BIRKBECK COLLEGE
(University of London)

BSc Examination
School of Business, Economics & Informatics

Probability and Statistics
EMMS098S5

30 credits

Friday 13 June 2014

10:00am–1:00pm

This examination contains two sections: Section A (8 questions) and Section B (4 questions). Questions in Section A are worth 5 marks each, and questions in Section B are worth 20 marks each.
Candidates should attempt **all** *of the questions in Section A and* **two** *questions from Section B.*

New Cambridge Statistical Tables are provided.

Candidates can use their own calculator, provided the model is on the circulated list of authorized calculators or has been approved by the chair of the Mathematics & Statistics Examination Sub-board.

Please turn over

Section A

1. Let A and B be two events such that $\Pr(A) = 0.6$, $\Pr(A \cap B) = 0.3$ and $\Pr(A \cup B) = 0.8$.

   (a) Are A and B mutually exclusive? Explain your answer. [1]

   (b) Calculate $\Pr(B)$. [2]

   (c) Are A and B statistically independent? Explain your answer. [2]

2. Of the commuters at a railway station, 40% use an Oyster card. At a given moment, there is a queue of eight commuters at a ticket barrier. Assume that each commuter uses an Oyster card independently of other commuters.

   (a) Suggest a probability distribution for X, the number of Oyster card users in the queue, including the values of any parameters. [3]

   (b) Calculate, correct to three decimal places, the probability that there are exactly three Oyster card users in the queue. [2]

3. Buses are scheduled to leave a bus station every fifteen minutes. Passengers arrive randomly at the bus station at a mean rate of one every 72 seconds. All passengers at the bus station when the next bus is due to leave, travel on that bus. Assume that the number of passengers arriving at the bus station during any fifteen minute period never exceeds the capacity of the bus.

   (a) Write down the probability distribution of X, the number of passengers on board a given bus as it leaves the bus station on time. [3]

   (b) Write down, correct to three decimal places, the probability that there are no more than 18 passengers on a given bus when it leaves the bus station on time. [2]

Please turn over

4. A species of squirrel has a tail length that is normally distributed with mean 21cm and variance 1.44cm^2.

   (a) Find the probability, correct to four decimal places, that the tail of a randomly chosen squirrel is longer than 20cm. [2]

   (b) Find the probability, correct to three decimal places, that the mean tail length of a random sample of 25 squirrels is between 20.5cm and 21.5cm. [3]

5. A continuous random variable X has probability distribution function f given by

$$f(x) = \frac{6}{x^2}, \quad x \geq 6.$$

   (a) Find the median of X. [3]

   (b) Find $\Pr(3 < X < 9)$. [2]

6. A museum shows an educational film regularly every ten minutes. It is suggested that the waiting times of visitors until the start of the next showing follow a discrete uniform distribution. The frequency distribution of the waiting times of 100 visitors to the nearest minute above is shown in the table.

Minutes	1	2	3	4	5	6	7	8	9	10
Frequency	11	26	13	7	9	7	6	8	1	12

   (a) Calculate an appropriate test statistic for conducting a goodness-of-fit test to test the hypothesis that the waiting times follow a discrete uniform distribution. [2]

   (b) Carry out the goodness-of-fit test, stating your conclusions clearly. [3]

Please turn over

7. A random sample of 160 families, each with two children, were assessed to see whether the gender of the younger child depended on the gender of the older child. The data are given in the table below.

	older boy	older girl
younger boy	45	36
younger girl	33	46

(a) Assuming that there is no association between the genders of the older and younger children, calculate the expected frequencies of each term in the table. [2]

(b) Use the chi-square test to investigate whether there is an association between the genders of the older and younger children. [3]

Please turn over

8. In cricket each player competes twice per match to make an individual score. It is believed that players' second scores are lower than their first scores. The scores of completed turns for ten players, P_1–P_{10}, in one match, recorded as 1st and 2nd, are given in the following table together with the difference between 1st and 2nd, and the sign of this difference.

player	1st	2nd	difference	sign
P_1	29	19	10	1
P_2	97	6	91	1
P_3	70	32	38	1
P_4	46	11	35	1
P_5	0	25	−25	−1
P_6	51	96	−45	−1
P_7	24	15	9	1
P_8	0	8	− 8	−1
P_9	0	3	− 3	−1
P_{10}	4	1	3	1

(a) Write down the distribution of the test statistic for the Sign Test, under the null hypothesis that, for a randomly selected player, the first score and second score are equally likely to be larger than the other. [1]

(b) If the Sign Test is to give positive evidence in favour of the second scores being lower than the first scores, at the 10% level of significance, what is the minimum value of the test statistic? [2]

(c) What further assumptions are necessary if, instead of the Sign Test, one wanted to use:

(i) the Signed Rank Test? [1]

(ii) the 1 Sample t-test? [1]

Please turn over

Section B

9. (a) Let $B_1, B_2, \ldots, B_k$ be a set of pairwise mutually exclusive and exhaustive events, all with non-zero probability, and let A be another event.

(i) State (without proof) the *Law of Total Probability* for $\Pr(A)$. [2]

(ii) Suppose, in addition, that $\Pr(A)$ is non-zero. State (without proof) *Bayes' Theorem.* [2]

(b) Three car manufacturers, Amethyst, Baxter and Chianti, charter a breakdown recovery service to attend to malfunctions of their cars during a one-year warranty period. Currently there are 5,000 Amethyst cars, 4000 Baxter cars and 3000 Chianti cars registered for the service. Malfunctions occur either because of an electrical fault or a mechanical problem. During the warranty period, Amethyst cars fail at the rate of eight per thousand for electrical faults and two per thousand for mechanical problems; Baxter cars fail at the rate of four per thousand for electrical faults and twelve per thousand for mechanical problems; Chianti cars fail at the rate of two per thousand for electrical faults and twelve per thousand for mechanical problems.

(i) What is the probability that a call-out is to a Chianti car with a mechanical problem? [2]

(ii) What is the probability that a call-out to a Baxter car is due to an electrical fault? [2]

(iii) The manufacturers contribute to the costs of the scheme in proportion to the frequencies of call-outs to their cars. What proportion of the costs do Chianti pay? [3]

(iv) The breakdown company allocates its training budget between electrical faults and mechanical problems in proportion to the expected number of call-outs for each malfunction. What fraction of their training budget do they spend on electrical faults? [3]

(v) An engineer is called out to a malfunction. Explaining your answer in terms of Bayes' theorem, calculate the probabilities of attending each make of car if she is told that the call-out is for a mechanical problem. [6]

Please turn over

10. Consider a continuous random variable X with probability density function f given by

$$f(x) = c(4-x)^2, \quad 0 \leq x \leq 4,$$

where c is a constant.

(a) Show that $c = \frac{3}{64}$. [2]

(b) Find the mean of X. [3]

(c) Find the variance of X. [4]

(d) Find the corresponding cumulative distribution function F, distinguishing clearly between the values of $F(x)$ for $x < 0$, $0 \leq x \leq 4$ and $x > 4$, respectively. [4]

(e) Find the median of X, correct to three decimal places. [3]

(f) Comment on the relationship of the median of X to the mean of X with reference to the shape of the probability density function. [2]

(g) Evaluate $\Pr(1 < X < 2)$. [2]

Please turn over

11. (a) In a study of the effects of smoking, samples of blood were tested from 36 randomly selected students and the levels of haemoglobin in g/dl were recorded. It was noted whether each participant was a smoker or a non-smoker. The results are shown in the table.

Smokers		Non-smokers	
14.1	14.5	16.8	18.1
13.5	15.1	16.1	13.7
13.9	11.6	14.1	14.8
13.5	13.5	13.5	15.5
12.8	15.9	15.0	15.5
15.0	16.0	15.1	13.2
13.4		16.5	18.6
14.9		14.3	11.7
14.4		16.8	17.7
12.4		16.8	17.5

The data were analysed as shown in the following Minitab output.

```
MTB > TwoSample 'Smokers' 'Non-smokers';
SUBC>   Pooled;
SUBC>   Alternative -1.
```

Two-Sample T-Test and CI: Smokers, Non-smokers

```
Two-sample T for Smokers vs Non-smokers

              N   Mean  StDev  SE Mean
Smokers      16  14.02   1.20     0.30
Non-smokers  20  15.56   1.83     0.41


Difference = mu (Smokers) - mu (Non-smokers)
Estimate for difference:  -1.544
95% upper bound for difference:  -0.646
T-Test of difference = 0 (vs <): T-Value = -2.91  P-Value = 0.003  DF = 34
Both use Pooled StDev = 1.5823
```

This question continues on the next page.

(i) State precisely the statistical model that is being used in the Minitab output, defining carefully any notation that you use. Specify the null and alternative hypotheses in terms of the model parameters. [5]

(ii) Write down the expression for the pooled sample variance for the two populations. [3]

(iii) Draw conclusions in the present case. [2]

(b) In a trial of a dietary iron supplement, which it is hoped will increase haemoglobin levels, 14 people were randomly selected to take the supplement. Their haemoglobin levels in g/dl were measured in blood tests before and after the course of supplements. The measurements are set out below:

Person	Before	After
1	16.3	16.9
2	17.3	16.9
3	15.3	15.7
4	12.7	14.4
5	15.5	16.3
6	14.5	14.6
7	17.2	17.1
8	18.9	19.4
9	15.3	16.3
10	12.9	13.4
11	13.4	12.4
12	14.3	13.5
13	16.3	16.9
14	14.7	15.3

The data were analysed as shown in the following Minitab output.

This question continues on the next page.

```
MTB > Paired 'Before iron' 'After iron';
SUBC>   Alternative 0.
```

Paired T-Test and CI: Before iron, After iron

```
Paired T for Before iron - After iron

              N    Mean  StDev  SE Mean
Before iron  14  15.329  1.770    0.473
After iron   14  15.646  1.857    0.496
Difference   14  -0.317  0.709    0.190


95% CI for mean difference: (-0.727, 0.093)
T-Test of mean difference = 0 (vs not = 0): T-Value = -1.67  P-Value = 0.118
```

(i) Explain why it would be inappropriate to perform the test used in part (a). [1]

(ii) State any distributional assumptions that are made and specify the null and alternative hypotheses used in the analysis. [3]

(iii) Comment on why the one-sided alternative hypothesis may be preferred over the two-sided hypothesis, in this case. Calculate the p-value using the null and the one-sided hypotheses. [6]

Please turn over

12. A country hosting a tennis tournament is permitted to enter eight players as wild-cards who do not have to qualify, and who do not play each other in the first round. The national association wishes to model the number of wild-card players who win their first-round matches in a year. It is proposed that the annual number of first-round wins for the last 40 years follows either a binomial or a Poisson distribution. The observed frequency distribution of the annual number of first-round wins is shown in the table.

No. of 1st-round wins in year	0	1	2	3	4	5	6	7	8	total
Observed frequency	15	12	6	3	0	3	1	0	0	40

(a) In the following Minitab output a goodness-of-fit test has been carried out.

```
MTB > PGoodness 'round1wins';
SUBC>   Frequencies 'frequency';
SUBC>   RTable.
```

Goodness-of-Fit Test for Poisson Distribution

```
Data column: round1wins
Frequency column: frequency

Poisson mean for round1wins = 1.35

                            Poisson                Contribution
round1wins  Observed  Probability  Expected     to Chi-Sq
0               15     0.259240   10.3696       2.06763
1               12     0.349974   13.9990       0.28544
2                6     0.236233    9.4493       1.25911
>=3              7     0.154553    6.1821       0.10821


 N  N*  DF   Chi-Sq  P-Value
40   0   2  3.72039    0.156
```

(i) State the null hypothesis that is being tested, write down a general formula for the test statistic that is being used, and state its approximate distribution under the null hypothesis. [3]

(ii) Explain why the amalgamation of frequencies into a category "≥ 3" has taken place. [3]

(iii) Comment on the Minitab output regarding the goodness-of-fit for a Poisson distribution. [3]

This question continues on the next page.

Please turn over

(b) An alternative to consider is that the number of wins each year for the eight wild-card players follows a binomial distribution with the same mean.

(i) Explain why the parameters of the binomial distribution, $B(n, p)$, being tested for are given by $n = 8$ and $p = 0.16875$. [2]

(ii) Suppose that a goodness-of-fit test is to be used to test whether the proposed binomial model is a good fit for the data. Calculate the expected frequency over the 40 year period in which none of the wild-card players will win, and hence calculate the associated contribution to the goodness-of-fit test statistic. Explain why, from this calculation alone, you can conclude that the binomial distribution is not a good fit to the data. [4]

(c) In the following Minitab output (as an alternative to the goodness-of-fit test) the dispersion test has been carried out.

```
MTB > let k1 = sum(round1wins*frequency)/sum(frequency)
MTB > name k1 'mean'
MTB > let k2 = sum(frequency*round1wins^2)-sum(frequency)*mean^2
MTB > name k2 'ss'
MTB > let k3 = ss/(sum(frequency)-1)
MTB > name k3 'variance'
MTB > Let k4 = SS/mean
MTB > Name k4 'index'
MTB > CDF 'index' k5;
SUBC> ChiSquare 39.
MTB > Let k6 = 1 - k5
MTB > Name k6 'p-value'
MTB > Print  k1-k6
```

Data Display

mean	1.35000
ss	101.100
variance	2.59231
index	74.8889
K5	0.999521
p-value	0.000478882

This question continues on the next page.

Please turn over

(i) State the property of the moments of the Poisson distribution that underlies the dispersion test. [1]

(ii) Write down a general formula for the index of dispersion that is used as the test statistic. [1]

(iii) In the present case, state the value of the index of dispersion, its approximate distribution under the null hypothesis, and its p-value, and draw conclusions. [3]