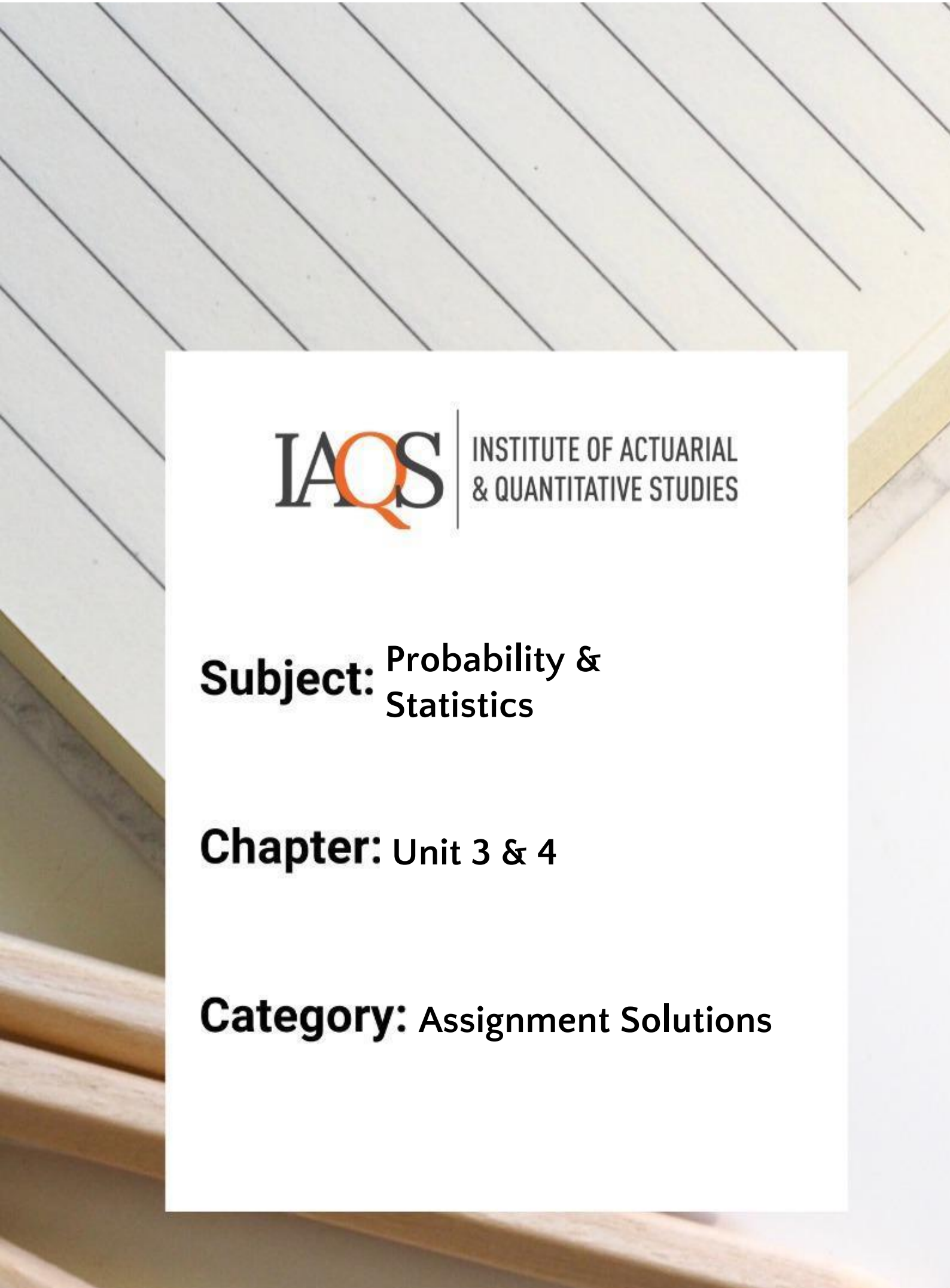

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability & Statistics

Chapter: Unit 3 & 4

Category: Assignment Solutions

1.

(i) Sample mean: $\bar{X} = \frac{\sum Xi}{n}$

$E[\sum Xi] = \sum E[Xi] = \sum \mu = n\mu$; since they are identically distributed

$Var[\sum Xi] = \sum Var[Xi] = n\sigma^2$; since they are iid

$E[\bar{X}] = \mu$.

$Var[\bar{X}] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$

(ii) Sample variance : $S^2 = \frac{1}{n-1}[\sum Xi^2 - n\bar{X}^2]$

$E[S^2] = \frac{1}{n-1}(\sum E[Xi^2] - n\,E[\bar{X}^2])$

$= \frac{1}{n-1}[\sum(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)]$

$= \frac{1}{n-1}[n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2] = \frac{1}{n-1}[(n-1)\sigma^2] = \sigma^2$

(iii) The sampling distribution of $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ when sampling from a normal population, with mean μ and variance σ^2

The variance of χ^2_k is 2k.

Hence, Var $[\frac{(n-1)S^2}{\sigma^2}] = 2(n-1) \Rightarrow Var[S^2] = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}$

(iv) It is given that $\sigma^2 = 100$ and n = 10.

We need to compute P $(50<S^2<150)$

We know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ and in this case it is χ^2_9

Let $Y = 9S^2 / 100$. Then, for $S^2 = 50$: $Y = 9(50)/100 = 4.5$;

and for $S^2 = 150$: $Y = 9(150)/100 = 13.5$

Thus, $P(50<S^2<150) = P(4.5<Y<13.5)$

$= P(Y<13.5) - P(Y<4.5)$ [from page 165 of Tables]

$= 0.8587 - 0.1245 = 0.7342$

2.

(i) The likelihood function is:

$$L(p) = C\,[(1-p)^4]^{86}\,[4p(1-p)^3]^{75}\,[6p^2(1-p)^2]^{16}\,[4p^3(1-p)]^2\,[p^4]^1$$

$$= C\,[(1-p)^{344}]\,[p^{75}(1-p)^{225}]\,[p^{32}(1-p)^{32}]\,[p^6(1-p)^2]\,[p^4]$$

$$= C\,(1-p)^{603}\,p^{117}$$

C is a constant.

Taking logs and differentiating with respect to p and setting equal to zero gives:

$dLn\,L/dp = -603/(1-p) + 117/p = 0$
$\Rightarrow p = 117/(603+117) = 117/720 = 0.1625$
Checking for maximum:
$d^2Ln\,L/dp^2 = -603/(1-p)^2 - 117/p^2 < 0$
$\Rightarrow$ Maximum value for Ln L is at $p = 0.1625$

(ii)

Goodness of fit test:

We are testing the following hypotheses using a $\chi 2$ goodness of fit test:
H_0: the probabilities conform to a Bin (4, p) distribution
H_1: the probabilities do not conform to a Bin (4, p) distribution

Using $\hat{p} = 0.1625$ from part (a), the probabilities for this binomial distribution are:

$P(X=0) = (1-p)^4 = 0.49197$
$P(X=1) = 4p(1-p)^3 = 0.38183$
$P(X=2) = 6p^2(1-p)^2 = 0.11113$
$P(X=3) = 4p^3(1-p) = 0.01437$
$P(X=4) = p^4 = 0.00070$

The expected values are 88.55, 68.73, 20.00, 2.59 and 0.13.
Combining the expected values less than 5 to third group, we get

No of Claims	0	1	2 or more
Observed O_i	86	75	19
Expected E_i	88.55	68.73	22.72

The degrees of freedom = 3-1-1=1.

$\chi 2 = \sum (O_i - E_i)^2 / E_i$

$= (86 - 88.55)^2 / 88.55 + (75 - 68.73)^2 / 68.73 + (19 - 22.72)^2 / 22.72$
$= 0.074 + 0.572 + 0.608 = 1.254$

This is less than the 5% critical value of 3.841.We have insufficient evidence at 5% level to reject Ho. Hence, the model is a good fit.

3.

(i)

From the Formulae and Tables for Actuarial Examinations this pdf corresponds to two parameter version of the distribution given by

$$f(x) = \frac{\alpha \lambda^{\alpha}}{(\lambda + x)^{\alpha+1}} \; ; x > 0 \, , \lambda > 0 \text{ and } \alpha > 0$$

$E[X] = \lambda / (\alpha - 1)$ and $\text{Var}[X] = \alpha \lambda^2 / ((\alpha - 1)^2 (\alpha - 2)); \quad \alpha > 2$

Thus the pdf of X, $f(x) = \frac{C \beta^3}{(x + \beta)^4}$; $x > 0$ and $\beta > 0$ can be identified with c=3 (α known) $\lambda = \beta$ (unknown).

$E[X] = \beta/2$

[0.5]

$\text{Var}[X] = 3/4 \, \beta^2$

[0.5]

If the students obtain the results using the first principles full credit to be given

$$f(x) = \frac{\alpha \lambda^{\alpha}}{(\lambda + x)^{\alpha+1}} \; ; x > 0 \, , \lambda > 0 \text{ and } \alpha > 0$$

$\int_0^{\infty} \frac{C \beta^3}{(x + \beta)^4} \, dx = 1$ implies $\left[-\frac{C}{3} \frac{\beta^3}{(x + \beta)^3} \right]_0^{\infty} = 1$ giving $c = 3$

Mean and Variance

(ii)

The method of moments estimator:

Equating sample mean to E[X] from (i) gives:

$\bar{X}_n = \beta / 2 \Rightarrow$ The moments estimator $\hat{\beta} = 2 \bar{X}_n$

The mean square error of $\hat{\beta}$:

$\text{MSE}[\hat{\beta}] = \text{Var}[\hat{\beta}] + (\text{Bias}[\hat{\beta}])^2$

$E[\hat{\beta}] = E[2 \bar{X}_n] = 2 E[\bar{X}_n] = 2 E[X] = 2 (\beta / 2) = \beta$ and $\text{Bias} = E[\hat{\beta}] - \beta = 0$

—This estimator is unbiased.

Using the fact that the individual values of Xi are independent:
Var $[\hat{\beta}] = \text{Var}\,[2\bar{X}_n] = 4\,\text{Var}\,[\bar{X}_n] = 4\,\text{Var}[X] / n = 4\,(3/4\,\beta^2) / n = 3\,\beta^2 / n$

Hence, MSE $[\hat{\beta}] = 3\,\beta^2 / n + 0 = 3\,\beta^2 / n$
This estimator is consistent since MSE tends to zero as $n \to \infty$

(iii)

MSE $[b\bar{X}_n] = \text{Var}\,[b\bar{X}_n] + (\text{Bias}\,[b\bar{X}_n])^2$

$$= b^2 \frac{3}{4}\frac{\beta^2}{n} + (E[b\bar{X}_n] - \beta)^2$$

$$= b^2 \frac{3}{4}\frac{\beta^2}{n} + (\beta(\frac{b}{2} - 1))^2$$

$$= \frac{\beta^2}{n}(\frac{3}{4}b^2 + n(\frac{b}{2} - 1)^2)$$

Differentiating the MSE respect to b:
$d\,(\text{MSE})/db = \frac{\beta^2}{n}\,(\frac{3}{2}b + n(\frac{b}{2} - 1))$
Setting this equal to 0 gives $b = 2n / (n + 3) = 2 / (1 + \frac{3}{n})$

Differentiating the MSE a second time with respect to b, we obtain:
$d^2(\text{MSE})/db^2 = \frac{\beta^2}{n}\,(\frac{3}{2} + \frac{n}{2})$ which is positive
=> a minimum value for the MSE is at $b = 2 / (1 + 3/n)$.

(iv)

It is seen in (ii) that $\hat{\beta} = 2\bar{X}_n$ is an unbiased estimator. The estimator $b\bar{X}_n$ in (iii) when $b = 2 / (1 + \frac{3}{n})$ is negatively biased estimator as $b < 2$ for $n \geq 1$.

As $n \to \infty$, b tends to 2. So, the estimator in (iii) is also consistent, in view of (ii)

4.

i) $E[X] = \frac{(a+b)}{2}$

$b = 2E[X] - a = 2\bar{x} - a$

$Var(X) = \frac{(b-a)^2}{12}$

$s^2 = \frac{(2\bar{x}-a-a)^2}{12} = \frac{(\bar{x}-a)^2}{3}$

$(\bar{x} - a)^2 = 3\,s^2$

$\hat{a} = \bar{x} - \sqrt{3}s$

ii) $\hat{b} = 2\bar{x} - (\bar{x} - \sqrt{3}s)$

$\hat{b} = \bar{x} + \sqrt{3}s$

iii) Sample: 1, 2, 3, 4, 50

$\bar{x} = 12\,; s = 21.27$

Method of moments estimates using above formulae are:

$\hat{a} = 12 - \sqrt{3}(21.27) = -24.84; \hat{b} = 12 + \sqrt{3}(21.27) = 48.84$

For U (a, b),the probability of a sample point being less than 'a' or greater than 'b' is zero and we have a sample value 50 that is greater than our estimate of 'b'. This highlights a potential weakness of the method of moments.

iv) Likelihood for a sample of size n is $L(b) = \frac{1}{b^n}$ if $b \geq \max(x_i)$,otherwise $L = 0$

Differentiation with respect to b does not work because in the range of x depends on b

We must find b that maximizes $L(b)\ for\ \max(x_i)$ given.. We want b to be as small as possible subject to the constraint that $b \geq \max(x_i)$.

Clearly the maximum is attained at $b = \max(x_i)$.

Hence $\hat{b} = \max(x_i)$.

5. :

i) P(Type I error) is the probability of rejecting H_0 when H_0 is true.

Let X be the no. of hits in first step 12 missiles and Y be the no. of hits in second step 12 missiles.

P(Type I error) = P (X≥ 3 | p = 0.1) + P (X = 0 | p = 0.1) P (Y ≥ 5 | p = 0.1) +
P (X = 1 | p = 0.1) P (Y ≥ 4 | p = 0.1) + P (X = 2 | p = 0.1) P (Y ≥ 3 | p = 0.1)

Using Actuarial Tables page 188 (probabilities for Binomial Distribution)

$$= (1 - 0.8891) + (0.2824)(1 - 0.9957) + (0.6590 - 0.2824)(1 - 0.9744) + (0.8891 - 0.6590)(1 - 0.8891)$$

$$= 0.14727 \approx 15\%$$

ii) Probability of rejecting the null hypothesis when p = 0.3

$$= P(X \geq 3 \mid p = 0.3) + P(X = 0 \mid p = 0.3)\, P(Y \geq 5 \mid p = 0.3) + P(X = 1 \mid p = 0.3)\, P(Y \geq 4 \mid p = 0.3) + P(X = 2 \mid p = 0.3)\, P(Y \geq 3 \mid p = 0.3)$$

Using Actuarial Tables page 188 (probabilities for Binomial Distribution)

$$= (1 - 0.2528) + (0.0138)(1 - 0.7237) + (0.0850 - 0.0138)(1 - 0.4925) + (0.2528 - 0.0850)(1 - 0.2528)$$

$$= 0.91253 \approx 91\%$$

iii) P(type II error) is the probability of accepting H_0 when H_0 is false.
$= 1 - 0.91253 = 0.08747 \approx 9\%$(when p=0.3)

iv) 10 Space agencies in aggregate fired 120 missiles and recorded 40 hits
Assuming that the sample comes from a binomial distribution, we know that the quantity

$$\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Here n = 120, X = 40. so $\hat{p} = \frac{40}{120} = \frac{1}{3} = 0.3333$

Using Actuarial Tables page162, $Z_{10\%} = 1.2816$
Lower bound of 90% right-tailed confidence interval for p is

$$\hat{p} - 1.2816\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.3333 - 1.2816\sqrt{\frac{0.3333(1-0.3333)}{120}} = 0.2782$$

v) Test whether p > 0.278 at 10% level of significance
H_0: P = 0.278 vs. H_1: p > 0.278
For one sample binomial model:

$$\frac{X - np_0}{\sqrt{np_0q_0}} \sim N(0,1)$$ with continuity correction.

$$\frac{39.5 - 120(0.278)}{\sqrt{120(0.278)(0.722)}} = 1.2511$$

We are carrying out a one-sided test. The value of the test statistic is less than

1.2816 (the upper 10% point of the N (0, 1) distribution) so we do not have sufficient evidence to reject H_0 at 10% level.

vi) Lower bound of Confidence interval implies that there is only a 10% chance of 'p' ≤ 0.2782, whereas from the hypothesis test, 'p' could be less than or equal to 0.2780 with probability more than 10% (approx 10.5% corresponding to 1.2511).
The minor disconnect between Confidence interval and Hypothesis testing at the same level is due to

- use of sample proportion $\hat{p}$to estimate population variance in calculating confidence interval, and
- applying continuity correction in hypothesis testing

vii) Test whether there is a difference in the mean scores
We assume that the samples come from normal distributions with the same variance and that the samples are independent.

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$.

The pivotal quantity is: $\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$

Given: $\bar{X}_1 = 65;\ \bar{X}_2 = 70;\ S_1 = 54;\ S_2 = 70;\ n_1 = 12; n_2 = 15$
The pooled variance is: $S_p^2 = \frac{1}{25}(11\,(2916) + 14\,(4900)) = 4027.04$

$$\frac{65-70-0}{63.459*\sqrt{\frac{1}{12}+\frac{1}{15}}} = -0.2034$$

This is within $\pm 2.060\ (= t_{25;2.5\%})$So we have insufficient evidence to reject H_0 at the 5% level. Therefore, it is reasonable to conclude that there is no significant difference in the mean scores for the populations associated with the two Institutes.

6.

i)

$$\sum X_{Public} = 270$$
$$(\bar{X}_{Public}) = 30$$
$$\sum X_{Private} = 315.5$$
$$(\bar{X}_{Private}) = 35.06$$
$$\sum X^2_{Public} = 8614.5$$

(subscript 1 refers public hospitals)

$$S_1^2 = (8614.5 - 9 \times 30^2)/8 = 64.31$$
$$\sum X^2_{Private} = 11599.25$$
$$S_2^2 = (11599.25 - 9 \times 35.06^2)/8 = 67.42$$

ii) Two sided t-test can be applied in case the samples come from populations with equal variances.

[1]

We are testing $H_0 : \sigma_1^2 = \sigma_2^2 \; vs \; H_1 : \sigma_1^2 \neq \sigma_2^2$

Test statistic is $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$

Value of statistic is $\frac{64.31}{67.42} = 0.9542$

$F_{8,8}$ values at 5% levels are 0.2256 and 4.433 . Since the value of the test statistic is between the above values, we have insufficient evidence to reject the hypothesis and conclude that the population variances are equal.

iii) We are testing $H_0 : \mu_1 = \mu_2 \; vs \; H_1 : \mu_1 < \mu_2$

[0.5]

Test statistic is $\frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2}$

Where $S_p^2 = \frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{(n_1+n_2-2)}$

Using the values in section (I) above,

$$S_p^2 = \frac{64.31\times8+67.42\times8}{16} = 65.86$$

value of test statistic is

$$\frac{(35.06-30)-0}{\sqrt{65.86(1/9+1/9)}} = 1.32$$

$$P(t_{16} > 1.32) = 20.5\%$$

This is higher than 95% hence we do not have sufficient evidence to reject the hypothesis and hence conclude that the cost of claims in private hospitals is similar to that in public hospital

7.

(i) $\hat{\theta}$ said to be unbiased when $E(\hat{\theta}) = \theta$

(ii) measure of the 'bias' is given by $E(\hat{\theta}) - \theta$

(iii) Mean Square Error (MSE) of this estimator $\hat{\theta} = (E(\hat{\theta}) - \theta)^2$

(iv) $\tilde{\theta}$ is efficient as an estimator with lower MSE is said to be more efficient than one with higher MSE.

(v) An estimator is termed as consistent if MSE converges to 0 as the sample size tends to ∞

(vi) θ can be estimated using:

a. Method of moments: the population moments are equated to the sample moments to estimate the parameters.

b. Maximum likelihood method: A maximum likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ is generated. A maximum likelihood estimate of the parameter is given by solution to $\frac{dL(\theta)}{d\theta} = 0$

c. Bootstrap method: This is computer intensive method that allows us to avoid making assumption about the sampling distribution by forming an empirical sampling distribution which is possible due to re-sampling based on the available sample.

8.

i) Variable t_k is defined as $t_k \Rightarrow \frac{N(0,1)}{\sqrt{\chi_k^2/k}}$

where k denotes the degrees of freedom
and the two random variables $N(0,1)$ and χ_k^2 are independent.

JARIAL TUDIES

ii) Mean and variance of t_k for k>2 are 0
and k/(k-2) respectively.

iii)

a) We know that for a sample from a normal population,

$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

For the given confidence level $t_9 = 3.25$,

Confidence interval for μ is thus $\left(50 - 3.25 \times \sqrt{\frac{48.667}{10}}, 50 + 3.25 \times \sqrt{\frac{48.667}{10}}\right)$

i.e. (42.83 , 57.17)

b)

From (ii) above, we know that $t_{n-1} \sim N\left(0, \sqrt{\frac{n-1}{n-3}}\right) \sim N(0 , \sqrt{\frac{9}{7}})$

i.e.

$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0 , \sqrt{\frac{9}{7}})$

$\frac{\bar{X}-\mu}{S/\sqrt{7n/9}} \sim N(0 ,1)$

Critical value for given level of confidence is 2.58

Confidence interval for μ is thus $\left(50 - 2.58 \times \sqrt{\frac{48.667}{10} \times \frac{9}{7}}, 50 + 2.58 \times \sqrt{\frac{48.667}{10} \times \frac{9}{7}}\right)$

i.e. (43.54 , 56.45)

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

9.

i) Type I error - Event of Rejecting the hypothesis when it is true

Let X be the random variable denoting the total number of claims on the portfolio. X thus follows Poi(nμ) i.e. Poi(3000) where μ is the Poisson parameter.

Null hypothesis H_0 is thus $X \sim Poi(3000)$

$P(reject\ H_0\ when\ H_0\ is\ true) = P(X > 3100\ when\ X \sim Poi(3000))$

Using normal approximation (as nλ is large enough)

$$X \sim N(3000, 3000)$$

$$P(X > 3100) = P\left(Z > \frac{3100-300}{\sqrt{3000}}\right) = 1 - P(Z < 1.825) = 3.39\%$$

ii) Type II error - Event of Accepting the hypothesis when it is false

iii) Power of a test - Probability of Rejecting the hypothesis when it is false

In terms of μ it is given by:

$$P(reject\ H_0\ when\ H_0\ is\ false) = P(X > 3100\ when\ X \sim Poi(n\mu) \sim N(n\mu, n\mu))$$

$$P(X > 3100) = 1 - \left(Z < \frac{3100-n\mu}{\sqrt{n\mu}}\right)$$

The value of power of test will depend on the value of parameter under alternate hypothesis.

iv) If $\hat{\mu}$ is the estimator of μ (the poisson parameter), $\hat{\mu}$ follows $N(\mu, \hat{\mu}/n)$

Hence $\frac{\hat{\mu}-\mu}{\sqrt{\hat{\mu}/n}}$ follows N(0,1)

Or $P\left(-2.5758 < \frac{2.9-\mu}{\sqrt{2.9/1000}} < 2.5758\right) = 0.99$

Hence the confidence interval for μ is (2.7613, 3.0387)

10.

(i) The Cramér-Rao Lower Bound result holds under very general conditions except where the range of the distribution involves the parameter, such as the uniform distribution in this case.

This is due to a discontinuity, so the derivative in the formula doesn't make sense.

(ii) We have $Y = \max_i X_i$.

Since each X_i lies between 0 and θ, the support for Y will be 0 and θ.

For $0 < y < \theta$,

$$P[Y < y] = P\left[\max_i X_i < y\right]$$

$$= P\left[\bigcap_{i=1}^{n}(X_i < y)\right]$$

$$= \prod_{i=1}^{n} P[X_i < y] \qquad [\because X_i s \text{ are independent}]$$

$$= \prod_{i=1}^{n}\left[\int_0^y \frac{dx}{\theta}\right] = \left(\frac{y}{\theta}\right)^n$$

Thus the probability density function of Y will be:

$$g_Y(y) = \frac{d}{dy}P[Y < y] = \frac{n}{\theta^n}.y^{n-1} \quad for\ 0 < y < \theta$$

Now, for any non-negative real number k,

$$E[Y^k] = \int_0^\theta y^k.\frac{n}{\theta^n}.y^{n-1}\,dy$$

$$= \frac{n}{n+k}\int_0^\theta \frac{n+k}{\theta^n}.y^{n+k-1}\,dy$$

$$= \frac{n}{n+k}.\frac{\theta^{n+k}-0}{\theta^n}$$

OF ACTUARIAL
ATIVE STUDIES

$$= \frac{n\,\theta^k}{n+k}$$

(iii) Bias of the estimator $\hat{\theta}(c)$ is given as:

$$Bias[\hat{\theta}(c)] = E[\hat{\theta}(c) - \theta]$$

$$= E[c\,Y - \theta]$$

$$= c.E[Y] - \theta$$

$$= c.\frac{n\,\theta}{n+1} - \theta \quad [using\ the\ results\ in\ (b) with\ k = 1]$$

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

Mean Square Error of the estimator $\hat{\theta}(c)$ is given as:

$$MSE[\hat{\theta}(c)] = E\left[\{\hat{\theta}(c) - \theta\}^2\right]$$

$$= E[\{cY - \theta\}^2]$$

$$= E[c^2\,Y^2 - 2c\theta\,Y + \theta^2]$$

$$= c^2.E(Y^2) - 2c\,\theta\,E[Y] + \theta^2$$

$$= c^2.\frac{n\,\theta^2}{n+2} - c.\frac{2n\,\theta^2}{n+1} + \theta^2 \quad [using\ (ii)\ with\ k = 1\ \&\ k = 2]$$

ACTUARIAL
VE STUDIES

(iv) For $\hat{\theta}(c)$ to be an unbiased estimator of θ, we need: $Bias[\hat{\theta}(c_u)] = 0$

$$\Rightarrow \left(\frac{c_u\,n}{n+1} - 1\right).\theta = 0 \quad \Rightarrow c_u = \frac{n+1}{n}.$$

(v) In order to minimize $MSE[\hat{\theta}(c)]$, we need:

$$0 = \frac{d}{dc}MSE[\hat{\theta}(c)] = \frac{d}{dc}\left[c^2.\frac{n\,\theta^2}{n+2} - c.\frac{2n\,\theta^2}{n+1} + \theta^2\right] = 2c.\frac{n\,\theta^2}{n+2} - \frac{2n\,\theta^2}{n+1}$$

Thus: $c_m = \frac{2n\theta^2}{n+1}.\frac{n+2}{2n\theta^2} = \frac{n+2}{n+1}.$

$\left(\text{Note: } \frac{d^2}{dc^2}MSE[\hat{\theta}(c)] = \frac{d}{dc}\left[2c.\frac{n\theta^2}{n+2} - \frac{2n\theta^2}{n+1}\right] = \frac{2n\theta^2}{n+2} > 0 \quad \Rightarrow minimum\right)$

(vi) In order to minimize error in estimation, it is preferred to opt for an estimator which has lower mean square error among different competing estimators. So here $\hat{\theta}(c_m)$ will be preferred over $\hat{\theta}(c_u)$.

As n becomes large, $\hat{\theta}(c_u) = \frac{n+1}{n}Y \to Y$. Similarly $\hat{\theta}(c_m) = \frac{n+2}{n+1}Y \to Y$
Thus the two estimators becomes one and same as $n \to \infty$.

11.

	1	2	3	4	5	6	7	8	9	10	Total
x	40	10	100	110	120	150	20	90	80	130	**850**
y	56	62	195	240	170	270	48	196	214	286	**1,737**
xy	2,240	620	19,500	26,400	20,400	40,500	960	17,640	17,120	37,180	**182,560**
x^2	1,600	100	10,000	12,100	14,400	22,500	400	8,100	6,400	16,900	**92,500**

$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i / n = 182560 - 850 * 1737/10 = 34915$

$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 92500 - 850^2/10 = 20250$

$\hat{b} = S_{xy} / S_{xx} = 1.72$

$\hat{a} = \bar{y} - b\bar{x} = (1737/10) - 1.72 * (850/10) = 27.14$

$y = 27.14 + 1.72x$

(b) Gradient represents the amount of hours per rupee spent

12.

i. Fitted Linear Regression Equation

The relevant summary statistics to fit the equation are:

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

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 12666.58 - 12 * \left(\frac{385.2}{12}\right)^2 = 301.66$$

$$S_{xy} = \sum xy - n\overline{xy} = 38191.41 - 12 * \left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right) = 875.16$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 119026.90 - 12 * \left(\frac{1162.5}{12}\right)^2 = 6409.71$$

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.90$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{1162.5}{12}\right) - 2.90 * \left(\frac{385.2}{12}\right) = 3.78$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90\,x}$

ii. Confidence interval for β

Assuming normal errors with a constant variance:

95% confidence interval for β: $\hat{\beta} \pm t_{n-2}(2.50\%) * s.e.(\hat{\beta})$

$$\text{Here: } s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}\right] = 387.07$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{387.07}{301.66}} = 1.13$$

95% confidence interval for β: $2.90 \pm 2.228 * 1.13 = (0.38, 5.42)$

iii. <u>95% confidence intervals for the mean IBM share price</u>

ACTUARIAL
IVE STUDIES

$$\hat{y}_{x_0} \pm t_{n-2}(2.50\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The Dell Share price is US $ 40 (x_0).

$$\hat{y}_{x_0} = 3.78 + 2.90 * 40 = 119.78$$

Thus, 95% Confidence interval:

$$= 119.78 \pm 2.228 * \sqrt{387.07 * \left[\frac{1}{12} + \frac{(40 - 32.1)^2}{301.66}\right]}$$

$$= 119.78 \pm 2.228 * 10.5989$$

$$= (96.17, 143.39)$$

13.

i)

The relevant summary statistics to compute correlation coefficient are:

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 207 - 10 * \left(\frac{39}{10}\right)^2 = 54.90$$

$$S_{xy} = \sum xy - n\overline{xy} = 2853 - 10 * \left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 40508 - 10 * \left(\frac{562}{10}\right)^2 = 8923.60$$

$$\text{Correlation Coefficient } r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{661.20}{\sqrt{54.90}\sqrt{8923.60}} = 0.945$$

ii)

Fitted Linear Regression Equation

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{562}{10}\right) - 12.04 * \left(\frac{39}{10}\right) = 9.23$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x}$

iii)

Relation: $SS_{TOT} = SS_{REG} + SS_{RES}$

$$SS_{TOT} = S_{yy} = 8923.60$$

$$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90} = 960.30$$

$$SS_{REG} = S_{TOT} - S_{RES} = 8923.60 - 960.30 = 7963.30$$

iv)

Coefficient of Determination:

$$R^2 = \frac{S_{xy}{}^2}{S_{xx}\,S_{yy}} = \frac{SS_{REG}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

For the simple linear regression model, the value of the coefficient of determination is the square of the correlation coefficient for the data, since,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} * S_{yy})^{0.5}} = \sqrt{R^2} = \sqrt{0.8924}$$

14.

i) $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 4789.42$

$S_{xy} = 2176.84$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2176.84}{4789.42} = 0.4545$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 17.895 - 0.4545 * 15.83 = 10.79$

The fitted regression equation is $\hat{y} = 10.79 + 0.4545 * x$

ii) $S_{xx} = 4789.42$ from result of part i

$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 1189.21$

$S_{xy} = 2176.84$ from result of part i

TE OF ACTUARIAL
TITATIVE STUDIES

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{9} * \left(1189.21 - \frac{2176.84^2}{4789.42}\right) = 22.20 \quad [2]$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{22.20}{4789.42}} = 0.0681 \quad [1]$$

To test $H_o: \beta = 0 \; v \; H_1: \beta \neq 0$, *the test statistic is*

$$\frac{\hat{\beta}-0}{s.e.(\hat{\beta})} = \frac{0.4545}{0.0681} = 6.674 \quad [1]$$

Under the assumption that the errors of the regression are i.i.d $N(0, \sigma^2)$ random variables, beta has a t distribution with n-2 degrees of freedom. [0.5]

Critical value for t-distribution with 9 degrees of freedom: $t_{9,0.025} = 2.68$. [0.5]

Since the critical value at 95% level of significance is less than the test statistic, there is sufficient evidence to reject the null hypothesis. Hence, it cannot be concluded that there is no statistically significant relationship between x and y. [0.5]

[6]

iii) Pearson's correlation coefficient is computed as: $\frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$ [0.5]

$S_{yy} = 1189.21$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{2176.84}{\sqrt{4789.42 * 1189.21}} = 0.91 \quad [1.5]$$

iv) The estimated value of y corresponding to $x^3 = 25$ *is* $10.79 + 0.4545 * 25 = 22.15$ [1]

$\hat{\sigma}^2 = 22.20$ ... from earlier workings

The variance of the estimator of the mean response is given by

$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.0889}{4789.42}\right] * 22.20 = 2.41$ [2]

The variance of the estimator of the individual response is given by

$\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = [1 + 0.1085] * 22.20 = 24.61$ [2]

Using t_9 distribution, the 95% confidence intervals for mean and individual responses are:

22.20 ± 2.262 *sqrt(2.41) and 22.20 ± 2.262 *sqrt(24.61) = (18.7,25.7) & (11.0,33.4) [2]

[7]

v) The residual plot shows a definite pattern. Although the correlation coefficient is high, the model does not seem to be appropriate. [1]
Using this model leads to underestimation of premium rates at low and high mortality ratings. [1]
[2]

15.

i) Factor analysis / Principal Component Analysis is - A method for reducing the dimensionality of data
It seeks to identify key components necessary to model and understand data [0.5]

Original variables may be
· correlated with each other [0.5]

While Newly identified principal components are chosen to be
· uncorrelated [0.5]
· linear combinations of the original variables of the data [0.5]
· which maximise the variance

ii)

Principal Component	Diagonal entry (PCi)	PCi/ (Sum(PCi) over 1 to 5)	
PC1	0.456	65.0%	of total variance explained by PC1
PC2	0.137	19.5%	of total variance explained by PC2
PC3	0.08	11.4%	of total variance explained by PC3
PC4	0.0165	2.4%	of total variance explained by PC4
PC5	0.012	1.7%	of total variance explained by PC5
	0.7015	**100.0%**	

Correct formula (1 mark)
Sum of PCi (0.5 marks)
Correct calculation (*2.5* Marks)

iii) As 1st 3 Principal components explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset [1]

The 1st 3 Principal Components can then be used for building further classification or regression modelling purpose [1]

[2]

IAQS | & QUANTITATIVE STUDIES

16.

i) The likelihood function for the given sample of observations is:

$$L(\alpha) = \left(\frac{1}{6} + \alpha\right)^7 \left(\frac{1}{2} - 3\alpha\right)^6 \left(\frac{1}{3} + 2\alpha\right)^{12}$$

Taking logs

$$\ln L(\alpha) = 7 \ln\left(\frac{1}{6} + \alpha\right) + 6 \ln\left(\frac{1}{2} - 3\alpha\right) + 12 \ln\left(\frac{1}{3} + 2\alpha\right)$$

ii) MLE:

$$\frac{d}{d\alpha} \ln L(\alpha) = \frac{7}{\left(\frac{1}{6} + \alpha\right)} - \frac{18}{\left(\frac{1}{2} - 3\alpha\right)} + \frac{24}{\left(\frac{1}{3} + 2\alpha\right)}$$

Equating this to zero, we have

$$\frac{7}{\left(\frac{1}{6}+\alpha\right)}-\frac{18}{\left(\frac{1}{2}-3\alpha\right)}+\frac{24}{\left(\frac{1}{3}+2\alpha\right)}=0$$

This gives $7\left(\frac{1}{2}-3\alpha\right)\left(\frac{1}{3}+2\alpha\right)-18\left(\frac{1}{6}+\alpha\right)\left(\frac{1}{3}+2\alpha\right)+24\left(\frac{1}{6}+\alpha\right)\left(\frac{1}{2}-3\alpha\right)=0$

$= (1+6\alpha)\left[\frac{7}{3}\left(\frac{1}{2}-3\alpha\right)-\frac{18}{3}\left(\frac{1}{6}+\alpha\right)+\frac{24}{6}\left(\frac{1}{2}-3\alpha\right)\right]=0$

Thus, either $(1+6\alpha)=0$ or $\frac{7}{3}\left(\frac{1}{2}-3\alpha\right)-\frac{18}{3}\left(\frac{1}{6}+\alpha\right)+\frac{24}{6}\left(\frac{1}{2}-3\alpha\right)=0$ giving

$\alpha=-\frac{1}{6}$ or $\frac{13}{6}-25\alpha=0$.

That is $\alpha=-\frac{1}{6}=-0.0167$ or $\alpha=\frac{13}{150}=0.0867$. [roots of log likelihood equation]

However, one of the roots $\alpha=-\frac{1}{6}$ is inadmissible as MLE since it does not lie in the range of α.

$\frac{\partial^2 \ln L(\alpha)}{\partial\alpha^2}=-\frac{7}{\left(\frac{1}{6}+\alpha\right)^2}-\frac{54}{\left(\frac{1}{2}-3\alpha\right)^2}-\frac{48}{\left(\frac{1}{3}+2\alpha\right)^2}<0$ for $\alpha=0.0867$ confirming the maximum.

[7]

iii) We have one unknown parameter, so we will use $E(X)=\bar{x}$.

$E(X)=-1\left(\frac{1}{6}+\alpha\right)+0\left(\frac{1}{2}-3\alpha\right)+1\left(\frac{1}{3}+2\alpha\right)=\frac{1}{6}+\alpha$

From the data, we have: $\bar{x}=\frac{-1(7)+0(6)+1(12)}{25}=\frac{5}{25}=0.2$

Therefore: $\frac{1}{6}+\alpha=0.2 \Rightarrow \hat{\alpha}=\frac{1}{30}=0.0333$

[2]

IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES