10/30/18
Wave
Moretto

Lecture #10: Measurement Error

- R-55 measurement error

$$\begin{bmatrix} 16\text{ oz} \\ 16 \\ \vdots \\ 16 \end{bmatrix} \xrightarrow[\text{for more sig figs}]{\text{turn dial}} \begin{bmatrix} 16.0\text{ oz} \\ 16.0 \\ \vdots \\ 16.0 \end{bmatrix} \rightarrow \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \begin{bmatrix} 15.97 \\ 16.01 \\ \vdots \\ \ \end{bmatrix} \updownarrow n$$

deterministic

probabilistic
(stochastic)

- Basic measurement error model

$y_1 = (\text{true value}) + (\text{bias}) + (\text{random error})_1$

$y_2 = (\text{true value}) + (\text{bias}) + (\text{random error})_2$

$\vdots$

$y_n = (\text{true value}) + (\text{bias}) + (\text{random error})_n$

IID
mean = 0
SD = σ

- Bias is a systematic tendancy to over or underestimate the truth
- Unbiased = (bias = 0) (no bias)

$$\underset{\text{pop}}{\begin{bmatrix} \ \\ \ \\ \ \end{bmatrix}} \xrightarrow[\text{random (IID)}]{\text{at}} \underset{\text{sample}}{\begin{bmatrix} \ \\ \ \\ \ \end{bmatrix}}$$

$15.97 = 16 + 0 - 0.03$
$16.01 + 16 + 0 + 0.01$
$\vdots$
$15.99 = 16 + 0 - 0.01$

unobservable

(true value) (bias) (random error)

observable

$y_1 = \theta + b + e_1$

$y_2 = \theta + b + e_2$

$\vdots$

$y_n = \theta + b + e_n$

$$\begin{bmatrix} 15.97 \\ 16.01 \\ 15.99 \end{bmatrix} \quad \theta = 16\text{ oz} \quad b = 0$$

sample mean $\rightarrow \bar{y} = \theta + b + \bar{e}_n$

$$\frac{e_1 + e_2 + \cdots + e_n}{n} = \frac{-0.03 + 0.1 + \ldots - 0.1}{n}$$

- cancellation of (+) and (−) errors will yield an $\bar{e}_n$ that is highly likely to be closer to 0 than any of the errors $e_1 \ldots e_n$ themselves

$$\bar{y}_n = \theta + b + \bar{e}_n$$

(sample mean) = (truth) + (bias) + (mean random errors)

- as $n\uparrow$ $\bar{e}_n \to 0$ highly likely
- Therefore as $n\uparrow$, $\bar{y}_n \to \theta + b$
- $\bar{y}_n \to$ (truth) θ, only when bias = 0
  - "get more good data"
    - unbiased

- 1936 FDR vs. Landon

Literary digest | 12 mil letters
 | 2.5 mil replies

result: Landon 60%, FDR 40% prediction

actual: FDR 60%, Landon 40%

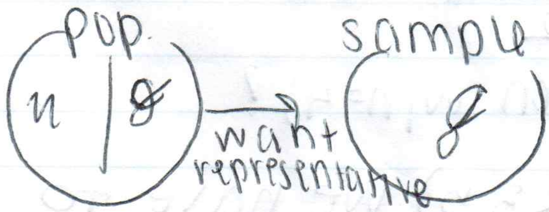
-20% point error

1 = FDR 0 = Landon

$$\bar{Y}_n = \underbrace{\theta + b}_{60\%\ (-20\%)} + \cancel{\bar{e}_n}$$

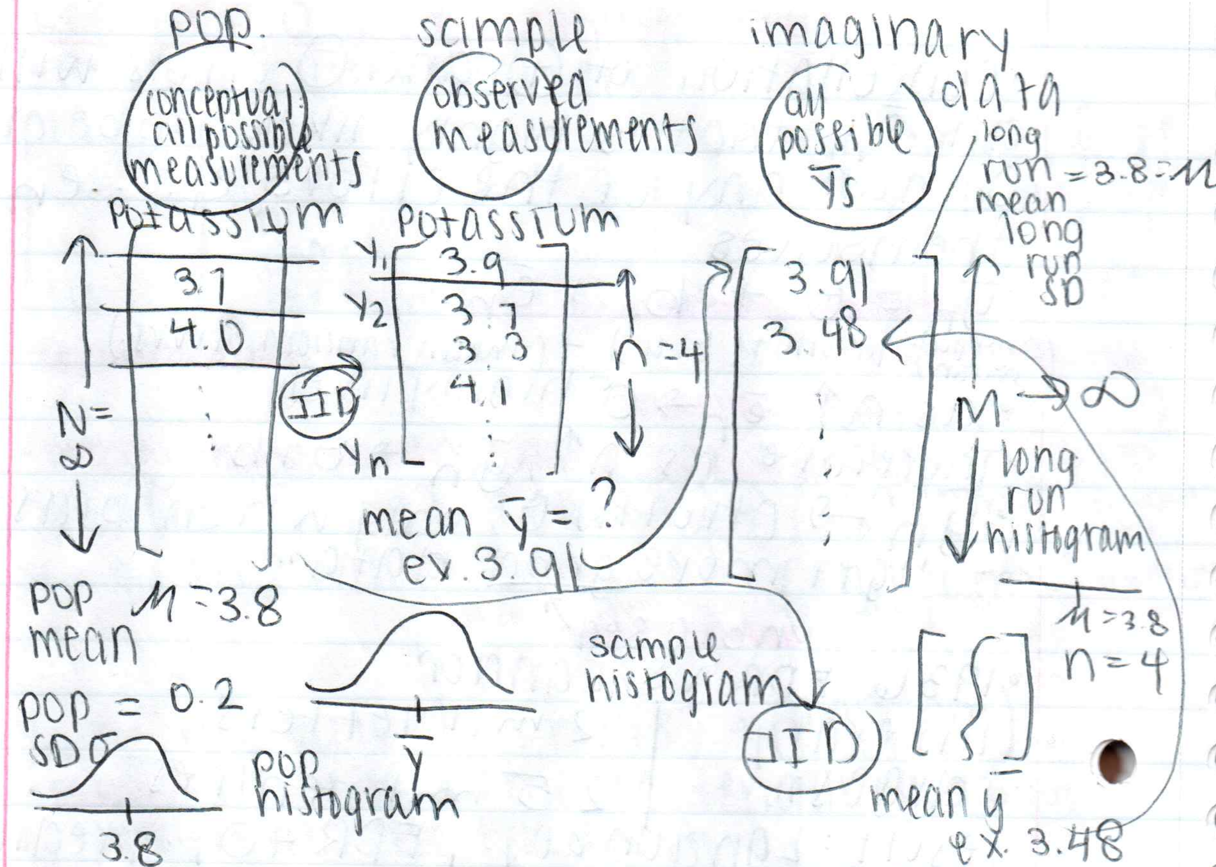
$$\begin{matrix} Y_1 \\ \vdots \\ Y_n \end{matrix} \begin{bmatrix} 1s \\ 0s \\ \\ \end{bmatrix} \quad n = 2.5\text{ mil}$$

mean 40%



- Phonebooks, club membership (country clubs)
  - oversampled rich ppl → Republican

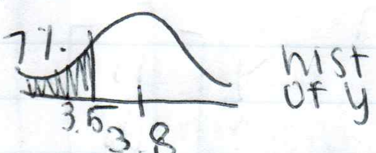
Probability model for measurement error



P(misclassification w| n=1) = 7%

too high

SD = 0.2



$$\frac{3.5-3.8}{0.2} = -1.5$$

P(misclassification w| n=4)?
$= P(\bar{y} < 3.5)$

- To estimate $P(\bar{y} < 3.5)$ we have to imagine getting lots of $\bar{y}$s and compute % of time $\bar{y} < 3.5$ in those repetitions

- Expected value $\bar{Y}$ = EV of $\bar{Y}$ = $E_{IID}(\bar{y}) = \mu = 3.8$