Raffles Institution
H2 Mathematics (9758)
2025 Year 6

Chapter S4 :Sampling

SYLLABUS INCLUDES

- Concepts of population and simple random sample
- Concept of the sample mean $\overline{X}$ as a random variable with $\mathrm{E}\left(\overline{X}\right)=\mu$ and $\mathrm{Var}\left(\overline{X}\right)=\dfrac{\sigma^2}{n}$.
- Distribution of sample means from a normal population
- Use of the Central Limit Theorem to treat sample mean as having normal distribution when the sample size is sufficiently large (e.g. $n \geq 30$)
- Use of unbiased estimates of the population mean and variance from a sample, including cases where the data are given in summarised form $\sum x$ and $\sum x^2$, or $\sum(x-a)$ and $\sum(x-a)^2$

PRE-REQUISITES

- Normal distribution

CONTENT

INTRODUCTION

Many real-world problems require the collection of data and studying/analyzing them. The objective of such statistical investigations is *inference* – that is, making decisions, estimations, generalizations or predictions about an unknown aspect of a population based on information obtained from a sample.

1 TERMINOLOGY

1.1 Population and Sample

A **population** is the entire collection of data (persons or items or individuals) that we want to study. In studying a population, we may focus on one or more characteristics (of the persons or items or individuals) in the population.

For example, we may be interested in the YouTube channels teenagers in Singapore subscribed to, so the population refers to all teenagers in Singapore while the characteristic refers to the various youtube channels.

If the population we wish to study is small in size, then it may be feasible to measure every unit in the population (this process is called a *census*). Very often, it is impossible or impractical to study an entire population by examining every unit in the population. This could be due to reasons such as:
(a) size of population is too big or infinite,
(b) population is dynamic and intractable,
(c) time and funding constraints.

Besides, if we are not too concerned with being exact, then it may not be worthwhile investing time and effort to the study of every data in the population. A more reasonable alternative would be to select and study a subset or a **sample** of the units in the population.

As in the previous example, a possible sample can be formed by selecting all the Year 6 students from Raffles Institution to obtain information about the YouTube channels they subscribed to.

In order to derive meaningful and reliable information about the population, it is necessary to select a sample from the population carefully. More importantly, it should be a **representative sample** of the population. For example, if we are studying heights of teenagers in Singapore, our sample should "resemble" the population and contain members with a good spread of heights. The method of selecting the sample is called the sampling method or sampling plan.

1.2 Random Samples

A sample is a random sample if it is chosen in such a way that:
1. every element in the population has an equal chance of being selected, and
2. the selections are independent of each other

A sample that is not random is known as a non-random sample.

Example 1
In a school there are 30 classes, each with 30 students, and 20 classes, each with 20 students. To find out students' perception of the food sold in the canteen, the school takes a random sample of 100 students. Explain what is meant in this context by the term 'a random sample'.

State, with a reason, whether each of the following methods would give a random sample.

Method 1: 2 students are randomly chosen from each of the 50 classes.

Method 2: Assign every student in the school a number, from 1 to 1300, by arranging their names in alphabetical order. Use a computer to generate 100 random numbers. The 100 students with the corresponding numbers would be the ones chosen to be in the sample.

Method 3: First 100 students who patronize the noodle stall are selected as the sample.

Solution

A random sample is a sample chosen in such a way that every student in the school has an equal chance of being selected, and the selection of students are independent of each other.

Method 1:
P(a particular student in a class of 20 being selected for the sample)

$$=\frac{1}{20}\cdot\frac{19}{19}\cdot 2=\frac{2}{20} \quad \text{or} \quad \frac{{}^{1}C_1\,{}^{19}C_1}{{}^{20}C_2}=\frac{1}{10}$$

P(a particular student in a class of 30 being selected for the sample)

$$=\frac{1}{30}\cdot\frac{29}{29}\cdot 2=\frac{2}{30} \quad \text{or} \quad \frac{{}^{1}C_1\,{}^{29}C_1}{{}^{30}C_2}=\frac{1}{15}$$

Since the probability of choosing any one student is not equal, method 1 does not give a random sample.

Method 2:
Since the children from all classes are grouped together and 100 children corresponding to the 100 numbers randomly selected by the computer forms a sample, each child has an equal chance of $\frac{100}{1300}$ of being selected and the selections are approximately independent* of each other.

Hence method 2 gives a random sample.

*Note:
In method 2, since 1300 is reasonably large, the selection of 100 students will be approximately independent of each other. For example,

P(a student is included in the sample) $=\dfrac{100}{1300}\approx 0.076923$

P(student A is included in the sample given that student B is included) $=\dfrac{99}{1299}\approx 0.076212$

Method 3:
It gives a non-random sample as students who do not eat noodles at all will not patronize the noodle stall and hence have zero probability of being included in the sample.

Having seen a few random and non-random samples, we will now discuss what to do with the information obtained from these samples. One of the main aims in the study of statistics is to be able to make inferences about the population based on sample information.

1.3 Population Parameters - Population Mean and Population Variance

A **population parameter** is a number that is associated with a population characteristic. In many cases, we are concerned with two population parameters, namely the **population mean** and the **population variance**. These are constants of a population and they are often unknown. The study of a population often involves finding estimates of these parameters.

Notation:
The population mean is denoted by μ and the population variance is denoted by σ^2.

1.4 Random Sample of size *n* and Sample Statistics

To define a sample statistic, let's first look at the definition of a random sample of size n.

A **random sample of size *n*** (from some given distribution) is a set of n random variables $X_1, X_2, X_3, \ldots, X_n$, satisfying the following 2 conditions:

(a) Each X_i ($i = 1, 2, 3, \ldots, n$) has exactly the same distribution.
(b) The set $X_1, X_2, X_3, \ldots, X_n$ are mutually independent.

Suppose $X_1, X_2, X_3, \ldots, X_n$ is a random sample of size n taken from some distribution, then a **sample statistic** is a function of these random variables. Thus once the random sample is known, the numerical value of the statistic is known, and its value differs from sample to sample.

Two examples of sample statistics are **sample mean,** $\bar{X}=\dfrac{X_1+X_2+\ldots+X_n}{n}$**, and sample variance,** $\dfrac{1}{n}\sum(X-\bar{X})^2$.

To summarise, a **sample statistic** is a **random variable** which contains information about the sample while any corresponding **population parameter** is a (possibly unknown) **constant**.

2 ESTIMATES OF POPULATION PARAMETERS

One of the main reasons for collecting a sample is to infer from it some information about a population parameter such as mean, variance or proportion.

Statistical inference refers to the methods by which one makes inferences or generalisations about the parameters of a population based on information obtained from samples selected from the population.

We obviously want a "good" estimate, one which is close to the actual (but unknown) population parameter. In statistics, an estimate is considered to be "good" if it's **unbiased**, i.e. the average value of the sample statistic (used to estimate the population parameter) for all possible samples gives the true value of the population parameter.

[For more information on mathematical definitions of estimator, estimates and unbiased estimator, refer to Appendix 1]

2.1 Unbiased Estimates of Population Mean and Variance

Let $x_1, \ x_2, \ x_3, \ldots, x_n$ be observed values from a random sample of size n taken from a population with **unknown** population mean μ and **unknown** population variance σ^2. Then,

1) Sample mean, $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$
2) Sample variance $= \dfrac{1}{n}\sum\left(x-\bar{x}\right)^2 = \dfrac{1}{n}\left[\sum x^2 - \dfrac{1}{n}\left(\sum x\right)^2\right]$

2.1.1 Unbiased Estimate of Population Mean

The sample mean, $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$ is an **unbiased estimate of μ**.

(Proof of this result is included in Section 3.

2.1.2 Unbiased Estimate of Population Variance

An **unbiased estimate** of σ^2 is

$$s^2 = \frac{n}{n-1}\left[\frac{\sum\left(x-\bar{x}\right)^2}{n}\right] = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]. \quad \text{[in Formula List]}$$

Note: From above, $s^2 = \dfrac{n}{n-1}\left(\text{sample variance}\right)$, so the sample variance is **NOT** an unbiased estimate of the population variance σ^2.

Example 2

A random sample of size 50 is taken from a population with mean μ and variance σ^2. The sample data are summarized by $\sum x = 134$ and $\sum x^2 = 1032$.

Calculate the sample mean, sample variance and unbiased estimates of μ and σ^2.

Solution:

Sample mean, $\bar{x} = \frac{1}{n}\sum x = \frac{134}{50} = 2.68$

Sample variance $= \frac{1}{n}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = 13.4576$

An unbiased estimate of μ is the sample mean, $\bar{x} = 2.68$

An unbiased estimate of σ^2 is $s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = 13.7$ (3 s.f.)

Or since sample variance have been found,

An unbiased estimate of σ^2 is $s^2 = \frac{n}{n-1}(\text{sample variance}) = \frac{50}{50-1}(13.4576) = 13.7$ (3 s.f.)

Consider the following values observed by taking a random sample of size 6.

1002, 1004, 1006, 1008, 1010, 1012 --- (I)

Rather than calculating the sample mean and sample variance directly, the above set of values may be transformed to the following values by subtracting 1000 from each value and then dividing by 2:

1, 2, 3, 4, 5, 6 --- (II)

It is easier to calculate unbiased estimates of the population mean and variance of (I) by calculating unbiased estimates using the values in (II) and then transforming them.

Useful Result:

Let $x_1, x_2, x_3, ..., x_n$ be observed values from a random sample of size n taken from a population with mean μ and variance σ^2.

If $y = ax + b$, where a and b are constants, then

(i) $\bar{y} = a\bar{x} + b$,

(ii) $s_y^2 = a^2 s_x^2$.

In particular, given the values of $\sum(x-b)$ and $\sum(x-b)^2$, we define $y = x - b$, and it follows that

(i) $\bar{y} = \bar{x} - b$, and

(ii) $s_y^2 = s_x^2$

Hence, an unbiased estimate of μ is $\bar{x}$, where $\bar{x} = \bar{y} + b$, and an unbiased estimate of σ^2 is s_x^2, where $s_x^2 = s_y^2$.

Example 3

The speeds of 120 randomly selected cars are measured as they pass a camera on a motorway. Denoting the speed by x km/h, the results are summarised by

$$\sum(x-100) = -221, \ \sum(x-100)^2 = 4708.$$

Find unbiased estimates of the population mean and population variance, giving your answers correct to 2 decimal places.

Solution:

Let $y = x - 100$. Then $\sum y = -221$, $\sum y^2 = 4708$.

$$\bar{y} = \frac{1}{n}\sum y = -\frac{221}{120} = -1.8417 \ \text{(4dp)}$$

$$\bar{y} = \bar{x} - 100$$
$$\bar{x} = \bar{y} + 100 \approx 98.16$$

Unbiased estimate of population mean is 98.16 (2 d.p.)

Unbiased estimate of population variance,

$$s_x{}^2 = s_y{}^2 = \frac{1}{n-1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]$$
$$= \frac{1}{119}\left[4708 - \frac{(-221)^2}{120}\right]$$
$$= 36.14 \ \text{(2 d.p.)}$$

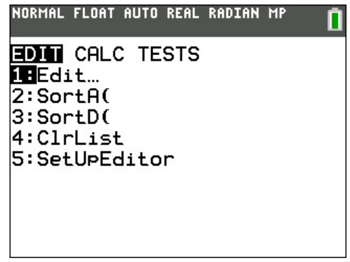
Use of GC to find Unbiased Estimates of Population Mean and Variance

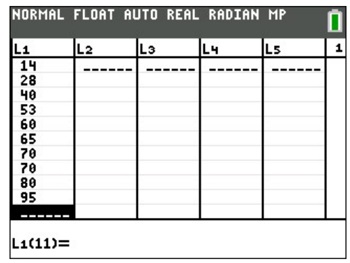
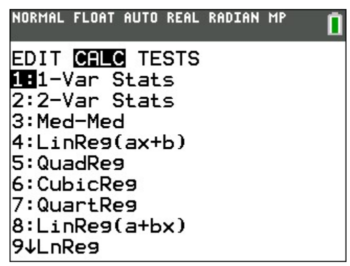
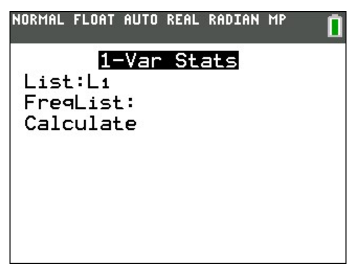
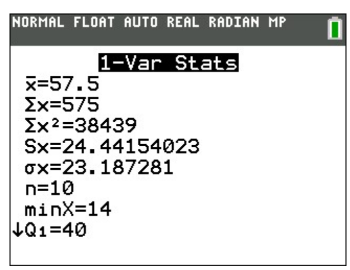
Example 4 (Individual Data)

A random variable has unknown population mean and variance. A sample of 8 data is drawn:

14 28 40 53 60 65 70 70 80 95

Find unbiased estimates of the population mean and variance.

<table>
<tr><td>1.</td><td>Press stat and select **1:Edit…**
Enter the data into L_1.
Return to the Home screen.</td><td>


</td></tr>
<tr><td>2.</td><td>Press stat and scroll to the right to select
CALC. Select **1: 1–Var Stats**.</td><td>
</td></tr>
<tr><td>3.</td><td>Enter **L_1** under **List.**
Leave the **FreqList** empty.
Scroll down to select **Calculate**.</td><td>
</td></tr>
<tr><td>4.</td><td>Press enter to obtain the following screen.
Unbiased estimate of population mean is $\overline{x} = 57.5$
Unbiased estimate of population variance is $s^2 \approx (24.442)^2 = 597$ (3 s.f.)</td><td>
</td></tr>
</table>

Note: In the graphing calculator TI84 plus,

1) $\bar{x}$ denotes an unbiased estimate of the population mean.
2) $s_x^{\ 2}$ denotes an unbiased estimate of the population variance i.e. $s_x^{\ 2} = \dfrac{1}{n-1}\left[\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right]$.
3) $\sigma_x^{\ 2}$ calculates the value of the sample variance, i.e. $\sigma_x^{\ 2} = \dfrac{1}{n}\left[\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right]$.

Example 5 (Grouped Data)

A sample of 80 boys was asked how many hamburgers each could eat for a meal and the results were tabulated. Calculate unbiased estimates for the population mean and variance.

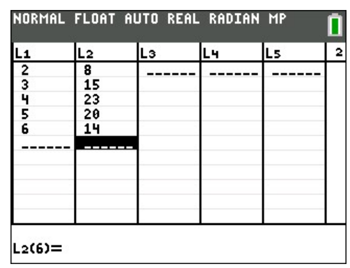
Number of hamburgers	2	3	4	5	6
Frequency	8	15	23	20	14

1. Press stat and select **1:Edit…**

   Enter the data for number of hamburgers into **L1**.

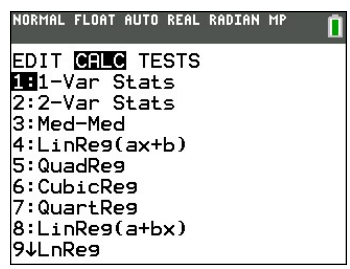
   Enter the frequency into **L2**.

   Return to the Home screen.



2. Press stat and scroll to the right to select **CALC**.

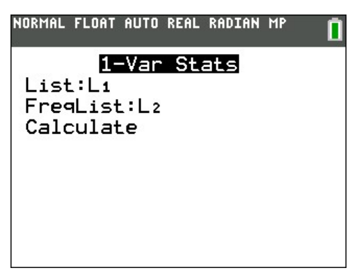
   Select **1: 1–Var Stats**.



3. Enter **L1** under **List**.

   Key in **L2** under **FreqList.**

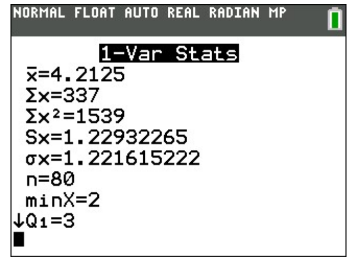
   Scroll down to select **Calculate**.



4. Press **enter** to obtain the following screen.

Unbiased estimate of population mean is $\bar{x} = 4.2125$

Unbiased estimate of population variance is $s^2 \approx (1.2293)^2 = 1.51$ (3 s.f.)



3 The Sample Mean, $\bar{X}$ as a Random Variable

Consider the example of a hospital administrator who is concerned that the hospital sometimes does not have enough beds to cater for all of its patients. She wants to know the average duration that a patient stays in the hospital. In a particular week, she takes a random sample of, say, 10 patients who are admitted, and records the length of time that each of them stays in the hospital.

From Section 2.1.1, the sample mean of the records of these 10 patients, $\bar{x} = \frac{1}{10}\sum x$, is an unbiased estimate of the population mean μ.

It is not difficult to see that we may get different values of $\bar{x}$ with different samples of 10 patients.

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size *n* taken from an infinite population (or finite population if sampling is done with replacement) with mean μ and variance σ^2.
Then the **sample mean** $\bar{X}$, defined by

$$\bar{X} = \frac{1}{n}\sum X = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n},$$

is a random variable with $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Proof:

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

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

since $X_1, X_2, \ ... \ , X_n$ are independent.

$$\begin{aligned}&= \frac{1}{n^2}(\sigma^2 + \sigma^2 + ... + \sigma^2) \\ &= \frac{1}{n^2}(n\sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

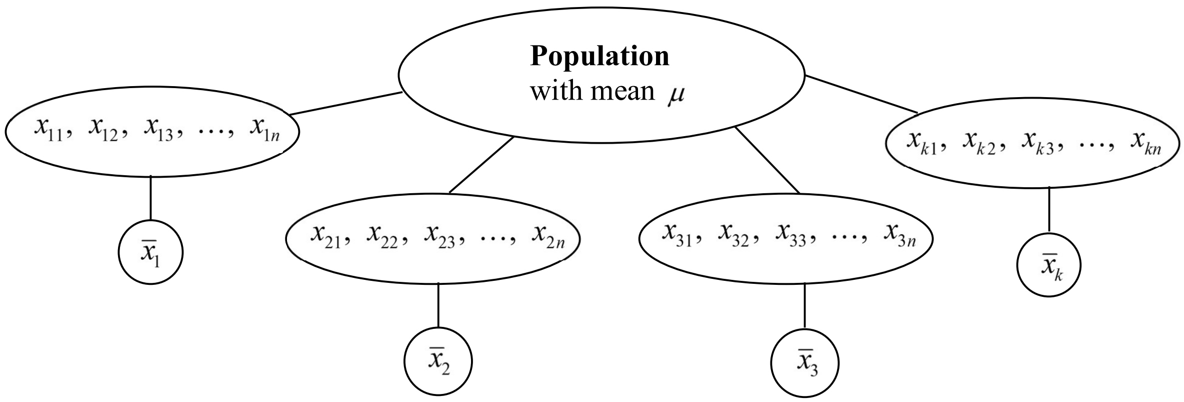
Remark: Could you see the difference between $\overline{X}$ and $\overline{x}$?

$\overline{X}$ is the random variable defined as $\overline{X} = \frac{1}{n}\sum X = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$, whereas $\overline{x}$ is an observed value of $\overline{X}$ which will differ for different samples.

As $\overline{X}$ is a random variable, we are interested in its distribution, which we will discuss in the next section.

4 The Distribution of the Sample Mean

For each sample, the values of the sample mean and variance can be calculated (see diagram below). As the values vary from sample to sample, the sample mean and sample variance (or standard deviation) are examples of random variables. Thus, if all possible samples of a given size are taken from a population with known mean and variance, then the *means of the samples* themselves will form a distribution, which are either discrete or continuous, known as the **distribution of the sample mean**. This sampling distribution depends on the distribution of the population and the sample size.



Example 6 (sampling distribution)
Consider a finite population of size $N = 3$, whose elements are 1, 2, 3, each printed on a piece of card. The distribution of X, the number obtained when a card is drawn at random is

x	1	2	3
Probability	1/3	1/3	1/3

(i) State the value of the population mean and find the population variance.

Samples of size $n = 2$ are drawn repeatedly with replacement. There are altogether 9 possible samples of size 2 from the population. For each sample, we calculate the sample mean $\bar{x}$ as shown. The probability of obtaining each sample mean is 1/9.

(x_1, x_2)	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)	(3, 1)	(3, 2)	(3, 3)
$\bar{x}$	1	1.5	2	1.5	2	2.5	2	2.5	3

(ii) Tabulate the probability distribution of $\bar{X}$.

(iii) Verify that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$.

Solution

(i) Population mean, $\mu = \mathrm{E}(X) = 2$ by symmetry

Population variance, $\sigma^2 = \mathrm{Var}(X) = 1^2(1/3) + 2^2(1/3) + 3^2(1/3) - 2^2 = 2/3$

(ii)

$\bar{x}$	1	1.5	2	2.5	3
Probability	$1/9$	$2/9$	$1/3$	$2/9$	$1/9$

Note: Since each card is replaced after it has been drawn, for each sample, the 2 draws are independent of each other and each of the sample variables X_1, X_2, have the same distribution as that for X. Effectively, we can consider the given finite population (with sample replacement) as an infinite one.

(iii) $\mathrm{E}(\bar{X}) = 1\,(1/9) + 1.5\,(2/9) + 2\,(1/3) + 2.5\,(2/9) + 3\,(1/9) = 2 = \mu$ (by symmetry)

i.e. Mean of the sampling distribution is the same as the mean of the population.

$$\mathrm{Var}(\bar{X}) = 1^2(1/9) + 1.5^2(2/9) + 2^2(1/3) + 2.5^2(2/9) + 3^2(1/9) - 2^2 = 1/3 = \frac{\mathrm{Var}\,X}{n}$$

i.e. Variance of the sampling distribution = $\dfrac{\sigma^2}{n}$ where n is the sample size.

[For information on the use of GC to simulate sampling from a normal population, refer to Appendix 2]

We will consider the sampling distribution for the following two cases:

- Sampling from a Normal Population
- Sampling from a Non-Normal Population

4.1 Sampling from a Normal Population

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n taken from a normal population with mean μ and variance σ^2.

Then **(1)** sample mean $\bar{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.

In other words, the sampling distribution of $\bar{X}$ is a normal distribution with mean μ and variance $\dfrac{\sigma^2}{n}$ if X is taken from a normal distribution.

(2) $X_1 + X_2 + X_3 + \ldots + X_n = n\bar{X} \sim \mathrm{N}(n\mu,\ n\sigma^2)$

Example 7

The height of a particular species of plant follows a normal distribution with mean 21 cm and standard deviation $\sqrt{90}$ cm. A random sample of 10 plants is taken and the mean height calculated. Find the probability that

(i) the sample mean lies between 18 cm and 27 cm,

(ii) $X_1 + X_2 + \ldots + X_{10}$ is more than 200 cm.

State what it means for a sample to be random in this context.

Solution:

Let X be the height of a plant in cm.
Then $X \sim \mathrm{N}(21, 90)$.

(i) Sample mean $\bar{X} \sim \mathrm{N}\left(21, \dfrac{90}{10}\right)$

$\therefore\ \mathrm{P}(18 < \bar{X} < 27) = 0.819$ (3 s.f.)

(ii) $\mathrm{P}(X_1 + X_2 + \ldots + X_{10} > 200)$

$= \mathrm{P}\left(\dfrac{X_1 + X_2 + \ldots + X_{10}}{10} > \dfrac{200}{10}\right)$

$= \mathrm{P}(\bar{X} > 20) = 0.631$ (3 s.f.)

OR Let $T = X_1 + X_2 + \ldots + X_{10}$

$\mathrm{E}(T) = 10\mathrm{E}(X) = 10 \times 21 = 210$

$\mathrm{Var}(T) = 10\mathrm{Var}(X) = 10 \times 90 = 900$

Then $T \sim \mathrm{N}(210, 900)$

$\mathrm{P}(T > 200) = 0.631$ (3 s.f.)

The sample is random if every plant has an equal chance of being selected, and the selection of plants are independent of each other.

Example 8

A large number of random samples of size n are taken from a normal population with mean 74 and variance 36 and the sample means are calculated. If at most 85% of the sample means exceed 72, find the largest possible value of n.

Solution:

Let $X \sim \mathrm{N}(74, 36)$.

Then sample mean $\bar{X} \sim \mathrm{N}\left(74, \dfrac{36}{n}\right)$

Given $\mathrm{P}(\bar{X} > 72) \le 0.85$,

$$\mathrm{P}\left(Z > \frac{72-74}{\sqrt{\frac{36}{n}}}\right) \le 0.85$$

$$\mathrm{P}\left(Z > \frac{-\sqrt{n}}{3}\right) \le 0.85$$

From GC, $\mathrm{P}\left(Z > -1.0364\right) = 0.85$

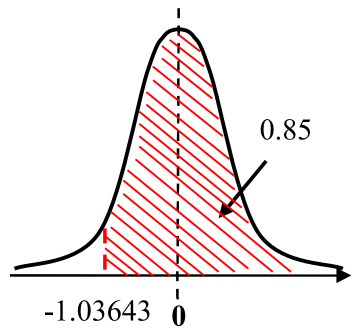
From the diagram,

$$-\frac{\sqrt{n}}{3} \ge -1.0364$$

$$1.0364 \ge \frac{\sqrt{n}}{3}$$

$$n \le 9.67$$

$\therefore$ Largest value of n is 9.



Alternatively: From GC,

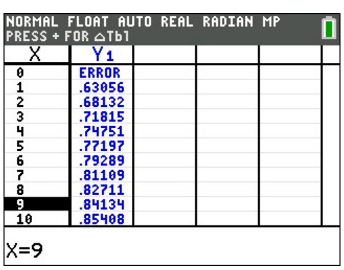
When $n = 9$, $\mathrm{P}(\bar{X} > 72) = 0.841 < 0.85$

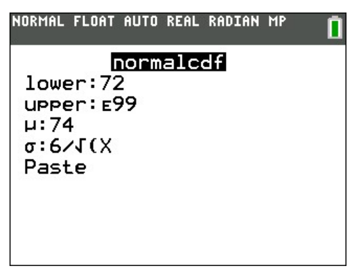
When $n = 10$, $\mathrm{P}(\bar{X} > 72) = 0.854 > 0.85$

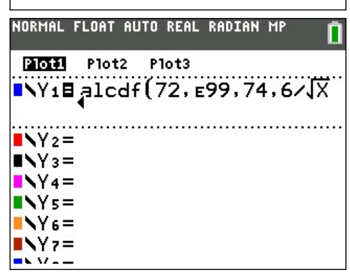
Using the Graphing Calculator for Example 8

Find n such that $\mathrm{P}(\bar{X} > 72) \le 0.85$.

1. Key $Y_1 = \textbf{normalcdf}(72, \mathrm{E}99, 74, 6/\sqrt{X})$
2. Look at the table of values and scroll down to find largest value of n such that $Y_1 = \mathrm{P}(\bar{X} > 72) \le 0.85$







4.2 Sampling from a Non-Normal Population

If the population does not follow a normal distribution, the sample mean also does not follow a normal distribution.

However, if the sample size is large, the distribution of the sample mean will be *approximately* normal. This important result is known as the **Central Limit Theorem**.

Central Limit Theorem

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n taken from a non-normal population with mean μ and variance σ^2.

Then

(1) Sample mean $\bar{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ ***approximately*** if sample size n is large, e.g. $n \geq 30$.

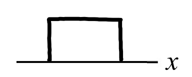
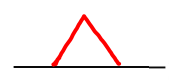
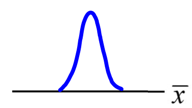
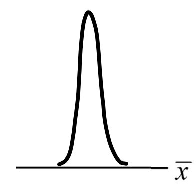
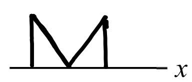
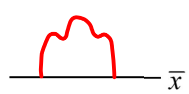
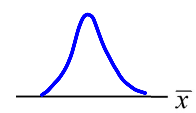
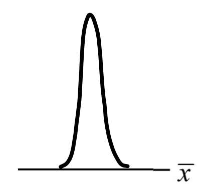
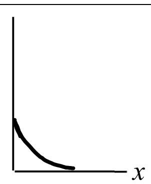
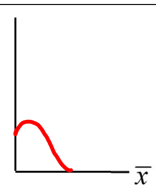
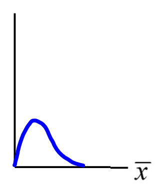
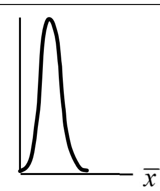
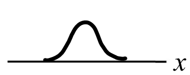
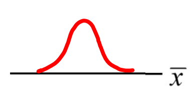
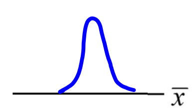
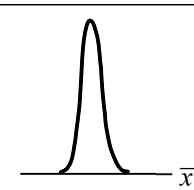
In other words, **for large *n***, the sampling distribution of $\bar{X}$ is ***approximately*** a **normal** distribution with mean μ and variance $\dfrac{\sigma^2}{n}$ if X is taken from a non-normal distribution. The approximation gets better as n gets larger.

(2) $X_1 + X_2 + X_3 + \ldots + X_n = n\bar{X} \sim \mathrm{N}(n\mu,\ n\sigma^2)$ ***approximately*** if sample size n is large, e.g. $n \geq 30$.

A natural question to ask: how large must the sample size n be?

The answer depends on the shape of the distribution of the population as shown by the table below.

Sampling Distributions of $\overline{X}$ for Different Populations and Different Sample Sizes

Original population	Sampling distribution of $\overline{X}$ for $n = 2$	Sampling distribution of $\overline{X}$ for $n = 5$	Sampling distribution of $\overline{X}$ for $n = 30$
			
			
			
			

If the population (i.e. distribution of X) has roughly the shape of a normal distribution, then n will not need to be large for the sampling distribution of $\overline{X}$ to be approximately normal. But if the population is very skewed, then n might need to be quite large. The greater the skewness of the population distribution, the larger the sample size must be before the normal distribution is an adequate approximation for the sampling distribution of $\overline{X}$.

To summarise , how large n needs to be for Central Limit Theorem to hold depends on the context of the problem. In general, $n \geq 30$ is sufficiently large for Central Limit Theorem to hold in most problems.

Example 9

Ten fair dice are thrown and X denotes the number of sixes obtained. If the ten dice are thrown 50 times, *estimate* the probability that the average value of X is less than 1.6.

Solution:

We have $X \sim \mathrm{B}\left(10, \frac{1}{6}\right)$, and $\mathrm{E}(X) = \frac{5}{3}$ and $\mathrm{Var}(X) = \frac{25}{18}$

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

$\bar{X} \sim \mathrm{N}\left(\frac{5}{3}, \frac{\frac{25}{18}}{50} = \frac{1}{36}\right)$ approximately

$\therefore \ \mathrm{P}(\bar{X} < 1.6) = 0.345$ (3 s.f.)

The required probability can be found exactly here. However, when the question states to "estimate" you are required to use the Central Limit Theorem to obtain an approximation as is done here.

Example 10

The standard deviation of the masses of articles in a large population is 4.55 kg. If random samples of sizes 100 are drawn from the population, estimate the probability that a sample mean will differ from the population mean by less than 0.8 kg.

Solution:

Let the population mean and sample mean be μ and $\bar{X}$ respectively.

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

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{4.55^2}{100} = 0.455^2\right) \text{approximately.}$$

$$\begin{aligned}\mathrm{P}\left(\left|\bar{X} - \mu\right| < 0.8\right) &= \mathrm{P}\left(\left|\frac{\bar{X} - \mu}{0.455}\right| < \frac{0.8}{0.455}\right) \\ &= \mathrm{P}\left(-\frac{0.8}{0.455} < Z < \frac{0.8}{0.455}\right) \\ &= 0.921 \text{ (3 s.f.)}\end{aligned}$$

For more information on sampling distribution, check out the following websites or book:

- *http://onlinestatbook.com/chapter7/sampling_distributions.html*
- *A First Course in Statistics* by McClave, J.T. & Sincich, T., Pearson Education (2006)

CONCLUSION

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n taken from a population with mean μ and variance σ^2, and $\bar{X}$ be the sample mean, where $\bar{X} = \frac{1}{n}\sum X = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$.

(A) Expectation and Variance of Sample Mean

> (1) $\mathrm{E}(\bar{X}) = \mu$,
>
> (2) $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

(B) Distribution of Sample Mean

The following table shows the distribution of $\bar{X}$ when the sample is drawn from
- (I) a normal population
- (II) a non-normal population with large sample size n, e.g. $n \geq 30$

Characteristics of population	Distribution of $\bar{X}$
(I) Normal Population	**(1)** $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. In other words, the sampling distribution of $\bar{X}$ is a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ if X is taken from a normal population. **(2)** $X_1 + X_2 + X_3 + \ldots + X_n \sim N(n\mu, n\sigma^2)$
(II) Non-Normal Population with large sample size n, e.g. $n \geq 30$	**Central Limit Theorem** **(1)** $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ ***approximately*** if sample size n is large, e.g. $n \geq 30$. In other words, **for large *n***, e.g. $n \geq 30$, the sampling distribution of $\bar{X}$ is ***approximately*** a **normal** distribution with mean μ and variance $\frac{\sigma^2}{n}$ if X is taken from a non-normal population. The approximation gets better as n gets larger. **(2)** $X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$ ***approximately*** if sample size n is large, e.g. $n \geq 30$.

Appendix 1: Unbiased Estimators of Population Mean and Variance

1.1 Estimator and Estimate

Suppose a population has an unknown parameter τ (tau). To estimate τ, we obtain a random sample from the population and derive a sample statistic T.

We say that T is an **estimator** of τ.
The numerical value taken by the estimator is called an **estimate** of τ.

For example, if the population mean μ is unknown, a sample of size n may be taken and the sample mean $\overline{X}$ calculated. $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ serves as an estimator of μ. $\overline{X}$ can take different values depending on the sample collected (i.e. $\overline{X}$ is a random variable). If the sample mean is 2 for a particular sample collected, then 2 is said to be an estimate of μ.

1.2 Unbiased Estimator and Unbiased Estimate

A statistic T is **an unbiased estimator** of an unknown parameter τ if and only if $\mathrm{E}(T) = \tau$. Given a sample, the value of T based on the given sample is an **unbiased estimate** of τ.

An estimator is considered to be "good" if it is unbiased. This is reasonable since "unbiasedness" means that the long run average value of the estimator T is τ. Putting it another way, the average value of T for all possible samples gives the true value of parameter τ.

Using the above definition, $\overline{X}$ is an unbiased estimator of μ, and S^2 is an unbiased estimator of σ^2, since it can be proven that

1) $\mathrm{E}(\overline{X}) = \mu$, and
2) $\mathrm{E}\left(S^2\right) = \sigma^2$.

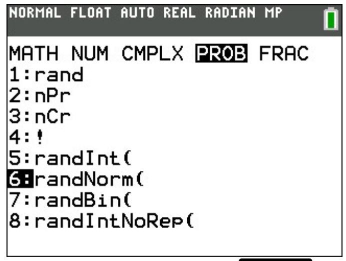
Let T_1 and T_2 be 2 different unbiased estimators for some unknown parameter τ of the population.
We would say that T_1 is a better or more efficient estimator than T_2 if

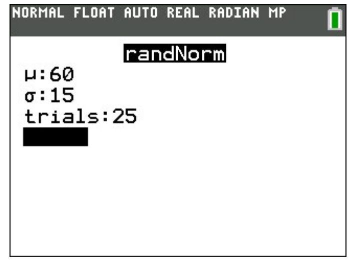
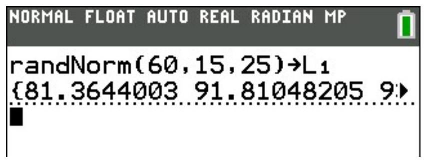
$$\mathrm{Var}\,(T_1) < \mathrm{Var}\,(T_2),$$

since with its smaller variance, T_1 is more likely (in the long run) to be closer to τ.

Appendix 2: Use of GC to simulate Sampling from a Normal Population

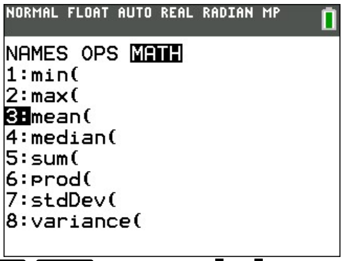
Activity 1a: Simulating a Random Sample from a Normal Population

<table>
<tr>
<td rowspan="2">To obtain a random sample of size $n = 25$ from a normal population with mean $\mu = 60$ and $\sigma = 15$</td>
<td>

Use Math function math</td>
<td>
</td>
</tr>
<tr>
<td colspan="2">

Store the 25 random samples generated into the list L_1 by using sto→</td>
</tr>
</table>

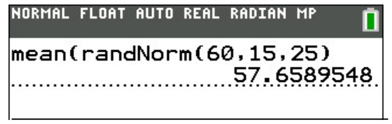
Activity 1b: Calculating Sample Mean of a Random Sample Simulated from a Normal Population



Use [2nd] [stat] to access [list], select

"Math" and **"3:mean"**

Follow the steps in **Activity 1a** to obtain "randNorm (60,15,25)" ,



or select "L_1" if you have stored the random sample values in L_1 earlier.

Note : The value of the sample mean that you have found is most likely not the same as the one that we have found above!

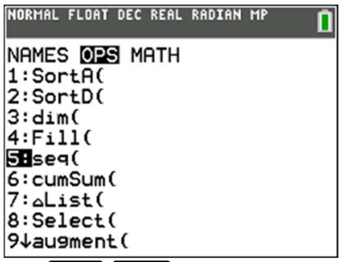
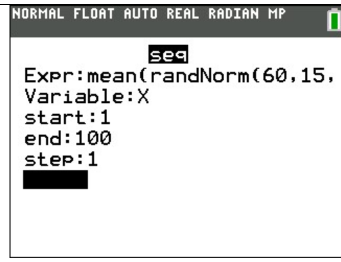
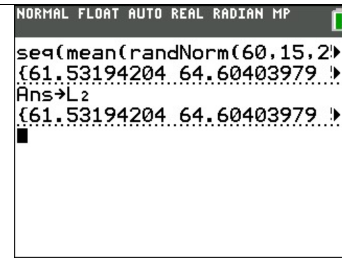
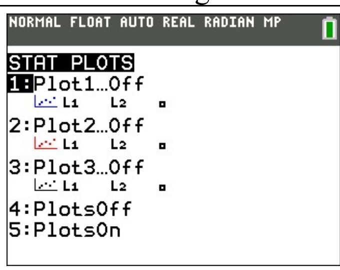
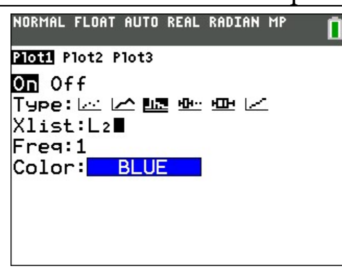
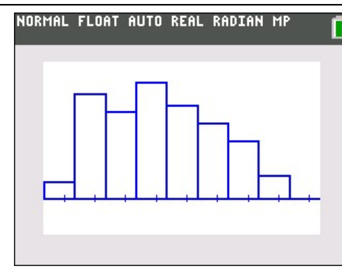
Activity 1c:

Generate a different set of 25 random samples from a normal distribution with the same parameters and calculate the sample mean. What do you notice & what can you conclude?

Ans : The second sample mean that you found is mostly likely not the same as the first sample mean that you found! The sample mean for different random samples of the same size drawn from the same population is highly likely to be different, thus the sample mean is a random variable.

Activity 2: The Sampling Distribution of $\bar{X}$

Use the graphing calculator to generate the mean for 100 samples. Each sample consists of 25 measurements from a normal distribution with $\mu = 60$ and $\sigma = 15$.

 Use 2nd stat to access [list], select **"OPS"** and **"5:seq"**	 Expr: mean(randNorm(60, 15, 25))	 Store the sequence generated into L_2.
To create a histogram for the mean taken from 100 samples:		
 Use 2nd y= to access [stat plot] and select **"1: Plot1..."**	 Turn on the Stat Plot function and select the picture of the histogram. Key in L_2 for the Xlist.	 Use zoom and select **"9↓ZoomStat"** to sketch the histogram

If a smooth curve is drawn through the top of the columns of the histogram, this will give a distribution which is roughly 'bell'-shaped. Is this of any surprise to you?

SUMMARY