STA304

Neil Montgomery

2016-03-17

test 2 information

Midterm details

The second test is on March 24. TA office hours will probably be the same as last time (and will be announced on the course website.)

The test will begin at 3:30 and will be designed to be 1 hour in length, but you will have until 5:00 to complete it if you wish.

The specific topics covered will be:

- SRS for proportions and counts
- Stratified random sampling
- Required reading topics added 2016-03-16 (required reading discussion on Monday)

Stratified example question

These sorts of questions are tedious, time consuming, and error prone, when done entirely by hand. So I'll probably try something like this.

Here's my spreadsheet solution to question 5.1 (a sample size question):

				Sum
N_i	112	68	39	219
c_i	9	25	36	
sigma^2_i	2.25	3.24	3.24	
W_i	0.511416	0.310502	0.178082	
W_i*sigma/sqrt_c_i	56	24.48	11.7	92.18
a_1	0.607507	0.265567	0.126926	
W_i^2sigma^2_i/a_i	0.968677	1.176251	0.809537	2.954465
W*sigma^2_i	1.150685	1.006027	0.576986	2.733699
			n	26.26596

Stratified example question

On a test I might produce a similar table but with some of the calculated entries obscured, like this (ALTHOUGH THE W_i ROW IS WRONG):

				Sum
N_i	112	68	39	219
c_i	9	25	36	
sigma^2_i	2.25	3.24	3.24	
W_i	0.511416	[redacted]	[redacted]	
W_i*sigma/sqrt_c_i	[redacted]	[redacted]	11.7	92.18
a_1	[redacted]	0.265567	[redacted]	
W_i^2sigma^2_i/a_i	0.968677	[redacted]	0.809537	[redacted]
W*sigma^2_i	1.150685	1.006027	0.576986	2.733699
			n	[redacted]

(Note: CORRECTED spreadsheet available in this lecture's github: `stratified.xlsx`)

back to ratio, regression, difference estimation

recap

If one has a sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ the following may be of interest:

1. To estimate the population ratio $R = \tau_y \big/ \tau_x$ using $\hat{R} = r = \bar{y} \big/ \bar{x}$.
2. To enable estimation of τ_y when N is unknown.
3. To enable improved estimation of τ_y or μ_y when y and x are correlated.

Formula summary:

r	$\hat{\tau}_y$	$\hat{\mu}_y$	$\hat{V}(r)$
$\bar{y} \big/ \bar{x}$	$r\tau_x$	$r\mu_x$	$s_r^2(1 - n/N)$

The error bounds are based on this new object:

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n - 1}$$

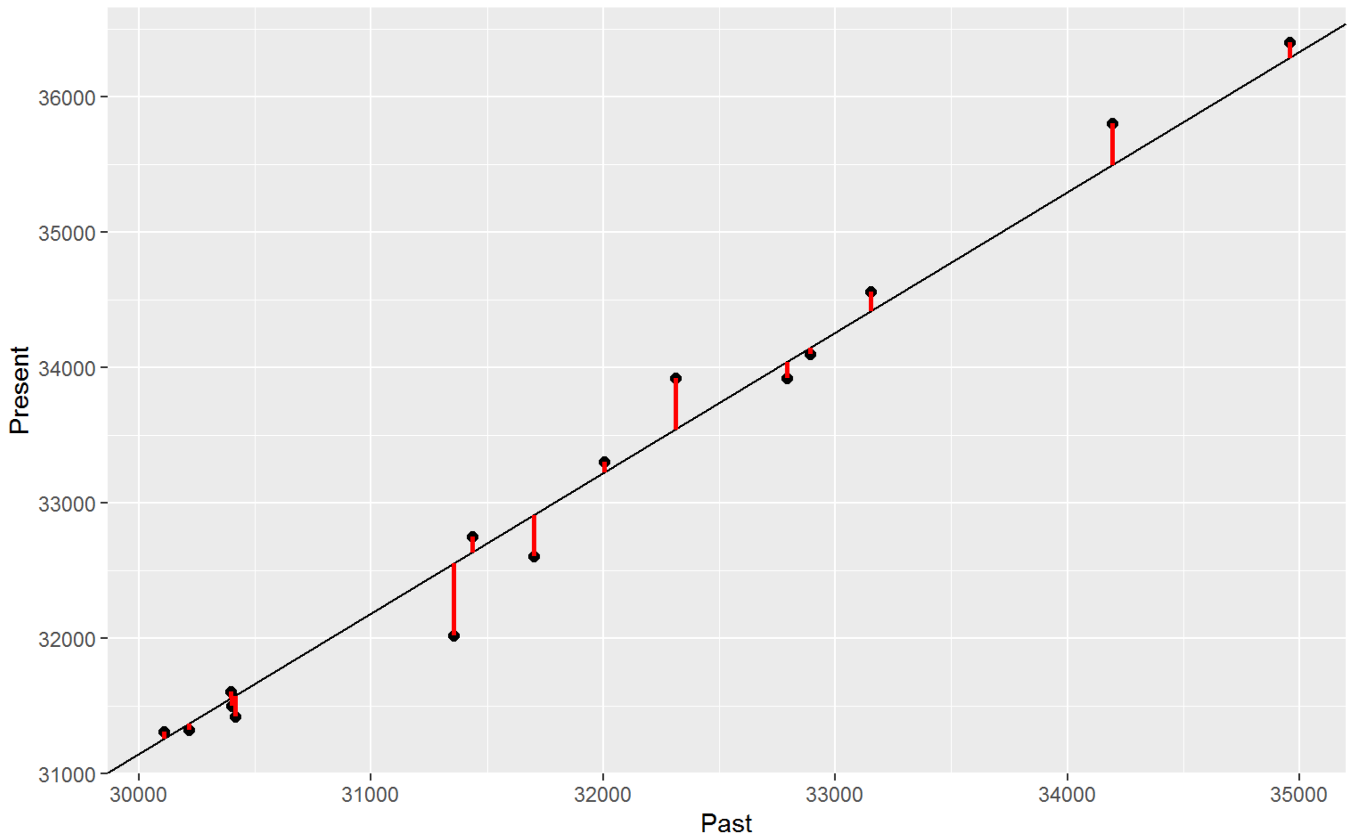
$r_i = y_i - rx_i$ is a "residual"

Reconsider the teacher salary example with $r = 1.0380283$. First 5 lines shown:

```
##   Teacher  Past Present Predicted = r*Past        r_i
## 1       1 30400   31500           31556.06  -56.06028
## 2       2 31700   32600           32905.50 -305.49706
## 3       3 32792   33920           34039.02 -119.02397
## 4       4 34956   36400           36285.32  114.68280
## 5       5 31355   32020           32547.38 -527.37730
```

Think of rx_i as the "predicted" value for y_i. Then $y_i - rx_i$ is in some sense a prediction error, or "residual". And s_r^2 is just the "average" of the squared residuals.

"residuals" plotted



from τ_y to μ_y

It was possible to estimate τ_y without knowing N (or τ_x). If they are known then it is possible to estimate μ_y using ratio techniques more accurately than with SRS observing y_i alone.

Last time we had explicit formulae for $\hat{\tau}_y$ and $\hat{V}(\hat{\tau}_y)$. They are easily adjusted to get:

$$\hat{\mu}_y = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}\mu_x = \frac{\hat{\tau}_y}{N}$$
$$\hat{V}(\hat{\mu}_y) = \left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}$$

such an example of improved estimation

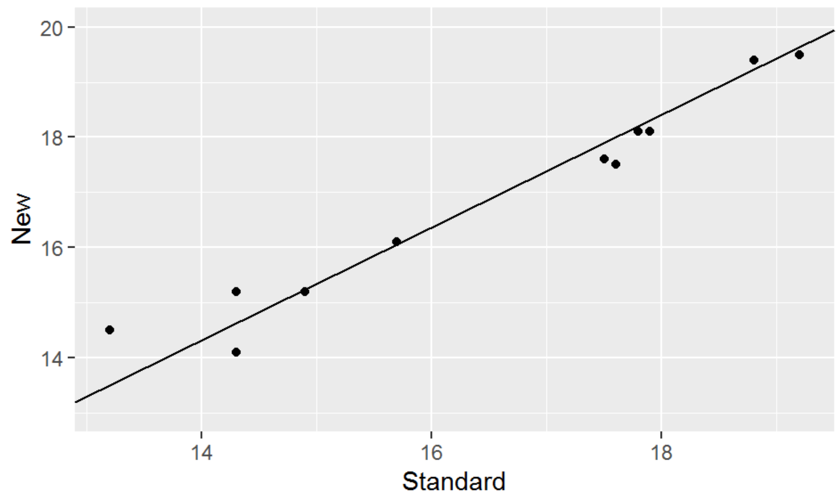
We have estimated a ratio and estimated τ_y using ratio techniques because there was no other option. In this example we'll estimate μ_y using ratio techniques simply to take advantage of the information contained in the x variable.

Consider question 6.6. "Rats doing mazes while on drugs". They have $N = 763$ who completed the maze on the standard drug in an average of $\mu_x = 17.2$ seconds.

A random sample of 11 rats are given a new drug. Their old times x_i were known from before and they complete the maze while on the new drug in time y_i .

The task is to estimate the average maze time μ_y for the new drug.

these are your rats on drugs



The estimated ratio is $r = 1.0226269$. The mean estimate using the ratio technique is:

$$\hat{\mu}_y = r\mu_x = 1.0226269 \cdot 17.2 = 17.5891834$$

bounding the estimation error

It turns out $s_r^2 = 0.2049424$. So the estimated variance of $\hat{\mu}_y$ is:

$$\hat{V}(\hat{\mu}_y) = \left(1 - \frac{11}{763}\right)\frac{0.2049424}{11} = 0.0183625$$

So the usual bound on the error of estimation would be $B = 2\sqrt{\hat{V}(\hat{\mu}_y)} = 0.2710168$

.

Equivalently a 95% confidence interval for μ_y is 17.5891834 ± 0.2710168.

how much better than SRS on y_i alone?

It would be perfectly correct to ignore the x_i and μ_x that are given and simply estimate μ_y with $\bar{y}$ using regular SRS theory.

If we did that, we would get $\bar{y} = 16.8454546$. The error bound would depend now on the old:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 3.6727272$$

and the SRS bound on the error of estimation would be based on::

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = 0.3290708$$

giving us a B_{SRS} of $1.1472938\ldots$substantially higher than 0.2710168.

so, when does the ratio technique improve $\hat{\mu}_y$?

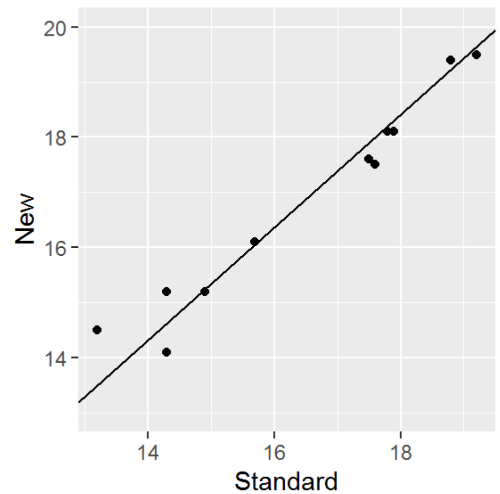
The math is complicated (see p. 177 and 6.8 for the gory details).

In practice it tends to work well in repeated surveys where the numbers are being updated, i.e. the y_i are new versions of the x_i.

These is a more technical summary of when an improvement is likely:

- when the relationship between y and x is linear (not curved) and "through the origin" (y's are directly proportional to the x's).
- when the correlation coefficient between the y_i and x_i is high enough, say $\hat{\rho} > 0.5$.

evaluating the rats data



Linear and "through the origin".

Also:

$$\hat{\rho} = 0.9787687$$

So the big improvement over SRS is not surprising.

sample size requirements for ratio techniques

Nothing more than a slight adjustment to the regular SRS formula.

In fact the sample size requirement to estimate the mean μ_y using ratio techniques to within a bound B (with 95% confidence) is unchanged at:

$$n = \frac{N\sigma^2}{(N-1)\frac{B^2}{4} + \sigma^2}$$

To estimate a ratio R to within B_R simply note that $\hat{V}(r) = \frac{1}{\mu_x^2}\hat{V}(\hat{\mu}_y)$, so just use $B = B_R\mu_x$ in the above formula.

IMPORTANT: note that here σ^2 is now the population variance of the ratios y_i/x_i, so…

possibly improved formula??

$$n = \frac{N\sigma_R^2}{(N-1)\frac{B^2}{4} + \sigma_R^2}$$

As usual σ_R^2 is unknown, so some prior information ought to be used - such as s_r^2 from either a prior survey or a pilot sample. (The old range/4 guess is probably a bad idea this time—why?)

sample size for ratio example

Suppose with the rats we wanted to estimate the ratio R with a bound B of 0.01. We can use $s_r^2 = 0.2049424$ as a guess for σ_R^2. The sample size required is:

$$n = \frac{763 \cdot 0.2049424}{(762)\frac{0.029584}{4} + 0.2049424} = 26.7726796$$