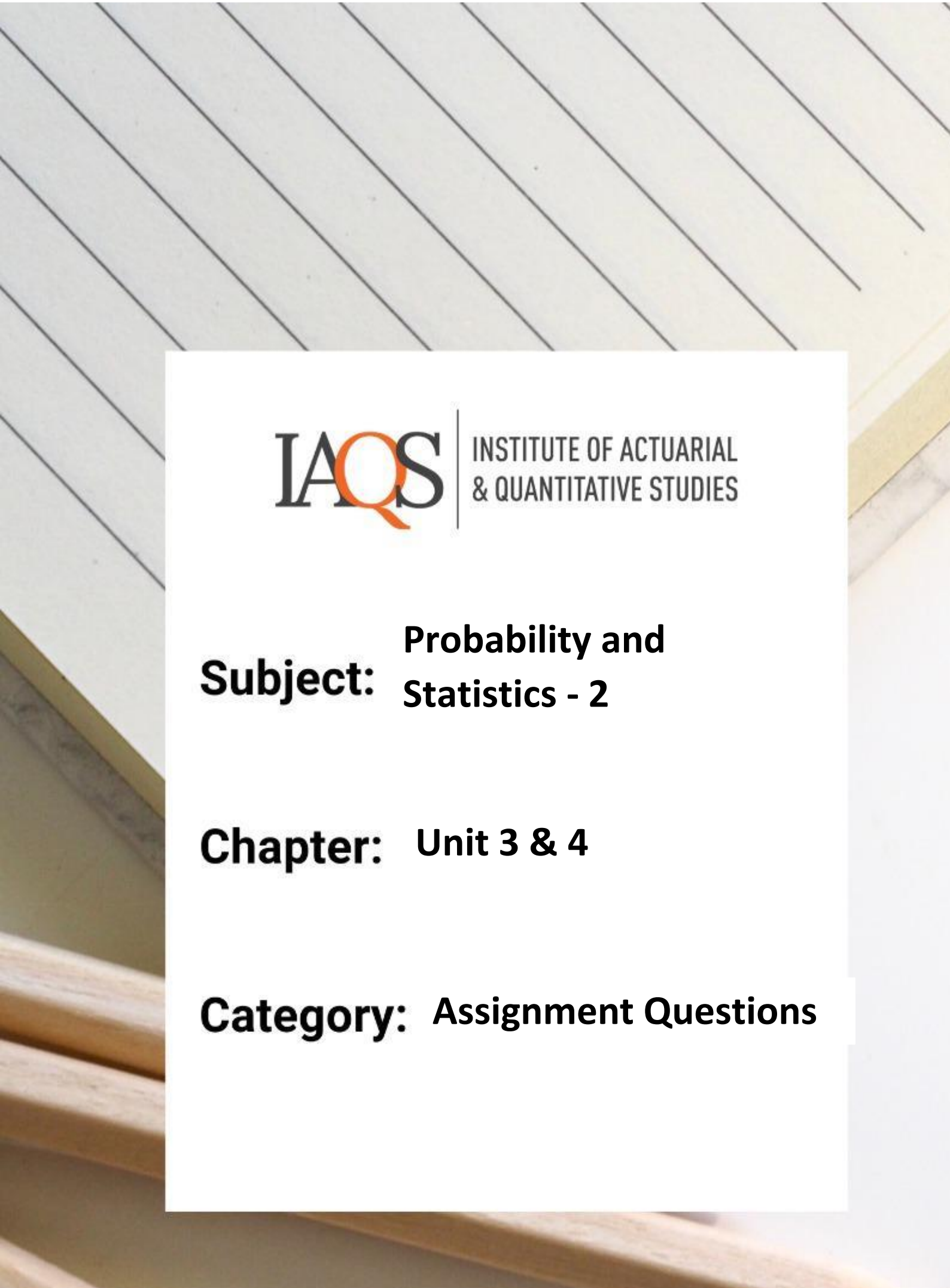


IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics - 2

Chapter: Unit 3 & 4

Category: Assignment Questions

1. Unit 3

Anand obtains cash from an ATM (cash machine) for his girlfriend. He suspects that the rate at which she spends cash is affected by the amount of cash he withdrew at his previous visit to an ATM.

To investigate this, he deliberately varies the amounts he withdraws. For the next 10 withdrawals, he records, for each visit to an ATM, the amount x (in Rs.) withdrawn, and the number of hours, y, until his next visit to an ATM.

Withdrawal	1	2	3	4	5	6	7	8	9	10
x	40	10	100	110	120	150	20	90	80	130
y	56	62	195	240	170	270	48	196	214	286

(a) Calculate the equation of the regression line of y on x

(b) Interpret, in context of the question, the gradient of the regression line

2. Unit 3

Dell and IBM are well known in the computer industry. If the computer industry is doing well then we may expect the stocks of these two companies as well to increase in value. If the industry goes down then we would expect both may go down as well.

The table below gives data on the share prices (in US $) of Dell (X) and IBM (Y) at the end of each month for a calendar year:

X	27.9	40.7	37.8	31.6	37.5	31.6	29.2	24.5	30.9	25.6	37.9	30.0
Y	97.4	105.0	145.5	126.2	114.2	106.7	76.7	65.6	68.9	82.2	95.6	78.5

$\sum x = 385.2;\ \sum x^2 = 12{,}666.58;\ \sum y = 1{,}162.5;\ \sum y^2 = 119{,}026.9;\ \sum xy = 38{,}191.41$

i) Calculate the least squares fit regression line in which IBM share price is modelled as the response and the Dell share price as the explanatory variable.

ii) Determine a 95% confidence interval for the slope coefficient of the model. State any assumptions made.

iii) Use the fitted model to construct 95% confidence intervals for the mean IBM share price when the Dell Share price is US $ 40.

3. Unit 4

The government's urban planning committee is concerned about the increasing traffic within the major cities and the time required for commuting within the city.

In order to understand the situation, it commissions an external agency to conduct a survey looking into the difference between the times taken to commute in the peak hours (i.e. office hours) compared with that in the non-peak hours (i.e. non-office hours).

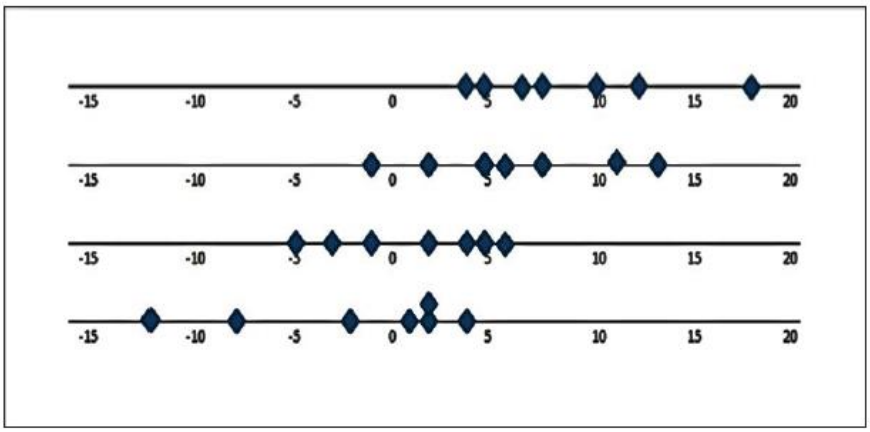
The study was carried out in 4 major cities with a randomly selected sample of 7 commuters in each city.

The table in the next page shows the difference between the time taken (in minutes) for each individual to commute to office in peak hours and non-peak hours:

Cities →	A	B	C	D	$\sum y_{ij}$	$\sum y_{ij}^2$
	18	2	2	-8	14	396
	7	11	-3	4	19	195
	8	13	-5	2	18	262
	4	-1	5	-12	-4	186
	10	5	4	1	20	142
	5	8	6	-2	17	129
	12	6	-1	2	19	185
Total	**64**	**44**	**8**	**-13**	**103**	**1,495**

The agency wanted to infer on the mean time differences among the 4 cities and thus decides to use an analysis of variance (ANOVA) approach.

i) The diagram given below compares the time differences in commutation in the four cities (in order). Suggest brief comments one can make on the basis of the plot.



ii) State the assumptions required to carry out the ANOVA.

iii) Carry out the ANOVA test and draw your conclusion at the 5% significance level.

iv) Carry out an analysis of the mean differences using a least significant difference approach at the 5% significance level

IAQS INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

4. Unit 4

(a) State the mathematical model of one-way ANOVA defining all notations and Assumptions.

(b) Random sample of claim amounts (in units if Rs.1,000 over a total 3-year period) under hospitalization reimbursement policies were taken from 5 different general insurance companies are shown below.

Company				
y_1	y_2	y_3	y_4	y_5
60	40	33	55	88
52	56	0	78	89
11	123	0	99	44
33	0	12	234	78
1	12	19	45	85
87	54	23	67	
110	77		67	
	24			

$\sum y_1 = 354$; $\sum y_2 = 386$; $\sum y_3 = 87$; $\sum y_4 = 645$; $\sum y_5 = 384$

$\sum y_1^2 = 27{,}184$; $\sum y_2^2 = 29{,}430$; $\sum y_3^2 = 2{,}123$; $\sum y_4^2 = 84{,}669$; $\sum y_5^2 = 30{,}910$

Perform ANOVA to test whether claims are equal or not.

(c) A marketing analyst claims that salaries of actuarial students do not depend on number of actuarial papers they have cleared. To test his claim, he collects data of 158 actuarial students distributed according to their annual salaries and number of papers cleared as shown below

Papers cleared	Salary per annum (in Rs. lacs)			
	3 - 5	5 - 8	8 - 10	10 - 12
0 - 3	45	20	6	5
4 - 6	7	20	9	6
7 - 9	5	8	15	12

Conclude your view statistically based on the above data using a x^2 test

5. Unit 4

The following data (x) are the number of germinations per square foot observed in an experiment where a particular type of plant seed was applied at four different rates.

	Rate of Application			
	1	2	3	4
	29	180	332	910
	13	90	444	880
	21	120	190	460
Mean	21	130	322	750
Variance	64	2,100	16,204	63,300

[i] State the assumptions required for a one-way analysis of variance (ANOVA) and whether these data appear to violate any of them. Give reasons.

[ii] An agricultural scientist working on this data applied the following three transformations to the data:

$\sqrt{x}, \quad \log_e(x) \quad \text{and} \quad (1/x).$

The following table contains values for the means and variances of the transformed data although he forgot to fill in a couple of them and thus is missing (the ones marked as ****):

Rate	Transformation					
	$\sqrt{x}$		$\log_e(x)$		$1/x$	
	mean	variance	mean	variance	mean	variance
1	****	0.79	2.992	0.1630	0.05301	0.0004721
2	11.29	3.94	4.827	****	0.00833	0.0000077
3	17.69	13.49	5.716	0.1861	0.00351	0.0000025
4	27.09	23.96	6.575	0.1479	0.00147	0.0000004

Complete the table for him by determining the missing values.

[iii] The scientist insisted (in order to perform a one-way ANOVA) we must consider the loge transformation of the data only. Argue heuristically why this is the case?

[iv] The scientist now decides to go ahead and performs the one-way ANOVA.
(a) State the ANOVA model for the transformed data along with the null hypothesis he is testing.
(b) Perform the ANOVA for the transformation chosen.
(c) What conclusions can be drawn from the ANOVA?

6. Unit 3

The insurance regulator has conducted a study to understand the relation between the number of branches a life insurance company operates with and the number of policies it sells. At the end of the month the regulator examined the records of 10 insurance companies. It obtained the total number of branches (x) and the number of policies (y) sold in the month.

The collected data is given in the table below;

Company	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Number of Branches (x)	5	9	2	3	3	1	1	6	5	4
Number of Policies (y)	73	120	34	46	35	24	26	93	45	66

A set of summarised statistics based on the above data is given below:

$$\sum x = 39; \sum x^2 = 207; \sum y = 562; \sum y^2 = 40{,}508; \sum xy = 2{,}853$$

i) Calculate the correlation coefficient between x and y.

ii) Calculate the fitted linear regression equation of y on x.

iii) Calculate the "total sum of squares" together with its partition into the "regression sum of squares" and the "residual sum of squares".

iv) Using the values in part (iii), calculate the coefficient of determination R^2. Comment briefly on its relationship with the correlation coefficient calculated in part (i).

7. Unit 3

You have been given the data set comprising of mortality rating and premium rates.

Mortality Rating (%)	25	50	75	100	135	180	215	245	300	365	400
Premium Rate	3.50	5.80	8.07	10.33	13.44	17.38	20.39	22.94	27.53	32.83	35.62

It has been specified that the premium rates can be expressed as a cubic function of mortality rating. You have been given the task of deriving a simple formula for calculation of premium rates at different mortality ratings.

Suppose, moving ahead in line with the proposed methodology, you decide to fit a simple linear regression with premium rates being regressed on cubic function of mortality rating. The transformed data is as below:

(Mortality Rating) : (X)	0.02	0.13	0.42	1.00	2.46	5.83	9.94	14.71	27.00	48.63	64.00
Premium Rate: (Y)	3.50	5.80	8.07	10.33	13.44	17.38	20.39	22.94	27.53	32.83	35.62

You are given the following summary statistics:

$$\sum x = 174.13, \sum y = 197.84 \sum x^2 = 7545.90, \sum y^2 = 4747.45 \sum (x - x^{-})(y - y^{-}) = 2176.84$$

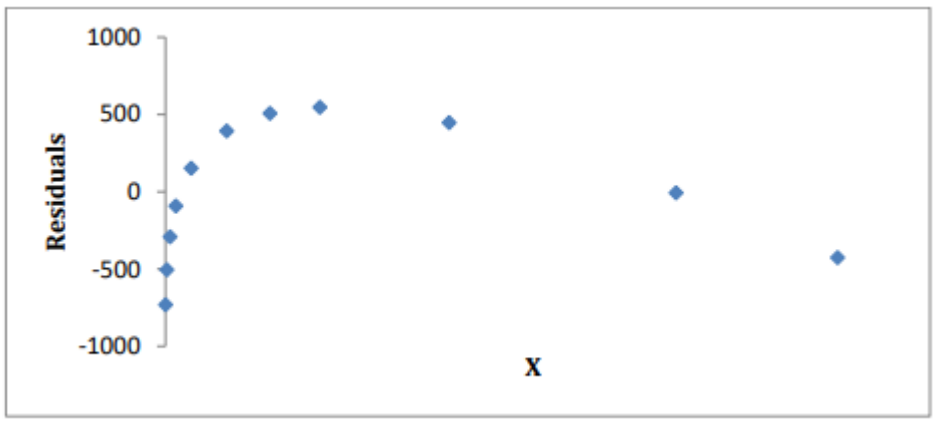
i) Derive the linear regression equation of premium rates Y on X.

ii) Perform a statistical test to investigate the hypothesis that there is no linear relationship between X and Y. State clearly all assumptions made.

iii) Calculate the sample correlation coefficient.

iv) Calculate the 95% confidence interval for the individual and mean responses corresponding to $x\text{^}3 = 25$.

v) Consider the residual plot of the fitted regression:



Comment on the fit of the model and any drawbacks of using the model rather than the full premium rate table.

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

8. Unit 3

i) What is the main purpose of performing Factor analysis? Comment on the original variables and newly identified Principal components.

ii) A student came up with 5 by 5 variance-covariance matrix of the Principal Components (PC1, PC2, PC3, PC4, PC5) with these 5 diagonal entries: 0.456, 0.137, 0.080, 0.0165 and 0.012 respectively. Identify the percentage of the total variance explained by each Principal Component.

iii) What can be concluded based on the results of part (ii).

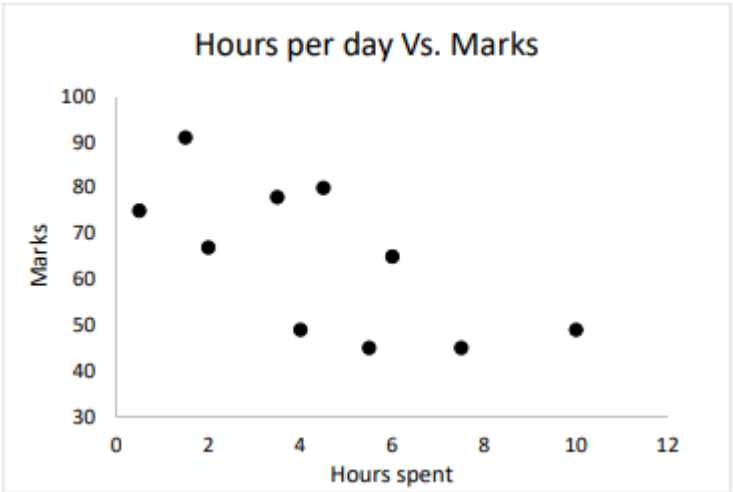
9. Unit 3

The below data gives the marks scored and time spent on social media in hours per day:

Hours x	6	5.5	0.5	3.5	7.5	4.5	4	1.5	2	10
Marks y	65	45	75	78	45	80	49	91	67	49

For this data: $\sum x = 45; \sum x^2 = 277.5; \sum y = 644; \sum y^2 = 43{,}956; \sum xy = 2{,}602$

Scatterplot for the above data is provided below:



i) Based on the above scatterplot, comment on the association between marks scored and hours spent per day on social media.

ii) Calculate the correlation co-efficient between the two variables and comment on the same.

iii) Investigate the hypothesis that there is a negative correlation using Fisher's transformation of the correlation co-efficient. You should clearly state the hypothesis of your test and any assumption made for the test to be valid.

iv) Fit a linear regression model to this data with marks being the response variable and hours spent as the explanatory variable.

v) Calculate the coefficient of determination for this model and interpret the same.

vi) Calculate the expected change in marks for every 1 additional hour spent on social media basis the model fit in (iv).

10. Unit 3

In a survey by Industry, 20 organizations from each industry were questioned on their attrition rate and a summary was prepared as below:

Industry	Manufacturing	IT / ITES	Consulting
Average resignation %	27%	36%	30%
Sample standard deviation	5%	10%	8%

Perform a one-way analysis of variance test to test the hypothesis that the Industry has no impact on resignation rate.

11. Unit 4

A random variable z has a binomial distribution with parameters n and μ and has the following density function:

$$f(z) = (nCz)\ \mu^{z}\ (1 - \mu)^{n-z}$$

where $0 < \mu < 1$

i) Show that the distribution function of Y = Z

n can be written in the standard form of the exponential family of distributions, stating the natural and scale parameters, θ and φ, and the associated functions of these parameters.

ii) Verify the mean and variance of the Binomial Distribution, using the expressions from part (i) together with the properties of the exponential family of distributions.

A researcher is investigating the number of students who pass in a particular examination. The researcher believes that the number of students who pass follows binomial distribution.

He also believes that probability of passing, μ, depends on the followings

- The number of assignment, N, submitted by the student
- The student's mark in the mock exam S
- Whether student attended tutorials or not (Yes/No)

The researcher specifies the following linear predictor, where αi, $\beta 1$ and $\beta 2$ are parameters to be estimated

$\eta(\mu) = \alpha_i + \beta_1 N + \beta_2 S$

Where αi takes one value for those attending tutorials (αY) and a different value for those who do not (αN).

The researcher then runs computer model that fits generalized linear model (using binomial canonical link function) basis of data collected from 30 observation points.

Parameters:	Estimate	Standard Error
Intercept, α_Y	-1.501	0.29190
Intercept, α_N	-3.196	0.13401
β_1, no. of assignment	0.5459	0.08352
β_2, mark in mock exam	0.0251	0.00156

iii) Explain, using the model output shown above, whether the variable "no. of assignment" is significant or not.

iv) Estimate using the fitted model, the probability of passing for a student who attends tutorials, submitted 4 assignments and scored 65 marks in the mock exam.

12. Unit 4

i) a) Poisson distribution $(e^{-mu} * mu^y) / (y!)$ can be written as exp(y*log(mu)-mu-log y!), in the form of a member of the exponential family.

Identify correct option indicating Variance function i.e. V(mu)
A. V(mu) = log(mu)
B. V(mu) = (1/mu) -1
C. V(mu) = (y/mu) -1
D. V(mu) = mu
E. V(mu) = 1/mu

b) Identify correct option indicating mean (mu) and variance as a function of 'mu' for a particular distribution when written in the form of a member of the exponential family having

- b(theta) = -log(-theta)
- theta = -1/mu
- a(phi) = 1

A. mean = mu and variance = mu
B. mean = mu^2 and variance = mu^2
C. mean = mu and variance = mu^2
D. mean = mu^2 and variance = mu
E. None of the above

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES