Chapter 6

Estimation

6.1 *basic concepts*

- In statistical inference, the problem of estimation is part of decision making process.
- In the problem of estimation determination we must determine the value of a parameter (s) from a possible continuum of alternatives.

- Suppose, we use a value of to estimate the mean of a population,
- and observed sample proportion to estimate the parameter P of binomial population or a value of S^2 to estimate a population variance.
- We, in each case use a point estimate of the parameter. These estimates are

Correspondingly as point estimators for instance, $\bar{x}$ may be used as a point estimator of μ and similarly s^2 may be used as a point estimator of σ^2, in which cases $\bar{x}$ and s^2 are point estimates of these parameters respectively.

6.2 desirable properties of estimators

- The particular properties of estimators are unbiased ness,
- minimum variance,
- efficiency,
- consistency,
- sufficiency and
- robustness.

1. Point Estimation

- It is a procedure that results in a single value as an estimate for a parameter.

2. Interval estimation

- It is the procedure that results in the interval of values as an estimate for a parameter, which is interval that contains the likely values of a parameter. It deals with identifying the upper and lower limits of a parameter. The limits by themselves are random variable

Definitions

- **Confidence Interval**: An interval estimate with a specific level of confidence
- **Confidence Level**: The percent of the time the true value will lie in the interval estimate given.
- **Consistent Estimator**: An estimator which gets closer to the value of the parameter as the sample size increases.
- **Degrees of Freedom**: The number of data values which are allowed to vary once a statistic has been determined.
- **Estimator**: A sample statistic which is used to estimate a population parameter.
- It must be unbiased, consistent, and relatively efficient.

- **Estimate**: Is the different possible values which an estimator can assumes.
- **Interval Estimate**: A range of values used to estimate a parameter.
- **Point Estimate**: A single value used to estimate a parameter.
- **Relatively Efficient Estimator**: The estimator for a parameter with the smallest variance.
- **Unbiased Estimator**: An estimator whose expected value is the value of the parameter being estimated

Point and Interval estimation of the population mean: μ

Point Estimation

- Another term for statistic is **point estimate**, since we are estimating the parameter value.
- A **point estimator** is the mathematical way we compute the point estimate.
- For instance, sum of over n is the point estimator used to compute the estimate of the population . This is

$$\overline{X} = \frac{\sum x_i}{n}$$

is a point estimator of the population mean.

Confidence interval estimation of the population mean

- We will have to be satisfied knowing that the statistic is "close to" the parameter. That leads to the obvious question, what is "close"?
- what is the probability that the parameter's value is within a certain range of the statistic's value? This range is the confidence interval.
- The **confidence level** is the *probability* that the value of the parameter falls within the range specified by the confidence interval surrounding the

There are different cases to be considered to construct confidence intervals.

- **<u>Case 1:</u>** ***If sample size is large or if the population is normal with known variance***
- By using *Central Limit Theorem*, which applies to the sampling distribution of the mean of a sample.
- The sampling distribution of will have a mean and a standard deviation , and *approaches a normal distribution as n gets large*. This allows us to use the normal distribution curve for computing confidence intervals.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$\Rightarrow \mu = \overline{X} \pm Z\,\sigma/\sqrt{n}$$

$$= \overline{X} \pm \varepsilon, \quad \text{where } \varepsilon \text{ is a measure of error.}$$

$$\Rightarrow \varepsilon = Z\,\sigma/\sqrt{n}$$

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

Where α = *is the probability that the parameter lies outside the interval*

$Z_{\alpha/2}$ = *standards for the standard normal variable to the right of which* $\alpha/2$ *probability lies, i.e* $P(Z > Z_{\alpha/2}) = \alpha/2$

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

$$\Rightarrow P(\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1 - \alpha$$

$\Rightarrow (\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n},\ \overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n})$ *is a* $100(1-\alpha)\%$ *confidence interval for* μ

- But usually σ is not known, in that case we estimate by its point estimator S^2

$$\Rightarrow (\bar{X} - Z_{\alpha/2} S/\sqrt{n},\ \bar{X} + Z_{\alpha/2} S/\sqrt{n}) \text{ is a } 100(1-\alpha)\% \text{ conifidence interval for } \mu$$

Here are the z values corresponding to the most commonly used confidence

$100(1-\alpha)\%$	α	$\alpha/2$	$Z_{\alpha/2}$
90	0.10	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.58

Case 2: *If sample size is small and the population variance, is not known.*

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad has\ t\ distribution\ with\ n-1\ degrees\ of\ freedom.$$

$$\Rightarrow (\bar{X} - t_{\alpha/2}\, S/\sqrt{n},\ \ \bar{X} + t_{\alpha/2}\, S/\sqrt{n})\ \ is\ a\ 100(1-\alpha)\%\ \ confidence\ interval\ \ for\ \mu$$

Example 1: If a random sample of size n = 40 form a normal population with the variance = 225 has the mean =64.3. construct a 0.95 or (95%) confidence interval for the population mean .

- **Solution**: Since n 30, then we use Z - table and form the Z-table

$$\bar{x} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow 64.3 - 1.96 \times \frac{15}{\sqrt{40}} < \mu < 64.3 + 1.96 \times \frac{15}{\sqrt{40}}$$

$$59.65 < \mu < 68.95$$

Example 2: A fertilizer manufacture wants to estimate the average number of tons of fertilizer sold per month. One year was monitored, and average monthly sales of 10 tons were recorded. If the sample variance is 4 tons, compute the confidence limits at the 95% level

S

Given: $\bar{x} = 10$ tons, $S^2 = 4$ tons $\rightarrow$ $S = 2$ tons,

n = one year $\rightarrow$ 12 months, confidence level = $1-\alpha = 0.95$, $\alpha/2 = 0.025$

There fore we use t − distribution and $t_{\alpha/2,n-1} = t_{\alpha/2(12-1)} = t_{\alpha/2,11} = 2.201$

Then the interval estimate is:

$$\bar{x} - t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}$$

$$10 - 2.201 \times \frac{2}{\sqrt{12}} < \mu < 10 + 2.201 \times \frac{2}{\sqrt{12}}$$

$$8.729 < \mu < 11.271$$

Estimation of differences between means

- If we have two populations with specified means and variances and n_1 and n_2 be two large samples each greater than or equal to 30 drawn from the first and the second populations respectively the sampling distribution of is normal with mean and standard

- The statistic $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ has the standard normal distribution.

  If we substitute the above expression for Z in to $P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$

  $$P(\bar{x}_1 - \bar{x}_2 - Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2} \times \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1 - \alpha$$

- In other words, if are the values of the means of independent random samples of size n_1 and n_2 from a normal populations with the known variances

$$(\overline{x_1}-\overline{x_2})-Z_{\alpha/2}\times\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\mu_1-\mu_2<(\overline{x_1}-\overline{x_2})+Z_{\alpha/2}\times\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

- is a (1-)100% confidence interval for the difference between the two population means.
- If & are not known & replace by respective sample standard deviation &

- On the other hand, when the sample sizes n_1 and n_2 are small, and & are unknown the procedure for estimating the difference between the means of the two normal population could be constructed by assuming that = ,
- If — then

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

- is a random variable having the standard normal distribution, and can be estimated as pooled estimator by pooling the squared deviations form the means of the two samples.

Assumption: The population is normal and the samples are independent

$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ is indeed, and unbiased estimator of σ^2.

Recalling the chi-square distributions, the independent random variables

$\frac{(n_1-1)S_1^2}{\sigma^2}$ *and* $\frac{(n_2-1)S_2^2}{\sigma^2}$ .have the chi-square distribution

With n_1-1 and n_2-1 degrees of freedom, and their sum

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

a chi-square distribution with n_1+n_2-2 degrees of freedom.

Since the random variables Z and Y are independent, then

$$T = \frac{Z}{\sqrt{\frac{y}{n_1+n_2-2}}} = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$ has a t-distribution

- with n_1+n_2-2 degrees of freedom and hence substituting this expression for t- in

$$P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-1}) = 1-\alpha,$$

- For if and S_1 & S_2 are the values of the means and the standard deviations of independent random samples of size n_1 and equal variances, then

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

- is a (1-α)100% confidence interval for

- **Example 3:** A study has been made to compare the nicotine contents of two brands of tea leaves. Ten packets of tealeaves of “wushwush” had an average nicotine content of 3mg with a variance of 0.25mg while eight packets of tea leaves of “Ambessa” had an average nicotine content of 2.8mg with a variance of 0.49mg. Assuming that the two sets of data are independent random samples from normal populations with equal variance, construct a 0.95 confidence interval for the difference between the mean nicotine contents of the two types of tea leaves.

- **Solution:**

Given: $\bar{x}_1 = 3mg; \bar{x}_2 = 2.8mg$, and

$n_1 = 10, n_2 = 8, S_1^2 = 0.25 \rightarrow S_1 = 0.5, S_2^2 = 0.49 \rightarrow S_R = 0.7,$

then $Sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10-1)0.25 + (8-1)0.49}{10+8-2}} =$

$= \sqrt{\frac{9 \times 0.25 + 7 \times 0.49}{16}} = 0.596$ and since $n_1 < 30$ and $n_2 < 30$ we use

$t-$ distribution, and $t_{\frac{\alpha}{2}, n_1 + n_2 - 2} = t_{0.025, 16} = 2.120$(from t-table), then

a 95% confidence interval:

$$(\overline{X_1} - \overline{X_2}) - t_{\alpha/2, n_1+n_2-2} \times Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2, n_1+n_2-2} \times Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$=(3-2.8) - 2.120 \times 0.596\sqrt{\frac{1}{10} + \frac{1}{8}} < \mu_1 - \mu_2 < (3-2.8) + 2.120 \times 0.596 \times \sqrt{\frac{1}{10} + \frac{1}{8}}$$

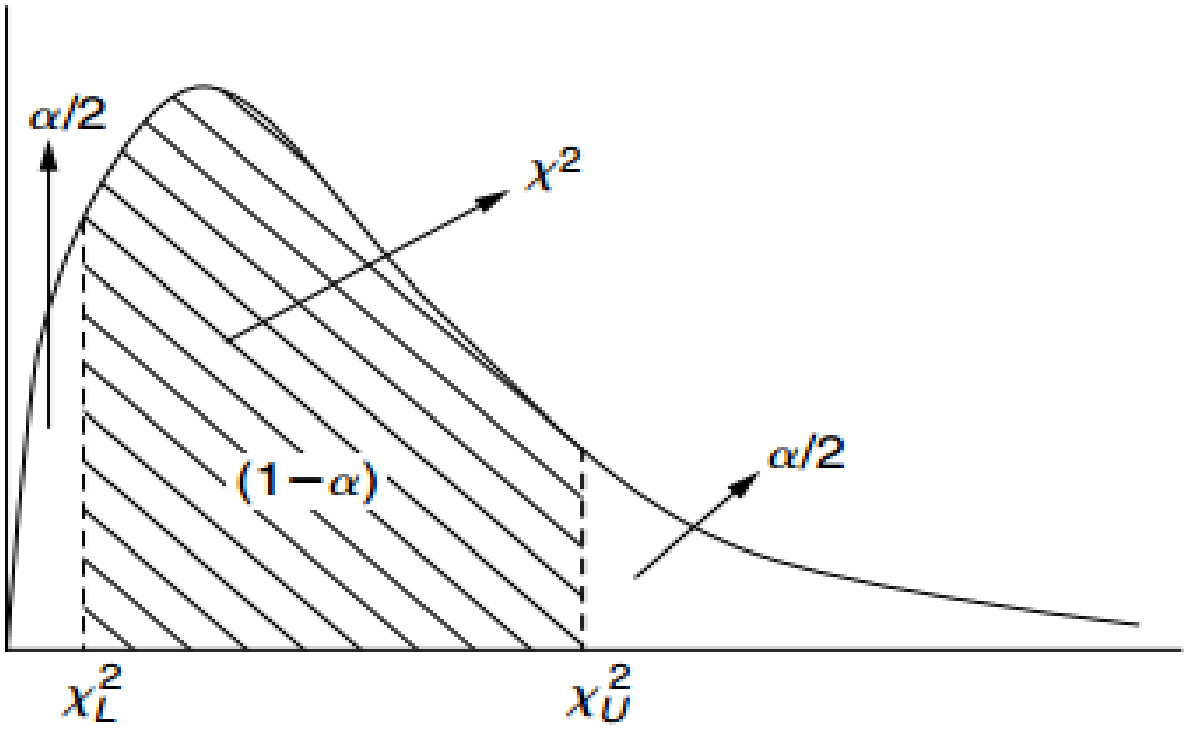
$$=0.2 - 1.26352 \times 0.474341649 < \mu_1 - \mu_2 < 0.2 + 1.26352 \times 0.474341649 =$$

$$-0.39934016 < \mu_1 - \mu_2 < 0.79934016$$

$$\Rightarrow -0.4 < \mu_1 - \mu_2 < 0.8$$

Estimation of Variances

- If we have n independent observations with the variance of a random sample from a normal population, then a random variable having a chi-square distributio



Thus, $P[X^2_{1-\alpha/2,n-1} < \chi^2 < X^2_{\alpha/2,n-1}] = 1-\alpha$

$$P[X^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < X^2_{\alpha/2,n-1}] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{X^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{X^2_{1-\alpha/2,n-1}}\right] = 1-\alpha,$$

Example 3: The length of the skulls of 10 fossil skeletons of an extinct species of bird has a mean of 5.68cm and a standard deviation of 0.29cm assuming that such measurements are normally distributed. Construct a 95% confidence interval for the true variance of the skull l .

Solution:

$$\text{Given: } n = 10,\ S = 0.29 \rightarrow S^2 = 0.0841$$

$$X^2_{0.025,9} = 19.023;\ X^2_{0.975,9} = 2.70039$$

$$\Rightarrow \frac{9(0.29)2}{19.023} < \sigma 2 < \frac{9(0.29)^2}{2.70039}$$

$$0.0398 < \sigma 2 < 0.2803$$


THANK
YOU