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 16 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.

Whichever version of the exam you wrote, all your questions are in here somewhere.

1. Look at the experiment depicted below.

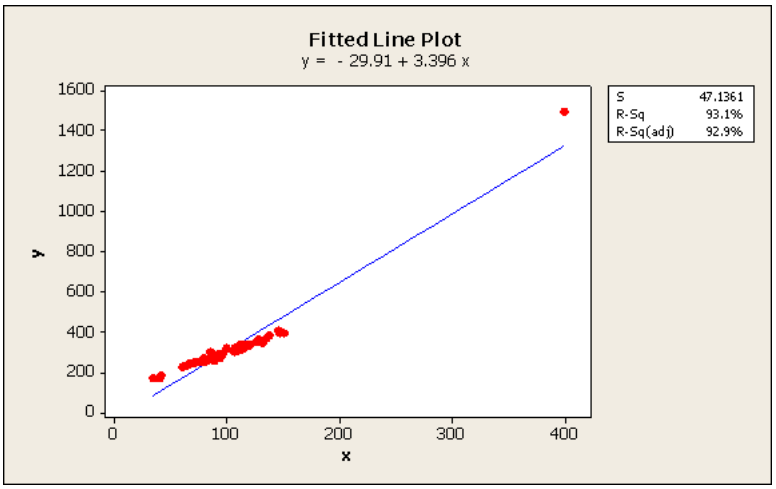
```
                          Treatment 1
                          Mike, Joe, Mary, Jill ----- record results
                        /
Subjects:              /
Mike, Joe, Mary, Jill \
                        \
                          Treatment 2
                          Mike, Joe, Mary, Jill -----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

   Who is Mike matched with?

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

   All four people do both treatments, so they are matched in some way. Under treatments 1 and 2, they are listed in the same order, so they are matched with themselves (it is the kind of experiment where you get two measurements from each person, like a "before" and "after"). So Mike is matched with Mike.

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

   The point over on the right is, in x terms, a long way from the other points, so it is likely to be influential on the regression line. This is confirmed if you look at the picture more carefully: a line going through the cloud of points on the left would be less steep than the line shown, so the line's slope is being dragged upwards by the isolated observation on the right.

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

   Ask yourself what would be the best line without the point on the right: one with a smaller slope. That offers (a) as an option. The correlation might go up or down; sometimes removing an influential point makes it bigger, sometimes smaller. So (c) isn't necessarily true.

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

   My last 10 journeys are only a sample of "all possible journeys", so this is a statistic. If I were talking about all journeys, then it would be a parameter. A sample is the *process* producing the mean, not the value that is (or might be) produced.

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).

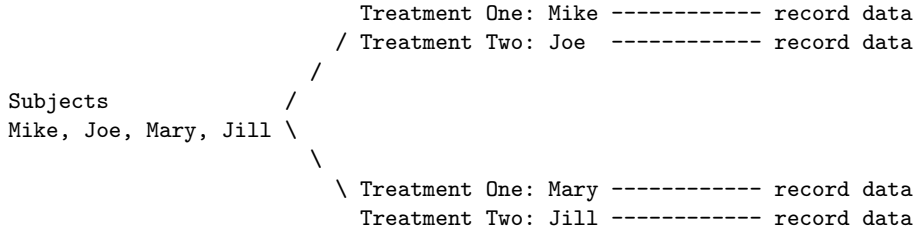
   Suppose the dairy inspectorate's farm sampling procedure is as follows: First, randomly select a sample of Ontario counties. Then, within each selected county, take a simple random sample of dairy farms. Then, visit each of the sampled dairy farms.

   What kind of sampling procedure is this?

   (a) * Multistage sample
   (b) Stratified sample
   (c) Systematic sample
   (d) Simple random sample
   (e) Voluntary-response sample

   What makes it multistage is the selection of counties first, and then farms from within the selected counties. (Compare making a list of all dairy farms in Ontario, and then selecting from that: this would be a simple random sample.)

6. Look at the experiment depicted below.



Which factor was blocked?

(a) subjects

(b) treatment two

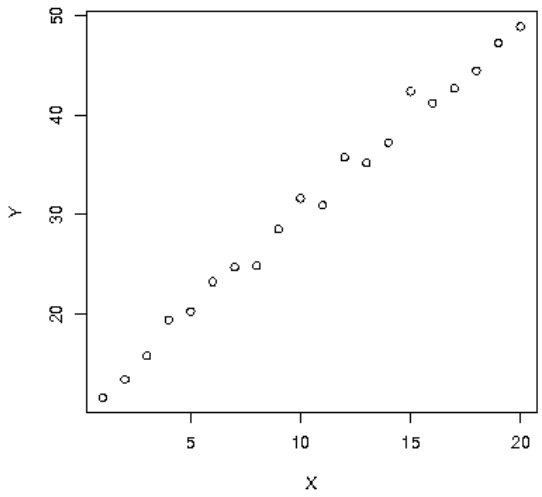
(c) recorded data

(d) * gender

(e) treatment one

There are two groups of people, in both of which one person receives treatment 1 and one receives treatment 2. The upper group of people is Mike and Joe, both males, and the lower group is Mary and Jill, both females, so what makes the two groups different is gender.

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

The correlation is positive, and too strong to be 0.55 (which would look much less clear as an upward trend).

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

Plug 10 into the regression equation: $\hat{y} = -39.5 + 15.5(10)$, which is about 115.

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.

This is what R-squared tells you, and the R-squared is "quite high" rather than "low". The slope doesn't tell us about accuracy of prediction, and we don't know that we are extrapolating (maybe we are, but "must" is too strong).

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

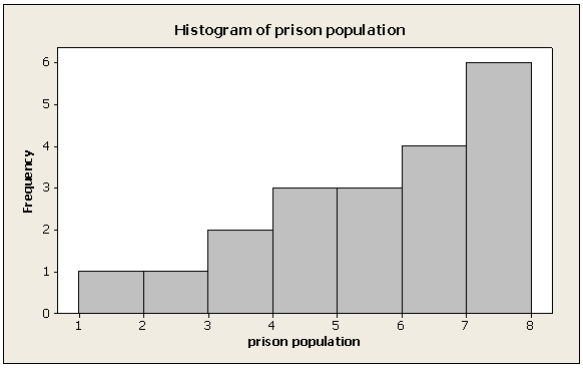
(c) 500

(d) 900

(e) * -900

Find the prediction first, and then see how the prediction differs from the observed value. Prediction is $\hat{y} = -39.5 + 15.5(80) = 1200$ days; residual is $280 - 1200 = -920$ days. Humans are very unlike the animals in this data set.

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

Pretty much a classical left skew.

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

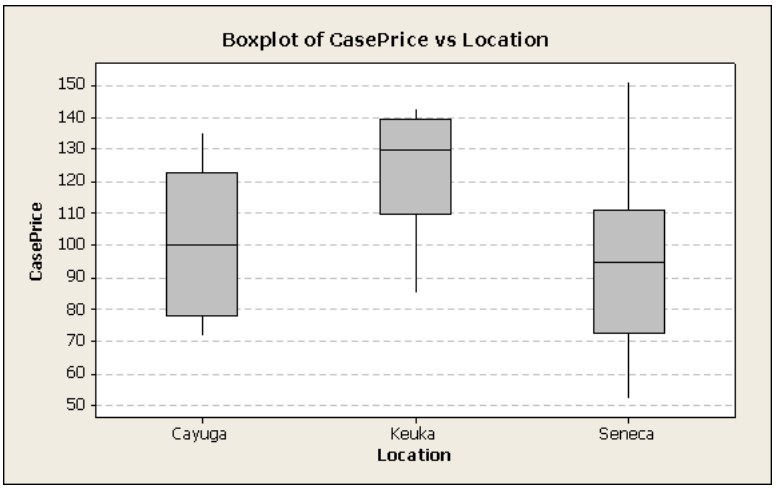
(c) 7.4

(d) * 6.1

(e) 3.5

Because of the left skew, or the long left tail, the mean will be pulled downwards further than the median is, so the median will be *larger* than the mean. Only 6.1 and 7.4 are larger than the mean, and 7.4 is too high to be the median (way more than 50% of the data values are less than 7.4).

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

Look for the highest bar across the *middle* of the rectangle in the boxplots.

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

Keuka's IQR is smallest (eg. by looking at it). Seneca's is about $110 - 72 = 38$, and Cayuga's is about $122 - 78 = 44$. Whatever precise values you have, you should be able to figure out that Seneca's IQR is a little less than 40, and Cayuga's is a little more.

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

    Look at the upper and lower parts of the rectangle, and the upper and lower whiskers (there are no outliers here). For Cayuga, the upper and lower parts of the rectangle are about the same, and the whiskers are about the same length: more or less symmetric. Seneca appears if anything right-skewed (look at the whiskers), while Keuka has a larger lower part to the box and a clearly longer lower whisker, so this is the one to pick.

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

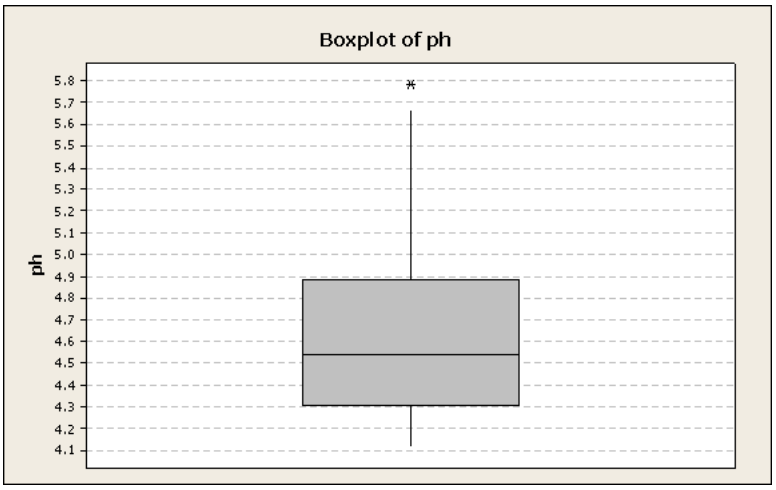
    The variable in question (first class/second class/third class/crew) is categorical, so a pie chart or bar chart are suitable. For judging fractions of a whole, you'd prefer a pie chart.

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

    By the same reasoning as above, we have a pie chart and a bar chart to choose from. Bsr charts are better at comparing the number of individuals in different categories.

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

Boxplots show median and quartiles, not the mean and SD. So if you want the mean, you need something other than a boxplot.

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

Top of the rectangle is about 4.9, bottom around 4.3, so the IQR is about $4.9 - 4.3 = 0.6$, and none of the other alternatives are anywhere close.

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

The one individually plotted observation with pH near 5.8.

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.

The upper part of the rectangle is bigger, the upper whisker is bigger, and there is an outlier at the upper end. (Even if "symmetric" had been a reasonable answer, a boxplot wouldn't have told you whether a normal distribution was a reasonable description of the shape.)

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%

4.9 is close to Q3, so there should be about a quarter of the data values above it. Q3 is close enough to 4.9 that the percentage wouldn't be as low as 10 or as high as 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

A stratified sample is two simple random samples side by side. According to the hint, you select your simple random sample of 2 males first, and then select the simple random sample of 2 females. Knowing this much eliminates all but (c) and (d): an overall simple random sample could have more or less than 2 females, but a stratified sample must have exactly 2. Since we are selecting from 6 people (both for males amd for females), use the random digits singly (in ones): for males, 2 (OK), 7 (too big), 2 (repeat), 6 (OK), so males 2 and 6 (Mike and Ajay). Then continue from where you left off to sample the females: 0 (no good), 9 (too big), 2 (OK since we haven't selected *female* number 2 before), 1 (OK). So we select females 2 and 1, Megan and Shelley. (Note that we had to select the males first, because we didn't know how many of the random digits we were going to need.)

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

Correlation is only meaningful for straight-line relationships, amd the relationship between a variable and its square is by definition not linear (it is a parabola, if you know about these things).

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

The median is between the 25th and 26th value (up from the bottom or down from the top). Working from the bottom, the median is between the first and second value on the line beginning with `(9)`. Since the leaves are 10s (and therefore the stems are 100s), this is 500 rather than 50. (The *mean* is the thing you can't easily obtain from a stemplot, unless you read off all the values and calculate the mean directly.)

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

Find Q1, then find Q3, then take the difference.

Q1 is the median of the lower 25 values, that is, the 13th one up from the bottom. This is the 4th one on the first line beginning with "17", that is, 450. Likewise, Q3 is the 13th value down from the top, which is the last value on the other line beginning "17", that is, 530. The IQR is $530 - 450 = 80$. (If you thought the quartiles were 53 and 45, you'd get an IQR of 8, so mark (d), 50, as being the closest value to that. This would be at least a hint to go back and check what you did.)

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

    The regression line's slope is rs_y/s_x. Here, though, we know the standard deviations and the slope (from the regression line), so we work backwards to solve for r. So $-2 = (40/10)r$, which means that the correlation had better be -0.5. (This was a hard question. We wanted to put in a few of them to see how you handle them.)

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.

    The bias will still be there if the sample size gets bigger, but a bigger sample will allow the sample statistic to get closer to whatever it is estimating: that is, its variability will decrease. If this didn't happen, there wouldn't be much point in using a large sample rather than a small one.

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  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Births/woman  26   0   2.854    0.149  0.761    1.500   2.275   2.750   3.275
Life Exp.     26   0  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

    Number of births per woman is y, don't forget. Figure out the slope first, which is $(-0.812)(0.761)/(4.150) = -0.1489$. Then get the intercept, which is $2.854 - (-0.1489)(74.5) = 13.95$. There's no good way to get the intercept except to figure out the slope first.

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

    The presence of an alternative containing "extrapolation" ought to lead you to check that one first, because if it's true you can save yourself a calculation. Extrapolation happens if you are trying to do a prediction outside the range of the data. The smallest life expectancy in the data is 64 years (see "min"), and 60 is smaller than that. So we *are* trying to predict outside the range of the data, and so we quickly mark (e) and go on. If you do the calculation, the predicted number of births would be $-0.15 + 13.95(60) = 5.01$, but the actual usefulness of this number depends on whether the straight line continues to hold, which we would have to take on faith. So the *best* answer here is (e); (b) is in some sense correct but is not to my mind the best answer.

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

    There wasn't actually a Table B on the exam, but it wouldn't have helped you if there had been one. What you need to know is that Table B contains 10 different digits (0 to 9), and each one is unpredictable even when you know what's come before. This requires there to be the same number "in the long run" of each digit, which would give a proportion of $1/10 = 0.1$ for each digit.

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.

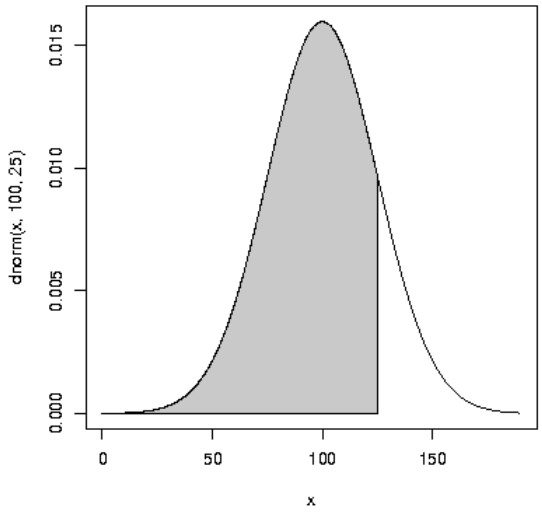
    (c) * Values below 20 or above 100.

    (d) Values below 20 or above 80.

    (e) Values below 25 or above 85.

    1.5 times IQR up from Q3 and down from Q1. IQR is $70 - 50 = 20$, and 1.5 times that is 30. Q1 minus 30 is $50 - 30 = 20$, and Q3 plus 30 is $70 + 30 = 100$. The median does not matter, which is fair enough because only unusually low or high values might be outliers, not values near the median.

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

A proportion had better be between 0 and 1, so (a) is "true" but not very insightful. The shaded piece goes from the bottom up to mean plus SD. If you like, you can figure out $z = (125 - 100)/25 = 1$, and find that the proportion less is about 0.84, or you can contort the 68-95-99.7 rule to find that the proportion between 75 and 125 is 68%, 32% is left in the two ends, so 16% is in each, and we have to add one of the ends back on (the bit below 75), which also gives 84%.

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

This is your standard normal proportion question. For 13.5, $z = (13.5 - 15)/1 = -1.5$, and proportion less is 0.0688; for 15.5, $z = (15.5-15)/1 = 0.5$, and proportion less is 0.6915. From here, there are a couple of ways to go. The easier way is to subtract the two proportions you just found. Alternatively, proportion greater than 15.5 is $1-0.6915 = 0.3085$, and everything that's not less than 13.5 or greater than 15.5 is what we want: $1 - 0.3085 - 0.0688$, leading to the same choice.

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

Use Table A backwards. If there are 10%=0.1000 of smelts longer than the value you are looking for, there are 90%=0.9000 shorter. Look up 0.9000 in the body of the table to get $z = 1.28$, and then "unstandardize" to get a length of $15 + (1)(1.28) = 16.28$ (or write $1.28 = (x - 15)/1$ and solve for x). The longest 10% of smelts are *longer* than this.

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

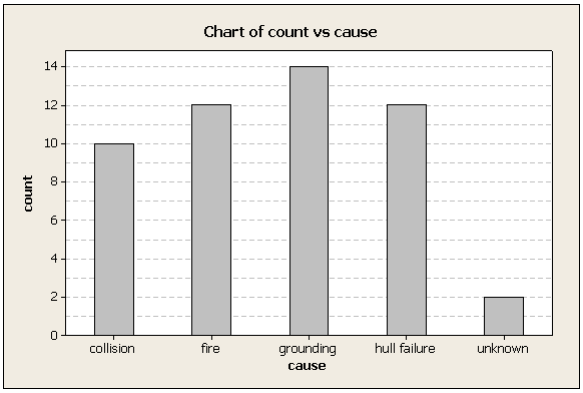
(c) none of the other alternatives

(d) 61 and 8

(e) 6 and 1

The rats are numbered up to 60, so choose 2 digits at a time. 61 is too big, 08 is OK (number 8), 12 is OK. This ought to be a straightforward use of a random digit table.

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

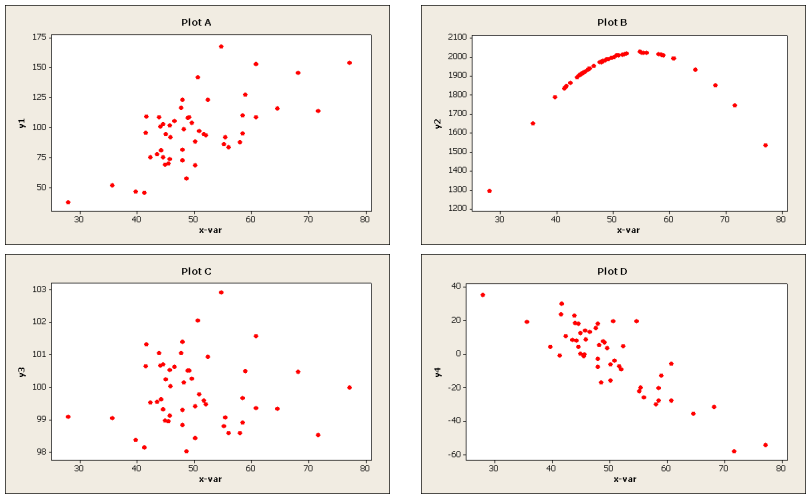
(c) 12

(d) * 20

(e) 14

10 out of 50 is 20%.

38. The four scatterplots below all show different correlations.

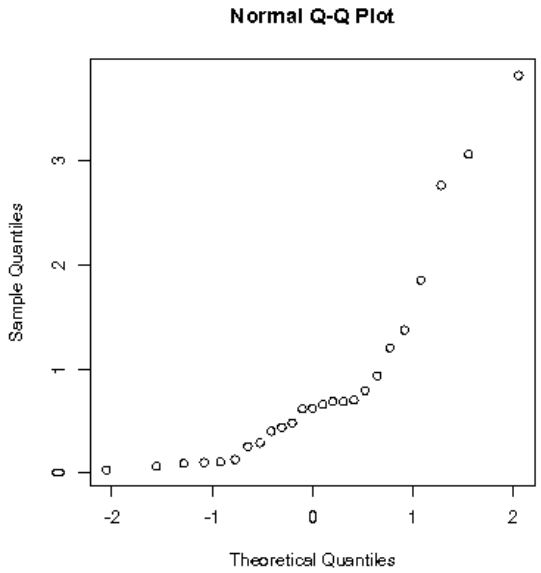


Which plot shows the highest positive correlation?

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

A is a moderate positive association (correlation of maybe 0.6). B is a strong association, but the association is curved, so the correlation will not be large (notice how it goes up then down almost the same distance, so the correlation will be close to 0. In C the correlation is close to 0, while in D the correlation is largish but negative (some value like -0.8).

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

This is a normal quantile plot, not a scatterplot, ruling out (b), so the issue is "is it straight?", and if so, a normal distribution describes the data. This is clearly not straight, so a normal distribution is not appropriate, and the curve means the data are skewed rather than symmetric. The only such alternative is (d); I don't even know what (a) means! If you had to choose between direction of skew, look for the axis containing the data (here the vertical axis). The highest values are spread out, and the lowest ones are close together, so it is indeed a right skew.

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

The *right* way to do this is to randomize who drinks and who does not, so one of the "bad" answers is called for. "Should always block for volunteers" seems like too sweeping a statement (always beware of "always" alternatives on multiple-choice!). Checking, the people who volunteer to drink might be different in some other important ways from those who do not (in ways that would affect their driving: for example, arrogant people might imagine they could drive well no matter how much they drink).