Data Validity

Outline

- The definition of data validity
- Different errors: data collection, data measurement, data preprocessing, model design…
- Case study:

  Google Flu Trend System (*https://time.com/23782/google-flu-trends-big-data-problems/*)

  Mice lab research

Validity and Reliability

- Validity refers to how accurately a method measures what it is intended to measure. If research has high validity, that means it produces results that correspond to real properties, characteristics, and variations in the physical or social world.
- Reliability refers to how consistently a method measures something. If the same result can be consistently achieved by using the same methods under the same circumstances, the measurement is considered reliable.
- High reliability is one indicator that a measurement is valid. If a method is not reliable, it probably isn’t valid.

Data Integrity and Data Validation

- Data Integrity is the assurance that information is unchanged from its source, and has not been accidentally (e.g. through programming errors), or maliciously (e.g. through breaches or hacks) modified, altered or destroyed. In another words, it concerns with the completeness, soundness, and wholeness of the data that complies with the intention of data creators.
- Data Validation is the tests and evaluations used to determine compliance with security specifications and requirements, in order to ensure correctness and reasonableness of data

Reliability & validity

- Reliability = Consistency
- Validity = Accuracy: Does the measure “hit the bullseye?”

There are four main types of validity:

- Construct validity: Does the test measure the concept that it's intended to measure?
- Content validity: Is the test fully representative of what it aims to measure?
- Face validity: Does the content of the test appear to be suitable to its aims?
- Criterion validity: Do the results correspond to a different test of the same thing?

Psychologists consider three types of reliability:

- over time (test-retest reliability)
- across items (internal consistency)
- across different researchers (inter-rater reliability)

There are four main types of validity:

- Construct validity: Does the test measure the concept that it's intended to measure?
- Content validity: Is the test fully representative of what it aims to measure?
- Face validity: Does the content of the test appear to be suitable to its aims?
- Criterion validity: Do the results correspond to a different test of the same thing?

Psychologists consider three types of reliability:

- over time (test-retest reliability)
- across items (internal consistency)
- across different researchers (inter-rater reliability)

- **Data collection** is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same.

Sample and Population

- A **population** is the entire group that you want to draw conclusions about.
- A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

The importance of ensuring accurate and appropriate data collection

1. Regardless of the field of study or preference for defining data (quantitative, qualitative), accurate data collection is essential to maintaining the integrity of research.
2. Consequences from improperly collected data include:

- inability to answer research questions accurately
- inability to repeat and validate the study
- distorted findings resulting in wasted resources
- misleading other researchers to pursue fruitless avenues of investigation
- compromising decisions for public policy
- causing harm to human participants and animal subjects

…

	Reliability	Validity
What does it tell you?	The extent to which the results can be reproduced when the research is repeated under the same conditions.	The extent to which the results really measure what they are supposed to measure.
How is it assessed?	By checking the consistency of results across time, across different observers, and across parts of the test itself.	By checking how well the results correspond to established theories and other measures of the same concept.
How do they relate?	A reliable measurement is not always valid: the results might be reproducible, but they're not necessarily correct.	A valid measurement is generally reliable: if a test produces accurate results, they should be reproducible.

Measurement theory

Rules that govern measurement

- Directness
- Levels
- Error
- Reliability
- Validity

Directness of measurement

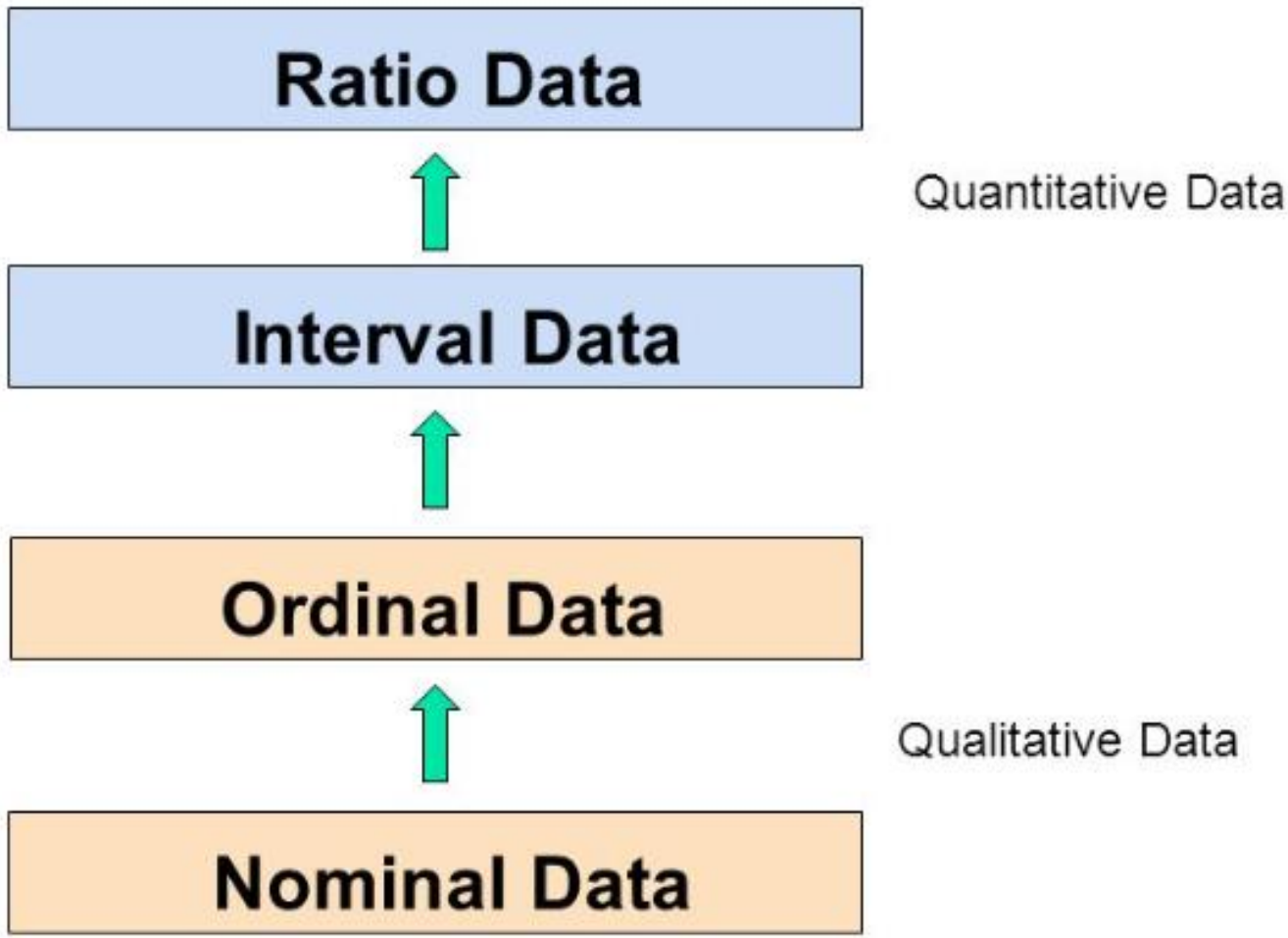
Direct

- Concrete
- Objective
- Examples: weight, demographics

Indirect

- Abstract
- Surrogate indicators
- Examples: depression, pain, symptom awareness

Types of Scales of Measurements



Primary Scales of Measurement

Scale	Basic Characteristics	Common Examples	Marketing Examples	Permissible Statisitics	
				Descriptive	*Inferential*
Nominal	Categorical variables, no quantitative meaning	Brand names, car model	Store types, brand nos.	Mode, percentages	Binomial test, chi-square
Ordinal	No quantitative meaning, have a definite order	Team rankings, quality ratings	Social class, market position	Mode Median, percentile	Friedman ANOVA, rank-order correlation
Interval	Have quantitative meaning & fixed order	Temperature (Fahrenheit)	Opinions, index, attitudes	Mode Median Standard, range, mean	Product-moment
Ratio	Quantitative meaning, definite order & fixed zero	Weight, length	Income, costs, sales, age	Harmonic mean, geometric mean	Coefficient of variation

Graphs

- Bar, Pie
- Bar, Pie, Stem and leaf
- Bar, Pie, Stem and leaf, Box plot, Histogram
- Histogram, Box plot

Sources of error

- Measurement error = Difference between the true measure and what is actually measured
  - Systematic error: the variation in measurement is in the same direction
  - Random error: the difference is without pattern

Data Collection Strategies

- How do researchers collect the data they need?
  - Tools include: surveys, measures, questionnaires, interviews, scales
- The strategy for data collection depends on research design
  - Qualitative study *vs.* Quantitative study

Summary

- Data collection strategies depend on the research design (qualitative vs. quantitative)
- Researcher must select (or develop) tools to measure variables of interest
  - Determine reliability and validity
  - Pilot testing needed if researcher develops new tool

Google Flu Trends (GFT)

- In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

Mice in lab research

- A lot of science is developed by studying animals, such as mice, in laboratories.
- It turns out that male and female mice have extremely different behaviors.
- However, the majority of research was done with solely male mice. But a significant minority on solely female mice.
- Also, various biological processes is not always the same across female and male mice.