LECTURE 5: SIMPLE RANDOM SAMPLING 3.4, 3.5, 3.7

3.5 Reliability of Estimates

Recall that reliability of an estimator can be measured by the size of its variance (equivalently, the standard error). Last time me proved that $\mathbb{E}(\bar{x}) = \bar{X}$, and

$$\mathbb{V}ar(\bar{x}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\left(\frac{N\sigma_X^2}{N-1}\right)$$

where the population variance

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2\,.$$

Since σ_X is unknown, we estimate $\mathbb{V}ar(\bar{x})$ by

$$\widehat{\mathbb{V}ar}(\bar{x}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)s_x^2$$

where the sample variance

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is an unbiased estimator of

$$\frac{N}{N-1}\,\sigma_X^2\,.$$

Hence,

$$\widehat{SE}(\bar{x}) = \frac{s_x}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}\,,$$

and a $100(1-\alpha)\%$ confidence interval for the population mean $\bar{X}$ under the simple random sampling is

$$\bar{x} \pm z_{1-\alpha/2}\,\frac{s_x}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}\,.$$

Likewise, it can be shown that a $100(1-\alpha)\%$ confidence interval for the population total $X = \sum_{i=1}^{N} X_i$ is

$$x' \pm z_{1-\alpha/2}\,\frac{N\,s_x}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

where the point estimate

$$x' = \frac{N}{n}\sum_{i=1}^{n} x_i \;\text{ with }\; SE(x') = \frac{N}{\sqrt{n}}\sqrt{\frac{N}{N-1}}\,\sigma_X\sqrt{1 - \frac{n}{N}}\,.$$

Special case. When the measurements are binary, the sample mean is the sample proportion

$$p_x = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

The population variance in this case is $\sigma_X^2 = P_X(1-P_X)$, and the sample variance

$$s_x^2 = \frac{np_x(1-p_x)}{n-1}$$

is an unbiased estimator of

$$\frac{N}{N-1}\sigma_X^2\,.$$

Thus, the estimated variance is

$$\widehat{\mathbb{V}ar}(p_x) \;=\; \frac{1}{n}\Big(1-\frac{n}{N}\Big)s_x^2 \;=\; \frac{1}{n}\Big(1-\frac{n}{N}\Big)\,\frac{np_x(1-p_x)}{n-1} \;=\; \Big(1-\frac{n}{N}\Big)\,\frac{p_x(1-p_x)}{n-1}\,.$$

A $100(1-\alpha)\%$ confidence interval for the population proportion has the form

$$p_x \,\pm\, z_{1-\alpha/2}\sqrt{1-\frac{n}{N}}\,\sqrt{\frac{p_x(1-p_x)}{n-1}}\,.$$

3.4 Coefficients of Variation of Estimated Population Parameters

Let $\hat{d}$ denote an estimator of a population parameter d. Define the coefficient of variation $V(\hat{d})$ as

$$V(\hat{d}) \;=\; \frac{SE(\hat{d})}{d}\,.$$

Recall that the population coefficient of variation is denoted by

$$V_X = \frac{\sigma_X}{\bar{X}}\,.$$

Proposition

$$\text{(i)}\;\; V(\bar{x}) \;=\; \frac{V_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\,,$$

$$\text{(ii)}\;\; V(x') \;=\; \frac{V_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\,,$$

and

$$\text{(iii)}\; V(p_x) = \sqrt{\frac{1-P_X}{n\,P_X}}\,\sqrt{\frac{N-n}{N-1}}\,.$$

Proof. (i) $SE(\bar{x})$ can be written as

$$SS(\bar{x}) \;=\; \sqrt{\mathbb{V}ar(\bar{x})} \;=\; \sqrt{\frac{1}{n}\Big(1-\frac{n}{N}\Big)\Big(\frac{N\sigma_X^2}{N-1}\Big)} \;=\; \frac{\sigma_X}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\,.$$

Thus,

$$V(\bar{x}) \stackrel{def}{=} \frac{SE(\bar{x})}{\bar{X}} = \frac{\sigma_X}{\bar{X}} \frac{1}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{V_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

(ii) $SE(x')$ is equal to

$$SE(x') = \frac{N}{\sqrt{n}} \sqrt{\frac{N}{N-1}}\, \sigma_X \sqrt{1-\frac{n}{N}} = \frac{\sigma_X\, N}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

The coefficient of variation

$$V(x') \stackrel{def}{=} \frac{SE(x')}{X} = \frac{1}{X} \frac{\sigma_X\, N}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{\sigma_X}{\bar{X}} \frac{1}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{V_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

(iii)

$$SE(p_x) = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{P_X(1-P_X)}{n}} \sqrt{\frac{N-n}{N-1}}.$$

Therefore,

$$V(p_x) \stackrel{def}{=} \frac{SE(p_x)}{P_X} = \frac{1}{P_X} \sqrt{\frac{P_X(1-P_X)}{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1-P_X}{n\,P_X}} \sqrt{\frac{N-n}{N-1}}. \quad \square$$

3.7 How Large a Sample Do We Need?

Consider a $100(1-\alpha)\%$ confidence interval for a population parameter d, $\hat{d} \pm z_{1-\alpha/2}\, SE(\hat{d})$. Suppose we would like the margin of error, $z_{1-\alpha/2}\, SE(\hat{d})$, not to exceed a pre-specified value $\varepsilon\, d$, that is, we want to find the sample size n such that

$$z_{1-\alpha/2}\, SE(\hat{d}) \le \varepsilon\, d\,.$$

Equivalently, we want to find n, which is the smallest integer satisfying

$$z_{1-\alpha/2}\, V(\hat{d}) \le \varepsilon\,.$$

Proposition (i) If $d = \bar{X}$ and $\hat{d} = \bar{x}$, then

$$n \ge \frac{N\, z^2_{1-\alpha/2}\, V_X^2}{(N-1)\,\varepsilon^2 + z^2_{1-\alpha/2}\, V_X^2}\,.$$

(ii) If $d = X$ and $\hat{d} = x'$, then

$$n \ge \frac{N\, z^2_{1-\alpha/2}\, V_X^2}{(N-1)\,\varepsilon^2 + z^2_{1-\alpha/2}\, V_X^2}\,.$$

(iii) If $d = P_X$ and $\hat{d} = p_x$, then

$$n \geq \frac{N\, z_{1-\alpha/2}^2\,(1-P_X)}{(N-1)\varepsilon^2\, P_X \,+\, z_{1-\alpha/2}^2\,(1-P_X)}\,.$$

PROOF. We will show only part (i). Parts (ii) and (iii) are left as the homework.
(i) If $d = \bar{X}$ and $\hat{d} = \bar{x}$, then n is the smallest integer such that

$$z_{1-\alpha/2}\,\frac{V_X}{\sqrt{n}}\,\sqrt{\frac{N-n}{N-1}} \leq \varepsilon\,,\ \text{ or }\ z_{1-\alpha/2}^2\,\frac{V_X^2}{n}\left(\frac{N-n}{N-1}\right) \leq \varepsilon^2\,,\ \text{ or }\ \frac{N}{n}-1 \leq \frac{(N-1)\,\varepsilon^2}{z_{1-\alpha/2}^2\,V_X^2}\,,$$

$$\text{or }\ \frac{N}{n} \leq \frac{(N-1)\,\varepsilon^2}{z_{1-\alpha/2}^2\,V_X^2} + 1 = \frac{(N-1)\,\varepsilon^2 \,+\, z_{1-\alpha/2}^2\,V_X^2}{z_{1-\alpha/2}^2\,V_X^2}\,,$$

$$\text{or }\ n \geq \frac{N\, z_{1-\alpha/2}^2\,V_X^2}{(N-1)\,\varepsilon^2 \,+\, z_{1-\alpha/2}^2\,V_X^2}\,. \quad \square$$

Remark From this proposition, if N is very large, then (i) if $\hat{d} = \bar{x}$ or $\hat{d} = x'$, an approximate required sample size satisfies

$$n \geq \left(\frac{z_{1-\alpha/2}\,V_X}{\varepsilon}\right)^2.$$

(ii) if $\hat{d} = p_x$,

$$n \geq \frac{z_{1-\alpha/2}^2\,(1-P_X)}{\varepsilon^2\, P_X}\,.$$

Example (on pages 73 – 75) We are given $N = 2500$, $\varepsilon = 0.1$, $z = z_{1-\alpha/2} = 3$ (if not specified explicitly). To estimate V_X, we use the given information: $N_0 = 1,000$, $\bar{x}_0 = 70$, and $s_x^2 = 14$. We have

$$\hat{V}_X^2 = \frac{\hat{\sigma}_X^2}{\bar{x}^2} = \frac{N_0 - 1}{N_0}\,s_x^2\,\frac{1}{\bar{x}^2} = \frac{(999/1000)(14^2)}{70^2} = 0.03996\,.$$

Thus,

$$n \geq \frac{(2500)(9)(0.03996)}{(2499)(0.1)^2 + (9)(0.03996)} = 35.47\,,$$

or $n = 36$. If we use the approximate formula, we get

$$n \geq \frac{(9)(0.03996)}{(0.1)^2} = 35.96\,,\ \text{ or }\ n = 36\,.$$