Chapter 6

Sampling and Sampling Distributions

6.1 Definitions

- A statistical population is a set or collection of all possible observations of some characteristic.

- A sample is a part or subset of the population.

- A random sample of size $n$ is a sample that is chosen in such a way as to ensure that every sample of size $n$ has the same probability of being chosen.

- A parameter is a number describing some (unknown) aspect of a population. (i.e. $\mu$)

- A statistic is some function of the sample observations. (i.e. $\bar{X}$)

- The probability distribution of a statistic is known as a sampling distribution. (How is $\bar{X}$ distributed)

- We need to distinguish the distribution of a random variable, say $\bar{X}$ from the realization of the random variable (i.e. we get data and calculate some sample mean say $\bar{X} = 4.2$)
CHAPTER 6. SAMPLING AND SAMPLING DISTRIBUTIONS

**Populations and Samples**

- **A Population** is the set of all items or individuals of interest
  - **Examples:**
    - All likely voters in the next election
    - All parts produced today
    - All sales receipts for November

- **A Sample** is a subset of the population
  - **Examples:**
    - 1000 voters selected at random for interview
    - A few parts selected for destructive testing
    - Random receipts selected for audit

Figure 6.1:
Figure 6.2:
Note on Statistics

- The value of the statistic will change from sample to sample and we can therefore think of it as a random variable with its own probability distribution.

- $\bar{X}$ is a random variable

- Repeated sampling and calculation of the resulting statistic will give rise to a distribution of values for that statistic.
A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population.
Chapter Outline

- Sampling Distributions
  - Sampling Distribution of Sample Mean
  - Sampling Distribution of Sample Proportion
  - Sampling Distribution of Sample Difference in Means

Figure 6.4:
6.2 Important Theorems Recalled

Suppose $X_1, X_2, \ldots, X_n$ are independent with $E[X_i] = \mu_i$ and $V[X_i] = \sigma_i^2 \; \forall i = 1, 2, \ldots, n$.

Suppose $Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + b$, then:

$$E[Y] = E\left[\sum a_i X_i + b\right] = a_1 E[X_1] + a_2 E[X_2] + \cdots + a_n E[X_n] + b$$

$$= a_1 \mu_1 + \ldots a_n \mu_n + b$$

$$= \sum a_i \mu_i + b$$

and

$$V[Y] = V\left[\sum a_i X_i + b\right] = a_1^2 V[X_1] + a_2^2 V[X_2] + \cdots + a_n^2 V[X_n] .$$

$$= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2$$

$$= \sum a_i^2 \sigma_i^2 \text{ because of independence}$$

and if $X_i$ is normal $\forall i$, i.e. $X_i \sim N(\mu_i, \sigma_i^2)$ independently $\forall i$:

$$Y \sim N\left(\sum a_i \mu_i + b, \sum a_i^2 \sigma_i^2\right)$$

6.3 Frequently used Statistics

6.3.1 The sample mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a population with mean $\mu$ and variance $\sigma^2$. The sample mean is:
\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

1. The **expected value of the sample mean** is the population mean:

\[ E[\bar{X}] = E\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} (\mu + \mu + \ldots + \mu) = \mu \]

2. The **variance of the sample mean** (\( X_i \)s independent):

\[ V[\bar{X}] = V\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \]

3. If we do not have **independence** it can be shown that

\[ V[\bar{X}] = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right) \text{ where } N \text{ is the population size} \]

\[ \left( \frac{N-n}{N-1} \right) \text{ is called the correction factor} \]

and if \( N \) is large relative to \( n \) then \( \left( \frac{N-n}{N-1} \right) \Rightarrow 1 \) so that \( V[\bar{X}] = \frac{\sigma^2}{n} \)

**Note on Sample Mean**

1. The use of the formulas for expected values and variances of sums of random variables that we saw in chapter 5.

2. The variance of the sample mean is a **decreasing** function of the sample size.

3. The **standard deviation of the sample mean** (under independence)

\[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \]
6.3.2 The sample variance

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a population with variance $\sigma^2$.
- The sample variance is:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

1. $E[s^2] = \sigma^2$ (we omit the proof)

6.4 Sampling distribution of the Sample Mean

Sampling from a Normal Population

- Let $\overline{X}$ be the sample mean of an independent random sample of size $n$ from a population with mean $\mu$ and variance $\sigma^2$.
- Then we know that $E[\overline{X}] = \mu$ and $V[\overline{X}] = \frac{\sigma^2}{n}$.
- If we further specify the population distribution as being normal, then

$$X_i \sim N(\mu, \sigma^2) \text{ for all } i$$

and we can write:

$$\overline{X} \sim N \left( \mu, \frac{\sigma^2}{n} \right).$$

6.5 Equation for the Standardized Sample Mean

Since $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ we can ask what transformation can give us to a standard normal

- Generically the approach is ALWAYS

$$Z = \frac{\text{Random Variable-Mean of Random Variable}}{\text{Standard Deviation of Random Variable}}$$

- What does that mean for $\overline{X}$:
• Random Variable = $\bar{X}$,

• Mean of Variable = $E[\bar{X}] = \mu$

• Standard Deviation of Variable = $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

• Put it all together

\[ Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}. \]
### Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of \( \bar{X} \):

\[
Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}
\]

where:
- \( \bar{X} \) = sample mean
- \( \mu \) = population mean
- \( \sigma \) = population standard deviation
- \( n \) = sample size

Figure 6.5:
Example of Standardizing for Sample Mean

The lengths of individual machined parts coming off a production line at Morton Metalworks are normally distributed around their mean of $\mu = 30$ centimeters. Their standard deviation around the mean is $\sigma = .1$ centimeter. An inspector just took a sample of $n = 4$ of these parts and found that $\bar{X}$ for this sample is 29.875 centimeters. What is the probability of getting a sample mean this low or lower if the process is still producing parts at a mean of $\mu = 30$?

Answer

- Given the population is normally distributed with mean $\mu = 30$ and standard deviation $\sigma = .1$, we know that $X_i \sim N(30,.1^2)$ for all $i$, so

\[ \bar{X} \sim N(30,.1^2/n). \]

Now want to apply our transformation stuff:
6.5. EQUATION FOR THE STANDARDIZED SAMPLE MEAN

\[ P(\bar{X} \leq 29.875) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{29.875 - 30}{1/\sqrt{n}}) \]

\[ = P(Z \leq \frac{29.875 - 30}{1/\sqrt{4}}) = P(Z \leq -2.50) = .0062. \]

Questions: 7.3.

6.5.1 Sampling from a Non Normal Distribution

- We have seen that we can obtain the exact sampling distribution for the sample mean if the individual \( X_i \) are all independent normal variates.

- What happens when the \( X_i \)'s are not normally distributed?
Developing a Sampling Distribution

- Assume there is a population ...
- Population size $N=4$
- Random variable, $X$, is age of individuals
- Values of $X$: 18, 20, 22, 24 (years)

Figure 6.7:
Developing a Sampling Distribution

Summary Measures for the Population Distribution:

\[
\mu = \frac{\sum X_i}{N}
\]

\[
= \frac{18 + 20 + 22 + 24}{4} = 21
\]

\[
\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236
\]

Figure 6.8:
6.5.2 The Central Limit Theorem

Now consider all possible samples of size $n = 2$.
Developing a Sampling Distribution

Sampling Distribution of All Sample Means

16 Sample Means

<table>
<thead>
<tr>
<th>1st Observation</th>
<th>2nd Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>18 20 22 24</td>
</tr>
<tr>
<td>20</td>
<td>19 20 21 22</td>
</tr>
<tr>
<td>22</td>
<td>20 21 22 23</td>
</tr>
<tr>
<td>24</td>
<td>21 22 23 24</td>
</tr>
</tbody>
</table>

Sample Means Distribution

P(\bar{X})

(no longer uniform)

Figure 6.10:
Let $X_1, X_2, \ldots, X_n$ be an independent random sample having identical distribution from a population of any shape with mean $\mu$ and variance $\sigma^2$.

Then if $n$ is large:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

approximately for large $n$

and similarly we can use the same transformation to standard form:

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

approximately for large $n$

Notes on the Central Limit Theorem

1. This result holds only for large $n$ and we refer to such results as holding asymptotically. In this case we say that $\bar{X}$ is asymptotically normally distributed

2. We have a short-hand way to write that the distribution of $X_i$ is independently and identically (iid) distributed with mean $\mu$ and variance $\sigma^2$

$$X_i \sim i.i.d(\mu, \sigma^2)$$

3. Identically implies that $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ for all $i$. That is the distribution of each observation ($i = 1, \ldots, n$) is the same.
As the sample size gets large enough... the sampling distribution becomes almost normal regardless of shape of population.
Example

- Suppose a population has mean $\mu = 8$ and standard deviation $s = 3$. Suppose a random sample of size $n = 36$ is selected.

- What is the probability that the sample mean is between 7.8 and 8.2?

Figure 6.12:
Example From Transformation to Standard Form when Sampling from a Non-Normal Distribution

- The delay time for inspection of baggage at a border station follows a bimodal distribution with a mean of $\mu = 8$ minutes and a standard deviation of $\sigma = 6$ minutes. A sample of $n = 64$ from a particular minority group has a mean of $\bar{X} = 10$ minutes.

- Is there evidence that this minority group is being detained longer than usual?

- How likely is it that a sample mean of $\bar{X} \geq 10$ if the population mean is 8?

Answer:

- Note from the question that this population from which we are sampling is non-normal.

- We know this is from the word bimodal since the normal has one mode.
• If the $X_i$ are iid, then by applying (or invoke) the central limit theorem:

$$P(\bar{X} \geq 10) \approx P(Z \geq \frac{10 - 8}{6/\sqrt{64}})$$

$$= P(Z \geq 2.67) = .0038,$$

approximately.

**Questions:** NCT 7.11, 7.18, 7.19, &7.20.

---

### 6.6 Sampling Distribution: Differences of Sample Means $\bar{X}_1 - \bar{X}_2$

• Let $\bar{X}_1$ and $\bar{X}_2$ be the means of two samples from two separate and independent populations.

$$E[\bar{X}_1] = \mu_1 \quad V[\bar{X}_1] = \frac{\sigma_1^2}{n_1}$$

$$E[\bar{X}_2] = \mu_2 \quad V[\bar{X}_2] = \frac{\sigma_2^2}{n_2}$$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent:

$$E[\bar{X}_1 - \bar{X}_2] = E[\bar{X}_1] - E[\bar{X}_2] = \mu_1 - \mu_2$$

$$V[\bar{X}_1 - \bar{X}_2] = V[\bar{X}_1] + V[\bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

#### 6.6.1 Normal Case for $\bar{X}_1 - \bar{X}_2$

If $X_{1,i}$ are i.i.d. as $N(\mu_1, \sigma_1^2)$, $X_{2,i}$ are i.i.d. as $N(\mu_2, \sigma_2^2)$, and $X_1$ and $X_2$ are independent,

then we have an exact normal result for any sample size:

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

• Recall a linear combination of normals is normal
6.6. **SAMPLING DISTRIBUTION: DIFFERENCES OF SAMPLE MEANS** $\bar{X}_1 - \bar{X}_2$

6.6.2 Non Normal Case for $\bar{X}_1 - \bar{X}_2$

If $X_{1,i}$ are i.i.d. with mean $\mu_1$ and variance $\sigma_1^2$, $X_{2,i}$ are i.i.d. with mean $\mu_2$ and variance $\sigma_2^2$, and $X_1$ and $X_2$ are independent, then using the central limit theorem, for large $n_1$, and $n_2$

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$ approximately or asymptotically

6.6.3 Example of Difference of Means with Non-Normal Population

Suppose that right-handed (RH) students have a mean IQ of 80 units with variance 1,400 and left-handed (LH) students have a mean IQ of 80 with variance 1,320. What is the probability that the sample mean IQ of RH students will be at least 5 units higher than the sample mean IQ of LH students if we take a sample of 100 RH students and 120 LH students?

Answer: Let $X_{1,i}$ be the IQ of the ith RH student and $X_{2,i}$ be the IQ of the ith LH student.

$$\mu_1 = 80, \quad \sigma_1^2 = 1400, \quad n_1 = 100$$

$$\mu_2 = 80, \quad \sigma_2^2 = 1320, \quad n_2 = 120$$

and we want to find $P(\bar{X}_1 - \bar{X}_2 \geq 5)$.

Since $n_1$ and $n_2$ are both large we can apply the central limit theorem,

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$ approximately

here:

$$\mu_1 - \mu_2 = 80 - 80 = 0 \quad \text{and} \quad \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = 25$$

Note: Formula for the standardizing transformation is:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$
where

\[
\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.
\]

So that

\[\bar{x}_1 - \bar{x}_2 \sim N(0, 25) \text{ approximately}\]

\[P(\bar{x}_1 - \bar{x}_2 \geq 5) = P(Z \geq 1) = .1587\]
6.7 Sampling Distribution of Sample Proportion

Let $X$ be a binomially distributed random variable (the number of successes in $n$ trials).

- Recall the sample proportion is
  $$\hat{p} = \frac{X}{n}$$
  is the fraction of successes in $n$ trials.

- In Chapter 6 we used the normal approximation to the binomial as the number of trials got large ($n\pi(1 - \pi) \geq 9$)

- This is another application of the **Central Limit Theorem**
  $$E(X) = \mu = n\pi$$ and $$Var(X) = \sigma^2 = n\pi(1 - \pi).$$

- So we might ask what is the $E[p]$ and $V[p]$

- $$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{E[X]}{n} = \frac{n\pi}{n} = \pi$$

- Recall trials are independent so that
  $$V[\hat{p}] = V\left[\frac{X}{n}\right] = \frac{V[X]}{n^2} = \frac{n\pi(1 - \pi)}{n^2} = \frac{\pi(1 - \pi)}{n}$$

- So we apply the generic formula

$$Z = \frac{\text{Random Variable-Mean of Random Variable}}{\text{Standard Deviation of Random Variable}}$$

- What is mean for $\hat{p}$
  - Random Variable is $\hat{p}$,
  - Mean of Variable is $E[\hat{p}] = p$

- Standard Deviation of Variable is $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$

- Put it all together

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ for large } n$$
6.8 Sampling Distribution of Sample Proportion: $\hat{p}_1 - \hat{p}_2$

- Consider two independent populations.

1. Population 1:

   \[ X_1 = \text{number of successes}, \ n_1 = \text{number in sample 1}, \ \text{so } \hat{p}_1 = \frac{X_1}{n_1}. \]

   Recall

   \[ E[\hat{p}_1] = p_1 \]

   and

   \[ V[\hat{p}_1] = \frac{\pi_1(1 - \pi_1)}{n_1}. \]

2. Population 2:

   \[ X_2 = \text{number of successes} \]

   \[ n_2 = \text{number in sample 2} \]

   \[ \hat{p}_2 = \frac{X_2}{n_2} \]

   \[ V[\hat{p}_2] = \frac{p_2(1 - p_2)}{n_2} \]

3. Form Difference of Sample Proportion: $\hat{p}_1 - \hat{p}_2$

   - If $n_1$ and $n_2$ are large, i.e. $n_1p_1(1 - p_1) \geq 9$ and $n_2p_2(1 - p_2) \geq 9$ then:

     \[ \hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}) \text{ approximately.} \]

   - This is another application of the central limit theorem

Questions: NCT 7.21, 7.22, 7.27 & 7.36.

Omit Sampling Distribution of the Sample Variance