Stat 145, Thu 11-Feb-2021 -- Thu 11-Feb-2021
Biostatistics
Spring 2021

Thursday, February 11th 2021

Wk 2, Th
Topic:: Center and spread
Read:: Lock5 2.2-2.3

Start/join an Etherpad at link
  https://pad.disroot.org/p/s145-11feb2021-gXX
with
  XX = 01 for Latvaitis, Morren, Aardema
     = 02 for Toldy, Bultje, Katje
     = 03 for Triezenberg, Pastoor, Lemon
     = 04 for Steen, Ching, Tanis
     = 05 for Nedd, Rai, Brink
     = 06 for Ochiagha, Anderson, Cheek
     = 07 for Arthur, Stob, Sytsema
     = 08 for Johnson, Opalewski, Haveman
     = 09 for Rudy, Krikke
     = 10 for Wolf, Schneider, Wakeman

Examples of bias (and further thoughts):
 - In surveys: scenarios
    "Local library is sponsoring talk by Planned Parenthood representative.
    Do you think our community should sanction baby-killers?"
      leading questions
    Ann Landers on whether parents would choose to have children in do-over
      voluntary response bias
    Literary digest survey leading into 1936 election
      poor sampling frame
    "Do you take elicit drugs?"
      embarrassing question
    "How old were you when you stopped taking baths?"
      imperfect recall
    "Do you prefer this first soft drink, or the second one?"
      order of presentation should be random to avoid bias

"Which candidate did you vote for?", asked outside only during hours 7-9 am

- convenience sample

SRS

- its features
- not subject to voluntary bias, poor sampling frame, others
- often impractical

3:16 (book by Donald Knuth)

Q1: Is the sampling method Knuth uses an SRS?

- In experiments
  - measuring instrument not calibrated
  - order of treatment

Summary thoughts on "issues" surrounding experiments

- variables
  - explanatory:
    - may be more than one
    - each explanatory variable called a "factor"
      - like any variable, it takes on values---referred to as the "level"
      - a "treatment" is comprised of one combination of levels among factors
    - must be things researcher can assign to members of a treatment group
- comparison
  - more than one treatment
  - often there is a treatment with levels set to zero (control group)
- randomized assignment to treatments
  - avoids biases like volunteers getting a certain treatment, researcher will tend to make groups equal as concerning other variables
    - confounding vars, do not play a role
    - when a difference in response is observed as significant, get causality
- replication: the larger the number in each treatment group,
  - the more generally similar groups should be w/ respect to other variables
  - the greater the power to detect a (small?) difference
- blind and double-blind
- blocking
  - identifying specific (non-factor) variables to even out
    - example: soil, sunlight in agricultural studies
    - example: sex, smoking status, age in drug studies
  - matched pairs: each "case" contributes two values
    - case might be a person: contributes "control" and "treatment" values

Field

A	B
B	A

only 2 treatments

factor synonym for explanatory variable

level " " value of a var.

case might be identical twins: one twin is "control" for the other
case might be "married couple": one spouse is "control" for the other

Q2: How many treatments in the Physician's Health Study?
What are the levels?

What are the factors?

{ Beta-carotene (or not)
Aspirin (or not)

In relation to observational studies

- both types of studies may have explanatory/response vars
- observational study does not attempt to assign explanatory values
  ==> when difference appears significant, cannot rule out lurking vars
  in presence of significant difference only say vars have an association

----

Mode, median, and mean

- said to be measures of "center" (or "central tendancy")
- what they are
  mode = location/value occurring most frequently
  meaningful for both categorical and quantitative variables
  median = 50th percentile
  meaningful for quantitative variables only
  sequence of values matters, but not size ==> resistant to outliers
  mean = average
  formula $\bar{x} = \frac{1}{n} \sum x_i$
  meaningful for quantitative variables only
  size of values matters ==> sensitive to outliers
- median and mode app
  how to visualize

Q3: 5-number summary has 4 other numbers besides the median.
Are these other numbers resistant to outliers, or are they sensitive?

Q_1, Q_3

min., max

Range, interquartile range (IQR), standard deviation

- said to be measures of "spread" (or "variation")
- valid only for a quantitative variable
- what they are
  range =

  IQR =

  variance =

factors = expl. vars.

Case	got BC?	$\longrightarrow$	resp. Heart attack
1	No		yes
	No		no
	No		yes
	Yes		no
	Yes		no
			⋮

factor: "Got BC?", it's levels: Yes, No (2 levels)

"Got Aspirin?", it's levels: Yes, No (2 levels)

# of Treatment = product of # of levels

= (2)(2) = 4 treatments

Treatment 1: No BC, No aspirin (control group)

2: No BC, Yes aspirin

3: Yes BC, No aspirin

4: Yes BC, Yes asp.