University of Toronto Scarborough
STAB22 Midterm Examination

March 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 15 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").

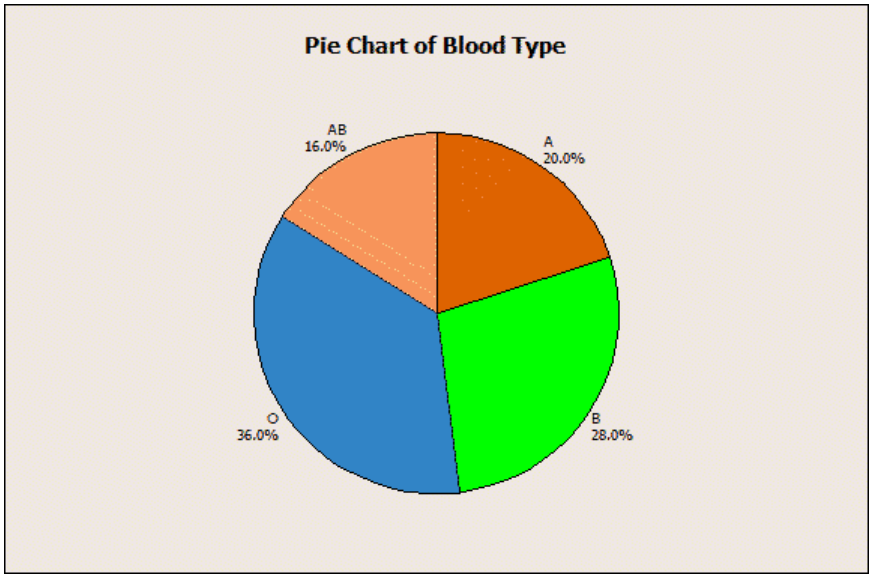
Mark in each case the best answer out of the alternatives given (which means the numerically closest answer if the answer is a number and the answer you obtained is not given.)

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.

This is the Pink version. If you did another version, all the questions you did are in here somewhere. Correct answers are marked by an asterisk, with brief explanation.

1. A random sample of 25 blood donors was given a blood test to determine their blood type. The pie chart below shows the distribution of the blood types of these 25 donors. (Note: A, B, O and AB are the blood types)



How many donors in this sample had blood type A?

(a) 10
(b) 15
(c) * 5
(d) 18
(e) 20

20% of 25.

2. Some researchers have proposed a new treatment for Alzheimer's disease. They propose to divide their subjects into two groups; one group gets this new treatment, while the other group gets the standard treatment. At the end of the study, the quality of life of all the subjects is assessed.

   What would you say about this study design?

   (a) the researchers didn't need to have a group of subjects receiving the standard treatment. They could have obtained equally good results with half the number of subjects.
   (b) the researchers could have used available data
   (c) * it enables the researchers to see whether the new treatment has more than a placebo effect
   (d) this is a case where a comparative experiment is not necessary
   (e) there is likely to be a nonresponse bias

   The group getting the standard treatment is the control group, and doing a comparative experiment using a control group is the best way to assess whether a treatment is really effective.

3. In the study of Question 2, suppose that the researchers obtain statistically significant results, with the new treatment offering a higher average quality of life. Are the researchers entitled to conclude that the new treatment is a cause of a higher quality of life?

(a) yes, because the results come from an observational study

(b) no, because the results come from an observational study

(c) * yes, because the results come from a statistical experiment

(d) no, because the results come from a statistical experiment

It's an experiment because a treatment was imposed on the subjects. Therefore it can produce evidence of cause and effect, and because statistically significant results were obtained, it does.

4. Two variables x and y are believed to have a straight-line relationship. We would like to predict y from x. Minitab tells us this about x and y:

```
Descriptive Statistics: x, y

Variable  N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
x         7   0   7.00     1.63   4.32     1.00   3.00    7.00  11.00    13.00
y         7   0  13.43     1.67   4.43     8.00  10.00   13.00  17.00    21.00

Correlations: x, y

Pearson correlation of x and y = 0.941
P-Value = 0.002
```

Use this information for this question and the two following.

What is the **intercept** of the regression line for predicting y from x?

(a) -5.3

(b) 6.4

(c) 0.9

(d) * 6.7

(e) 1.0

Slope is $(0.941)(4.43/4.32) = 0.96$, intercept is $13.43 - 0.96(7) = 6.7$.

5. Using the information given in Question 4, what is the predicted value of y when $x = 10$? The slope of the regression line is 1.0.

(a) 12.2

(b) 13.4

(c) * 16.7

(d) 7.7

$6.7 + (1)(10) = 16.7$. Or: $x = 10$ is higher than average for x, and the correlation is positive, so y should be higher than average for y too. The only alternative that is is 16.7.

6. In Question 4, some information is given about two variables x and y. From the information given, does it make sense to find the regression line?

(a) No, because x and y have outliers

(b) * Yes

(c) No, because the correlation is not a good measure of the relationship between x and y

(d) No, because the relationship is not a straight line

You can do a quick $1.5 \times IQR$ check to verify that there are no outliers. The correlation *is* a good measure of the relationship if it is a straight line, and there is nothing in the output to suggest that a curve would be better.

7. PTC is a compound that has a strong bitter taste for some people ("tasters") and no taste at all for others ("nontasters"). The ability to taste PTC is an inherited trait. A study of people in Ireland and Portugal gave the numbers of tasters and nontasters shown:

	Ireland	Portugal
Tasters	550	350
Nontasters	220	100

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

One of the conditional proportions is 0.714. Which one?

(a) the proportion of tasters overall

(b) Out of the people who are tasters, the proportion who are from Portugal

(c) Out of the people who are tasters, the proportion who are from Ireland

(d) Out of the people from Portugal, the proportion who are tasters

(e) * Out of the people from Ireland, the proportion who are tasters

Work them out. If you look only at the people from Ireland, the proportion of tasters is $550/(550 + 220) = 0.714$.

8. From the information in Question 7, which one of the following statements is true?

(a) people from Portugal are less likely to be tasters than people from Ireland

(b) people who are tasters are more likely to be from Portugal

(c) in this study, the marginal proportion of tasters is less than 0.50.

(d) * people from Portugal are more likely to be tasters than people from Ireland

$350/(350 + 100)$ is bigger than $550/(550 + 220)$. The other statements are all false.

9. The price of seafood varies with species and time. Data was collected on the prices (cents per pound) received by fishers for several species in 1970 and 1980, and a regression was carried out to predict the 1980 price from the 1970 price. Some Minitab output is shown below; some items have been deleted.

```
Descriptive Statistics: 1970 price, 1980 price

Variable     N  N*   Mean  SE Mean  StDev  Minimum    Q1  Median     Q3
1970 price  14   0   41.1     10.1   37.7     1.80  11.5    33.0   56.8
1980 price  14   0  109.8     28.1  105.2     4.50  26.2    94.9  154.6

Variable    Maximum
1970 price    135.6
1980 price    404.2

Regression Analysis: 1980 price versus 1970 price

The regression equation is
1980 price = - 1.2 + 2.70 1970 price

Predictor     Coef  SE Coef      T      P
Constant     -1.23    11.26  -0.11  0.915
1970 price  2.7016   0.2053  13.16  0.000

Unusual Observations
```

```
      1970
Obs  price  1980 price     Fit  SE Fit  Residual  St Resid
 11     95      190.00  deleted  13.29    ... deleted ...
 13    136      404.20  365.10   20.79     39.10      2.10X

X denotes an observation whose X value gives it large influence.
```

Use this information for this question and the one following.

American lobsters sold for 95 cents per pound in 1970 and 190 cents per pound in 1980. What is the residual for this observation?

(a) 255

(b) 0

(c) * -65

(d) 65

(e) 190

Predicted 1980 selling price is $-1.2 + 2.70(95) = 255$, so residual is $190 - 255 = -65$.

10. Observation 13 is sea scallops, which sold for 135.6 cents per pound in 1970 and 404.2 cents per pound in 1980. Why does Minitab mark this observation with an X?

(a) * The 1970 price is unusually high compared to the other species.

(b) Minitab just felt like marking this observation with an X.

(c) The 1970 price is unusually low compared to the other species.

(d) The 1980 price is unusually low compared to the other species.

(e) The 1980 price is unusually high compared to the other species.

X means "unusual X-value", which means the 1970 price is unusual. The mean of the 1970 prices is 41.1, and 135.6 is higher, so this must be an unusually high 1970 price.

11. The United Kingdom contains 4 countries, England, Scotland, Wales and Northern Ireland. Some people who live in Scotland feel especially strongly that Scotland should be independent of the United Kingdom. (The United Kingdom is dominated by England, because there are more people living in England than Scotland). A survey is to be taken about attitudes towards independence of Scotland. The survey will cover all of the United Kingdom. What would be the best way to take the survey?

(a) set up a website where people can post their opinions

(b) use Table B

(c) * a stratified sample using the 4 countries as strata

(d) a simple random sample

Because different parts of the population can be expected to give different responses to the survey, a stratified sample is called for (and will give better results than a simple random sample). The website will produce a voluntary-response sample (bad). Table B could be used at some point, but this would be true whatever kind of random sampling is used.

12. A statistical experiment is carried out on 30 male rats. 15 of the rats, chosen at random, are given an energy drink, while the other 15 rats maintain their usual diet. All of the rats then run through a maze. The rats who had the energy drink run the maze in a mean time of 37.2 seconds, while the rats who did not have the energy drink run the maze in a mean time of 41.3 seconds.

Are the numbers 37.2 and 41.3 parameters or statistics?

(a) * 37.2 and 41.3 are both statistics.

(b) 37.2 and 41.3 are both parameters.

(c) 37.2 is a parameter and 41.3 is a statistic.

(d) 37.2 is a statistic and 41.3 is a parameter.

They both come from samples, so they are both statistics.

13. A statistical experiment is to be carried out. There are 20 subjects available, numbered 1–20. 10 of these subjects are to be selected for the treatment group, and 10 for the control group. Randomization is to be used to determine which subjects are selected for each group. Table B will be used for the randomization. Out of the methods described below, one of them is both a valid method of selecting the 10 subjects to be in the treatment group, and is the most efficient use of Table B (it uses the fewest numbers from the table). Which one?

    (a) * Select 2-digit groups. If the number formed is greater than 20, repeatedly subtract 20 from it until a result of 20 or less is obtained. The subject with the resulting number is chosen for the treatment group. If the selected subject has already been selected for the treatment group, ignore the 2 digits chosen and move on to the next 2 digits.

    (b) Start with the first subject. Select a single digit. If that digit is between 0 and 4 (inclusive), the first subject goes into the treatment group. Otherwise, the first subject goes into the control group. Repeat for all 20 subjects.

    (c) Select 2-digit groups. If the 2 digits form a number that is 20 or less, select that subject for the treatment group. If the selected subject has already been selected for the treatment group, ignore the 2 digits chosen and move on to the next 2 digits.

    (d) Select single digits. Select the subject numbered by this digit for the treatment group. If the selected subject has already been selected for the treatment group, ignore the digit chosen and move on to the next digit.

    (b) might produce more or less than 10 people in the treatment group (it is like tossing a coin to make the decision for each person). (d) can only select people numbered up to 9, but there are 20 people altogether. (c) is the way we learned in class, but it is rather wasteful of digits because every time you choose a pair of digits beginning with 3 or more, you cannot use them. For (a), does every number between 1 and 20 have an equal chance to be chosen? Yes (imagine writing out which subject eventually gets chosen from any initial pair of digits and you will see that, for example, subject 12 would be chosen if the initial digits are 12, 32, 52, 72 or 92 and there are always 5 ways to select each subject between 1 and 20. This doesn't waste any digits at all, so it is more efficient than (c).

14. In the experiment described in Question 13, what is the most efficient valid way to select the 10 subjects to be in the control group?

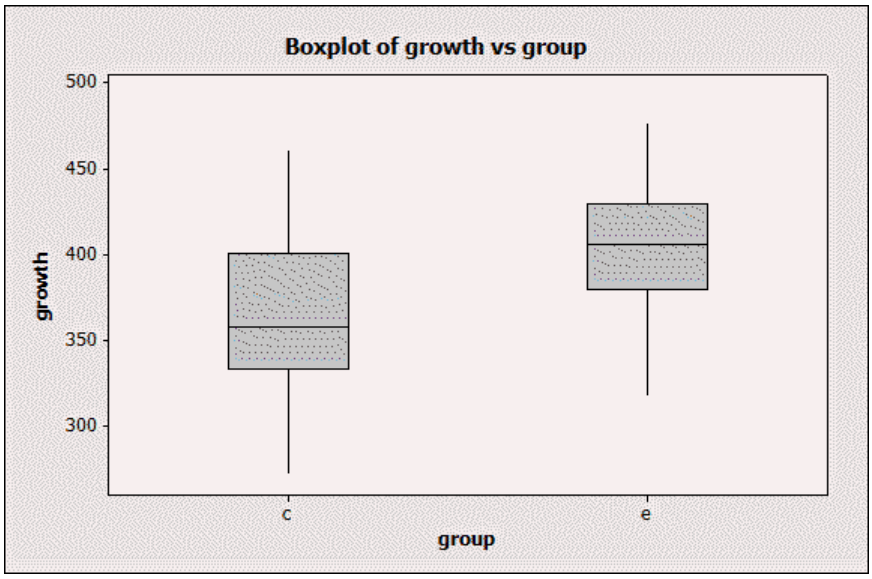
    (a) Since only 10 subjects remain, number the remaining subjects 0–9 and select single digits from Table B.

    (b) * Select the remaining 10 subjects who were not chosen for the treatment group

    (c) Repeat the process used in Question 13.

    No reason to be clever here: if they're not in the treatment group, they must be in the control group.

15. In a study, the growth of chicks fed normal corn and of chicks fed a new variety of corn was compared. The values recorded were weight gains (in grams) from the beginning of the study to the end. The results are shown as boxplots below. The control group (on the left of the boxplot) was fed the normal corn, and the experimental group (on the right) was fed the new corn.



Some statements are given below about these data. From the information contained in the boxplots, only one of these statements is definitely true. Which one?

(a) The median weight gain for the chicks fed the normal corn is higher than the chicks fed the new corn.

(b) There are outliers among the chicks fed the normal corn.

(c) * The weight gains for the chicks fed the normal corn have a larger interquartile range than those for the chicks in the experimental group.

(d) The mean weight gain for the chicks fed the normal corn is lower than the chicks fed the new corn.

(e) The weight gains for the chicks fed the new corn have a distribution that is definitely skewed to the right.

Compare the heights of the boxes. The other statements are all false. (Boxplots don't tell you about means.)

16. The mean of a data set is equal to zero. Which of the following statements regarding this data set must be true?

(a) * none of the other statements is necessarily true

(b) 50% of the values in the data set are negative and 50% are positive

(c) the distribution of the values in the data set is positively skewed

(d) the median of the data set must also be zero

(e) each value in the dataset must be equal to zero

The other statements may be true, but they don't have to be. The data set $-1, -1, -1, 3$ has mean 0 but median -1, with values that are not all zero (and 75% negative values).

17. A city is going to use its property tax records to select people living in that city to take part in a focus group on recycling in the city. (Everyone who owns a house or apartment in the city will be listed in the property tax records.) These focus groups have been very popular, and it turns out that everyone selected to take part in the focus group actually does take part. The city's aim is that everyone who lives in the city has a chance to take part in the focus group. What problems, if any, do you see with this sampling method?

(a) there are no problems with this sampling method

(b) nonresponse

(c) response bias

(d) * undercoverage

(e) lack of realism

People who are not in the property tax records cannot be sampled (for example, people who rent their homes rather than owning them).

18. For a certain data set, the interquartile range is equal to zero. Which of the following statements regarding this data set must be true?

(a) * none of the other statements is necessarily true

(b) all the values in the data set are identical

(c) the distribution of the values in the data set is positively skewed

(d) the mean of the data set must also be zero

(e) 50% of the values in the data set are negative and 50% are positive

The middle 50% of the values must be zero, but the ones at the upper end could be positive and the ones at the lower end could be negative. The skewness could go either way, and if (for example) the positive values are bigger than the negative ones, the mean would be positive. (e) must be false, since at least 50% of the values must be 0.

19. The MINITAB summary statistics of the IQ scores of a group of students are given below:

```
Descriptive Statistics: iq

Variable   N  N*    Mean  StDev  Minimum      Q1  Median      Q3  Maximum
IQ        78   0  110.00  13.00    72.00  103.00  110.00  118.00   136.00
```

Assume that no two students in this group had the same IQ score.

What percentage of students in this group had IQ score greater than or equal to 103?

(a) 25

(b) * 75

(c) 50

(d) 10

(e) 95

25% of the values are less than Q1, so 75% must be greater.

20. Assume that the distribution of IQ scores in Question 19 above is approximately normally distributed. What percentage, approximately, of students in this group have an IQ score of 97 or greater?

(a) 2.5

(b) 16

(c) * 84

(d) 95

(e) 68

$z = (97 - 110)/13 = -1$; table gives about 0.16, so $1 - 0.16 = 0.84$. Or: 68% of values between $110 - 13 = 97$ and $110 + 13 = 123$. Of remaining 32%, half (16%) below 97. Rest (100%-16%) are above 97.

21. We transform the IQ scores in Question 19 above by multiplying each students IQ score by 0.9 and then adding 5. (Example: If a student has 80 before transformation, his score after the transformation is $(0.9 \times 80) + 5$). Calculate the value of the IQR (interquartile range) of the IQ scores after this transformation.

    (a) 18.5
    (b) 64
    (c) 15
    (d) 20
    (e) * 13.5

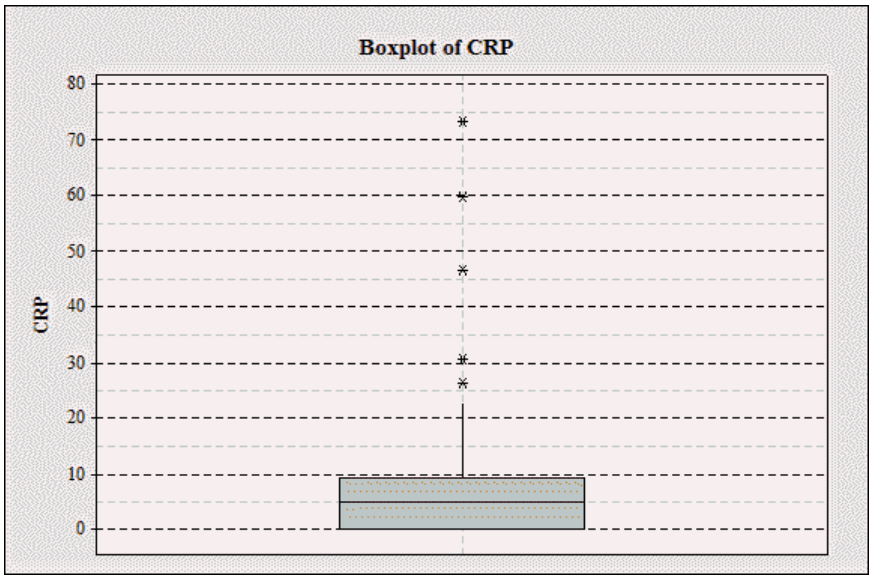
    It's a measure of spread, so adding 5 doesn't affect it. Thus the IQR becomes $(0.9)(118 - 103) = 13.5$.

22. Suppose we are interested in the mean height of all students in this class. We want to know the sampling distribution of the sample mean height based on samples of size 10. Which of the methods below would produce a histogram of this sampling distribution?

    (a) Use a sample of 20 students, because 20 students will estimate the mean height of all the students in the class better than 10 will.
    (b) Measure the height of every student in the class, and draw a histogram of the values.
    (c) * Make a list of all possible samples of 10 students from the class. For each one, find the sample mean height. Draw a histogram of these values.
    (d) Take a simple random sample of 10 students from the class. Find the height of each student in the sample, and draw a histogram of the values.

    Only (c) gets at the idea of repeated sampling ("all possible samples of size 10"). (a) is true but irrelevant, (b) is the population distribution, and (d) is only one possible sample.

23. C-reactive protein (CRP) is a substance that can be measured in the blood. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of healthy children aged 6 to 60 months in Papua New Guinea, CRP was measured in a sample of 40 children. The units are milligrams per litre (mg/l). Here is the boxplot of the data from this sample of 40 children:



Use this information for this question and the 4 questions following.

Looking at the boxplot, what is the interquartile range, in mg/l?

(a) 15
(b) 0
(c) * 10
(d) 20
(e) 5

Height of the box, top to bottom.

24. Using the information in Question 23 above, which of the following is the most appropriate measure of spread in the CRP values in the sample given?

(a) median
(b) standard deviation
(c) range (maximum minus minimum)
(d) * IQR (interquartile range)
(e) all the other options are equally good for measuring spread in the sample given

The shape is skewed (or has outliers), so the SD and range will be affected by the outlying values when the IQR won't. The median is not even a measure of spread!

25. Using the information in Question 23 above, which of the following is the most appropriate measure of centre of the CRP values in the sample given?

(a) mode
(b) mean
(c) IQR (interquartile range)
(d) * median
(e) all the other options are equally good for measuring the centre in the sample given

Median is better than mean, for the same reasons as above.

26. Using the information in Question 23 above, how many children in this sample had CRP measurement between 5 and 10 mg/l?

(a) 25
(b) 15
(c) 20
(d) 5
(e) * 10

The upper half of the middle 50%, so 25% of 40.

27. Using the information in Question 23 above, how would you describe the shape of the distribution of CRP values?

(a) skewed to the left
(b) * skewed to the right
(c) like a normal curve

(d) approximately symmetric

There are outliers at the upper end, and the upper whisker is longer than the non-existent lower whisker.

28. What is Q3 (the third quartile) for a normal distribution with mean 70 and SD 15?

    (a) 90

    (b) cannot calculate from the information given.

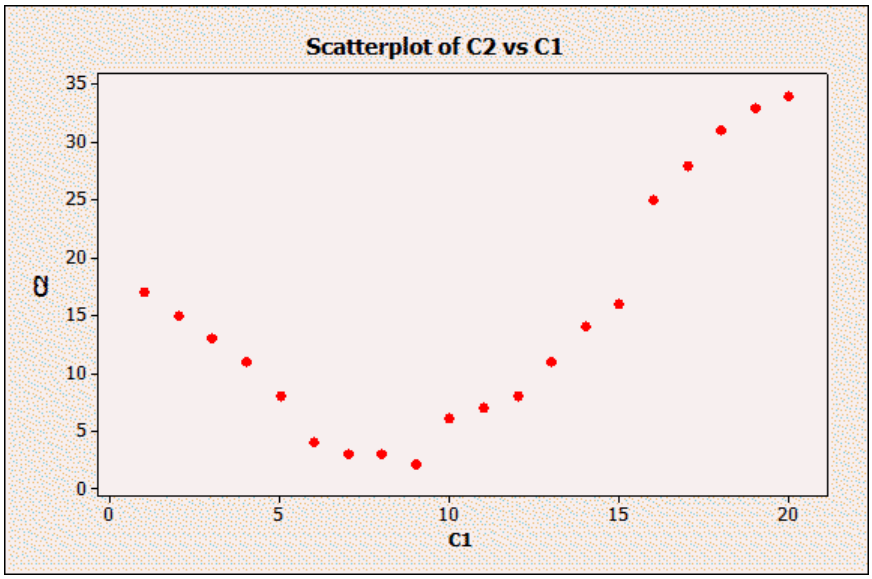
    (c) 70

    (d) * 80

    (e) 60

    $z = 0.67$ has 0.75 below and 0.25 above. Converting to the mean and SD given, Q3 is $(0.67)(15) + 70 = 80$.

29. The scatterplot below shows the relationship between two variables. What is the correlation between these variables?



    (a) * 0.6

    (b) -0.9

    (c) -0.6

    (d) 0

    (e) 0.9

    It's a curved trend, but more up than down, so the correlation is not 0 and not negative. The curve reduces the correlation, so 0.6 is better than 0.9.

30. The blood cholesterol levels for young women (aged 20 to 34) follow a normal distribution with mean 185 milligrams per deciliter (mg/dl) and standard deviation 39 mg/dl. Cholesterol levels above 240 mg/dl demand medical attention. Using the normal distribution, approximately what percentage of young women have cholesterol levels above 240 mg/dl?

    (a) 68

    (b) * 8

(c) 16

(d) 92

(e) 4

$z = (240 - 185)/39 = 1.41$; from table, proportion less is about 0.92, so proportion more is 0.08.

31. Some summary statistics, calculated by Minitab, and a stemplot for the test scores of 150 students enrolled in a certain course are given below. Use this information for this question and the one following.

```
Descriptive Statistics: score

Variable    N    Mean  SE Mean  TrMean  StDev      Q1      Q3
score     150  66.160    0.654  67.060  8.004  62.000  72.000

Stem-and-Leaf Display: score

Stem-and-leaf of score  N  = 150
Leaf Unit = 1.0

 1    3  2
 2    3  8
 4    4  34
 7    4  555
 13   5  012333
 23   5  5567778899
 44   6  000001111122222333334
(42)  6  55566666666666677777888888888999999999999
 64   7  000000000001111111111222222222222233333333334444444444444444
 4    7  5555
```

Which of the following numbers is closest to the median score?

(a) 6.8

(b) 68

(c) * 69

(d) 6.9

(e) 67

Median is between 75th and 76th value (from top or bottom). From top is easier: there are 64 values in the bottom 2 rows, so you need 11 or 12 more from the end of the previous row. Both of these are 69 (not 6.9 because leaves are 1s and stems are 10s).

32. Using the information on scores given in Question 31 above, how many outliers are there according to the "$1.5 \times IQR$" criterion?

(a) * more than 3

(b) none

(c) exactly 2

(d) exactly 1

(e) exactly 3

IQR is $72 - 62 = 10$, so $1.5 \times IQR$ is 15. Anything below $62 - 15 = 47$ or above $72 + 15 = 87$ is an outlier: there's nothing above 87, but all those values 45 and less are outliers.

33. Some investors believe in a "January indicator" for the stock market: that is, if the stock market is up in January, it will be up for the rest of the year (and if the market is down in January, it will be down for the rest of the year). Historical data from a stock market is as follows:

	Rest of year	
January	Up	Down
Up	42	16
Down	21	23

Use this information for this question and the next one.

Find the marginal distribution of ups and downs for January. What is the marginal proportion of times the stock market went up?

(a) 0.5
(b) * 0.6
(c) 0.3
(d) 0.7
(e) 0.4

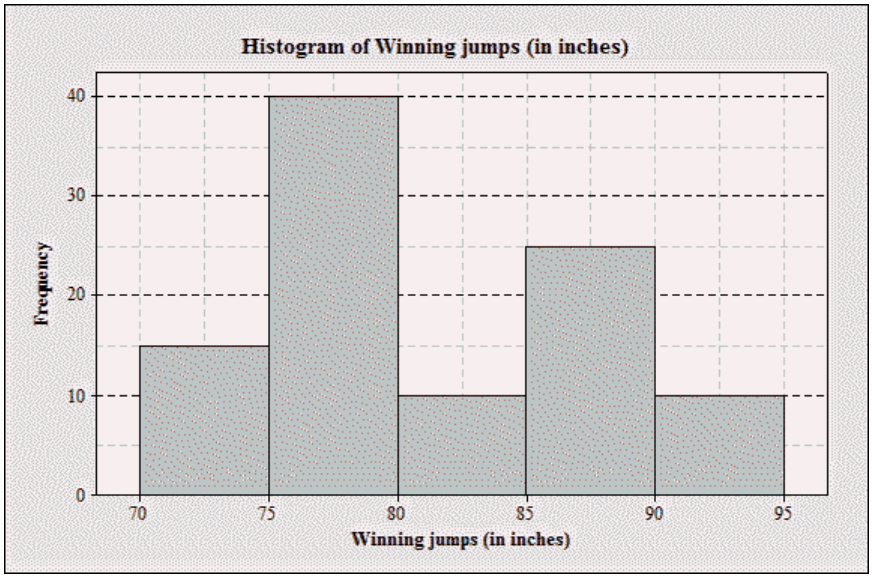
Add up over the rest of the year: 58 ups and 44 downs in January, so the marginal proportion of ups is just below 0.6.

34. Using the information in Question 33, find the conditional distribution of ups and downs for the rest of the year when the market went up in January. What is the conditional proportion of times the market went up the rest of the year?

(a) 0.4
(b) 0.5
(c) * 0.7
(d) 0.6
(e) 0.3

Ignore the times the market went down in January, so 42 out of $42 + 16$.

35. The histogram below represents the height (in inches) of the gold medal-winning high jumps for the Olympic games up to Sydney 2000.



Which of the following class intervals will contain the median?

(a) 80 to 85

(b) 90 to 95

(c) 70 to 75

(d) 85 to 90

(e) * 75 to 80

There are $15 + 40 + 10 + 25 + 10 = 100$ observations in total, so the median is between the 50th and 51st. There are 15 observations below 75, and $15 + 40 = 55$ below 80, so the median is just below 80: in the interval 75 to 80.

36. Which of the descriptions below best describes simple random sampling?

(a) Each individual has the same chance to be in the sample.

(b) A newspaper contains an advertisement for people's opinions on a certain issue. The sample consists of those people who reply to the advertisement.

(c) Subgroups of the population (such as males and females) are guaranteed to be properly represented in the sample.

(d) * Each individual has the same chance to be in the sample, independently of other individuals.

(a) is true of simple random sampling, but it is true of other kinds of sampling too; you need to add the independence to make the best description of simple random sampling. (b) is a voluntary response sample, while (c) describes stratified sampling.

37. The Graduate Record Examinations (GRE) are used to help predict the performance of applicants to graduate schools. The examinations are designed so that the mean score is 550 and the standard deviation of scores is 100. A certain graduate school will only accept applicants whose score on the GRE is in the top 3%. What score does an applicant need to achieve to be in the top 3%?

(a) 550 or above

(b) 900 or above

(c) 600 or above

(d) 360 or above

(e) * 740 or above

Assuming that the scores follow a normal distribution, suppose the score in question is x. 3% of the scores need to be bigger than x, and the other 97% smaller. Looking up 0.9700 in Table A gives $z = 1.88$, so x has to be $550 + (1.88)(100) = 738$. The best alternative is 740.

38. A consumer group surveyed the prices for white cotton extra-long twin sheet sets in five different department stores and reported the mean price as \$16. We visited four of the five stores, and found the prices to be \$12, \$15, \$17, and \$22. Assuming that the consumer group is correct, what is the price of the item at the store that we did not visit?

(a) \$10

(b) none of the other answers is correct

(c) \$15

(d) \$17

(e) * \$14

If the mean of the 5 prices is \$16, the total is $5 \times \$16 = \$80$. The total of the 4 given prices is \$66, so the missing one is $\$80 - \$16 = \$14$.

39. For the list of numbers

$$13, 15, 12, 16, 34,$$

what is the median?

(a) 18

(b) * 15

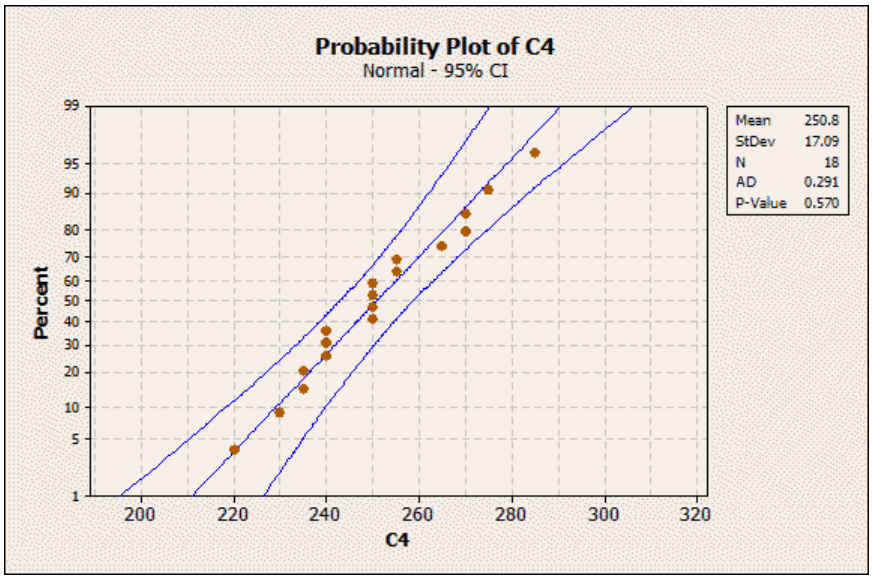
(c) 13

(d) 12

(e) 16

Don't forget to sort the values in order first!

40. On a college football team, the weights of all the players were recorded. One of the positions on a football team is Defensive Lineman. This college team had 18 defensive linemen, and a normal quantile plot of their weights is shown below.



What can you say about the distribution of weights of these defensive linemen?

(a) It is not described by a normal distribution because it is skewed to the left.

(b) It is not described by a normal distribution because it is skewed to the right.

(c) * It is described reasonably well by a normal distribution.

(d) There is a strong positive association between weight and percent of tackles made.

(e) There are outliers among the weights.

The plot is reasonably straight and stays within the outer lines. The endmost points are right on the centre line, so there are no outliers. The "percent" on this graph is the usual "percent" on a normal quantile plot: it has nothing to do with percent of tackles made!