12 Chapter 8 - Large Sample Estimation

> Whenever we take a sample, we do so with the idea of learning something about the population from which the sample is drawn. Provided that the sample is drawn in an unbiased manner we believe that it may be taken as representative of the parent population. But representatives are not all equally authoritative. Spokesmen - even official spokesmen - do not always tell a reliable tale, and it is necessary in retelling a story secondhand from such a source that we indicate the degree of confidence which may be placed in what the spokesman has said. Just as the journalist tries to emphasize for his readers the difference between rumours and 'usually well informed sources', so too the statistician has to attempt a similar thing. ... Given large sample, the problem is easily enough disposed of intuitively. ... But, when the samples are small, we have to face not only the possibility of bias but also the fact that the average, standard deviation, or proportion found in the sample may differ be quite appreciably from the population parameters it is sought to estimate through the sample. It is evident that there can be no possibility of finding a method of estimation which will guarantee us a close estimate under all conditions. All we can hope for is a method which will be the best possible in the sense that it will have a high probability of being correct in the long run.
>
> M.J. Moroney
> *Facts from Figures*

1. Points covered in this chapter
   (a) Two approaches to estimate population parameters (e.g., to estimate the mean and variance of the population (for a normal distribution) and the proportion (for a binomial distribution) when these values are unknown).
   (b) Properties of sound estimators.
   (c) Calculation of the margin of error and confidence intervals.
   (d) How to choose a sample size.
2. Prior material used in this section
   (a) Standard error measurement
   (b) z scores
   (c) Tchebysheff's Theorem and Empirical Rule.

(d) Central Limit Theorem - if the sample size is large (e.g, if n is large), the sampling distribution will be approximately normal. If the sample is normal, we have a large set of statistical tools to our disposal.

3. Types of estimators

   (a) **Point Estimator** - what is the best single value that can be used to estimate a population parameter.

   (b) **Interval Estimator** - what is the best interval (refer to as a confidence interval) that contains the population estimate. Tied to the notion of the confidence interval is the confidence coefficient $(1-\alpha)$ where $(0.01 \leq \alpha \leq 0.10)$

4. Properties of point estimators

   (a) **Unbiased** - average values of the estimated parameter equals the population parameter.

   (b) **Consistent** - Estimators from sample converge to the true value as the sample size increases.

   (c) **Efficient** - Estimator with smallest sampling variance.

5. Univariate Analysis

   (a) Estimating point estimator

      i. For population mean (μ). $\rightarrow \bar{x}$
      Margin of error $\rightarrow$ 1.96* Standard error of the estimator or

$$\pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

      If σ is unknown and n $\geq$ 30, one can substitute s for σ.

      ii. For population proportion (p). $\rightarrow \hat{p} = \frac{x}{n}$
      Margin of error

$$\pm 1.96\sqrt{\frac{pq}{n}}$$

      estimated as

$$\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

      Recall: $n\hat{p} > 5$ and $n\hat{q} > 5$

   (b) Estimating interval estimator

i. General function - two tail test

$$\text{Point estimator} \pm z_{\frac{\alpha}{2}} * \text{Standard Error}$$

A. Population mean

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

When n > 30

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

B. Population proportion

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

ii. General function - left tail test (one-sided confidence interval)

$$\text{Point estimator} - z_{\alpha}\text{Standard Error}$$

iii. General function - right tail test (one-sided confidence interval)

$$\text{Point estimator} + z_{\alpha}\text{Standard Error}$$

iv. values of $z_{\frac{\alpha}{2}}$ and z_{α} for given values of α.

α	$z_{\frac{\alpha}{2}}$ **- two tail**	z_{α} **-one tail**	**Confidence**
0.010	2.58	2.33	99.0%
0.020	2.33	2.055	98.0%
0.025	2.24	1.96	97.5%
0.050	1.96	1.645	95.0%
0.100	1.645	1.28	90.0%

6. Bivariate Analysis - This type of analysis works with two samples each drawn from different populations. For this form of bivariate analysis, the research question is, 'Are the populations different?'. Using the example from the textbook, one would want to test if the average MCAT scores for biochemistry and biology majors are the same. If there is no difference between these two populations (biochemistry and biology students), the difference in their population means ($\mu_1 - \mu_2$) would equal 0. This research question will be addressed briefly in this section and in more detail in Chapters 9 and 10. Right now, we wish to deal with the point and interval estimates from the samples drawn from two populations.

   There are two kinds of data sets used for bivariate analysis

   (a) Data sets are not paired (or data sets are independent from one another). There is no relationship between the two parameters differenced (e.g., MCAT scores for Biochemistry and Biology Majors).

(b) Data sets are paired (e.g., there is a relationship between the two data sets). Examples, comparing the differences in gas mileage when a car is first given one type and then another type of gasoline, test scores of trainees before and after viewing an instructional video.

(a) Properties of Sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ - not paired
Mean and Standard error

$$\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$$

$$SE = \sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Margin of Error

$$\pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Confidence interval (two-tail)

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

i. If sampled populations are normally distributed, then the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is normally distributed regardless of size.
ii. If the sampled populations are not normally distributed, then the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ is approximately normally distributed when n_1 and n_2 are large, due to the Central Limit Theorem.
iii. If σ_1^2 and σ_2^2 are unknown, but both n_1 and n_2 are greater than or equal to 30, you can substitute the sample variances for the population variances.
iv. Use z values found in section 5.B.iv (on previous page).

(b) Properties of Sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ - not paired
Mean and Standard Error

$$(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$$

$$\mu_{(\hat{p}_1 - \hat{p}_2)} = (p_1 - p_2)$$

$$SE = \sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Margin of Error

$$\pm 1.96\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

Confidence Interval (two-tail)

$$(\hat{p}_1-\hat{p}_2)\pm z_{\frac{\alpha}{2}}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

i. The sampling distribution of $(\hat{p}_1-\hat{p}_2)$ is approximately normally distributed when n_1 and n_2 are large, due to the Central Limit Theorem.

ii. n_1 and n_2 must be sufficiently large so that the sampling distribution of $(\hat{p}_1-\hat{p}_2)$ can be approximated by a normal distribution. n_1p_1, n_1q_1, n_2p_2 and $n_2q_2 > 5$.

iii. Use z values found in section 5.B.iv.

(c) Properties of Sampling distribution of $(\bar{x}_1-\bar{x}_2)$ - paired
Mean and Standard error

$$\mu_{(\bar{x}_1-\bar{x}_2)}=\mu_1-\mu_2$$

$$SE=\sigma_{(\bar{x}_1-\bar{x}_2)}=\sqrt{\frac{\sigma_D^2}{n}}$$

where σ_D^2 is the variance of the differenced data and $n_1=n_2=$ n.
Margin of Error

$$\pm 1.96\sqrt{\frac{\sigma_D^2}{n}}$$

Confidence interval (two-tail)

$$(\bar{x}_1-\bar{x}_2)\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_D^2}{n}}$$

(d) Properties of Sampling distribution of $(\hat{p}_1-\hat{p}_2)$ - paired
Will not be covered in this class.

7. Choosing a sample size
Choosing a sample size is an application of the point and interval estimation techniques. Suppose you want to generate a sample such that the margin of error is equal to some value, let's call it B. You also want a sample such that 95% of repeated sampling will given you a margin of error less than or equal to B.

For univariate and bivariate analyses, each margin of error function is a function of n. Here is the case for the population mean, univariate case.

$$B \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

If I rearrange the function above, one finds the function for computing the sample size.

$$n \geq \left(\frac{1.96}{B}\right)^2 \sigma^2$$

If σ is not known, the sample standard deviation can be used or a value based on the range of the values divided by 4.

In order to prepare a sample with a different degrees of confidence, just replace the margin of error function with the confidence interval function, two tail version. Below is a table of the set of function one can use to determine the sample size. B is equal to margin of error.

Analysis	**Estimator**	**Minimum sample size**
Univariate	$\bar{x}$	$n \geq (z^2_{\alpha/2}\sigma^2)/B^2$
	$\hat{p}$	$n \geq (z^2_{\alpha/2}pq)/B^2$
Bivariate - not paired	$\bar{x}_1 - \bar{x}_2$	$n \geq (z^2_{\alpha/2}(\sigma_1^2 + \sigma_2^2))/B^2$
	$\hat{p}_1 - \hat{p}_2$	$n \geq (z^2_{\alpha/2}(p_1q_1 + p_2q_2)/B^2$

For the Bivariate functions $n_1 = n_2 = n$.
B is the acceptable margin of error.
If σ is not known, the sample standard deviation can be used or a value based on the range of the values divided by 4.