DMPA ASSIGNMENT-2.

Q1) Answer the following :-

a) What do you mean by hypothesis testing for some parameter in a population ?

ans) Hypothesis testing is a statistical method used to make decision about a population parameter (like mean or proportion) based on sample data.

It begins with two statements :

- Null hypothesis (H_0) : The parameter has a specific value.
- Alternative Hypothesis (H_1) : The parameter differs from that value.

Sample evidence is used to decide whether to reject the null hypothesis.

b) How is a confidence interval used for hypothesis testing?

ans) A confidence interval provides a range of plausible values for a population parameter.

To test a hypothesis :

- If the hypothesized parameter value (H_0) lies inside the confidence interval → Fail to reject H_0
- If it lies outside the interval → Reject H_0

Q2) The duration of customer service calls to an insurance company is normally distributed, with mean 20 minutes and standard deviation 5 minutes. For the following sample sizes, construct a 95% confidence interval for the population mean duration of customer service calls.

a) $n = 25$
b) $n = 100$
c) $n = 200$

ans) $CI = \bar{x} \pm Z_{0.025} \frac{6}{\sqrt{n}}$

$z = 1.96$

Population mean = 20

a) $n = 25$

$$SE = \frac{5}{\sqrt{25}} = 1$$

$$CI = 20 \pm 1.96\ (1) = (18.04, 21.96)$$

b) $n = 100$

$$SE = \frac{5}{10} = 0.5$$

$$CI = 20 \pm 1.96\ (0.5) = (19.02, 20.98)$$

c) $n = 200$

$$SE = \frac{5}{\sqrt{200}} = 0.3536$$

$$CI = 20 \pm 1.96\ (0.3526) = (19.31, 20.69)$$

Name:

Q4) Discuss about Type I and Type II errors in the context of hypothesis testing with an example.

ans) **Type I Error** (False Positive)

Rejecting a true null hypothesis.

Example: Concluding a customer churn rate has increased when it actually has not.

Type II Error (False Negative)

Failing to reject a false null hypothesis.

Example: Concluding churn rate has NOT increased when in reality it has.

Q5) In the churn problem discussed in Chapter 3, recall that 483 of 3333 customers in the sample had churned the company. Using level of significance $\alpha = 0.10$, test whether the population proportion π differs from 0.15 by computing the p-value from the Z data.

ans)

$\hat{p} = 0.1449$

$H_0 : p = 0.15$

$H_1 : p \neq 0.15$

$$z = \frac{\hat{p} - 0.15}{\sqrt{\frac{0.15(1-0.15)}{3333}}} = -0.8246$$

Given: $P(z > 0.8246) = 0.2048$

Two tailed p-value:

$p = 2(0.2048)$

$= 0.4096$

Since $p > 0.10 \rightarrow$ Fail to reject H_0

No evidence that actual churn proportion differs from 0.15.

Name:_____

Regd. Number:_____

Q3) Of 1000 customers who received promotional materials for a marketing campaign, 100 responded to the promotion. For each of the confidence levels, construct a confidence interval for the population proportion who would respond to the promotion.

a) 90% b) 95% c) 99%

ans) $n = 1000$

$\hat{p} = 0.10$ [100 responded]

$$CI = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$SE = \sqrt{\frac{0.1(0.9)}{1000}} = 0.00949$$

a) 90% $\rightarrow$ $z = 1.645$

$$CI = 0.1 \pm 1.645\,(0.00949)$$
$$= (0.0844, 0.1156)$$

b) 95% $\rightarrow$ $z = 1.96$

$$CI = 0.1 \pm 1.96\,(0.00949)$$
$$= (0.0814, 0.1186)$$

c) 99% $\rightarrow$ $z = 2.576$

$$CI = 0.1 \pm 2.576\,(0.00949)$$
$$= (0.0755, 0.1245)$$

Regd. Number:_____

Q6) A sample of 100 donors to a charity has a mean donation amount of $55 with a sample standard deviation of $25. Test using $\alpha = 0.05$ whether the population mean donation amount exceeds $50.

a) Define the null hypothesis and the alternatives hypothesis on μ.

ans) $n = 100$

$$\bar{x} = 55, s = 25, \alpha = 0.05$$

We test:

$$H_0 : \mu = 50$$

$$H_1 : \mu > 50$$

b) What is the rejection rule ?

ans) Reject H_0 if

$$T > t_{0.05, 99} = 1.660$$

c) What is the meaning of the test statistic T ?

ans) T measures how many standard errors the sample mean is above the hypothesized mean.

d) What are the values of the test statistic T_{data} and the p-value in this example ?

ans)
$$T = \frac{55 - 50}{25 - \sqrt{100}} = \frac{5}{2.5} = 2$$

$$T = 2 > 1.660 \rightarrow p < 0.05.$$

e) What is our conclusion after comparing the p-value with the level of significance ?

ans) Reject H_0.

f) Interpret our conclusion so that non-specialist could understand it.

ans) There is a strong evidence that average donation amount is greater than $50.

Q7) Test whether the partition is valid for this variable using $\alpha = 0.10$.

ans) Training

$$\bar{x}_1 = 20.5, \quad s_1 = 5.2, \quad n_1 = 2000$$

Test

$$\bar{x}_2 = 20.4, \quad s_2 = 4.9, \quad n_2 = 600$$

Test statistics

$$T = 0.4322$$

Given:

$$P(T > 0.4322) = 0.3328$$

Since $p = 0.3328 > 0.10 \longrightarrow$ fail to reject H_0

The partition is valid for this variable.

Q9) Suppose a multinomial variable "movie choices" has the following values {'Romantic', 'Science fiction'}. We have a set of 1000 males and a set of 250 females with the following frequencies:

	Romantic	Scientific fiction	Total
Male	350	650	1000
Female	175	75	250
Total	525	725	1250

Test whether significant differences exist between the multinomial proportions of the two gender groups, given that $P(\chi^2 < 102.17) = 0.9999$

ans) Romantic Proportion = 525/1250 = 0.42

Sci-fi proportion = 725/1250 = 0.58

Expected

Male: (420, 580)

Female: (105, 145)

$\chi^2 = 102.17$

Given: $P(\chi^2 < 102.17) = 0.9999$

Meaning χ^2 is extremely large → proportion differ significantly.

Conclusion: Movie preferences differ significantly between male and female groups.

Name:

Regd. Number:

Q8) The multinomial variable payment preference takes the values credit card, debit card, and cheque. Now, suppose we know that 50% of the customers in our population prefer to pay by credit card, 20% by debit card and 30% to pay by cheque. We have taken a sample of size 200 shows 125 customers preferring to pay by credit card, 25 by debit card and 50 by cheque. Test whether the sample is representative of the population using $\alpha = 0.05$. Given that $P(\chi^2 > 38.82) = 10^{-6}$.

ans) Population Proportion : (0.5, 0.2, 0.3)

Sample counts : (125, 25, 50), $n = 200$

Expected :

(100, 40, 60)

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 10.416$$

Given : $P(\chi^2 > 38.82) = 10^{-6}$

Since 10.416 is far below this threshold → fails to reject H_0

Sample is representative of the population.

Q10) Describe the difference between the training set, test set, and validation set.

ans)

Dataset	Purpose	When used	What helps
• Training set	To train the model by learning parameters	During model building	Helps the model learn patterns and relationships in the data.
• Validation set	To tune hyperparameter and choose the best model	During model selection, after training	Prevents overfitting assists in choosing complexity and settings.
• Test set	To evaluate final model performance	Only after the model is fully trained and tuned	Measures generalization on unseen data, gives unbiased performance estimate.

Q11) How is the bias-variance tradeoff related to the issue of overfitting and underfitting? Is high bias associated with overfitting and underfitting and why?

ans) The bias-variance trade-off explains how model complexity affects prediction accuracy.

- Bias is the error that occurs when a model is too simple and cannot capture the true pattern of the data. High bias leads to underfitting because the model makes strong assumptions and misses important relationships.

Variance is the error that occurs when a model is too complex and becomes highly sensitive to the training data.

High variance leads to overfitting where the model fits the noise in the training data and performs poorly on new data.

Thus:

- Underfitting → High bias, Low variance
- Overfitting → Low bias, High variance

Therefore, high bias is associated with underfitting, not overfitting. because a high-bias model is too simple to learn the underlying structure of data.