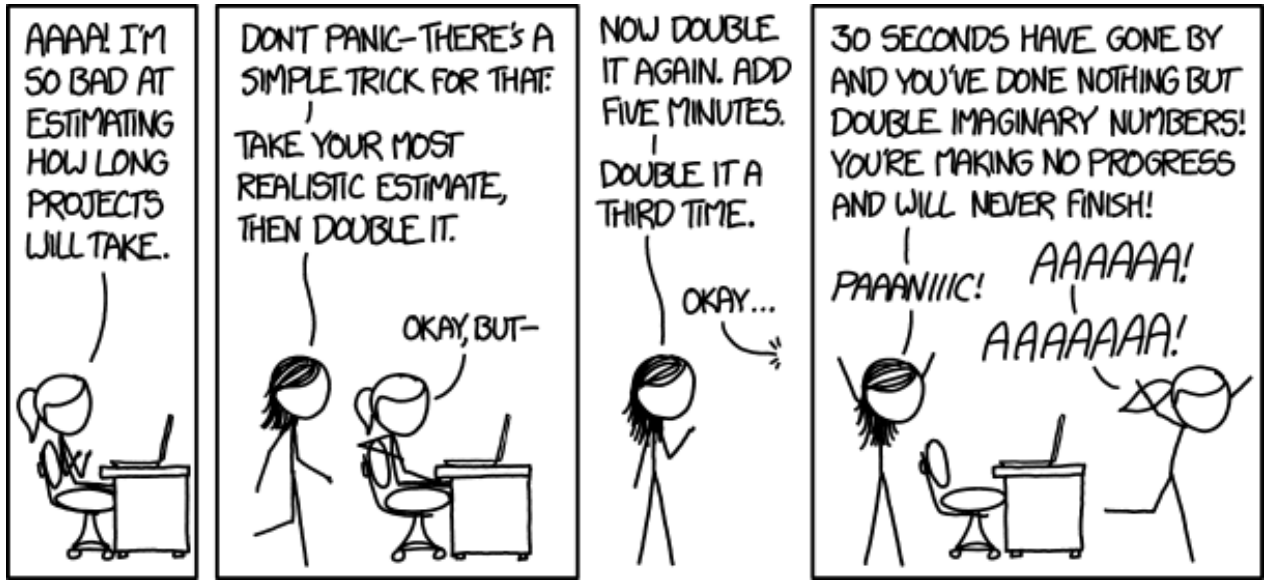
Estimation I: *Estimating the mean of the distribution*

Your Challenge: Estimate the mean, μ, of a random variable Y. Both the mean, μ, of Y and its variance, σ^2, are unknown.

Your Strategy: Sample independently n times from the distribution; use the data to estimate the unknown mean.



Random Sampling, Estimators and Estimates

***Ex Ante* (estimators as random variables) v. *Ex Post* (estimates as data)**:

- ***E Ante*:** n indept. random draws from the dist. Y. Each draw is a random variable, Y_i.
  a. The Y_i's are ***iid*** (independently and identically distributed) with distribution Y.
  b. A ***point estimator*** of μ is some function of the observed values of the Y_i's; it's a random variable, *ex ante*… and a number, *ex post*.
  c. **Estimators are *rules***, which assign different estimates to different drawn samples.
- ***Ex Post:* Estimates (data; observations)**: After we have the sample data $\{y_1, y_2, \ldots, y_n\}$, the ***estimate*** is the value of the point estimator for the given sample.

Here's an Estimator: The Sample Mean

- ***ex ante***, the sample mean estimator is a random variable, and takes on different values with different probabilities, reflecting the random nature of the sampling process:

$$M(Y_1, Y_2, ..., Y_n) = \bar{Y} = \frac{1}{n}\sum Y_i \ .$$

- ***ex post***, and after the sample data are drawn, the sample mean estimator provides us with a point estimate: $M(y_1, y_2, ..., y_n) = \bar{y} = \frac{1}{n}\sum y_i \ .$

Linear and Unbiased Estimators (LUEs)

- ***ex ante:***
  a. ***Linear Estimators***: $M = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_n Y_n$.
  b. ***Unbiased***: $E(M) = \mu$
  c. ***Linear Unbiased Estimators (LUEs)*: *Linear & Unbiased***
- ***ex post***:
  a. An estimate, given your sample: $m = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + ... + \beta_n y_n$.
  b. Close or not? No idea! But…
  c. Estimates generated in this fashion will on average equal μ (if M is an unbiased estimator of μ **)**

LUEs for the unknown mean

- ***Linear Estimators***: $M = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_n Y_n$.
- ***Expectation***: $E(M) = \beta_0 + \beta_1 \mu + \beta_2 \mu + ... + \beta_n \mu = \beta_0 + \sum \beta_i \mu = \beta_0 + \mu \sum \beta_i$.
- ***Unbiased:*** $E(M) = \mu \rightarrow \beta_0 + \mu \sum \beta_i = \mu \rightarrow \beta_0 = 0$ and $\sum_{i=1}^{n} \beta_i = 1$
- ***LUEs***: So the class of *LUE*s (*Linear Unbiased Estimators*) for our estimation problem.:

  $M = \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_n Y_n$, where $\sum_{i=1}^{n} \beta_i = 1$.

Variance of the LUEs

- Since the Y_i's are pairwise independent, $Cov(Y_i, Y_j) = 0$ *for* $i \neq j$, and $Var(M) = \beta_1^2 Var(Y_1) + \beta_2^2 Var(Y_2) + ... + \beta_n^2 Var(Y_n)$.
- Since $Var(Y_i) = \sigma^2$ for each i, $\boxed{Var(M) = \sigma^2 \sum \beta_i^2}$…
- The variance of M depends on the β_i's and differs across the LUEs.

Best Linear Unbiased Estimators (BLUE)

BLUE I:

- The ***Best Linear Unbiased Estimator*** (***BLUE***) will be the estimator in the class of LUEs with minimum variance.

- **LUE candidates**: $M = \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_n Y_n$, where $\sum_{i=1}^{n} \beta_i = 1$,

BLUE: Find the $\{\beta_i\}$ that minimize the variance within this group/class of estimators.

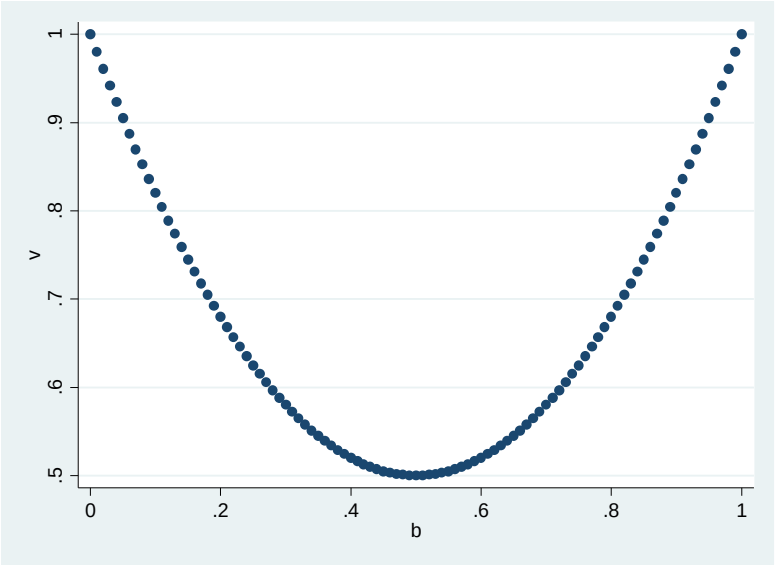
LUEs are not Alone!

- **There are lots (an infinite number!) of candidate LUEs!**
- $M_1 = Y_1$: $E(M_1) = \mu$ and $Var(M_1) = \sigma^2$
- $M_{2.1} = .5Y_1 + .5Y_2$: $E(M_{2.1}) = \mu$ and

  $Var(M_{2.1}) = .5^2\sigma^2 + .5^2\sigma^2 = \frac{2}{4}\sigma^2 = \sigma^2/2$
- $M_{2.2} = \beta Y_1 + (1-\beta)Y_2$:

  a. $E(M_{2.2}) = \mu$ and $Var(M_{2.2}) = \beta^2\sigma^2 + (1-\beta)^2\sigma^2$
  $= \sigma^2\left\{\beta^2 + (1-\beta)^2\right\}$

  b. So to minimize $Var(M_{2.2})$, set $\beta = .5$.

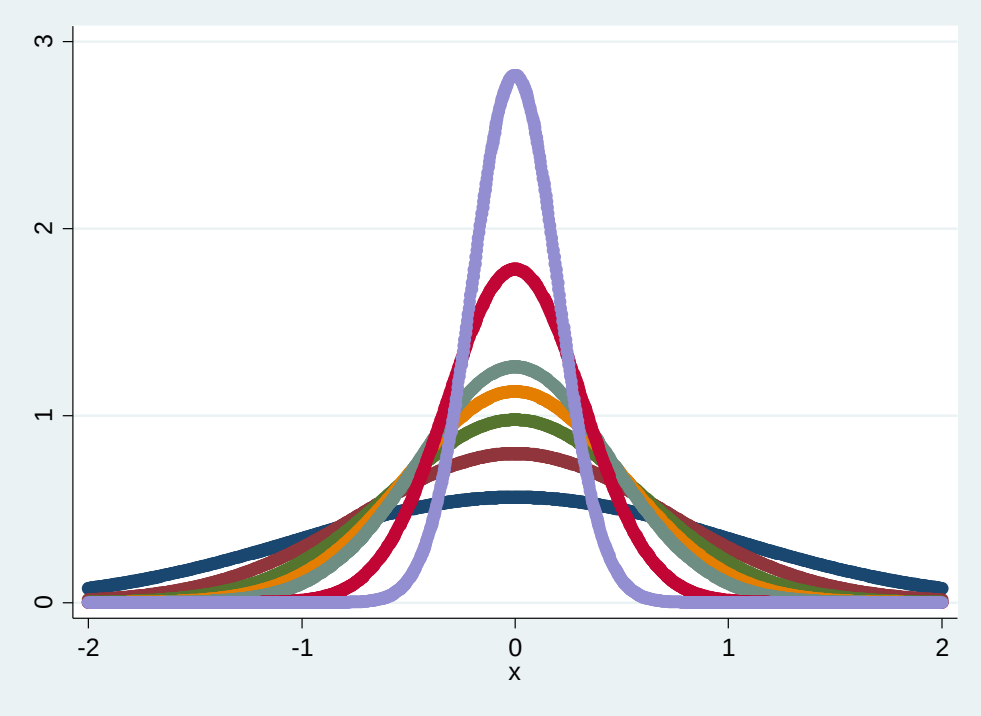


More Candidate LUEs

- $M_{3.1} = (1/3)Y_1 + (1/3)Y_2 + (1/3)Y_3$: $E(M_{3.1}) = \mu$ and

  $Var(M_{3.1}) = (1/3)^2\sigma^2 + (1/3)^2\sigma^2 + (1/3)^2\sigma^2 = \frac{3}{9}\sigma^2 = \sigma^2/3$

- Here's what the distributions of the equally weighted sample means (with different sample sizes… so $M_1, M_{2.1}, M_{3.1}$...) look like…. Assuming $Y \sim N(0,1)$:

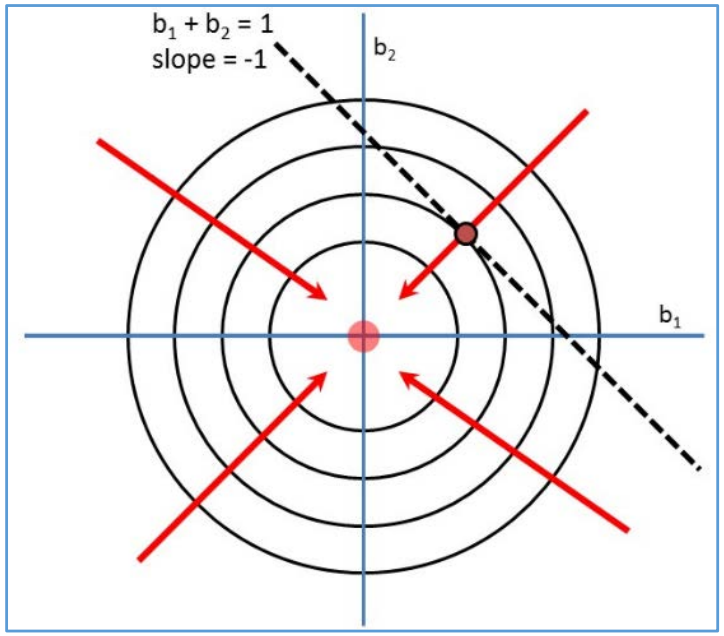


As the sample size increases, the distribution of the sample mean becomes more and more tightly concentrated around the true unknown mean, μ .

BLUE II: The optimization problem

- To find the BLUE estimator of the unknown mean μ, we need to solve the following optimization problem:

$$\min Var(M) = \sigma^2 \sum \beta_i^{\,2} \text{ subject to } \sum_{i=1}^{n} \beta_i = 1.$$



- Solution: $\beta_i^* = \frac{1}{n}$ *for* $i = 1, ..., n$, and $M = \bar{Y} = \frac{1}{n}\sum Y_i$.

- ***The Sample Mean is BLUE!***

BLUE III: Wrapup/Review

- Since the Sample Mean is unbiased and has minimum variance in the class of LUE's, it is a ***BLUE***
- ***(ex ante)*** Estimators are random variables, taking on different values depending on the drawn sample.

  a. $M = \bar{Y} = \frac{1}{n}\sum Y_i$, is a random variable (values depend on the actual sample)

- ***(ex post)*** Estimates are numbers, the value of the estimator for our particular drawn sample (set of observations).

  a. For a particular sample $\{y_1, y_2, \ldots, y_n\}$, $m = \bar{y} = \frac{1}{n}\sum y_i$. Close to μ? ***No clue!***

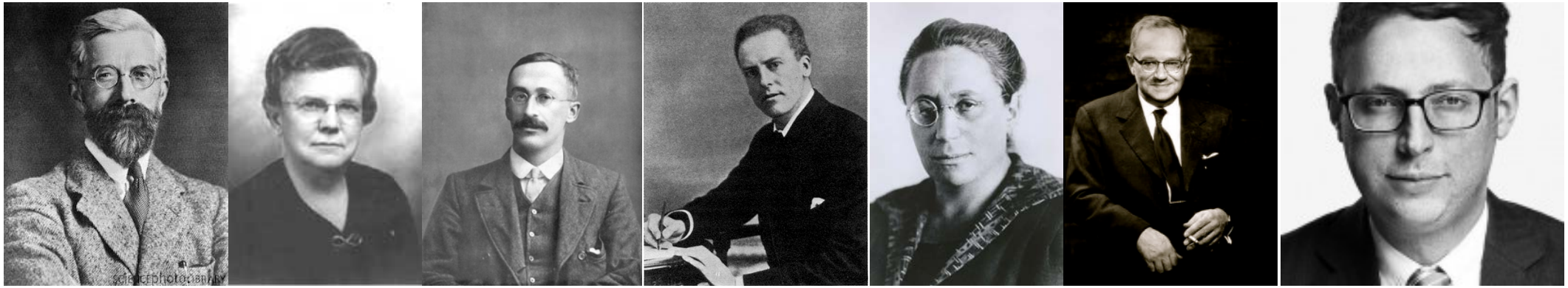
  b. But on average, estimates generated in this fashion will equal μ, since $M = \bar{Y}$ is an unbiased estimator of μ (or $E(M) = \mu$).

Blessed estimators generate, by definition, blessed estimates

- We bless estimates not because we know them to be specifically praiseworthy, but rather because we praise the process/rule/estimator that generated the estimate.

Or put differently:

- We can say something about the quality of estimator… but we don't have much to say about the quality of specific estimates, unless of course, we perhaps know something about the representativeness of our sample.



Confidence intervals as interval estimators

- ***Interval estimators***: Estimates are intervals (a range of values) rather than points (specific values).
- The most common interval estimator is the ***Confidence Interval***:

  a. $L = l\left(Y_1, Y_2, \ldots Y_n\right)$, a point estimator; the lower bound of the confidence interval

  b. $U = u\left(Y_1, Y_2, \ldots Y_n\right)$ is its upper bound.

  c. L and U are both random variables, taking on different values with different samples

  d. The randomly generated confidence interval [L,U] is an ***interval estimator***

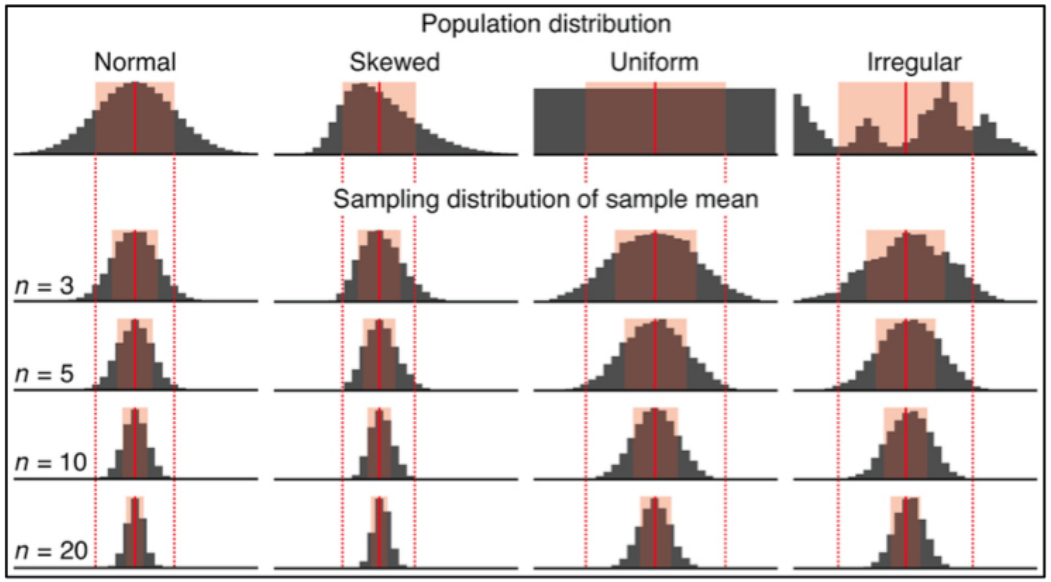
  e. If say, 95% of the time intervals generated in this fashion contain the true mean μ, then we say that we have a **95% *Confidence Interval*** (estimator) for μ.

  f. Is μ in the particular CI given your sample, $\left[\hat{L}, \hat{U}\right]$? ***No idea! But 95% of the time it will be!***

Sample statistics as estimators

- ***Mean***: The ***sample mean***, $\bar{Y} = \frac{1}{n}\sum Y_i$, is an unbiased estimator (in fact, it's BLUE) of the mean of Y, μ . $E(\bar{Y}) = E(Y) = \mu$.

  a. The particular estimated sample mean is $\bar{y} = \frac{1}{n}\sum y_i$.



More Estimators: Sample variance and standard deviation

- ***Variance***: The ***sample variance***, $S_{YY} = S_Y^2 = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2$

  a. *A*n unbiased estimator of the variance of Y, σ^2, when the mean of Y is unknown; $E(S_{YY}) = Var(Y) = \sigma^2$.

  b. We divide by n-1 to generate an unbiased estimator.

  c. The particular estimated sample variance is $S_{yy} = S_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$.

- ***Standard deviation***: The **sample standard deviation** is the square root of the sample variance, $S_Y = \sqrt{S_{YY}} = \sqrt{\frac{1}{n-1}\sum (Y_i - \bar{Y})^2}$.

  a. Generally a biased estimator of the standard deviation of Y, σ_Y.

  b. The particular estimated standard deviation is $S_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2}$.

Even More Estimators: Sample covariance & correlation

- ***Covariance***: The ***sample covariance***, $S_{XY} = \frac{1}{n-1}\sum (X_i - \bar{X})(Y_i - \bar{Y})$

  a. It is an unbiased estimator of the covariance of X and Y, $E(S_{XY}) = Cov(X,Y) = \sigma_{XY}$, when the means of X and Y (μ_X *and* μ_Y) are unknown.

  b. The particular estimated sample covariance is $S_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$.

- ***Correlation***: The ***sample correlation*** estimator, $\rho_{XY} = \frac{S_{XY}}{S_X S_Y}$

  a. Generally a biased estimator of the correlation of X and Y, $corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.

The particular estimated sample correlation is $\rho_{xy} = \frac{S_{xy}}{S_x S_y}$.