유용한 확률모형

Jong-June Jeon [1]

[1]Department of Statistics
University of Seoul

August 28, 2017

확률분포

누적분포함수

- 확률공간이 정의되었다고 가정하자. 즉, 어떤 사건 A에 대해서 $P(A)$를 항상 생각할 수 있다고 가정하자.
- 어떤 확률변수 X 주어졌을 때 $\Pr(X \leq x) = P(X \in (-\infty, x])$ 로 정의하면 $\Pr(X \leq x)$ 의 값을 모든 x에 대해 생각할 수 있다.
- $F(x) = \Pr(X \leq x)$ 이라고 놓으면 임의의 x에 대해서 $F(x)$ 는 값을 가지며, F는 실수에서 $[0, 1]$에 대응되는 함수다.
- 여기서 F를 확률변수 X의 누적분포함수라고 한다.

누적분포함수

- 누적분포함수를 알면 확률변수 X를 통해 얻어지는 임의의 사건에 대한 확률를 구할 수 있다.
  - $\Pr(a < X \leq b) = F(b) - F(a)$
  - 특별히 $\Pr(X = x) = F(x) - F(x-)$ (단, $F(x-) = \lim_{h \downarrow 0} F(x - h)$)
- 즉, 누적분포함수는 확률변수 X를 통해 얻어지는 불확실성에 대한 모든 정보(확률)를 제공한다.

확률밀도함수

- 연속형 확률변수에 대해서는 다음 조건을 만족시키는 함수 f 가 존재한다.
  - $F(x) = \int_{-\infty}^{x} f(t)dt$
- 사실 연속형 확률변수에 대해서는 $F'(x) = f(x)$ 가 성립한다.
- 따라서, 확률밀도함수 f를 안다는 것과 F를 안다는 것은 같다.
- 즉, 확률밀도함수는 확률변수 X를 통해 얻어지는 불확실성에 대한 모든 정보(확률)를 제공한다.

유용한 확률분포

- 일변량 분포
  - 베르누이 분포
  - 정규분포
  - 포아송분포
  - 감마분포
  - 베타분포
- 다변량 분포
  - 다변량 정규분포
  - 다항분포
  - 디리클레분포

분포에 대한 이해

- 물리적인 의미가 있나?
- 확률변수가 다른 확률변수로부터 유도된 것인가?
- 확률변수가 가질 수 있는 값은 무엇인가?
- 확률변수의 분포를 결정하는 모수는 무엇인가?
- 평균과 분산?

베르누이 분포

- ‘성공’ 혹은 ‘실패’와 같은 두 가지 결과만을 가지는 실험
- 기본적인 확률변수
- 확률변수는 0 또는 1의 값을 가짐
- 확률밀도함수는 $P(X=x)=\theta^x(1-\theta)^{1-x}$와 같이 주어지고

$$X \sim \text{Bernoulli}(\theta)$$

  (모수: $\theta \in (0,1)$ 로 표기한다.
- $E(X)=\Pr(X=1)=\theta$, $\text{Var}(X)=\theta(1-\theta)$

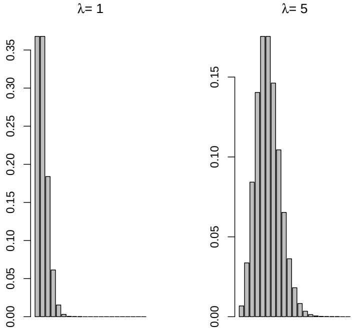
포아송 분포

- 단위시간 동안 랜덤하게 발생한 사건의 건수
- 지수분포와 관계가 있음
- 확률변수는 0을 포함한 자연수의 값을 가짐
- 확률밀도함수는

$$P(X = x) = \frac{\lambda^x \exp(-\lambda x)}{x!}$$

  와 같이 주어지고 $X \sim \text{Poisson}(\lambda)$ (모수: $\lambda > 0$) 로 표기한다.
- $\text{E}(X) = \lambda$, $\text{Var}(X) = \lambda$

포아송 분포



probability distribution of Poisson random variables

R을 이용한 분포함수 그리기: poisson distribution

- 누적분포함수: ppois
- 확률밀도함수: dpois
- 분위수: qpois
- 랜덤넘버생성: rpois

정규 분포

- 독립인 확률변수를 많이 더한 값의 분포
- 중심극한 정리
- 확률변수는 실수 값을 가짐
- 확률밀도함수는

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

와 같이 주어지고 $X \sim \mathrm{N}(\mu, \sigma^2)$ (모수: $\mu \in \mathbb{R}$, $\sigma > 0$) 로 표기한다. 누적확률은 아래와 같이 주어진다.

$$\Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$

정규 분포: 표본평균의 분포

```
> n = 1e+4
> z = rexp(n)
> x = c()
> for (i in 1:n)
+ {
+   idx = sample(1:n,25)
+   x[i] = mean(z[idx])
+ }
> hist(x)
```

```
> n = 1e+4
> z = runif(n)
> x = c()
> for (i in 1:n)
+ {
+   idx = sample(1:n,25)
+   x[i] = mean(z[idx])
+ }
> hist(x)
```

R을 이용한 분포함수 그리기: normal distribution

- 누적분포함수: pnorm
- 확률밀도함수: dnorm
- 분위수: qnorm
- 랜덤넘버생성: rnorm

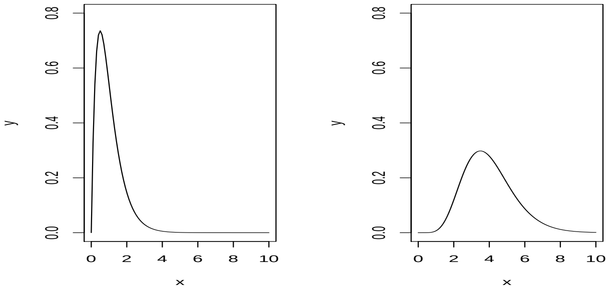
감마 분포

- 어떤 시스템에서 서비스 처리 시간에 대한 분포
- 독립인 지수분포의 합
- 확률변수는 양의 실수 값을 가짐
- 확률밀도함수는 다음과 같이 주어짐

$$f(x) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^{\alpha}}$$

  이 때, $X \sim \mathrm{Gamma}(\alpha, \beta)$ 로 표기함.
- $E(X) = \alpha\beta$, $Var(X) = \alpha\beta^2$

감마 분포



Left is the pdf with $\alpha = 2, \beta = 0.5$; Right is the pdf with $\alpha = 8, \beta = 0.5$

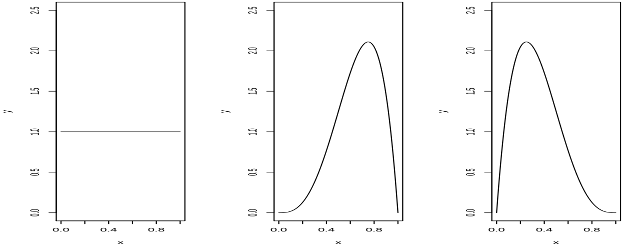
베타 분포

- 0과 1사이의 값을 가지는 어떤 데이터의 불확실성을 모형화하기 위해 사용
- 같은 규모모수를 가지고 독립인 감마분포를 따르는 두 확률변수의 비 $X \sim \text{Gamma}(\alpha_1, \beta)$ 고 $Y \sim \text{Gamma}(\alpha_2, \beta)$, X 와 Y가 독립일 때, $X/(X+Y)$ 의 분포.
- 확률변수는 0과 1사이의 값을 가짐
- 확률밀도함수는 다음과 같이 주어짐

$$f(x) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1-1}(1-x)^{\alpha_2-1}$$

  이 때, $X \sim \text{Beta}(\alpha_1, \alpha_2)$ 로 표기함.
- $\text{E}(X) = \alpha_1/(\alpha_1+\alpha_2)$, $\text{Var}(X) = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}$

베타 분포



Left is the pdf with $\alpha_1 = 1, \alpha_2 = 1$; center is $\alpha_1 = 4, \alpha_2 = 2$;
Right is the pdf with $\alpha_1 = 2, \alpha_2 = 4$

다변량분포

- 랜덤벡터: $X_1, \cdots, X_p$가 확률변수인 경우 $\mathbf{X} = (X_1, \cdots, X_p)'$ 를 랜덤벡터라고 한다.
- 랜덤벡터의 평균: $\mu_j = \mathrm{E}X_j$ 일때, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$를 랜덤벡터 $\mathbf{X}$의 평균벡터라고 한다.
- 공분산 행렬: $\mathrm{Cov}(X_j, X_k)$ 를 j행 k열의 원소로 갖는 행렬 $\boldsymbol{\Sigma}$를 $\mathbf{X}$의 공분산 행렬이라고 한다.
    - 공분산 행렬 $\boldsymbol{\Sigma}$의 대각원소는 무엇인가?
    - 공분산 행렬은 대칭행렬인가?

다변량 정규분포 $\mathbf{X} = (\mathbf{X_1}, \cdots, \mathbf{X_p})'$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)$, 그리고 $\boldsymbol{\Sigma}$ 를 $\mathbf{X}$ 의 공분산 행렬이라고 하자.

- 평균이 $\boldsymbol{\mu}$, 분산이 $\boldsymbol{\Sigma}$인 다변량 정규분포의 확률밀도함수는 다음과 같이 주어진다.

$$f(\mathbf{x}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

여기서 $|A|$는 정방행렬 A에 대한 행렬식으로 R에서는 det(A) 로 구할 수 있다.

R code for multivariate normal distribution

```
>   library(mvtnorm)
>   n = 50
>   mu.vec = c(1,1/2)
>   Sigma.mat = matrix( c(1,0.5,0.5,2),2,2)
>   x1 = x2 = seq(-3,3, length = n)
>   z <- matrix(0,n,n)
>   for (i in 1:n)
+     for (j in 1:n)
+       z[i,j] <- dmvnorm(c(x1[i],x2[j]), mu.vec, Sigma.mat)
>   contour(x1,x2,z)
```

다항분포

- 여러 개의 사건 중 하나의 사건이 발행하는 경우, 이를 묘사하는 확률 모형.
- 베르누이 분포의 확장
- p개의 사건 중 k번째 사건 발생 유무를 나타내는 확률변수를 $X_k \in \{0, 1\}$, $\Pr(X_k = 1) = \theta_k$ 라고 하자. $\mathbf{X} = (X_1, \cdots, X_p)'$ 라고 하면, 정의에 의해 항상 $\sum_{j=1}^{p} X_j = 1$ 이다.
- 확률밀도함수는 다음과 같이 주어진다.

$$\Pr(\mathbf{X} = (x_1, \cdots, x_p)') = \prod_{j=1}^{p} \theta_j^{x_j}$$

(단, $\sum_{j=1}^{p} x_j = 1$, $\sum_{j=1}^{p} \theta_j = 1$)

예시

- 타석에 들어선 타자의 기록: 1루타, 2루타,...
- 문서의 주제가 주어진 경우 하나의 단어의 출현 빈도: Latent Diriclet allocation 참조

디리클렛 분포

- 양의 값을 가지고 합이 1이 되는 랜덤벡터 (심플렉스: simplex)에 대한 분포
- 베르누이분포$\Rightarrow$ 베타분포 vs 다항분포 $\Rightarrow$ 디리클렛 분포
- $Y_j \sim \text{Gamma}(\alpha_j, \beta)$ for $j = 1, \cdots, p$ independently.

$$X_j = \frac{Y_j}{\sum_{k=1}^{p} Y_j} \quad (j = 1, \cdots, p)$$

Then, $\mathbf{X} = (X_1, \cdots, X_p)' \sim \text{Diriclet}(\boldsymbol{\alpha})$ $(\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)')$

예시

- 어떤 문서는 p개 주제로 이루어진다고 하자. 예를 들면, 특정 문서에서 우리가 생각할 수 있는 주제는 정치, 경제, 사회, 문화, 연예라고 하자.
- A 문서는 정치 80%, 경제 20% 의 주제로 이루어져 있다.
- B 문서는 경제 50%, 사회 30%, 문화 20% 로 이루어져 있다.
- 다섯개의 주제의 비율을 랜덤하게 생성하여, 특정 문서의 주제집합을 생성하고자 한다. 어떠한 확률분포 모형을 사용할까?

모수의 추론

기대값과 적률

어떤 확률변수 X의 확률밀도함수를 $f(x)$라고 하자.

- 기대값: $\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx$
- k차 적률: $\mathrm{E}X^k = \int_{-\infty}^{\infty} x^k f(x)dx$
- 일반적으로 $\mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$ 로 정의한다.

대수의 법칙

같은 분포를 따르는 독립인 확률변수의 평균은 참 평균으로 수렴한다.즉, $X_i \sim F$: iid 이고 $\mathrm{E}|X| < \infty$ 이면

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathrm{E}X$$

한편 만약 우리가 $\mathrm{E}(|g(X)|) < \infty$ 라는 것을 안다면

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathrm{E}g(X)$$

라는 것을 알 수 있다.

적률근사를 이용한 모수의 추론

- 정규분포 $X \sim N(\mu, \sigma^2)$
  - $\mathrm{E}X = \mu$
  - $\mathrm{Var}X = \mathrm{E}X^2 - (\mathrm{E}X)^2 = \sigma^2$
  - 만약 $N(\mu, \sigma^2)$를 따르는 랜덤샘플 n개를 관찰했다면, $\mathrm{E}X$ 와 $\mathrm{E}X^2$를 각각 $\frac{1}{n}\sum_{i=1}^{n} X_i$, $\frac{1}{n}\sum_{i=1}^{n} X_i^2$으로 근사할 것이다.
  - μ와 σ^2을 추정하는데, 적률의 근사값을 이용할 수 있을 것이다.
- 감마분포에 대한 모수 추론을 해 보자.

부록

지수, 로그, 자연대수

- 자연대수의 정의

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \simeq 2.718282$$

- 한편,

$$e^a = \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n$$

- e^a를 $\exp(a)$ 로 표기한다.
- 연습: $x = 0.1$, $\beta_0 = -1$, $\beta_1 = 2$ 일 때, $\exp(\beta_0 + \beta_1 x)$ 의 값을 구하시오

지수, 자연로그, 자연대수

- 자연로그의 정의

$$\log(b) = \int_1^b \frac{1}{t} dt, \quad b > 0$$

- $\log(b) = 2$ 가 되는 b는 무엇인가?
- $\log b = a$ 가 되는 b를 $\exp(a)$ 라고 정의한다. 한편 $\log b = a$ 가 되는 b는 유일하므로 $\exp(a) = b$ 가 되도록 해주는 a는 $\log b$다. 즉,

$$\log(\exp(b)) = \exp(\log(b)) = b$$

- $\log(\exp(\beta_0 + \beta_1 x))$ 는 얼마인가?

행렬

- 행렬의 표현 (n행, p열)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pp} \end{pmatrix}$$

  여기서 x_{ij} 는 행렬 $\mathbf{X}$의 i행, j열 원소를 나타낸다.
- 여기서 $(x_{i1}, x_{i2}, \cdots, x_{ip})$ 를 $\mathbf{X}$의 i 번째 행 벡터라고 한다.
- 한편

$$\begin{pmatrix} x_{j1} \\ \vdots \\ x_{jn} \end{pmatrix}$$

  을 행렬 $\mathbf{X}$의 j 번째 열 벡터라고 한다.

- 행렬의 전치 (transpose): 앞서 주어진 행렬 **X** 에 대하여 **X**의 전치행렬 $\mathbf{X}'$ 는 다음과 같이 주어진다.

$$\mathbf{X}' = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1p} & x_{2p} & \cdots & x_{pp} \end{pmatrix}$$

- **X**가 n행 p열 행렬이면, $\mathbf{X}'$은 p행 n열 행렬이다. $\mathbf{X}'$ 행렬의 i행, j열 원소는 **X** 행렬의 j행 i열 원소와 같다.

연습

- 행렬 $\mathbf{X}$ 가 다음과 같이 주어져있다.

$$\mathbf{X} = \begin{pmatrix} 1 & 3 & 0 & 0 \\ 1 & 2 & 5 & -1 \\ 4 & -2 & 3 & 0 \end{pmatrix}$$

  이 때, $\mathbf{X}'$ 를 구하여라.
- $\mathbf{X}'$ 의 2열과 $\mathbf{X}$ 의 2행이 같음을 확인하여라.
- $\mathbf{X}'$ 의 i열과 $\mathbf{X}$ 의 i행이 같음을 확인하여라.
- 위 문제에 이어서 $(\mathbf{X}')$ 의 전치행렬 즉, $(\mathbf{X}')'$ 를 구하여라.

행렬의 종류

- 정방행렬 (rectangular matrix): 행과 열의 수가 같은 행렬
- 대각행렬 (diagonal matrix): 정방행렬 중 대각원소를 제외한 나머지 원소가 모두 0인 행렬; $\mathrm{diag}(x_1, \cdots, x_p)$
- 단위행렬 (identity matrix): 대각행렬 중 모든 대각원소가 1인 행렬; I로 표기

행렬의 연산: R프로그래밍 행렬 참고

- 행렬의 덧셈
- 행렬의 곱셈
- 행렬의 스칼라 곱

역행렬

- 정방행렬 A 에 대해서

$$AB = BA = I$$

  를 만족하는 B가 존재하는 경우 행렬 B를 A의 역행렬이라고 부르고 A^{-1}로 표기한다.
- A^{-1}이 존재한다면 그것은 유일하다.

R 연습 a를 p행 1열인 열벡터, $\mathbf{X}$를 n행 p열 벡터라고 하자.

- $a'a$의 계산
- $\mathbf{X}'\mathbf{X}$ 의 계산
- $(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$의 계산
- $a'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{a}$ 의 계산

다변량정규분포의 pdf 계산 $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Mean vector is given by $\boldsymbol{\mu} = (0, 1, -1)'$
- Covariance matrix is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{pmatrix}$$

- 다변량정규분포의 pdf:

$$f(\mathbf{x}) \propto \mathbf{g}(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

$\mathbf{x} = (1, 0, 1/2)'$ 일 때, $g(\mathbf{x})$의 값을 R을 이용하여 계산하여라.

행렬식(determinant)

- 어떤 $\{1, \cdots, p\}$ 위에서 정의된 순열(permutation) σ하나를 생각하자. 예) $(1,2,3) \underset{\sigma}{\rightarrow} (3,2,1)$
- 위 예에서 $\sigma(1) = 3$, $\sigma(2) = 2$, $\sigma(3) = 1$ 이다.
- 다음과 같은 행렬 $\mathbf{X}$를 생각해보자.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0.5 \\ 1 & 1 & -1 \\ -1 & 2 & 4 \end{pmatrix}$$

  여기서 $x_{1\sigma(1)} \times x_{2\sigma(2)} \times x_{3\sigma(3)}$ 을 계산해보자.
- 가능한 σ의 종류는 총 몇가지인가?

행렬식(determinant): 순열 σ의 부호

- 호환 (transposition)은 두 개의 위치만 바꾸는 순열이다.
  - $(1,2,3) \underset{\sigma}{\rightarrow} (3,2,1)$은 호환인가?
  - $(1,2,3) \underset{\sigma}{\rightarrow} (2,3,1)$은 호환인가?
- 모든 순열은 호환을 여러번 연산함으로써 모두 표현 가능하다. 어떤 순열을 표현하기 위해 짝수번의 호환이 필요한 경우 그 순열이 양의 부호를 가진다고 하고, 홀수 번의 호환이 필요한 경우 음의 부호를 가진다고 한다.
- sign(σ) 라고 표현한다.

행렬식(determinant)의 정의
정방행렬 **X**에 대해서 행렬식은

$$|X| = \sum_{\sigma} \mathrm{sign}(\sigma) \prod_{i=1}^{p} x_{i\sigma(i)}$$

와 같이 정의한다.

2 × 2 행렬의 행렬식을 정의대로 계산해보자.

양의 정부호행렬(positive definite matrix)

- 0 이 아닌 모든 열벡터 a에 대해서 정방행렬 $\mathbf{X}$ 가 $a'\mathbf{X}a > 0$ 를 만족하면 $\mathbf{X}$를 양의 정부호행렬이라고 부른다.
- 공분산 행렬 $\mathbf{\Sigma}$ 가 양의 정부호행렬이면 항상 역행렬 $\mathbf{\Sigma}^{-1}$가 존재한다.
- 만약 $a'\mathbf{X}a \geq 0$면, 양의 준정부호행렬 (non-negative definite matrix)라고 한다.