# University of Toronto Scarborough
# STAB22 Term Test
# Yellow version

Olga Chilina Srishta Chopra

February 6, 2014

For this examination, you are allowed one handwritten letter-sized (8.5 × 11 inches) sheet of notes (both sides) prepared by you, a non-programmable, non-communicating calculator, and writing implements.

This question paper has 16 numbered pages of questions, with 2 blank pages and statistical tables at the back. Before you start, check to see that you have all the pages. You should also have a Scantron sheet on which to enter your answers. If any of this is missing, speak to an invigilator.

This examination is multiple choice. Each question has equal weight, and there is no penalty for guessing. To ensure that you receive credit for your work on the exam, fill in the bubbles on the Scantron sheet for your correct student number (under "Identification"), your last name, and as much of your first name as fits. If you do not fill in the bubbles for your name and student number, you risk getting a **zero** for the exam.

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.)

If you need paper for rough work, use the blank sheets at the back of this question paper. You may detach these pages or the tables.

Before you begin, two more things:

- 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.
- Complete the signature sheet, but *sign it only when the invigilator collects it.* The signature sheet shows that you were present at the exam.

At the end of the exam, you *must* hand in your Scantron sheet (or you will receive a mark of zero for the examination). You will be graded *only* on what appears on the Scantron sheet. You are responsible for making sure that an invigilator receives your Scantron. You may take away the question paper after the exam, but whether you do or not, anything written on the question paper will *not* be considered in your grade.

*The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.*

1. 75% of students scored below your score of 82 in Course A, and 35% of students scored above your score of 86 in Course B. In which course have you scored better?
   A. Course B. Because z-score value for Course B is greater than that of Course A.
   B. Course B. Because 86 in Course B is greater than 82 in Course A.
   C. Cannot say. Need more information about mean and standard deviation to conclude.
   D. Course A. Because z-score value for Course A is greater than that of Course B.
   **E. Course A. Because 75% of students scored below you which is better than 65% of students scoring below you in Course B.**

2. What is the 40th percentile for the given data: 1, 6, 3, 7, 5, 5, 11, 8, 6?
   A. 1 **B. 5** C. 6 D. 3 E. 7

   **Solution:** 1, 3, 5, 5, 6, 6, 7, 8, 11. The 40th percentile is 5.
   There are 9 data values, and $9(40\%) = 3.6$, so the 40th percentile should be between the 3rd and 4th values, which are both 5.

3. What is the main advantage of boxplots over stemplots and histograms?
   A. boxplots show skewed distributions, whereas stemplots and histograms show only symmetric distributions
   B. only boxplots can show outliers
   C. boxplots use the five-number summary, whereas stemplots and histograms use the mean and standard deviation
   D. boxplots show more detail about the shape of the distribution
   **E. boxplots make it easy to compare several distributions**

4. A company wants to analyze the impact of a new drug for cancer. It wishes to study the relationship of the gender of various individuals and their response to the drug as "positive" or "negative" by displaying as a scatterplot. Which of the following statements are true about such a scatterplot?
   A. The response depends on gender.
   B. The response remains unchanged with gender.
   C. The scatterplot is not possible for quantitative variables.
   **D. The scatterplot is not possible for categorical variables.**

   **Solution:** The two variables are categorical, so we cannot display them on a scatterplot.

5. For the given contingency table, which numbers represent the marginal distribution of the variable "burritos" (with values "ate" and "did not eat")?

| | Got GI illness | Did not get GI illness | Totals |
|---|---|---|---|
| Ate burritos | 8 | 5 | 13 |
| Did not eat burritos | 6 | 33 | 39 |
| Totals | 14 | 38 | 52 |

Question 5 continues...

A. 0.61, 0.38

B. 0.61, 0.15

**C. 0.25, 0.75**

D. 0.15, 0.10, 0.12, 0.63

E. 0.27, 0.73

**Solution:** $13/52 = 0.25, 39/52 = 0.75$.

6. A supermarket display of ground beef has 27 packages. The weights of the packages (in kg) have median 1.06, first quartile 0.92, and third quartile 1.18. The lightest package weighs 0.75 kg and the heaviest weighs 1.41 kg. Are the lightest and heaviest packages outliers, in terms of weight, compared to the others? Use the usual criterion based on the inter-quartile range.

   A. Both the lightest and heaviest packages are outliers.

   B. There is not enough information to decide whether these packages are outliers.

   C. The heaviest package is an outlier, but the lightest is not.

   D. The lightest package is an outlier, but the heaviest is not.

   **E. Neither the lightest nor the heaviest packages are outliers.**

**Solution:** Work out $R = 1.5 \times IQR = 1.5(1.18 - 0.92) = 0.39$. Q1 minus $R$ is $0.92 - 0.35 = 0.57$. The lightest package is not smaller than this, so it is not an outlier. Q3 plus $R$ is $1.18 + 0.39 = 1.57$. The heaviest package is not bigger than this, so it is not an outlier either.

7. A consumer group surveyed the prices for a certain item in five different stores, and reported the average price as \$15. We visited four of the five stores, and found the prices to be \$10, \$15, \$15, and \$25. Assuming that the consumer group is correct, what is the price of the item at the store that we did not visit, in dollars?

   A. 15 B. 20 C. 30 D. 25 **E. 10**

**Solution:** $(10 + 15 + 15 + 25 + x)/5 = 15$, thus $65 + x = 75$ or $x = 10$.

8. Sales of the most popular item at a company last year had median \$2200 and interquartile range \$370. Sales have changed from last year to this year by multiplying by 3 and adding 400 (dollars). Use this information for this question and the next one.

   What is the median of sales of this item this year?

   **A. 7000** B. 19800 C. 6600 D. 2600 E. 20200

**Solution:** For the median, multiply by 3 and add 400: $3(2200) + 400 = 7000$.

9. Refer to the information given in Question 8. What is the interquartile range of this year's sales?

   A. 1510 B. 3730 C. 3330 **D. 1110** E. 770

**Solution:** The interquartile range is a measure of spread, so just multiply by 3 to get $3(370) = 1110$.

10. In a study, a researcher finds that scores on an English vocabulary test, $x$, have mean 210 and standard deviation 15, and scores on a Math test, $y$, have mean 360 and standard deviation 120. The regression equation is found to be $\hat{y} = b_0 - 2x$, where the value $b_0$ is not known.

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

    What can you say about the correlation between $x$ and $y$?

    A. There is no association between the two variables.

    **B. The coefficient of correlation is $-0.25$ which is a weak and negative association.**

    C. The coefficient of correlation is $-0.52$ which is a strong and negative association.

    D. It is a very strong and positive association.

    E. The coefficient of correlation is 0.14 which is a weak and positive association.

**Solution:** Slope is $rs_y/s_x$, so $-2 = r(120/15)$, so $r = -2/8 = -0.25$.

11. Using the information in Question 10, calculate the value of $b_0$.

    A. -60

    B. 870

    **C. 780**

    D. 60

**Solution:** $b_0 = \bar{y} - b_1\bar{x} = 360 - (-2)(210) = 780$.

12. Using the information in Question 10, what can you say about the slope of the regression?

**A. For every one-point increase in the English vocabulary test score, the Math test score is estimated to decrease by 2.**

B. For every one-point increase in the Math test, the English vocabulary test score is estimated to increase by 2.

C. For every one-point increase in the Math test score, the English vocabulary test score is estimated to decrease by 2.

D. For every one-point increase in the English vocabulary test score, the Math test score is estimated to increase by 2.

**Solution:** Definition of slope, since English vocab test score is $x$ and Math test score is $y$, the response.

13. There are three children in a room, ages three, four, and five. If a four-year-old child enters the room which of the statistics will change and how?

A. mean age and standard deviation will stay the same

**B. mean age will stay the same but the standard deviation will decrease**

C. only median will change

D. mean age and standard deviation will increase

E. mean age will stay the same but the standard deviation will increase

14. Which one of the following variables is NOT categorical?

A. gender of a person

B. clothes size of a person (small, medium, large)

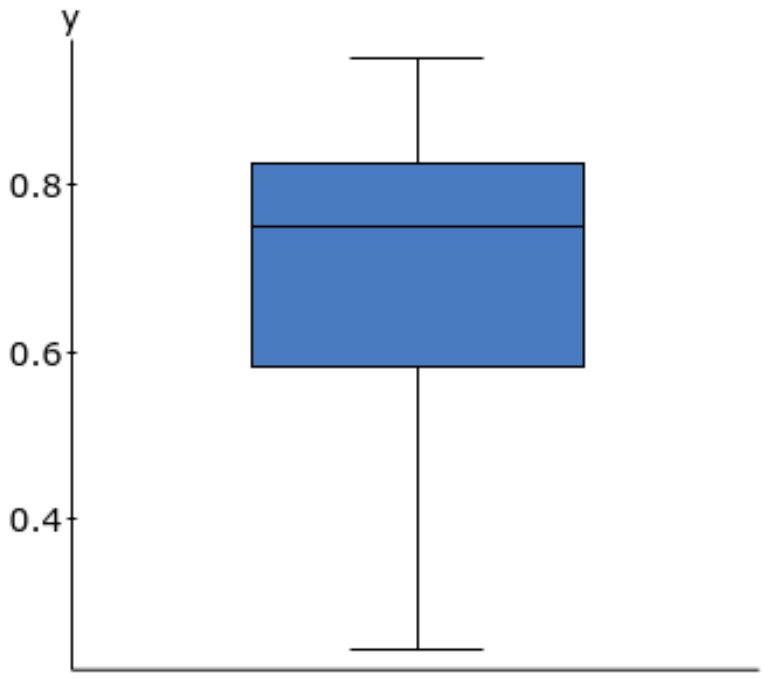
C. eye color of a person

D. marital status of a person

**E. age of a person**

**Solution:** Age is the only quantitative variable here.

15. What term would best describe the shape of the distribution in the boxplot shown below?




A. bimodal

**B. left-skewed**

C. normal

D. symmetric

E. right-skewed

16. On a STAB22 test, the middle 95% of students score between 46 and 82. The scores have a normal distribution. Calculate the mean score and the standard deviation of the scores. You may wish to use the 68-95-99.7 rule. In the choices below, the mean is given first in each case, and then the standard deviation.

A. 18, 23

B. 64, 18

**C. 64, 9**

D. 18, 32

E. 64, 32

**Solution:** The difference between bottom and top is $82 - 46 = 36$. This is *four* times the standard deviation, since the rule says to go up *and* down 2 SDs. So the SD is 9. The mean is in the middle of the two values given, $(82 + 46)/2 = 64$.

17. The United Nations keeps records of illiteracy rates in its member countries. Below is a stemplot of the percent of adult males who are illiterate in each of 142 countries. The rates are in percent, with the stems being units and the leaves being tenths of a percent.

```
0| 00000000000000111111111122223344444444
0| 5555556666777778888999 9
1| 0000111123344
1|5677788899
2| 0000111234
2| 556677899
3| 00011344
3| 6777888
4| 024
4| 56668
5| 0111234
6| 002
6|
7| 1
7| 9
```

The mean of this distribution is

A. very close to the median

B. clearly less than the median

C. very close to the first quartile

D. cannot say because the mean is random

**E. clearly greater than the median**

**Solution:** The distribution is right-skewed, so the mean is greater than the median.

18. In a study, 74% of the variation in the response variable, $y$ is explained by the least squares regression line of the response variable, $y$ on the explanatory variable, $x$. There is an upward trend in the relationship between $x$ and $y$. Which of the following statements is true?

A. Correlation coefficient is 0.74 and the coefficient of determination is 0.86

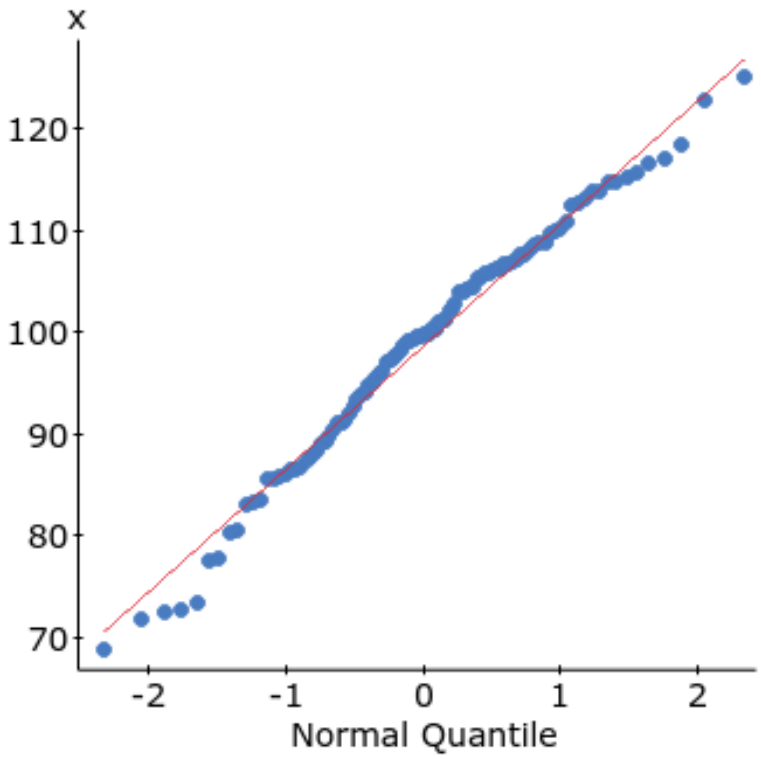
B. Coefficient of determination is 0.74 and the correlation coefficient is $-0.86$

C. Correlation coefficient is 0.74 and the coefficient of determination is $-0.86$

**D. Coefficient of determination is 0.74 and the correlation coefficient is 0.86**

**Solution:** The 74% is R-squared, which is the same as the coefficient of determination. The correlation is whatever value that, when squared, would give 0.74, ie. $\sqrt{0.74} = 0.86$. It must be the positive square root, not the negative one, because we are told that the trend is upward.

19. The plot below was obtained for some data $x$:




Which one of the following statements is the best conclusion from the plot?

**A. The data follow an approximately normal distribution.**

B. The data follow a distribution that is skewed to the left.

C. The data follow a distribution that is skewed to the right.

D. The line of best fit is quite good.

E. The data have a non-linear relationship.

20. We are conducting a clinical study. The variables of interest are patients' age, weight, height, and whether or not patients smoke. Which of the following graphical displays would be most appropriate to show the distribution of the smoking variable?

A. stemplot

B. histogram

C. boxplot

**D. pie chart**

E. scatterplot

**Solution:** The smoking variable is categorical with values "yes" or "no", and the other choices are for quantitative variables.

21. A social skills training program was implemented for seven students with mild disabilities. Each student was assessed for social skills, both before the program ("pre") and after ("post"). The pre-program scores had mean 96.7 and standard deviation 9.3; the post-program scores had mean 102.7 and standard deviation 11.1. The correlation between pre and post scores was 0.76. Use this information for this question and the next one.

What is the *slope* of the regression line for predicting post-program score from pre-program score?

A. 1.2 B. 0.6 **C. 0.9** D. 0.8

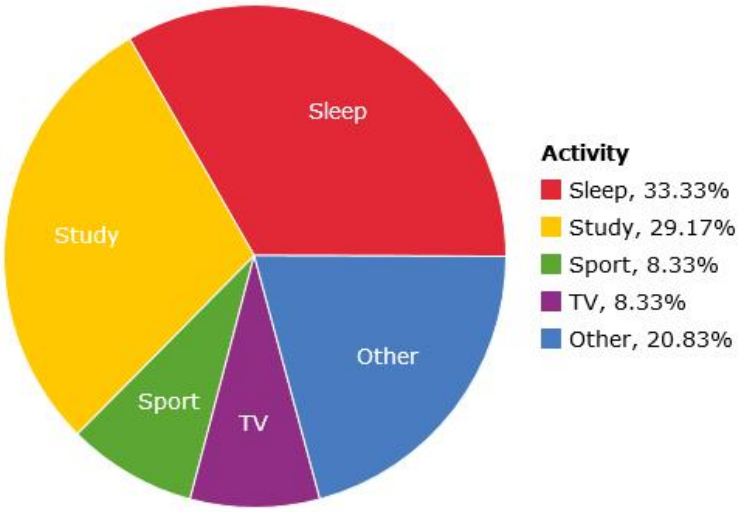
**Solution:** Use the formula: slope is $rs_y/s_x = (0.76)(11.1/9.3) = 0.907$.

22. Refer back to the information in Question 21. What is the *intercept* of the regression line for predicting post-program score from pre-program score?

A. 31 **B. 15** C. 3.5 D. 41

**Solution:** $\bar{y} - b_1\bar{x} = 102.7 - 96.7(0.907) = 14.983$.

23. John recorded the amount of time he spends on different activities over a twenty-four-hour period and drew the pie chart given below.

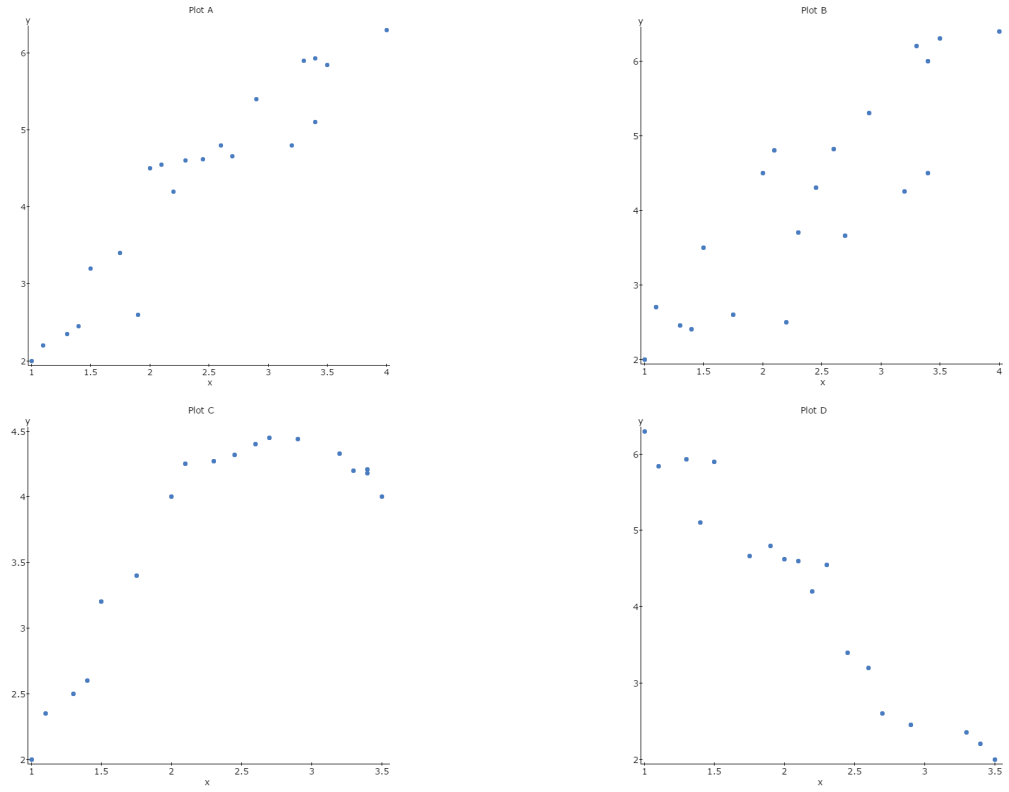



Approximately how many hours per day does John sleep?

A. 7 **B. 8** C. 10 D. 6 E. 9

**Solution:** 33.33% of 24 hours is 8

24. Following are the scatterplots from four different studies. Use the information here to answer this question and the next question.




Which of the plots above show *positive* correlation?

A. Only plots A and B

B. Only plots A and C

**C. Only plots A, B and C**

D. Only plot D

**Solution:** D is going downhill but all the rest are going uphill.

25. Refer to the plots in Question 24. Which of the plots would be described by a straight line with positive slope?

A. Only plot D

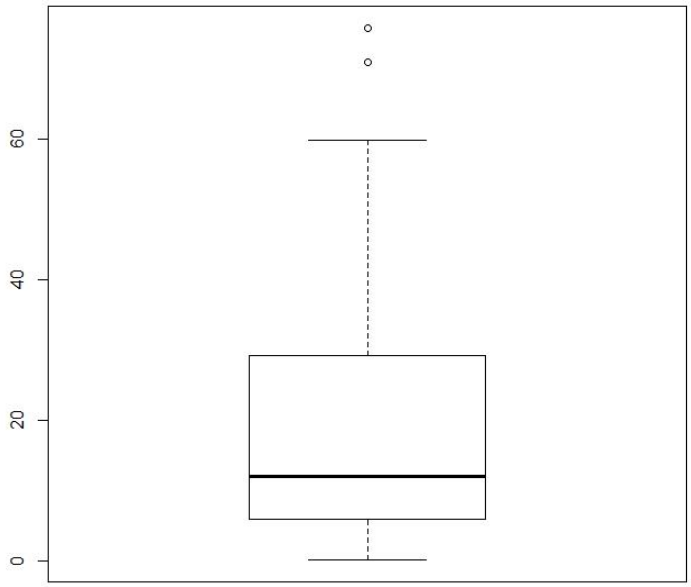
**B. Only plots A and B**

C. Only plots A, B and C

D. Only plots A and C

**Solution:** C is upward, but a curve, leaving only A and B as upward straight trends.

26. A (modified) boxplot is shown below.




For the data in the boxplot, which statement is **TRUE**?

A. removing the two points outside the upper fence would increase the median

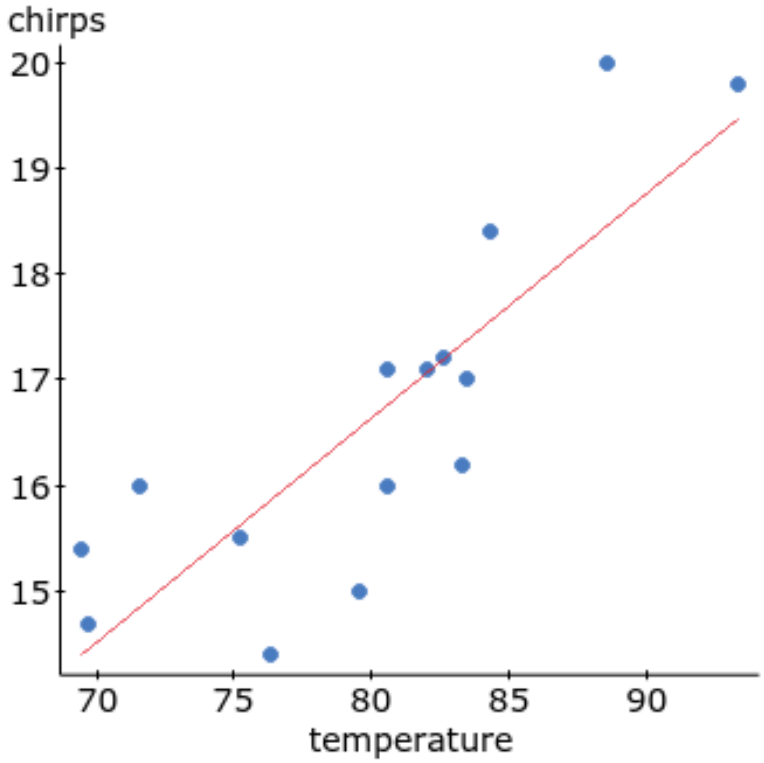
B. removing the two points outside the upper fence would increase the mean

C. more than one of the other statements is true

**D. the third quartile is approximately 30**

E. the IQR is greater than 40

27. Crickets are insects that make a chirping sound by rapidly sliding one wing over another. Scientists believe that crickets will chirp more rapidly when it is warmer. Some data were obtained of chirping rate at 15 different temperatures, as shown below. The variable `chirps` is the number of chirps per second.




Use this information for this question and the next one.

What do you learn from the plot?

**A. Crickets tend to chirp more rapidly when it is warmer.**

B. There is no relationship between chirping rate and temperature.

C. Crickets tend to chirp less rapidly when it is warmer.

D. There is an almost perfect straight-line relationship between chirping rate and temperature.

E. There is a relationship between chirping rate and temperature, but it is not a straight line.

28. Look again at the plot in Question 27. Which one of the numbers below is closest to the correlation between the chirping rate and the temperature?

**A. 0.8** B. −0.1 C. 0.3 D. −0.7 E. 0.95

**Solution:** The actual correlation is 0.83. (0.3 as a correlation looks almost non-existent, and this correlation definitely is stronger than that.)

29. What is the five-number summary for the given data set: 2, 4, 22, 6, 1, 4, 1, 5, 7, 4?

A. 1, 4, 5, 7, 22

B. 2, 22, 3, 5, 4

**C. 1, 2, 4, 6, 22**

D. 1, 2, 4, 6, 7

Question 29 continues...

E. 22, 6, 4, 2, 1

**Solution:** Arrange the numbers in order to get 1, 1, 2, 4, 4, 4, 5, 6, 7, 22. There are 10 values altogether. The median is halfway between the 5th and 6th values, which are both 4. The smallest value is 1 and the largest is 22. Only one of the choices matches this (and you can check that the quartiles, found the way described in the text, are 2 and 6). The five-number summary has to be given with the smallest number first.

30. A survey asked people how often they exceed speed limits. The data are then categorized into the following contingency table of counts showing the relationship between age group and response.

| | Exceed Limit if Possible? | | |
|---|---|---|---|
| **Age** | **Always** | **Not Always** | **Total** |
| **Under 30** | 100 | 100 | 200 |
| **Over 30** | 40 | 160 | 200 |
| **Total** | 140 | 260 | 400 |

Use this information for this question and the next one.

Among people with age over 30, how likely is a person chosen at random to always exceed the speed limit?

A. 0.10 **B. 0.20** C. 0.25 D. 0.35 E. 0.50

**Solution:** $40/200 = 0.20$

31. Refer to the table in Question 30. What is the chance of encountering a person who always exceeds the speed limit?

A. 0.20 B. 0.25 C. 0.10 **D. 0.35** E. 0.50

**Solution:** $140/400 = 0.35$

32. A researcher studies students in an elementary school and finds a strong positive linear association between their heights and the quality of the decisions they make. Which of the following statements best describe the association?

A. Taller heights cause wise decisions.

**B. The observed association can be explained by a lurking variable.**

C. There is an error in the study. The observed association cannot be explained.

D. Heights and wisdom are confounded variables.

**Solution:** The lurking variable in question might be age: older children would tend to be both taller and able to make better decisions.

33. A variable $x$ has a distribution with median 5, first quartile 3 and third quartile 6. The variable $y$ is calculated from $x$ by the formula $y = 30 - 2x$. What is the *first quartile* of $y$?

A. 24 **B. 18** C. 3 D. 6 E. 20

**Solution:** Multiplying by a negative number will switch the order around, so that the new first quartile will be the transformed *third* quartile. This is easy enough to check: the transformed quartiles and median are $30 - 2(3) = 24$, $30 - 2(5) = 20$ and $30 - 2(6) = 18$, so the new first quartile had better be 18.

34. Which one of these statistics is **not** affected by outliers?

A. standard deviation

**B. interquartile range**

C. mean

D. correlation

E. range

**Solution:** The interquartile range is based in the quartiles, which are not affected by outliers.

35. In the given data set: 2, 4, 22, 6, 1, 4, 1, 5, 7, 4, the value 22 is an outlier. Which of the statistics below would change if we replaced it with the value 8?

A. minimum value

B. third quartile

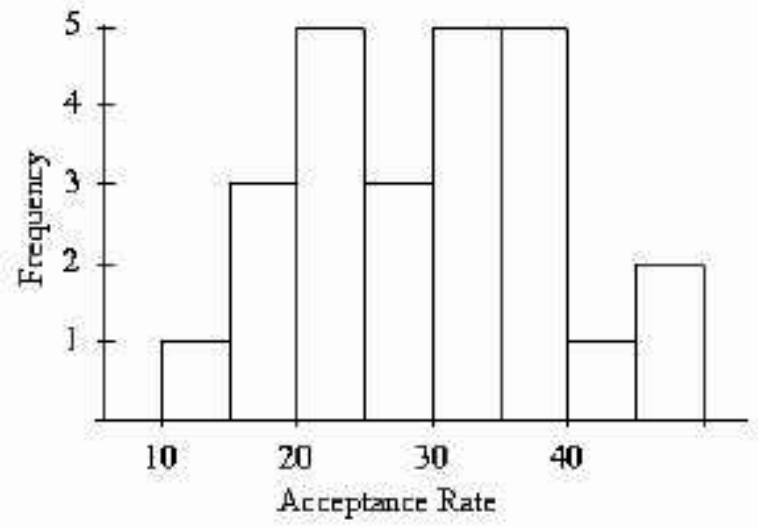
C. interquartile range

D. median

**E. mean**

**Solution:** All the other choices are not affected by outliers. The mean would become noticeably smaller.

36. The following histogram represents the distribution of acceptance rates (percent accepted) among 25 business schools in 1997. In each class interval, the left endpoint is included but not the right. Which interval contains fewer than half of all the observations?




A. (20%, 35%)

B. (30%, 50%)

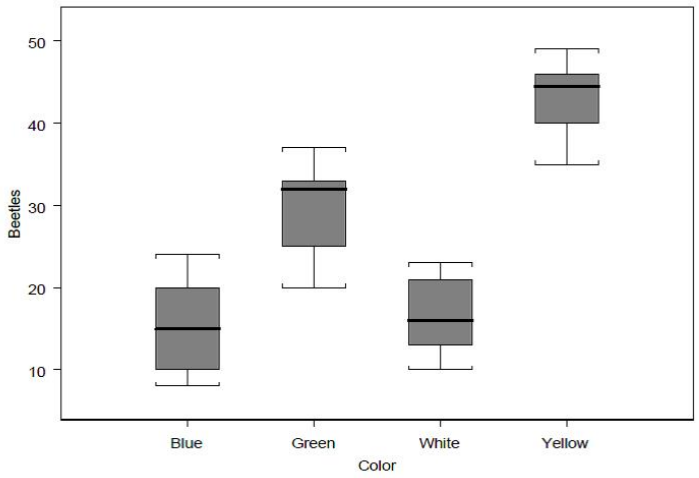
**C. (0%, 30%)**

D. none of the other alternatives

E. (25%, 40%)

**Solution:** There are $1 + 3 + 5 + 3 = 12$ observations between 0 and 30, which is just less than half of the 25 observations. All the other actual alternatives contain 13 observations, which is more than half.

37. The following graph depicts the number of beetles captured on boards of different colors. Which statement is **FALSE**?




A. the median for the distribution of beetles captured on blue boards is close to the mean of the same distribution

B. the shape of the distribution of beetles captured on yellow boards is skewed to the left

**C. the shape of the distribution of beetles captured on green boards is skewed to the right**

D. the median for the distribution of beetles captured on white boards is close to the median for the distribution of beetles captured on blue boards

E. the interquartile range for the distribution of beetles captured on blue boards is close to 10

38. As part of its quality control program, the Autolite Battery Company conducts tests on battery life. For a particular D cell alkaline battery, the mean useful life is 19 hours. The useful life of the battery follows a normal distribution with a standard deviation of 1.2 hours. Use this information to answer this question and the next question.

What proportion of batteries have a useful life greater than 22.5 hours?

A. 0.292 **B. 0.002** C. 0.922 D. 0.998

**Solution:** $z = (22.5-19)/1.2 = 2.93$. Proportion less is 0.9983, so proportion greater is $1-0.9983 = 0.0017$.

39. Using the information in Question 38, what proportion of batteries have a useful life between 16 and 20.5 hours?

A. 0.901 B. 0.375 C. 0.894 **D. 0.888**

**Solution:** $z$ for 16 hours is $z = (16 - 19)/1.2 = -2.5$, and for 20.5 is $z = (20.5 - 19)/1.2 = 1.25$. In table, these give respectively 0.0060 and 0.8944, so proportion between is difference, $0.8944 - 0.0060 = 0.8884$.

40. Just before the Canadian penny was taken out of circulation, a bank employee recorded the ages of 50 pennies, in years. The results are shown in the stemplot below.

```
Variable: penny

Decimal point is 1 digit(s) to the right of the colon.
Leaf unit = 1
0 : 00000000000011111122233344
0 : 55556899
1 : 0
1 : 677999
2 : 00123
2 : 558
3 :
3 : 6
```

You may care to note that there are 26 pennies in the first row of the boxplot and 8 pennies in the second row.

What is the *median* age of these pennies, in years?

A. 0.4 B. 40 **C. 4** D. 13 E. 1.3

**Solution:** There are 50 pennies altogether, so the median is the average of the 25th and 26th penny ages: that is, of the last two values in the first line of the stemplot. So the median age is 4 years.