This time | intro; types of variables; samples & populations

AMS 7 29 Sep 15

(1)

next time: | numerical & graphical descriptive methods

dichotomous → binary

sample

disease?

0 ~~No~~
1 ~~Y~~
0 ~~N~~
0 ~~N~~
0 ~~N~~
⋮
1 ~~Y~~
⋮
0 ~~N~~

1 ← Y = yes

0 ← N = no

1 row for each deer

N

sample size n = 93

1 row for each subject

% of deer with disease

Y = 4

N = 89

n = 93

$\frac{4}{93} = 4.3\%$

average = $\frac{\text{sum}}{n}$ = mean

1 column for each variable

sum of 1s & 0s = # deer with disease

population ($1 = Y$, $0 = N$) sample

all UCSC deer as of 31 Dec 2006

the observed deer

disease?

population size

$N \doteq 800$

1s & 0s

like a 1 row for each deer

disease?

0
1
0
0
⋮
1
⋮
0

$n = 93$

"theta hat"

$\bar{y} = \hat{\theta}$ sample estimate of θ

$\bar{y}$ is a good estimate of θ

mean $= \boxed{\theta} = ?$

unknown

a pop. numerical summary

$\equiv$ a parameter

mean $\frac{4}{93} \doteq 4.3\% = \bar{y}$

sample mean

about equal to (approx)

---

y

$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{93} = y_n \end{bmatrix}$

"y sub 1" or "y 1"

$n = 93$

mean $=$

$$\frac{y_1 + y_2 + \dots + y_{93}}{93}$$

$$= \frac{y_1 + y_2 + \dots + y_n}{n}$$

mean $\bar{y}$ ← "y bar"

mean $= \bar{y} = \dfrac{y_1 + y_2 + \dots + y_n}{n}$

$$= \frac{1}{n}\Big(\underbrace{y_1 + y_2 + \dots + y_n}_{\sum_{i=1}^{n} y_i}\Big)$$

capital Sigma

sum

index of summation

$$\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \qquad (y_1 - \bar{y}) + (y_2 - \bar{y}) + \dots + (y_n - \bar{y})$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})$$

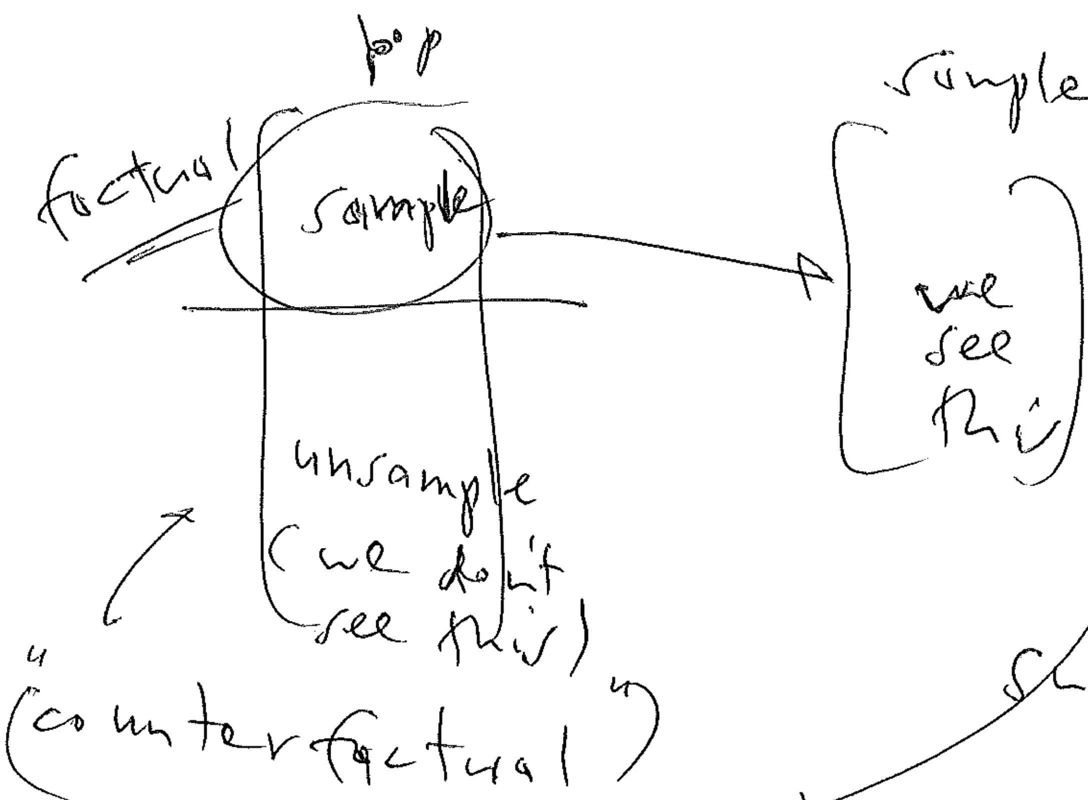
$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\frac{1}{n}(y_1 + y_2 + \dots + y_n)$$

mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

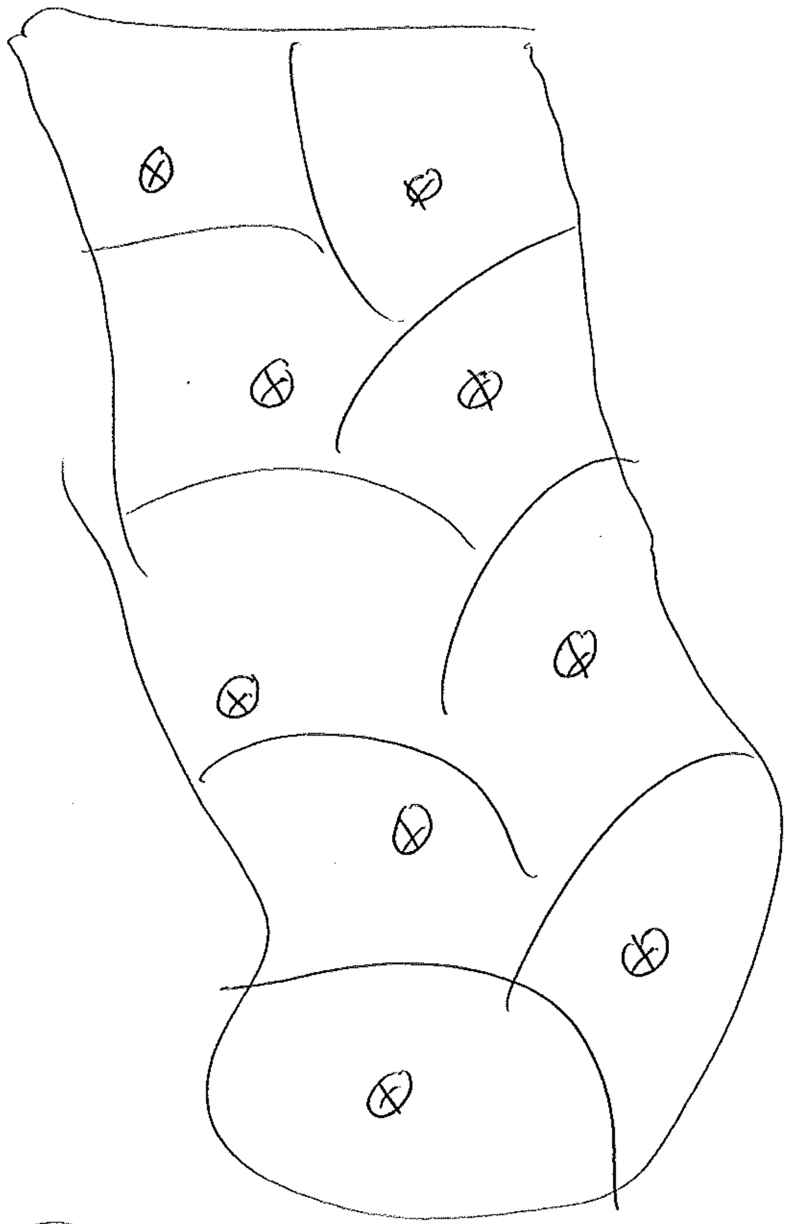
basic sampling principle:

try to make sampled & unsampled subjects in pop. as similar as possible in all relevant ways



how do this?

choose sample at random from pop.



pop

$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ ~~at~~ random

sample

$\begin{bmatrix} Y_1 = 9 \\ Y_2 \end{bmatrix}$ $n = 2$

math easier

at random with replacement =

independent identically distributed (IID) sampling at random without replacement = simple random sampling (SRS)

more informative than IID

variable		possible values
		qualitative binary dichotomous
eye color		blue (1), brown (0)
hair color		black, brown, red, white
plant size	height (cm)	7.42, 8.20, ...
	# of leaves	

quantitative