STAT 131
27 May 20

this time: useful distributions

next time: catalog

read JS ch. 5

catch-up lecture

people in U.S. $\doteq$ 328M

eligible & motivated voters $\doteq$ 120M = T

P

θ = population proportion favoring Clinton, 2 weeks before election

if can achieve random or like-at-random sampling from P

pop. all voters, favor C?

$T \doteq \begin{bmatrix} 1s \\ \vdots \\ 0s \end{bmatrix}$ 120M — mean θ

like-at-random →

the observed voters sample, favor C?

$\begin{bmatrix} 1s \\ 0s \end{bmatrix}$ n — mean $\hat{\theta}$, sum S

IID

$\hat{\theta} = \frac{S}{n}$

$(S \mid n, \theta) \sim$ Binomial (n, θ)

$$E(\hat{\theta}) = E\left(\frac{S}{n}\right) = \frac{1}{n} \underbrace{E(S)}_{n\theta} = \theta$$

So $\hat{\theta}$ is an unbiased estimate of θ

$$V(\hat{\theta}) = V\left(\frac{S}{n}\right) = \frac{V(S)}{n^2}$$

$$= \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} \quad \text{and}$$

$$SD(\hat{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}} \quad \text{as } n \uparrow$$

$SD \uparrow$ uncertainty about θ based on $\hat{\theta}$

$\downarrow$ as $n \uparrow$

$= O\left(\frac{1}{\sqrt{n}}\right)$

square root law

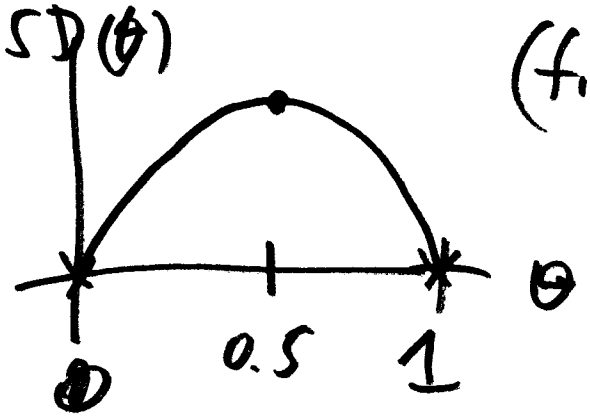
to cut $SD(\bar{\theta})$ in half, need to quadruple the sample size

---

Q: How big does n have to be to get $SD_{IID}(\bar{\theta}) \doteq 1\%$?

A: $SD_{IID}(\bar{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}} = .01$

given 2 candidates & polarization of electorate, $\theta = 0.5$



(fix n)

$$SD_{\substack{IID \\ worst\ case}}(\bar{\theta}) = \sqrt{\frac{0.5 \cdot (1-0.5)}{n}}$$

$$= \frac{0.5}{\sqrt{n}}$$

$$\frac{0.5}{\sqrt{n}} = .01 \rightarrow n = \left(\frac{0.5}{0.01}\right)^2 = 2500\ (!)$$

$T = 120m$

$n_{\text{worst case IID}} = n_{\text{worst case SRS}}$

big issue, however: really hard to achieve like-at-random

~~(T, C)~~ (T, C, no response)

$$RMSE(\hat{\theta}) = \sqrt{V_{IID}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2}$$

(error due to ~~unlucky~~ random sampling)

(if $\hat{\theta}$ is biased) $= SD_{IID}(\hat{\theta})$ with no bias

Literary Digest (LD) poll

1936 < (D) FDR
(R) Alf Landon

---

in summer of 1936, mailed out 10M letters, only 2.27M responded;

~~result~~: $\hat{\theta}$ = 60% Landon, truth: 59% R(!)
estimate

---

error: 19 percentage points (!)

---

how is this possible with $n = 2.27M$?

$$SD_{(IID)}(\hat{\theta}) = \sqrt{\frac{(0.6)(0.4)}{2.27M}} \doteq .00025 \doteq .03\%$$

(IID ← their mistake)

$$RMSE(\hat{\theta}) = \sqrt{\underbrace{V_{IID}(\hat{\theta})}_{\doteq 0} + \underbrace{(bias(\hat{\theta}))^2}_{\doteq .19}}$$

($\hat{\theta}$ ← used)

George Gallup: correctly estimated GD result to within 1 percentage point before election, correctly estimated FDR's vote share was about 59%

GD sources of bias

① non-response bias: people who ② chose to mail postcard back were systematically different in voting behavior than people who chose not to

② how obtain list of valid addresses in

1936? phonebooks → not every body had phone. bias toward wealthy → London (R)

+ club membership lists → country clubs, ... bias toward wealthy → London (R)

$$RMSE(\hat{\theta}) = \sqrt{V_{IID}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2}$$

→ 0 as $n \uparrow$

typical amount by which $\hat{\theta}$ misses θ

real enemy in (data) science

A sample is <u>representative</u> of the population from which it is taken

$\longleftrightarrow$

The sample is like-IID or at least like at-random