UNIVERSITY OF LONDON

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

Module Leader: Dr James Abdey

3.1 Classify your variables!

Many of the questions for which people use statistics to help them understand and make decisions involve types of variables which can be measured. Obvious examples include height, weight, temperature, lifespan, rate of inflation and so on. When we are dealing with such a variable – for which there is a generally recognised method of determining its value – we say that it is a **measurable variable**. The numbers which we then obtain come ready-equipped with an order relation, i.e. we can always tell if two measurements are equal (to the available accuracy) or if one is greater or less than the other.

Data[1] are obtained on any desired **variable**. For most of this course, we will be dealing with variables which can be partitioned into two types.

1. **Discrete** data: things you can *count*. Examples include the number of passengers on a flight and the number of telephone calls received each day in a call centre. Observed values for these will be $0, 1, 2, \ldots$ (i.e. non-negative integers).
2. **Continuous** data: things you can *measure*. Examples include height, weight and time which can be measured to several decimal places.

Of course, before we do any sort of data analysis, we need to collect data. Week 4 will discuss a range of different techniques which can be employed to obtain a sample. For now, we just consider some examples of situations where data might be collected.

- A pre-election opinion poll asks 1,000 people about their voting intentions.
- A market research survey asks how many hours of television people watch per week.
- A census[2] interviewer asks each householder how many of their children are receiving full-time education.

[1] Note that the word 'data' is plural, but is very often used as if it were singular. You will probably see both forms written in texts.

[2] A census is the total enumeration of a population, hence this would not be a sample.

Categorical vs. measurable variables

A polling organisation might be asked to determine whether, say, the political preferences of voters were in some way linked to their job type – for example, do supporters of Party X tend to be blue-collar workers? Other market research organisations might be employed to determine whether or not users were satisfied with the service which they obtained from a commercial organisation (a restaurant, say) or a department of local or central government (housing departments being one important instance).

This means that we are concerned, from time to time, with *categorical variables* in addition to measurable variables. So we can count the frequencies with which an item belongs to particular categories. Examples include:

(a) the total number of blue-collar workers (in a sample)

(b) the total number of Party X supporters (in a sample)

(c) the number of blue-collar workers who support Party X

(d) the number of Party X supporters who are blue-collar workers

(e) the number of diners at a restaurant who were dissatisfied/indifferent/satisfied with the service.

In cases (a) and (b) we are doing simple *counts*, within a sample, of a single category, while in cases (c) and (d) we are looking at some kind of cross-tabulation between variables in two categories: worker type vs. political preference in (c), and political preference vs. worker type in (d) (they are not the same!).

There is no unambiguous and generally agreed way of putting worker types in order (in the way that we can certainly say that $1 < 2$). It is similarly impossible to *rank* (as the technical term has it) many other categories of interest: for instance in combating discrimination against people organisations might want to look at the effects of gender, religion, nationality, sexual orientation, disability etc. but the whole point of combating discrimination is that different 'varieties' within each category cannot be ranked.

In case (e), by contrast, there is a clear ranking – the restaurant would be pleased if there were lots of people who expressed themselves satisfied rather than dissatisfied. Such considerations lead us to distinguish two main types of variable, the second of which is itself subdivided.

- **Measurable variables** are those where there is a generally recognised method of measuring the value of the variable of interest.
- **Categorical variables** are those where no such method exists (or, often enough, is even possible), but among which:
  - some examples of categorical variables can be put in some sensible order (case (e)), and hence are called **ordinal** (categorical) variables
  - some examples of categorical variables cannot be put in any sensible order, but are only known by their name, and hence are called **nominal** (categorical) variables.

Nominal categorical variables

For a nominal variable (like gender), the numbers (values) serve only as labels or tags for identifying and classifying cases. When used for **identification**, there is a strict one-to-one correspondence between the numbers and the cases. For example, your passport or driving licence number uniquely identifies you.

Any numerical values do not reflect the amount of the characteristic possessed by the cases. Counting is the only arithmetic operation on values measured on a nominal scale, and hence only a very limited number of statistics, all of which are based on frequency counts, can be determined (such as percentages and the mode, considered in Sections 3.2 and 3.3).

Ordinal categorical variables

An ordinal variable has a **ranking scale** in which numbers are assigned to cases to indicate the relative extent to which the cases possess some characteristic. It is possible to determine *if* a case has more or less of a characteristic than some other case, but not *how much* more or less.

Any series of numbers can be assigned which preserves the ordered relationships between the cases. In addition to the counting operation possible with nominal variables, ordinal variables permit the use of statistics based on centiles such as percentiles, quartiles and the median (introduced in Section 3.3).

Interval measurable variables

Interval-level variables have scales where numerically equal distances on the scale represent equal value differences in the characteristic being measured. For example, if the temperatures on three days were 0, 10 and 20 degrees, then there is a constant 10-degree differential between 0 and 10, and 10 and 20. This allows comparisons of differences between values. The location of the **zero point** is not fixed – both the zero point and the units of measurement are arbitrary. For example, temperature can be measured in different (arbitrary) units, such as degrees Celsius and degrees Fahrenheit.

Any **positive linear transformation** of the form $y = a + bx$ will preserve the properties of the scale, hence it is not meaningful to take ratios of scale values. Statistical techniques which may be used include all of those which can be applied to nominal and ordinal variables. In addition statistics such as the mean and standard deviation (introduced in Section 3.3) are applicable.

Ratio measurable variables

Ratio-level variables possess all the properties of nominal, ordinal and interval variables. A ratio variable has an absolute zero point and it is meaningful to compute ratios of scale values.

Only **proportionate transformations** of the form $y = bx$, where b is a positive constant, are allowed. All statistical techniques can be applied to ratio data.

Example

Consider the following three variables describing different characteristics of countries. Later, we consider a sample of 155 countries in 2002 for these variables.

- Region of the country.
  - This is a *nominal* variable which could be coded (in alphabetical order) as follows: 1 = Africa, 2 = Asia, 3 = Europe, 4 = Latin America, 5 = Northern America, 6 = Oceania.
- The level of democracy, i.e. a democracy index, in the country.
  - This could be an 11-point *ordinal* scale from 0 (lowest level of democracy) to 10 (highest level of democracy).
- Gross domestic product per capita (GDP per capita) (i.e. per person, in \$000s) which is a *ratio* scale.

Region and the level of democracy are discrete, with the possible values of $1, 2, \ldots, 6$, and $0, 1, 2, \ldots, 10$, respectively. GDP per capita is continuous, taking any non-negative value.

Many discrete variables have only a *finite* number of possible values. The region variable has 6 possible values, and the level of democracy has 11 possible values.

The simplest possibility is a **binary**, or **dichotomous**, variable, with just *two* possible values. For example, a person's gender could be recorded as 1 = female and 2 = male. A discrete variable can also have an unlimited number of possible values. For example, the number of visitors to a website in a day: $0, 1, 2, 3, 4, \ldots$.

The levels of democracy have a meaningful ordering, from less democratic to more democratic countries. The numbers assigned to the different levels must also be in this order, i.e. a larger number = more democratic.

In contrast, different regions (Africa, Asia, Europe, Latin America, Northern America and Oceania) do not have such an ordering. The numbers used for the region variable are just labels for different regions. A different numbering (such as 6 = Africa, 5 = Asia, 1 = Europe, 3 = Latin America, 2 = Northern America and 4 = Oceania) would be just as acceptable as the one we originally used.