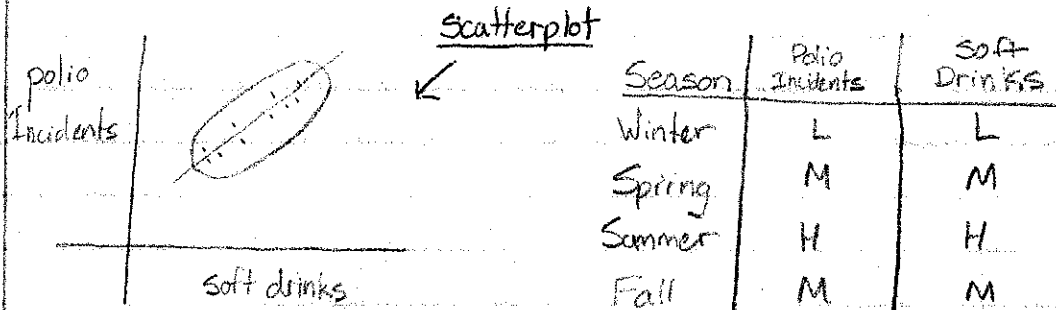


9/24 Intro, populations, + samples



L = Low
M = Medium
H = High

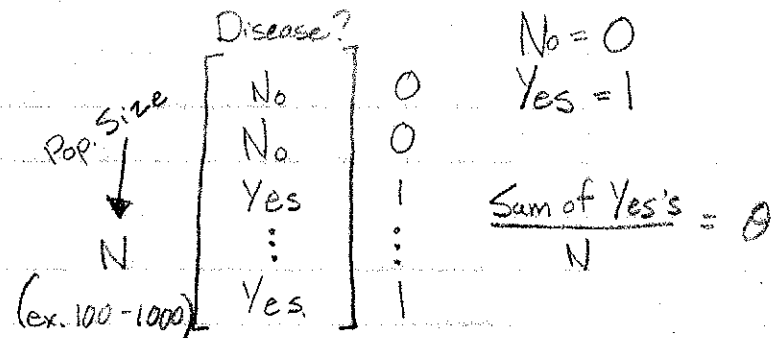
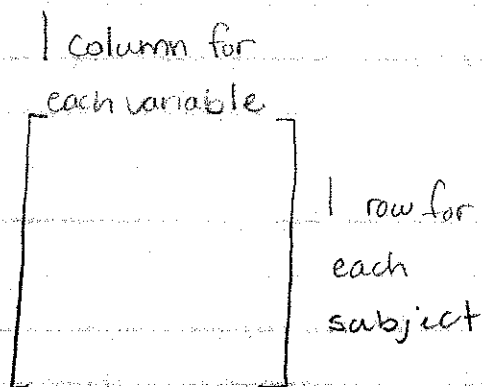
• There was a strong positive association between soft drink consumption and polio incidence. Not the real cause. Real cause = Non-Chlorinated swimming pools. More swimming in summer = more polio incidences.

9/29 Introduction + Descriptive methods

Uncertainty: incomplete information, θ is the uncertain fact.
Gather Info

Pop.
all wisc deer
on 31 Dec. 2006

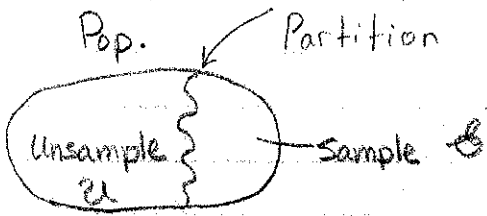
Deer Pop.



Average = Mean = Arithmetic mean

Parameter, is a numerical summary of a population, such as θ in the example above.

People don't do complete census. rather they use a subset or sample.



- use the sample to estimate the total population percentage of diseased deer

Pop.
all UCSC Deer
on Dec. 31 2006

Sample
The observed
deer

disease?
N = 800
1's
+
0's

Chosen
at
Random

disease?
sample size
n = 150
1's
+
0's

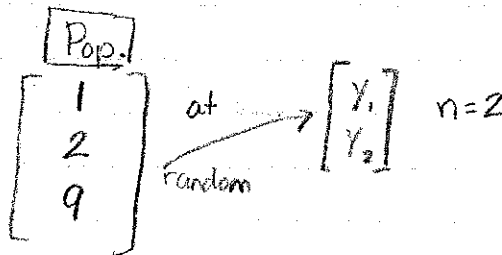
mean θ
unknown parameter

mean $\hat{\theta}$ = sample estimate of θ

If samples good, $\hat{\theta}$ is a good estimate of θ ; representative sample

The goal in sampling is to make the sample and the unsample similar in all relevant ways

Random sampling is one way to do this, is relatively simple exp.



① At random with replacement = independent identically distributed sampling (IID)

② At random without replacement = simple random sampling (SRS)

③ In practice people typically do SRS (or equivalent) but the math is easier for IID ① if n is a lot smaller than N ($n \ll N$), SRS \approx IID

SRS: Neyman (1920's) figured out random sampling usefulness

One way to sample randomly the deer, is to partition the campus into areas and test the 1st animal you discover in each area
 - Not useful if the disease is in pockets, not a real random sample

but close enough.

Variable	Values
Eye Color	brown, blue
Success in maze running	1 (very slow) 2 (slow) 3 (medium) 4 (fast) 5 (very fast)

Qualitative Nominal
 ← two values: **Dichotomous** binary
Qualitative not numerical
 This version is qualitative
 #'s stand in for qualitative or categorical variable

Variable	Values
Size of plant = Height	15.2 cm
# of leaves	64
Growing temp. w/ most buds	19°C

(Continuous) = Hypothetically or conceptually on # line
 (Quantitative) values on number line
 Quantitative (discrete) meas.
 Quantitative (quant.) Continuous (Cont.)

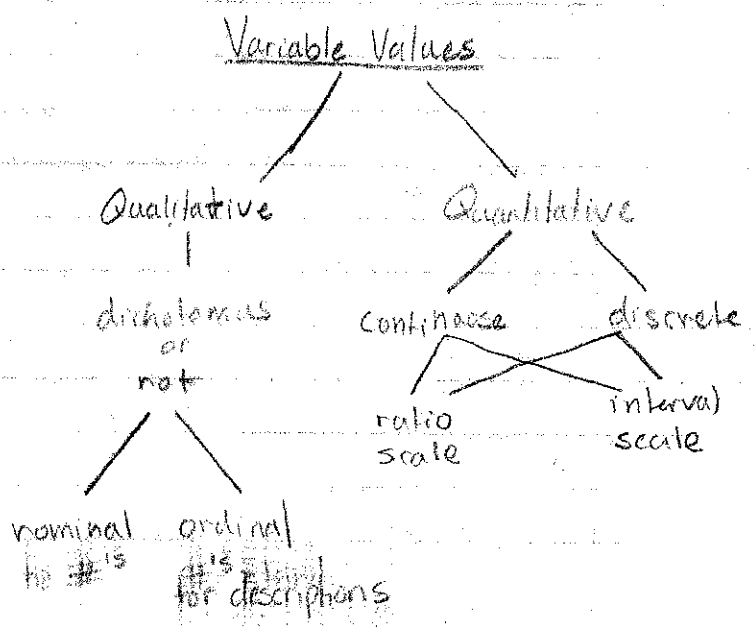
Structural Gaps between possible values

cm = has proper 0 no height exp.
 can use to make ratios,
ratio scale

temp. = 0 is arbitrary,
 can't use ratios, interval

Extra Vocab

ordinal = categorical



Histograms have no gaps in bars
bar graphs have gaps.

Wing length
cm

4.4
3.6
⋮
3.9

n = 24 butterfly

SORT THEM

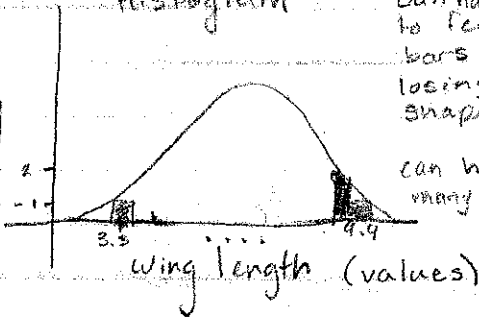
3.3
3.5
3.6
3.6
⋮
4.5

Raw Frequency

Values	Count	Relative Frequency %
3.3	1	4% = $(1/24) \cdot 100\%$
3.4	0	0%
3.5	1	4%
⋮		
4.4	1	
4.5	1	
+ 1		
24		

Raw Frequency Histogram

Relative Raw Frequency
Frequency
8%
4%



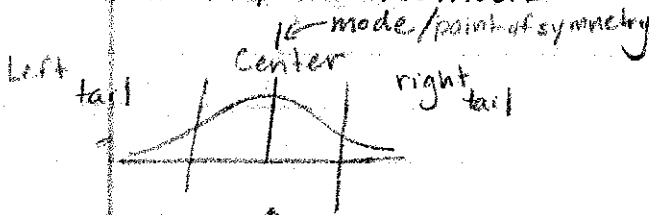
can have to few bars losing shape or can have to many = too noisy

Raw Frequency distribution for wing length

Histograms = a special bar graph for a

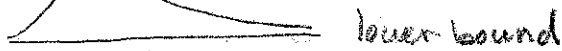
9/31

Descriptive Methods



Symmetric distribution unimodal + (0 skew)

Income

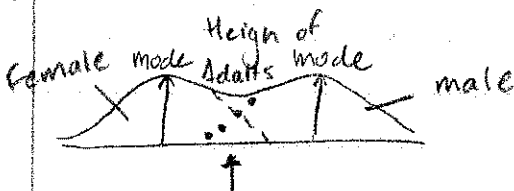


skewed (long right hand tail or positive skewed)

mid term grades

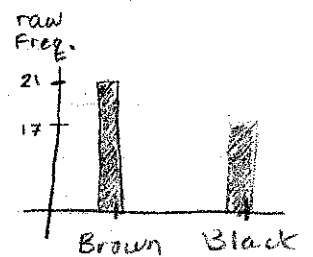


upper bound (long left hand tail negative skewed)



Quantitative var.

Hair Color	Freq
brown	21
black	17



Bar graph = a graph of qualitative var.