Chapter 8

One and two sample estimation problems

We know that statistical engineering consists of two categories

1- Descriptive statistics
2- Inferential statistics

In previous chapters, we collected information about first part by some methods for presenting and showing results.

Our methods are Frequency, Relative Frequency, Frequency Table, Bar Graph, Line Graph, Pie Graph, and Histogram …

İn second part (Inferential Statistics), we will talk about finding point estimator and constructing confidence interval for parameters of one or two population and also hypothesis testing in different situations.

İn this part, there is one population with some fix parameters like mean μ, variance σ^2, standard deviation σ. Because of different reasons, we have only one sample with size n from that population and by statistics of this sample we want to reach parameter point estimator (Peer to Peer)

Point estimator: point estimator of population parameter like θ is a single value like $\hat{\theta}$ that calculates from a sample. For example the value of $\bar{x}$ (mean value of a sample) is a point estimator of μ. similarly, $\hat{p} = \frac{x}{n}$ is a point estimator of the true proportation P for a binomial experiment.

Unbiased estimator: a statistics $\hat{\theta}$ is to be unbiased estimator of the parameter θ, if $E(\hat{\theta}) = \theta$. Also biasedness is $b(\theta) = E(\hat{\theta}) - \theta$.

Example: if x has the binomial distribution with parameters n and p, show the sample proportion of x/n is an unbiased estimator of p.

$$E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{np}{n} = p$$

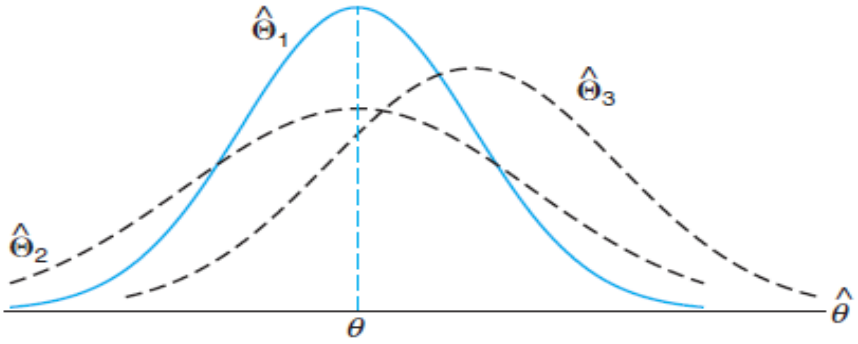
Example: a sample of n independent observations $(x_1, x_2, \dots, x_n)$ taken. Under what condition $\hat{\mu} = \sum_{i=1}^{n} a_i x_i$ will be an unbiased estimator of the mean of the population?

$$E(\hat{\mu}) = \mu$$

$$E\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} E(a_i x_i) = \sum_{i=1}^{n} a_i E(x_i) = \sum_{i=1}^{n} a_i\,\mu \;\; should\ be\ equal\ to\ \mu \rightarrow \sum_{i=1}^{n} a_i = 1$$

Variance of a point estimator

- İf $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same population parameter θ, we want to choose the estimator whose sampling distribution has the smaller variance. hence, if $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, we say that $\hat{\theta}_1$ is more efficient estimator of θ than $\hat{\theta}_2$
- İf we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the most efficient estimator of θ



$\hat{\theta}_1, \hat{\theta}_2, \text{and } \hat{\theta}_3$, all estimating θ. İt is clear that only $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased, since their distributions are centered at θ. The estimator $\hat{\theta}_1$ has a smaller variance than $\hat{\theta}_2$ and is therefore more efficient. hence, our choice for an estimator of θ, amoung the three considered, would be $\hat{\theta}_1$.

Properties of expectation value:

- $E(cX) = c.E(X)$ x is random variable and c is constant value
- $E(X + c) = c + E(X)$
- $E(X \pm Y) = E(X) \pm E(Y)$ X & Y are random variables

$$E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E(x_i)$$

Properties of variance:

- $var(cX) = c^2.var(X)$ x is random variable and c is constant value
- $var(X + c) = var(X)$
- $var(X \pm Y) = var(X) + var(Y)$ X & Y are independent random variables
- $var[\sum_{i=1}^{n} x_i] = \sum_{i=1}^{n} var(x_i)$
- $Standard\ Deviation\ (SD) of\ X = \sqrt{var(X)}$

Example: if $x_1, x_2, and x_3$ constitute a random sample size of n=3 from a normal population with the mean μ and the variance σ^2. Assume that $\theta_1, \theta_2, \theta_3$ are estimators

$$\theta_1 = \frac{x_1 + x_2 + x_3}{2} \qquad \theta_2 = \frac{x_1 + 2x_2 + x_3}{4} \qquad \theta_3 = \frac{x_1 + x_2 + x_3}{3}$$

- Which one is an unbiased estimator of μ?
- Calculate biasness of bias point estimator?
- Find which one is more efficient?

$$E(\theta_1) = E\left(\frac{x_1 + x_2 + x_3}{2}\right) = \frac{1}{2}E(x_1 + x_2 + x_3) = \frac{1}{2}(E(x_1) + E(x_2) + E(x_3)) = \frac{1}{2}(\mu + \mu + \mu) = \frac{3\mu}{2} \neq \mu$$

$$\rightarrow \theta_1 \text{ is biased estimator}$$

$$biasness = E(\hat{\theta}) - \theta = \frac{3\mu}{2} - \mu = \frac{\mu}{2}$$

$$E(\theta_2) = E\left(\frac{x_1 + 2x_2 + x_3}{4}\right) = \frac{1}{4}E(x_1 + 2x_2 + x_3) = \frac{1}{4}(E(x_1) + 2E(x_2) + E(x_3)) = \frac{1}{4}(\mu + 2\mu + \mu)$$

$$= \frac{4\mu}{4} = \mu \rightarrow \theta_2 \text{ is unbiased estimator}$$

$$E(\theta_3) = E\left(\frac{x_1 + x_2 + x_3}{3}\right) = \frac{1}{3}E(x_1 + x_2 + x_3) = \frac{1}{3}(E(x_1) + E(x_2) + E(x_3)) = \frac{1}{3}(\mu + \mu + \mu) = \frac{3\mu}{3} = \mu$$

$$\rightarrow \theta_3 \text{ is unbiased estimator}$$

$$V(\theta_2) = V\left(\frac{x_1 + 2x_2 + x_3}{4}\right) = \frac{1}{16}V(x_1 + 2x_2 + x_3) = \frac{1}{16}(V(x_1) + 4V(x_2) + V(x_3))$$

$$= \frac{1}{16}(\sigma^2 + 4\sigma^2 + \sigma^2) = \frac{6\sigma^2}{16}$$

$$V(\theta_3) = V\left(\frac{x_1 + x_2 + x_3}{3}\right) = \frac{1}{9}V(x_1 + x_2 + x_3) = \frac{1}{9}(V(x_1) + V(x_2) + V(x_3)) = \frac{1}{9}(\sigma^2 + \sigma^2 + \sigma^2)$$

$$= \frac{3\sigma^2}{9} = \frac{\sigma^2}{3} \qquad V(\theta_3) < V(\theta_2) \rightarrow \theta_3 \text{ is more efficient}$$

Example: If x_1, x_2, x_3, x_4 constitute a random sample size of $n = 4$ from a normal population with the mean μ and the variance σ^2.

$$\theta_1 = \frac{3x_1 + 3x_2 + 3x_3 - 3x_4}{6}, \qquad \theta_2 = \frac{2x_1 + 2x_2 + 2x_3 - x_4}{6}, \qquad \theta_3 = \frac{2x_1 + 2x_2 - 2x_2 + 3x_4}{5}$$

- Which one is/are unbiased point estimator of μ? .
- Calculate biaseness of biased point estimator of μ.
- Which one is more efficient estimator of μ? Why? .

Confidence interval

There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the parameter. Such an interval is called an interval estimate

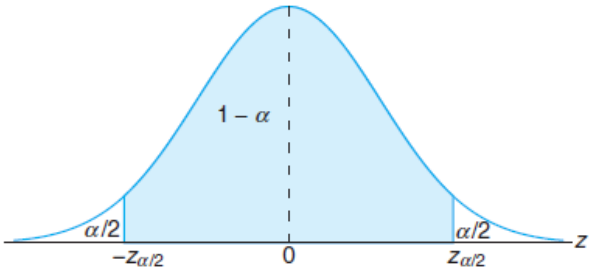
Interpretation of interval estimates:

Since different samples will generally yield different values of $\hat{\theta}$ and, therefore, different values for $\hat{\theta}_L \; and \; \hat{\theta}_U$, these end points of the interval are values of corresponding random variables $\hat{\theta}_L \; and \; \hat{\theta}_U$. If for instance, we find $\hat{\theta}_L \; and \; \hat{\theta}_U$ such that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha \qquad for \; 0 < \alpha < 1$$

Then we have a probability $1 - \alpha$ of selecting random sample that will produce an interval containing θ. The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, computed from the selected sample, is called a $100(1 - \alpha)\%$ confidence interval.

Confidence interval of single sample



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

From result of CLT, we know that $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ and by puting z to formula we will reach to

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

Multiplying each term in the inequality by $\sigma/\sqrt{n}$ and then subtracting $\bar{x}$ from each term and multiplying by -1 (reserve the sense of the inequalities) , we obtain

$$P\left(\bar{x} - Z\alpha_{/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + Z\alpha_{/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$$

Where $Z\alpha_{/2}$ is the z value leaving an area of $\alpha/2$ to the right $(n \geq 30)$

Example: the average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per millimeter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per millimeter

$$P\left(\bar{x} - Z\alpha_{/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + Z\alpha_{/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$$

$$P\left(2.6 - 1.96\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 1.96\frac{0.3}{\sqrt{36}}\right) = 0.95$$

$$P(2.6 - 0.1 < \mu < 2.6 + 0.1) = 0.95$$

$$P(2.5 < \mu < 2.7) = 0.95$$

$$P\left(\bar{x} - Z\alpha_{/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + Z\alpha_{/2}\,\sigma/\sqrt{n}\right) = 1 - \alpha$$

$$P\left(2.6 - 2.575\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 2.575\frac{0.3}{\sqrt{36}}\right) = 0.99$$

$$P(2.6 - 0.13 < \mu < 2.6 + 0.13) = 0.99$$

$$P(2.47 < \mu < 2.73) = 0.99$$

1- Confidence interval on μ, σ^2 *Known*

If $\bar{x}$ is the mean of random sample of size n from a population with known variance $\boldsymbol{\sigma^2}$, a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\mu}$ is given by

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

Where $z_{\alpha/2}$ is the z value leaving an area of $\alpha/2$ to the right $(n \geq 30)$

Example: One organization wishes to have information on the mean income of managers in the industry. A random sample of 256 managers reveals a sample mean of \$45,240. The standard deviation of this population is \$2,050. The association would like answers to the following questions.

- What is the point estimator of mean's population?
- What is the reasonable range of values for the population mean with confidence coefficient of 0.95
- What do these results mean?

$$\hat{\mu} = \bar{x} = 45240$$

$$P\left(\bar{x} - Z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}\right) = 1 - \alpha$$

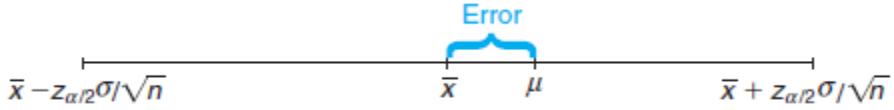
$$P\left(45240 - 1.96\frac{2050}{\sqrt{256}} < \mu < 45240 + 1.96\frac{2050}{\sqrt{256}}\right) = 0.95$$

$$P(45240 - 251 < \mu < 45240 + 251) = 0.95$$

$$P(44989 < \mu < 45491) = 0.95$$

If we select many samples of 256 managers, and for each sample, we calculate the mean value and then construct a 95 percent confidence interval. We could expect that about 95 percent of these confidence intervals to contain the real population mean value.

Note: the $100(1-\alpha)\%$ confidence interval provides an estimate of the accuracy of our point estimate. İf μ is actually the center value of the interval, then $\bar{x}$ estimates μ without error. But most of the time there will be an error. The size of this error will be the absolute value of difference between μ and $\bar{x}$, and we can be $100(1-\alpha)\%$ confident that this difference will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We can readily see this if we draw a diagram of a hypothetical confidence interval, as



In the previous example, we are 95% confident that the sample mean $\bar{x} = 45{,}240$ differs from the true mean μ by an amount less than $\frac{(1.96)(2050)}{\sqrt{256}} = 251.125$ and 99% confident that the difference is less than $\frac{(2.575)(2050)}{\sqrt{256}} = 329.922$.

Theorem: if $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not be exceed a specified amount *e*when the sample size is $n = (\frac{z\alpha/_2 \sigma}{e})^2$

Example: How large a sample is required if we want to be 95% confident that our estimate of μ in the previous example is off by less than 50?

is $n = (\frac{z\alpha/_2 \sigma}{e})^2 = (\frac{1.96\times 2050}{50})^2 = 6457$

1-1 One sided confidence bounds on μ, σ^2

If $\bar{x}$ is the mean of random sample of size n from a population with known variance $\boldsymbol{\sigma^2}$, a $100(1-\alpha)\%$ confidence bounds for $\boldsymbol{\mu}$ is given by

$Lower\ one-sided\ bound \qquad \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$

$Upper\ one-sided\ bound \qquad \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$

Example: In previous example construct Lower and Upper bound for the population mean with confidence coefficient of 0.95 and 0.99

$$Upper: P\left(\mu < \bar{x} + Z_\alpha {}^{\sigma}/_{\sqrt{n}}\right) = 1-\alpha \rightarrow P\left(\mu < 45240 + 1.645\frac{2050}{\sqrt{256}}\right) = 0.95 \rightarrow P(\mu < 45450) = 0.95$$

$$lower: P\left(\mu > \bar{x} - Z_\alpha {}^{\sigma}/_{\sqrt{n}}\right) = 1-\alpha \rightarrow P\left(\mu > 45240 - 1.645\frac{2050}{\sqrt{256}}\right) = 0.95 \rightarrow P(\mu > 45030) = 0.95$$

$$UpperP\left(\mu < \bar{x} + Z_\alpha {}^{\sigma}/_{\sqrt{n}}\right) = 1-\alpha \rightarrow P\left(\mu < 45240 + 2.33\frac{2050}{\sqrt{256}}\right) = 0.99 \rightarrow P(\mu < 45537) = 0.99$$

$$lower: P\left(\mu > \bar{x} - Z_\alpha {}^{\sigma}/_{\sqrt{n}}\right) = 1-\alpha \rightarrow P\left(\mu > 45240 - 2.33\frac{2050}{\sqrt{256}}\right) = 0.99 \rightarrow P(\mu > 44943) = 0.99$$