Z-score

- 데이터가 평균에서 표준편차의 몇 배나 떨어져 있는지 수치화
- 데이터가 평균에서 떨어져 있는 정도를 표준편차의 배수로 표현

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \bar{x}}{s}$$

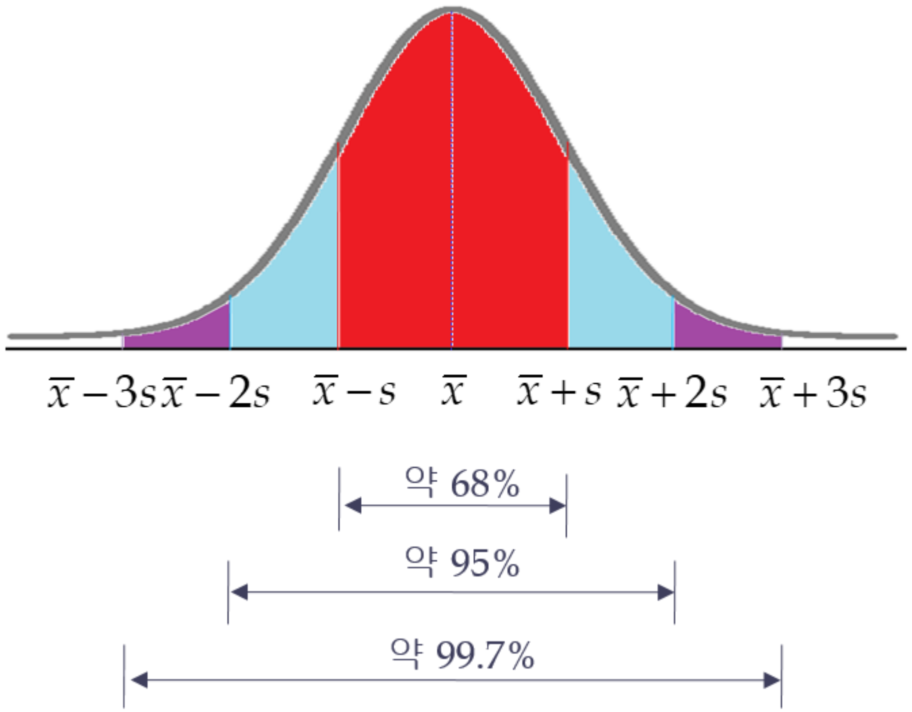
어떤 데이터에 대한 Z 점수가 2일 때를 해석하면 ?

특별한 경우

- 가운데가 불룩한 대칭 히스토그램
- 데이터가 평균에서 표준편차의 몇 배나 떨어져 있는지 수치화

수치화한 값이

- 표준편차의 1배 이내: 68.3%
- 표준편차의 2배 이내: 95.4%
  - 표준편차의 1.96배 이내: 95%
- 표준편차의 3배 이내: 99.7%



일반적으로

- 임의의 자료집단에 대해,
- 적어도 75%의 자료가 구간 $(\bar{x} - 2s, \bar{x} + 2s)$ 안에 놓인다.
- 적어도 88.9%가 구간 $(\bar{x} - 3s, \bar{x} + 3s)$ 안에 놓인다.
- 적어도 96%가 구간 $(\bar{x} - 5s, \bar{x} + 5s)$ 안에 놓인다.
- 일반적으로 $(1 - 1/k^2) \times 100\%$가 구간 $(\bar{x} - ks, \bar{x} + ks)$안에 놓인다. $(k > 1)$

이런 건 누가 알아냈을까?

Chebyshev's Inequality

체비셰프

$$P(|X-u| \leq k\sigma) \geq 1-\frac{1}{k^2}$$

어떻게 알아냈을까? 이건 나중에...

[예제 12]

자료수가 100인 아래 자료집단에 대하여 평균은 $\bar{x} = 30.138$이고 표준편차는 $s = 1.991$이다. 구간 $(\bar{x} - 2.5s,\ \bar{x} + 2.5s)$ 안에 최소한 몇 개의 자료값이 놓이는지 체비세프 정리를 이용하여 구하고, 실제 자료집단을 이용하여 그 개수를 구하라.

30.74	28.44	30.20	32.67	33.29	31.06	30.08	30.62	27.31	27.88	26.03	29.93	31.63
28.13	30.62	27.80	28.69	28.14	31.62	30.61	27.95	31.62	29.37	30.61	31.80	29.32
29.92	31.97	30.39	29.14	30.14	31.54	31.03	28.52	28.00	28.46	30.38	30.64	29.51
31.04	27.00	30.15	29.13	27.63	30.87	28.67	27.39	33.20	29.52	30.86	34.01	29.41
31.18	34.59	33.35	33.73	28.39	26.82	29.53	32.55	30.34	32.44	27.09	29.51	31.36
31.61	31.24	28.83	31.88	32.24	31.72	28.34	29.89	30.27	31.42	29.11	29.36	32.24
29.56	31.72	30.67	28.85	30.87	27.17	30.85	28.75	25.84	28.79	31.74	34.59	32.69
26.23	28.20	31.62	33.48	28.00	33.86	29.22	26.50	30.89				

$100(1 - 1/2.5^2) = 84\%$, 체비셰프에 따르면 84개, 실제로는 96개

공분산

covariance

두 자료집단에 대한 분산

- 자료 2쌍

요일	월	화	수	목	금	토	일
최저	21	22	22	23	23	24	24
최고	27	28	29	30	31	32	31

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
	y_1	y_2	y_3	y_4	y_5	y_6	y_7

계산식

x에 대한 **분산**

σ_x^2

$$\sigma_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(x_i - \mu_x)$$

x, y에 대한 **공분산 ?**

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

(1) 최저온도의 평균과 분산은 다음과 같다.

$$\bar{x} = \frac{1}{7}\sum x_i = 22.714, \quad s_x^2 = \frac{1}{6}\sum (x_i - 22.714)^2 = \frac{7.42857}{6} = 1.238$$

(2) 최고온도의 평균과 분산은 다음과 같다.

$$\bar{y} = \frac{1}{7}\sum y_i = 29.714, \quad s_y^2 = \frac{1}{6}\sum (y_i - 29.714)^2 = \frac{19.4286}{6} = 3.238$$

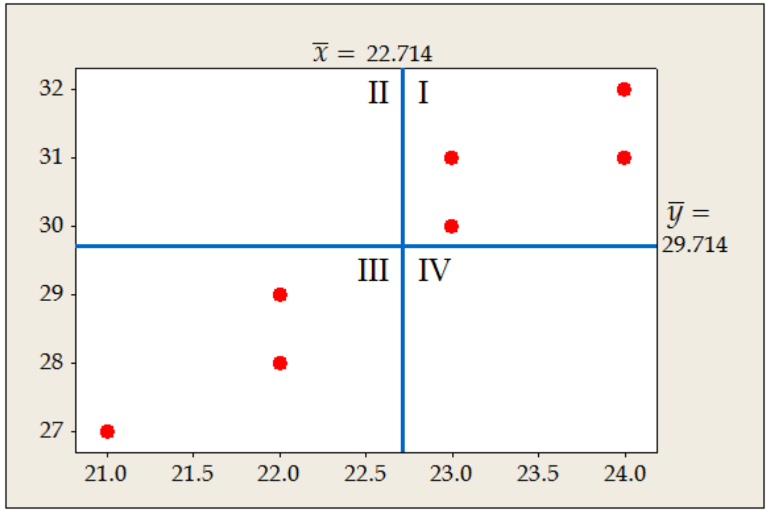
(3) 다음 표로부터 $\sum (x_i - \bar{x})(y_i - \bar{y}) = 11.4286$ 이므로 공분산은 다음과 같이다.

$$s_{xy} = \frac{11.4286}{6} = 1.9048$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
21	27	-1.714	-2.714	4.6518
22	28	-0.714	-1.714	1.2238
22	29	-0.714	-0.714	0.5098
23	30	0.286	0.286	0.0818
23	31	0.286	1.286	0.3678
24	32	1.286	2.286	2.9398
24	31	1.286	1.286	1.6538

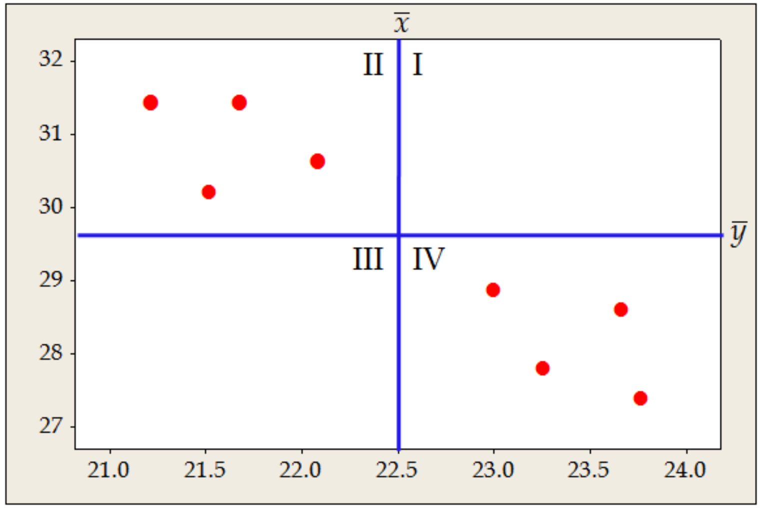
상관관계

$s_{xy} > 0$



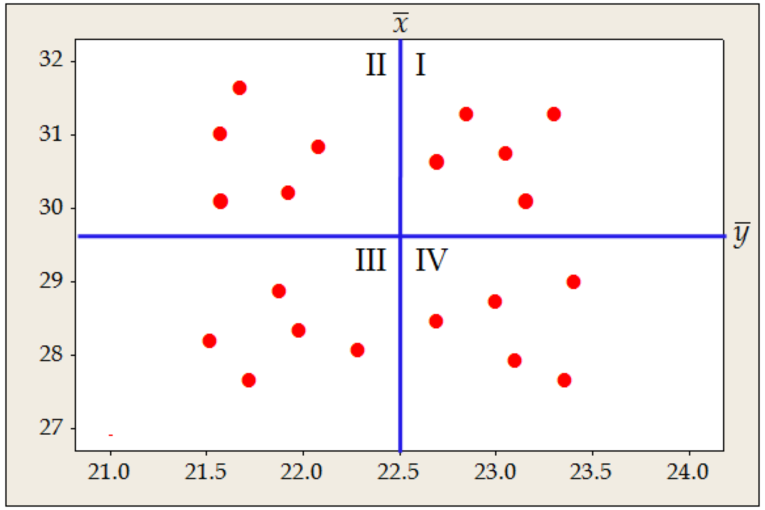
양의
상관관계

$s_{xy} < 0$



음의
상관관계

$s_{xy} = 0$



상관관계
없음

상관계수; correlation coefficient

- $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- $r_{xy} = \frac{s_{xy}}{s_x s_y}$

범위를 -1 ~ +1 사이로 만든 것

fit

1. $-1 \leq r_{xy} \leq 1$
2. $r_{xy} > 0$이면 양의 상관관계를 가지며, 양의 기울기를 갖는 적합선이 존재한다.
3. $r_{xy} < 0$이면 음의 상관관계를 가지며, 음의 기울기를 갖는 적합선이 존재한다.
4. $r_{xy} = 0$이면 두 자료집단은 무상관이다.
5. $r_{xy} = 1$이면 두 자료집단은 완전 양의 상관관계를 갖는다.
6. $r_{xy} = -1$이면 두 자료집단은 완전 음의 상관관계를 갖는다.

이 적합선 (fitting line)은 어떻게 찾을까?

나중에 배운다. 회귀

공분산 행렬; covariance matrix

자료 집단이 3개 이상 있다.

X,Y에 대한 공분산 = Y,X에 대한 공분산
X,Z에 대한 공분산 = Z,X에 대한 공분산
Y,Z에 대한 공분산 = Z,Y에 대한 공분산

	X	Y	Z
X	σ_{xx}	σ_{xy}	σ_{xz}
Y	σ_{yx}	σ_{yy}	σ_{yz}
Z	σ_{zx}	σ_{zy}	σ_{zz}

$\sigma_x^2 = \sigma_{xx}$

$\sigma_y^2 = \sigma_{yy}$

$\sigma_z^2 = \sigma_{zz}$

공분산행렬 이라고 부르자.

공분산 행렬 계산

- 자료가 2쌍 이상
- $\boldsymbol{x}_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \boldsymbol{x}_{\mathrm{n}} = \begin{pmatrix} x_{\mathrm{n}} \\ y_{\mathrm{n}} \end{pmatrix}$
- $\overline{\boldsymbol{x}} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$
- $\boldsymbol{\acute{x}}_1 = \boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \overline{\boldsymbol{\acute{x}}} = \boldsymbol{0}$
- $\boldsymbol{X} = (\boldsymbol{\acute{x}}_1 \boldsymbol{\acute{x}}_2 \dots \boldsymbol{\acute{x}}_{\mathrm{n}})$
- $\boldsymbol{X}\boldsymbol{X}^T$

코딩:
앞 예제에서는 2X2 행렬이 나온다.
공분산 값 하나하나 계산한 결과와 행렬 곱으로 계산한 결과 비교

```
>>>import numpy as np

>>>x=np.array([21,22,22,23,23,24,24]
>>>y=np.array([27,28,29,30,31,32,31]
>>>x_bar = np.mean(x)
>>>nx = x – x_bar
>>>a = np.mean(nx)
>>>'{:0.2f}'.format(a)
>>>f'{a:0.2f}'
>>>np.matmul(xy, xy.transpose())/6
array([[1.23809524, 1.9047619 ],
       [1.9047619 , 3.23809524]])
```

```
import pandas as pd

x=[21, 22, 22, 23, 23, 24, 24]
y=[27, 28, 29, 30, 31, 32, 31]

X = pd.Series(x)
Y = pd.Series(y)

X.cov(Y)
1.9047619047619044
```

```
0     21
1     22
2     22
3     23
4     23
5     24
6     24
dtype: int64
```

```
xy = np.stack([x,y])

array([[21, 22, 22, 23, 23, 24, 24],
       [27, 28, 29, 30, 31, 32, 31]])

xy.shape
(2,7)


df = pd.DataFrame(xy.transpose(), columns=['X', 'Y’])

 X   Y
0  21  27
1  22  28
2  22  29
3  23  30
4  23  31
5  24  32
6   24  31

df.cov()
```

	X	Y
X	1.238095	1.904762
Y	1.904762	3.238095

https://wikidocs.net/book/7188

참고자료

- 공학인증을 위한 확률과 통계, 이재원, 이욱기, 북스힐