Analysis of Biological Data - Lecture Notes

Introduction to Statistics

Statistics is the study of methods to describe and measure aspects of nature from samples. Crucially, statistics gives us tools to quantify the uncertainty of these measures.

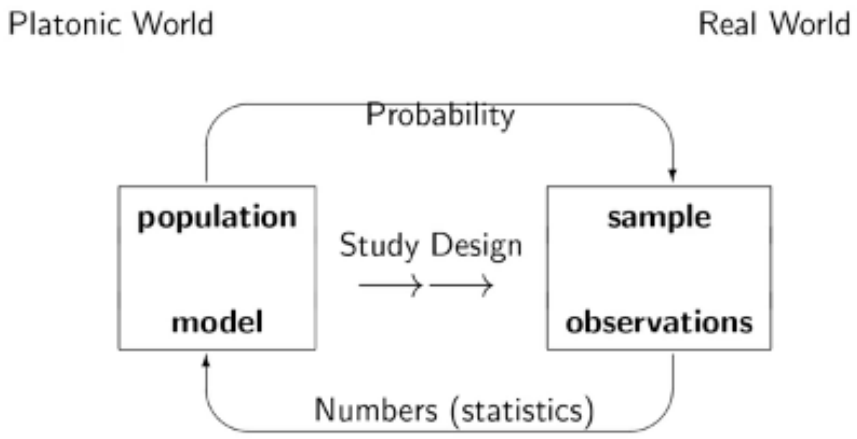
Statistics is about **estimation**, the process of inferring an unknown quantity of a target population using sample data. Properly applied, the tools for estimation allow us to approximate almost everything about populations using only samples.

Examples range from the average flying speed of bumblebees to the risks of exposure to cell phones. Most importantly, we can assess differences between groups and relationships between variables. For example, we can estimate the effects of different drugs on the possibility of recovery.

All of these quantities describing populations - namely, averages, proportions and measures of variation, and measures of relationship - are called **parameters**. Statistical methods tell us how to best estimate these parameters using our measurements of a sample. Hence, a parameter is a quantity describing a population, whereas an estimate or statistic is a related quantity calculated from a sample.

Statistics is also about hypothesis testing. A **statistical hypothesis** is a specific claim regarding a population parameter. Hypothesis testing uses data to evaluate evidence for or against statistical hypotheses. An example is 'The mean effect of this new drug is not different from that of its predecessor'.

Statistics starts with a platonic ideal and imagines that this ideal is observed or measured with some error or uncertainty. As a consequence, all we can ever observe in the real world is some estimate of some platonic ideal.



Case Study: Fisher's Sex Ratio Theory

Fisher wondered why half the population are male; Fisher pointed out that it is much more efficient for a population to have fewer males because few males can fertilise a very large number of females. However, if there are few males in a population, then it becomes in the female's best interest to produce sons (these sons will tend to have greater fitness because there are so few of them).

Example 1
Platonic World:
Let X = 'sex of randomly caught snake' so that the random outcome X could be x = male or x = female. For the probability model, assume:

$$p = Pr(X=male),\ 0 \leq p \leq 1$$
$$1 - p = Pr(X=female)$$
$$p=0.5$$

Real World:
54 out of 89 sampled snakes are males. Thus an estimate of the unknown quantity p is 0.61. How close is this number to the real value of p? Can we conclude p does not equal 0.5?

To summarise, statistics is the discipline (mostly science, sometimes art) concerned with:

1. Designing experiments and other data collection procedures.
2. Summarising information to aid understanding.
3. Drawing conclusions from data.
4. Estimating the present or predicting the future.

The discipline of statistics is critical for data analysis because it allows us to link the real world to simplified and interpretable descriptions (probability models). Statistics develops methods which tell us about the most plausible model. Such most plausible models can be then used to make scientific statements, test theories and make predictions. Statistics also helps us to describe observations succinctly, both graphically and numerically, and to quantify the variation in our data.

Sampling Populations
A **population** is the entire collection of individual units that a researcher is interested in. Ordinarily, a population is composed of a large number of individuals - so many that it is not possible to measure them all. For example, a population is all children in Melbourne suffering from asthma.

A **sample** is a much smaller set of individuals selected from the population. The research uses this sample to draw conclusions that, hopefully, apply to the whole population. For example, a selection of 50 children in Melbourne suffering from asthma.

Sometimes, the basic unit of sampling is a group of individuals (e.g. a gene), in which case a sample consists of a set of such groups. Scientists use several terms to indicate the sampling unit, such as unit, individual, subject or replicate.

Estimates based on samples are to defer from the true population characteristics simply by chance. This chance difference from the truth is called **sampling error**. The spread of estimates resulting from sampling error indicates the **precision** of an estimate. The lower the sampling error, the higher the precision. Larger samples are less affected by chance and so, all else being equal, larger samples will have lower sampling error and higher precision than smaller samples.

Ideally, our estimate is accurate or **unbiased**, meaning that the average of estimates that we obtain is centred on the true population value. If a sample is not properly taken, measurements made on it might systematically underestimate (or overestimate) the population parameter. This is a second kind of error called **bias**. Bias is a systematic discrepancy between the estimates we would obtain, if we could sample a population again and again, and the true population characteristic. Hence, the major goal of sampling is to minimise sampling error and bias in estimates.

A **random sample** is a sample from a population that fulfils two criteria. First, every unit in the population must have an equal chance of being included in the sample. Second, the selection of units must be independent. That is, the selection of any one member of the population must in no way influence the selection of any other member. Hence, random sampling minimises bias and makes it possible to measure the amount of sampling error.

A random sample can be obtained by using the following procedure:

1. Create a list of every unit in the population of interest, and give each unit a number between one and the total population size.
2. Decide on the number of units to be sampled (call this number n).
3. Using a random-number generator, generate n random integers between one and the total number of units in the population.
4. Sample the units whose numbers match those produced by the random-number generator.

One undesirable alternative to the random sample is the **sample of convenience**, a sample based on individuals that are easily available to the research. The researchers must assume that a sample of convenience is unbiased and independent like a random sample. The main problem is that it is bias.

Human studies must deal with **volunteer bias**, which is a bias resulting from a systematic difference between the pool of volunteers and the population to which they belong. The problem arises when the behaviour of the subjects affects their chance of being sampled. Compared with the rest of the population, volunteers might be more health conscious/proactive, low-income, more ill, and so on.

Types of Data

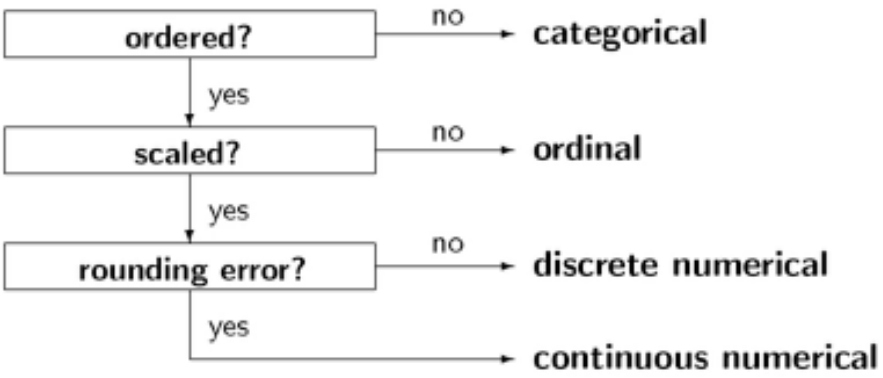
A **variable** is any characteristic or measurement that differs from individual to individual. Examples include running speed, reproductive rate and genotype. Estimates (e.g. average running speed of a random sample of 10 lizards) are also variables, because they differ by chance from sample to sample. **Data** are the measurements of one or more variables made on a sample of individuals.

Categorical variables describe membership in a category or group. They describe qualitative characteristics of individuals that do not correspond to a degree of difference on a numerical scale. For example, survival (alive or dead) or sex chromosome genotype. A categorical variable is nominal if the different categories have no inherent order (i.e., named). In other words, there is no obvious order to the categories (name only). In contrast, the values of an ordinal categorical variable can be ordered. Categorical data is always discrete.

A **numerical variable** is when measurements of individuals are quantitative and have magnitude. These variables are numbers. Measurements that are counts, dimensions, angles, rates and percentages are numerical. Examples include core body temperature or age at death.

Numerical data are either continuous or discrete. **Continuous data** can take on any real number value within some range. Between any two values of a continuous variable, an infinite number of other values are possible. In contrast, **discrete data** come in indivisible units. Number of amino acids in a protein is an example of one.

The taxonomy data types using a dichotomous key is shown below.



Different types of variables require different methods of treatment, and variable types have different inherent information content:

Categorical Variable = Category
Ordinal Variable = Category + Order
Numerical Variable = Category + Order + Scale

Thus, a numerical variable can be treated as an ordinal variable (ignoring the scaling) or as a categorical variable (ignoring the ordering and the scaling).

The main difference between data and variables is that variables are the attributes we measure. These are interpreted as potential data (before the actual data collection) since they may have different multiple values with some degree of uncertainty. Data are the observations that accumulate for each attribute. For example, 'toe length' is a variable; 12.5, 23.6,… are the data that might accumulate for that variable.

Often when association between two variables is investigated, a goal is to assess how well one of the variables, deemed the **explanatory variable**, predicts or affects the other variables, called the **response variable**. When conducting an experiment, the explanatory variable is the treatment variable, and the measured effect of the treatment is the response variable.

Different individuals in a sample will have different measurements. The **frequency** of a specific measurement in a sample is the number of observations having a particular value of the measurement. The **frequency distribution** describes the number of times each value of a variable occurs in a sample.

The distribution of a variable in the whole population is called its **probability distribution**. That is, it describes the number of times each value occurs in a population. The real probability distribution of a population in nature is almost never known. Researchers typically use theoretical probability distributions to approximate the real probability distribution. Typically, continuous variables are approximated by a theoretical probability distribution know as the **normal distribution**.

Types of Studies

Data in biology are obtained from either an experimental study or an observational study. In an **experimental study**, the researcher assigns different treatment groups or values of an explanatory variable randomly to the individual units of study. In an **observational study**, on the other hand, the research has no control over which units falls into which groups. Hence, a study is experimental if the researcher assigns treatments randomly to individuals, whereas a study is observational if the assignment of treatments is not made by the researcher.