LG 전자 고급 빅데이터 전문가 과정
확률통계 특강

임채영

서울대학교 통계학과

강의에서 다룰 내용

- 확률의 개념 - 확률의 공리, 조건부 확률, 확률의 법칙, 독립, 베이즈정리
- 확률변수 및 확률분포 - 확률변수 (이산, 연속), 확률분포 (기대값, 분산, 표준편차), 결합확률분포, 주변확률분포, 독립성, 공분산, 상관계수, 이산, 연속확률분포의 예
- 표본분포 - 표본분포의 개념 및 성질, 표본평균의 분포
- 통계적추론 - 점추정, 구간추정, 최대가능도법, 유의성 검정

확률의 개념

확률(Probability)이란?

어떤 사건(event)이 일어날 가능성을 나타내는 개념

- 확률을 정의하기 위해 필요한 개념

  - 표본공간(Sample Space, $\mathcal{S}$) : 어떤 시행 (Experiement)에서 얻을 수 있는 가능한 모든 결과(outcome)들의 집합
    예 : 하나의 주사위를 던지고, 나오는 눈의 수를 관찰할 때
    표본공간 $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
  - 사건(event, 사상) : 표본공간의 부분집합으로 보통 집합
    $A, B, C..$ 등으로 표현

표본공간과 사건 : 예제

- 두 개의 동전을 동시에 던져서 나오는 면의 순서쌍 (앞면 H, 뒷면 T)
  - 표본공간, $\mathcal{S} = \{(H,H),(H,T),(T,H),(T,T)\}$
  - 앞면이 적어도 한번 나오는 사건을 A 라 하자.

$$A = \{(H,H),(H,T),(T,H)\},\ A \subset \mathcal{S}$$

- 고객센터에 전화를 했을때 기다려야 하는 시간을 조사하기 위해 한 명의 고객이 기다린 시간(분)을 관측할때,
  - 표본공간 $\mathcal{S} = \{t | t \geq 0\}$
  - 기다린 시간이 3분 이상인 사건을 B 라하자.
  - $B = \{t \,|\, t \geq 3\}$

사건의 연산

집합 연산의 기호를 사용

- 합사건 : $A \cup B$
- 곱사건 : $A \cap B$
- 여사건 : A^C
- 배반사건 : $A \cap B = \emptyset$ 이면 A와 B는 서로 배반

확률의 정의 - 등확률모형의 경우

사건 A가 일어날 확률 $P(A)$는 다음과 같이 정의

$$P(A) = \frac{\text{사건 A에 속하는 원소의 개수}}{\text{표본공간 전체의 원소의 개수}}$$

- 예: 두 개의 동전을 동시에 던졌을 때, 앞면이 적어도 한 번 나올 확률

확률측도를 통한 확률의 정의

다음과 같은 성질을 만족하는 $P(\cdot)$를 확률측도(Probability Measure) 라고 한다 (Axiom of Probability)

(1) 표본공간 $\mathcal{S}$에서 임의의 사건 A에 대하여 $0 \leq P(A)$

(2) $P(S) = 1$

(3) 서로 배반인 사건 $A_1, A_2, \ldots$에 대하여

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

위의 공리로 부터 나오는 성질

- $P(\varnothing) = 0$
- $A \subset B$ 이면 $P(A) \leq P(B)$
- $0 \leq P(A) \leq 1$
- $P(A^C) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

조건부 확률(Conditional probability)

- 사건 A가 주어졌을 때 사건 B의 조건부확률은 $P(B|A)$로 나타내고 $P(A) > 0$이라는 가정하에 다음과 같이 정의

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- 사건 A를 새로운 축소된 표본공간으로 간주했을 때, 사건 B가 일어날 확률
- 예제 : 세개의 동전을 차례로 던지는 경우, 앞면이 나온 수가 2 (A)일때, 첫번째 던지기에서 앞면이 나올 (B) 확률은?

곱셈법칙

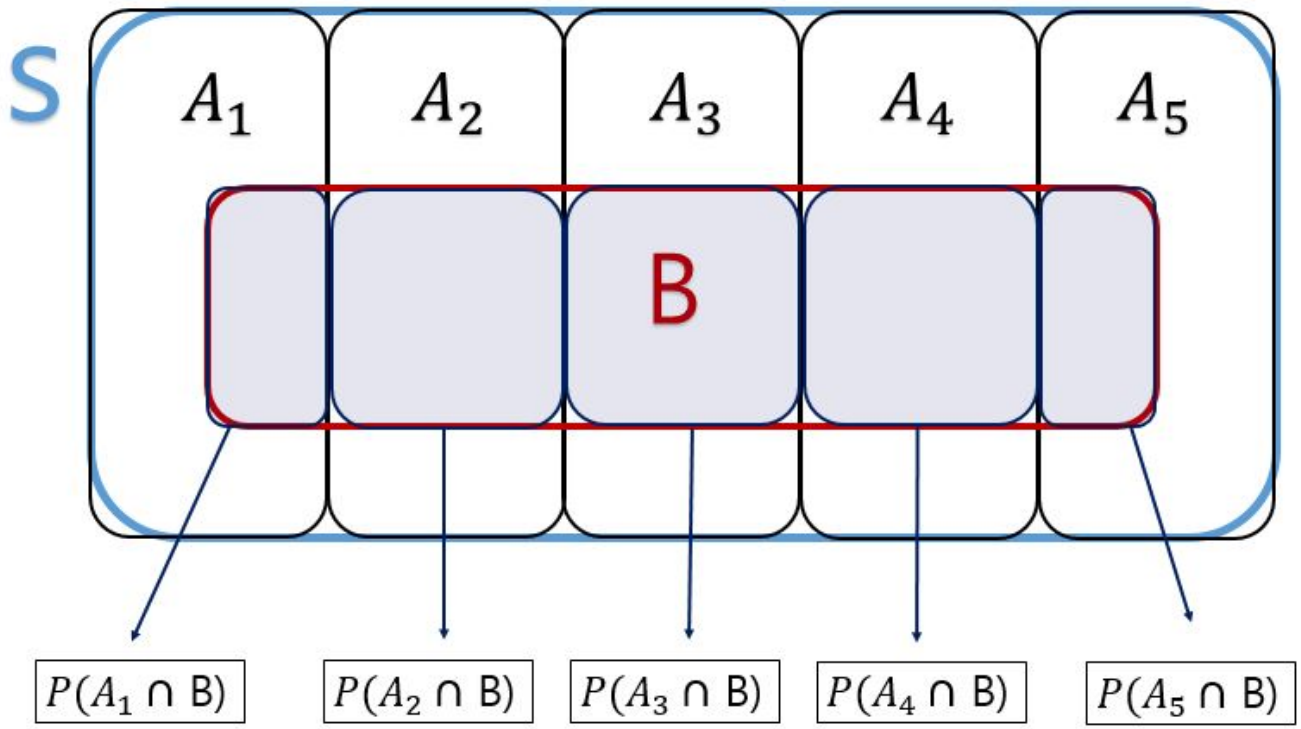
$P(A) > 0, P(B) > 0$이면

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

예제 : 빨간 구슬 10개와 파란구슬 90개가 들어있는 상자에서 2개를 단순 랜덤추출할 때, 2개 모두 빨간구슬일 확률을 구하여라.

전확률공식 (law of total probability)

어떤 사건 B 의 확률 $P(B)$을 구할때, 표본공간의 분할정보를 이용하는 공식



$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) + P(B|A_5)P(A_5)$$

- 표본공간 $\mathcal{S}$의 분할 $\{A_1, \cdots, A_n\}$을 생각하자. 표본공간의 분할은 다음을 만족한다.

  $A_i \cap A_j = \emptyset \ (i \neq j)$, $A_1 \cup A_2 \cup \cdots \cup A_n = \mathcal{S}$

- 이때, 전확률공식은

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)$$

전확률공식 : 예제

어떤 대학의 통계학과 학생의 30%는 1학년, 25%는 2학년, 25%는 3학년, 20%는 4학년 학생이다. 그런데 1학년의 50%, 2학년의 30%, 3학년의 10%, 4학년의 2%가 수학 과목의 수강생이다.
통계학과 학생 중 한 학생을 단순랜덤추출했을 때 그 학생이 수학 과목의 수강생일 확률을 구하여라.

독립(Independence)

서로 독립(mutually independence)

- 사건 A가 일어났다고 하더라도 사건 B가 일어날 확률에 아무런 영향을 미치지 않는 것
- 사건 A와 B가 서로 독립

  $P(B|A) = P(B)$ 또는 $P(A \cap B) = P(A)P(B)$
- 두 사건 A와 B가 독립이 아니면 종속이라고 한다.
- 참고 : $A \cap B = \varnothing$인 두 사건 A와 B는 서로 배반(mutually disjoint), 즉 두 사건이 동시에 일어날 수 없음을 의미하고 A와 B는 종속 사건이다.

예제

불량품 20개와 양호품 80개로 구성된 lot에서 2개의 제품을 단순랜덤추출할 때, 첫번째 제품이 불량품일 사건을 A, 두번째 제품이 불량품일 사건을 B라 하면 A와 B는 독립인가?

베이즈 정리(Bayes Theorem)

- $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$
- 두 사건 A, B의 확률 P(A),P(B), 조건부 확률 $P(B|A) = \frac{P(A \cap B)}{P(A)}$ 를 알고 있을때, $P(A \mid B)$를 구함.
- $P(B)$ 는 전확률공식을 이용하여 $P(B|A)P(A) + P(B|A^c)P(A^c)$로 바꿀수 있다.

베이즈 정리 : 예제

- 독감(infulenza) 감염 여부를 90%의 확률(독감이 걸렸을때 양성으로 판정할 확률 90%, 독감이 걸리지 않았을때 음성으로 판정할 확률 90%)로 진단하는 진단 시약 A가 있다. 사람이 독감에 걸릴 확률이 1%라 하자. 이 때, B라는 사람이 이 시약 A를 통한 진단 결과 독감 양성 판정을 받았을 때, 실제로 B가 독감에 걸렸을 확률은 얼마일까?

확률변수 및 확률분포

(실) 확률변수 (random variable)

표본공간의 각 원소를 하나의 실수로 대응하는 함수

$$c \in \mathcal{S},\ X(c) = x \in \mathbb{R}$$

예제

- 동전 1개를 던졌을 때 표본공간은 $\mathcal{S} = \{H, T\}$ (앞면 : H, 뒷면 : T)
- $P(H) = P(T) = \frac{1}{2}$
- $X(H) = 1,\ X(T) = 0$
- 함수 X의 정의역은 $\mathcal{S} = \{H, T\}$, 치역은 $\{0, 1\}$

확률과 확률변수

확률변수가 가지는 값에 대한 확률의 의미?

- 앞의 예제에서 $X = 1$일 확률은?
  - 동전 1개를 던졌을 때, 앞면이 나올 확률과 같다.
  - 따라서, 확률변수 X가 1의 값을 가질 확률을 다음과 같이 표현한다.
  - $P(X = 1) = P(\{c \in \mathcal{S} \mid X(c) = 1\}) = P(\{H\}) = 1/2$

확률분포 (probability distribution)

확률변수 X의 확률분포 (probability distribution)란: 확률변수 X가 가질 수 있는 값과 해당하는 확률에 대해 나타낸 것으로, 확률을 계산 할 수 있는 정보를 제공.

- 이산확률변수: X가 취할 수 있는 값이 $x_1, x_2, x_3, \ldots$와 같이 이산 일 때
  - 해당 값과 대응하는 확률을 제공
- 연속확률변수: X의 취할 수 있는 값이 셀 수 없이 많을 때
  - 특정 구간에 속하는 확률을 계산할 수 있는 정보를 제공.

이산확률변수 (discrete random variable)

확률분포는 다음과 같은 확률질량함수 (probability mass function) $p(x)$로 표현 가능

$$p(x) = P(X = x) = \begin{cases} P(X = x_i) & , x = x_i \text{일 때} (i = 1, 2, \dots) \\ 0 & \textit{otherwise}. \end{cases}$$

- $0 \leq p(x) \leq 1$
- $\sum_{all\ x} p(x) = 1$
- $P(a < X \leq b) = \sum_{a<X\leq b} p(x)$

예제

15개의 상품 중 5개가 불량품이다. 3개를 단순랜덤추출하였을 때, 그 중 불량품의 개수를 X라 하자. 확률변수 X의 확률분포를 구하여라.

연속확률변수 (continuous random variable)

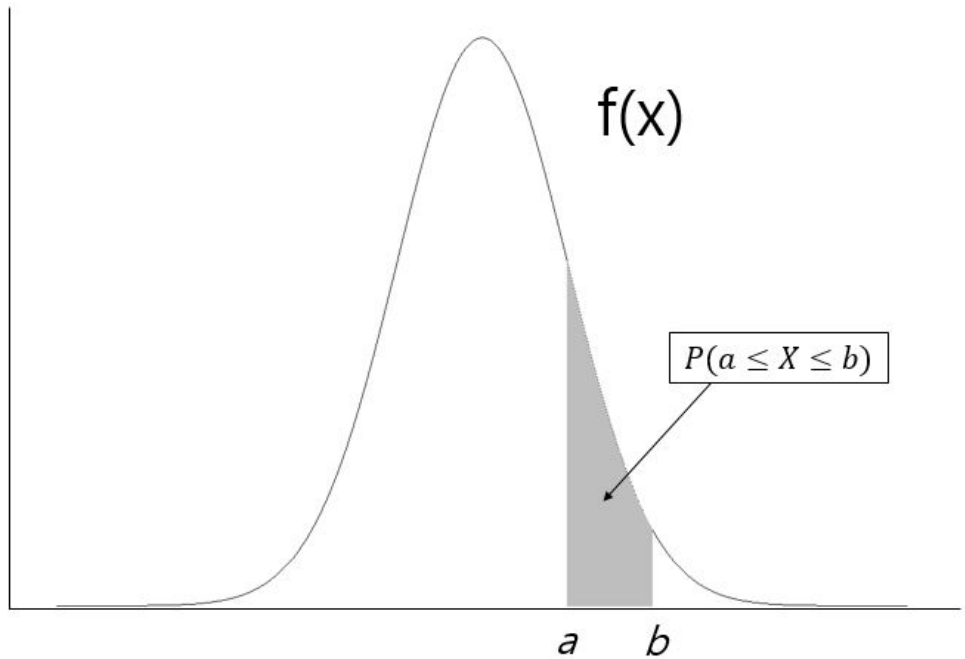
확률분포는 확률밀도함수 (probability density function) $f(x)$를 도입하여 X의 값이 $a < X \leq b$일 확률로 표현

$$P(a < X \leq b) = \int_a^b f(x)dx$$

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_a^b f(x)dx$

연속확률변수의 성질

- 연속확률변수의 한 점에서의 확률은 0이다. $P(X = a) = 0$
- $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$
  $= P(a < X < b)$



예제

연속확률변수 X의 확률밀도함수가

$$f(x) = \begin{cases} bx(1-x) & (0 \le x \le 1\text{일 때}) \\ 0 & (x < 0 \text{ 또는 } x > 1\text{일 때}) \end{cases}$$

로 주어질 때, 상수 b와 확률 $P(0 \le X \le \frac{3}{4})$를 구하여라.

기대값(expected value)

확률변수 X의 중심을 나타내는 값, 평균

$$\mu = E(X) = \begin{cases} \sum_x xp(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} xf(x)dx & \text{(연속확률변수)} \end{cases}$$

예제: 동전을 2회던지는 실험에서 앞면의 개수를 X라고 할 때, X의 기대값을 구하여라.

- 확률변수 X의 함수 $g(X)$의 기대값

$$E(g(X)) = \begin{cases} \sum_x g(x)p(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{(연속확률변수)} \end{cases}$$

- 기대값의 성질 : 선형성
  - $E(aX + b) = aE(X) + b$ (a, b는 상수)
  - $E[ag(X) + bh(X)] = aE(g(X)) + bE(h(X))$ (a, b는 상수)

분산(variance)과 표준편차(standard deviation)

X의 평균을 μ라고 하자.

(1) 분산

$$Var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 p(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx & \text{(연속확률변수)} \end{cases}$$

(2) 표준편차

$$sd(X) = \sqrt{Var(X)}$$

- $Var(X) = E(X^2) - [E(X)]^2$
- $Var(aX+b) = a^2 Var(X)$

예제

X의 확률밀도함수가 $f(x) = \begin{cases} 1 & , 0 \leq x \leq 1 \\ 0 & , otherwise \end{cases}$ 일 때,

X의 평균, 분산, 표준편차를 구하여라.

결합분포(joint probability distribution)

두 개의 확률변수가 취할 수 있는 값들의 모든 쌍의 확률을 나타낸 것

- 이산형 결합확률밀도함수

$$p(x, y) = P(X = x, Y = y)$$

  - $0 \leq p(x, y) \leq 1$
  - $\sum_x \sum_y p(x, y) = 1$
  - $P(a < X \leq b, c < Y \leq d) = \sum_{a<x\leq b} \sum_{c<y\leq d} p(x, y)$

- 연속형 결합확률밀도함수

$$P(a < x \leq b, c < y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$$

  - $f(x, y) \geq 0$
  - $\int f(x, y) dx dy = 1$
  - $P(a < x \leq b, c < y \leq d) = \int_c^d \int_a^b f(x, y) dx dy$

예제

서로 다른 동전 A,B,C를 동시에 던지는 실험에서 확률변수

$$X = \begin{cases} 1 & , \text{동전 A가 } H \\ 0 & , otherwise \end{cases}$$

$$Y = \begin{cases} 1 & , \text{동전 A,B가 } H \\ 0 & , otherwise \end{cases}$$

$$Z = \begin{cases} 1 & , \text{동전 B,C가 } H \\ 0 & , otherwise \end{cases}$$

Table: 표본공간과 확률변수

S	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	1	0	0	0	1	0	0	0

Table: X와 Y의 결합확률분포

y \ x	0	1	행의 합
0	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{3}{4}$
1	0	$\frac{1}{4}$	$\frac{1}{4}$
열의 합	$\frac{2}{4}$	$\frac{2}{4}$	1

주변확률밀도함수(Marginal PDF)

- 이산형 : $p_X(x) = \sum_y p(x, y)$
- 연속형 : $f_X(x) = \int f(x, y)dy$

예제: 앞의 예제에서

Table: X의 주변확률분포

x	0	1	계
$p_X(x)$	$\frac{1}{2}$	$\frac{1}{2}$	1

두 확률변수의 함수의 기대값

두 확률변수 X, Y의 함수의 기대값

- $E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y)p(x,y) & \text{(이산형)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y) & \text{(연속형)} \end{cases}$
- $E[ag(X,Y) + bh(X,Y)] = aE[g(X,Y)] + bE[h(X,Y)]$ (a, b는 상수)

두 확률변수의 독립성

두 확률변수 X, Y 가 다음을 만족할때:
모든 x, y에 대해

$$\begin{aligned} p(x,y) &= p_1(x)p_2(y) \text{ (이산형)} \\ f(x,y) &= f_1(x)f_2(y) \text{ (연속형)} \end{aligned}$$

- X와 Y는 서로 독립이면, $E(XY) = E(X)E(Y)$

공분산(Covariance)과 상관계수(Correlation coefficient)

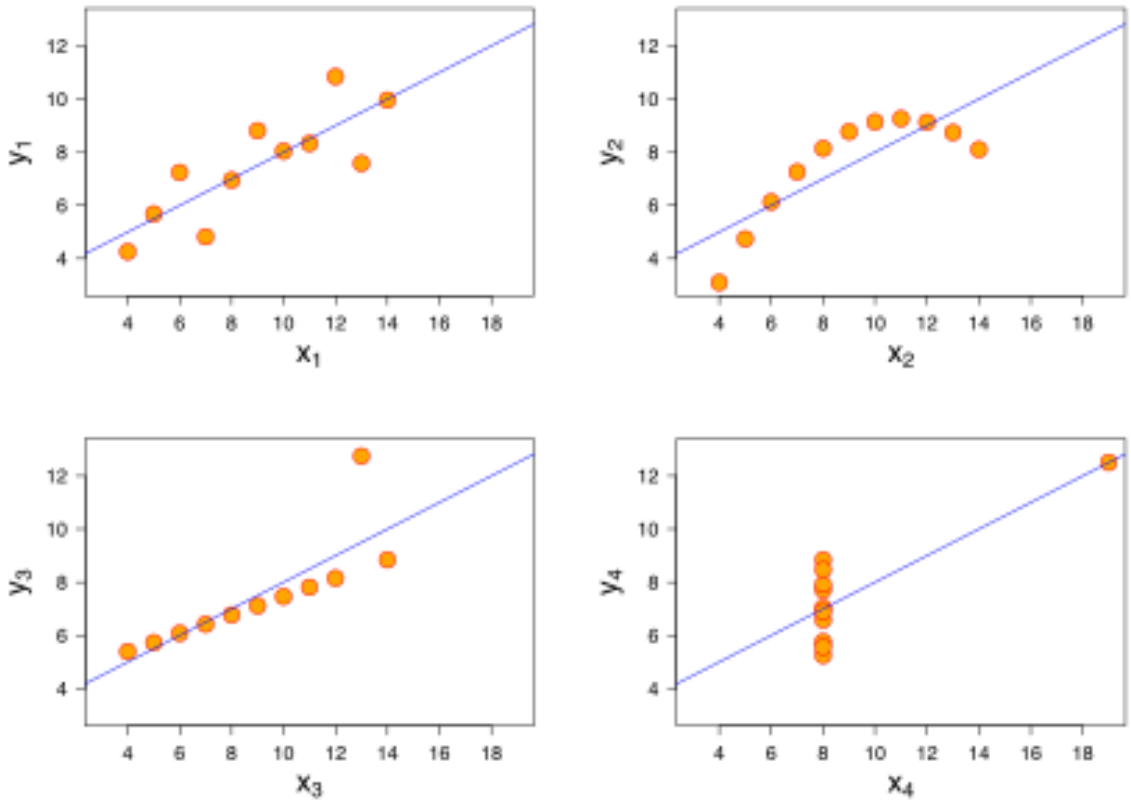
- 공분산(Covariance)

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_1)(Y - \mu_2)] \\ &= E(XY) - \mu_1\mu_2 = E(XY) - E(X)E(Y) \end{aligned}$$

- 상관계수(Correlation coefficient) -선형의 연관성을 나타냄

$$Corr(X, Y) = \rho_{XY} = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

Anscombe's quartet



From Wikipedia

공분산과 상관계수의 성질

확률변수 X, Y에 대해 다음과 같은 성질들이 있다.

- $Cov(aX + b, cY + d) = acCov(X, Y)$
- $Corr(aX + b, cY + d) = sign(ac)Corr(X, Y)$
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$
- $-1 \leq \rho \leq 1$
- $Y = a + bX$ 이면 $\rho = \pm 1$

확률변수 X, Y가 독립일 경우

- $E(XY) = E(X)E(Y)$
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
- $Cov(X, Y) = 0$, $Corr(X, Y) = 0$
  (주의 : $Cov(X, Y) = 0$인 것이 X, Y의 독립을 의미하지 않음)
- $Var(X \pm Y) = Var(X) + Var(Y)$

베르누이 분포(Bernoulli distribution)

베르누이 시행 (Bernoulli trial)

- 실험의 결과 두 가지 중의 하나로 나오는 시행
- 표본 공간 $\mathcal{S} = \{\text{성공}(s), \text{실패}(f)\}$
- 성공 확률 $p = P(\{s\})$

베르누이 확률변수 (Bernoulli random variable)

- 베르누이 시행의 결과를 0 또는 1의 값으로 대응시키는 확률변수
- $X(s) = 1, X(f) = 0$인 확률변수

- 베르누이 확률변수의 확률분포를 베르누이 분포라 한다
- $X \sim Ber(p)$
- $p(x) = p^x(1-p)^{1-x}$, $x = 0, 1$.
- $E(X) = p$
- $Var(X) = E(X^2) - [E(X)]^2 = p(1-p)$

이항분포 (Binomial distribution)

베르누이 시행을 n번 시행할 때 성공횟수의 분포

- $X \sim B(n, p)$ 또는 $Bin(n, p)$.
- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, x=0, $\cdots$, n.
- $n = 1$이면 베르누이 분포

이항분포의 성질

$X \sim B(n, p)$

- $X = X_1 + X_2 + \ldots X_n$, $X_i \sim i.i.d.\ Ber(p)$
- $E(X) = E(X_1 + X_2 + \ldots X_n) = np$
- $Var(X) = Var(X_1 + X_2 + \ldots X_n) = np(1 - p)$

예제

5개 중 하나를 택하는 선다형 문제가 20문항 있는 시험에서
랜덤하게 답을 써 넣는 경우
X : 20문항 중 정답의 수

- $X \sim B(20, 0.2)$
- 정답이 하나도 없을 확률 :
- 8개 이상의 정답을 맞힐 확률 :
- $E(X) = 20 \times 0.2 = 4$
- $Var(X) = 20 \times 0.2 \times 0.8 = 3.2$

포아송 분포 (Poisson distribution)

- 주어진 기간동안 독립적으로 일어나는 사건들의 횟수에 대한 확률변수
- $X \sim Poisson(\lambda)$
- $f(x) = \frac{\lambda^x}{x!} e^{-\lambda},\ x = 0, 1, 2, \cdots,$
- $E(X) = \lambda,\ Var(X) = \lambda$

예제

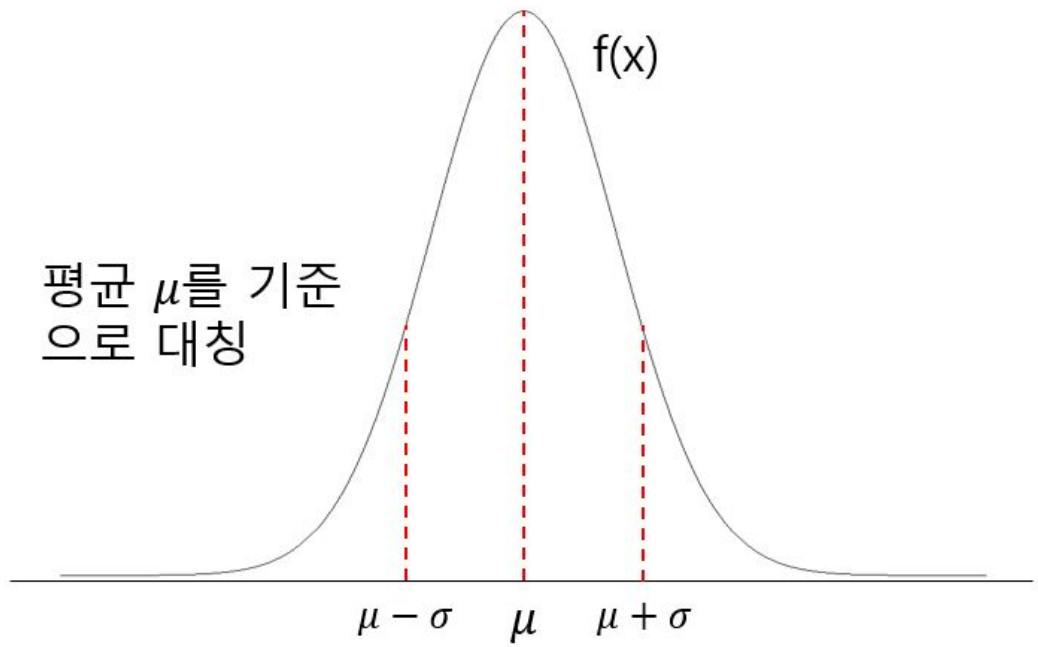
기업 A의 노트북의 한달동안 반품되는 수량은 평균 5인 포아송 분포를 따른다고 하자. 이번달에 반품되는 노트북이 1대 이하일 확률은?

정규분포 (Normal distribution)

- 가우스(Gauss, 1777-1855)에 의해 제시된 분포로서 가우스분포(Gaussian distribution)라고도 불린다.
- 물리학 실험 등에서 오차에 대한 확률분포를 연구하는 과정에서 발견된 연속확률분포.
- 통계학 초기 발전 단계에서 모든 자료의 히스토그램이 가우스분포의 형태와 유사하지 않으면 비정상적인 자료라고 믿어서 "정규(normal)"라는 이름이 붙게 되었다.

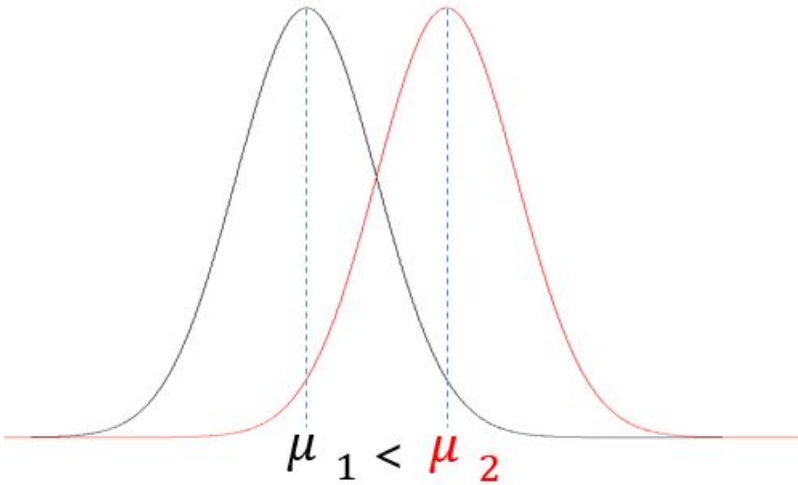
- $X \sim N(\mu, \sigma^2)$, μ: 평균, σ^2: 분산

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big), -\infty < x < \infty, \ \sigma > 0$$

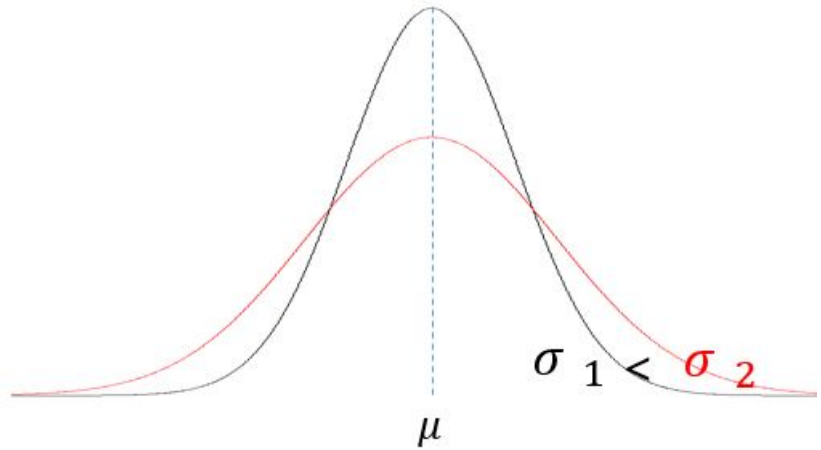


정규분포의 성질

표준편차는 같고 평균이 다른 두 정규분포



평균은 같고 표준편차가 다른 두 정규분포



정규분포의 성질

- $X \sim N(\mu, \sigma^2)$

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- $X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_1, \sigma_2^2)$, X_1, X_2는 서로 독립

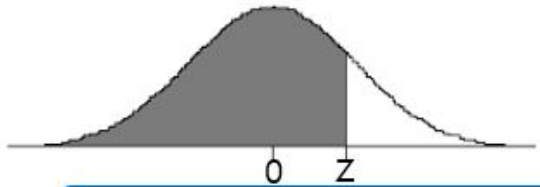
$$a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$$

표준정규분포(standard normal distribution)

- 평균이 0이고 표준편차가 1인 정규분포를 표준정규분포 (standard normal distribution)라고 한다.
- 보통 Z로 표기.

표준정규분포표

- 표준정규분포표($P(Z \leq z)$)



소수점 둘째 자리

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

소수점 첫째 자리

- $P(Z \leq 0.93) = 0.8238$

정규분포에서 확률 구하기

- 일반적인 정규분포 $X \sim N(\mu, \sigma^2)$의 확률 계산시에 표준정규분포를 이용한다
- 표준화(standardization) :
  $X \sim N(\mu, \sigma^2)$일 때 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$

  $X = \mu + \sigma Z$

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right)$$
$$= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

예제

학생들의 통계학 성적의 분포가 근사적으로 $N(60, 10^2)$을 따른다고 한다. 45점 이하인 학생에게 F학점을 준다고 할 때, F학점을 받게 될 학생의 비율을 근사적으로 구하여라.

표본 분포

표본분포 소개를 위한 몇가지 개념

- 모집단 (population): 정보를 얻고자 하는 대상이 되는 집단 전체
- 모수 (parameter): 모집단의 특성을 나타내는 대표값
- 표본 (sample): 모집단에서 추출한 부분집합
- 통계량 (Statistic): 표본자료의 특성을 나타내는 값.
  - 통계량은 표본으로 구하므로, 표본의 함수라고 볼 수 있다.
- 추정량 (Estimator): 모수의 추정을 위해 구해진 통계량

표본분포란?

- 통계량의 확률 분포
  - 랜덤표본의 값에 따라 통계량의 값 역시 정해진다. 이 때, 통계량 역시 확률변수로서 특정한 확률분포를 따르게 되고, 그 분포는 모집단의 분포와 관계가 있다.

모수와 추정량 예

5개의 시리얼 박스에 쿠폰이 한장씩 있다고 하자. 당첨, 당첨, 당첨,탈락,탈락으로 이루어졌다고 할때 이 모집단에서 크기 3인 표본을 단순 랜덤 비복원추출로 뽑아 모비율(당첨비율) p (이 예제에서는 0.6)을 추정하는 상황(당첨=1, 탈락=0)을 생각해보자.

Table: 가능한 표본과, 표본비율, 그 확률

가능한 표본	표본비율 ($\hat{p}$)	확률
당첨 3, 탈락 0	1	$\binom{3}{3}\binom{2}{0} \Big/ \binom{5}{3}$
당첨 2, 탈락 1	2/3	$\binom{3}{2}\binom{2}{1} \Big/ \binom{5}{3}$
당첨 1, 탈락 2	1/3	$\binom{3}{1}\binom{2}{2} \Big/ \binom{5}{3}$

- ’표본비율’은 모수인 모비율을 표본으로 추정한 추정량이다.
- 표본비율은 표본에 따라 다른 값을 가진다.

- 이 예에서는 해당하는 표본이 나오는 확률이 표본비율의 값에 대응되는 확률로, 표본비율의 표본분포가 된다.
- 만약 모비율이나 표본의 크기가 달라진다면 표본비율의 분포 역시 달라진다. 즉, 표본분포는 모집단의 분포와 표본 추출 방식의 영향을 받는다.

표본평균의 분포

- 표본평균 (sample mean), $\bar{X}$
  - 표본의 중심경향성을 나타내는 통계량.
  - 모집단의 평균 (모평균)을 μ라고 하면, 표본평균은 μ의 추정량 (estimator)이다.
  - 표본 $\{X_1, X_2, \cdots, X_n\}$가 모평균 μ, 모분산 σ^2인 모집단에서 추출된 랜덤표본일때,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- 무한모집단에서 추출된 랜덤표본일 경우,

$$E(\bar{X}) = \mu, \ Var(\bar{X}) = \frac{\sigma^2}{n}, \ sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

큰수의 법칙 (Law of large numbers)

- 표본의 크기 n 이 커질수록 표본평균의 분산은 0에 가까워진다.
- 표본평균의 기대값은 모평균과 같고, 분산이 작아지므로, $\bar{X}$는 모평균 μ의 근처에 밀집되어 분포함을 알 수 있다.
- 이러한 결과를 큰수의 법칙 (Law of Large Numbers)이라고 한다.

중심극한정리 (Central limit theorem)

- 임의의 모집단에 대해 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$의 분포는 표준정규분포 $N(0,1)$에 근사한다.
- 유한모집단의 경우, 모집단의 크기 N과 표본의 크기 n이 충분히 크면(단 $N \gg n$) $\frac{N-n}{N-1}$의 값이 1에 근사하므로, 위의 성질이 성립한다.
- 중심극한정리를 통해, 모집단의 분포가 어떤 형태이든지 표본의 크기가 크면 표본평균의 분포를 정규분포로 근사할 수 있다.

  즉, $\bar{X}$의 분포 $\approx N\left(\mu, \frac{\sigma^2}{n}\right)$.

이항분포의 정규분포 근사

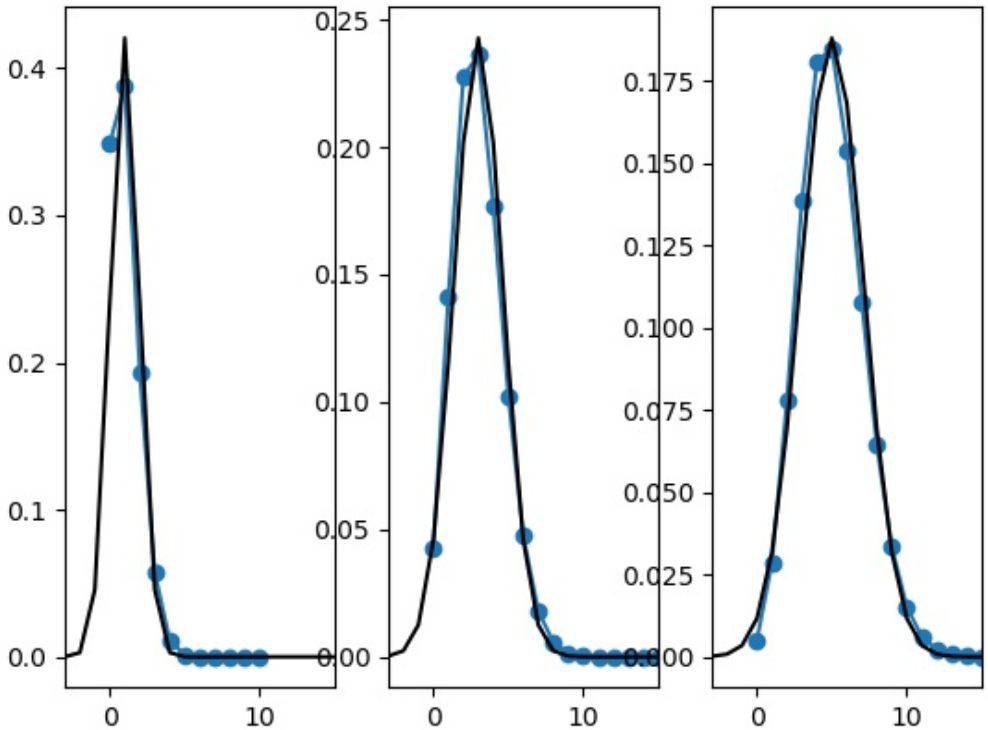
- $X_1, X_2, \cdots, X_n$이 성공률이 p 인 베르누이분포를 따르는 무한모집단의 랜덤표본이라고 하자
- 이 경우, $S = \sum_{i=1}^{n} X_i$ 은 이항분포 $B(n, p)$을 따른다.
- 중심극한정리를 적용하면, n이 충분히 클 때

$$\frac{S - np}{\sqrt{(np(1-p))}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

의 분포는 표준정규분포 $N(0, 1)$에 근사한다.
($\hat{p}$= 베르누이분포의 표본비율 $\frac{S}{n}$)

- 즉, n이 충분히 크고, np가 적당한 값이면, $B(n, p)$를 이용하는 확률계산을 $N(np, np(1-p))$를 이용하여 근사할수 있다.

이항분포의 정규근사 그래프



파랑: p=0.1인 이항분포, 검정: 평균이 np, 분산이 np(1-p)인 정규분포. 왼쪽부터 차례대로 n=10,30,50

통계적 추론

통계적 추론 (Statistical Inference)

- 표본으로부터의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어내는 과정
- 추정(Estimation)
- 유의성 검정(Significance test, 또는 가설 검정(Hypothesis test))

추정 (Estimation)

표본으로부터 모집단의 특성값(모수)에 대한 추측값과 오차를 제시

- 모수(Population parameter, θ) : 모집단의 특징을 나타내는 대표값 (예 : 모평균 μ, 모분산 σ^2)
- 랜덤표본 : 서로 독립이고 동일한 확률분포를 따르는 확률변수들을 말하며, 실제로 표본을 추출하여 얻은 값들을 관측값(Observation)이라 한다.

모수의 추정 (Parameter estimation) - 점추정 (Point estimation), 구간추정 (Interval estimation)

점추정(Point estimation)

표본으로부터 계산한 모수의 추정값을 제시

추정량의 예

- 모평균의 추정량 : 표본평균 $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 모분산의 추정량 : 표본분산 $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

추정량의 평가

추정량(Estimator)을 평가하는 몇 가지 기준이 있다.

- 불편추정량 (Unbiased estimator)
  - $E(\hat{\theta}) = \theta$를 만족하는 추정량 $\hat{\theta}$

    예 : 표본평균과 표본분산은 각각 모평균과 모분산의 불편추정량이다.
  - Bias $(\hat{\theta}) = E(\hat{\theta}) - \theta$
- 표준오차 (Standard error), $se(\hat{\theta})$
  - 추정량의 표준편차

    예 : $X_1, \cdots, X_n \overset{i.i.d.}{\sim} (\mu, \sigma^2)$,

$$se(\hat{\mu}) = sd(\hat{\mu}) = \sqrt{var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

구간추정 (Interval estimation)

- 모수를 추정값을 구간으로 제공
- 구간추정의 일반적 방법: 신뢰구간 (Confidence Interval, CI)
- 신뢰수준 (Confidence level)이 $100(1-\alpha)\%$인 신뢰구간 (L, U)는 다음을 만족한다.

$$P(L \leq \theta \leq U) = 1 - \alpha$$

  - L, U는 표본으로부터 구해짐. 즉, $L \equiv L(X_1, \cdots, X_n)$, $U \equiv U(X_1, \cdots, X_n)$
  - 따라서 (L, U) 는 확률 변수로 이루어진 구간 (random interval)
- $1-\alpha$ 는 포함확률(coverage probability)이라고 부름

신뢰구간 예제

모분산 σ^2를 알 때 정규모집단의 모평균 μ의 구간추정

- 신뢰수준 $100(1-\alpha)\%$인 μ의 신뢰구간

$$\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$로부터

$$\begin{aligned} 1-\alpha &= P\left(-Z_{1-\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

- $Z \sim N(0,1)$일때, $P(Z \le z_p) = p$
- 오차의 한계: 신뢰구간에서 허용하는 가장 큰 오차. 신뢰구간 길이의 $\frac{1}{2}$. 즉, $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

신뢰구간의 의미

μ의 100(1-α)% 신뢰구간 : 100번의 표본 추출을 통해 얻어진 100개의 신뢰구간

$$\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

에서, 100(1-α)% 개 정도의 신뢰구간이 모평균을 포함할 거라 기대함

- $n = 25$, σ = 10일 때 모평균의 90% 신뢰구간
- $(\bar{X} - Z_{0.95}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{0.95}\frac{\sigma}{\sqrt{n}})$
  $= (\bar{X} - 1.645\frac{10}{\sqrt{25}}, \bar{X} + 1.645\frac{10}{\sqrt{25}})$ $(\because Z_{0.95} \simeq 1.645)$
- 표본을 새로 추출할 때마다 $\bar{X}$가 달라지므로 신뢰구간 역시 길이를 유지한 채 표본에 따라 값들이 달라짐

최대 가능도 추정법 (Maximum Likelihood Estimation)

가능도 (우도, likelihood)를 최대가 되게 하는 모수 값을 그 추정량으로 하는 방법

- 가능도:
  - 확률변수의 관측값이 주어졌을때 해당 확률의 정도
  - 모수를 가지는 확률 분포의 경우 모수값이 가능한 정도
  - 일반적으로, 확률밀도함수에 확률변수의 관측값을 대입한 값

- 예제: $Ber(p)$를 따르는 확률변수 X_1, X_2, X_3가 독립적으로 다음과 같이 관측되었다고 하자. $X_1 = 1, X_2 = 0, X_3 = 0$.
  - 결합 확률질량함수에 관측값을 대입한 값은 $p(1,0,0) = p_1(1)p_2(0)p_3(0) = p \times (1-p) \times (1-p)$
  - 따라서 성공확률 p에대한 가능도는 $L(p) = p(1-p)^2$, $0 \leq p \leq 1$.
- 일반적으로, 모수 θ에 대한 가능도함수는

$$L(\theta) = L(\theta|X_1, \cdots, X_n) = f(X_1, \cdots, X_n; \theta)$$

- 이산확률변수인 경우 가능도는 관측값이 주어졌을때의 해당 확률이 된다.
- 확률변수들이 독립적으로 관측되었을 경우 가능도는 다음과 같이 주변확률분포의 곱으로 표현.

$$L(\theta) = L(\theta|X_1, \cdots, X_n) = \prod_{i=1}^{n} f(X_i; \theta)$$

최대가능도 추정의 원리 - 예

- 공장 A의 한 생산라인의 제품 불량률 (p)을 추정하기 위해 10개의 제품을 랜덤 추출하였다고 하자. 불량인 경우를 1로 할 경우, 10개의 불량정보는 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0$이라고 하자.
- 각 제품의 불량 유무를 1또는 0을 가지는 확률변수 X로 봤을때 $X_1, \cdots, X_{10} \overset{i.i.d.}{\sim} Ber(p)$

- 불량률 p에 대한 가능도: $L(p) = p^2(1-p)^8, 0 \leq p \leq 1$

  즉, 불량률이 p일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)
- $L(0.5) = 0.5^2 0.5^8 = 0.00098$: 불량률이 $p = 0.5$일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)
- $L(0.2) = 0.2^2 0.8^8 = 0.0067$: 불량률이 $p = 0.2$일때 $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)
- $(0,0,0,1,0,1,0,0,0,0)$를 관측할 확률(정도)를 최대가 되게 하는 p

최대가능도 추정의 원리

- 주어진 관측값에서 모수의 가능도를 최대가 되게 하는 값으로 모수를 추정하는것이 최대가능도 추정법(maximum likelihood estimation, MLE)이다.

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta | X_1, \cdots, X_n)$$

로그 가능도

- 가능도 $L(\theta)$는 θ에 관하여 복잡한 함수의 형태로 나타나고, 모수들의 집합인 Ω (Parameter space)가 조건에 따라 bounded set인 경우도 있다.
- 로그변환을 시킬 경우 좀 더 다루기 편한 함수가 되기도 한다.
- 독립인 관측값들이 주어졌을때,

$$\ell(\theta) = \log(L(\theta)) = \log(\prod f(X_i;\theta)) = \sum_{i=1}^{n} \log(f(X_i,\theta)).$$

$$\hat{\theta} = \arg\max_{\theta\in\Omega} L(\theta) = \arg\max_{\theta\in\Omega} \ell(\theta)$$

예제 1

- 앞에서 예제로 소개하였던 문제에서 제품의 불량률 p의 최대가능도 추정량을 구하여라.

예제 2

- 치즈회사에서 새로운 치즈(A) 를 선보여 다른 두 경쟁 제품 (B,C)과의 선호도 차이를 알아보고자 50명에게 blind test를 실시 하여 선호도를 조사한 결과, 20, 18, 12 명이 각각의 치즈를 골랐다. 선호비율을 p_1, p_2, p_3라고 할때 주어진 자료를 가지고 최대가능도 추정량을 구하여라

MLE 의 성질

몇가지 가정하에,

- 일치성:

$$\hat{\theta} \to \theta \text{ in probability}$$

- 점근적 정규성:

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \mathcal{I}^{-1}(\theta)) \text{ in distribution,}$$

여기서 $\mathcal{I}(\theta) = E\left[\left(\frac{d}{d\theta} \log f(X;\theta)\right)^2\right] = -E\left(\frac{d^2}{d\theta^2} \log f(X;\theta)\right)$, 는 피셔정보 (*Fisher Information*)라고 부른다.

- 즉, 점근적으로 (n이 커질때), MLE는 평균이 θ, 분산이 $\frac{1}{n\mathcal{I}(\theta)}$인 정규분포를 따른다.

최대가능도 추정량의 계산

- 많은 경우 가능도가 최대가 되게 하는 θ를 찾기 위해 미분을 이용
- 즉, $\dot{\ell}(\theta; x) = 0$ (가능도 방정식, likelihood equation) 의 해 중에 최대가능도 추정량을 찾음.
- 가능도방정식을 통해 추정량의 형태을 직접 구하거나
- 그렇지 않은 경우, 수치적으로 근을 찾는 방법을 사용.

- 수치적으로 근을 찾는 방법중, Newton-Raphson 방법

$$\hat{\theta}^{(r+1)} = \hat{\theta}^{(r)} + [-\ddot{\ell}(\hat{\theta}^{(r)})]^{-1}\dot{\ell}(\hat{\theta}^{(r)}),\ r = 0, 1, 2. \cdots$$

- 초기값으로는 흔히 적률추정량을 사용한다.

MLE를 이용한 신뢰구간

- MLE의 점근분포를 이용하여 θ의 신뢰구간을 구할 수 있다.
- $\sqrt{n\mathcal{I}(\theta)}(\hat{\theta} - \theta) \sim \mathcal{N}(0, 1)$로부터,
- 신뢰수준 $100(1-\alpha)\%$인 θ의 CI:

$$\left(\hat{\theta} - Z_{1-\alpha/2} \frac{1}{\sqrt{n\mathcal{I}(\hat{\theta})}}, \ \hat{\theta} + Z_{1-\alpha/2} \frac{1}{\sqrt{n\mathcal{I}(\hat{\theta})}} \right).$$

- $Z \sim N(0, 1)$일때, $P(Z \leq z_p) = p$

유의성 검정

- 기존의 이론이나 법칙을 부정하는 것으로 보이는 현상이 관측되었을 때, 이를 유지할지 부정할지를 결정하는데 사용.
- 반증을 찾기 위해 설정된 가설 (주로 '기존의 가설'): 귀무가설 (Null hypothesis, H_0)
- 귀무가설의 대안으로 상정되는 가설: 대립가설(Alternative hypothesis, H_1)
- 귀무가설에 대한 반증의 강도를 제공하는 과정을 유의성 검정 (Test of significance)이라 한다.

예제1

건물의 소화용으로 사용되는 살수장치가 섭씨 55도에서 작동되도록 제조하려고 한다. 제조공정의 이상 여부를 판단하기 위해 생산품 중에서 표본을 추출하여 작동 시작 온도를 조사하고자 한다. 이러한 조사에서 공정에 이상이 있다는 증거가 뚜렷하면 후속되는 기술적 조치를 하려 한다.

- 살수장치의 평균 작동 시작 온도 : μ
- 공정에 이상이 있는 경우 : $\mu \neq 55$
- $H_0 \ : \ \mu = 55, H_1 \ : \ \mu \neq 55$

만약, 9개의 표본을 관측한 결과 표본평균 $\bar{x} = 55.63$ 이라는 결과가 나왔다면, 어떻게 해석해야 할까?

예제 2

- 흰색과 빨간색이 섞인 10개의 구슬이 주머니에 들어있다. 흰구슬이 5개 또는 7개가 들어있다고 알려져 있다.
- 구슬을 10번 복원추출하여 나온 값으로 흰 구슬이 5개인지 7개인지 결정하고자 한다.
- 이때 생각할수 있는 가설 두 개는 다음과 같다.

$$\begin{aligned} H_0 &: p = 0.5 \ (= p_0) \\ H_1 &: p = 0.7 \ (= p_1) \end{aligned}$$

  여기서 p는 한번 추출할때 흰구슬이 나올 확률이다.
- 만약 10번의 복원추출 결과 2개가 흰색이 나왔다면, 어느 가설을 골라야 할까?

예제 2- 계속

- X를 10번의 복원추출한 후의 흰구슬의 개수라고 하자. 이를 이용하여 가설을 고르는 규칙(test or rule)을 만들어 보자.
- 먼저 임의로 test 1을 $X \geq 6$이면 귀무가설을 기각하고 대립가설을 채택하는것으로 정하자.
- 만약 관측값이 2인경우, 즉 $X = 2$, test 1에 의하면 귀무가설을 채택한다.
- 이제 test 2를 $X \geq 1$이면 귀무가설을 기각하는것으로 정하자.
- 이 경우, 관측값 $X = 2$는 test 2에 의하여 귀무가설을 기각하고 대립가설을 채택한다.
- 어떤 test를 골라야 할까?

필요한 용어의 정리

- 단순가설 (simple hypothesis): $\theta = \theta_0$, $p = 0.5$와 같이 모수를 특정값으로 가정하는 가설
- 복합가설 (composite hypothesis): 모수값이 하나보다 많은경우를 가정하는 가설
- 단측가설 (one-sided): $\theta > \theta_0$, 또는 $\theta < \theta_0$와 같이 비교하는 값의 한 쪽에 대해서만 제시되는 가설
- 양측가설 (two-sided): $\theta \neq \theta_0$와 같이 양 쪽에 대해서 제시되는 가설

- 1종 오류 (type I error): 귀무가설이 옳은 상황에서 귀무가설을 기각함으로 인해 생기는 오류
- 2종 오류(Type II error): 귀무가설이 틀린 상황에서 귀무가설을 기각하지 못함으로 인해 생기는 오류
- 두가지의 오류를 동시에 작게 하기 어렵기 때문에, 보통 1종 오류가 일어날 확률을 정하고 (controlling type I error), 그중에 2종 오류가 일어날 확률이 적은 test를 고려한다.

실제현상 / 검정결과	H_0 참	H_1참
H_0채택	옳은 결정	**제 2종 오류**
H_1채택	**제 1종 오류**	옳은 결정

- 유의수준 (significance level): 1종 오류가 일어날 확률, α
  - $\alpha = 0.05$라 함은, 귀무가설이 참인데 기각할 오류를 5% 이하로 하겠다는 것이다.
- 검정력 (power): 귀무가설이 거짓일때 test가 귀무가설을 기각할 확률
  - 2종 오류가 일어날 확률을 β라고 하면, 검정력은 $1 - \beta$
- 주어진 유의수준하에서 검정력이 가장 큰 (2종 오류가 제일 작은) test를 most powerful test라고 부른다.

- 검정통계량 (Test statistics): 가설 검정에 사용되는 통계량
- 기각역 (Critical region): 귀무가설 H_0을 기각시킬 수 있는 검정통계량의 관측값의 영역.
- 앞의 예제2에서 검정 통계량과 기각역, 1종오류 확률등을 찾아보자.

- 유의확률 (P 값): 검정 통계량의 관측값을 가지고 귀무가설이 기각되게하는 가장작은 유의수준.
  - 또는 검정통계량의 관측값을 포함하는 기각역의 최소확률
- 표본으로부터 구한 검정통계량의 관측값으로 구한 유의확률이 지정된 유의수준 이하로 나타나면 **통계적으로 유의하다**라고 표현
- 예제2에서 관측값이 $X = 2$인 경우에 대한 유의확률을 구해보자.

가설검정에서 알아두어야 할 사항

- 검정통계량이 다르면 다른 test
- 기각역의 형태는 대립가설의 영향을 받음
- P값(유의확률) 또는 기각역을 구하기 위해서는
  귀무가설하에서 검정통계량의 분포를 알아야 함
- 어떤 가설을 귀무가설로?
  - 기존에 믿어오던(알려져 있던) 사실
  - 오류의 위험이 더 큰 경우를 1종 오류가 되도록 정함

예제: 정규모집단에서 모평균의 가설 검정

- $H_0 \ : \ \mu = \mu_0$
- 검정통계량: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
- 검정통계량의 관측값 : z_0

$Z \sim N(0,1)$일때, $P(Z \leq z_p) = p$

대립가설 H_1	유의확률	유의수준 α의 기각역
$\mu > \mu_0$	$P(Z > z_0)$	$Z > z_{1-\alpha}$
$\mu < \mu_0$	$P(Z < z_0)$	$Z < -z_{1-\alpha}$
$\mu \neq \mu_0$	$P(\|Z\| > \|z_0\|)$	$\|Z\| > z_{1-\alpha/2}$

유의성 검정의 절차

- 귀무가설, 대립가설, 유의수준을 설정한다.
- 표본을 추출하고 검정통계량의 값을 계산한다.
- 가설을 기각할 수 있는지 없는지를 판단하고, 결론을 이끌어낸다.
  - 유의수준으로부터 기각역을 찾아 검정통계량의 값이 기각역에 속하는지 또는
  - 검정통계량의 값으로 유의확률을 계산하여 유의수준과 비교

예제

A사에서 생산중인 고양이 사료 캔의 열량은 평균이 1,200kcal, 표준편차가 100kcal로 알려져 있다. 이제, 사료의 열량을 늘리기 위해 재료를 일부 변경하여 만든 시제품을 25개 생산하여 조사한 결과 평균 열량이 $\bar{x} = 1240$kcal이었다. 새로운 재료로 만든 사료 열량의 표준편차가 100kcal로 유지된다고 할 때, 이 조사 결과는 사료의 열량을 늘리기 위한 재료 변경이 성공적임을 뜻하는가? 유의 수준 0.05에서 검정해보자.