이론 통계학

4장 기댓값

2021년 봄학기

전북대학교 통계학과

기댓값

기댓값

- 확률밀도함수나 확률분포함수는 확률변수의 전체적인 성격 설명
- 몇 개의 수치들 (평균, 분산,...)로 확률분포의 성질 요약

확률변수 X의 기댓값 또는 평균: $E(X)$, μ_X

확률변수 X (분포)의 중심위치 측도

$$E(X) = \begin{cases} \sum_{\text{all } x_i} x_i f_X(x_i) & X : \text{discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & X : \text{continuous} \end{cases}$$

(단, $E(|X|) < \infty$일 때 정의됨)

확률변수의 함수의 기댓값

$X \sim f_X(x), \quad g(X)$: a function of X

$g(X)$의 기댓값

$$E[g(X)] = \begin{cases} \sum_{\text{all } x_i} g(x_i) f_X(x_i) & X : \text{discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & X : \text{continuous} \end{cases}$$

- $X_1, X_2, \ldots, X_n \sim f(x_1, x_2, \ldots, x_n)$
- $g(X_1, X_2, \ldots, X_n)$: a function of $X_1, X_2, \ldots, X_n$

$g(X_1, X_2, \ldots, X_n)$의 기댓값

$$E[g(X_1, X_2, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

예제

‘성공’확률이 p인 베르누이 시행을 3번 독립적으로 반복하여, i번째 시행에서 성공이면 1, 아니면 0이 되는 확률변수를 $X_i, \quad i = 1, 2, \ldots, n$이라 하자.

- $E(X)$
- $E(e^{tX}]$
- $E[\sum_{i=1}^{n} X_i]$

분산과 표준편차

- 확률변수 X (분포)의 변동 측도
- X가 μ_X로부터 멀리 떨어져 있는 경향이 많을수록 $(X - \mu_X)^2$의 값도 커지는 경향

분산과 표준편차

$$\sigma_X^2 = Var(X) = E(X - \mu_X)^2 = E(X^2) - [E(X)]^2$$

$$\sigma_X = \sqrt{Var(X)} \text{ 또는 } SD(X)$$

- $Var(X)$의 단위는 X의 단위의 제곱으로 X의 변동 측도로 비합리적
- 분산의 제곱근으로 단위가 통일된 표준편차 사용하여 변수의 흩어진 정도 측정

표준화 (standardization)
평균이 0이고 표준편차가 1인 변수로 변환

$$Z = \frac{X - \mu_X}{\sigma_X}$$

예제
베르누이 확률변수를 표준화 하시오.

공분산

공분산 (covariance)

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

- $Cov(X, X) = Var(X)$
- 공분산은 X와 Y의 선형관계의 측도로 사용
- 측정 단위의 영향을 받음 $\rightarrow$ 이러한 단점을 보완하기 위해 상관계수 제안됨

상관계수 (correlation coefficient)

$$\rho = Corr(X, Y) = Cov\left[\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right]$$

- $-1 \leq \rho \leq 1$
- X와 Y의 선형관계의 측도로 사용
- 정리 4.8

상관계수

교재 예제 4.8
다음과 같은 결합확률밀도함수를 따르는 확률변수 X와 Y의 상관계수를 구하시오.

$$f(x,y) = (x+y)I(0 \le x \le 1)I(0 \le y \le 1)$$

확률변수의 선형 결합

통계적인 연구에서 자주 사용되는 통계량은 다음과 같은 확률변수의 선형결합 형태.

확률변수의 선형결합
확률변수 $X_1, \ldots, X_n$와 상수 $a_1, \ldots, a_n$에 대하여

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

확률변수의 선형 결합

$E(X_i^2) < \infty, E(Y_j^2) < \infty$인 확률 변수 $X_1, \ldots, X_n$과 $Y_1, \ldots, Y_m$, 상수 $a_1, \ldots, a_n$과 $b_1, \ldots, b_m$에 대하여

$$U_1 = \sum_{i=1}^{n} a_i X_i \text{ and } U_2 = \sum_{j=1}^{m} b_j Y_j$$

라고 하자.

- $E(U_1) = \sum_{i=1}^{n} a_i E(X_i)$
- $Var(U_1) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum\sum_{i<j} a_i a_j Cov(X_i, X_j)$
- $Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$

확률변수의 선형 결합

상수 a, b, c, d에 대하여

$$E(c) = c, \quad E(aX + b) = aE(X) + b$$

$$Var(c) = 0, \quad Var(aX + b) = a^2 Var(X)$$

$$Cov(aX + b, cY + d) = acCov(X, Y)$$

$$Corr(aX + b, cY + d) = sgn(ac)Corr(X, Y)$$

- 두 확률변수 X와Y가 서로 독립이면

$$E(XY) = E(X)E(Y), \text{ 즉 } Cov(X, Y) = 0$$

- 확률변수 $X_1, X_2, \ldots, X_n$이 서로 독립이면

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

적률생성함수

r차 적률 (rth moment)
확률변수 X의 함수 $g(X) = X^r$의 기대값

$$\mu_r' = E(X^r) = \int_{-\infty}^{\infty} x^r f(x)dx$$

적률생성함수(moment generating function: mgf)
확률변수 X의 함수 $g(X) = e^{tX}$의 기대값 $\Rightarrow$ t의 함수

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

- 적률생성함수를 보면 확률분포를 알 수 있음
- 모든 확률분포에 대하여 적률생성함수가 존재하는 것은 아니나, 우리가 다루는 분포에 대해서는 모두 존재.

적률생성함수(mgf)의 성질

- $M(0) = 1, M'(0) = E(X), \ldots, M^{(k)}(0) = E(X^k)$
- $Y = a + bX$에 대하여 $M_Y(t) = e^{at} M_X(bt)$
- X와 Y가 독립이면, $Z = X + Y$에 대하여

$$M_Z(t) = M_X(t) M_Y(t)$$

확률변수의 기댓값에 근거한 확률부등식

마코프 부등식
실함수 $u(x) > 0$라고 할 때, 확률변수 X는 임의의 상수 $c > 0$에 대하여

$$P[u(X) \geq c] \leq \frac{E[u(X)]}{c}$$

체비셰프 부등식
확률변수 X의 평균이 μ이고 분산이 $\sigma^2 \leqslant \infty$이면, 임의의 $k > 0$에 대하여

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

- 마코프 확률부등식에서 $u(x) = (x - \mu)^2, c = k^2\sigma^2$
- 표준편차 σ가 존재하는 변수의 평균과의 차이에 대한 확률의 한계 제공: $P[|X - \mu| < k\sigma] \geq 1 - \frac{1}{k^2}$

Cauchy-Schwarz inequality
두 확률변수 X와 Y에 대해서 $E(X^2) < \infty, E(Y^2) < \infty$가 만족되면,

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

$h(t) = E[(tX - Y)^2] \geq 0$이용