Frequently Asked Questions

1. Explain Interval estimates for the difference of population means with examples.

Answer:

The confidence interval formula for the difference between two population means yields a range (interval) within which we feel with some confidence the difference between two populations mean is located.

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 diameters of plants grown on two different types of nutrients

2. Discuss the parameters and estimators used in the interval estimation of the difference between the means.

Answer:

To obtain an interval estimate for the difference between the two population means we need two populations. Suppose the first with mean and variance μ_1 and σ_1^2 and the second with mean and variance μ_2 and σ_2^2. 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. Finally, the estimates of the population parameters are calculated from the sample data using estimators as shown in the table below:

	Population I	Population II
Mean	μ_1	μ_2
Variance	σ_1^2	σ_2^2

	Sample I	Sample II
Mean	$\bar{x}$	$\bar{y}$
Variance	s_1^2	s_2^2
Sample size	m	n

3. What is the best point estimator for the difference of population means that can be used to construct a confidence interval?

Answer:

Intuitively the difference between the two sample means would provide the maximum information about the actual difference between two population means and this is in fact the case. The best point estimator of the difference ($\mu_1 - \mu_2$) between the population means is ($\bar{x} - \bar{y}$)

4. What are the conditions to be satisfied to get an interval estimate of the difference of population means?

Answer:

Estimation Requirements

The interval estimation approach 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

5. When the sampling distribution will be normally distributed?

Answer:

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 15 or less
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40
- The sample size is greater than 40, without outliers

6. How do you compute the standard deviation of the sampling distribution of the difference?

Answer:

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:

$$(\sigma = \mathrm{S.D}(\overline{x} - \overline{y}) = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Where, σ1 is the standard deviation of the population 1, σ2 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 equation below.

$$\mathrm{S.E} = (\overline{x} - \overline{y}) = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

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

- In real-world analyses, the standard deviation of the population is seldom known. Therefore, $\mathrm{S..E}(\overline{x} - \overline{y})$ is used more often than $\mathrm{S.D}(\overline{x} - \overline{y})$.

7. Derive an interval estimate for the difference of means of two populations with Known Standard Deviations.

Answer:

Suppose $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ are the random samples of size m and n drawn from the populations of size N_1 and N_2 with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 we know that

We know that $\overline{x} = \dfrac{\sum_{i=1}^{m} x_i}{m}$ and $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

$$Z = \frac{(\overline{x} - \overline{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

We can always find two quantities $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ from Standard Normal variate tables such as

$P[-Z_{\alpha/2} \le Z \le Z_{\alpha/2}] = 1-\alpha$

$$P\left[-Z_{\alpha/2} \le \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \le Z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}} \le (\bar{x}-\bar{y})-(\mu_1-\mu_2) \le Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right] = 1-\alpha$$

$$P\left[Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}} > \mu_1-\mu_2 > -Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right] = 1-\alpha$$

$$P\left[(\bar{x}-\bar{y})+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}} > (\mu_1-\mu_2) > (\bar{x}-\bar{y})-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right] = 1-\alpha$$

$$P\left[(\bar{x}-\bar{y})-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}} \le (\mu_1-\mu_2) \le (\bar{x}-\bar{y})+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right] = 1-\alpha$$

Therefore 100 (1- α) % C.I for the difference of means of two populations with Known Standard Deviations is given by

$$\left[(\bar{x}-\bar{y})-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}},\ (\bar{x}-\bar{y})+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right]$$

8. Derive an interval estimate for the difference of means of two populations with unknown but common Standard Deviations.

Answer:

Suppose $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ are the random samples of size m and n drawn from Normal populations of size N_1 and N_2 with unknown means μ_1 and μ_2 and variances σ^2 we know that

$X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ where σ^2 is unknown

We know that $\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$\bar{x} \sim N(\mu_1, \sigma^2/m)$ and $\bar{y} \sim N(\mu_2, \sigma^2/n)$

Therefore, $\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, \sigma^2/m + \sigma^2/n)$

$$Z=\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{m}+\frac{\sigma^2}{n}}}\sim N(0,1)$$

$$\frac{(m-1)s_1^2}{\sigma^2}\sim\chi^2(m-1) \quad \text{and} \quad \frac{(n-1)s_2^2}{\sigma^2}\sim\chi^2(n-1)$$

$$\Rightarrow\frac{(m-1)s_1^2}{\sigma^2}+\frac{(n-1)s_2^2}{\sigma^2}\sim\chi^2(m+n-2)$$

$$t=\frac{Z}{\sqrt{\frac{\chi^2}{d.f}}}=\frac{\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{m}+\frac{\sigma^2}{n}}}}{\sqrt{\frac{\frac{(m-1)s_1^2}{\sigma^2}+\frac{(n-1)s_2^2}{\sigma^2}}{m+n-2}}}\sim t_\alpha(m+n-2)$$

$$t=\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t_\alpha(m+n-2) \text{ where } S_p=\sqrt{\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}}$$

We can always find two quantities – $t_\alpha(m+n-2)$ and $t_\alpha(m+n-2)$ from Student's t variate table such that

$P[-t_\alpha(m+n-2) \le t \le t_\alpha(m+n-2)] = 1-\alpha$

$$P\left[-t_\alpha(m+n-2) \le \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \le t_\alpha(m+n-2)\right] = 1-\alpha$$

$$P\left[-t_\alpha(m+n-2)\, S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \le (\bar{x}-\bar{y})-(\mu_1-\mu_2) \le t_\alpha(m+n-2)\, S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right] = 1-\alpha$$

$$P\left[t_\alpha(m+n-2)\, S_p\sqrt{\frac{1}{m}+\frac{1}{n}} > \mu_1-\mu_2 > -t_\alpha(m+n-2)\, S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right] = 1-\alpha$$

P[($\bar{x}$ - $\bar{y}$)+ t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$ > ($\mu_{1-}\mu_{2)}$ > ($\bar{x}$ - $\bar{y}$)- t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$] = 1- α

P[($\bar{x}$ - $\bar{y}$)- t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$ ≤ ($\mu_{1-}\mu_{2)}$ ≤ ($\bar{x}$ - $\bar{y}$)+ t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$] = 1- α

Therefore 100 (1- α) % C.I for the difference of means of two populations with unknown but common Standard Deviations is given by

[($\bar{x}$ - $\bar{y}$)- t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$, ($\bar{x}$ - $\bar{y}$)+ t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$]

9. Derive a Confidence Interval for the mean in case of correlated (dependent) variables.

Answer:

Let (x_i,y_i) , i=1,2,...,n be n pairs of observations which are correlated. Let the deviation $d_i=x_i-y_i$

Assume that $d_i \sim N(\theta, \sigma_d^2)$

Then $\bar{d} \sim N(\theta, \frac{\sigma_d^2}{n})$

$$Z= \frac{\bar{d}-\theta}{\frac{\sigma_d}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{d}-\theta)}{\sigma_d}$$

We know that $\frac{(n-1)s_d^2}{\sigma_d^2} \sim \chi^2(n-1)$ where $s_d^2 = \frac{\sum(d_i-\bar{d})^2}{n-1}$

$$\text{We have } \frac{Z}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} = \frac{\frac{\sqrt{n}(\bar{d}-\theta)}{\sigma_d}}{\sqrt{\frac{(n-1){s_d}^2}{{\sigma_d}^2(n-1)}}} = \frac{\sqrt{n}(\bar{d}-\theta)}{s_d} \sim t_\alpha(n-1)$$

We can always find two quantities – t_α(n-1) and t_α(n-1) from Students t-distribution table such that

P[-t_α(n-1) ≤ t ≤ t_α(n-1)] = 1- α

$$P[t_\alpha(n-1) \le \frac{\sqrt{n}(\bar{d}-\theta)}{s_d} \le t_\alpha(n-1)] = 1-\alpha$$

$$P[-t_\alpha(n-1)\frac{s_d}{\sqrt{n}} \le (\bar{d}-\theta) \le t_\alpha(n-1)\frac{s_d}{\sqrt{n}}] = 1-\alpha$$

$$P[\bar{d} - t_\alpha(n-1)\frac{s_d}{\sqrt{n}} \le \mu \le \bar{d} + t_\alpha(n-1)\frac{s_d}{\sqrt{n}}] = 1-\alpha$$

Therefore 100 (1- α) % C.I for the population mean θ in case of correlated variables is given by

$$[\bar{d} - t_\alpha(n-1)\frac{s_d}{\sqrt{n}},\ \bar{d} + t_\alpha(n-1)\frac{s_d}{\sqrt{n}}] \text{ where } s_d^2 = \frac{\sum(d_i-\bar{d})^2}{n-1} \text{ and } \bar{d} = \frac{\sum d_i}{n}$$

10. What are the steps to be followed for the construction of the C.I for the difference between means of two populations when the population variances are known?

Answer:

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. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
- Find the margin of error= critical value (table value) into standard deviation.= $Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}$

  When the population variances are known z score is used as the critical value. Since it does not require computing degrees of freedom, the z score is a little easier.
- Then 100 (1- α) % C.I for the difference of means of two populations with Known Standard Deviations is given by
- $[(\bar{x} - \bar{y})-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}},\ (\bar{x} - \bar{y})+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}]$

11. What are the steps to be followed for the construction of the C.I for the difference between Means of two populations when the population variances are unknown?

Answer:

Steps to be followed for the construction of the C.I for the difference between Means of two populations when the population variances are unknown

- Identify a sample statistic. Use the difference between samples means to estimate the difference between population means.
- Select a confidence level. One can choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
- Find the margin of error. = critical value (table value) into standard deviation.= $t\alpha(m+n-2)$ $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$
- 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 40), use a t score for the critical value.
    - If you use a t score, you will need to compute degrees of freedom (DF).
    - If we are working with a pooled standard deviation (see above), DF = m + n - 2.
- Then, 100 (1- α) % C.I for the difference of means of two populations with unknown but common Standard Deviations is given by

$$\left[(\bar{x} - \bar{y}) - t_{\alpha}(m+n-2)\ S_p\sqrt{\frac{1}{m}+\frac{1}{n}},\ (\bar{x} - \bar{y}) + t_{\alpha}(m+n-2)\ S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right]$$

Where, $S_p = \sqrt{\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}}$

12. Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.
What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions with common variances in both schools?

Answer:

Given m=15, n=20, $\bar{x}$ = 1000, $\bar{y}$ = 950, s1=100, s2=90, 1-α=0.90

$$\bar{x}-\bar{y}=1000-950=50$$

100 (1- α) % C.I for the difference of means of two populations with unknown but common Standard Deviations is given by

[($\bar{x}$ - $\bar{y}$)- t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$, ($\bar{x}$ - $\bar{y}$)+ t_α(m+n-2) $S_p\sqrt{\frac{1}{m}+\frac{1}{n}}$]

- Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.

  t_α(m+n-2)= $t_{0.1}$(33)=1.282

  $$S_p=\sqrt{\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}}$$

  $$S_p=\sqrt{\frac{(15-1)(100)^2+(20-1)(90)^2}{15+20-2}}$$

  - =2.8598

  By substituting the above in the interval we get

  [50-(1.282)(2.8598) $\sqrt{\frac{1}{15}+\frac{1}{20}}$,50+(1.282)(2.8598) $\sqrt{\frac{1}{15}+\frac{1}{20}}$]

  [48.74755, 51.25245]

Therefore, the 90% confidence interval is 48.74755 to 51.25245.

13. The local baseball team conducts a study to find the amount spent on refreshments at the ballpark. Over the course of the season, they gather simple random samples of 50 men and 100 women. For men, the average expenditure was \$20, with a standard deviation of \$3. For women, it was \$15, with a standard deviation of \$2.
What is the 99% confidence interval for the spending difference between men and women? Assume that the two populations are independent and normally distributed.

Answer:

- Identify a sample statistic. Since we are trying to estimate the difference between population means, we choose the difference between sample means as the sample statistic. Thus, x_1 - x_2=

  $20 - $15 = $5.

  - Find standard error. The standard error is an estimate of the standard deviation of the difference between population means. We use the sample standard deviations to estimate the standard error (SE).

- S.E$(\overline{x} - \overline{y}) = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$

  S.E$(\overline{x} - \overline{y}) = \sqrt{\frac{3^2}{50} + \frac{2^2}{100}}$ = 0.47

  - Find critical value. The critical value is a factor used to compute the margin of error. Because the sample sizes are large enough, we express the critical value as a z score.

  - $Z_{0.01/2}$ =$Z_{0.05}$=2.58

  - Compute margin of error (ME): ME = critical value * S.E
  - = $Z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$

    = 2.58 * 0.47 = 1.21

Therefore 100 (1- α) % C.I for the difference of means of two populations with unknown Standard Deviations is given by

$$\left[(\overline{x} - \overline{y}) - Z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}\ ,\ (\overline{x} - \overline{y}) + Z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}\right]$$

=[5-1.21, 5+1.21]= [3.79, 6.21]

Therefore, the 99% confidence interval is $3.79 to $6.21. That is, we are 99% confident that men outspend women at the ballpark by about $5 ± $1.21.

14. How to find tα from the table of t distribution?

Answer:

In the table of values of t we have the 'α' values at the top and degrees of freedom at the left margin. For desired degree of freedom and the value of α we can get t_α from the body of the table. The degree of freedom is infinity when the test statistic follows Normal distribution.. When the test statistic follows t- distribution degrees of freedom depends on the sample size (May be (n-1) or (m+n-2). The value of α depends on the confidence coefficient. For example: If we want 95% confidence Interval 1-α=0.95 Then α=0.05 similarly for 99% confidence interval 1-α=0.99 and α=0.01 and so on

15. Marks in a test before and after coaching of 9 students are given below.

Assuming that the difference in marks is from Normal distribution find a

90% Confidence Interval for the difference in the mean performance.

Marks before coaching: 60 12 38 45 19 72 70 50 48

Marks after coaching: 62 28 40 41 30 75 70 54 50

Answer:

The given set of observations is correlated variables. Hence, we find di for the given values.

x_i	y_i	$d_i=x_i-y_i$	d_i^2
60	62	-2	4
12	28	-16	256
38	40	-2	4
45	41	4	16
19	30	-11	121
72	75	-3	9
70	70	0	0
50	54	-4	16
48	50	-2	4
Total		-36	430

Therefore, 100 (1- α) % C.I for the population mean θ in case of correlated variables is given by

[$\bar{d}$ - $t_\alpha(n-1)\ \frac{s_d}{\sqrt{n}}$, $\bar{d}$ + $t_\alpha(n-1)\ \frac{s_d}{\sqrt{n}}$] where $s_d^2 = \frac{\sum(d_i - \bar{d})^2}{n-1}$ and $\bar{d} = \frac{\sum d_i}{n}$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-36}{9} = -4$$

$$s_d^2 = \frac{\sum(d_i - \bar{d})^2}{n-1} = \frac{\sum di^2 - n\bar{d}^2}{n-1} = 35.75$$ and $s_d = 5.9791$

100 (1- α) %=90% which implies α=0.1

From the table of t-distribution we get $t_\alpha(n-1) = t_{0.1}(8)=1.86$

Therefore 90 % C.I for the difference in mean performance is given by

[$\bar{d}$ - $t_\alpha(n-1)\ \frac{s_d}{\sqrt{n}}$, $\bar{d}$ + $t_\alpha(n-1)\ \frac{s_d}{\sqrt{n}}$]

[-4 – 1.86 $\frac{5.9791}{\sqrt{9}}$, -4 + 1.86 $\frac{5.9791}{\sqrt{9}}$,] = [-7.707,-0.2930]

Hence, 90% Confidence Interval for the difference in the mean performance is [-7.707,-0.2930]