통계학입문
Introduction to Statistics

통계학입문
Introduction to Statistics

김동일
홍익대학교

Philosophy & Art

통계학입문

저자 | 김동일
발행인 | 이미애
발행처 | Philosophy & Art

출판등록 | 2008년 1월 8일 제152호
주소 | 대전시 유성구 도룡동 380-39
전화 | 070-7893-4471
홈페이지 | http://philosophy-art.com

값 20,000원

ISBN 978-89-961425-1-5 93310
2008년 8월 20일 1판 1쇄 발행

차 례

표 차례

그림 차례

1 통계학과 통계프로그램

사람들은 같은 사물을 제각기 다른 모습으로 인식한다.
통계학은 사물의 인식에 따르는 불확실성을 평가하고, 그를 통해 사물의 참 모습을 과학적으로 추론한다.

1.1 통계학이란 무엇인가?

정의 1.1. *(*모집단, 표본, 표집*)* 관심의 대상이 되는 전체를 모집단*(population)*이라고 하며, 모집단에서 관측된 부분을 표본*(sample),* 모집단으로부터 표본을 뽑는 것을 표집 또는 표본추출*(sampling)*이라고 한다.

정의 1.2. *(*통계, 통계적 추론, 통계학*)* 표본의 데이터에 산술적 연산을 적용한 결과를 통계*(statistic)*라고 하며, 표본의 데이터에서 통계를 만들고, 그 통계를 분석하여 표본의 모집단에 대해 통계적 추론*(statistical inference)*을 하는 학문을 통계학*(statistics)*이라고 한다.

만약 모집단 전부를 관측할 수 있다면, 즉 표본이 바로 모집단이라면, 표본의 데이터를 정리하여 통계를 만드는 것으로 통계학의 임무는 끝난다. 그러나 일반적으로 모집단 전체를 관측하는 것은 매우 비효율적이거나 또는 아예 불가능하여, 모집단의 극히 작은 일부인 표본만을 관측하는 경우가 대부분이다. 관측된 표본의 데이터로부터 관측되지 않은 전체 모집단의 특성에 대해 통계적 추론을 하는 경우 언제나 불확실성이 따르게 되는데 이 불확실성을 과학적으로 분석하는 것이 통계학의 또 다른 임무이다. 그림 1.1은 모집단, 표본, 표본집, 통계, 통계적 추론의 관계와 표본의 데이터로부터 통계를 만들고 모집단에 대해 통계적 추론을 하는 통계학의 역할을 잘 요약하고 있다.

1.2 통계프로그램

현재 우리나라에서 가장 많이 사용되고 있는 통계프로그램으로는 Excel, Minitab, Eviews, SPSS, SAS, Matlab, R 등을 들 수 있다.

그림 1.2는 Excel과 Minitab의 첫 실행화면을 보여준다. Excel은 1987년 Microsoft가 스프레드시트(spreadsheet) 프로그램으로 개발하였지만, 통계프로그램 기능을 지원하며 무엇보다도 MS Office에 포함되어 누구나 쉽게 구할 수 있기 때문에 통계프로그램으로도 널리 사용되고 있다. Minitab은 1972년 펜실베이니어주립대학

그림 1.1: 모집단, 표본, 표본추출, 통계, 통계적 추론, 통계학

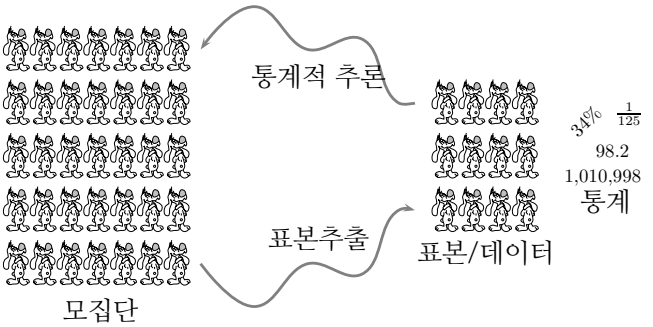
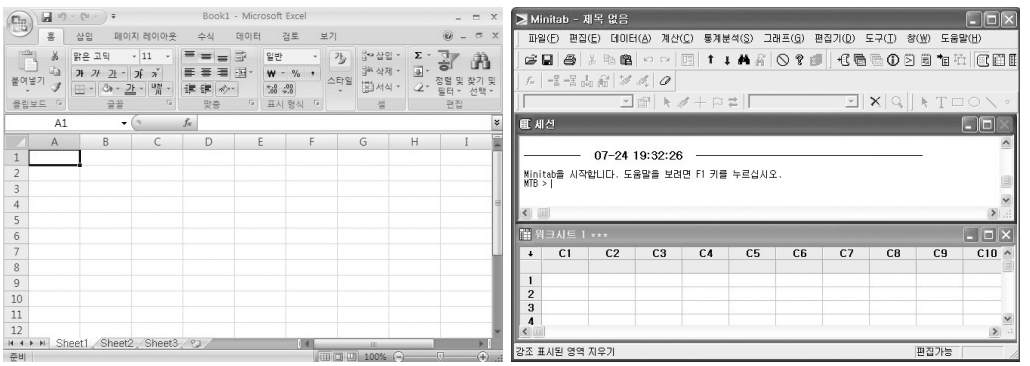


그림 1.2: Excel과 Minitab



(a) Excel (b) Minitab

(Pennsylvania State University)의 연구원들이 교육용 통계프로그램으로 개발하였는데, 통계에 자주 사용되는 명령문이 아이콘 형태로 툴바에 알기 쉽게 정리되어 있어서 교육용으로 널리 사용되고 있다.

그림 1.3은 Eviews의 실행화면을 보여준다. Eviews는 1994년 Quantitative Micro Software가 개발하였는데, 시계열의 분석에 유용한 도구가 많이 내장되어 있어 경제 시계열분석 통계프로그램으로 널리 사용되고 있다.

그림 1.4는 SPSS와 SAS의 실행화면을 보여준다. 사회과학용 통계패키지(Sta-

그림 1.3: Eviews

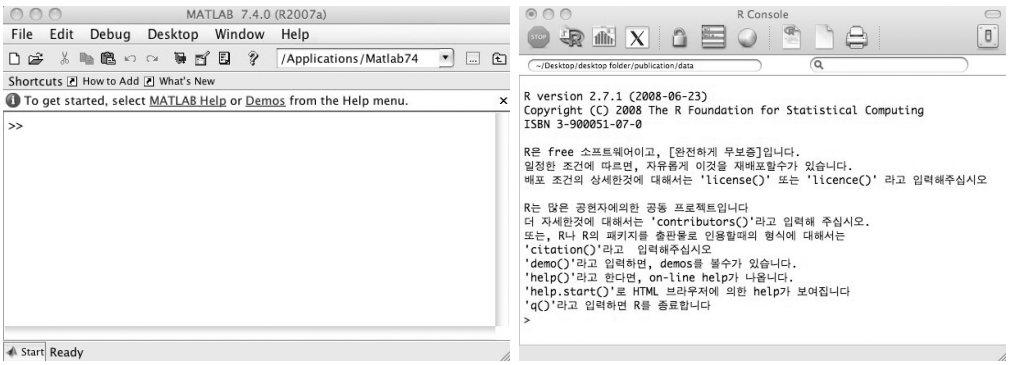
그림 1.4: SPSS와 SAS

(a) SPSS (b) SAS

tistical Package for the Social Sciences)란 의미를 가진 SPSS는 1968년 SPSS 회사가 개발하였으며, 설문조사기관, 정부기관, 교육기관, 금융기관 등에서 본격적인 통계프로그램으로 널리 사용되고 있다. 통계분석시스템(Statistical Analysis System)란 의미를 가진 SAS는 1966년 SAS Institute가 개발하였다. SAS는 통계프로그램 이외에도 데이터웨어하우징(data warehousing)과 데이터마이닝 등의 기능을 가지고 있어서 큰 사이즈의 데이터를 다루는데 적합하여, 설문조사기관, 정부기관, 교육기관, 금융기관 등에서 널리 사용되고 있다.

그림 1.5는 MATLAB과 R의 실행화면을 보여준다. 행렬실험실(matrix labo-

그림 1.5: MATLAB과 R



(a) MATLAB (b) R

ratory)이란 의미를 가진 MATLAB은 1970년대 말에 뉴멕시코대학(University of New Mexico)의 컴퓨터과학 교수인 Cleve Moler가 Fortran을 사용하지 않고 행렬 계산을 할 수 있는 프로그램으로 처음 개발하였으며, 교육기관, 이미지프로세싱 관련 산업에서 널리 사용되고 있다. R은 1997년 Ross Ihaka와 Robert Gentleman가 통계계산 및 그래프 프로그램으로 개발하였으며, 두 개발자의 이름 첫자를 따서 이름이 지어졌다. R 프로그램은 GNU 일반공중라이선스(GNU General Public License)에 따라 무료로 배포되고 있으며, 통계프로그램 개발과 데이터 분석에 널리 사용되고 있다.

Excel, Minitab, Eviews, SPSS, SAS, Matlab, R의 통계프로그램들은 서로 다른 장단점을 갖고 있기 때문에, 작업의 성격에 따라 보다 효율적인 통계프로그램을 선택하여 사용하는 것이 바람직하다. 예를 들어, Excel은 대부분의 컴퓨터에 설치되어 있어 어디서든 사용하기 쉽다는 장점이 있고, 교육용으로 개발된 Minitab은 본격적인 통계프로그램 중에서는 가장 쉽게 배울 수 있으며, SPSS는 마케팅이나 설문조사기관에서 사용하기 편하게 특화되어 있고, SAS는 큰 데이터를 다룰 수 있는 뛰어난 데이터마이닝(data mining) 기능이 있고, Matlab과 R은 수학적 연산을 자유롭게 할 수 있다. 그러나 이 책에서 다루는 통계학의 기초적인 내용은 어떤 통계프로그램으로도 쉽게 다룰 수 있기 때문에 굳이 특정 통계프로그램을 선택할 이유는 없다. 이 책은 위의

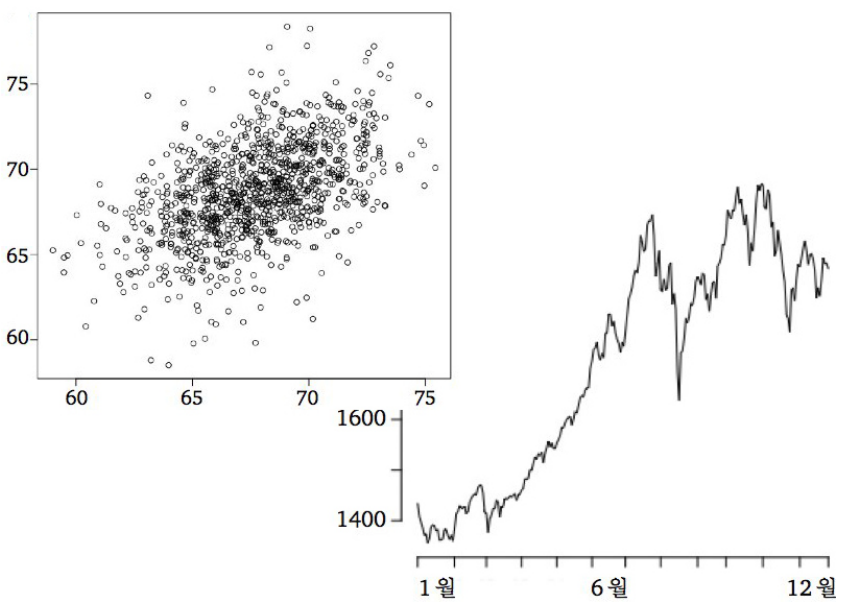
어떤 통계프로그램으로도 통계학을 학습할 수 있도록 모든 통계프로그램을 소개할 것이다.

연습문제

문제 1.1 모집단, 표본, 통계의 예를 들어 보라.

2 데이터의 요약

우리는 데이터를 통해서 사물을 인식한다.
데이터는 우리가 세상을 바라보는 창이다.

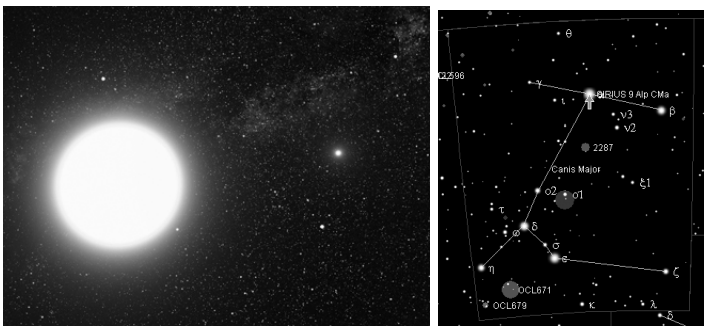


2.1 그래프와 표를 이용한 데이터의 요약

도수분포와 히스토그램

천체의 외관상 밝기는 외관등급(apparent magnitude)으로 측정되는데, 외관등급이 낮을수록 더 밝게 보이며, 외관등급이 1단위 작으면 약 2.512배 더 밝게 보이는 것을 의미한다. 태양의 외관등급은 -26.73, 달의 외관등급은 -12.6, 금성의 외관등급은 -3.7, 도시에서 육안으로 볼 수 있는 가장 희미한 별의 외관등급은 3, 가장 완벽한 조건에서 육안으로 볼 수 있는 가장 희미한 별의 외관등급은 6.5이다. 그림 2.1은 지구의 밤하늘에서 가장 밝은 별인 큰개자리(Canis Major)의 시리우스(Sirius)인데, 외관등급은 -1.44이다.[1] 별들 중에서 시리우스가 가장 밝게 보이는 것은 실제로 가장

그림 2.1: 시리우스와 큰개자리



(a) 시리우스 (b) 큰개자리

밝기 때문이 아니라 지구에서 가깝기 때문이다. 별의 실제 밝기는 절대등급(absolute magnitude)으로 측정되는데, 외관등급과 삼각시차(parralax, 단위는 arcsec 또는 $''$이며 $1'' = \frac{1}{3600}\,^\circ$)의 함수로 다음과 같이 주어지며,

$$\text{절대등급} = \text{외관등급} + 5(\log_{10} \text{삼각시차} + 1) \tag{2.1}$$

1) 그림 (a)는 NASA, ESA, 그림 (b)는 Zwergelstern가 만들었으며, 공용도메인이다.

지구에서 10pc(10parsec, 약 32.616광년) 떨어진 거리에 있을 경우 별의 외관등급을 나타낸다.

Hipparcos 폴더의 Sirius.csv 파일은 시리우스를 중심으로 하는 밤하늘의 일부에서 도시에서 육안으로 볼 수 있는 22개 별의 히파르코스 고유번호(HIP), 외관등급(Vmag), 삼각시차(Plx)의 데이터이다.[2] 표 2.1은 Sirius 데이터를 보여주는데, 절

표 2.1: Sirius 데이터

HIP	Vmag	Plx	Amag
23875	2.78	36.7	0.60
24436	0.18	4.2	-6.69
25336	1.64	13.4	-2.72
25606	2.81	20.5	-0.63
25930	2.25	3.6	-4.99
25985	2.58	2.5	-5.40
26241	2.75	2.5	-5.30
26311	1.69	2.4	-6.38
26634	2.65	12.2	-1.93
26727	1.74	4.0	-5.26
27366	2.07	4.5	-4.65
27989	0.45	7.6	-5.14
30324	1.98	6.5	-3.95
32349	-1.44	379.2	1.45
33579	1.50	7.6	-4.10
34444	1.83	1.8	-6.87
35264	2.71	3.0	-4.92
35904	2.45	1.0	-7.51
36188	2.89	19.2	-0.70
37279	0.40	285.9	2.68
39429	2.21	2.3	-5.95
39757	2.83	52.0	1.41

대등급(Amag)은 식 (2.1)에 따라 계산된 것이다.[3] 시리우스의 HIP는 32349인데,

2) 1997년 유럽우주기구(European Space Agency)는 그리스 천문학자 히파르코스(Hipparchus)의 이름을 딴 인공위성 히파르코스(Hipparcos, High Precision Parallax Collecting Satellite)를 이용하여 약 12만개 별의 외관등급과 삼각시차 등을 측정한 히파르코스 카탈로그(Hipparcos Catalogue)를 발표하였다. Hipparcos 폴더의 Hipparcos.csv는 히파르코스 카탈로그의 118,218개의 모든 별들에 대한 데이터이며, Sirius.csv는 시리우스를 중심으로 좌우 상하로 30도 이내의 밤하늘에서 도시에서 육안으로 볼 수 있는, 외관등급 3이하의 별들을 고른 표본이다.

3) 히파르코스 카탈로그의 삼각시차의 단위는 $\frac{1}{1000}''$이므로, $\text{Amag} = \text{Vmag} + 5(\log_{10}\frac{\text{Plx}}{1000} +$

외관상 밝기를 나타내는 외관등급은 -1.44로 표본에서 가장 밝지만, 실제 밝기를 나타내는 절대등급은 1.45로 표본에서 두번째로 어두운 별임을 알 수 있다.

표 2.2는 절대등급(Amag)의 도수(frequency, 변수값이 관측된 횟수)와 상대도수(relative frequency, 도수의 비율)를 구간별로 기록한 도수분포(frequency distribution)이다. 구간별 도수분포는 정보의 손실이 있는 대신 데이터의 특성을

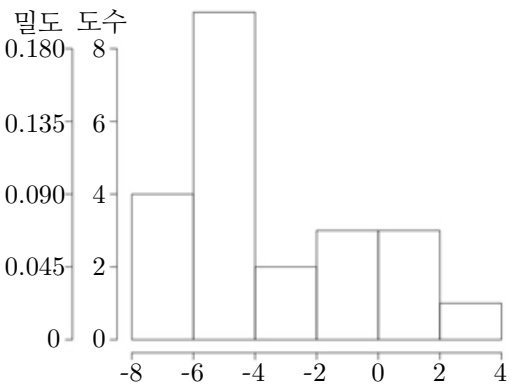
표 2.2: Sirius 데이터의 절대등급의 도수분포

절대등급	도수	상대도수
[-8, -6)	4	0.18
[-6, -4)	9	0.41
[-4, -2)	2	0.09
[-2, 0)	3	0.14
[0, 2)	3	0.14
[2, 4)	1	0.05

보다 알기 쉽게 요약하는 장점이 있다.

구간별 도수분포를 그래프로 나타낸 것을 히스토그램(histogram)이라고 한다. 그림 2.2는 표 2.2의 도수분포의 히스토그램이다. 히스토그램의 가로축에는 변수값의

그림 2.2: Sirius 데이터의 절대등급의 히스토그램



1)로 계산된다.

구간을 표시하고, 세로축에는 도수 또는 밀도를 표시하고, 구간별로 그에 해당하는 높이의 막대를 그린다. 밀도(density)는 상대도수를 구간의 폭으로 나눈 것이다. 히스토그램을 그릴 때에는 막대의 면적이 상대도수에 비례하도록 그려야 구간의 상대적인 비중에 대해 올바른 정보를 전달할 수 있다. 밀도의 높이로 막대를 그릴 경우, 막대의 면적은 구간의 폭과 밀도를 곱한 것이다. 그런데 밀도는 상대도수를 구간의 폭으로 나눈 것이기 때문에, 막대의 면적은 상대도수와 일치하고, 따라서 구간의 상대적인 비중에 대해 올바른 정보를 전달할 수 있다. 표 2.2의 구간별 도수분포는 구간의 폭이 일정하다. 이런 경우에는 도수의 높이로 막대를 그려도 막대의 면적이 상대도수에 비례하는 면적을 가지게 되어 밀도의 높이로 막대를 그린 경우와 마찬가지로 구간의 상대적인 비중에 대해 올바른 정보를 전달할 수 있다.

산점도와 두변수의 도수분포

1896년 통계학자 피어슨(Pearson, K.)은 키의 유전에 대한 우생학자 갈톤(Galton, F.)의 주장을 확인하기 위해 영국의 1078명의 아버지와 아들의 키를 조사하였다. Pearson 폴더의 Pearson.csv 파일은 피어슨이 조사한 아버지의 키(Fheight, 단위는 인치)와 아들의 키(Sheight)의 데이터이다.

두 변수의 분포를 그래프로 나타낸 것을 산점도(scatter plot)라고 한다. 그림 2.3은 Pearson 데이터의 두 변수의 산점도로, 가로축은 아버지의 키, 세로축은 아들의 키를 나타낸다.

표 2.3은 아버지의 키와 아들의 키의 도수를 구간별로 기록한 도수분포이다. 가운데 셀의 숫자는 특정 조합의 아버지의 키와 아들의 키의 구간이 관측되는 도수이다. 맨 오른쪽 열은 세로축 변수인 아들의 키의 구간별 도수분포이며, 맨 아래쪽 행은 가로축 변수인 아버지의 구간별 도수분포이다. 표 2.3과 같은 두 변수의 구간별 도수분포는 3차원 히스토그램으로 나타낼 수 있지만, 일반적으로 널리 사용되지 않는다.

그림 2.3: Pearson 데이터의 아버지의 키와 아들의 키의 산점도

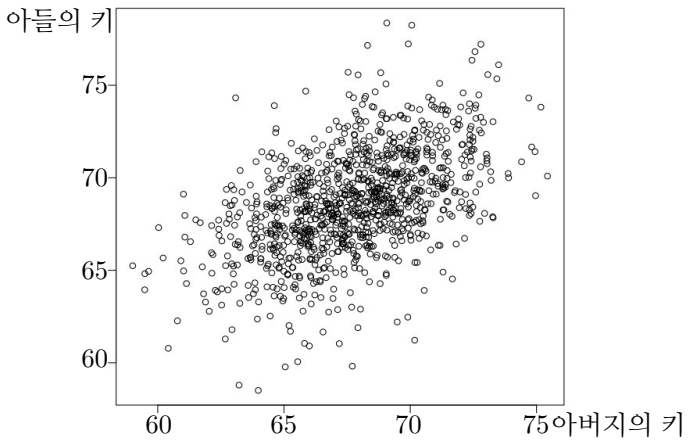


표 2.3: Pearson 데이터의 아버지의 키와 아들의 키의 도수분포

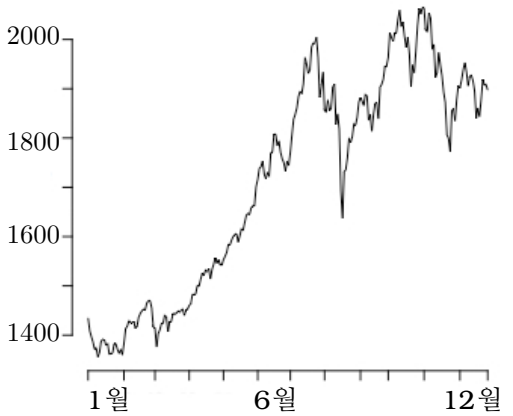
아버지의 키	아들의 키					
	[55,60)	[60,65)	[65,70)	[70,75)	[75,80)	
[55,60)	0	3	1	0	0	4
[60,65)	2	36	130	13	0	181
[65,70)	2	47	438	177	7	671
[70,75)	0	4	86	122	8	220
[75,80)	0	0	0	2	0	2
	4	90	655	314	15	1078

시계열그림

Stock폴더의 Stock.csv 파일은 2007년 한국종합주가지수(Kospi)와 코스닥지수(Kosdaq)의 일별 데이터인데, 이렇게 시간에 순서에 따라 관측된 데이터를 시계열데이터(time series data)라고 한다.

시계열데이터의 경우 시간에 따른 추이를 이해하는 것이 매우 중요한데, 시계열데이터의 추이를 보여주는 그래프를 시계열그림(time series plot)이라고 한다. 그림 2.4은 2007년 한국종합주가지수의 추이를 보여주는 시계열그림이다. 시계열그림의 가로축은 시간, 세로축은 변수값을 나타내고, 변수값들은 시간의 순서대로 선으로 연결하여

그림 2.4: 2007년 한국종합주가지수의 시계열그림



그린다.

2.2 통계를 이용한 데이터의 요약

변수의 특성을 결정하는 가장 중요한 요소는 변수값의 위치(location)와 스캐일(scale)이다. 위치와 스캐일에 대한 측도는 모집단과 표본에 대해 각각 따로 정의되는데, 표본의 데이터로 정의되는 측도(measure)가 통계이다.

위치에 대한 통계

위치에 대한 측도는 평균(mean), 중위수(median), 사분위수(quartile) 등이 있고, 변수값의 위치를 측정한다. 표본의 평균은 다음과 같이 정의되고,

정의 2.1. *(표본의 평균)* 표본 $\{X_i\}_{i=1}^n$의 평균 $\bar{X}$은 다음과 같이 정의된다.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.2}$$

표본의 중위수, 제1사분위수, 제2사분위수, 제3사분위수는 다음과 같이 정의된다.

정의 2.2. *(표본의 중위수와 사분위수)* 표본 $\{X_i\}_{i=1}^n$의 관측치가 오름순으로 $X_1 \le \ldots \le X_n$와 같이 정렬되어 있다면, 표본의 중위수 X_M*(*또는 동등하게 제*2*사분위수 X_{Q2}*)*는 다음과 같이 정의된다.

$$X_M = \begin{cases} X_{\frac{1}{2}(n+1)} & n\text{이 홀수} \\ \frac{1}{2}\left(X_{\frac{1}{2}n} + X_{\frac{1}{2}(n+1)}\right) & n\text{이 짝수} \end{cases} \tag{2.3}$$

한편 만약 $n+1$이 *4*의 배수이면 제*1*사분위수 X_{Q1}와 제*3*사분위수 X_{Q3}는 다음과 같이 정의되고,

$$X_{Q1} = X_{\frac{1}{4}(n+1)} \tag{2.4}$$

$$X_{Q3} = X_{\frac{3}{4}(n+1)} \tag{2.5}$$

그렇지 않은 경우에는 여러 가지 종류의 선형보간법*(linear interpolation)*으로 정의된다.*a)*

a) *Excel*에서는 표본 $\{X_i\}_{i=1}^n$의 X_i를 $\frac{i-0.5}{n}$ 분위수*(quantile)*로, *Minitab*과 *SPSS*에서는 $\frac{i}{n+1}$ 분위수로, *SAS*와 *Matlab*에서는 $\frac{i-1}{n-1}$ 분위수로 가정하고, *0.25, 0.5, 0.75* 주위의 두 분위수의 변수값을 이용하여 선형보간법으로 사분위수를 구한다. *R*에서는 기본적으로 *Excel*에서의 방법으로 구하지만 옵션을 지정하여 세 방법을 선택할 수 있다.

스캐일에 대한 통계

스캐일에 대한 측도로는 분산(variance), 표준편차(standard deviation), 범위(range), 사분위범위(interquartile range) 등이 있고, 변수값의 퍼진정도(dispersion) 또는 변동성(variability)을 측정한다. 표본의 분산과 표준편차는 다음과 같이 정의되고,

정의 2.3. *(표본의 분산과 표준편차)* 표본 $\{X_i\}_{i=1}^n$ 의 분산 s^2 과 표준편차 s 는 다음과 같이 정의된다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{2.6}$$

$$s = \sqrt{s^2} \tag{2.7}$$

표본의 범위와 사분위범위는 다음과 같이 정의된다.

정의 2.4. *(표본의 범위와 사분위범위)* 표본 $\{X_i\}_{i=1}^n$ 의 관측치가 오름순으로 $X_1 \leq \ldots \leq X_n$ 와 같이 정렬된다고 하면, 표본의 범위 X_R 와 사분위범위 X_{IQR} 는 다음과 같이 정의된다.

$$X_R = X_n - X_1 \tag{2.8}$$

$$X_{IQR} = X_{Q3} - X_{Q1} \tag{2.9}$$

선형상관에 대한 통계

선형상관(linear correlation)은 두 변수간의 관계의 특성을 나타내는 중요한 요소이며, 이에 대한 측도는 공분산(covariance)과 상관계수(correlation coefficient)가 있다. 표본의 공분산은 표본에서 두 변수의 관측치가 선형상관의 관계를 가지는지 나타내며, 양의 숫자이면 양의 선형상관을 음의 숫자이면 음의 선형상관을 가짐을 의미한다. 공분산은 선형상관의 존재여부와 방향을 나타내지만 선형상관의 크기를 나타내지는 않는다. 상관계수는 공분산을 두 변수의 표준편차로 나눈 것으로, -1에서 1의 값을 가지고, 선형상관의 방향과 함께 크기도 나타낸다. 표본의 공분산과 상관계수는 다음과 같이 정의된다.

정의 2.5. *(표본의 공분산과 상관계수)* 표본 $\{X_i, Y_i\}_{i=1}^n$ 의 공분산 s_{XY} 과 상관계수 r_{XY} 는 다음과 같이 정의된다.

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{2.10}$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \tag{2.11}$$

표 2.4는 Pearson 데이터의 아버지의 키와 아들의 키의 위치와 스캐일, 선형상관에 대한 통계이다. 앞에서 설명한 바와 같이 제1사분위수와 제3사분위수, 그리고

표 2.4: Pearson 데이터의 아버지의 키와 아들의 키의 요약통계

통계	아버지의 키			아들의 키		
평균	67.69			68.68		
최소값	59.01			58.51		
제1사분위수	65.79	65.78	65.79	66.93	66.93	66.93
중위수	67.77			68.62		
제3사분위수	69.60	69.60	69.60	70.47	70.47	70.47
최대값	75.43			78.36		
분산	7.53			7.92		
표준편차	2.74			2.81		
범위	16.43			19.86		
사분위범위	3.82	3.82	3.82	3.53	3.54	3.54
공분산	3.87			3.87		
상관계수	0.50			0.50		
	(1)	(2)	(3)	(1)	(2)	(3)

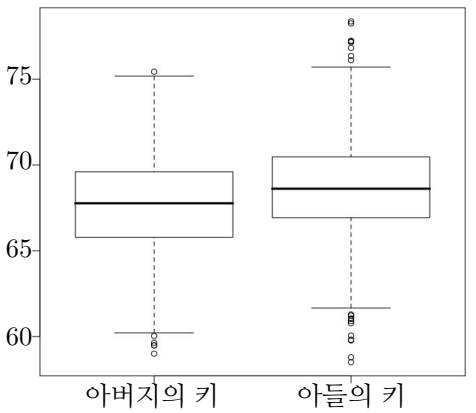
주 : (1) Excel, R, (2) Minitab, SPSS, (3) SAS, Matlab

사분위범위를 구하는 방법은 선형보간법에 따라 달라지고, 때문에 통계프로그램에 따라 그 결과가 달라진다.

상자그림

변수의 위치와 스케일에 대한 측도 중에서, 최소값, 제1사분위수, 중위수, 제2사분위수, 최대값, 범위, 사분위범위를 그래프로 나타낸 것을 상자그림(box plot)이라고 한다. 그림 2.5는 Pearson 데이터의 아버지의 키와 아들의 키의 상자그림이다. 상자그림을

그림 2.5: Pearson 데이터의 아버지의 키와 아들의 키의 상자그림



그릴 때는, 일반적으로 제1사분위수보다 1.5배의 사분위범위보다 작은 관측치, 또는 제3사분위수보다 1.5배의 사분위범위보다 큰 관측치는 이상점(outlier)으로 간주하여 따로 표시하고, 그 외의 관측치에서 최소값과 최대값을 선택해 범위를 표시한다.

연습문제

문제 2.1 IQEnglish 폴더의 C1.csv는 2004년 서울의 어느 남자중학교 3학년 1반의 IQ 테스트와 영어성적(1학기 중간고사 성적, English)의 데이터이다. IQ 테스트는 어휘력(Language), 추리력(Reasoning), 수리력(Math), 지각력(Spatial) 등 4 항목의 테스트로 이루어지며, IQ는 이 점수의 합계를 2로 나눈 것이다. 통계프로그램을 이용하여 다음의 질문에 답하라.

(1) IQ변수를 만들고, IQ변수의 구간별 도수분포를 구하고, 히스토그램을 그려라.

(2) IQ와 영어성적의 산점도를 그리고, 구간별 도수분포를 구하라.

(3) IQ와 영어성적의 요약통계(평균, 최소값, 제1사분위수, 중위수, 제3사분위수, 최대값, 분산, 표준편차, 범위, 사분위범위, 공분산, 상관계수)를 구하고, 상자그림을 그려라.

문제 2.2 Stock 폴더 Stock.csv 파일의 코스닥지수(Kosdaq)의 일별 데이터의 시계열그림을 그려라.

찾아보기

저자소개

위스콘신-매디슨대학(University of Wisconsin-Madison) 경제학학사
시카고대학(University of Chicago) 경제학박사
현재, 홍익대학교 조교수