Course: Mathematical statistics

Week 5: interval Estimation

Lecturer: Nagulama Moses

Kumi University

April 17, 2025

Outline

1. Interval Estimation
2. Confidence interval for μ when the variance is known
3. Confidence interval for μ when the variance is unknown
4. Confidence Interval for the Difference Between Two Population Means
5. CI for $\mu_1 - \mu_2$ when σ_1, σ_2 are unknown $n_1 \leq 30, n_2 < 30$

Intended learning outcomes

- Differentiate between point estimation and interval estimation
- Use the standard normal distribution (Z-distribution) to construct confidence intervals for μ.
- Use the t-distribution to construct confidence intervals for μ.
- Interpret the meaning of a confidence interval for $\mu_1 - \mu_2$ in context

Interval estimation

- locating an interval $(\hat{\theta_1}, \hat{\theta_2})$ which is believed to contain a population parameter θ is called interval estimation
- if $\hat{\theta_1}$ and $\hat{\theta_2}$ are estimates resulting from a set of sample values such that the $P(\hat{\theta_1} < \theta < \hat{\theta_2}) = 1 - \alpha, 0 < \alpha < 1$ then $(\hat{\theta_1}, \hat{\theta_2})$ is called the $(1 - \alpha)100\%$ confidence interval for the parameter θ and the fraction $1 - \alpha$ is called the confidence coefficient or the degree of confidence.
- The end points θ_1 and θ_2 are called the confidence limit θ_1- lower confidence limit and θ_2- upper confidence limit.

Confidence interval for μ when the variance is known

- we know that $\bar{x}$ is unbiased estimator of the population parameter μ
- Basic problem is to determine constant k in the interval $\bar{x} - k < \mu < \bar{x} + k$ such that $P(\bar{x} - k < \mu < \bar{x} + k) = 1 - \alpha, 0 < \alpha < 1$
- from

$$\bar{x} - k < \mu < \bar{x} + k..............(i)$$

$$\bar{x} - k < \mu, \bar{x} < \mu + k.......(ii)$$

$$\mu < \bar{x} + k, \mu - k < \bar{x}.........(iii)$$

from (ii) and (iii)

$$\mu - k < \bar{x} < \mu + k$$

$\bar{x} \pm k$ contain μ, the point estimate $\bar{x}$ will fall in the interval $\mu \pm k$ and therefore

$$P(\bar{x} - k < \mu < \bar{x} + k) = P(\mu - k < \bar{x} < \mu + k) = 1 - \alpha$$

by CLT sampling distribution of $\bar{x}$ is normal with $\mu_{\bar{x}} = \mu$ and $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

The quantity

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

is standard normal variable μ

from $P(\mu - k < \bar{x} < \mu + k) = 1 - \alpha$

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

if $\bar{x} = \mu + k$

$$z_{\frac{\alpha}{2}} = \frac{\mu + k - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$k = z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

$(1-\alpha)100\%$ C.I for μ

$$\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

if you take many samples from a population whose mean is unknown then the probability is $(1-\alpha)100\%$ of the different intervals that we calculate

NB: we have considered samples taken from a known distribution with a known variance.

suppose the distribution is not normal and a sample variance is not known provided that $n < 30$ we will replace the population standard deviation σ by a sample standard deviation s

Example

A random sample of size n=12 and a sample mean of 64.75 is taken from a normal population with standard deviation of 2.79. find

(i) the 95% confidence interval

(ii) 99% confidence interval for the true population mean

$$n = 12, \bar{x} = 64.75, \sigma = 2.79$$

population normally distributed

$$(1-\alpha)100\% = \bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

$$95 = (1-\alpha)100, \alpha = 0.05$$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

$$64.75 - 1.96 \times \frac{2.79}{\sqrt{12}} < \mu < 64.75 + 1.96 \times \frac{2.79}{\sqrt{12}}$$

$$63.2 < \mu < 66.3$$

(ii) 99% CI

$$(1-\alpha)100 = 99, \alpha = 0.01$$

$$z_{0.005} = 2.576$$

$$64.75 - 2.576 \times \frac{2.79}{\sqrt{12}} < \mu < 64.75 + 2.576 \times \frac{2.79}{\sqrt{12}}$$

$$62.7 < \mu < 66.8$$

99% is longer than the 95% confidence interval. there is a higher level of confidence in the 99% confidence interval
in general for a fixed sample size and standard deviation there is a higher confidence level he longer the confidence limit.

Confidence interval for μ when the variance is unknown

we need to determine the constant k in $\bar{x} - k < \mu < \bar{x} + k$

$$P(\bar{x} - k < \mu < \bar{x} + k) = 1 - \alpha, 0 < \alpha < 1$$

$$\bar{x} - k < \mu, \mu < \bar{x} + k$$

$$\bar{x} < \mu + k, \mu - k < \bar{x}$$

$$\mu - k < \bar{x} < \mu + k$$

$$P(\mu - k < \bar{x} < \mu + k) = 1 - \alpha$$

but if $n < 30$ sampling distribution of sample mean doesn't follow the z-distribution
if the sample are taken from a distribution that is normal then the sampling distribution of a sample means follow a student t-distribution with $\mu_{\bar{x}} = \mu$ and $\sigma^2_{\bar{x}} = \frac{s^2}{n}$
consequently

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

with (n-1) degree of freedom

Properties of t-distribution

- The values for a t-distribution is different for different values of sample size n.
- Like the z-distribution the t-distribution is centred and symmetric about zero
- The area under the curve of a t-distribution is 1
- As t becomes large the graph of a t-distribution but never becomes zero similarly as $t \to \infty$, the graph also approaches but not to zero

- as the sample size n increases, the density curve for t-distribution get closer and closer to the standard normal density curve

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t_{\frac{\alpha}{2}} = \frac{\mu + k - \mu}{\frac{s}{\sqrt{n}}}$$

$$k = t_{\frac{\alpha}{2}}, n - 1 \times \frac{s}{\sqrt{n}}$$

$$(1 - \alpha)100\% CI$$

$$\bar{x} - t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

Example

A random sample of 8 cigarettes of a certain brand has an average nicotine of content of 2.6mg and standard deviation of 0.9mg. construct the 99% confidence interval for the true average nicotine content of this particular brand of cigarette. clearly state an assumption.

solution

$$n = 8, \bar{x} = 2.6, s = 0.9$$

nature of the distribution of the population?
Assumption: the distribution of nicotine content is normally distributed

$$99\% CI, \alpha = 0.01$$

$$\alpha/2 = 0.005, df = 8 - 1 = 7$$

$$t_{0.005,7} = 3.499$$

$$2.6 - 3.499 \times \frac{0.9}{\sqrt{8}} < \mu < 2.6 + 3.499 \times \frac{0.9}{\sqrt{8}}$$

$$1.61 < \mu < 3.71$$

Confidence Interval for the Difference Between Two Population Means

- if independent samples of sizes n_1, n_2 are drawn from any population with μ_1, μ_2 and σ_1^2, σ_2^2 and the sampling distribution for the difference between the sample mean will be approximately normal for the $\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$ and $\sigma^2_{\overline{x_1}-\overline{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  case 1
  CI for $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known

$$P((\overline{x_1} - \overline{x_2}) - k < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + k) = 1 - \alpha, 0 < \alpha < 1$$

if the two population are normal by CLT, sampling distribution of $\overline{x_1} - \overline{x_2}$ is normal

$$\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$$

$$\sigma^2_{\overline{x_1}-\overline{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z_{\frac{\alpha}{2}} = \frac{(\mu_1 - \mu_2) + k - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$k = z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(1-\alpha)100\%$$

$$(\overline{x_1} - \overline{x_2}) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example

a sample of 150 brand of type A light bulbs showed a mean life of 1400hrs and standard deviation of 120hr. A sample of 200 brands of type B light bulbs showed a mean life of 1200hrs with a standard deviation of 80hrs. find

(a) 95% confidence interval

(b) 99% confidence interval for the difference between the true mean life times of the population brands of type A and B.

solution

Given

$$n_1 = 15, n_2 = 200, \overline{x_1} = 1400, \overline{x_2} = 1200, s_1 = 120, s_2 = 80$$

Because samples are large sampling distribution of $\overline{x_1} - \overline{x_2}$ is approximately normal

95% CI $\alpha = 0.05, z_{\frac{\alpha}{2}} = 1.96$

$$= (\overline{x_1} - \overline{x_2}) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$\mu_1 - \mu_2 = (1400 - 1200) \pm 1.96\sqrt{\frac{14400}{150} + \frac{6400}{200}}$$

$$177.8 < \mu_1 - \mu_2 < 222.2$$

99% CI

$$z_{\frac{\alpha}{2}} = 2.576$$

$$170.8 < \mu_1 - \mu_2 < 229.2$$

CI for $\mu_1 - \mu_2$ when σ_1, σ_2 are unknown $n_1 \leq 30, n_2 < 30$

Definition

Suppose s_1^2 and s_2^2 are sample variances resulting from independent random samples of sizes n_1 and n_2 arising from normal population x and y respectively have a common variance

$$s_p^2 = \frac{(n-1)s_1^2 + (n-1)s_2^2}{n_1 + n_2 - 2}$$

pooled sample variance

$$\frac{\Sigma_{i=1}^{n}(x_i - \overline{x_1})^2 + \Sigma_{i=1}^{n}(x_i - \overline{x_2})^2}{n_1 + n_2 - 2}$$

Theorem

pooled sample variance is unbiased estimator of the population variance σ^2

$$E(s_p^2) = \sigma^2$$

$$E\left[\frac{(n-1)s_1^2 + (n-1)s_2^2}{n_1 + n_2 - 2}\right]$$

$$\frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)E(s_1^2) + (n_2 - 1)E(s_2^2)\right]$$

$$= \frac{1}{n_2 + n_2 - 2}\left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}[n_1 - 1 + n_2 - 1]\sigma^2$$

$$= \sigma^2$$

Standard error for difference between population means

$$\sigma_{\overline{x_1} - \overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

two population have common variance

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

To get $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ we replace σ by s_p
we use a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Theorem

$(1-\alpha)100\%$ CI
for $\mu_1 - \mu_2$. when the two population have a common variance with $n_1 < 30, n_2 < 30$ is given by

$$\mu_1 - \mu_2 = (\overline{x_1} - \overline{x_2}) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example
Twelve randomly selected palm trees have a mean height of 13.8m with a standard deviation of 1.2m. 15 randomly selected palm trees of another variety have a mean height of 11.9m with a standard deviation of 1.5m. Assuming random samples were selected from normal populations with equal variances,. construct a 95% CI for the difference between two average heights of the two kinds of palm tress.
solution

$$n_1 = 12, n_2 = 15, \overline{x_1} = 13.8, \overline{x_2} = 11.9, s_1 = 1.2, s_2 = 1.5$$

sample sizes are small. samples taken from normal population having equal variances

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{11(1.44) + 14(2.25)}{12 + 15 - 2}} = 1.376$$

$$t_{\frac{\alpha}{2}, n_1+n_2-2} = t_{0.025,25} = 2.060$$

$$\mu_1 - \mu_2 = (\overline{x_1} - \overline{x_2}) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\mu_1 - \mu_2 = (13.8 - 11.9) \pm (2.060)(1.376)\sqrt{\frac{1}{12} + \frac{1}{15}}$$

$$= 1.9 \pm 1.098$$

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: Tests of hypothesis