Lecture 1 – An introduction to statistics in Ichthyology and Fisheries Science

What is statistics and why do we need it?

Statistics attempts to make inferences about unknown values that are common to a population based on information gleaned from a subset of that population - a sample. These inferences are phrased in two ways; either as estimates of the respective values such as central tendency, spread and confidence intervals, or as tests of hypotheses about their real values.

Is statistics necessary?

When is statistics used?

Question	Answer
The heights of Rhodes Zoology 3 students are measured and assumed to be representative of the Rhodes student community. These measurements are compared against a sample from the Zoology 3 class that is assumed to be representative of the UCT student body. Are Rhodes students taller than UCT students?	Yes, statistics is necessary. Only samples of the student bodies are taken and not the entire population.
A researcher wants to estimate fish abundance in a saltmarsh and see if abundance changes over time. The saltmarsh is block netted at spring-high tide and all the fish collected when it drains at low-tide. This experiment is conducted for each austral season.	No, statistics is not necessary. The whole saltmarsh population was sampled on each sampling occasion.
A researcher is sampling a river and measuring fish to see if fish length changes over time.	Yes, statistics is needed. The researcher will never know what the average length of the fish population is and has to use the samples to draw inference about it.
A researcher is assessing three different diets to see which one provided optimum growth in a particular species of fish.	Yes, statistics is needed. The researcher will be using samples of fish and cannot realistically test the entire fish population.
The weights of male and female Ichthyology 3 students are measured to see if males are heavier than females.	No, statistics is not necessary. The whole class. If the averages differed by even 0.0001g there was a difference!

Data and data generation

Data (or singular *datum*) that we observe come from somewhere – be they heights, weights, sizes etc.

There are four main categories of data.

Categorical: Theses data are simply non-numerical categories without a specific order. Examples include; sexes (male and female) or colours (red, green or blue).

Ordinal: These data are a special category of data that are not numeric but have some underlying ordering. Examples are good, average, bad or high, middle, low.

Continuous: These are numerical data that can take a decimal – in other words decimal data. Examples are; height, weight, distance etc. The decimal fraction can simply be related to the accuracy of the measuring instrument therefore if a measuring board in centimeters only then if a fish was 151 mm then a measurement would say 15 cm and not necessarily 15.1 cm.

Discrete: These are integer or count data. Examples are; number of offspring, incidents of fish aggression to a conspecific, the number of spines in a fish's dorsal fin, or the number of growth zones counted from a otolith.

These data come from somewhere – from some form of data generating function, a black box that takes some known values such as a population average and spread, and then generates observations randomly. If all the population values are known then these values will be known, but as we usually take a subset of the values in the form of a sample then we can make educated guesses only.

This function, the data generating function, can be approximated by a mathematical function. In short, a mathematical model of the data generating process.

Properties of data generating functions:

- All data generated from them have a common central tendency (or average) and a common spread (or variance)
- All of the probabilities of different outcomes must add to one.

A data generating function is also known as a *probability distribution function* – or *pdf* for short.

Identifying the data generating function and then trying to estimate its parameters is what we are trying to achieve in statistics. It takes time and practice but it can be done.

The most common data generating function, and the only one examined in this course, is the Normal distribution. The Normal distribution is arguably the most important function. Not only do most features of the natural world tend to follow a normal distribution, but there is good theoretical reason to expect that many things would (see The Central Limit Theorem later). As a result, most traditional statistical tests were designed to work with data that follows a normal distribution. It is the infamous "bell-shaped curve."

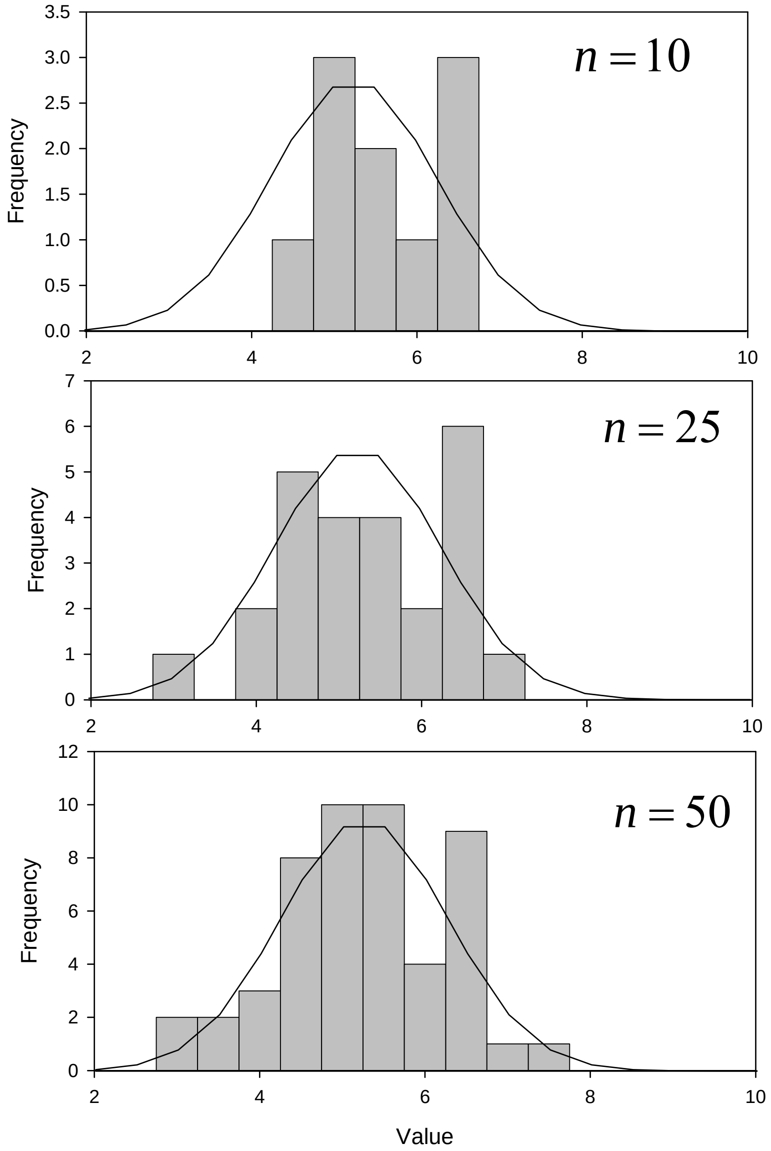


Figure 1.1: Three examples of samples of different size from a common Normal distribution. The data have all be generated randomly with a common average of 5 and a common spread of 1. If an extremely large number of observations were collected in a sample, say for instance 100000, then the sample would look like the line superimposed on top of each histogram.

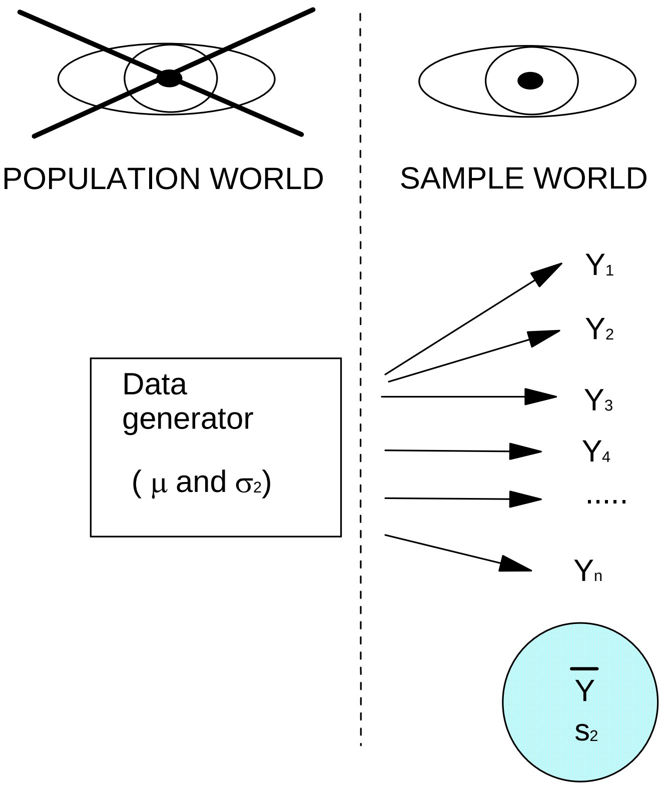
Populations and Samples

As mentioned earlier, when we wish to infer something about a phenomenon in nature, we normally cannot measure every instance of that phenomenon, but instead are limited to a few cases. As a result, we are left trying to infer about the whole from a relatively small subset – the sample.

Populations vs samples

In statistical terms, we refer to the whole group of individuals about which inferences are to be made as the population. Starting any investigation requires us to know if we are sampling the population or collecting all data about the population. We assume that we only have a sample.

DO NOT CONFUSE THE TWO



Parameters vs estimates

Similarly, for every value which we want to estimate has a true value in the population, which we call a *parameter*. Our best guess of the value that parameter actually takes which we get from our sample is called a *statistic*.

Therefore:

Populations ↔ **P**arameters UNOBSERVED VALUES COMMON TO EACH DATUM

Samples ↔ **S**tatistics OBSERVED VALUES COMMON TO EACH DATUM

Following common practice - Greek letters (because it is Greek to us) are used to represent parameters, and roman letters to represent sample statistics.

	Population parameters	Sample statistics
Mean	μ	$\bar{y}$
Variance	σ^2	s^2
Standard deviation	σ	s

The problem of sampling

If we take a sample from a population, unless we measure every member of that population, the sample will not exactly have the same properties as the population. This deviation of the sample statistic from the parameter is called sampling error.

A very basic concept in statistics is that a smaller sample will, on average, have more sampling error than a larger sample.

It is easy to see why this is even if the argument is taken to an extreme. A sample of only one individual is obviously different from the population, in the sense that every individual is different from the mean in some respects. Therefore, a very small sample will usually deviate substantially from the population. In contrast, if we take an extremely large sample, the sample will almost recreate the population, and the sample statistic is unlikely to vary much from the population parameter. This difference between a sample of size one and a huge sample holds true for less extreme cases as well.

Properties of a good sample

For example, sampling fish 10 000 times from a single dam to ask questions about the South Africa ichthyofauna fauna is like. Each dam has its own attributes, which cause the samples to be more similar than they would otherwise be, if they were randomly and independently chosen. Lastly, a sufficiently large number of fish should be sampled to be answer questions more precisely.

Independence

The members of the sample should be independent of one another. In other words, the inclusion of one individual in the sample should not change the probability of choosing another particular individual.

Random

The members of the sample should be a fair representation of *all* of the population, not just a biased subset.

Sufficiently Large

As we learned above, larger samples are more reliable, because the sampling error is likely to be much less. As a result, samples must be sufficiently large to be useful.