ST2334 - Chapter 6 - Estimation Based on Normal Distribution

6.1 - Point Estimation of Mean and Variance

6.2 - Interval Estimation

6.3 - Confidence Intervals for the Mean

6.4 - Confidence Intervals for the Difference between Two Means

6.5 - C.I. for Variances and Ratio of Variances

6.1 – Point Estimation of Mean and Variance

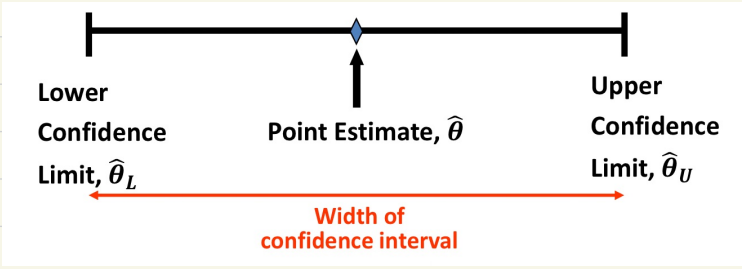
- Estimation of unknown parameter θ in $f_X(x;\theta)$ can be made in two ways: point estimation and interval estimation
- Point estimation: let the value of some statistic, say $\hat{\Theta} = \hat{\Theta}(X_1, X_2, \ldots, X_n)$ to estimate the unknown parameter
- Point estimator: the statistic $\hat{\Theta} = \hat{\Theta}(X_1, X_2, \ldots, X_n)$
- Statistic: a function of the random sample which does not depend on any unknown parameters
- Interval estimation: to define two statistics, say, $\hat{\Theta}_L$ and $\hat{\Theta}_U$ where $\hat{\Theta}_L < \hat{\Theta}_U$ so that $(\hat{\Theta}_L, \hat{\Theta}_U)$ constitutes a random interval for which the probability of containing the unknown parameter θ can be determined
- **Definition 6.2**: A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if

(Unbiased estimator)

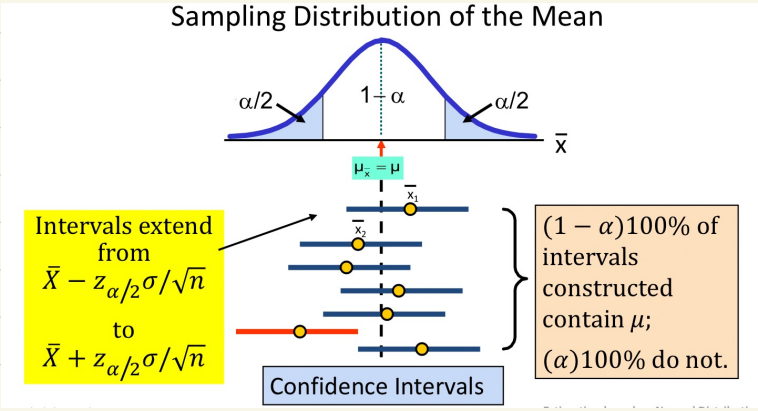
$$E(\hat{\Theta}) = \theta$$

6.2 - Interval Estimation

- An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_u$, where $\hat{\theta}_L$ and $\hat{\theta}_u$ depend on:
  1. the value of the statistic $\hat{\Theta}$ for a particular sample
  2. the sampling distribution of $\hat{\Theta}$



- We shall seek a random interval $(\hat{\Theta}_L, \hat{\Theta}_u)$ containing θ with a given probability $1-\alpha$, i.e. $Pr(\hat{\Theta}_L < \theta < \hat{\Theta}_u) = 1-\alpha$
- Then the interval $\hat{\theta}_L < \theta < \hat{\theta}_u$, computed from the selected sample is called a $(1-\alpha)100\%$ confidence interval for θ
- Confidence coefficient / degree of confidence: the fraction $(1-\alpha)$
- Lower and upper confidence limits: the end points $\hat{\theta}_L$ and $\hat{\theta}_u$ respectively
- If samples of the same size n are taken, in the long run, $(1-\alpha)100\%$ of the intervals will contain the unknown parameter θ
- With a confidence of $(1-\alpha)100\%$, we can say that the interval covers θ



6.3 - Confidence Intervals for the Mean

6.3.1 - Known variance case

- Confidence interval for mean with

① known variance

② the population is normal / n is sufficiently large (say $n \geq 30$)

- When population is normal / by Central Limit Theorem, we can expect that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. $\therefore$ $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Thus $Pr\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$

  or $Pr\left(\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha$
- If $\bar{X}$ is the mean of a random sample of size n from a population with known variance σ^2, a $(1-\alpha)\,100\%$ confidence interval for μ is given by:

  $$\left(\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$$
- Most of the time, $\bar{X}$ will not be exactly equal to μ and the point estimate is in error
- Size of error : $|\bar{X}-\mu|$
- $Pr\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \longrightarrow Pr\left(|\bar{X}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$
- Margin of error : e
- We want the error $|\bar{X}-\mu|$ to not exceed margin error e with probability larger than $1-\alpha$, i.e. $Pr(|\bar{X}-\mu| \leq e) \geq 1-\alpha$
- Since $Pr\left(|\bar{X}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$, $\therefore$ $e \geq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\therefore$ For a given margin of error e, the sample size is given by:

  $$n \geq \left(z_{\alpha/2}\frac{\sigma}{e}\right)^2$$

6.3.2 – Unknown variance case

- Confidence interval for mean with

① unknown population variance

② population is normal / very close to a normal distribution

③ sample size is small

- Let $T = \frac{(\bar{X}-\mu)}{S/\sqrt{n}}$ where S^2 is the sample variance
- We know $T \sim t_{n-1}$
- $\therefore \Pr(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}) = 1-\alpha$

  or $\Pr(-t_{n-1;\alpha/2} < \frac{(\bar{X}-\mu)}{S/\sqrt{n}} < t_{n-1;\alpha/2}) = 1-\alpha$

  or $\Pr(-t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}}) < \bar{X}-\mu < t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}})) = 1-\alpha$

  or $\Pr(\bar{X} - t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}}) < \mu < \bar{X} + t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}})) = 1-\alpha$
- If $\bar{X}$ and S are the sample mean and standard deviation of a random sample size of $n < 30$ from an approximate normal population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by:

  $$\bar{X} - t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}}) < \mu < \bar{X} + t_{n-1;\alpha/2}(\frac{S}{\sqrt{n}})$$
- For large n (say $n > 30$), the t-distribution is approximately the same as the $N(0,1)$ distribution. $\therefore$ when

① σ^2 is unknown

② population is normal

③ $n > 30$,

a $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\bar{X} - z_{\alpha/2}(\frac{S}{\sqrt{n}}) < \mu < \bar{X} + z_{\alpha/2}(\frac{S}{\sqrt{n}})$$

6.4 - Confidence Intervals for the Difference between Two Means

- If we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, then $\bar{X}_1 - \bar{X}_2$ is the point estimator of $\mu_1 - \mu_2$

6.4.1 - Known variances

- σ_1^2 and σ_2^2 are known and not equal, and the two populations are normal, or σ_1^2 and σ_2^2 are known and not equal, but n_1, n_2 are sufficiently large ($n_1 \geq 30$, $n_2 \geq 30$)
- According to Section 5.5, we have $(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
- We can assert that

$$Pr\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha$$

- which leads to the following $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

6.4.2 - Large sample C.I. for unknown variances

- σ_1^2 and σ_2^2 are unknown
- n_1, n_2 are sufficiently large ($n_1 \geq 30$, $n_2 \geq 30$)
- We may replace σ_1^2 and σ_2^2 by their estimates, S_1^2 and S_2^2
- A $(1-\alpha)100\%$ confidence interval for $(\mu_1 - \mu_2)$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}}$$

6.4.3 – Unknown but equal variances

- σ_1^2 and σ_2^2 are unknown but equal and the two populations are normal
- Small sample sizes ($n_1 \leq 30$ and $n_2 \leq 30$)
- Let $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

- $\therefore$ We obtain a standard normal variable in the form

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- σ^2 can be estimated by the pooled sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

with S_1^2 and S_2^2 being the sample variances of the first and second samples respectively

- Note that if the two populations are normal with the same variance σ^2, then

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{and} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\therefore \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

- Substituting S_p^2 for σ^2, we obtain the statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

- We can assert that

$$Pr\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2;\alpha/2}\right) = 1 - \alpha$$

- $\therefore$ A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where S_p is the pooled estimation of the population standard deviation and $t_{n_1+n_2-2;\alpha/2}$ is the value from the t-distribution with the degrees of freedom $n_1 + n_2 - 2$, leaving an area of $\alpha/2$ to the right

[i.e. $Pr(W > t_{n_1+n_2-2;\alpha/2}) = \alpha/2$ where $W \sim T_{n_1+n_2-2}$.]

- Note that for large samples s.t. $n_1 \geq 30$ and $n_2 \geq 30$, we can replace $t_{n_1+n_2-2;\alpha/2}$ by $z_{\alpha/2}$ in the previous formula
- $\therefore$ A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

6.4.4 - C.I. for the difference between two means for paired (dependent) data

- We must consider the differences $d_i \ (= x_i - y_i)$ of paired observations
- These differences are the values of a random sample $d_1, d_2, \cdots, d_n$ from a population that we shall assume to be normal with mean μ_D and unknown variance σ_D^2
- $\mu_D = \mu_1 - \mu_2$ and the point estimate of μ_D is given by:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)$$

- The point estimate of σ_D^2 is given by:

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$$

- A $(1-\alpha)100\%$ confidence interval for μ_D can be established by writing:

$$Pr(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}) = 1 - \alpha, \text{ where } T = \frac{\bar{d} - \mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$$

- A $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by:

$$\bar{d} - t_{n-1;\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right) < \mu_D < \bar{d} + t_{n-1;\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right)$$

- For sufficiently large sample, we may replace $t_{n-1;\alpha/2}$ by $z_{\alpha/2}$
- A $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by:

$$\bar{d} - z_{\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right) < \mu_D < \bar{d} + z_{\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right)$$

6.5 - C.I. for Variances and Ratio of Variances

6.5.1 - Confidence intervals for a variance (of a normal population)