BIRKBECK

(University of London)

BSc EXAMINATION FOR INTERNAL STUDENTS
SCHOOL OF BUSINESS, ECONOMICS AND INFORMATICS

Probability and Statistics
EMMS098S5

Tuesday, 2nd June, 2015
Morning, 10.00 am – 1.00 pm

This examination contains two sections: Section A and Section B. 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 the four in 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. Three events, A, B and C, are such that $\Pr(A) = 0.2$, $\Pr(B) = 0.4$ and $\Pr(C) = 0.5$. Furthermore, A and B are mutually exclusive, A and C are independent, and $\Pr(B \cap C) = 0.3$. Evaluate the following, giving your answers as fractions.

   (a) $\Pr(A \cup B)$ [1]

   (b) $\Pr(A \cap C)$ [1]

   (c) $\Pr(A \cup C)$ [1]

   (d) $\Pr(A \cup B | C)$ [2]

2. The text in a certain codex (an ancient form of hand-written book) is written in columns, of which there are several hundred. The text of the original scribe has been corrected in places by a later scribe. It has been observed from analysis of the whole text that 34% of the columns contain some corrections.

   A random sample of 20 columns is taken from the codex for more detailed analysis.

   (a) State the probability distribution of the number X of columns in the sample that contain corrections, including the values of any parameters. [1]

   (b) Evaluate the expected value of X. [1]

   (c) Evaluate, correct to 3 decimal places, the standard deviation of X. [1]

   (d) Find, correct to 3 decimal places, the probability that $X \leq 5$. [1]

   (e) Find, correct to 3 decimal places, the probability that $X \geq 10$. [1]

Please turn over

3. Cars that pass through a particular level crossing on a quiet country road, in a given direction, may for most of the day be regarded as arriving in a random process, at an average rate of 15 per hour. Occasionally, when a train is crossing, the barriers at the crossing are lowered for exactly ten minutes, which results in a queue of cars building up at the crossing. Let X denote the total length of the queue of cars that builds up before the barrier is raised again.

   (a) State the probability distribution that should be used for the distribution of X, including the value of any parameter. [2]

   (b) Evaluate, correct to 3 decimal places, the probability that $X = 0$. [1]

   (c) Evaluate, correct to 3 decimal places, the probability that $X = 3$. [1]

   (d) Evaluate, correct to 3 decimal places, the probability that $X \geq 4$. [1]

4. A company produces kilogram bags of potatoes whose weight is normally distributed with mean 1001 grams and standard deviation 2 grams.

   (a) Find the probability, correct to 4 decimal places, that the weight of a randomly chosen bag exceeds 1000 grams. [2]

   (b) If 5 such bags are selected at random, find the probability, correct to 4 decimal places, that the total weight of the bags exceeds 5000 grams. [3]

Please turn over

5. A random sample of 24 hospital patients in the age-range from 70-79 years has been taken and a platelet count (in thousands per cubic millimetre of blood) recorded for each. The sample mean is 193.54 and the sample variance is 2620.17. It is assumed that the data are drawn from a normal distribution with unknown mean μ and unknown variance σ^2.

   (a) Calculate a 95% confidence interval for the mean μ. [2]

   (b) Calculate a 95% confidence interval for the variance σ^2. [3]

6. The abilities of two new radar systems, System A and System B, to detect packages dropped from an aeroplane are being compared. In a series of independent trials, System A detected the packages being dropped 34 times out of 40 and System B detected the packages being dropped 36 times out of 60.

   The null hypothesis to be tested is that there is no difference between the two systems in their ability to detect the packages being dropped.

   (a) Under the null hypothesis, calculate an estimate of the long-term proportion of packages detected. [1]

   (b) Calculate a test statistic to test the null hypothesis, evaluate the corresponding p-value, correct to 4 decimal places, and draw conclusions. [4]

Please turn over

7. Tests were made on random samples to compare the ability of three materials, A, B and C, to withstand extreme temperature changes, with the following observed frequencies:

	Material A	Material B	Material C
Number crumbled	43	28	23
Number not crumbled	77	52	77

The data are to be analysed to test whether the chances of crumbling are the same for all three materials.

(a) Write down a table of expected frequencies, correct to 1 decimal place, given the row and column totals, under the null hypothesis that the chances of crumbling are the same for all three materials. [2]

(b) The value of the chi-square test statistic for testing the null hypothesis turns out to be 4.86. Find the corresponding p-value, to 2 decimal places, and draw conclusions. [3]

8. A company is trying out a new recipe, proposed by its research team, for a tinned soup that it produces.

As part of the company's market research, a random sample of 20 individuals are each given the old and the new soup to taste, in a random order. Of the 20 individuals, 5 say that they prefer the old soup, 13 say that they prefer the new soup and 2 have no preference.

Carry out an appropriate non-parametric test procedure to test whether this provides sufficient evidence to justify the claim by the research team that the new recipe would be more popular than the old. Calculate a corresponding p-value and draw conclusions. [5]

Please turn over

Section B

9. (a) Consider a number of mutually exclusive hypotheses, $H_0, H_1, \ldots, H_k$, with prior probabilities $\Pr(H_i)$ $(0 \leq i \leq k)$ such that $\sum_{i=0}^{k} \Pr(H_i) = 1$. Let E be a piece of evidence for which we know the probabilities $\Pr(E|H_i)$ $(0 \leq i \leq k)$. Given that E is observed, write down Bayes' Theorem for the posterior probabilities, $\Pr(H_i|E)$ $(0 \leq i \leq k)$. [2]

   (b) A chemical company has to pay particular attention to the impurity levels of the chemicals that it produces. Previous experience shows that about 1% of its batches of a certain chemical have an impurity level that is too high.

   To help with quality control, the company has invested in a new instrument for recording whether the impurity level is too high. However, this instrument is not perfect and will from time to time give an erroneous result. Its manufacturers warn that it will falsely give a reading of a high impurity level for about 3% of batches that actually have satisfactory impurity levels. On the other hand, it will falsely indicate a satisfactory impurity level for about 2% of batches that have impurity levels that are too high.

   Provide answers to the following questions and comments from the company management.

   i) If a high impurity reading is obtained from the instrument, what, correct to 4 decimal places, is the probability that the impurity level really is too high? [8]

   ii) If a satisfactory impurity reading is obtained, what, correct to 4 decimal places, is the probability that the impurity level really is satisfactory? [7]

   iii) The management is surprised that your answer to i) is much smaller than they expected and that most of the batches giving a high impurity reading are in fact satisfactory.
   How might you comment on and attempt to explain this apparently surprising result? [3]

Please turn over

10. Consider a continuous random variable X that takes values in the interval (0,3) with cumulative distribution function F given by

$$F(x) = \frac{x(1+x)}{12} \qquad (0 \leq x \leq 3).$$

(a) Find $\Pr(1 \leq X \leq 2)$. [3]

(b) Find the median of X. [2]

(c) Find an expression for the probability density function f of X. [2]

(d) Evaluate the mean of X. [5]

(e) Evaluate the variance of X. [8]

In parts (a),(b),(d) and (e), give your answers as fractions.

Please turn over

11. In an experiment to investigate the effect of aspirin on blood clotting time, for each of 12 subjects, a measurement is made of the time in seconds to clot formation before and then three hours after taking two aspirin tablets.

The results are given in the table below.

Blood clotting time (in seconds)

Subject	Before Aspirin	After Aspirin
1	12.3	12.0
2	12.0	12.3
3	12.0	12.5
4	13.0	12.0
5	13.0	13.0
6	12.5	12.5
7	11.3	10.3
8	11.8	11.3
9	11.5	11.5
10	11.0	11.5
11	11.0	11.0
12	11.3	11.5

In the Minitab output on the next page, two alternative test procedures are carried out to test whether the taking of aspirin has an effect on clotting time.

Let $x_1, x_2, \ldots, x_{12}$ denote the observations before the taking of aspirin for subjects $1, 2, \ldots, 12$, respectively, and let $y_1, y_2, \ldots, y_{12}$ denote the observations after the taking of aspirin for subjects $1, 2, \ldots, 12$, respectively.

(a) i) For **both** of the procedures in the output, specify precisely the underlying statistical model and the null and alternative hypotheses that are being tested. [8]

ii) For **both** of the procedures in the output, write down precisely a formula for the test statistic and state its distribution under the null hypothesis. [7]

(b) State with reasons which of the two procedures is the right one to use in the present case and why it is wrong to use the other one. [3]

(c) Using the test procedure that you identified as the correct one in part (b), draw conclusions. [2]

...continued

Please turn over

```
MTB > Paired 'Before' 'After'.

Paired T-Test and CI: Before, After

Paired T for Before - After

             N    Mean  StDev  SE Mean
Before      12  11.892  0.705    0.204
After       12  11.783  0.748    0.216
Difference  12   0.108  0.507    0.146


95% CI for mean difference: (-0.214, 0.431)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.74  P-Value = 0.475

MTB > TwoSample 'Before' 'After';
SUBC>   Pooled.

Two-Sample T-Test and CI: Before, After

Two-sample T for Before vs After

         N    Mean  StDev  SE Mean
Before  12  11.892  0.705     0.20
After   12  11.783  0.748     0.22


Difference = mu (Before) - mu (After)
Estimate for difference:  0.108
95% CI for difference:  (-0.507, 0.724)
T-Test of difference = 0 (vs not =): T-Value = 0.37  P-Value = 0.719  DF = 22
Both use Pooled StDev = 0.7269
```

12. A slide counting chamber is used is to count the numbers of cells in suspensions of bacteria. The part of the slide from which the cells are counted is partitioned by a grid of lines into 80 equally sized small squares. If the bacterial cells are distributed randomly over the plate, the numbers of bacterial cells per square should follow a Poisson distribution. The observed frequency distribution of the number of bacterial cells in each square is tabulated below.

No. of bacterial cells in square	0	1	2	3	4	5	6	7	total
Observed frequency	4	18	23	15	12	5	2	1	80

(a) i) In the following Minitab output a goodness-of-fit test has been carried out. 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) Draw conclusions in the present case. [2]

```
MTB > PGoodness 'Cells';
SUBC>   Frequencies 'Freq';
SUBC>   RTable.

Goodness-of-Fit Test for Poisson Distribution

Data column: Cells
Frequency column: Freq

Poisson mean for Cells = 2.5125

                    Poisson                Contribution
Cells  Observed  Probability  Expected     to Chi-Sq
0             4     0.081065    6.4852      0.952372
1            18     0.203677   16.2941      0.178592
2            23     0.255869   20.4695      0.312828
3            15     0.214290   17.1432      0.267939
4            12     0.134601   10.7681      0.140938
5             5     0.067637    5.4110      0.031212
>=6           3     0.042861    3.4289      0.053649

 N  N*  DF   Chi-Sq  P-Value
80   0   5  1.93753    0.858

1 cell(s) (14.29%) with expected value(s) less than 5.
```

(b) i) State the reason for the presence of the comment in the output:
`1 cell ...  with expected value less than 5.` [2]

. . . continued

Please turn over

ii) Carry out an amalgamation of the last two categories in the Minitab output into a category "≥ 5", calculating the corresponding observed frequency, Poisson probability, expected frequency and contribution to the chi-square statistic. Calculate the corresponding value of the test-statistic for testing the null hypothesis, state the degrees of freedom, and use the tables to evaluate the corresponding p-value to 3 decimal places. [8]

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

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]

```
MTB > Let k1 = sum(Cells*Freq)/80
MTB > Name k1 'mean'
MTB > Print 'mean'.

Data Display
mean    2.51250

MTB > Let k2 = sum(Cells*Cells*Freq) - 80*mean*mean
MTB > Name k2 'SS'
MTB > Let k3 = SS/79
MTB > Name k3 'variance'
MTB > Print 'variance'

Data Display
variance    2.25301

MTB > Let k4 = SS/mean
MTB > Name k4 'index'
MTB > CDF 'index' k5;
SUBC> ChiSquare 79.
MTB > Let k6 = 1 - k5
MTB > Name k6 'p-value'
MTB > Print  'index' 'p-value'

Data Display
index      70.8408
p-value    0.732164
```