

**Topic#1: Sampling distribution and confidence interval estimation**

**Definitions**

**Population**: the totality of items or individuals under consideration. A population can be *finite* (can be counted) or *infinite* (can not be counted).

**Sample**: portion of the population that is selected for the analysis. A sample is characterized by its size $n$, which refers to the number of observations, items composing the sample. The size of a sample is always lower than the size of the population.

**Parameter**: summary measure that is computed to describe a characteristic of the entire population.

**Statistic**: summary measure used to approximate a parameter. A statistic is derived from a sample.

**Sampling Distributions of the Sample Mean:**

The mean $\bar{x}$ has a distribution defined as follows:

$$\mu_x = \mu \text{ and } \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

**Central Limit Theorem**: states that the sampling distribution of the mean is approximately normal for large $n$ (n is the sample size) or for approximately normally distributed populations with mean $\mu$ and variance $\sigma^2/n$.

**Statistical Inference**: making estimations about the true parameters of the populations using sample data

**Estimation**: determining the approximate value of a population parameter on the basis of a sample statistic.

**Point estimator**: draw inferences about a population by estimating the value of an unknown parameter using a *single point or value*.

**Interval estimator**: draw inferences about a population by estimating the value of an unknown parameter by using an *interval*. This interval, called *confidence interval*, is derived using a certain level of confidence.

**Significance level**: given by $\alpha$, margin of error allowed for the estimation or for the hypothesis test; i.e. $\alpha=5\%$ implies that up to 5% of the times the true of the population $\mu$ would lie outside the confidence interval.

**Level of confidence for the estimation**: given by $(1-\alpha)$, measure the probability that the parameters of the population lies between the *upper and lower bounds* of the confidence interval.

**Lower Confidence Limits (LCL)**: is the lower bound of the confidence interval.

**Upper Confidence Limits (UCL)**: is the upper bound of the confidence interval.

**Bootstrapping**: technique that aims at obtaining better estimation of the parameters of the populations by drawing an initial sample of size $n$ out of a population frame of size $N$ ($n<<N$) and repeatedly resampling it $m$ different times ($m=100$ or $m=1000$) by replacing the $n$ observations in the original sample by $n$ observations from the population frame.
Confidence intervals estimations for the Population Mean

Case 1: standard deviation of population known

\[ UCL = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad LCL = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]

Consequently the \((1-\alpha)\) confidence interval estimate is

\[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \]

The critical values \(z_{\alpha/2}\) are given by the table for the standard normal distribution. Typically \(\alpha\) is set either to 10%, 5%, 1%. The corresponding critical values are

<table>
<thead>
<tr>
<th>Level of significance (\alpha)</th>
<th>Confidence level ((1-\alpha))</th>
<th>(\alpha/2)</th>
<th>(z_{\alpha})</th>
<th>(z_{\alpha/2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>90%</td>
<td>5%</td>
<td>1.282</td>
<td>1.645</td>
</tr>
<tr>
<td>5%</td>
<td>95%</td>
<td>2.5%</td>
<td>1.645</td>
<td>1.96</td>
</tr>
<tr>
<td>1%</td>
<td>99%</td>
<td>.5%</td>
<td>2.32</td>
<td>2.57</td>
</tr>
</tbody>
</table>

Thus, the 95% confidence interval for the mean is

\[ \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \]

meaning that for the standardized variable \(z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\)

\[ \text{Prob} \left( -z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right) = 1 - \alpha \]

Case 2: standard deviation of population unknown but sample standard deviation available

In this case, the true variance is approximated with the standard deviation of the sample \(s\).

\[ UCL = \bar{x} + t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \quad \text{and} \quad LCL = \bar{x} - t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \]

Consequently the \((1-\alpha)\) confidence interval estimate is

\[ \bar{x} - t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \]

The critical value \(t_{\alpha/2,n-1}\) is given by the table for the Student’s distribution. Unlike the normal distribution, the values of the Student’s distribution depends upon the degree of freedom \((n-1)\). Again, \(\alpha\) is set either to 10%, 5%, 1%.

Thus, assuming \(n=30\), the 95% confidence interval for the mean is
\[
\bar{x} - t_{0.025,29} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{0.025,29} \frac{s}{\sqrt{n}}
\]

with \( t_{0.025,29} = 2.0452 \) (look in the table to degree of freedom=29 and upper tail area=.025).

Finally, for the standardized variable \( t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \)

\[
\text{Prob} \left( -t_{\nu,n-1} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{\nu,n-1} \right) = 1 - \alpha
\]

**Case 3: population and sample standard deviation both not available**

In this case, approximately set the sample standard deviation to the range divided by 4.

\[
s \equiv \frac{\text{Range}}{4} \equiv \frac{(\text{Max} - \text{Min})}{4}
\]

Then proceed as in Case 2.

**Confidence interval estimations for an individual observation**

Assume now that we want to generate a confidence interval estimate for an individual observation (versus the mean of the whole population). All of the above apply with the single distinction that \( \sqrt{n} \) is removed from the formula.

Thus, if \( \sigma \) is known the confidence interval is

\[
\bar{x} - z_{\alpha/2} \sigma \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma
\]

Thus, if \( \sigma \) is unknown the confidence interval is

\[
\bar{x} - t_{\alpha/2,n-1} s \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} s
\]

**Confidence interval estimation for the proportion**

Let \( \pi \) be the true population proportion having a specific feature.

To establish a confidence interval for the unknown parameter \( \pi \), let \( p, X, \) and \( n \) be the ratio of the number of successes in the sample, the number of successes, and the sample size respectively.

\[
p = \frac{X}{n}
\]

The confidence interval for \( \pi \) is

\[
p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}
\]