Introduction to Statistical Data Analysis

데이터

- 척도

종류		설명
수치형 데이터 (양적 데이터)	간격 척도	▪ 대소 관계 및 차이에도 의미 부여 ▪ 비율은 의미 없음 ▪ 예) 연도, 온도
	비례 척도	▪ 대소 관계, 차이, 비율 등에 모두 의미 부여 ▪ 예) 길이, 무게
범주형 데이터 (질적 데이터)	명의 척도	▪ 단순히 분류하기 위한 데이터로 변수의 동일성 여부에만 의미 부여 ▪ 대소관계, 차이, 비율 등은 의미 없음 ▪ 예) 학번, 전화번호, 성별
	순서 척도	▪ 순서 관계 또는 대소 관계에 의미 부여 ▪ 차이, 비율 등은 의미 없음 ▪ 예) 석차, 제품 만족도

- 비례 척도와 간격 척도를 구분하는 방법 : 0이 없음을 나타내는 지 확인
  → 길이 0은 길이가 없음을 의미(비례척도)
  → 온도 0은 온도가 없음을 의미하지 않지만(간격 척도)

데이터

- 종류

종류		설명
수치형 데이터 (양적 데이터)	연속형 데이터	▪ 두 값 사이에 무한한 개수의 값이 있는 숫자 ▪ 예) 부품 길이, 제품 중량
	이산형 데이터	▪ 두 값 사이에 셀 수 있는 개수의 값이 있는 숫자 ▪ 예) 고객 불만 수, 결점 또는 결함의 수
범주형 데이터 (질적 데이터)	순위형 데이터	▪ 정렬하거나 순위화 할 수 있는 데이터 ▪ 예) T-Shirt Size : XL > L > M
	명목형 데이터	▪ 정렬하거나 순위화 할 수 없는 데이터 ▪ 예) T-Shirt Color : Red, Green, Blue
	이진 데이터	▪ 범주의 값으로 두 개의 값 만을 갖는 특수한 경우 ▪ 예) 참/거짓, 0/1

기초 통계량

- 대푯값

종류	설명
평균	▪ 전체 데이터의 합을 데이터의 개수로 나눈 값 ▪ 편차와 분포를 반영하지 못하는 문제
중앙값	▪ 모든 데이터를 크기 순서로 정렬했을 때 가운데 위치한 값 ▪ 정도의 차이는 있으나 평균과 마찬가지로 분포를 반영하지 못하는 문제

- 산포도

종류	설명
분산	▪ 각 데이터와 평균 사이의 편차를 제곱한 값의 평균 ▪ 편차에 음의 값이 존재하고 편차의 평균이 0이 되므로 제곱의 평균 사용
분위 수	▪ 전체 데이터의 몇 %에 위치하는지 표시 ▪ 대표적으로 4분위 수 사용 (최소, 25%, 50%, 75%, 최대) ▪ 3Q(75%)와 1Q(25%)의 차이(범위)는 IQR(Interquartile Range) ▪ Q3 + IQR * 1.5, Q1 + IQR * 1.5를 이상 값의 기준으로 사용
표준 편차	▪ 분산의 제곱근을 구한 값 ▪ 개별 데이터와 같은 단위 사용

데이터 정규화

- 데이터를 통일된 지표로 변환하는 것

- 평균, 분산에 의존하지 않고 상대적인 위치 관계를 파악할 수 있는 지표

- 표준화
  - » (데이터 - 평균) / 표준편차
  - » 표준화된 데이터를 표준화 변량 또는 z-score로 표현

$$z_i = \frac{x_i - \bar{x}}{S}$$

- T Score
  - » 50 + 10 * (데이터 - 평균) / 표준편차
  - » 평균 50, 표준편차가 10이 되도록 정규화한 값
  - » z-score를 자연수와 백분위수로 표현해서 데이터의 가독성 향상

$$z_i = 50 + 10 \times \frac{x_i - \bar{x}}{S}$$

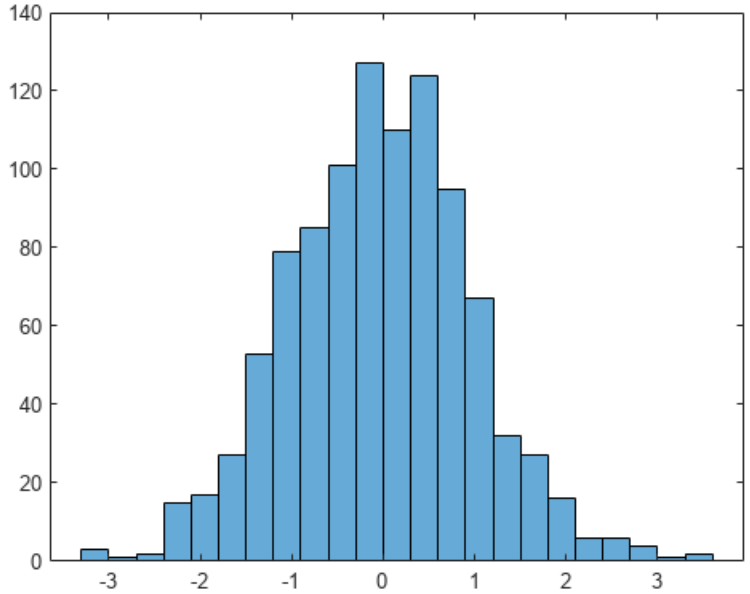
1차원 데이터 시각화

- 도수 분포표
  - » 전체 데이터를 몇 개의 구간으로 나누고 각 구간에 포함되는 데이터의 개수를 정리한 표

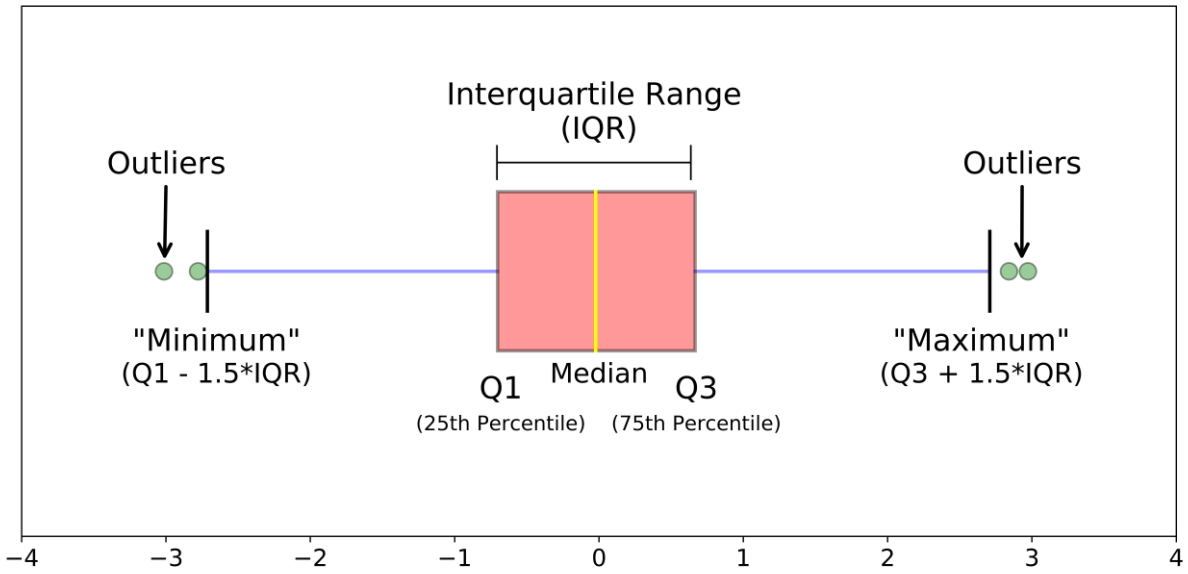
종류	설명
계급	▪ 데이터의 구간
도수	▪ 구간에 포함된 데이터 개수
상대도수	▪ 전체 데이터에서 해당 계급의 데이터가 차지하는 비율
누적도수/누적상대도수	▪ 시작 계급부터 해당 계급까지의 도수 또는 상대도수의 합
계급폭	▪ 각 구간의 크기
계급수	▪ 분할된 계급의 총 개수
계급값	▪ 계급을 대표하는 값으로 주로 계급의 중앙값 사용

1차원 데이터 시각화

- 히스토그램 (histogram)
  - » 도수분포표를 막대그래프로 표시
  - » 데이터의 분포 상태를 시각적으로 확인 가능



- 상자 그림 (box plot)
  - » 데이터의 산포도를 표현하는 시각화 도구
  - » 1사분위, 중앙값(2사분위), 3사분위 및 IQR 값을 사용해서 데이터의 분포와 이상값 시각화



두 데이터 사이의 관계 지표

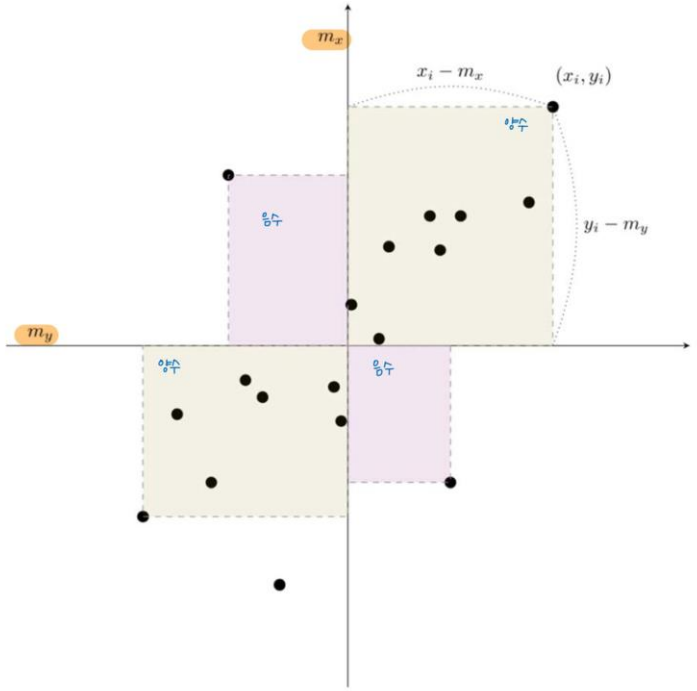
- 공분산
  - » 두 데이터의 편차를 곱한 값으로 선형 관계를 표현

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

$\sigma_{xy} > 0$: X와 Y가 양의 선형 관계

$\sigma_{xy} < 0$: X와 Y가 음의 선형 관계

$\sigma_{xy} = 0$: X와 Y는 선형적 관계를 갖지 않음

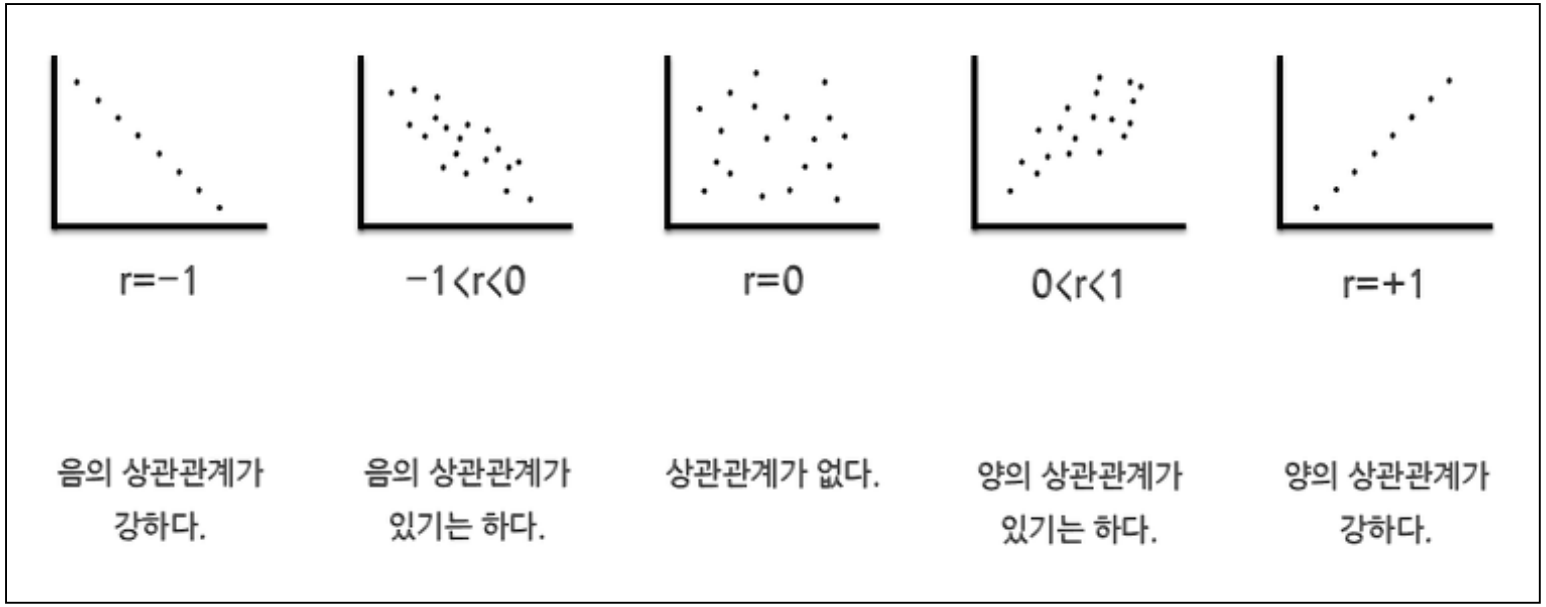


- 상관계수
  - » 공분산은 데이터의 단위에 영향을 받기 때문에 서로 다른 데이터의 상관성 정도를 비교하기 어려움
  - » 데이터의 단위에 영향 받지 않는 표준화된 관계 지표로 상관계수 사용
  - » 공분산을 각 데이터의 표준 편차로 나누어서 도출

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

두 데이터 사이의 관계 지표

- 상관계수 해석
  - » 상관관계의 정도를 파악하는 상관계수(Correlation coefficient)는 두 변수 사이의 연관 정도를 나타낼 뿐 인과관계를 설명하는 것은 아님
  - » 회귀분석을 통해 두 변수 사이에 존재하는 인과관계의 방향, 정도, 수학적 모델 확인 가능
  - » $0 < r \leq +1$ 이면 양의 상관, $-1 \leq r < 0$ 이면 음의 상관, $r = 0$이면 무상관을 의미
  - » 0인 경우 상관이 없다는 것이 아니라 선형의 상관관계가 아니라는 의미



두 데이터 사이의 관계 지표

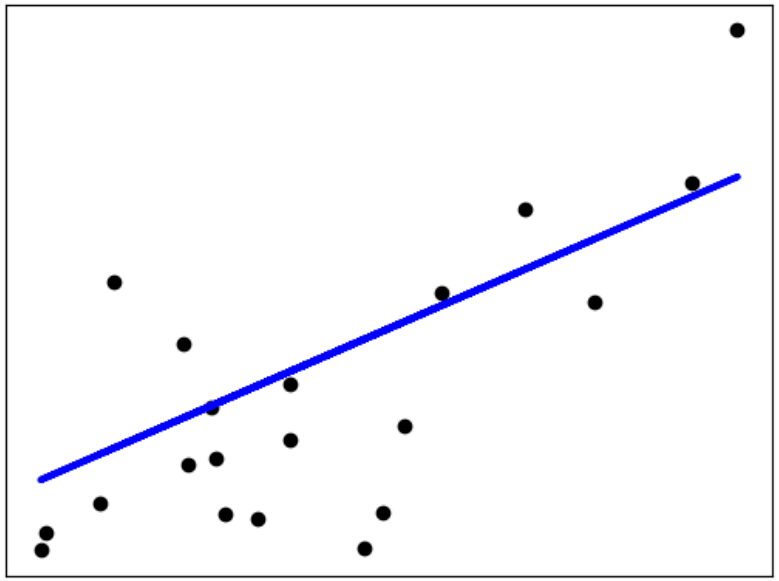
- 스피어만 상관계수
  - » 상관계수를 계산할 두 데이터의 실제 값 대신 두 값의 순위를 사용해 상관계수 비교
  - » 이산형 데이터 및 순서형 데이터 적용 가능
  - » 예)국어점수-영어점수 관계는 피어슨 / 국어석차-영어석차는 스피어만

- 켄달의 순위 상관계수
  - » (X, Y) 형태의 순서쌍 데이터에 대해 x1 < x2 에 대해 y1 < y2가 성립하면 concordant, 성립하지 않으면 discordant라고 정의

- 상관계수 검정
  - » scipy.stats 모듈의 pearsonr, spearmanr, kendalltau 함수를 사용해서 상관 계수의 유의성 판단
  - » 귀무가설은 상관계수가 0인 가설

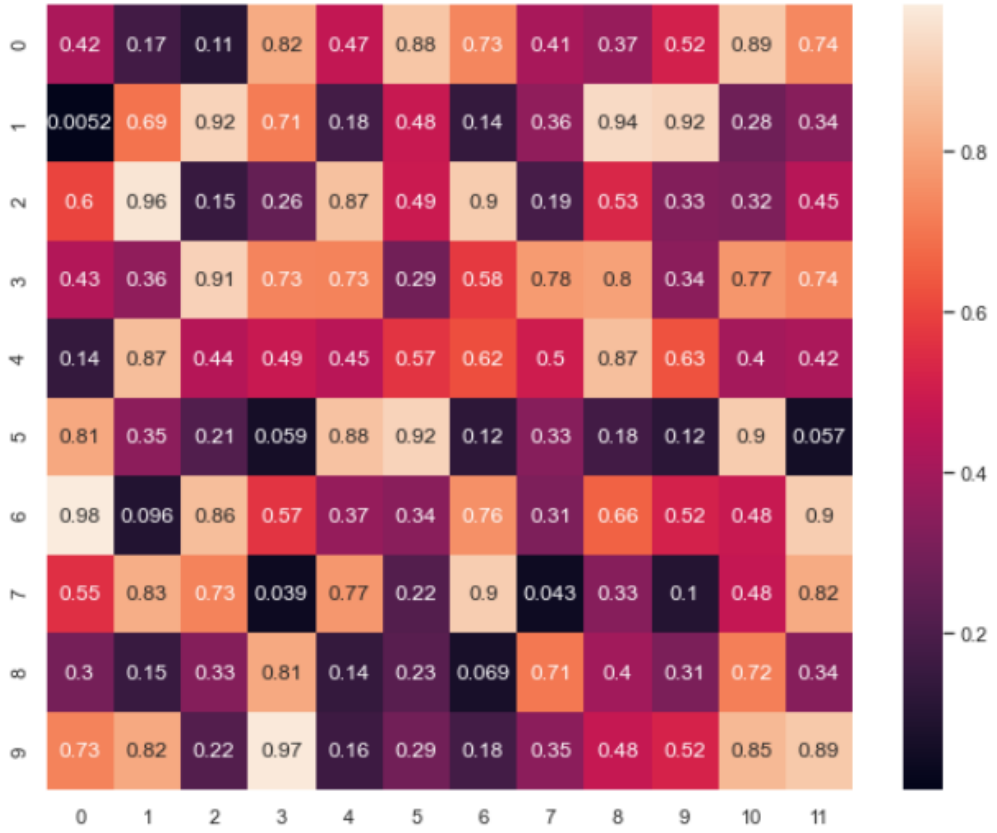
2차원 데이터 시각화

- 산점도 그래프와 회귀 직선



회귀 직선 → 두 데이터 사이의 선형 관계를 표현하는 시각화 도구

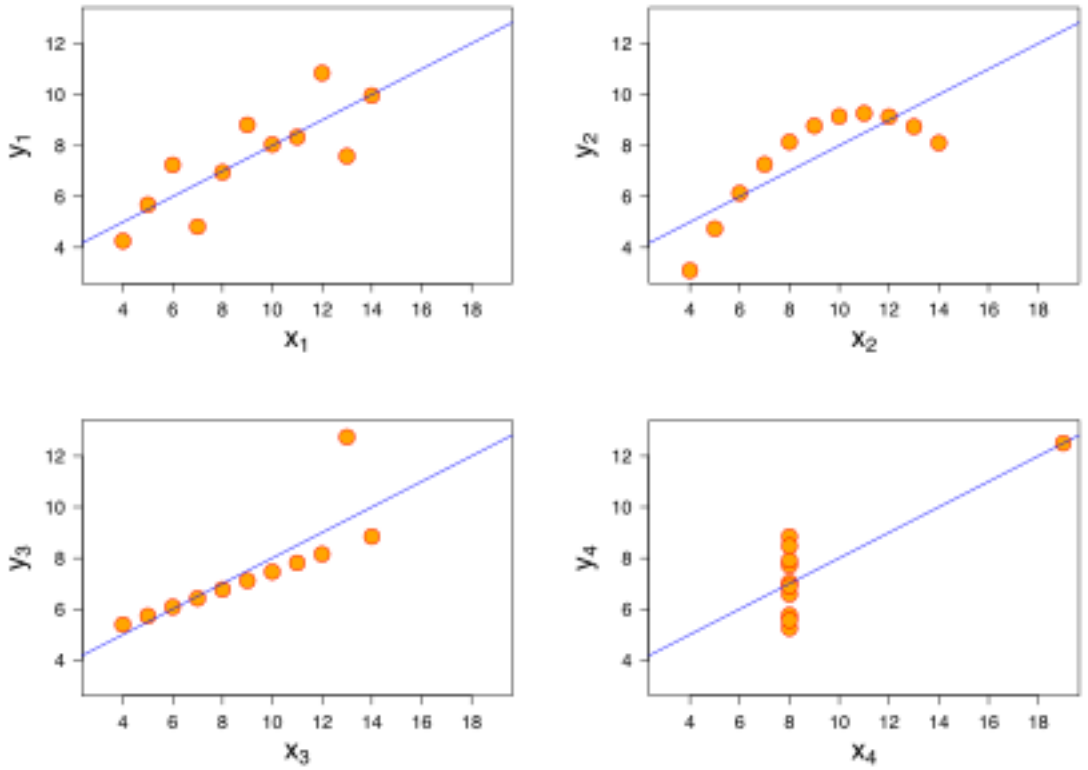
- 히트맵



히스토그램의 2차원 버전으로 색을 이용해 표현하는 시각화 도구

Anscombe’s quartet

- 평균, 표본분산, 선형회귀선, 결정계수 등의 기술 통계량은 동일하지만 분포나 그래프를 이용하여 시각화 하면 전혀 다른 특성이 나타나는 4개의 데이터 세트
- 시각화의 중요성을 보여주기 위한 예시



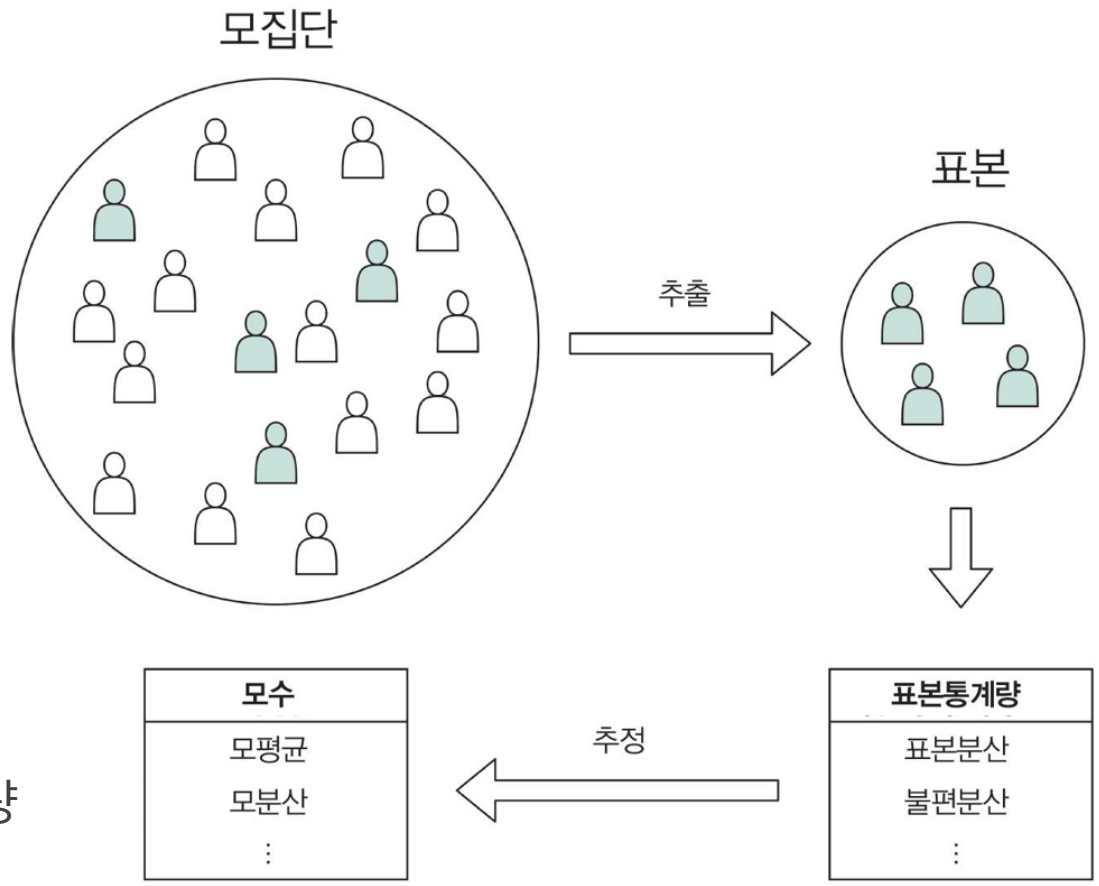
기술 통계와 통계적 추론

- 기술 통계
  - » 수집한 데이터의 정리, 요약, 해석, 표현 등을 통해 데이터의 특성을 규명하는 통계적 방법
  - » 평균, 표준편차 등의 수치와 히스토그램, 박스플롯 등의 시각화 표현 활용

- 통계적 추론
  - » 모집단에 대한 어떤 미지의 양상을 알기 위해 통계학을 이용하여 추측하는 과정
  - » 추정(estimation)
    - › 표본을 통해 모집단 특성을 추측하는 과정
    - › 하나의 값을 추정하는 것은 점추정, 구간으로 추정하는 것은 구간추정
    - › 예) 표본평균 계산을 통해 모집단평균을 추측
  - » 가설검정(testing hypothesis)
    - › 표본 데이터를 이용해서 모집단에 대한 주장(가설)이 맞는지 판정하는 과정
    - › 모집단의 통계적 성질에 대한 가설을 세우고 그 가설의 참/거짓을 통계적으로 판정

모집단과 표본

- 모집단
  - » 추측하고 싶은 관측 대상 전체
- 표본
  - » 추측에 사용하는 관측 대상의 일부분
- 표본 추출
  - » 모집단에서 표본을 골라내는 작업
- 모수
  - » 모집단의 평균, 분산, 상관계수 등의 통계량
- 표본 통계량
  - » 표본을 바탕으로 계산한 평균, 분산, 상관계수 등의 통계량



표본추출 방법

- 무작위추출
  - » 모집단을 잘 반영할 수 있도록 편향되지 않은 임의의 표본 데이터 추출

- 복원추출
  - » 한 번 뽑은 데이터를 다시 추출할 대상에 복원한 후 다음 데이터를 추출하는 방법
  - » 같은 데이터를 여러 번 뽑을 수 있는 방법

- 비복원추출
  - » 한 번 뽑은 데이터를 추출할 대상에서 제거하고 다음 데이터를 추출하는 방법
  - » 같은 데이터는 한 번만 뽑는 방법

확률 모형

- 무작위추출은 실행해볼 때까지는 어떤 결과가 나올지 알 수 없으며 실행할 때마다 다른 결과가 도출되는 불확정성이 있음

- 불확정성을 수반한 현상을 해석하기 위해 확률 사용

- 확률을 사용한 무작위추출을 통계적(수학적)으로 모델링한 것 → **확률 모형**

확률

- 확률변수
  - » 항상 결과를 정확하게 맞힐 수는 없지만 결과값이 나올 확률이 결정되어 있는 변수
  - » 확률적인 결과에 따라 결과값이 바뀌는 변수
  - » 일정한 확률에 따라 일어나는 사건에 수치가 부여된 것

- 시행
  - » 확률변수의 결과를 관측하는 것
  - » 시행에 의해 관측되는 값 → 실현 값
  - » 시행 결과로 나타날 수 있는 값 → 사건
  - » 두 개 이상의 사건이 동시에 발생할 수 없는 경우 → 상호배반
  - » 확률은 사건에 대해 정의
    - › 주사위 눈이 1인 사건에 대한 확률 $1/6$, 주사위 눈이 홀수인 사건에 대한 확률 $1/2$

확률분포

- 정의
  - » 확률변수가 어떻게 움직이는지 나타낸 것
  - » 확률변수가 특정한 값을 가질 확률을 나타내는 함수
  - » 주사위 눈 사례

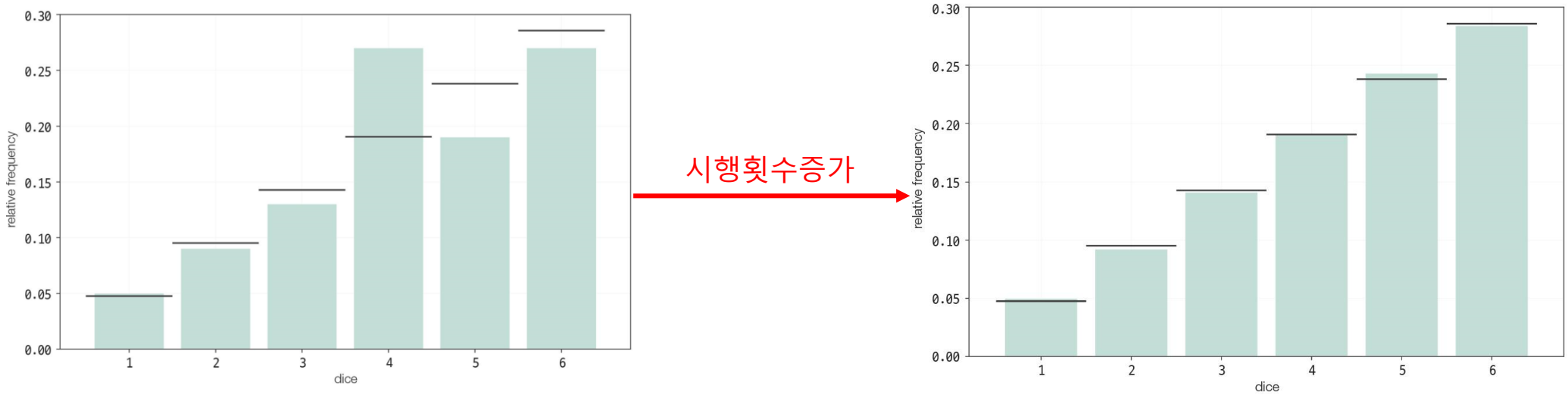
동일확률

눈	1	2	3	4	5	6
확률	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

불공정확률

눈	1	2	3	4	5	6
확률	$\frac{1}{21}$	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$	$\frac{6}{21}$

- 시행 횟수가 증가하면 관찰된 값의 상대도수는 확률분포에 수렴



통계적 추론에서 확률

- 무작위추출로 얻은 표본을 모집단의 확률분포를 따르는 확률 변수로 간주
  - » 통계적 추론에서 다루는 데이터는 확률 변수의 실현 값 (관측 값)
  - » 모집단으로부터 무작위추출을 수행하면 모집단의 각 값의 확률분포에 따라 표본데이터 발생

- 표본의 크기가 커지면 (무작위추출 시행 횟수가 많아지면) 표본 데이터의 상대도수는 실제의 확률분포에 근사

- 표본 하나하나가 확률변수이므로 표본들의 평균으로 계산되는 표본평균도 확률변수
  - » 1. 무작위추출로 표본 크기가 n개인 표본을 추출하고
  - » 2. 표본평균을 계산하는 작업을 여러 번 실행해서 얻은
  - » 3. 표본평균 값들의 평균은 모평균에 근사하고
  - » 4. 표본의 크기가 충분히 크면 모평균을 중심으로 정규분포에 근사

확률분포

- 확률변수가 어떤 종류의 값을 가지는가에 따라 이산 확률 분포와 연속 확률 분포로 구분

- 이산 확률 분포
  - » 확률 변수가 가질 수 있는 값이 셀 수 있는 제한된 개수인 경우
  - » 확률 질량 함수를 통해 표현
  - » 누적 분포 함수로 표현할 경우 비약적 불연속으로 증가
  - » 종류 → 이산균등 분포, 푸아송 분포, 베르누이 분포, 이항 분포, 다항 분포 등

- 연속 확률 분포
  - » 확률 변수가 가질 수 있는 값이 연속형인(셀 수 없는 무제한의 개수인) 경우
  - » 확률 밀도 함수를 이용해 표현
  - » 종류 → 정규 분포, 연속 균등 분포, 카이제곱 분포 등

이산형 확률 분포

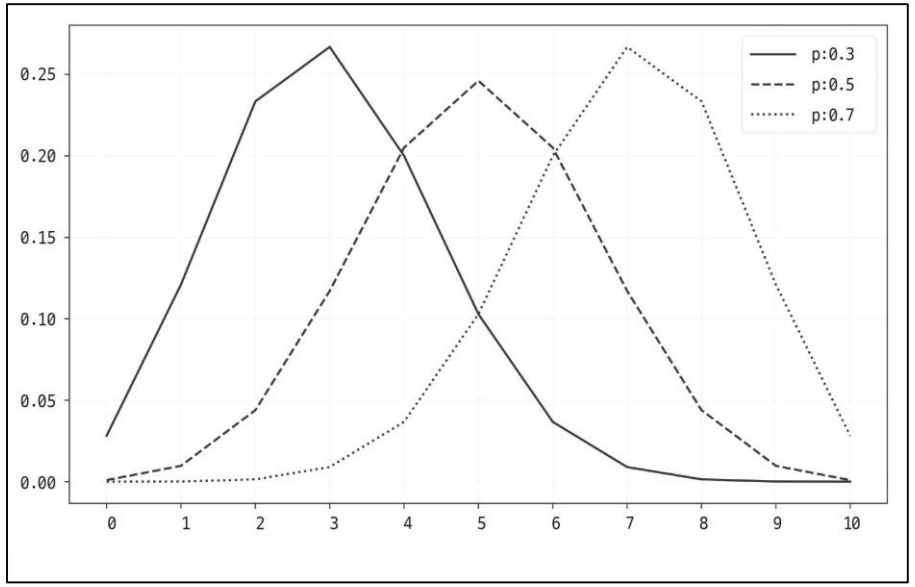
- 베르누이 분포
  - » 확률변수가 취할 수 있는 값이 0과 1밖에 없는 분포

파라미터	p
취할 수 있는 값	$\{0, 1\}$
확률함수	$p^x(1-p)^{(1-x)}$
기댓값	p
분산	$p(1-p)$
scipy.stats	bernoulli(p)

이산형 확률 분포

- 이항분포
  - » 성공 확률이 p인 베르누이 시행을 n번 했을 때 성공 횟수가 따르는 분포

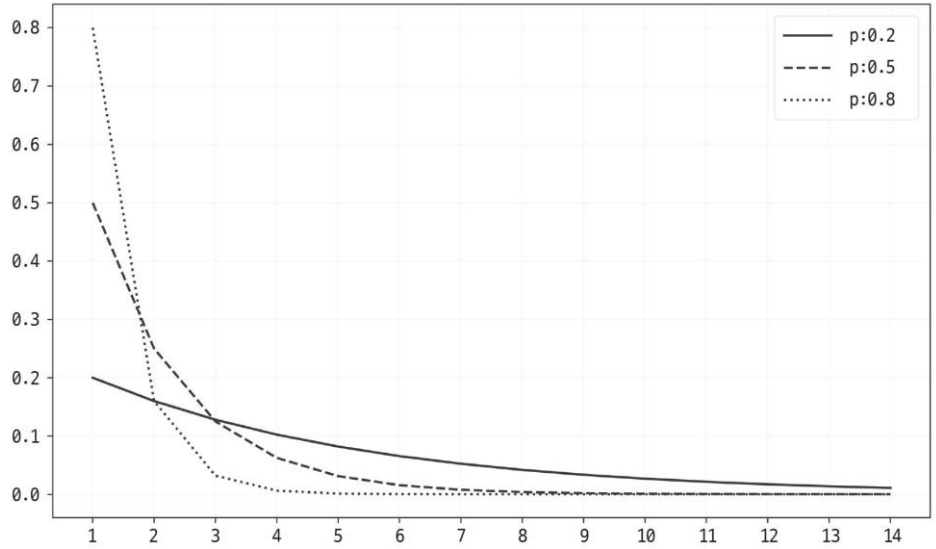
파라미터	n, p
취할 수 있는 값	$\{0, 1, \cdots, n\}$
확률함수	${}_nC_x p^x (1-p)^{(1-x)}$
기댓값	np
분산	$np(1-p)$
scipy.stats	binom(n, p)



이산형 확률분포

- 기하분포
  - » 베르누이 시행에서 처음 성공할 때까지 반복한 시행 횟수가 따르는 분포

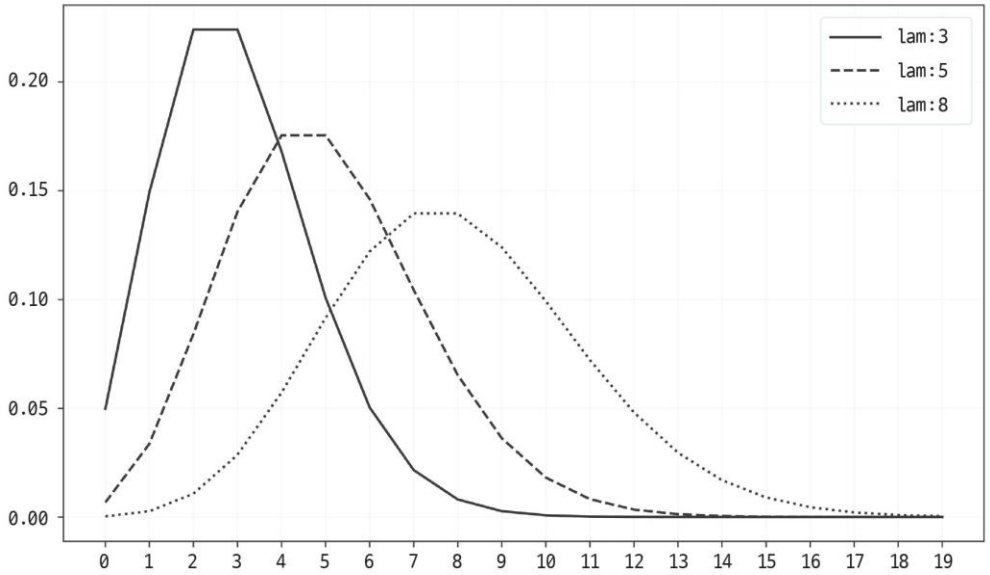
파라미터	p
취할 수 있는 값	$\{1, 2, 3, \cdots\}$
확률함수	$(1-p)^{(x-1)}p$
기댓값	$\frac{1}{p}$
분산	$\frac{(1-p)}{p^2}$
scipy.stats	geom(p)



이산형 확률분포

- 포아송분포
  - » 임의의 사건이 단위 시간당 발생하는 건수가 따르는 확률분포

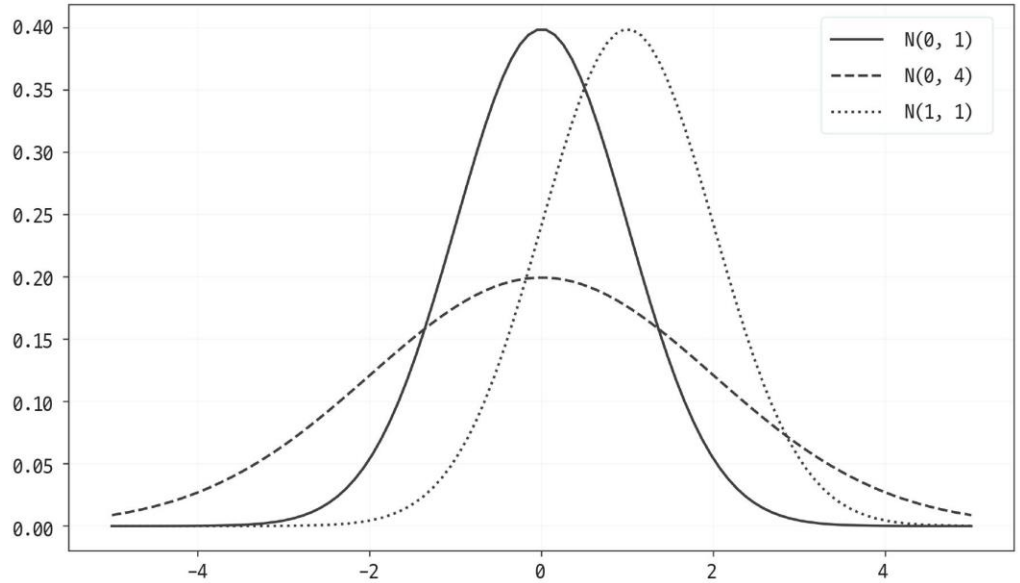
파라미터	p
취할 수 있는 값	$\{0, 1, 2, \cdots\}$
확률함수	$\frac{\lambda^x}{x!} \cdot e^{-\lambda}$
기댓값	λ
분산	λ
scipy.stats	poisson(λ)



연속형 확률분포

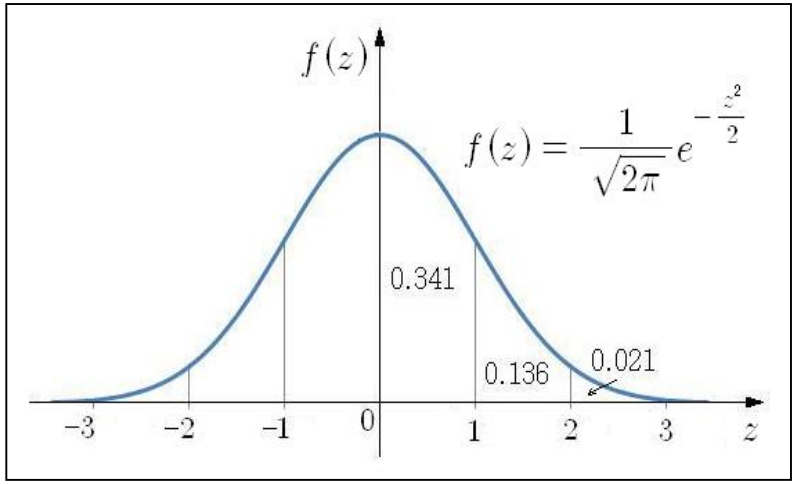
- 정규분포 (가우스분포)
  - » 자연 현상에서 나타나는 숫자를 확률 모형으로 설명할 때 많이 사용
  - » 평균과 표준편차 2개의 매개변수에 의해 모양 결정

파라미터	μ, σ
취할 수 있는 값	실수 전체
밀도함수	$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
기댓값	μ
분산	σ^2
scipy.stats	norm(μ, σ)



연속형 확률분포

- 표준정규분포
  - » -(무한대) ~ +(무한대)의 모든 수치 데이터로 구성
  - » 데이터의 평균 값을 기준으로 좌.우 대칭형 분포
  - » 평균은 0, 표준편차는 1
  - » 표준편차의 1배 범위의 상대도수는 0.6826, 2배 범위의 상대도수는 0.9545
  - » 일반적으로 1.96배 사용 → 상대도수 0.95



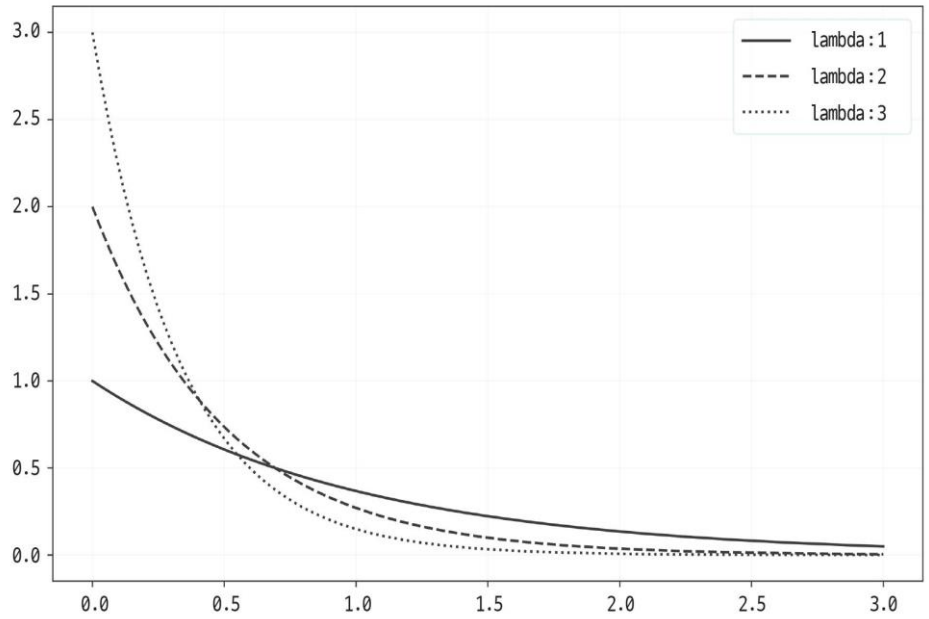
- 일반정규분포
  - » 표준정규분포의 데이터에 표준편차를 곱한 후 평균을 더한 데이터 분포
    - › X = Zσ + μ
  - » 일반정규분포 데이터를 표준정규분포 데이터로 변환

$$Z = \frac{X - \mu}{\sigma}$$

연속형 확률분포

- 지수분포
  - » 어떤 사건이 발생하는 간격이 따르는 분포
  - » 간격이라는 시간이 따르는 분포이므로 확률변수가 취할 수 있는 값은 0 이상의 실수

파라미터	λ
취할 수 있는 값	양의 실수
밀도함수	$\lambda e^{-\lambda x}$
기댓값	$\frac{1}{\lambda}$
분산	$\frac{1}{\lambda^2}$
scipy.stats	$\texttt{expon(scale=}\frac{1}{\lambda}\texttt{)}$

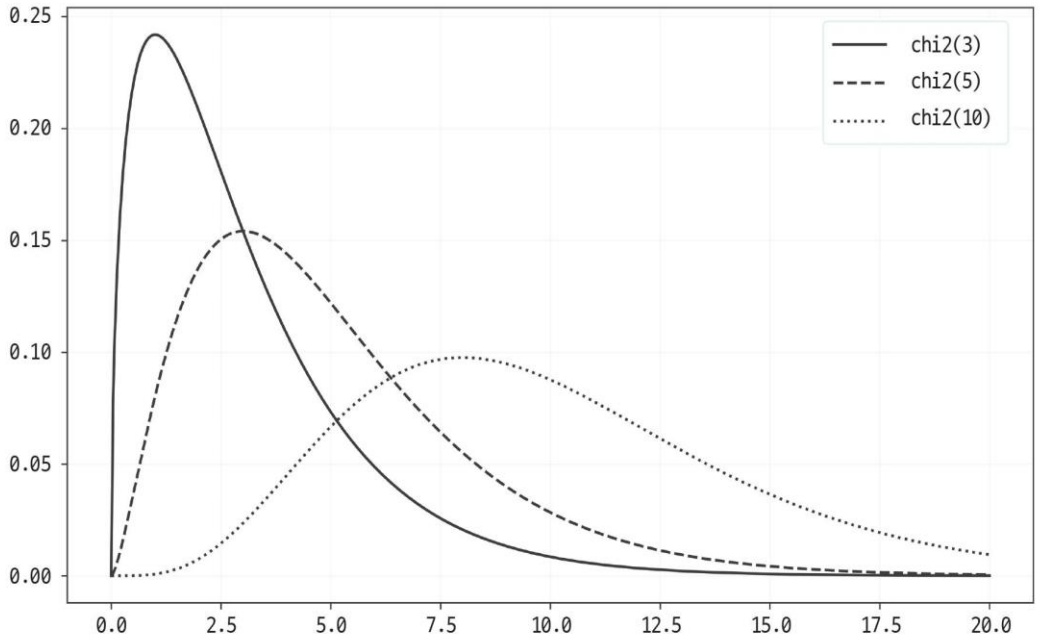


연속형 확률분포

- 카이제곱분포
  - » 분산의 구간추정 또는 독립성 검정에 사용되는 확률분포
  - » K개의 서로 독립적인 표준정규 확률변수를 각각 제곱한 다음 합해서 얻어지는 분포
  - » 좌우비대칭으로 왼쪽으로 치우치고 오른쪽으로 넓어짐
  - » 자유도가 커지면 좌우대칭에 가까워짐
  - » 자유도의 값 가까이에 분포의 정점 위치

$$Q = \sum_{i=1}^{k} X_i^2$$

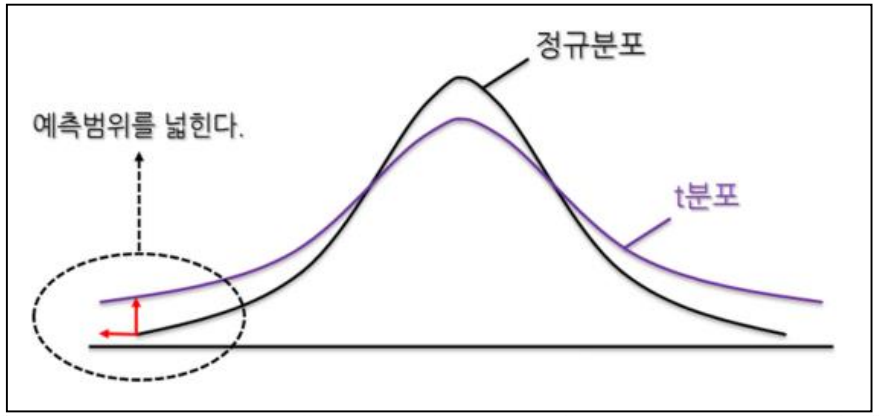
파라미터	n
취할 수 있는 값	음수가 아닌 실수
scipy.stats	chi2(n)

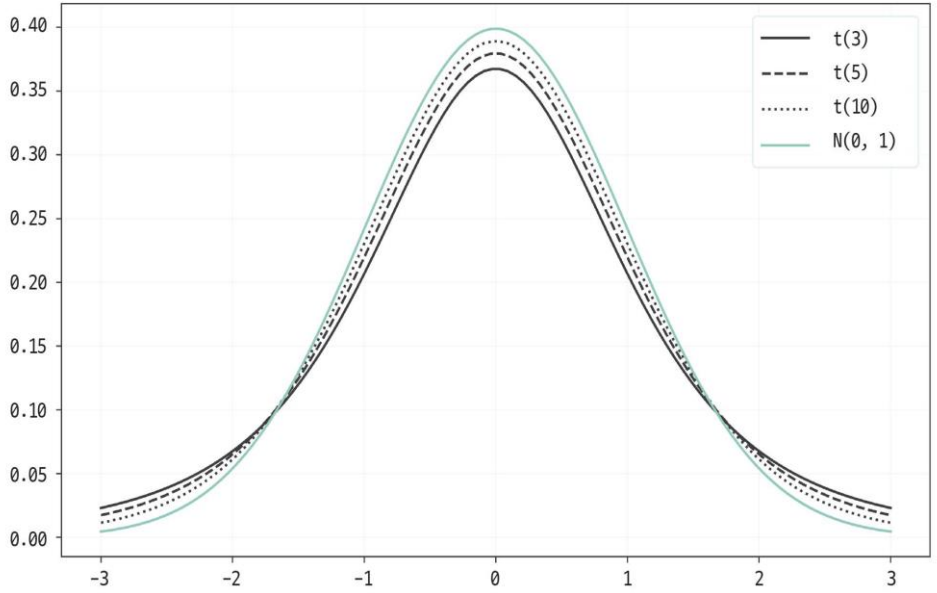


연속형 확률분포

- t분포
  - » 정규분포에서 모평균의 구간추정 등에 사용하는 확률분포
  - » 좌우대칭인 분포
  - » 표준정규분포보다 양쪽 끝이 두꺼움
  - » 도가 커지면 표준정규분포에 가까워짐

파라미터	*n*
취할 수 있는 값	실수전체
scipy.stats	*t(n)*

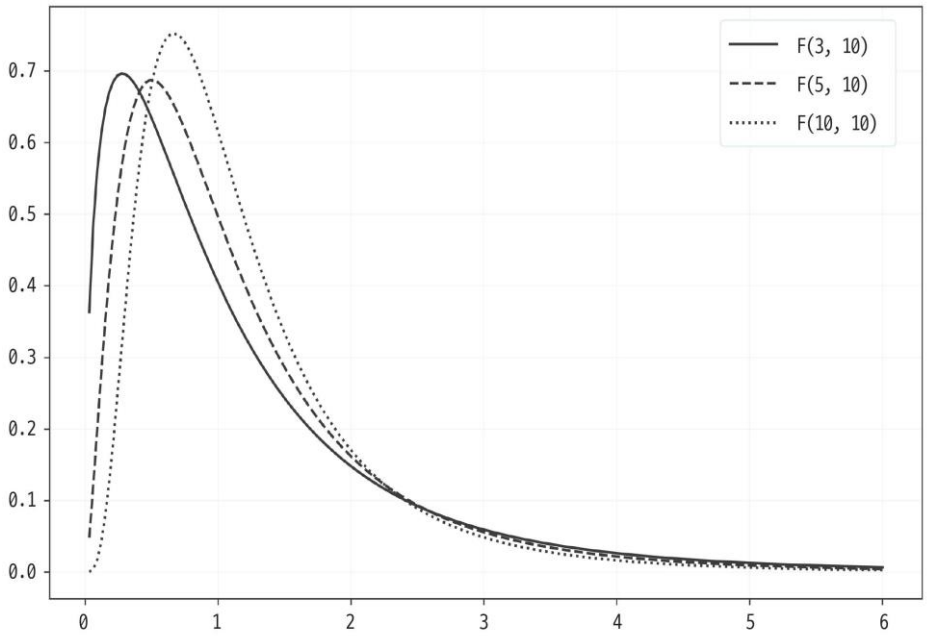




연속형 확률분포

- f분포
  - » F-검정, 분산분석 등에서 사용되는 확률분포
  - » 좌우비대칭으로 왼쪽으로 치우치고 오른쪽으로 넓은 분포
  - » 분포의 정점은 0에 가까움

파라미터	$n_1,\ n_2$
취할 수 있는 값	음수가 아닌 실수
scipy.stats	f $(n_1,\ n_2)$



독립성

- 확률변수의 독립성이란 2개 이상의 확률변수가 서로 영향을 끼치지 않으며 관계가 없음을 의미

- 2차원 확률변수 X, Y에 대해 다음과 같은 관계가 성립할 때 X와 Y는 서로 독립

$$f(x, y) = f_X(x) f_Y(y)$$

확률변수가 독립일 때 결합확률분포는 주변확률분포의 곱과 동일

- N차원 확률변수 X1, X2, …, Xn에 대해 다음을 만족할 때 X1, X2, …, Xn은 서로 독립

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

함수 f가 이산형이면 확률함수, 연속형이면 밀도함수 의미

- 독립성과 무상관성
  - » 두 확률변수가 서로 관계 없음을 의미
  - » 공분산이나 상관계수가 0인 경우 무상관이라 하고 이는 두 확률변수 사이의 선형 관계가 없다는 의미
  - » 독립인 경우 무상관이 되지만 무상관인 경우 항상 독립이 되지는 않음

표본평균의 분포

- 확률변수 $X_1, X_2, \cdots, X_n$ 의 표본평균 $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ 이 따르는 확률분포
- 모평균의 구간추정 또는 모평균의 검정 등에서 사용하는 분포
- 중심극한정리, 대수의 법칙 등 통계적 추론에서 중요한 성질 포함
- 표본평균의 기댓값과 분산
  - » 확률변수 $X_1, X_2, \cdots, X_n$ 이 서로 독립이고 기댓값이 μ, 분산이 σ^2 인 확률분포를 따를 때

$$
\begin{aligned}
E(\overline{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} \\
&= \frac{n\mu}{n} \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
V(\overline{X}) &= V\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{V(X_1) + V(X_2) + \cdots + V(X_n)}{n^2} \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

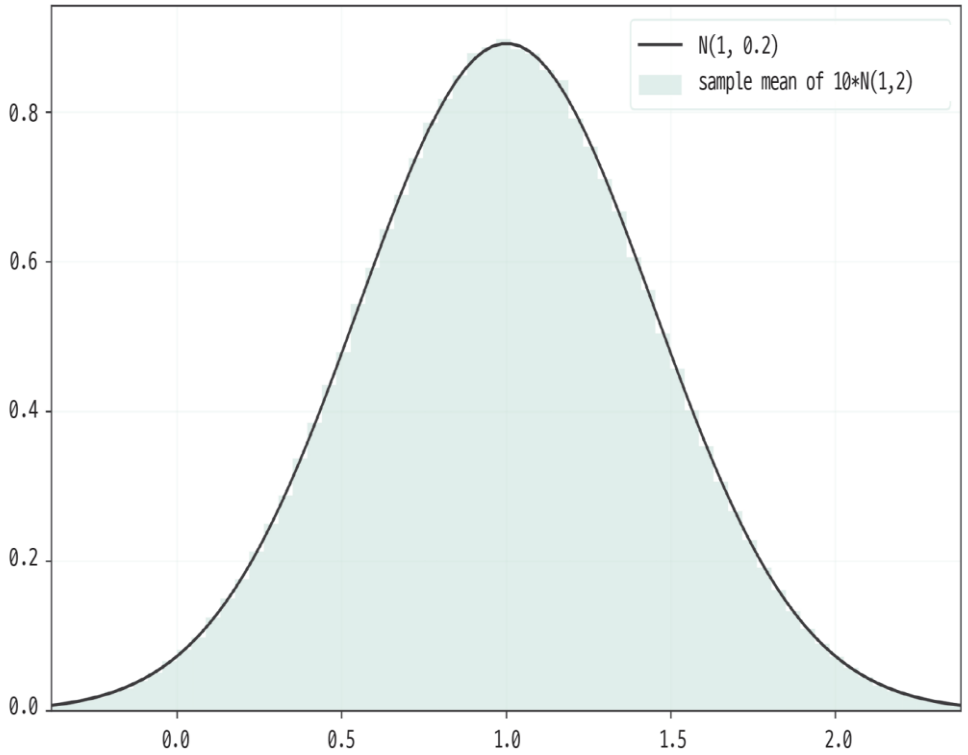
$$E(\overline{X}) = \mu$$

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

표본평균의 분포

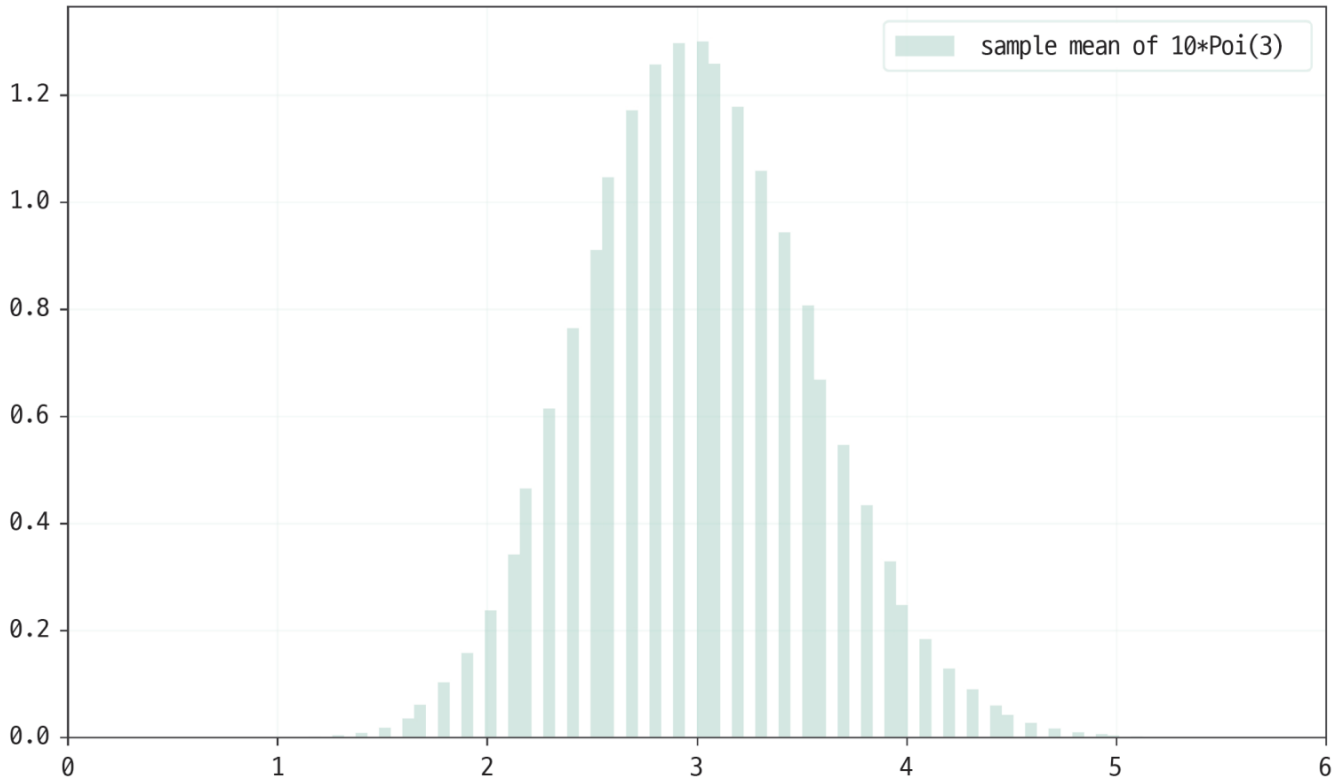
- 정규분포의 표본평균 분포도 정규분포
  - » 평균 μ, 분산 σ^2 인 정규분포의 표본평균의 분포는 다음과 같음

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



표본평균의 분포

- 포아송 분포의 표본평균의 분포는 정규분포에 근사



중심극한정리

- 확률 변수 X1, X2, …, Xn이 서로 독립이고 기댓값이 μ, 분산이 σ^2 인 확률분포를 따를 때 n이 커짐에 따라 표본평균의 분포는 다음의 정규분포에 가까워짐

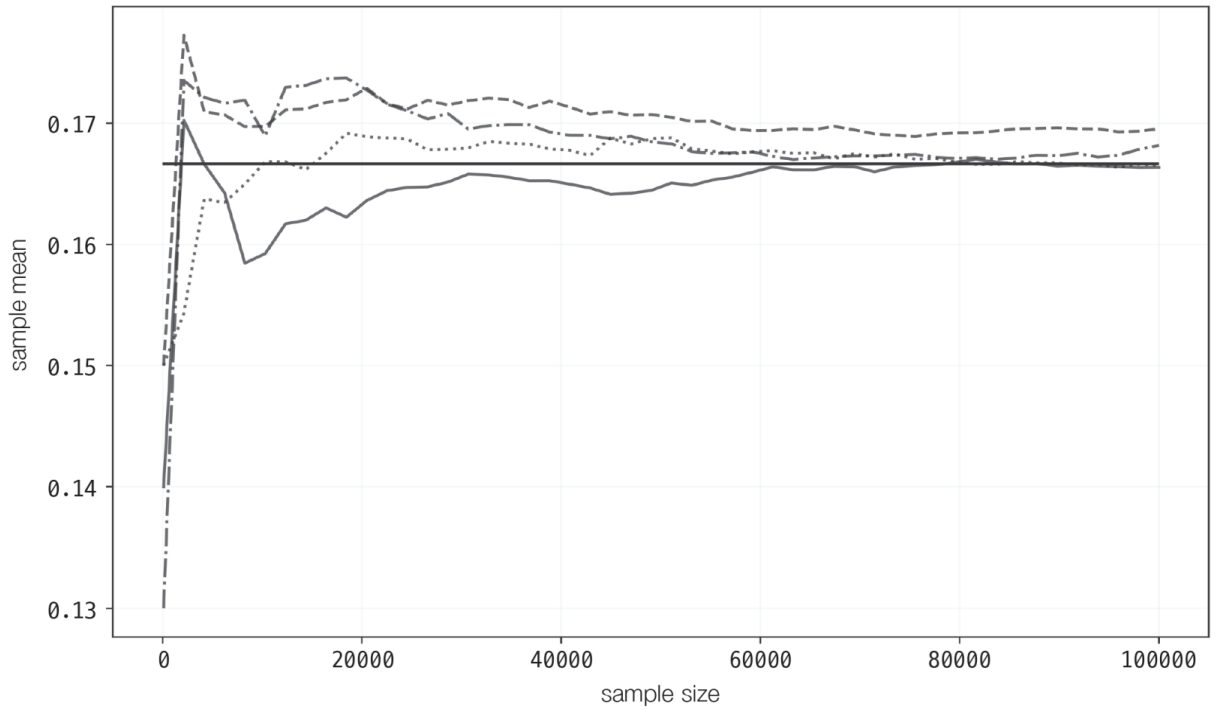
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- 동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 충분히 크다면 정규분포에 가까워진다는 정리

- 원래 분포와 상관 없이 표본평균의 분포는 정규분포에 가까워진다는 것

대수의 법칙

- 큰 수의 법칙
- 표본의 크기가 커지면 표본평균은 모평균에 수렴하는 원리



통계적 가설검정

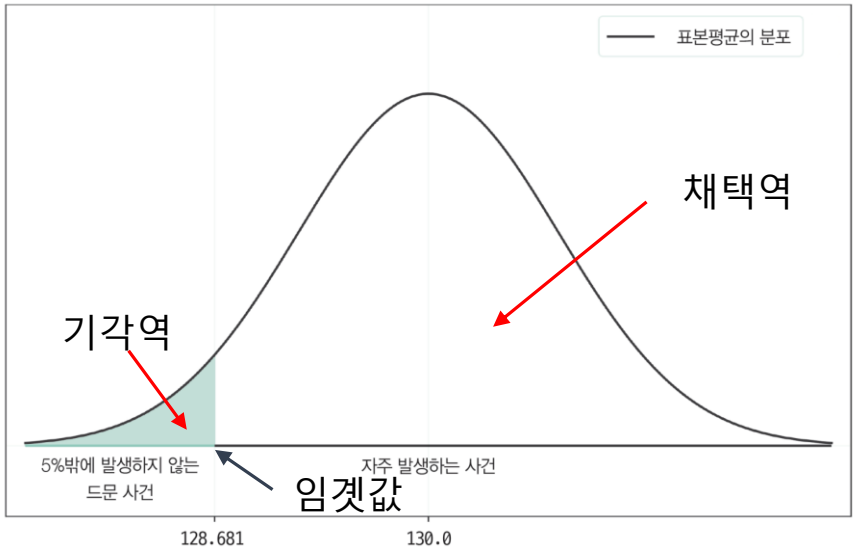
- 모집단의 모수에 관하여 두 가지 가설을 세우고 표본으로부터 계산되는 통계량을 이용하여 어느 가설이 옳은지 판단하는 통계적 기법

- 용어

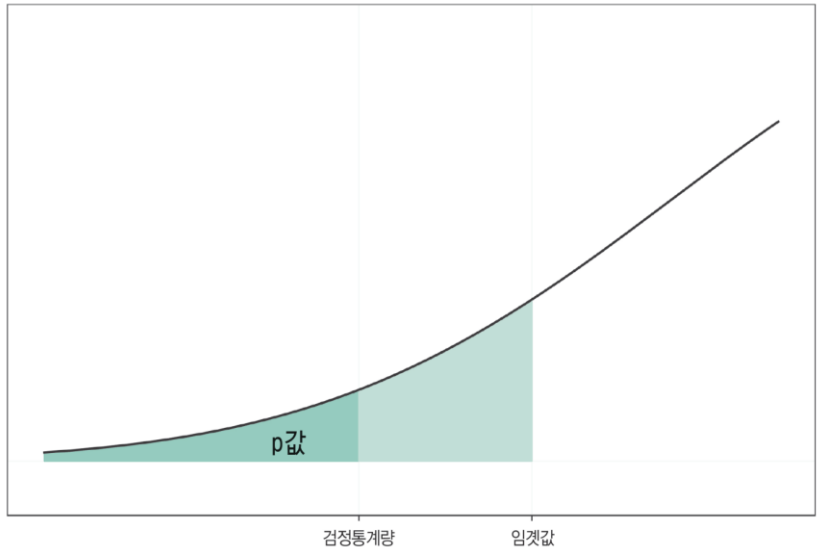
종류	설명	종류	설명
검정 통계량	▪ 검정에 사용되는 통계량	기각역	▪ 귀무가설이 기각되는 구간
귀무가설 (H_0)	▪ 주장하는 가설의 반대 가설 ▪ 예) 차이가 없다, 효과가 없다 등	채택역	▪ 귀무가설이 채택되는 구간
대립가설 (H_1)	▪ 주장하는 가설 ▪ 예) 차이가 있다, 효과가 있다 등	유의수준	▪ 기각역의 면적 (확률)
귀무가설 기각	▪ 귀무가설이 옳지 않음	p-value	▪ 검정통계량의 면적 (확률)
귀무가설 채택	▪ 귀무가설이 옳지 않다고 판단할 수 없음	제1종 오류	▪ 귀무가설이 옳을 때 귀무가설을 기각하는 오류
기각역	▪ 귀무가설이 기각되는 구간	제2종 오류	▪ 대립가설이 옳을 때 대립가설을 기각하는 오류
채택역	▪ 귀무가설이 채택되는 구간		

통계적 가설검정

- 표본으로부터 통계량을 계산하여 귀무가설과 대립가설의 검정 수행
  - » 검정 통계량의 발생 확률(p-value)이 유의수준보다 낮은 경우 귀무가설 기각



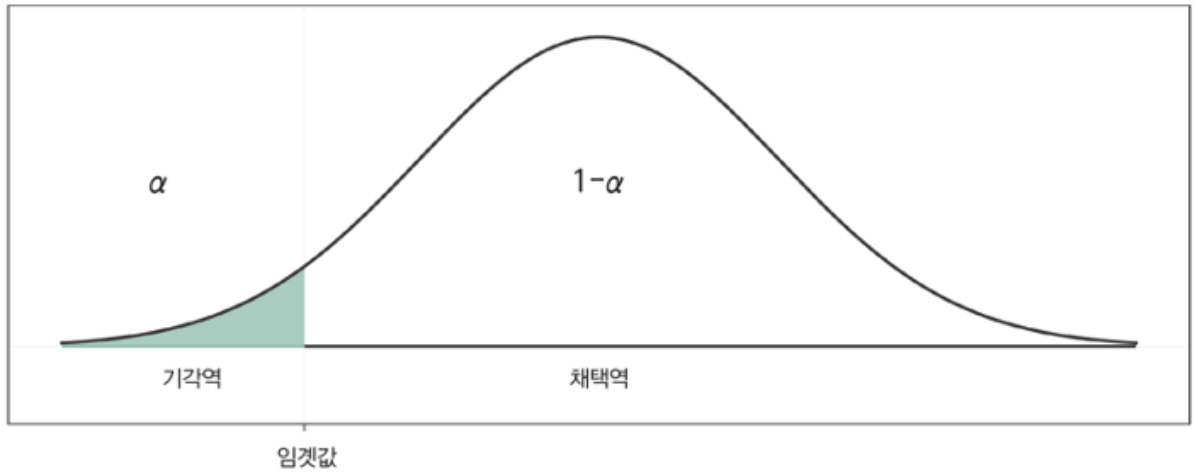
유의수준 5%

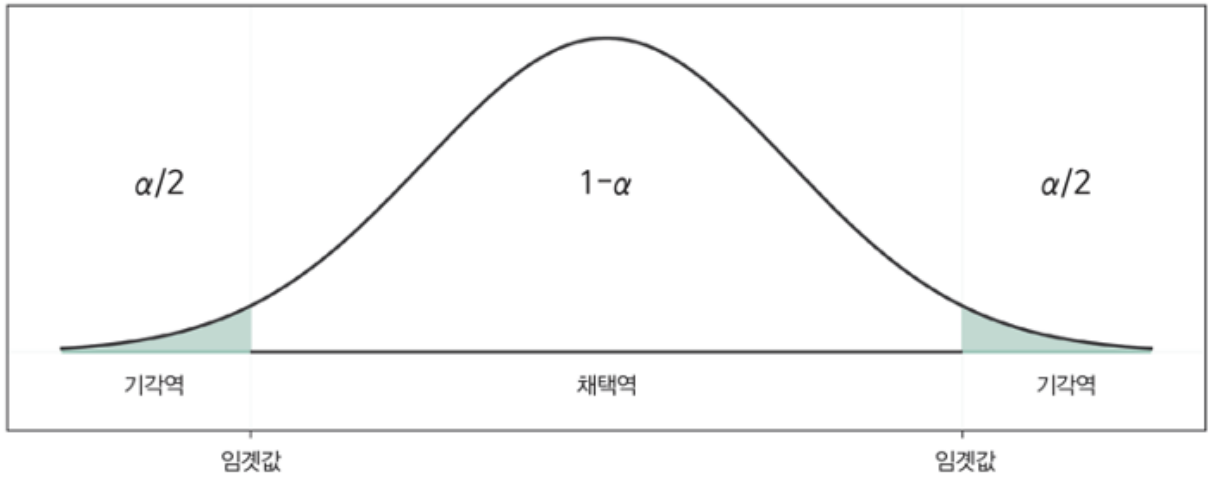


검정통계량: 표본평균

통계적 가설 검정

- 단측 검정과 양측 검정





1표본 t검정

- 모분산을 알지 못하는 상황에서 정규분포의 모평균에 대한 검정

- 검정통계량
  - » 자유도가 n-1인 t분포를 따르는 통계량

$$t = (\overline{X} - \mu_0) / \sqrt{\frac{s^2}{n}}$$

- 가설검정
  - » 귀무가설 기각 → $t < t_{1-\alpha/2}(n-1)$ 또는 $t_{\alpha/2} < t$
  - » 귀무가설 채택 → $t_{1-\alpha/2}(n-1) \leq t \leq t_{\alpha/2}(n-1)$

2표본 문제에 관한 가설검정

- 2 모집단에 관한 다양한 관계성 검정

- 종류

	정규분포를 가정할 수 있음	정규분포를 가정할 수 없음
대응표본	대응비교 t 검정	윌콕슨의 부호순위검정
독립표본	독립비교 t 검정	만 · 위트니의 U 검정

» 대응표본
› 두 데이터에서 서로 대응하는 동일한 개체에 대해 각각 다른 조건으로 측정한 표본
› 예) 피검자에게 약을 투여하기 전후에 측정한 두 데이터
» 독립표본
› 두 데이터에 서로 대응이 없는 표본
› 예) 두 고등학교의 수학 시험 성적

2표본 문제에 관한 가설검정

- 대응비교 t 검정
  - » 대응하는 데이터가 있고 데이터 차이에 정규분포를 가정할 수 있는 경우의 평균값 차이에 대한 검정

- 독립비교 t 검정
  - » 대응하는 데이터가 없고 독립된 2표본 모집단에 정규분포를 가정할 수 있는 경우 평균값의 차이에 대한 검정

- 윌콕슨의 부호순위검정
  - » 대응표본에서 차이에 정규분포를 가정할 수 없는 경우 중앙값의 차이에 대한 검정
  - » 절댓값이 작은 것부터 순서대로 부여된 순위에 의해 검정 수행

- 만.위트니의 U 검정
  - » 대응되는 데이터가 없는 2표본 모집단에 정규분포를 가정할 수 없는 경우 중앙값의 차이에 대한 검정
  - » 윌콕슨의 순위합검정

독립성 검정

- 범주형 두 변수 X와 Y에 관해서 "X와 Y가 독립이다" 라는 귀무가설과 "X와 Y가 독립이 아니다" 라는 대립가설을 기반으로 수행하는 검정

- 카이제곱분포 사용 → 카이제곱검정

- 교차표 작성

광고	구입 하지 않았다	했다
A	351	49
B	549	51

- 검정통계량

$$Y = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

→ Y는 카이제곱분포를 근사적으로 따름