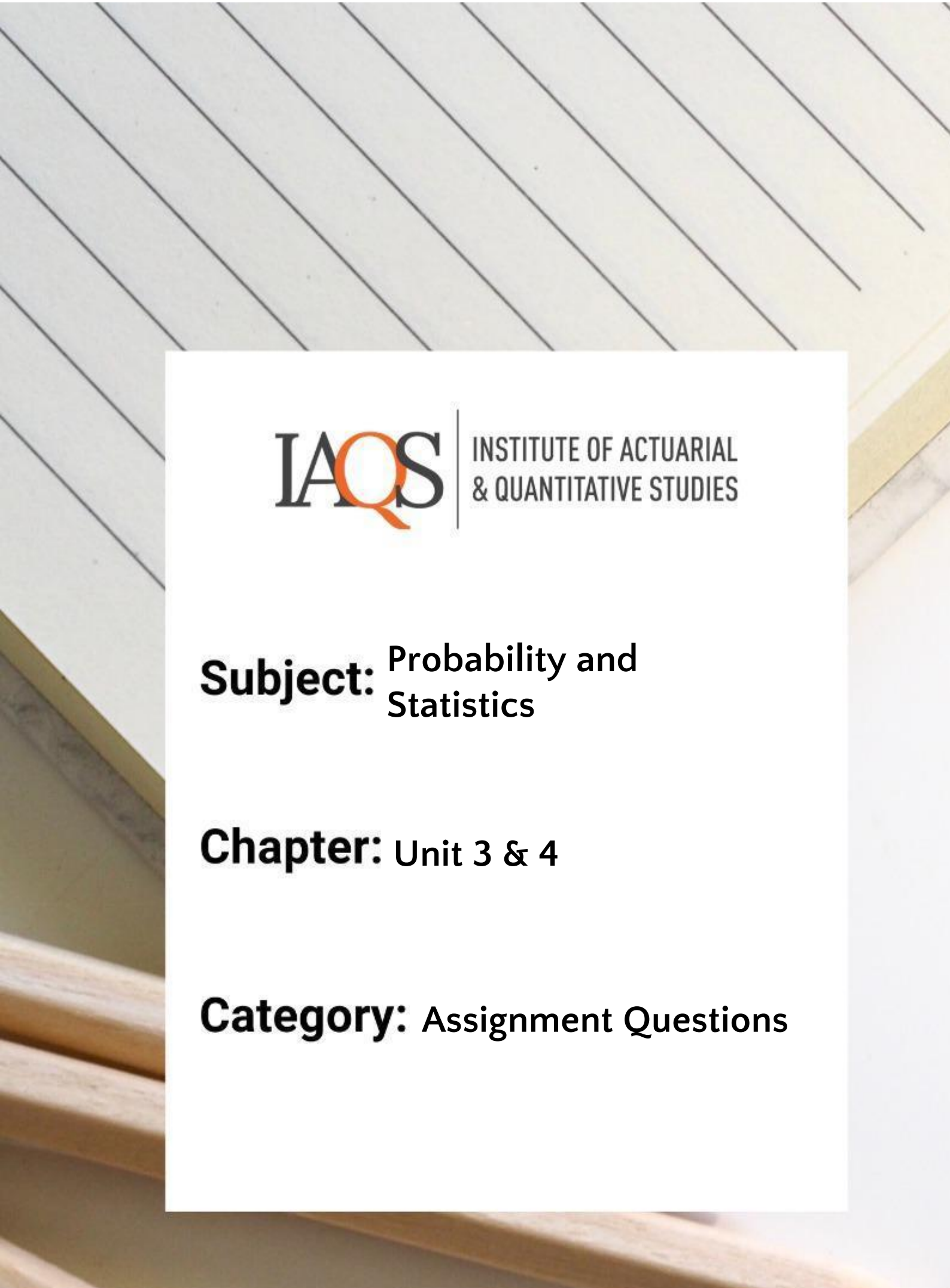


IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics

Chapter: Unit 3 & 4

Category: Assignment Questions

1. Suppose X_i: i=1,2,…*n are n* independent and identically distributed random variables, each with mean μ and variance σ^2. Let $\bar{X} = \sum X_i/n$ be the sample mean and

$S^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2$ be the sample variance

i) Show that the mean and variance of $\bar{X}$ is μ and σ^2/n respectively

ii) Show that $E(S^2)=\sigma^2$.

Assume X_i 's are from a normal population. Using the distribution of $(n-1)S^2/\sigma^2$ as chi square

iii) Show that variance of S^2 is $2\sigma^4/(n-1)$.

iv) Find the probability that S^2 will fall between plus or minus 50% of its expected value when n=10 and $\sigma^2=100$.

2. An insurer believes that the distribution of the number of claims on a particular type of policy is binomial with parameters *n=4* and *p*. A random sample of 180 policies revealed the following information.

No. of claims	0	1	2	3	4
No. of policies	86	75	16	2	1

i) Obtain the maximum likelihood estimate of p.
ii) Carry out a goodness of fit test for the number of claims on each policy conforms to the binomial model.

3. The probability density function of a random variable X , is given by

$f(x) = \frac{c\beta^3}{(x+\beta)^4}$; $x > 0, \beta > 0$

where c is a constant and β is a parameter.

(i) Determine the value of c and calculate the mean and variance of X as a function of β by using Formulae and Tables for Actuarial Examinations or otherwise.

It is required to estimate β based on a random sample $X_1, X_2, \ldots, X_n$
ii) Show that the method of moments estimator $\hat{\beta}$ is $2\bar{X}_n$ and verify the unbiasedness and consistency of this estimator.

UNIT 3 & 4

ASSIGNMENT QUESTIONS

iii) Consider the set of estimators of the form $b\bar{X}_n$ where b is a constant. Show that the value of b that minimizes the MSE of $b\bar{X}_n$ is $2/(1+3/n)$

iv) Compare the unbiasedness and consistency of the estimator in (iii) with minimum b using the corresponding properties of the estimator in (ii).

4. i) Show that the method of moments estimate for ' a ' of a continuous uniform distribution *U(a,b) is* $\hat{a} = \bar{x} - \sqrt{3}s$ where $\bar{x}$ is sample mean and *s* is sample standard deviation.

ii) State a formula for $\hat{b}$ (method of moments estimate for ' b ') in terms of sample mean and sample standard deviation.

iii) Create a sample of five observations and use the sample to demonstrate potential weakness of the method of moment's estimation of *' a ' and ' b ' for U(a,b).*

iv) Find the maximum likelihood estimate of *b for U(a,b)* based on a sample $x_1, x_2, \ldots, x_n$, when *a=0.*

5. ABC Space agency, responsible to protect a planet from asteroid collision, developed new space-to-space missiles to be loaded in satellites.

The agency planned missile trial in two steps for testing H_0: p=0.1 versus H_1: p>0.1, where p is the proportion of hits of missiles, each missile targeted at similar asteroid.

At the first step, 12 missiles will be fired. If three or more independent hits are observed among the (first) 12 missiles, H_0 is rejected, the study is terminated, and no more missiles are fired.

Otherwise, another 12 missiles will be fired in the second step. If a total of five or more independent hits are observed among the 24 missiles fired in the two steps, then H0 is rejected.

i) Calculate the probability of Type I error for the two step testing procedure.

ii) Calculate the probability of rejecting the null hypothesis H_0 when p=0.3.

iii) Calculate the probability of Type II error when p=0.3.

Suppose that 10 Space agencies have developed missiles similar to ABC Space agency. They have performed only the first step trials and observed an aggregate of 40 hits out of 120 independent missiles fired.

iv) Calculate an approximate lower bound of 90% right-tailed confidence interval for ' p '. (show up to four decimal places)

v) Test the hypothesis that H_0: p=0.278 against H_1: p>0.278 at 10% level of significance.

vi) Comment on your results of confidence interval obtained in (iv) and hypothesis testing in (v).

Aptitude tests were conducted by ABC Space agency at their two Institutes (1 and 2) that provide special training on missile technology. The test scores are shown in the table below:

	Institute 1	**Institute 2**
No. of trainees	12	15
Mean scores	65	70
Standard deviation	54	70

vii) Test the equality in mean scores of the populations associated with the two Institutes. State any assumptions made.

6. A study into the average claim (in Rs. '000) per health insurance policy was performed for the claims incurred in public and private hospitals. Data for some cities is given below:

	City 1	City 2	City 3	City 4	City 5	City 6	City 7	City 8	City 9
Public	24	45	29	33	20	40	26.5	25	27.5
Private	30	54.5	30	40	28.5	36	30.5	30.5	35.5

i) Determine the sample mean and sample variance of average claim size in both the type of hospitals.

ii) State the primary condition that needs to be true for testing equal mean and verify whether that condition is satisfied in the above example (You may assume that the samples come from a normal population).

iii) Test whether the treatments in private hospitals result in higher claim size at 95% level.

7. If $\hat{\theta}$ is an estimator of parameter θ, answer the following:

i) Define unbiased estimator

ii) Define 'bias'.

iii) Define Mean Square Error (MSE) of this estimator $\hat{\theta}$

There exist another estimator $\tilde{\theta}$ of the same parameter θ, such that $\hat{\theta}$ has no bias but higher MSE than $\tilde{\theta}$ while $\tilde{\theta}$ has a positive bias.

iv) State giving reason, which estimator is 'efficient'?

v) When would either of the two estimators be termed as consistent?

vi) Outline (in one sentence each) any two methods of estimating

8. i) Define the variable t_k used in the t-test for sampling distribution of sample mean describing all the symbols used.

ii) State the mean and variance of t_k for k>2.

iii) A sample of 10 numbers from normal population has sample mean and sample variance as 50 and 48.667 respectively.

Determine the confidence interval for the population mean at 99% confidence level-
a) Using the t-test tables
b) Assuming a Normal distribution with parameters as the results of part (ii) above

9. Number of claims in a year on an insurance policy is believed to follow a Poisson distribution. Claims on portfolio of 1000 such policies were observed for one year. It was suggested that the value of Poisson parameter is 3 . If the observed number of claims in that one year is less than 3100 then the suggested value of 3 for the Poisson parameter is accepted else rejected.

You may use the result that probability distribution of summation of ' n ' Poisson variables with parameter μ is Poi (nμ).

i) Define Type I error and estimate it for the above case.

ii) Define Type II error.

iii) Define power of a test and determine the power of test in terms of μ in above case.

iv) If the actual observed number of claims is 2900 , determine the confidence interval for the Poisson parameter at 99% confidence level.

10. Let $X_1, X_2 \ldots X_n$ be a random sample from Uniform distribution over $(0,\theta)$, where θ is an unknown parameter (>0).

[i] Outline why the Cramer-Rao lower bound for the variance of unbiased estimators of θ does not apply in this case.

Consider an estimator of θ : $\hat{\theta}(c)=cY$ for some constant c where $Y=X_i$
[ii] Show that the probability density function of Y is given as:

$g_Y(y)=(n/\theta^n)\cdot y^{n-1}$ for $0<y<\theta$

UNIT 3 & 4

ASSIGNMENT QUESTIONS

Hence, show that:

$E[Y^k] = (n\theta^k)/(n+k)$ for any non negative real number k

[iii] Show that the bias and mean square error (MSE) of the estimator $\hat{\theta}$ (c) are given as follows:

$$Bias[\hat{\theta}(c)] = \left(\frac{c\,n}{n+1} - 1\right).\theta$$

$$MSE[\hat{\theta}(c)] = c^2.\frac{n\,\theta^2}{n+2} - c.\frac{2n\,\theta^2}{n+1} + \theta^2$$

[iv] Find the value of $c(=c_u)$ for which $\hat{\theta}$ (c) becomes an unbiased estimator of θ.

[v] Find the value of $c(=c_m)$ for which the mean square error of $\hat{\theta}$(c) is minimised.

[vi] Which of the two estimators $\hat{\theta}(c_u)$ or $\hat{\theta}$ (c_m) will you prefer for estimating θ ? Give reasons. What happens when n is large?

11. 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

12. Dell and IBM are well known in 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. (4)

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.

13. 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. (4)

ii) Calculate the fitted linear regression equation of y on x. (2)

iii) Calculate the "total sum of squares" together with its partition into the "regression sum of squares" and the "residual sum of squares". (2)

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). (2)

14. 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:

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

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

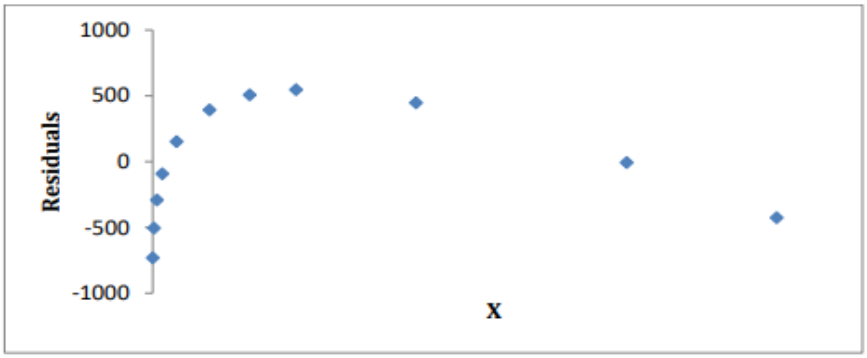
ii) Perform a statistical test to investigate the hypothesis that there is no linear relationship between X and Y.

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 $\hat{x}^3 = 25$.

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.

15. i) What is the main purpose of performing Factor analysis? Comment on the original variables and newly identified Principal components. (3)

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). (2)

16. A discrete random variable X assumes values -1, 0, 1 each with non-zero cell probability as under:

X	-1	0	1
$P(X=x)$	$\frac{1}{6} + \alpha$	$\frac{1}{2} - 3\alpha$	$\frac{1}{3} + 2\alpha$

$(-\frac{1}{6} < \alpha < \frac{1}{6}$, is a parameter).

A random sample of 25 observations gave respective frequencies of 7, 6 and 12.

i) Show that the log likelihood function for the given sample of observations is.

$\ln L(\alpha) = 7 \ln(1/6 + \alpha) + 6 \ln(1/2 - 3\alpha) + 12 \ln(1/3 + 2\alpha)$

ii) Determine the maximum likelihood estimate and explain why one of the two roots of the log likelihood equation is rejected as a possible estimate of α.

iii) Using method of moments, determine the estimate of α.