UNIVERSITY OF LONDON

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

Module Leader: Dr James Abdey

3.3 Descriptive statistics – measures of central tendency

Frequency tables, bar charts and histograms aim to summarise the *whole* sample distribution of a variable. Next we consider descriptive statistics which summarise (describe) *one* feature of the sample distribution in a single number: **summary (descriptive) statistics**.

We begin with **measures of central tendency**. These answer the question: where is the 'centre' or 'average' of the distribution?

We consider the following measures of central tendency:

- **mean (i.e. the average, sample mean or arithmetic mean)**
- **median**
- **mode**.

Notation for variables

In formulae, a generic variable is denoted by a single letter. In this course, usually X. However, any other letter (Y, W etc.) could also be used, as long as it is used consistently. A letter with a subscript denotes a single observation of a variable.

We use X_i to denote the value of X for unit i, where i can take values $1, 2, 3, \ldots, n$, and n is the sample size. Therefore, the n observations of X in the dataset (the sample) are $X_1, X_2, X_3, \ldots, X_n$. These can also be written as X_i, for $i = 1, \ldots, n$.

Let $X_1, X_2, \ldots, X_n$ (i.e. X_i, for $i = 1, \ldots, n$) be a set of n numbers. The sum of the numbers is written as:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n.$$

This may be written as $\sum_i X_i$, or just $\sum X_i$.

Other versions of the same idea are:

- infinite sums: $\sum_{i=1}^{\infty} X_i = X_1 + X_2 + \cdots$
- sums of sets of observations other than 1 to n, for example:

$$\sum_{i=2}^{n/2} X_i = X_2 + X_3 + \cdots + X_{n/2}.$$

The sample mean

The **sample mean** ('arithmetic mean', 'mean' or 'average') is the most common measure of central tendency. The sample mean of a variable X is denoted $\bar{X}$.

It is the 'sum of the observations' divided by the 'number of observations' (sample size) expressed as:

$$\boldsymbol{\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.}$$

For example, the mean $\bar{X} = \sum_i X_i/n$ of the numbers 1, 4 and 7 is:

$$\frac{1+4+7}{3} = \frac{12}{3} = 4.$$

The frequency table of the level of democracy is:

Level of democracy X_j	Frequency f_j	%	Cumulative %
0	35	22.6	22.6
1	12	7.7	30.3
2	4	2.6	32.9
3	6	3.9	36.8
4	5	3.2	40.0
5	5	3.2	43.2
6	12	7.7	50.9
7	13	8.4	59.3
8	16	10.3	69.6
9	15	9.7	79.3
10	32	20.6	100
Total	155	100	

If a variable has a small number of distinct values, $\bar{X}$ is easy to calculate from the frequency table. For example, the level of democracy has just 11 different values which occur in the sample $35, 12, 4, \ldots, 32$ times each, respectively.

Suppose X has K different values $X_1, X_2, \ldots, X_K$, with corresponding **frequencies** $f_1, f_2, \ldots, f_K$. Therefore, $\sum_{j=1}^{K} f_j = n$ and:

$$\bar{X} = \frac{\sum_{j=1}^{K} f_j X_j}{\sum_{j=1}^{K} f_j} = \frac{f_1 X_1 + \cdots + f_K X_K}{f_1 + \cdots + f_K} = \frac{f_1 X_1 + \cdots + f_K X_K}{n}.$$

In our example, the mean of the level of democracy (where $K = 11$) is:

$$\bar{X} = \frac{35 \times 0 + 12 \times 1 + 4 \times 2 + \cdots + 32 \times 10}{35 + 12 + 4 + \cdots + 32} = \frac{0 + 12 + 8 + \cdots + 320}{155} \approx 5.3.$$

Why is the mean a good summary of the central tendency?

Consider the following small dataset:

		Deviations:			
		from $\bar{X}$ $(= 4)$		from the median $(= 3)$	
i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$X_i - 3$	$(X_i - 3)^2$
1	1	-3	9	-2	4
2	2	-2	4	-1	1
3	3	-1	1	0	0
4	5	$+1$	1	$+2$	4
5	9	$+5$	25	$+6$	36
Sum	20 $\bar{X} = 4$	0	40	$+5$	45

We see that the sum of deviations from the mean is 0, i.e. we have:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0.$$

The mean is 'in the middle' of the observations $X_1, \ldots, X_n$, in the sense that positive and negative values of the **deviations** $X_i - \bar{X}$ cancel out, when summed over all the observations.

Also, the smallest possible value of the sum of *squared* deviations $\sum_{i=1}^{n} (X_i - C)^2$ for any constant C is obtained when $C = \bar{X}$.

The sample median

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the sample values of X when *ordered* from the smallest to the largest, known as the **order statistics**, such that:

- $X_{(1)}$ is the smallest observed value (the *minimum*) of X
- $X_{(n)}$ is the largest observed value (the *maximum*) of X.

The **(sample) median**, q_{50}, of a variable X is the value which is 'in the middle' of the ordered sample.

If n is odd, then $q_{50} = X_{((n+1)/2)}$.

- For example, if $n = 3$, $q_{50} = X_{(2)}$: (1) **(2)** (3).

If n is even, then $q_{50} = (X_{(n/2)} + X_{(n/2+1)})/2$.

- For example, if $n = 4$, $q_{50} = (X_{(2)} + X_{(3)})/2$: (1) **(2) (3)** (4).

For our country data $n = 155$, so $q_{50} = X_{(78)}$. From a table of frequencies, the median is the value for which the cumulative percentage first reaches 50% (or, if a cumulative % is *exactly* 50%, the average of the corresponding value of X and the next highest value).

The ordered values of the level of democracy are:

	(.0)	(.1)	(.2)	(.3)	(.4)	(.5)	(.6)	(.7)	(.8)	(.9)
(0.)		0	0	0	0	0	0	0	0	0
(1.)	0	0	0	0	0	0	0	0	0	0
(2.)	0	0	0	0	0	0	0	0	0	0
(3.)	0	0	0	0	0	0	1	1	1	1
(4.)	1	1	1	1	1	1	1	1	2	2
(5.)	2	2	3	3	3	3	3	3	4	4
(6.)	4	4	4	5	5	5	5	5	6	6
(7.)	6	6	6	6	6	6	6	6	6	6
(8.)	7	7	7	7	7	7	7	7	7	7
(9.)	7	7	7	8	8	8	8	8	8	8
(10.)	8	8	8	8	8	8	8	8	8	9
(11.)	9	9	9	9	9	9	9	9	9	9
(12.)	9	9	9	9	10	10	10	10	10	10
(13.)	10	10	10	10	10	10	10	10	10	10
(14.)	10	10	10	10	10	10	10	10	10	10
(15.)	10	10	10	10	10	10				

For the level of democracy, the median is 6.

The median can be determined from the frequency table of the level of democracy:

Level of democracy X_j	Frequency f_j	%	Cumulative %
0	35	22.6	22.6
1	12	7.7	30.3
2	4	2.6	32.9
3	6	3.9	36.8
4	5	3.2	40.0
5	5	3.2	43.2
6	12	7.7	50.9
7	13	8.4	59.3
8	16	10.3	69.6
9	15	9.7	79.3
10	32	20.6	100
Total	155	100	

Sensitivity to outliers

For the following small ordered dataset, the mean and median are both 4:

$$1, \quad 2, \quad 4, \quad 5, \quad 8.$$

Suppose we add one observation to get the ordered sample:

$$1, \quad 2, \quad 4, \quad 5, \quad 8, \quad 100.$$

The median is now 4.5, and the mean is 20.

In general, the mean is affected much more than the median by **outliers**, i.e. unusually small or large observations. Therefore, you should identify outliers early on and investigate them – perhaps there has been a data entry error, which can simply be corrected. If deemed genuine outliers, a decision has to be made about whether or not to remove them.

Due to its sensitivity to outliers, the mean, more than the median, is pulled toward the longer tail of the sample distribution.

- For a positively-skewed distribution, the mean is larger than the median.
- For a negatively-skewed distribution, the mean is smaller than the median.
- For an exactly symmetric distribution, the mean and median are equal.

The sample mode

The **(sample) mode** of a variable is the value which has the highest frequency (i.e. appears most often) in the data.

For our country data, the modal region is 1 (Africa) and the mode of the level of democracy is 0.

The mode is not very useful for continuous variables which have many different values, such as GDP per capita.

A variable can have several modes (i.e. be multimodal). For example, GDP per capita has modes 0.8 and 1.9, both with 5 countries out of the 155. The mode is the only measure of central tendency which can be used even when the values of a variable have no ordering, such as for the (nominal) region variable.