통계학

모집단 : 연구의 대상이 되는 모든 개체들을 모은 집합 ( 모든 대상을 분석하는 것 불가능 )

표본 : 모집단의 일부분의 관측값들

모수(모집단의 요약값) : 수치로 표현되는 모집단의 특성

통계량 : 표본의 관측값들에 의해서 결정되는 양

* 통계량을 통해 모수를 알아내는 것이 통계학의 목표
* 자료가 달라지면 행하는 분석이 달라진다

수치형( 연속형 / 이산형 )자료 | 범주형( 순위형 / 명목형 )자료

중심 경향값(대표값)

* 평균(Mean)
* 중앙값(Median) : 크기순으로 정렬시켜 중앙에 위치한 값
* 최빈값(Mode) : 가장 자주 나오는 값

산포도 (퍼진정도)

* 분산 : 분산이 작으면 위로 솟은 모양, 분산이 크면 낮아지는 모양
* 사분위수 범위

정규분포

* 위치는 평균에 의해, 모양은 분산에 의해 결정

분포도

* 왜도(Skewness)

분포의 비대칭 정도 ( 꼬리가 긴쪽의 skewness ) 왼쪽이 길면 left-skewed

* 첨도

분포의 꼬리 부분의 비중에 대한 측도

K = 0 뾰족한 정도가 정규분포와 동일

표본평균, 표본분산 ( 모집단의 파라미터를 알기위해 표본에서부터 얻은 추정량 )

표본분산을 구할 때 표본평균을 이용한다. 모집단의 평균을 모르기 때문에. ?? 분모가 n-1인 이유는?

< 확률의 기초 >

주사위를 던진다.

표본공간(sample space) : S = {1,2,3,4,5,6}

근원사건(sample outcom) 표본공간의 원소 : 1,2,3,4,5,6

사건(event) : 짝수가 나오는 사건 A = {2,4,6} p(A) = 1/2

조건부확률

* P(A|B) = P(A N B)/P(B)

독립

* P(A|B) = P(A), P(B|A) = P(B) => P(A N B) = P(A)P(B)

확률변수의 기대값(평균)

* 확률변수의 중심경향값. 흔히 평균이라 칭함( 각각이 나올확률변수와 확률을 곱하여 총 더한다) [ E(X) ]

확률변수의 분산

* Var(X) = E(X-u)^2

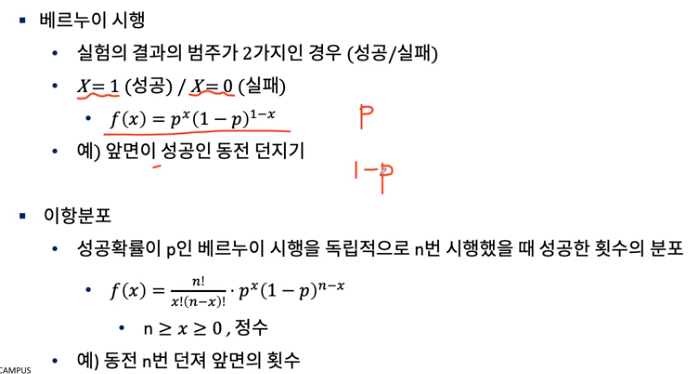
공분산

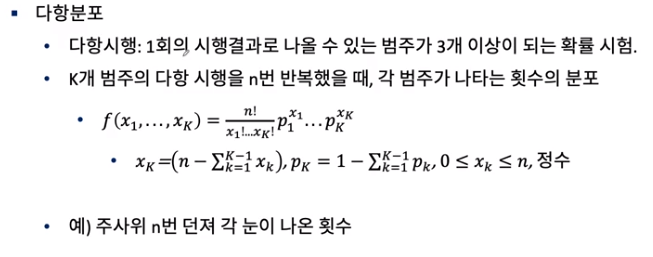
* 변수가 두개가 있을 때, 이 둘의 관계가 어떤가를 나타낸 축도
* Cov(X,Y) = E(X-ux)(Y-uy)
* X,Y가 독립이면 Cov(X,Y)=0

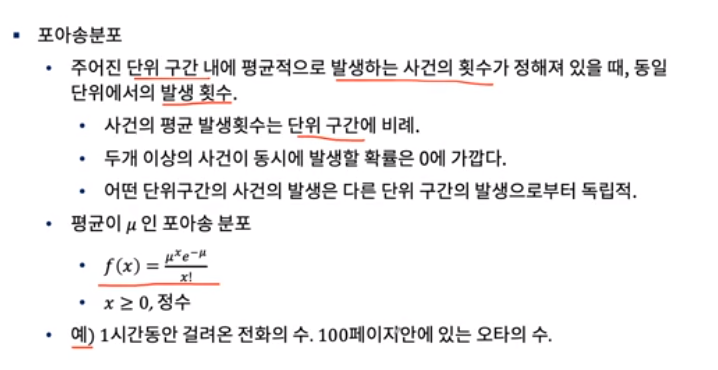
상관계수(공분산의 단점 크다 작다를 못나타낸다, 따라서 상관계수를 사용하여 공분산을 단위화)

* P = Cov(X,Y)/root(Var(X)Var(Y)), -1<= p <= 1

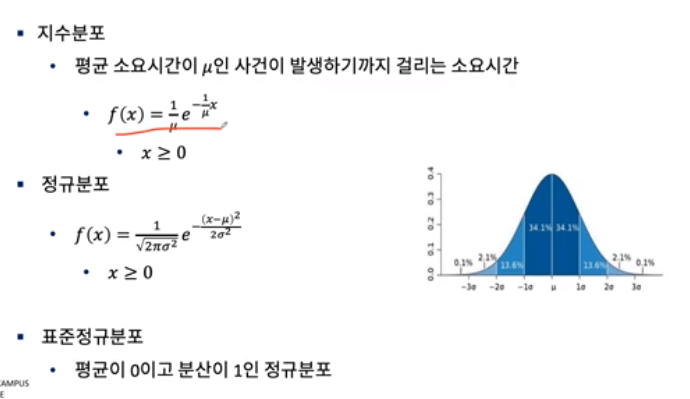
< 이산형 확률분포 >







< 연속형 확률분포 >

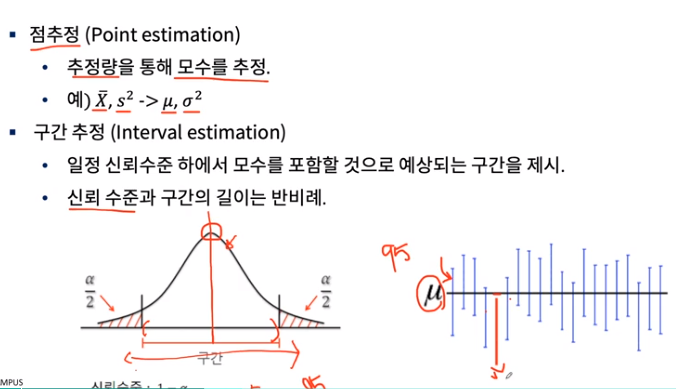


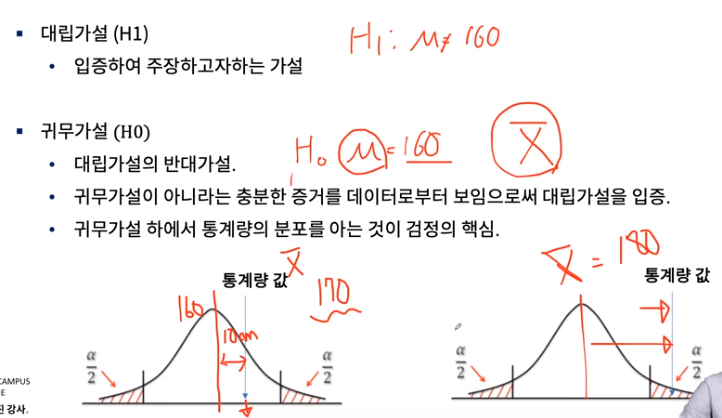
* 정규분포 1시그마 : 68.4% 2(1.96배)시그마 : 약 95%

< 통계적 추론 >

추론 : 모집단에 대한 정보를 알고싶은데 알수가없다. 표본으로부터 모수를 알아내는 과정이 추론

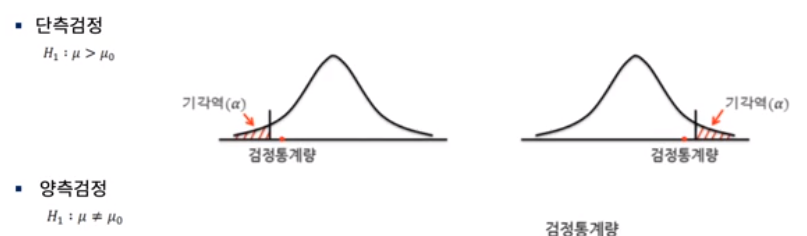
통계량값을 이용한다



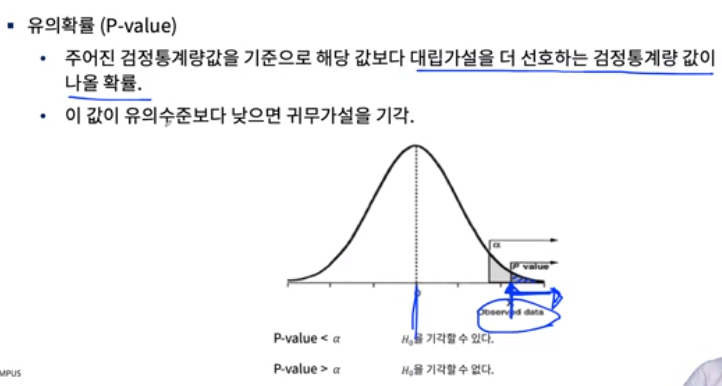


 1종오류를 최대한 지양한다

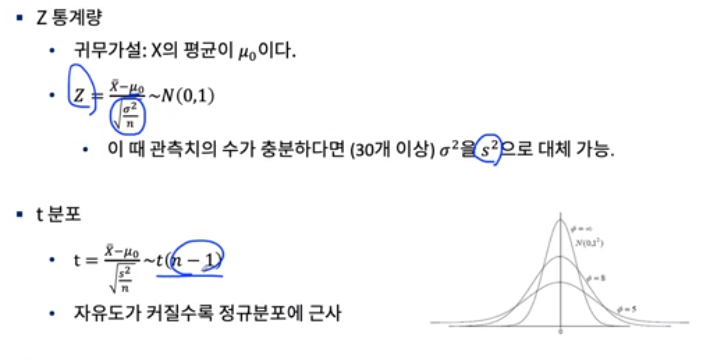
Ex) 키가 160이 평균이 아니다 라고 했을 때 실제 160이라고 하는(오류)를 지양한다



단측검정으로 양의 방향으로 선택하면 160보다 작은 값이 크게 차이가 나더라도 기각에 영향을 끼치지 않는다



< 검정통계량의 분포 >



관측지가 적으면 t분포 사용 ( 커질수록 정규분포에 가까워진다 )

N(0,1) : 표준정규분포를 뜻한다

+ 카이제곱분포 , F분포

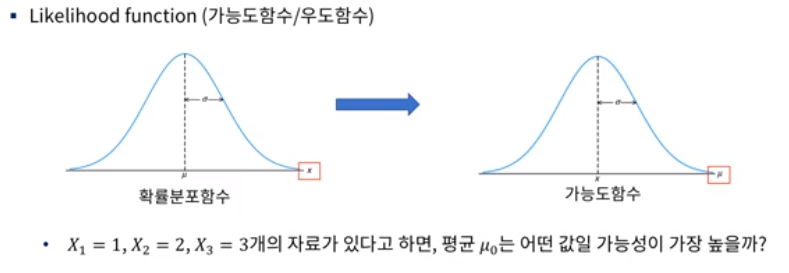
< Likelihood >

가능도(우도)

같은식을 변수를 다른 관점에서 보느냐에 따라 확률분포함수 or 가능도함수로 나타내진다

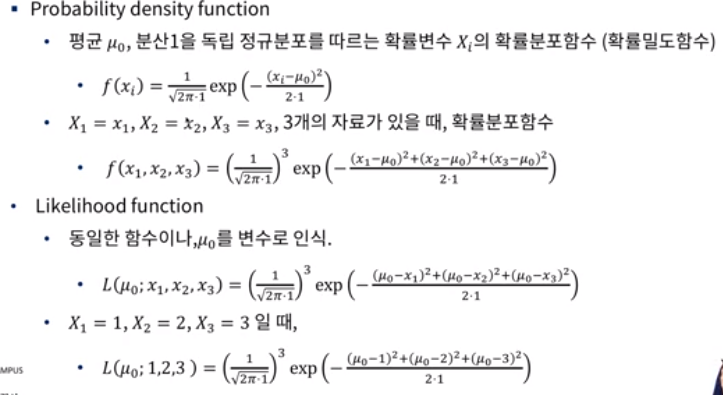
우리가 가지고있는 자료의 가능값을 가장 잘 설명하는 과정 ( 가능도함수 )

:우리가 가진 데이터를 가장 잘 설명하는 통계량을 표현해서 그것으로 모수를 추정하는게 타당하다



* 확률분포함수 : 모수를 알 때, 확률변수의 실현값을 예측하고자 함

확률밀도함수 : 연속형 확률변수의 확률 분포함수 (pdf)



보통 Likelihood를 가장 크게 만드는 값을 얻어오는 것이 목표(ex 1,2,3을 대입해보고 최대인 likelihood가 2임을 알수있다. ( MLE(maximum likelihood estimator) : likelihood를 최대로 만드는 모수의 값 u0hat)

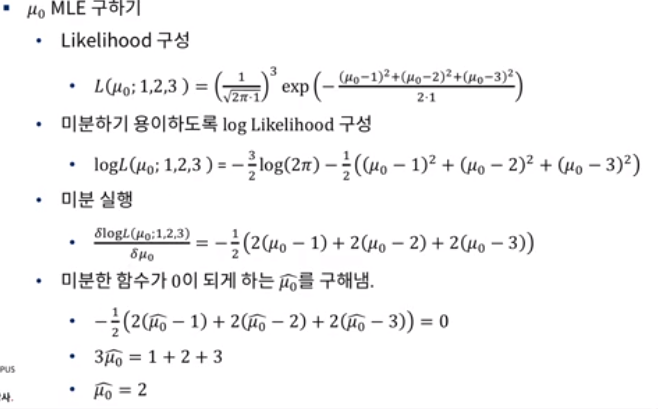
확률질량함수 : 이산형 확률변수의 확률 분포함수 (pmf)

누적분포함수 : 누적 확률 분포함수 (cdf)

* 가능도함수 : 확률변수의 실현값을 알 때( 데이터가 있을 때), 모수를 추정하고자 함

< MLE 구하기 >

미분에 용이하도록 log Likelihood 단조증가함수이기 때문에 최대가되는 미분값을 찾기 편함



미분실행에서 log(2pi)는 u와 관련이 없기에 소거

미분을 배웠다 – MLE를 구하기위해 – MLE를 왜구하나? – 모집단을 모른다. 모집단의 일부를 가지고 모수를 추정하는데 자료를 가장 잘 설명하는 likelihood를 최대로 만드는 모수값을 MLE라 하고 그것을 찾기위해 미분을한다

