University of Toronto Scarborough
STAB22 Midterm Examination

March 2008

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").

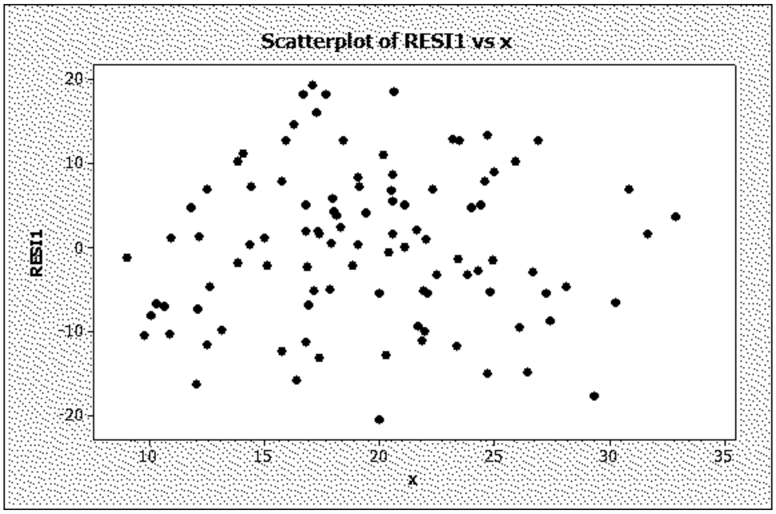
Mark in each case the best answer out of the alternatives given (which means the numerically closest answer if the answer is a number and the answer you obtained is not given.)

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.

Whichever version of the exam you had, all these questions will have been on it, though most likely with the alternative answers in a different order.

1. A linear regression was carried out for predicting one variable y from another variable x. The residuals from the regression were calculated and plotted against x. The plot is shown below.



What do you conclude from this plot?

(a) There is no evidence for a straight-line relationship at all between y and x from this graph.

(b) * A straight-line relationship between y and x is a satisfactory fit.

(c) The relationship between y and x is obviously curved.

(d) In the relationship between y and x, y is predicted more accurately for smaller values of x than for larger values.

(e) To assess the strength and form of the relationship between y and x, it is enough to look at the correlation; there is no need to look at the residual plot.

This *residual* plot is a formless mess of nothing, which is the ideal. So the straight line that was fitted to produce these residuals does describe the data satisfactorily. (Don't confuse this with a scatterplot of y vs. x: if this picture had been that scatterplot, then there would indeed have been no evidence of a relationship.

2. Use the information below for this question and the following two questions.

Scores on a standardized test for children have mean 50 and standard deviation 10, and they follow a normal distribution.

What proportion of children will score above 65 on this test? (Mark the closest answer below if your answer does not appear.)

(a) 0.50

(b) 0.93

(c) 0.20

(d) 0.80

(e) * 0.07

$z = (65 - 50)/10 = 1.5$; the proportion below is 0.93 (from table A), so proportion above is one minus that.

3. Using the information in Question 2, what proportion of children will score between 45 and 65? (Mark the closest answer below if your answer does not appear.)

   (a) 0.94
   (b) 0.31
   (c) 0.50
   (d) 0.08
   (e) * 0.63

   For 65, $z = 1.5$ as above. For 45, $z = (45 - 50)/10 = -0.5$. Look these both up in Table A and subtract: $z = 1.5$ gives 0.93 and $z = -0.5$ gives 0.30, and subtracting gives 0.63.

4. Using the information in Question 2, the lowest 5% of children will score less than what value?

   (a) 58
   (b) 41
   (c) 67
   (d) 50
   (e) * 33

   Start with a proportion and end with a value this time, so do everything backwards. Looking up 0.0500 in the table gives $z = -1.65$ (or -1.64). Converting this back to a score gives $(-1.65)(10) + 50$, about 33.

5. Some people seem not to gain weight even when they overeat. This might be explained by fidgeting and other "non-exercise activity" (NEA). In an experiment, researchers deliberately overfed healthy young adults for 8 weeks. They measured fat gain (in kilograms) and the increase in energy use (in calories) form activities other than deliberate exercise (NEA).

   The NEA increase values had mean 324.8 and SD 257.66 calories; the fat gains had mean 2.388 kg and SD 1.1389 kg. The correlation between fat gain and NEA increase was -0.7786. What is the intercept of the least-squares regression line for predicting fat gain from NEA increase?

   (a) -1.2
   (b) 1.2
   (c) -0.003
   (d) cannot be calculated because necessary information is missing
   (e) * 3.5

   Calculate the slope of the regression line first, and then get the intercept from it. All the necessary information is here. $b = (-0.7786)(1.1389/257.66) = -0.00344$; $a = \bar{y} - b\bar{x} = 2.388 - (-0.00344)(324.8) = 3.5$.

6. For this question and the next, what graphical display would be most appropriate for the variable described?

   The number of hours per week students study during a semester?

   (a) Pie chart
   (b) Scatterplot
   (c) Bar chart
   (d) * Histogram

(e) Both pie charts and bar charts are equally good for displacing the distribution of the variable involved here.

This is a quantitative variable, ruling out most of the alternatives, and just one variable (not two), ruling out the scatterplot.

7. Which radio stations are the students favorites?

    (a) Stemplot
    (b) * Bar chart
    (c) Boxplot
    (d) Scatterplot
    (e) Histogram

    There is an implied "using the previous question" here, so "which graph can best be used" for this categorical variable. The other ones are all used for quantitative variables.

8. Scores on an exam are normally distributed with a mean of 68 and a standard deviation of 9. Using the 68-95-99.7 rule, what percentage of students score above 77?

    (a) 5%
    (b) 32%
    (c) 2.5%
    (d) * 16%
    (e) 68%

    68% of the values are between $68 - 9 = 59$ and $68 + 9 = 77$. Of the other 32%, half are below 59 and half are above 77. So the answer is half of 32%.

9. The correlation between two variables x and y is 0.5. Which of these statements is true?

    (a) A larger value of x is the cause of a larger value of y.
    (b) A larger value of x is the cause of a smaller value of y.
    (c) * Larger values of x tend to go with larger values of y.
    (d) Larger values of x tend to go with smaller values of y.

    A positive correlation, and correlation does not imply causation.

10. Aspirin is believed to aid in the prevention of heart attacks. Much of the evidence for this belief comes from the "Physicians' Health Study", in which 22,000 male physicians were randomly divided into two groups. One group received aspirin, and the other group a placebo. At the end of the study, the aspirin group had fewer heart attacks than the placebo group, and the difference was statistically significant.

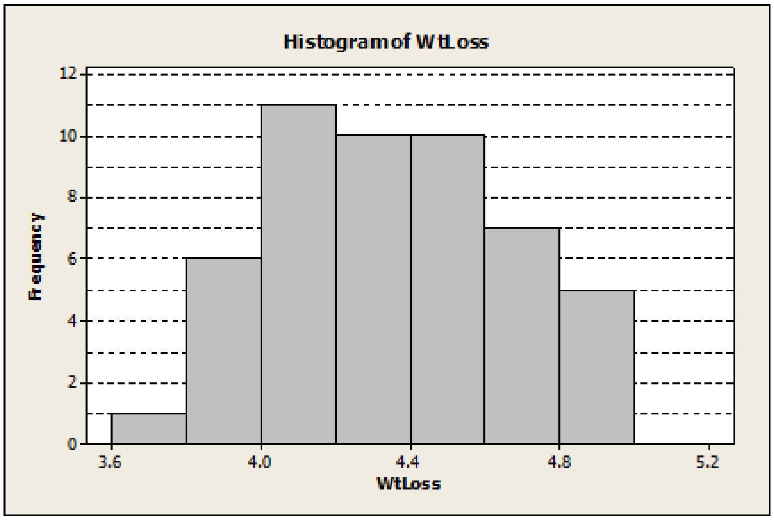
    What does "statistically significant" mean here?

    (a) * The difference in numbers of heart attacks was larger than we would expect to see by chance.
    (b) The difference in numbers of heart attacks seen here could easily have occurred by chance.
    (c) The difference in numbers of heart attacks was medically meaningful.
    (d) Because the groups were chosen at random, it is impossible to say whether this difference could have occurred by chance.

    "Statistically significant" means *only* "larger than you would expect to see by chance". It may or may not be meaningful.

    It actually *is* impossible to say whether a difference occurred by chance, but your instructions were to pick the *best* answer to the question, which this is not.

11. Shown below is a histogram (generated from MINITAB) for the weight losses (in grams) of laboratory rats 24 hours after they were injected with an experimental drug. Use this histogram to answer this question and the following question.



What percent of rats in the sample lost more than 4.0 grams? Choose the closest answer from the options below. You may assume that there were no data falling exactly at the class boundaries.

(a) 55%

(b) * 85%

(c) 95%

(d) 65%

(e) 75%

Total of $1+6+11+10+10+7+5 = 50$ rats, of which $1+6 = 7$ lost less than 4 grams, so the percent losing more was $43/50 \times 100\% = 86\%$.

12. Using the information from Question 11 above, the class that contains the median weight loss is:

(a) * (4.2, 4.4)

(b) (4.6, 4.8)

(c) (4.4, 4.6)

(d) (4.0, 4.2)

(e) (4.8, 5.0)

We want the class containing the 25th and 26th rats. The first 3 classes contain 18 rats between them, and the first 4 contain 28. So the median is in the fourth class.

13. Suppose that the weights of packages of lettuce coming off a packaging line have a normal distribution with mean 8.2 ounces and standard deviation 0.16 ounces. If every package is labeled 8 ounces, what percent of the packages weigh less than the labeled amount? Choose the closest answer from the options below.

(a) * 10%

(b) 20%

(c) 25%

(d) 5%

(e) 15%

$z = (8 - 8.2)/0.16 = -1.25$, and the proportion less than this is about 10%.

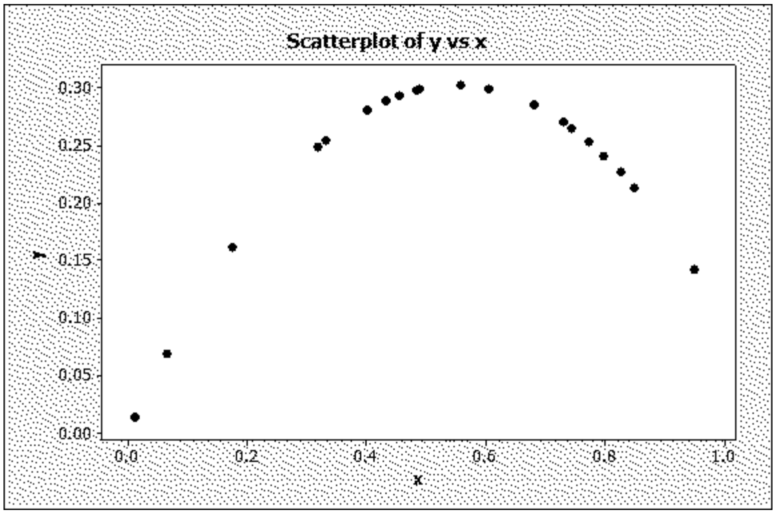
14. Using the information in Question 13 above, what is the first quartile of the distribution of the weights of packages of lettuce coming off this packaging line? Choose the closest answer from the options below.

    (a) * 8.1 ounces
    (b) 8.2 ounces
    (c) 8.3 ounces
    (d) 8.0 ounces
    (e) 7.9 ounces

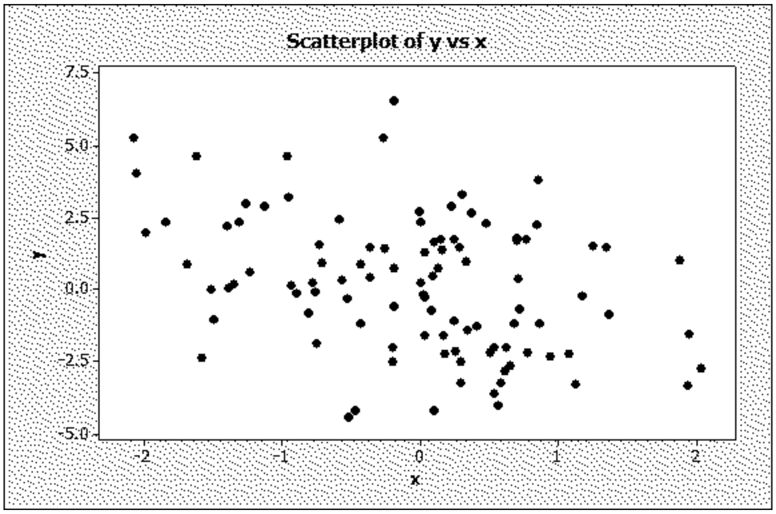
    The first quartile is the weight w such that 25% of the packages weigh less than w (and the other 75% weigh more). 25% goes with $z = -0.67$, and this goes with a weight of $(-0.67)(0.16) + 8.2$, which is closest to 8.1.

15. Five scatterplots are shown below.

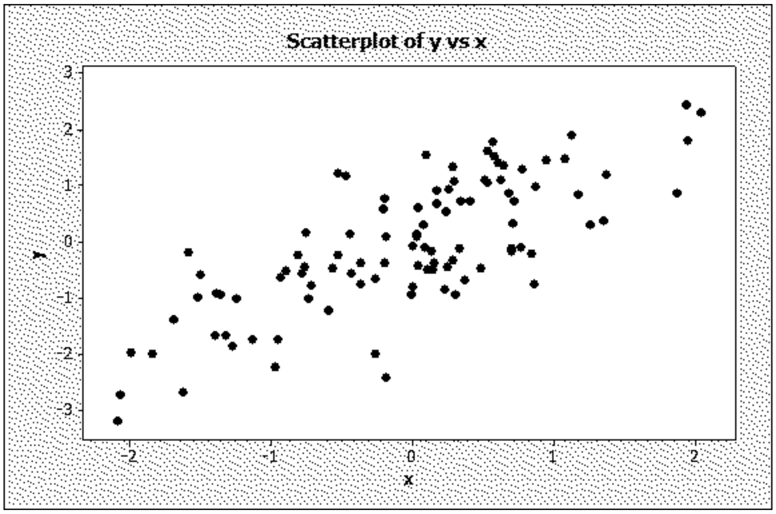
    Scatterplot 1:



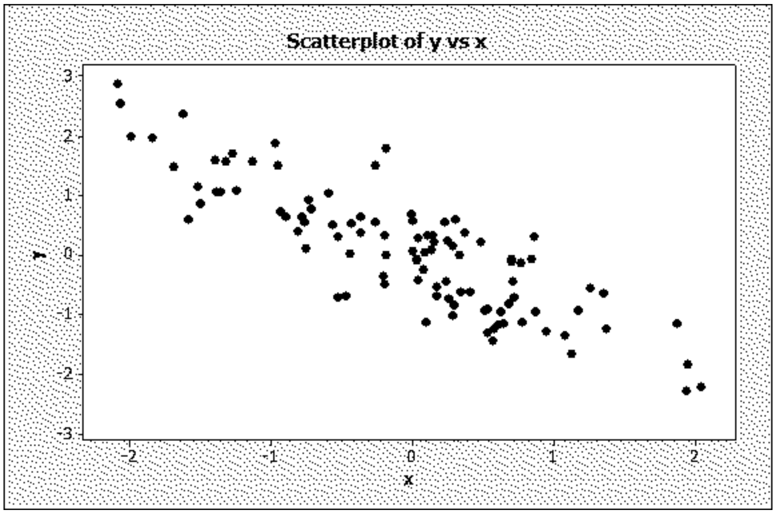
Scatterplot 2:



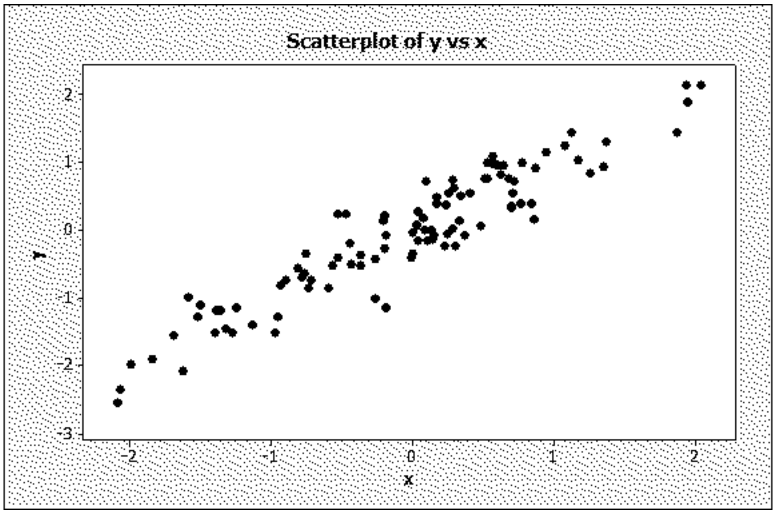
Scatterplot 3:



Scatterplot 4:



Scatterplot 5:



Which of the scatterplots above shows a correlation of -0.85?

(a) Scatterplot 1
(b) Scatterplot 3
(c) * Scatterplot 4
(d) Scatterplot 5
(e) Scatterplot 2

The relationships (apart from plot #1) get progressively stronger as you go through. -0.85 has to be a strongish downward trend, stronger than #2. 0.75 has to be upward, but not as strong a trend as #4. #3 looks stronger than 0.4. #1 is a curve, going more up than down, so it is plausible that this would have correlation 0.4.

16. Look again at the scatterplots shown in Question 15. Which scatterplot shows a correlation of 0.75?

(a) Scatterplot 5
(b) * Scatterplot 3
(c) Scatterplot 1
(d) Scatterplot 4
(e) Scatterplot 2

17. Look again at the scatterplots shown in Question 15. Which scatterplot shows a correlation of 0.4?

(a) * Scatterplot 1
(b) Scatterplot 2
(c) Scatterplot 5
(d) Scatterplot 4
(e) Scatterplot 3

18. A small university has 150 male and 100 female faculty members. The Human Resources department is commissioning a survey on working conditions, and wants to sample 70 faculty members. It is believed that males and females will have similar opinions. It is intended to contact the sampled faculty members in their offices (using an interviewer who will walk around campus). What kind of sampling method would be most appropriate?

(a) * Simple random sample
(b) Stratified sample
(c) Multistage sample

The aim here is to give you a reason to rule out the other two. There is believed to be no difference between the two subgroups, so there is no reason to use a stratified sample, and it is equally easy to contact any sampled faculty members, ruling out the use of any kind of multistage sample.

19. A government agency in Ontario wants to know how many licensed drivers have never been in an accident. One clerk looks at the records of all the licensed drivers in the province, and finds that 40% of licenced drivers have never been in an accident. Another clerk, anxious to save time, randomly selects 200 driver records, and finds that 37% of these drivers have never been in an accident.

Are the numbers 40% and 37% parameters or statistics?

(a) 40% and 37% are both statistics.
(b) 40% is a statistic and 37% is a parameter.
(c) 40% and 37% are both parameters.
(d) * 40% is a parameter and 37% is a statistic.

40% refers to all the drivers in the province, while 37% comes from a sample.

20. In an experiment in ecology, researchers are attempting to predict the proportion of perch (in a pen) eaten by bass, using as explanatory variable the number of perch in the pen before the bass were let in. Supposing that y is the proportion of perch eaten, and x is the initial number of perch, the regression equation turned out to be

$$y = 0.120 + 0.0086x.$$

How much would you expect the proportion of perch eaten to change if the initial number of perch increases by 1?

(a) decrease by -0.120

(b) * increase by 0.0086

(c) decrease by 0.0086

(d) increase by 0.120

(e) increase by 0.1286

This is the definition of the slope of the regression line.

21. The stemplot and some descriptive statistics (generated from MINITAB) of the grades of a statistics class (not STAB22) are given below. Use this information for this question and the two questions following.

```
Stem-and-Leaf Display: grade

Stem-and-leaf of grade  N  = 90
Leaf Unit = 1.0


 38   5  00000011111111111222333333333334444444
(21)  5  555567777777888889999
 31   6  0001122233344
 18   6  68899
 13   7  0111233
 6    7  667
 3    8
 3    8  69
 1    9
 1    9  5

Descriptive Statistics: grade

Variable   N  N*      Q1      Q3
grade     90   0  53.211  63.223
```

What is the median score of this statistics class? Choose your answer from the options below. If none of the options below is equal to the median, choose the one closest to the median.

(a) 56

(b) * 57

(c) 55

(d) 58

(e) 54

The median is the average of the 45th and 46th values. There are 38 in the first class "the low 50s", so we need 7 or 8 from the next class. The 45th and 46th values are both 57, so this is the median.

22. Using the information from Question 21 above, what percent of students in this class will receive an A grade if a score of 80 or above qualifies for an A? Choose the closest answer from the options below.

    (a) 25%
    (b) 6%
    (c) * 3%
    (d) 10%
    (e) 1%

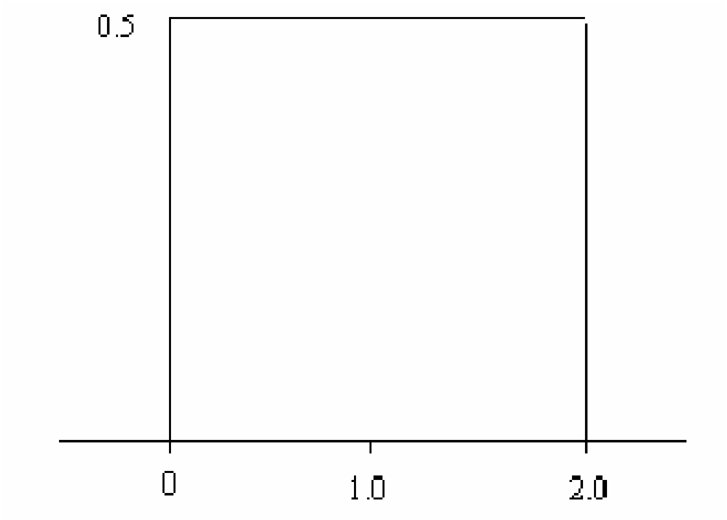
    3 out of 90.

23. Using the information from Question 21 above, how many of these scores are identified as outliers by the 1.5 × IQR rule?

    (a) Only one outlier
    (b) Only two outliers
    (c) More than three outliers
    (d) * Only three outliers
    (e) No outliers

    IQR is $63.2 - 53.2 = 10$, and $1.5 \times IQR = 15$. Count the values less than $53.2 - 15 = 38.2$ (none), and those more than $63.2 + 15 = 78.2$ (three).

24. The density curve of a random variable is given below.



    Use this information to answer this question and the following question.

    What is the interquartile range of the distribution of this random variable?

    (a) 1.25
    (b) 0.50
    (c) * 1.00
    (d) 1.50

(e) 0.75

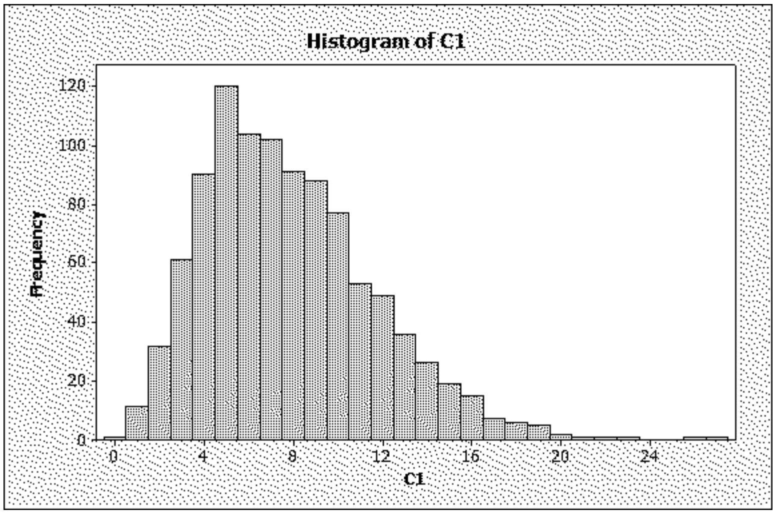
The lines up the side just show you that the density curve is for values between 0 and 2 (and the height is 0.5). The first quartile is the value for which the area to the left is 0.25 (0.5 will do it) and the third quartile has area to the left being 0.75 (1.5). So the IQR is $1.5 - 0.5 = 1$.

25. Using the density curve given in Question 24 above, what percent of observations of this distribution lie below 0.6?

    (a) 40%
    (b) * 30%
    (c) 20%
    (d) 25%
    (e) 50%

    The area of the rectangle whose width is 0.6 (from 0 to 0.6) and whose height is 0.5.

26. The histogram below is of the sampling distribution of a sample mean, based on samples of size 20.



    What does this sampling distribution tell you?

    (a) There is an error here because the population size is not given.
    (b) There are no values over 40 in the population.
    (c) This cannot be the picture of a sampling distribution because sampling distributions always have a normal shape.
    (d) * Most of the possible samples of size 20 will have a sample mean between 3 and 10.
    (e) Most of the values in the population lie between 3 and 10.

    A sampling distribution always talks about some quantity calculated from a sample, and how that quantity might vary.

27. What is the median of the numbers 3, 2, 0, 7, 7?

    (a) 7
    (b) * 3
    (c) 4.4
    (d) 0

You remembered to sort the numbers first, didn't you?

28. A researcher believes that taking vitamin C helps to prevent a person from developing a cold. The researcher collects a number of people who have colds, and matches each person with someone else of the same sex, similar age and diet who does not have a cold. The researcher finds that for those people who have a cold, very few of them took vitamin C, and for those who do not have a cold, a large proportion had taken vitamin C.

    Which statement below best describes this situation?

    (a) This is not a statistical experiment, so the researcher is not allowed to conclude that vitamin C helps to prevent colds.

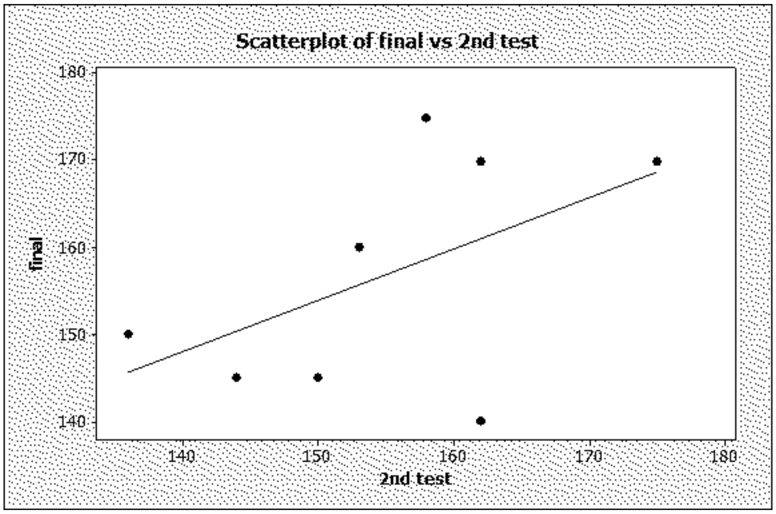
    (b) This study is all mixed up, because taking vitamin C is the response and having (or not having) a cold is the factor.

    (c) * Because each person with a cold was matched with a similar person without a cold, the researcher is entitled to conclude that taking vitamin C helps to prevent colds.

    (d) The researcher needed to randomize who took vitamin C and who did not in order to draw any valid conclusions from the study.

    This is a case-control study, and because of the matching, we have some evidence for cause and effect even though it's not an experiment. So an answer that says "it is not an experiment" is OK, but the indicated answer is better.

29. A certain course at a university has two term tests and a final exam. There are 8 students in the course. The scatterplot below shows the scores on the 2nd test and the final exam for these students. The regression line has been drawn on the scatterplot.



    The student who scored 150 on the second test scored approximately how many on the final exam?

    (a) 135

    (b) 140

    (c) 170

    (d) * 145

    (e) 152

    The observed value, so read it off the graph. (There was nothing asking for a predicted value.)

30. For the data described in Question 29, what is the residual for the student who scored 150 on the second test?

(a) 0

(b) -10

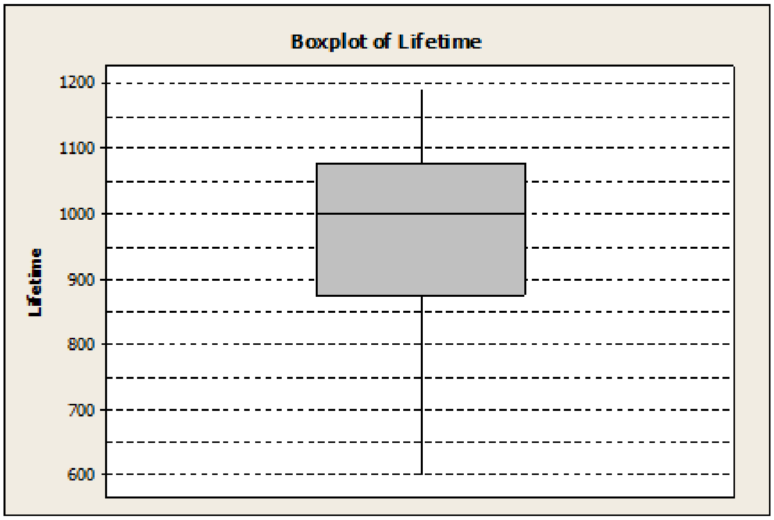
(c) * -7

(d) 10

(e) 5

The predicted value is hard to see with accuracy, but it is definitely less than 155, and possibly as small as 152. (Go up to the line and then go across.) So not -10 and not 0 either.

31. Given below is the boxplot (generated from MINITAB) of lifetimes (in hours) of a sample of 20 incandescent lamps. Use the boxplot to answer this question and the three following.



Which of the following numbers is the closest to the median lifetime of this sample?

(a) 1050 hours

(b) 950 hours

(c) 850 hours

(d) * 1000 hours

(e) 900 hours

The bar across the middle of the box.

32. Using the information from Question 31 above, which of the following numbers is the closest to the first quartile of the lifetimes in this sample?

(a) 600 hours

(b) 1175 hours

(c) * 875 hours

(d) 1000 hours

(e) 1075 hours

The bottom of the box.

33. Using the information from Question 31 above, which of the following numbers is the closest to the interquartile range of the lifetimes in this sample?

(a) 600 hours

(b) 150 hours
(c) * 200 hours
(d) 400 hours
(e) 75 hours

The height of the box.

34. Based on information from Question 31 above, which of the following statements is true?

(a) The distribution of the lifetime is right skewed.
(b) * More than 25% of the lamps in this sample had lifetimes over 1050 hours.
(c) The shortest lifetime observed in this sample is about 875 hours
(d) No lamp in this sample had a lifetime over 1100 hours.

The shape of the boxplot suggests left-skewed. The shortest lifetime is at the bottom of the lower whisker (about 600), and the longest lifetime is the top of the upper whisker (not quite 1200). The third quartile is about 1075, and we know that 25% of the lifetimes are bigger than this (so there must be more than 25% over 1050).

35. Pine trees that grow in the dry forests of Arizona may be better able to resist drought if they can grow in the shade. To test this, an experiment was carried out. Investigators planted pine seedlings in a greenhouse in either full light or reduced light (light reduced to 5% of normal by shade cloth). At the end of the study, they weighed the young trees.

In this experiment, what are the experimental units or subjects?

(a) The light conditions in the greenhouse
(b) * The pine seedlings
(c) The weight of the young trees
(d) Ability of the seedlings to grow in shade

The things on which the experiment was conducted.

36. In the experiment of Question 35, what are the treatments?

(a) The pine seedlings
(b) Ability of the seedlings to grow in shade
(c) The weight of the young trees
(d) * The light conditions in the greenhouse (full light or reduced light)

What you do to the pine seedlings (subject them to different light conditions).

37. In the experiment of Question 35, what is the response?

(a) Full light or reduced light in the greenhouse
(b) Whether the seedlings can grow in shade
(c) One of the pine seedlings
(d) * The weight of the young trees

The thing you measure at the end.