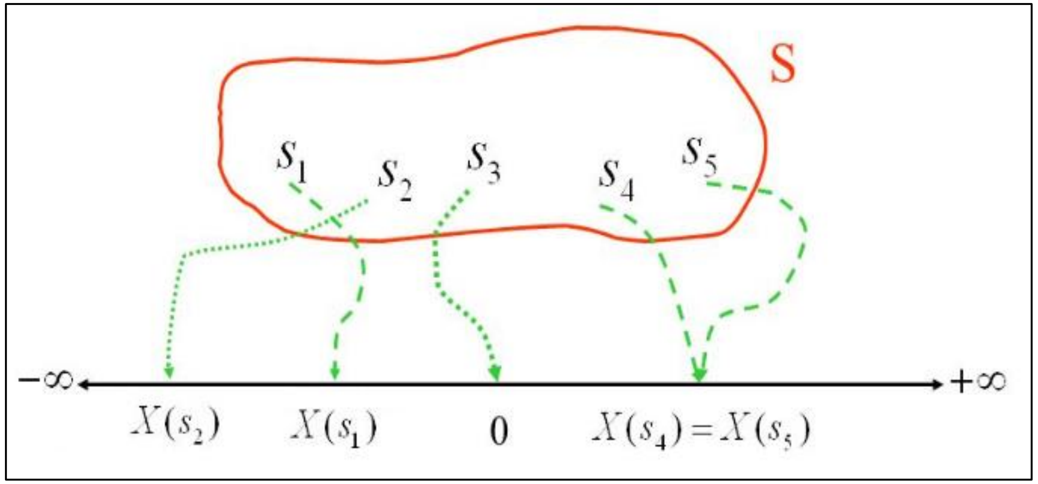
3. 확률변수와 확률분포함수

확률 변수와 확률 분포함수

- **확률변수와 분포함수**
  - **확률 실험** (Random Experiment)
    - 시행 마다 그 결과에 불확실성이 있는 실험.
  - **표본 공간** (Sample Space)
    - 확률 실험에서 모든 가능한 사건의 결과를 포함하는 집합. S로 표기.
    - $S = \{O_1, \dots, O_k\}$
    - 표본 공간의 원소는 서로 상호 배반적(mutually exclusive) 이고 완비적(exhaustive).
    - 표본 공간의 각 원소인 개별 사건이 발생할 확률이 정의될 수 있음.

확률 변수와 확률 분포함수

- **확률변수**(Random Variable) : 표본공간에서 정의된 실수값 함수.



  - **이산형 확률변수** (discrete random variable) : 확률변수가 취할 수 있는 모든 값은 하나씩 셀 수 있으며, 모두 유한개 또는 가산무한개인 경우.
  - **연속형 확률변수** (continuous random variable) : 주어진 구간에서 모든 값을 취할 수 있어 그 값의 수가 무한개인 경우.

- **확률분포함수**(probability distribution function) : 확률변수 X가 취할 수 있는 값 x와 그 확률을 대응시키는 관계.

확률 변수와 확률 분포함수

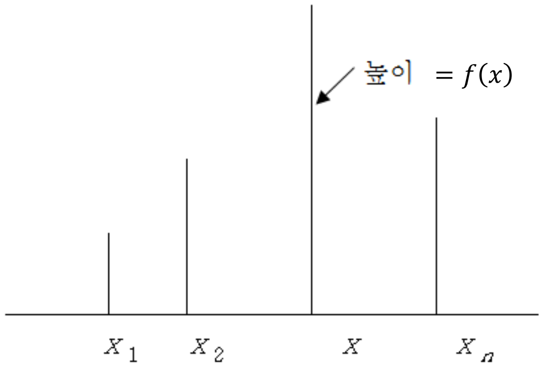
- **이산형 확률변수의 확률분포함수**
  - **확률질량함수**

    확률변수 X가 이산형인 경우 X가 취할 수 있는 값 $x_1, x_2, \ldots, x_n$의 각각에 대하여 확률 $P[X = x_1]$, $P[X = x_2]$,…을 대응시켜 주는 관계를 X의 **확률질량함수** (probability mass function, **pmf**) 라고 하며 $f(x)$로 표기한다.

$$f(x_i) = P[X = x_i], \qquad i = 1,2, \ldots, n$$

◆ **확률질량함수의 성질**

1) 모든 $i = 1,2, \ldots, n$에 대해 $0 \leq f(x_i) \leq 1$
2) $\sum_{i=1}^{n} f(x_i) = 1$



확률 변수와 확률 분포함수

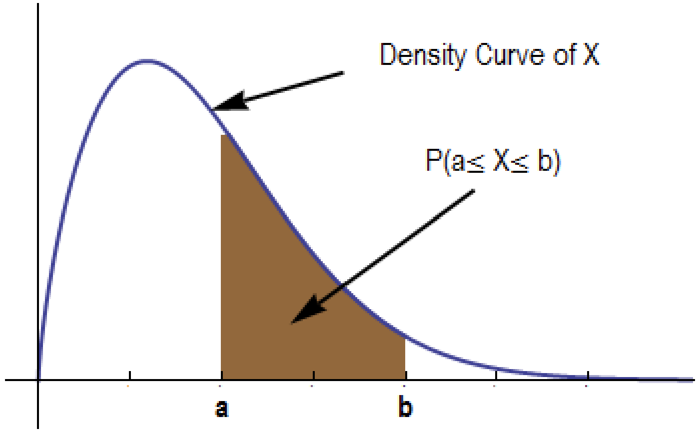
- **연속형 확률변수의 확률분포함수**
  - **확률밀도함수**

    확률변수 X가 연속형인 경우 X가 가질 수 있는 구간 $(-\infty, \infty)$위에서의 함수 $f(x)$ 가 다음을 만족할 때, 이를 X의 **확률밀도함수**(probability density function, **pdf**)라고 한다.

$$\int_a^b f(x)\,dx = P[a \le X \le b] \quad (단, -\infty < a < b < \infty)$$

◆ **확률밀도함수의 성질**

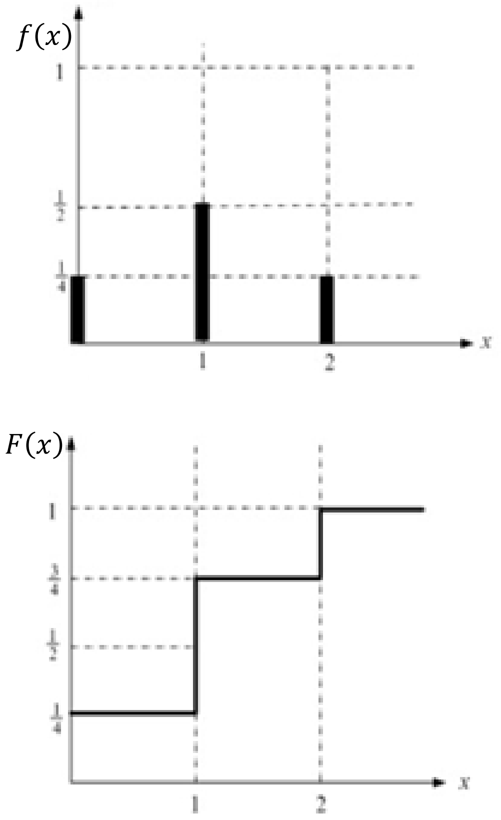
1) 모든 a, b 에 대해 $0 \le \int_a^b f(x)\,dx \le 1$
2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$



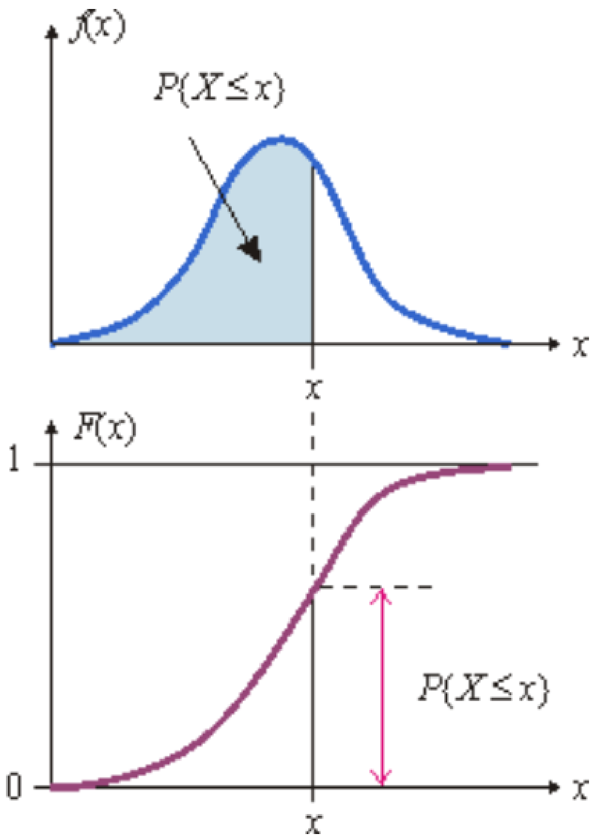
확률 변수와 확률 분포함수

- **누적분포함수** (cumulative distribution function, **cdf**)
  - $F(x) = P[X \le x]$ 로 정의됨

- 이산형



- 연속형



확률 변수와 확률 분포함수

- 예제

  A, B, C, D 기업 가운데, 어느 날에 주가가 상승한 기업의 수를 X라고 할 때, X는 다음과 같은 확률분포를 가진다고 한다.

X	0	1	2	3	4
$P[X = x]$	1/16	4/16	6/16	4/16	?

- 어떤 날 A, B, C, D 네 기업의 주가가 모두 상승할 확률을 구하여라.

- 어떤 날 주가가 상승한 기업의 수가 1보다 작거나 같을 확률을 구하여라.

확률 변수와 확률 분포함수

- 예제

  어느 주유소에서 매일 판매되는 휘발유의 양은 최소 2,000 갤런에서 최대 5,000 갤런 사이에 균일하게 분포한다고 하자.

  - 이 주요소의 일일 판매량이 2,500 ~ 3,000 갤런일 확률은 얼마인가?

  - 주유소가 최소 4,000 갤런을 판매 할 확률은 얼마인가?

  - 주유소가 정확히 2,500 갤런을 판매할 확률은 얼마인가?

확률변수에 관한 특성값

- **확률 분포의 특성값**

  이산형 확률변수 X가 확률질량함수 $f(x)$를 가진다고 할 때, 혹은 연속형 확률변수 X가 확률밀도함수 $f(x)$를 가진다고 할 때,

  - **기대값** : 분포의 무게중심, 중심 위치를 나타내는 특성값.

$$E[X] = \mu = \begin{cases} \sum_{all\ x} x f(x), & \text{이산형} \\ \int_{-\infty}^{\infty} x f(x)\, dx, & \text{연속형} \end{cases}$$

확률변수에 관한 특성값

- **기대값 연산자의 성질**

  - $E[g(X)] = \begin{cases} \sum_{all\ x} g(x)f(x), & \text{이산형} \\ \int_{-\infty}^{\infty} g(x)f(x)\,dx, & \text{연속형} \end{cases}$

  - $E[aX + b] = aE[X] + b$, a, b 는 상수

확률변수에 관한 특성값

- **분산** : 분포의 산포를 나타내는 특성값.

$$V[X] = \sigma^2 = E[(X-\mu)^2]$$

◆ **분산 연산자의 성질**

- $V[X] = E[X^2] - \mu^2$
- $V[aX+b] = a^2V[X]$, a, b 는 상수

확률변수에 관한 특성값

- **표준편차** :분산의 제곱근. 단위가 보정됨.

$$S[X] = \sigma = \sqrt{V[X]}$$

확률변수에 관한 특성값

- **예제**

  1년 뒤 주가를 현재 주가에 비교해서 받는 금액이 달라지는 금융상품이 있다고 하자. 이 금융상품에 100만원을 예금하였을 때, 1년 뒤 받을 수 있는 금액을 X(단위: 만원)라고 할 때, 다음 물음에 답하여라.

1년 뒤 상황	금액 (X)	확률
주가 상승률 < 0%	100	0.5
0% ≤ 주가 상승률 <15%	103	0.3
15% ≤ 주가 상승률	110	0.2

- X의 기대값 $E[X]$를 구하여라.

- X의 분산 $V[X]$를 구하여라.

확률변수에 관한 특성값

```
> x <- c( 100, 103, 110 )
> prb <- c( 0.5, 0.3, 0.2 )
> EX <- sum( x * prb )
> EX
[1] 102.9
> VX <- sum ( x^2 * prb ) - EX^2
> VX
[1] 14.29
```

확률변수에 관한 특성값

- 예제
  - 어느 애널리스트가 한 기업의 수익을 추정하였다. 대상이 되는 기업은 주당수익이 3000원이 될 가능성이 0.72, 주당 수익이 2000원이 될 가능성이 0.16, 주당수익이 1000원이 될 가능성이 0.12라고 한다. 이 기업 주식을 15주 보유하였을 때, 15주에 대한 수익의 기대값과 표준편차를 구하여라.

확률변수에 관한 특성값

```
> x <- c( 100, 103, 110 )
> x <- c( 3000, 2000, 1000 )
> prb <- c( 0.72, 0.16, 0.12 )
> EX <- 15 * sum( x * prb )
> EX
[1] 39000
```

확률변수에 관한 특성값

- 예제
  - 어느 컴퓨터 매장의 월별 판매금액의 평균 $25,000이고 표준 편차는 $4,000이다. 이 매장의 이익은 판매금액에 30%를 곱하고 $6,000의 고정 비용을 차감한 것으로 계산된다. 이 매장의 월 이익의 평균과 표준 편차를 구하여라.

두 개의 확률변수에 관한 확률분포

- **두 개의 확률변수에 관한 확률분포**

  두 개의 이산형 확률변수 X, Y가 주어지고 X가 가질 수 있는 값을 x, Y가 가질 수 있는 값을 y로 표기한다고 할 때,

  - **결합 확률분포**
    - 모든 가능한 (x, y)의 조합에 대해, $f_{X,Y}(x, y) = P(X = x \text{ and } Y = y)$를 만족하는 함수를 X, Y의 **결합확률질량함수**(joint probability mass function)로 정의.

    - **결합 확률질량함수의 성질**

      1) 모든 (x, y)의 조합에 대해 $0 \leq f_{X,Y}(x, y) \leq 1$
      2) $\sum_{all x} \sum_{all y} f_{X,Y}(x, y) = 1$

두 개의 확률변수에 관한 확률분포

- 예시

		Y	
		1	2
X	1	0.1	0.2
	2	0.3	0.4

두 개의 확률변수에 관한 확률분포

- **주변 확률분포**
  - X, Y의 결합확률질량함수$f_{X,Y}(x, y)$를 이용하면 X, Y 각각의 확률질량함수를 유도할 수 있으며, 이 때 $f_X(x)$와 $f_Y(y)$를 **주변확률질량함수**(marginal probability mass functions)라고 함.

$$f_X(x) = P(X = x) = \sum_{all y} f_{X,Y}(x, y) \qquad \& \qquad f_Y(y) = P(Y = y) = \sum_{all x} f_{X,Y}(x, y)$$

※ 연속형의 경우도 이와 유사한 방식으로 결합 및 주변 확률 밀도 함수(joint pdf, marginal pdf)가 정의됨.

두 개의 확률변수에 관한 확률분포

- 예제

Xavier 와 Yvette 은 어느 지역에 있는 두 명의 부동산중개업자이다. Xavier가 한달 동안 판매한 집의 개수를 X, Yvette 이 한달 동안 판매한 집의 개수를 Y라고 할 때, 과거 경험에 의하면 X와 Y는 다음과 같은 결합확률을 가진다고 하자.

		X		
		0	1	2
Y	0	.12	.42	.06
	1	.21	.06	.03
	2	.07	.02	.01

- Xavier가 한달 동안 판매한 집의 개수에 관한 주변확률분포를 표로 정리하여라.

- Yvette이 한달 동안 판매한 집의 개수에 관한 주변확률분포를 표로 정리하여라.

두 개의 확률변수에 관한 확률분포

```
> Pxy <- matrix( c( 0.12, 0.21, 0.07, 0.42, 0.06, 0.02, 0.06, 0.03, 0.01),
+                nrow=3, ncol=3 )
> rownames( Pxy ) <- c(0, 1, 2)
> colnames( Pxy ) <- c(0, 1, 2)
> Pxy
     0    1    2
0 0.12 0.42 0.06
1 0.21 0.06 0.03
2 0.07 0.02 0.01
> Px <- apply( Pxy, 2, sum )
> Px
  0   1   2
0.4 0.5 0.1
> Py <- apply( Pxy, 1, sum )
> Py
  0   1   2
0.6 0.3 0.1
```

두 개의 확률변수에 관한 특성값

- **두 개의 확률변수에 관한 특성값**
  - **기대값, 분산, 표준편차**

    - 확률변수 X와 Y의 기대값
      - $\mu_X (= E[X])$: X의 기대값
      - $\mu_Y (= E[X])$: Y의 기대값

    - 확률변수 X와 Y의 분산
      - σ_X^2 $(= V[X])$: X의 분산
      - $\sigma_Y^2 (= V[Y])$: Y의 분산

    - 확률변수 X와 Y의 표준편차
      - $\sigma_X (= S[X])$: X의 표준편차
      - $\sigma_Y (= S[Y])$: Y의 표준편차

두 개의 확률변수에 관한 특성값

- **두 개의 확률변수에 관한 특성값**
  - **공분산(covariance)**

    두 개의 확률변수 X, Y에 관한 선형적 연관성을 나타내는 특성값.

    $$COV[X,Y] = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

    - **공분산의 성질**
      - $COV[X,Y] = \sigma_{XY} = E[XY] - \mu_X\mu_Y$
      - $COV[aX,bY] = ab \cdot COV[X,Y]$

두 개의 확률변수에 관한 특성값

◆ **공분산의 성질**

1) 선형관계의 **방향**을 파악할 수 있음.
   - $\sigma_{XY} > 0$: X, Y 는 양의 선형 관계를 가짐
   - $\sigma_{XY} < 0$: X, Y 는 음의 선형 관계를 가짐

2) 선형관계의 **강도**는 공분산 만으로 파악할 수 없음. $COV[X, Y]$가 가질 수 있는 값의 범위는 X와 Y의 스케일에 의존하기 때문.
   - $-\sigma_X\sigma_Y \le \sigma_{XY} \le \sigma_X\sigma_Y$, by Cauchy−Schwarz inequality

3) 자료의 단위를 변경하면 공분산 값도 변함.
   - $X' = aX + b$ 이고 , $Y' = cY + d$ 일 때, $\sigma_{X'Y'} = ac \cdot \sigma_{XY}$

두 개의 확률변수에 관한 특성값

- **상관계수(correlation coefficient)** :
  두 개의 확률변수 X, Y의 선형적 연관성의 방향과 강도를 동시에 파악할 뿐 아니라 자료의 단위에도 의존하지 않도록 공분산을 보정한 지표.

$$CORR[X,Y] = \rho_{XY} = \frac{COV[X,Y]}{S[X]S[Y]} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- **상관계수의 성질**
  - $-1 \le \rho_{XY} \le 1$

두 개의 확률변수에 관한 특성값

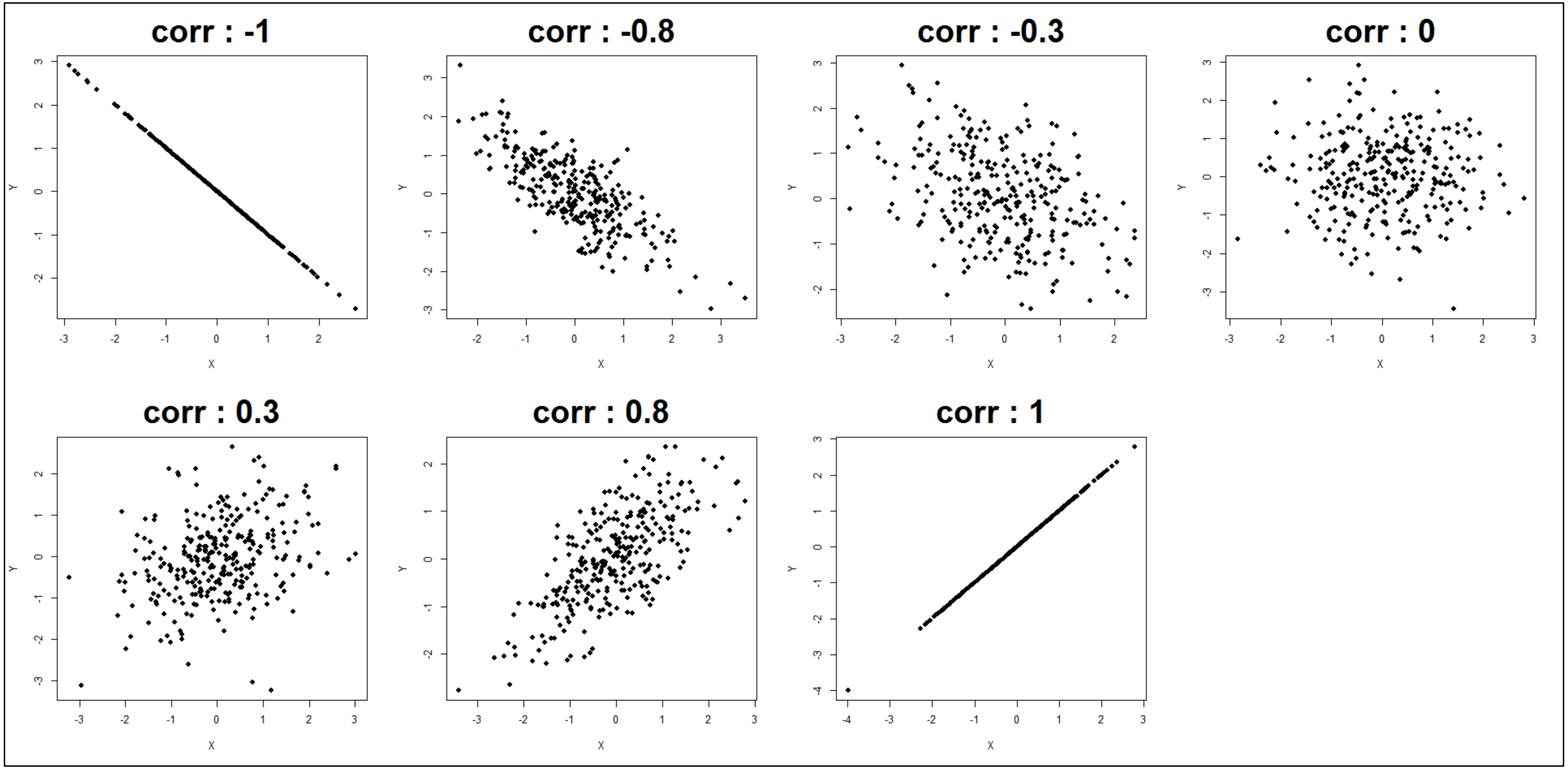
◆ 상관계수의 성질

1) 선형관계의 **방향**을 파악할 수 있음.
   - $\rho_{XY} > 0$: X, Y 는 양의 선형 관계를 가짐
   - $\rho_{XY} < 0$: X, Y 는 음의 선형 관계를 가짐

2) 선형관계의 **강도**를 파악할 수 있음.
   - $|\rho_{XY}| \approx 0$: 강도가 약함.
   - $|\rho_{XY}| \approx 1$: 강도가 강함.

3) 자료의 단위를 변경해도 상관계수 값은 변하지 않음.
   - $X' = aX + b$ 이고 , $Y' = cY + d$ 이고, $ac > 0$ 일 때, $\rho_{X'Y'} = \rho_{XY}$

두 개의 확률변수에 관한 특성값



두 개의 확률변수에 관한 특성값

- 두 확률변수 X 와 Y의 합 또는 차에 관한 특성값
  - $E[X \pm Y] = E[X] \pm E[Y]$
  - $V[X \pm Y] = V[X] + V[Y] \pm 2COV[X, Y]$
  - $V[X \pm Y] = V[X] + V[Y]$, 단, X 와 Y는 서로 독립.

두 개의 확률변수에 관한 특성값

- n개의 확률변수 $X_1, X_2, \dots, X_n$의 합에 관한 특성값
  - $E[\sum_{i=1}^{n} X_i] = E[\sum_{i=1}^{n} E[X_i]]$
  - $V[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V[X_i] \pm \sum_{i \neq j} COV[X_i, X_j]$
  - $V[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V[X_i]$, 단, $X_1, X_2, \dots, X_n$은 서로 독립.

두 개의 확률변수에 관한 특성값

- Xavier 와 Yvette에 관한 예제 (계속)

		X		
		0	1	2
Y	0	.12	.42	.06
	1	.21	.06	.03
	2	.07	.02	.01

- Xavier가 한달 동안 판매한 집의 개수에 대한 기대값과 표준편차를 구하여라.

- Yvette이 한달 동안 판매한 집의 개수에 대한 기대값과 표준편차를 구하여라.

두 개의 확률변수에 관한 특성값

- Xavier 와 Yvette에 관한 예제 (계속)

		X		
		0	1	2
Y	0	.12	.42	.06
	1	.21	.06	.03
	2	.07	.02	.01

  - Xavier와 Yvette의 한달 동안 판매한 집의 개수에 대한 공분산과 상관계수를 구하여라.

두 개의 확률변수에 관한 특성값

- Xavier 와 Yvette에 관한 예제 (계속)

		X		
		0	1	2
Y	0	.12	.42	.06
	1	.21	.06	.03
	2	.07	.02	.01

- 두 중개업자에 의해 한달 동안 판매된 집의 총 개수에 대한 기대값과 분산을 구하여라.

두 개의 확률변수에 관한 특성값

```
> Ex <- sum( (0:2) * Px )
> Sx <- sqrt( sum( (0:2)^2 * Px ) - Ex^2 )
> Ex
[1] 0.7
> Sx
[1] 0.6403124
> Ey <- sum( (0:2) * Py )
> Sy <- sqrt( sum( (0:2)^2 * Py ) - Ey^2 )
> Ey
[1] 0.5
> Sy
[1] 0.6708204
```

두 개의 확률변수에 관한 특성값

```
> COVxy <- sum( outer( 0:2, 0:2 ) * Pxy ) - Ex*Ey
> CORxy <- COVxy / ( Sx * Sy )
> COVxy
[1] -0.15
> CORxy
[1] -0.3492151
> Ex + Ey
[1] 1.2
> Sx^2+Sy^2+2COVxy
[1] 0.56
```

두 개의 확률변수에 관한 특성값

- 예제
  - 주식 A의 수익률을 X라고 할 때 X의 기대값은 4, 분산은 100이다. 또한 주식 B의 수익률을 Y라고 할 때 Y의 기대값은 7, 분산이 144이다. 두 주식 A와 B의 수익률은 서로 독립이라고 하자. $Z = 0.5X + 0.5Y$로 정의되는 Z의 기대값과 분산은 얼마인가?

두 개의 확률변수에 관한 특성값

- 예제
  - X, Y, Z 3사의 주식으로 포트폴리오가 구성되어 있으며, 주식의 구성비율 및 기대수익률과 이들 3사 수익률 간의 분산과 공분산이 다음과 같다고 하자. 이 포트폴리오의 기대수익률과 분산을 구하시오.

기업	가중치(주식구성비)	기대수익률
X	50%	8%
Y	10%	12%
Z	40%	16%

$$V[X] = 0.20, V[Y] = 0.30, V[Z] = 0.25,$$

$$COV[X, Y] = 0.15, COV[X, Z] = 0.10, COV[Y, Z] = 0.12$$

두 개의 확률변수에 관한 특성값

```
> wt <- c(0.5, 0.1, 0.4)
> Evec <- c(0.08, 0.12, 0.16)
> Epf <- sum( wt * Evec )
> Epf
[1] 0.116
> Vvec <- c(0.2, 0.3, 0.25)
> Cmat <- diag( Vvec )
> Cmat[1, 2] <- Cmat[2, 1] <- 0.15
> Cmat[1, 3] <- Cmat[3, 1] <- 0.10
> Cmat[2, 3] <- Cmat[3, 2] <- 0.12
> Cmat
     [,1] [,2] [,3]
[1,] 0.20 0.15 0.10
[2,] 0.15 0.30 0.12
[3,] 0.10 0.12 0.25
> Vpf <- wt %*% Cmat %*% wt
> Vpf
       [,1]
[1,] 0.1576
```