2. 다양한 확률분포

데이터 및 데이터분석에 사용되는 다양한 분포

임채영

서울대학교 통계학과

이번 강의에서 다룰 내용

- 확률분포의 예- 데이터의 분포로 사용될 수 있는 확률분포
- 다변량 확률변수와 확률분포- 여러 종류의 데이터의 불확실성을 동시에 표현하는 도구

확률분포의 예

베르누이 분포(Bernoulli distribution)

베르누이 시행 (Bernoulli trial)

- 실험의 결과가 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 과 1의 값으로 대응시키는 확률변수
- $X(s) = 1, X(f) = 0$인 확률변수

베르누이 확률변수의 확률분포를 베르누이 분포라 한다

- $X \sim Ber(\theta)$ 또는 $Bernoulli(\theta)$
- $p(x) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$.
- $E(X) = \theta$
- $Var(X) = E(X^2) - [E(X)]^2 = \theta(1-\theta)$
- 불량 여부, 질병 감염 여부 등 이원적 결과를 가지는 데이터에 적용할 수 있다.

이항분포 **(Binomial distribution)**

베르누이 시행을 n번 독립적으로 시행할 때 성공횟수의 분포

- $X \sim Bin(n, \theta)$ 또는 $Binomial(n, \theta)$.
- $p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, $x = 0, \cdots, n$.
- $n = 1$이면 베르누이 분포
- 제품 중 불량품의 갯수, 전체 환자중 특정 증상을 가지는 환자 수 등 고정된 범위 내에서 일어나는 횟수로 나타나는 데이터에 적용할 수 있다.
- $X \sim Bin(n, \theta)$, $E(X) = n\theta$, $Var(X) = n\theta(1-\theta)$

예제

10개의 제품을 묶어서 1박스로 판매하는 경우를 생각하자. 하나의 제품이 불량일 확률이 10%라고 하고 각 제품의 불량여부는 서로 독립이라고 하자. 1박스에 불량품이 1개 초과하는경우 반품이 가능하다고 하자.
X를 20개의 박스중에 반품되는 박스의 수라고 하는경우

- $X \sim B(20, \theta)$
- $\theta =$?
- 반품되는 박스가 하나도 없을 확률 :
- 8개 이상의 반품될 확률 :
- $E(X) =$?
- $Var(X) =$?

포아송분포(Poisson distribution)

- 일정 기간 또는 특정 공간상에서 일어나는 독립적인 사건들의 횟수를 모형화 한 분포

  $p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, \cdots. \lambda > 0.$
- $X \sim Poisson(\lambda)$
- $E(X) = \lambda$, $Var(X) = \lambda$.
- 1년동안 특정 도로에서 발생하는 교통사고 횟수, 하룻동안 방문하는 고객수 등 범위가 제한되지 않은 횟수로 나타나는 데이터에 적용할 수 있다.

예제

생산라인 A에서 하루동안 생산되는 제품중 불량품의 개수는 평균이 1인 포아송 분포를 근사적으로 따른다고 하자.

- 오늘 생산되는 제품중 불량품이 없을 확률은?
- 이틀동안 생산된 제품중 불량품이 1개 이하일 확률은?

균일분포(Uniform distribution)

- 확률변수 X가 a와 b 사이에서 같은 정도로 값을 가질 때 균일분포(균등분포)를 따른다고 한다.
- $X \sim Uniform(a, b)$
- $f(x) = \frac{1}{b-a},\ a < x < b$
- $E(X) = \frac{a+b}{2},\ Var(X) = \frac{(b-a)^2}{12}$

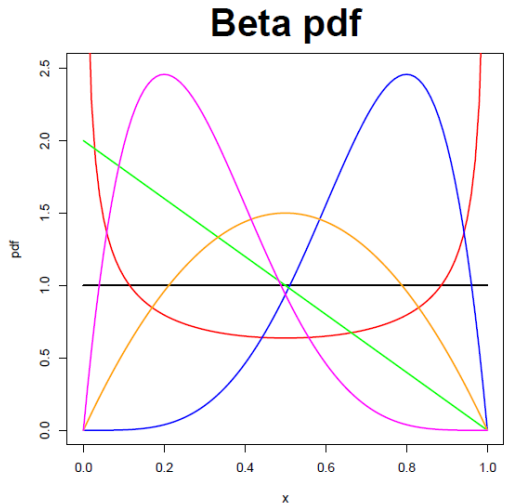
예제

핸드폰 배터리의 수명은 3년과 5년 사이의 균일분포를 따른다고 하자. 배터리의 평균 수명과 분산을 구하고 배터리가 4.5년 이상의 수명을 가질 확률을 구하여라.

베타 분포(Beta distribution)

- 연속확률분포중의 하나로 $0 \leq X \leq 1$인 확률변수가 다음의 확률밀도함수를 가지는 경우이다.
- $f(x) = f(x|\alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $x \in [0,1]$, $\alpha > 0, \beta > 0$.
- $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ 는 정규화 상수 (normalizing constant). 여기서 $\Gamma(x)$는 감마함수.
- $X \sim Beta(\alpha, \beta)$
- $E(X) = \frac{\alpha}{\alpha+\beta}$
- $\alpha = \beta = 1$이면 베타분포는 균일분포와 같다.

두 모수 α, β의 값에 따라 다양한 형태의 확률밀도함수가 나온다.



지수분포(Exponential distribution)

- 하나의 사건이 일어난 후 독립인 그 다음 사건이 일어날 때까지 기다리는 시간 (waiting time)을 모형화 한 분포
- $X \sim Exp(\lambda)$
- $f(x) = \lambda \exp(-\lambda x)$, $x > 0$. λ: rate parameter
- 또는 $f(x) = \frac{1}{\rho} \exp(-x/\rho)$. ρ: scale parameter
- $E(X) = \frac{1}{\lambda} = \rho$
- $Var(X) = \frac{1}{\lambda^2} = \rho^2$
- Memoryless property: $P(X > s + t | X > s) = P(X > t)$, $s, t > 0$.

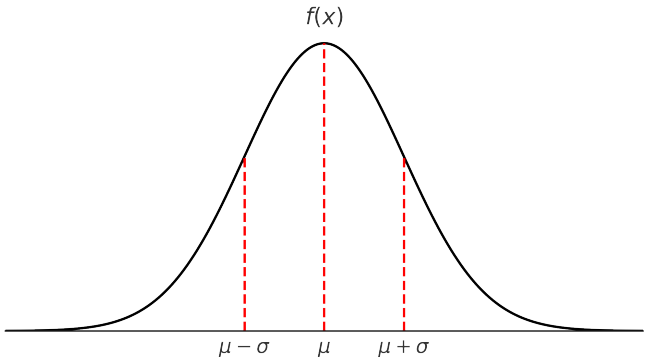
정규분포(Normal distribution)

- 가우스(Gauss, 1777-1855)에 의해 제시된 분포로서 가우시안분포(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.

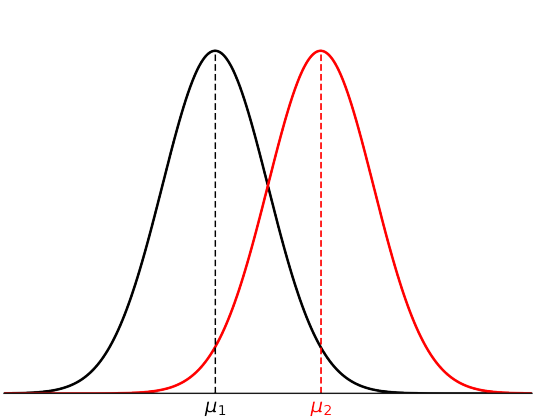
정규분포의 확률밀도함수

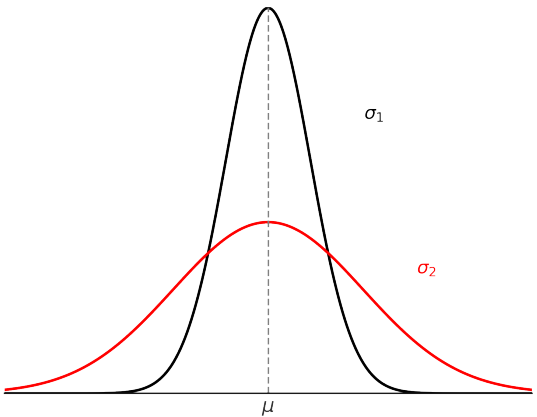
- $X \sim N(\mu, \sigma^2)$, μ: 평균, σ^2: 분산, $\tau^2 = 1/\sigma^2$: *precision*

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



정규분포의 성질





정규분포의 성질

- 선형변환을 해도 정규분포이다.

$$X \sim N(\mu, \sigma^2) \Longrightarrow aX + b \sim N(a\mu + b, a^2\sigma^2), a \neq 0$$

- 평균이 0이고 표준편차가 1인 정규분포를 표준정규분포(Standard Normal Distribution)라고 한다.

  보통 Z로 표기하므로, $Z \sim N(0, 1)$

- 표준화(Standardization)

$$X \sim N(\mu, \sigma^2) \Longrightarrow (X - \mu)/\sigma \sim N(0, 1)$$

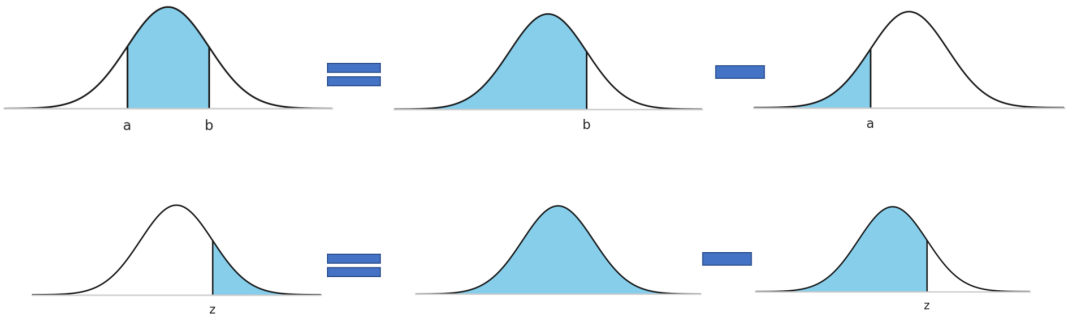
표준정규분포표($P(Z \leq z)$)

- 예) $P(Z \leq 0.93) = 0.8238$

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
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

표준정규분포에서 확률 구하기

- $P(a \leq Z \leq b) = P(Z \leq b) - P(Z < a)$
- $P(Z \geq z) = 1 - P(Z < z)$



정규분포에서 확률 구하기

- 일반적인 정규분포 $X \sim N(\mu, \sigma^2)$의 확률 계산시 표준화를 통해 표준정규분포표를 활용할 수 있다.

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

$(\alpha \times 100)$-th percentile (백분위수)

- αth 백분위수는 $P(Z \leq z) = \alpha$를 만족하는 z를 말하며 z_α로 표기하자.
- $Z_{0.975} = 1.96$. 즉, $P(Z < 1.96) = 0.975$.
- $Z_{0.995} = 2.58$
- 정규분포가 아니더라도, 같은 방식으로 정의할 수 있다.
- 교재에 따라 $P(Z > z) = \alpha$인 z를 z_α로 표기하기도 한다. 따라서, 문제에서 주어지는 z_α의 정의에 유의한다.

다변량 확률변수와 확률분포

여러종류의 데이터가 있을 때 불확실성을 측정하는 도구

- 여러 종류의 데이터가 관측되는 경우 (ex. 기온과 바람의 세기, 수학과 영어성적) 각각의 불확실성을 따로 보는것 보다는 같이 보는것이 필요하다.
- 이렇게 여러 종류의 데이터가 있는 경우 확률변수 여러개를 모아놓은 다변량 확률변수 (Multivariate Random Variable)의 실현값(또는 관측값)으로 본다.
- 다변량 확률변수의 분포를 다변량 확률분포(Multivariate Probability Distribution) 이라고 한다.

다변량 확률변수

- p개의 확률변수를 모아 놓은 다변량 확률변수는 벡터로 볼 수 있다.
- 각 원소 X_i 가 확률변수인 크기가 $p \times 1$인 벡터 $\boldsymbol{X} = (X_1, \cdots, X_p)^T$를 확률벡터 (Random Vector)라고 부른다.
- 확률벡터의 확률분포는 여러 확률변수들이 동시에 가질 수 있는 값들과 확률을 계산할 수 있게 해주는 것으로 결합확률분포(Joint Probability Distribution)이라고 한다.
- 먼저 두 개의 확률변수 X, Y로 이루어진 경우를 생각해 보자.

이산형 결합확률분포

- 확률변수 X, Y가 둘 다 이산형 확률변수 일 때의 확률분포로 결합확률질량함수(Joint pmf)를 생각할 수 있다.
- 결합확률질량함수는 다음과 같이 정의된다.

$$p_{X,Y}(x,y) = P(X = x, Y = y)$$

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

연속형 결합확률분포

- 확률변수 X, Y가 둘 다 연속형 확률변수 일 때의 확률분포로 결합확률밀도함수(Joint pdf)를 생각할 수 있다.
- 연속형 결합확률밀도함수는 다음과 같이 정의된다.

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy dx$$

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

* 결합누적확률분포함수(Joint CDF)로도 결합확률분포를 나타낼 수 있다.
$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$.

주변확률밀도함수(Marginal pdf)

- 여러 확률변수로 이루어져 있는 확률벡터의 결합확률분포를 알면 각 원소인 개별 확률변수의 분포도 알 수 있다.
- 결합확률분포와 구분하여 주변확률분포라고 부른다.
- 이산형 : $p_X(x) = \sum_y p(x, y)$
- 연속형 : $f_X(x) = \int f(x, y)dy$

결합확률분포의 예

서로 다른 동전 A,B,C를 동시에 던지는 실험에서 다음의 확률변수들을 생각해보자

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

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

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

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

Table: X의 주변확률분포

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

결합확률분포의 요약

- 결합확률분포의 요약으로 확률벡터의 기댓값과 분산을 생각할 수 있다.
- 두 개의 확률 변수로 이루어진 확률벡터 $\boldsymbol{Z} = (X, Y)$를 생각 해 보자.
- 기댓값은 확률변수가 가질 수 있는 값들의 중심으로 확률벡터 $\boldsymbol{Z}$의 기댓값은 각 원소의 기댓값의 벡터로 정의할 수 있다.

  $E(\boldsymbol{Z}) = (E(X), E(Y))$
- 분산은 확률변수가 가질수 있는 값들이 얼마나 퍼져 있는지를 나타내는 것으로 확률벡터의 경우 원소 X, Y가 각각 퍼져있는 정보와 같이 움직이는 정보가 필요하다. 이를 위해 두 확률 변수의 공분산(같이 움직이는 정도)을 정의한다.

공분산(Covariance)

- 두 확률 변수 X, Y의 공분산은 다음과 같이 정의 한다.

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

- 이때 $E(XY)$는 다음과 같이 구 할 수 있다.
- $E[XY] = \begin{cases} \sum_x \sum_y xyp(x, y) & \text{(이산형)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) & \text{(연속형)} \end{cases}$
- 이를 일반화하여 두 확률변수의 함수의 기댓값도 마찬가지로 구할 수 있다.
- $E[g(X, Y)] = \begin{cases} \sum_x \sum_y g(x, y)p(x, y) & \text{(이산형)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y) & \text{(연속형)} \end{cases}$

공분산 행렬(Covariance Matrix)

- 확률벡터 $\boldsymbol{Z} = (X, Y)$의 분산은 X, Y의 분산과 X, Y 사이의 공분산으로 이루어진 공분산 행렬(Covariance Matrix)로 표현할 수 있다.
- $Cov(\boldsymbol{Z}) = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{pmatrix}$

p개의 원소로 이루어진 확률벡터의 기댓값

- 두 개 이상의 확률변수로 이루어진 확률벡터의 기댓값도 두 개로 이루어진 확률벡터의 기댓값의 자연스러운 확장으로 다음과 같이 정의 된다.
- 확률벡터 $\boldsymbol{X} = (X_1, \cdots, X_p)^T$의 기댓값

$$E(\boldsymbol{X}) = E\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu},$$

여기서 $\mu_i = E(X_i)$.

p개의 원소로 이루어진 확률벡터의 공분산 행렬

- 마찬가지로 두 개 이상의 확률변수로 이루어진 확률벡터 $\boldsymbol{X}$의 공분산 행렬 Σ는 다음과 같이 정의한다.

  $Cov(\boldsymbol{X}) = E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)$

- $var(X_i) = \sigma_i^2$, $cov(X_i, X_j) = \sigma_{ij}$라고 하고, $\sigma_{ii} = \sigma_i^2$ 라고 하면, 공분산 행렬은 다음과 같이 표현된다.

$$\Sigma = cov(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

- Σ^{-1}: Precision matrix

상관계수(Correlation coefficient)

- 두 확률변수 X, Y의 공분산의 값은 X, Y의 단위에 의존하므로 단위에 의존하지 않으면서 두 확률변수가 같이 움직이는 정도를 나타내는 값으로 상관계수(Pearson's Correlation Coefficient)를 다음과 같이 정의한다.

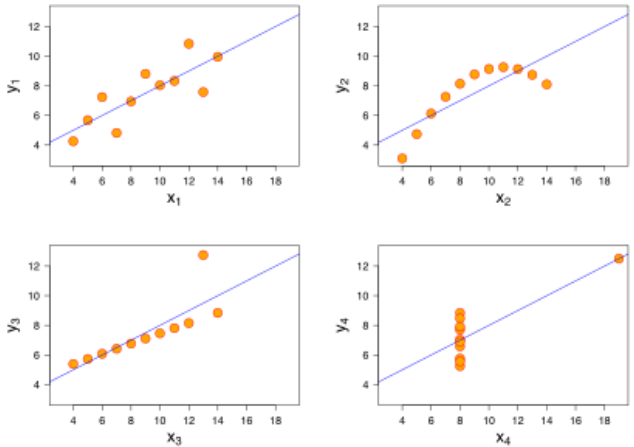
$$Corr(X, Y) = \rho_{XY} = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

공분산과 상관계수의 성질

확률변수 X, Y에 대한 공분산과 상관계수는 다음과 같은 성질들이 있다.

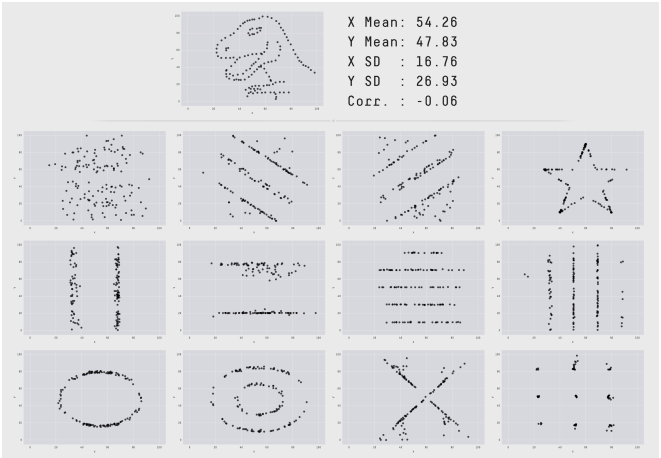
- $Cov(aX + b, cY + d) = acCov(X, Y)$
- $Corr(aX + b, cY + d) = sign(ac)Corr(X, Y)$
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$
- $-1 \leq \rho \leq 1$
- $Y = a + bX$ 이면 $\rho = \pm 1$: 즉, 상관계수는 두 확률변수의 선형관계의 정도를 나타낸다.

Anscombe's quartet



From Wikipedia

Dinosaurs



www.autodeskresearch.com/publications/samestats

두 확률변수의 독립성

- 여러 개의 데이터가 있을 때, 서로 독립적으로 관측되었다라는 개념은 확률변수들의 독립성으로 설명 할 수 있다.
- 두 확률변수 X, Y 가 다음을 만족할때 X와 Y는 서로 독립이라고 한다.
  모든 x, y에 대해

$$\begin{aligned} p(x,y) &= p_X(x)p_Y(y) \text{ (이산형)} \\ f(x,y) &= f_X(x)f_Y(y) \text{ (연속형)} \end{aligned}$$

- 두 사건 A, B가 서로 독립이라는 개념으로부터 유도된다.

두 확률변수가 독립일 때 성질

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

조건부 확률분포

- 하나의 확률변수의 값이 주어졌을때, 다른 확률변수의 확률분포는 조건부 확률분포 (Conditional Probability Distribution)로 설명할 수 있다.

  예) X는 어느 공장에서 하루동안 생산되는 제품중 불량품의 갯수, Y는 해당 공장의 생산라인 1에서의 하루동안 생산되는 제품중 불량품의 갯수. 이 공장에서 하루동안 생산되는 제품중 불량품의 갯수가 n개일때, 불량품중 생산라인 1에서 생산된 제품의 갯수의 분포는?

- 두개의 이산 확률변수 X, Y에 대하여 $X = x$가 주어졌을때의 Y의 조건부 확률질량함수:

$$p(y|x) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

- $p(y|x)$는 확률질량함수이다.

- 두개의 연속 확률변수 X, Y에 대하여 $X = x$가 주어졌을때의 Y의 조건부 확률밀도함수:

$$f(y|x) = \frac{f(x,y)}{f(x)}$$

- $f(y|x)$는 확률밀도함수이다.
- 하나가 이산 확률변수이고, 다른 하나가 연속 확률변수여도 상관없다.

조건부 확률분포 예제

- 불량 배터리의 사용시간은 평균 4시간이고 표준편차가 1시간인 정규분포를 따른다고 하고, 정상 배터리의 사용시간은 평균 5시간, 표준편차가 1시간인 정규분포를 따른다고 하자. 배터리가 불량일 확률은 1%라고 알려져 있다. 하나의 배터리를 랜덤추출하여 조사하였을때, 사용시간이 5시간 이상일 확률은?
- X :불량 여부: 1 (불량) or 0 (정상)

  Y : 배터리 사용시간

  $Y|X=1 \sim N(4,1)$, $Y|X=0 \sim N(5,1)$
- $P(Y \geq 5) = P(Y \geq 5|X=1)P(X=1) + P(Y \geq 5|X=0)P(X=0) = P(N(4,1) \geq 5) \times 0.01 + P(N(5,1) \geq 5) \times 0.99 = P(Z \geq 1) \times 0.01 + P(Z \geq 0) \times 0.99 = 0.1587 \times 0.01 + 0.5 \times 0.99 = 0.4966.$

두 개 이상의 확률변수들의 결합확률분포

- 두 개 이상의 확률변수로 이루어진 확률벡터 $\boldsymbol{X} = (X_1, \cdots, X_n)^T$의 결합확률분포로서 결합확률밀도함수와 결합누적확률분포함수를 다음과 같이 쓸 수 있다.

$$\begin{aligned} f(x_1, \cdots, x_n) & \\ F(x_1, \cdots, x_n) &= P(X_1 \leq x_1, \cdots, X_n \leq x_n) \end{aligned}$$

- 여러 개의 확률변수를 가지고도 조건부 확률분포를 얘기할 수 있다.

$$f(x_1, \cdots, x_k | x_{k+1}, \cdots, x_n)$$

데이터가 여러 개 있을 때

- n개의 데이터가 있다고 할 때 데이터들이 서로 독립적으로 같은 분포로부터 관측된 값이라고 가정하는 경우가 많다.
- 이때, 이론적인 내용이나 방법론들을 소개하기 위해서 데이터를 확률변수 자체로 보고 다음과 같이 표현한다.

  $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} f$ (or F)

- i.i.d. : independent identically distributed

- 이 경우 결합확률분포는 다음과 같이 표현된다.

  $f(x_1, \cdots, x_n) = f(x_1) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$

데이터가 여러 개 있을 때

- 데이터가 서로 독립이 아닐 때, 예를 들어 시간에 따라 관측되는 시계열 데이터를 생각해보자.
- $X_1, \cdots, X_{n-1}$이 시간 $t = 1, \cdots, n-1$일때의 값이라고 하면, 과거의 데이터 $X_1, \cdots, X_{n-1}$가 주어졌을때 그 다음 시간 $t = n$일때의 값인 X_n의 분포는 $f(x_n|x_{n-1}, \cdots, x_1)$ 로 표현될수 있다.
- 또한, 조건부 확률분포의 성질을 이용하면 결합확률분포는 다음과 같이 표현할 수 있다.

$$f(x_1, \cdots, x_n) = f(x_n|x_{n-1}, \cdots, x_1) f(x_{n-1}|x_{n-2}, \cdots, x_1) \times \cdots \times f(x_2|x_1) f(x_1)$$

다변량 정규분포(Multivariate Normal Distribution)

- 크기가 p인 확률벡터 $\boldsymbol{X} = (X_1, \cdots, X_p)$의 각 원소가 정규분포를 따르는 경우, $\boldsymbol{X}$의 분포를 다변량 정규분포라고 하고 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ 로 표시한다.
- 확률밀도함수는 다음과 같이 정의된다.

$$\begin{aligned} f(x_1, \cdots, x_p) &= f(x_1, \cdots, x_p | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \end{aligned}$$

- $|\Sigma|$는 Σ의 행렬식 (determinant)이다.
- 공분산 행렬 Σ는 양의 정 부호 행렬 (Positive Definite Matrix)이다.

- 각 원소가 표준정규분포이고 서로 독립이면 , $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$로 표현된다. 이때, $\boldsymbol{I}$는 단위행렬 (Identity Matrix)이다.
- 공분산 행렬은 $\Sigma = \boldsymbol{A}\boldsymbol{A}^T$로 분해(Cholesky Decomposition) 될 수 있다. 이 경우, $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$인 $\boldsymbol{X}$는 $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu}$의 관계가 성립한다.
- $\sigma_{ij} = E((X_i - \mu_i)(X_j - \mu_j)) = 0$이면, 즉 Σ의 (i, j)원소가 0이면, X_i, X_j는 서로 독립이다.

혼합 분포(믹스쳐 분포, **Mixure Distribution**)

- 여러 개의 분포의 선형결합으로 이루어진 분포를 혼합분포라고 한다.
- 이산확률분포에서는 k개의 이산확률분포의 선형결합으로 이루어진 다음과 같은 확률질량함수를 가진다. $p(x) = w_1 p_1(x) + \cdots + w_k p_k(x) = \sum_{i=1}^{k} w_i p_i(x)$
- 이때 $p_k(x)$는 확률질량함수이고, $w_i \geq 0$, $\sum w_i = 1$을 만족한다.
- 연속확률분포에서는 다음과 같은 확률밀도함수를 가진다.
  $f(x) = w_1 f_1(x) + \cdots + w_k f_k(x) = \sum_{i=1}^{k} w_i f_i(x)$.

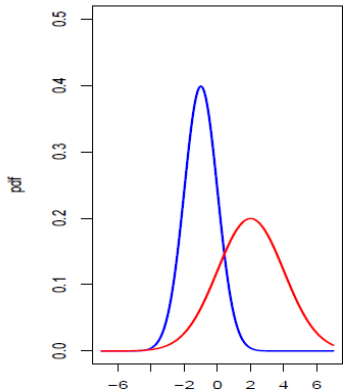
가우시안 혼합 분포(Gaussian Mixure Distribution)

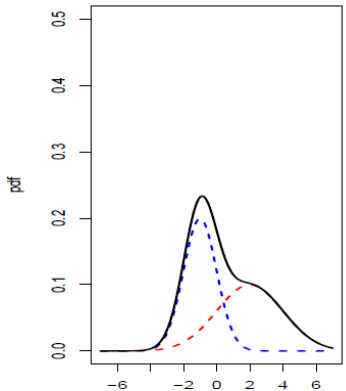
- $f_i(x)$들이 가우시안 확률밀도함수인 경우 가우시안 혼합 분포라고 한다.
- $\phi(x)$를 표준정규분포의 확률밀도함수라고 하자. 즉, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.
- $X \sim N(\mu, \sigma^2)$인 경우, X의 확률밀도함수는 $\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$로 표현할 수 있다.
- 이 경우 k개의 구성원을 가지는 가우시안 혼합 분포의 확률밀도함수는 다음과 같이 쓸 수 있다.
- $f(x) = \sum_{i=1}^{k} w_i \frac{1}{\sigma_i}\phi(\frac{x-\mu_i}{\sigma_i})$.

- $k = 2$인 경우 $f(x) = w_1 \frac{1}{\sigma_1}\phi(\frac{x-\mu_1}{\sigma_1}) + (1 - w_1)\frac{1}{\sigma_2}\phi(\frac{x-\mu_2}{\sigma_2})$
- $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f(x) = \sum_{i=1}^{k} w_i \frac{1}{\sigma_i}\phi(\frac{x-\mu_i}{\sigma_i})$, 즉, 가우시안 혼합 분포를 따르는 랜덤 추출된 데이터가 있다고 할때, 각 X_j는 w_i의 확률로 $N(\mu_i, \sigma_i^2)$을 따른다고 해석할 수 있다.
- 군집분석의 모델로 사용할 수 있다.

가우시안 혼합 분포 예

- 왼쪽: 파란선 $N(-1, 1^2)$, 빨간선 $N(2, 2^2)$
- 오른쪽: 파란점선 $0.5 \times N(-1, 1^2)$, 빨간점선 $0.5 \times N(2, 2^2)$ 까만선: $0.5 \times N(-1, 1^2) + 0.5 \times N(2, 2^2)$





The End