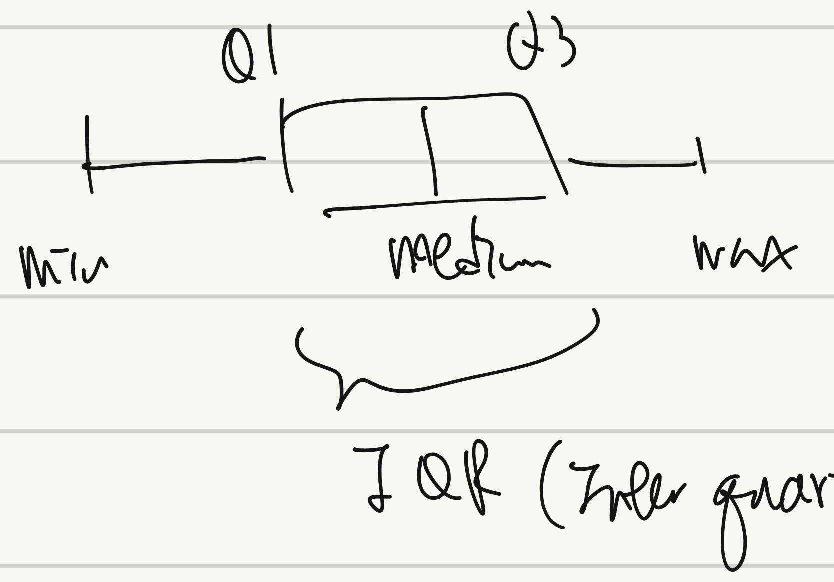
DAY 2 30 MAR 2021

ㅇ 5숫자 요약



IQR = Q3 - Q1

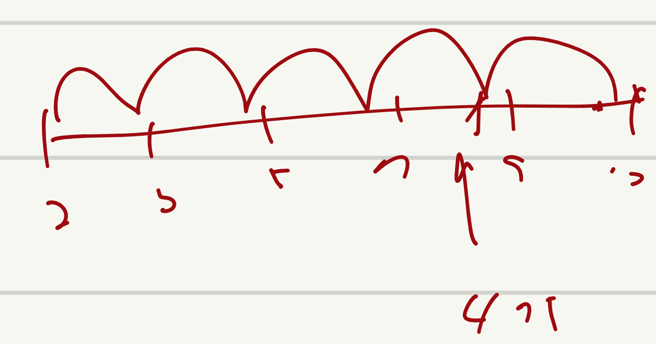
Outlier : Q3 + 1.5 IQR.

(e.g.) df ← c(2, 3, 5, 7, 9, 10)

- median : 6
- Q1 : ~~2.5~~ 3.5
- Q3 : ~~9.5~~ 8.5
- IQR : ~~7~~ 5

R에서 사용하는 IQR 계산법 (type : 7)

median (Q2) = (5+7)/2 = 6



[Q1] 5개의 구간에서 0.25는 얼마인가.

=> 5×0.25 = 1.25 +1 (data point 시작)

= 2.25

=> 3 + (5-3) × 0.25 = 3.5

[Q3] 5개의 구간에서 0.75

=> 5 × 0.75 = 3.75 + 1

= 4.75

=> 7 + (9-7) × 0.75 = 7 + 1.5 = 8.5

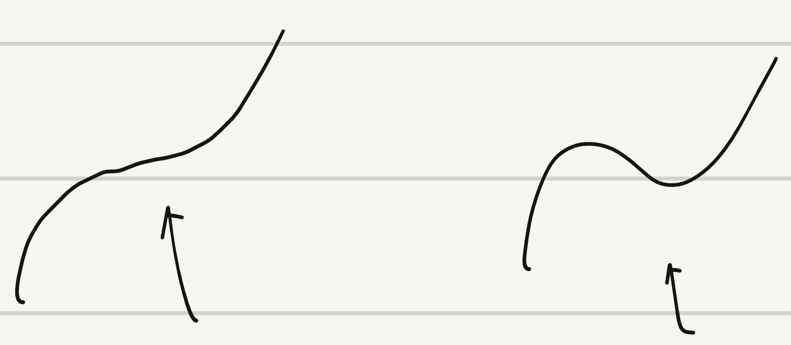
* 표준편차, 범위, 사분위 비교
  - 중심 측정값 부터 Sample mean 을 사용 → Standard deviation 사용, 자료가 정규분포
  - 중위 측정값 부터 median 을 사용 → IQR 사용.

* 두 변수간의 관계 분석.

- 분할표 : 두 데이터가 discrete 할 때 가능.

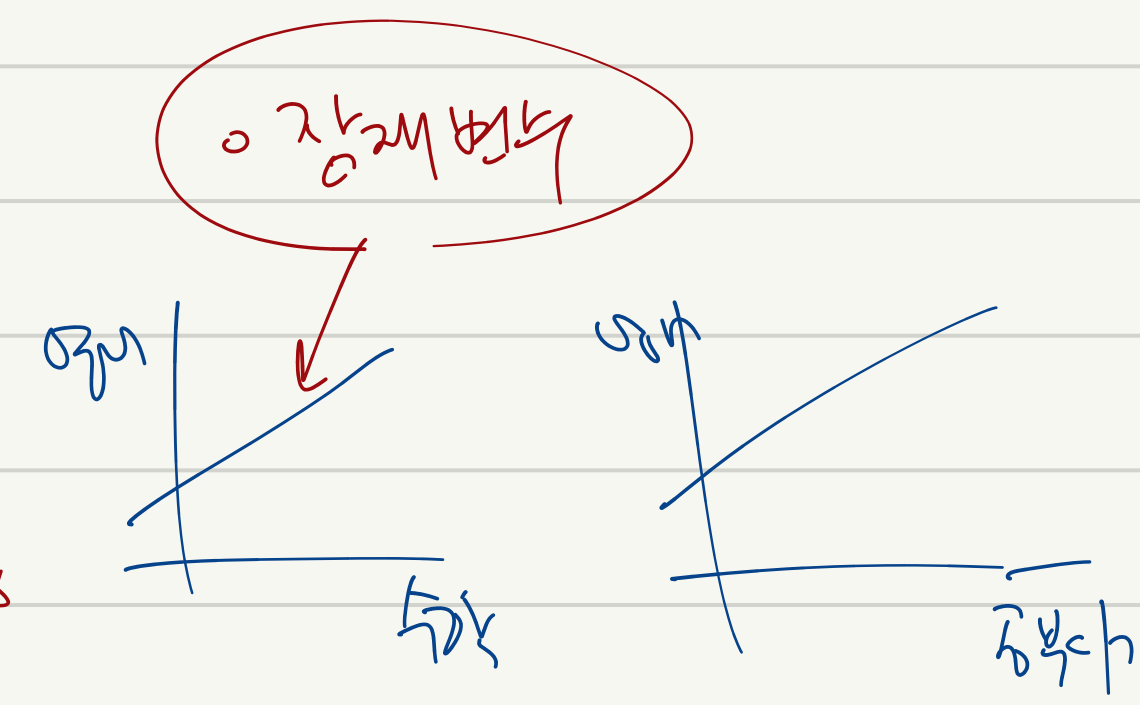
* 데이터가 이산 vs. 연속 부정확할 때는 실험으로 확인

- 상관 분석 (correlation)
  ① 두 연속형 변수 간 선형적 연결 관계가 있는지 분석
  ② 상관 계수 (correlation coefficient)는 선형적 관계의 강도
  ③ 모집단 상관 계수는 ρ, 표본집단의 상관계수는 r로 표기.



: 최소한 강도는 x

* 인과관계는 확인할 수 없다



연관관계 → correlation

인과관계 → regression

- Pearson's r

$$\mu \leftarrow \hat{\mu} \leftarrow \bar{x}$$
$$\sigma^2 \leftarrow \hat{\sigma}^2 \leftarrow s^2$$
$$\rho \leftarrow \hat{\rho} \leftarrow r$$

⇒ 두 연속형 변수가 정규분포를 따를 때.
⇒ r 값이 $+$: 양의 상관관계.
⇒ $-1 < r < 1$
⇒ $0.7 < |r| < 0.9$: 높은 상관 관계
$0.4 < |r| < 0.7$: 중증도 상관관계
} depending on domains

$$r_p = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}, \quad Cov(X,Y) = E(X-\mu_x)E(Y-\mu_y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}$$

※ 확률

- 사건의 확률
  - 표본 공간 (sample space) Ω : 한 실험에서 나올 수 있는 결과들의 모임 $\{1, 2, 3, 4, 5, 6\}$
  - 근원 사건 (Elementary outcomes) $\omega_1, \omega_2, \ldots$: 표본 공간을 구성하는 개개의 결과.
  - 사건 (event) A, B : 표본 공간의 부분집합 $\{2, 4, 6\}$, $\{5, 6\}$
  - 동일한 실험을 무한히 반복했을 때, 나타나는 사건의 상대도수 의 ← 확률.

(ex) $\Omega = \{0, \ldots, 143\}$ (아무도 위반 x ... 모두 위반)

Q. 143에 10% 미만 사건 A = ?

- 확률의 법칙
  - 모든 A 에 대해 $0 \le P(A) \le 1$
  - $P(A) = \sum_{\omega \in A} P(\omega)$
  - $P(\Omega) = \sum_{\omega_i \in \Omega} P(\omega_i) = 1$

- 확률의 계산
  - $P(A^c) = 1 - P(A)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - $P((A \cup B)^c) = P(A^c \cap B^c)$
  - $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, $P(B) > 0$
  - 두 사건 A, B 가 독립 → $P(A \cap B) = P(A)P(B)$ ← joint probability 결합확률

드모르간의 법칙

$(A \cup B)^c = A^c \cap B^c$

marginal probability 주변확률
(관심 아닌 것들 제외)
모델링

※ 확률 변수 (Random variable)

= 실험의 결과가 확률적으로 나타나는 실험에서 얻어진 결과는 실수값에 대응된 결과

※ 범주(이산)형 확률변수와 확률분포.

- Probability distribution : 확률 변수가 갖는 값들과 그에 대응하는 값.
- 이산확률변수 X의 확률질량함수 P(x)는
  - 모든 x_i 에 대해 $0 \le P(x_i) \le 1$
  - $\sum_{x_i} P(x_i) = 1$

nCr

(ex)

$x=0$	1	2	3	4
⋮	⋮	⋮	⋮	⋮

$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$

이항분포.

도수분포표로부터 모분산 구하기.

※ how to calculate $_nC_x = \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$

★ 유의 수준과 유의 확률의 정의와 개념.

① 귀무가설과 대립가설 (Null Hypothesis, H0) & (Alternative hypothesis, H1)

- H_0 : Not H_1.
- H_1 : 연구자가 연구를 통해 입증하고자 하는 주장.

$P\left(\text{우리의 실험 나온 확률} \mid P_0 = \frac{1}{2}\right) < 0.05$ 이면 $P_0 = \frac{1}{2}$이 틀렸다.

유의 확률 　 귀무가설 　 유의수준 (5%로 하기 !!)

n 수를 늘리면 가설을 기각 할 수도 있다.

⇒ 유의 확률이 유의수준보다 낮기 때문에 우리의 가정이 틀렸다.

$$P\left(x=1 \mid P_0 = \frac{1}{2}\right) = \binom{4}{1} 0.5^1 (1-0.5)^3$$
$$= 4 \times 0.5 \times 0.5^3$$
$$= 4 \times 0.5^4$$
$$= \frac{1}{4}$$

- 유의수준 = α value = 기각역
  = 1종 오류
- 유의 확률 = p-value

대립가설을 세우고 → H_0 (귀무가설) 가설 생성 → 귀무가설 검정