

# Appendix: Summary Measures for Distributions

This appendix contains information about how to calculate simple measures of central tendency and dispersal.

## Mode and Variation ratio (nominal variables)

The **mode** is simply defined as the response category with the highest frequency and can be read straight from a frequency table. Sometimes two or more responses will occur with equal frequency: in that case, the distribution is bimodal or multi-modal.

The **variation ratio** is defined as “the percentage of responses not in the modal category”:

$$\text{Variation Ratio (v)} = 100 - \% \text{ of responses in the modal category}$$

The higher the variation ratio, the greater the variation among respondents and the less useful the use of the mode as a description of the ‘typical’ or ‘normative’ response.

## Median, Range and Inter-quartile Range (ordinal variables)

The **median** is defined as “the middle case in a ranked set of cases.”<sup>1</sup> If there is an even number of cases, the median will be the mid-point between the two middle-most cases (if these cases happen to be in different response categories).

To calculate the median by hand with a small dataset:

- exclude all missing and invalid responses (for example, ‘don’t know responses must be excluded since they lack the properties of ordinal data);
- sort the cases by the relevant variable (this ranks the cases according to the variable of interest);
- count down to the middle case (or mid-most cases): the response category to which it belongs is the median.

The **range** identifies the lowest and highest ranked valid responses that received responses. Where the range is used to measure dispersal in interval

---

<sup>1</sup> David de Vaus, *Surveys in Social Research*, 361.

level data or ordinal-level data with a numerical value (such as age-band or income band information) it may also be expressed as a figure:

**Range = Value of the highest response category – value of the lowest response category.**

The range is limited as a measure of dispersal because a few extreme responses at either end of the scale can suggest a wider pattern of dispersal than is actually the case. This can be addressed by using the inter-quartile range.

The **inter-quartile range** measures the range between the 25<sup>th</sup> percentile (the first quartile) and 75<sup>th</sup> percentiles (the third quartile). The 50<sup>th</sup> percentile (the second quartile) is simply the median. Examining only the middle 50% of responses – the inter-quartile range – has the effect of removing the potentially distorting effect of extreme responses at either end of the scale. As a general rule, the smaller the inter-quartile range, the more useful the median is as a summary of the distribution.

The **median, range** and **inter-quartile range** can also be read from frequency tables that include a **cumulative frequency** column:

- the median is found by identifying the first category over the 50<sup>th</sup> percentile;
- the inter-quartile is found by identifying the first category over the 25<sup>th</sup> percentile and the first category over the 75<sup>th</sup> percentile: the middle 50% of responses will fall within these categories.

## **Means (or average) and Standard Deviations (interval variables)**

The mean or average is calculated by adding together the value of all of the cases and dividing by the total number of cases. It is computed as follows:

$$\bar{X} = \frac{x_1 + x_2 \dots + x_n}{N}$$

where  $\bar{X}$  = the mean,  $x_1$  is the value of the first case,  $x_2$  the value of the second case,  $x_n$  the value of the last case and  $N$  is the total number of valid cases.

For example, take the following set of scores: 2, 4, 6, 8, 10, 12. To calculate the mean, add the scores together and divide by the total number of scores (N)

$$\bar{X} = \frac{2 + 4 + 6 + 8 + 10 + 12}{6}$$

$$\bar{X} = \frac{42}{6}$$

$$\bar{X} = 7$$

Before calculating the mean:

- check that the variable being examined is an interval variable (i.e. response categories can be ordered in a non-arbitrary way, the difference between categories can be quantified and is identical across the range);
- exclude missing and invalid responses as this will incorrectly inflate or deflate the mean, rendering the measure useless.

The **standard deviation** is the main measure of dispersal used with interval variables. The standard deviation is a summary measure used to describe how the cases are distributed about the mean. The smaller the standard deviation, the more closely clustered the cases are around the mean and the better it functions as a description of the 'typical' or 'normative' response.<sup>2</sup>

The formula for the standard deviation is given below:

$$s = \sqrt{\frac{\sum (x - \bar{X})^2}{N}}$$

where 's' is the standard deviation,  $\sum$  is 'the sum of', x is the value of an individual response  $\bar{X}$  is the mean and N is the number of valid cases. The process for working out the standard deviation by hand is cumbersome as it involves calculating the distance of each case from the mean. When calculating the standard deviation, check that the variable being examined is an interval variable and that missing and invalid responses have been excluded.

---

<sup>2</sup> David de Vaus, *Surveys in Social Research*, 226.