데이터과학과 AI를 위한 파이썬

10강. 확률분포의 종류와 특성

세종사이버대학교

김명배 교수

학습내용

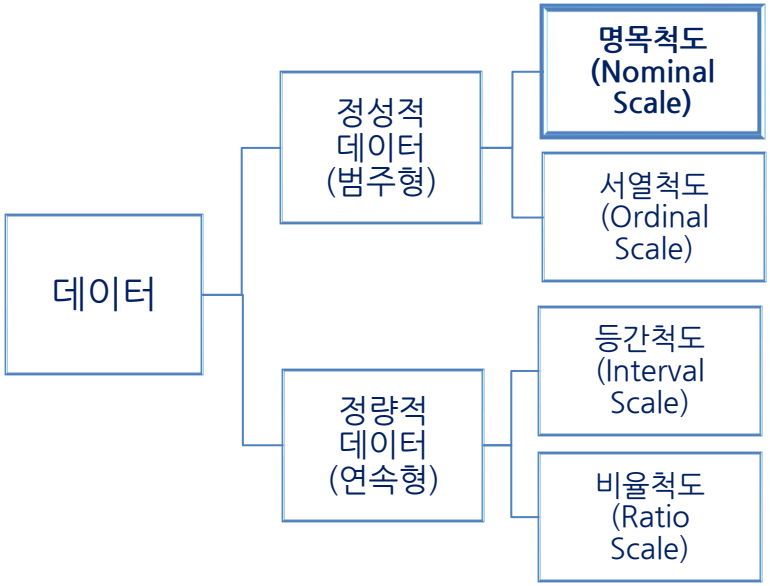
- 변수의 구분
- 분포특성의 표현
- 확률분포의 종류

학습목표

- 변수의 유형을 나열하고 데이터를 보고 분류할 수 있다.
- 분포의 특성을 파악하는 방법을 설명할 수 있고, 각 기술통계량의 의미를 설명할 수 있다.
- 이산형 확률분포와 연속형 확률분포의 종류를 설명하고 파이썬으로 시뮬레이션 할 수 있다.

1. 변수의 구분

1) 변수의 유형에 따른 분류



▪명목척도 (Nominal Scale)

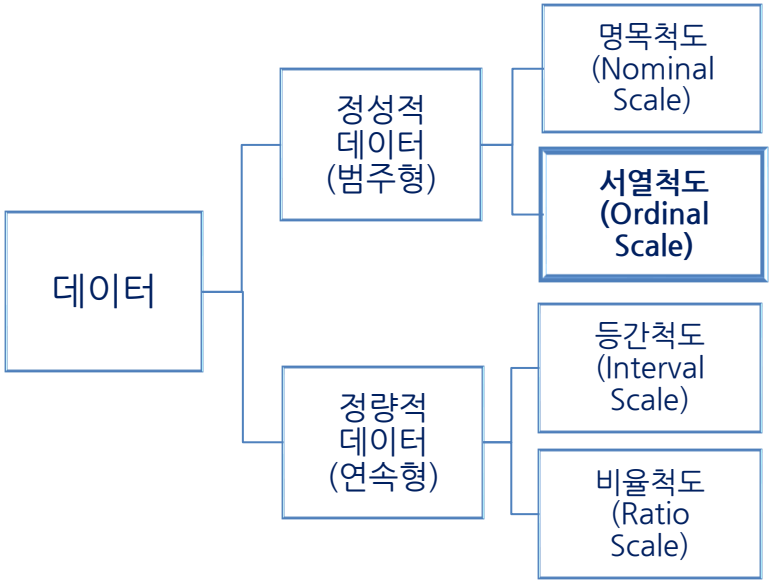
- 범주형 데이터로 측정된 현상을 상호배타적인 범주(category)로 수치를 부여한 척도

(예시) 성별 (남자=1, 여자=2), 혈액형 (A=1, B=2, AB=3, O=4) 등

- 순서의 개념이나 가감승제의 수학적 연산 기능의 의미가 없는 척도

1. 변수의 구분

1) 변수의 유형에 따른 분류



▪서열척도(Ordinal Scale)
= 순서척도, 순위척도

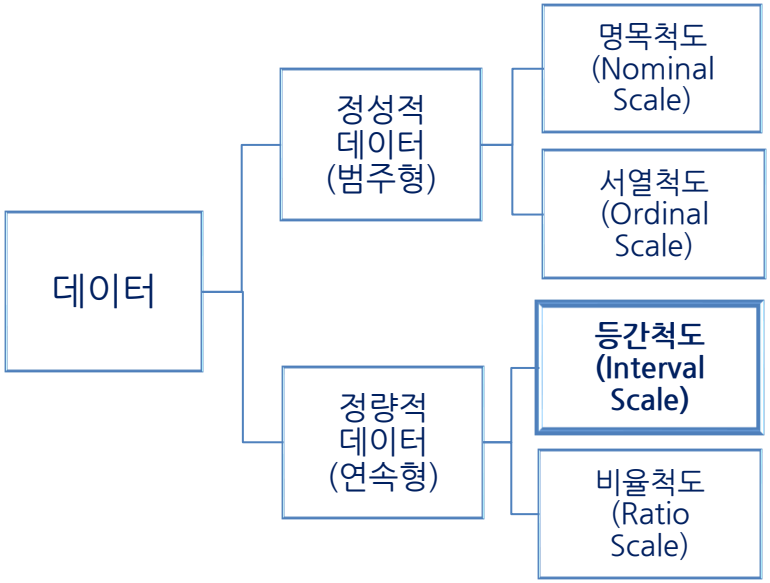
- 범주형 데이터로 명목척도의 기능 뿐만 아니라 각 범주간의 대소관계, 서열성에 관하여 수치를 부여한 척도

(예시) 건강상태 (나쁨=1, 보통=2, 양호=3) 등

- 수학적 의미 : A > B, A < B , A = B

1. 변수의 구분

1) 변수의 유형에 따른 분류



▪등간척도 (Interval Scale)=구간척도

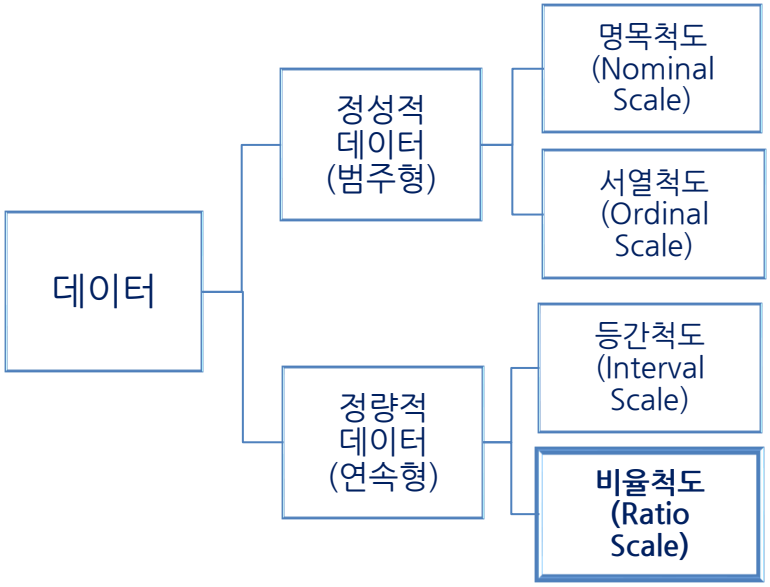
- 연속형 데이터로 절대적 원점(Absolute zero)이 없음
- 양적인 정도의 차이에 따라 등(等)간격으로 수치를 부여한 척도

(예시) 온도 (섭씨 0℃, 50℃, 100℃), 물가지수, 생산지수 등

- 수학적으로 가감(+,-)의 조작이 가능
- 승제의 조작은 불가능한 척도

1. 변수의 구분

1) 변수의 유형에 따른 분류



▪비율척도 (Ratio Scale)

- 절대적 원점이 존재하며, 비율계산이 가능한 수치를 부여한 척도

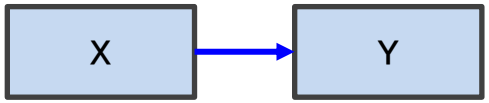
(예시) 광고비, 판매량, 매출액, 무게, 가격, 소득 등

- 수학적으로 가감승제(+,-,×,÷)의 조작이 모두 가능한 척도

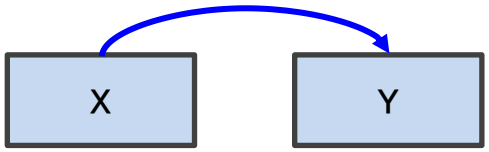
1. 변수의 구분

2) 역할에 따른 변수의 분류

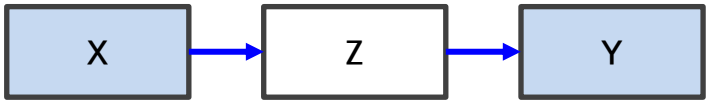
① 인과 관계

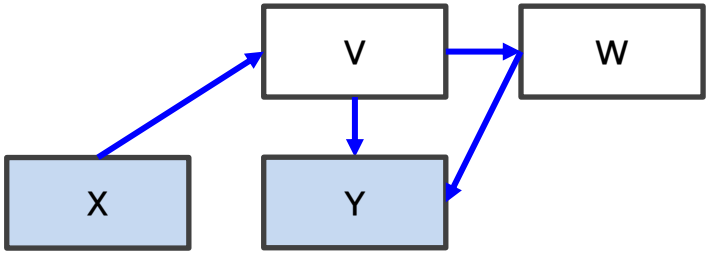


② 상관 관계



③ 제 3의 변수와의 관계



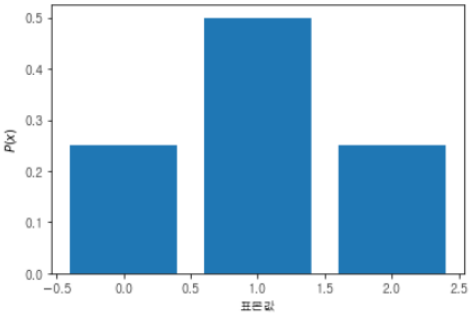


2. 분포특성의 표현

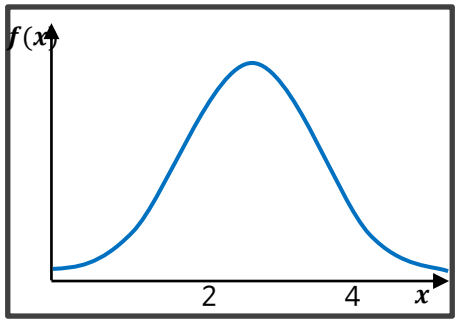
1) 그래프를 통한 분포의 표현

- 분포란 확률적 데이터에서 어떠한 값이 자주 나오고 어떠한 값이 드물게 나오는가를 나타내는 정보
- 그래프를 활용하여 분포를 표현할 수 있음

범주형 확률변수



연속형 확률변수



2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

- 분포의 특징을 나타내는 여러 가지 숫자를 계산하여 그 숫자로 분포를 나타내는 값을 기술 통계(descriptive statistics)라고 함

① 표본 평균(mean, average)

- 관찰치 전체를 합한 후에 자료의 관찰치 총 개수로 나눈 값
- 데이터 분포의 대략적인 중심을 나타내는 대표값
- 특이치/이상치 같은 극단값에 영향을 많이 받는다.

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

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

② 표본 중앙값(median)

- 데이터를 크기별로 정렬했을 때, 가장 중앙에 위치한 값
- 평균과 함께 데이터 분포의 대략적인 중심을 나타내는 대표값
- 특이치/이상치 같은 극단값에 영향을 덜 받는다.

  - N이 홀수 : (N+1)/2 번째 값
  - N이 짝수 : (N/2 번째 값 + N/2+1 번째 값)/2

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

③ 표본 최빈값(mode, most frequent value)

- 데이터 중에서 가장 많이 발생한 값
- 명목형 척도에 주로 사용함

④ 합계(sum)

- 전체 개체들의 총합

$$\sum_{i=1}^{N} x_i$$

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

⑤ 백분위수(Percentile)

- 동일빈도를 가지도록 100개의 구간으로 나누어 표현(0~100%)

⑥ 사분위수(Quartile)

- 백분위수와 비슷한 의미로, 자료를 순서화하여 특정 위치에 값을 나타냄
- 25%(1사분위 수), 50%(중앙값), 75%(2사분위 수), 100%(최대값) 비율에 해당하는 값

⑦ 사분위 범위(IQR; Interquartile Range)

- 극단값을 제외한 범위를 나타냄

  3사 분위수 - 1사분위수

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

⑧ 분산(variance)과 표준편차(SD; Standard deviation)

- 분포의 퍼짐정도를 나타내는 정보
- 데이터가 얼마나 변동하고 있는지의 정도를 표현하는 정보

표준편차(s) = $\sqrt{\text{분산}(v)}$

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

⑨ 변동계수(CV; coefficient of variation)

- 표준편차(σ)를 산술평균(μ)을 기준으로 표준화시킨 것

$$CV = \frac{\sigma}{\mu}$$

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

⑩ 최대값(Max)과 최소값(Min)

- 데이터의 가장 큰값과 작은 값을 나타내는 정보
- 유효범위 내에 있는지 살펴볼 수 있음

⑪ 범위(range)

- 데이터의 가장 큰 차이가 얼마인지를 나타냄
- 최대값 - 최소값

2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

⑫ 평균의 표준오차(SEM; standard error of the mean)

- 표본 평균들의 표준편차

$$SEM = \sqrt{Var(\bar{x})} = \frac{s}{\sqrt{n}}$$

⑬ 신뢰구간

- 모집단에서 추출한 표본에서 모수를 추정하려고 할 때, 정확도를 나타내는 것으로 신뢰도의 구간을 의미한다.
- 모집단 평균이 표본평균에 의해서 얼마나 정확히 추정될 수 있는지를 반영하는 지표

[예시] 모평균 μ에 대한 95% 신뢰구간이 $(\bar{x} - d, \bar{x} + d)$라는 의미

만일 표본을 똑같은 방법으로 100번 추출하면 모수 μ가 신뢰구간에 95번은 포함된다는 의미($d = Z - score \times \frac{s}{\sqrt{n}}$)

2. 분포특성의 표현

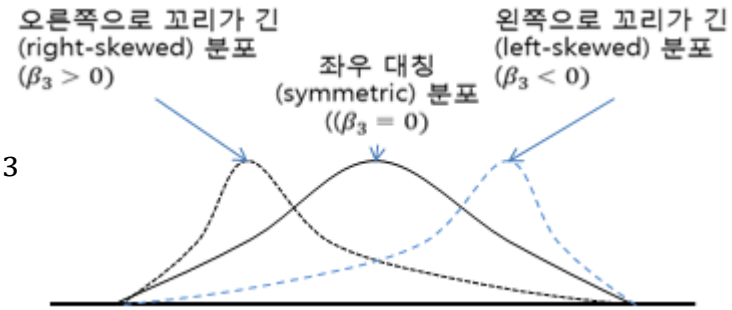
2) 기술통계를 통한 분포의 표현

⑭ 왜도(비대칭도, skewness)

- 평균과의 거리의 세제곱을 하여 구한 특징값을 왜도(비대칭도)라고 함
- 0 : 좌우 대칭임
- '–' : 외쪽이 꼬리
- '+' 오른쪽 꼬리

$$Skewness = \beta_3$$

$$= \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$



출처 : [R 분석과 프로그래밍] http://rfriend.tistory.com

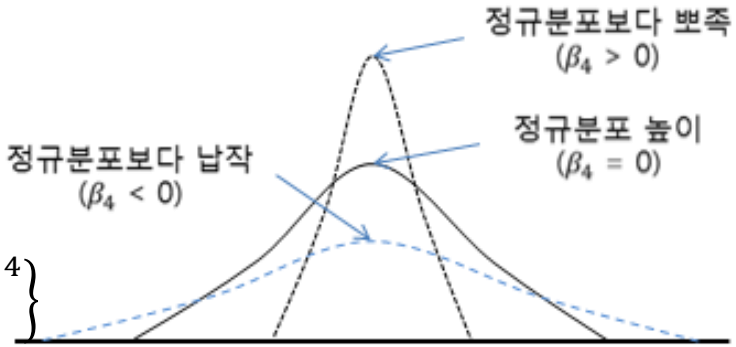
2. 분포특성의 표현

2) 기술통계를 통한 분포의 표현

⑮ 첨도(kurtosis)

- 평균과의 거리의 네제곱을 이용하여 구한 특징값을 첨도라고 함
- 분포의 뾰족한 정도를 나타냄
- 0 : 정규분포와 유사한 형태
- '+': 좁게 밀집
- '–': 넓게 밀집

$$Skewness\ \beta_4 = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$



출처 : [R 분석과 프로그래밍] http://rfriend.tistory.com

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

① 베르누이 분포(Bernoulli Distribution)

- 베르누이 시행 : 2가지 결과 중 정확히 하나로만 나오는 실험이나 시행

  Ex) 예 또는 아니오, 사망 또는 생존, 성공 또는 실패

- : 베르누이 실행 결과를 0 또는 1로 표현한 베르누이 확률변수를 베르누이분포를 따른다고 함

$$X \sim Bern(x;\mu) \quad f(x;\mu) = \begin{cases} \mu & ,if\ x=1 \\ 1-\mu & ,if\ x=0 \end{cases} \Rightarrow \mu^x(1-\mu)^{(1-x)}$$

- 평균 : μ , 분산 : $\mu(1-\mu)$

[파이썬] 싸이파이의 stats.bernoulli(), pmf() 적용

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

② 이항분포(Binomial Distribution)

- 성공확률이 μ인 베르누이 시행을 N번 시행했을 때, 성공한 횟수 X를 확률변수를 이항분포를 따르는 확률변수라고 함

$$X \sim Bin(x; N, \mu) \qquad f(x; N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$$

Ex) 동전을 10번 던져 4번 앞면이 나올 확률 (성공 : 앞면)
100개의 제품 중 불량의 수 (성공 : 불량)

- 평균 = $N\mu$, 분산 = $N\mu(1-\mu)$
- 이항분포에서 N=1인 경우 베르누이 분포가 된다.

[파이썬] 싸이파이의 stats.binom(), pmf() 적용

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

③ 카테고리분포(categorical Distribution)

- 실험의 결과가 k개인 경우

  $x = (x_1, x_2, \cdots, x_k)$, $x_i = 0$ 또는 1, $\sum_{i=1}^{K} x_i = 1$

  를 만족하는 확률변수를 카테고리 확률변수라고 하며, 카테고리 분포를 따른다고 함(주사위 던지기)

  $\mathrm{X} \sim Cat(x_1, x_2, \ldots, x_k; \mu_1, \mu_2, \ldots, \mu_K) = Cat(x; \mu)$

  $f(x; \mu) = \prod_{i=1}^{K} \mu_i^{x_i}$

- $0 \le \mu_i \le 1$, $\sum_{i=1}^{K} \mu_i = 1$
- 평균 : μ_k, 분산 : $\mu_k(1 - \mu_k)$

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

③ 카테고리분포(categorical Distribution)

- 원핫코딩(one-hot-encoding)

$x = 1 \rightarrow x = (1, 0, 0, 0, 0, 0)$

$x = 2 \rightarrow x = (0, 1, 0, 0, 0, 0)$

$x = 3 \rightarrow x = (0, 0, 1, 0, 0, 0)$

$x = 4 \rightarrow x = (0, 0, 0, 1, 0, 0)$

$x = 5 \rightarrow x = (0, 0, 0, 0, 1, 0)$

$x = 6 \rightarrow x = (0, 0, 0, 0, 0, 1)$

[파이썬] 싸이파이의 stats.multinomial(), pandas.get_dumies() 적용

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

④ 다항분포(multinomial Distribution)

- 카테고리 확률변수의 반복 횟수가 여러 개 인경우 다항분포를 따름

  [예시] 주사위를 N번 던져 각 면이 나오는 횟수의 분포

  $X \sim Mu(x; N, \mu)$ $\quad f(x; N, \mu) = \binom{N}{x} \prod_{k=1}^{K} \mu_k^{x_k}$

- 평균 : $N\mu_k$, 분산 : $N\mu_k(1-\mu_k)$

[파이썬] 싸이파이의 stats.multinomial(), pmf() 적용

3. 확률분포의 종류

1) 이산형확률분포(discrete probability distribution)

⑤ 포아송분포(Poisson Distribution)

- 어떤 희귀한 현상의 발생횟수의 확률변수는 포아송분포를 따름

[예시] 고속도로의 일정한 지역에서 주당 교통사고 건수

보험사 사고(보상)건수

- 이항분포에서 시행횟수 N이 매우 크고, 성공확률 μ가 아주 작은 경우 포아송 분포로 근사함

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, \cdots$$

- λ는 단위시간에 발생하는 평균횟수
- 평균 : λ, 분산 : λ

[파이썬] 싸이파이의 stats.poisson(), pmf() 적용

4. 연속형 확률분포

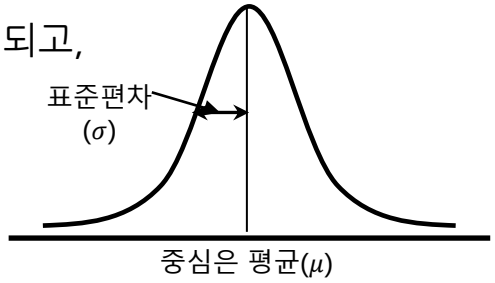
1) 연속형 확률분포(continuous probability distribution)

① 정규 분포 (Normal Distribution)

- 정규분포는 혹은 가우스(Gaussian) 정규분포는 자연 현상에서 나타나는 숫자를 확률 모형으로 모형화할 때 많이 사용함
- 모수는 평균(μ)과 분산(σ^2)임

$$X \sim N(\mu, \sigma^2), \quad f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- 평균(μ)을 중심으로 좌우 대칭으로 종모양의 확률밀도함수를 가짐
- 평균(μ)에 따라 그래프의 위치가 결정되고, 분산(σ^2)에 따라 퍼진 모양이 결정



4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

② 표준 정규 분포 (Normal Distribution)

- μ=0, σ^2=1인 정규분포를 표준정규분포라고 함
- $X \sim N(\mu, \sigma^2)$인 확률변수 X에 대해 $Z = \frac{x_i - \mu}{\sigma}$ 는 표준정규분포를 따름
- 즉 Z는 x_i를 평균이 0, 분산이 1인 분포로 변환시키는 표준화 개념임

[파이썬] 사이파이의 stats.norm() 사용

4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

[중심극한정리]

- '모집단이 평균이 μ, 표준편차가 σ인 임의의 분포를 따른다고 할 때,
  이 모집단으로부터 추출된 표본의 표본의 크기 N이 충분히 크면
  표본평균들은 평균이 μ, 표준편차가 $\sigma/\sqrt{N}$인 정규분포에 근사한다.'
- 실제로 발생하는 많은 현상을 정규분포를 따른다고 보고 분석할 수 있음
- 여러 가지 선형 모형들은 정규분포를 가정하고 있음
  예시) 선형회귀분석, 분산분석(F-test)

4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

③ 스튜던트 t-분포(student's t Distribution)

- 정규분포에서 생성된 표본 데이터 집합에 여러 수식으로 값을 변환한 데이터가 따르는 통계량 분포 중 하나
- 정규분포와 유사하지만 양 끝단의 비중이 더 큰 분포

$$t(x;\mu,\lambda,v)=\frac{\sqrt{\lambda}}{\sqrt{v\pi}}\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\left(1+\lambda\frac{(x-\mu)^2}{v}\right)^{-\frac{v+1}{2}}$$

λ는 정규분포의 정밀도 $(\sigma^2)^{-1}$에 대응되는 개념

$\Gamma(x)$는 감마 함수 $=\int_0^\infty u^{x-1}e^{-u}\,du$

v는 자유도(degree of freedom)

[파이썬] 사이파이의 stats.t() 사용

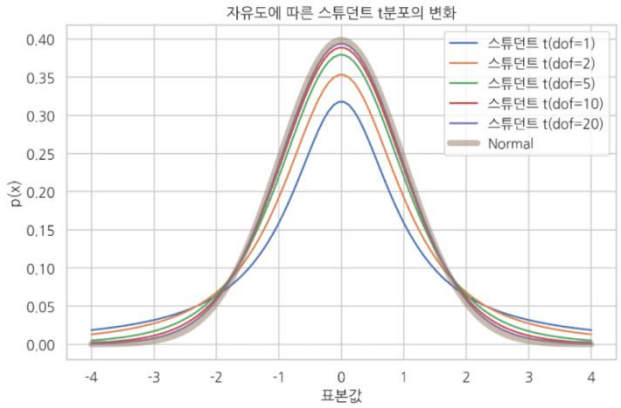
4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

③ 스튜던트 t-분포(student's t Distribution)

- t 통계량

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}} \sim t(x; 0,1, N-1)$$



4. 연속형 확률분포

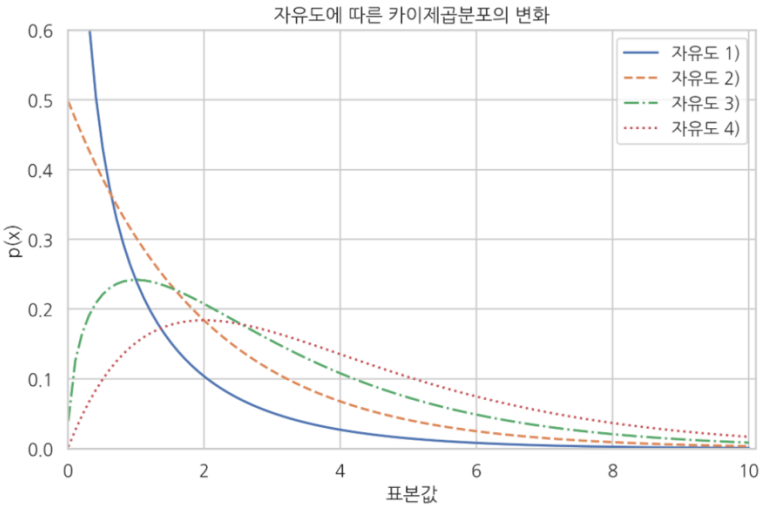
1) 연속형 확률분포(continuous probability distribution)

④ 카이제곱 분포(χ^2 Distribution)

- 화률변수 X의 N개의 표본을 제곱하여 더하면 양수값만 가지는 카이제곱 분포

$$\sum_{i=1}^{N} x_i^2 \sim \chi\text{^}2(x; v = N)$$

$$\chi^2(x; v) = \frac{x^{(v/2-1)} e^{-x/2}}{2^{v/2}\Gamma(\frac{v}{2})}$$



[파이썬] 사이파이의 stats.chi2() 사용

4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

⑤ F 분포(F Distribution)

- 카이제곱분포를 따르는 독립저긴 두 확률변수 표본을 각각의 자유도(N)으로 나눈 뒤 비율을 구하면 F분포가 됨

$$x_1 \sim \chi^2(N_1), \qquad x_2 \sim \chi^2(N_2) \rightarrow \frac{x1/N_1}{x_2/N_2} \sim F(x; N_1, N_2)$$

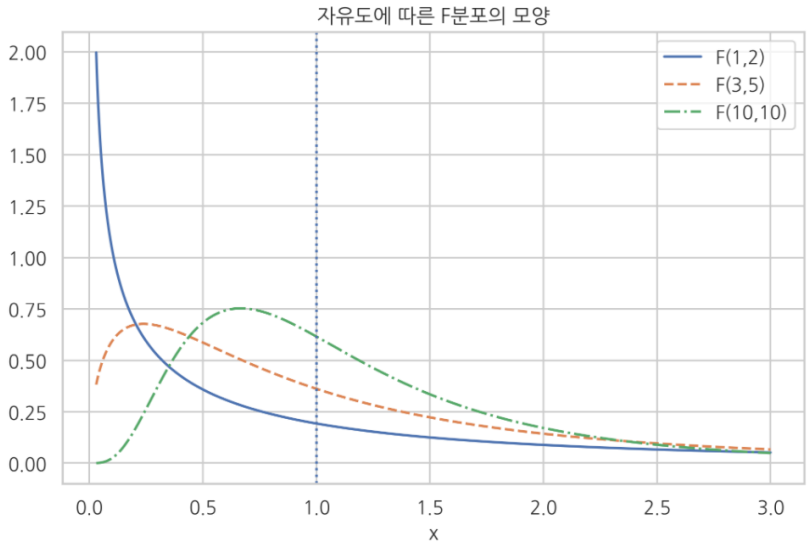
$$f(x; N_1, N_2) = \frac{\sqrt{\frac{(N_1 x)^{N_1} N_2^{N_-2}}{(N_1 x + N_2)^{N_1+N_2}}}}{x\ \mathrm{B}(\frac{N_1}{2}, \frac{N_2}{2})}, \qquad \mathrm{B}(x)\text{는 베타함수}$$

[파이썬] 사이파이의 stats.f() 사용

4. 연속형 확률분포

1) 연속형 확률분포(continuous probability distribution)

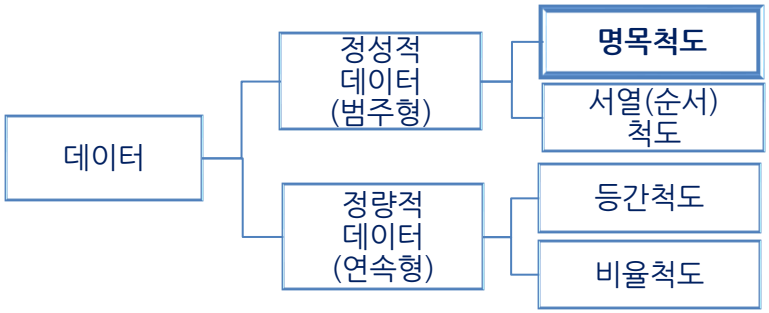
⑤ F 분포(F Distribution)



정리하기

1. 변수의 구분

- 변수 유형에 따른 분류



- 역할에 따른 변수의 분류

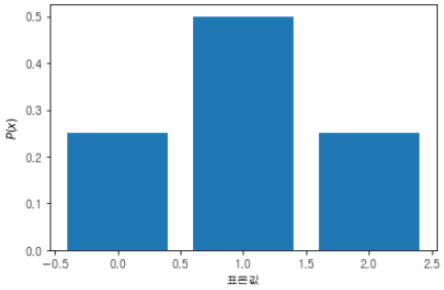
  인과관계(→), 상관관계(↔), 제 3의 변수와의 관계

정리하기

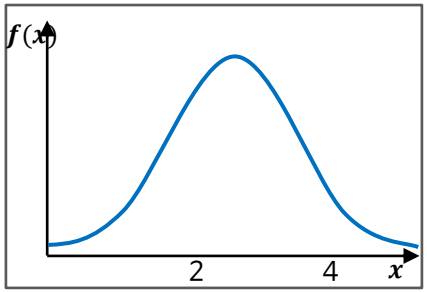
2. 분포특성의 표현

- 그래프를 통한 분포의 표현

범주형 확률변수



연속형 확률변수



- 기술통계량을 통한 분포의 표현
  - 중심 척도 : 평균, 중앙값, 최빈값, 합계
  - 산포측도 : 분위수, 최솟값, 최댓값, 범위, 사분위 범위, 분산, 변동계수
  - 분포측도 : 왜도(비대칭성), 첨도

정리하기

3. 확률분포의 종류

[이산형확률분포]

- **베르누이분포**
  - 베르누이 시행 결과를 0 또는 1로 표현한 베르누이 확률변수의 분포
  - 모수 : 성공확률(μ)
  - 파이썬 : 싸이파이의 stats.bernoulli()
- **이항분포**
  - 베르누이 시행을 N번 반복할 때, 성공횟수의 분포
  - 모수 : 반복횟수(N), 성공확률(μ)
  - 파이썬 : 싸이파이의 stats.binom(), pmf() 적용
- **카테고리분포**
  - 주사위 던지기와 같이 실험의 결과가 k개인 경우의 각 숫자가 나올 확률분포
  - 모수 : 각 결과의 성공확률(μ_k)
  - one-hot-endcoding을 사용하여 단순하게 표현
  - 파이썬 : 싸이파이의 stats.multinomial(), pandas.get_dumies() 적용

정리하기

3. 확률분포의 종류

[이산형확률분포]

- **다항분포**
  - 카테고리 확률변수의 반복 횟수가 여러 개인 경우의 분포
  - 모수 : 반복횟수(N), 성공확률(μ)
  - 파이썬 : 싸이파이의 stats.multinomial()
- **포아송분포**
  - 어떤 희귀한 현상의 발생횟수의 확률변수의 분포
  - 모수 : 단위시간에 발생하는 평균횟수(λ)
  - 파이썬 : 싸이파이의 stats.poisson()

정리하기

3. 확률분포의 종류

[연속형확률분포]

- **정규분포**
  - 자연 현상에서 가장 흔히 나타나는 이상적인 분포
  - 모수 : 평균(μ)과 분산(σ^2)
  - 파이썬 : 싸이파이의 stats.norm()
- **표준정규분포**
  - 평균이 0, 분산이 1인 정규분포
  - 모수 : 없음(고정값)
  - 파이썬 : 싸이파이의 stats.norm()
- **중심극한정리**

- '모집단이 평균이 μ, 표준편차가 σ인 임의의 분포를 따른다고 할 때,
  이 모집단으로부터 추출된 표본의 표본의 크기 N이 충분히 크면 표본평균들은 평균이 μ, 표준편차가 $\sigma/\sqrt{N}$인 정규분포에 근사한다.'

정리하기

3. 확률분포의 종류

[연속형확률분포]

- **t분포**
  - 정규분포와 유사하지만 양 끝단의 비중이 더 큰 분포
  - 모수 : 평균(μ), 정밀도 개념(λ), 자유도(ν)
  - 파이썬 : 싸이파이의 stats.t()
- **카이제곱분포**
  - 화률변수 X의 N개의 표본을 제곱하여 더한 값의 분포
  - 모수 : 자유도(ν)
  - 파이썬 : 사이파이의 stats.chi2()
- **F분포**
  - 카이제곱분포를 따르는 독립적인 두 확률변수 표본을 각각의 자유도(N)으로 나눈 비율값의 분포
  - 모수 : 자유도1(N_1), 자유도2(N_2)
  - 파이썬 : 싸이파이의 stats.f()