Module 03
Data Discovery
Lecture 1

Wan Fang

Southern University of Science and Technology

Agenda

- Data Quality Assessment

- Statistical Analysis Step by Step
  - Data Collection
  - Probability Distribution
  - Statistical Tests

Data Quality Assessment

Wan Fang
Southern University of Science and Technology

[Adapted from 10.5334/dsj-2015-002 by Li Cai and Yangyong Zhu]

Data Quality Assessment

10.5334/dsj-2015-002

> garbage in, garbage out
>
> [garbage in, garbage out]
>
> DEFINITION
>
> used to express the idea that in computing and other fields, incorrect or poor-quality input will produce faulty output.
>
> *Data from Oxford Languages*

"Garbage in, Garbage out …"

- **High-quality data are the precondition for analyzing and using big data and for guaranteeing the value of the data.**

- Features of big data (Katal, Wazid, & Goudar, 2013)
  - Volume
    - refers to the tremendous volume of the data. We usually use TB or above magnitudes to measure this data volume.
  - Velocity
    - means that data are being formed at an unprecedented speed and must be dealt with in a timely manner.
  - Variety
    - indicates that big data has all kinds of data types, and this diversity divides the data into structured data and unstructured data. These multityped data need higher data processing capabilities.
  - Value
    - represents low-value density. Value density is inversely proportional to total data size, the greater the big data scale, the less relatively valuable the data.

The Challenges of Data Quality

The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.

- Big data sources are very wide, including:
  - 1) data sets from the internet and mobile internet (Li & Liu, 2013);
  - 2) data from the Internet of Things;
  - 3) data collected by various industries;
  - 4) scientific experimental and observational data: such as high-energy physics experimental data, biological data, and space observation data.

- These sources produce rich data types.
  - *unstructured* data: documents, video, audio, etc, occupies more than 80% of the total amount of data .
  - *semi-structured* data: software packages/modules, spreadsheets, and financial reports.
  - *structured* data.

- As for enterprises, obtaining big data with complex structure from different sources and effectively integrating them are a daunting task (McGilvray, 2008).
  - *conflicts* and *inconsistent* or *contradictory* phenomena among data from different sources.
  - In the case of small data volume, the data can be checked by a manual search or programming, even by ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform).
  - However, these methods are useless when processing PB-level even EB-level data volume.

The Challenges of Data Quality

Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.

Decimal		
Value		Metric
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

- In 2011, the amount of global data created and copied reached 1.8 ZB.
  - After the industrial revolution, the amount of information dominated by characters doubled every **ten years**.
  - After 1970, the amount of information doubled every **three years**.
  - Today, the global amount of information can be doubled **every two years**.

- A great challenge to the existing techniques of data processing quality.
  - It is difficult to collect, clean, integrate, and finally obtain the necessary high-quality data within a reasonable time frame.
  - Unstructured data in big data is very high, it will take a lot of time to transform unstructured types into structured types and further process the data.

The Challenges of Data Quality

Data change very fast, and the "timeliness" of data is very short, which necessitates higher requirements for processing technology.

- Due to the rapid changes in big data, the "timeliness" of some data is very short.
    - If companies can't collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information.
- Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making mistakes by governments or enterprises.
- At present, real-time processing and analysis software for big data is still in development or improvement phases.

10.5334/dsj-2015-002

The Challenges of Data Quality

No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun.

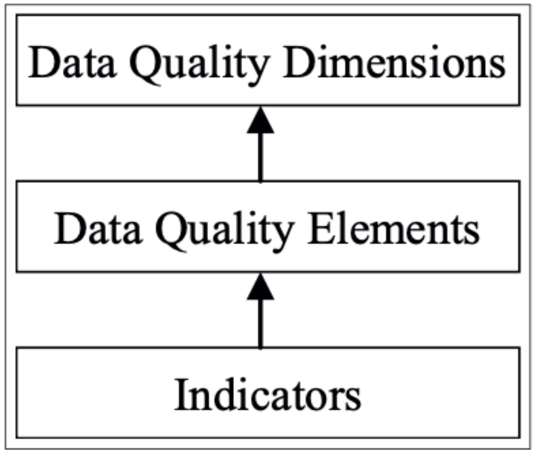
- In order to guarantee the product quality and improve benefits to enterprises, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards.
  - Nowadays, there are more than 100 countries and regions all over the world actively carrying out these standards.
  - This implementation promotes mutual understanding among enterprises in domestic and international trade and brings the benefit of eliminating trade barriers.

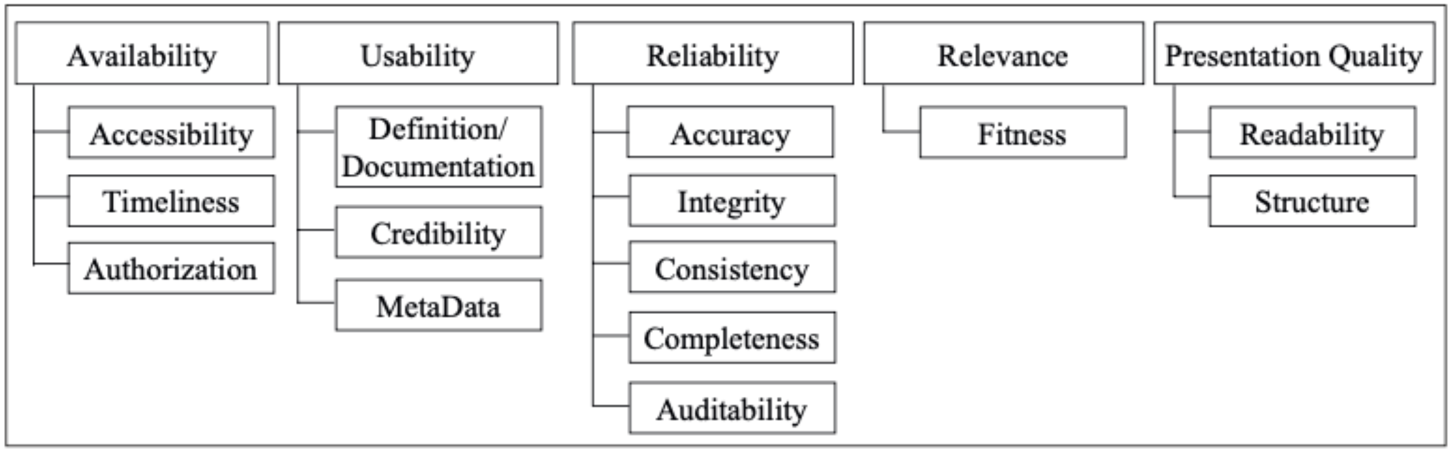
- By contrast, the study of data quality standards began in the 1990s, but not until 2011 did ISO published ISO 8000 data quality standards (Wang, Li, & Wang, 2010).
  - At present, more than 20 countries have participated in this standard, but there are many disputes about it.
  - The standards need to be mature and perfect.

10.5334/dsj-2015-002

Quality Criteria of Big Data

- Academia hasn't made a uniform definition of its data quality and quality criteria
- But one thing is certain:
  - Data quality depends not only on its own features but also on the business environment using the data, including business processes and users.
- Only the data that conform to the relevant uses and meet requirements can be considered qualified (or good quality) data
  - A hierarchical data quality standard from the perspective of the users





10.5334/dsj-2015-002

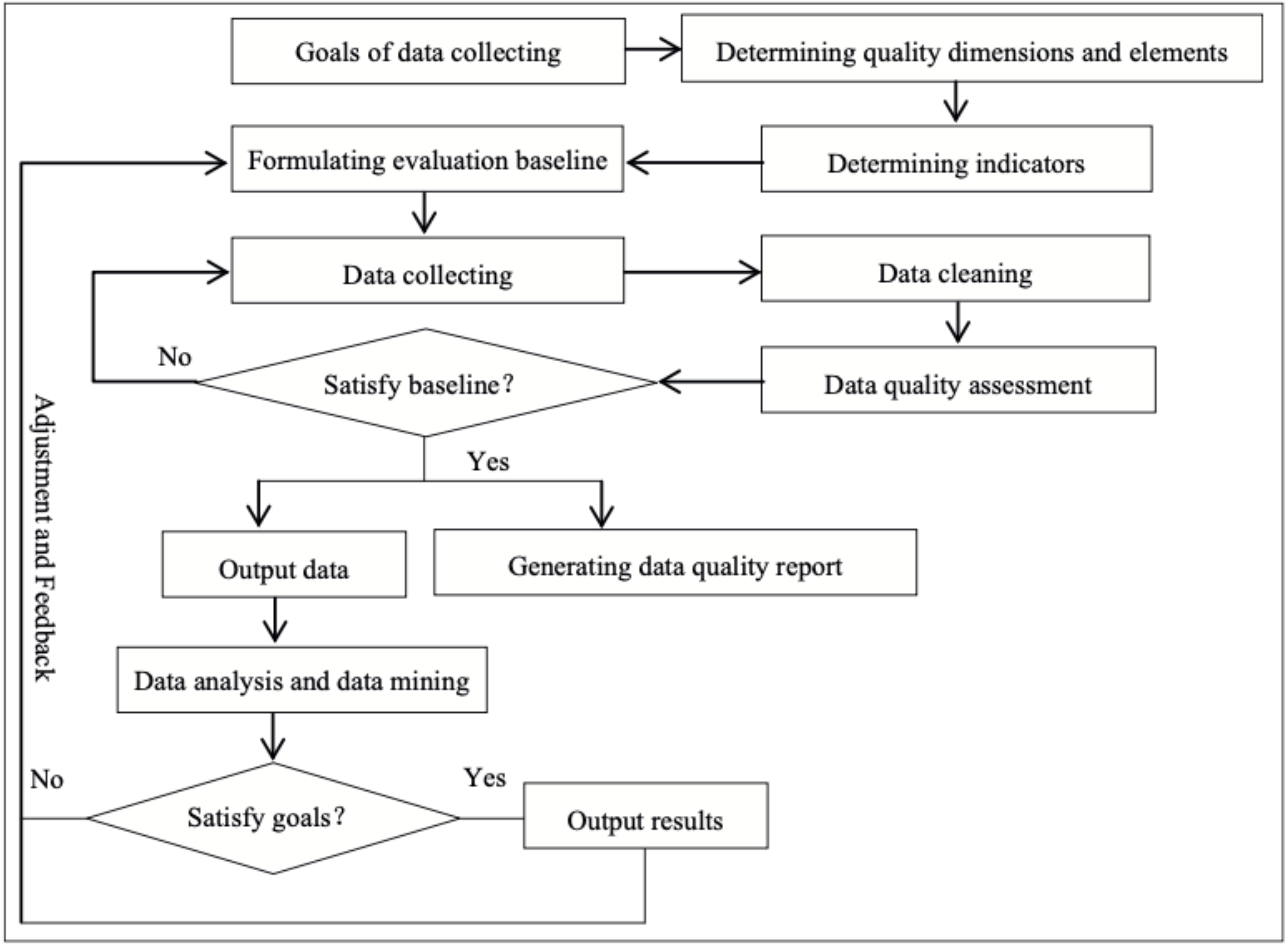
A Hierarchical Big Data Quality Assessment Framework

Discussion: Which elements are important in evaluating **social media data**?

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	▪ Whether a data access interface is provided ▪ Data can be easily made public or easy to purchase
	2) Timeliness	▪ Within a given time, whether the data arrive on time ▪ Whether data are regularly updated ▪ Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	▪ Data come from specialized organizations of a country, field, or industry ▪ Experts or specialists regularly audit and check the correctness of the data content ▪ Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	▪ Data provided are accurate ▪ Data representation (or value) well reflects the true state of the source information ▪ Information (data) representation will not cause ambiguity
	2) Consistency	▪ After data have been processed, their concepts, value domains, and formats still match as before processing ▪ During a certain time, data remain consistent and verifiable ▪ Data and the data from other data sources are consistent or verifiable
	3) Integrity	▪ Data format is clear and meets the criteria ▪ Data are consistent with structural integrity ▪ Data are consistent with content integrity
	4) Completeness	▪ Whether the deficiency of a component will impact use of the data for data with multi-components ▪ Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	▪ The data collected do not completely match the theme, but they expound one aspect ▪ Most datasets retrieved are within the retrieval theme users need ▪ Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	▪ Data (content, format, etc.) are clear and understandable ▪ It is easy to judge that the data provided meet needs ▪ Data description, classification, and coding content satisfy specification and are easy to understand

10.5334/dsj-2015-002

Quality Assessment Process For Big Data



Source of Sample Data

salesforce

+tableau public | Create | Resources

Tableau Public will be unavailable from 3/19/2023 00:00 to 04:00 GMT+8 for maintenance. Thanks for your patience w

Resources

Explore how-to videos, sample data, and community resources to help you get started or to take your skills to the next level.

Learn | Sample Data | Community Resources

Explore these sample data sets, data sources, and web data connectors to get started on your next visualization project. Down to start creating. Data sets may be available in English only.

Business

Superstore Sales	Contains information about products, sales, and profits that you can use to identify key areas of improvement within this fictitious company.	**Dataset (xls)**
The 2014 Inc. 5000	The Inc. 5000 is Inc. Magazine's annual list of the 5000 fastest growing private companies in the United States. The list is compiled by measuring each company's percentage revenue growth over a four-year period.	**Dataset (csv)**

Sources for Data Sets

Explore publicly available data sets. Don't forget to check that the data is well-structured!

- Makeover Monday
- data.world
- Data Is Plural
- UN Data
- Data.gov
- Kaggle
- NOAA
- Reddit
- The World Factbook
- UN Environment Programme GRID-Geneva
- World Health Organization

Find More Data Sources

Web Data Connectors

Connect to data housed in a cloud database. To learn how to use web data connectors, see Creators: Connect to Data on the Web.

- English Premier League
- Fitbit
- NYT Best Sellers
- Google Places
- USGS Earthquake Data
- Facebook Page Feed
- Facebook Page Insights
- Twitter

See More on Github

Statistical Analysis Step by Step

Wan Fang
Southern University of Science and Technology

[Adapted from Statistics by Scribbr]

A Beginner's Guide to Statistical Analysis

- Investigating trends, patterns, and relationships using quantitative data.
  - An important research tool used by scientists, governments, businesses, and other organizations.
- Statistical analysis planning.
  - Specify your hypotheses and make decisions about your *research design*, *sample size*, and *sampling procedure*.
  - **After collecting data** from your sample, you can organize and summarize the data using ***descriptive statistics***.
  - **Then**, you can use ***inferential statistics*** to formally test hypotheses and make estimates about the population.
  - **Finally**, interpret and generalize your findings.

Example: Causal research question

- Can meditation improve exam performance in teenagers?

Example: Correlational research question

- Is there a relationship between parental income and college grade point average (GPA)?

Step 1: Write your hypotheses and plan your research design

Step 2: Collect data from a sample

Step 3: Summarize your data with descriptive statistics

Step 4: Test hypotheses or make estimates with inferential statistics

Step 5: Interpret your results

Step 1: Write your hypotheses and plan your research design

- **Writing statistical hypotheses**
  - The goal of research is often to investigate a relationship between variables within a population.
  - You start with a prediction and use statistical analysis to test that prediction.

- *A statistical hypothesis is a formal way of writing a prediction about a population.*
  - Every research prediction is rephrased into ***null*** and ***alternative*** hypotheses that can be tested using sample data.
  - While the null hypothesis always predicts no effect or no relationship between variables, the alternative hypothesis states your research prediction of an effect or relationship.

Step 1: Write your hypotheses and plan your research design

Example: Statistical hypotheses to test an effect

- ***Null hypothesis***: A 5-minute meditation exercise will have no effect on math test scores in teenagers.
- ***Alternative hypothesis***: A 5-minute meditation exercise will improve math test scores in teenagers.

Example: Statistical hypotheses to test a correlation

- ***Null hypothesis***: Parental income and GPA have no relationship with each other in college students.
- ***Alternative hypothesis***: Parental income and GPA are positively correlated in college students.

- *While the **null hypothesis** always predicts no effect or no relationship between variables, the **alternative hypothesis** states your research prediction of an effect or relationship.*

Step 1: Write your hypotheses and plan your research design

- **Planning your research design**
  - A strategy for data collection and analysis.
  - It determines the statistical tests you can use to test your hypothesis later on.

- **Decide which is your research design.**
  - In an ***experimental*** design, you can *assess a cause-and-effect relationship* using statistical tests of comparison or regression.
    - E.g., the effect of meditation on test scores
  - In a ***correlational*** design, you can *explore relationships between variables* without any assumption of causality using correlation coefficients and significance tests.
    - E.g., parental income and GPA
  - In a ***descriptive*** design, you can *study the characteristics of a population or phenomenon* using statistical tests to draw inferences from sample data.
    - E.g., the prevalence of anxiety in U.S. college students

- **Whether you'll compare participants at the group level or individual level, or both.**
  - In a ***between-subjects*** design, you compare the *group-level outcomes* of participants who have been exposed to different treatments.
    - E.g., those who performed a meditation exercise vs those who didn't
  - In a ***within-subjects*** design, you compare *repeated measures from participants* who have participated in all treatments of a study.
    - E.g., scores from before and after performing a meditation exercise
  - In a ***mixed (factorial)*** design, *one variable is altered between subjects* and *another is altered within subjects*.
    - E.g., pretest and posttest scores from participants who either did or didn't do a meditation exercise

Step 1: Write your hypotheses and plan your research design

Example: Variables (experiment)

- You can perform many calculations with quantitative age or test score data, whereas categorical variables can be used to decide groupings for comparison tests.

Example: Variables (correlational)

- The types of variables in a correlational study determine the test you'll use for a correlation coefficient.
- A parametric correlation test can be used for quantitative data, while a non-parametric correlation test should be used if one of the variables is ordinal.

Variable	Type of data
Age	Quantitative (ratio)
Gender	Categorical (nominal)
Race or ethnicity	Categorical (nominal)
Baseline test scores	Quantitative (interval)
Final test scores	Quantitative (interval)

Variable	Type of data
Parental income	Quantitative (ratio)
GPA	Quantitative (interval)

Step 1: Write your hypotheses and plan your research design

Example: **Experimental research design**

- You design a within-subjects experiment to study whether a 5-minute meditation exercise can improve math test scores. Your study takes repeated measures from one group of participants.
- First, you'll take baseline test scores from participants. Then, your participants will undergo a 5-minute meditation exercise. Finally, you'll record participants' scores from a second math test.
- In this experiment, the independent variable is the 5-minute meditation exercise, and the dependent variable is the math test score from before and after the intervention.

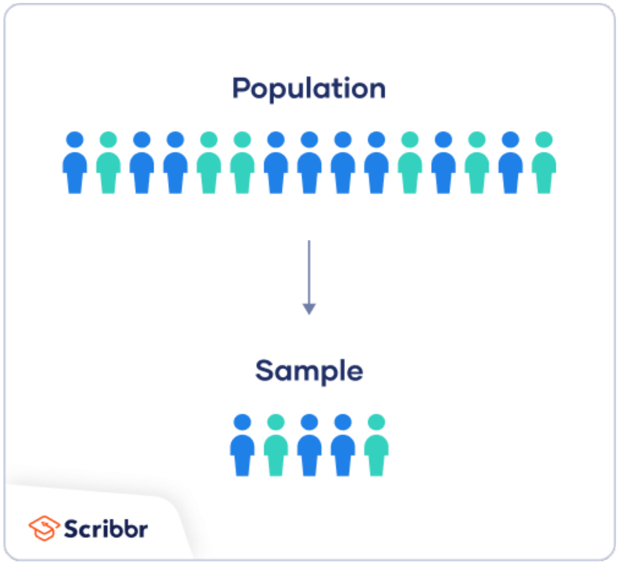
Example: **Correlational research design**

- In a correlational study, you test whether there is a relationship between parental income and GPA in graduating college students. To collect your data, you will ask participants to fill in a survey and self-report their parents' incomes and their own GPA.
- There are no dependent or independent variables in this study, because you only want to measure variables without influencing them in any way.

Step 2: Collect data from a sample

- In most cases, it's too difficult or expensive to collect data from every member of the population you're interested in studying. Instead, you'll collect data from a sample.

- Statistical analysis allows you to apply your findings beyond your own sample as long as you use appropriate sampling procedures.
  - Aim for a sample that is representative of the population.



Sampling for statistical analysis: Two main approaches to selecting a sample.

- **Probability sampling**: every member of the population has a chance of being selected for the study through random selection.

- **Non-probability sampling**: some members of the population are more likely than others to be selected for the study because of criteria such as convenience or voluntary self-selection.

Step 2: Collect data from a sample

Create an appropriate sampling procedure

- Based on the resources available for your research, decide on how you'll recruit participants.
  - Will you have resources to advertise your study widely, including outside of your university setting?
  - Will you have the means to recruit a diverse sample that represents a broad population?
  - Do you have time to contact and follow up with members of hard-to-reach groups?

- **Example: Can meditation improve exam performance in teenagers?**
  - The population you're interested in is high school students in your city. You contact three private schools and seven public schools in various districts of the city to see if you can administer your experiment to students in the 11th grade.
  - Your participants are self-selected by their schools. **Non-probability** sample.

- **Example: Is there a relationship between parental income and college grade point average (GPA)?**
  - Male college students in the US. Using social media advertising, you recruit senior-year male college students from a smaller subpopulation: seven universities in the Boston area.
  - Your participants volunteer for the survey. **Non-probability** sample.

Step 2: Collect data from a sample

Calculate sufficient sample size

- Before recruiting participants, decide on your sample size either by looking at other studies in your field or using statistics. A sample that's too small may be unrepresentative of the sample, while a sample that's too large will be more costly than necessary.

- There are many sample size calculators online. Different formulas are used depending on whether you have subgroups or how rigorous your study should be (e.g., in clinical research). As a rule of thumb, a minimum of 30 units or more per subgroup is necessary.

- To use these calculators, you have to understand and input these key components:
  - **Significance level (alpha):** the risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.
  - **Statistical power**: the probability of your study detecting an effect of a certain size if there is one, usually 80% or higher.
  - **Expected effect size**: a standardized indication of how large the expected result of your study will be, usually based on other similar studies.
  - **Population standard deviation**: an estimate of the population parameter based on a previous study or a pilot study of your own.

Step 3: Summarize your data with descriptive statistics

- Once you've collected all of your data, you can inspect them and calculate descriptive statistics that summarize them.

Inspect your data

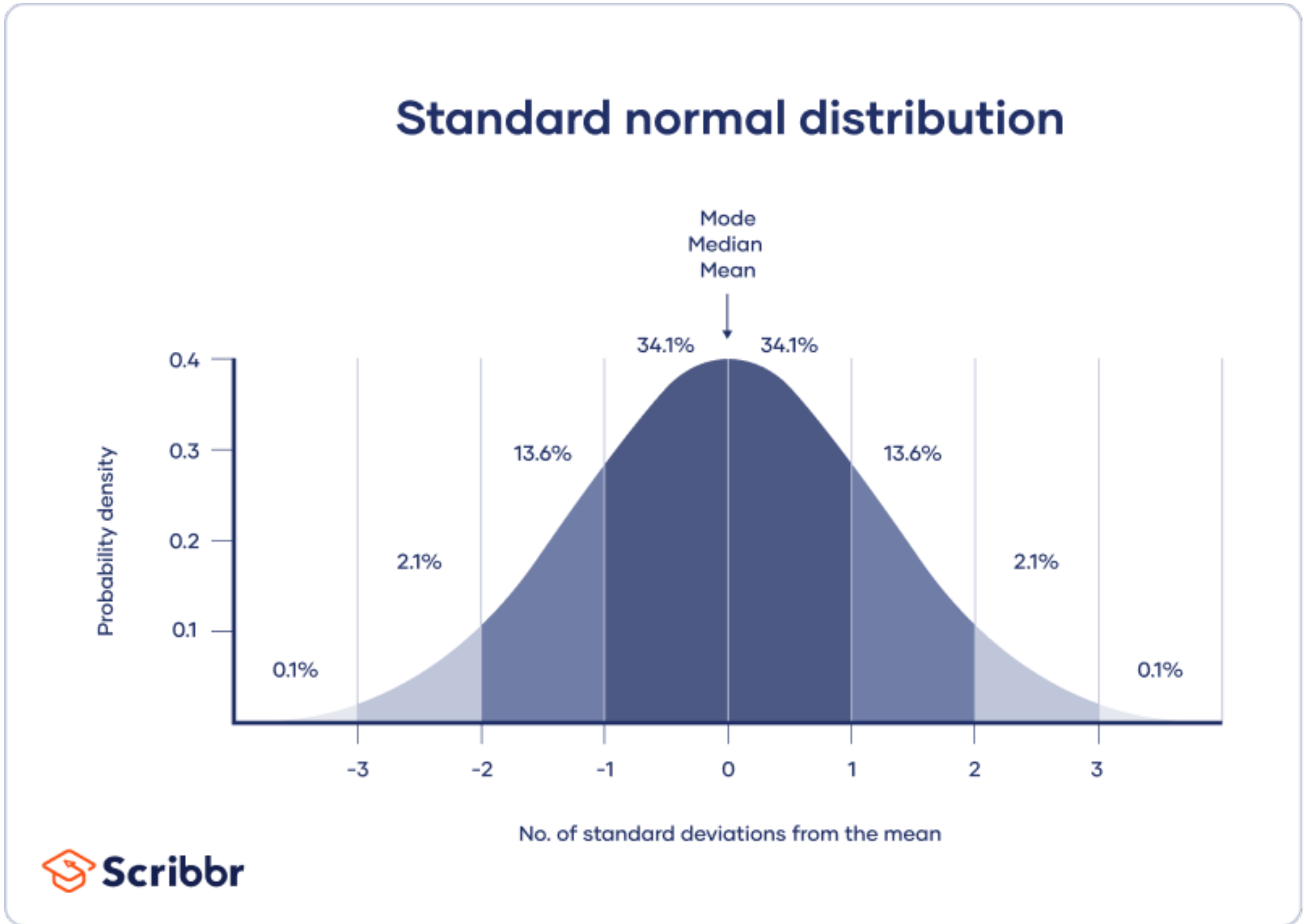
- There are various ways to inspect your data, including the following:
  - Organizing data from each variable in frequency distribution tables.
  - Displaying data from a key variable in a bar chart to view the distribution of responses.
  - Visualizing the relationship between two variables using a scatter plot.

- By visualizing your data in tables and graphs, you can assess whether your data follow a skewed or normal distribution and whether there are any outliers or missing data.

Step 3: Summarize your data with descriptive statistics

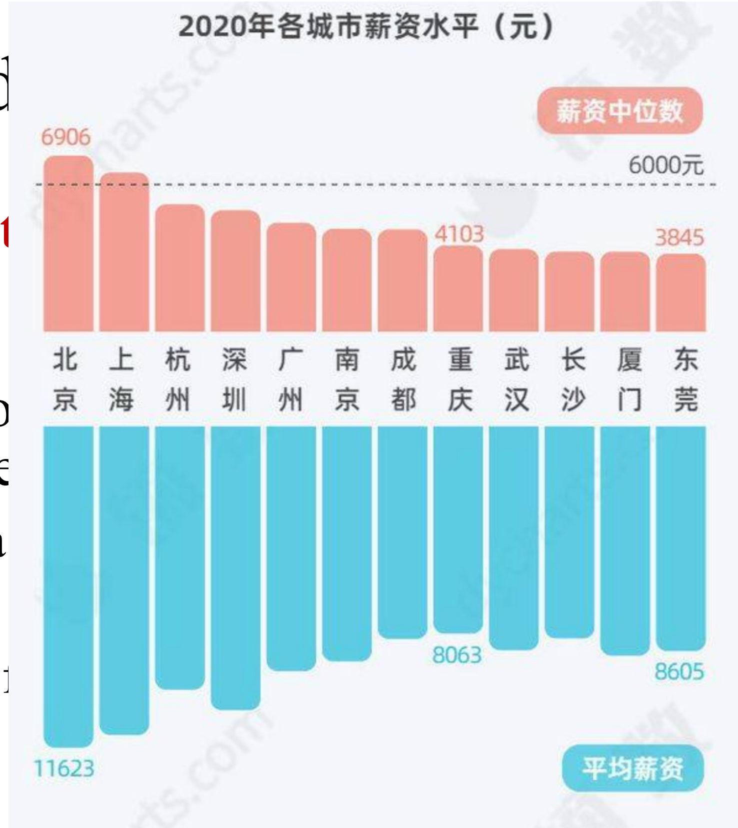
- A **normal distribution** means that your data are symmetrically distributed around a center where most values lie, with the values tapering off at the tail ends.

- In contrast, a **skewed distribution** is asymmetric and has more values on one end than the other. The shape of the distribution is important to keep in mind because only some descriptive statistics should be used with skewed distributions.



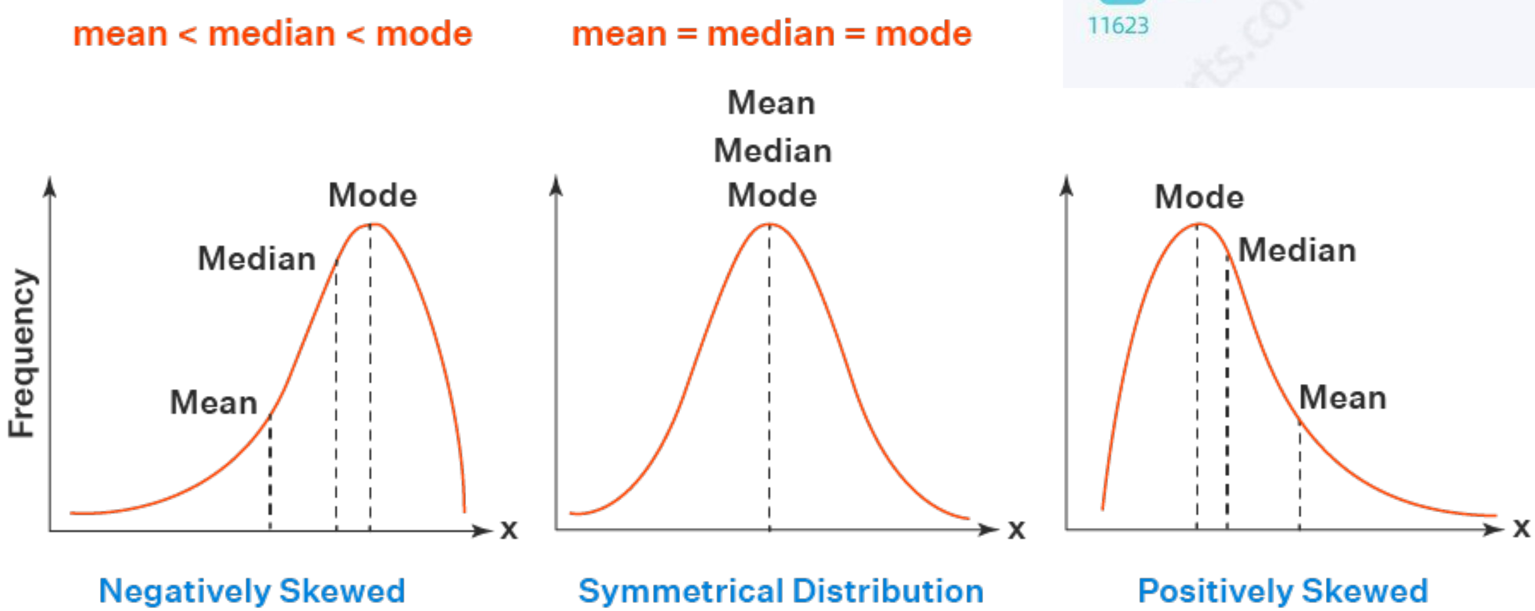
Extreme outliers can also produce misleading statistics, so you may need a systematic approach to dealing with these values.

Step 3: Summarize your data with d



Calculate measures of central t

- Measures of ***central tendency*** describe where most o Three main measures of central tendency are often re
  - **Mode**: the most popular response or value in the data
  - **Median**: the value in the exact middle of the data set
  - **Mean**: the sum of all values divided by the number o



Step 3: Summarize your data with descriptive statistics

Calculate measures of variability

- Measures of *variability* tell you how spread out the values in a data set are.

- Four main measures of variability are often reported:
    - **Range**: the highest value minus the lowest value of the data set.
    - **Interquartile range**: the range of the middle half of the data set.
    - **Standard deviation**: the average distance between each value in your data set and the mean.
    - **Variance**: the square of the standard deviation.

- Once again, the shape of the distribution and level of measurement should guide your choice of variability statistics.
    - The interquartile range is the best measure for skewed distributions, while standard deviation and variance provide the best information for normal distributions.

Step 3: Summarize your data with descriptive statistics

Example: Descriptive statistics (experiment)

- After collecting pretest and posttest data from 30 students across the city, you calculate descriptive statistics.
- Because you have normal distributed data on an interval scale, you tabulate the mean, standard deviation, variance and range.

	Pretest scores	**Posttest scores**
Mean	68.44	75.25
Standard deviation	9.43	9.88
Variance	88.96	97.96
Range	36.25	45.12
N	30	

- Using your table, you should check whether the units of the descriptive statistics are comparable for pretest and posttest scores.
  - For example, are the variance levels similar across the groups? Are there any extreme values?
  - If there are, you may need to identify and remove extreme outliers in your data set or transform your data before performing a statistical test.
- From this table, we can see that the mean score increased after the meditation exercise, and the variances of the two scores are comparable.
  - Next, we can perform a statistical test to find out if this improvement in test scores is statistically significant in the population.

Step 3: Summarize your data with descriptive statistics

Example: Descriptive statistics (correlational study)

- After collecting data from 653 students, you tabulate descriptive statistics for annual parental income and GPA.

	Parental income (USD)	**GPA**
Mean	62,100	3.12
Standard deviation	15,000	0.45
Variance	225,000,000	0.16
Range	8,000–378,000	2.64–4.00
N	653	

- It’s important to check whether you have a broad range of data points.
  - If you don’t, your data may be skewed towards some groups more than others (e.g., high academic achievers).

Step 4: Test hypotheses or make estimates with inferential statistics

- A number that describes a sample is called a **statistic**, while a number describing a population is called a **parameter**.
  - Using inferential statistics, you can make conclusions about population parameters based on sample statistics.

- Two main methods (simultaneously) to make inferences in statistics.
  - **Estimation**: calculating population parameters based on sample statistics.
  - **Hypothesis testing**: a formal process for testing research predictions about the population using samples.

Step 4: Test hypotheses or make estimates with inferential statistics

Estimation

- You can make two types of estimates of population parameters from sample statistics:
  - **A point estimate**: a value that represents your best guess of the exact parameter.
  - **An interval estimate**: a range of values that represent your best guess of where the parameter lies.

- If your aim is to infer and report population characteristics from sample data, it's best to use both point and interval estimates in your work.
  - You can consider a sample statistic a point estimate for the population parameter when you have a representative sample.
  - There's always error involved in estimation, so it's good to provide a confidence interval.

Step 4: Test hypotheses or make estimates with inferential statistics

Hypothesis testing

- Using data from a sample, you can *test hypotheses* about relationships between variables in the population.

- Statistical tests determine where your sample data would lie on an expected distribution of sample data if the null hypothesis were true. These tests give two main outputs:
  - A ***test statistic*** tells you how much your data differs from the null hypothesis of the test.
  - A ***p value*** tells you the likelihood of obtaining your results if the null hypothesis is actually true in the population.

Step 5: Interpret your results

Statistical significance

- In hypothesis testing, statistical significance is the main criterion for forming conclusions.
  - You compare your *p* value to a set significance level (usually 0.05) to decide whether your results are statistically significant or non-significant.
- Statistically significant results are considered unlikely to have arisen solely due to chance.

Example: Interpret your results (experiment)

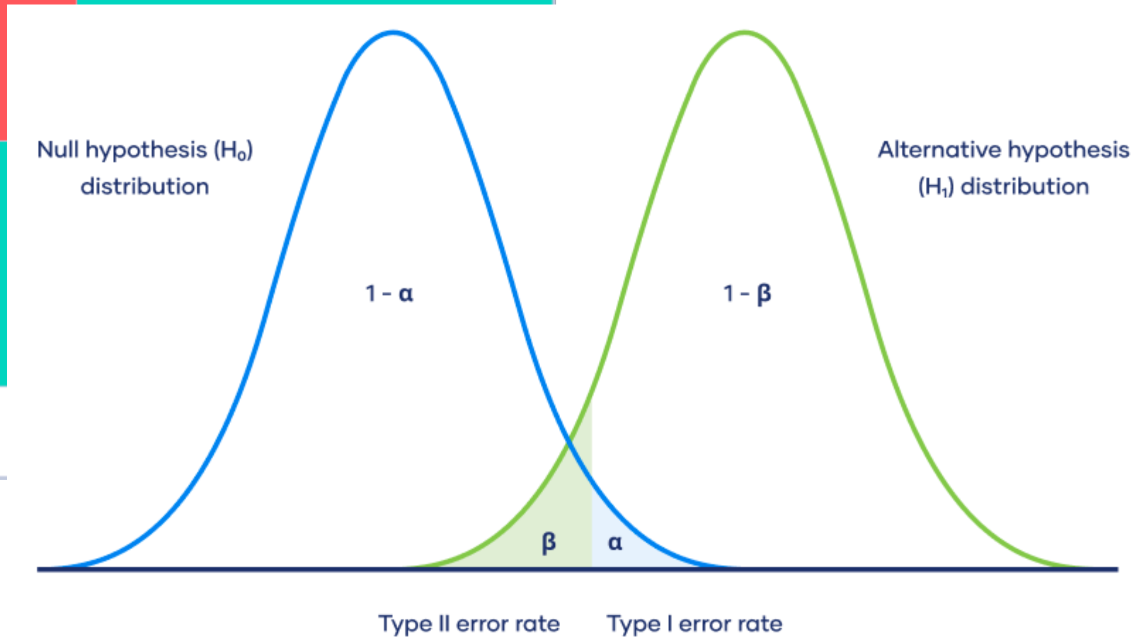
- You compare your *p* value of 0.0027 to your significance threshold of 0.05. Since your p value is lower, you decide to reject the null hypothesis, and you consider your results statistically significant.
- This means that you believe the meditation intervention, rather than random factors, directly caused the increase in test scores.

Step 5: Interpret your results

Decision errors

- Type I and Type II errors are mistakes made in research conclusions.

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision
Not rejected	Correct decision True negative Probability = 1 - α	



Thank you~

Wan Fang
Southern University of Science and Technology