

Lecture 1

Introduction to

Study design & R

Dr. Wendy R

Today

Here's the plan

Finish syllabus stuff + Slido poll	Create community guidelines and grading guidelines
Lecture	Data terms, types of variables, parameters vs. statistics, types of studies
R Markdown	Introduction to R Markdown documents



Slido poll

Join at
slido.com
#2383 391

Why statistics?

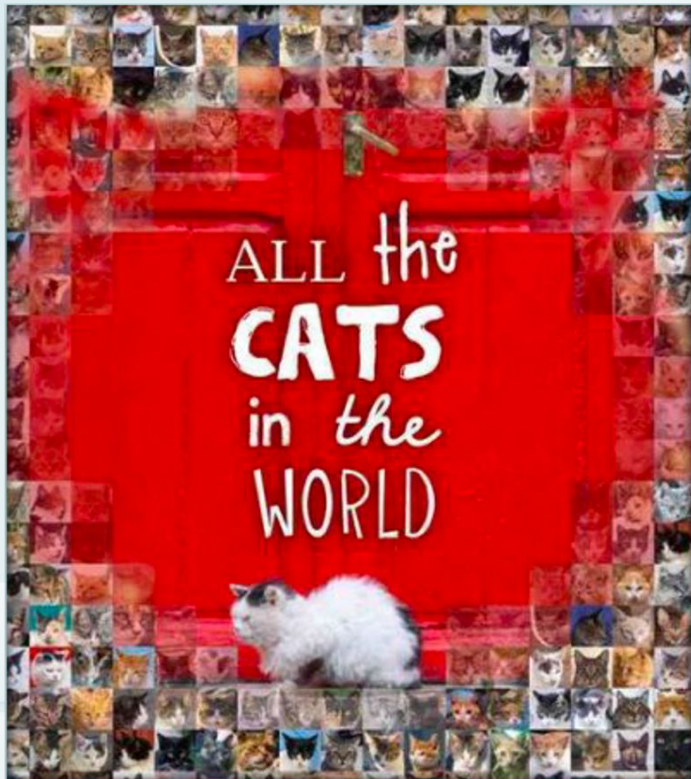
Statistics help us understand the world around us.

E.g., [Covid tracking](#), misinformation, social justice movements, social media influence, and so much more!

What does a statistician do?

“Quantify uncertainty”

1. define the problem, and formulate research questions
2. design the sampling procedure or experiment for collecting the data
3. explore and analyze the data
4. formulate conclusions and communicate the results



Population

a collection of people,
items, or events about
which you want to
make inferences

kitten I fostered last year

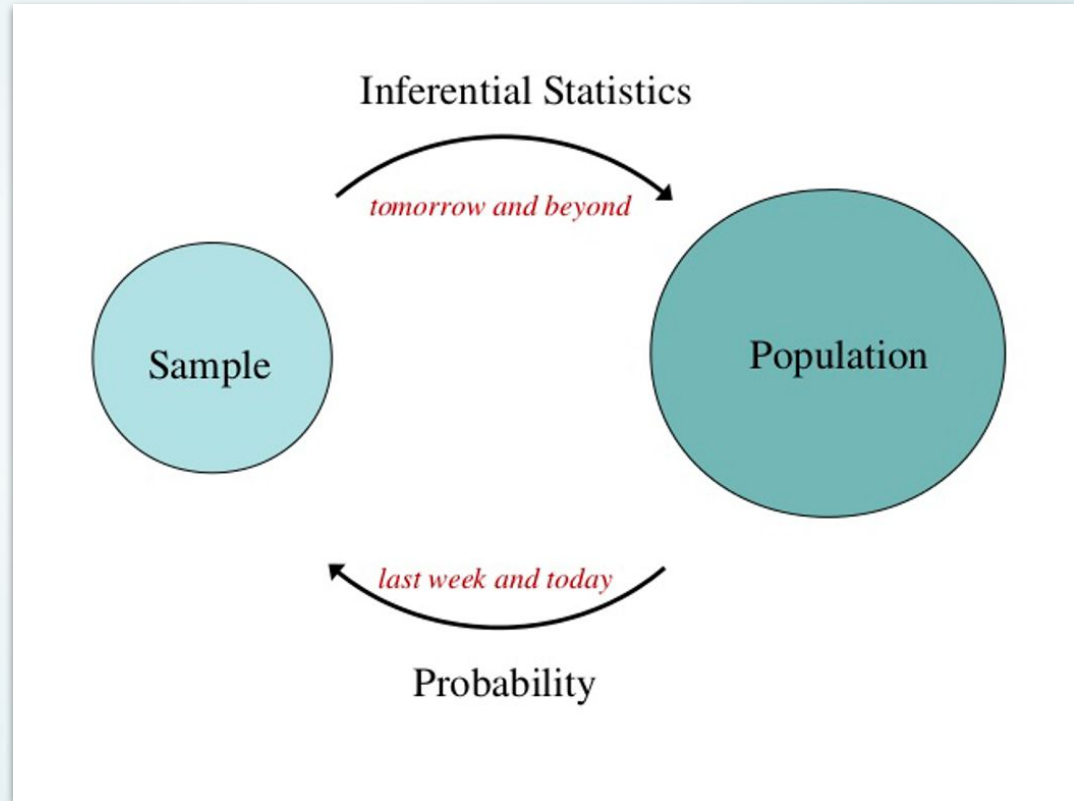


Sample

a subset of the population

e.g., 50 randomly sampled foster kittens

Probability vs. Statistics



statistics vs. parameters

Parameter: value that describes an entire population

typically notated with a Greek letter

- μ
- σ
- β_0
- β_1
- ρ

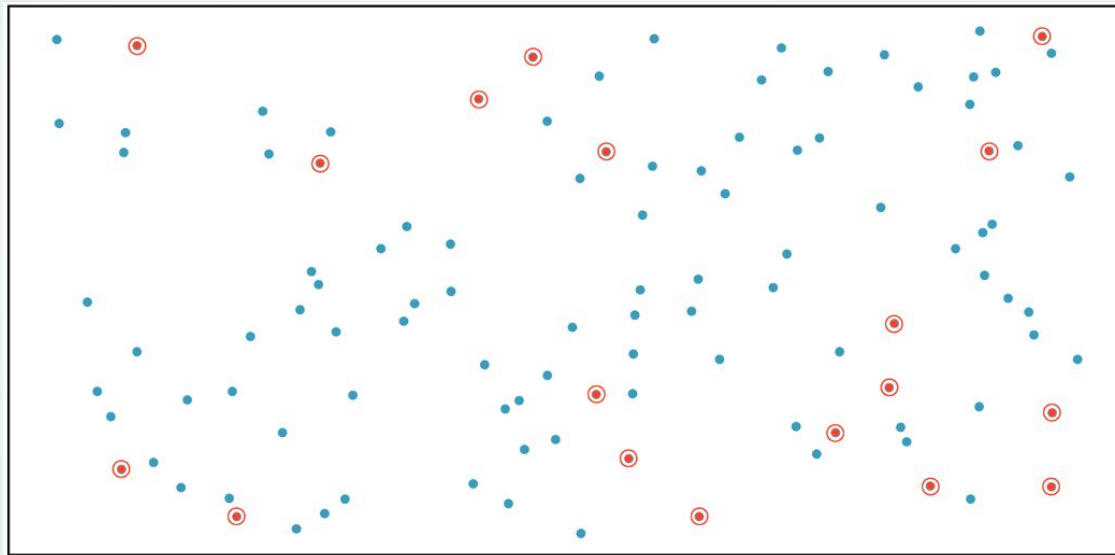
Statistic: value that describes a sample

typically notated with a lowercase letter or hats

- \bar{x}
- s
- $\hat{\beta}_0$
- $\hat{\beta}_1$
- $\hat{\rho}$

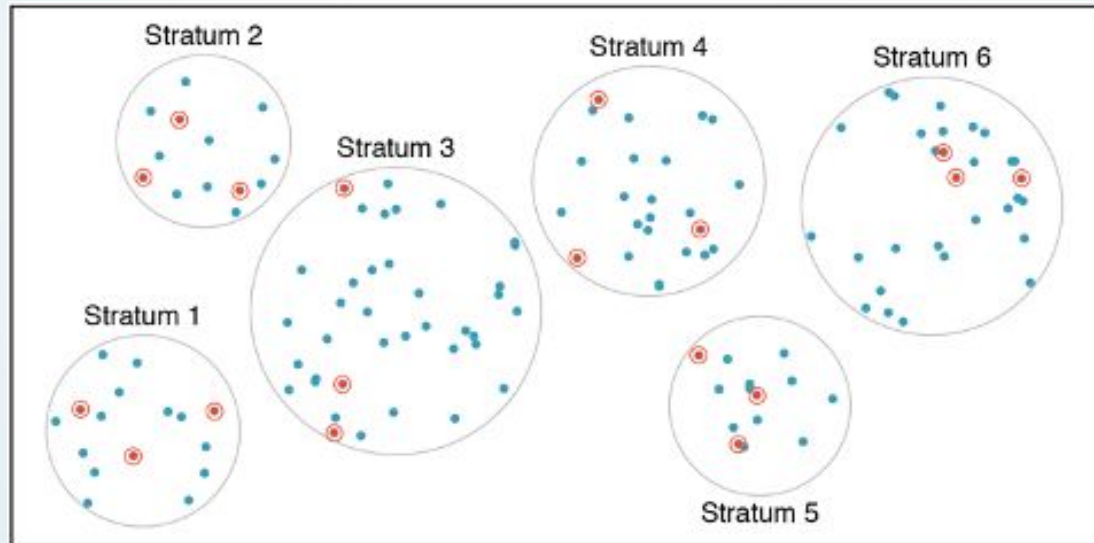
SRS

simple random sample (srs): each case (observational unit) in the population has an equal chance of being included in the final sample



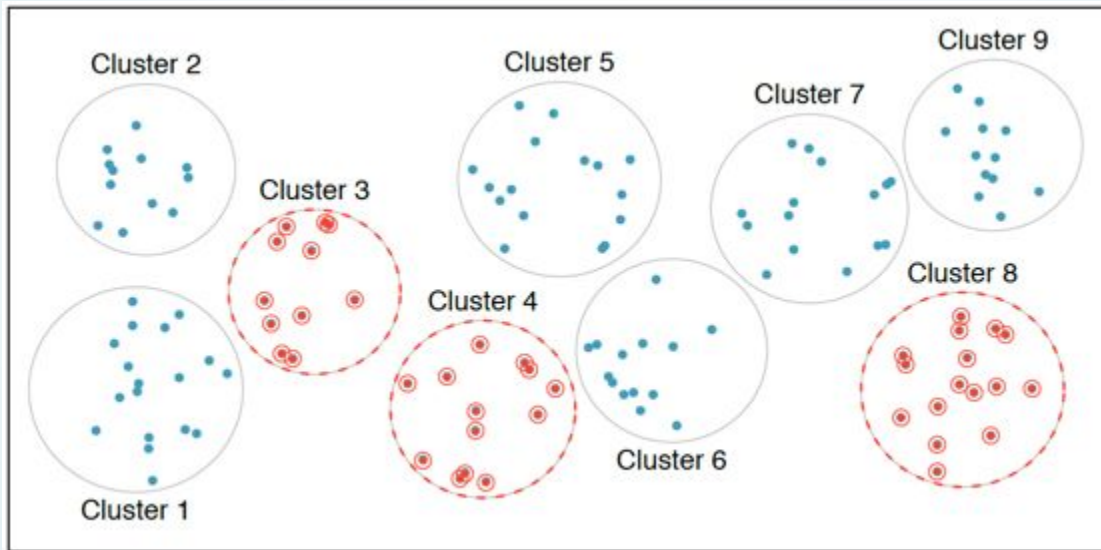
stratified sampling

stratified sampling: the population is divided into strata according to some variables that are thought to be related to the variable of interest; take an SRS from each group



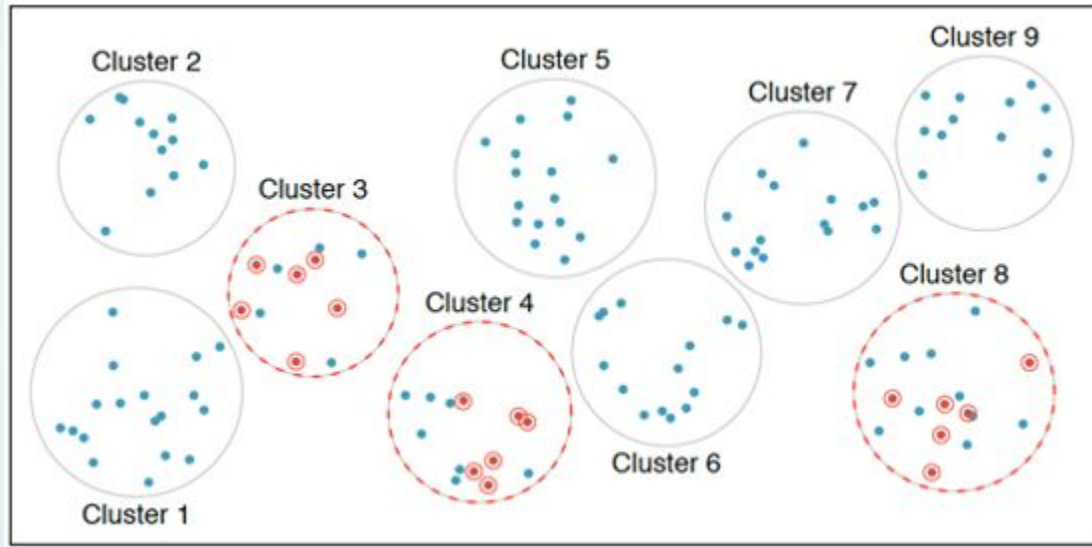
cluster sampling

cluster sampling: the population is divided into clusters and a sample of clusters is taken



multistage sampling

multistage sampling: break up population into clusters, select a sample of clusters, and apply an srs within each chosen cluster



convenience sampling

convenience sampling: individuals who are easily accessible are more likely to be included in the sample

example

- sitting outside of the student union and asking students to answer your survey questions

voluntary response sample

voluntary response sample: non-probability sample made up of individuals who *volunteer* to be included in the sample

examples


- Mail-in survey
- Call survey
- Social media poll

mishaps in sampling

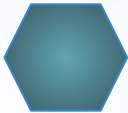
bias: occurs when a sample statistic under/overestimates a population parameter



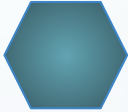
sampling bias



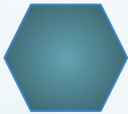
the sampling frame was non-random and does not reflect the characteristics of the population being tested (i.e., they are not **representative** of the population)



non-response: individuals chosen are unwilling/unable to participate



voluntary response: observations are self-selected volunteers



convenience: individuals who are easily accessible are sampled

types of studies



observational

observational units
are not manipulated
in any way that will
affect the outcome of
the study



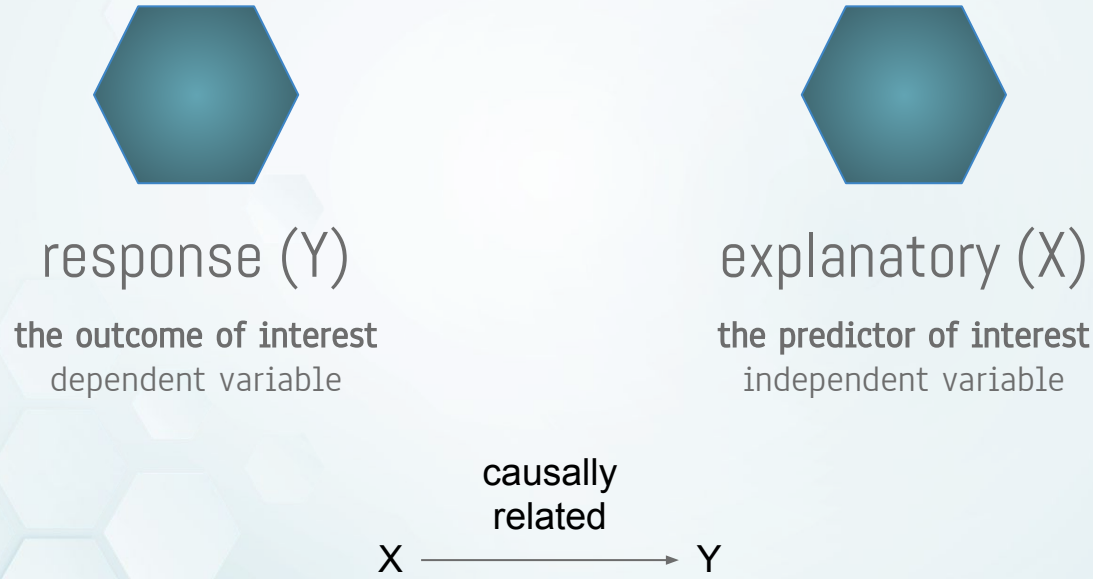
experiment

where researchers
manipulate something
and measure the
effect of the
manipulation on some
outcome of interest

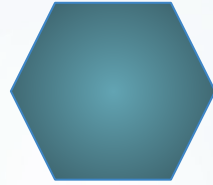
generalizability



relationships between variables

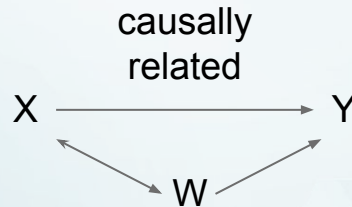


relationships between variables



confounding variable (W)

a variable that is associated with the predictor of interest (X)
and causally related to the outcome of interest (Y)



RStudio terms to know

01

R script

where you will type code and code only

02

console

where all the output goes when you run code

03

packages

“applications” containing special functions, datasets, and maybe more

04

R Markdown

where you can type code, text, and more!

05

environment

where all your variables and data frames live

06

plots/Viewer

where you can view plots and RMarkdown docs

R terms to know

01

variable

(in R) a named object that
represents some sort of data

02

data frame

a rectangular dataset

04

assignment operator

`<-`
how to save data as a variable

05

vector

a collection of similar types
of data

types of variables

variable: any characteristic, number, or quantity that can be measured or counted

quantitative

consist of meaningful numeric values
taken on each observational unit

discrete

counted (only integer values)

continuous

measured (any real number)

categorical

consist of categories or group names
measured on an observational unit

ordinal

categories have a logical ordering

nominal

do not necessarily have a logical ordering

types of variables in R

quantitative

num (numeric)

discrete

int: integer

continuous

dbl: double

categorical

fct: factor

ordinal

fct: factor

nominal

fct: factor

text

chr: character

(a single letter/special character)

string

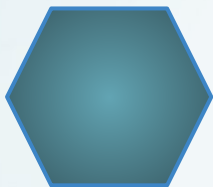
string

(made up of characters)

logical

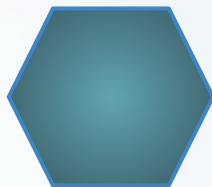
TRUE/FALSE

types of objects in R



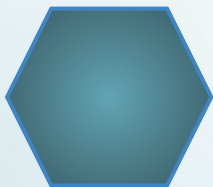
data frame

rectangular
dataset



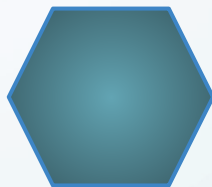
matrix

2-dimensional
array



vector

sequence of a similar
type of data



list

can contain
heterogeneous data types