STAB22F: 1999 Midterm Solutions

Name:

Student Number:

Time: 1 hour, 50 minutes
Aids: The exam is open book. Students may use any notes, books and calculators in writing this exam.

1. (a) (5 marks) With a categorical variable what is the appropriate method for recording its distribution in a dataset and for plotting this distribution?

Typically it is best to tabulate the categorical variable in a table recording each possible value of the variable and the count or proportion of times that value occurred in the data set. The bar chart is the plot used to depict this distribution.

(b) (5 marks) What is the fundamental difference between categorical and quantitative variables?

The fundamental difference is that with quantitative variables you can perform artithmetic; i.e. add, multiply, etc, on the values of the variable while it doesn't make any sense to do this with categorical variables as they only serve to classify.

2. Suppose a population consists of 10 objects and there are two quantitative variables X and Y taking the following values.

Object	1	2	3	4	5	6	7	8	9	10
X	1	1	2	1	3	2	3	3	1	3
Y	0	-1	0	-1	0	0	-1	0	0	0

(a) (5 marks) Record the empirical distribution functions of X and Y.

The empirical distribution function for X is given by

X	1	2	3
F_{10}	.4	.6	1.0

and the empirical distribution function for Y is given by

Y	-1	0
F_{10}	.3	1.0

.

(b) (5 marks) Record the conditional emprical distribution functions of X given Y.

The conditional empirical distribution function for X when $Y = 0$ is given by

X	1	2	3
F_{10}	2/7	4/7	1

and the conditional empirical distribution function for X when $Y = -1$ is given by

X	1	2	3
F_{10}	2/3	2/3	1

.

(c) (5 marks) Is there a relationship between the variables? Justify your answer.

Yes there appears to be a relationship between X and Y because the conditional distributions of X given Y change as we change the value of Y.

3. (a) (5 marks) Suppose that you have the values of a quantitative variable in a dataset stored in C1 in a Minitab worksheet. Indicate how you would use Minitab to sort these values from largest to smallest and store the values in C2.

```
MTB >sort c1 c2;
SUBC>descending c1.
```

(b) (5 marks) Suppose you are asked to compute the area under a $N(6, 3)$ density curve over the interval on the x-axis from -2 to 8. Show how you would do this using Minitab.

```
MTB >cdf 8;
SUBC>normal 6 3.
```

```
Cumulative Distribution Function
Normal with mean = 6.00000 and standard deviation = 3.00000
 x P( X <= x)
 8.0000 0.7475
MTB >let k1=.7475
MTB >cdf -2;
SUBC>normal 6 3.
Cumulative Distribution Function
Normal with mean = 6.00000 and standard deviation = 3.00000
 x P( X <= x)
 -2.0000 0.0038
MTB >let k2=.0038
MTB >let k3=k1-k2
MTB >print k3
Data Display
K3 0.743700
```

(c) (5 marks) Using Minitab commands show how you compute the point on the x-axis such that the area under the $N(6, 3)$ density curve to the left of this point equals .67.

```
MTB >invcdf .67;
SUBC>normal 6 3.
Inverse Cumulative Distribution Function
Normal with mean = 6.00000 and standard deviation = 3.00000
 P( X <= x) x
 0.6700 7.3197
```

The point is 7.3197.

4. Suppose the values for a quantitative variable in a dataset stored in C1 in a Minitab worksheet give the boxplot shown in Figure 1.

(a) (5 marks) What are approximate values for the median and quartiles for this variable?

The median is approximately 0, the first quartile is approximately -.5 and the third quartile is approximately .5.

(b) (5 marks) Based on this boxplot does it seem reasonable to approximate the distribution of the variable by a normal density curve? Justify your answer.

Yes it appears reasonable the distribution appears to be symmetric about its centre.

5. Suppose we are interested in whether or not a relationship exists between two variables X and Y defined on a population of students where X is the average

number of television hours watched weekly and Y is the GPA earned at the end of the year.
(a) (5 marks) Identify each of the following kinds of studies
(i) A set of students is randomly selected from the population and their (X, Y) values are recorded.
(ii) A set of students is randomly selected from the population and each student is assigned a number of television hours that they watch each week and their (X, Y) values are recorded.

(i) sampling study
(ii) experiment

(b) (5 marks) If you want to assert that a cause-effect relationship exists between these variables which study is appropriate? Justify your answer.

An experiment must be conducted if we want to assert that a cause-effect relationship exists between the variables. This is because we must be sure that there are no other variables which cause X and Y to change simultaneously. By controlling the assignment of the values of the X variable we prevent this.

6. Suppose the correlation between two quantitative variables X and Y is .6.
(a) (5 marks) What is the correlation between Y and $3X$?

The correlation between Y and $3X$ is .6 as the correlation between variables is scale invariant.

(b) (5 marks) Can Y be accurately predicted, when X is known, by an expression of the form $a + bX$? Justify your answer.

Only $.6^2 = .36$ of the variation in Y is explained by linear variation in X. So Y cannot be accurately predicted from an expression of the form $a + bX$.

(c) (5 marks) Can the correlation coefficient be used to measure the degree of relationship existing between two categorical variables? Justify your answer.

No it does not make sense to use the correlation coefficient to measure the degree of relationship existing between two categorical variables. This is because the correlation coefficient is computed using arithmetic on the values of the variables and this doesn't make sense with categorical variables.

7. Suppose two quantitative variables are stored in C1 and C2 in a Minitab worksheet. the following command is given to Minitab

```
MTB >regress c1 1 c2
```

and the following (partial) output obtained

```
Regression Analysis:  C1 versus C2
The regression equation is
```

```
C1 = 0.992 + 2.00 C2
```

(a) (5 marks) Explain how you would predict the value of C1 when C2 = 3. Provide suitable Minitab commands.

```
MTB >let k1= 0.992 + 2.00*3.0
MTB >print k1
```

(b) (5 marks) In Figure 2 there is a plot of the standardized residuals against the values of the predictor varibale. What is the purpose of this plot and what comments can you make based on its appearance?

We examine this plot to determine if the assumptions; e.g. the assumption of a linear relationship, underlying the analysis make sense. In this case the plot seems reasonable as there is no discernible pattern in it and all the standardized residulas lie in the interval $(-3, 3)$.

(c) (5 marks) Explain what is meant by the least-squares line that is computed from the data values $(X_1, Y_1), \ldots, (X_n, Y_n)$.

The least-squares line is the line given by $Y = \hat{a} + \hat{b}X$ where a and b are the values minimizing the sum of squared deviations of the observed values of Y from their predicted values; i.e. we minimize

$$\sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

for a and b.

8. (10 marks) Write out the Minitab commands to generate a sample of 100 values of the statistic $T(X_1, X_2, \ldots, X_{10}) = X_1^2 + X_2^2 + \cdots + X_{10}^2$ when $X_1, X_2, \ldots, X_{10}$ are a sample from the distribution given by the table below with $P(X = x)$ denoting the proportion of population members with X taking the value x.

X	1	2	3	4
$P(X = x)$	.2	.4	.1	.3

```
MTB >read c1 c2
DATA>1 .2
DATA>2 .4
DATA>3 .1
DATA>4 .3
DATA>end
 4 rows read.
MTB >random 100 c3-c12;
SUBC>discrete c1 c2.
MTB >rssq c3-c12 c13
```