임상연구 설계와 분석을 위한 통계 방법

Boncho Ku, Ph.D., Senior researcher

16th November, 2017

KM Fundamental Research Division, Korea Institute of Oriental Medicine

Chapter I: Overview of Statistics

Famous quotes about statistics

There are three types of lies: lies, damn lies, and ***STATISTICS*** (Benjamin Disraeli)

Fact are stubborn things, but ***STATISTICS*** *are pliable* (Mark Twain)

and so on ...

Huge number of quotes about statistics commented it in sarcastic tone

→ mostly hard to refute

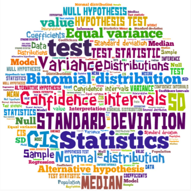
Really?

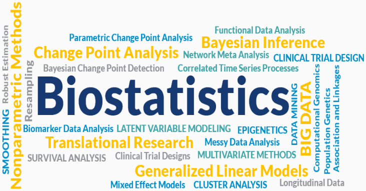
However ...

Statistics itself always provides useful information and allows us to maintain objective perspective based on DATA

What is Statistics?

So then, what is statistics??





† Each wordcloud was cited from Trident University International and Augusta University, respectively.

Statistics

Concerning with **collection**, **organization**, **summarization** and **analysis** of **DATA**

Main Pillars of Statistics

The most important things in statistics

1. Data (sample)
   - Investigation, experiment, and survey
   - Gathering numbers (for quantitative analysis)
2. Description or Summarization
   - Table, chart, and so on
   - Based on summarized statistics (e.g. mean, standard deviation, median, . . .)
3. Inference
   - Numerous statistical tests and models based on probability theory
   - e.g. two-sample t-test, ANOVA, ANCOVA, regression, and so on

Why should we collect data (sample)?

Measure everything from POPULATION

- Benefits
  - You will get exactly correct answer
  - No need to meet an awkward statistician LIKE ME
- If you had a plenty of
  - Money (typing “SHOW ME THE MONEY” may help your budget)
  - Time (TOO SHORT TO COLLECT data of entire population)

Inferential approach based on SAMPLE

- If we have a proper sample that represents the whole population, you can get NEARLY the correct answer
- Estimation and hypothesis testing

Parameter vs. Estimates

Parameter

Parameters exist somewhere in the universe → the true value representing the target population

From the view of *frequentist*,

- Parameters are fixed → never changing
- Parameters exists but we never know the true value of them
- But we can “guess” them from sample

Parameter vs. Estimates

Estimates

- Estimating parameters based on the given samples (data)
- Estimates have a variation in accordance with different samples or data

The data is an aspect of the real world we have captured

- How good is our estimation?
  - Estimation inevitably involves **ERROR**
  - Error measures: standard error (SE) → reliability of an estimate

$$\mathrm{SE} = \frac{\sigma}{\sqrt{N}}$$

Measurement is ubiquitous → then error is also ubiquitous.

Type of variables

Data consist of a set of independent sample and measured variables

Table 1: Types of variable based on their scales

Scale	Example	Operation
Qualitative (질적변수)		
Nominal (명목)	sex, marital status, blood type, race, eye colour, religion, ...	counting
Ordinal (순서)	grade, education level, preference, severity, ...	counting, ranking
Quantitative (양적변수)		
Interval (구간)	temperature, IQ, SAT score, ...	counting, ranking, $+$, $-$
Ratio (비율)	distance, length, height, weight, BMI, blood pressure, ...	counting, ranking, $+$, $-$, $\times$, $\div$

Can we separate types of variable clearly?

Continuous variable is limited by the precision of the measurement

Example

- Height: measured to the nearest centimeter → continuous variable?
- Age: measured to the year but theoretically, measured to any level of precision (e.g. month, day, and time)

In practice, all variables are discrete but some variables can be treated as continuous when its distribution can be well approximated by a continuous distribution.

How to express your data?

Data themselves are just a bunch of numbers → how to extract meaningful information from data?

Descriptive statistics

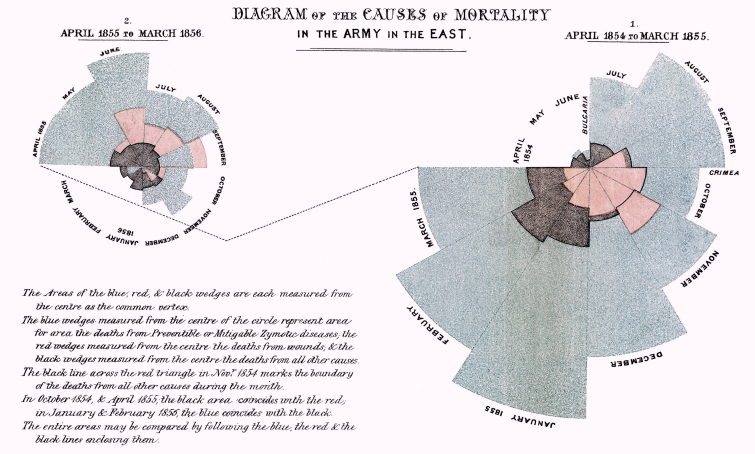
- Summary statistics: all information of data are represented by a certain type of numbers
  - Example: mean, median, proportion, standard deviation, interquartile range, percentile, . . . → developing "*statphobia*"

Table 2: Descriptive statistics of "mpg" dataset

Variable	n	Min	q_1	$\widetilde{x}$	$\bar{x}$	q_3	Max	s	IQR
displ	234	1.6	2.4	3.3	3.5	4.6	7	1.3	2.2
year	234	1999.0	1999.0	2003.5	2003.5	2008.0	2008	4.5	9.0
cyl	234	4.0	4.0	6.0	5.9	8.0	8	1.6	4.0
cty	234	9.0	14.0	17.0	16.9	19.0	35	4.3	5.0
hwy	234	12.0	18.0	24.0	23.4	27.0	44	6.0	9.0

How to express your data?

Example of polar area diagram by Florence Nightingale (1820 ∼ 1910)



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

How to express your data?

Example of data visualization

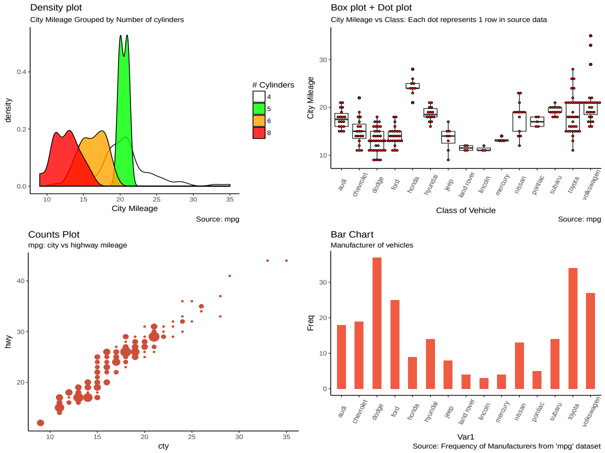


Figure 1: Data visualization examples for “mpg” dataset. All plots are available at http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

How to express your data?

Data visualization

Sometimes, a graph provides us more useful information than complex tables

Various types of statistical graphs

- Histogram, boxplot, Q-Q plot, scatterplot, . . .
- Do NOT rely only on NUMBERS, Do draw a PLOT!!

Is description of data fully enough?

Again, data are the small aspect of the real world.

Statistical inference provides us more reasonable interpretation regarding to the uncertainty of data.

Two main category of statistical inference

- **Estimation**
- **Hypothesis testing**

Meaning of 95% confidence interval

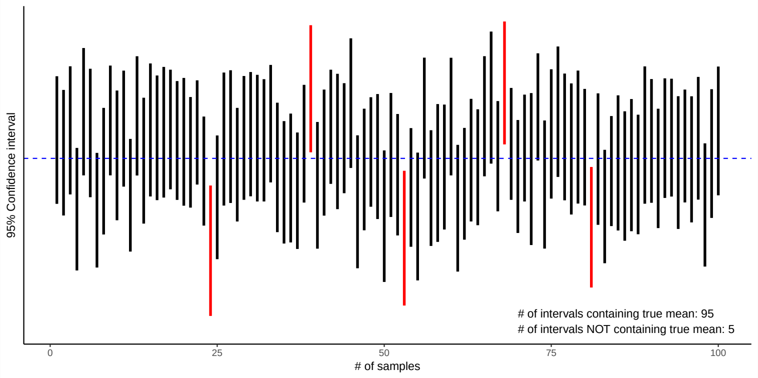


Figure 2: 95% confidence intervals for 100 independent drawn samples with $n = 100$.

Clinical Research

Research or trial?

Research

자료의 수집과 분석 목적이 학술적 목적에 국한된 모든 종류의 연구 및 실험

Trial

자료의 수집과 분석 목적이 이윤추구 또는 허가에 목적이 있는 임상시험

Observational Study

Cross-sectional study (단면적 관찰연구)

1. prevalence study
2. Diagostic test
3. Ecological study
4. Validity, Reliability, and agreement study

Longitudinal study (종단적 관찰연구)

1. Prospective study
2. Retrospective study

Experimental Study

Randomized controlled trial

Pilot study

Exploratory study

Confirmative study

Type of outcome variables

Primary outcomes

Secondary outcomes

Surrogate variables

Global assessment variable

Sample size calculation

Overview

Two approaches

1. Based on the marginal error rate → population based observational study
2. Based on the effectiveness between concerning groups → experimental study

Both approaches are based on previous studies

Is your study entirely new?

Observational study

Observational study: prevalence study

Observational study: prevalence study

Parallel design

2×2 cross-over design

Factorial design

Multiple comparison

What makes data significant?

1. Data themselves contain unexpected errors
2. Bias
3. Just conincidence
4. Our hypothesis is working

Torturing data

Statistical Analysis

Overview

Independent two sample t-test

1. Too easy, but very useful methodology for the comparison of sample means between two groups

Analysis of Variance (ANOVA)

Analysis of Covariance (ANCOVA)

Simple or multiple regression

Repeated Measures ANOVA

Linear mixed effects model

Reliability analysis

Cohen's κ

Cronbach's α

Intra Class Correlation (ICC)