CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: A COMPARISON OF TWO RATIO ESTIMATORS AND A REGRESSION ESTIMATOR FOR A FINITE POPULATION MEAN

AUTHOR: Shashank Nath

SUPERVISOR: Patrick Farrell

DATE: May 6, 2019

Introduction:

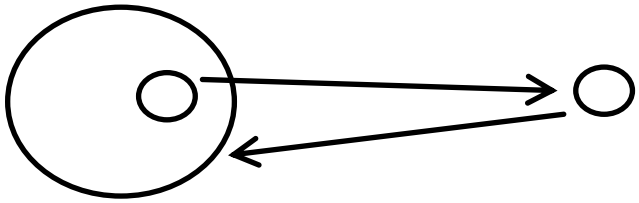
Often in practical and contemporary society, the only way to obtain data is through the use of surveys. Surveys are very important, as within the ever-expanding technological society obtaining data, while simpler with the use of data mining and machine learning, often crosses the threshold of legality and ethics. Look no further than online targeted ads recording an individual's search history by using data from their browser to make marketing more personalized. For websites where purchases will be made, such as Amazon, this may not seem like a problem; however, for the majority of situations where the individual is not making a purchase, they are left wondering how the algorithm knows what to sell them, even though it has no way to be privy to the information. Obtaining data in this way is very brilliant, yet, ethically, the individual never agreed to have themselves documented. This is where the importance of **surveys** come in. They often (not always) can avoid the ethical dilemma, as the individual is clearly notified that they are being documented and data collection is kept to a very rudimentary process. When analyzing data from the surveys, say that one ends up making the wrong choice, perhaps an individual makes a choice on an **estimator** when there is a superior one that goes unnoticed. How much worse will the analysis of the data be using the chosen estimator against the deemed superior estimator?

Background:

Definition: “A **survey** is the process of measuring a characteristic of some of the members of a population with the purpose of making quantitative generalizations about the population as a whole”.

Definition: “A **sample** is the members of the population whose characteristics are being measured. The sample is, in general, a subset of the population of interest.”

The basic use of surveys comes down to using information from a sample to draw inference about the population from which the sample is collected.



Population Sample

A survey usually aims to estimate certain population attributes, such as population means, totals and proportions. In accordance with the prior listed, it is the analysis of **bias** and **variability.**

Definition for **Sampling Bias**: “Is the average sampling error if the survey under consideration is repeated many times independently under the same condition.”

Definition for **Sampling Variability**: “The extent to which the sampling error varies around its average value if the survey under consideration were repeated many times independently under the same conditions.”

To assess the sampling bias and variability, a **probability sample**, which is defined as “the selection of units from a population according to known probabilities,” must be selected. One common probability sample used is a **simple random sample** (**SRS**).

Definition of **SRS**: "Draw items from a population *one at a time without replacement*, so at every stage, each remaining item from the population has the same probability of being chosen. "

Looking for the SRS as an estimator, one would proceed as follows:

First, one wishes to estimate the mean, μ, of a finite population given by

$$\mu = \sum_{j=1}^{N} \frac{u_j}{N}$$

Now, using the **sample mean**, $\bar{y}$, can be written as

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

The **population variance**, σ^2, can be used to quantify the variability of the u_j values in the population

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2$$

Suppose that an SRS of size n is selected. If $\bar{Y}$ is a random variable for the mean of a simple random sample given by

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{Y_1 + \cdots + Y_n}{n}$$

Since $E(\bar{Y}) = \mu$, it must be the case that $E(\bar{Y} - \mu) = 0$, implying that $\bar{Y}$ is an unbiased estimator for μ.

Likewise, the sampling variability in $\bar{Y}$, $V(\bar{Y} - \mu) = V(\bar{Y})$ (Since μ is a constant, then $V(\mu)$ is non-random, and equal to 0), is given by

$$V(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) + \left(\frac{1}{n}\right)\left(\frac{1}{n}\right)\sum_{i=1}^{n}\sum_{i=1}^{n} Cov(Y_i, Y_k) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Now, the application of the **Central Limit Theorem (CLT)**, which states that:

If n large enough, then $\bar{Y}$ has an approximate normal distribution with

$$E(\bar{Y}) = \mu \qquad \text{and} \qquad V(\bar{Y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

allows for the construction of a (1-α)100% **Confidence Interval (CI).**

Definition of **Confidence Interval (CI)**: "A range of values so defined that there is a specified probability that the value of a parameter lies within it." (Baird, S., n.d.)

Thus, an approximate (1-α)100% confidence interval for μ is given by

$$\bar{y} \pm z_{\alpha/2}\sqrt{\mathrm{V}(\bar{Y})} \qquad \text{or} \qquad \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

where the $\alpha/2$ upper critical value of the standard normal distribution is represented by $z_{\alpha/2}$, and $\sigma = \sqrt{\mathrm{V}(\bar{Y})}$ is the standard deviation or standard error of $\bar{Y}$ with the **point estimate** (generally known as the pivot) being $\bar{y}$.

Typically, σ^2 is unknown, thus, an analogous estimate for σ^2 with the sample variance, s^2, given by:

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

In turn,

$$\mathrm{V}(\bar{Y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

is estimated by,

$$\widehat{\mathrm{V}}(\bar{Y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

Therefore, an approximate (1-α)100% Confidence Interval for μ is given by

$$\bar{y} \pm z_{\alpha/2}\sqrt{\hat{\mathrm{V}}(\bar{Y})} \qquad \text{or} \qquad \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$$

The quantity $\left(1-\frac{n}{N}\right)$ is known as the **finite population correction factor (fpc)**, which may be ignored when the ratio of the sample to the population $\left(\frac{n}{N}\right) < 0.05$.

This is the computation of a basic simple random sample (SRS) and will serve as a foundation for the upcoming calculations and analysis of estimations for **ratio, regression,** and, **mean of ratio,** all of which serve as the focal point of the paper.

Premise:

Proceeding with the knowledge of how to compute an SRS, comes the ability to judge more estimators and how they fair when compared to one another. In particular, the comparison of the aforementioned estimators of **ratio, regression,** and, **mean of ratio**, and what would occur if one were chosen in spite of another being a superior estimator**.** The motivation for this paper came about with the computation of the following problem:

A useful theoretical model to compare various ratio estimators is given by

$$Y_i = \beta x_i + D_i$$

where D_i represents the deviation from the line. For a given value of x, we assume that the values of Y are scattered about the line so that the expected value and the variance of the D_i's are

$$E(D_i|x_i) = 0 \qquad V(D_i|x_i) = \sigma^2 x_i^{2a}$$

For example, if $a = 1$, then the scatter about the line increases as x increases. Consider a general estimator for β of the form $\widehat{\beta} = \sum_{i=1}^{n} c_i Y_i$ where c_i depends on x_i.

a) Using the model given above, find a condition on the c_i so that $\hat{\beta}$ is an unbiased estimator for β, given the observed x_i.

Solution:

$$E(\hat{\beta}|xi) = E(\sum_{i=1}^{n} c_i Y_i \,|xi)$$

$$= \sum_{i=1}^{n} c_i \beta x_i + \sum_{i=1}^{n} E(D_i|x_i)$$

$$= \beta \sum_{i=1}^{n} c_i x_i$$

Note: $E(D_i|\, x_i) = 0 \rightarrow \sum_{i=1}^{n} E(D_i|x_i) = 0$. Thus, $\widehat{\beta}$ is conditional unbiased estimator for β if $\sum_{i=1}^{n} c_i x_i = 1$.

b) Determine an expression for the variance of $\hat{\beta}$ as a function of a, conditional on the observed x_i's (Assume that all of the x_i's are greater than zero).

Solution:

$$V(\widehat{\beta}\,|xi) = V(\sum_{i=1}^{n} c_i Y_i|x_i)$$

$$= \sum_{i=1}^{n} c_i^2 V(Y_i|X_i)$$

$$= \sum_{i=1}^{n} c_i^2 V(\beta x_i + D_i|X_i)$$

$$= \sum_{i=1}^{n} c_i^2 V(D_i|X_i)$$

$$= \sum_{i=1}^{n} c_i^2 \sigma^2 x_i^{2a}$$

c) Using your results in (a) and (b), for a given value of a, find the unbiased estimator in the above class with minimum conditional variance.

Solution:

Need to find the minimum conditional variance. Thus, we need to minimize

$V(\widehat{\beta}\,|xi) = \sum_{i=1}^{n} c_i^2 \sigma^2 x_i^{2a}$ with respect to $\sum_{i=1}^{n} c_i x_i = 1$

Using Lagrange,

$$f(c_i\,, \lambda) = \sum_{i=1}^{n} c_i^2 \sigma^2 x_i^{2a} + \lambda\left(\sum_{i=1}^{n} c_i x_i - 1\right)$$

Note: $\sum_{i=1}^{n} c_i x_i - 1 = 0$

We have,

$$\frac{\partial f(c_i, \lambda)}{\partial c_i} = 2c_i \sigma^2 x_i^{2a} + \lambda x_i$$

Setting $\frac{\partial f(c_i,\lambda)}{\partial c_i} = 0$, and solving for c_i gets us: $c_i^* = -\frac{\lambda x_i}{2\sigma^2 x_i^{2a}}$

Solving this we get,

$\sum_{i=1}^{n} c_i^* x_i = \sum_{i=1}^{n} -\frac{\lambda x_i}{2\sigma^2 x_i^{2a}} x_i = -\frac{\lambda}{2\sigma^2}\sum_{i=1}^{n} x_i^{2-2a} = 1$. **Note**: $\sum_{i=1}^{n} c_i x_i = 1$

$\lambda = -\frac{2\sigma^2}{\sum_{i=1}^{n} x_i^{2-2a}}$ and $c_i^* = -\frac{\lambda x_i}{2\sigma^2 x_i^{2a}} = \frac{x_i^{1-2a}}{\sum_{i=1}^{n} x_i^{2-2a}}$

$\hat{\beta} = \sum_{i=1}^{n} c_i^* Y_i = \frac{(\sum_{i=1}^{n} x_i^{1-2a})Y_i}{\sum_{i=1}^{n} x_i^{2-2a}}$ with the smallest conditional variance.

d) For $a = 0$, which estimator has the smallest conditional variance? What about for $a = 1/2$? For $a = 1$?

Solution:

The solution to this is what is used as a foundation for this paper. It is what is assessed when comparing between the estimators.

$\rightarrow a = 0$, we have $\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i^{1-2(0)}\right)Y_i}{\sum_{i=1}^{n} x_i^{2-2(0)}} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$ (**Regression**)

$\rightarrow a = 1/2$, we have $\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i^{1-2(1/2)}\right)Y_i}{\sum_{i=1}^{n} x_i^{2-2(1/2)}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}$ (**Ratio of Means**)

$\rightarrow a = 1$, we have $\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i^{1-2(1)}\right)Y_i}{\sum_{i=1}^{n} x_i^{2-2(1)}} = \frac{\sum_{i=1}^{n} Y_i x_i^{-1}}{\sum_{i=1}^{n} 1} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{x_i}$ (**Mean of Ratios)**

e) Comment on the consequences of this analysis for ratio and regression estimation of μ_y.

Solution:

- If conditional variance of Y given the observed x_i's is **constant** $\rightarrow$ **Regression**
- If conditional variance of Y given the observed x_i's is **proportional** $\rightarrow$ **Ratio (means of ratio, or ratio of means)**

Now, lets examine a scenario where an individual is presented with a set of data, and must decide as to which estimator they wish to use. Without loss of generality, assume that they use the **regression** estimator; however, what if the superior estimator here were to be **Ratio of means**? Meaning that the confidence interval when using Ratio of means would be smaller than that of Regression. How much worse off is an individual when using the estimator.

Ratio

Ratio estimation is most appropriate when

1) The relationship between y and x is a **straight line through the origin.**
2) The variance of y around the line is proportional to the value of x

Definition of **Population Ratio**: Consider a bivariate population of *N* units $(x_{(j)}, y_{(j)}), j = 1, \dots, N$. Let the $x_{(j)}$ have mean μ_x, total τ_x, and variance σ_x^2, and the $y_{(j)}$ have mean μ_y, total τ_y, and the variance σ_y^2. Then the population ratio is the desired estimate:

$$r = \frac{\sum_{j=1}^{N} y_{(j)}}{\sum_{j=1}^{N} x_{(j)}} = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$$

Definition of **Sample Ratio**: To estimate r, an SRS of n of the $(x_{(j)}, y_{(j)})$ pairs are taken, yielding (x_i, y_i), for $i = 1, \dots, n$. Thus, resulting in a bivariate random sample.
Take (X_i, Y_i) to be a pair of random variables for the i-th pair selected. The sample ratio as an estimator for r is:

$$\hat{R} = \frac{\sum_{j=1}^{N} Y_i}{\sum_{j=1}^{N} X_i} = \frac{\bar{Y}}{\bar{X}}$$

The resulting point estimate being

$$\hat{r} = \frac{\sum_{j=1}^{N} y_i}{\sum_{j=1}^{N} x_i} = \frac{\bar{y}}{\bar{x}}$$

Keeping in mind that $E(\bar{X}) = \mu_x$, and the analogues $E(\bar{Y}) = \mu_y$, the respective variances are:

$$\mathrm{V}(\bar{X}) = \frac{\sigma_x^2}{n}\left(\frac{N-n}{N-1}\right) \quad \text{and} \quad \mathrm{V}(\bar{Y}) = \frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right)$$

Where

$$\sigma_x^2 = \frac{1}{N}\sum_{j=1}^{N}(x_{(j)} - \mu_x)^2 \quad \text{and} \quad \sigma_y^2 = \frac{1}{N}\sum_{j=1}^{N}(y_{(j)} - \mu_y)^2$$

However, $E(\hat{R})$ is **not** an unbiased estimator, so proceeding with $E(\bar{Y}) = \mu$, as done before, would be incorrect. Thus, we consider a **biased** estimator. Such estimators can be good; when bias associated with the estimator is NOT large and variance is also small.

To accomplish this, one must compare a biased estimator with an unbiased estimator or with another biased estimator. This leads to the computation of the **mean square estimators (MSE).**

Definition of **Mean Square Error**: Suppose that $\hat{\mu}$ is a biased estimator for μ such that

$$E(\hat{\mu}) = m \neq \mu$$

The bias in $\hat{\mu}$ is a biased estimator for μ, such that

$$E(\hat{\mu} - \mu) = E(\hat{\mu}) - \mu = m - \mu$$

The mean square estimator in $\hat{\mu}$ as an estimator for μ is

$$MSE = E[(\hat{\mu} - \mu)^2] = E[[(\hat{\mu} - m) + (m - \mu)]^2] = E[(\hat{\mu} - m)^2] + E[(m - \mu)^2]$$
$$= V(\hat{\mu}) + (m - \mu)^2 = V(\hat{\mu}) + (Bias\ in\ \hat{\mu})^2$$

Thus, MSE for $\hat{R}$ is

$$MSE\ for\ \hat{R} = V(\hat{R}) + \left[E(\hat{R} - r)\right]^2$$

For the computation of the **sampling bias** in $\hat{R}$ as an estimator for r, consider

$$\hat{R} - r = \frac{\bar{Y}}{\bar{X}} - r = \frac{1}{\bar{X}}(\bar{Y} - r\bar{X})$$

Taking the expectation on both sides, results in the sampling bias in $\hat{R}$ as an estimator for r

$$E(\hat{R} - r) = E\left(\frac{\bar{Y}}{\bar{X}}\right) - r$$

Now for the computation of $E\left(\frac{\bar{Y}}{\bar{X}}\right)$. Consider again, as done for the simple case, a Taylor series expansion of $\frac{1}{\bar{X}}$ about $\bar{X} = \mu_x$:

$$\frac{1}{\bar{X}} \approx \frac{1}{\mu_x} - \frac{1}{\mu_x^2}(\bar{X} - \mu_x)$$

Thus, subbing into the following equation

$$\hat{R} - r = \frac{\bar{Y}}{\bar{X}} - r = \frac{1}{\bar{X}}(\bar{Y} - r\bar{X}) \approx \frac{1}{\mu_x}(\bar{Y} - r\bar{X}) - \frac{1}{\mu_x^2}(\bar{X} - \mu_x)(\bar{Y} - r\bar{X})$$

Taking the expectation, the sampling bias in $\hat{R}$ as an estimator for r is approximated by:

$$E(\hat{R} - r) \approx -\frac{1}{\mu_x^2}E[\bar{Y}(\bar{X} - \mu_x) - r\bar{X}(\bar{Y} - r\bar{X})]$$

Note that,

$$E[\bar{Y}(\bar{X}-\mu_x)-\ r\bar{X}(\bar{Y}-r\bar{X})]=Cov(\bar{X},\bar{Y})-rV(\bar{X})$$

Resulting in

$$E\left(\hat{R}-r\right)\approx-\frac{1}{\mu_x^2}[Cov(\bar{X},\bar{Y})-rV(\bar{X})]$$

For further computation, an expression for $Cov(\bar{X},\bar{Y})$ is required.

Recall, for any two random variables

$$V(X+Y)=V(X)+V(Y)+2Cov(X,Y)$$

This can be extended to random samples, i.e.

$$V(\bar{X}+\bar{Y})=V(\bar{X})+V(\bar{Y})+2Cov(\bar{X},\bar{Y})$$

Recall, since bivariate sampling is being used, the sample pairs consist of:

$$(x_1,y_1),\dots,(x_n,y_n)\quad\Rightarrow\qquad\bar{x}+\bar{y}=\frac{\sum_{i=1}^{n}x_i}{n}+\frac{\sum_{i=1}^{n}y_i}{n}=\frac{\sum_{i=1}^{n}(x_i+y_i)}{n}=\overline{x+y}$$

Therefore,

$$\bar{X}+\bar{Y}=\overline{X+Y}\qquad\text{so}\qquad V(\bar{X}+\bar{Y})=V(\overline{X+Y})$$

Thus, the following is obtained,

$$V(\overline{X+Y})=\frac{\sigma_{x+y}^2}{n}\left(\frac{N-n}{N-1}\right)\qquad\text{given}\quad\sigma_{x+y}^2=\frac{\sum_{j=1}^{N}[(x_{(j)}+y_{(j)})-(\mu_x-\mu_y)]^2}{N}$$

Expanding out the expression for σ_{x+y}^2, and solving

$$\sigma_{x+y}^2=\sigma_x^2+2Cov(X,Y)+\sigma_y^2$$

Recall the expression for covariance of two random variables in terms of the correlation coefficient ρ is:

$$\rho=\frac{Cov(X,Y)}{\sigma_x\sigma_y}$$

Isolating the above for $Cov(X,Y)$ and subbing into the expression for σ_{x+y}^2, results in,

$$\sigma_{x+y}^2=\sigma_x^2+2\rho\sigma_x\sigma_y+\sigma_y^2$$

Finally, reintroducing σ_{x+y}^2 back into the $V(\bar{X}+\bar{Y})=\ V(\overline{X+Y})$,

$$V(\bar{X}+\bar{Y})=\ V(\overline{X+Y})=\frac{\sigma_{x+y}^2}{n}\left(\frac{N-n}{N-1}\right)=\frac{\sigma_x^2+2\rho\sigma_x\sigma_y+\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right)$$

Recall

$$V(\bar{X}+\bar{Y}) = V(\bar{X}) + V(\bar{Y}) + 2Cov(\bar{X},\bar{Y})$$

Rearranging,

$$2Cov(\bar{X},\bar{Y}) = V(\bar{X}+\bar{Y}) - V(\bar{X}) - V(\bar{Y})$$

After substituting the analogues quantities,

$$Cov(\bar{X},\bar{Y}) = \frac{\rho\sigma_x\sigma_y}{n}\left(\frac{N-n}{N-1}\right)$$

Finally, the **sampling bias** $\hat{R}$ as an estimator for r is approximated by:

$$E\left(\hat{R}-r\right) \approx -\frac{1}{\mu_x^2}[Cov(\bar{X},\bar{Y}) - rV(\bar{X})] \approx \frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\left(\frac{N-n}{N-1}\right)r\sigma_x^2 - \rho\sigma_x\sigma_y$$

Next, the **sampling variability** $\hat{R}$ as an estimator for r. Recall,

$$\hat{R} - r = \frac{\bar{Y}}{\bar{X}} - r = \frac{1}{\bar{X}}(\bar{Y} - r\bar{X})$$

Taking the variance of both sides,

$$V\left(\hat{R}-r\right) = V(\hat{R}) = V\left(\frac{\bar{Y}}{\bar{X}} - r\right) = V\left[\frac{1}{\bar{X}}(\bar{Y} - r\bar{X})\right]$$

Now, a replacement of $\bar{X}$ with μ_x, gives

$$V\left(\hat{R}\right) = V\left[\frac{1}{\mu_x}(\bar{Y} - r\bar{X})\right] = \frac{1}{\mu_x^2}V(\bar{Y} - r\bar{X})$$

Consider a new random variable W_i, such that $W_i = Y_i - rX_i \rightarrow \bar{W} = \bar{Y} - r\bar{X}$. Thus, resulting in

$$V\left(\hat{R}\right) \approx \frac{1}{\mu_x^2}V(\bar{Y} - r\bar{X}) = \frac{1}{\mu_x^2}V(\bar{W}) = \frac{1}{\mu_x^2}\frac{\sigma_w^2}{n}\left(\frac{N-n}{N-1}\right)$$

- As the units are selected according to SRS, $V(\bar{W}) = \frac{\sigma_w^2}{n}\left(\frac{N-n}{N-1}\right)$
- $\sigma_w^2 = \frac{1}{N}\sum_{j=1}^{N}(w_{(j)} - \mu_w)^2$
- $\mu_w = \frac{1}{N}\sum_{j=1}^{N} w_{(j)} = \frac{1}{N}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right) = \mu_y - r\mu_x$
- Recall,

$$r = \frac{\mu_y}{\mu_x} \Rightarrow \mu_y - r\mu_x = 0 \Rightarrow \mu_w = 0 \Rightarrow \sigma_w^2 = \frac{1}{N}\sum_{j=1}^{N} w_{(j)}^2 = \frac{1}{N}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2$$

Thus, the **sampling variability in $\widehat{R}$ as an estimator for r**:

$$V(\hat{R}) \approx \frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\frac{1}{N}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2\left(\frac{N-n}{N-1}\right)$$

Now, an expression for $\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2$, can be obtained by

$$\begin{aligned}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2 &= \sum_{j=1}^{N}\left(y_{(j)} - \mu_y + \mu_y - rx_{(j)}\right)^2 \\ &= \sum_{j=1}^{N}\left(y_{(j)} - \mu_y\right)^2 + r^2\sum_{j=1}^{N}\left(x_{(j)} - \mu_x\right)^2 - \sum_{j=1}^{N} r\left(x_{(j)} - \mu_x\right)\left(y_{(j)} - \mu_y\right) \\ &= N\sigma_y^2 + Nr^2\sigma_x^2 - 2rNCov(X,Y)\end{aligned}$$

- Since: $Cov(X,Y) = \rho\sigma_x\sigma_y$

Finally, an expression for the variance is given by

$$V(\hat{R}) \approx \frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\frac{1}{N}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2\left(\frac{N-n}{N-1}\right) = \frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\left(\sigma_y^2 + r^2\sigma_x^2 - 2r\rho\sigma_x\sigma_y\right)\left(\frac{N-n}{N-1}\right)$$

An approximate (1 - α)100% CI for r is given by

$$\hat{r} \pm z_{\alpha/2}\sqrt{\frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\frac{1}{N}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2\left(\frac{N-n}{N-1}\right)}$$

With μ_x known an estimator for $V(\hat{R})$ is required. Thus

$$\hat{V}(\hat{R}) \approx \frac{1}{\mu_x^2}\left(\frac{1}{n-1}\right)\frac{1}{n}\sum_{j=1}^{N}\left(y_{(j)} - rx_{(j)}\right)^2\left(1 - \frac{n}{N}\right) \approx \frac{1}{\mu_x^2}\left(\frac{1}{n}\right)\left(S_y^2 + \hat{r}^2S_x^2 - 2\hat{r}\hat{\rho}S_xS_y\right)\left(1 - \frac{n}{N}\right)$$

- $S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$
- $S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}$
- $S_{xy} = \frac{\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}}{n-1}$
- $\hat{\rho} = \frac{S_{xy}}{S_xS_y}$

Note: As the experiments in the paper are preformed under design-based estimation, the above result of $\hat{V}(\hat{R})$ will be used

Therefore, with μ_x known, $(1-\alpha)100\%$ CI for r is given

$$\hat{r} \pm z_{\alpha/2}\sqrt{\frac{1}{\mu_x^2}\left(\frac{1}{n-1}\right)\frac{1}{n}\sum_{j=1}^{N}\left(y_{(j)}-rx_{(j)}\right)^2\left(1-\frac{n}{N}\right)}$$

If μ_x is unknown however, an adjustment using $\bar{x}$ as an approximation can be made.

However, a confidence interval for a finite population mean, μ_y, is obtained by multiplying the lower and upper limits of the C.I, for r by μ_x.

Means of Ratio

The means of ratio, or the mean of the sample ratio is denoted by:

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

Thus, Sample Ratio as an estimator for r is:

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}$$

This estimator is another variation on the ratio estimator. It comes with the benefit of being **unbiased**, however, it is susceptible to large variability. Therefore, the critical regions for when a confidence interval is constructed can often be pretty wide, which means that coverage will often be very good (Rao, T., 2002) (Formenti, M., 2014).

The variance of the estimator, $\hat{R}$, is

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

- $\sigma_R^2 = \frac{\sum_{j=1}^{N}\left(\frac{y_{(j)}}{x_{(j)}} - \mu_R\right)^2}{N}$
- $\mu_R = \frac{\sum_{j=1}^{N}\frac{y_{(j)}}{x_{(j)}}}{N}$

The estimated variances using sample variances

$$\hat{V}(\hat{R}) = \frac{S_R^2}{n}\left(1-\frac{n}{N}\right) \qquad \text{where} \qquad S_R^2 = \frac{\sum_{i=1}^{n}\left(\frac{y_i}{x_i} - \hat{r}\right)^2}{n-1}$$

Therefore, a known, $(1-\alpha)100\%$ CI for r is given

$$\hat{r} \pm z_{\alpha/2}\sqrt{\frac{S_R^2}{n}\left(1-\frac{n}{N}\right)}$$

Multiplying by multiplying by μ_x to get the C.I. for μ_y

Regression

The motivation behind using regression estimation is one of making sure that:

1) There is evidence of a linear relationship between y and x that **does not necessarily pass through the origin**.
2) The variance of y around the line is constant for all values of x.

Afterwards, the use of any other information provided by the variable x may be considered using a regression estimator.

Again, as under ratio estimation, one must have knowledge of $\mu_x(\bar{x})$.

Consider the **model**:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \quad \text{where} \qquad E_M(\varepsilon_i) = 0 \quad \text{and} \quad V_M(\varepsilon_i) = \sigma_Y^2$$

Firstly, a few definitions.

Define **Sum of Squares**: "A **square** is determined by squaring the distance between a datapoint and the regression line. A sum of squares is just the addition of all these squared distances."

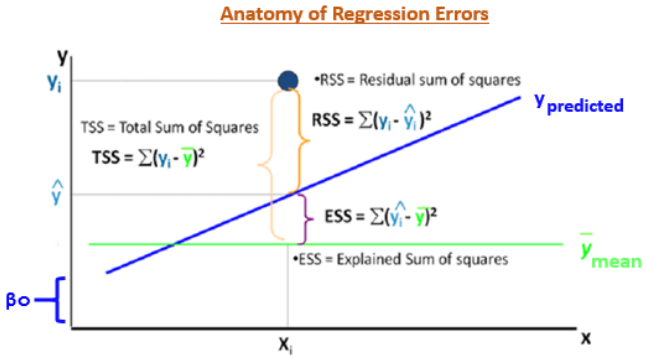


Figure 1 Anatomy of Regression Errors (Saraswat, M., 2016)

Definition of **Least squares**: "Least squares is a statistical method used to determine a line/curve of best fit. This is accomplished by minimizing the **sum of squares** which is created by a mathematical function." (Weisstein, E., n.d.).

Definition of **Linear least squares**: “A mathematical process of finding the best-fitting **line** to a given set of points by minimizing the sum of squares”.

Definition of **Ordinary Least Squares (OLS)**: “A type of linear least square process which is used for estimating the unknown parameters in a linear regression model. Using the principle of Least squares, the OLS picks the parameters of linear function on a set of variables.” (xlstat, n.d.)

A **fitted regression model**, using ordinary least squares, is given by:

$$\hat{y} = b_0 + b_1 x$$

where,

- $b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \hat{\rho}\frac{S_y}{S_x}$
- $b_0 = \bar{y} - b_1\bar{x}$

Recall: $\hat{\rho}$ is the sample correlation between X and Y. A proof of the two quantities above will be given later on.

To obtain an estimate for the population mean, one could rewrite the model above as

$$\hat{y} = b_0 + b_1 x = \bar{y} + b_1(x - \bar{x})$$

Now, regression estimates of μ_y is the predicted value of y from the fitted regression model when $x = \mu_x$, resulting in

$$\bar{y}_{reg} = \bar{y} + b_1(\mu_x - \bar{x})$$

With the associated estimator being

$$\bar{Y}_{reg} = \bar{Y} + \hat{\beta}_1(\mu_x - \bar{X})$$

Note: As per the ratio estimator, the regression estimator for μ_y is also **biased**. To show this,

$$E(\bar{Y}_{reg}) = \mu_y + \mu_x E(\hat{\beta}_1) - E(\hat{\beta}_1 \bar{X})$$

Hence with $\mu_x = E(\bar{X})$

$$E(\bar{Y}_{reg}) - \mu_y = -\left[E(\hat{\beta}_1 \bar{X}) - E(\hat{\beta}_1)E(\bar{X})\right] = -Cov(\hat{\beta}_1, \bar{X})$$

Note: If the regression line passes through **all the values of** $(x_{(j)}, y_{(j)})$ for the entire population, then the bias, $-Cov(\hat{\beta}_1, \bar{X}) = 0$, as $\hat{\beta}_1 = \beta_1$ for every sample. Also, if the sample drawn is large enough, the bias can be disregarded again.

Under the assumption that a large sample is drawn; $\hat{\beta}_1 \approx \beta_1 \approx b_1$ is constant for the j-th unit in the population.

Let,

$$d_{(j)} = y_{(j)} + b_1(\mu_x - x_{(j)})$$

then,

$$\mu_d = \frac{\sum_{j=1}^{N} d_{(j)}}{N} = \mu_y + b_1(\mu_x - \mu_x) \rightarrow \mu_y$$

Thus, for any i-th until from the sample,

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \bar{y} + b_1(\mu_x - \bar{x}) = \bar{y}_{reg}$$

Now, for the associated estimator, let

$$\bar{D} = \bar{Y} + \hat{\beta}_1(\mu_x - \bar{X}) = \bar{Y}_{reg}$$

Therefore, the accompanying variance is

$$V(\bar{Y}_{reg}) = V(\bar{D}) = \frac{\sigma_d^2}{n}\left(\frac{N-n}{N-1}\right)$$

Given

$$\sigma_d^2 = \frac{\sum_{j=1}^{N}(d_{(j)} - \mu_d)^2}{N} = \sigma_y^2 + b_1^2\sigma_x^2 - 2b_1\rho\sigma_x\sigma_y$$

However, this is under the model-based scenario. In a practical scenario, the calculation of σ_d^2 isn't plausible, as again, it is often not known. Thus, an estimate of $V(\bar{Y}_{reg})$ is required.

First, an expression for the $V(\bar{Y}_{reg})$ is required, with the variance needing to be minimum. This is due to the fact that each time an iteration of the estimator is calculated, it is closer to its expectation, thus being unbiased.

Definition of **Minimum Variance Unbiased Estimator** (**MVUE**): "An unbiased estimator that has lower variance than all other unbiased estimator for all possible values of the parameter."

As the imposed condition of $\hat{\beta}_1 \approx \beta_1 \approx b_1$ holds, the **unbiased** condition holds. Then, the expression for b_1 that **minimizes** $V(\bar{Y}_{reg})$ is equivalent to one which **minimizes** σ_d^2:

$$\frac{\partial\, \sigma_d^2}{\partial b_1} = \frac{\partial(\sigma_y^2 + b_1^2\sigma_x^2 - 2b_1\rho\sigma_x\sigma_y)}{\partial b_1} = 2b_1\sigma_x^2 - 2\rho\sigma_x\sigma_y$$

Setting $\frac{\partial\, \sigma_d^2}{\partial b_1}$ equal to 0, and solving for b_1 results in:

$$b_1 = \frac{\rho\sigma_y}{\sigma_x}$$

This is the **minimum** as when taking $\frac{\partial^2\, \sigma_d^2}{\partial {b_1}^2}$ (the second derivative) results in $2\sigma_x^2$, a positive value. Which, by optimization laws from calculus, predicate a minimum.

- Recall: $\hat{y} = b_0 + b_1 x$ and $b_1 = \hat{\rho}\frac{s_y}{s_x}$

Thus, an estimate of $V(\bar{Y}_{reg})$ is

$$\hat{V}(\bar{Y}_{reg}) = \frac{s_d^2}{n}\left(1 - \frac{n}{N}\right)$$

With σ_d^2 estimated by:

$$s_d^2 = \frac{\sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-1}$$

Where e_i is known as the residuals associated with fitting the regression model.

For a large enough sample size n however,

$$s_d^2 = MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Which is said to be the estimate of the variance of the error terms in the regression model.

Thus, if n large, then $\bar{Y}_{reg}$ has an approximate normal distribution.

Then an approximate $(1-\alpha)100\%$ CI for μ_y is

$$\bar{y}_{reg} \pm z_{\alpha/2}\sqrt{\hat{V}\left(\bar{Y}_{reg}\right)} \qquad \text{or} \qquad \bar{y}_{reg} \pm z_{\alpha/2}\sqrt{\frac{s_d^2}{n}\left(1-\frac{n}{N}\right)}$$

Simulation Study:

Now, generating values for x and using said values to generate values for y, a data set of (x_i, y_i) pairs are created. The creation of x_i is done on the basis of three different distributions: Normal, Uniform, and Exponential. This dataset is then analysed under the three techniques of: **Ratio of means (Rat1)**, **means of ratio (Rat2)**, and **regression (Reg)**. Upon running the analysis, a comparison is made to check the coverage as well as the width of each interval, in an attempt to compare and contrast the three different estimators.

A simulation was run with different parameters to test the choice of "Given that an estimator is known to be optimal, is there another which might be just as good, if not better?"

The data generated from the three different distributions were examined under three different assumptions; each assigning either **Rat1**, **Rat2**, or **Reg** to be the optimal choice. Then a simulation was run, where under each of the distributions, different parameters are considered.

Estimate μ_Y for a finite population size N, using the three estimators **Rat1**, **Rat2**, **Reg**. Two population sizes, namely, N=100, and N=1000 were used. The generation of N was done in accordance to

$$Y_i = Bx_i + D_i \qquad \text{with} \qquad B = 5, \text{for } i = 1, \dots, N.$$

For a given population size N, the x_j's were generated using three distributions. Specifically

- Uniform with lower limits $a = 16$ and $b = 65$
- Normal with mean 41 and s.d. 8
- Exponential with mean 8

The D_i's were generated from normal distributions with mean 0. There different assumptions on the conditional variance of D_i given x_i were made. Specifically,

- $V(D_i|x_i) = \sigma^2$
- $V(D_i|x_i) = x_i\sigma^2$
- $V(D_i|x_i) = x_i^2\sigma^2$

Three different values of σ^2 were chosen: 1, 625,and 1000. The response variable values for the entire population were then calculated using $Y_i = 5x_i + D_i$.

200 samples of size $n = 10$ drawn from the finite population, and C.I. for μ_y computed using each estimator. “Coverage” and “avg. width” of the C.I’s reported for each estimator. A total of 54 combinators were instigated. An indicator was used to see how often each of the three estimators actually landed within the confidence interval, thus giving the “coverage” (i.e. the proportion of times it hit the 95% CI). It was also noted how variable the widths of the CI’s were from one another, by taking the standard deviations of each CI and checking it against the other 199 which were simulated alongside it.

Coverage rates and confidence interval widths with variance:

$$V(D_i|x_i) = \sigma^2 x_i$$

For the three estimators for different error term distributions, error term variance, distribution for the error terms, and population sizes. Note variance minimization on an unbiased estimator for B leads back to **Rat1** in this case (refer back to the motivation for this paper).

$Dist.of\ x's$	σ^2	N	n	$Coverage\ (rat1)$	$Coverage\ (rat2)$	$Coverage\ (reg)$	$Width\ (rat1)$	$Width\ (rat2)$	$Width\ (reg)$
Expo.	1	100	10	0.930	0.935	0.910	1.460	1.603	1.4696
	1	1000	10	0.920	0.930	0.865	1.540	1.659	1.519
	625	100	10	0.935	0.945	0.875	44.63	32.95	43.11
	625	1000	10	0.930	0.940	0.885	33.30	36.30	32.91
	2500	100	10	0.905	0.900	0.835	65.64	65.64	64.43
	2500	1000	10	0.905	0.905	0.820	68.64	72.60	68.30
Unif.	1	100	10	0.890	0.920	0.850	2.467	1.989	2.475
	1	1000	10	0.915	0.925	0.870	1.944	2.153	1.962
	625	100	10	0.910	0.905	0.870	52.44	58.24	51.79
	625	1000	10	0.915	0.940	0.880	49.21	61.14	49.87
	2500	100	10	0.965	0.950	0.945	79.00	98.40	75.97
	2500	1000	10	0.940	0.950	0.925	86.66	108.61	84.99
Normal	1	100	10	0.905	0.905	0.870	1.854	1.886	1.774
	1	1000	10	0.905	0.925	0.885	1.672	1.597	1.676
	625	100	10	0.945	0.945	0.905	57.92	48.62	56.04
	625	1000	10	0.945	0.945	0.915	46.09	47.36	46.60
	2500	100	10	0.930	0.940	0.915	73.10	74.63	74.47
	2500	1000	10	0.900	0.915	0.880	96.78	98.58	93.87

Coverage rates and confidence interval widths:

$$V(D_i|x_i) = \sigma^2$$

For the three estimators for different error term distributions, error term variance, distribution for the error terms, and population sizes. Note variance minimization on an unbiased estimator for B leads back to **Reg** in this case (refer back to the motivation for this paper).

$Dist. of\ x's$	σ^2	N	n	$Coverage\ (rat1)$	$Coverage\ (rat2)$	$Coverage\ (reg)$	$Width\ (rat1)$	$Width\ (rat2)$	$Width\ (reg)$
Expo.	1	100	10	0.905	0.905	0.870	0.2443	0.3278	0.2343
	1	1000	10	0.910	0.910	0.860	0.2590	0.3404	0.2544
	625	100	10	0.895	0.925	0.845	6.881	9.214	7.246
	625	1000	10	0.895	0.915	0.865	7.184	9.200	7.384
	2500	100	10	0.940	0.930	0.900	13.851	20.59	14.120
	2500	1000	10	0.920	0.925	0.860	14.99	18.80	14.45
Unif.	1	100	10	0.920	0.925	0.850	0.2729	0.5085	0.2648
	1	1000	10	0.900	0.920	0.820	0.2692	0.4785	0.2570
	625	100	10	0.935	0.945	0.900	5.462	7.231	5.537
	625	1000	10	0.935	0.940	0.910	6.878	11.887	7.026
	2500	100	10	0.905	0.915	0.880	11.844	18.85	12.030
	2500	1000	10	0.910	0.890	0.875	14.38	22.21	14.15
Normal	1	100	10	0.890	0.890	0.870	0.3040	0.3659	0.3104
	1	1000	10	0.895	0.910	0.875	0.2990	0.3483	0.2942
	625	100	10	0.920	0.910	0.880	6.137	8.158	6.171
	625	1000	10	0.925	0.925	0.895	6.527	7.896	6.440
	2500	100	10	0.925	0.920	0.875	12.897	14.73	12.361
	2500	1000	10	0.900	0.895	0.870	14.010	16.21	14.46

Coverage rates and confidence interval widths:

$$V(D_i|x_i) = \sigma^2 x_i^2$$

For the three estimators for different error term distributions, error term variance, distribution for the error terms, and population sizes. Note variance minimization on an unbiased estimator for B leads back to **Rat2** in this case (refer back to the motivation for this paper).

$Dist.of\ x's$	σ^2	N	n	$Coverage\ (rat1)$	$Coverage\ (rat2)$	$Coverage\ (reg)$	$Width\ (rat1)$	$Width\ (rat2)$	$Width\ (reg)$
Expo.	1	100	10	0.905	0.930	0.835	8.354	6.241	8.166
	1	1000	10	0.915	0.945	0.825	9.075	6.379	7.685
	625	100	10	0.925	0.945	0.840	14.2	10.7	14.4
	625	1000	10	0.890	0.905	0.845	16.2	13.2	14.2
	2500	100	10	0.925	0.930	0.845	26.3	25.2	25.5
	2500	1000	10	0.905	0.925	0.835	36.5	24.8	30.2
Unif.	1	100	10	0.905	0.925	0.870	13.083	11.485	12.645
	1	1000	10	0.885	0.915	0.835	0.998	0.815	0.995
	625	100	10	0.905	0.925	0.865	25.4	18.8	26.0
	625	1000	10	0.925	0.930	0.870	24.6	19.9	24.2
	2500	100	10	0.910	0.940	0.855	44.5	33.1	42.6
	2500	1000	10	0.900	0.900	0.850	47.1	36.6	46.6
Normal	1	100	10	0.880	0.875	0.870	10.702	9.340	10.421
	1	1000	10	0.920	0.930	0.895	11.945	11.193	11.778
	625	100	10	0.925	0.925	0.895	302.0	268.1	305.3
	625	1000	10	0.900	0.905	0.890	353.1	304.8	345.3
	2500	100	10	0.930	0.930	0.905	482.8	459.9	465.7
	2500	1000	10	0.935	0.930	0.880	647.1	603.8	635.3

Qualitative analysis:

Rat1:

Under the assumption that the "ratio of means (Rat 1)" is **correct**, the behaviour of the data is contingent on distribution of the x_i, as well as the magnitude of the variance.

For data that follows an exponential distribution, it can be noted both **Rat1** and **Rat2** seem to have equivalent levels of coverage, as both are hitting approximately 90% coverage for each test. However, the coverage of **Reg** does not fare so well. The width of the intervals is contingent on the variance of the x_i generated. Therefore, lower levels of variance will lead to lower levels of variation between confidence levels.

For data that follows a uniform distribution, it can be noted that both the coverages of the aforementioned **Rat1** and **Rat2** seem to drop in terms of coverage, as both are hitting lower than 90%, with **Reg** being even worse off than it was in the previous case. However, once the variability is increased, all three seem to preform quite well, hitting almost the full 95%. This seems to indicate that the combination of uniformly distributed x's with high variance leads to more adequate coverage. As per before, the width is contingent on the variance. Thus, a higher variance leads to higher variability among the confidence levels.

For data that follows a Normal distribution, it can be noted that all three perform well with respect to high levels of variability. However, not too high, as a level of 2500 leads to drops in coverage. **Reg** again preforms a little worse than the other 2, but still holds up well (approximately 90%). The widths of the intervals are proportionally higher than when the x's were distributed in accordance to exponential and uniform, therefore the variability between all of them is quite high. It is also interesting to note that as the population size increases (**N=1000**), with a variance of 2500, the level of coverage dips as well.

A general consensus regarding the width of the intervals is that, other than a few scenarios, the width among intervals is least variable for the **Reg** estimation. **Rat1** is consistently more variable than **Reg** among its CI's, however, it is less variable than **Rat2**, as its CI's are much more volatile. This lends credence to the fact that **Rat2**, while being unbiased, is very susceptible to high levels of volatility. Therefore, just because it is an unbiased estimator, does not mean it is the "best" estimator.

Reg:

For data that follows an exponential distribution, **Rat1** and **Rat2**, once again perform better than **Reg**, as the former both hit approximately a 90% CI while the latter hits maybe 85% of the time. As the variance increases, the variability among the confidence intervals also increases.

For data that follows a uniform distribution, **Rat1** and **Rat2** once again perform better than **Reg**. A key aspect to note is that as variance increases, the coverage of the confidence also increases. However, for a variance of 2500, both **Rat1** and **Rat2** dip proportionally more in coverage than **Reg** does. When looking at the width of the intervals, it can be noted that between then three, **Rat1** seems to have the least variance among its CI's, with **Reg** following closely and **Rat2** trailing.

For data that follows a Normal distribution, **Rat2** fairs the best out of the three estimations, followed by **Rat1** and then **Reg**. An important observation to be made here is that for variance of 2500, the dip in coverage for **Rat2** is quite low, and the dip for **Rat1** isn't even prevalent. With an increase in population size of 1000, however, leads to coverage dip for all, with **Reg** being the lowest. The variance among the confidence intervals is led by **Rat2**, followed by **Rat1** and then **Reg**.

A general consensus regarding the width of the intervals is that, when under exponential and uniform, the variance among the confidence intervals is led by **Rat2** once

again. This is to be expected as it is very volatile to variability; however, CI's under **Rat1** performed better than they had under **Reg** (the widths of the CI's were less variable for **Rat1** comparative to **Reg**). Only under the Normal distribution did **Reg** once again have the least amount of variability among its CI's.

This is key, as this implies that even though the correction assumption is that **Reg** is the correct estimator here, on average, **Rat1** has better coverage and the variability among its CI's is smaller. Meaning that the choice of **Rat1** here is not a "bad" decision for an estimator.

Rat2:

For data that follows an exponential distribution, **Rat2** has a better coverage in comparison to the other two in every circumstance. Even when the variance increases, the dip for **Rat2** is proportionally less than that of the other two. When looking at the width, **Rat2** has a smaller variance in every single instance. While it's true that as the variance increases, the volatility between the confidence intervals increases, it remains that **Rat2** has the smallest variance when compared to the other two. It is superior in every way. However, this is not to say that the choice of **Rat1** would be necessarily a "bad" choice. **Rat1** covers 90% of the time in almost all circumstances. The issue is that comparatively, it has the largest variation of width among its confidence intervals.

For data that follows a uniform distribution, again, as before **Rat2** regimes supreme. With very high coverages in almost all instances, only dipping for a high variance and large population. Analogously, **Rat1** follows next with covering, on average, about 90% of the time. Again, the width of the intervals is smallest with **Rat2**, then **Reg**, and then **Rat1**.

For data that follows a normal distribution, **Rat2** once again preforms very well, however, the other two also preform decent, with **Rat1** hitting a coverage of over 90% on average and **Reg** hitting a coverage of just under 90% on average. Therefore, under a Normal distribution, the choice of selecting **Rat1** or **Reg** over **Rat2**, whilst **Rat2** is a superior estimator is

not too detrimental. Once again, the width of the intervals is smallest with **Rat2**, then **Reg**, and then **Rat1**.

It is important to note that under the assumption that **Rat2** is correct, the variation among the width of the confidence intervals is massive. From exponential to uniform to normal, the width of each interval is increasingly more variable. This would mean that a situation where **Rat2** would be optimal is a situation where the data generated itself is extremely volatile. Therefore, the use of **Rat1**, which has less susceptibility to variation might actually be a good alternative.

Conclusion:

As seen above, the choice of estimators isn't always so cut and dry. There might be scenarios where the choice of an estimator which is known to be better, like in the **Reg** case, might be outdone by another. It might also be possible to have a situation where the data itself is so volatile, like in the **Rat2** case, that the choice of another might be a better choice, as it might not do as well as the superior estimator, but it might not be weighed down by the intrinsic traits which demand more conventions. It might also be the case that the coverage is not the most optimal, as in all cases, **Reg** didn't cover as well as the other two, however, it always had smaller volatility among its CI's which may be preferable in circumstance which might demand for such a task. The choice of an estimator is depended on a lot of things, and therefore, there isn't really a clear-cut selection in every circumstance. Just those which preform the task at hand; and a different choice doesn't necessarily have to be a wrong one.

Appendix:

```
gmacro
ygeneration
# k1 is N
# k2 is variance
# k6 is Beta
# k10 is the number of samples drawn
# k11 is true population mean
# k12 is sample size
name c2 "D"
name c3 "Y"
name c4 "xsample"
name c5 "ysample"
name c6 "y over x"
name c7 "YbarRat1"
name c8 "VarNoFiniteCorrRat1"
name c9 "Vhat(YbarRat1)"
name c10 "LowerRat1"
name c11 "UpperRat1"
name c14 "Ybar Rat2"
name c15 "VarNoFiniteCorrRat2"
name c16 "Vhat(YbarRat2)"
name c17 "LowerRat2"
name c18 "UpperRat2"
name c21 "YbarReg"
name c22 "VarNoFiniteCorrReg"
name c23 "Vhat(YbarReg)"
name c24 "LowerReg"
name c25 "UpperReg"
#base 46
let k1=1000
let k2=2500
let k6=5
let k10=200
let k12=10
random k1 c1;
#---Exponential---
#exponential 8.
#let c1=round(c1,0)+16
#---Uniform---
#uniform 16 65.
#let c1=round(c1,0)
#---Normal---
normal 41 8.
let c1=round(c1,0)
#---
do k3=1:k1
#------The follow are the 3 settings for the assummtions of Rat1 Rat2 and Reg
#let k4=c1(k3)*k2             #Rat1
#let k4=k2                    #Reg
let k4=(c1(k3)**2)*k2         #Rat2
let k5=sqrt(k4)
random 1 c30;
normal 0.0 k5.
let c2(k3)=c30(1)
```

```
enddo
erase c30
let c3=k6*c1+c2
do k9=1:k10
Sample k12 c1 c3 c4 c5.
let c6=c5/c4
let k7=sum(c5)/sum(c4)
let k14=stde(c4)
let k15=stde(c5)
corr c4 c5 m1.
copy m1 c30-c31
let k16=c31(1)
let k17=k16*(k15/k14)
erase c30 c31
let k8=mean(c6)
let c7(k9)=k7*mean(c1)
let c8(k9)=(1/k12)*((k15**2)+k7*k7*(k14**2)-2*k7*k16*k14*k15)
let c9(k9)=c8(k9)*(1-k12/k1)
let c10(k9)=c7(k9)-1.96*sqrt(c9(k9))
let c11(k9)=c7(k9)+1.96*sqrt(c9(k9))
let c14(k9)=k8*mean(c1)
let c15(k9)=(mean(c1)**2)*(1/k12)*(stde(c6)**2)
let c16(k9)=c15(k9)*(1-k12/k1)
let c17(k9)=c14(k9)-1.96*sqrt(c16(k9))
let c18(k9)=c14(k9)+1.96*sqrt(c16(k9))
let c21(k9)=mean(c5)+k17*(mean(c1)-mean(c4))
let c22(k9)=(1/k12)*(k15**2)*(1-(k16**2))
let c23(k9)=c22(k9)*(1-k12/k1)
let c24(k9)=c21(k9)-1.96*sqrt(c23(k9))
let c25(k9)=c21(k9)+1.96*sqrt(c23(k9))
enddo
let k11=mean(c3)
let c12=(c10 le k11 and c11 ge k11)
let c19=(c17 le k11 and c18 ge k11)
let c26=(c24 le k11 and c25 ge k11)
tally c12 c19 c26
let c13=c11-c10
let c20=c18-c17
let c27=c25-c24
desc c13 c20 c27
endmacro
```

Works Cited:

Baird, S. (n.d.). Confidence Interval. Retrieved from: https://www.processmodel.com/blog/faq-items/confidence-interval/

Formenti, M. (2014, September 3). Mean of Ratios or Ratio of Means: statistical uncertainty applied to estimate Multiperiod Probability of Default. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1409/1409.4896.pdf

Rao, T. (2002, March). Mean of ratios or ratio of means or both? Retrieved from: https://www.researchgate.net/publication/222776446_Mean_of_ratios_or_ratio_of_means_or_both

Weisstein, E (n.d.). Least Squares Fitting. Retrieved from: http://mathworld.wolfram.com/LeastSquaresFitting.html

Xlstat (n.d.). Ordinary Least Squares regression (OLS). https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols