제1장 기초통계이론

1.1 기본 용어

1.1.1 모집단과 표본

1. 모집단(population)
① 얻고자 하는 정보와 관련된 모든 개체로부터 얻을 수 있는 모든 관측값들의 집합
② 우리가 얻고자 하는 정보를 가지고 있는 하나의 연구대상

2. 표본(sample)
① 모집단의 일부분
② 원하는 정보를 얻기 위해 수행한 관측과정을 통하여 실제로 얻어진 관측결과의 집합

3. 표본추출(sampling): 대표성 있는 표본

1.1.2 변수의 분류

1. 변수(variable): 분석에 이용하기 위하여 관측되는 모집단 원소의 특성들

2. 변수의 분류
(1) 질적[범주형] 자료(qualitative[categorical] data)
① 명목척도자료(nominal scaling data)
: 단지 구분하기 위한 부호로 표시된 자료 [예] 김씨: 1, 이씨: 2, 박씨: 3
② 서수척도자료(ordinal scaling data)
: 자료들 사이의 크기를 비교하여 내림차순 또는 오름차순으로 숫자(순위)를 부여한 것으로 실제값보다는 순서를 나타낸 자료
[예 A^+: 4.5, A: 4.0, B^+: 3.5, B: 3.0, C^+: 2.5, C: 2.0, D^+: 1.5, D: 1.0, F: 0.0
(2) 양적 자료(quantitative[measurement] data)
① 구간척도자료 (Interval scaling data)
: 자료들 사이의 크기가 의미를 가지는 자료.
자료가 나타내는 숫자 자체만 보지 말고 숫자가 나타내는 의미를 보아야 한다.
[예] 성적, 온도

② 비율척도자료(Ratio scaling data)

: 절대 0점이 있어서 비율로 이야기할 수 있는 자료

[예] 몸무게 - A는 B보다 몸무게가 두 배가 된다.

1.2 자료의 기초적 분석

1.2.1 표와 그림을 이용한 자료분석

1. 변수의 분포(distribution): 변수가 가질 수 있는 값들이 어떠한 형태를 하고 있는가

2. 분포는 식이나 그림으로 나타낼 수 있다.

3. 표나 그림을 이용하여 분포를 추정하는 방법

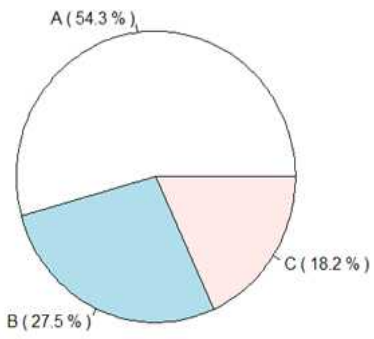
■도수분포표(frequency table) ■막대그래프(bar graph) ■히스토그램(histogram)
■원형 차트(pie chart) ■점도표(dot diagram)
■줄기-잎 그림(stem-and-leaf plot) ■상자그림(box plot) ■산점도(scatter plot)

(1) 도수분포표

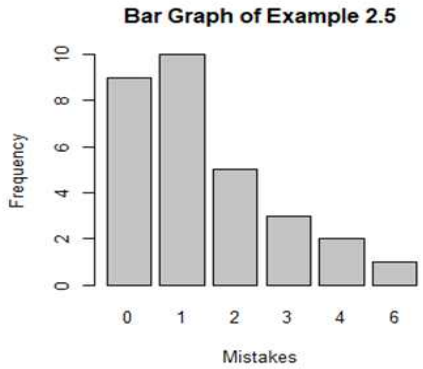
[예제] 어느 대학교의 학생회장 선거에서 세 명의 학생(A, B, C)이 입후보하여 투표를 실시한 결과 후보 A는 1,520표를, 후보 B는 770표를, 후보 C는 510표를 얻었다. 결과를 도수분포표로 나타내어라.

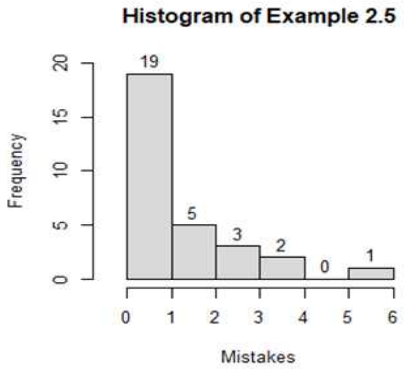
후보자	도수	상대도수
A	1520	1520 / 2800 = 0.543
B	770	770 / 2800 = 0.275
C	510	510 / 2800 = 0.182
계	2800	1

(2) 원형 차트



(3) 막대그래프, 히스토그램





(4) 점도표

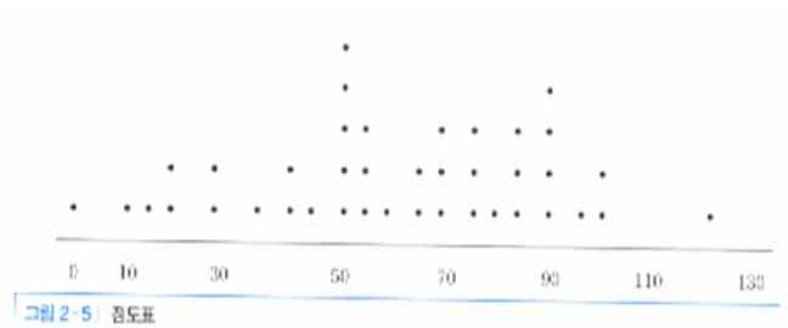


그림 2-5 점도표

(5) 줄기-잎 그림

[예제] 다음의 자료는 어느 대학 통계학과 신입생 51명의 키를 센티미터 단위로 기록한 것이다. 이 자료에 대한 줄기와 잎 그림을 그려라.

181	161	170	160	158	169	162	179	183	178	171	177	163
158	160	160	158	174	160	163	167	165	163	173	178	170
167	177	176	170	152	158	160	160	159	180	169	162	178
173	173	171	171	170	160	167	168	166	164	174	180	

15 |	15 | 888289	15 | 288889
16 |	16 | 1092300037537009207864	16 | 0000000122333456777899
17 |	17 | 0981743807608331104	17 | 0000111333446778889
18 |	18 | 1300	18 | 0013
1단계	2단계	3단계

그림 2-12 통계학과 신입생의 키에 대한 줄기-잎 그림

(6) 상자그림

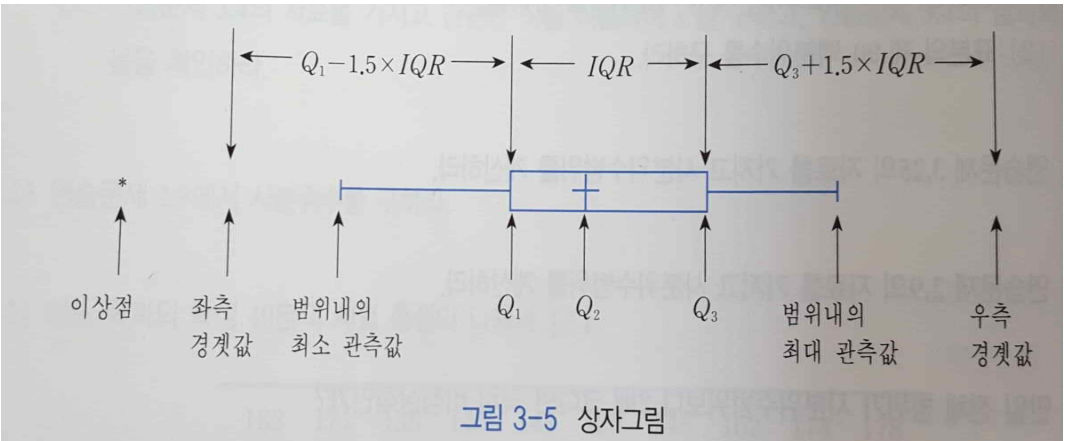
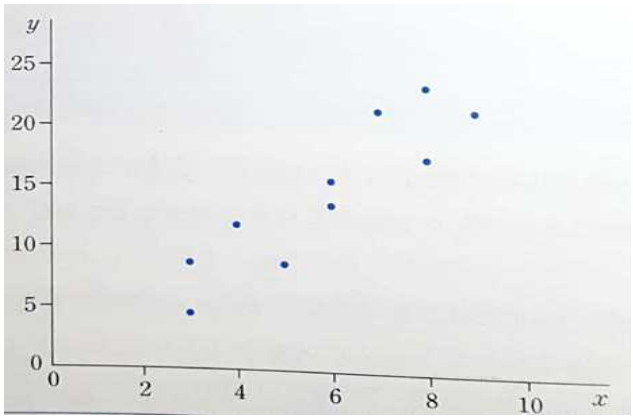


그림 3-5 상자그림

(7) 산점도



1.2.2 기본적 척도를 이용한 자료분석

1. 자료구조의 중심을 나타내는 척도

(1) 평균 (mean[average])

① 정의: $x_1, x_2, \cdots, x_n$의 평균. $\bar{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$

② 단점: 모든 관측값이 반영되므로 표본평균은 극단적으로 아주 큰 값이나 아주 작은 값에 영향을 많이 받는다.

(2) 중앙값 (median)

① 정의: 자료들을 크기 순으로 정렬하였을 때 순서에 따라 가장 가운데 있는 값

② 중앙값 구하는 방법

□ 자료의 개수(n)가 홀수: $\frac{n+1}{2}$번째 관측값

□ 자료의 개수(n)가 짝수: $\frac{n}{2}$번째 관측값과 $\frac{n}{2}+1$번째 관측값 사이의 중간값 또는 평균

③ 장점: 관측값들의 변화에 민감하지 않고 특히 아주 큰 관측값이나 아주 작은 관측값에 영향을 받지 않는다.

[예제] 크기 순으로 정의된 다음의 표본에서 표본평균과 표본중앙값을 구하라. 그리고 끝값인 15가 조사단위를 잘못하여 150으로 바뀌면 어떻게 되겠는가?

1, 3, 4, 6, 6, 7, 8, 8, 9, 10, 15

【풀이】

□ 평균: $\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{77}{11}=7$

□ 중앙값: $n=11$(홀수), 중앙값은 $\frac{n+1}{2}=\frac{11+1}{2}=6$ 번째 관측값$=7$

→ 이 경우 평균과 중앙값은 같은 값인 7을 갖는다.

만약 끝값이 15에서 150으로 변경되면

□ 평균: $\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{212}{11}=19.2727$

□ 중앙값: $n=11$(홀수), 중앙값은 $\frac{n+1}{2}=\frac{11+1}{2}=6$ 번째 관측값$=7$

→ 이 경우 평균은 이상치(150)의 영향을 많아 7에서 19.2727로 변경되지만, 중앙값은 상기 경우와 같은 7을 갖는다.

(3) 절사평균 (trimmed mean)

① $(100\times\alpha)\%$ 절사평균의 정의: 자료를 크기 순으로 나열하고 자료의 아래쪽 $(100\times\alpha)\%$와 위쪽 $(100\times\alpha)\%$를 버린 나머지 자료들의 평균

② $(100\times\alpha)\%$ 절사평균 구하는 방법: $\overline{x}_{\alpha}=\frac{x_{([na]+1)}+\cdots+x_{(n-[na])}}{n-2[na]}$

③ 장점

▫ 아주 큰 관측값이나 아주 작은 관측값에 영향을 받지 않는다.

▫ 평균과 중앙값의 성질을 모두 갖고 있다.

[예제] 10만 가구가 살고 있는 어떤 마을의 1년 소득을 조사한 자료(단위: 만원)가 다음과 같을 때, 15% 절사평균을 구하라.

950 1,050 10,310 760 1,470 1,530 1,170 1,240 1,090 1,020

【풀이】

재정렬: 760 950 1,020 1,050 1,090 1,170 1,240 1,470 1,530 10,310

$15\% \rightarrow \alpha = 0.15,\ n = 10$

$$\bar{x}_{0.15} = \frac{x_{([na]+1)} + \cdots + x_{(n-[na])}}{n - 2[na]}$$

$$= \frac{x_{([10\times 0.15]+1)} + \cdots + x_{(10-[10\times 0.15])}}{10 - 2[10\times 0.15]} = \frac{x_{(1+1)} + \cdots + x_{(10-1)}}{10-(2\times 1)} = \frac{x_{(2)} + \cdots + x_{(9)}}{8}$$

$$= \frac{9520}{8} = 1190$$

2. 자료구조의 퍼짐을 나타내는 척도

(1) 편차 (deviation)

① 정의: 표본평균을 중심위치 측도로 사용할 때 각 관측값과 평균의 차이

② 공식: $x_i - \bar{x} \qquad (i = 1, \cdots, n)$

③ 성질: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ → '퍼진 정도의 측도'로 부적합

(2) 분산 (variance)

① 공식: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ 여기서 $n-1$은 자유도(degree of freedom)

② 간편공식: $s^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$

여기서

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

(3) 표준편차 (standard deviation): $s = \sqrt{s^2}$

[예제] 다음 자료에 대한 표본분산과 표본표준편차를 구하라.

【풀이】

	자료											합계
X	1	3	4	6	6	7	8	8	9	10	15	77
X^2	1	9	16	36	36	49	64	64	81	100	225	681

$$\text{표본분산} = s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]$$

$$= \frac{1}{10}\left[681 - \frac{77^2}{11}\right] = 14.2$$

$$\text{표본표준편차} = s = \sqrt{s^2} = \sqrt{14.2} = 3.7683$$

(4) 표본범위 (sample range)

① 정의: 관측값에서 가장 큰 값과 가장 작은 값의 차이

② 공식: 범위$= \max - \min$

③ 장점: 간편하게 구할 수 있고 해석이 용이하다.

④ 단점: 양 끝점에 의해서만 결정되기 때문에 중간에 위치한 관측값들이 어떻게 퍼져 있는가 하는 것은 전혀 고려되지 않는다. 특히 극단적으로 큰 값이나 작은 값이 있는 경우 그 관측값이 미치는 영향이 매우 클 수 있다.

(5) 백분위수 (percentile)

① 정의: 전체를 백 부분으로 나누어 각 경계선에 해당하는 값

② 예: 제25백분위수$= Q_1$, 제50백분위수$= Q_2$(중앙값), 제75백분위수$= Q_3$

[예제] 서울의 한 전철역에서 인천의 한 전철역까지 소요되는 시간을 기록한 자료가 다음과 같다(단위: 분). 제 50 백분위수인 중앙값과 제 20 백분위수를 구하라.

42 40 38 37 43 39 78 38 45 44 40 38 41 35 31 44

【풀이】

상기 관측값을 순서대로 재배열하면

31 35 37 38 38 38 39 40 40 41 42 43 44 44 45 78

$n = 16$ (짝수)

① $p = 0.5 \rightarrow np = 16 \times 0.5 = 8 \rightarrow$ 제50백분위수$= \dfrac{x_{(8)} + x_{(9)}}{2} = \dfrac{40 + 40}{2} = 40$

② $p = 0.2 \rightarrow np = 16 \times 0.2 = 3.2$ (정수 아님) $\rightarrow m = 3 + 1 = 4$

$\rightarrow$ 제20백분위수$= x_{(m)} = x_{(4)} = 38$

(6) 사분위수 (quartile)

① 정의: 전체를 네 부분으로 나누어 각 경계선에 해당하는 값

② 구성: 제25백분위수$= Q_1$, 제50백분위수$= Q_2$(중앙값), 제75백분위수$= Q_3$

(7) 표본사분위수범위 (sample interquartile range; IQR) $= Q_3 - Q_1$

① 정의: 제3사분위수와 제1사분위수 사이의 거리

② 공식: $Q_3 - Q_1$

③ 장점: 극단값에 영향을 받지 않고, 한쪽으로 치우친 분포에서 극단값을 제외한 퍼진 정도를 알려고 할 사용된다.

④ 단점: 사분위수범위에 대한 이론적 추론이 어렵기 때문에 분산이나 표준편차만큼 퍼진 정도의 측도로 많이 쓰이지는 않는다.

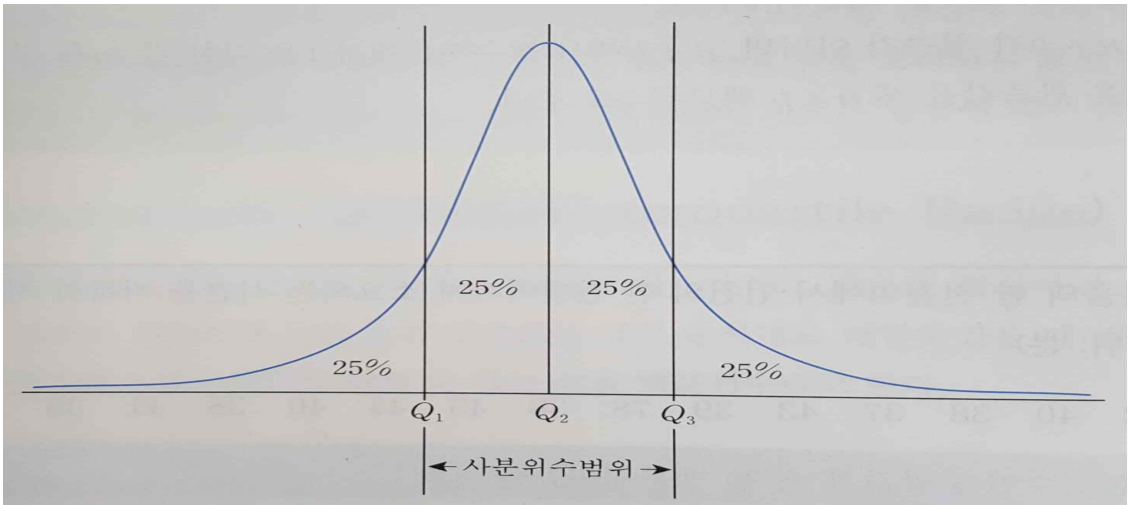


그림 3-4 사분위수의 위치와 사분위수범위

[예제] 아래 자료에 대한 표본사분위수범위를 구하라.

【풀이】

$$1,\ 3,\ 4,\ 6,\ 6,\ 7,\ 8,\ 8,\ 9,\ 10,\ 15$$

$$Q_1 = 4,\ Q_3 = 9$$

$$\text{표본사분위수범위} = Q_3 - Q_1 = 9 - 4 = 5$$

(8) 다섯숫자 요약 (5-number summary)

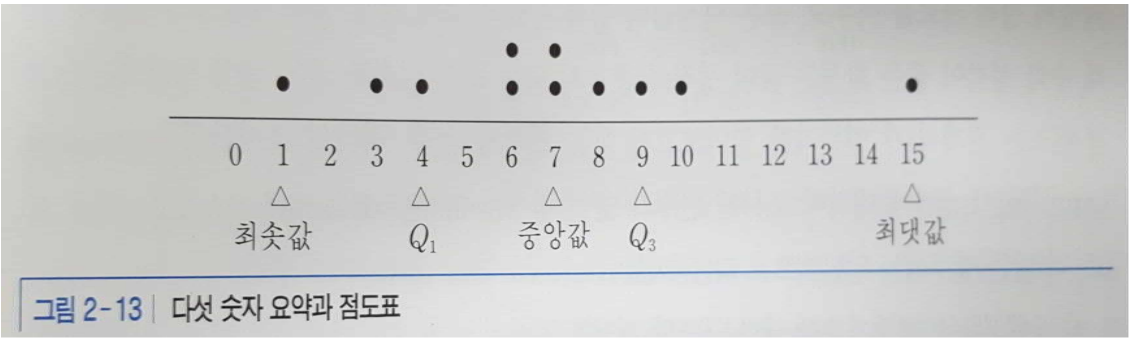


그림 2-13 | 다섯 숫자 요약과 점도표

1.3 확률변수와 확률표본

1.3.1 확률변수

1. 정의: 확률변수 X란 실험의 결과들에 수치를 대응시키는 것이다.

$[X = a] = \{\omega \in \Omega | X(\omega) = a\}$ for any real number a

[예제] 동전을 세 번 던지는 실험에서 앞면이 나오는 횟수를 X라고 하자. X가 가지는 값들과 이에 대응하는 결과들을 나열하라.

X: 근원사상	X값
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

X가 가질 수 있는 각각의 값에 따라서 대응되는 사상

X값	각 X값에 대응하는 사상
$[X=0]$ =	$\{TTT\}$
$[X=1]$ =	$\{HTT,\ THT,\ TTH\}$
$[X=2]$ =	$\{HHT,\ HTH,\ THH\}$
$[X=3]$ =	$\{HHH\}$

2. 성질

① 서로 다른 X값들에 대응하는 사상들은 배반적이다.

② 이런 사상들의 합사상은 전체 표본공간이 된다.

1.3.2 확률분포의 기댓값과 표준편차

1. 평균 (mean): $E(X)=\mu=\sum(\text{값}\times\text{확률})=\sum x_i f(x_i)$

[예제] 공정한 동전을 세 번 던져서 앞면이 나오는 횟수를 X라고 할 때, X의 평균을 구하라.

【풀이】

x	$f(x)$	$xf(x)$
0	1/8	0
1	3/8	3/8
2	3/8	6/8
3	1/8	3/8
합계	1	$E(X)=\mu=\sum x_i f(x_i)=1.5$

2. 분산 (variance): $\sigma^2=Var(X)=\sum(x_i-\mu)^2 f(x_i)=\sum x_i^2 f(x_i)-\mu^2$

3. 표준편차 (standard deviation): $\sigma=sd(X)=\sqrt{Var(X)}$

[예제] 위 예제의 확률분포를 하에서 σ^2의 간편식을 이용하여 분산과 표준편차를 구하라.

【풀이】

x	$f(x)$	$xf(x)$	$x^2f(x)$
0	0.1	0.0	0.0
1	0.2	0.2	0.2
2	0.4	0.8	1.6
3	0.2	0.6	1.8
4	0.1	0.4	1.6
합계	1	2.0	5.2

$\sigma^2 = Var(X) = \sum (x_i - \mu)^2 f(x_i) = \sum x_i^2 f(x_i) - \mu^2 = 5.2 - (2.0)^2 = 1.2$

$\sigma = sd(X) = \sqrt{Var(X)} = \sqrt{1.2} = 1.095$

1.3.3 상관계수

1. 상관계수에 대한 정의

n개의 자료쌍 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$에 대하여

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\ \sqrt{S_{yy}}} \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

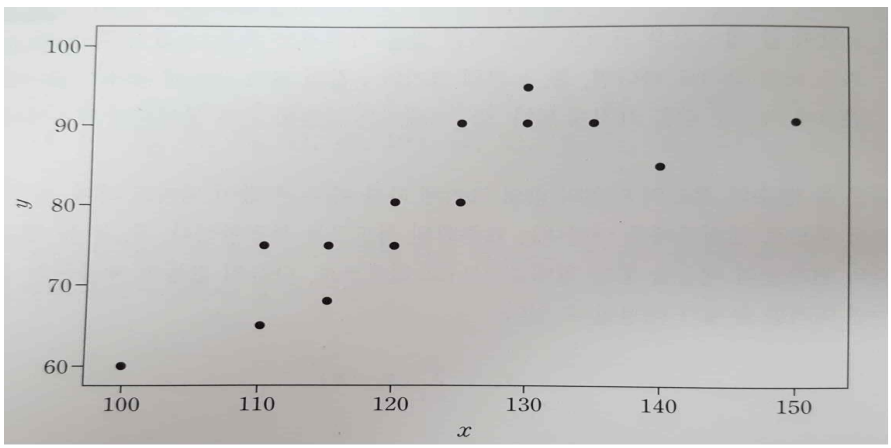
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

[예제] 중학교 1학년 학생들에 대하여 IQ와 성적의 관계를 알아보기 위해 15명의 학생을 임의추출하여 다음의 자료를 얻었다. 산점도를 그리고 상관계수를 구하라.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IQ	110	130	125	120	115	120	125	130	150	140	100	110	115	120	135
성적	75	90	80	80	70	75	90	95	90	85	60	65	75	75	90

【풀이】



학생	$X(IQ)$	Y(성적)	X^2	Y^2	XY
1	110	75	12,100	5,625	8,250
2	130	90	16,900	8,100	11,700
3	125	80	15,625	6,400	10,000
4	120	80	14,400	6,400	9,600
5	115	70	13,225	4,900	8,050
6	120	75	14,400	5,625	9,000
7	125	90	15,625	8,100	11,250
8	130	95	16,900	9,025	12,350
9	150	90	22,500	8,100	13,500
10	140	85	19,600	7,225	11,900
11	100	60	10,000	3,600	6,000
12	110	65	12,100	4,225	7,150
13	115	75	13,225	5,625	8,625
14	120	75	14,400	5,625	9,000
15	135	90	18,225	8,100	12,150
계	1,845	1,195	229,225	96,675	148,525

$$S_{xx}=\sum_{i=1}^{15}x_i^2-15\overline{x}^2=229,225-(1,845\times1,845)/15=2,290.00$$

$$S_{yy}=\sum_{i=1}^{15}y_i^2-15\overline{y}^2=96,675-(1,195\times1,195)/15=1,473.34$$

$$S_{xy}=\sum_{i=1}^{15}x_iy_i-15\overline{x}\,\overline{y}=148,525-(1,845\times1,195)/15=1,540.00$$

이므로 상관계수는 $r=\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}=0.8384$가 된다. ■

2. 상관계수의 성질

(1) r은 변수의 종류나 특정단위에 관계없는 측도로 -1과 +1 사이의 값을 가지며, r의 값이 +1에 가까울수록 강의 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 나타내며, r의 값이 0에 가까울수록 상관관계는 약해진다.

(2) X와 Y의 대응되는 모든 값들이 한 직선 상에 위치하면 r의 값은 -1(직선의 기울기가 음인 경우)이나 +1(직선의 기울기가 양인 경우)의 값을 가진다.

(3) 상관계수 r은 단지 두 변수 간의 선형관계만을 나타내는 측도이다. 그러므로 $r=0$인 경우에 두 변수의 선형상관관계는 없지만 다른 관계는 가질 수 있다.