

# **S\_01 Descriptive Statistics**

## **Statistical Analysis**

통계분석의 첫 강의로 Descriptive Statistics, 즉, 기술통계에 대해 강의하겠습니다.

## References

- Text and figures from the DATAtab site (<https://datatab.net/>) and the book "Statistics made easy" published by DATAtab.



- Statistics Page from Scribbr site (<https://www.scribbr.com/category/statistics/>)



- LLM: ChatGPT, Claude 3

이 강의는 Datatab (데이터탭) 과 Scribbr (스크리버)를 주로 참고하여 제작 되었습니다.  
강의에서 다루지 않는 디테일들이 필요한 경우 이 사이트들을 참고하실 수 있습니다.  
또한 Claude 3나 ChatGPT 같은 AI LLM 서비스를 활용하는 것도 많은 도움이 됩니다.

## Population and Sample

- Population (모집단)
  - 통계의 대상이 되는 모든 개체를 포함하는 집단
  - ex) 대한민국 사람 전체, 서울시민전체, 대전소재 21세 여성 전체, ...
  - Population에 속한 전체 개체의 data를 얻기는 사실상 불가
- Sample (표본)
  - Population 으로부터 sampling된 Sample 을 대상으로 한 통계가 흔히 사용됨
  - Sample은 sampling된 개체들의 집단을 의미하며, "Sample Set" 이라 부르지 않음

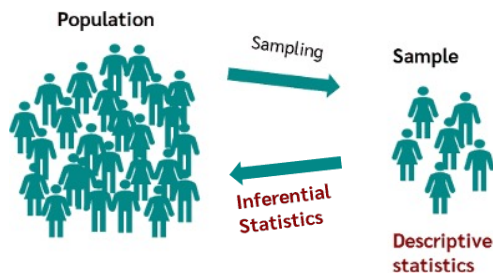


3

1. Population, 즉, 모집단은 통계의 대상이 되는 모든 개체를 포함하는 집단을 말합니다.
2. 예를 들면, 대한민국 사람 전체, 서울 시민 전체, 대전 거주 21세 여성 전체 등이 될 수 있을 것입니다.
3. 그러나 population에 속한 전체 개체의 data를 얻기는 사실상 불가능 합니다. 한가지 예로, 대한민국 고3 학생의 평균 수면 시간을 조사하려면, 전수조사, 즉 대한민국 고3 학생 모두의 수면시간을 조사하여 평균을 내야 할 것이지만, 그것은 사실상 불가능합니다.
4. 따라서 우리는 흔히 Sample, 즉, 모집단으로부터 샘플링된 표본을 사용합니다.
5. 통계학에서 Sample은 sampling된 개체들의 집단을 의미하며, 결코 개체 하나를 지칭하는 말이 아닙니다. 따라서 우리는 Sample을 Sample Set 이라 부르지 않는

다는 것을 기억해 주기 바랍니다.

## Descriptive vs Inferential Statistics



- Descriptive statistics (기술 통계): Sample의 통계를 describe 하기 위해 사용
  - ex) 전국민 평균 수면 시간
- Inferential statistics (추론 통계): Sample의 통계로부터 Population의 통계를 추론 (inference) 해 내기 위해 사용

4

1. Descriptive statistics, 즉, 기술통계는 Sample의 통계 정보를 기술하기 위해 사용하는 기법들을 말합니다. 예를 들어 Population인 전국민의 수면시간을 조사할 수 없으니 Sample인 수천명에 대해서만 수면시간을 조사했다고 가정합니다. 모아진 데이터로부터 수면시간의 평균, 표준편차 등을 구하고, histogram이나 다른 visualization 방법으로 데이터를 요약 표현하는 등, Sample에 대한 통계를 describe할 수 있을 것입니다. 이런 기술 단계로 끝나는 것을 기술 통계라 하는 것입니다.
2. Descriptive statistics와 대비하여 Inferential statistics, 즉, 추론통계는 Sample의 통계로부터 Population의 통계를 inference, 즉, 추론해 내는 과정을 말합니다. 기술 통계는 데이터를 정리하여 보여주는 수준이므로 비교

적 어렵지 않게 이해할 수 있으나, 추론통계는 사이즈가 작은 Sample의 통계로부터 훨씬 큰 Population의 통계를 추론해내야 하므로 고려해야 할 조건들이나 계산 과정이 좀 더 복잡하다고 볼 수 있겠습니다.

## Types of Variables (by characteristics of value)

- Categorical Variables
  - Nominal: value들을 순서 없이 구분만 가능
  - Ordinal: value들의 순서 있음
- Quantitative Variables (= Metric Variables)
  - Continuous: value의 domain이 real number
  - Discrete: value의 domain이 integer

5

1. 본격적으로 Descriptive Statistics에 대해 알아보기 전에 먼저 그 value의 특징에 따라 variable들의 type을 구분해 볼 필요가 있습니다. Variable들 중 Categorical variable의 value들은 어떤 양을 표현할 수 없습니다.
2. Categorical variable의 하나인 Nominal variable은 value간의 순서를 정할 수 없으며, 오직 value들 간을 구분하는 것만이 가능합니다.
3. 또 다른 categorical variable로는 Ordinal variable이 있는데, Ordinal variable은 그 value들 간의 순서를 정할 수는 있습니다. 하지만 value 간의 차이가 어떤 수치로 명확히 표현될 수 있는 것은 아닙니다.
4. Quantitative variable은 metric variable이라고도 하는데, value를 숫자로 나타낼 수 있는 양 일 경우를 말합니다.

5. 그 중에서도 continuous variable은 value의 domain이 real number이고
6. discrete variable은 domain이 integer 입니다.



## Nominal Variables

- Operations: **equal, unequal**
- No ranking and order
- ex) Binary (dichotomous) – value가 두 개 뿐인 경우

### Examples:

Gender	Marital status	Preferred newspaper:
1 = male	single	The Washington Post
2 = female	married	The New York Times
	divorced	USA Today
	widowed	...

6

1. 이제 각각의 variable type들에 대해 좀 더 자세히 알아보겠습니다. 먼저 Nominal variable에 적용할 수 있는 operation은 equal과 unequal 뿐입니다. 즉, nominal variable들은 그 value가 같은지 다른지만 확인할 수 있을 뿐입니다.
2. variable간에 어떤 ranking이나 order가 존재하지 않습니다.
3. 특별히 두 가지 값 중 하나를 가질 수 있는 Binary variable은 Nominal variable의 special case로 볼 수 있는데, 예를 들면 "Yes/No", "Presence/Absence 의 출결상태" 등이 binary variable 이 될 수 있습니다.
4. 이 표의 예에서 Gender는 male 또는 female 중의 하나의 value를 가지는 binary nominal variable이며

5. Marital status는 single, married, divorced, widowed의 value를 가질 수 있고
6. 선호하는 신문의 이름도 nominal variable 이라 할 수 있습니다.

## Ordinal Variables

- Operations: equal, unequal, **greater**, **smaller**
- Ranking (hierarchy) exists

### Examples:

#### Frequency of television:

- 1 = daily
- 2 = several times a week
- 3 = less frequently
- 4 = never

#### The government is doing a good job:

- 1 = agree with
- 2 = undecided
- 3 = disagree with

- But no exact distance between different values

7

1. Ordinal variable들은 Nominal variable에 적용되는 equal과 unequal에 더해 greater와 smaller의 두 operation들을 추가로 더 사용할 수 있습니다.
2. 즉, variable의 value들 간에 ranking 또는 hierarchy를 정할 수 있는 것이지요.
3. Example 들로는 "television을 보는 빈도" 에는 매일, 한 주에 몇 차례, 거의 안봄, 전혀 안봄의 value들이 있을 수 있는데, 이 value들은 빈도가 잦은 순서대로 나열할 수 있습니다.
4. 또 "정부가 일을 잘하고 있다" 에 동의하는 정도는 동의한다, 잘 모르겠다, 동의하지 않는다 의 value들을 가질 수 있으며, 이 들은 동의하는 정도 측면에서 순서를 정할 수 있는 것입니다.
5. 비록 순서를 정할 수는 있지만 ordinal variable의 서로

다른 value들 간에 어떤 수치적인 차이를 명확하게 계산할 수 있는 것은 아닙니다.

## Quantitative (Metric) Variables

- Equal, unequal, greater, smaller, **difference, sum**
- 즉, value들 간의 수치적인 차이를 정확히 계산 가능

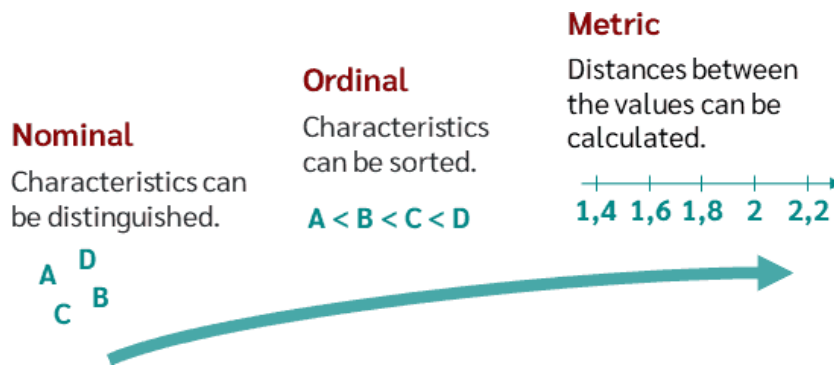
### Examples:

Income	Weight	Age	Electricity consumption
1820 \$	81 kg	18 years	520 kWh
3200 \$	70 kg	27 years	470 kWh
800 \$	68 kg	64 years	340 kWh
...	...	...	...

8

1. Quantitative (또는 Metric) variable들은 Ordinal variable들의 operation들에 추가적으로 difference, sum 까지 사용할 수 있는 것들입니다.
2. 차와 합을 계산할 수 있다는 것은, value들 간의 수치적인 차이를 정확하게 계산할 수 있다는 뜻입니다.
3. 예를 들어 수입, 체중, 나이, 전기사용량 등이 quantitative variable의 예 입니다.

## Level of Measurement



9

이러한 variable들의 종류를 Level of Measurement, 즉, 측정의 수준 이라는 측면에서 본다면, Nominal, Ordinal, Metric의 순서로 더 level이 높아진다고 볼 수 있습니다.

Nominal variable의 value들은 서로 구분될 수 있습니다. Ordinal은 variable의 value들은 순서대로 sorting될 수 있다.

Metric variable은 value들 간의 거리가 계산될 수 있습니다.

이러한 측면에서 측정의 수준은 nominal, ordinal, metric의 순서로

점점 더 높아지고 있으며,

측정의 수준이 더 올라갈 수록,

비교나 수치계산 operation들을 더 많이 적용할 수 있게

됩니다.

## Metric Variable – Ratio Scale

- Ratio Scale (비율척도)
- Data value들 간 가치의 비율이 의미가 있음
- 절대 영점 (absolute zero) 이 반드시 존재
- ex)
  - 마라톤 기록: 1등의 기록이 꼴찌의 기록보다 두배 빠르다. (절대 영점: 마라톤 시작 시간)
  - 10kg 은 20kg 무게의 1/2 이다. (절대 영점: 0kg)



10

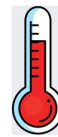
Metric Variable들 중에서도 ratio scale, 즉, 비율척도를 제공하는 것들을 따로 분류해 낼 수 있습니다. 이 경우는 data value들 간 차이의 비율이 반드시 의미가 있는 경우이며, absolute zero, 즉, 절대 영점이 반드시 존재하는 경우입니다. 예를 들면 마라톤 완주 기록에서 "1등의 기록이 꼴찌의 기록보다 두배 빠르다" 라고 말할 수 있다면 이 것은 ratio scale이라 할 수 있습니다. 이 경우, absolute zero는 마라톤 시작 시간이며, 따라서 모든 완주기록들 간의 비율이 의미가 있어집니다. 또 다른 예는 "10kg이 20kg의 0.5배 이다" 라 할 수 있는



것도  
ratio scale이기 때문인데,  
이 때 absolute zero는 0kg이 됩니다.

## Metric Variable – Interval Scale

- Interval Scale (간격 척도)
- Data value들 간의 차이만 계산 가능
- 절대 영점 (absolute zero) 이 없음, 비율 계산 불가
- ex)
  - 마라톤 시작 때 눌렀던 스톱워치를 분실한 경우
    - 선두와 2등 간의 시간 간격 측정만 가능
    - 선두가 2등보다 두배 빠르다 (ratio scale) 는 표현은 불가능
  - 온도
    - 20도와 10도의 차이는 10도
    - 그러나 20도가 10도의 두 배는 아님
    - 0도는 물이 어는 온도이지만 절대적인 “없음” 의 개념은 아님
  - 연도
    - 2024년이 1012년의 두배는 아님
    - 두 연도 간의 간격은 계산 가능



11

Metric variable의 다른 scale 중 하나는 Interval Scale, 즉, 간격 척도 입니다.  
이 경우 우리는 data value들 간의 차이를 계산 가능하지만, 그 차이들 간의 비율이 어떤 의미가 있다고 말하기 어렵습니다.  
이 경우, data의 절대 영점이 없기 때문에 비율을 계산하는 것은 불가능합니다.  
예를 들어 마라톤 시작 때 눌렀던 스톱워치를 분실한 경우를 가정해 봅시다.  
이 때 도착점에 도착한 각 선수의 도착 시간은 기록할 수 있기 때문에 선두와 2등 간의 시간 간격 (interval) 은 측정 할 수 있습니다.

그러나 스타트 시간을 모르고 있기 때문에  
어떤 선수라도 자신이 완주하는데 얼마나 걸렸는지를 모  
르기 되며  
따라서 선두가 2등보다 두배 빠르다는 식의  
Ratio scale의 표현은 불가능 합니다.  
온도의 경우, 20도와 10도의 차이는 10도 입니다.  
그러나 20도가 10도의 두배라고 말할 수는 없습니다.  
0도가 물이 어는 온도이긴 하지만  
절대적인 "없음"을 나타내는 수치는 아니기 때문입니다.  
마찬가지로 연도를 고려할 때  
2024년이 1012년의 두배라고 말할 수는 없습니다.  
다만 두 연도 간의 간격만 계산이 가능합니다.

## Level of Measurement: Examples

		Scale level
1	States of the USA	nominal
2	Product rating on a scale from 1 to 5	ordinal
3	religious confession    고해묵록	nominal
4	CO2 emissions in the year	metric, ratio scale
5	IQ-Score of students	metric, interval scale
6	examination grades from 1 to 5	ordinal
7	telephone numbers of respondents	nominal
8	care level of a patient	ordinal
9	Living space in m <sup>2</sup>	metric, ratio scale
10	job satisfaction on a scale from 1 to 4	ordinal

12

Variable들의 측정 레벨을 다양한 예로 알아 보겠습니다.  
 먼저 미국의 주 이름은 nominal,  
 1 부터 5 까지의 값을 가지는 상품의 품질은 ordinal 입니  
 다.

“고해묵록”이란 카톨릭에서 고해성사를 할 때  
 자신의 죄를 되돌아 보기 위해 자문자답을 하는 리스트를  
 말하는데  
 nominal 입니다.

연간 이산화탄소 방출량은 metric ratio scale입니다.  
 이산화탄소 방출량의 절대 영점은 0이기 때문입니다.

IQ-score는 metric interval scale입니다.

IQ-score의 절대 영점을 정할 수는 없습니다.

시험 성적을 1, 2, 3, 4, 5 등급으로 표시했다면 이것은  
 ordinal 입니다.

응답자들의 전화번호는 nominal 입니다.  
전화번호의 의미적 순서를 정할 수는 없기 때문입니다.  
환자의 돌봄 수준은 ordinal,  
스퀘어 미터 단위로 나타낸 주거 면적은 metric ratio scale,  
1, 2, 3, 4 중 하나로 표시된 업무 만족도는 ordinal 입니다.

## Descriptive Statistics

- 통계적 특성 (characteristics), 차트 (chart), 그래픽 (graphics) 또는 표 (tables)를 사용하여 데이터를 설명하는 (describing) 통계적 방법 (statistical methods)
- Descriptive Statistics 방식의 구분
  - Location parameters (위치 매개변수)
  - Dispersion parameters (분산 매개변수)
  - Tables
  - Charts

13

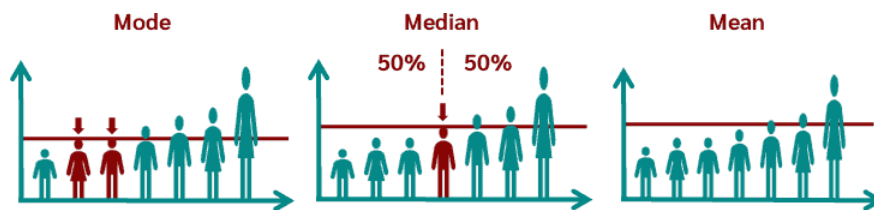
기술통계를 좀 더 구체적으로 정의하면,  
통계적 특성이나 차트, 그래픽 또는 표와 같은 방법을 사용하여

통계 결과를 visualize 하거나 설명하는 통계적 방법이라  
할 수 있습니다.

우리는 이 강의에서 기술통계 방식들을  
location parameter들 (즉, 위치 매개 변수),  
dispersion parameter들 (즉, 분산 매개 변수),  
Table과 Chart로 나누어 설명할 것입니다.

## Location Parameters

- Measures of central tendency (중심화 경향)
- Data distribution의 "center" 의 위치에 대한 정보
- Mean, Median, Mode



14


Location parameter는 다른 말로 measures of central tendency, 즉, 중심화 경향이라 부릅니다. 이 것은 데이터의 분포에서 "중심", 즉, center의 위치에 대한 정보들을 말하며, mean, median, mode가 있습니다.

## Mean

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mean Value      Number of values      Value at the i-th position

Arithmetic Mean



1	2	3	4	5
21	25	10	8	11

$\frac{21 + 25 + 10 + 8 + 11}{5} = 15$

Data들이 반드시 양을 나타내는 숫자로 표현되는 **metric data**이어야 함

15

Mean은 말 그대로 평균을 말합니다.  
Mean은 모든 데이터 값을 다 더하여  
이를 데이터의 개수로 나누어 구하며, 산술 평균이라 불립니다.  
예를 들어 학생이 받은 다섯 과목의 성적의 mean은  
다섯개의 성적을 모두 더해 5로 나누어 구합니다.  
Mean을 계산하기 위해서는 모든 데이터가  
양을 나타내는 숫자로 표현될 수 있어야 합니다.  
즉 data를 나타내는 variable이 metric (quantitative) 이어야 합니다.



## Root Mean Square

### Root Mean Square

$$\bar{x}_{RMS} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

- = RMS, Quadratic Mean
  - 절대값의 mean 측정과 유사
  - 그러나, RMS는 미분가능
  - 양수와 음수가 섞인 데이터들의 합보다 그 절대값, 즉, 변동폭이 더 중요한 경우

- ex) 머신러닝 모델의 예측값(output)과 진실값(ground truth) 간의 RMS 오차 측정
  - 예측값: [3, 5, 2.5, 7], 진실값: [3, 5.5, 2, 8]
  - $\{(3-3)^2 + (5-5.5)^2 + (2.5-2)^2 + (7-8)^2\} / 4 = 1.5 / 4 = 0.375$
  - $\sqrt{0.375} = 0.612$

16

Root Mean Square 는 흔히 RMS라고 불리는데, Quadric Mean이라고 불리며,  
각 개체의 제곱의 산술평균에 루트를 씌운 것입니다.  
RMS는 절대값의 mean을 측정하는 의미를 가지고 있다 할 수 있으나,  
RMS 함수가 절대값 함수와 달리 미분 가능하기 때문에 훨씬 유리합니다.

일반 산술평균의 경우 음수와 양수를 더하면 서로 상쇄 되므로,  
예러 측정과 같은 용도로는 적당하지 않습니다.  
RMS는 기준점으로부터의 거리 (오차)를 나타내는데 많이 쓰입니다.  
Example에서와 같이 machine learning 모델의 ouput인  
prediction value와 진실값 ground truth 간의 오차를 RMS로 많이  
측정합니다.

예를 들어, 예측값이 {3, 5, 2.5, 7}, 진실값이 {3, 5.5, 2, 8} 이라면  
차의 제곱의 평균은  $\{(3-3)^2 + (5-5.5)^2 + (2.5-2)^2 + (7-8)^2\} /$

4  
=  $1.5 / 4 = 0.375$  가 되고,  
root를 씌우면 0.612 가 RMS 오차가 됩니다.

## Geometric Mean

### Geometric Mean

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

- Geometric Mean (기하평균)
  - n개 데이터를 모두 곱한 후, 그 값의 n번째 root를 구함
  - 비율이나 성장률의 평균을 구할 때 사용

- ex) 3년 간 주식 평균 수익 율 계산
  - 1년차: 10%, 2년차: 20%, 3년차: -10%
  - 소수점으로 변환하면: 1.10, 1.20, 0.90
  - 모두 곱함:  $1.10 \cdot 1.20 \cdot 0.90 = 1.188$
  - 세제곱근 (3개 데이터이므로) =  $\sqrt[3]{1.188} \approx 1.058$
  - 즉  $(1.058 - 1.0) = 0.058 = 5.8\%$  수익

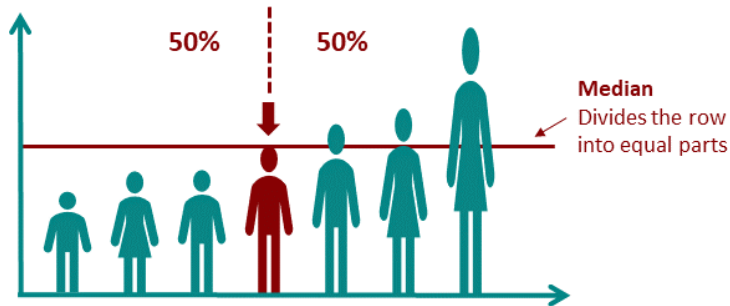
17

Geometric Mean, 즉, 기하평균은  
n개 데이터를 모두 곱한 값의 n번째 root를 구합니다.  
GM은 비율이나 성장률의 평균을 구할 때 사용합니다.

예를 들어 3년간 주식 평균 수익율을 계산한다고 가정합니다.  
1년 차에 수익율은 10%, 2년 차에는 20%, 3년 차에는 -10%로 주어졌다고 할 때  
각 연차의 원금과 수익을 합한 금액을 소수점으로 변환하면: 1.10, 1.20, 0.90이 됩니다.  
이를 모두 곱하면:  $1.10 \cdot 1.20 \cdot 0.90 = 1.188$   
3개의 데이터이므로, 1.188의 세제곱근을 구하면 1.058이 됩니다.  
따라서 3년간의 평균 수익율은  $(1.058 - 1.0) = 0.058 = 5.8\%$ 가 됩니다.

## Median

- Data를 sorting하여 순서대로 세웠을 때 중간에 오는 값
- Data가 반드시 metric data일 필요는 없으나, 순서를 정할 수 있어야 함 (Ordinal data)



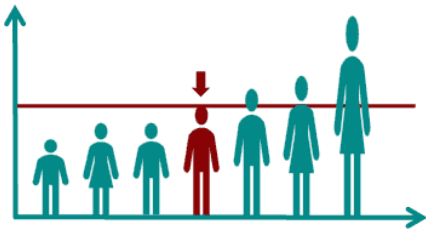
18

Median은 data를 크기 순서대로 sorting하여 세웠을 때, 중간에 오는 값을 말합니다.  
따라서 data가 반드시 metric일 필요는 없으나 순서를 정할 수는 있어야 하며, 최소 ordinal 이상의 data여야 합니다.

## Median – Odd and Even Number of Values

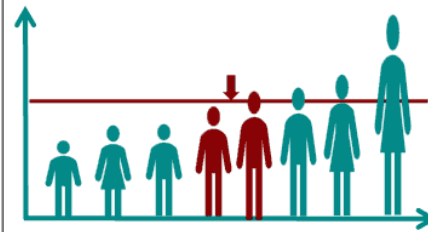
### Odd number of values

The median is a value that actually occurs.



### Even number of values

The mean value of the two middle values (둘 사이의 평균)

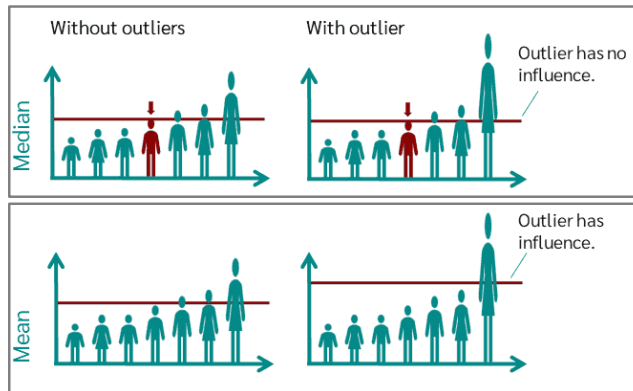


19

데이터의 수가 홀수라면 median은 유일하게 하나입니다.  
그러나 짝수개의 데이터라면 딱 중간이 될 후보가 두개 있을 것이고,  
우리는 이 둘의 평균을 구해 median 값으로 사용합니다.

## Mean vs Median

- Median 은 scattering에 robust함. Outlier가 median에 주는 영향이 적음

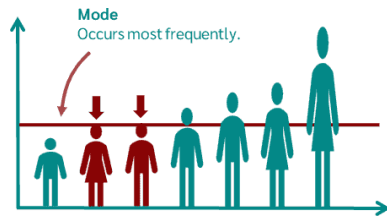


20

Median은 순서만을 따지기 때문에  
특출하게 값이 크거나 작은 outlier가  
median 값을 구하는데 영향을 덜 미치게 됩니다.  
그러나 그림에서 보듯 mean은 모든 값이 계산에 참여하기  
때문에  
outlier의 영향을 많이 받습니다.

## Mode (Modal Value)

- Most common value (가장 많이 출현하는 값) = Most frequent value
- Data는 구별되기만 하면 됨 (Nominal, Ordinal, Metric)



Car brand	Daimler	BMW	VW	Audi
Frequency	20	25	10	15

Frequency Table

Mode는 가장 많이 출현하는 값을 말합니다.  
따라서 순서와는 전혀 무관하며,  
데이터가 구별되기만 하면 되기 때문에  
Nominal, Ordinal, Metric data가 모두 사용될 수 있습니다.  
그림의 예에서는 어떤 집단의 사람들이 소유한  
자동차 브랜드의 빈도 수를  
frequency table로 나타낸 것입니다.  
이렇게 정리해 놓으면 Mode를 찾기가 쉬워질 것입니다.

## Comparisons

	Advantages	Disadvantages	Data
Mean	Most used	Sensitive to outlier	Metric
Median	Robust against outliers	Not utilizing all the information in the data	Metric, Ordinal
Mode	Computable for Non numerical data	Not reflect the characteristics of entire data	Metric, Ordinal, Nominal

22

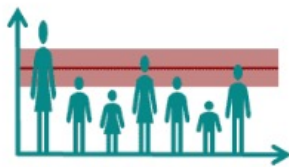
이제 Mean, Median, Mode를 서로 비교해 보겠습니다.  
Mean은 가장 많이 사용되고 있는 익숙한 개념입니다.  
그러나 mean은 outlier에 취약하다는 단점이 있으며,  
다루는 데이터도 반드시 metric 이어야 합니다.  
Median은 outlier에 영향을 덜 받으나,  
데이터의 전체적인 정보를 사용하지 않고  
단순히 순서만을 사용한다는 단점이 있습니다.  
하지만 Ordinal 데이터에서도 median을 구할 수 있으며  
이는 mean에 비해 장점이 될 수 있습니다.  
Mode는 Nominal 즉 순서가 없는 데이터에도 사용될 수  
있으나,  
개수만을 세는 것이므로  
데이터의 정보를 가장 덜 활용하게 된다는 단점이 있습니  
다.



## Dispersion parameter

- Describe the **scatter of values** of a sample around a location parameter

Standard deviation



Average distance of all measured values from the mean value

Range



Distance between lowest and highest value of a distribution

Interquartile range

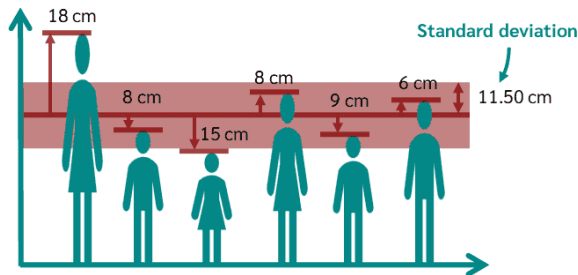


Spectrum in which the middle 50% of the values lie. Difference between first and third quartile

Dispersion parameter, 즉, 분산 파라미터는 location parameter를 중심으로 데이터 값이 흩어져 있는 상태를 표현합니다. 표준편차 (standard deviation), 구간 (range), 사분위수 (interquartile range) 범위 등이 있습니다.

## Standard Deviation

- Mean deviation (root mean square) of all measured values from the mean
- Indicates the spread of a variable around its mean



std of population:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$n$  is the number of persons  
 $x_i$  is the size of the individual  
 $\bar{x}$  is the mean value of all persons

std of sample:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

24

표준편차는 data value들이 mean으로부터 떨어져 있는 평균 deviation을 말합니다.

즉, 모든 측정값들의 root mean square, RMS 입니다.

표준편차는 mean을 중심으로 variable이 얼마나 퍼져 있는가를 나타냅니다.

Population의 표준편차는 한 개체가 가지는 value와 population mean간의 차이의 제곱의 평균을 구해 거기에 root를 씌우면 됩니다.

그러나 Population 전체를 대상으로 측정값과 평균값을 구하는 것은 불가능하기 때문에

Sample의 표준편차를 구하게 되는데,

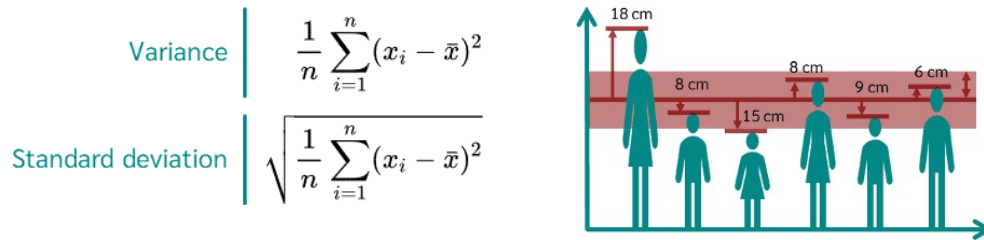
이 식을 보면 root 안의 제곱을 평균을 구하는 데 Sample size인  $n$ 이 아닌

$(n-1)$ 로 sum을 나누고 있습니다.

이 이유에 대해서는 잠시 뒤에 설명하도록 하겠습니다.

## Variance

- measures the squared average distance from the mean



25

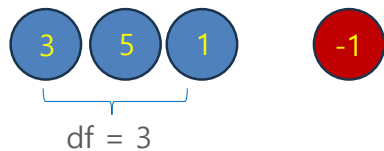
Variance, 즉, 분산은 data value와 mean과의 차이의 제곱의 평균입니다.

따라서 Variance에 root를 씌우면 표준편차가 됩니다.

앞 장의 슬라이드에서 보았듯이 Population이 아닌 Sample의 variance는  $n$ 이 아닌  $(n-1)$ 로 나누어야 합니다.

## Why Divide by (n-1) in Sample Variance?

- Degree of Freedom (df)
  - The number of ways that observed data values can vary independently
  - data를 자유롭게 선택할 수 있는 개수
  - $(\# \text{total data}) - (\# \text{parameters being estimated})$
  - ex)  $n$ 개 data - 1 (평균): 분산에서는 평균을 먼저 정해 놓고 계산
  - ex)  $n = 4$  개의 수를 선택하되 그들의 평균이  $m = 2$  이 되도록
    - $n - 1 = 3$  개 수 까지는 자유롭게 선택 가능
    - 나머지 한 개는 자유롭게 선택 불가. 평균을  $m = 2$  로 만들어야 하므로



- 결국  $x = -1$  로 고정되며,  $df = 3$  이 됩니다.

26

Population이 아닌 Sample의 variance를 구할 때에는 Sample size  $n$ 이 아닌  $(n-1)$ 로 나눈다고 하였습니다. 그 이유는 무엇일까요?

이를 이해하기 위해서는 "Degree of Freedom", 즉, 자유도의 개념을 이해해야 합니다. 자유도는 약자로 df라고 씁니다. 자유도는 독립적으로 정해질 수 있는 관찰값의 갯수, 간단히 말하면, 데이터를 자유롭게 선택할 수 있는 갯수를 말합니다.

df를 계산하는 식은 전체 데이터 갯수에서 현재 이미 정해져 있다고 가정한 어떤 estimated parameter의 갯수를 빼면 됩니다. 예를 들면 Sample variance의 계산에서는 미리 Sample mean을 계산해 둔 후 variance 계산에 들어

갑니다.

이 때 데이터들이 무엇인지를 전혀 모르고

다만 sample mean만 아는 상황이라 가정해 봅시다.

필요한  $n$ 개 데이터 중  $(n-1)$  개 까지는 자유롭게 정하는 것이 가능합니다.

그러나 마지막 하나의 value는 자유롭게 정할 수 없습니다.

이미 sample mean의 값을 정해 놓았기 때문에

마지막 value는 어떤 값으로 고정될 수 밖에 없는 것입니다.

따라서 sample variance의 계산에서 자유도는  $(n-1)$  이 됩니다.

따라서 sample variance의 계산에서는  $n$ 으로 나누는 것이 아니라

자유도인  $(n-1)$ 로 나누게 되는 것입니다.

## Range

$$R = x_{max} - x_{min}$$

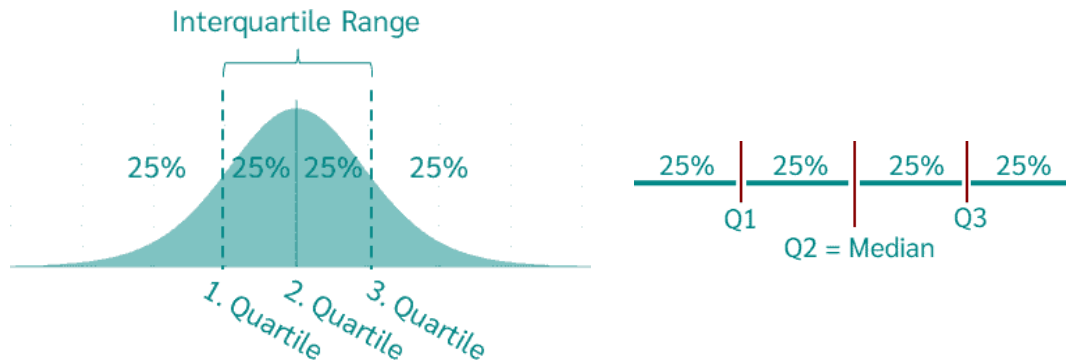


27

Range는 value의 최대값에서 최소값을 뺀 구간을 말합니다.

Range도 중요한 dispersion parameter가 됩니다.

## Quartile, Interquartile Range (IQR)



28

Quartile Range 또는 Interquartile Range,  
줄여서 IQR은 전체 range의 1분위 point Q1과  
3분위 point Q3 사이의 range를 말합니다.  
그림에서 보듯 Q2는 전체 데이터 갯수 중 딱 절반을 가르  
는 point이며,  
이것은 median 과 같습니다.  
Q1과 Q3 사이에는 전체 데이터 중 50%의 데이터가 존재  
하게 됩니다.



## Frequency Table

Car brand	Frequency	%	Valid %
VW	3	25%	27.27%
Ford	3	25%	27.27%
BMW	2	16.67%	18.18%
Opel	2	16.67%	18.18%
Daimler	1	8.33%	9.09%
Total	11	91.67%	100%
Invalid	1	8.33%	
Total	12	100%	

29

이제 Descriptive Statistics에서 사용되는 table의 종류를 살펴보겠습니다.


이 table은 어떤 주택단지 거주민들이 소유하고 있는 자동차의 브랜드들의 빈도를 나타내고 있습니다.

자동차 브랜드명, 빈도, 퍼센테이지와 invalid brand를 제외하였을 때의 퍼센테이지를 column들로 하고 있습니다.

빈도의 합은 주택단지의 자동차 대수와 같아야 하며, 퍼센테이지 컬럼의 합은 100%가 되어야 함을 알 수 있습니다.

## Contingency Table

- Frequency Table을 다른 조건 분류로 세분화



	Cake	Ice	Donut	Total
Female	4	3	6	13
Male	5	7	9	21
Total	9	10	15	34

Female and without a degree occurs 6 times in the data

	Female	Male
Without graduation	6	7
College	13	16
Bachelor	16	15
Master	8	11
Total	43	49

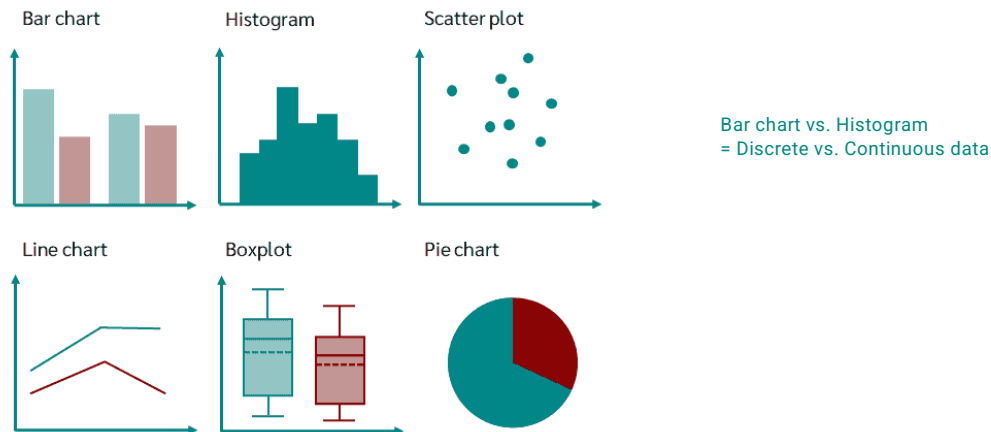
With umbrella				
		yes	no	Total
Gender	female	5	7	12
	male	5	5	10
	Total	10	12	22

30

Frequency Table은 목적에 따라 더 세분화 된 table로 나타내어 질 수도 있습니다. 첫번째 table은 성별에 따라 선호하는 디저트들의 frequency를 정리해 놓은 것입니다. 그러니까 앞 장의 슬라이드의 table이 오로지 하나의 기준인 자동차 브랜드를 기준으로 정리한 frequency table이었다면, 이 table은 성별과 디저트라는 두 개의 variable들로 정리해 놓은 것이라 볼 수 있겠습니다. 오른쪽의 두번째 table은 성별과 학력이라는 두 variable들을 기준으로 정리한 frequency table입니다. 학위없음, 전문학사, 학사, 석사로 학력을 구분하였습니다.

마지막 table은 5분간 특정 위치를 지나는 사람들을 관찰  
한 결과를  
성별과 우산소지유무라는  
두 개의 variable을 기준으로 정리한  
frequency table입니다.

## Charts



31

Chart는 특정한 하나 또는 여러 개의 variable의 분포를 도식화 하여 이해하기 쉽게 또는 비교하기 쉽게 그려주는 기법입니다.

우리가 이미 많이 접해왔듯이

Chart에는 Bar chart, Histogram, Scatter plot, Line chart, Boxplot, Pie chart 등이 있습니다.

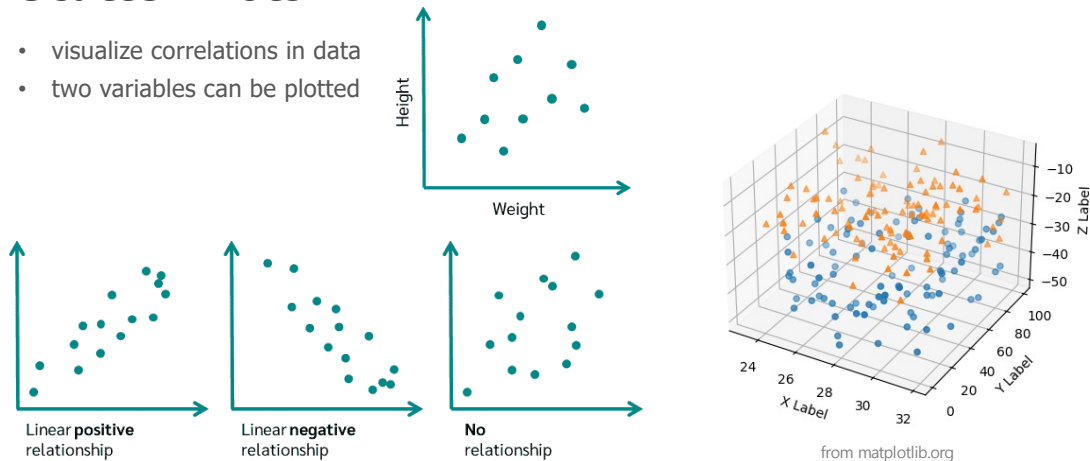
특히 Bar chart와 Histogram은 매우 유사합니다.

Bar chart는 discrete data를 표현하는 것이기 때문에 하나의 bar가 variable의 특정한 값의 개수를 나타낸다고 볼 수 있습니다.

반면에, Histogram은 continuous data를 위한 것으로, 하나의 bar가 어떤 특정 구간에 속해 있는 값의 개수를 표현하는 것입니다.

## Scatter Plots

- visualize correlations in data
- two variables can be plotted



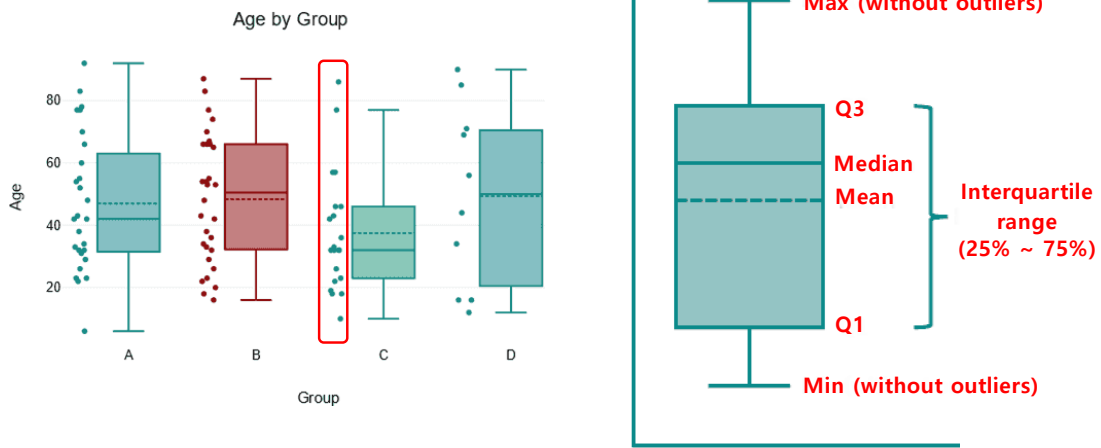
32

특별히 Scatter Plot은 여러 개의 variable들의 value들 간의 관계를 잘 보여주는 chart입니다. 이러한 variable들 간의 관계를 흔히 correlation이라 하는데, plot된 point들이 직선에 더 가까워지면 correlation이 높다고 볼 수 있습니다. 또, positive와 negative correlation이 있을 수 있는데 plot들의 set이 나타내는 직선의 기울기가 positive인지 negative인지의 여부에 따라 부르는 말입니다. positive인 경우는 하나의 variable이 증가할 수록 다른 쪽도 증가하지만, negative일 때는 한쪽이 증가할 때 다른 쪽은 감소하게 됩니다.

세개의 variable들이 참여하는 3D scatter plot의 경우도 있는데  
이 경우에는 3D volume에 point들을 나타내게 됩니다.

## Boxplot

- Compare and contrast two or more groups



33

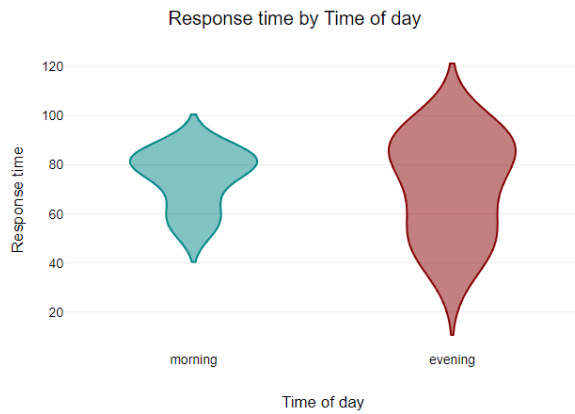
Boxplot은 흔히 두 개 이상의 group들간의 데이터를 비교하여 보여주기 위해 많이 사용됩니다.  
그림에서는 Boxplot들이 A, B, C, D의 네개의 group들에서 각각 구성원들의 나이 분포를 나타내고 있습니다.  
Boxplot 왼쪽의 점들은 각각의 data point를 나타내고 있습니다.  
Boxplot이 표시하는 정보로는 먼저 Outlier가 있습니다.  
이것들은 측정된 data의 일부이지만 평균에서 너무 많이 동떨어져 있는 비정상 데이터라고 할 수 있으며, 평균 등 통계 데이터 계산에서는 제외됩니다.  
Max와 Min은 Outlier들을 제외한 데이터 중에서

가장 크고, 가장 작은 값을 말하며  
Min과 Max로 부터 box에 이르는 세로줄을 수염과 같이  
그립니다.  
가운데 box의 가장 아래 값은 Q1, 즉, 25% 선의 data 값을  
말하며  
box의 가장 위의 값은 Q3, 즉, 75% 선의 data 값을 말합니  
다.  
따라서 box 내부의 영역은 데이터의 중간 부분의  
50% 데이터를 포함합니다.  
Box 내부에는 Median과 Mean 값을 각각 직선과 점선으로  
표시합니다.



## Violin Plot

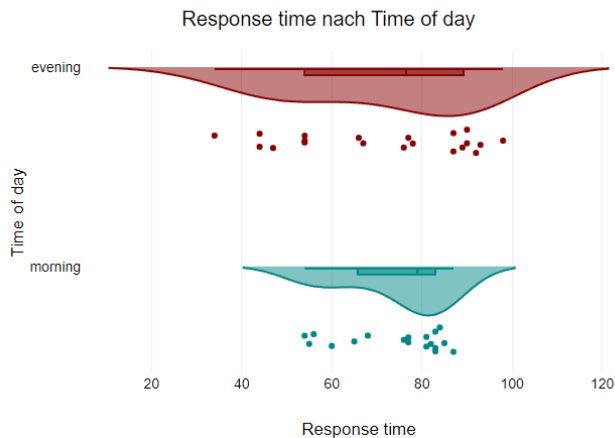
- Similar to boxplot, showing probability distribution



34

Violin Plot은 boxplot과 유사하지만 사각형이 아니라 실제 데이터의 분포를 나타내는 폐곡선을 사용합니다.

## Raincloud Plot



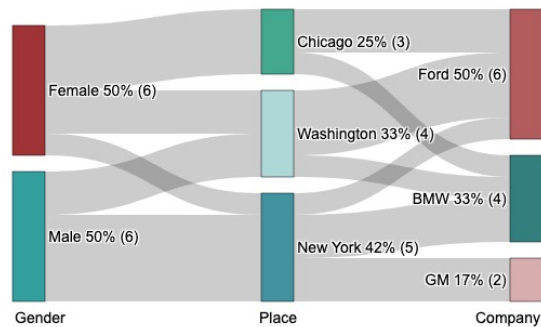
35

Raincloud Plot은 위쪽에 구름과 같은 모양의 데이터 분포를 나타내는 곡선 영역이 있고, 그 아래에 실제 데이터 포인트의 위치를 점으로 찍어 나타냅니다. 이 모습이 마치 구름에서 비가 떨어지는 모양처럼 보이기 때문에 Raincloud plot이라는 이름을 가지게 되었습니다.

## Sankey diagram

- 어떤 value들의 set에서 다른 value set으로 value의 flow를 표시
- 여러 단계의 프로세스 간의 에너지, 재료, 또는 비용의 flow를 시각화
- Flow arrow의 너비는 Flow 양에 비례

Gender	Place	Company
Female	Chicago	BMW
Female	Chicago	Ford
Male	New York	BMW
Male	New York	BMW
Female	Chicago	Ford
Female	Washington	Ford
Male	Washington	Ford
Male	Washington	Ford
Female	New York	Ford
Male	New York	GM
Female	Washington	BMW
Male	New York	GM



36

Sankey Diagram의 경우 어떤 value들의 set에서 다른 value set으로 value의 flow를 표시 합니다.  
여러 단계의 프로세스가 있을때  
이 프로세스 간의 에너지, 재료, 또는 비용의 flow를 시각화 합니다.  
Flow arrow의 너비는 Flow 양에 비례합니다.  
그림의 예에서는 처음에 성별을 기준으로 거주 도시로의 flow,  
다시 거주 도시에서 소유한 자동차의 브랜드로의 flow를 표시하고 있습니다.

## Python Code

- <https://github.com/iklee99/StatCode>
  - 01\_descriptive.py
    - packages needed: pandas numpy matplotlib seaborn scipy
    - for installing package(s): pip install package\_name1 package\_name2 ... (in console)
  - seaborn Package (document): <https://seaborn.pydata.org/>

37

지금까지 학습한 Descriptive Statistics의 대부분은 python programming으로 어렵지 않게 계산할 수 있습니다.

01\_descriptive.py 를 한 가지 코딩의 예로 받아보시기 바랍니다.

필요한 package들은 numpy, matplotlib, seaborn, scipy 등입니다.

특히 seaborn package는 이 강의에서 보여준 것 이외에도 매우 다양한 형태의 chart들을 쉽게 visualize할 수 있게 해주며,

color, 간격, 범위 등을 customize할 수 있는 기능들이 매우 강력하므로,

꼭 한번 그 document를 보기를 추천합니다.

## Free Utilities

- JASP (<https://jasp-stats.org/>)
  - 무료 (Open-source)
  - Frequentist Analyses (지금 이 note들에서 다루는 방식) – 가설검정: p-value 기반
  - Bayesian Analyses (확률을 주관적 신념, 사전 정보로 정의) – 가설검정: 확률적 추론을 사용하여 특정 가설이 참일 확률을 계산할 수 있음. ex) Null hypothesis 가 true일 확률은 얼마인가?
  - Free textbook: “Learning Statistics with JASP”
- SOFA (<https://www.sofastatistics.com/home.php>)
- GNU PSPP (<https://www.gnu.org/software/pspp/>): Alternative to SPSS
- SCI Labs (<https://www.scilab.org/software/scilab/statistics>)
- Jamovi (<https://www.jamovi.org/>)
- MacAnova (<https://www.stat.umn.edu/macanova/macanova.home.html>)
- Devele (<https://develve.net/>)
- InVivoStat (<https://invivostat.co.uk/>)
- IBM SPSS (<https://www.ibm.com/spss>)