University of Toronto Scarborough
STAB22 Midterm Examination solutions

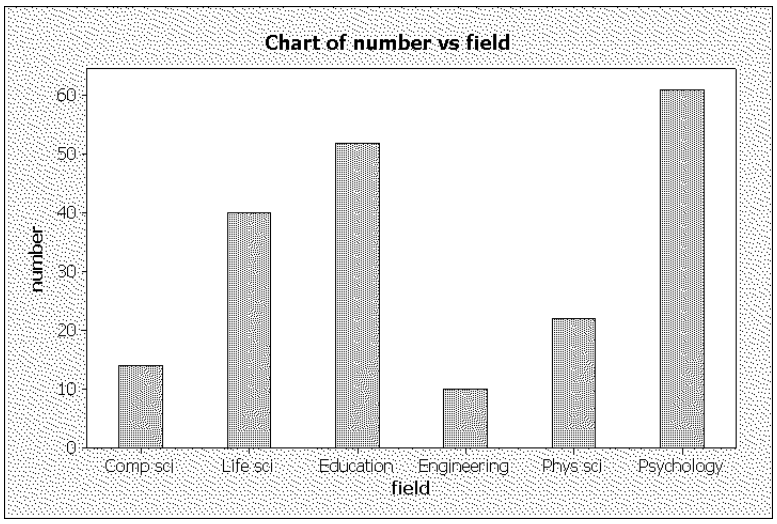
June 2007

This examination contains a mixture of short-answer and multiple choice questions. For the short-answer questions, write your answers in the spaces provided; for the multiple-choice questions, circle the best answer out of the alternatives given.

Marks for each question are as shown.

Correct answers for multiple-choice questions are shown by a *.

1. (5 marks) At a certain university in 1993, a record was kept of the number of women graduating with doctoral degrees in each of six different fields of study.



   (a) What is this kind of graph usually called?

      i. Histogram
      ii. * Bar chart (the variable "field of study" is categorical)
      iii. Stem-and-leaf plot
      iv. Pie chart

   (b) Which field of study had the most women graduating with doctoral degrees in 1993?

      Psychology

   (c) Approximately how many women graduated with doctoral degrees in life sciences in 1993?

      40 (or 39 if that's what you think it is)

2. (8 marks) A political scientist takes a large sample of registered voters, and measures a number of variables, as shown below. Are each of the variables listed below categorical or quantitative?

   (a) Gender

      i. * Categorical
      ii. Quantitative

   (b) Age

      i. Categorical
      ii. * Quantitative

   (c) Household income

      i. Categorical

ii. * Quantitative

(d) Party voted for at last election

i. * Categorical

ii. Quantitative

3. (5 marks) What is the median of the following set of numbers: 7, 6, 10, 9, 5?

(a) 7.4

(b) * 7 (remember to arrange the numbers in order first!)

(c) 10

(d) 6

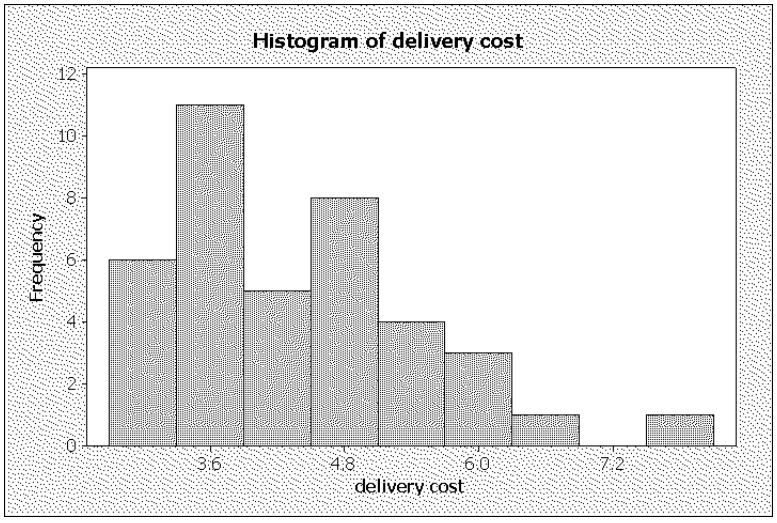
4. (7 marks) The speeds of 57 cars were measured (in km/h) on a city street. A numerical summary of the results is given below:

```
Descriptive Statistics: speed

Variable   N  N*   Mean  SE Mean  StDev
speed     57   0  45.53     1.64  12.38

Variable   Minimum     Q1  Median     Q3  Maximum
speed        25.60  36.80   43.20  51.20    83.20
```

Do a calculation to decide whether the highest recorded speed, 83.2 km/h, is an outlier in this data set. What do you conclude?

IQR is $51.20-36.80 = 14.4$; $1.5 \times$ IQR is 21.6; Q3 plus this is $51.2+21.6 = 72.8$. The maximum value is 83.2, bigger than this, so is an outlier by this criterion.

5. (3 marks) A delivery company recorded the delivery costs for 39 small packages it delivered one day last week. A histogram of the delivery costs is shown below.



(a) Describe the shape of this histogram.

Skewed to the right. (I want to see both "skewed" and "to the right", but I don't want you to fill the space with other stuff!)

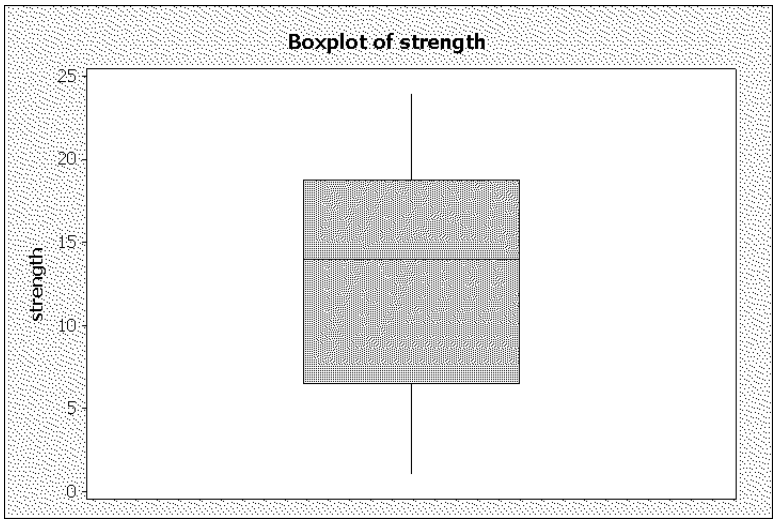
(b) A manager at the delivery company recorded the mean and median delivery cost, but unfortunately lost the paper on which the values were recorded. However, the manager can remember that one value was about \$4.00 and the other was about \$4.40. Which value is the mean and which is the median?

i. \$4.00 is the mean and \$4.40 is the median.

ii. * \$4.40 is the mean and \$4.00 is the median. (because mean is bigger than median when the distribution is skewed to the right)

iii. It's impossible to tell which is which.

6. (9 marks) All the third-graders at a certain elementary school were given a physical-fitness strength test. The test scores are shown in the boxplot below.



(a) What is the median test score, approximately?

14 (where the centre line of the box goes across)

(b) What is the interquartile range of test scores, approximately? Show your calculations.

Q1 is about 7, Q3 is about 19 (bottom and top of the box), so IQR is about $19 - 7 = 12$. The exact values are not important; the point is to get reasonable values for the quartiles and then find the IQR in the right way.

(c) Are there any unusually high or low test scores? If there are, indicate the unusual values on the boxplot.

No. (They would have been plotted separately as asterisks on the plot.)

7. (8 marks) A set of exam marks has mean 70, median 65, inter-quartile range 25 and SD 15 marks. It is decided to subtract 10 from all the marks. For the new set of marks,

   (a) what is the mean?

   (b) what is the median?

   (c) what is the inter-quartile range?

   (d) what is the SD?

   Mean and median are the original values minus 10 (60 and 55), IQR and SD are measures of spread so they are unchanged (25 and 15).

8. (2 marks) A set of data has quartiles 30 and 75, and median 40. What would you conclude about the shape of the data distribution?

   Skewed to the right (because the median is closer to Q1 than Q3).

9. (12 marks) For a particular group of adult males, the distribution of cholesterol readings is normal with mean 210 and SD 15. Use Table A to find the following, showing your calculation in each case:

   (a) The proportion of males in this group with cholesterol reading less than 240.

   For 240, $z = (240 - 210)/15 = 2$; proportion less (from table) is 0.9772.

   (b) The proportion of males in this group with cholesterol readings between 200 and 240.

   For 200, $z = (200 - 210)/15 = -0.67$; proportion less (from table)=0.2514. Proportion less than 240 (from (a)) is 0.9772, so proportion between is $0.9772 - 0.2514 = 0.7258$.

   (c) The cholesterol reading that 20% of males in this group are higher than.

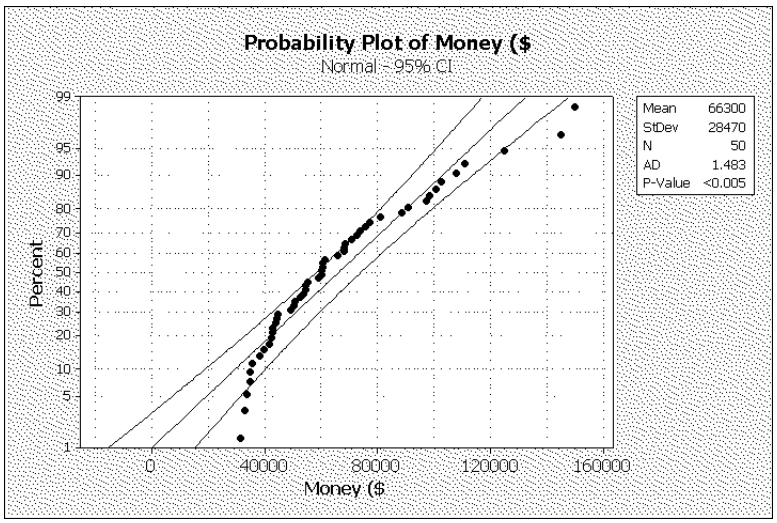
   This was admittedly not stated as clearly as it might have been. Another way to ask it is: "Suppose x is a cholesterol reading, and 20% of males have higher cholesterol reading than x. What is x?"

   Table gives you less than, so 20% higher means 80% lower. Look up 0.8000 backwards in the body of the table, to get $z = 0.84$ (closest). Then the required level is $(0.84)(15) + 210 = 222.6$. (Or solve $0.84 = (x - 210)/15$ for x if you prefer that way.)

   (d) The first quartile of cholesterol readings for males in this group.

   Like (c), but now we want 25% lower (that's what the first quartile is). Looking up 0.2500 in the body of the table gives $z = -0.67$ approximately. Thus the first quartile is $(-0.67)(15) + 210 = 200$. (Compare the first part of (b).)

10. (3 marks) The 1998 winnings of 50 professional golfers were recorded. A normal quantile plot of the distribution is shown below.



From the plot, do you conclude that these data are described well by a normal distribution? In one sentence, explain your conclusion.

No. The normal quantile plot should show a straight line, but it shows a curve. (A minimal answer would be "No. It's curved"; you certainly don't need to say more than I did above.)

11. (3 marks) The correlation between two variables is very close to 0.

(a) Does it follow that there is no relationship of any kind between these two variables?

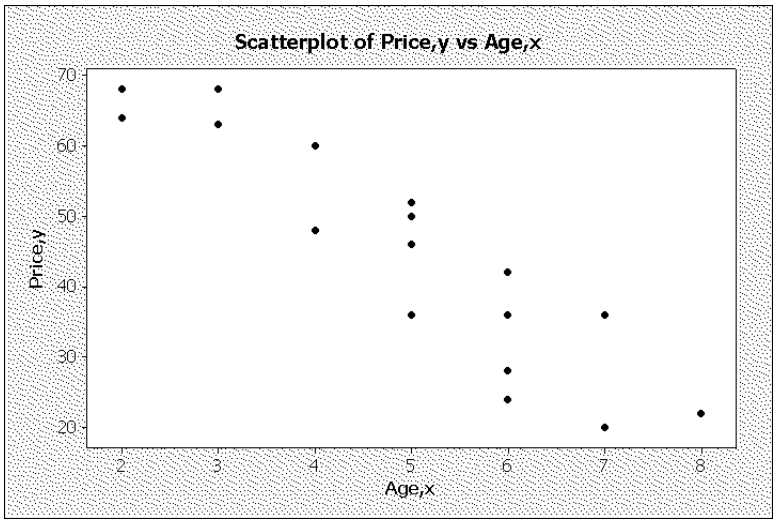
i. yes, there cannot be any relationship

ii. * no, there could be some kind of relationship

(b) If you answered "no" to part (a), what kind of relationship could there be? Answer in one sentence only.

It could be a curved relationship. (A curved trend that goes down then up can have a very small correlation, even if the trend itself is very clear.)

12. (3 marks) The scatterplot below shows, for 19 used foreign compact cars, the age of the car, and the asking price (in hundreds of dollars).



Choose the best value for the correlation between these two variables from the list below.

(a) * -0.9

(b) -0.4

(c) 0

(d) 0.4

(e) 0.9

It must be a negative correlation, and a correlation of -0.4 would look a lot more scattered. So -0.9 it is.

13. (3 marks) Water flowing across farmland washes away soil. Researchers released water across a test area at different flow rates and measured how much soil was washed away (amount of eroded soil). In this case, which is the explanatory variable and which is the response?

(a) Eroded soil is the explanatory variable and flow rate is the response.

(b) * Eroded soil is the response and flow rate is the explanatory variable. (Flow of water leads to soil being eroded.)

(c) There is no explanatory variable or response in this situation.

14. (7 marks) A study was made of some popular fast-food items. In particular, a regression analysis was done for predicting calorie content from the amount of fat in a food item. Some regression output from Minitab is given below:

```
Regression Analysis: Calories versus Fat

The regression equation is
Calories = 242 + 7.35 Fat


Predictor     Coef  SE Coef      T      P
Constant   242.03    68.44   3.54  0.003
Fat         7.353    2.860   2.57  0.021


S = 152.030   R-Sq = 29.2%   R-Sq(adj) = 24.8%
```

(a) What is the value of the slope of this regression line?

7.35.

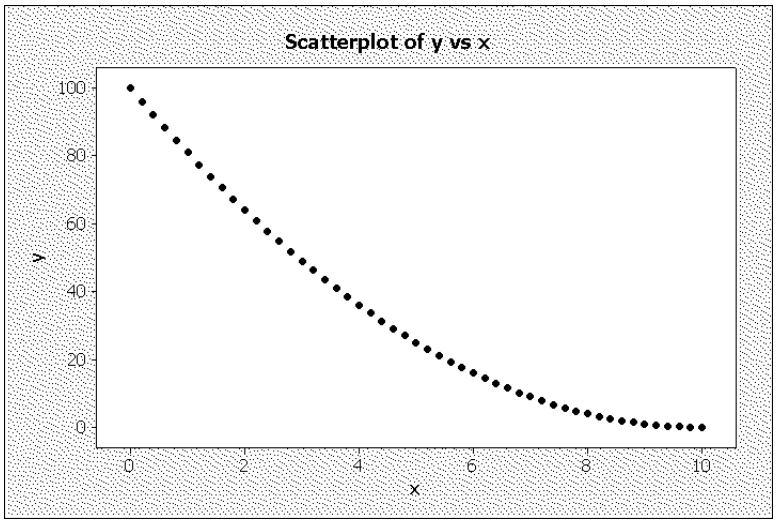
(b) Which is the best interpretation of the slope of this regression line?

i. The calorie content of an item that has no fat
ii. The fat content of an item that has no calories
iii. * The increase in calories associated with a one-unit increase in fat
iv. The increase in fat associated with a one-calorie increase.

(c) Predict the calorie content of an item containing 15 units of fat.

$242 + (7.35)(15) = 352.25$. The important part is showing that you knew the right calculation to do, though getting it right is nice!

15. (3 marks) The plot below shows a scatterplot for two variables x and y.



Which of the statements below about the correlation between x and y is most accurate?

(a) The correlation is +1 because there is a perfect positive association.

(b) The correlation is a little less than +1 because the trend is slightly curved.

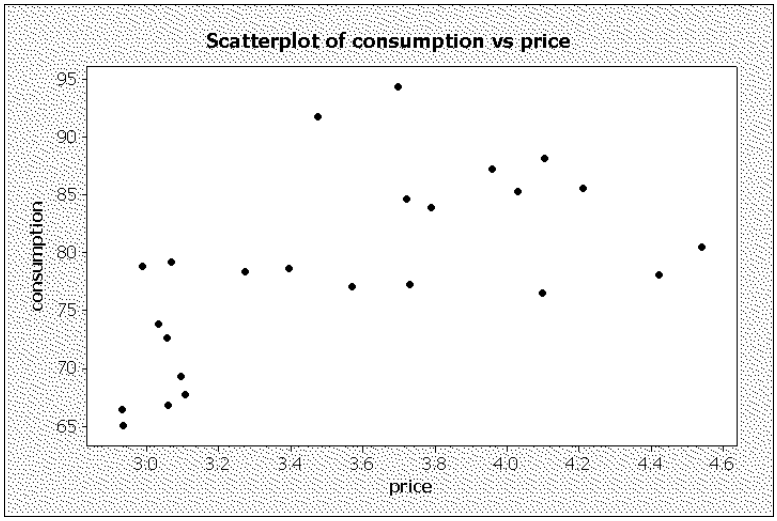
(c) The correlation is 0 because the trend is not linear.

(d) * The correlation is a little greater than −1 because the trend is slightly curved. (There is clearly some kind of downward trend, so the correlation ought to be negative; having a non-linear relationship will tend to make the correlation closer to 0 than it would otherwise have been, but won't make it *exactly* 0 except in special cases (compare q. 11)).

(e) The correlation is −1 because there is a perfect negative association.

16. (7 marks) A study was made of the price (dollars per pound) and consumption (pounds per person per year) of beef in the United States for each year 1970–1993. The prices were adjusted to 1993 dollars, because everything got more expensive in this time period.

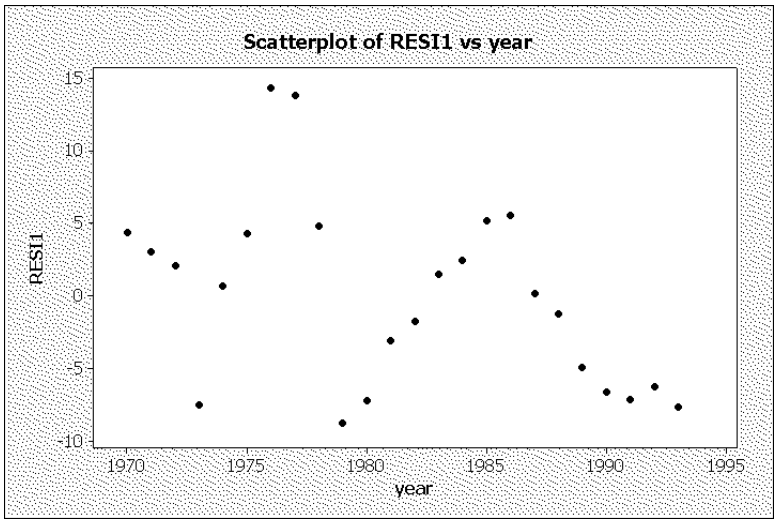
(a) Economic theory says that if the price of an item is higher, consumption of that item will be less. A scatter plot for predicting consumption from price is shown below.



Does this scatterplot support the economic theory? Explain briefly (1 sentence).

No. There should be a downward trend, but here the trend is nonexistent or slightly upwards.

(b) A regression was done for predicting consumption from price. The residuals were plotted against the year, as shown below.



Do you see any problems in this plot? Explain briefly (1 sentence).

Yes: I see a steady increase in residuals between 1979 and 1986, and a steady decrease after that. (The residual plot should show a random, formless scatter of points, so any patterns you can describe in a few words are a problem. I think this is a bigger problem than the two large residuals in 1976 and 1977.)

17. (7 marks) Breast cancer that is detected in its early stages can be treated. In the past, the preferred treatment was mastectomy (removal of the breast); now, it is usual to remove the tumour, and have the patient undergo radiation. A medical team compares the survival times after surgery of all women who have had either treatment.

    (a) What is the explanatory variable here?

    Treatment: that is, mastectomy or removal of tumour plus radiation.

    (b) What is the response variable here?

    Survival time.

    (c) Is this study (choose one):

    i. * an observational study

    ii. an experiment?

    (d) Will the medical team be able to conclude that the better of the two treatments causes a longer average survival time (circle your preferred answer)?

    i. yes

    ii. * no (an observational study won't let you conclude cause and effect, not at least without extra work. If you thought it was an experiment, I gave you a point for answering "yes", on the grounds that an experiment *can* let you infer cause and effect.

18. (5 marks) A statistical experiment is to be done to compare three treatments. 9 subjects are available, named: Alomar, Bikalis, Cranston, Durr, George, Han, Imrani,

Lawless, Zhang. Three equal-sized groups of subjects will be used. The numbers 1–9 are randomly rearranged as follows:

```
6    2    9    1    8    5    3    4    7
```

Name the subjects that will receive the 2nd treatment, and explain how you came to your conclusion.

The easiest solution is to number the subjects 1–9 (they are already in alphabetical order) and note that there will be 3 groups of 3 subjects each. Using the random rearrangement given, subjects 6, 2 and 9 form group 1, and 1, 8 and 5 (Alomar, Lawless, George) form group 2.

Or you can assign subject 6 to treatment 1, 2 to 2, 9 to 3, 1 to 1, 8 to 2 and so on (so Bikalis, Lawless and Durr form group 2).

Beware of randomizing twice so that you end up un-randomizing: I saw some answers where subjects 4, 5 and 6 on the list ended up in group 2, which doesn't look very random!