Data collection and wrangling

The ASTA team

Contents

1 Data collection

1.1 Data collection

- Getting numbers to report is easy
- Getting sensible and trustworthy numbers to report is orders of magnitude more difficult

Ronald Fisher (1890-1962):

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Said about Fisher:

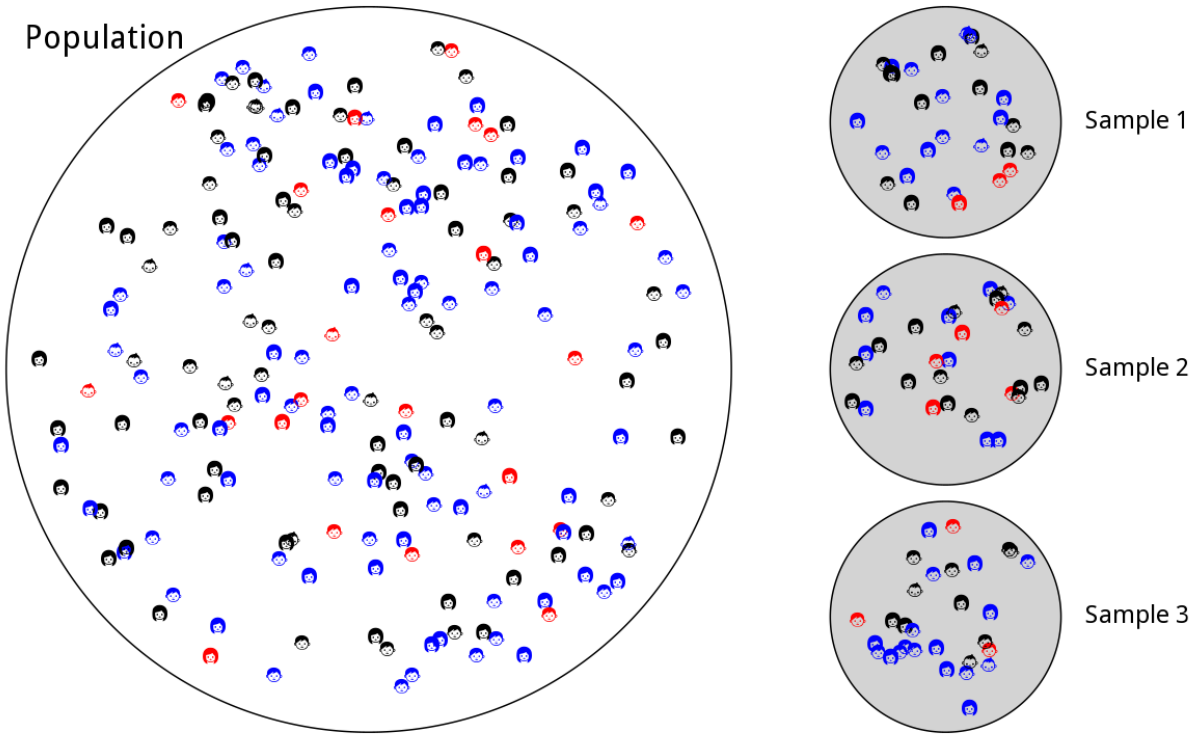
- Anders Hald (1913-2007), Danish statistician: "*a genius who almost single-handedly created the foundations for modern statistical science*"
- Bradley Efron (b. 1938): "*the single most important figure in 20th century statistics*"

1.2 Data collection

- Competences, ideally:
  - Statistics, both conceptually and analyses
  - Data wrangling (loading data; right format for analyses, tables, figures; ...)
  - Visualizations
  - Knowledge about subject (best with access to experts)
- Not just downloading a spreadsheet!
  - Population vs sample
  - Descriptives of the sample (e.g. mean)
  - Statistical inference about population (how close is sample's mean to population's mean)
- Do collect and analyze data, but know about pitfalls and limitations in generalisability!

2 Population and sample

2.1 Population and sample



Sample 3 of size $n = 30$:

shape	color	n_sample	p_sample	p_pop	p_diff
baby	black	2	0.07	0.04	-0.03
baby	blue	1	0.03	0.04	0.01
baby	red	0	0.00	0.01	0.01
man	black	5	0.17	0.12	-0.05
man	blue	8	0.27	0.22	-0.05
man	red	3	0.10	0.08	-0.02
woman	black	3	0.10	0.23	0.13
woman	blue	8	0.27	0.22	-0.05
woman	red	0	0.00	0.02	0.02

- Descriptive vs statistical inference.

3 Sample bias and non response bias

3.1 Example: United States presidential election, 1936

(Based on Agresti, this and this.)

- Current president: Franklin D. Roosevelt
- Election: Franklin D. Roosevelt vs Alfred Landon (Republican governor of Kansas)
- Literary Digest: magazine with history of accurately predicting winner of past 5 presidential elections

3.2 Example: United States presidential election, 1936

- Literary Digest poll ($\hat{\pi}$ and $1 - \hat{\pi}$): Landon: 57%; Roosevelt: 43%
- Actual results (π and $1 - \pi$): Landon: 38%; Roosevelt: 62%
- Sampling error: 57%-38% = 19%
  - Practically all of the sampling error was the result of **sample bias**
  - Poll size of > 2 mio. individuals participated – extremely large poll

3.3 Example: United States presidential election, 1936

- Mailing list of about 10 mio. names was created
  - Based on every telephone directory, lists of magasine subscribers, rosters of clubs and associations, and other sources
  - Each one of 10 mio. received a mock ballot and asked to return the marked ballot to the magazine
- "respondents who returned their questionnaires represented only that subset of the population with a relatively intense interest in the subject at hand, and as such constitute in no sense a random sample ... it seems clear that the minority of anti-Roosevelt voters felt more strongly about the election than did the pro-Roosevelt majority" (*The American Statistician*, 1976)
- Biases:
  - Selection bias
    - List generated towards middle- and upper-class voters (e.g. 1936 and telephones)
    - Many unemployed (club memberships and magazine subscribers)
  - Non-response bias
    - Only responses from 2.3/2.4 mio out of 10 million people
    - Cannot force people to participate: but mail may be junk (phone, interviews, online, pay/paid, ...)

4 Survivership bias

4.1 Example: Bullet holes of honor

(Based on this.)

- World War II
- Royal Air Force (RAF), UK
  - Lost many planes to German anti-aircraft fire
- Armor up!
  - Where?
  - Count up all the bullet holes in planes that returned from missions
    - Put extra armor in the areas that attracted the most fire

4.2 Example: Bullet holes of honor

- Hungarian-born mathematician Abraham Wald:
  - If a plane makes it back safely with a bunch of bullet holes in its wings: holes in the wings aren't very dangerous

    - **Survivorship bias**
  - Armor up the areas that (on average) don't have any bullet holes
    - They never make it back, apparently dangerous

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.80

5 Response bias

5.1 Example: New York Times/CBS News poll on attitude to increased fuel taxes

- "Are you in favour of a new gasoline tax?" - 12% said yes.
- "Are you in favour of a new gasoline tax to decrease US dependency on foreign oil?" - 55% said yes.
- "Do you think a new gas tax would help to reduce global warming?" - 59% said yes.

5.2 Example: Order of questions matter

US study during cold war asked two questions:

1 "Do you think that US should let Russian newspaper reporters come here and sent back whatever they want?"

2 "Do you think that Russia should let American newspaper reporters come in and sent back whatever they want?"

The percentage of yes to question 1 was 36%, if it was asked first and 73%, when it was asked last.

6 Theory: Biases / sampling

6.1 Biases

Agresti section 2.3:

- Sampling/selection bias
  - Probability sampling: each sample of size n has same probability of being sampled
    - Still problems: undercoverage, groups not represented (inmates, homeless, hospitalized, ...)
  - Non-probability sampling: probability of sample not possible to determine
    - E.g. volunteer sampling
- Response bias
  - E.g. poorly worded, confusing or even order of questions
  - Lying if think socially unacceptable
- Non-response bias
  - Non-response rate high; systematic in non-responses (age, health, believes)

6.2 Sampling

Agresti section 2.4:

- Random sampling schemes:
  - Simple sampling: each possible sample of equal size equally probable
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling
  - ...

7 Data wrangling

7.1 Data wrangling

This will be illustrated with two specific cases.

The material is on Moodle.