STA304

Neil Montgomery
2016-03-10

ratio, regression, difference estimation

overview

In most actual samples, more than one measurement is available on each unit sampled.

Let's consider the case of two measurements, generically called y and x.

The population is now $\{(y_1, x_1), \ldots, (y_N, x_N)\}$ and the sample is $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ (with the usual abuse of notation.)

Sometimes the population parameter of interest is still the total or mean of the y. Let's call these numbers τ_y and μ_y. We'll also make use of τ_x and μ_x.

If the y_i and the x_i in the population are *correlated*, it is possible to get improved estimates of τ_y and μ_y.

Also, sometimes *the actual quantity of interest* is the population ratio itself $R = \tau_y/\tau_x = \mu_y/\mu_x$.

a note, and some examples

Note: stratified sampling was a way to use additional information about the units *to design the sampling itself in a different way that can be better*. This part of the course concerns using additional information *in the estimation procedures*.

Examples:

- Consumer Price Index (CPI): a "basket" of consumer goods has its prices measured each month. The y variable is this month's price and the x variable is the previous month's price. The ratio of the total basket price is the proportion of increase/decrease in prices.
- "The average amount of screen time per child in households." Unit: household. y: total screen time. x: number of children.
- (Section 6.2): Total amount of sugar in a truckload of oranges. y is sugar per orange and x is the weight of an orange.

estimating the population ratio

I'll use the salary data (a sample from a population of $N = 750$) from question 6.10 in the textbook as an example to motivate the concepts.

Teacher	Past	Present
1	30400	31500
2	31700	32600
3	32792	33920
4	34956	36400
5	31355	32020
6	30108	31308
7	32891	34100
8	30216	31320
9	30416	31420
10	30397	31600
11	33152	34560
12	31436	32750
13	34192	35800
14	32006	33300
15	32311	33920

estimating the population ratio

x is the *Past* salary and y is the *Present*. The goal is to estimate the rate of salary increase for the population. This is the ratio:

$$R = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$$

The obvious way to estimate R is to use the ratio of the sample totals or sample means:

$$\hat{R} = \frac{\hat{\tau}_y}{\hat{\tau}_x} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{\mu}_y}{\hat{\mu}_x}$$

which the book mystifyingly calls r rather than $\hat{R}$, so I'll respect that notation.

The basic challenge is that the denominator is random.

variance of r - I

It turns out r is slightly biased for R, but mainly with very small sample sizes. **So we'll go with the assumption that the sample size is not very small**, giving $E(r) \approx R$.

The end goal is to get $V(r) \approx E\left[(r-R)^2\right]$. Start with:

$$r - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} \approx \frac{\bar{y} - R\bar{x}}{\mu_x}$$

Then:

$$V(r) \approx E\left[\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\right)^2\right] = \frac{1}{\mu_x^2} E\left[\left(\bar{y} - R\bar{x}\right)^2\right]$$

variance of r - II

$\left(\bar{y} - R\bar{x}\right)$ is actually a pretty simple object. Let (NOTE: changed this to capital R_i.):

$$R_i = y_i - Rx_i$$

Then $\{R_1, R_2, \ldots, R_n\}$ is a SRS and $\overline{R} = \left(\bar{y} - R\bar{x}\right)$ is a sample average with mean 0 and variance that can be copied from the basic SRS theory:

$$V(\overline{R}) = \frac{s_R^2}{n}\left(1 - \frac{n}{N}\right)$$

with

$$s_R^2 = \frac{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2}{n-1}$$

(This isn't practically useful yet…we don't know R!)

variance of r - III

Putting it all together and we get:

$$V(r) \approx \frac{1}{\mu_x^2}\frac{s_R^2}{n}\left(1 - \frac{n}{N}\right)$$

There are a few unknowns in there. R is never known, and μ_x may or may not be known. The conclusion is:

$$\hat{V}(r) \approx \frac{1}{\mu_x^2}\frac{s_r^2}{n}\left(1 - \frac{n}{N}\right)$$

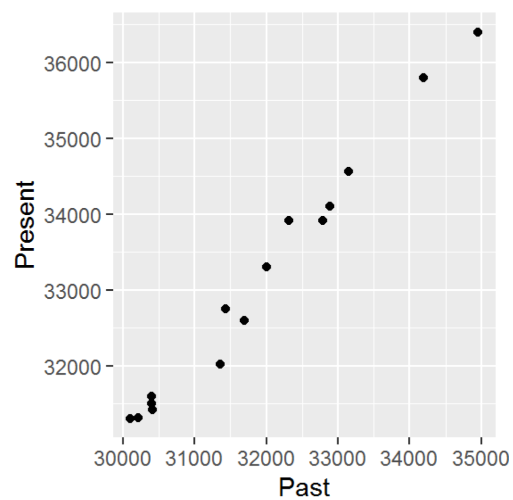
where s_r^2 is s_R^2 with r used in place of R and $\overline{r}$ used in place of $\overline{R}$, and use $\overline{x}$ if μ_x is unknown.

salary example

Let's estimate the population salary rate of increase and place the usual bound on the error of estimation $\left(B = 2\sqrt{\hat{V}(r)}\right)$. Here are the details.

x_bar	y_bar	r	s^2_r	B
31888.53	33101.2	1.038028	51086.04	0.0036234

(See the file `6_10.xlsx` on github with these slides for a spreadsheet solution to the problem.)



using x and ratio technique for estimation of τ_y

Ratio estimation might make estimating τ_y *better*, or it might even be the only feasible way to do it at all, if N isn't known and is too difficult to determine, but τ_x *is* known.

An example of the latter case the total amount of sugar τ_y in a truck of oranges example. (Section 6.2 and Example 6.2). Here τ_x is the total weight of the oranges.

In general the usual estimator of τ_y is $N\overline{y}$, but always requires N. Another option is:

$$\hat{\tau}_y = r\tau_x$$

If the y_i and x_i are correlated then this is a better estimator, and if N is unknown this can be the only feasible estimator.

estimated variance of $\hat{\tau}_y$

Option 1:

$$\hat{V}(\hat{\tau}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n}$$

Option 2 (N unknown)::

$$\hat{V}(\hat{\tau}_y) = (\tau_x^2) \frac{1}{\mu_x^2} \frac{s_r^2}{n} \left(1 - \frac{n}{N}\right)$$

with $\overline{x}$ for μ_x if required.

sugar in oranges example

(Raw data not available). A sample of $n = 10$ oranges was taken and the sugar content y_i and weight x_i measured. The total weight τ_x of all oranges was 1800 pounds.

The estimate of the total sugar in pounds is:

$$\hat{\tau}_y = r\tau_x = \frac{\sum y_i}{\sum x_i}\tau_x = \frac{0.246}{4.35}(1800) = 101.79$$

The bound $B = 2\sqrt{\hat{V}(\hat{\tau}_y)}$ requires also to know that $s^2 = (0.0024)^2$. Plugging all into the formula gives a bound of 6.3