Survey Sampling

Summary of Ch. 7

Def: Simple Random Sampling

Each particular sample of size n has the same probability of occurrence; that is each of the $\binom{N}{n}$ possible samples of size n taken without replacement has the same probability.

Remark: Any statistic computed from a random sample is a random variable and has an associated sampling distribution.

The sampling distribution of $\bar{X}$ determines how accurately an $\bar{x}$ estimates μ. Roughly specking, the more tightly the sampling distribution is centered around μ, the better the estimate.

CLT: $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$

As a measure of the center of the sampling distribution we use $E[\bar{X}] = \mu$ and as a measure of the dispersion of the sampling distribution we will use $SD(\bar{X}) = \sigma/\sqrt{n}$.

Thm A: With SRS $E[\bar{X}] = \mu$.

So $\bar{X}$ is an unbiased estimator of μ.

Lemma B: With SRS, without replacement

$$Cov(X_i, X_j) = -\frac{\sigma^2}{N-1} \qquad i \neq j \tag{1}$$

Thm B:

$$\begin{aligned} Var(\bar{X}) &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) & (2)\\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right) & (3)\\ &= \frac{\sigma^2}{n}(f.p.c.) & (4) \end{aligned}$$

Where $f.p.c.$ is the finite population correction.

So if N is large, then $Var(\bar{X}) \approx \sigma^2/n$.

Estimation of the Population Variance σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

a biased estimator.

Thm:

$$E(\hat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)\left(\frac{N}{N-1}\right) \tag{6}$$

<u>Cor:</u> An unbiased estimate of $Var(\bar{X})$ is

$$s^2_{\bar{X}} = \frac{\hat{\sigma}^2}{n}\left(\frac{n-1}{n}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{7}$$

$$= \frac{s^2}{n}\left(1-\frac{n}{N}\right) \tag{8}$$

If N is large then $s^2_{\bar{X}} \approx s^2/n$.

In practice we disregard the finiteness of the population and assume $n << N$. This implies independence in the sampling and

$$Cov(X_i, X_j) \approx 0 \tag{9}$$

<u>CLT:</u> Normal Approximation to the Sampling Distribution of $\bar{X}$.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{10}$$

The spread of the sampling distribution and therefore the precision of $\bar{X}$ are determined by the sample size n and not by the populations size N.

<u>Confidence Intervals:</u>

$100(1-\alpha)\%$ C.I. for μ, large $n \geq 30$

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{11}$$

$100(1-\alpha)\%$ C.I. for μ, small $n < 30$

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{12}$$

$100(1-\alpha)\%$ C.I. for p, large n

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}} \tag{13}$$

<u>Inference Procedures</u>

<u>Point Estimation</u>

Assuming there is a population with parameter μ a simple random sample (SRS) is taken to produce a sample of size n, $x_1, x_2, ..., x_n$. From the sample we calculate the sample mean $\bar{x}$ as an unbiased estimate of μ.

<u>Interval Estimation</u>

Starting before the data from a SRS, $X_1, X_2, ..., X_n$, is collected from a population with unknown mean μ and known standard deviation σ^2, a $100(1-\alpha)\%$ confidence interval is a random interval.

$$
\begin{aligned}
P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) &= 1-\alpha \\
P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) &= 1-\alpha \\
P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \\
P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha
\end{aligned}
$$

Since $\bar{X}$ is a random variable the endpoints of the last interval are random.

So before we collect our data the confidence interval

$$
\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{14}
$$

has a 95% probability of including μ. After we collect our data, the confidence interval

$$
\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{15}
$$

can be interpreted as follows:

"We are $100(1-\alpha)$% confident that our interval includes the population mean μ."

Note: We do not use the word probability when discussing a single confidence interval after it has been computed. A single interval either includes μ or it does not, we do not know.

Starting before the data from a SRS, $X_1, X_2, ..., X_n$, is collected from a population with unknown mean μ and unknown standard deviation σ^2, a $100(1-\alpha)$% confidence interval is a random interval.

$$
\begin{aligned}
P\left(-t_{\alpha/2} \le T \le t_{\alpha/2}\right) &= 1-\alpha \\
P\left(-t_{\alpha/2} \le \frac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{\alpha/2}\right) &= 1-\alpha \\
P\left(\mu - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \bar{X} \le \mu + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) &= 1-\alpha \\
P\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) &= 1-\alpha
\end{aligned}
$$

So the confidence interval for μ when σ^2 is unknow is

$$
\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{16}
$$

Aside: Sample size calculation. "Forethought in Statistics."

Find the sample size n needed to have a margin-or-error of .03 with 95% confidence when estimating the population proportion π.

The $100(1-\alpha)\%$ confidence interval for π.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{17}$$

So

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .03 \tag{18}$$

A conservative estimate of π is to let $\hat{p} = 0.5$.

$$\begin{aligned} 1.96\sqrt{\frac{(.5)^2}{n}} &= .03 \\ \left[\frac{(1.96)(.50^2}{.03}\right]^2 &= n \\ n &= 1067 \end{aligned}$$

Hypothesis Testing

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned}$$

For a SRS $X_1, X_2, ..., X_n$ then

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{19}$$

Therefore we reject H_0 is Z falls in the tails of the distribution.