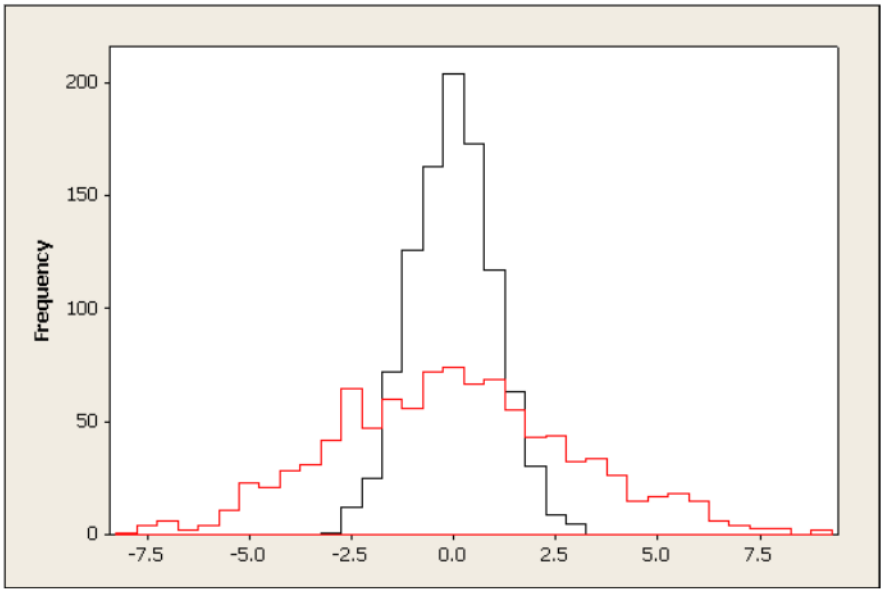
UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

3.4 Descriptive statistics – measures of spread

Central tendency is not the whole story. The two sample distributions shown below have the same mean, but they are clearly not the same. In one (red) the values have more **dispersion** (variation) than in the other.



One might imagine these represent the sample distributions of the daily returns of two stocks, both with a mean of 0%.

The black stock exhibits a smaller variation, hence we may view this as a safer stock – although there is little chance of a large *positive* daily return, there is equally little chance of a small *negative* daily return. In contrast, the red stock would be classified as a riskier stock – now there is a non-negligible chance of a large *positive* daily return, however this coincides with an equally non-negligible chance of a large *negative* daily return, i.e. a loss!

Example

A small example determining the **sum of the squared deviations from the (sample) mean**, used to calculate common measures of dispersion.

			Deviations from $\bar{X}$	
i	X_i	X_i^2	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	1	1	-3	9
2	2	4	-2	4
3	3	9	-1	1
4	5	25	$+1$	1
5	9	81	$+5$	25
Sum	20 $\bar{X} = 4$	120 $= \sum X_i^2$	0	40 $= \sum (X_i - \bar{X})^2$

Variance and standard deviation

The first measures of dispersion, the **sample variance** and its square root, the **sample standard deviation**, are based on $(X_i - \bar{X})^2$, i.e. the squared deviations from the mean.

The sample variance of a variable X, denoted S^2 (or S_X^2), is defined as:

$$\boldsymbol{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.}$$

The sample standard deviation of X, denoted S (or S_X), is the positive square root of the sample variance:

$$\boldsymbol{S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.}$$

These are the most commonly-used measures of dispersion. The standard deviation is more understandable than the variance, because the standard deviation is expressed in the same units as X (rather than the variance, which is expressed in squared units).

A useful rule-of-thumb for interpretation is that for many symmetric distributions, such as the 'normal' distribution (covered in the next section):

- about 2/3 of the observations are between $\bar{X} - S$ and $\bar{X} + S$, that is, within one (sample) standard deviation about the (sample) mean
- about 95% of the observations are between $\bar{X} - 2 \times S$ and $\bar{X} + 2 \times S$, that is, within two (sample) standard deviations about the (sample) mean.

Remember that standard deviations (and variances) are *never* negative, and they are zero *only* if all the X_i observations are the same (that is, there is no variation in the data).

Example

Consider the following simple dataset:

i	X_i	X_i^2	Deviations from $\bar{X}$ $X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	1	1	-3	9
2	2	4	-2	4
3	3	9	-1	1
4	5	25	$+1$	1
5	9	81	$+5$	25
Sum	20 $\bar{X} = 4$	120 $= \sum X_i^2$	0	40 $= \sum (X_i - \bar{X})^2$

We have:

$$S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2 = \frac{40}{4} = 10 = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{120 - 5 \times 4^2}{4}$$

and $S = \sqrt{S^2} = \sqrt{10} = 3.16$.

Sample quantiles

The median, q_{50}, is basically the value which divides the sample into the smallest 50% of observations and the largest 50%. If we consider other percentage splits, we get other (sample) **quantiles (percentiles)**, q_c. Some special quantiles are given below.

- The **first quartile**, q_{25} or Q_1, is the value which divides the sample into the smallest 25% of observations and the largest 75%.
- The **third quartile**, q_{75} or Q_3, gives the 75%–25% split.
- The extremes in this spirit are the **minimum**, $X_{(1)}$ (the '0% quantile', so to speak), and the **maximum**, $X_{(n)}$ (the '100% quantile').

These are no longer 'in the middle' of the sample, but they are more general measures of *location* of the sample distribution. Two measures based on quantile-type statistics are the:

- **range**: $X_{(n)} - X_{(1)}$ = maximum − minimum
- **interquartile range (IQR)**: IQR $= q_{75} - q_{25} = Q_3 - Q_1$.

The range is, clearly, extremely sensitive to outliers, since it depends on nothing but the extremes of the distribution, i.e. the minimum and maximum observations. The IQR focuses on the middle 50% of the distribution, so it is completely insensitive to outliers.

Boxplots

A boxplot (in full, a box-and-whiskers plot) summarises some key features of a sample distribution using quantiles. The plot is comprised of the following.

- The line inside the box, which is the median.
- The box, whose edges are the first and third quartiles (Q_1 and Q_3). Hence the box captures the middle 50% of the data. Therefore, the length of the box is the interquartile range.
- The bottom whisker extends either to the minimum or up to a length of 1.5 times the interquartile range below the first quartile, whichever is closer to the first quartile.
- The top whisker extends either to the maximum or up to a length of 1.5 times the interquartile range above the third quartile, whichever is closer to the third quartile.
- Points beyond 1.5 times the interquartile range below the first quartile or above the third quartile are regarded as outliers, and plotted as individual points.

A much longer whisker (and/or outliers) in one direction relative to the other indicates a skewed distribution, as does a median line not in the middle of the box. The boxplot below is of GDP per capita using the sample of 155 countries.