Quiz 2 case study | STAT 206 22 Jan 21 | AM disc. sec.

pop: all 18+ people in SC in mid-Jan 2021

sample: the observed people

1 = yes, 0 = not: full empl.?

full empl.?

$N \doteq 54k$

$\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$

(actually ~~like~~ IID)

$I_1 \; y_1$ ... $I_n \; y_n$ $\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$

$\underline{n = 921}$

mean $\theta = ?$

SD $\sigma = \sqrt{\theta(1-\theta)} = ?$

pop PMF =

0 1

$\text{mean} = \frac{S}{n} = \hat{\theta}_n = \bar{y}_n = \frac{830}{921} \doteq 0.901$

hypothetical IID

repeated sampling possible $\{\hat{\theta}_n\}$ ①

$\begin{pmatrix} 0.901 \\ 0.897 \\ 0.916 \\ \vdots \end{pmatrix}$

want big $M \to \infty$

long run mean

by WLLN ✓

$E_{IID}(\hat{\Theta}_n) = \theta$

$\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$ $n = 921$ (ex. 0.897)

hyp. IID

mean $\hat{\theta}_n = ?$

long run $SD_{IID}(\hat{\Theta}_n)$ ✱ (up to θ)

$\begin{pmatrix} \\ \\ \end{pmatrix}$ mean $\hat{\theta}_n$ 0.916

$\hat{\Theta}_n = \bar{I}_n = \frac{1}{n} \sum_{i=1}^{n} I_i$ RS $= S/n$

random variable story:

$(I_i \mid \theta) \overset{IID}{\sim} \text{Bernoulli}(\theta)$

$i = 1, \dots, n$

$(S \mid n, \theta) \sim \text{Binomial}(n, \theta)$

we're interested in $(\hat{\Theta}_n - \theta)$

with realized value $(\hat{\theta}_n - \theta)$;

$P_F(|\hat{\Theta}_n - \theta| \le \epsilon) = ?$ want this to be big

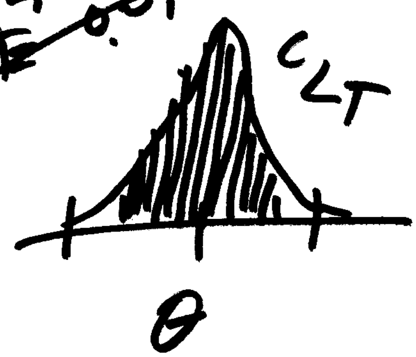
($\epsilon > 0$ small)

IID sample from the sample data set:

(frequentist) (Efron (1979)):

the bootstrap (*)

SE → 0.01



long run PMF/PDF of $\hat{\Theta}_n$

$E_{IID}(\hat{\Theta}_n) = EV_{IID}(\hat{\Theta}_n)$

STAT 131

$E_{IID}(\bar{Y}_n) = \theta$

STAT 131

$\underline{SD}_{IID}(\bar{Y}_n) = \frac{\sigma}{\sqrt{n}}$

⌜to cut $SD_{IID}(\bar{Y}_n)$ in half, need to quadruple (multiply by $2^2 = 4$) the sample size⌝

$$\widehat{SD}_{IID}(\bar{Y}_n) = \widehat{SD}_{IID}(\hat{\theta}_n) = \frac{\hat{\sigma}}{\sqrt{n}}$$

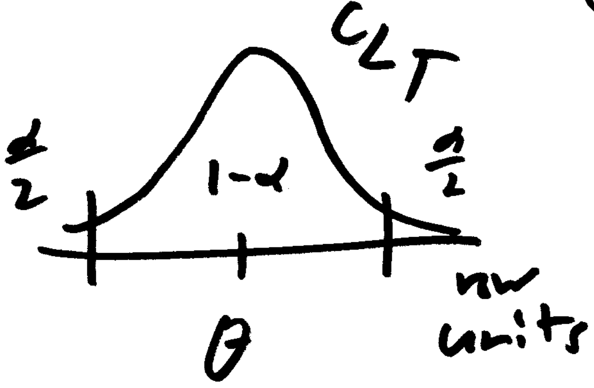
$$\hat{\sigma} = \sqrt{\hat{\theta}_n(1-\hat{\theta}_n)} \qquad = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \doteq \widehat{SE}_{IID}(\hat{\theta}_n)$$

standard error

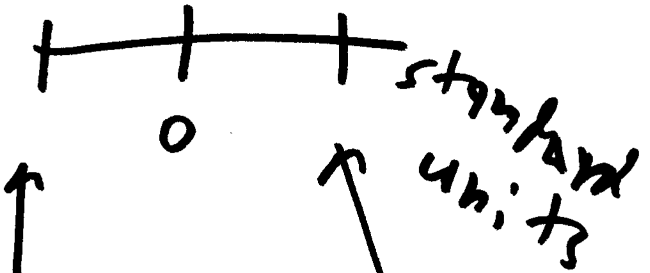
(3) we think θ is around $\hat{\theta}_n \doteq 0.901$ (observed value of $\hat{\theta}_n$), give or take about $\underset{0.01}{SE_{IID}(\hat{\theta}_n)}$

$$\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \doteq 0.0098$$

(*)



PMF/approx PDF of $\hat{\Theta}_n$



$100(1-\alpha)\%$ confidence interval (CI)

$$P_F\left(\theta - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot \hat{SE}(\hat{\Theta}_n) \le \hat{\Theta}_n \le \theta + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot \hat{SE}(\hat{\Theta}_n)\right) \doteq 1-\alpha$$

$$= P_F\left(\underbrace{\hat{\Theta}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot \hat{SE}}_{\text{random left endpoint}} \le \theta \le \underbrace{\hat{\Theta}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{SE}}_{\text{random right endpoint}}\right)$$

Mr. Neyman: $100(1-\alpha)\%$ CI for θ (nominal confidence level) in this sampling model

is defined to be

fixed known constant

$$\hat{\Theta}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

99.9% CI for θ $\quad \alpha = 0.001$

0.869 $\quad$ 0.901 $\quad$ 0.934

fixed unknown constant

$$P_F(0.869 \le \theta \le 0.934) = \cancel{0.999}$$

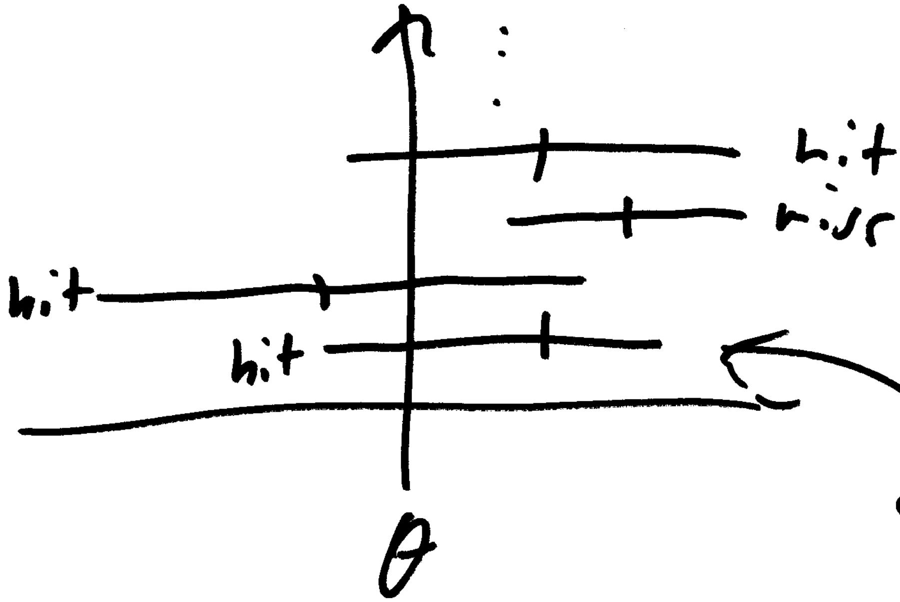
$$= \text{undefined}$$

fixed known constant

$$P_B(0.869 \le \theta \le 0.934 \mid \bar{y}, \text{low info. prior}) \doteq 0.999$$

Bayes: like a realization of a r.v.

✓



Mr. Neyman says you should have a hit rate (**)

each IID dataset yields this CI

(**) of about $100(1-\alpha)\%$

Mr. Neyman's confidence is in the process of CI construction, not the outcome of any one data-gathering activity (sample)