this time: measurement error

next time: statistical inference

read: DD(B) ch. 11
LN pp. 137–156

AMS 7
30 Oct 18

today: LN pp. 127 → 1

R (55) measurement error

deterministic

$$\begin{bmatrix} 16.02 \\ 16 \\ \vdots \\ 16 \end{bmatrix}$$

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

probabilistic (stochastic)

$$\begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} \begin{bmatrix} 15.97 \\ 16.01 \\ \vdots \\ 1 \end{bmatrix} \; n$$

basic measurement error model

IID mean 0 SD σ

$$y_1 = (\text{true value}) + (\text{bias}) + (\text{random error})_1$$

$$y_2 = (\text{true value}) + (\text{bias}) + (\text{random error})_2$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots$$

$$y_n = (\text{true value}) + (\text{bias}) + (\text{random error})_n$$

bias a systematic tendency to over- or underestimate the truth

unbiased = (bias = 0) (no bias)

obs. #	(true value)	(bias)	(random errors)
y_1	$= \theta$	$+ b$	$+ e_1$
y_2	$= \theta$	$+ b$	$+ e_2$
⋮	⋮	⋮	⋮
y_n	$= \theta$	$+ b$	$+ e_n$

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

$\bar{y}$: sample mean

$\bar{e}_n$: $\dfrac{e_1 + e_2 + \dots + e_n}{n}$

$$\begin{bmatrix} 15.97 \\ 16.01 \\ \vdots \\ 15.99 \end{bmatrix}$$

$\theta = 16$ oz

$b = 0$

$15.97 = 16 + 0 + (-0.03)$

$16.01 = 16 + 0 + (+.01)$

$\vdots$

$15.99 = 16 + 0 + (-.01)$

$$\frac{e_1 + e_2 + \dots + e_n}{n} = \frac{(-.03) + (.01) + \dots + (-.01)}{n}$$

Cancellation of $\oplus$, $\ominus$ errors will yield an $\bar{e}_n$ that is highly likely to be closer to 0 than any of the errors $e_1, \dots, e_n$ themselves

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

$$\begin{pmatrix}\text{sample}\\ \text{mean}\end{pmatrix} = \overset{(\text{truth})}{\theta} + \overset{(\text{bias})}{b} + \begin{pmatrix}\text{mean of}\\ \text{random}\\ \text{errors}\end{pmatrix}$$

as $n \uparrow$ $\bar{e}_n$ is highly likely to $\to 0$

therefore as $n \uparrow$, $\bar{y}_n \to \theta + b$

$\bar{y}_n \to (\text{truth})\ \theta$ only when

bias $= 0$ "get more good data"

(1936) FDR vs. Landon — unbiased

Literary Digest	12 million letters 2.5 million replies

result: Landon 60% FDR 40%
prediction

actual: FDR 60% Landon 40%

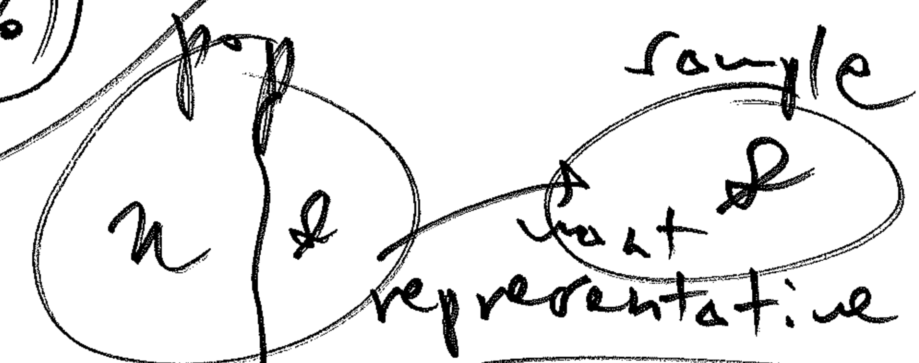
20 percentage point error

1 = FDR
0 = Landon

$y_1, \ldots, y_n$ [1s & 0s], n = 2.5 million, mean 40%

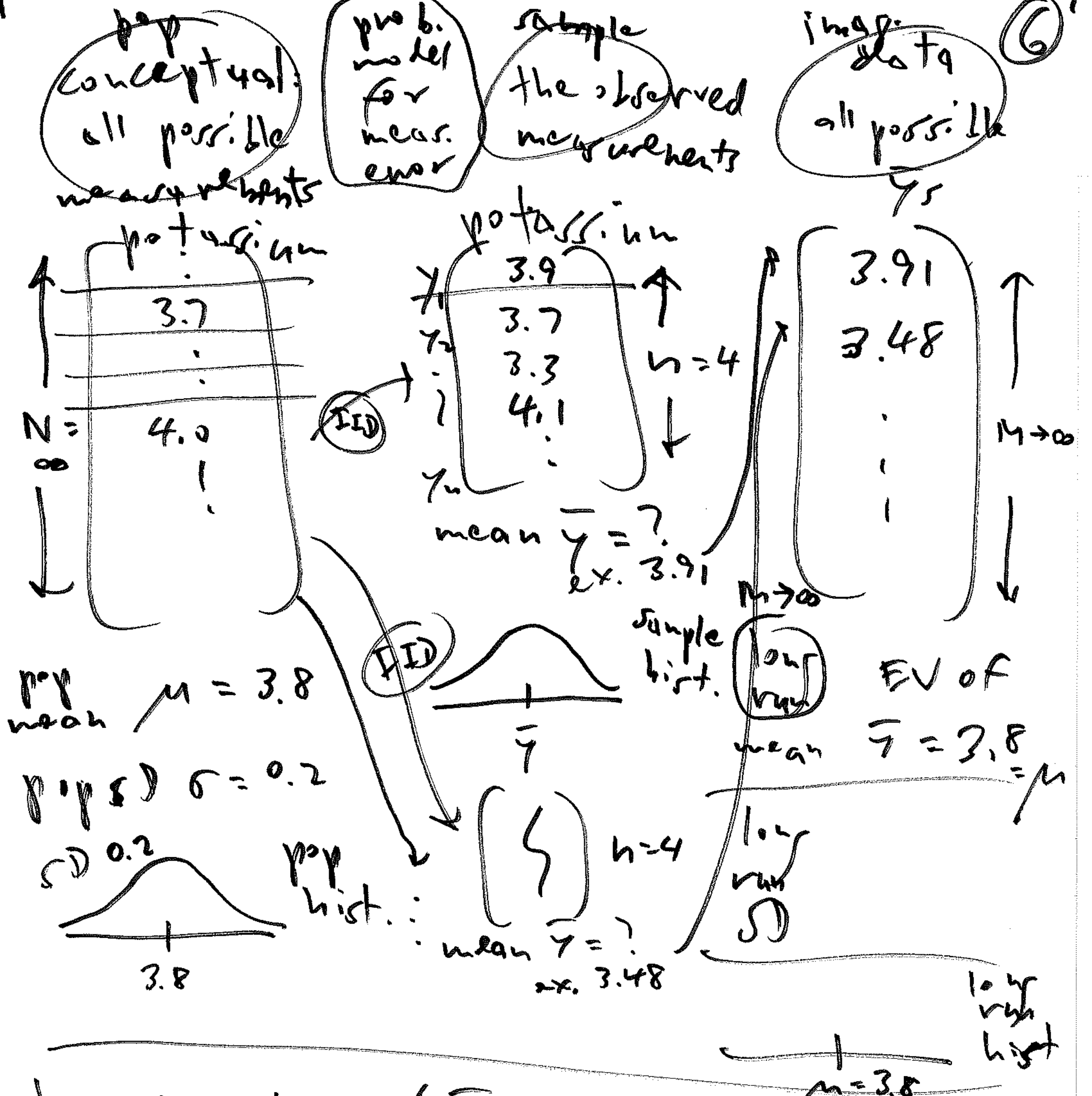
$$\bar{y}_n = \theta + b + e_n$$

$$40\% = 60\% + (-20\%)$$



phone books
club membership lists
↓ country clubs

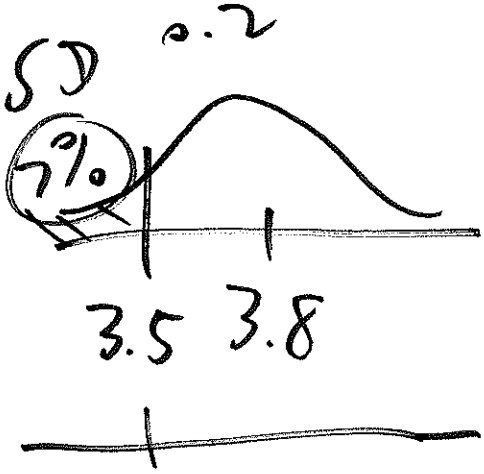
oversampled rich people + R



to estimate $P(\bar{y} < 3.5)$, we have to imagine getting lots of $\bar{y}$s & compute % of time $\bar{y} < 3.5$ in those repetitions

p(misclassification with $n=1$) ⑦

= 7%



hist of $\bar{Y}_1$

(12.46) too high

$\leftarrow \frac{3.5 - 3.8}{0.2} = -1.5$

p(misclassification with $n=4$)

= ?

$= P(\bar{Y} < 3.5)$

expected value of $\bar{Y}$ = EV of $\bar{Y}$

$= E_{IID}(\bar{Y}) = \mu = 3.8$