MAS113 Fundamentals of Statistics I
Lecture Notes

Lawrence Pettit
School of Mathematical Sciences
Queen Mary, University of London
January 3, 2007
1 Introduction

1.1 What is Statistics?

Statistics is concerned with the process of finding out about real phenomena by collecting and making sense of data. Its focus is on extracting meaningful patterns from the variation which is always present in the data. An important feature is the quantification of uncertainty so that we can make firm decisions and yet know how likely we are to be right.

1.1.1 Problems and Questions

Statistical methods are applied in an enormous diversity of problems in such fields as:

- Agriculture (which varieties grow best?)
- Genetics, Biology (selecting new varieties, species)
- Economics (how are the living standards changing?)
- Market Research (comparison of advertising campaigns)
- Education (what is the best way to teach small children reading?)
- Environmental Studies (do strong electric or magnetic fields induce higher cancer rates?)
• Meteorology (is global warming a reality?)
• Medicine (which drug is best?)
• Psychology (how are shyness and loneliness related?)
• Social Science (comparison of people’s reaction to different stimuli)

Questions which arise in an investigation should be posed in non-statistical terms to keep subject matter priorities first; ”translating” these questions into the language of statistics usually means to answer the following:
- What should I measure?
- How should I measure it?

1.2 Ideas of Statistical modelling

In this section we are going to discuss some of the ideas of Statistical Modelling. We start with a real life problem. We think about what to measure and how to measure it. We decide how to collect some data. This may be via a survey, an experiment or carrying out an observational study. We have to design the method of data collection. For example by thinking carefully about questionnaire wording or in what way we decide experimental units receive different treatments or deciding which
variables to measure. We should also think of an appropriate statistical model for our data. This will often be of the form

$$\text{Observed data} = f(x, \theta) + \text{error},$$

where $x$ are variables we have measured and $\theta$ are parameters of our model. Data often exhibit great variability. The relationship we are assuming here is NOT deterministic. that is why the error term is there. We usually make some assumptions about the error term but we should use our data to check if those assumptions seem justified. If not we should go back and revise our model.

Statistical model building is an iterative process. we entertain a tentative model but we are ready to revise it if necessary. Only when we are happy with our model should we stop. We can then use our model, sometimes to understand our current set of data, sometimes to help us predict what may happen in the future. We must be ready to translate what the model is telling us statistically to the client with the real life problem.

1.3 Populations and Samples

When we carry out a statistical investigation we want to find out about a population.
Definition 1  A population is the collection of items under discussion. It may be finite or infinite; it may be real or hypothetical.

Sometimes although we have a target population in mind the study population we can actually find out information about may be different.

We are interested in measuring one or more variables for the members of the population but to record observations for everyone would be costly. The government carries out such a census of the population every ten years but also carries out regular surveys based on samples of a few thousand.

Definition 2  A sample is a subset of a population.

The sample should be chosen to be representative of the population because we usually want to draw conclusions or inferences about the population based on the sample. Samples will vary and the question of whether our sample is compatible with hypotheses we may have about the population will be a large concern in this course.

We will not concern ourself much with the mechanics of how the sample is chosen, this is a topic for the course Samples, Surveys and Simulation which some of you may do next year.
But the following examples give you some idea of the sorts of problems:

1. A city engineer wants to estimate the average weekly water consumption for single-family dwellings in the city. The population is single-family dwellings in the city. The variable we want to measure is water consumption. To collect a sample if the dwellings have water meters it might be best to get lists of dwellings and annual usage directly from the water company. If not then the local authority should have lists of addresses which can be sampled from. Note we should collect data through the year as water consumption will be seasonal.

2. A political scientist wants to determine if a majority of voters favour an elected House of Lords. The population is voters in the UK. Electoral rolls provide a list of those eligible to vote. What we want to measure is their opinion on this issue using a neutral question. (It would be easy to bias the response by asking a leading question.) We could choose a sample using the electoral roll and then ask the question by post, on the telephone or face to face but all these methods have problems of non-
response and/or cost.

3. A medical scientist wants to estimate the average length of time until the recurrence of a certain disease.

The population is people who are suffering from this disease or have done in the past. What we want to measure are the dates of the last bout of disease and the new bout of disease. We could take a sample of patients suffering the disease now and follow them until they have another bout. This may be too slow if the disease doesn’t recur often. Alternatively we could use medical records of people who suffered the disease in one or more hospitals but records can be wrong and there may be biases introduced.

4. An electrical engineer wants to determine if the average length of life of transistors of a certain type is greater than 5000 hours.

The population is transistors of this type. We want to record the length of time to failure by putting a sample of transistors on test and recording when they fail. Note that for such experiments where the items under test are very reliable it may be necessary to use an ”accelerated” test where we subject the items to higher currents than usual.
In other parts of the course we may not emphasize the underlying population or exactly how we collect a sample but remember these questions have had to be considered.