AMS 7
5 Apr 18

this time: samples & populations; histograms

next time: measures of center & spread

ar of ① next Mon, go to discussion section

principle: to decrease your uncertainty about something unknown (or not completely known) to you, gather new good data

population P

all deer living in UCSC campus on 31 Mar 18

sample of the observed deer

≐ 800

1 column for each variable

chronic wasting disease? (CWD)

no
no = 0
yes = 1
no
⋮
no

1 row for each element

N

population size

(0 0 1 0 ⋮ 0) ← binary, N

(1 = yes, 0 = no)

sum S = # deer with CWD

mean $\frac{S}{N} = ? = \theta = p = ?$

parameter → $\frac{S}{N}$

parameters

CWD?

$y_1, y_2, \dots, y_n$

(1s & 0)

at random

all deer have same chance of being sampled

sample size $n = 91$

sum $s = 4$

mean $\frac{s}{n} = \frac{4}{91} \doteq 4.4\%$

$= \bar{y} = \hat{\theta} = \hat{p}$

our estimate of θ = $\hat{\theta}$

theta hat

Graphical & numerical summaries of data sets: descriptive statistics

goal of sampling: try to make $\mathcal{S}$ and $\mathcal{U}$ as similar possible in all relevant ways

pop. $\mathcal{P}$

$\mathcal{U}$ unsample | $\mathcal{S}$ sample

(partition of $\mathcal{P}$)

this goal: how to achieve: choose $\mathcal{S}$ at random

2 simple random sampling methods:

① at random with replacement (independent identically distributed (IID))

② at random without replacement

(simple random sample (SRS))

SRS is more informative than IID, but IID has easier math

---

when $n \ll N$ ← is a lot smaller than

(sample size) (pop. size)

,

SRS $\doteq$ IID

---

perfect similarity & at every time

randomization can't guarantee between & → (in all relevant ways)

① but the bigger n gets, the more likely that $\hat{p}_A$ & $\hat{p}_B$ are (relatively similar)

② we will learn methods to estimate how often randomization yields bad samples → unrepresentative of p

since sampling was at random,

$$\hat{\theta} = \hat{p} = \bar{y} = \frac{\Sigma}{n} = 4.4\%$$

is a good estimate of $\theta = p = \frac{\Sigma}{N}$

how good? we think θ is around 4.4%, give or take ?%

variables & the values they take on

variable	possible values
eye color	brown, blue
hair color	brown, black, red, white
success in maze running	very slow, slow, moderate, fast

dichotomous

ordered categorical (ordinal)

very fast

categorical = (non-numerical) = qualitative

← its values do not have unique places on number line