Chapter 3

Statistical Inference, Populations and Samples

Contents

3.1 Introduction

In this chapter we

- consider the problem of statistical inference: drawing conclusions about "populations", from what we observe in a random sample from the population;
- consider two 'types' of population: finite and infinite, and how we describe the population and define its characteristics in each case;
- show that by modelling a sampled observation as a random variable, we can (begin) to use probability theory from part I in this module to help us perform statistical inference.

3.2 What is statistical inference?

Statistical inference is the process of estimating characteristics of a **population** when, as is often the case, it is only possible to observe a subset or **sample** of members from the population.

3.2.1 Examples of statistical inference problems

Which country has the best education system for teaching schoolchildren?

Every few years, the Organisation for Economic Co-operation and Development (OECD) conducts a survey, known as the Programme of International Student Assessment (PISA), to compare school systems across different countries. In the 2015 survey, 72 countries were compared, and about half a million 15-year-old children took a test. But this only represents a sample out of about 28 million (15-year-old) children who could have taken the test. What conclusions can be drawn if (approximately) only 2% of children have been tested?

Which mobile phone brand/model is the most reliable?

You are considering buying a particular phone, and ask your friends who own the same model whether they've had any problems with theirs. One friend tells you her phone is great and has been trouble-free, another complains that the touchscreen on his phone doesn't work properly some of the time. Was he just unlucky, or is the problem common? It would be helpful to know the proportion of *all* phones of the brand/model that have reliability problems such as touchscreen issues.

A consumer magazine has conducted a survey of its readers, and compares 9 brands given responses from 2950 readers. This still only represents a very small proportion of the number of phones produced by each manufacturer: will their findings be reliable?

Does having a degree improve your career (and earning) prospects?

It may be of interest to know what proportions of graduates and non-graduates are currently employed (perhaps in 'highly skilled' jobs), and what the average earnings are in the two groups. No-one actually knows the correct values of any of these figures! Any figures you hear reported will only ever be *estimates*.

The UK Government reports graduate labour market statistics every year, based on the Labour Force Survey (LFS), a sample of about 40,000 UK households and 100,000 individuals

per quarter (which includes graduates and non-graduates). How can estimates of employment and earnings for *all* graduates and non-graduates be obtained from this sample?

Can playing 'brain-training' games make you more intelligent?

A company has developed a game which, it is claimed, will boost your intelligence if you play it regularly. How can we tell whether the claim is correct or not? As with any other claim (or theory/hypothesis) we need suitable data, either from a designed experiment or a suitable observational study.

Redick et al. (2013) conducted an experiment to test the effects of particular type of memory game ("dual n-back training") on other aspects of intelligence. In their study, a sample of 24 people were given training with the memory game, another 29 people received a different type of training, and 20 people had no training at all. How do we draw conclusions about the population of *everyone* who might play the memory game, based on this (apparently small) sample?

3.3 Populations

Given the investigation of interest, the next step is to define what it is we want to know: how we represent the population, and what characteristics of the population we want to estimate. There are two ways that we might represent and describe a population: **finite** and **infinite**.

3.4 Finite populations

We think of a finite population as a list of values. We have N members of the population, and we denote the ith member's value of interest by y_i. Two examples:

1. In the PISA testing of 15-year-old schoolchildren, suppose we are just interested in the UK, so that the population of interest is all 15-year-olds in the UK. Suppose there are currently 800,000 15-year-olds, so we have $N = 800,000$. We define y_i to be the score that the ith 15 year old would get, if he/she were to take the test, so that the full population of interest is $y_1, y_2, \ldots, y_{800000}$.

2. Suppose there are currently 12,000,000 graduates (below retirement age) in the UK, and we are interested in how many of them are currently employed. We can define

$$y_i = \begin{cases} 1 & \text{if graduate } i \text{ is currently employed} \\ 0 & \text{if graduate } i \text{ is currently unemployed} \end{cases} \tag{3.1}$$

   so that $y_1, \ldots, y_{12,000,000}$ describes the employment status of all graduates in the population.

3.4.1 The distribution of a finite population

By the **distribution** of a population, we mean the percentage of population members taking particular values (or values in particular ranges, if the population values are largely all distinct.) A (fictitious) example in the PISA testing would be

- 11% of the values $y_1, y_2, \ldots, y_{800000}$ are in the interval $[0, 200)$,
- 38% of the values $y_1, y_2, \ldots, y_{800000}$ are in the interval $[200, 400)$,
- 42% of the values $y_1, y_2, \ldots, y_{800000}$ are in the interval $[400, 600)$,
- 9% of the values $y_1, y_2, \ldots, y_{800000}$ are in the interval $[600, 800)$,

and similarly for any other set of intervals we might choose. The distribution will usually be *unknown*, because we won't know what all the values $y_1, y_2, \ldots, y_{800000}$ are.

3.4.2 The mean and variance of a finite population

Finite population mean

We define the finite population mean to be

$$\mu_{fin} := \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{3.2}$$

Note that in the example above of graduate employment and the definition of y_i in equation (3.1), μ_{fin} gives the population *proportion* of graduates who are employed. Hence finite population proportions can be expressed as finite population means.

Finite population variance

We define the finite population variance[1] to be

$$\sigma^2_{fin} := \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_{fin})^2 \tag{3.3}$$

(The symbols μ and σ^2 are often used to represent means and variances in different contexts, so we have included the subscript 'fin' to make it clear that we are using μ_{fin} and σ^2_{fin} to represent the mean and variance of a *finite* population. Later on, we will drop these subscripts.)

We will often be interested in estimating the mean of a population. In the PISA example, countries are ranked based on their mean test scores. The variance will often be of interest too. For example, it may be interesting to know if test scores vary more in some countries than others. Hence one example of a statistical inference problem is to estimate μ_{fin} and σ^2_{fin} based on knowing only a *subset* of the values $y_1, \ldots, y_N$.

3.4.3 Drawing a random sample from a finite population

The key idea that will enable us to perform statistical inference is to choose the subset *randomly*, to obtain what we call a **random sample** from the population. In particular, each randomly sampled value can be thought of as an observation of a *random variable*. This will allow us to use probability theory both to justify our choice of estimates, and understand how accurate they are likely to be.

Define a random variable X to be the outcome of picking one member of the population and observing its value. In particular:

[1] Some textbooks use the denominator $N-1$ in the definition of a finite population variance, for reasons that we don't need to worry about here. For large N, we have $\frac{1}{N} \simeq \frac{1}{N-1}$ in any case.

1. we suppose each population member has the same probability $1/N$ of being selected;

2. if member j of the population is selected, the observed value of X will be y_j.

In practice, it's not always possible to achieve (1): some members of the population may be difficult to reach. More complicated cases where the probabilities are unequal are considered in MAS370, but will not be covered here.

3.4.4 The expectation of a randomly sampled observation from a finite population

Theorem 3.1. *For a random variable X as defined in Section 3.4.3, we have*

$$\mathbb{E}(X) = \mu_{fin}. \tag{3.4}$$

But before we prove this result...

Confusion alert number 1!
Means, means and means...

A major source of confusion when studying probability and statistics is the word "mean": it is used in different contexts to mean different (but often related) things. Try to avoid using the word "mean" on its own, and make sure you are always clear precisely what is meant by the word "mean" in any situation.

1. **The arithmetic mean**
   If we have a list or sequence of n numbers, the arithmetic mean is their sum, divided by n: the arithmetic mean of 64, 14, 21, 32 is
   $$\frac{64+14+21+32}{4} = 32.75. \tag{3.5}$$

2. **Finite population mean**
   We have defined the finite population mean to be the arithmetic mean of all the population values:
   $$\mu_{fin} := \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3.6}$$

3. **Expectation or mean of a random variable**
   For any random variable X, its expectation $\mathbb{E}(X)$ is **also** known as its mean, but this is **not** the same thing as an arithmetic mean. For a discrete random variable X, its mean is defined as
   $$\mathbb{E}(X) := \sum_{x \in R_x} x p_X(x), \tag{3.7}$$
   and for a continuous random variable X, its mean is defined as
   $$\mathbb{E}(X) := \int_{-\infty}^{\infty} x f_X(x) dx \tag{3.8}$$
   Try to use the word "expectation" rather than "mean" in this context.

You can, however, **interpret** an expectation of a random variable in terms of an arithmetic mean: if a very large number of random variables $X_1, X_2, \ldots$ were to be observed, each with same expectation so that $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \ldots$, informally, the value of their arithmetic mean would approximately be equal to the value of this expectation.

Note that Theorem 3.1 is **not** saying that $\mathbb{E}(X)$ is *defined* to be μ_{fin}; its definition is given in (3.7). The theorem tells us that if we apply the definition (3.7), the result we will get is equal to μ_{fin}.

Now back to the proof:

Proof. Starting with the definition of expectation, we have

$$\mathbb{E}(X) = \sum_{x \in R_x} xP(X = x), \tag{3.9}$$

where R_x is the range of X, the set of possible values that X can take. Writing out the range is a little awkward, as some values in the population could be duplicated, but it is not difficult to see, for example, that if three members of the population all had the value 528, then $P(X = 528) = \frac{3}{N}$. The corresponding term in the summation above would be

$$528 \times \frac{3}{N} = 528 \times \frac{1}{N} + 528 \times \frac{1}{N} + 528 \times \frac{1}{N}, \tag{3.10}$$

and so we can write the expectation as

$$\begin{aligned} \mathbb{E}(X) &= \sum_{i=1}^{N} (\text{value of population member } i) \times P(\text{population member } i \text{ selected}) && (3.11)\\ &= \sum_{i=1}^{N} y_i \times \frac{1}{N} && (3.12)\\ &= \mu_{fin}, && (3.13) \end{aligned}$$

□

3.4.5 The variance of a randomly sampled observation from a finite population

We find the variance of X via $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Again, instead of working with

$$\mathbb{E}(X^2) = \sum_{x \in R_x} x^2 P(X = x), \tag{3.14}$$

we write

$$\begin{aligned} \mathbb{E}(X^2) &= \sum_{i=1}^{N} (\text{value of population member } i)^2 \times P(\text{population member } i \text{ selected})\\ &= \sum_{i=1}^{N} y_i^2 \times \frac{1}{N}. && (3.15) \end{aligned}$$

Using this result we can prove the following

Theorem 3.2.

$$Var(X) = \sigma_{fin}^2. \tag{3.16}$$

3.4.6 A random sample from a finite population

We now think of drawing a random sample of n observations as observing n random variables $X_1, \ldots, X_n$ as described above: X_j represents the jth instance of picking one member of the population at random. In practice, we wouldn't want to choose the same population member twice, but in this module, we will ignore this possibility, and suppose that the n members of the population are chosen independently of each other. (Picking the same member twice will be very unlikely if N is large and n is relatively small). We have $X_1, \ldots, X_n$ independent and identically distributed, with

$$E(X_j) = \mu_{fin}, \tag{3.17}$$
$$Var(X_j) = \sigma^2_{fin} \tag{3.18}$$

for $j = 1, \ldots, n$. Shortly, we will see how we can use these results both to estimate population means and variances from samples, and how to quantify uncertainty in these estimates.

3.5 Infinite populations

In many cases, the idea of a finite population does not satisfactorily describe the situation of interest. Typically, this will be the case when the population we are interested includes items that may 'come into existence' in the future, in addition to items that currently exist or existed in the past; we can't conceive of a finite population size N.

3.5.1 Limitations of the finite population model: an example

A casual runner runs a 5K race most weeks. He judges his general fitness to have been 'constant' over these races. He has recorded all his times, and has an average running time of 23 minutes. His fastest time was 21:49 minutes, and his slowest time was 24:22 minutes. A friend suggests that drinking an energy drink two hours before each race will improve his times (a little). For the next three races, he does this, and his running times are 22:36, 22:48, 23:06. Did the drink have an effect?

Given that his times vary anyway, we would not be convinced of anything after only three races. What we would like to know is: what will his *mean* running time be, over a 'population' of all the races that he *could* run, if he drinks the energy drink each time (assuming no other changes to his fitness). Is this mean value less than 23 minutes?

We might think of three new running times as a *sample* from this population. But how exactly do we represent this population of running times, and in what sense are those three observed times a sample from it? The idea of sampling from a finite population doesn't make sense here:

- the three observed times were not randomly *selected* from some 'list' of running times;
- he has only run three races using the energy drink so far; no other races have happened yet and so other population values have *not yet been determined*;
- we can't specify *now* a population size N; it has not been determined how many races he will run in the future.

3.5.2 Modelling infinite populations with probability distributions

In the infinite population case, we again think of our observed data as observations of random variables $X_1, X_2, \ldots$, but now we simply *suppose* that these are 'random draws' from a probability distribution. This probability distribution will *represent* the population.

An example: normally distributed running times

We might suppose that each running time we observe is a random draw from a normal distribution $N(\mu, \sigma^2)$, and we say that "the runner's times are normally distributed": the population distribution of running times is given by the $N(\mu, \sigma^2)$ distribution. We visualise this in Figure 3.1.

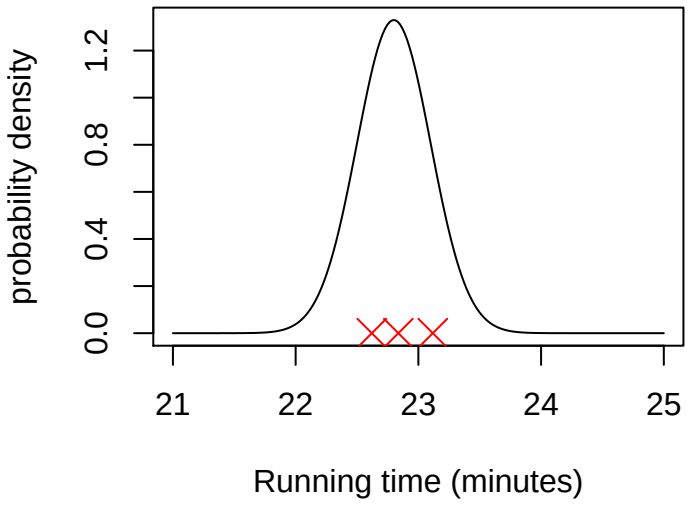

Figure 3.1: We suppose the population distribution of running times (using the energy drink) is described by some probability distibution (the solid line). The three running times we observed (the red crosses) are treated as random draws from this distribution.

More specifically, we suppose that the population distribution of running times can be represented by a density function f_X, where we have supposed that this density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

You have, of course, met the normal distribution and this density function before. But two things are little different in this context:

1. we are using a density function to describe a *population* of running times, not just one single 'random' running time;

2. the parameters μ and σ^2 will be *unknown*. In particular, we may wish to *estimate* them given the observed data of the three running times.

3.5.3 The mean and variance of an infinite population

The (infinite) population mean and variance are defined to be the mean and variance of the corresponding probability distribution. If this population has probability density function f_X, then the population mean and variance are defined as

$$\mu_{inf} := \int_{-\infty}^{\infty} x f_X(x) dx, \tag{3.19}$$

$$\sigma^2_{inf} := \int_{-\infty}^{\infty} (x - \mu_{inf})^2 f_X(x) dx. \tag{3.20}$$

(If the population values are discrete, the population mean and variance would be defined in terms of a probability mass function, with summation rather than integration).

Note that, by definition, for any 'random draw' X_i from the population, we have $\mathbb{E}(X_i) = \mu_{inf}$ and $Var(X_i) = \sigma^2_{inf}$.

3.5.4 Population proportions for infinite populations

For a finite population, the proportion of population members talking values between some limits a and b is obtained by counting members of the population. For an infinite population, we suppose it is defined by the integral

$$p_{[a,b]} := \int_a^b f_X(x) dx. \tag{3.21}$$

3.5.5 Interpreting the mean, variance and population proportions for an infinite population

The definitions we have given for the population mean, variance and proportion all match the definitions of $\mathbb{E}(X)$, $Var(X)$ and $P(a \leq X \leq b)$ for a single random variable X that you have already met in Part 1: Probability of this module So why, in this context, have we called the above integrals the *population* mean, variances and proportion? To understand this, think about what happens when we have a large number of identically distributed (and independent) random variables $X_1, X_2, \ldots$

Suppose we could obtain a very large number of observations from an infinite population. We think of this as observing a large number of independent and identically distributed random variables $X_1, \ldots, X_N$. Denote the values we actually get by $x_1, \ldots, x_N$. If we now treat these values $x_1, \ldots, x_N$ as a *finite* population we would find that:

- the finite population mean of $x_1, \ldots, x_N$, calculated using (3.2), would be approximately equal to μ_{inf}, defined in (3.19);
- the finite population variance of $x_1, \ldots, x_N$, calculated using (3.3) would be approximately equal to σ^2_{inf}, defined in (3.20);
- the proportion of members in the finite population taking values between the limits of a and b would be approximately $p_{[a,b]}$ as defined in (3.21);
- a histogram of $x_1, \ldots, x_N$, scaled to have total area 1, would look very similar to the infinite population density function f_X.

An illustration using a simulation experiment

Suppose the population of interest concerns the length of time (in days) each patient will stay in a particular type of hospital ward. Suppose this population is represented by an infinite population, described by the $Exp(rate = 0.2)$ distribution. We say that 'the length of time will be exponentially distributed'.

Note that for $X \sim Exp(rate = 0.2)$, we have $\mathbb{E}(X) = 5$, $Var(X) = 25$, so we have population mean $\mu_{inf} = 5$, and the population variance $\sigma^2_{inf} = 25$. We also have

$$p_{[5,10]} = \int_5^{10} 0.2 \exp(-0.2x)dx = 0.233.$$

How do we interpret these values? In R, we do the following

1. We generate N random draws from the $Exp(rate = 0.2)$ distribution, for some large value of N. Denote these generated values by $x_1, \ldots, x_N$.

2. We now treat $x_1, \ldots, x_N$ as a separate finite population, and calculate the finite population mean and variance using (3.2) and (3.3).

3. We compare the finite population mean and variance calculated in step 2 with $\mu_{inf} = 5$ and $\sigma^2_{inf} = 25$.

4. We compare a histogram of the finite population $x_1, \ldots, x_N$, scaled to have total area 1, with the $Exp(rate = 0.2)$ probability density function $f_X(x) = 0.2 \exp(-0.2x)$ (for $x > 0$).

5. We compare the proportion of $x_1, \ldots, x_N$ taking values in $[5, 10]$ with $p_{[5,10]} = 0.233$.

We conduct this experiment in R as follows.

```
par(mar=c(5, 4, 1, 2) + 0.1)
x <- rexp(1000000, rate = 0.2)
mean(x)

## [1] 4.995208

var(x)

## [1] 25.06156

sum(x>=5 & x<=10)/1000000

## [1] 0.231775

hist(x, prob = TRUE, xlim=c(0, 40), breaks=0:ceiling(max(x)),
     xlab="length of stay (days)", main = "")
curve(dexp(x, rate = 0.2),from = 0, to = 40 ,col  ="red", add = TRUE)
```

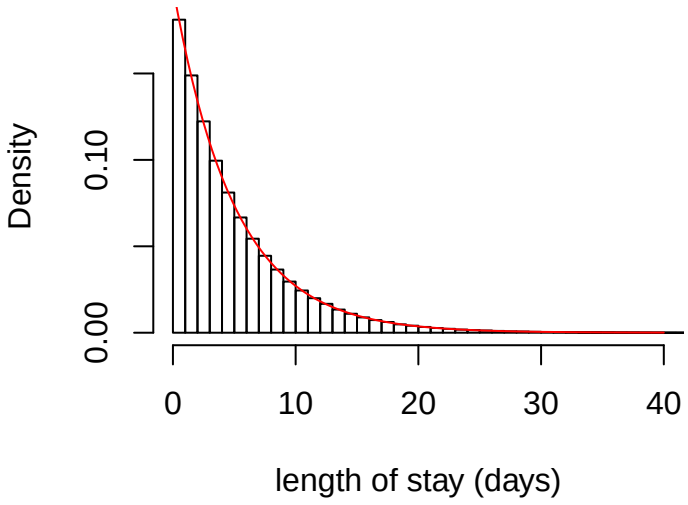

Figure 3.2: A histogram of a finite population with $N = 1000000$, generated from a probability distribution with density function given by the red curve.

Observe how the histogram lines up with the density function, note how the finite population mean and variance are very close to μ_{inf} and σ^2_{inf}, and how the observed proportion was close to 0.233. Hence, informally, μ_{inf}, σ^2_{inf} and $p_{[a,b]}$ tell us what we would *observe* in a very large sample.

Comment: why "infinite" population?

Informally, f_X, μ_{inf} and σ^2_{inf} can be thought of as the limiting case of the finite population distribution, mean and variance as the population size $N \to \infty$. Another way to think about it is that in an infinite population model, we can define a mean, variance and proportion without any reference to a population size N, *unlike* the finite population case.

3.6 Summary

We have two ways of thinking of a population: finite or infinite. In both cases, we will think of obtaining a sample from the population as observing a set of independent and identically distributed random variables $X_1, \ldots, X_n$.

- In the finite population case, we have

$$E(X_j) = \mu_{fin}, \tag{3.22}$$
$$Var(X_j) = \sigma^2_{fin}, \tag{3.23}$$

  where μ_{fin} and σ^2_{fin} are the finite population mean and variance respectively, defined in equations (3.2) and (3.3).

- In the infinite population case, we have

$$E(X_j) = \mu_{inf}, \tag{3.24}$$
$$Var(X_j) = \sigma^2_{inf} \tag{3.25}$$

where μ_{inf} and σ^2_{inf} are the infinite population mean and variance respectively, defined in equations (3.19) and (3.20).

In this regard, the distinction between finite and infinite population is unimportant, and from now on we will simply write

$$E(X_j) = \mu, \tag{3.26}$$
$$Var(X_j) = \sigma^2, \tag{3.27}$$

where μ is the population mean and σ^2 is the population variance.

Regarding the interpretation of μ and σ^2:

- in a finite population, μ will be the arithmetic mean of each population member's value (e.g., exam score, income, weight etc.);
- in an infinite population, μ will, approximately, be the arithmetic mean of a very large number of population members' values,

and we can interpret σ^2 in a similar way.