Section 2, 9/10/19

Survey Sampling

Motivation: We wish to learn properties (statistics) about the population but it is unrealistic to have data from the entire population. So, we must infer these statistics from a subset of the population, a sample.

Population statistics

Assuming we had the entire population, we would be able to know the statistics. Let $x_1, \ldots, x_N$ be measurements from the entire population of size N.

Population mean: $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

Population Total: $\tau = N\mu$

Population Variance: $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

Note that $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 = \mathbb{E}(x^2) - \mathbb{E}(x)^2$

The standard deviation is σ and measure the "spread" of the distribution around the mean.

Simple Random Sampling

We only have access to a subset of the population measurements, a sample of size $n < N$ in which each sample is sampled without replacement and with equal probability from the population (sampling distribution). Since the sample is random, the statistics themselves are random variables.

Let our random sample be $X_i, \ldots, X_n$ and note that x_i and X_i are not the same since x_i is the fixed measurement of the ith individual and X_i is the ith randomly sampled measurement.

Sample Mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Population Total: $T = N\bar{X}$ (we assume N is known)

We can examine properties of $\bar{X}$ such as its variance, which is determined by the sampling distribution.

Ex. A

Suppose we have 393 hospitals and wish to estimate their statistical properties using $n = 16$ sample hospitals. However, the sampling distribution is unknown. It could be calculated by considering all $\binom{393}{16}$ possible samples, but that is computationally infeasible. So, we run a **simulation** by drawing m samples of size n and using those samples to estimate the distribution.

Estimate Properties

First, $\bar{X}$ is "centered" at μ. We call it an **unbiased** estimator of μ. This proof easily extends to the estimated population total.

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

What is the variance of the estimate? Start by using its relation to the covariance.

$$\begin{aligned} Var(\bar{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) \end{aligned}$$

Since the mean squared error $MSE = variance + bias^2$, the MSE of these unbiased estimators is just their variance.

With Replacement (Simple random sampling)

If sampling is done with replacement, then $Cov(X_i, X_j) = 0$ for $i \neq j$ since they are independent and thus

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Cov(X_i, X_i) = \frac{\sigma^2}{n} = \sigma_{\bar{X}}^2$$

The standard error of our estimate $\bar{X}$ is $\sigma_{\bar{X}}$

Sampling Without Replacement

Thus we can solve for the variance of $\bar{X}$ in the case of sampling without replacement. Through some algebra considering the covariance terms,

$$Var(\bar{X}) = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$$

which is the sample variance scaled by the **finite population correction**. Since n/N is typically small, the standard error $\sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}}$.

Ex. B

Back to the hospital example, say we sample from $n = 32$ hospitals,

$$\begin{aligned} \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \\ &= \frac{589.7}{\sqrt{32}}\sqrt{1 - \frac{31}{392}} \\ &= 104.2 \times 0.96 \\ &= 100.0 \end{aligned}$$

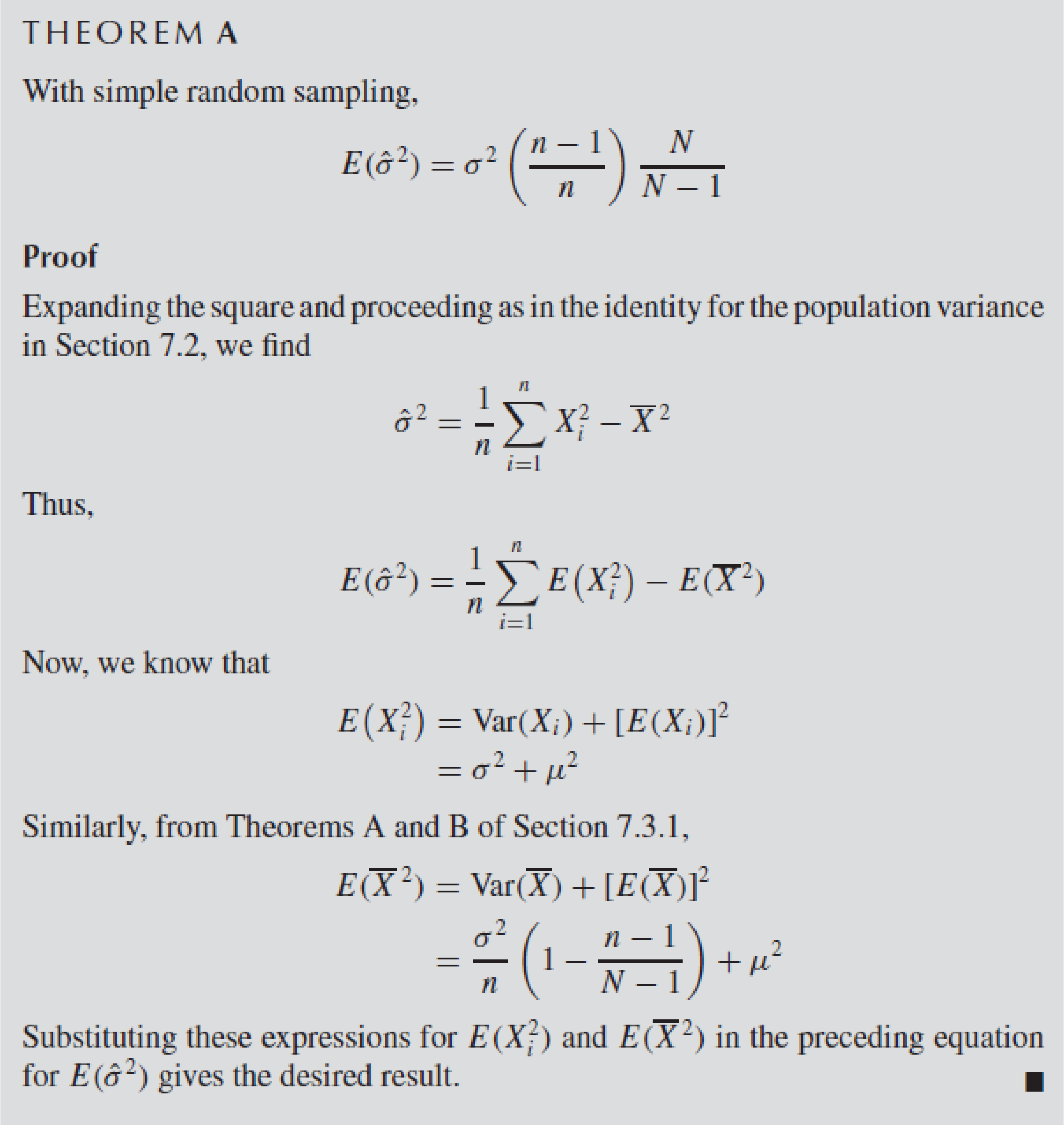
This is the variance of our estimate and can be verified by the distribution from the simulation in example A.

Ex. C

Say that a proportion $p = 0.654$ of the hospitals had fewer than 1000 discharges. Since a proportion deals with measurements that are binary (either there were fewer or greater than 1000 discharges), the variance is the bernoulli variance $\sigma^2 = p(1-p)$. If we wish to find an estimate $\hat{p}$ of this statistic using a sample, then the variance of our estimate is

$$\begin{aligned}\sigma_{\hat{p}} &= \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} \\ &= \frac{\sqrt{p(1-p)}}{\sqrt{32}}\sqrt{1-\frac{31}{392}} \\ &= 0.08\end{aligned}$$

Expectation of estimated variance



THEOREM A

With simple random sampling,

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

Proof

Expanding the square and proceeding as in the identity for the population variance in Section 7.2, we find

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Thus,

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E(\overline{X}^2)$$

Now, we know that

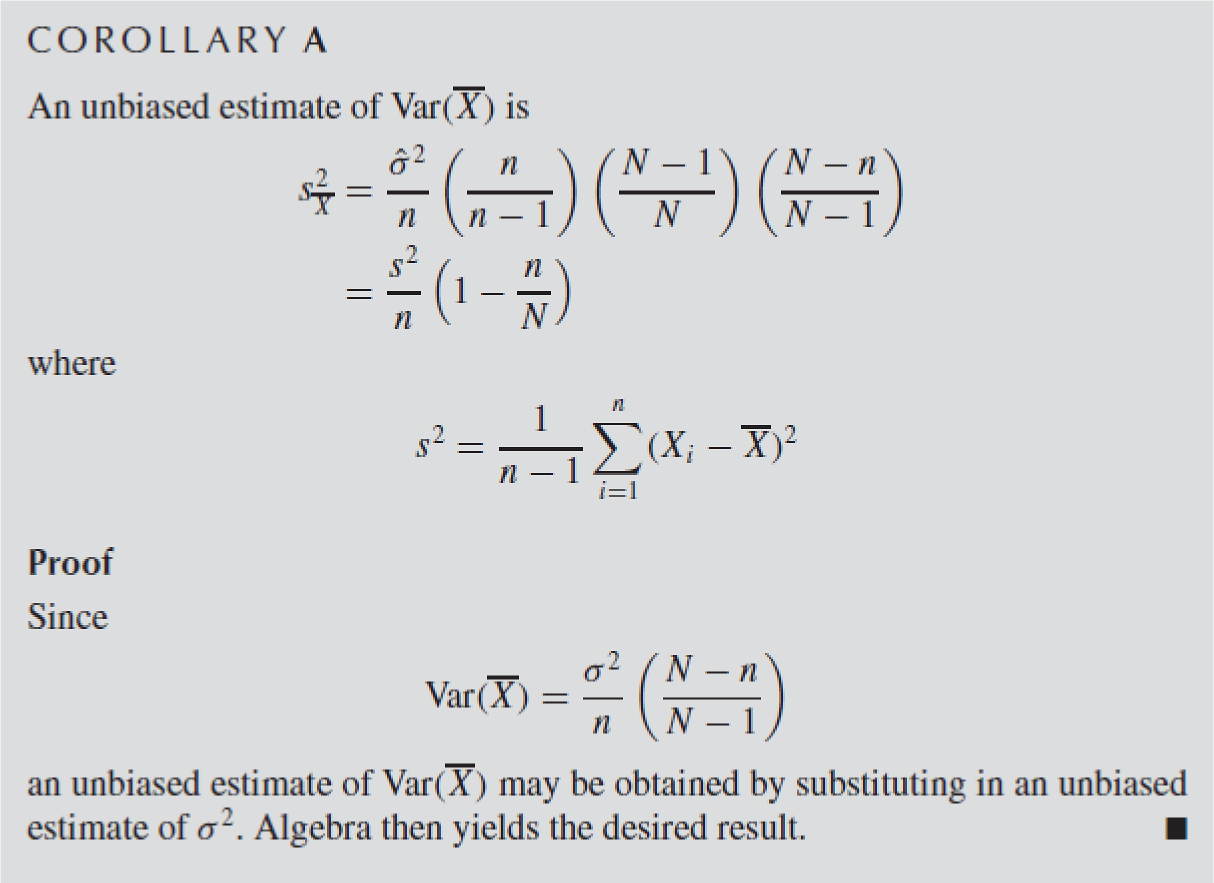
$$\begin{aligned}E\left(X_i^2\right) &= \mathrm{Var}(X_i) + [E(X_i)]^2 \\ &= \sigma^2 + \mu^2\end{aligned}$$

Similarly, from Theorems A and B of Section 7.3.1,

$$\begin{aligned}E(\overline{X}^2) &= \mathrm{Var}(\overline{X}) + [E(\overline{X})]^2 \\ &= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right) + \mu^2\end{aligned}$$

Substituting these expressions for $E(X_i^2)$ and $E(\overline{X}^2)$ in the preceding equation for $E(\hat{\sigma}^2)$ gives the desired result. ■

Thus it's a biased estimator. To find an unbiased estimator, we multiply by constant terms.



COROLLARY A

An unbiased estimate of $\mathrm{Var}(\overline{X})$ is

$$s_{\overline{X}}^2 = \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right)$$
$$= \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

where

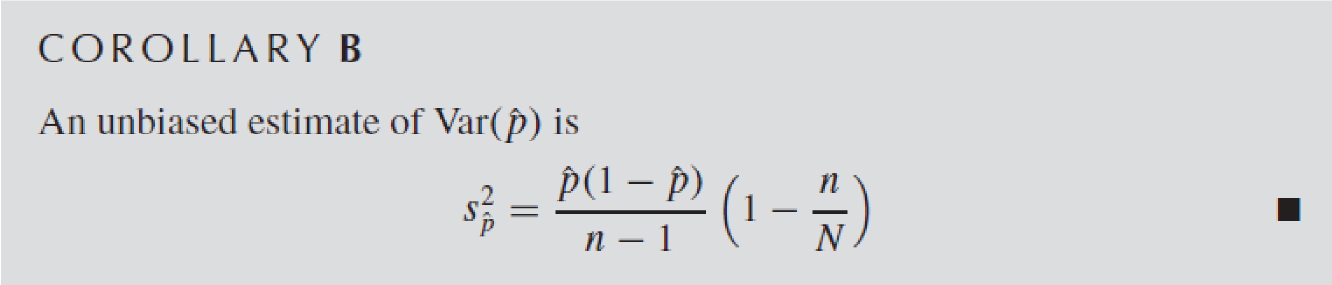
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Proof

Since

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

an unbiased estimate of $\mathrm{Var}(\overline{X})$ may be obtained by substituting in an unbiased estimate of σ^2. Algebra then yields the desired result. ■

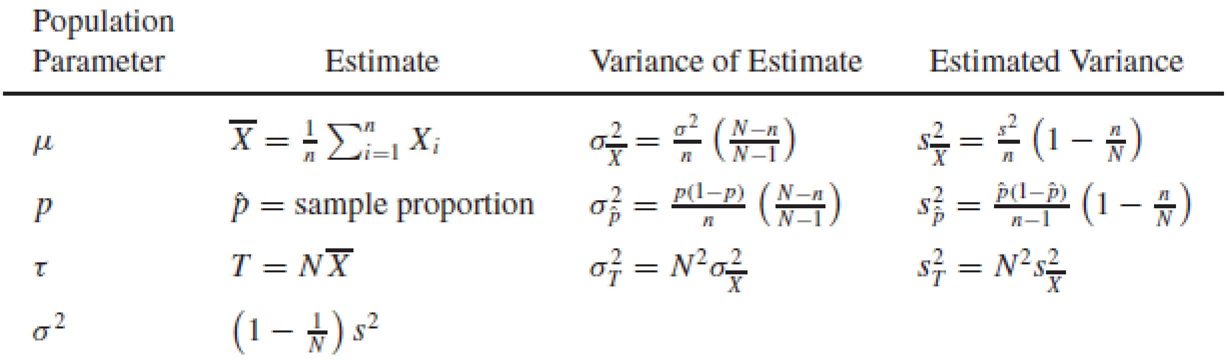


COROLLARY B

An unbiased estimate of $\mathrm{Var}(\hat{p})$ is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$$ ■

Memorize this table on page 214 (dichotomous case can be derived easily)



Population Parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1-\frac{n}{N}\right)$
p	$\hat{p}$ = sample proportion	$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$
τ	$T = N\overline{X}$	$\sigma_T^2 = N^2\sigma_{\overline{X}}^2$	$s_T^2 = N^2 s_{\overline{X}}^2$
σ^2	$\left(1-\frac{1}{N}\right)s^2$		

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Ex. A

Back to the 392 hospitals, suppose we take a random sample of 50 of them and find that $\bar{X} = 938.5$. Thus we also have $x^2 = 614.53$. We wish to know the estimated variance of $\bar{X}$, which is

$$s_{\bar{X}}^2 = \frac{s^2}{n}\big(1 - \frac{n}{N}\big) = 6592$$

Ex. B

From the estimated average number of discharges per hospital, we can calculate the estimated number of population discharges as $T = N\bar{X} = 368,831$. The estimated standard deviation of this estimate is $s_T = Ns_{\bar{X}} = 31,908$. This approximates our estimation error.

Ex. C

Back to proportions, we know the population proportion of fewer than 1000 discharges, $p = 0.654$. In the sample from A, 26 of the 50 hospitals had fewer than 1000 discharges. Thus $\hat{p} = 0.52$. The variance of our estimate is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\big(1 - \frac{n}{N}\big) = 0.0045$$

Normal Approximations

We know $\mathbb{E}(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$. So for large $n < N$ the central limit theorem that

$$P\left(\frac{\bar{X}_n - \mu}{\sigma\sqrt{n}}\right) \to \Phi(z)$$

as $n \to \infty$ where Φ is the cdf of the standard normal. Recall that $\sigma_{\bar{X}}$ converges to σ for n large but small relative to N. This lets us derives probabilistic bounds on the estimate of the mean and generate confidence intervals.

Ex. C

From the prior example C, we found for a sample of size 50 an estimated discharge proportion of $\hat{p} = 0.52$. Let the population proportion be 0.65 for a difference of 0.13. We wish to understand the probability of this difference occurring. We start by estimating the variance of our estimate as

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1 - \frac{n-1}{N-1}} = 0.064$$

Then we can calculate

$$\begin{aligned} P(|p - \hat{p}| > 0.13) &= 1 - P(|p - \hat{p}| \leq 0.13) \\ &= 1 - P\left(\frac{|p - \hat{p}|}{\sigma_{\hat{p}}} \leq \frac{0.13}{\sigma_{\hat{p}}}\right) \\ &= 2[1 - \Phi(2.03)] = 0.4 \end{aligned}$$