very unlikely if H_o is true is considered to be good evidence that the claim H_o is not true. Consider H_o:$\mu \leq 3.2$ H_a: $\mu > 3.2$. In this situation we would reject H_o for very large values of $\bar{x}$. This is called a one tail test. If, for this test, our data gives $\bar{x} = 15$, it would be highly unlikely that finding $\bar{x}$ this different from 3.2 would occur by chance and so we would probably reject the null hypothesis in favor of the alternative hypothesis.

Testing the Null Hypothesis

If we were to test the hypothesis that the seniors had a mean SAT score of 1100 our null hypothesis would be that the SAT score would be equal to 1100 or:

$$H_o: \mu = 1100$$

We test the null hypothesis against an alternative hypothesis, which is given the symbol H_a and includes the outcomes not covered by the null hypothesis. Basically, the alternative hypothesis states that there is a difference between the hypothesized population mean and the sample mean. The alternative hypothesis can be supported only by rejecting the null hypothesis. In our example above about the SAT scores of graduating seniors, our alternative hypothesis would state that there is a difference between the null and alternative hypotheses or:

$$H_a: \mu \neq 1100$$

Testing Hypotheses: P-Values

Deciding Whether to Reject the Null Hypothesis: One-Tailed and Two-Tailed Hypothesis Tests

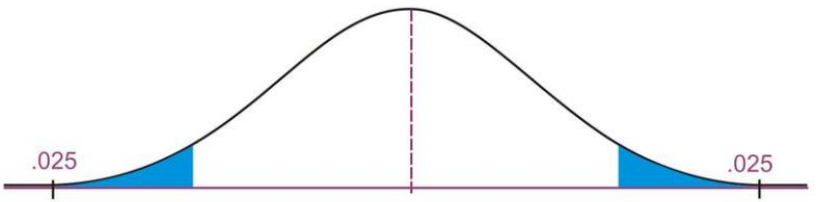
When a hypothesis is tested, a statistician must decide on how much evidence is necessary in order to reject the null hypothesis. For example, if the null hypothesis is that the average height of a population is 64 inches a statistician wouldn't measure one person who is 66 inches and reject the hypothesis based on that one trial. It is too likely that the discrepancy was merely due to chance.

We use statistical tests to determine if the sample data give good evidence against the claim (H_o). The numerical measure that we use to determine the strength of the sample evidence we are willing to consider strong enough to reject H_o is called the level of significance and it is denoted by α. If we choose, for example, $\alpha = .01$ we are saying that we would get data at least as unusual as the data we have collected no more than 1% of the time when H_o is true.

The most frequently used levels of significance are 0.05 and 0.01. If our data results in a statistic that falls within the region determined by the level of

significance then we reject H_o. The region is therefore called the **critical region**. When choosing the level of significance, we need to consider the consequences of rejecting or failing to reject the null hypothesis. If there is the potential for health consequences (as in the case of active ingredients in prescription medications) or great cost (as in the case of manufacturing machine parts), we should use a more 'conservative' critical region with levels of significance such as $.005$ or $.001$.

When determining the critical regions for a two-tailed hypothesis test, the level of significance represents the extreme areas under the normal density curve. We call this a two-tailed hypothesis test because the critical region is located in both ends of the distribution. For example, if there was a significance level of 0.95 the critical region would be the most extreme 5 percent under the curve with 2.5 percent on each tail of the distribution.



Visual Representation of Critical Regions

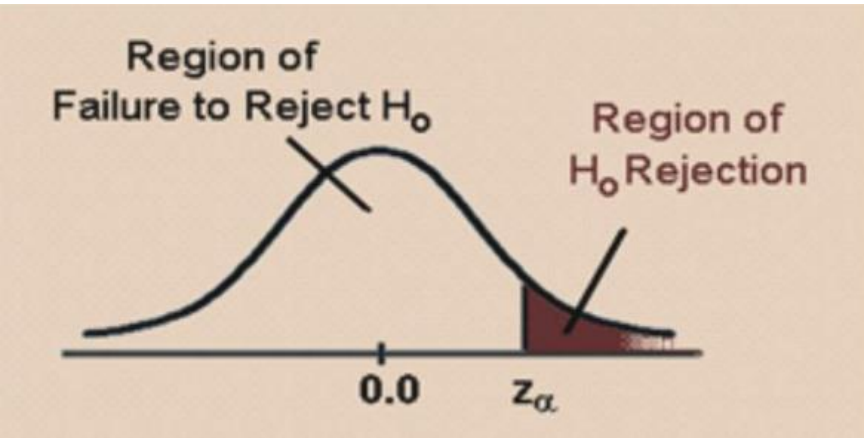
Therefore, if the mean from the sample taken from the population falls within one of these critical regions, we would conclude that there was too much of a difference between our sample mean and the hypothesized population mean and we would reject the null hypothesis. However, if the mean from the sample falls in the middle of the distribution (in between the critical regions) we would fail to reject the null hypothesis.

We calculate the critical region for the single-tail hypothesis test a bit differently. We would use a single-tail hypothesis test when the direction of the results is anticipated or we are only interested in one direction of the results. For example, a single-tail hypothesis test may be used when evaluating whether or not to adopt a new textbook. We would only decide to adopt the textbook if it improved student achievement relative to the old textbook. A single-tail hypothesis simply states that the mean is greater or less than the hypothesized value.

When performing a single-tail hypothesis test, our alternative hypothesis looks a bit different. When developing the alternative hypothesis in a single-tail hypothesis test we would use the symbols of greater than or less than. Using our example about SAT scores of graduating seniors, our null and alternative hypothesis could look something like:

$$H_o: \mu = 1100$$
$$H_a: \mu > 1100$$

In this scenario, our null hypothesis states that the mean SAT scores would be equal to 1100 while the alternate hypothesis states that the SAT scores would be greater than 1100. A single-tail hypothesis test also means that we have only one critical region because we put the entire region of rejection into just one side of the distribution. When the alternative hypothesis is that the sample mean is greater, the critical region is on the right side of the distribution. When the alternative hypothesis is that the sample is smaller, the critical region is on the left side of the distribution (see below).



To calculate the critical regions, we must first find the critical values or the cut-offs where the critical regions start. To find these values, we use the critical values found specified by the z–distribution. These values can be found in a table that lists the areas of each of the tails under a normal distribution. Using this table, we find that for a 0.05 significance level, our critical values would fall at 1.96 standard errors above and below the mean. For a 0.01 significance level, our critical values would fall at 2.57 standard errors above and below the mean. Using the z–distribution we can find critical values (as specified by standard z scores) for any level of significance for either single-or two-tailed hypothesis tests.

Determining Critical Values

Determine the critical value for a single-tailed hypothesis test with a 0.05 significance level.

Using the z distribution table, we find that a significance level of 0.05 corresponds with a critical value of 1.645. If alternative hypothesis is the mean is greater than a specified value the critical value would be 1.645. Due to the symmetry of the

normal distribution, if the alternative hypothesis is the mean is less than a specified value the critical value would be -1.645.

Calculating the Test Statistic

Before evaluating our hypotheses by determining the critical region and calculating the test statistic, we need confirm that the distribution is normal and determine the hypothesized mean μ of the distribution.

To evaluate the sample mean against the hypothesized population mean, we use the concept of z–scores to determine how different the two means are from each other. Based on the **Central Limit theorem** the distribution of $\bar{x}$ is normal with mean, μ and standard deviation, $\frac{\sigma}{\sqrt{n}}$. As we learned in previous lessons, the z score is calculated by using the formula:

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

Where:

z= standardized score

$\bar{x}$= sample mean

μ= the population mean under the null hypothesis

σ= population standard deviation. If we do not have the population standard deviation and if $n \geq 30$, we can use the sample standard deviation, s. If $n < 30$ and we do not have the population sample standard deviation we use a different distribution which will be discussed in a future lesson.

Once we calculate the z score, we can make a decision about whether to reject or to fail to reject the null hypothesis based on the critical values.

Following are the steps you must take when doing an hypothesis test:

1. Determine the null and alternative hypotheses.
2. Verify that necessary conditions are satisfied and summarize the data into a test statistic.
3. Determine the α level.
4. Determine the critical region(s).
5. Make a decision (Reject or fail to reject the null hypothesis)
6. Interpret the decision in the context of the problem.

1) College A has an average SAT score of 1500. From a random sample of 125 freshman psychology students we find the average SAT score to be 1450 with a standard deviation of 100. We want to know if these freshman psychology students are representative of the overall population. What are our hypotheses and the test statistic?

a. Let's first develop our null and alternative hypotheses:

$$H_o: \mu = 1500$$
$$H_a: \mu \neq 1500$$

b. The test statistic is $z = \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{1450-1500}{\frac{100}{\sqrt{125}}} \approx -5.59$

c. Choose $\alpha = .05$

d. This is a two sided test. If we choose $\alpha = .05$, the critical values will be -1.96 and 1.96. (Use invNorm $(.025, 0, 1)$ and the symmetry of the normal distribution to determine these critical values) That is we will reject the null hypothesis if the value of our test statistic is less than -1.96 or greater than 1.96.

e. The value of the test statistic is -5.59. This is less than -1.96 and so our decision is to reject H_o.

f. Based on this sample we believe that the mean is not equal to 1500.

Finding the P-Value of an Event

We can also evaluate a hypothesis by asking "what is the probability of obtaining the value of the test statistic we did if the null hypothesis is true?" This is called the p−value.

The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the p-value, the more likely you are to reject the null hypothesis.

Type I and Type II Errors

When we decide to reject or not reject the null hypothesis, we have four possible scenarios:

- The null hypothesis is true and we reject it.
- The null hypothesis is true and we do not reject it.
- The null hypothesis is false and we do not reject it.
- The null hypothesis is false and we reject it.

Two of these four possible scenarios lead to correct decisions: accepting the null hypothesis when it is true and rejections the null hypothesis when it is false.

Two of these four possible scenarios lead to errors: rejecting the null hypothesis when it is true and accepting the null hypothesis when it is false.

Which type of error is more serious depends on the specific research situation, but ideally both types of errors should be minimized during the analysis.

Below is a table outlining the possible outcomes in hypothesis testing:

	H_0 is true	H_0 is false
Accept H_0	Good Decision	Error (type II)
Reject H_0	Error (type I)	Good Decision

The general approach to hypothesis testing focuses on the Type I error: rejecting the null hypothesis when it may be true. The level of significance, also known as the alpha level, is defined as the probability of making a Type I error when testing a null hypothesis. For example, at the 0.05 level, we know that the decision to reject the hypothesis may be incorrect 5 percent of the time.

$$\alpha = P(rejecting\ H0|H0\ is\ true) = P(making\ a\ type\ I\ error)$$

Calculating the probability of making a Type II error is not as straightforward as calculating the probability of making a Type I error. The probability of making a Type II error can only be determined when values have been specified for the alternative hypothesis. The probability of making a type II error is denoted by β.

$$\beta = P(accepting\ H0|H0\ is\ false) = P(making\ a\ type\ II\ error)$$

Once the value for the alternative hypothesis has been specified, it is possible to determine the probability of making a correct decision $(1-\beta)$. This quantity, $1-\beta$, is called the power of the test.

The goal in hypothesis testing is to minimize the potential of both Type I and Type II errors. However, there is a relationship between these two types of errors. As the level of significance or alpha level increases, the probability of making a Type II error (β) decreases and vice versa.

Often we establish the alpha level based on the severity of the consequences of making a Type I error. If the consequences are not that serious, we could set an alpha level at 0.10 or 0.20. However, in a field like medical research we would set the alpha level very low (at 0.001 for example) if there was potential bodily harm to patients. We can also attempt minimize the Type II errors by setting higher alpha levels in situations that do not have grave or costly consequences.

Calculating the Power of a Test

The power of a test is defined as the probability of rejecting the null hypothesis when it is false (that is, making the correct decision). Obviously, we want to maximize this power if we are concerned about making Type II errors. To determine the power of the test, there must be a specified value for the alternative hypothesis.

Suppose that a doctor is concerned about making a Type II error only if the active ingredient in the new medication is greater than 3 milligrams higher than what was specified in the null hypothesis (say, 250 milligrams with a sample of 200 and a standard deviation of 50). Now we have values for both the null and the alternative hypotheses.

$$H_o: \mu = 250$$
$$H_a: \mu = 253$$

By specifying a value for the alternative hypothesis, we have selected one of the many values forH_a. In determining the power of the test, we must assume that $H_a$is true and determine whether we would correctly reject the null hypothesis

Calculating the exact value for the power of the test requires determining the area above the critical value set up to test the null hypothesis when it is re centered around the alternative hypothesis. If we have an alpha level of $.05$ our critical value would be 1.645 for the one tailed test. Therefore,

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

$$1.645 = \frac{(\bar{x} - 250)}{\frac{50}{\sqrt{200}}}$$

Solving for $\bar{x}$ we find: $\bar{x} = 1.645\left(\frac{50}{\sqrt{200}}\right) + 250 \approx 255.8$

Now, with a new mean set at the alternative hypothesis $H_a: \mu = 253$ we want to find the value of the critical score when centered around this score when we center this $\bar{x}$ around the population mean of the alternative hypothesis, $\mu = 253$. Therefore, we can figure that:

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(255.8 - 253)}{\frac{50}{\sqrt{200}}} \approx 0.79$$

Recall that we reject the null hypothesis if the critical value is to the right of .79. The question now is what is the probability of rejecting the null hypothesis when, in fact, the alternative hypothesis is true? We need to find the area to the right of 0.79. You can find this area using a z table. The probability is .2148. This means that since we assumed the alternative hypothesis to be true, there is only a 21.5% chance of rejecting the null hypothesis. Thus, the power of the test is .2148. In other words, this test of the null hypothesis is not very powerful and has only a 0.2148 probability of detecting the real difference between the two hypothesized means.

There are several things that affect the power of a test including:

- Whether the alternative hypothesis is a single-tailed or two-tailed test.
- The level of significance α.
- The sample size.

Hypothesis Testing about Population Proportions by Applying the Binomial Distribution Approximation

We could perform tests of population proportions to answer the following questions:

- What percentage of graduating seniors will attend a 4-year college?
- What proportion of voters will vote for John McCain?
- What percentage of people will choose Diet Pepsi over Diet Coke?

To test questions like these, we make hypotheses about population proportions. For example,

- H_o:35% of graduating seniors will attend a 4-year college.
- H_o:42% of voters will vote for John McCain.
- H_o:26% of people will choose Diet Pepsi over Diet Coke.

To test these hypotheses we follow a series of steps:

- Hypothesize a value for the population proportion P like we did above.
- Randomly select a sample.
- Use the sample proportion $\hat{p}$ to test the stated hypothesis

To determine the test statistic we need to know the sampling distribution of the sample proportion. We use the binomial distribution which illustrates situations in which two outcomes are possible (for example, voted for a candidate, didn't vote for a candidate), remembering that when the sample size is relatively large, we can use the normal distribution to approximate the binomial distribution. The test statistic is

$$z = \frac{sample\ estimate - value\ under\ the\ null\ hypothesis}{standard\ error\ under\ the\ null\ hypothesis}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

p_0 is the hypothesized value of the proportion under the null hypothesis
n is the sample size

Determining Hypotheses and Test Statistics

We want to test a hypothesis that 60 percent of the 400 seniors graduating from a certain California high school will enroll in a two or four-year college upon graduation. What would be our hypotheses and the test statistic?

Since we want to test the proportion of graduating seniors and we think that proportion is around 60 percent, our hypotheses are:

$$H_o: p = .6$$
$$H_o: p = .6$$

The test statistics would be $z = \frac{\hat{p}-.6}{\sqrt{\frac{.6(1-.6)}{n}}}$ to complete this calculation we would have to have a value for the sample size (n).

Testing a Proportion Hypothesis

Similar to testing hypotheses dealing with population means, we use a similar set of steps when testing proportion hypotheses.

- Determine and state the null and alternative hypotheses.
- Set the criterion for rejecting the null hypothesis.
- Calculate the test statistic.
- Decide whether to reject or fail to reject the null hypothesis.
- Interpret your decision within the context of the problem.

Significance Testing for Means

Evaluating Hypotheses for Population Means using Large Samples

When testing a hypothesis for the mean of a normal distribution, we follow a series of six basic steps:

1. State the null and alternative hypotheses.
2. Choose an α level
3. Set the criterion (critical values) for rejecting the null hypothesis.
4. Compute the test statistic.
5. Make a decision (reject or fail to reject the null hypothesis)
6. Interpret the result

If we reject the null hypothesis we are saying that the difference between the observed sample mean and the hypothesized population mean is too great to be attributed to chance. When we fail to reject the null hypothesis, we are saying that the difference between the observed sample mean and the hypothesized population mean is probable if the null hypothesis is true. Essentially, we are willing to attribute this difference to sampling error.

Example:

The school nurse was wondering if the average height of 7th graders has been increasing. Over the last 5 years, the average height of a 7th grader was 145 cm with a standard deviation of 20 cm. The school nurse takes a random sample of 200 students and finds that the average height this year is 147 cm. Conduct a single-tailed hypothesis test using a .05 significance level to evaluate the null and alternative hypotheses.

First, we develop our null and alternative hypotheses:

$$H_o: \mu = 145$$
$$H_a: \mu > 145$$

Choose $\alpha = .05$. The critical value for this one tailed test is 1.64. Any test statistic greater than 1.64 will be in the rejection region.

Next, we calculate the test statistic for the sample of $7th$ graders.

$$z = \frac{147 - 145}{\frac{20}{\sqrt{200}}} \approx 1.414$$

Since the calculated $z-$score of 1.414 is smaller than 1.64 and thus does not fall in the critical region. Our decision is to fail to reject the null hypothesis and conclude that the probability of obtaining a sample mean equal to 147 if the mean of the population is 145 is likely to have been due to chance.

Testing a Mean Hypothesis Using P-values

We can also test a mean hypothesis using p-values. The following examples show how to do this.

A sample of size 157 is taken from a normal distribution, with a standard deviation of 9. The sample mean is 65.12. Use the 0.01 significance level to test the claim that the population mean is greater than 65.

We always put equality in the null hypothesis, so our claim will be in the alternative hypothesis.

$$H_o: \mu = 65$$
$$H_a: \mu > 65$$

The test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{65.12 - 65}{9/\sqrt{157}} \approx \frac{0.12}{0.718} = 167$$

Now we will find the probability of observing a test statistic at least this extreme when assuming the null hypothesis. Since our alternative hypothesis is that the mean is greater, we want to find the probability of z scores that are greater than our test statistics. The p-value we are looking for is:

p-value=$P(z > 0.17) = 1 - P(z < 0.17)$

Using a z-score table:
p-value
$= P(z > 0.0.167) = 1 - P(z < 0.167) = 1 - 0.6064 = 0.3936 > 0.01$

The probability of observing a test statistic at least as big as the z=0.17 is 0.3936. Since this is greater than our significance level, 0.01, we fail to reject the null hypothesis. This means that the data does not support the claim that the mean is greater than 65.

Testing a Mean Hypothesis When the Population Standard Deviation is Known

We can also use the standard normal distribution, or z-scores, to test a mean hypothesis when the population standard deviation is known. The next two examples, though they have a smaller sample size, have a known population standard deviation.

1. A sample of size 50 is taken from a normal distribution, with a known population standard deviation of 26. The sample mean is 167.02. Use the 0.05 significance level to test the claim that the population mean is greater than 170.

We always put equality in the null hypothesis, so our claim will be in the alternative hypothesis.

$$H_0: \mu = 170$$
$$H_a: \mu > 17$$

The test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{167.02 - 170}{26/\sqrt{50}} \approx \frac{2.98}{3.67} = 0.18$$

Now we will find the probability of observing a test statistic at least this extreme when assuming the null hypothesis. Since our alternative hypothesis is that the mean is greater, we want to find the probability of z scores that are greater than our test statistics. The p-value we are looking for is:

p-value

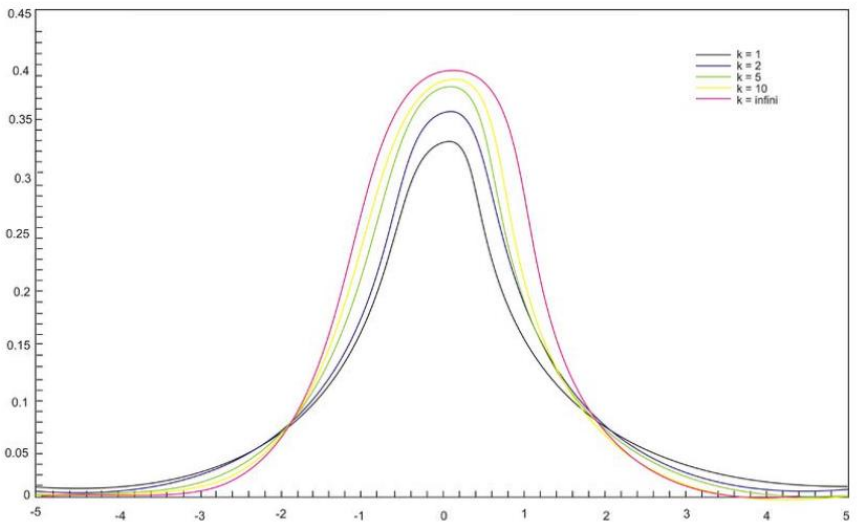
$$= P(z > 0.811) = 1 - P(z < 0.811) = 1 - 0.791 = 0.209 > 0.05$$

The probability of observing a test statistic at least as big as the z=0.81 is 0.209. Since this is greater than our significance level, 0.05, we fail to reject the null hypothesis. This means that the data does not support the claim that the mean is greater than 170.

Student's t-Distribution

Hypothesis Testing with Small Populations and Sample Sizes

T−distributions are a family of distributions that, like the normal distribution, are symmetrical and bell-shaped and centered on a mean. However, the distribution shape changes as the sample size changes. Therefore, there is a specific shape or distribution for every sample of a given size (see figure below; each distribution has a different value of k, the number of degrees of freedom, which is 1 less than the size of the sample).



We use the Student's t−distribution in hypothesis testing the same way that we use the normal distribution. Each row in the t distribution table (see link below) represents a different t−distribution and each distribution is associated with a unique number of degrees of freedom (the number of observations minus one). The column headings in the table represent the portion of the area in the tails of the distribution – we use the numbers in the table just as we used the z−scores.

As the number of observations gets larger, the t−distribution approaches the shape of the normal distribution. In general, once the sample size is large enough - usually about 30 - we would use the normal distribution or the z−table instead. Note that usually in practice, if the standard deviation is known then the normal distribution is used regardless of the sample size.

In calculating the t−test statistic, we use the formula:

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

where:

t is the test statistic and has n−1 degrees of freedom.
$\bar{x}$ is the sample mean
μ_0 is the population mean under the null hypothesis.
s is the sample standard deviation
n is the sample size
$\frac{s}{\sqrt{n}}$ is the estimated standard error

Significance Testing

The high school athletic director is asked if football players are doing as well academically as the other student athletes. We know from a previous study that the average GPA for the student athletes is 3.10. After an initiative to help improve the GPA of student athletes, the athletic director samples 20 football players and finds that the average GPA of the sample is 3.18 with a sample standard deviation of 0.54. Is there a significant improvement? Use a .05 significance level.

First, we establish our null and alternative hypotheses.

$$H_o: \mu = 3.10$$
$$H_a: \mu \neq 3.10$$

Next, we use our alpha level of .05 and the t−distribution table to find our critical values. For a two-tailed test with 19 degrees of freedom and a .05 level of significance, our critical values are equal to ±2.093.

In calculating the test statistic, we use the formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.18 - 3.10}{\frac{.54}{\sqrt{20}}} \approx 0.66$$

This means that the observed sample mean 3.18 of football players is .66 standard errors above the hypothesized value of 3.10. Because the value of the test statistic is less than the critical value of 2.093, we fail to reject the null hypothesis.

Therefore, we can conclude that the difference between the sample mean and the hypothesized value is not sufficient to attribute it to anything other than sampling error. Thus, the athletic director can conclude that the mean academic performance of football players does not differ from the mean performance of other student athletes.

Testing a Hypothesis for Dependent and Independent Samples

We have learned about hypothesis testing for proportion and means with both large and small samples. However, in the examples in those lessons only one sample was involved. In this lesson we will apply the principals of hypothesis testing to situations involving two samples. There are many situations in everyday life where we would perform statistical analysis involving two samples. For example, suppose that we wanted to test a hypothesis about the effect of two medications on curing an illness. Or we may want to test the difference between the means of males and females on the SAT. In both of these cases, we would analyze both samples and the hypothesis would address the difference between two sample means.

In this Concept, we will identify situations with different types of samples, learn to calculate the test statistic, calculate the estimate for population variance for both samples and calculate the test statistic to test hypotheses about the difference of proportions or means between samples.

Dependent and Independent Samples

When we are working with one sample, we know that we need to select a random sample from the population, measure that sample statistic and then make hypothesis about the population based on that sample. When we work with two independent samples we assume that if the samples are selected at random (or, in the case of medical research, the subjects are randomly assigned to a group), the two samples will vary only by chance and the difference will not be statistically significant. In short, when we have independent samples we assume that the scores of one sample do not affect the other.

Independent samples can occur in two scenarios.

Testing the difference of the means between two fixed populations we test the differences between samples from each population. When both samples are randomly selected, we can make inferences about the populations.

When working with subjects (people, pets, etc.), if we select a random sample and then randomly assign half of the subjects to one group and half to another we can make inferences about the populations.

Dependent samples are a bit different. Two samples of data are dependent when each score in one sample is paired with a specific score in the other sample. In short, these types of samples are related to each other. Dependent samples can occur in two scenarios. In one, a group may be measured twice such as in a pretest-posttest situation (scores on a test before and after the lesson). The other scenario is one in which an observation in one sample is matched with an observation in the second sample.

To distinguish between tests of hypotheses for independent and dependent samples, we use a different symbol for hypotheses with dependent samples. For dependent sample hypotheses, we use the delta symbol δ to symbolize the difference between the two samples. Therefore, in our null hypothesis we state that the difference of scores across the two measurements is equal to 0;δ=0 or:

$$H_o: \delta = \mu_1 - \mu_2$$

Calculating the Pooled Estimate of Population Variance

When testing a hypothesis about two independent samples, we follow a similar process as when testing one random sample. However, when computing the test statistic, we need to calculate the estimated standard error of the difference between sample means,

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

Where $n_1$and $n_2$are the sizes of the two samples $s^2$is the pooled sample variance, which is computed as $s = \frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n_1+n_2-2}$ Often, the top part of this formula is simplified by substituting the symbol SS for the sum of the squared deviations. Therefore, the formula often is expressed by $s^2 = \frac{SS_1+SS_2}{n_1+n_2-2}$.

Calculating s^2 Suppose we have two independent samples of student reading scores.

The data are as follows:

Sample 1	Sample 2
7	12
8	14
10	18
4	13
6	11
	10

From this sample, we can calculate a number of descriptive statistics that will help us solve for the pooled estimate of variance:

Descriptive Statistic	Sample 1	Sample 2
Number n	5	6
Sum of Observations $\sum x$	35	78
Mean of Observations $\bar{x}$	7	13
Sum of Squared Deviations $\sum_{i=1}^{n}(x_i - \bar{x})^2$	20	40

Using the formula for the pooled estimate of variance, we find that

$$s^2 = 6.67$$

We will use this information to calculate the test statistic needed to evaluate the hypotheses.

Testing Hypotheses with Independent Samples

When testing hypotheses with two independent samples, we follow similar steps as when testing one random sample:

- State the null and alternative hypotheses.
- Choose α
- Set the criterion (critical values) for rejecting the null hypothesis.
- Compute the test statistic.
- Make a decision: reject or fail to reject the null hypothesis.
- Interpret the decision within the context of the problem.

When stating the null hypothesis, we assume there is no difference between the means of the two independent samples. Therefore, our null hypothesis in this case would be:

$$H_0: \mu_1 = \mu_2 \; or \; H_0: \mu_1 - \mu_2 = 0$$

Similar to the one-sample test, the critical values that we set to evaluate these hypotheses depend on our alpha level and our decision regarding the null hypothesis is carried out in the same manner. However, since we have two samples, we calculate the test statistic a bit differently and use the formula:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s.e(\bar{x}_1 - \bar{x}_2)}$$

where:

$\bar{x}_1 - \bar{x}_2$ is the difference between the sample means

$\mu_1 - \mu_2$ is the difference between the hypothesized population means

$s.e(\bar{x}_1 - \bar{x}_2)$ is the standard error of the difference between sample means

Evaluating the Difference Between Two Samples

The head of the English department is interested in the difference in writing scores between remedial freshman English students who are taught by different teachers. The incoming freshmen needing remedial services are randomly assigned to one of two English teachers and are given a standardized writing test after the first semester. We take a sample of eight students from one class and nine from the other. Is there a difference in achievement on the writing test between the two classes? Use a 0.05 significance level.

First, we would generate our hypotheses based on the two samples.

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

This is a two tailed test. For this example, we have two independent samples from the population and have a total of 17 students that we are examining. Since our sample size is so low, we use the $t-$distribution. In this example, we have 15 degrees of freedom (number in the samples minus 2) and with a .05 significance level and the t distribution, we find that our critical values are 2.131 standard scores above and below the mean.

To calculate the test statistic, we first need to find the pooled estimate of variance from our sample. The data from the two groups are as follows:

Sample 1	Sample 2
35	52
51	87
66	76
42	62
37	81
46	71
60	55
55	67
53	

From this sample, we can calculate several descriptive statistics that will help us solve for the pooled estimate of variance:

Descriptive Statistic	Sample 1	Sample 2
Number n	9	8
Sum of Observations $\sum x$	445	551
Mean of Observations $\bar{x}$	49.44	68.875
Sum of Squared Deviations $\sum_{i=1}^{n}(x_i - \bar{x})^2$	862.22	1058.88

Therefore:

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = 128.07$$

and the standard error of the difference of the sample means is:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{128.07\left(\frac{1}{9} + \frac{1}{8}\right)} \approx 5.50$$

Since -3.53 is less than the critical value of 2.13, we decide to reject the null hypothesis and conclude there is a significant difference in the achievement of the students assigned to different teachers.

Application

Determine whether each of the following is a null or an alternative hypothesis.

1. The average weight of golden retriever dogs is the same as the average weight of pit bull dogs.
2. The proportion of books in the library that are novels is higher than the proportion of books in the library that are nonfiction.
3. The average price of wool coats in San Francisco is lower in the summer than in the winter.

KEY POINTS

- ✓ Alternative hypothesis, H1, is the clear and concise statement of the initial claim.
- ✓ Hypothesis is the tentative explanation to a scientific question that can be tested.
- ✓ Hypothesis testing involves testing the difference between a hypothesized value of a population parameter and the estimate of that parameter which is calculated from a sample.
- ✓ The null hypothesis, H_o, is the opposite of what you are hoping to claim.
- ✓ A confidence interval is the interval within which you expect to capture a specific value. The confidence interval width is dependent on the confidence level.
- ✓ Critical regions (or rejection regions) describe the entire area of values that indicate you reject the null hypothesis.
- ✓ Critical values are the values that indicate the edge of the critical region
- ✓ The numerical measure that we use to determine the strength of the sample evidence we are willing to consider strong enough to reject H_o is called the level of significance and it is denoted by alpha.
- ✓ A p-value is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that H_o is true.
- ✓ The power of a test is defined as the probability of rejecting the null hypothesis when it is false (that is, making the correct decision).
- ✓ A single-tail hypothesis test also means that we have only one critical region because we put the entire region of rejection into just one side of the distribution.
- ✓ A two-tailed test has tails on both ends of the normal distribution curve. The null hypothesis will be in the form $H_o = 5$, for example.
- ✓ A Type I error occurs when a true null hypothesis is rejected.
- ✓ A Type II error occurs when a false null hypothesis is not rejected.
- ✓ The z -score of a value is the number of standard deviations between the value and the mean of the set.
- ✓ The hypothesized population proportion is given the symbol p_o in the z-score calculation.
- ✓ A test statistic is a single measure of some attribute of a sample (i.e. a statistic) used in statistical hypothesis testing. The z-score is one kind of test statistic that is used to determine the probability of obtaining a given value.
- ✓ Degrees of freedom are essentially the number of samples that have the 'freedom' to change without necessarily affecting the sample mean. Degrees of freedom has the formula $df = n - 1$.

- ✓ A Student's t-distribution is a distribution similar to the normal distribution, with slightly greater spread and thicker tails. It is commonly used in the calculation of confidence intervals when $n < 30$.
- ✓ Dependent samples occur when you have two samples that do affect one another.
- ✓ Independent samples occur when you have two samples that do not affect one another.
- ✓ The likelihood is the test statistic (t) associated with two dependent samples.
- ✓ The z-score is the test statistic associated with two independent samples used when testing the proportion associated with two independent samples.

END OF MODULE ASSESSMENT

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. On every module's "End of Module Assessment" (this part), write your answers on a one whole sheet of yellow pad paper.

LOOKING AHEAD

Congratulations for making it till the end of this module! The next topic will deal on Regression and Correlation! Happy learning!

REFERENCES

https://www.ck12.org/book/ck-12-advanced-probability-and-statistics-concepts/section/2.0/

https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/box-plot/

https://www.scribbr.com/statistics/p-value/#:~:text=The%20p%2Dvalue%20is%20a,to%20reject%20the%20null%20hypothesis.

END OF MODULE ASSESSMENT

(Please do not forget to provide information on this part)

Name: _____ Course& Year: _____
Module Number and Title: _____ Date accomplished: _____

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. Write your answers on the blank space provided for each question.

Task 1. Write your answer on a one whole sheet of yellow paper.

1. In hypothesis testing, we have scenarios that have both dependent and independent samples. Give an example of an experiment with (1) dependent samples and (2) independent samples.
2. True or False: When we test the difference between the means of males and females on the SAT, we are using independent samples.
3. True or False: When we fail to reject the null hypothesis, we are saying that the difference between the observed sample mean and the hypothesized population mean is probable if the null hypothesis is true.

Task 2. For each of the following scenarios, state which one is more likely to lead to the rejection of the null hypothesis?

1. A one-tailed or two-tailed test
2. .05 or .01 level of significance
3. A sample size of n=144 or n=444

Regression and Correlation

Module 9

OVERVIEW

In previous Module, you have learned about descriptive statistics, probabilities, and distributions of random variables. In this Module, you will learn how to recognize whether two variables have a relationship; where one variable can accurately predict the values of another variable. This relationship is called correlation, and in this Module you will learn about linear correlation and writing equations of regression lines, which can be used to predict values of one variable based upon an explanatory variable.

LEARNING OUTCOMES

By the end of this module, you should be able to:

- Understand the concepts of bivariate data and correlation, and the use of scatterplots to display bivariate data.
- Calculate the linear correlation coefficient and coefficient of determination of bivariate data, using technology tools to assist in the calculations.
- Predict values using bivariate data plotted on a scatterplot.

LEARNING EXPERIENCES & SELF-ASSESSMENT ACTIVITIES (SAA)

Abstraction

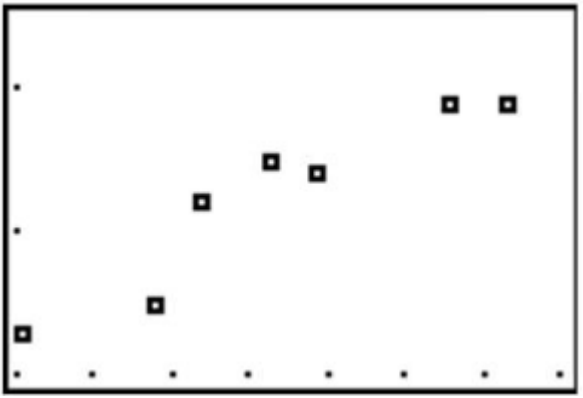
Scatter Plots and Linear Correlation

Linear Correlation

Correlation measures the relationship between bivariate data. Bivariate data are data sets in which each subject has two observations associated with it.

Scatterplots display these bivariate data sets and provide a visual representation of the relationship between variables. In a scatterplot, each point represents a

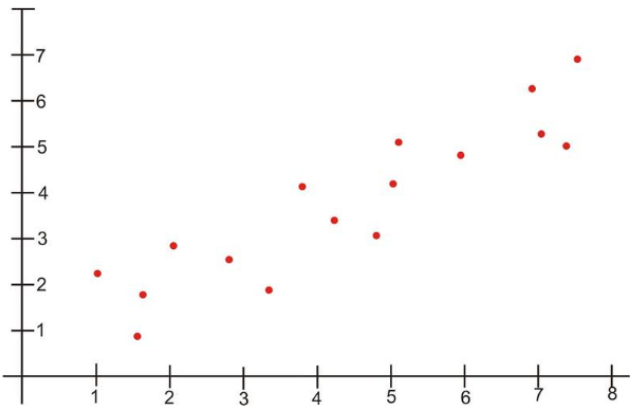
paired measurement of two variables for a specific subject, and each subject is represented by one point on the scatterplot.



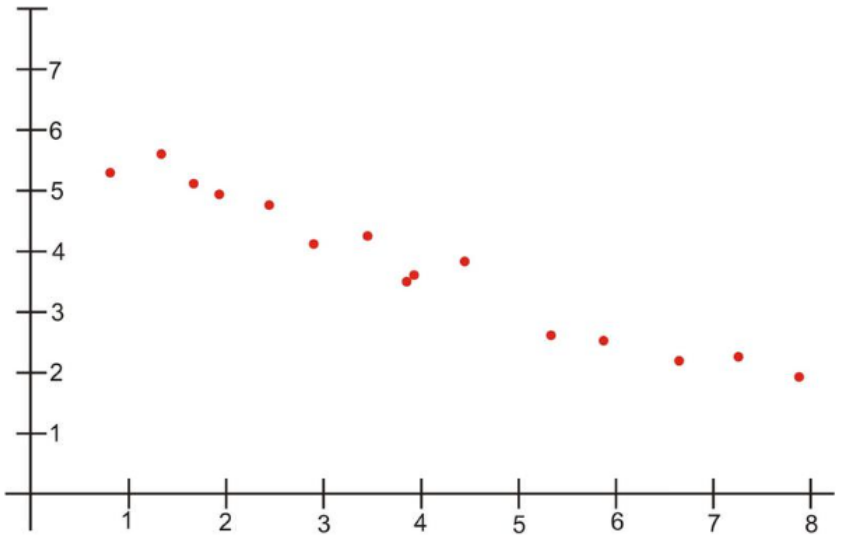
Correlation Patterns in Scatterplot Graphs

Examining a scatterplot graph allows us to obtain some idea about the relationship between two variables.

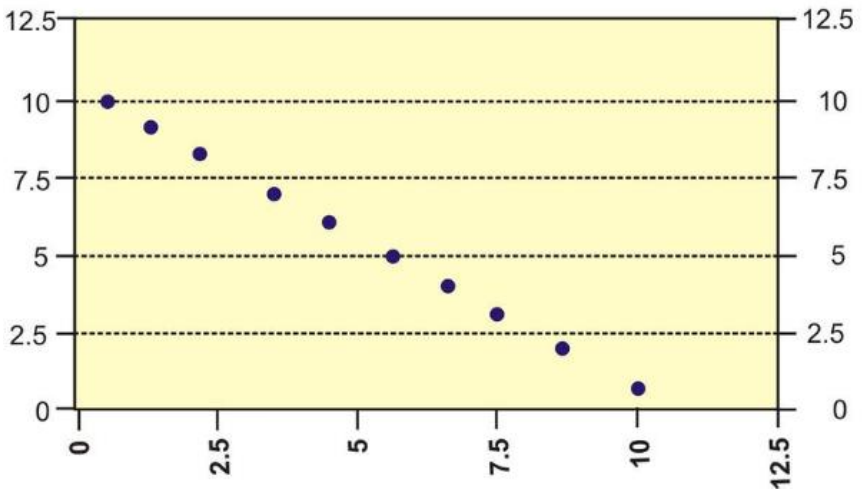
When the points on a scatterplot graph produce a lower-left-to-upper-right pattern (see below), we say that there is a positive correlation between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be high as well, and vice versa.



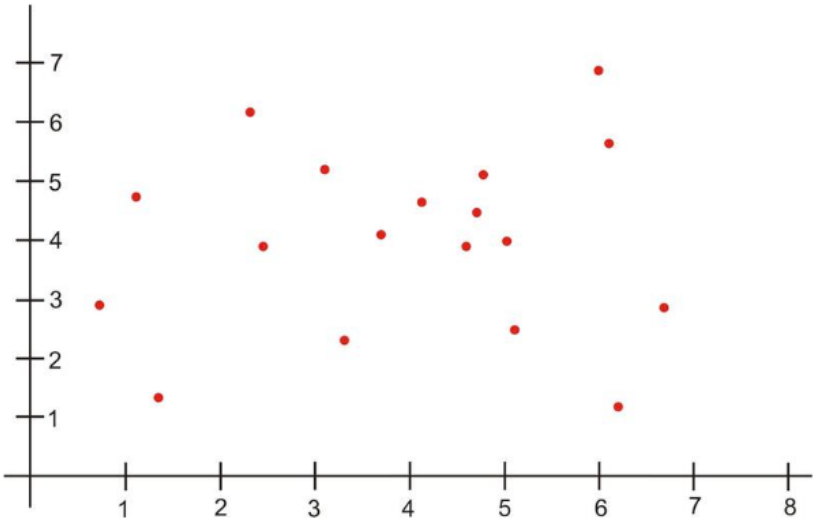
When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a negative correlation between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low, and vice versa.



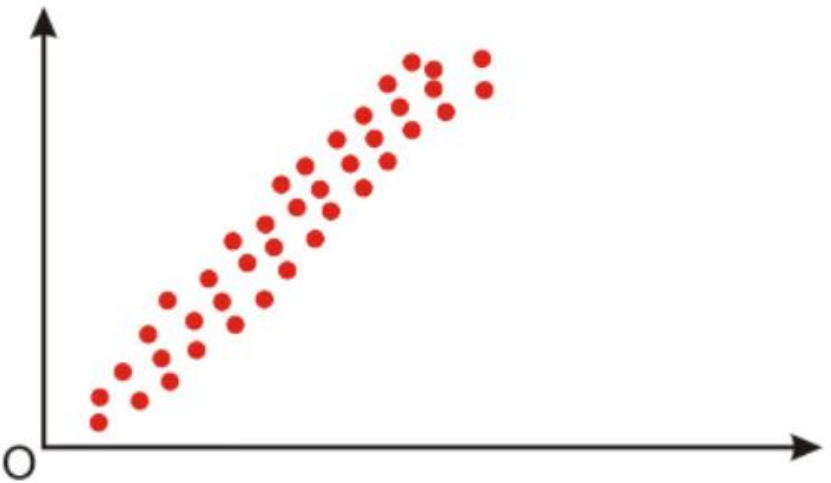
When all the points on a scatterplot lie on a straight line, you have what is called a perfect correlation between the two variables (see below).



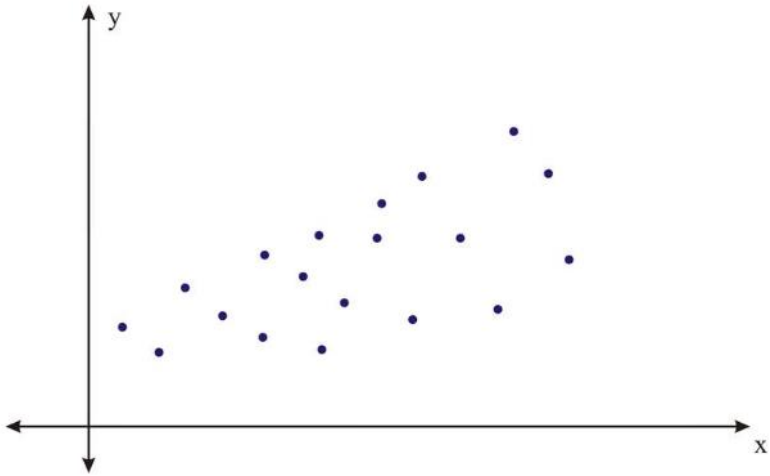
A scatterplot in which the points do not have a linear trend (either positive or negative) is called a zero correlation or a near-zero correlation (see below).



When examining scatterplots, we also want to look not only at the direction of the relationship (positive, negative, or zero), but also at the magnitude of the relationship. If we drew an imaginary oval around all of the points on the scatterplot, we would be able to see the extent, or the magnitude, of the relationship. If the points are close to one another and the width of the imaginary oval is small, this means that there is a strong correlation between the variables (see below).



However, if the points are far away from one another, and the imaginary oval is very wide, this means that there is a weak correlation between the variables (see below).



Correlation Coefficients

While examining scatterplots gives us some idea about the relationship between two variables, we use a statistic called the **correlation coefficient** to give us a more precise measurement of the relationship between the two variables. The correlation coefficient is an index that describes the relationship and can take on values between **−1.0** and +1.0, with a positive correlation coefficient indicating a positive correlation and a negative correlation coefficient indicating a negative correlation.

The absolute value of the coefficient indicates the magnitude, or the strength, of the relationship. The closer the absolute value of the coefficient is to 1, the stronger the relationship. For example, a correlation coefficient of 0.20 indicates that there is a weak linear relationship between the variables, while a coefficient of −0.90 indicates that there is a strong linear relationship.

The value of a perfect positive correlation is 1.0, while the value of a perfect negative correlation is −1.0.

When there is no linear relationship between two variables, the correlation coefficient is 0. It is important to remember that a correlation coefficient of 0 indicates that there is *no linear relationship*, but there may still be a strong relationship between the two variables. For example, there could be a quadratic relationship between them.

The Pearson product-moment correlation coefficient is a statistic that is used to measure the strength and direction of a linear correlation. It is symbolized by the letter r. To understand how this coefficient is calculated, let's suppose that there is a positive relationship between two variables, X and Y. If a

subject has a score on X that is above the mean, we expect the subject to have a score on Y that is also above the mean. Pearson developed his correlation coefficient by computing the sum of cross products. He multiplied the two scores, X and Y, for each subject and then added these cross products across the individuals. Next, he divided this sum by the number of subjects minus one. This coefficient is, therefore, the mean of the cross products of scores.

Pearson used standard scores (z-scores, t-scores, etc.) when determining the coefficient.

Therefore, the formula for this coefficient is as follows:

$$r_{XY} = \frac{\sum z_X z_Y}{n-1}$$

In other words, the coefficient is expressed as the sum of the cross products of the standard z-scores divided by the number of degrees of freedom.

An equivalent formula that uses the raw scores rather than the standard scores is called the raw score formula and is written as follows:

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

Again, this formula is most often used when calculating correlation coefficients from original data. Note that n is used instead of $n-1$, because we are using actual data and not z-scores. Let's use our example from the introduction to demonstrate how to calculate the correlation coefficient using the raw score formula.

Calculating the Pearson Product-Moment Correlation Coefficient

What is the Pearson product-moment correlation coefficient for the two variables represented in the table below?

The table of values for this example.

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

In order to calculate the correlation coefficient, we need to calculate several pieces of information, including xy, x^2, and y^2. Therefore, the values of xy, x^2, and y^2 have been added to the table.

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

Applying the formula to these data, we find the following:

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

$$= \frac{(7)(13262) - (3970)(22.7)}{\sqrt{[(7)(2321400) - 3970^2][(7)(76.01) - 22.7^2]}}$$

$$= \frac{2715}{2864.22} \approx 0.95$$

The correlation coefficient not only provides a measure of the relationship between the variables, but it also gives us an idea about how much of the total variance of one variable can be associated with the variance of the other. For example, the correlation coefficient of 0.95 that we calculated above tells us that to a high degree, the variance in the scores on the verbal SAT is associated with the variance in the GPA, and vice versa. For example, we could say that factors that influence the verbal SAT, such as health, parent college level, etc., would also contribute to individual differences in the GPA. The higher the correlation we have between two variables, the larger the portion of the variance that can be explained by the independent variable.

The calculation of this variance is called the **coefficient of determination** and is calculated by squaring the correlation coefficient. Therefore, the coefficient of determination is written as r2. The result of this calculation indicates the proportion of the variance in one variable that can be associated with the variance in the other variable.

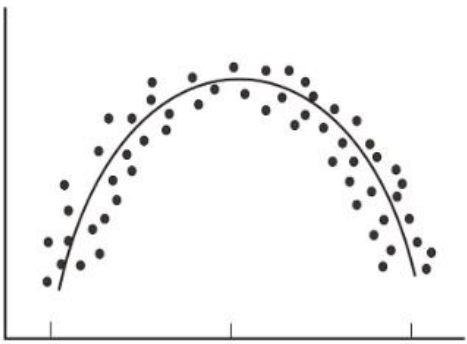
The Properties and Common Errors of Correlation

Correlation is a measure of the linear relationship between two variables-it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to remember that we are interpreting the variables and the variance not as causal, but instead as relational.

When examining correlation, there are three things that could affect our results: linearity, homogeneity of the group, and sample size.

Linearity:

As mentioned, the correlation coefficient is the measure of the linear relationship between two variables. However, while many pairs of variables have a linear relationship, some do not. For example, let's consider performance anxiety. As a person's anxiety about performing increases, so does his or her performance up to a point. (We sometimes call this good stress.) However, at some point, the increase in anxiety may cause a person's performance to go down. We call these non-linear relationships **curvilinear relationships**. We can identify curvilinear relationships by examining scatterplots (see below). One may ask why curvilinear relationships pose a problem when calculating the correlation coefficient. The answer is that if we use the traditional formula to calculate these relationships, it will not be an accurate index, and we will be underestimating the relationship between the variables. If we graphed performance against anxiety, we would see that anxiety has a strong affect on performance. However, if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient is not always the best statistic to use to understand the relationship between variables.



Homogeneity of the Group:

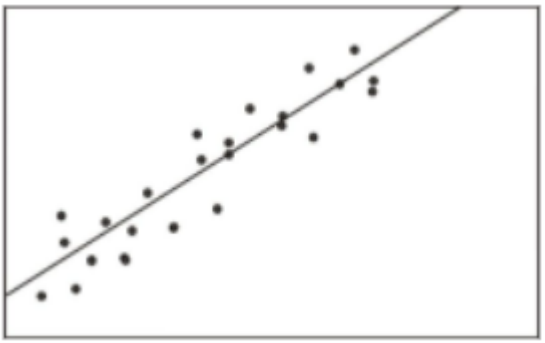
Another error we could encounter when calculating the correlation coefficient is homogeneity of the group. When a group is homogeneous, or possesses similar characteristics, the range of scores on either or both of the variables is restricted.

Sample Size:

Finally, we should consider sample size. One may assume that the number of observations used in the calculation of the correlation coefficient may influence the magnitude of the coefficient itself. However, this is not the case. Yet while the sample size does not affect the correlation coefficient, it may affect the accuracy of the relationship. The larger the sample, the more accurate of a predictor the correlation coefficient will be of the relationship between the two variables.

Least-Squares Regression

Least squares line, or the linear regression line

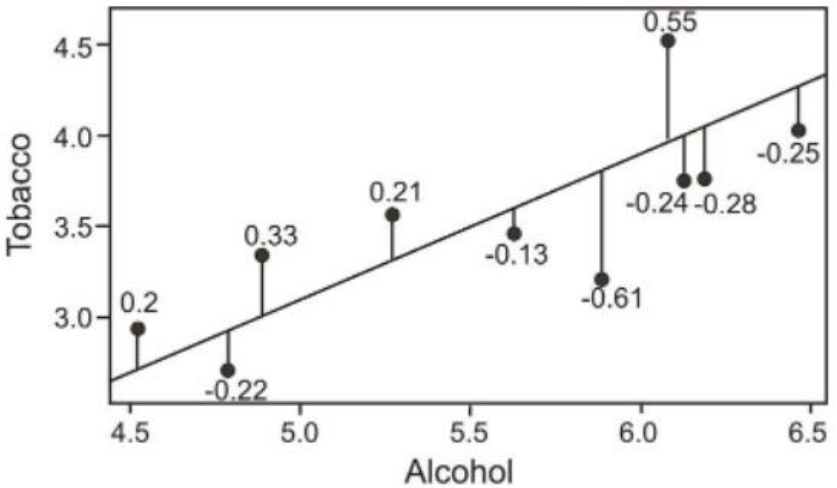


Calculating and Graphing the Regression Line

Linear regression involves using data to calculate a line that best fits that data and then using that line to predict scores. In linear regression, we use one variable (the **predictor variable**) to predict the outcome of another (the **outcome variable**, or **criterion variable**). To calculate this line, we analyze the patterns between the two variables.

We are looking for a line of best fit, and there are many ways one could define this best fit. Statisticians define this line to be the one which minimizes the sum of the squared distances from the observed data to the line.

To determine this line, we want to find the change in X that will be reflected by the average change in Y. After we calculate this average change, we can apply it to any value of X to get an approximation of Y. Since the regression line is used to predict the value of Y for any given value of X, all predicted values will be located on the regression line, itself. Therefore, we try to fit the regression line to the data by having the smallest sum of squared distances possible from each of the data points to the line. In the example below, you can see the calculated distances, or residual values, from each of the observations to the regression line. This method of fitting the data line so that there is minimal difference between the observations and the line is called the method of least squares, which we will discuss further in the following sections.



As you can see, the regression line is a straight line that expresses the relationship between two variables. When predicting one score by using another, we use an equation such as the following, which is equivalent to the slope-intercept form of the equation for a straight line:

$$Y = bX + a$$

where:

Y is the score that we are trying to predict.

b is the slope of the line.

a is the y-intercept, or the value of Y when the value of X is 0.

To calculate the line itself, we need to find the values for b (the **regression coefficient**) and a (the **regression constant**). The regression coefficient explains

the nature of the relationship between the two variables. Essentially, the regression coefficient tells us that a certain change in the predictor variable is associated with a certain change in the outcome, or criterion, variable. For example, if we had a regression coefficient of 10.76, we would say that a change of 1 unit in X is associated with a change of 10.76 units of Y. To calculate this regression coefficient, we can use the following formulas:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Or

$$b = (r)\frac{s_Y}{s_X}$$

where:

r is the correlation between the variables X and Y.

sY is the standard deviation of the Y scores.

sX is the standard deviation of the X scores.

In addition to calculating the regression coefficient, we also need to calculate the regression constant. The regression constant is also the y-intercept and is the place where the line crosses the y-axis. For example, if we had an equation with a regression constant of 4.58, we would conclude that the regression line crosses the y-axis at 4.58. We use the following formula to calculate the regression constant:

$$a = \frac{\sum y - b\sum x}{n} = \bar{y} - b\bar{x}$$

Finding the Line of Best Fit

Find the least squares line (also known as the linear regression line or the line of best fit) for the example measuring the verbal **SAT** scores and **GPAs** of students that was used in the previous section.

SAT and GPA data including intermediate computations for computing a linear regression.

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

Using these data points, we first calculate the regression coefficient and the regression constant as follows:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{(7)(13{,}262) - (3{,}970)(22.7)}{(7)(2{,}321{,}400) - 3{,}970^2}$$

$$= \frac{2715}{4889900}$$

$$\approx 0.0056$$

$$a = \frac{\sum y - b\sum x}{n} = 0.094$$

Note: If you performed the calculations yourself and did not get exactly the same answers, it is probably due to rounding in the table for xy.

Application

Compute the correlation coefficient for the following data:

X values	-5	1	3	9
Y values	6	5	-1	1

KEY POINTS

- ✓ Bivariate data has two variables
- ✓ Correlation is a statistical method used to determine if there is a connection or a relationship between two sets of data.
- ✓ Non-linear relationships are called curvilinear relationships.
- ✓ If the line on a line graph rises to the right, it indicates a direct relationship.
- ✓ When a group is homogeneous, or possesses similar characteristics, the range of scores on either or both of the variables is restricted.
- ✓ If the line on a line graph falls to the right, it indicates an indirect relationship.
- ✓ A linear relationship appears as a straight line either rising or falling as the independent variable values increase.
- ✓ A negative correlation appears as a recognizable line with a negative slope.
- ✓ A non-linear relationship may take the form of any number of curved lines but is not a straight line.
- ✓ A positive correlation appears as a recognizable line with a positive slope
- ✓ A scatter plot is a plot of the dependent variable versus the independent variable and is used to investigate whether or not there is a relationship or connection between 2 sets of data.
- ✓ Slope is a measure of the steepness of a line. A line can have positive, negative, zero (horizontal), or undefined (vertical) slope. The slope of a line can be found by calculating "rise over run" or "the change in the y over the change in the x." The symbol for slope is m
- ✓ Two variables with a strong correlation will appear as a number of points occurring in a clear and recognizable linear pattern.
- ✓ Trends in data sets or samples are indicators found by reviewing the data from a general or overall standpoint
- ✓ Two variables with a weak correlation will appear as a much more scattered field of points, with only a little indication of points falling into a line of any sort.
- ✓ The outcome variable in a linear regression is the criterion variable.
- ✓ An influential point in regression is one whose removal would greatly impact the equation of the regression line.

- ✓ The least squares method is a statistical method for calculating the line of best fit for a scatterplot.
- ✓ If we draw a straight line to represent the change in one variable associated with the change in the other. This line is called the linear regression line.
- ✓ The line of best fit on a scatterplot is a draw line that best represents the trend of the data.
- ✓ An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
- ✓ Prediction statement that tells what will happen under certain conditions.
- ✓ A predictor variable is a variable that can be used to predict the value of another variable (as in statistical regression).
- ✓ The regression coefficient (or the slope) explains the nature of the relationship between the two variables.
- ✓ The differences between the actual and the predicted values are called residual values.
- ✓ tests for linearity, a rule of thumb that says a relationship is linear if the difference between the linear correlation coefficient (r) and the nonlinear correlation coefficient (eta) is small

END OF MODULE ASSESSMENT

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. On every module's "End of Module Assessment" (this part), write your answers on a one whole sheet of yellow pad paper.

LOOKING AHEAD

Congratulations for making it till the end of this module! Happy learning!

REFERENCES

https://www.ck12.org/book/ck-12-advanced-probability-and-statistics-concepts/section/3.0/

END OF MODULE ASSESSMENT

(Please do not forget to provide information on this part)

Name: _____ Course& Year: _____
Module Number and Title: _____ Date accomplished: _____

Now that you have finished the review of the various concepts outlined above, it is now time for an assessment to see how far you have improved. Write your answers on a one whole sheet of yellow paper.

Task 1. Answer the following questions.

1. Give 2 scenarios or research questions where you would use bivariate data sets.
2. For the following table of values calculate the regression line and then calculate $\hat{y}$ for each data point.

X	4	4	7	10	10
Y	15	11	19	21	29