UNIVERSITY OF TORONTO SCARBOROUGH
Department of Computer and Mathematical Sciences
Midterm Test March 2014

STAB22H3 Statistics I
Duration: 1 hour and 45 minutes

Last Name:________________________ First Name: ________________________

Student number: ________________________

Aids allowed:

- One handwritten letter-sized sheet (both sides) of notes prepared by you
- Non-programmable, non-communicating calculator

Standard normal distribution tables are attached at the end.

This test is based on multiple-choice questions. The are 30 questions. All questions carry equal weight. On the Scantron answer sheet, ensure that you enter your last name, first name (as much of it as fits), and student number (in "Identification").

Mark in each case the best answer out of the alternatives given (**which means the numerically closest answer if the answer is a number and the answer you obtained is not given.**)

Also before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

There are 16 pages including this page. Please check to see you have all the pages.

Good luck!!

1. A consumer reporting magazine published an article comparing 10 models of infant car seats in Canada. This article provided information on the brand, cost, age limit, weight limit, and overall safety rating. Which of the following choices correctly identifies the W's (Note in this question is only interested in three W's: Who, What and Why. Please identify the choice that identifies all these three W's correctly.)

   A) Who: the 10 infant car seat models; What: Overall safety rating; Why: To provide information to readers

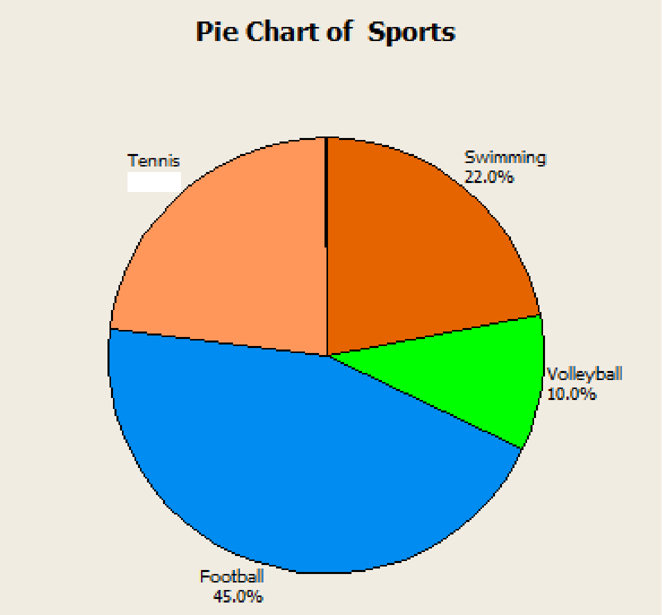
   B) Who: Magazines; What: Articles; Why: To provide information to readers

   C) Who: Consumer reporting magazine; What: Infant car seat models; Why: To provide information to readers

   D) Who: the 10 infant car seat models; What: Brand, cost, age limit, weight limit, and overall safety rating; Why: To provide information to readers

   E) Who: Consumers; What: Consumer reporting magazine; Why: To provide information to the editors of this magazine

2. Four hundred members of a sports club were asked, "What is your favourite sport?". The pie chart shows the results:



   How many members chose Tennis as their favourite sport?

   A) 23 members

   B) 92 members

   C) 100 members

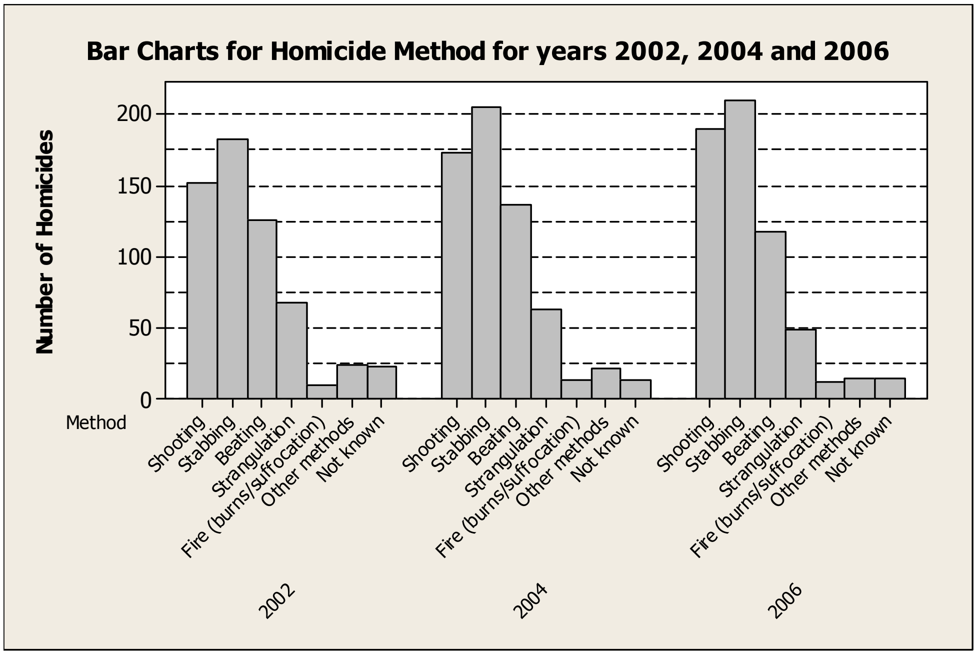
   D) 50 members

Question 2 continues on the next page...

E) 108 members

Solution: The number of members who chose Tennis as their favourite sport = $400 \times (1 - 0.45 - 0.22 - 0.10) = 92$

3. Are homicides increasing in Canada? Do the methods used change over time? Trying to figure out the answers to these questions, an investigator analyzed some homicide counts for three years (2002, 2004 and 2006) obtained from Statistics Canada reports: The bar charts obtained from this analysis is given below:



One of the statements below is **NOT** supported by the information in this bar chart. Which one?

A) The number of homicides by shooting appears to increase over time.

B) The number of homicides by stabbing appears to increase over time.

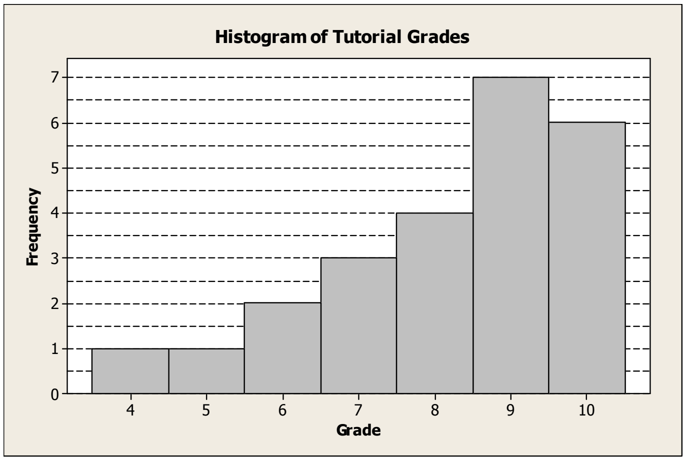
C) The number of homicides by strangulation appears to increase over time.

D) In each of these three years, the highest number of homicides have happened by stabbing.

E) All the above statements are supported by the information in the bar chart above.

4. The histogram below shows the distribution of tutorial grades in a Statistics class.

Question 4 continues on the next page. . .



Consider the following statements regarding the distribution of tutorial grades in this class, each of which is either true or false.

I The distribution of tutorial grades is skewed to the right.

II For these grades, the class average must be greater than the class median.

III For this distribution, the median is a more appropriate measure of centre than the mean.

Which of these statements is (are) true?

A) Only statement I is true

B) Only statement II is true

C) Only statement III is true

D) All three statements are true

E) None of the statements is true

5. The stemplot of the survival times (in months) after a treatment for 28 patients with severe chronic left-ventricular heart failure is given below:

```
Variable: Survival Time

Decimal point is 1 digit(s) to the right of the colon.
0 : 579
1 : 00234
1 : 55666779
2 : 0
2 : 7799
3 : 0123
3 : 569
```

What is the median survival time?

Question 5 continues on the next page. . .

A) 1.6 months

B) 16 months

C) 1.7 months

D) 17 months

E) 19 months

Solution: median = average of the 14th and 15th observation = $(17 + 17)/2 = 17$.

6. Here are some summary statistics for the recent exam in Statistics: lowest score = 31, mean score = 67, median = 81.2, range = 79, $IQR = 60$, $Q1 = 27$. Between what two values are the middle 50% of the scores found?

A) 31 and 110

B) 67 and 81.2

C) 27 and 87

D) 16.75 and 50.25

E) 20.3 and 60.9

Solution: This is question 3 , p7, Chap 6, Deveaux (question bank). The middle 50% of the scores will be between $Q1$ and $Q3$. $Q1 = 27$ and $Q3 = Q1 + IQR = 27 + 60 = 87$

7. A national achievement test is administered annually to 3rd graders. The test has a mean score of 100 and a standard deviation of 15. If Jane's z-score is 1.20, what was her score on the test?

A) 82

B) 88

C) 100

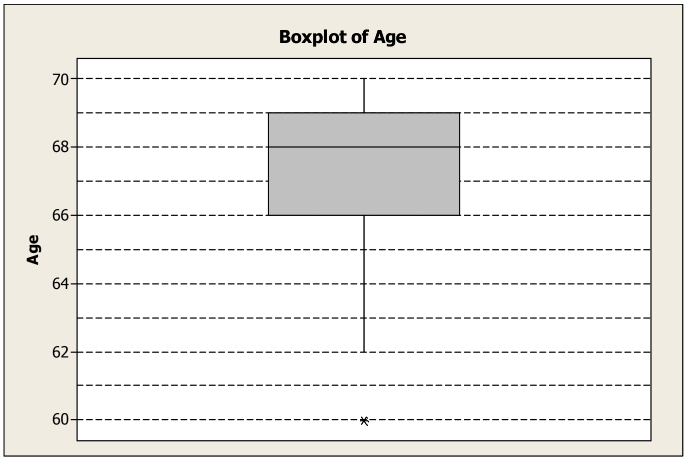
D) 112

E) 118

Solution: Score $= 100 + 15 \times 1.20 = 118$

8. Age at onset of Parkinsons disease, a degenerative brain disorder, was determined for a sample of adults between 60 and 70 years old. The boxplot of their ages is shown below:

Question 8 continues on the next page. . .



Consider the following statements regarding the distribution of the age at onset of Parkinsons disease, each of which is either true or false.

I The distribution of the age at onset of Parkinsons disease is skewed to the right.

II The ***mean*** age of the adults in this sample is 68 years.

III The interquartile range is 3 years.

Which of these statements is (are) true?

A) Only statement I is true

B) Only statement II is true

C) Only statement III is true

D) All three statements are true

E) None of the statements is true

9. A study was conducted on students in a university. The investigators recorded the following four variables:

X1 = Type of car the student owns
X2 = Number of courses taken during that semester
X3 = The time (in minutes) the student waited in line at the bookstore to pay for his/her textbooks
X4 = Home state (or province) of the student

How many categorical variables are there in the above list of variables?

A) There are no categorical variables in the above list.

B) There is only one categorical variables in the above list.

C) There are only two categorical variables in the above list.

D) There are only three categorical variables in the above list.

Question 9 continues on the next page. . .

E) All four variables in the above list are categorical variables.

10. Which one of the following statistics is most affected by outliers?

    A) median

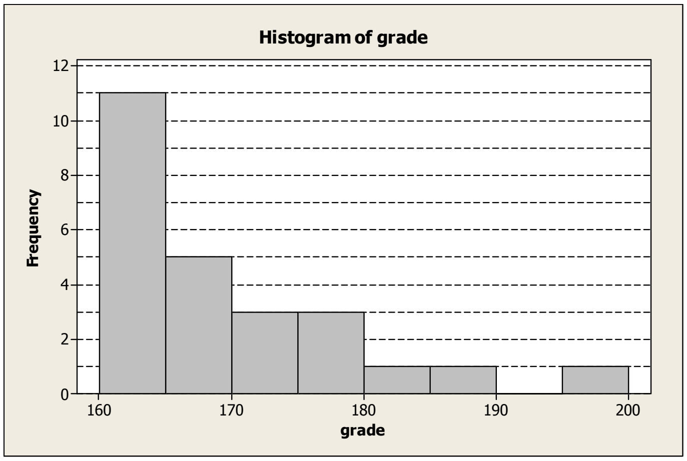
    B) the first quartile

    C) the third quartile

    D) standard deviation

    E) interquartile range

11. The histogram below shows the grades of a class of 25 students (you may assume that there were no grades exactly equal to any of the class boundaries of the histogram):



Which of the above statements about the median grade of this class is true?

A) The median is between 160 and 165

B) The median is between 165 and 170

C) The median is between 170 and 175

D) The median is between 175 and 180

E) The median is greater than 180.

Solution: $n = 25$ and so median is the 13th observation in the ordered data set which is in the second class.

12. Which of the following regression equations represents the strongest linear relationship between x and y?

    A) $\hat{y} = 2 + 0.3x$

    B) $\hat{y} = 2 + 1.3x$

Question 12 continues on the next page. . .

C) $\hat{y} = 2 + 1.8x$

D) $\hat{y} = 2 + 3x$

E) The strength of the relationship cannot be determined from the regression equation.

13. Many nutritional experts have expressed concern about the high levels of sodium in prepared foods. The stemplot below shows the data on sodium content (in milligrams) per frozen meal(Boston Globe, April 24, 1991). Note that there are 27 observations in this data set.

```
Variable: Sodium Content

Decimal point is 2 digit(s) to the right of the colon.

 3 : 0499
 4 : 0255
 5 : 023
 6 : 0389
 7 : 268
 8 : 0558
 9 : 1459
10 : 5
```

Calculate the interquartile range(IQR).

A) 4

B) 40

C) 100

D) 400

E) 750

Solution: For odd n, $Q1$ is the average of the 7th and the 8th observation from the bottom $= (450 + 450)/2 = 450$ and $Q3$ is the average of the 7th and the 8th observarion from the top $= (850+850)/2 = 850$ and so $IQR = Q3-Q1 = 850-450 = 400$.

14. The StatCrunch summary statistics and the stemplot of systolic blood pressure values of a group of individuals are shown below:

```
Summary statistics:
```

Question 14 continues on the next page. . .

```
Column                    n  Mean      Median  Min  Max   Q1   Q3
Systolic Blood Pressure 31 119.06     118     92  165   109  124

Variable: Systolic Blood Pressure

Decimal point is 1 digit(s) to the right of the colon.
 9 : 2
 9 :
10 : 24
10 : 66888
11 : 000000
11 : 68888
12 : 0022
12 : 6
13 : 024
13 : 6
14 : 2
14 :
15 :
15 :
16 : 0
16 : 5
```

Based on the $1.5 \times IQR$ rule, **how many** outliers are there in this data?

A) There are no outliers

B) There is only one outlier

C) There are only two outliers

D) There are only three outliers

E) There are more than three outliers

Solution: Upper fence $= 124 + 1.5 \times (124 - 109) = 146.5$, Lower fence $= 109 - 1.5 \times (124 - 109) = 86.5$ and so only two outliers (160 and 165).

15. A researcher wishes to determine whether the rate of water flow (in liters per second) over an experimental soil bed can be used to predict the amount of soil washed away (in kilograms). The researcher measures the amount of soil washed away for various flow rates, and from these data calculates the least-squares regression line to be:

$$\text{amount of eroded soil} = 0.4 + 1.3 \times (\text{flow rate})$$

Consider the following three statements, each of which is either true or false:

Question 15 continues on the next page...

I The correlation between amount of eroded soil and flow rate is $\frac{1}{1.3}$.

II In this study, the response variable is amount of eroded soil.

III When the flow rate increases by one liters/sec., the erosion increases by 0.4 kilograms.

Which of these statements is (are) true?

A) Only statement I is true

B) Only statement II is true

C) Only statement III is true

D) All three statements are true

E) None of the statements is true

16. The StatCrunch output (with some values deleted) below, was obtained from a study of the relationship between the weight (x) (in 1000 lbs.) and highway fuel efficiency (y) (in miles/gallon) for a sample of 13 cars.

```
Simple linear regression results:
Dependent Variable: y
 Independent Variable: x
y = 50.436616 - 6.2270665 x
Sample size: 13
R (correlation coefficient) = deleted
R-sq = 0.80598054
```

What is value of the sample correlation between x and y? Choose the closest.

A) 0.8

B) -0.8

C) 0.9

D) -0.9

E) -0.6

Solution: $r = -\sqrt{0.8059} = 0.8977193325 \approx -0.9$. Correlation is negative because the slope of the regression line is negative.

17. If the slope of the regression line of y on x is calculated to be 2.5 and the intercept 16 then the predicted value of y when $x = 4$ is:

A) 26

B) 16

Question 17 continues on the next page...

C) 2.5

D) 66.5

E) 18.5

Solution: $\hat{y} = 16 + 2.5 \times 4 = 26$

18. Family income is normally distributed with mean $25,000 and standard deviation $10,000. If the poverty level is $10,000, what percentage of the population lives in poverty?

A) About 7%

B) More than 7%

C) 5% or less

D) 2.5% or less

E) More than 10%

Solution: Let X = Family income. We want to find $P(X \leq 10,000)$. let $Z = (X - 25,000)/10,000$. If $x = 10,000$, then $z = (10,000 - 25,000)/10,000 = -1.5$. So, $P(X \leq 10,000) = P(Z \leq -1.5) = .0668$. Hence, around 7% of the population lives in poverty.

19. A radio station claims that the amount of advertising per hour of broadcast time has an average of 17 minutes and a standard deviation equal to 2.2 minutes. You listen to the radio for 1 hour, at a randomly selected time and carefully observe that the amount of advertising time is 15 minutes. Calculate the z-value for this amount of advertising time.

A) 0.91

B) -0.91

C) -0.75

D) 0.75

E) 1.96

Solution: $z = \frac{15-17}{2} \approx -0.91$

20. Which of the following is NOT used for the “five-number summary” of a data set?

A) The sample median.

Question 20 continues on the next page...

B) The smallest value in the sample.

C) The first quartile, Q1.

D) The third quartile, Q3.

E) The sample mean.

21. The top 5% of applicants (as measured by GRE scores) will receive scholarships. If GRE scores are normally distributed with mean 500 and standard deviation 100, how high does your GRE score have to be to qualify for a scholarship?

A) 335 or higher

B) Between 335 and 665

C) 665 or higher

D) 565 or higher

E) 925 or higher

Solution: Let X =GRE score. We want to find x such that $P(X \geq x) = .05$ Let Z be a z-score for X. Find value z so that $P(Z \geq z) = .05$. From the table $z = 1.65$ (approximately). Thus, $x = 500 + (1.65 \times 100) = 665$. Thus, your GRE score needs to be 665 or higher to qualify for a scholarship.

22. The children in a school are to have extra swimming lessons if they cannot swim. The table below gives information about the children in Years 7, 8 and 9 (total of 534 children).

	Can swim	Cannot swim
Year 7	120	60
Year 8	168	11
Year 9	172	3

What percentage of all children can swim?

A) 14%

B) 86%

C) 34%

D) 66%

E) 26%

Solution: $(120 + 168 + 172)/534 = 460/534 = 0.86 = 86\%$

23. Using the information in question 22 above, what proportion of those who cannot swim are in year 8?

    A) 6%

    B) 94%

    C) 15 %

    D) 2%

    E) 33%

Solution: $11/(60 + 11 + 3) = 11/74 = 0.148 = 15\%$

24. The **variance** of exam marks in a course is 16. If the instructor adds 2 marks to each student, what will be the **standard deviation** of the new scores?

    A) 16

    B) 4

    C) 2

    D) 6

    E) 20

Solution: Standard deviation doesn't change.

25. The middle 95% of students at school are between 1.1m and 1.7m tall. Assuming these data are Normally distributed, what are the mean and standard deviation of the heights of these students?

    A) The mean is 1 and standard deviation is 0.1

    B) The mean is 1.4 and standard deviation is 0.1

    C) The mean is 1.3 and standard deviation is 0.2

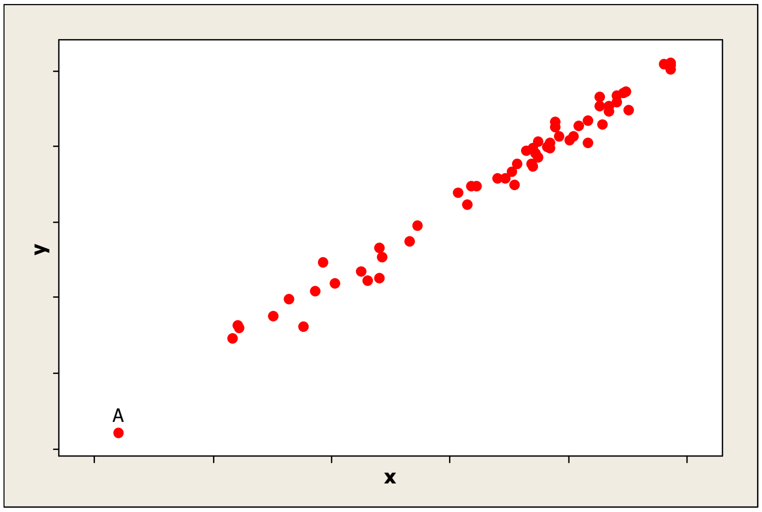
    D) The mean is 1.4 and standard deviation is 0.15

    E) Not enough information to determine

Solution: The mean is halfway between 1.1m and 1.7m: Mean = (1.1m + 1.7m) / 2 = 1.4m 95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so 1 standard deviation = (1.7m-1.1m) / 4 = 0.6m / 4 = 0.15m

26. Consider the following scatterplot:

Question 26 continues on the next page…



Read the following statements labeled I, II and III carefully.

I. Point A is influential.

II. Point A has high leverage.

III. Point A has a large residual.

Which of the above statements is/are true?

A) Only I is true.

B) Only II is true.

C) Only III is true.

D) Only I and III are true.

E) Only II and III are true.

27. Your score in a recent test was 0.5 standard deviations above the average. Which statement about other students' scores is correct? (We assume that the scores are normally distributed.)

A) 95% of people scored lower than you did

B) 59% of people scored lower than you did

C) 40% of people scored lower than you did

D) 69% of people scored lower than you did

E) Not sufficient information to say anything about other students' scores

Solution: $P(Z \leq 0.5) = 0.6915$, so about 69% scored lower

28. For which of the following correlations would the data points be clustered most closely around a straight line?

    A) 0.1

    B) 0.5

    C) -0.9

    D) 0.8

    E) There is no relationship between the value of r and how close the data points are to the line

29. You measure the heights in inches and the weights in pounds of a group of students. Based on these measurements the correlation between height and weight was 0.6. If you decide to convert the weights of these students to kilograms (1 lb=0.45 kg), then the correlation between height in inches and weight in kilograms will be

    A) 0.6 multiplied by 0.45

    B) 0.6 divided by 0.45

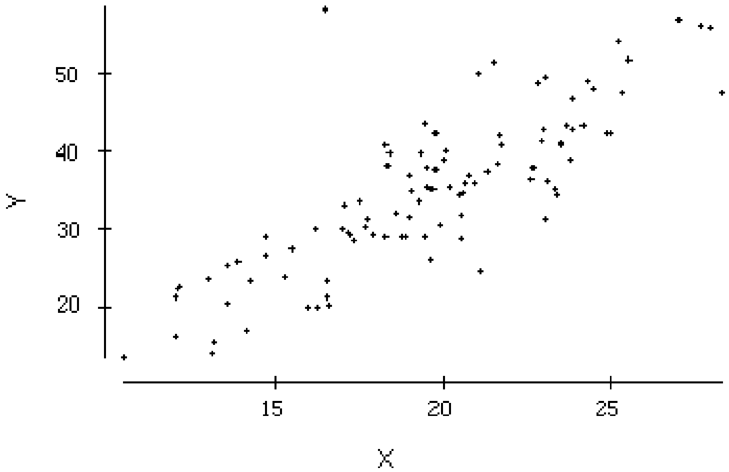
    C) 0.6 multiplied by $(0.45)^2$

    D) 0.6 divided by $(0.45)^2$

    E) 0.6

Solution: The correlation does not change when we change units of measurement.

30. Consider the following scatterplot.



The correlation between X and Y is approximately

A) 0.999

B) 0.3

C) -0.75

D) 0.8

E) 0

Solution: 0.999 implies a very strong correlation with points very close to a straight line.

0.3 and 0 imply a weak relationship and the relationship on this scatterplot is not weak. We can clearly see that Y increases as X increases.

-0.75 implies a negative relationship.

END OF TEST