Stats 7 - class 12

11-05

R-55 to decrease uncertainty, get (more) good data

$\hookrightarrow$ unbiased measuring

$$\begin{bmatrix} 16oz \\ 16 \\ \vdots \\ 16 \end{bmatrix} \quad \begin{bmatrix} 16.0 \\ 16.0 \\ \vdots \\ 16.0 \end{bmatrix} \quad \begin{bmatrix} 15.98 \\ 16.01 \\ 16.00 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \updownarrow n$$

deterministic

probabilistic (stechastic)

(obs 1) y_1 = True value + bias + random error #1

(obs 2) y_2 = true value + bias + random error #2

$\vdots$

(obs n) y_n = true value + bias + random error n

systematic tendency to over/under-estimate the truth

IID draws from a population with mean 0 and SD σ

$$y_1 = \theta + b + e_1$$

$$y_2 = \theta + b + e_2$$

$$y_n = \theta + b + e_n$$

take the mean

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

Is $\bar{y}$ always likely to be closer to θ than (e.g.) y_1?

$$\begin{bmatrix} 15.98 \\ 16.01 \\ 16.00 \\ \vdots \\ 16.03 \end{bmatrix} \updownarrow 10$$

$15.98 = 16.0 + 0 + (-.02)$

$16.01 = 16.0 + 0 + .01$

$\vdots$

$16.03 = 16.0 + 0 + .03$

suppose truth $\theta = 16.0$
& no bias $(b=0)$

$\bar{y} = 16.0 + 0 + \bar{e}$

$$\bar{e} = \frac{-.02 + .01 + \ldots + .03}{10}$$

cancellation of $\oplus$, $\ominus$ errors

- $\bar{e}$ is highly likely to be a lot closer to 0 than any one of the errors $e_1 \ldots e_n$

- as $n \uparrow$, $\bar{y}_n \rightarrow \theta + b$
  $\bar{e}_n \rightarrow 0$

- Only way measurement error can be ~~da~~ 0 is to make measuring process unbiased

---

Literary Digest Poll
2.1 million replies

$$\begin{bmatrix} 1s \\ \cancel{0s} \\ 0s \end{bmatrix} n = 2.1 \text{ million}$$

FDR got 58% of votes

wrong by 18 percentage points

mean $.4 = 40\%$

bias in LD poll:

- How did they get the addresses?
- Response bias (12.1mil ↓ to 2.1mil)

- used phone books (x)
- club membership lists (country(golf)) (x)

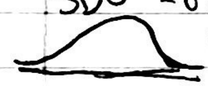
midterm prob 4

(possible blood) pop, $n = \infty$ $\xrightarrow{IID}$ (blood) (potassium) sample, $n = 4$

imag. data set (possible values of $\bar{y}$): 3.9, 3.65, ⋮ ; $M \to \infty$

mean $\mu = \theta$ truth $\mu = 3.8$ | mean $\bar{y}$ | long run mean / expected value of $\bar{y}$

SD $\sigma = 0.2$ | | long run SD



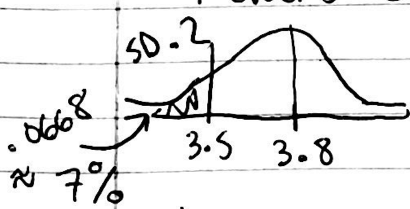
long run hist

P(missclassification) = 7% ← error rate

n=1

P(______)
n=4

P(mis-class, n=1) = P(y_1 < 3.5)



hist of y_1 = pop hist

EV of $\bar{y} = E_{IID}(\bar{y}) = \frac{E_{II}(S)}{n}$

$= \mu = 3.8$

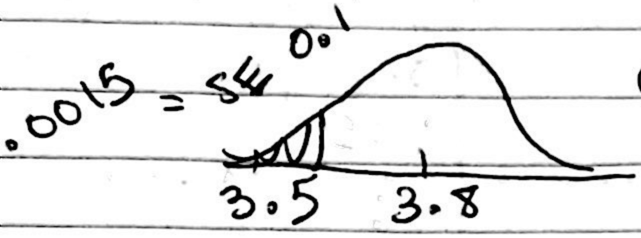
z

$\frac{3.5 - 3.8}{.2} = -1.5$

$$SE_{IID}(\bar{y}) = SE_{IID}\left(\frac{S}{n}\right) = \frac{1}{n} SE_{IID}(S)$$

$$= \frac{1}{n}(\sigma\sqrt{n}) = \boxed{\frac{\sigma}{\sqrt{n}}} = \frac{\text{pop SD}}{\sqrt{\text{\# of draws}}}$$

$$\frac{0.2}{\sqrt{4}} = 0.1$$



.15% n=4

$$\frac{3.5-3.8}{0.1} = -3$$

n	P(error)	Cost
1	7%	$25
4	0.15%	$100

benefit