Data Discovery
Lecture 8

Wan Fang

Southern University of Science and Technology

Agenda

- Inferential Statistics
- Statistical Analysis Step by Step

Inferential Statistics

Wan Fang

Southern University of Science and Technology

Descriptive vs. Inferential

https://www.scribbr.com/category/statistics/

- When you have collected data from a sample, you can use **inferential statistics** *to understand the larger population from which the sample is taken*.

- Inferential statistics have two main uses:
  - **Making estimates** about populations
    - (for example, the mean SAT score of all 11th graders in the US).
  - **Testing hypotheses** to draw conclusions about populations
    - (for example, the relationship between SAT scores and family income).

Descriptive versus Inferential Statistics

to *describe* vs. to *make inferences*

Descriptive statistics

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** concerns how spread out the values are.
- In descriptive statistics, there is **no uncertainty** – the statistics precisely describe the data that you collected.

Example: Descriptive statistics

- You collect data on the SAT scores of all 11th graders in a school for three years.
- You can use descriptive statistics to get a quick overview of the school's scores in those years. You can then directly compare the mean SAT score with the mean scores of other schools.

Inferential statistics

- Most of the time, you can only acquire data from samples, because it is too difficult or expensive to collect data from the whole population that you're interested in.
- Inferential statistics **use your sample to make reasonable guesses about the larger population.**
- **Use random and unbiased sampling methods**. .

Example: Inferential statistics

- You randomly select a sample of 11th graders in your state and collect data on their SAT scores and other characteristics.
- You can use inferential statistics to make estimates and test hypotheses about the whole population of 11th graders in the state based on your sample data.

Hypothesis Testing

- Hypothesis testing is ***a formal process of statistical analysis using inferential statistics***.
  - The goal is to compare populations or assess relationships between variables using samples.

- Statistical tests can be **parametric** or **non-parametric**.
  - Parametric tests are considered more statistically powerful because they are more likely to detect an effect if one exists.

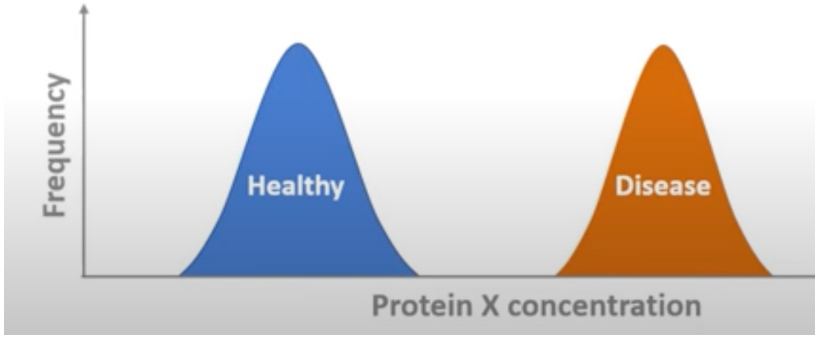
- **Parametric tests** make assumptions that include the following:
  - the population that the sample comes from follows a normal distribution of scores
  - the sample size is large enough to represent the population
  - the variances of each group being compared are similar

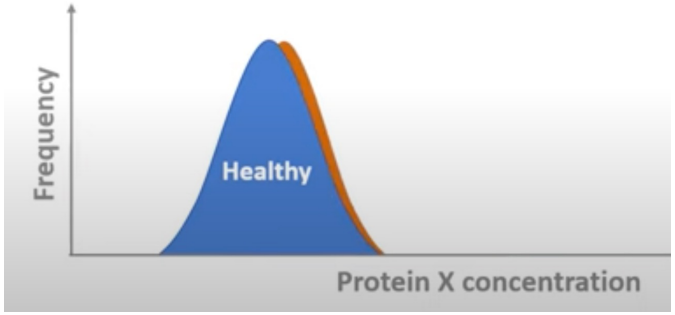
- When your data violates any of these assumptions, **non-parametric tests** are more suitable.
  - Non-parametric tests are called "distribution-free tests" because they don't assume anything about the distribution of the population data.

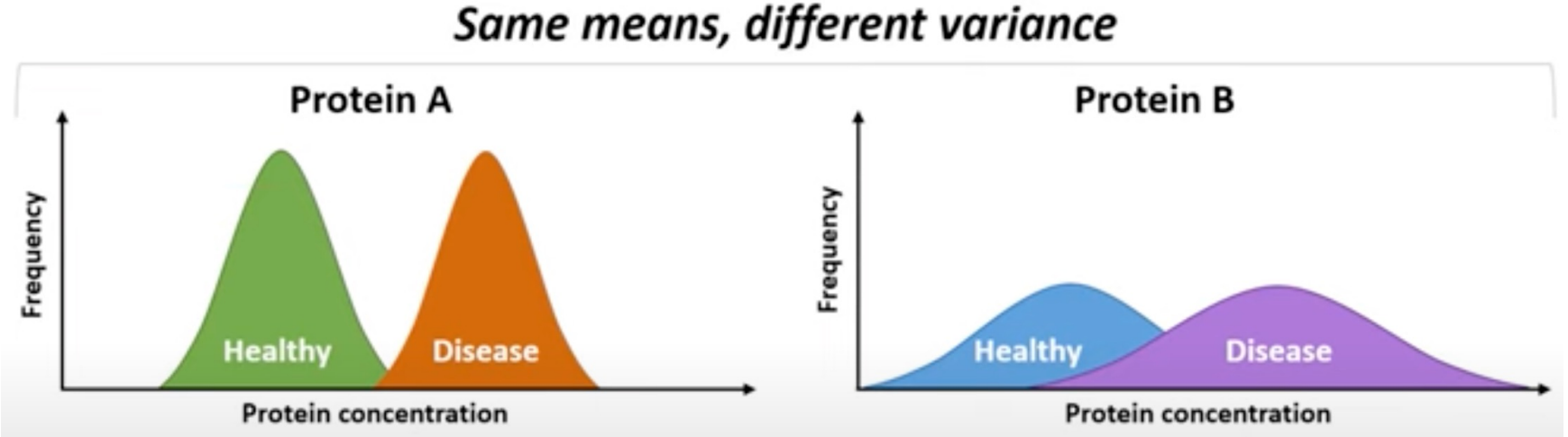
- Statistical tests come in three forms: ***tests of comparison***, ***correlation*** or ***regression***.

Comparison Tests

- Comparison tests assess whether there are differences in means, medians or rankings of scores of two or more groups.







Comparison Tests

- Comparison tests assess whether there are differences in means, medians or rankings of scores of two or more groups.
  - To decide which test suits your aim, consider whether your data meets the conditions necessary for parametric tests, the number of samples, and the levels of measurement of your variables.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Comparison test	Parametric?	What's being compared?	Samples
t test	Yes	Means	2 samples
ANOVA	Yes	Means	3+ samples
Mood's median	No	Medians	2+ samples
Wilcoxon signed-rank	No	Distributions	2 samples
Wilcoxon rank-sum (Mann-Whitney *U*)	No	Sums of rankings	2 samples
Kruskal-Wallis *H*	No	Mean rankings	3+ samples

Correlation Tests

- Correlation tests determine the extent to which two variables are associated.
  - Although **Pearson's *r*** is the most statistically powerful test, Spearman's r is appropriate for interval and ratio variables when the data doesn't follow a normal distribution.
  - The chi square test of independence is the only test that can be used with nominal variables.

Correlation test	Parametric?	Variables
Pearson's *r*	Yes	Interval/ratio variables
Spearman's *r*	No	Ordinal/interval/ratio variables
Chi square test of independence	No	Nominal/ordinal variables

- Null hypothesis (H_0): Variable 1 and variable 2 are **not related** in the population; The proportions of variable 1 are **the same** for different values of variable 2.
- Alternative hypothesis (H_a): Variable 1 and variable 2 are **related** in the population; The proportions of variable 1 are **not the same** for different values of variable 2.

Regression Tests

- Regression tests demonstrate whether changes in predictor variables cause changes in an outcome variable.
- Most of the commonly used regression tests are parametric.
  - If your data is not normally distributed, you can perform data transformations.

Regression test	Predictor	Outcome
Simple linear regression	1 interval/ratio variable	1 interval/ratio variable
Multiple linear regression	2+ interval/ratio variable(s)	1 interval/ratio variable
Logistic regression	1+ any variable(s)	1 binary variable
Nominal regression	1+ any variable(s)	1 nominal variable
Ordinal regression	1+ any variable(s)	1 ordinal variable

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.

Step 1: Write your hypotheses and plan your research design (*sample size*, and *sampling procedure*)

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

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

- **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.

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.

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)

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
Parental income	Quantitative (ratio)
GPA	Quantitative (interval)

Step 1: Write your hypotheses and plan your research design

Example: **Experimental research design**

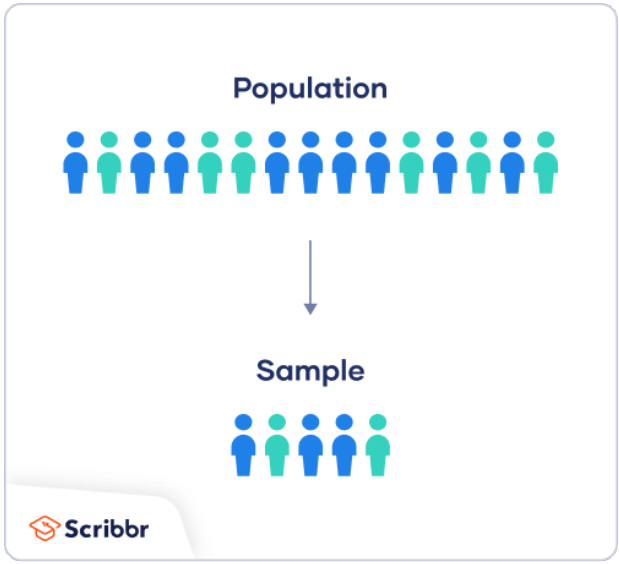
- You design a within-subjects experiment to study whether a 5-minute meditation exercise can improve math test scores. Your study takes repeated measures from one group of participants.
- First, you'll take baseline test scores from participants. Then, your participants will undergo a 5-minute meditation 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

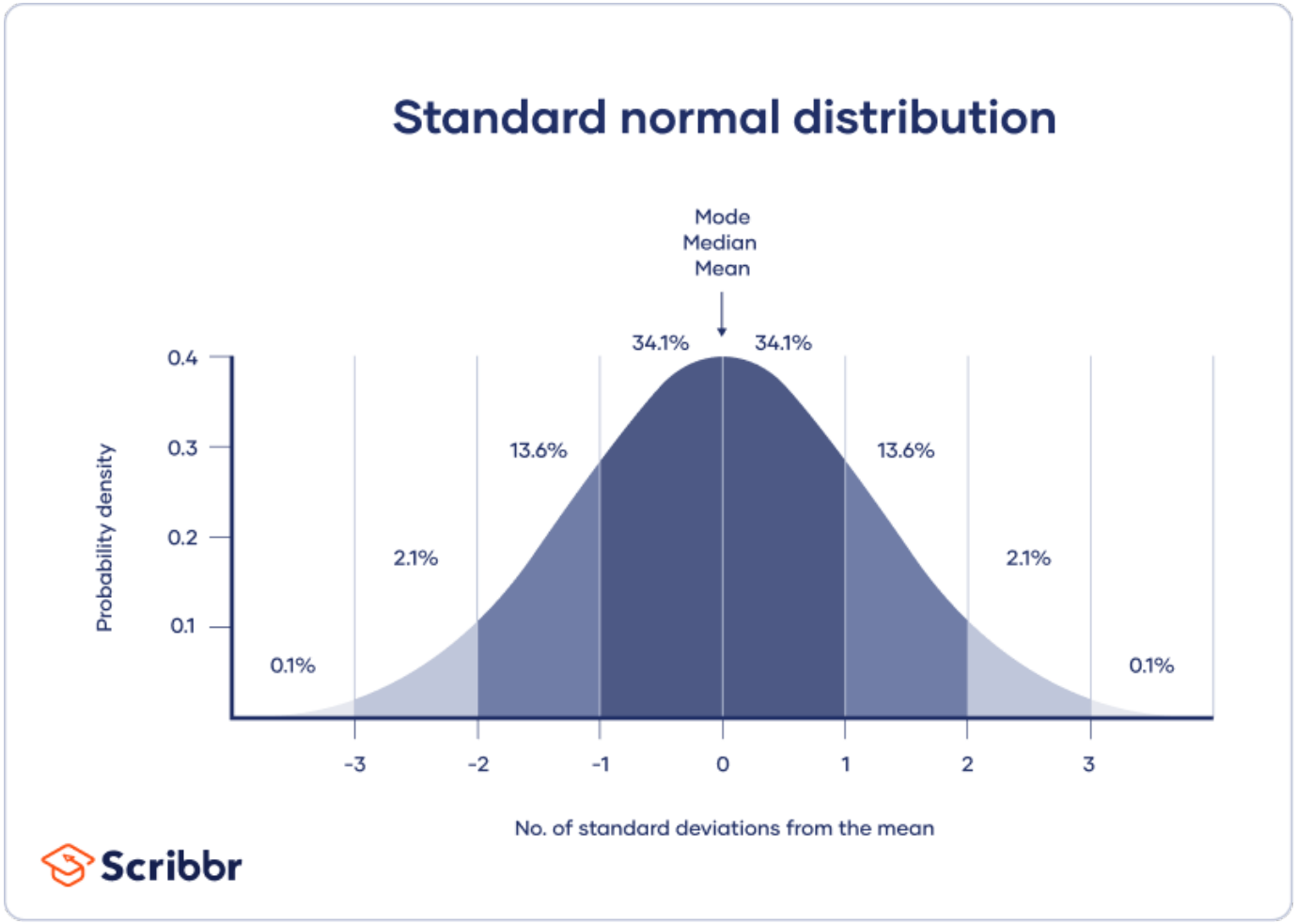
- There are various ways to inspect your data, including the following:
  - Organizing data from each variable in frequency distribution tables.
  - Displaying data from a key variable in a bar chart to view the distribution of responses.
  - Visualizing the relationship between two variables using a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

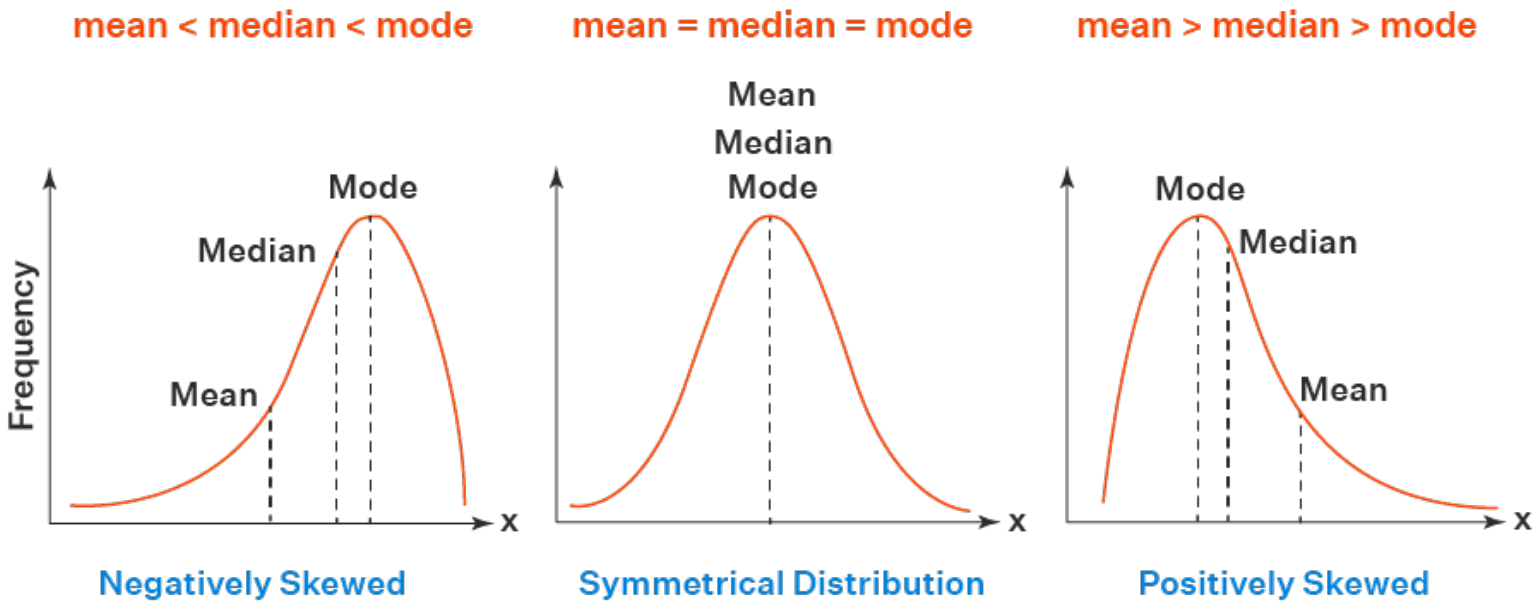


Extreme outliers can also produce misleading statistics, so you may need a systematic approach to dealing with these values.

Step 3: Summarize your data with descriptive statistics

Calculate measures of central tendency

- Measures of ***central tendency*** describe where most of the values in a dataset lie. Three main measures of central tendency are often reported:
  - **Mode**: the most popular response or value in the data set.
  - **Median**: the value in the exact middle of the data set when ordered from low to high.
  - **Mean**: the sum of all values divided by the number of values.



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.

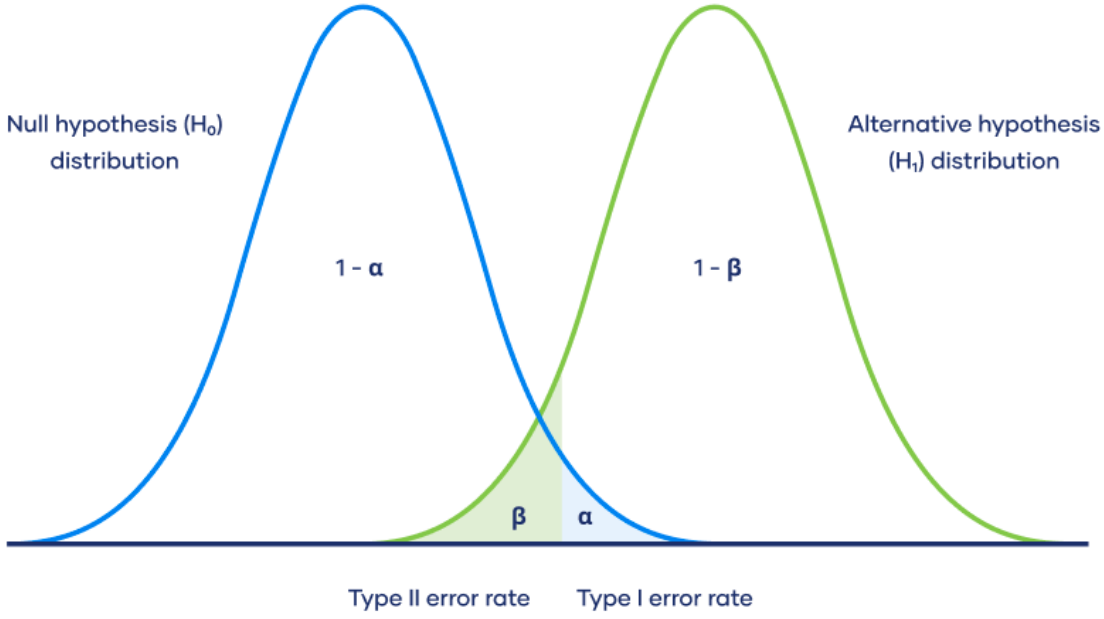
Step 5: Interpret your results

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	

Scribbr



https://ds363.ancorasir.com/

Thank you~

Wan Fang
Southern University of Science and Technology