# Chapter 5 (Statistics) Sampling (Teacher's copy)

## Objectives

At the end of the chapter, you should be able to:

(a) understand the concepts of population, simple random sample;

(b) understand the use of and calculate the unbiased estimates of the population mean and variance, including cases where the data are given in summarised form $\sum x$ and $\sum x^2$, or $\sum (x-a)$ and $\sum (x-a)^2$;

(c) understand that the sample mean $\overline{X}$ is a random variable with $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$ ;

(d) use the fact that $\overline{X}$ has a normal distribution if $X$ has a normal distribution;

(e) apply the sampling distribution $\mathrm{N}(\mu, \dfrac{\sigma^2}{n})$ to solve statistical problems in real-world situations;

(f) use the Central Limit Theorem to treat sample mean as having a normal distribution when the population is not normally distributed and the sample size is sufficiently large. [Note: 'Large' samples will usually be of size at least 30, but students should know that using the approximation of normality can sometimes be useful with samples that are smaller than this.]

## Content

## 5.1 Introduction and Definitions

In real life situations, we often need to conduct a statistical enquiry for different purposes. For example, a school may want to gather information about the studying habits of the students. A manufacturer of batteries may want to find out the length of lifespan of his product.

For such purposes, we need information to draw valid conclusions about a group of individuals or objects. This group of individuals or objects is called a population.

A **population** is defined as the entire collection of objects which a statistician is interested to study on. The size of a population can be finite or infinite. In most situations, it is impossible or impractical to examine every individual or object of the whole population for various reasons, such as:

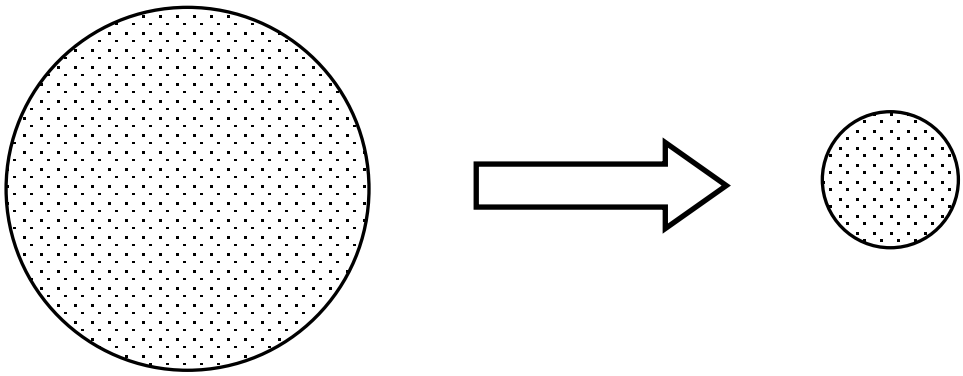
(a) The population is large or infinite.
E.g. the number of people who earned less than US$600 a month.

(b) The collection of information may destroy the sample.
E.g. when testing batteries, fireworks, electric fuses etc.

(c) It may be impossible to gain access to every member of the population.
E.g. measuring the lengths of ants of a particular species.

Therefore it is more practical to conduct a careful study of a sample. A **sample** is a portion or subset of objects drawn from the population.

For example,




This process of obtaining samples of the population is called **sampling** and each object of the population is called a sampling unit. i.e**. Sampling units** are the individual members of the target population whose characteristics are to be measured. In the above example, a sampling unit is a student from SAJC.

A **sampling frame** is a complete list of all the members of the population. i.e. the list of all the objects from which the sample is to be chosen. Examples are school registers, membership lists,

the electoral register etc. A problem, of course, is that the list may not be up to date. In some cases, a list may not even exist.

Usually, the objective of taking samples is to obtain estimates of **population parameters**, such as the population mean $\mu$ or the population variance $\sigma^2$. For example, a population parameter can be the height of a student in SAJC.

Sample estimates of population parameters are called **statistics**. A sample statistic is said to be useful if it can be used to estimate an unknown parameter of the population.

For example in the above example, to estimate the average height of students from SAJC, we can take a sample and compute the mean; this sample mean is a statistic that estimates the population mean (population parameter).

The accuracy of our estimates will depend on the method of taking the sample and the sample size.

This process of drawing valid conclusions about the population based on the results found from the sample is known as **statistical inference** (specific $\rightarrow$ general).

**What makes a good sample?**

To draw valid conclusions about the population, a sample must be **free of selection bias** and **representative**.

A sample is said to be **unbiased** if every individual in the population has an equal chance of being selected.
Eg: If we are interested in the distribution of the shoe sizes of male students in SAJC, then we should not stand at the finish line of a cross-country race and survey the first 50 boys who cross the line. The sample will be biased towards good runners.

A sample is said to be **maximally representative** if the characteristics of the sample represent (as accurately as possible) the entire population.
Eg: If we are interested in the characteristics of army officers, we should make sure each different type of officer is represented in order for the sample to be maximally representative.

## 5.1.1 Simple Random Samples

**Random Samples**
A sample is said to be a random sample (or a randomly chosen sample) if the method of selecting the sample is such that every member of the population has an ***equal*** probability of being selected and each selection is ***independent*** of each other. When carrying out a **random sample**, you must ensure that all possible samples are equally likely to be chosen.

The simplest type of random sample is a *simple random sample* of size *n* which is drawn from a population of size *N* in such a manner that every possible sample of size *n* is equally likely to be selected. By this method, **all members of the population** have **an equal chance of being selected**. Random sampling methods ensure that bias does not exist.

## Exercise 1

1. [H1 Math/N2011/A-level/Q7 modified]
Two thousand students travel to college either by car, by bicycle or on foot. Any given student travels by the same method each day. A researcher carries out a survey to investigate the length of students' journey times to college, using a random sample of 100 students.
Explain what is meant in this context by the term 'a random sample'.

**Solution:**
A random sample is a sample drawn in such a way that each student who travels to college must have an **equal chance** of being selected as a member of the sample and the students must be selected **independently** i.e. one student being chosen does not have any effect on the chances of any other student then being chosen. In other words, the first student in the sample is chosen at random from the population of 2000 students. Then for the second member of the sample, each of the remaining 1999 students has an equal chance of being selected, and so on.

Note:
[From Examiner's report]
Few candidates were able to give an appropriate explanation of the term 'random sample'. The majority of answers consisted of either reversing the request and talking about a sample that was chosen randomly, or explaining how to find a random sample rather than explaining the meaning of the term. Some candidates did refer to there being an equal opportunity of selection, but hardly any mentioned the need for independence.

2. [H2 Math/N2019/A-level/Q6 (i) and (ii)]

In a certain country there are 100 professional football clubs, arranged in 4 divisions. There are 22 clubs in Division One, 24 in Division Two, 26 in Division Three and 28 in Division Four.

(i) Alice wishes to find out about approaches to training by clubs in Division One, so she sends a questionnaire to the 22 clubs in Division One, Explain whether these 22 clubs form a sample or a population.

**Solution:**
(i) These 22 clubs form a **population**, since Alice was sending the questionaire to all the clubs in Division one. she was interested in.

{Note to students : It is not a sample as a sample would comprise a fraction of the clubs in Division One, rather than *all* the clubs.}

(ii) Dilip wishes to investigate the facilities for supporters at the football clubs, but does not want to obtain the detailed information necessary from all 100 clubs. Explain how he should carry out his investigation, and why he should do the investigation in this way.

**Solution:**

(ii) The **sampling frame** is made up by the 100 professional football clubs.

Dilip should carry out a random sample of the 100 football clubs by first numbering the clubs 1 to 100.

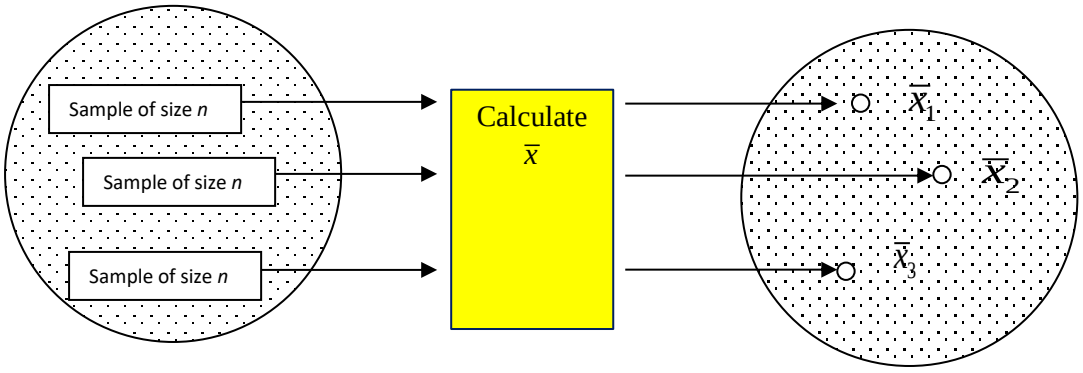
If he wants to investigate 20 clubs to investigate, he should randomly draw 20 distinct numbers from 1 to 100 and investigate those clubs whose numbers were drawn. This is to ensure that the sample he took is random so that he could avoid bias.

[A level examiner report: explicit knowledge on how to to obtain the sample is not required.]

## 5.2 The Sample Mean as a Random Variable

Take a random sample of *n* independent observations from a population. Calculate the mean of these *n* sample values. This is known as the sample mean. Now repeat the procedure until you have taken all possible samples of size *n*, calculating the sample mean of each one. Form a distribution of all the sample means calculated.

The distribution of the sample mean that would be formed is called the sampling distribution of the sample mean.




Population with

Mean = E($X$) = $\mu$

Standard deviation

= $\sqrt{\text{Var}(X)} = \sigma$

Repeat this procedure.

Sampling distribution of $\bar{X}$

Mean = $\text{E}(\bar{X})$

Standard Deviation = $\sqrt{\text{Var}(\bar{X})}$

### 5.2.1 Mean and Variance of the Distribution of Sample Mean

Since the sample mean $\bar{X}$ is a random variable, it makes sense to talk about the expected value of the sample mean, E( $\bar{X}$ ) and variance of the sample mean, Var( $\bar{X}$ ).

Consider a population $X$ in which

$$\text{E}(X) = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2.$$

Take $n$ independent observations $X_1, X_2, ..., X_n$ from $X$. We know that each $X_i$ has the same probability distribution as $X$ and so

$$\mathrm{E}(X_i) = \mathrm{E}(X) = \mu \qquad \text{and} \qquad \mathrm{Var}(X_i) = \mathrm{Var}(X) = \sigma^2$$

Then the **sample mean** $\overline{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$ is a random variable with mean $\mu$ and variance $\dfrac{\sigma^2}{n}$, i.e. $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$

and the **sample sum** $X_1 + X_2 + \ldots + X_n$ is a random variable with mean $n\mu$ and variance $n\sigma^2$.
i.e. $\mathrm{E}(X_1 + X_2 + ... + X_n) = n\mu$ and $\mathrm{Var}(X_1 + X_2 + ... + X_n) = n\sigma^2$

Proof:

$$\begin{aligned}
&\mathrm{E}\left(\overline{X}\right) \\
&= \mathrm{E}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\
&= \mathrm{E}\left(\frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)\right) \\
&= \frac{1}{n}\mathrm{E}\left(X_1 + X_2 + \ldots + X_n\right) \\
&= \frac{1}{n}\left(\mathrm{E}\left(X_1\right) + \mathrm{E}\left(X_2\right) + \ldots + \mathrm{E}\left(X_n\right)\right) \\
&= \frac{1}{n}\left(n\mathrm{E}\left(X\right)\right) \\
&= \mathrm{E}\left(X\right) \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Var}\left(\overline{X}\right) \\
&= \mathrm{Var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\
&= \mathrm{Var}\left(\frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(X_1 + X_2 + \ldots + X_n\right) \\
&= \frac{1}{n^2}\left(\mathrm{Var}\left(X_1\right) + \mathrm{Var}\left(X_2\right) + \ldots + \mathrm{Var}\left(X_n\right)\right) \\
&= \frac{1}{n^2}\left(n\mathrm{Var}\left(X\right)\right) \\
&= \frac{1}{n}.\mathrm{Var}\left(X\right) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Note:

$$X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i = \sum X$$

## 5.3 The Distribution of the Sample Mean and Sample Sum

The distribution of $\overline{X}$ is called the sampling distribution of the sample mean.
The distribution of $X_1 + X_2 + \ldots + X_n$ is called the sampling distribution of the sample sum.

The sampling distribution depends on the distribution of the population and the sample size.

We will study the following two types, namely

- Sampling from a Normal population and
- Sampling from a Non – Normal population.

## 5.3.1 Distributions of Sample Mean and Sample Sum from a Normal population

If $X_1, X_2, \ldots\ldots, X_n$ is a random sample of $n$ independent observations taken from a ***normally distributed population,*** $X$ with mean $\mu$ and variance $\sigma^2$ , i.e. $X \sim \mathrm{N}(\mu, \sigma^2)$, then

1. $X_1 + X_2 + \ldots + X_n \sim \mathrm{N}\left(n\mu, n\sigma^2\right)$ ← Distribution of Sample Sum
2. $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ ← Distribution of Sample Mean

**Example 1**

The heights of a particular species of plant follow a normal distribution with mean 21 cm and standard deviation $\sqrt{90}$ cm. $\overline{X}$ is the mean height of 10 randomly chosen plants and $T$ is the total height of 20 randomly chosen plants. Find the distributions of $\overline{X}$ and $T$.

**Solution:**

Let $X$ be the random variable denoting height (in cm) of a plant.

$X \sim \mathrm{N}(21, 90)$

$\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_{10}}{10}$

$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\left(X\right) = 21$

$\mathrm{Var}\left(\overline{X}\right) = \dfrac{\mathrm{Var}\left(X\right)}{10} = \dfrac{90}{10} = 9$

$\therefore \overline{X} \sim \mathrm{N}(21, 9)$

$T = X_1 + X_2 + \ldots + X_{20}$

$\mathrm{E}\left(T\right) = 20\left(\mathrm{E}\left(X\right)\right) = 20\left(21\right) = 420$

$\mathrm{Var}\left(T\right) = 20\left(\mathrm{Var}\left(X\right)\right) = 20\left(90\right) = 1800$

$\therefore T \sim \mathrm{N}\left(420, 1800\right)$

**Example 2**

Given that the random variables $X$ is normally distributed with mean 40 and variance 9, and $Y$ is normally distributed with mean 20 and variance 16. 10 independent observations of $X$ and 8 independent observations of $Y$ are taken. State the sampling distribution of

(i) $\overline{X} - \overline{Y}$ (ii) $5\overline{X} + 3\overline{Y}$

**Solution:**

Given $X \sim \mathrm{N}(40, 9)$, $Y \sim \mathrm{N}(20, 16)$.

$\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_{10}}{10}$, $\therefore \overline{X} \sim \mathrm{N}(40, \dfrac{9}{10})$. i.e. $\overline{X} \sim \mathrm{N}(40, 0.9)$.

Also, $\overline{Y} = \dfrac{Y_1 + Y_2 + \cdots + Y_8}{8}$, $\therefore \overline{Y} \sim \mathrm{N}(20, \dfrac{16}{8})$. i.e. $\overline{Y} \sim \mathrm{N}(20, 2)$.

(i) $\mathrm{E}(\overline{X} - \overline{Y}) = 40 - 20 = 20$

$\mathrm{Var}(\overline{X} - \overline{Y}) = 0.9 + 2 = 2.9$

$\therefore \overline{X} - \overline{Y} \sim \mathrm{N}(20, 2.9)$

(ii) $\mathrm{E}(5\overline{X} + 3\overline{Y}) = 5(40) + 3(20) = 260$

$\mathrm{Var}(5\overline{X} + 3\overline{Y})$

$= 5^2(0.9) + 3^2(2) = 40.5$

$\therefore 5\overline{X} + 3\overline{Y} \sim \mathrm{N}(260, 40.5)$

**Example 3**

A large number of random samples of size $n$ are taken from a normal population with mean 75 and variance 36. Find the largest sample size if at most 80% of the sample means exceeds 74.

**Solution:**

$X \sim \mathrm{N}(75, 36)$

$\bar{X} \sim \mathrm{N}\left(75, \frac{36}{n}\right)$

$$\mathrm{P}\left(\bar{X} > 74\right) \le 0.80$$

$$\mathrm{P}\left(Z > \frac{74-75}{\sqrt{\frac{36}{n}}}\right) \le 0.80$$

$$\mathrm{P}\left(Z > -\frac{\sqrt{n}}{6}\right) \le 0.80$$

From GC,

$$-\frac{\sqrt{n}}{6} \ge -0.84162$$

$$\frac{\sqrt{n}}{6} \le 0.84162$$

$$\sqrt{n} \le 5.0497$$

$$n \le 25.5$$

Thus, largest sample size is 25.

**Alternative Method**

$\mathrm{P}\left(\bar{X} > 74\right) \le 0.80$

Using GC,

| $n$ | $\mathrm{P}\left(\bar{X} > 74\right)$ |
|---|---|
| 24 | 0.79289 < 0.80 |
| 25 | 0.79767 < 0.80 |
| 26 | 0.80229 > 0.80 |

$\therefore$ Largest $n$ is 25.

## Exercise 2

1. Given that the random variables $X$ is normally distributed with mean 30 and variance 18, and $Y$ is normally distributed with mean 20 and variance 16. 15 independent observations of $X$ and 8 independent observations of $Y$ are taken. State the sampling distribution of

(i) $\bar{X}$ (ii) $5\bar{X}-3\bar{Y}$ (iii) mean of the sum of all observations

[Ans: (i) $\bar{X} \sim \mathrm{N}(30,1.2)$ (ii) $5\bar{X}-3\bar{Y} \sim \mathrm{N}(90,48)$

(iii) $\dfrac{X_1+\cdots+X_{15}+Y_1+\cdots+Y_8}{23} \sim \mathrm{N}\left(\dfrac{610}{23},\dfrac{398}{529}\right)$]

**Solution:**

$X \sim \mathrm{N}(30,18)$

$Y \sim \mathrm{N}(20,16)$

(i) $\bar{X} \sim \mathrm{N}\left(30,\dfrac{18}{15}\right)$

i.e. $\bar{X} \sim \mathrm{N}(30,1.2)$

(ii) $\bar{Y} \sim \mathrm{N}\left(20,\dfrac{16}{8}\right)$

i.e. $\bar{Y} \sim \mathrm{N}(20,2)$

$5\bar{X}-3\bar{Y} \sim \mathrm{N}\left(5(30)-3(20),5^2(1.2)+3^2(2)\right)$

$5\bar{X}-3\bar{Y} \sim \mathrm{N}(90,48)$

(iii) $\dfrac{X_1+\cdots+X_{15}+Y_1+\cdots+Y_8}{23} \sim \mathrm{N}\left(\dfrac{15(30)+8(20)}{23},\dfrac{15(18)+8(16)}{23^2}\right)$

$\dfrac{X_1+\cdots+X_{15}+Y_1+\cdots+Y_8}{23} \sim \mathrm{N}\left(\dfrac{610}{23},\dfrac{398}{529}\right)$

2. $X_1, X_2, X_3, ..., X_8$ is a random sample of size 8 drawn from the distribution N(1,6). $Y_1, Y_2, Y_3, Y_4$ is a random sample of size 4 drawn from the distribution N(2,4). Find

(i) $\mathrm{P}(\bar{Y} \geq 1)$

(ii) $\mathrm{P}\left(2\bar{X} \geq 1-\bar{Y}\right)$.

[Ans: 0.841, 0.933]

**Solution:**

(i) $\bar{Y} \sim N(2,\dfrac{4}{4})$ i.e. $\bar{Y} \sim N(2,1)$

$\mathrm{P}(\bar{Y} \geq 1) = 0.841$

(ii) $\bar{X} \sim N(1,\dfrac{6}{8})$ i.e. $\bar{X} \sim N(1,0.75)$

$2\bar{X}+\bar{Y} \sim N(2(1)+2,4(0.75)+1)$

$2\bar{X}+\bar{Y} \sim N(4,4)$

$\mathrm{P}\left(2\bar{X} \geq 1-\bar{Y}\right) = \mathrm{P}\left(2\bar{X}+\bar{Y} \geq 1\right) = 0.933$

3. If $X_1, X_2, ..., X_n$ is a random sample from $X \sim \mathrm{N}(\mu, 1)$, state the distribution of the sample mean $\bar{X}$. Find the least sample size required to ensure that the probability that $\bar{X}$ is within 0.1 from $\mu$ is greater than 0.95. [Ans: 385]

**Solution:**

$X \sim \mathrm{N}(\mu, 1)$

$\bar{X} \sim \mathrm{N}(\mu, \frac{1}{n})$

$\mathrm{P}(\mu - 0.1 < \bar{X} < \mu + 0.1) > 0.95$

$\mathrm{P}\left(-\frac{0.1}{1/\sqrt{n}} < \frac{\bar{X}-\mu}{1/\sqrt{n}} < \frac{0.1}{1/\sqrt{n}}\right) > 0.95$

$\mathrm{P}(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) > 0.95$

From GC, $0.1\sqrt{n} > 1.9599$

$\sqrt{n} > 19.599$

$n > 384.12$

$\therefore$ Least sample size is 385.

Question Prompt

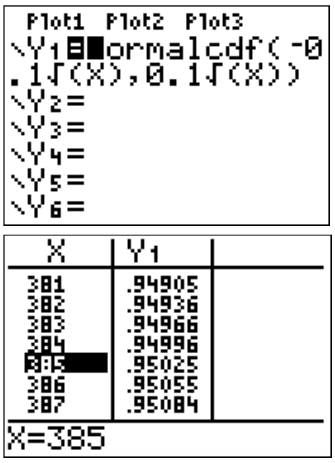
Is there any other way to solve $P(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) > 0.95$ other than the above algebraic method of using InvNorm?

Ans:

Yes. We can also use GC "forming a table" method to solve (as shown on the right).

**Alternative Method**

$P(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) > 0.95$




From GC,

When $n = 384$,

$P(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) = 0.94996 < 0.95$

When $n = 385$,

$P(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) = 0.95025 > 0.95$

When $n = 386$,

$P(-0.1\sqrt{n} < Z < 0.1\sqrt{n}) = 0.95055 > 0.95$

$\therefore$ Least sample size is 385.

4. [N2003/II/27]
The random variable $X$ has the distribution $\mathrm{N}(1, 20)$. A random sample of $n$ observations of $X$ is taken. Given that the probability that the sample mean exceeds 1.5 is at most 0.01, find the set of possible values of $n$.

[Ans: $\{n \in \mathbb{Z}: n \geq 433\}$]

**Solution:**

$$\bar{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \sim \mathrm{N}\left(1, \frac{20}{n}\right)$$

$$\mathrm{P}\left(\bar{X} > 1.5\right) \leq 0.01$$

$$\mathrm{P}\left(Z > \frac{1.5 - 1}{\sqrt{\frac{20}{n}}}\right) \leq 0.01$$

$$\mathrm{P}\left(Z > \frac{0.5\sqrt{n}}{\sqrt{20}}\right) \leq 0.01$$

From GC,

$$\frac{0.5\sqrt{n}}{\sqrt{20}} \geq 2.3263$$

$$n \geq 432.934$$

Since $n$ is an integer value, the set of possible values of $n$ is $\{n \in \mathbb{Z}: n \geq 433\}$.

**Alternative Method**

$$\mathrm{P}\left(\bar{X} > 1.5\right) \leq 0.01$$

Using GC,

| $n$ | $\mathrm{P}\left(\bar{X} > 1.5\right)$ |
|---|---|
| 432 | $0.0101 < 0.01$ |
| 433 | $0.01 \leq 0.01$ |
| 434 | $0.0099 \leq 0.01$ |

$\therefore$ Since $n$ is an integer value, the set of possible values of $n$ is $\{n \in \mathbb{Z}: n \geq 433\}$.

## 5.3.2 Distributions of Sample Mean and Sample Sum from a Non-Normal population

In this case we have to use the **Central Limit Theorem** which states that:

> If $X_1, X_2, \ldots, X_n$ is a random sample of $n$ independent observations taken from a population of ***any distribution*** that is ***non-normal***, $X$ with mean $\mu$ and variance $\sigma^2$, then
>
> 1. $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ approximately if the sample size $n$ is large ($n \geq 30$)
> 2. $X_1 + X_2 + \ldots\ldots + X_n \sim \mathrm{N}(n\mu\ ,\ n\sigma^2)$ approximately if the sample size $n$ is large ($n \geq 30$)

Note:

1. Approximation improves as the sample size $n$ gets larger.
2. When the sampling is done from a normal distribution, we do not use Central Limit Theorem.

**Example 4**

10 fair dice are thrown and the number of sixes obtained is recorded. If the 10 dice are thrown 50 times, find the probability that the

(i) mean number of sixes obtained is more than 2.

(ii) total number of sixes obtained is less than 100.

**Solution:**

Let $X$ the number of sixes obtained out of 10 dice being thrown.

$$X \sim \mathrm{B}\left(10, \frac{1}{6}\right)$$

$$\mathrm{E}(X) = 10\left(\frac{1}{6}\right) = \frac{5}{3}$$

$$\mathrm{Var}(X) = 10\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \frac{25}{18}$$

(i) Let $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_{50}}{50}$.

Since $n = 50$ is large, by Central Limit Theorem,

$$\overline{X} \sim \mathrm{N}\left(\frac{5}{3}, \frac{\frac{25}{18}}{50}\right) \text{ approximately}$$

$$\overline{X} \sim \mathrm{N}\left(\frac{5}{3}, \frac{1}{36}\right) \text{ approximately}$$

$$\mathrm{P}\left(\overline{X} > 2\right) = 0.0228$$

(ii) Let $T = X_1 + X_2 + \cdots + X_{50}$.

Since $n = 50$ is large, by Central Limit Theorem,

$$T \sim \text{N}\left(50\left(\frac{5}{3}\right), 50\left(\frac{25}{18}\right)\right) \text{ approximately}$$

$$T \sim \text{N}\left(\frac{250}{3}, \frac{625}{9}\right) \text{ approximately}$$

$\text{P}(T < 100) = 0.977$

**Example 5**

If a random sample of size 50 is taken from $X \sim \text{B}(9, 0.5)$, find the probability that the

(i) sample mean exceeds 5

(ii) sum of the sample is at most 230

Estimate the sample size, $n$ that should be taken if 5% of the sample means are less than 4.25.

**Solution:**

(i) $X \sim \text{B}(9, 0.5)$

$\text{E}(X) = 9(0.5) = 4.5$ and $\text{Var}(X) = 9(0.5)(1-0.5) = 2.25$

Since $n = 50$ is large, by Central Limit Theorem,

$$\bar{X} \sim \text{N}(4.5, \frac{2.25}{50}) \text{ approximately.}$$

$\text{P}(\bar{X} > 5) = 0.00921$

(ii) Since $n = 50$ is large, by Central Limit Theorem,

$X_1 + X_2 + ... + X_{50} \sim \text{N}(50(4.5), 50(2.25))$

i.e. $X_1 + X_2 + ... + X_{50} \sim \text{N}(225, 112.5)$

$\text{P}(X_1 + X_2 + ... + X_{50} \leq 230)$

$= 0.681$

Assuming $n$ is large, by Central Limit Theorem,

$$\bar{X} \sim \text{N}(4.5, \frac{2.25}{n}) \text{ approximately.}$$

$$P(\bar{X} < 4.25) = 0.05$$

$$P\left(Z < \frac{4.25 - 4.5}{\sqrt{\frac{2.25}{n}}}\right) = 0.05$$

$$P\left(Z < \frac{-0.25\sqrt{n}}{\sqrt{2.25}}\right) = 0.05$$

From GC,

$$\frac{-0.25\sqrt{n}}{\sqrt{2.25}} = -1.6449$$

$$n = \left(\frac{(-1.6449)\left(\sqrt{2.25}\right)}{-0.25}\right)^2 = 97.4$$

$\therefore$ $n$ = 97 (nearest integer)

**Alternative Method**

$P(\bar{X} < 4.25) = 0.05$

Using GC,

| $n$ | $P(\bar{X} < 4.25)$ |
|---|---|
| 96 | 0.0512 |
| 97 | 0.0503 |
| 98 | 0.0495 |

$\therefore$ Nearest integer $n$ is 97.

**Example 6**

The rate of consumption of petrol of an "SA" car is known to have a mean of 13 km per litre and standard deviation of 2 km per litre. A random sample of 50 "SA" cars is taken.

(i) Is this sampling distribution of the mean normally distributed? Explain briefly. State the mean and variance of this sampling distribution.

(ii) Find the probability that the mean consumption rate is more than 13.8.

(iii) Find the probability that the total consumption of petrol is less than 675.

**Solution:**

(i) Let $X$ be the rate of consumption of petrol of an "SA" car.
Yes. Since $n$ = 50 is large, by Central Limit Theorem,
$\bar{X}$ is normally distributed approximately.

$$\mathrm{E}\left(\bar{X}\right)=13$$

$$\mathrm{Var}\left(\bar{X}\right)=\frac{2^2}{50}=0.08$$

(ii) Since $\bar{X} \sim \mathrm{N}(13, 0.08)$ approximately,

$\mathrm{P}(\bar{X} > 13.8) = 0.00234$

(iii) Since $n$ = 50 is large, by Central Limit Theorem,

$X_1 + X_2 + ... + X_{50} \sim \mathrm{N}(650, 200)$ approximately

$\mathrm{P}(X_1 + X_2 + ... + X_{50} < 675) = 0.961$

## Exercise 3

1. A large number of random samples of size $n$ are taken from B(20, 0.2).
   (i) 90% of the sample means are less than 4.354. Estimate $n$.
   (ii) 90% of the sum of the sample are less than 200. Estimate $n$.

[Ans: (i) 42 (ii) 46]

**Solution:**

$X \sim \mathrm{B}(20, 0.2)$

$\mathrm{E}(X) = 20(0.2) = 4$

$\mathrm{Var}(X) = 20(0.2)(0.8) = 3.2$

(i) Since $n$ is large, by Central Limit Theorem,

$\bar{X} \sim \mathrm{N}(4, \frac{3.2}{n})$ approximately.

$\mathrm{P}(\bar{X} < 4.354) = 0.9$

$$\mathrm{P}\left(Z < \frac{4.354-4}{\sqrt{\frac{3.2}{n}}}\right) = 0.9$$

From GC, $\dfrac{4.354-4}{\sqrt{\frac{3.2}{n}}} = 1.2816$

$$\sqrt{\frac{3.2}{n}} = 0.27622$$

$$n = 41.94$$

$$n \approx 42$$

(ii) Since $n$ is large, by Central Limit Theorem,
$X_1 + \cdots + X_n \sim \mathrm{N}(4n, 3.2n)$ approximately.

$\mathrm{P}(X_1 + \cdots X_n < 200) = 0.9$

$$\mathrm{P}\left(Z < \frac{200 - 4n}{\sqrt{3.2n}}\right) = 0.9$$

From GC, $\dfrac{200 - 4n}{\sqrt{3.2n}} = 1.2816$

From GC, $n = 46.108$

$n \approx 46$

**Alternative Method**

$\mathrm{P}\left(\bar{X} < 4.354\right) = 0.9$

Using GC,

| $n$ | $\mathrm{P}\left(\bar{X} < 4.354\right)$ |
|---|---|
| 41 | 0.8974 |
| 42 | 0.9002 |
| 43 | 0.9028 |

$\therefore$ Nearest integer $n$ is 42.

2. [ACJC/2014/Prelim]
In an island resort, there are trams that go around the island which stopped over at many stations. The probability that a passenger on board will alight at Aquafront Station is 0.52. Assume that each passenger alights at a destination independently of one another. On a particular weekday, there are 60 fully occupied trams, each with a seating capacity of 20 passengers, stopping over at Aquafront Station. Find the probability that the mean number of passengers alighting from a tram exceeds 11.

[Ans: 0.0188]

**Solution:**

Let $X$ be the number of passengers alighting at Aquafront Station, out of 20.

$$X \sim \mathrm{B}(20,\ 0.52)$$

$$\mathrm{E}(X) = (20)(0.52) = 10.4$$

$$\mathrm{Var}(X) = (20)(0.52)(0.48) = 4.992$$

Since 60 is large, by Central Limit Theorem,

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_{60}}{60} \sim N\left(10.4,\ \frac{4.992}{60}\right) \text{ approximately}$$

i.e. $\bar{X} \sim N(10.4, 0.0832)$ approximately

$$\mathrm{P}(\bar{X} > 11) = 0.0188$$

3. Two firms, *A* and *B*, manufacture similar components with a mean breaking strength of 6 kN and 5.5 kN and standard deviations of 0.4 kN and 0.2 kN respectively. If random samples of 100 components from firm *A* and of 50 from *B* are tested, find the probability that the mean breaking strength of the components from firm *A* will be between 0.45 kN and 0.55 kN more than the mean of those from firm *B*.

[Ans: 0.693]

**Solution:**

Let $A$ be the breaking strength (in kN) of components from firm *A*.
Let $B$ be the breaking strength (in kN) of components from firm *B*.

Since $n$ = 100 is large, by Central Limit Theorem,

$$\bar{A} \sim \mathrm{N}\left(6, \frac{0.4^2}{100}\right) \text{approximately}$$

Since $n$ = 50 is large, by Central Limit Theorem,

$$\bar{B} \sim \mathrm{N}\left(5.5, \frac{0.2^2}{50}\right) \text{ approximately}$$

$$\bar{A} - \bar{B} \sim \mathrm{N}\left(6 - 5.5, \frac{0.4^2}{100} + \frac{0.2^2}{50}\right)$$

$$\bar{A} - \bar{B} \sim \mathrm{N}(0.5, 0.0024)$$

$$P(0.45 < \bar{A} - \bar{B} < 0.55) = 0.693$$

4. The content of a randomly chosen packet of a particular drink $A$ is found to have a mean of 200 ml and a standard deviation of 15 ml. A random sample of 50 packets of the drink is taken. Calculate the probability that the mean content of the sample exceeds 195 ml.

   Another type of drink $B$ is such that its mean is found to be 200 ml and its standard deviation is 20 ml. If a random sample of 50 packets of each type of drink were taken, what is the probability that the difference between the total amounts of content of the samples for each type is more than 100 ml?

   [Ans: 0.991, 0.444]

**Solution:**

Let $X$ be the content (in ml) of a packet of drink $A$.
Since $n = 50$ is large, by Central Limit Theorem,

$$\overline{X} \sim \text{N}(200, \frac{15^2}{50}) \text{ approximately}$$

$$\text{P}(\overline{X} > 195) = 0.991$$

Since $n = 50$ is large, by Central Limit Theorem,

$$S = X_1 + X_2 + ... + X_{50} \sim \text{N}(50(200), 50(15)^2) \text{ approximately}$$

Let $Y$ be the content (in ml) of a packet of drink $B$.
Since $n = 50$ is large, by Central Limit Theorem,

$$T = Y_1 + Y_2 + ... + Y_{50} \sim \text{N}(50(200), 50(20)^2) \text{ approximately}$$

$$\therefore S - T \sim \text{N}(0, 31250) \text{ approximately}$$

$$\begin{aligned} \text{P}(|S - T| > 100) &= 1 - \text{P}(|S - T| \leq 100) \\ &= 1 - \text{P}(-100 \leq S - T \leq 100) \\ &= 0.572 \end{aligned}$$

Question Prompt

When we calculate the difference between the total amounts of content of the samples for each type of drink, should we consider $S - T > 100$ or $T - S > 100$ ? Why?

Ans:

We should consider **both** $S - T > 100$ and $T - S > 100$ in our calculation. Because the question did not state whether $S$ is greater or $T$ is greater, hence to calculate the difference, we would need to consider both cases.

## 5.4 Estimation

Wouldn't it be great if you could tell what a population was like just by taking one sample? In this section, you will learn how to use samples to accurately predict what the population will be like (e.g. population mean) and come up with a way of saying how reliable your predictions are.

Suppose a population has an unknown parameter (such as the mean or variance), then an estimate of the unknown parameter can be made from information supplied by a random sample.

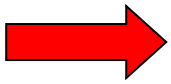
### 5.4.1 Some Definitions

1. An **estimator** is a sample statistic used to estimate the value of an unknown population parameter and it is denoted by capital letters e.g. $\overline{X}$

   Recall: A **sample statistic** is a quantity describing a characteristic of a random sample. Examples include the mean and variance of a sample, denoted by $\overline{X}$ and $S_x^2$ respectively

2. The numerical value of the estimator is called an **estimate** and is denoted by small letters, eg: $\bar{x}$

In a statistical enquiry, the value of a sample statistic is used to estimate the corresponding population parameter.

| **Sample** | | **Population** |
|---|---|---|
| Sample mean: $\bar{x}$<br>Sample standard deviation: $s_x$ |  | Population mean: $\mu$<br>Population standard deviation : $\sigma$ |

3. A statistic is said to be an **unbiased estimator** of a given parameter when **the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated.** For example, the mean of a sample is an unbiased estimate of the mean of the population from which the sample was drawn, i.e. $\mathrm{E}\left(\overline{X}\right) = \mu$.

## 5.4.2 Unbiased Estimates of Population Mean and Variance

In the above section, we saw that the mean of the sampling distribution is the same as the population mean since E( $\overline{X}$ ) = $\mu$ . Therefore we say that the sample mean, $\overline{X}$ is the **unbiased estimator** of the population mean, $\mu$. In other words,

Unbiased estimate for population mean, $\mu$ is $\bar{x} = \dfrac{\sum x}{n}$

OR $\bar{x} = \dfrac{\sum (x-a)}{n} + a$, where $a$ is a constant

However, the sample variance is ***NOT*** an unbiased estimator of population variance.

Let $X_1$, $X_2$, ……, $X_n$ be a random sample of size $n$ from a population with unknown variance $\sigma^2$ and let the sample variance be denoted by $s_x^2$ .

Sample variance, $s_x^2 = \dfrac{\sum (x-\bar{x})^2}{n} = \dfrac{\sum x^2}{n} - (\bar{x})^2$

Then an

Unbiased estimate for population variance, $\sigma^2$ is $s^2 = \dfrac{n}{n-1} s_x^2$

Note: Use the above formula when the sample variance, $s_x^2$ is given.

Unbiased estimate for population variance is

$$s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 = \frac{1}{n-1}\left[\sum (x^2) - \frac{(\sum x)^2}{n}\right]$$ **(given in MF26)**

Note:

1. Although the estimator $\overline{X}$ is an unbiased estimator for population mean, the individual estimate $\bar{x}$ is usually not equal to the population mean. It does not make sense that every sample chosen will yield the same mean value as the population mean!

2. The sample variance for a sample of $n$ measurements is the sum of the squared distances from the mean divided by $n$. This measures the variability or the spread of the sample.

3. The sample standard deviation for a sample is defined as the positive square root of the sample variance. i.e.

$$\text{Sample standard deviation, } s_x = \sqrt{\text{sample variance}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

4. For grouped data (i.e. data presented as a frequency table), we use

$$\bar{x} = \frac{\sum xf}{n} \quad \text{and} \quad s^2 = \frac{1}{n-1}\left(\sum x^2 f - \frac{\left(\sum xf\right)^2}{n}\right) \quad \text{or} \quad \frac{1}{n-1}\sum\left(x-\bar{x}\right)^2 f$$

where $n = \sum f$ .

**In summary:**

| | **Population Parameter** | **Sample Parameter** | **Unbiased Estimate for the population parameter if it is unknown** |
|---|---|---|---|
| **Mean** | $\mu$ | $\bar{x}$ | $\bar{x} = \frac{\sum x}{n}$<br>OR<br>$\bar{x} = \frac{\sum(x-a)}{n} + a$, where $a$ is a constant |
| **Variance** | $\sigma^2$ | $s_x^2$ | $s^2 = \frac{n}{n-1}s_x^2$<br>OR<br>$s^2 = \frac{1}{n-1}[\sum(x^2) - \frac{(\sum x)^2}{n}]$<br>OR<br>$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$ |

**Example 7**

Obtain the unbiased estimates of the population mean and population variance from which the following sample is drawn: $n = 12$, $\bar{x} = 23.5$, $\sum(x-\bar{x})^2 = 48.72$.

**Solution:**

Unbiased estimate for population mean is $\bar{x} = 23.5$

Unbiased estimate for population variance is $s^2 = \frac{1}{n-1}[\sum(x-\bar{x})^2] = \frac{1}{11}(48.72) = 4.43$

**Example 8**

A random sample drawn from a large population contains 20 observations, $x_1, x_2, \ldots, x_{20}$ such that $\Sigma x = 22.8$ and $\Sigma x^2 = 27.55$.
Obtain unbiased estimates for the population mean and population variance.

**Solution:**

$$\bar{x} = \frac{\sum x}{n} = \frac{22.8}{20} = 1.14$$

Unbiased estimate for population mean is $\bar{x} = 1.14$

Unbiased estimate for population variance is $s^2$

$$= \frac{1}{n-1}[\sum (x^2) - \frac{(\sum x)^2}{n}]$$
$$= \frac{1}{19}[27.55 - \frac{(22.8)^2}{20}]$$
$$= 0.082$$

**Example 9**

A random sample of 10 bulbs is taken and their lifetimes, $x_i$ are obtained. The results are summarized by $\Sigma(x - 1000) = 1890$; $\Sigma(x - 1000)^2 = 362050.2$. Find unbiased estimates of the mean and standard deviation of $X$.

**Solution:**

$$\bar{x} = \frac{\sum (x-1000)}{n} + 1000$$
$$= \frac{1890}{10} + 1000$$
$$= 1189$$

Unbiased estimate for population mean is $\bar{x} = 1189$

Unbiased estimate for population variance is $s^2$

$$= \frac{1}{n-1}[\sum (x-1000)^2 - \frac{(\sum (x-1000))^2}{n}]$$
$$= \frac{1}{9}[362050.2 - \frac{(1890)^2}{10}]$$
$$= 537.8$$

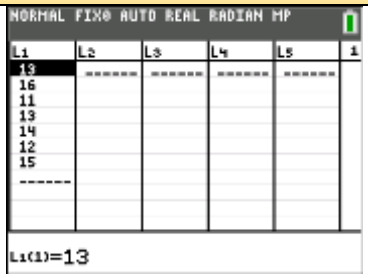
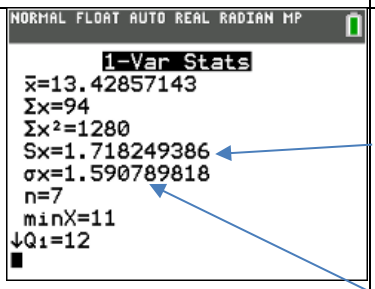
Unbiased estimate for population standard deviation is $s$

$$= 23.2$$

**Example 10**

Consider a group of 7 children who sat for a test and their scores were 13, 16, 11, 13, 14, 12, 15 respectively. Using a graphic calculator, calculate the sample mean, sample standard deviation and unbiased estimate of population standard deviation.

**Solution:**

| Steps | Screenshot | Remarks |
|---|---|---|
| Press **stat**<br><br>choose **1:Edit**<br><br>key in the scores under $\mathbf{L_1}$ |  | |
| Press **stat**<br><br>choose **CALC**<br><br>followed by **1:1-Var Stats,**<br><br>**Enter** |  | Note that **FreqList** is left empty as the frequency for all data is 1. |
| |  | Note the different notations here…<br><br>Unbiased estimate for population std. deviation, previously denoted by $s$<br><br>Sample std. deviation, previously denoted by $s_x$ |

From GC,

Sample mean is 13.4,

Sample standard deviation is 1.59,

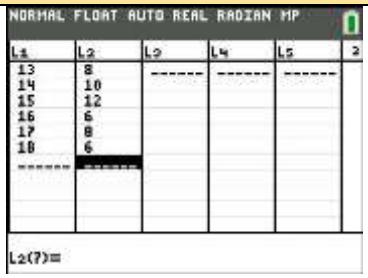
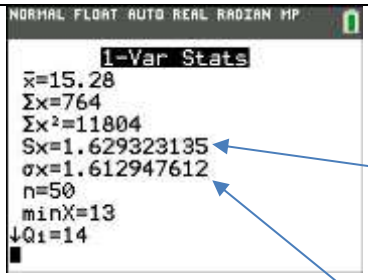
Unbiased estimate of population standard deviation is 1.72

**Example 11**

Consider a sample of 50 children who sat for a test and their scores are tabulated below. Using a GC, calculate the sample mean, sample standard deviation and unbiased estimate of population standard deviation.

| Scores | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|
| Frequency | 8 | 10 | 12 | 6 | 8 | 6 |

**Solution:**

| Steps | Screenshot | Remarks |
|---|---|---|
| Press **stat**<br>choose **1:Edit**<br>key in the scores under $L_1$ and the frequency under $L_2$ |  | |
| Press **stat**<br>choose **CALC**<br>followed by **1:1-Var Stats,**<br>**Enter** |  | Note that **FreqList** can be left empty if the frequency for all data is 1. |
| |  | Note the different notations here…<br>Unbiased estimate for population std. deviation, previously denoted by $s$<br>Sample std. deviation, previously denoted by $s_x$ |

From GC,
Sample mean is 15.28,
Sample standard deviation is 1.61,
Unbiased estimate of population standard deviation is 1.63

**Example 12**

The amount of donation collected by each student in a particular college is $\$x$. Data is collected from a random sample of 50 students and the results are summarized by

$$\sum x = 4000 \text{ and } \sum x^2 = 360250.$$

(i) Find unbiased estimates of the population mean and variance of the amount of donation collected by the students.

(ii) State clearly the distribution of the mean amount collected by the students, giving its mean and variance.

(iii) Find the probability that the mean amount collected exceeds \$85.

(iv) Find the smallest possible sample size, $n$ if the probability that the mean amount exceeds \$85 is less than 0.03. (Assume $n$ is large)

**Solution:**

(i) Unbiased estimate of the population mean, $\bar{x} = \dfrac{4000}{50} = 80$

Unbiased estimate of the population variance,

$$s^2 = \frac{1}{49}\left(360250 - \frac{4000^2}{50}\right) \approx 821.43 = 821 \text{ (3 s.f.)}$$

(ii) Since $n$ = 50 is large, by Central Limit Theorem,

$$\bar{X} \sim \text{N}\left(80, \tfrac{821.43}{50}\right) \text{ approximately.}$$
$$\bar{X} \sim \text{N}(80, 16.427) \text{ approximately.}$$

(iii) $\text{P}(\bar{X} > 85) = 0.109$

(iv) Let $n$ be the sample size.

Since $n$ is large, by Central Limit Theorem,

$$\bar{X} \sim \text{N}\left(80, \frac{821.43}{n}\right) \text{ approximately}$$

$$\text{P}(\bar{X} > 85) < 0.03$$

$$\text{P}\left(Z > \frac{85-80}{\sqrt{\frac{821.43}{n}}}\right) < 0.03$$

From GC,

$$\frac{5}{\sqrt{\frac{821.43}{n}}} > 1.8808$$

$$5 > 1.8808\frac{\sqrt{821.43}}{\sqrt{n}}$$

$$\sqrt{n} > \frac{1.8808\sqrt{821.43}}{5} = 10.781$$

$$n > 116.2$$

$\therefore$ Smallest sample size is 117.

**Alternative Method**

$\text{P}(\bar{X} > 85) < 0.03$

From GC,

| $n$ | $\text{P}(\bar{X} > 85)$ |
|---|---|
| 116 | 0.0301 > 0.03 |
| 117 | 0.0296 < 0.03 |
| 118 | 0.029 < 0.03 |

$\therefore$ Smallest $n$ is 117.

## Exercise 4

1. In each of the following cases, find unbiased estimates of the population mean and population variance of $X$. Give your answers to 2 decimal places.

(a) Sample size = 80, $\sum x = 13560$, $\sum x^2 = 2388670$

(b) Sample size = 15, $\sum x = 45$, $\sum (x-3)^2 = 90$

(c) Sample Size = 90, $\sum (x-120) = -29$, $\sum (x-120)^2 = 419$

(d) Sample Size = 30, $\sum x = 136.45$, Sample variance, $s_x^2 = 20.4$

(e)

| $x$ | 10 | 11.5 | 12 | 12.5 | 13 |
|---|---|---|---|---|---|
| **Frequency** | 13 | 14 | 13 | 7 | 3 |

[Ans: (a) 169.5; 1142.41 (b) 3; 6.43 (c) 119.68; 4.60 (d) 4.55; 21.10 (e) 11.47; 0.94]

**Solution:**

(a) Unbiased estimate of the population mean of $X$, $\bar{x} = \frac{\sum x}{n} = \frac{13560}{80} = 169.5$

Unbiased estimate of the population variance of $X$,

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = \frac{1}{79}\left[2388670 - \frac{(13560)^2}{80}\right] = 1142.41$$

(b) Unbiased estimate of the population mean of $X$, $\bar{x} = \frac{\sum x}{n} = \frac{45}{15} = 3$

Unbiased estimate of the population variance of $X$,

$$s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 = \frac{1}{14}(90) = 6.43$$

(c) Unbiased estimate of the population mean of $X$,

$$\bar{x} = \frac{\sum (x-120)}{n} + 120 = \frac{-29}{90} + 120 = 119.68$$

Unbiased estimate of the population variance of $X$,

$$s^2 = \frac{1}{n-1}\left[\sum (x-120)^2 - \frac{\left(\sum (x-120)\right)^2}{n}\right] = \frac{1}{89}\left[419 - \frac{(-29)^2}{90}\right] = 4.60$$

(d) Unbiased estimate of the population mean of $X$, $\bar{x} = \frac{136.45}{30} = 4.55$

Unbiased estimate of the population variance of $X$, $s^2 = \frac{30}{29}(20.4) = 21.10$

(e) From GC,

Unbiased estimate of the population mean of $X$, $\bar{x} = 11.47$

Unbiased estimate of the population variance of $X$, $s^2 = 0.97106^2 = 0.94$

2. A fruit importer claims that the durians he imports has an average mass of 1.5 kg each. A random sample of 8 durians is examined and the mass of each durian is recorded as 1.4 kg, 1.5 kg, 1.6 kg, 1.4 kg, 1.6 kg, 1.5 kg, 1.8 kg and 1.7 kg. Find unbiased estimates of the population mean and variance.

[Ans: 1.5625; 0.0198]

**Solution:**

Unbiased estimate of the population mean, $\bar{x} = 1.5625$

Unbiased estimate of the population variance, $s^2 = 0.14079^2 = 0.0198$

3. The following table illustrates the number of cars sold in a day during a period of 50 days.

| No. of cars sold | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| No. of days | 18 | 8 | 8 | 5 | 7 | 4 |

(i) Find unbiased estimates of the mean and variance of the number of cars sold.
(ii) Find the probability that the mean number of cars sold per day is greater than 3 over a period of 70 days.

[Ans: (i) 2.74; 2.97 (ii) 0.103]

**Solution:**

(i) Unbiased estimate of the population mean, $\bar{x} = 2.74$

Unbiased estimate of the population variance,

$s^2 = 1.723901602^2 = 2.9718 (5 \text{ s.f.}) = 2.97 (3 \text{ s.f.})$

(ii) Since $n = 70$ is large, by Central Limit Theorem,

$\bar{X} \sim \text{N}(2.74, \frac{2.9718}{70})$ approximately

$\text{P}(\bar{X} > 3) = 0.103$

## Practice Questions

1. [ACJC/2007/Prelim/P2/Q10]
In the 2006 graduation ceremony of a particular university, a survey was done. From the survey of the 50 graduates, $x$ represents the number of hours the graduates spent at the central library. The summarised data are as follows:

$$\sum(x-30)=64 \qquad \sum(x-30)^2=786$$

(i) Calculate the unbiased estimates of the population mean and variance.
(ii) Estimate the probability that a random sample of 50 graduates spent an average of at least 32 hours at the central library.

[Ans: (i) 31.28, 14.4 (ii) 0.0896]

**Solution:**

(i) Unbiased estimate for population mean: $\bar{x}=\frac{64}{50}+30=31.28$

Unbiased estimate for the population variance

$$=\frac{1}{n-1}\left[\sum(x-30)^2-\frac{\left(\sum(x-30)\right)^2}{n}\right]=\frac{1}{49}\left[786-\frac{64^2}{50}\right]=14.369\approx 14.4$$

(ii) Since $n=50$ is large, by Central Limit Theorem,

$\overline{X}$ ~ N(31.28, $\frac{14.369}{50}$) approximately

$P(\overline{X}\geq 32)=0.0896$

2. At a certain supermarket, spinach and kangkong are sold in bundles. Bundles of spinach have weights which are normally distributed with a mean of 150 g and a standard deviation of 10 g, and bundles of kangkong have weights which are normally distributed with a mean of 180 g and a standard deviation of 15 g. The weights of the bundles of spinach are independent of the weights of the bundles of kangkong.

Find the probability that
(i) the average weight of 5 randomly chosen bundles of kangkong is at least 20 g more than the average weight of 10 randomly chosen bundles of spinach.
(ii) the average weight of 3 randomly chosen bundles of spinach and 2 randomly chosen bundles of kangkong is less than 160 g.

[Ans: (i) 0.911 (ii) 0.358]

**Solution:**

Let $X$ g and $Y$ g be the weights of a bundle of spinach and kangkong respectively.

$X$ ~ N(150, $10^2$), $Y$~N(180, $15^2$)

(i) Let the average weight of 10 bundles of spinach and 5 bundles of kangkong be $\overline{X}$ g and $\overline{Y}$ g respectively.

$$\overline{X} = \frac{X_1 + X_2 + ..... + X_{10}}{10} \sim \text{N}\left(150, \frac{10^2}{10}\right).$$

$$\overline{Y} = \frac{Y_1 + Y_2 + ..... + Y_5}{5} \sim \text{N}\left(180, \frac{15^2}{5}\right)$$

$$\overline{Y} - \overline{X} \sim \text{N}\left(180 - 150, \frac{15^2}{5} + \frac{10^2}{10}\right)$$

i.e., $\overline{Y} - \overline{X} \sim \text{N}(30, 55)$

Required probability = $\text{P}(\overline{Y} - \overline{X} \geq 20)$

= 0.911 (3 s.f.)

(ii) Let $\overline{T} = \dfrac{X_1 + X_2 + X_3 + Y_1 + Y_2}{5}$

$\overline{\text{T}} \sim N(\frac{810}{5}, 30)$

$\text{P}(\overline{T} < 160) = 0.358$

3. [9233/N2008/II/Q30 EITHER]

The masses of valves produced by a machine are normally distributed with mean $\mu$ and standard deviation $\sigma$. 12% of the valves have mass less than 86.50 g and 20% have mass more than 92.25 g. Find $\mu$ and $\sigma$.

The setting of the machine is adjusted so that the mean mass of the valves produced is unchanged, but the standard deviation is reduced. Given that 80% of the valves now have a mass within 2 g of the mean, find the new standard deviation.

After the machine has been adjusted, a random sample of *n* valves is taken. Find the smallest value of *n* such that the probability that the sample mean exceeds $\mu$ by at least 0.50 g is at most 0.1.

[Ans: 89.9, 2.85; 1.56; 16]

**Solution:**

Let $X$ grams be the masses of valves produced by the machine.

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

$\mathrm{P}(X < 86.5) = 0.12$

$\mathrm{P}\left(Z < \dfrac{86.5-\mu}{\sigma}\right) = 0.12$

Using GC, $\dfrac{86.5-\mu}{\sigma} = -1.174987$

$\mu - 1.174987\sigma = 86.5 \quad ------- (1)$

$\mathrm{P}(X > 92.25) = 0.2$

$\mathrm{P}\left(Z > \dfrac{92.25-\mu}{\sigma}\right) = 0.2$

Using GC, $\dfrac{92.25-\mu}{\sigma} = 0.841621$

$\mu + 0.841621\sigma = 92.25 \quad ------- (2)$

Solving $(1)\,\&\,(2)$,

$\mu \approx 89.850 = 89.9$ (3 s.f.)

$\sigma \approx 2.8513 = 2.85$ (3 s.f)

After the machine has been adjusted, the mean mass $\mu$ is unchanged but the standard deviation is reduced.

Let $a$ be the new standard deviation.

$\therefore \; X \sim \mathrm{N}(\mu, a^2)$

Given $\mathrm{P}(-2 < X - \mu < 2) = 0.8.$

$\mathrm{P}\left(\dfrac{-2}{a} < \dfrac{X-\mu}{a} < \dfrac{2}{a}\right) = 0.8$

$\mathrm{P}\left(\dfrac{-2}{a} < Z < \dfrac{2}{a}\right) = 0.8$

Using GC, $\dfrac{2}{a} = 1.28155$

$a \approx 1.5606 = 1.56$ (3 s.f.)

The new standard deviation is 1.56. (3 s.f.)

$X \sim \mathrm{N}(\mu, 1.5606^2)$

$\overline{X} \sim \mathrm{N}\left(\mu, \dfrac{1.5606^2}{n}\right)$

Given $\mathrm{P}(\overline{X} - \mu \geq 0.5) \leq 0.1.$

$$\mathrm{P}\left(\frac{\overline{X}-\mu}{\sqrt{\frac{1.5606^2}{n}}} > \frac{0.5}{\sqrt{\frac{1.5606^2}{n}}}\right) \le 0.1$$

$$\mathrm{P}\left(Z > \frac{0.5\sqrt{n}}{1.5606}\right) \le 0.1$$

$$\frac{0.5\sqrt{n}}{1.5606} \ge 1.28155$$

$$n \ge 15.999$$

Hence the least value of $n$ is 16.

4. [9233 Specimen Paper/II/Q29 EITHER]

The mass of a certain kind of casting produced in an iron foundry has a normal distribution. Each of 150 castings in a random sample was weighed and the value of $x$, its mass in kg, was recorded. The results are summarised by

$$\sum x = 3712.75 \text{ and } \sum x^2 = 92515.4, \text{ correct to 6 significant figures.}$$

(i) Calculate the sample mean mass and an unbiased estimate of the population variance, each correct to 2 decimal places.

(ii) Assuming that the population mean mass is 25.00 kg, estimate the probability that a random sample of 150 castings will have a sample mean mass less than that calculated above. Give your answer correct to 3 decimal places.

(iii) Explain briefly why the probability calculation in part (ii) is only an estimate, and whether the calculation would still be valid if the masses of the castings had not been known to have a normal distribution.

[Ans: (i) 24.75, 4.15 (ii) 0.066]

**Solution:**

(i) Sample mean mass, $\overline{x} = \frac{\sum x}{n} = \frac{3712.75}{150} = 24.752 = 24.75 \text{ (to 2 d.p.)}$

Unbiased estimate of the population variance $\sigma^2$ is

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

$$= \frac{1}{149}\left[92515.4 - \frac{(3712.75)^2}{150}\right]$$

$$= 4.1520$$

$$= 4.15 \text{ (to 2 d.p.)}$$

(ii) $\overline{X} \sim \mathrm{N}\left(25.00, \frac{4.1520}{150}\right)$

$P\left(\overline{X} < 24.75\right) = 0.066$ (3 d.p.)

(iii) The calculation in (ii) is an estimate because since the population variance is unknown, we used the unbiased estimate of the population variance in our calculation.

If the mass of the castings is not known to have a normal distribution, the calculation will still be valid because as $n$ = 150 is large, Central Limit Theorem can be applied.

5. [2008/TJC/J2 Mid-Year/Q10(ii)(iii)]

A survey on the amount of pocket money each student receives in a month is to be conducted by filling in questionnaires. From previous surveys done, it is known that the mean amount of pocket money received in a month is \$200 and the standard deviation \$80.

(i) A sample size of 1000 students is selected randomly from schools all over Singapore. Find the probability that the mean pocket money received in a month is between \$195 and \$225.

Give a reason why it is not necessary to assume that the amount of pocket money received in a month is normally distributed in order to calculate the probability.

(ii) A sample size of $n$ students is selected randomly from schools all over Singapore. Assuming that $n$ is large, find the least value of $n$ such that the probability of these students receiving a total amount of pocket money of more than \$16800 in a month is more than 0.95.

[Ans: (i) 0.976 (ii) 91]

**Solution:**

(i) Let $X$ be the amount of pocket money received by a student in a month.

Since the sample size $n = 1000$ is large, by Central Limit Theorem,

$\overline{X} \sim \text{N}\left(200, \frac{80^2}{1000}\right)$ approximately.

Hence, $\text{P}\left(195 < \overline{X} < 225\right) = 0.976$

Since $n = 1000$ is large, we can approximate the mean amount of pocket money received in a month to follow normal distribution by Central Limit Theorem. As such, it is not necessary to assume the amount of pocket money received is normally distributed.

(ii) Let $T = X_1 + X_2 + ... + X_n$

Since sample size $n$ is large, by Central Limit Theorem,

$T \sim \text{N}\left(200n, 80^2 n\right)$ approximately

$\text{P}(T > 16800) > 0.95$

**Alternative Method**

$\text{P}(T > 16800) > 0.95$

Using GC,

| $n$ | $\text{P}(T > 16800)$ |
|---|---|
| 90 | 0.9431 < 0.95 |
| 91 | 0.9667 > 0.95 |
| 92 | 0.9815 > 0.95 |

$$\mathrm{P}\left(Z > \frac{16800 - 200n}{80\sqrt{n}}\right) > 0.95$$

$$\frac{16800 - 200n}{80\sqrt{n}} < -1.645.$$

$$16800 - 200n < -131.6\sqrt{n}$$

$$200n - 131.6\sqrt{n} - 16800 > 0$$

Using GC, $\sqrt{n} > 9.5$ or $\sqrt{n} < -8.84$ (rejected $\because \sqrt{n} > 0$)

$\Rightarrow n > 90.25$

$\therefore$ Least $n$ is 91.

6. In a glass factory, it is found that 20% of glass panels produced by machine *A* are more than 3 mm thick. Given that the thickness of glass panels produced by machine *A* follows a normal distribution with mean 2.56 mm and standard deviation $\sigma$ mm, find $\sigma$.

(i) For the manufacture of a certain type of windscreen, two of the glass panels produced by machine *A* are used to form a double panel. Find the probability that the thickness of a double panel is between 4 mm and 6 mm.

(ii) Five glass panels produced by machine *A* are chosen at random and tested for their thickness. Find the probability that the mean thickness is greater than 3 mm.

(iii) The thickness of glass panels produced by machine *B* for a certain type of shower screen has a normal distribution with mean 5.9 mm and standard deviation 0.35 mm. Find the probability that the average thickness of 2 glass panels produced by machine *A* and 3 glass panels produced by machine *B* is at least 4.2 mm.

[Ans: 0.523 (i) 0.818 (ii) 0.0299 (iii) 0.972]

**Solution:**

Let *X* mm be the thickness of glass panels produced by machine *A*.

$$X \sim \mathrm{N}(2.56, \sigma^2)$$

$$\mathrm{P}(X > 3) = 0.2$$

$$\mathrm{P}\left(Z > \frac{3 - 2.56}{\sigma}\right) = 0.2$$

From GC, $\frac{0.44}{\sigma} = 0.84162$

$\sigma = 0.52280 = 0.523$ (to 3 s.f.)

(i) $X \sim \mathrm{N}(2.56, 0.27332)$

$$\mathrm{E}(X_1 + X_2) = 2\mathrm{E}(X) = 5.12$$
$$\mathrm{Var}(X_1 + X_2) = 2\mathrm{Var}(X) = 0.54664$$
$$X_1 + X_2 \sim \mathrm{N}(5.12, 0.54664)$$
$$\mathrm{P}(4 < X_1 + X_2 < 6) = 0.81812 = 0.818 \text{ (to 3 s.f.)}$$

(ii) Let $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_5}{5}$

$$\overline{X} \sim \mathrm{N}\left(2.56, \frac{0.27332}{5}\right)$$
$$\mathrm{P}(\overline{X} > 3) = 0.029923 = 0.0299 \text{ (to 3 s.f.)}$$

(iii) Let $Y$ mm be the thickness of glass panels produced by machine $B$.

$$Y \sim \mathrm{N}(5.9, 0.35^2)$$

Let $\overline{T} = \dfrac{X_1 + X_2 + Y_1 + Y_2 + Y_3}{5}$.

$$\mathrm{E}(\overline{T}) = \frac{1}{5}\left[2\mathrm{E}(X) + 3\mathrm{E}(Y)\right] = 4.564$$

$$\mathrm{Var}(\overline{T}) = \frac{1}{5^2}\left[2\mathrm{Var}(X) + 3\mathrm{Var}(Y)\right] = 0.0365656$$

$$\overline{T} \sim \mathrm{N}(4.564, 0.0365656)$$

$$\mathrm{P}(\overline{T} \geq 4.2) = 0.972 \text{ (to 3 s.f.)}$$

## Summary

**Notation:**

| | Population | Sample | Unbiased Estimate |
|---|---|---|---|
| **Mean** | $\mu$ | $\bar{x}$ | $\bar{x}$ |
| **Variance** | $\sigma^2$ | $s_x{}^2$ | $s^2$ |

Note :

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

$\mathrm{Var}(\bar{X}) = \dfrac{\mathrm{Var}(X)}{n} = \dfrac{\sigma^2}{n}$

**Sampling distribution:**

| **Sample is taken from a population with distribution** | |
|---|---|
| Normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ | If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathrm{N}(\mu, \dfrac{\sigma^2}{n})$<br>If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\sum X \sim \mathrm{N}(n\mu,\ n\sigma^2)$ |
| Any other distribution e.g. Binomial | By Central Limit Theorem, $\bar{X} \sim \mathrm{N}(\mu, \dfrac{\sigma^2}{n})$ approximately if $n$ is large.<br>By Central Limit Theorem, $\sum X \sim \mathrm{N}(n\mu,\ n\sigma^2)$ approximately if $n$ is large. |

**Unbiased estimates of population mean and variance**

For populations with unknown mean and variance, we may use the following to estimate the mean and variance:

- An ***unbiased estimate for population mean*** is $\bar{x} = \dfrac{\sum x}{n}$
- An ***unbiased estimate for population variance*** is $s^2 = \dfrac{1}{n-1}\left(\sum x^2 - \dfrac{(\sum x)^2}{n}\right) = \dfrac{1}{n-1}\sum (x - \bar{x})^2$

Note that $s^2 = \dfrac{n}{n-1} \times$ sample variance (use this formula when **sample variance** is given)

## Checklist

**I am able to:**

- ☐ understand that the sample mean can be regarded as a random variable, and use the fact that $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$ ;
- ☐ apply the sampling distribution $\mathrm{N}(\mu, \frac{\sigma^2}{n})$ to solve statistical problems in real – world situations;
- ☐ use the fact that $\overline{X}$ and $\sum X$ are normally distributed if $X$ is normally distributed;
- ☐ use Central Limit Theorem to approximate $\overline{X}$ and $\sum X$ as normal distributions when $X$ is not normally distributed and the sample size is sufficiently large;
- ☐ calculate the unbiased estimates of the population mean and population variance.