Math 408 - Mathematical Statistics

Lecture 15. Accuracy of estimation of the population mean $\overline{X}_n \approx \mu$

February 25, 2013

In Lecture 12, we discussed the basic mathematical framework of survey sampling:

- We have the target population of size N (N is very large).
- A numerical value of interest x_i (age, weight, income, etc) is associated with i^{th} member of the population.
- We are interested in population parameters:
  - Population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We estimate μ by the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ is a sample drawn from the population using the simple random sampling.

We proved that $\overline{X}_n$ is an unbiased estimate of μ:

$$\boxed{\mathbb{E}[\overline{X}_n] = \mu}$$

In other words, on average $\overline{X}_n \approx \mu$.

Our next goal is to investigate how variable $\overline{X}_n$ is

As a measure of the dispersion of $\overline{X}_n$ about μ, we will use the standard deviation of $\overline{X}_n$, $\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$.

Thus, we want to find

$$\boxed{\mathbb{V}[\overline{X}_n] = ?}$$

$$\mathbb{V}[\overline{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right]$$

<u>Remark:</u> If sampling were done with replacement then X_i would be independent, and we would have:

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

In simple random sampling, we do sampling without replacement.
This induces dependence among X_i. And therefore

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \neq \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i]$$

Recall Lecture 6:

$$\mathbb{V}\left[\sum_{i=1}^{n} \alpha_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \mathrm{Cov}(X_i, X_j)$$

Thus, we have:

$$\mathbb{V}[\overline{X}_n] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

So, we need to find $\mathrm{Cov}(X_i, X_j)$.

Lemma

If $i \neq j$, then the covariance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

Theorem

The variance of $\overline{X}_n$ is given by

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

Important observations:

- If $n << N$, then

$$\mathbb{V}[\overline{X}_n] \approx \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

  $\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction.

- To double the accuracy of $\mu \approx \overline{X}_n$, the sample size must be quadrupled
- If σ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if σ is large, then a larger sample will be required to obtain the same accuracy.

Summary

- The main result of this lecture is the expression for the variance of $\overline{X}_n$:

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

- The corresponding standard deviation

$$\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$$

measures the dispersion of $\overline{X}_n$ about μ.