Properties of the Sample Mean

Charlie Gibbons
Economics 140
University of California, Berkeley

Summer 2011

Outline

Sample analogue

Population concept: $\mathbb{E}(Y) = \mu$.

Sample concept: $\bar{y} \equiv \frac{1}{N} \sum y_i = \hat{\mu}$.

We estimate the population mean by using the sample analogue.

$\bar{y}$ is an *estimator*, a function that takes the data and turns it into an *estimate*, the value of an estimator for a particular data set.

Question: Is this a good strategy?

Goals

What makes a good strategy?

- Our estimator is right on average—it is accurate.
- Our estimator doesn't vary a lot—it is precise.

Ultimate goal: Find an estimator for μ and determine the statistical properties of that estimator.

Unbiasedness

The expected value of our estimator is

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum y_i\right] &= \frac{1}{N}\mathbb{E}\left[\sum y_i\right] \\
&= \frac{1}{N}\sum \mathbb{E}\left[y_i\right] \\
&= \frac{1}{N}\sum \mu \\
&= \frac{1}{N} N\mu \\
&= \mu.
\end{aligned}$$

An estimator for μ is called *unbiased* when the expected value of the estimator is μ.

Consistency

The *law of large numbers* states that, as N goes to ∞,

$$\frac{1}{N}\sum y_i \to \mathbb{E}[Y] = \mu;$$

that is, averages turn into expectations as the amount of data gets large.

An estimator for μ is said to be *consistent* when the estimator converges to μ when the number of observations goes to infinity.

Comparing the two

Unbiasedness tells us that, no matter how much data we have, our estimator is right on average.

Consistency tells us that, if we have infinite data, our estimator gives the desired answer.

We do not know how big N has to be in order to be "close enough" to infinity to get the right answer.

Though our estimator here is both unbiased and consistent, that isn't always the case.

Variance of the estimator

Next, given that $\mathrm{Var}(y_i) = \sigma^2$ and that the y_i's are uncorrelated, calculate the variance of the estimator:

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}\sum y_i\right) &= \frac{1}{N^2}\mathrm{Var}\left(\sum y_i\right) \\
&= \frac{1}{N^2}\sum \mathrm{Var}(y_i) \\
&= \frac{1}{N^2} N\sigma^2 \\
&= \frac{\sigma^2}{N}.
\end{aligned}$$

Interpretation

What does it mean that the estimator has a variance?

If we sample N people and calculate their average y, we get one value. If we draw another sample, we get a different average. The variance of our estimator tells us how spread out these averages will be.

Because our estimator is a function of a random variable (*e.g.*, Y), *our estimator is a random variable* and thus has a mean and variance.

Note that the variance goes down when N goes up.

Best unbiased

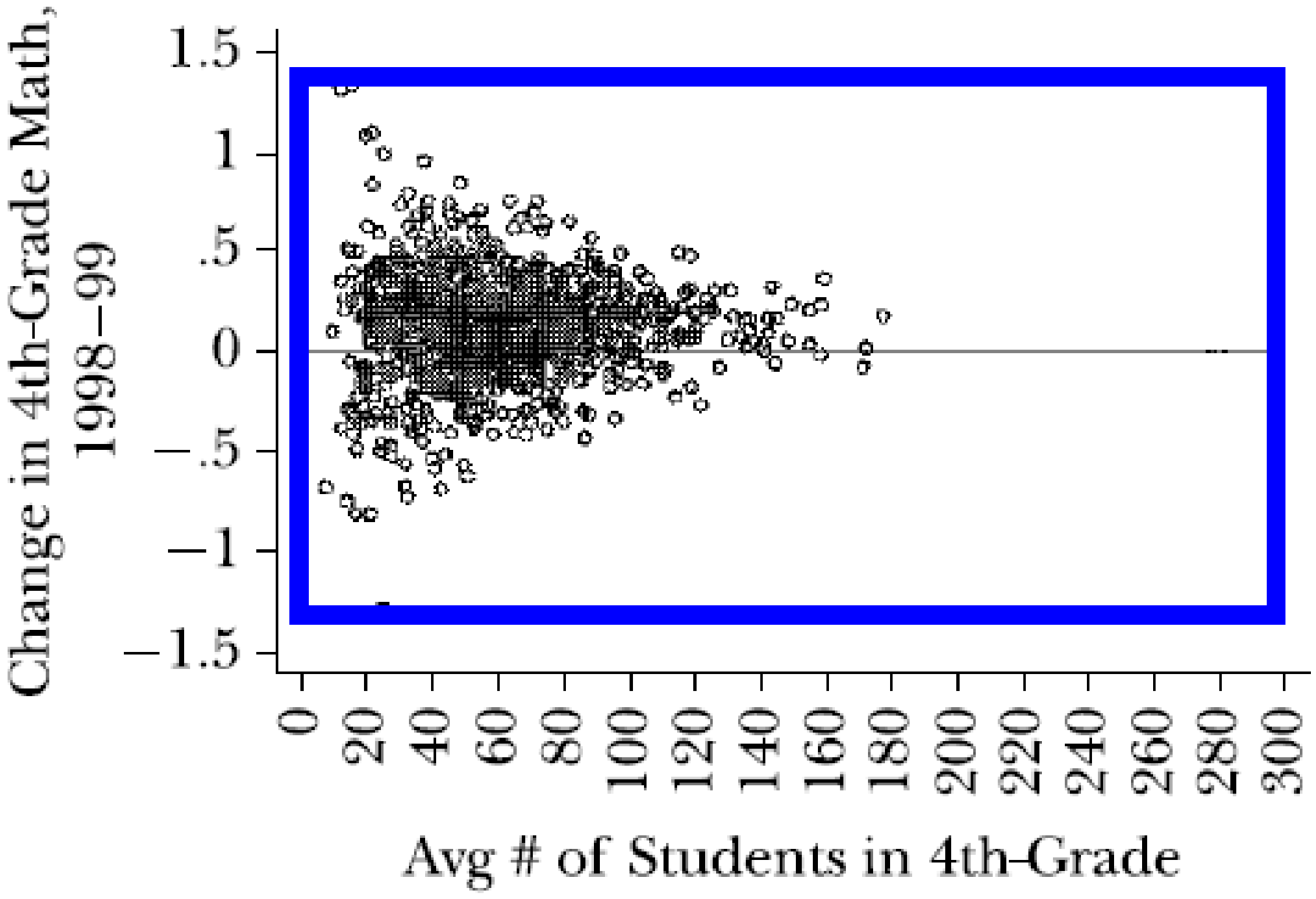
No other unbiased estimator for μ has a smaller variance than the sample mean. If an estimator has the smallest variance of all estimators, it is called *best*.

The sample mean is the best unbiased estimator.

School size and top performers

School Size	*Percentage Ever "Top 25" 1997–2000*
Smallest decile	27.7%
2nd	11.8
3rd	8.2
4th	3.6
5th	2.4
6th	3.6
7th	4.8
8th	7.1
9th	0
Largest decile	1.2

Variance of the mean across sample sizes



Central limit theorem

The *central limit theorem* builds upon the law of large numbers. It states that

$$\sqrt{N}\left(\bar{y}-\mu\right)=\sqrt{N}\left(\frac{1}{N}\sum y_i-\mu\right)\overset{N\to\infty}{\longrightarrow}\mathrm{N}(0,\sigma^2)$$

or

$$\bar{y}\overset{N\to\infty}{\longrightarrow}\mathrm{N}\left(\mu,\frac{\sigma^2}{N}\right).$$

The sample mean has an *asymptotically* normal distribution.

In finite samples, the normal distribution is only an approximation.

Notice that, as N goes to infinity, the variance of the estimator goes to 0.

Standardized estimator

Hence, as N goes to infinity,

$$\frac{\bar{y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim \mathrm{N}(0, 1).$$

If we standardize our estimator, it has a standard normal distribution.

Note that the "hat" on the standard deviation emphasizes that we must estimate this parameter.

Biased but consistent

Consider an estimator for the sample variance s^2:

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

Is it unbiased?

A biased estimator

$$
\begin{aligned}
\mathbb{E}[s^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}((y_i-\mu)-(\bar{y}-\mu))^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(y_i-\mu)^2\right] - 2\frac{1}{N}\mathbb{E}\left[(\bar{y}-\mu)\sum_{i=1}^{N}(y_i-\mu)\right] \\
&\quad + \mathbb{E}\left[(\bar{y}-\mu)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(y_i-\mu)^2\right] - \mathbb{E}\left[(\bar{y}-\mu)^2\right] \\
&= \mathrm{Var}(y_i) - \frac{\mathrm{Var}(y_i)}{N} = \frac{N-1}{N}\mathrm{Var}(y_i).
\end{aligned}
$$

Biased, but consistent

s^2 is a biased estimator of the variance of y_i—we have

$$\mathbb{E}[s^2] = \frac{N-1}{N}\sigma^2.$$

But, since $\frac{N-1}{N}$ goes to 1 as N goes to infinity, this is a consistent estimator.

Unbiased, but not consistent

Suppose that, no matter how much data we collect, we just use the first value that we got as our estimator $\hat{\mu} = y_1$. This is unbiased:

$$\mathbb{E}\left[\hat{\mu}\right] = \mathbb{E}[y_1] = \mu.$$

It is not consistent, however. No matter how much data we have, our estimator is not guaranteed to have the true value.

One of the requirements of consistency is that the variance of the estimator has to go to 0 as N goes to infinity.