# UNIVERSITY OF DUBLIN

## TRINITY COLLEGE

**Faculty of Engineering and Systems Sciences**
**Department of Statistics**

**B. A. (Mod) Computer Science**
**JS Examination** **Trinity Term 2003**

**Statistics and Numerical Analysis**

**Thursday 29th May** **Sam. Beckett Room** **09.30 – 12.30**

**Dr Mosurski, Prof Byrne**

**Answer five questions at least one of which is from section B. Use separate answer books for each section. Statistical tables are available from the invigilator. Statistical formulas are attached.**

### Section A

1. A large market survey of users of home computers was carried out on behalf of a PC manufacturer. The purpose behind the survey was to design an appropriate salable package. The respondents were asked to tick three options from a list peripherals, expansions etc. that they would most like to see in the package.

The table overleaf shows the estimated probabilities of user preferences of what should be included in the package for the three most frequently asked for options.

Three items are considered:

| | ticked | not ticked |
|---|---|---|
| RAM expansion | $R$ | $\overline{R}$ |
| High quality printer | $P$ | $\overline{P}$ |
| Scanner | $S$ | $\overline{S}$ |

| | | $S$ | $\overline{S}$ |
|---|---|---|---|
| $R$ | $P$ | 0.10 | 0.08 |
| | $\overline{P}$ | 0.15 | 0.14 |
| | | | |
| $\overline{R}$ | $P$ | 0.17 | 0.19 |
| | $\overline{P}$ | 0.15 | 0.02 |

The entries in the table give the probabilities of the appropriate combination of the three events. Eg $P(P \cap R \cap S) = 0.10$ , $P(R \cap P \cap \overline{S}) = 0.08$ etc.

(a) Compute the following probabilities and say what they mean in the context of the survey and its aims. (12 marks)

(i) $P(P)$

(ii) $P(P \cup S)$

(iii) $P(S \mid P)$

(iv) $P(P \mid S)$

(v) $P(S \mid P \cup R)$

(vi) $P(S \cap P \mid \overline{R})$

(vii) Are the events P and S independent? Justify.

(viii) If it is decided that only two options are going to be included, what is the most favoured combination?

Note: "$\cap$" means AND , "$\cup$" means OR.

(b) The manufacturer decides that everyone will get, the RAM upgrade (for logistical reasons) and customers will be allowed to choose either a printer or a scanner as they wish. Assuming that customers, that ticked neither or both of these options, are equally likely to request either (they will always choose something), compute the

expected proportion of all customers that will choose a printer. Is the assumption sensible? Discuss briefly. (8 marks)

2. A SPAM filter when set a certain threshold has the following characteristics: 1% of genuine messages get classified as SPAM, 95% of SPAM messages get classified as SPAM.

(a) Express these characteristics in terms probabilities involving the events G – genuine message, $\overline{G}$ - SPAM message and S – message classified as SPAM, $\overline{S}$ - message classified as genuine. (4 marks)

(b) For a certain email account $P(G) = 0.6$. Suppose this filter is applied to this account, compute $P(\overline{G} \mid S)$ and $P(G \mid \overline{S})$. Interpret these probabilities. (8 marks)

(c) Ten randomly selected messages are received by this account. What is the appropriate model for X, the number of SPAM messages amongst these 10? Give expressions, Excel or otherwise for:

P(none of the messages are SPAM)
P(all of the messages are SPAM) (3 marks)

(d) On average the account receives 30 messages a day. Why might Y, the number of messages in a day, be Poisson distributed? What assumptions underlie a Poisson process model for message arrivals? Briefly discuss whether you think it is reasonable here. Give an expression, Excel or otherwise, for the probability that no messages arrive on a particular day.

If the Poisson process model is an appropriate description of the number of messages and the proportion of SPAM messages is 0.6, what would be an appropriate model for the number of SPAM messages per day? (5 marks)

3. (a) The number of PCs sold by a retail outlet follows the distribution tabulated below: (10 marks)

| Daily Sales | Probability |
|---|---|
| 0 | 0.1 |
| 1 | 0.2 |
| 2 | 0.4 |
| 3 | 0.2 |
| 4 | 0.1 |

(i) What are the expected value, E(X), and the variance, V(X), of daily sales?

(ii) The retail outlet is part of a retail chain of 50 such shops. Assuming that the distribution of daily sales in each of the shops is identical with the one above and due to geographic distances the daily sales are independent. What are the mean and variance of the total daily sales by this chain?

(iii) Briefly explain why a Normal distribution would provide an approximation to actual daily sales. Compute the approximate probability that the number of sales exceeds 120.

(b) The manager of one of the outlets has observed that demand exceeds supply and he feels that he could increase his sales if could hold more stock. He has collected data on daily demand for PCs. To establish the appropriate stock level he wants to conduct a simulation of the system.

Illustrate how this could be achieved by simulating two values for the demand distribution given in the table on the next page using the random numbers provided:.

You may use any appropriate method.

| Demand | Probability |
|---|---|
| 0 | 0.10 |
| 1 | 0.20 |
| 2 | 0.30 |
| 3 | 0.10 |
| 4 | 0.10 |
| 5 | 0.10 |
| 6 | 0.05 |
| 7 | 0.05 |

Random Numbers (assume independent U(0,1))

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.753 | 0.147 | 0.001 | 0.302 | 0.005 | 0.941 | 0.211 |
| 0.368 | 0.808 | 0.536 | 0.200 | 0.600 | 0.618 | 0.593 |
| 0.781 | 0.715 | 0.358 | 0.366 | 0.074 | 0.020 | 0.603 |

(10 marks)

4. (a) Two OCR systems are compared as follows. Eight randomly selected pages from the Irish Times are scanned and converted to text using the Acuread software. The error rates per 1000 characters are recorded. Another (different) eight pages from the same publication are scanned, the images are converted to text using the Besttex software. (8 marks)

The data with their summaries appear in the table below.

| | Acuread | Besttex |
|---|---|---|
| | 25.41 | 40.97 |
| | 39.39 | 32.82 |
| | 29.56 | 29.03 |
| | 27.87 | 36.14 |
| | 38.91 | 28.02 |
| | 28.93 | 22.52 |
| | 26.92 | 33.02 |
| | 37.32 | 23.83 |
| $\bar{X}$ | 31.789 | 30.794 |
| $\hat{\sigma}$ | 5.757 | 6.199 |

(i) Assuming that error rates are Normally distributed, compare the mean error rates for the two systems.

(ii) Give a brief discussion of the circumstances in which one-tail tests and two-tail tests are appropriate. Discuss the choice of significance level.

(b) A statistician is consulted as to the analysis in (a) above. She suggests that a matched comparison should have been carried out. She explains what this means: the same pages should have been scanned by both systems and a test of the differences should have been conducted. (12 marks)

(i) Explain why this design is preferable.

(ii) It is decided that to conduct the experiment again. This time 4 pages of the Irish Times are scanned and the number of errors recorded. The number of errors for each page using each system is given in the table below.

| | Acuread | Besttex |
|---|---|---|
| Page 1 | 491 | 489 |
| Page 2 | 430 | 413 |
| Page 3 | 410 | 394 |
| Page 4 | 507 | 482 |

Again assuming that the number or errors in Normally distributed carry out a formal comparison.

Discuss the results of these two analyses.

5. CompSolutions is a company that takes on contracts for a variety of IT problems. The company accountant is concerned that the rates they charge for different types of work might vary. To investigate this she randomly selects a number of jobs and classifies them according as one of the following types:

| | |
|---|---|
| Web development | (Web) |
| Data base implementation | (Dbase) |
| Custom solutions. | (Custom) |

She then obtains a profit per man-hour figure by dividing the price charged by the number of man-hours devoted to that contract. This results in the following table:

| Profit per man-hour | | |
|---|---|---|
| Web | Dbase | Custom |
| 223.91 | 207.58 | 219.09 |
| 212.75 | 241.09 | 196.18 |
| 250.32 | 242.39 | 207.83 |
| 208.29 | 219.13 | 175.34 |
| 213.01 | 224.34 | 211.33 |
| 230.04 | 200.68 | |
| 212.25 | 241.20 | |
| 245.10 | | |
| 235.73 | | |
| 221.02 | | |

A one-way analysis of variance was carried out on the data in MINITAB.

(a) Write down the linear model for the one-way layout and explain the meaning of each term. (2 marks)

(b) Reconstruct the ANOVA table by replacing the "?" by appropriate values in the output below. Interpret the results. (12 marks)

**One-way ANOVA: Web, Dbase, Custom**

```
Analysis of Variance
Source     DF        SS        MS        F        P
Factor      ?         ?         ?        ?        ?
Error       ?      4849         ?
Total      21      6941

                                    Individual 95% Cis For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
Web        10    225.24     14.67                      (------*------)
Dbase       7    225.20     17.09                     (-------*--------)
Custom      5    201.95     17.01   (---------*---------)
                                   ------+---------+---------+---------+
Pooled StDev =    15.97             195       210       225       240
```

(Contd...)

(c) The graphical display of confidence intervals suggests that "Custom solutions" jobs are less profitable than the other two types. If the 17 jobs classified as Web or Dbase are treated as a single group it is observed that $\bar{X}$ =225.23, $\hat{\sigma}$ = 15.18. Carry out a standard t-test to compare the mean with that for 'custom solutions'; use a 5% significance level. (6 marks)

6. To investigate the effect of prolonged keyboard/mouse use on RSI (Repetitive Stress Injury), data were collected on 30 individuals who had a varying amounts of computer usage in the work place.

The X variable is the average number of hours per day of computer use.
The Y variable is the RSI index (0 –100) as scored by a health professional.

A regression analysis was carried out in Data Desk. The output follows the end of the question.

(a) Write down the model equation and interpret it. What is the meaning of the associated "prob" values? Comment on the values observed in this case. (5 marks)

(b) What does $R^2$ measure? Comment on the value. (3 marks)

(c) I use my computer an average of 4 hours per day. What does the model predict as my RSI index value? Provide an interval estimate of this. (9 marks)

Note: $(\bar{X} = 3.65, SSX = 54.5)$

(Contd...)

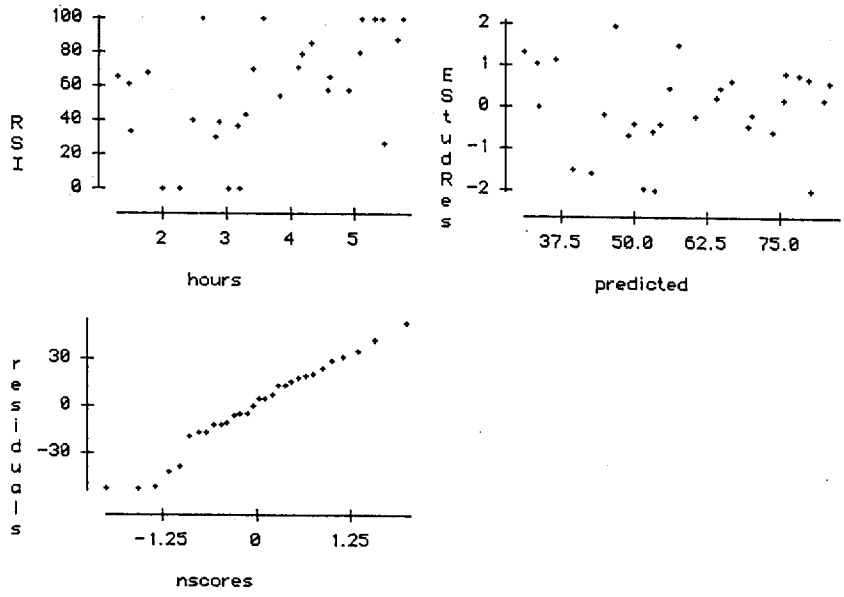
(d) Scatter, residual and probability plots appear below the output. Why is it important to examine these plots? Discuss the plots in this case. (3 marks)

Dependent variable is:RSI
No Selector

R squared = 24.3% R squared (adjusted) = 21.6%
s = 28.86 with 30 - 2 = 28 degrees of freedom

| Source | Sum of Squares | df | Mean Square | F-ratio |
|---|---|---|---|---|
| Regression | 7505.22 | 1 | 7505.22 | 9.01 |
| Residual | 23318.6 | 28 | 832.809 | |

| Variable | Coefficient | s.e. of Coeff | t-ratio | prob |
|---|---|---|---|---|
| Constant | 15.8884 | 15.21 | 1.04 | 0.3053 |
| hours | 11.734 | 3.909 | 3 | 0.0056 |



## Section B

7. Write down Newton's method for finding a root of a single non-linear equation f(x).

5 marks

Show that the method has quadratic convergence provided f'(x) <> 0.

10 marks

If it is known that f(x) has one double root outline how you would establish that Newton's method has linear convergence. You need not do all the algebraic manipulations.

5 marks

8. What is linear programming?

2 marks

Write down the equations for the second primal form.

2 marks

(a) Describe, in general, the simplex method using the exchange algorithm. In particular describe how the variables to be exchanged are chosen.

7 marks

(b) Show how to solve a linear programme by the graphical method. What happens at each step of the simplex method?
Sketch 3 situations which can arise. 7 marks

(c) Given m constraints and n variables derive a formula for the complexity of the simplex algorithm.

2 marks

## Technical Formulae

One sample

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

Two samples

Pooled $\hat{\sigma}^2 = s^2$ from two groups with sizes n and m and standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$

$$\hat{\sigma}^2 = s^2 = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{n+m-2}$$

z-statistic

$z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$, where $\sigma^2$ is known, is distributed as N(0,1).

t- statistic

$t = \dfrac{\bar{X} - \mu}{\dfrac{\hat{\sigma}}{\sqrt{n}}}$ is distributed as Student's t with n-1 degrees of freedom

f-statistic

$f = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are two independent estimates of the same variance with $df_1$ and $df_2$ degrees of freedom respectively is distributed as F on $df_1$, $df_2$ .

Excel Functions for probabilities (discrete random variables)

Binomial(n,p)

$P(X = x)$ =BINOMDIST(x,n,p,0)

$P(X \le x)$ =BINOMDIST(x,n,p,1)

Hypergeometric

X in sample of size n, from population with R in total size N

$P(X = x) =$ HYPERGEOMDIST(x,n,R,N)

Poisson with mean $\lambda$

$P(X = x) =$ POISSON(x, $\lambda$ ,0)

$P(X \leq x) =$ POISSON(x, $\lambda$ ,1)

Linear regression:

$SSX = \sum x^2 - n\bar{X}^2$, $SSE = \sum r_i^2$,

$a =$ estimate of intercept, $b =$ estimate of slope, $\hat{y}(x_h) = a + bx_h$

$$\hat{\sigma}^2 = \frac{SSE}{n-2},$$

$$V(b) = \frac{\hat{\sigma}^2}{SSX}$$

$$V(\hat{y}(x_h)) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSX}\right) \qquad \text{(mean)}$$

$$V(\hat{y}(x_h)) = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSX}\right) \text{ (individual value)}$$