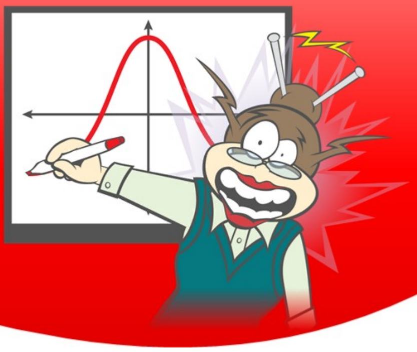
STUDENT REVISION SERIES



Data Analysis

Each of the questions included here can be solved using the TI-Nspire CAS.

Scan the QR code or use the link: Bitly link: https://bit.ly/FM_DataAnalysis

Univariate data analysis

Question: 1

Compared to the year 2018, the percentage change in the average daily circulation (sales) of 12 Australian newspapers for the year 2019 were as follows:

-4.4%, -2.2%, -2.1%, -0.9%, -3.8%, -2.6%, -2.4%, 2.5%, 0.1%, -1.1%, -3.9%, -1%

(a) Determine the five-number summary.

(b) Determine the IQR and hence determine the value of the

(i) lower fence $\text{Q1}-1.5\times\text{IQR}$

(ii) upper fence $\text{Q3}+1.5\times\text{IQR}$

(c) Determine whether any outliers exist by the $'1.5\times\text{IQR}'$ measure.

(d) Construct a

(i) box plot

(ii) box plot with outliers displayed

Question: 2

At a weather station, the wind speed (measured in knots) at 9am was recorded for 40 consecutive days.

The data obtained from the weather station is displayed below.

15, 22, 14, 12, 21, 34, 19, 11, 13, 0, 16, 4, 23, 8, 12, 18, 24, 17, 14, 3

10, 12, 9, 15, 20, 5, 19, 13, 17, 11, 16, 19, 24, 12, 7, 14, 17, 10, 14, 23

(a) What type of variable is wind speed?

(b) Construct a histogram using an appropriate interval.

(c) Describe the distribution’s main features, such as number of peaks and general shape.

(d) Determine whether any outliers exist by the $'1.5\times\text{IQR}'$ rule.

(e) Determine the proportion of wind speeds at 9am that were 20 knots or more.

(f) Estimate the mean wind speed at 9am for this sample.

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Author: Peter Flynn

Question: 3

The heights of females aged 20 to 29 are known to be normally distributed with mean 162.5 cm and standard deviation 6.9 cm.

(a) Between what heights do approximately 95% of female heights aged 20 to 29 years lie?

(b) Determine the approximate percentage of females aged 20 to 29 that are taller than 169.4 cm.

(c) If the heights of 1,000,000 randomly chosen females aged 20 to 29 were measured, determine approximately how many would be expected to be taller than 183.2 cm.

(d) If a female aged 20 to 29 years has a height of 200 cm, determine how many standard deviations above the mean this is.

(e) According to this model, is it possible for a female to be taller than 220 cm?

Question: 4

The playing time of a game of AFL football is known to be approximately normally distributed with a mean of 120 minutes and a standard deviation of 5 minutes.

(a) A randomly chosen game of AFL football is selected and found to have a playing time of 143 minutes.

(i) Convert this playing time to a *z*-score.

(ii) Comment on how extreme this playing time might be.

(b) (i) Determine the playing time that corresponds to a *z*-score of -2.

(ii) Determine the approximate percentage of games that have a playing time less than the playing time found in part (b) (i).

(c) Determine the approximate percentage of games that have a playing time of less than 120 minutes.

(d) Determine the playing times between which approximately 68% of football games are expected to lie.

(e) For $-3 \le z \le 3$, determine the approximate lower and upper bound values for playing time.

(f) In a sample of 1500 games, determine approximately the number of games that would be expected to have a playing time of less than 105 minutes.

Question: 5

The weights of medium-sized strawberries are known to be normally distributed with a mean of 14.1 grams and a standard deviation of 0.3 grams.

(a) A medium-sized strawberry is randomly chosen, weighed and found to be 12.9 grams.

(i) Convert this weight to a *z*-score.

(ii) Comment on how extreme this weight might be.

(b) Determine what weight corresponds to a *z*-score of 2.

(c) Determine the weights between which approximately 68% of medium-sized strawberries are expected to lie.

(d) Quality control staff at a strawberry farm suggest that only medium-sized strawberries that satisfy the condition $-3 \le z \le 3$ should be accepted. Determine the approximate lower and upper weights for accepted medium-sized strawberries.

(e) Determine the approximate percentage of medium-sized strawberries that are expected to weigh more than 15 grams.

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Bivariate data analysis

Question: 6

A researcher conducts a preliminary study to determine whether mothers who smoke during pregnancy give birth to babies with lower birth weights. The following table shows the birth weights, in kilograms, of babies from 14 mothers who did not smoke during pregnancy and 15 mothers who did smoke during pregnancy.

The sample of 29 mothers were all known to the researcher.

Non-smoker	3.65	3.9	3.45	3.8	4.33	3.36	3.32	4.07	3.66	3.69	3.66	3.69	3.62	3.89	
Smoker	3.1	2.87	3.41	2.73	2.27	2.67	3.21	3.55	2.9	2.76	2.87	3.15	2.73	2.93	3.01

(a) State the two variables in the study and state what type they are.

(b) Comment on the selection of the researcher's sample for the study. Is it likely to be representative of the population?

(c) For the babies in each group, calculate the

(i) mean birth weight

(ii) standard deviation of birth weights

(iii) five-number summary of birth weights

(d) Construct parallel box plots (one for each group), showing any outliers by the $'1.5 \times \text{IQR}'$ rule.

(e) Describe briefly what these parallel box plots show.

Question: 7

A random sample of 13 adults with a driver's licence were selected to see whether alcohol consumption affected reaction time.

Each driver's reaction time was measured in a laboratory before and after drinking a specified amount of alcohol. The reaction times, measured in seconds, were as follows:

Reaction time (before)	0.67	0.70	0.67	0.65	0.66	0.74	0.73	0.74	0.74	0.76	0.69	0.68	0.63
Reaction time (after)	0.71	0.68	0.72	0.72	0.67	0.79	0.79	0.74	0.76	0.75	0.75	0.69	0.67

(a) State the two variables in the study and state what type they are.

(b) Determine the five-number summary for reaction time (before) and reaction time (after).

(c) Use parallel box plots to display this data.

(d) Use the box plots to discuss whether this data provides evidence that alcohol causes an increase in reaction time.

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Question: 8

In a study of animal behaviour, researchers collected information on the average hours that various animal species spend in dreaming and non-dreaming sleep. The data for a selected group of 14 of these animals is shown in the following table.

Species	Dreaming (hours)	Non-dreaming (hours)
African giant pouched rat	2.0	6.3
Asian elephant	1.8	2.1
Baboon	0.7	9.1
Big brown bat	3.9	15.8
Brazilian tapir	1.0	5.2
Cat	3.6	10.9
Chimpanzee	1.4	8.3
Chinchilla	1.5	11.0
Cow	0.7	3.2
Desert hedgehog	2.7	7.6
Eastern American mole	2.1	6.3
European hedgehog	4.1	6.6
Golden hamster	3.4	11.0
Mole rat	2.4	8,2

(a) State the two variables in the study and state what type they are

(b) Construct parallel box plots for this data.

(c) Use the box plots to compare and contrast the two distributions.

(d) Calculate the ratio of non-dreaming sleep to dreaming sleep for each of the 14 animals.

(e) Identify the animal that spends the highest percentage of their sleeping time in 'dreaming sleep'?

(f) Comment on this distribution of the ratio of non-dreaming sleep to dreaming sleep in comparison to the variation within the original two variables.

(g) Identify the animal that is an outlier in the distribution described in part (f).

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Answers

Question: 1

(a)

$$\boxed{-4.4 \quad -3.2 \quad -2.15 \quad -0.95 \quad 2.5}$$

(b) IQR $= 2.25$

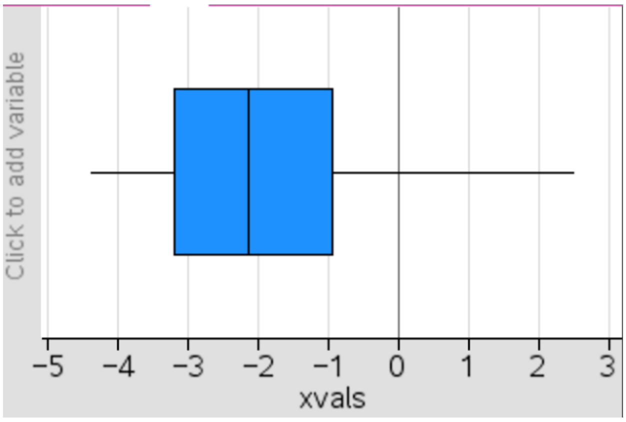
(i) -6.575

(ii) 2.425

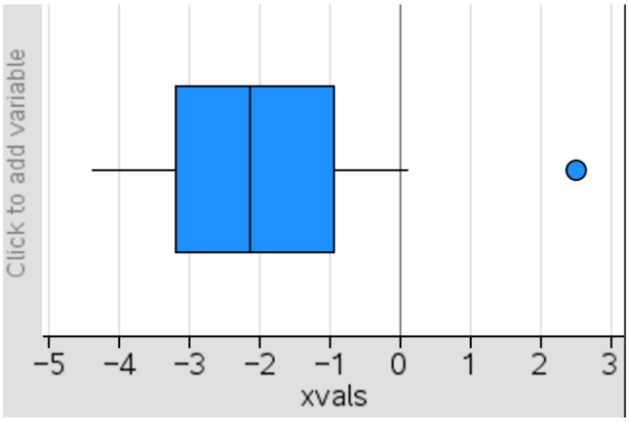
(c) 2.5% is an outlier

(d)

(i)



(ii)



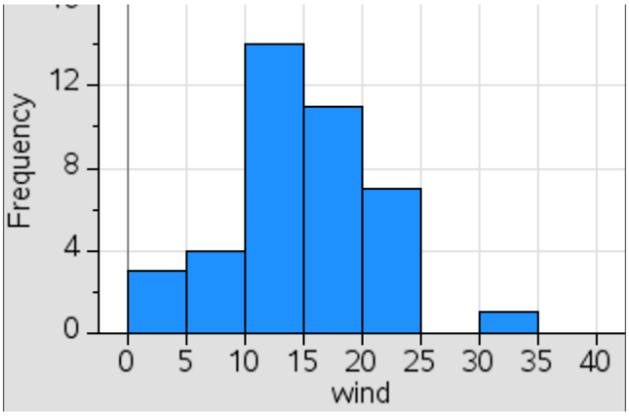
© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Author: Peter Flynn

Question: 2

(a) Numerical, continuous

(b)



(c) A single peak between 10-15 knots and reasonably symmetric. There is a possible outlier in the 30-35 interval.

(d) 34 knots is an outlier

(e) 20%

(f) Approximately 15 knots.

Question: 3

(a) [148.7,176.3]

(b) 16%

(c) 1500

(d) 5.43 standard deviations

(e) $z = 8.33$; extremely unlikely

Question: 4

(a) (i) $z = 4.6$

(ii) Very extreme, nearly 5 standard deviations from the mean.

(b) (i) 110 minutes

(ii) 2.5%

(c) 50%

(d) [115, 125]

(e) [105, 135]

(f) Approximately 2 (0.15%)

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Author: Peter Flynn

Question: 5

The weights of medium-sized strawberries are known to be normally distributed with a mean of 14.1 grams and a standard deviation of 0.3 grams.

(a) (i) $z = -4$

(ii) This is extremely light for a medium-sized strawberry. Its weight is well below the weights that 99.7% of medium-sized strawberries would exceed.

(b) 14.7 grams

(c) [13.8 grams, 14.4 grams]

(d) [13.2 grams, 15.0 grams]

(e) Approximately 0.15%.

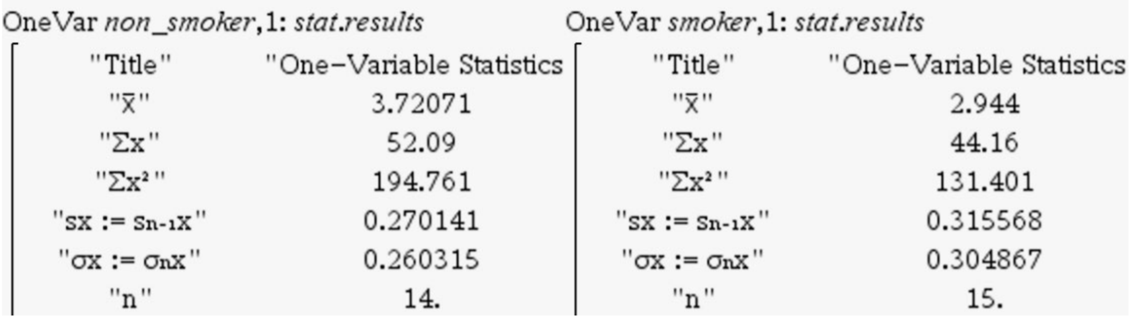
Question: 6

(a) The two variables are birth weight (continuous numerical) and smoker status (categorical).

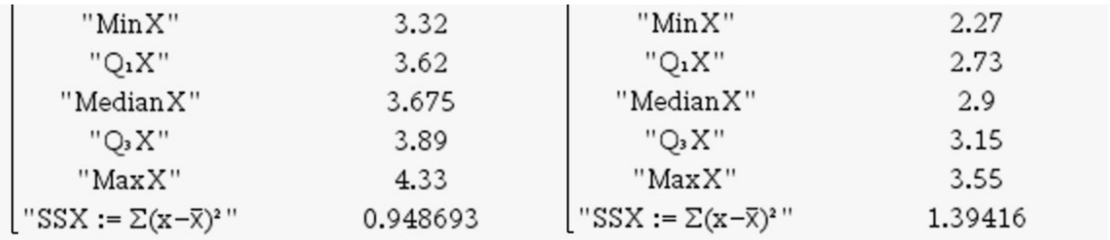
(b) The sample chosen is not random. Women were chosen on the basis of being ‘known’ by the researcher. This makes any conclusions questionable. The sample size is not large enough to make any more definitive conclusions about the impact of smoking on birth weights.

(c) (i), (ii) and (iii): with the data entered in column A (non-smoker) and column B (smoker), use of **One-Variable Statistics** gives the following:

Non-smokers Smokers



```
OneVar non_smoker,1: stat.results          OneVar smoker,1: stat.results
 "Title"         "One-Variable Statistics    "Title"         "One-Variable Statistics
 "x̄"             3.72071                     "x̄"             2.944
 "Σx"            52.09                       "Σx"            44.16
 "Σx²"           194.761                     "Σx²"           131.401
 "sx := sn-1x"   0.270141                    "sx := sn-1x"   0.315568
 "σx := σnx"     0.260315                    "σx := σnx"     0.304867
 "n"             14.                         "n"             15.
```

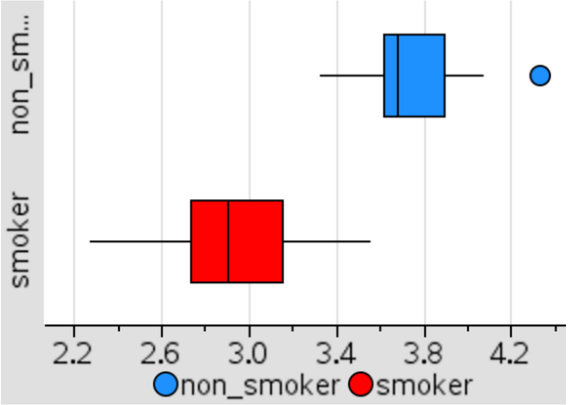


```
 "MinX"              3.32                     "MinX"              2.27
 "Q₁X"               3.62                     "Q₁X"               2.73
 "MedianX"           3.675                    "MedianX"           2.9
 "Q₃X"               3.89                     "Q₃X"               3.15
 "MaxX"              4.33                     "MaxX"              3.55
 "SSX := Σ(x-x̄)²"    0.948693                 "SSX := Σ(x-x̄)²"    1.39416
```

The summary statistics shows higher median and mean birth weights for babies whose mothers did not smoke. They also had less weight variation (lower range, IQR and standard deviation).

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

(d)



(e) The results are quite striking and mirror the comments made in part (c). The outlier in the non-smoking data accentuates the difference between the distributions, further highlighting the relatively low variability in the birth weights for babies whose mothers did not smoke.

Question: 7

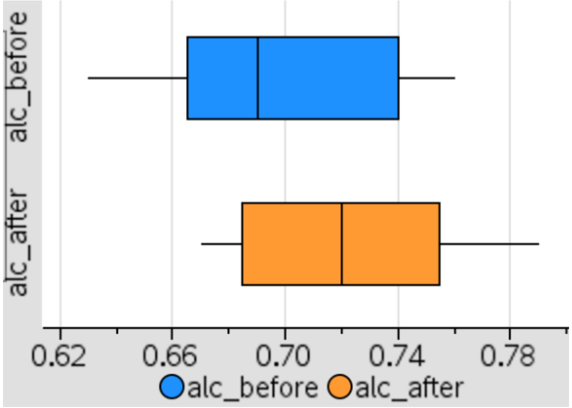
(a) The two variables are reaction time (continuous numerical) and alcohol status (categorical).

(b)

0.63 0.665 0.69 0.74 0.76

0.67 0.685 0.72 0.755 0.79

(c) Parallel box plots



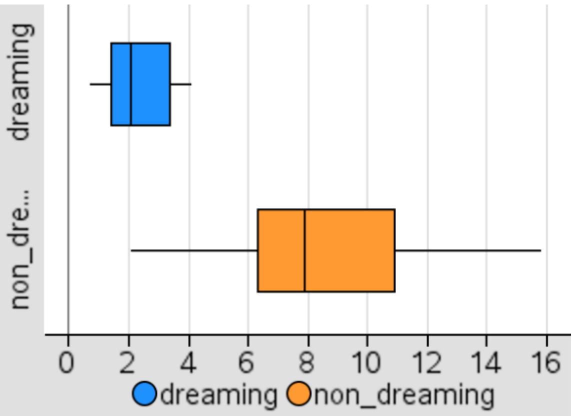
(d) The distribution for ‘after alcohol’ is centred at a higher reaction time, but more data would be needed before making any stronger conclusions.

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.

Question: 8

(a) The two variables are sleeping time (continuous numerical) and sleeping status (categorical).

(b)



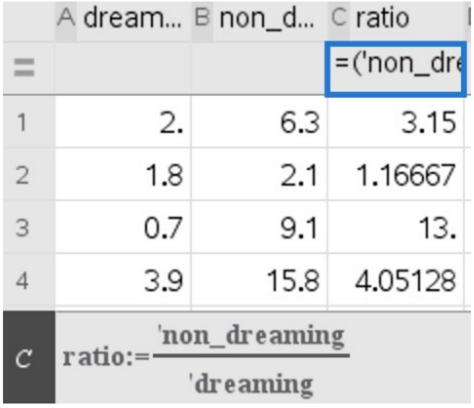
(c) All the animals sampled spent more time in non-dreaming sleep than dreaming sleep with the dreaming sleep ranging from 1 hour to 4 hours.

The plots show that the distribution of dreaming sleep is slightly positively skewed and concentrated around 1.5 hours to 3.5 hours (50% of animals in this range).

The median for non-dreaming sleep was 7.9 hours.

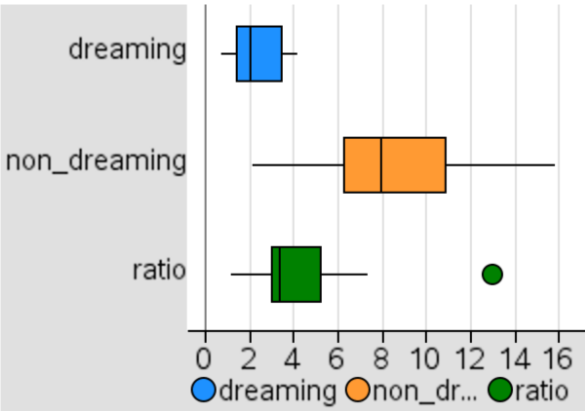
There was greater variation (spread) in the amount of non-dreaming sleep hours and was more affected (than dreaming sleep) by the total number of hours particular animals sleep each day.

(d)



	A dream...	B non_d...	C ratio
=			=('non_dr
1	2.	6.3	3.15
2	1.8	2.1	1.16667
3	0.7	9.1	13.
4	3.9	15.8	4.05128

C $\text{ratio} := \frac{\text{'non\_dreaming}}{\text{'dreaming}}$



(e) Asian elephant (approx. 46%). This was calculated using 1.8/(1.8+2.1) and then multiplying by 100 to convert to a percentage.

(f) This ratio is positively skewed (the median ratio 3.33 is closer to the Q1 ratio which is 3) with 25% of the animals spending around 3 times as much in non-dreaming sleep compared with dreaming sleep ($Q1 = 3$).

(g) The outlier is the baboon, which spends 13 times as much in non-dreaming sleep compared with dreaming sleep.

© Texas Instruments 2020. You may copy, communicate and modify this material for non-commercial educational purposes provided all acknowledgements associated with this material are maintained.