For each question, the percentage of students getting it correct is shown.

1. (80%) 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) 20

   (b) 25

   (c) 50

   (d) 60

   (e) * 75

   103 is the first quartile, so 25% of students are below and 75% above. (If there were any students who scored exactly 103, that would have fouled up the "75% more".)

2. (68%) 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) 39

(b) 4

(c) 365

(d) * 28

(e) 24

A 15% increase means multiplying by 1.15, so measures of centre and spread all get multiplied by 1.15. (Or, more briefly, they all increase by 15% too.) So the new IQR is $1.15(53.34 - 29.00) = 27.991$. Mark 28.

3. (76%) 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) 58 grams

   (b) 50 grams or less

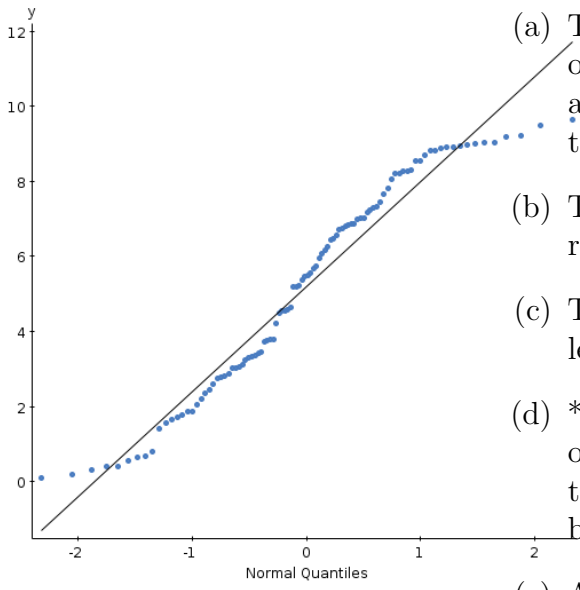
   (c) 66 grams or more

   (d) * 54 grams

   (e) 62 grams

   This is the normal table backwards. If 20% of eggs are heavier than x, 80% must weigh less than x. Look up 0.8000 in the body of Table A, giving $z = 0.84$ (0.7995 is as close as you can get). Turn this into a weight: $x = 50.9 + 3.7(0.84) = 54.008$. Or you can estimate what the answer ought to be: 68% of all eggs ought to weigh between $50.9 - 3.7$ and $50.9 + 3.7$, so half the remaining 32%, that is 16%, is above $50.9 + 3.7 = 54.6$. The value of x for 20% is a bit less than 54.6, so 54 ought to be the answer (50 is below the mean, so it can't be right.)

4. (24%) 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 farther apart than a normal distribution.

(b) The data are skewed to the right.

(c) The data are skewed to the left.

(d) * The values at the extremes of the distribution are closer together than a normal distribution.

(e) Any apparent failure to have a normal distribution is random variation.

The plot has a shape like an S (it curves *twice*) so it's not a simple case of skewness. Also, it's a clear pattern (look particularly at the points at the ends of the S: they are far from the line). To see what kind of pattern it is, imagine where the data would fall on the data scale (the y axis): the values at the top end are bunched up near 9, and the values at the bottom end are bunched up just above 0. So (d) is what's going on. (I actually took these values from a *uniform* distribution, whose density curve is shaped like a rectangle, and doesn't *have* tails.)

5. 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 0.1%

   (b) less than 1% but more than 0.1%

   (c) * between 1% and 3%

   (d) about 5%

   (e) 10% or more

   OK, I screwed up this one. The right answer is this: $z = (200 - 184)/9 = 1.78$; 0.9625 of Dutch men will have height less than this, and $1 - 0.9625 = 0.0375$ will be taller. This isn't one of the alternatives! The instructions are to mark the *best* (numerically closest) of the answers given; I thought this could reasonably be either the answer shown or "about 5%", so you got the point if you marked either one of those.

6. (75%) 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) 2
   (c) * 3
   (d) 12
   (e) 6

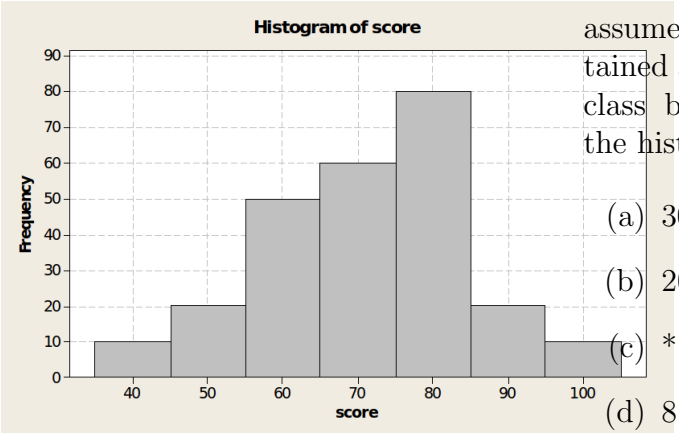
   Much the easiest way is to use 68–95–99.7. The two heights are 6 inches above and below the mean, which is twice the SD (because of the 95%), so the SD had better be 3. If you want to use the table, you have to figure out that half of the remaining 5% of heights go each end, so the proportion less than 61 is 0.0250. This goes with $z = -1.96$. 61 as a z-score is $(61 - 67)/\sigma$; put these equal and you find that σ is a smidgen more than 3 (3.06).

7. (59%) 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 shows the most extreme observations that are not outliers
   (b) It shows where the quartiles are
   (c) It describes the centre of the distribution
   (d) * It has no significance
   (e) It describes the spread of the distribution

   As we draw them, only heights on boxplots matter. (If you look at how StatCrunch draws boxplots, you see that they are skinnier when you have several of them side by side, but this doesn't matter.)

8. (66%) 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) 20

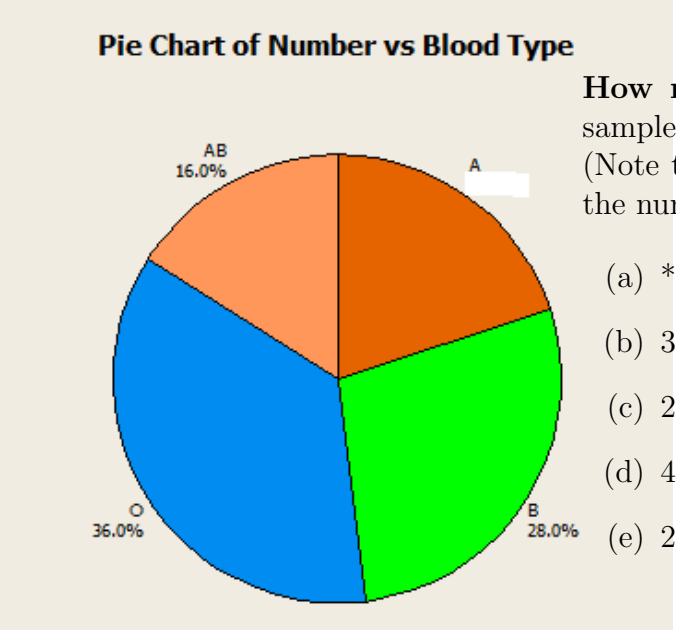
(c) * 12

(d) 8

(e) 50

We don't know how many students there are in total, so first add up the heights of all the bars: $10 + 20 + 50 + 60 + 80 + 20 + 10 = 250$. Of these, the first two bars are the students less than 55, and there are $10 + 20 = 30$ of these. This makes $30/250 \times 100 = 12\%$.

9. (87%) 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) 3

(c) 20

(d) 4

(e) 2

The four percentages in the pie chart should add up to 100%; the ones shown add up to 80%, so the missing one (A) is 20%. This makes sense, since the A slice is a little bigger than the AB slice but definitely smaller than the B slice. 20% of 25 is 5.

10. 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) 475

    (b) 200

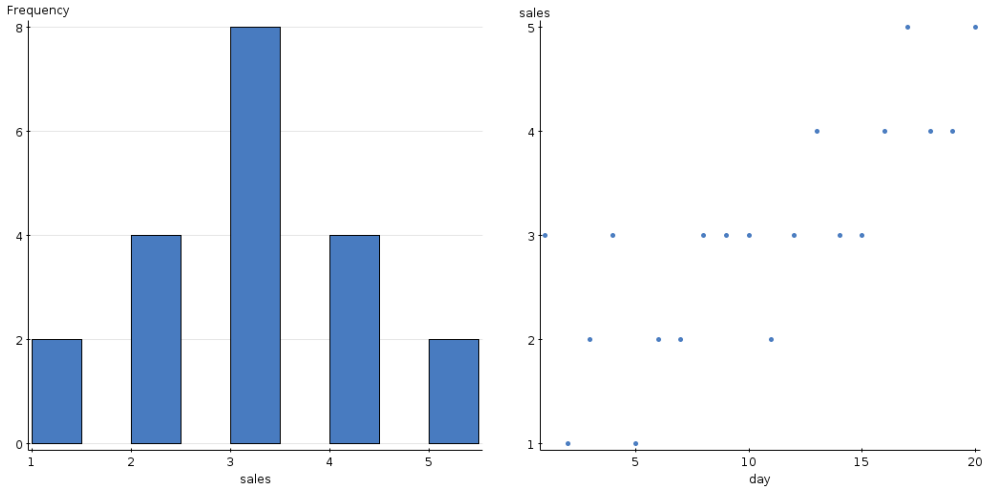
    (c) * 340

    (d) 68

    (e) 95

    Never mind about the 500 students for a minute. 135 and 165 minutes are one SD below and above the mean, so about 68% of all the students should take that long to finish. (Or you can turn 135 and 165 into z-scores (-1 and 1 respectively) and get a more accurate figure of $0.8413 - 0.1587 = 0.6826$ from the table.) Finally, 68% of 500 is 340. (If you used tables, you would have had 341.3, so 340 is clearly the best answer.)

11. (86%) 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) There is a curved relationship between sales and days.

(b) The distribution of sales is skewed.

(c) Sales have approximately a normal distribution.

(d) * Sales appear to be increasing over time.

(e) Sales appear to be decreasing over time.

A scatterplot tells you about the association between two variables (here sales and days since opening), while a histogram tells you about the distribution of sales without reference to the number of days since opening. The scatterplot shows a positive association (maybe not a hugely strong one, but a positive association nonetheless), so sales are increasing over time. "Sales have approximately a normal distribution" is something you could deduce from the histogram but not the scatterplot, which is the wrong way around.

12. (68%) 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) * 190

(b) 100

(c) 15

(d) 0

(e) -10

Get the slope first, and then the intercept. Slope is $0.947(144.2/9.8) = 13.934$, and intercept is $472.7 - 20.6(13.934) = 185.65$.

13. 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) stratified sample

    (b) voluntary-response sample

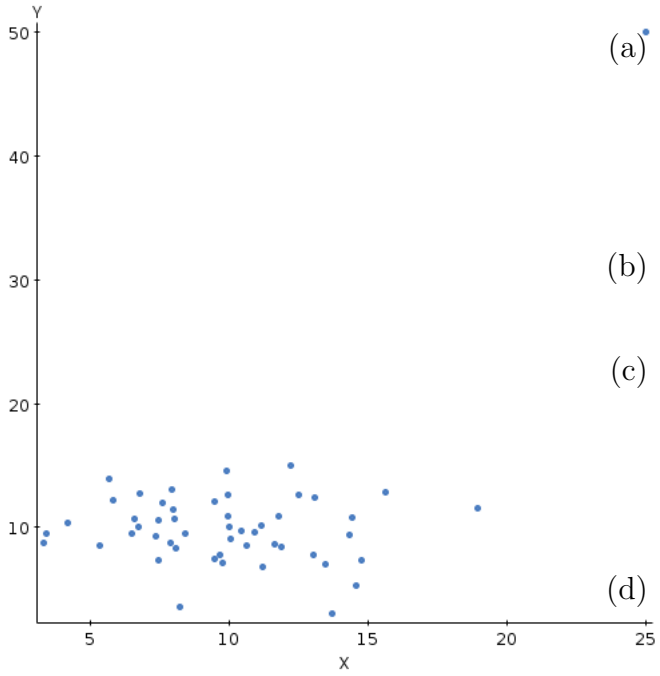
    (c) simple random sample

    (d) * multi-stage sample

    (e) systematic sample

    This is one of the questions we omitted. But: the procedure of first selecting cases, and then only looking at bags within the selected cases (not at any other bags) makes it a multi-stage sample.

14. (27%) 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 correlation between x and y would be much less if the stray point were removed.

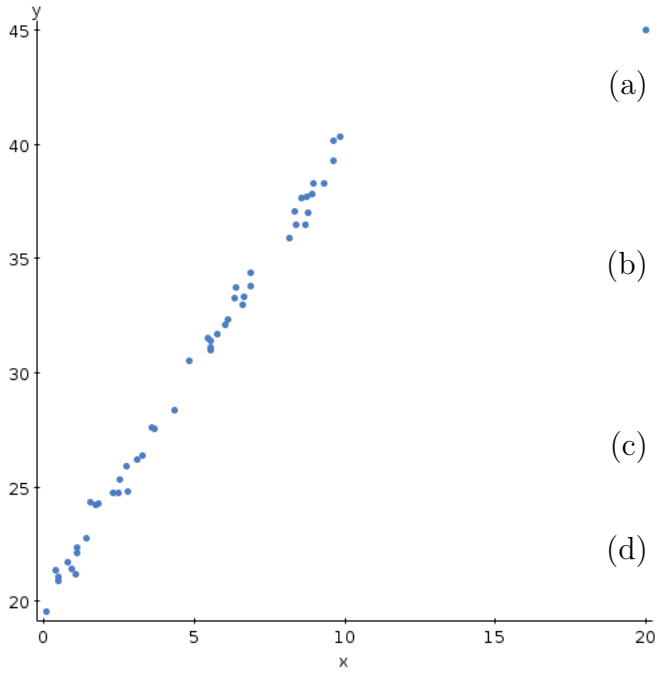
(b) The stray point has a large residual.

(c) The correlation between x and y would be much larger if the stray point were removed.

(d) The stray point is not influential.

The stray point is the one top right. It has clearly the largest x-value, so it is influential, meaning that it pulls the line towards itself. That rules out (d), and (b) as well. If that point were taken away, there would be almost no relationship between x and y (the correlation would be very small), pointing you towards (a) rather than (c).

15. (71%) See Question 14 for instructions.



(a) If the stray point were removed, the correlation between x and y would not change.

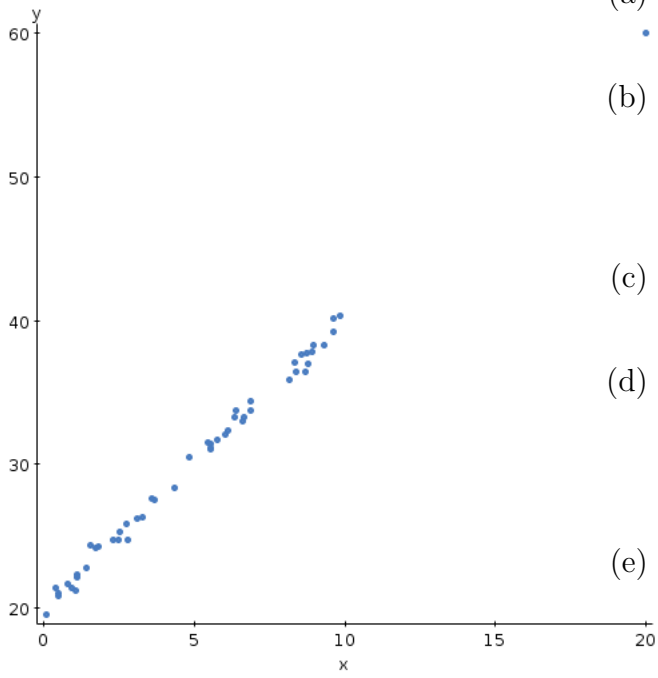
(b) If the stray point were removed, the correlation between x and y would decrease.

(c) The stray point has a large residual.

(d) * The stray point is influential.

The stray point is again top right and again influential (clearly the largest value of x). So the line will go close to the stray point, rather than following the trend of the other points. Taking the stray point away would make the correlation *very* close to 1, whereas right now it is something less than that.

16. (62%) See Question 14 for instructions.



(a) The stray point is not influential.

(b) If the stray point were removed, the correlation between x and y would increase.

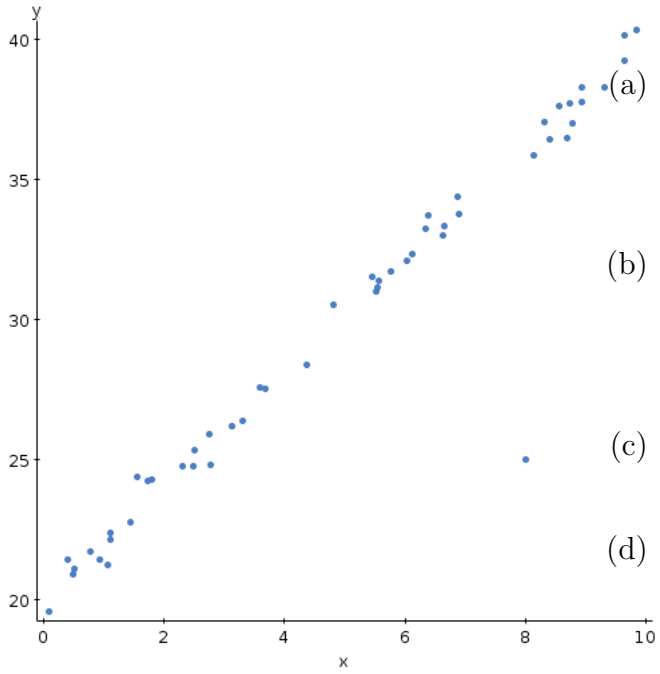
(c) The stray point has a large residual.

(d) If the stray point were removed, the correlation between x and y would decrease.

(e) * If the stray point were removed, the slope of the line would hardly change.

This time the stray point (again top right, again influential) is on the trend, so even though it drags the line closer to itself, the line goes about where it would go without the stray point. Neither the slope nor the correlation change much.

17. (39%) See Question 14 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.

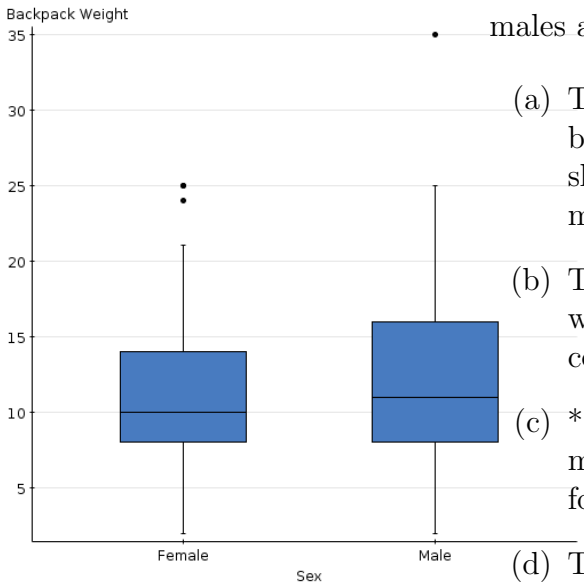
(c) * The stray point has a large residual.

(d) The stray point is influential.

This time the stray point is bottom right, a long way off the trend of the other points. Its x value is not unusual, so it won't be influential, and taking it away will make the association look stronger. Because it is a long way off the trend without being influential, it will have a large residual.

18. (80%) 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 distribution of male backpack weights is more skewed to the left than for females.

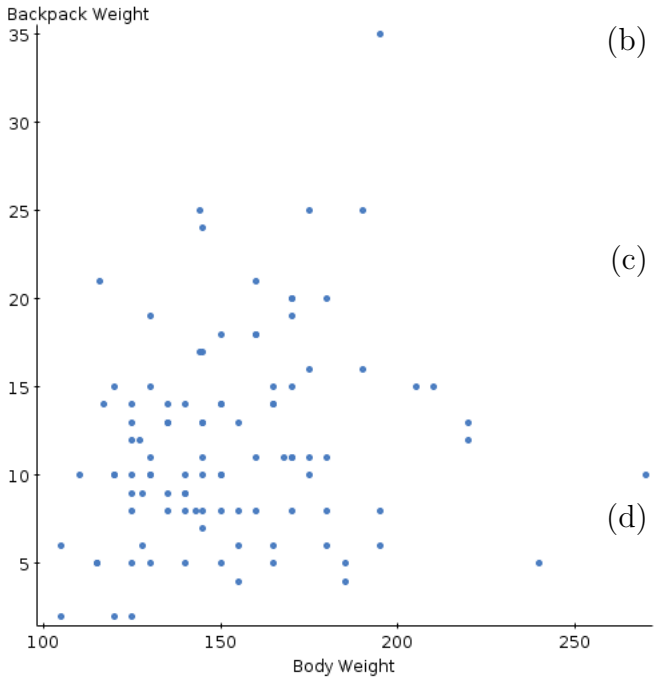
(b) The distributions of backpack weights differ substantially in centre.

(c) * The backpack weights for males have a greater spread for males than for females.

(d) There is a substantial difference in the number of outliers between males and females.

There is a tiny difference between the medians, and the males have one outlier as against the females' two (though the male one is further out), but I think the most striking difference here is that the boxplot is taller for the males: that is, the backpack weights for males are more spread out than for females.

19. (67%) Refer to the description of the backpackers data in Question 18. A scatterplot is shown below of each hiker's backpack weight (response) against body weight (explanatory).



What do you conclude from this plot?

(a) Backpack weights do not have a normal distribution.

(b) There is a fairly strong linear relationship between backpack weight and body weight.

(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.

Does knowing body weight tell you anything about backpack weight? Not very much. If anything, people who are heavier have heavier backpacks, but the relationship is very weak, and there's certainly no evidence of the relationship being anything other than a straight line, if it exists at all.

20. (61%) 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:

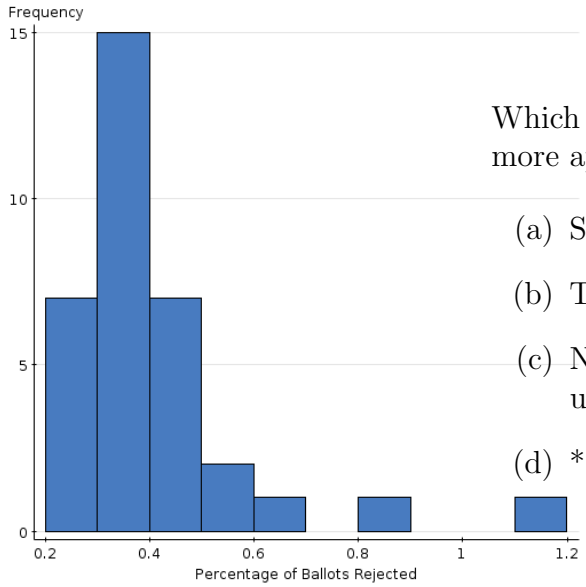
Summary statistics:

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) Neither of them should be used

(d) * Summary 1

The histogram shows a distribution that is skewed to the right (or has outliers at the top end, if you prefer). Either way, the mean and SD will be larger than they ought to be, and we should prefer a summary based on the median and IQR. (Summary 1 is a five-number summary, so is just the thing.)

21. (62%) Question 20 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 mean and SD would change substantially, while the median and IQR would barely change at all.

    (b) The mean and median would change substantially, while the SD and IQR would not change at all.

    (c) The data would become less spread out, so the IQR and SD would both decrease.

    (d) Something would happen that is not described in the other alternatives.

    (e) The median and IQR would change substantially, while the mean and SD would barely change at all.

    3.70 would be a high outlier, so it would change anything that is affected by outliers, ie. the mean and SD. The median and IQR are barely affected by one unusual (or wrong) value, which is why we have a rule based on the quartiles and IQR for assessing outliers with.

22. (21%) Suppose a boxplot had been drawn of the data in Question 20. The upper whisker would extend to what value? (You may assume that outliers are plotted separately on the boxplot.)

    (a) 0.635

    (b) * between 0.44 and 0.635, but it is impossible to tell exactly what without seeing the data values

    (c) 0.24

    (d) 0.44

    (e) 1.1

    The upper whisker of a boxplot extends to the *highest value that is not an outlier*, not to the limit beyond which a value is declared to be an outlier, which is $0.44 + 1.5(0.44 - 0.31) = 0.635$. Whatever the value is, it'll be somewhere between Q3 (0.44) and 0.635 (possibly including either value), but only by looking at the data can we tell what. Even looking at the histogram doesn't help: there is one value between 0.6 and 0.7, which might be an outlier or it might not.

23. (68%) 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) * 2

(b) 4 or more

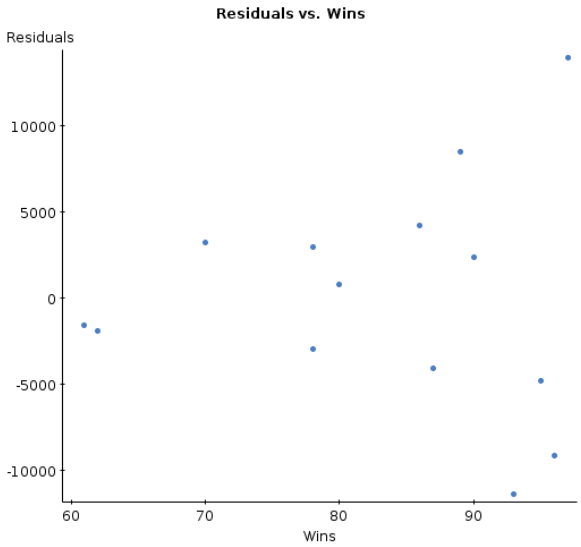
(c) 3

(d) 1

(e) 0

1.5 times IQR is $1.5(520 - 489) = 46.5$, so anything below $489 - 46.5 = 442.5$ or above $520 + 46.5 = 566.5$ is an outlier. There are no outliers at the top end (the largest values are 550, certainly less than 560) but the two lowest values 410 and 420 are outliers. A closer look at the stemplot reveals that they are quite a bit lower than the rest.

24. (41%) 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 residuals should have a normal distribution, and they do not.

(b) * The predictions become less accurate as the number of wins increases.

(c) There are no problems with this residual plot.

(d) There is little or no relationship between the number of wins and attendance.

(e) The relationship between number of wins and attendance is actually curved, not a straight line.

This is a *residual* plot, so if a straight-line regression is OK, the residual plot should show no pattern. But this one does: as you go across to the right, the residuals tend to get further from zero, which is a fanning-out, and therefore the predictions get less accurate as the number of wins gets larger.

The regression, which wasn't shown, said that attendances got a bit larger as the number of wins increased. There are a

couple of possible reasons for the fanning-out: maybe larger attendances are harder to predict (and maybe the *percentage* errors have constant variability), or maybe it really depends on whether a team is still in contention for the playoffs (a team with a lot of wins might be in a tough division, so they wouldn't make the playoffs even with a lot of wins).

25. (60%) 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) 2 or more

(b) 1

(c) * 0.1

(d) 0.001 or less

These questions require careful thinking about what is out of what. This one says "how many people are both Inuit and from Nunavut, and what is that out of all Canadians? That

is 22560/29639030 = 0.00076, which is a bit less than 0.1%, but not nearly as small as 0.001%.

26. (65%) Refer to the table in Question 25. What percentage of Aboriginal people are Inuit from Nunavut?

    (a) 0.02
    (b) * 2
    (c) 20
    (d) less than 0.01
    (e) more than 30

    Out of all Aboriginal people, how many of them are both Inuit and from Nunavut? You would guess a bigger answer than the previous question (because it's out of fewer people): 22560/976305 = 0.023, about 2%.

27. (91%) Refer to the table in Question 25. What percentage of the population of Nunavut is Inuit?

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

    Out of all the people in Nunavut, how many are Inuit? This is 22560/26665 = 0.846, which is nearly 85%. (Did this surprise you?)

28. (83%) Refer to the table in Question 25. What percentage of Inuit are from Nunavut?

    (a) 70

    (b) * 50

    (c) 85

    (d) 10

    (e) 20

    Read carefully: this is *not* the same as the previous one. Out of all the people who are Inuit, how many of them come from Nunavut? This is 22560/45070 = 0.5005, right around 50%. In the light of the previous question, this might seem surprisingly small, but: most people who live in Nunavut are Inuit (previous question), about half of the people who are Inuit live in Nunavut (this one). There are quite a few Inuit in Quebec and in NL, but there aren't very many "southerners" in Nunavut.

29. (73%) 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) * use a larger sample

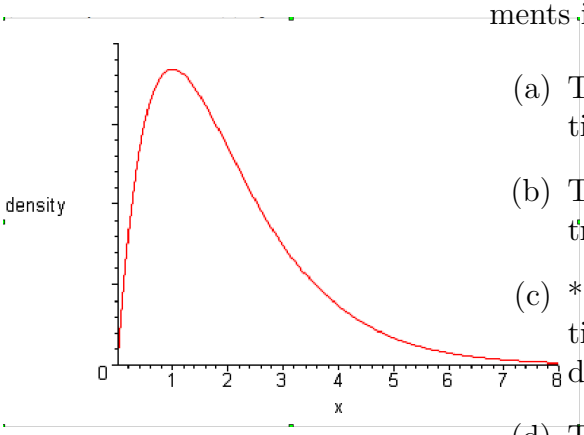
    (c) use a voluntary-response sample

    (d) sample from a larger population

    (e) sample from a smaller population

    The population size doesn't matter, but having a larger *sample* is good (and "more accurate estimation" is the reason why). If we were to use a voluntary response sample, we wouldn't be able to say anything about how accurately our sample mean might estimate the population mean.

30. (74%) 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 third quartile of the distribution is 6.

(c) * The mean of the distribution is greater than the median.

(d) The distribution is left-skewed.

(e) The total area under the density curve is greater than 1.

4 and 6 might be halfway (and 3/4 of the way) along the range of possible values, but the fact that the density curve is higher at the left means that lower values of X are more likely. So the median is less than 4 and Q3 is less than 6. (Or ask yourself: where is the area split in half? Maybe 2 or 2.5; and where is the rightmost quarter of the area? Maybe above 3.5 or 4. Definitely less than the values given.) The distribution is skewed to the right, so the mean *will* be greater than the median. The area under a density curve is always 1 exactly, so that statement must be false.

31. (35%) 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) Yes, because the sample is large.

    (b) It's a good estimate of the fraction for *all visitors to that web site*.

    (c) Yes, because a voluntary-response sample was used.

    (d) * No, because this is not a random sample.

    (e) No, because there is always sampling variability.

    I think I was missing a "not" — "a good estimate of the fraction of all people who do *not* use emoticons". Given that, the estimate would be good if it had been some kind of random sample, because the sample is large. But it isn't: it's a voluntary response sample. So I don't think we made any material difference by missing out the "not": the important point is that it was a voluntary-response sample (the people who happened to visit the website *and* complete the survey), and that when you have a voluntary-response sample you can't say anything about how close your sample statistic might be to the population parameter.

32. (42%) 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
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

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.)

```
88063 56513 31056 32105 08993
```

(a) * Some province or territory not given in the other alternatives

(b) Nfld & Labrador

(c) Northwest Territories

(d) Prince Edward Island

(e) The list of random digits is not long enough

There are 13 provinces/territories to sample from, so take the random digits in 2's (13 has 2 digits) and expect to reject a lot of them. The random numbers are 88, 06, 35, 65, 13, 31, 05, 63, 21, 05, 08. The only ones of these you can use are 06, 13, 05, 08 (you have to reject the second 05 as well, since you can't choose an individual twice). 08 is Nunavut, so mark (here) (a).

33. (68%) 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) The shot put, because the distance is longer

    (b) * The long jump

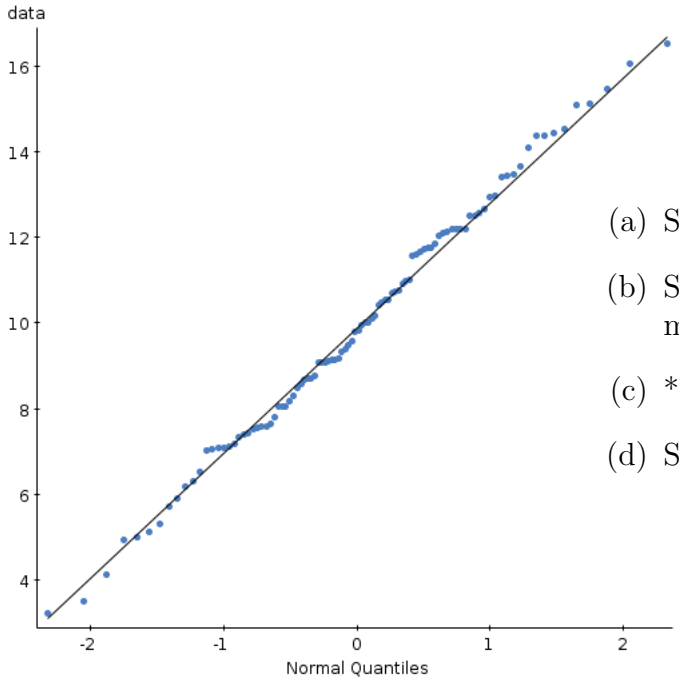
    (c) The shot put, but not just because the distance is longer

    (d) Both events represent the same performance

    Turn both performances into z-scores. The long jump gives $(6.78-6.17)/0.23 = 2.70$ and the shot put gives $z = (14.77-13.29)/1.24 = 1.19$. The long jump has a higher z-score, so it represents a better performance. (The fact that this athlete put the shot further than she long-jumped is neither here nor there; everyone can do that.) Also, shot-put performances are more variable than long-jump performances, and computing a z-score allows for this.

    Hepthathlons (and decathlons) are scored by referring each performance to a table, which awards points by comparing each performance with a "standard" one for each event. You'd get more points for long-jumping 50 cm above the standard than you would for putting the shot 50 cm above *its* standard, because shot-putting performances are more variable.

34. (74%) 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) Skewed to the left

(b) Symmetric but not normal

(c) * Normal

(d) Skewed to the right

This is as good a straight line as you'll see on one of these plots. These data are well described by a normal distribution.

35. (61%) 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 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 with one more square foot would cost about $370 more

    (b) a condo that costs nothing would have about 0.37 square feet.

    (c) a condo that is 0 square feet in size would cost about $370.

    (d) a condo with one more square foot would cost about $0.37 more

    (e) a condo that costs 1 thousand dollars more would have about 0.37 more square feet

    0.37 is the slope. The interpretation of the slope is that when you increase the explanatory variable by 1 (1 square foot), you increase the response variable by whatever the slope is (0.37 thousand dollars) on average.

    This is kind of hard to conceptualize; it also means that increasing the square footage by 100 increases the selling price by $370(100)=$37,000, which sounds about right.

36. (95%) In Question 35, 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) $370,000

(b) * $490,000

(c) $790,000

(d) more than $3,000,000

(e) less than $100,000

Substitute $x = 1200$ into the regression line to get $\hat{y} = 49.30 + 0.37(1200) = 493.3$, or $493,300. For downtown Toronto, that seems more or less reasonable. (1200 square feet is about the size of a biggish two-bedroom apartment.)

37. (84%) 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

Sketching even a very rough scatterplot should convince you that as x goes up, y goes rather clearly down. Only one of the alternatives is negative, and its size looks as if it should be about right.

38. 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) Stratified sample

    (b) Simple random sample

    (c) Convenience sample

    (d) * Multi-stage sample

    (e) Systematic sample

    We skipped this one, but: there is a two-stage process of picking the players, first selecting a team, and then selecting some of the players on that team. This is therefore a multi-stage sample.

    Note that taking a simple random sample of 10 players in the league could end up with 1 player on this team, 2 players on that team, and so on, and the drug-testing crew would have a lot of travelling to do to test all the players in the sample. This way, all 10 players are in the same place.

39. Question 38 described a baseball league's drug testing procedure. Why do you think this kind of sampling method was used?

    (a) It was simpler to understand than other methods.

    (b) It would give more accurate results than other methods.

    (c) * It was more convenient than other methods.

    (d) It was not convenient to obtain a list of all registered players in the baseball league.

    (Skipped also) Multi-stage sampling is for convenience, not primarily for accuracy. (Even if you picked another answer to the previous question, it seems reasonable to conclude that the prime virtue of this sampling method is its convenience.) I don't think it's simpler to understand than a simple random sample (why go to the trouble of picking a team first?), and players registered for a baseball league are going to be on a list somewhere.

40. (76%) 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

    If you knew the correlation (call it r), you'd find the slope by calculating $r(15/0.8)$. But this has to be equal to 15, so r has to be 0.8.