STAT 7
5 Nov 19

this time: measurement error;
next time: statistical inference

quiz 5 due tonight

midterm 1 due Sun night

quiz 6 due a week from now

R-55

to decrease your uncertainty,

get (more) **good** data

unbiased measuring

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

deterministic   probabilistic ("stochastic")

bias: systematic tendency to over- or under-estimate the truth

$\binom{\text{obs.}}{\#1}\ Y_1 = \text{true value} + \text{bias} + \left(\begin{array}{c}\text{random}\\ \text{error}\\ \#1\end{array}\right)$

$\binom{\text{obs.}}{2}\ Y_2 = \text{true value} + \text{bias} + \left(\begin{array}{c}\text{random}\\ \text{error 2}\end{array}\right)$

$\vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$

$\binom{\text{obs.}}{n}\ Y_n = \text{true value} + \text{bias} + \left(\begin{array}{c}\text{random}\\ \text{error } n\end{array}\right)$

IID draws from a population with mean 0 and SD σ

$$Y_1 = \theta + b + e_1$$

$$Y_2 = \theta + b + e_2$$

$$\vdots$$

$$Y_n = \theta + b + e_n$$

take mean

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

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

$$\begin{bmatrix} 15.98 \\ 16.01 \\ 16.00 \\ \vdots \\ 16.03 \end{bmatrix} \quad 10 \qquad \begin{aligned} 15.98 &= 16.0 + 0 + (-0.02) \\ 16.01 &= 16.0 + 0 + (+0.01) \\ &\vdots \\ 16.03 &= 16.0 + 0 + (+0.03) \end{aligned}$$

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

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

$$\bar{e}_n = \frac{(-0.02) + (+0.01) + \dots + (0.03)}{10}$$

cancellation of $+$, $-$ errors:

math fact: $\bar{e}_n$ is highly likely to be a lot closer to 0 than any one of the $e_1, \dots, e_n$

math fact: as $n \uparrow$, $\bar{y}_n \to \theta + b$

$\bar{e}_n \to 0$

bottom line: only way we can damp measurement error down to 0 is to make measuring process unbiased

1936 D: FDR (low income) R: Alf Landon

Literary Digest poll favored by

12 million addresses

2.1 million replies

actual result: FDR got 58% of vote 18 percentage points error

FDR?

$\left[\begin{matrix} 1s \\ \& \\ 0s \end{matrix} \right]$

$n = 2.1$ million

mean $0.4 = 40\%$

bias in LD poll:

- how get addresses?
- response bias ($12.1m \to 2.1m$)

(1) phone books (X)

(2) club membership lists (X)

↓ country (golf)

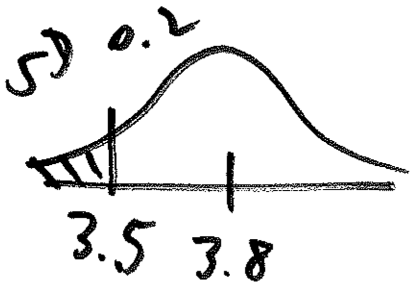
$P(\text{mis-classification}) \doteq 7\%$

$n = 1$

error rate

$P\left(\frac{\qquad}{n=4}\right) = P(\bar{y} < 3.5) = 0.15\%$

$P(\text{mis-class}, n=1) = P(\underline{\underline{y_1}} < 3.5)$



hist of y_1 = pop. hist.

pop.
all possible blood samples from you
potassium

sample (midterm problem 4)
the observed blood samples from you
potassium

imag. data
possible values of $\bar{y}$

$N = \infty$

actual
like IID

$n = 4$

sum $S = ?$

truth
mean $\bar{y} = ?$
(ex. 3.9)

3.9
3.65
⋮

$M \to \infty$

hyp. IID

mean $\mu = \theta = 3.8$

SD $\sigma = 0.2$

pop. hist.

$n = 4$

mean $\bar{y}$
(ex. 3.65)

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

long run SD | standard error of $\bar{y}$ $= \frac{0.2}{\sqrt{4}} = 0.1$

long-run hist

(*)

$\left(\text{expected value of } \bar{y} \right) = \left(\text{EV of } \bar{y} \right) = E_{IID}(\bar{y})$

$= E_{IID}\left(\frac{S}{n}\right) = \frac{E_{IID}(S)}{n}$

$= \frac{\cancel{n} \cdot \mu}{\cancel{n}} = \mu$ ✓

$.0668 \doteq 6.7\% \doteq 7\%$

SD 0.2

3.5 3.8

hist. of y,

z $(3.5 - 3.8)/0.2 = \frac{-0.3}{0.2} = -1.5$

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

SE of $\bar{y}$

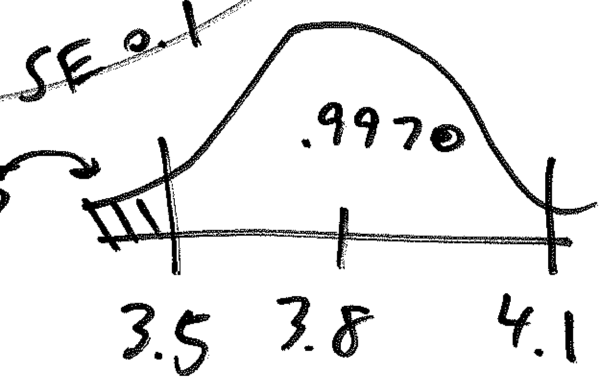
$$= \frac{1}{n} SE_{IID}(S)$$

$$= \frac{1}{n}(\sigma \sqrt{n})$$

$$\boxed{SE_{IID}(\bar{y}) = \frac{\sigma}{\sqrt{n}}} = \frac{(\text{1-Y SD})}{\sqrt{(\#\text{ draws})}}$$

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

0.15% = .0015



i.e. dsn hist of $\bar{y}$, n=4

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

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

benefit