2021

BOTANY — HONOURS

Paper : DSE-A-1

(Biostatistics)

Full Marks : 50

The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable.

1. Answer any five questions :

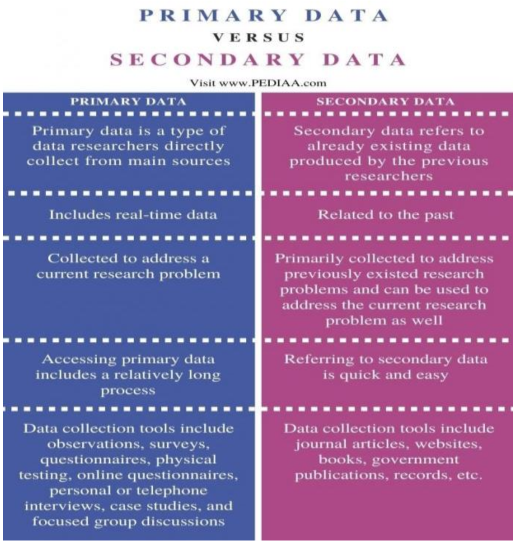
(a) State two limitations of biometry.

(a) Two limitations of biometry are:

1. All the conclusions about statistical analysis depend on the availability of sample data. If the sampling is biased, the analysis will be erroneous[1].
2. Statistics can be used and applied only on collective data, not on individual data[1].

(b) Distinguish between primary and secondary data.

(b)



PRIMARY DATA VERSUS SECONDARY DATA

Visit www.PEDIAA.com

PRIMARY DATA	SECONDARY DATA
Primary data is a type of data researchers directly collect from main sources	Secondary data refers to already existing data produced by the previous researchers
Includes real-time data	Related to the past
Collected to address a current research problem	Primarily collected to address previously existed research problems and can be used to address the current research problem as well
Accessing primary data includes a relatively long process	Referring to secondary data is quick and easy
Data collection tools include observations, surveys, questionnaires, physical testing, online questionnaires, personal or telephone interviews, case studies, and focused group discussions	Data collection tools include journal articles, websites, books, government publications, records, etc.

(c) What is alternative hypothesis?

© The alternative hypothesis is a statement used in statistical inference experiment. It is contradictory to the null hypothesis and denoted by Ha or H1. We can also say that it is simply an alternative to the null. In hypothesis testing, an alternative theory is a statement which a researcher is testing. This statement is true from the researcher's point of view and ultimately proves to reject the null to replace it with an alternative assumption[3].

(d) Define 'goodness of fit'.

(d) Goodness of fit evaluates how well observed data align with the expected values from a statistical model. A high goodness of fit indicates the observed values are close to the model's expected values. A low goodness of fit shows the observed values are relatively far from the expected values[4].

(e) Compare variable and variate

(e) A random variable or stochastic variable is a variable whose value is subject to variations due to chance. A random variate is a particular outcome of a random variable: the random variates which are other outcomes of the same random variable would have different values[5].

(f) What is central tendency? Write about one measure of central tendency.

(f) Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution[6]. One measure of central tendency is the mean, which is the sum of all values in a dataset divided by the number of values in that dataset.

(g) Compare simple random sampling and non-random sampling

(g) Random sampling is a method where each sample has an equal chance of being selected. It ensures a representative and unbiased sample[7]. Non-random sampling is a method where the selection of samples is based on factors such as convenience, judgment, and experience of the researcher. It offers convenience and allows researchers to target specific subgroups[7].

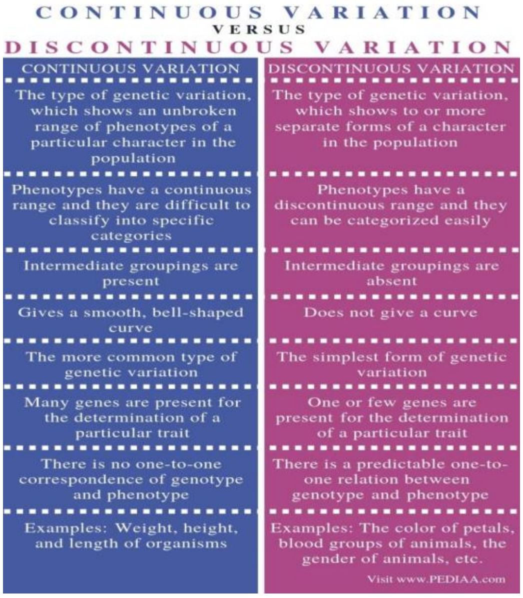
(h) What will be the value of 'probability of not happening' when the value of 'probability of happening' is 0·8?

(h) The value of 'probability of not happening' when the value of 'probability of happening' is 0.8 is 0.2. This is because the sum of the probabilities of all possible outcomes is always 1. Therefore, if the probability of an event happening is 0.8, then the probability of it not happening is 1 - 0.8 = 0.2.

2. Answer any two of the following :

(a) Distinguish between (any two) : 2½×2

(i) Continuous variable and Discontinuous variable.

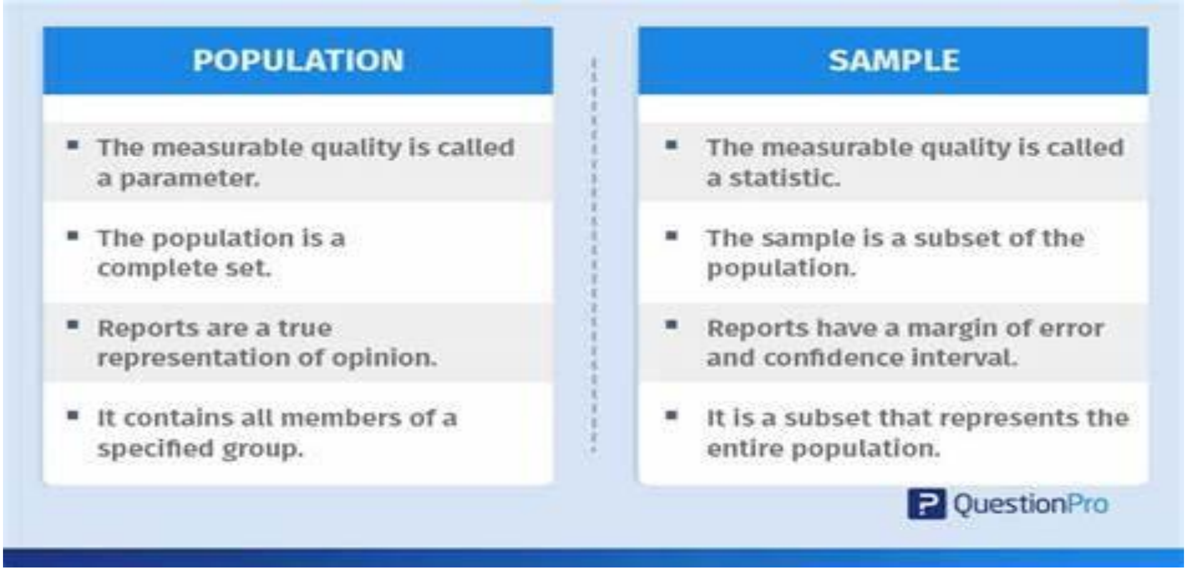


CONTINUOUS VARIATION VERSUS DISCONTINUOUS VARIATION

CONTINUOUS VARIATION	DISCONTINUOUS VARIATION
The type of genetic variation, which shows an unbroken range of phenotypes of a particular character in the population	The type of genetic variation, which shows to or more separate forms of a character in the population
Phenotypes have a continuous range and they are difficult to classify into specific categories	Phenotypes have a discontinuous range and they can be categorized easily
Intermediate groupings are present	Intermediate groupings are absent
Gives a smooth, bell-shaped curve	Does not give a curve
The more common type of genetic variation	The simplest form of genetic variation
Many genes are present for the determination of a particular trait	One or few genes are present for the determination of a particular trait
There is no one-to-one correspondence of genotype and phenotype	There is a predictable one-to-one relation between genotype and phenotype
Examples: Weight, height, and length of organisms	Examples: The color of petals, blood groups of animals, the gender of animals, etc.

Visit www.PEDIAA.com

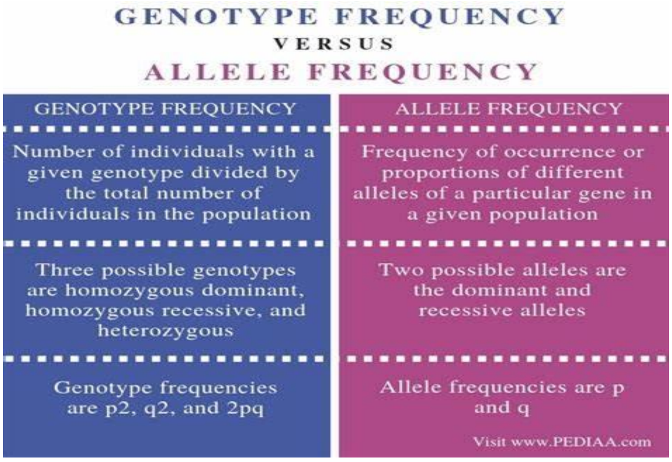
(ii) Population parameter and Sample statistic.



POPULATION	SAMPLE
▪ The measurable quality is called a parameter.	▪ The measurable quality is called a statistic.
▪ The population is a complete set.	▪ The sample is a subset of the population.
▪ Reports are a true representation of opinion.	▪ Reports have a margin of error and confidence interval.
▪ It contains all members of a specified group.	▪ It is a subset that represents the entire population.

QuestionPro

(iii) Genotype frequency and Allele frequency.



GENOTYPE FREQUENCY VERSUS ALLELE FREQUENCY

GENOTYPE FREQUENCY	ALLELE FREQUENCY
Number of individuals with a given genotype divided by the total number of individuals in the population	Frequency of occurrence or proportions of different alleles of a particular gene in a given population
Three possible genotypes are homozygous dominant, homozygous recessive, and heterozygous	Two possible alleles are the dominant and recessive alleles
Genotype frequencies are p2, q2, and 2pq	Allele frequencies are p and q

Visit www.PEDIAA.com

(b) Define mean, median and mode. Explain which one is more acceptable in statistics .

Mean:

The mean, often referred to as the average, is a measure of central tendency that is calculated by adding up all the values in a data set and then dividing by the number of observations. Mathematically, the mean ($\bar{X}$) is expressed as:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where X_i represents each individual value in the data set, and n is the number of observations.

- **Median**: The median is the middle value in an ordered dataset[2]. When the data is arranged in ascending or descending order, the median is the value that separates the higher half from the lower half of the data sample[3]. If the dataset has an odd number of observations, the median is the middle number. If the dataset has an even number of observations, the median is the average of the two middle numbers.
- **Mode**: The mode is the value that appears most frequently in a dataset[4]. In other words, it is the value that is most likely to be sampled[5].

As for which measure is more acceptable in statistics, it depends on the nature of the data and the specific circumstances:

- The **mean** is often used when the data is symmetrically distributed with no skew, and there are no outliers[6]. It includes every value in your data set as part of the calculation and is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero[6].
- The **median** is a more appropriate measure of central tendency for skewed distributions or when there are outliers in the data[7]. This is because the median is robust to outliers, meaning outliers do not impact the median value significantly compared to the mean[7].
- The **mode** is useful when dealing with categorical data, or when it's important to identify the most common value[8].

In summary, all three measures of central tendency (mean, median, and mode) have their uses and can provide valuable insights about a dataset. The choice of which one to use depends on the characteristics of the data and the purpose of the analysis.

(c) Briefly mention the factors affecting gene frequency.

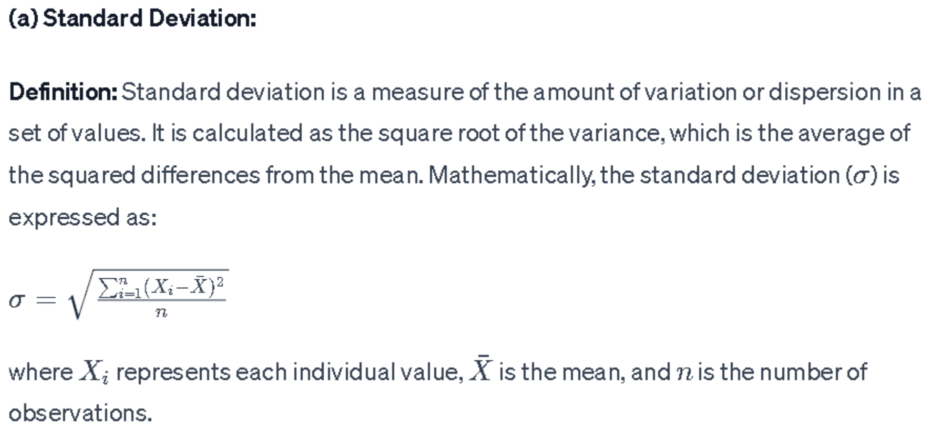
Gene frequency, also known as allele frequency, can be influenced by several factors[12]:

1. **Mutation**: Mutations introduce new alleles into a population. Although the rate of mutation is generally low, these rare events can produce considerable genetic variability over time[12].
2. **Migration or Gene Flow**: The movement of individuals between populations can significantly impact gene frequencies. When individuals migrate from one population to another, they carry their genes with them, thereby altering the gene frequencies in both the source and destination populations[12].
3. **Genetic Drift**: This is a random change in allele frequencies that occurs in small populations. Genetic drift can lead to genetic variation within populations and can result in alleles being completely eliminated or becoming fully dominant over time[12].
4. **Natural Selection**: This is the process by which certain traits become more or less common in a population due to consistent effects upon the survival or reproduction of their bearers. It is a key mechanism of evolution[12].
5. **Recombination during Sexual Reproduction**: This can create new combinations of genes and can significantly affect gene frequencies[2].

These factors can cause changes in gene frequencies, which are the basis for understanding the processes of evolution and population genetics[1].

3. Answer any three of the following :

(a) What is standard deviation? Mention its merits and demerits. The plant height of a rice cultivar is as follows : Plant height (cm) 80 85 86 90 91 95 96 100 No. of plants 3 5 8 4 – – – – Calculate the mean and standard error of the height of cultivar.



(a) Standard Deviation:

Definition: Standard deviation is a measure of the amount of variation or dispersion in a set of values. It is calculated as the square root of the variance, which is the average of the squared differences from the mean. Mathematically, the standard deviation (σ) is expressed as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

where X_i represents each individual value, $\bar{X}$ is the mean, and n is the number of observations.

Merits of Standard Deviation:

1. **Sensitive to Variability:** Standard deviation takes into account the spread of values, making it sensitive to variability within the data set.
2. **Useful in Comparisons:** It is useful for comparing the degree of dispersion between different data sets.
3. **Incorporates all Data Points:** Standard deviation considers every data point in its calculation, providing a comprehensive measure of variability.

Demerits of Standard Deviation:

1. **Affected by Outliers:** Standard deviation is influenced by extreme values or outliers in the data, which can skew its interpretation.
2. **May Be Misleading for Skewed Distributions:** In the case of skewed distributions, the mean and standard deviation may not accurately represent the central tendency and variability.

Calculation for the Given Data:

$$\text{Mean}(\bar{X}) = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(3\times80)+(5\times85)+(8\times86)+(4\times90)}{20}$$

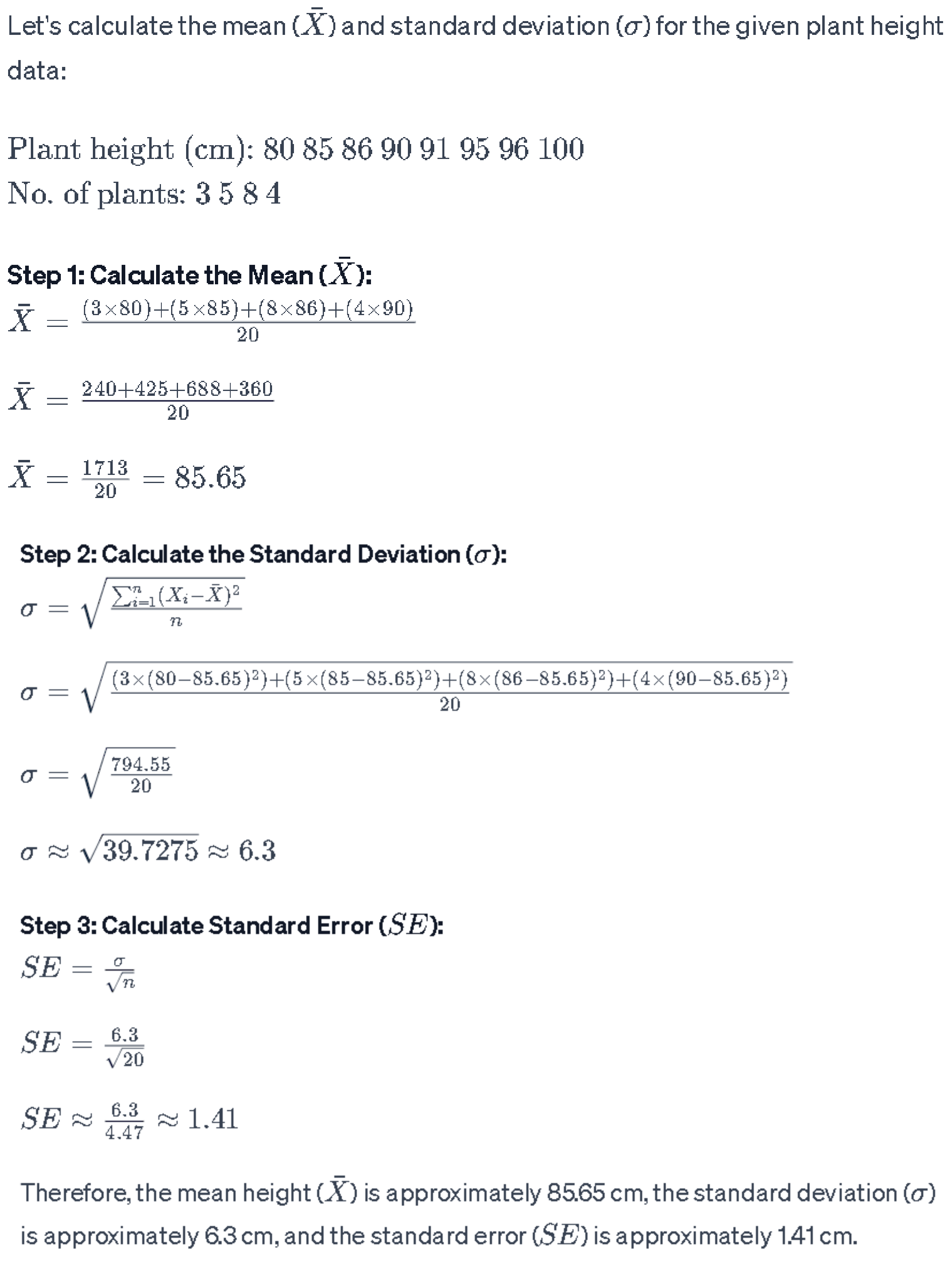
$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}}$$

Now, plug in the values to find the mean and standard deviation for the given data set.

Once the mean is calculated, the standard error of the mean (SE) can be obtained using the formula:

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation, and n is the number of observations.



Let's calculate the mean ($\bar{X}$) and standard deviation (σ) for the given plant height data:

Plant height (cm): 80 85 86 90 91 95 96 100
No. of plants: 3 5 8 4

Step 1: Calculate the Mean ($\bar{X}$):

$$\bar{X} = \frac{(3\times80)+(5\times85)+(8\times86)+(4\times90)}{20}$$

$$\bar{X} = \frac{240+425+688+360}{20}$$

$$\bar{X} = \frac{1713}{20} = 85.65$$

Step 2: Calculate the Standard Deviation (σ):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}}$$

$$\sigma = \sqrt{\frac{(3\times(80-85.65)^2)+(5\times(85-85.65)^2)+(8\times(86-85.65)^2)+(4\times(90-85.65)^2)}{20}}$$

$$\sigma = \sqrt{\frac{794.55}{20}}$$

$$\sigma \approx \sqrt{39.7275} \approx 6.3$$

Step 3: Calculate Standard Error (SE):

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE = \frac{6.3}{\sqrt{20}}$$

$$SE \approx \frac{6.3}{4.47} \approx 1.41$$

Therefore, the mean height ($\bar{X}$) is approximately 85.65 cm, the standard deviation (σ) is approximately 6.3 cm, and the standard error (SE) is approximately 1.41 cm.

(b) What is dispersion? State the properties of a normal distribution curve. The leaflet length of Cassia sophera is as follows : Leaflet length (cm) 5.1 6.0 6.1 7.0 7.1 8.0 8.1 9.0 9.1 10.0 No. of leaflets 2 12 25 13 3 – – – – – Calculate the coefficient of variation and comment on it.

(b) Dispersion:

Definition: Dispersion in statistics refers to the spread or variability of a set of values. It measures how much the individual values in a data set differ from the central tendency (mean, median, etc.).

Properties of a Normal Distribution Curve:

1. **Symmetry:** The normal distribution curve is symmetrical around its mean. This means that the left and right halves of the curve are mirror images of each other.

2. **Bell-Shaped Curve:** The curve has a characteristic bell shape, with a single peak at the mean. As you move away from the mean in either direction, the frequency of observations decreases.

3. **Mean, Median, and Mode Coincide:** In a normal distribution, the mean (average), median, and mode are all located at the center of the distribution, and they coincide.

4. **Empirical Rule:** A large percentage of the data falls within a certain number of standard deviations from the mean:

- About 68% of the data falls within one standard deviation.
- About 95% falls within two standard deviations.
- About 99.7% falls within three standard deviations.

5. **Continuous and Unbounded:** The normal distribution is a continuous probability distribution that extends infinitely in both directions.

Calculation for the Given Leaflet Length Data:

Leaflet length (cm): 5.1 6.0 6.1 7.0 7.1 8.0 8.1 9.0 9.1 10.0

No. of leaflets: 2 12 25 13 3

Coefficient of Variation (CV):

\[CV = \frac{\sigma}{\bar{X}} \times 100\% \]

where \(\sigma\) is the standard deviation and \(\bar{X}\) is the mean.

Step 1: Calculate the Mean (\(\bar{X}\)):

\[\bar{X} = \frac{(2 \times 5.1) + (12 \times 6.0) + (25 \times 6.1) + (13 \times 7.0) + (3 \times 7.1) + (8 \times 8.0) + (1 \times 8.1) + (9 \times 9.0) + (1 \times 9.1) + (10 \times 10.0)}{75} \]

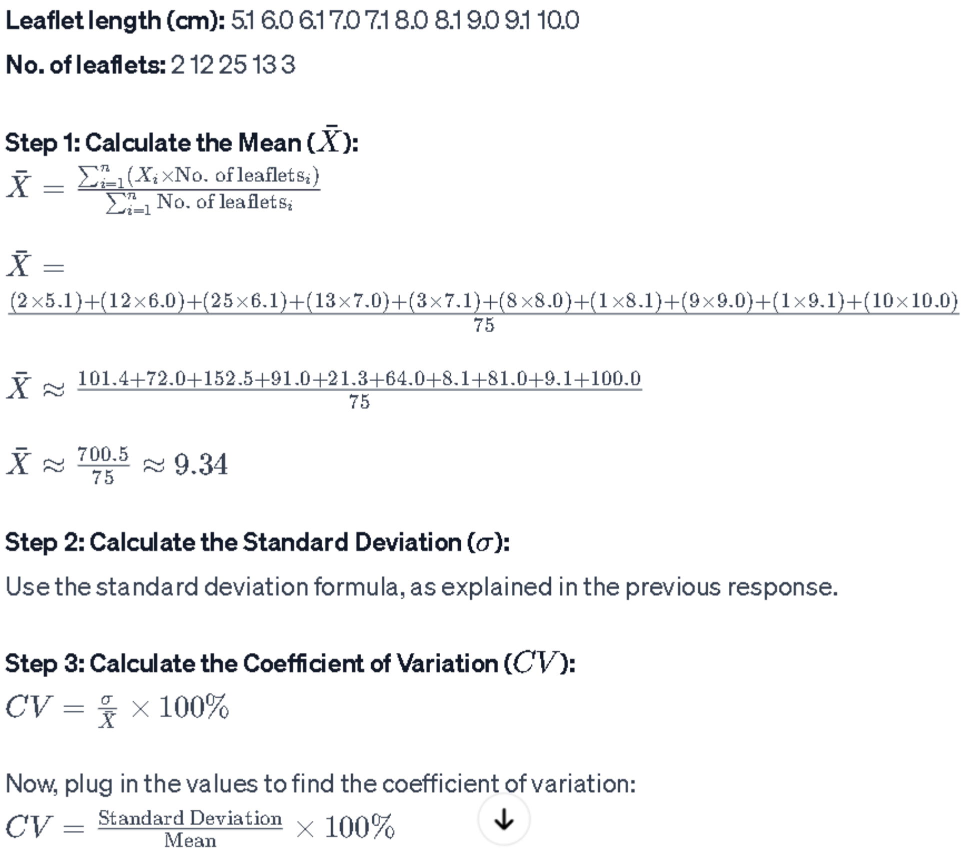
Step 2: Calculate the Standard Deviation (\(\sigma\)):

Use the formula for standard deviation as explained in the previous response.

Step 3: Calculate the Coefficient of Variation (\(CV\)):

\[CV = \frac{\sigma}{\bar{X}} \times 100\% \]

Now, plug in the values to find the mean, standard deviation, and coefficient of variation for the given leaflet length data set.



Leaflet length (cm): 5.1 6.0 6.1 7.0 7.1 8.0 8.1 9.0 9.1 10.0

No. of leaflets: 2 12 25 13 3

Step 1: Calculate the Mean ($\bar{X}$):

$$\bar{X} = \frac{\sum_{i=1}^{n}(X_i \times \text{No. of leaflets}_i)}{\sum_{i=1}^{n}\text{No. of leaflets}_i}$$

$$\bar{X} = \frac{(2\times5.1)+(12\times6.0)+(25\times6.1)+(13\times7.0)+(3\times7.1)+(8\times8.0)+(1\times8.1)+(9\times9.0)+(1\times9.1)+(10\times10.0)}{75}$$

$$\bar{X} \approx \frac{101.4+72.0+152.5+91.0+21.3+64.0+8.1+81.0+9.1+100.0}{75}$$

$$\bar{X} \approx \frac{700.5}{75} \approx 9.34$$

Step 2: Calculate the Standard Deviation (σ):

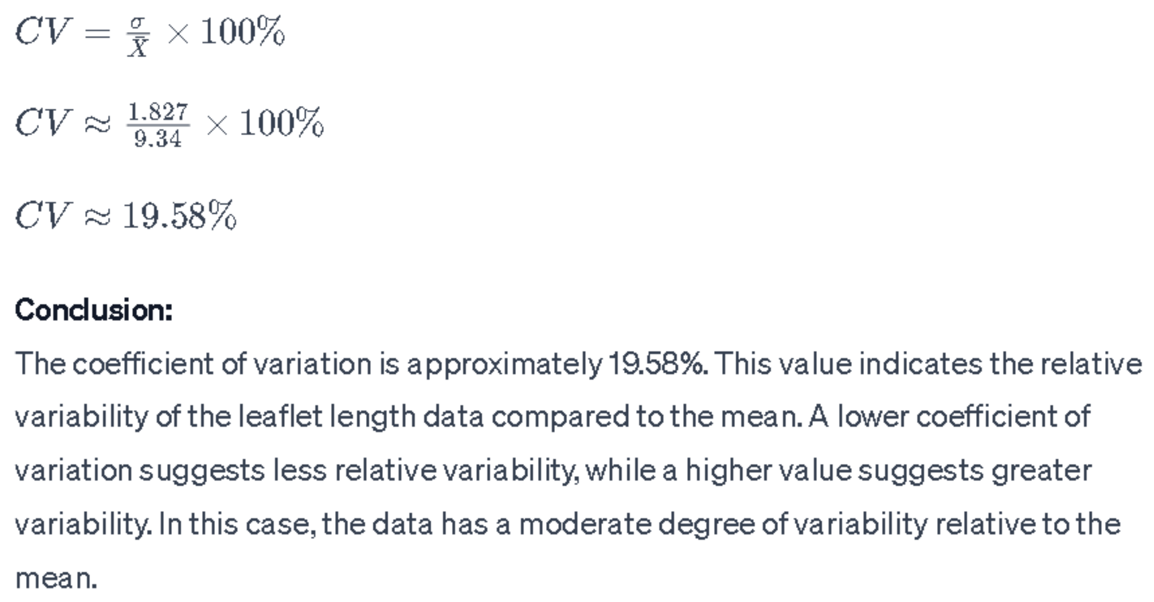
Use the standard deviation formula, as explained in the previous response.

Step 3: Calculate the Coefficient of Variation (CV):

$$CV = \frac{\sigma}{\bar{X}} \times 100\%$$

Now, plug in the values to find the coefficient of variation:

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$



$$CV = \frac{\sigma}{\bar{X}} \times 100\%$$

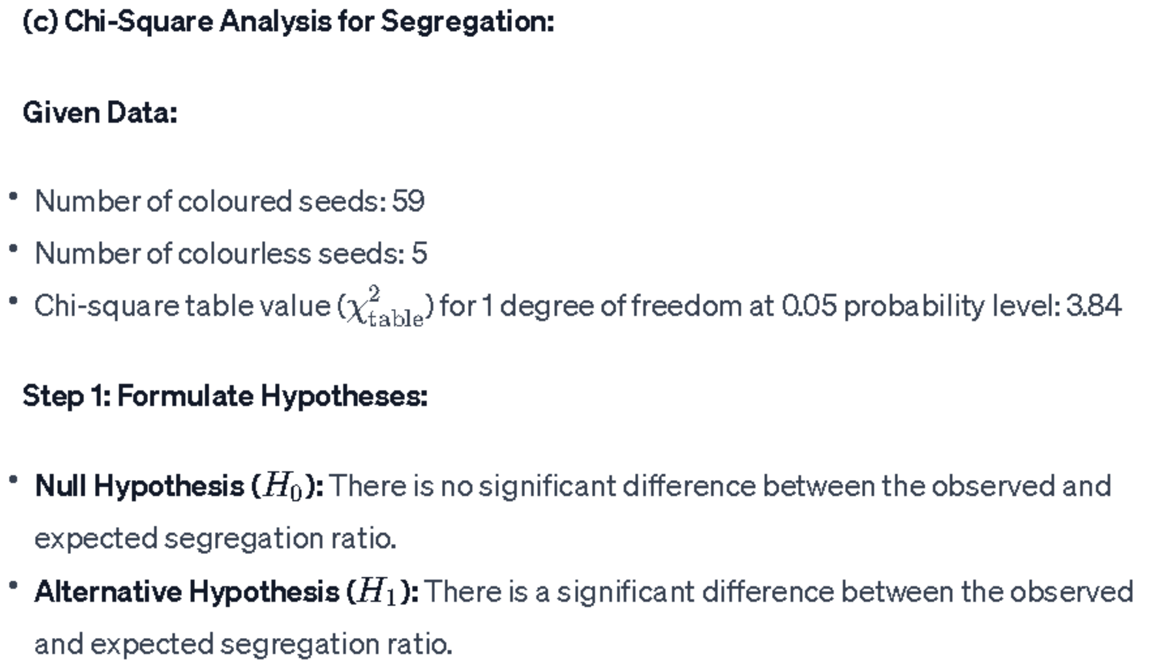
$$CV \approx \frac{1.827}{9.34} \times 100\%$$

$$CV \approx 19.58\%$$

Conclusion:

The coefficient of variation is approximately 19.58%. This value indicates the relative variability of the leaflet length data compared to the mean. A lower coefficient of variation suggests less relative variability, while a higher value suggests greater variability. In this case, the data has a moderate degree of variability relative to the mean.

(c) Selfing of a hybrid produced a population with 59 coloured and 5 colourless seeds. The chi-square table value is 3.84 for 1 degree of freedom at 0.05 probability level. Find out the segregation ratio and test the goodness of fit using the chi-square analysis. Comment on the nature of segregation



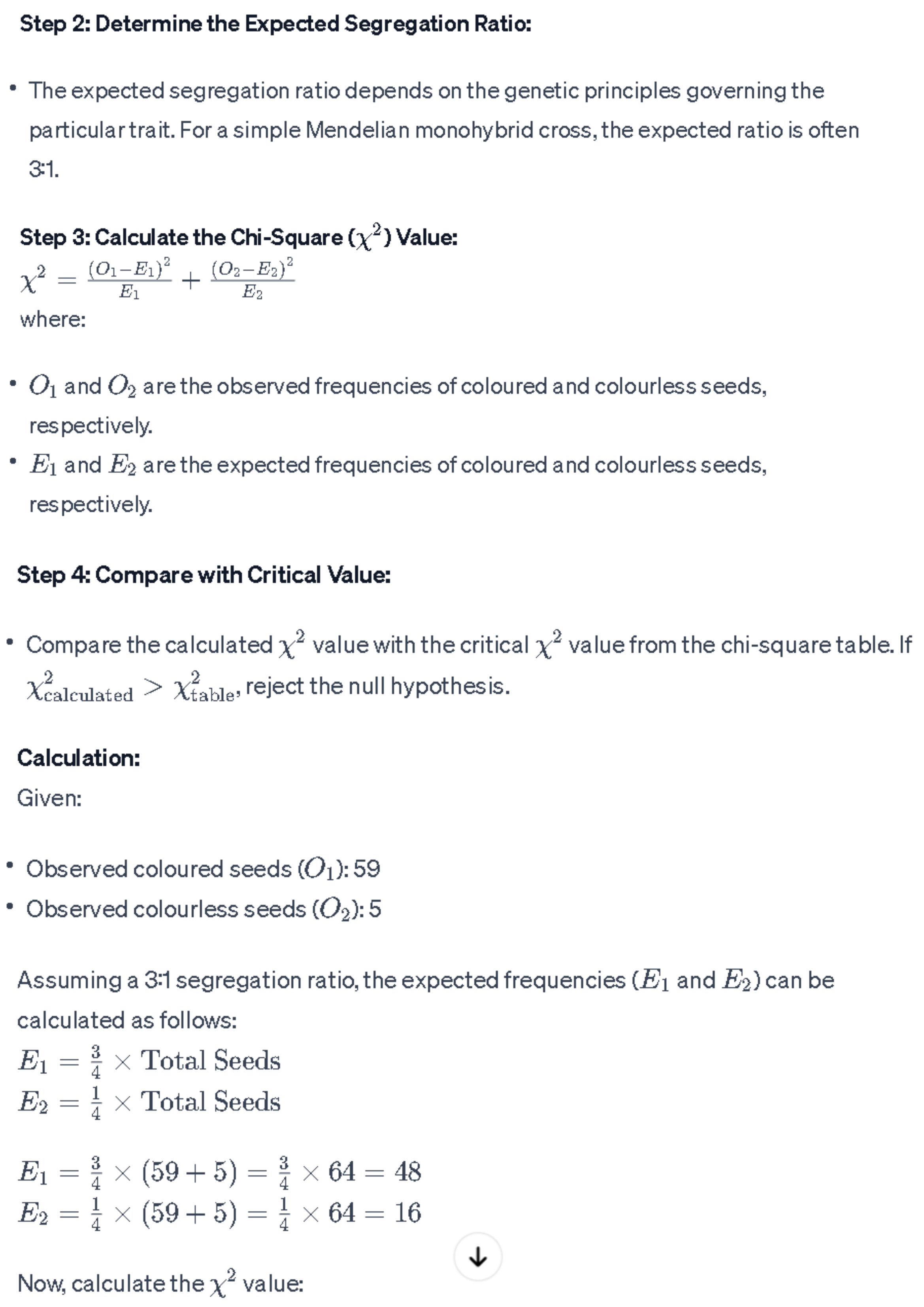
(c) Chi-Square Analysis for Segregation:

Given Data:

- Number of coloured seeds: 59
- Number of colourless seeds: 5
- Chi-square table value (χ^2_{table}) for 1 degree of freedom at 0.05 probability level: 3.84

Step 1: Formulate Hypotheses:

- **Null Hypothesis (H_0):** There is no significant difference between the observed and expected segregation ratio.
- **Alternative Hypothesis (H_1):** There is a significant difference between the observed and expected segregation ratio.



Step 2: Determine the Expected Segregation Ratio:

- The expected segregation ratio depends on the genetic principles governing the particular trait. For a simple Mendelian monohybrid cross, the expected ratio is often 3:1.

Step 3: Calculate the Chi-Square (χ^2) Value:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

where:

- O_1 and O_2 are the observed frequencies of coloured and colourless seeds, respectively.
- E_1 and E_2 are the expected frequencies of coloured and colourless seeds, respectively.

Step 4: Compare with Critical Value:

- Compare the calculated χ^2 value with the critical χ^2 value from the chi-square table. If $\chi^2_{\text{calculated}} > \chi^2_{\text{table}}$, reject the null hypothesis.

Calculation:

Given:

- Observed coloured seeds (O_1): 59
- Observed colourless seeds (O_2): 5

Assuming a 3:1 segregation ratio, the expected frequencies (E_1 and E_2) can be calculated as follows:

$$E_1 = \frac{3}{4} \times \text{Total Seeds}$$
$$E_2 = \frac{1}{4} \times \text{Total Seeds}$$

$$E_1 = \frac{3}{4} \times (59 + 5) = \frac{3}{4} \times 64 = 48$$
$$E_2 = \frac{1}{4} \times (59 + 5) = \frac{1}{4} \times 64 = 16$$

Now, calculate the χ^2 value:

(d) State the Hardy-Weinberg equation. Explain its utility in measuring gene frequency. MN blood types were tested in 100 people. The genotypic data was MM = 66, MN = 20 and NN = 14. Prove that the population is in Hardy-Weinberg equilibrium.

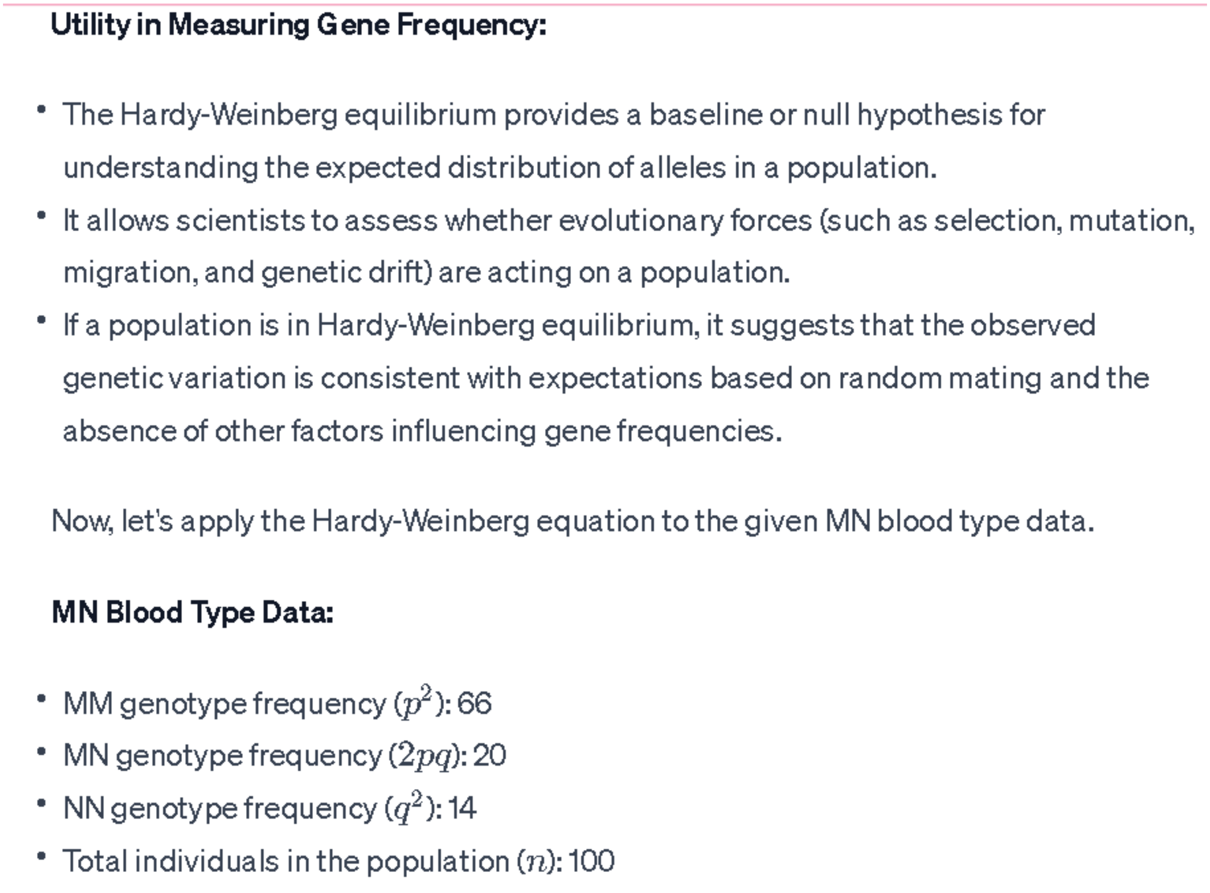
(d) Hardy-Weinberg Equation and its Utility:

Hardy-Weinberg Equation:
The Hardy-Weinberg equation is a mathematical expression that describes the relationship between the frequencies of alleles and genotypes in a population under certain conditions of no selection, no migration, no mutation, random mating, and a large population. The equation is represented as:

$$p^2 + 2pq + q^2 = 1$$

where:

- p^2 represents the frequency of the homozygous dominant genotype (AA),
- $2pq$ represents the frequency of the heterozygous genotype (Aa),
- q^2 represents the frequency of the homozygous recessive genotype (aa),
- The sum of these three terms equals 1, representing the total frequency of alleles in the population.



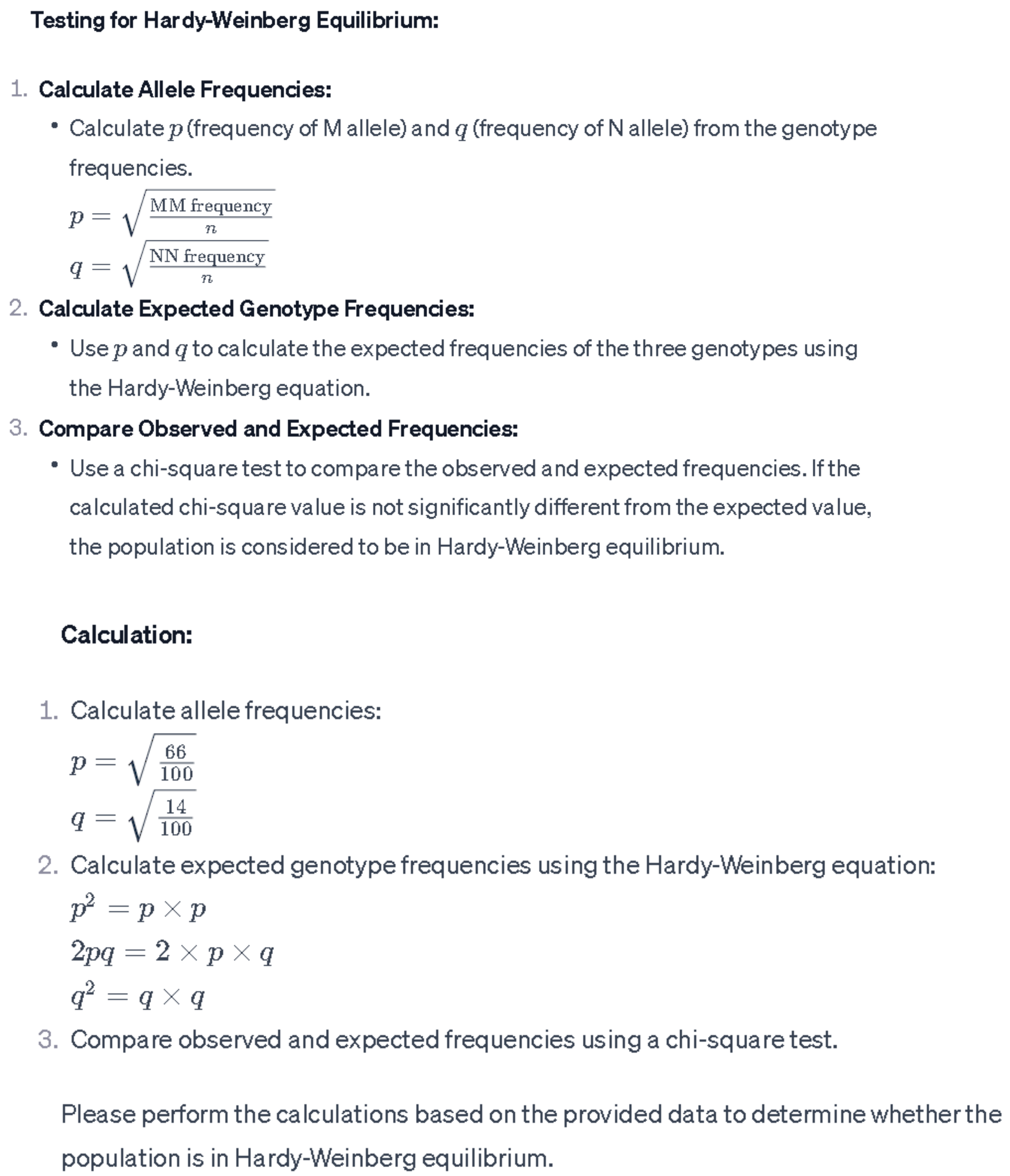
Utility in Measuring Gene Frequency:

- The Hardy-Weinberg equilibrium provides a baseline or null hypothesis for understanding the expected distribution of alleles in a population.
- It allows scientists to assess whether evolutionary forces (such as selection, mutation, migration, and genetic drift) are acting on a population.
- If a population is in Hardy-Weinberg equilibrium, it suggests that the observed genetic variation is consistent with expectations based on random mating and the absence of other factors influencing gene frequencies.

Now, let's apply the Hardy-Weinberg equation to the given MN blood type data.

MN Blood Type Data:

- MM genotype frequency (p^2): 66
- MN genotype frequency ($2pq$): 20
- NN genotype frequency (q^2): 14
- Total individuals in the population (n): 100



Testing for Hardy-Weinberg Equilibrium:

1. **Calculate Allele Frequencies:**
   - Calculate p (frequency of M allele) and q (frequency of N allele) from the genotype frequencies.
     $p = \sqrt{\frac{\text{MM frequency}}{n}}$
     $q = \sqrt{\frac{\text{NN frequency}}{n}}$
2. **Calculate Expected Genotype Frequencies:**
   - Use p and q to calculate the expected frequencies of the three genotypes using the Hardy-Weinberg equation.
3. **Compare Observed and Expected Frequencies:**
   - Use a chi-square test to compare the observed and expected frequencies. If the calculated chi-square value is not significantly different from the expected value, the population is considered to be in Hardy-Weinberg equilibrium.

Calculation:

1. Calculate allele frequencies:
   $p = \sqrt{\frac{66}{100}}$
   $q = \sqrt{\frac{14}{100}}$
2. Calculate expected genotype frequencies using the Hardy-Weinberg equation:
   $p^2 = p \times p$
   $2pq = 2 \times p \times q$
   $q^2 = q \times q$
3. Compare observed and expected frequencies using a chi-square test.

Please perform the calculations based on the provided data to determine whether the population is in Hardy-Weinberg equilibrium.

(e) Explain the addition and multiplication rules of probability. What is the probability of getting a king or a joker from a pack of 52 cards (with 4 kings and 2 jokers)? What is the combined probability of getting a king in 4 consecutive drawals from this pack of cards without replacing cards after each drawal?

(e) Addition and Multiplication Rules of Probability:

Addition Rule:

The addition rule of probability is used to calculate the probability of the occurrence of either of two mutually exclusive events. For two events, A and B, the rule is expressed as:

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

where $P(A \text{ and } B)$ represents the probability of both events A and B occurring together.

Multiplication Rule:

The multiplication rule of probability is used to calculate the probability of the occurrence of two or more independent events. For two independent events, A and B, the rule is expressed as:

$P(A \text{ and } B) = P(A) \times P(B)$

If there are more than two events, the multiplication rule extends to:

$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C)$

and so on.

Now, let's apply these rules to the card problem:

Probability of Getting a King or a Joker:

- Number of Kings = 4
- Number of Jokers = 2
- Total Number of Cards = 52

$$P(\text{King or Joker}) = P(\text{King}) + P(\text{Joker}) - P(\text{King and Joker})$$

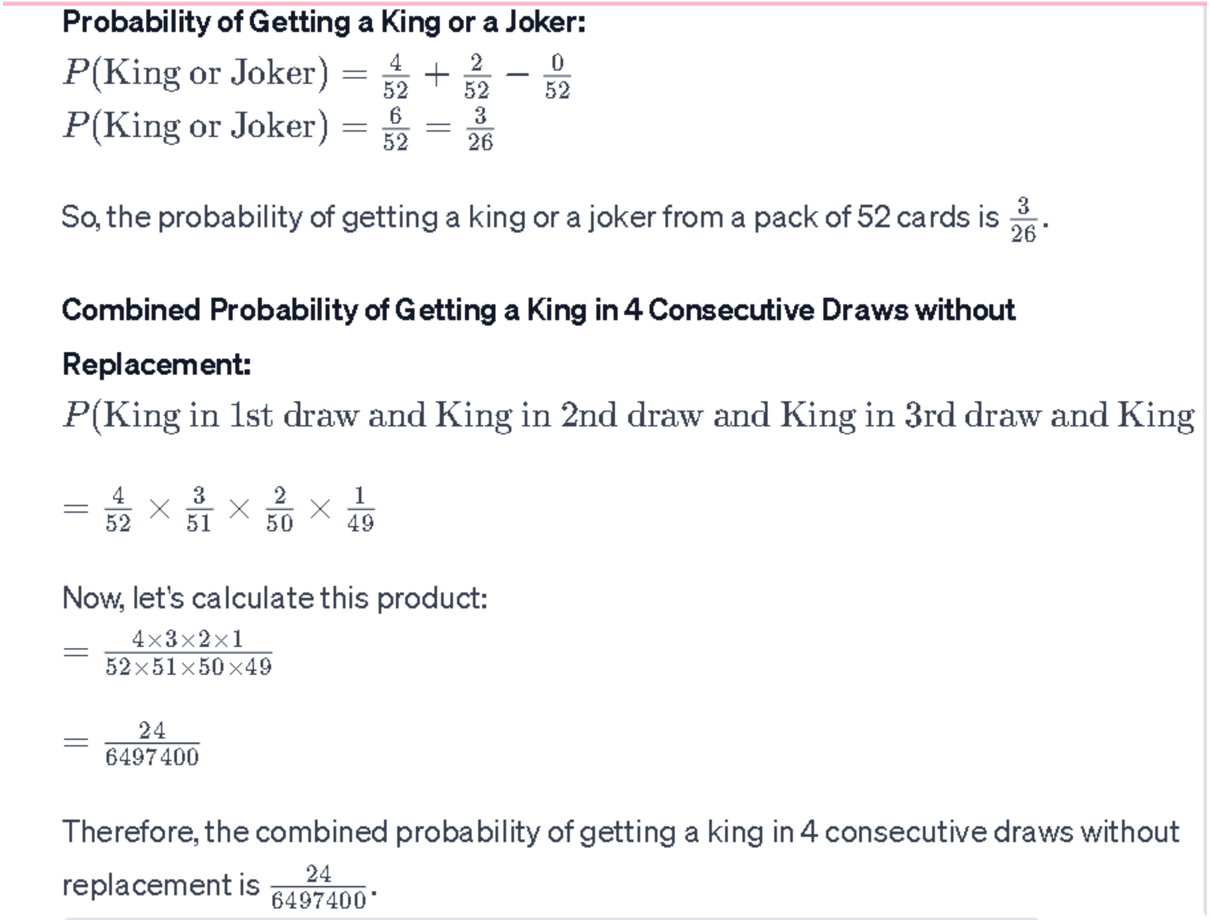
$$P(\text{King or Joker}) = \frac{4}{52} + \frac{2}{52} - \frac{0}{52}$$
$$P(\text{King or Joker}) = \frac{6}{52} = \frac{3}{26}$$

Combined Probability of Getting a King in 4 Consecutive Draws without Replacement:
When cards are drawn without replacement, the events are not independent. The probability changes with each draw.

$$P(\text{King in 1st draw and King in 2nd draw and King in 3rd draw and King i}$$

$$= P(\text{King in 1st draw}) \times P(\text{King in 2nd draw after the 1st draw}) \times$$
$$P(\text{King in 3rd draw after the 2nd draw}) \times$$
$$P(\text{King in 4th draw after the 3rd draw})$$

$$= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}$$



Probability of Getting a King or a Joker:

$$P(\text{King or Joker}) = \frac{4}{52} + \frac{2}{52} - \frac{0}{52}$$
$$P(\text{King or Joker}) = \frac{6}{52} = \frac{3}{26}$$

So, the probability of getting a king or a joker from a pack of 52 cards is $\frac{3}{26}$.

Combined Probability of Getting a King in 4 Consecutive Draws without Replacement:

$$P(\text{King in 1st draw and King in 2nd draw and King in 3rd draw and King}$$

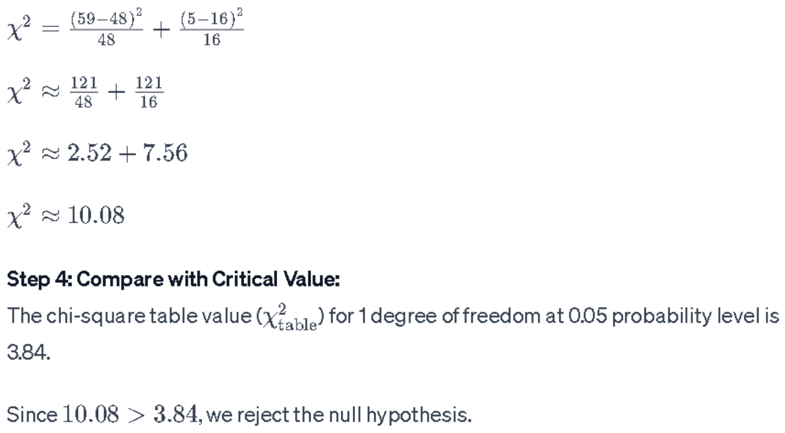
$$= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49}$$

Now, let's calculate this product:

$$= \frac{4 \times 3 \times 2 \times 1}{52 \times 51 \times 50 \times 49}$$

$$= \frac{24}{6497400}$$

Therefore, the combined probability of getting a king in 4 consecutive draws without replacement is $\frac{24}{6497400}$.



$$\chi^2 = \frac{(59-48)^2}{48} + \frac{(5-16)^2}{16}$$

$$\chi^2 \approx \frac{121}{48} + \frac{121}{16}$$

$$\chi^2 \approx 2.52 + 7.56$$

$$\chi^2 \approx 10.08$$

Step 4: Compare with Critical Value:

The chi-square table value (χ^2_{table}) for 1 degree of freedom at 0.05 probability level is 3.84.

Since $10.08 > 3.84$, we reject the null hypothesis.

Conclusion: The observed segregation ratio significantly differs from the expected 3:1 ratio. The nature of segregation suggests that there may be a deviation from the expected Mendelian inheritance pattern for the particular trait under consideration. Further genetic analysis or investigation may be warranted to understand the underlying factors influencing the observed segregation.