Final - 2023

Q2.

(a) The data is summarized in the following 2 x 2 contingency table:

Contraceptive practice	Myocardial infarction		Total
	Yes	No	
Users	23 (a)	34 (b)	57 (a+b)
Non-users	35 (c)	132 (d)	167 (c+d)
Total	58 (a+c)	166 (b+d)	224 (n)

a) Difference of proportion (Risk difference, RD):

Incidence among users, $\hat{P}_1 = \frac{a}{a+b} = \frac{23}{57}$,

$= 0.4035$

Incidence among non-users, $\hat{P}_2 = \frac{c}{c+d}$

$= \frac{35}{167}$

$= 0.2096$

∴ Difference of Proportion $= \hat{P}_1 - \hat{P}_2$

$= 0.4035 - 0.2096$

$= 0.1939$

Comment: The positive risk difference (0.194) indicates that the absolute risk of myocardial infarction is about 19.4 percentage points higher among oral contraceptive users compared to non-users.

Relative Risk

$$RR = \frac{\hat{P}_1}{\hat{P}_2} = \frac{0.4035}{0.2096}$$

$$= 1.925$$

Comment: The relative risk (1.925) indicates that oral contraceptive users are about 1.9 times more likely to experience myocardial infarction compared to non-users.

Odds ratio:

$$OR = \frac{a/b}{c/d} = \frac{23/34}{35/132}$$

$$= 2.551$$

Comments: the odds of Myocardial infarction are 2.55 times larger in users than in non-users.

(b) The null hypothesis (H_0) in all three cases is that there is no association between oral contraceptive use and mypcardial infarction. This implies

$H_0: RD = 0$

$H_0: RR = 1$

$H_0: OR = 1$

Test for difference of Proportions:

$$S.E = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Here,
$\hat{P}_1 = 0.4035; \hat{P}_2 = 0.2096$
$n_1 = 57, \; n_2 = 167$

$$S.E(RD) = \sqrt{\frac{0.4035(1-0.4035)}{57} + \frac{0.2096(1-0.2096)}{167}}$$

$$= \sqrt{0.7946} = \sqrt{0.005214}$$

$$= 0.892 \approx 0.0722$$

$\therefore$ Test statistic, $z = \frac{0.194}{0.892} \quad \frac{0.194}{0.0722}$

$$\approx 2.69$$

This z-value is compared to the standard normal distribution. $|z| > 1.96$ leads to rejection of H_0 at a 5% significance level. i.e. The difference is statistically significant.

Test of Relative Risk (RR)

$$\ln(RR) = \ln(1.925)$$

$$= 0.65499 \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(\hat{p}(1-\hat{p})}{n}}$$

$$SE(\ln(RR)) = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+c} - \frac{1}{c+d}}$$

$$= \sqrt{\frac{1}{23} + \frac{1}{35} - \frac{1}{57} - \frac{1}{167}}$$

$$\simeq \sqrt{0.048517}$$

$$\simeq 0.220$$

Test statistic, $z = \frac{0.6549}{0.220}$

$$\simeq 2.97$$

$\therefore$ 2.97 > 1.96, so RR significantly different from 1.

Test for odds ratio:

$$\ln(OR) = \ln(2.551) = 0.9365$$

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$= \sqrt{\frac{1}{23} + \frac{1}{34} + \frac{1}{35} + \frac{1}{132}}$$

$$= \sqrt{0.1091} = 0.330$$

Test statistic, $z = \frac{0.9365}{0.330}$

≈ 2.84

∴ Since $2.84 > 1.96$ OR significantly different from 1

∴ All three test indicate a statistically significant association between oral contraceptive use and myocardial infarction at conventional $\alpha = 0.05$ is ± 1.96

(c) 95% CI for Risk Difference (RD)

$$SE(RD) = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{a+b} + \frac{\hat{P}_2(1-\hat{P}_2)}{c+d}}$$

$$= \sqrt{\frac{(0.4035 \times 0.5965)}{57} + \frac{(0.2096 \times 0.7904)}{167}}$$

$$= \sqrt{0.00521}$$

$$= 0.0721$$

$\therefore$ 95% CI = RD $\pm$ 1.96 $\times$ SE(RD)

$= 0.1939 \pm 1.96 \times 0.0722$

$= 0.1939 \pm 0.1415$

$\therefore$ CI = (0.0524, .3354)

Comment: The 95% CI for the RD (0.0524, .3354) does not include zero. this means we are 95% confident that the true absolute increase in risk for users lies between 5.2% and 33.5% percentage points.

95% CI for Relative Risk (RR):

$\ln(RR) \pm 1.96 \times SE(\ln(RR))$

$= 0.6549 \pm 1.96 \times 0.220$

$= 0.6549 \pm 0.4312$

$= (0.224, 1.086)$

$\therefore$ CI for RR $= (e^{0.224}, e^{1.086})$

$= (1.251, 2.962)$

Comment: The 95% CI for the ~~Odds Ratio~~ RR (1·25, 2·96) does not include 1. this means we are 95% confident that the true relative risk is between 1·25 and 2·96

95% CI for odds Ration (OR):

$$\ln(OR) \pm 1.96 \times SE(\ln(OR))$$

$$= 0.9365 \pm 1.96 \times 0.330$$

$$= 0.9365 \pm 0.6468$$

$$= (0.2897, 1.5833)$$

$\therefore$ CI for $OR = (e^{0.2897}, e^{1.5833})$

$= (1.336, 4.871)$

Comment: The 95% CI for the OR (1·336, 4·871) does not include 1. this means we are 95% confident that the true odds ratio is between 1·336 and 4·871.

Q3.

(a) Computing Confidence Intervals for Association parameters:

Association parameters in contingency tables measures the strength and direction of association between categorical variables.

To compute a ~~cond~~ confidence interval for an odds ratio (OR). ~~for example~~:

For a 2×2 table with cells

	Yes	No
Row 1	a	b
Row 2	c	d

$$OR = \frac{(a/b)}{(c/d)} = \frac{ad}{bc}$$

1. Calculate the odds ratio:

$$OR = \frac{(a/b)}{(c/d)} = \frac{ad}{bc}$$

2. Compute the standard error (SE) of the natural log of OR:

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

3. A $100(1-\alpha)\%$ confidence interval for $\ln(OR)$ is: $\ln(OR) \pm 1.96 \times SE(\ln(OR))$

Now, Back-transformation to obtain the CI for OR

$$CI = \left(e^{\ln(OR) - Z_{\alpha/2} SE(\ln(OR))}, e^{\ln(OR) + Z_{\alpha/2} \cdot SE(\ln(OR))} \right)$$

Importance of small sample tests in Contingency Tables:

d	c	Row 1
b	a	Row 2

i) They provide accurate p-value without relying on large-sample approximations.

ii) They avoid Type I errors that can occur with chi-square tests in small samples.

iii) They are essential for sparse tables or stratified analyses with limited data.

(b) The data are stratified by income level:

1. Low income stratum:

Education	Yes	No	Total
High school	25	15	40
College	35	25	60
Total	60	40	100

2. Medium income stratum:

Education	Yes	No	Total
High school	40	20	60
College	50	30	80
Total	90	50	140

The Mantel-Haenszel estimate of common odds ratio is

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

where a_i, b_i, c_i, d_i are the cell counts and N_i is the total for stratum i:

For Low income ($i = 1$):

$a_1 = 25, \quad b_1 = 15, \quad c_1 = 35, \quad d_1 = 25, \quad N_1 = 100$

$$\frac{a_1 d_1}{N} = \frac{25 \times 25}{100}$$

$$= 6.25$$

$$\frac{b_1 c_1}{N} = \frac{15 \times 35}{100}$$

$$= 5.25$$

For medium income ($i = 2$):

$a_2 = 40, \quad b_2 = 20, \quad c_2 = 50, \quad d_2 = 30, \quad N_2 = 140$

$$\frac{a_2 d_2}{N} = \frac{40 \times 30}{140} = 8.572$$

$$\frac{b_2 \times c_2}{N} = \frac{20 \times 50}{140} = 7.143$$

$$\therefore \quad OR_{MH} = \frac{6.25 + 8.572}{5.25 + 7.143} = 1.20$$

Interpretation:

Since $OR_{MH} = 1.20 > 1$, it indicates that across both income levels, individuals with college education have higher odds of supporting the policy compared to those with high school education. The effect is consistent in both strata. To formally test significance the Mantel-Haenszel chi-square test can be used but the OR_MH value alone suggests a positive association.

(c) Bayesian inference for categorical data involves the following steps:

1. Specify a Prior Distribution:

For multinomial data, the Dirichlet distribution is used as a conjugate prior for the cell probabilities. For binomial proportions, ~~Beta~~ Beta distributions are used.

2. Likelihood function:

The likelihood is typically multinomial for count data.

3. Posterior Distribution via Bayes' Theorem:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

The posterior distribution combines prior knowledge with observed data.

4. Computation:

For simple models, the posterior is analytic.

For complex models, Markov chain Monte Carlo (MCMC) methods are used to simulate from the posterior.

5. Inference:

Use the posterior to compute point estimates (posterior mean, median), credible intervals (e.g. 95% credible interval) and predictive probabilities.

Q1. (a) Categorical variable: A categorical variable is a variable whose values represent categories, labels or groups, not numerical magnitudes. The values indicate types or class and arithmetic operations are not meaningful for these variables.

Examples:

Gender: (Male, Female)

Blood group: (A, B, AB, O)

Marital status: (single, married, divorced)

Type of residence: (Rural, Urban.

Major types of categorical data:

1. Nominal data;
2. Ordinal data;
3. Binary data:

Nominal data: Nominal data consist of categories that have no natural order or ranking.

The characteristics of nominal data is:

i) Categories are names or labels only.

(ii) Differences between ranks are not equal or known.

Examples:

i) Categories are names or labels only.

ii) No ordering among the categories.

Examples:

Gender: Male, Female

Blood group: A, B, AB, O

Religion: Islam, Hinduism, Christianity

Ordinal data:

Ordinal data consist of categories that can be arranged in a meaningful order, but the difference between successive categories cannot be measured numerically.

The characteristics of ordinal data is:

i) Categories have a logical order.

ii) Differences between ranks are not equal or known.

Examples:

Educational qualification: Primary, Secondary, Graduate

Customer satisfaction: Poor, average, good, excellent

Socio-economic status: low, middle, high

Binary data: Binary data are a special type of categorical data that contain only two possible categories.

The characteristics of binary data is:

i) Exactly two categories.

ii) May be nominal or ordinal in nature.

Examples:

yes/no, Pass/fail, success/failure,

disease status: Present/absent

1|(b): A contingency table is a tabular method used to summarize and analyze categorical data by showing the frequency distribution of two or more categorical variables simultaneously. When two categorical variables are involved, the table is called a two-way contingency table.

In a two-way contingency

Rows represent the categories of one categorial variables.

Columns represents the categories of another categorical variables.

Each cell shows the frequency for the corresponding combination of categories.

Consider

Example:

Consider two categorical variables:

1. Test result: {Positive, negative}
2. Disease status: {Disease, Non-disease}

Test result	Disease	Non-disease	Total
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Explanation:

Here,

a: number of individuals who have the disease and test positive

b: number of individuals who do not have the disease but test positive.

c: number of individuals who have the disease but test negative.

d: number of individuals who do not have the disease and test negative.

Row totals and column totals gives the marginal distributions

Each cell represents the frequency corresponding to a particular combination of categories. The row and column totals provide marginal distributions.

1/c(ii) Given,

Estimated regression coefficient, $\hat{\beta}_1 = 0.98$

Standard error, $SE(\hat{\beta}_1) = 0.2$

Hypotheses:

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

The Wald test statistic is

$$W = \left(\frac{\hat{\beta}_1}{S.E(\hat{\beta}_1)}\right)^2$$

$$W = \left(\frac{0.8}{0.2}\right)^2$$

$$= 16$$

Under H_0, $W \sim \chi^2_1$

At 5% level of significance level,

$$\chi^2_{0.05,1} = 3.84$$

Since $16 > 3.84$, we reject H_0

Conclusion:

The coefficient $\hat{\beta}_1$ is statistically significant at the 5% level. Hence x_1 has a significant effect on the response variable.

(iii) Likelihood Ratio Test:

Given,

log-likelihood of reduced model (with x_1, x_2):

$l_0 = -102.4$

log-likelihood of full model (with x_1, x_2, x_3):

$l_1 = -98.6$

(ii) Likelihood Ratio Test

Log-likelihood of reduced model: $l_0 = -102.4$

Log-likelihood of full model: $l_1 = -98.6$

The likelihood ration (LR) test statistic is

$$LR = -2(l_0 - l_1)$$

$$= -2(-102.4 + 98.6)$$

$$= -2 \times -3.8$$

$$= 7.6$$

Under H_0, $LR \sim \chi^2_1$

At 5% level, $\chi^2_{0.05,1} = 3.84$

Since $7.6 > 3.84$, H_0 is rejected

~~The~~ Conclusion:

The prediction X_3 significantly improves the model. Hence adding X_3 leads to be a better fitting logistic regression model.

Final-2022

1/a) Types of response variables used in regression-Type models

The choice of regression model depédent depends on the nature of the response (dependent) variable. Common types are:

i) continuous response:

The response variable takes real numerical values.

Model: Linear regression.

Example: Income, height, GDP

ii) Binary response:

The response variable has two categories.

Model: Logistic regression.

Example: Disease/No-disease, Yes/No

iii) Count response:

The response variable represents counts.

Model: Poisson or Negative binomial regression.

Example: Number of accidents, number of visits.

iv) Ordinal response:

The response variable has ordered categoies.

Hypothesis testing in a Multiple Regression model

Consider a multiple regression model with three parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

We test the hypothesis

$$H_0 : \beta_1 = \beta_2 \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_2$$

This can be written as a linear restriction:

$$H_0 : \beta_1 - \beta_2 = 0$$

Wald Test:

The Wald test is based on the unrestricted estimates of the parameters. The Wald statistic is

$$W = \left(\frac{(\hat{\beta}_1 - \hat{\beta}_2)}{S.E(\hat{\beta}_1 - \hat{\beta}_2)} \right)^2$$

where,

$$S.E(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta})}$$

Under H_0

$$W \sim \chi_1^2$$

If W exceeds the critical value of χ_1^2 at the chosen significance level, H_0 is rejected.

Lagrange Multiplier (LM) Test:

The Lagrange multiplier (LM) test is based on the restricted model, estimated under $H_0: \beta_1 = \beta_2$

The LM test statistic is

$$LM = S' I^{-1} S$$

where S is the score vector evaluated at the restricted estimates and I is the information matrix.

Under H_0,

$$LM \sim \chi_1^2$$

If the LM statistic exceeds the critical value, H_0 is rejected.

1/b] Suppose the observations $Y_1, \ldots, Y_n$ follows a Poisson distribution. Two nested models are considered:

Model A (reduced model): one parameter λ_1

Model B (full model): two parameters λ_1, λ_2

The objective is to test whether inclusion of the second parameter improves the model fit

Hypotheses

$H_0: \lambda_2 = 0$ (Model A is adequate)

$H_1: \lambda_2 \neq 0$ (Model B is better)

Likelihood Ratio Test

Let,

~~$l_0 = \ln L_0$~~ b

$l_0 = \ln L_0$ be the log-likelihood under Model A,

$l_1 = \ln L_1$ be the log-likelihood under Model B.

The likelihood ratio test statistic is

$$LR = -2(l_0 - l_1)$$

$$= 2(l_1 - l_0)$$

Under the null hypothesis H_0, the test statistic has an asymptotic chi-square distribution with degrees of freedom equal to the difference in the number of parameters.

$$LR \sim \chi^2_1$$

Decision Rule:

At significance level α, compare the calculated value of LR with the critical value $\chi^2_{\alpha,1}$:

If $LR > \chi^2_{\alpha,1}$, reject H_0, otherwise do not reject H_0.

2/a] Spearman's rank correlation and Kendall's tau-b are non-parametric measures of association used for ordinal data. The following assumptions are required for their measurement.

Assumptions for Spearman's Rank correlation (r_s)

i) The variables are measured on an ordinal scale (or can be meaningfully ranked)

ii) Observations are independent of each other.

iii) The relationship between the two variables is monotonic (either consistently increasing or decreasing)

(iv) The data need not follow a normal distribution.

(v) Tied ranks may occur; appropriate tie-correlation correlations can be applied.

Assumptions for Kendall's Tau-b (τ_b):

(i) The variables are measured on an ordinal scale.

(ii) Observations are independent.

(iii) The association between variables is monotonic.

(iv) Kendall's tau-b explicitly accounts for tied

observations in both variables.

(v) No assumption of normality or linearity is required

2/b] Given data.

Participants	Exercise (X)	Fitness (Y)
1	2	3
2	1	2
3	3	4
4	3	3
5	2	3
6	3	4

Number of observations: (2,3), (1,2), (3,4), (3,3), (2,3), (3,4)

$n = 6$

total pairs $= \binom{6}{2} = 15$

Kendall's tau-b formula

$$\tau_b = \frac{P - Q}{\sqrt{(P+Q+T)(P+Q+U)}}$$

where

P = number of concordant pairs.

Q = numbers of discordant pairs.

T = number of pairs tied only in X

U = number of pairs tied only in Y

Now, Let $\Delta X = X_i - X_j$, $\Delta Y = Y_i - Y_j$

Concordant if $(\Delta X)(\Delta Y) > 1$

Discordant if $(\Delta X)(\Delta Y) < 0$

Tie in x only if $\Delta X = 0, \Delta Y \neq 0$

Tie in Y only if $\Delta Y = 0, \Delta X \neq 0$

All pairs and their type:

Pair	(X_i, Y_i) vs (X_j, Y_j)	ΔX	ΔY	Type
	(2,3) vs (1,3)	1	1	Concordant
	(2,3) vs (3,4)	-1	-1	Concordant
	(2,3) vs (3,3)	-1	0	Tie in Y only
	(2,3) vs (2,3)	0	0	Tie in both
	(2,3) vs (3,4)	-1	-1	Concordant
	(1,2) vs (3,4)	-2	-2	Concordant
	(1,2) vs (3,3)	-2	-1	concordant
	(1,2) vs (2,3)	-1	-1	concordant
	(1,2) vs (3,4)	-2	-2	concordant
	(3,4) vs (3,3)	0	1	Tie in X only
	(3,4) vs (2,3)	1	1	Concordant
	(3,4) vs (3,4)	0	0	Tie in both
	(3,3) vs (2,3)	1	0	Tie in X only
	(3,3) vs (3,4)	0	-1	Tie in X only
	(2,3) vs (3,4)	-1	-1	concordant

From the table we have.

Number of concordant pairs, $P = 9$

Number of discordant pairs, $Q = 0$

no. of Ties in X only $T = 2$

number of Ties in Y only $U = 2$

numbers of ties in both $= 2$

$$\therefore \tau_b = \frac{9 - 0}{\sqrt{(9+0+2) \times (9+0+2)}}$$

$$= \frac{9}{\sqrt{121}}$$

$$\therefore \tau_b = 0.818$$

The value of 0.818 indicates a strong positive association between exercise frequency and perceived fitness level.

2/c] Given data,

Method A:

Number of students $n_A = 150$, passed $x_A = 85$

Method B:

Number of students $n_B = 120$, passed $x_B = 80$

Sample

Step 1: Sample proportions

$$\hat{P}_A = \frac{85}{150} = 0.567$$

$$\hat{P}_B = \frac{80}{120} = 0.667$$

Difference in sample proportions.

$$\hat{P}_A - \hat{P}_B = 0.567 - 0.667 = -0.1$$

Step-2: Standard error of the difference

$$S.E = \sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n_A} + \frac{\hat{P}_B(1-\hat{P}_B)}{n_B}}$$

$$= \sqrt{\frac{0.567(1-0.567)}{150} + \frac{0.667(1-0.667)}{120}}$$

$$= \sqrt{0.00164 + 0.00185}$$

$$\therefore S.E = 0.059$$

Step 3: 95% confidence interval

For a 95% confidence level, $z_{0.025} = 1.96$

$$\therefore (\hat{P}_A - \hat{P}_B) \pm 1.96 \times S.E$$

$$= -0.1 \pm (1.96 \times 0.059)$$

$$= -0.1 \pm 0.11564$$

$$= (-0.216, 0.016)$$

Interpretation: The 95% confidence interval for the difference in pass rates $(P_A - P_B)$ is $(-0.216, 0.016)$. Since the interval contains 0, there is no statistically significant difference between the pass rates of teaching methods A and B at the 5% significance level.

3/a] To construct a confidence interval for the population odds ratio, we first form a 2×2 contingency table.

We compared Schizophrenia with Non-schizophrenia

Step 1: Form a 2×2 table

Diagnosis	Drugs	No Drugs	Total
Schizophrenia	100	12	112
Non-schizophrenia	30	29	59

Step 2: Sample odds ratio

Let $a = 100$, $b = 12$, $c = 30$, $d = 29$

$$\hat{OR} = \frac{ad}{bc} = \frac{100 \times 29}{12 \times 30} = 8.06$$

Step 3: Standard error of $\ln(\hat{OR})$

$$S.E\{\ln(\hat{OR})\} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$= \sqrt{\frac{1}{100} + \frac{1}{12} + \frac{1}{30} + \frac{1}{29}}$$

$$= 0.402$$

Step 4: 95% confidence interval for $\ln(OR)$

$$\ln(\hat{OR}) = \ln(8.06) = 2.087$$

For 95% confidence, $z_{0.025} = 1.96$:

$$\ln(\hat{OR}) \pm z_{0.025} \times S.E\{\ln(OR)\}$$

$$= 2.087 \pm 1.96 \times 0.402$$

$$= 2.087 \pm 0.788$$

$$= (1.299, 2.875)$$

Step 5: Confidence interval for OR

$$(e^{1.299}, e^{2.875}) = (3.665, 17.725)$$

∴ 95% CI for the population odds ratio

$$= (3.665, 17.725)$$

Interpretation: Since the 95% confidence interval for the odds ratio does not include 1, patients with schizophrenia have significantly higher odds of being prescribed drugs compared to patients with other diagnosis.

3(b) A study compares surgery and radiation therapy for controlling cancer of the larynx. The data are:

Treatment	Cancer controlled	Cancer Not controlled	Total
Surgery	4	3	7
Radiation therapy	3	2	5
Total	7	5	12

We test whether the odds ratio equals 1

Hypotheses

$H_0 : \theta = 1$ against $H_1 : \theta \neq 1$

Fisher's Exact test

Given the fixed marginal totals, the probability of the observed table under H_0 is

$$P = \frac{\binom{7}{4}\binom{5}{3}}{\binom{12}{7}}$$

$$= \frac{35 \times 10}{792}$$

$$= 0.442$$

Possible tables as or more extreme than the observed one give probabilities whose sum constitutes the p-value.

Here, the resulting two-sided P-value is large (>, 0.05)

Since the p-value is greater than 0.05, we fail to reject H_0

Conclusion: There is no ~~signifi~~ significant difference between surgery and radiation therapy in controlling cancer of the larynx at the 5% significance level.

3/c] A study examines the relationship between Gender and Smoking status for a sample of 500 individuals.

The observed data are

Gender	Smoker	Non-Smoker	Total
Male	60	140	200
Female	40	260	300
Total	100	400	500

Hypotheses

H_0: Gender and smoking status are independent

H_1: Gender and smoking status are associated

Expected Frequencies

$$E_{ij} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

For male smoker:

$$E_{11} = \frac{200 \times 100}{500} = 40$$

male: non-smoker: $E_{12} = \frac{200 \times 400}{500} = 160$

Female-smokers, $E_{21} = \frac{300 \times 100}{500} = 60$

Female non-smokers, $E_{22} = \frac{300 \times 400}{500} = 240$

Chi-square test statistic

Gender	Smokers	Non-Smokers	Total
Male	40	160	200
Female	60	240	300
Total	100	400	500

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(60-40)^2}{40} + \frac{(140-160)^2}{160} + \frac{(40-60)^2}{60} + \frac{(260-240)^2}{240}$$

$$= 10 + 2.5 + 6.667 + 1.667$$

$$= 20.84$$

Degrees of freedom

$$df = (2-1) \times (2-1) = 1$$

At 5% significance level.

$$\chi^2_{0.05,1} = 3.84$$

Since $20.84 > 3.841$, we reject H_0. So there is statistically significant association between gender and smoking status in the sample.

4/a] Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) extend the classical linear regression framework to accomodate response variables that do not follow a normal distribution such as binary, count or skewed continuous data. Unlike ordinary linear regression, GLMs allow the respons variable to follow any distribution belonging to the exponential family.

Components of a GLM

i) Random component:

The response variable Y follows a distribution from the exponential family, such as normal, binomial, poisson or gamma.

ii) Systematic component:

The explanatory variables enter the model through a linear predictor:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

iii) Link function:

A know monotonic function $g(\cdot)$ such that relates

$$g(\mu) = g(E[Y]) = \eta$$

Classical linear regression assumes a continuous response variable with normally distributed errors and constant variance.

GLMs extend this framework by:

i) Allowing non-continuous response variable (binary, count, categorical),

ii) Modeling the mean-variance relationship appropriate to the chosen distribution,

iii) Using a link function to connect the mean response to the predictors.

In a GLM, regression coefficients are not directly interpretable on the original scale of the response variable because the model is defined on the link function scale. The coefficients measures changes in the transformed mean $g(E[Y])$, not in $E[Y]$ itself.

For example:

- In logistic regression, coefficients represents in log-odds.

· In Poisson regression, coefficients represent changes in the log of the expected count

Therefore, interpretation on the response scale requires applying the inverse link function.

4/b] Let Y be a binary response variable, with

$\pi = P(Y=1)$

In a Generalized linear model, different link functions may be used to relate π to the linear predictor η

Logit link:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta$$

the logit link is symmetric about $\pi = 0.5$ and allows interpretation of regression coefficients in terms of odds ratios.

Preferred when: Interpretation in terms of odds ratios is important and probabilities are approximately symmetric

Probit link:

Probit $(\pi) = \Phi^{-1}(\pi) = \eta$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

The probit link is also symmetric and assumes an underlying normally distributed latent variable.

Preferred when: a latent normal response mechanism is appropriate.

Complementary log-log link:

$$\text{cloglog}(\pi) = \log[-\log(1-\pi)] = \eta$$

This link is asymmetric and is suitable for modeling rare events or skewed probabilities.

Preferred when: The probability of success is small or data arise from a hazard-type process.

4(c) The overall fit of a generalized linear model (GLM) can be assessed by examining the deviance, which measures the discrepancy between the fitted model and saturated model.

Assessing overall model fit:

The residual deviance is compared with its degrees of freedom. Under a correctly specified model, the residual deviance approximately follows a chi-square distribution with degrees of freedom equal to the residual degrees of freedom.

i) If the residual deviance is small relative to its degrees of freedom, the model provides an adequate fit.

ii) A large deviance indicates lack of fit and suggests that the model may be inadequate.

Additionally, comparing the deviance of nested models using a likelihood ratio test helps assess whether additional predictors significantly

improve the model fit.

Limitation of deviance as a measure of fit:

One limitation of using deviance is that it is sensitive to sample size. In large samples, even small departures from the model assumptions may result in a significant deviance, leading to rejection of an otherwise reasonable model.

5/(a) Logistic Regression:

Logistic regression is a statistical method used to model the relationship between one or more explanatory variables and a binary response variable, which takes only two possible outcomes such as success / failure or presence / absence of an event.

Let

$$Y = \begin{cases} 1; & \text{if the event occurs} \\ 0; & \text{otherwise} \end{cases}$$

and let $\pi = P(Y=1)$ denote the probability of occurance of the event.

Logistic regression models the log-odds (logit) of the event as a linear function of the predictors:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Solving for π,

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}},$$

which ensures that the estimated probabilities lie

between 0 and 1

Difference between Linear Regression and Logistic Regression:

Linear Regression	Logistic Regression
1) Response variable is continuous.	1) Response variable is binary.
2) Errors are normally distributed	2) Errors follows a binomial distribution.
3) Models the mean of the response directly	3) Models the log-odds of the response
4) Predicted values can lie outside [0,1]	4) Predicted values lie in [0,1]
5) Estimated using least squares	5) Estimated using maximum likelihood

Modeling the probability of an event

Logistic regression models the probability of an event by:

i) Transforming the probability using the logit function,

ii) Expressing the transformed probability as a linear predictor,

iii) Applying the inverse logit function to obtain the event probability.

The regression coefficients represents changes in the log-odds of the event, and their exponentials are interpreted as odds ratios.

5/b] The logistic regression model is fitted using 400 observations. The overall ~~significan~~ significance of the model is assessed using the likelihood ratio test, which yield.

$$LR\ \chi^2(5) = 11.35,\ P = 0.0448$$

Since the p-value is less than 0.05, the null hypothesis that all slope coefficients are zero is rejected. Hence, the model provides a statistically significant improvement over the null model. The pseudo $R^2 = 0.0460$ indicates modest explanatory power, which is typical

in logistic regression models.

Interpretation of Regression Coefficients (Odds Ratios)

Retirement condition (Yes):

The odds ratio is 1.618, indicating that retired individuals have higher odds of the outcome compared to non-retired individuals. However, this effect is not statistically significant (P=0.191), as the 95% confidence interval includes 1.

Medical support from family:

Relative to the reference category, individuals receiving fair support have an odds ratio of 0.71, while those receiving no support have an odds ratio of 1.17. Both effects are statistically insignificant (P > 0.05), suggesting no evidence of association with the outcome.

Social Involvement (Yes):

The odds ratio of 0.488 suggests that socially

involved individuals have lower odds of the outcome compared to those not socially involved. This effect is marginally significant (P=0.079), indicating a possible protective association.

PHQ score:

The odds ratio is 1.07, implying that each one-unit increase in PHQ score increases the odds of the outcome by approximately 7%. This effect is marginally significant (P=0.058), indicating a positive association.

Constant term:

The intercept represents the baseline odds of the outcome when all predictors are at their reference categories and is statistically significant

6/a] A log-linear model is a statistical model used to analyze and described the relationship between two or more categorical variables, typically in the form of a contingency table. It is commonly used to model the relationships between categorical variables. the logarithmic transformation helps in ~~linearity~~ linearizing relationships and making certain statistical analysis more tractable. The goal is to model the association or relationship between these variables while accounting for the expected cell counts. For a three-way table with expected frequency μ_{ijk},

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

where the λ's are unknown parameters.

Interpretation of papameters:

i) Main effect parameters ($\lambda_i^X, \lambda_j^Y, \lambda_k^Z$). Represents the effect of individual categorical variables on the expected cell frequencies.

ii) Two-way interaction parameters (e.g., λ_{ij}^{XY}) measures the association between two variables,

controlling for the remaining variables.

iii) Higher-order interaction parameters (e.g., λ_{ijk}^{XYZ}) represents joint interaction among three or more variables.

Joint and conditional associations:

Log-linear models explain joint associations through interaction terms. Conditional independence between two variables, given another variable, is indicated by the absence of the corresponding interaction term in the model.

Comparison between Logistic regression and Log-linear model is below:

Logistic regression	Log-linear model
1) One variable is treated as the response	1) No variable is designated as response.
2) Models conditional probability or odds	2) Models expected cell frequencies.
3) Focuses on conditional association.	3) Focuses on joint association.
4) Used for binary or multinomial outcomes.	4) Used for multi-way contingency tables.

5) Parameters interpreted via odds ratios	5) Parameters interpreted via interaction effects

Appropriate situations:

- Logistic regression is appropriate when the objective is to model the probability of a specific outcome given explanatory variables.
- Log-linear models are appropriate when the objective is to study associations among categorical variables symmetrically, without identifying a response variable.

6/b/(i) The independence model for the contingency table is

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

where, X = aspirin use

Y = Myocardial Infarction

Since this is an independent model so, $\lambda_{ij}^{XY} = 0$

Goodness of fit test:

Group	Myocardial Infarction		
	Yes (Y=1)	No (Y=0)	total
Placebo (X=0)	189	10845	11034
Aspirin (X=1)	104	10933	11037
total	293	21778	22071

Expected frequencies are.

$$\mu_{11} = \frac{(189+104) \times (189*10845)}{189+10845+104+10933}$$

$$= \frac{293 \times 11034}{22071}$$

$$= 146.48$$

$$\mu_{12} = \frac{11034 \times 21778}{22071} = 10887.52$$

$$\mu_{21} = \frac{293 \times 11037}{22071} = 146.52$$

$$\mu_{22} = \frac{21778 \times 11037}{22071} = 10890.48$$

H_0: Aspirin use and myocardial infraction are independent

H_1: Aspirin use and myocardial infraction are dependent

the test statistic is.

$$\chi^2 = \sum \frac{(O_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

$$= \frac{(189 - 146.48)^2}{146.48} + \frac{(10845 - 10887.52)^2}{10887.52} +$$

$$\frac{(104 - 146.52)^2}{146.52} + \frac{(10933 - 10890.48)^2}{10890.48}$$

$$\therefore \chi^2 = 25.014$$

and, $\chi^2_{tab(0.05,1)} = 3.841$

Since $\chi^2_{cal} > \chi^2_{tab\,(0.05,1)}$ then we reject the null hypothesis. that means that aspirin use X and myocardial infraction are dependent

Interpretation of $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$

Under the independence model,

$$\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$$

represents the log ratio of expected frequencies for myocardial infarction outcomes. Since the independence model does not fit, this difference alone cannot explain the observed association.

(ii) The saturated log-linear model is

$$\log_2(u_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

λ_{ij}^{XY} = the probability of joint occurance of level i and j for the variables X and Y.

Now, Under the saturated model, the interaction parameters satisfy

$$\log(OR) = \lambda_{11}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} + \lambda_{22}^{XY}$$

the sample odds ratio is

$$OR = \frac{189 \times 10933}{10845 \times 104} = 1.83$$

Interpretation: Since the odds ratio is greater than 1 it suggests an increased likelihood of myocardial infraction with aspirin used.

7/a] Saturated model: The saturated log linear model is the most complex model that perfectly fits the observed data when analyzing contingency tables. It includes parameters for all possible interactions between the categorical variables, resulting in a model that perfectly reproduces the observed cell counts. The saturated log linear model for a two way table is

$$\log(u_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

In the context of a contingency table,

a saturated model would have a parameter for each cell, resulting in a perfect fit to the observed counts.

Unsaturated model: The unsaturated log linear model is a less complex model compared to the saturated model. It involves a subsets of the interaction terms included in the saturated model. It used for hypothesis testing, assessing the significance of specific interactions and finding a balance between model complexity and goodness of fit. The unsaturated log linear model for a two way table is

$$\log(u_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Consider a three way contingency table formed by the following categorical variables:

A: Gender (male, female)

B: Smoking status (smoker, non-smoker)

C: Disease status (Yes, no)

Let n_{ijk} denote the observed frequency in the cell corresponding to the ith level of A, jth level of B and k^{th} level of C. Let,

$\mu_{ijk} = E(n_{ijk})$ be the expected cell count.

A saturated log linear model is the most general model that can be fitted to a contingency table. It include all possible main effects and all possible interaction terms among the variable. For the three way table (A, B, C) the saturated model is-

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Because this saturated model has a many parameters as there are independent cell counts. It fits the data perfectly. Consequently, there are zero degrees of freedom and no goodness of fit test can be used to assess the adequacy of the model.

An unsaturated log linear model is obtained by remaining removing one or more interaction terms from the saturated model. Such models impose structure on the data and we used to test substantive hypothesis about independence or conditional indepence.

For example, consider the model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

which excludes the three way interaction term λ_{ijk}^{ABC}. This unsaturated model assumes that although the variance may be associated pairwise, there is no three-way interaction.

7| b| The degrees of freedom equal the number of cell counts minus the number of model parameters. The df values decreases as the model become more complex. The saturated model has $df = 0$.

Assuming a standard $2 \times 2 \times 2$ table the total number of cells is $2 \times 2 \times 2 = 8$

Model	d.f	G^2
(X, Y, Z)	8-(1+1+1+1) = 4	137.93
(XY, Z)	8-(1+1+1+1) = 3	131.96
(XY, YZ)	2	7.91
XZ, Y	3	120.36
XYZ	0	0.00

7/c] For each log linear model, we test

H_0: The expected frequency satisfy the given log linear model

H_1: The model does not fit the data.

Given the likelihood ratio statistic, we have:

Model	df	G^2	$\chi^2_{\alpha, df}$
(X, Y, Z)	4	137.93	9.49
(XY, Z)	3	131.96	7.81
(XY, YZ)	2	7.91	5.991
(XZ, Y)	3	120.36	7.81
(XYZ)	0	0.00	-

Model evaluation:

The saturated model (XYZ) has $G^2 = 0$; indicating a perfect fit but it not parsimonious.

The model (X, Y, Z), (XY, Z) and (XZ, Y) have very large G^2 value indicating poor fit, So their null hypothesis we reject.

The model (XY, YZ) has small $G^2 = 7.91$, suggesting that it fits the data adequately.

8/a/ Matched pair data:

Matched pair data arise when observations are collected in pairs that are naturally related or deliberately matched, so that the two observations within each pair are not independent. Each pair shares similar characteristics, and comparisons are made within pairs rather than between unrelated subjects.

Matched pairs commonly occur in:

i) Before-after studies on the same subject.
ii) Case-control studies with matched controls.
iii) Studies involving twins or siblings.

Models for matched pair data:

1) Paired t-test:

Used when the response variable is continuous and approximately normally distributed. The analysis is based on the difference within each pair.

2) Wilcoxon signed-rank test:

A non-parametric alternative to the paired

t-test, used when the normality assumption is not satisfied.

3) McNemar's test:

Used for binary matched-pair data to test marginal homogeneity.

4) Conditional logistic regression:

Used when the response is binary and there are additional covariates, accounting for the matched-pair structure.

8|b| A matched-pair clinical trial compares control treatment with a new treatment for the same patients. the observed data are:

	Treatment improved	Treatment not improved
Control improved	20	25
control not improved	30	28

Step 1: Identifying discordant pairs

McNemar's test is based only on discordant pairs:

b: Control improved, Treatment not improved = 25

c: Control not improved, treatment improved = 30

Step-2: State hypotheses

H_0: No. difference between control and treatment ($b = c$)

H_1: There is a difference between control and treatment.

Step-3: McNemar's test statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c} = \frac{(25-30)^2}{25+30} = \frac{25}{55}$$

$$\therefore \chi^2 = 0.455$$

Step-4:

$$d.f = 1$$

Critical value at 5% level is

$$\chi^2_{0.05,1} = 3.84.$$

Since $0.455 < 3.84$, we fail to reject H_0.

So, there is no statistically significant difference between the control treatment and the new treatment in terms of improvement.

8|c| For binary matched pair data, a logistic regression model is formulated by conditioning on each matched pair so that the within-pair dependence is properly accounted for. This is commonly known as conditional logistic regression.

Model Formulation:

Let Y_{ij} denote the binary outcome for subject j $(j=1,2)$ in matched pair i where

$$Y_{ij} = \begin{cases} 1; & \text{if the outcome is present} \\ 0; & \text{otherwise} \end{cases}$$

Let X_{ij} be an indicator variable for treatment status. The conditional logistic regression model is given by

$$\log\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)}\right) = \alpha_i + \beta X_{ij},$$

where α_i is a pair-specific effect and β is the treatment effect parameter.

Key components of the model:

i) Dependent variable: Binary out come Y_{ij}

ii) Independent variable: Treatment indicator X_{ij}

iii) Parameters:

α_i accounts for unobserved pair specific characteristics, and

β measures the effect of treatment on the log-odds of the outcome.

Explanation:

The model captures the relationship between treatment and the binary outcome by modeling the log-odds of the outcome as a function of treatment status, while accounting for the matched-pair structure. The inclusion of a pair-specific effect controls for all characteristics common to each pair. The treatment coefficient represents the log odds ratio comparing treatment to control within pairs, and its exponential gives the odds ratio, thereby quantifying the treatment effect on the binary outcome.