

경영 통계학

[illegible]

학습 목표

1. 중심, 변동성, 형태 등의 개념을 이해하고 설명한다
2. 중심, 변동성에 대한 일반적 측정수단을 계산하고 해석한다
3. 체비셰프 정리를 적용한다
4. 경험 법칙을 적용하여 특이값을 파악하고, 데이터 세트를 표준화 값으로 전환한다
5. 사분위수와 백분위수, 상자그림, 상관계수와 공분산, 그룹 데이터의 평균과 표준 데이터를 계산할 수 있다.
6. 표본의 왜도와 첨도를 평가한다.

숫자적 기술


- 이 장에서는 데이터의 숫자적 기술에 대해 설명한다. 데이터의 시각적 표현이 많은 물음에 답을 주긴 하지만, 더 자세한 숫자적 기술이 필요하기도 하다.
- 다음의 세가지 데이터의 특성에 관심을 갖는다.

특성	해석
중심	데이터 값들의 중심은 어디인가? 어떤 데이터 값이 전형적인, 또는 중간 값인가? 중심경향이 있는가?
변동성 형태	데이터의 퍼짐이 어느 정도인가? 데이터 값들이 얼마나 퍼져 있는가? 특이한 값들이 있는가? 데이터 값들의 분포가 대칭적인가? 비대칭인가? 중심이 아주 뾰족한가? 평평한가? 봉우리가 두 개인가?

숫자적 기술

- J.D. Power and Associates는 자동차 품질 등급을 매년 발표한다. 이 평가는 소비자, 딜러, 메이커 모두에게 중요한 정보이다. 표 4.2는 33개 자동차 브랜드의 2010년형 기준 결함률을 나타낸 것이다. 나와 있는 결함률은 표본에 기준하였다. 우리는 이와 같은 데이터를 숫자형 통계량을 이용하여 어떻게 요약하는지 공부한다.

표 4.2

자동차 100대당 결함  JDPower

브랜드	결함	브랜드	결함	브랜드	결함
Acura	86	Hyundai	102	MINI	133
Audi	111	Infiniti	107	Mitsubishi	146
BMW	113	Jaguar	130	Nissan	111
Buick	114	Jeep	129	Porsche	83
Cadillac	111	Kia	126	Ram	110
Chevrolet	111	Land Rover	170	Scion	114
Chrysler	122	Lexus	88	Subaru	121
Dodge	130	Lincoln	106	Suzuki	122
Ford	93	Mazda	114	Toyota	117
GMC	126	Mercedes-Benz	87	Volkswagen	135
Honda	95	Mercury	113	Volvo	109

출처: J.D. Power and Associates 2010 Initial Quality Study™. 평가는 교육 목적으로만 제시된 것이며 소비자 의사결정의 지표로 사용되어서는 안 됨.

숫자 데이터의 예

- 측정척도가 연속형(continuous)이라는 것에 주의해야 한다(예: 3대의 차량에서 4개의 결함이 발견될 경우 결함률은 1.333333 혹은 100대당 133이 된다).
- 결함률은 매년 바뀌기도 하고, 아마 같은 연도 내에서도 달라지기 때문에, 조사 시점이 언제냐에 따라 결과가 달라질 수도 있다.
- 데이터로 가장 먼저 해봐야 하는 것이 정렬(sorting)이다.

브랜드	결함	브랜드	결함	브랜드	결함
Porsche	83	Audi	111	Chrysler	122
Acura	86	Cadillac	111	Suzuki	122
Mercedes-Benz	87	Chevrolet	111	GMC	126
Lexus	88	Nissan	111	Kia	126
Ford	93	BMW	113	Jeep	129
Honda	95	Mercury	113	Dodge	130
Hyundai	102	Buick	114	Jaguar	130
Lincoln	106	Mazda	114	MINI	133
Infiniti	107	Scion	114	Volkswagen	135
Volvo	109	Toyota	117	Mitsubishi	146
Ram	110	Subaru	121	Land Rover	170

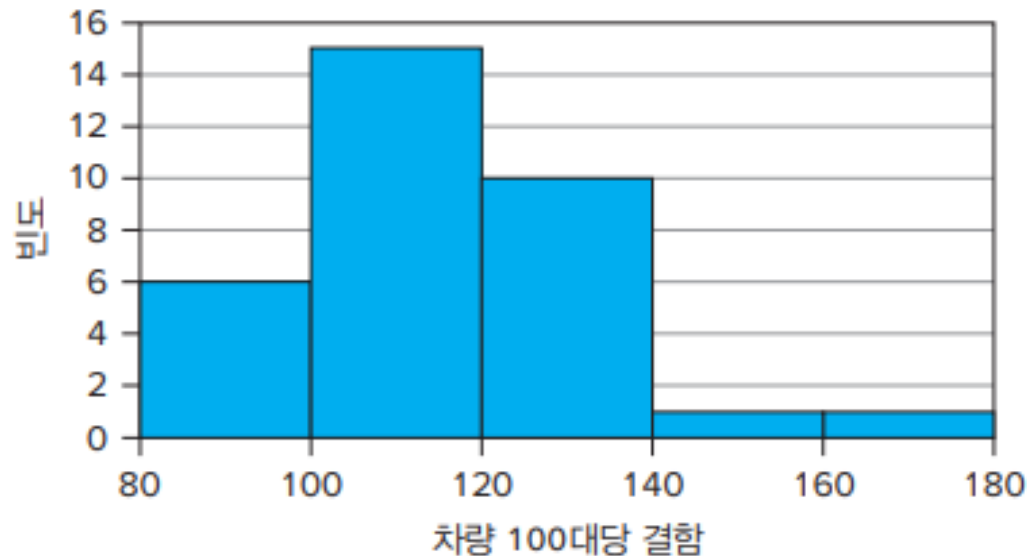
숫자 데이터의 예

- 앞의 표에 있는 정렬된 데이터는 중심과 변동성에 대해 시사점을 제공한다. 데이터 값들의 범위가 83(Porsche)에서 173(Land Rover)까지이며,中间的 값은 110에서 120 사이인 것으로 보인다.
- 아래 점그림은 특이값의 존재를 보여준다.



숫자 데이터의 예

- 또 다른 시각적 표현은 히스토그램으로서 그림 4.3에 나와 있다. 110과 120 사이의 최빈계급(modal class)이 중심을 보여준다. 히스토그램의 형태는 오른쪽으로 기운(대부분의 데이터가 왼쪽에 있고 오른쪽 꼬리가 긴) 형태이다.



중심측정

표 4.4

중심에 대한 6개 측도

통계량	공식	엑셀 함수*	장점	단점
평균	$\frac{1}{n} \sum_{i=1}^n x_i$	=AVERAGE(Data)	익숙하며 표본의 모든 정보를 이용	극단값의 영향을 받음
중앙값	정렬된 열에서 중간값	=MEDIAN(Data)	극단값의 영향을 받지 않음	극단값을 고려치 않으며, 값들간 간격이 영향을 줌
최빈값	가장 발생빈도가 높은 값	=MODE.SNGL(Data)	속성 데이터나 값이 다양하지 않은 이산형 데이터에 유용	복수의 최빈값이 가능, 연속형 데이터에 부적절
범위의 중앙	$\frac{x_{\min} + x_{\max}}{2}$	=0.5*(MIN(Data) + MAX(Data))	이해 및 계산이 쉬움	대부분의 값을 무시, 극단값의 영향을 받음
기하평균(G)	$\sqrt[n]{x_1 x_2 \dots x_n}$	=GEOMEAN(Data)	성장률의 평균에 유용, 큰 극단값의 영향을 완화	익숙하지 않으며 양(+)의 데이터만 가능
절사평균	상하위 각 k%를 제거한 평균값	=TRIMMEAN(Data, Percent)	극단값의 영향을 완화	의미 있는 데이터가 배제될 수 있음

*R offers equivalent functions mean(x), median(x), and trimmed mean mean(x, trim=p%) but not the others on this list. See **Appendix J** for a side-by-side list of R and Excel functions.

평균(mean)

- 평균은 가장 익숙한 통계적 중심측정치이다.
- 데이터 값을 모두 합하고 데이터의 갯수로 나눈다.
- 모집단 평균은 μ 로 표기하고 표본은 \bar{x} 로 표기한다.

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (\text{모집단 평균}) \qquad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (\text{표본평균})$$

평균(mean)의 특성

- 산술평균은 가장 익숙한 “평균”
- 평균은 표본 원소 모두에게 영향을 받음
- X축을 지렛대로 생각하고 모든 데이터를 같은 무게를 갖는 것으로 가정하면
평균은 데이터분포의 균형점, 혹은 받침점에 해당됨



중앙값(median)

- 중앙값(M으로 표기)은 표본 데이터의 50백분위수(50th percentile) 또는 중간점(midpoint)을 찾은 것
- 정렬된 관측치가 중앙값을 경계로 두 개로 나누어짐



- 중앙값은 n 이 홀수일 때는 정렬된 데이터의 중앙 관측치이나, n 이 짝수일 때는 중앙에 있는 2개 관측치의 평균값



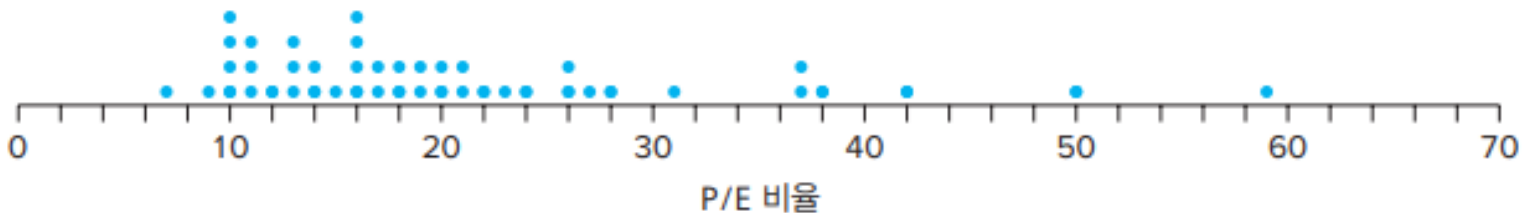
최빈값(mode)

- 가장 빈번하게 발생하는 데이터 값
- 여러 개 있거나 전혀 없는 경우도 있을 수 있음
- 범주형 혹은 이산형 변수에 가장 유용하지만, 연속형 데이터의 경우 유용하지 않은 경우가 일반적임

83	86	87	88	93	95	102	106	107	109	110
111	111	111	111	113	113	114	114	114	117	121
122	122	126	126	129	130	130	133	135	146	170

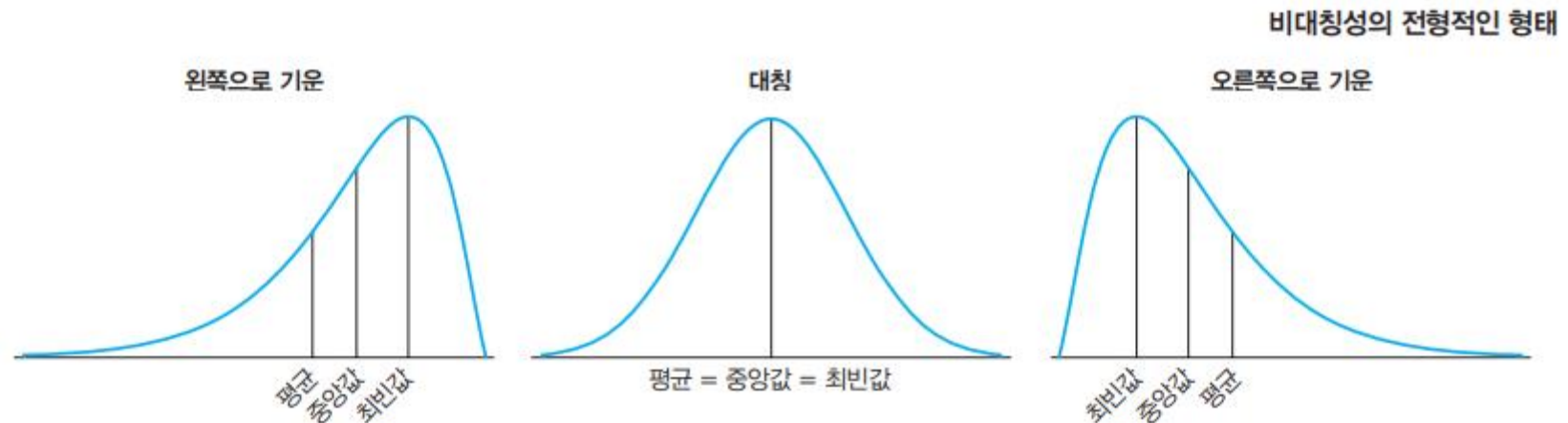
최빈값(mode)

- 아래 점그림에서 최빈값은 10과 16(각각 4번 발생)
- 그러나 다른 한편으로 보면 11과 13도 세 번 발생하는 등 최빈값이 중심에 대한 정밀한 측도가 아니라는 것을 보여주고 있음



형태(shape)

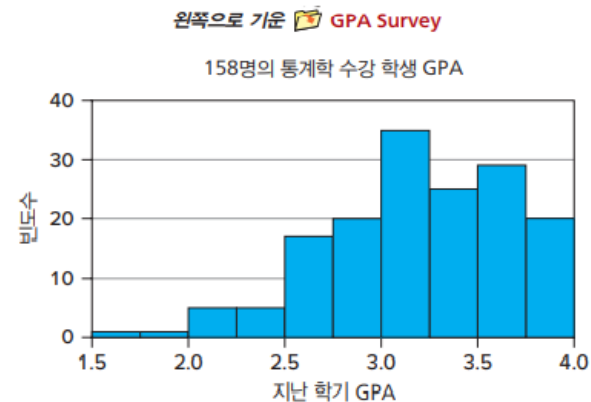
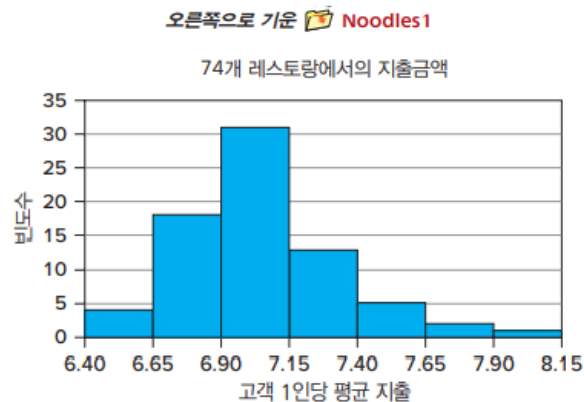
- 어떤 분포의 형태는 히스토그램을 그려보거나 또는 평균과 중앙값을 비교함으로써 판단할 수 있음. 대칭성(skewness)의 정도(이를 왜도라고도 함)가 다름



형태(shape)

분포의 형태	히스토그램 모양	통계량
왼쪽으로 기운 (음의 왜도)	히스토그램의 긴 꼬리가 왼쪽에 있음(대부분의 데이터가 오른쪽에 몰려 있으나 일부 매우 작은 값이 왼쪽에 존재)	평균 < 중앙값
대칭	히스토그램의 꼬리가 균형을 이룸(작은 값과 큰 값이 상쇄됨)	평균 \approx 중앙값
오른쪽으로 기운 (양의 왜도)	히스토그램의 긴 꼬리가 오른쪽에 있음(대부분의 데이터가 왼쪽에 몰려 있으나 일부 매우 큰 값이 오른쪽에 존재)	평균 > 중앙값

비대칭성 히스토그램 예시



기하평균(geometric mean)

- 기하평균(G)은 곱셈에 의한 평균
- 데이터 값을 모두 곱하고 n 제곱근으로 구함

$$G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- 예를 들어 $X = 2, 3, 7, 9, 10, 12$ 의 기하평균은

$$G = \sqrt[6]{(2)(3)(7)(9)(10)(12)} = \sqrt[6]{45,360} = 5,972$$

성장률

$$GR = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$$

- 기하평균을 약간 변형시켜 어떤 시계열의 평균 성장률(average growth rate)을 계산할 수 있음
- 예를 들어 2011년부터 2015년까지 JetBlue항공사의 수익이 표와 같다면
- 평균성장률은 금년과 전년도의 비율을 가지고 기하평균 처리
- 연평균성장률은 9.2%

연도	수익
2011	4,504
2012	4,982
2013	5,441
2014	5,817
2015	6,416

$$GR = \sqrt[4]{\left(\frac{4982}{4504}\right)\left(\frac{5441}{4982}\right)\left(\frac{5817}{5441}\right)\left(\frac{6416}{5817}\right)} - 1 = \sqrt[4]{\frac{6416}{4504}} - 1 = 1.09248 - 1 = 0.09248 \approx 9.2\%$$

범위의 중앙(midrange)

- 범위의 중앙은 변수 X의 최솟값과 최댓값의 중간점.
- 계산하기는 쉬우나, 극단적 값에 영향을 받기 때문에 중심경향의 좋은 척도는 아님

$$\text{범위의 중앙} = \frac{x_{\min} + x_{\max}}{2}$$

- J.D. Power데이터의 예:

$$\text{범위의 중앙} = \frac{x_1 + x_{33}}{2} = \frac{83 + 170}{2} = 126.5$$

절사평균(trimmed mean)

- 일반적인 평균과 동일하게 계산하되, 최상위 및 최하위 각각 k%를 제거한 다음 평균을 구한다는 차이가 있음
- 절사평균은 양극단의 데이터값이 주는 영향을 중화시킴
- J.D. Power 데이터($n = 33$)에서 5% 절사평균은 $0.05 \times 33 = 1.65 = 1$ (소수점을 없앤 정수)이기 때문에 양끝에서 한 개씩의 관측치를 없앤 후 31개의 관측치로 평균을 구함

중심측정

J.D. Power 데이터에 대한 엑셀의 중심 측도

평균:	=AVERAGE(Data)	= 11470
중앙값:	=MEDIAN(Data)	= 113
최빈값:	=MODE.SNGL(Data)	= 111
기하평균:	=GEOMEAN(Data)	= 11335
범위의 중앙:	=(MIN(Data)+MAX(Data))/2	= 1265
5% 절사평균:	=TRIMMEAN(Data,0.10)	= 11394

변동성 측정

- 데이터의 변동성(Variation)은 데이터 값들이 중심에서 퍼져 있는 정도를 의미
- 경제학 강의를 듣는 학생들의 공부 시간에 관한 다음 자료를 보자



- 각 그림은 같은 평균을 가지고 있지만 평균주변으로 퍼진 정도는 다름

변동성 측정

표본의 변동성에 대한 5개 측도

통계량	공식	엑셀 함수*	장점	단점
범위(R)	$X_{\max} - X_{\min}$	=MAX(Data)- MIN(Data)	계산이 용이하고 해석이 쉬움	극단값에 민감
표본분산(s^2)	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$	=VAR.S(Data)	수리통계학에서 핵심적 역할	직관적 의미전달이 어려움
표본표준편차(s)	$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$	=STDEV.S(Data)	가장 보편적 측도, 원데이터와 단위가 동일(\$, £, ¥, grams 등).	직관적 의미전달이 어려움
변동계수(CV)	$100 \times \frac{s}{\bar{x}}$	없음	퍼센트로 측정한 상대적 변동 측도, 데이터간 비교 가능	데이터가 음수가 아니어야 함
평균절대편차(MAD)	$\frac{\sum_{i=1}^n x_i - \bar{x} }{n}$	=AVEDEV(Data)	이해하기 쉬움	이론적 성질이 '좋은' 편이 아님

범위(range)

- 범위는 최대값과 최소값의 차이

$$\text{범위} = x_{\max} - x_{\min}$$

- J.D. Power자료에서 범위는

$$\text{범위} = 170 - 83 = 87$$

- 범위는 계산하기 쉽지만, 양극단의 두 값만을 가지고 계산한다는 것이 단점

분산(variance)

- 모분산(population variance, σ^2 으로 표기. 여기서 σ 는 그리스문자 시그마의 소문자)은 평균으로부터의 편차 제곱을 합하여 모집단 원소 수로 나눈 것으로 정의함
- 표본이라면 (대부분의 경우에 해당되지만) μ 를 \bar{x} 로 대체하여 표본분산(sample variance, s^2 으로 표기)을 구함

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

모분산

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

표본분산

표준편차(standard deviation)

- 표준편차(standard deviation)은 분산에 제곱근을 취한 것으로 정의
- 표준편차의 측정단위는 변수의 측정단위와 같음

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

모표준편차

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

표본표준편차

- 변동성의 기술에 있어서 표준편차가 가장 많이 사용
- 표준편차는 데이터세트에서 각 데이터값이 평균으로부터 얼마나 변동하는지를 보여주는 단일 수치

표준편차(standard deviation)

위험을 비교하는 데 사용되는 표준편차

“일반적으로 표준편차는 투자 수익이 평균에서 사적으로 얼마나 벗어났는지를 나타낸다. 통계적 측도는 보통 최근 3년간의 월별 수익을 사용하여 계산된다. [...] 표준편차는 다양한 뮤추얼 펀드의 투자 위험을 비교할 때 유용하다. 펀드의 평균 수익률은 비슷하지만 표준편차가 다르면 두 펀드 중 표준편차가 높은 펀드가 변동성이 더 크다.”

출처: T. Rowe Price Investor, September 2009, p. 8.

변동계수(coefficient of variation)

- 측정단위가 서로 다를 때(예: 킬로그램과 온스), 또는 평균이 서로 다를 때 데이터의 퍼진 정도를 비교하기 위한 것이 변동계수(CV)
- 단위와 무관한 측도

$$CV = \frac{s}{\bar{x}} \times 100\%$$

평균절대편차(mean absolute deviation)

- 퍼진 정도의 또 다른 척도
- 이 통계량은 중앙으로부터의 평균 거리를 나타냄. 다음과 같이 편차에다 절대값을 취하여 그것들의 평균을 구한 것

$$MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- MAD 는 간단하고 직관적이기 때문에 호소력이 있음. 지렛대의 예를 상기하면 MAD 는 개별 데이터 점들이 지레받침(균형점)에서 평균적으로 얼마나 떨어져 있는가는 의미함

체비셰프 정리(Chebyshev's theorem)

- 어떤 데이터든, 그리고 그 분포가 무엇이든 상관없이 평균으로부터 k 표준편차 이내 (즉, $\mu \pm k\sigma$ 이내)에 전체 관측치 중 최소한 $100[1 - 1/k^2]\%$ 가 존재함

$k = 2$; 관측치 중 최소한 75.0%가 $\mu \pm 2\sigma$ 안에 있음

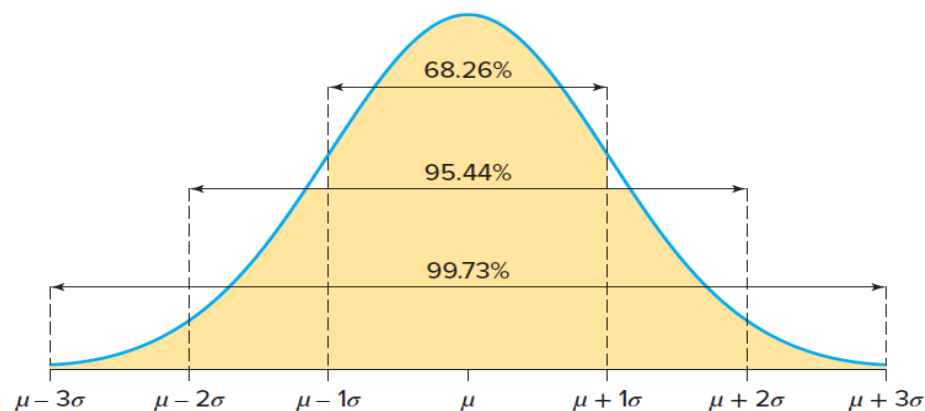
$k = 3$; 관측치 중 최소한 88.9%가 $\mu \pm 3\sigma$ 안에 있음

$k = 4$; 관측치 중 최소한 93.8%가 $\mu \pm 4\sigma$ 안에 있음

- 이 정리는 어떤 데이터에도 적용된다는 장점이 있지만 한도가 지나치게 넓음

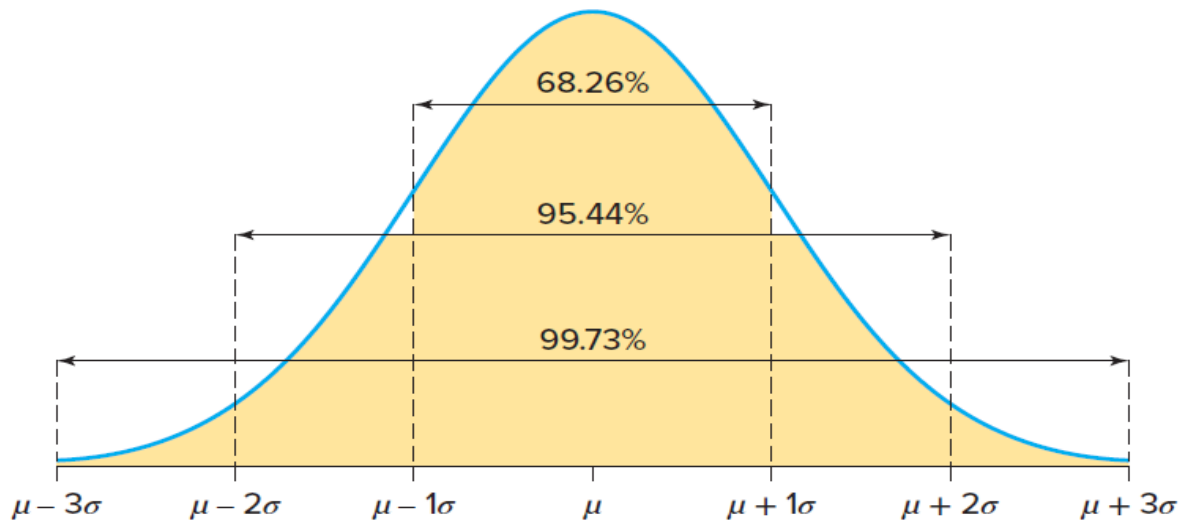
경험법칙(The Empirical Rule)

구간 $\mu \pm k\sigma$ 안에 어떤 정해진 비율의 데이터가 있다는 것임. 즉:



경험법칙(The Empirical Rule)

- 예를 들어 학생 80명이 시험을 치렀다고 하자. 어느 정도의 학생이 평균으로부터 2표준편차 이내에 들어 있을까? 시험성적이 정규분포 혹은 종 모양의 분포를 따른다고 하면 경험법칙에 의거하여 예상이 가능하다. 즉, 대략 76명($95.44\% \times 80$)의 학생이 평균에서 2표준편차 이내에 들어 있을 것이다. 정규분포는 평균을 중심으로 대칭적이기 때문에 약 2명의 학생은 평균보다 2표준편차 이상의 점수를, 또 다른 2명은 평균보다 2표준편차 이하의 점수를 받은 것으로 예상된다.



표준화 데이터

- 이상(unusual) 관측치를 판단하는 일반적인 방법은 각 관측치를 표준편차에 대한 평균으로부터의 거리로 다시 정의하여 정규화된 자료로 만드는 것
- 표준화 값(standardized value; z점수라고도 부름)은 각 관측치와 평균과의 거리가 표준편차의 몇 배인지를 나타냄

모집단

$$z_i = \frac{x_i - \mu}{\sigma}$$

z값이 음수이면 관측치가 평균의 왼쪽에 있음을 의미한다.

표본

$$z_i = \frac{x_i - \bar{x}}{s}$$

z값이 양수이면 관측치가 평균의 오른쪽에 있음을 의미한다.

이상 관측치

- 특이값(Outlier) if $|z_i| > 3$ ($\mu \pm 3\sigma$ 이상)
- 이상값(Unusual) if $|z_i| > 2$ ($\mu \pm 2\sigma$ 이상)
- .
- 대규모 표본(예를 들어 $n=1000$)에 대해서는 3 표준편차 밖의 범위에 데이터가 있어도 이상할 것이 없음
- 소규모 표본(예를 들어 $n < 30$)에 대해서는 표본의 분포를 정규분포와 비교하는 것은 위험함. 왜냐하면 분포의 형태에 대해서 많은 정보가 없기 때문임

시그마 추정

- 정규분포의 경우 거의 모든 관측치가 $\mu \pm 3\sigma$ 이내에 있기 때문에 관측치의 범위가 대략 $6\sigma(\mu - 3\sigma$ 에서 $\mu + 3\sigma$ 까지)임
- 따라서 범위 $R(x_{\max} - x_{\min})$ 을 안다면 표준편차를 $s = R/6$ 으로 추정할 수 있음
- 이 룰은 범위만 알고 있을 때 표준편차를 근사적으로 손쉽게 구할 수 있는 방법
- 물론 이 추정치는 정규분포를 따른다는 가정을 바탕으로 하고 있음

백분위수(percentile)

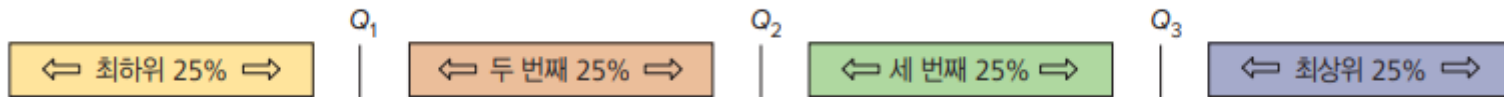
- 백분위수는 데이터를 100개 그룹으로 나눈 것
- 예를 들어 자신의 성적이 83백분위수라면 시험을 치른 사람 중 83%가 자신보다 성적이 낮은, 상위 17%에 속해 있다는 것을 의미
- 데이터를 10개 그룹(십분위수; decile),
5개 그룹(오분위수; quintile),
4개 그룹(사분위수; quartile)으로 나눌 수 있음

백분위수(percentile)

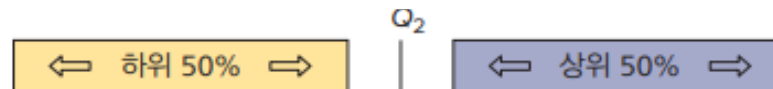
- 백분위수를 계산함에 있어서 많은 경우에 두 데이터 값의 중간을 보간(interpolation)으로 채워야만 함
예를 들어 표본의 크기가 $n = 73$ 인 상황에서 95백분위수를 구할 때,
 $0.95 \times 73 = 69.35$ 이기 때문에 95백분위수를 얻기 위해서는 69번째와 70번째 관측치(즉, x_{69} 와 x_{70}) 사이를 보간해야 함
- 인사관리에서는 직원에 대한 성과평가 혹은 보수의 벤치마킹 등에 백분위수 이용
- 나누는 그룹의 개수는 일의 성격이나 표본의 크기에 달렸지만, 사분위수의 경우 표본의 크기가 아주 작아도 계산할 수 있기 때문에 좀 더 자세히 논의할 가치가 있음

사분위수(quartiles)

- 사분위수는 정렬된 데이터를 대략 동일한 크기로 4개 그룹으로 나누었을 때의 분기점들이다. Q_1 , Q_2 , Q_3 로 표기하며 이는 곧 25, 50, 75백분위수를 의미하는 셈

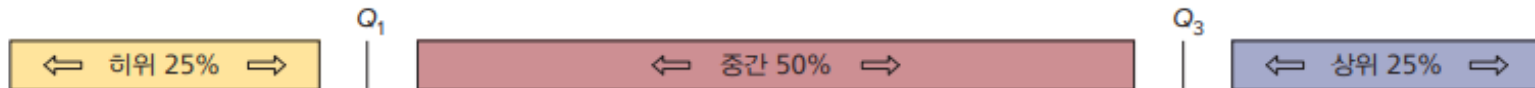


- 2분위인 Q_2 는 중앙값. 아래 위로 같은 수의 데이터가 있기 때문에 중앙값은 중요한 중심의 지표임



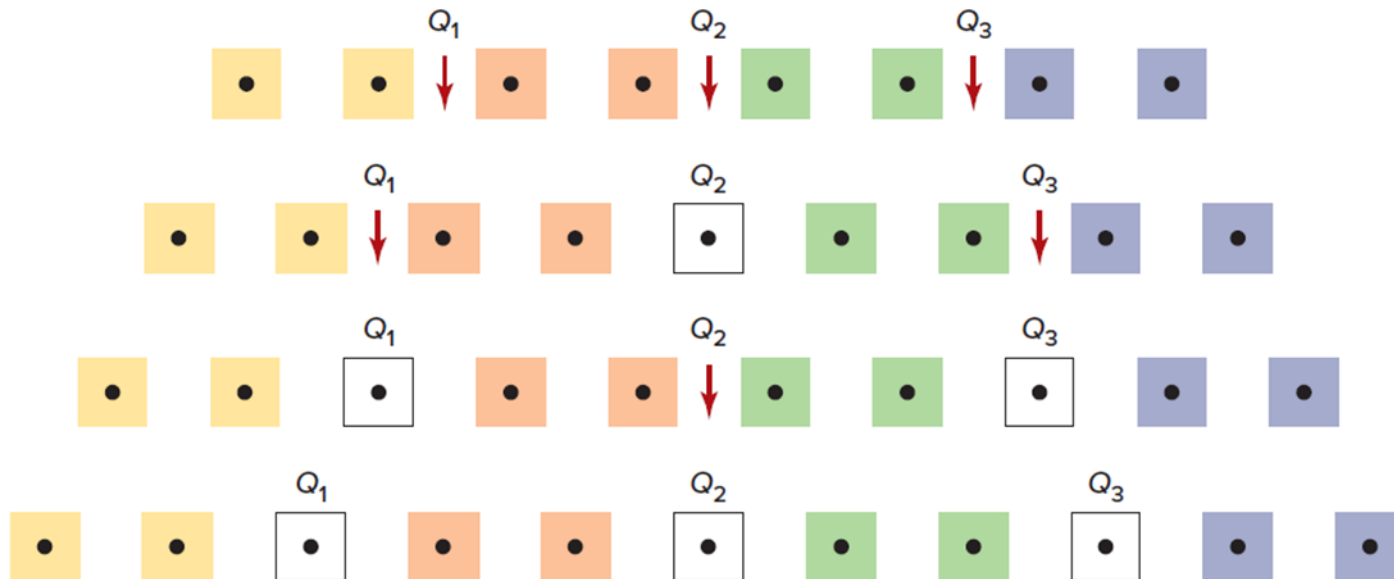
사분위수(quartiles)

- Q1과 Q3는 중간 50%의 범위를 의미하므로 중심을 나타내기도 함. Q1과 Q3는 퍼진 정도를 가리키기도 하는데, 그 이유는 사분위범위(interquartile range) 거리인 $Q3 - Q1$ (IQR)이 중간 50% 데이터의 퍼진 정도를 나타내기 때문임



사분위수(quartiles) – 중앙값 방법

- 1사분위수(Q1)는 Q2보다 작은 값의 중앙값이며, 3사분위수(Q3)는 Q2보다 큰 값의 중앙값임
- 표본의 크기 n 에 달려 있지만 사분위수 Q1, Q2, Q3가 정확히 어떤 데이터 값이 될 수도 있고, 아니면 어떤 2개의 데이터 값 사이에 놓일 수도 있음



사분위수(quartiles) – 중앙값 방법

- 데이터세트의 크기가 작은 경우 중앙값 방법을 이용하여 사분위수를 구할 수 있음
- Step 1: 관측치 정렬
- Step 2: 중앙값 Q2를 구함
- Step 3: Q2보다 작은 데이터만을 이용하여 중앙값을 구함
- Step 4: Q2보다 큰 데이터만을 이용하여 중앙값을 구함

사분위수(quartiles) – 중앙값 방법

- 어떤 재무분석가가 12종목의 에너지장비 관련 주식을 가지고 있음. 이들의 최근 주가수익(P/E) 비율 데이터를 이용하여 사분위수를 구해보자. 우선 데이터를 오름차순으로 정렬한 다음, 한 가운데 두 데이터 값의 중간으로 Q2(중앙값)을 구한 다음에 Q1과 Q3(각각 위쪽 절반과 아래쪽 절반의 중앙값)을 구함

그림 4.20

중앙값 방법

회사

정렬된 P/E

Maverick Tube
BJ Services
FMC Technologies
Nabors Industries
Baker Hughes
Varco International
National-Oilwell
Smith International
Cooper Cameron
Schlumberger
Halliburton
Transocean

7
22
25
29
31
35
36
36
39
42
46
49



Q_1 는 x_3 과 x_4 의 중간이기 때문에
 $Q_1 = (x_3 + x_4)/2 = (25 + 29)/2 = 27.0$



Q_2 는 x_6 과 x_7 의 중간이기 때문에
 $Q_2 = (x_6 + x_7)/2 = (35 + 36)/2 = 35.5$



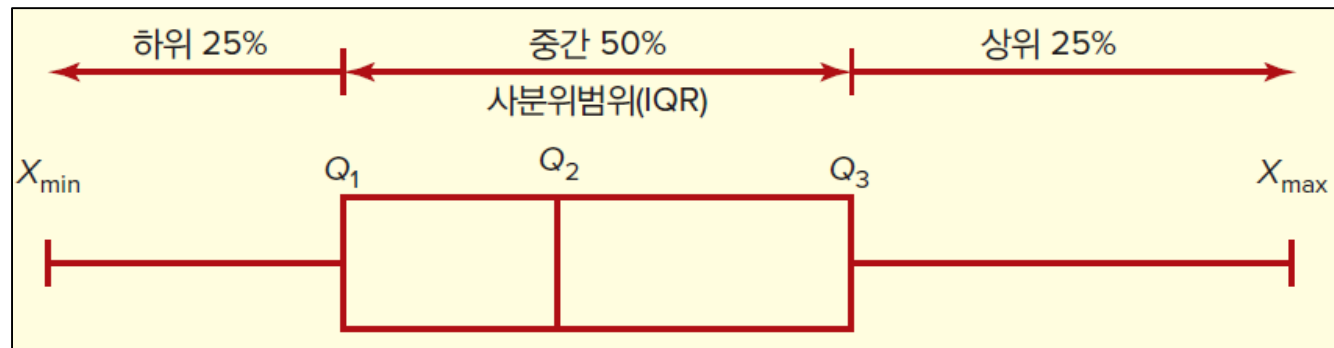
Q_3 는 x_9 과 x_{10} 의 중간이기 때문에
 $Q_3 = (x_9 + x_{10})/2 = (39 + 42)/2 = 40.5$

상자그림(box plot)

- 상자그림(box plot) 또는 상자수염그림(box-and-whisker plot)은 탐색적 데이터분석(EDA: exploratory data analysis)에서 유용한 분석도구이다. 이는 다음의 다섯숫자요약(five-number summary)에 의해 만들어진다.:

$$x_{\min}, Q_1, Q_2, Q_3, x_{\max}$$

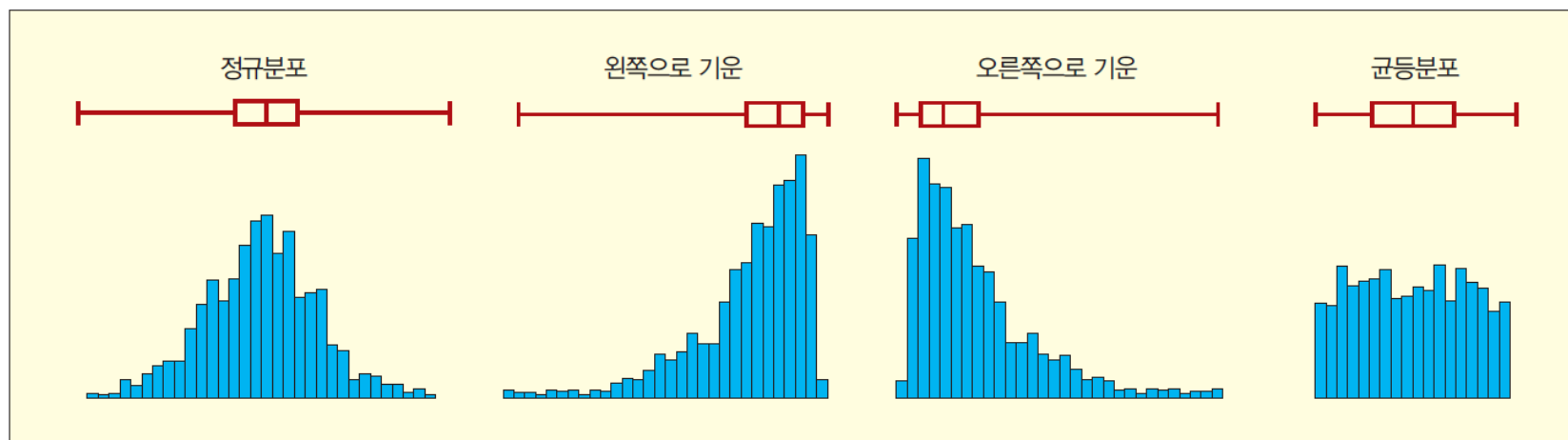
- 상자그림은 다음과 같이 그려진다.



상자그림(box plot)

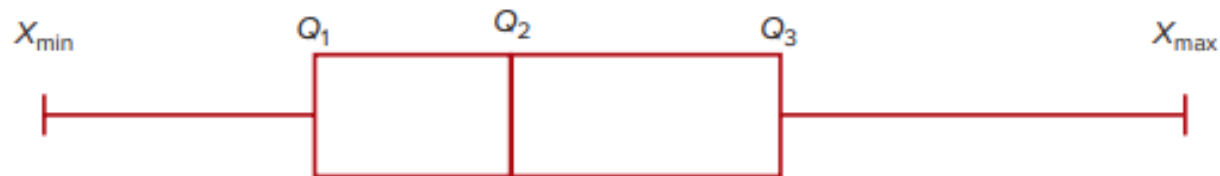
- 상자그림은 중심과 변동성을 나타냄

네 개의 모집단($n = 1,000$)에서 추출한 표본에 대한 상자그림



상자그림(box plot)

- 상자그림 아래쪽에는 변수의 값을 보여주는 눈금표시가 있음. 상자의 바깥쪽 세로줄은 Q_1 , Q_3 임. 상자 안의 세로줄은 중앙값(Q_2)이다. 상자 양쪽의 긴 선은 “수염”으로 오른쪽 수염이 길면 오른쪽으로 기운 분포임. 이는 중앙값이 왼쪽으로 치우친 것으로도 확인할 수 있음



상자그림(box plot)

- 사분위수를 이용하여 이상값을 판별할 수도 있음. 이를 판단하는 울타리(fence)는 다음과 같이 사분위범위(IQR)인 $Q_3 - Q_1$ 을 기초로 함

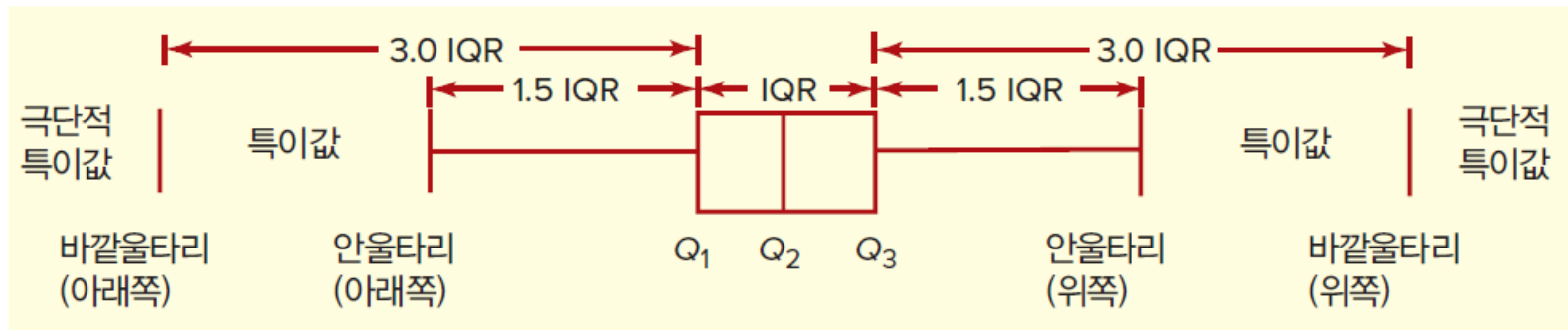
	안울타리	바깥울타리
아래쪽 울타리	$Q_1 - 1.5(Q_3 - Q_1)$	$Q_1 - 3.0(Q_3 - Q_1)$
위쪽 울타리	$Q_3 + 1.5(Q_3 - Q_1)$	$Q_3 + 3.0(Q_3 - Q_1)$

- MegaStat 등의 통계패키지에서는 상자그림에서 안울타리(inner fence) 밖에 있는 관측치를 특이값(outlier), 그리고 바깥울타리(outer fence) 밖에 있는 관측치를 극단적 특이값(extream outlier)라고 함

	안울타리	바깥울타리
아래쪽 울타리	$107 - 1.5(126 - 107) = 78.5$	$107 - 3.0(126 - 107) = 50$
위쪽 울타리	$126 + 1.5(126 - 107) = 154.5$	$126 + 3.0(126 - 107) = 183$

상자그림(box plot): 울타리와 이상값

- 상자그림에서 안울타리 밖에 있는 관측치를 특이값(outlier), 바깥울타리 밖에 있는 관측치를 극단적 특이값(extreme outlier)으로 정의함



상자그림(box plot): 중간영역중심(midhinge)

- 사분위수를 이용하면 특이값의 영향을 받지 않는 또 다른 중심 측도를 만들 수 있음
- 중간영역중심은 1사분위수와 3사분위수의 평균

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2}$$

- 중간영역중심은 항상 Q1과 Q3의 한 가운데 있는데 반해 중앙값 Q2는 상자의 어디든 위치할 수 있음. 이를 이용해 왜도(기울어짐)을 판단할 수 있는 또 다른 방법이 있음

중앙값 < 중간영역중심	⇒ 오른쪽으로 기운(오른쪽 꼬리가 김)
중앙값 \approx 중간영역중심	⇒ 대칭(양쪽 꼬리가 거의 동일)
중앙값 > 중간영역중심	⇒ 왼쪽으로 기운(왼쪽 꼬리가 김)

공분산(covariance)

- 두 확률변수 X 와 Y 의 공분산은 $\text{Cov}(X, Y)$ 또는 σ_{XY} 으로 표기. 공분산은 X 와 Y 가 같은 방향으로 변하는 정도를 측정함

모집단 공분산

$$\sigma_{XY} = \frac{\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{N}$$

표본 공분산

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

- 공분산은 측정 단위가 애매함. 왜냐하면 X 와 Y 의 측정 단위가 서로 다를 수 있기 때문임. 이 때문에 일반적으로 공분산 값을 표준화시켜 항상 -1 에서 $+1$ 사이의 값을 갖도록 만든 상관계수를 이용함

상관계수(correlation coefficient)

- X 와 Y 의 공분산을 각 표준편차(모집단은 σ_X 와 σ_Y 으로, 표본은 s_X 와 s_Y 으로 표기)의 곱으로 나눈 것
- 모집단은 그리스 문자 ρ (rho)로 적고, 표본에 대해서는 r 로 표기

$$\text{모집단 상관계수} \quad \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$\text{표본 상관계수} \quad r = \frac{s_{XY}}{s_X s_Y}$$

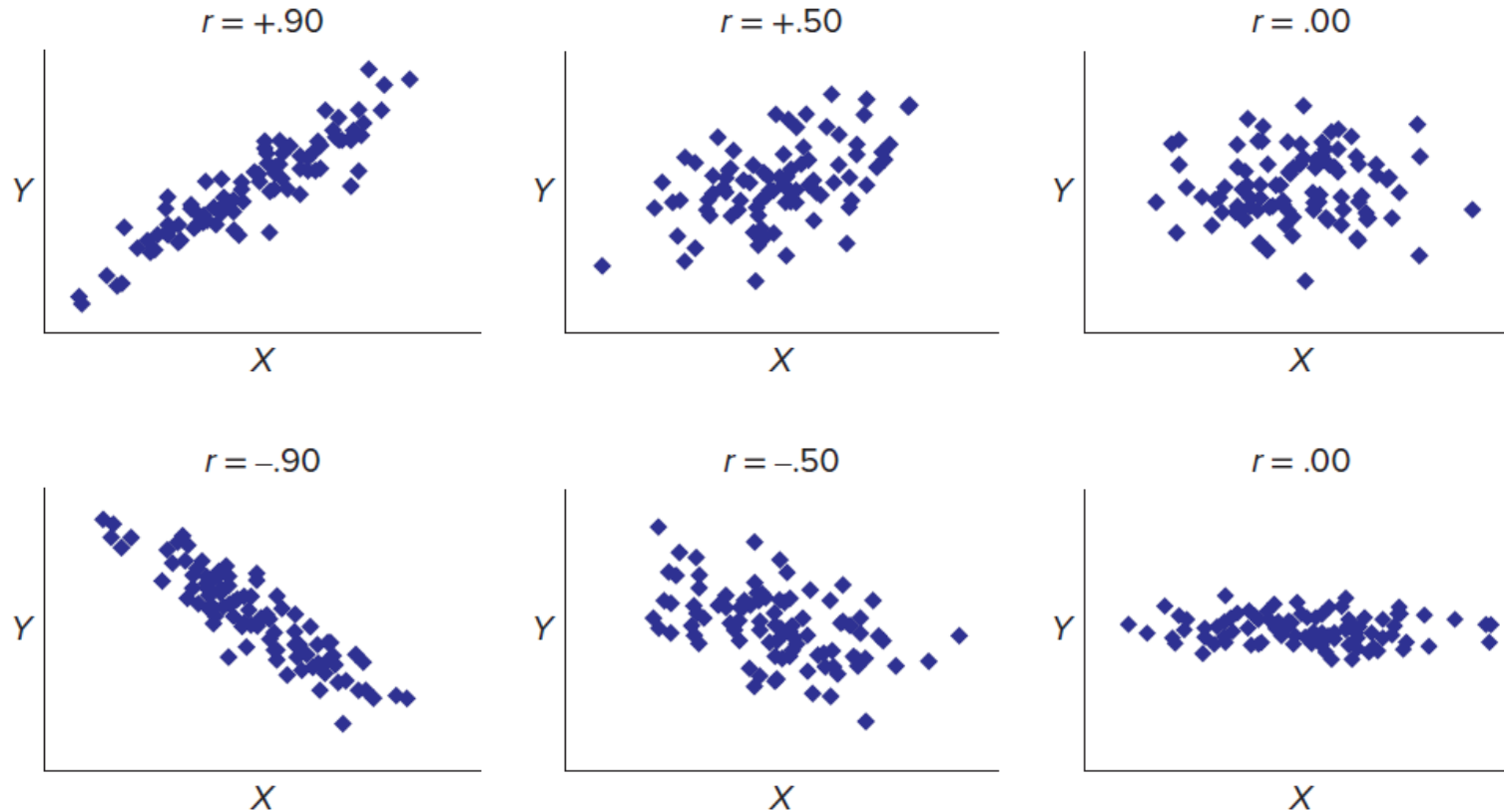
표본상관계수(sample correlation coefficient)

- X, Y 두 개의 정량 변수가 쌍으로 있을 때 둘 사이의 선형 관련성을 측정하는 통계량

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- r의 범위: $-1 \leq r \leq +1$
- r이 0에 가까울수록: X,Y 사이에는 선형관계가 없음
- r이 +1에 가까울수록: 강한 양의 관계
- r이 -1에 가까울수록: 강한 음의 관계

상관계수(correlation coefficient)



가중평균(weighted mean)

- 가중평균은 각 데이터 값에 가중치 w_j 를 부여한 다음 합한 것으로, 각 가중치는 전체에서 차지하는 비중을 나타냄(따라서 모든 가중치의 합은 1이어야 함)

$$\bar{x} = \sum_{j=1}^k w_j x_j \quad \text{Where} \quad \sum_{j=1}^k w_j = 1.00$$

- 예를 들어 수강생들의 평가기준이 과제 30%, 중간고사 20%, 기말고사 40%, 프로젝트 10% (가중치 합이 1)라고 하자. 각각의 성적이 85, 68, 78, 90이라고 하면, 성적의 가중평균은

$$\bar{x} = \sum_{j=1}^k w_j x_j = .30 \times 85 + .20 \times 68 + .40 \times 78 + .10 \times 90 = 79.3$$

그룹 평균(group mean)

- 가중평균의 개념을 이용해 그룹으로 이루어진 관측치를 처리할 수 있음
- 각 간격 j 마다 중간점 m_j 와 빈도수 f_j 가 있음. 평균을 추정하기 위해서는 각 계급의 중간점에다 빈도수를 곱하여 그것을 모든 k 개 계급에 대해 합친 다음 표본의 크기 n 으로 나눔

$$\bar{x} = \frac{\sum_{j=1}^k f_j m_j}{n}$$

그룹 표준편차(group standard deviation)

- 표준편차 추정에서는 평균으로부터의 편차 제곱의 합을 구하기 위해 각 계급의 중간점에서 평균 추정치를 빼고 그것의 제곱을 취한 다음 빈도수를 곱하여 그것을 모든 계급에 대해 합친 것. 그것을 $n - 1$ 로 나눈 다음 제곱근을 취함
- 각 중간점에서 평균 추정치를 뺄 때 “소수점을 없애는” 실수를 범해서는 안됨

$$s = \sqrt{\sum_{j=1}^k \frac{f_j (m_j - \bar{x})^2}{n - 1}}$$

그룹 데이터

콜레스테롤약 그룹 데이터 워크시트($n = 47$) 📁 RxPrice

부터	까지	f_j	m_j	$f_j m_j$	$m_j - \bar{x}$	$(m_j - \bar{x})^2$	$f(m_j - \bar{x})^2$
60	65	6	62.5	375.0	-10.42553	108.69172	652.15029
65	70	11	67.5	742.5	-5.42553	29.43640	323.80036
70	75	11	72.5	797.5	-0.42553	0.18108	1.99185
75	80	13	77.5	1,007.5	4.57447	20.92576	272.03486
80	85	5	82.5	412.5	9.57447	91.67044	458.35220
85	90	0	87.5	0.0	14.57447	212.41512	0.00000
90	95	1	92.5	92.5	19.57447	383.15980	383.15980
	합	47	합	3,427.5		합	2,091.48936
			평균(\bar{x})	72.9255		표준편차(s)	6.74293

$$\bar{x} = \frac{\sum_{j=1}^k f_j m_j}{m} = \frac{3,427.5}{47} = 72.9255$$

$$s = \sqrt{\sum_{j=1}^k \frac{f_j (m_j - \bar{x})^2}{n-1}} = \sqrt{\frac{2,091.48936}{47-1}} = 6.74293$$

왜도(skewness)

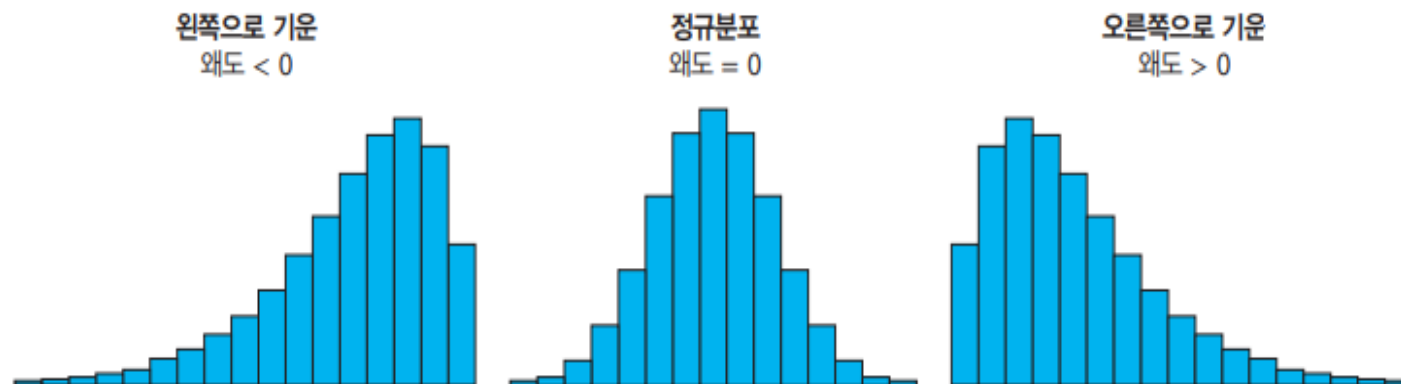
- 일반적으로 왜도(skewness; 그림 4.30)는 표본 히스토그램을 검토하거나 평균과 중앙값을 비교함으로써 판단할 수 있음
- 그러나 이러한 방법은 부정확하고 표본의 크기를 고려하지 않는다는 문제점이 있음
- 더 높은 정확도가 필요할 때는 표본의 왜도계수(skewness coefficient)로 판단

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

왜도(skewness)

그림 4.30

왜도의 전형적인 형태



왜도(skewness)

- 엑셀에서 왜도계수를 구하기 위해서는 데이터>데이터 분석>기술통계법을 이용하거나 혹은 함수 =SKEW(Data)를 사용하면 된다.
- 왜도를 측정하는 또 다른 방법은 피어슨 2 왜도계수(Pearson2 skewness coefficient; Sk_2 로 표기)

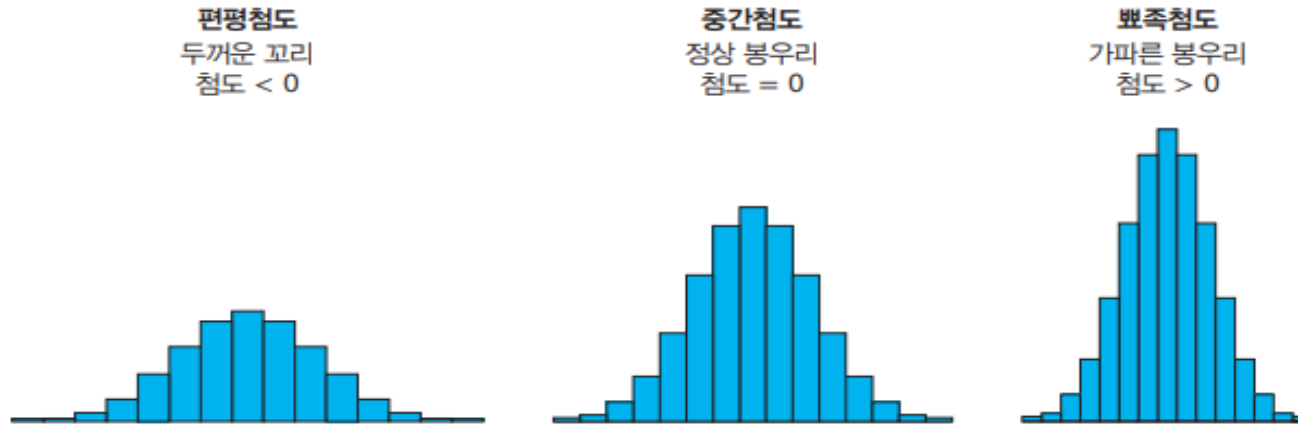
$$Sk_2 = \frac{3(\bar{x} - m)}{s}$$

- 평균과 중앙값의 차이를 표준편차 기준으로 표현

첨도(kurtosis)

- 분포에서 꼬리의 두터움 또는 (동일한 의미이지만) 중심 부분의 뾰족함의 정도
- 종모양의 정규분포를 기준으로 하며 이를 중간첨도(mesokurtic)라 함
- 정규분포보다 중앙부분이 더 편평한(즉 꼬리가 더 가는) 것을 편평첨도(platykurtic)라고 하고, 정규분포보다 중앙부분이 더 뾰족한(즉 꼬리가 더 두터운) 것을 뾰족첨도(leptokurtic)라 함
- 첨도와 변동성이 종종 혼동되기도 하지만 둘은 서로 같지 않음

첨도(kurtosis)



- 히스토그램으로 첨도를 파악하는 데는 한계가 있음. 척도나 축의 비율 등이 다르기 때문이다. 따라서 다음과 같이 숫자로 된 통계량이 필요함

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$