5 Samples and populations

In this section we consider the fundamental statistical concepts of **populations** and **samples**. We investigate the properties of samples, particularly as the sample size gets large, via computer simulation. This material is intended to reinforce the ideas studied in MAS2901.

Example 5.1: A factory has been depositing mercuric waste into a river near its estuary. It is feared that this waste may have caused the mercury content in the prawns there to have risen to a level which may be toxic for human consumption. The maximum permissible level is one part per million and the factory claims the average mercury content in the prawns does not exceed this level. This claim is to be tested.

It is obviously impractical to measure the mercury content of all the prawns. A random sample of 20 prawns is collected. The concentration of mercury (parts per million) in each of the 20 prawns is measured, giving the following results: the observed sample mean is $\bar{x} = 1.2$, and the observed sample standard deviation $= s = 0.5$.

Given these measurements do we believe the factory's claim?

5.1 Populations

Suppose we measure some random quantity X: X can adopt a range of possible values, and some values are more likely than others. This is, of course, the familiar idea of a **distribution** for X. Usually we do not know this distribution exactly, but we want to learn something about it. The unknown distribution is called the **population distribution**.

Referring to the prawn example:

- the population consists of the prawns in the estuary;
- the random quantity X is the mercury concentration in a randomly selected prawn; and
- the population distribution is the distribution of X.

We are usually interested in key properties of the population distribution such as:

- the expectation of X – usually called the **population mean**;
- the variance of X – usually called the **population variance**; or
- the 95th percentile of X (for example).

Often we make some simplifying assumptions about the population distribution. For example, we might assume:

(a) X is normally distributed with unknown mean and variance;

(b) X is exponentially distributed with rate parameter λ, where λ is unknown but lies on the interval $(0, 1)$;

(c) X is normally distributed with unknown mean and variance $\sigma^2 = 5$.

A set of assumptions like this is referred to as a **model**. Typically a model involves some unknown parameters. In the examples above we have the unknown rate λ in (b), and unknown mean μ in (c).

In some situations – usually rather artificial ones – we know the population distribution exactly. For example:

- let X be the score obtained from rolling a fair die; or
- let X be the number on a card drawn at random from a full deck. (Assume Jack, Queen, King numbered 11, 12, 13 respectively.)

5.2 Samples

Typically we do not know everything about the population distribution. We learn about the population distribution by drawing a **sample**. Referring to the prawn example: we cannot measure the mercury concentration in every prawn in the estuary, so we sample 20 prawns at random and analyse those.

A sample of size n corresponds to taking n independent measurements from the distribution. Each measurement is itself a random variable with the same distribution as X: the sample measurements denoted $X_1, X_2, \ldots, X_n$. So in our example, X_1 represents the mercury concentration in the first prawn collected – and this is a random quantity. The actual measurements obtained are denoted $x_1, x_2, \ldots, x_n$.

The distinction between the population distribution and how we learn about the population from limited samples is probably the most **important concept** in statistics.

5.3 Estimators

Suppose we wish to learn about some aspect of the population distribution: the population mean or the population variance for example. To do this we construct an **estimator** for the quantity of interest. For example, if we wish to learn about the population mean, a good estimator is the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(We won't address what a 'good' estimator might be in this module.) Formally, an estimator is defined to be some **function** of the sample, i.e. an estimator S is a random variable for the form

$$S = g(X_1, X_2, \ldots, X_n)$$

for some function g. It follows that every estimator is a **random variable**. When we observe some measurements $X_1 = x_1, \ldots, X_n = x_n$ then we can compute the corresponding value of the estimator:

$$s = g(x_1, x_2, \ldots, x_n).$$

This value is called an **estimate**. To summarize: an estimator S is a random quantity which is a function of the sample; whereas an estimate is an observed value of that random quantity.

Since any estimator S is a random variable it makes sense to talk about its **distribution**. The rest of this section is concerned with studying the distribution of some simple estimators.

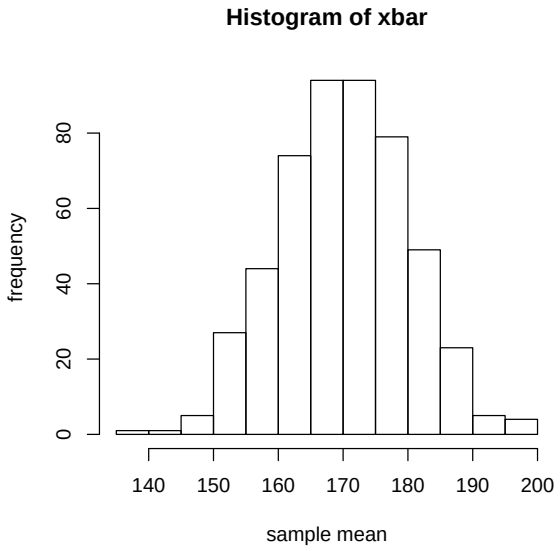
Example 5.2: Suppose the population distribution is normal, and we wish to estimate the population mean. Suppose the sample size is $n = 4$ and our estimator is $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$. What is the distribution of $\bar{X}$ when the population distribution is $N(170, 20^2)$?

Simulation can be used to answer this question approximately: we can simulate lots of samples and look at the resulting distribution.

```
simulate.xbar = function(n) {
    xbar = vector(mode = 'numeric', length = n)
    for (i in 1:n) {
        x = rnorm(4, 170, 20) # Generate a sample of size 4
        xbar[i] = 0.25*sum(x)
    }
    xbar
}
```

Histogram of $\bar{X}$:

```
test = simulate.xbar(500)
hist(test, xlab = 'sample mean', ylab = 'frequency')
```



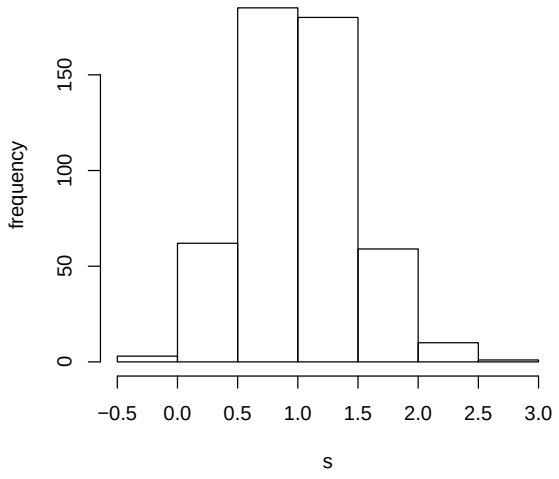
Example 5.3: Suppose the population distribution is normal, and we wish to estimate the 90th percentile using a sample of size 10. A sensible estimator is to define S to be the second largest value in the sample (i.e. the 9th value when the samples are ordered from smallest to largest). What is the distribution of S when the population distribution is $N(0, 1)$?

Code to simulate the distribution of S:

```
simulate.percentile = function(n) {
    s = vector(mode = 'numeric', length = n)
    for (i in 1:n) {
        x = rnorm(10, 0, 1) # Generate a sample of size 10
        x = sort(x)
        s[i] = x[9] # Get 9th value on sorted list
    }
    s
}
```

Histogram of S:

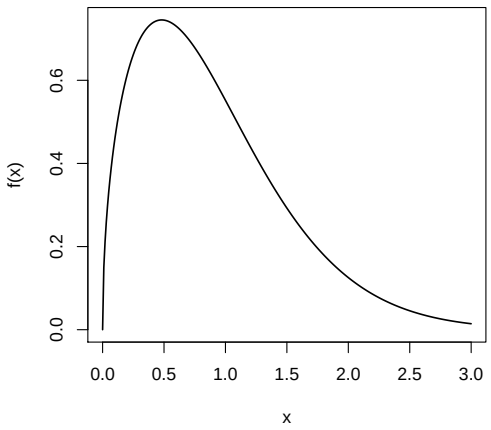
```
S = simulate.percentile(500)
hist(S, xlab = 's', ylab = 'frequency', main = '')
```



5.4 Simulation experiment

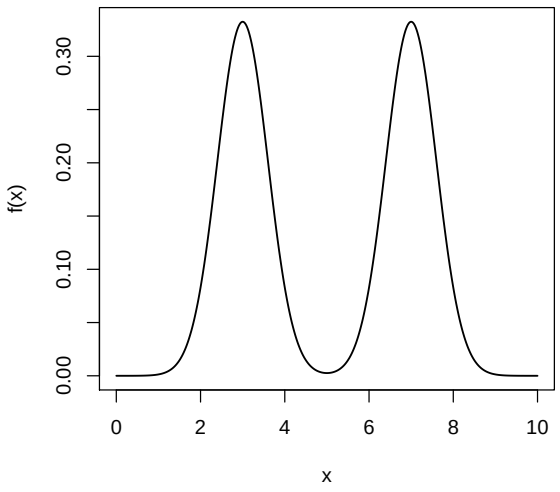
Consider the following two examples for the density of the population distribution. For each example, decide which histogram on the slides (A, B, C or D) is most likely to represent the distribution of the sample mean $\bar{X}$ when the sample size is 10.

Example 5.4: Suppose X (the population quantity) has the following PDF:



Most likely histogram:

Example 5.5: Suppose X (the population quantity) has the following PDF:



Most likely histogram:

Conclusion:

- The sample mean is distributed around the population mean.
- The distribution of sample mean values 'forgets' the underlying shape of the population distribution.
- As n increases we expect the distribution of $\bar{X}$ to become more clustered around the true value.

5.5 The Central Limit Theorem

The Central Limit Theorem is the most important theorem in statistics. It describes the distribution of the sample mean $\bar{X}$ as the sample size increases.

Suppose $X_1, X_2, \ldots, X_n$ are independent and **identically distributed** random variables with common mean μ and variance σ^2 which are both finite. Define

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

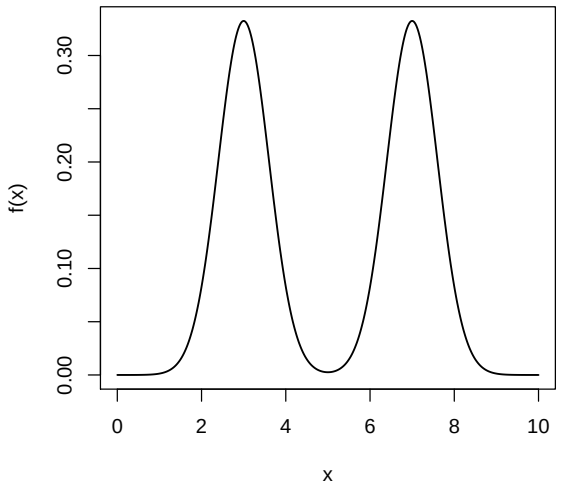
Then as $n \to \infty$ the distribution of Z tends to $N(0, 1)$.

While this results describes the approximate distribution of the sample mean as n gets large, it can also be adapted and extended to understand the asymptotic behaviour of other estimators.

Remarks:

1. Equivalently, the Central Limit Theorem shows that for large n, $\bar{X} \sim N(\mu, \sigma^2/n)$ approximately.
2. It follows that $\bar{X}$ is distributed somewhere around the population mean μ.
3. As n increases, the variance of the distribution of $\bar{X}$ decreases. In other words, $\bar{X}$ is more likely to lie closer to μ – "larger samples give better estimates".
4. The underlying population distribution is forgotten as n increases, as $\bar{X}$ has a normal distribution.

Example 5.6: Suppose X has the bimodal distribution considered in example 5.5:



This is a normal mixture distribution, and the following function can be used to sample values $\bar{X}$ from the distribution when the sample size is n.

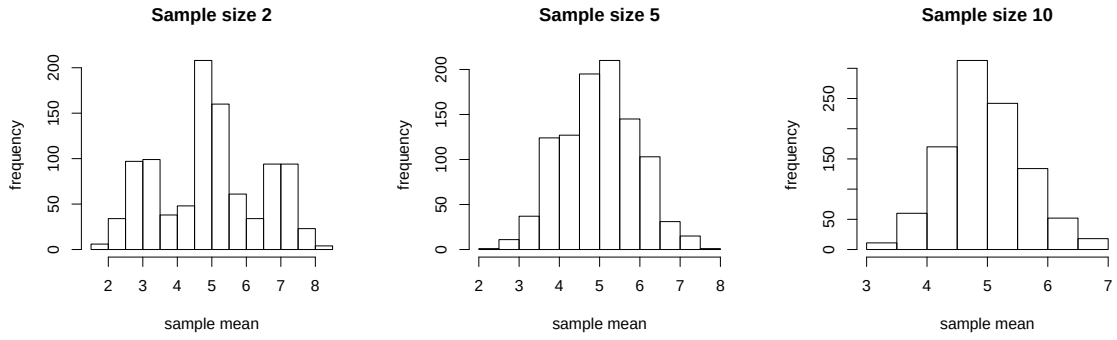
```
simulate.bimod = function(k, n) {
    # Generate k samples of size n
    s = vector(mode = 'numeric', length = k)
    for (i in 1:k) {
        u = rnorm(n, 3, 0.6)
        v = rnorm(n, 7, 0.6)
        r = runif(n)
        x = c(u[r>0.5], v[r<=0.5])
        s[i] = mean(x)
    }
    s
}
```

Use this function to study the distribution of $\bar{X}$ for sample sizes of $n = 2$, $n = 5$ and $n = 10$. The population mean is $\mu = 5$ and $\sigma^2 = 4.3$.

Generate samples:

```
samples.size.2 = simulate.bimod(1000, 2)
samples.size.5 = simulate.bimod(1000, 5)
samples.size.10 = simulate.bimod(1000, 10)
```

Histograms:



We can compare the observed means and variances to the values predicted from the Central Limit Theorem:

Sample size n	μ	σ^2/n	Observed mean of $\bar{X}$	Observed variance of $\bar{X}$
2	5.0	2.15	4.94	2.27
5	5.0	0.86	4.98	0.862
10	5.0	0.43	4.96	0.443

The observed mean and variance of $\bar{X}$ for $n = 2$ are calculated by:

```
> mean(samples.size.2)
[1] 4.939386
> var(samples.size.2)
[1] 2.270115
```

Further examples will be studied in the practicals.