University of Toronto Scarborough
STAB22 Midterm Examination

November 1, 2014

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 18 numbered pages; before you start, check to see that you have all the pages. There are statistical tables at the back, along with two blank pages for rough work (these will not be graded). You may take away this question paper after the exam, which means that you may detach any of its pages that you wish.

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").

Only your Scantron sheet will be graded. It is *your* responsibility to complete the Scantron sheet during the exam time. You will *not* be allowed extra time to do this.

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 and detach the signature sheet from the front of the exam, but *sign it only when the invigilator collects it.* The signature sheet shows that you were present at the exam.

1. The red blood cell count of a healthy person was measured on each of 15 days. The number recorded is in units of million cells per microlitre. What would be an appropriate graph to display the red blood cell counts?

   (a) Pie chart

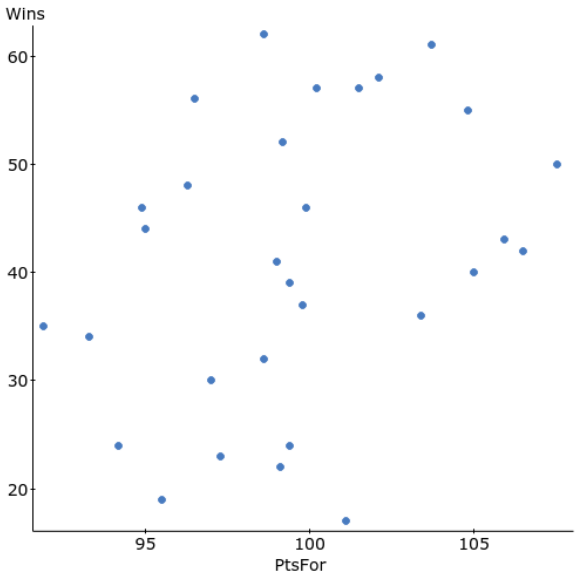
   (b) Bar chart

   (c) * Stemplot

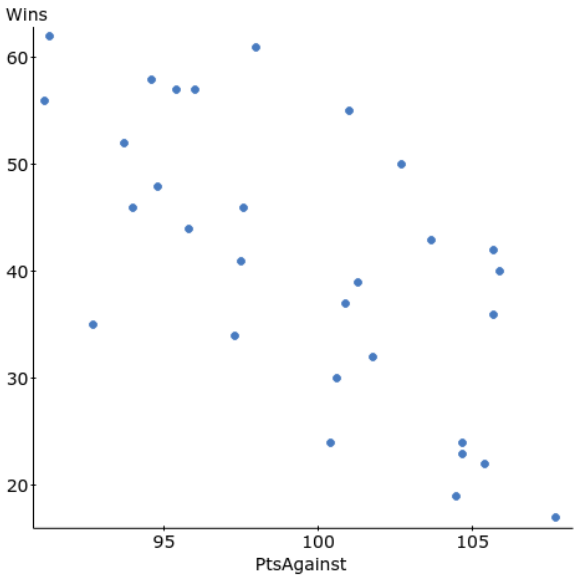
   (d) Scatterplot

2. Sports fans often say that what is more important for winning games is not scoring lots of points (goals, whatever), but stopping the other team from scoring lots of points. You will hear things like "defence wins championships". To assess whether that was true for professional basketball, a fan took results from a season's worth of NBA (National Basketball Association) games and recorded, for each team, the number of wins and losses (each team played 82 games), along with the average number of points per game scored by the team and the average number of points per game conceded by the team (that is, how many points per game the team's opponents scored).

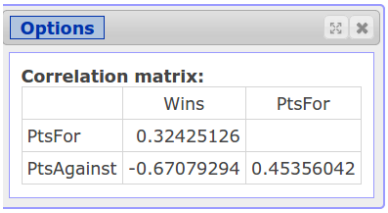
   A scatterplot of wins by points scored per game is shown below:



   Next, a scatterplot of wins by points allowed per game:



   and finally a table of correlations between these variables:



Which is a better predictor of games won: points scored, or points allowed?

(a) There is no way to choose based on this information.

(b) Points scored, because the correlation is positive.

(c) Points scored, because the relationship is weaker.

(d) Points allowed, because the correlation is negative.

(e) * Points allowed, because the relationship is stronger.

3. A variable has median 20, first quartile 17 and third quartile 24. The lowest few values are 3, 4, 7, 8 and the highest few values are 28, 31, 32. We are going to draw a boxplot for these data. Use this information for this question and the next one.

   How far should the upper whisker extend?

   (a) 40.0

   (b) 24.0

   (c) * 32.0

   (d) 34.5

   (e) 36.5

4. In the boxplot of the previous question, how far should the lower whisker of the boxplot extend?

   (a) * 7.0

   (b) 3.0

   (c) 6.5

   (d) 8.0

   (e) 4.0

5. The heights of men from the Netherlands ("Dutch men") have a normal distribution with mean 70 inches and standard deviation 4.5 inches. Use this information for this question and the next 3 questions.

   What fraction of Dutch men are shorter than 62 inches?

   (a) 0.20
   (b) 0.08
   (c) 0.12
   (d) * 0.04
   (e) 0.16

6. Using the information in Question 5, what fraction of Dutch men are between 72 and 77 inches tall?

   (a) 0.67
   (b) 0.06
   (c) 0.12
   (d) 0.45
   (e) * 0.27

7. Using the information in Question 5, and according to the 68–95–99.7 rule, 95% of Dutch men are between which heights, in inches?

   (a) * 61.0 and 79.0
   (b) 58.8 and 77.7
   (c) 65.5 and 74.5
   (d) 56.5 and 83.5

8. Using the information in Question 5, the tallest 10% of Dutch men are taller than what height, in inches?

   (a) 74
   (b) 72
   (c) 70
   (d) 64
   (e) * 76

9. What is the **median** of the following set of numbers?

$$15, 7, 12, 16, 8, 5, 11$$

(a) 16

(b) 7.5

(c) 13.5

(d) 10.5

(e) * 11

10. Wildlife ecologists monitored the breeding habitats of mountain plovers (birds) for three years. In each year, they noted how many of the plovers' nests were in agricultural fields, prairie dog habitat, or grassland. The results are shown below. Use this information for this question and the following 2 questions.

Location	2004	2005	2006	Total
Agricultural field	21	29	26	76
Prairie dog habitat	17	38	12	67
Grassland	5	6	9	20
Total	43	73	47	163

What percentage of all mountain plovers nested in grassland in 2005?

(a) 6

(b) * 4

(c) 10

(d) 12

(e) 15

11. Using the information in Question 10, what is the marginal percentage of mountain plovers that nested in prairie dog habitat?

(a) 50

(b) * 40

(c) 30

(d) 25

(e) 45

12. Using the information in Question 10, out of the nests recorded in 2006, what is the conditional percentage of mountain plovers that nested in agricultural fields?

    (a) 40

    (b) * 55

    (c) 35

    (d) 50

    (e) 45

13. Two hospitals conducted a study of whether patients lived or died. The patients were also classified as being in "good condition" or "poor condition". The results are shown below.

Patients in good condition:

	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	800
% died	1.0	1.3

Patients in poor condition:

	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1600	200
% died	3.5	4.0

All patients regardless of condition:

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800
% died	3.0	2.0

Why is it that more of hospital A's patients died overall, even though fewer of hospital A's patients in good condition and fewer of hospital A's patients in poor condition died, compared to hospital B?

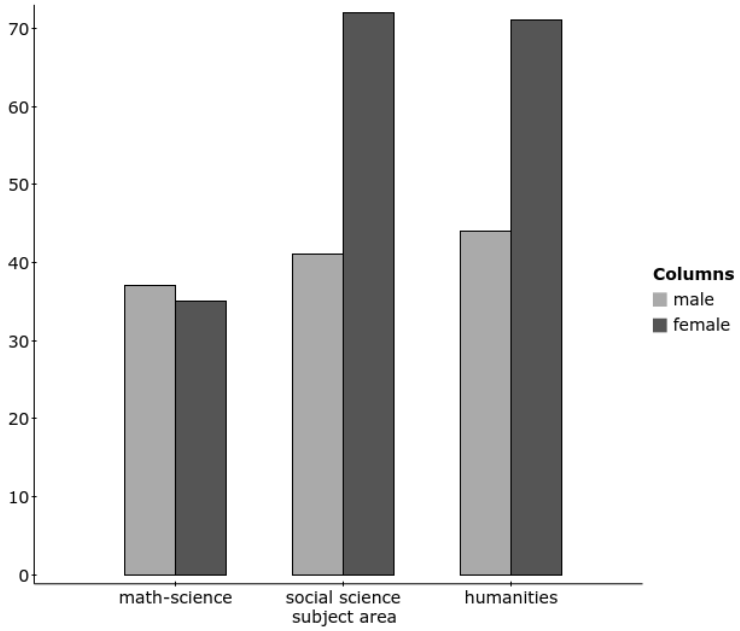
(a) This kind of result with the percentages can never happen.

(b) There was a mistake in the numbers.

(c) * More of the poor-condition patients went to hospital A.

(d) More of the good-condition patients went to hospital A.

14. 300 students were identified as male or female, and each one was asked whether they preferred taking courses from math and science, social science or humanities. A bar chart is shown below.



Use this graph to answer this question and the next two questions.

About how many males prefer math-science courses?

(a) 41

(b) 35

(c) 45

(d) * 37

(e) 70

15. Look at the graph in Question 14. In terms of social science and humanities courses, which one of the following is true?

(a) More females prefer taking social science courses, but more males prefer taking humanities courses.

(b) * More females than males prefer taking courses in both subject areas.

(c) The numbers of males and females preferring to take courses in both subject areas are about the same.

(d) More males prefer taking social science courses, but more females prefer taking humanities courses.

(e) Fewer females than males prefer taking courses in both subject areas.

16. Look at the graph in Question 14. Which of the following best describes the difference between these male and female students?

    (a) There is no real difference between the kinds of courses that male and female students prefer.

    (b) * Male students are more likely to prefer math and science courses than female students are.

    (c) About the same proportion of female as male students prefer math and science courses.

    (d) Male students are more likely to prefer social science courses than female students are.

    (e) Male students are more likely to prefer humanities courses than female students are.

17. There are 26 tutorials in an introductory statistics class. A random sample of 6 tutorials is taken, and a random sample of 5 students is taken from each sampled tutorial. What kind of sampling method is this?

    (a) stratified sample

    (b) * multistage sample

    (c) simple random sample

    (d) voluntary response sample

18. A student organization has 55 members. Five students are selected using a random number table. What kind of sampling method is this?

    (a) multistage sample

    (b) * simple random sample

    (c) stratified sample

    (d) voluntary response sample

19. In an introductory psychology class, separate samples of male and female students are selected to receive a one-week additional instruction method. What kind of sampling method is this?

    (a) * stratified sample

    (b) multistage sample

    (c) voluntary response sample

    (d) simple random sample

20. Here are the GPAs of 16 graduates of a postgraduate program at the University of Toronto in 2013, given as a stem-and-leaf plot.

```
Variable: var1

Decimal point is at the colon.
Leaf unit = 0.1

0 : 7
1 : 899
2 : 56
3 : 12567889
4 : 00
```

Use this information to answer this question and the next two questions.

Comment on the skewness of the data.

(a) * left skewed

(b) symmetric

(c) right skewed

(d) cannot tell

21. Using the information in Question 20, obtain the mean and median (in that order below) for these data.

(a) * 3 and 3.35

(b) 30 and 33.5

(c) 3.35 and 3

(d) 33 and 35

22. Using the information in Question 20, what are the best measures of centre and spread for these data?

(a) mean and standard deviation

(b) mean and interquartile range

(c) median and variance

(d) *median and interquartile range

(e) median and standard deviation

23. Two polls were conducted on the question "Do you think bus service in Toronto should be improved?". One poll was conducted in January and the other in May. The results are shown below.

Opinion	January	May	Total
Agree	723	779	1502
Disagree	277	321	598
Total	1000	1100	2000

Suppose that the polling organization wants to compare the proportions of people agreeing in the two different polls. Which percentages should the organization calculate?

(a) conditional percents.

(b) * column percents

(c) row percents

(d) joint percents

24. According to the US Environmental Protection Organization, the composition of the 254 million tons of solid municipal waste in 2007 was as shown in the pie chart below. The top right pie slice is labelled "other".

Which of the following statements is most clearly true for this pie chart?

(a) Other waste forms less than a quarter of all waste.

(b) Plastics form about a quarter of all waste.

(c) Food waste and yard waste together form more than a third of all waste.

(d) * Paper waste is about a third of all waste.

(e) There is less yard waste than food waste.

25. A fair octahedral (8-sided) die is rolled once. Use this information to answer this question and the next two questions.

    What is the probability of getting a number which is both an even number and strictly greater than 5?

    (a) 0.125

    (b) * 0.250

    (c) 0.500

    (d) 0.375

26. Using the information in Question 25, what is the probability of getting a number which is either an odd number or strictly less than 3?

    (a) * 0.625

    (b) 0.875

    (c) 0.750

    (d) 0.500

27. The fair 8-sided die described in Question 25 is rolled once. Let A be the event "an odd number is rolled", let B be the event "a number strictly less than 2 is rolled", and let C be the event "a number less than or equal to 2 is rolled". Four statements about events A, B and C are given below. Only one of them is **true**. Which one?

    (a) Events A and B^C are disjoint.

    (b) Events A and B are disjoint.

    (c) * Events A^C and B are disjoint.

    (d) Events B^C and C are disjoint.

28. A study was conducted to investigate how people with a hearing impairment communicate with their conversational partners. Each of 13 hearing-impaired subjects, all fitted with a cochlear implant, participated in a structured communication interaction with a familiar conversational partner (a family member) and with an unfamiliar conversational partner (who was instructed not to take the initiative to repair breakdowns in the communication). The total number of words used by the subjects in each of the two conversations was recoded. The purpose of the study was to predict number of words in conversation with an unfamiliar partner (y) from number of words in conversation with a familiar partner (x). The regression equation is $\hat{y} = 21.13 + 0.62x$.

    The coefficient of determination (R-squared) is 58.4%. This means:

    (a) A conversation with a familiar partner accounts for 58.4% of any conversation.

    (b) * Number of words in conversation with a familiar partner accounts for 58.4% of variation in number of words in conversation with an unfamiliar partner.

    (c) Number of words in conversation with an unfamiliar partner accounts for 58.4% of variation in number of words in conversation with a familiar partner.

    (d) 58.4% of conversation remained unclear in this study.

29. Refer to the information in question 28. The slope of the regression line is 0.62. This means that:

    (a) * For each additional word in conversation with a familiar partner, the mean number of words in conversation with an unfamiliar partner is estimated to increase by 0.62

    (b) For each additional word in two conversations (familiar or unfamiliar partner), the mean number of words spoken is estimated to increase by 0.62

    (c) The estimated mean number of words spoken in any conversation is 0.62

    (d) For each additional word in conversation with an unfamiliar partner, the mean number of words in conversation with a familiar partner is estimated to increase by 0.62

30. Refer to the information in question 28. $x = 0$ is in the range of the observed data. How can we interpret the number 21.13 in the regression equation?

    (a) * The mean number of words in a conversation with an unfamiliar partner is estimated to be 21.13 when the number of words in conversation with a familiar partner is 0

    (b) The mean number of words in a conversation with a familiar partner is estimated to be 21.13 when the number of words in conversation with an unfamiliar partner is 0

    (c) The mean number of words in a conversation with an unfamiliar partner is estimated to be 0 when the number of words in conversation with a familiar partner is 0

    (d) There is no meaningful interpretation.

31. Refer to the information in question 28. For a subject in this study, the number of words in conversation with a familiar partner is 65, and the number of words in conversation with an unfamiliar partner is 47. The residual (comparing the observed and the predicted value) is:

    (a) 61.43

    (b) * –14.43

    (c) 3.57

    (d) 14.43

32. In the general population, 6% of all people have the O-negative blood type. These people are called "universal donors" because any patient can receive a donation of O-negative blood. 4 people appear at random to give blood. What is the probability that at least one of them has the O-negative blood type?

    (a) 0.78

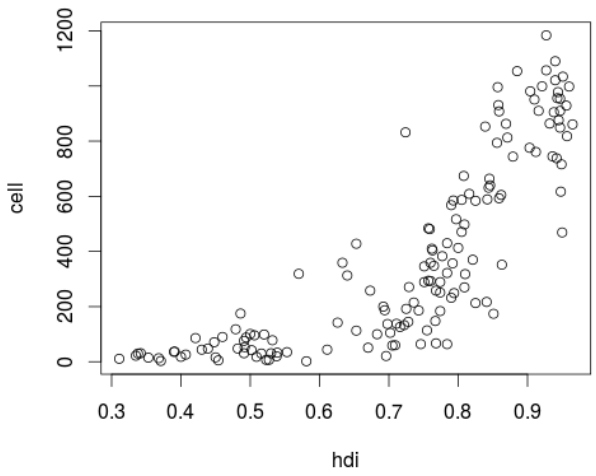
    (b) 0.50

    (c) 0.06

    (d) * 0.22

    (e) 0.94

33. The Human Development Index is a measure, for each country, of whether it has the attributes of a developed country like Canada. It is a value between 0 and 1, with a higher value being better. Do more highly-developed countries have a higher usage of cellphones? For a recent data set of Human Development Index values and cell phone use by country, the scatterplot shown below was obtained:



What re-expression is likely to be best?

(a) * Logarithm of `cell`.

(b) Square root of `hdi`.

(c) Square `cell`.

(d) Use `cell` as it is.

34. Suppose a couple of your friends go to a particular restaurant in Toronto on either Monday or Friday each week, not both. 26% of the time they go on Monday. On Mondays, the probability of receiving good service is 0.72. On Fridays, probability of receiving good service is only 0.13. Use this information for this question and the next two questions.

What is the probability that they went to that restaurant on Monday and received good service?

(a) 0.72

(b) * 0.19

(c) 0.26

(d) 1.00

35. Refer to the information in Question 34. What is the probability that they received good service at that restaurant last week?

    (a) 0.85

    (b) 0.72

    (c) * 0.28

    (d) 1.00

36. Refer to the information in Question 34. Suppose that you don't know which day they went last week, but they tell you they received good service. What is the probability that they went on Monday?

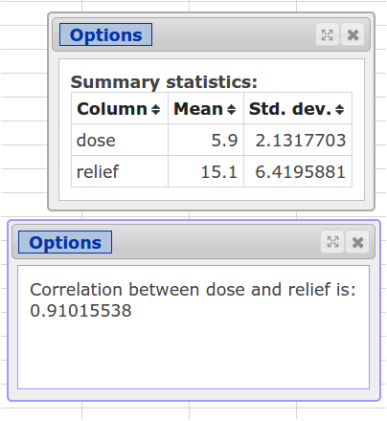
    (a) 0.26

    (b) * 0.66

    (c) 0.72

    (d) 1.00

37. One of the stages in the development of a new drug is the determination of the appropriate dose of the drug. In an experiment, ten patients were given various doses of an allergy-relieving drug. The patients were asked to report back when the allergy symptoms returned, and the number of hours of allergy relief was recorded. Summary statistics for the data are shown below.



What is the **intercept** of the regression line for predicting hours of allergy relief from dose?

(a) * –1.1
(b) 0.9
(c) 2.7
(d) –0.5
(e) 1.1

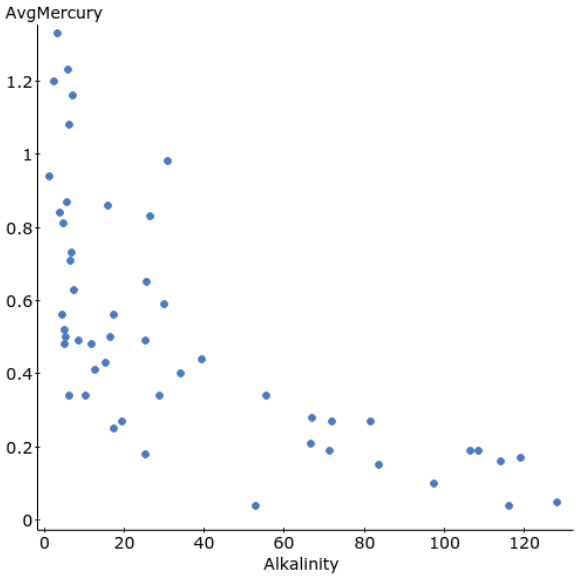
38. In a regression, the response variable has been re-expressed using its square root. The resulting regression is

$$\widehat{\sqrt{y}} = 12 + 1.3x.$$

When $x = 9$, what is the predicted value of y?

(a) * 560
(b) 120
(c) 24
(d) 4.8
(e) 14,000

39. Data were collected on 53 lakes in Florida. Interest was in `Alkalinity` (concentration of calcium carbonate in milligrams per litre) and `AvgMercury` (the average mercury level of a sample of fish collecctd from the lake). A scatterplot of mercury level (response) vs. alkalinity (explanatory) is shown below.



Use this information for this question and the next one.

Why might you have doubts about fitting a regression to these data?

(a) Because there is a very weak relationship between the two variables.

(b) * Because the relationship is a curve, not a straight line.

(c) Because the correlation is negative.

(d) Because there are clear outliers.

40. In Question 39, a plot was shown of the relationship between mercury in fish and alkalinity of the lake water. What might you try, if anything, to make this relationship more suitable for regression?

(a) Instead of using `AvgMercury`, predict the square of `AvgMercury` from `Alkalinity`.

(b) The shape of the relationship is such that a transformation (re-expression) will not help.

(c) Instead of using `Alkalinity`, predict `AvgMercury` from the square root of `Alkalinity`.

(d) * Make a transformation (re-expression) of `AvgMercury`.

(e) Make a transformation (re-expression) of `Alkalinity`.