탐색적 데이터 분석:
그래프에 의한 기술통계

Key words

#바차트 #파이차트 #히트맵
#스택트컬럼차트 #히스토그램
#박스플롯 #QQ플롯 #산점도

데이터 시각화 개요

그래프를 이용한 자료의 정리

자료의 유형에 맞는 그래프를 이용하여
한눈에 알아볼 수 있게 자료를 시각화할 수 있음.

질적 자료인 경우

- 1개 변수 : 바차트(막대그림), 파이차트
- 2개 변수 : 히트맵, 스택드컬럼차트

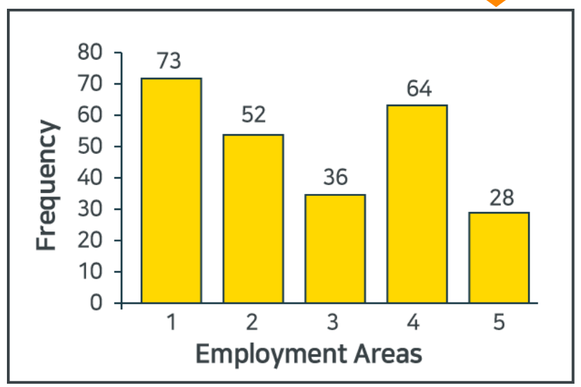
양적 자료인 경우

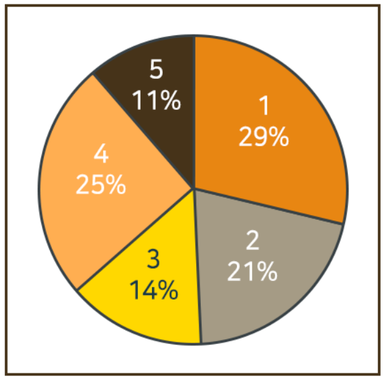
- 1개 변수 : 히스토그램, 박스플롯(상자그림), 라인차트, QQ플롯
- 2개 변수 : 산점도

바차트와 파이차트

Frequency and Relative Frequency Distributions

Area	Frequency	Relative Frequency
Accounting	73	28.9%
Finance	52	20.6%
General management	36	14.2%
Marketing/Sales	64	25.3%
Other	28	11.1%
Total	253	100%



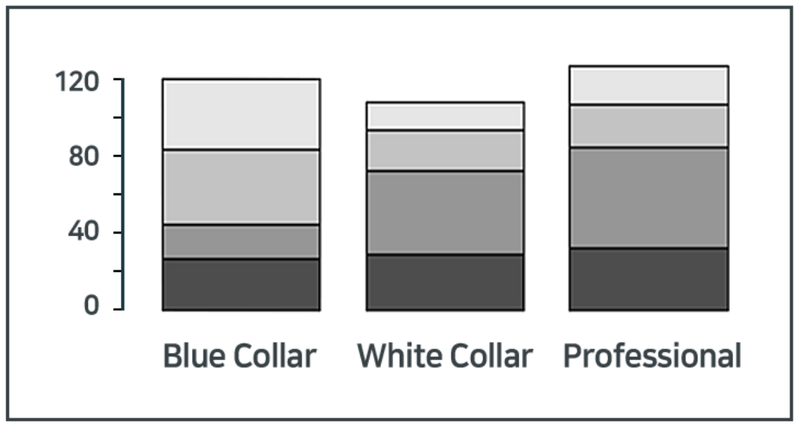


스택트컬럼차트

Contingency Table of Frequencies

Reader	Newspaper	Occupation
1	2	2
2	3	1
3	1	2
.	.	.
.	.	.
352	2	3
353	3	1
354	3	2

	Occupation			
Newspaper	Blue Collar	White Collar	Professional	Total
G&M	27	29	33	89
Post	18	43	51	112
Star	38	21	22	81
Sun	37	15	20	72
Total	120	108	126	354

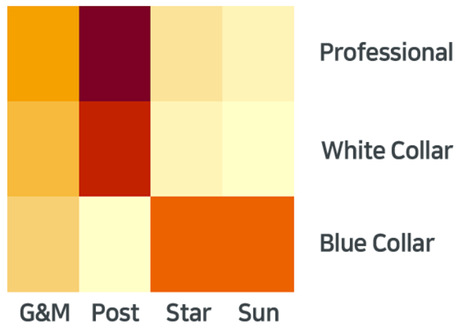


히트맵

Contingency Table of Frequencies

Reader	Newspaper	Occupation
1	2	2
2	3	1
3	1	2
.	.	.
.	.	.
352	2	3
353	3	1
354	3	2

	Occupation			
Newspaper	Blue Collar	White Collar	Professional	Total
G&M	27	29	33	89
Post	18	43	51	112
Star	38	21	22	81
Sun	37	15	20	72
Total	120	108	126	354

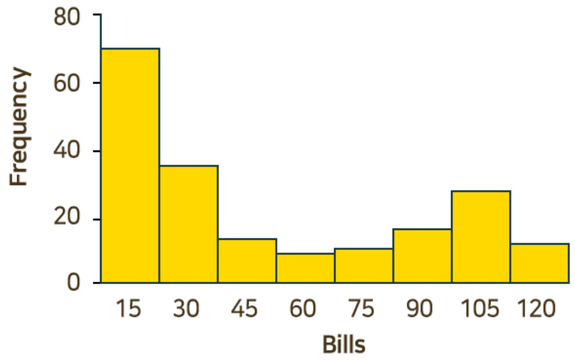


히스토그램

Subscribers	Bills
1	42.19
2	38.45
3	29.23
4	89.35
5	118.04
...	...

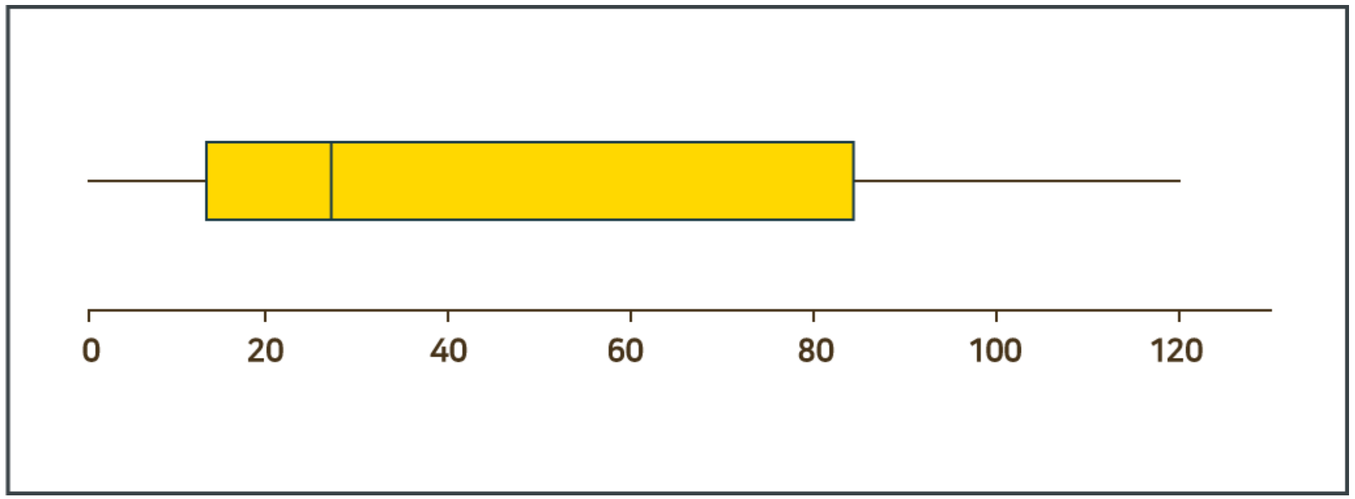
Frequency Distribution of the Long-Distance Bills

Class Limits	Frequency
0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200



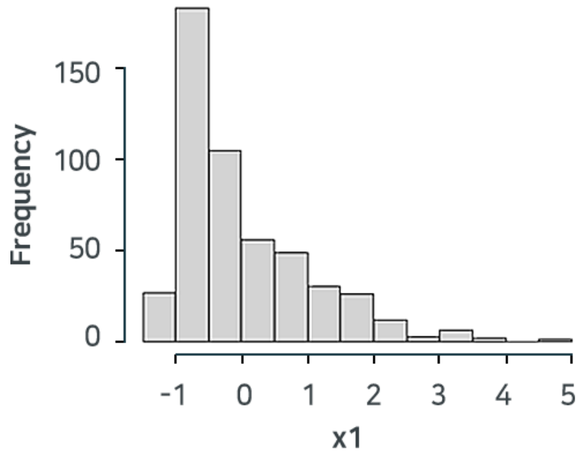
박스플롯

Subscribers	Bills
1	42.19
2	38.45
3	29.23
4	89.35
5	118.04
...	...

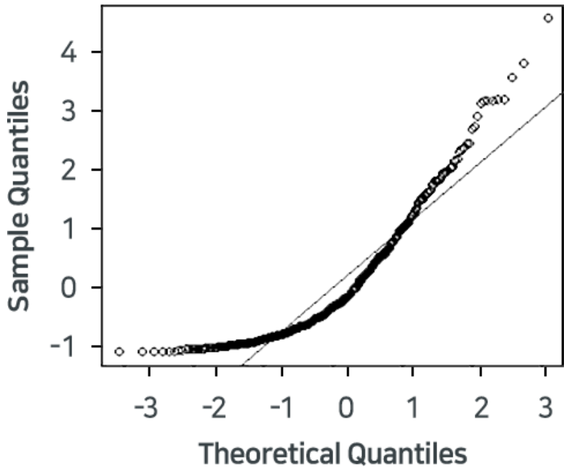


정규 QQ 플롯

Histogram of x1

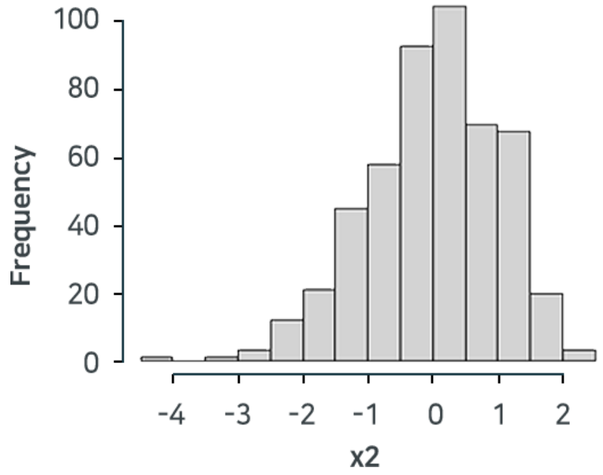


Normal Q-Q Plot

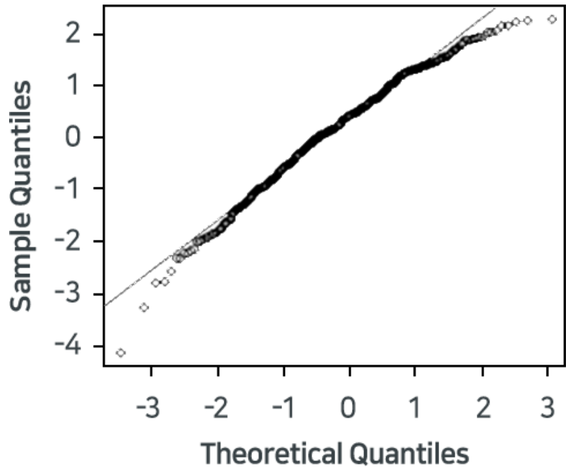


정규 QQ 플롯

Histogram of x2

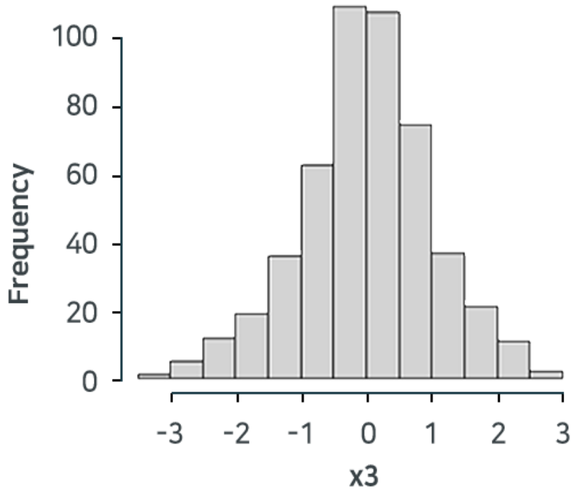


Normal Q-Q Plot

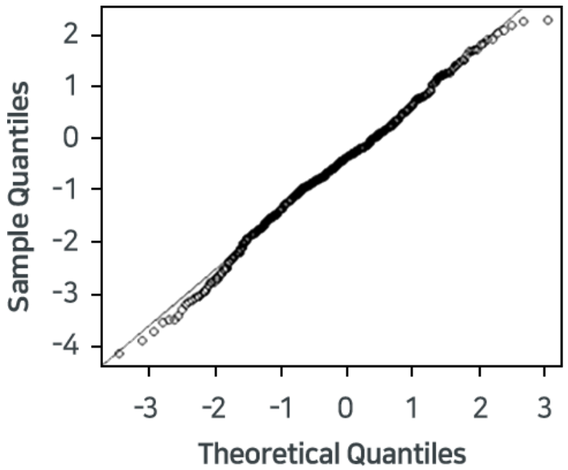


정규 QQ 플롯

Histogram of x3

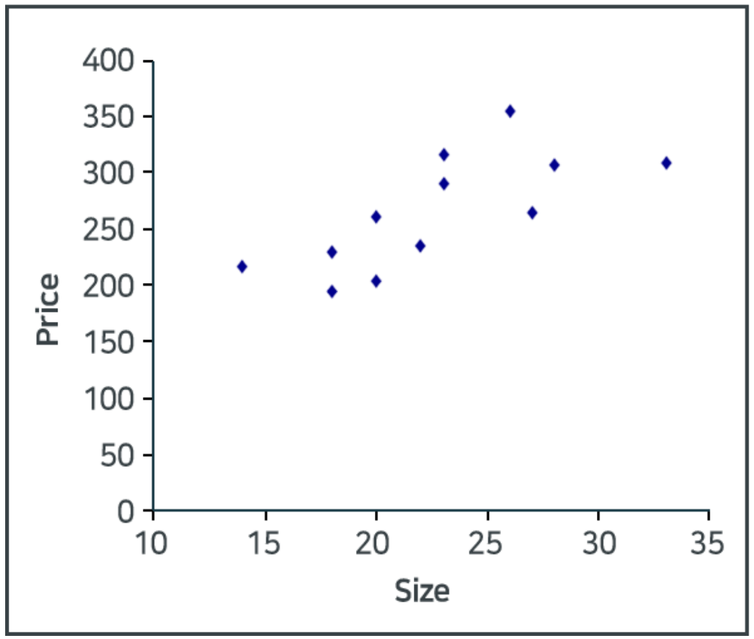


Normal Q-Q Plot



산점도

Size(100ft²)	Price($1,000)
23.54	315
18.07	229
26.37	355
20.54	261
22.41	234
14.89	216
33.77	308
28.25	306
23.02	289
20.68	204
27.15	265
18.33	195



탐색적 데이터 분석:
수치적 기술통계 ①
위치, 변이, 모양통계량

Key words

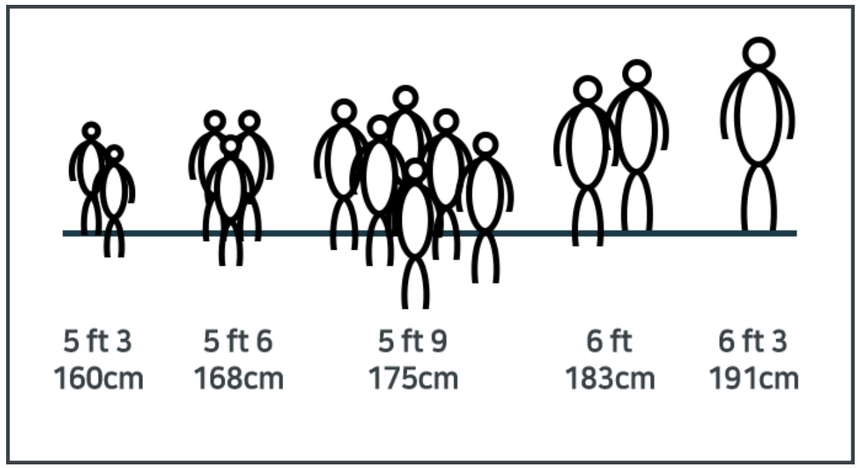
#평균 #중앙값 #중위수 #최빈값
#사분위수 #범위 #사분위간 범위(IQR)
#분산 #표준편차 #왜도 #첨도

중심 위치 척도

평균(Mean)

- 표본 자료 $x_1, \ldots, x_n$이 주어졌을 때,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



중심 위치 척도

중위수, 중앙값 (Median)

- 표본 자료 $x_1, \ldots, x_n$을 오름차순 정렬하였을 때,

$$x_{med} = \begin{cases} \left(\frac{n+1}{2}\right) \text{번째 자료} & , n \text{ 이 홀수} \\ \left(\frac{n}{2}\right) \text{번째와 } \left(\frac{n}{2}+1\right) \text{번째 자료의 평균} & , n \text{ 이 짝수} \end{cases}$$

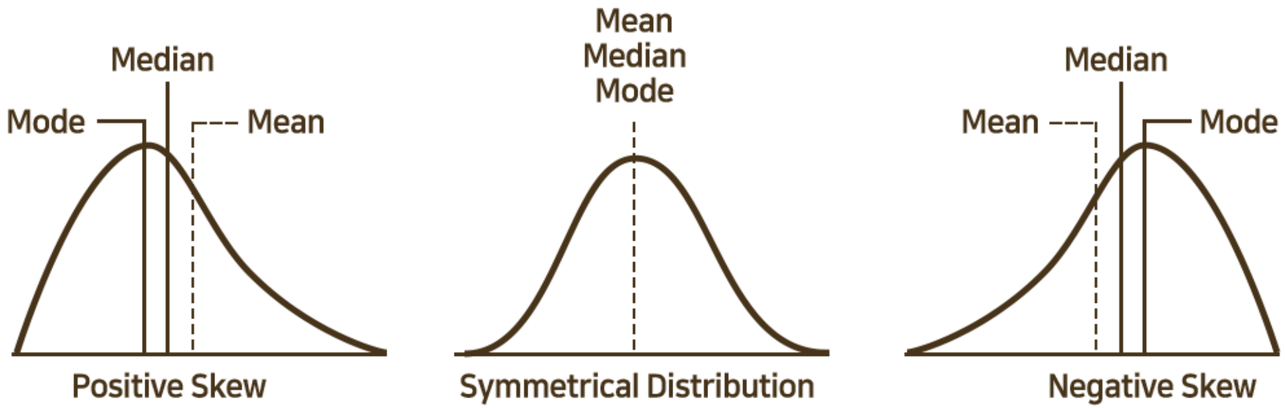
중심 위치 척도

최빈값 (Mode)

- 가장 빈도가 높은 값 또는 구간

중심 위치 척도

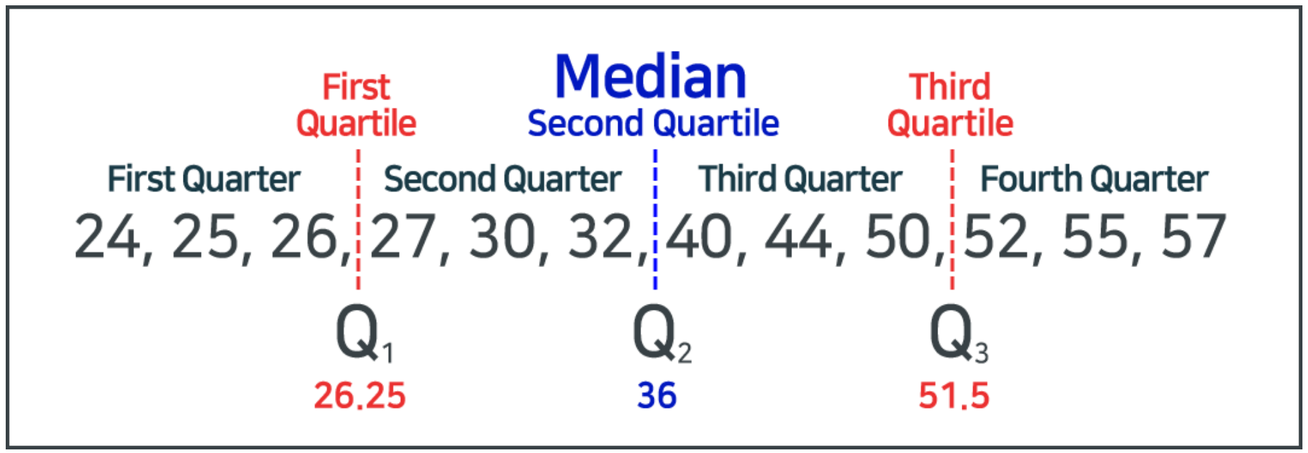
분포의 치우침 유형 별 중심위치 척도들 간의 관계



상대적 위치 척도

사분위수 (Quartile) : Q1, Q2, Q3

- Q1$\left(= x_{((n+1)/4)}\right)$: 25^{th} percentile, 1사분위수
- Q2$\left(= x_{((n+1)/2)}\right)$: median, 50^{th} percentile, 2사분위수
- Q3$\left(= x_{(3(n+1)/4)}\right)$: 75^{th} percentile, 3사분위수

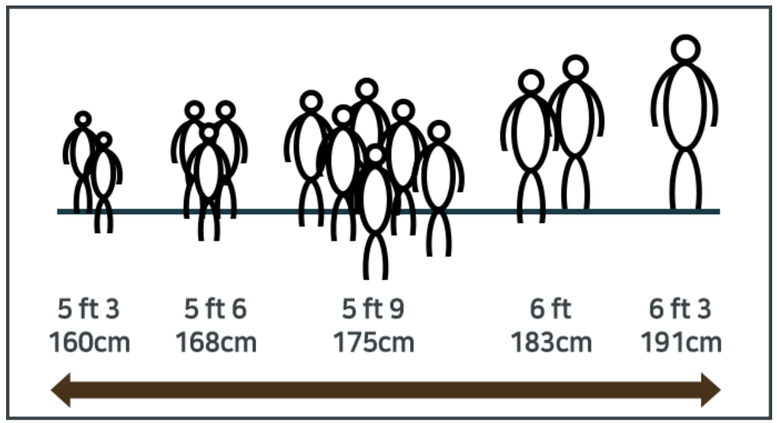


변동성 척도

범위 (Range)

- 표본 자료 $x_1, \ldots, x_n$이 주어졌을 때,

$$max(x_i) - min(x_i)$$



변동성 척도

사분위간 범위 (IQR, Inter Quartile Range)

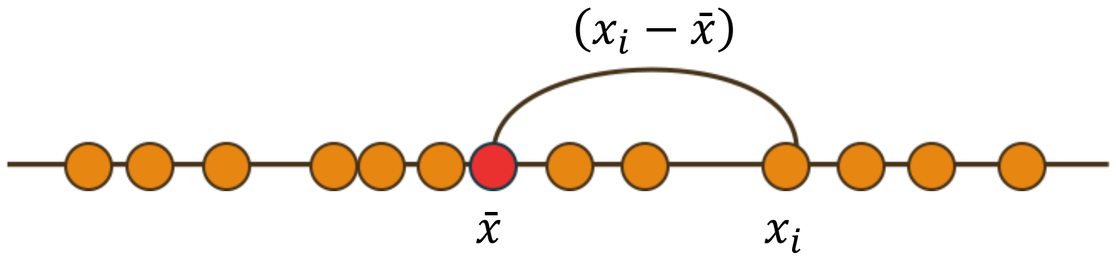
- Q3 – Q1

변동성 척도

표본 분산 (Sample Variance)

- 표본 자료 $x_1, \ldots, x_n$이 주어졌을 때,

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



변동성 척도

표본 표준편차 (Sample Standard Deviation)

- $s = \sqrt{s^2}$

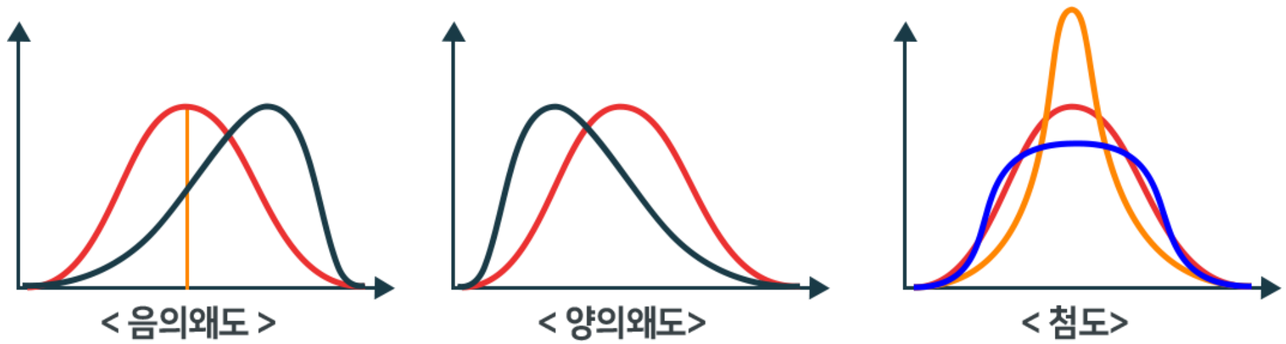
변동계수 (Coefficient of Variation)

- $cv = s/\bar{x}$

형태 척도

분포의 형태에 관한 척도

- 왜도: 분포의 비대칭 정도를 나타내는 척도.
- 첨도: 분포의 중심에서 뾰족함 정도를 나타내는 척도.



탐색적 데이터 분석:
수치적 기술통계 ②
연관성

Key words

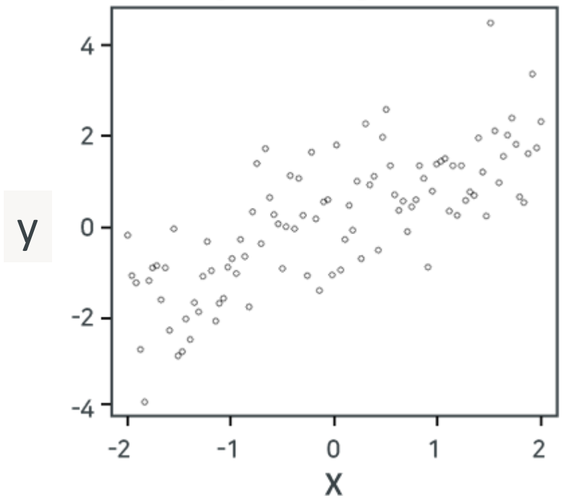
#공분산 #피어슨 상관계수
#스피어만 상관계수 #켄달의 상관계수

선형적 연관성

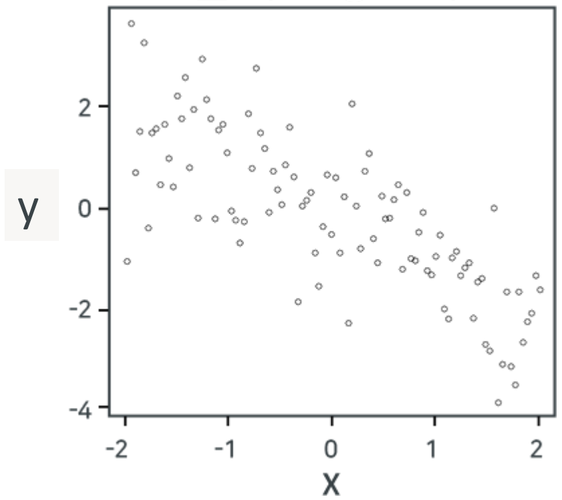
두 숫자형 변수의 선형적 연관성

- 선형적 연관성의 방향

< 양의 선형관계 >



< 음의 선형관계 >

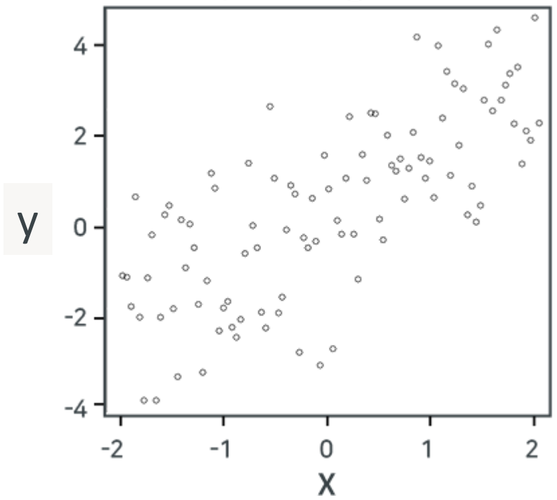


선형적 연관성

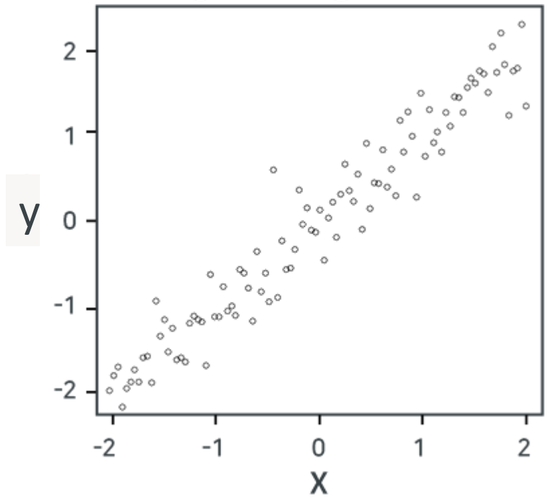
두 숫자형 변수의 선형적 연관성

- 선형적 연관성의 강도

< 약한 선형관계 >



< 강한 선형관계 >



선형적 연관성 척도

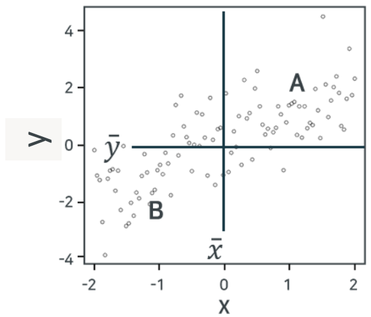
표본 공분산 (Sample Covariance)

- n쌍의 표본 자료 $(x_1, y_1), \ldots, (x_n, y_n)$이 주어졌을 때,

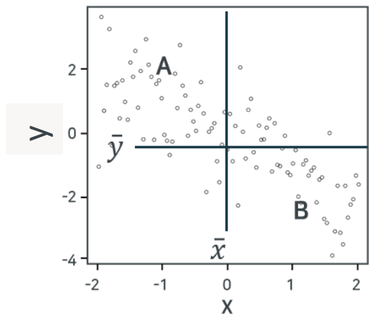
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- 선형관계의 방향

1) $s_{xy} > 0$: 양의 선형 관계, 비례관계
2) $s_{xy} < 0$: 음의 선형 관계, 반비례관계



< 양의 선형관계 >



< 음의 선형관계 >

선형적 연관성 척도

표본 공분산 (Sample Covariance)

- 선형관계의 강도

  $-s_x s_y \leq s_{xy} \leq s_x s_y$ by Cauchy-Schwarz 부등식

- 표본 공분산은 x와 y 의 측정 단위에 의존하는 지표임.

  $x' = ax + b$ 이고 $y' = cy + d$인 경우에, $s_{x'y'} = ac \cdot s_{xy}$

선형적 연관성 척도

표본 상관계수 (Sample Correlation, 피어슨의 상관계수)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- $-1 \leq r_{xy} \leq 1$, by Cauchy-Schwarz 부등식

선형적 연관성 척도

표본 상관계수 (Sample Correlation, 피어슨의 상관계수)

- 선형관계의 방향

  1) $r_{xy} > 0$: 양의 선형관계
  2) $r_{xy} < 0$: 음의 선형관계

- 선형관계의 강도

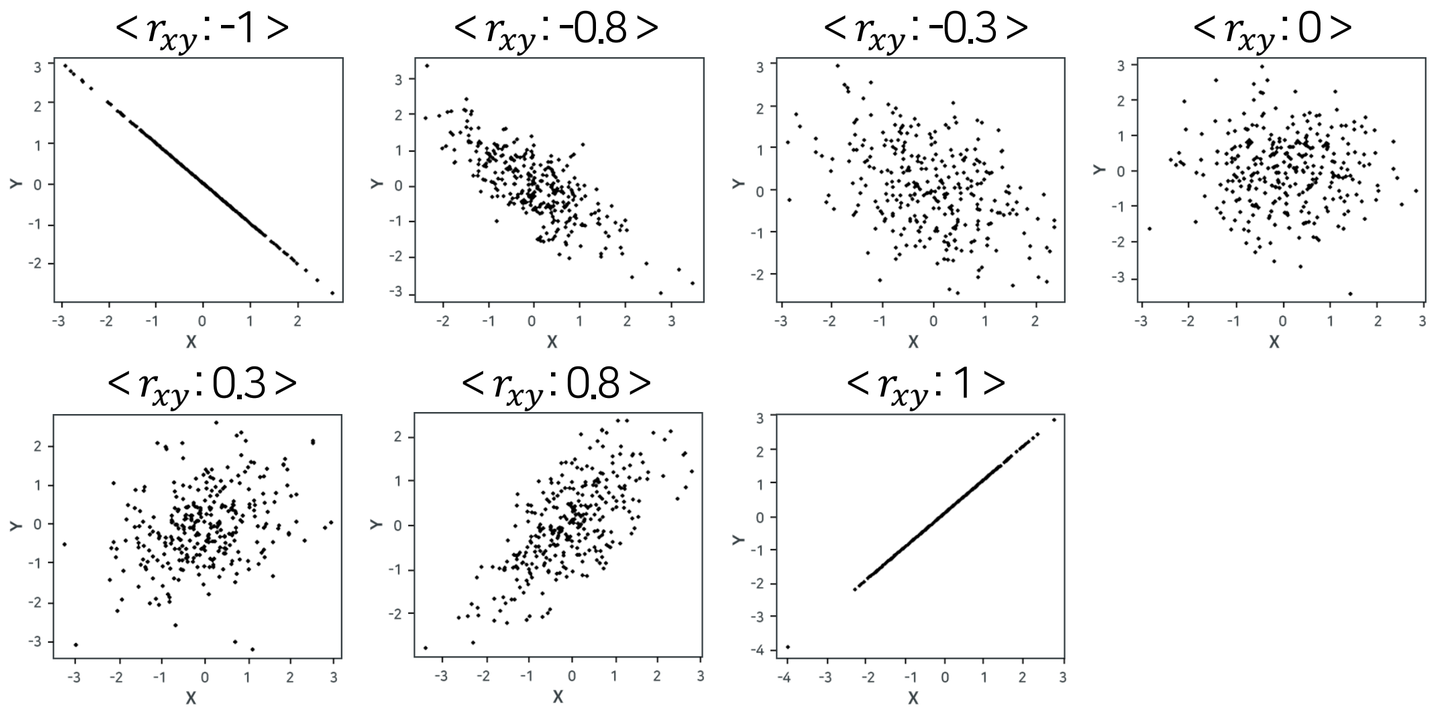
  1) $|r_{xy}| \approx 0$: 강도가 약함.
  2) $|r_{xy}| \approx 1$: 강도가 강함.

- 표본 상관계수는 x와 y 의 측정 단위에 의존하지 않음.

  - $x' = ax + b,\ y' = cy + d$이고, $ac > 0$ 인 경우에, $r_{x'y'} = r_{xy}$

선형적 연관성 척도

표본 상관계수 (Sample Correlation, 피어슨의 상관계수)



선형적 연관성 척도

순위를 이용한 상관계수

- 서열 척도이거나, 정규분포를 심하게 벗어나는
  두 숫자형 변수의 연관성 파악.

- 스피어만 상관계수 (Spearman's correlation coefficient)
  - 원 자료값의 순위를 구한 뒤, 순위에 대하여 피어슨의 상관계수를 구한 것.
  - -1에서 1 사이의 값을 가지며, 절대값이 클수록 강한 상관관계를 나타냄.

선형적 연관성 척도

순위를 이용한 상관계수

- 켄달 상관계수(Kendall rank correlation coefficient)
  - 두 변수 순위의 일치 정도를 측정.
  - 한 변수의 순위가 증가할 때 다른 변수의 순위도 함께 증가하는 경우가 그렇지 않은 경우에 비해 얼마나 큰지를 측정하는 방식.