X4L SDiT

Survey Data in Teaching

enhancing critical thinking and data numeracy

GLOSSARY

July 2004

UK Data Archive, University of Essex

x4l@essex.ac.uk
x4l.data-archive.ac.uk

Version 1.0

Investigating Crime
Glossary

Adjustments
'Adjusting the figures' sounds like 'cooking the books' or further evidence that 'there are lies, damned lies and statistics[1].' However it can be a legitimate operation if it allows the statistician to measure the impact of an external factor that can influence the measurement. In this case, the Home Office statisticians know what the impact of the change to the operational definition of crime will be on how much crime is observed.. Inter-year comparisons demand that measured crime reflect the same underlying phenomenon.

Axes
Graphs typically have two *axes*. By convention, the vertical axis describes the event that changes, while the horizontal axis describes what produces the change. In Figure 1.1, the event that changes is the 'Total recorded offences in millions'. The graph shows whether this changes from one year to the next, thus 'Year' is described on the horizontal axis. The 'blocks' on the graph represent the joint-occurrence of the year and the 'Total recorded offences in millions' in that year. So in 1981 there were about 3,000,000 recorded offences while in 1995 the police recorded roughly 5,000,000 offences.

[1] A saying often attributed to Benjamin Disraeli, even though there is no evidence that he really said it.

Case
An instance or occurrence of something. In a survey it usually refers to a respondent.

Census
The word 'census', which is now used to describe the process of observing everyone who is eligible to be observed, comes from the Latin *censere* – to assess. Roman authorities enumerated everyone who was eligible for property tax. In the United Kingdom, the population census is conducted once every 10 years and counts everyone who was resident at a UK address on a particular date.

Central Limit Theorem
The Central Limit Theorem, which underpins our use of samples to estimate population characteristics (parameters) states:

> The means taken from random samples drawn from large populations will be normally distributed around the original population mean with standard deviation inversely proportional to the *square root* of sample size and directly proportional to the population standard deviation. See also sample size

Correlation
Correlation describes the relationship between two different variables. If one increases when the other increases then there is a correlation between them. For example there is a correlation between the speed at which a car travels and the

risk of death in an accident. The fact that there is a correlation does not necessarily mean that one causes the other. They could both be due to some third common factor.

(from www.sciencemuseum.org.uk/on-line/genetics/glossary.asp)

Ecological fallacy
An association that applies to a group does not necessarily apply to the individuals in the survey. This is known as the ecological fallacy:

> 'The mistake of assuming that where relationships are found among aggregate data, these relationships will also be found among individuals or households.'
>
> (from hds.essex.ac.uk/g2gp/gis/sect101.asp)

For more on the ecological fallacy see www.jratcliffe.net/research/ecolfallacy.htm

Estimates
Estimates derived from sample surveys are called 'statistics'.

Frequency table
A frequency table summarises data. It records how often groups of values of a variable occur.

Horizontal axis
The horizontal axis runs across the page, from left to right.

Line graph
Line graphs, like the one presented in Figure 1.1, can illustrate rates of change efficiently. However, although a "picture can be worth a 1000 words," graphs can also mislead the reader. The most common mistake is to be fooled by changing the width of the intervals on the graph's axes. For example, in Figure 1.1 the reader would get a much more dramatic portrayal of inter-year change if the *physical* distance between the categories on the vertical axis were doubled. However halving the space between the intervals would lessen the reader's perception of inter-year change.

Margin of Error
The amount by which we would expect an estimate to vary from the parameter.

Mean
The mean is the arithmetic average, which can be calculated for each sample to give an overall summary of how much of the characteristic was present in the sample.

Normal Distribution
The normal distribution is a bell-shaped curve which is symmetrical around the real population value (parameter).This statement implies that if you drew many random samples, their individual estimates of the population parameter would cluster around that value.

Official Statistics
Official statistics are statistics produced by government agencies to:

- shed light on economic and social conditions;
- develop, implement and monitor policies;
- inform decision making, debate and discussion both within government and the wider community;.

Government and its administrative arms need official statistics for policy development, implementation and evaluation. The public at large have similar information needs in order to evaluate government policy, to ensure public accountability, and to be adequately informed about social and economic conditions.

Operational Definition
'Crime' is an abstract concept that we cannot observe directly. Thus if we want to observe how much crime exists in a particular year, we have to specify observable acts that constitute crime. The operational definition of crime describes (the operation of) how crime will be measured. Clearly, there is scope for disagreement about what activities should be included in the operational definition, to say nothing of whether the observed acts should be weighted equally. For example, should parking on a double yellow line be given the same weight as a murder? Moreover, as the law changes, the collection of observable

acts that comprise the operational definition used to measure crime must also change. That of course will have an impact on how much "crime" is observed in a particular year.

All phenomena described by official statistics require operational definitions to make abstract phenomena concrete. Counting the number of people who are unemployed offers a good example. The standard operational definition of unemployment is to identify someone as unemployed if they receive unemployment benefits. However while that provides a consistent measure, the rule governing eligibility often change. Consequently, a person may go from being identified (and thus counted) as unemployed to being not counted as unemployed because of a change in the rules of eligibility – there have been over 30 changes to these rules since 1975.

Parameter
The population's true value is called a parameter.

Pearson's r
'Pearson's r is a correlation coefficient specially computed to show the covariance of metric variables. Pearson's r ranges between -1.0 and +1.0. Its absolute value reflects the covariance between the variables, while the sign indicates whether the correlation is positive or negative. Values approaching -1.0 indicate a perfect negative correlation. Correspondingly, values approaching +1.0 indicate a perfect positive correlation. Values approaching 0.0 indicate that there is no covariance between the two variables.'

(from NSDStat Help File)

Population
In this context, *population* means everyone in a group. For example, the population of England and Wales is everyone who normally resides in England and Wales. The target population is the group that the study wishes to find out about. The study population is the group that the study actually finds out about.

Random
Random selection means that everyone in the population that the sample is going to represent has a known probability of being selected.

Random Sample
Probability samples are often called random samples. A simple random sample has a strict technical meaning which implies both that each unit selected for the sample is chosen independently of all the other units chosen for the sample and that each member of the population that the sample represents has an equal chance of selection.

Rates
All things being equal, rates of change measure the amount that a phenomenon has changed from one year to the next. There are several ways of calculating rates of change. The easiest method is to (a) subtract the number of crimes recorded last year from the number of crimes committed this year and then (b) divide the result by the number of crimes committed last year and (c) multiply the result by 100 to calculate the percentage change.

Perhaps an example will clarify the calculations. Using the information in Table 1.1 for 1994 and 1995 (and rounding) , the calculations are:

(a) 5,100 – 5,252 = -152
(b) -152 ÷ 5,252 = -0.028
(c) -0.028 x 100 = -2.8%

In other words, in 1995, the incidence of crime as recorded by the police dropped by 2.8%.

Respondents
So-called because they *respond* to the questions in the questionnaire.

Sample
A sample is used to observe a representative selection of the eligible population. For example, doctors analyse a few drops of blood to assess the entire system. Of course, a doctor could drain all the blood from a patient but that would be costly. Fortunately, it is also unnecessary.

Similarly, social scientists who are interested in describing what is happening in a country, could observe everyone in the country by conducting a census. However the process is extremely costly both in terms of money and time. Fortunately, as in the case of the doctor's diagnosis, it is not necessary. The answers to a survey questionnaire given to a correctly-constructed sample of the population can provide an accurate estimate of the population's characteristics in much less time and much lower cost then a census would require.

Every year, the British Crime Survey asks a sample of 40,000 people, chosen to represent the population of England and Wales* who are 16 years or older, about their experience of several activities that are considered to be outside the law. Given that the population of England and Wales numbered more than 50 million in 2001, this a much more feasible exercise then asking everyone in the population of England and Wales about their experiences.

Nonetheless, the economies that sampling yields has a corresponding cost – sampling error. Every sample will provide a different estimate of the population's true value. Several steps can be taken to minimize the extent of this error.

1. The most important is to use a probability sample. This ensures that everyone in the population that the sample is meant to represent has a known probability of being selected.
2. Choosing a larger sample will lessen the error.

Sample size – effect of increasing
The Central Limit Theorem tells us that the improvement forthcoming from increasing the sample size is only the square root of the increase. For example, boosting a sample size from 1000 to 5000 increases the precision of the estimate from 37.6 ($\sqrt{1000}$) to 70.7 ($\sqrt{5000}$). So if, for example, there were just over 70% of a sample that thought crime had increased, increasing the sample size from 1000 to 5000 decreases the margin of error from about 2.8% to about 1.3%. So a fivefold increase in the number of observations would only double the precision of the estimate.

Standard deviation
The standard deviation measures how varied (or heterogeneous) the population is on the characteristic being measured – as such it is a measure of dispersion.

Time index
A time index compares every year's figure with a base year, which is made 100. The point of this is that two different trends can then be compared.

Time series
A time series is a sequence of data collected over a period of time.

Variable
Something that can alter from one case to the next.

Variation
Variation describes how homogeneous a population is. It is often measured by predicting how much of a characteristic everyone in the population would have if everyone was the same. This predicted value is then compared to the value that each individual in the population is observed to have. The difference (or "deviation") between the predicted value and the observed value is added across all individuals – the final sum measures how varied the members of the population are.

Vertical Axis
The vertical axis goes up and down the page.

* Scotland and Northern Ireland have their own Crime Surveys. The Scottish sample size is 5,000 people and the Northern Irish sample size is 3000.