Measurement Error, Bias, Probability models for means 11/5/19

Quiz 5 due Tonight
Midterm due sun night
Quiz 6 due a week from now R-55

---

to decrease your uncertainty, get (more) good data

unbiased measuring

$$\begin{bmatrix} 16 \text{ oz} \\ 16 \\ \vdots \\ 16 \end{bmatrix} \quad \begin{bmatrix} 16.0\cancel{0} \text{ oz} \\ 16.0 \\ \vdots \\ 16.0 \end{bmatrix} \quad \begin{bmatrix} 15.98 \text{ oz} \\ 16.01 \\ 16.00 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} n$$

deterministic

probabilistic ("stochastic")

Bias: systematic tendency to over or under estimate the truth

Scanned with CamScanner

$\begin{pmatrix}\text{obs.}\\ \#1\end{pmatrix}$ $y_1 =$ true value + bias + (random error #1)

$\begin{pmatrix}\text{obs.}\\ \#2\end{pmatrix}$ $y_2 =$ true value + bias + (random error #2)

⋮

$\begin{pmatrix}\text{obs.}\\ n\end{pmatrix}$ $y_n =$ true value + bias + (random error n)

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 mean

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

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

Scanned with CamScanner

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

$$15.98 = 16.0 + 0 + (-0.02)$$
$$16.01 = 16.0 + 0 + (+0.01)$$
$$\vdots$$
$$16.03 = 16.0 + 0 + (+0.03)$$

suppose truth $\theta = 16.0$ &

~~bi~~ no bias ($b = 0$)

$\bar{y}_n = 16.0$

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

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

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

$\boxed{\text{MATH FACT}}$ $\bar{e}_n$ is highly likely to be a lot closer to 0 than any one of the $e_1, \ldots, e_n$

$\boxed{\text{MATH FACT}}$ as $n \uparrow$, $\bar{y}_n \to \theta + b$

$\bar{e}_n \to 0$

Scanned with CamScanner

BOTTOM LINE

only way we can damp measurement over down to 0 is to make measuring process unbiased

Literary Digest pool

1936 D: FDR (Favored by low income)

R: Alf London

12 million addresses

2.1 million replies	Actual result	FDR
	FDR got 58% of vote	1s
		2
points' error	18%	0s

n = 2.1 million

mean 0.4 = 40%

Scanned with CamScanner

Bias in LD poll

- how get addresses?
- Response Bias (12.1m → 2.1m)

① phone books (x)

② club membership lists (x)
↓
country (golf)

P(misclassification) = 7%

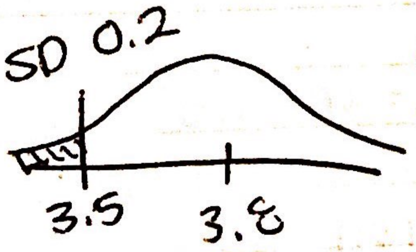
n=1

error rate

$P(\overline{\quad n=4 \quad}) = p(\bar{y} < 3.5)$

$= 0.15\%$

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



hist of y_1 = pop hist

Midterm problem 4

<u>population</u>	<u>sample</u>	<u>imag. data</u>
all possible blood samples from you potassium	the observed blood samples from you potassium	possible values of $\bar{y}$

$n = \infty$

actual

~~like IID~~

$n = 4$

3.9

3.65

$M \to \infty$

sum $s' = ?$

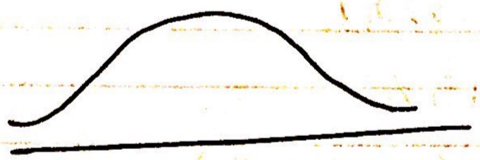
truth mean $\bar{y} = ?$

(ex 3.9)

mean $\mu = \theta = 3.8$

hyp IID

SD $\sigma = 0.2$



pop. hist.

$n = 4$

mean $\bar{y}$

(ex. 3.65)

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

long run standard SD error of $\bar{y}$

$= \frac{0.2}{\sqrt{4}} = 0.1$ long-run hist

(*)

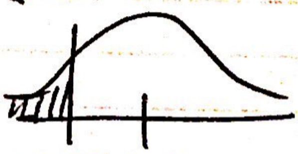
$$\begin{pmatrix}\text{expected}\\ \text{value of}\\ \bar{y}\end{pmatrix} = \begin{pmatrix}\text{EV of}\\ \bar{y}\end{pmatrix} = E_{IID}(\bar{y}) = E_{IID}\left(\frac{S}{n}\right) = \frac{E_{IID}(S)}{n}$$

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

$0.0668 = 6.7\% = 7\%$

SD 0.2

hist of y_1



3.6 3.8

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

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

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

SE of $\bar{y}$

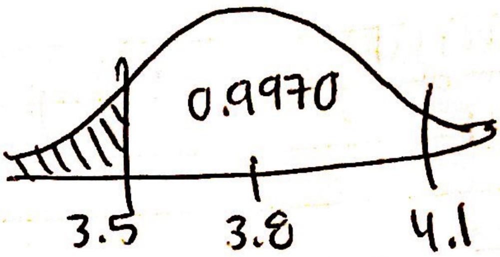
$$= \frac{1}{n}(\sigma\sqrt{n})$$

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

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

SE 0.1

0.15% = 0.0015



long run hist of $\bar{y}$

EV

$n=4$

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

n	P(char)	Lost
1	7%	\$25
4	0.15%	\$100

benefit