AMS 7 - 5/3/18

THIS TIME: Prob. Models for means

NEXT TIME: Inference

* Strike next week → will be updated by email
* Case study (medicine): Hypokalemia

R-55 → How to cope with uncertainty:

GET MORE GOOD (Unbiased) DATA!

* Blocks of Butter (weight/mass) in oz

Non-random (Deterministic)

$$\begin{bmatrix} 16 \\ 16 \\ 16 \\ \vdots \\ 16 \end{bmatrix}$$

Non-random (Deterministic)

$$\begin{bmatrix} 16.0 \\ 16.0 \\ 16.0 \\ \vdots \\ 16.0 \end{bmatrix}$$

$$\begin{bmatrix} 16.03 \\ 15.99 \\ 15.95 \\ \vdots \end{bmatrix}$$

Probabilistic (random) (stochastic)

* Astronomy (1600s)
  - Basic measurement error model

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

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

$$\vdots$$

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

IID draws from a normal curve with mean 0 & SD σ same

theta

$$y_1 = \theta + b + e_1$$
$$y_2 = \theta + b + e_2$$
$$\vdots$$
$$y_n = \theta + b + e_n$$

avg

$$\bar{y} = \theta + b + \begin{pmatrix} \text{mean of} \\ n \text{ IID} \\ \text{draws} \\ \bar{e} \end{pmatrix} \text{(mean)}$$

each with mean 0

(oz) Unbiased

$$16.03 = 16.0 + 0 + (+.03)$$
$$15.99 = 16.0 + 0 + (-0.01)$$
$$\vdots$$
$$15.95 = \theta + b + (-0.05)$$

$$\frac{(+0.03) + (-0.01) + \dots + (-0.05)}{n}$$

(2)

→ $\bar{e}$ will (with high prob.) be closer to 0 than any of the e_i

truth + bias

→ as $n\uparrow$, $\bar{e}\downarrow$ to 0 (with high prob), therefore, $\bar{y}$ will be close, when n is large, to $(\theta + b)$

→ To make $\bar{y}$ get arbitrarily close to Θ, we need 2 things:

1) n should get ↑ and...
2) $b = 0$ (measuring process is unbiased)

* you cannot make bias ↓

▷ Literary Digest (LD) - Surveying Organization (Magazine)

- 1936 - Rosevelt (D) vs. Landon (R)
  - ↳ 24,000 letters (Pre-stamped post card)
    Got back 16,000 post cards
    LD estimated that Landon 60% vs Rosevelt 40%.
  - ↳ **truth:** Landon 40% vs. Rosevelt 60%
    * 20-percentage point error
  - ↳ **George Gallup** (Iowa State) (1,000 people)
    - found out real values & how wrong LD was
    - LD data was deeply biased
      * Addresses (1936)
        - landowner records
        - telephone books
        - club membership lists

        Republicans
        RICH people

▷ Case study: Medicine (HYPOKalemia)

(conceptual) Population	Sample	Imaginary Data Set
All possible measur.	the obs. measurements	Hyp. repetitions
Potassium level	Potassium level	All possible $\bar{y}$ values

$$N = \infty \quad \left[\ \cdot\ \right] \xrightarrow[\text{IID}]{\text{like}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_4 \end{bmatrix} \begin{array}{l} n=4 \\ \bar{y}=? \end{array} \qquad \begin{array}{c} \begin{bmatrix} 3.92 \\ 3.77 \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \bar{y}_1 \\ \bar{y}_2 \\ M\to\infty \end{array}$$

M = long run mean

L-133

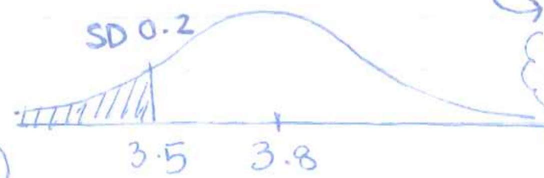
Population

mean $\mu = 3.8$

SD $\sigma = 0.2$

Population Histogram

↳ Histogram of one reading at time (y_1)



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

P(misdiagnosis w/ n=1)

↳ $\doteq 7\%$

P(misdiagnosis w/ n=4)

↳ $P(\bar{y}_{(n=4)} < 3.5) = ?$

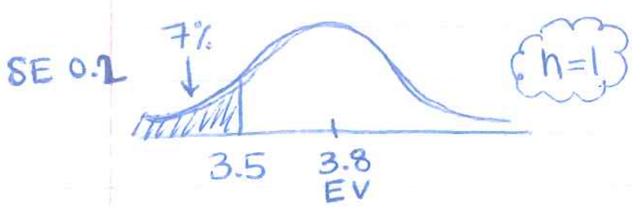
high

Imaginary Data Set

* long run mean → expected value of $\bar{y}$

$$\left(\text{expected value of } \bar{y}\right) = \left(\text{EV of } \bar{y}\right) = \underbrace{E_{IID}(\bar{y}) = \mu}_{\text{MATH FACT}} = 3.8$$

long run histogram $(\bar{y})$



long run SD

↳ standard error of $\bar{y}$

$(\text{SE of } \bar{y}) = SE_{IID}(\bar{y}) = ?$

Ingredients	belong?
N	X
μ	X
σ	$\sigma\uparrow$, $SE(\bar{y})\uparrow$
n	$n\uparrow$, $SE(\bar{y})\downarrow$

* Square root law:
  to cut $SE(\bar{y})$ in half, need to quadruple n

L-124

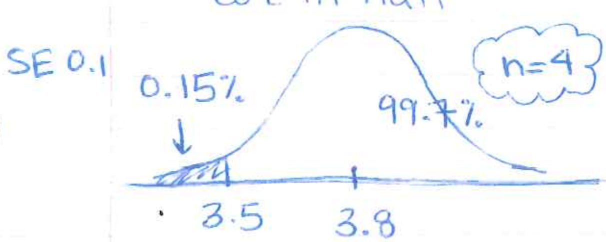
CLT

$SE(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{\sqrt{4}} = 0.1$

↳ cut in half

$\cancel{\frac{\sigma}{n}}$ → Perfect world

$\frac{\sigma}{\sqrt{n}}$ → uncertainty goes down w/ n, but only at a $\sqrt{n}$ rate (more data)



$(3.5 - 3.8)/0.1 = -3$

* This is called the... SQUARE ROOT LAW

4

n	P(misdiagnosis)	cost
1	7%	$25
4	0.15%	$100

*cost-benefit trade off
In this case, 1st row better

→ Just eat a banana!
Save the money b/c not that bad of a misdiagnosis in this case.