Lecture - 12 : Interval Estimation

Section 1: Confidence Intervals

What we have learned till now in point estimation. In order to estimate the value of a population parameter, we just draw a random sample of a reasonable size and construct a statistic which can best estimate that population parameter, in some sense, either as unbiased or as the best maximum likelihood estimator. But if we are drawing a sample from a probability density function the probability that a value of the statistic will coincide with population parameter is zero. This motivates us to take the view that it might be better to be able to construct an interval around the statistic so that the population parameter value falls in that interval with certain probability.

Let us draw a sample from the population with pdf $f_X(x,\theta)$, and let $X_1 \ldots X_n$ be a random sample of sizes n drawn from it, and let Θ be a statistic, $\Theta = \Theta(X_1 \ldots X_n)$ constructed to estimate, θ. We may for example want to know, if

$$\Theta - c < \theta < \Theta + c$$

with probability say γ; i.e.

$$P(\Theta - c < \theta < \Theta + c) = \gamma$$

Then $(\Theta - c, \Theta + c)$ forms a confidence interval associated with θ. This is more formally defined as follows.

Definition 1: If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are values of the random variables themselves statistics, such that $\hat{\Theta}_1, \hat{\Theta}_2$, which are also

$$P(\hat{\Theta}_1 < \alpha < \hat{\Theta}_2) = 1 - \alpha$$

for some specified probability $1 - \alpha$. We shall refer to the interval $(\hat{\theta}_1, \hat{\theta}_2)$ as a $(1-\alpha)100\%$ confidence interval for the population parameter θ. The value $1 - \alpha$ is called degree of confidence & θ_1 & θ_2 the upper and lower confidence limits.

limits. So if $\alpha = 0.05$, we say the degree of confidence is 0.95 and we get a 95% confidence interval. It is also important to note at the very beginning that the confidence intervals for θ need not be unique as we will just see.

Section 2: Interval Estimation of means

In this section we will study first the interval estimation of the mean μ of a normal population whose variance σ^2 is known. Next we will see how to deal the case when σ^2 is unknown. It will be important to note how $\hat{\Theta}_1$ and $\hat{\Theta}_2$ will be very different in these two scenarios.

Case 1: σ^2 is known.

Let $X_1, \ldots X_n$ be a random sample of size n drawn from a population $N(\mu, \sigma^2)$, where σ^2 is known but μ is not. In order to estimate the $(1-\alpha)100\%$ confidence interval of μ we proceed as follows.

Note that $E(\bar{X}) = \mu$ & $Var(\bar{X}) = \frac{\sigma^2}{n}$

where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$.

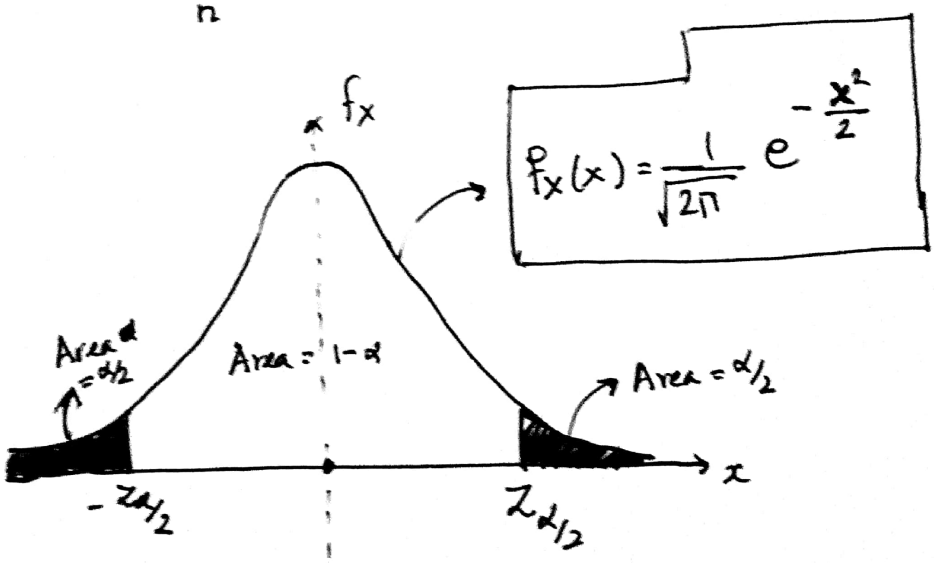


Fig 1: The point $z_{\alpha/2}$

Let us look at diagram above and it is seen how the point $Z_{\alpha/2}$ is defined, i.e.

$$\int_{Z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \alpha/2$$

$\therefore$ From diagram if $Z \sim N(0,1)$ then

$$P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1-\alpha \longrightarrow (A)$$

$\therefore$ Since we have drawn a normal population

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{[Check Chapter Lecture 10]}$$

$\therefore$ Set $Z = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Hence from (A) we have

$$P\left(-Z_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\alpha/2}\right) = 1-\alpha$$

$$\boxed{P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha}$$

Here $\hat{\theta}_1 = \bar{X} - Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ & $\hat{\theta}_2 = \bar{X} + Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

Now we will again consider studying a normal population but this time we shall consider that the population variance σ^2 is unknown. In this situation we have to use sample variance S^2 in order to try to develop a statistic using it. Let us set

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Let us assume that the sample is small, i.e. $n < 30$, then it better to proceed by noting that $T \sim t$ distribution with $n-1$ degrees of freedom (see Lecture 10). To find out the $(1-\alpha)100\%$ confidence interval we must have two points q_1 & q_2, such that

$$P(q_1 < T < q_2) = 1 - \alpha$$

$$\Rightarrow P\left(q_1 < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < q_2\right) = 1 - \alpha$$

$$\Rightarrow P\left(\frac{S}{\sqrt{n}} q_1 < \bar{X} - \mu < \frac{S}{\sqrt{n}} q_2\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X} - q_2 \frac{S}{\sqrt{n}} < \mu < \bar{X} - q_1 \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Thus in this case the confidence interval for the mean is given by

$$\left(\bar{X} - q_2 \frac{S}{\sqrt{n}} < \mu < \right.$$

$$\left(\bar{X} - q_2 \frac{S}{\sqrt{n}}, \quad \bar{X} - q_1 \frac{S}{\sqrt{n}}\right)$$

Hence the length of this interval $= (q_2 - q_1) \frac{S}{\sqrt{n}}$

or $L = (q_2 - q_1) \frac{S}{\sqrt{n}}$.

From a more practical point of view we want $(q_{\nu_2} - q_{\nu_1})$ to be as small as possible. However note that L is ing in general a random variable. However for a given sample no of size say n, we need to actually solve the optimization problem

$$\min L = (q_{\nu_2} - q_{\nu_1}) \frac{S}{\sqrt{n}} \longrightarrow (E)$$

Subject to $$\int_{q_{\nu_1}}^{q_{\nu_2}} f_T(q) \, dq = 1 - \alpha \longrightarrow (F).$$

where f_T is the pdf of the t-distribution with $(n-1)$ degrees of freedom. Let us rewrite (F) using the dummy variable t, ie

$$\int_{q_{\nu_1}}^{q_{\nu_2}} f_T(t) \, dt = 1 - \alpha. \longrightarrow (F')$$

Since α is fixed from a theoretical point of view one can always express q_{ν_2} as a function of q_{ν_1}. Let us first differentiate (F') with respect to q_{ν_1}. In order to do so we need to use the Lebnitz formula for differentiating under the integral sign. We shall write it ## in the following box.

Lebnitz rule : For differentiating under the integral sign.

Let $a(x)$ and $b(x)$ are differentiable function. Then consider the integral

$$\int_{a(x)}^{b(x)} \varphi(x,t) \, dt$$

$$\frac{d}{dx} \int_{a(x)}^{b(x)} \varphi(x,t) \, dt = \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} \varphi(x,t) \, dt + \varphi(x, b(x)) \frac{db}{dx} - \varphi(x, a(x)) \frac{da}{dx}$$

Note that if we go back to (P'), f_T depends on t and not on q_1. Thus we will have

$$\frac{d}{dq_1} \int_{q_1}^{q_2} f_T(t)\, dt = 0$$

By Leibnitz rule we

$$\int_{q_1}^{q_2(q_1)} \frac{\partial}{\partial q_1} f_T(t)\, dt + f_T(q_2) \frac{dq_2}{dq_1} - f_T(q_1) \frac{dq_1}{dq_1} = 0$$

$\Rightarrow$ As $\frac{\partial}{\partial q_1} f_T(t) = 0$ we have from the above expression

$$f_T(q_2) \frac{dq_2}{dq_1} - f_T(q_1) = 0 \longrightarrow (a)$$

Now when we replace q_2 as a function of q_1 in L then the problem become unconstrained and to minimize L we first differentiate with respect to q_1

$$\left(\frac{dq_2}{dq_1} - 1 \right) \frac{s}{\sqrt{n}} = 0$$

$$\Rightarrow \left(\frac{f_T(q_1)}{f_T(q_2)} - 1 \right) \frac{s}{\sqrt{n}} = 0$$

$$\text{or} \quad f_T(q_1) = f_T(q_2)$$

So, if q_1 is the minimizer then $f_T(q_1) = f_T(q_2)$.

Of course one solution is $q_1 = q_2$, but that will give $1 - \alpha = 0$ from (P'), which is a contradiction.

But as the graph of t distribution is symmetric we have $q_{1-\alpha} = -q_\alpha$ as the solution. So from the statistical tables we have to find q_α such that

$$P(T > q_\alpha) = \frac{\alpha}{2}, \quad \text{where } T \sim t \ (n-1 \text{ degrees of freedom})$$

In general one writes $q_\alpha = t_{\alpha/2,\, n-1}$

What happens if we have a large sample but we draw a sample from a normal distribution. Then using what is known as "central limit theorem", we can deduce that if

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

where n is the sample size then as $n \to \infty$, the distribution function of Z_n coincides with that of the standard normal distribution $N(0,1)$. Thus for n very large $Z_n \sim N(0,1)$ "approximately". So I leave it to the reader to deduce that the confidence $(1-\alpha)100\%$ confidence interval of μ is given as

$$\left(\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

Of course the next question one may ask is what about the variance. If we are studying a normal population with mean μ and variance σ^2, what is the confidence interval for σ^2, assuming of course σ^2 is not known. This is what we will study in the next section.

3. Interval Estimation for the Variance

As before let us draw a sample $X_1, \ldots, X_n$ from a normal distribution with mean μ and variance σ^2. Here the variance is unknown. In order to proceed we will introduce what is known as a pivotal quantity.

Defn 3.1 (Pivotal quantity)

Let $X_1 \ldots X_n$ be a random sample of size n drawn from a density $f(\cdot, \theta)$; where θ as before is the parameter of the distribution. Let $Q = q_f(X_1 \ldots X_n, \theta)$, ie Q is a function of the sample observations and θ. If Q is a probability density that does not depend on θ, then Q is called a pivotal quantity.

E.g: In the examples of estimating mean $Z_n = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is a pivotal quantity.

So the key idea of estimation is to find q_1 & q_2 such that

$$P[q_1 < Q < q_2] = (1 - \alpha).$$

In the particular case of estimating variance, the pivotal quantity is

$$Q = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

Now $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2$ with $(n-1)$ degrees of freedom.

$$P[q_1 < Q < q_2] = 1 - \alpha$$

$$\Rightarrow P\left[q_1 < \frac{(n-1)s^2}{\sigma^2} < q_2\right] = 1 - \alpha$$

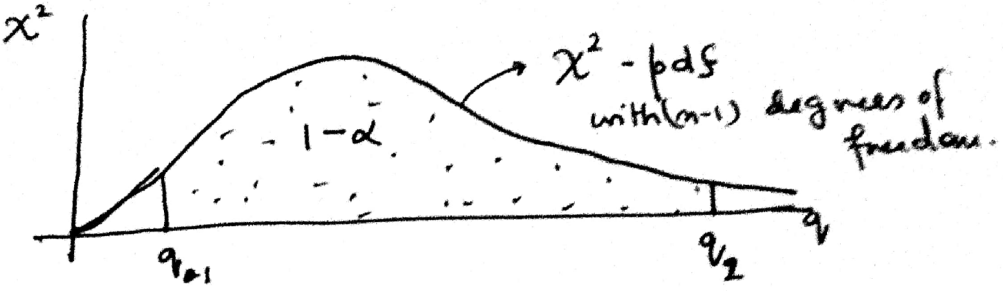
(s^2 is the observed value of the r.v. S^2)

$$P\left[\frac{(n-1)s^2}{q_2} < \sigma^2 < \frac{(n-1)s^2}{q_1}\right] = 1-\alpha$$

∴ The $(1-\alpha)100\%$ confidence intervals are then given as

$$\left(\frac{(n-1)s^2}{q_2}, \frac{(n-1)s^2}{q_1}\right)$$

Remember that χ^2 with $(n-1)$ degrees of freedom is not a symmetric distribution



The key idea is to find q_1 and q_2 such that

$$P[Q < q_1] = \alpha/2 \quad \& \quad P[Q > q_2] = \alpha/2.$$

This is equal called equal tails estimation. In practice one should figure this out from the statistical tables related to χ^2 distribution but as before one may also obtain this by numerically solving the optimization problem.

$$\min \min L = (n-1)s^2\left[\frac{1}{q_1} - \frac{1}{q_2}\right]$$

$$\text{Subject to} \int_{q_1}^{q_2} f_Q(q)\,dq = 1-\alpha.$$

Estimation of the Confidence Interval for the difference of the means of two population

In this case we will be drawing samples from two normal populations.

Sample 1 will be drawn from $N(\mu_1, \sigma_1^2)$

Sample 2 will be drawn from $N(\mu_2, \sigma_2^2)$

We will assume that σ_1^2 and σ_2^2 are known. We want to ~~use the~~ estimate the difference of the means μ_1 and μ_2, ie $\mu_1 - \mu_2$ to know how different these two normal populations are. Size of sample 1 is n_1

The pivotal quantity in this case is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{X}_1$ is the mean of the first sample and $\bar{X}_2$ is the mean of the ~~small~~ second sample. One can show that

$$Z \sim N(0,1)$$

Observe that

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

As $\bar{X}_1$ & $\bar{X}_2$ are independent

It can be proved using any approach of finding the sampling distribution of $\bar{X}_1 - \bar{X}_2$, that

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Thus $Z \sim N(0,1)$.

Hence if we are looking for a $(1-\alpha)100\%$ ~~interval~~ confidence interval then

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

If the samples are large and σ_1^2 & σ_2^2 are not known then replace them by s_1^2 and s_2^2 in the expression of Z and proceed using the central limit theorem.

In that case we have

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

However if the sample size is small ~~th~~ ie ($n_1 < 30$ or $n_2 < 30$ or both), then the confidence interval can be used if both population is assumed to have the same variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ even ~~though~~ though its exact value is not known. Then the key is to consider the pooled sample variance

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

This is nothing but an weighted average of s_1^2 & s_2^2; i.e

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(x_i^1 - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i^2 - \bar{x}_2)^2}{n_1+n_2-2}$$

You will show in the assignments that $E(s_p^2) = \sigma^2$

Construct a new random variable

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

Now $\frac{(n_1-1)}{\sigma^2}s_1^2 \sim \chi^2$ with n_1-1 degrees of freedom

and $\frac{(n_2-1)}{\sigma^2}s_2^2 \sim \chi^2$ with n_2-1 degrees of freedom.

Thus $Y \sim \chi^2$ with n_1+n_2-1 degrees of freedom. Now construct the pivotal variable

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Of course it can be shown that T follows the t distribution with n_1+n_2-2 degrees of freedom. Once this is known we can go ahead and write down the $(1-\alpha)100\%$ confidence intervals. But how did we get T.

Note that let us set

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Of course $Z \sim N(0,1)$. It can also be shown that Z and Y are independent. (Note that ~~we have~~ even one can show that $\bar{x}$ and S^2 are independent).

$$\therefore \quad T = \frac{Z}{\sqrt{\frac{Y}{n_1+n_2-2}}}$$

Hence $T \sim t$ with (n_1+n_2-2) degrees of freedom. This is known from our study of sampling distributions. The fact that $T \sim t$ with (n_1+n_2-2) degrees of freedom precisely needs the fact that Z and Y are independent

$$\therefore \quad T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hence that $(1-\alpha)100\%$ confidence interval is given as

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