Chapter 7 Sampling and Sampling Distribution

Example:

There are 2500 managers in the electronic associates. The average annual salary of 2500 managers is \$51800. That is,

$$\mu = \frac{\sum_{i=1}^{2500} y_i}{2500} = 51800 = \text{population mean},$$

where y_i is the *i'th* manager's annual salary. Also, assume the population variance of the salary data is

$$\sigma^2 = \frac{\sum_{i=1}^{2500}(y_i - \mu)^2}{2500} = \frac{\sum_{i=1}^{2500}(y_i - 51800)^2}{2500} = 4000^2.$$

Assume that 1500 of the 2500 managers have completed the training program. Then,

$$p = \frac{1500}{2500} = 0.6 = \text{population proportion of completing the program}$$

Objective: we try to just use part of the data (because it is too costly and time consuming to use all the data) and thus obtain the accurate guess for the population mean, variance and proportion.

7.1 Simple Random Sampling

Simple random sampling from finite population:

A simple random sample of size *n* from a finite population of size *N* is a sample selected such that each possible sample of size *n* has the same probability of being selected.

Note: the total number of random samples is $\binom{N}{n} = \frac{N!}{n!(N-n)!}$**. Thus, the probability of a specific random sample being selected is** $\frac{1}{\binom{N}{n}}$**.**

Note: the above sampling is also called a random sample without replacement since we did not place a selected element back into the population. That is, the sample element can not be selected twice.

Simple random sampling from infinite population:

- **Each element selected comes from the same population.**
- **Each element is selected independently.**

7.2 Point Estimate

Example:

Suppose we randomly select 30 managers as a sample. Let $x_1, x_2, \ldots, x_{30}$ be the annual salaries of the 30 managers (a sample). Suppose 19 of them have completed the training program. Thus,

$$\bar{x} = \frac{\sum_{i=1}^{30} x_i}{30} = 51814 = \text{sample mean}$$

and

$$\bar{p} = \frac{19}{30} = 0.63 = \text{sample proportion}$$

Note: $\bar{x}$ and $\bar{p}$ are sensible estimates of $\mu = 51800$ and $p = 0.6$.

Point estimate:

A point estimate is a statistic based on a sample of size n (not necessary to be simple random sample) from a finite population of size N. Suppose $x_1, x_2, \ldots, x_n$ is the sample. Then,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{sample mean}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \text{sample variance}$$

and the sample proportion $\bar{p}$ are point estimates of the population mean μ, the population variance σ^2, and the population proportion *p*, respectively.

Note: $\bar{x}$, s^2 and $\bar{p}$ are not the only estimates. They are just some sensible estimates.

7.3 Sampling Distributions

Example:

Suppose we sample 500 times and let

$$\begin{aligned} x_1^1, \quad x_2^1, \quad \dots, x_{30}^1 \quad &\Rightarrow \bar{x}_1 \\ x_1^2, \quad x_2^2, \quad \dots, x_{30}^2 \quad &\Rightarrow \bar{x}_2 \\ \vdots \quad \vdots \quad \vdots \quad &\vdots \\ x_1^{500}, x_2^{500}, \dots, x_{30}^{500} \quad &\Rightarrow \bar{x}_{500} \end{aligned}$$

be 500 simple random samples of 30 managers. ***Let $\bar{X}$ be the random variable representing the average salary of a random sample of 30 managers.*** Then, $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{500}$ are 500 possible values of $\bar{X}$. Note there are $\binom{2500}{30}$ **possible $\bar{x}$ for random variable $\bar{X}$.**

Properties of $\bar{X}$:

Let $\bar{X}$ be the random variable representing the average of a random sample. Then,

1. $E(\bar{X}) = \mu \equiv$ the population mean
2. ***For finite population with population size N and the sample size n, then the***

standard deviation of $\overline{X}$ ***is***

$$\sqrt{Var(\overline{X})} = \sqrt{E(\overline{X}-\mu)^2} = \sigma_{\overline{X}} = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}.$$

For infinite population (the infinite population size),

$$\sqrt{Var\ (\overline{X})} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Note: ***as*** $\frac{n}{N} \le 0.05$, ***then*** $\sqrt{\frac{N-n}{N-1}} \approx 1$. ***Thus,*** $\sigma_{\overline{X}} = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}} \approx \frac{\sigma}{\sqrt{n}}$. ***That is, the standard deviation of*** $\overline{X}$ ***for the finite population is approximately equal to the one for the infinite population.***

Sampling Distribution of $\overline{X}$:

As $n \ge 30$ ***or the population is normally distributed, then***

$$\overline{X} \approx N(\mu, \sigma^2_{\overline{X}})$$

, where $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. ***Thus, for some constants*** $-\infty \le c \le d \le \infty$,

$$P(c \le \overline{X} \le d) = P(c-\mu \le \overline{X}-\mu \le d-\mu) = P(\frac{c-\mu}{\sigma_{\overline{X}}} \le \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \le \frac{d-\mu}{\sigma_{\overline{X}}})$$

$$\approx P(\frac{c-\mu}{\sigma_{\overline{X}}} \le Z \le \frac{d-\mu}{\sigma_{\overline{X}}}),$$

where Z is the standard normal random variable.

Example:

What is the probability of the difference between the sample mean and the population mean will be less or equal to 500 as the sample size $n = 30$?

[solution]

$\mu = 51800, \sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.30$. Thus,

$\bar{X} \approx N(51800, (730.30)^2) \Rightarrow \frac{\bar{X} - 51800}{730.30} \approx Z \equiv$ the standard normal random variable

$$\Rightarrow P(|\bar{X} - 51800| \le 500) = P(-500 \le \bar{X} - 51800 \le 500)$$

$$= P\left(\frac{-500}{730.30} \le \frac{\bar{X} - 51800}{730.30} \le \frac{500}{730.30}\right)$$

$$\approx P(-0.68 \le Z \le 0.68) = 0.5036$$

$\Rightarrow$ There is 50.36% chance that the difference between the sample mean and the population mean is not more than 500.

Sample Size and Sampling Distribution:

Since $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, ***increasing the sample size will decrease the standard error!!***

Thus, the larger the sample size is, the larger $P(c \le \bar{X} \le d)$ ***is (since the interval***

$\left[\frac{c - \mu}{\sigma_{\bar{X}}}, \frac{d - \mu}{\sigma_{\bar{X}}}\right]$ ***is larger than the one with smaller sample size)!!***

Example:

What is the probability of the difference between the sample mean and the population mean will be less or equal to 500 as the sample size $n = 100$?

[solution]

$\mu = 51800, \sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400 \le 730.30$ $(n = 30)$. Thus,

$\bar{X} \approx N(51800, (400)^2) \Rightarrow \frac{\bar{X} - 51800}{400} \approx Z \equiv$ the standard normal random variable

$$\Rightarrow P(|\bar{X}-51800|\le 500)=P(51300\le \bar{X}\le 52300)$$

$$=P\left(\frac{51300-51800}{400}\le\frac{\bar{X}-51800}{400}\le\frac{52300-51800}{400}\right)$$

$$\approx P(-1.25\le Z\le 1.25)=0.7888\ge 0.5036\,(n=30)$$

$\Rightarrow$ As the sample size increases to 100, there is 78.88% chance that the difference between the sample mean and the population mean is not more than 500. That is, the larger sample size will provide a higher probability that the value of the sample mean will be within a specific distance of the population mean.

Properties of $\bar{P}$:

Let $\bar{P}$ be the random variable representing the proportion of a random sample (the sample proportion of a random sample is one possible value of $\bar{P}$). Then,

1. $E(\bar{P}) = p \equiv$ the population proportion
2. ***For finite population with population size N and the sample size n, then the standard deviation of $\bar{P}$ is***

$$\sqrt{Var(\bar{P})}=\sqrt{E(\bar{P}-p)^2}=\sigma_{\bar{P}}=\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}.$$

For infinite population (the infinite population size),

$$\sqrt{Var\,(\bar{P})}=\sigma_{\bar{P}}=\sqrt{\frac{p(1-p)}{n}}.$$

Note:

As $\frac{n}{N}\le 0.05$, ***then*** $\sqrt{\frac{N-n}{N-1}}\approx 1$. ***Thus,*** $\sigma_{\bar{X}}=\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}\approx\sqrt{\frac{p(1-p)}{n}}$.

That is, the standard deviation of $\bar{P}$ for the finite population is approximately equal to the one for the infinite population.

Sampling Distribution of $\bar{P}$:

As $np\ge 5$ and $n(1-p)\ge 5$,

$$\overline{P} \approx N(p, \sigma_{\overline{P}}^2)$$

, where $\sigma_{\overline{P}}^2 = \frac{p(1-p)}{n}$ ***. Thus, for some constants*** $0 \le c \le d \le 1$

$$P(c \le \overline{P} \le d) = P(c - p \le \overline{P} - p \le d - p) = P(\frac{c-p}{\sigma_{\overline{P}}} \le \frac{\overline{P} - p}{\sigma_{\overline{P}}} \le \frac{d-p}{\sigma_{\overline{P}}})$$

$$\approx P(\frac{c-p}{\sigma_{\overline{P}}} \le Z \le \frac{d-p}{\sigma_{\overline{P}}}),$$

where Z is the standard normal random variable.

Note: ***Since*** $\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}}$, ***increasing the sample size will decrease the standard error!! Thus, the larger the sample size is, the larger*** $P(c \le \overline{P} \le d)$ ***is (since the interval*** $\left[\frac{c-p}{\sigma_{\overline{P}}}, \frac{d-p}{\sigma_{\overline{P}}}\right]$ ***is larger than the one with smaller sample size)!!***

Example:

What is the probability of the difference between the sample proportion and the population proportion will be less or equal to 0.05 as the sample size $n = 30$? What is the probability as we increase the sample size to 100?

[solution]

$p = 0.6, \sigma_{\overline{P}} \approx \sqrt{\frac{0.6(1-0.6)}{30}} = 0.0894$. Thus,

$\overline{P} \approx N(0.6, (0.0894)^2) \Rightarrow \frac{\overline{P} - 0.6}{0.0894} \approx Z \equiv$ the standard normal random variable

$$\Rightarrow P(|\overline{P} - 0.6| \le 0.05) = P(-0.05 \le \overline{P} - 0.6 \le 0.05)$$

$$= P(\frac{-0.05}{0.0894} \le \frac{\overline{P} - 0.6}{0.0894} \le \frac{0.05}{0.0894})$$

$$\approx P(-0.56 \le Z \le 0.56) = 0.4246$$

$\Rightarrow$ There is 42.46% chance that the difference between the sample proportion and

the population proportion is not more than 0.05 as . $n = 30$.

As sample size is increased to 100, then

$p = 0.6, \sigma_{\bar{P}} \approx \sqrt{\frac{0.6(1-0.6)}{100}} = 0.0490 \le 0.0894 \ \ (n = 30)$. Thus,

$\bar{P} \approx N(0.6, (0.0490)^2) \Rightarrow \frac{\bar{P} - 0.6}{0.0490} \approx Z \equiv$ the standard normal random variable

$$\Rightarrow P(|\bar{P} - 0.6| \le 0.05) = P(0.55 \le \bar{P} \le 0.65)$$

$$= P(\frac{0.55-0.6}{0.0490} \le \frac{\bar{P} - 0.6}{0.0490} \le \frac{0.65-0.6}{0.0490})$$

$$\approx P(-1.02 \le Z \le 1.02) = 0.6922 \ge 0.4246 \, (n = 30)$$

$\Rightarrow$ There is 69.22% chance that the difference between the sample proportion and the population proportion is not more than 0.05 $n = 100$. That is, the larger sample size will provide a higher probability that the value of the sample proportion will be within a specific distance of the population proportion.

7.4 Other Sampling Methods

1. Stratified Random Sampling:

- The population is divided into groups of elements called strata according to some "characteristic" of the data.
- A simple random sample is taken from each stratum.

How to determine the sample size in each stratum?

- ***According to the size of the stratum.***
- ***According to the variance of each stratum.***

2. Cluster Sampling:

- The population is divided into several separate groups of elements called clusters.
- A simple random sample of the clusters is taken. ***All*** elements within these ***sampled*** or ***"selected"*** clusters are in the sample.

Note: ***one of the primary applications of cluster sampling is area sampling, where clusters are city blocks or other well-defined areas!!***

3. Systematic Sampling:

- Select randomly one of the first $\frac{N}{n}$ elements, where n and N are the sample size and the population size, respectively.
- Starting from the first selected element, select every $\left(\frac{N}{n}\right)$'th element after the first element.

Example:

Suppose $n = 50, N = 5000,$ and $y_1, y_2, \ldots, y_{5000}$ are the elements in the population. Since $\frac{N}{n} = \frac{5000}{50} = 100$, by using systematic sampling, we should select randomly one from the first 100 elements first. Suppose the third element is selected, i.e. $x_1 = y_3$. Then, select every 100'th element after y_3, thus

$$x_2 = y_{103}, x_3 = y_{203}, \ldots, x_{50} = y_{4903}.$$