STAB22F:2000 (summer) Midterm Solutions

Time: 1 hour
Instructions: Only use No. 2 black lead pencil on the Answer Sheet. First fill-in your student number and name on the Answer Sheet. Indicate your answer to each question by thoroughly filling in the appropriate circle on the Answer Sheet. Each question is worth 5 marks.
Aids: The exam is open book. Students may use any notes, books and calculators in writing this exam..

The correct answer is marked with a *.

1. Suppose we have a variables X defined on a population Π. To approximate the distribution of X by a density curve f means that

(a) X is a quantitative variable.
(b)* the proportion of population members having their X measurement in any interval (a, b) is approximately equal to $\int_a^b f(x)\,dx$.
(c) the values of $f(x)$ are close to a histogram of the values of X..

2. Suppose that we have two categorical variables X and Y defined on a population. It is possible for X and Y to have correlation equal to 0.

(a) True
(b)* False

3. Suppose that we have a variable X defined on a population Π. Suppose that the mean of X is 10.2 while the median is 8.4. Which of the following is the correct summary to report of the location and scale characteristics of the distribution of X.

(a) The mean is 10.2 and the standard deviation is 2.2.
(b)* The median is 8.4 and the interquartile range is 4.3.
(c) The mean is 10.2 and the interquartile range is 4.3.
(d) The median is 8.4 and the standard deviation is 2.2.

4. Suppose that a random variable X follows the distribution given by

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{12}$

Suppose that a sample of size 2 is taken from this distribution obtaining X_1, X_2. Then $P(X_1 = X_2)$ equals

(a)* $\frac{19}{72}$.
(b) 1.
(c) $\frac{1}{3}$.

5. In repeated sampling the long-run average value of X in **4** is

(a) 0
(b)* -$\frac{7}{12}$
(c) $\frac{7}{12}$

6. Suppose that the joint distribution of X and Y is given by $P(X = 0, Y = 0) = 1/2, P(X = 1, Y = 0) = 1/2$. Then X and Y are statistically independent.

(a)* True
(b) False

7. Suppose that the empirical distribution function of a variable in a dataset takes the following values.

$$\begin{array}{ll} x & F_n(x) \\ \vdots & \vdots \\ 2.3 & .6 \\ 3.1 & .65 \\ 3.2 & .70 \\ \vdots & \vdots \end{array}$$

What is the sample size n?
(a) 3
(b) 10
(c)* 20

Figure 1: Histogram.

8. Suppose that two symmetric dice are rolled and two even numbers are obtained. What is the conditional probability that the sum of the dice equals 6?

(a) $\frac{1}{3}$.
(b)* $\frac{2}{9}$.
(c) $\frac{7}{36}$.

9. Suppose we plotted the histogram in Figure 1. Choose the most appropriate statement.
(a) The histogram tells us that we have a somewhat left-skewed distribution.
(b)* The histogram is not appropriate because the intervals are of different lengths and we didn't plot a density histogram.
(c) From the histogram we can obtain the number of observations that fall in each of the intervals.

10. The boxplot in Figure 2 indicates that the first quartile is approximately

(a) 0.0.
(b) 2.2.
(c)* -0.5.
(d) -1.6.

Figure 2: Boxplot.

11. Suppose that we have a sample of n from a distribution and we compute the sample standard deviation. Which of the statements below is the most appropriate when discussing the sampling distribution of this quantity?

(a) The sampling distribution of the sample standard deviation is not defined.
(b) The sampling distribution of the sample standard deviation tells us about the population distribution that we are sampling from.
(c)* The sampling distribution of the sample standard deviation tells us what proportion of sample standard deviations will fall in each interval when we repeatedly take samples of size n.
(d) The sampling distribution of the sample standard deviation is a normal distribution.

12. The following Minitab commands

```
MTB >random 1000 c1-c20;
SUBC>binomial 1 .3.
MTB >let c21 = rmean(c1-c20)
MTB >histogram c21;
SUBC>density.
```

(a) compute an approximation to the mean of a Binomial$(1, .3)$ distribution.
(b)* compute an approximation to the sampling distribution of the proportion of heads when a coin is tossed 20 times and the probability of an individual head is .3.
(c) compute a density curve approximation to the Binomial$(1, .3)$ distribution.

13. If we take a sample of 100 from a Binomial$(10, .5)$ distribution then the sample mean has standard deviation

(a) 10.
(b)* $\sqrt{2.5}/10$.
(c) .25.
(d) .5.

14. Suppose that 5 students are randomly selected from a class containing 40 females and 60 males. If after each selection the selected student is returned to the population then the total number of females selected is distributed

(a)* Binomial$(5, .4)$.
(b) $N(5, .4)$.
(c)Binomial$(100, .4)$.

15. Suppose that the values of a response variable are stored in C3 and the values of a predictor variable are stored in C2. The following Minitab command and output

```
MTB >regress c3 1 c2
The regression equation is
C3 = 3.09 + 4.91 C2
Predictor    Coef   SE Coef      T      P
Constant   3.0858    0.6762   4.56 0.000
      C2   4.9112    0.1090  45.07 0.000
S = 3.130 R-Sq = 95.4%
```

indicates that

(a) there is a weak linear relationship existing between the response and the predictor variable.
(b)* there is a strong linear relationship existing between the response and the predictor variable.
(c) there is no linear relationship existing between the response and the predictor variable.

16. Given the output in **15** suppose that we wanted to predict a future response when the predictor variable takes the value 0. This prediction is given by

(a)* 3.09.
(b) 4.91.

(c) 3.09 + 4.91.

17. If we carry out a regression analysis and have concluded that a relationship exists between a response variable and a predictor variable then we can safely conclude that this is a cause-effect relationship due to the power of the regression technique.

(a) True.
(b)* False.

18. Suppose that I decide to study the relationship between gender and grades in introductory statistics courses and for my study I use the current STAB22 class. Which of the following statements is most appropriate?

(a) By applying rigorous statistical methods I can make inferences about the relationship between these variables in the general population of students at U of T.
(b) If I conclude that there is a relationship between these variables then I cannot conclude that it is a cause-effect relationship because I can't control gender.
(c)* This is only an observational study and it is very risky to extrapolate conclusions from such a study to a wider population.

19. The essential feature of an experiment when compared to a sampling study is

(a) the random allocation of treatments to experimental units.
(b)* we can assign the values of the predictor variable to experimental units.

20. Suppose that we have two categorical variables X and Y. The most relevant plot for determining whether or not a relationship exists between these variables is

(a) a histogram.
(b) a scatter plot.
(c)* bar charts of the conditional distributions of one variable given the values of the other..
(d) side-by-side boxplots.