Negin Navaei

AMS 7 - Lecture 4·5·18

THIS TIME: SAMPLES OF POPULATIONS; HISTOGRAMS
NEXT TIME: MEASURES OF CENTER & SPREAD

Announcements

* HW Due Tue 17 April 2018 by 11:59 pm (100 points)
  ↳ @canvas.ucsc.edu
  Note: always include a brief description to support your answers for full credit! AND never leave a problem blank
* As of next monday, go to your discussion section
  ↳ preferably assigned section
* Waitlisted students will get permission codes via email if you don't receive one, the other class has space.
* Webcast: webcast.ucsc.edu
  ↳ ID: ams-7-1 PSWD: Uncertainty-Quantification

Chapter 1 ▷ INTRODUCTION & DESCRIPTIVE METHODS

1.1 • Introduction

→ Statistics
the study of uncertainty
How to measure it & What to do about it

→ Uncertainty
a state of incomplete or imperfect information about something of interest

* example
the percentage θ (theta=P) of the deer who live on the UCSC campus as of 31 Mar 2018 who have Chronic Wasting disease (CWD)
↳ θ is small Since most deer seem healthy, but there is substantial uncertainty about its Precise value
* uncertainty can be reduced by gathering data (data set)

4.5.18

PRINCIPLE: To decrease uncertainty about something unknown (or not completely known), gather new GOOD data

→ The set / collection

* P = {the deer who live on the UCSC campus as of 31 Mar 2018}
  ↳ finite number of deer ≈ 800
  ↳ an example of a population

→ Population (P)

a collection of subjects or elements of interest
*CWD or healthy?

Things that can be measured on population subjects of interest

← 1 column for each variable →

Pop Size = N
aka how many rows we have
800 deer

↑ 1 row for each element ↓ (N)

	DICHOTOMOUS	BINARY
	NO	0
	NO	0
⇒	Yes	1
	NO	0
	⋮	⋮
	NO	0

BINARY CODING
yes = 1
NO = 0
*more useful to assign binary values

Can also be dichotomous (yes, NO)

Sum Σ = # deer with CWD mean = $\frac{\Sigma}{N}$ = ? = θ = P

* Graphical vs. Numerical Characters
• graphical & numerical summaries of data sets:
  ↳ Descriptive statistics

→ Parameter

a numerical summary of a population
*could be θ, P, N, Σ, etc. or any numerical summary

▷ Choose a subset ($\mathcal{S}$) of P & evaluate the variable(s) of interest only on the population subjects in the subset
  ↳ such a $\mathcal{S}$ is called a sample from the population P
    *sample $\mathcal{S}$ ⇒ the observed deer

→ sample

a subset of a pop. that you use to make intelligent guesses about the whole population

4.5.18

Sample $\mathcal{S}$

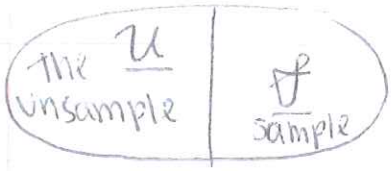
The observed deer

measured variable → CWD?

$$\begin{matrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{matrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} \quad n = \text{Sample size}$$

n = the # of deer in sample

Population $\mathcal{P}$



* Partition of $\mathcal{P}$

* You can have good & bad samples
* Goal of Sampling:
  Try to make $\mathcal{S}$ & $\mathcal{U}$ as similar as possible in all relevant ways
  ↳ similarity vs. relevance
  ↳ Choose the sample at random so that all deer have the same chance of being sampled

To achieve this goal, choose $\mathcal{S}$ at random

▷ Should sampling be w/ replacement or without?!?

* TWO SIMPLE RANDOM SAMPLING METHODS
  1) At random with replacement
     ↳ Independent identically distributed (IID) → doesn't require tagging deer
  2) At random without replacement
     ↳ simple random sample (SRS)
     * Can be more informative than IID but IID has easier math

* when $n \ll N$ (Sample size n a lot smaller than N ↳ pop. size) SRS is approximately same as IID

▷ randomization can't guarantee perfect similarity (in all relevant ways) between $\mathcal{S}$ & $\mathcal{U}$ everytime
↳ The bigger "n" gets, the more likely that $\mathcal{S}$ & $\mathcal{U}$ are relatively similar
↳ We will learn methods to estimate how often randomization yields BAD sample (unrepresentative of the population $\mathcal{P}$)

4.5.18

$n = 91$ CWD?

Sum $= 4$

mean $\frac{s}{n} = \frac{4}{91} \doteq 4.4\%$ (estimate of θ) $= \bar{Y} = \hat{\theta} = \hat{p}$

* Since sampling was at random, this is a good estimate of $\theta = p = \frac{s}{N}$

↳ How Good?

We think θ is around 4.4% give or take ? %

1.2 • Data TYPes

Variables → the values they take on

* Genetics (Phenotypes)

variable	Possible values
eye color	brown, blue ② ① Dichotomous if only possible eyecolors
Hair color	brown, black, red, white
Success in maze running	very slow, slow, moderate, fast

Qualitative
Different scale used
Catagorical / nonnumerical

↳ Ordered Catagorical

→ Qualitative

Its values do not have unique places on the # line