

Variables and Histograms

9/30/01

- Terms

- Statistics - study of uncertainty
 - how to measure uncertainty
 - how to make choices in the face of it
- uncertainty - state of incomplete or imperfect information

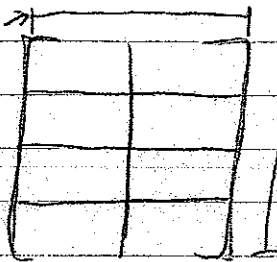
- ex. percent(%) of deer on UCSC who have chronic wasting disease on 7/24/08

$$P = \{ \text{deer who live on UCSC campus on 7/24/08} \} \leftarrow \text{the set}$$

where P is the population

- Population - collection of subjects or elements of interest
- Variable - things that can be measured on a population
 - dichotomous (binary) variable - one which only takes on two values (yes/no)

one column for each variable



one row for each subject

$$N = ? \quad (P \sim 1,000)$$

- sum of 0s and 1s equal to # of infected deer

Deer with Disease

N	N=0	0
N	Y=1	0
Y		1
⋮		⋮
N		0

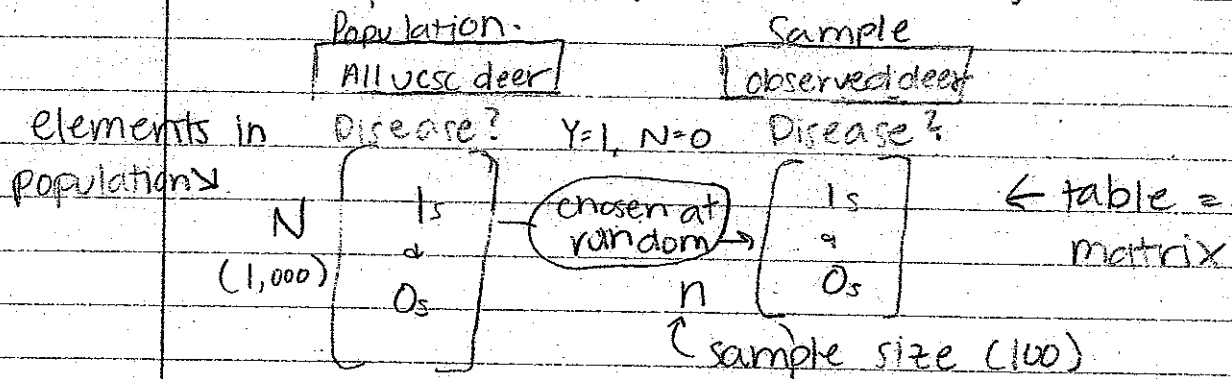
qualitative (categorical) variable

- mean of 0s and 1s equal to proportion of "yes" (1) values



- Populations and Samples

- Parameter - numerical summary of a population
 - ex. θ (percentage of deer w/ CWD)
- Subset (s) - sample of a population used to evaluate variables of that population
- Statistic - an estimate ($\hat{\theta}$) of a population parameter (θ)



mean $\theta = ?$

mean $\bar{y} = \hat{\theta} = 2/100 = .02 = 2\%$ $\hat{\theta}$ = "theta hat"

\bar{y} is good estimate of θ

N = # subjects in population P

n = # of subjects in sample s

- To fill out table:

- Title with what you observe
- Title each column w/ the variable it represents
- Identify N
- Repeat for sample column
- Sample must be as similar as possible to the population in all relevant ways
 - un-sample (all other population members) similar to sample as well

- Neyman, Fisher = Statisticians

- random sample achieves this because all subsets have an equal chance of being chosen

* write sampling method on diagram

- random with replacement = independent identically distributed (IID) sampling

- random without replacement = simple random sampling (SRS)

- more informative than IID

- more complicated mathematically

- SRS + IID function about the same when $N \gg n$ (SRS \approx IID)

- to choose random sample (n):

- make a list of all population subjects with tags

- choose n of these tags

- computers have a pseudo-random number generator

- measure the variable on these subjects with chosen tags

= Variables (Data Types)

Variable	Value	Notes
Nominal: Eye color in animal	brown, blue	- dichotomous - qualitative
Ordinal: Success in maze	1-5 (slow \rightarrow fast)	- qualitative
Ratio: Size of a plant (height cm)	4.71, 6.02, etc	- quantitative - continuous
Interval: temp. at which plants bud	73°, 22°C	- quantitative



look at flow chart for more info:

- Nominal - dichotomous variables with no distinct place on the number line
 - qualitative, categorical
- Ordinal - natural order to variable, no distinct place on number line
 - qualitative, categorical
 - ordered categorical variable
- Ratio - variables with a constant size interval and a true zero which allows us to make statements about ratios
 - quantitative, numerical
 - have place on number line
- Interval - has a constant size interval, but no true zero
 - cannot make ratio statements (ie 40°F is not twice as hot as 20°F because 0°F does not represent an absence of heat)
- Discrete - quantitative variable where gaps occur between possible values
 - ex. # of leaves: 1, 2, 3
- Continuous - quantitative variable with no gaps between values
 - ex. length of hair: 1 in, 1.19 in

color
brown=0
black=1
red=2
white=3

	Animal #	45	333	...	5011	Mean
	Age	2	0	...	3	1.7
	Color	0	2	...	1	1.2

- The means of "Animal #" and "color" are meaningless because they are qualitative (non-dichotomous) variables which use arbitrary numbers



- The mean of "age" is actually the average age of the animals

Graphical Descriptive Methods

- Ex. Monarch Butterfly wing length

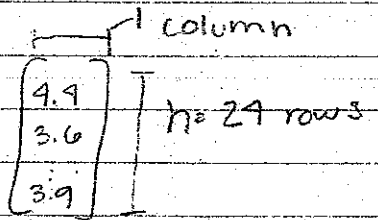
$n = 24$ immature butterflies

Variable = wing length (cm)

smallest = 3.3

largest = 4.5

arrange data from smallest \rightarrow largest



(as long as the order isn't relevant)

• Raw Frequency Distribution

Value	Frequency
3.3	1
3.4	0
3.5	1
4.5	1
Total	$n = 24$

describes how data is distributed on a number line (also "y distribution")

• Raw Frequency Histogram (graphical display)

