Chapter S6

Sampling and Estimation Theory

Objectives:

At the end of this chapter, students should be able to:

(a) understand the concepts of population, random and non-random samples;

(b) understand the concept of sample mean $\overline{X}$ as a random variable with $\mathrm{E}\left(\overline{X}\right)=\mu$ and $\mathrm{Var}\left(\overline{X}\right)=\dfrac{\sigma^2}{n}$

(c) understand that the sample mean (denoted by $\overline{X}$) is a random variable which follows a normal distribution if the population of X follows a normal distribution;

(d) use of the Central Limit Theorem to treat sample mean as having normal distribution when the sample size is sufficiently large;

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

(f) solve real-life problems involving the sampling distribution.

6.1 Introduction

Statistics can be broadly classified into two major categories: **descriptive statistics** and **inferential statistics**.

Descriptive statistics uses the data to provide descriptions of the population, through either numerical calculations or graphs or tables. Typically, there are two general types of statistic that are used to describe data: measures of central tendency (e.g. mean) and measures of spread (e.g. variance).

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

In the 2015 Singapore General Election, The Election Department of Singapore decided to made public, the results of sample count, after some trials.
What is sample count?

https://www.youtube.com/watch?v=vXnjf5MyghA

One reason for this move is to prevent unnecessary speculation and reliance on unofficial sources of information before all the votes are tallied and the final results are announced. Sample counts proved accurate in the 2015 general election, with actual results coming well within the error margin around what was predicted, especially for larger wards.

In this chapter, we will discuss what is meant by random (or probabilistic) and non-random (non-probabilistic) sampling.

6.2 Some Terminologies

A ***population*** is the entire set of items/group of individuals under consideration. The population size can be small, large or infinite.

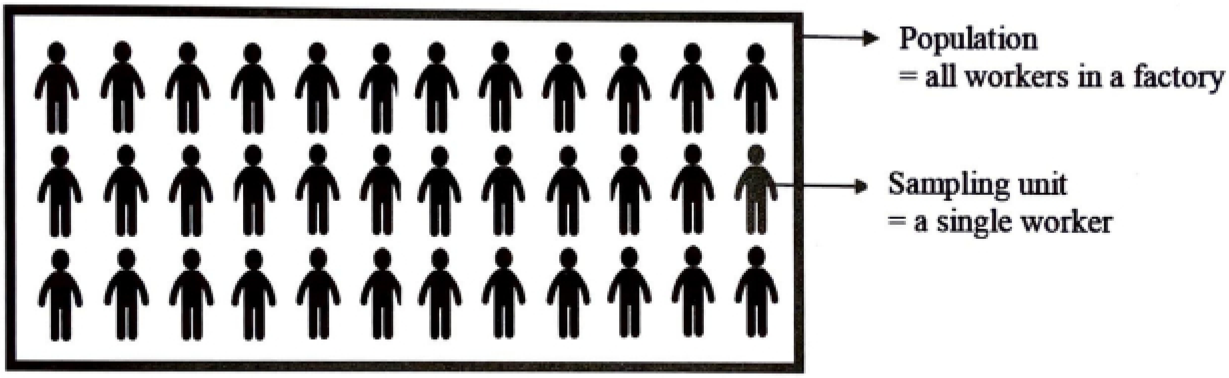
A ***sample*** is a subset of the population that is available for analysis. The purpose of sampling is to obtain a *representative* sample of a population and use it to make an inference about the population. For example, to determine the mean lifetime of a batch of batteries, the manufacturer tests only a sample of batteries rather than the entire batch.

The ***sample size*** is the number of observations obtained for the sample.

A ***sampling unit*** is that element or set of elements considered for selection in some stage of sampling.

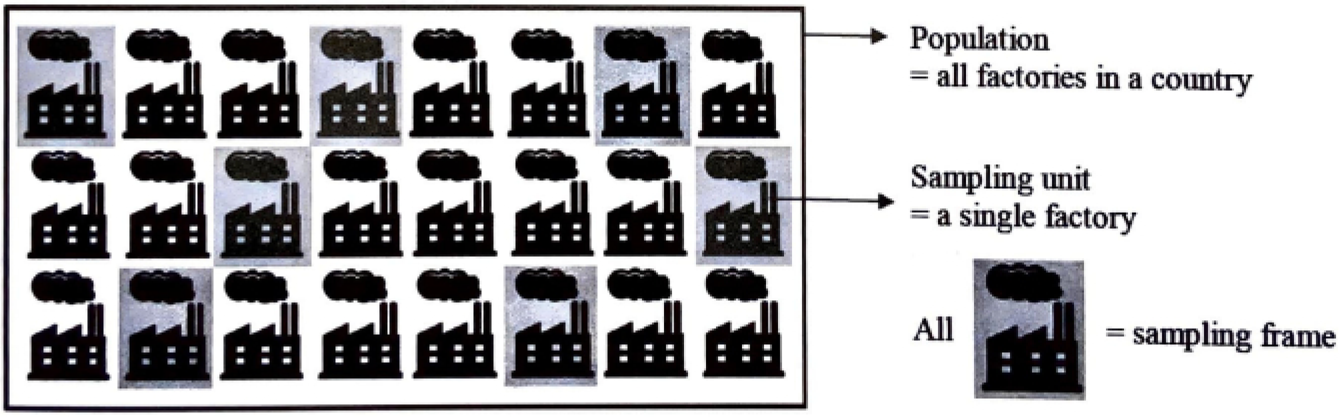
A ***sampling frame*** is a complete list of all the sampling units of the population. The sampling frame provides a base for the selection of the sample.

As an illustration,



Population
= all workers in a factory

Sampling unit
= a single worker



Population
= all factories in a country

Sampling unit
= a single factory

All = sampling frame

To make valid inferences about the population under study, it is important that we obtain a **representative** sample which reflects the population characteristics. Acquiring a representative sample requires the use of random sampling techniques that avoid bias in the sample. **Biasedness** occurs when some members of the population are more likely to be selected than other members of the population, causing some group(s) to be over represented. **Random sampling** ensures that every member of a population has an **equal chance** of being selected and that the sample obtained is not biased.

Example 1:

The NYJC Math Department wants to find out the average battery life of a particular graphing calculator model, and needs to acquire a sample from approximately 700 students. Assume that every student owns the same model of a graphing calculator. Consider the following sampling methods, how can a random sample be formed?

1. The department randomly picks 2 students from each Supplementary Programme class.
2. The department randomly picks a student from each Lecture Theatre group. You may assume that the population is evenly distributed into all three LTs.
3. A Google Form is set up and publicized on the Math noticeboard, accepting the first 50 responses.
4. All math tutors randomly pick a student from each class.

Example 2:

A company sells a certain brand of baby milk powder and wishes to randomly reward 5 customers with free milk vouchers through a lucky draw. Suppose that 2000 customers qualify for the draw. Explain (with appropriate calculations) if there will be equal probability of a particular customer being the first to be selected or the third to be selected for the free milk vouchers.

Solution:

(i)

P(a particular customer is the first to be selected to win the prize) =

P(a particular customer is the third to be selected to win the prize) =

Hence, there will be equal probability of a particular customer being the first to be selected or the third to be selected for the free milk vouchers.

6.3 Introduction to Estimation Theory

Suppose we want to obtain the mean yearly income of all working Singaporeans in the 21-60 age group. A direct way is to get all the Singaporeans in this age group to report their yearly income and then calculate the (actual) mean yearly income. However, this method is infeasible as one has to deal with a large amount of data and some may not be easily obtainable (especially Singaporeans who work overseas). An alternative is to **estimate** the mean yearly income using a **sample**. From this sample, one can easily compute the mean yearly income which should turn out to be a reasonably good estimation of the actual mean we are interested to find out. This leads us to the next discussion on *Estimation Theory*.

6.3.1 Population Parameters and the Sample Statistic

A **population parameter** is a value that describes a population. The population mean, denoted by μ and the population variance, denoted by σ^2 are two common examples of a population parameter. A **sample statistic** is a random variable that describes a sample. The sample statistic is used to estimate the population parameter when the parameter is unknown.

The sample mean, denoted by $\overline{X}$ and the sample variance are two examples of a sample statistic.
Note on notation: $\overline{X}$ (uppercase) is a random variable used to denote the sample mean and $\overline{x}$ (lowercase) denotes the value of $\overline{X}$ based on the sample chosen.

6.3.2 Sample Mean and Variance

For a given sample of n observations $x_1, x_2, \ldots, x_n$, (i.e. sample size is n)

(1) the **sample mean**, $\bar{x}$, is given by

$$\bar{x} = \frac{\sum x}{n},$$

$$\text{or } \bar{x} = \frac{\sum(x-a)}{n} + a,$$

where a is a constant (called the **assumed mean**)

(2) the **sample variance**,

$$\text{Sample variance} = \frac{\sum x^2}{n} - \bar{x}^2,$$

$$\text{equivalently, sample variance} = \frac{1}{n}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right],$$

or

$$\text{sample variance} = \frac{1}{n}\sum(x-\bar{x})^2 .$$

6.3.3 Unbiased Estimators of the Population Mean and Variance

Let T be a sample statistic derived from a sample used to estimate a population parameter θ. We say that T is an **unbiased estimator** of θ if $\mathrm{E}(T) = \theta$, i.e. the mean of the distribution of the statistic T is equal to the parameter θ.

For example, $\bar{X}$ is an unbiased estimator of μ since $\mathrm{E}(\bar{X}) = \mu$, i.e. the sample mean is an unbiased estimator of the population mean (refer to Annex for the proof). Note that the sample variance, however, is not an unbiased estimator of the population variance. This is because the sample variance underestimates the population variance and an adjustment is needed to give an unbiased estimate of the population variance. The proof is beyond the scope of the syllabus and we will just apply the formula when it is appropriate for all questions.

Let's consider a random sample of size n be taken from a population.

1. The ungrouped data $x_{1,}, x_2, x_3, \ldots, x_n$ is usually presented in summarized form: $\sum x$, $\sum x^2$ or $\sum(x-\bar{x})^2$.

   I. The **unbiased estimate of the population mean** μ is the sample mean given by

$$\bar{x} = \frac{\sum x}{n}.$$

II. The **unbiased estimate of the population variance** σ^2, denoted by s^2, is given by

$$s^2 = \frac{n}{n-1} \times \text{sample variance},$$

equivalently $$s^2 = \frac{n}{n-1}\left[\frac{\sum x^2}{n} - \bar{x}^2\right],$$

or $$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$ (in List MF 26)

or $$s^2 = \frac{n}{n-1}\frac{\sum(x-\bar{x})^2}{n}$$ (in List MF 26)

2. Often, the summary data may be given in the form $\sum(x-a)$ and $\sum(x-a)^2$ where a is the assumed mean.

In this case,

unbiased estimate of the population mean μ,

$$\bar{x} = \frac{\sum(x-a)}{n} + a$$

unbiased estimate of the population variance σ^2,

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

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

Remark: We can remember the formula for s^2 by "replacing" x by $x - a$.

Example 4:

Find the unbiased estimates of the population mean and variance for the following:

(a) $n = 8,\ \sum x = 152.98,\ \sum x^2 = 2927.1$

(b) $n = 50,\ \sum x = 4537,\ \sum(x-\bar{x})^2 = 4825.62$

(c) $n = 50,\ \sum(x-20) = 30,\ \sum x^2 = 21300$

(d) $n = 12,\ \sum(x-10) = 23.5,\ \sum(x-10)^2 = 48.72$

(e) The observed values are

17 5 3 10 4 3 10 5 2 14

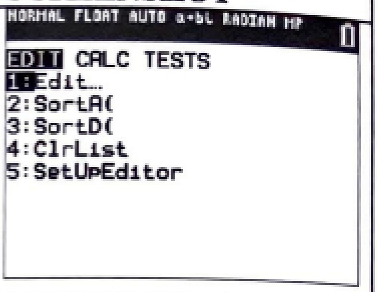
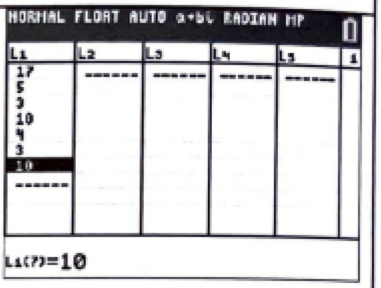
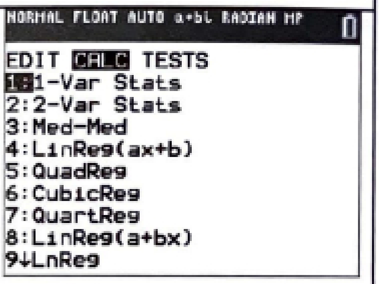
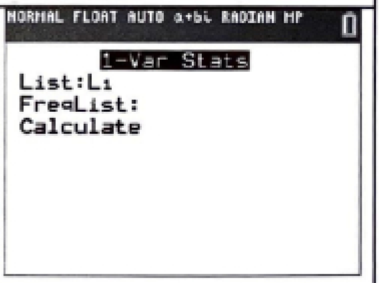
(f) The observed values are

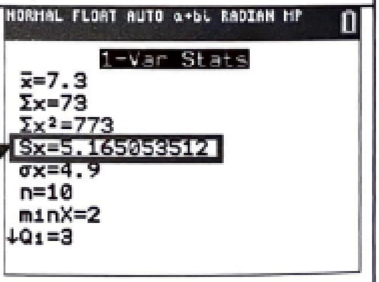
x	17	5	3	10	4	3	10	5	2	4
frequency, f	1	2	4	3	2	1	1	6	3	1

Solution:

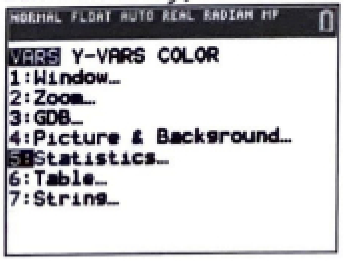
(a) $\bar{x} = \frac{\Sigma x}{n} =$ $s^2 = \frac{1}{n-1}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right] =$ $\approx 0.24856 = 0.249$ (3 s.f.)	Sample mean $\bar{x} = \frac{\sum x}{n}$. (Not found in MF26) Unbiased variance estimate formulae in MF 26, pg 6.
(b) $\bar{x} = \frac{\sum x}{n} = \frac{4537}{50} = 90.74 = 90.7$ (3 s.f.) $s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 =$	
(c) $\bar{x} =$ $\bar{x} = \frac{\sum x}{n} \Rightarrow$ $s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] =$ $= 1.6735 \approx 1.67$ (3 s.f.)	Understand that 20 is subtracted from every data, thus adding 20 back to the mean is logical.
(d) $\bar{x} = \frac{\Sigma(x-10)}{n} + 10 = \frac{23.5}{12} + 10 = 11.96 = 12.0$ (3 s.f.) $s^2 = \frac{1}{n-1}\left\{\Sigma(x-10)^2 - \frac{\left[\Sigma(x-10)\right]^2}{n}\right\} =$ $= 0.24538$ $= 0.245$ (3 s.f.)	What is the similarity of this formula with that in the MF26?

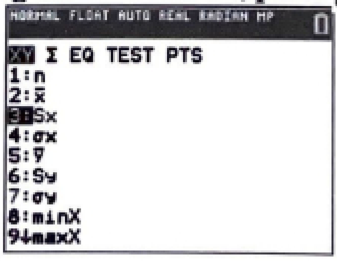
(e) Instead of doing it manually, we can compute the estimates by using GC.

PROCEDURE	SCREENSHOT
Press STAT and select **1:Edit**.	 NORMAL FLOAT AUTO a+bi RADIAN MP EDIT CALC TESTS 1:Edit… 2:SortA(3:SortD(4:ClrList 5:SetUpEditor
Enter the data into L1.	 17 5 3 10 4 3 10 ------ L1(7)=10
Press STAT › to **CALC** and select **1:1-Var Stats**	 NORMAL FLOAT AUTO a+bi RADIAN MP EDIT CALC TESTS 1:1-Var Stats 2:2-Var Stats 3:Med-Med 4:LinReg(ax+b) 5:QuadReg 6:CubicReg 7:QuartReg 8:LinReg(a+bx) 9↓LnReg
Press ⌄ ⌄ and select **Calculate** and press ENTER	 List:L1 FreqList:

Thus $\bar{x} = 7.30$ (3 *s*.f.)
$s^2 = (5.1651)^2 = 26.678 = 26.7$ (3 *s*.f.)
Take this value to find the unbiased estimate of the population variance, s^2.
	
	x̄=7.3
	Σx=73
	Σx²=773
	Sx=5.165053512
	σx=4.9
	n=10
	minX=2
	↓Q1=3

Alternatively, after entering the data into L1, press vars to quickly access $\bar{x}$ and S*x*.

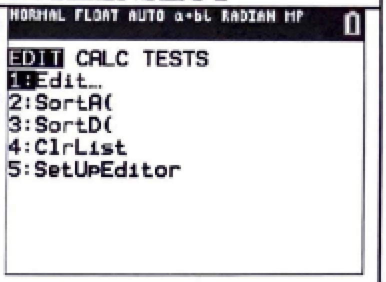




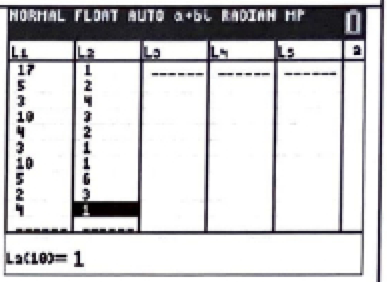
(f) Similar to (e), we can compute the estimates by GC.

PROCEDURE | **SCREENSHOT**

Press STAT and select **1:Edit**.



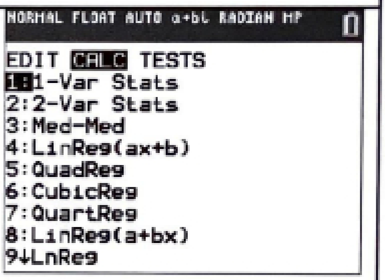
Enter the data into L1 and the corresponding frequency into L2 .



L1	L2	L3	L4	L5
17	1	------	------	------
5	2			
3	4			
10	3			
4	2			
3	1			
10	1			
5	6			
2	3			
4	1			
------	------			

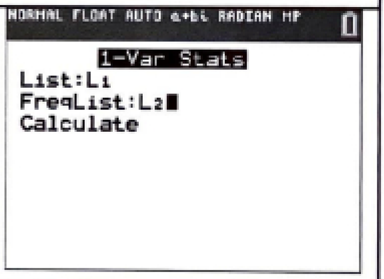
L2(10)= 1

Press STAT ▶ to **CALC** and select **1:1-Var Stats**



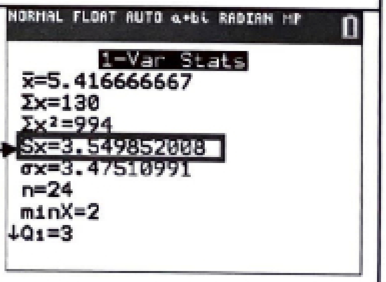
Press ▼ ▼ and select **Calculate** and press ENTER

Under FreqList:, key in L2 (by pressing 2nd STAT to access the 'List' menu)



List:L1
FreqList:L2
Calculate

Take this value to find the unbiased estimate of the population variance, s^2.



```
x̄=5.416666667
Σx=130
Σx²=994
Sx=3.549852008
σx=3.47510991
n=24
minX=2
↓Q1=3
```

Thus $\bar{x} = 5.42$ (3 s.f.)

$$s^2 = (3.5499)^2 = 12.601 = 12.6 \quad (3\ s.f.)$$

6.4 Distribution of the Sample Mean $\overline{X}$

In the previous section, we discussed the use of just one sample to estimate the population mean and variance. In reality, if we repeat the sampling process several times on the same population, we are able to obtain many samples of the same size but with different sample means. Hence, it will be interesting to study the distribution of the sample mean which is denoted by $\overline{X}$.

To find the distribution of the sample mean of a discrete random variable, we perform the following:

- Take a random sample of n independent observations (i.e. sample size n) from a population.
- Calculate the mean of these n sample values.
- Repeat the above 2 steps until we have taken all possible samples of size n and calculated the sample mean of each one.
- Form a distribution of all the sample means we have thus obtained.

The distribution thus formed is called the **(probability) distribution of the sample mean.**

To illustrate this process, let us consider a random variable X with the following distribution:

X	1	2	3
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

We shall demonstrate how to obtain the distribution of $\overline{X}=\dfrac{X_1+X_2+X_3}{3}$.

We have to consider all possible samples of 3 observations. The values of the 3 observations in each sample and the corresponding sample mean are tabulated as shown.

Sample Observation	$\overline{X}$
(1, 1, 1)	1
(1, 1, 2)	$\frac{4}{3}$
(1, 1, 3)	$\frac{5}{3}$
(1, 2, 1)	$\frac{4}{3}$
(1, 2, 2)	$\frac{5}{3}$
(1, 2, 3)	2
(1, 3, 1)	$\frac{5}{3}$
(1, 3, 2)	2
(1, 3, 3)	$\frac{7}{3}$

Sample Observation	$\overline{X}$
(2, 1, 1)	$\frac{4}{3}$
(2, 1, 2)	$\frac{5}{3}$
(2, 1, 3)	2
(2, 2, 1)	$\frac{5}{3}$
(2, 2, 2)	2
(2, 2, 3)	$\frac{7}{3}$
(2, 3, 1)	2
(2, 3, 2)	$\frac{7}{3}$
(2, 3, 3)	$\frac{8}{3}$

Sample Observation	$\overline{X}$
(3, 1, 1)	$\frac{5}{3}$
(3, 1, 2)	2
(3, 1, 3)	$\frac{7}{3}$
(3, 2, 1)	2
(3, 2, 2)	$\frac{7}{3}$
(3, 2, 3)	$\frac{8}{3}$
(3, 3, 1)	$\frac{7}{3}$
(3, 3, 2)	$\frac{8}{3}$
(3, 3, 3)	3

Next we need to calculate the probability corresponding to each value taken by $\overline{X}$.

For example, the sample mean $\bar{x}=\frac{5}{3}$ is obtained from samples (1, 1, 3), (1, 2, 2), (1, 3, 1), (2, 1, 2), (2, 2, 1), (3, 1, 1).

Thus

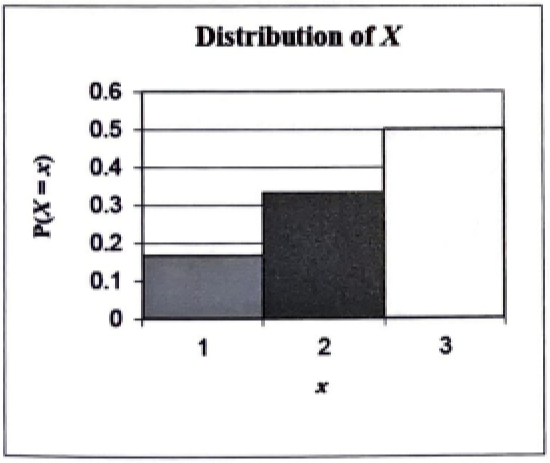
$$\begin{aligned}
\mathrm{P}\left(\bar{X}=\frac{5}{3}\right) &= \mathrm{P}(X_1=1,X_2=1,X_3=3)\ +\ \mathrm{P}(X_1=1,X_2=2,X_3=2)\ +\ \mathrm{P}(X_1=1,X_2=3,X_3=1)\\
&\quad +\ \mathrm{P}(X_1=2,X_2=1,X_3=2)\ +\mathrm{P}(X_1=2,X_2=2,X_3=1)\ +\ \mathrm{P}(X_1=3,X_2=1,X_3=1)\\
&= 3\mathrm{P}(X_1=1,X_2=1,X_3=3)+3\mathrm{P}(X_1=1,X_2=2,X_3=2)\\
&= 3\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{2}\right)+3\left(\frac{1}{6}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\\
&= \frac{7}{72}
\end{aligned}$$

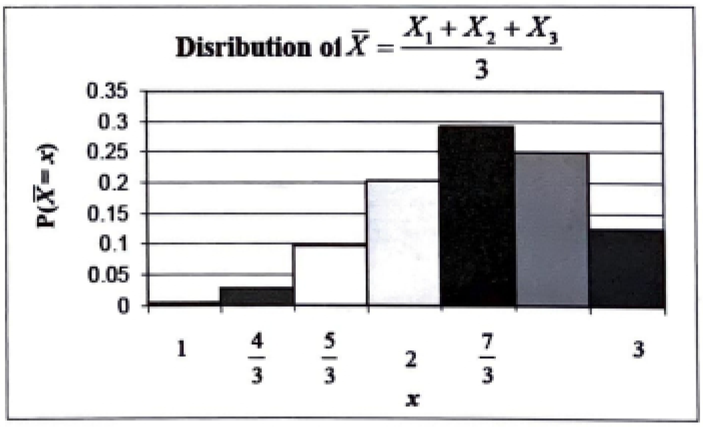
By repeating the above procedure for each possible value of $\bar{X}$, we will be able to tabulate the probability distribution for $\bar{X}$:

$\bar{x}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$	$\frac{8}{3}$	3
$\mathrm{P}(\bar{X}=\bar{x})$	$\frac{1}{216}$	$\frac{1}{36}$	$\frac{7}{72}$	$\frac{11}{54}$	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{8}$

Notice that the calculation of the distribution of the sample mean is a tedious process. However, in most practical cases, we are only interested to know the mean and variance of $\bar{X}$.

The distribution of X and $\bar{X}$ can be shown graphically as below.





We observe that the distribution of $\bar{X}$ roughly follows a 'bell-shaped' curve, reminiscent of the normal distribution we encountered in Chapter 5 even though it is a discrete random variable. It can be shown that if the sample size n gets larger and larger, the distribution of $\bar{X}$ can be well approximated by the normal distribution with the same mean. This motivates the discussion of the distribution of $\bar{X}$ when the sample is drawn from a population that is normally distributed and from a population that is not normally distributed.

6.4.1 Distribution of the Sample Mean taken from a Normal Distribution

If X is taken from a normal distribution, then $\bar{X}$ is also normally distributed, i.e.

if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$.

Remarks:

1. $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where $i = 1, 2, 3, \ldots, n$, is the i-th observation of X.
   We say that the X_i's are **independent and identically distributed (i.i.d.)**.

2. $\bar{X}$ is said to be the mean (average) of n **independent observations** of X.

This result says that if $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a *Normal distribution* with known mean μ and variance σ^2, then the sample mean $\bar{X}$ will also follow a *Normal distribution* given by $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Proof:

Let $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where $i = 1, 2, \ldots\ldots, n$.

Then,

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n}\mathrm{E}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n}\left[\mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)\right] \\
&= \frac{1}{n}(\mu + \mu + \ldots + \mu) \qquad \text{(since } X_i\text{'s are identically distributed)} \\
&= \frac{1}{n}(n\mu) = \mu
\end{aligned}$$

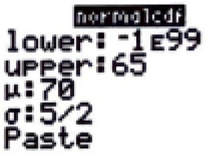
$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}\left[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right] \\
&= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] \quad \text{(since } X_i\text{'s are identically distributed)} \\
&= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}
\end{aligned}$$

Recall that a linear combination of independent normal random variables will also follow a normal distribution. Hence result follows.

Example 5:

At a certain college, the masses of male students are known to follow a normal distribution with mean mass 70 kg and standard deviation 5 kg. Four male students are chosen at random. Find the probability that their mean mass is less than 65 kg.

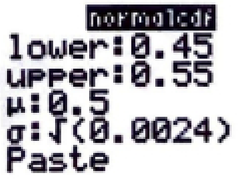
Solution:

Let X denote the mass of a male student. Thus $X \sim \mathrm{N}(70, 5^2)$.	Define the random variable.
Hence, $\bar{X} \sim N\left(70, \frac{5^2}{4}\right)$ $\therefore P(\bar{X} < 65) = 0.022750 = 0.0228$	Press **[2ND] [VARS]** for **DISTR** and select **2:normalcdf(**  normalcdf lower: -1E99 upper:65 μ:70 σ:5/2 Paste

Example 6:

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 both distributions are normal and random samples of 100 components from manufacturer *A* and of 50 from *B* are tested, find the probability that the mean breaking strength of the components from manufacturer *A* will be between 0.45 kN and 0.55 kN more than the mean of those from manufacturer *B*.

Solution:

Let $\bar{X}$ denote the mean breaking strength of the components from manufacturer *A*, then $\bar{X} \sim N\left(6, \frac{0.4^2}{100}\right)$ i.e. $\bar{X} \sim N(6, 0.0016)$	Define the random variables accordingly.
Let $\bar{Y}$ denote the mean breaking strength of the components from manufacturer *B*, then $\bar{Y} \sim N\left(5.5, \frac{0.2^2}{50}\right)$ i.e. $\bar{Y} \sim N(5.5, 0.0008)$	
$\mathrm{E}(\bar{X}-\bar{Y}) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = 6 - 5.5 = 0.5$ $\mathrm{Var}(\bar{X}-\bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = 0.0016 + 0.0008$ $= 0.0024$ We have $\bar{X}-\bar{Y} \sim N(0.5, 0.0024)$ Thus, $\mathrm{P}(0.45 < \bar{X}-\bar{Y} < 0.55) = 0.69257 = 0.693$ (to 3 sf)	 normalcdf lower:0.45 upper:0.55 μ:0.5 σ:√(0.0024) Paste

a) $X \sim N(1003, 42)$ $\bar{X} \sim (1003, \frac{42}{20})$ $P(\bar{X}<1000) = 0{\cdot}019216\cdots$ $0{\cdot}0192$ (to 3sf)

Self-Review 1 :

(a) The amount, X cm^3, of liquid in cartons of orange juice is normally distributed with mean 1003 and variance 42. A random sample of 20 cartons is taken. Find the probability that the sample mean volume will be less than 1000 cm^3. [0.0192]

(b) A factory produces ball bearings with diameter X cm, where $X \sim \mathrm{N}(1, 0.0002)$. An independent sample of size 30 is taken. Find the value of k so that the factory can be 95% sure that the sample mean will be within k cm of 1 cm. [$k = 0.00506$]

b) $\bar{X} \sim N(1, \frac{0.0002}{30})$ $P(|\bar{X}-k|<1) = 0{\cdot}95$ $P(-1<\bar{X}-k<1) = 0{\cdot}95$

6.4.2 Distribution of the Sample Mean taken from a non-Normal Distribution

We now state without proof the most important theorem in Statistics – *The Central Limit Theorem* (or CLT for short), which states that the distribution of $\bar{X}$ approaches a normal distribution when n is large even though the distribution of X is not normal.

The Central Limit Theorem

Let X follows any **non-normal** distribution with mean μ and variance σ^2.
If n is **large** (say $n \geq 30$), then $\bar{X}$ is approximately normally distributed, that is,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately, where } \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Equivalently,

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots\ldots + X_n \sim \mathrm{N}(n\mu, n\sigma^2) \text{ approximately.}$$

Remark: As a guideline, a sample size that is at least 30 is considered large. However, some distributions can be well approximated by the standard normal distribution for small values of the sample size (a symmetric distribution). For our syllabus, we must assume the sample size is reasonably large in order for us to apply Central Limit Theorem.

Example 7 (2012/RJC/II/5):
The monthly rainfall in a region is modelled by a random variable with mean 159 mm and standard deviation 45 mm. Find the probability that the average monthly rainfall over five years is less than 150 mm. State an assumption you have used in your calculations.

Solution:	**Define** the random variable in the context of the question.
Let X be the monthly rainfall in the region. $\mathrm{E}(X) = 159$ and $\mathrm{Var}(X) = 45^2$	**Identify** the distribution – note that the distribution is not stated in the question.
	Note the key word "average" in the question. To determine the value of n, ask what we are averaging over – average monthly rainfall over 5 years (60 months). Hence $\bar{X} = \frac{X_1 + X_2 + \cdots + X_{60}}{60}$
Assume that the rainfall in each month is independent and identically distributed.	
	Compute the probability. Any logical answer in the context of the question.

Example 8:

X is the number of heads obtained when an unbiased coin is tossed nine times. Fifty random observations are taken and the sample mean is calculated. Find the probability that the sample mean exceeds 5.

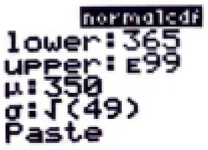
Solution:

Given $X \sim \mathrm{B}\left(9, \frac{1}{2}\right)$, we have $\mu = np = 9(\frac{1}{2}) = 4{\cdot}5$ and $\sigma^2 = np(1-p) = 9(\frac{1}{2})(\frac{1}{2})$ $= 2{\cdot}25$ since $n=50$ is large, By central limit theorem, $\bar{X} \sim N(4{\cdot}5, \frac{2{\cdot}25}{50})$ approximately. $\therefore P(\bar{X}>5) \approx 0{\cdot}00921$ (to 3 sf)	What is E(X) and Var (X) for a Binomial distribution?

Example 9:

The mean time spent by children of age twelve on watching TV programmes in one day is 3.5 hours with a standard deviation of 0.7 hours. Find the probability that in a random sample of 100 children, the total time spent on watching TV in one day exceeds 365 hours.

Solution:

Let X_i be the amount of time the i^{th} child spends watching TV in one day.	**Define** the random variable in the context of the question.
$\mathrm{E}(X_1 + X_1 + ... + X_{100}) = 100E(X)$ $= 100(3.5)$ $= 350$ $\mathrm{Var}(X_1 + X_2 + ... + X_{100}) = 100\mathrm{Var}(X)$ $= 100(0.7)^2$ $= 49$	**Identify** the distribution – note that the distribution is not stated in the question. Note the key words "total time" in the question. To determine the value of n, ask what we are summing over – total time spent by 100 children in one day.
Since $\mu = 3{\cdot}5$, $\sigma = 0{\cdot}7$ and $n=100$ is large, by CLT, $Y = \sum_{i=1}^{100} X_i \sim N(100 \times 3{\cdot}5, 100 \times 0{\cdot}7^2)$ i.e. $Y \sim N(350, 49)$ approximately $P(Y>365) \approx 0{\cdot}0161$ (3sf)	**Compute** the probability.  normalcdf lower:365 upper:E99 μ:350 σ:√(49) Paste

Self-Review 2 :

The mass of cockatoos in a bird sanctuary is found to have a mean of 6.7 kg and a standard deviation of 3.1 kg. A random sample of 300 cockatoos is taken.

(i) The mean mass of the 300 cockatoos is denoted by $\overline{X}$. State, with justification, the approximate distribution of $\overline{X}$, giving its mean and variance. [N(6.7,3.1²/300)]

(ii) Find the probability that the mean found in (i) is between 6.5 kg and 6.8 kg. [0.580]

(iii) Find the probability that the total mass of the cockatoos exceeds 2040 kg. [0.288]

(iv) If there is a probability of more than 0.75 that the mean mass of a random sample of n cockatoos is greater than 6.5 kg, obtain the least value of n. [110]

Example 10: (Self-reading)

A game is played in which an unbiased die is tossed 50 times and the number of 6's is counted.

(i) Find the expected number of 6's in a game.

(ii) 100 games are played. Find the probability that the average number of 6's obtained in a game is more than 10.

Solution:

(i) Let X be the number of 6's obtained out of 50 tosses of the die (in a game) Then $X \sim \mathrm{B}\left(50,\frac{1}{6}\right)$ Expected number of 6's in a game, $\mathrm{E}(X)=50\left(\frac{1}{6}\right)=\frac{25}{3}$ (ii) Since $n=100$ is large, by CLT, $\overline{X}\sim \mathrm{N}\left(\frac{25}{3},\frac{\frac{125}{18}}{100}\right)$ approximately $\overline{X}\sim \mathrm{N}\left(\frac{25}{3},\frac{5}{72}\right)$ approximately $\mathrm{P}\left(\overline{X}>10\right)\approx 1.2754\times 10^{-10}$ $=1.28\times 10^{-10}$ (3 s.f.)	Recall the Binomial Distribution. Note the key word "average" in the question. To determine the value of n, ask what we are averaging over – average number of 6's *per game*. Hence $\overline{X}=\frac{X_1+X_2+\cdots+X_{100}}{100}$

6.5 Miscellaneous Examples

Example 11:
A baker, on average, sells 4 out of 5 muffins that he prepares for a day. It is known that he prepares exactly 200 muffins a day.

(i) In any randomly chosen day, find the probability that he can sell between 160 and 196 muffins inclusively.

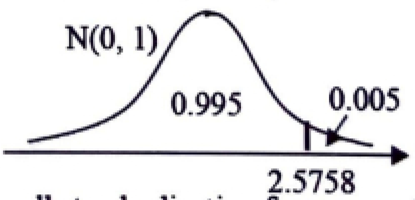
(ii) Calculate the probability that the mean number of muffins sold per day is at least 159 over a period of 60 days.

Solution:

(i) Let X be the number of muffins sold per day, out of 200 muffins . Then $X \sim \mathrm{B}\left(200, \frac{4}{5}\right)$ $\mathrm{P}(160 \le X \le 196) = \mathrm{P}(X \le 196) - \mathrm{P}(X \le 159) = 0.542$. (ii) $\mu = E(X) = np = 200\left(\frac{4}{5}\right) = 160$ $\sigma^2 = Var(X) = npq = 200\left(\frac{4}{5}\right)\left(\frac{1}{5}\right) = 32$ Since $n=60$ is large, by CLT . $\bar{X} \sim N\left(160, \frac{32}{60}\right)$ approx , $P(\bar{X} \ge 159) \approx 0.915$.	Recall the Binomial Distribution. Look out for the key words in this part: ' ***mean*** number of muffins sold ***per day***'

Example 12:
Suppose the distribution of the random variable X is N(25, 340.48) and the mean of a random sample of size n drawn from this distribution is denoted by $\bar{X}$. Find the value of n, correct to the nearest integer, given that $\mathrm{P}(\bar{X} > 28)$ is approximately 0.005.

Solution:

$\bar{X} \sim \mathrm{N}\left(25, \frac{340.48}{n}\right)$ $\mathrm{P}(\bar{X} > 28) \approx 0.005$ $\mathrm{P}(\bar{X} \le 28) \approx 0.995$ $\Rightarrow \mathrm{P}\left(Z \le \frac{3\sqrt{n}}{\sqrt{340.48}}\right) \approx 0.995$ $P\left(Z \le \frac{28-25}{\sqrt{\frac{340.07}{n}}}\right)$	Should we apply CLT for this question? If not necessary, why? Question has stated that X is N(25, 340.48) , thus the sample mean $\bar{X}$ is also normally distributed.  Recall standardization from normal distribution.

Remark: If the population parameters are not given, we will need to use the sample mean and unbiased estimate for the population variance to estimate the parameters. For a normal distribution, we will use $\bar{x}$ and s^2 to estimate μ and σ^2 respectively. The following example illustrates this. The distribution of the sample mean in this case will still be approximately normal for a large sample size.

Example 13:
A mathematics test for JC2 students produces scores which are normally distributed on a scale from 0 to 100. A random sample of 160 JC2 students were assessed, and each of their individual test scores, x, was recorded. The results are summarised by

$$\sum x = 9120, \qquad \sum (x-\bar{x})^2 = 35775.$$

(i) Calculate the unbiased estimates of the population mean and variance.

(ii) Assuming that the unbiased estimate of the population mean obtained above is equal to the population mean, find the value of c such that there is a probability of 0.95 that a sample of 160 scores has a sample mean that differs from its mean score by less than c.

(iii) Comment on the validity of the calculations if the scores were not normally distributed.

Solution:

(i) Unbiased estimate of the population mean μ,

$$\bar{x} = \frac{\sum x}{n} = \frac{9120}{160} = 57.$$

Unbiased estimate of the population variance σ^2,

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{35775}{159} = 225$$

(ii) Let $\bar{X}$ denote the random variable representing the mean test scores of a sample of 160 randomly chosen JC2 students.

Then $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{160})$.

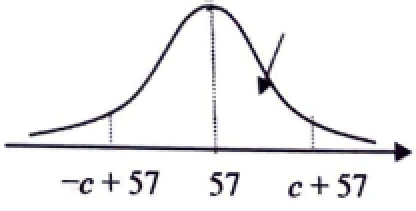
Assuming $\mu = 57$ and estimating σ^2 using s^2 ($= 225$) from (i),

we have $\bar{X} \sim \mathrm{N}\left(57, \frac{225}{160}\right)$ ~~approximately~~.

$P(|\bar{X}-57|<c)=0.95$

P

Note that the region is "center".



Use invNorm(0.025,57, $\sqrt{225/160}$, CENTER)

Alternative Approach (Using standardization)

$$\mathrm{P}\left(\left|\overline{X}-57\right|<c\right)=0.95$$

$$\Rightarrow \mathrm{P}\left(\left|\frac{\overline{X}-57}{\sqrt{225/160}}\right|<\frac{c}{\sqrt{225/160}}\right)=0.95$$

$$\Rightarrow \mathrm{P}\left(\left|Z\right|<0.84327c\right)=0.95$$

$$\Rightarrow \mathrm{P}\left(-0.84327c<Z<0.84327c\right)=0.95$$

$$\Rightarrow \mathrm{P}(Z<-0.84327c)=0.025 \quad \text{(by symmetry)}$$

$$\Rightarrow -0.84327c=-1.95996$$

Hence $c = 2.32$ (3 s.f.)

(iii) If the scores were not normally distributed, we can still apply the Central Limit Theorem to conclude that $\overline{X}$ is approximately normally distributed since $n = 160 > 30$ is large. Hence, the calculations in (ii) would still be valid.

Perform standardisation to Z:

$$Z=\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$$

Observe the symmetry by drawing a normal distribution curve.

Annex

Proof for

(i) $\mathrm{E}(\bar{X}) = \mathrm{E}(X)$ and $\mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X)$

(ii) the sample mean $\bar{x}$ and s^2 (where $s^2 = \frac{n}{n-1} \times \text{sample variance}$) are the **unbiased estimate of the population mean** and **unbiased estimate of the population variance** respectively.

Theorem:

Let X be a random variable with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Suppose $\bar{X}$ is the sample mean of a random sample of size n. Then

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

(ii) $\mathrm{E}(S^2) = \sigma^2$, where S^2 is the unbiased estimator of the population variance.

Proof:

(i) Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, thus

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) \quad \text{(using properties of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_1) \quad \text{(since } X_i\text{'s are identically distributed)} \\
&= \frac{1}{n} \cdot n\mathrm{E}(X_1) \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \quad \text{(since } X_i\text{'s are independent - } \textit{random} \text{ sample)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \quad \text{(using properties of variance)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_1) \quad \text{(since } X_i\text{'s are identically distributed)} \\
&= \frac{1}{n^2} \cdot n\mathrm{Var}(X_1) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

(ii) Note that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Let us first consider $\mathrm{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2 \\
&= \sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2 \\
&= \sum_{i=1}^{n}\left[(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\left(\sum_{i=1}^{n}X_i-n\mu\right)+\sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2 \quad \left(\text{since } \sum_{i=1}^{n}X_i=n\bar{X}\right)
\end{aligned}$$

Since $\mathrm{E}(X_i)=\mu$, thus $\mathrm{E}(X_i-\mu)^2=\mathrm{Var}(X_i)=\sigma^2$. Using (i), since $\mathrm{E}(\bar{X})=\mu$, thus $\mathrm{E}(\bar{X}-\mu)^2=\mathrm{Var}(\bar{X})=\frac{\sigma^2}{n}$. Thus

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) &= \mathrm{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-2n\mathrm{E}(\bar{X}-\mu)^2+\mathrm{E}\left[\sum_{i=1}^{n}(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}\mathrm{E}(X_i-\mu)^2-2n\frac{\sigma^2}{n}+\sum_{i=1}^{n}\mathrm{E}(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(\sigma^2)-2\sigma^2+\sum_{i=1}^{n}\left(\frac{\sigma^2}{n}\right) \\
&= n\sigma^2-2\sigma^2+\sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Thus $\mathrm{E}(S^2)=\frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\sigma^2$.

Note that sample variance $=\frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$ and

$$s^2 = \frac{n}{n-1}\times\text{sample variance}$$

NANYANG JUNIOR COLLEGE

DEPARTMENT OF MATHEMATICS

H2 Mathematics | **Tutorial S5 Normal Distribution** | **Year 2/2020**

1. **[DIY]** Given that $X \sim \mathrm{N}(123, 45^2)$, find the probabilities:

(a) $\mathrm{P}(100 < X < 145)$ **(b)** $\mathrm{P}(X < 125)$ **(c)** $\mathrm{P}(X > 185)$

(d) $\mathrm{P}(|X - 100| < 145)$ **(e)** $\mathrm{P}(|X - 88| > 30)$

2. **[DIY]** Given that $X \sim \mathrm{N}(67, 89)$, find the value(s) of k such that

(a) $\mathrm{P}(X < k) = 0.55$ **(b)** $\mathrm{P}(X \le k) > 0.55$ **(c)** $\mathrm{P}(X < k) \le 0.55$

(d) $\mathrm{P}(X > k) = 0.777$ **(e)** $\mathrm{P}(X > k) > 0.777$ **(f)** $\mathrm{P}(60 < X < k) = 0.444$

(g) $\mathrm{P}(|X - 67| < k) = 0.444$

3. Observations are made of the speeds of cars on the Pan Island Expressway during the peak hours. It is found that, on average, 1 in 20 cars is traveling at a speed exceeding 90 km/h, and 1 in 10 is traveling at a speed less than 60 km/h. Assuming normal distribution, find the mean and the standard deviation of this distribution. A random sample of 100 cars is taken; find the probability that at most 18 cars will be traveling at a speed less than 60 km/h.

4. A manager claims that the time spent by each customer at his supermarket follows a normal distribution with mean 35 minutes and standard deviation 30 minutes. A statistician comments that the distribution $\mathrm{N}(35, 30^2)$ will not provide an adequate model.

(i) Do you agree with the statistician's comment? Give a reason to support your answer.

(ii) Suppose the time spent by each customer follows a normal distribution and its standard deviation is 10 minutes instead of 30 minutes, find the probability that the total time spent by 2 randomly chosen customers in the supermarket is more than 3 times that of another randomly chosen customer.

5. [2007/H1/I/Q12(modified)]

Men and women have masses, in kg, that are normally distributed with means and standard deviations as shown in the following table.

	Mean mass	Standard deviation
Men	75	12.5
Women	55	10.5

(i) One man and one woman are chosen at random. Find the probability that the man has mass more than 90kg and the woman has mass less than 90kg.

(ii) Two men are chosen at random. Find the probability that one of the men has mass more than 90kg and the other has mass less than 90 kg.

(iii) One man and one woman are chosen at random. Find the probability that the woman's mass is greater than the man's.

(iv) Find the probability that the average weight of five men exceeds 80kg. State any assumption(s) you have used in your calculation.

The safety limit for a hotel elevator is 530kg.

(v) Six men are chosen at random. Find the probability that their total mass is greater than 530kg.

(vi) Six male hotel guests enter the elevator, at a time when a large number of sumo wrestlers are staying at the hotel. Give two reasons why the probability that their total mass exceeds 530kg may be different from the value calculated in part **(v)**.

6. It is given that $X \sim \mathrm{N}(\mu, \sigma^2)$ and $\mathrm{P}(X < 3) = \mathrm{P}(X > 7)$. Write down the value of μ. It is also given that $2\mathrm{P}(X < 2) = \mathrm{P}(X < 8)$.

(i) Find σ^2.

(ii) Three observations of X are taken. Determine the probability that one will be more than 7 and the other two will be between 3 and 6 inclusive.

The random variable Y is the sum of n independent observations of X.

(iii) State the approximate value of $\mathrm{P}(Y > 3.5n)$ as n becomes very large, justifying your answer.

7. **[1989/II/Q9 (modified)]**

The random variable Y has a normal distribution with mean 10 and variance 4.

Find the possible value(s) of a if

(i) $\text{P}(10-a<Y<10+a)=0.95$.

(ii) the probability that Y is between a and 12 is 0.15.

Two independent observations of Y are denoted by Y_1 and Y_2. Find

(iii) $\text{P}(Y_1 + Y_2 > 18)$

(iv) probability that the difference between Y_1 and Y_2 is less than 5.

Three values of Y are taken at random. Calculate the probability that

(v) their sum does not exceed 40.

(vi) the largest value does not exceed 14.

8 The Body Mass Index (BMI) is a measure of body fat based on height and weight that applies to men and women. Typically, the BMI of eighteen year old well-built male and female students are modelled as having normal distributions with means and standard deviations as shown in the table.

Gender	Mean BMI	Standard deviation
Male	μ	σ
Female	25	1.2

(i) Let M denote the BMI of an eighteen year old well-built male student. Given that $\text{P}(M>25)=\text{P}(M<31)=0.97725$, state the value of μ and show that $\sigma=1.5$.

(ii) Find the probability that the mean BMI of three randomly chosen male and five randomly chosen female eighteen year old well-built students is more than 26.

(iii) Four eighteen-year-old well-built female students are randomly chosen. Find the probability that exactly two of them have BMI less than 26.

(iv) State, with a reason, whether or not a normal model is likely to be appropriate for the BMI reading of a combined group of male and female students in a school.

9. A telecommunication company finds that the duration of calls made by its customers to City A are normally distributed with mean 8 minutes and standard deviation 1.5 minutes. The duration of calls made by its customers to City B are normally distributed with mean 10 minutes and standard deviation 1.8 minutes. Assume that the duration of the calls to City A and City B are independent.

(i) Find the probability that the total duration of three randomly selected calls to City A differs from the total duration of two randomly selected calls to City B by more than three minutes.

The company charges a rate of 22 cents per minute for every call to City A and a rate of 30 cents per minute for every call to City B.

(ii) Find the probability that the cost of one call made to City A is at least \$1.50 and the cost of one call made to City B is at least \$3.00.

(iii) Find the probability that the total cost of one call made to City A and one call made to City B is at least \$4.50.

(iv) Explain why the answer to **(iii)** is greater than the answer to **(ii)**.

(v) Three calls were made to City A. Find the probability that exactly one call costs more than \$1.80 and exactly one call costs less than \$1.70.

(vi) John is planning to set the budget for overseas calls for the week at \$$a$. Based on his record, a total of 12 calls are made to City A and 8 calls to City B in any week. Find the least integer value of a if he wants to be at least 99% sure that it will not exceed his budget.

10. [ACJC/2016/II/6]

A factory manufactures rectangular glass panels. The length and breadth of each panel, in cm, are modelled as having independent normal distributions with means and standard deviations as shown in the table.

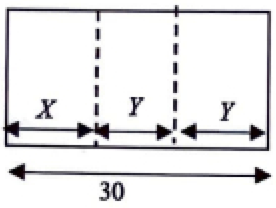
Glass Panel	Mean (cm)	Standard Deviation (cm)
Length	300	0.5
Breadth	150	0.2

The probability that the total perimeter of 2 randomly selected glass panels exceeds the mean length of n randomly selected glass panels by more than 1501cm is less than 0.2576. Find the least value of n.

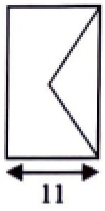
11. [AJC/2016/II/9]

A secretary types letters onto sheets of paper 30 cm long and folds the letters as shown.

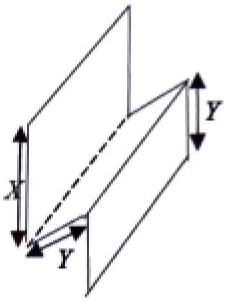
Unfolded sheet



Envelope



Folding of sheet



The first fold is X cm from one edge. The second fold, Y cm from the other edge, is exactly in the middle of the remaining part of the paper, so that $Y = \frac{1}{2}(30 - X)$.

The length X cm is normally distributed with mean 10.2 cm and standard deviation 1.2 cm. The letters have to fit into envelopes 11 cm wide.

(i) Find $\mathrm{P}(11 < Y < 15)$.

(ii) Find the probability that a randomly chosen letter will fit into the envelope.

(iii) By expressing $X - Y$ in terms of X, verify that

$$\mathrm{Var}(X - Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y) \ .$$

Explain why the rule $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$ does not apply in this case.

12. Every morning, Peter has to take a bus to school. On average, he reaches the bus stop at 6:30 am each day. His arrival time at the bus stop is normally distributed with a standard deviation of 10 minutes. The first bus will leave the bus stop at 6:40 am sharp and the second bus will leave at 6:50 am sharp. Regardless of the time the buses leave the bus stop, the bus journey will follow a normal distribution with mean 45 minutes and standard deviation 20 minutes. Peter will be late if he arrives at school after 7:30 am or if he misses the 6:50 am bus.

Assuming that the time Peter reaches the bus stop and the time taken for the bus journey are independent, find the probability that

(i) Peter will miss the 6:50 am bus.

(ii) Show the probability that Peter will be late for school is 0.442.

Assignment:

1. **[DHS/2016/II/9(a) (modified)]**

In this question you should state clearly all distributions that you use, together with the values of the appropriate parameters.

The queuing time, in minutes, for flight passengers at the Economy and Business class check-in counters have independent normal distributions with means and standard deviations as shown in the table.

Check-in Counter	Mean queuing time	Standard deviation
Economy class	11.6	4.2
Business class	3.2	0.9

(i) Find the probability that the queueing time of a randomly chosen Economy class passenger is within 5 minutes of the total queueing time of 2 randomly chosen Business class passengers.

(ii) The queueing time of 8 randomly chosen Business class passengers are taken. Find the probability that

(a) the shortest queuing time among all 8 passengers is no less than 2 minutes and the longest queuing time is no more than 5 minutes,

(b) at least half of the passengers will have queuing time of more than 4 minutes.

2. **[RI/2016/II/8]**

(a) S and W are independent random variables with the distributions $\mathrm{N}(20, 25)$ and $\mathrm{N}(\mu, \sigma^2)$ respectively. It is known that $\mathrm{P}(W < 10) = \mathrm{P}(W > 13)$ and $\mathrm{P}(S > 2W) = 0.43$. Calculate the values of μ and σ correct to three significant figures.

(b) A small hair salon has two hairstylists Joe and Joan attending to customers wanting an express haircut. For Joe, the time taken to attend to a customer follows a normal distribution with mean 10 minutes and standard deviation 42 seconds. For Joan, the time taken to attend to a customer follows a normal distribution with mean 10.2 minutes and standard deviation 45 seconds.

(i) Find the probability that among three randomly chosen customers attended to by Joe, one took more than 10.5 minutes while the other two each took less than 10 minutes.

(ii) Joe and Joan each attended to two customers. Find the probability that the difference in the total time taken by Joe and Joan to attend to their two customers respectively is more than 3 minutes. State any assumption(s) that you have used in your calculation.

Tutorial Questions Answers

1	(a) 0.383 (b) 0.518 (c) 0.0841 (d) 0.997 (e) 0.619
2	(a) $k = 68.2$ (b) $k > 68.2$ (c) $k \leq 68.1$ (d) $k = 59.8$ (e) $k < 59.8$ (f) $k = 71.2$ (g) $k = 5.56$
3	73.1, 10.3, 0.995
4	0.146
5	(i) 0.115 (ii) 0.204 (iii) 0.110 (iv) 0.186 (v) 0.00449
6	$\mu = 5$, (i) $\sigma^2 = 48.5$ (ii) 0.0336 (iii) 1
7	(i) $a = 3.92$ (ii) $a = 11.0$ or 14.8 (iii) 0.760 (iv) 0.923 (v) 0.998 (vi) 0.933
8	(i) 28 (ii) 0.606 (iii) 0.156
9	(i) 0.635 (ii) 0.392 (iii) 0.659 (v) 0.140 (vi) 50
10	7
11	(i) 0.0334 (ii) 0.714
12	0.0228

Assignments Answers

1	(i) 0.472 (ii) (a) 0.380 (b) 0.0451
2	(a) $\mu = 11.5$, $\sigma = 8.13$ (b) (i) 0.178 (ii) 0.0461

Extra Practice Questions

1 **ACJC Prelim 9758/2018/02/Q9**

A shop sells two brands of car battery, A and B. The battery life, in months, of each brand of car battery have independent normal distributions. The means and standard deviations of these distributions, are shown in the following table.

	Mean battery life	Standard deviatio n
Brand A	30	8
Brand B	25	σ

(i) Explain why the battery life of a randomly chosen Brand B battery cannot be normally distributed if $\sigma = 22$. [1]

(ii) The probability that the battery life of a randomly chosen Brand B battery is within 5 months of the mean battery life of Brand B batteries is 0.8. Find the variance of the distribution of Brand B battery life. [3]

Use $\sigma = 4$ for the rest of the question.

(iii) Sketch the distributions of the battery life of the Brand A battery and the Brand B battery on a single diagram. [1]

(iv) Find the probability that the battery life of a randomly chosen Brand A car battery is exactly 26 months. [1]

(v) Find the probability that the difference between the mean battery life of 3 randomly chosen Brand B batteries and 75% of the battery life of a randomly chosen Brand A battery is less than 3 months. State the parameters of any distribution you use. [4]

The manufacturer of Brand B battery replaces for free all batteries that fail within the warranty period of k months. If they are willing to replace for free less than 1% of all batteries sold, find the longest warranty period (to the nearest integer), that the manufacturer can offer. [2]

[(ii) 15.2, (iv) 0, (v) $\bar{X}-0.75A \sim N(2.5,\frac{124}{3})$, 0.335, (iv) 15]

2 AJC Prelim 9758/2018/02/Q7

Soya bean drink is sold in cups of two sizes – small and large. For each size, the amount of content, in ml, of a randomly chosen cup is normally distributed with mean and variance as given in the table. The selling prices are also given in the table.

	Mean (ml)	Variance (ml^2)	Selling Price ($)
Small	202	21	1.10
Large	405	74	2.20

The amount of content in any cup may be assumed to be independent of the amount of content in any other cup.

(i) A small cup and two large cups are selected at random. Find the probability that the total amount of content in the two large cups is less than four times the amount of content in the small cup. [2]

(ii) A boy needs at least 600 ml of soya bean drink but he only has $3.80. In what way should he make his purchase so that he has the highest probability of getting at least 600 ml? Support your answer with clear workings. [3]

(iii) A random sample of 20 small cups and n large cups of soya bean drink is taken. Find the least value of n such that there is a probability of more than 0.8 that the mean amount of content in these cups is more than 350 ml. [4]

[(i) $\text{P}(L_1 + L_2 - 4S < 0) = 0.464$ (ii) He should buy 3 small cups (iii) Least value of $n = 55$]

3 DHS Prelim 9758/2018/02/Q8

An interactive simulation ride allows a group of 5 riders to take the ride at a time. The ride time, X minutes, follows a normal distribution with mean μ minutes and standard deviation 2 minutes. The ride starts promptly at 10 am daily with no wait time between any groups of 5 riders. There are only 4 scheduled rides every morning. At 10 am on a particular morning, there are already 20 people queuing for the ride. It is assumed that all the people in the queue will take the ride based on the sequence of the queue and the ride times are independent.

(i) Show that $\mu = 14$, correct to the nearest integer, if $\text{P}(\mu < X < 16) = 0.35$. [2]

For the rest of the question, use $\mu = 14$ for your calculations. A ride is considered long if it has a ride time of at least 15 minutes.

(ii) Find the probability that the 12th person in the queue took the ride before 10.30 am on that morning. [3]

(iii) Show that the probability of having at least 2 long rides on that morning is 0.363. [2]

(iv) Given that there are at least 2 long rides on that morning, find the probability that none of these long rides are consecutive. [2]

[(ii) 0.760 (3 s.f.) (iv) 0.376 (3 s.f.)]

4 **EJC Prelim 9758/2018/02/Q10**

[In this question, you should state clearly the distribution of any random variables that you define.]

The volume, S, in ml, of perfume in a randomly chosen small bottle has mean 20 and variance σ^2.

(i) If $\sigma = 15$, explain why S may not be appropriately modelled by a normal distribution. [2]

It is now assumed that S follows a normal distribution

(ii) Given that 6.68% of the small bottles contains more than 23 ml of perfume, find the value of σ. [2]

For the rest of the question, the volume of perfume, S, in ml, in a randomly chosen small bottle follows the distribution $\mathrm{N}(20,4)$ and the volume of perfume, L, in ml, in a randomly chosen large bottle follows the distribution $\mathrm{N}(100,25)$.

(iii) Calculate the probability that 6 randomly chosen small bottles and 9 randomly chosen large bottles contain a total volume of at least 1 litre of perfume. [3]

(iv) Calculate the probability that the volume of perfume in a randomly chosen large bottle differs from 6 times the volume of perfume in a randomly chosen small bottle by more than 25ml. [3]

(v) State, in this context, an assumption needed for your calculations in parts **(iii)** and **(iv)**. [1]

$[\sigma = 2.00 \text{ (iii) } 0.898 \text{ (iv) } 0.351]$

5 **HCI Prelim 9758/2018/02/Q10**

Each morning, Tony drives from his home to his office and has to pass 4 traffic lights on his way. He has to reach his office by 8.50 am. The driving time to his office and the time held up at a traffic light junction, in minutes, may be assumed to follow normal distributions with means and standard deviations as summarized below:

	M	Standard deviation
Driving time	14	2.1
Time held up at a traffic light junction	μ	0.2

Tony leaves home at 8.30 am. If the probability that Tony is not late is 0.713, show that $\mu = 1.2$, correct to 1 decimal place. State an assumption needed in your calculations. [4]

(i) For 10 mornings, Tony leaves home at 8.30 am. Find the probability that he arrives late at his office for the third time on the 10th day. [2]

(ii) Find the probability that Tony's driving time to his office is less than 10 times the time he is held up at a traffic light junction. [2]

(iii) Find the probability that Tony's driving time to his office and the total time he is held up at the 4 traffic light junctions differs by more than 8 minutes. [3]

Assume μ is unknown.

(iv) The time held up at a traffic light junction is recorded on n randomly chosen occasions. Find the smallest n so that it is at least 98% certain that the sample mean time Tony is held up at a traffic light junction is within 5 seconds of μ. [3]

[(i) 0.0797; (ii) 0.245; (iii) 0.713 ; (iv) Smallest $n = 32$]

6 JJC Prelim 9758/2018/02/Q9

Durians and melons are sold by weight. The masses, in kg, of durians and melons are modelled as having independent normal distributions with means and standard deviations as shown in the table.

	Mean Mass	Standard Deviation
Durians	2.1	0.25
Melons	0.6	0.16

Durians are sold at \$15 per kg and melons at \$6 per kg.

(i) Find the probability that the mass of a randomly chosen durian is less than four times the mass of a randomly chosen melon. [3]

(ii) Two durians and eight melons are randomly selected. Find the probability that the average mass of these ten fruits exceeds 1 kg. [4]

(iii) Find the probability that the total selling price of a randomly chosen durian and a randomly chosen melon is less than \$40. [4]

(iv) Without any further calculation, explain why the probability of the event that both a randomly chosen durian has a selling price less than \$35 and a randomly chosen melon has a selling price less than \$5 is less than the answer to part **(iii)**. [1]

[(i) 0.669, (ii) 0.0408, (iii) 0.897]

7 MJC Prelim 9758/2018/02/Q11

(a) The time T (in minutes) taken by Kathy to drive from her house to the nearest supermarket has a mean of 5 and a variance of 8.

Explain why T is unlikely to be normally distributed. [2]

(b) In this question you should state clearly the values of the parameters of any normal distribution you use.

In a supermarket, the masses in kilograms of chickens have the distribution $\mathrm{N}(2.4, 0.5^2)$ and the masses in kilograms of ducks have the distribution $\mathrm{N}(4.3, 1.8^2)$.

(i) Find the probability that the total mass of 2 randomly chosen chickens is more than 5kg.[2]

(ii) Find the probability that the mean mass of a randomly chosen duck and 2 randomly chosen chickens is more than 3.20kg. [3]

Chickens are sold at \$7 per kilogram and ducks are sold at \$13 per kilogram. The supermarket is conducting their annual sale and has a discount of 10% and 25% off the prices of chickens and ducks respectively.

(iii) Find the probability that the total cost of 2 randomly chosen chickens and a randomly chosen duck is less than \$80. [4]

[(b)(i) 0.389, (b)(ii) 0.398, (b)(iii) 0.667]

8 NYJC Prelim 9758/2018/02/Q8

The masses of manufactured links of a chain are normally distributed with mean 800 grams and standard deviation 20 grams.

(i) Find the probability that the mass of a randomly chosen link is more than 805 grams. [1]

To close the gate opening in a link, a locking sleeve is attached to it and it increases the mass of a link by 10%.

(ii) By writing down the distribution of the masses of links with locking sleeves, find the probability that the mass of a randomly chosen link with a locking sleeve is between 865.35 and 895.5 grams. [3]

Hooks with mean masses 750 grams are manufactured to attach to the links. The masses of hooks are normally distributed such that 15% of them have mass less than 735.6 grams.

(iii) Find the standard deviation of the masses of hooks. [2]

Five independent links with locking sleeves and a hook are packed in a wooden box with a fixed mass of 1 kilogram. The probability that the mean mass of n such wooden boxes with its contents more than 6190 grams exceeds 0.013.

(iv) Find the greatest value of n, stating the parameters of any distribution that you use. [4]

[(i) 0.401, (ii) 0.507, (iii) 13.9, (iv) 8]

9 RI Prelim 9758/2018/02/Q8

In this question you should state clearly the values of the parameters of any normal distribution you use.

The masses in grams of apples have the distribution $N(80, 3^2)$ and the masses in grams of oranges have the distribution $N(100, 5^2)$.

(i) Find the probability that the total mass of 2 randomly selected oranges is less than 205 grams.[2]

(ii) Find the probability that the total mass of 4 randomly selected apples is more than three times the mass of 1 randomly selected orange. [3]

During the Community Health Week Event, visitors who take part in the activities are given gift boxes. Each gift box contains 3 apples and 2 oranges that are individually machine wrapped. The mass of wrapper used for each fruit is dependent on the mass of the fruit resulting in the mass of each wrapped fruit being 7% more than the mass of the fruit. The fruits are packed in a gift box and the mass of an empty gift box is normally distributed with mean 50 grams and standard deviation 2 grams. During packing, parts of the gift box is cut and removed. This process reduces the mass of each empty gift box by 10%.

(iii) The probability that the total mass of a randomly selected gift box is less than k grams is 0.8. Find the value of k.

Find the probability that, at a particular collection point, the 25th gift box that is given is the 19th gift box whose total mass is less than k grams. [5]

[(i) 0.760 (ii) 0.892 (iii) $k = 524$, 0.124]

10 RVHS Prelim 9758/2018/02/Q10

Gary likes to take part in duathlons where there is a total running distance of 15 km and a total cycling distance of 36 km. His timings, in minutes, for running and cycling are modelled as having independent normal distributions with means and standard deviations as shown in the table.

	Mean	Standard Deviation
Timing for running (min)	100	15
Timing for cycling (min)	μ	σ

(i) The probability that he completes his cycling in less than an hour is equal to the probability that he completes his cycling after 2 hours. Write down the value of μ. [1]

(ii) The probability that he completes his cycling under 80 minutes is 0.158655. Show that $\sigma = 10$. [2]

(iii) Find the probability that thrice the time needed for him to complete cycling is more than an hour from twice the time needed for him to complete the running. [2]

(iv) Find the probability that his timing for cycling is less than 80 minutes and his timing for running is less than 90 minutes. [2]

(v) Show that the probability of Gary finishing the duathlon under 170 minutes is 0.134. Give a reason why this probability is greater than the probability calculated in part **(iv)**. [2]

Gary's target timing for completing a duathlon is under 2 hours 50 minutes. Over the past two years, he has already completed 9 duathlons.

(vi) Find the probability that he achieves his target timing in at least 3 but fewer than 6 of the duathlons over the last two years. State an assumption needed for your calculation to be valid.[4]

[(i) $\mu = 90$, (iii) 0.593, (iv) 0.0401, (vi) 0.108]

11 SRJC Prelim 9758/2018/02/Q8

The length of time that Somesong phones last in between charges has a normal distribution with mean 20 hours and standard deviation 2 hours. The length of time that Apfel phones last in between charges has a normal distribution with mean μ hours and standard deviation σ hours.

(a) The average length of time two randomly chosen Somesong phones and one randomly chosen Apfel phone will last in between charges is equally likely to be less than 19 hours or more than 23 hours, with the probability known to be 0.02275.

(i) Calculate the values of μ and σ [3]

(ii) Find the probability that twice the length of time a Somesong phone lasts in between charges differs from 40 hours by at least 2 hours. [2]

(b) Six randomly chosen Somesong phones are examined. Find the probability that the sixth Somesong phone is the fourth Somesong phone that lasts more than 22 hours in between charges. [2]

[(a)(i) $\mu = 23$, $\sigma = 1$, (a)(ii) 0.617, (b) 0.00449]

12 TJC Prelim 9758/2018/02/Q7

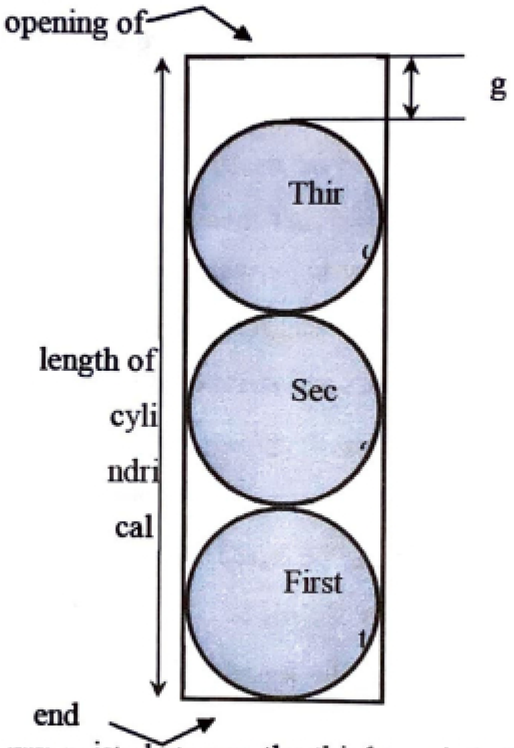
A company manufactures tennis balls with radii that are normally distributed with mean 3.3 cm and standard deviation 0.2 cm.

(i) Find the probability that the radius of a randomly selected tennis ball lies between 3.135 cm and 3.465 cm.

Without any further calculation, explain with the aid of a diagram how the answer obtained would compare with the probability that the radius lies between 3.465 cm and 3.795 cm. [3]

(ii) 3 tennis balls are randomly selected. Find the probability that the largest tennis ball has radius less than 3.4 cm. [2]

The tennis balls are packed into cylindrical tubes for sale. The cylindrical tubes have lengths that are normally distributed with mean 20 cm and standard deviation 0.3 cm. 3 tennis balls are randomly selected and packed into a randomly selected cylindrical tube such that the first tennis ball is in contact with the end of the tube and each subsequent ball is in contact with its neighbouring ball as shown in the diagram below.



(iii) The probability that a gap exists between the third tennis ball and the opening of the tube and that the gap is at least k cm long is at least 0.15. Assuming that the centres of all the tennis balls are vertically aligned, find the possible range of values of k. [4]

[(i) 0.591, (ii) 0.331, (iii) $0 < k \leq 0.982$]

13 **VJC Prelim 9758/2018/02/Q9**

AppleC is a fruit farm that produces Cameo apples. It is known that 8.1% of the apples have a mass more than 125 g and 14.7% of the apples have a mass less than 90 g. It is also known that the masses of this batch of Cameo apples follows a normal distribution.

(i) Show that the mean and standard deviation of the masses for this batch of Cameo apples, correct to 3 significant figures, are 105 g and 14.3 g respectively. [4]

The Cameo apples are packed into bags of 8 apples.

(ii) Find the probability that the mass of a randomly chosen bag of Cameo apples is less than 845 g. State the distribution used and its parameters. [3]

A local distributor imports Cameo apples from AppleC. The distributor also imports Jazz apples from another supplier. It is known that the masses of bags of Jazz apples follow a normal distribution with mean 700 g with standard deviation of 35 g. The retail price of Cameo apples and Jazz apples are $8.50 per kg and $7.30 per kg respectively.

(iii) Find the probability that the cost of three bags of Jazz apples is more than twice the cost of a bag of Cameo apples by at least $1. [4]

[(ii) 0.549, $\mathrm{P}(Y > 1) = 0.524$]

NANYANG JUNIOR COLLEGE

DEPARTMENT OF MATHEMATICS

H2 Mathematics **Tutorial S6 Sampling and Estimation** **Year 2/2020**

1 **(DIY)** Find unbiased estimates of the population mean and variance from which each of the following samples is drawn:

(i) 35, 42, 38, 55, 70, 69

(ii) $n = 34, \sum x = 330, \sum x^2 = 23700$

(iii) $n = 8, \sum x = 120, \sum (x - \bar{x})^2 = 302$

(iv) $n = 250, \sum (x - 76) = 683, \sum (x - 76)^2 = 26132$

(v) $n = 40$, $\sum (x - 40) = -13$ and $\sum x^2 = 87532$,

(vi) The observed values are

x	1	2	3	5	7	6	10	8
Frequency, f	1	2	4	3	2	1	6	5

[(i) 51.5, 241 (ii) 9.71, 621 (iii) 15, 43.1 (iv) 78.7, 97.5 (v) 39.68, 629.94 (vi) 6.33, 9.19]

2 **N2019/II/6**

In a certain country there are 100 professional football clubs, arrange 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.

(ii) Dilip wishes to investigate the facilities for supports 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.

3 An automatic machine is used to fill bottles with liquid. It is found in practice, that there are slight variations from bottle to bottle, and that the mass of liquid in a bottle is a random variable having a normal distribution. A sample of ten bottles gave the following excess masses over 100g (measured in milligrammes) :

13 28 12 -2 27 8 5 20 18 15

Find the unbiased estimate of the mean and variance of the mass of liquid in a bottle. Explain the meaning of the term 'unbiased estimate' in the context of the question.
Find the probability that the mean mass of liquid in a random sample of twenty selected bottle exceeds 100.02g.

[100.0144, 0.8827×10^{-4} , 0.00384]

4 A random sample of size 100 is taken from the distribution B(50,0.6). Find the probability that the mean is (i) less than 30.2, and (ii) greater than 31.

[(i) 0.718, (ii) 0.00195]

5 **2011/RVHS/II/7 (modified)**

A factory produces packets of peanuts. The mass of peanuts in a packet has mean 605g and standard deviation 6g. A sample of sixty packets is chosen. Find the probability that the mean mass of peanuts in a packet from this sample is between 600g and 606g. State the assumptions that you have made. Find the probability that the total mass of the 60 packets exceeds 36.2kg.

[0.902, 0.984]

6 Let $X_1, X_2, X_3, \ldots, X_n$ be n independent observations drawn from a normal distribution with mean μ and standard deviation σ. State the distribution of the sample mean $\overline{X}$. Is it necessary to apply the Central Limit Theorem in this case? Why?

Given that $\sigma = 2$ and the probability that $\overline{X}$ is within ± 0.1 of the population mean μ is at least 0.95, find the least value of the sample size required. [1537]

7 **2017/NJC/II/7**

There are three identically shaped balls, numbered from 1 to 3, in a bag. Balls are drawn one by one at random and with replacement. The random variable X is the number of draws needed for any ball to be drawn a second time. The two draws of the same ball do not need to be consecutive.

(i) Show that $\mathrm{P}(X=4)=\frac{2}{9}$ and find the probability distribution of X. [3]

(ii) Show that $\mathrm{E}(X)=\frac{26}{9}$ and find the exact value of $\mathrm{Var}(X)$. [3]

(iii) The mean for forty-four independent observations of X is denoted by $\overline{X}$. Using a suitable approximation, find the probability that $\overline{X}$ exceeds 3. [3]

[44/81, 0.159]

8 **2016/MJC/II/8**

The mass of a randomly chosen bar of body soap manufactured by a factory has a normal distribution with mean 110 grams and standard deviation 1.5 grams.

(i) Find the probability that the difference in sample means between any two random samples of 20 bars of body soap each, is within 0.5 grams. [4]

(ii) Five randomly chosen bars of body soap are liquefied and separated into four equal portions, which are each placed into a bottle. Find the probability that the mass of liquid body soap in a randomly chosen bottle exceeds 140 grams. [4]

[0.708, 0.00143]

9 **N84/II/12**

The random variables X and Y are related to two biased coins A and B in the following way. If when A is tossed and a head appears, which it does with probability p, then $X = 1$, and if a tail appears, then $X = 0$. The random variable Y is defined in the same way for coin B, for which the probability of a head appearing is $2p$, where $0<p<\frac{1}{2}$. For all values of λ, it is given that the random variable T, defined by $T=\lambda X+\frac{1}{2}(1-\lambda)Y$ is an unbiased estimator of p.

(i) Show that $\mathrm{Var}(T)=\lambda^2 p(1-p)+\frac{1}{2}(1-\lambda)^2 p(1-2p)$.

(ii) Show that, when $p=\frac{1}{3}$, the minimum value of Var(T) is obtained by taking $\lambda=\frac{1}{5}$.

(iii) Hence obtain, for $p=\frac{1}{3}$, the least possible value of $\mathrm{P}\left(\left|\overline{T}-\frac{1}{3}\right|>0.05\right)$, where $\overline{T}$ is the random variable denoting the mean value of T obtained from 100 tosses of the two coins.

[0.0177]

Assignment

1. At an early stage in analysing the marks scored by the large number of candidates in an examination, the Examination Board takes a random sample of 250 candidates and finds that the marks, x, of these candidates give $\sum x = 11872$ and $\sum x^2 = 646{,}193$. Find the unbiased estimates for the population mean and variance. Hence, find the probability that, on average, the candidates score between 45 to 55 marks.

2 **2003/II/13**

A random variable X has the distribution 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.

3 **2018/Prelim/ACJC/II//Q6**

On average, 40% of Christmas tree light bulbs manufactured by a company are red and the rest are blue. The lights are sold in boxes of 20. Assume that the number of red light bulbs in a box has a binomial distribution.

(i) A random sample of 50 boxes is chosen. Using an approximate distribution, find the probability that the mean number of red light bulbs in 1 box will not exceed 7.5.

(ii) A customer selects boxes of light bulbs at random from a large consignment until she finds a box with fewer red lights than blue. Give a reason why a binomial distribution is not an appropriate model for the number of boxes selected.

Extra Practice Questions

The questions will not be discussed during tutorials. Full solutions will be uploaded to portal.

1 **2018/Prelim/HCI/II/Q10**

Each morning, Tony drives from his home to his office and has to pass 4 traffic lights on his way. He has to reach his office by 8.50 am. The driving time to his office and the time held up at a traffic light junction, in minutes, may be assumed to follow normal distributions with means and standard deviations as summarized below:

	Mean	Standard deviation
Driving time	14	2.1
Time held up at a traffic light junction	μ	0.2

Tony leaves home at 8.30 am. If the probability that Tony is not late is 0.713, show that $\mu = 1.2$, correct to 1 decimal place. State an assumption needed in your calculations. [4]

Assume μ is unknown.

The time held up at a traffic light junction is recorded on n randomly chosen occasions. Find the smallest n so that it is at least 98% certain that the sample mean time Tony is held up at a traffic light junction is within 5 seconds of μ. [3]

2 **2018/Prelim/IJC/II/Q7**

The discrete random variable X takes values 0, 1, 2 and 3 only. The probability distribution of X is shown in the table, where p is a constant and $0 < p < \frac{1}{10}$.

x	0	1	2	3
$\mathrm{P}(X = x)$	$1-6p$	$3p$	$2p$	p

(i) Given that $\mathrm{Var}(X) = 0.75$, find the value of $\mathrm{E}(X)$. [3]

(ii) The random variable S is the sum of n independent observations of X, where n is large. Given that the probability that S exceeds 150 is at least 0.75, find the set of possible values of n. [3]

3 **2018/Prelim/NYJC/II/Q8**

The masses of manufactured links of a chain are normally distributed with mean 800 grams and standard deviation 20 grams.

(i) Find the probability that the mass of a randomly chosen link is more than 805 grams. [1]

To close the gate opening in a link, a locking sleeve is attached to it and it increases the mass of a link by 10%.

9 2015/Prelim/NJC/II/Q9

The weight, in grams, of a handphone is a random variable with the distribution $\mathrm{N}\left(\mu, 5^2\right)$, where μ is a constant. The weight, in grams, of a tablet is a random variable with the independent distribution $\mathrm{N}\left(330, 6^2\right)$.

(a) Suppose $\mu = 130$. Calculate the probability that the total weight of 5 randomly selected handphones is less than twice the weight of one randomly selected tablet. **[3]**

(b) Suppose instead that μ is unknown. A random sample of n handphones is taken. There is a probability of at most 0.04 that the mean weight of these handphones differs from the population mean weight of the handphones by more than 1 gram. Calculate the least value of n. **[4]**

10 2014/Prelim/MJC/II/Q10

The masses, in kilograms, for a randomly chosen pumpkin and a randomly chosen watermelon, are normally distributed with means 4.2 and 8.5, and standard deviations 1.6 and 2.2 respectively.

(i) Find the range of values of k such that the average mass of five randomly chosen watermelons exceeds k kilograms occurs at most 10% of the time. **[3]**

(ii) Find the probability that the total mass of five randomly chosen pumpkins differs from three times the average mass of five randomly chosen watermelons by more than 5 kg. **[4]**

(iii) A fruit store has 70 batches, each consisting of n pumpkins. It is known that, on average, 3 in 20 pumpkins are rotten. Given that in the 70 batches, the probability of having an average of at least 4 rotten pumpkins per batch is more than 0.7. Determine the minimum value of n. **[4]**

Answer

1	smallest $n = 32$						
2 (i)	$\mathrm{E}(X) = \dfrac{1}{2}$	**(ii)**	$\{n : n \in \mathbb{Z}^+,\ n \geq 321\}$				
3(i)	0.401	**(ii)**	0.507	**(iii)**	13.9	**(iv)**	8
4(i)	0.982	**(ii)**	smallest $n = 43$				
5(i)	0.32	**(ii)**	0.360				
6(ii)	32.7	**(iii)**	0.958	**(iv)**	1.50		
7(i)	0.0359	**(iii)**	6				
8(i)	0.0000514	**(ii)**	0.998				
9(a)	0.729	**(b)**	106				
10(i)	$k \geq 9.76$	**(ii)**	0.477	**(iii)**	28		