양적자료 – 수치로 표현할 수 있음 ex)성적, 기온

질적자료 – 수치로 표현할 수 없음 ex)혈액형

로버스트 – 극단값에 영향을 적게받으면 로버스트한 놈임. 비교적인 개념.

대푯값

산술평균 – 일반적인 평균. 양적자료에만 활용가능. 극단값에 매우 민감하게 작용함.

중앙값 – n이 홀수개인 경우 (n+1)/2번째 값　　n이 짝수개인 경우 n/2와 (n+1)/2과의 평균

최빈값 – 이거 모르면 때려치우자

사분위수(25, 50, 75) – 변량X의 n개의 자료에 대해 n*0.25, n*0.5, n*0.75에 해당되는 순서의 값.

각 값이 소수점으로 떨어지면 +1을 해줌 ex) n*0.25=3.21일 경우 25%백분위수는 4번째 값.

x%절사평균 – 자료의 양 끝에서 x%만큼을 제외하고 구한 평균.

산포도(상위개념) – 변량이 흩어져 있는 정도를 나타내는 것.

범위 – 최댓값 - 최솟값

사분위수 범위 IQR (Q3-Q1) – 자료 집합의 중간 50%에 해당되는 자료의 산포도를 나타냅니다.

분산 – 자료들이 퍼져있는 정도를 나타냄. (편차^2의 평균)

표준편차 – 분산의 양의 제곱근. 분산의 단위가 자료 측정단위의 제곱이기 때문에 실 자료 측정에는 양의 제곱근인 표준편차를 사용한다.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

〈 모분산 〉

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

〈 표본분산 〉

(모표준편차, 표본표준편차)

표본은 모집단의 전체를 대표하지 않기 때문에 모분산을 과소추정하는 경향이 있습니다. 이를 보정하기 위하여 표본분산을 계산할 때는 n-1로 나누어 계산합니다. 그러나 모집단이 매우 많고 표본이 모집

단의 개수에 근접할만큼 많다면 이 보정법은 큰 의미를 가지지 못합니다.

표본분산 = 모분산/(n-1)

변동계수 – 평균이 큰 차이로 차이나거나 측정기준이 다른 경우 사용함.

평균을 중심으로 상대적으로 흩어진 정도를 나타냄.

CV= 표준편차/평균*100%

상자그림(box plot)

최댓값, 최솟값, 중앙값(Me), 사분위수 Q1, Q2, Q3를 사용하여 그린다.

안 울타리 1.5(Q3-Q1)

바깥 울타리 3(Q3-Q1)

각 울타리 인접값 표시.

챕터2 확률

2.1 표본공간과 사건

표본공간 – 어떤 시행에서 일어날 수 있는 모든 결과의 집합.

사건 – 표본공간의 임의의 부분집합

근원사건 – 한 개의 원소만으로 이루어진 사건.

전사건 – 표본공간의 모든 원소를 포함하는 사건. (사건 전체)

표본점 – 표본공간을 구성하는 개개의 원소

공사건 – 표본점을 하나도 포함하지 않는 사건.

이산 표본공간 – 표본공간의 원소 개수가 유한하거나 셀 수 있음 (생산 불량품 개수).

연속 표본공간 – 수학적으로 셀 수 없음 (전구의 평균 수명)

합사건 – A, B중 적어도 한 사건이 일어나는 사건

곱사건 – A, B가 동시에 일어나는 사건

여사건 – 사건 A가 일어나지 않는 사건 A^c

배반사건 – 두 사건이 절대 동시에 일어나지 않을 때 서로 배반임. P(A교B)=공집합

수학적 확률 – 각 근원사건의 확률이 같을 때

통계적 확률 – (lim무한대) 시행을 극한으로 보내면 확률은 수렴함.

확률의 성질 - P(A ∪ B) = P(A) + P(B) – P(A ∩ B),

P(A ∩ B) = P(A) + P(B) - P(A ∪ B)

판별식 D = b^2-4ac

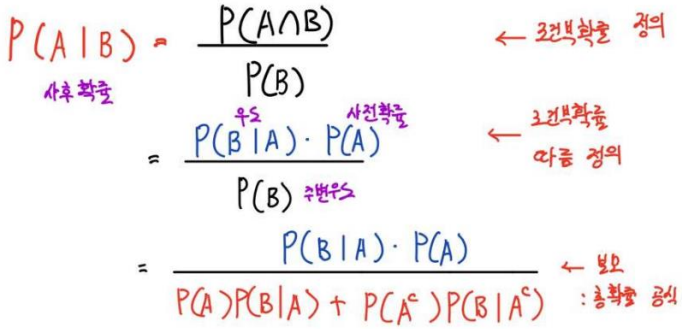
조건부 확률 – 사건 B가 일어났다는 조건 하에서 사건 A가 일어날 확률 P(A|B)

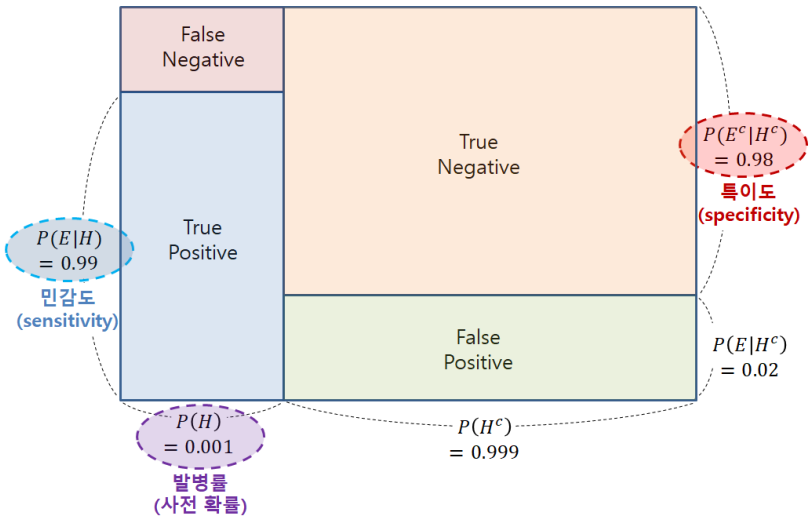
P(A|B) = P(A∩B) / P(B)

독립사건 – 두 사건 A, B에서 P(A ∩ B) = P(A)*P(B) 를 만족하면 독립. 아니면 종속임.

사전확률 = P(A) , 사후확률 = P(A | B). on 베이즈 정리

기초통계학 - 확률3 : 베이즈 정리(Bayes' theorem)

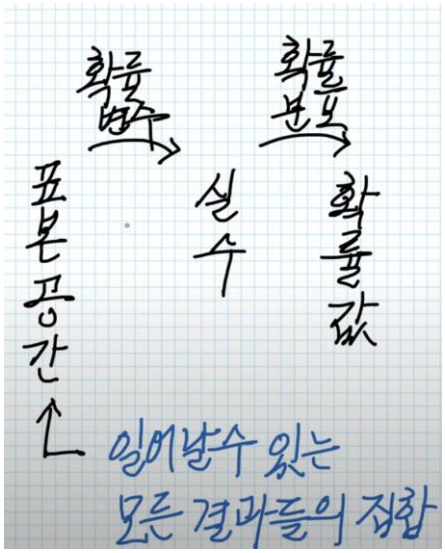




확률변수와 확률분포

확률변수 – 특정 시행에서 표본공간의 원소를 넣으면 하나의 실숫값을 뱉어내는 **함수**.

확률변수 예시) 동전을 두개 던졌을 때 앞면이 나올 횟수. 0 1 2



이산확률변수 – 확률변수의 치역이 셀 수 있을 경우. Ex)주사위의 눈

확률질량함수 – 이산확률변수 X에 대하여 X가 임의의 실수 a를 취할 확률에 대응하는 함수.

연속확률변수 – 확률변수의 치역이 연속적이라서 셀 수 없는 경우. Ex)사람 키

확률밀도함수 – 확률변수의 분포를 나타내는 함수. 확률의 존재를 그래프로 나타낼 수 있음

연속확률분포의 누적분포함수 –> 확률밀도함수를 적분하면 나옴.

즉 확률밀도함수의 미분 = 누적분포함수.

이산확률분포의 분포함수 – 누적확률분포함수 F(x) = P(X<=x) 즉 F(x) = P(x) – F(x-1)

확률분포 – 각 확률변수가 특정 값일 확률. Ex) P(X=2)일 확률.

확률분포표 – 확률변수와 확률분포를 표처럼 표기한 것.

확률변수의 기댓값과 분산

일반적인 경우 E(X) = x*f(x)

분산 = Var(X) = E(X^2) – E(X)^2

베르누이분포 B(1,p), E(X)=p, Var(X)=pq

이항분포 B(n,p), 기댓값 = np, 분산=npq

정규분포 N(평균,분산)에서 표본n개를 뽑으면 정규분포N(평균, 분산/n)을 따름.

공분산(Cov) = 두 확률변수 X, Y가 같은 방향으로 변화하는지의 척도.

Cov(X, Y) = E(XY) – E(X)E(Y)

Cov(aX+b , cY+d) = acCov(X,Y)

상관계수(Corr) = 공분산을 사용하여 두 자료를 비교하기 곤란할 때 두 자료를 비교하기 위한 척도.

Corr(X, Y) = Cov(X, Y) / X표준편차 Y표준편차

확률변수의 독립

E(XY) = E(X)E(Y) 이면 독립이다.

Z 검정

표본 평균이 모집단 평균과 얼마나 다른지를 검정하는 데 사용됩니다. 이 검정은 주로 표본 크기가 크고 모집단의 표준 편차를 알고 있을 때 사용됩니다. Z-검정은 **정규 분포를 기반**으로 하며, **표본 평균과 모집단 평균의 차이를 표준 편차로 나눈 Z-값**을 사용하여 가설을 평가합니다.

1. **귀무가설 (H0):** 모집단 평균과 표본 평균 간에는 유의한 차이가 없다.

2. **대립가설 (H1 또는 H_a):** 모집단 평균과 표본 평균 간에는 유의한 차이가 있다.

Z-검정은 주로 평균에 대한 가설 검정에 활용되며, 특히 큰 표본 크기에서는 중심극한정리에 의해 정규 분포를 따르기 때문에 유용하게 적용될 수 있습니다.

T 검정

t-검정은 통계학에서 두 집단 간의 평균 차이를 비교하는 데 사용되는 가설 검정 방법 중 하나입니다. 특히, 표본 크기가 작거나 모집단의 표준 편차를 모를 때 주로 사용됩니다.

t-검정은 일반적으로 다음과 같은 두 가지 가설을 검정하는 데 사용됩니다:

1. **귀무가설 (H0):** 두 집단 간에는 평균 차이가 없다.
2. **대립가설 (H1 또는 H_a):** 두 집단 간에는 평균 차이가 있다.

1. **독립 t-검정 (Independent t-test): 두 독립적인 표본**간의 평균 차이를 비교합니다. 예를 들어, 어떤 처리를 받은 그룹 A 와 그렇지 않은 그룹 B 간의 차이를 검정할 때 사용됩니다.
2. **대응 t-검정 (Paired t-test):** 동일한 A 그룹의 서로 다른 조건에서 얻은 두 관측값 간의 차이를 비교합니다. 예를 들어, **동일한 표본**에서 전후의 측정값을 비교하여 어떤 처리의 효과를 검정할 때 사용됩니다.

t-검정에서는 t-값을 계산하고, 이를 t-분포의 t-표에 비교하여 유의수준에 따른 임계치를 확인합니다. t-값이 임계치를 넘으면 귀무가설을 기각하게 되며, 이는 두 집단 간에 통계적으로 유의미한 평균 차이가 있다는 증거로 해석됩니다.

1. **제1종 오류 (Type I Error):**

   - **정의:** 귀무가설이 참인데도 불구하고, 귀무가설을 기각하는 오류입니다.

(내말이 맞다고!! 시발 맞다니까? -> 대립가설 채택)

2. **제2종 오류 (Type II Error):**

   - **정의:** 귀무가설이 거짓인데도 불구하고, 귀무가설을 채택하는 오류입니다.

(내 말이 맞나? 아닌것 같은데? 그럼 그냥 귀무가설 채택)

제1종 오류와 제2종 오류는 서로 반비례 관계에 있습니다. 즉, 하나를 감소시키면 다른 하나의 확률이 증가하게 됩니다.