02-24-00201
Probability and Statistics II

DR. AHMED TAYEL

Department of Engineering Mathematics and Physics, Faculty of Engineering, Alexandria University

ahmed.tayel@alexu.edu.eg

Outline

1. Introduction.
2. Some Important Statistics.
3. Choice of the sample.
4. Distribution of $\bar{X}$.
5. Case of two populations.

1. Introduction

Random Sampling

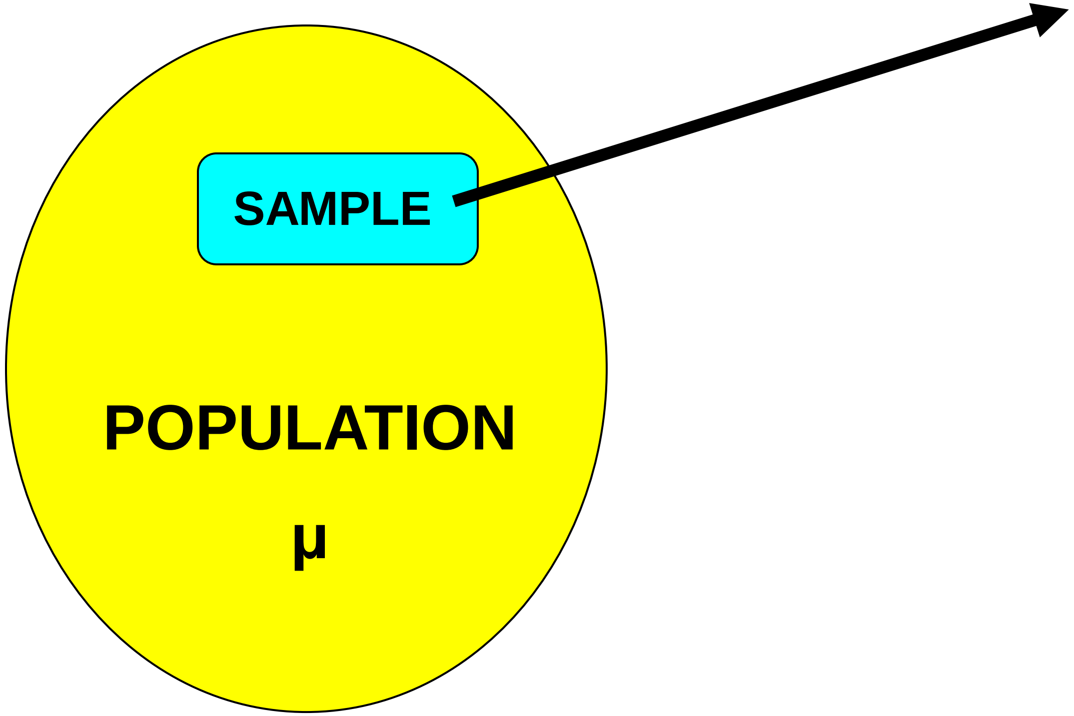
Definition 1:

A population consists of the totality of the observations with which we are concerned. (Population=Probability Distribution)

Definition 2:

A sample is a subset of a population.

The Sampling Process



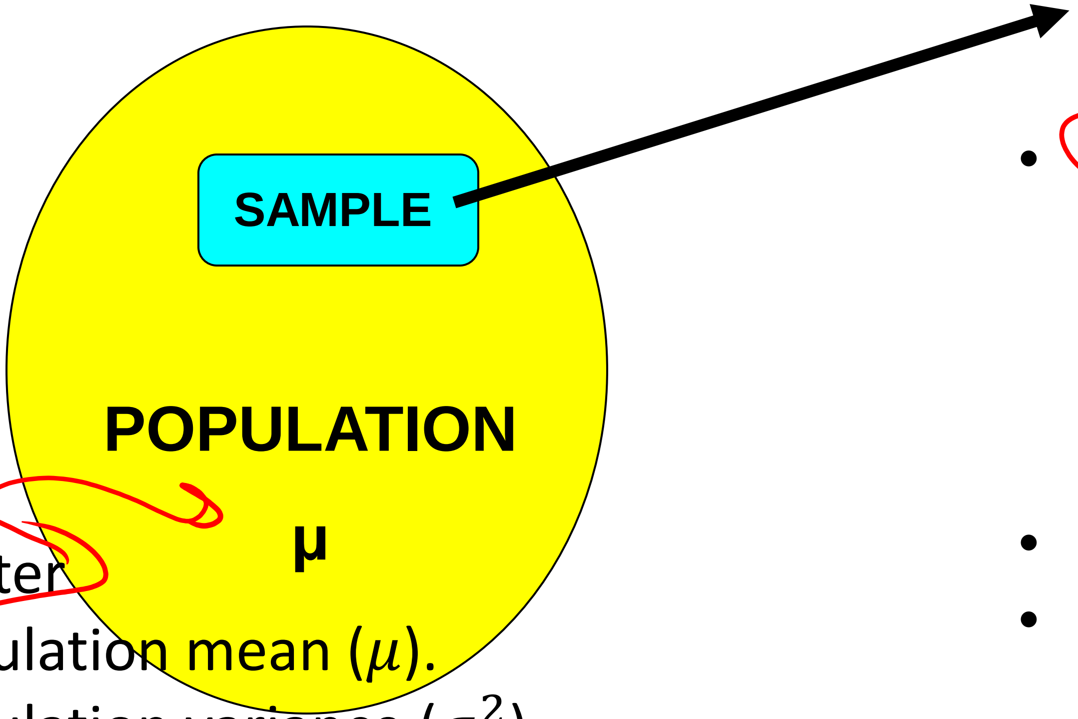
$$\{X_1, X_2, \ldots, X_n\}$$

Some Important Statistics

Definition:

Any function of the random sample $X_1, X_2, \ldots, X_n$ is called a **statistic**.

The Sampling Process

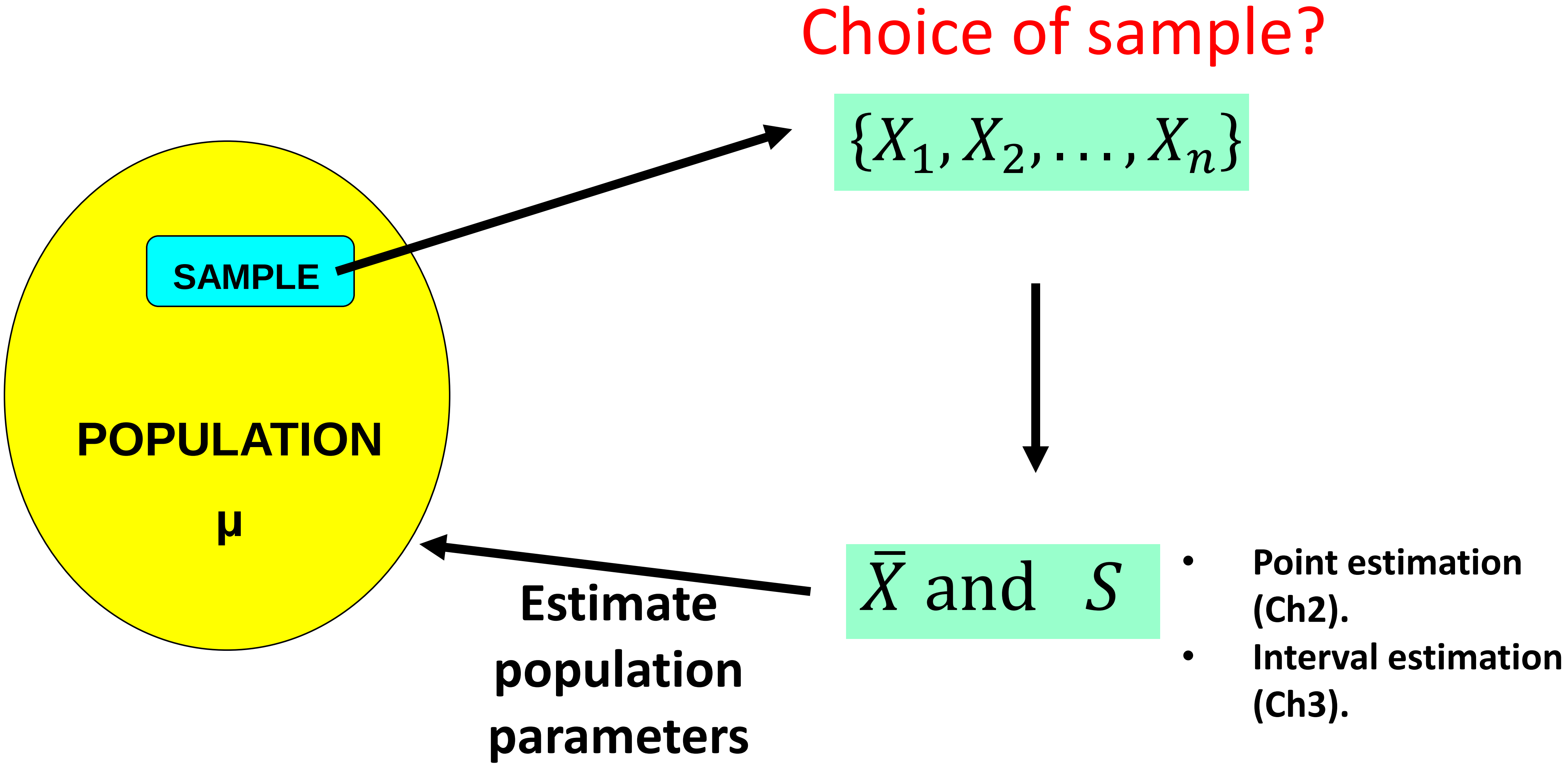


$$\{X_1, X_2, \ldots, X_n\}$$

- Parameter
  - Population mean (μ).
  - Population variance (σ^2).
  - Population standard deviation (σ).
- Hard to observe.

- Statistic
  - Sample mean $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$
  - Sample variance $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
- Much easier.
- Less cost.

The Sampling Process



2. Some Important Statistics

Central Tendency in the Sample

Definition:

If $X_1, X_2, \dots, X_n$ represents a random sample of size n, then the sample mean is defined to be the statistic:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Note:

- $\bar{X}$ is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$.
- $\bar{X}$ has same unit of $X_1, X_2, \ldots, X_n$.
- $\bar{X}$ measures the central tendency in the sample (location).

Variability in the Sample

Definition:

If $X_1, X_2, \dots, X_n$ represents a random sample of size n, then the sample variance is defined to be the statistic:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \cdots + (X_n-\overline{X})^2}{n-1} \ (\text{unit})^2$$

Q: Why $n-1$?

Note:

- S^2 is a statistic because it is a function of the random sample $X_1, X_2, \dots, X_n$.
- S^2 measures the variability in the sample.

Definition:

The sample standard deviation is defined to be the statistic:

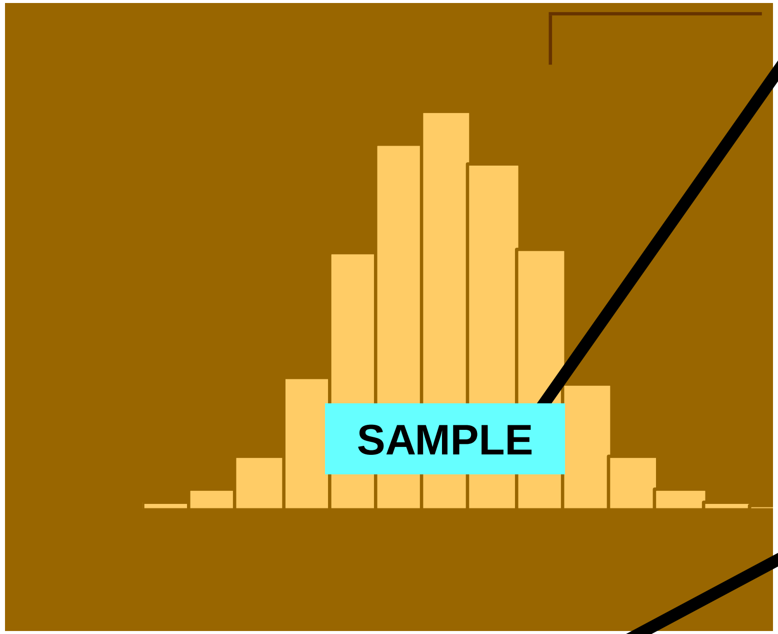
$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \text{ (unit)}$$

Sample $\overline{X}$ S^2 estimate μ " σ^2 Population

3. Choice of the sample

The Sampling Process

30, 42, 48, 49, 61, 54, 41, 38, 59, 57



Calculate Mean = 47.9

This sample mean is an ESTIMATE of the population mean.

POPULATION

$\mu =$ **Mean = 50**

Parameter

We should not be surprised that the estimate does not equal the true mean for the population!

The Sampling Process

Choice of sample? $\{X_1, X_2, \ldots, X_n\}$

independent and representative

$$\{X_1, X_2, \ldots, X_n\}$$

independent

Identically distributed

The Sampling Process

Choice of sample? $\{X_1, X_2, \ldots, X_n\}$

independent

$$f(X_1, X_2, \ldots, X_n) = f(X_1)f(X_2)\ldots f(X_n)$$

Joint pdf **Marginal pdf**

Selection of $\boldsymbol{X_i}$ does not affect selection of $\boldsymbol{X_{i+1}}$

Identically distributed

$X_1, X_2, \ldots, X_n$ have the same distribution

- For a population $\sim$ distr(μ, σ^2)
- $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$ (the population mean)
- $V(X_1) = V(X_2) = \cdots = V(X_n) = \sigma^2$ (the population variance)

The Sampling Process

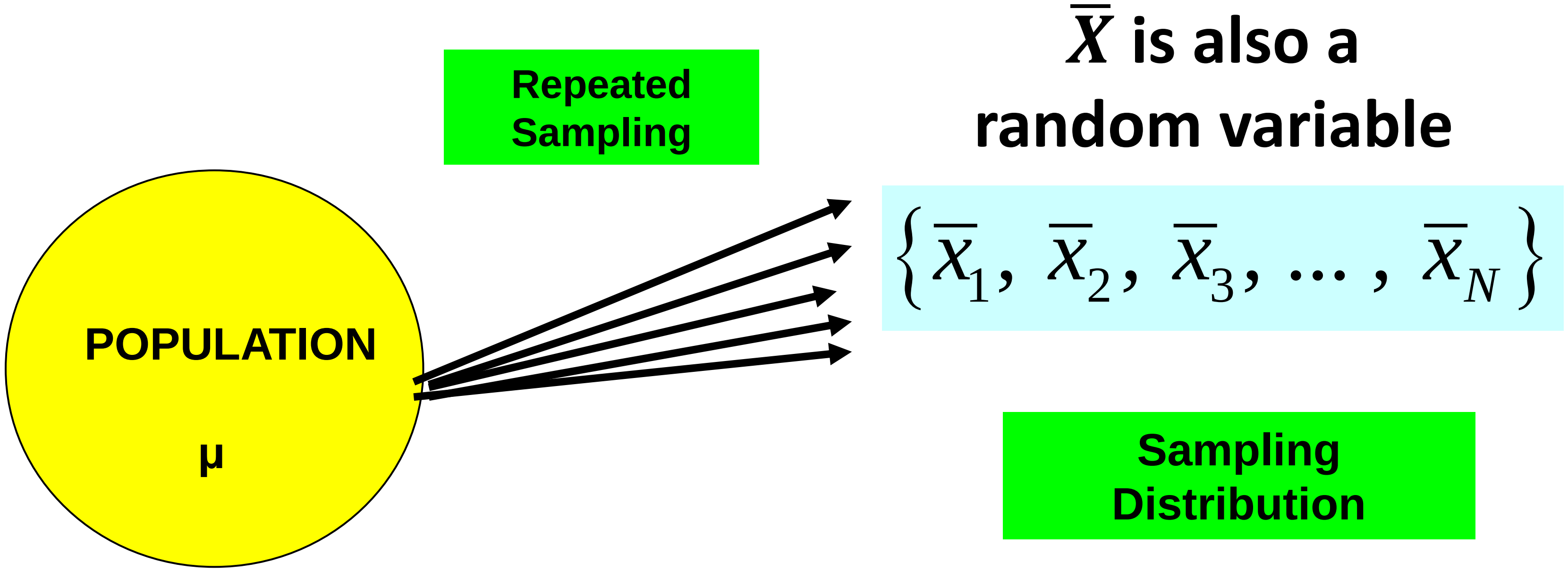
Choice of sample?

$\{X_1, X_2, \ldots, X_n\}$ are i.i.d. random variables

i.i.d. : independent and identically distributed

4. Distribution of $\bar{X}$

The Sampling Distribution



Sampling distribution

Definition:

The probability distribution of a statistic is called a sampling distribution.

Example: If $X_1, X_2, \dots, X_n$ represents a random sample of size n, then the probability distribution of $\bar{X}$ is called the sampling distribution of the sample mean $\bar{X}$.

Sampling Distribution of Means

$\frac{\sum x_i}{n} = \bar{X}$ estimate μ

$\frac{\sum (x_i-\bar{x})^2}{n-1} = S^2$ $\sim$ σ^2

Result:

If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a **normal distribution** with mean μ and variance σ^2, i.e. $N(\mu, \sigma^2)$, then the sample mean $\bar{X}$ has a normal distribution with mean

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

and variance

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Proof: Since $\bar{X}$ is the mean of a r.s. , then $X_1, X_2, \ldots, X_n$ are i.i.d. with common M.G.F.,

$M_x(t) = E[e^{tx}]$

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

The moment generating function corresponding to the normal probability density function $N(x; \mu, \sigma^2)$ is the function $M_x(t) = \exp\{\mu t + \sigma^2 t^2/2\}$.

Now, the M.G.F. of $\bar{X}$ is,

$$M_{\bar{X}}(t) = E\left[e^{\bar{X}t}\right] = E[e^{(\frac{1}{n}\sum_{i=1}^{n} X_i)t}]$$

$$= E[e^{X_1(\frac{t}{n})}]\,E[e^{X_2(\frac{t}{n})}] \ldots E[e^{X_n(\frac{t}{n})}]$$

$e^{x_1+x_2+\cdots+x_n} = e^{x_1} e^{x_2} \cdots e^{x_n}$

$E(x_1 x_2) = E(x_1)E(x_2)$

indep.

$$= M_{X_1}\left(\frac{t}{n}\right) M_{X_2}\left(\frac{t}{n}\right) \ldots M_{X_n}\left(\frac{t}{n}\right)$$

$$= \left[M_X\left(\frac{t}{n}\right)\right]^n$$

$$= \left(e^{\mu(\frac{t}{n}) + \frac{1}{2}\sigma^2(\frac{t}{n})^2}\right)^n$$

$$= e^{\mu t + \frac{1}{2}(\frac{\sigma^2}{n})t^2}$$

$\sim N(\mu, \frac{\sigma^2}{n})$

which is the M.G.F. of $N(\mu, \sigma^2/n)$. Therefore $\bar{X} \sim N(\mu, \sigma^2/n)$ i.e. $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

From this theorem we also conclude that the mean and variance of $\bar{X}$ are given by

$$\mu_{\bar{X}} = E(\bar{X}) = \mu, \quad \text{and} \quad \sigma^2_{\bar{X}} = \operatorname{var}(\bar{X}) = \frac{\sigma^2}{n}$$

<u>Proof:</u>

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}\,E\left(\sum X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}\,E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\left(\cancelto{\mu}{E(X_1)} + \cancelto{\mu}{E(X_2)} + \cdots + \cancelto{\mu}{E(X_n)}\right)$$

$$E(\bar{X}) = \frac{1}{n}\,n\,\mu = \mu$$

$V(ax+b) = a^2\,V(x)$

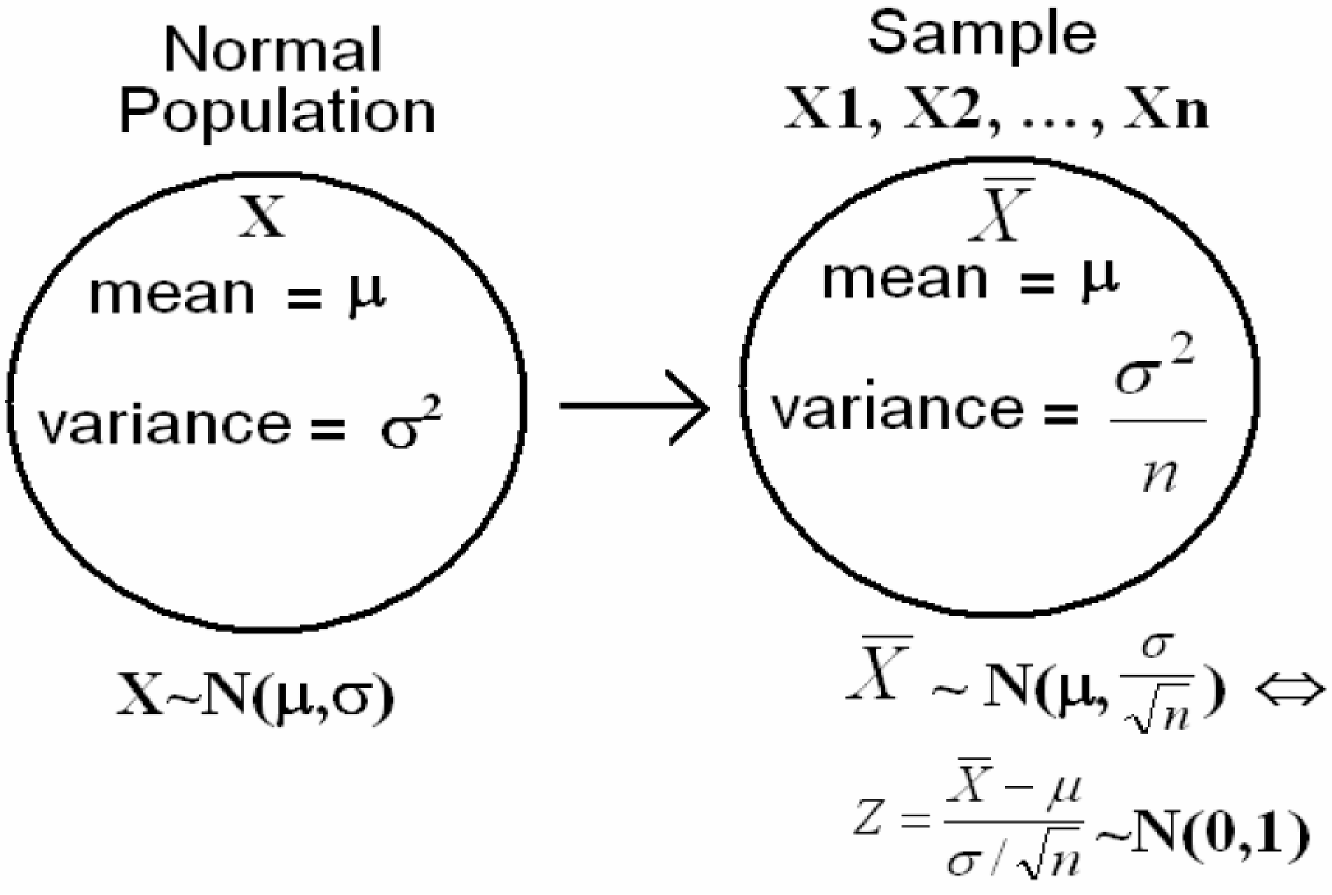
$X_i \sim N(\mu, \sigma^2)$

$$V(\bar{X}) = V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2}\,V\left(\sum X_i\right)$$

$$V(\bar{X}) = \frac{1}{n^2}\,V(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}\left(\cancelto{\sigma^2}{V(X_1)} + \cancelto{\sigma^2}{V(X_2)} + \cdots + \cancelto{\sigma^2}{V(X_n)}\right)$$

$$E(\bar{X}) = \frac{1}{n^2}\,n\,\sigma^2 = \frac{\sigma^2}{n}$$

- If $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\mu,\sigma)$, then $\overline{X} \sim N(\mu_{\overline{X}}, \sigma_{\overline{X}})$ or $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.
- $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$



What if the population is not normally distributed?

Theorem: (Central Limit Theorem)

If X_1, X_2, …, X_n is a random sample of size n from any distribution (population) with mean μ and finite variance σ^2, then, if the sample size n is large, the random variable

$n \geq 30$

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

is approximately standard normal random variable, i.e.,

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ approximately.

Example

$\bar{X} \sim$ Normal

An electric firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

μ σ n

$P(\bar{X} < 775)$ المطلوب

Solution:

X= the length of life

$\mu=800$, $\sigma=40$

$X \sim N(800, 40)$

$n=16$

$$\mu_{\overline{X}} = \mu = 800$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$$

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(800, 10)$$

$Var(\bar{x}) = \frac{\sigma^2}{n}$

St. dev. $(\bar{x})$

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$

$\bar{x} \sim N$

$P(\bar{x} < 775)$

$$\Leftrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = Z = \frac{\bar{X} - 800}{10} \sim N(0,1)$$

$$P(\bar{X} < 775) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{775 - \mu}{\sigma / \sqrt{n}} \right)$$

$$\frac{-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \qquad \frac{-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

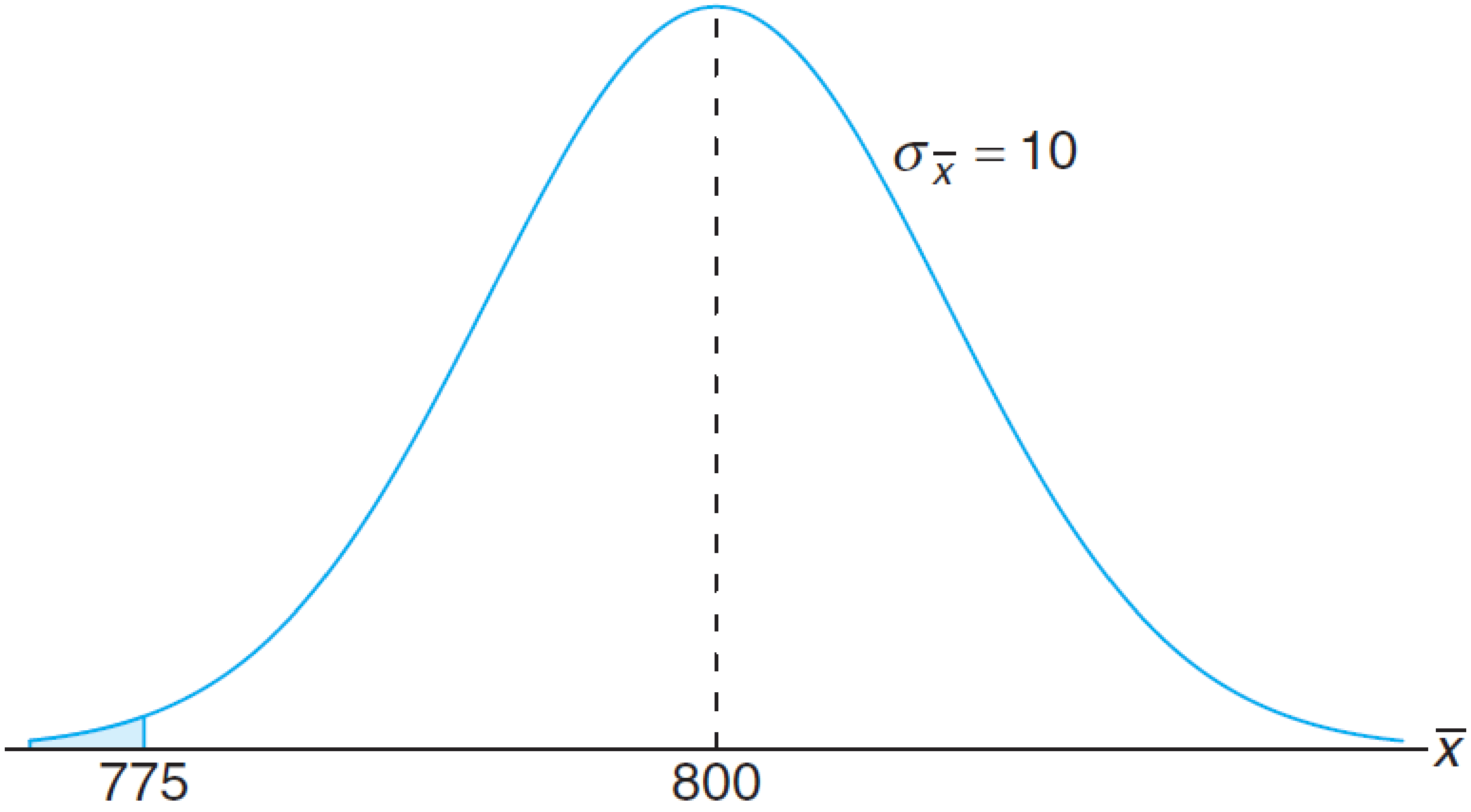
$$= P\left(\frac{\bar{X} - 800}{10} < \frac{775 - 800}{10} \right)$$

$$= P\left(Z < \frac{775 - 800}{10} \right)$$

minitab R

$$= P(Z < -2.50) = 1 - \Phi(2.5)$$

$$= 0.0062$$


$\sigma_{\bar{x}} = 10$
775
800
$\bar{x}$

Example

Certain tubes manufactured by a company have a

μ σ

mean lifetime of 900 hr and standard deviation of 50 hr. Find the probability that a random sample of 64 ≥ 30

C.L.T.

tubes taken from the group will have a mean

Normal

lifetime between 895 and 910 hrs.

Solution:

Here we have $\mu = 900$, $\sigma = 50$.

Since n = 64 is large enough, then by the central limit theorem

$\Phi(1.6) - \Phi(-0.8)$

$-(1 - \Phi(0.8))$

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

$$P(895 < \bar{X} < 910) = P\left(\frac{895 - 900}{50/8} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{910 - 900}{50/8}\right) = P(-8.0 < Z < 1.6)$$

$$= \Phi(1.6) - \Phi(-0.8) = \Phi(1.6) - 1 + \Phi(0.8) = 0.733$$

Case of σ^2 is unknown

If $\boxed{n \geq 30}$, the central limit theorem (CLT) is still valid.

If we replace σ^2 by s^2

Population Samples

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

Summary

Given the random sample $X_1, X_2, \ldots, X_n$.
We have **three cases** for the distribution of $\bar{X}$

[1]

- Sample taken from a normal population
- σ^2 is known

$$\bar{X} \sim Norm\left(\mu, \frac{\sigma^2}{n}\right)$$

[2]

- $n \geq 30$ and sample taken from any distribution

$$\bar{X} \sim Norm\left(\mu, \frac{\sigma^2}{n}\right)$$

σ known

σ unknow

$$\bar{X} \sim Norm\left(\mu, \frac{s^2}{n}\right)$$

[3]

- $n < 30$
- σ^2 is unknown
- Next lecture