Quantitative Social Science in the Age of Big Data and AI

Lecture 2: Measurement

Zhaopeng Qu
Hopkins-Nanjing Center
March 11 2025

This Week's Agenda

1. Measurement: Conceptualization and Operationalization

2. Data Management in R

Measurement in Social Sciences

Measurement in Social Sciences

- Quantitative social science is about understand social phenomena using **quantitative methods and theories to analyze social data**.
- The first step to quantitative analysis in social science is **measurement**, which is the process by which we describe social phenomena in terms of observable and measurable variables.
- A obivious difference between social science and natural science is that social phenomena are not always easy to observe and measure.
    - eg. the height of a person vs people's attitudes or opinions.

Measurement in Social Sciences

- Kaplan.A (1964) argues that objects in social science can be classified into three classes:
  - **direct observables**: can be measured directly, such as height, weight, income, schooling, etc.
  - **indirect observables**: can be measured indirectly, such as intelligence, personality, etc.
  - **constructs**: cannot be measured directly or indirectly, such as love, happiness, democracy, autocracy, etc.
- In social science,especially political science and sociology, **constructs** are the most common objects.

Kaplan, A. (1964). The conduct of inquiry: Methodology for behavioral science. San Francisco, CA: Chandler Publishing Company, p.55

Measurement: from Phenomena to Concepts

- A phenomenon is a collection of observable events or objects. eg. We could observe some people:
  - Talk a lot about men and women being equal.
  - Give speeches or write articles about men and women being equal.
  - Go to rallies or protests about the equal rights of men and women.
- When the phenomenon become more common, some people develop a **concept** to describe it:
  - **Feminism** or **Gender Equality**
- But in reality, concepts are constructed by people, and they do not exist in the objective world. They are the agreed-upon meanings we assign to terms.
  - Though everyone may have their own concept of "feminism".
  - Still have some consensus on the **definition** of "feminism".

Measurement: Conceptualization

- Because concepts are not objective, we need to **conceptualize** them in a way that is clear and concise.
  - **Conceptualization**: the process of defining concepts clearly and concisely.
  - Normally, you can find it in the literature or textbook of your field.
  - For example, the concept of "democracy" is often operationalized as the **degree of electoral participation** or the **extent of political freedoms**.
- Some concepts may have multiple dimensions, which means that they can be operationalized in multiple ways.
  - For example, the concept of "democracy" has multiple dimensions, such as **electoral participation**, **political freedoms**, **civil liberties**, and **economic development**.
- The point here is that the concept is accepted by the community of your field.

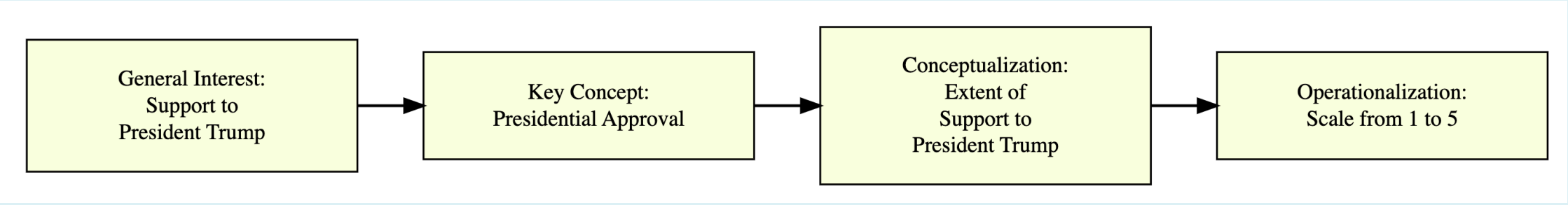
Measurement: Operationalization

- **Operationalization**: the process of translating concepts into measurable variables.
- It is a process which highly related to how you collect your data and analyze your data.
  - Survey data: questionaire design is the process of operationalization.
- Normally, you have to identify(some time have to create) **specific indicators** to measure the concept, which is called **variables**.
- For example, the concept of "democracy" can be operationalized as the **degree of electoral participation** or the **extent of political freedoms**.

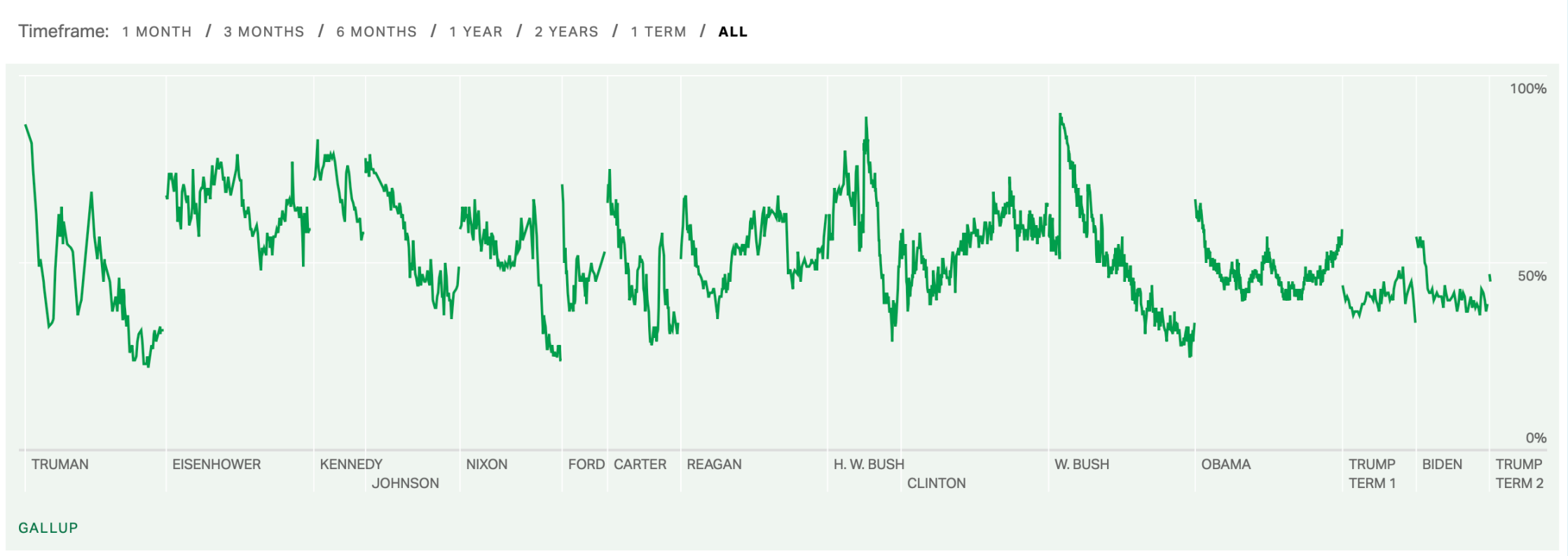
Example: Presidential Support

- **Topic**: Presidential Support
- **Concept**: presidential job approval
- **Conceptualization**:
  - Extent to which US adults support the actions and policies of the current US president
- **Operationalization**:

> "On a scale from 1 to 5, where 1 is least supportive and 5 is most supportive, how much would you say you support the job that Donald Trump is doing as president?"



Example: Presidential Approval in the US by Gallup



Presidential job approval 1945-2024

- Gallup poll

Measurement: Level of Measurement

- **Variables** or **features** are the characteristics or properties of the observations in the data set, which can be broadly classified into two types:
- **Quantitative**: continuous or discrete numerical values. These numbers represent a quantity or amount.
  - Interval(定距): the difference between two values is *meaningful and consistent*, but only can be used to calculate with **addition and subtraction**.
  - Ratio(定比): the difference between two values is *meaningful and consistent*, and can be used to calculate with **addition, subtraction, multiplication and division**.
- **Qualitative**: specific numbers used to represent different groups or categories.
  - Nominal(定类): the difference between two values is *meaningless or arbitrary*, and no order.
  - Ordinal(定序): the difference between two values is also *meaningless or arbitrary*, but the order is meaningful.

Quantitative Variables v.s. Qualitative Variables

TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

- **Quantitative Variables**: wage, education(years), experience are all **interval** variables.
- **Qualitative Variables**: gender, race are all **nominal** variables.

Variables or Features in Data

- The number of variables is the number of columns in the data set,represents by p,
  - which is normally **less than** the number of observations in the **traditional data set**, represents by N

$$p \ll N$$

  - However, in the **big data** era, the number of variables can be **larger than** the number of observations, represents by

$$p \gg N$$

- eg. Text data, image data, video data, audio data, etc.

Features in Big Data

TABLE 14.1 Variables in the 817-Predictor School Test Score Data Set

Main variables (38)

- Fraction of students eligible for free or reduced-price lunch
- Fraction of students eligible for free lunch
- Fraction of English learners
- Teachers' average years of experience
- Instructional expenditures per student
- Median income of the local population
- Student–teacher ratio
- Number of enrolled students
- Fraction of English-language proficient students
- Ethnic diversity index
- Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
- Number of teachers
- Fraction of first-year teachers
- Fraction of second-year teachers
- Part-time ratio (number of teachers divided by teacher full-time equivalents)
- Per-student expenditure by category, district level (7)
- Per-student expenditure by type, district level (5)
- Per-student revenues by revenue source, district level (4)

+ Squares of main variables (38)

+ Cubes of main variables (38)

+ All interactions of main variables $(38 \times 37/2 = 703)$

Total number of predictors $= k = 38 + 38 + 38 + 703 = 817$

Measurement Quality

- **Question**: How do we know that our measures are any good?
- **Answer**: **Reliability** and **Validity** are two key concepts in measurement quality.
- **Reliability**: the consistency of a measurement, which normally refers to the consistency of a measurement across time.
  - eg, **the height of a person** v.s **happiness of a person**.
- **Validity**: the accuracy of a measurement, which normally refers to the extent to which a measurement measures what it is intended to measure.
  - eg. "Have you ever had a **problem** with alcohol?"
  - different people may have different understanding of what "problem" means.

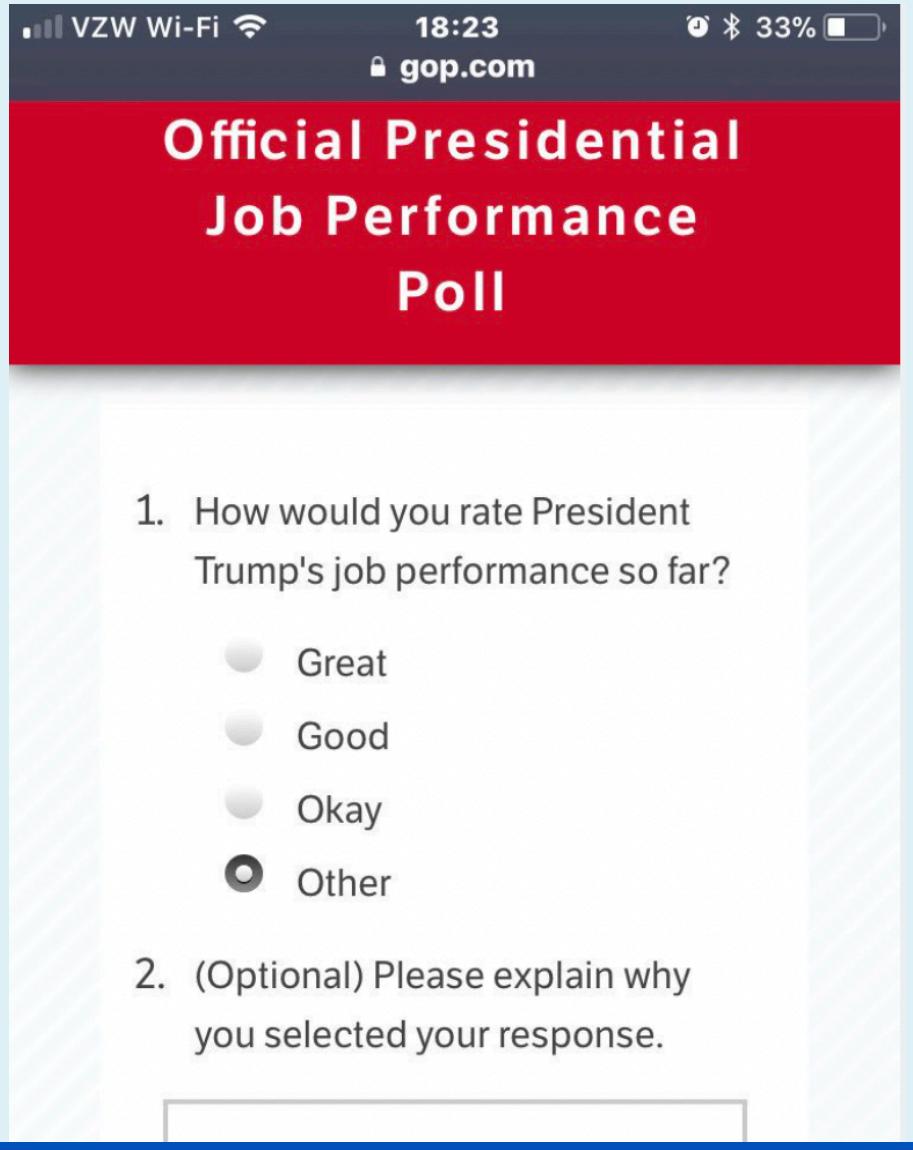
Measurement: Errors vs Bias

- **Measurement error**: chance variation in our measurements, normally can be seen as random noise.
  - Individual measurement = exact value + chance error
  - Chance errors tend to cancel out when we take averages when the sample size is large.
- No matter how careful we are, chance error can always exist in a measurement.
  - Eg. Gallup poll of 19,000 respondents: someone reported being a citizen in 2010 and then a non-citizen in 2012
    - Data entry errors or misreporting or other reasons?

Measurement: Errors vs Bias

- **Bias**: systematic errors for all units in the same direction.
  - individual measurement = exact value + bias + chance error
- Bias can be caused by many reasons,
  - The most common one is **response bias**, which is caused by the way we ask questions.
  - eg. "What did you eat yesterday evening" answers may be underreported of unhealthy food.

Measurement: Presidential Approval

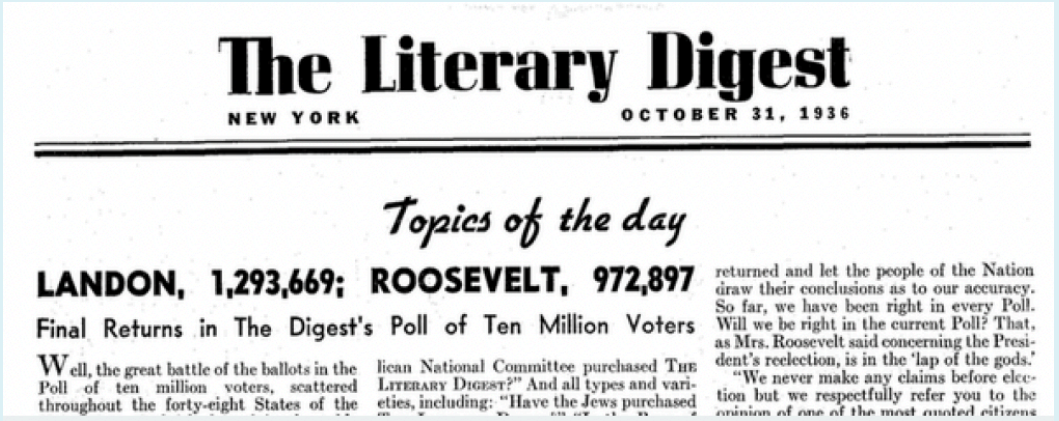


Measurement: Sampling is critical

- Always need to figure out what are **population** and **sample** in the measurement process.
  - **Population**: the entire group of individuals or objects that we are interested in.
  - **Sample**: a subset of the population that we actually study.
- **Sampling**: How do we select a subset of the population to study, which is supposed to be representative of the population.

A Failed Sampling: 1936 Literary Digest Poll

- Literary Digest used **mail-in polls** to get the sample.
  - Source of addresses: automobile registrations, phone books, etc.
  - In 1936, sent out 10 million ballots, over 2.3 million returned.
- George Gallup used only 50,000 respondents but a more careful sampling method.



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

1948 Election

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- Gallup's failure:
  - **Quota sampling**: fixed quota of certain respondents for each interviewer base on the population distribution.
  - The population distribution in 1948 is different from the census in 1940. And republicans easier to find within quotas (phones, listed addresses).

Sampling Methods

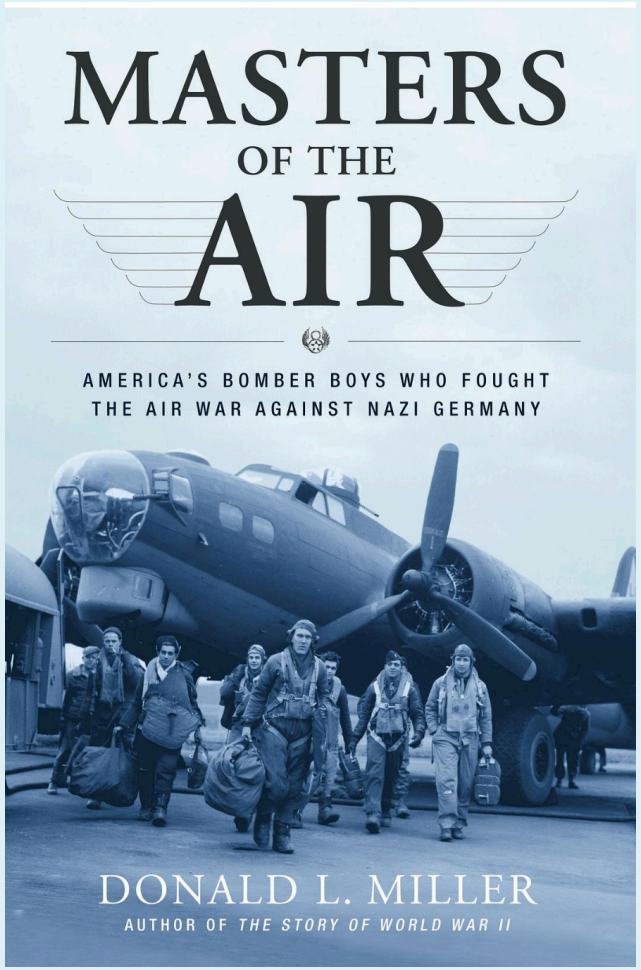
- **Non-probability sampling**: every unit in the population does not have a known, non-zero probability of being selected into sample.
    - eg. Convenience sampling, quota sampling, snowball sampling, etc.
- **Probability sampling**: every unit in the population has a known, non-zero probability of being selected into sample.
    - eg.**Simple random sampling**: every unit has an equal selection probability.
- How to select a sample randomly from a population?
    - eg, Random digit dialing strategy in a telephone survey: Take a particular area code + exchange

Sampling Methods: Probability Sampling

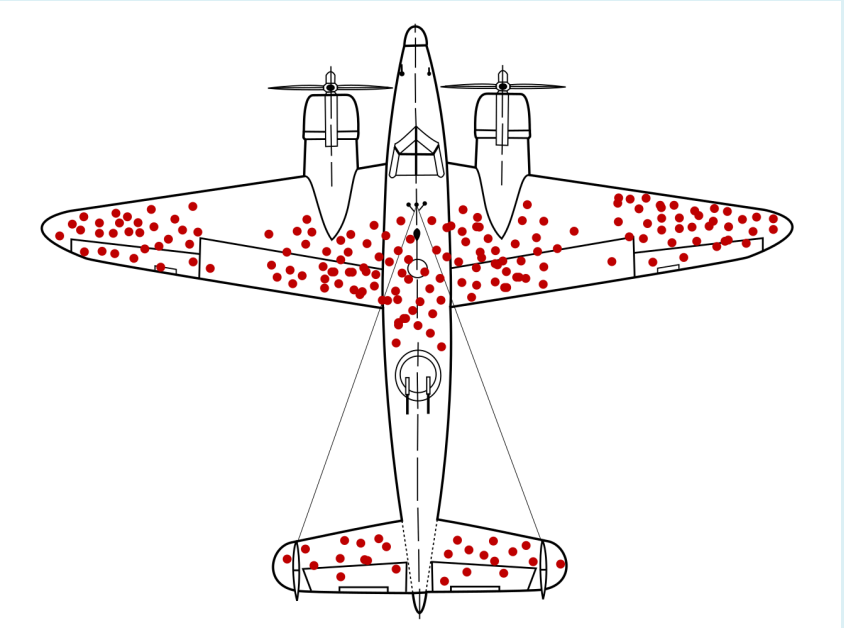
Sampling.Type	Description
Simple random	Researcher randomly selects elements from sampling frame.
Systematic	Researcher selects every kth element from sampling frame.
Stratified	Researcher creates subgroups then randomly selects elements from each subgroup.
Cluster	Researcher randomly selects clusters then randomly selects elements from selected clusters.

- In the reality, the sampling methods in large surveys are generally more complex than the above four types.
- Whatever the sampling method is, the goal is to get a **representative sample** from the population.
- Always need to be aware of the potential **sampling bias** in the data, especially in the big data era.

Survivorship Bias from WWII Aircrafts



- The bullet holes of a bomber that, crucially, survived



- **Question**: How to reinforce the armor to increase the survival of allied bombers?

More bias on Data Collection

- Even if we have a representative sample, we still suffer from potential **sampling bias** in the data.
- Respondents vs Non-respondents in samples
  - The non-respondents may be different from the respondents.
- Item completion and Item non-response
  - Some respondents may refuse to answer some certain questions.
- All will induce to **Missing Data**, which is a big problem in the big data era.
- Missing data is very common in data collection. Almost all data sets(survey data, big data, etc.) suffer from missing data more or less.

Missing Data

- The most important thing is to be aware the reason of missing data:
  - **Missing completely at random (MCAR)**: the missingness is **completely random** and unrelated to the values of the variables.
    - Normally, we can ignore MCAR data, just delete them.
  - **Missing at random (MAR)**: the missingness is related to the values of variables or some other observable variables.
    - eg. wealthy people who are more likely to refuse to answer the question about their income.
    - Delete the missing data may induce bias. Normally we can use **multiple imputation** to deal with MAR data.
  - **Missing not at random (MNAR)**: the missingness is related to some unobserved variables.
    - In this case, we have to use more sophisticated methods to deal with the missing data.

Data Management in R