RAFFLES INSTITUTION
H2 Mathematics 9758
2023 Year 6 Term 3 Revision 15 (Summary and Tutorial)

Topic: Sampling and Estimations, Hypothesis Testing, Correlation&Regression

---

Summary for Sampling and Estimations

Definitions - A **population** is the entire collection of data (persons or items or individuals) that we want to study e.g. apples produced by a farm.

- A sample is **random** if every element in the population has *an equal chance* of being selected, and the selection of an element is independent of another. e.g. 'Every biscuit bar has an equal chance of being selected, and the selection of one biscuit bar is not affected or influenced by the selection of another biscuit bar'.

  [Note that it is **not** sufficient to say 'each biscuit bar has an equal chance of being selected']

- A sample is **non-random** if each element in the population *does not have an equal chance of being selected*, resulting in certain segments of the population being over-represented, as some members are "systematically or deliberately excluded" from the study and the sample being biased.

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

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

The Distribution of the Sample Mean

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	Let $X_1, X_2, X_3, \ldots, X_n$ be a **large** random sample of size *n* taken from a **non-normal** population with mean μ and variance σ^2. Then since **sample size *n* is large**,
(1) $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$	**(1)** $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ ***approximately*** by **Central Limit Theorem**
(2) $X_1 + X_2 + X_3 + \ldots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$	**(2)** $X_1 + X_2 + X_3 + \ldots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$ ***approximately*** by **Central Limit Theorem**

General Tips

1. To identify questions on Central Limit Theorem, look out when there is no mention of 'normally distributed' or 'normal population' e.g. "The random variable X is thought to have a mean of 50 but it is known that the standard deviation is 14.5." and the question asks for 'find the probability that the **sample mean / average value of *X* / sum** …' , 'by using a **suitable approximation**' or '**estimate** the probability'.
2. Do not make the mistake of writing $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ at the first sighting of mean and variance/standard deviation in the question. Mean and variance applies to any population, not just normal, so look out for the phrases in point 1 above.
3. Also, writing $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ approximately by Central Limit Theorem is **wrong** as the theorem applies to $\bar{X}$.
4. When the population variance σ^2 is unknown, the unbiased estimate of population variance s^2 will be used and the notation will change accordingly. This is especially important when we write the distribution of $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\cancel{\sigma^2}\, s^2}{n}\right)$ approximately under hypothesis testing.

Unbiased Estimates of Population Mean and Population Variance.

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.

In particular, the average value of $\bar{X}$ gives the true value of μ, that is, $\mathrm{E}\left(\bar{X}\right) = \mu$.

Population	Estimate from sample	Unbiased estimate?	How?
μ	Sample mean $\bar{x}$	Yes, sample mean is an unbiased estimate of population mean.	$\bar{x} = \dfrac{\sum x}{n}$ or $\bar{x} = \dfrac{\sum (x-a)}{n} + a$
σ^2	Sample variance	No, sample variance is NOT an unbiased estimate of population variance.	We will have to calculate the unbiased estimate of population variance s^2 by using one of the following: (1) $s^2 = \dfrac{n}{n-1}\left[\dfrac{\sum\left(x-\bar{x}\right)^2}{n}\right]$ (2) $s^2 = \dfrac{1}{n-1}\left[\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right]$ (3) $s^2 = \dfrac{n}{n-1}\left(\text{sample variance}\right)$ (4) $s^2 = \dfrac{1}{n-1}\left[\sum (x-a)^2 - \dfrac{\left(\sum (x-a)\right)^2}{n}\right]$

Summary for Hypothesis Testing

Performing a Hypothesis Test

Step 1	Understand the given question and write down the **null hypothesis** H_0 and the **alternative hypothesis** H_1
Step 2	Write down the **level of significance** α (usually given in the question)
Step 3	Decide on the **test statistic** to be used and determine its **distribution**
Step 4	Use the GC to calculate the ***p*-value**
Step 5	**Reject** H_0 if p-value $\leq \alpha$, OR **Do not reject** H_0 if p-value $> \alpha$ Write down the **conclusion** in the context of the question

Example:

Let X be the IQ of a student in ABC University.

Remember to always define X if they are not defined in the question.

Step 1: To test $H_0 : \mu = 118$ vs $H_1 : \mu > 118$

This is an example. Read question carefully to decide >, < or ≠.

Step 2: Perform a 1-tail / 2-tail test at 5% level of significance.

Step 3: (Sample from a Normal population of known variance)

Under H_0, $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, where $\mu_0 = 118$ and $\sigma = 12$.

From the sample, $\bar{x} = 121$, $n = 50$.

OR

Step 3 (Large sample from a Normal population of unknown variance):

Under H_0, $\bar{X} \sim N\left(\mu_0, \frac{s^2}{n}\right)$ **approximately**, with $\mu_0 = 118$.

From the sample, $\bar{x} = 121$, $s = \sqrt{97.454}$

OR

Step 3 (Large sample from a non-Normal population of unknown variance):

Under H_0, **since** $n = 60$ **is large,** $\bar{X} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right)$ **approximately, by Central Limit Theorem**, with $\mu_0 = 118$.

From the sample, $\bar{x} = 121$, $s = \sqrt{97.454}$

$$\begin{aligned}
&\text{"}\mathrm{H}_1 : \mu > \mu_0\text{"} \Rightarrow \mathrm{P}\left(\bar{X} \geq \bar{x}\right) \\
&\text{"}\mathrm{H}_1 : \mu < \mu_0\text{"} \Rightarrow \mathrm{P}\left(\bar{X} \leq \bar{x}\right) \\
&\text{"}\mathrm{H}_1 : \mu \neq \mu_0\text{"} \Rightarrow 2\mathrm{P}\left(\bar{X} \leq \bar{x}\right) \text{ if } \bar{x} < \mu_0 \\
&\qquad\qquad\qquad\quad \text{or } 2\mathrm{P}\left(\bar{X} \geq \bar{x}\right) \text{ if } \bar{x} > \mu_0
\end{aligned}$$

Step 4: Using a *z*-test, $p\text{-value} = \mathrm{P}(\bar{X} \geq 121) = 0.0385$ (3 s.f.)

Step 5: Since $p\text{-value} = 0.0385 < 0.05$, we **reject** H_0 and conclude that there is **sufficient** evidence, at 5% level of significance, to support the claim that the mean IQ of students in ABC University is greater than 118.

OR

(if test is performed at 1% level of significance)

Step 5: Since $p\text{-value} = 0.0385 > 0.01$, we **do not reject** H_0 and conclude that there is **insufficient** evidence, at 5% level of significance, to support the claim that the mean IQ of students in ABC University is greater than 118.

[Notice the only difference are the two phrases in bold and the end of the sentence is to describe the alternative hypothesis H_1.]

General Tips

- Take note of the nature of the "claim" in the question and the respective conclusion e.g. if the claim is "the mean IQ is equal to 118 or at least 118 or at most 118", then when $\mathrm{H}_0 : \mu = 118$ is rejected, there is "sufficient evidence at … to conclude that the claim is **invalid**"; whereas if the claim is "the mean IQ is higher or lower than 118", then when $\mathrm{H}_0 : \mu = 118$ is rejected, there is "sufficient evidence at … to conclude that the claim is **valid**".
- If H_0 is rejected at 5% level of significance for a certain *p*-value, H_0 will also be rejected at 10% level of significance or any level higher than 5%. On the contrary, H_0 may or may not be rejected at 1% or any level lower than 5%.

Definitions

1. The **level of significance** (or significance level) of a hypothesis test, denoted by α, is defined as the **probability of rejecting** H_0 **when** H_0 **is true** e.g. "5% level of significance" means "there is a 0.05 probability of wrongly concluding that the mean IQ of the students is more than 118 when in fact it is 118".
2. The p-value $= P(\bar{X} \geq 121)$ in this context refers to the probability of getting a sample with average IQ at least 121. This is also the value we will put down on our script if question asked for the smallest level of significance at which H_0 can be rejected in favour of H_1.

Important Cases where conclusion is given and we are asked to find:

Case 1: Level of significance $\alpha\%$

Carry on as if we are performing a test. After finding the p-value:

If given H_0 is rejected, p-value $\leq \frac{\alpha}{100}$; if given H_0 is not rejected, p-value $> \frac{\alpha}{100}$.

Case 2: Sample mean $\bar{x}$ (No standardization required)

Example: For a test at 5% level of significance and under H_0, $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, where $\mu_0 = 118$ and $\sigma = 12$. From the sample, $n = 50$.

If given H_0 **is rejected**, use: [GC invNorm: key in standard deviation $= \frac{12}{\sqrt{50}}$, not 12]

1-tail $(\text{e.g. } H_1 : \mu > 118)$	2-tail $(\text{e.g. } H_1 : \mu \neq 118)$
p-value ≤ 0.05 $P(\bar{X} \geq \bar{x}) \leq 0.05$ Using GC, $P(\bar{X} \geq 120.791) = 0.05$ $\bar{x} \geq 121$	p-value ≤ 0.05 $\mathbf{2}\,P(\bar{X} \geq \bar{x}) \leq 0.05$ or $\mathbf{2}\,P(\bar{X} \leq \bar{x}) \leq 0.05$ $P(\bar{X} \geq \bar{x}) \leq 0.025$ or $P(\bar{X} \leq \bar{x}) \leq 0.025$ $\vdots$

Case 3: μ_0 or n or σ^2 or s^2

All steps are similar to that of Case 2 above except we have to standardize in order to solve for the unknown parameter.

Example: (to find μ_0)

1-tail $\left(\text{e.g. } \mathrm{H}_1 : \mu > \mu_0;\ \text{from sample, } \bar{x} = 121\right)$
$p\text{-value} \le 0.05$ $\mathrm{P}\left(\bar{X} \ge 121\right) \le 0.05$ $\mathrm{P}\left(Z \ge \dfrac{121-\mu_0}{\sqrt{\dfrac{12^2}{50}}}\right) \le 0.05$ Using GC, $\mathrm{P}\left(Z \ge 1.6449\right) = 0.05$ $\dfrac{121-\mu_0}{\sqrt{\dfrac{12^2}{50}}} \ge 1.6449$ $\vdots$

Summary for Correlation and Regression

Terminology

- An **independent variable** is the variable whose change will have an effect on the **dependent variable**. Sometimes the independent variable can be **controlled** so that the variable only assumes a set of predetermined values.

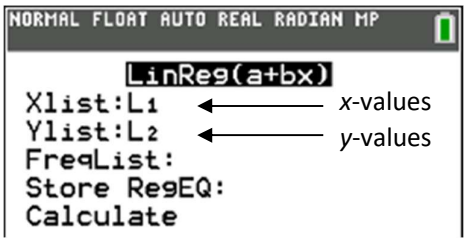
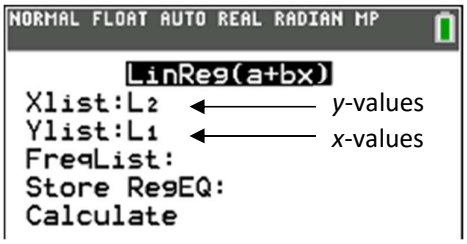
- A **scatter diagram** is a two-dimensional plot, with the values of one variable plotted along each axis. We plot the independent variable along the horizontal axis. A scatter diagram is used to show visually the relation between two variables, and it helps to identify outliers.

- The **product moment correlation coefficient**, denoted by r, is a measure of the strength of the linear relation between two variables. The value of r is independent of the units of the variables, and $-1 \leq r \leq 1$.

Description of Scatter Diagram	**Relation between variables**	***r* –value**
Points lie close to a straight line of positive gradient.	Positive linear correlation between variables	$r > 0$
Points lie close to a straight line of negative gradient.	Negative linear correlation between variables	$r < 0$
All points lie on a straight line of positive gradient.	Perfect positive linear correlation between variables.	$r = 1$
All points lie on a straight line of negative gradient.	Perfect negative linear correlation between variables.	$r = -1$
Points are spread randomly without visible trend.	No clear relation between variables	$r \approx 0$
Points lie close to a curve.	Non-linear relation between variables	Depends on how close the curve is to a straight line.

Note the product moment correlation coefficient merely gives an idea of the linear relationship between the variables. It does not imply any cause-and-effect relationship between the variables. There may be intermediate variables involved in the relationship which we do not know about, or there may even be more than one explanation to the linear relation.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to a set of observed data.

Regression Line	**GC**
The **least squares regression line of *y* on *x*** has the form $y = a + bx$. It is used when - x is the independent variable and y is the dependent variable; or - the independent/dependent variable cannot be determined and you want to estimate y for a given value of x.	
The **least squares regression line of *x* on *y*** has the form $x = c + dy$. It is used when - y is the independent variable and x is the dependent variable; or - the independent/dependent variable cannot be determined and you want to estimate x for a given value of y.	

Note:
- Both the regression lines of y on x and x on y pass through the point $(\overline{x}, \overline{y})$.
- In general, the regression line of y on x is <u>not</u> the same as the regression line of x on y.
- The regression lines of y on x and x on y are the same if and only if the product moment correlation coefficient between x and y is 1 or -1. The closer the coefficient is to either of these values, the closer are the lines to each other.

Estimating a Value using a Regression Line
Using the regression line to estimate a value within the range of data is known as **interpolation**. Using the line to estimate values outside the range of data is known as **extrapolation**. Extrapolating using values beyond the given range is unreliable as the regression model may not be applicable outside the range.

An estimate obtained using the linear regression line is reliable if
- there is linear correlation between the variables, and
- the estimate is obtained by interpolation.

Linearization of Data
There are cases where the relation between the variables are non-linear. However, through a suitable transformation on the data, it may still be possible to find a linear relation between the variables for the transformed data, e.g. x and y have a non-linear relation if $y = ax^2 + b$, but there is a linear relation between x^2 and y.

Reminder when using GC:
For the GC to give the value of r together with the equation of the linear regression line, ensure that STAT DIAGNOSTICS is turned ON.

Revision Tutorial Questions

Source of Question: TPJC 2016/Prelim/02/11

1 A researcher claims that children in schools spend 3 hours per week outdoors. The time, x hours, spent per week outdoors by a random sample of 80 children is summarised by

$$\sum x = 268, \qquad \sum (x - \bar{x})^2 = 195.$$

(i) Find unbiased estimates of the population mean and variance. **[2]**

(ii) Test, at the 5% significance level, whether the researcher's claim is valid. **[5]**

A physical education teacher at school A claims that the children in school A spend more than 3 hours per week outdoors. The standard deviation of the number of hours spent per week outdoors by the children in school A is known to be 1.25 hours. A random sample of 100 students from School A is chosen and the mean number of hours spent per week outdoors of this sample is m hours. A test at the 5% significance level indicates that the teacher's claim is valid.

(iii) Find the least possible value of m, giving your answer correct to 2 decimal places. **[3]**

Solution:

1(i) [2]	Unbiased estimate of population mean $\bar{x} = \dfrac{\sum x}{80} = \dfrac{268}{80} = 3.35$. Unbiased estimate of population variance, $s^2 = \dfrac{1}{80-1}\sum (x - \bar{x})^2 = \dfrac{195}{79}$
1(ii) [5]	Let μ denote the population mean time spent per week outdoors by children (in hours). To test $\mathrm{H}_0 : \mu = 3$ vs $\mathrm{H}_1 : \mu \neq 3$ Perform a 2-tail test at 5% significance level. Under H_0, since $n = 80$ is large, $\bar{X} \sim \mathrm{N}\left(\mu_0, \dfrac{s^2}{n}\right)$ approximately by Central Limit Theorem, where $\mu_0 = 3$. From the sample, $\bar{x} = 3.35$, $s = \sqrt{2.4684}$. Using a z-test, $p\text{-value} = 2\mathrm{P}(\bar{X} \geq 3.35) = 0.0463$ Since p-value $= 0.0463 < 0.05$, we reject H_0 and conclude that there is sufficient evidence, at 5% level of significance, that the researcher's claim is not valid.

(iii) [3]	Let Y be the random variable that represents the time, in hours, spent per week outdoors by children in School A. $H_0: \mu = 3$ vs $H_1: \mu > 3$ Perform a 1-tail test at 5% significance level. Under H_0, since $n = 100$ is large, $\overline{Y} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ approximately by Central Limit Theorem, where $\mu_0 = 3$, $\sigma = 1.25$ From the sample, $\overline{y} = m$. For teacher's claim to be valid, we reject H_0. p-value ≤ 0.05 $P\left(\overline{Y} \geq m\right) \leq 0.05$ $P\left(\overline{Y} \leq m\right) \geq 0.95$ By GC, $m \geq 3.2056$ Least value of $m = 3.21$ (2 d.p)

Source of Question: SRJC Prelim 9758/2017/02/Q8

2 A company manufactures bottles of iced coffee. Machines A and B are used to fill the bottles with iced coffee.

(i) Machine A is set to fill the bottles with 500 ml of iced coffee. A random sample of 50 filled bottles was taken and the volume of iced coffee (x ml) in each bottle was measured. The following data was obtained

$$\sum x = 24965 \quad \sum (x - \bar{x})^2 = 365 .$$

Calculate unbiased estimates of the population mean and variance. Test at the 2% level of significance, whether the mean volume of iced coffee per bottle is 500 ml. [6]

(ii) The company claims that Machine B filled the bottles with μ_0 ml of iced coffee. A random sample of 70 filled bottles was taken and the mean is 489.1 ml with standard deviation 4 ml. Find the range of values of μ_0 for which there is sufficient evidence for the company to have overstated the mean volume at the 2% level of significance. [5]

Solution:

2(i) [6]

Unbiased estimate of population mean , $\bar{x} = \dfrac{24965}{50} = 499.3$

Unbiased estimate of population variance

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{50}{49}\left(\frac{365}{50}\right) = \frac{365}{49}$$

Let μ denote the population mean volume of iced coffee bottles (in ml) from machine *A*.

To test $\mathrm{H}_0 : \mu = 500$ vs $\mathrm{H}_1 : \mu \neq 500$

Perform a 2-tail test at 2% level of significance

Under H_0, since $n = 50$ is large,

$\bar{X} \sim \mathrm{N}\left(\mu_0, \dfrac{s^2}{n}\right)$ approximately, by Central Limit Theorem,

with $\mu_0 = 500$

From the sample, $\bar{x} = 499.3$, $s = \sqrt{\dfrac{365}{49}}$

Using a *z*-test, p-value $= 2\mathrm{P}\left(\bar{X} \leq 499.3\right) = 0.0697$

Since p-value $= 0.0697 > 0.02$, we do not reject H_0 and conclude that there is insufficient evidence, at 2% level of significance, that the mean volume is not 500ml.

(ii) [5]

Let *Y* be the random variable denoting the volume of a randomly chosen iced coffee bottle in ml from Machine *B*.

Unbiased estimate for population variance = $\dfrac{70}{69}\left(4^2\right) = 16.232$

$\mathrm{H}_0 : \mu = \mu_0$

$\mathrm{H}_1 : \mu < \mu_0$

	Perform a 1-tail test at 2% level of significance Under H_0, since $n = 70$ is large, $\overline{Y} \sim N\left(\mu_0, \frac{s^2}{n}\right)$ approximately, by Central Limit Theorem From the sample, $\overline{y} = 489.1$, $s = \sqrt{16.232}$ For H_0 to be rejected, p-value ≤ 0.02 $P(\overline{Y} \leq 489.1) \leq 0.02$ $P\left(Z \leq \frac{489.1 - \mu_0}{\sqrt{\frac{16.232}{70}}}\right) \leq 0.02$ From GC, $\frac{489.1 - \mu_0}{\sqrt{\frac{16.232}{70}}} \leq -2.053748911$ $\mu_0 \geq 490$ (to 3 s.f.)

Source of Question: ACJC Prelim 9758/2017/02/Q7(modified)

3 It has been suggested that the optimal pH value for shampoo should be 5.5, to match the pH level of healthy scalp. Any pH value that is too low or too high may have undesirable effects on the user's hair and scalp. A shampoo manufacturer wants to investigate if the pH level of his shampoo is at the optimal value, by carrying out a hypothesis test at the 10% significance level. He measures the pH value, x, of n randomly chosen bottles of shampoo, where n is large.

(a) In the case where $n = 30$, it is found that $\sum x = 178.2$ and $\sum x^2 = 1238.622$.

(i) Find unbiased estimates of the population mean and variance, and carry out the test at the 10% significance level. [6]

(ii) Explain if it is necessary for the manufacturer to assume that the pH value of a bottle of shampoo follows a normal distribution. [1]

(b) In the case where n is unknown, assume that the sample mean is the same as that found in **(a)**.

Given that n is large and that the population variance is found to be 6.5, find the greatest value of n that will result in a favourable outcome for the manufacturer at the 10% significance level. [4]

Solution:

3(a) **(i)** **[6]**	Unbiased estimate of population mean $= \bar{x}$ $= \dfrac{178.2}{30}$ $= \underline{5.94}$ Unbiased estimate of population variance $= s^2$ $= \dfrac{1}{29}\left(1238.622 - \dfrac{178.2^2}{30}\right)$ $= 6.21083$ $= \underline{6.21}$ (3 s.f.) Let μ denote the population mean pH value of the shampoo. To test $\mathrm{H}_0 : \mu = 5.5$ vs $\mathrm{H}_1 : \mu \neq 5.5$ Perform a 2-tail test at 10% significance level Under H_0, since $n = 30$ is large, $\bar{X} \sim \mathrm{N}\left(\mu_0, \dfrac{s^2}{n}\right)$ approximately, by Central Limit Theorem, with $\mu_0 = 5.5$ From the sample, $\bar{x} = 5.94$, $s = \sqrt{6.21083}$ Using a z-test, p-value $= 2\mathrm{P}\left(\bar{X} \geq 5.94\right) = 0.334$ Since p-value $= 0.334 > 0.1$, we do not reject H_0 and conclude that there is insufficient evidence, at 10% level of significance, that the population mean pH value of the shampoo is not 5.5.
(ii) **[1]**	It is not necessary to assume X is normally distributed. As the sample size is large, by Central Limit Theorem, $\bar{X}$ is approximately normally distributed.
(b) **[4]**	Under H_0, since n is large, $\bar{X} \sim \mathrm{N}\left(\mu_0, \dfrac{\sigma^2}{n}\right)$ approximately, by Central Limit Theorem, with $\mu_0 = 5.5$ and $\sigma = 6.5$. From the sample, $\bar{x} = 5.94$ For a favourable outcome at 10% significance level, do not reject H_0, p-value > 0.1

$2\mathrm{P}(\bar{X} \geq 5.94) > 0.1$

$\mathrm{P}(\bar{X} \geq 5.94) > 0.05$

$$\mathrm{P}\left(Z \geq \frac{5.94 - 5.5}{\sqrt{\frac{6.5}{n}}} \right) > 0.05$$

$$\mathrm{P}\left(Z \leq \frac{5.94 - 5.5}{\sqrt{\frac{6.5}{n}}} \right) < 0.95$$

From GC,

$$\frac{0.44\sqrt{n}}{\sqrt{6.5}} < 1.64485$$

$$\sqrt{n} < \frac{1.64485\sqrt{6.5}}{0.44}$$

$$n < \left(\frac{1.64485\sqrt{6.5}}{0.44} \right)^2$$

$n < 90.837$

Hence largest $n = \underline{90}$

Source of Question: DHS Prelim 9758/2017/02/Q7

4 The company Snatch provides a ride-hailing service comprising taxis and private cars in Singapore. Snatch claims that the mean waiting time for a passenger from the booking time to the time of the vehicle's arrival is 7 minutes.

To test whether the claim is true, a random sample of 30 passengers' waiting times is obtained. The standard deviation of the sample is 2 minutes. A hypothesis test conducted concludes that there is sufficient evidence at the 1% significance level to reject the claim.

(i) State appropriate hypotheses and the distribution of the test statistic used. [3]

(ii) Find the range of values of the sample mean waiting time, $\bar{t}$. [3]

(iii) A hypothesis test is conducted at the 1% significance level whether the mean waiting time of passengers is more than 7 minutes. Using the existing sample, deduce the conclusion of this test if the sample mean waiting time is more than 7 minutes. [2]

Solution:

4(i) **[3]**	Let X be the waiting time for a passenger from the booking time to the time of the vehicle's arrival. Let μ the population mean of X. To test $\mathrm{H}_0: \ \mu = 7$ vs $\mathrm{H}_1: \ \mu \neq 7$ Perform a 2-tail test at 1% level of significance Under H_0, since $n = 30$ is large, $\overline{T} \sim \mathrm{N}\left(\mu_0, \frac{s^2}{n}\right)$ approximately, by Central Limit Theorem, with $\mu_0 = 7$. From the sample, $s^2 = \frac{30}{29}(\text{sample variance}) = \frac{30}{29}(4) = \frac{120}{29}$.
(ii) **[3]**	To reject H_0 at 1% level of significance, p-value ≤ 0.01 $2\mathrm{P}(\overline{T} \geq k) \leq 0.01$ or $2\mathrm{P}(\overline{T} \leq k) \leq 0.01$ $\mathrm{P}(\overline{T} \geq k) \leq 0.005$ or $\mathrm{P}(\overline{T} \leq k) \leq 0.005$ From GC, $\mathrm{P}(\overline{T} \geq 7.96) = 0.005$ or $\mathrm{P}(\overline{T} \leq 6.04) = 0.005$ $\overline{t} \geq 7.96$ or $\overline{t} \leq 6.04$ Range of values of $\overline{t}$ is $\overline{t} \geq 7.96$ or $\overline{t} \leq 6.04$.
(iii) **[2]**	From the two tail test, we know that p-value (two tail) ≤ 0.01. For a one-tail test, p-value(one tail) $= \frac{p\text{-value (two tail)}}{2} \leq 0.005 < 0.01$, therefore we reject H_0 and conclude that there is sufficient evidence at 1% significance level to say that mean waiting time is more than 7 minutes.

Source of Question: 9740/2015/02/Q8(modified)

5 A market stall sells pineapples which have masses that are normally distributed with standard deviation 0.08 kg. The stall owner claims that the mean mass of the pineapples is at least 0.9 kg. Nur buys a random selection of 8 pineapples from the stall. The 8 pineapples have masses, in kg, as follows.

0.80 1.00 0.82 0.85 0.93 0.96 0.81 0.89

Find unbiased estimates of the population mean and variance of the mass of pineapples. Test at the 10% level of significance whether there is any evidence to doubt the stall owner's claim. [7]

Solution:

5 **[7]**

Let X be the mass of a pineapple in kg. $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$

From GC,
Unbiased estimate of μ, $\bar{x} = 0.8825$
Unbiased estimate of σ^2, $s^2 = 0.074785^2 \approx 0.00559$

Let μ the population mean of X.

To test $\mathrm{H}_0: \quad \mu = 0.9 \quad$ vs $\quad \mathrm{H}_1: \quad \mu < 0.9$

Perform a 1-tail test at 10% level of significance

Under H_0, $\bar{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$, where $\mu_0 = 0.9$ and $\sigma = 0.08$

From the sample, $\bar{x} = 0.8825$
Using a z-test, p-value $= \mathrm{P}\left(\bar{X} \leq 0.8825\right) = 0.268$

Since p-value = 0.268 > 0.1, we do not reject H_0 and thus we do not have sufficient evidence at 10% level of significance to doubt the claim of the stall owner that the mean mass of the pineapples is at least 0.9kg.

Source of Question: MI PU3 Mid-Year CT 9758/2018/02/Q11(modified)

6 At a food fair, several brands of chocolate are available.

The masses, in grams (g), of packets of Bodiva chocolate and Buylian chocolate are modelled as having independent normal distributions. The means and standard deviations of these distributions, and the selling prices, in dollars per gram, are shown in the table.

	Mean (g)	Standard Deviation (g)	Selling price (dollars per g)
A packet of Bodiva	184.2	5.2	0.03
A packet of Buylian	80	3.1	0.02

(i) State, with a reason, whether a normal model is likely to be appropriate for the mass of a packet of chocolate chosen at random from all the packets available. **[1]**

(ii) Find the probability that both a randomly chosen packet of Bodiva has a selling price exceeding \$5.50 and a randomly chosen packet of Buylian has a selling price exceeding \$1.50. **[3]**

(iii) Find the probability that the total selling price of a randomly chosen packet of Bodiva and a randomly chosen packet of Buylian is more than \$7.00. **[3]**

(iv) Explain why the answer to part **(iii)** is greater than the answer to part **(ii)**. **[1]**

(v) Packets of Neuhaus chocolate have mass with mean 75 g and standard deviation 2.1 g. A random sample of 50 packets of Buylian chocolate and a random sample of 50 packets of Neuhaus chocolate are taken. Find the probability that the difference in the sample means is less than 4 g. **[3]**

Solution:

6(i) **[1]**	A normal model is inappropriate because when we combine the two types of chocolate with different normal distributions, it results in a bi-modal distribution (i.e. with two modes) which contrasts with a normal distribution which has only one mode.
(ii) **[3]**	Let X be the mass of a randomly chosen packet of Bodiva. Then $X \sim \mathrm{N}\left(184.2, (5.2)^2\right)$. Let Y be the mass of a randomly chosen packet of Buylian. Then $Y \sim \mathrm{N}\left(80, (3.1)^2\right)$. Let U be the price of a randomly chosen packet of Bodiva. $U = 0.03X$ Then $U \sim \mathrm{N}\left(0.03 \times 184.2, 0.03^2 \times (5.2)^2\right) = \mathrm{N}\left(5.526, 0.024336\right)$. Let V be the price of a randomly chosen packet of Buylian.

	$V = 0.02Y$ Then $V \sim \mathrm{N}\left(0.02\times 80, 0.02^2 \times (3.1)^2\right) = \mathrm{N}(1.6, 0.003844)$. Required probability $= \mathrm{P}(U > 5.5) \times \mathrm{P}(V > 1.5)$ $= 0.53595 = 0.536 \ (3\text{sf})$
(iii) **[3]**	$U + V \sim \mathrm{N}(5.526 + 1.6, 0.024336 + 0.003844)$ $= \mathrm{N}(7.126, 0.02818)$ $\mathrm{P}(U + V > 7) = 0.77354 = 0.774 \ (3\text{sf})$
(iv) **[1]**	Event in part **(ii)** is a subset of the event in **(iii)**. For example, the case where $U = 5$ and $V = 2$ is in part **(iii)** but not in part **(ii)**.
(v) **[3]**	$Y \sim \mathrm{N}\left(80, (3.1)^2\right)$ $\bar{Y} \sim \mathrm{N}\left(80, \frac{(3.1)^2}{50}\right)$ Let W be the mass of a randomly chosen packet of Neuhaus. $\mathrm{E}(W) = 75$ $\mathrm{Var}(W) = 2.1^2$ Since $n = 50$ is large, by Central Limit Theorem, $\bar{W} \sim \mathrm{N}\left(75, \frac{2.1^2}{50}\right)$ approximately. $\bar{Y} - \bar{W} \sim \mathrm{N}\left(80 - 75, \frac{(3.1)^2}{50} + \frac{2.1^2}{50}\right)$ approximately, i.e. $\bar{Y} - \bar{W} \sim \mathrm{N}(5, 0.2804)$ approximately, $\mathrm{P}\left(\left\lvert \bar{Y} - \bar{W} \right\rvert < 4\right) = \mathrm{P}\left(-4 < \bar{Y} - \bar{W} < 4\right)$ $= 0.029481$ $= 0.0295 \ (3\text{sf})$

Source of Question: VJC Prelim 9758/2018/02/Q10

7 In an agricultural experiment, a certain fertilizer is applied at different rates to ten identical plots of land. Seeds of a type of grass are then sown and several weeks later, the mean height of the grass on each plot is measured. The results are shown in the table.

Rate of application of fertilizer, x g/m^2	10	20	30	40	50	60	70	80	90	100
Mean height of grass, y cm	6.2	11.4	13.2	14.8	15.8	17.0	19.4	19.4	20.6	20.8

(i) Draw the scatter diagram for these values, labelling the axes clearly. [1]

It is thought that the mean height of grass, y cm, can be modelled by one of the formulae

$$y = ax + b \quad \text{or} \quad y = c\ln x + d$$

where a, b, c and d are constants.

(ii) Find, correct to 4 decimal places, the value of the product moment correlation coefficient between

(a) x and y, **(b)** $\ln x$ and y. [2]

(iii) Use your answers to parts **(i)** and **(ii)** to explain which of $y = ax + b$ or $y = c\ln x + d$ is the better model. [2]

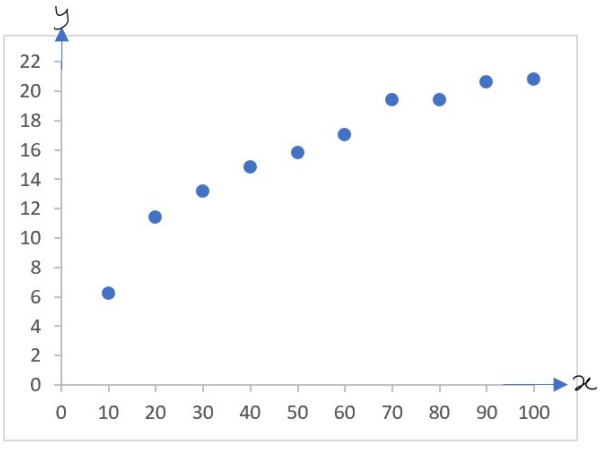
It is required to estimate the value of x for which $y = 17.2$.

(iv) Explain why neither the regression line of x on y nor the regression line of $\ln x$ on y should be used. [1]

(v) Find the equation of a suitable regression line and use it to find the required estimate. [3]

Solution:

i



ii a	$r = 0.9541$ (4 d.p)
b	$r = 0.9942$ (4 d.p)
iii	From the scatter diagram, as x increases, y increases by decreasing amounts. In addition, the product moment correlation coefficient between $\ln x$ and y, 0.9942, is closer to 1 as compared to that between x and y, 0.9541. Hence $y = c + d\ln x$ is the better model.
iv	Since x is the independent variable, neither the regression line of x on y nor the regression line of $\ln x$ on y should be used to estimate the value of x when $y = 17.2$.
v	Equation of regression line of y on $\ln x$ is $y = 6.3074\ln x - 8.1904$ $y = 6.31\ln x - 8.19$ When y = 17.2, $17.2 = -8.1904 + 6.3074\ln x$ $x = 56.008$ $= 56.0$

Source of Question: YJC Prelim 9758/2018/02/Q9

8 A new Internet service provider in the market, Y-Fai, decides to investigate the effect of the distance from its router on its wifi signal. The most convenient way to express wifi signal strength is by using dBm, which stands for decibels relative to a milliwatt. The signal strength measured in dBm is negative in value and a value closer to 0 signifies a stronger wifi signal. An employee measures the signal strength (y dBm) at various fixed distances away from its router (x m) as follows.

x	1	1.5	2	2.5	3	10	15
y	–30	–60	–100	–76	–79	–83	–88

(i) Draw a scatter diagram for these values. On your diagram, circle the data point that seems to be unexpected and suggest a possible reason for the anomaly. [2]

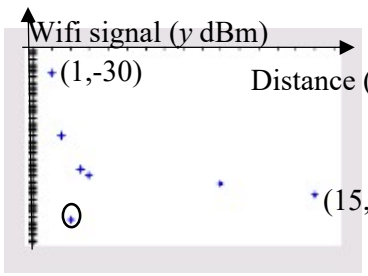
For parts **(ii)** and **(iii)** of this question, you should **exclude** the anomaly.

(ii) Explain from your scatter diagram why the relationship between x and y should not be modelled by an equation of the form $y = ax + b$. [1]

(iii) Which of the formulae $y = \frac{c}{x} + d$ and $y = \frac{e}{x^2} + f$, where c, d, e and f are constants, is the better model for the relationship between y and x? Explain fully how you decided, and find the constants for the better formula. [4]

(iv) Use the formula you chose from part **(iii)** to estimate the signal strength when the distance away from its router is 5 m. Explain why you would expect this estimate to be reliable. [2]

Solution:

(i)	 There could be an obstruction which reduces the strength of the wifi signal or There could be a wifi disruption.
(ii)	The data points do not seem to lie on a straight line.
(iii)	For $y = \frac{c}{x} + d$, $r = 0.970$ For $y = \frac{e}{x^2} + f$, $r = 0.997$ Since the r value for the second model is closer to 1, $y = \frac{e}{x^2} + f$ is the better model. $e = 55.81195... \approx 55.8\,(3\text{ s.f})$ $f = -85.42578... \approx -85.4\,(3\text{ s.f})$
(iv)	When $x = 5$, $y = \frac{55.81195}{5^2} - 85.42578$ $= -83.2\,(3\text{sf})$ The required signal strength is -83.2 dBm This estimate is expected to be reliable as $x = 5$ is within the given data range, and $r = 0.997$ is close to 1.

Source of Question: HCI Prelim 9758/2018/02/Q7

9 An engineering team from a car manufacturer wants to test their cars' braking system. The car travels along a stretch of road with speed v km/h. When the brakes are applied, the car comes to rest after travelling a further distance of d metres. A random sample of 6 pairs of values of v and d collected by a trainee mechanic from the engineering team is shown below.

v	30	40	50	60	70	80
d	5.00	5.30	6.45	8.50	17.00	25.90

(i) Draw a scatter diagram for these values, labelling the axes clearly. [2]

It is thought that the distance travelled d can be modelled by one of the following models.

Model I: $d = av + b$ or

Model II: $d = \mathrm{e}^{pv+q}$

where a, b, p and q are constants.

(ii) Find the value of the product moment correlation coefficient between

(a) v and d,

(b) v and $\ln d$. [2]

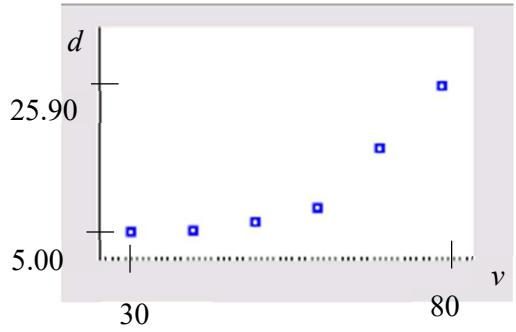
(iii) The trainee mechanic proposed that Model II is a better model than Model I. Use your answers to parts **(i)** and **(ii)** to explain why the trainee mechanic is right. [2]

(iv) Find the equation of the regression line of $\ln d$ on v. [1]

(v) Using the regression line in part **(iv)**, find the value of v if the driver applies his brakes immediately upon seeing an obstacle that is 10 metres away and stops just in time before crashing into it. [1]

(vi) The original data set contains 7 pairs of data with regression line $d = 0.4256v - 11.74$. The trainee mechanic found that he does not have the value of d when $v = 75$ from his record. Find the missing value of d correct to 2 decimal places. [3]

Solution:

(i)	
(ii)(a)	0.902 (3 s.f.)
(ii)(b)	0.955 (3 s.f.)
(iii)	The points on the scatter diagram are not close to a straight line, the d -values increase at an increasing rate as v increases. For Model II, the r value is closer to 1 as compared to Model I. Thus Model II is better.
(iv)	$d = e^{cv+d}$ $\ln d = cv + d$ Regression line of ln d on v: $\ln d = 0.0342758025v + 0.34294554177$ $\ln d = 0.0343v + 0.343$ (3 s.f.)
(v)	When $d = 10$, $\ln 10 = 0.0342758025v + 0.34294554177$ $v = 57.2$ km/h
(vi)	$d = 0.4256v - 11.74$ $(\bar{d}, \bar{v})$ satisfies the regression line. $\Rightarrow \bar{d} = 0.4256\bar{v} - 11.74$ Let d be the distance travelled after brakes are applied. $\Rightarrow \frac{(68.15 + d)}{7} = 0.4256\frac{(330 + 75)}{7} - 11.74$ $\Rightarrow 68.15 + d = (0.4256)(405) - (11.74)(7)$ $\Rightarrow d = 22.04$ metres

Source of Question: RVHS Prelim 9758/2018/02/Q8

10 Verde wants to investigate the time taken for different volumes of water to cool to room temperature. He prepared a few samples of different volume and heated the samples to their boiling point, and then recorded the time taken for the water to cool to room temperature. The results are given in the table.

Volume (x/cm^3)	100	200	300	400	500	600	700
Time (t/min)	14	23	47	83	A	172	293

(i) It is known that the regression line of t on x is $t = -65.1429 + 0.4318x$. Show that $a = 121$. [2]

(ii) Draw a scatter diagram for the data and find the product moment correlation coefficient between x and t. [2]

(iii) Comment whether the regression line is appropriate based on

(a) the scatter diagram; [1]

(b) the context. [1]

Verde considers using one of the following two models:

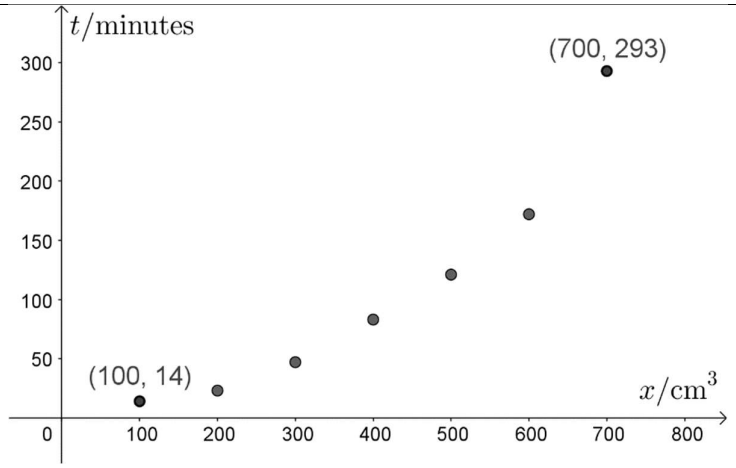
$$\text{A: } t = a + bx^2, \quad \text{B: } t = ae^{bx},$$

where $a, b \in \mathbb{R}$, for the relationship between x and t.

(iv) Explain which is the better model and find the equation of a suitable regression line for that model. [2]

(v) Estimate the time taken for 450 cm^3 of water to cool from boiling point to room temperature. Comment on the reliability of the estimate. [3]

Solution:

(i)	$\overline{x} = 400$ $\overline{t} = -65.1429 + 0.4318(400)$ $\overline{t} = 107.5771$ $14 + 23 + 47 + 83 + a + 172 + 293 = 7 \times \overline{t}$ $632 + a = 753.0397$ $a \approx 121$ (shown)
(ii)	 $r = 0.941$ (3sf)
(iii)	(a) Although the r value is 0.941 which is close to 1 suggesting that there is a strong linear correlation, the scatter diagram shows that the relationship is more of a curvilinear one, with t increasing at an increasing rate as x increases. Thus, the regression line is not appropriate. (b) The regression line predicts that for volume less than a certain amount, it takes negative time for the water to cool to room temperature which is not possible in the context. Thus, the regression line is not appropriate.
(iv)	Calculating the r- values for models A and B, Model A: r-value = 0.987 while Model B: $t = ae^{bx} \Rightarrow \ln t = \ln a + bx$, r-value = 0.994, Since the $\lvert r \rvert$-value for model B is closer to 1 compared to model A, model B is the better model. Regression line of $\ln t$ on x is $\ln t = 2.224858298 + 0.0050332105x$ $\ln t = 2.22 + 0.00503x$ (3 sig fig)
(v)	When x = 450, $\ln t = 2.224858298 + 0.0050332105(450)$ $t = 89.1$ minutes (3sf) The estimate is reliable as r = 0.994 which is close to 1, suggesting a strong linear correlation between $\ln t$ and x; and x = 450 is within the given data range of 100 to 700.