UNIVERSITY OF LONDON

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

Module Leader: Dr James Abdey

3.2 Data visualisation

Statistical analysis may have two broad aims.

1. **Descriptive statistics**: summarise the data which were collected, in order to make them more understandable.
2. **Statistical inference**: use the observed data to draw conclusions about some broader population.

Sometimes '1.' is the only aim. Even when '2.' is the main aim, '1.' is still an essential first step.

Data do *not* just speak for themselves. There are usually simply too many numbers to make sense of just by staring at them. Descriptive statistics attempt to summarise some key features of the data to make them understandable and easy to communicate. These summaries may be **graphical** or **numerical** (tables or individual summary statistics).

The statistical data in a sample are typically stored in a **data matrix**:

Countries.mtw ***

→	C1-T	C2	C3	C4	C5
	Country	region	democ	GDP	
1	Norway	3	10	37.8	
2	USA	5	10	37.8	
3	Switzerland	3	10	32.7	
4	Denmark	3	10	31.1	
5	Austria	3	10	30.0	
6	Canada	5	10	29.8	
7	Ireland	3	10	29.6	
8	Belgium	3	10	29.1	
9	Australia	6	10	29.0	
10	Netherlands	3	10	28.6	
11	Japan	2	10	28.2	
12	UK	3	10	27.7	
13	France	3	9	27.6	
14	Germany	3	10	27.6	
15	Finland	3	10	27.4	
16	Sweden	3	10	26.8	
17	Italy	3	10	26.7	
18	Singapore	2	2	23.7	

Rows of the data matrix correspond to different **units** (subjects/observations). Here, each unit is a country.

The number of units in a dataset is the **sample size**, typically denoted by n. Here, $n = 155$ countries.

Columns of the data matrix correspond to **variables**, i.e. different characteristics of the units. Here, region, the level of democracy, and GDP per capita are the variables.

Sample distribution

The **sample distribution** of a variable consists of:

- a list of the values of the variable which are observed in the sample
- the number of times each value occurs (the *counts* or *frequencies* of the observed values).

When the number of different observed values is small, we can show the whole sample distribution as a **frequency table** of all the values and their frequencies.

The observations of the region variable in the sample are:

```
3 1 1 4 2 6 3 2 2 2 3 3 1 2 4
1 4 3 1 2 1 1 2 1 5 1 4 2 4 1
1 4 1 3 4 2 3 3 1 4 2 4 1 4 1
1 3 1 6 3 3 1 1 2 3 1 3 4 1 1
4 4 4 3 2 2 2 2 3 2 3 4 2 2 2
1 2 2 2 3 1 1 1 3 3 1 1 2 1 1
1 4 3 2 1 1 2 1 2 3 4 1 1 3 6
2 2 4 4 4 2 6 3 3 2 3 3 1 1 2
2 1 3 1 2 3 3 3 2 1 1 3 3 2 2
2 1 2 1 4 1 2 2 2 1 3 3 4 5 2
4 2 2 1 1
```

We may construct a frequency table for the region variable as follows:

Region	Frequency (count)	Relative frequency (%)
		$100 \times (48/155)$
(1) Africa	48	31.0
(2) Asia	45	29.0
(3) Europe	34	21.9
(4) Latin America	23	14.8
(5) Northern America	2	1.3
(6) Oceania	3	1.9
Total	155	100

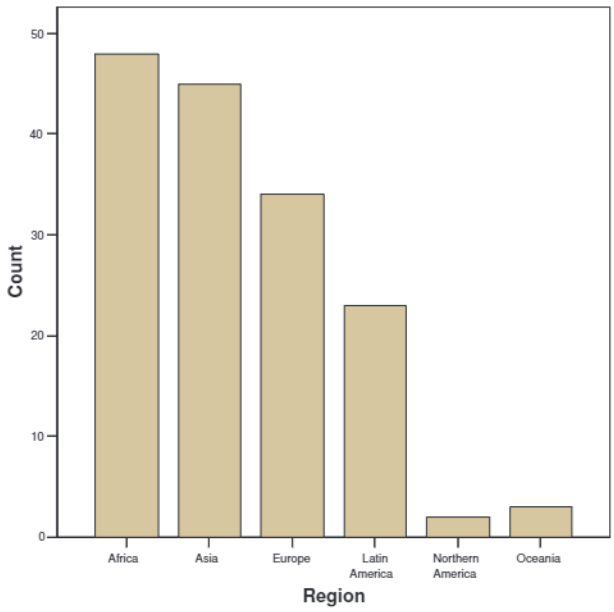
Here '%' is the percentage of countries in a region, out of the 155 countries in the sample. This is a measure of **proportion** (that is, **relative frequency**).

Similarly, for the level of democracy, the frequency table is:

Level of democracy	Frequency	%	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	

'Cumulative %' for a value of the variable is the sum of the percentages for that value and all lower-numbered values.

A **bar chart** is the graphical equivalent of the table of frequencies. The next figure displays the region variable data as a bar chart. The relative frequencies of each region are clearly visible.



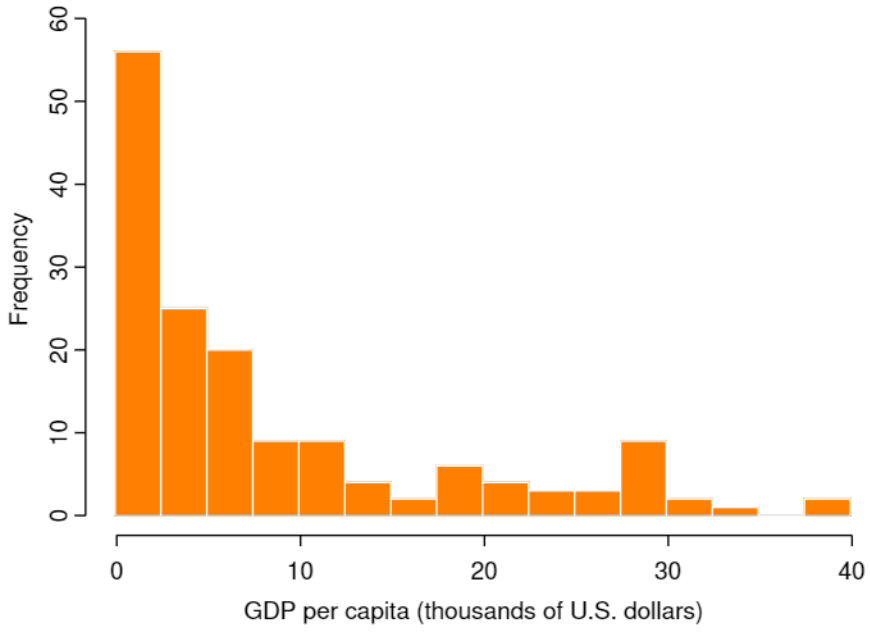
If a variable has many distinct values, listing frequencies of all of them is not very practical. A solution is to group the values into non-overlapping **intervals**, and produce a table or graph of the frequencies within the intervals.

The most common graph used for this is a **histogram**. A histogram is like a bar chart, but without gaps between bars, and often uses more bars (intervals of values) than is sensible in a table. Histograms are usually drawn using statistical software, such as Minitab, R or SPSS. You can let the software choose the intervals and the number of bars.

A table of frequencies for GDP per capita where values have been grouped into non-overlapping intervals is shown below.

GDP per capita (in $000s)	Frequency	%
[0, 2)	49	31.6
[2, 5)	32	20.6
[5, 10)	29	18.7
[10, 20)	21	13.5
[20, 30)	19	12.3
[30, 50)	5	3.2
Total	155	100

The next figure shows a histogram of GDP per capita with a greater number of intervals to better display the sample distribution.



Associations between two variables

So far, we have tried to summarise (some aspect of) the sample distribution of *one* variable at a time. However, we can also look at two (or more) variables together. The key question is then whether some values of one variable tend to occur frequently together with particular values of another, for example high values with high values. This would be an example of an **association** between the variables. Such associations are central to most interesting research questions, so you will hear much more about them in the future.

Some common methods of descriptive statistics for two-variable associations are introduced here, but only very briefly now and mainly through examples.

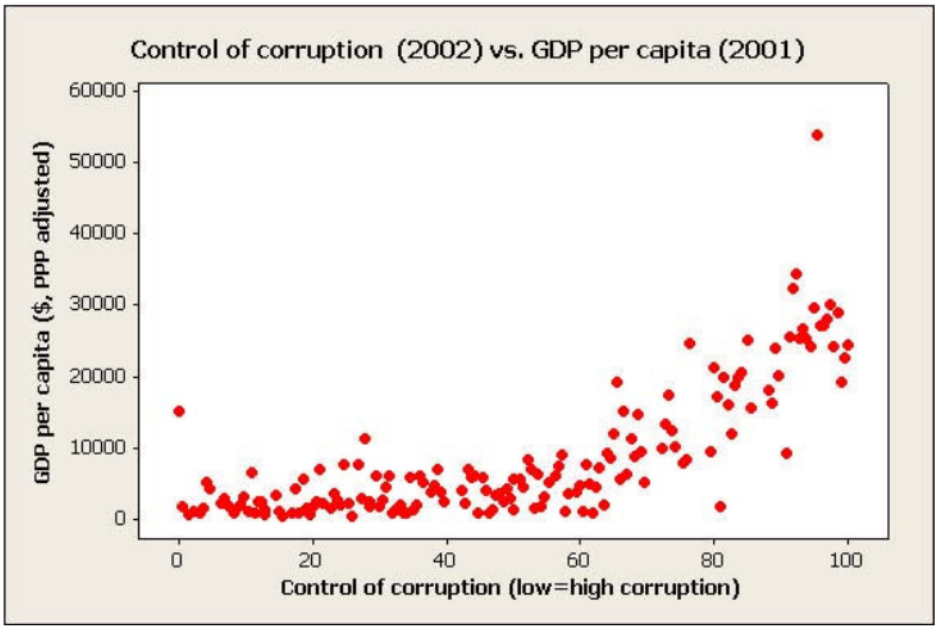
The best way to summarise two variables together depends on whether the variables have 'few' or 'many' possible values. We illustrate one method for each combination, as listed below.

- 'Many' versus 'many': scatterplots.
- 'Few' versus 'many': side-by-side boxplots.
- 'Few' versus 'few': two-way contingency tables (cross-tabulations).

A **scatterplot** shows the values of two *measurable* variables against each other, plotted as points in a two-dimensional coordinate system.

A plot of data for 164 countries is shown below which plots the following variables.

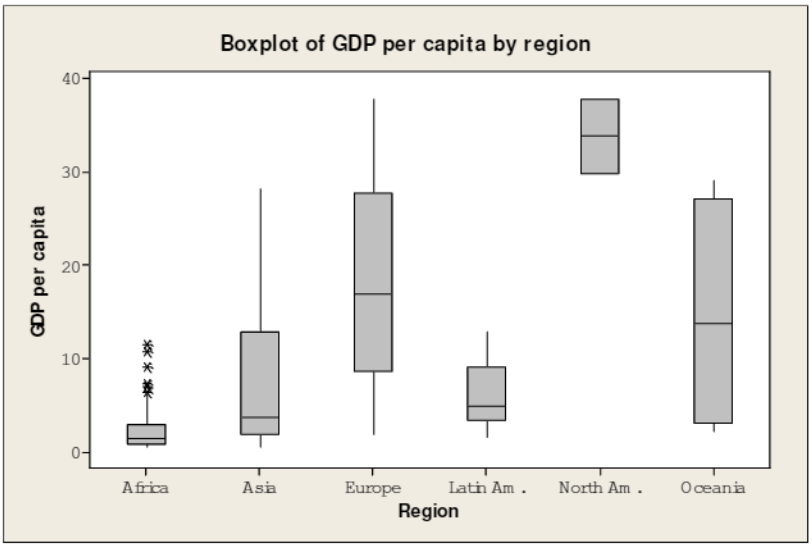
- On the horizontal axis (the x-axis): a World Bank measure of 'control of corruption', where *high* values indicate *low* levels of corruption.
- On the vertical axis (the y-axis): GDP per capita in \$.



Interpretation: it appears that virtually all countries with high levels of corruption have relatively low GDP per capita. At lower levels of corruption there is a positive association, where countries with very low levels of corruption also tend to have high GDP per capita.

Boxplots (explained in Section 3.4) are useful for *comparisons* of how the distribution of a measurable variable varies across different groups, i.e. across different levels of a categorical variable.

The figure below shows side-by-side boxplots of GDP per capita for the different regions.



- GDP per capita in African countries tends to be very low. There is a handful of countries with somewhat higher GDPs per capita (shown as *outliers* in the plot).
- The median for Asia is not much higher than for Africa. However, the distribution in Asia is very much skewed to the right, with a tail of countries with very high GDPs per capita.
- The median in Europe is high, and the distribution is fairly symmetric.
- The boxplots for Northern America and Oceania are not very useful, because they are based on very few countries (two and three countries, respectively).

A (two-way) **contingency table** (or **cross-tabulation**) shows the frequencies in the sample of each possible *combination* of the values of two categorical variables. Such tables often show the percentages within each *row* or *column* of the table.

The table below reports the results from a survey of 972 private investors. The variables are as follows.

- Row variable: age as a categorical, grouped variable (four categories).
- Column variable: how much importance the respondent places on short-term gains from his/her investments (four levels).

	Importance of short-term gains				
Age group	Irrelevant	Slightly important	Important	Very important	Total
Under 45	37	45	38	26	146
	(25.3)	(30.8)	(26.0)	(17.8)	(100)
45–54	111	77	57	37	282
	(39.4)	(27.3)	(20.2)	(13.1)	(100)
55–64	153	49	31	20	253
	(60.5)	(19.4)	(12.3)	(7.9)	(100)
65 and over	193	64	19	15	291
	(66.3)	(22.0)	(6.5)	(5.2)	(100)
Total	494	235	145	98	972
	(50.8)	(24.2)	(14.9)	(10.1)	(100)

Numbers in parentheses are percentages within the rows. For example, $25.3 = (37/146) \times 100$.

Interpretation: look at the row percentages. For example, 17.8% of those aged under 45, but only 5.2% of those aged 65 and over, think that short-term gains are 'very important'. Among the respondents, the older age groups seem to be less concerned with quick profits than the younger age groups.