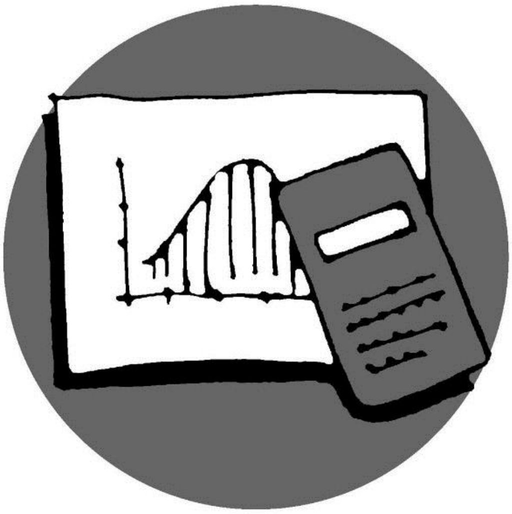
Descriptive Statistics

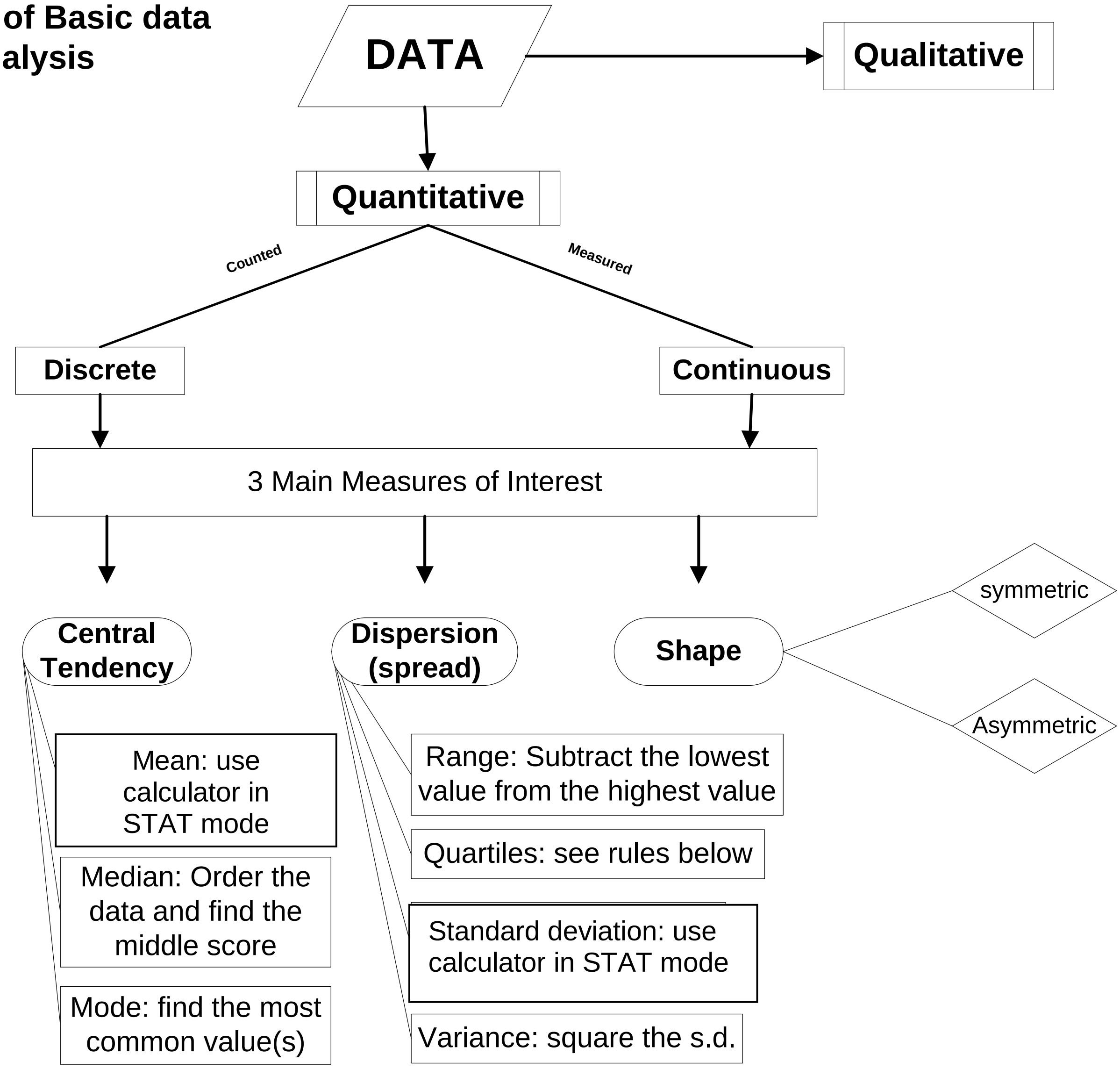


Library, Teaching & Learning
2014

Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND

New Zealand's specialist land-based university

Summary of Basic data Analysis


DATA
Qualitative
Quantitative
Counted
Measured
Discrete
Continuous
3 Main Measures of Interest
Central Tendency
Dispersion (spread)
Shape
symmetric
Asymmetric
Mean: use calculator in STAT mode
Median: Order the data and find the middle score
Mode: find the most common value(s)
Range: Subtract the lowest value from the highest value
Quartiles: see rules below
Standard deviation: use calculator in STAT mode
Variance: square the s.d.

To find the median and quartiles

- For the lower quartile, median and upper quartiles respectively ie the 25th, 50th and 75th percentiles:

  Calculate 25% 50% or 75% of n.
- if the answer is an integer, the quartile or median is the average of scores which correspond to this order data value and the next.
- if the answer is not an integer, the quartile or median is the score which corresponds to the next data value in order.

Interquartile Range:
The difference between the upper and lower quartiles (ie UQ - LQ)

Semi-interquartile Range :
Half the interquartile range

Outliers:
Any values that are ***less*** than L.Q. minus 1.5 times the IQR or any values that are ***greater*** than U.Q. plus 1.5 times the IQR

General terms and symbols you should be conversant with.

Binomial
Variation
Normal
Probability
Z- score

Continuous
Contingency
Central tendency
Range
Population

Discrete
Variable
Mean
Standard Normal
Sample

Coefficient
Random
Measures of spread
Standard Score
Variance

These are words that have a quite specific and rigorous meaning in the world of mathematics and statistics. Each will need learning as it applies in the context in which you meet it

USEFUL METHODS OF DATA DISPLAY

In the following example, heights of trees in two orchards are measured (in metres). The results of the two samples (in order) are:

Orchard A	Orchard B
10.0, 11.2, 11.4, 12.5, 12.7, 13.8, 13.8, 13.9, 14.0, 14.0, 14.3, 14.7, 14.7, 14.7, 15.1, 15.1, 15.5, 15.6, 15.8, 15.9	9.8, 9.9, 10.1, 10.3, 11.0, 11.2, 11.4, 11.6, 12.3, 12.4, 12.5, 12.7, 12.8, 13.2, 13.3, 13.7, 14.0, 14.1, 14.6, 15.3

To construct a **stem plot** (stem and leaf display), each measurement is considered to be made up of two parts

- the stem, which consists of the leading digit(s) and
- the leaf, which is usually the last digit

To represent the values 15 and 115 in a stem and leaf plot, we classify the leading digit, ‘1’ in 15 as the stem and ‘5’ as the leaf, while for 115, the leading pair ‘11’ would be the stem and ‘5’ the leaf.

In doing so, we will get the following stem plot (a) for the trees of orchard B and, in addition, (b), a **back to back** stem plot for the data from both orchards.

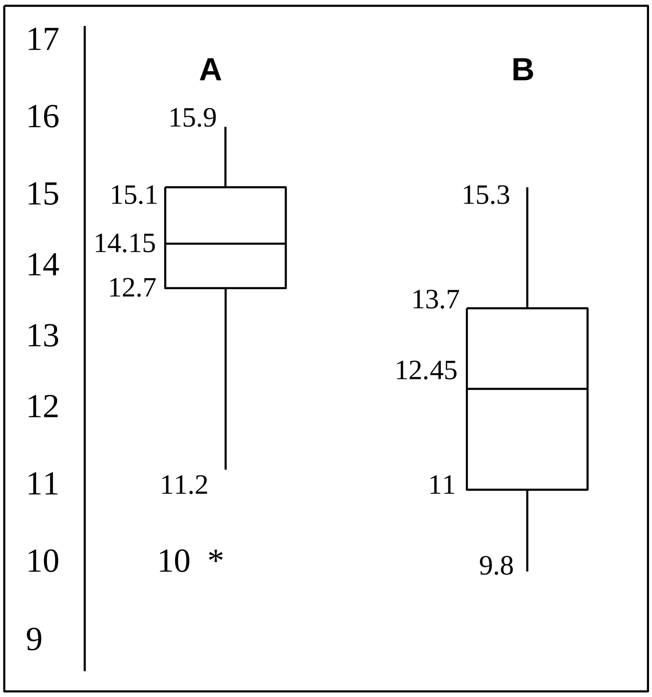
(a) stem plot – one distribution (b) stem plot – two distributions

Orchard B	
9	8 9
10	1 3
11	0 2 4 6
12	3 4 5 7 8
13	2 3 7
14	0 1 6
15	3

Orchard A		Orchard B
	9	8 9
0	10	1 3
4 2	11	0 2 4 6
7 5	12	3 4 5 7 8
9 8 8	13	2 3 7
7 7 7 3 0 0	14	0 1 6
9 8 6 5 1 1	15	3

This enab t “feel” for the diffe orchards. i.e. we can compare the distributions of the two samples by looking at the general shape of each as they lie next to one another.

This display also allows us to very easily find the 5 summary statistics to draw a **box and whisker** graph – an even more informative graphical display and comparison:



By comparing the ranges, (note the outlier as shown in orchard A) inter-quartile ranges, medians, we can see distinct differences in the two orchards.

To Draw the Box and Whisker:

MEDIAN: $n = 20 \Rightarrow \frac{n+1}{2} = \frac{21}{2} = 10.5$, median is the average of the 10th and 11th scores – that is for Orchard A $\frac{14.0 + 14.3}{2} = 14.15$ and for Orchard B $\frac{12.4 + 12.5}{2} = 12.45$

L.Q.: $n = 20 \Rightarrow n+1 = 21$

$\frac{21}{4} = 5.25\, \textit{rounded down} = 5$

ie for Orchard A, 5th score = 12.7

and for Orchard B 5th score = 11.0

U.Q.: $n = 20 \Rightarrow n+1 = 21$

$3 \times \frac{21}{4} = 15.75\, \textit{rounded up} = 16$

ie for Orchard A, 16th score = 15.1

and for Orchard B 16th score = 13.7

Hence box and whisker diagram is drawn as on the previous page.

Also, since mean for Orchard A = 13.935 < median, distribution is positively skewed,
and mean for Orchard B = 12.31, ≈ median, distribution is almost zero skewed.

FURTHER STATISTICS CALCULATED: Variance and Standard Error

Variance

The variance is related to the standard deviation.   The following process may help to explain this. (Note this is to SHOW you the process – YOU will let your calculator will do it for you!) To calculate the variance:

- **Calculate the mean of the distribution:** **e.g.** $x = 1, 2, 3, 4, 5$

$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$ $\qquad$ $\bar{x} = \frac{1+2+3+4+5}{5} = 3$

(add all the scores and divide by the number of scores)

- **Find the difference between each score and the mean:**

$(x_1 - \bar{x}), (x_2 - \bar{x}), etc$ $\qquad$ **e.g.** $(1-3) = -2,\ (2-3) = -1,\ (3-3) = 0,$ $(4-3) = 1,\ (5-3) = 2$

(subtract the mean from each individual score in the data set)

- **Square each of these results:**

$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, etc$ $\qquad$ **e.g.** $(-2)^2 = 4,\ (-1)^2 = 1,\ 0^2 = 0$ $1^2 = 1,\ 2^2 = 4$

- **Sum these squared deviations (differences)**

$\Sigma(x_i - \bar{x})^2$ $\qquad$ **e.g.** $4+1+0+1+4 = 10$

- **Calculate the 'average' of the squared deviations (divide by *n-1*)**

$\frac{\Sigma(x_i - \bar{x})^2}{n-1} = s^2$ $\qquad$ **e.g.** $\frac{10}{4} = 2.5$

This is called the variance. In words, it is the "*average squared deviation from the mean".*

One further step is required to get the standard deviation. Since the variance is the average *squared* deviation, simple take the square root of the variance and you have the standard deviation.

That is: $\sqrt{\left(\frac{\Sigma\left(x_1-\bar{x}\right)^2}{n-1}\right)}=s$ **e.g.** $\sqrt{2.5}=1.5811$

Your calculator will take you straight to the standard deviation. If you need the variance, get the s.d. and then just square that answer.

Standard Error

This is the standard deviation of the sampling distribution. Roughly, it is the average difference between the sample statistic and the population parameter. Each sample statistic will have a standard error, and a formula to go with it. Following from the above, the sample statistic would be the mean, and the corresponding standard error would be the *standard error of the mean*. Calculate the standard error of the mean as $s.e.=\frac{s}{\sqrt{n}}$. That is, get the standard deviation of the sample and divide it by the square root of the sample size. You will see this referred to as s.e, sem or s.e.m. Learn to recognise it in all its forms.

To calculate the standard error in the above worked example, $s.e.=\frac{1.5811}{\sqrt{5}}=0.7071$

PRACTICE QUESTIONS FOR DESCRIPTIVE STATISTICS

1. An auditor for Abbott & Kasteler found billing errors in invoices. In some cases the customer was billed too much (+ values) and in other cases the customer was billed too little (- values).

Customer

Size of error ($)

A Simpson	32.10
Lea & Kean	-45.00
Machine Assoc.	66.00
Fort Systems	2.53
Tension Mills	-8.10
Bridge Inc.	-51.25
Overland Toy	12.10
Valley Run	18.00
Down Kingston	21.00
Forty Nine South	-31.00
P J Smith	2.10

For these data calculate:

(a) Mean (b) Variance (c) Range (d) Median

(e) Lower quartile (f) Upper quartile (g) Standard error of the mean

2. The glorious Canterbury Crusaders Super 12 rugby team won all 12 matches to reach the final (which they also won!) Here are the winning margins (i.e., the number of points they won each game by) for the 12 games leading to the finals:

2 7 19 7 1 34 20 3 7 28 77 11

For these data points, calculate the following:

(a) Mean (b) Variance (c) Mode (d) Range
(e) Median (f) Lower quartile (g) Inter Quartile Range (IQR
(h) Standard error of the mean

3. The starting incomes ($000’s) for 11 recent BCM graduates are shown:

20.5 19.2 21.0 31.5 29.7 19.2 27.3 22.8 18.5 35.8 30.5

For these data points, calculate the following:

(a) Mean (b) Variance (c) Mode (d) Range
(e) Median (f) Lower quartile (g) Inter Quartile Range (IQR
(h) Calculate whether the upper value is an outlier

Use the following information for the next ***two*** *questions.*

A study of air pollution in a New Zealand city yielded the following daily readings of sulphur dioxide concentrations.

The stem and leaf display below details the concentrations for 20 consecutive days.

n=20
stem unit = 10.00

0	3 3 4 5 6
1	0 3 4 6 6 7 8 9
2	2 3 6 8
3	1 6
4	4

4. Calculate the mean, median and upper quartile.

5. The scientist is concerned that the highest value could be an outlier. Calculate whether it is.

---

6. The universe or ‘totality of items or things’ under consideration is called:

A a sample B a population C a parameter D a statistic

7. A summary measure that is computed to describe a characteristic from only a sample of the population is called:

A a census. B the scientific method C a parameter D a statistic

8. Which of the following is a continuous quantitative variable?

A the colour of a student's eyes.
B the number of employees of an insurance company.
C the amount of milk produced by a cow in one 24-hour period.
D the number of litres of milk sold at the local supermarket yesterday.

9. To monitor campus security, the campus security guards are taking a survey of the number of students in a parking lot each 30 minutes of a 24-hour period. If *X* is the number of students in the lot each period of time, then *X* is an example of:

A a categorical random variable. B a discrete random variable.
C a continuous random variable. D a statistic.

10. The classification of student status (e.g. 1st year, 2nd year, 3rd year, postgraduate) is an example of:

A a categorical random variable. B a discrete random variable.
C a continuous random variable. D a parameter.

11. Which of the following statistics is not a measure of central tendency?
A mean B median C mode. D Q_3 (upper quartile)

12. Health care issues are receiving much attention in both academic and political arenas. A sociologist recently conducted a survey of citizens over 60 years of age whose income is too high to qualify for a community services card. The ages of 25 senior citizens without a community services card were as follows:

60 61 62 63 64 65 66 68 68 69 70 73 73 74 75 76 76 81 81 82 86 87 89 90 92

Using the above data, find the median age of the senior citizens who do not qualify for a community services card.

A 73.04 B 73 C 68 D 13

13. Which of the following is most likely a population as opposed to a sample?

A respondents to a newspaper survey B every third person to arrive at the bank
C the first five students completing an assignment D registered voters in a county

14. Which of the following statements about the median is NOT true?

A It is more affected by extreme values than the mean.
B It is a measure of central tendency.
C It is equal to Q2.
D It is equal to the mode in bell-shaped "normal" distributions.

15. A survey was conducted to determine how people rated the quality of programming available on television. Respondents were asked to rate the overall quality from 0 (no quality at all) to 100 (extremely good quality). The Stem-and-Leaf display of the data is shown below.

Stem	Leaves
3	24
4	03478999
5	0112345
6	12566
7	01
8	
9	2

Referring to the Stem-and-Leaf Plot above, what percentage of the respondents rated overall television quality with a rating of 80 or above?

A 0% B 4% C 96% D 50%

16. The smaller the spread of data around the mean,

A the smaller the interquartile range B the smaller the standard deviation
C The smaller the coefficient of variation D All the above

17. Which of the following is NOT sensitive to extreme values?

A The range B The mean C The interquartile range D The coefficient of variation

18. The vice-chancellor of a university was concerned about alcohol abuse on campus and wanted to find out the proportion of students at his university who visited campus bars every weekend. His advisor took a random sample of 250 students.

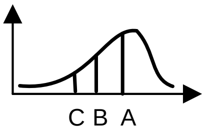
The proportion of students in the sample who visited campus bars every weekend is an example of a:

A discrete random variable B statistic C parameter D categorical random variable

19. In a right-skewed distribution the:

A median equals the mean B mean is less than the median
C mean is greater than the median D mode is less than the mean

20. For the distribution drawn here, identify the mean, median and mode



A A= Mean, B= median, C= mode
B A= Median, B= mean, C= mode
C A= Mode, B= median, C= mean
D A= Mean, B= mode, C= median

Use the following information to answer the NEXT TWO questions.

The Lincoln University Students Association wants to estimate the average amount of money spent by students on food each week. The weekly food expenditure from a sample of 7 students is shown:

$50 $35 $45 $60 $20 $45 $30

21. What is the mean?

A $285.00 B $45.00 C $40.71 D $47.50

22. What is the variance?

A $13.36 B $178.58 C $12.37 D $285.00

Solutions to PRACTICE QUESTIONS FOR DESCRIPTIVE STATISTICS

1. Data in order:

66	32.1	21	18	12.1	2.53	2.1	-8.1	-31	-45	-51.25

Answers:

a	1.68	b	1195.6	c	117.25	d	2.53
e	-31	f	21	g	10.425		

2. Data in order:

77	34	28	20	19	11	7	7	7	3	2	1

Answers:

a	18	b	456.7	c	7	d	76
e	9	f	3	g	25	h	6.17

3. Data in order:

35.8	31.5	30.5	29.7	27.3	22.8	21	20.5	19.2	19.2	18.5

Answers:

a	25.1	b	36.7	c	19.2	d	17.3
e	22.8	f	19.2	g	11.3	h	35.8 is not an outlier

4. Mean = 17.7, median = 16.5; Upper Quartile = 26

5. 44.4 is not an outlier

Solutions to questions 6-22

6.	B
7.	D
8.	C
9.	B
10.	A
11.	D
12.	B
13.	D
14.	A
15.	B
16.	D
17.	C
18.	B
19.	C
20.	C
21.	C
22.	B