THE FIRST STAGE OF DATA ANALYSIS:

IDENTIFICATION OF THE FORM OF DATA

ANTONY UGONI BSc (Hons) *

Abstract: In this paper, four different types of data are presented and their properties discussed. Also, presented is a brief summary of the statistical issues that need to be addressed when performing quantitative research.

Key Words: Data, study design

1. INTRODUCTION

Choice of study design and selection of appropriate analyses will usually make or break a piece of research. While the former is made from prior education and experience, the latter can be made easier if the structure of data is well understood. It is the purpose of this paper to discuss the concepts of data at length, and sound statistical protocol in general.

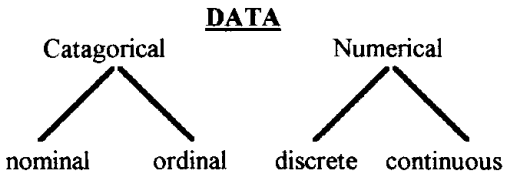
2. DATA

Of most importance to any quantitative study is the data itself. The modern scientific community rarely embraces hypotheses without relevant research. Without the correct data, the viability of the hypothesis can never be tested.

But what is data? Thought of simply as 'numbers' , data loses its lustre and value. It is in fact, (usually) numerical information that will enlighten the researcher with regard to a particular population at large.

Most data can be classified into 2 categories, categorical and numerical. These in turn can then be classified into 2 sub levels. Categorical data may be classified into "nominal" and "ordinal" data, numerical into "discrete" and "continuous" (1).

To illustrate the different types of data, consider the fictional data of people who attend Metropolis



* DEPARTMENT OF SOCIAL AND PREVENTATIVE MEDICINE
MONASH UNIVERSITY. COMMERCIAL ROAD. PRAHRAN.

Chiropractic and Osteopathic clinic after having brachial neuralgia attacks, presented in table 1.

Table 1.

Person	Sex	Age	Condition after 6 Weeks	AAGS* KGF	# of attacks
1	M	34	Good	44.4	3
2	M	56	Stable	35.7	2
3	F	64	Excellent	35.1	2
4	M	51	Good	34.8	2
5	F	76	Stable	23.3	1
6	M	66	Excellent	43.9	1
7	F	82	Poor	15.0	4
8	F	33	Stable	24.2	1
9	F	57	Excellent	44.7	2
10	M	60	Poor	15.6	3
11	F	71	Stable	23.2	1

*Affected Arm Grip Strength.

Let us first consider the data as a whole. The above table provides information concerning some characteristics of 11 people. These characteristics are known as variables (1).

Consider first, the data 'Sex' and 'Condition after 6 weeks', both are examples of categorical data. These variables use only descriptive words or letters, and the words/letters are known as levels. The levels of the variable 'Sex' are 'Male' and 'Female'.

Both 'Sex' and 'Condition after 6 weeks' are categorical, but a difference exists between the nature of these two variables. 'Condition after 6 weeks' has an ordering or ranking to it. Obviously 'Excellent' is more favourable than 'Good' which is more favourable than 'Stable', which is more favourable than 'Poor'. In contrast, the two levels of 'Sex', 'Male and 'Female', have no particular ranking to them. Due to this difference, 'Condition after 6 weeks' is known as an 'Ordinal' categorical variable, and 'Sex' is known as a 'Nominal' categorical variable.

Other examples of Ordinal variables are:

stage of cancer (I,II,III,IV)
severity of pain (none, mild, bearable, severe)
degree of satisfaction (none, little, total)

Other examples of Nominal variables are:

nationality (Greek, Chinese, Australian)
blood type (A, B, AB, O)
religion (Catholic, Moslem, Hindu)

Now consider the variables 'Age', '# of attacks' and 'AAGS'. In the first two cases, the data are represented as whole numbers, for example, person 7 is 82 years old and suffered 4 heart attacks. As with the two categorical variables however a subtle difference exists between these two continuous variables.

The variable 'Age' is (usually) rounded down to age at the person's last birthday. For example, a person born on January 1, 1970, would record their age as being 20 on June 30, 1990. Strictly speaking, this is incorrect since the person ignored the extra 6 months they had lived after their 20th birthday. The correct age would be 20.5 years. Even a breakdown to the number of months would be incorrect. June 30 is the 181st day of a non leap year. A more accurate estimate of this person's age would be $20\frac{181}{365}$ years ($\approx$20.49589 years).

If hours, minutes, seconds, etc. were recorded at the time of data collation a more accurate estimate of age would be available to the researcher. Of course this is usually impossible to do, and in the case of a variable such as 'Age', borders on trivial. It does, however, illustrate the concept of a numerical 'Continuous' variable.

In general, the accuracy of a numerical continuous variable is limited only by the measuring instrument used. For example, AAGS with our instrument would only be measured to one decimal place.

This is not the case with the variable '# of attacks'. A person could have had 1, 2, 3, ... brachial neuralgia attacks. People cannot have 2.7 brachial neuralgia attacks. Due to this restriction, '# of attacks' is known as a numerical 'discrete' variable.

Other examples of continuous data are:

weight (kg)
height (cm)
lung capacity (litres)

Other examples of discrete data are:

number of white blood cells
number of vaccination boosters
number of bones broken

Recognising the types of data in a data set is a crucial element in selecting an appropriate statistical procedure to use.

3. OTHER ISSUES

The issues of 'statistical protocol' will now be briefly reviewed.

Planning/Design

This is perhaps the most crucial stage of any study, and fortunately there is usually more than one way in which to design a study of interest. The motivation behind any study is some hypothesis, and care should be taken to collect data that will directly address that hypothesis. No amount of high powered analyses will shed any light on a hypothesis when the data is not relevant. Care should also be made to select (via some random mechanism) a sample which will be representative of the population of interest. Other questions to be addressed at this stage are:

How should the data be sampled?
How large should the sample size be?
Should the study be retrospective or prospective?
What are the exclusion/inclusion criteria?
What analyses are to be used when the data is collected?
Will anything bias the results, and can these be controlled for?

All these questions (usually) have answers, however the solutions are the subjects of many advanced statistical courses, and won't be discussed here.

Data collection

When collecting data to be analysed, the preferred method is to use the standard form with boxes provided for each digit. An example is given below.

Patient Number	□□□□
Age	□□
Weight	□□□
Sex	□
Number of treatments	□□□

Completeness of data is of paramount importance. When even one missing piece of data is missing for (say) a patient, some analyses will not be able to use *any* of that patients' data.

Data analysis

Contrary to popular belief, this is not the most important part of a study. Without proper study design and without good data no amount of analyses will yield meaningful results. Of course it is still imperative to do the analysis correctly, and researchers should take care to do so when analysing data. New software makes analysing data as easy as a keystroke, but also makes the possibility of a mistake greater than before. The analysis should be performed at least twice. If identical results are not achieved, then the researcher should start over.

Another pitfall to be avoided when using such software, is using analyses without having a thorough understanding of the theory involved. Many statistical analyses make prior assumptions about the data, and when these assumptions are not met, the results will often be meaningless.

Presentation

Having analysed your data, it is easy to fall into the trap that with minimum information, you assume other people will immediately know what has been done. It is important to report the study design and analyses used in a clear and concise manner.

The two more common ways to present results are in tables and graphs. While both can be a very effective way in which to display results, overuse can detract from the effectiveness of the main results. Any secondary results should be briefly mentioned in the body of the results section of presentations, whether they are posters, abstracts or publications.

Other points to consider are:

When quoting confidence intervals, quote the percentage coverage.
Do not use (for example) $p<0.05$ if the exact p-value is available.
If only one of either the confidence interval or p-value are to be displayed, quote the confidence interval, as it contains more information.

Interpretation

Having collected and analysed the data, interpretations of results are required. In order to make this easier, it would aid the researcher to remember the underlying principle when testing hypotheses. That is, the hypothesis that 'nothing happens' or the null hypothesis (2) is usually the basis for the test. Important to note also, is that statistical significance does not imply clinical significance. With a large enough sample size, any hypothesis will be rejected, and the practical importance of any result should be assessed via estimates of magnitudes of key questions.

4. DISCUSSION

The four most common types of data and their properties, and the basis for statistical involvement in research was discussed. With regards to the data, there is no substitute for becoming familiar with data, than manipulating 'real life' data. The statistical involvement in research, however, comes with the appropriate education and experience. It is the author's recommendation that people relatively new to research and wanting to pursue this area, consult a statistician to provide the technical and theoretical support required before they commence their research.

Acknowledgments

The author wishes to thank Andrew Forbes for suggestions made to every draft of this paper.

References

1. Altman, D.G. *Practical Statistics for Medical Research.* Chapman and Hall, London, 1992.
2. Ugoni, A. *On the Subject of Hypothesis testing.* COMSIG Review, Volume 2, Number 2, July, 1993.