This samples & populations; time: histograms
next: measures of center & spread
as of 1 next Mon.
go to discussion section

principle: to decrease your uncertainty about something unknown to you, gather new good data

Sample of the observed deer

N = 80

For each element
no
no = 0
yes = 1

Chronic wasting disease: CWD

Population size

Sample size

\[ n = 91 \]

Mean \[ \bar{y} = 4 \]

Estimated \[ \theta = \frac{4}{91} = 4.49\%

Sample variance

Binary

Logistic regression

Parameter \( \theta \)
Graphical & numerical summaries of data sets: descriptive statistics

Goal of sampling:
- Try to make \( \mathcal{S} \) and \( \mathcal{U} \) as similar as possible in all relevant ways.

This goal: choose \( \mathcal{S} \) at random

2 simple random sampling methods:
- At random with replacement (independent identically distributed (IID))
2) at random without replacement
(simple random sample (SRS))

SRS is more informative than II), but II) has easier math

when \( n \ll N \) smaller

\[
\frac{\text{sample size}}{\text{pop. size}}
\]

\[\text{SRS} = \text{IID}\]

randomization can't guarantee perfect similarity between S & II every time (in all relevant ways)
1. The bigger a set, the more likely that if we are relatively similar

2. We will learn methods to estimate how often randomization yields

bad samples → unrepresentative

since sampling is at random,

\[ \hat{\theta} = \hat{p} = \frac{\gamma}{n} = 4.4\% \]

Is a good estimate of \( \theta = p = \frac{3}{N} \),

How good? we ask \( \theta \) is around 4.4%, give or take ?%
variables

- categories
- nominal
- qualitative
- Squalaluria
- A

- variables
- hair color
- eye color
- middle

- unique
de numbers
- not
- yes
- no

- test
- very
- moderate
- very

- brown
- black
- white
- blue

- short