University of Toronto Scarborough
STAB22 Midterm Examination

October 2009

For this examination, you are allowed one handwritten letter-sized sheet of notes (both sides) prepared by you, a non-programmable, non-communicating calculator, and writing implements.

This question paper has 14 numbered pages; before you start, check to see that you have all the pages. There is also a signature sheet at the front and statistical tables at the back.

This examination is multiple choice. Each question has 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.)

Before you begin, check that the colour printed on your Scantron sheet matches the colour of your question paper. If it does not, get a new Scantron from an invigilator.

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.

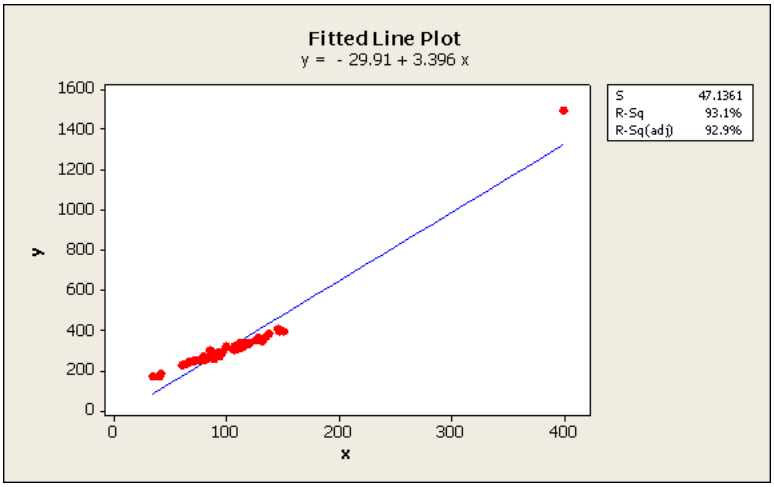
1. Look at the experiment depicted below.

```
                          Treatment 1
                          Mike, Joe, Mary, Jill ----- record results
                        /
Subjects:             /
Mike, Joe, Mary, Jill \
                        \
                          Treatment 2
                          Mike, Joe, Mary, Jill ----- record results
```

   Who is Mike matched with?

   (a) Mary
   (b) treatment one
   (c) himself
   (d) treatment two
   (e) both treatments one and two

2. The scatterplot below shows the association between a variable x and a variable y, with the regression line superimposed. Use the scatterplot to answer this question and the one following.



   How would you describe the point with $x = 400$?

   (a) Having a large negative residual
   (b) Influential
   (c) Outlier

3. In the scatterplot of Question 2, what would happen if the point with $x = 400$ were removed?

   (a) The slope must become less
   (b) The slope would not change
   (c) The correlation must become lower
   (d) The slope must become greater

4. When I ride the bus to school, I note how many minutes the journey takes. My last 10 journeys had a mean length of 37 minutes. Which of the following names describes the 37 minutes?

   (a) parameter

   (b) statistic

   (c) census

   (d) sampling variability

   (e) sample

5. Dairy inspectors visit Ontario farms unannounced and take samples of the milk. If the milk is found to contain dirt, antibiotics or other foreign matter, the day's milk output from the farm is destroyed (and the farm re-inspected until the purity of the milk is satisfactory).

   Suppose the dairy inspectorate's farm sampling procedure is as follows: First, randomly select a sample of Ontario counties. Then, within each selected county, take a simple random sample of dairy farms. Then, visit each of the sampled dairy farms.

   What kind of sampling procedure is this?

   (a) Multistage sample

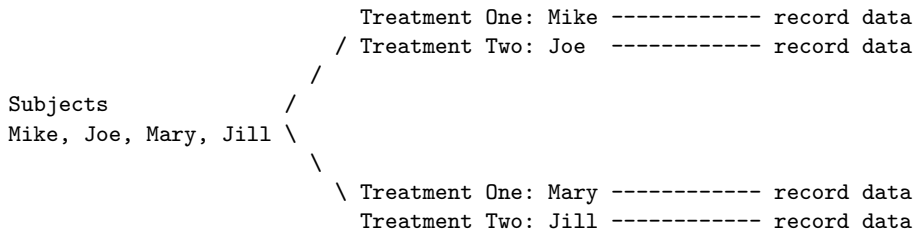
   (b) Stratified sample

   (c) Systematic sample

   (d) Simple random sample

   (e) Voluntary-response sample

6. Look at the experiment depicted below.



   Which factor was blocked?

   (a) subjects

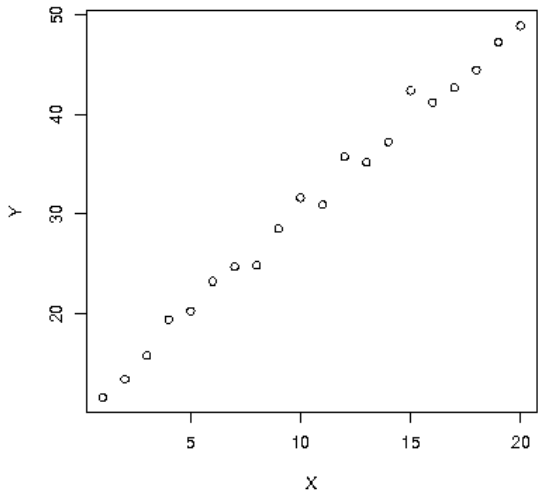
   (b) treatment two

   (c) recorded data

   (d) gender

   (e) treatment one

7. Look at the scatter plot below.



The correlation between X and Y is closest to:

(a) -0.99
(b) -0.55
(c) 0.55
(d) 0.99

8. For human women, a pregnancy lasts about 9 months. For other animals, the gestation period (average length of pregnancy) is different. A researcher believes that longer-lived animals generally have longer gestation periods. The researcher measured life expectancy in years and gestation period in days, and did a regression for predicting gestation period from life expectancy. The regression line had intercept -39.5 and slope 15.5, with an R-squared of 72.2%. Use this information for this question and the following two.

What is the predicted gestation period, in days, for an animal with life expectancy 10 years?

(a) 115
(b) 200
(c) 3
(d) 89

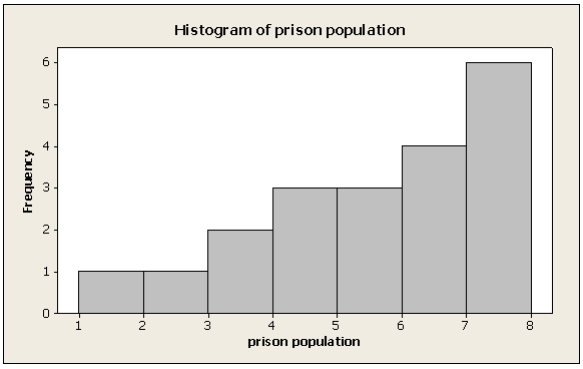
9. Would you guess that your prediction in Question 8 was reasonably accurate or not, based on the information given in that question?

(a) no, because the slope is small.
(b) yes, because the slope is positive.
(c) no, because we must have been extrapolating.
(d) yes, because R-squared is quite high.
(e) no, because R-squared is low.

10. Humans have an average life expectancy of 80 years and their gestation period is 280 days. Using the information in Question 8, what is the residual when using that regression equation to predict human gestation period from human life expectancy?

    (a) 0
    (b) -500
    (c) 500
    (d) 900
    (e) -900

11. A report from the US Department of Justice gave the percent increases in prison populations in 20 northeastern and midwestern states. These are shown in the histogram below. Use the information in the histogram for this question and the next one.



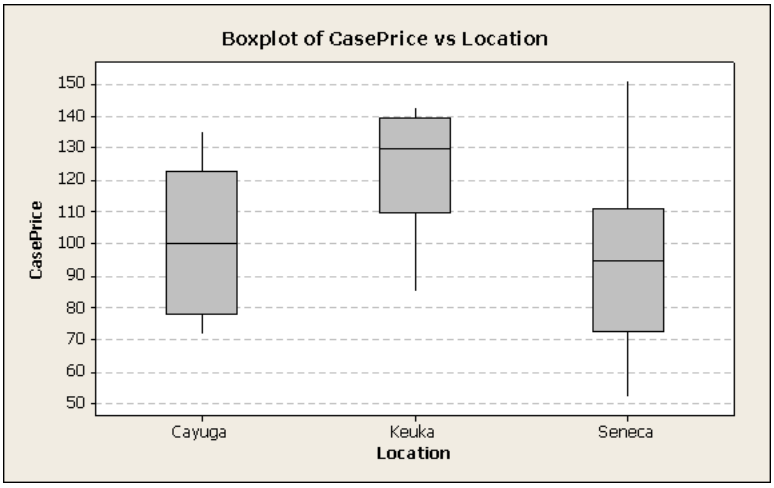
    How would you describe the *shape* of this histogram?

    (a) Approximately symmetric
    (b) Skewed to the left
    (c) Like a normal distribution
    (d) Skewed to the right

12. Look again at the histogram in Question 11. The mean percentage increase is 5.6. Which of the following values could be the median percentage increase?

    (a) 4.6
    (b) 5.6
    (c) 7.4
    (d) 6.1
    (e) 3.5

13. The boxplots below show the display case prices (in dollars) of varieties of wine produced by vineyards along three different lakes. Use this information for this question and the 2 following.



Which distribution of wine prices has the largest median?

(a) Cayuga

(b) Keuka

(c) Seneca

(d) two or more are tied for the largest median

14. In the boxplots of Question 13, which distribution of wine prices has the largest spread (as measured by the interquartile range)? (Use the boxplots as accurately as you can.)

(a) Seneca

(b) two or more are tied for the largest spread

(c) Cayuga

(d) Keuka

15. In the boxplots of Question 13, which distribution of wine prices is clearly skewed to the left?

(a) Keuka

(b) Seneca

(c) Cayuga

(d) none of them

(e) two or more of them

16. The people on the ship *Titanic* when it sank were passengers, either in first class, second class or third class, or crew members. Records were kept so that we know how many people were in each group. Use this information for this question and the one following.

Suppose we wanted to know whether the third class passengers made up more or less than a quarter of all the people on the ship. Which graph would be most appropriate for showing this?

(a) a histogram

(b) a pie chart

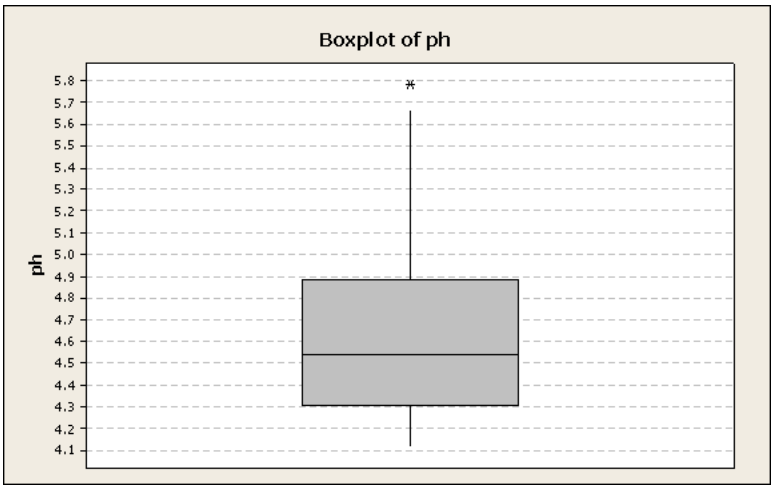
(c) a stemplot

(d) a bar chart

17. Using the information in Question 16 above, suppose that we wanted to know which group (first class passengers, second class passengers, third class passengers, crew) had fewest people in it. Which graph would be most appropriate for showing this?

    (a) a bar chart
    (b) a histogram
    (c) a stemplot
    (d) a pie chart

18. In a study of acid rain, two researchers measured the pH of water collected from rain and snow in Allegheny County, Pennsylvania. (pH is a measure of acidity; a value of 7 is neutral, and values below 7 are acidic). The results are shown in the boxplot below. Use this information for this question and the four questions following.



    What is the mean pH value?

    (a) Between 5.6 and 5.7
    (b) 0.6
    (c) Cannot determine mean from a boxplot
    (d) Between 4.5 and 4.6

19. Look again at the boxplot in Question 18. What is the inter-quartile range of pH values?

    (a) 0.6
    (b) 4.9
    (c) 5.8
    (d) 4.55
    (e) 1.6

20. Look again at the boxplot in Question 18. How many outliers are shown on the boxplot?

    (a) A boxplot does not say anything about outliers.
    (b) 2 or more
    (c) Impossible to say, because some of the values in the upper whisker could be outliers.
    (d) 1
    (e) None

21. How would you describe the shape of the distribution of pH values as shown in Question 18?

    (a) Cannot conclude anything about shape from a boxplot.
    (b) Skewed to the right.
    (c) Approximately symmetric.
    (d) Like a normal distribution.
    (e) Skewed to the left.

22. In the boxplot of Question 18, about what percentage of the data values are above 4.9?

    (a) 10%
    (b) 25%
    (c) 50%
    (d) 40%

23. A certain population has 12 people in it, as below:

	Males		Females
1	Ken	1	Shelley
2	Mike	2	Megan
3	Zengxin	3	Amy
4	Mark	4	Ming
5	Siavash	5	Tharshini
6	Ajay	6	Janine

    It is desired to sample 4 people from this population, but it is also desired to select an equal number of males and females, so a stratified sample will be used. Below is an excerpt from Table B:

    ```
    27260 92145 39974 234
    ```

    Use this excerpt from Table B to select your stratified sample. (If you have to choose, select the males first.) Which females did you sample?

    (a) no females
    (b) Amy only
    (c) Shelley and Ming
    (d) Megan and Shelley
    (e) 3 or more females

24. You would expect the correlation between student IQ scores and squared student IQ scores to be

    (a) 1
    (b) meaningless
    (c) both "meaningless" and "need more information" are correct
    (d) need more information

25. A survey was conducted on the amount of gasoline used per person in each US state (measured in US gallons). The results are shown in the stemplot below. Use the stemplot for this question and the next one.

```
Stem-and-leaf of gas usage  N  = 50
Leaf Unit = 10


 1   2  9
 1   3
 2   3  2
 2   3
 2   3
 3   3  8
 5   4  01
 9   4  2333
 17  4  44555555
 22  4  66677
 24  4  89
(9)  5  000011111
 17  5  22233
 12  5  44444555
 4   5  667
 1   5  8
```

What is the median amount of gasoline used per person, according to the stemplot?

(a) 50

(b) between 25 and 26

(c) impossible to obtain median from stemplot

(d) 500

26. Use the stemplot shown in Question 25 to find the inter-quartile range. What value do you get?

(a) 80

(b) 530

(c) 200

(d) 50

(e) 450

27. In a regression calculation, a researcher finds that the explanatory variable x has mean 100 and SD 10, and the response variable y has mean 250 and SD 40. The regression equation is found to be $\hat{y} = 450 - 2x$. What is the correlation between x and y?

(a) cannot tell from the information available

(b) -0.8

(c) -0.5

(d) 0.4

(e) 0.1

28. Consider the sampling distribution of a sample statistic. If the sample size is increased, which of the following will happen?

(a) the bias of the statistic will decrease.

(b) the variability of the statistic will decrease.

(c) the variability of the statistic will increase.

(d) the sampling distribution will have a less normal shape.

29. A study was made of the association between (female) life expectancy and the average number of children born per woman in a number of different countries. Some information about these two variables is given below. Use this information for this question and the one following.

```
Descriptive Statistics: Births/woman, Life Exp.

Variable       N     Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Births/woman  26    2.854    0.149  0.761    1.500   2.275   2.750   3.275
Life Exp.     26   74.500    0.814  4.150   64.000  71.000  74.500  78.000

Variable      Maximum
Births/woman    4.700
Life Exp.      82.000

Pearson correlation of Births/woman and Life Exp. = -0.812
```

What is the *intercept* of the regression line for predicting number of births per woman from the female life expectancy?

(a) 87

(b) -4.4

(c) -0.15

(d) 14

30. Using the information in Question 29, what would be your predicted number of births per woman in a country where female life expectancy is 60 years?

(a) less than 0

(b) about 5

(c) about 2

(d) about 3

(e) should not do a prediction because this is extrapolation

31. The proportion of 4's in table B (the table of random digits) should be closest to

(a) 0.2

(b) 0.4

(c) 0.1

(d) 0.3

32. A data set has median 55, first quartile 50 and third quartile 70. Which of the following statements will correctly identify the outliers in the data set, according to the rule for outliers learned in class?

(a) Values below 50 or above 70.

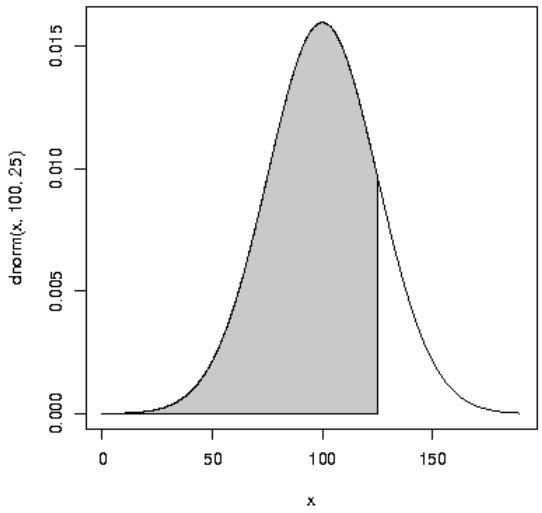
(b) Values below 30 or above 90.

(c) Values below 20 or above 100.

(d) Values below 20 or above 80.

(e) Values below 25 or above 85.

33. Below is the normal density curve with mean $\mu = 100$ and standard deviation $\sigma = 25$.



What proportion of the curve is the shaded area?

(a) less than 125
(b) greater than 0.90
(c) greater than 0.50
(d) less than 0.50

34. A smelt is a type of food fish. Smelt lengths are normally distributed with mean 15 cm and standard deviation 1 cm. Use this information for this question and the next one.

What proportion of smelts are between 13.5 and 15.5 cm long?

(a) 0.62
(b) 0.26
(c) 0.31
(d) 0.82

35. Using the information in Question 34, how long are the longest 10 percent of smelts?

(a) bigger than 10.14 cm
(b) bigger than 16.28 cm
(c) 10.14 cm
(d) less than 16.28 cm
(e) 16.28 cm

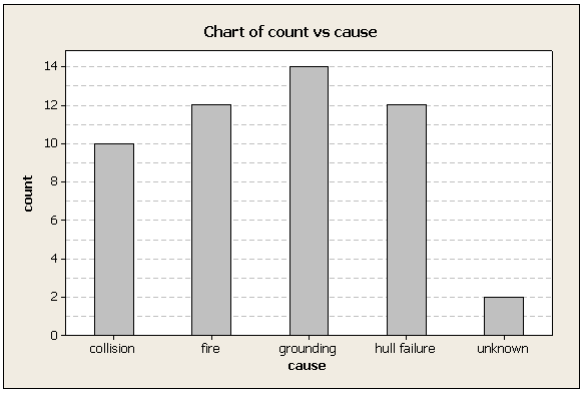
36. Researchers are studying the effect of diet on lab rats' ability to run a maze. There are 60 rats available, numbered 1–60. The researchers are studying two new diets plus a standard diet, and it is desired to have the same number of rats in each group. Below is an excerpt from Table B.

```
61081 29987 74578 34167
```

Use this excerpt from Table B to select the first two rats to get Diet 1. What are the numbers of the rats you selected?

(a) 61 and 34
(b) 8 and 12
(c) none of the other alternatives
(d) 61 and 8
(e) 6 and 1

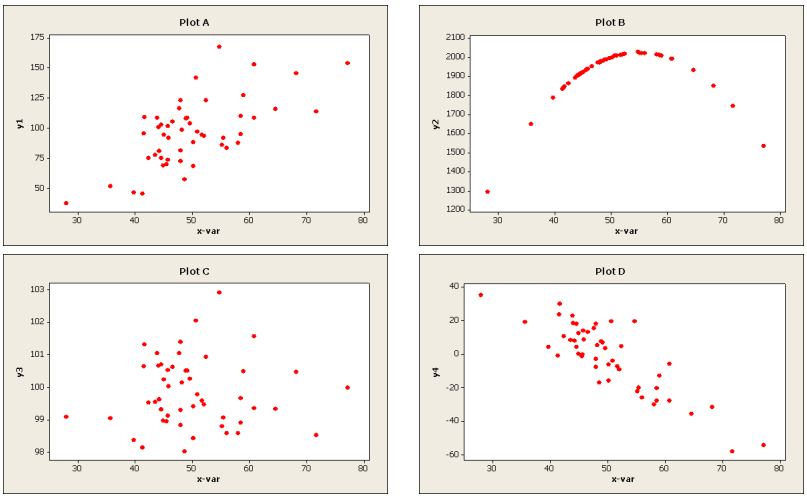
37. Designers of oil tankers want to improve the structural design to decrease the likelihood of an oil spillage. To understand the reasons for oil spillages, 50 major oil spills were analyzed. The reasons for the spillages are summarized in the bar chart below.



Approximately what **percentage** of oil spills were caused by collisions?

(a) 10
(b) 24
(c) 12
(d) 20
(e) 14

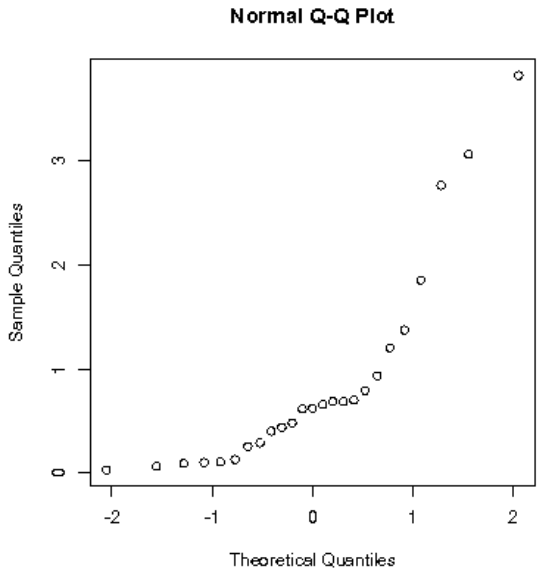
38. The four scatterplots below all show different correlations.



Which plot shows the highest positive correlation?

(a) Plot D
(b) Plot A
(c) Plot C
(d) None of the plots show a positive correlation.
(e) Plot B

39. A normal quantile plot is shown below.



What do you conclude from the plot?

(a) we should extrapolate the sample quantiles
(b) there is a positive association between the two variables
(c) the data is approximately normally distributed
(d) the data is skewed to the right

40. Among 8 subjects, 4 volunteered to drink alcohol while the remaining 4 became the control group (they were alcohol-free for the duration of the study). All the subjects then had to perform some driving tasks on a test track. The quality of each subject's driving was measured. Which of the following best describes the design?

    (a) good, response will reflect dissimilarity of groups

    (b) bad, should always block for volunteers

    (c) good, 4 volunteers matched with 4 control subjects

    (d) bad, response might reflect dissimilarity of groups