## 0.1 Confidence Interval For The Mean

Suppose that you wish to estimate the mean sales amount perretail outlet for a particular consumer product during the past year. The number of retail outlets is large. Determine the 95 percent confidence interval given that the sales amounts are assumed to be normally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.
Ans. $3; 346 : 60to3$;503:40

Determine the 95 percent confidence interval given that the population is assumed to benormally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.
Ans. $3; 342 : 44to3$;507:56

## 0.2 Confidence Intervals for Means

The structure of a confidence interval for the mean is as follows:

Sample Mean ± (Quantile × Std. Error)

**Quantiles:**
For large samples (i.e. greater than 30) where a normal distribution can be assumed, the quantiles are as follows

| | |
|---|---|
| 90% | 1.645 |
| 95% | 1.96 |
| 99% | 2.576 |

**Sample Mean:**
The sample mean $\bar{x}$ is usually given in the question.
(Remark: Sample mean is a type of ***point estimate***).

**Standard Error :**
The standard error is computed using the sample standard deviation ($s$) and the sample size ($n$).

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}}$$

## 0.3 Confidence Intervals: Example

- The length of life of a type of battery is estimated from a sample of 100 test items taken from a large population.
- Sample results show that the mean length of life is 57.4 hours with a standard deviation of 15.1 hours.
- Construct a 95% confidence interval for the mean length of life of all of these batteries.

With a sample standard deviation of 15.1 and a sample size of 100, the standard error is as follows:

$$S.E.(\bar{x}) = \frac{15.1}{\sqrt{100}}$$

With a sample standard deviation of 15.1 and a sample size of 100, the standard error is as follows:

$$57.4 \pm (1.96 \times 1.51)$$

### Confidence Intervals (Revision)

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.
- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%, 99% and 99.9%.

### Confidence Level (Revision)

- A confidence level for an interval is denoted to $1 - \alpha$ (in percentages: $100(1 - \alpha)\%$) for some value $\alpha$.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1 - \alpha)\% = 100(1 - 0.05)\% = 100(0.95)\% = 95\%$
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for $\alpha$ is important when determining quantiles.

### Using the $t-$distribution for large samples

- The $t-$distribution is used for computing quantiles in the case of small samples (i.e. when sample size $n \leq 30$).
- A key value in the $t-$distribution is the degrees of freedom, denoted *df* (or sometimes $\nu$). For small samples

  $$df = n - 1$$

  .
- The $t-$distribution is used for computing quantiles in the case of large samples too, as an alternative to using the $Z$ distribution.
- In this case , use the value $\infty$ as the degrees of freedom (see bottom row of table 7).

  $$df = \infty$$
- This means that we can use the $t-$ distribution for finding the quantiles of all types of confidence intervals.

### Small samples

- We indicated that use of the normal distribution in estimating a population mean is warranted for any large sample ($n > 30$).
- For a small sample ($n \leq 30$) only if the population is normally distributed **and** $\sigma$ is known, the standard normal distribution can be used compute quantiles. In practice, this case is unusual.
- Now we consider the situation in which the sample is small and the population is normally distributed, but $\sigma$ is not known.

**Student's $t-$distribution (1)**

- Student's $t-$distribution is a variation of the normal distribution, designed to factor in the increased uncertainty resulting from smaller samples.al
- The distribution is really a family of distributions, with a somewhat different distribution associated with the degrees of freedom (*df*). For a confidence interval for the population mean based on a sample of size n, $df = n - 1$.

**Student's $t-$distribution (2)**

- With increasing sample size, the $t-$distribution approaches the form of the standard normal ('Z') distribution.
- In fact the standard normal distribution can be thought of as the $t-$distribution with $\infty$ degrees of freedom.
- For computing quantiles, we will consider the 'Z' distribution in this way.
- For values of $n$ greater then 30, the difference between using $df = n - 1$ and $df = \infty$ is negligible.
- As this will be relevant later, remember that a confidence interval is a **two-tailed** procedure, i.e. $k = 2$.

[fragile] **Student's $t-$distribution (3)**

- Student's t- values are determined using the `t` family of commands (e.g. `qt, pt, dt`).
- To compute quantiles, use the code below.
- The degrees of freedom must be additionally be specified. Degrees of freedom are computed as sample size minus one ($n - 1$)
- As the degrees of freedom gets larger and larger, the student t distribution converges to the Z distribution.

**Confidence Interval for a Mean (Small Sample)**

- The mean operating life for a random sample of $n = 10$ light bulbs is $\bar{x} = 4,000$ hours, with the sample standard deviation $s = 200$ hours.
- The operating life of bulbs in general is assumed to be approximately normally distributed.
- We estimate the mean operating life for the population of bulbs from which this sample was taken, using a 95 percent confidence interval as follows:

$$4,000 \pm (2.262)(63.3) = (3857, 4143)$$

- The point estimate is 4,000 hours. The sample standard deviation is 200 hours, and the sample size is 10. Hence

$$S.E(\bar{x}) = \frac{200}{\sqrt{10}} = 63.3$$

- From last slide, the t quantile with $df = 9$ is 2.262.

**Independent Samples (New Section)**

- Two samples are referred to as independent if the observations in one sample are not in any way related to the observations in the other.
- This is also used in cases where one randomly assign subjects to two groups, i.e. in give first group treatment A and the second group treatment B and compare the two groups.
- Often we are interested in the difference between the mean value of some parameter for both groups.
- SE = $\sqrt{[p_1 \times (1 - p_1)/n_1] + [p_2 \times (1 - p_2)/n_2]}$
- SE = $\sqrt{[0.40 \times 0.60/400] + [0.30 \times 0.70/300]}$
- SE = $\sqrt{[(0.24/400) + (0.21/300)]} = \sqrt{(0.0006 + 0.0007)} = \sqrt{0.0013} = 0.036$

$$(\bar{X} - \bar{Y}) \pm \left[\text{Quantile } \times S.E(\bar{X} - \bar{Y})\right]$$

- If the combined sample size of X and Y is greater than 30, even if the individual sample sizes are less than 30, then we consider it to be a large sample.
- The quantile is calculated according to the procedure we met in the previous class.
- Assume that the mean ($\mu$) and the variance ($\sigma$) of the distribution of people taking the drug are 50 and 25 respectively and that the mean ($\mu$) and the variance ($\sigma$) of the distribution of people not taking the drug are 40 and 24 respectively.

**Difference in Two means** For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ().

$$S.E.(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- $s_x^2$ and $s_x^2$ is the variance of both samples.
- $n_x$ and $n_y$ is the sample size of both samples.

The degrees of freedom is $n_x + n_y - 2$.

**CI for Proportion: Example (1)**

- $\hat{p} = 0.62$

- Sample Size $n = 250$
- Confidence level $1 - \alpha$ is 95%

**CI for Proportion: Example (2)**

- First, lets determine the quantile.
- The sample size is large, so we will use the Z distribution.
- (Alternatively we can uses the $t-$ distribution with $\infty$ degrees of freedom.

Although the sample mean is useful as an unbiased estimator of the population mean, there is no way of expressing the degree of accuracy of a point estimator. In fact, mathematically speaking, the probability that the sample mean is exactly correct as an estimator of the population mean is $P = 0$.

A confidence interval for the mean is an estimate interval constructed with respect to the sample mean by which the likelihood that the interval includes the value of the population mean can be specified.

The *level of confidence* associated with a confidence interval indicates the long-run percentage of such intervals which would include the parameter being estimated.

- Confidence intervals for the mean typically are constructed with the unbiased estimator $\bar{x}$ at the midpoint of the interval.
- The $\pm Z\sigma_x$ or $\pm Zs_x$ frequently is called the ***margin of error*** for the confidence interval.

We indicated that use of the normal distribution in estimating a population mean is warranted for any large sample ($n > 30$), **and** for a small sample ($n \leq 30$) only if the population is normally distributed and $\sigma$ is known.

- Now we consider the situation in which the sample is small and the population is normally distributed, but $\sigma$ is not known.
- The distribution is a family of distributions, with a somewhat different distribution associated with the degrees of freedom (*df*). For a confidence interval for the population mean based on a sample of size n, $df = n - 1$.

**Computing the Standard Error**

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{\hat{(p)} \times (100 - \hat{p})}{n}}$$

---

$$\hat{p} = 144/200 \times 100\% = 0.72 \times 100\%. = 72$$

$100\% - \hat{p} = 100\% - 72\% = 28\%$

**Computing the Standard Error**

$$S.E.(\hat{p}) \;=\; \sqrt{\frac{72 \times 28}{200}}$$

**Difference in proportions** This lesson describes how to construct a confidence interval for the difference between two sample proportions, p1 - p2. **Estimation Requirements** The approach described in this lesson is valid whenever the following conditions are met:

- Both samples are simple random samples.
- The samples are independent.
- Each sample includes at least 10 successes and 10 failures.
- The samples comprises less than 10% of their respective populations.

**Standard Error for Difference of Proportions**

$$S.E.(\hat{P}_1 - \hat{P}_2) = \sqrt{[P_1 \times (1 - P_1)/n_1] + [P_2 \times (1 - P_2)/n_2]}$$

- $\hat{P}_1$ and $\hat{P}_2$ are the sample proportions of groups 1 and 2 respectively.
- $n_1$ and $n_2$ are the sample sizes of groups 1 and 2 respectively.

N.B. This formula will be provided in the exam paper.

- SE $= \sqrt{[p_1 \times (1 - p_1)/n_1] + [p_2 \times (1 - p_2)/n_2]}$
- SE $= \sqrt{[0.40 \times 0.60/400] + [0.30 \times 0.70/300]}$
- SE $= \sqrt{[(0.24/400) + (0.21/300)]} = \sqrt{(0.0006 + 0.0007)}$ = sqrt(0.0013) = 0.036

$$(\bar{X} - \bar{Y}) \pm \left[\text{Quantile } \times S.E(\bar{X} - \bar{Y})\right]$$

- If the combined sample size of X and Y is greater than 30, even if the individual sample sizes are less than 30, then we consider it to be a large sample.
- The quantile is calculated according to the procedure we met in the previous class.
- Assume that the mean ($\mu$) and the variance ($\sigma$) of the distribution of people taking the drug are 50 and 25 respectively and that the mean ($\mu$) and the variance ($\sigma$) of the distribution of people not taking the drug are 40 and 24 respectively.

**Difference in Two means** In order to construct a confidence interval, we are going to make three assumptions:

- The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
- The populations are normally distributed.
- Each value is sampled independently from each other value.

**Difference in Two means** For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ().

$$S.E.(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- $s_x^2$ and $s_x^2$ is the variance of both samples.
- $n_x$ and $n_y$ is the sample size of both samples.

The degrees of freedom is $n_x + n_y - 2$.

**Difference in Two means**

| Group | sample size | mean | variance |
|---|---|---|---|
| X | 17 | 5.353 | 2.743 |
| Y | 17 | 3.882 | 2.985 |

- Point estimate : $\bar{x} - \bar{y} = 1.4699$
- Standard Error: 0.5805
- Quantile : 1.96

$$1.4699 \pm (1.96 \times 0.5805) = (0.33212, 2.60768)$$

This analysis provides evidence that the mean for Y is higher than the mean for X, and that the difference between means in the population is likely to be between 0.29 and 2.65.