Confidence interval for a population mean with known variance

Prerequisites

You should be familiar with (1) the central limit theorem and (2) application of the central limit theorem to test whether a sample mean is equal to a population mean.

Example (1)

The diameter X of widgets produced at a manufacturing company is expected to be normally distributed with mean 4.3 mm and variance 0.45 mm^2. The management wishes to test whether the machine producing these widgets needs recalibrating. A sample of size 15 has mean 3.88 mm. Formulate a suitable hypothesis and test at the 5% level whether the sample mean is has differed from the population mean of 4.3 mm.

Solution

Let $\mu =$ the true sample mean.

H_0 $\quad \mu = 4.3$

H_1 $\quad \mu \neq 4.3$ $\quad$ Two-tailed test

Significance level = $\alpha = 0.05$

As this is a two-tailed test the significance level for each tail is $\frac{0.05}{2} = 0.025$

$X \sim N(4.3,\ 0.45)$ then $\overline{X} \sim N\left(4.3, \frac{0.45}{15}\right)$ [By the central limit theorem]

The experimentally determined value of the sample mean is

$\overline{X}_{\text{test}} = 3.88$

$$z_{\text{test}} = \frac{3.88 - 4.3}{\sqrt{\left(\frac{0.45}{15}\right)}} = 2.422 \text{ (3 d.p.)} \qquad \Phi(2.422) = 0.9922$$

p-value $= P\left(\overline{X} < 3.88\right) = P(Z < -2.422) = 1 - 0.9922 = 0.0078$

p-value $= 0.0078 < 0.025 =$ significance level for each tail

Reject H_0, accept H_1

The machines do need recalibrating.

Systematic hypothesis testing

Suppose in example (1) that the manufacturing company regularly take samples of size 15 of their widgets as part of their quality control management processes. Then it is not necessary for them to determine on each occasion the *p*-value associated with the sample mean. The management is able to determine in advance an interval for the sample mean such that if the value of the sample mean falls inside this interval then the null hypothesis shall be accepted, otherwise it shall be rejected.

Example (1) continued

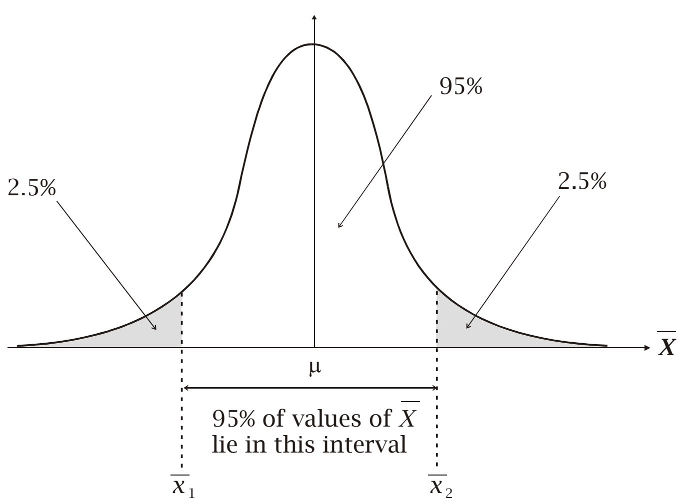
The diameter X of widgets produced at a manufacturing company is expected to be normally distributed with mean 4.3 mm and variance 0.45 mm². As part of a regular process of quality control the management of the company routinely take samples of widgets of sample size $n = 15$. Determine a two-tailed interval equivalent to a significance level of 5% for this sampling process such that the null hypothesis

$H_0 \qquad \mu = 4.3$

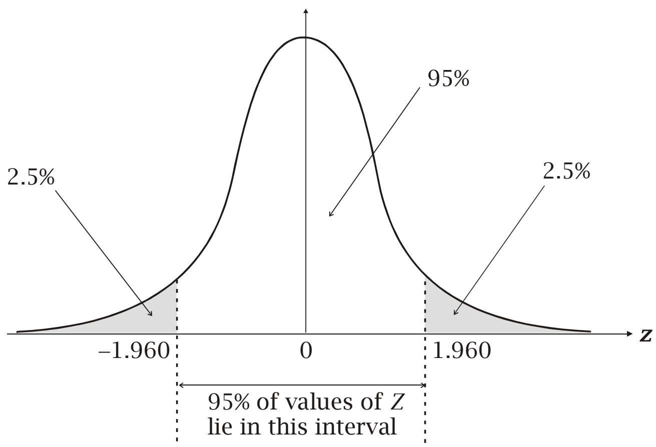
where μ is the true sample mean shall be accepted if the test value of the sample mean falls within this interval.

Solution

The following diagram illustrates the solution.



In this question the population mean is $\mu_0 = 4.3$. Assuming that the true sample mean is $\mu = 4.3$ we are seeking an interval $[\overline{x}_1, \overline{x}_2]$ such that there is a 95% probability that value of a randomly determined test sample mean $\bar{X}_{\text{test}}$ shall lie within it. In fact, we shall reject the hypothesis that the true sample mean is $\mu = 4.3$ if the test sample mean lies outside this interval. To find this interval we can establish the usual equivalence between this interval and the standard normal distribution. $Z \sim N(0,1)$



As the diagram indicates the values of Z that correspond to an interval with a probability of 95% are $z_1 = -1.960$ and $z_2 = 1.960$. This is because $\Phi(1.960) = 0.975$. We need to find the values of the sample mean $\bar{X}$ that correspond to these z-values. To do this we substitute into

$$z = \frac{\bar{X} - \mu}{\sqrt{\left(\frac{\sigma^2}{n}\right)}}$$

Here $\mu = 4.3$, $\sigma^2 = 0.45$ and $n = 15$

$$-1.960 = \frac{\bar{x}_1 - 4.3}{\sqrt{\left(\frac{0.45}{15}\right)}} \qquad 1.960 = \frac{\bar{x}_2 - 4.3}{\sqrt{\left(\frac{0.45}{15}\right)}}$$

$$\bar{x}_1 = 4.3 - 0.340 = 3.960 \qquad \bar{x}_2 = 4.3 + 0.340 = 4.460 \quad (3 \text{ d.p.})$$

The interval that we require is $3.960 \le \bar{X} \le 4.460 \quad (3 \text{ d.p.})$.

In this example, the manufacturers will conclude that the true sample mean is equal to the population mean if the test sample mean lies within this interval (the null hypothesis), and conclude that the true sample mean is not equal to the population mean (the alternative hypothesis) if the test sample mean lies outside this interval.

Remark

Throughout this section we are making use of distinctions between the following.

(1) The population mean, which we denote by μ_0

(2) The true sample mean, which we denote by μ

(3) The sample mean, which is a variable that we denote by $\bar{X}$.
It is based on the idea that we may repeatedly take samples of size *n*.

(4) A single instance of the variable $\bar{X}$ which we denote by $\bar{X}_{\text{test}}$

In this section we are given the population mean (μ_0) to start with. We are testing whether the true mean (μ) of a sample is drawn from this population or not. Therefore, we form the null hypothesis that the true sample mean is equal to the population mean

$H_0 \qquad \mu = \mu_0$

which forms the basis of the test. In this context we are testing this against an alternative hypothesis

$H_1 \qquad \mu \neq \mu_0$

which is a two-tailed test. In practice the test value is found from observational data. A particular sample of size *n* is taken and the value of the test mean $\bar{X}_{\text{test}}$ for this sample is calculated in the usual way as the average of all the observations in the sample. We have found that we can summarise the effect of repeated tests (of the same sample size) each producing a different test value by accepting the null hypothesis if the particular test value falls within a certain interval, and rejecting it otherwise.

Generalising further

The calculation of the interval in example (1) is clearly capable of generalisation. Given the background population $X \sim N(\mu_0, \sigma^2)$ the 95% interval for the sample mean is $\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$

What this says is that, given that the population really is $X \sim N(\mu_0, \sigma^2)$, then it is expected that the mean $(\bar{X})$ of 95% of samples of size *n* shall lie in this interval. We may apply this to hypothesis testing by further adding that if a particular test sample mean $\bar{X}_{\text{test}}$ does not lie in this interval, then we shall reject the null hypothesis at the 5% significance level

$H_0 \qquad \mu = \mu_0$

in favour of the alternative hypothesis

$H_1 \qquad \mu \neq \mu_0$

Example (2)

Given $X \sim N(3.52,\ 0.44)$ find a 95% interval for the sample mean where the sample size is $n = 20$. Give your answer to 3 significant figures.

Solution

$$\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$$

Here $\mu_0 = 3.52,\ \sigma^2 = 0.44$ and $n = 20$. Hence

$$3.52 - 1.96 \times \left(\sqrt{\frac{0.44}{20}}\right) \le \bar{X} \le 3.52 + 1.96 \times \left(\sqrt{\frac{0.44}{20}}\right)$$

$$3.23 \le \bar{X} \le 3.81 \quad (3 \text{ s.f.})$$

In the formula

$$\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$$

the value of 1.96 derives from the fact that it is the *z*-value corresponding to a two-tailed significance test of 5%. That is, the critical region of each tail of the test has probability 2.5% and $\Phi(1.960) = 0.975 = 97.5\% = 1 - 2.5\%$. Thus a different interval corresponds to each level of significance. The level of significance in the test corresponds to a *critical z-value*. The general formula is

$$\mu_0 - z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right) \le \bar{X} \le \mu_0 + z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

The critical *z*-values are found from the table of *z*-values in general to be as follows.

Confidence interval	75%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%
Significance level	25%	10%	5%	2.5%	1%	0.5%	0.25%	0.1%
p-value of tail $p = \Phi(z)$	0.125	0.05	0.025	0.0125	0.005	0.0025	0.00125	0.0005
z_{critical}	1.282	1.645	1.960	2.326	2.576	2.897	3.090	3.291

Example (3)

Given $X \sim N(67.35,\ 2.58)$ find a 99% interval for the sample mean where the sample size is $n = 100$. Give your answer to 2 decimal places.

Solution

The formula is

$$\mu_0 - z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right) \le \bar{X} \le \mu_0 + z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

The critical value corresponding to a 99% interval is $z_{\text{critical}} = 2.576$

Also $\mu_0 = 67.35$, $\sigma^2 = 2.58$ and $n = 100$. On substitution

$$67.35 - 2.576 \times \left(\sqrt{\frac{2.58}{100}}\right) \le \bar{X} \le 67.35 + 1.96 \times \left(\sqrt{\frac{2.58}{100}}\right)$$

$$66.94 \le \bar{X} \le 67.76 \quad (2 \text{ d.p.})$$

Estimating a population mean

In the previous section we were given a population $X \sim N\left(\mu_0, \sigma^2\right)$ and testing whether the mean of a sample was drawn from this population. In this process we start by being given the population mean, μ_0, and form a hypothesis on the basis of it. However, this process is somewhat artificial. In many practical situations we have no way of knowing with certainty what the population mean really is. In fact, we use data drawn from a sample to *estimate* the mean of the population. The natural way to estimate a population mean is to take the sample mean as an *estimator* for it. It can be shown also that this is the theoretically correct approach to adopt.

Example (4)

A population with variable X is thought to be normally distributed such that $X \sim N\left(\mu, \sigma^2\right)$.

The results of 10 observations of X are given in the following table.

Observation, i	1	2	3	4	5	6	7	8	9	10
Value, x_i	6.5	6.3	7.1	8.5	5.9	9.6	8.2	7.8	7.9	7.3

Find an estimate for the population mean, μ.

Solution

The value of the sample mean is

$$\bar{x}_{\text{sample}} = \frac{\text{Sum of all the values}}{\text{sample size}} \qquad \left[\bar{x}_{\text{sample}} = \frac{\sum x_i}{n}\right]$$

$$= \frac{6.5+6.3+7.1+8.5+5.9+9.6+8.2+7.8+7.9+7.3}{10}$$

$$= \frac{75.1}{10}$$

$$= 7.51$$

We shall use this value as an estimate of the population mean, $\hat{\mu}_0$. Hence

$$\hat{\mu} = \bar{x}_{\text{sample}} = 7.51$$

Remarks

(1) Here we introduce another symbol $\hat{\mu}$ to stand for the estimator of the population mean. The estimator of the population mean is the mean of the sample

$$\hat{\mu} = \bar{X}$$

(2) From the sample data given in example (4) we could have also calculated the sample variance. However, in this chapter the population variance shall always we given in advance, and we shall not attempt to estimate the population variance. This is not only a case of taking things one-step at a time, for in fact it would be a theoretical mistake to use the sample variance as an estimator of the population variance. So estimating population variance properly belongs to subsequent chapter.

Confidence interval for a sample mean of known variance

When we are estimating a true population mean we would like also an indication of how exact the estimate is. We can never say for certainty that an estimate is exact. Furthermore, neither could we establish an interval for the true population mean such that the true population mean must lie with certainty in this interval. The best that we can do is to establish a *confidence interval* for the true population mean based on the estimate drawn from the sample mean. Each confidence interval shall be associated with a measure of probability. For instance, a 95% confidence interval for a population mean shall be an interval such that there is a 95% probability that the true population mean lies in this interval. The converse of this is that there is a 5% probability that the true population mean does **not** lie in this interval.

We use the same table as before to determine the critical z-values that correspond to each confidence level.

Confidence interval	75%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%
Significance level	25%	10%	5%	2.5%	1%	0.5%	0.25%	0.1%
p-value of tail $p=\Phi(z)$	0.125	0.05	0.025	0.0125	0.005	0.0025	0.00125	0.0005
z_{critical}	1.282	1.645	1.960	2.326	2.576	2.897	3.090	3.291

Denote the true population mean by μ and the background distribution by $X \sim N(\mu,\sigma^2)$. The variance, σ^2, shall be given. A sample of size n is taken. We denote the sample mean by $\bar{X}_{\text{sample}}$ which we use as an estimator of the population mean. This estimator is denoted by $\hat{\mu} = \bar{X}_{\text{sample}}$. Then a 95% confidence interval for μ_0 is

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}.$$

In general, the confidence interval associated with a critical z-given value is given by

$$\hat{\mu} - z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \hat{\mu} + z_{\text{critical}} \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

Example (5)

Twelve successive train times (in minutes) between the same two British cities were recorded as follows

63.5 78.9 65.7 61.2 76.8 72.5 77.1 81.2 64.9 71.3 73.4 75.8

Assume that these observations form a random sample from a normal distribution with standard deviation 6.

(*a*) Calculate a 95% confidence interval for the mean time taken by the trains between the two cities. Give your answer to 1 decimal place.

(*b*) An almanac states that the mean time is 70 minutes. State with a reason whether or not your results support this belief.

Solution

(*a*) The value of the sample mean is

$$\bar{x}_{\text{sample}} = \frac{\sum x_i}{n} = \frac{862.3}{12} = 71.86 \quad (4 \text{ s.f.})$$

The estimate of the population mean is

$\hat{\mu} = \bar{x}_{\text{sample}} = 71.86 \quad (4 \text{ s.f.})$

A 95% confidence interval is given by

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}$$

Here $\sigma = 6$ and $n = 12$. Hence

$$71.86 - 1.96 \times \frac{6}{\sqrt{12}} \le \mu \le 71.86 + 1.96 \times \frac{6}{\sqrt{12}}$$

$68.5 \le \mu \le 75.3$ minutes (1 d.p.)

(*b*) The value 70 lies within the 95% confidence interval established in part (*a*). Therefore, there is no evidence at the 5% significance level to reject the hypothesis that the true mean is given by $\mu_0 = 70$.

Remarks

(1) In this last example the claim by the almanac may be based on other data or an alternative theoretical justification. For example, trains may be scheduled to take 70 minutes between the two cities. On the other hand, if the only data available for estimating the mean time of the train journey is the data provided in this example, then the best estimator for that mean time is $\hat{\mu} = \bar{X}_{\text{sample}}$ which in this case has value $\bar{x}_{\text{sample}} = 71.9$ minutes (1 d.p.) and not 70 minutes.

(2) In the preceding section we made a distinction between

(i) The true population mean, which we denote by μ_0

(ii) The true sample mean, which we denote by μ

There we were specifically testing the hypothesis that the true sample mean had deviated from the population mean. Therefore, we needed to distinguish between the two concepts and needed two symbols. In this section the population mean is **not given**. It is assumed that the true sample mean is the same as the population mean and we in fact use the test sample mean as an estimate for both. Therefore, in this section $\mu_0 = \mu$ and only one symbol is required.

Example (6)

The weights in kilograms of ten pheasants brought to a market in one day by local gamekeepers are given as follows.

1.25 1.38 0.95 1.68 1.50 1.62 1.39 1.05 1.43 1.38

(*a*) Assuming that this is a random sample from a normal distribution with mean μ and standard deviation 0.2 calculate a 90% confidence interval for μ.

(*b*) How many observations would be required in order to halve the width of this confidence interval?

Solution

(*a*) $\bar{x}_{\text{sample}} = \frac{\sum x_i}{n} = \frac{13.63}{10} = 1.363 \quad (4 \text{ s.f.})$

The estimate of the population mean is

$\hat{\mu} = \bar{x}_{\text{sample}} = 1.363 \quad (4 \text{ s.f.})$

A 90% confidence interval is given by

$$\hat{\mu} - 1.645\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + 1.645\frac{\sigma}{\sqrt{n}}$$

Here $\sigma = 6$ and $n = 12$. Hence

$$1.363 - 1.645 \times \frac{0.2}{\sqrt{10}} \le \mu \le 1.363 + 1.645 \times \frac{0.2}{\sqrt{10}}$$

$$1.363 - 0.104... \le \mu \le 1.363 + 0.104...$$

$$1.26 \le \mu_0 \le 1.47 \text{ kg} \quad (2 \text{ d.p.})$$

(*b*) The interval of the confidence interval in part (*a*) is $2 \times 0.104...$

Half the width of the interval is 0.104..., and

$$1.645 \times \frac{\sigma}{\sqrt{n}} = \frac{1}{2} \times 1.645 \times \frac{0.2}{\sqrt{10}}$$

$$\sqrt{n} = 2\sqrt{10}$$

$$n = 40$$

Summary

95% confidence interval for the sample mean.

Given

(1) The population mean μ_0 and population variance σ are **both known**.

(2) A sample of size n is to be taken.

Then a 95% confidence interval for the sample mean $\bar{X}$ is

$$\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$$

Hypothesis test of the sample mean at the 5% significance level

If it is suspected that the true sample mean μ is not the same as the population mean μ_0 then the null hypothesis is

$H_0 \qquad \mu = \mu_0$

which is tested against the two-tailed alternative hypothesis

$H_1 \qquad \mu \neq \mu_0$

Then the null hypothesis shall be accepted at the 5% significance level if the test sample mean $\bar{X}_{\text{test}}$ lies in the interval

$$\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}_{\text{test}} \le \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}$$

95% confidence interval for the population mean

Given

(1) The population mean μ is **unknown** and population variance σ is **known**.

(2) A sample of size n is taken. The estimator of the population mean is the sample mean $\hat{\mu} = \bar{X}$

Then a 95% confidence interval for the population mean is

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}$$