# Section 4, 9/23/19

## Variance of the Ratio Estimate

(page 221)

We are interested in our ratio estimate $R = \frac{\bar{Y}}{\bar{X}}$. As shown in class we have the expected value of $R$. But now we are interested in the variance of our estimate.

### Ratio Estimates

First, we need to define $Var(\bar{X}), Var(\bar{Y})$, and $Cov(\bar{X}, \bar{Y})$. We know the former. The **population covariance** is

$$Cov(X,Y) = \sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_i)(y_i - \mu_i)$$

For reference, the **population correlation coefficient** is $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. This lies between $-1$ and $1$ and a large positive value indicated a strong positive linear relation.

And so our the covariance of our estimated means is

$$Cov(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}(1 - \frac{n-1}{N-1})$$

Once we have these, we can appeal to approximation methods to approximate $Var(R)$ and $E(R)$.

## Variance Approximation

Consider the arbitrary function $Z = g(X,Y)$. Let $\mu$ indicate the point $(\mu_X, \mu_Y)$ upon which we center our first order multivariate Taylor expansion. Then

$Z = g(X,Y) \approx g(\mu) + (X - \mu_X)\frac{\partial g(\mu)}{\partial x} + (Y - \mu_Y)\frac{\partial g(\mu)}{\partial y}$

And

$$\begin{aligned} Var(Z) &\approx Var(X\frac{\partial g(\mu)}{\partial x} + Y\frac{\partial g(\mu)}{\partial y}) \\ &= \left(\frac{\partial g(\mu)}{\partial x}\right)^2 Var(X) + \left(\frac{\partial g(\mu)}{\partial y}\right)^2 Var(Y) + 2Cov(X,Y)\left(\frac{\partial g(\mu)}{\partial x}\right)\left(\frac{\partial g(\mu)}{\partial y}\right) \end{aligned}$$

Back to our ratio case, $Z = g(x,y) = y/x$. Calculating all of the partial derivatives, we find that

**THEOREM A**

With simple random sampling, the approximate variance of $R = \overline{Y}/\overline{X}$ is

$$\begin{aligned}\mathrm{Var}(R) &\approx \frac{1}{\mu_x^2}\left(r^2\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 - 2r\sigma_{\overline{XY}}\right) \\ &= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right) \quad \blacksquare\end{aligned}$$

## Estimated Population Parameters for Ratio Estimates

Clearly we can't calculate the population covariance from a sample. Instead, we calculate the estimated population covariance as

$$\begin{aligned}s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}\right)\end{aligned}$$

The estimated population correlation is $\hat{p} = \frac{s_{xy}}{s_x s_y}$. We can sub these values into the prior variance equation to obtain an estimate of the ratio variance, $s_R^2$. Additionally, we can generate an $\alpha$ confidence interval for the true raatio $r$, $R \pm z(\alpha/2)s_R$.

## Univariate Analysis with Ratio Estimates

If we know properties of the ratio AND **one** of the variables, we can make inferences of the values of the **other** variable. For instance, define the **ratio estimate** $\bar{Y}_R = \mu_X R = \frac{\mu_X}{\bar{X}}\bar{Y}$. Then Var($\bar{Y}_R$) = $\mu_X^2$Var(R). This is an alternative to the traditional $\bar{Y}$, not necessarily better. Thus

**COROLLARY A**

The approximate variance of the ratio estimate of $\mu_y$ is

$$\mathrm{Var}(\overline{Y}_R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y) \quad \blacksquare$$

Similarly, from Theorem B, we have another corollary.

**COROLLARY B**

The approximate bias of the ratio estimate of $\mu_y$ is

$$E(\overline{Y}_R) - \mu_Y \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x}(r\sigma_x^2 - \rho\sigma_x\sigma_y) \quad \blacksquare$$

Clearly there appears to be some bias to using this estimate. However, the variance may be better (lower). When is this better? We know that $Var(\bar{Y}) = \frac{\sigma_y^2}{n}$. Thus our ratio estimate has a smaller variance if $r^2\sigma_x^2 - 2r\rho\sigma_x\sigma_y < 0$. Once again, we don't actually know these population parameters. We can use their estimates to estimate this variance.

**COROLLARY C**

The variance of $\overline{Y}_R$ can be estimated by

$$s_{\overline{Y}_R}^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$$

and an approximate $100(1-\alpha)\%$ confidence interval for $\mu_y$ is $(\overline{Y}_R \pm z(\frac{\alpha}{2})s_{\overline{Y}_R})$. $\blacksquare$

## Example: Bias-variance tradeoff

Back to the hospital example with population parameters $\mu_x = 274.8, \mu_y = 814.6, r = 2.96, \sigma_x = 213.2, \sigma_y = 589.7, \rho = 0.91$. Say we measure a set of samples and that we are interested in an estimate of $\mu_y$. We can either appeal to $\bar{Y}$ or $\bar{Y}_R$. To start with, we consider $\bar{Y}_R$ which is slightly (on the order of 0.25%) biased. The variance of it as an estimate, however, is

$Var(\bar{Y}_R) \approx \cdots = \frac{68697.4}{n}$

And so $\sigma_{\bar{Y}_R} \approx \frac{262.1}{\sqrt{n}}$. With the finite correction factor and $n = 64$, $\sigma_{\bar{Y}_R} = 30.0$.

Meanwhile, $\bar{Y}$ is unbiased as an estimator, but its variance $\sigma_{\bar{Y}} = \cdots = 66.3$, more than twice the variance of the ratio estimate. In fact, in this case using the ratio estimate requires 80% less data to obtain the same variance as compared to $\bar{Y}$.