

PROBABILITY: quantification of uncertainty

- to decrease uncertainty, gather more data
- when you must make a decision w/ incomplete info, you make a prediction

DECISION MAKING: predicting the future under different sets of conditions & choosing the favorite output.

- to get more data, design experiments or sample surveys

It is possible to have too little data to make a decision, but also possible to have too much data.

- too much data = waste of time & money

Power calculations & sample size calculations tell us how much data we need to gather.

"STATISTICS IS COMMON SENSE REDUCED TO CALCULATIONS"

### EXAMPLE STATS PROBLEM

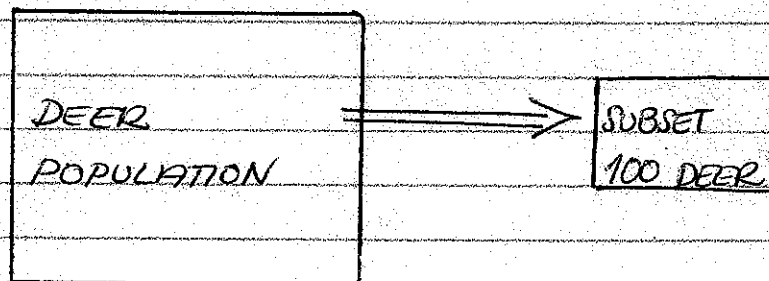
What percent of the deer on campus have Chronic Wasting Disease?

- Don't know exactly  $\rightarrow$  impression that deer on campus are pretty healthy. You can assume that the percentage is small.

Examining every deer on campus = too much information.

TAKE A SUBSET OF THE POPULATION

- as is done for voting polls  $\rightarrow$  about 1000 people polled to get an idea about entire population's preference



**DRAW A TABLE**

- If the variable of interest takes on only 2 values (ex: yes/no) = binary
- 1 row for each deer (100 rows, 1 column)

CHRONIC WASTING DISEASE?

no  
no  
no  
yes  
:  
no

RECODE TO BINARY  
CODING

CHRONIC WASTING DISEASE?

0  
0  
0  
1  
:  
0

yes = 1  
no = 0

(Computers store information by coding in #'s)

Binary is useful: automatically gives proportions

- From data set, add all numbers (1's & 0's) together & divide by population size in the subset.

1 deer w/ Chronic Wasting Disease = 1% of deer in subset have disease  
100 deer

Proportions in the subset not the same as % of total population w/ disease, but the # can be used as an estimate

Diseased % of deer population  $\theta$

$\neq$

Diseased % of subset  $\bar{y}$

But the process eliminates much uncertainty

- To eliminate all uncertainty, must find all the deer on campus and record disease status

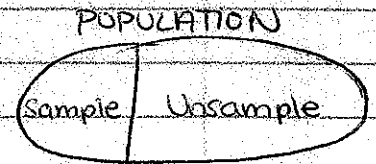
The subset is an estimate of the parameter (numerical summary of a population)  $\bar{y} = \hat{\theta}$

- Sample must be chosen well if it is to give a good estimate of the parameter

NOTE: population of deer w/ disease =  $\theta$   
% of deer w/ disease in sample =  $\bar{y}$   
 $\bar{y}$  = estimate for total population =  $\hat{\theta}$   
 $\hat{\theta}$  = estimate for  $\theta$

A good estimate can be made if THE SAMPLE POPULATION & THE UNSAMPLE POPULATION (TOTAL POPULATION) ARE AS SIMILAR AS POSSIBLE IN ALL RELEVANT WAYS.

- called a representative sample
- Good sample = Random Sample
- each deer has an equal chance of being chosen



RANDOM SAMPLING DOES NOT MEAN EACH SAMPLE WILL GIVE A GOOD ESTIMATE. IT ONLY PROMOTES A GOOD ESTIMATE

## 2 TYPES OF SAMPLING

- IID (Independently Identically Distributed) Sampling
  - at random w/ replacement: when using multiple samples, you replace the individuals from the previous sample back into the population before choosing another sample at random
  - easier math but less informative
- SRS (Simple Random Sampling)
  - when taking multiple samples, individuals from the previous sample are NOT replaced into the population. They cannot show up again in subsequent samples.
  - harder math but more informative

\* Although SRS is more informative & tends to produce better estimates, when the sample population is a lot smaller than the unsample, SRS & IID produce similar results  
if  $n \ll N$ ,  $SRS \approx IID$

Random Sampling can be achieved in several ways

① tag all deer & enter their ID's into computer; computer chooses a random sample using SRS method

- although this technique gives a true random sample, it is time-consuming & difficult

• If you plan to find all deer on campus to ID them, why not examine each deer when you find it & get exact % of population w/ Chronic Wasting Disease?

② Pseudo-Random Sampling - works well if individuals are evenly distributed throughout an area

- Partition campus into 100 areas of even size & send out 100 volunteers on the same day to examine the first deer they encounter.

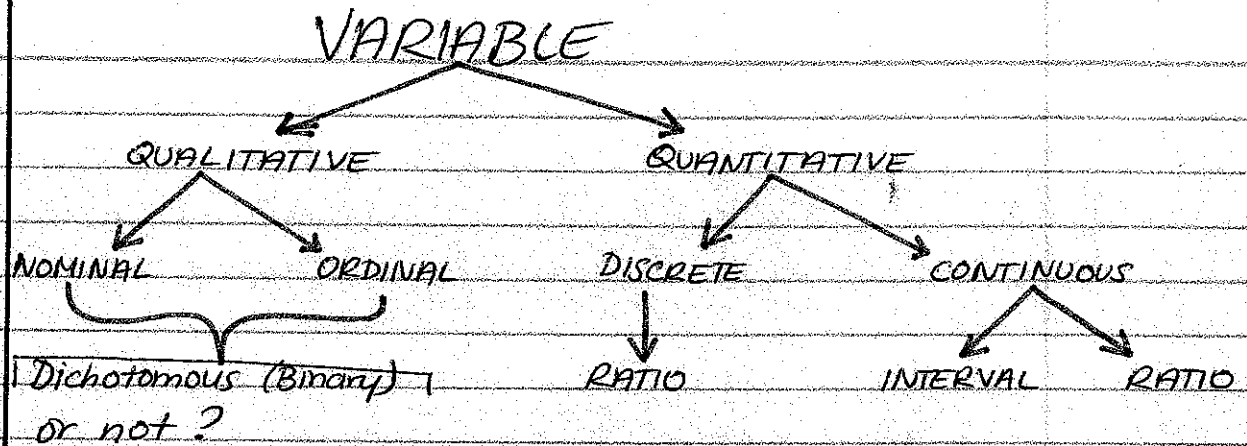
### SAMPLE VARIABLES

VARIABLE	VALUES & TYPES
eye color	blue, brown - qualitative - nominal - dichotomous (binary)
mouse's success in running a maze	very slow, slow, moderate, fast, etc. - qualitative, ordinal
height of plant	100.39 cm - quantitative, continuous, ratio
# of leaves on a plant	43 - quantitative, discrete, ratio
growing $T^{\circ}$ @ which most buds on a plant are produced	25.24 $^{\circ}\text{C}$ - quantitative, continuous, interval

# MAYA SERTIC

## TYPES OF VARIABLES

P. 5  
01/08/09



**QUALITATIVE (CATEGORICAL)**: variables that do NOT have a unique place on the number line. If data not numerical, then it is qualitative.

**NOMINAL**: non-numerical variables  $\&$  do not have a natural ordering to them (ex: hair color)

**ORDINAL**: variables which, although they do not have unique places on the number-line, they do have a natural ordering (ex: small, medium, lg)

**QUANTITATIVE** - variables that have a definite place on the number-line

**DISCRETE**: quantitative variables w/ gaps of the same size between all of the values. These values are usually whole numbers. The value stays the same, no matter how finely you measure it. (ex. number of leaves on a plant = 34; cannot = 34.3)

**CONTINUOUS**: quantitative variables w/ no conceptual gaps between them. Your measurements can always be finer  $\&$  finer, utilizing all the numbers on the number-line.

**RATIO**: there is a true zero on the measurement scale  $\Rightarrow$  zero means the individual is non-existent. This allows us to make meaningful statements about ratios.

ex: plant = 100 cm tall, twice as tall as 50 cm plant

**INTERVAL**: no true zero  $\Rightarrow$  zero does NOT mean individual is non-existent. Cannot make statements that are valid about ratios (ex:  $T^{\circ} = 0^{\circ}C$  does not mean temperature doesn't exist;  $80^{\circ}F$  is NOT twice as hot as  $40^{\circ}F$ .)