The Practice of Statistics 1

Week 1: Summarising Data

Video 7: Some Features of Data

In this lecture we will summarise some of the features of data in general.

An **observational unit** is the person or thing on which measurements are taken. Observational unit is also called **case**, **subject** (when a person), **experimental unit**, etc. Examples of observational units that we have encountered already include country or territory for the life expectancy data, skeleton in the anthropological data, and soccer player in the data of salaries of New York Red Bulls players.

A **variable** is a measured characteristic on an observational unit. In the life expectancy data this includes life expectancy and region of the world. In our anthropological data, variables include age at death, sex, BMI, the estimated age at death, the difference between the actual and estimated age at death.

An observed value or **observation** is our data. It is the actual value of a variable for one of the observational units in a dataset.

Observation	Sex	BMI	Age	DGEstimate	DGDifference
1	2	underweight	78	44	-34
2	1	normal	44	32	-12
3	1	overweight	72	32	-40
4	1	overweight	59	44	-15
5	1	normal	60	32	-28
6	1	underweight	34	25	-9
7	1	overweight	50	32	-18
8	1	underweight	73	50	-23
9	1	normal	70	39	-31
10	1	normal	60	44	-16
11	1	normal	58	32	-26
12	1	overweight	61	32	-29
13	2	overweight	52	44	-8
14	1	normal	67	44	-23

Figure 1: How data are arranged

Data are arranged in rows and columns as shown in Figure 1. There is one row for each observational unit and a column for each variable. Many different file formats are possible for data, such as spreadsheets and comma separated values. In this course, we are supplying the data in plain text files with spaces separating the observed values of the variables. Note that categorical variables can be coded in multiple ways, including as numbers. In Figure 1, a sex of 1 indicates a male and 2 indicates a female.

We can classify variables as quantitaive or categorical. A **categorical variable** records into which of several categories an observation falls. We saw some examples of categorical variables in the skeletons dataset such as sex and BMI. A variable is **quantitative** if it takes numerical values for which arithmetic operations make sense. For a quantitative variable, saying that one observation is twice as large as another has a practical meaning. If one

soccer player's salary is twice the salary of another player there is a clear meaning. Similarly, saying one soccer player's salary is $10,000 more than another player's salary has a precise meaning.

Ordinal variables have a natural order. However, unlike quantitative variables, the differences between two values of an ordinal variable may not be meaningful. Sometimes ordinal variables are treated as quantitative variables. When doing so we need to be careful. An example of an ordinal variable we have seen is the BMI classifications of underweight, normal, overweight, and obese.

In earlier lectures, we talked about **outliers** and their effect on calculations of summary statistics. The defining feature of an outlier is that it is separated from the rest of the data. Outliers may have an overly influential effect on summary statistics that are not robust such as the mean and standard deviation. Outliers should be treated with care in the analysis. It is not unusual to report findings both with and without outliers. Outliers should not be ignored or routinely removed from the data since it is possible that they are errors. Cleaning data by correcting errors is an important first step in dealing with a new dataset. It is important to try to find out the reason for unusual observations. When outliers are not mistakes, they might be the most interesting feature of the data.

Sometimes, a dataset can include **missing values**. We need to consider carefully whether missing values might lead to different results than we would have if the data had not been missing. Observations that are missing for a reason can **bias** the results.

For example, let's consider the the anthropology dataset. Skeletal age estimation using the Di Gangi method is based on features of the first rib. If a skeleton has a missing or damaged first rib, it may be impossible to obtain a Di Gangi age estimate. The age estimate for this skeleton would be missing. For our data, we do not expect that a missing first rib has any relation to the age of the skeleton at time of death, or to any other features of the skeleton, such as sex, or BMI. It may be reasonable then to assume that excluding skeletons for which we have a missing value for the Di Gangi age estimate does not have any effect on our analysis and we can continue to work with the dataset without them.

In contrast, sometimes the fact that a data value is missing tells us something about what it would have been had it been observed. Imagine a scale for measuring a person's weight that only goes up to 100 kilograms. Any larger subjects who were part of a study using this scale would have a missing value for their weight. Not including them in the analyses would bias the results by, for example, having a mean weight that is smaller than what the actual mean weight should be.