Measures of Spread

The mean and median measure the "location" or "centre" of a distribution or data set.

We can measure the "spread" — how widely values vary around the centre — in several ways:

- The **range** of a data set is the largest value minus the smallest value.
- The **interquartile range** (abbreviated to IQR) is the 3rd quartile minus the 1st quartile. (The textbook distinguishes between the IQR and the "Q-spread", but that is just a slight variation on the IQR.)
- The **mean absolute deviation** is the average amount by which values differ from the mean value.
- The **standard deviation** is the most common and the most theoretically important measure of spread. It's square is called the **variance**.

The Sample Variance and Sample Standard Deviation

The variance of the data points $x_1, \ldots, x_n$, denoted by s^2, is the average squared distance from the sample mean, $\bar{x}$:

$$s^2 \;=\; \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Why do we divide by $n-1$ rather than n? We'll see later; don't worry about it now.

The sample standard deviation, denoted by s, is the square root of the sample variance.

Note: It is the standard deviation that has the right units for interpretation. For example, if $x_1, \ldots, x_2$ are measures of time in seconds, s will be a measure of how spread out these times are, in seconds. But the variance has units of "seconds squared".

So the variance is mostly of interest for theory, not practice.

Mean and Standard Deviation vs. Median and Interquartile Range

We can measure the "centre" and "spread" of a set of numbers by either the sample mean and standard deviation, or the sample median and interquartile range (IQR = 3rd quartile - 1st quartile). How do these measures differ?

- What happens to them if we move one data point by a lot? (Ie, how *resistant* are these measures?)
- When is the sample standard deviation zero? When is the IQR zero?

Most importantly: The mean and median measure different things. You need to ask which is appropriate for your purpose.

Effect of Transformations

Suppose we linearly transform our data $x_1, \ldots, x_n$ to give $y_1, \ldots, y_n$, where

$$y_i \;=\; a \,+\, b\,x_i$$

How do the mean, median, standard deviation, and IQR of the new data relate to those of the old data?

$$\begin{aligned}
\bar{y} &= a \,+\, b\,\bar{x} \\
\text{median}(y) &= a \,+\, b\,\text{median}(x) \\
SD(y) &= |b|\,SD(x) \\
IQR(y) &= |b|\,IQR(x)
\end{aligned}$$

Extreme Observations

Some data points may be quite far from the others. Such *outliers* might be due to:

1) A mistake, either in measurement, or in recording the data.

2) An extraordinary occurrence that we're not interested in. Eg, we are measuring how noisy classrooms are; one day a bomb goes off in the building.

3) An extraordinary occurrence we *are* interested in. Eg, we are counting deaths of chickens on a farm; one week, many die from a disease epidemic.

4) An extreme value that occurs in the normal course of events. Eg, we are counting theatre attendance; one day, lots of people come, for no special reason.

What to Do About Outliers

If an outlier is just a mistake, we should fix it, or ignore the bad data if we can't fix it.

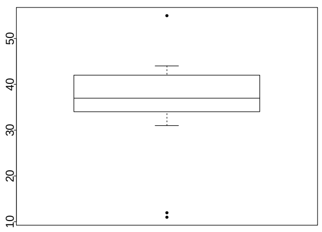
We should also ignore an outlier that is due to an extraordinary occurrence that we aren't interested in.

Other outliers should **not** be ignored. But they may cause us problems. Eg, if we have records for only one disease epidemic, it will be hard to judge how much such epidemics contribute to the total death rate.

Identifying Outliers

We may suspect that an observation is an outlier if it is far from most of the others. One "rule of thumb" is to suspect observations that are more than 1.5 times the IQR above the 3rd quartile, or below the 1st quartile.

Boxplots sometimes show such suspect observations separately, as dots or lines:



Data: 11 12 31 34 34 35 37 37 38 39 41 42 43 44 55

BUT: You should *not* automatically delete points that are this far out as being "bad". You have to **think** about these points.