

University of Toronto Scarborough STAB22 Midterm Examination

October 2011

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 20 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.

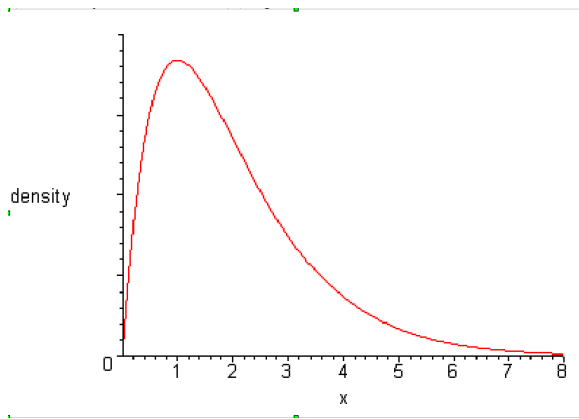
This examination is multiple choice. Each question has equal weight. On the Scantron answer sheet, ensure that you enter your last name, first name (as much of it as fits), and student number (in “Identification”).

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.

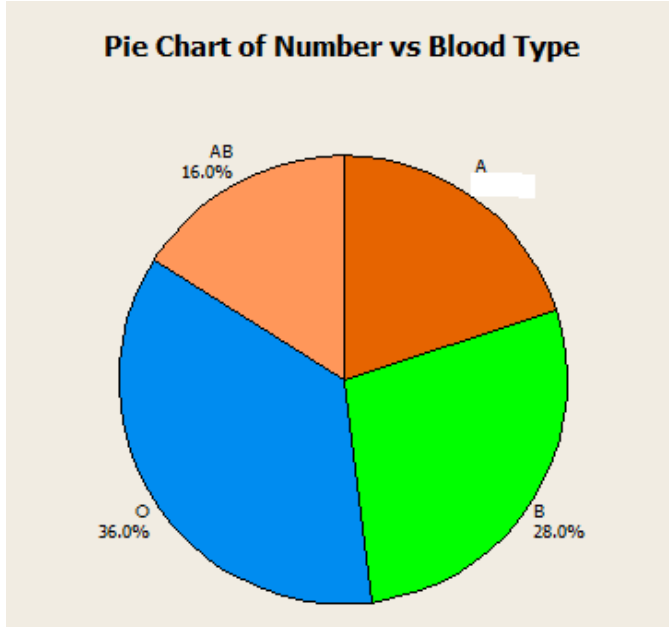
1. The density curve of a variable X is given below:



Five statements are given below about the density curve shown to the left. Each statement is either true or false. Only one of the statements is true. Which one?

- (a) The median of the distribution is 4.
- (b) The distribution is left-skewed.
- (c) * The mean of the distribution is greater than the median.
- (d) The third quartile of the distribution is 6.
- (e) The total area under the density curve is greater than 1.

2. A random sample of 25 blood donors was given a blood test to determine their blood type. The pie chart below, displays 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? (Note this question requires the number of donors):

- (a) * 5
- (b) 20
- (c) 3
- (d) 4
- (e) 2

3. The table below shows the population of each province and territory of Canada, showing also the aboriginal population in each case. The aboriginal population is divided into North American Indian, Métis, Inuit and “other” (not shown). Use the table for this question and the three following.

▼ Name ▲	Total population ▼ ▲	Aboriginal population ¹ ▼ ▲	North American Indian ▼ ▲	Métis ▼ ▲	Inuit ▼ ▲	Non-Aboriginal population ▼ ▲
Canada !	29,639,030	976,305	608,850	292,305	45,070	28,662,725
Newfoundland and Labrador	508,080	18,775	7,040	5,480	4,560	489,300
Prince Edward Island	133,385	1,345	1,035	220	20	132,040
Nova Scotia	897,565	17,010	12,920	3,135	350	880,560
New Brunswick	719,710	16,990	11,495	4,290	155	702,725
Quebec !	7,125,580	79,400	51,125	15,855	9,530	7,046,180
Ontario !	11,285,545	188,315	131,560	48,340	1,375	11,097,235
Manitoba !	1,103,700	150,045	90,340	56,800	340	953,655
Saskatchewan !	963,155	130,185	83,745	43,695	235	832,960
Alberta !	2,941,150	156,225	84,995	66,060	1,090	2,784,925
British Columbia !	3,868,875	170,025	118,295	44,265	800	3,698,850
Yukon Territory	28,520	6,540	5,600	535	140	21,975
Northwest Territories !	37,100	18,730	10,615	3,580	3,910	18,370
Nunavut !	26,665	22,720	95	55	22,560	3,945

The “!” next to some of the province/territory names above are of no significance.

What percentage of Canadians are Inuit from Nunavut?

- (a) * 0.1
 - (b) 2 or more
 - (c) 0.001 or less
 - (d) 1
4. Refer to the table in Question 3. What percentage of Aboriginal people are Inuit from Nunavut?
- (a) 20
 - (b) * 2
 - (c) more than 30
 - (d) less than 0.01
 - (e) 0.02

5. Refer to the table in Question 3. What percentage of the population of Nunavut is Inuit?

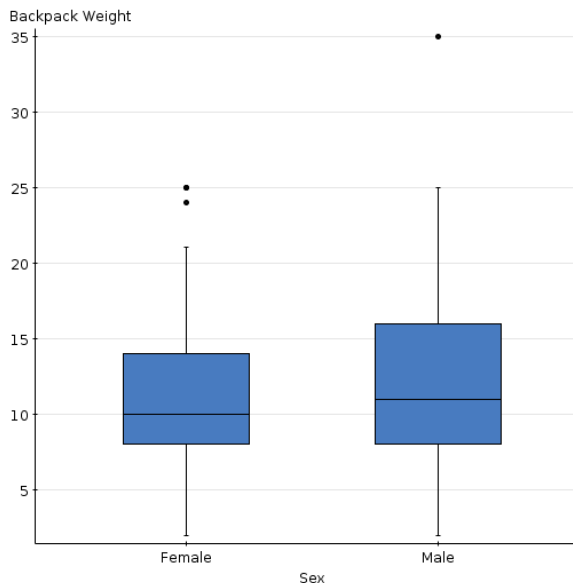
- (a) 50
- (b) 70
- (c) 20
- (d) 10
- (e) * 85

6. Refer to the table in Question 3. What percentage of Inuit are from Nunavut?

- (a) * 50
- (b) 85
- (c) 20
- (d) 10
- (e) 70

7. 100 backpackers went on a group hike. For each backpacker, their body weight was recorded, along with the weight of their backpack and whether they were male or female.

A pair of boxplots is shown below. These show the distribution of backpack weights for males and females.



What is the *most important* difference between backpack weights for males and females?

- (a) * The backpack weights for males have a greater spread for males than for females.
- (b) The distribution of male backpack weights is more skewed to the left than for females.
- (c) The distributions of backpack weights differ substantially in centre.
- (d) There is a substantial difference in the number of outliers between males and females.

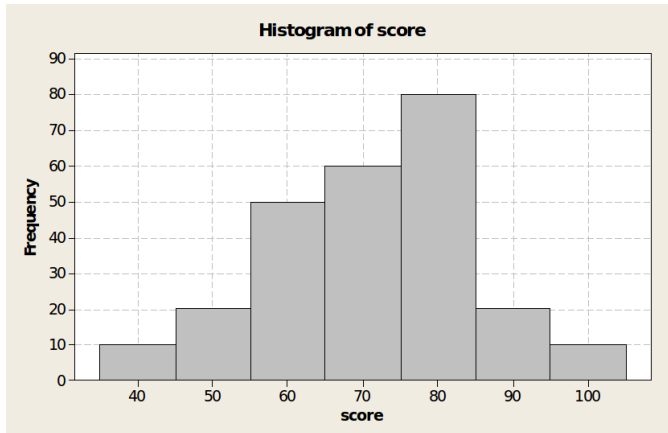
8. The summary statistics of the annual salaries (in thousands of dollars) of a group of 100 employees in a large company are given below:

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
salary	100	45.00	19.00	20.00	29.00	43.00	53.34	112.00

If all employees in this group receive a 15% increase in salary, what will be the IQR of the new salaries, in thousands of dollars?

- (a) 4
 - (b) * 28
 - (c) 39
 - (d) 24
 - (e) 365
9. A company that packages snack foods does its quality control by selecting 10 cases from each day's production, and opening two bags from each case and inspecting the contents. What kind of sampling procedure is this?
- (a) * multi-stage sample
 - (b) simple random sample
 - (c) systematic sample
 - (d) voluntary-response sample
 - (e) stratified sample

10. The histogram below displays the scores of a group of students in an examination. You may assume that no student scored exactly at the class boundaries of the histogram below.



What percentage of students scored below 55? You may assume that no student obtained a score exactly at the class boundaries shown on the histogram.

- (a) 30
(b) * 12
(c) 20
(d) 50
(e) 8
11. A baseball league tests players to see whether they are using performance-enhancing drugs. Officials select a team at random, and a drug-testing crew shows up unannounced at a training session and tests a randomly chosen 10 players. Use this information for this question and the next one.
- What kind of sampling method is this?
- (a) * Multi-stage sample
(b) Stratified sample
(c) Systematic sample
(d) Simple random sample
(e) Convenience sample
12. Question 11 described a baseball league's drug testing procedure. Why do you think this kind of sampling method was used?
- (a) It was not convenient to obtain a list of all registered players in the baseball league.
(b) It would give more accurate results than other methods.
(c) It was simpler to understand than other methods.
(d) * It was more convenient than other methods.

13. A web site had a survey: “Do you ever use emoticons when you type online?”. (The web site had other content as well.) Of the 87,262 respondents, 27% said that they did not. Do you think this value 27% is a good estimate of the fraction of all people who use emoticons? Why?
- (a) No, because there is always sampling variability.
 - (b) * No, because this is not a random sample.
 - (c) Yes, because a voluntary-response sample was used.
 - (d) It’s a good estimate of the fraction for *all visitors to that web site*.
 - (e) Yes, because the sample is large.

14. A school teacher plans to have some of his students make a poster about a Canadian province or territory. The teacher makes a list of the provinces and territories and numbers them as below:

01 Alberta	Use the random digits below to choose four different provinces and territories for the four students who will make posters. Which is the fourth province or territory chosen? (Note that you do not need Table B for this.)
02 British Columbia	
03 Manitoba	
04 New Brunswick	
05 Nfld & Labrador	
06 Northwest Territories	
07 Nova Scotia	
08 Nunavut	
09 Ontario	
10 Prince Edward Island	
11 Quebec	
12 Saskatchewan	
13 Yukon	

88063 56513 31056 32105 08993

- (a) Nfld & Labrador
- (b) The list of random digits is not long enough
- (c) Prince Edward Island
- (d) Northwest Territories
- (e) * Some province or territory not given in the other alternatives

15. A 2008 real estate report listed the asking price (in thousands of dollars) and size (in square feet) of condos under 1500 square feet in downtown Toronto. A regression analysis gives the predicted price \hat{y} in terms of the size x as $\hat{y} = 49.30 + 0.37x$. Use this information for this question and the next one.

How would you interpret the value 0.37?

- (a) a condo that costs nothing would have about 0.37 square feet.
 - (b) a condo with one more square foot would cost about \$0.37 more
 - (c) a condo that is 0 square feet in size would cost about \$370.
 - (d) a condo that costs 1 thousand dollars more would have about 0.37 more square feet
 - (e) * a condo with one more square foot would cost about \$370 more
16. In Question 15, some information was given about square footage and asking prices of condos in downtown Toronto. What asking price would you predict for a 1200 square foot condo in this market?
- (a) * \$490,000
 - (b) less than \$100,000
 - (c) more than \$3,000,000
 - (d) \$790,000
 - (e) \$370,000

17. Data on two variables x and y are shown below.

Row	x	y
1	4	14
2	1	25
3	2	17
4	7	9
5		

The correlation between x and y is very close to which of the values shown on the right?

- (a) 0.7
- (b) * -0.9
- (c) 0
- (d) 0.9
- (e) 0.5

18. The Program for International Student Assessment reported average scores on a standardized math test for students in 32 different (industrialized) nations. A five-number summary and a stemplot are shown below:

Summary statistics:

Column	n	Min	Q1	Median	Q3	Max
Ave Score	32	416	489	500	520	558

Variable: Ave Score

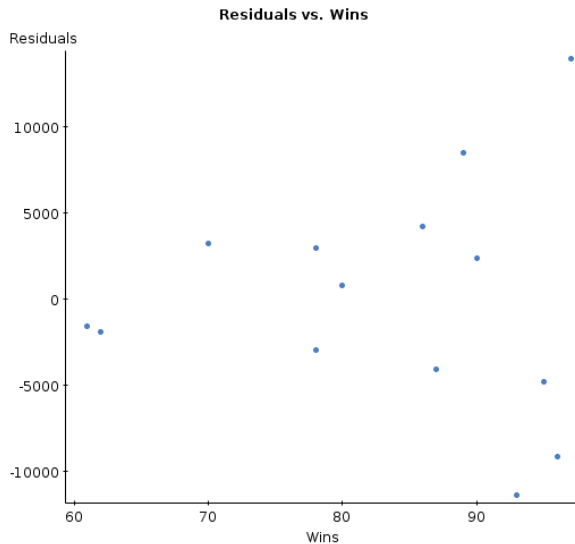
Decimal point is 2 digit(s) to the right of the colon.

```
4 : 12
4 : 6677889999
5 : 000000011112222333
5 : 55
```

How many outliers are there, using the usual rule?

- (a) 3
 - (b) 1
 - (c) 0
 - (d) * 2
 - (e) 4 or more
19. 500 students wrote an exam. The mean time to finish the exam was 150 minutes, the standard deviation was 15 minutes, and the distribution of the time taken to finish the exam was normal. Approximately **how many** students took between 135 and 165 minutes to finish the exam?
- (a) * 340
 - (b) 475
 - (c) 95
 - (d) 200
 - (e) 68

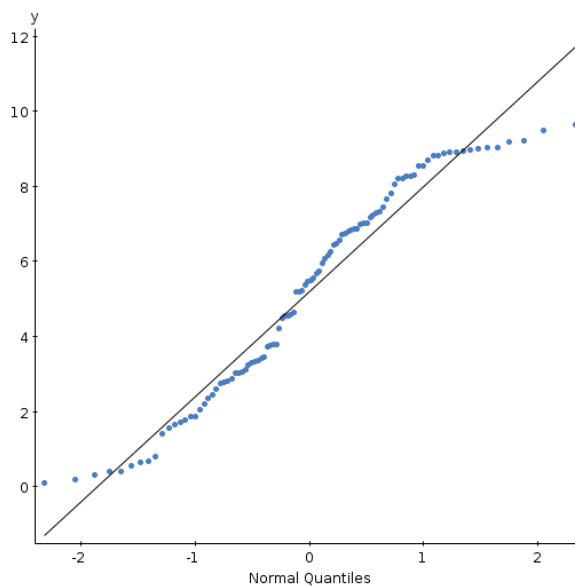
20. A study was made of whether average home attendance was higher for baseball teams that had more wins over the season. A regression was carried out predicting the average attendance from the number of wins for each team. A plot was made of the residuals from this regression against the number of wins, as shown below:



What do you conclude from this plot?

- (a) The relationship between number of wins and attendance is actually curved, not a straight line.
- (b) There is little or no relationship between the number of wins and attendance.
- (c) The residuals should have a normal distribution, and they do not.
- (d) There are no problems with this residual plot.
- (e) * The predictions become less accurate as the number of wins increases.

21. A normal quantile plot of some data y is shown below.



In what way do these data fail to have a normal distribution?

- (a) * The values at the extremes of the distribution are closer together than a normal distribution.
- (b) The data are skewed to the right.
- (c) Any apparent failure to have a normal distribution is random variation.
- (d) The data are skewed to the left.
- (e) The values at the extremes of the distribution are farther apart than a normal distribution.

22. In a regression for predicting a variable y from another variable x , the means and SDs of x and y are as shown:

	x	y
Mean	4	50
SD	0.8	15

The least-squares regression line for predicting y from x was $\hat{y} = -10 + 15x$. What must be the correlation between x and y ?

- (a) -0.2
 - (b) * 0.8
 - (c) 0
 - (d) 0.5
 - (e) 1
23. A heptathlon contest has a number of track and field events. We focus on the long jump and shot put at one contest. The long jump distances had a mean of 6.16 metres and an SD of 0.23 metres; the shot put distances had a mean of 13.29 metres and an SD of 1.24 metres. Assume that distances achieved in both events are normally distributed.
- An athlete long-jumps 6.78 metres and puts the shot 14.77 metres. Which of the two performances is better relative to the competition?
- (a) Both events represent the same performance
 - (b) The shot put, because the distance is longer
 - (c) * The long jump
 - (d) The shot put, but not just because the distance is longer

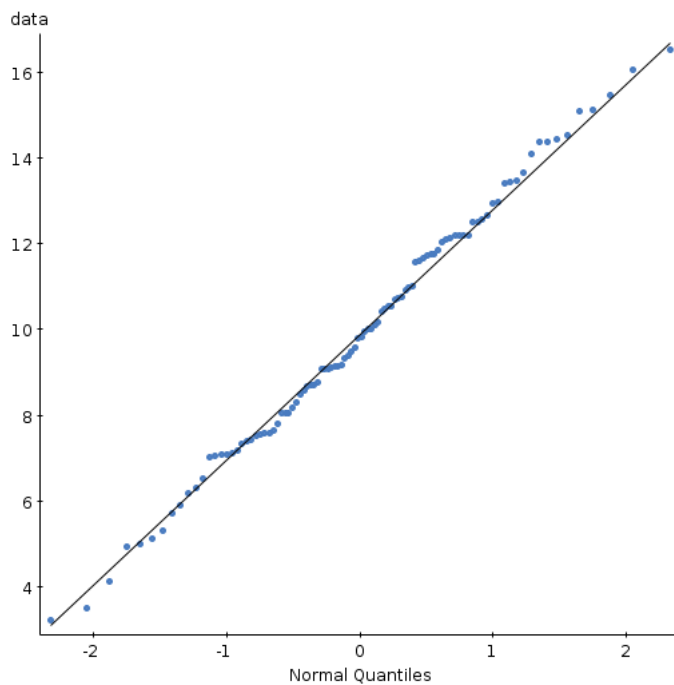
24. The summary statistics of the IQ scores of a group of students are given below.

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
iq	79	110.00	13.00	72.00	103.00	110.00	118.00	136.00

What *percentage* of students in this group scored over 103? (You may assume that no two students in this group have the same IQ score.)

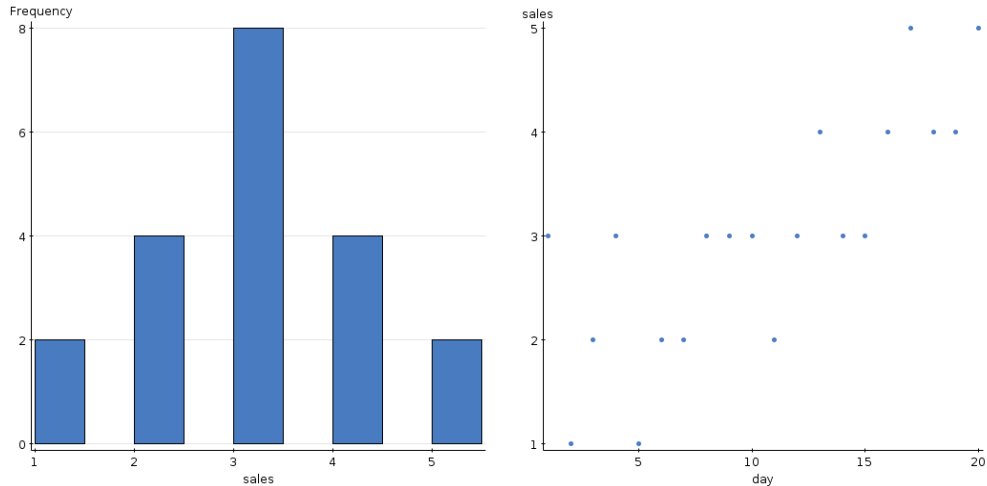
- (a) 25
 - (b) 50
 - (c) * 75
 - (d) 60
 - (e) 20
25. The Dutch are among the tallest people in the world. Heights of Dutch men follow a normal distribution with mean 184 cm and SD 9 cm. What percentage of Dutch men will be over 2 metres (200 cm) tall?
- (a) less than 1% but more than 0.1%
 - (b) less than 0.1%
 - (c) about 5%
 - (d) 10% or more
 - (e) * between 1% and 3%
26. A researcher is planning to take a simple random sample of 100 people out of a population of 1 million people, to estimate the population mean. Which of the following modifications to the sampling procedure would lead to a more accurate estimation?
- (a) use a smaller sample
 - (b) sample from a larger population
 - (c) * use a larger sample
 - (d) sample from a smaller population
 - (e) use a voluntary-response sample

27. Your instructor received some data whose nature is a closely-guarded secret. A normal quantile plot was drawn, as shown below. What should your instructor conclude about the distribution of the data from this plot?



- (a) Symmetric but not normal
- (b) Skewed to the right
- (c) * Normal
- (d) Skewed to the left

28. Owners of a new coffee shop kept track of sales (in hundreds of dollars) in the first 20 days after opening. They made a histogram of sales, and a scatterplot of sales against days (since opening). These are shown below.



What conclusion can you draw from the scatterplot but *not* from the histogram?

- (a) The distribution of sales is skewed.
 - (b) Sales appear to be decreasing over time.
 - (c) Sales have approximately a normal distribution.
 - (d) There is a curved relationship between sales and days.
 - (e) * Sales appear to be increasing over time.
29. Tests on 11 brands of fast-food chicken sandwiches revealed a more or less linear relationship between fat and calories. Some summary statistics were calculated, as follows:

	Fat (grams)	Calories
Mean	20.6	472.7
SD	9.8	144.2

The correlation between fat and calorie content for the 11 brands is 0.947.

Calculate the *intercept* of the least-squares regression line for predicting calories from fat. What do you get?

- (a) 0
- (b) * 190
- (c) -10
- (d) 100
- (e) 15

30. In a Canadian federal election, a ballot paper where it is not clear which candidate the voter intended to vote for is called “spoiled”. There were 34 ridings in British Columbia in the 2000 federal election. The *percentage* of spoiled ballots was recorded. Two numerical summaries of the data are shown below, and a histogram is shown below that.

Use this information for this question and the two following.

Summary 1:

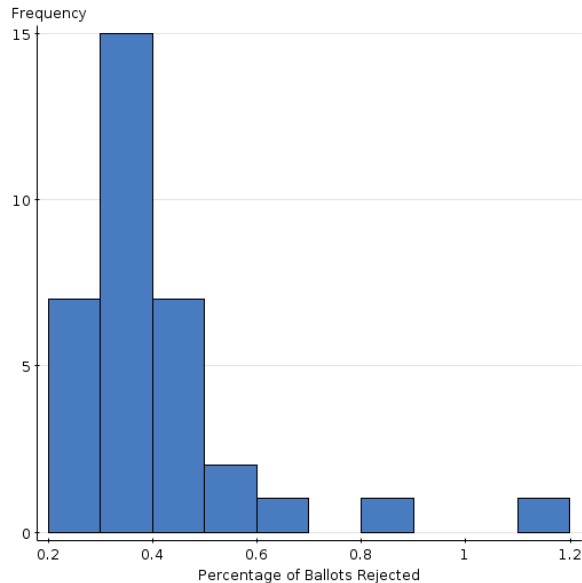
Summary statistics:

Column	Min	Q1	Median	Q3	Max
Percentage of Ballots Rejected	0.24	0.31	0.34	0.44	1.1

Summary 2:

Summary statistics:

Column	Mean	Std. Dev.
Percentage of Ballots Rejected	0.40529412	0.17706762



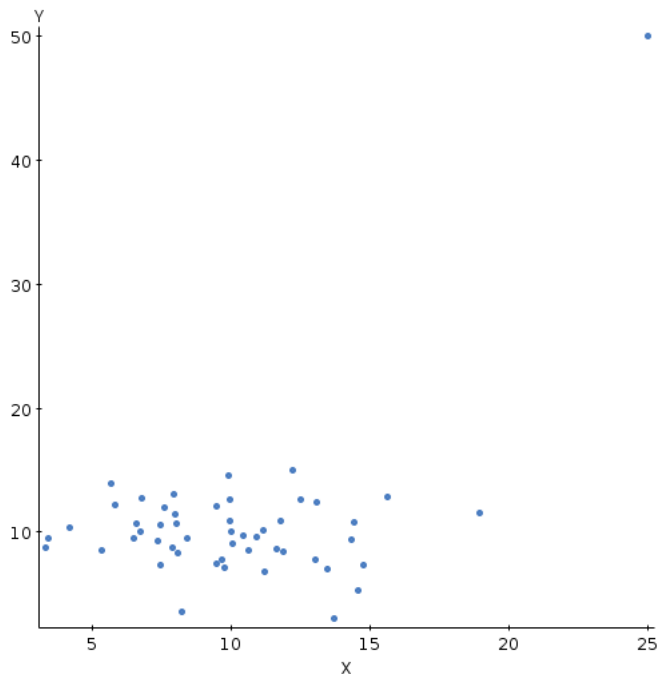
Which of the two summaries is more appropriate?

- (a) Summary 2
- (b) They are both equally good
- (c) * Summary 1
- (d) Neither of them should be used

31. Question 30 concerned the percentage of spoiled ballots by riding in British Columbia in 2000. The percentage of spoiled ballots in Victoria was 0.37%. Suppose this had been incorrectly recorded as 3.70%. What effect would this have on the summary statistics?
- (a) The median and IQR would change substantially, while the mean and SD would barely change at all.
 - (b) The data would become less spread out, so the IQR and SD would both decrease.
 - (c) * The mean and SD would change substantially, while the median and IQR would barely change at all.
 - (d) The mean and median would change substantially, while the SD and IQR would not change at all.
 - (e) Something would happen that is not described in the other alternatives.
32. Suppose a boxplot had been drawn of the data in Question 30. The upper whisker would extend to what value? (You may assume that outliers are plotted separately on the boxplot.)
- (a) 1.1
 - (b) 0.24
 - (c) 0.44
 - (d) * between 0.44 and 0.635, but it is impossible to tell exactly what without seeing the data values
 - (e) 0.635
33. The distribution of the heights of students in a large class is approximately normal with a mean height of 67 inches. Approximately 95% of the heights are between 61 and 73 inches. What, approximately, is the standard deviation of the distribution of heights, in inches?
- (a) 9
 - (b) * 3
 - (c) 6
 - (d) 12
 - (e) 2

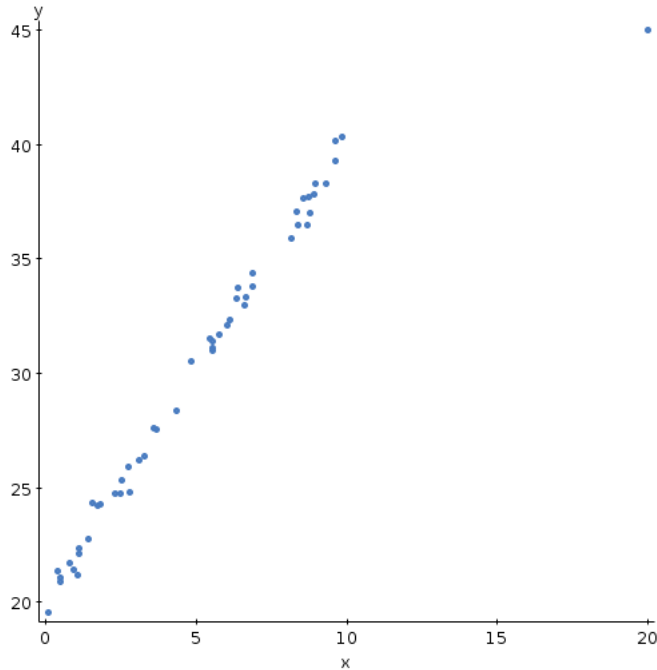
34. Hens usually begin laying eggs when they are six months old, but the eggs they produce are often too small to sell. The weights of the eggs (from hens of this age) have a normal distribution with mean 50.9 grams and SD 3.7 grams. What is the weight x such that the heaviest 20% of eggs laid by these hens are heavier than x ?
- (a) 66 grams or more
 - (b) 62 grams
 - (c) * 54 grams
 - (d) 58 grams
 - (e) 50 grams or less

35. In this question and the three questions that follow it, you will see a scatterplot showing a cluster of points and one “stray” point. In each question, you are given a number of statements about the association with the stray point. Mark the most correct one in each case.



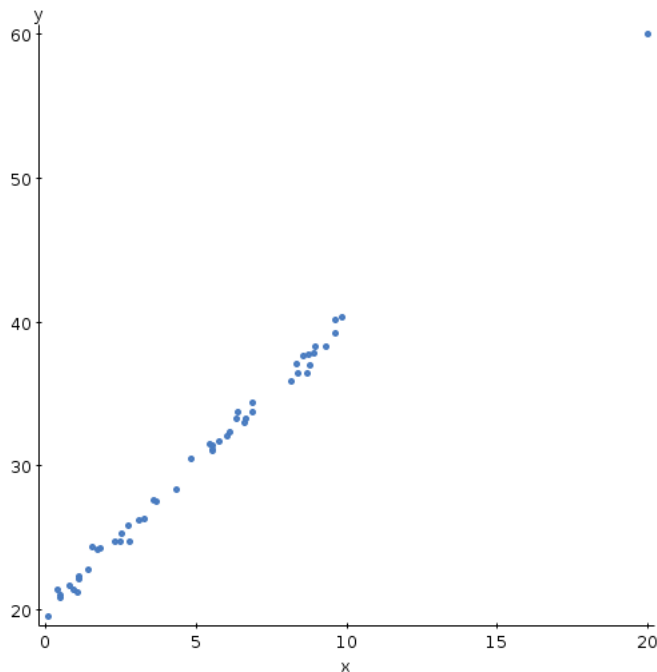
- (a) The stray point is not influential.
- (b) * The correlation between x and y would be much less if the stray point were removed.
- (c) The stray point has a large residual.
- (d) The correlation between x and y would be much larger if the stray point were removed.

36. See Question 35 for instructions.



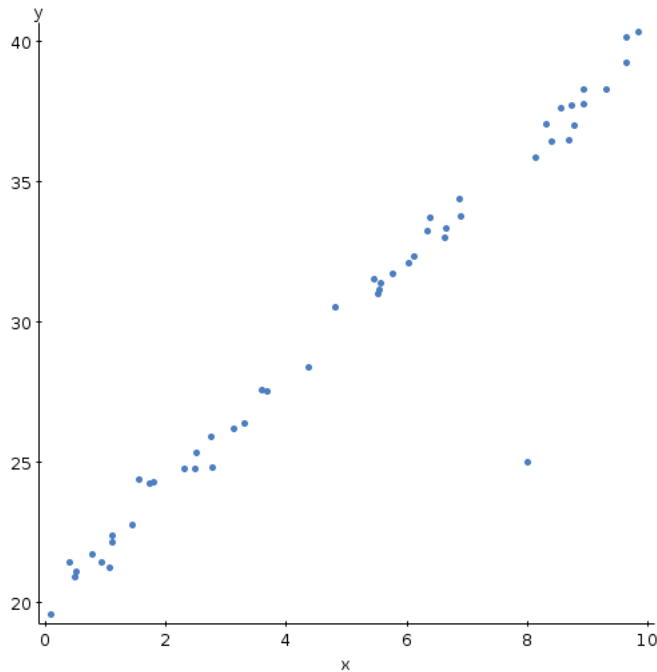
- (a) The stray point has a large residual.
- (b) * The stray point is influential.
- (c) If the stray point were removed, the correlation between x and y would not change.
- (d) If the stray point were removed, the correlation between x and y would decrease.

37. See Question 35 for instructions.



- (a) The stray point is not influential.
- (b) If the stray point were removed, the correlation between x and y would increase.
- (c) * If the stray point were removed, the slope of the line would hardly change.
- (d) If the stray point were removed, the correlation between x and y would decrease.
- (e) The stray point has a large residual.

38. See Question 35 for instructions.



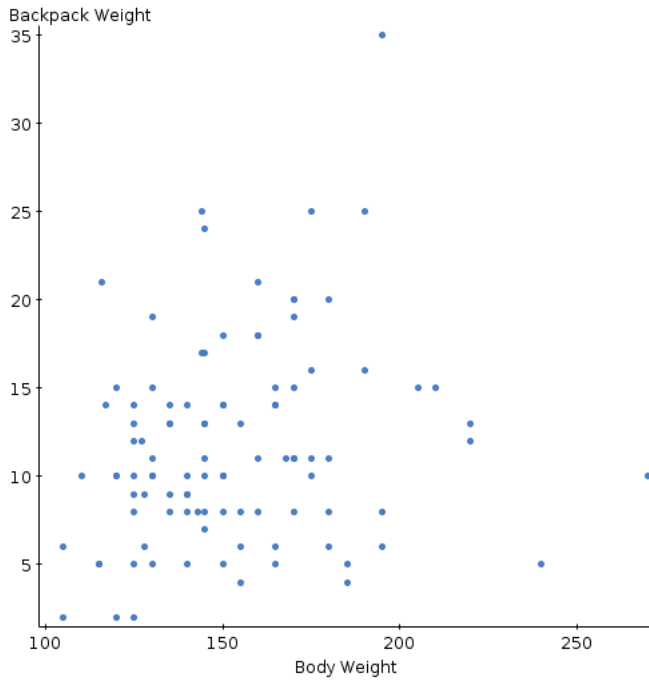
- (a) If the stray point were removed, the slope of the regression line would not change.
- (b) If the stray point were removed, the correlation between x and y would become smaller.
- (c) The stray point is influential.
- (d) * The stray point has a large residual.

39. A boxplot is drawn vertically (so that any outliers are at the top and bottom of the plot). What is the significance of the *width* of the box on the boxplot?

- (a) It describes the spread of the distribution
- (b) * It has no significance
- (c) It shows where the quartiles are
- (d) It shows the most extreme observations that are not outliers
- (e) It describes the centre of the distribution

40. Refer to the description of the backpackers data in Question 7. A scatterplot is shown below of each hiker's backpack weight (response) against body weight (explanatory).

What do you conclude from this plot?



- (a) There is a fairly strong linear relationship between backpack weight and body weight.
- (b) Backpack weights do not have a normal distribution.
- (c) A hiker with larger body weight tends to carry a backpack that weighs less, but the relationship is not very strong.
- (d) * There is at most a weak relationship between backpack weight and body weight.
- (e) There is a fairly strong relationship between backpack weight and body weight but it is not linear.