Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 4 Spring 2012

What is Statistics about?

We collect data, then analyse the data, and then interpret the results, to find out about real-world phenomena. There is aways variability in the data: we need to extract meaningful patterns. Because of the variability, our conclusions cannot be certain. We need to quantify the uncertainty, so that we can make definite decisions and judge how likely it is that we are right.

Here are some areas of application, with some typical problems.

Agriculture	Which varieties of wheat make the best bread?
Manufacturing	Can we make a cheaper detergent that is just as effective as the current one?
Health	Does an aspirin a day protect against stroke? If so, are there any side-effects?
Education	What is the best way to teach young children mental arithmetic?
Biology	How does biodiversity affect the enviroment?
Social science	Do people live in better houses than they did 20 years ago?
Economics	How is the credit crunch affecting food prices?
Market research	What sort of advertising campaign is most effective?
Environmental studies	Are people who live near mobile-phone masts more likely to get cancer?
Meteorology	Is global warming a reality?
Psychology	Are shyness and loneliness related?

Before starting any of these investigations, we need to stop and ask:

What do we want to investigate?
What should we measure?
How should we measure it?

Three methods of collecting data

1) Take a *sample* from a *population*.

   First we have to define the population. (It is just a set, and there is no need for its elements to be people.)

   How do we choose the sample? At random, or for convenience, or to make it representative?

   Do we ask questions (in which case, how do we word the questionnaire?), or take "objective" measurements, such as blood pressure?

   This is called a *survey*. If the sample is the whole population, it is called a *census*.

2) Design an *experiment*.

   This means that we apply different treatments to different experimental units, and then measure something to see if there is a difference between the treatments.

   How do we choose the treatments?

   How do we choose the experimental units?

   How do we decide who or what is given which treatment?

   (See *MTH6116 Design of Experiments*.)

3) If it is impractical or unethical to impose our choice of treatments, we may do an *observational study*. We might compare the effect of things that people can change themselves (for example, diet, or whether they go to the gym), or things that they cannot change (such as height, or place of birth).

Populations, Samples, Random Variables and Models

There are three ways of thinking about the source of the data.

1) Sampling with replacement from a (large? but finite) data set

For example, the population could be the set of all women living in the UK. Each has a definite number of children. Choose one woman at random; this means that all women are equally likely to be chosen. Define the random variable X_1 to be her number of children.

Return her to the population, and repeat the process. The second choice gives a random variable X_2. And so on, until we have random variables $X_1, \ldots, X_n$. Then

(i) $X_1, X_2, \ldots, X_n$ all have the same distribution;

(ii) $X_1, X_2, \ldots, X_n$ are independent in the sense that

$$\mathbb{P}(X_n = x_n \mid X_1 = x_1,\ \ldots,\ X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n).$$

Of course, in practice we don't waste resources by samplng the same person or thing twice!

2) Simple random sampling of n objects from a population of size N

This is sampling without replacement. All subsets of size n are equally likely to be chosen. Now $X_1, \ldots, X_n$ still have the same distribution but they are no longer independent. However, if n is very much smaller than N then they are almost independent, and we pretend that they are.

3) Sampling from a set of independent random variables

Imagine a set of individuals, and a *random variable* (not just a *number*) associated with each one.

For example,

$$\begin{aligned} \text{probability space} &= \{\text{all years 1975–2025}\} \\ X_i(2003) &= \text{number of days that person } i \text{ is ill during 2003} \\ P(X_i = 10) &= \text{proportion of years in which person } i \text{ is ill for exactly 10 days.} \end{aligned}$$

If we choose people with the same background, their random variables may reasonably be assumed to be identically distributed. If we choose people who are unrelated and live far apart, we may assume that their random variables are independent. So, choosing any n unrelated but similar individuals, and recording the number of days ill in one given year, gives n independent random variables with the same distribution, even if we do not choose the individuals randomly.

Particularly in this last approach, we usually assume something about the distribution of the random variables. In the last example, it would be reasonable to assume that

$$X_i \sim \text{Poisson}(\lambda).$$

This assumption is called a *model*, and λ is called a *parameter*.

(i) We might suppose that we think that $\lambda = 2$. Then we collect some data. Do our data look like a random sample from a Poisson(2) distribution? The statistical way of answering this question is called a *goodness-of-fit* test.

(ii) We might suppose that $X_i \sim \text{Poisson}(\lambda)$ but that we do not know what λ is. We collect some data. We use the data to *estimate* the true value of λ. We shall develop statistical methods for deciding when a method of estimation is good, and for quantifying the uncertainty in the estimate.

(iii) We might have two groups of people, with an inherent difference (such as male/female) or a chosen difference (such as taking Vitamin C pills or not).

We assume that

$$\begin{aligned} X_i &\sim \text{Poisson}(\lambda_1) \qquad \text{if person } i \text{ is in group 1} \\ X_i &\sim \text{Poisson}(\lambda_2) \qquad \text{if person } i \text{ is in group 2.} \end{aligned}$$

We collect some data from both groups. We develop a statistical method of answering the question "Is $\lambda_1 = \lambda_2$?": this is called a *hypothesis test*. If we decide that λ_1 and λ_2 are different, we probably go on to *estimate* the difference $\lambda_2 - \lambda_1$.

Some practical examples

1. The BBC wants to know how many people watch each of its programmes.

$$\begin{aligned} \text{Population} &= \text{all people in UK} \\ \text{Sample} &= \textit{panel} \text{ of people, chosen to be representative.} \end{aligned}$$

Each member of the panel keeps a diary recording all their TV viewing for a week, then sends it to the BBC. The BBC uses this data to estimate the total number of people who watched each programme. This is a survey.

2. Health researchers want to know which lifestyle factors affect the chances of getting various diseases. The UK BioBank has recently recruited 500,000 volunteers. The UK Biobank people collect some information now (for example, "Do you drink full-cream milk?"), then follow the person's medical records until they die, recording which diseases they get. They will then be able to test hypotheses such as "If you cycle daily, you are less likely to have a stroke". This is an observational study.

3. A marine engineer wants to know if a new sort of paint protects pier supports from corrosion. He paints ten metal beams, and leaves a further ten beams unpainted (why?). He puts all the beams in a tank of sea water for three months, then he measures the amount of corrosion in each beam. This is an experiment.