Imperial College
London

[C245 2014]

BEng and MEng EXAMINATIONS 2014

PART II : STATISTICS (COMPUTING C245)

Date : Friday 16th May 2014 10 - 12

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

Answer ALL questions

Statistical data sheets are provided

[*Before starting, please make sure that the paper is complete; there should be a total of FOUR questions. Ask the invigilator for a replacement if your copy is faulty.*]

1. *Choose one answer for each part. Partial credit may be awarded for working if an incorrect answer is selected. There is no negative marking.*

   (i) For the data below, calculate the sample median, $x_{(\{n+1\}/2)}$; mean, $\bar{x} = \sum_{i=1}^{n} x_i/n$; interquartile range (IQR), $x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}$; and standard deviation, $s_n = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}$.

   0 2 3 3 4 7 9.

   Which one of the following statements about these sample statistics is true?

   (a) The median is less than the mean and the IQR is less than the standard deviation;
   (b) The mean is less than the median and the IQR is less than the standard deviation;
   (c) The median is less than the mean and the standard deviation is less than the IQR;
   (d) The mean is less than the median and the standard deviation is less than the IQR;
   (e) The median is equal to the mean and the IQR is equal to the standard deviation.

   (ii) For two events A and B, we have $\mathrm{P}(A) = 0.2$, $\mathrm{P}(B) = 0.3$, and $\mathrm{P}(\overline{A}|\overline{B}) = 0.8$. What is $\mathrm{P}(\overline{B}|A)$ equal to?

   (a) 0.3; (b) 0.4; (c) 0.5; (d) 0.6; (e) 0.7; (f) 0.8.

   (iii) A die is rolled three times. What is the probability that we observe at least one 3?

   (a) 0.5; (b) 0.58; (c) 0.42; (d) 0.33; (e) 0.67.

   (iv) A bag contains 5 red balls and 1 black ball. Balls are randomly drawn from the bag until the black ball is found, and the number of draws which were required is noted. How much greater is the expected number of draws required if any red ball drawn from the bag is put back in the bag rather than set aside?

   (a) 0; (b) 1; (c) 2; (d) 2.5; (e) 3.

   (v) Suppose $X_1, X_2, \ldots, X_n$ are n independent random variables which all follow the same distribution P_X which has mean μ and variance σ^2. Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of the n samples, and define $T_n = S_n/\sqrt{n}$. By the Central Limit Theorem, which of the following is an approximate distribution for T_n?

   (a) $\mathrm{N}(\sqrt{n}\mu,\sigma^2)$; (b) $\mathrm{N}(\mu,n\sigma^2)$; (c) $\mathrm{N}(\frac{\mu}{\sqrt{n}},\sigma^2)$; (d) $\mathrm{N}(\sqrt{n}\mu,\frac{\sigma^2}{\sqrt{n}})$; (e) $\mathrm{N}(\mu,\frac{\sigma^2}{n})$.

2. If X is a discrete random variable on the positive integers $\{1, 2, 3, \ldots\}$, it can be shown that

$$\mathrm{E}(X) = \sum_{x=1}^{\infty} \mathrm{P}(X \geq x).$$

The Geometric distribution describes such a random variable. Let $X \sim \text{Geometric}(p)$, so that X has probability mass function

$$p(x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

(i) Find $\mathrm{P}(X \geq x)$ and hence show that $\mathrm{E}(X) = \dfrac{1}{p}$. [*Hint: Recall that the infinite sum of a geometric progression with first term a and common ratio r is $a/(1-r)$.*]

(ii) Find the number of rolls of a die we would expect to require until we finally observe a 6.

(iii) Suppose that after five rolls of the die we have not yet observed a 6. If we decide to roll the die another five times, what is the probability that we will have observed a 6?

(iv) If X is a discrete random variable on the positive integers, prove the formula $\mathrm{E}(X) = \sum_{x=1}^{\infty} \mathrm{P}(X \geq x)$. [*Hint: Express* $\mathrm{P}(X \geq x)$ *as a summation, and then change the order of the two summations.*]

3. The number of emails sent to an IT helpdesk in a small department between the hours of 10 a.m. and 11 a.m. was recorded each day for a period of 100 working days. The data are summarised in the following table.

Number of emails	0	1	2	3	≥4
Frequency	14	39	29	11	7

Before collecting the data, it had been believed that the number of emails in an hour was well described by a Poisson distribution, and that two emails per hour were to be expected on average.

Recall that a Poisson random variable with mean parameter λ has probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots.$$

(i) For a sample $X_1, \ldots, X_n$ from the Poisson distribution, show that the sample mean is the maximum likelihood estimator (MLE) for λ.

(ii) Calculate the value of the MLE for λ for the data above. When calculating the sample mean of the data, treat the category "≥4" as taking fixed value 4.5.

(iii) Using the MLE for λ, perform a goodness-of-fit test at the 95% significance level to assess whether the number of emails per hour does indeed follow a Poisson distribution.

(iv) Suppose that instead of estimating λ we assume 2 emails per hour are received on average. Repeat the Poisson goodness-of-fit test under this assumption, and comment on the comparison.

4. Let $X_1, \ldots, X_n$ be a random sample of independent observations from the continuous uniform distribution U(0,b), where the parameter b is unknown. The probability density function for each X_i is, thus,

$$f(x) = \begin{cases} \frac{1}{b}, & 0 \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

   (i) Find the cumulative distribution function, $F(x) = \mathrm{P}(X_i < x)$ for $x \in [0, b]$.

   (ii) Let $X_{(n)} = \max_i X_i$ be the largest value in the sample. Find the cumulative distribution function $\mathrm{P}(X_{(n)} < x)$ for $X_{(n)}$. Hence derive the probability density function of $X_{(n)}$. [*Hint: Notice that* $X_{(n)} < x \iff \forall i, X_i < x$].

   (iii) Calculate the expected value $\mathrm{E}(X_{(n)})$ as a function of b and n.

   (iv) Find the bias of the maximum sample value as an estimator for the upper limit of the population b. From this result, suggest a revised estimator of b using $X_{(n)}$ which is unbiased. Remember in constructing your estimator that b is unknown.

END OF PAPER

	EXAMINATION QUESTIONS/SOLUTIONS 2013-2014	Course **Comp245**
Question 1.		Marks & seen/unseen
Parts		

(i) $x_{(\{n+1\}/2)}=3$, $\bar{x}=4$, IQR$=x_{(6)}-x_{(2)}=5$, $s_n=2\sqrt{2}$. seen sim.
Answer is (c).

(ii) Using Bayes Theorem,

$$\mathrm{P}(\overline{B}|A)=\frac{\mathrm{P}(A\cap\overline{B})}{\mathrm{P}(A)}=\frac{\mathrm{P}(A|\overline{B})\mathrm{P}(\overline{B})}{\mathrm{P}(A)}=\frac{(1-\mathrm{P}(\overline{A}|\overline{B}))(1-\mathrm{P}(B))}{\mathrm{P}(A)}$$
$$=\frac{(1-0.8)(1-0.3)}{0.2}=0.7.$$

Answer is (e). seen sim.

(iii) The probability of not observing a 3 is $(5/6)^3\approx 0.58$, so the probability of at least one 3 is $1-0.58=0.42$. seen sim.
Answer is (c).

(iv) Let X be the number of draws until the black ball is found. Without replacement, X is equally likely to be any number between 1 and 6, so E$(X)=3.5$. With replacement, $X\sim$ Geometric$(1/6)$, so E$(X)=6$. unseen
Answer is (d).

(v) By the Central Limit Theorem, an approximate distribution for S_n is N$(n\mu,n\sigma^2)$. The distribution of $T_n=S_n/\sqrt{n}$ is then, approximately, N$(\sqrt{n}\mu,\sigma^2)$. unseen
Answer is (a).

4 marks each

Question 2. | Marks & seen/unseen

Parts

(i) $\mathrm{P}(X \geq x) = \sum_{k=x}^{\infty} p(k) = p \sum_{k=x}^{\infty} (1-p)^{k-1} = p \times \frac{(1-p)^{x-1}}{1-(1-p)} = (1-p)^{x-1}.$ — seen, 4 marks

We then have $\mathrm{E}(X) = \sum_{x=1}^{\infty} \mathrm{P}(X \geq x) = \sum_{x=1}^{\infty} (1-p)^{x-1} = \frac{1}{1-(1-p)} = \frac{1}{p}.$ — unseen, 4 marks

(ii) If X is the number of rolls required, then $X \sim \text{Geom}\left(\frac{1}{6}\right)$, so $\mathrm{E}(X) = \frac{1}{\frac{1}{6}} = 6.$ — seen sim., 3 marks

(iii) This is the same as the probability of failing to roll a 6 in five attempts; the first five unsuccessful rolls have no bearing on subsequent attempts (the Geometric is memoryless).

The probability is $\mathrm{P}(X \leq 10 | X > 5) = 1 - \mathrm{P}(X \geq 11 | X > 5) = 1 - \left(1 - \frac{1}{6}\right)^5 = 0.598.$ — seen sim., 4 marks

(iv) $\sum_{x=1}^{\infty} \mathrm{P}(X \geq x) = \sum_{x=1}^{\infty} \sum_{k=x}^{\infty} p(k) = \sum_{k=1}^{\infty} \sum_{x=1}^{k} p(k) = \sum_{k=1}^{\infty} k p(k) = \mathrm{E}(X).$ — unseen, 5 marks

Question 3. | Marks & seen/unseen

Parts

(i)

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \implies \ell(\lambda) = (\sum_{i=1}^{n} x_i)\log(\lambda) - n\lambda - \log(\prod_{i=1}^{n} x_i!)$$

$$\implies \frac{d}{d\lambda}\ell(\lambda) = \frac{\sum_{i=1}^n x_i}{\lambda} - n \implies \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

This is a maximum since

$$\implies \frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{\sum_{i=1}^n x_i}{\lambda^2}$$

is negative for all $\lambda > 0$.

seen

7 marks

(ii) $\hat{\lambda} = \bar{x}$ =1.615.

3 marks

(iii) Assuming a Poisson(1.61) distribution, we would have the following expected counts.

Number of emails (i)	0	1	2	3	≥4
Observed Frequency (O_i)	14	39	29	11	7
Expected Frequency (E_i)	19.89	32.12	25.94	13.96	8.09

These give a chi-square statistic of

$$X^2 = \sum_{i=0}^{4} \frac{(O_i - E_i)^2}{E_i} = 4.35$$

which is less than $\chi^2_{3,0.95}$=7.81. So insufficient evidence to reject the null hypothesis that the email counts follow a Poisson distribution.

seen sim.

5 marks

(iv) Now assuming a Poisson(2) distribution, we would have the following expected counts.

Number of emails (i)	0	1	2	3	≥4
Expected Frequency (E_i)	13.53	27.07	27.07	18.04	14.29

Here,

$$X^2 = \sum_{i=0}^{4} \frac{(O_i - E_i)^2}{E_i} = 11.88$$

which is greater than $\chi^2_{4,0.95}$=9.49, so we reject the null hypothesis that the email counts follow a Poisson(2) distribution. This suggests the mean number of emails per hour is not close to 2.

seen sim.

5 marks

Question 4. | Marks & seen/unseen

Parts

(i) $X_i \sim \mathrm{U}(0,b)$ has cumulative distribution function — seen, 2 marks

$$F_{X_i}(x) = \begin{cases} 0, & x \le 0 \\ \frac{x}{b}, & 0 < x < b \\ 1, & x \ge b. \end{cases}$$

(ii) By independence of the $\{X_i\}$, — unseen, 3 marks

$$F_{X_{(n)}}(x) = \mathrm{P}(X_{(n)} < x) = \prod_{i=1}^{n} \mathrm{P}(X_i < x) = \left(\frac{x}{b}\right)^n, \quad 0 \le x \le b.$$

The density is the first derivative of this distribution function, giving — seen, 4 marks

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = \frac{d}{dx}\left(\frac{x}{b}\right)^n = \frac{n}{b}\left(\frac{x}{b}\right)^{n-1}, \quad 0 \le x \le b.$$

(iii) — seen, 5 marks

$$\mathrm{E}(X_{(n)}|b) = \int_{-\infty}^{\infty} x\, f_{X_{(n)}}(x)dx = n\int_0^b \left(\frac{x}{b}\right)^n dx = \frac{n}{n+1}\frac{x^{n+1}}{b^n}\Big|_0^b = \left(\frac{n}{n+1}\right)b.$$

(iv) As an estimator of b, — unseen, 3 marks

$$\mathrm{Bias}(X_{(n)}) = \mathrm{E}(X_{(n)}|b) - b = -\frac{b}{n+1}.$$

Since $\mathrm{E}(X_{(n)}|b) = \left(\frac{n}{n+1}\right)b$, consider a revised estimator $T = \left(\frac{n+1}{n}\right)X_{(n)}$. — unseen, 3 marks
Then clearly $\mathrm{E}(T|b) = b$, so T is now an unbiased estimator for b.