STAB22F:1998 Midterm Solutions

Name:

Student Number:

Tutorial:

Time: 2 hours
Aids: The exam is open book. Students may use any notes, books and calculators in writing this exam.

1. (5 marks)
Identify the types of following variables defined on the population of students at the University of Toronto.

height
hair color
age
blood type
degree program

2. In a population of businesses the following tabulation is made:

Sales	Have a web-site	Don't have a web-site
Less than $\$10^6$	1000	2000
Between $\$10^6$ and $\$10^7$	2500	2500
Greater than $\$10^7$	200	300

(a) (5 marks) Record the conditional distributions of incidence of web-site given sales.

The 3 conditional distributions are given in the following table.

Sales	Have a web-site	Don't have a web-site
Less than $\$10^6$	$\frac{1000}{3000} = .33$	$\frac{2000}{3000} = .66$
Between $\$10^6$ and $\$10^7$	$\frac{2500}{5000} = .5$	$\frac{2500}{5000} = .5$
Greater than $\$10^7$	$\frac{200}{500} = .4$	$\frac{300}{500} = .6$

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

Since the conditional distributions differ considerably it would appear that there is a relationship between sales and incidence of web-site.

3. Suppose you have 990 values of a quantitative variable.
(a) (5 marks) Indicate how you would use Minitab to compute the empirical distribution of this data.

Suppose the data is stored in c1. Then the commands

```
MTB > sort c1 c1
MTB > set c2
DATA> 1:990
DATA> end
MTB > let c2=c2/990
MTB > print c1 c2
```

print out the empirical distribution function evaluated at the data values.

(b) (5 marks) Suppose that part of the print out of the empirical distribution function gave the following results (with the data values in c1 and the empirical distribution function in c2). What is the value of the empirical distribution function at 57?

C1	C2
56	0.00101
57	0.00202
57	0.00303
58	0.00404
58	0.00505
58	0.00606
59	0.00707

The value of the empirical distribution function at 57 is .00303.

(c) (5 marks) What is the value of the .005 quantile in the data?

The value of the .005 quantile is 58 as the empirical distribution function at 58 is .00606 and this is closest to .005.

4. Suppose the density histogram in Figure 1 is obtained for a variable stored in c1.

Figure 1

(a) (5 marks) Using the histogram estimate the proportion of the data less than or equal to 9.

The proportion less than or equal to 9 is approximately

$$(.01)(6-5)+(.15)(8-6)+(.25)(9-8)=.56$$

(b) (5 marks) What does it mean to approximate a density histogram by a density curve?

A density histogram is approximated by a density curve f when the proportion of data values lying in any interval (a,b) is approximately equal to

$$\int_a^b f(x)\,dx$$

namely the area under the density curve f over the interval (a,b).

5. Suppose we have observations of the quantitative variables (X, Y) with the values of X in c1 and the values of Y in c2. A linear relationship is considered possible for these data. The least-squares line is computed via Minitab to be:

```
The regression equation is
C2 = 18.1 + 0.94 C1
```

(a) (5 marks) Indicate how you would predict the value of Y when $X = 5$ (use Minitab).

The prediction is given by the value of k1.

```
MTB > let k1=18.1+.94*5
MTB > print k1
Data Display
K1 22.8000
```

(b) (5 marks) A plot of the standardized residuals is given in Figure 2. Comment on the adequacy of the fit.

Figure 2

The distribution of the standardized residuals appears to be somewhat assymmetrical about 0 with a preponderance of positive residuals with several of these being relatively large. This plot indicates that there are some problems with the model that has been fitted.

(c) (5 marks) It is claimed for this data that the standardized residuals (which are stored in c5) are approximately distributed $N(0,1)$. To assess the validity of this claim Figure 3 was plotted. How was the plot carried out using Minitab? What comments can you make concerning the validity of the claim?

Figure 3

The Minitab commands for carrying this plot are

```
MTB > nscores c5 c6
MTB > plot c6*c5
```

It would not appear that the claim of normal variation is valid as there is a definite bend in the plot.

6. Suppose you have a variable X defined on a population and the probability function for the variable under simple random sampling is given by:

X	1	2	3
f_X	.3	.5	.2

(a) (5 marks) Indicate how you would use Minitab to compute the mean and standard deviation of X.

The following commands print out the mean as k1 and the standard deviation as k2.

```
MTB > read c1 c2
DATA> 1 .3
DATA> 2 .5
DATA> 3 .2
DATA> end
 3 rows read.
MTB > erase c5 c6
MTB > let c3=c1*c2
MTB > let k1=sum(c3)
MTB > let c4=c1-k1
MTB > let c5=c2*c4*c4
MTB > let k2=sum(c5)
MTB > let k2=sqrt(k2)
MTB > print k1 k2
Data Display
K1 1.90000
K2 0.700000
```

(b) (10 marks) Give the session commands you would use to carry out a simulation to assess the sampling distribution of the statistic $T(X_1, \ldots, X_n) = \bar{X} - 3S$ when taking a sample of size $n = 15$ from the population.

The following commands generate 1000 samples of size 15 and plot a density histogram of $\bar{X} - 3S$ for these 1000 samples.

```
MTB > random 1000 c3-c17;
SUBC> discrete c1 c2.
MTB > rmean c3-c17 c18
MTB > rstdev c3-c17 c19
MTB > let c20=c8-3*c19
MTB > histogram c20;
SUBC> density.
```

(c) (5 marks) How would you assess the accuracy of your simulation?

One method of assessing the accuracy of the simulation is to increase the number of samples generated say from 1000 to 5000 to 10000 and see if the density histogram begins to satbilize.

7. Suppose random variable X follows a $N(0,1)$ distribution and random variable Y follows a $Uniform(0,2)$ distribution and let $Z = X/Y$.
(a) (5 marks) Indicate how you would approximate $P(Z > 4)$ using Minitab.

The following commands compute such an estimate in k1.

```
MTB > random 1000 c1;
SUBC> normal 0 1.
MTB > random 1000 c2;
SUBC> uniform 0 2.
MTB > let c3=c1/c2
MTB > let c4=c3 > 4
MTB > let k1=mean(c4)
MTB > print k1
Data Display
K1 0.0500000
```

(b) (10 marks) Indicate how you would approximate $SD[Z] = \sqrt{VAR[Z]}$ using Minitab.

The following commands compute such an estimate in k2.

```
MTB > random 1000 c1;
SUBC> normal 0 1.
MTB > random 1000 c2;
SUBC> uniform 0 2.
MTB > let c3=c1/c2
MTB > let k1=mean(c3)
MTB > let c4=c3-k1
MTB > let c4=c4**2
MTB > let k2=mean(c4)
MTB > let k2=sqrt(k2)
MTB > print k2
Data Display
K2 22.6119
```

8. A study is to be conducted to determine if there is a relationship between coffee drinking and blood pressure amongst adults over 50 in Canada. Coffee drinking is measured in milliliters consumed daily.
(a) (5 marks) What is the population? What is the response and what is the predictor variable?

The population consists of all Canadians over 50 in Canada.
The response is blood pressure measured in some units.
The predictor variable is the amount of coffee consumed daily.

(b) (5 marks) Comment on the various types of study that could be carried out. In particular comment on the feasibility of determining whether or not a cause-effect relationship exists between these variables in this population.

One could conduct an observational study where one took some group of Canadians over 50 and measured their blood pressure and questioned them about the amount of coffee they consumed daily. A better approach, to avoid selection effects, would be to conduct a sampling study where we randomly selected a sample of n individuals from this population so that each subset of size n had the same probability of being selected and then measured their blood pressure and coffee consumption. It is probably not feasible to do a truly random sample here so some sort of compromise will have to be made. It is not possible to determine if a cause-effect relationship exists because it is not possible to conduct an experiment; i.e. we cannot randomly select individuals from this population and then assign the amount of coffee they will consume daily.