mass amounts of
Anecdotal evidence isn't reliable.

- small number of observations
  - ↳ natural variation can be small

- selection bias — participants might be predisposed in their participation. } reason for participation

- Confirmation Bias - participants may be more inclined to contribute in a way that re-affirms their view. } what they say cite.

- inaccuracy } People, memories and their psyche are fallible and error prone.

To solve for the anecdotal unreliability.

Can use stats tools:

1. Data Collection — of statistically valid/collected inference.

2. Descriptive Statistics — summarize data and vizualize with statistical method.

3. Exploratory Data Analysis — look for patterns, differences and features that address our question.
   Also sanity check for inconsistencies and limitations.

4. Estimation — sample data — then estimate

characteristics of the general population.

5. Hypothesis testing - where we see apparent effects like difference between two groups - determine if valid or not.

These steps allow for the extraction of meaningful justifiable determinations about data.

Surveys.

Cross sectional - snapshot of a group in time

(in gen. meant to be representative)

longitudinal study - over time observing a group.

Cycle - the number of times the survey was conducted.

Population - the target of the survey, which

group you are trying to
learn about.

<u>Sample</u> - a subset of the population which you collect data from.

<u>Respondents</u> - the people which participate in a survey.

<u>Representative</u> - where every member of the target population has an equal chance of participating.

<u>Oversampled</u> - to record certain groups deliberately more so than others.

→ This may allow for truer inference about those groups but ruins the ability to make conclusions about the general population.

[illegible]

Codebooks — detail the methodology of the survey.

Data Frames

a. datastructure apart of the Pandas library

An interface to.

- access via row & schema, variable name
- modification

Sample of API,

- .columns
- general access will dump an extract of the table's rows & columns

---

Variables — columns/metrics in the data.

recodes — not raw data → calculated raw data.

Based on logic that checks the consistency and accuracy of the data.

Usually worth using.

Transformation

after importing data, must correct the data. Called "data cleaning".

- check for errors
- deal with special values
- convert data into different formats.
- perform calculations

Validation

Validations

it is important to validate your assumptions about the data.

- can spring from.
  1. misunderstanding
  2. how the data has been handled.

Save time and effort by doing this before your projects.

⟹ Easy check is to compare the data & published counts.

Interpretation

Be a statician. but also remember the context of the data — and have empathy.