Course Overview

Jeongyoun Ahn

KAIST

Statistics, the Grammar of Science

Life and most aspects of it is inherently variable.

Statistics allows us to make informed decisions in the face of uncertainty.

Statistics, as a field, lays ground rules for

- how data should be collected.
- how decision can be made based on the data.

Population and Samples

Population is the entire collection of objects on which an investimation is focused.

A **census** measures every member of the population.

A **sample** is any subset of the population. Our decisions are only as good as our sample.

A **variable** is a characteristic of interest for the objects in a population.

Descriptive vs. Inferential Statistics

Descriptive statistics uses graphical or numerical methods to describe the sample.

Inferential statistics draws inference from the sample about the population.

- **Frequentist** approach is the focus of this course. We interpret probability as the long-run chance of an outcome's occuring in repeated trials.

- **Bayesian** approach is a popular alternative to the classical inference.

Statistical Procedure

1. Set the goal - What do we want to show?
2. Collect data (experimental design) - What kind and how much data need to be collected?
3. Describe the data - summarize and describe the prominant features of data. (e.g., histogram, scatter plot, mean, variance, etc.)
4. Analyze the data (inferential statistics)
   - estimation, prediction, test, decision
   - generalize from a sample to a population
5. Conclusion based on the goal - how to assess the strength of the conclustion?

Parameter and Statistic

Paramter: a numerical summary of a population

- Population mean: average of a numerical measure
- Population proportion: fraction having a particular characteristic

Statistic: a numerical summary of a sample

- Sample mean
- Sample proportion

Inferences depend on the sample being representative of the population.

Classification of a variable I

- Qualitative
  - measurement is a set of unorderd categories.
- Quantitative
  - values of the variable differ in magnitude
- Ordinal
  - values are categories but with natural ordering

Classification of a variable II

- Discrete
  - takes finite (countable) number of values

- Continuous
  - can take any value within an interval, infinite possibilities

* Remark
  - all categorial variables are discrete
  - quantitative variables could be discrete or continous
  - sometimes it depends on a situation

Descriptive study of data

Want to estimate population distribution using sample distribution.

- tabular/grapical representation
  - quantitative: frequency table, dot diagram, histogram, line diagram, stem-and-leaf display

  - categorical: contingency table, pie chart, bar chart

Descriptive study of data

- numerical representation
  - center: mean, meadian, percentiles, trimmed mean, Winsorized mean

  - variaion: variance, standard deviation, range, interquartile range

  - empirical rule:

Descriptive study of data

- Boxplot
  - Graphical display of five-number summary
  - location, variation, skewness, outliers