0장 통계학 기초 (1)

통계학과 자료

- 통계학이란?

  불확실성을 다루는 학문으로 자료(data)를 수집, 체계화, 분석, 해석, 표현하는 학문

- 통계학 학습의 궁극적인 목적은?

  모집단(population)의 통계적 성질을 찾아내는 것

  예) 모수(parameter) 찾기, 모집단 분포(population distribution) 찾기

- 자료(data or data set)의 3요소

  (1) 변수(variable): 주체가 가지고 있는 특성 예) 키, 몸무게, 성별, 연봉

  (2) 주체(subject): 변수의 주인으로 사람(somebody) 또는 사물(something) 등

  (3) 관찰 값(observation): 해당 변수의 구체적인 값 예) 키(변수): 165cm, 172cm, 169cm

자료 예

연번	이름	직책	연봉	성별	나이
1	마이클	전문가	80K	M	28
2	이유미	인사총괄	100K	F	47
3	다이아나	비서	70K	F	33
4	김길동	DB관리	85K	M	40
5	이다희	전문가	70K	F	24
6	데이빗	수습사원	35K	M	25
7	김길동	대표이사	250K	M	56

- 주체는 몇 명(개)인가?
- 변수는 몇 개이며 무엇인가?
- [이름]은 변수인가?
- 연번은 변수인가?

변수의 종류

- 정성변수(categorical variable):
  관찰 값이 주로 문자(verbal label)로 표현되는 변수로 성별, 이름, 자동차 브랜드 등이 예이다. 때로는 관찰 값을 숫자로 표기(코드화)하기도 하는데, 코드화된 숫자는 **수학연산이 불가능**한 경우이며, 코드화된 변수도 여전히 정성변수임.

➢ 자료 예

연번	성별
1	남성
2	여성
3	남성
4	여성

연번	성별(코드화)
1	0
2	1
3	0
4	1

변수의 종류

- 정량변수(numerical variable):
  관찰 값이 숫자(numerical value)로 표현되는 변수로 이산변수 (discrete variable)와 연속변수(continuous variable)로 구분된다.

- 이산변수: 관찰 값(관찰치)이 연산이 가능한 숫자로 표현되며, **셀 수 있는 형태**인 경우이며 지정된 값 이외의 경우는 관찰할 수 없음.

➢ 자료 예

연번	형제자매의 수
1	1
2	2
3	3
4	2
5	4

연번	교과목 시수
1	3학점
2	0학점
3	3학점
4	1.5학점

변수의 종류

- 연속변수: 관찰 값(관찰치)이 연산이 가능한 숫자로 표현되며, 관찰 값이 **셀 수 없는 형태**인 경우

➢ 자료 예

연번	키(cm)
1	178.3
2	159.1
3	170.8
4	165.3
5	172.4

- 셀 수 없는 형태란?

  키는 실수(real number)로 예를 들면, 150~190cm 사이에 관찰 가능한 값이 몇 개인지 셀 수가 없다. 왜냐하면, 실수는 소수 이하 무한대의 값이 존재하기 때문임.

모집단과 표본

- 모집단(population)

  모집단은 **변수명**과 **주체**를 이용해서 정의할 수 있다. 주체의 개수를 N으로 표기한다. N은 반드시 큰 값일 필요는 없고, 연구자가 관심을 갖는 집단이 모집단이 된다. 아래의 예 '하버드 경제학과 교수들의 연봉'에서 '연봉'이 변수 명이고, '하버드대 경제학과 교수들'이 주체가 된다. 해당 대학 학과의 교수의 수가 N이 된다.

- 표본집단 또는 표본(sample)

  표본집단도 모집단과 마찬가지로 **변수명**과 **주체**를 이용해서 정의할 수 있는데, 표본집단의 주체는 반드시 **모집단 주체**의 **부분집합**이어야 한다. 표본집단의 주체의 수를 n으로 표기하는데, $n < N$ 의 관계가 성립한다.

모집단과 표본

➢ 자료 예

모집단	표본집단
하버드대 경제학과 교수들의 연봉	하버드대 경제학과 조교수들의 연봉
KOSPI 상장기업의 연 수익	KOSPI 상장기업 중 조선업종 기업의 연 수익
뉴욕대 학생들의 GPA	무작위로 추출한 100명의 뉴욕대 학생의 GPA

- 위의 모집단에 대한 표본집단의 예를 들어보시오

기술통계와 추론통계

- 기술통계(descriptive statistics)

  모집단이나 표본 집단을 이용해서 구한 값이나 그림, 도표를 기술통계라고 한다.

- 추론통계(inferential statistics)

  표본 집단을 이용해서 구한 값, 그림, 도표로 모집단의 특성을 찾아가는 과정을 추론통계라고 한다. 추론통계방법은 크게 추정(estimation)과 검정(test)으로 구분된다. 추정은 다시 점추정과 구간추정으로 구분된다.

모수와 통계량

- 모수(parameter)

  모집단을 구성하는 모든 자료를 이용해서 구한 값을 모수라고 한다.

  예를 들어서, 모집단의 평균(모평균), 모집단의 분산(모분산) 등이 있다.

  모수는 고정된(fixed) 값으로 간주한다.

- 통계량(statistic)

  표본 집단을 구성하는 자료를 이용해서 구한 값을 통계량이라고 한다.

  예를 들어서 표본평균, 표본분산 등이 있다. 통계량은 변화하는(random) 값이다.

  추정에 사용되는 통계량을 추정량(estimator)이라고 하며, 검정에 사용되는

  통계량을 검정통계량(test statistic)이라고 부른다. 실제 관측된 추정량의 값을

  추정치(estimate)라고 부른다.

모수, 통계량, 기술통계, 추론통계

- 모집단의 평균과 표본집단의 평균은 기술통계 기법인가?
- 모집단의 평균과 표본집단의 평균을 이용하여 추론통계를 설명해 보시오
- 모집단의 평균은 모수인가 통계량인가?
- 표본집단의 평균 모수인가 통계량인가?
- 모수와 통계량을 이용하여 추론통계를 설명해 보시오

분포의 정의와 분석 기법

- 분포의 정의

  변수에 대해서 그 변수가 가질 수 있는 값(관찰치)과 그 값이 나올 수 있는 확률에 대한 정보를 나타내는 것을 해당변수에 대한 분포라고 한다. 따라서, 분포는 변수에 대해서 정의되어 있고, 구체적으로 해당 변수의 관찰치를 이용하여 표현할 수 있다.

  분포는 그림이나 도표를 이용한 시각화 기법과 특정 값(모수 또는 통계량)을 계산하는 수치화 기법이 있다.

- 변수의 종류에 따른 기법

	시각화 기법	수치화 기법
정성변수	막대그래프(bar graph), 파이차트(pie chart)	비율계산, 개수산정
정량변수	히스토그램, 빈도분포표, 줄기-잎 그림, 상자그림, 산포도	평균, 중앙값, 퍼센타일, 트림드 미인, 레인지(range), IQR, 분산, 표준편차, 공분산

정량변수의 수치화 기법

- 정량변수의 수치화

  정량변수의 분포를 수치로 표현하는 기법은 다양하다. 분포의 어떠한 특성을 표현하느냐에 따라서 여러 가지 기법으로 나뉜다. 분포의 특성은 크게,

  1) 중심경향치(central tendency),

  2) 퍼짐정도(spreadness 또는 dispersion),

  3) 치우침 정도(skewness) 등이 있다.

정량변수의 수치화 기법

- 중심경향치(cnetral tendency)

  중심경향치는 "정량변수가 갖는 분포의 대푯값은 무엇인가?"와 같은 질문의 대답으로 사용될 수 있다. 중심경향치를 표현하는 방법은 평균, 중앙값 등이 있다.

- 퍼짐정도(spreadness)

  분포의 퍼짐정도는 자료의 변동성은 어느 정도인가? 분포가 어떻게 펼쳐져 있는가? 이상치가 있는가? 와 같은 물음에 대답으로 사용될 수 있다. 퍼짐정도의 측정방법으로는 분산, 표준편차, IQR 등이 있다. 주로 0과 같거나 양수로 표시

- 치우침정도(skewness)

  치우침 정도는 분포의 모양을 표현하는 것으로, skewness 값을 이용하여 측정할 수 있다.

중심경향치: 평균

- 평균(mean, average)

  평균은 모집단을 이용해서 구한 평균인 모평균과 표본집단을 이용해서 구한 평균인 표본평균이 있다. 산식은 아래와 같다.

모평균	표본평균
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

모평균 공식은 모집단 전체의 관찰 값의 크기인 N 을 이용하는 반면, 표본평균은 표본집단 전체의 관찰 값의 크기인 n 을 이용한다. N 과 n은 $N > n$ 의 관계를 만족한다.

퍼짐정도: 분산

- 분산(variance)

  분산은 모집단을 이용해서 구한 분산인 모분산과 표본집단을 이용해서 구한 분산인 표본분산이 있다. 산식은 아래와 같다.

모분산	표본분산
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  모분산 공식은 모집단 전체의 관찰 값의 크기인 N 을 이용하는 반면, 표본분산은 표본집단 전체의 관찰 값의 크기인 n 을 이용한다. 표본분산의 공식을 자세히 살펴보면, n이 아닌 n-1 을 이용하고 있다.

- 분산 값의 범위는 최소 0에서 최대 무한대에 해당된다.
- 분산이 0인 경우는?

퍼짐정도: 표준편차

- 표준편차(standard deviation)

  표준편차는 분산에 square root를 이용하여 변환을 시킨 값이다. 모 표준편차와 표본 표준편차 공식은 아래와 같다.

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

- 퍼짐정도에 대해서 '분산'이라는 좋은 측정방법이 있는데, 굳이 '표준편차'를 사용하는 이유는 무엇인가?

0장 통계학 기초 (2)

기본용어

- 확률실험(random experiment)

  결과 치를 100% 확신을 가지고 예측할 수 없는 실험이 확률실험이다.

  예를 들면, 동전를 던져서 잡은 뒤에 나올 수 있는 결과에 대해서 예측을 하는 실험을 생각해 보자. 나올 수 있는 결과는 앞면과 뒷면으로 한정할 수 있지만, 동전 던지기를 했을 때, 앞면이 나올지 뒷면이 나올지에 대해서 미리 100% 확신을 가지고 예측할 수는 없다. 이러한 실험이 확률실험이다.

기본용어

- 표본공간(sample space)

확률실험을 통해서 나올 수 있는 결과물(outcomes)의 집합을 표본공간이라고 한다. 일반적으로 S를 이용해서 표현한다. 예를 들어서, 동전 던지기를 하였을 때 나올 수 있는 결과를 앞면과 뒷면이라고 할 때, 앞면이 나올 때 H로 표기하고, 뒷면이 나올 때 T로 표기하면, 표본공간 S={ H, T }로 표기할 수 있다.

표본공간을 구성하는 결과물(outcome)을 셀 수 있을 경우에는 이를 이산 표본공간이라 하며, 셀 수 없는 경우는 연속 표본공간이라고 한다. 연속 표본공간의 예를 들어보면, 전화 통화시간을 T라고 하면, 표본공간 S={ T | T ≥ 0}로 표기할 수 있다.

이산 표본공간은 개개의 결과물들(outcomes)을 이용하여 집합의 형태로 표현할 수 있다.

S={E1, E2, ..., Ek}

기본용어

- 사건(event)

확률실험을 통해서 나올 수 있는 결과물의 부분집합(subset)이 사건(event)이다.
구체적인 사건은 연구자의 관심사항에 따라서 정의될 수 있다. 예를 들어서 동전 던지기를 하였을 때, 앞면이 나오는 사건을 A라고 하면, A={H}로 표기할 수 있다.

하나의 결과물(outcome)로 구성되어 있는 사건을 elementary event라고 하며,
하나 이상의 결과물로 구성되어 있는 사건을 complimentary event라고 한다.
후자의 예는, 주사위 던지기를 하였을 때 짝수가 나오는 사건을 B라고 하면,
B={2,4,6}으로 표기할 수 있다.

- Equally likely

확률실험에서 나올 수 있는 하나하나의 결과물이 선택될 기회(chance)가 모두 같은 경우를 equally likely라고 한다.

예) 동전이 정상적인(fair, balanced) 동전이라면, 동전 던지기에서 나올 수 있는 결과물인 H와 T가 나올 기회는 같다.

확률

- 확률의 정의

확률은 어떤 사건(event)이 발생할 기회 및 상대적 가능도(relative likelihood)를 숫자로 표기한 것이다. 사건은 집합(set)의 형태로 표기되는데 확률이라는 함수를 통해서 숫자(number)로 변환시키는 것이다. A라는 사건이 발생할 확률은 P(A)로 표기하고, P(A)의 범위는 $0 \le P(A) \le 1$이다.

P(A)=0 이면, A라는 사건은 절대 발생하지 않는다는 것이고,
P(A)=1 이면, A라는 사건은 항상 발생한다는 것이다.
P(A)가 1에 가까울수록 A 사건의 발생 가능성은 큰 것이고, 0에 가까울수록 작은 것이다.

표본공간이 이산인 경우에, 표본공간의 확률은 $P(S)=P(E1)+P(E2)+\ldots+P(Ek)$이 성립한다.

확률

- 확률의 종류

1) 경험적 확률(empirical probability)

기존의 데이터나 실험을 통해서 얻은 자료를 이용해서 얻은 확률이 경험적 확률에 해당된다. 빈도 분포 표에서 상대빈도수(relative frequency)가 경험적 확률에 해당된다.

예1) 쌍둥이가 태어날 확률: 전체 신생아 출생 수 가운데, 쌍둥이인 경우의 수를 비율로 계산

예2) 학생 대출의 부도확률: 전체 학생 대출 건 가운데 부도난 건의 비율로 계산

확률

2) 전통적(이론적) 확률(classical or theoretical probability)

기존의 데이터가 없거나 실험을 하지 않더라도 연역적 방법을 이용해서 구할 수 있는 확률로 동전 던지기에서 앞면이 나올 확률은 누구나 0.5라고 답할 수 있다. 이러한 답을 얻는 과정에서 기존의 데이터나 실험은 필요하지 않고 논리적 사고와 연산만 정확히 할 수 있다면 누구나 같은 대답을 도출해 낼 수 있다.

3) 주관적 확률(subjective probability)

기존에 발생한 적이 없어서 자료가 준비되어 있지 않고, 실험을 통해서 얻을 수 없으며, 연역적 추론을 통해서도 구할 수 없는 경우에는 종종 해당 분야 전문가의 전문성에 기대어 확률을 계산할 수도 있다. 이러한 확률은 누구나 같은 결론을 도출해 내기 어렵고 다분히 전문가 주관적인 성격이 강하다. 이렇게 얻은 확률을 주관적 확률이라고 한다.

확률

- 대수의 법칙(law of large number)

  경험적 확률은 (표본자료 관찰 값의 크기 또는 실험의 횟수)이 커짐에 따라서 이론적 확률에 점점 근접하게 된다.

  동전 던지기를 경험적 확률 방법으로 구하려면, 동전 던지기 시행 횟수 가운데 앞면이 나올 경우의 수를 비율로 계산하여야 한다. 예를 들어서, 3번 동전을 던져서 1번 앞면이 나오면, 앞면이 나올 확률은 1/3이고, 10번 동전을 던져서 5번 앞면이 나오면, 앞면이 나올 확률은 1/2이 된다.

  동전 던지기 시행을 할 때 마다 앞면이 나올 확률은 이론적 확률인 0.5에 가까워지거나 멀어질 수도 있으나, 시행횟수를 증가시킴에 따라서 경험적 확률은 점점 이론적 확률에 근접하게 된다.

확률변수의 정의

- 확률변수(random variable)

  확률실험(random experiment)을 하였을 때 나올 수 있는 표본공간(sample space)의 결과물(outcome)에 숫자(numerical value)를 부여하는 함수(function)를 확률변수라고 한다. 확률변수(명)는 대문자 X 또는 Y로 표현하고, 구체적인 값은 소문자 x 나 y 등으로 표기한다. 확률변수 X = x로 표기한다.

- 이산 확률변수, 연속 확률변수

  확률변수(X)가 가질 수 있는 값(x)가 셀 수 있으면 이산 확률변수라고 부르고, 셀 수 없으면 연속 확률변수라고 부른다. 위의 예에서처럼 X가 동전 던지기를 3번 해서 앞면이 나오는 개수라고 하면, 나올 수 있는 값(x)은 0,1,2,3 으로 4개의 값을 가질 수 있으므로 이때의 X는 이산 확률변수이다. 반면에, 확률변수(Y)를 몸무게라고 하자. Y가 가질 수 있는 값(y)이 50kg에서 120kg이라고 할 때, 50~120kg 사이에 관찰 가능한 값의 개수는 셀 수가 없다. 따라서 몸무게는 연속 확률변수에 해당된다.

확률변수의 확률계산

➢ 이산 확률변수의 확률계산 방법과 연속 확률변수의 확률 계산 방법은 다르다. 이산 확률변수에 정의되는 확률을 확률질량함수(probability mass function)라고 부른다.

▪ 이산 확률변수

확률은 원래 사건(event)에 대해서 정의되어 있다. 사건은 표본공간(sample space)의 결과물(outcome)의 부분집합이다. 예를 들어서, 사건을 A로 표기할 때, A가 발생할 확률은 P(A)로 표기한다. 사건은 표본공간의 부분집합이므로 집합(set)의 형태를 갖는다.

확률변수는 확률실험의 표본공간(sample space)의 결과물(outcome)에 대하여 수(numerical value)를 할당하는 함수로, 집합의 형태가 아니다. 그렇다면, 집합이 아닌데도 불구하고 어떻게 확률을 계산할 수 있을까?

확률변수의 확률계산

➢ 예1) 확률변수 X를 동전 던지기 1회를 하여 앞면이 나오는 개수라고 하자.

결과물(outcome)	X	P(outcome)	P(X=x)
T	0	P(T)=0.5	P(X=0)
H	1	P(H)=0.5	P(X=1)

➢ 예2) 확률변수 X를 동전 던지기 3회를 하여 앞면이 나오는 개수라고 하자.

결과물(outcome)	X	P(outcome)	P(X=x)
TTT	0	P(TTT)=1/8	P(X=0)
HTT, THT, TTH	1	P(HTT∪THT∪TTH)=3/8	P(X=1)
HHT, THH, HTH	2	P(HHT∪THH∪HTH)=3/8	P(X=2)
HHH	3	P(HHH)=1/8	P(X=3)

확률변수의 확률계산

- 확률질량함수(probability mass function: pmf)

이산 확률변수에 정의되는 확률을 확률질량함수(pmf)라고 부르며, 이는 확률의 일반적인 성질을 가지고 있다. 확률변수 X가 어떤 값 x를 가질 때, $0 \le P(X=x) \le 1$을 만족한다. 확률변수 X가 가질 수 있는 모든 x에 대하여 확률질량함수를 계산하여 합하면 1이 된다.

즉,

$$\sum_{all\ x} P(x) = 1$$

이는 사건(event)에 대해서 정의되어 있는 확률의 성격과 부합한다.

Note: 사건 A에 대해서 $0 \le P(A) \le 1$이며, $P(S)=1$을 만족한다.

확률변수의 확률계산

- 누적(확률)분포함수(cumulative (probability) distribution function: cdf)

  확률변수의 값을 최솟값부터 특정 값까지의 구간 개념으로 정의하고, 이에 대한 확률을 계산하고자 할 때, 누적분포함수를 이용한다. 확률변수 X의 값을 최솟값부터 특정 값 까지의 누적분포함수는 다음과 같이 표기한다.

$$F(x_0) = P(X \le x_0)$$

  누적분포함수는 구간에 대한 확률을 정의한 개념이므로 최솟값은 0이고, 최댓값은 1이다. 이산확률변수의 경우, 아래의 공식을 이용하여 계산한다.

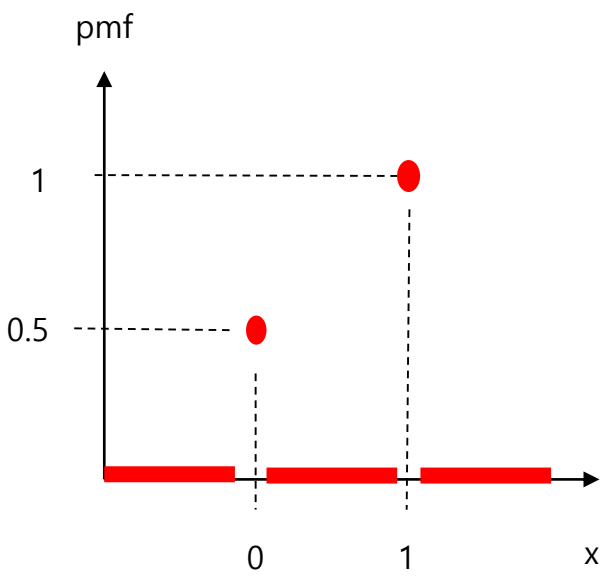
$$F(x_0) = \sum_{x \le x_0} P(x)$$

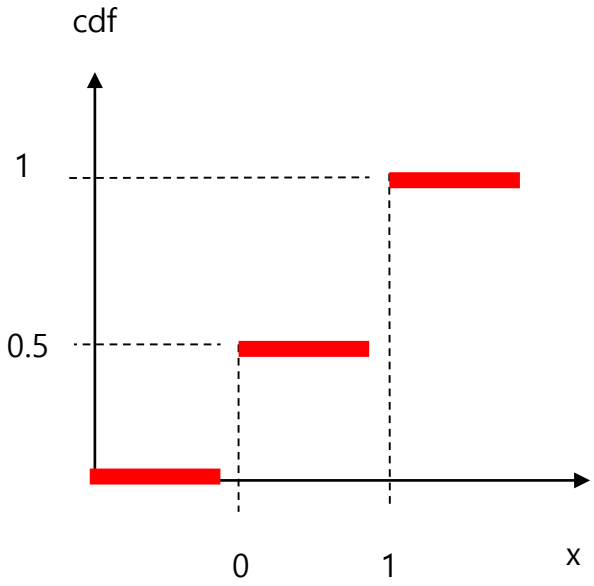
➢ 예) X가 동전 던지기 1회 하였을 때 앞면의 개수일 때, F(0), F(1), F(0.5)=?

확률변수의 확률계산

예) X가 동전 던지기 1회 하였을 때 앞면의 개수일 때, F(0), F(1), F(1.5)=?

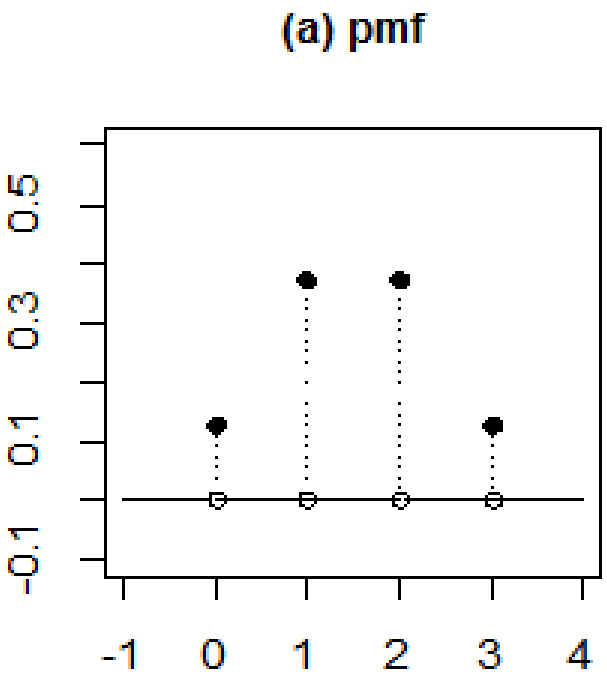
결과물(outcome)	X	P(X=x)	F(x)=$P(X \leq x)$
T	0	P(X=0)=0.5	$P(X \leq 0)$=0.5
H	1	P(X=1)=0.5	$P(X \leq 1)$=1

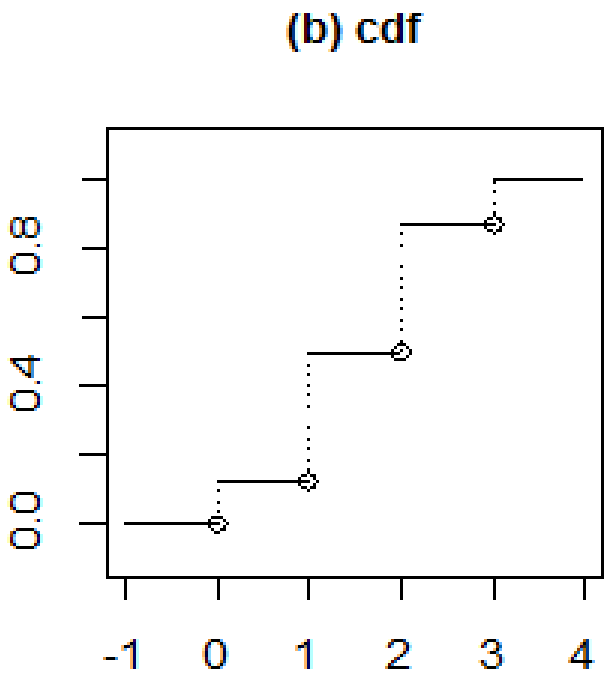




확률변수의 확률계산

예) X가 동전 던지기 3회 하였을 때 앞면의 개수일 때, F(0), F(1), F(2)=?





확률변수의 확률계산

- 연속 확률변수

연속 확률변수의 경우에 확률변수 X가 특정 값 (point) x를 가질 확률을 계산하면 항상 0이다. 즉, P(X=x)=0.

연속 확률변수는 셀 수 없다고 정의되어 있는데, 예를 들어서 X를 키라고 하면, 160cm와 170cm 사이에는 무수히 많은 관찰 값이 존재한다고 할 수 있으며, 정확히 몇 개가 존재하냐는 물음에는 답할 수 없다. 키는 실수(real number)로 표현할 수 있으므로 소수 이하 무한대의 값으로 표현 가능하며, 따라서 셀 수가 없는 것이다. 이렇게 무한히 많은 값들 가운데 하나의 값이 나올 상대적 가능도는 0이라고 해도 무방하다.

그렇다면, 연속 확률변수의 경우에는 확률 계산이 아예 불가능한 것인가? 연속확률변수의 경우에는 확률을 구간을 이용해서 계산한다. 예를 들어서 여러 명 가운데 한 명을 선택하여 측정하였을 때 키가 172.3(하나의 값: point)이 될 확률은 0이지만, 172cm에서 173cm 사이에 있을 확률은 0이 아니다. 구간 172~173cm는 비록 1cm 밖에 되지 않지만, 이 범위 안에 역시 무한히 많은 관찰 값이 존재하기 때문에 0으로 단정 지을 수 없다.

확률변수의 확률계산

- 이산 확률변수에서는 하나의 값(point)에 대한 확률을 구할 수 있었고 이를 확률질량함수라고 불렀는데, 연속 확률변수에서는 하나의 값에 대한 확률은 항상 0이므로 이를 직접 구할 수 없고 구간을 이용해서 구해야 한다. 이 때 사용하는 함수를 확률밀도함수라고 한다. 확률밀도함수는 확률질량함수와 달리 확률의 기본적인 성격을 포함하고 있지 않다.

- 확률밀도함수(probability density function: pdf)

확률밀도 함수는 f(x)로 표기하고, 0≤f(x)의 특성을 갖는다. 상한 값이 1로 지정되어 있지 않으므로, 확률질량함수와 달리 확률의 의미를 갖고 있지는 않다. 다만, 확률밀도함수를 그린 후 0을 통과하는 수평선과 확률밀도함수 선 아래쪽 사이의 면적을 계산해 보면 항상 1이 되어야 한다. 이를 수식으로 표현하면, 다음과 같다.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

확률밀도함수 아래쪽의 면적은 적분(integration)을 이용해서 구할 수 있고, 변수의 범위를 실수라고 정의하면 실수 모든 구간에 대한 적분을 계산할 수 있다.

확률변수의 확률계산

- 연속변수의 누적분포함수

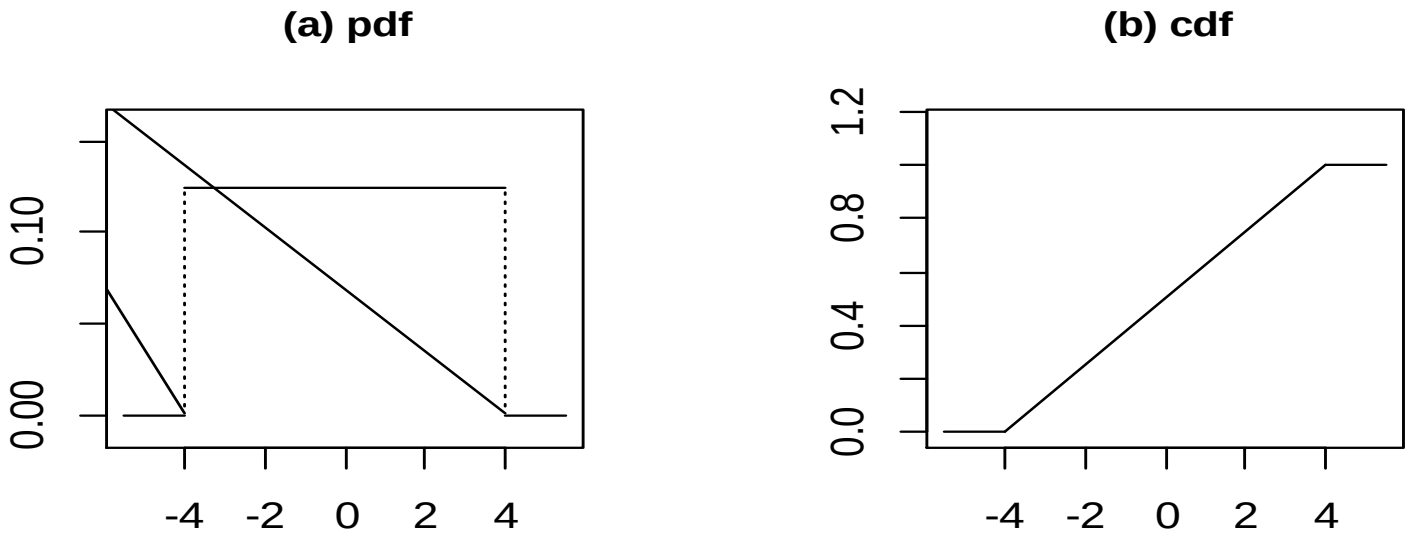
  앞서 설명한 바와 같이 연속변수가 특정 값을 가질 확률은 항상 0이므로 연속 확률변수의 확률은 구간을 이용하여 계산하여야 하는데, 구간에 대해서 확률이 정의되어 있는 누적분포함수(cdf)를 이용하여 계산하게 된다.

  연속변수의 누적분포함수는 이산변수에서와 마찬가지로 $F(x_0) = P(X \le x_0)$ 으로 정의되어 있으나, 계산방식은 아래와 같다.

$$F(x_0) = \int_{-\infty}^{x_0} f(x)dx$$

확률변수의 확률계산

➢ 예) 균등분포를 따르는 연속변수에 대한 확률밀도함수와 누적분포함수



확률변수의 확률계산

- 연속변수의 확률계산

연속변수의 경우, 하나의 값(point)에 대한 확률은 항상 0이므로 구간의 개념을 이용하여 확률을 계산한다.

예를 들어서 $P(a < X \leq b)$ 을 구하고자 할 경우에는 확률밀도함수와 누적확률분포함수의 정의를 이용하여 $P(X \leq b) - P(X \leq a)$ 로 구할 수 있고, 이는 적분을 이용하여 구할 수 있다.

참고로, $P(X = a) = 0$ 이므로 다음의 관계가 성립한다.

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b)$$

확률변수의 기댓값

- 확률변수의 기댓값

  확률변수 X의 기댓값은 E(X)로 표기하고, 모평균 μ_x 를 의미함.

  계산방식은 아래와 같음.

이산변수	연속변수
$\sum_{all\ x} xP(x)$	$\int_{-\infty}^{\infty} xf(x)dx$

- 확률변수의 기댓값 일반화

이산변수
$E[g(X)] = \sum_{all\ x} g(x)P(x)$

  - 예) $g(X) = X^2$

확률변수의 분산

- 확률변수의 분산: $g(X) = (X - \mu_x)^2$

$$E(X - \mu_x)^2 = \sum_{all\ x} (x - \mu_x)^2 P(x)$$

$$V(X) = \sigma^2 = E(X^2) - \mu_x^2$$

확률변수의 기댓값 연산규칙

- 확률변수 X, Y와 상수 $a,\ b$

1) $E(a) = a,\ \ V(a) = 0$

2) $E(a + bX) = a + bE(X)$

$V(a + bX) = b^2V(X)$

3) $E(aX + bY) = aE(X) + bE(Y)$

$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$

❖ $Cov(X, Y)$은 무엇인가?

두 확률변수의 기댓값

- 두 확률변수의 기댓값 일반화

  확률변수 X와 Y의 특정함수 g(X,Y)의 기댓값은 E[g(X,Y)]로 표기하고,
  계산방식은 아래와 같음. 여기서 P(x,y)는 두 변수의 결합확률분포 함수임.

이산변수

$$E[g(X,Y)] = \sum_{all\ x} \sum_{all\ y} g(x,y)P(x,y)$$

- 예) $g(X,Y) = XY$

확률변수의 공분산과 상관계수

- 두 확률변수의 공분산: 두 변수 사이의 선형관계의 강한 정도를 측정

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

$$= \sum_{all\ x} \sum_{all\ y} (x - \mu_x)(y - \mu_y) P(x,y)$$

$$-\infty < Cov(X,Y) < \infty$$

- 두 확률변수의 상관계수

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

$$-1 \leq Corr(X,Y) \leq 1$$

두 확률변수의 독립과 상관계수 0

- 확률변수의 독립과 $Cov(X, Y) = 0$

  두 확률변수 X와 Y가 독립이면, 이들의 상관계수가 항상 0이다.

  그러나, 상관계수가 0이라고 항상 독립이라고 할 수는 없다.