this time: probability & statistical
next time: models for means

Read: DD(8) ch. 11, LN pp. L-137 + L-144

AMS 7 10 May 17

1

today: LN p. L-128 +

I have an extra office hour today: 1.15–2.15 pm in Jack's Lounge

Hypokalemia case study

p. R-55

basic measurement error model:

$$\underbrace{\begin{pmatrix}\text{obs.}\\ 1\end{pmatrix}}_{Y_1} = \underbrace{\begin{pmatrix}\text{true}\\ \text{value}\end{pmatrix}}_{\theta} + \underbrace{(\text{bias})}_{b} + \underbrace{\begin{pmatrix}\text{random}\\ \text{error } 1\end{pmatrix}}_{e_1}$$

mean 0, SD $\sigma$, IID normal

$$\underbrace{\begin{pmatrix}\text{obs.}\\ 2\end{pmatrix}}_{Y_2} = \begin{pmatrix}\text{true}\\ \text{value}\end{pmatrix} + (\text{bias}) + \underbrace{\begin{pmatrix}\text{random}\\ \text{error } 2\end{pmatrix}}_{e_2}$$

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

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

---

$$\underbrace{\begin{pmatrix}\text{mean of}\\ n \text{ obs.}\end{pmatrix}}_{\bar{Y}} = \underbrace{\begin{pmatrix}\text{true}\\ \text{value}\end{pmatrix}}_{\theta} + \underbrace{(\text{bias})}_{b} + \underbrace{\begin{pmatrix}\text{mean of } n \text{ random errors}\\ \frac{e_1 + e_2 + \cdots + e_n}{n}\end{pmatrix}}_{\bar{e}}$$

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

mean of n obs. | truth | bias | mean of n ~~measure~~ random errors

assume that your measuring process is unbiased:

i.e., assume $b = 0$

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

ex. hypokalemia

pretend we know $\theta = 3.8$

obs

$$3.9 = 3.8 + 0 + (+0.1)$$
$$3.5 = 3.8 + 0 + (-0.3)$$
$$\vdots$$
$$3.8 = 3.8 + 0 + (+0.0)$$

$$\bar{y} = 3.8 + 0 + \frac{(+0.1) + (-0.3) + \dots + (+0.0)}{n}$$

we will get to "enjoy" cancellation of $\oplus$ & $\ominus$ errors, with the result that the typical size $|\bar{e}|$ of $\bar{e}$ will be smaller than any of the $e_i$

as $n \uparrow$, the expected value of $|\bar{e}|$ goes to $0$, so $\bar{y} \to \theta + b + 0$ as $n \to \infty$

i.e., if $b = 0$, $\bar{y}$ gets closer & closer to the truth ($\theta$) as $n$ increases

but if $b \neq 0$, $\bar{y} \to \theta + b$ not $\theta$: with a biased measuring process, you cannot make the bias go away just by taking more obs. & averaging them

$\left( \begin{array}{c} \text{to decrease} \\ \text{uncertainty} \\ \text{about } \theta \end{array} \right)$, get more & more good data

unbiased measuring process

pop

conceptual: all possible blood from pigs

pretend truth $\theta$ is 3.8

sample

the observed blood samples from peterson

from you

repeated sampling imaginary data set

potassium from you

$N = \infty$

at random with replacement IID

3.9
3.6
⋮

$n = 4$

mean $\bar{y} = ?$ (ex. 3.85)

3.85 $\leftarrow \bar{y}$
3.67
⋮

$M = \infty$

mean $\mu = \theta = 3.8$

SD $\sigma = 0.2$

hyp IID

$n = 4$

mean $\bar{y} = ?$ (ex. 3.67)

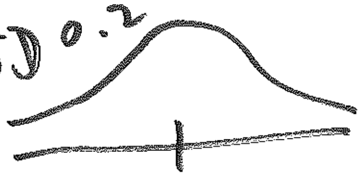
⋮

hyp IID

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

long run SD SE of $\bar{y}$ $= 0.1$

SD 0.2



pop hist

3.8

long run hist SE 0.1 CLT part III

3.5 3.8 EV

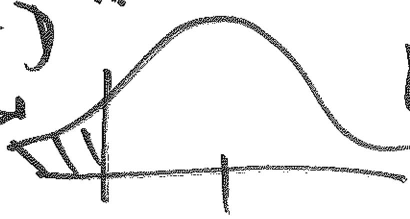
here $SE(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{\sqrt{4}} = \frac{0.2}{2} = 0.1$

$P$(misdiagnosis with $n = 1$)

$= P(Y_1 < 3.5) = 7\%$

7%

5) 0.2



i.e., rel hist (hist of any single $Y_i$)

3.5 3.8

$-1.50$

$$\frac{3.5 - 3.8}{0.2} = \frac{-0.3}{0.2} = -1.50$$

---

$P$(misdiagnosis with $n = 4$)

$= P(\bar{Y}$ based on $n=4$ obs. $< 3.5) = ?$

---

EV of $\bar{Y}$ = $\boxed{E_{IID}(\bar{Y}) = \mu}$

math fact

$E_{IID}(S) = n\mu$ but $\bar{Y} = \frac{S}{n}$

$$\text{SE of } \bar{Y} = \boxed{SE_{IID}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}}$$

~~math fact~~

⑥

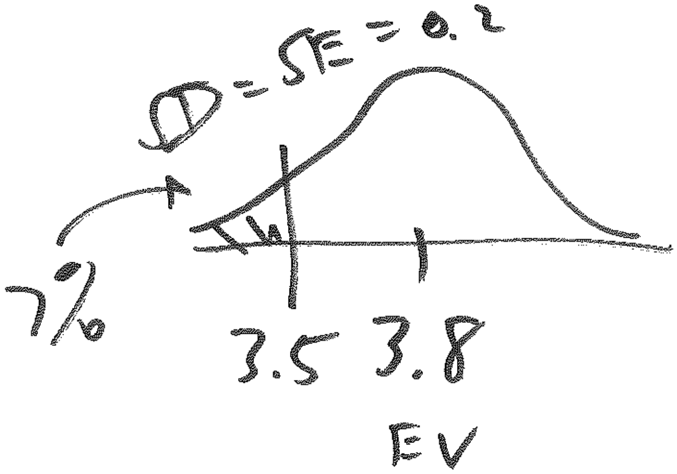
| | |
|---|---|
| $N$ | X |
| $\mu$ | X |
| $\sigma$ | $\sigma \uparrow$ $SE(\bar{Y}) \uparrow$ |
| $n$ | $n \uparrow$ $SE(\bar{Y}) \downarrow$ |
| $M$ | X |

$\bar{Y}$ is our best <u>estimate</u> of $\mu$; SE of $\bar{Y}$ represents how much uncertainty (noise) there is in the process of using $\bar{Y}$ to estimate $\theta$

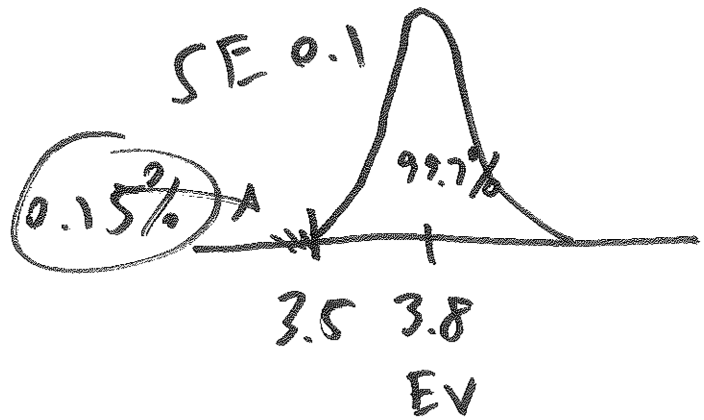
uncertainty about $\mu$ on the basis of $\bar{Y}$ as an estimate of $\mu$ goes <u>down</u> with $n$, but only at a $\sqrt{n}$ rate

the square root law

i.e., unfortunately, to cut $SE(\bar{Y})$ in half you have to quadruple the sample size




long run list of $\bar{y}$ with $n = 1$




long run list of $\bar{y}$ with $n = 4$

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

| n | cost | P(mis-diagnosis) |
|---|---|---|
| 1 | \$25 | 7% |
| 4 | \$100 | 0.15% |

cost benefit