

S_01 Descriptive Statistics

Statistical Analysis

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/>)

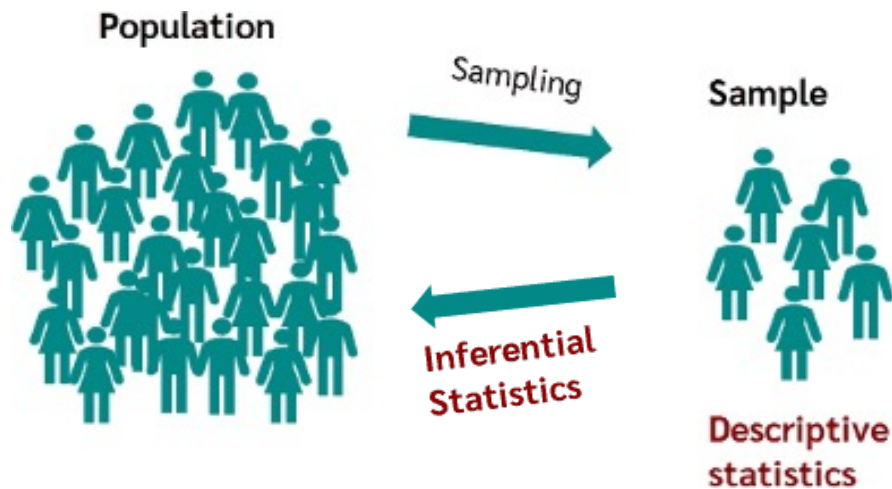


Population and Sample

- 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



Descriptive vs Inferential Statistics



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

Types of Variables

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

Nominal Variables

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

Examples:

Gender

1 = male

2 = female

Marital status

single

married

divorced

widowed

Preferred newspaper:

The Washington Post

The New York Times

USA Today

...

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

Quantitative (Metric) Variables

- Equal, unequal, greater, smaller, **difference**, **sum**

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
...

Level of Measurement

Nominal

Characteristics can be distinguished.

A D
C B

Ordinal

Characteristics can be sorted.

$A < B < C < D$

Metric

Distances between the values can be calculated.

1,4 1,6 1,8 2 2,2



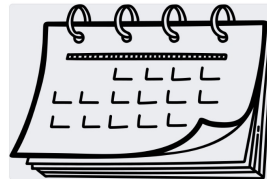
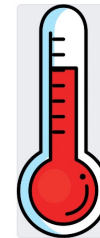
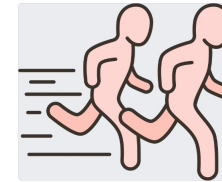
Metric Variable – Ratio Scale

- Ratio Scale (비율척도)
- Data value들 간의 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)



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



Level of Measurement: Examples

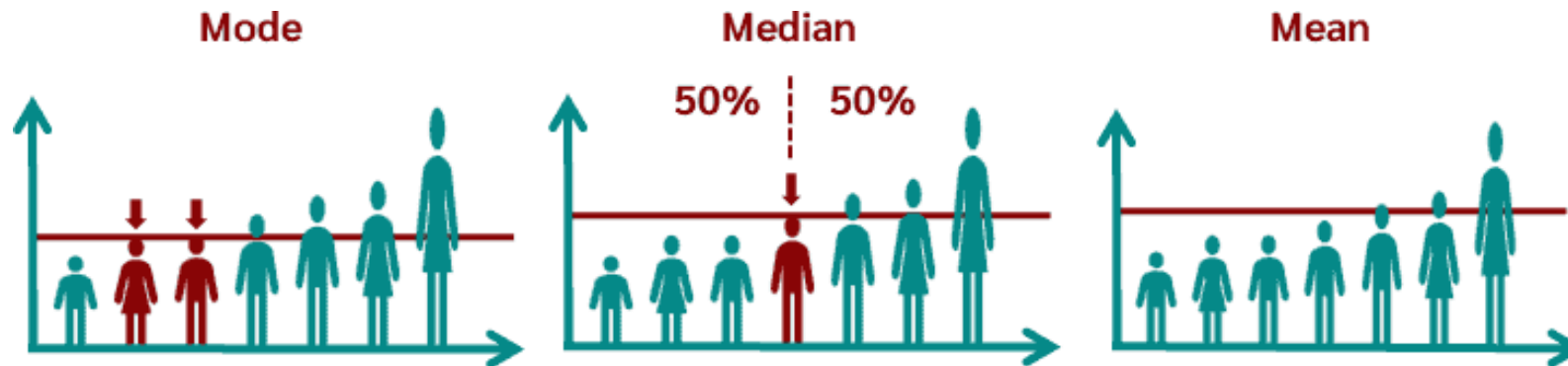
		Scale level
1	States of the USA	nominal
2	Product rating on a scale from 1 to 5	ordinal
3	religious confession 고해목록	nominal
4	CO2 emissions in the year	metric, ratio scale
5	IQ-Score of students	metric, interval scale
6	examination grades from 1 to 5	ordinal
7	telephone numbers of respondents	nominal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

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

Location Parameters

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



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

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

Geometric Mean

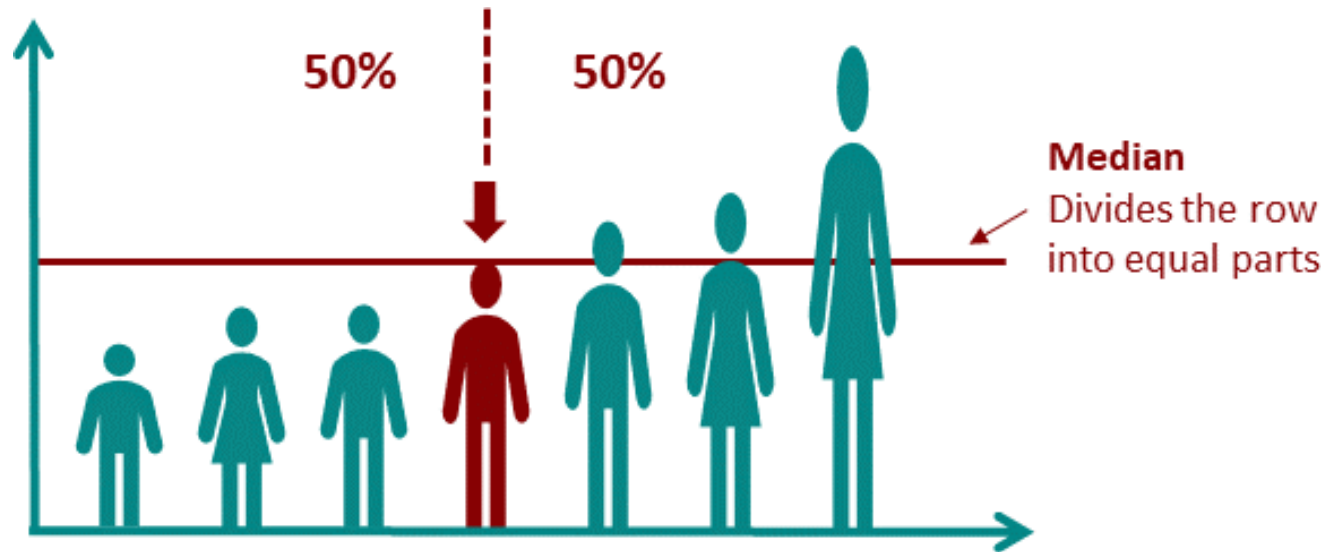
Geometric Mean

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

- 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
 - 1년치 10%, 2년치 20%, 3년치 -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

Median

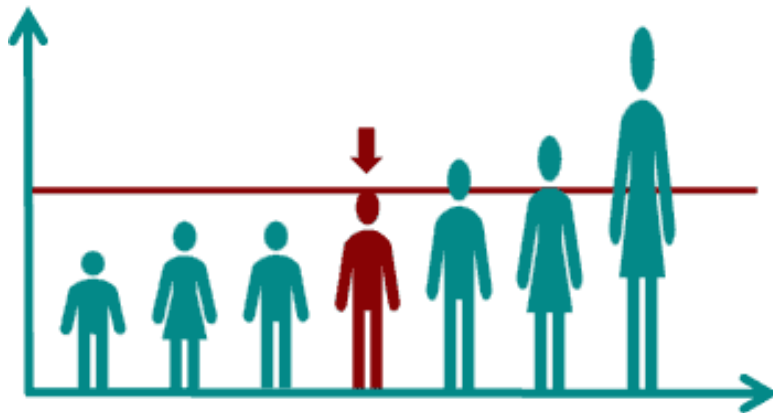
- Data를 sorting하여 순서대로 세웠을 때 중간에 오는 값
- Data가 반드시 metric data일 필요는 없으나, 순서를 정할 수 있어야 함 (Ordinal data)



Median – Odd and Even Number of Values

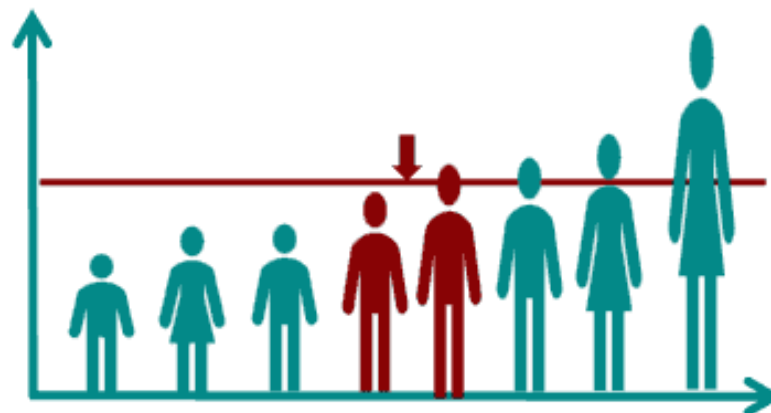
Odd number of values

The median is a value that actually occurs.



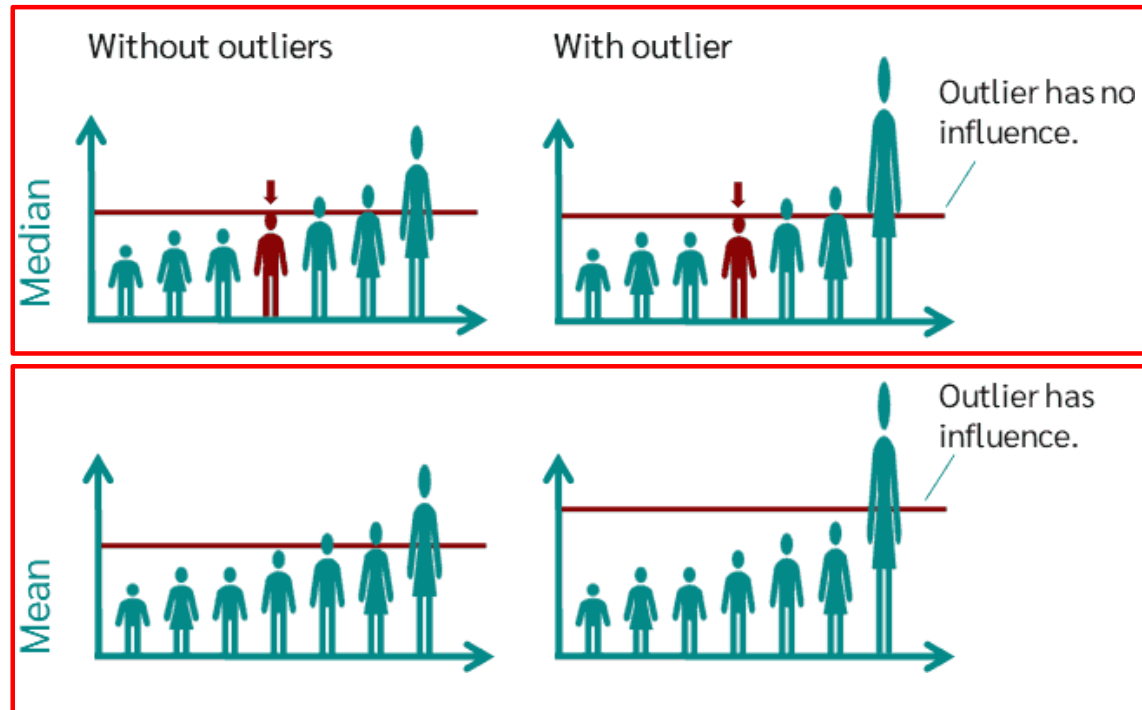
Even number of values

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



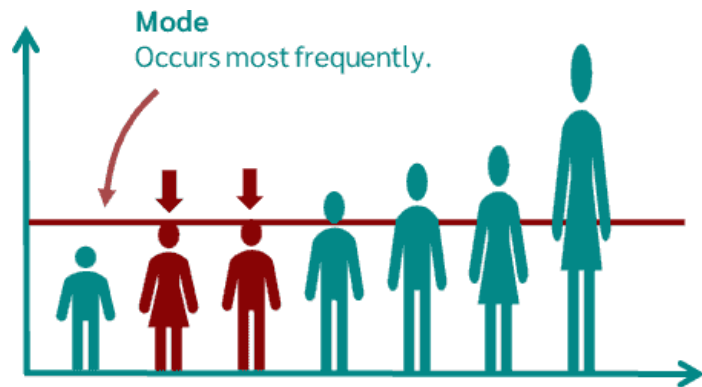
Mean vs Median

- Median 은 scattering에 robust함. Outlier가 median에 주는 영향이 적음



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

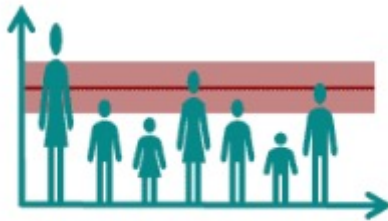
Comparisons

	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

Dispersion parameter

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

Standard deviation



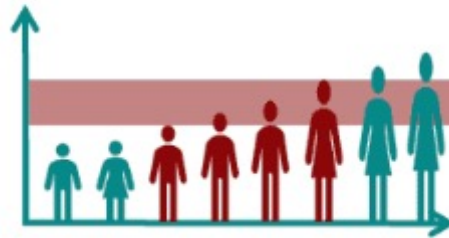
Average distance of all measured values from the mean value

Range



Distance between lowest and highest value of a distribution

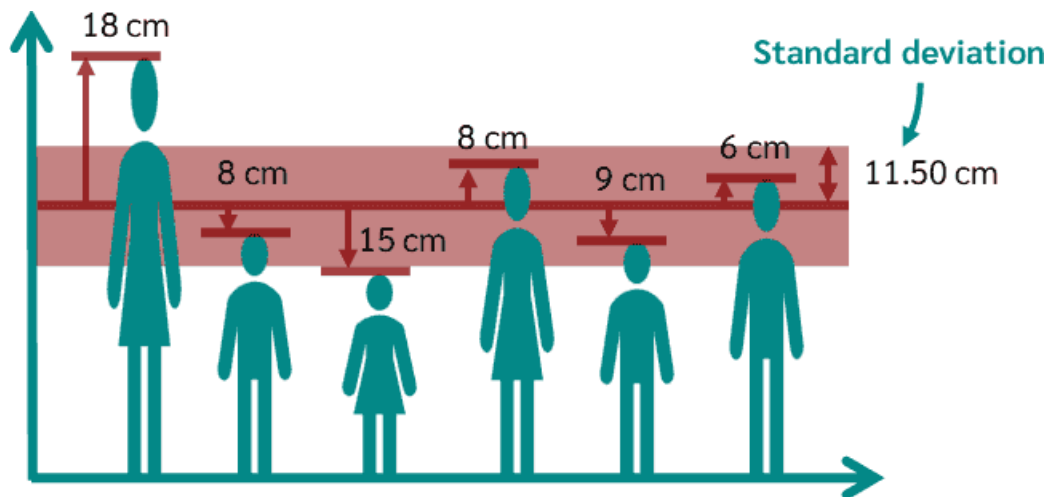
Interquartile range



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

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:

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

Variance

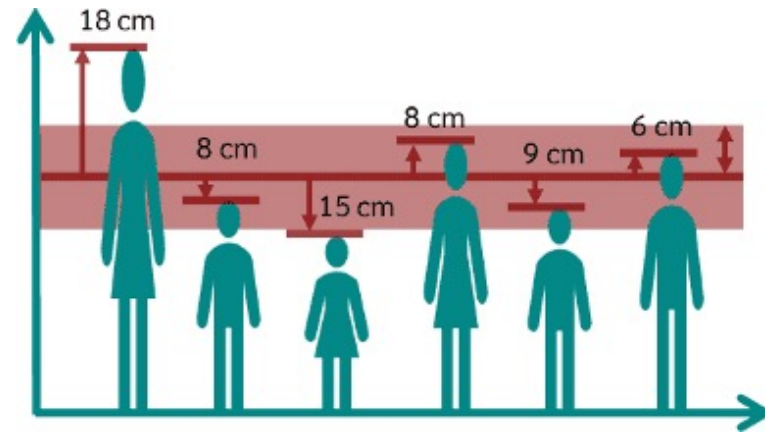
- measures the squared average distance from the mean

Variance

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

Standard deviation

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



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

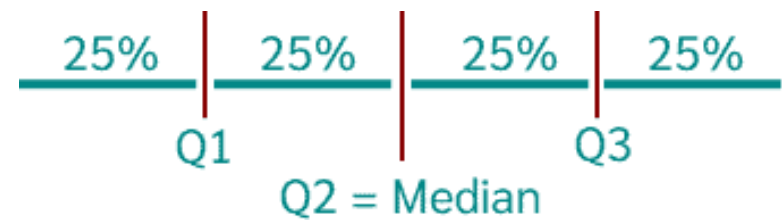
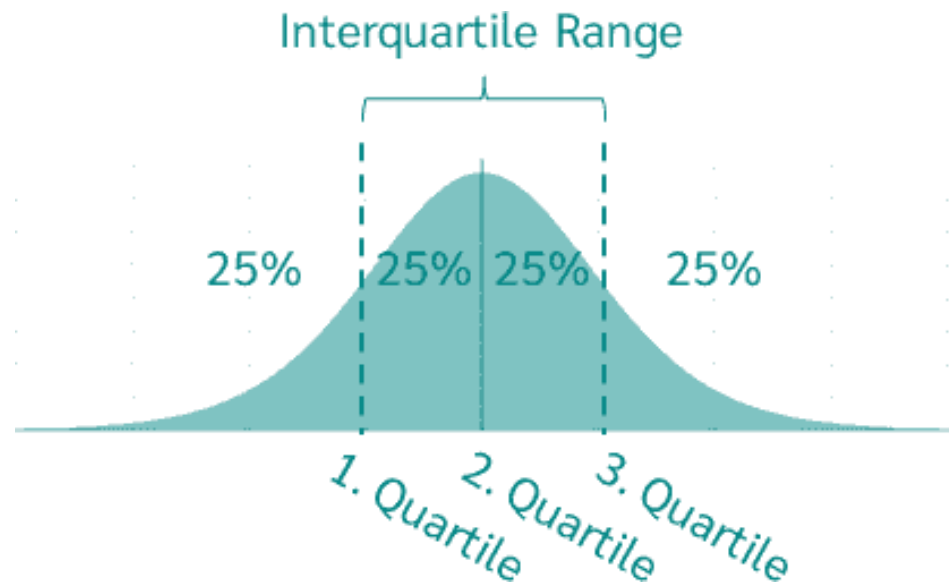
- Degree of Freedom (df)
 - The number of ways that observed data values can vary independently
 - data를 자유롭게 선택할 수 있는 개수
 - (#total data) – (#parameters being estimated)
 - ex) n 개 data – 1 (평균): 분산에서는 평균을 먼저 정해 놓고 계산
 - ex) n 개의 수를 선택하되 그들의 평균이 m 이 되도록 하는 조건으로.
 - $n - 1$ 개 수 까지는 자유롭게 선택 가능
 - 나머지 한 개는 자유롭게 선택 불가. 평균을 m 으로 만들어야 하므로
- ex) df in sample variance
 - Variance
 - Uses “sample mean” as the estimated parameter
 - So, $df = n - 1$
 - So, sample variance:
$$\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

Range

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



Quartile, Interquartile Range (IQR)





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%	

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

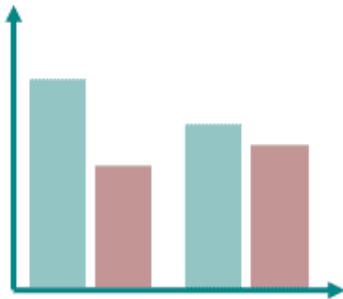
	Female	Male
Without graduation	6	7
College	13	16
Bachelor	16	15
Master	8	11
Total	43	49

With umbrella

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

Charts

Bar chart



Histogram

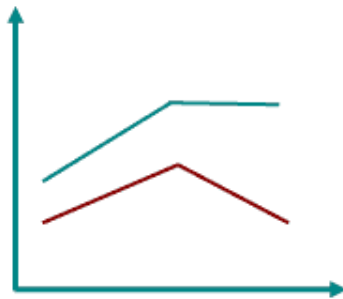


Scatter plot

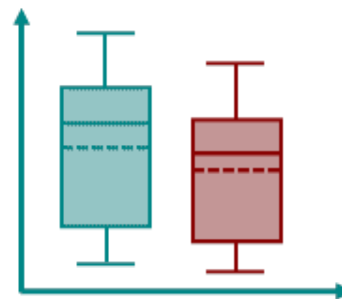


Bar chart vs. Histogram
= Discrete vs. Continuous data

Line chart



Boxplot

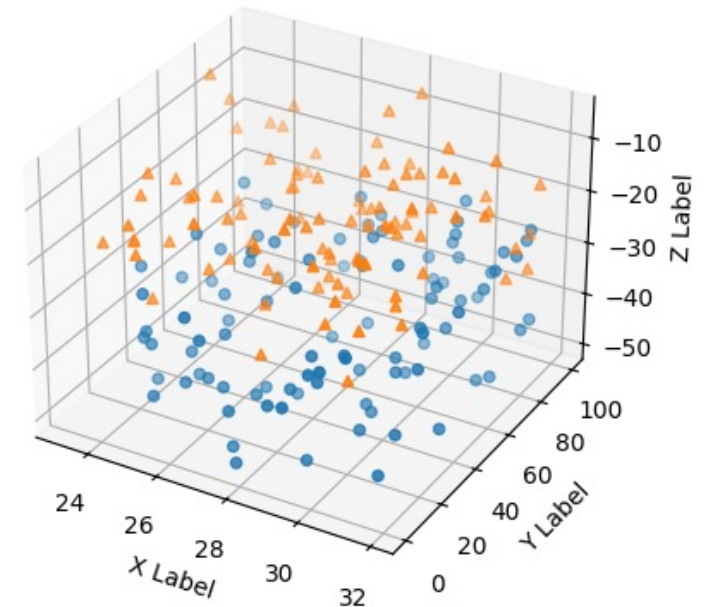
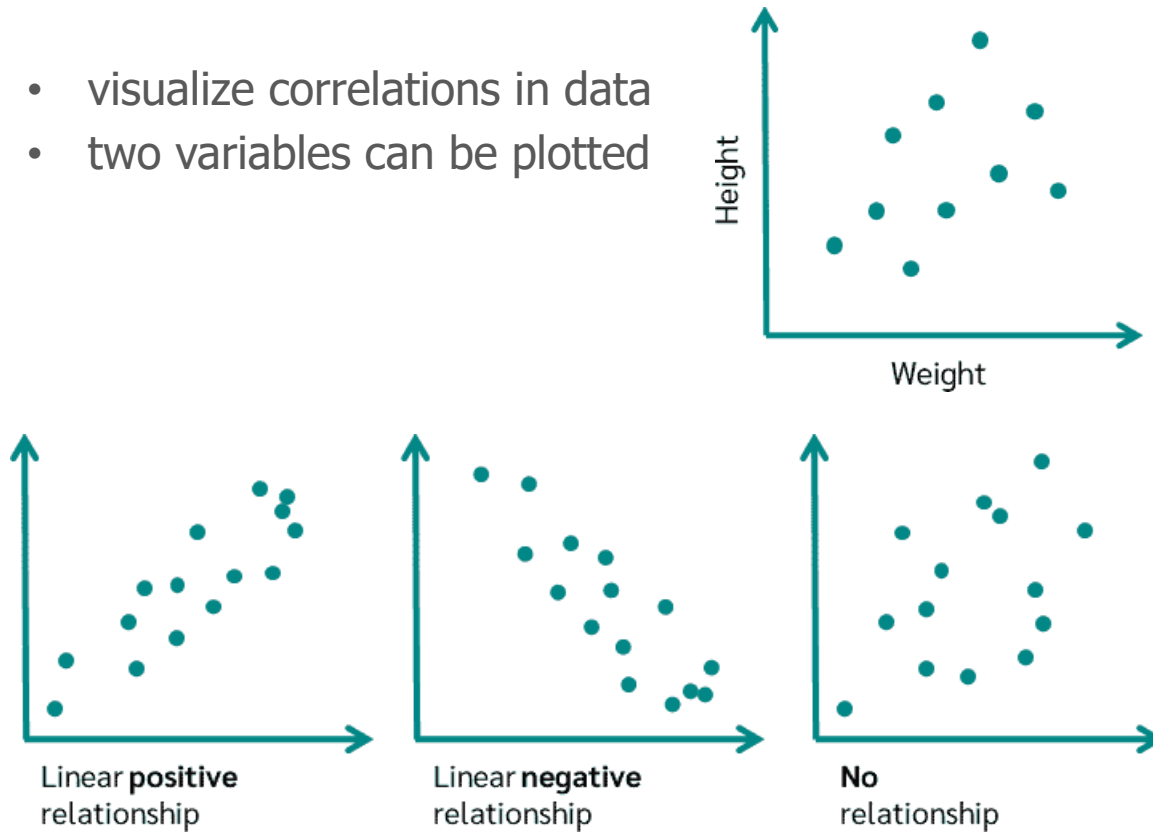


Pie chart



Scatter Plots

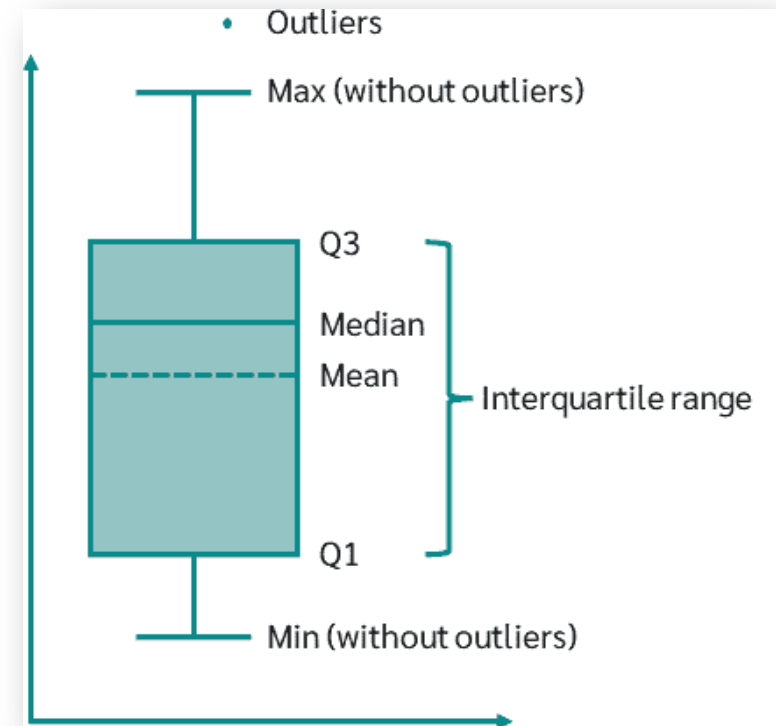
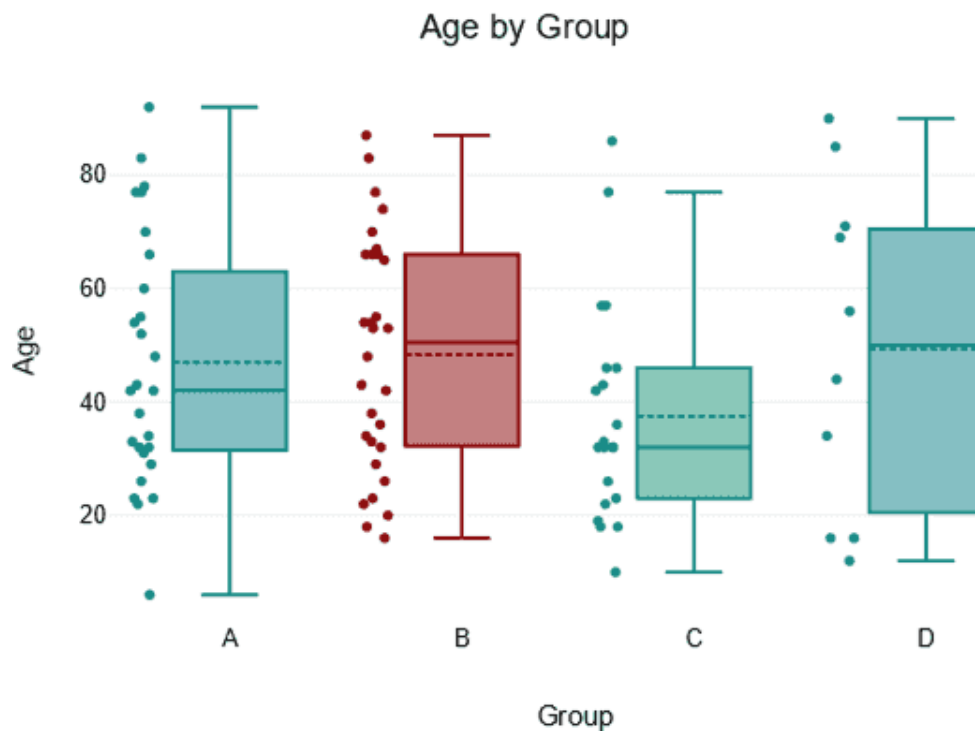
- visualize correlations in data
- two variables can be plotted



from matplotlib.org

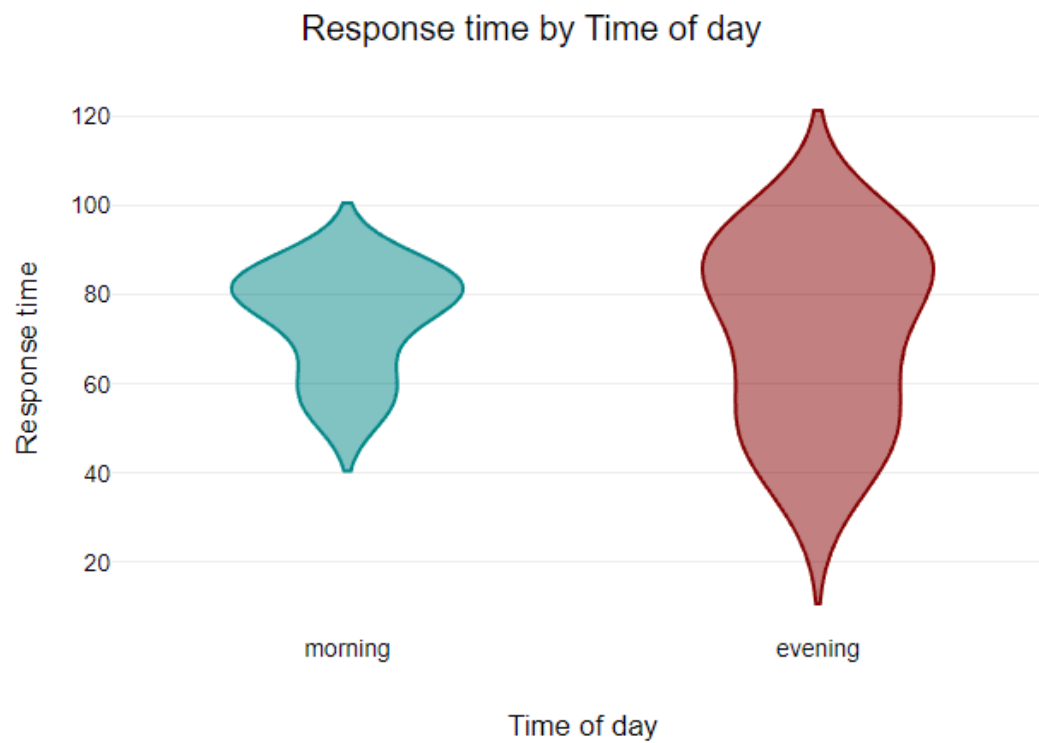
Boxplot

- Compare and contrast two or more groups

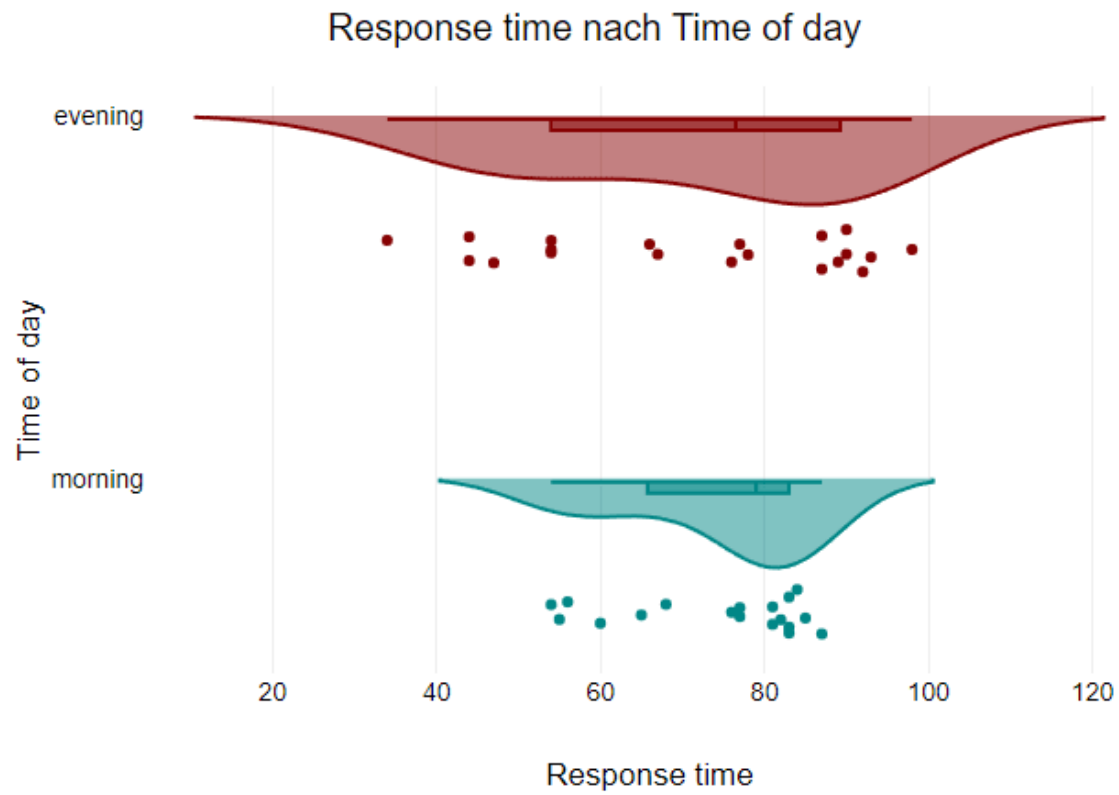


Violin Plot

- Similar to boxplot, showing probability distribution



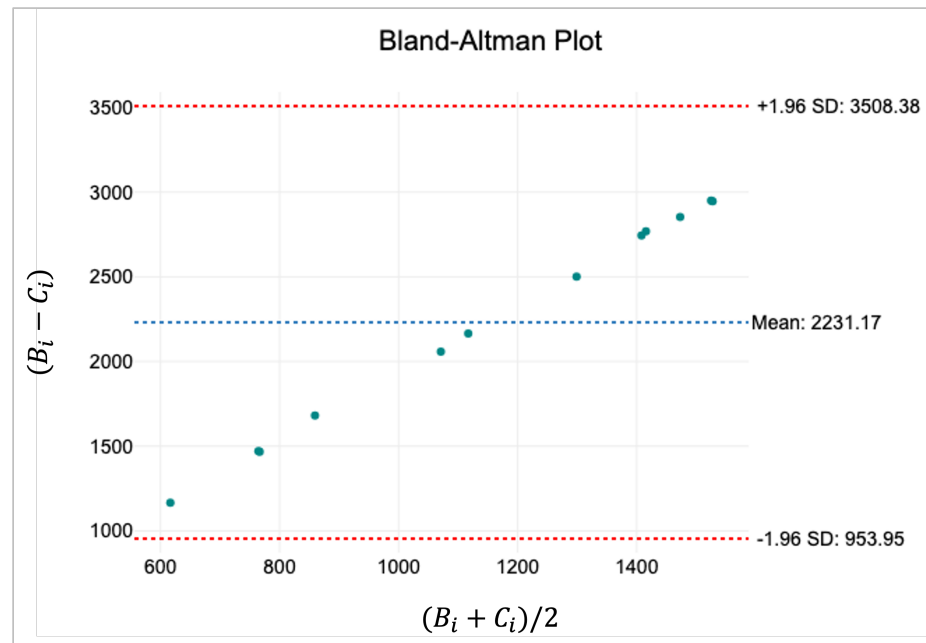
Raincloud Plot



Bland-Altman Plot (1/2)

- 같은 양을 측정하기 위해 두가지 서로 다른 방법에 의해 측정된 값
- 두 방식을 각각 B와 C라 할 때
- Vertical axis: $B_i - C_i$
- Horizontal axis: $(B_i + C_i) / 2$

B	C	$(B+C)/2$	$(B-C)$
1500	33	766.5	1467.0
1200	33	616.5	1167.0
2200	34	1117.0	2166.0
2100	42	1071.0	2058.0
1500	29	764.5	1471.0
1700	19	859.5	1681.0
3000	50	1525.0	2950.0
3000	55	1527.5	2945.0
2800	31	1415.5	2769.0
2900	46	1473.0	2854.0
2780	36	1408.0	2744.0
2550	48	1299.0	2502.0



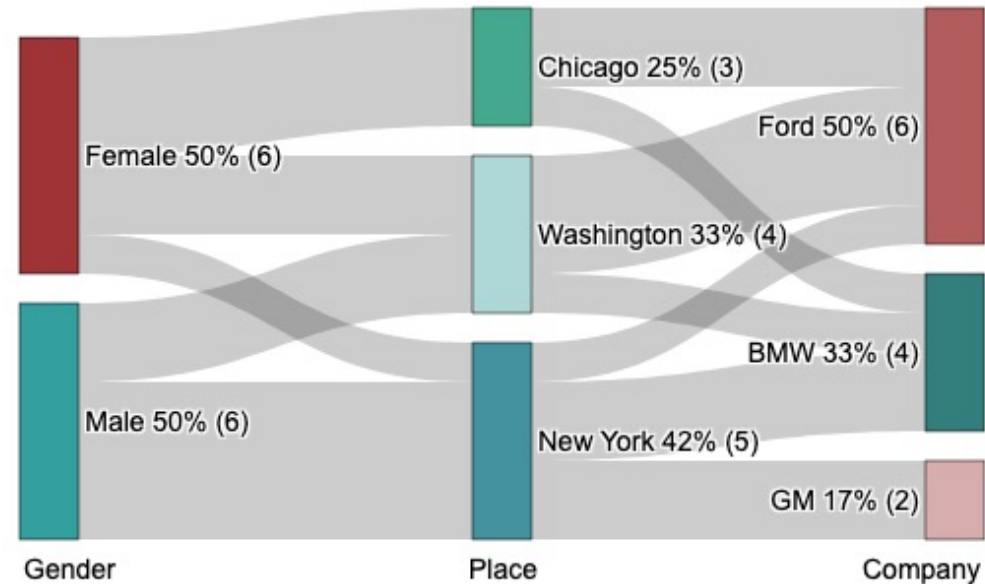
Bland-Altman Plot (2/2)

- 점들이 직선에 가깝게 분포할 때
 - 두 측정방식 간에 비례 바이어스 (proportional bias) 가 존재하지 않음
 - 두 측정방식에서 측정된 동일 데이터 간의 차이는 각 측정방식의 mean 간의 차이와 거의 일치
- 점들이 x축을 따라 증가하거나 감소할 때
 - 두 방식 간에 비례 바이어스가 존재
 - 측정값의 크기에 따라 두 방식의 일치도가 차이가 남
 - 두 측정방법 간의 차이를 각 측정방식의 mean 간의 차이로 해석 할 수 없음
- Example)
 - 수은혈압계 vs 전자혈압계
 - 같은 학생에 대한 a와 b 심사위원의 점수

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



CODE and UTILs

- <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/>
- DATAtab Online Statistics Calculator
 - <https://datatab.net/statistics-calculator/descriptive-statistics>