Ratio Estimation

Statistics 110

Summer 2006

Ratio Estimation

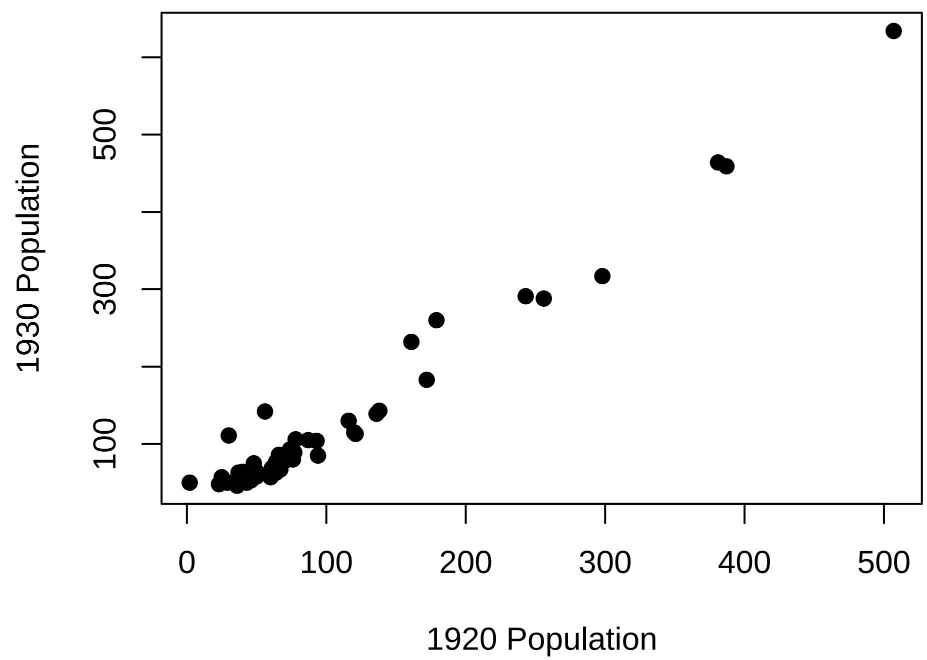
Another problem of interest involves two random variable X and Y, in particular the ratio of their two means (or equivalently, the ratio of their totals)

$$r = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

One example would be where y_i is population in year 1930 and x_i is the population in 1920 for city i (population in 1000's).

The plot shows the populations for 49 large cities for the two years in question.

r describes how much the population changes over the 10 year period.



A second example would have y_i be the annual soy bean production and x_i be the area in acres of farm i. Then r is the mean yield per acre in the population of farms.

One important thing to note is that

$$r \neq \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{x_i}$$

As before, we want to use a sample to estimate r. So suppose we sample n pairs (X_i, Y_i) and estimate r with

$$R = \frac{\bar{Y}}{\bar{X}}$$

Since R is a random quantity, is would be useful to determine $E[R]$ and $\mathrm{Var}(R)$. Since the ratio is a nonlinear function of $\bar{X}$ and $\bar{Y}$, getting exact values for these is difficult, but we can approximate them via the Taylor series methods discussed earlier.

Before doing that, we need two facts. The first is

Definition. *The population covariance of $\{x_i\}$ and $\{y_i\}$ is*

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

and the population correlation coefficient is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The second is

Theorem. *If* $(X_1, Y_1), \ldots, (X_n, Y_n)$, *then*

$$\mathrm{Cov}(\bar{X}, \bar{Y}) = \frac{\sigma_{XY}}{n}\left(1 - \frac{n-1}{N-1}\right)$$

Theorem. *If* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *is a SRS, then*

$$E[R] \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_X^2}(r\sigma_X^2 - \rho\sigma_X\sigma_Y)$$

and

$$\begin{aligned}\mathrm{Var}(R) &\approx \frac{1}{\mu_X^2}(r^2\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 - 2r\sigma_{\bar{X}\bar{Y}}) \\ &= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_X^2}(r^2\sigma_X^2 + \sigma_Y^2 - 2r\rho\sigma_X\sigma_Y)\end{aligned}$$

The proof of this an application of the general results for ratio of two random variables (Example C section 4.6).

A couple of comments on these formulas.

- First the bias and variance decrease as n increases
- The bias and variance are large if μ_X are small. This isn't too surprising, since small changes in x can lead to big changes in $\frac{1}{x}$ if x is small.
- The more variable X and Y are, the bigger the variance of R.
- The bias and variance decrease if ρ is positive

As before, we need to estimate the variance of R since none of the population variances and covariances are usually known.

The usual estimate of the population covariance is

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

leading to the estimate of the correlation of

$$\hat{\rho} = \frac{S_{XY}}{S_X S_Y}$$

Combining these, an estimate of $\mathrm{Var}(R)$ is

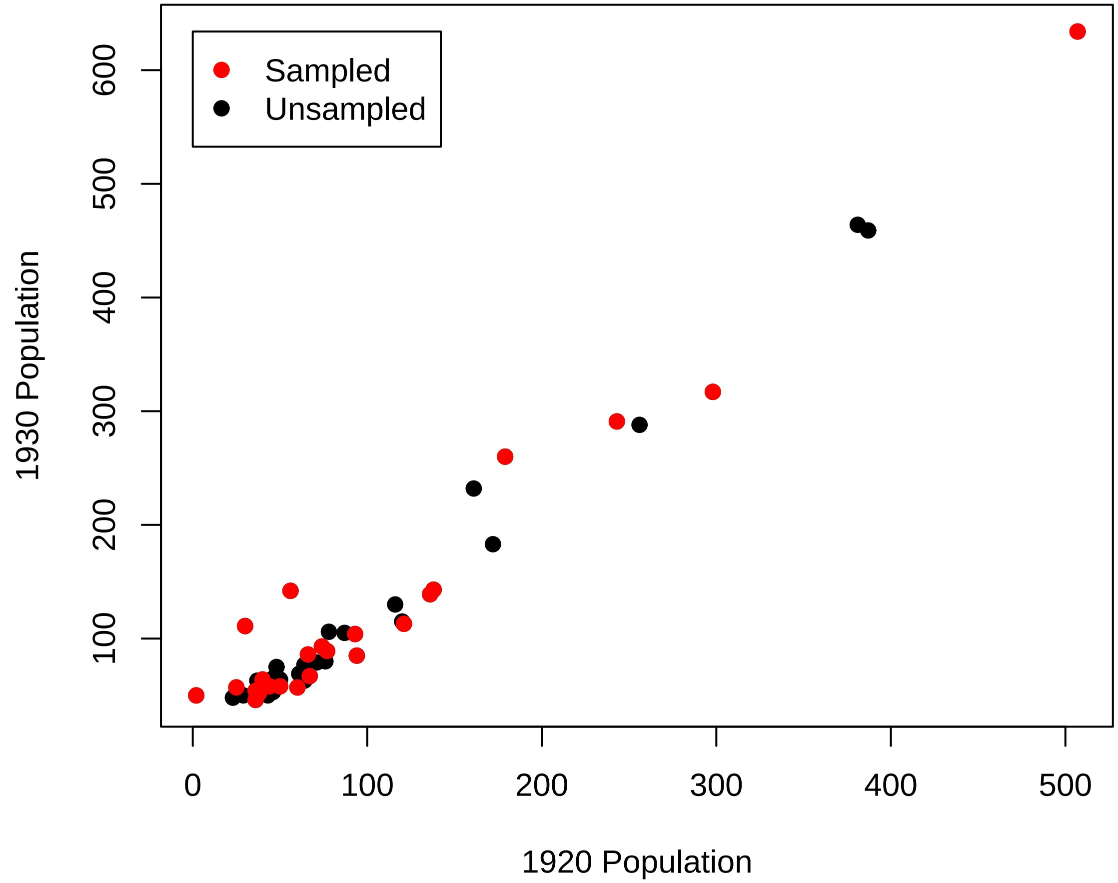
$$s_R^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\bar{X}^2}(R^2 S_X^2 + S_Y^2 - 2R S_{XY})$$

Finally, a $100(1-\alpha)\%$ CI for r is

$$R \pm z(\alpha/2) s_R$$

Example: Population Growth

$$\mu_X = 103.14$$
$$\mu_Y = 127.80$$
$$r = 1.239$$
$$\sigma_X = 103.33$$
$$\sigma_Y = 121.86$$
$$\rho = 0.982$$



The sample estimates of these quantities (based on a sample of $n = 25$ cities) are

$$\bar{X} = 102.0 \qquad \bar{Y} = 129.2 \qquad R = 1.267$$
$$S_X = 109.30 \qquad S_Y = 129.11 \qquad \hat{\rho} = 0.982$$

$$
\begin{aligned}
s_R^2 &= \frac{1}{25}\left(1-\frac{24}{48}\right)\frac{1}{102.0^2}\times \\
&\qquad (1.267^2 109.3^2 + 129.11^2 - 2\times 1.267\times 0.982\times 109.3\times 129.11) \\
&= 0.001972
\end{aligned}
$$

So a 95% CI for r is

$$1.267 + 1.96\times\sqrt{0.001972} = 1.267 \pm 0.087 = (1.180, 1.354)$$

Note that the extremely high correlation between X and Y allows us to estimate the ratio very precisely.

We can use this property to estimate other quantities more precisely.

Conceptually is similar to using to doing prediction with $E[Y|X = x]$ instead $E[Y]$. The dependency allows us to make more precise statements.

In particular we can use it to get better estimates of μ_Y, assuming that we know μ_X. While initially this idea might seem a bit surprising, in some situations it can work.

For example take the soy bean example where y_i is the soy bean yield and x_i is the area of farm i. While y_i might take some work to get, x_i is often easy to get through public records or may have already been collected.

The ratio estimate of μ_Y is

$$\bar{Y}_R = \frac{\mu_X}{\bar{X}}\bar{Y} = \mu_X R$$

Lets suppose that $\rho > 0$ and $\bar{X} < \mu_X$. In this case, it is likely that $\bar{Y}$ will also be $< \mu_Y$, so this estimator will bump things up, hopefully closer to μ_Y.

Theorem. *If* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *is a SRS, then*

$$E[\bar{Y}_R] \approx \mu_Y + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_X}(r\sigma_X^2 - \rho\sigma_X\sigma_Y)$$

and

$$\mathrm{Var}(\bar{Y}_R) = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)(r^2\sigma_X^2 + \sigma_Y^2 - 2r\rho\sigma_X\sigma_Y)$$

The ratio estimator is more precise if $\mathrm{Var}(\bar{Y}) > \mathrm{Var}(\bar{Y}_R)$ or equivalently if

$$r^2\sigma_X^2 - 2r\rho\sigma_X\sigma_Y < 0$$

which, if given $r > 0$

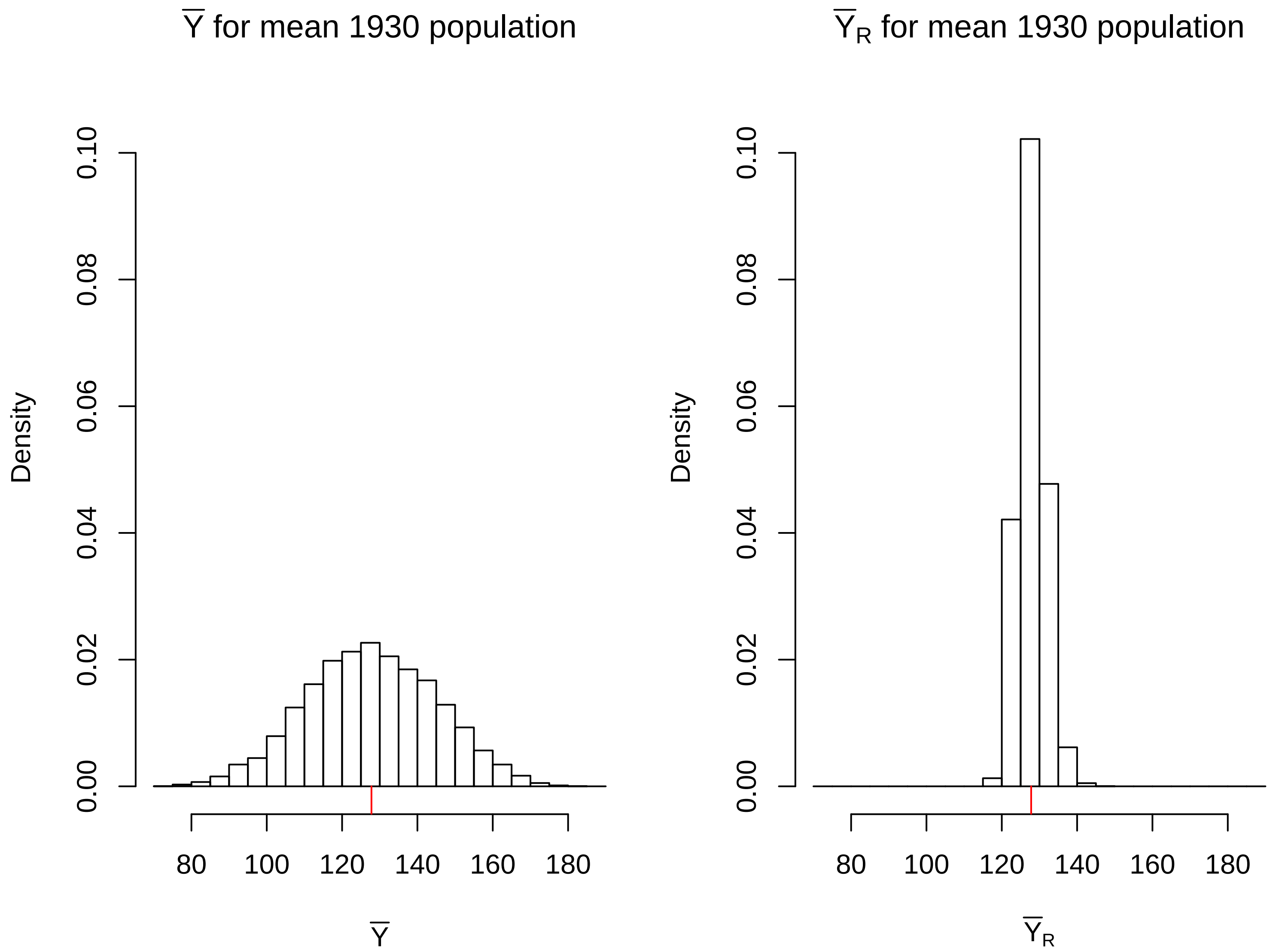
$$2\rho\sigma_Y > r\sigma_X$$

This is equivalent to

$$\rho > \frac{1}{2}\frac{C_X}{C_Y}$$

where $C_X = \frac{\sigma_X}{\mu_X}$ is the **coefficient of variation**

Note the the coefficient of variation is a relative standard deviation and is dimensionless (e.g. it is the same whether you measure in pounds or kilograms)


$\overline{Y}$ for mean 1930 population
$\overline{Y}_R$ for mean 1930 population
Density
0.00
0.02
0.04
0.06
0.08
0.10
80
100
120
140
160
180
$\overline{Y}$
$\overline{Y}_R$

Based on this ratio estimator, there is a second CI for μ_Y of

$$\bar{Y}_R \pm z(\alpha/2) s_{\bar{Y}_R}$$

where

$$s^2_{\bar{Y}_R} = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(R^2 S^2_X + S^2_Y - 2RS_{XY}\right)$$

For the population example,

$$\begin{array}{lll} \bar{X} = 102.00 & \bar{Y} = 129.20 & R = 1.267 \\ S_X = 109.30 & S_Y = 129.11 & \hat{\rho} = 0.982 \end{array}$$

These give (with $\mu_X = 103.14, \mu_Y = 127.80$)

$$\bar{Y}_R = \frac{103.14}{102.00} \times 129.20 = 130.64$$

$$s_{\bar{Y}} = \frac{129.11}{\sqrt{25}}\sqrt{1 - \frac{25}{49}} = 18.07$$

$$
\begin{aligned}
s^2_{\bar{Y}_R} &= \frac{1}{25}\left(1 - \frac{24}{48}\right)(1.267^2 109.30^2 + 129.11^2 - 2 \times 1.267 \times 13857.71) \\
&= 20.52
\end{aligned}
$$

95% CI based on $\bar{Y}$:

$$129.20 \pm 1.96 \times 18.07 = 129.20 \pm 35.42$$

95% CI based on $\bar{Y}_R$:

$$130.64 \pm 1.96\sqrt{20.52} = 130.64 \pm 8.88$$

So for this example, it ends up the ratio estimate is slightly worse (though usually we can never know this). However we can see the big advantage with this estimator, a much narrower confidence interval.