Chapter 3. Numerical Summaries

Once a sample of data has been drawn from the population of interest, the first task of the statistical analyst might be to calculate various numerical summaries of these data.

This procedure serves two purposes:

- The first is exploratory. Calculating statistics which characterise general properties of the sample, such as location, dispersion, or symmetry, helps us to understand the data we have gathered. This aim can be greatly aided by the use of graphical displays representing the data.
- The second, as we shall see later in Chapters 9 and 10, is that these summaries will commonly provide the means for relating the sample we have learnt about to the wider population in which we are truly interested.

In this chapter we introduce some common numerical summaries used in statistics.

3.1 Summary Statistics

3.1.1 Measures of Location

The **arithmetic mean** (or just mean for short) of a sample of real values $(x_1, \ldots, x_n)$ is the sum of the values divided by their number. That is,

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

This is often colloquially referred to as the average.

Example The mean of $(8, 3, 2, 12, 5)$ is

$$\frac{8+3+2+12+5}{5} = \frac{30}{5} = 6.$$

■

For a sample of real values $(x_1, \ldots, x_n)$, define the i^{th} **order statistic** $x_{(i)}$ to be the i^{th} smallest value of the sample.

So

- $x_{(1)} \equiv \min(x_1, \ldots, x_n)$ is the smallest value;
- $x_{(2)}$ is the next smallest, and so on, up to

$\vdots$

- $x_{(n)} \equiv \max(x_1, \ldots, x_n)$ being the largest value.

Example For the sample $(8, 3, 2, 12, 5)$ we have

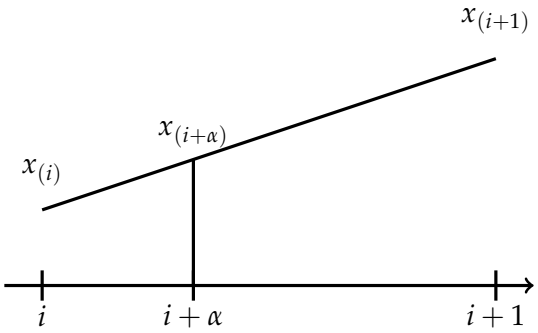
$$x_{(1)} = 2, \qquad x_{(2)} = 3, \qquad x_{(3)} = 5, \qquad x_{(4)} = 8, \qquad x_{(5)} = 12.$$

■

Furthermore, in an abuse of notation it will be useful to define $x_{(i+\alpha)}$ for integer $1 \leq i < n$ and non-integer $\alpha \in (0,1)$ as the linear interpolant

$$x_{(i+\alpha)} = (1-\alpha)\ x_{(i)} + \alpha\ x_{(i+1)},$$

where the order statistics $x_{(i)}$ are defined as before.



Example

$$x_{(4.2)} = 0.8 \times x_{(4)} + 0.2 \times x_{(5)}.$$

■

The **median** of a sample of real values $(x_1, \ldots, x_n)$ is the middle value of the order statistics. That is, using our extended notation,

$$\text{median} = x_{(\{n+1\}/2)} = \begin{cases} x_{(\{n+1\}/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example

- The median of $(7, 2, 4, 12, 5)$ is 5.
- The median of $(7, 2, 4, 12, 5, 15)$ is 6.

■

The mean is sensitive to outlying points, whilst the median is not.

Example

$(1, 2, 3, 4, 5)$ has median = mean = 3

$(1, 2, 3, 4, 40)$ has median = 3, but now mean = 10

■

The arithmetic mean is the most commonly used *location* statistic, followed by the median.

The **mode** of a sample of real values $(x_1, \ldots, x_n)$ is the value of the x_i which occurs most frequently in the sample.

Example The mode of $(3, 5, 7, 2, 10, 14, 12, 2, 5, 2)$ is 2. ■

Note Some data sets are *multimodal*.

Two other useful measures of location (other *averages*) are the geometric and harmonic mean.

For positive data, the **geometric mean** is given by

$$x_G = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Note It is easy to show that $x_G = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log x_i\right\}$, the exponential of the arithmetic mean of the logs of the data. This implies that geometric mean is less severely affected by exceptionally large values.

The **harmonic mean** is given by

$$x_H = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}\right\}^{-1}.$$

which is most useful when averaging rates.

Note For positive data $(x_1, \ldots, x_n)$,

$$\text{Arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean}.$$

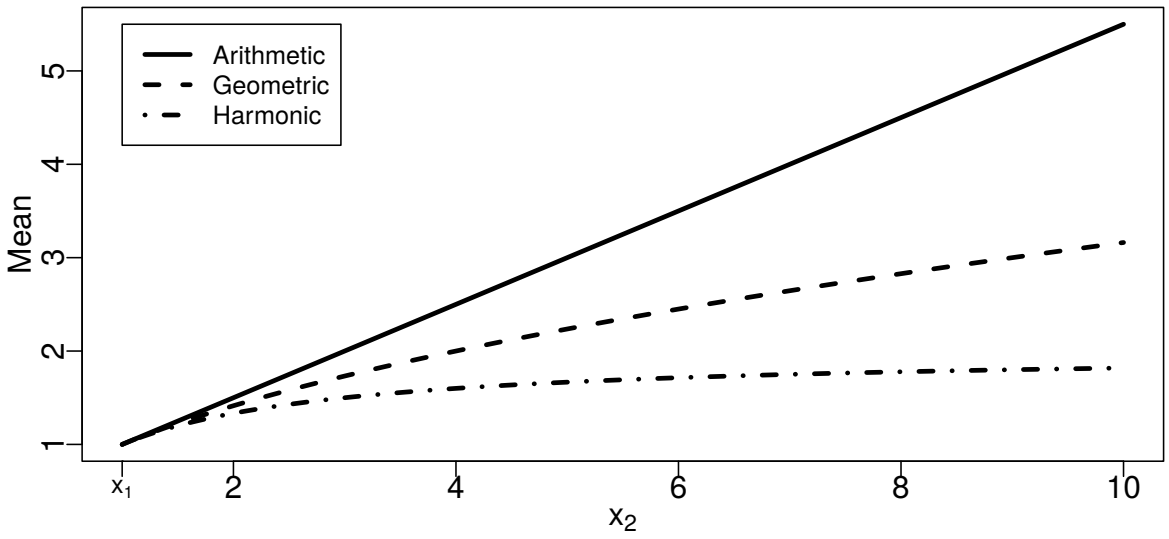


Figure 3.1: Arithmetic, geometric and harmonic means for two data points points (x_1, x_2), where $x_1 = 1$.

3.1.2 Measures of Dispersion

The **range** of a sample of real values $(x_1, \ldots, x_n)$ is the difference between the largest and the smallest values. That is

$$\text{range} = x_{(n)} - x_{(1)}$$

Example The range of $(7, 1, 4, 15, 5)$ is $15 - 1 = 14$. ■

Consider again the order statistics of a sample, $\left(x_{(1)}, \ldots, x_{(n)}\right)$.

We defined the *median* so that it lay approximately $\frac{1}{2}$ of the way through the ordered sample — not necessarily exactly or uniquely since there may be tied values or n even.

Similarly, we can define the **first** and **third quartiles** respectively as being values $\frac{1}{4}$ and $\frac{3}{4}$ of the way through the ordered sample:

$$\begin{aligned}\text{first quartile} &= x_{(\{n+1\}/4)}\\ \text{third quartile} &= x_{(3\{n+1\}/4)}\end{aligned}$$

and thus we define the **interquartile range** as the range of the data lying between the first and third quartiles,

$$\begin{aligned}\text{interquartile range} &= \text{third quartile} - \text{first quartile}\\ &= x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}\end{aligned}$$

The five point summary of a set of data lists, in order:

- The minimum value in the sample
- The lower quartile
- The sample median
- The upper quartile
- The maximum value

The most widely used measure of dispersion is based on the squared differences between the data points and their mean, $(x_i - \overline{x})^2$. The average (the mean) of these squared differences is the **mean square** or **sample variance**

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Equivalently, it is often more convenient to rewrite this formula as

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

That is, the mean of the squares minus the square of the mean.

The square root of the variance is the **root mean square** or **sample standard deviation**

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Unlike the variance, the standard deviation is in the same units as the x_i.

We can see analogies between the numerical summaries for location and dispersion, and their robustness properties are comparable.

	Least Robust	More Robust	Most Robust
Location	$\frac{x_{(1)} + x_{(n)}}{2}$	$\overline{x}$	$x_{(\{n+1\}/2)}$
Dispersion	$x_{(n)} - x_{(1)}$	s^2	$x_{(3\{n+1\}/4)} - x_{(\{n+1\}/4)}$

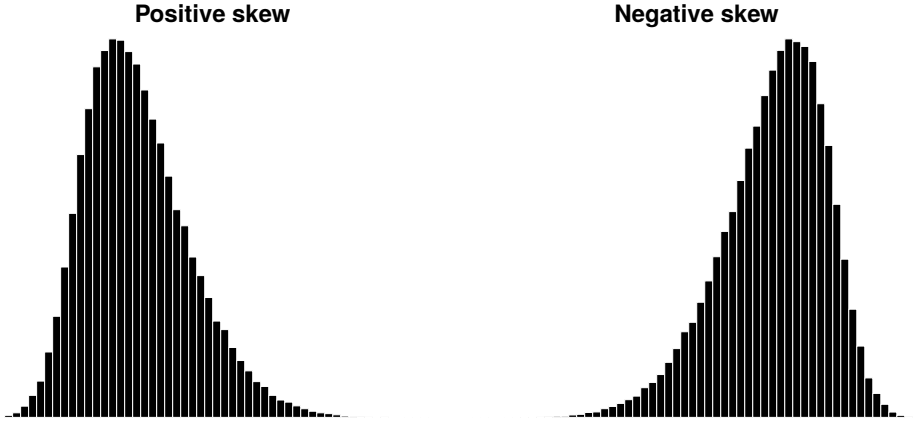
(where $\frac{x_{(1)} + x_{(n)}}{2}$ would be the midpoint of our data halfway between the minimum and maximum values in the sample, which provides another alternative descriptor of location.)

3.1.3 Skewness

Skewness is a measure of asymmetry. The **skewness** of a sample of real values $(x_1, \ldots, x_n)$ is given by

$$\text{skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3.$$

A sample is **positively** (**negatively**) or **right** (**left**) **skewed** if the upper tail of the histogram of the sample is longer (shorter) than the lower tail.



Since the mean is more sensitive to outlying points than is the median, one might choose the median as a more suitable measure of 'average value' if the sample is skewed.

We expect skewness for example when the data can only take positive (or only negative values) and if the values are not far from zero.

We can remove skewness by transforming the data. In the case above, we need a transformation which has larger effect on the larger values: e.g. square root, log (though beware 0 values).

Note For a positively skewed sample the mean is greater than the median.

3.1.4 Covariance and Correlation

Suppose we have a sample made up of ordered pairs of real values $((x_1, y_1), \ldots, (x_n, y_n))$. The value x_i might correspond to the measurement of one quantity x of individual i, and y_i to another quantity y of the same individual e.g. an individuals weight and height.

The **covariance** between the samples of x and y is given by

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

It gives a measurement of relatedness between the two quantities x and y.

The covariance can be rewritten equivalently as

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy}.$$

Note that the magnitude of s_{xy} varies according to the scale on which the data have been measured. The **correlation** between the samples of x and y is defined to be

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{n s_x s_y}.$$

where s_x and s_y are the sample standard deviations of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ respectively.

Unlike covariance, correlation gives a measurement of relatedness between the two quantities x and y which is scale-invariant.

3.2 Related Graphical Displays

3.2.1 Box-and-Whisker Plots

Based on the five point summary.

- Median – middle line in the box
- 3rd & 1st Quartiles – top and bottom of the box
- 'Whiskers' – extend out to any points which are within $(\frac{3}{2} \times$ interquartile range$)$ from the box
- Any extreme points out to the maximum and minimum which are beyond the whiskers are plotted individually.

Example Figure 3.2 are box plots of the counts of insects found in agricultural experimental units treated with six different insecticides (A-F).

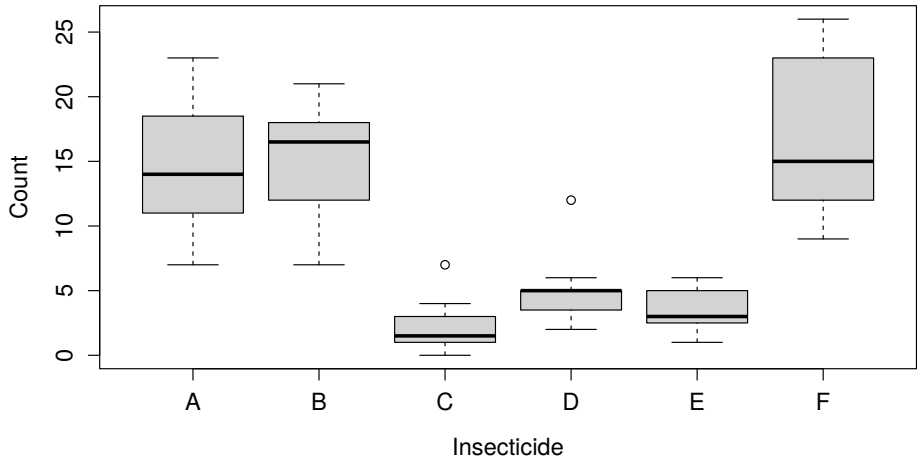


Figure 3.2

■

3.2.2 Empirical CDF

The **empirical cumulative distribution function** (**CDF**) of a sample of real values $(x_1, \ldots, x_n)$ is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

For any real number x, $F_n(x)$ returns the proportion of the data having values which do not exceed x. Note this is a step function, with change points at the sampled values.

Example Figure 3.3 is a plot of the empirical CDF of the insecticide data across the different treatments.

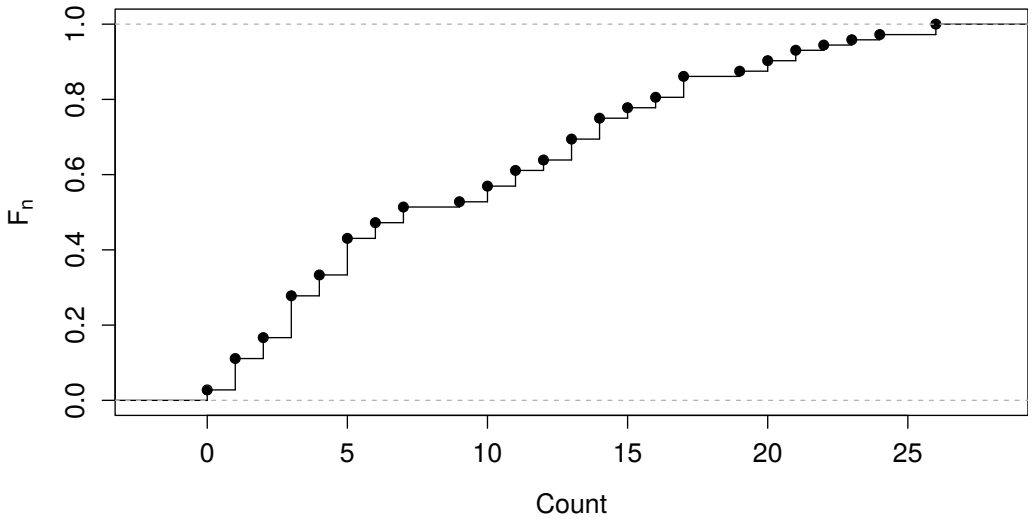


Figure 3.3

■