1. Introduction

Welcome to the series of E-learning modules on Confidence Intervals for the difference between the means.

By the end of this session, you will be able to:

- Explain the confidence interval for the difference between the averages of two populations when the population standard deviations are known
- Explain the confidence interval for the difference between the averages of two populations when the population standard deviations are common but unknown
- Explain the confidence interval for the difference of means of two dependent samples

The problem for a quantitative population is the comparison of two population means, which is as important as the estimation of a single population mean mu.

We may be interested to make comparisons like the following:

- The average scores in the medical college admission test for students whose major was Biochemistry and those whose major was Biology
- The average yield in a chemical plant using new raw materials furnished by two different suppliers
- The average stem diameter of plants grown on two different types of nutrients

For each of these examples there are two populations:
The first with mean and variance mu one and sigma one square and the second with mean and variance mu two and sigma two square.
A random sample of size m is drawn from the population I and a sample of size n is drawn independently from the population II.

Intuitively, the difference between the two sample means would provide the maximum information about the actual difference between two population means. The best point estimator of the difference (mu 1 minus mu 2) between the population means is (x bar minus y bar)

The confidence interval formula yields a range (interval) within which the difference between two populations mean is located.
This topic describes how to construct a confidence interval for the difference between two means.

Estimation Requirements
The approach described in this topic is valid whenever the following conditions are met:

- Both samples are simple random samples
- The samples are independent
- Each population is at least 10 times larger than its respective sample
- The sampling distribution of the difference between means is approximately normally distributed

Generally, the sampling distribution will be approximately normally distributed if each sample is described by at least one of the following statements:

- The population distribution is normal
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is fifteen or less
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between sixteen and forty
- The sample size is greater than forty, without outliers

The variability of the difference between sample means

To construct a confidence interval, we need to know the variability of the difference between sample means. This means we need to know how to compute the standard deviation of the sampling distribution of the difference.

• If the population standard deviations are known, the standard deviation of the sampling distribution is:

Sigma is equal to standard deviation of (x bar minus y bar) which is equal to square root of sigma 1 square by m plus sigma 2 square by n

Where, sigma 1 is the standard deviation of the population 1, sigma 2 is the standard deviation of the population 2, and m is the size of sample 1, and n is the size of sample 2

When the standard deviation of either population is unknown and the sample sizes (m and n) are large, the standard deviation of the sampling distribution can be estimated by the standard error, using the below equation:

Standard error of (x bar minus y bar) is equal to square root of s1 square by m plus s2 square by n

Where, s1 is the standard deviation of the sample 1, s2 is the standard deviation of the sample 2, m is the size of sample 1, and n is the size of sample 2.

2. CI's for the difference of Means of Two Populations with Known SD's

1. Confidence Intervals for the difference of means of two populations with known standard deviations

Assumptions:

- Populations standard deviations are known
- Population is normally distributed
- If population is not normal, use large sample

Suppose x one, x two, up to xm and y one, y two, up to yn are the random samples of size m and n drawn from the populations of size N1 and N2 with means mu 1 and mu 2 and variances sigma 1 square and sigma 2 square, we know that
x bar is equal to summation xi by m and y bar is equal to summation yi by n
Z is equal to (x bar minus y bar) minus (mu 1 minus mu 2) by square root of sigma 1 square by m plus sigma 2 square by n which follows normal with mean zero and variance 1

We can always find two quantities minus Z alpha by 2 and plus Z alpha by 2 from standard normal variate tables such as
Probability of minus Z alpha by 2 less than or equal to Z less than or equal to plus Z alpha by 2 is equal to 1 minus alpha
Probability of minus Z alpha by 2 less than or equal to (x bar minus y bar) minus (mu 1 minus mu 2) by square root of sigma 1 square by m plus sigma 2 square by n less than or equal to plus Z alpha by 2 is equal to 1 minus alpha
Probability of minus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n less than or equal to (x bar minus y bar) minus (mu 1 minus mu 2) less than or equal to plus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n is equal to 1 minus alpha

Probability of Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n greater than minus (x bar minus y bar) plus (mu 1 minus mu 2) greater than minus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n is equal to 1 minus alpha
Probability of (x bar minus y bar) plus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n greater than (mu 1 minus mu 2) greater than (x bar minus y bar) minus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n is equal to 1 minus alpha
Probability of (x bar minus y bar) minus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n less than or equal to (mu 1 minus mu 2) less than or equal to (x bar minus y bar) plus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n is equal to 1 minus alpha

Therefore, hundred into (1 minus alpha) confidence interval for the difference of means of

two populations with known standard deviations is given by
[(x bar minus y bar) minus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n, (x bar minus y bar) plus Z alpha by 2 into square root of sigma 1 square by m plus sigma 2 square by n]

3. CI's for the difference of Means of Two Populations with Unknown but Common SD's

2. Confidence Interval for the difference of means of two populations with unknown but common standard deviations

Assumptions:

- Populations standard deviations are unknown but common
- Population is normally distributed
- If population is not normal, use large sample

Suppose x one, x two, up to xm and y one, y two, up to yn are the random samples of size m and n drawn from the normal populations of size N1 and N2 with unknown means mu 1 and mu 2 and common variance sigma square,
We know that
Xi follows Normal with mean mu 1 and variance sigma square and
Yi follows Normal with mean mu 2 and variance sigma square
Where, sigma square is unknown
We know that x bar is equal to summation xi by m and y bar is equal to summation yi by n
X bar follows Normal with mean mu 1 and variance sigma square by m and y bar follows Normal with mean mu 2 and variance sigma square by n
Therefore, x bar minus y bar follows Normal with mean mu1 minus mu 2 and variance sigma square by m plus sigma square by n

Z is equal to (x bar minus y bar) minus (mu 1 minus mu 2) by square root of sigma square by m plus sigma square by n which follows normal with mean zero and variance 1
(m minus 1) into s1 square by sigma square follows chisquare with m minus 1 degrees of freedom and (n minus 1) into s2 square by sigma square follows chisquare with n minus 1 degrees of freedom
Which implies (m minus 1) into s1 square by sigma square plus (n minus 1) into s2 square by sigma square follows chi square with m plus n minus 2 degrees of freedom

t is equal to Z by square root of chi square by degrees of freedom which is equal to(x bar minus y bar) minus (mu 1 minus mu 2) by square root of sigma square by m plus sigma square by n divided by
square root of (m minus 1) into s1 square by sigma square plus (n minus 1) into s2 square by sigma square by (m plus m minus 2) follows t alpha distribution with (m plus n minus 2) degrees of freedom
t is equal to (x bar minus y bar) minus (mu 1 minus mu 2) by Sp into square root of (1 by m plus 1 by n) follows t alpha (m plus n minus 2)
Where, Sp is equal to square root of (m minus 1) into s1 square plus (n minus 1) into s2 square by (m plus n minus 2)

We can always find two quantities minus t alpha (m plus n minus 2) and t alpha (m plus n minus 2) from student's t variate table such that
Probability of minus t alpha (m plus n minus 2) less than or equal to t less than or equal to plus t alpha (m plus n minus 2) is equal to 1 minus alpha
Probability of minus t alpha (m plus n minus 2) less than or equal to (x bar minus y bar) minus (mu 1 minus mu 2) by Sp into square root of (1 by m plus 1 by n) less than or equal to plus t alpha (m plus n minus 2) is equal to 1 minus alpha
Probability of minus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) less than or equal to (x bar minus y bar) minus
(mu 1 minus mu 2) less than or equal to plus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) is equal to 1 minus alpha

Probability of t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) greater than minus (x bar minus y bar) plus
(mu 1 minus mu 2) greater than minus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) is equal to 1 minus alpha
Probability of (x bar minus y bar) plus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) greater than
(mu 1 minus mu 2) greater than (x bar minus y bar) minus t alpha (m plus n minus 2) into Sp into square root of(1 by m plus 1 by n) is equal to 1 minus alpha
Probability of (x bar minus y bar) minus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) less than or equal to
(mu 1 minus mu 2) less than or equal to (x bar minus y bar) plus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n) is equal to 1 minus alpha

Therefore, hundred into (1 minus alpha) percent Confidence Interval for the difference of means of two populations with unknown but common standard deviations is given by
[(x bar minus y bar) minus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n), (x bar minus y bar) plus t alpha (m plus n minus 2) into Sp into square root of (1 by m plus 1 by n)]

4. CI's for the Mean in case of Correlated Variables

3. Confidence Interval for the mean in case of correlated (dependent) variables
Let (xi, yi), is equal to 1, 2, up to n be n pairs of observations which are correlated. Let the deviation di is equal to xi minus yi,
Assume that di follows normal with mean theta and variance sigma d square.
Then, d bar follows normal with mean theta and variance sigma d square by n
Z is equal to d bar minus theta by sigma d by root n which is equal to root n into (d bar minus theta) by sigma d

We know that (n minus 1) into sd square by sigma d square follows chi square with (n minus 1) degrees of freedom, where sd square is equal to summation (di minus d bar) whole square by (n minus 1)
We have Z by root of chi square into (n minus 1) by n minus 1 which is equal to root n into (d bar minus theta) by sigma d divided by square root of (n minus 1) into s d square by sigma d square into (n minus 1) which is equal to root n (d bar minus theta) by sd which follows t alpha with (n minus 1) degrees of freedom

We can always find two quantities minus t alpha (n minus 1) and t alpha (n minus 1) from students t-distribution table such that
Probability of minus t alpha (n minus 1) less than or equal to t less than or equal to plus t alpha (n minus1) is equal to 1 minus alpha
Probability of minus t alpha (n minus 1) less than or equal to root n (d bar minus theta) by sd less than or equal to plus t alpha (n minus 1) is equal to 1 minus alpha
Probability of minus t alpha (n minus 1) into sd by root of n less than or equal to d bar minus theta less than or equal to plus t alpha (n minus 1) into sd by root n is equal to 1 minus alpha

Probability of d bar minus t alpha (n minus 1) into sd by root n less than or equal to theta less than or equal to d bar plus t alpha (n minus 1) into sd by root n is equal to 1 minus alpha
Therefore, hundred into (1 minus alpha) percent confidence interval for the population mean theta in case of correlated variables is given by
[d bar minus t alpha (n minus 1) into sd by root n, d bar plus t alpha (n minus 1) into sd by root n, where s d square is equal to summation (di minus d bar) whole square by n minus 1 and d bar is equal to summation di by n

5. Steps to Construct CI's for the Difference between Means

Steps to construct confidence interval for the difference between means

- Identify a sample statistic. Use the difference between sample means to estimate the difference between population means
- Select a confidence level. Often, researchers choose ninety percent, ninety-five percent or ninety nine percent confidence levels, but any percentage can be used
- Find the margin of error which is equal to the table or critical value into standard error of (x bar minus y bar) and then use the derived rule to obtain the confidence interval

When the sample size is large, you can use a t score or a z score for the critical value. Since it does not require computing degrees of freedom, the z score is a little easier. When the sample sizes are small (less than thirty), use a t score for the critical value.

Here's a summary of our learning in this session, where we understood:

- The derivation of confidence limits for the difference of population means when the variances are known and unknown
- The derivation of confidence limits for the population mean in case of correlated variables