Chapter 5: Exploratory Data Analysis

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

Types of univariate data

The following data types are based on distinctions that are very widely or almost universally held.

Continuous data *Continuous data* represents the values or observations of variable that can take any value in a continuous metric space such as the real line or some interval thereof. A person's height, weight, or even age are usually taken to be continuous variables, as are variables like speed, time, distance.

Categorical data *Categorical data* is where each value takes on one of a (usually, but not necessarily) finite number of values that are categorically distinct and so are not ordered, nor do they exist as points on some interval or metric space. Examples include a person's nationality, country of residence, occupation.

Types of univariate data

Ordinal data *Ordinal data* represent values of a variable that can be ordered but have no natural or uncontroversial sense of distance between them. First, second, third, and so on, are values with a natural order, but there is no general sense of distance between them. For example, knowing that three students scored in first, second, and third place, respectively, on an exam tells us nothing about how far apart their scores were.

Count data *Count data* are tallies of the number of times something has happened or some value of a variable has occurred. The number of cars stopped at a traffic light, the number of people in a building, the number of questions answered correctly on an exam, and so on, are all counts.

Characterizing univariate distributions

We can describe any univariate distribution in terms of three major features: *location*, *spread*, and *shape*. We will explore each of these in detail below through examples, but they can be defined roughly as follows.

Location The *location* or *central tendency* of a distribution describes, in general, where the mass of the distribution is located on an interval or along a range of possible values. More specifically, it describes the typical or central values that characterize the distribution. Adding a constant to all values of a distribution will change the location of the distribution, by essentially shifting the distribution rigidly to left or to the right.

Characterizing univariate distributions

Dispersion The *dispersion*, *scale*, or *spread* of a distribution of a distribution tells us how dispersed or spread out the distribution is. It tells us roughly how much variation there is in the distribution's values, or how far apart are the values from one another on average.

Shape The shape of a distribution is roughly anything that is described by neither the location or spread. Two of the most important shape characteristics are *skewness* and *kurtosis*. Skewness tells us how much asymmetry there is in the distribution. A left or negative skew means that the tail on the left (that which points in the negative direction is) is longer than that on the right, and this entails that the center of mass is more to the right and in the positive direction. Right or positive skew is defined by a long tail to the right, or in the positive direction, and hence the distribution is more massed to the left. Kurtosis is a measure of how much mass is in the center versus the tails of the distribution.

Measures of location or central tendency

Let us assume that we have a sample of n univariate values denoted by $x_1, x_2 \ldots x_i \ldots x_n$. Three commonly used measures of central tendency, at least in introductory approaches to statistics, are the arithmetic mean, the median, and the mode. Let us examine each in turn.

Arithmetic Mean

The arithmetic mean, or usually known as simply *the* mean, is defined as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

It can be seen as the centre of gravity of the set of values in the sample.

As an example, consider the following human reaction time (in milliseconds) data.

```
rt_data <- c(567, 1823,  517,  583,  317,  367,  250,  503,
              317,  567,  583,  517,  650,  567,  450,  350)
```

The mean of this data is

```
mean(rt_data)
#> [1] 558
```

Median

The median of a finite sample is defined as the middle point in the sorted list of its values. If there is an odd number of values, there is exactly one point in the middle of the sorted list of values. If there is an even number of values, there there are two points in the middle of the of the sorted list. In this case, the median is the arithmetic mean of these two points. In the `rt_data` data set, the median is as follows.

```
median(rt_data)
#> [1] 517
```

Mode

The sample mode is the value with the highest frequency. When dealing with random variables, the mode is clearly defined as the value that has the highest probability mass or density. While the mode is clearly defined for random variables, for finite samples, it is in fact not a simply matter. For example, using our `rt_data` data above, we can calculate the frequency of occurrence of all values as follows.

```
table(rt_data)
#> rt_data
#>  250  317  350  367  450  503  517  567  583  650 1823
#>    1    2    1    1    1    1    2    3    2    1    1
```

Clearly, in this case, most values occur exactly once. We can identify the value that occurs most often as follow.

```
which.max(table(rt_data)) %>% names()
#> [1] "567"
```

Measures of dispersion: Variance, standard deviation

The standard measure of the dispersion of distribution is the *variance* or *standard deviation*. These two measures should be seen as essentially one measure given that the standard deviation is simply the square root of the variance.

For a finite sample, the variance is defined as

$$\text{variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

One issue with the sample variance as just defined is that it is a *biased* estimator of the variance of the probability distribution of which $x_1, x_2 \ldots x_n$ are assumed to be a sample. An unbiased estimator of the population's variance is defined as follows.

$$\text{variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Measures of dispersion: Variance, standard deviation

Applied to our `rt_data` data, the variance is as follows.

```
var(rt_data)
#> [1] 127933.6
```

The standard deviation is the square root of this number, which is calculated by R's `sd` function, as we see in the following code.

```
var(rt_data) %>% sqrt()
#> [1] 357.6781
sd(rt_data)
#> [1] 357.6781
```

Measures of dispersion: Median absolute deviation

A much more robust alternative to the variance and standard deviation is the *median absolute deviation from the median* (MAD). As the name implies, it is the median of the absolute differences of all values from the median, and so is defined as follows.

$$\text{mad} = \text{median}(|x_i - m|)$$

In the case of the normal distribution, mad $\approx \sigma/1.48$, where σ is the distribution's standard deviation. Given this, the MAD is often scaled by approximately 1.48 so as to act as a robust estimator of the standard deviation. In R, the function the built-in command, part of the `stats` package, for calculating the MAD is `mad` and this is by default calculated as follows.

$$\text{mad} = 1.4826 \times \text{median}(|x_i - m|)$$

```
mad(rt_data)
#> [1] 98.5929
```

Range estimates of dispersion

By far the simplest measure of the dispersion of a set of values is the *range*, which the difference between the maximum and minimum values.

On the other hand, the range from the 25th to the 75th percentile, which gives the 50% inner range, is known as the *interquartile range* (IQR), which can also be calculated using the built-in `IQR` command in R.

```
IQR(rt_data)
#> [1] 208.25
```

Just as with MAD, in normal distributions, there is a constant relationship between the IQR and the standard deviation. Specifically, iqr $\approx 1.349 \times \sigma$, and so iqr/1.349 is a robust estimator of the standard deviation.

```
IQR(rt_data) / 1.349
#> [1] 154.3736
```

Measure of skewness

Skewness is a measure of the asymmetry of a distribution of numbers. In a finite sample of n values, the skewness is calculated as follows:

$$\text{skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3},$$

where $\bar{x}$ and s are the sample mean and sample standard deviation, respectively. In R, this is available as `skew` in the `psych` package.

```
psych::skew(rt_data)
#> [1] 2.66389
```

A slight variant is also available as `skewness` in the package `moments`.

```
moments::skewness(rt_data)
#> [1] 2.934671
```

In this version, the standard deviation is calculated based on a denominator of n rather than $n - 1$.

Measures of kurtosis

Kurtosis is often described as measuring how *peaky* a distribution is. However, this is a misconception and kurtosis is better understood as relating to the heaviness of a distribution's tails.

The sample kurtosis is defined as follows.

$$\text{kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{4}(x_i - \bar{x})^4}{s^4},$$

where $\bar{x}$ and s are the mean and standard deviation. This simplifies to the following.

$$\text{kurtosis} = \frac{1}{n}\sum_{i=1}^{n} z^4,$$

where $z_i = (x_i - \bar{x})/s$. This function is available from `moments::kurtosis`.

Excess kurtosis

In a normal distribution, the kurtosis, as defined above has a value of 3.0. For this reason, it is conventional to subtract 3.0 from the kurtosis function, both the population and sample kurtosis functions. This is properly known as *excess kurtosis*, but in some implementations, it is not always clearly stated that the excess kurtosis rather than kurtosis per se is being calculated.

- Distributions with zero or close to zero excess kurtosis are known as *mesokurtic*.
- Distributions with positive excess kurtosis are known as *leptokurtic*.
- Distribution with negative excess kurtosis are known as *platykurtic*.