5 SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM

5.1 Introduction

Problem
Obtain information about populations. Examples are

1. Mean income of Australians.
2. Proportion voting Labor.
3. Market share.
4. Proportion of faulty items produced.
5. Mean weight of a packet of cereal.

- The above are population quantities — we want: the mean income of ALL Australians; the proportion that is faulty of ALL items produced.
- How can this be done? One way is to make measurements on every unit in the population – check each item for fault; ask every voter their voting intention.
- The problems with this approach are:
  (i) Expense. It is costly to make measurements on all units in a population — census.
  (ii) Takes too much time.
  (iii) Difficult to conduct (census!)
  (iv) Measurements may be destructive – how do you test a match for quality?
  (v) Impossible in some situations. For example:
    - How much gold/oil is present in this deposit?
    - What is the population of tuna, lions, elephants, turtles, whales, seals?

How do we proceed?

Solution
Sampling. Make observations on a subset of the population, then generalise the results to the whole population. THIS REQUIRES CARE!
• How should the sample be selected? The sample should “represent” the population. Inappropriate sampling techniques can lead to bias in results.

BEWARE OF BIAS IN THE SAMPLE!

Examples are:

1. TV polls — Voluntary response survey. Only the viewers of that channel, with strong feelings (and SMS capability) will respond. Such polls are VERY UNRELIABLE.

2. Telephone polls. Only people with telephones and who are in at the time of calling will be in the sample. Non-response (no one at home or unwilling to participate) is a major problem here.

The common sources of bias is samples are:

1. Location bias. Only certain locations are sampled from. This can either be a physical location or a more general one, such as viewers of a particular TV station. Thus only a part of the population is sampled from.

2. Time bias. Samples are taken at fixed times or in a narrow window of time. Again only a certain part of the population is sampled from.

3. Incorrect population is sampled from. An obvious example of this is asking Perth residents what they think of house prices in Sydney.

Deliberately selecting a biased sample to obtain favourable outcomes is common and unethical.

Example 1. In the 1936 Presidential elections there were two candidates — Democratic incumbent Franklin Roosevelt and Republican challenger Alg Landon. The magazine Literary Digest mailed out questionnaires to 10 million American voters. Based on 2.3 million responses, the magazine predicted the results of the election. Below are the magazine’s prediction and the actual results.

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landon</td>
<td>57%</td>
<td>38%</td>
</tr>
<tr>
<td>Roosevelt</td>
<td>43%</td>
<td>62%</td>
</tr>
</tbody>
</table>

Why was the prediction so wrong, despite being based on a sample of 2.3 million? The survey design is very poor because it is based on voluntary response sample. Of the 10 million people in the survey, only 23% responded. In this survey, those voters who were happy for Roosevelt to serve another term felt far less motivated to respond. Even a large sample can be very misleading of the method of data collection is dubious.

Some Definitions

• A population is a set of units of interest.
• A **parameter** is any population measure, such as mean, variance, proportion or population size.

• A **sample** is a subset of the population.

Several sampling schemes exist. An important idea in sampling theory is **randomisation**, that is, each unit in the sample is picked at random from the population.

We need to define probability models for the above concepts so that probability theory can be used in estimation and inference for population parameters.

**PROBABILITY MODEL**

• We want to measure a quantity $X$ (such as salary, voting preference) for units in a population. We need a probability distribution for $X$; we call this the **population distribution**.

• A **parameter** is any quantity associated with a population distribution, and is usually a summary measure of the population.

• A **sample** is a set of independent and identically distributed (iid) random variables $X_1, X_2, \cdots, X_n$, having the same distribution as $X$ (that is, the population distribution).

• A **simple random sample** (srs) is one in which every unit in the population has the same probability of being selected.

• In practice, the form of the population distribution is known, but some parameters will be unknown.

• The only information available about the unknown parameters is the data $x_1, x_2, \cdots, x_n$, which are observations on the random variables $X_1, X_2, \cdots, X_n$ in the sample.

• **ALL estimation** and **inference** is based on the population model and the data.

**Example 2. Market Share**

To determine market share, a business selects a srs of size $n$ and asks each person if she/he uses their product. Let $X$ denote the number who say ‘yes’. Then $X \sim \text{Bin}(n, p)$. This is the population distribution, and the parameter in this model is $p = \text{proportion of market share}$.

A **statistic** is a random variable the observed value of which depends only on the observed data. A statistic is usually a summary measure of the data.

Two common statistics are:

**Sample Mean** \[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i; \text{ observed value } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \]

**Sample Variance** \[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2; \text{ observed value } s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]

Note that $\bar{X}$ and $S^2$ are random variables — their observed values ($\bar{x}$ and $s^2$) depend on the sample selected.
Example 3. Mean product weight

To determine the mean weight of a packed product, \( n \) items are selected at random from a warehouse. Let \( X \) denote the weight of a randomly selected item. Then \( X \sim N(\mu, \sigma^2) \) is the assumed population model, where the model parameters are the mean weight \( \mu \) and variance \( \sigma^2 \) of the weights.

Let \( X_1, X_2, \ldots, X_n \) be the random variables denoting the weights of the items in the sample. Then \( X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2), i = 1, 2, \ldots, n. \)

Problem The weights of boxes of a cereal are normally distributed with mean 500 g and standard deviation 5 g. What is the probability that a random sample of 25 boxes has a sample mean weight less than 495 g?

Solution

We need \( P(\overline{X} < 495) \). TO COMPUTE THIS WE REQUIRE THE PROBABILITY DISTRIBUTION OF \( \overline{X}! \)

**IMPORTANT RESULT**

Let \( X_1, X_2, \ldots, X_n \) be iid random variables with mean \( \mu \) and variance \( \sigma^2 \), that is, \( E(X_i) = \mu \) and \( \text{Var}(X_i) = \sigma^2 \) for \( i = 1, 2, \ldots, n. \) Let \( \overline{X} \) be the sample mean,

\[
\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]

Then

\[
E(\overline{X}) = \mu_{\overline{X}} = \mu
\]

and

\[
\text{Var}(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}.
\]

Proof
Thus if $n$ is large, $\sigma_X$ is small, so $\bar{X}$ is likely to be close to its mean $\mu$.

5.2 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

In many sampling situations the population mean $\mu$ is unavailable and is the parameter of interest. The natural way to estimate $\mu$ is by the sample mean $\bar{X}$. However, the sampling distribution of the sample mean depends on the population variance $\sigma^2$. Different sampling distributions result depending on whether $\sigma^2$ is known or not. Additionally, the population being sampled may not be normal.

We consider three cases below. These three cases from the basis of estimation and inference and will be used throughout the rest of the course.

<table>
<thead>
<tr>
<th>CASE 1 Normal Population — $\sigma$ known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let $X_1, X_2, \cdots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. Then</td>
</tr>
<tr>
<td>$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$</td>
</tr>
<tr>
<td>or</td>
</tr>
<tr>
<td>$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$</td>
</tr>
<tr>
<td>This follows from our earlier results that sums and linear scaling of normal random variables are also normal.</td>
</tr>
</tbody>
</table>

Example 4. The time it takes to server a customer at a bank teller is normal with mean 150 secs and variance 900 secs$^2$.

(i) What is the probability that a random sample of 25 customers will have a mean service time greater than 160 seconds?

(ii) Repeat (i) above for a sample of 100 customers.

Solution
Example 5. A sample of size 100 is drawn from a $N(\mu, 16)$ population. Find

(i) $P(\bar{X} < \mu)$.

(ii) $P(|\bar{X} - \mu| < 0.2)$.

(iii) $k$ such that $P(|\bar{X} - \mu| < k) = 0.95$.

Solution
CASE 2 Normal Population — $\sigma$ Unknown

Now we estimate the population standard deviation $\sigma$ by the sample standard deviation $S$. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where $t_{n-1}$ is a $t$ distribution with $n - 1$ degrees of freedom.

Notes

1. $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is the estimator for $\sigma^2$. The divisor in this expression is $n - 1$, which is the degrees of freedom for the $t$ distribution. ALWAYS the divisor in the estimator for the variance gives the degrees of freedom for the corresponding $t$ distribution for the problem.

2. $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$ or $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, and this is estimated by

$$\frac{S}{\sqrt{n}} = \text{standard error (SE) of the mean}.$$

3. The $t$ distribution is similar in shape to the N(0, 1) distribution, but has **thicker tails** and is **flatter**.
$t$ tables list the $t$-values for only the tail probabilities for given degrees of freedom. The $t$-tables available on line will be the one used in examinations.

**Example 6.** Complete the table of critical values below by referring to the $t$ tables.

<table>
<thead>
<tr>
<th>df ($\nu$)</th>
<th>1</th>
<th>9</th>
<th>60</th>
<th>500</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\nu}^{0.05}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_{\nu}^{0.025}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For small values of df, the $t$ values are fairly different from the $N(0, 1)$ values. As $\nu \to \infty$, $t_{\nu} \to N(0, 1)$.

**Example 7.** A sample of size $n = 10$ gives $s^2 = 9$. find $k$ such that

$$P (|\bar{X} - \mu | > k) = 0.05.$$  

**Solution**
CASE 3 Population Distribution not Normal

CENTRAL LIMIT THEOREM (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are iid random variables with mean $\mu$ and variance $\sigma^2$. If $n$ is large enough, then

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

Notes

1. We consider $n \geq 30$ to be large enough for the approximation to be good. As $n$ increases the approximation improves.

2. If $\sigma$ is unknown then

$$T = \frac{\overline{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

**BUT we must have** $n \geq 30$. If the sample size is large (a few hundred), then the $t_{n-1}$ can be approximated by $N(0, 1)$ distribution. This is useful for computing probabilities.

Example 8. A particular long-life light bulb has a mean life of 1500 hours with a standard deviation of 300 hours. What is the probability that a random sample of 200 bulbs has a mean life less than 1450 hours?

Solution
Summary

1. A population is modelled by a random variable \( X \).

2. A sample is a set of iid random variables \( X_1, X_2, \ldots, X_n \), each with the same distribution as that of \( X \). The data \( x_1, x_2, \ldots, x_n \) are the observed values of \( X_1, X_2, \ldots, X_n \).

3. Sample mean = \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \).

\[
E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n},
\]

where \( \mu \) = population mean and \( \sigma^2 \) = population variance.

4. (a) If the population is normally distributed, that is, \( X \sim N(\mu, \sigma^2) \), then
   i. if the population standard deviation \( \sigma \) is known, then
      \( Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \).
   ii. if the population standard deviation \( \sigma \) is unknown, then estimate it by the sample standard deviation \( S \), and then
      \( T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \).

   (b) If the population distribution is not normal or is unknown and \( n \geq 30 \), then
   i. if the population standard deviation \( \sigma \) is known, then
      \( Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \).
   ii. if the population standard deviation is unknown, estimate it by the sample standard deviation \( s \), and then
      \( \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \).

If \( n \) is large (a few hundred), then the t-distribution can be approximated by \( N(0,1) \).

Note If the sample size \( n \) is small \((n < 30)\) then we must assume that the population is normal, and use either 4 (a) (i) (if \( \sigma \) is known) or (ii) (if \( \sigma \) is unknown).