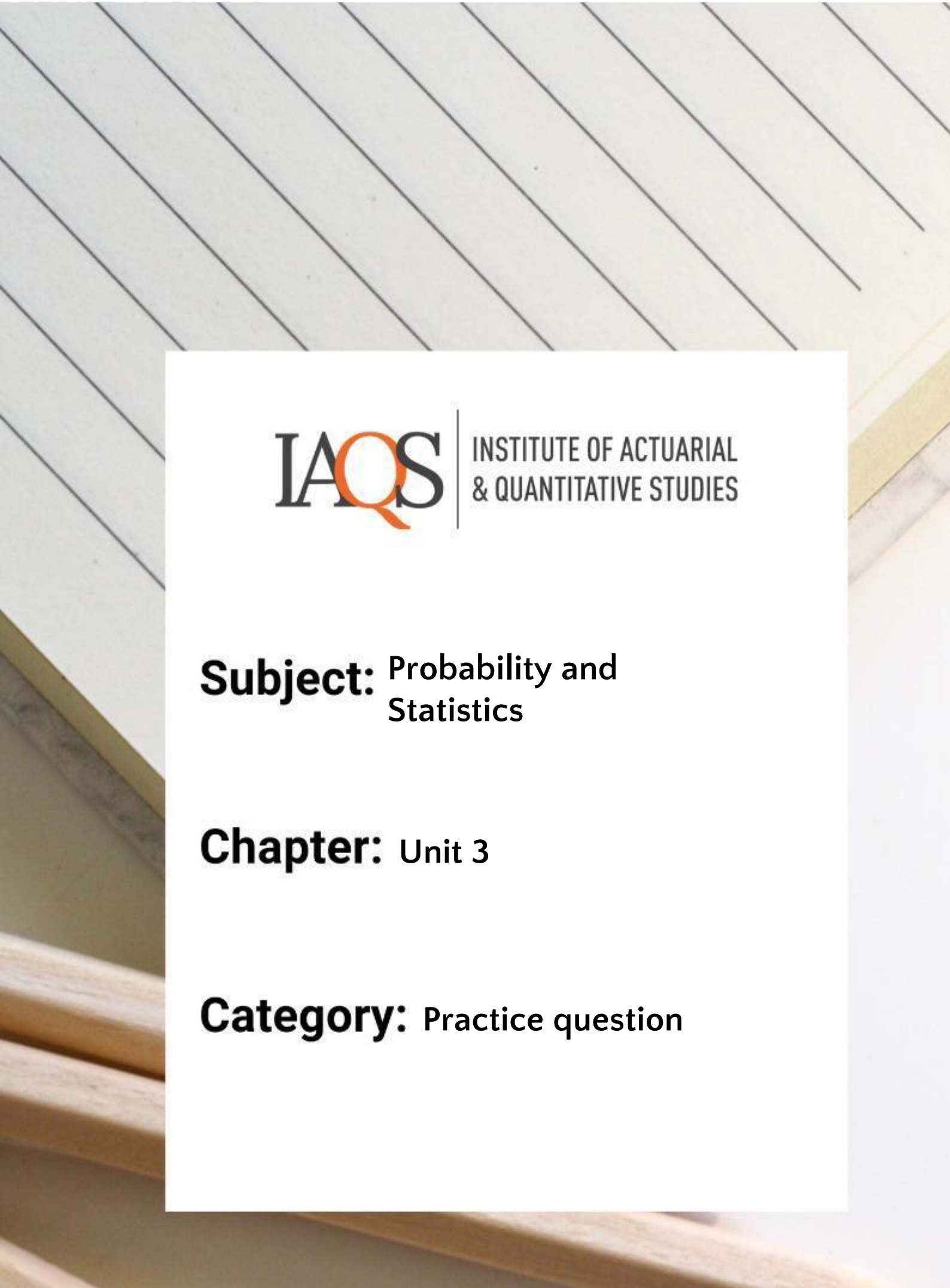


INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics

Chapter: Unit 3

Category: Practice question

Part 1

- **Central Limit Theorem**
- **Sampling Distributions**
- **Theory of Estimation 1**

1. CT3 April 2013 Question 4

Consider a random sample, X_1, X_2 … X_n, from a normal $N(\mu, \sigma^2)$ distribution, with sample mean $\underline{X}$ and sample variance S^2.

i. Define carefully what it means to say that X_1, X_2 … X_n is a random sample from a normal distribution.

ii. State what is known about the distributions of $\underline{X}$ and S^2 in this case, including the dependencies between the two statistics.

iii. Define the t-distribution and explain its relationship with $\underline{X}$ and S^2.

Ans:

(i) The random variables $X1$,… , Xn are independent and identically distributed with $Xi \sim N(\mu,\sigma^2)$

(ii) $\underline{X}$ and S^2 are independent

$$\underline{X} \sim N(\mu,\sigma^2 / n)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

(iii) $t_k = N(0,1)/\sqrt{\chi^2_k/k}$ where N(0,1) and χ^2_k are independent

This result can be applied here, and we get $\frac{\frac{X-\mu}{S}}{\sqrt{n}} \sim t_{n-1}$

2. CT3 April 2013 Question 7

A regulator wishes to inspect a sample of an insurer's claims. The insurer estimates that 10% of policies have had one claim in the last year and no policies had more than one claim. All policies are assumed to be independent.

i. Determine the number of policies that the regulator would expect to examine before finding 5 claims.

On inspecting the sample claims, the regulator finds that actual payments exceeded initial

estimates by the following amounts:
£35 £120 £48 £200 £76

ii. Find the mean and variance of these extra amounts.

It is assumed that these amounts follow a gamma distribution with parameters α and λ.
iii. Estimate these parameters using the method of moments.

Ans:
(i) Expected no of inspected policies =50
(ii) Mean = 95.8, Variance = 4454.2
(iii)α =2.06, λ= 0.0215

3. CT3 October 2013 Question 2
An insurance company experiences claims at a constant rate of 150 per year. Find the approximate probability that the company receives more than 90 claims in a period of six months.

Ans: 0.037

4. CT3 October 2013 Question 3
The random variable X has a distribution with probability density function given by:

$$f(x)=\begin{cases}\frac{2x}{\theta^2} & ; \quad 0\le x\le\theta \\ 0 & ; \quad x<0 \text{ or } x>\theta\end{cases}$$

where θ is the parameter of the distribution.
i. Derive expressions in terms of θ for the expected value and the variance of X.

Suppose that $X_1, X_2 \ldots X_n$ is a random sample, with mean, $\underline{X}$ from the distribution of X.
ii. Show that the estimator $\hat{\theta} = \frac{3\underline{X}}{2}$ is an unbiased estimator of θ.

Ans:
(i) E(X) = $\frac{2\theta}{3}$, V(X) = $\frac{\theta^2}{18}$, (ii) -

Unit 3

PRACTICE QUESTION

5. CT3 October 2013 Question 4

An actuary is considering statistical models for the observed number or claims, X, which occur in a year on a certain class of non-life policies. The actuary only considers policies on which claims do actually arise. Among the considered models is a model for which:

$$P(X=x)=-\frac{1}{\log(1-\theta)}\frac{\theta^x}{x}, \quad x=1, 2, 3, \ldots$$

where θ is a parameter such that $0 < \theta < 1$.

Suppose that the actuary has available a random sample $X_1, X_2 \ldots, X_n$, with sample mean $\underline{X}$.

i. Show that the method of moments estimator (MME), $\tilde{\theta}$ satisfies the equation:

$$\bar{X}(1-\tilde{\theta})\log(1-\tilde{\theta})+\tilde{\theta}=0.$$

ii. (a) Show that the log likelihood of the data is given by:

$$l(\theta) \propto -n\log\{-\log(1-\theta)\}+\sum_{i=1}^{n} x_i \log(\theta).$$

(b) Hence verify that the maximum likelihood estimator (MLE) of θ is the same as the MME.

iii. Suggest two ways in which the MLE of θ can be computed when a particular data set is given.

6. CT3 October 2013 Question 5

Consider a random sample consisting of the random variables $X_1, X_2 \ldots, X_n$ with mean μ and variance σ^2. The variables are independent of each other.

i. Show that the sample variance, S^2, is an unbiased estimator of the true variance σ^2.

Now consider in addition that the random sample comes from a normal distribution, in which case it is known that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

ii. (a) Derive the variance of S^2 in terms of σ and n.

(b) Comment on the quality of the estimator S^2 with respect to the sample size n.

Unit 3

PRACTICE QUESTION

Ans:

(i) –

(ii) a) var $(S^2) = \frac{2\sigma^4}{n-1}$

b) Estimator gets better (more accurate) as n increases, as its variance reduces.

7. CT3 April 2014 Question 8

Let $X_1, X_2 \ldots, X_n$ be a random sample from a distribution with parameter θ and density function:

$$f(x) = \begin{cases} \frac{2x}{\theta^2} & ; \quad 0 \le x \le \theta \\ 0 & ; \quad x < 0 \text{ or } x > \theta \end{cases}$$

Suppose that $\underline{x} = (x_1, x_2, \ldots, x_n)$ is a realization of $X_1, X_2 \ldots, X_n$.

i. (a) Derive the likelihood function $L(\theta; \underline{x})$ and produce a rough sketch of its graph.

(b)Use the graph produced in part (i)(a) to explain why the maximum likelihood estimate of θ is given by $X_{(n)} = \max\{X_1, X_2 \ldots, X_n\}$.

Let $X_{(n)} = \max\{X_1, X_2 \ldots, X_n\}$ be the estimator of θ, that is the random variable corresponding to $x_{(n)}$.

ii. (a) Show that the cumulative distribution function of the estimator $X_{(n)}$ is given by:

$$F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^{2n}$$

for $0 \le x \le \theta$.

(b) Hence, derive the probability density function of the estimator $X_{(n)}$.

(c) Determine the expected value $E(X_{(n)})$ and the variance $V(X_{(n)})$.

(d) Show that the estimator $\frac{2n+1}{2n} X_{(n)}$ is an unbiased estimator of θ.

iii. (a) Derive the mean square error of the estimator given in part (ii)(d).

(b) Comment on the consistency of this estimator.

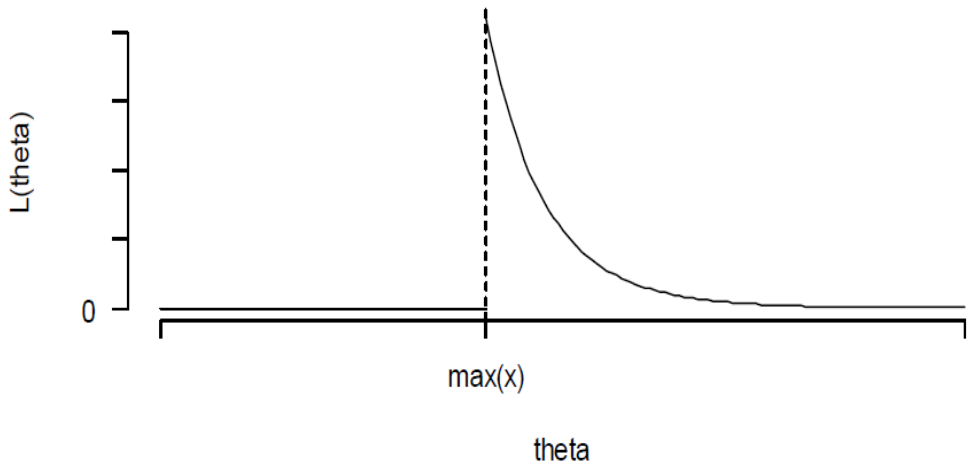
Ans:

(i)

(a) Likelihood is given as

$$L(\theta;\underline{x}) = \begin{cases} \Pi_{i=1}^{n} f(x_i;\theta) = \frac{2^n x_1 x_2 \cdots x_n}{\theta^{2n}} & \text{if } \theta \geq x_{(n)} = \max\{x_1, x_2, \cdots, x_n\} \\ 0 & \text{if } \theta < x_{(n)}. \end{cases}$$

Its graph is given below:



b. From the graph, the likelihood is maximised at $\theta = x_n = \max\{x_1, x_2, \ldots x_n\}$

(ii) a. –

$$f_{X_{(n)}}(x) = \begin{cases} \frac{2nx^{2n-1}}{\theta^{2n}} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

b.

c. $E(X_n) = \frac{2n\theta}{2n+1}$, $V(X_n) = \frac{n\theta^2}{(n+1)(2n+1)^2}$

d. –

(iii) a. $MSE = \frac{\theta^2}{4n(n+1)}$

b. We have $MSE \to 0$ as $n \to \infty$, therefore the estimator is consistent.

Unit 3

PRACTICE QUESTION

8. CT3 September 2014 Question 7

Consider the following discrete distribution with an unknown parameter p for the distribution of the number of policies with 0, 1,2, or more than 2 claims per year in a portfolio of n independent policies.

number of claims	0	1	2	more than 2
Probability	2p	p	0.25p	1-3.25p

We denote by X_0 the number of policies with no claims, by X_1 the number of policies with one claim and by X_2 the number of policies with two claims per year. The random variable $X = X_0 + X_1 + X_2$ is then the number of policies with at most two claims.

i. Derive an expression for the maximum likelihood estimator $\hat{p}$ of parameter p in terms of X and n.

ii. Show that the estimator obtained in part (i) is unbiased.

The following frequencies are observed in a portfolio of n = 200 policies during the year 2012:

number of claims	0	1	2	more than2
observed frequency	123	58	13	6

A statistician proposes that the parameter p can be estimated by $\tilde{p}$= 58/200 = 0.29 since p is the probability that a randomly chosen policy leads to one claim per year.

iii. Estimate the parameter p using the estimator derived in part (i).

iv. Explain why your answer to part (iii) is different from the proposed estimated value of 0.29.

An alternative model is proposed where the probability function has the form:

number of claims	0	1	2	more than2
probability	P	2p	0.25p	1-3.25p

v. Explain how the maximum likelihood estimator suggested in part (i) needs to be adapted

to estimate the parameter p in this new model.

(iv) Suggest a suitable test to use to make a decision about which of the two models should be used based on empirical data.

Ans:

(i) $\hat{p} = \frac{X}{3.25n}$

(ii) –

(iii) $\hat{p}$ =0.2985

(iv) The MLE in part (iii) takes the structure of the entire probability function into account while the estimator 58/200 only considers the number of policies with one claim.

(v) No change required, since the MLE $\hat{p}$ turns out to dependent only on the total number of policies with less than three claims.

(vi) χ^2 - test

9. CT3 April 2015 Question 6

Let X_1, X_2, … , X_6 be a random sample from a population following a Gamma(2,1) distribution. Consider the following two estimators of the mean of this distribution:

$$\hat{\theta}_1 = \bar{X} \text{ and } \hat{\theta}_2 = \frac{9}{30}(X_1 + X_2 + X_3) + \frac{1}{30}(X_4 + X_5 + X_6)$$

where $\underline{X}$ is the mean of the sample.

i. Determine the sampling distribution of $\underline{X}$ using moment generating functions.

ii. Derive the bias of each estimator $\widehat{\theta_1}$ and $\widehat{\theta_2}$

iii. Derive the mean square error of each estimator $\widehat{\theta_1}$ and $\widehat{\theta_2}$

iv. Compare the efficiency of the two estimators $\widehat{\theta_1}$ and $\widehat{\theta_2}$

Ans:

(i) Gamma(12, 6) distribution.

(ii) Bias $(\widehat{\theta_1}) = 0$ Bias$(\widehat{\theta_2})= 0$

(iii) MSE$(\widehat{\theta_1}) = 0.333$, MSE$(\widehat{\theta_2})= 0.547$

(iv) $\widehat{\theta_1}$ has smaller MSE, and therefore is more efficient than $\widehat{\theta_2}$.

10. CT3 October 2015 Question 10

The random variables X_1, X_2,…, X_n are independent from each other and all follow a Poisson distribution with parameter λ.

i. Derive the maximum likelihood estimator of λ based on X_1, X_2,…, X_n. You are not required to verify that your answer corresponds to a maximum.

Unit 3

PRACTICE QUESTION

11. CT3 April 2016 Question 5

Players A and B play a game of "heads or tails", each throwing 50 fair coins. Player A will win the game if she throws 5 or more heads than B; otherwise, B wins. Let the random variables X_A and X_B denote the numbers of heads scored by each player and $D = X_A - X_B$.

i. Explain why the approximate asymptotic distribution of D is normal with mean 0 and variance 25.

ii. Determine the approximate probability that player A wins any particular game, based on your answer in part (i).

Ans:

(i) From CLT: $D \sim N(0, 25)$, (ii) 0.1841

12. CT3 April 2016 Question 6

A statistician is sent a summary of some data. She is told that the sample mean is 9.46 and the sample variance is 25.05. She decides to fit a continuous uniform distribution to the data.

i. Estimate the parameters of the distribution using the method of moments.

The full data are sent later and are given below:

3.5 5.4 7.3 8.5 9.2 10.3 11.4 20.1

ii. Comment on the results in part (i) in the light of the full data.

Ans:

(i) $\hat{a} = 0.791$, $\hat{b} = 18.129$

(ii) The largest observation is greater than our estimate of b in part (i). This would suggest the uniform distribution is not a good fit to this data, or the largest observation is a mistaken observation. This also highlights a potential weakness of the method of moments.

13. CT3 April 2016 Question 7

A random sample is taken from an exponential distribution with parameter λ. The sample contains some censored observations for which we only know that the value is greater than 3. The observed values are given in the following table:

I	1	2	3	4	5	6	7	8	9	10
X_i	1.3	1.8	2.1	2.2	2.2	2.4	>3	>3	>3	>3

Estimate the parameter λ using the method of maximum likelihood. You are not required to verify that your answer corresponds to the maximum.

Ans: $\hat{\lambda} = 0.25$

14. CT3 October 2016 Question 9

A statistical model is used to describe the total loss, S (in pounds), experienced in a certain portfolio of an insurance company over a period of one year. The total loss is given by:

$S = X_1 + X_2 + \cdots + X_N$

where X_i gives the size of the loss from claim i = 1,…,N.

N is a random variable representing the number of claims per year and follows a Poisson distribution. The $X_i's$ are independent, identically distributed according to a gamma distribution with parameters α and λ, and are also independent of N.

Data from previous years show that the average number of claims per year was 14, while the average size of claims was £500 and their standard deviation was £150.

i. Estimate the parameters α and λ using the method of moments.

ii. Estimate the mean and the variance of the total loss S using the information from the data above.

Now suppose that the value of parameter α is known to be equal to α* and n = 5 claims have been made in a particular year with average size again £500.

(iii) (a) Derive an expression for the maximum likelihood (ML) estimate of the parameter λ in terms of α*. You should verify that your answer corresponds to a maximum.
(b) Derive the asymptotic distribution of the ML estimator of the parameter λ in terms of α*.
(c) Comment on the validity of the distribution in part (iii)(b).

Now suppose that the values of both parameters α and λ are unknown and n claims have been made in a particular year.

Unit 3

PRACTICE QUESTION

(iv) (a) Show that the ML estimate, $\hat{\alpha}$ of the parameter α needs to satisfy the equation:

$$\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log(\bar{x}) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

where Γ'($\hat{\alpha}$) denotes the first derivative of Γ'($\hat{\alpha}$) with respect to $\hat{\alpha}$

(b) Comment on how the ML estimates of the parameters α and λ can be obtained in this case.

Ans:

(i) $\tilde{\alpha}$ =11.1111, $\tilde{\lambda}$ = 0.022

(ii) E(S) = 7,000, V(S) =3,815,000

(iii) a. $\hat{\lambda} = \frac{\alpha*}{500}$

b. CRLB = $\frac{\lambda^2}{n\alpha*}$, $\hat{\lambda} \sim N(\lambda, \lambda^2 / 5\alpha^*)$

c. The approximation of the distribution relies on a very small sample (n = 5) and therefore may not be valid.

(iv) a. –

b. The equation for $\hat{\alpha}$ cannot be solved analytically, so a numerical solution is required. Then $\hat{\alpha}$ can be substituted in the equation for $\widehat{\lambda}$.

15. CT3 April 2017 Question 5

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, identically distributed random variables with finite mean μ and finite (non-zero) variance σ^2.

i. State the central limit theorem (CLT) in terms of the sum $\sum_{i=1}^{n} X_i$

Assume now that each X_i, i = 1, 2, …, 50, follows an exponential distribution with parameter λ = 2 and let $Y = \sum_{i=1}^{50} X_i$

ii. Determine the approximate distribution of Y together with its parameters using the CLT.

iii. State the exact distribution of Y together with its parameters.

iv. Comment on the shape of the distribution of Y based on your answers to parts (ii) and (iii).

Ans:

(i) The CLT states that as $n \to \infty$, approximately, $\sum_{i=1}^{n} X_i$ approaches the N($n\mu$, $n\sigma^2$) distribution.

Unit 3

PRACTICE QUESTION

(ii) $Y = \sum_{I=1}^{50} X_i \sim N(25, 12.5)$

(iii) $Y = \sum_{I=1}^{50} X_i \sim$ Gamma(50,2)

(iv) Generally, the gamma distribution is an asymmetric distribution. Here, as n is large, the CLT suggests that the distribution of Y is approximately normal, and therefore symmetric.

16. CT3 April 2017 Question 7

An investigation at a large airport focuses on the delay with which i flights arrive. The delay time X, in minutes, is the difference between the actual time of arrival and the scheduled arrival time of delayed flights. Assume that X has an exponential distribution with parameter $\lambda > 0$.

i. Derive the estimator $\hat{\lambda}$ for λ using the method of moments.

The following table shows the observed values of X for a random sample of ten delayed flights.

45 20 120 90 60 30 45 90 60 150

ii. Estimate the value of λ for this sample using the method of moments.

To gain further insight into the distribution of flight delays, it is suggested that the time at which a flight is scheduled to arrive during a day has an impact on the delay.

Therefore, assume now that X_i has an exponential distribution with a parameter λ that depends on the scheduled arrival time as follows:

$X_i \sim Exp(\lambda_i)\ with\ \lambda_i = \theta Z_i$

where the random variable Z_i describes the scheduled arrival time (in minutes) after midnight on the day of arrival for the i^{th} randomly selected delayed flight and $\theta > 0$ is a parameter in this model.

iii. Derive the maximum likelihood estimator $\hat{\theta}$ for the parameter θ. You should show that your solution is indeed a maximum.

Ans:

(i) $\hat{\lambda} = \frac{1}{\bar{X}}$

(ii) $\hat{\lambda} = 0.014085$

(iii) $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} Z_i X_i}$

Unit 3

PRACTICE QUESTION

17. CT3 April 2017 Question 8

An actuary models the number of claims X per year per policy as a discrete random variable with the following distribution:

Number of claims	0	1	2	3	More than 3
Probability	*	P	p/2	p/4	p/8

where p is an unknown parameter.

i. Show that P[X=0] $=\frac{8-15p}{8}$

ii. Determine the range of possible values of p.

In a sample of n independent policies there are N_0 policies with no claims during a year, N_1 policies with one claim, N_2 policies with two claims and N_3 policies with three claims. There are also some policies with more than three claims.

iii. Show that the maximum likelihood estimator $\hat{p}$ for p based on observations of N_0, … , N_3 in a sample of n independent claims is given by:

$$\hat{p}=\frac{8}{15}\frac{n-N_0}{n}$$

You do not need to check that your solution is a maximum.

iv. Explain why the distribution of N_0 is a Binomial distribution specifying its parameters.

v. Verify that $\hat{p}$ is an unbiased estimator for p.

Assume that in a sample of size n = 300 there were 100 policies with no claims during the previous year.

vi. Determine the value of the variance of the estimator $\hat{p}$.

The insurance company has now decided to limit the maximum number of claims per year to four per policy, but otherwise continue to use the distribution above. The claim amount of any individual claim is assumed to have a normal distribution with expectation 100 and standard deviation 20. Let S denote the total amount claimed in a portfolio of 300 independent policies during a year. We assume that claim amounts are independent of each other and independent of the number of claims.

Let X be the number of claims per policy per year and Y be the total number of claims per year.

vii. (a) Show that E(X) = 3.25p and Var(X) = 7.25p − 10.5625p^2.

Assume now that p = 0.2

(b) Determine E(Y) and Var(Y).
(c) Determine the expected value and the standard deviation of S.

Ans:
(i) –
(ii) p $\epsilon(0,\frac{8}{15})$
(iii) –
(iv) N_0 has a binomial distribution since it counts the outcome "no claim" in n independent trials. The distribution of N_0 is B(n, $\frac{8-15p}{8}$)
(v) –
(vi) $\hat{p}$=0.0002107
(vii) a. –
b. E(Y) = 195, Var(Y) = 308.25
c. E(S)= 19,500 , SD(S) = 1,777.78

18. CT3 September 2017 Question 8

The two random variables X_1 and X_2 are independent from each other and follow a uniform $U(-\theta, \theta)$ distribution, where $\theta > 0$ is a parameter.

Let $\widehat{\theta_1} = 3Z$ denote a possible estimator of θ , where $Z = \max(X_1, X_2)$.

i. Show that the probability density function of Z is given by $f_Z(z) = \frac{z+\theta}{2\theta^2}$ by first deriving its cumulative distribution function.

ii. Show that $E(Z) = \frac{\theta}{3}$

iii. a.Derive the bias of $\widehat{\theta_1}$

b.Derive an expression for the mean squared error (MSE) of $\widehat{\theta_1}$ in terms of the unknown parameter θ.

Let $\widehat{\theta_2} = 2Z$ denote a different estimator of θ, where again Z = max(X_1, X_2).

iv. a.Show that bias $(\widehat{\theta_2}) = \frac{-\theta}{3}$

b.Show that MSE$(\widehat{\theta_2}) = \theta^2$

v. Comment on how good the two estimators are, based on your answers in parts (iii) and (iv).

Ans:

(i) –

(ii) –

(iii) a. bias $(\widehat{\theta_1})$=0

b. MSE $(\widehat{\theta_1})= 2\theta^2$

(iv) a. –

b. –

(v) $\widehat{\theta_1}$ is unbiased, but has much larger MSE compared to $\widehat{\theta_2}$ (by factor of 2). On the other hand $\widehat{\theta_2}$ has considerable bias, equal to a third of the true value of the parameter.

19. CS1 April 2019 Question 5

i.State the central limit theorem for independent identically distributed random variables $X_1, X_2 \ldots, X_n$ with finite mean μ and finite (non-zero) variance σ^2.

ii.Show that if the random variable B has the binomial distribution with parameters (n,p), then $\frac{B-np}{\sqrt{np(1-p)}}$

approximately follows a standard normal distribution for large n, using the central limit theorem.

Two players have played a large number of independent games. In a sample of 100 of these games, one player has won 57 games and the other player has won 43.

iii.Derive a 95% confidence interval for the probability p that the first player wins a given game, using the normal approximation in part (ii).

Unit 3

PRACTICE QUESTION

Ans:

(i) The distribution of $\frac{\underline{X}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal distribution, $N(0,1)$ as n tends to infinity.

(ii) –

(iii) (0.473, 0.667)

20. CS1 September 2019 Question 2

Let $X_1, X_2, \ldots, X_n$ be a random sample consisting of independent random variables with mean μ and variance σ^2. Consider the sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

i. Derive the expected value of $\underline{X}$.

ii. Derive the variance of $\underline{X}$

iii. Comment on the variance of variable $\underline{X}$ as compared to the variance of X_i.

An actuary is interested in exploring the difference in the size of claim losses from two insurance portfolios, and can take samples of claims from these portfolios.

iv. Explain how the answer to part (iii) can affect the precision of the actuary's comparison.

Ans:

(i) $E(\underline{X}) = \mu$

(ii) $V(\underline{X}) = \frac{\sigma^2}{n}$

(iii) The variance of the sample mean is smaller compared to the variance of individual variables.

(iv) Individual values are less precise than the average of a sample. Larger sample leads to smaller variance.

21. CS1A April 2021 Q1

A random variable, X, is modeled using a gamma distribution with parameters $\alpha = 50$ and $\lambda = 0.25$.

i. Calculate an approximate value for P(X > 270) using the chi-square distribution.

ii. Calculate an approximate value for P(X > 270) using the central limit theorem.

iii. Comment on the difference between your answers to parts (i) and (ii).

Unit 3

PRACTICE QUESTION

Ans:

(i) 0.01

(ii) 0.0066642

(iii) The gamma distribution converges to the normal distribution as $\alpha \to \infty$. But for α=50, the gamma distribution exhibits positive skew, and gives a higher tail probability than the symmetric normal distribution.

22. CS1A April 2021 Q4

Consider a random sample of size n = 25 from a Normal distribution with mean 10, variance 4 and sample variance S^2.

i. Write down the sampling distribution of S^2.

ii. Calculate, using your answer in part (i), the expected value of S^2.

iii. Calculate, using your answer in part (i), the variance of S^2.

Ans:

(i) $6S^2 \sim \chi^2_{24}$, (ii) $E[S^2]=4$, (iii) $\text{var}[S^2]=4/3$

23. CS1A April 2021 Q6

A tutor believes that the number of exams passed by students sitting three different exams follows a binomial distribution with parameters n = 3 and p. A random sample of 120 students showed the following results:

Number of exams passed	0	1	2	3
Number of students	40	60	15	5

(i) (a) Identify which one of the following corresponds to the log likelihood function of p given the observed data:

A $\log L \propto 255\log(1-p) + 105\log(p)$

B $\log L \propto 115\log(1-p) + 80\log(p)$

C $\log L \propto 265\log(1-p) + 115\log(p)$

D $\log L \propto 175\log(1-p) + 85\log(p)$

(b) Show, using your answer to part (i)(a), that the maximum likelihood estimate for p is $\hat{p}$ = 0.2917. You are not required to check that it is a maximum.

Ans:

(i) a. A

b.-

Unit 3

PRACTICE QUESTION

24. CS1A September 2021 Q1

A random sample of size 15 is taken from a Normal distribution with mean 19 and variance 2.

i. Write down the sampling distribution of S^2.

ii. Explain why your answer in part (i) is valid for this random sample.

Ans:

(i) $7S^2 \sim \chi^2_{14}$

(ii) The underlying sample is from the Normal distribution, hence the chi-squared distributional assumption for the sample variance holds true.

25. CS1A April 2022 Q4

i. Describe what is meant by each of the following:

(a) A random sample

(b) A statistic.

A new political party is interested in the level of support it would have among the voters in a particular country. The random variable X is defined as:

$$X = \begin{cases} 1, & \text{if the voter would support the party,} \\ 0, & \text{otherwise.} \end{cases}$$

A random sample of 50 voters are presented with a simple summary of the party's policies and asked if they would support this new party. The random sample is represented by X_1, X_2, …, X_{50}.

ii. (a) Identify a suitable population together with a possible parameter of interest.

(b) Determine, using your answer to part (ii)(a), the sampling distribution of the statistic:

$$Y = \sum_{i=1}^{50} X_i$$

Ans:

(i) a. A random sample is made up of independent and identically distributed random v variables, typically denoted as $X_1,\ldots X_n$.

Unit 3

PRACTICE QUESTION

b. A statistic is a function of random variables. It will be a random variable itself and will have
distribution, its sampling distribution. A statistic does not involve any unknown parameters.

(ii) a. A suitable population in this case is the set of all voters. In terms of the random variable X it will consist of a series of 1s and 0s depending on whether an individual voter would or would not support the new party. The parameter of interest is pp, representing the proportion of 1s in the population. i.e. the proportion of voters in the population that support the party.

b. Y is the number of voters who would support the party. Since the sample is random, therefore each observation is independent, p is constant and the responses are either success (1) or failure (0), then Y will have a binomial distribution with n=50 and parameter p , i.e. $Y \sim Bin(50,p)$.

26. CS1A APRIL 2022 Q5

Let X_1, X_2, …, X_n be independent identically distributed random variables following a Poisson(m) distribution. Suppose that, rather than observing the random variables precisely, only the events $X_i = 0$ or $X_i > 0$ are observed for i = 1, 2, …, n.

Let Y be a random variable with:

$$Y_i = \begin{cases} 0, & X_i = 0 \\ 1, & X_i > 0 \end{cases}$$

for i = 1, 2, …, n.

i. Explain why the distribution of Y_i is a Bernoulli (p) distribution with parameter p = $1 - e^{-m}$.

ii. Identify which one of the following expressions gives the correct likelihood function based on observations y_1, …, y_n in terms of $\underline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and the unknown parameter m.

A $L(m) = (1+e^{-m})^{n\bar{y}}(e^{m})^{n-n\bar{y}}$

B $L(m) = (1-e^{m})^{n\bar{y}}(e^{-m})^{n-n\bar{y}}$

C $L(m) = (1-e^{-m})^{n\bar{y}}(e^{-m})^{n-n\bar{y}}$

D $L(m) = (1-e^{-m})^{n\bar{y}}(e^{-m})^{n+n\bar{y}}$

Unit 3

PRACTICE QUESTION

iii. Derive an expression for the Maximum Likelihood Estimate (MLE) $\hat{m}$ of m in terms of $\underline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

iv. State the condition that $\hat{m}$ and L(m) must satisfy for $\hat{m}$ to maximise the likelihood function.

Ans:

(i) Y_i are independent random variables with only two outcomes. The two outcomes are 0 and 1 with 1 having a "success" probability of $p = 1 - e^{-m}$. This is the definition of the Bernoulli(p) distribution.

(ii) Option C

(iii) $\hat{m} = -\log\log\left(1 - \underline{y}\right)$

(iv) The second derivative of L evaluated at $\hat{m}$ must be strictly negative, that is $\frac{\partial^2}{\partial m^2} L(\hat{m}, y_1, y_2, \ldots\ldots y_n) < 0$

27. CS1A APRIL 2022 Q6

The size of claims on a certain type of motor insurance policy are modelled as a random variable X with Probability Density Function (PDF)

$$f(x; \alpha, \beta) = \alpha \frac{\beta^{\alpha}}{x^{\alpha+1}}, \qquad x \geq \beta, \quad \alpha, \beta > 0.$$

i. Identify which one of the following expressions gives the correct log likelihood function in terms of a random sample $(x_1, x_2, \ldots, x_n)$ and the unknown parameters α and β:

A $\quad l(\alpha, \beta) = n\log\alpha + n\alpha\log\beta + (\alpha+1)\sum_{i=1}^{n}\log x_i$

B $\quad l(\alpha, \beta) = \log\alpha + n\alpha\log\beta - (\alpha+1)\sum_{i=1}^{n}\log x_i$

C $\quad l(\alpha, \beta) = n\log\alpha + n\log\beta - (\alpha+1)\sum_{i=1}^{n}\log x_i$

D $\quad l(\alpha, \beta) = n\log\alpha + n\alpha\log\beta - (\alpha+1)\sum_{i=1}^{n}\log x_i$

ii. Derive the MLE $\hat{\alpha}$ of parameter α as a function of parameter β, for a random sample $(x_1, x_2, \ldots, x_n)$.

iii. Comment on the behaviour of the PDF of X when β increases.

Unit 3

PRACTICE QUESTION

iv. Determine the MLE $\hat{\beta}$ of parameter β based on your comment in part (iii).

The values (in $) of a sample of 10 claims are given in the table below:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
10,000	12,000	8,000	16,000	20,000	19,000	17,000	22,000	18,000	5,000

v. Calculate the mean and standard deviation of the natural logarithm of the sample.

vi. Calculate the MLEs $\hat{\alpha}$ and $\hat{\beta}$ based on the sample.

Ans:

(i) Option D

(ii) $\text{MLE}(\hat{\alpha})=\frac{n}{-n\log\beta+\sum_{i=1}^{n}\log\log x_i}$

(iii) The PDF increases as β increases. Also, the support of the PDF (i.e. $\{x:x\geq\beta\}$) moves to the right.

(iv) The MLE of β is the smallest value of x_i.

(v) Mean (log)= 9.508, SD(log)= 0.476

(vi) $\hat{\alpha}$=1.009, β=5000

28. CS1A APRIL 2023 Q4

Statisticians A and B obtain independent samples $X_1, \ldots, X_{10}$ and $Y_1, \ldots, Y_{17}$ respectively, both from a Normal distribution with expectation μ and variance σ^2, with both μ and σ unknown. The variance, σ^2, can be estimated by the sample variance of each sample, denoted as S_X^2 and S_Y^2 .

(i) Identify which one of the following options gives the correct probability that S_X^2 exceeds $1.5\sigma^2$:

A $P(S_X^2 > 1.5\sigma^2) = 0.14$

B $P(S_X^2 > 1.5\sigma^2) = 0.18$

C $P(S_X^2 > 1.5\sigma^2) = 0.10$

D $P(S_X^2 > 1.5\sigma^2) = 0.21$.

(ii) Calculate the probability that S_Y^2 exceeds $1.5\sigma^2$.

Unit 3

PRACTICE QUESTION

Ans:

(i) option A

(ii) 0.09

29. CS1A APRIL 2023 Q7

Let X_1, Y_1, …, (X_n, Y_n) be pairs of random variables with each pair (X_i, Y_i) being independent of all other pairs. The distribution of X_i is $N(0, 1)$, for $i = 1, \ldots, n$. The conditional distribution of Y_i, given that X_i takes a particular value x_i, is $N(x_i\theta, 1)$, for $i = 1, \ldots, n$ where $\theta \in (-\infty, +\infty)$ is an unknown parameter.

(i) Identify which one of the following options gives the correct expression of the likelihood function:

A $$L(\theta) = \prod_{i=1}^{n} \exp\left[\frac{(y_i + x_i\theta)^2 - x_i^2}{2}\right]$$

B $$L(\theta) = \prod_{i=1}^{n} \frac{1}{2\pi} \exp\left[-\frac{(y_i - x_i\theta)^2 + x_i^2}{2}\right]$$

C $$L(\theta) = \prod_{i=1}^{n} \frac{\pi}{2} \exp\left[-\frac{(y_i - x_i\theta)^2 - x_i^2}{2}\right]$$

D $$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_i + x_i\theta)^2 + x_i^2}{2}\right].$$

(ii) Determine the maximum likelihood estimator $\hat{\theta}$ of θ. You do not need to check that your solution is a maximum of the likelihood function.

(iii) Determine the Cramer–Rao lower bound for $\hat{\theta}$

(iv) Write down the asymptotic distribution of $\hat{\theta}$.

Ans:

(i) Option B

(ii) $\hat{\theta}.=\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

(iii) CRLB = 1 /n

(iv) $\hat{\theta} \sim N(\theta, 1/n)$

Part 2

- **Theory of estimation 2**
- **Hypothesis testing**

1. CT3 April 2011 Question 9

Claims on a certain type of policy are such that the claim amounts are approximately normally distributed.

i. A sample of 101 such claim amounts (in £) yield a sample mean of £416 and sample standard deviation of £72. For this type of policy:

a) Obtain a 95% confidence interval for the mean of the claim amounts.

b) Obtain a 95% confidence interval for the standard deviation of the claim amounts.

The company makes various alterations to its policy conditions and thinks that these changes may result in a change in the mean, but not the standard deviation, of the claim amounts. It wants to take a random sample of claims in order to estimate the new mean amount with a 95% confidence interval equal to:

sample mean ±£10.

ii. Determine how large a sample must be taken, using the following as an estimate of the standard deviation:

a) The sample standard deviation from part (I).

b) The upper limit of the confidence interval for the standard deviation from part (i)(b).

iii. Comment briefly on your two answers in (ii)(a) and (ii)(b).

Ans:

(i) a. (402.0, 430.0)
b. (63.2, 83.6)

(ii) a. $n \geq 200$
b. $n \geq 269$

(iii)Assuming a larger value of s results in a larger standard error, so a larger sample size is required to achieve the same width of confidence interval.

2. CT3 October 2011 Question 9

In a recent study of attitudes to a proposed new piece of consumer legislation ("proposal X") independent random samples of 200 men and 200 women were asked to state simply whether they were "for" (in favour of), or "against", the proposal.

Unit 3

PRACTICE QUESTION

The resulting frequencies, as reported by the consultants who carried out the survey, are given in the following table:

	Men	*Women*
For	138	130
Against	62	70

i. Carry out a formal chi-squared test to investigate whether or not an association exists between gender and attitude to proposal X.

Note: in this and any later such tests in this question you should state the P-value of the data and your conclusion clearly.

At a subsequent meeting to discuss these and other results, the consultants revealed that they had in fact stratified the survey, sampling 100 men and 100 women in England and 100 men and 100 women in Wales. The resulting frequencies were as follows:

	England		*Wales*	
	Men	*Women*	*Men*	*Women*
For	82	66	56	64
Against	18	34	44	36

A chi-squared test to investigate whether or not an association exists between gender and attitude to proposal X in England gives $x2 = 6.653$, while an equivalent test for Wales gives $x2 = 1.333$.

ii. a) Find the P-value for each of the chi-squared tests mentioned above and state your conclusions regarding possible association between gender and attitude to proposal X in England and in Wales.

b) Discuss the results of the survey for England and Wales separately and together, quoting relevant percentages to support your comments.

iii. A different survey of 200 people conducted in each of England, Wales, and Scotland gave the following percentages in favour of another proposal:

	England	*Wales*	*Scotland*
% in favour of proposal	62%	53%	58%

A chi-squared test of association between country and attitude to the proposal gives $x^2 = 3.332$ on 2 degrees of freedom, with P-value 0.189.

Unit 3

PRACTICE QUESTION

Suppose a second survey of the same size is conducted in the three countries and results in the same percentages in favour of the proposal as in the first survey. The results of the two surveys are now combined, giving a survey based on the attitudes of 1,200 people.

a) State (or find) the results of a second chi-squared test for an association between country and attitude to the proposal, based on the overall survey of 1,200 people.

b) Comment briefly on the results.

Ans:

(i) $x^2 = 0.724$, p-value= 0.395
No evidence against H_0 – we conclude that no association exists between gender and attitude to proposal X

(ii) a. For England:
P-value = $P(\chi_1^2 > 6.653) = 0.010$
Evidence against H_0 – we reject it at the 1% level of testing and conclude that an association exists between gender and attitude to proposal X in England.
For Wales:
P-value = $P(\chi_1^2 > 1.333) = 0.248$
No evidence against H_0 – we conclude that there is no association between gender and attitude to proposal X in Wales.

b. England: there is evidence of an association – 82% of men and only 66% of women support proposal X – these proportions are significantly different.
Wales: there is no evidence of an association – 56% of men and 64% of women support proposal X – these proportions are not significantly different.
The effects are in different directions and cancel out to some extent when the data are combined: now there is no evidence of an association – overall 69% of men and 65% of women support proposal X – these proportions are not significantly different.
The combined data give a misleading message – they hide the effect of the factor "country" and fail to reveal that there is an association in England.

(iii) a. P-value = $P(\chi_1^2 > 6.664) = 0.0357$
Conclusion: reject "no association" at the 3.6% level of testing and conclude that an association does exist.
b. Comment: having more data with the same proportions provides strong enough evidence to justify claiming that an association exists.

Unit 3

PRACTICE QUESTION

3. CT3 April 2012 Question 12

Consider a random sample $X_1, \ldots, X_4$ of size k = 400 . Statistician A wants to use a x2 - test to test the hypothesis that the distribution of X_i is a binomial distribution with parameters n = 2 and unknown p based on the following observed frequencies of outcomes of X_i:

Possible realisation of X_i	0	1	2
Frequency	90	220	90

i. Estimate the parameter p using the method of moments.

ii. Test the hypothesis that X_i has a binomial distribution at the 0.05 significance level using the data in the above table and the estimate of p obtained in part (i).

Statistician B assumes that the data are from a binomial distribution and wants to test the hypothesis that the true parameter is p_0= 0.5.

iii. Explain whether there is any evidence against this hypothesis by using the estimate of p in part (i) and without performing any further calculations.

Statistician C wants to test the hypothesis that the random variables X_i have a binomial distribution with known parameters n= 2 and p = 0.5.

iv. Write down the null hypothesis and the alternative hypothesis for the test in this situation.

v. Carry out the test at the significance level of 0.05 stating your decision.

vi. Explain briefly the relationship between the test decisions in parts (ii), (iii) and (v), and in particular whether there is any contradiction.

Ans:

(i) $\hat{p}$ = 0.5

(ii) H_0 is rejected since the (1− α) -quantile (α = 0.05) of the x^2 -distribution with one degree of freedom is 3.841 < C = 4, where C is x^2 -distributed with 3−1−1 = 1 degree of freedom.

Unit 3

PRACTICE QUESTION

(iii) Since the estimated value is 0.5, any reasonable test will not reject that value, since the value 0.5 will always be in the acceptance region of the test. In other words, 0.5 will always be in any confidence interval around the estimate 0.5.
(iv) $H_0 : X_i \sim Bin(2, 0.5)$ VS $H_1 : X_i$ does not follow $Bin(2, 0.5)$
(v) H_0 is NOT rejected at a 5%-level since the $(1-\alpha)$ -quantile ($\alpha = 0.05$) of the x^2 - distribution with two degrees of freedom is 5.991 > 4.
(vi) The result in part (ii) states that a binomial distribution does not fit the data well and is rejected. However, in part (iii) we found that, under the assumption of a binomial distribution, $p0 = 0.5$ cannot be rejected. A specific binomial distribution with parameter $p = 0.5$ is not rejected in part (v) for the same data. The reason is that the additional degree of freedom in part (v) allows for a larger value of the test-statistic under the null.

4. CT3 October 2013 Question 6

A researcher obtains samples of 25 items from normally distributed measurements from each of two factories. The sample variances are 2.86 and 9.21 respectively.

i. Perform a test to determine if the true variances are the same.

ii. For each factory calculate central 95% confidence intervals for the true variances of the measurements.

iii. Comment on how your answers in parts (i) and (ii) relate to each other.

Ans:

(i) Reject H_0 at 1% significance level and conclude that the variances are not same.
(ii) Confidence interval 1 = (1.74, 5.54)
Confidence interval 2 = (5.61, 17.83)
(iii)Confidence intervals don't overlap i.e. agree with result in (i) that variances are different.

5. CT3 April 2014 Question 6

In an opinion poll, a sample of 100 people from a large town were asked which candidate they would vote for in a forthcoming national election with the following results:

Candidate	A	B	C
Supporters	32	47	21

i. Determine the approximate probability that candidate B will get more than 50% of the vote.

A second opinion poll of 150 people was conducted in a different town with the following results:

Candidate	A	B	C
Supporters	57	56	37

ii. Use an appropriate test to decide whether the two towns have significantly different voting intentions.

Ans:

(i) 0.274

(ii) Test statistic = 2.315, df = 2, p-value (χ^2_2) = between 0.30 and 0.32 (0.314 from interpolation.) we fail to reject H_0 that towns have the same voting intentions.

6. CT3 September 2014 Question 6

In a medical study conducted to test the suggestion that daily exercise has the effect of lowering blood pressure, a sample of eight patients with high blood pressure was selected. Their blood pressure was measured initially and then again, a month later after they had participated in an exercise programme.

The results are shown in the table below:

Patient	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before	155	152	146	153	146	160	139	148
After	145	147	123	137	141	142	140	138

i. Explain why a standard two-sample t-test would not be appropriate in this investigation

to test the suggestion that daily exercise has the effect of lowering blood pressure.

ii. Perform a suitable t-test for this medical study. You should clearly state the null and alternative hypotheses.

Ans:

(i) The two samples are from the same patients, so they are clearly not independent.

(ii) T-test stat = 3.855, t_7 (0.005) = 3.499 and t_7 (0.001) = 4.785. Therefore, we have strong evidence against H_0 (P-value < 0.5%), and conclude that daily exercise has the effect of lowering blood pressure.

7. CT3 April 2015 Question 9

An insurance company has calculated premiums assuming that the average claim size per claim for a certain class of insurance policies does not exceed £20,000 per annum. An actuary analyses 25 such claims that have been randomly selected. She finds that the average claim size in the sample is £21,000 and the sample standard deviation is £2,500. Assume that the size of a single claim is normally distributed with unknown expectation α and variance σ^2.

(i) Calculate a 95% confidence interval for α based on the sample of 25 claims.

(ii) Perform a test for the null hypothesis that the expected claim size is not greater than £20,000 at a 5% significance level.

(iii) Discuss whether your answers to parts (i) and (ii) are consistent.

(iv) Calculate the largest expected claim size, α_0, for which the hypothesis $\alpha \leq \alpha_0$ can be rejected at a 5% significance level based on the sample of 25 claims.

The insurer is also concerned about the number of claims made each year. It is found that the average number of claims per policy was 0.5 during the year 2011. When the analysis was repeated in 2012 it was found that the average number of claims per policy had increased to 0.6. These averages were calculated on the basis of random samples of 100 policies in each of the two years.

Assume that the number of claims per policy per year has a Poisson distribution with unknown expectation λ and is independent from the number of claims in any other year or for any other policy.

Unit 3

PRACTICE QUESTION

v. Perform a test at 5% significance level for the null hypothesis that λ= 0.6 during the year 2011.

vi. Perform a test to decide whether the average number of claims has increased from 2011 to 2012.

Ans:

(i) (19.968, 22.032)

(ii) Test Stat = 2, $t_{0.05,24}$ = 1.711

We reject the null hypothesis.

(iii) The confidence interval in part (i) corresponds to a two-sided test. We found in part (i) that 20 is contained in the confidence interval, and we can therefore not reject the null hypothesis H_0: α= 20 at a 5% significance level. However, the one-sided test rejects H_0: $\alpha \leq$ 20 since only positive differences $\underline{X} - \alpha_0$ are considered. Answers are consistent.

(iv) α_0 = 20.1445

(v) z = -1.29, $z_{0.025}$ = -+1.96

The null hypothesis H_0: λ= 0.6 cannot be rejected for the year 2011.

(vi) Test stat = 0.9535, $z_{0.5}$ = 1.64

The null hypothesis H_0: $\lambda_{2012} \leq \lambda_{2011}$ cannot be rejected at the 5% level. Therefore, we do not have empirical evidence to suggest that the alternative $\lambda_{2012} > \lambda_{2011}$ is true.

8. CT3 October 2015 Question 9

A survey team is using satellite technology to measure the height of a mountain. This is an established technology and the variability of measurements is known. On each satellite pass over a mountain, they get a measurement that they know lies within ±5m of the true height with a 95% probability. The survey height is given by the mean of the measurements. They assume that all measurements are independent and follow a normal distribution with mean equal to the true height.

i. a) Show that the standard deviation of a single measurement is 2.551m.

b) Determine how many satellites passes over a mountain are required to have a 95% confidence interval for the true height with width less than 1m.

In a mountain range there are two summits which appear to have a similar height.

The team manages to get 20 measurements for each summit and finds there is a difference of 1.6m between the mean survey height of the two summits.

ii. Perform a statistical test of the null hypothesis that the summits' true heights are the same, against the alternative that they are different.

At the same time the team is testing a new system on these two summits. They again get 20 measurements on each summit with an estimated standard deviation on the first summit of 2.5m and on the second of 2.6m. This system also measures the difference in survey heights between the two peaks to be 1.6m.

iii. Perform a statistical test of the same hypotheses as in part (ii) when heights are measured by the new system, justifying any assumptions you make.

iv. Comment on your answers to parts (ii) and (iii).

Ans:

(i) a. $\sigma = 2.551$
b. $n > 100$

(ii) p-value = 0.048. Therefore, reject H_0 at the 5% significance level.

(iii) Test stat = 1.984, $t_{38,\ 0.025} = 2.024$
So do not reject H_0 : no difference in means at 5% significance level.

(iv) Both systems gave the same estimate of difference and almost the same standard deviation, with the second being lower. However, the tests gave different results. We did not reject that there was no difference for the test in part (iii) as there was greater uncertainty since we did not know the standard deviation beforehand.

9. CT3 April 2016 Question 10

Consider a large portfolio of insurance policies and denote the claim size (in £) per claim by X. A random sample of policies with a total of 20 claims is taken from this portfolio and the claims made for these policies are reported in the following table:

Claim i	1	2	3	4	5	6	7	8	9	10
Claim size x_i	130	164	170	173	173	175	177	183	183	184

Claim i	11	12	13	14	15	16	17	18	19	20
Claim size x_i	185	186	197	202	208	213	215	229	233	272

For these data: $\sum x_i = 3{,}852$ and $\sum x^2 = 759{,}348$.

i. Calculate the mean, the median and the standard deviation of the claim size per claim in this sample.

Unit 3

PRACTICE QUESTION

ii. Determine a 95% confidence interval for the expected value E[X] based on the above random sample, stating any assumptions you make.

iii. Determine a 95% confidence interval for the standard deviation of X based on the above random sample.

iv. Explain briefly why the confidence interval in part (iii) is not symmetric around the estimated value of the standard deviation.

An actuary assumes that the number of claims from each policy has a Poisson distribution with an unknown parameter λ. In a new sample of 50 policies the actuary has observed a total of 80 claims yielding an estimated value of $\hat{\lambda}=1.6$ for the parameter λ.

v. Determine a 95% confidence interval for λ using a normal approximation.

vi. Determine the smallest required sample size n for which a 95% confidence interval for λ has a width of less than 0.5. You should use the same normal approximation as in part (v), and assume that the estimated value of λ will not change.

Now assume that the true value of λ is 1.6 and the values calculated in part (i) are the true values. Also assume that all claims in the portfolio are independent and the claim sizes are independent of the number of claims.

vii. Determine the expected value and the standard deviation of the total amount of all claims from a portfolio of 5,000 insurance policies.

Ans:

(i) Mean=$\underline{x}$ = 192.6
Median = 184.5
SD: s = 30.31
(ii) [178.4, 206.8]
(iii) [23.05, 44.27]
(iv) The x^2 distribution is not symmetric.
(v) [1.25,1 .95]
(vi) $n \geq 99$
(vii) $E(S)$ = 1,540,800, SD(S) = 17,438.68

Unit 3

PRACTICE QUESTION

10. CT3 October 2016 Question 4

Consider two portfolios, A and B, of insurance policies and denote by X_A the number of claims received in portfolio A and by X_B the number of claims received in portfolio B during a calendar year. The observed numbers of claims received during the last calendar year are 134 for portfolio A and 91 for portfolio B. X_A and X_B are assumed to be independent and to have Poisson distributions with unknown parameters β_A and β_B.

Determine an approximate 99% confidence interval for the difference $\beta_A - \beta_B$. You may use an appropriate normal distribution.

Ans:

[4.4, 81.6]

11. CT3 April 2017 Question 6

We consider the impact that different types of cars have on the amount spent on fuel per month. Three different types of cars are considered: small, medium and large. For each type of car a group of 15 drivers are asked about the amount of money (in £) spent on fuel per month. The results are summarised in the following table

Type of car	*Small*	*Medium*	*Large*
Sample mean	70	75	83
Sample standard deviation	16	19	16

For example, the 15 drivers of medium sized cars spent on average £75 per month with a sample standard deviation of £19.

i. Perform a one-way analysis of variance to test the hypothesis that the type of car has no impact on the monthly amount spent on fuel. For some further investigation, only the difference between small and large cars is considered.

ii. Determine a 95% confidence interval for the difference between the average amount spent on fuel for small cars and large cars, stating any assumptions you make.

iii. Test the null hypothesis that the average fuel costs for small and large cars are the same at a 5% significance level against the alternative that the fuel costs for small and large cars are different.

Unit 3

PRACTICE QUESTION

Ans:

(i) Test stat = 2.216, $F_{2,42,0.05}$= 3.22
the null hypothesis is not rejected. We conclude that there is no evidence that the type of car has an impact on the monthly amount of money spent on petrol.

(ii) (1.035, 24.965)

(iii) we reject the null hypothesis of equal amounts spent on petrol for large cars and small cars since '0' is not present in the CI above.

12. CT3 April 2017 Question 9

A statistician is examining the survey methodology of a country's national statistics department. It conducts much of its data collection by telephoning individuals selected at random and asking them questions.

i. Comment on whether this methodology will give a random sample.

ii. Comment on whether this methodology will give a representative random sample of the population.

Ans:

(i) A random sample should be independent and identically distributed. As people are chosen at random the methodology should give a random sample.

(ii) While the sample chosen will be independent, they will not necessarily be representative of the population as a whole. In many places phone ownership may be restricted by economic, cultural or geographic limitations so some parts of the population may be excluded.

13. CT3 September 2017 Question 5

In an election between two candidates A and B in a large district, a sample poll of 100 voters chosen at random, indicated that 55% were in favour of candidate A.

i. Calculate a 95% confidence interval for the proportion of all voters in favour of candidate A based on the above sample.

A candidate is elected if they win more than 50% of the votes. We want a test in which the alternative hypothesis is that support for candidate A is such that she will win the election.

ii. a) Write down the hypotheses for this test in terms of a suitable parameter.

b) Explain whether or not the confidence interval in part (i) can be used to test the hypothesis in part (ii)(a) at the 5% level of significance.

It has been reported in the news that a new poll estimates support for candidate A at 52%, with a margin of error of no more than ±2% with confidence 95%.

iii. Determine the minimum size of the sample of voters that was taken in this new poll.

Ans:

(i) (0.4525, 0.6475)

(ii) a. If p is proportion voting for candidate A, we want
H_0: $p = 0.5$ (or $p \leq 0.5$) v. H_1: $p > 0.5$
b. CI in part (i) is 2-sided, so cannot be used here.

(iii) The sample size must be at least 2398.

14. CT3 April 2018 Question 10

A large pension scheme regularly investigates the lifestyle of its pensioners using surveys. In successive surveys it draws a random sample from all pensioners in the scheme and it obtains the following data on whether the pensioners smoke.

Survey 1: Of 124 pensioners surveyed, 36 were classed as smokers.
Survey 2: Of 136 pensioners surveyed, 25 were classed as smokers.

An actuary wants to investigate, using statistical testing at a 5% significance level, whether there have been significant changes in the proportion of pensioners, p, who smoke in the entire pension scheme.

i. Perform a statistical test, without using a contingency table, to determine if the proportion p has changed from the first survey to the second.

When a third survey is performed it is found that 26 out of the 141 surveyed pensioners, are smokers.

ii. Perform a statistical test using a contingency table to determine if the proportion p is different among the three surveys.

iii. a) Calculate the proportion of smokers in the third survey.
b) Comment on your answers to parts (i) and (ii).

Unit 3

PRACTICE QUESTION

Ans:

(i) test stat =2.014, As is $Z_{0.3}$ less than the test statistic we reject H_0: $p_1=p_2$ at a 5% significance level.

(ii) Test stat = 5.69, df = 2,
The 95% point of χ^2_2= 5.991. As test statistic is lower, do not reject that the proportion of smokers is equal.

(iii) a.$\widehat{p_3}$ = 0.184
b. In the first case the test rejected that the proportions were the same, but in the second it did not reject that they were, as the proportion in the third survey is almost identical to that in the second.

15. CS1A April 2019 Q3

The number of claims on a certain type of policy follows a Poisson distribution with claim rate l per year. For a group of 200 independent policies of this type, the total number of claims during the last calendar year was 82.

Determine an approximate 95% confidence interval for the true annual claim rate for this type of policy based on last year's claims.

Ans:

(0.321, 0.499)

16. CS1A September 2019 Q9

An actuary wants to model a particular type of claim size and has been advised to use a Gamma distribution with probability distribution function:

$$f(x,\alpha,\theta)=\frac{x^{\alpha-1}}{\Gamma(\alpha)\theta^{\alpha}}e^{-\frac{x}{\theta}},\quad 0<x<\infty,\quad \alpha>0,\quad \theta>0.$$

i. Show, using moment generating functions, that:

(a) $E(X)=\alpha\theta$
(b) $E(X^2)=\alpha(\alpha+1)\theta^2$
(c) $E(X^3)=\alpha(\alpha+1)(\alpha+2)\theta^3$.

The shape parameter alpha is assumed to be α = 4.

ii. (a) Determine the variance of the claim size distribution in terms of θ.

(b) Calculate the coefficient of skewness of the claim size distribution, which is defined as:

$$\frac{E\left[\left(X-E(X)\right)^3\right]}{\left\{E\left[\left(X-E(X)\right)^2\right]\right\}^{1.5}}$$

Let $X_1, X_2, \ldots, X_n$ be a random sample of n claim sizes for such claims.

iii. Show that the maximum likelihood estimator (MLE) of θ is given by:

$$\hat{\theta} = \frac{\bar{X}}{4}$$

iv. Show that $\hat{\theta}$ is an unbiased estimator of θ.

A sample of n = 100 claim sizes yields

$$\sum x_i = 796.2 \text{ and } \sum x_i^2 = 8{,}189.4.$$

v. Calculate the MLE of θ.

vi. (a) Calculate the sample variance.

(b) Compare the result in part (vi)(a) with the variance of the distribution evaluated at $\hat{\theta}$.

The sample coefficient of skewness is given as 1.12.

vii. Comment on its comparison with the coefficient of skewness of the distribution, calculated in part (ii)(b).

viii Calculate an appropriate 95% confidence interval for q by using an approximate 95% confidence interval for the mean of the distribution of the claim size.

ix. (a) Determine the variance of the distribution of q at both lower and upper limits of the confidence interval calculated in part (viii).

(b) Comment on the result in part (ix)(a) with reference to your answer in part (vi)(a) above.

Unit 3

PRACTICE QUESTION

Ans:

(i) –

(ii) a. $\sigma^2 = 4\theta^2$

b. Coefficient of skewness = 1

(iii) -

(iv) -

(v) 1.9905

a.$s^2 = 18.69$

b. s^2 is a bit larger than the variance at $\hat{\theta}$.

(vii) Sample coefficient 1.12 is close to the distribution value 1.

17. CS1A April 2021 Q7

A telecommunications company has performed a small empirical study comparing phone usage in rural and urban areas, collecting data from a total of 35 people who use their phones independently. The average number of hours that each person spent using their phone during a week is denoted by Y.

In the following table, $\underline{Y}$, denotes the sample mean of Y in rural and urban areas, and S_Y denotes the sample standard deviations; that is

$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

	Sample size n	$\overline{Y}$	S_Y
Rural areas	15	3.7	2.1
Urban areas	20	4.4	1.9

A statistical test is to be performed, at the 5% significance level, to determine whether the null hypothesis that mean phone usage in rural areas is the same as mean phone usage in urban areas, i.e. for:

H_0: phone usage is equal versus

H_1: phone usage is not equal.

i. State a suitable distribution for the test statistic with its parameter(s).

ii Justify any assumption(s) required to perform this test.

iii. Identify which one of the following options gives the correct value of the test statistic for this test:

A –1.031

B –0.519

C –3.019

D –1.455.

iv. Write down the conclusion of the test including the relevant critical value(s) from the Actuarial Formulae and Tables.

v. Determine a 95% confidence interval for the mean phone usage (hours per week) for rural areas, stating any assumption(s) you make.

Ans:

(i) t distribution would be suitable, with 33 df.

(ii) Assumed that the variances (rural and urban) are equal
Equal variances seem to be justified given the S_y values for rural and urban areas are similar given the small sample sizes
Assumption of Normality

(iii) Option A

(iv) we conclude that the null hypothesis of equal phone usage being equal cannot be rejected

(v) [2.537,4.863]

18. CS1A September 2021 Q5

The probability that a claim is made on a car insurance policy in a particular year is 0.06. The policies are assumed to be independent among them. 500 of these policies are selected at random.

i. Calculate the probability that no more than 40 of these policies will result in a claim during the year, stating any approximations you make.

Past data from the insurer indicate that the standard deviation of claim amounts is £75. The insurer wishes to construct a 95% confidence interval for the mean claim amount, with an interval width of £10.

Unit 3

PRACTICE QUESTION

ii. Calculate the sample size needed to achieve this level of accuracy for a 95% confidence interval.

Ans:

(i) 0.97599

(ii) sample size of 865

19. CS1A September 2021 Q10

Total yearly aggregate claims in a particular company are modelled as a random variable X, where X is assumed to follow a Normal distribution with unknown mean μ and variance $\sigma^2 = 12{,}000^2$.

Aggregate claims from the last 5 years are as follows:

146,000 142,000 153,000 127,000 132,000

An analyst wishes to estimate the unknown parameter μ.

i. Identify which one of the following gives the correct expression of the derivative of the log-likelihood function:

A $\frac{dl(\mu)}{d\mu} = -\sum_{i=1}^{n}(x_i - \mu)$

B $\frac{dl(\mu)}{d\mu} = \sum_{i=1}^{n}(x_i - \mu)$

C $\frac{dl(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$

D $\frac{dl(\mu)}{d\mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$

ii. Calculate the maximum likelihood estimate for μ, using your answer to part (i).

iii. Calculate a 95% confidence interval for μ.

Ans:

(i) Option C

(ii) $\hat{\mu}$=140,000

(iii) (129,481.54,150,518.46)

Unit 3

PRACTICE QUESTION

20. CS1A September 2022 Q8

Following a recent Climate Action Plan (CAP) conference, a particular model was agreed for measuring global earth tremors. A series of n positive measurements are to be taken, which are assumed to be independent observations of a random variable that are Uniformly distributed on $(0, \upsilon)$, where $\upsilon > 0$.

A Climate Risk Actuary adopts the model agreed by the CAP conference. The Actuary knows only that the number, R, of the measurements that are less than 1 is r with the remaining $n - r$ being greater than 1

(i) a. Show that the probability for a single measurement to be less than 1 is $1/\upsilon$.

b. Show that the maximum likelihood estimate of υ is $\hat{\upsilon} = n/r$.

c. Identify which one of the following expressions gives the Cramer–Rao lower bound for estimating υ .

A $\dfrac{\upsilon(1-\upsilon)^2}{n}$

B $\dfrac{\upsilon^2(1-\upsilon)^2}{n}$

C $\dfrac{(\upsilon-1)}{n}$

D $\dfrac{\upsilon^2(\upsilon-1)}{n}$

d. Write down the asymptotic distribution of $\hat{\upsilon}$, using your answer from part (i)(c).

In a random sample of 500 measurements, exactly 75 measurements are less than 1.

(ii) a. Calculate an estimate for the standard error of $\hat{\upsilon}$.

b. Determine an approximate two-sided 99% confidence interval for υ.

c. Perform a test for the following hypotheses, using your asymptotic distribution of $\hat{\upsilon}$ from part (i)(d):

H_0: $\upsilon = 10$ vs H_A: $\upsilon < 10$

Ans:

(i) a.-

b.-

c. option D

(ii) a. $S.E.(\hat{\upsilon}) = 0.70972$

Unit 3

PRACTICE QUESTION

b.(4.83857 ,8.49477)

c. test stat = −2.48452, Critical value at 1% is -2.32635

Therefore reject H_0 at the 1% level (and conclude $v<10$).

21. CS1A April 2023 Q9

An insurer models the number of cars owned by a single policyholder as a discrete random variable with the following distribution, where p is an unknown parameter:

Number of cars	0	1	2	3	More than 3
Probability	1/2p	p	1/4p	1/4p	1-2p

In an empirical study, the number nk of policyholders owning k cars is recorded, for k = 0, …, 3, and n_4 is the number of policyholders with more than three cars.

(i) Identify which **one** of the following functions is the log likelihood function for p using the recorded numbers n_0, …, n_4, where C is a constant independent

A $(n_0+n_1+n_2+n_3)\frac{1}{4}\log p+n_4\log(1-2p)+C$

B $(n_0+n_1+n_2+n_3)\log\frac{1}{4}\log p+n_4\log(1-2p)+C$

C $\log(n_0+n_1+n_2+n_3)\log p+\log n_4\log(1-2p)+C$

D $(n_0+n_1+n_2+n_3)\log p+n_4\log(1-2p)+C.$

(ii) Derive the maximum likelihood estimator for the parameter p. You do not need to check that your solution is a maximum of the likelihood function.

In a larger study, the insurer has to restrict the information held for each policyholder due to data protection. The insurer records only the number of policyholders with no car, m_0, and the number of policyholders with at least one car, m_1

(iii) Show that the log likelihood function for the parameter p based on the observations m_0 and m_1, and the above distribution for the number of cars is given by:

$l(p)=m_0\log(\frac{1}{2}p)+m_1\log(\frac{2-p}{2})$

(iv) Identify which **one** of the following estimators is the maximum likelihood estimator for the parameter p:

Unit 3

PRACTICE QUESTION

A $\hat{p} = \frac{m_0}{2m_0 + m_1}$

B $\hat{p} = \frac{2m_0}{2m_0 + m_1}$

C $\hat{p} = \frac{2m_0}{m_0 + m_1}$

D $\hat{p} = \frac{m_0}{m_0 + m_1}$.

The following data have been observed:
$n_0 = 50$, $n_1 = 37$, $n_2 = 17$, $n_3 = 16$, $n_4 = 10$

(v) Estimate the value of p using the estimator derived in part (ii).

(vi) Estimate the value of p using the estimator found in part (iv).

Ans:

(i) Option D
(ii) $\hat{p}=\frac{n-n_4}{2n}$
(iii) –
(iv) Option C
(v) $\hat{p}$= 0.4615
(vi) $\hat{p}$= 0.769

22. CS1A April 2024 Q9

A small vehicle repair shop, Repair Shop 1, has recently merged with another nearby vehicle repair shop, Repair Shop 2, and the owner wishes to analyse the number of customers that visit each shop.
The owner has recorded the following data over the past 30 days:

$\sum_{i=1}^{30} repair\ 1=781$ $\sum_{i=1}^{30} repair\ 2=707$, $n_1 = n_2 = 30$

(i) Calculate the sample means of the number of daily customers for each repair Shop.

The owner believes that the number of customers arriving at Repair Shop 1 and

Repair Shop 2 per day follow separate Poisson distributions with unknown parameters k_1 and k_2, respectively.

(ii) Calculate a 95% confidence interval for the difference between the parameters k_1 and k_2, stating any assumptions you make.

(iii) Test at the 5% significance level whether k_1 and k_2 are equal.

The owner later believes it may instead be more appropriate to consider the recorded data as being paired data.

(iv) Explain why it may be more appropriate to treat the data as paired.

The sample standard deviation of the paired differences has been recorded as 6.55.

(v) Test at the 5% significance level whether k_1 and k_2 are equal, assuming the data is paired, stating any additional assumptions you make.

(vi) Comment on your answers to parts (iii) and (v).

Ans:

(i) $\underline{X_1}$=26.03, $\underline{X_2}$= 23.57

(ii) (-0.06,4.98)

(iii) there is insufficient evidence at the 5% significance level to reject the hypothesis that k_1 and k_2

(iv) The samples are recorded for each shop on the same day and may not be independent, in which case the data would be paired. This could happen because the shops are nearby so would attract similar customers. A single customer may visit both shops on the same day for quotes.
External factors may also affect both shops' customer numbers in the same way, such as: the weather, roadworks or closures limiting access to the shops

(v) Test stat = 2.06, Critical value = 2.045, we have sufficient evidence to reject H_0 at the 5% level, and conclude that $k_1 \neq k_2$.

(vi) In part (iii) we treated the data as unpaired and failed to reject the hypothesis that k_1 and k_2 are equal.
In part (v) we treated the data as paired and found sufficient evidence to reject the same hypothesis.

Unit 3

PRACTICE QUESTION

If the samples are not truly independent, then treating them as such would produce invalid results. However, treating them as paired always produces valid results, although it would make inefficient use of the data.
However, the evidence in part (v) was not very strong, as the test statistic was only slightly above the critical value.
In both cases we used a normal approximation to the Poisson distribution, which would have introduced some error in the testing. Using the exact distributions may have produced different conclusions.