University of Toronto Scarborough
STAB22 Midterm Examination

June 11, 2012

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 32 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.

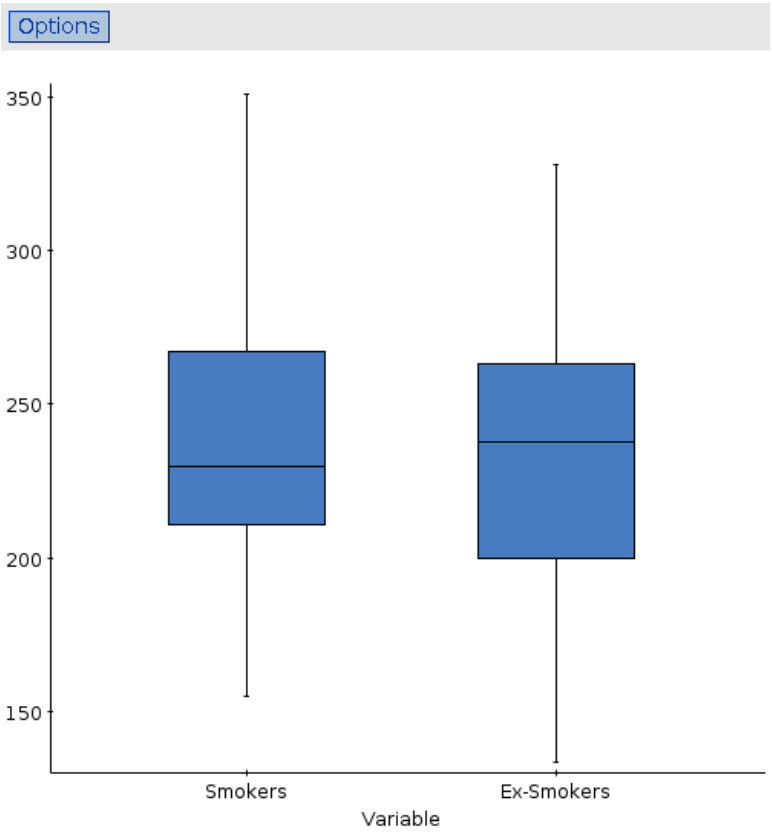
This examination is multiple choice. Each question has equal weight. On the Scantron answer sheet, ensure that you enter your last name, first name (as much of it as fits), and student number (in "Identification").

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. A study examining the health risks of smoking measured the cholesterol levels of people who had smoked for at least 25 years and people of similar ages who had smoked for no more than 5 years and then stopped. Side-by-side boxplots of the two groups' cholesterol levels are shown below.



Which one of the statements below is true?

(a) The smokers' cholesterol levels are more variable.

(b) * The two distributions differ in shape.

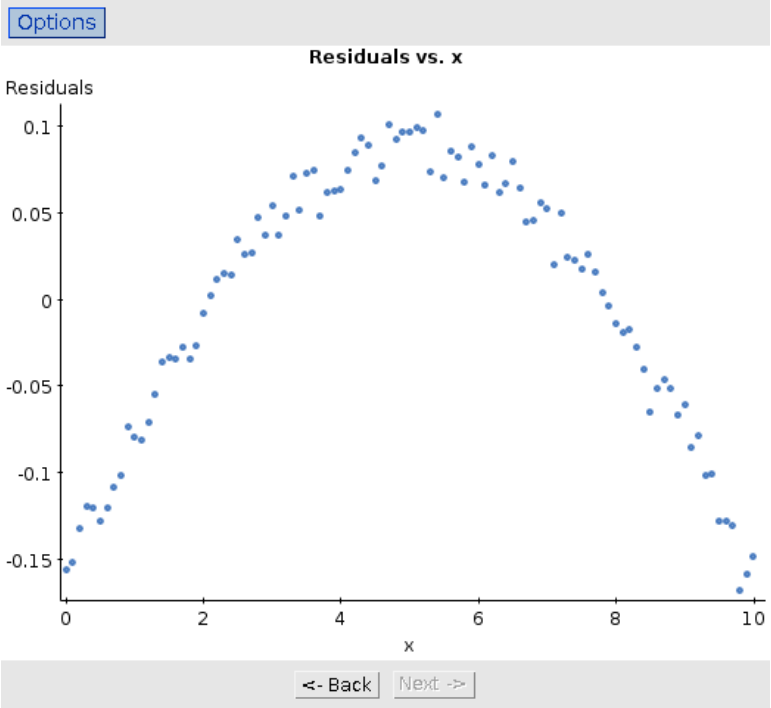
(c) There is an outlier in one or both distributions.

(d) One or both groups has median larger than 250.

(e) The mean cholesterol level for ex-smokers is higher than that for smokers.

Smokers' cholesterol levels are skewed to the right (a little) while ex-smokers' are more or less symmetric. If anything, the ex-smokers' box is taller (so that distribution is more variable). Boxplots say nothing about *means*. Outliers would be plotted separately, and the medians are somewhere around 230 and 240.

2. A linear regression is performed for predicting the variable y from the variable x, and the residuals from the regression are plotted against x below:



What do you conclude from this plot?

(a) * The data should be modelled with a curved relationship.

(b) A straight line model is appropriate for these data.

(c) The data are not normally distributed.

(d) The spread of the response variable is not the same for all values of the explanatory variable.

(e) There is no relationship between x and y.

When the residual plot has a pattern, something is wrong. When the pattern is a curve, the original relationship was a curve. (More precisely, if the relationship modelled was a curve, a residual plot like this indicates that the *wrong* curve was used for the model.) This is not a normal probability plot, so (c) doesn't apply, and (d) and (e) are just plain wrong.

3. A company announces openings for three positions as regional managers. There are 22 applicants, 12 male and 10 female. After the interviews, the company announces that all three new regional managers are women. The male applicants complain of discrimination.

   To assess whether there is a case of discrimination, a simulation is carried out. The simulation works under the assumption that every applicant is equally likely to be chosen as a sales manager. The results of the simulation are that 14% of the simulated sets of regional managers are all male, 43% contain 2 males and 1 female, 35% contain 1 male and 2 females, and 8% contain 3 females.

   What do you conclude about discrimination?

   (a) The simulation is unrealistic, since no-one believes that all of the applicants are equally likely to be selected as regional managers,

   (b) There is no evidence whatever that the males are being discriminated against, since it is highly likely that three females will end up being selected as sales managers given the numbers of applicants.

   (c) * There is only a small amount of evidence that the males are being discriminated against, since it is somewhat unlikely that the chosen regional managers will be all female, but far from impossible.

   (d) The males are definitely being discriminated against, because there are fewer female applicants than male, so there cannot possibly be more female regional managers than males.

   (a) is plausible, but for a simulation you have to start somewhere, and the usual place is to assume that things are equally likely (any collection of 3 people, ahead of time (before you know anything about them) is equally likely to end up being selected). So the question the simulation is asking is "*if* the candidates were all equally likely to be chosen, then how likely is it that all three that were chosen are women?". Note the "if". I wouldn't call 8% "highly likely", but nor is it "impossible". (d) is just wrong: even if you had 3 females and 300 males, it is *possible*, though very very unlikely, that the three females could be the ones chosen. I'd say (c) is the best answer; if the simulation had given something like 1% instead of 8%, I'd say the company would have a case to answer, but 8% is more likely than tossing 4 coins and getting 4 heads.

4. The red blood cell count of a healthy person was measured on each of 15 days. The value is measured in number of hundred thousand cells per microlitre. It is believed that a normal distribution does not apply for these data, and that there is no trend over time.

   Which of the graphs listed below would be most appropriate for displaying these data?

   (a) normal probability plot

   (b) bar chart

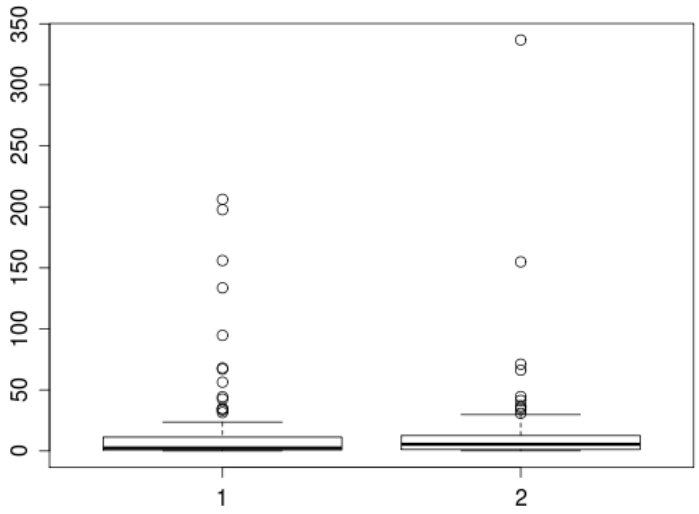
   (c) scatter plot

   (d) * stem and leaf plot

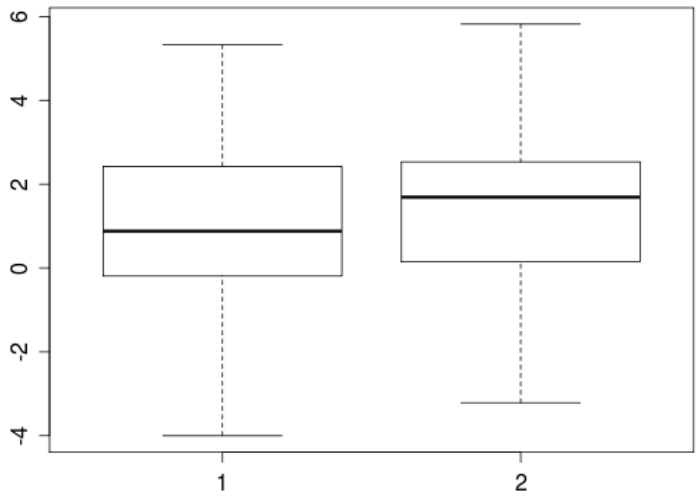
   (e) pie chart

   Quantitative variable, eliminating (b) and (e). One quantitative variable, not two, eliminating (c), and we are not interested in testing normality, eliminating (a). (I put in the bit about "no trend over time" because I didn't want to get into time plots, and also if you were interested in time trends you might have thought of making a scatterplot against time.)

5. A boxplot of a variable for two groups is shown below:



After re-expression, the boxplots become these:



Which of the following was *not* achieved by the re-expression?

(a) reducing the number of outliers

(b) making the spreads the same

(c) * making the centres the same

(d) making the distributions more symmetric

The goal of re-expression in this situation is to *compare* centres without being distracted by things like unequal spreads and differing skewness, so we *hope* that the re-expression will do at least some of (a), (b) and (d), which it certainly seems to have done. On the top boxplots, it's very difficult to compare the medians, but on the bottom ones, it's obvious that group 2 has a slightly higher median. You would suspect, because the distributions are now symmetric and without outliers, that group 2 would have a slightly higher mean as well, which turns out to be true. (In other words, (c) is not a typical goal of re-expression, so you might guess that this would be the answer without even looking at the pictures!)

6. A normal distribution is such that 16% of it is smaller than 13, and 2.5% of it is larger than 22. What is the *mean* of this normal distribution? (Hint: use the 68-95-99.7 rule and *draw a picture*, or maybe two.)

   (a) 17

   (b) * 16

   (c) 15

   (d) 19

   (e) 18

   A tricky one. If you had some practice with 68-95-99.7, you might recognize that 16% is half of what's left over from 68%, and 2.5% is the same for 95%. If not, then when you draw a picture, and think about the fact that when you apply this rule, you're always talking about the *middle* 68%, 95% or whatever, you'll be able to see that you have to think about 16% or 2.5% being the other end as well.

   So, from your picture, you see that 13 is one standard deviation below the mean and 22 is two standard deviations above. In symbols, $13 = \mu - \sigma$ and $22 = \mu + 2\sigma$. If you subtract these equations, you get $9 = 3\sigma$ so $\sigma = 3$, so $\mu = 16$ as you can check from both those equations.

   If push comes to shove, you haul out Table Z and find that 13 goes with a z-score of -0.99 and 22 goes with a z-score of 1.96. Then $(13-\mu)/\sigma = -0.99$ and $(22-\mu)/\sigma = 1.96$, and with some difficulty you solve for σ and μ.

   When I was in school in England, the examiners loved to state some possibly-helpful fact (like I did in my hint above), and then say "hence or otherwise do so-and-so", and you just *knew* that the "hence" way would be hard to see but easy to do when you saw it, and the "otherwise" way would be easy to see but almost impossible to do!

7. A study was carried out at Penn State University of how many ear pierces a sample of 137 female students had. The five-number summary is given below:

Min	Q1	Median	Q3	Max
0	2	4	7	15

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

How would you describe the *shape* of this distribution?

(a) skewed to the left

(b) can't tell from the five-number summary

(c) * skewed to the right

(d) like a normal distribution

(e) approximately symmetric

A boxplot would have a larger upper part of the box and longer upper whisker. Or, Q3 and max are further above the median than Q1 and min are below. (Thinking about what a boxplot would look like is often helpful with this kind of question.)

8. Using the information in Question 7, determine whether the maximum value in the data set is a suspected outlier. What do you conclude?

(a) Yes, because it is bigger than 12.

(b) * Yes, because it is bigger than 14.5.

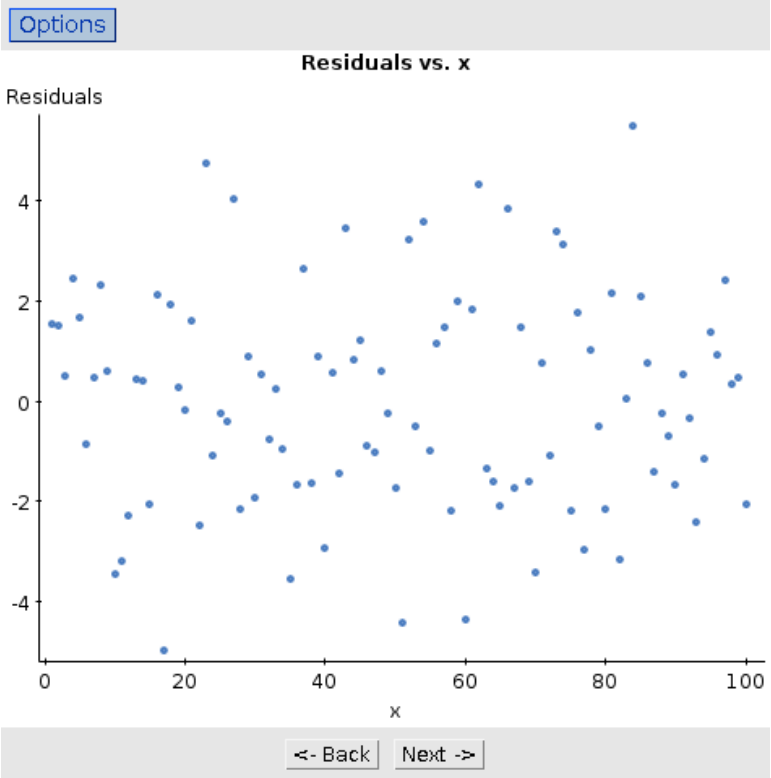
(c) No, because it is bigger than 14.5

(d) No, because it is not bigger than 17.

(e) Yes, because it is bigger than 7.

Use the 1.5 times IQR rule. Here this gives $IQR = 7-2 = 5$; $1.5 \times IQR = 7.5$; $Q3 + 7.5 = 7 + 7.5 = 14.5$. At the top end, anything *bigger* than the value you calculate is a suspected outlier (and deserves investigating). The last option cannot possibly be correct, because in *any* dataset, a quarter of the values will be bigger than the third quartile!

9. A linear regression is performed for predicting the variable y from the variable x, and the residuals from the regression are plotted against x below:



What do you conclude from this plot?

(a) The data are not normally distributed.

(b) The spread of the response variable is not the same for all values of the explanatory variable.

(c) * A straight line model is appropriate for these data.

(d) The data should be modelled with a curved relationship.

(e) There is no relationship between x and y.

This is pretty much a classical featureless residual plot, so there's nothing wrong here.

10. An article modelled the relationship between putting distance x (feet), and success rate y (percent) for professional golfers as

$$\hat{y} = 76.5 - 3.95x,$$

for putting distances between 5 and 15 feet. Use this information for this question and the next three questions.

Predict the success rate, in percent, for a putting distance of 10 feet.

> Just substitute 10 for the putting distance and see what the regression gives you for success rate. $76.5 - 3.95(10) = 76.5 - 39.5 = 37$. This regression could easily give you a nonsense success rate if you stray outside the range of the data, but 10 is right in the middle, and a 37% success rate is perfectly reasonable.

(a) 72

(b) 3

(c) 99

(d) 51

(e) * 37

11. In the situation described in Question 10, what is the best interpretation of the *slope*?

(a) The success rate increases by about 4% for each extra foot of putting distance.

(b) The putting distance decreases by about 4 feet for each extra percentage point of success rate.

(c) The success rate for a putt of 0 feet would be about 77%.

(d) * The success rate decreases by about 4% for each extra foot of putting distance.

(e) The putting distance increases by about 4 feet for each extra percentage point of success rate.

> Slope is change in response (success rate) when the explanatory variable (putting distance) increases by 1. Since the slope is negative, the success rate *decreases* as putting distance increases (it's more difficult to hole a putt from further away). Getting explanatory and response the right way around points to (d).

12. In the situation described in Question 10, what is the residual for a professional golfer who has a 50% success rate from a putting distance of 8 feet?

    (a) 0

    (b) -10

    (c) * 5

    (d) -5

    (e) 10

    First predict what the success rate would be for a putting distance of 8 feet: $\hat{y} = 76.5 - 3.95(8) = 44.9$. Then work out the residual as observed minus predicted: $50 - 4.9 = 5.1$. (c) is obviously the best answer.

13. In the situation described in Question 10, what do you predict the success rate, in percent, to be for putting distance 20 feet?

    (a) 0

    (b) -2.5

    (c) * should not do the prediction because that would be extrapolation

    (d) 2.5

    (e) 5

    You can try substituting without thinking, and get -2.5. But as a success rate that makes no sense. Or, if you're smart, you can remember that the regression only applies to putting distances between 5 and 15 feet, so that you shouldn't use it outside that range. What I was trying to get at in this question is that unthinking substitution doesn't do anyone any good.

14. Suppose you decide to play a certain lottery and see whether you win a prize or not. (The chances of winning *a* prize are the same every week, even though the jackpot might vary.) What makes this process random, as we defined "random"?

    (a) Whether you win in any given week is unpredictable.

    (b) You might win and you might not; there's no saying which.

    (c) * Whether or not you win in any given week is unpredictable, but in the long run you can work out what fraction of weeks you should win.

    (d) In the long run you can work out what fraction of weeks you'll win.

    Randomness has two parts: short-term unpredictable, long-term predictable. (a) and (d) get one of those, but only (c) gets both. (b) might have been *your* definition of "random" coming into the course, but I wanted to steer you into something more precise than that.

15. For female bears, the correlation between height and chest girth (the distance around the outside of the chest) is 0.92. Assume that calculating the correlation is reasonable here. When comparing female bears of different heights, what can you say about their chest girths?

    (a) The taller bear will probably have a smaller chest girth.

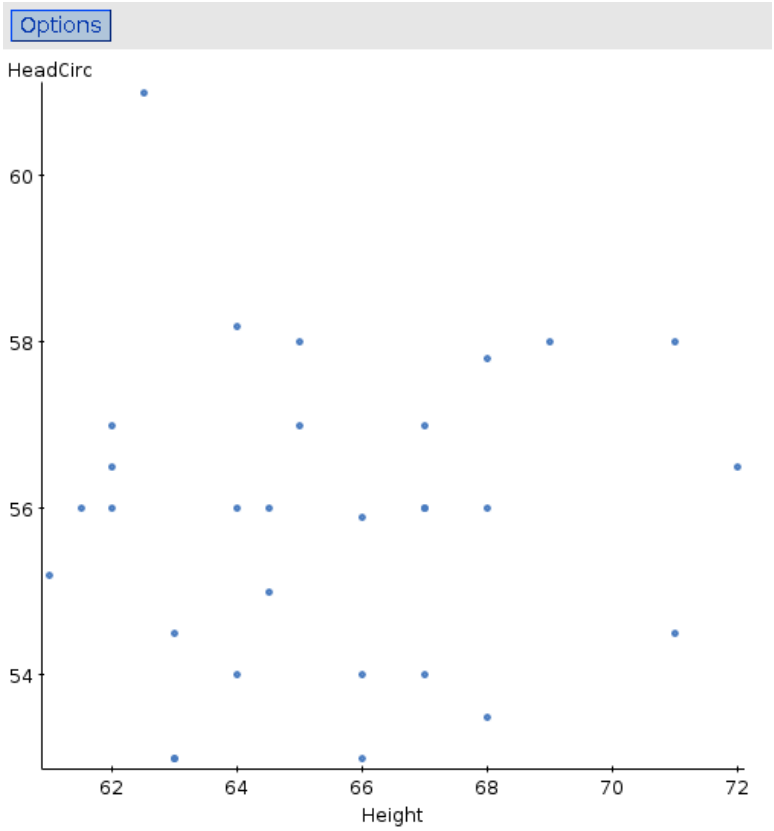
    (b) Cannot say anything about the chest girths.

    (c) The two bears will probably have about the same chest girth.

    (d) * The taller bear will probably have a larger chest girth.

    The correlation is high and positive, so larger x goes with larger y. Not for certain, since the correlation is not 1, but often enough.

16. The scatterplot below shows head circumference (cm) vs. height (inches) for a sample of 30 females.



Use this information for this question and the one following.

What direction of association does this plot show between the two variables?

(a) Positive association

(b) * No association

(c) Curved association

(d) Negative association

Except possibly for the one observation top left, knowing height tells you nothing about head circumference. Basing an association on one observation is too flimsy, so there's no association here.

17. Refer to Question 16. There is one outlier on the plot. What is the head circumference, in cm, for this individual?

    (a) 58

    (b) * 61

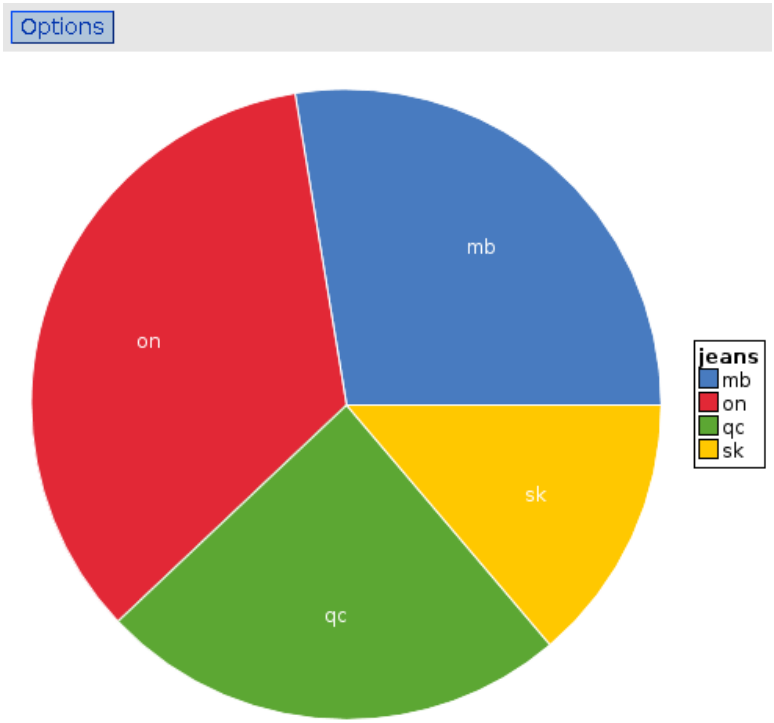
    (c) 54

    (d) 53

    (e) 56

    Height 63 (or 62.5), head circumference 61 is the point at the top left, which doesn't fit the overall non-trend. (Head circumference is definitely more than 60, which eliminates the other alternatives.)

18. A jeans manufacturer has plants in Quebec (QC), Manitoba (MB), Ontario (ON) and Saskatchewan (SK). A sample is taken of jeans made by the manufacturer, and the plant is recorded where each pair of jeans was made. A pie chart of the results is shown below.



Use this information for this question and the one following.

About *what percentage* of the jeans were made in Quebec?

(a) 50%

(b) * 25%

(c) 15%

(d) 33%

(e) 10%

The QC pie slice goes from about "5 o'clock" to "8 o'clock", so is about a quarter of the whole. (The colours didn't show up on your exam paper, but the letters distinguishing the provinces did — I checked.) Even if you didn't think of the clock face thing, the QC slice is definitely more than a sixth (15%) and definitely less than a third (33%).

19. Refer to the pie chart in Question 18. There were 29 pairs of jeans in the data set. About *how many* of them were made in Ontario?

    (a) 5

    (b) 15

    (c) * 10

    (d) 20

    (e) none

    What fraction is Ontario of the whole? About a third (8 o'clock to 12 o'clock), but definitely more than a quarter and less than a half. So the number of jeans should be 29 times a third, or 9.7, which is closest to 10. (Or, more than $29(1/4) \simeq 7$ and less than $29(1/2) = 14.5$, which eliminates the other possibilities.)

20. Yearly rainfall totals for a certain city in Northern California have a normal distribution with mean 18 inches and standard deviation 6 inches. Use this information for this question and the 2 following.

    In what proportion of years is the annual rainfall greater than 25 inches?

    (a) * 0.12

    (b) 0.22

    (c) 0.88

    (d) 0.32

    (e) 0.48

    Turn 25 into a z-score and look it up in the table: $z = (25 - 18)/6 = 1.17$, giving 0.8783 for proportion *less*; we want more, so take it away from 1, giving 0.1217.

21. Refer to the information in Question 20. In what proportion of years is the annual rainfall between 10 and 15 inches?

    (a) 0.78

    (b) 0.12

    (c) * 0.22

    (d) 0.88

    (e) 0.48

    Turn 10 and 15 into z-scores, look them up in the table, and subtract: 10 has z-score $(10 - 18)/6 = -1.33$, for which the table gives 0.0912; 15 has z-score $(15 - 18)/6 = -0.50$, for which the table gives 0.3085. Subtracting gives 0.2173.

    If you prefer the other way: proportion less than 10 is 0.0912 (as above); proportion more than 15 is $1 - 0.3085 = 0.6915$; everything else is between, which is $1 - 0.0912 - 0.6915 = 0.2173$.

22. Refer to the information in Question 20. In the driest 3% of years, the rainfall is less than how many inches?

    (a) 15

    (b) 12

    (c) * 7

    (d) 29

    (e) 22

    The driest 3% of years have the *smallest* rainfall. Use the normal table backwards to find that the bottom 3%=0.0300 of the standard normal distribution is less than $z = -1.88$, and then turn that value of z back into an amount of rainfall: $18 - 1.88(6) = 6.72$, which is closest to 7 inches. Or: see what amount of rainfall x would give you a z of -1.88 by $(x-18)/6 = -1.88$ and solve for x to get the same answer. (The answer 29 is the rainfall for the *wettest* 3% of years, when there is the *most* rainfall.)

    You can eyeball this by saying that 3% is close to 2.5%, which is what's left at the bottom after you take out the middle 95%, and therefore the right amount of rainfall should be close to 2 SD's below the mean: $18 - 2(6) = 6$. The right answer should be a bit more than this, which points you (correctly) to 7 inches.

23. In basketball, a player can attempt a shot from close to the basket, scoring 2 points if it succeeds, or can attempt a shot from farther away from the basket, scoring 3 points if it succeeds. Basketball statisticians keep track of the number of attempts at each type of shot for each player on a team. Below are data for two players, Morgan and Lisa, for three different seasons.

Table 1

	Morgan			Lisa		
	Successes	Shots	%	Successes	Shots	%
Two-pointers	20	40	50%	44	95	46%
Three-pointers	8	32	25%	4	30	13%
Total	28	72	39%	48	125	38%

Table 2

	Morgan			Lisa		
	Successes	Shots	%	Successes	Shots	%
Two-pointers	20	40	50%	44	95	46%
Three-pointers	8	32	25%	3	19	16%
Total	28	72	39%	47	114	41%

Table 3

	Morgan			Lisa		
	Successes	Shots	%	Successes	Shots	%
Two-pointers	20	40	50%	16	30	53%
Three-pointers	1	5	20%	9	40	23%
Total	21	45	47%	25	70	36%

Which table illustrates Simpson's Paradox?

(a) * More than one of them

(b) Table 3 only

(c) None of them

(d) Table 2 only

(e) Table 1 only

For Simpson's paradox, you want something misleading to come out of the overall percentage, here of shots made. In table 1, Morgan's overall shooting percentage is higher, and her shooting percentage is also higher for each type of shot. No paradox here. But in tables 2 and 3, there is an apparent contradiction: in table 2, Lisa has a higher overall success rate, but a lower success rate on each type of shot, which seems to be impossible, but has a rational explanation (which is what a paradox is). Likewise, in Table 3, Morgan is better overall, but worse at each type of shot. So there are two cases of Simpson's paradox.

24. Refer again to the situation described in Question 23. Which of the following scenarios makes it most likely that Simpson's paradox will occur?

    (a) The player with the lower success rate on each type of shot takes mostly difficult shots.

    (b) Both players take mostly easy shots.

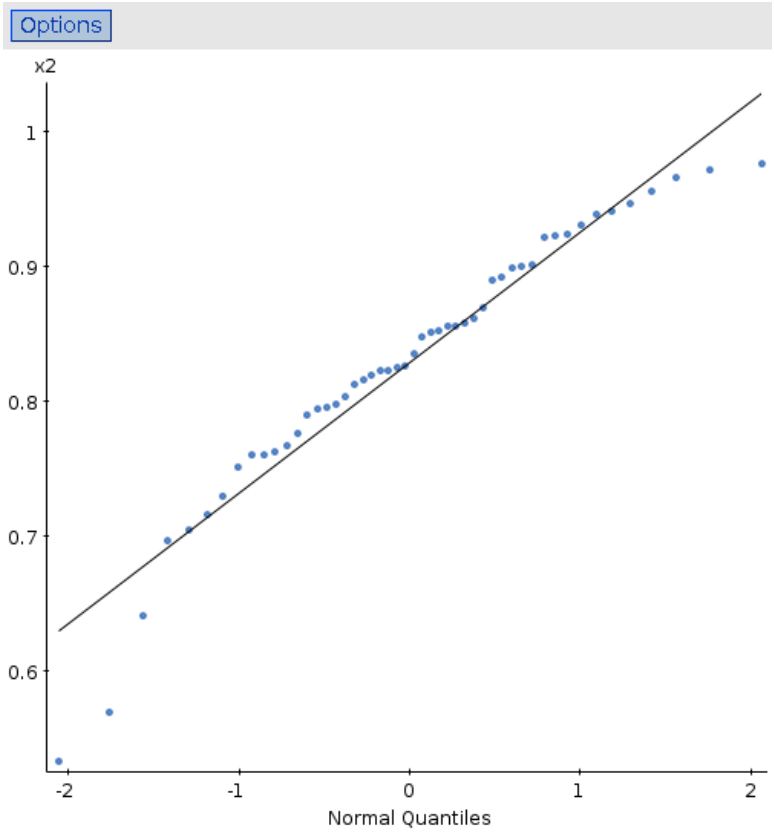
    (c) * The player with the lower success rate on each type of shot takes mostly easy shots.

    (d) The players both take similar numbers of difficult shots.

    (e) Simpson's paradox is impossible.

    In Table 2, Lisa takes mostly easy shots, and ends up with a higher overall shooting percentage, and a lower success rate on each type of shot. In Table 3, it's Morgan who does the same thing. What they have in common is (c). Does that make sense? Well, yes: the player with the lower success rate on each type of shot has their overall success rate "inflated" by having taken mostly easy 2-point shots.

25. A normal probability plot is shown below.



What do you conclude from this plot?

(a) A normal model is reasonable for these data.

(b) There is a strong association between the two variables.

(c) The data are skewed to the right.

(d) * The data are skewed to the left.

(e) The data are symmetric but not normal.

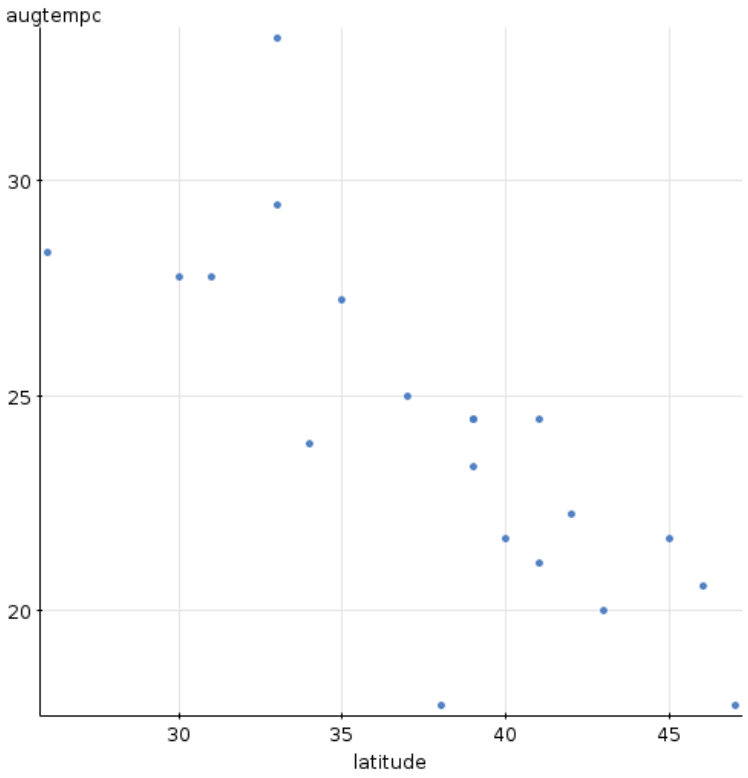
This is a curve. Maybe the points in the middle are rather near a straight line, but the points at the ends stray enough off the line in consistent directions for me to call this a curve.

So the data are skewed. Which way? Look at the *vertical* axis, which is where the data are. The bottom-most values are spread out (they are lower than they ought to be), while the top-most values are bunched up (they don't go above about 0.97, where, if the normal were correct, you'd expect two or three of them to be bigger than that. Bunched at the top and spread-out at the bottom points to skewed left.

(b) is irrelevant because it's not a scatterplot.

26. It seems natural that the further away a city is from the equator, the cooler it should be in summer. Data were collected from 20 US cities, including the latitude (a larger value means further north) and the average August temperature (in degrees Celsius). A scatterplot and some summary statistics are shown below.

Options



Options

Summary statistics:

Column	n	Mean	Std. Dev.
augtempc	20	24.11111	3.999675
latitude	20	37.95	5.586403

Options

Correlation between augtempc and latitude is:
-0.78072876

Assuming that a straight-line relationship is appropriate, calculate the slope and intercept of the least-squares regression line for predicting mean August temperature from latitude. What is the *intercept*?

(a) 20

(b) 0

(c) * 50

(d) 150

(e) 80

Just apply the formulas. Slope is $-0.7807(4)/5.586 = -0.559$ and intercept is $24.111 - 37.95(-0.559) = 45.32505$, not forgetting that minus a minus is a plus.

You can also eyeball this one. A latitude of 45 appears to go with a summer temperature of 20, and a latitude of 30 appears to go with a temperature of 30. So an increase of 15 in latitude goes with about a 10 degree decrease in temperature. That means the slope ought to be about $-10/15 \simeq 0.7$, and decreasing the latitude by two more steps of 15 gives an intercept of 50. This is close enough to get to the answer. (We are not extrapolating here, because we are not doing a prediction; we figured the intercept by saying "*if* the straight line continues to work, the temperature would be at about 50 when latitude is zero". Using this to predict the temperature at the equator (latitude zero) is another matter.)

27. Six vehicles are selected from those that have campus parking permits, and the following data are recorded:

Vehicle	Type	Make	Carpool?	Commute distance (km)	Vehicle age (years)
1	Car	Honda	No	23.6	6
2	Car	Toyota	No	17.2	3
3	Truck	Toyota	No	10.1	4
4	Van	Dodge	Yes	31.7	2
5	Motorcycle	Harley-Davidson	No	25.5	1
6	Car	Chevrolet	No	5.4	9

Use the information above for this question and the 2 following.

What type of variable is "Make"?

(a) * Categorical

(b) Quantitative

(c) Discrete

(d) Continuous

I *hope* you got this one!

28. How many quantitative variables are there, not counting any identifier variables?

    (a) 1
    (b) 3
    (c) 5
    (d) 6
    (e) * 2

    Ignore the 1–6 down the left side, which just identify the individual vehicles. Commute distance and vehicle age are the only two quantitative variables.

29. What type of vehicle is the newest one in this data set?

    (a) Car
    (b) Truck
    (c) Van
    (d) * Motorcycle

    The 1-year-old vehicle is a Harley-Davidson motorcycle.

30. When a distribution of a variable is skewed to the right, which of the following re-expressions would you try first to make the distribution more symmetric?

    (a) * logarithm ("power 0")
    (b) square ("power 2")
    (c) reciprocal ("power -1")
    (d) no re-expression will work

    Go down the ladder of powers from 1 (which is no re-expression). The first of these values you hit is 0, which is the logarithm. Square root would also be a good answer, but it's not one of the alternatives here.

31. The relationship between two variables x and y first increases and then decreases. Which re-expression of the response variable y would you try, to make the relationship more like a straight line?

    (a) square ("power 2")

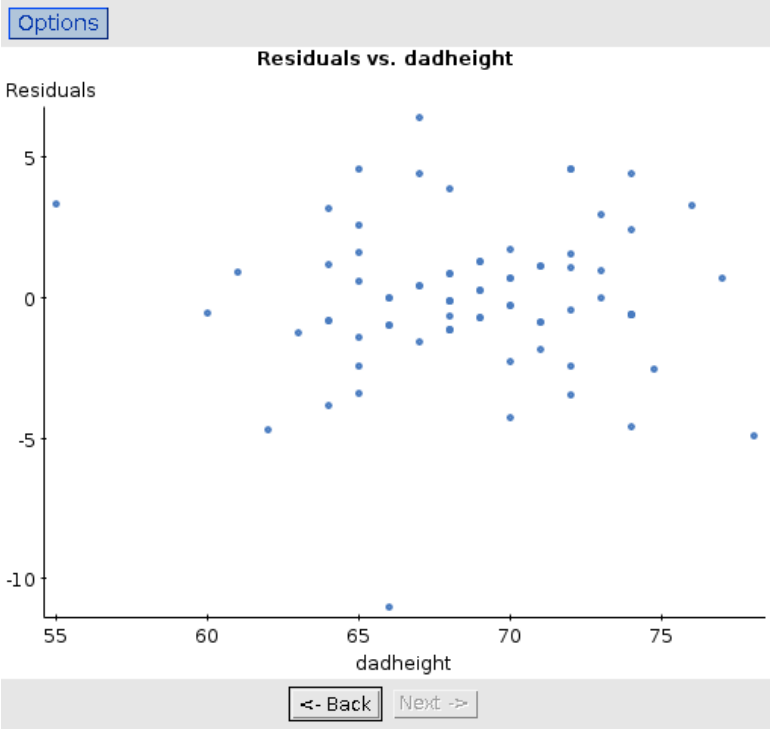
    (b) reciprocal ("power -1")

    (c) * no re-expression will work

    (d) logarithm ("power 0")

    The given re-expressions only straighten out associations that go in just one direction: either always up or always down. This isn't one of those, so at this point there's nothing we can do. (There is a way of fitting a parabola, which *does* go up and then down, but that's part of STAB27's agenda, and we don't deal with it.)

32. A regression was carried out to predict the height of 76 male college students from their father's heights. A residual plot from the regression is shown below.



What do you conclude from this plot?

(a) The regression is satisfactory.

(b) The association between father's height and son's height is curved.

(c) * One of the students is much shorter than predicted by the regression.

(d) There is at best a weak association between father's and son's heights.

(e) One of the students is much taller than predicted by the regression.

This residual plot looks perfect *except* for the one point at the bottom, with father's height 66 (inches) and residual -11. The negative residual means that the son's observed height was less than his predicted height, and by quite a bit, this being the farthest-from-zero residual of all. You might see a curve in the "cloud" of points, but the two outlying observations (the one I just mentioned plus the son with the very shortest father) don't support that curve at all.

33. In growing oranges, as the number of oranges per tree (x) increases, the average weight y (pounds) of an orange tends to decrease. The relationship between x and y is curved, but the relationship between x and $1/y$ is very close to linear. Information from this regression is shown below.

Options

Simple linear regression results:
Dependent Variable: 1/y
Independent Variable: x
1/y = 1.6026461 + 0.0011218862 x
Sample size: 14
R (correlation coefficient) = 0.9991
R-sq = 0.99826306
Estimate of error standard deviation: 0.012822304

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	1.6026461	0.006521398	≠ 0	12	245.7519	<0.0001
Slope	0.0011218862	1.3509196E-5	≠ 0	12	83.04612	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	1.1338896	1.1338896	6896.657	<0.0001
Error	12	0.0019729377	1.6441147E-4		
Total	13	1.1358626			

Predict the weight per orange (in pounds) on a tree with 700 oranges.

(a) 2.4

(b) 1.6

(c) 0.8

(d) 0.2

(e) * 0.4

Two steps in doing a prediction for a re-expressed regression: predict the re-expressed response (here $1/y$), and then undo the re-expression. The regression says $\widehat{1/y} = 1.602 + 0.001122(700) = 2.3874$. This is a predicted number of oranges per pound, so we need to flip it over to get what we want: predicted pounds per orange is $1/2.3874 = 0.419$. If you forgot to undo the re-expression, you would have gotten (a) instead of (e)!

34. A certain normal distribution has the middle 95% of its values between 20 and 34. Use this information for this question and the next one.

    What is the mean of this normal distribution?

    (a) 20
    (b) * 27
    (c) 31
    (d) 29
    (e) 25

    This one you can guess, since we're talking about the middle 95% of a symmetric distribution, so the mean ought to be halfway between 20 and 34.

    Or you can remember the rule and let μ be the mean and σ be the SD you're trying to find, write $\mu + 2\sigma = 34$ and $\mu - 2\sigma = 20$, and solve to find μ and σ (which gets you the answer to the next question as a bonus).

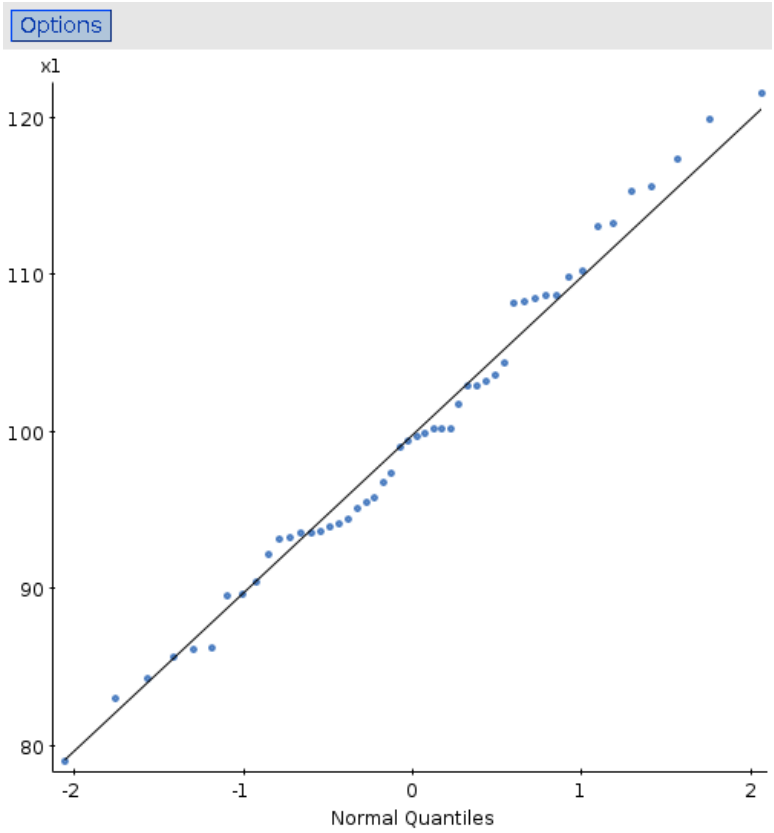
35. Question 34 gave some information about a normal distribution. What is the standard deviation of this normal distribution?

    (a) 7.0
    (b) 14.0
    (c) 27.0
    (d) 1.5
    (e) * 3.5

    Mean plus twice SD is 34, and the mean is 27, so twice the SD must be 7, so the SD itself must be 3.5. Or use mean minus twice SD is 20, with the same result.

    Or use the mathematics of the previous question to get both answers at once.

36. A normal probability plot is shown below.



What do you conclude from this plot?

(a) * A normal model is reasonable for these data.

(b) The data are skewed to the left.

(c) There is a strong association between the two variables.

(d) The data are skewed to the right.

(e) The data are symmetric but not normal.

This one is as straight as you could wish for, all the way along. You can look at the upper and lower ends, where the points are pretty much on the line, and nothing untoward is happening in between. (The data from which this plot came actually *were* normally distributed.)

"Strong association" might be valid for a scatterplot, but this is not one of those.

37. A survey was carried out to assess 212 people's perception of their own weight. Each respondent was categorized by sex and by their perception of their own weight (overweight, about right, underweight, don't know). The data were as follows:

Sex	Overweight	About right	Underweight
Female	39	87	3
Male	3	64	16

The principal aim of the study was to see whether males and females differed in their attitudes towards their own weights. Which percentages would be the best ones to use for this comparison?

(a) Marginal percentages of sex

(b) Marginal percentages of attitudes

(c) Overall percentages (out of the grand total)

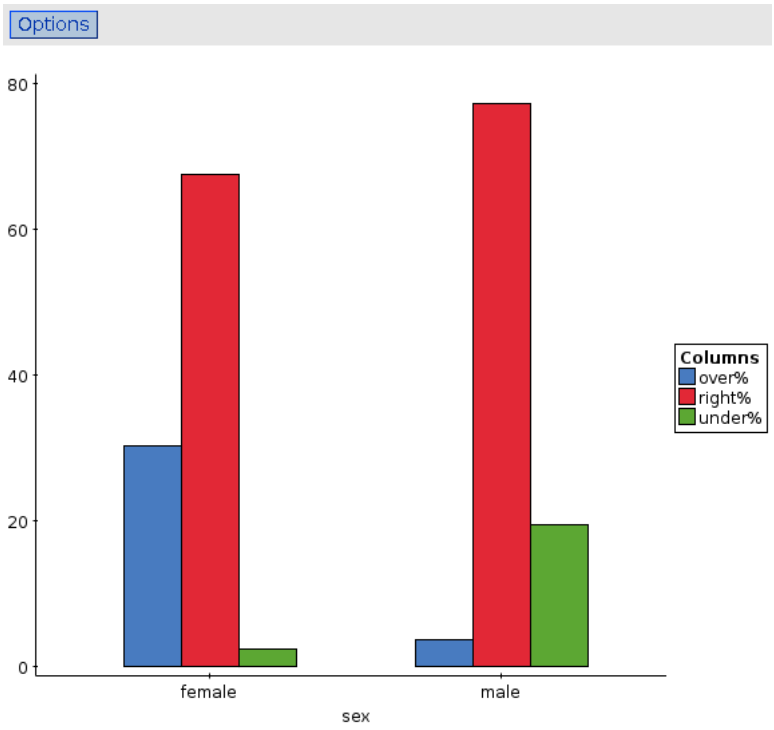
(d) * Row percentages

(e) Column percentages

Attitudes towards weight are the response variable here, in the *columns*, so the percentages you want are *row* percentages. The marginal percentages would tell you only whether there were more males than females altogether, regardless of attitude, or which attitude had the most people in it, regardless of sex. The overall percentages (joint distribution) might help some, but there are more females than males, so it wouldn't be easy to compare.

38. Consider again the survey described in Question 37.

A chart is shown below, with the vertical axis being appropriate percentages.



Which of the statements below are best supported by the chart?

(a) A similar percentage of males and females consider themselves to be of the right weight, but more males than females consider themselves to be overweight.

(b) More males than females consider themselves to be of the right weight, but fewer males than females consider themselves to be underweight.

(c) The majority of females consider themselves to be overweight.

(d) * A similar percentage of males and females consider themselves to be of the right weight, but more females than males consider themselves to be overweight.

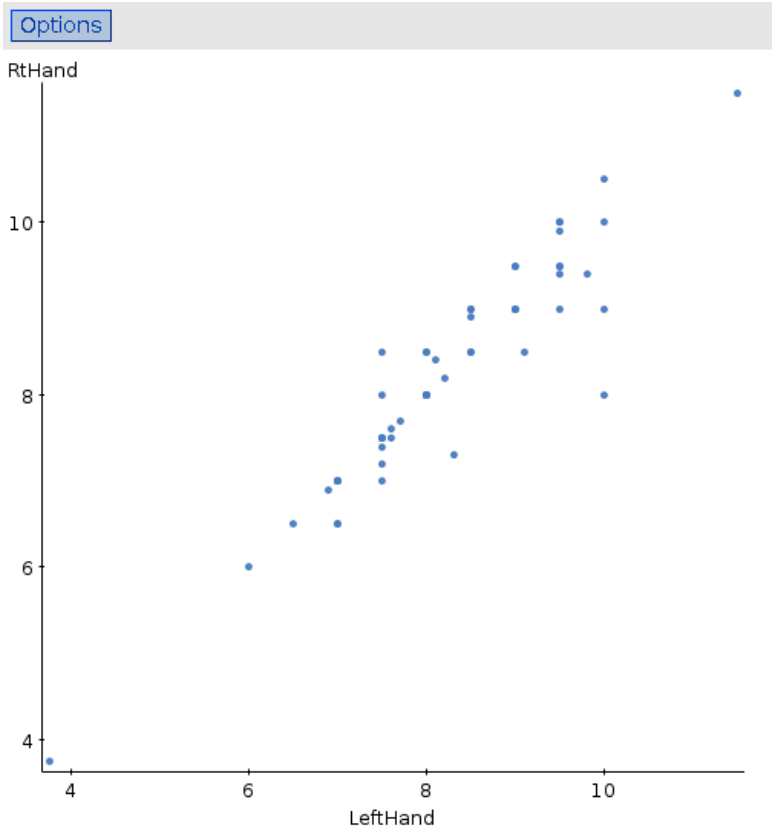
(e) The majority of males consider themselves to be underweight.

Yes, I know the fancy coloured chart didn't print very well in black and white. That means it's up to you to ask if you're not clear about which bar is which.

The "majority of" ones you can eliminate right away because most people consider themselves to be of the right weight, males and females both, and you certainly can't say that more than 50% consider themselves to be either overweight or underweight.

In (a), the second part fails; in (b) the second part fails again; in (d) the tall (red) bars are of similar height (you could reasonably call them "similar" or say that the male bar is taller than the female one), and it's true that the blue bar for females is higher than the blue bar for males.

39. A survey of hand sizes produced the results shown below for right hand and left hand sizes (in inches).



Which of the values below is the correlation between right hand and left hand sizes?

(a) * 0.9

(b) 0.99

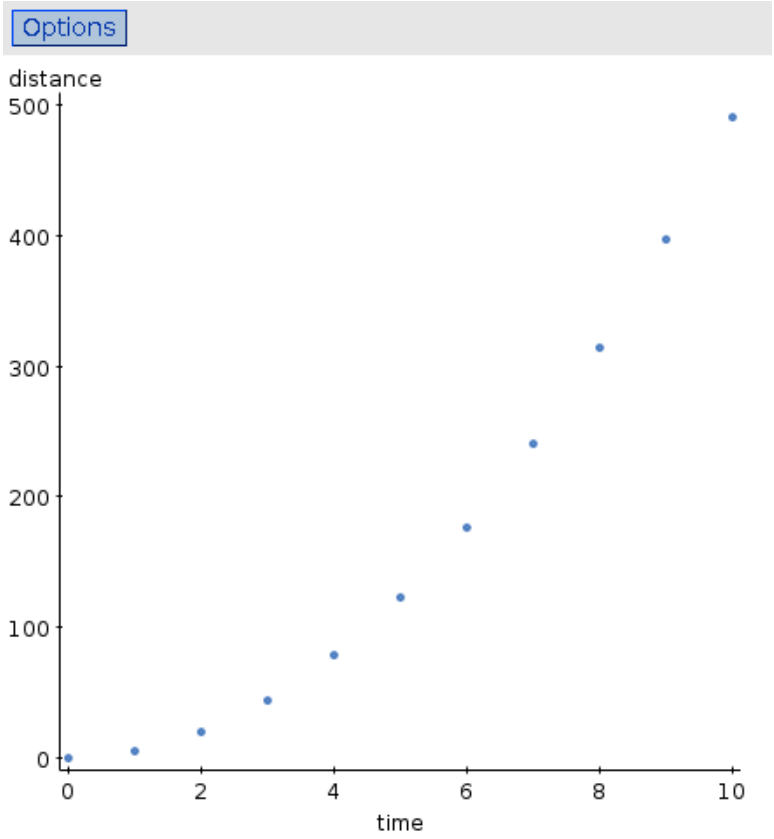
(c) 0.6

(d) -0.8

(e) we should not calculate the correlation for these data

It's a good linear relationship with a bit of scatter, heading upwards (eliminating the negative value). It's not strong enough to be 0.99, but 0.6 would look a lot weaker. (The actual correlation was about 0.91.)

40. A rock is dropped from a tall building. For each number of seconds, the distance that the rock has fallen is recorded. A scatterplot of distance against time is shown below.



You are asked to estimate the correlation between distance and time. What is your reaction?

(a) The correlation is close to 0 but not exactly 0.

(b) The correlation is about 0.9.

(c) * The relationship is not a straight line so the correlation should not be calculated.

(d) The correlation is very close to 1.

(e) The correlation is 0 because the relationship is not a straight line.

I hope your first reaction was "Curve!" and your second one was "don't calculate the correlation!".

The correlation actually *is* somewhere near 0.9 but that doesn't summarize what's going on, because, according to the physics, the association is a perfect parabola (curve), and if you have the equations of motion, you can get this association exactly right (not in this course, though). Even though it's a perfect association, it's not a straight line, so the correlation won't be close to 1. Nor is the correlation 0, or close to it, just because the relationship is curved. When you have a curved relationship, the (irrelevant) correlation can be more or less anything.