Basic concepts

Different formats of the data

- Numerical data

  (e.g.) height, weight, age, blood pressure

- Categorical data
  - Nominal categorical data

    (e.g.) race, gender, disease status
  - Ordinal categorical data

    (e.g.) survey answers, age group

Different types of data

- Univariate data
- Longitudinal data
- Functional data (daily temperature)
- Spatial data (Covid case map)
- Imaging data (e.g., neuro imaging)
- Spatio-temporal data (spatial data over multiple years, e.g., HIV spread over years)
- Text data (Amazon reviews)

Statistical inference

1. Specify your **claim**

   (i.e., state null and alternative hypotheses)
2. Choose **proper test** and calculate test statistic

3. Check **P-value**

   Any types of inferential test will provide p-value

   We make a conclusion based on this.
4. Make a **conclusion**

Statistical inference

- All topics we will cover in the class are examples of inferential test.
  - One-sample t-test: test mean value
  - Two-sample t-test: comparing means of two groups
  - ANOVA: comparing means of more than 2 groups
  - Regression: test linear relationship
  - Logistic regression: test relationship between binary response variable and predictor

Statistical inference examples

- ✓ One-sample t-test

  (e.g., Covid vaccine is effective)

- ✓ Two-sample t-test

  (e.g., test if Moderna and Pfizer vaccines have the same effectiveness)

- ✓ ANOVA test (simply speaking, comparing group means among more than two groups)

  (e.g., test among Moderna, Pfizer, and J&J)

Statistical inference examples

- Covid vaccine Phase 1 (Moderna)
  - ✓ Dose-escalation, open-label trial including
  - ✓ Compare antibody responses for
    does of 25 mg, 100 mg, and 250 mg
  - ✓ Link for the article:

  https://www.nejm.org/doi/full/10.1056/NEJMoa2022483?fbclid=IwAR1rhWsyuRN2bet0V8UHCJ1MqKQXo1h0NUzm3kQCALlmwr5jEqLF_AiBRKU

- Covid vaccine Phase 3
  - ✓ Record symptoms and temperature readings
  - ✓ Group of 100mg vaccine vs. Group of placebo
  - ✓ link for details:

  https://www.nih.gov/news-events/news-releases/phase-3-clinical-trial-investigational-vaccine-covid-19-begins

What is p-value?

- P-value means the **probability of finding observed result or extreme when null hypothesis is true.**
- If P-value is .01 then it means that this result occurs only 1 time among 100 trials when null is true. We regard it as too small and believe it rarely happens. Probably observed result is not from null hypothesis.
- The degree of statistical evidence

Make a conclusion

- The smaller the p-value, the **stronger** the **evidence** that you should **reject the null hypothesis**
  - If p-value is smaller than significance level (α), we reject the null hypothesis
  - If p-value is larger than α, we do not have enough evidence to reject the null hypothesis and we fail to reject the null hypothesis

- **0.05 is not a magic number!!**

Note about statistical inference

- **What is difference between statistically significant difference and difference?**

- **P-value** and **Effect size**
  - Small p-value implies large effect size?
  - (e.g.) with p-value 0.0001, galaxy phone battery has significantly longer life span than i-phone battery has.

Statistical inference

- **Parametric test**
  - E.g., t-test, ANOVA, regression…
    : Normality assumption is required
  - Established theory
  - Simple calculation of test-statistic

- **Non-parametric test**
  - No assumption is required. Flexible
  - More computation
  - E.g., bootstrap test or permutation test

Data mining/ Stat learning

- Main goal is prediction
  - Not an inferential conclusion
  - E.g., classification, Principal Component Analysis..
  - E.g., application in face recognition

- What you are going to learn in Algorithm 2

Philosophies in Statistics

- Frequentist
  - Parameter is fixed, but unknown
  - We estimate the “true” unknown parameter
  - Main stream

- Bayesian
  - Parameter is not fixed, but follows the distribution
  - There is no “true” parameter and we estimate the parameter given our data set
  - The counter part of (almost) frequentist method can be found under Bayesian perspective

https://medium.com/analytics-vidhya/a-short-story-on-bayesian-vs-frequentist-statistics-27f55ae56253