Course Overview

Jeongyoun Ahn

KAIST

Statistics, the Grammar of Science (Karl Pearson)

Life and most aspects of it is inherently variable. (random) "in the sense of repeated trials"

Ch 2 ~ Ch 6 → Uncertainty from the nature. (cannot be reduced)

Ch 7 ~ → Uncertainty from not enough information (= data)

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. — impossible

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.

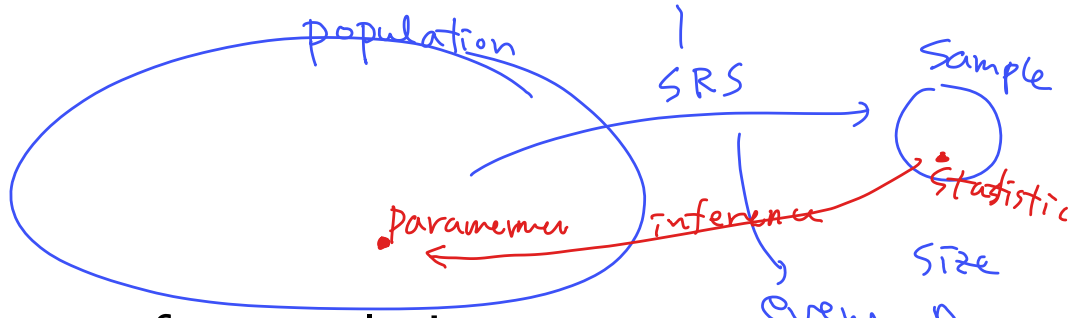
Statistical Procedure

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? What pop. do your data represent?
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?

Statistics — a field of study
plural form of statistic

simple random sampling

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.

every n size subset of the pop. has the same chance to be a sample.

ideally SRS

Classification of a variable I

- Qualitative
  - measurement is a set of unorderd categories.

- Quantitative — value matters
  - 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

e.g. education levels, blood types, no. books read

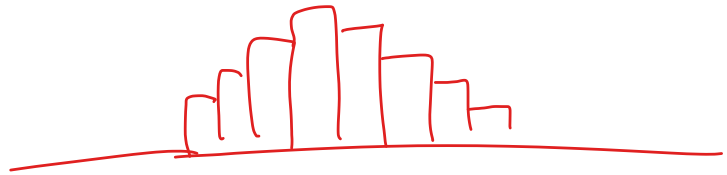
- Continuous
  - can take any value within an interval, infinite possibilities

e.g. weight, ...

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

age

Descriptive study of data

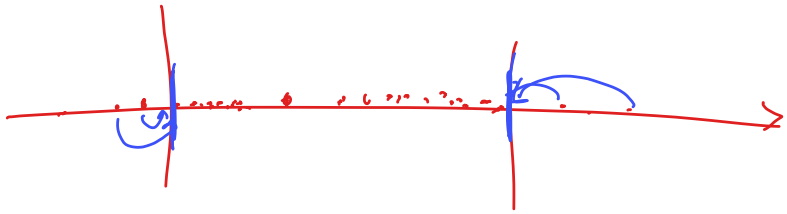


Want to estimate population distribution using sample distribution.

- 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, median, percentiles, trimmed mean, Winsorized mean
  - variation: variance, standard deviation, range, interquartile range
  - empirical rule:

mean = 50
stdev = 10

If data are bell-shape,

mean ± 1.stdev contains 68%

2. " 95%

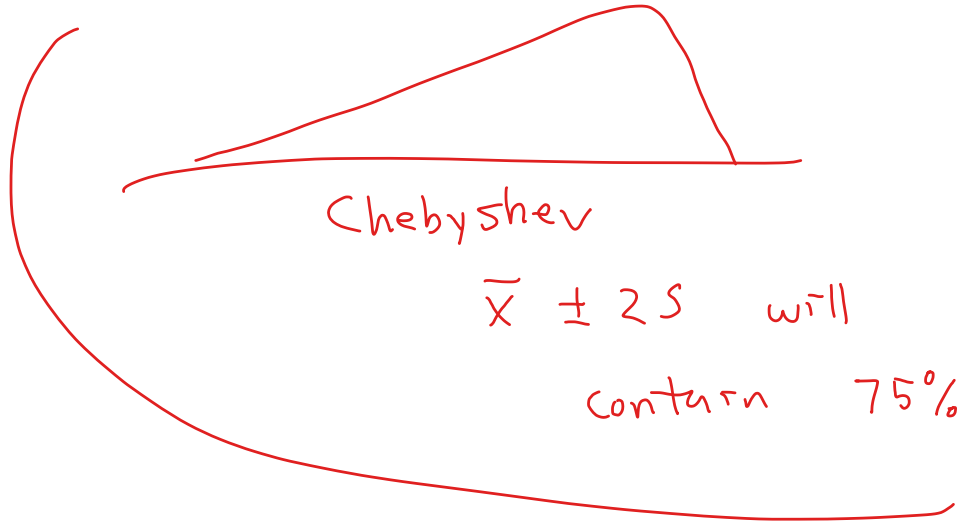
3 " " 99.7%

Quartiles

- Q_1 – 25th percentile
- Q_2 = median
- Q_3 = 75th percentile

IQR = $Q_3 - Q_1$

Descriptive study of data



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