UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

4.4 Sampling distributions

A **simple random sample** is a sample selected by a process where every possible sample (of the same size, n) has the same probability of selection. The selection process is left to chance, therefore eliminating the effect of **selection bias**. Due to the random selection mechanism, we do not know (in advance) which sample will occur. Every population element has a known, non-zero probability of selection in the sample, but no element is certain to appear.

Example

Consider a population of size $N = 6$ elements: A, B, C, D, E and F.

We consider all possible samples of size $n = 2$ (*without replacement*, i.e. once an object has been chosen it cannot be selected again).

There are 15 different, but equally likely, such samples:

AB, AC, AD, AE, AF, BC, BD, BE,

BF, CD, CE, CF, DE, DF, EF.

Since this is a simple random sample, each sample has a probability of selection of 1/15.

A population has particular characteristics of interest such as the mean, μ, and variance, σ^2. Collectively, we refer to these characteristics as **parameters**. If we do not have population data, the parameter values will be *unknown*.

'Statistical inference' is the process of estimating the (unknown) parameter values using the (known) sample data.

We use a statistic (called an **estimator**) calculated from sample observations to provide a **point estimate** of a parameter.

Returning to our example, recall there are 15 different samples of size 2 from a population of size 6. Suppose the variable of interest is monthly income, such that:

Individual	A	B	C	D	E	F
Monthly income in £000s	3	6	4	9	7	7

If we seek the population mean, μ, we will use the sample mean, $\bar{X}$, as our estimator where, for a sample of size n, we have:

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

For example, if the observed sample was 'AB', the sample mean is:

$$\frac{3000 + 6000}{2} = £4{,}500.$$

Clearly, different observed samples will lead to different sample means.

Consider the *values* of $\bar{X}$, i.e. $\bar{x}$, for all possible samples (in £000s):

Sample	AB	AC	AD	AE	AF	BC	BD	BE
Values	3, 6	3, 4	3, 9	3, 7	3, 7	6, 4	6, 9	6, 7
$\bar{x}$	4.5	3.5	6	5	5	5	7.5	6.5

Sample	BF	CD	CE	CF	DE	DF	EF
Values	6, 7	4, 9	4, 7	4, 7	9, 7	9, 7	7, 7
$\bar{x}$	6.5	6.5	5.5	5.5	8	8	7

So, the *values* of $\bar{X}$ vary from 3.5 to 8, depending on the sample values.

Since we have the population data here, we can actually compute the population mean, μ, in £000s, which is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{3+6+4+9+7+7}{6} = 6.$$

So, even with simple random sampling, we sometimes obtain $\bar{x}$ values which are far from μ. Here, in fact, only one sample (AD) results in $\bar{x} = \mu$.

Let us now consider the maximum possible **absolute deviations** of the sample mean from the population mean, i.e. the distance $|\ \bar{x} - \mu\ |$.

max $\lvert\ \bar{x} - \mu\ \rvert$	Range of $\bar{x}$	Number of samples	Probability
0	$\bar{x} = 6$	1	0.067
0.5	$5.5 \leq \bar{x} \leq 6.5$	6	0.400
1	$5 \leq \bar{x} \leq 7$	10	0.667
1.5	$4.5 \leq \bar{x} \leq 7.5$	12	0.800
2	$4 \leq \bar{x} \leq 8$	14	0.933
2.5	$3.5 \leq \bar{x} \leq 8.5$	15	1.000

So, for example, there is an 80% probability of being within 1.5 units of μ (in £000s).

We now represent this as a **frequency distribution**. That is, we record the frequency of each possible value of $\bar{x}$.

$\bar{x}$	Frequency	Relative frequency
3.5	1	$1/15 = 0.067$
4.5	1	$1/15 = 0.067$
5.0	3	$3/15 = 0.200$
5.5	2	$2/15 = 0.133$
6.0	1	$1/15 = 0.067$
6.5	3	$3/15 = 0.200$
7.0	1	$1/15 = 0.067$
7.5	1	$1/15 = 0.067$
8.0	2	$2/15 = 0.133$

This is known as the **sampling distribution** of $\bar{X}$. The sampling distribution is a central and vital concept in statistics. It can be used to evaluate how 'good' an estimator is. Specifically, we care about how 'close' the estimator is to the population parameter of interest.

As we have seen, different samples yield different sample mean values, as a consequence of the random sampling procedure. Hence estimators (of which $\bar{X}$ is an example) are random variables.

So, $\bar{X}$ is our estimator of μ, and the observed value of $\bar{X}$, denoted $\bar{x}$, is a point estimate.