

Research Methods 2

Week 3: Document 1

Data Description: using numerical summaries

Background

As will be seen, the principal example used in the first part of this course (roughly weeks 3 to 7) concerns the heights of five-year-old boys. Using such a commonplace measurement has several advantages. First, it is readily understood and can easily be pictured by the reader. Second, it is one of the most widely studied of variables and its properties are well understood. This will allow the exposition of certain points to be made with greater confidence. Third, it is a variable that is widely used in clinical practice, at least where children are concerned.

It is this last point which provides the most tangible link with the practice of oncology or palliative care. The main purpose of childhood is to grow, in all its manifestations, and virtually any childhood cancer will interfere with the normal processes of growth. In many instances the treatment of the disease will also compromise growth. Consequently the patient will emerge from their treatment shorter than they would have been otherwise and with growth potential that may be less than their peers. Such children will need follow-up care which will include careful monitoring of their height and, on occasion, intervention to encourage their height to reach more acceptable levels. This can only be done satisfactorily if the nature of the distribution of heights of children can be quantified in the way we will now pursue.

Descriptive Statistics: the five number summary

Once a set of data has been collected, one of the first tasks is to describe the data. The following table contains the heights of 99 five-year-old British boys in cm.

117.9	110.2	112.9	115.9	108.0	104.6	107.1	117.9	111.8
106.3	111.0	100.4	112.1	109.2	101.0	105.4	99.4	110.1
103.3	106.9	108.2	119.3	112.0	106.2	105.9	106.9	109.3
105.9	110.0	106.7	108.5	107.7	114.3	108.6	104.6	113.7
116.7	103.5	96.1	110.8	97.2	109.6	110.5	105.9	106.2
107.4	114.9	110.3	104.8	99.2	119.2	111.4	103.0	110.1
105.8	101.5	105.9	107.6	97.1	113.3	109.4	109.4	110.8
106.3	108.1	109.6	102.4	110.4	110.1	115.3	102.9	111.2
99.4	105.7	119.5	109.3	112.8	108.2	117.0	106.8	105.4
108.7	109.2	97.1	103.3	108.8	116.3	115.5	114.9	101.1
104.1	110.8	112.7	105.6	99.9	111.1	109.4	109.1	110.7

The immediate impression is of an indigestible mass of numbers. Some of the numbers are under 100 cm, a few are above 115 cm but most seem to be between 100 cm and

115 cm. Little more can be said from this display. Some progress can be made by re-arranging the table so that the heights are in numerical order, as in the following:

96.1	101.5	105.4	106.3	108.1	109.3	110.3	111.8	115.3
97.1	102.4	105.4	106.3	108.2	109.4	110.4	112.0	115.5
97.1	102.9	105.6	106.7	108.2	109.4	110.5	112.1	115.9
97.2	103.0	105.7	106.8	108.5	109.4	110.7	112.7	116.3
99.2	103.3	105.8	106.9	108.6	109.6	110.8	112.8	116.7
99.4	103.3	105.9	106.9	108.7	109.6	110.8	112.9	117.0
99.4	103.5	105.9	107.1	108.8	110.0	110.8	113.3	117.9
99.9	104.1	105.9	107.4	109.1	110.1	111.0	113.7	117.9
100.4	104.6	105.9	107.6	109.2	110.1	111.1	114.3	119.2
101.0	104.6	106.2	107.7	109.2	110.1	111.2	114.9	119.3
101.1	104.8	106.2	108.0	109.3	110.2	111.4	114.9	119.5

This is much better and the investigator can glean a good deal of information from this presentation, such as whether there are any unusual values in the sample, as well as getting a better appreciation of the distribution of these 99 heights. However, it can hardly be said to be a succinct way to present the data, and when giving presentations, or publishing results or comparing different data sets (such as a set of heights of children from another country) and for many other purposes, more economical summaries are needed.

A common way to summarise a data set is the *five number summary*, and for these heights the five numbers are the ones highlighted in the above table.

The number in the cell with the triple outline, 108.7 cm, is the ‘middle’ value of the sample when it is placed in ascending order: it is the 50th largest value, so 49 values are smaller than it and 49 values are larger. It is known as the *median* and is a widely used measure of the *location* of a sample.

Samples can be located similarly but be quite different because they can be more or less dispersed around their location. It is therefore useful to have a measure of the *spread* of the sample. There are several possible measures and a widely used one is provided by the quartiles, which are the numbers with the double outlines in the table. The *lower quartile*, 105.6 cm, is defined as the number which is a quarter of the way from the smallest to the largest value in the sample. The *upper quartile*, 111.1 cm is three quarters of the way from the smallest to the largest value in the sample. The inter-quartile range (IQR) is defined as the difference between the figures, i.e. 5.5 cm. An alternative measure of spread is the *range*, which is the difference between the *maximum* in the sample (119.5 cm) and the *minimum* (96.1 cm), i.e. 23.4 cm.

The exact definitions of these quantities need care so that several awkward technicalities are overcome consistently. In the present example there is a value that is unequivocally the ‘middle’ value because there is an odd number of observations in the sample. Had there been an even number, for example 100 observations, there would be problems of definition: the 50th largest number would exceed 49 values but be

exceeded by 50 values, whereas the 51st largest number would exceed 50 values but be exceeded by only 49 values, so neither would be exactly in the middle, but each would have an equal claim to this status. The solution is to define the median as being half-way between the 50th and 51st values. Similar issues attend the quartiles and the requisite formulae are given in [Appendix 1](#).

The minimum, lower quartile, median, upper quartile and maximum collectively comprise the *five number summary*, which is an important way of summarising a set of data.