# 제 2 강

# 통계적 기초



## 확률변수 (Random Variable)

**확률변수 (r.v.):**
관측되기 전까지는 그 값이 알려지지 않은 변수.
확률변수의 값은 **확률적** 실험으로부터 결과된다.

확률적 실험은 실제 수행할 수 있는 실험 뿐 아니라 가상적 실험도 포함함 (ex. 주사위 던지기, [0,1] 실선에 점 던지기 )

확률변수는 그 변수의 모든 가능한 값들의 집합에 대해 정의된 알려지거나 알려지지 않은 어떤 확률분포의 존재가 연계됨

반면에, 임의의 변수는 그 값들에 연계되어 있는 확률분포를 가지지 않는다.



## 이산확률변수 (Discrete Random Variable)

이산확률변수:
이산확률변수는 정수를 이용해서 셀 수 있는 값들을 갖는 확률변수임

<u>예</u>: 다음의 복권으로 부터 얻을 수 있는 상금은
이산확률변수임:
일등: 1억원
이등: 1천만원
삼등: 1백만원
이 확률변수는 오직 네 가지 가능한 결과(값)들을 갖는다.(즉 1, 2, 3, 4, 로 셀 수 있음) :
0원; 1백만원; 1천만원; 1억원



## 연속확률변수 (Continuous Random Variable)

연속확률변수:
연속확률변수는 실선(real line)상의 구간(들)의 실수값들을 갖는 확률변수임

<u>예</u>:
GNP
통화공급량
이자율
쌀 가격
가계소득
의류에 대한 지출



## 더미변수 (Dummy Variable)

두 개의 가능한 값(대개 0과 1)만을 갖는 이산 확률변수를 더미변수 (또는, 이원변수, 모의변수)

더미변수들은 <u>질적(qualitative)</u> 차이를 나타내기도 함
성별 (0=남성, 1=여성),
고용 (0=실업, 1=취업),
거주지 (0=비서울., 1=서울),
소득수준 (0=저소득, 1=고소득).



## 확률(밀도)함수 (Probability (Density) Function)

이산확률변수

이산확률변수가 취하는 모든 가능한 값들에 대해 해당 값의 발생 확률을 대응시켜주는 함수를 확률(밀도)함수(probability function)라고 함

| 주사위 | x | f(x) |
|---|---|---|
| one dot | 1 | 1/6 |
| two dots | 2 | 1/6 |
| three dots | 3 | 1/6 |
| four dots | 4 | 1/6 |
| five dots | 5 | 1/6 |
| six dots | 6 | 1/6 |

## 확률(밀도)함수 (Probability (Density) Function)

계량경제학 2.7

**이산확률변수**

이산확률변수 X의 확률함수 f(x)는 확률변수 X가 x라는 값을 가질 확률을 다음과 같이 줌

$$f(x) = P(X=x)$$

따라서, $0 \leq f(x) \leq 1$

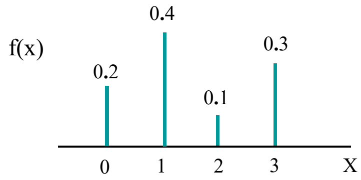
X가 n 개의 값들: $x_1, x_2, \ldots, x_n$을 가질 경우
$f(x_1) + f(x_2) + \ldots + f(x_n) = 1.$

## 확률(밀도)함수 (Probability (Density) Function)

계량경제학 2.8

**이산확률변수**

이산확률변수 X가 x라는 값을 취할 확률f(x)는 다음과 같이 높이로 나타낼 수 있음

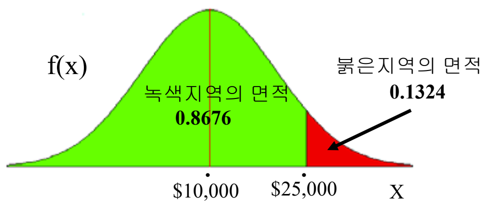



## 확률밀도함수 (Probability Density Function)

계량경제학 2.9

**연속확률변수**

연속확률변수는 확률을 나타내기 위해 높이가 아니라 f(x)가 나타내는 곡선 아래의 면적(area)을를 이용한다.




우리나라의 일인당 소득 X

## 확률밀도함수 (Probability Density Function)

계량경제학 2.10

**연속확률변수**

연속확률변수는 셀수없이 무한한 **(uncountably infinite)** 수의 값들을 가지며, 따라서 특정 값을 취할 확률은 **0**이다.

$$P[X = a] = P[a \leq X \leq a] = 0$$

확률은 면적으로 표현되나, 높이만으로는 면적을 갖지 않음

f(x)의 곡선 아래에 면적을 갖기 위해서는 X가 취하는 값의 구간이 필요함

## 확률밀도함수 (Probability Density Function)

계량경제학 2.11

**연속확률변수**

곡선 아래의 면적은 그 곡선을 만들어 낸 함수에 대한 적분값임:

$$P[a < X < b] = \int_a^b f(x)\,dx$$

연속확률변수의 경우 f(x) 그 자체가 아니라 f(x)의 적분이 면적을 정의하며 따라서 확률을 정의함 – f(x)를 연속확률변수 X의 확률밀도함수(pdf)라고 부름

## 누적분포함수 (Cumulative Distribution Function)

계량경제학 2.12

**누적분포함수**

확률변수 X의 누적분포함수(cdf)는 다음과 같이 정의된다. $F(x) \equiv P[X \leq x]$

이산적 r.v : $F(x) \equiv P(X \leq x) = \sum_{x_i \leq x} f(x_i)$

연속적 r.v : $F(x) \equiv P(X \leq x) = \int_{-\infty}^{x} f(x)dx$

cf) 모든 확률변수에 대해 cdf는 존재하지만, pdf가 존재하지 않는 확률변수도 있음.

## 누적분포함수 (Cumulative Distribution Function)

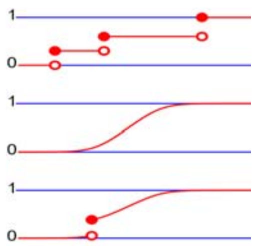


**누적분포함수**

이산적 r.v :
계단함수(step function)

연속적 r.v :
연속함수

이산_연속적 r.v :




누적분포함수는 non-decreasing function이며, 우측연속(right continuous)이다.

## 합산법칙 (Rule of Summation)



Rule 1: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

Rule 2: $\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$

Rule 3: $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

Σ는 선형작용자(linear operator)임을 의미함

## 합산법칙 (Rule of Summation)



Rule 4: $\sum_{i=1}^{n} (a x_i + b y_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$

Rule 5: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$

Rule 5에서 주어진 $\bar{x}$ 의 정의는 다음의 중요한 사실을 의미함

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

## 합산법칙 (Rule of Summation)



Rule 6: $\sum_{i=1}^{n} f(x_i) = f(x_1) + f(x_2) + \ldots + f(x_n)$

표기법: $\sum_{x} f(x_i) = \sum_{i} f(x_i) = \sum_{i=1}^{n} f(x_i)$

Rule 7: $\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i, y_j) = \sum_{i=1}^{n} [\, f(x_i, y_1) + f(x_i, y_2) + \ldots + f(x_i, y_m)]$

합산의 순서는 문제되지 않음을 의미 :

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i, y_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} f(x_i, y_j)$$

## 기대값 (Expected Value)



이산 확률변수 X의 기대값은 X의 모든 가능한 값을 대응되는 확률함수의 값으로 가중하여 합한 값임

⇨ $E[X] = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_n f(x_n)$

$= \sum_{i=1}^{n} x_i f(x_i)$

연속확률변수? : 합산기호 ⇨ 적분기호

$$E[X] = \int x f(x)\, dx$$

## 기대값 (Expected Value)



**경험적(Empirically) vs. 분석적(Analytically)**

경험적 (표본) 기대값 또는 평균:

$$\bar{x} = \sum_{i=1}^{T} x_i / T$$

단, **T**는 표본관측값들의 수

분석적(수학적) 평균:

$$E[X] = \sum_{i=1}^{n} x_i f(x_i)$$

단 **n** 은 X의 가능한 값들의 수.

## 확률변수 함수의 기대


X의 기대값:

$$EX = \sum_{i=1}^{n} x_i f(x_i)$$

X-제곱의 기대값:

$$EX^2 = \sum_{i=1}^{n} x_i^2 f(x_i)$$

확률변수의 함수가 취하는 값이 달라질 뿐 거기에 대응되는확률 $f(x_i)$ 는 변하지 않음에 주의!

X-세제곱의 기대값

$$EX^3 = \sum_{i=1}^{n} x_i^3 f(x_i)$$

## 확률변수 함수의 기대


$$EX = 0\,(.1) + 1\,(.3) + 2\,(.3) + 3\,(.2) + 4\,(.1) = 1.9$$

$$EX^2 = 0^2(.1) + 1^2(.3) + 2^2(.3) + 3^2(.2) + 4^2(.1) = 0 + .3 + 1.2 + 1.8 + 1.6 = 4.9$$

$$EX^3 = 0^3(.1) + 1^3(.3) + 2^3(.3) + 3^3(.2) + 4^3(.1) = 0 + .3 + 2.4 + 5.4 + 6.4 = 14.5$$

## 확률변수 함수의 기대


$$E[g(X)] = \sum_{i=1}^{n} g(x_i)\, f(x_i)$$

$$g(X) = g_1(X) + g_2(X)$$

$$E[g(X)] = \sum_{i=1}^{n} [g_1(x_i) + g_2(x_i)]\, f(x_i)$$

$$E[g(X)] = \sum_{i=1}^{n} g_1(x_i)\, f(x_i) + \sum_{i=1}^{n} g_2(x_i)\, f(x_i)$$

$$E[g(X)] = E[g_1(X)] + E[g_2(X)]$$

## 분산(Variance)


var(X) = X의 기대값을 중심으로 X가 취하는 값의 편차의 제곱의 기대값

$$\boxed{var(X) = E[(X - EX)^2]}$$

$$= E[X^2 - 2XEX + (EX)^2]$$

$$= E(X^2) - 2\,EX\,EX + E\,(EX)^2$$

$$= E(X^2) - 2\,(EX)^2 + (EX)^2$$

$$\boxed{= E(X^2) - (EX)^2}$$

## 분산(Variance)


이산확률변수 X의 분산:

$$var\,(X) = \sum_{i=1}^{n} (x_i - EX)^2 f(x_i)$$

표준편차(standard deviation) 는 분산의 제곱근임

## 결합확률밀도함수(Joint pdf)


결합확률밀도함수 f(x,y)는확률변수 X와 Y의 모든 가능한 값들의 쌍(pair)의 발생에 대응되는 확률을 제공함

## 결합확률밀도함수(Joint pdf)



결합 pdf
$f(x,y)$

| 자가주택 여부 \ 보유 자가용 수 | Y = 1 | Y = 2 |
|---|---|---|
| X = 0 | f(0,1) .45 | f(0,2) .15 |
| X = 1 | .05 f(1,1) | .35 f(1,2) |

## 결합확률밀도함수(Joint pdf)



실제 계산 예

$$E(XY) = \sum_i \sum_j x_i y_j \ f(x_i,y_j)$$

$$E[g(X,Y)] = \sum_i \sum_j g(x_i,y_j) \ f(x_i,y_j)$$

E(XY) = (0)(1)(.45)+(0)(2)(.15)+(1)(1)(.05)+(1)(2)(.35)=.75

## 한계확률밀도함수(Marginal pdf)



이산확률변수 X와 Y에 대한 한계확률(밀도)함수f(x) and f(y)는 각각 f(x,y)를 Y의 값들에 대해 합하거나(f(x)) X의 값들에 대해 합하여 구함(f(y))

$$f(x_i) = \sum_j f(x_i,y_j) \qquad f(y_j) = \sum_i f(x_i,y_j)$$

## 한계확률밀도함수(Marginal pdf)



| | Y = 1 | Y = 2 | X의 한계pdf : |
|---|---|---|---|
| X = 0 | .45 | .15 | .60 f(X = 0) |
| X = 1 | .05 | .35 | .40 f(X = 1) |
| Y의 한계 pdf : | .50 f(Y = 1) | .50 f(Y = 2) | |

## 조건부확률밀도함수(Conditional pdf)



Y=y로 주어졌을 때 X의 조건부확률밀도함수 f(x|y)와 X=x로 주어졌을 때 Y의 조건부확률밀도함수 f(y|x)는 각각f(x,y)를 f(y)로 나누거나 (f(x|y)), f(x)로 나누어 (f(y|x)) 얻음.

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

## 조건부확률밀도함수(Conditional pdf)



f(Y=1|X = 0)=.75 f(Y=2|X= 0)=.25

| | Y = 1 | Y = 2 | |
|---|---|---|---|
| X = 0 | .75 .45 | .25 .15 | .60 |
| f(X=0\|Y=1)=.90 f(X=1\|Y=1)=.10 | .90 .10 | .30 .70 | f(X=0\|Y=2)=.30 f(X=1\|Y=2)=.70 |
| X = 1 | .05 .125 | .35 .875 | .40 |
| | .50 | .50 | |

f(Y=1|X = 1)=.125 f(Y=2|X = 1)=.875

## 독립인 확률변수(Independent r.v.)



X와 Y의 결합pdf f(x,y)가 그 한계pdf f(x)와 f(y)의 곱으로 표시될 경우 X와 Y는 독립인 확률변수임

$$f(x_i,y_j) = f(x_i)\ f(y_j)$$

독립성을위해서 이 등식이 모든 i와 j의 쌍에 대해 성립해야 함

## 공분산(Covariance)



두 확률변수 X와 Y의 공분산은 이들 두 확률변수들 간의 선형관계의 정도를 측정함

$$cov(X,Y) = E[(X - EX)(Y-EY)]$$

분산은 공분산의 특별한 경우임에 주의.

$$cov(X,X) = var(X) = E[(X - EX)^2\ ]$$

## 공분산(Covariance)



$$cov(X,Y)\ = E\ [(X - EX)(Y-EY)]$$

$$\begin{aligned} cov(X,Y)\ &= E\ [(X - EX)(Y-EY)] \\ &= E\ [XY\ - X\ EY - Y\ EX + EX\ EY] \\ &= E(XY)\ - EX\ EY - EY\ EX + EX\ EY \\ &= E(XY)\ - 2\ EX\ EY + EX\ EY \\ &= E(XY)\ - EX\ EY \end{aligned}$$

$$cov(X,Y)\ =\ E(XY)\ - EX\ EY$$

## 상관(Correlation)



두 확률변수 X와 Y의 상관은 그들의 공분산을 각각의 표준편차의 곱으로 나누어 준 것임

$$\rho(X,Y)\ = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

상관(계수)는 단위와 무관한 값으로 -1과1 사이의 값

## 영의 공분산 및 상관 (Zero covariance and correlation)



독립인 확률변수들은 0의 공분산을 가지며 따라서 0의 상관을 가짐

그 역(converse)은 사실이 아님

## 확률변수 가중합계의 평균



E 역시 선형작용자임

확률변수들의 가중합의 기대값은 개별 항의 기대값들의 가중합과 같음

$$E[c_1X + c_2Y] = c_1EX + c_2EY$$

일반적으로 확률변수 $X_1, \ldots, X_n$ 에 대해:

$$E[c_1X_1+\ldots+ c_nX_n] = c_1EX_1+\ldots+ c_nEX_n$$

## 확률변수 가중합계의 분산



확률변수들의 가중합의 분산은 개별 항의 분산에 가중치의 제곱을 곱한 값들의 합에다 모든 확률변수들의 쌍의 공분산에 그 가중치들의 곱을 곱하고 2를 곱한 것의 합임

두 확률변수의 가중합:

$$V(c_1X + c_2Y) = c_1^2\,V(X) + c_2^2\,V(Y) + 2c_1c_2Cov(X,Y)$$

두 확률변수의 가중차:

$$V(c_1X - c_2Y) = c_1^2\,V(X) + c_2^2\,V(Y) - 2c_1c_2Cov(X,Y)$$

일반화:

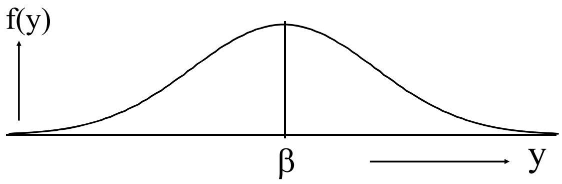
$$V(\sum_i c_iX_i) = \sum_i\sum_j c_ic_jCov(X_i,X_j) = \sum_i c_i^2V(X) + \sum\sum_{i\neq j} c_ic_jCov(X_i,X_j)$$

## 정규분포 (Normal Distribution)



$$Y \sim N(\beta,\sigma^2)$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y-\beta)^2}{2\sigma^2}\right]$$




## 정규분포 (Normal Distribution)



표준정규분포

$$Z = (Y - \beta)/\sigma$$

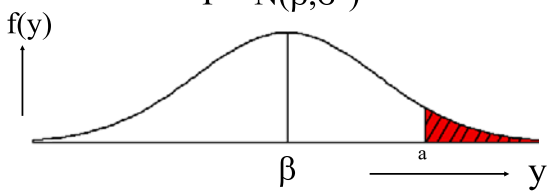
$$Z \sim N(0,1)$$

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{-z^2}{2}\right]$$

## 정규분포 (Normal Distribution)



$$Y \sim N(\beta,\sigma^2)$$

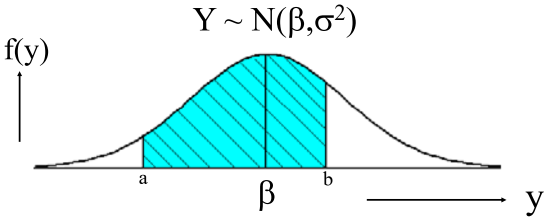



$$P[Y \geq a] = P\left[\frac{Y-\beta}{\sigma} \geq \frac{a-\beta}{\sigma}\right] = P\left[Z \geq \frac{a-\beta}{\sigma}\right]$$

## 정규분포 (Normal Distribution)



$$Y \sim N(\beta,\sigma^2)$$




$$P[a \leq Y \leq b] = P\left[\frac{a-\beta}{\sigma} \leq \frac{Y-\beta}{\sigma} \leq \frac{b-\beta}{\sigma}\right]$$

$$= P\left[\frac{a-\beta}{\sigma} \leq Z \leq \frac{b-\beta}{\sigma}\right]$$

## 정규분포 (Normal Distribution)



정규분포를 하는 확률변수들의 선형결합은 정규분포를 함

$$Y_1 \sim N(\beta_1,\sigma_1^2),\ Y_2 \sim N(\beta_2,\sigma_2^2),\ \ldots,\ Y_n \sim N(\beta_n,\sigma_n^2)$$

$$W = c_1Y_1 + c_2Y_2 + \ldots + c_nY_n$$

$$W \sim N[\,E(W),\ var(W)\,]$$

## 카이제곱 분포 (Chi-square Distribution)



$Z_1, Z_2, \ldots, Z_m$ 이 m개의 독립인 N(0,1) 확률변수들이고,
$V \equiv Z_1^2 + Z_2^2 + \ldots + Z_m^2$ 이면 $V \sim \chi^2_{(m)}$
즉 V 는 m의 자유도를 갖는 카이제곱분포임

평균: $E[V] = E[\chi^2_{(m)}] = m$

분산: $var[V] = var[\chi^2_{(m)}] = 2m$

## t 분포 (Student-t Distribution)



$Z \sim N(0,1),\ V \sim \chi^2_{(m)}$ 이고 Z와 V가 독립이면,

$$t \equiv \frac{Z}{\sqrt{V/m}} \sim t_{(m)}$$

즉 t는 m의 자유도를 갖는 t분포임

평균: $E[t] = E[t_{(m)}] = 0$ ,(m>1) 0에 대해 대칭임

분산: $var[t] = var[t_{(m)}] = m/(m-2)$, (m>2)

## *F*분포 (*F* Distribution)



$V_1 \sim \chi^2_{(m_1)},\ V_2 \sim \chi^2_{(m_2)}$ 이고 $V_1$과 $V_2$ 가 독립이라면,

$$F \equiv \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)}$$

즉 F는 $m_1$ 의 분자 자유도와 $m_2$ 의 분모 자유도를 갖는 F분포임

$F \sim F_{m,n} \rightarrow \frac{1}{F} \sim F_{n,m},\ t_m^2 \sim F_{1,m}$