7.3 분산과 표준편차

이 절에서는 확률분포함수의 모양을 설명하는 두 번째 특성인 분산을 공부한다. 분산은 확률분포함수에서 확률이 모여있는지 퍼져있는지를 나타내는 값이다. 기댓값이 확률변수에서 어떤 값이 나올지를 예측한 것이라면 분산은 그 예측의 정확도 혹은 신뢰성을 표현한 것이라고 볼 수 있다.

확률분포의 분산

확률밀도함수 $p(x)$의 수식을 알고 있다면 이론적인 분산을 구할 수 있다. 분산을 구하는 연산은 영어 Variance의 앞글자를 따서 $\mathrm{Var}[\cdot]$로 표기하고 이 연산으로 계산된 분산값은 σ^2으로 표기한다.

$$\sigma^2 = \mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2] \tag{7.3.1}$$

이산확률변수의 분산은 평균으로부터 표본 데이터까지 거리의 제곱을 확률질량함수 $p(x)$로 가중하여 더한 값이다.

$$\sigma^2 = \sum_{x_i \in \Omega} (x_i - \mu)^2 p(x_i) \tag{7.3.2}$$

연속확률변수의 분산은 평균으로부터 표본 데이터까지 거리의 제곱을 확률밀도함수 $p(x)$로 가중하여 적분한 값이다.

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \tag{7.3.3}$$

분산의 성질

분산은 다음과 같은 성질을 만족한다.

- 분산은 항상 0 또는 양수이다.

$$\mathrm{Var}[X] \geq 0 \tag{7.3.4}$$

- 확률변수가 아닌 상수 값 c에 대해 다음 식이 성립한다.

$$\mathrm{Var}[c] = 0 \tag{7.3.5}$$

$$\mathrm{Var}[cX] = c^2 \mathrm{Var}[X] \tag{7.3.6}$$

또한 기댓값의 성질을 이용하여 다음 성질을 증명할 수 있다.

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[X^2] - \mu^2 \tag{7.3.7}$$

또는

$$\mathrm{E}[X^2] = \mu^2 + \mathrm{Var}[X] \tag{7.3.8}$$

(증명)

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(X-\mu)^2] \\ &= \mathrm{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathrm{E}[X^2] - 2\mu\mathrm{E}[X] + \mu^2 \\ &= \mathrm{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathrm{E}[X^2] - \mu^2\end{aligned} \tag{7.3.9}$$

두 확률변수의 합의 분산

두 확률변수 X, Y의 합의 분산은 각 확률변수의 분산의 합과 다음과 같은 관계가 있다.

$$\mathrm{Var}\left[X+Y\right] = \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right] + 2\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{7.3.10}$$

마지막 항은 양수도 될 수 있고 음수도 될 수 있다.

이 식의 증명은 다음과 같다. 우선 확률변수 $X+Y$의 기댓값은 기댓값의 성질로부터 각 확률변수의 기댓값의 합과 같다.

$$\mathrm{E}[X+Y] = \mu_X + \mu_Y \tag{7.3.11}$$

분산의 정의와 기댓값의 성질로부터 다음이 성립한다.

$$\begin{aligned}\mathrm{Var}\left[X+Y\right] &= \mathrm{E}\left[(X+Y-(\mu_X+\mu_Y))^2\right] \\ &= \mathrm{E}\left[((X-\mu_X)+(Y-\mu_Y))^2\right] \\ &= \mathrm{E}\left[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)\right] \\ &= \mathrm{E}\left[(X-\mu_X)^2\right] + \mathrm{E}\left[(Y-\mu_Y)^2\right] + 2\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right] + 2\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]\end{aligned} \tag{7.3.12}$$

확률변수의 독립

두 확률변수가 서로 독립(independent)이라는 것은 두 확률변수가 가질 수 있는 모든 사건의 조합에 대해 결합사건의 확률이 각 사건의 확률의 곱과 같다는 뜻이다. 쉽게 생각하면 **두 확률변수가 서로에게 영향을 미치지 않는다**라는 의미로 생각해도 된다. 예를 들어 주사위를 두 번 던져 각각 나오는 값을 나타내는 확률변수 X_1과 X_2는 서로 독립이다.

독립의 반대, 즉 **두 확률변수에서 하나의 확률변수의 값이 특정한 값이면 다른 확률변수의 확률분포가 영향을 받아 변하게 되면 종속(dependent)**이라고 한다. 쉽게 생각하면 **두 확률변수가 서로에게 영향을 미치는 경우**이다. 예를 들어 주사위를 두 번 던져 나오는 값의 합은 각각의 주사위에서 나온 값에 종속적이다.

연습 문제 7.3.1

1. 서로 독립이라고 생각되는 두 확률변수의 예를 들어라.
2. 서로 종속이라고 생각되는 두 확률변수의 예를 들어라.

두 확률변수 X, Y가 서로 독립이면 다음 식이 성립한다.

$$\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = 0 \tag{7.3.13}$$

왜 이 등식이 성립하는가는 추후 설명하기로 한다. 이 등식을 이용하면 서로 독립인 두 확률변수의 합의 분산은 각 확률변수의 분산의 합과 같다는 것을 보일 수 있다.

$$\mathrm{Var}\left[X+Y\right] = \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right] \tag{7.3.14}$$

연습 문제 7.3.2

1. 넘파이를 사용하여 숫자 100개를 무작위로 생성하여 표본집합을 구한다. 이 표본집합을 확률변수 X_1의 표본이라고 하자.
2. 같은 방식으로 숫자 100개를 생성하며 확률변수 X_2의 표본집합을 구한다.
3. 두 확률변수의 표본 쌍의 값을 더하여 확률변수 $X_1 + X_2$의 표본집합을 구한다.
4. $X_1 + X_2$의 표본분산과 X_1, X_2의 표본분산값의 합을 각각 계산하여 두 값이 비슷함을 보여라.

표본평균의 분산

확률변수 X의 표본평균 $\bar{X}$도 확률변수이고 그 기댓값 $\mathrm{E}[\bar{X}]$은 원래 확률변수 X의 기댓값 $\mathrm{E}[X]$과 같다는 것을 증명한 적이 있다.

$$\mathrm{E}[\bar{X}] = \mathrm{E}[X] \tag{7.3.15}$$

표본평균 $\bar{X}$의 분산 $\mathrm{Var}[\bar{X}]$은 원래 확률변수 X의 분산 $\mathrm{Var}[X]$과 다음 관계를 가진다.

$$\mathrm{Var}[\bar{X}] = \frac{1}{N}\mathrm{Var}[X] \tag{7.3.16}$$

따라서 **표본평균을 계산한 표본 개수가 커지면 표본평균의 값의 변동은 작아진다**. 표본의 수가 무한대가 되면 표본평균의 값은 항상 일정한 값이 나온다. 즉 확률적인 값이 아니라 결정론적인 값이 된다.

증명은 다음과 같다.

$$
\begin{aligned}
\mathrm{Var}[\bar{X}] &= \mathrm{E}\left[\left(\bar{X} - \mathrm{E}\left[\bar{X}\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(\bar{X} - \mu\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{N}N\mu\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{N}\left(\sum_{i=1}^{N} X_i - N\mu\right)\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - \mu)(X_j - \mu)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left[(X_i - \mu)(X_j - \mu)\right]
\end{aligned}
\tag{7.3.17}
$$

i번째 표본값은 j번째($i \neq j$) 표본값에 영향을 미치지 않으므로 X_i와 X_j ($i \neq j$)는 독립이다. 따라서

$$
\mathrm{E}\left[(X_i - \mu)(X_j - \mu)\right] = 0 \;\; (i \neq j) \tag{7.3.18}
$$

라는 사실을 이용하면 $i = j$인 항, 즉 제곱항만 남는다.

$$
\begin{aligned}
\mathrm{Var}[\bar{X}] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}\left[(X_i - \mu)^2\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}\left[(X - \mu)^2\right] \\
&= \frac{1}{N^2}N\mathrm{E}\left[(X - \mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[(X - \mu)^2\right] \\
&= \frac{1}{N}\mathrm{Var}[X]
\end{aligned}
\tag{7.3.19}
$$

위 식이 의미하는 바는 다음과 같다.

- 데이터를 생성하는 확률변수 X의 기댓값을 구하려면 확률밀도함수 $p(x)$의 수식을 알아야 한다.
- 그런데 우리는 데이터를 생성하는 확률변수 X의 확률밀도함수 $p(x)$의 수식을 정확히 알지 못한다.
- 하지만 표본평균이라는 새로운 확률변수 $\bar{X}$의 기댓값 $\mathrm{E}[\bar{X}]$은 원래 확률변수 X의 기댓값 $\mathrm{E}[X]$과 같으므로 표본평균 $\bar{x}$는 원래 확률변수 X의 기댓값 $\mathrm{E}[X]$과 비슷한 값이 나올 것이다. 하지만 정확한 값은 아니다.
- 만약 표본 개수 N이 크면 표본평균 $\bar{x}$의 분산이 아주 작아지므로 표본평균의 값 $\bar{x}$은 항상 표본평균의 기댓값 $\mathrm{E}[\bar{X}] = \mathrm{E}[X]$ 근처의 거의 일정한 값이 나올 것이다.
- 따라서 **표본 개수 N가 크면 표본평균 $\bar{x}$은 원래 확률변수 X의 기댓값 $\mathrm{E}[X]$의 근삿값**이라고 할 수 있다.

연습 문제 7.3.3

(1) 넘파이를 사용하여 숫자 100개를 무작위로 생성하여 표본집합을 구한다. 이 표본집합을 확률변수 X_1의 표본이라고 하자. X_1의 표본분산을 계산한다.

(2) 같은 작업을 50번 반복하여 확률변수 $X_2, X_3, \ldots, X_{50}$의 표본집합을 구한다.

(3) 확률변수 X_i의 표본집합의 표본평균 $\bar{x}_i$를 각각 계산한다. 이 값들은 표본평균 확률변수 $\bar{X}$의 표본집합이다.

(4) 확률변수 $\bar{X}$의 표본분산값을 계산하고 X_1의 표본분산과의 비율을 계산한다.

표본분산의 기댓값

앞에서 표본평균의 기댓값을 구하면 이론적인 평균 즉, 기댓값과 같아진다는 것을 증명했다.

그런데 표본분산 S^2의 기대값을 구하면 이론적인 분산 σ^2과 같아지는 것이 아니라 이론적인 분산값의 $\frac{N-1}{N}$배가 된다. 즉 표본분산값이 이론적인 분산값보다 더 작아진다.

$$\mathrm{E}[S^2] = \frac{N-1}{N}\sigma^2 \tag{7.3.20}$$

증명은 다음과 같다.

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right] = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{(X_i-\mu) - (\bar{X}-\mu)\right\}^2\right] \\
&= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{(X_i-\mu)^2 - 2(X_i-\mu)(\bar{X}-\mu) + (\bar{X}-\mu)^2\right\}\right] \qquad (7 \\
&= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right] - 2\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] + \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2\right]
\end{aligned}$$

이때 첫 번째 항은

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N}N(X-\mu)^2\right] \\
&= \mathrm{E}\left[(X-\mu)^2\right] \\
&= \mathrm{Var}[X] \\
&= \sigma^2
\end{aligned} \tag{7.3.22}$$

두 번째 항은

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\left(\frac{1}{N}\sum_{j=1}^{N}X_j-\mu\right)\right] \\
&= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\left(\frac{1}{N}\sum_{j=1}^{N}(X_j-\mu)\right)\right] \\
&= \mathrm{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i-\mu)(X_j-\mu)\right]
\end{aligned} \tag{7.3.23}$$

X_i와 X_j $(i \neq j)$는 독립일 때,

$$\mathrm{E}\left[(X_i-\mu)(X_j-\mu)\right] = 0 \quad (i \neq j) \tag{7.3.24}$$

라는 성질을 이용하면

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] &= \mathrm{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}(X_i-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[\frac{1}{N}N(X-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[(X-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{Var}[X] \\
&= \frac{\sigma^2}{N}
\end{aligned} \tag{7.3.25}$$

세 번째 항은 다음과 같아진다.

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2\right] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}X_j-\mu\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}(X_j-\mu)\right)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}(X_j-\mu)(X_k-\mu)\right]
\end{aligned} \tag{7.3.26}$$

X_j와 X_k $(j \neq k)$는 독립일때,

$$\mathrm{E}\left[(X_j-\mu)(X_k-\mu)\right] = 0 \quad (j \neq k) \tag{7.3.27}$$

라는 성질을 이용하면

$$
\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2\right] &= \mathrm{E}\left[\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_j-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N^3}N\sum_{j=1}^{N}(X_j-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{N^2}\sum_{j=1}^{N}(X_j-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{E}\left[\frac{1}{N}\sum_{j=1}^{N}(X_j-\mu)^2\right] \\
&= \frac{1}{N}\mathrm{Var}[X] \\
&= \frac{\sigma^2}{N}
\end{aligned}
\tag{7.3.28}
$$

따라서 세 항의 합은 다음과 같다.

$$
\mathrm{E}[S^2] = \sigma^2 - \frac{2\sigma^2}{N} + \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2 \tag{7.3.29}
$$

그러므로 표본분산의 기대값이 정확하게 σ^2이 되려면 평균과의 거리의 제곱의 평균을 구할 때 분모가 N이 아니라 $N-1$으로 써야 한다.

$$
\begin{aligned}
\sigma^2 &= \frac{N}{N-1}\mathrm{E}[S^2] \\
&= \frac{N}{N-1}\mathrm{E}\left[\frac{1}{N}\sum(X_i-\bar{X})^2\right] \\
&= \mathrm{E}\left[\frac{1}{N-1}\sum(X_i-\bar{X})^2\right]
\end{aligned}
\tag{7.3.30}
$$

따라서 기댓값이 정확한 분산값과 일치하는 비편향 표본분산은 다음처럼 정의한다.

$$
S^2_{\text{unbiased}} \equiv \frac{1}{N-1}\sum(X_i-\bar{X})^2 \tag{7.3.31}
$$

이렇게 표본분산이 실제 분산보다 작아지는 이유는 다음과 같다.

1. 표본분산을 계산할 때 사용하는 표본평균의 값이 데이터가 많이 몰려있는 쪽으로 편향되게 나온다.
2. 이렇게 데이터가 몰려있는 위치에 있는 표본평균을 기준으로 각 데이터까지의 거리를 계산하면 원래의 기댓값으로부터의 거리보다 작게 나올 수 있다.

실제 데이터로 예를 들어 살펴보자. 기댓값 $\mu = 0$, 분산이 $\sigma^2 = 1$인 정규분포로부터 7개의 표본을 뽑는다.

In [1]:

```
np.random.seed(15)
N = 7
data = np.sort(np.random.normal(size=(N)))[::-1]
```

이 표본의 표본평균은 약 -0.46이다. 우연히 음수인 표본이 많이 나오는 바람에 원래의 기댓값 0에서 음수쪽으로

떨어진 값이 나왔다.

In [2]:

```
mean = np.mean(data)
mean
```

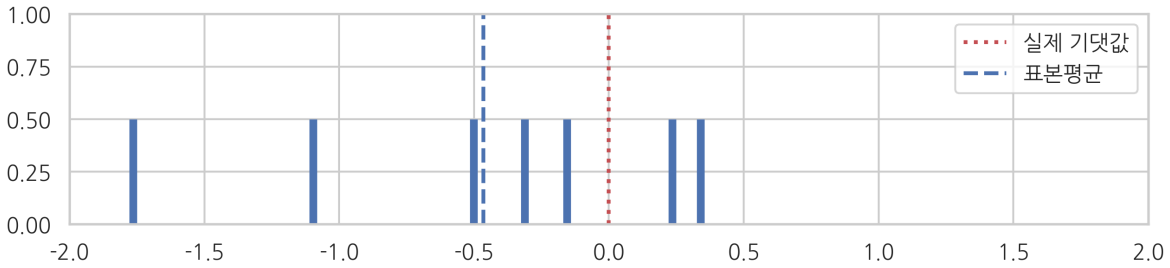
Out[2]:

-0.46494862738581794

데이터와 표본평균의 위치를 그림으로 그리면 다음과 같다.

In [3]:

```
plt.figure(figsize=(10, 2))
sns.rugplot(data, height=0.5, linewidth=4)
x = np.linspace(-3, 3, 100)
plt.axvline(x=0, ls=":", c="r", linewidth=2, label="실제 기댓값")
plt.axvline(x=mean, ls="--", c="b", linewidth=2, label="표본평균")
plt.legend()
plt.xlim(-2, 2)
plt.show()
```



표본표준편차는 표본평균으로부터 각 데이터가 떨어진 거리(의 제곱)의 평균이다.

In [4]:

```
distance_from_sample_mean = data - mean
distance_from_sample_mean
```

Out[4]:

```
array([ 0.80423333,  0.70051752,  0.30904009,  0.15262015, -0.03684105,
       -0.63091342, -1.29865663])
```

이 거리들은 진짜 평균(기댓값)으로부터 각 데이터가 떨어진 거리보다 평균적으로 작게 나온다. 그 이유는 우리가 생각한 기댓값인 표본평균이 우연히 왼쪽으로 몰려나온 데이터들 중간에 있기 때문이다.

In [5]:

```
sample_variance = (distance_from_sample_mean ** 2).mean()
sample_variance
```

Out[5]:

0.4774618257836171

따라서 표본분산값은 정확한 분산값인 1보다 작은 값이다. $N-1$로 나누어 편향 보정한 값은 다음과 같다,

In [6]:

```
sample_variance * N / (N - 1)
```

Out[6]:

0.5570387967475533

주의할 점은 표본분산의 기댓값이 원래의 분산값보다 작은 값이 나오는 경향이 있다는 것이지 항상 원래의 분산값보다 작게 나온다는 뜻은 아니다.

비대칭도와 첨도

비대칭도(skew)는 3차 모멘트 값에서 계산하고 확률밀도함수의 비대칭 정도를 가리킨다. 비대칭도가 0이면 확률분포가 대칭이다.

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \tag{7.3.32}$$

첨도(kurtosis)는 4차 모멘트 값에서 계산하며 확률이 정규분포와 대비하여 중심에 모여있는지 바깥으로 퍼져있는지를 나타낸다.

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} \tag{7.3.33}$$

모멘트

앞서 구한 기댓값이나 분산은 확률분포의 **모멘트(moment)**의 하나다.

$$\mu_n = \mathrm{E}[(X-\mu)^n] = \int (x-\mu)^n p(x) dx \tag{7.3.34}$$

모멘트는 확률분포에서 계산한 특징값이다. 만약 두 확률분포 X, Y가 있고 1차부터 무한대 차수에 이르기까지 두 확률분포의 모든 모멘트값이 서로 같다면 두 확률분포는 같은 확률분포다.

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}[Y] \\
\mathrm{E}[(X-\mu_X)^2] &= \mathrm{E}[(Y-\mu_Y)^2] \\
\mathrm{E}[(X-\mu_X)^3] &= \mathrm{E}[(Y-\mu_Y)^3] \\
\mathrm{E}[(X-\mu_X)^4] &= \mathrm{E}[(Y-\mu_Y)^4] \\
\mathrm{E}[(X-\mu_X)^5] &= \mathrm{E}[(Y-\mu_Y)^5] \\
&\vdots
\end{aligned} \tag{7.3.35}$$

이면

$$X \overset{d}{=} Y \tag{7.3.36}$$

이다. $\stackrel{d}{=}$ 는 두 확률변수가 같은 분포(distribution)를 가진다는 것을 표시하는 기호다.