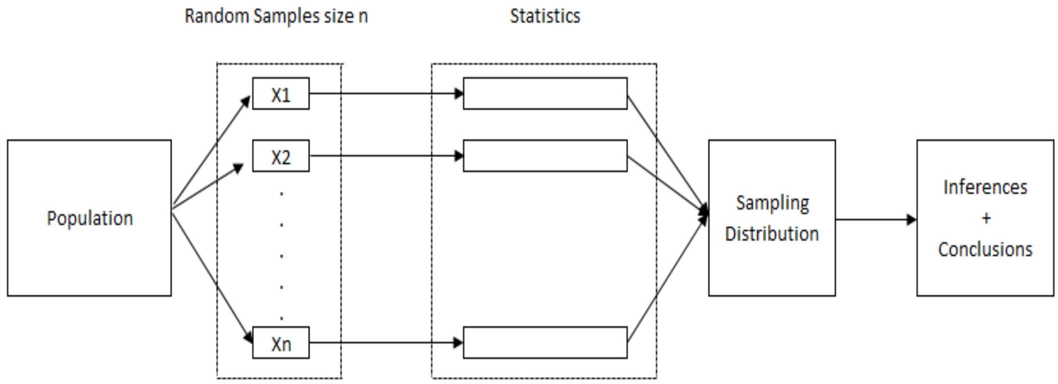
Ch 7: Sampling Distributions and Point Estimation:

Introduction:

Often we are interested in drawing some valid conclusions (inferences) about a large group of individuals or objects (called population in statistics). Instead of examining (studying) the entire group (population, which may be difficult or even impossible to examine), we may examine (study) only a small part (portion) of the population (entire group of objects or people). Our objective is to draw valid inferences about certain facts for the population from results found in the sample; a process known as statistical inferences. The process of obtaining samples is called sampling and theory concerning the sampling is called sampling theory. This leads us to what is known as sampling distributions which allows us to make decisions about population parameters (attributes) on the basis of information of statistics obtained from samples drawn from population.



Example: We may wish to draw conclusions about the percentage of defective bolts produced in a factory during a given 6-day week by examining 20 bolts each day produced at various times during the day. Note that all bolts produced in this case during the week comprise the population, while the 120 selected bolts during 6-days constitutes a sample.

In business, medical, social and psychological sciences etc., research, sampling theory is widely used for gathering information about a population. The sampling process comprises several stages:

- Defining the population of concern
- Specifying the sampling frame (set of items or events possible to measure)
- Specifying a sampling method for selecting the items or events from the sampling frame
- Determining the appropriate sample size
- Implementing the sampling plan
- Sampling and data collecting
- Data which can be selected

When studying the characteristics of a population, there are many reasons to study a sample (drawn from population under study) instead of entire population such as:

1. **Time:** as it is difficult to contact each and every individual of the whole population
2. **Cost:** The cost or expenses of studying all the items (objects or individual) in a population may be prohibitive
3. **Physically Impossible:** Some population are infinite, so it will be physically impossible to check the all items in the population, such as populations of fish, birds, snakes, mosquitoes. Similarly it is difficult to study the populations that are constantly moving, being born, or dying.
4. **Destructive Nature of items:** Some items, objects etc are difficult to study as during testing (or checking) they destroyed, for example a steel wire is stretched until it breaks and breaking point is recorded to have a minimum tensile strength. Similarly, different electric and electronic components are check and they are destroyed during testing, making impossible to study the entire population as time, cost and destructive nature of different items prohibits to study the entire population.
5. **Qualified and expert staff:** For enumeration purposes, highly qualified and expert staff is required which is some time impossible. National and International research organizations, agencies and staff is hired for enumeration purposive which is some time costly, need more time (as rehearsal of activity is required), and some time it is not easy to recruiter or hire a highly qualified staff.
6. **Reliability:** Using a scientific sampling technique the sampling error can be minimized and the non-sampling error committed in the case of sample survey is also minimum, because qualified investigators are included.

Statistic: any quantity whose value can be calculated from sample data, or by any function of the observations in a random variable.

Random Sample: The random variables $X_1, X_2, \dots, X_n$ are a random sample of size n if:

(a) the X_i's are independent random variables with equal chance of being selected
(b) every X_i has the same probability distribution. i.e. X_i's are independent and identically distributed (iid)

Sampling conditions for random sampling are satisfied if:

a. Samples have equal chance of being selected
b. Sampling is performed with replacement (if possible) or from an infinite population size N so that N>>n and $\frac{n}{N} \leq 0.05$

Sampling Methods:

- With replacement (most frequent): number of possible samples N^n permutations
- Without replacement (sometimes necessary): number of possible samples = $\frac{N!}{n!(N-n)!}$ Combinations

Sampling Distribution: It is the probability distribution of a statistic. It depends on the distribution of the population, the size of the sample and the way the samples are selected.

7.2 Estimation of parameters:

statistical estimation procedures provide estimates of population parameter with a desired degree of confidence. This can be controlled in part

- by the size of the sample (larger sample → greater accuracy)
- by the type of estimate made. The estimation of population parameters is divided into two parts;
  1. point estimation
  2. Interval estimation.

Point Estimation:

Objective is to obtain a single number from the sample which will represent the unknown value of the population parameter.

population parameters are estimated from corresponding sample statistics.

ex. sample mean, variance $\xrightarrow{\text{estimate}}$ population mean and σ^2

A statistic used to estimate a parameter is called a point estimator, the value obtained is an estimate

Point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$ which is called estimator:

Population mean μ. Based on sample estimate is $\hat{\mu}$ and $\hat{\mu} = \bar{x}$; the sample mean. $\bar{X}$ is the estimator which is a random variable with its own sampling probability distribution. It is called sampling distribution of the mean.

Variance of a single population σ^2:

the estimate is $\hat{\sigma}^2 = s^2$; the sample variance

S^2: sampling distribution of variance.

- Interval Estimation:

point estimate provides a single number as an estimate of the population parameter, which cannot be expected to be exactly equal to the population parameter. Therefore we estimate an interval of values within which the population parameter is expected to lie with a certain degree of confidence. The purpose of an interval estimate is to provide information about how close the point estimate is to the true value.

Which of the two types of estimation do you like the most, and why?

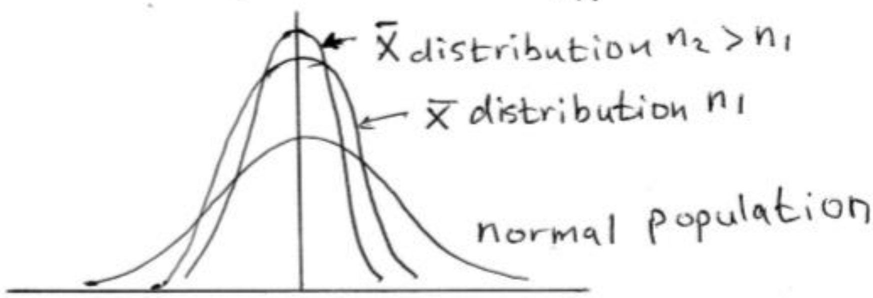
- Point estimation is nice because it provides an exact point estimate of the population value. It provides you with the single best guess of the value of the population parameter.

- Interval estimation is nice because it allows you to make statements of confidence that an interval will include the true population value.

7.3 Sampling distribution of the mean.

Let $X_1, X_2, \ldots X_n$ be a random sample from a normal population with mean μ and standard Deviation σ. Then for any n $\bar{x}$ is normally distributed with:

$$\text{mean} = \mu_{\bar{x}} = \mu \qquad SD = \frac{\sigma}{\sqrt{n}}$$



7.4 Central Limit Theorem:

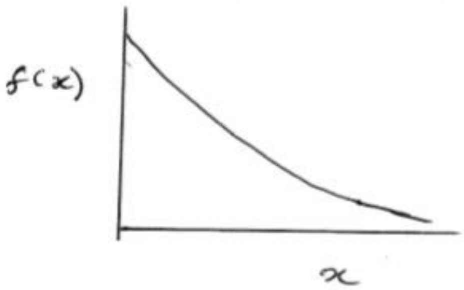
As $n \to \infty$ all sample statistics can be approximate by normal distributions

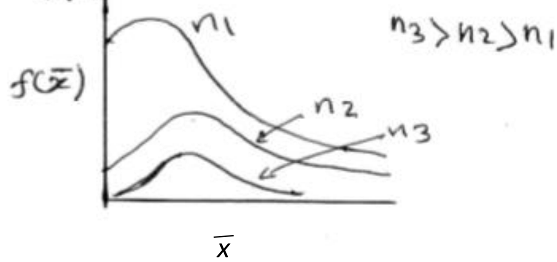
The mean of iid rv's will be approximately normally distributed regardless of the parent population's distribution.

$\therefore$ samples $x_1, x_2 \ldots x_n$ drawn from a population with μ and σ will have normal means with

$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Also $\quad Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$





Example 7.1

An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than 95 ohms.

Note that the sampling distribution of $\overline{X}$ is normal, with mean $\mu_{\overline{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Therefore, the desired probability corresponds to the shaded area in Fig. 7-1.

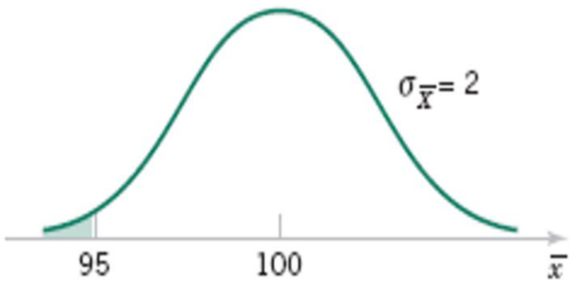


Figure 7-2 Probability for Example 7-1

Standardizing the point $\bar{X} = 95$ using $$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

$$z = \frac{95 - 100}{2} = -2.5$$

and therefore,

$$P(\overline{X} < 95) = P(Z < -2.5) = 0.0062$$

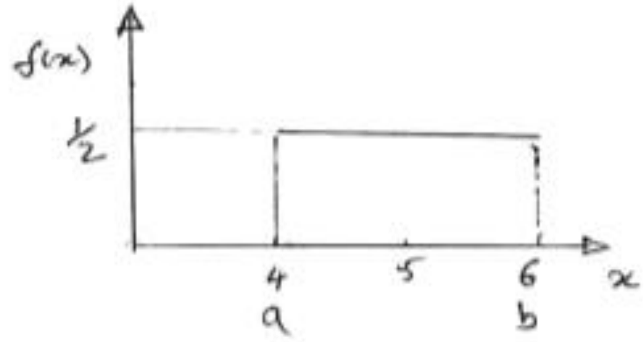
Therefore we can say that it is a rare event for $\bar{X}$ to be < 95 Ohms.

Ex 7.2

Random variable X with PDF given by

$$f(x) = \begin{cases} \frac{1}{2} & 4 \leq x \leq 6 \\ 0 & \text{otherwise.} \end{cases} \qquad \text{uniform PDF}$$

Find the distribution for $\bar{X}$ for $n=40$



- Required: Find $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$.

- Solution

Using the central Limits Theorem, we can say that $\bar{X}$ is approximately normal with

$$\mu_{\bar{x}} = \mu_x \quad \text{and} \quad \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}$$

$$\mu_x = E(X) = \int_4^6 f(x).x\,dx \qquad \boxed{\mu = \frac{b+a}{2}}$$

$$= \frac{6+4}{2}$$

$$\underline{\mu = 5}$$

$$V(x) = \int_4^6 (x-\mu_x)^2 . f(x)\,dx$$

$$= \frac{(6-4)^2}{12} = \frac{4}{12} \qquad \boxed{V(x) = \frac{(b-a)^2}{12}}$$

$$\therefore \underline{\sigma^2_x = \frac{1}{3}}$$

$$\therefore \mu_{\bar{x}} = 5 \quad \text{and} \quad \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n} = \frac{1}{3}\left[\frac{1}{40}\right] = \frac{1}{120}$$