Sampling and Central Limit Theorem

Week 9

Prepared by: Nurazrin Jupri
Key Definitions

- A **population** (universe) is the collection of things under consideration.
- A **sample** is a portion of the population selected for analysis.
Data are often collected from a sample rather than from a population.

If the whole population is examined, the survey is called a census.
Population and Sample

Inference on the population from the sample

Use statistics to summarize features
Sampling methods

Simple Random
A sampling frame is constructed. So every item has an equal chance of being selected.

Stratified Random
Population is divided into strata or categories and random samples taken from each strata.

Quota
Interviewers interview all the people they meet up to a certain quota.

Cluster
Non-random, selects one definable subsection of population as the sample.

Multistage
Divides population into sub-populations then selects sample of sub-populations at random.
Random sampling

• Items selected in *no systematic way* (‘at random’)
• All items potentially have *same ‘chance’* of selection
• *Random number tables* can help here: e.g. if telephone directory is sampling frame, then such tables identify each page and line of the directory to select each person for sample
• Useful for statistical inference
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>61424</td>
<td>20419</td>
<td>86546</td>
<td>00517</td>
<td></td>
</tr>
<tr>
<td>90222</td>
<td>27993</td>
<td>04952</td>
<td>66762</td>
<td></td>
</tr>
<tr>
<td>50349</td>
<td>71146</td>
<td>97668</td>
<td>86523</td>
<td></td>
</tr>
<tr>
<td>85676</td>
<td>10005</td>
<td>08216</td>
<td>25906</td>
<td></td>
</tr>
<tr>
<td>02429</td>
<td>19761</td>
<td>15370</td>
<td>43882</td>
<td></td>
</tr>
<tr>
<td>90519</td>
<td>61988</td>
<td>40164</td>
<td>15815</td>
<td></td>
</tr>
<tr>
<td>20631</td>
<td>88967</td>
<td>19660</td>
<td>89624</td>
<td></td>
</tr>
<tr>
<td>89990</td>
<td>78733</td>
<td>16447</td>
<td>27932</td>
<td></td>
</tr>
</tbody>
</table>
Select five random numbers from 1 to 25. The resulting sample consists of population elements 3, 7, 9, 16, and 24. Note, there is no element from Group C.
Stratified random sampling

- dividing the population into *strata* or *categories*.
- Random samples are then taken from *each* stratum or category.
- It takes *more time* than simple random sampling but samples should be *more representative* and sample error should be reduced.
Randomly select a number from 1 to 5 for each stratum, A to E. The resulting sample consists of population elements 4, 7, 13, 19 and 21. Note, one element is selected from each column.
Quota sampling

- **cheapness and administrative simplicity.**
- Investigators are told to interview all the people they *meet up to a certain quota.*
- The advantages - a much **larger sample** can be studied, **more information** can be gained at a **faster speed** for a given outlay.
Example: Quota sampling

An investigator’s quotas might be as follows.

<table>
<thead>
<tr>
<th>Category</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partnerships</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Public companies</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Private companies</td>
<td>30</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td>Public practice</td>
<td>30</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>250</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using quota sampling, the investigator would interview the first 30 male cost and management accountants in partnerships that he met, the first 20 female cost and management accountants in partnerships that he met and so on.
Multistage sampling

- probability sampling method which involves dividing the **population into a number of sub-populations** and then selecting a small sample of these sub-populations at **random**.
- Each sub-population is then **divided further**, and then a small sample is again selected at random.
- **fewer investigators** are needed and it is **not so costly** to obtain a sample.
- However, there is the **possibility of bias**
- not truly random
Example: Multistage sampling

- A survey of spending habits is being planned to cover the whole of Britain.
- A fair approximation to a random sample can be obtained.
- Thus, we might choose a random sample of eight areas, and from each of these areas, select a random sample of five towns. From each town, a random sample of 200 people might be selected so that the total sample size is $8 \times 5 \times 200 = 8,000$ people.
Cluster sampling

- a **non-random** sampling method that involves selecting one *definable subsection* of the population as the sample,
- that subsection taken to be representative of the population in question.
- For example, the pupils of one school might be taken as a cluster sample of all children at school in one county.
- it is **inexpensive** to operate. However, there is potential for considerable **bias**.
Sample size

• Sample size of 30 or more permits important statistical tests to be used

• The greater the sample size, the more confident we can be that any sample statistic represents the ‘population’ from which the sample is drawn
Distribution of sample means

• When the population is normal,
  – the *distribution of sample means* is normal
  – have the *same mean* as for the population,
  – have a *different ‘standard deviation’*, which we call the *standard error* (SE)

\[
SE = \frac{\sigma}{\sqrt{n}} \text{ or } \frac{s}{\sqrt{n}}
\]
Distribution of sample means

\[
\text{SD} = \sigma
\]

\[
SE = \frac{\sigma}{\sqrt{n}}
\]
Z Statistic

- We can still use our Z statistics and the Z tables since the distribution of sample means is normal

\[ Z = \frac{X_i - \mu}{\sigma / \sqrt{n}} \]  

- or

\[ Z = \frac{X_i - \mu}{s / \sqrt{n}} \]
Standardising the Normal Distribution

The tables in the book give this area

What is the probability sales are greater than 450 units?

\[
z = \frac{x - \mu}{\sigma} \quad z = \frac{450 - 400}{50} = +1, \text{ Therefore from the tables } p = 0.1587, \text{ and that is the required result.}
\]
What is the probability sales are less than 450 units?

\[ z = \frac{X - \mu}{\sigma} \]

\[ z = \frac{450 - 400}{50} = +1 \]

Therefore from the tables \( p = 0.1587 \), but the answer is now \( 1 - 0.1587 = 0.8413 \)
Standardising the Normal Distribution

When the calculated value of $Z$ is not a whole number for example $Z = 0.13$

A combination of Row and Column has to be used

Probability associated with a particular $z$ value, in this case $z=0.13$ and the answer is 0.45,

You then have to decide how the answer relates to the question

<table>
<thead>
<tr>
<th>$Z$</th>
<th>0</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>0.3</td>
<td>0.33</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.35</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.42</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>0.4</td>
<td>0.4</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>0.1</td>
<td>0.46</td>
<td>0.46</td>
<td>0.45</td>
<td>0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
<td>0.43</td>
<td>0.42</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.49</td>
<td>0.49</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
<td>0.47</td>
<td>0.47</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Question

- A production line makes units which are normally distributed with a mean weight of 200 grams and a standard deviation of 9 grams. What is the probability of a sample of 36 units having a (sample) mean weight of 203 grams or more?
\[ Z_1 = \frac{203 - 200}{9/\sqrt{36}} = \frac{3}{1.5} \]

\[ Z_1 = +2 \]

\[ P_1 = 0.0228 \]
Solution \[ P_1 = 0.0228 \]
Central Limit Theorem

- Even if the population is not normal, if:
  - sampling is random
  - the sample size \((n)\) is \(\geq 32\),
  - then the distribution of sample means can be regarded as approximately normal

- Use Z statistic and Z tables for calculating probabilities for this distribution of sample means
Central Limit Theorem
Question

• The output of glass panels has a mean thickness of 4cm and a standard deviation of 1cm. If a random sample of 100 glass panels is taken, what is the probability of the sample mean having a thickness of between 3.9cm and 4.2cm?
$Z_1 = \frac{3.9 - 4.0}{1/\sqrt{100}} = \frac{-0.1}{0.1} = -1$

$P_1 = 0.1587$

$A_1 = 0.5000 - 0.1587$

$A_1 = 0.3413$
\[ Z_2 = \frac{4.2 - 4.0}{1/\sqrt{100}} = \frac{0.2}{0.1} = +2 \]

\[ P_2 = 0.0228 \]
\[ A_2 = 0.5000 - 0.0228 \]
\[ A_2 = 0.4772 \]
Solution
\[ A_1 + A_2 = 0.8185 \]
Confidence intervals

• A confidence interval is a range of values within which we can have a certain level of confidence that a particular value of a variable will lie

• 95% and 99% confidence intervals are the most usual
95% confidence interval for sample mean
Question

• A large number of random samples of size 100 are taken from the production line of glass panels which is thought to produce panels with mean thickness 4cm and standard deviation 1cm

Find

• a) the 95% and
• b) the 99% confidence intervals for the sample mean
(a) Finding the 95% confidence interval for the sample mean

(b) Finding the 95% confidence interval for the sample mean

\[ SE = \frac{\sigma}{\sqrt{n}} \]

\[ SE = \frac{1}{\sqrt{100}} = 0.1 \]
SE = \frac{\sigma}{\sqrt{n}}
SE = \frac{1}{\sqrt{100}} = 0.1

(b) Finding the 99% confidence interval for the sample mean
Solutions

a) $4 \pm 1.96 \frac{1}{\sqrt{100}}$

$4 \pm 0.196$

3.804 to 4.196 cm

b) $4 \pm 2.58 \frac{1}{\sqrt{100}}$

$4 \pm 0.258$

3.742 to 4.258 cm