Brittany Guest
(lec. 2)

Variable Types + Histograms

AMS 7 9/30

*read cha. 1 + 2
*reader available from Juan Carlos at back of room → $20 cash; no more in class handouts
1 week from today

1.1 Introduction

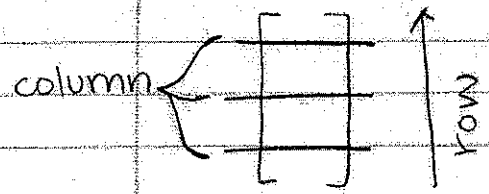
* the percentage θ (theta) of the deer who live on UCSC as of 27 July 2008 who have chronic wasting disease

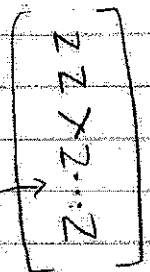
→ Don't know the value of θ, we think it is small

The Set:

P = the deer who lived on UCSC 27 July 2008 is an example of a population: a collection of subjects or elements of interest.

1 row for each subject, + 1 column for each variable.





N = NO Y = Yes

n = ? (around 1,000)

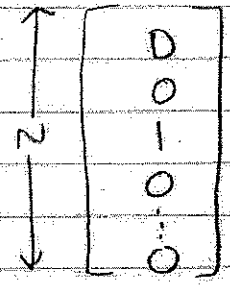
qualitative (categorical)

population (disease?)

Things that can be measured on a population subjects are variables

* recoding from qualitative to quantitative:

N = 0
Y = 1



← quantitative

mean (θ) = proportion of yes values.

θ = example of one possible parameter of interest

* In practice we are not able to perform a complete census of population P

* Instead it is natural to choose a subset S of P and evaluate the variable(s).

* Such a subset is called a sample from population P. If sample chosen well use data to make a good estimate

Population (P) — all deer on UCSC July 2009 — 1 = Yes, 0 = No — Sample (S) — the observed deer

N = 1,000

disease?

[1s, ⋮, 0s]

How chosen? → AT RANDOM

disease?

[1s, ⋮, 0s]

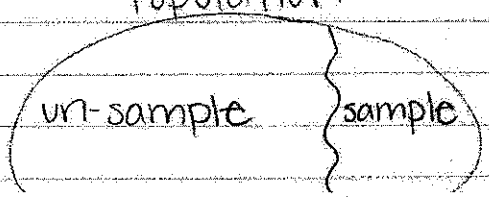
n = ? (ex. 100) ↕ sample size

mean θ

mean $\bar{y} = \frac{2}{100} = 0.02 = 2\%$

$\bar{y}$, or $\hat{\theta}$ (theta hat), is a good estimate of θ

Want a representative sampling method:
Sample should be similar to unsample in All relevant ways



How do we achieve this goal?

* simplest way to achieve this goal
Neyman and Fisher: choose the sample at random

Population → Sample

$N = 3$ $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ → $\begin{bmatrix} \quad \end{bmatrix}$ $n = 2$

At random with replacement: independent identically distributed (IID) sampling
At random without replacement: simple random sampling (SRS)

☆ SRS is more informative than IID, so SRS is what is actually used often, but the math is easier for IID. If N is alot bigger than n ($N >> n$) SRS & IID are about the same (SRS ≈ IID)

In practice, divide campus into subsets & take data from one deer in each subset. (NOTE: deers must be evenly distributed throughout campus).

	variable	values	
nominal = no order	eye color	brown, blue	qualitative dichotomous (two possible values)
ordinal = ⊛ have order	success in running in maze	1. very slow 2. slow 3. moderate 4. fast 5. very fast	} qualitative
	size of plant (ht. in cm)	4.77 cm, 6.024 cm; 0cm (true zero) → ratio	} quantitative
	growing temp@ which most productive	73°F 23°C; 0° (no absolute value) → interval	quantitative
	# of leaves	0, 1, 2	quantitative

quantitative: numerical + unique place on number line

⊛ maze success has order to it: ordinal (ordered categorical)

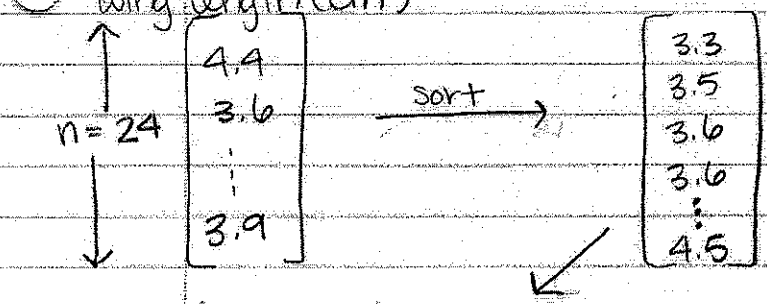
grow temp + # of leaves: discrete; only whole #'s ∴ gaps

size of plant: continuous; no gaps
↳ true zero: ratio (absence of leaves or height)
growing temp: no absolute zero: interval

*never meaningful to take the mean of a qualitative variable

Graphical Descriptive Methods

Ex. butterfly wing length



value	count (raw frequency)
3.3	1
3.4	0
3.5	1
3.6	2
⋮	⋮
4.5	1
	24 = n

raw frequency distribution of wing length variable