Sampling Distribution and Central Limit Theorem

Sampling distribution and Central Limit Theorem

Learning Outcomes

At the end of this lesson you should be able to:

1. Define the sampling distribution of the sample means with reference to inferential statistics.
2. Calculate probabilities involving the sample mean by using the standard normal distribution and the central limit theorem.
3. Apply probabilities of the sample mean values in real-life business problems by applying concepts on the sampling distribution of sample means

Introduction to Sampling Distribution

Your company's core business is in packaging Delicious Oatmeal Cereals into boxes. Thousands of boxes are packed every day. As a Quality Production Manager, you need to ensure that mean weight of each box of cereal is 360 gm.

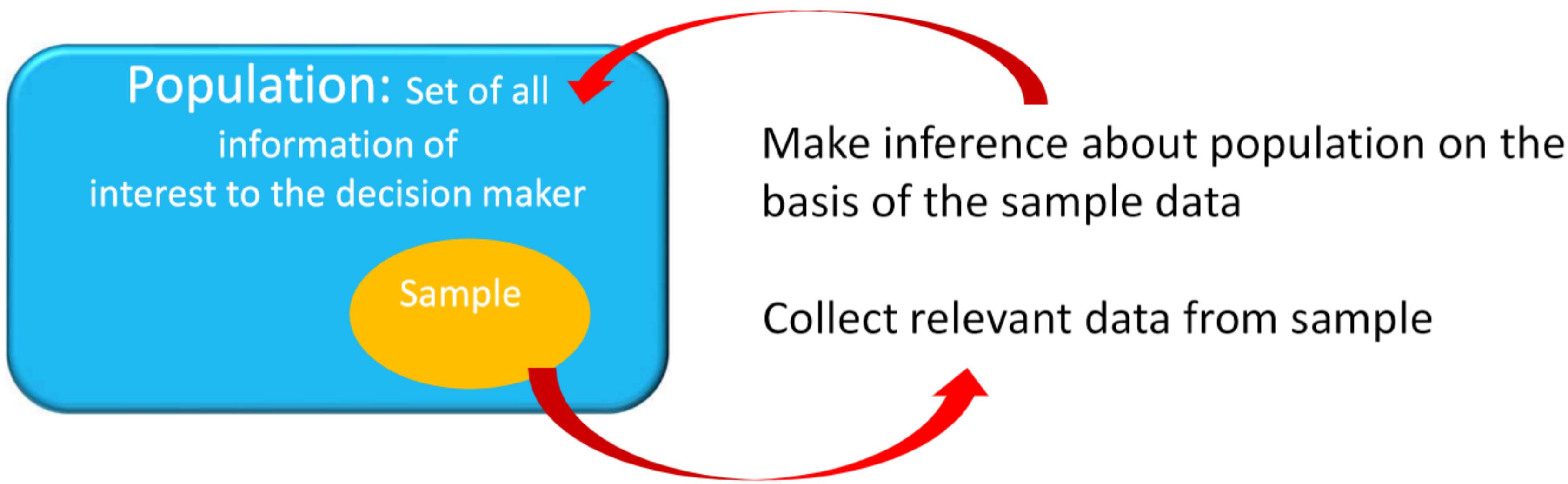
What can you do?

You can weigh all the boxes, but it will be to costly and time consuming. Instead, you collect data from samples taken from the production. Based on the data, you will decide whether to continue or stop the production.

- This case study is an example of application of sampling.
- In this topic, you will apply concepts of sampling distribution to solve real life problems.

What is Sampling?

- The purpose of **inferential statistics** is to find something about a population based on a sample.



Reasons for Sampling

- There are many reasons for sampling, some of which are listed here:
  - ✓ Too time-consuming
  - ✓ Too costly to study all the items in the population
  - ✓ Impossible to locate all the items in the population
  - ✓ Results of a sample may adequately estimate the value of the population parameter, saving time and money

Sampling Error

- It is unlikely the mean of a sample will be exactly equal to the mean of the population. For example, in the Quaker Life Cereals example, it is unlikely that the sample mean of the 30 boxes is the same the as the population mean of 360 gm.

Sampling Error

Difference between a sample mean, $\bar{x}$ and its corresponding mean, μ.

- Example: In the Quaker Life Cereals example, the sample mean $\bar{x}$ of the 30 boxes is 380 gm. Hence, the sampling error $(\bar{x} - \mu) = 380 - 360 = 20$ gm.

Sampling Distribution of the Sample Mean

- In the Quaker Life Cereals example, if we take sample boxes in production and weigh them, we will have many sample means

$\bar{x}_1$

$\bar{x}_2$

$\bar{x}_3$

...... $\bar{x}_n$.

- $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ ...... $\bar{x}_n$. will very likely have different values, since the $\bar{x}$ of each sample will depend on the weight of each box taken from the population.
- Hence in general, the sample mean, $\bar{X}$, is a random variable (as we cannot determine the actual value of x for a randomly chosen sample.)

Sampling Distribution of the Sample Mean

A probability distribution of all possible sample means of a given sample size

Properties of a Sampling Distribution of the Sample Mean

When the population distribution of X is **normal** with mean μ and standard deviation, σ then

- ✓ The mean of the sample means, $\mu_{\bar{x}}$, is the same as the population mean μ i.e. $\boldsymbol{\mu_{\bar{x}} = \mu}$
- ✓ The standard deviation of the sample mean, $\sigma_{\bar{x}}$, is given by $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$ where n is the sample size. $\boldsymbol{\sigma_{\bar{x}}}$ is also called **the standard error of the mean**.

Sampling from Normal Populations

- We have seen from previous slide that, when the population distribution of X is **normal** with mean μ and standard deviation σ, then the sampling distribution of the sample mean is also **normally** distributed with mean, $\mu_{\bar{x}}$, and standard deviation, $\sigma_{\bar{x}}$.
- This is denoted as follows:

$$X \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$$

where mean of the sample means, $\mu_{\bar{x}} = \mu$

standard deviation of the sample means, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

n is the sample size

Sampling from Normal Populations

- You have learnt in the topic on Normal Probability Distribution that we can convert any *normal* variable to a *standard normal* variable using

$$Z = \frac{x - \mu}{\sigma}$$

- In Sampling Distribution, the following formula will be used:

$$Z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}}$$

$\boldsymbol{\mu_{\overline{x}}}$ = mean of the sample mean

$\boldsymbol{\sigma_{\overline{x}}}$ = standard deviation of the sample mean

$\boldsymbol{\sigma_{\overline{x}}}$ is also called the standard error of the mean

Sampling from Normal Populations

- Example1: The weight of onion bulb produced by farm Green Life is approximately normally distributed with a mean of 60 g and standard deviation of 5 g. Find the probability that a random sample of 16 garlic bulbs will have an average weight of less than 62 g.

Let X represent the weight of onion bulb.

$\mu = 60 \quad \sigma = 5 \quad n = 16$

$X \sim N(60, 5^2) \Rightarrow \bar{X} \sim N(\mu_{\bar{x}}, {\sigma_{\bar{x}}}^2)$

where $\mu = \mu_{\bar{x}} = 60$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{16}} = 1.25$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{62 - 60}{1.25} = 1.6$$

$P(\bar{X} \leq 62) = P(Z \leq 1.6) = \underline{0.9452}$

Sampling from Normal Populations

- Example 2: Physical fitness score of a certain population of executives is normally distributed with mean and standard deviation of 75 and 10 respectively. What is the probability that a random sample of 25 such executives has a mean score between 70 and 78?

Let X represent the physical fitness score.

$\mu = 75 \quad \sigma = 10 \qquad n = 25$

$X \sim N(75,10^2) \Rightarrow \bar{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}{}^2)$

where $\mu = \mu_{\bar{x}} = 75$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}}$

Sampling from Normal Populations

$$\bar{x}_1 = 70;\ z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{70 - 75}{2} = -2.5$$

$$\bar{x}_2 = 78;\ z_2 = \frac{\bar{x}_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{78 - 75}{2} = 1.5$$

$$P(70 \le \bar{X} \le 78) = P(-2.5 \le Z \le 1.5)$$

$$= P(Z \le 1.5) - P(Z \le -2.5)$$

$$= 0.9332 - 0.0062$$

$$= \underline{0.9270}$$

Sampling from Normal Populations

Example 3: A bank calculates that its individual savings accounts are normally distributed with a mean of $2,000 and a standard deviation of $600. If the bank takes a random sample of 100 accounts, what is the probability that sample mean will lie between $1,900 and $2,050? (Assume population is infinite.)

Let X represent amount in individual savings account.

$\mu = 2000$ $\sigma = 600$ n = 100

Sampling from Normal Populations

$X \sim N(2000, 600^2) \Rightarrow \bar{X} \sim N(\mu_{\bar{x}}, {\sigma_{\bar{x}}}^2)$

where $\mu = \mu_{\bar{x}} = 2000$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{600}{\sqrt{100}} = 60$

$\bar{x}_1 = 1900;\ z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1900 - 2000}{60} = -1.67$

$\bar{x}_2 = 2050;\ z_2 = \frac{\bar{x}_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{2050 - 2000}{60} = 0.83$

$P(1900 \le \bar{X} \le 2050) = P(-1.67 \le Z \le 0.83)$

$= P(Z \le 0.83) - P(Z \le -1.67)$

$= 0.7967 - 0.0475$

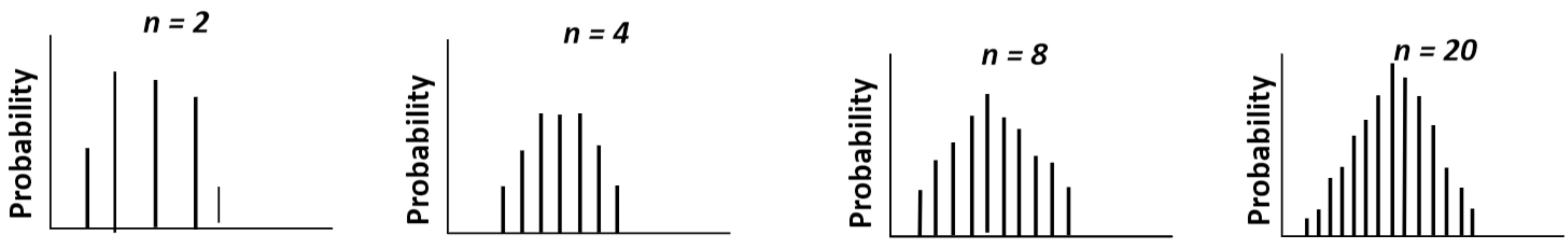
$= \underline{0.7492}$

Central Limit Theorem

- In the last 3 examples, the population is normally distributed. For any sample size, the sampling distribution of the sample mean will also be normally distributed. However, in real life, we know that in many cases, the population is **not** normally distributed.

Central Limit Theorem(CLT)

As sample size gets *large enough*(generally, $n \geq 30$), the sampling distribution will follow an approximately normally distribution



Sampling from a Non-Normal Population Shape of Sampling Distribution

Significance of Central Limit Theorem

- The CLT permits us to use sample statistics to make inferences about population parameters **without knowing anything about the shape of the probability distribution** of the population.
- If we repeatedly take sufficient large samples from the population, from say, a exponential distribution, the resulting sampling distribution will follow an approximately normally distribution.
- However, if the population being samples is normal, then the CLT is not necessary, as the sampling distribution will be normally distributed.
- When CLT is applied, we use the notation:

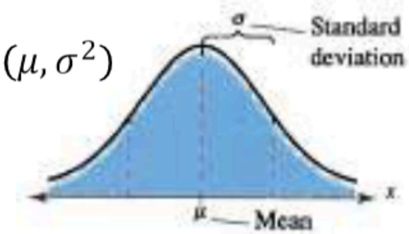
By CLT, $\overline{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$ where $\mu_{\bar{x}} = \mu$

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

Summary of Central Limit Theorem

Population Distribution - Normal

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



Distribution of Sample Means, (any *n*)

$\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$

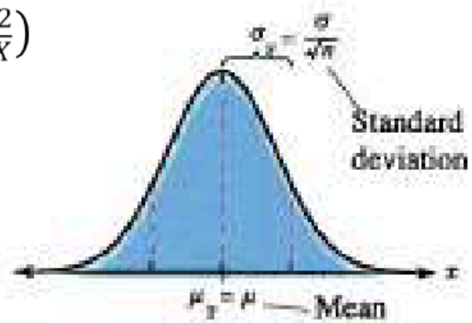
$where$

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

$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$

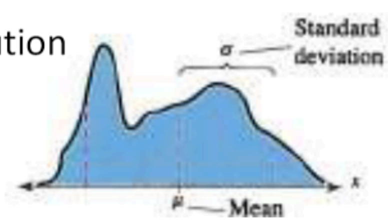
$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$=$ standard error of the mean



Population Distribution - Not Normal

$X \sim$any distribution



Distribution of Sample Means, ($n \geq 30$)

$\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$

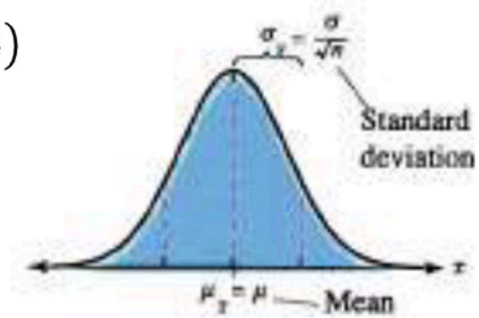
$where$

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

$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$

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

$=$ standard error of the mean



Central Limit Theorem

Example 4: On weekdays at noon, the number of people in a cafeteria has a mean, $\mu = 140$ and standard deviation, $\sigma = 50$. The manager counts the number of noontime customers on 100 randomly selected days and calculates the sample mean. What is the probability that the sample mean is between 130 and 150? (**Note that the population distribution is unknown) .**

Central Limit Theorem

Let X represent the number of customers at noontime on weekdays

$\mu = 140 \qquad \sigma = 50 \qquad n = 100$

Since n=100 > 30, by CLT, $\bar{X} \sim N(\mu_{\bar{x}}, {\sigma_{\bar{x}}}^2)$

where $\mu_{\bar{x}} = \mu = 140$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5$

$\bar{x}_1 = 130; \; z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{130 - 140}{5} = -2.0$

$\bar{x}_2 = 150; \; z_2 = \frac{\bar{x}_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{150 - 140}{5} = 2.0$

Central Limit Theorem

$P(130 \leq \bar{X} \leq 150) = P(-2.0 \leq Z \leq 2.0)$

$= P(Z \leq 2.0) - P(Z \leq -2.0)$

$= 0.9772 - 0.0228$

$= \underline{0.9544}$

Central Limit Theorem

Example 5: The mean life of a certain saw blade is 41.5 hours, with a standard deviation of 2.5 hours. What is the probability that a simple random sample of 50 blades has a mean of between 40.5 and 42 hours?

Let X represent the life (in hours) of the saw blade

$\mu = 41.5$ $\sigma = 2.5$ $n = 50$

Since n=50 > 30, by CLT, $\bar{X} \sim N(\mu_{\bar{X}}, {\sigma_{\bar{X}}}^2)$

where $\mu = \mu_{\bar{x}} = 41.5$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{50}} = 0.3536$

Central Limit Theorem

$$\bar{x}_1 = 40.5;\ z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{40.5 - 41.5}{0.3536} = -2.83$$

$$\bar{x}_2 = 42.0;\ z_2 = \frac{\bar{x}_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{42 - 41.5}{0.3536} = 1.41$$

$$P(40.5 \le \bar{X} \le 42.0) = P(-2.83 \le Z \le 1.41)$$

$$= P(Z \le 1.41) - P(Z \le -2.83)$$

$$= 0.9207 - 0.0023$$

$$= \underline{0.9184}$$