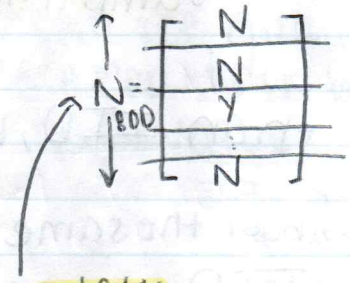
Wave Moretto 10/2/18

Stats Lecture #2: Variable Types + Histograms

P = population (all deer at UCSC as of 31 Aug 2018)

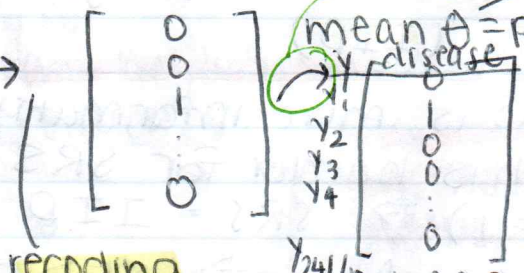
sample (8) (the observed deer)

disease? Y (yes) (1)
N (no) (0)



population size

recoding (numerical scale)



focus on this → parameter

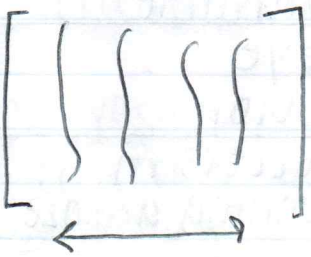
mean θ = p = ? (Y=1, N=0)

disease?

n = sample size = 241

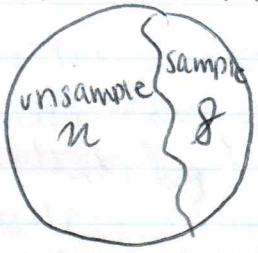
mean = $\bar{y}$ ("ybar") = $\frac{2}{241}$ = 0.8%

- General data set (matrix)
  - 1 row for each element/individual (subjects)
  - 1 column for each variable (things we measure on subjects)



- Any variable that takes on only 2 possible values is dichotomous (binary)
- Mean of a column of numbers = average of that column = $\frac{\text{sum of all \#s}}{\text{how many \#s there are}}$

Population



- Goal: try to sample so that sample and unsample are as similar as possible in all relevant ways
- 1650 - Pascal, Fermat began studying probability/statistics

- Bayes (1760) started looking at statistics
- Neyman (1925): proposed using true random sampling
- simplest method: choose sample at random
  - at random with replacement = independent identically distributed sampling (IID)
  - at random w/o replacement = simple random sampling (SRS)

- SRS is more informative than IID, but math is harder for SRS

$(n = 1) \Rightarrow$ SRS = IID

$n \ll N \Rightarrow$ SRS $\doteq$ IID (about the same)

($\ll$: is a lot smaller than; SRS: sample; IID: analysis)

variable	possible values for variable
eye color	brown, blue (dichotomous) — nominal — qualitative (qual) categorical
success in running a maze	very slow (1), slow (2), moderate (3), fast (4), very fast (5) — (no place on number line) — qual — ordinal
size of a plant	height (cm) (2.54 cm/in) ex: 4.0 cm — continuous — quantitative (quan) numerical
	# leaves ex: 12, 13, 14 — discrete
growing temp. (°C) (°F) (K) produces most bugs	78 °F