Comparing sets of data

Back-to-back stem-and-leaf plots

Some of the most useful and interesting statistical investigations involve the comparison of two sets of data.
Back-to-back stem-and-leaf plots are useful to compare the distribution of two similar sets of data. This is particularly useful in the situation of controlled experiments. The two sets of data use the same central stem. One set of leaves is set to the right of the stem and the other to the left. Care must be taken when arranging the
data of the left set. Place the smallest numeral closest to the central margin, then rang outwards as the data size increases. The key generally relates to data which are presented on the right of the plot.

An example of such a plot is presented below. The data show the lifetimes of a sample of 40 batteries of each of two brands when fitted into a standard children's toy. Some of the toys are fitted with an ordinary brand battery and some with Brand X. Which battery is best?

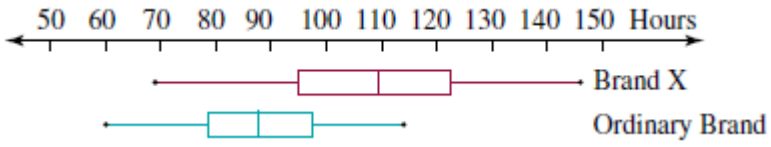
Key: 6|9 = 69

Ordinary brand Leaf	Stem	Brand X Leaf
8 6 2 0 0	6	9
9 9 9 8 8 6 4 0	7	3 5
8 8 7 5 3 1 1 1 0	8	2 4 8
9 6 6 4 2 2 2 0 0	9	0 1 4 5 5 9
8 7 5 3 1 1 1	10	0 0 2 5 8 8 9 9
4 2	11	0 0 1 1 3 3 6 7 9
	12	1 4 6 6 6 7 8 8
	13	3 5
	14	6

The spread of each set of data can be seen graphically from the stem-and-leaf plot. In this case it can be seen that, although Brand X showed a little more variability than the ordinary brand, it generally gave a longer lasting performance.

Side-by-side boxplots

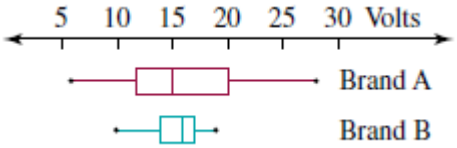
Two or more sets of data may be compared by using *side-by-side boxplots*. The boxplots share a common scale. Numerical comparisons can be made between the sets of data based upon the size and position of the range, interquartile range and median. This is a strong feature of a boxplot. In general, a histogram or stem-and-leaf plot is better than a boxplot at giving the reader information about the distribution of a set of scores, but boxplots have greater scope for making quantitative comparisons. In the case of the battery test data above, the following side-by-side boxplot would result. (Quartiles and medians are found in the usual way.)



We can make the following comparisons between the sets of data:
1. Brand X showed more variability in its performance (that is, its lifetime) than the ordinary brand. (Brand X range = 77, ordinary brand range = 54; Brand X interquartile range = 27.5, ordinary brand interquartile range = 19).
2. The longest lifetime recorded was that of a Brand X battery (146).
3. The shortest lifetime recorded was that of an ordinary brand battery (60).
4. Brand X batteries' median lifetime was better than that of the ordinary brand (Brand X median = 109.5, ordinary brand median = 87.5).
5. Over one-quarter of Brand X batteries were better performers than the best ordinary brand battery (that is, had longer lifetimes than the longest of the ordinary brand batteries' lifetimes).

Exercise Set 1

Q1. The boxplots below show the results of testing two brands of transistor by applying increasing voltage.



a) Which brand had the transistor which could withstand the highest voltage?

b) Which brand had the transistor which could withstand the least voltage?

c) Which brand gave the best median performance?

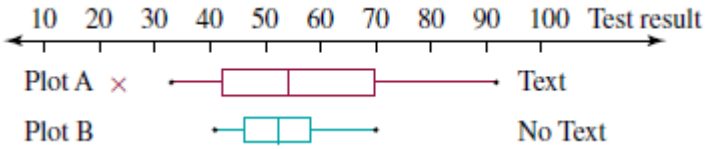
d) Which brand showed most variability in terms of range?

e) Which brand showed most variability in terms of interquartile range?

f) Which brand would you recommend to a manufacturer of electronic equipment if they requested a transistor that reliabily worked at its expected voltage?

g) Which brand would you recommend to a manufacturer of electronic equipment if they requested a transistor that was likely to withstand higher voltages?

Q2. The boxplots at right were drawn by a teacher who was trying to assess the effects of allowing students the privilege of an 'open book' exam. Plot A gives information about the results of students who were allowed to use their textbook in an end of unit test. Plot B gives information about the results of students who did the test without the aid of the textbook.



a) Which group had the student who gained the best test result?

b) Which group did best on median result?

c) Compare the variability in the performance of each group.

d) Does the use of the textbook get a better test performance? Explain.

e) What other things need to be taken into account when drawing these conclusions?

Q3. Draw side-by-side boxplots for the following pair of five figure summaries.

Group X: 14, 18.5, 21.5, 27.5, 33

Group Y: 11, 17.5, 21, 26.5, 35

Q4. The following stem-and-leaf plots give the age at marriage of a group of 10 women and a group of 10 men.

Key: 1|8 = 18 years old

Men Leaf	Stem	**Women** Leaf
8 7	1	8 8
9 8 7 5 1	2	0 2 3 4 4 5
6 3	3	0 1
0	4	

a) Draw side-by-side boxplots of the data.

b) Make comparisons about the distribution of the sets of data.

Q5. The number of words in each of the first 12 sentences is counted in each of 3 different types of book: a children's book, a grade 12 geography text, and a major daily newspaper. The results are as follows:

Children's book	6	8	12	15	6	8	10	8	5	11	10	8
Geography text	16	18	25	13	10	25	29	18	7	22	28	22
Newspaper	12	6	8	14	18	7	12	10	21	17	16	8

a) Draw side-by-side boxplots of the data.

b) Make comparisons about the sentence length of each type of publication. Use statistics in your answer.

Q6. A new spray treatment has been developed to improve the budding of apple trees. Twenty-five trees are subjected to the spray treatment while 25 others are kept as a control. The number of apples that form on each tree is recorded below.

Group A (sprayed)

35	62	71	21	34	42	76	45	48	32
29	85	73	28	34	59	62	66	27	29
33	38	54	42	61					

Group B (unsprayed)

44	42	65	41	39	68	63	62	58	51
43	47	49	45	40	52	56	50	71	39
35	37	38	62	68					

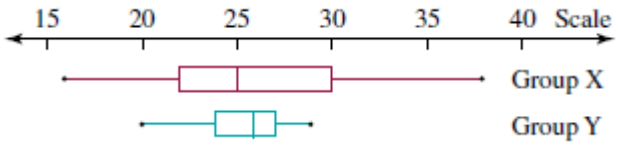
a) Detail the data on a back-to-back stem-and-leaf plot. Use a class size of 10.

b) Prepare side-by-side boxplots of the data.

c) Make comparisons of the data. Use statistics in your answer.

d) Comment on the overall effect of the spray. Would you recommend that orchardists use the spray?

Q7. Multiple Choice



Which of the following statements is a correct comparison of the data?

A Group X has a higher median and shows more variability than Group Y.

B Group X has a lower median and shows more variability than Group Y.

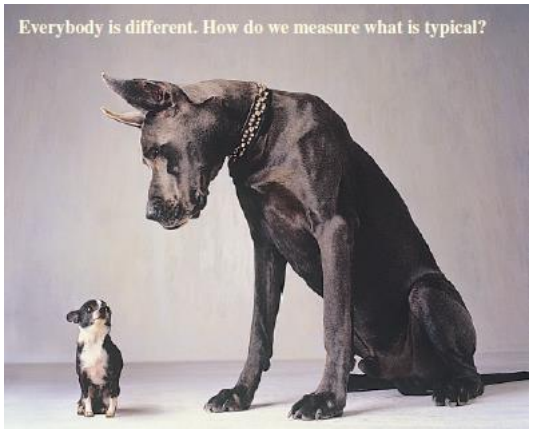
C Group X has a higher median and shows less variability than Group Y.

D Group X has a lower median and shows less variability than Group Y.

E It is impossible to make comparisons like this without seeing the data displayed on a stem-and-leaf plot.

Measures of central tendency

One of the main tasks of the statistician is to summarise large volumes of data. It is useful to be able to find one score which is typical of a whole set of data, or a few figures which can describe its distribution.



Finding the mean, median and mode are three different methods of arriving at a score which is typical or central to a set of data. Mean, median and mode are often called *measures of central tendency*.

We have already looked at the median.

The mean

The most commonly used measure is the *mean*.

The mean is the average score in the set. To find the mean all the scores are added together, then the total is divided by the number of scores. The symbol $\bar{x}$ is commonly used to denote the mean. The operation of finding the mean could be written as a formula:

$$\bar{x} = \frac{\sum x_i}{n}.$$

The symbol $\sum$ is the Greek capital letter sigma. In maths it is used to signify 'the sum of'.

So this formula $\bar{x} = \frac{\sum x_i}{n}$. could be read as: 'The mean equals the sum of all the scores (x) divided by the number of scores (n)'.

The following data give the number of pets kept in each of 10 different households:

3, 5, 4, 4, 2, 3, 0, 1, 4, 5

We could find the mean as follows:

$$\text{Mean} = \bar{x} = \frac{\sum x_i}{n} = \frac{3+5+4+4+2+3+0+1+4+5}{10} = 3.1$$

The mode

The *mode* of a group of scores is the score that occurs most often. That is, it is the score with the highest frequency.
In the case of the household pets data, a score of 4 occurs three times, making it the most frequently occurring score. It is therefore the mode of the data. In some cases there will be two or more scores which occur equally 'most often'. In such cases all of them are modes. (Don't attempt to average them.)

Example

The following data give the number of hours spent on homework by 8 students:

2, 2, 3, 0, 1, 1, 5, 1

a) Mean $\text{Mean} = \dfrac{2+2+3+0+1+1+5+1}{8}$

$= 1.875$

b) Arranging in order 0, 1, 1, 1, 2, 2, 3, 5 gives a median of 1.5

c) The mode is 1 (there are 3 one's)

If the data have been presented as a frequency table then our techniques have to be adapted a little in order to find the mean, median and mode.

Example

The following data show the number of cinema visits made by each of 20 students.

Number of visits	0	1	2	3	4
Frequency	6	7	4	2	1

This data can be arranged in a frequency table.

No of visits	**Freq**	$f \times x$	**Cumul freq**
0	6	0	6
1	7	7	13
2	4	8	17
3	2	6	19
4	1	4	20
Total		25	

The column $f \times x$ allows to easily calculate the sum of the scores, in this case, 25.

The cumulative frequency (cumul freq) column allows us to find the median score.

a) To find the mean of the data, divide the total of all scores by the number of students.

Mean = 25 ÷ 20 = 1.25

b) The scores are already arranged in order on the frequency distribution table. The cumulative frequency column tells us that the 6th score was a 0. Then all the scores up to the 13th score were all the number 1. We are interested in only the 10th and 11th scores which both must be 1. So the median is 1.

c) The mode is the score with the highest frequency. This is easy to find when the data are presented as a frequency distribution table. For the cinema data the score of 1 occurred 7 times and no other score occurred as often. The mode is 1.

The effect of extreme values

An *extreme value* (or *outlier*) is a score which is considerably different from the rest of the data in the set. The presence of extreme values may affect the representative nature of any statistics calculated.
Consider again the number of household pets data that were introduced earlier in this section: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5.
We have already determined the mean, median and mode:
Mean = 3.1 Median = 3.5 Mode = 4.
Let's include an extra household in this survey — the household of dear old Mrs
Kindheart and her 99 cats! How would this extra piece of data affect the mean, median and mode?
There would be 11 pieces of data: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5, 99.
We would find that:
Mean = 11.82 Median = 4 Mode = 4. (Check these calculations for yourself.)
The inclusion of an extreme value has dramatically increased the mean of the data, marginally increased the value of the median and left the mode unchanged. The important point to learn from this illustration is that when a set of data includes extreme values then the mean may not be truly representative of typical scores in the set. In the case of the household pets data none of the houses had a number of pets like 11.82.
Under such circumstances the median or mode may be a more useful way of relating the notion of a typical score.

Exercise Set 2

Q1. Find the mean, median and mode of each of the following sets of data.

a) 6, 8, 4, 7, 6, 7, 6

b) 102, 99, 106, 108, 101, 103, 102, 105, 102, 101

c) 5.6, 5.2, 5.4, 5.3, 5.8, 5.4, 5.3, 5.4

d) 25, 27, 29, 27, 26, 28, 29, 28, 27, 28

Q2. The number of daily errors that a machine makes is recorded over a 50-day period.
The following frequency table is obtained.

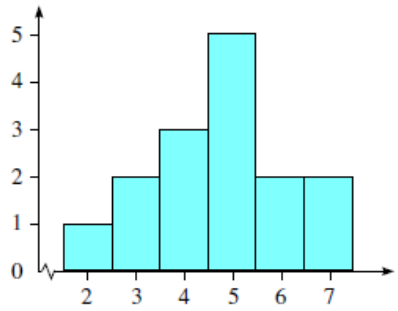
Number of errors per day	0	1	2	3	4	5
Frequency	9	18	13	6	3	1

Construct a frequency table to find the following.

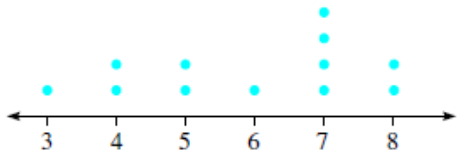
a) The mean number of errors that the machine makes in a day.
b) The median number of errors that the machine makes in a day.
c) The modal number of errors that the machine makes in a day.

Q3. For each of the following frequency distributions, find:
(i) the number of scores represented
(ii) the mean (to 1 decimal place)
(iii) the median
(iv) the mode

A.



B



Q4. Students were surveyed on the number of compact discs (CDs) they had purchased in the last 6 months, with the results shown.

No. of CDs	No. of students
0	6
1	7
2	8
3	10
4	9
5	5
6	5

Add two columns to this table to help you answer the following questions.

a) How many students were included in the survey?

b) What was the modal number of CDs purchased?

c) Calculate the mean number of CDs purchased.

d) What was the median number of CD's purchased?

e) Construct a histogram of the results and comment on the distribution.

Investigating Averages

The word **average**, as we have seen, can refer to either the mean, median or mode:
_ The **mean** is the best 'average' when there are no outliers or when a distribution is fairly symmetrical.
_ The **median** is the 'best average' when there are outliers or the distribution is skewed.
_ The **mode** is the 'best average' when referring to the most popular category.

Example
Read each statement and decide whether the mean, median or mode is the best measure of the 'average'. Justify your answer.

a) The average mark in a Maths test was 63.

Use the mean—all scores are included in this average.

b) Minh takes an average shoe size.

Mode—average here means most popular size.

c) Sydney's average house price is $300 000.

Use the median—high and low values do not affect this average.

Exercise Set 3

Q1. Decide whether the mean, median or mode is the best measure of the average in each case. Try and give a reason for your choice.

a) jeans size of the average teenage girl

b) average height of people gathered for a family party

c) average life of a light globe

d) average test mark in Mathematics

e) average daily temperature in Alice Springs in June

f) average life span for insurance purposes

g) men's average shirt size

h) average number of children per family