Statistics : extracting meaning from "incomprehensible/raw/undigested" data

Population 모집단 : the complete collection / entire set
Sample 표본 : a subset selected from the population (randomly → no systemmatic difference between samples)

* 표본or모집단의 특성을 나타내는 값 (특징 정보 얻을 수 있음)
Population Parameter 모수/모수치 : 모집단 전체로부터 얻은 통계 수치 (e.g. average) - 구하기 사실상 거의 불가능
Sample statistic (표본)통계치/통계량 : 표본들로부터 얻은 통계치 - 경험적으로 얻어짐

 # Descriptive Statistics 기술 통계
to summarize, organize, and simplify data (표본, 모집단에 다 사용 가능)
- Frequency Distributions 도수 분포
- Central Tendency Measures 중심경향치 (e.g. single number average (예:85))
- Variability Measures 산포도

 # Inferential Statistics 추론 통계
표본 연구 및 모집단에 일반화하기 위함 (표본에만 사용 가능)
- Estimation 견적
- Hypothesis Testing 가설 검정 (e.g. t-test, ANOUA, regression)

* Basics of 추론 통계 : Standard Scores, Probability, Normal Distributions, Sampling Distributions

Variable 변수/변인 : different values among the individuals of the sample/population (예: Motivation, Performance)
Constant 상수 : same, fixed value (예: 해당 학생이 PSY2002를 수강했는지 아닌지)

*** Scale of measurement 측정 정도**
measuring = assigning numbers to individuals
예시1 A(170cm)'s variable value : 170 / B(160cm)'s variable value : 160
예시2 A(taller than avg.)'s variable value : 1 / B(smaller than avg.)'s variable value : 2
→ 측정 수준(Level of measurement)에 따라 참가자들의 value가 달라짐 → 통계의 성격도 달라짐

when "categorizing observations",
① Nominal Scale 명목 척도
- labeling
- observations 간에 아무 양적 차이가 없음

예 : Sex(male=1 female=2), Ethnicity(group들에 1,2,3, … 임의의 숫자 부여)
② Ordinal Scale 서열 척도
- rank observations in terms of size or magnitude
- numbering from high to low / differences between the numbers is not interpretable

예 : A3 B1 C5 D2 E4 → 실제 preference 간격이 다를 수 있음
③ Interval Scale 등간 척도
- rank observations and equal differences between numbers reflect equal differences in magnitude
- Zero point != 0
- Does not permit statement about ratios of measurements

예 : Temperature, IQ score
④ Ratio Scale 비율 척도
- has an absolute zero point (=0)
- permits statement about ratios

예 : Height, Weight, Time to complete a task (measured in minutes), Annual rainfall (measured in millimeters)

Quantitative(양적) Data ③④ vs Qualitative(질적) Data ①②
(-> do not have any quantitative meaning)

<Descriptive Statistics 기술 통계>

1. Frequency distributions 빈도 분포 - table(표) / graph(그래프)

* grouped frequency distribution table
  intervals(구간) 존재 : range(범위) = Max-Min / interval width(구간 너비)
  -> 구간의 lower bound (bottom score) & upper bound (top score)은 minimum measurement unit에 의해 정해짐
* bar graph
  nominal/ordinal : bars끼리 띄기 (touching) -> Regular frequency table / Bar graph
  interval/ratio : bars끼리 붙이기 (no space)
  -> Regular(적을때) or Grouped(많을때) frequency table / Histogram, Frequency polygon

f = frequency

p = proportion(relative frequency) = f/n (n=$\sum f$=total frequency)

%f = 100p ($\sum \%f$=always 100%)

(if necessary,) **Cumulative f 누적 빈도** (가장 윗 칸의 cum f = n)

Nominal data는 order info가 아님

→ nominal : f, p, %

→ ordinal : f, p, % + cumf

2. Central tendency measures 중심경향치/중앙경향치 -> a single value (center)

"effectively summaraize, but might not capture all the important features" 예) 분포

- **Mean 평균** = arithmetic avg. 산술평균

N = population size / u = population mean

n = sample size / $\overline{X}$ = sample mean

편차 = Xi - $\overline{X}$

- **Median 중앙값**
- **Mode 최빈치**

3. Variability measures 변동성/변산성

"if measured in the same unit" (단위 다르면 makes no sense)

1) 항상 양수

2) nominal/ordinal(x) interval/ratio(o)

- **Range 범위** : Max - Min -> 두 극단치에만 의존 -> crude(not very sophisticated)
- **Variance 분산** : avg. squared distance from the mean

Population variance = σ^2(sigma-squared) = SS/N (sum of squared diviations / N)

Sample variance = s^2 = SS/N (sum of quared diviations / n)

sum of 편차 = 0

sum of 분산 = 0

avg. of 분산 = 0

- **Standard Deviation 표준편차**

Population standard deviation = σ

Sample standard deviation = s

* one formula for population variable
* two formulas for sample variable