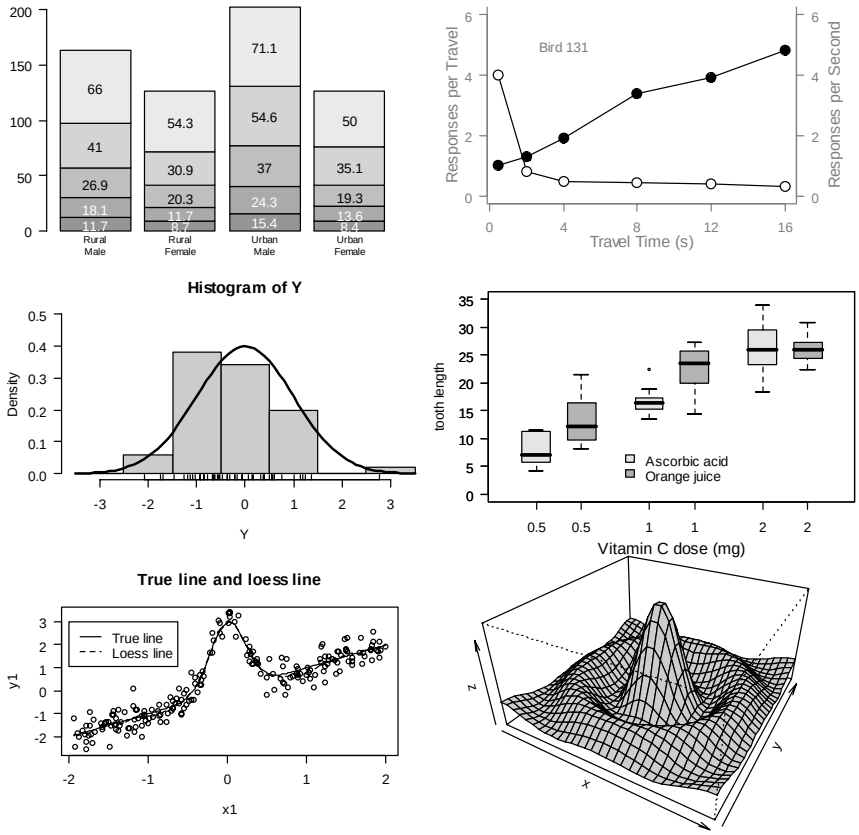
비전공자를 위한 통계학 이야기

통계 마인드



통계교육원

머 리 말

21세기를 살아가는 사람들에게는 정보를 분석하고 판단하는 능력이 매우 중요하다. 날마다 홍수처럼 쏟아지는 숫자를 올바르게 해석할 수 있다면 그들의 삶의 질이 달라질 것이기 때문이다. 통계는 다양한 분야에서 중요한 역할을 수행한다. 전달에 비해 실업률이 3.5%, 소비자물가가 4% 상승이라고 하는 뉴스를 숫자 자체로 생각할 수 있다. 하지만 관련된 자료를 자신의 업무에 적용해야 하는 사람들은 이 통계를 어떻게 조사하고 분석하는지를 알아야만 일을 할 수 있다. 한마디로 통계를 생산하기 위해서도, 또한 통계를 이용하기 위해서도 통계이론이 필요하다는 점을 강조하고 싶다.

통계학은 불확실한 상황 하에서 정보의 수집, 분석, 그리고 추정과 검정을 통하여 올바른 의사결정의 방법을 연구하는 학문이다. 이와 같이 올바른 의사결정을 내리기 위해서는 모집단으로부터 추출한 표본으로부터 필요한 자료를 수집해야 하고, 수집된 정보를 정리하고 분류하여 유용한 자료로 만든 다음 통계적 방법을 이용하여 의사결정을 하게 된다. 그러나 통계적 방법은 매번 크고 작은 오류를 범할 가능성을 내포하게 된다. 예를 들어 확률의 잘못된 해석, 여론조사 결과의 오해, 분석방법의 잘못 등이 그 예이다. 또한 많이 변화하고 있지만 지금까지의 통계교육은 공식에 의한 문제 해결에 초점을 두고 있어 매우 어렵게 느끼고 있는 것이 사실이고, 실생활과 밀접한 필수 학문이라는 사실을 알려주는 방법을 택하고 있기에 무엇인가를 강요한다는 느낌이 든다.

본 서는 다음과 같은 의도로 기획되었다. 첫째, 업무상 통계 지식이 반드시 필요한 사람들을 위한 책으로 구성하였다. 둘째, 통계학을 전공하지 않은 비전공자를 위한 통계학 개론서로 제작되었다. 셋째, 사회 각계 분야에서 자신의 일을 소중하게 생각하며 열심히 근무하는 직업인을 대상으로 하였다. 넷째, 언론에 보도되는 통계를 자신의 일상생활과 비교해 볼 수 있는 능력을 부여한다는 점에서 국민들을 대상으로 제작하였다. 한마디로 **통계 지식이 필요한 어른들을 위한 통계학 개론서**이다.

본 서는 통계학을 학습하는 순서대로 총 4부로 제작되었다. 처음부터 정독을 하는 사람들을 위한 배려이다 보니 대학에서 사용하는 통계학 개론서와 많이 닮아있기도 하다. 다른 점이 있다면 나름대로 풍부한 예제, 신문기사 등을 함께 보여줌으로써 현장감을 높인 점이다. 또한 통계와 통계학의 존재 이유, 필수개념을 설명하는 과정에서도 업무에 자주 활용하는 개념을 설명할 때에는 쉽게 이해할 수 있도록 수식을 줄여 말로 설명하고 있다는 것 등의 특징을 갖고 있다. 덧붙이자면 통계학의 개념을 충실히 설명하되, 실제 업무에 많이 활용하는 예제를 중심으로 설명하고 실습문제를 제시하였다는 의미이다. 혼자서 통계학을 공부하기에는 어려운 점이 많다. 이를 참고도서로 활용하면 교육 효과는 물론 업무에 적용할 때에도 그 효용성이 극대화될 수 있다고 생각된다.

본 서는 통계교육원에서 추진하는 표준교재 개발의 일환으로 탄생하였다. 과제를 수행해 주신 분들은 장대흥 교수(연구책임자, 부경대학교 수리통계학부), 김영일 교수(중앙대학교 정보시스템학과), 이태림 교수(방송통신대학교 정보통계학과), 강명희 교수(이화여자대학교 교육공학과)이다. 넉넉하게 연구할 시간과 적정한 사업비를 가지고 진행했어야 할 사업임에도 그렇지 못해서 안타까웠으나 저자 분들도 사업목적을 공감하고 최선을 다해 이 책을 만들어 주셨기에 지면을 빌어 존경과 감사의 말씀을 드린다.

통계교육원은 통계에 관한 국민의 관심이 높아지도록 지속적인 노력을 할 것이며 이때에 이 책이 유용한 자료로 활용될 수 있음을 확신한다. 끝으로 결과를 정리하고 편집한 통계교육원 직원들에게도 고마움을 전한다.

2008년 7월

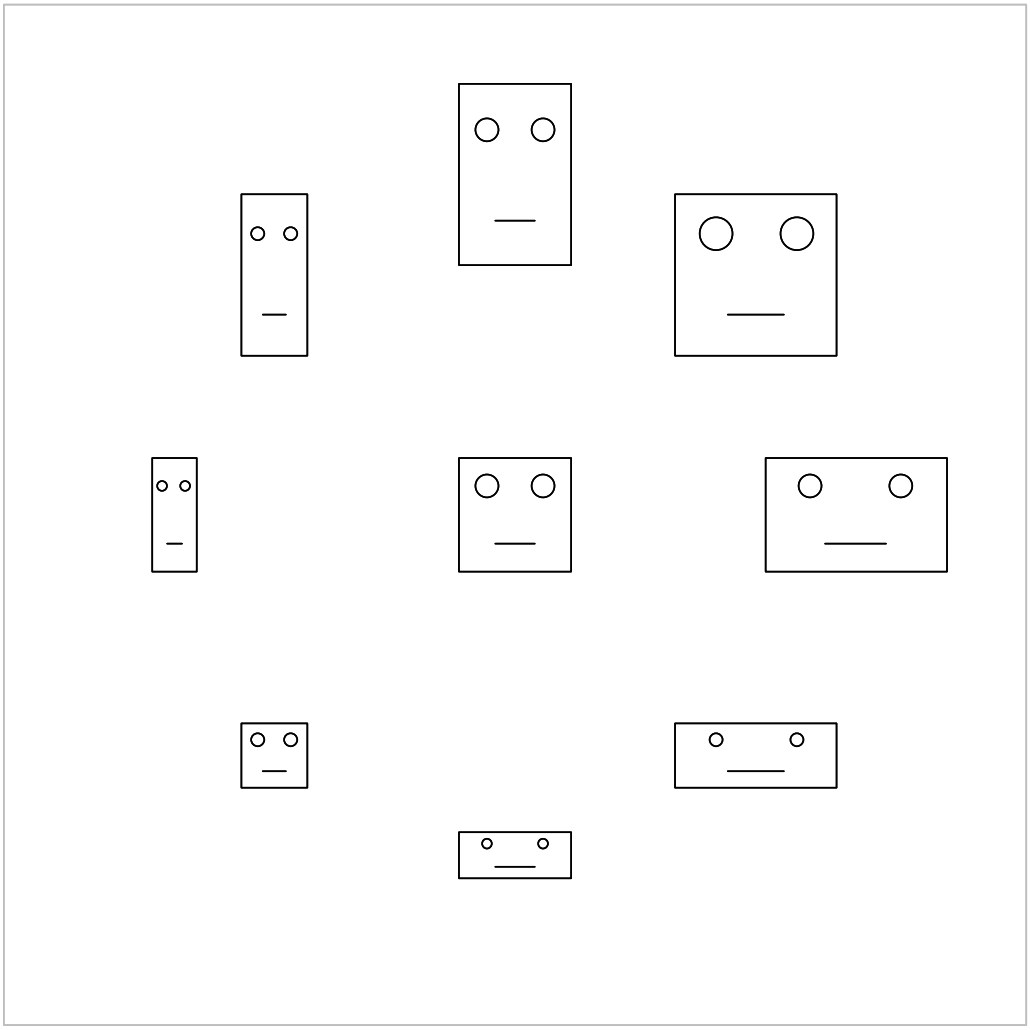
통계교육원장 신 승 우

차 례

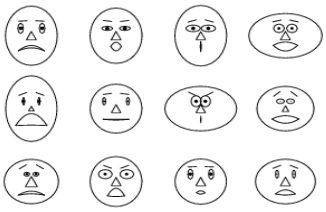
쉬어가기 차례

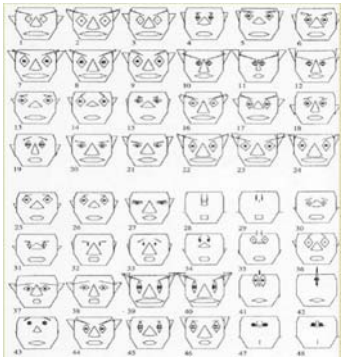
제 1 부

자료와 통계학



- 앞의 그림은 사람의 얼굴을 이용하여 다변량자료(변수가 여러 개 있는 자료)를 나타내는 체르노프 얼굴(Chernoff face)의 변형그림이다. 눈의 크기, 입의 크기, 얼굴의 좌우, 위아래 길이에 각각 변수들을 대응시키면 총 4개의 변수를 표현할 수 있다. 다음 그림들은 또 다른 체르노프얼굴들이다. 왼쪽 체르노프얼굴에서는 눈의 크기, 눈의 기울기, 눈동자의 크기, 눈썹의 기울기, 입의 크기, 입의 기울기, 코의 크기, 얼굴의 크기에 각각 변수들을 대응시키면 총 8개의 변수를 표현할 수 있다.





- 카이스트의 박경수 교수 웹페이지 http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm에 가보면 움직이는 체르노프얼굴을 볼 수가 있다. 다변량 시계열자료에 이 움직이는 체르노프 얼굴을 사용할 수 있다.


All 0.

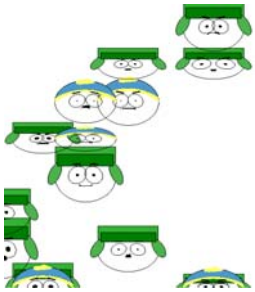

All 0.5.


All 1.


Randomly selected parameters.
Change face.

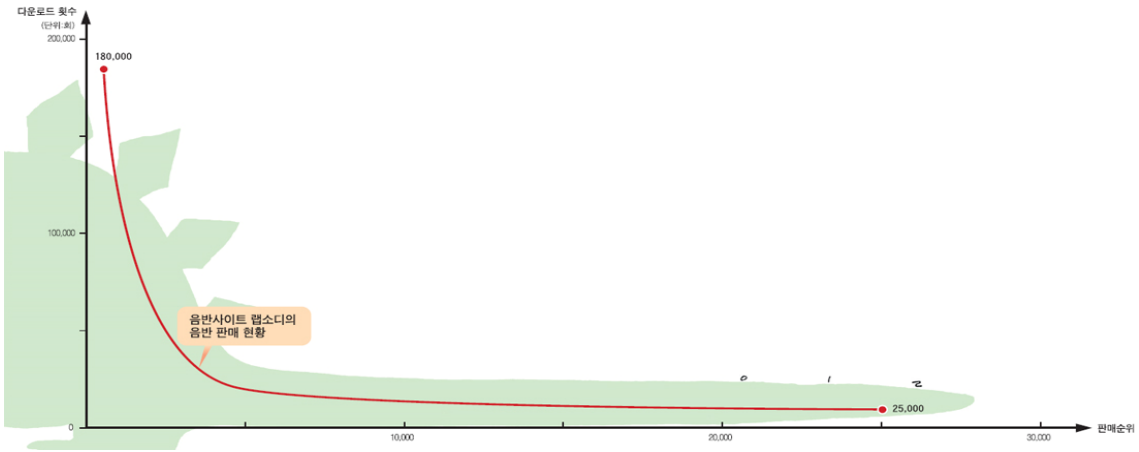

Animation with random parameters.

- Flury와 Riedwyl이 제시한 비대칭 체르노프 얼굴(asymmetrical Chernoff face)도 있다. 전체 얼굴을 좌우로 나누고 각 얼굴에서 최대 18개 변수까지 표현한다. 총 최대 36개의 변수를 표현할 수 있다.
- 아래 그림은 유명한 만화애니메이션 'South Park' 등장인물을 이용하여 만든 Chernoff Park plot이다.



제 1 장

통계는 선택이 아닌 필수이다.



차 례

학습목표

우리는 일상생활에서 여러 가지 통계자료에 접하게 된다. 이러한 통계자료를 분석하여 유용한 정보를 얻기 위하여 통계학을 사용한다. 정보화 사회에서 통계학의 중요성은 점점 커지고 있는 반면 통계문맹도 심각한 문제를 야기하고 있다. 통계학의 필요성과 통계문맹에 대하여 학습한다. 통계자료분석의 과정을 살펴보고 실험연구와 관측연구의 차이에 대하여 학습한다.

1.1 통계학이란 무엇인가?

지식기반사회에서는 평범한 사람들도 매일 매일의 생활에서 수많은 데이터를 만나게 된다. 이러한 데이터는, 이 데이터가 우리들에게 무엇을 말해주고 있는 지를 판단하는 판단력을 요구하게 된다. 요즈음의 데이터는 유비쿼터스화되고 있다. 데이터들이 사방에 널려있는 환경 속에서 살고 있다. 그래프, 차트, 비율, 퍼센트, 확률, 평균, 예측 등이 우리들의 일상생활에 들어와 있다. 우리들은 매스컴에서 수시로 시행하여 발표하는 여론조사들의 결과를 만나게 되고 광고, 정치 및 사회 분석가, 경제예측전문가들의 메시지를 접하게 된다. 예를 들어, 정당지지율, 부패인식지수, 뇌물공여지수, 정보화지수, 여성개발지수, 이혼율, 자살률, 행복지수, 고객만족지수, 물가지수, 경제성장률, 인구증가율, 실업률, 쌀생산량, 산업생산지수, 국민총생산고(GNP) 등 일일이 열거할 수 없을 정도로 많다. 예로 질병의 하나인 담배 의존증(흡연)과 관계된 다음과 같은 2008.02.08 연합뉴스 기사를 보자.

"금연 지출비 고작 담배세 수입의 500분의 1" <WHO>

금세기 흡연사망자 10억명 추산..`MPOWER 캠페인' 발족

각국 정부들이 해마다 천문학적인 담배세를 거둬들이면서도 금연 대책을 위해서는 담배세 수입의 500분의 1에 불과한 비용을 지출하고 있다고 유엔 산하 세계보건기구(WHO)가 7일 밝혔다.
마거릿 찬 WHO 사무총장은 이날 뉴욕에서 마이클 블룸버그 뉴욕시장과 함께 가진 기자회견에서 그 같은 내용을 담은 `'글로벌 담배 유행병 리포트(Global Tobacco Epidemic Report)'를 발표하고 금연 운동에 지속 가능한 재정을 지원할 것을 촉구했다고 WHO가 전했다.
보고서에 따르면, 고소득 국가의 경우 금연 대책 관련 지출비가 담배세 수입의 340분의 1 이하인 반면에 중소득과 저소득 국가들은 각각 4천분의 1과 9천분의 1에도 미치지 못하는 있다. 중.저소득 국가들의 경우 665억 달러(62조6천억원)를 거둬 고작 1천400만 달러를 금연 대책에 투입한 것으로 조사됐다.
자기 나라에 포괄적인 금연 관련 법률이 있어 흡연으로부터 보호를 받는 인구는 단지 세계 인구의 5%에 그쳤고, 병원이나 학교에서도 흡연을 허용하고 있는 국가들이 무려 전체의 40%에 이르고 있다.
담배 의존증에 대한 완벽한 치료 서비스가 가능한 나라는 전 세계 인구의 5%에 해당하는 9개국에 불과하며, 국가에서 담배 광고 및 판매촉진를 일괄 금지하는 나라들의 인구도 전체의 5%에 그쳤다.
또한 세계 인구의 6%에 해당하는 15개국만이 담배를 포장하면서 `'그림 경고문'을 게재하고 있었다.
WHO는 이날 회견을 통해 담배 반대를 위해 ▲담배 사용 및 예방정책 모니터(M) ▲담배 연기로부터 보호(P) ▲금연을 위한 지원 제공(O) ▲담배의 위험성 경고(W) ▲담배 광고· 판매촉진· 후원 금지 강화(E) ▲담배세 인상(R) 등 6가지의 전략을 담은 `MPOWER' 캠페인을 발족시켰다.
찬 총장은 "이 6가지 전략은 부유하든 가난하든 모든 나라가 활용할 수 있다"면서 "이들 전략을 한꺼번에 사용한다면 (담배라는) 이 유행병의 확산을 역전시킬 수 있는 최선의 기회를 얻게 될 것"이라고 말했다.
이번 보고서 작성에 재정 지원을 했던 블룸버그 뉴욕시장은 "MPOWER 정책들을 완벽히 이행하는 나라는 하나도 없으며, 그 중 한 가지 정책을 완벽하게 이행하는 나라도 80%에 그치고 있다"면서 "이제야 처음으로 담배라는 유행병을 막는 동시에 우리 모두가 책임있게 나서는데 도움이 되는 확실한 데이터들이 확보됐다"고 강조했다.

보고서는 또한 개도국의 청소년 및 성인을 타깃으로 삼아 매년 수백만명이 담배에 중독되도록 하는 글로벌 담배산업의 전략에 따라 담배라는 유행병이 개도국으로 옮겨가고 있다고 지적한 뒤, 담배 관련 연간 사망자 800여만 명의 80%가 2030년 즈음에는 개도국에서 발생할 것으로 예상된다고 덧붙였다.
나아가 보고서는 특히 젊은 여성들을 타깃으로 삼은 전략은 "이 담배라는 유행병의 증가와 관련해 가장 불길한 잠재 요소 중 하나"라면서 개도국 정부들이 담배세 인상과 흡연 경고 등을 통해 적극 대처하지 않을 경우 21세기에는 담배 관련 사망자가 20세기의 10배인 10억 명이 될 것이라고 경고했다.
끝으로 보고서는 중국이 현재 세계 최대의 담배 생산국인 동시에 소비국이라고 지적하고 중국 남성의 60% 가까이가 아직도 흡연을 하고 있다고 덧붙였다.

우리는 타이틀을 포함한 이 기사 내용에서 다음 [표 1.1]과 같이 총 23가지의 데이터들이 언급되어 있음을 알 수 있다. 또한 이러한 데이터들을 통하여 금연운동에 대한 미흡한 재정지원이 전 세계적인 현상임을 알 수 있다. 부차적으로 환율적용을 941원/$(=626,000/665)로 하고 있음도 알 수 있다.

언급된 수의 종류	내용(괄호 안은 중복된 빈도수)
배율	500분의 1(2), 340분의 1, 4천분의 1, 9천분의 1, 10배
퍼센트	5%(3), 6%, 40%, 60%, 80%(2)
사람수	수백만명, 800여만명, 10억명(2)
금액	665억달러, 62조 6천억원, 1천400만달러
년도, 일자	2030년, 20세기, 21세기, 7일
나라수	9개국, 15개국
경우수	6가지(2)

[표 1.1] 연합뉴스 기사에 나타나는 데이터들

일상생활에서 쓰이는 데이터들에 대하여 몇 가지 그림을 더 보자.

예제 1.1 다음 [그림 1.1]은 2007년 9월 초 두바이에서 발행되는 한 신문에 실린 광고이다. 지면 광고의 카피를 보면 '매년 5백 4십 만명이 흡연 관련 질병으로 사망합니다. 9/11 테러 희생자의 2천배에 달하는 숫자입니다.'라고 되어 있다. 이 광고는 두바이의 광고회사 퍼셉트걸프(PerceptGulf, Dubai, UAE)가 제작한 것인데, 이 회사의 크리에이티브 디렉터(Creative Director)인 Prashant Sankhe는 이 광고가 '흡연이라는 테러에 사로잡히면 9/11 테러와 똑같은 일이 몸에서 일어날 수 있다'는 사실을 보여주는 작품이라고 설명하였다고 한다. 이 광고가 미국 네티즌들 사이에서 큰 논란과 화제를 일으키고 있는데 광고 속 불타는 담배 이미지가 여객기와 충돌한 세계 무역 센터를 연상시키기 때문이다. 그러면 9/11 테러 희생자수는 몇 명일까? 정확히 2,973명이다. 그러므로 매년 흡연 관련 질병으로 인한 사망자수는 9/11 테러 희생자의 1, 816배, 대략 2천배가 된다. 이렇듯 우리의 일상생활은 온통 수많은 숫자로 뒤덮여 있고 이러한 숫자가 우리의 인생을 좌지우지 할 때가 많다. ■

(제공처: farrukh.wordpress.com)

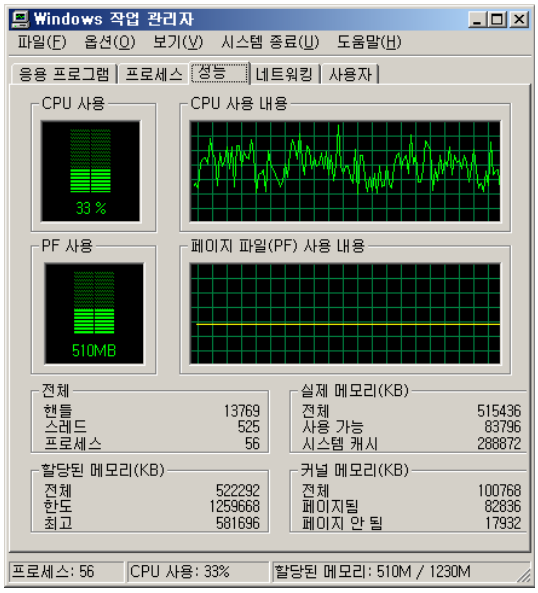
[그림 1.1] 두바이 신문에 실린 광고

예제 1.2 다음 [그림 1.2]는 2007.06.14부터 2007.12.14까지의 6개월간 우리나라의 코스피 지수와 거래량을 각각 꺾은선그래프와 막대그래프로 나타낸 그림이다. 코스피지수의 급등락 현상이 자본시장에서 얼마나 큰 영향을 주고 있는지를 보면 코스피지수라는 하나의 수치가 우리나라 경제의 역동성, 보는 시각에 따라서는 불안전성을 적나라하게 나타내고 있다고 하겠다. ■

(제공처: 증권선물거래소, www.krx.co.kr)

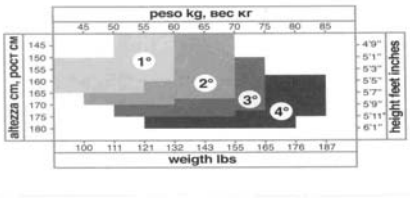
[그림 1.2] 2007년 하반기 6개월간 코스피지수와 거래량

예제 1.3 다음 [그림 1.3]은 우리가 컴퓨터를 사용할 때 운용시스템으로 주로 사용하는 MS 윈도우 상에서의 윈도우 작업관리자 창을 나타내고 있다. CPU 사용(퍼센트로 나타냄)을 숫자와 막대그래프로 실시간 나타내고 있고 CPU 사용이력(history)을 꺾은선그래프로 나타내고 있다. 이러한 숫자들과 그래프들로 인하여 자기가 쓰고 있는 컴퓨터 CPU 사용 상황을 실시간으로 확인하여 볼 수 있다. ■



[그림 1.3] MS 윈도우 상에서의 윈도우 작업관리자 창

예제 1.4 다음 [그림 1.4]는 여성용 스타킹 포장지에 그려져 있는 그림이다. x축은 몸무게, y축은 키를 나타내는 산점도의 변형이라 볼 수 있다. 4 종류의 스타킹 중 자기의 몸무게와 키에 해당하는 스타킹을 고르면 된다. ■



[그림 1.4] 여성용 스타킹 포장지에 그려져 있는 그림

통계자료가 우리 생활 깊숙이 들어와 있음을 실감할 수 있는 또 다른 예를 살펴보자. J씨는 2007.10.02 통계청으로부터 한 통의 휴대폰 메시지를 받았다. 통계청이 단문메세지서비스(SMS)를 통하여 J씨 개인 휴대폰으로 보낸 휴대폰 메시지에는 다음과 같이 내용이 쓰여 있었다.

'(통계청) 2007년 9월 소비자물가 전월대비 0.6%, 전년 동월대비 2.3% 상승'
'(통계청) 2007년 8월 서비스업 생산은 전년 동월대비 7.3% 증가'

이 간단한 메시지를 통하여 J씨는 우리나라 경제에 대한 통계자료를 아주 쉽게 습득하게 된다. 통계청에서는 매월 보도·공표되는 10개 주요통계지표를 실시간으로 공표즉시 이동통신단말기(휴대폰, 휴대폰 기능을 가진 PDA)를 가진 다수의 이용자에게 단문메세지서비스를 통하여 전송해 알려줌으로써 이용자에게 찾아가는 서비스를 실시하고 있다. 또한 무선인터넷을 통하여 통계청의 새소식(보도자료, 공지사항, 주요일정), 주요통계지표(서비스업, 소비자물가, 실업률, 산업 등 분야별), 민원안내(업무검색, 이름검색)를 이동통신단말기로 제공함으로서 이용자의 편의성을 도모하고 있다.

통계학, 너는 누구냐?

통계학(statistics)의 어원이 라틴어에서 국가라는 의미를 갖는 'status'에서 유래되었다는 사실이나 'statistics'라는 단어가 나타나기 전 쓰였던 단어가 'political arithmetic(정치산술)'이라는 사실에서 알 수 있듯이 오랜 시간동안 통계라는 것은 한 국가의 지표로서 경제, 인구, 정치상황을 자료나 도표로 나타내는 것과 동일시 되어왔다. '통계학'을 한자로 쓰면 '統計學'이 되는데 여기서 '統'은 거느릴 통(govern)이고 '計'는 셈할 계(device)이다. 통계학이라는 용어에도 국가를 위한 집계라는 의미가 강하게 있다고 하겠다.

그러면 통계학이란 무엇인가? 통계학(statistics)을 한 문장으로 표현한다면

> 우리의 일상생활에서 얻어지는 다양한 통계자료를 **수집**, **정리**하고 **평가**하고 의미있는 **결론**을 이끌어내는 작업을 수행하는 학문

이라고 정의할 수 있다. 미국통계협회(ASA) 회장을 역임한 Jon Kettenring은 통계학의 성격에 대하여 다음과 같이 말하고 있다(1997).

"I like to think of statistics as the science of learning from data...It presents exciting opportunities for those who work as professional statisticians. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels."

통계학을

> 데이터로부터 배우는(**데이터의 지식화**) 과학

이라고 정의하였고 통계학을 국가나 기업, 나아가 모든 현대 교육과정의 중요 구성요소라고 강조하였다. 또한 미국통계협회(ASA) 홈페이지에는 통계학을 다음과 같이 정의하고 있다.

"Statistics is the scientific application of mathematical principles to the collection, analysis, and presentation of numerical data. Statisticians contribute to scientific inquiry by applying their mathematical and statistical knowledge to the design of surveys and experiments; the collection, processing, and analysis of data; and the interpretation of the results. Statisticians may apply their knowledge of statistical methods to a variety of subject areas, such as biology, economics, engineering, medicine, public health, psychology, marketing, education, and sports. Many economic, social, political, and military decisions cannot be made without statistical techniques, such as the design of experiments to gain Federal approval of a newly manufactured drug."

통계학이란 조사와 실험에 대한 설계, 데이터의 수집, 처리, 분석, 결과의 해석을 행하는 과학이라 정의하고 있다. 또한, 통계의 본질에 대하여 다음과 같이 설명하고 있다.

"The Nature of Statistics

Statistics provides the reasoning and the methods for producing and understanding data. Statisticians are specialists, but statistics by its nature demands that they be generalists also.

Mathematics and Computers Are Involved ...

Statistics uses mathematics, but it is not abstract or isolated: statisticians work with people from other professional backgrounds to solve practical problems. Statistics uses modern computing to organize and analyze data, and statisticians command specialized tools, but the emphasis is on the data to be understood and the problem to be solved rather than on computing for its own sake.

... But Understanding the Data Is Crucial

Statisticians must know more than statistics. A statistician who works in medicine or in a manufacturing plant or in market research must learn enough medicine or engineering or marketing to understand the data in their setting. Statisticians need the ability to work with other people, to listen, and to communicate."

통계학은 데이터를 생산하고 이해하는 논리와 방법들을 제공한다. 통계학에서 수학과 컴퓨터를

사용하기는 하나 추상적인 학문도 아니고 홀로 존재하는 학문도 아니다. 통계학자들은 현실적인 문제들을 해결하기 위하여 다른 분야의 전문가들과 같이 일한다. 그러므로 통계학자들은 이러한 분야들에 대한 식견이 있어야 하고 다른 분야 전문가들의 이야기를 듣고 공동작업을 할 능력이 있어야 한다. 통계학에서는 컴퓨터의 계산 자체 보다는 이해해야 할 데이터 자체와 풀어야 할 현실 문제가 더 중요하다.

그러면 단수로 쓰이는 statistics(통계학)과 복수로서 쓰이는 statistics(통계량들, 통계치들)의 차이는 무엇일까? 이 두 단어를 구별하기 위하여 우선 Wikipedia에 가보자.

statistics(통계학, 統計學): a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.
statistic(통계량 統計量, 통계치 統計値): the result of applying a function(statistical algorithm) to a set of data

단수로 쓰이는 statistics는 '통계학'이라는 학문(자료를 수집, 분석, 해석, 설명, 표현하는 수리과학)을 가리키고 복수로서 쓰이는 statistics는 통계량(자료에 함수를 적용한 결과, 실제로 관찰이 이루어지기 전에 취할 수 있는 모든 값과 그에 대응하는 가능성을 총칭하는 양)이나 통계치(통계량의 실제 관측값)의 복수형이다. 우리가 보통 '통계'라고 말하는 단어가 바로 이 통계량이나 통계치를 가리킨다.

인터넷백과사전 Microsoft Encarta (encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx)에서는 'statistics'와 'statistic'을 어떻게 표현하고 있을까?

statistics

statistics

sta·tis·tics [stə tístiks]

noun

Definition:

branch of mathematics: a branch of mathematics that deals with the analysis and interpretation of numerical data in terms of samples and populations (*takes a singular verb*)

plural noun

Definition:

collection of numerical data: a collection of numerical data (*takes a plural verb*)
- *this month's sales statistics*

[Late 18th century. < German *Statistik*< Latin *status* (see state)]

statistic

statistic

sta·tis·tic [stə tístik] (*plural* sta·tis·tics)

noun

Definition:

1. element of data: a single element of data from a collection

2. numerical value or function: a numerical value or function, e.g. a mean or standard deviation, used to describe a sample or population

3. piece of information: somebody or something treated as a piece of data or information

[Late 19th century. Back-formation <statistics]

- **sta·tis·ti·cal** *adjective*
- **sta·tis·ti·cal·ly** *adverb*

단수로 쓰이는 statistics는 '통계학'이라는 학문(표본과 모집단으로부터 나오는 수치자료를 분

석하고 파악하는 일을 하는 수학의 한 분야)이고, 복수로서 쓰이는 statistics는 '수치자료의 총체'라고 정의하고 있다.

두산백과사전에서는 통계학과 통계량을 다음과 같이 구별하고 있다.

◎ 통계학: 집단현상(集團現象)을 수량적(數量的)으로 관찰하고, 분석하는 방법을 연구하는 학문
◎ 통계량: 표본의 특성을 보이는 특성치(特性値)

통계량(우리는 통상 통계라고 부른다)과 통계학은 직관적으로 어떻게 다른가? 영문으로는 모두 statistics이다. 그러나 통계는 정리되어서 주기적으로 발표되는 일종의 정리된 정보인 반면, 통계학은 자료를 어떻게 체계적으로 정리하고 요약하는가에 대한 학문이다. 통계를 잘 알려면 해당 분야의 내용을 잘 알아야 하지만 통계학을 알 필요는 없다. 그러나 통계학을 잘하려면 수학 등 통계적 방법에 대한 이해가 필요하다. 예를 들어 경제성장률의 의미를 알기 위해서는 국민계정의 작성과정만 알면 되지만, 경제성장률을 정량적으로 예측한다면 작성과정 보다는 예측 관련 통계학을 이해할 필요가 있다.

자, 그러면 이러한 통계학의 활용 예들을 간단히 들어보자. 실험연구의 한 예로서 어느 제약회사에서 혈압을 내리는 고혈압치료제를 새로 개발하였다고 하자. 이 회사에서는 새로운 약품의 효과를 조사해 보기 위하여 500명의 고혈압 환자들에게 일정기간동안 새로운 약을 복용시킨 후 혈압을 측정한 결과, 새로 개발된 치료제가 과거의 치료제보다 평균 20mHg 만큼 혈압을 떨어뜨리는 효과가 있었다고 하자. 여기서 평균 20mHg 만큼 혈압이 강하하였다는 것은 오차를 포함할 수 있는 측정 자료로부터 얻어진 정보이므로 이 효과가 과연 새로 개발된 치료제가 진정으로 약효가 뛰어나서 나타난 필연적인 결과인지, 아니면 약효가 없는데도 불구하고 오차에 의하여 나타난 우연적 결과인지에 대한 명확한 결론을 이끌어 낼 필요가 있는데, 통계학은 이러한 경우에 합리적이고도 과학적인 결론을 유도할 수 있는 방법을 제공한다. 또 다른 예로서, 통계학에서는 수천 명의 표본조사를 통하여 각 정당의 지지도를 여론조사하거나 가구당 월수입을 조사하기 위해서, 어떻게 표본설계를 하여야 지지도나 평균 수입의 참값을 보다 정확하게 추정할 수 있는가를 연구하게 된다. 또한, 과거의 여러 통계자료를 이용하여 앞으로의 경제에 대한 예측을 하는데 여러 가지 통계적 수법이 이용되고 있으며, 기업체에서는 생산제품이나 서비스의 수요예측을 하고 투자 포트폴리오를 작성하기 위하여 통계학 기법들을 적용하며 통계적 품질관리 기법을 실시하여 제품의 품질을 향상시키고, 심리학에서는 적성검사의 결과를 통계적으로 처리하여 좀 더 합리적인 결론을 유도할 수 있다.

이와 같이 통계학은 자료의 수집과정을 설계하고, 수집된 자료를 요약하고 해석하여 불확실한 사실에 대하여 합리적인 결론을 이끌어 내거나 일반화하는 과학적인 원리와 방법론을 제공하여 준다. 오늘날 자료의 수집과 통계적 방법이 이용되는 분야는 일일이 열거할 수 없을 정도로 다양하다. 우리가 일상적으로 접하는 일기예보, 여론조사, 국가통계 등에서 중요한 역할을 담당하고 있을 뿐 아니라, 경영학, 경제학, 자연과학, 사회과학, 공학, 의학 등 거의 모든 학문분야에서 통계학이 활발히 응용되고 있다. 특히 최근에는 컴퓨터의 발달로 대량 자료의 신속한

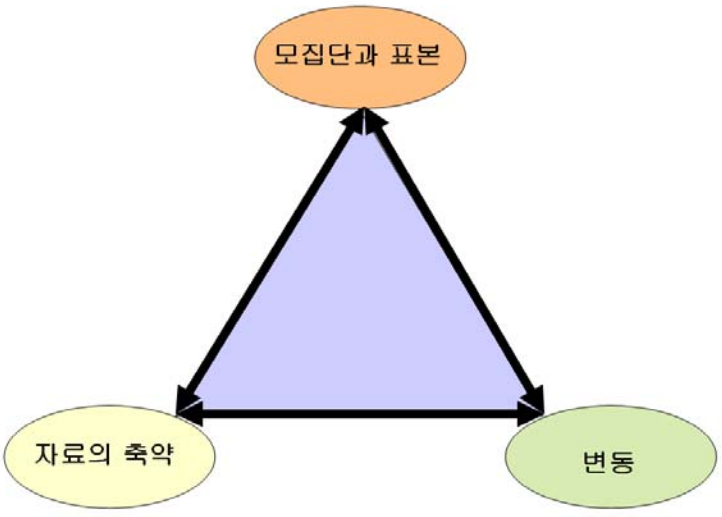
처리가 가능해짐에 따라 통계학의 응용분야와 그 역할이 점점 확대되고 있는 추세이다. 그 예로서 데이터마이닝, 생물정보학, 금융공학, 공간통계학 등에서의 통계학의 역할을 들 수 있다.

통계학의 3가지 주제

통계학은 크게 보면 다음과 같은 3가지 주제를 연구하는 학문이라 할 수 있다(Fisher(1935)).

1. 모집단(population)과 표본(sample)
2. 자료의 축약방법(methods of data reduction)
3. 변동(variation)

이 세 가지 주제는 모든 자료분석 시 동시에 발생하는 문제들이다. 모집단은 변동을 포함하고 있다. 즉 모집단을 구성하고 있는 개체는 모두 다르므로 변동을 품고 있다. 우리는 모집단을 총체적으로 확인할 수 없으므로 표본에 의지하게 된다. 모집단에 대한 정보를 요약하기 위하여 우리는 표본을 축약하여야 한다. 이 세 가지 주제를 하나의 그림으로 표현하면 다음 [그림 1.5]와 같다.



[그림 1.5] 통계학의 3가지 주제

이 세 가지 주제에 대하여 좀 더 자세히 생각하여 보자.

모집단과 표본

앞으로 이 책을 통하여 통계학을 공부하면서 '모집단'과 '표본'이라는 말을 자주 접하게 되는데, 통계학 전반에 관한 정확한 이해를 위해서는 이들에 대한 정의를 올바로 이해하는 것과 이들의 차이를 명확하게 파악하는 것이 중요하다.

많은 현실적인 문제에서 관심의 대상이 되는 모든 자료를 수집한다는 것은 실제로 불가능하거나 불합리한 경우가 많다. 따라서 많은 경우 관심의 대상이 되는 모든 자료를 수집하는 대신 이들을 잘 대표한다고 판단되어지는 일부만을 추출하여 조사하게 되는 경우가 많다. 예를 들어서, 대선에서 특정후보에 대한 지지율을 조사하고자 하는 경우 정확한 지지율의 파악을 위하여서는 전국의 유권자들을 한 사람도 빠짐없이 모두 조사하여야만 정확한 지지율을 파악할 수 있지만 현실적으로 그렇게 하는 것은 거의 불가능함과 동시에 시간이나 비용 면에서 현실성이 없는 조사방법이라고 할 수 있다. 따라서 이 경우 전국에 있는 유권자들을 모두 조사하는 대신에 전국의 유권자를 대표한다고 생각되어지는 그 일부만을 추출하여 지지율을 추측하는 것이 일반적이다. 또한, 어떤 형광등을 생산하는 회사에서 생산되는 형광등의 평균 수명을 조사하여 이를 소비자에게 공개함으로써 타 회사에 비해서 자 회사의 형광등의 품질이 뛰어남을 홍보하고 싶다고 하자. 이 경우 정확한 형광등의 평균 수명을 알기 위해서는 생산되는 모든 형광등을 모두 조사하여 고장 날 때까지의 시간을 측정하고 이들의 평균을 산출해야겠지만 이렇게 하는 것이 현실적으로 불가능하므로 생산되는 형광등 전체를 잘 대표한다고 판단되어지는 일부 형광등을 추출하여 이들의 평균 수명을 조사함으로써 전체 형광등의 평균 수명을 추측하게 된다. 이들 예에서 보듯이 연구대상이 되는 가능한 관측값이나 측정값의 집합을 모집단(population)이라 하며, 통계적 처리를 위하여 모집단에서 실제로 추출한 관측 값이나 측정값의 집합을 표본(sample)이라 한다. 앞의 예에서는 전국의 유권자 전체의 자료가 모집단이 되며, 전국의 유권자를 잘 대표하도록 추출된 일부 유권자의 자료가 표본이 된다. 뒤의 예에서는 생산되는 모든 전구 전체 자료의 집합이 모집단이 되며, 전구 전체를 잘 대표하도록 추출된 일부 전구의 자료가 표본이 된다. 모집단과 표본에 대한 더 자세한 이야기는 2장에서 언급하도록 하자.

모집단 전체를 조사하는 것을 전수조사(census)라고 하는 데 시간이나 비용 같은 문제들 때문에 주로 국가기관에서 실시한다. 성경 민수기(Numbers)에 보면 B. C. 15000년경 이스라엘 민족이 이집트를 탈출(Exodus)하여 나라를 세우기 위하여 팔레스타인으로 들어가기 전에 두 번에 걸쳐 인구조사를 행하는 장면이 나오고 성경 누가복음(Luke)에 보면 예수의 탄생이야기를 통하여 B. C. 5년 로마황제 아우구스투스에 의한 인구조사가 언급되어 있다. 그러나 세계적으로 보면 우리가 알고 있는 정규적인 인구조사는 17세기에나 가서야 비로소 시작되었다. 우리나라 통계청에서는 현재 인구 및 주택센서스를 5년마다 한 번씩 시행하고 있다. 반면에 표본을 이용하여 모집단의 모습을 살피는 조사를 표본조사(sampling survey)라고 한다. 대중매체가 많이 실시하는 여론조사(opinion survey)나 기업체에서 많이 행하는 시장조사(market research, marketing survey)가 표본조사의 대표적인 예이다.

그러면 표본조사가 왜 타당성이 있을까? 즉 모집단의 일부분인 표본을 이용하여 표본의 정

보를 얻은 후 이를 확대하여 모집단의 정보로 여기는 것이 타당성이 있느냐는 것이다. 이 의문에 답하는 것이 아래와 같은 Glinko-Cantelli 정리이다. 이 정리를 혹자는 통계학에서의 근본정리(fundamental theorem in statistics)라 하여 9.4절에서 언급할 중심극한정리(central limit theorem)만큼 중요하게 언급하기도 한다.

'모집단에서 iid(independent and identically distributed) 성질을 만족하게 표본을 뽑으면 경험적 누적분포함수(empirical cumulative distribution function)는 표본의 수가 커짐에 따라 점점 이론적 누적분포함수(theoretical cumulative distribution function)에 다가간다.'

이 정리를 쉽게 풀어 쓰면 다음과 같다.

'모집단에서 표본을 골고루 섞어 잘 뽑으면 표본의 수가 커짐에 따라 점점 표본은 모집단을 닮아간다.'

2002년 대통령 선거 개표를 앞두고 방송 3사(KBS, MBC, SBS)가 실시한 출구조사 결과 노무현 후보의 당선을 다음과 같이 정확히 예측하였다. 표본만 잘 뽑으면 모집단에 대한 예측이 정확함을 확인하는 한 예가 될 수 있다.

방송사	출구조사		개표결과	
	노무현	이회창	노무현	이회창
KBS	49.1%	46.8%	48.92%	46.59%
MBC	48.4%	46.8%		
SBS	48.2%	46.7%		

2007년에도 대통령 선거 개표를 앞두고 방송 3사(KBS, MBC, SBS)가 실시한 출구조사 결과 이명박 후보의 당선을 다음과 같이 비교적 정확히 예측하였다. MBC(코리아 리서치)와 KBS(미디어 리서치)는 전국 250개 투표구에서 유권자 7만 명에게 출구조사를 했고 SBS는 여론조사 회사인 TNS 코리아와 합동으로 19일 오전 6시부터 오후 6시까지 전국 233개 투표소에서 투표를 마친 유권자 10만여 명을 대상으로 단독 출구조사를 했다. 총투표자수는 23,732,854명이었으니 KBS와 MBC 공동출구조사는 총투표자수의 0.3%, SBS 출구조사는 총투표자수의 0.4%를 표본조사한 셈이다.

방송사	출구조사		개표결과	
	이명박	정동영	이명박	정동영
KBS & MBC	50.3%	26.0%	48.42%	26.02%
SBS	51.3%	25.0%		

예제 1.5 세계 최대의 독립 PR컨설팅사인 에델만의 한국지사, 에델만코리아(www.edelman.co.kr, 사장 김원규)는 2007년 9월 12일 '2007 한국 블로거 성향 조사'라는 이름의 보고서를 발표했다. 에델만 코리아가 한국과학기술원(KAIST) 바이오 및 뇌공학과 정재승 교수팀과 공동으로 진행한 이번 조사는 2006년 12월부터 2007년 2월까지 총 59일 동안 온라인 설문 방식으로 실시됐으며, 블로그와 미니홈피를 운영하고 있는 국내 블로거 총 347명이 참여했다. 조사의 표본오차는 95% 신뢰수준에서 ±4.3% 포인트다.

여기서 모집단은 우리나라 블로거 전체이고 조사에 참여한 우리나라 블로거 347명이 표본이 된다. ■

자료의 축약방법

통계학에서 하는 큰 작업 중의 하나가 '자료의 축약'이다. 통계학에서의 많은 방법들이 '자료를 어떻게 줄이느냐?'하는 문제와 관련이 있다. 한 예로 히스토그램을 그리는 경우를 살펴보자. 100개의 자료가 얻어졌다고 하자. 이 자료가 갖고 있는 구조(분포)를 어떻게 볼 수 있을까? 우리는 이 100개의 자료를 이용하여 10개의 계급구간을 갖는 도수분포표(frequency table)를 계산할 수 있고 이를 이용하여 10개의 기둥을 갖는 히스토그램을 그릴 수 있다. 이 히스토그램을 보면서 우리는 자료의 구조를 보게 된다. 이 자료의 구조를 보게 되는 과정이 바로 '자료의 축약'이다. 100개의 자료라는 정보에서 10개의 기둥이라는 정보로 자료가 축약되었다. 이 때 우리가 명심하여야 할 것은 우리가 단순히 100개의 자료에서 90개를 버리고 나머지 10개의 자료를 선택한 것이 아니고 100개의 자료를 대상으로 최소의 비용(정보유실)으로 최대효과(자료의 구조를 알아냄)를 얻기 위하여 100개의 정보를 축약시키는 것이다. 이 과정에서 당연히 정보가 유실되게 된다. 즉 우리는 도수분포표나 히스토그램만을 보고 거꾸로 원 자료를 복구할 수는 없게 된다. 그러나 우리는 히스토그램을 그리는 과정을 통하여 정보가 유실되는 비용을 치르더라도 자료의 구조를 알아내는 목적을 달성할 수 있게 되는 것이다. 더 나아가 우리는 히스토그램을 이용하여 산술평균과 분산(또는 표준편차)을 구하게 되는 데 이 두 개의 값은 자료의 중심경향과 자료의 흩어진 정도를 나타내는 수치적 측도들이다. 100개의 자료에서 10개의 기둥으로, 다시 2개의 수치 값으로 자료의 축약이 이루어지는 것이다. 여러 집단을 비교할 때 히스토그램을 서로 비교하는 것도 좋지만 각 집단에 대한 산술평균과 분산을 비교하는 것이 더 용이할 때가 많다.

예제 1.6 다음 사진은 미국 국립공원인 Yellow Stone Park 내에 있는 간헐온천(Old Faithful geyser)을 보여주고 있다. 이러한 간헐천에서 과학자들의 관심을 끈 것은 온천물이 나오는 지속시간(duration)과 온천물이 분출한 후 다음 온천물이 쏟아질 때까지의 간격시간(interval)이다.

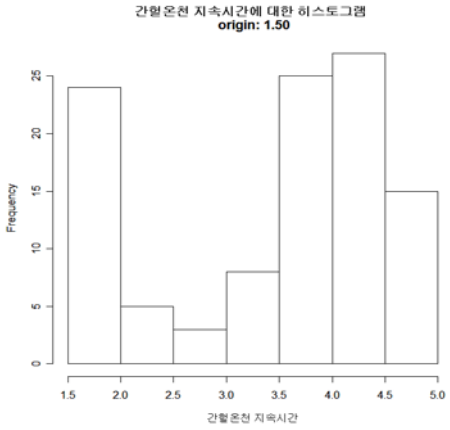
다음 자료는 온천물의 지속시간을 분 단위로 잰 자료(107개)이다.

4.37, 3.87, 4.00, 4.03, 3.50, 4.08, 2.25, 4.70, 1.73, 4.93, 1.73, 4.62, 3.43, 4.25, 1.68, 3.92, 3.68, 3.10, 4.03, 1.77, 4.08, 1.75, 3.20, 1.85, 4.62, 1.97, 4.50, 3.92, 4.35, 2.33, 3.83, 1.88, 4.60, 1.80, 4.73, 1.77, 4.57, 1.85, 3.52, 4.00, 3.70, 3.72, 4.25, 3.58, 3.80, 3.77, 3.75, 2.50, 4.50, 4.10, 3.70, 3.80, 3.43, 4.00, 2.27, 4.40, 4.05, 4.25, 3.33, 2.00, 4.33, 2.93, 4.58, 1.90, 3.58, 3.73, 3.73, 1.82, 4.63, 3.50, 4.00, 3.67, 1.67, 4.60, 1.67, 4.00, 1.80, 4.42, 1.90, 4.63, 2.93, 3.50, 1.97, 4.28, 1.83, 4.13, 1.83, 4.65, 4.20, 3.93, 4.33, 1.83, 4.53, 2.03, 4.18, 4.43, 4.07, 4.13, 3.95, 4.10, 2.72, 4.58, 1.90, 4.50, 1.95, 4.83, 4.12

이 자료를 다음과 같이 정렬하여 보자. 아직 이 자료의 분포를 알기가 어렵다.

1.67 1.67 1.68 1.73 1.73 1.75 1.77 1.77 1.80 1.80 1.82 1.83 1.83 1.83 1.85 1.85 1.88 1.90 1.90 1.90 1.95 1.97 1.97 2.00 2.03 2.25 2.27 2.33 2.50 2.72 2.93 2.93 3.10 3.20 3.33 3.43 3.43 3.50 3.50 3.50 3.52 3.58 3.58 3.67 3.68 3.70 3.70 3.72 3.73 3.73 3.75 3.77 3.80 3.80 3.83 3.87 3.92 3.92 3.93 3.95 4.00 4.00 4.00 4.00 4.00 4.03 4.03 4.05 4.07 4.08 4.08 4.10 4.10 4.12 4.13 4.13 4.18 4.20 4.25 4.25 4.25 4.28 4.33 4.33 4.35 4.37 4.40 4.42 4.43 4.50 4.50 4.50 4.53 4.57 4.58 4.58 4.60 4.60 4.62 4.62 4.63 4.63 4.65 4.70 4.73 4.83 4.93

위의 자료를 이용하여 도수분포표를 작성한 후 히스토그램을 그리면 다음과 같다. 자료의 분포를 알 수 있다. 봉우리가 두 개인 쌍봉분포이다. 107개의 자료를 축약하여 7개의 기둥을 만드니 자료의 구조가 보이는 것이다. 더 나아가 이 히스토그램을 이용하여 산술평균과 분산을 구하니 각각 3.432와 1.151이었다. 이 두 개의 수치적 측도 중 산술평균은 자료의 중심경향을 알 수 있는 수치적 측도로서 쓰이고 분산은 자료의 흩어진 정도를 알 수 있는 수치적 측도로서 쓰이게 된다. 위의 자료에서는 자료의 중심경향을 알 수 있는 수치적 측도로서 최빈값(4.25)도 동시에 제시하는 것이 좋다. 우리는 107개의 자료를 축약하여 7개의 기둥을 만들어 자료의 구조를 파악하고 다시 2개의 수치적 측도(산술평균과 분산)로 자료를 축약하여 다른 집단 간의 비교(comparison)에 이 수치들을 이용하게 되는 것이다. ■



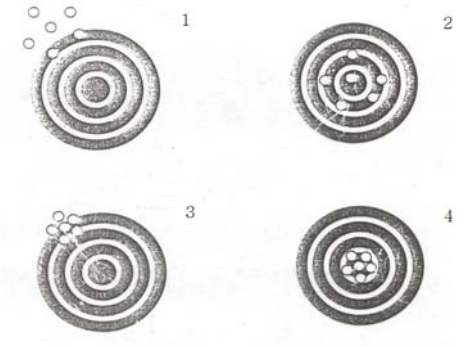
변동

어느 통계학자는 통계학을 다른 학문과 비교하며 비유하기를 다음과 같이 '통계학이란 한 마디로 변동에 대하여 연구하는 학문이다.'라고 하였다.

> Economics is about...*Money*(*and why it is good*).
> Psychology: *Why we think what we think*(*we think*).
> Biology: *Life*.
> Anthropology: *Who*?
> History: *What, where, and when*?
> Philosophy: *Why*?
> Engineering: *How*?
> Accounting: *How much*?
> In such a caricature, Statistics is about...*Variation*.

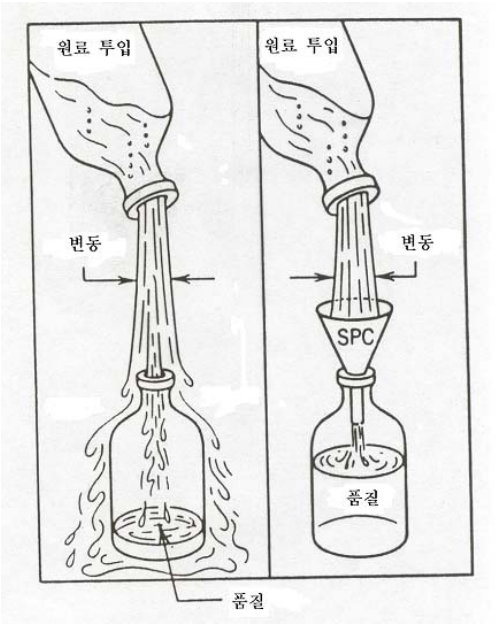
'변동'이라는 문제는 현대생활에서 아주 중요한 통계학의 주제가 되고 있다. 기업체들이 자주 언급하는 식스시그마(6σ) 운동도 변동에 초점을 맞춘 개념이다. 예로 키가 모두 큰 부모 하에서 주로 키 큰 자녀들이 나오지만 키가 작은 자녀도 나타나고 키가 모두 작은 부모 밑에서 주로 키가 작은 자녀들이 나오지만 키가 큰 자녀도 나타나는 현상도 변동의 문제인 것이다. 다음 [그림 1.6]을 보자. 사격장에서 6발 사격을 한 후 총알의 흔적을 나타내는 그림이다. 2번과 4번을 보면 총알의 흔적의 중심과 동심원들의 중심이 일치한다. 이러한 경우를 품질공학에서는 정확하다(accurate)고 하고 통계학에서는 비편향하다(unbiased)고 표현한다. 반면 3번과 4번을 보면 총알의 흔적이 조밀하게 모여 있음을 알 수 있다. 이러한 경우를 품질공학에서는 정밀하다(precise)고 하고 통계학에서는 유효하다(efficient)고, 또는 변동이 적다고 표현한다. 그러면 2번과 3번 중 4번으로 개선하기가 수월한 것은 어느 것인가? 2번에서 4번으로 바꾸는 것보다 3번에서 4번으로 바꾸는 것이 훨씬 수월함을 경험상 알 수 있다. 이러한 현상을 제품의 품질에 적용하여 보면 제품 품질의 변동이 크면 설혹 제품품질의 중심이 목표값과 같더라도 제품품질의 변동을 줄이기가 어려우나 제품 품질의 변동이 적으면 제품품질의 중심을 목표값으로

이동하기가 훨씬 수월하다는 것이다. 제조업체 중심으로 많이 거론되는 식스시그마 운동의 핵심도 제품의 '변동'에 있다고 하겠다.



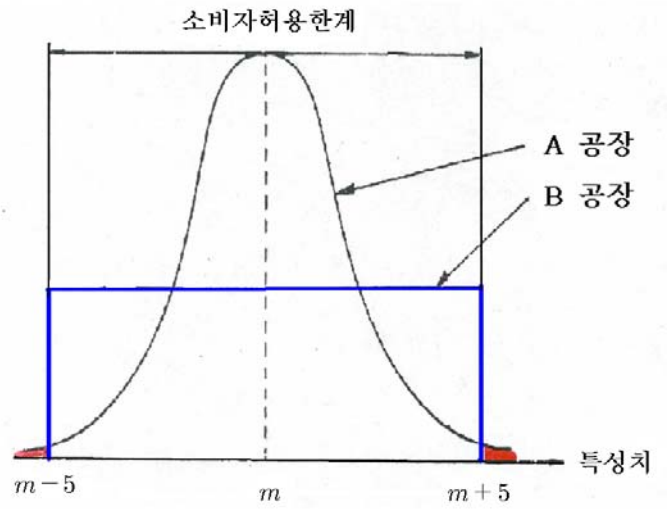
[그림 1.6] 정확성과 정밀성

다음 [그림 1.7]을 보자. 마치 주둥이가 큰 병에 담겨 있는 물을 주둥이가 작은 병으로 옮길 때 깔때기(funnel)를 이용하면 전혀 물을 흘리지 않고 옮길 수 있다는 비유이다. 이를 통하여 제조업체에서 원료로 제품을 만들 때 통계적 공정관리(statistical process control)를 통하여 제품의 변동을 줄임으로써 제품의 품질을 보장하여 생산성을 높일 수 있음을 알 수 있다.



[그림 1.7] 통계적 공정관리에 의한 변동의 감소

예제 1.7 다음 [그림 1.8]은 어느 텔레비전 제조회사의 두 개의 공장(A, B)에서 생산하는 텔레비전 색상밀도(color density) 품질의 분포곡선을 비교한 그림이다. m은 색상밀도에 대한 목표값이다. 색상밀도가 $m-5$보다 작거나 $m+5$보다 크면 불량품이 된다. A공장은 평균이 m이고 표준편차가 $\frac{5}{3}$인 정규분포를 이루고 B공장은 평균이 m이고 표준편차가 $\frac{5}{\sqrt{3}}$인 균등분포를 이룬다. A공장은 불량률이 0.25%임에 반해 B공장은 불량률이 0%이다. 어느 공장이 더 좋은 공장이라 할 수가 있을까? 제품의 품질을 정의할 때 제품이 소비자에 끼치는 총이익이 많을수록 품질이 좋다는 서양인들의 시각과 달리 일본인들은 제품의 품질을 '제품이 출하된 시점으로부터 성능특성치의 변동과 부작용 등으로 인하여 사회에 끼친 총손실(total loss)'로 정의하였다. 제품이 소비자에게 끼치는 총손실이 적은 제품일수록 좋은 제품이라는 것이다. 손실함수를 2차식으로 정의하고 이 손실을 계산하여 보면 A공장의 기대손실이 1.11인 반면 B공장의 기대손실은 3.33이 된다. A공장은 불량률이 0.25%이나 소비자에게 끼치는 손실의 측면에서는 B공장보다 1/3배가 된다는 것이다. 이러한 결과가 나오는 것은 A공장의 변동이 B공장의 변동보다 작다는 사실에서 기인한다. ■



[그림 1.8] 텔레비전 색상밀도 분포

통계학의 유형

통계학을 크게 두 가지 유형으로 구분하면 기술통계학(descriptive statistics)과 추측통계학(inferential statistics)으로 구분할 수 있다.

기술통계학(Descriptive Statistics)

자료에서의 측정값은 조사자에게 중요한 정보이지만, 자료의 양이 방대한 경우에는 이의 전반적인 내용을 한 눈에 파악하기가 쉽지 않다. 따라서 대표값, 변동의 크기, 분포의 형태 등을 요약해 놓으면 방대한 자료집합의 특징을 쉽게 알아볼 수 있다. 이와 같이 자료를 수집하고 정리하여 도표나 표를 만들거나 자료를 요약하여 대표값이나 변동의 크기 등을 구하는 방법을 다루는 분야를 기술통계학이라 한다.

추측통계학(Inferential Statistics)

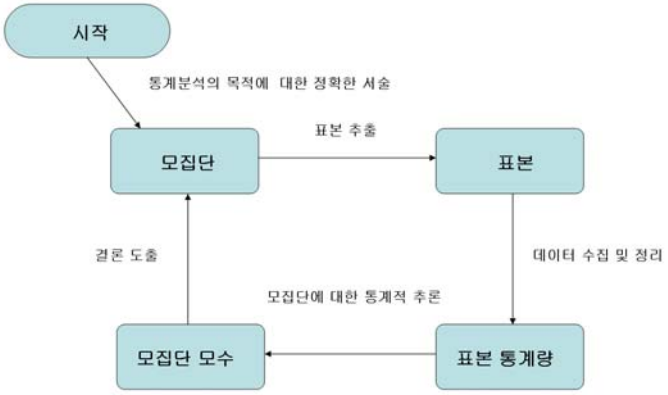
추측통계학은 자료에 내포되어있는 정보를 분석하여 불확실에 대한 추론을 다루는 분야이다. 즉, 추측통계학에서는 통계적 모형을 설정하고, 또한 설정된 모형이 합리적인지 평가하며, 자료로부터 얻어지는 정보를 근거로 하여 미지의 특성에 대한 결론을 내리고 미래에 일어날 현상에 대한 예측을 하게 된다. 추측통계학은 모집단으로부터 얻은 표본자료를 사용하여 모집단 전체에 대한 특징을 추측해 내고, 모집단에 대한 일련의 의사결정방법을 연구하는 분야로 현대통계학에서 핵심이 되는 분야이다.

기술통계학과 추측통계학을 요약하면 다음과 같은 [표 1.2]로 정리할 수 있다.

유 형	특 징
기술통계학	• 자료의 수집(collection), 정리, 요약(summary) • 자료의 전반적인 구조(overall structure)
추측통계학	• 통계적 모형(statistical model) • 모수(parameter)에 대한 통계적 추론(statistical inference) • 통계적 추측과 예측(statistical prediction and statistical forecasting)

[표 1.2] 기술통계학과 추측통계학

다음 [그림 1.9]는 통계분석의 과정을 나타내는 그림으로서 이 그림을 통하여 기술통계학과 추측통계학의 관계를 알 수 있다.



[그림 1.9] 통계분석 과정

통계자료분석에서 기술통계학과 추측통계학은 다음 [그림 1.10]과 같이 서로 보완적인 관계에 있고 자료분석의 사이클을 이룬다. 즉 자료분석의 초기 단계에서는 자료의 수집, 정리, 요약을 통하여 자료의 구조를 살펴보고 이를 통하여 적절한 통계적 모형을 설정하고 모수에 대한 추론 및 예측을 행하게 된다. 통계적 모형의 타당성을 검증하기 위하여 새로운 자료를 수집하게 되고 우리는 다시 자료분석의 사이클을 돌리게 된다.

[그림 1.10] 기술통계학과 추측통계학의 관계

1.2 통계학의 필요성

통계학의 응용분야는 매우 다양하다. 통계학이 데이터의 지식화를 위한 과학이므로 데이터가 있는 모든 학문분야에 통계학이 쓰인다. 통계학의 응용분야는 다음과 같은 것들이 있다 (Wikipedia 참조).

국가통계(Government statistics)
보험수리학(Actuarial science)
응용정보경제학(Applied information economics)
경영통계학(Business statistics)

데이터마이닝(Data Mining(applying statistics and pattern recognition to discover knowledge from data))
수리경제학(Economic statistics(Econometrics))
에너지통계학(Energy statistics)
환경통계학(Environment statistics)
생물통계학(Biostatistics)
생물정보학(Bioinformatics)
면역학(Epidemiology)
지리학, 지리정보시스템, 공간통계학(Geography and geographic information systems, more specifically in spatial analysis)
인구학(Demography)
심리통계학, 계량심리학(Psychological statistics)
사회통계학(Social statistics)
범죄통계학(Crime statistics)
공업통계학(Engineering statistics)
품질경영, 품질관리(Quality management, Quality control)
공정분석과 수리화학(Process analysis and chemometrics(for analysis of data from analytical chemistry and chemical engineering))
신뢰도공학/생존분석(Reliability engineering/Survival analysis)
화상처리(Image processing)
통계교육(Statistics education)과 통계해독력(Statistical literacy)
표본론, 여론조사, 시장조사(Statistical surveys)
스포츠통계학(Statistics in various sports)

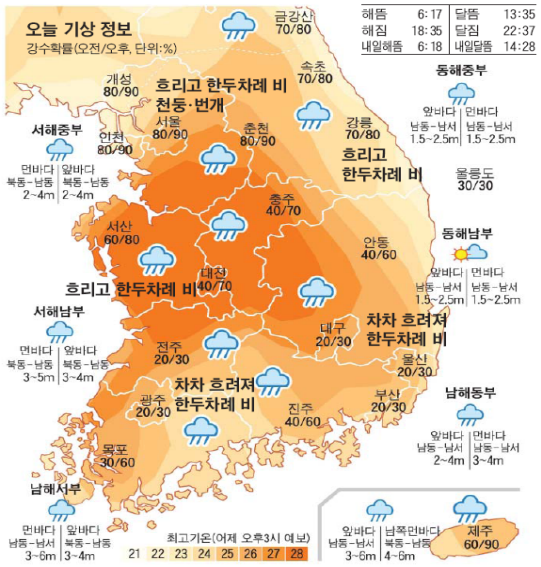
우리 생활에서 느끼는 통계학의 필요성에 대하여 다음과 같은 두 가지 예를 들어보자.

예제 1.8 e-shopping mall(인터넷쇼핑몰)이나 e-market place(온라인전자장터)의 예를 보자. 요즈음 인터넷쇼핑몰이나 온라인전자장터에서 물품을 구입하는 소비자들이 점점 늘고 있다. 이러한 인터넷쇼핑몰에서는 고객의 거래가 발생할 때마다 고객의 정보, 거래, 판매량 등의 자료가 회사의 DB에 기록되게 된다. 이러한 정보를 통하여 고객의 취향을 파악함으로써 고객의 미래구매를 예측하게 되고 각각의 고객의 구매를 일으킬 만한 맞춤형 광고를 고객들의 컴퓨터로 보내게 되는 것이다. 이러한 일련의 작업들은 '온라인 고객관계관리(eCRM)'를 통하여 행해지게 된다. 예로 인터넷쇼핑몰 음악 CD store에서 고객의 거래가 발생하면 다음과 같은 자료가 만들어진다.

이름	성별	나이	직업	구입 일자	상품코드	아티스트

자료의 가치는 그 자료의 context에 있는데 자료의 중요한 정보로서는 5W(who, what, when, where, why), 1H(how)가 있다. 이러한 5W 1H에 답하기를 통하여 자료의 가치에 대한 context를 제공한다. 자료를 분석하기 위하여서는 최소한 who와 what을 알아야 한다. 이러한 자료들이 일정한 기간 모이면 자료를 요약하게 되고 고객의 미래구매를 예측하게 된다. 이러한 과정에서 다양한 통계학 기법들이 쓰이게 된다. ■

예제 1.9 일간 신문에 매일 등장하는 '비올 확률'에 대하여 알아보자. 2007년 9월 태풍 '나리'로 인하여 제주도에 쏟아진 엄청난 양의 폭우로 제주도에 엄청난 재난을 가져다주었다. 기상청의 일기예보는 우리 생활에 점점 중요성을 더하고 있다. 다음 [그림 1.11]은 2007.09.19 조선일보에서 제공한 일기예보이다. 예로 '서울' 밑에 80/90이라 적혀 있다. 서울지역에서 비올 확률이 오전에는 80%, 오후에는 90%라는 말이다. 반면에 부산지역에서는 비올 확률이 오전에는 20%, 오후에는 30%이다.



(제공처: 조선일보)

[그림 1.11] 일기예보

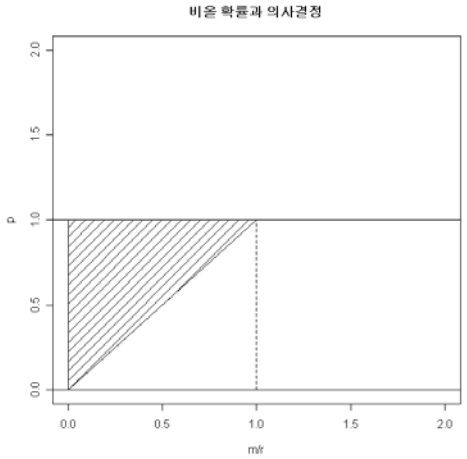
Rao(2003)는 통계학을 정의하며 '불확실성(uncertainty)의 모형화', '불확실성 길들이기'라는 표현들을 사용하고 있다. 불확실성을 내포한 지식에 불확실성의 계량화를 시행하면 사용가능한 지식이 된다. Rao(2003)의 설명을 다시 음미하여 보자. 비올 확률이 50%이면 집을 나설 때 우산을 가지고 가야 하나 말아야 하나? 이를 해결하기 위해서 우산을 가지고 외출해서 겪게 되는 불편이 m원이고 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실을 r원이라 하자. 비올 확률이 p라 할 때 의사결정시의 예상손실을 정리하면 다음 표와 같다.

의사결정	예상손실
우산을 가져간다.	m
우산을 가져가지 않는다.	$p\times r+(1-p)\times 0=pr$

그러면 우리는 다음 표와 같이 항상 손실을 최소화하는 방향으로 의사결정을 한다.

비교	의사결정
$m \le pr \Leftrightarrow \frac{m}{r} \le p$	우산을 가져간다.
$m > pr \Leftrightarrow \frac{m}{r} > p$	우산을 가져가지 않는다.

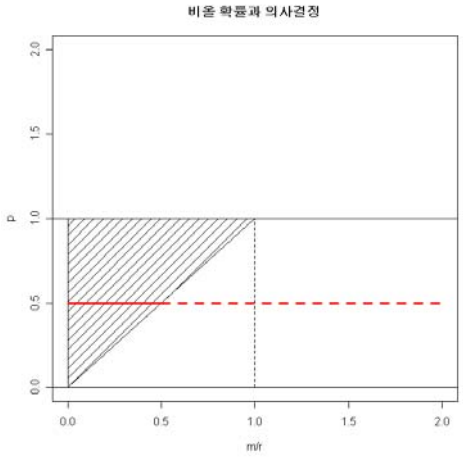
이 두 가지 경우를 다음과 같이 [그림 1.12]로 나타내어 보자. 이 그림에서 빗금친 삼각형 부분이 '우산을 가져간다.'이고 흰색 부분이 '우산을 가져가지 않는다.'이다.



[그림 1.12] 의사결정

극단적인 경우인 $\frac{m}{r} > 1$이면 즉 $m > r$이면 항상 우산을 가져가지 않는다. 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실보다 우산을 가지고 외출해서 겪게 되는 불편을 더 크게 생각하는 사람에게는 비올 확률이 크든 작든 항상 우산을 가져가지 않으므로 비올 확률이 유용한 정보가 되지 못한다. 그러나 보통의 사람들에게는 우산을 가지고 외출해서 겪게 되는 불편보다 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 더 크다. 앞의 질문인 '비올 확률이 50%이면 집을 나설 때 우산을 가지고 가야 하나 말아야 하나?'에 대하여 다시 생각하여 보자. $0 < \frac{m}{r} < \frac{1}{2}$라고 여기는 사람, 즉 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 우산을 가지고 외출해서 겪게 되는 불편의 2배 이상이 된다고 여기는 사람은 우산

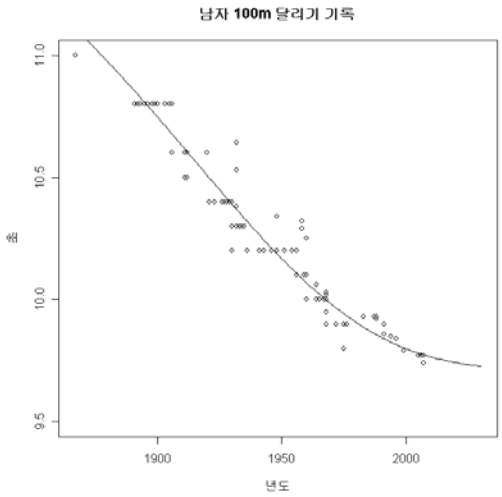
을 가지고 가게 된다. 이것을 다음의 [그림 1.13]에서 빨간 직선으로 표시하였다. $\frac{1}{2} < \frac{m}{r} < 1$ 라고 여기는 사람, 즉 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 우산을 가지고 외출해서 겪게 되는 불편의 2배 이하가 된다고 여기는 사람은 우산을 가져가지 않게 된다. 이것을 다음 그림에서 빨간 점선으로 표시하였다. 이처럼 비올 확률이라는 정보가 우리들이 합리적인 의사 결정을 하는 데 중요한 역할을 하게 된다. 불확실성을 내포한 지식에 불확실성의 계량화를 시행함으로써 사용가능한 지식으로 바꿀 수가 있게 되는 것이다. ■



[그림 1.13] 비올 확률 50%에 대한 해석

통계학의 응용분야는 매우 다양하다고 앞에서 언급하였다. 양적 정보가 있는 곳이면 항상 통계학이 쓰인다고 하겠다. 다음 예제를 통하여 스포츠영역에서 통계학이 어떻게 쓰이는 지를 간단히 알아보자.

예제 1.10 다음 [그림 1.14]는 1867년부터 2007년까지 남자 100m 달리기 기록을 점으로 나타내고 이 정보를 이용하여 Gompertz곡선을 구한 후 2030년까지 연장시켜 본 그림이다. Gompertz곡선$(9.712 + 2.328e^{-e^{0.014(x-1915)}})$이 어떤 수렴값으로 수렴하는 경향이 있음을 알 수 있다. 이 수렴값이 9.728초이므로 앞으로 남자 100m 달리기 기록은 계속 갱신은 되겠으나 대략 9.728초 근방에 머물 것이라고 우리는 짐작할 수 있다. ■



[그림 1.14] 남자 100m 달리기 기록을 나타내는 산점도와 Gompertz곡선

1.3 통계문맹

수문맹(innumeracy)이란 numerical illiteracy에 대한 합성어로서 숫자나 수학적 개념에 대한 사고 능력이 부족하거나 없는 현상을 가리킨다. 이 수문맹이라는 단어는 인지과학자인 Douglas Hofstadter에 의하여 만들어졌고 수학자인 John Allen Paulos가 1989년에 쓴 책 'Innumeracy: Mathematical Illiteracy and its Consequences'에서 언급되며 대중화되었다. 이러한 수문맹과 밀접한 관련을 갖고 있는 것이 통계문맹이다.

통계문맹(statistical illiteracy)이란 자료를 다루는 능력이나 통계 개념에 대한 사고 능력이 부족하거나 없는 현상을 가리킨다. 통계해독력(statistical literacy)은 정보화시대에 신문, 텔레비전, 인터넷 같은 대중매체의 기사내용을 이해하는데 꼭 필요한, 도시민들이 갖추어야 할 능력으로서 수해독력(numerical literacy)의 핵심요소가 된다. 통계해독력은

> 통계적 정보나 메시지에 대하여 해석하고
> 비판적으로 평가하고 의사소통할 수 있는 능력

을 말한다. Smith(2002)는 통계해독력을 다음 작업과 관련된 능력이라고 정의하였다.

- 통계자료를 이해하고 해석하기
- 통계정보와 데이터관련 논의를 비판적으로 평가하기
- 매일의 생활에서 정보를 사용하기
- 데이터에 대한 서로의 반응에 대하여 토의하고 정보교환하기

Gal(2002)은 통계해독력 모형을 다음 [표 1.3]과 같이 나타내었다.

<table>
<tr><td rowspan="7">통계해독력: 다양한 환경 하에서 효과적인 데이터프로슈머 (data prosumer)로 활동할 능력</td><td rowspan="5">지식 요소</td><td>해독력 기술</td><td rowspan="5">• 다양한 통계정보, 데이터와 관련된 논의, 확률론적 현상을 해석하고 비판적으로 평가하는 능력
• 통계적 메시지를 이해하고, 해석하고, 비판적으로 평가하고, 반응하는 능력</td></tr>
<tr><td>통계적 지식</td></tr>
<tr><td>수학적 지식</td></tr>
<tr><td>문맥 지식</td></tr>
<tr><td>비판적 질문</td></tr>
<tr><td rowspan="2">성향 요소</td><td>믿음과 태도</td><td rowspan="2">• 통계정보에 대한 각자의 반응(통계정보의 의미 파악, 통계정보의 함축에 대한 의견, 주어진 결론을 용인하는 것에 대한 관심)을 토의하고 의사소통하는 능력</td></tr>
<tr><td>비판적 자세</td></tr>
</table>

[표 1.3] 통계해독력 모형

통계문맹에 대하여 한 예를 들어보자. 우리는 각종 신문사나 방송사에서 실시하는 대선관련 여론조사 결과를 수시로 대중매체에서 만나고 있다. 다음 예는 주간조선(1970호, 2007년 9월3일 발행) 대선관련 여론조사 기사내용에서 발췌한 한 부분이다.

'2002년 대선에서는 동서(東西)로 갈리는 지역변수와 아울러, 30대 이하와 50대 이상의 정당 선호가 뚜렷하게 갈리는 세대변수도 큰 영향을 미쳤다. 하지만 최근 여론조사에서는 올해 처음으로 대통령선거의 투표권을 행사하게 되는 정치 신인류(新人類) 1924세대(대선에 처음 참여하는 19~24세 유권자)의 보수화 경향이 뚜렷해지면서 세대대결 양상에 변화 조짐이 보이고 있다.
Weekly Chosun은 여론조사 전문기관인 한국리서치에 의뢰해 지난 8월 22일 전국의 19~24세 500명을 대상으로 이들의 정치·사회 의식을 측정하는 전화 여론조사를 실시했다.(표본오차 : 95% 신뢰수준에서 ±4.4%포인트) 연말 대선을 앞두고 1924세대의 표심을 파악하기 위해 이들만을 대상으로 정밀한 여론조사를 실시한 것은 국내 언론에서 이번이 처음이다.'

이 기사 내용은 여론조사 방법과 결과를 어느 정도까지 신뢰할 수 있는 지를 나타내고 있다. 이 기사내용에서 중요한 통계개념은 다음과 같다.

1. '표본오차'란 무엇인가?
2. '95% 신뢰수준'이란 무엇을 의미하는가?
3. '95% 신뢰수준에서 ±4.4%포인트'란 무엇을 의미하는가?
4. '%포인트'란 무엇인가?

이러한 4가지 통계개념을 모르면 위의 기사내용을 상당부분 이해하지 못하게 된다. 즉 통계

문맹이 일어나게 된다. 이러한 통계문명은 우리의 일상생활에서 수시로 일어나고 통계문맹으로 인한 개인적, 집단적 피해도 크다고 하겠다. 이러한 통계개념을 우리는 10.4절에서 배우게 될 것이다.

예제 1.11 미국은 영토가 너무 넓어 중요한 장거리 교통수단으로 미국인들은 비행기를 선호한다. 2007년은 특히 비행기 활주로대기, 비행기 이착륙지연, 초만원탑승 등으로 얼룩진 한 해였다. 2007년 12월 6일에서 9일까지 시행한 미국갤럽조사(www.gallup.com/poll/103237/Airline-Satisfaction-Remains-High.aspx)에서 지난 1년 동안 비행기를 타본 경험이 있는 473명에게 행한 '비행기 서비스에 대하여 전반적으로 찬성하느냐?'라는 질문에 대한 결과를 1999년과 2000년과 비교하여 아래와 같이 표로 정리하였다.

년도	만족함	만족하지 않음	무응답
2007	72	24	3
2000	69	29	2
1999	65	32	3

(단위: %, 표본오차: 95% 신뢰수준에서 ±5%포인트)

이 예에서 우리가 알아야 할 중요한 통계개념 및 알고 싶은 내용은 다음과 같다.

1. '표본오차'란 무엇인가?
2. '95% 신뢰수준'이란 무엇을 의미하는가?
3. '95% 신뢰수준에서 ±5%포인트'란 무엇을 의미하는가?
4. '%포인트'란 무엇인가?
5. 2007년의 만족도는 1999년이나 2000년과 비교할 때 차이가 없다고 할 수 있나? ■

우리는 통계문맹을 없애기 위하여 우리는 통계적 사고(statistical thinking)를 함양하여야 한다. 품질공학(Quality Engineering)에서 통계적 사고란 다음과 같은 근본 원리에 기초하여 학습하거나 행동하는 철학을 말한다.

1. 우리들이 만드는 모든 일들은 상호관련이 있는 처리과정 시스템 하에서 일어난다.
2. 모든 처리 과정에는 변동이 존재한다.
3. 변동을 이해하고 줄이는 노력이 성공의 지름길이다.

통계적 사고는 모든 일의 처리 과정을 전체적으로 바라볼 수 있는 수단을 제공한다.

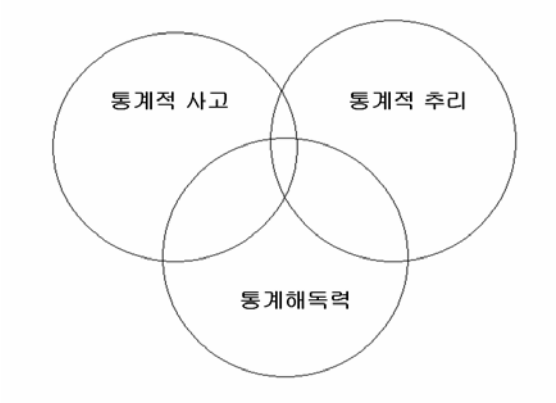
우리는 종종 통계해독력, 통계적 추리(statistical reasoning), 통계적 사고를 다음 [표 1.4]와 같이 구분한다.

구분	통계해독력	통계적 추리	통계적 사고
정의	• 통계학의 기본언어 및 수단을 이해하고 사용하기 • 통계용어의 의미를 이해하기 • 통계심볼의 사용을 이해하기 • 데이터의 표현을 인식하고 해석하기	• 통계 개념으로 판단하는 방법 • 통계정보의 의미를 취하기 • 통계 개념들을 서로 연결하기 • 통계 개념들을 자료/확률과 결합시키기 • 통계 처리과정을 이해하고 설명하기 • 통계처리 결과를 해석하기	• 통계조사/연구가 왜 시행되고 어떻게 수행되는 지를 이해하기 • 전 통계조사/연구과정을 인식하고 이해하기 • 무작위현상을 통계모형이 어떻게 흉내 내는 지를 이해하기 • 통계적 추론 수단을 어떻게, 언제, 왜 사용하는 지를 인지하기 • 통계조사/연구를 계획하고 평가하고 결론을 이끌어내는 것을 이해하기
평가 작업과 관련된 용어	• 식별하기 • 기술하기 • 번역하기 • 해석하기 • 읽기 • 계산하기	• '왜?'에 대해 설명하기 • '어떻게?'에 대해 설명하기	• 적용하기 • 비판하기 • 평가하기 • 일반화하기
예	• 2007.07.19 부산지역 중학생 49명의 키를 측정하니 평균이 161.2cm, 표준편차가 7.3cm이었다. 통계학에 대해 잘 모르는 사람에게 표준편차의 의미에 대하여 설명하여 줄 수가 있는가?	• 다음 그림은 우리나라 67개 지점에 대하여 조사한 1971 ~ 2000년 평균기온(^0C, 년평균값)을 이용하여 그린 줄기-잎 그림이다. The decimal point is at the | 6 | 4 7 | 8 | 5 9 | 9 10 | 01124788899 11 | 1222446667789 12 | 00112223333556777889 13 | 00135578889 14 | 001348 15 | 55 16 | 2 계산기를 이용하지 말고 평균값이 중앙값보다 크겠느냐, 작겠느냐, 아니면 같겠느냐? 그 이유는 무엇인가?	• 종합대학교 한 곳을 선택한 후 이공계 1학년 학생 30명을 무작위로 뽑아 기초미적분학 시험을 치루니 평균이 81.7점, 표준편차가 11.45점이었다. 이 결과를 본 어느 전문대학 교수가 이 대학 어느 교양수업에 등록한 이공계 학생 321명 중 1학년 전체(53명)에 대하여 기초미적분학 시험을 치루니 95% 신뢰구간이 (69.47, 75.72)이었다. 다음 중 어느 문장이 옳으냐? 그 이유를 설명하여라. 물론 이 두 문장이 다 틀릴 수도 있다. 1. 이 전문대학 1학년 전체 학생의 평균시험성적이 종합대학교 평균시험성적보다 낮다. 2. 이 전문대학 교양수업에 등록한 1학년 학생 53명의 평균시험성적이 종합대학교 평균시험성적보다 낮다.

(참조: http://app.gen.umn.edu/artist/glossary.html)

[표 1.4] 통계해독력, 통계적 추리, 통계적 사고

이러한 통계해독력, 통계적 추리, 통계적 사고는 서로 떨어져있는 데이터의 지식화 산출물이 아니라 다음 [그림 1.15]와 같이 겹쳐 있다.

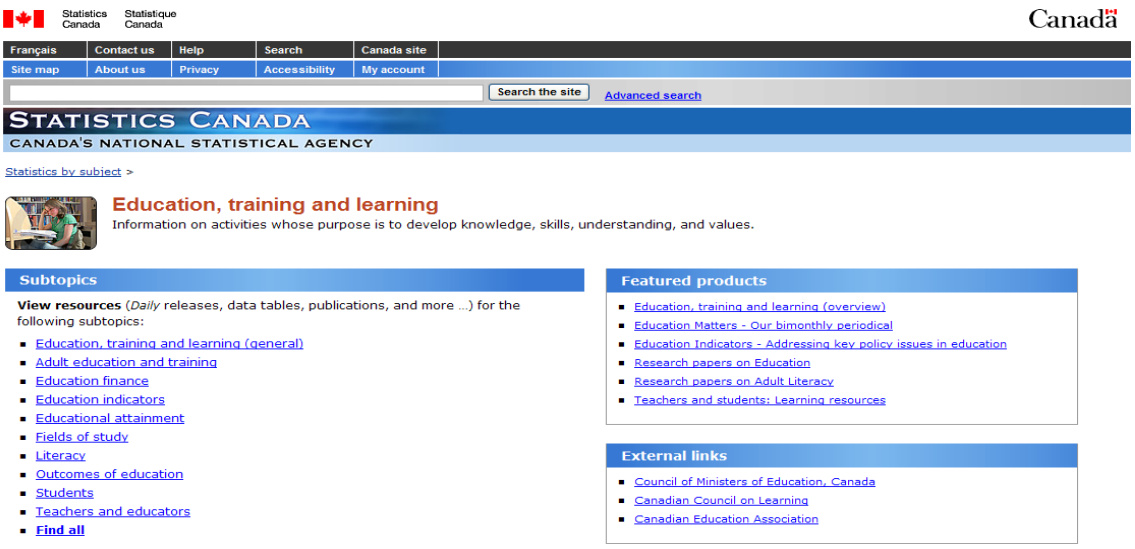


[그림 1.15] 통계해독력, 통계적 추리, 통계적 사고의 관계

세계 각국은 통계문맹을 퇴치하기 위한 노력을 기울이고 있는 데 캐나다의 Statistics Canada (우리나라의 통계청과 같은 기관)는 학교 학생들에게 통계의 본질에 대하여 가르치는 프로그램을 시행하고 있다. 다음 화면은 Statistics Canada 영문홈페이지(www.statcan.ca/menu-en.htm) 중 “Education, training and learning” 메뉴화면이다. 이 메뉴를 통하여 통계문맹을 퇴치하고 젊은이들을 양적정보에 대하여 좋은 소비자가 되도록 도와주기 위한 노력을 엿볼 수 있다. 이 노력 중 ‘Education Outreach Program’이라는 프로그램이 있는데 다음과 같은 두 가지 기둥으로 이루어져 있다.

1. 대중적인 전파를 위한 온라인 학습자원포털
   (1) E-STAT: Statistics Canada가 갖고 있는 사회경제적 자료에 대한 데이터웨어하우스 (Cansim)와 센서스데이터(최근, 과거)
   (2) Statistics: Power from Data!: 기초통계학을 내용으로 하는 모듈중심 훈련패키지(13개의 독립적인 모듈로 구성)

2. Statistics Canada가 갖고 있는 인적자원/전문가그룹과 지방협회/기관(예로 교사연합회, 교육청, 교육과정/IT전문가그룹 등)과의 네트워크 형성: 캐나다 전국을 대상으로 훈련, 지도, 지식전달을 전담하는 전문가들이 Statistics Canada를 대신하여 지방협회/기관들과 네트워크를 형성하고 활동함



외국 민간기관에서도 통계문맹을 퇴치하기 위한 노력으로서 학교 학생들 및 일반인들에게 통계학을 가르치는 프로그램을 시행하고 있다. 예로 미국통계협회(American Statistical Association) 홈페이지 메뉴 중 Education 메뉴를 선택하면 서브메뉴가 나타난다. 각 그룹(초·중·고등학교, 대학교, 대학원·전문가·평생교육)별로 나누어 통계교육에 대한 정보를 제공하고 있다.

가장 큰 특징 중 하나가 미국통계협회와 NCTM(전미수학교사협의회, National Council of Teachers of Mathematics) 사이에 공동위원회(joint committee)를 1982년부터 만들어 학교현장에서 초·중·고등학교 학생들에게 통계교육을 어떻게 효과적으로 가르치고 학생들의 통계적 사고를 어떤 방법으로 함양할 것인가에 대한 고민과 해결방안을 시도하고 있다는 것이다. 이 공동위원회에서는 초·중·고등학교 수학교사들 간의 의사소통을 위하여 STN(Statistics Teacher Network)이라는 소식지도 발간하고 있다.

통계교육에 유용한 웹사이트도 다음 [표 1.5]와 같이 제공하고 있다(2007년 4월 수정).

일반 사이트	Exploring Data	http://exploringdata.cqu.edu.au/
	The CHANCE Project	http://rtmouth.edu/~chance
	CAUSE(The Consortium for the Advancement of Undergraduate Statistics Education) web	http://www.causeweb.org
	International Statistical Literacy Project	http://www.stat.auckland.ac.nz/~iase/islp/priteach http://www.stat.auckland.ac.nz/~iase/islp/priclass
	NCTM Illuminations web site	http://illuminations.nctm.org/
	JSE(Journal of Statistics Education) Information Service	http://www.amstat.org/publications/jse/jse_info_service.html
	ARTIST (Assessment Resource Tools for Improving Statistical Thinking)	http://app.gen.umn.edu/artist/index.html
	STN(Statistics Teacher Network)	http://amstat.org/education/stn/
데이터 소스	DASL(The Dataset and Story Library)	http://lib.stat.cmu.edu/DASL/
	Chance Project Datasets	http://www.dartmouth.edu/~chance/teaching_aids/data.html
	U.S. Census Bureau home page	http://www.census.gov/
	FEDSTATS	http://www.fedstats.gov
	WWW Virtual Library	http://www.stat.ufl.edu/vlib/statistics.html
	UCLA Statistics Data Sets	http://www.stat.ucla.edu/data/
	UCLA Statistics Case Studies	http://www.stat.ucla.edu/cases/
	Rice Virtual Lab in Statistics	http://onlinestatbook.com/rvls.html
기타 사이트	Robin Lock's page at Saint Lawrence University	http://it.stlawu.edu/~rlock/
	Al Coons at Buckingham, Browne and Nichols School	http://www.bbns.org/us/math/ap_stats
	Herkimer's Hideway, prepared by Sanderson Smith	http://www.herkimershideaway.org
	M&M's web site	http://www.mms.com/
	The North Carolina School of Science and Mathematics Statistics Summer Institutes	http://courses.ncssm.edu/math/Stat_Inst/links_to_all_stats_institutes.htm

[표 1.5] 통계교육에 유용한 웹사이트

외국민간기관에서는 통계교육 관련 잡지 및 소식지도 다음 [표 1.6]과 같이 발간하고 있다.

저널명	발행기관	대상
Numeracy - 전자저널(무료) http://services.bepress.com/numeracy/	NNN(National Numeracy Network)	교사, 통계교육전문가
JSE(Journal of Statistics Education) - 전자저널(무료) http://www.amstat.org/publications/jse/	ASA	교사, 통계교육전문가, 학생
Chance	ASA	교사, 통계교육전문가, 학생
STATS	ASA	학생
Teaching Statistics	blackwell 출판사	교사(9-19세 학생대상 통계교육), 통계교육전문가
SERJ(Statistics Education Research Journal) - 전자저널(무료) http://www.stat.auckland.ac.nz/~iase/publications.php?show=serj	IASE(the International Association for Statistical Education)	교사, 통계교육전문가
TISE(Technology Innovations in Statistics Education) - 전자저널(무료) http://repositories.cdlib.org/uclastat/cts/tise/	The UCLA Dept. of Statistics Center for Teaching Statistics	교사, 통계교육전문가

[표 1.6] 통계교육 관련 잡지들

우리나라 통계청에서도 통계문맹을 퇴치하기 위한 노력을 다각도로 기울이고 있다. 초등학생들을 위하여 어린이통계동산(mirae.nso.go.kr) 페이지를 운영하여 어릴 때부터 통계적으로 사고하는 훈련을 함양할 수 있는 데이터베이스를 제공하고 있다. 통계청산하 통계교육원에서는 매년 두 가지 통계경진대회를 실시하는 데 초등학교 4~6학년 재학생을 대상으로 하는 전국어린이 통계활용대회와 중학교 재학생을 대상으로 하는 전국중학생 통계활용대회가 있다. 그리고 어린이통계교실을 통하여 지방통계청별로 일상생활 속에서 쉽게 접할 수 있는 재미있는 통계를 찾아보고 통계퀴즈, 컴퓨터를 통한 통계체험 등 다양한 프로그램을 실습하고 있다. 또한 초등학교를 대상으로 매년 2개 통계교육연구학교를 설치하고 운영하는 데 도움을 주고 있다. 또한 통계청 홈페이지에는 '통계체험하기'라는 코너가 있는 데 '우리집 씀씀이', '물가변동 알아보기', '우리집 물가지수' 등이다. 이를 통하여 우리는 우리생활에 밀접한 통계를 체험하게 된다.

국가통계포털인 KOSIS(www.kosis.kr)에 가보면 통계교실 메뉴가 있는 데 다음 [표 1.7]과 같이 일반인 통계교실, 청소년 통계교실, 어린이 통계교실, 통계법제도 4가지 부메뉴로 구성되어 있다. 3개의 부메뉴(일반인 통계교실, 청소년 통계교실, 어린이 통계교실)에는 각 수준에 맞는 재미있는 애니메이션이 제공된다. 이러한 일반인 통계교실, 청소년 통계교실, 어린이 통계교실의 내용이 단계별 통계교육 표준지침이 될 정도의 풍부한 콘텐츠를 담도록 보강할 필요가 있다.

메 뉴	부메뉴	내 용
통계교실	일반인통계교실	애니메이션(국가통계의 의미와 통계품질관리) 인구, 고용통계, 물가, 가구소득/소비, 경제지표, 표본조사
	청소년통계교실	애니메이션(말발굽에 치어죽은 군인은 얼마나 될까?) 통계이야기, 통계로 본 세상, 청소년 통계
	어린이통계교실	애니메이션(쉬운 통계의 세계로 빠져보세요!) 통계만들기, 통계표와 도표, 통계의 이용
	통계법제도	통계제도, 법령

[표 1.7] KOSIS 내의 통계교실 메뉴

초·중·고등학교와 대학교에서 통계교육을 실시하는 주된 이유는 우리들의 실생활에서 통계문맹을 없애고 통계적 사고를 함양하는데 있다. 통계학을 한 마디로 정의한다면 '데이터로부터 배우는(데이터의 지식화) 과학'이라고 앞에서 정의하였다. 통계학에서 가장 중요한 것이 바로 '데이터(자료)'인 것이다. 그러므로 통계교육에서도 통계교육 시 '어떤 자료를 사용하느냐?'하는 것이 통계교육의 성패를 가름할 정도로 중요한 문제가 된다. 그런데 이러한 문제는 통계교육을 책임지고 있는 교사나 강사, 교수의 국가통계에 대한 개인적인 관심과 활용의지에 크게 좌우된다. 모든 통계자료에는 자료제공처(제공자), 자료 작성일 등이 꼭 명시되어 있어야 한다. 자신이 직접 실험하거나 관측된 자료가 아닌 타인의 자료를 이용할 때 그 자료의 출처를 밝히는 훈련은 학생들에게 매우 중요한 훈련이 된다.

초등학교 '수학'과 '수학익힘책' 교과서(1종)에 나타나는 국가통계 자료의 활용 예에서 자료의 출처와 관련된 사항은 다음과 같다.

1. 국가통계 자료를 제공한 기관에 대한 언급이 없는 사례가 55%(20건 중 11건)나 된다.
2. 국가통계 자료를 제공한 기관에 대한 홈페이지 주소를 언급한 사례는 1건(기상청)뿐이다.
3. 국가통계 자료를 제공하면서 자료제공기관, 홈페이지 주소, 제공년도가 전부 없는 사례가 20%(20건 중 4건)나 된다.

중·고등학교 수학교과서(7-10단계 16종), 실용수학(4종)과 수학 Ⅰ(12종) 수학교과서에 나타나는 국가통계 자료의 활용 예에서 자료의 출처와 관련된 사항은 다음과 같다.

1. 다양한 국가통계자료를 활용하고 있었고 자료의 원천이 되는 웹 홈페이지 주소를 명기하고 있다. 그러나 교과서마다 웹 자료에 대한 언급에 편차가 심했다.
2. 교과서에서 인용되는 그래프와 도표 중 자료 제공처를 명기한 비율이 매우 저조하다. 예로 실용수학에서는 최소 3.3%, 최대 17.5%이었고, 수학 Ⅰ에서는 최소 0%, 최대 8.9%이었다. 또한, 수학 Ⅰ에서 자료는 제공하지 않고 웹 홈페이지 주소만 언급한 경우가 42번 중 25번이나 되었다.

교과서에 제시된 자료에 자료제공처(제공자)가 명시되어 있지 않는 경우 교사는 자료제공처(제공자), 자료 작성일 등에 대하여 정보를 수집하여 학생들에게 제시하여야 한다. 통계 자료를 얻기 위해서는 먼저 자료를 수집하는 주체, 자료 수집 일자 및 기간, 자료 수집 방법 등이 결정되어야 하며, 이러한 정보가 통계자료의 질에 큰 영향을 주게 된다. 모든 교과서에는 통계자료 수집의 결과만 나와 있으므로 이러한 자료를 얻게 되기까지의 선행 작업에 대한 언급이 필히 있어야 한다. 교사가 사전 학습자료 준비를 통하여 '실생활 문제'와 관련된 자료들을 수집하여 학생들에게 제시하여야 한다. 통계영역 교수-학습활동에서는 학생들이 다양한 자료를 접하기 위하여 신문 및 인터넷을 적극 활용하여야 한다. 이 때 되도록 시의 적절한 자료를 활용하도록 한다. 통계청의 홈페이지 내의 자료광장에 있는 통계 DataBase(국가통계포털(KOSIS)), 통계표준분류, e-나라지표, 마이크로 통계(KMDSS), 통계지리정보시스템(GIS)을 이용하면 '실생활 문제'와 관련되어있고 시의 적절한 통계자료를 얻을 수 있다. 물론 통계청이 아닌 정부기관 및 지정기관의 홈페이지를 통해서도 많은 국가통계 자료를 얻을 수 있다. 여기서 한 가지 지적할 사항은 국가통계자료는 통계교육 시 중요한 보조수단이 됨으로 국가통계작성기관에서는 그 기관이 책임지고 생산하는 국가통계자료가 통계품질을 결정하는 6개 요소인 1) 정확성, 2) 시의성, 3) 관련성, 4) 접근성, 5) 비교성 및 6) 효율성을 만족하는 통계자료가 되도록 통계시스템의 개혁을 위하여 부단히 노력하여야 한다는 것이다.

1.4 통계자료분석의 과정

통계자료분석의 과정은 다음과 같다.

1. 문제(problem)의 파악
2. 자료의 수집(the collection of data)
3. 자료의 요약(data summary) 및 분석(data analysis)
4. 결론(conclusion)

각 과정에 대하여 살펴보자.

문제의 파악

통계자료분석에 있어서 '이 자료분석의 목적(objective)이 무엇인가'를 아는 것이 중요하다. 경험이 없는 통계전문가는 '이 자료분석의 목적이 무엇인가?' 또는 '이 자료가 목적하는 분석에 적절한가?'라는 고려도 없이 무작정 복잡하고 난해한 자료분석 방법으로 곧장 몰두하여 버리는 과오를 자주 범한다. 우리는 뛰기 전에 보아야 한다(Look before you leap!). Albert Einstein이 한 다음 이야기는 통계자료분석에도 그대로 적용할 수 있는 이야기이다.

"The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill."

문제를 정형화(formulation)하기 위하여 필요한 행동 4가지를 열거하면 다음과 같다.

1. 물리적 배경을 이해하라. 자료분석가는 자주 타 분야의 사람들과 협조를 하며 일을 하게 된다. 이 때 그 사람들의 전문분야에 대하여 이해를 하여야 한다. 이 전문분야에 대한 지식을 배우는 일을 허드렛일로 보지 말고 기회로 여겨라.
2. 목적을 이해하라. 다시 한번 말하지만 자료분석가는 자주 '이 자료분석의 목적이 무엇인지'를 모르는 타 분야의 사람들과 협동으로 일을 하게 된다. 낚시여행을 생각하여 보라! 낚시에서 물속을 잘 보아야 물고기를 낚을 수가 있지 우연히 물고기가 낚이는 것이 아니다.
3. 자료분석을 맡기는 의뢰자가 무엇을 원하는 지 확실히 알아야 한다. 우리는 같은 자료인 데도 불구하고 아주 다른 분석을 할 수가 있다. 정말로 의뢰자가 요구하는 정도보다 더 지나쳐 아주 복잡한 자료분석을 행하는 경우가 가끔 있다. 기술통계량의 계산 정도만 필요할 때도 있는 것이다.
4. 문제를 통계적 용어로 변환하여라. 이 과제는 도전적인 작업이다. 우리는 가끔 이러한 작업에서 돌이킬 수 없는 과오를 범하기도 한다. 문제가 통계적 용어로 번역이 되기만 하면 문제 풀이는 거의 기계적 절차라 할 수 있다.

자료의 수집

통계자료분석 시 문제를 파악한 후에는 데이터를 수집하여야 한다. 쓰레기 같은 데이터를 이용하면 자료분석 방법이 아무리 좋아도 쓰레기 같은 결론만 얻어낼 뿐이다(Garbage in, garbage out). 우리가 이 단계에서 명심할 것은 '데이터는 거짓말을 하지 않는다.'라는 것이다. 이 단계에서는 '어떻게 자료가 수집되었는가?'를 이해하는 것이 중요하다.

자료수집 시 유의할 사항은 다음과 같다.

1. 데이터가 관측이나 실험이 가능한가? 이 표본은 편의표본(convenience sample)인가? 이 데이터는 관측자료인가, 아니면 실험자료인가? '어떻게 수집되었는가?'하는 것이 '어떤 결론을 내릴 것인가?'를 결정하는 데 아주 중요한 영향을 준다.
2. 무응답(non-response)은 없는가? 우리가 보고 있는 자료만큼 우리가 보지 못하고 숨겨 있는 자료도 중요하다.
3. 결측값(missing value)이 있는가? 이 문제는 매우 골치가 아프고 해결하는 데 시간을 많이 요구하는 문제이다.
4. 데이터를 어떻게 코딩할 것인가? 특히 질적 자료(qualitative data)들을 어떻게 처리할 것인가? 질적 자료에 대한 설명은 4.2절에서 다루게 된다.

5. 측정단위(measurement unit)는 무엇인가? 측정에는 타당성(validity)과 정확성(accuracy)이 보장되어야 한다.
6. 자료입력시의 오류와 자료의 오염(corruption)에 대하여 주의하라. 중간 규모 이상의 데이터 셋에서는 너무나 자주 일어나는 현상이다. 미친(insane) 데이터는 없는가 점검하라! 자료의 오염이 있을 때는 자료세탁(data cleaning)이 필요하다.

자료의 요약 및 분석

자료를 그림이나 수치를 이용하여 요약하고 자료에 대한 구조를 살펴보고 목적과 자료의 성격에 따라 이에 맞는 자료분석 방법을 선택한다.

결론

자료분석의 결과를 이용하여 통계적 결론을 내리고 자료분석을 맡긴 의뢰자와 함께 상의하여 전문분야에서 필요한 결론을 내린다.

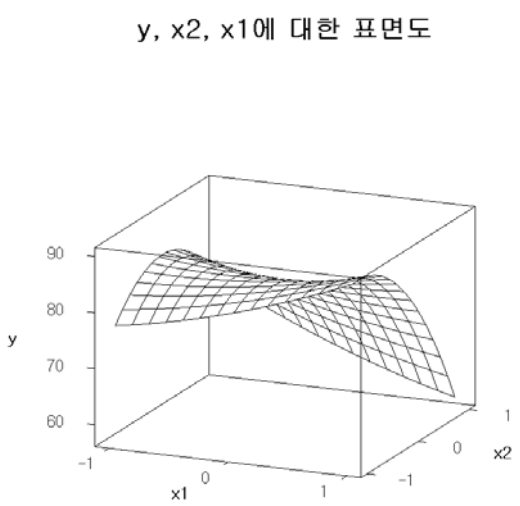
1.5 실험연구와 관측연구

데이터를 수집하는 데 필요한 대표적인 통계방법이 표본조사와 실험계획이다. 과학적 연구에 활용되는 데이터는 실험이나 관측을 통하여 얻게 된다. 표본조사에 대한 이야기는 2장과 9장에서 다루므로 본 절에서는 실험연구와 관측연구에 대하여 간략히 언급하고자 한다.
과학적 연구는 방법에 따라 실험연구(experiment study)와 관측연구(observational study)로 나뉜다. 관측연구란 연구대상의 상태와 특성을 관측하여 자료를 수집한 후 분석하는 연구이고, 실험연구는 연구자가 관심이 있는 처리를 연구대상에게 실시한 다음 효과를 측정하여 자료를 구한 후 주어진 처리와 그에 따른 반응 간의 인과관계를 밝히는 연구이다.

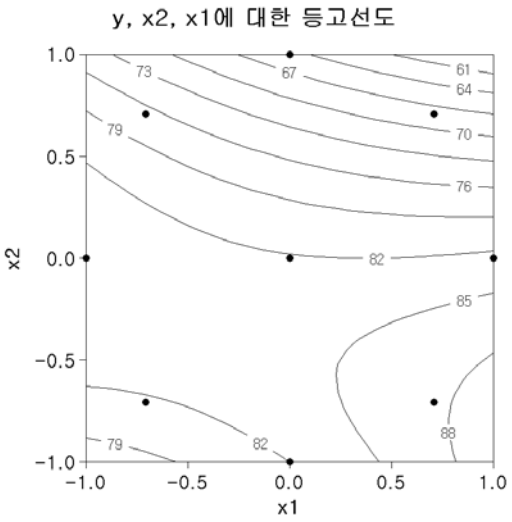
실험연구

변수 사이의 인과관계를 밝히고자 하는 경우에 실험연구를 행한다. 실험연구에서는 변수들이 반응변수와 설명변수로 이루어지는 데 반응변수란 처리에 의하여 변화하는 우리가 연구하고자 하는 변수이고 설명변수는 반응변수에 영향을 주는 변수이다. 처리란 실험단위에 적용되는 특정한 실험조건이다.

예제 1.12 (실험연구) 화학물질 Mercapatobenzothiazole(MBT)의 합성과정에 있어서 중요한 변수로 합성시간(ξ_1)과 합성온도(ξ_2)가 있다. 이 두 변수가 MBT의 합성률(y)에 어떠한 영향을 주는지 알고자 합성시간(ξ_1)과 합성온도(ξ_2)을 각각 표준화(x_1과 x_2)한 후 12개의 실험조건을 결정한 다음 실험을 행하였다. 다음 [그림 1.16]은 등고선도(contour plot)이고 [그림 1.17]은 표면도(surface plot)이다. [그림 1.16]에서 12개의 점들(실험점 중 원점 (0,0)에서 3번 반복 실험되고 (-1,0)에서 2번 반복 실험)이 12개의 실험조건을 나타낸다.



[그림 1.16] 등고선도



[그림 1.17] 표면도

최적조건을 찾으면 최적조건이 $x_1 = 1, x_2 = -1$이 됨을 알 수 있다. 원 설명변수로 표시하면 ξ_1(합성시간)이 20시간, ξ_2(합성온도)가 220도가 되고 이 최적조건에서의 합성률은 89.6%가 된다. ■

관측연구

실험을 하기 어렵거나 비효율적인 경우 실험 대신 관측연구를 행하게 된다. 관측대상에게 인위적인 처리를 가하지 않고 있는 그대로 관측하여 얻은 자료를 이용하여 자료분석을 하는 연구이다. 대표적으로 사례-대조연구(case-control study)와 코호트연구(cohort study)가 있다.

예제 1.13 (관측연구) 다음 표는 흡연과 심근경색의 관련성에 대한 사례-대조 연구의 결과이다. 심장병으로 진료 받은 중년 이하의 부인 환자들이 사례집단이고, 대조집단은 같은 병원에서 그 밖의 다른 급성질병으로 진료를 온 사람들이다. 과거 흡연량(일 평균)에 따라 사례 및 대조 표본들을 분류하였다. 우리는 흡연 수준에 따라 범주 점수를 각각 0, 7.5, 19.5, 30을 부여하여 통계모형을 적합시키고 효과에 대한 통계분석을 수행할 수 있다.

그룹	흡연량	사례수	대조수
1	0	90	346
2	1-14	57	91
3	15-24	65	48
4	≥ 25	40	18

통계분석을 수행하여 승산비(odds ratio)를 비교하여 보면 다음 표와 같다. 어떤 사건이 발생할 확률이 p라면 승산(odds)는 $\frac{p}{1-p}$로 정의하고 승산비란 하나의 그룹의 승산을 또 다른 그룹의 승산으로 나눈 값이다. 예로 그룹 4 대 그룹 1의 승산비가 8.543이란 그룹 4가 그룹 1에 비하여 상대적으로 심근경색발생위험이 8배 이상 높다는 것을 의미한다.

비교	승산비
그룹 2 대 그룹 1	2.408
그룹 3 대 그룹 1	5.206
그룹 4 대 그룹 1	8.543

담배를 많이 피울수록 심근경색발생위험이 커짐을 알 수 있다. 흡연과 심근경색의 관련성을 볼 수 있다. 그러나 여기서 주의할 사항은 사례집단에 있는 사람들(심장병 환자)은 과거 흡연습관을 후회하는 심리적 상태에 있기 쉬우므로 과거 흡연량을 다소 과다 보고할 수 있고 대조집단에 있는 사람들(다른 급성질병)은 과거 흡연량에 대하여 가볍게 생각해서 다소 과소 보고할 수 있다. 그러므로 실제 흡연이 심장병 발생과 아무 관련이 없는 상황에서도 사람들은 흡연이 심장병 발생과 관계가 있다고 생각하기 때문에 사례-대조연구 결과가 흡연이 심장병 발생과 관계가 있다고 나올 수 있다. ■

학습요약

1. 우리는 우리생활에 필요한 수많은 자료에 둘러싸여 있고 이러한 자료가 폭발적으로 양이 많아짐에 따라 통계학의 필요성이 점점 커지고 있다.

2. 통계학이란
(1) 우리의 일상생활에서 얻어지는 다양한 통계자료를 수집, 정리하고 평가하고 의미있는 결론을 이끌어내는 작업을 수행하는 학문이다.
(2) 조사와 실험에 대한 설계, 데이터의 수집, 처리, 분석, 결과의 해석을 행하는 과학이다.
(3) 데이터로부터 배우는(데이터의 지식화) 과학이다.

3. 통계학에서 다루는 중요한 3가지 주제는 '모집단과 표본', '자료의 축약방법', '변동'이다.
(1) 연구대상이 되는 가능한 관측값이나 측정값의 집합을 모집단이라 하며, 통계적 처리를 위하여 모집단에서 실제로 추출한 관측 값이나 측정값의 집합을 표본이라 한다. 우리는 모집단을 잘 반영하도록 표본을 뽑아야 한다.
(2) 우리는 자료를 축약시킴으로써 자료의 구조를 보아야 한다.
(3) 통계학이란 한 마디로 변동에 대하여 연구하는 학문이다.

4. 통계학을 크게 두 가지 유형으로 구분하면 기술통계학과 추측통계학으로 구분할 수 있다.

5. 통계학이 데이터의 지식화를 위한 과학이므로 데이터가 있는 모든 학문분야에 통계학이 쓰인다.

6. 우리는 통계해독력, 통계적 추리, 통계적 사고를 함양함으로써 통계문맹을 퇴치할 수 있다.

7. 통계자료분석의 과정은 문제의 파악 -> 자료의 수집 -> 자료의 정리 및 분석 -> 결론 순이다.

8. 과학적 연구에 활용되는 데이터는 실험이나 관측을 통하여 얻게 된다. 과학적 연구는 방법에 따라 실험연구와 관측연구로 나뉜다.

1장 연습문제

1.1 (모집단과 표본) 어느 대기업에서 직원들의 직장생활 만족도를 조사하기 위하여 전체 직원들 중 100명을 뽑아서 이들을 조사한 다음 이를 바탕으로 전체 직원들의 만족도를 알아보고자 한다. 이 경우 모집단과 표본이 무엇인지 답하라.

1.2 (모집단과 표본) (1) 어느 대도시에서 실업률을 알아보고자 그 대도시의 전체 노동 인구 중 1,000명을 추출하여 이들의 실업률을 조사함으로써 그 대도시의 전체 실업률을 알아보고자 한다. 이 경우 모집단과 표본이 무엇인지 답하라.
(2) 갑과 을 두 기관이 각각 1,000명씩을 추출하여 실업률을 조사한 결과 갑의 표본으로부터 얻어진 실업률은 5.0%이였고, 을의 표본으로부터 얻어진 실업률은 6.0%이였다. 두 표본으로부터 구한 실업률이 서로 다른 이유에 대하여 설명하여 보라.

1.3 다음 자료는 환경부 홈페이지(www.me.go.kr)에 있는 2004년 전국시도별 대기오염물질 배출량(단위: 톤) 자료이다. 국가 대기오염배출량에서는 다양한 대기오염물질 중 환경기준 대기오염물질인 일산화탄소(CO), 질소산화물(NOx), 황산화물(SOx), 먼지농도(TSP, PM10) 및 VOC(Volatile Organic Compound : 휘발성 유기화합물) 등에 대한 배출량 자료를 제공한다. 대기 중 먼지 농도를 나타내는 통상적인 표현방법으로는 TSP, PM10, PM2.5 등이 있다. TSP(total suspended particulate)는 대기 중 부유상태에 있는 총먼지의 양이고, PM10은 직경 10㎛ 이하인 먼지의 양이며, PM2.5는 직경이 2.5㎛ 이하인 먼지의 양이다. 현재 우리나라 환경기준법에서는 TSP와 PM10에 대하여 농도 기준이 제시되어 있다.

시도명	CO	NOx	SOx	TSP	PM10	VOC
서울특별시	161,154	103,549	6,462	4,585	4,424	77,694
부산광역시	50,187	73,486	22,554	3,469	2,994	35,013
대구광역시	41,013	41,446	5,711	2,405	2,154	42,400
인천광역시	48,694	70,380	10,367	2,983	2,635	58,600
광주광역시	18,363	17,054	1,265	861	828	14,248
대전광역시	22,299	22,497	1,423	1,085	1,052	14,195
울산광역시	31,049	64,512	59,230	13,266	8,737	84,708
경기도	147,336	201,078	31,387	10,287	9,346	161,266
강원도	31,842	73,605	21,564	5,865	3,749	19,236
충청북도	35,832	52,147	13,265	5,442	3,333	26,759
충청남도	44,087	234,958	62,936	5,784	4,568	50,754
전라북도	32,547	49,099	12,633	2,906	2,513	28,797
전라남도	37,842	98,485	58,024	7,858	5,656	62,523
경상북도	55,696	80,348	33,791	7,436	5,582	46,387
경상남도	49,801	182,292	103,959	5,301	4,433	69,708
제주도	9,213	12,589	2,233	549	486	4,949

(1) 이 자료를 구한 목적이 무엇이라고 생각하는가?
(2) 이 자료는 실험연구를 통하여 얻어졌는가? 관측연구를 통하여 얻어졌는가?
(3) 당신은 자료분석을 어떻게 하겠는가?

1장 실습문제

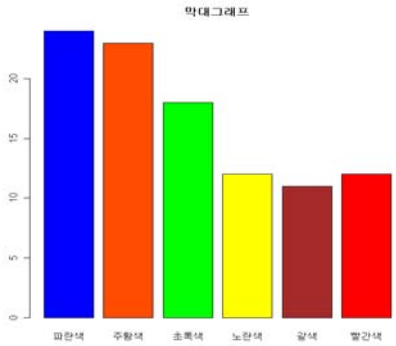
1.1 M&M milk chocolate는 6가지 색깔의 단추모양의 밀크쵸코캔디의 상표명이다. 미국 M&M 홈페이지(http://us.mms.com/us/about/products/milkchocolate/)에 가보면 여섯 가지 색깔의 밀크쵸코캔디의 비율은 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13이라고 나와 있다. 그러나 한 봉지 안에 들어있는 이 여섯 가지 색깔의 밀크쵸코캔디의 개수와 비율은 봉지마다 다름을 알 수 있다. 원생들이 구입한 M&M milk chocolate를 개봉하게 하여 다음과 같은 실습을 행한다.

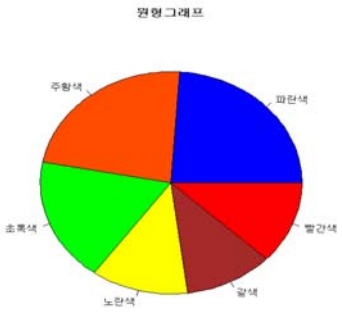
1. 6가지 색깔을 갖는 밀크쵸코캔디 각각의 개수와 총 개수를 세어 기록하게 한다. 이를 막대그래프와 원형그래프로 그려본다.
2. 원생 전체의 결과를 수집하여 분석하여 본다. 총 개수의 분포는 어떤 구조를 갖는지, 6가지 색깔을 갖는 밀크쵸코캔디 각각의 개수는 어떠한 비율로 나뉘는 지를 막대그래프와 원형그래프를 통하여 확인한다.

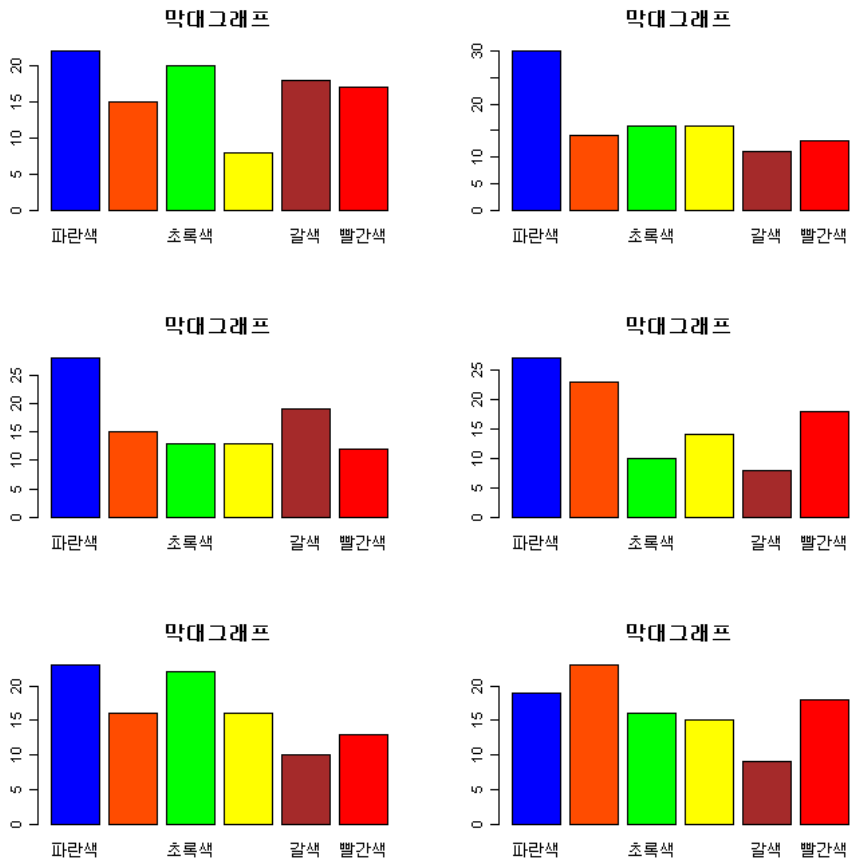
<목적> '변동(variation)'을 체험하게 한다.

<참고> 이러한 현상을 우리는 통계적 시뮬레이션으로 확인하여 볼 수 있다. 여섯 가지 색깔의 밀크쵸코캔디의 비율을 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13으로 포장한다고 가정하고 편의상 한 봉지 안의 밀크쵸코캔디의 개수는 100개라 하자. 그러면 이 분포는 다항분포가 된다. 우리는 통계패키지를 이용하여 통계적 시뮬레이션을 행하여 볼 수 있다. 통계적 시뮬레이션 결과 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색의 비율이 0.24: 0.23: 0.18: 0.12: 0.11: 0.12가 나왔다. 즉 여섯 가지 색깔의 밀크쵸코캔디의 비율을 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13을 지키며 랜덤하게 하나의 봉지에 100개의 밀크쵸코캔디을 담은 결과 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=24개: 23개: 18개: 12개: 11개: 12개가 나왔다는 것이다. 왜 그렇게 나왔을까? 이것을 이해하는 것이 '변동'을 이해하는 것이 된다.

위의 시뮬레이션 결과를 막대그래프와 원형그래프로 그리면 다음과 같다.







'변동'이라는 현상을 더 자세히 보기 위하여 위와 같은 시뮬레이션을 6번 반복 시행하여 보면 우리는 6번 반복시행이 조금씩 다 다른 모습을 이루는 것을 볼 수 있다. 이러한 현상이 바로 '변동'이라는 현상이다. 이 시뮬레이션을 통하여서도 변동을 확인할 수 있다. 6번 반복시행 모두 여섯 가지 색깔의 밀크쵸코캔디의 비율을 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색 =0.24: 0.20: 0.16: 0.14: 0.13: 0.13을 지키며 시뮬레이션을 행하였는데도 결과는 다 다르다.

1.2 다음과 같은 실습을 행한다.

(1) 원생 각자 A4 용지로 종이비행기를 접는다.
(2) 일정한 거리를 두고(약 3m) 직경 50cm 원을 그린다.
(3) 원생 각자 원의 중심을 향하여 종이비행기를 던진다.
(4) 종이비행기들의 흩어진 모습을 보고 '변동'이라는 현상에 대하여 토의를 행한다.

<목적> '변동'을 체험하게 한다.

쉬어가기

현대생활에 필요한 통계상식에 대하여 몇 가지 알아보자.

1. 퍼센트와 퍼센트 포인트

다음 대화는 2007.11.10 SBS TV 저녁 8시 뉴스 내용 중 일부이다.

<앵커>
그렇다면 국민들이 차기 대통령에게 기대하는 모습은 어떤 것일까요? 조사 결과, 우리 국민들은 경제 못지 않게 제대로 된 사회복지 시스템을 원하는 것으로 나타났습니다. 이현식 기자입니다.

<기자>
차기 대통령이 가장 우선해야 할 국정과제가 무엇인지 국민에게 물었습니다.
경제라고 답한 사람이 76.6%로 제일 많았고 사회·복지·문화가 두 번째였습니다.
외교, 안보, 통일이나 정치, 행정을 우선해야 한다는 대답은 상대적으로 적었습니다.
그런데 가장 바람직한 국가는 어떤 나라라고 생각하느냐는 질문에는 '경제가 풍요로운 나라'라는 답변이 21.1%인데 비해 '빈부 격차가 작고 사회 복지가 잘된 나라'라는 답변이 그 두 배가 넘는 49.1%나 됐습니다.
[배지영/직장인 : 아직 그 크기에 걸맞는 만큼의 혜택을, 경제적인 혜택을 사람들이 많이 못 누리는 것 같아서 복지 문제가 더 중요한 것 같습니다.]
특히 빈부격차 해소와 사회 복지를 가장 원한다는 답변은 한창 일할 나이인 30대에서 가장 많았습니다.
젊은이들이 미래에 대한 불안감에 위축되서 장·노년층보다도 양극화 현상을 더 우려하는 것으로 해석됩니다.
이 조사는 전국 만 19세 이상 남녀 천 명을 상대로 전화 면접방식으로 실시됐으며, 95% 신뢰 수준에 오차한계는 플러스 마이너스 3.1 퍼센트 포인트입니다.
성장이냐 분배냐, 또는 경제냐 복지냐라는 이분법을 넘어서 근심 없이 꿈을 추구할 수 있는 나라를 만들어 달라는 것이 국가 리더십에 대한 국민들의 바람입니다.

위 대화 중에 나오는 “퍼센트포인트(%P)”란 어떤 뜻인가? 우리가 변동률을 계산할 때 퍼센트(%)와 퍼센트포인트(%P)를 혼동하는 경우가 많다. 퍼센트는 보통 100을 기준으로 차지하는 비율을 의미하고 퍼센트포인트는 그런 퍼센트 사이의 차이를 의미한다. 예를 들어 콜금리가 3.2%에서 3.5%로 올랐다면 콜금리는 총 0.3%만큼 차이가 나게 올랐다. 이 경우 0.3% 인상되었다고 하지 않고 0.3%P 인상되었다고 표현한다. 퍼센트(%)와 퍼센트포인트(%P)는 다음과 같이 간단하게 정의할 수가 있다.

* 퍼센트 : 서로 다른 수치간의 변화율(비율)
* 퍼센트 포인트 : 퍼센트 값 사이의 차이(변화량), 뺀 값

2. 기여율과 기여도

다음 기사는 2007.11.29 매일경제신문에 나타난 글이다. “기여율”이라는 용어가 나온다.

내수 중 투자 부문의 경우 선진국들의 과거 2만달러 진입 시점의 총투자가 GDP에서 차지하는 비중을 한국이 약간 웃돌고 있다. 그러나 최근 한국의 투자는 설비투자보다는 주택 경기 호조에 따르는 건설투자에 의해 주도됐다는 점에 문제가 있다.
우리나라 설비투자의 경제성장에 대한 기여율은 70년대 연평균 31.3%에 달했으나, 2000년대에 들어서는 2.9%로 크게 낮아졌다.
반면 2000년대 건설투자의 기여율은 14.0%로 과거보다는 낮아졌으나 설비투자 기여율의 4배를 넘어서고 있다. 이런 이유로 성장 잠재력과 보다 밀접한 관련이 있는 설비투자의 침체는 향후 한국 경제의 성장 속도가 미약할 것이라는 우려를 낳고 있다.

다음 기사는 2007.11.27 뉴시스에 나타난 글이다. “기여도”라는 용어가 나온다. 기여율과 기여도의 차이는 무엇일까?

대구경북지역 건설산업의 침체가 지역 경제의 발목을 잡고 있는 것으로 지적됐다.

한국은행 대구경북본부는 27일 발표한 ‘대구경북지역의 건설업 현황과 특징’ 자료를 통해 지역 건설업의 생산 및 고용 비중은 각각 8.6%와 7.1%로 점차 하락하고 있고 건설시장규모도 외환위기 직전인 1997년과 비슷한 수준에 머물러 있다고 밝혔다.
실제로 지난해 대구경북지역의 건설발주액은 총 9.2조원으로 외환위기 직전인 1997년(9.1조원)보다 고작 0.6% 늘어나는데 그쳤다.
하지만 일반건설업체수는 2000년말 923개사에서 9월말 현재 1311개사로 42.0% 늘어나고 전문건설업체도 2000년말 3054개사에서 4044개사로 32.4% 증가해 공사물량은 변함이 없는데 업체가 난립해 과당경쟁 양상을 띠고 있다.
최근 대구경북지역 건설업의 지역경제성장에 대한 기여도는 2001년 0.77%p에서 2005년 -0.13%p로 하락한 데 이어 작년과 올해도 계속 떨어진 것으로 추정된다.
또 2007년도 시공능력평가 결과 전국 100위 이내인 지역 업체는 4개이며 1사당 평균 시공능력 평가액은 292억원으로 수도권업체 평균(830억원)에 비해 현저히 낮은 수준으로 지역 건설업의 위상이 급격히 추락했다.

기여율[寄與率]이란 합계값 또는 평균값의 변화(증감)에 대하여, 그 명세인 각 항목이 전체를 증감시키는 데 어느 정도 공헌하고 있는가를 나타내는 지표이다. 각 항목의 변화의 크기를 전체의 증감에 대한 백분율로 나타낸다. 만약 전체가 증가한 경우에 명세 항목 중에 감소한 것이 있으면, 그 항목은 마이너스의 공헌을 한 셈이 된다. 가령 A·B·C 세 가지 상품을 팔고 있는 상점에서, 당일의 매상액이 전날에 비해 50만 원 증가하고 그 내역이 A가 40만 원 증가, B가 20만 원 증가, C가 10만 원 감소하였다고 가정한다. 이 경우의 기여율은 A가 80%(=40÷50×100), B가 40%(=20÷50×100)이며, C는 마이너스 20%(=-10÷50×100)가 된다.

기여도[寄與度]는 물가상승, 하락이나 GNP성장률 등에 대해 특정 항목이 얼마나 기여하고 있는가를 나타낸 것이다. 예컨대 소비자물가지수가 전월비 1% 상승하고, 이 가운데 쌀가격의 상승만으로 소비자물가지수가 0.25% 올랐다면 쌀가격의 기여도는 0.25%이며 기여율은 25%라는 계산이 된다.

정리하면 기여율은 기준시점의 통계치를 구성하는 각 요소의 증가분을 전체의 증감분에 대한 백분비로 표시한 것이고, 기여도는 통계치를 구성하는 각 요소가 전체증감률에 어느 정도 기여하는 지를 나타내는 것이다.

3. 각종 생활지수

다음 기사는 어느 날 아침 mbn 매일경제 TV에 방영된 내용이다. 본문 중 "생활지수"라는 단어가 나온다. 어떤 뜻일까?

아침부터 마음이 뜨거워지시죠? 2012년 여수 세계 박람회 개최가 결정이 됐는데요. 하지만 오늘 아침 피부로 느껴지는 공기는 차갑습니다. 어제 아침보다 춥습니다. 중부지역은 기온이 영하권으로 내려간 곳이 많으니까요. 출근길 따뜻하게 입으셔야겠습니다. 한 낮에도 어제보다 기온이 2~5도 가량 낮아지면서 쌀쌀한 하루가 되겠습니다. 하지만 하늘은 맑아서 오늘도 햇볕이 따뜻하겠습니다.
지역별 날씨입니다. 오늘 중부지역은 구름 한 점 없이 깨끗한 하늘이 예상되고요. 남부지역에는 구름이 많겠습니다. 또 울릉도, 독도는 차차 흐려져서 오후에 비가 내리겠습니다.
현재기온 서울은 영하 1.1도 기록하고 있고 중부지역은 중심으로 쌀쌀한 아침입니다.
낮기온은 서울이 7도, 전주 9도, 부산 15도로 한 낮에도 어제보다 기온이 2~5도 가량 낮아지겠습니다.
다음은 오늘의 생활지수입니다. 오늘은 빨래나 세차 하셔도 되겠습니다. 당분간 들리는 비소식은 없는데요. 하지만 쌀쌀해진 날씨에 건강 관리는 잘 하셔야겠습니다.
당분간 특별한 비소식은 없습니다. 서울의 아침기온은 0도에서 1도를 유지하면서 한 주 동안 다소 쌀쌀한 날씨가 이어지겠습니다.
제 142회차 세계박람회기구 총회가 열린 프랑스 파리는 하늘이 조금 흐리다고 하는데요. 그 밖의 다른 나라들은 어떤지 오늘의 세계 날씨 보시죠.

생활 지수(指數)란 날씨와 기온이 우리 생활에 미치는 영향의 정도를 1 ~ 100까지의 지수로 표시한 것을 생활 지수라고 한다. 날씨에 따른 생활 지수는 불쾌지수(不快指數)외에도 난방 지수, 빨래 지수, 운동 지수, 외출 지수(나들이 지수), 세차 지수, 불조심 지수 등 여러 가지가 있다. 나들이, 세차, 빨래, 운동 지수는 높을수록 쾌적하며 좋은 상태를 나타내고, 불쾌지수는 높을수록 불쾌감이 높은 것을 나타내며, 난방지수는 높을수록 난방이 필요한 것을 나타낸다. 다음 [그림 1.18]은 어느 날 포털사이트 네이버 상의 뉴스>날씨(weather.news.naver.com)에 나타난 생활 지수이다.

[그림 1.18] 각종 생활 지수

4. 경제활동인구와 실업률

다음 기사는 2008.01.28 한경비지니스에 실린 칼럼(저자: 조준모) 기사 중 일부이다.

전두환 정권 시절 대학의 졸업 정원제가 실시되면서 대학 졸업자가 크게 증가했다. 그 이후 김대중 정권하에서 대학 설립 규제가 완화되면서 대학 수가 급증했고 노무현 정권 들어서는 급기야 청년 가운데 대졸자의 비중이 80%를 상회하는 국면에 이르게 됐다. 청년 대졸자의 공급은 이렇게 증가한 반면 외환위기 이후 대졸자 노동력에 대한 수요는 공급 대비 정체 상태에 있다. 이런 형국에서도 정부는 낮은 청년 실업률에만 매달려, 늘어나는 청년 비경제활동인구화 문제에 적절하게 대응하지 못하고 있다는 비판을 받고 있다.

현재의 청년 노동시장 문제를 이해하기 위해 간단한 숫자를 예로 들어보자. 시장에 공급되는 대졸자를 위한 좋은 일자리 수는 50명 분인데 100명이 시장에 공급된다면 100명은 50명 안에 들기 위해 해외 연수, 인턴, 토익 점수, 자격증 등 노동시장에 추가 신호(signal)를 보내기 위해 투자할 수밖에 없다. 좋은 일자리에 취업하지 못한 나머지 50명은 이전에 고졸자가 일하는 일자리에 취직을 하거나 실업 혹은 비경제활동인구화하는 상황이 벌어지는 것이다. 여기서 발생하는 것이 대졸자의 하향 취업과 전공 불일치 취업 문제, 그리고 청년 니트(NEET) 문제다.

청년 니트(NEET: Not in Employment, Education or Training)란 취직도 하지 않고 교육이나 훈련 과정에도 없는 실업자나 비경제활동인구화된 청년층을 말한다. 청년 니트는 다시 네 가지 유형으로 세분화된다. 첫 번째 유형은 전통적인 청년 실업자로, 2007년 8월 기준으로 31만6000명이다. 두 번째 유형은 고시족, 공시족을 포함하는 함정형 니트로, 41만7000명을 점한다. 세 번째 유형은 현실 회피형 니트로, 일본의 은둔 외톨이(히키코모리) 니트족과 유사하며 30만5000명에 해당된다. 마지막으로 통계치에 포착되지 않은 네 번째 유형은 가족 노동형 니트로, 구직 무급 종사자와 가사 노동자를 포함하며 30만 명 이상

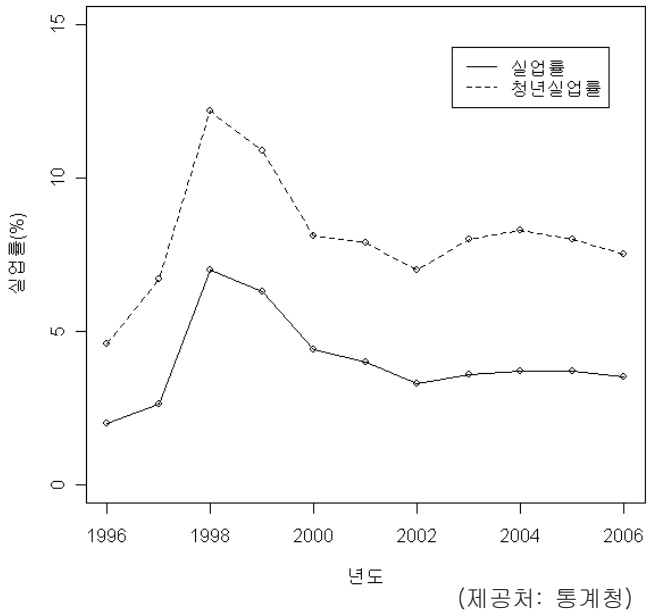
을 점할 것으로 예상된다.
이들 가운데 고시족과 공시족, 은둔 외톨이형 니트는 구직 활동을 하지 않아 사회적으로 소외될 가능성이 높다. 실증 분석 결과에 의하면 대체로 교육 수준이 높은 청년일수록, 그리고 여성일수록 청년 실업자화 확률은 낮아지지만 청년 니트화할 가능성은 높아지는 것으로 나타난다.
청년 실업자 수는 31만6000명인데 비해 청년 니트는 최소 133만8000명에 달해 청년 실업자 수가 청년 고용 문제를 제대로 반영하지 못하고 있다는 지적이 나오고 있다. 특히 구직 활동을 하지 않는 함정형 니트와 현실 회피형 니트의 경우 경기가 호전돼도 지속적으로 니트 상태에 머물러 있을 가능성이 높다.

이 글의 앞부분에 나오는 '청년실업률'이나 '비경제활동인구'란 무슨 뜻일까? 이를 알기 위하여 우선 통계청에서 실시하고 있는 경제활동인구조사(대상기간: 15일이 속한 1주간)와 관련된 용어를 일부 정리하면 다음과 같다.

- 15세이상 인구(Population 15 years & over) : 노동가능인구, 생산연령인구, 생산(활동)가능인구 등으로도 불림. 대한민국에 상주하는 만 15세 이상인 자 (군복무자, 교도소 수감자, 외국인 등은 제외), 생산가능인구 = 경제활동인구 + 비경제활동인구
- 경제활동인구(Economically active population) : 만 15세 이상 인구 중 조사대상기간 동안 상품이나 서비스를 생산하기 위하여 실제로 수입이 있는 일을 한 취업자와 일을 하지는 않았으나 구직활동을 한 실업자를 말함
- 비경제활동인구(Non economically active population) : 만 15세 이상 인구 중 조사대상기간에 취업도 실업도 아닌 상태에 있는 사람을 말하는 데 이들은 주된 활동상태에 따라 가사, 육아, 취업준비, 통학, 연로, 심신장애, 기타 등으로 구분됨
- 실업자(Unemployed person) : 15세 이상 인구 중 조사대상기간에 일할 의사와 능력을 가지고 있으면서도 전혀 일을 하지 못 하였으며 일자리를 찾아 적극적으로 구직활동을 하였던 사람으로서 즉시 취업이 가능한 사람을 말함
- 취업자(Employed person) : 15세 이상 인구 중 조사대상기간에 소득, 이익, 봉급, 임금 등의 수입을 목적으로 1시간 이상 일한 자를 말함. 18시간이상 일한 무급가족종사자도 포함 됨
- 경제활동참가율(Labor force participation rate) : 만 15세 이상 인구 중 경제활동인구(취업자+실업자)가 차지하는 비율을 말함. $\text{경제활동참가율}(\%) = \frac{\text{경제활동인구}}{\text{15세 이상 인구}} \times 100$
- 고용률(Employment-population ratio) : 만 15세 이상 인구 중 취업자가 차지하는 비율을 말함. $\text{고용률}(\%) = \frac{\text{취업자}}{\text{15세 이상 인구}} \times 100$
- 실업률(Unemployment ratio) : 실업자가 경제활동인구(취업자+실업자)에서 차지하는 비율을 말함. $\text{실업률}(\%) = \frac{\text{실업자}}{\text{경제활동인구}} \times 100$

우리는 노동시장의 고용상태를 실업률로 판단하곤 한다. 그러나 요즈음에는 청년실업률이 더 심각한 사회문제로 등장하고 있다. 청년실업률이 실업률의 약 2배로 고착화되어가고 있음을 다

음 [그림 1.19]에서 알 수 있다. 청년실업의 심각성은 청년실업률보다는 고용률 내지는 구인배율이 노동시장의 고용상태 현실을 더 잘 반영할 수 있다. 구인배율이란 기업체의 구인수를 일자리를 찾는 구직자수로 나눈 수치로서 인력수급의 지표로 쓰인다. 2006년 실업률은 3.5%로서 일본의 3.8%보다도 낮으나 구인배율이 0.48 밖에 되지 않는다.



(제공처: 통계청)

[그림 1.19] 실업률과 청년실업률

제 2 장

표본조사로 충분하다.

차 례

학습목표

어느 기업에서 행하는 기업이미지 광고의 효과를 알기 위해 광고를 맡고 있는 부서가 일반인의 의견을 묻는 조사를 시행한다고 하였을 때 먼저 이 부서는 어떤 집단(흔히 '모집단' 이라 한다.)에게 물어야 하고 어떤 방법을 이용하여 설문을 하여야 하는지에 대한 구체적인 틀을 세워야 한다. 그러나 우리가 관심을 가지고 있는 모집단의 크기는 일반적으로 매우 크다. 따라서 모든 개체를 접촉하여 우리가 알고자 하는 사항을 구하는 것은 시간과 비용이 많이 드는 매우 번거로운 작업이 된다. 대부분의 경우 모집단에서 표본을 추출하여 표본에 담겨져 있는 정보를 이용하여 모집단의 성질에 대해 알고자 할 것이다. 이것이 통계적 추론이다. 좋은 추론을 얻기 위해서는 모집단을 닮은 좋은 표본이 있어야 함은 물론이다. 제 2장에서는 모집단과 표본에 대한 일반적인 정의를 하고 사례를 통해 표본 추출방법 및 조사방법의 중요성에 대해 알아보도록 하자. 구체적인 표본추출방법은 제 9장에서 다루기로 한다.

2.1 모집단과 표본설계

표본에 관한 기본 용어

모집단(population) :

연구나 조사 목적이 정해지면 관심의 대상인 모집단이 정확히 정의되어야 한다. 예를 들어 모 회사에서 생산한 핸드폰의 품질에 대한 소비자 의견을 조사하기 위해 모집단을 정의하려면 최근 한 달간 생산된 핸드폰이 대상인지 1년간 생산된 핸드폰이 대상인지 또는 우리나라에서 생산된 제품인지 아니면 같은 제품으로 중국에서 생산된 것까지를 관심대상으로 하여 설명하고자 하는 것인지에 대한 구분을 명확히 해야 한다.

모집단(population) : 관심 대상이 되는 모든 개체의 집합.

모집단은 구성요소의 특징에 따라 유한모집단과 무한모집단으로 나눈다. 무한히 생산되는 알약이나 분필, 담배 등과 같이 모집단의 크기가 무한인 경우 무한모집단(infinite population)이 되고 로트(lot)단위로 포장된 1,000켤레 운동화나 100대 선풍기 등 유한한 경우에 유한모집단(finite population)으로 정의한다.

유한모집단(finite population) : 모집단의 크기가 유한한 경우 무한모집단(infinite population) : 모집단의 크기가 무한한 경우

표본(sample) :

모집단을 가장 잘 대표할 수 있는 일부를 뽑아 조사하거나 측정할 때 조사되는 일부를 표본(sample)이라 한다. 모집단 전체를 조사하지 못하고 표본조사를 하게 되는 이유는 다음과 같다.

- 전수조사가 시간적 또는 경제적 여건상 불가능한 경우
- 때에 맞추어 조사결과가 제시되어야 조치가 가능한 경우
- 관심 특성치가 파괴를 해야만 얻을 수 있는 자료인 경우
- 전수조사를 함으로써 오차 개입이 커져서 정확도를 오히려 떨어뜨리는 경우

표본(sample) : 실제 조사되거나 측정되는 모집단의 일부

통계적 조사의 목적은 미지의 모집단에 대한 정보를 알아보려고 하는 것이다. 모집단의 정보란 우리가 알고자 하는 모집단의 성질을 말하며 이러한 모집단의 성질을 모수(parameter)라 한다. 위에서 언급한 이유로 모집단 전체를 조사하는 것은 불가능하기 때문에 모집단의 성질은 표본이 가지고 있는 정보를 이용하여 추정하여야 한다.

표본설계의 기본 블록

모집단을 잘 대표할 수 있는 표본을 뽑는 방법은 매우 많다. 그러나 기본적으로 임의추출법(random sampling)은 많은 방법의 기본 블록을 형성한다. 간단하게 말하면 임의추출법이란 모집단의 구성요소 하나하나가 표본으로 뽑힐 확률이 같게끔 표본을 뽑는 방법이다. 따라서 이러한 방법에 의해 만들어진 표본은 모집단을 잘 대표할 수 있는 근거가 되는 것이다. 다음은 임의추출법을 적용한 예이다.

- 복권추첨 시 대상번호를 모두 통속에 넣어 골고루 섞어서 어떤 번호나 뽑힐 확률이 같은 상황에서 번호를 뽑는 경우
- 선거인명부를 놓고 일련번호를 매긴 뒤 난수표를 이용해서 일부 사람들을 뽑아 전화로 지지 후보에 대해서 조사하는 경우
- 100명의 회원 중 5명을 뽑아 오페라 입장권을 준다고 할 때 100명의 이름이나 번호를 카드에 적어서 통속에 놓고 골고루 섞은 후 눈을 감고 5명을 뽑는 경우

임의추출법(random sampling) : 모집단의 구성요소 하나하나가 표본으로 뽑힐 확률이 같은 상황에서 표본을 뽑는 방법

표본설계의 절차

일반적으로 표본설계는 다음과 같은 절차에 따라 진행된다.

1. 표본조사의 목표설정

표본조사에서의 첫 단계는 조사의 목적을 명확히 기술하는 것이다. 조사의 목적은 주어진 비용, 인력과 시간 등의 가용자원에 부합되는 현실적인 것이어야 한다.

2. 모집단의 정의

표본이 뽑히게 될 모집단은 애매하지 않고 명확히 정의되어야 한다. 모집단을 정의할 때는 “2008년 11월 1일 0시 현재 서울특별시에 상주하고 있는 모든 사람”과 같이 시간, 공간 및 속성이 규정되어야 조사에 포함되는 범위가 명확해 진다.

3. 추출프레임과 추출단위의 결정

표본설계를 할 때 추출틀(줄어서 프레임)을 마련하는 작업은 대단히 중요하다. 표본조사의 추출틀은 실제 표본이 뽑히는 추출단위의 목록이다. 추출틀에는 모든 기본단위가 중복되거나 누락됨이 없이 반영되어 있어야 한다. 또한 추출단위를 정할 때에는 미리 추출틀의 작성 가능성을 염두에 두어야 하고 조사현장에서 조사자가 쉽게 식별할 수 있도록 정해야 한다.

4. 표본추출방법의 선정

조사의 정확도, 표본크기, 추출틀, 조사비용, 조사일정 등 조사의 현실적인 측면을 고려하여 표본추출방법을 선정한다.

5. 현지조사

표본조사에서 목적달성의 상당부분은 현지조사의 정확성에 달려있다. 조사목적을 달성하기 위해서는 현지조사가 치밀한 감독 하에 정밀하게 이루어져야 한다.

6. 데이터 분석 및 보고서 작성

표본조사의 마지막 단계인 표본자료의 분석과 결과해석은 중요한 부분으로 주의 깊게 다뤄져야 한다. 보고서 작성단계에서는 조사결과의 기술과 실행 가능한 결론을 제시한다.

2.2 조사와 오차

조사오차 왜 생기나?

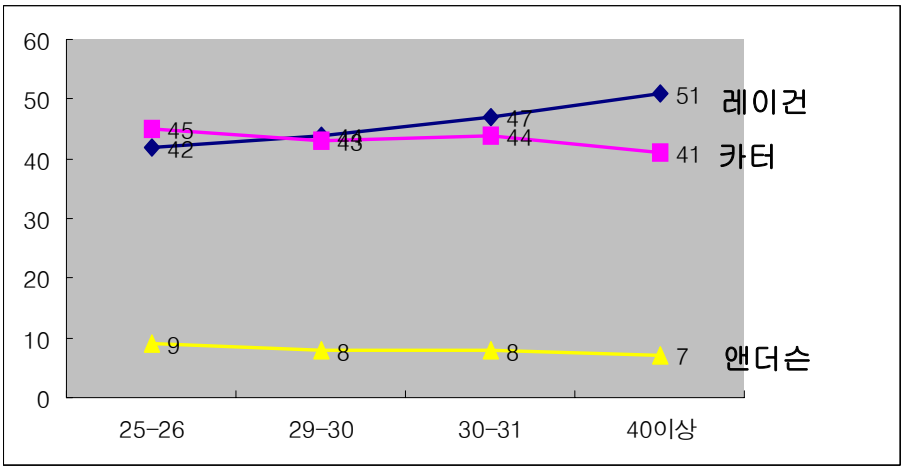
유력한 여론조사 기관들이 백중하리라는 예상과는 달리 미국의 40대 대통령선거에서 레이건은 카터를 51% 대 41%로 누르고 압도적 승리를 거뒀다. 최종선거예측에서 갤럽은 47%대 44%로 레이건의 우세를 예측했고 해리스와 ABC뉴스는 45%대 40%, CBS뉴스와 뉴욕 타임스는 44%대 43%로 레이건이 우세하다고 발표했다. 따라서 레이건을 기준으로 할 때 갤럽은 4%, 해리스는 6%, CBS는 7%의 오차를 보였다. 지난 39대 미국 대통령선거에서 갤럽의 평균오차가 1.2%, 이번 선거 이전까지의 4회의 선거에서 평균오차가 0.8%였음을 감안할 때 이번의 오차 4%는 꽤 큰 것이다. 왜 이런 오차가 생겼을까?

첫째, 갤럽이 11월3일에 발표한 최종조사결과는 투표 3일전에 조사된 것이다. 따라서 조사 후에 생기게 마련인 여론의 흐름을 파악할 수가 없게 된다.

현재 갤럽조사는 미국전역에서 2백 개의 조사지점을 무작위로 뽑아 1천5백 명 안팎을 상대로 면접원이 가정 방문하여 개별적으로 면접하는 방식을 취한다. 전국에서 설문지가 프리스턴의 본부에 도착하여 컴퓨터 통계처리과정을 거쳐 발표되기까지는 3일이 걸린다.

최근에는 투표전날 전화 면접하여 바로 전산처리하는 방법도 개발됐으나 여기에는 표본선택상의 오차와 응답상의 오차 등 전화면접의 한계가 있어 갤럽은 이 방식을 사용하지 않는다.

그러면 마지막 순간에 여론이 어떻게 변했는가를 도표로 살펴보자. 갤럽이 조사한 결과를 바탕으로 그린 [그림 2.1]에서 투표를 앞둔 11일간 여론의 흐름을 보면 10월 25, 26일에는 "카터우세"였던 것이 28일의 TV 토론 다음날인 29일에는 "레이건 우세"로 역전되었다. 10월 25일부터 투표일까지 레이건 지지율은 매일 평균 1.2%씩 늘어나고 있다.



[그림 2.1] 미국 대통령 후보 지지율 추이

그 원인은 TV토론이 결정적인 역할을 했고 다음으로 기대해온 이란에서의 인질석방이 투표날까지 이루어지지 못한데서 오는 실망과 좌절 때문으로 볼 수 있다. 이보다 앞서 행한 갤럽조사에서 보면 미국유권자의 3분의 1이상이 카터가 인질문제를 정치적으로 이용하고 있다고 믿고 있었다. 이 도표를 보면 레이건 지지율이 매일 1.2%씩 늘어나고 있으니 갤럽은 10월 30, 31일의 조사결과에다 투표일까지 매일 1.2%의 증가를 반영하여 레이건 지지율을 51%라고 해야 한다고 생각할 수도 있다. 그러나 여론조사는 일정시점에서 여론의 한 단면을 보여줄 뿐이며 예언이나 주관적인 분석을 하지 않는다.

둘째, 여론조사는 국민전체가 아니라 표본조사를 하기 때문에 1천5백 명을 대상으로 할 때 3% 안팎의 오차가 생길 수 있다.

셋째, 여론조사의 정확도가 투표율에 영향을 받는다는 사실을 지적할 수 있다.

1952년 이후 6차례에 걸친 대통령선거에서 기권은 40%에 이르렀다. 만일 기권자의 투표성향이 투표자의 비율과 같다면 문제가 없으나 과거 12회의 미국선거결과를 보면 기권자 중에는 민주당 지지가 80%가 많았다. 최근 들어 선거여론조사가 보다 정확해진 것은 갤럽조사소의 전문가인 폴 캐리 박사가 면접하는 사람 중 기권자를 미리 가려내는 방법을 고안해냈기 때문이다. 그러나 투표일의 날씨나 선거분위기의 변화 등 예측할 수 없는 요인으로 오차가 발생할 수도 있다. 이밖에도 조사응답과 실제투표가 달라질 수도 있고 부동표도 있어 오차를 낳는 요인은 너무나 많다. 40대 선거에서 조사기관들은 레이건의 우세는 예측했지만 투표 직전의 급격한 여론변동은 포착하지 못했다.

(통계적) 신뢰성과 타당성

조사결과는 조사대상 전체가 아니라 일부 즉 표본에 의거해 나온 결과이므로 어쩔 수 없이 표본오차, 즉 우리가 알고자 했던 성질과 표본에서 얻어진 값과의 차이인 표본오차가 항상 발생되기 마련이다. 통계적 신뢰성은 이러한 표본오차와 관련된 문제이다. 표본오차가 작으면 작을수록 조사결과는 신뢰성이 있다고 이야기 할 수 있는데 이러한 오차를 과학적으로 관리를 하기 위해서는 표본추출은 확률적으로 이루어져야 한다. 그러나 그 표본오차는 조사방법과 표본크기에 의하여 영향 받는다. 어떤 조사방법에 따라 표본이 추출 되었느냐에 따라 같은 비용과 시간으로도 훨씬 효율적인 결과, 즉 표본오차의 폭을 크게 줄일 수 있는 여지가 나온다. 그리고 표본의 크기는 크면 클수록 이런 표본오차의 크기는 줄어든다. 그러나 비용, 그리고 시간관계상 표본의 크기를 무작정 크게 할 순 없다. 또한 표본의 크기를 늘린다 하더라도 오차의 폭이 그렇게 줄어들지 않는 현상이 벌어지는 시점이 온다. 즉, 비용과 신뢰성의 타협(트레이드오프, trade-off)이 발생되는데 대략적으로 1,000명 내지 1,500명의 표본의 크기를 가지고 여론조사는 실시된다. 이에 대한 논의는 후에 한다.

통계적 타당성이란 통계조사가 과연 의도하는 것을 측정하고 있느냐 하는 것인데 이점에서의 우리 선거여론에서의 가장 큰 문제는 투표참여의향과 관련된 추계를 하지 못하고 있다는 것이다. 각 후보의 지지율을 추정하고자 하는 경우에는 우선 개별 응답자가 투표할 사람인지

그렇지 않은지를 알아내야 한다. 투표할 의사가 전혀 없는 응답자가 보인 반응을 각 후보 지지율에 산입해서는 안 될 것이다. 그러나 지금 우리나라에서 행해지는 일부 선거여론조사는 그렇게 하고 있지 않은데서 오차가 발생하게 된다.

표본조사에서 표본추출방법 및 표본의 크기에 따라 달라지는 신뢰성도 중요한 것이기는 하지만 타당성 역시 간과해서는 안 된다. 표본오차 보다는 비표본오차에 더 신경을 써야 한다. 그런데 여론조사 결과를 보면 이러한 문제는 무시한 채 통계적 유의성만 부각하는 경우가 간혹 있다.

2.3 표본크기

표본이 클수록 정확한가?

우리사회의 통계불신은 어제 오늘의 일이 아니다. 정부가 제시해온 각종의 부실했던 통계들은 국민에게 그 같은 불신감을 안겨준 한 가지 요인이었다고 할 수 있다. 특히 최근에 이르러서는 문화, 학술적인 사회조사의 조사방법과 조사결과를 두고도 그 정확도를 둘러싼 논란이 일고 있어 관심을 끌고 있다. 통계나 사회조사가 오늘을 사는 많은 사람들에게 너무나도 중요한 정보의 원천이며 정책결정이나 학문연구에 있어서도 결정적인 자료가 된다는 점에서 사회조사에 대한 국민의 정확한 인식과 계몽이 시급한 형편이다.

조사나 통계의 정확성은 아무리 강조해도 지나치지 않는다. 틀리거나 잘못된 조사는 하지 않는 것보다 더 큰 피해를 주기 때문이다. 그럼에도 정보나 각 기관에서 발표되고 있는 통계에 대해 불신이나 회의를 품고 있는 사람들이 어느 정도 있다.

조사의 정확성과 관련하여 조사를 실시하는 쪽이나 조사결과를 읽는 쪽에서 품고 있는 오해가 하나 있다. 그것은 조사의 정확성이 조사대상의 수에 비례할 것이라는 생각이다. 예컨대 천명을 조사 할 때보다 1만 명을 대상으로 하면 10배정도 더 정확하리라고 막연하게 생각하고 있는 점이다.

어떤 경우는 조사대상수가 많아질수록 오차가 클 가능성이 높아진다. 미국 갤럽조사의 일반여론조사에 있어서 대상자는 2천명 전후이다. 이 정도의 숫자만을 조사하여도 최근 4회에 걸친 대통령선거예측에서의 평균오차는 0.8%였다. 어떻게 이렇게 적은 수로 전 미국인의 의견을 정확하게 파악할 수 있는가? 이 원리는 조지-갤럽박사는 다음과 같이 쉽게 설명한다.

- "주부들은 냄비에 국을 끓여 간을 볼 때 잘 저은 뒤 한 두 숟갈 떠서 맛을 본다. 백배나 더 큰 가마솥에 국을 끓였더라도 잘 저어졌다면 백 숟갈이나 떠서 맛을 보지 않는다."
- "여기 흰 공, 검은 공 10만개를 7만개, 3만개의 비율로 잘 섞어놓고 또 다른 상자엔 천개의 공을 700개, 300개의 비율로 섞어놓았다. 각 상자에서 백 개씩 집어내도록 한다. 한쪽에 백배나 더 많은 공이 들어있는 것이 확실하나 집어내는 확률은 어느 쪽도 같은 것이다."

표본조사의 원리는 이와 같은 통계적 기본개념에 기초를 두고 있다. 표본의 크기와 오차와의 관계를 보면 6백 명에서 1천2백 명으로 늘려 조사하면 오차가 4%에서 2.9%로 감소하며 2천4백 명으로 늘릴 경우 2% 아주 적게 감소할 뿐이다. 이 경우 인구가 천만이든 억 만이든 관계없이 표본의 크기는 언제나 거의 일정하다. 이런 원리를 잘 이해하지 못하고 가능한 많은 수를 조사하는 것이 정확도를 재는 척도인 것처럼 오인되어 막대한 시간과 돈이 낭비되는 경우가 흔한 것이다. 또 어떤 사람은 전체를 하나도 빠짐없이 조사하면 100% 정확하다고 생각할지 모르나 이 말은 이론이지 현실은 아니다. 경우에 따라서 0.1% 오차라도 허용해서는 안 되는 아주 정확한 조사를 해야 할 때는 5만 명을 대상으로 할 수도 있다. 그러나 이때 그 실사과정에서 1%이상 오차가 발생할 확률은 더욱 커질 수 있음에 주목할 필요가 있다. 즉 조사는 수가 아니라 질이 문제인 것이다. 어떻게 전체를 대표하는 표본을 뽑는가가 중요하며 더욱 중요한 것은 설문지의 구성, 면접원의 태도 등 실사과정이 오히려 오차를 더욱 좌우하는 요인임을 명심해 둘 필요가 있다.

외국의 경우 일반국민의 여론이나 태도, 의식을 추정하는 데는 표본 2천 명 내지 3천 명이 보통이다. 그러나 우리나라의 경우 청소년 의식조사나 어떤 가치관조사는 1만 명 혹은 그 이상의 수를 대상으로 하고 있다. 만일 표본이 잘못 설계되면 1만 명이 아니라 1천만 명을 대상으로 했다 하더라도 엉터리가 된다. 농수산통계나 국민보건통계 같은 정확한 실태파악이 요구되는 조사는 표본오차에 얽매일 것이 아니라 조사 실시과정에서 일어날 수 있는 오차에 더욱 주목해야 한다.

조사를 실시하는데 있어서 가장 중요한 것은 시간과 비용이다. 적은 비용으로 많은 수를 조사하려고 할 때엔 머리수를 채우기 위해 건성으로 조사할 수 밖에 없게 된다. 글을 읽을 줄 모르는 사람에게까지 설문지를 맡겨두고 얼마 후에 회수하는 방식으로 적당하게 조사한다면 어떻게 되겠는가? 나아가 실제 조사가 어떻게 진행되는가를 제대로 지켜보지 않고 공문서 하나로 지방행정부서에 지시하고 중앙에서 집계하여 발표하는 통계조사는 1천만 명을 대상으로 했다 하더라도 믿기 힘든 것이다. 보다 정확한 조사나 통계를 위해서는 인원수의 많고 적음이 문제라기 보다는 표본의 설계, 면접방법, 면접원의 수준 및 감독 등 각 단계마다 숨어있는 오차에 더욱 주목해야 할 것이다.

표본크기의 결정

표본의 크기란 통계적으로 믿을만한 추정치를 얻기 위해 조사해야 하는 조사단위의 수를 의미하는데 통계조사에서 표본의 크기는 조사목적, 부분집단별 통계치의 필요성 여부, 전체적인 조사비용과 계획 등 여러 요인을 고려해서 결정해야 한다.

앞서 설명한 것처럼 표본크기가 늘면 표본오차는 줄지만 데이터 수집, 데이터처리, 분석 등 조사의 전 과정에서의 비용이 증가하게 된다. 거기에 표본크기가 늘어나면 조사원의 업무량과 조사과정에 대한 관리 감독도 어려워져서 표본조사에 따른 총 오차가 증가하게 되는 경우도 있다. 따라서 표본크기를 결정할 때는 전체적인 조사비용과 계획을 고려해서 결성하여야 한다. 또한 표본크기는 표본오차에 영향을 미치는 가장 중요한 요소이지만 좋은 통계조사가 되기 위

한 여러 조건 중의 하나라는 사실이다.

표본오차의 크기는 모집단 크기에 따라 좌우되는 것이 아니라 표본크기에 좌우된다는 점을 염두에 두어야 한다. 예를 들어 국회의원 선거구 한 지역에서 뽑은 200명의 표본을 조사한 경우와 우리나라 전체에서 뽑은 200명을 조사한 경우의 표본오차는 비슷하다.

여론조사를 의뢰하는 사람은 조사자가 내놓은 조사결과를 평가할 때 표본크기를 반드시 고려해야 한다. 전체 모집단을 대표할 만큼 표본이 충분히 확보되었는가? 조사자가 내놓은 조사결과의 차이가 오차의 한계 범위 내에 있는가? 등을 살펴보아야 한다. 다음으로 표본크기를 결정할 때 중요하게 고려해야 할 사항은 조사목적에 따라 조사하기 전 자료의 분석계획이다. 이 조사에서 필요로 하는 것이 모집단 전체에 대한 모수추정인지 아니면 모집단 내의 작은 부분집단에 대한 통계를 작성하는 것인지에 따라 표본크기를 결정하는 방법에서 차이가 생긴다.

예를 들어 어떤 연구자가 우리나라 성인을 대상으로 투표성향을 조사한다고 가정할 때 이 조사에서 우리나라 전체에 대한 신뢰성 있는 통계작성을 목적으로 할 수도 있지만 각 시도별로 어떤 차이가 있는지 알고자 하는 것이 더 중요할 수 있다. 여기서 각 시도별로 신뢰성 있는 통계자료가 필요하다면 단순히 우리나라 전국에 대한 신뢰성 있는 통계자료의 작성에 요구되는 표본보다 훨씬 많은 수의 표본이 필요할 것이다.

2.4 사례로 본 조사 이야기

한국인의 현재 삶에 대한 가치관과 국제 비교

우리나라 사람들의 대다수는 돈과 물질 뿐만 아니라 권위와 권력은 중시되지 않으면서 가정이 중심이 되는 사회를 가장 바람직한 미래 사회로 꼽고 있는 것으로 나타났다. 그러나 일상생활에 있어서는 절대다수가 “눈 코 뜰 새 없이 바쁘다”, “지겹다”, “우울하다”는 부정적인 경험을 우선순위로 들고 있어 “성취감”, “흥미로움” 등 밝은 면을 내세운 다른 나라 사람에 비해 부정적인 감정의 지배를 받고 있는 것으로 조사되어 대조를 보이고 있다.

이와 같은 사실은 한국갤럽조사연구소가 한국인의 인간가치관을 미국, 캐나다, 일본, 영국, 프랑스 등 세계 18개국과 동일한 질문항목으로 실시한 여론조사결과 밝혀졌다(1990. 12.20 한국일보). 이 연구를 위해 지난 10년간 갤럽국제조사기구의 협력을 받아 18개국 2만 4천명을 대상으로 여론조사를 실시했다.

한국인은 바람직한 미래사회로 돈이나 물질적 재산에 그다지 집착하지 않고(68%), 권위와 권력이 중요시 되지 않으며(72%), 개인능력개발(86%)과 기술개발(86%)의 중요성을 강조하고 있다.

반면, 현실의 삶에 대한 태도에 있어서 한국의 부정적인 견해는 다른 나라에 비해 크게 높은 것으로 밝혀져 주목된다. 한국인의 과반수 이상(53%)도 삶에 대한 자신감이 없다고 답변했는데 자신감의 정도를 10점 만점으로 환산했을 때 평균점수 5.11점으로 18개국 가운데 가장

낮았다.

저금을 할까 빚을 낼까?

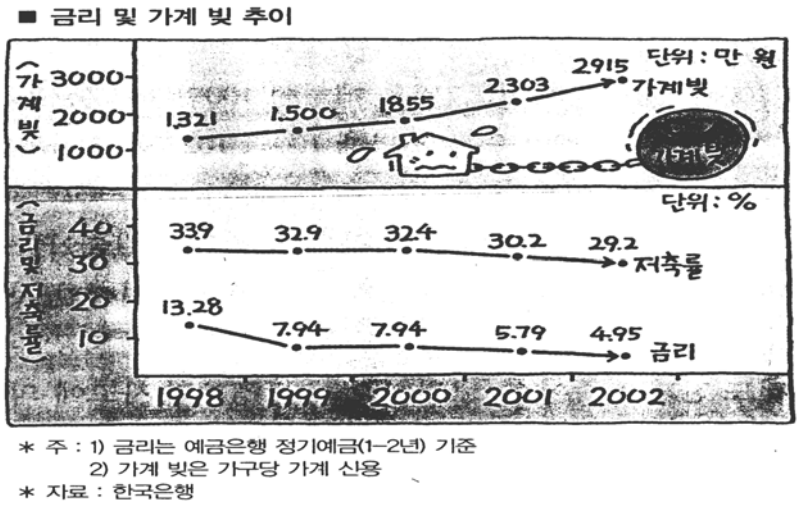
현재의 100만원과 1년 후의 100만원의 가치는 초등학생이라도 현재의 100만원의 가치가 더 크다는 것을 알 것이다. 은행에 100만원을 예금하면 1년 후에는 원금 100만원에 이자가 붙어서 100만원 이상이 되기 때문이다. 이자율이 연 10%라고 가정하면 이자가 10만원이므로 현재의 100만원은 1년 후의 110만원과 같다. 그리고 1년 후의 100만원은 현재의 91만원과 같다.

현재의 100만원이 1년 후의 100만원보다 가치가 크다면 현재의 100만원은 1달 후의 100만원보다 크고 아침의 100만원 또한 저녁의 100만원보다 크다고 할 수 있다. 실제로 은행 등 금융기관들은 한 나절이나 하루 정도 돈을 빌리거나 빌려줄 때도 이자를 주고받는다. 이때의 이자율을 콜금리라고 하는데 연 3.75%로 가정해 보자. 아침에 100만원을 빌린 후 오후에 갚는다면 원금 100만원에 이자 102.7원을 더하여 100만 102.7원을 갚아야 한다. 말하자면 아침의 100만원은 저녁때 100만 102.7원의 가치가 있는 것이다.

위에서 살펴본 것처럼 이자율을 적용하면 현재의 금액을 미래로 환산할 수 있고 또한 미래의 금액을 현재가치로 환산할 수도 있다. 그래서 사람들은 이자율을 보고 돈을 지금 쓰는 것이 좋은지 아니면 저금하는 것이 좋을지 또는 빚을 내는 것이 좋을지 등을 판단한다. 예를 들어 이자율이 높으면 돈을 당장 쓰기보다는 저금하는 것이 유리하고 빚을 내기보다는 어렵더라도 참는 것이 낫다고 판단된다.

사람들은 이자율에 따라 자신에게 유리한 쪽으로 행동한다. 금리가 높으면 소비보다는 저금을 많이 하고 반대로 금리가 낮으면 소비를 많이 하고 빚을 내는 것을 별로 두려워하지 않는다. 사람들의 이러한 합리적인 행동은 통계자료에도 그대로 나타나 있다. 다음 [그림 2.2]를 보면 금리가 지속적으로 낮아지면서 저축률은 감소하였고, 가계 빚은 같은 기간에 1,321만원에서 2,915만원으로 증가하였다. 이는 금리가 낮아지면서 사람들이 저축보다는 소비를 많이 하고 빚을 내는 것을 두려워하지 않은 결과라고 해석할 수 있다.

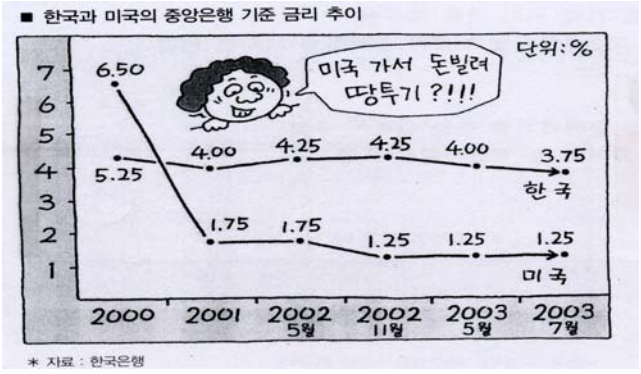
1998년 은행 정기예금 금리가 연 13.28%였을 때 1억원을 은행에 저금하면 1년에 이자를 1,328만원 받을 수 있었다. 하지만 2003년 3월의 은행 정기예금 금리는 연 4.30%에 불과하기 때문에 1억원을 은행에 저금하면 이자가 1년에 430만원 밖에 되지 않는다. 그래서 은행 예금은 저축수단으로 매력이 크게 떨어진다.



[그림 2.2] 금리와 가계 빚

금리는 은행예금이나 집값, 주가 등에만 영향을 미치는 것이 아니라 한나라의 경제에도 큰 영향을 미친다. 일반적으로 금리가 낮으면 가계의 소비와 기업의 투자가 증가하여 경기가 좋아지고 반대로 금리가 높으면 가계의 소비와 기업의 투자가 감소하여 경기가 나빠진다. 그래서 각국 정부는 경기가 나쁘면 금리를 내리고 경기가 과열되면 금리를 올리는 정책을 쓴다.

[그림 2.3]을 보면(잠깐! 이 그림은 문제가 있는 그림이다. 어떤 문제가 있는 지는 4장에서 배우게 될 것이다.) 미국의 연방기금 금리는 2000년 말 6.50%에서 2002년 11월 1.25%로 크게 하락했다. 이를 통해서 우리는 그동안 미국 경제가 좋지 않았다는 것을 알 수 있다. 금리인하는 경기를 살리기 위해 쓰는 정책수단이기 때문이다. 반면 우리나라는 미국에 비해 금리수준도 상대적으로 높았고 금리인하도 적었다. 이는 우리나라의 경기가 그 동안 그렇게 나쁘지 않았다는 것을 알 수 있다.



[그림 2.3] 기준금리 추이

자, 그러면 2003년 이후 미국의 연방기금 금리에는 어떠한 변화가 있었을까? 2008.01.19 문화일보 기사 중 일부를 보자.

앨런 그린스펀 전 미국 연방준비제도이사회(연준·FRB) 의장 재직 시절 미국의 기준금리인 연방기금금리는 사상 최저 수준인 1.00%(2003년 6월 26일~2004년 6월 29일)까지 떨어졌다. 금리가 이렇게 사상 최저 수준으로 낮아지면서 시중에 풀리는 돈(유동성)도 급격하게 늘었다. 이에 따라 시중에 풀린 돈이 부동산으로 몰리면서 부동산 가격이 급등했다. 이런 상황에서 미국 금융회사들은 신용도가 낮은 사람들에게까지 주택담보대출을 무분별하게 늘렸다. 서브프라임 모기지 액수가 엄청나게 늘어난 것도 바로 이 시기다. 문제는 2004년 6월 30일 미 연준이 다시 금리를 올리기 시작하면서 발생했다. 저금리 기조가 오래 지속되면서 부동산 등 자산 가격이 지나치게 올라 경제운용을 하기 어려워지자 정책당국이 금리를 올리기 시작했던 것이다. 그래서 2년 뒤인 2006년 6월 30일에는 미국의 기준금리인 연방기금금리가 5.25%까지 높아졌다. 이처럼 단기간에 금리가 급등하자 주택담보대출을 받은 사람들의 이자도 급격히 늘기 시작했다. 반대로 중앙은행이 금리를 올리면서 시중의 돈이 줄자 부동산 가격은 하락하기 시작했다. 이자 부담이 늘어도 부동산 가격이 오르면 부동산을 담보로 금융회사로부터 추가 대출을 받을 수 있지만, 이것마저도 불가능해졌다. 결국 신용도가 낮은 서브프라임 모기지에서부터 연체율이 높아지기 시작하면서 미국 전체가 서브프라임 모기지 부실을 둘러싼 금융위기에 휩싸이게 된 것이다. ... 미국의 경기침체를 촉발시켜 실물 부문으로 급속도로 확산되고 있으며, 전 세계를 경기 침체의 공포로 몰아넣고 있다. ... 서브프라임 모기지 부실 사태는 자산 가격의 지나친 상승이 경제에 얼마나 부정적인 영향을 미치는지 잘 보여준 사례이다.

이처럼 금리는 한 나라의 국가경제뿐만 아니라 세계경제에도 지대한 영향을 끼친다. 2003년에서 2006년까지 한국은행 콜금리 변동추이를 보면 다음 표와 같다. 미국에 비하면 상대적으로 금리의 변화가 적었음을 알 수 있다.

기간	콜금리
2003.05.13	4.00
2003.07.10	3.75
2004.08.12	3.50
2004.11.11	3.25
2005.10.11	3.50
2005.12.08	3.75
2006.02.09	4.00
2006.06.08	4.25
2006.08.10	4.50

그린스펀 전 미국 연방준비제도이사회(FRB, Federal Reserve Bank) 의장 재직 시절인 2003년 6월에서 2004년 6월까지 미국의 기준금리인 연방기금금리는 사상 최저 수준인 1.00%까지 떨어졌다. 금리가 이렇게 사상 최저 수준으로 낮아지면서 유동성도 급격하게 늘었다. 이에 따라 시중에 풀린 돈이 부동산으로 몰리면서 부동산 가격이 급등했다. 이런 상황에서 미국 금융회사

들은 신용도가 낮은 사람들에게까지 주택담보대출을 무분별하게 늘렸다. 당연히 서브프라임 모기지(비우량 주택담보대출) 액수가 엄청나게 늘어났다. 문제는 2004년 6월 FRB가 다시 금리를 올리기 시작하면서 발생했다. 저금리 기조가 오래 지속되면서 부동산 등 자산 가격이 지나치게 올라 경제운용을 하기 어려워지자 정책당국이 금리를 올리기 시작하여 2년 뒤인 2006년 6월에는 연방기금금리가 5.25%까지 높아졌다. 이처럼 단기간에 금리가 급등하자 주택담보대출을 받은 사람들의 이자도 급격히 늘기 시작했다. 반면 FRB가 금리를 올리면서 시중의 돈이 줄자 부동산 가격은 하락하기 시작했다. 이자 부담이 늘어도 부동산 가격이 오르면 부동산을 담보로 금융회사로부터 추가 대출을 받을 수 있지만, 부동산 가격이 하락하니 이것마저도 불가능해졌다. 결국 신용도가 낮은 서브프라임 모기지에서부터 연체율이 높아지기 시작하면서 미국 전체가 서브프라임 모기지 부실을 둘러싼 금융위기에 휩싸이게 되었고 미국의 경기침체를 촉발시켜 실물 부문으로 급속도로 확산되고 있으며, 전 세계를 경기 침체의 공포로 몰아넣고 있다.

서브프라임 모기지 부실 사태는 자산 가격의 지나친 상승이 경제에 얼마나 부정적인 영향을 미치는지 잘 보여준 사례이다.

우리를 우울하게 하는 숫자들

"나에게 우울한 숫자로 말하지 마라. 인생은 헛된 꿈일 뿐이다."(H. W. Longfellow)

우리는 식습관, 운동, 흡연, 음주습관, 직업과 일상생활에서 받는 스트레스 등이 어떤 식으로 좋고 나쁜 영향을 끼치는지에 대해 신문이나 잡지 또는 기타 다른 언론매체 등을 통하여 끊임없이 접하게 된다. 이런 정보는 증감을 나타내는 어떤 단위의 수치로 주어진다. 다음 [표 2.1]은 Cohen과 Lee(1979)로부터 얻은 자료표인데 몇 가지 우리를 우울하게 하는 수치들이 있다.

원 인	일 수	원 인	일 수
미혼(남성)	3,500	평균적 직업, 사고	74
왼손잡이	3,285	화기에 의한 사고	11
흡연(남성)	2,250	자연 방사능	8
미혼(여성)	1,600	병원 X-레이	6
30% 과체중	1,300	커피	6
20% 과체중	900	경구피임약	5
흡연(여성)	800	다이어트 음료	2
시가 흡연	330	자궁암 검사	-4
위험한 직업, 사고	300	집안의 연기경보	-10
파이프 흡연	220	자동차 에어백	-50
알코올	130	이동식 심장치료장치	-125

[표 2.1] 여러 가지 원인에 따른 기대수명의 감소

이런 숫자들을 어떻게 해석해야 할까? 그들은 어떤 메시지를 전하고 있는가? 그들을 어떻게

이용하여 보다 나은 행복을 위해 자신의 생활방식을 실현할 수 있을까?

표에 나타난 첫 번째 숫자 3,500을 보면 이 숫자는 미혼남성의 경우 기대수명이 감소한다는 것을 보여준다. 이 숫자를 사망자의 성, 결혼유무, 사망한 나이 등에 관한 사망기록정보에서 얻을 수 있다. 남성들의 사망기록에서 기혼인 사람들과 미혼인 사람들의 평균 사망나이를 간단하게 계산할 수 있다. 이렇게 계산된 평균 사망 나이의 차이가 3,500일(day)이라는 숫자이다. 이 사실을 아마도 독신으로 지내는 것에 따른 위험요소를 암시하는 것이며, 결혼제도를 좋게 얘기할 수 있는 근거가 될 것이다. 게다가 결혼을 빨리하면 10년을 더 오래 살 수 있다고 조언 할 수 있는 확실한 사례가 될 것이다.

그럼에도 불구하고 이런 사실이 누구에게나 적용될 수 있는 원인(결혼하는 것)과 결과(10년을 더 오래 사는 것)를 암시하는 것은 아니다. 어쩌면 특정 개인에게는 결혼하는 것이 자살행위처럼 여겨질 수도 있을 것이다. 즉 개인에 따른 특성에 따라 분류하여 사망기록에 대한 세밀한 표를 작성하면 더 유익한 정보를 얻을 수 있을 것이다. 집단이 다르면 기대수명의 감소와 증가로 각각 다른 값을 가질 것이다. 어떤 특정개인에 대하여 자기 자신의 성격을 분석하여 자신의 성격과 유사한 특성을 지닌 사람들 집단에서의 관련된 숫자를 참고할 수 있을 것이다.

왼손잡이가 오른손잡이보다 9년 더 일찍 사망한다는 사실을 표에서 볼 수 있는데 이것은 왼손잡이에게 유전적으로 나쁜 무엇인가가 있음을 의미하는 것일까? 아마도 그렇지 않을 것이다. 그 차이는 대부분의 시설들이 오른손잡이에 맞추어 만들어져 있어 왼손잡이인 사람들이 살아가는데 따른 불이익 때문일 것이다. 그러나 통계적 정보는 일어날 수 있는 위험에 대비하여 왼손잡이들이 자기 자신을 보호할 수 있는 어느 정도의 유용함을 준다.

일반적으로 평균은 모집단의 특성을 총괄적으로 암시해준다. 평균은 모집단들을 비교하는데 유용한 용도로 쓰인다. 따라서 우리는 한 달 평균수입이 1,000달러인 사람들의 집단이 한달 평균수입이 500달러인 다른 집단보다 형편이 더 좋다고 말할 수 있다. 그러나 평균은 개인의 소득 차이에 대해서는 어떠한 사실도 말해 주지 못한다. 예를 들어 개인소득이 20달러에서 100,000달러까지 차이가 나지만 평균차이는 1,000달러가 될 수 있다. 한 집단 내 개인의 소득차는 여러 집단을 비교하는데 적합하다. 대부분의 경우 평균과 변이의 측도는 실제로 의미있는 정보를 제공한다. 개인에 대한 판단을 내릴 때 평균 자체만으로는 신뢰하기 힘들며 유용하지 않다. 수영을 못하는 군인에게 그의 키가 강의 평균수심보다 크니까 강을 걸어서 건너도 된다고 조언해주는 경우를 생각해 보면 평균의 허망함을 알 수 있다.

선거 조사의 예

13대 대통령 선거 때 한국 최초로 개표 전 전국 남녀 유권자 32,290,416명(제주도, 도서지역, 부재자 제외)을 대상으로 2,234명의 표본을 3단 층화무작위추출법으로 1:1 가구방문 개별면접을 실시하여 예측한 결과가 [표 2.2]와 같았다(한국 갤럽 실시, 1987년 12월 16일 6시 발표).

후보자	노태우	김영삼	김대중	김종필	기타
실제선거결과	36.6	28.0	27.1	8.1	0.2
선거예측(A)	34.4	28.7	28.0	8.4	0.5
선거예측(B)	35.3	28.4	27.5	8.3	0.5

(A) : 제주도와 도서지역 투표자를 제외한 조사예측

(B) : 제주도와 도서지역 투표자 포함 조사예측(최종 공식발표)

[표 2.2] 13대 대통령 후보에 대한 사전조사결과 및 실제 선거결과(단위: %)

이 발표에서의 선거 예측값과 실제 투표결과는 당선자를 기준으로 2.2%, 2위와는 0.7%, 3위와는 0.9%, 4위와는 0.3%의 오차를 보였다. 각 후보자의 순위는 물론 적중했다. 이 선거예측이 이처럼 작은 오차로 적중할 수 있었던 것은 전문 조사원을 활용하여 지켜야 할 조사의 틀을 제대로 지킨데 있었고 무응답의 투표경향을 분류하여 예측한 것을 꼽을 수 있다고 알려져 있다.

조사는 처음부터 끝까지 오차와의 싸움이라고 할 수 있는데 모집단(population)의 규정, 묻는 방식(wording), 면접방법, 면접원의 성실성, 응답의 솔직성 등 어느 한 단계라도 잘못하면 조사의 결과는 믿지 못하게 된다.

통계학회에서 시행된 "선거예측과 오류"라는 심포지엄에서 지적된, 선거 즈음에 실시된 여러 조사의 문제점이 다음과 같이 지적되었다.

- 선거예측조사는 전국을 모집단(National Representative Survey)으로 해야 함에도 일부 지역을 대상으로 한 조사결과가 많았다.
- 선거여론조사에서 지지도를 밝히는 조사로서 전화조사나 자기기입식 질문지조사는 적합하지 않다고 할 수 있다. 우리나라 전화보급률이 71%인 당시 전화조사를 실시할 경우 29%의 전화 미보유자는 모집단에서 제외되며, 자기기입식 질문지의 경우 글을 쓸 수 없는 사람이 제외되어 조사결과가 고학력층으로 기울어지게 된다.
- 조사결과를 잘못 해석하거나 보도하는 경향으로 일부 지역을 대상으로 한 조사임을 밝혔음에도 전국 여론조사로 착각하거나 잘못 해석한 경우이다.
- 조사결과의 단순치가 곧 선거예측과 통하리라는 착각으로 무응답자의 투표성향을 밝혀내는 분석 작업이 이루어져야 했음에도 그렇지 못했다.
- 뽑힌 표본의 특성을 검토하고 모집단의 비율에 맞추어 가중치(Weighting)를 주어야 했었다. 무작위로 표본조사를 실시했을 경우라도 남녀의 비율이나 교육수준, 나이 등이 모집단을 있는 그대로 대표하지 않는 경우가 있다. 시계열 비교분석을 위해서는 가중치 작업이 필수적이다.

부동층의 해석에 대해

선거여론조사에서 관심이 되는 것은 부동층의 방향이다. 각 언론들은 이러한 부동층의 크기와 내용에 대해 지역별로 또는 각 계층별로 분석하는데 초점을 모았다. 부동층이란 여론조사시 "말할 수 없다, 모르겠다, 무응답" 등을 의미하는 것으로 14대 대통령선거 때 한국 갤럽의

경우 부동층이 어떻게 투표로 나타날 것인가를 다음 [표 2.3]과 같이 예측했다.

	선거결과(a)	조사결과(b)	a-b(구성비)	한국갤럽의 예측
김영삼	42.0	24.6	17.4(55.6)	39.5
김대중	33.8	24.1	9.7(31.0)	31.1
정주영	16.3	10.5	5.8(18.5)	15.7
박찬종	6.4	8.1	-1.7( )	12.4
기타	1.5	1.4	0.1(0.3)	1.2
무응답	-	31.3		
계	100.0	100.0		

주 : 구성비는 무응답률(31.3%)을 100으로 보았을 때 <선거결과-조사결과>의 비율을 나타낸 것.

[표 2.3] 부동층 분석(단위: %)

[표 2.3]에 의하면 무응답자 중 실제로 김영삼 후보에게 투표한 경우는 55.6%, 김대중에게는 31.0%, 정주영에게는 18.5%로 이동해서 부동층 가운데 김영삼 후보에 투표한 비율이 가장 높은 것으로 밝혀졌다.

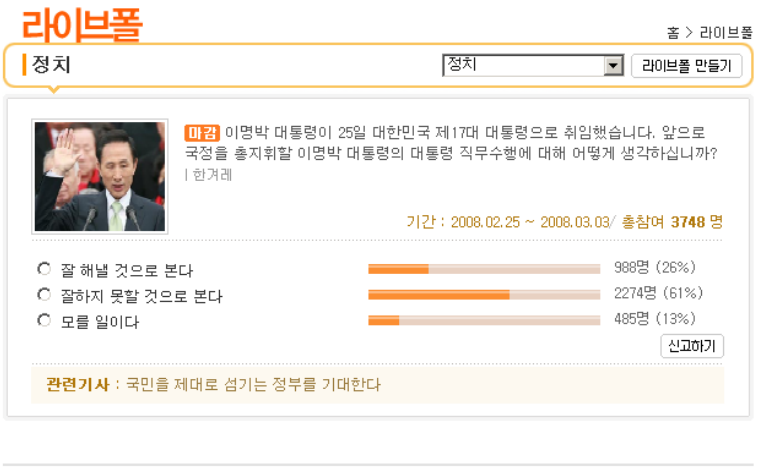
학습요약

제 2장은 모집단과 표본의 정의에서부터 출발하였다. 우리가 알고자 하는 모집단의 성질을 정확하게 파악하는 것은 우리의 관심이 되는 집단인 모집단을 정확하게 규정하는 것부터 출발한다. 모집단을 대표하는 표본을 추출하는 것 역시 매우 중요하다. 표본이 모집단을 대표하지 못한다면 표본에 의거 나오는 결과는 신뢰성이 전혀 없기 때문이다. 조사결과는 신뢰성 및 타당성의 문제가 있다. 신뢰성은 오차의 크기를 확률적인 표본 추출로부터 어느 정도 극복할 수 있으나 타당성은 다른 문제이다. 타당성은 우리가 원하는 성질을 향해 가고 있는 방향과 직결된 문제이기 때문이다. 이러한 내용은 제 9장에서 다시 언급이 될 것으로 표본추출의 기본적인 블록인 임의추출법에 대한 실습과 더불어 표본이 가지고 있는 정보의 성질을 배울 것이다. 제 2장에서는 이에 앞서 표본의 중요성을 몇 가지 사례와 더불어 살펴보았다.

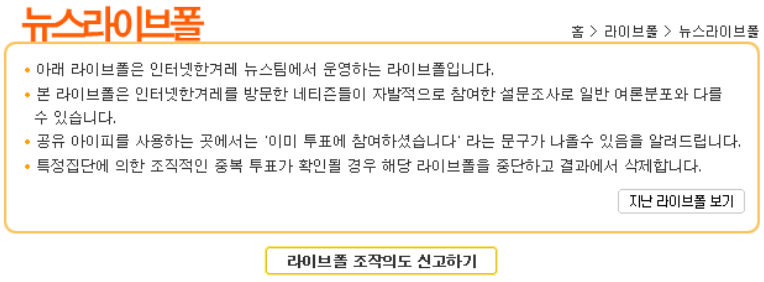
2장 연습문제

2.1 한 TV 프로그램에서 시청자들에게 국민들의 지대한 관심사 중 하나에 대해서 사회자가 질문을 던지고 ARS로 전화응답하여 줄 것을 부탁하였다. 100,000명이 전화응답하였고 그 중 '찬성'이 65%이었다. 같은 질문에 대하여 전국에서 임의로 추출한 성인 1,000명에게 물으니 '찬성'이 75%이었다. 일반적으로 100,000명의 전화응답자의 의견보다 임의로 선택된 1,000명의 의견이 국민들의 의견을 더 잘 반영하는 이유는 무엇인지 설명하여라.

2.2 다음 화면은 한겨레 인터넷신문에서 시행한 online-poll의 결과화면이다.



(1) 다음 화면은 online-poll의 화면에 나타나는 경고화면이다. 이러한 경고화면을 보여주는 이유는 무엇이라 생각하는가?



(2) 이러한 방식의 여론조사의 문제점은 무엇인지 두 가지 측면에서 생각하여 보라.

- 조사에 참여하는 사람들이 무작위로 추출된 표본인가?
- 인터넷 여론조사의 결과가 특정 계층을 중심으로 편향되어있지 않은가?

2장 실습문제

2.1 각종 매스컴에 나타나는 최근의 여론조사를 찾아내어 해당 여론조사에서의 모집단과 표본의 관계, 표본의 개수, 표본오차, 표본추출방법 등에 대하여 조사하고 서로 발표하고 토의하여라.

2.2 Census의 사례를 들고 표본조사를 하지 않는 이유에 대해서 토론해 보라.

2.3 통계청의 표본조사 사례를 3가지 정도 조사하고 각 사례의 모집단 정의, 모수가 무엇인지 살펴보라.

쉬어가기

조사결과를 볼 때 검토할 사항

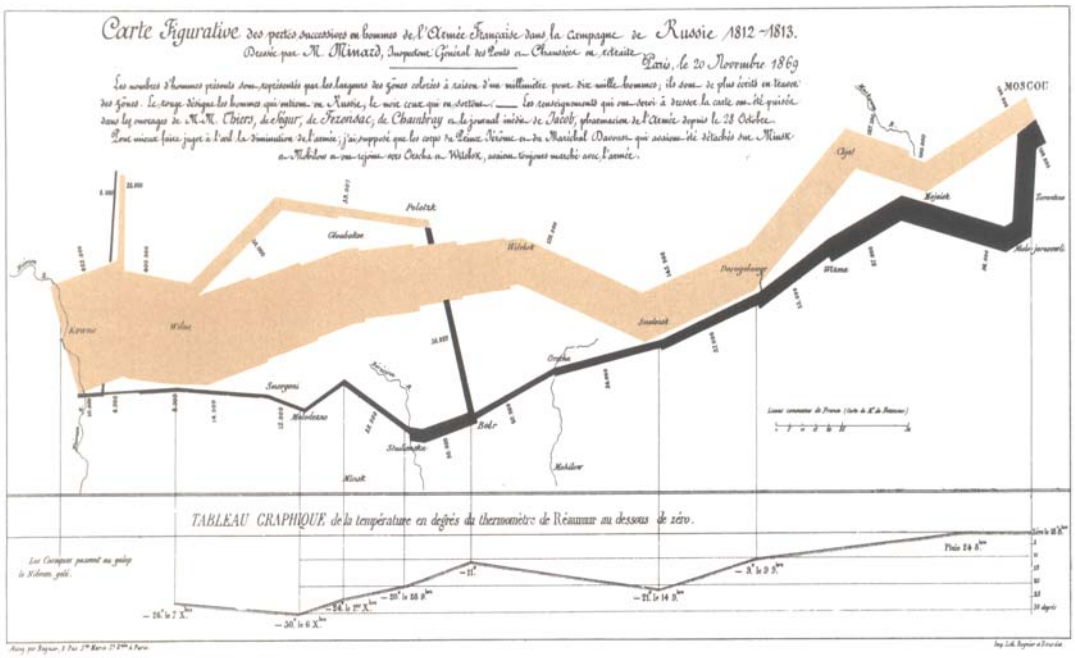
과학적 방법인 표본조사라 하더라도 드물게 사람을 속이는 경우가 발생하기도 한다. 조사 결과를 볼 때는 다음과 같은 사항을 검토해 본다.

1. 모집단은 무엇인가? 조사의 목표집단과 추출된 표본이 일치되는지를 확인해 보아야 한다.
2. 누가 조사를 시행하였는가? 전문기관에 위촉하지 않고 당사자가 직접 조사한 경우라면 의심해 봐야 한다.
3. 표본이 어떻게 추출되었는가? (조사하기 쉬운 편의표본을 사용하지 않았는가?)
4. 표본의 크기는 얼마인가? 오차한계는?
5. 어떤 조사방법이 사용되었으며, 무응답률은 얼마인가?
6. 언제 조사되었는가? 조사에 무언가 영향을 받을 수 있는 시기였는지를 검토해 봐야 한다.
7. 설문은 무엇이었는가? 를 검토해 보아야 한다. 조사는 설문 문구에 영향을 많이 받는다. 세금을 덜 내는데 찬성하십니까? 식의 질문은 곤란하다. 당연 찬성하지 않겠는가?

대중매체에 의한 조사 결과의 발표는 위에서 언급한 항목들이 반드시 보도되어야 한다. 그러나 여러 가지 이유로 누락되고 잘 지켜지지 않는 경우가 많으나, 접하는 사람이 비판적으로 선별 수용할 필요가 있다고 하겠다.

제 3 장

자료는 어떻게 구분되나?



차 례

학습목표

우리가 정보로 활용하는 자료들은 다양한 종류가 있고 나름대로의 독특한 특징들을 갖는다. 우리가 자료를 잘 활용하려면 이러한 자료의 종류와 특징들을 잘 파악하고 있어야 한다. 자료의 종류에 따라 통계적 분석 틀이 달라지기 때문에 자료의 특징을 파악하지 못하면 우리들은 통계적 오용을 범하기 쉽게 된다. 본 3장에서는 자료의 4가지 종류와 특징에 대하여 배우게 된다.

* 앞쪽 그림 설명: 나폴레옹은 1812년 6월 422,000명의 군사들을 이끌고 러시아를 침공하였다. 모스크바에 입성하는 데는 성공하였으나 러시아 군대의 끈질긴 저항으로 퇴각하고 말았다. 살아 돌아온 병사의 숫자가 10,000 정도였으니 40만명이 넘는 병사들이 제대로 싸우지도 못하고 얼어 죽거나 굶어 죽었다. 앞 쪽 그림은 Minard(1781-1870)가 나폴레옹군대의 러시아 침공과 퇴각을 나타낸 그래프이다. 뛰어난 시공간 그래프로서 총 6개의 변수(나폴레옹 군대의 병사수, 군대의 위치, 군대 이동 방향, 모스크바 퇴각 시의 날짜와 기온)를 하나의 그림으로 나타내고 있다. 침공 시는 연한색으로, 퇴각 시에는 진한색으로 표시하였다.

3.1 자료이야기

우리는 삶을 영유하며 수많은 자료를 만들어내고 있다. 다음과 같은 예들을 보자.

- 우체국 우편물의 이동: 미국의 우체국인 UPS(The United Parcel Service)의 database크기는 17 terabytes인데 미국 국회의사당 도서관 책 내용 분량에 해당한다.
- 식품의약품안전청이 실시하는 유전자조작농산물이나 새로운 의약품에 대한 검사
- 인터넷쇼핑몰이나 온라인전자장터에서 하루에 거래되는 거래건수
- 각 나라의 AIDS의 새로운 발생건수
- 각종 여론조사
- '비타민 C가 질병을 예방하는가?'라는 질문에 답하기 위한 실험
- '충실한 조기교육이 아이들의 학업성취도에 영향을 주는가?'를 알기 위한 조사
- 금융기관에서 잠재적 불량고객 선별하기

이러한 인간의 활동으로 인하여 자료가 발생하고, 쌓이는 자료들을 모은 데이터베이스를 분석함으로서 우리의 미래 삶에 대한 예측을 하게 된다. 우리가 정보로 활용하는 자료들은 다양한 종류가 있고 나름대로의 독특한 특징들을 갖는다. 자료를 잘 활용하려면 이러한 자료의 종류와 특징들을 잘 파악하고 있어야 한다. 자료의 종류에 따라 통계적 분석 틀이 달라지기 때문에 자료의 특징을 파악하지 못하면 우리들은 통계적 오용을 범하기 쉽게 된다.

예제 3.1 다음 글은 2008.02.11 조선일보 기사와 그래프를 나타내고 있다.

● 해양硏 정밀 환경조사

동해정 24개 지점 바다밑 수질 20%가 최하등급에도 못미쳐
납·카드뮴등 중금속 대거 포함 "자연 정화, 100년 넘게 걸려"

동해가 '폐기물 해양투기 구역'의 환경오염으로 골병들고 있다. 온갖 폐기물을 바다에 갖다 버리는 바람에 해양 투기구역 내 일부 지역의 밑바닥 부근 바닷물은 공업용수로도 쓰지 못할 만큼 오염된 것으로 나타났다. 해양 전문가들은 "지금 당장 폐기물 해양투기를 중단해도 이 구역이 자연적으로 복원되려면 최소 100년은 걸릴 것"이라고 경고했다.

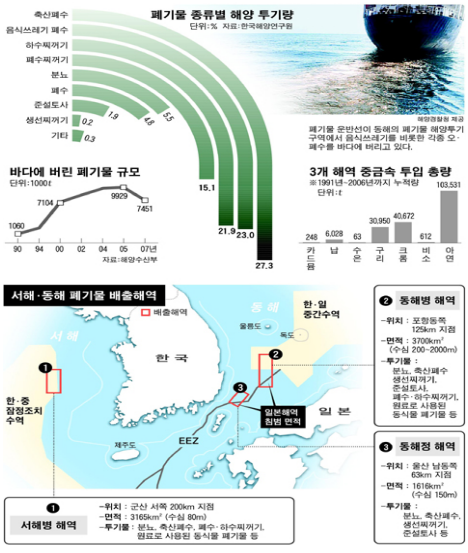
◆ 시화호에 버금갈 만큼 오염됐다

한국해양연구원은 작년 한 해 동안 울산 남동쪽 63㎞ 지점에 있는 '동해정 해양투기 구역'(총면적 1616㎢)에 대해 정밀 해양환경 조사를 실시했다. 동해정 구역이 해양오염방지법상 '폐기물 투기장'으로 설정된 것은 1993년이지만, 그 이전인 1988년부터 이미 해양투기는 시작됐기 때문에 이 구역에 대한 환경조사는 20년 만에 처음 이뤄진 것이다.

해양연구원이 동해정 구역 24개 지점에서 바다 밑바닥 근처의 수질상태를 조사한 결과, 총질소(TN) 기준으로 5개 지점(20.8%)의 수질이 '등급 외' 수준이었다. 해역수질등급은 총 질소의 농도가 1PPM 이하일 경우 1~3등급으로 분류하는데, 21%가량은 최하등급에도 못미칠 만큼 수질이 나쁜 상태라는 것이다.
9개 지점(37.5%)은 선박 정박용이나 공업용수로 쓸 수 있는 3등급이었고, 물고기들의 서식이나 양식, 그리고 해수욕에 적합한 1등급과 2등급 수질은 각각 1곳과 9곳으로 모두 합해 41.7%에 불과했다. 해양연구원 관계자는 "이 구역에 질소성분이 많이 포함된 음식쓰레기가 대거 버려졌기 때문"이라며 "또 다른 투기허용해역인 동해병 구역과 서해병 구역도 비슷한 수준으로 오염된 상태"라고 말했다.
또 다른 수질오염 지표인 화학적산소요구량(COD)도 사정은 마찬가지다. 등급 외(4PPM 초과)와 3등급(2~4PPM 이하)인 경우가 9개 지점(37.5%)에 달했다. 동해 바다 한가운데의 수질등급이 시화호의 수질(2006년 현재 COD 4.7PPM)에 버금갈 정도로 악화된 상황인 것이다.

◆ "자연적인 오염 개선에 100년 넘게 걸려"

문제 구역의 바닷물 수질보다 더 심각한 문제는 바다 밑바닥 퇴적물이다. 해양연구원에 따르면 1991년부터 작년까지 17년간 동해정 구역에 버려진 폐기물의 총량은 2352만3000t. 15t 트럭으로 치면 157만대 분량의 폐기물이 그간 바다 밑바닥에 깔렸거나 해수에 희석돼 바다 속을 떠돌고 있는 셈이다. 문제는 이런 폐기물 속에는 납과 카드뮴, 수은 같은 인체와 생태계에 치명적인 유해 중금속이 대거 포함돼 있다는 점이다.
미국 해양대기청(NOAA)에 따르면 이런 유해 중금속은 퇴적물 1㎏에 수십㎎만 포함돼도 생태계에 심각한 부작용을 초래한다. 해양연구원 관계자는 "최근 16년간 동해정 지역에만 4만2800t 분량의 중금속이 바다 밑바닥에 깔려 있거나, 플랑크톤이나 물고기의 몸속으로 들어갔을 것으로 추정됐다"며 "장래에 생태계에 어떤 영향을 끼칠지 가늠하기 어려울 정도"라고 말했다.
오염된 폐기물 투기해역을 복원시키려면 얼마나 많은 시간이 걸릴까? 포항에서 동쪽으로 125㎞ 떨어진 동해병 구역(총면적 3700㎢)에 대해 해양연구원이 '환경복원 타당성' 조사를 실시한 결과, 폐기물을 더 버리지 않더라도 원래대로 되돌리려면 최소한 100년 이상 걸릴 것이라는 분석이 나왔다. 광주대 양성렬 교수는 "동해와 서해에 대한 폐기물 투기를 즉각 중단해야 하며, 환경 복원을 위한 신기술 개발 같은 조치가 병행돼야 한다"고 말했다.



이 글을 보면 폐기물 종류별 해양 투기량(%), 해역수질등급, 각 해역의 면적 및 수심 같은 서로 다른 성격의 자료들이 나타난다. 이 자료들을 어떻게 구분할 수 있는지를 살펴보자! ■

3.2 자료의 종류와 특징

우리는 자료를 다음과 같이 4 가지의 척도(scale)로서 종종 구분한다.

1. 명목척도(명명척도, nominal scale)
2. 순서척도(서열척도, 순위척도, ordinal scale, rank scale)
3. 구간척도(등간척도, interval scale)
4. 비율척도(비척도, ratio scale)

1. 명목척도

2007.10.28 오후 3시경 경남 하동군 화개면 부춘리 검두마을 앞 국도 19호선 도로 상에서 화개 방면에서 하동읍 방향으로 달리던 관광버스가 마주오던 승합차 등 차량 4대와 잇따라 충돌, 모두 6명이 숨지는 사고가 일어났다. 이 사고는 관광버스가 마주오던 야생동물을 피하려다 일어난 것임으로 밝혀졌다. 이 사건을 통하여 야생동물이 도로에 나타나 피하는 과정에서 일어나는 사고와 야생 동물과 직접 부딪히는 사고(이를 통상 '로드킬'이라 한다.)의 위험성이 부각되고 있다. 이러한 사고는 전국 고속도로와 국도에서 점차 늘고 있다. 다음 기사는 2007.01.29 조선일보에 실린 로드킬 현황(2004년 7월에서 2006년 12월까지 지리산 4개 도로 현장에서 수행한 로드킬 조사)에 대한 기사이다. 기사 밑에 있는 자료들(종별 로드킬 현황, 법정보호종 로드킬 현황, 도로별 로드킬 현황)을 우리는 명목척도라 부른다.

'길 위의 죽음' 야생동물 로드킬 30개월 현장기록

지리산 주변 4개도로서 5769마리 희생
삵·소쩍새 등 법정보호동물도 311마리

로드킬(road-kill·차량 사고로 죽은 야생동물)의 실태를 2년6개월에 걸쳐 현장 조사한, 보기 드문 연구성과가 나왔다. 이 기간에 지리산 주변 4개 도로에서만 법정 보호종 311마리를 비롯해 모두 5769마리의 야생동물이 로드킬로 숨졌다. 일부 도로에선 정부기관 등이 발표한 공식 통계보다 최고 7배 이상 많은 로드킬이 발생했다. 그동안 로드킬 규모가 실제보다 턱없이 축소된 채 알려져 온 셈이다.

이 같은 내용은 서울대 환경대학원 박종화 교수팀이 2004년 7월부터 지난해 12월까지 지리산 4개 도로 현장에서 수행한 '로드킬 실태 조사 결과'에 담겼다. 박 교수팀은 오는 31일 서울대에서 열리는 '로드킬 현황과 대책' 세미나에서 조사결과를 공식 발표한다.

이번 연구는 차량사고가 생태계의 먹이사슬과 사람이 살아가는 환경에 어떤 영향을 끼치는지 규명하는 첫 단추가 될 것으로 전망된다.

사람의 입장에서는 고라니와 멧돼지 같은 덩치 큰 동물이 차량과 부딪힐 경우 운전자 안전이 크게 위협받기 때문에 오래전부터 로드킬 실태조사의 필요성이 제기돼 왔다.

조사 결과, 지리산을 두른 전체 도로(320㎞)의 절반에 못 미치는 4개 도로(88고속도로, 19번 강변·산업도로)에서 한 해 평균 2308마리씩, 모두 5769마리의 야생동물이 숨졌다. 포유류 1792마리, 양서류 1604마리, 조류 1329마리, 파충류 970마리 등 순이었다.

박 교수는 "법정보호종을 비롯한 야생동물이 이렇게 많이 숨진다는 것은 충격적"이라며 "사람의 안전과 생태계 보호를 위해 로드킬을 줄이기 위한 보호대책 마련이 시급하다"고 말했다.

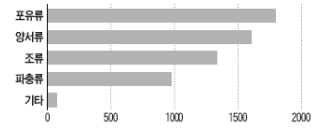
멸종위기종과 천연기념물로 지정된 법정보호 동물은 4개 도로에서 모두 16종, 311마리(5.4%)로 집계됐다. 이 중 삵(멸종위기종)이 103마리로 가장 많았고, 소쩍새와 큰소쩍새(천연기념물)가 각각 102마리와 49마리였다. 하늘다람쥐와 남생이, 솔부엉이, 수달 등 다른 법정보호종은 1~13마리씩이었다.

지리산과 섬진강 사이에 난 19번 강변국도(33㎞)는 야생동물에겐 '죽음의 도로'나 마찬가지였다. 전체의 절반이 넘는 3000마리가 이 구간에서 로드킬을 당했다. 88고속도로의 로드킬 규모는 그동안 한국도로공사의 공식 집계보다 훨씬 많은 것으로 집계됐다. 한국도로공사는 "1988년부터 지난해 6월까지 8년6개월간 88고속도로(함양~남원 간 44㎞)에서 모두 864마리가 숨졌다"고 밝혔지만 서울대팀 조사에선 2년6개월간 1845마리나 됐다. 연 평균으로 환산하면 각각 102마리와 738마리로 서울대팀의 조사가 7.2배 많았다. 도로공사 측의 그간 조사가 부실했다는 비판에 봉착할 수도 있는 대목이다.

법정 보호종 가운데 로드킬 1위로 집계된 삵은 4개 도로 가운데 유독 천은사에서 성삼재를 잇는 861번 산악도로에선 2년6개월간 단 한 건도 로드킬이 발생하지 않았다. 최태영 선임연구원은 "861번 도로가 높은 고개를 오르내리도록 건설돼 있기 때문에 차량의 통행속도가 아주 느리고, 통행량도 적기 때문으로 보인다"고 말했다. 최 연구원은 "대형맹수가 없는 우리나라에서 삵은 농촌지역 생태계를 지탱하는 거의 유일한 육식성 야생동물"이라며 "설치류 같은 동물의 포식자 역할을 하면서 생태계 유지에 중요한 역할을 한다는 점을 감안해 앞으로 삵이 돌아다니는데 문제가 없도록 대책이 필요하다"고 말했다.

박은호기자 unopark@chosun.com
박명진 인턴기자(고려대 영문과 2학년)

종(種)별 로드킬 현황 단위:마리



*서울대 환경계획연구소 (2004년 7월-2006년 12월)

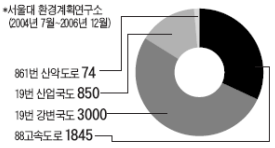
법정보호종 로드킬 현황 단위:마리

삵	103	수리부엉이	5
소쩍새	102	황조롱이	4
큰소쩍새	49	쇠족제비	3
하늘다람쥐	13	조롱이	3
남생이	11	기타(올빼미 등 5종)	5
솔부엉이	8	합계	311
수달	5		

*서울대 환경계획연구소 (2004년 7월-2006년 12월)

도로별 로드킬 현황 단위:마리

*서울대 환경계획연구소 (2004년 7월-2006년 12월)



명목척도 : 가장 원시적인 자료(변수)로서 자료의 정보가 가장 적은 자료이다. 서열을 알 수 없는 각 범주(category)에 대응되는 도수(frequency)만 주어지는 자료이다.

(특징) 1. 자료의 중심경향을 측정하는 측도로서 최빈값(mode)만 가능하고 중앙값(median)이나 산술평균(arithmetic mean)은 구할 수 없다.

2. 자료의 흩어진 정도를 측정하는 측도로서 다양성지수(diversity index)만 가능하고 분산(variance), 표준편차(standard deviation), 범위(range) 등은 계산할 수 없다.

다양성지수는 다음과 같이 정의되는 수치이다. 이 다양성지수 값은 제일 작은 값이 0이고 제일 큰 값이 1이 된다. 즉 다양성지수 값이 0이면 각 범주에 속하는 도수가 한 범주에 몰리게 되고 다양성지수 값이 1이면 각 범주에 속하는 도수가 모든 범주에 똑같게 된다.

$$J' = \frac{H'}{H'_{\max}} = \frac{\dfrac{nlogn - \sum_{i=1}^{k} f_i \log f_i}{n}}{\log k}$$

, 여기서 n : 표본의 크기, k : 범주의 개수, f_i : i번째 범주에서의 도수

예제 3.2 다음과 같은 서양인들의 머리색에 대한 자료에서

색깔	검정	갈색	빨강	금색	합계
도수	108	286	71	127	592

최빈값은 "갈색"이고 다양성지수는 0.899이다. 머리색의 분포를 다음과 같이 3가지로 가정하여 보면 각각의 다양성지수는 다음과 같다.

경우	색깔	검정	갈색	빨강	금색	합계	다양성지수
1	도수	108	286	71	127	592	0.899
2	도수	592	0	0	0	592	0
3	도수	148	148	148	148	592	1

두 번째 자료는 다양성지수가 0이므로 산포도가 제일 작다. 즉, 592명이 모두 머리색이 검정색으로 산포도가 0이 된다. 세 번째 자료는 다양성지수가 1이므로 산포도가 제일 크다. 즉, 592명이 4개의 머리색에 골고루 나뉘어 148명으로 동일하다. ■

예제 3.3 앞에서 언급한 로드킬 자료 중 법정보호종 로드킬현황을 다음과 같이 표로 작성하여 보자. 최빈값은 “삵”이고 다양성지수는 0.689이다. ■

법정보호종	마리수
삵	103
소쩍새	102
큰소쩍새	49
하늘다람쥐	13
남생이	11
솔부엉이	8
수달	5
수리부엉이	5
황조롱이	4
쇠족제비	3
조롱이	3
기타	5
합계	311

[표 3.1] 법정보호종 로드킬현황

2. 순서척도

다음 기사는 2007.11.20 매일경제 기사 내용이다.

대학생 2명 중 1명 “난 유행에 관한한 캔비족”

대학생 2명중 1명은 스스로 유행에 관한 한 '캔비족'이라고 생각하는 것으로 나타났다.
아르바이트 구인·구직 포털 아르바이트천국(www.alba.co.kr)이 대학생 380명을 대상으로 '유행 민감 정도'에 대한 설문조사에서 대학생 2명 중 1명은 자신이 유행에 민감하다고 생각하는 것으로 나타났다.
민감하다는 의견이 36%로 가장 많았고 보통(29%), 매우 민감(19%) 순이었다. 민감하지 않다는 의견은 전체의 16%에 불과했다.
유행에 민감한 이유로는 '예쁘고 멋져보여서'가 49%로 가장 많았고 '남에게 뒤지지 싫어서' 20%, '연예인처럼 되고 싶어서'라는 의견이 17%였다.
이런 최신 유행 스타일들은 외국 패션 잡지나 연예인을 통해 먼저 접할 수 있기 때문에 자신도 유행스타일을 따라가면 연예인들처럼 예쁘고 멋져 보일 것이라는 환상을 갖고 있는 것으로 나타났다.
이런 이들을 '캔비족'(can be·될 수 있다)이라고 부르는데 유명 연예인의 옷과 액세서리 등 패션을 모방하며 스스로의 가치를 연예인과 동격화하려는 사람들을 말한다.
패션·헤어스타일이 대표적이며 심지어는 다이어트 방법이나 성형까지 그 영역이 확대되고 있다.

‘유행에 민감한가?’에 대한 답변을 표로 정리하면 다음과 같다. 이 자료를 순서척도라 한다.

민감하지 않다	보통	민감하다	매우 민감하다	합계
16%	29%	36%	19%	100%

순서척도 : 명목척도에 순위라는 정보를 더 갖는 자료이다. 범주 사이의 서열이 존재하는 자료이다.

(특징) 1. 자료의 중심경향을 측정하는 측도로서 최빈값 외에 중앙값도 계산이 가능하다. 그러나 산술평균은 구할 수 없다.
2. 자료의 흩어진 정도를 측정하는 측도로서 다양성지수 외에 사분위수간 범위(IQR, inter-quartile range)가 가능하다. 그러나 분산, 표준편차, 범위 등은 계산할 수 없다.

예 3.4 다음 표는 100명의 panelist에게 식품을 맛보게 한 후 3개의 관능점수(sensory score)를 부여하게 하는 3점 판별법을 시행한 결과이다.

범주	나쁘다	그저 그렇다	좋다
도수	20	25	55

3 개의 범주 사이에 “좋다” > “그저 그렇다” > “나쁘다”라는 서열이 존재하게 된다. “좋다”라는 범주가 최빈값이면서 중앙값이 된다. ■

예 3.5 앞에서 언급한 유행민감도 조사에서 ‘유행에 민감한가?’에 대한 답변은 “민감하지 않다” < “보통” < “민감하다” < “매우 민감하다”라는 서열이 매겨진다. “민감하다”라는 범주가 최빈값이면서 중앙값이 된다. ■

3. 구간척도

다음 기사는 2007.09.27 매일경제 TV 기사 내용이다.

"소설"인 오늘 전국 곳곳 흐리고 눈비

오늘 아침 어제보다 기온이 크게 올랐습니다. 포근하다고 느껴질 정도로 며칠동안 이어진 영하권 추위를 무색하게 하는 날씨인데요. 또 오늘은 절기상 "소설"입니다. 보통 겨울의 첫 추위가 나타나는 때라고 하는데요. 그런데 오늘 추위는 없고 대신 소설을 알리는 비가 내리겠습니다. 지금도 중부지역에 약하게 비가 내리고 있는데요. 오늘 흐리고 곳곳에 오락가락 약하게 비가 이어지겠습니다. 내리는 비의 양은 많지 않겠지만 우산 챙기셔야겠습니다.
구름모습입니다. 중부지역에 비구름이 들어와 있습니다. 밤에는 이 구름이 남부지역까지 내려오겠습니다. 지역별 오늘 날씨입니다. 오늘 중부지역은 흐리고 비가 내리겠습니다. 호남과 경북지역은 흐려져서 밤 한때 비가 내리겠습니다.
현재기온 서울이 5도, 전주와 광주가 6도로 어제보다 기온이 5도 이상 크게 올랐습니다. 낮기온은 서울이 8도, 강릉 11도, 부산 16도로 어제와 비슷하거나 조금 더 높겠습니다. 내일 차차 날씨가 맑아지겠습니다. 이번 주말에도 포근한 날씨는 계속 되겠고요. 월요일에는 중북부지역에 비가 예상됩니다.
날씨였습니다.

기온에 해당되는 자료가 구간척도가 된다.

구간척도 : 순위척도에 ‘차이(difference)’라는 정보가 부여된 자료이다.
(특징) 대다수의 대표값과 산포도가 가능하다. 그러나 절대영점(absolute zero point)이 없기 때문에 비율(ratio)이 지켜지지 않는다. 절대영점은 0의 값이 없는 것을 나타내느냐의 개념이다. 온도가 0^0C라면 없다는 이야기가 아니라 단지 얼음이 어는 기준을 나타낼 뿐이다.

예 3.6 섭씨(centigrade) 온도계와 화씨(Fahrenheit) 온도계

섭씨온도 5^0C와 10^0C는 5^0C의 차이는 있으나 10^0C가 5^0C보다 2배 더 온도가 높다고 할 수 없다. 화씨온도로 바꾸면 당장 알 수 있다. 그러나 섭씨온도와 화씨온도 사이의 차이는 유지가 된다. 온도에서 비율을 비교하려면 절대온도(^0K)를 사용하여야 한다. ■

$$5^0C, 10^0C (d = 5^0C) \iff 41^0F, 50^0F (d = 9^0F)$$

$$20^0C, 25^0C (d = 5^0C) \iff 68^0F, 77^0F (d = 9^0F)$$

예 3.7 앞에서 언급한 날씨 기사에서 "부산의 낮기온 16^0C은 서울의 낮기온 8^0C보다 온도가 8^0C 높다"고 표현할 수 있으나 "부산의 낮기온 16^0C은 서울의 낮기온 8^0C보다 온도가 2배 온도가 높다"고 표현할 수는 없다. ■

4. 비율척도

다음 기사는 2007.09.27 한국일보 기사 내용이다.

"CEO 가족 사망하면 실적 떨어져"

최고경영자(CEO)의 사생활과 경영을 연결짓는 연구가 잇따라 외신에 공개되고 있다. 속설에 가까운 이런 연구는 CEO 역할을 과대평가한다는 지적을 낳고 있다.

먼저 CEO 가족의 애사는 회사 실적을 악화시킨 것으로 나타났다. 뉴욕대 다니엘 올펜존 교수 등 3명은 덴마크 7만5,000개 기업실적과 CEO 가족사망의 관련성을 추적했다.

그 결과 CEO의 자녀가 숨진 회사는 2년 뒤 총자산이익률(ROA)이 21.4% 하락했다. 숨진 자녀가 어릴수록 하락 폭은 더 컸다. 배우자가 사망하면 14.7%, 부모가 숨지면 7.7% 가량 이익률이 떨어졌다. 여성이 CEO인 회사에서 그 정도는 더욱 심했다. 특이한 것은 장모가 숨지면 반대로 이익률이 올라간다는 점이다.

1997년 아들이 살해되는 슬픔을 겪은 타임워너의 CEO 제럴드 레빈은 "나는 이에 영향 받지 않고 25시간 일한다"고 강조했다. 하지만 이후 3년간 상승하는 주식시장에서 타임워너 주가는 하락했다.

다른 연구에선 CEO가 대저택에 살면 회사 주가가 시장평균 이하로 하락했다. 미 애리조나주립대의 예르맥, 크로커 류 교수가 2004년 S&P500 기업의 CEO 집 크기를 조사한 결과 중간치는 520㎡였다.

그런데 집이 929㎡를 넘거나 대지가 4만㎡ 이상이면 회사 주가는 하락하기 시작했고, 3년 뒤 그 폭은 25%에 이르렀다. 이는 CEO가 현재 자리에 만족해 부를 만들기보다 즐기려 한다고 시장이 믿기 때문으로 풀이됐다. 대저택에선 정원, 인테리어 등에 많은 시간을 쏟아야 해 회사에 대한 관심이 전보다 줄어든다는 점도 지적된다.

그래서인지 97년 힐튼호텔 CEO 스테판 볼렌바흐가 1,194㎡의 대저택을 산 뒤 S&P500지수가 75% 오르는 동안 회사 주가는 70%나 빠졌다. 테넷 헬스케어의 경우 2005년 CEO가 929㎡ 저택을 구매한 이후

주가는 60% 이상 하락했다.
외부에서 CEO가 '베스트'로 평가 받거나 수상을 해도 회사 실적과 주가가 떨어지는 사례가 나타났다. 역시 스타 CEO는 경영보다 바깥일에 신경을 더 쓰기 때문으로 분석됐다. 언론을 좋아하고 자기애가 강한 CEO는 리스크를 감수하는 경향도 높아 회사 실적의 변동성이 큰 것으로 조사됐다.
또 세계 최대 CEO 조직체인 비스타지 회원을 대상으로 한 조사에서 CEO 1,582명 중 43%는 맏이였다. 맏이가 장악력이나 책임감 등에서 다른 형제보다 뛰어나 조직의 수장이 되는 경향이 높다는 것이다. 막내는 23%, 중간에 낀 형제는 33%에 불과했다.

집크기에 해당되는 자료가 비율척도가 된다.

비율척도 : 구간척도에 절대영점이 부여된 자료이다.

(특징) 자료 사이에 차이뿐만이 아니라 비율도 비교할 수 있게 된다.

예 3.8 몸무게가 80kg인 사람은 몸무게가 40kg인 사람보다 몸무게가 40kg 더 나간다(차이)고 할 수도 있고 2배 더 나간다(비율)고 할 수도 있다. ■

예 3.9 앞에서 언급한 최고경영자(CEO)의 사생활과 경영을 연결짓는 연구에서 집크기 929 m^2은 520m^2보다 409m^2 더 크다(차이)고 할 수도 있고 1.79배 더 크다(비율)고 할 수도 있다. ■

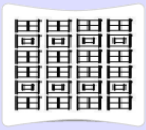
우리는 명목척도와 순서척도를 묶어 범주형자료(categorical data) 또는 질적 자료(qualitative data)라고 부르고 구간척도와 비율척도를 묶어 수치자료(numerical data) 또는 양적 자료(quantitative data)라고 부른다. 수치자료는 관측 가능한 값이 연속적인 연속형 자료(continuous data)와 관측 가능한 값이 이산적인(관측 가능한 값을 셀 수 있는) 이산형 자료(discrete data)로 구분하기도 한다. 자료가 어떤 종류이냐에 따라 자료에 대한 통계분석 방법이 달라지므로 우리는 자료의 성격을 잘 파악하여야 한다.

예 3.10 다음 한자들은 분야별 세계최고를 선정하는 웹사이트 '레코드컵'(recordcup.com/ranking/287/hardest_chinese_ever.html?&b=0)이 선정한, 가장 읽기 어렵고 이해하기 힘든 한자를 누리꾼들의 투표로 뽑은 한자들이다(2007년 10월 27일, 한겨레신문 참조). 이 한자들을 쓸 때의 획수가 이산형 자료의 예가 된다. 자 획수를 세어 보자! ■

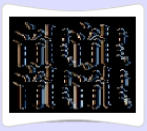
귀신쫓기, 굿

2007.11.26 순위는 다음과 같다. 3개가 바뀌었음을 알 수 있다.

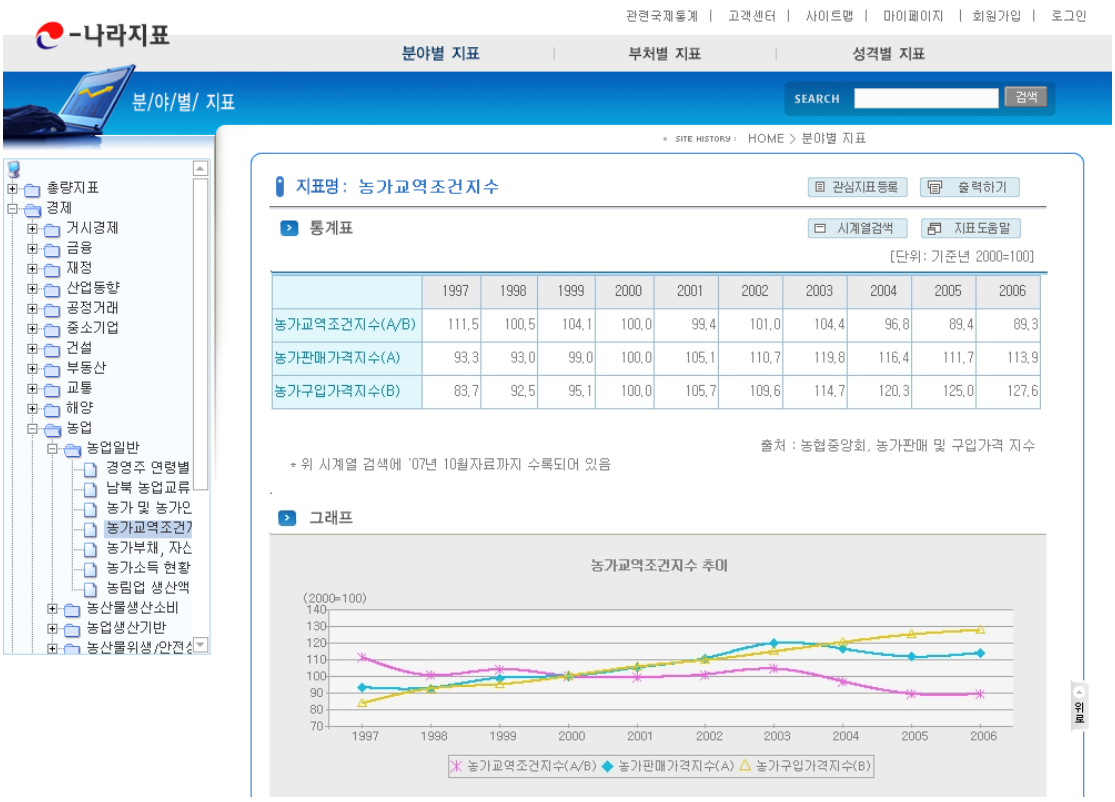
Showing **1st** to **5th** place

Rank	Title	Item	Score	Changes
1	This is the word "Exorcism"		131 Cups	↑ 1
2	This is the word "one"		88 Cups	↑ 1
3	This is the word "Flying Dragon"		79 Cups	↑ 1
4	This is the word "noodle"		71 Cups	↓ 2
5	This is the word "thunder"		45 Cups	↑ 1

Showing **6th** to **10th** place

Rank	Title	Item	Score	Changes
6	This is the word "cheat"		43 Cups	↓ 2
7	This is the word "flying squirrel"		36 Cups	↓ 1
8	This is the hardest character		25 Cups	↑ 1
9	This is the word "commend"		24 Cups	↑ 1
10	This is the word "peace"		23 Cups	↑ 1

예 3.11 다음은 통계청 홈페이지(www.nso.go.kr) 화면 오른쪽에 있는 자료광장 중 e-나라지표를 선택한 후 농업 메뉴에서 선택한 농가교역조건지수에 대한 화면이다. 이 지수들이 연속형 자료의 예가 된다. 2000년의 농가교역조건지수를 100으로 할 때 2006년의 농가교역조건지수는 89.3으로 10.7만큼 줄어 농가교역조건이 악화되었음을 알 수 있다. ■



학습요약

1. 우리는 자료를 다음과 같이 4가지의 척도로서 종종 구분한다.

(1) 명목척도: 가장 원시적인 자료(변수)로서 자료의 정보가 가장 적은 자료이다. 서열을 알 수 없는 각 범주에 대응되는 도수만 주어지는 자료이다. 자료의 중심경향을 측정하는 측도로서 최빈값만 가능하고 중앙값이나 산술평균은 구할 수 없다. 자료의 흩어진 정도를 측정하는 측도로서 다양성지수만 가능하고 분산, 표준편차, 범위 등은 계산할 수 없다. 성별은 명목척도의 한 예이다.

(2) 순서척도: 명명척도에 순위라는 정보를 더 갖는 자료이다. 범주 사이의 서열이 존재하는 자료이다. 자료의 중심경향을 측정하는 측도로서 최빈값 외에 중앙값도 계산이 가능하다. 그러나 산술평균은 구할 수 없다. 자료의 흩어진 정도를 측정하는 측도로서 다양성지수 외에 사분위수간 범위가 가능하다. 그러나 분산, 표준편차, 범위 등은 계산할 수 없다. 리커트 척도(아주 좋아한다, 좋아한다, 보통이다, 싫어한다, 아주 싫어한다 등)가 순서척도의 한 예이다.

(3) 구간척도: 순위척도에 '차이'라는 정보가 부여된 자료이다. 대다수의 대표값과 산포도가 가능하다. 그러나 절대영점이 없기 때문에 비율이 지켜지지 않는다. 온도는 구간척도의 한 예이다.

(4) 비율척도: 등간척도에 절대영점이 부여된 자료이다. 자료 사이에 차이뿐만이 아니라 비율도 비교할 수 있게 된다.

2. 자료의 종류에 따라 통계적 분석 틀이 달라지기 때문에 자료의 특징을 파악하지 못하면 통계적 오용을 범하기 쉽게 된다.

3장 연습문제

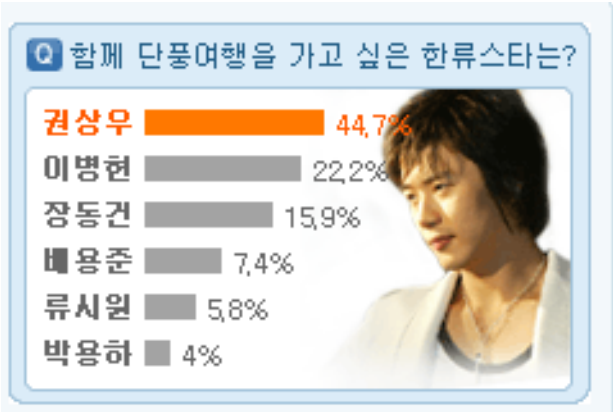
3.1 (명목척도, 순서척도, 구간척도, 비율척도)

(1) 다음 기사(2007.10.19 메디컬투데이/뉴시스 제공)에 나타나는 데이터들은 4개의 척도(명목척도, 순서척도, 구간척도, 비율척도) 중 어느 척도에 속하는지 설명하여라.

162 cm 이하 키 작은 남성 '불행하다'

키가 작은 사람들이 평균 신장의 사람들에 비해 정신기능과 신체 건강 저하를 호소할 가능성이 큰 것으로 나타났다.
덴마크 연구팀이 14,000명 이상을 대상으로 진행한 연구결과 162cm이하 남성과 151cm이하 여성들이 이 보다 키가 큰 사람들에 비해 웰빙지수가 저하된 것으로 나타났다.
이번 연구에서 사람들은 키가 더 클수록 자신의 건강이나 웰빙에 대해 긍정적으로 생각하는 것으로 나타났다.
연구결과 남성의 경우 7cm, 여성의 경우 6cm 키가 더 클 경우 건강과 연관된 삶의 질이 6.1% 향상될 수 있는 것으로 나타났다.
연구팀은 이 같은 효과가 비만인 사람이 체중을 10-15kg 가량 줄이는 것과 유사하다고 말했다.
연구팀은 키가 작은 사람들이 교육, 고용, 인간관계에 있어서 정상키의 사람들보다 어려움을 더 많이 겪는다고 말하며 이번 연구결과 키가 작은 것이 신체 건강의 저하를 직접 유발하지는 않지만 키가 작은 사람들일수록 건강과 연관된 삶의 질이 떨어졌다고 스스로 느낄 가능성이 큰 것으로 나타났다고 말했다.
'임상내분비학저널'에 발표된 이번 연구결과를 바탕으로 왜 키가 작은 사람일수록 자신이 건강에 대해 나쁘게 느끼는지를 정확히 알기 위한 추가 연구가 필요하다고 말했다.
이에 대해 전문가들은 키가 더 클수록 더욱 건강하고 더 오래 산다는 생물학적 증거들이 많다며 영양상태가 더 좋고 질병이 없는 등 건강할수록 성장 상태가 좋아 키가 클 가능성이 크다고 말했다.
전문가들은 그러나 키가 크다는 것이 반드시 건강하다는 것을 의미하는 것은 아니라고 말했다.

(2) 다음 그림은 2007.10.21 스포츠조선이 시행한 online poll의 결과를 나타나는 화면이다. 4개의 척도(명목척도, 순서척도, 구간척도, 비율척도) 중 어느 척도에 속하는가?



(3) 다음 표는 한국직업능력개발원이 지난 7~8월 한국・일본・미국・독일 등 4개국 취업자 1,200명을 대상으로 10개 주요 직업에 대한 직업의식을 조사한 결과이다(2007.11.20 서울경제신문 참조). 직업위세는 직업의 권위・중요성・가치・존경에 대한 인식 정도를 나타내는 용어이다. 표에 나타나는 데이터들은 4개의 척도(명목척도, 순서척도, 구간척도, 비율척도) 중 어느 척도에 속하는지 설명하여라.

● 국가별 직업위세에 대한 평가점수 및 순위

구분	한국	일본	미국	독일
국회의원	4.21(1)	3.88(1)	3.15(6)	3.54(4)
약사	3.76(2)	3.57(2)	3.70(3)	3.71(2)
중고교사	3.67(3)	3.17(5)	3.50(5)	3.24(7)
중소기업간부	3.35(5)	3.09(6)	3.57(4)	3.60(3)
기계공학엔지니어	3.18(7)	3.38(4)	3.72(2)	3.49(5)
소프트웨어개발자	3.39(4)	3.49(3)	3.82(1)	3.77(1)
은행사무직원	3.30(6)	2.99(7)	2.99(7)	3.39(6)
공장근로자	2.03(8)	2.32(8)	2.79(9)	2.17(9)
음식점종업원	1.77(9)	2.17(9)	2.72(10)	2.18(8)
건설일용근로자	1.56(10)	1.77(10)	2.91(8)	1.61(10)

※자료:한국직업능력개발원. 괄호 안은 순위

3.2 (이산형 자료와 연속형 자료)

다음 기사(2007.12.23 연합뉴스 제공)에 나타나는 데이터들은 이산적 자료와 연속적 자료 중 어디에 속하는지 설명하여라.

로드킬 급감..네비게이션 안내방송 '효과'

대구지방환경청이 야생동물 로드킬(Road Kill)을 감소시키기 위해 올해 시범사업으로 실시한 네비게이션 로드킬 안내방송이 큰 효과를 거둔 것으로 나타났다.

대구환경청은 지난 7월부터 4개월간 일부 운전자들에게 로드킬 빈발 구역을 사전에 알려주는 네비게이션 안내 방송을 한 결과 로드킬 개체 수가 대폭 감소한 것으로 나타났다고 23일 밝혔다.

환경청은 소프트웨어를 업그레이드한 운전자들에게 대구·경북지역 31개 구간에 대한 로드킬 정보를 제공했으며, 안내 방송이후 로드킬 감소여부를 모니터링했다.

모니터링 결과 24마리의 야생동물이 목숨을 잃은 것을 확인했는데 이는 지난해 같은 기간의 80마리에 비해 70% 감소율을 보였다. 환경청은 이에 따라 시범 실시한 네비게이션 업체 외에도 다른 업체들의 협조를 얻어 로드킬 위험 구간 안내 방송을 전국적으로 확대 실시할 방침이다.

대구환경청 관계자는 "로드킬이 급격히 감소한 것은 차량 운전자들이 네비게이션을 통해 나오는 안내방송을 듣고 야생동물의 출현에 대비, 운행속도를 줄이거나 전방을 주시하는 등 안전운전을 한 결과로 분석된다."고 말했다.

3.3 (명목척도, 순서척도, 구간척도, 비율척도)

다음 설문지에 나타나는 척도들에 대하여 설명하여라.

안녕하십니까? 바쁘신 시간 가운데 설문에 응해 주셔서 대단히 감사 합니다. 본 조사는 OOO 놀이시설 이용자의 서비스 만족을 조사하기 위한 설문입니다. 본 자료는 통계자료 이외에 다른 어떤 목적으로도 이용하지 않을 것임을 약속합니다. 여러분의 성의 있는 자료 작성이 통계에 있어 귀중한 자료가 될 것입니다. 감사합니다.

※작성요령
각 질문에 대한 만점은 5점이며 최하점은 1점입니다. 여러분이 생각하시는 만족도 점수에 표시를 해주시면 됩니다.

⑤아주 좋다 ④좋다 ③보통이다 ②나쁘다 ①매우 나쁘다
또는
⑤매우 그렇다 ④그렇다 ③보통이다 ②아니다 ①매우 아니다

1. 종업원들은 친절하고 고객편의를 위한 노력을 보였다.
2. 볼거리와 즐길 거리가 많이 있었다.
3. 각종 편의시설이 잘 정비되어 갖추어져 있었다.
4. 각종 시설은 안전했으며 안내도 잘되어 있었다.
5. 음식 및 시설 이용료 가격은 적당하였다.
6. 시설이용을 위해 기다리는 불편함이 없었다.
7. 과거 경험은 OOO 놀이시설 방문에 도움이 되었다.
8. 일상에서 벗어날 수 있는 시간이 될 수 있었다.
9. 주변 사람의 방문경험이 긍정적 영향이 되었다.
10. 교통수단과 이동시간 등 접근성이 편리 했다.
11. OOO 놀이시설 이용에 대해 만족한다.
12. OOO 놀이시설 이용을 다름 사람에게 추천할 생각이 있다.
13. 다음 기회에 다시 이용할 의사가 있다.

개인 신상에 관한 사항입니다.
(본 자료는 통계자료 외에 다른 용도로 사용하지 않음을 약속드립니다.)

1. 당신은 성별은 무엇입니까?　　① 남　　② 여

2. 당신의 연령은 어떻게 되십니까?

3. 당신의 학력은 어떻게 되십니까? ①중졸이하 ②고졸 ③대학생 ④대학졸업 ⑤대학원이상

4. 현재 거주하시는 곳은 어디 입니까?

①서울 ②경기 ③광역시 ④강원, 충청, 전라, 경상권역 ⑤기타(해외 등)

5. 주로 누구와 방문 하십니까?　①가족 및 친지 ②친구 ③연인 ④업무 및 단체 ⑤기타

6. 최근 10년 내 몇 번 방문하셨습니까?

7. 한 달 수입은 얼마이십니까?
①100만원미만　②100만원 ~ 199만원 ③200만원 ~ 299만원 ④300만원 ~ 399만원
⑤400만원이상

감사합니다.

3장 실습문제

3.1 다음 화면은 에너지관리공단 홈페이지 내의 연비·등급표시(bpm.kemco.or.kr/transport) 메뉴 하의 연비등급확인 부메뉴를 선택하였을 때 나타나는 화면이다.

이 자료를 보면 10개의 변수(차모델명. 업체, 유종, 배기량, 공차중량, 변속형식, 연비, 등급, CO_2발생량(g/km), 연간예상연료비(원))가 있다. 각 변수에 대하여 정렬(sort)이 가능하다.

(1) 각 변수가 4개의 척도(명목척도, 순서척도, 구간척도, 비율척도) 중 어느 척도에 속하는 지 지적하여라.
(2) 각 변수들이 어떤 값들로 형성되어 있는지 말하여 보아라. 예로 등급은 1에서 5등급까지 있다.

3.2 www.chinese-tools.com/characters/new.html에 나타나는 한자들의 획수를 세어 보라. 어느 척도에 속한다고 하였는지 기억하는가?

쉬어가기

현대사회의 몬도가네 (Mondo Cane)

다음 기사는 2007.05.13 일간스포츠기사이다. 여성에게 성형이 일반화되어가는 추세에 맞추어 유명인을 대상으로 성형비용에 대하여 언급한 기사이다.

빅토리아 베컴 성형 견적은 $37,100

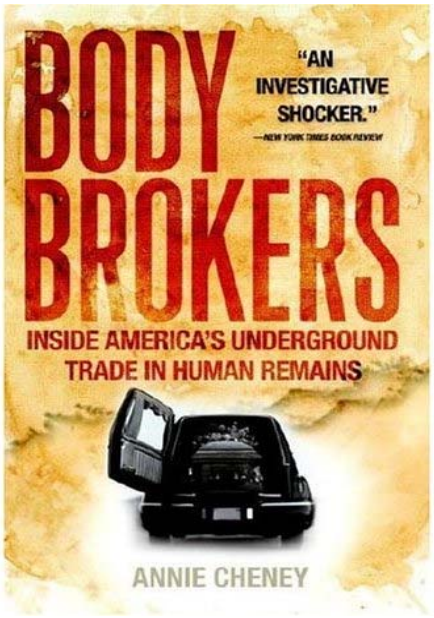
'US 위클리' 최신호가 특집으로 게재한 '포시가 되는 성형' 견적 기사. 왼쪽이 1993년 데뷔 당시의 자연 그대로의 모습이고 오른쪽은 현재의 화려한 사진이다.

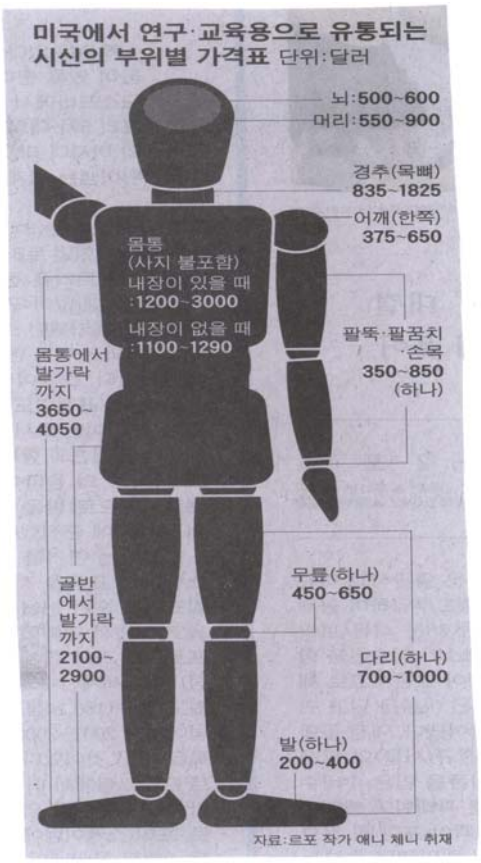
1993년 18세의 나이에 스타가 되고자 사진을 찍었던 댄서가 1994년 '스파이스 걸'이 되더니 현재는 세계적인 패션 아이컨으로 변신해 있다.
1999년 축구 스타 데이비드 베컴과의 결혼으로 날개를 단 빅토리아 베컴을 두고 하는 말이다. 그녀의 별명은 '포시(Posh)'. 포시가 되려면 일단 성형에만 최소한 3만7100달러(약 3,500만원)를 투자해야 한다는 견적이 나왔다.
'US 위클리' 최신호가 1993년 빅토리아 베컴의 성형 전 모습과 현재의 사진을 비교하면서 얼굴부터 가슴까지 각 부위별 성형 비용을 산출해 소개했다.
거의 대부분이 달라졌다. 뺨에는 3,000달러, 입술은 2,000달러, 머리와 피부에는 3,100달러, 눈은 8,000달러, 코는 11,000달러, 가슴은 10,000달러의 성형 비용이 들어간 것으로 집계됐다.
그러나 더 중요한 것은 평소에 쓰는 비용이다. 그녀는 1년에 의상비로 1억원에 가까운 기본 10만 달러를 지출하고 있다. 빅토리아 베컴은 인터뷰에서 '나는 내 모습이 결코 늙어가도록 놓아두지 않겠다. 주름도

없을 것"이라고 밝혔다.
가슴을 키우는 수술을 한 것도 부인하고 있기는 하다. 그러나 할리우드의 성형 전문가들은 정확하게 견적을 내놓았다.

다음 그림은 우리나라에서 여성이 성형을 할 때 성형부위별 성형비용을 나타낸 그림이다. 4개의 척도(명목척도, 순서척도, 구간척도, 비율척도) 중 어느 척도에 속하는가? 총비용은 얼마나 드는가? 가장 비싼 부위와 가장 싼 부위의 차이는? 이 그림을 보면서 당신은 어떤 느낌이 드는가?

몸에 대한 또 다른 이야기를 보자! 2007.11.26 조선일보에서는 미국 '시체시장'을 폭로한 책을 저술한 Annie Cheney와의 인터뷰를 실었다. 웹사이트를 이용하여 직접 찾아보는 것도 숫자공부에 도움이 될 것이다. 다음 왼쪽 사진은 미국 '시체시장'을 폭로한 책(Body Brokers, Annie Cheney 저, Broadway출판사(2007)의 겉표지를 나타낸다. 오른쪽은 조선일보에 실린 시신 부위별 가격표를 나타낸 그림이다. 앞의 부위별 성형비용을 나타낸 그림에 나타나는 숫자와 비교하여 보자. 서로 다른 점이 무엇인가? 양쪽 그림을 보면서 당신은 어떤 느낌이 드는가?

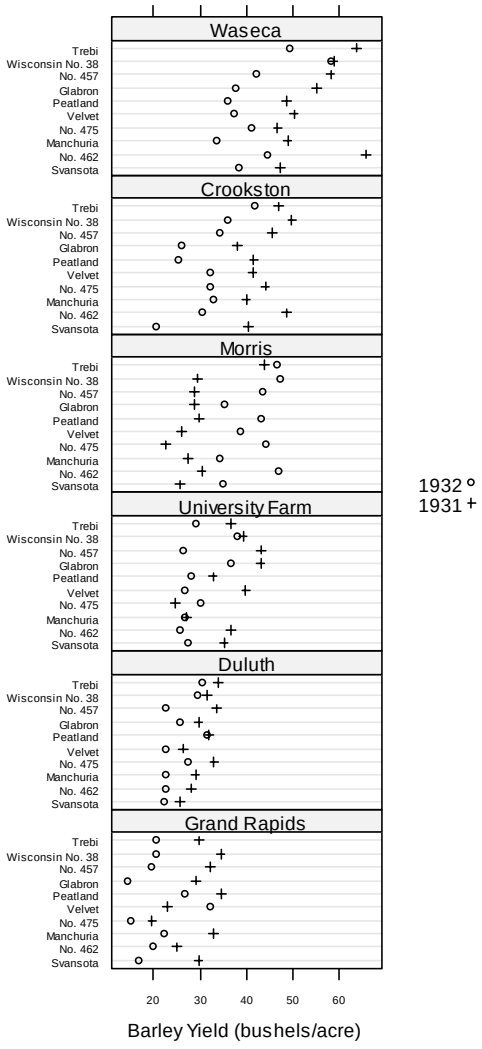




(제공처: www.amazon.com과 조선일보)

제 2 부

자료의 탐색


Waseca
Crookston
Morris
University Farm
Duluth
Grand Rapids
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
1932
1931
20
30
40
50
60
Barley Yield (bushels/acre)

1931년과 1932년 2년에 걸쳐 미국 미네소타주 농경학자들이 경작실험을 실시하였다. 6군데 경작지(Waseca, Crookston, Morris, University Farm, Duluth, Grand Rapids)에 10종류의 보리(Trebi, Wisconsin No. 38, No. 457, Glabron, Peatland, Velvet, No. 475, Manchuria, No. 462, Svansota)를 경작하여 얻은 수확량(총 120개의 자료)을 비교하였다. 여러 학자들이 통계분석을 행하였으나 이 자료가 품고 있는 문제점을 발견하지 못하였다. 앞의 그림(다중점차트)을 보면 6군데 경작지 중 Morris만 유일하게 1931년의 수확량이 1932년의 수확량보다 적다. 같은 미네소타주 경작지들인데 Morris만 다른 결과가 나왔다는 것은 의심의 여지가 있다. 우리는 이러한 문제점을 다중점차트를 통하여 쉽게 확인할 수가 있다.

제 4 장

그래프는 몇 마디 말보다 낫다.

차 례

학습목표

자료에 담겨져 있는 정보를 하나의 그림으로 정확하게 요약한다면 통계학의 이론은 많은 경우 사실 필요가 없다. 숫자로 정보를 요약하기 전에 그림을 그려 자료가 가지고 있는 정보를 파악하는 것은 가장 먼저 해야 할 일이다. 물론 자료가 가지고 있는 다양한 다차원적인 정보를 그림으로 파악하는 것은 생각보다 쉽지가 않다. 이를 위해서는 고차원적인 통계학 이론 뿐 아니라 특수한 통계소프트웨어에 대한 이해를 습득하여야 한다. 그러나 단순한 그림이라도 일상적인 업무를 하는 과정에서 숫자가 제공하는 정보보다 더 유용하게 정보를 제공하는 경우가 많아 본 4장에서는 그림에 대한 기본적인 이해를 돕고자 한다. 그림의 오남용을 막고 정보를 정확하게 전달하려면 일부 통계학의 이론이 접목되는 것이 필수적이다. 시계열 자료 그림에 대해서도 언급하여 보고 마지막으로 그림을 제시할 때 잘못된 정보를 줄 수 있는 잘못된 습관을 지적한다.

4.1 통계학과 그림

통계학은 자료에 관한 학문이다. 주어진 자료에서 정확한 정보를 추출하여 표현하는 것이 통계학의 일차적인 목적이라면 그림은 이를 위한 첫 번째 도구가 된다. 그러나 자료를 그림으로 표현하는 것은 신중을 기하여야 하고 정확한 정보가 그림에 담겨져 있는지 확인하여야 한다. 예를 들어 설명하여 보자.

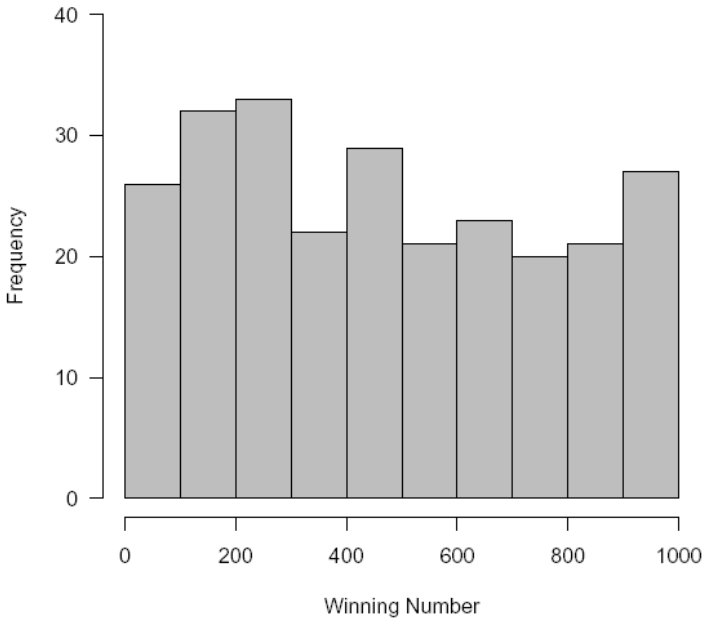
예제 4.1 미국 뉴저지에서 발행되는 한 복권의 형태를 보도록 하자. 매일 시행되는 복권 구매자는 000부터 999까지의 세 자리 숫자 중 한 숫자를 선택하여 복권을 산다. 그리고 매일밤 구매가 종료된 시점에서 하나의 숫자가 무작위로 선택되어 발표되는데 구매한 복권의 숫자와 일치하면 일치한 많은 사람들과 복권액수를 나누어 가지는 형태이다. 매일 당첨 숫자와 복권금액은 발표된다. [그림 4.1]은 지난 254일 자료 중 일부이다.

(810, $190.0), (156, $120.5), (140, $285.5), (542, $184.0), (507, $384.5),
(972, $324.5), (431, $114.0), (981, $506.5), (865, $290.0), (499, $869.5),
(020, $668.5), (123, $83.0), (356, $188.0), (015, $449.0), (011, $289.5),
(160, $212.0), (507, $466.0), (779, $548.5), (286, $260.0), (268, $300.5),
(698, $556.5), (640, $371.5), (136, $112.5), (854, $254.5), (069, $368.0),
(199, $510.0), (413, $102.0), (192, $206.5), (602, $261.5), (987, $361.0),
(112, $167.5), (245, $187.0), (174, $146.5), (913, $205.0), (828, $348.5),
(539, $283.5), (434, $447.0), (357, $102.5), (178, $219.0), (198, $292.5),
(406, $343.0), (079, $332.5), (034, $532.5), (089, $445.5), (257, $127.0),
(662, $557.5), (524, $203.5), (809, $373.5), (527, $142.0), (257, $230.5),
(008, $482.5), (446, $512.5), (440, $330.0), (781, $273.0), (615, $171.0),
(231, $178.0), (580, $463.5), (987, $476.0), (391, $290.0), (267, $176.0),
(808, $195.0), (258, $159.5), (479, $296.0), (516, $177.5), (964, $406.0),
(742, $182.0), (537, $164.5), (275, $137.0), (112, $191.0), (230, $298.0),
(310, $110.0), (335, $353.0), (238, $192.5), (294, $308.5), (854, $287.0),
(309, $203.5), (026, $377.5), (960, $211.5), (200, $342.0), (604, $259.0),
(841, $231.0), (659, $348.0), (735, $159.0), (105, $130.5), (254, $176.0),
(117, $128.5), (751, $159.0), (781, $290.0), (937, $335.0), (020, $514.0),
(348, $191.0), (653, $304.5), (410, $167.0), (468, $257.0), (077, $640.0),
(921, $142.0), (314, $146.0), (683, $356.0), (000, $96.0), (963, $295.0),

[그림 4.1] 복권 숫자와 당첨금액

이러한 정보로부터 복권 구매전략을 세울 수 있는지 알아보도록 하자. 먼저 숫자를 가지고 인간이 전략을 세우기에는 인간의 시각적인 능력이 많이 부족하다. 대부분의 인간은 몇 가지의 숫자를 기억하기에도 벅차기 때문이다. 그러므로 그림을 그려 자료가 가지고 있는 정보를 캐내려고 하는 것이다. 먼저 어떤 숫자를 선택하여 복권을 사야 하는가? 이를 위해 히스토그램을 그려보자. 히스토그램은 후에 자세히 그 용도에 대해 알아보겠지만 여기서는 0부터 99까지

에서 행운의 복권이 나온 도수, 그리고 100부터 199까지에서 행운의 복권이 나온 도수 등을 기록하여 y축에 막대로 그림을 그린 것을 의미한다. 히스토그램을 그리면 다음 [그림 4.2]와 같다.



[그림 4.2] 행운의 복권을 가져다 준 숫자

일반인은 이러한 히스토그램에서 어떠한 정보를 얻을까? 아무런 통계적인 지식이 없다면 아래와 같은 추론을 할 것이다.

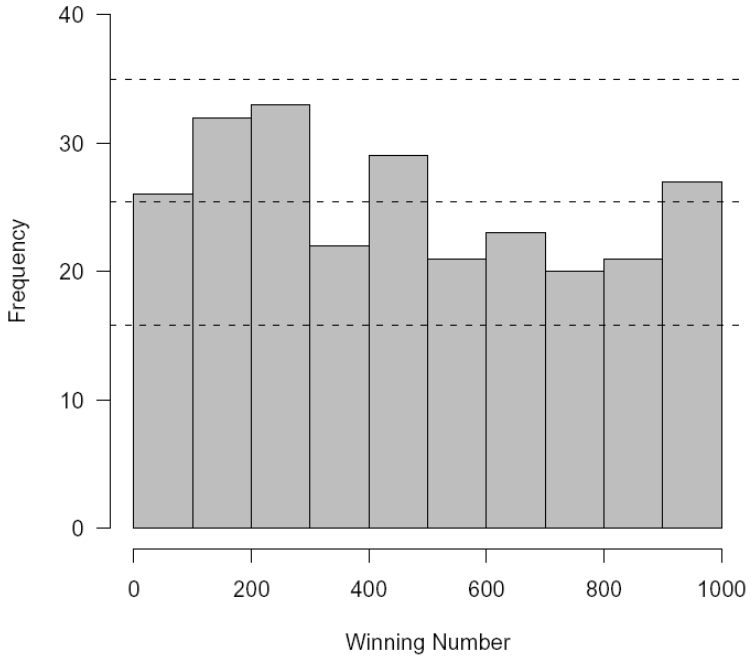
- 100부터 300까지의 숫자가 다른 영역에서의 숫자보다 더 행운의 복권을 가져다 준 숫자이다.
- 따라서 앞으로 복권을 사는 경우는 1부터 300까지의 숫자를 선택하여 산다.

그러나 행운의 복권의 숫자는 000부터 999까지의 숫자 중 무작위로 선택되어 발표가 되었는데 일반인들은 괴리를 느낄 수밖에 없다. 무작위로 선택되어 발표가 된다는 것은 254개의 숫자 중에서 우연치 않게 100부터 300까지 숫자가 몰려 있을 가능성도 없지 않기 때문이다. 이를 위해서는 [그림 4.2]의 정보 뿐 아니라 통계학의 이론을 접목시켜야 할 필요가 있다. 후에 자세한 설명이 뒤따르겠지만 먼저

- [그림 4.2]의 한 셀에 들어가는 도수의 기댓값은 254 x 1/10 = 25.4일 것이다.
- 그리고 당첨 숫자가 무작위로 발표가 된다면 각 셀에서의 변동수준(level of variability)은 다음과 같이 계산되어 진다.

$$\sqrt{254 \text{x} \frac{1}{10} \text{x} \frac{9}{10}} = 4.78$$

- 이 숫자의 2배까지의 폭의 차이는 우연한 변동이라 이야기할 수 있다. 따라서 이러한 사실을 [그림 4.2]에 접목하게 되면 [그림 4.3]과 같다.

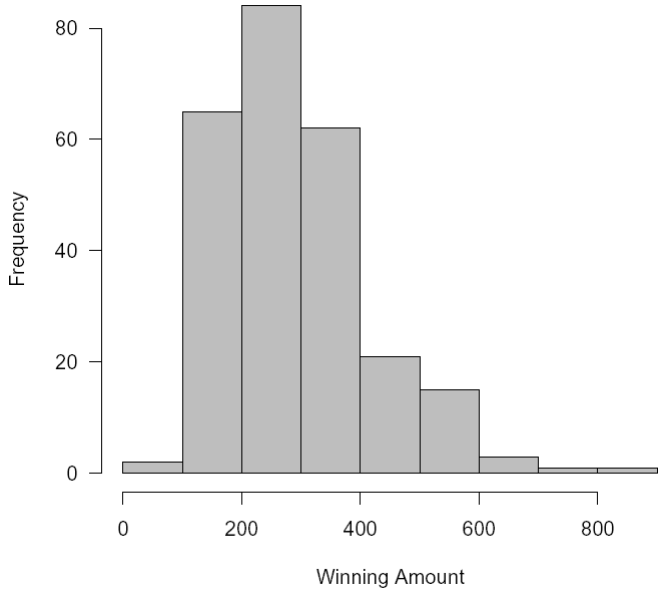


[그림 4.3] 통계이론이 접목된 히스토그램

결론적으로 다음과 같이 이야기할 수 있다.

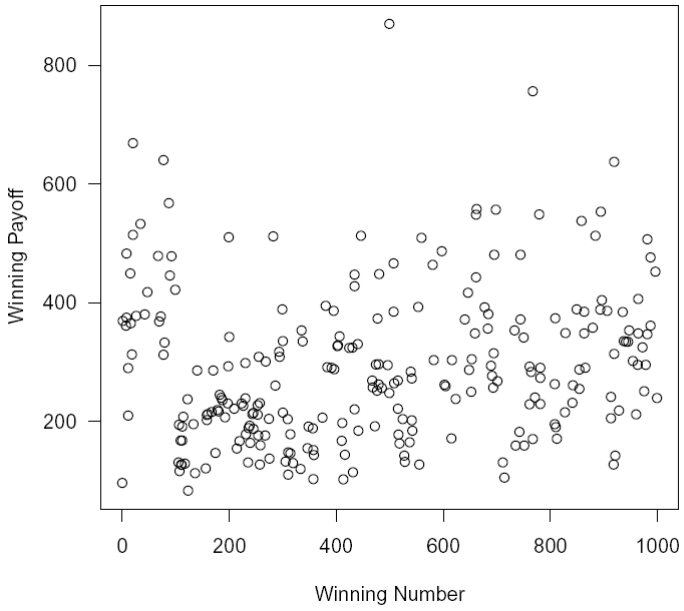
- 행운의 복권을 가져다주는 숫자는 무작위로 선택된다는 사실이다. 즉 어떤 특정의 숫자가 다른 숫자들에 비해 행운을 가져다주지 않는다.

행운의 복권을 가져다주는 숫자를 선별할 수 없으므로 복권의 금액에 대해 언급하여 보자. [그림 4.4]는 복권 금액에 대한 히스토그램이다.



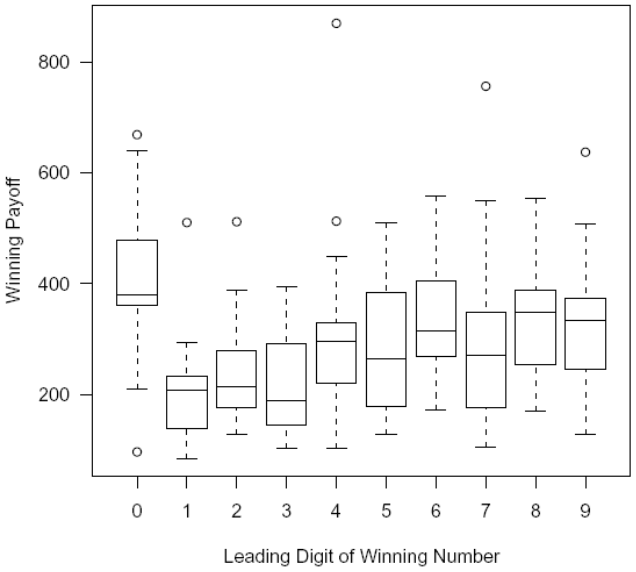
[그림 4.4] 복권 금액에 대한 히스토그램

[그림 4.4]에서 당첨 복권금액의 범위는 매우 넓게 퍼져 있다. 따라서 많은 복권금액을 획득할 수 있는 숫자를 선택할 수 있는 방법이 있는지 여지가 남아 있다. 이를 위해 [그림 4.5]에 당첨 복권 금액과 당첨숫자를 산점도를 그려 보았다. 즉 x-축에는 숫자를 y-축에는 당첨금액을 적어 놓았다.

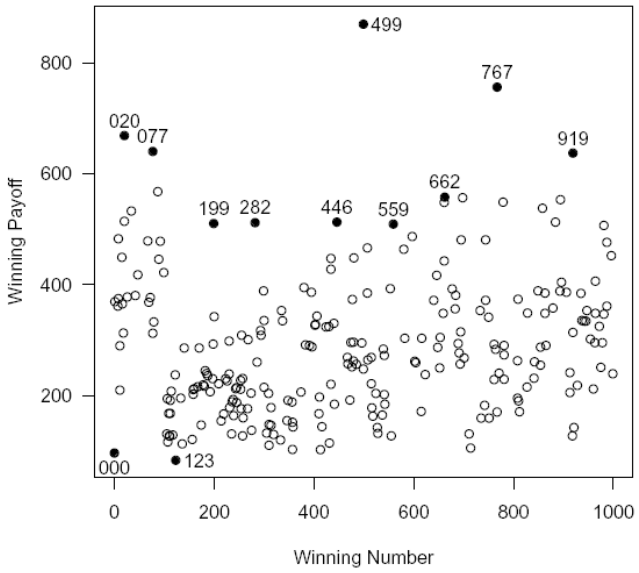


[그림 4.5] 숫자와 복권당첨 금액의 산점도

이러한 산점도에서 보다시피 맨 왼쪽에 있는 상금액이 다른 쪽보다는 높다는 사실을 짐작할 수 있는데 이러한 산점도를 10개의 구간으로 나누어 [그림 4.6]을 그려 보았다. 이러한 그림에 대해서는 5장에서 이야기할 기회가 주어진다. 이러한 그림을 상자그림이라 한다. 상자 가운데 값이 중앙값인데 맨 왼쪽의 값이 다른 값들에 비해 매우 높게 나온다. 그리고 각 구간별로 이상점들이 나타나는데 이러한 점들의 복권 숫자를 명시하여 [그림 4.7]처럼 그림을 그려 보아도 좋다.



[그림 4.6] 상자그림



[그림 4.7] 이상점을 표시한 숫자와 복권당첨 금액의 산점도

따라서 이러한 결과로 미루어 볼 때 다음과 같은 결론을 얻어 낸다.

- 다른 사람들이 선택할 가능성이 높지 않은 숫자를 선택하라. 만약 복권이 당첨된다면 당신은 더 많은 상금액을 얻을 것이다.
- 그리고 반복되는 숫자가 섞인 3자리 숫자를 얻으라.
- 명백한 숫자인 000, 123 등은 피하라. ■

주어진 자료에서 정보를 얻어 내는 작업은 통계학의 이론과 순차적이고 논리적인 사고가 필요하다. 이어지는 절에서는 일상에서 우리가 많이 사용하는 방법들을 소개하고 그 주의 점에 대해 알아보도록 하자.

4.2 히스토그램과 유사그림

구간척도나 비율척도로 수집된 자료인 경우 자료의 정보를 표현하는데 있어 기본적인 방법인 방법을 소개한다. 쉽게 생각 할 수 있는 자료 표현방법이다. 주어진 변수가 하나인 경우 많이 쓰인다.

예제 4.2 [그림 4.8]은 1869년부터 1957년까지 미국 뉴욕주의 강수량 자료(단위: 인치)이다.

43.6	37.8	49.2	40.3	45.5	44.2	38.6	40.6	38.7	46.0
37.1	34.7	35.0	43.0	34.4	49.7	33.5	38.3	41.7	51.0
54.4	43.7	37.6	34.1	46.6	39.3	33.7	40.1	42.4	46.2
36.8	39.4	47.0	50.3	55.5	39.5	35.5	39.4	43.8	39.4
39.9	32.7	46.5	44.2	56.1	38.5	43.1	36.7	39.6	36.9
50.8	53.2	37.8	44.7	40.6	41.7	41.4	47.8	56.1	45.6
40.4	39.0	36.1	43.9	53.5	49.8	33.8	49.8	53.0	48.5
38.6	45.1	39.0	48.5	36.7	45.0	45.0	38.4	40.8	46.9
36.2	36.9	44.4	41.5	45.2	35.6	39.9	36.2	36.5	

[그림 4.8] 뉴욕 강수량 자료

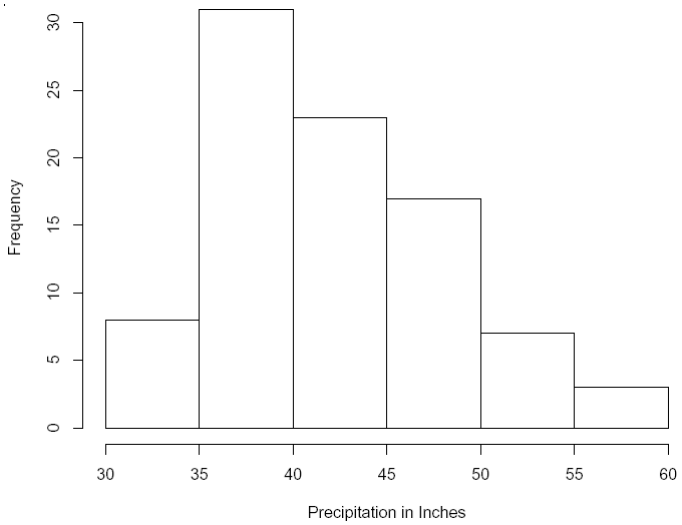
현재 뉴욕 강수량에 대해 아무런 정보를 가지고 있지 않은 상태에서 정보를 얻고자 한다면 제일 간단한 방법은 강수량을 크기 순으로 나열하는 것이다. 단 자료를 줄기(자료의 중요한 부분)와 잎(자료의 덜 중요한 부분)에 해당하는 부분으로 나누어 표현하면 좋을 것이다. 이런 아이디어로 나온 방법이 줄기-잎 그림이다. [그림 4.9]의 첫 번째 그림은 줄기가 일단위로 크기가 2씩 상승함을 알 수 있는 반면 두 번째 그림은 줄기가 일단위로 5씩 상승함을 알 수 있다.

예를 들어 첫 번째 그림 중 세 번째 줄 마지막 88중 8은 37.8에 해당하므로 자료에는 이와 같은 37.8이 두 개 있음을 알 수 있다.

```
32 | 7578
34 | 147056
36 | 1225778991688
38 | 3456670034445699
40 | 1346684577
42 | 4016789
44 | 2247001256
46 | 0256908
48 | 552788
50 | 380
52 | 025
54 | 45
56 | 11


3 | 344444
3 | 5566666777777788888999999999
4 | 000000011112222334444444
4 | 55555666677778999
5 | 0000113344
5 | 666
```

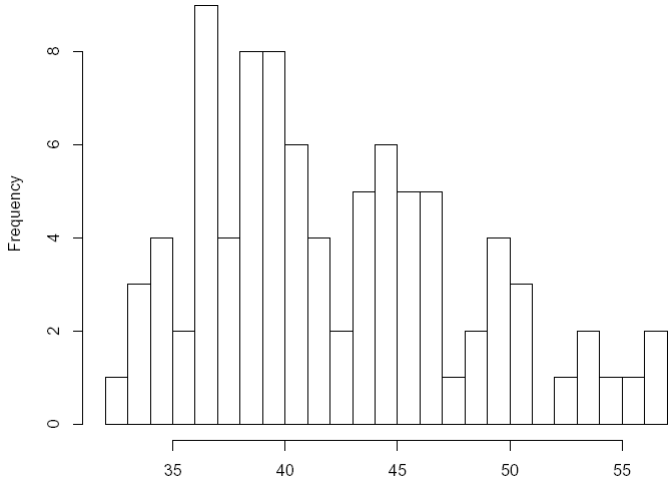
[그림 4.9] 줄기-잎 그림

이와 같은 줄기 잎 그림은 자료의 개수가 그렇게 크지 않을 때 사용하는 간편한 기법이다. 줄기-잎 그림을 통해 자료가 어디에 많이 몰려 있으며, 어느 줄기에 자료(잎)가 제일 많이 분포되어 있고 이상점은 없는지, 자료는 대칭인지 혹은 비대칭인지 등에 대한 정보를 알아 볼 수 있을 것이다. 그러나 자료의 개수가 많은 경우에는 줄기-잎 그림보다는 히스토그램을 더 자주 즐겨 쓴다. 히스토그램은 자료의 범위를 중복되지 않는 구간으로 나누어 각 구간에 자료의 개수가 몇 개가 들어가는지 파악한 후 막대를 이용해 그 개수를 표현하는 그림이다. [그림 4.10]은 구간의 크기를 5로 하여 그린 히스토그램이다.



[그림 4.10] 구간의 크기 5인 히스토그램

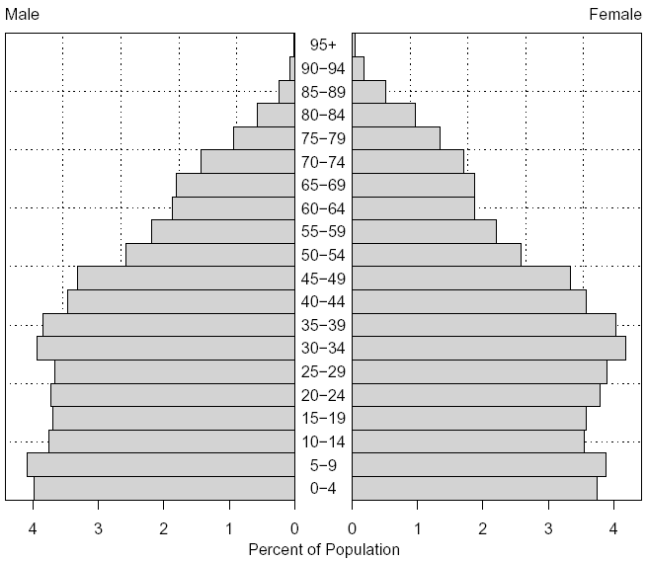
자료의 정확성을 전달하기 위하여 히스토그램의 구간은 중복되거나 구간의 크기가 일정치 않으면 안 된다. 히스토그램은 y-축이 도수로도 표현이 되지만 경우에 따라서는 상대도수, 즉 전체 도수를 전체 자료개수로 나눈 수로도 표현이 가능하다. 그러나 히스토그램은 구간의 크기에 따라 모양이 결정되기 때문에 구간의 크기를 정하는 문제는 신중해야 한다. 구간의 크기를 작게 하거나 너무 크게 하게 되면 자료 전달이 왜곡되거나 부정확하게 표현되기 때문이다. [그림 4.11]은 구간의 크기를 1로 하여 그린 히스토그램이다. [그림 4.10]과 비교하여 보라. ■



[그림 4.11] 구간의 크기가 1인 히스토그램

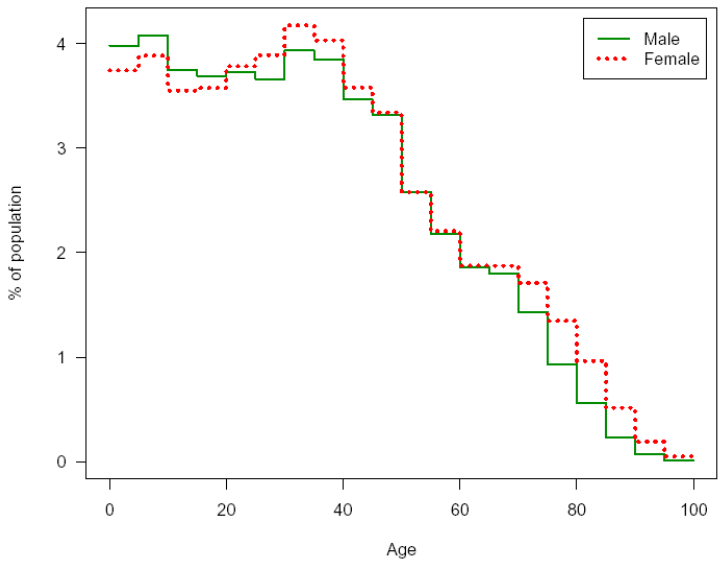
히스토그램은 두 그룹 간의 비교 목적으로 활용이 되는데 예를 들면 성별이나 나이, 지역 등으로 나누어 자료를 비교하는 경우를 들 수 있다. 다음 [그림 4.12]는 성별로 나눈 우리나라

국민의 나이에 관한 히스토그램이다. x-축은 총 인구대비 비율을 나타낸다.



[그림 4.12] 피라미드 히스토그램

그러나 이러한 피라미드 형태의 히스토그램은 비교를 좌우측으로 늘어진 막대의 길이로 하여야 하는데 보는 사람들로 하여금 약간 피곤하게 만드는 성향이 있다. 이보다 좋은 방법은 하나의 그림에 다른 그림을 덧붙이는 형태의 그림이다. [그림 4.13]을 보기 바란다. 히스토그램의 두 그림을 하나의 공동 스케일로 보게 되면 비교가 훨씬 용이하다.



[그림 4.13] 겹쳐진 두 개의 히스토그램

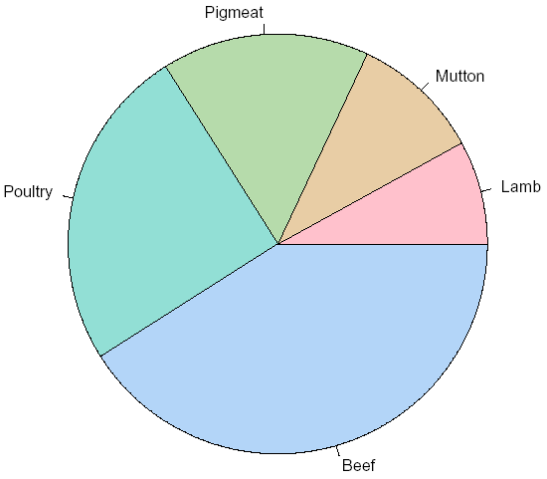
히스토그램의 약점은 구간의 크기를 정하는 문제이다. 구간의 크기 혹은 구간의 개수의 문제는 명확한 기준이 없는 관계로 독자적으로 판단하거나 통계소프트웨어가 지정한 대로 그려야 하는데 그만큼 주의를 하여야 한다는 의미이다.

4.3 또 다른 그림

언급한 히스토그램은 구간척도나 비율척도로 수집된 자료인 경우 제일 먼저 접하는 그림이다. 그러나 주어진 자료가 다른 척도인 명목척도나 순서척도로 수집된 경우는 다른 표현방법이 존재한다. 주로 빈도(count)나 비율(proportion)과 같은 자료에 많이 적용이 된다. 먼저 파이차트부터 알아보자.

원형그래프(파이차트):

다음 [그림 4.14]는 어느 나라 국민의 육류소비에서 차지하고 있는 각종 육류의 비율을 표시한 원형그래프이다.

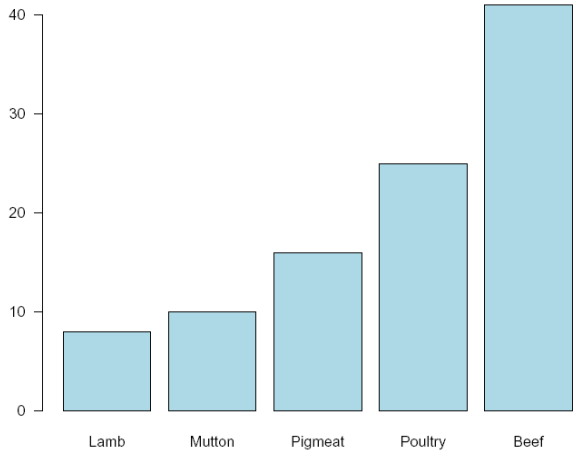


[그림 4.14] 원형그래프

원형그래프는 각도 혹은 면적으로 항목이 차지하고 있는 비율을 보여 주고 있기 때문에 정보를 표현하는데 있어 완벽하지 않다. 가끔 원형그래프 외곽에 숫자를 기록하여 보여 주고 있는 경우도 있지만 이런 단점을 극복하고자 하는 시도이다. 원형그래프는 흔히 비율에만 적합할 뿐 다른 자료에는 사용하지 않는 것이 원칙이다.

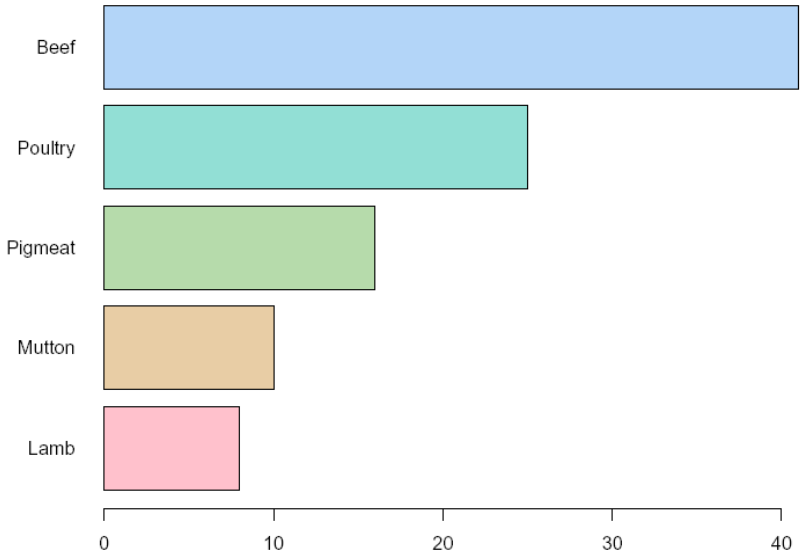
막대그래프(막대차트):

다음 [그림 4.15]는 앞에서 언급한, 같은 자료에 대한 막대그래프이다. 막대 차트는 공동 척도를 이용해 항목 간 비교를 하기 때문에 파이차트보다 훨씬 자료를 전달하는데 객관적이다. 그리고 비율뿐 아니라 다른 자료인 경우에도 적용된다.



[그림 4.15] 막대그래프

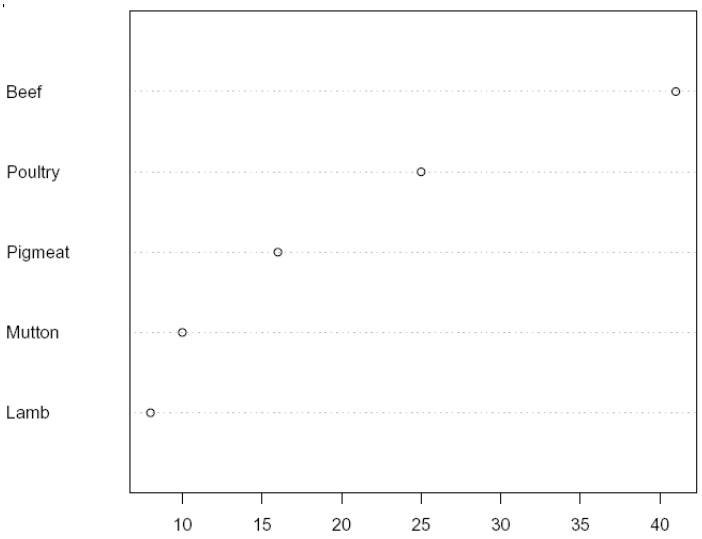
혹은 [다음 4.16]처럼 막대들을 눕히고 제일 큰 막대를 위에 놓는 차트가 있는데 항목의 순서 중요성을 부각하기 위한 방법이다. 이름을 굳이 붙인다면 **수평막대그래프**이다.



[그림 4.16] 수평막대차트

점차트:

이 그림은 막대를 점으로 표시하는 것인데 점을 이어주는 수평선을 그리는 것이 특징이다. 이는 수평막대그래프와 유사하다.



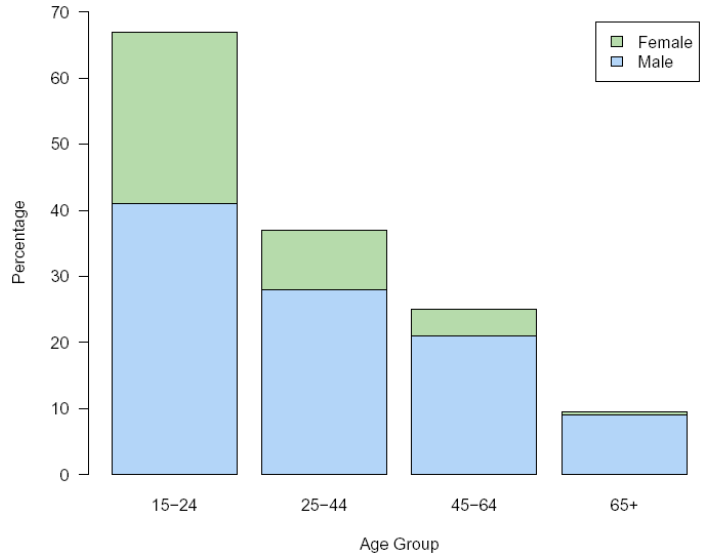
[그림 4.17] 점차트

항목이 두 개 이상인 경우에도 막대그래프는 유용하게 쓰인다. [표 4.1]은 모 보험회사에서 연령별 성별로 분류한 인구 범주에서 매우 위험한 운전 습관을 가진 비율이 얼마나 되는지를 보여주는 분할표이다.

성별	나이			
	15-24	25-44	45-64	65+
남성	41	28	21	9
여성	26	9	4	5

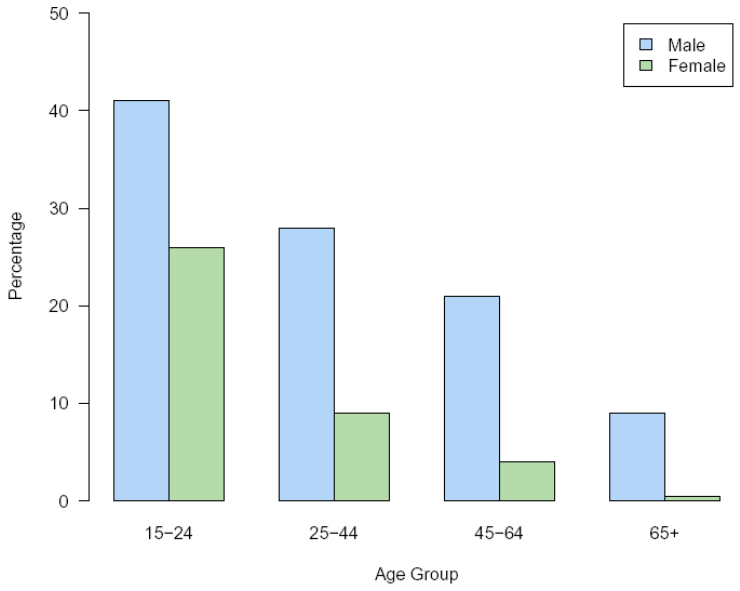
[표 4.1] 분할표

분할막대그래프 : [표 4.1]을 이용하여 분할막대그래프를 그리면 다음 [그림 4.18]과 같다.



[그림 4.18] 분할막대그래프

이런 그래프는 다음 [그림 4.19]와 같이 병렬로 비교를 원활히 하는 경우도 있다.

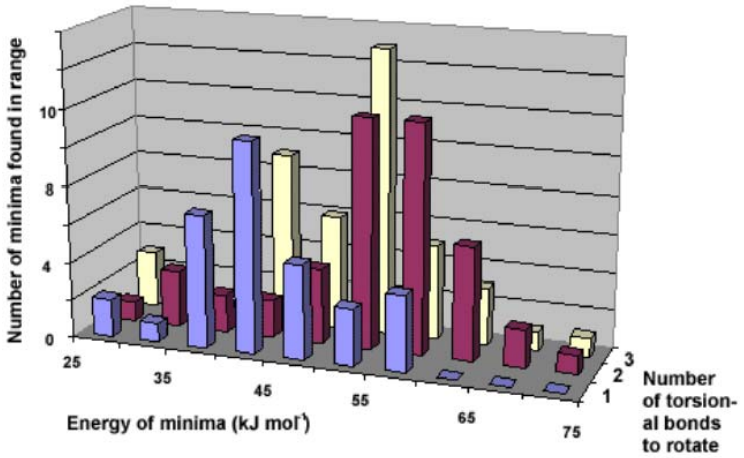


[그림 4.19] 병렬막대그래프

비교목적이라면 [그림 4.19]가 [그림 4.18]보다 더 좋다. 그러나 [그림 4.18]에서 합쳐진 막대의 크기에 대한 의미가 부여되면 [그림 4.18]을 쓰는 것도 생각해 봄 직하다.

이상이 명목척도나 순서척도로 만들어진 자료에 대해 기본적으로 많이 활용되는 그림들인데 이를 변형한 많은 그림들이 존재한다. 차원을 추가하는 등의 기법 등을 통해 표현하는데 많은

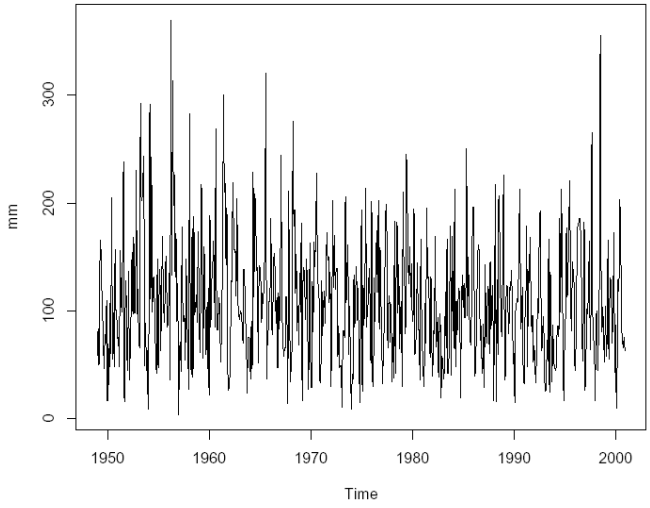
경우 이는 불필요하다. [그림 4.20]은 항목을 3차원으로 분류하여 만든 막대그래프이다. 과연 얼마만큼의 표현 목적을 달성할 수 있는지 의문시된다.



[그림 4.20] 3차원 막대차트

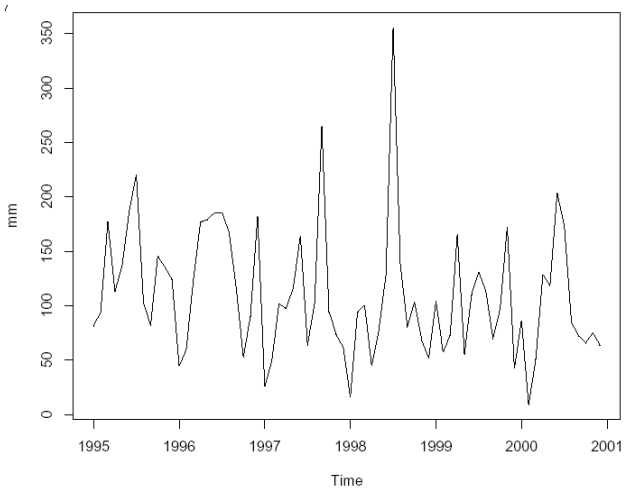
4.4 시계열그림

시계열 자료는 언급하였듯이 시간대별로 수집된 자료를 의미한다. 시계열 자료를 그림으로 표현하는 방법에 대해 알아보도록 하자. 시간별로 수집되었기 때문에 시간의 크기는 그림을 그리는데 매우 중요한 역할을 한다. [그림 4.21]은 우리나라의 1950년부터 2001년까지의 월강우량 자료이다.



[그림 4.21] 월강우량 자료를 나타내는 시계열그림(1)

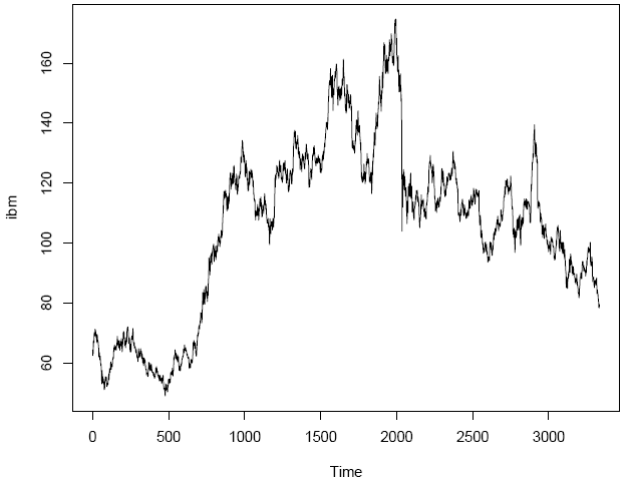
시간축이 너무 방대하게 설정되어 있기 때문에 이런 현상이 벌어진다. 이런 경우는 시간 축을 나누어 보는 것이 바람직하다. [그림 4.22]는 1995년부터의 기록이다.



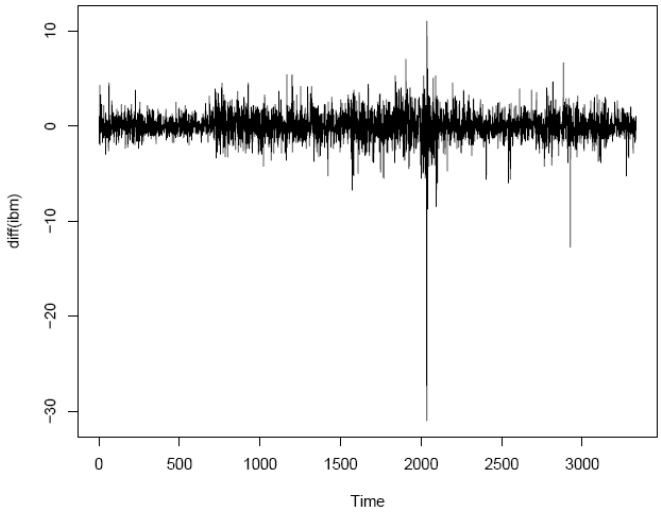
[그림 4.22] 월강우량 자료를 나타내는 시계열그림(2)

확대하여 본 결과 시계열 자료가 가지고 있는 특징이 나타난다.

[그림 4.23]은 1980년 1월 1일부터 1992년 10월 8일까지 IBM의 주가 자료이다. 어느 주식이나 거의 비슷한 패턴을 보인다. 주가자료는 이론에 의하면 주식시장이 효율시장(efficient market)이라면 랜덤워크(random work)라고 알려져 있는데 주가는 어느 날이고 상관없이 독립적으로 오르는 확률과 내리는 확률이 같다는 의미이다. 이를 탐색하기 위해서는 원자료를 시계열로 표현하는 대신 인접한 날짜 간의 주가 차이를 표현하여야 한다. [그림 4.24]는 이론이 바탕이 된 그림이 되는 것이다. 효율성을 이해하기 위해서는 [그림 4.24]의 자료를 분석하여야 한다.



[그림 4.23] IBM의 주가 자료를 나타내는 시계열그림

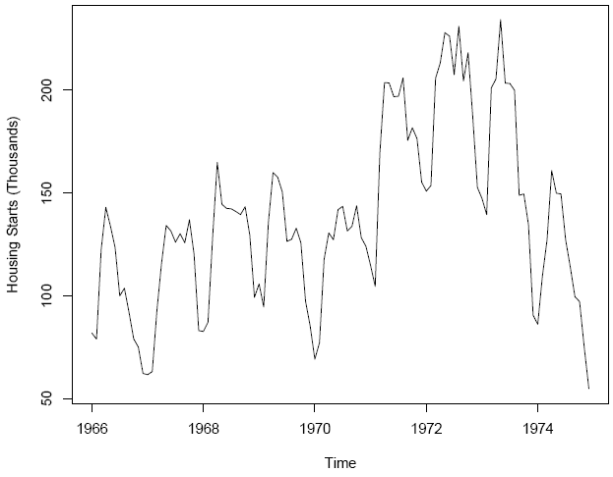


[그림 4.24] 주가의 차이를 나타내는 시계열그림

시계열 자료는 많은 구성 성분으로 이루어진 자료이다. 많은 시계열 자료는 추세요인과 계절적인 요인 그리고 통제할 수 없는 요인으로 이루어져 있다고 믿고 있다. 한 예로 시계열 자료를 x_t 라 한다면

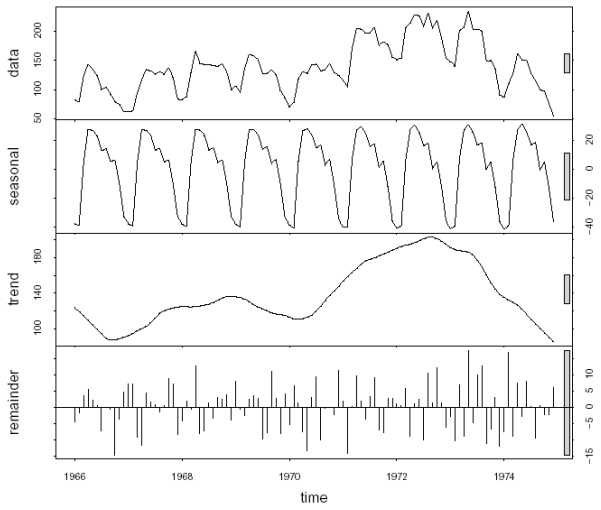
$$x_t = T_t + S_t + I_t, t = 1,2,\cdots,t$$

으로 구성이 될 수 있다. 여기서 T_t 는 추세요인, S_t 는 계절적인 요인, 그리고 I_t 는 통제할 수 없는 요인을 의미한다. [그림 4.25]는 미국에서 발표한 경제지표로 발표한 1966년 1974년까지의 매월 신축하는 집의 수에 대한 그림이다. 집은 경기가 좋지 않다고 판단되면 짓지 않기 때문에 이러한 지표는 경기 선행지수로 활용된다. [그림 4.25]를 보게 되면 이러한 세 가지 요인이 다 섞여 있는 그림이다.



[그림 4.25] 신축하는 집의 수에 대한 시계열그림

물론 이런 그림은 그 자체로도 의미가 있겠지만 통계적 분석을 통해 이러한 세 가지 요인을 분해할 수 있다면 좋을 것이다. [그림 4.26]이 한 예이다. 많은 공공부문의 자료들은 시계열 자료이기 때문에 분석 후의 결과를 그림으로 표현하는 것은 매우 중요하다.



[그림 4.26] 분해후의 시계열그림

4.5 좋은 그림과 나쁜 그림

요즈음 우리나라 매스컴에서 정보전달의 목적으로 통계그래픽스(statistical graphics)를 활용하는 작업이 부쩍 많아지고 있다. 단순한 정보전달이나 자료분석의 결과를 시각화하는 그래프를 통하여 일반 시청자나 독자가 흥미를 갖고 단순, 명료하게 보도록 하는 어려운 작업일 것이다. 이 작업의 실패로 말미암아 그래프 작성 담당자의 실력에 관계없이 정보나 자료분석에 대한 불신이나 더 나아가서는 매스컴 자체에 대한 불신을 가지고 올 가능성이 높다. 통계적 정보전달은 말(words), 표(table), 그래픽스(graphics)라는 수단을 통하여 효과적으로 전달되는 데 이 중 그래픽스는 통계적 정보에서의 관계나 경향을 간단한 시각 형태로 전달하는 데 필수적이다. Mahon(1977)은 이러한 말, 표, 그래픽스를 각각 재래식 군대의 보병, 포병, 기병에 비유하였다. 서로 보완적인 이 세 가지 수단 중 그래픽스가 갖는 특성을 재래식 군대에서 제일 기동력이 있는 기병에 비유하였던 것이다. 이처럼 통계 그래픽스는 컴퓨터 그래픽스의 발전과 더불어 통계적 정보전달의 강력한 도구로서 점점 그 중요성이 더해 가고 있다. 그러나 그 중요성에 비하여 매스컴에서는 통계 그래픽스를 잘못 적용, 강요하고 매스컴을 보는 일반 대중은 잘못 인식, 해독하는 악순환의 고리가 끊어지지 않고 되풀이되고 있다. 시각화 작업이 강한 인상을 주는 만큼 그로 인한 피해도 클 수밖에 없다. 강하지만 잘못된 메시지를 남기는 통계 그래픽스로 인하여 현실에 대한 편견이 생기는 것이다.

자료분석은 탐색의 단계와 확증의 단계로 구분하는데 탐색적 자료분석으로 자료의 구조와 특징 등을 파악한다. 이때 특히 유용한 것이 통계 그래픽스이다. 매스컴에 나타나는 그래프는 자료분석의 결과를 전달하거나 전시하는데 쓰이는 정보그래픽스(informational graphics) 위주로 구성되어 있고 정보그래픽스에서도 막대그래프(bar chart와 column chart), 원형그래프(pie chart), 꺾은선그래프(line chart), 그림그래프(pictogram), 통계지도(statistical map) 5개의 그래프 위주로 구성되어 있다.

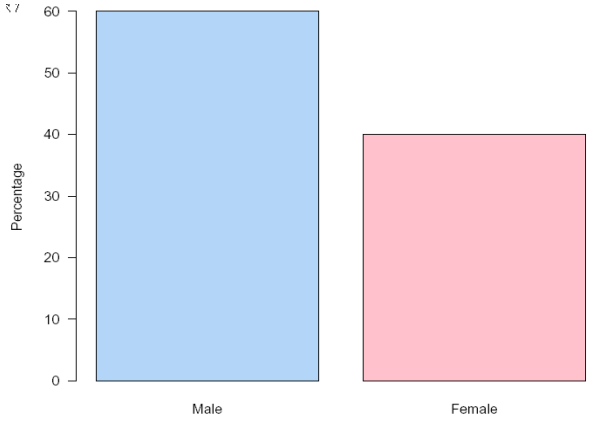
정보그래픽스에 대한 통계적 오용의 형태로서 우리는 다음과 같이 17가지를 나열할 수 있다.

(1) 자료의 크기와 그래프의 크기가 비례하지 않는다.
(2) 시계열그래프에서 시간축의 눈금 간격이 일정하지 않다.
(3) 같은 시간대의 비교를 하지 않는다.
(4) 그래프에서 자료가 빠져 있다.
(5) 그래프에서 수직축의 눈금이 중간에서 변경되거나 그래프의 줄임표시가 있다.
(6) 그래프를 입체화하기 위하여 투영도법을 무리하게 적용한다.
(7) 그래프에서 배경그림이 자료에 대한 해독(decoding)을 방해한다.
(8) 시계열그래프에서 수평축인 기준선을 공유하지 않는다.
(9) 중요한 그림 요소를 빠뜨리거나 불필요한 그림 요소를 첨가시켜 그래프 상의 수치 비교를 방해한다.
(10) 자료에 맞지 않는 그래프를 선택하여 그린다.
(11) 자료에 대한 오차를 표현할 때 아래의 통계적 변동 중 어떤 오차인지를 밝히지 않는다.
  (i) 표본평균±표본표준편차
  (ii) 표본평균±표준오차
  (iii) 모평균에 대한 $100(1-\alpha)$% 신뢰구간(여기서, α는 유의수준)
(12) 자료값이 범위를 갖는 경우에 표시를 하지 않는다.
(13) 독자들의 시각적 암시(visual metaphor)를 무시한다.
(14) 배경그림만 있고 그래프가 없다.
(15) 부주의하여 그림 요소에 대한 설명이 틀리거나 빠져 있다.
(16) 원형그래프를 변형하여 타원, 사각형, 육각형그래프 등을 사용할 때 각 범주의 면적을 고려하지 않는다.
(17) 시계열그래프에서 자료점들을 연결할 때 흥미를 끌기 위하여 직선이 아닌 다른 표현을 쓴다.

여러 가지 정보그래픽스에 대한 통계적 오용의 형태에 대하여 알아보자. 이를 통하여 좋은 그림을 그리는 기본적인 원칙을 설명하고자 한다. 또한 원칙을 위배하지 않는 기준을 마련할 수 있다.

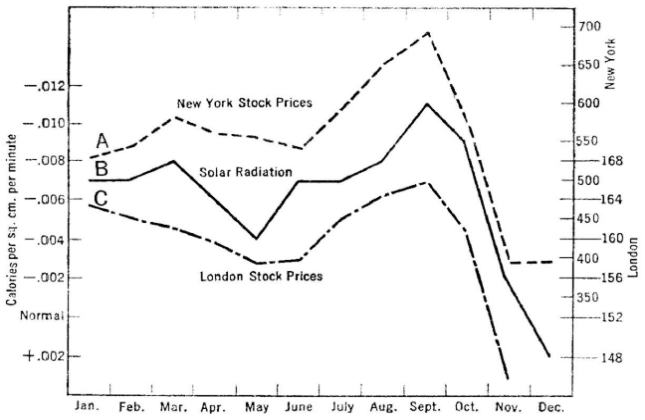
• 자료가 작은 경우 그림을 그릴 필요는 없다. [그림 4.27]과 같은 그림은 그림일 뿐 그림이

가지고 있는 목적, 즉 그림으로 자료의 정보를 정확하게 전달하고자 하는 목적을 상실한 그림이다. 단순하게 명시만 해도 될 사항을 그림으로 표시할 필요는 없다.



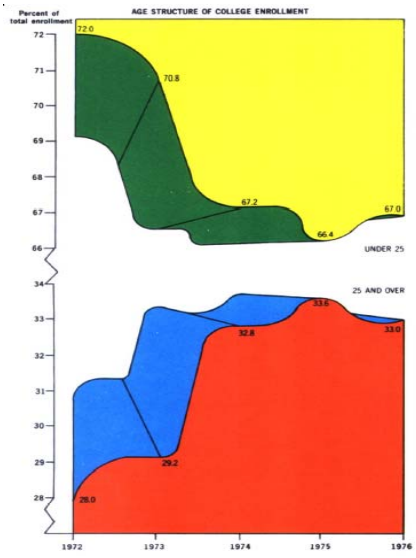
[그림 4.27] 막대그래프

- 그림은 자료만큼만 정확하다. 애매모호한 자료로부터 훌륭한 그림은 절대 나오지 않는다. 그리고 그런 노력을 할 필요가 없다. [그림 4.28]은 세 변수의 시계열 자료를 하나의 그림으로 나타내어 아무런 연관성이 없는 자료를 마치 연관이 있는 것처럼 그린 그림이다.



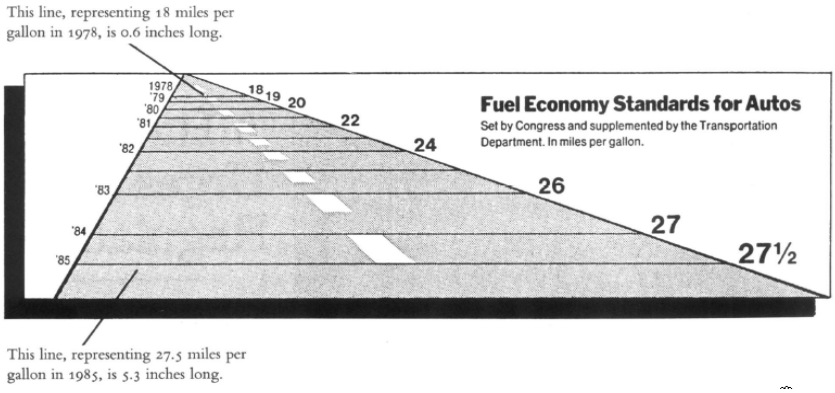
[그림 4.28] 꺾은선그래프

- 그림은 복잡성을 유발할 필요가 없다. 그림이 가지고 있는 복잡성만큼만 복잡하면 되지 쓸데없이 색깔을 칠하거나 불필요한 장식을 하거나 3차원적으로 그릴 필요가 없다. [그림 4.29]는 1974년부터 1978년까지의 모 대학에 등록한 25세 이상의 학생을 비율을 그림으로 표현한 것인데 우리가 필요한 숫자를 정확하게 전달하는데 실패한 그림이다.



[그림 4.29] 꺾은선그래프의 변형

- 그림을 그릴 때에는 절대로 자료를 왜곡해서는 안 된다. 왜곡은 우발적이거나 의도적으로 발생이 되는데 의도적으로 진실을 부각하거나 은폐하기 위해 쓰이는 경우가 많다. 그림그래프를 사용할 때 제일 주의할 사항은 자료의 크기와 그래프의 크기가 비례하여야 한다는 사실이다. [그림 4.30]은 1978년부터 1985년까지 갤런 당 마일리지로 정의된 적정 자동차의 효율성 기준이다.

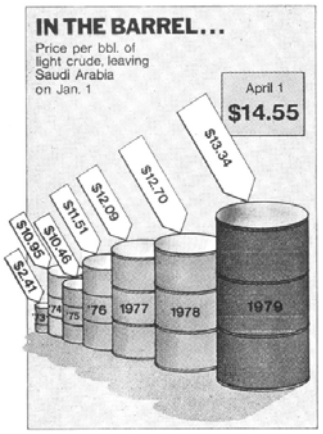


[그림 4.30] 자동차 효율성 기준을 나타내는 그래프

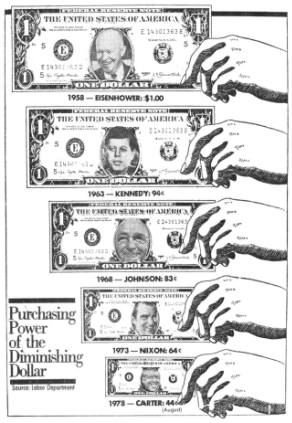
자료의 크기와 그래프의 크기가 비례하지 않을 때 인식왜곡지수 PDI(perceptual distortion index, 또는 lie factor)는 다음과 같이 정의하고 이 값이 100이 아니면 자료의 왜곡이 생긴다.

$$PDI = \frac{\text{그래픽에서 나타나는 효과의 크기}}{\text{자료의 효과 크기}} \times 100$$

미국 예일 대학의 Tufte교수(2001)에 의하면 [그림 4.30]에서 데이터의 크기의 효과는 (27.5-18)/18 = 0.53이며 그림의 크기에 의한 효과는 7.83이므로 이 그림이 가지고 있는 거짓말 효과(lie factor)는 7.83/0.53 = 14.8이다. 왜곡의 극치를 보여주는 그림이라 할 수 있다. 이러한 왜곡은 삼차원의 그림을 그리는 과정에서 부피로 잘못된 비교를 하게 하는 경우와 잘못 계산된 면적을 보여주는 경우에도 나타난다. [그림 4.31] 및 [그림 4.32]를 참조하기 바란다.

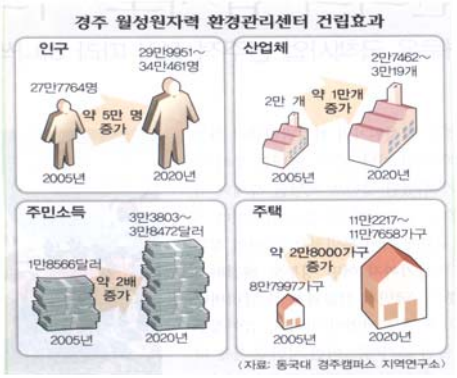


[그림 4.31] 원유가격을 나타내는 그림그래프



[그림 4.32] 달러의 구매력을 나타내는 그림그래프

또 다른 예를 살펴보자. 다음 [그림 4.33]은 2007년 발행된 국정홍보처 책자인 코리아플러스에 나타난 그림그래프들이다. 이들은 동국대 경주캠퍼스 지역연구소에서 방폐장 유치 이후 경주지역사회의 발전적 변화에 대한 연구발표를 인용하여 그린 결과이다. 증가효과는 주민소득 항목에서는 돈다발그림의 높이로 느끼게 되나 인구, 산업체, 주택 항목에서는 각각 사람그림, 공장그림, 주택그림의 면적이나 체적으로 느끼게 되어 자료의 크기와 그래프의 크기가 비례하지 않는다.



[그림 4.33] 환경관리센터 건립효과를 나타내는 그림그래프

다음 [표 4.2]는 각 항목에 대하여 실제 증가비율, 그림그래프에서의 증가비율, PDI 값을 비교한 표이다. 주택 항목에서 왜곡이 제일 크고 주민소득 항목에서는 왜곡이 적다.

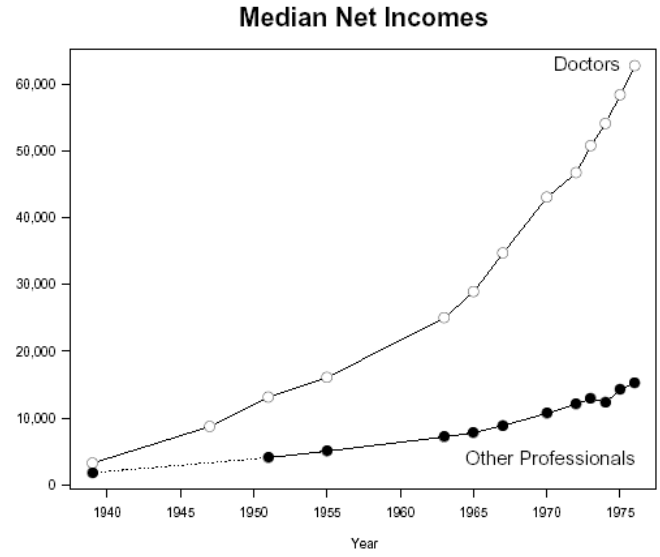
항목	자료에서의 증가비율		그래프에서의 증가비율		PDI 값	
	최소	최대	면적	체적	최소	최대
인구	1.08	1.23	$1.53^2 = 2.33$	$1.53^3 = 3.56$	189.4	329.6
산업체	1.37	1.50	$1.63^2 = 2.64$	$1.63^3 = 4.29$	176.0	313.1
주민소득	1.82	2.07	1.73	1.73	83.6	95.1
주택	1.27	1.33	$2.40^2 = 5.76$	$2.40^3 = 13.8$	433.1	1088.2

[표 4.2] 환경관리센터 건립효과를 나타내는 그림그래프에 대한 평가

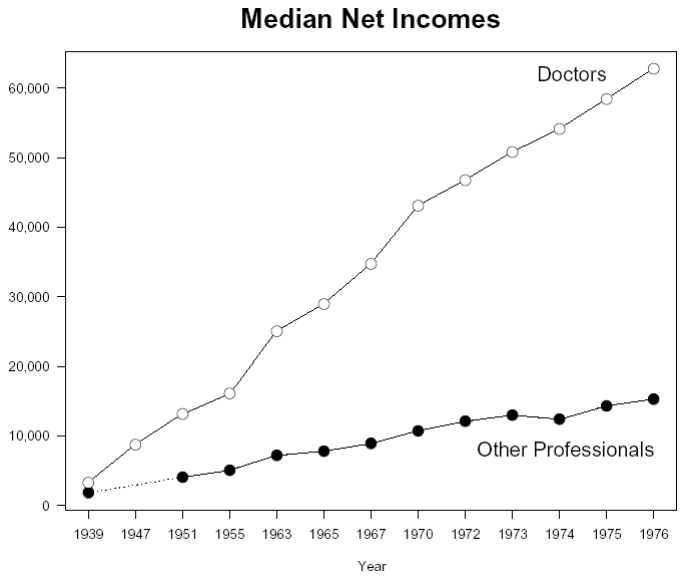
• 우리가 꺾은선그래프를 그릴 때 특히 주의 할 사항은 다음과 같다.

(1) 꺾은선그래프에서 시간축(가로축)의 눈금 간격이 일정한가?
(2) 세로축에 물결선이 꼭 필요한가?

[그림 4.34]는 정상적인 시계열그림인데 [그림 4.35]는 의도적으로 시계열 구간을 같은 크기로 하지 않음으로써 의사들의 봉급수준을 왜곡하는 결과를 초래하였다.



[그림 4.34] 꺾은선그래프(정상)

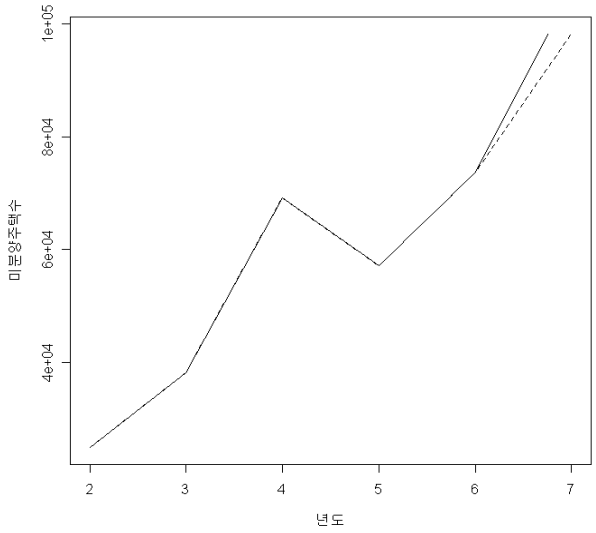


[그림 4.35] 꺾은선그래프(왜곡)

또 다른 예를 하나 더 들어보자. 다음 [그림 4.36]은 2007.11.21 조선일보에 나타난 미분양 주택 추이를 나타내는 꺾은선그래프이다. 데이터가 2002년 12월부터 2006년 12월까지는 1년 단위 간격이나 마지막 데이터는 9개월 간격인 데도 1년 단위 간격을 유지하고 있다. 즉 시계열그래프에서 시간축의 눈금 간격이 일정하지 않다. 마지막 데이터에 9개월 간격을 지키면 꺾은선의 기울기가 더 가파르게 된다.

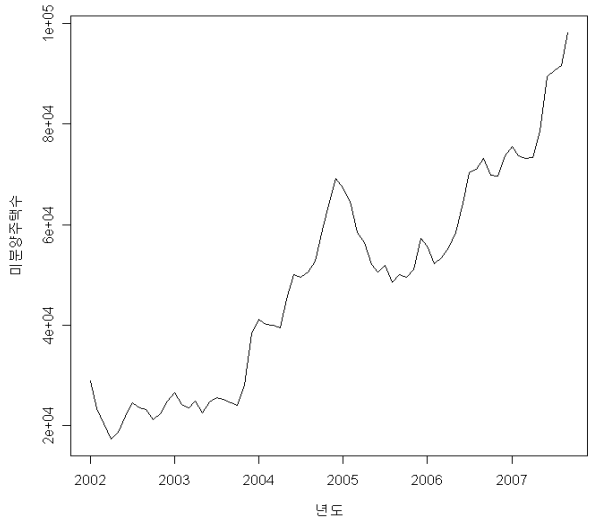
[그림 4.36] 미분양주택 추이를 나타내는 꺾은선그래프

다음 [그림 4.37]은 원그림(점선)과 수정한 그림(실선)을 겹쳐 그린 그림이다. 심각한 차이는 아니나 수정한 그림(실선)이 원그림(점선)보다 기울기가 더 가파르다는 것을 알 수 있다.



[그림 4.37] 미분양주택 추이를 나타내는 꺾은선그래프(정상과 왜곡)

다음 [그림 4.38]은 국토해양부 법령자료 메뉴/통계부메뉴에서 2000년 1월부터 2007년 9월까지의 미분양주택 현황 자료를 이용하여 그린 시계열 그림이다.



[그림 4.38] 미분양주택 추이를 나타내는 시계열그림

미분양주택 추이가 2002년 12월부터 2007년 9월까지 지수적 증가추세를 보이고 있고 2005년도의 피크점이 특이하다.

꺾은선그래프에서 물결선을 쓸 때 우리는 특히 주의하여야 한다. 물결선 사용여부는 전적으로 자료의 특성과 꺾은선그래프의 작성목적에 달려있다. 즉 '자료가 어떤 자료이냐'와 '무엇을 꺾은선그래프에서 나타내고자 하느냐?'에 따라 물결선을 사용할 수도 있고, 안 되는 경우도 있다는 것이다.

꺾은선그래프는 다른 정보그래픽스인 막대그래프나 원그래프와는 달리 이차원 직각좌표계를 쓴다. 즉 수평축에 시간이라는 변수를 사용하고 수직축에 관심의 대상이 되는 변수를 사용하여 각 시계열 자료를 이차원 직각좌표계 안의 한 점으로 표시하게 된다. 그러므로 꺾은선그래프를 작성하고 분석할 때는 패턴(특히 점들 사이의 기울기)이 중요하고 수평축과 수직축을 동시에 중요하게 다루어야 한다.

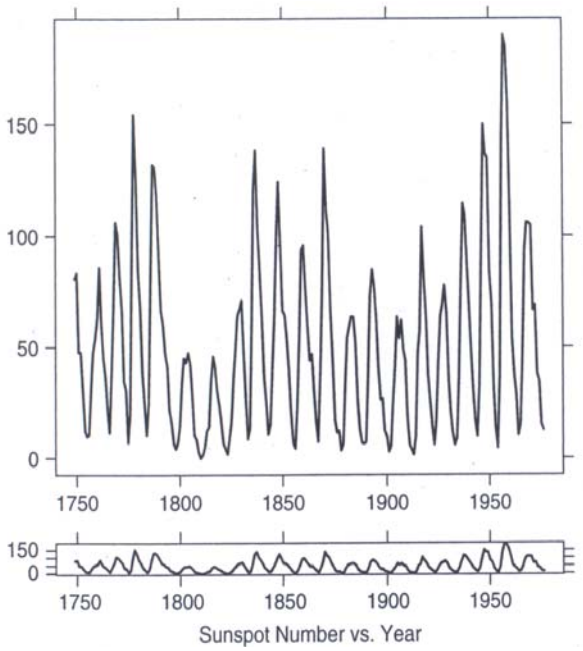
통계그래픽스, 특히 꺾은선그래프에서 모양모수의 영향은 매우 중요하다. 그래프의 전체적인 인상을 좌우할 뿐만 아니라 자료에 대한 시각적인 평가에도 영향을 미치기 때문이다. 그래프에서 모양모수(shape parameter)란 다음과 같이 정의된다.

$$sp = \frac{w}{h}$$

여기서, sp는 모양모수, h는 그래프의 높이(그래프에서 세로의 길이), w는 그래프의 폭(너비, 그래프에서 가로의 길이, 이 폭을 그래프의 길이라고도 함.)이다. 이 모양모수의 크기에 따라 꺾은선그래프 상의 점들 사이의 기울기가 다르게 나타나고 전체적인 패턴도 다르게 느껴질 수 있다. 우리들은 하나의 시계열 자료를 이용하여 다양한 모양모수를 적용하여 꺾은선그래프들을 그려보고 어떤 모양모수가 적당한지를 결정하도록 하는 훈련이 필요하다. Tufte(2001)는 모양모수에 대하여 다음과 같이 말하고 있다. "Playfair가 출간한 6개의 책에 나타나는 89개의 그

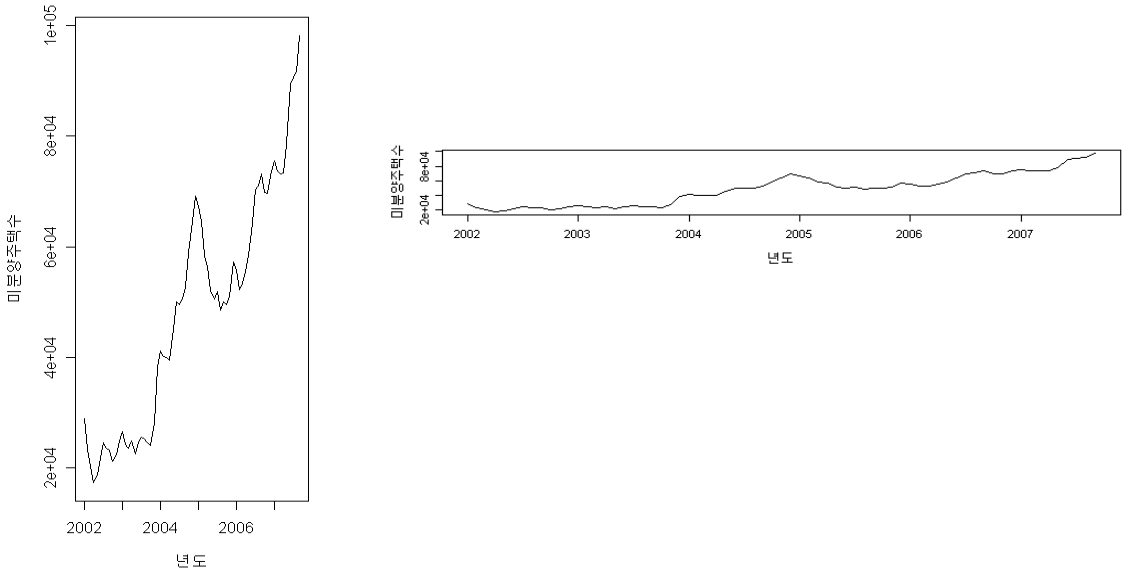
래프 중 92%가 높이보다 폭이 더 길다. (중략) 89개의 그래프 중 약 2/3가 모양모수가 1.4와 1.8 사이에 있고 이 사이에 든 그래프들 중 많은 그래프들이 폭의 길이 대 높이의 비가 황금비보다 더 크다." Playfair는 통계그래픽스의 아버지라고 일컫는 사람이다. 위의 인용 글에서 "높이보다 폭이 더 길다"라는 말은 '모양모수 값이 1보다 크다'라는 의미이다. 여기서 흥미로운 수인 황금비 $\phi = (\sqrt{5}+1)/2 \approx 1.618$가 언급되고 있다. 우리는 대표적인 모양모수 값으로서 $1, \sqrt{2} = 1.414, \phi = 1.618, \sqrt{3} = 1.732, 2$ 등을 생각하여 볼 수 있다.

다음 [그림 4.39]는 태양흑점 자료에 대한 꺾은선그래프이다. 아래 그래프의 모양모수 값은 18.25, 위 그래프의 모양모수 값은 1.16이다. 아래 그래프에서는 흑점 주기를 알기가 쉽고 위 그래프에서는 흑점 주기 안에서의 미세한 변동들을 알기가 쉽다. 이렇듯 모양모수가 달라지면서 그래프에서 우리가 얻을 수 있는 자료의 패턴이 달라진다.



[그림 4.39] 태양흑점 자료에 대한 꺾은선그래프

다음 [그림 4.40]은 앞에서 2000년 1월부터 2007년 9월까지의 미분양주택 현황 자료를 이용하여 그린 시계열 그림(모양모수=1.1)을 모양모수가 각각 0.4와 9.5로 만든 그림들이다. 모양모수가 0.4인 그림은 원 그림보다 기울기의 변화가 더 심하게 느껴지나 모양모수가 9.5인 그림은 원 그림보다 기울기의 변화가 많이 약화되어 있음을 알 수 있다.

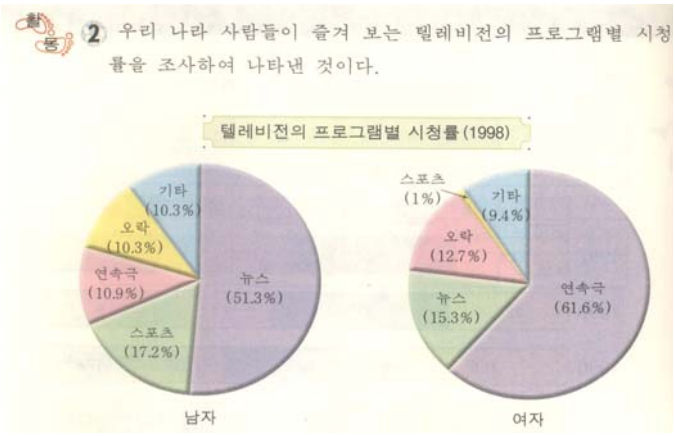


[그림 4.40] 미분양주택 현황 자료를 이용하여 그린 시계열 그림

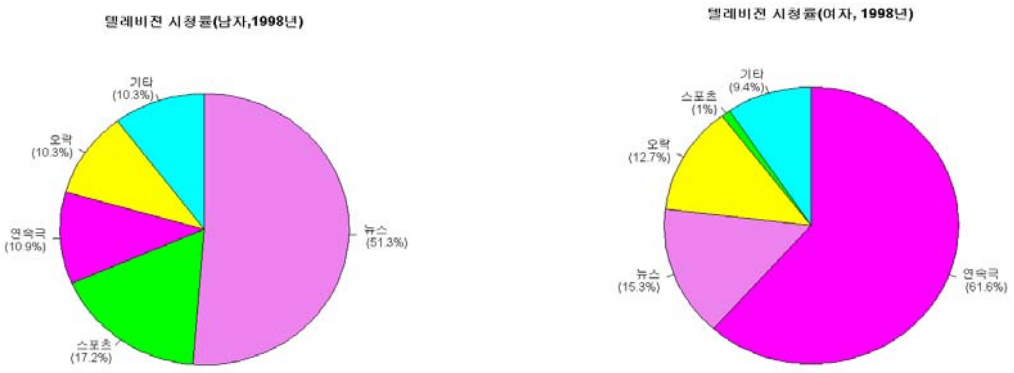
- 막대그래프에 줄임표시가 있으면 안된다. 다음 [그림 4.41]은 한미 FTA 농업부문 재정지원 내역을 나타내는 막대그래프들이다. 이 그래프들 중 농업체질개선에 관한 맞춤형 농정에 대응되는 막대에 줄임표시가 있다. 막대그래프에 이런 줄임표시를 하여 막대 길이를 줄이면 안된다. 막대그래프의 길이를 줄이면 자료의 왜곡이 생겨 오해의 소지가 있다.

[그림 4.41] 한미 FTA 농업부분 재정지원에 대한 막대그래프

• 초등학교 교과서에 보면 다음 [그림 4.42]와 같은 두 개의 원형그래프가 나타나 있다. 남녀 각각 순위대로 색깔을 배정하다보니 남녀 사이의 범주를 서로 비교하는 데 색깔이 방해를 하고 있다. 다음 [그림 4.43]과 같이 남녀 사이의 같은 범주에 같은 색깔을 배당하여 원그래프를 작성하는 것이 그래픽인식 작업상 더 수월하다.



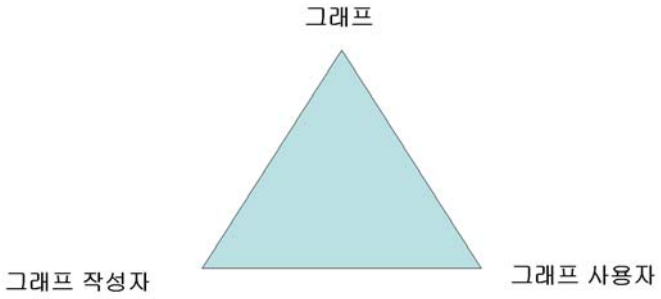
[그림 4.42] 텔레비전의 프로그램별 시청률을 나타내는 원형그래프



[그림 4.43] 텔레비전의 프로그램별 시청률을 나타내는 원형그래프(수정)

• 결론적으로 그림으로 자료를 표현하는 경우에는 자료가 간단하면 그림도 간단하여야 한다. 설사 자료가 복잡하여도 그림은 간단하게 보여야 한다. 그리고 자료를 왜곡시키려 들지 말라.

시각 정보 전달과정에서 그래프의 효용성은 그래프 사용자의 경험이나 지식, 또는 시각적 인지도에 의하여 좌우되기도 하지만 그래프의 질에 의해서도 좌우된다. 시각적 정보전달과정의 세 요소인 그래프, 그래프 작성자, 그래프 사용자 사이에는 다음 [그림 4.44]와 같은 관계가 성립한다.



[그림 4.44] 그래프, 그래프 작성자, 그래프 사용자 사이의 관계

그래프 작성자와 그래프 사용자 사이의 연결(link)이 가까우면 가까울수록 그래프 사용자가 추출해 낼 수 있는 정보의 양이 많아지고 정보의 질이 좋아질 것이다. 그래프 작성자가 질이 떨어지는 그래프를 작성하거나, 그래프 사용자가 대상 그래프에 대한 사전지식 또는 경험이 없으면 시각정보 전달과정은 심하게 파괴되거나 왜곡되고 만다. 매스컴은 광범위한 시청자나 독자층을 대상으로 하기 때문에 그래프 사용자가 갖는 약점을 항상 내포하고 있다. 그러므로 그래프 작성자는 더욱 질 좋은 그래프를 만들어야 하고 그러기 위해서는 그래프에 대한 지식은 물론이고 디자인에 대한 전문기술도 확고하여야 한다.

Schmid(1992)는 좋은 그래프가 가져야 할 성질을 다음과 같이 5가지로 요약하였다.

1. 정확성(accuracy)
2. 단순성(simplicity)
3. 명료성(clarity)
4. 꾸밈새(appearance)
5. 잘 디자인된 구조(well-designed structure)

이에 입각하여 좋은 그래프를 그리기 위하여 다음과 같은 일반규칙을 나열할 수 있다.

1. 자료를 나타내어라.
2. 그래프를 정확, 단순, 명료하게 그려라.
3. 독자들의 흥미를 끌게 하라.

또한, 그래프를 평가하는 수치적 측도들로서

1. 자료밀도지수(data density index, 자료의 개수/cm^2)
2. 모양모수
3. 인식왜곡지수

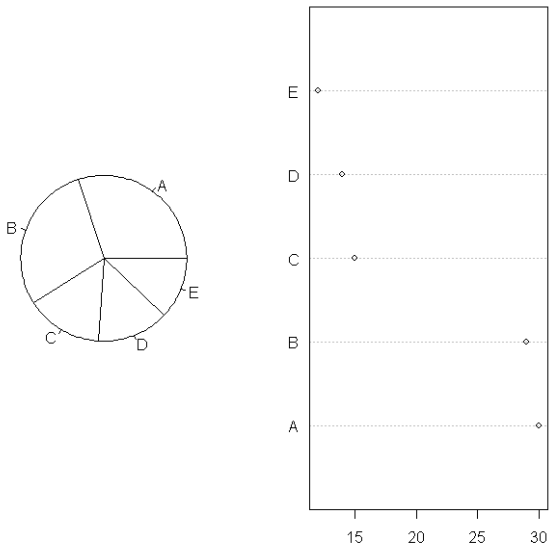
4. 자료잉크비(data ink ratio, 자료를 나타내는 데 쓰인 잉크의 양/그래프를 그리는 데 쓰인 잉크의 양)
5. 그림인식 작업을 위한 패러다임에서의 기본코드 서열

등을 고려하여 그래프를 작성하여야 한다. 그래프 담당자가 그래프를 그리면서 기호화한 정보를 매스컴의 시청자나 독자가 시각적으로 해독하는 것이 그림인식 작업인데 Cleveland(1985, 1990), Cleveland와 McGill(1984, 1986, 1987)은 그림인식 작업을 위한 패러다임을 제시하고 양적 자료를 기호화하는 것은 10개의 기본코드(elementary code)라고 주장하였다. 다음 [표 4.3]은 이 10개의 기본코드의 종류와 코드 사이의 서열(이 서열이 높을수록 그래프에서 양적 정보를 시각적으로 추출하기가 쉽다.)을 나열한 표이다.

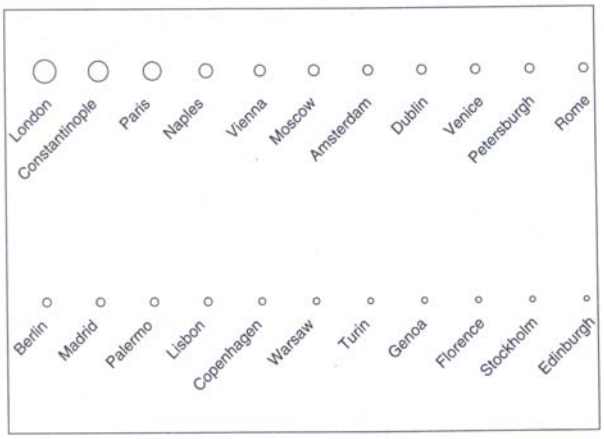
서열	코드
1	같은 축에 나열하기
2	동일하나 다른 축에 나열하기
3	길이
4	각도
5	기울기
6	면적
7	체적
8	밀도
9	채도(彩度)
10	색조

[표 4.3] 그림인식작업을 위한 기본코드의 종류와 코드 사이의 서열

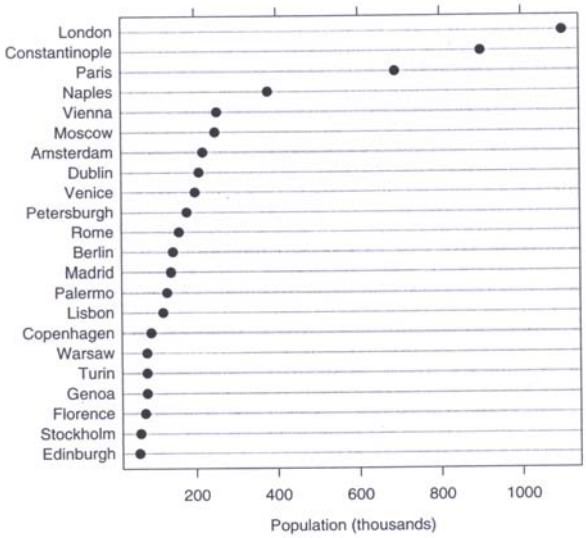
기본코드에 대한 예를 들어 보자. 다음 그래프 중 왼쪽의 원형그래프에서 각 범주의 크기(A=30, B=29, C=15, D=14, E=12)가 부채꼴의 각도가 나타난다. 우리는 각도의 크기를 잘 판정할 수가 없다. A가 큰가, B가 큰가? C가 큰가, D가 큰가? 이 자료를 점차트로 그리면 오른쪽 그래프와 같다. 각 범주의 크기를 쉽게 알 수 있다. 왜 이런 현상이 나타났을까? 원형그래프(기본코드 '각도'의 서열: 4)에서 점차트(기본코드 '같은 축에 나열하기'의 서열: 1)로 바뀌면서 서열이 4에서 1로 올라감으로 각 범주의 크기를 쉽게 알 수 있게 된 것이다.



또 다른 예를 보자. 다음 그래프는 Playfair(2005, 초판 1801)에 나타나는 19세기 초 유럽 여러 도시의 인구수를 나타내는 거품그래프이다. 거품들의 면적을 비교하기가 쉽지 않다. 특히 작은 거품들의 면적은 더 어렵다.



이 자료를 점차트로 그리면 다음 그래프와 같다. 각 인구수의 크기를 쉽게 알 수 있다. 거품그래프(기본코드 '면적'의 서열: 6)에서 점차트(기본코드 '같은 축에 나열하기'의 서열: 1)로 바뀌면서 서열이 6에서 1로 올라감으로 각 인구수의 크기를 쉽게 알 수 있게 된 것이다.



매스컴의 시청자나 독자들은 앞에서 언급한 정보그래픽스에 대한 통계적 오용의 형태를 숙지하고 매스컴에 나타나는 정보그래픽스에 대하여 비판적인 시각과 안목을 길러야 한다.

학습요약

제 4장에서는 자료에 담겨져 있는 정보를 그림으로 표현하는 방법에 대해 알아보았다. 통계학의 이론이 뒷받침되는 그림을 그리는 것은 많은 노력이 필요하다. 따라서 정보를 추출하는 작업은 어렵다. 구간척도나 비율척도로 수집된 자료에 쓸 수 있는 히스토그램을 비롯한 몇 가지 기법을 알아보았고 그 문제점을 지적하였다. 간단한 히스토그램이라도 사용자에 따라 왜곡될 수 있는 사실은 매우 중요하다. 그리고 명목척도이거나 순서척도로 수집된 경우 자주 쓰이는 방법인 파이차트나 막대차트에 대해서도 언급되었다. 시계열 자료인 경우도 살펴보았다. 자료가 가지고 있는 정보를 정확하게 추출하는 그림을 그리는 기준으로 몇 가지를 제안하였는데 제일 문제가 되는 것은 자료가 가지고 있는 만큼만 정보를 표현하여야 하는데 왜곡이 의도적으로 일어난다는 사실이다. 제 6장에서는 변수 간의 연관성을 나타내주는 그림에 대해서도 언급이 있을 것이다.

4장 연습문제

4.1 다음 자료는 2005년 4월 서울시에서 실시한 강동 지역 환경시설 유치에 대한 주민들 의견 중 일부를 발췌한 자료이다. 30명의 표본이 표에 기록이 되어 있으며 변수는 6개로 자료로 구성되어 있다. 변수는 응답자의 나이(age), 성별(sex), 지역(region), 자녀수(children), 봉급(salary, 단위: 만원), 의견(opinion)으로 구성되어 있다. 나이는 범주형(35세미만, 60대미만, 60대 이상) 자료(age_cat)로 그리고 성별은 여자인 경우는 1 그리고 남자인 경우는 2로 다시 기록(sex_mod)하였다. 그리고 의견은 1부터 5까지의 척도로 점수가 높으면 높을수록 환경시설 유치에 반대하는 의견이 강하다. <주민의견.xls>

	A	B	C	D	E	F	G	H	I
1	서울 강동지역에 대한 환경시설유치에 대한 주민의견								
2									
3	age	sex	region	Children	salary('10000)	opinion		age_cat	sex_mod
4	61	F	강동	2	6,200	1		elderly	2
5	37	M	강동	2	5,200	5		middle-aged	1
6	32	F	강동	3	8,140	1		young	2
7	65	F	강동	2	4,960	1		elderly	2
8	40	M	강동	3	4,770	4		middle-aged	1
9	32	F	강동	1	5,990	4		young	2
10	38	F	강동	2	3,900	2		middle-aged	2
11	48	M	강동	1	6,150	2		middle-aged	1
12	40	M	강동	1	4,450	3		middle-aged	1
13	44	M	강동	2	4,520	3		middle-aged	1
14	57	F	강동	2	3,670	4		middle-aged	2
15	21	F	강동	2	5,430	2		young	2
16	49	M	강동	1	6,210	4		middle-aged	1
17	34	M	강동	0	7,800	3		young	1
18	38	M	강동	1	4,330	1		middle-aged	1
19	35	M	송파	1	6,540	5		middle-aged	1
20	35	M	송파	0	6,320	3		middle-aged	1
21	33	F	송파	3	4,630	5		young	2
22	45	M	송파	1	4,590	5		middle-aged	1
23	57	M	송파	1	4,810	4		middle-aged	1
24	38	F	송파	0	5,810	3		middle-aged	2
25	37	F	송파	2	5,600	1		middle-aged	2
26	42	F	송파	2	5,340	1		middle-aged	2
27	49	M	송파	0	4,320	5		middle-aged	1
28	52	M	송파	1	4,410	3		middle-aged	1
29	27	M	송파	3	4,540	2		young	1
30	40	M	송파	0	5,900	4		middle-aged	1
31	63	M	송파	2	5,390	1		elderly	1
32	48	F	송파	2	3,100	4		middle-aged	2
33	40	M	송파	0	3,770	1		middle-aged	1

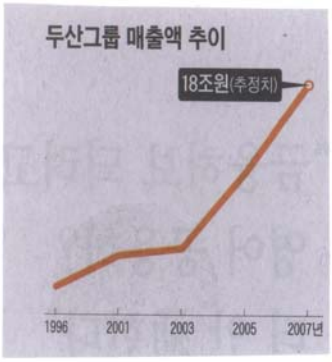
각각의 변수에 대해 제 4장에서 배운 내용을 바탕으로 의미있는 그림을 그려보라. 지역별 의견의 차이, 성별 의견의 차이, 천만원 단위로 구분한 봉급수준에 따른 의견의 차이, 어린아이 수에 따른 의견의 차이 등이 있다.

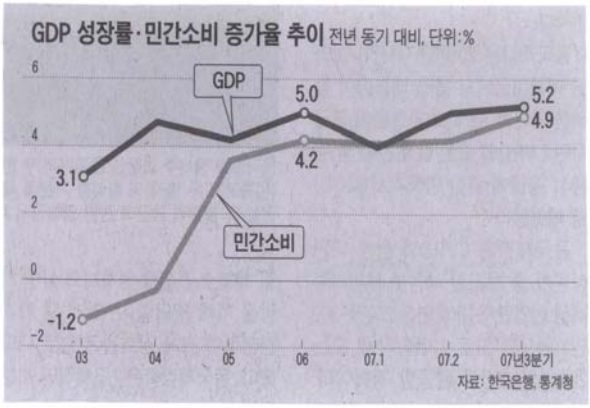
4.2 다음 자료는 도시 근로자 중 화이트 컬러 직장에 다니는 미혼 남성에게 물어본 조사 자료 중 일부를 무작위로 발췌하여 기록한 것이다. 봉급에서 차지하고 있는 세 변수의 비율에 대해 파이차트를 그려보라. 그리고 봉급을 제외한 세 변수에 대한 히스토그램을 그려보아라.

<소비형태.xls>

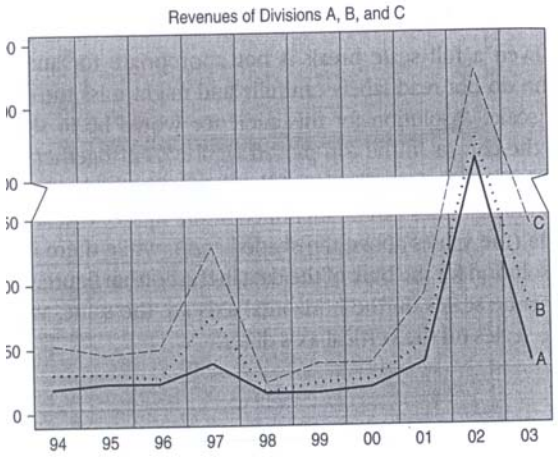
	A	B	C	D
1	**100가구 소득 소비형태**			
2				
3	봉급	문화	스포츠	외식
4	54,600	1,020	990	1,510
5	57,500	1,100	460	1,180
6	53,300	900	780	1,590
7	43,500	570	860	1,750
8	57,200	900	1,390	2,120
9	63,400	820	1,880	3,090
10	58,500	1,340	710	1,540
11	55,600	1,250	680	1,800
12	61,300	1,190	1,220	2,330
13	61,100	640	1,480	2,670
14	77,200	900	820	2,850
15	58,800	710	1,080	2,200
16	62,900	1,240	1,230	2,430
17	61,900	1,270	1,000	2,110
18	76,500	1,180	690	1,820
19	50,300	810	1,490	2,100
20	45,900	840	730	920
21	61,900	1,290	1,050	2,480
22	56,700	780	970	1,930
23	43,300	910	1,120	1,720
24	63,000	560	1,570	1,990
25	39,000	1,100	830	1,420
26	55,000	610	1,260	1,890
27	51,600	930	980	1,470
28	48,600	800	1,300	1,740
29	68,000	1,010	1,410	3,100
30	65,400	790	1,520	2,060

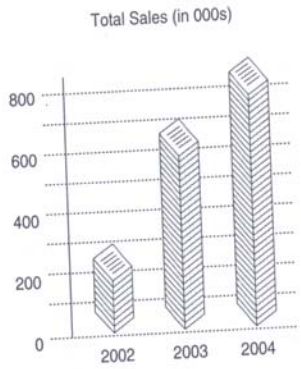
4.3(그래프의 오용) 다음 그래프들을 보고 어떤 문제점이 있는지 밝혀라.

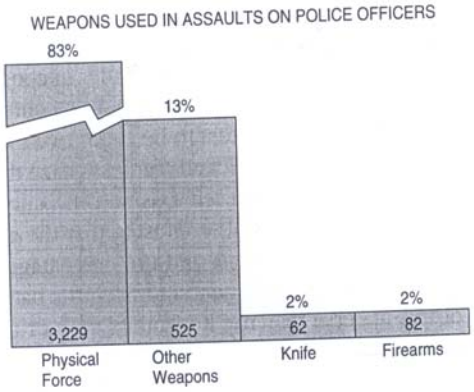


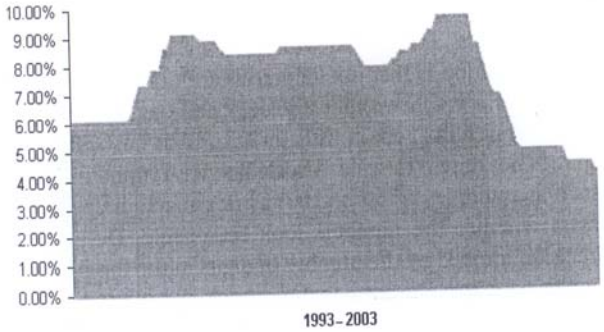


4.4(그래프의 오용) 다음 그래프들을 보고 무엇이 잘못 되었는지 밝혀라.









4장 실습문제

4.1 가정에서(또는 인터넷에서) 구독하고 있는 최근 한 달 분의 신문기사들을 수집하여 다음과 같은 실습을 해보자.

1) 각종 그래프들 중 잘못 그려진 그래프들을 찾아 스크랩북을 만들고 오용의 패턴에 대하여 기록하라.

2) 신문기사 중 통계의 오용이 일어났다고 판정되는 신문기사들을 찾아 스크랩북을 만들고 오용의 패턴에 대하여 기록하라.

3) 각자의 스크랩북을 서로 교환하여 보고 토론을 실시하라.

<목적> 통계의 오용은 우리의 일상생활에서 빈번하게 일어난다. 이러한 사실을 깨닫게 하고 이를 개선할 방법을 강구하도록 한다.

4.2 [그림 4-12]의 피라미드 히스토그램을 통계청 자료를 다운받아 그려보라. 어떻게 그릴 수 있겠는가?

<목적> 실제 자료로 직접 실용적인 그래프를 그릴 수 있게 한다.

4.3 다음은 미국의 전설적인 야구선수 베이브 루스와 마크맥과이어의 현역시절 홈런수이다. 적절한 그래프를 그려 두 선수의 홈런수를 비교해 보라.

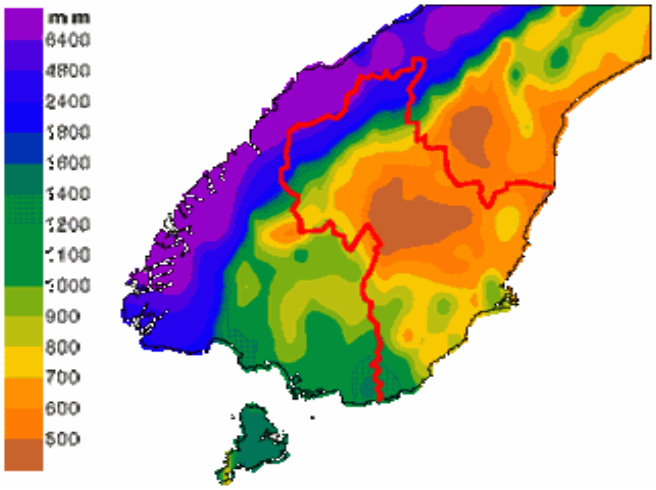
베이브 루스 : 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22
마크 맥과이어 : 49 32 33 39 22 42 9 9 39 52 58 70 65

쉬어가기

통계적 그림도 진화한다.

요즘은 자료를 시각적인 상태로 표현하는 분야가 매우 발달한 상태이다. 컴퓨터의 기술적 진화와 함께 예전에는 생각하지 못한 기법들이 보편화되기 시작했다. 예를 들어 지역간 평균 온도를 표시한다고 한 때 단지 지역을 명목변수로 하여 막대그림을 보여 주는 것으로는 많은 관리자는 만족해하지 않는다.

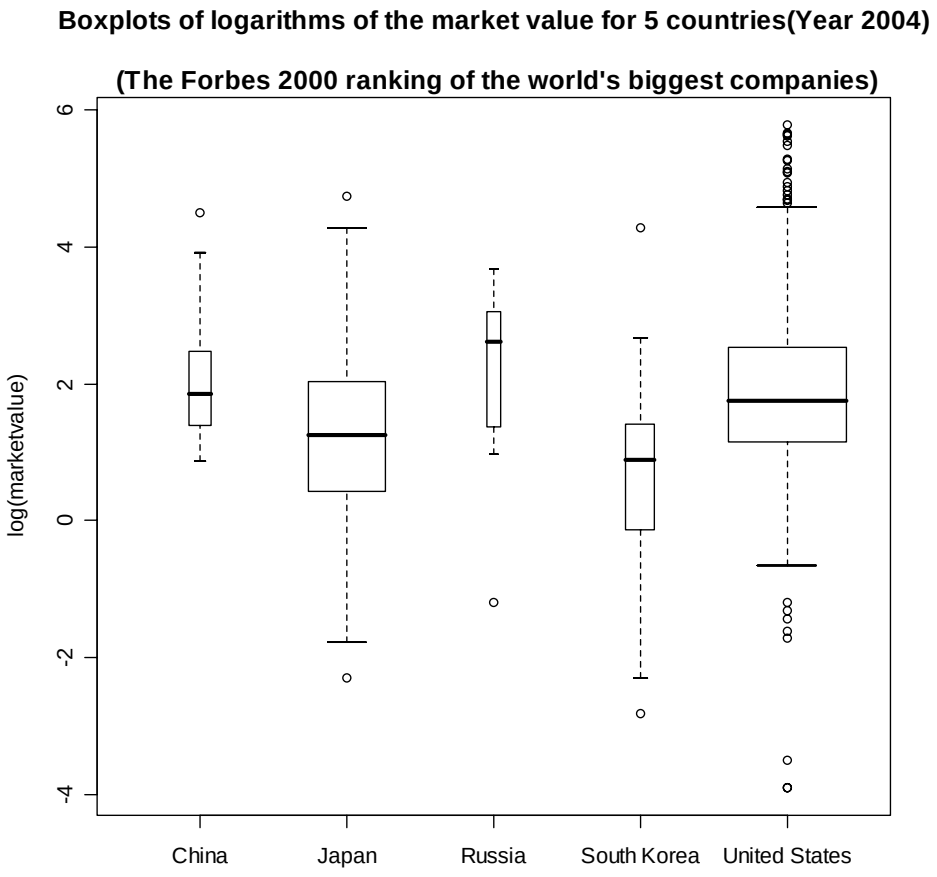
아래 [그림 4.45]는 지리 정보를 이용한 지역간 평균을 각기 다른 색깔로 표현한 것이다. 이런 기법은 공간정보를 이용한 자료의 표현이라 한다. 그러나 이런 호사스러운 기법을 이용하지 않더라도 자료에 담겨 있는 정보는 얼마든지 추출할 수 있다. 사실 많은 관리자는 평균의 의미를 제대로 인식하지 못하는 경우도 많다.



[그림 4.45] 색깔로 표현한 지역의 평균 강우량

제 5 장

자료의 분석도구와 의미



차 례

학습목표

시험을 치른 후 시험문제 및 답안지를 확인하는 많은 사람들은 제일 처음 하는 질문으로 “평균이 얼마나 됩니까?”라고 질문을 던질 것이다. 이와 같이 질문하는 것은 두 가지 목적이 있다. 첫째는 반 점수의 평균을 확인함으로써 전체적인 점수의 분위기를 파악하고자 하는 것이고 두 번째는 평균하고 자신의 점수와의 비교를 통해 자신이 얼마나 평균에서 멀리 떨어져 있는지 위치를 파악하는 것이다. 왜냐하면 사람들은 평균이라는 점수가 반의 점수를 대표한다고 믿고 있기 때문이다. 이렇듯 모든 점수의 분포를 알고 있지 않더라도 자신들의 시험 결과가 얼마나 될 수 있는지를 파악할 수 있다. 또한 점수 폭을 알기 위해 성적의 최저 점수 및 최대 점수를 알고자 할 것이다. 이와 같이 몇 개의 요약된 숫자만 추출하여도 자료의 형태를 파악할 수 있다. 5장에서는 자료를 파악하는 기술적인 방법을 소개할 것이다.

5.1 중심경향 측정을 위한 수치적인 요약

이번 장에서 설명하는 수치적인 요약(numerical summary)에 관련된 방법들은 대부분 한 개의 변수의 분포 특징을 설명하는 목적을 가지고 있다. 변수의 개수가 2개 이상일 경우 고려하여야 할 사항들은 다음 장에서 다룬다.

먼저 중심경향(central tendency)에 관련되어 많이 쓰이는 요약 수단을 보도록 하자. 여기에는 기본적으로 3가지가 있다. 산술평균(mean), 중앙값(median), 그리고 최빈값(mode)이다.

산술평균: 산술평균(arithmetic mean)은 통상적으로 $\bar{x}$ 로 표기한다. 그리고 줄여서 평균이라 부른다. 이는 변수의 모든 값들의 평균(average)을 의미한다. 자료가 모집단 전체인 경우는 모집단평균(population mean)이라 하고 자료가 모집단에서 추출된 표본이라면 표본평균(sample mean)이라 한다. 이 구분은 그렇게 중요하지 않다. 여기에 대해서는 후에 통계적 추론을 논할 때 다시 언급하도록 한다. 평균의 공식은 식 (5.1)과 같다.

$$\bar{x} = \frac{\sum_{i}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad (5.1)$$

여기서 x_i 는 i 번째 관측 값, 그리고 n은 관측값의 개수이다. 식 (5.1)은 단순히 모든 관측값을 모두 더한 다음 관측값의 개수로 나누어주는 것이다. 모든 관측값이 평균을 계산하는데 관여함을 알 수 있을 것이다.

예제 5.1 다음 [표 5.1]은 2005년 4월 서울시에서 실시한 강동 지역 환경시설 유치에 대한 주민들 의견 중 일부를 발췌한 자료이다. 30명의 표본이 표에 기록이 되어 있으며 변수는 6개로 응답자의 나이(age), 성별(sex), 지역(region), 자녀수(children), 봉급(salary, 단위: 만원), 의견(opinion)등으로 구성되어 있다. 마지막으로 나이는 범주형(35세미만, 60대미만, 60대 이상) 자료(age_cat)로 다시 기록하였고 성별은 여자인 경우는 1, 그리고 남자인 경우는 2로 다시 기록(sex_mod)하였다. 그리고 의견은 1부터 5까지의 척도로 조사하였는데 점수가 높으면 높을수록 환경시설 유치에 동의하는 의견이 강하다. <주민의견.xls>

	A	B	C	D	E	F	G	H	I
1	서울 강동지역에 대한 환경시설유치에 대한 주민의견								
2									
3	age	sex	region	Children	salary('10000)	opinion		age_cat	sex_mod
4	61	F	강동	2	6,200	1		elderly	2
5	37	M	강동	2	5,200	5		middle-aged	1
6	32	F	강동	3	8,140	1		young	2
7	65	F	강동	2	4,960	1		elderly	2
8	40	M	강동	3	4,770	4		middle-aged	1
9	32	F	강동	1	5,990	4		young	2
10	38	F	강동	2	3,900	2		middle-aged	2
11	48	M	강동	1	6,150	2		middle-aged	1
12	40	M	강동	1	4,450	3		middle-aged	1
13	44	M	강동	2	4,520	3		middle-aged	1
14	57	F	강동	2	3,670	4		middle-aged	2
15	21	F	강동	2	5,430	2		young	2
16	49	M	강동	1	6,210	4		middle-aged	1
17	34	M	강동	0	7,800	3		young	1
18	38	M	강동	1	4,330	1		middle-aged	1
19	35	M	송파	1	6,540	5		middle-aged	1
20	35	M	송파	0	6,320	3		middle-aged	1
21	33	F	송파	3	4,630	5		young	2
22	45	M	송파	1	4,690	5		middle-aged	1
23	57	M	송파	1	4,810	4		middle-aged	1
24	38	F	송파	0	5,810	3		middle-aged	2
25	37	F	송파	2	5,600	1		middle-aged	2
26	42	F	송파	2	5,340	1		middle-aged	2
27	49	M	송파	0	4,320	5		middle-aged	1
28	52	M	송파	1	4,410	3		middle-aged	1
29	27	M	송파	3	4,540	2		young	1
30	40	M	송파	0	5,900	4		middle-aged	1
31	63	M	송파	2	5,390	1		elderly	1
32	48	F	송파	2	3,100	4		middle-aged	2
33	40	M	송파	0	3,770	1		middle-aged	1

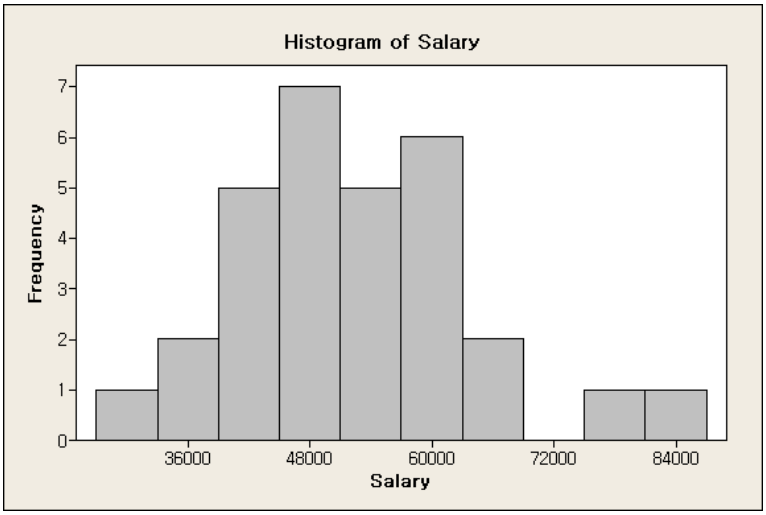
[표 5.1] 강동지역 환경시설 유치자료

[표 5.1]에서 변수 봉급(salary)에 대한 평균을 구하게 되면, 즉 모든 값을 다 더한 다음 자료의 개수 30으로 나누면 5,266.33이 나온다. 봉급 단위는 만원이므로 5,266.33만원이 평균봉급수준이 된다. 만약 엑셀을 통해 이를 구해보면 커서를 적절한 위치에 놓고 평균을 구하고자 하는 자료의 적절한 셀 범위(cell range)를 지정한 다음

=average(cell range)

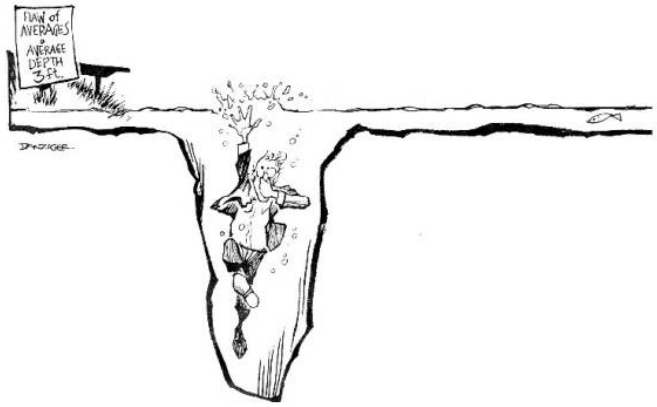
을 입력하면 된다. 실제 파일에서는 =average(E4:E33)이다. 아니면 범위 E4:E33이 salary로 이미 이름이 지정되었기 때문에 =average(salary)를 수행하면 된다. ■

평균이 봉급이라는 변수를 대표(representative)한다고 이야기할 수 있을까? 평균은 기본적으로 생김이 대칭일 때 쓸 수 있다. 그러나 본 봉급의 히스토그램인 [그림 5.1]은 약간 오른쪽으로 왜도가 있는 모양을 하고 있었기 때문에 평균이 봉급의 대표적인 값이라고 이야기하기에는 무리가 따른다. 이런 경우는 평균보다는 다른 중심을 측정하는 값이 필요하다.



[그림 5.1] 봉급에 대한 히스토그램

자료에 아주 크거나 작은 관측값(들)이 섞여 있는 경우에는 평균은 부풀려지거나 축소되어 나타나기 때문이다. [그림 5.2]에서 연못의 깊이를 평균으로 계산하여 팻말을 붙여 놓았다면 이는 매우 잘못된 기준이 되는 것이다.



[그림 5.2] 평균의 오류

평균의 오류에서 어느 정도 벗어나는 요약방법이 중앙값(median)이다.

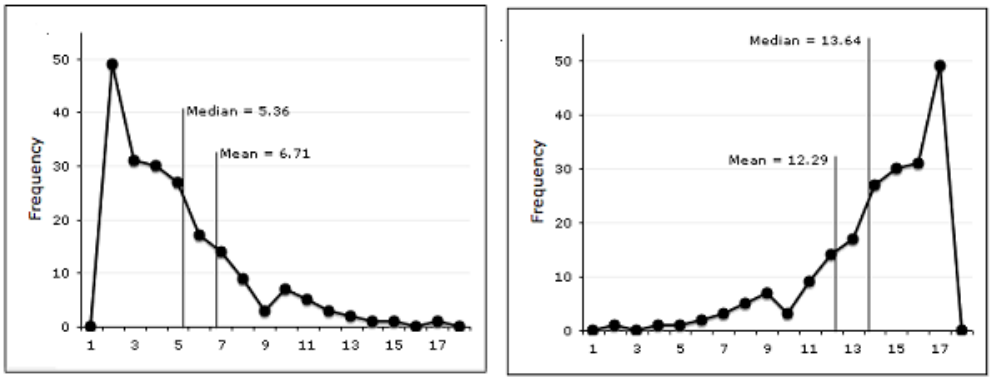
중앙값: 중앙값(median)은 자료가 작은 순서부터 큰 순서로 나열하였을 때 가운데(middle)에 있는 관측값을 의미한다. 관측값의 개수가 홀수이면 중앙값은 가운데 값이 된다. 예를 들어 관측값의 개수가 9이면 5번째로 작은 값(혹은 5번째로 큰 값)이 중앙값이 된다. 만약 관측값의 개수가 짝수이면 가운데 있는 두 개 관측값의 평균이 중앙값이 된다. 예를 들어 관측값의 개수

가 10이라면 5번째와 6번째로 작은 값의 평균이 중앙값이 된다. [표 5.1]에서 봉급변수의 관측값의 개수는 30이므로 15번째와 16번째 작은 값의 평균이 중앙값이 된다. 엑셀에서

=median(salary)

으로 명령문을 입력하여 실행시키면 5,080만원의 값을 얻을 것이다.

봉급이라는 변수가 좌우 대칭의 모양을 하고 있었다면 평균과 중앙값은 일치한다. 그러나 봉급이라는 변수는 오른쪽으로 왜도가 있었기 때문에 중앙값은 평균의 왼쪽에 위치하고 있다. 만약 왼쪽으로 왜도였다면 중앙값은 평균의 오른쪽에 위치하고 있었을 것이다. [그림 5.3]을 참조하기 바란다.



[그림 5.3] 평균과 중앙값의 비교

자료 중 한 관측값이 입력하는 과정에서 착오를 일으켜 굉장히 큰 값으로 기록되었다 하더라도 중앙값은 평균과 달리 전혀 영향을 받지 않고 분포의 가운데 있는 값으로 대표되는 점이 평균과 다르다. 그러나 중앙값은 순서를 정하는 과정에서 자료의 원래 값들을 무시할 수밖에 없다. 즉, 자료가 가지고 있는 정보가 일부 잃어버리는 결과를 초래한다. 따라서 중앙값은 순서척도인 경우에 제일 적합한 요약 방법이 된다.

예제 5.2 남자는 여자보다 성적 파트너를 더 많이 가진다.

2007년 6월에 발표된, 1992년부터 2002년까지 조사된 20세부터 69세까지의 미국 남녀 성인들을 대상으로 지금까지 몇 명의 이성과 성적 행위를 가졌나를 분석하는 보고서를 뉴욕타임즈의 한 과학자가 비판하면서 기사화했던 이야기다. 보고서 안에는 남자는 7명, 그리고 여자는 이보다 적은 4명의 이성과 성적 파트너를 가졌다는 내용이 있다. 그러면 이 기사에서 숫자가 의미하는 것은 무엇인가? 보고된 숫자가 평균이라면 서로 다른 이성이 갖는 성적 파트너의 숫자는 정확히 일치해야 한다, 혼자서만 이성적인 성적 행위를 할 수 없기 때문이다. 즉, 평균을

가지고 두 숫자를 비교한다면 아무런 의미가 없다.

그러면 이 보고서에서 남자와 여자의 성적 행위의 차이를 나타내는 숫자는 무엇인가?

예를 들어 보자. 어느 마을에 100명의 여자와 100명의 남자가 있다. 여자의 90명은 처녀이고 10명은 성관계를 가진 여자이다. 그리고 남자 100명은 모두 다 성적 경험을 가진 사람이다. 남자는 단 한명의 여자와 성적 관계를 가졌다면 성관계를 가진 여자는 10명의 서로 다른 남자와 성적인 관계를 가진 것이다. 남자의 평균은 100회 /100명 = 1이 나온다. 그러면 여자의 평균은 역시 100회/100명 = 1이다. 그렇다면 중앙값은 어떻게 되는가? 남자는 당연히 1이고 여자인 경우에는 0이다. 그렇지 않은가? 위의 보고서는 남자와 여자의 값이 다른 것으로 보다 중앙값으로 발표한 것이 아닌가 생각된다. ■

평균을 쓸 것인가, 중앙값을 쓸 것인가는 분석자가 자료의 내용에 따라 달리 선택해야 하는 문제이다.

예제 5.3 **중앙값이 극단값을 이긴다. - 유머 한마디**

아들의 테니스 실력을 향상시켜주기 위해 아버지는 아들에게 “나하고 테니스를 치든지 코치하고 테니스를 치든지 선택하여 3경기 중 두 번 연속해서 이기면 상금을 준다”고 약속하였다.

단 아버지와 치면 다음에는 코치와 테니스를 쳐야 한다. 마찬가지로 코치와 치면 다음에는 아버지와 쳐야 한다. 그럼 아들은 다음과 같은 두 가지 선택권이 주어진다.

아버지-코치-아버지
코치-아버지-코치

아들은 어떤 선택을 하여야 하는가? 테니스를 잘못 치는 아버지를 상대로 두 번 경기를 하여야 하는가? 아님 자기보다 실력이 월등한 코치를 상대로 두 번 경기를 하여야 하는가? 이 문제는 테니스 실력이 별로 없는 아버지를 중앙에 위치하고 코치와 두 번 겨루는 것이 상금을 탈 확률, 즉 두 번 연속해서 이길 확률이 높다. 항상 그렇다. ■

최빈값: 최빈값(mode)은 제일 빈도가 많이 일어나는 관측값을 의미한다. 그러나 변수가 연속형이라면 최빈값은 의미가 없다. 왜냐하면 기껏 해보아야 한 두 개의 빈도가 우연치 않게 나타나는 관측값이 최빈값일 공산이 크기 때문이다. 따라서 [표 8.1]에서 봉급이라는 연속형 변수 말고 의견(opinion)이란 변수를 가지고 보자. 의견이라는 변수는 수치형이지만 범주형 변수이다. 의견이라는 변수 중 제일 많이 나타난 값은 무엇인지 엑셀을 통해서 보면

=mode(opinion에 해당하는 셀 범위)

로 나타나는 최빈값은 1이다. 즉 30명의 응답자 중 제일 많은 의견은 전혀 동의하지 않음(“strongly disagree")이라는 의견이었다.

요약하면 평균은 자료가 비율척도이거나 구간척도에 제일 적합한 중심경향을 대표하는 값이지만 자료가 대칭이 아니거나 혹은 자료 중에 정상적인 자료로 분류되지 못하는 매우 큰 값이나 작은 값이 섞여 있으면 자료를 대표하는 값이라고 말할 수는 없다. 중앙값은 자료가 비율척도나 구간척도인 경우도 쓸 수 있으나 중앙값의 정의상 순서척도인 경우 제일 적합하다. 그리고 최빈값은 연속형 자료에는 사용치 못하고 이산형 자료에는 사용 가능하다. 물론 연속형 자료라도 히스토그램을 작성하였을 때는 도수가 가장 많은 계급값(히스토그램에서 높이가 가장 높은 기둥의 가운데값)이 최빈값이 된다. 그리고 순서 척도에도 사용이 가능하나 명목척도에 제일 적합하다.

이 세 가지 외에도 자주 언급되는 값들이 있어 명시한다.

기하평균: 어느 은행원의 작년도 임금 인상률이 5%이고 올해의 인상률이 15%라 하면 2년간 평균 인상률은 어느 평균을 사용하여야 하는지 계산하여 보자. 결론적으로 이 경우는 산술평균이 아닌 기하평균의 개념으로 평균인상률을 결정해야 한다.

예를 들어 재작년 봉급이 3,000(천만 원)이라면 작년은 봉급이 3,000 × (1+0.05) 인 3,150이고 올해 봉급은 3,150 × (1+0.15) = 3,622.50이 된다. 그렇다면 2년간 평균인상률에 의한 계산은 어떻게 이루어지는가?

3,622.50 = 3,000 × (1+평균인상률) × (1+평균인상률)

이 식에서 평균인상률은

$$\text{평균인상률} = \sqrt{3{,}622.50/3{,}000} - 1$$

인 9.8863%가 된다. 혹은

$$\sqrt{(1+0.05)(1+0.15)} - 1 \tag{5.2}$$

로 구해진다. 식 (5.2)의 첫 번째 항이 기하평균에 대한 일반적인 식이다.

만약 자산가치의 배수의 평균을 구한다면 구해야하는 개념이 기하 평균인 것이다. 즉, 첫해에는 1.05배 자산이 증가했으며 두 번째 해는 자산이 1.15배 증가했으므로 자산은 평균적으로 1.098863 배 커졌다고 이야기할 수 있다.

산술 평균인 (1.05+1.15) / 2 = 1.1과 비교하면 기하 평균은 항상 산술평균보다 적은 값을 가진다. 따라서 매우 극단적인 값에 의한 영향을 축소시켜 주는 역할을 한다. 이는 수학적으로 1.05와 1.15를 로그를 취한 다음 평균을 구하고 다시 역으로 값을 구하는 것과 같다. 엑셀에서 다음과 같은 명령문을 입력하여 확인하여 보아라.

= exp(average(ln(1.05), ln(1.15)))

참고로 n기간 동안 적용한 평균 인상률은 다음과 같이 계산하면 된다.

$$\sqrt[n]{\frac{\text{value at ending period}}{\text{value at beginning period}}} - 1$$

이와 같은 기하평균은 이자율이나 성장률을 구하는데 국한하지 않고 의외로 광범위하게 많이 쓰인다. 예를 들면 환경을 모니터링하는데 있어 박테리아의 함유율 같은 자료는 매우 극단값을 많이 포함하고 있기 때문에 로그를 취한 다음 평균을 내고 역으로 함유율을 구하는 절차가 일반화되어 있다. 비율이라든지, 퍼센티지 등과 같은 자료에 있어서는 흔한 일이다. 측정한 단위를 로그 단위로 바꾼 다음 평균을 내고 다시 역을 취하여 원래의 단위로 돌아간다면 기하평균인 것이다.

조화평균 : 자료 $x_1, x_2, x_3, \cdots, x_n$의 조화 평균 H는 다음과 같이

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}{n} \tag{5.3}$$

정의된다.

간단하게 설명하여 보자. 두 값 a, b의 조화평균은 식 (5.3)에 대입하여 정리하면 2ab/(a+b)로 구해지는데 음악의 음계와 관련된다.

조화평균은 피아노의 음계와 연관이 있다.

예를 들어 현악기에서 원래 현의 길이를 1이라고 할 때 길이를 1/2로 줄이면 한 옥타브 높은 음이 된다. 여기서 1과 1/2의 조화평균을 구하면 =1/(1+1/2)=2/3가 된다. 그리고 현악기의 현의 길이를 2/3로 하면 5도 높은 음을 얻게 되는데, 1도와 5도, 즉 '도'와 '솔'은 잘 어울리는 음이다. 이처럼 조화평균(harmonic mean)은 하모니를 이루는 조화로운 음을 만든다는 의미에서 붙여진 이름이다.

다른 예를 들어 보자. 일정한 거리(1km)를 두 자동차가 통과한 시간으로 기록을 하였다면 두 자동차의 분당 속도의 평균은 얼마나 될까? 속도는 단위가 분당 km인데 반해 자료는 거리당 시간이다. 첫 번째 자료가 1분 그리고 두 번째 자료가 1분 30초라면 분당속도는 각각 1킬로와 1/1.5=2/3 킬로가 된다. 그리고 평균을 구하면 (1+2/3)/2 =5/6가 되는데 이것의 역이 조화평균이다. 즉, 1 킬로를 달린 두 자동차의 평균시간은 6/5 분이 되는 것이다. 산술평균은 1.25분, 그리고 기하평균은 1.224745분으로 모두 조화평균보다는 높게 나온 수치이다.

예제 5.4 숫자는 표현에 따라 다르다.

2006년 6월 25일 LA 타임스 신문기사에 의하면 "지금까지 이라크 전쟁이 발발한 후 죽은 민간인의 수는 약 5만명이다."라고 발표하였다. 이는 지난 3년 동안 월 평균 1,389명에 해당하는 숫자이며 이는 인구 비례로 미국에서 월 15,833명이 죽어나가는 것과 비슷하다고 발표하

였다. 이라크의 인구가 약 2,600만 명이므로 이라크에서 지난 3년간 죽은 민간인의 비율은 약 0.19 퍼센트가 된다. 그러면 왜 퍼센트를 가지고 발표하지 않고 월평균으로 발표하거나 더 나아가 미국 인구에 비례하여 사망자를 발표하는가?

또 다른 사례를 보자. 위 신문기사가 나가고 정확히 3일 후 깅그리치란 위원은 Meet the press란 시사 프로그램에 나와 중동내전의 심각성을 알리기 위해 "만약 마이애미 시가 로켓 공격을 매일 받는다고 생각하여 보라"라고 운을 띄면서 팔레스타인의 헤즈볼라 공격으로 인해 이스라엘에서 일 평균 8명이 죽는데 이는 미국 시민이 매일 500명씩 죽는 것과 같다고 하였다.

왜 이런 비유를 하는 것일까? 반대로 르완다에서 4명 중의 3명이 내전으로 죽어 나간다는 기사는 사망 인구수를 직접 발표하거나 인구 비례하여 발표하지 않는다. 무슨 차이가 있는 것일까?

헤즈볼라 공격을 받은 도시는 하이파란 도시인데 인구 27만명이다. 8명이 죽었으므로 0.003 퍼센트이다. 인용한 도시 마이애미 시는 인구 362,470명인데 비례로 계산하여도 겨우 11명이 계산되어서 나오기 때문이다. 이런 값들은 너무 작기 때문에 일반인들이 그 충격을 이해하기 힘들기 때문에 전체 인구 비례하여 그 충격을 과장하여 일반인에게 전달하는 것이다. 왜냐하면 0.0003 퍼센트가 주는 의미는 일반인이 이해하기 힘들기 때문이다.

그러면 이런 비유는 타당한 것인가? 우선 이스라엘의 생명이 미국의 생명보다 더 존엄하지 않는 한 이러한 비교는 하지 않아야 한다. 그리고 이러한 비교는 이스라엘의 인구대비 죽은 사람의 비율에 따른 충격과 미국의 인구대비의 죽어나가는 사람의 비율의 충격이 같다고 가정하기 때문에 올바른 비교가 아니다. 이런 의미로 비교는 로켓 공격을 받은 도시와 마이애미 시를 비교하여야 하는데 겨우 11명이 나오므로 인구대비로 발표한 것이다. 반면 르완다의 경우는 비례를 할 필요가 없다. 이러한 비율은 비례를 하지 않더라도 일반인들은 그 충격을 쉽게 이해한다. 이런 경우는 사망자의 명수보다 훨씬 효과적이다. 주위 사람 4명 중 3명이 죽는다고 가정하여 보라.

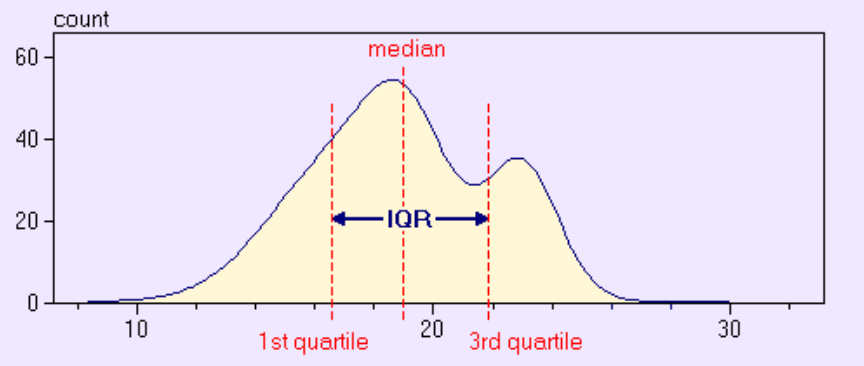
숫자를 발표하는 과정에서 어느 숫자를 이용하여 정보를 전달하는지에 따라 발표자의 주관에 따라 정보의 왜곡이 있을 수 있는 예를 살펴보았다.

5.2 사분위수와 백분위수

중앙값은 자료를 이등분하는 개념이라면 사분위수(quartile)는 자료를 4등분하는 개념이다. 그리고 백분위수(percentile)는 자료를 100등분하는 개념이다. 이러한 개념으로 보면 중앙값은 제 2사분위수(second quartile)이며 제 50백분위수가 된다. 예제 5.1을 다시 보도록 하자. 변수 봉급에 대한 평균, 그리고 중앙값은 이미 본 바 있다. 역시 엑셀을 이용해 같은 변수에 대해 제 1사분위수, 제 3사분위수를 구한다면

=quartile(salary,1),
=quartile(salary,3)

과 같이 구한다. 봉급이라는 변수의 제 1사분위수는 4,467.5만원이 나오는데 이는 봉급 값이 이 값 이하로 25%가 있다는 의미이다. 제 3사분위수는 5,967.5만원이다.



사분위수는 4등분된 피자 파이와 같다.

이러한 사분위수를 이용하면 또 다른 평균의 개념이 나오는데 다음과 같이 정의된 삼평균(trimean)이 그 것이다.

삼평균(trimean) = (제 1사분위수 + 2 × 제 2사분위수 + 제 3사분위수) / 4

이를 이용하면 중앙값과 마찬가지로 극단적인 이상값에 덜 민감한 대표값을 얻어 낼 수 있다. 위의 봉급 변수에 대해 계산을 하게 되면

삼평균 = (4,467.5 + 2 × 5,080 + 5,967.5)/4

인 5,148.75만원이 나온다.

그리고 백분위수도 엑셀에서 쉽게 구해진다. 예를 들어 제 5백분위수와 제 95백분위수는 각각

=percentile(salary,0.05)
=percentile(salary,0.95)

와 같이 구한다. 제 5백분위수는 3,715만원이다. 봉급변수 관측값의 5%만이 이보다 작은 (혹은 95%가 이보다 큰) 관측값이다. 제 95백분위수는 7,233만원이 나온다.

백분위수 자체도 의미가 있지만 백분위수간의 함수로 만들어지는 값도 매우 유용하게 쓰인다. 그 첫 번째가 제 3사분위수와 제 4사분위수의 차이(difference)인 **사분위수(간)범위**

(inter-quartile range: IQR)

이다. 크기순서대로 배열한 자료의 가운데 50%에서 제일 작은 값과 제일 큰 값의 차이를 의미한다. 봉급 변수의 IQR은 1,500만원(=5967.5 - 4467.5)으로 나왔다.

또한 상위 봉급소득과 하위 봉급소득 집단간의 차이를 비율로 표시한

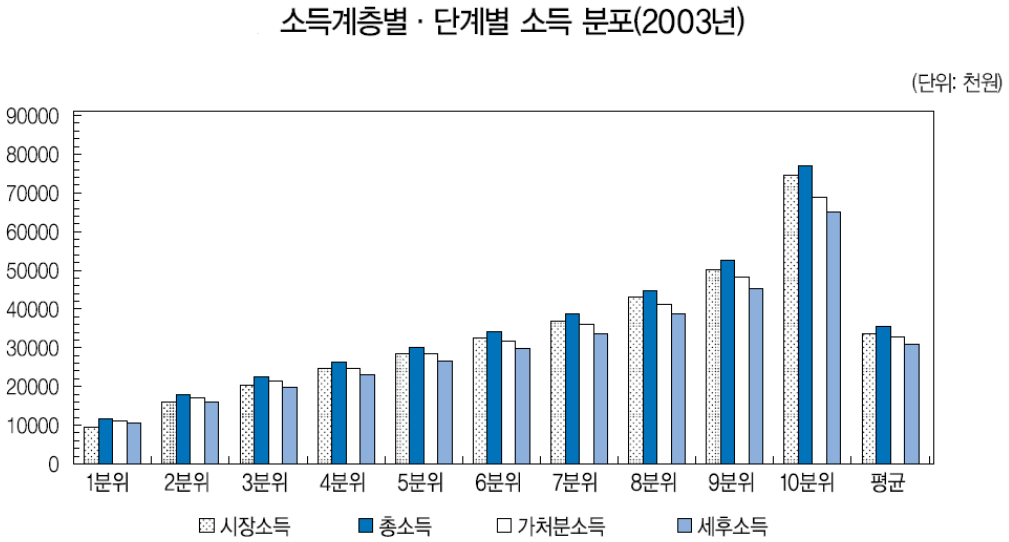
백분위수 비율 = X번째 백분위수/ Y번째 백분위수

이다. 소득 변수에 이 값을 구하면 (X=95, Y=5)에 대해 7,233/3,715 = 1.947이 나온다. 물론 이 값이 크면 클수록 봉급소득간의 차이가 주민 사이에 높다는 의미이다. 물론 X, Y를 어떻게 정하느냐는 것은 관리자의 몫이 된다. 국민소득의 불균형 정도를 나타내는 수치로 소득 5분위 배수가 있는 데 하위계층 20%에 대한 상위계층 20%의 소득배수를 말한다. 즉,

소득 5분위배수 = 80번째 백분위수/ 20번째 백분위수

로 정의한다. 이 값이 상승하면 국민소득의 불균형 정도가 심해짐을 알 수 있다.

예제 5.5 백분위수의 일종인 십분위수의 예를 들어보자. 소득계층을 10등분하여 1분위에서 10분위까지 구분한 후 소득계층별로 네 가지 소득(시장소득, 총소득, 가처분소득, 세후소득)의 분포를 다음 그림과 같이 그려보면 1-9분위까지는 완만하게 소득이 증가하다가 10분위에서 급증하는 패턴을 보이고 있다. ■



(출처: 재정포럼2006년 9월호)

그리고 백분위수 및 사분위수를 이용한 평균의 개념도 존재한다.

5.3 최소값, 최대값, 범위

최소값(minimum)은 그 밑으로 값이 존재하지 않는 값을 의미하고 최대값(maximum)은 이보다 큰 값이 존재하지 않는다는 의미를 가지고 있다. 예제 5.1에서 언급한 봉급변수에 대해 이를 구하여 보자. 엑셀에서는 다음과 같은

=min(salary)
=max(salary)

명령문을 입력하면 된다. 3,000 만원, 8,140 만원이 각각 봉급의 최소값과 최대값이다.

그리고 범위(range)는 「최대값-최소값」으로 정의된다. 봉급 변수의 범위는 5,040만원(=8,140-3,100)이다. 모든 봉급 값은 폭이 5,040 만원의 구간 안에 포함이 되어있다는 의미이다. 따라서 관측값의 범위는 최대값과 최소값과 더불어 사용되면 자료가 어디서부터 어디까지 얼마만큼의 폭에서 발생이 되었는지 쉽게 알아 볼 수 있다. 그러나 관측값 중 일부가 매우 큰 값이나 작은 값이 섞여 있는 경우는 이러한 값들은 매우 민감한 반응을 보이기 때문에 쓰는데 조심하여야 한다. 특히 범위는 자료의 퍼짐 정도를 대표하는 값으로 쓰기에는 부족함이 많다. 그럼에도 불구하고 이 세 가지 값들은 실제 유용하게 쓰인다.

5.4 변동 측정을 위한 도구

자료를 보다 잘 파악하기 위해서는 중심경향을 측정을 위한 값들 이외에도 변동의 개념을 이해하여야 한다. 다른 예를 들어보자.

예제 5.6 부품을 공급하는 두 하청업체를 생각하여 보자. 회사는 공급업체에게 부품인 볼트의 반지름 치수는 평균적으로 6mm를 유지하여야 하고 5.95mm에서 6.05mm사이에서 만들어져야 한다고 원칙을 통보하였다. 따라서 공급되는 볼트의 반지름은 이 구간 안에 있고 평균치가 6mm라면 품질로서의 최소자격은 만족이 되는 것이다. 그러나 한 회사는 5.95mm에서 6.05mm 구간에서 물건을 만들어 왔고 한 회사는 5.96mm에서 6.04mm 구간에서 물건을 공급하였다면 당연히 공급을 받는 입장에서 보게 되면 두 번째 회사의 볼트를 선호할 것이다. 이와 같이 위에서 언급한 범위(range)라는 값을 가지고 비교할 수 있겠지만 이 보다 좀 더 나은 변동측정을 위해 만든 값이 분산이며 표준편차이다. ■

분산 및 표준편차 : 제일 보편적으로 많이 쓰는 변동의 측정도구가 분산(variance) 및 표준편차(standard deviation)이다. 분산은 관측값들이 평균으로부터 얼마나 벗어났는지, 즉 편차(deviation)(= 개개의 관측값- 평균)를 구하고 이 편차를 제곱한 다음 평균을 내는 개념이다. 분산은 두 가지 버전이 존재한다. 모집단의 자료인 경우와 표본자료인 경우로 나누어진다. 이

를 식으로 쓰면 식 (5.4)와 식 (5.5)와 같다.

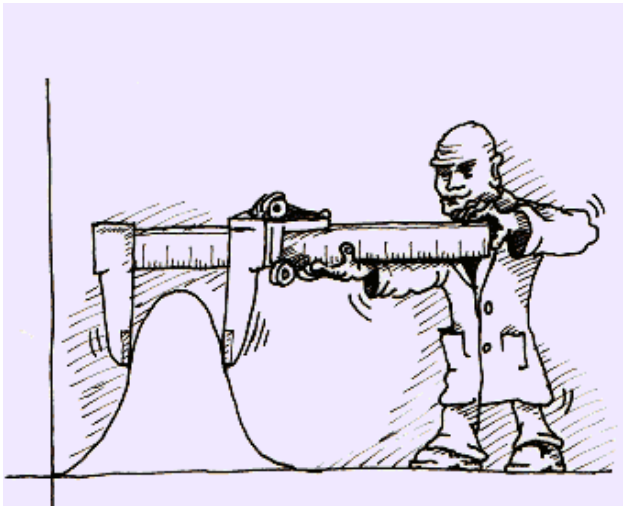
$$\sigma^2 = \frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n} \tag{5.4}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{5.5}$$

이 식들은 분모에서 n 이냐 $n-1$ 의 차이이지 기본적으로 같은 공식이다. 그리고 n 이 충분히 커지면 두 공식은 같아진다. 만약 엑셀을 위해 분산을 구한다면 모집단 분산인 경우는 varp 함수로 표본분산과 차별시키지만 대부분 우리가 접하는 자료는 표본인 관계로 특별한 경우가 아니면 var 함수만 이용한다. 예제 5.1에서 봉급 변수에 대한 분산은

=var(salary)

으로 구해진다. 1320810.23이다. 그런데 이렇게 구해진 분산의 단위는 식 (5.4)나 (5.5)을 보게 되면 원래의 봉급의 단위인 만원이 아니고 만원의 제곱임을 알 수 있다. 따라서 제곱단위로 자료의 변동을 논하는 것은 바람직하지 않다. 따라서 분산의 제곱근의 형태인 표준편차를 쓰는 것이 일반화되어 있다.



표준편차는 변동 측정도구이다.

표준편차 = $\sqrt{\text{분산}}$

엑셀에서는 다음과 같은 명령문으로 표준편차를

=stdev(salary)

구할 수 있다. 물론 표준편차는 =var(salary)의 제곱근 명령문인 =sqrt(var(salary))를 통해 구할 수도 있다. 봉급변수에 대한 표준편차는 1,149.265만원이 나온다.

설명 1: 표준편차를 구하는 과정을 보면 편차(deviation)에 대해 제곱을 한 다음 평균을 구하고 제곱근을 하는 형태를 취하였다. 왜냐하면 편차에 대한 평균을 구하면 0이 되기 때문이다.

설명 2: 혹은 편차에 절대값을 취한 다음 평균을 내는 공식을 Mean Absolute Deviation(MAD)을 사용하기도 하나 이는 절대값 함수이므로 여러 가지 이유로 불편한 것이 많다. 오히려 제곱을 한 다음 제곱근을 취하는 것이 좋다.

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

설명 3 : 평균을 중심으로 개개의 관측값이 얼마나 벗어나는지 편차를 구했으나 평균이 아닌 다른 값을 집어넣어도 된다. 특히 평균이 아주 큰 값이나 작은 값에 영향을 받는 경우라면 그렇다. 예를 들어 중앙값을 대입하여 편차를 구하여도 무방하나 기본적으로 평균을 집어넣는 이유는 구한 편차의 제곱의 합이 다른 어떤 경우보다도 제일 적게 되기 때문이다. 절대값 기준이라면 중앙값을 집어넣어야만 편차의 합이 제일 적게 나온다. 따라서 위의 MAD도 $\overline{x}$ 대신에 중앙값을 넣어도 좋다. 수학적으로 이를 표현하면 다음과 같다. 증명은 생략한다. 그러나 간단한 미적분의 지식만 가지고 있으면 해결되는 문제이다.

$\sum_{i=1}^{n}(x_i - a)^2$ 을 최소화하는 a 는 $\overline{x}$ 가 나오며, $\sum_{i=1}^{n}|x_i - a|$ 을 최소화하는 a 는 중앙값이다.

이렇듯 분산과 표준편차는 자료의 변동을 알려주는 지표가 된다. 분산 혹은 표준편차가 작으면 작을수록 자료는 평균을 중심으로 몰려 있다는 사실이다.

여기서 다음과 같은 경험법칙(rule of thumb)을 소개할 필요가 있다. 이는 매우 응용성이 높은 규칙이다.

경험법칙: 자료가 산이나 종의 모양을 하고 있을 때, 즉 좌우대칭의 모양을 가지고 있는 히스토그램을 통해 자료를 표현할 수 있을 때 표준편차의 해석은 다음과 같은 경험법칙을 통해 이해의 폭을 더 높일 수 있다.

1. 평균을 중심으로 1개의 표준편차구간, $\overline{x} \pm s$ 혹은 $[\overline{x}-s, \overline{x}+s]$ 에는 자료의 약 68%가 포함되어 있다.
2. 평균을 중심으로 2개의 표준편차구간, $\overline{x} \pm 2s$ 에는 자료의 약 95%가 포함되어 있다.
3. 평균을 중심으로 3개의 표준편차구간, $\overline{x} \pm 3s$ 에는 자료의 약 99%가 포함되어 있다.

봉급변수는 평균이 5,266만원, 표준편차가 1,149만원이 산출되었다. 예를 들어 이러한 경험법칙에 대한 개념을 설명하여 보자. [표 5.2]를 참조하기 바란다. <경험법칙.xls>

	A	B	C	D	E	K	L	M
1	Age	Gender	State	Children	Salary	contains	contains	contains
2	61	Female	강동	2	62,000	1	1	1
3	37	Male	강동	2	52,000	1	1	1
4	32	Female	강동	3	81,400	0	0	1
5	65	Female	강동	2	49,600	1	1	1
6	40	Male	강동	3	47,700	1	1	1
7	32	Female	강동	1	59,900	1	1	1
8	38	Female	강동	2	39,000	0	1	1
9	48	Male	강동	1	61,500	1	1	1
10	40	Male	강동	1	44,500	1	1	1
11	44	Male	강동	2	45,200	1	1	1
12	57	Female	강동	2	36,700	0	1	1
13	21	Female	강동	2	54,300	1	1	1
14	49	Male	강동	1	62,100	1	1	1
15	34	Male	강동	0	78,000	0	0	1
16	38	Male	강동	1	43,300	1	1	1
17	35	Male	송파	1	65,400	0	1	1
18	35	Male	송파	0	63,200	1	1	1
19	33	Female	송파	3	46,300	1	1	1
20	45	Male	송파	1	45,900	1	1	1
21	57	Male	송파	1	48,100	1	1	1
22	38	Female	송파	0	58,100	1	1	1
23	37	Female	송파	2	56,000	1	1	1
24	42	Female	송파	2	53,400	1	1	1
25	49	Male	송파	0	43,200	1	1	1
26	52	Male	송파	1	44,100	1	1	1
27	27	Male	송파	3	45,400	1	1	1
28	40	Male	송파	0	37,700	0	1	1
29	63	Male	송파	2	53,900	1	1	1
30	48	Female	송파	2	31,000	0	1	1
31	40	Male	송파	0	59,000	1	1	1
32								
33				average	52263.33333	0.766667	0.933333	1
34				stdev	11492.65083			
35								
36			1 stdev	lower limit	40770.68251			
37				upper limit	63755.98416			
38			2stdev	lower limit	29278.03168			
39				upper limit	75248.63498			
40			3stdev	lower limit	17785.38086			
41				upper limit	86741.28581			

[표 5.2] 보편적인 법칙 설명

여기에서는 실제 봉급 변수를 이용하여 각각의 구간에 실제 자료가 몇 %나 포함되어 있는지 알아 보았다. 열 K, L, M이 각각 하나 표준편차 구간, 둘 표준편차 구간, 세 표준편차 구간에 해당하는 값이 포함되어 있으면 1이고 그렇지 않으면 0으로 표시하였다. 그리고 평균을 낸 결과 각 구간에 포함되는 자료의 비율은 76.7%, 93.3%, 100%로서 보편적인 법칙에서 제시하는

값과 약간의 차이가 나타났다. 봉급 변수에 대한 히스토그램은 좌우 대칭이 아니고 오른쪽으로 왜도가 있었던 것을 기억하기 바란다. 그러나 일반적으로 자료가 평균을 중심으로 좌우대칭인 경우, 이 경험법칙은 잘 작동하는 편이다.

변동계수: 자료의 산포도를 측정하는 다른 도구로 다음과 같이 정의되는 변동계수가 있다.

변동계수 = 표준편차 / 산술평균

먼저 변동계수의 분자와 분모의 단위는 원래 수집된 단위와 같기 때문에 서로 상쇄되어 단위가 없어진다. 예제 5.1에서 봉급변수를 이용하여 변동계수를 구하면

변동계수 = 1,149.3/5,266.3 =0.218

이 나오는데 이는 표준편차의 크기를 평균으로 고려하여 답을 하는 격이 된다. 이 예제에서는 그렇게 큰 의미가 없다. 설령 두 지역(송파, 강동)을 나누어 비교한다 하더라도 두 지역의 평균 차이는 그렇게 크게 나타나지 않기 때문이다. 그러나 두 지역의 소득의 차이가 많이 나타나는 경우는 변동계수의 의미가 나타난다. 왜냐하면 소득이 커질수록 가구원의 소득의 편차는 상대적으로 커지는 경향이 있기 때문에 이러한 효과를 없애준다.

아직 이해가 안 되는가? 신생아와 성인의 몸무게의 표준편차를 직접 비교하는 것 보다는 평균을 감안한 편차의 크기를 논하는 것이 좋을 것이다. 왜냐하면 성인의 몸무게가 늘어날수록 표준편차의 크기는 덩달아 늘어나기 때문이다. 두 표준편차의 크기를 직접적으로 비교하는 것은 아무런 의미가 없다.

5.5 왜도와 첨도

중심경향을 논할 때 중앙값과 평균의 위치를 파악하는 방법으로 왜도가 있다.

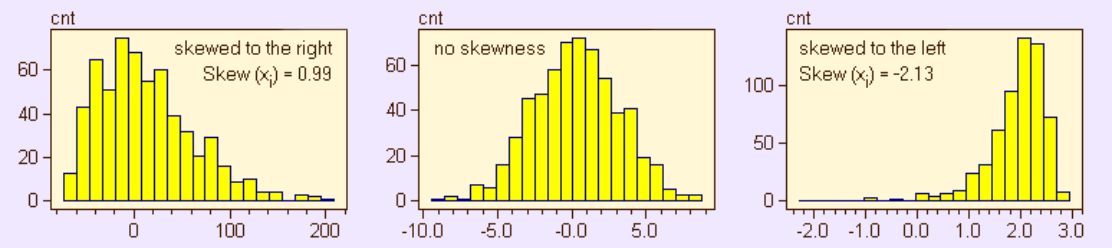
왜도: 왜도(skewness)는 기본적으로 편차의 3제곱의 개념으로 설명이 가능하다. 식 (5.6)은 왜도에 대한 정의인데 편차를 표준편차로 나눈 후 세 제곱을 한 다음 평균을 내는 개념이다.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3 \qquad (5.6)$$

따라서 분자와 분모의 단위가 같아지므로 왜도의 단위는 없다. 그리고 왜도의 부호는 편차의 세 제곱에 따른 부호에 따른다. 자료가 왼쪽으로 왜도가 발생한 경우는 부호가 음이 될 것이고 반대로 오른쪽으로 왜도가 발생되면 양의 부호가 된다. 물론 좌우 대칭인 경우는 왜도는 0이 된다. [표 5.1]에서 봉급변수에 대한 왜도를 엑셀 명령문을 사용하여 구한다면

=skew(salary)

으로 구할 수 있다. 0.64312로 나왔다. 봉급변수는 오른쪽으로 왜도가 발생한 자료임을 기억하기 바란다.

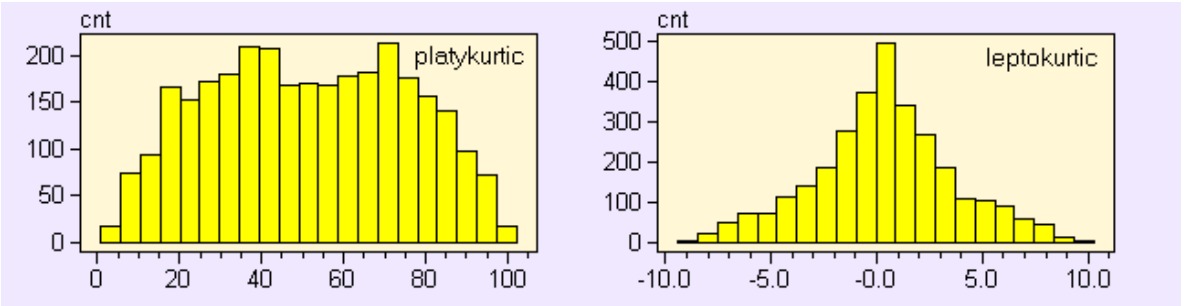


왜도의 크기

왜도의 값은 그 자체보다는 두 자료의 왜도를 상대적으로 비교하는데 많이 쓰인다. 기준 점은 물론 대칭인 자료인 경우의 왜도는 0이다.

첨도: 첨도(kurtosis)는 편차의 4제곱의 개념으로 설명이 가능하다. 식 (8.7)은 왜도에 대한 정의인데 편차를 표준편차로 나눈 후 4제곱을 한 다음 3을 빼는 개념이다. 여기서 3을 빼는 이유는 기준이 되는 정규분포인 경우 값이 3이기 때문이다.

$$\left\{\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4\right\}-3 \qquad (5.7)$$



첨도의 크기

자료의 중심에 있는 부분이 정규분포에 비해 얼마만큼 뾰족하냐(peakedness)를 측정하는 방법이다. 역시 분자와 분모의 단위와 같기 때문에 첨도의 단위는 없다. 판정은 다음과 같이 하면 된다.

첨도가 0보다 작으면 중심부분이 짧고 뚱뚱하다(short and fat).

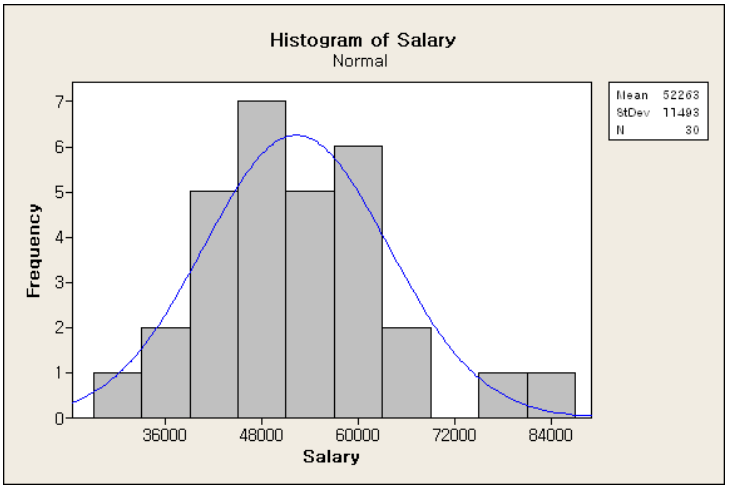
첨도가 0이면 정규분포의 정상부근의 모양새이다(normal).
첨도가 0이상이면 중심부분이 길고 마른(tall and skinny) 형태의 자료이다.

[표 5.1]의 봉급변수에 대한 첨도를 엑셀에서 구한다면 다음과 같은 명령문

=kurt(salary)

을 통해 구하면 된다. 값이 0.5646이 나왔다.

이 경우에는 0.5646이 나왔으므로 3번째 경우에 해당된다고 볼 수 있다. 첨도는 봉우리가 하나인 분포의 뾰족한 정도를 정규분포와 비교 목적으로 만든 값이다. 아직 정규분포에 대해서는 언급이 되지 않았다. 그러나 당분간은 산의 모양, 혹은 종의 모양을 연상하기 바란다. 후에 자세한 언급이 이어진다. 아래 [그림 5.4]는 히스토그램에 봉급변수의 평균 52,263, 표준편차 11,493을 가지고 있는 정규분포를 덧붙여 그린 그림이다. 그림에서 정규분포의 가운데 부분보다 조금 뾰족하게 올라 있는 막대가 있다. 물론 자료의 개수가 적기 때문에 판단하기는 어렵다.



[그림 5.4] 첨도

5.6 상자그림

봉급변수에 대해 지금까지 나온 모든 값들을 하나로 모아 보자. 그러나 일일이 개별적인 명령문을 시행하지 않고 일괄로 구하는 메뉴가 엑셀과 같은 소프트웨어에는 있다.

도구> 데이터분석> 기술통계법을 눌러 결과를 본다. [표 5.3]이 그 결과이다. 이 중 표준오차는 후에 설명할 내용이다. 그리고 최빈값은 연속형 자료이기 때문에 구하지 못하고 #N/A 표시가 되었음을 주지하기 바란다.

	A	B
1	Salary	
2		
3	평균	52263.33
4	표준 오차	2098.261
5	중앙값	50800
6	최빈값	#N/A
7	표준 편차	11492.65
8	분산	1.32E+08
9	첨도	0.5646
10	왜도	0.64313
11	범위	50400
12	최소값	31000
13	최대값	81400
14	합	1567900
15	관측수	30

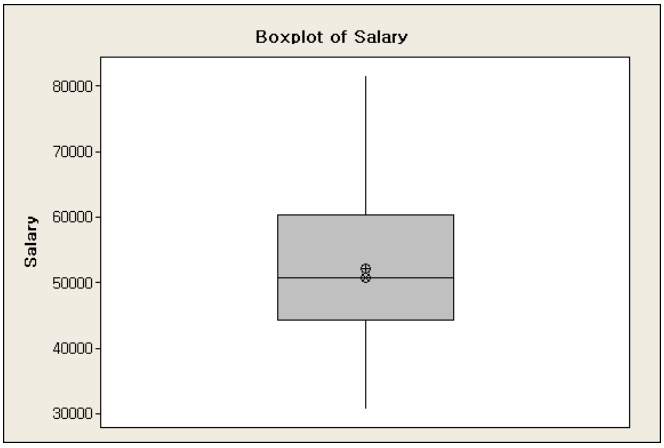
[표 5.3] 봉급변수에 대한 기술통계량들

봉급변수에 대한 수치적인 요약방법을 통해 어느 정도 자료가 어떤 모양새를 하고 있는지 굳이 히스토그램을 그리지 않더라도 알아 볼 수 있을 것이다. 이와 같은 수치적 요약 정보라 하더라도 그림으로 일목요연하게 표시하는 방법을 많이 쓰는데 상자-수염 그림(box-whisker plot), 상자그림(box plot)이라 줄여서 부르기도 한다. 미국의 통계학자 J. Tukey가 제안한 방법이다.

이 그림에 있는 상자를 그리기 위해서는 다음과 같은 (1)-(4)까지의 절차를 따르면 된다.

(1) 먼저 제 1사분위수와 제 3사분위수를 파악한다. 상자의 오른쪽 면은 제 3사분위수, 그리고 왼쪽은 제 1사분위수가 된다. 그러면 IQR의 값이 상자의 길이가 된다.
(2) 중앙값을 상자 안의 있는 수선(vertical line)으로 표시한다. 평균은 상자내의 점으로 표시한다.
(3) 상자바깥의 가운데 지점에서 출발하여 수평선을 그리되 IQR의 1.5배 내의 범위에서 자료가 있는 위치까지 뻗어서 그린다. 이는 범위 안에 최소값과 최대값이 있다면 최소값과 최대값까지 뻗어서 그린다는 의미이다.
(4) 그렇지 않고 일부 관측값이 1.5 × IQR 의 범위를 벗어나고 1.5 × IQR- 3.0 × IQR의 범위안에 있으면 보통이상점(mild outlier)라 하고 3.0 × IQR을 벗어나면 극단이상점(extreme outlier)이라 하고 표시가 다른 점(들)으로 달리 구분한다.

엑셀에서는 chart 기능을 약간 손질하면 원시적이기는 하지만 상자그림을 그릴 수는 있다. 그러나 불편하므로 전문적인 통계프로그램을 이용하는 것이 편하다. 다음 [그림 5.5]는 미니탭이란 통계소프트웨어를 이용하여 봉급 변수에 대한 상자그림을 그려보았다.

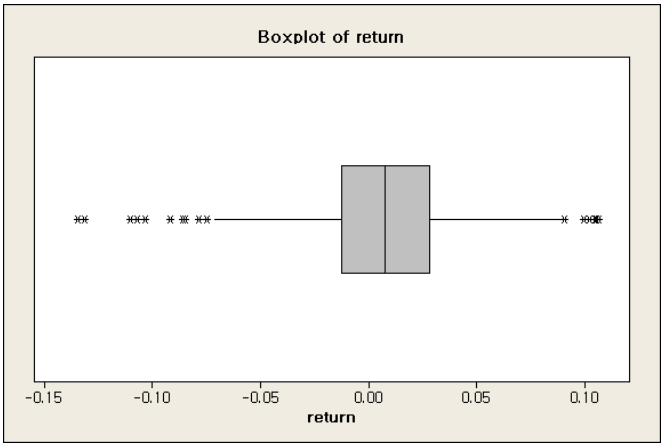


[그림 5.5] 봉급변수에 대한 상자그림

[그림 5.5]의 형태를 보도록 하자. 상자 가운데는 중앙값이다. 중앙값이 상자의 좌측, 즉 제 1사분위수에 가깝게 위치하고 있어 가운데 자료 50%를 기준으로 하였을 때 약간 오른쪽으로 왜도가 있다. 만약 중앙값이 제 1사분위수와 제 3사분위수 가운데 있었다면 좌우대칭일 것이다. 수염의 길이도 같지 않다. 이는 오른쪽으로 꼬리 부분이 길게 늘어져 있다는 점을 반영한다. 평균의 위치를 적어 놓지 않아도 되지만 [그림 5.5]에서 확인하다시피 평균은 중앙값보다 위에 있다.

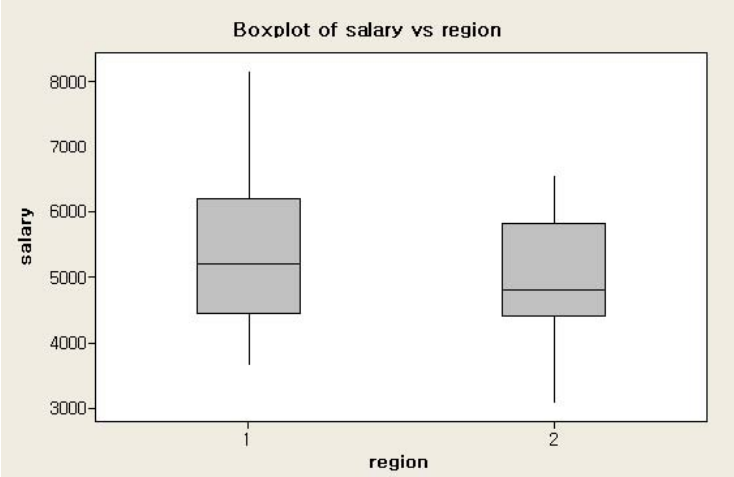
예제 5.7 다음 [그림 5.6]은 1947년 1월부터 1993년 1월까지 미국 다우존스의 월 수익률에 대해 상자그림을 그린 것이다. [그림 5.5]와 달리 수익률은 1.5 × IQR을 벗어나지만 3 × IQR는 넘어 가지 않는 보통이상점이 총 16개가 존재한다. 왼쪽(아래쪽)으로 10개, 오른쪽(위쪽)으로 6개가 존재함을 알 수 있다. 상자그림의 형태로 보아 좌우대칭의 성격이 강하게 나타난다. ■

<상자그림.xls>



[그림 5.6] 극단 이상값이 있는 상자그림

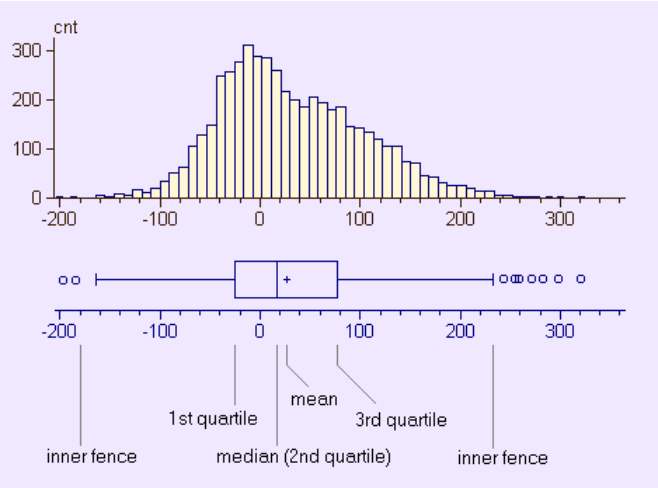
그러나 상자그림의 유용성은 다른데 있다. 알고 있듯이 [표 5.1]의 자료는 강동지역과 송파지역에 살고 있는 주민들을 대상으로 실시한 것이다. 만약 지역별로 봉급의 분포가 어떻게 달라지는지 일목요연하게 보기 위해서 [그림 5.7]과 같이 두 상자그림을 한 그림에 그리는 방법도 제시될 수 있다.



[그림 5.7] 두 지역 비교 상자그림

여기서 1은 강동지역, 2는 송파지역을 의미한다. 송파지역에 비해 상대적으로 강동지역의 봉급의 분포가 위쪽으로 분포가 되어 있음을 알 수 있다.

- 상자그림도 진화한다. 아래 [그림 5.8]은 히스토그램과 상자그림을 덧붙여 그린 그림인데 너무 호사스럽지 않은가?



[그림 5.8] 히스토그램과 상자그림의 만남

5.7 통계의 오용과 방지책

통계학이 우리 사회에서 중요한 역할을 하기 시작하면서 통계의 오용(misuse of statistics)도 늘고 있고 사회적 문제를 일으키기도 한다. 두 가지 예를 들어보자. 첫 번째 예로서 다음 [그림 5.9]는 2005.07.06 MBC 뉴스데스크 시간에 노무현 대통령의 국정운영과 정당지지도에 대한 여론조사(20세 이상 성인 남녀 938명, 코리아리서치센터)의 결과를 막대그래프로 발표한 화면이다. 여론조사 결과 한나라당 38.1%, 열린우리당 21.9%, 민주노동당 18.9%, 민주당 5.4% 순이었다. 그런데 화면을 보면 정당지지도가 큰 순으로 배열되어 있지도 않지만 더 큰 문제는 열린우리당의 정당지지도는 과대평가되었고 민노당은 과소평가되었다. 화면에서 점선이 정당지지도에 맞는 막대그래프의 정상위치이다. 의도적이든 실수이든 대중적인 방송매체의 통계의 오용을 나타내는 전형적인 예라 할 수 있다.

[그림 5.9] 정당지지도를 나타내는 막대그래프

두 번째 예로서 다음 [그림 5.10]은 2005.08.31 정부(행정자치부)의 잘못된 부동산통계를 질타한 신문 지면을 나타낸다. 부동산 관련 통계를 발표하면서 정부 입맛에 맞춰 자료를 부풀리거나 조사대상이나 기준을 자의적으로 설정하여 실상을 왜곡한 예이다. 토지소유율을 계산할 때는 통상 인구수가 아닌 가구수를 이용하는데 인구수를 이용하여 토지소유율을 계산함으로써 의도적으로 토지소유 집중이 심화되었다는 그릇된 정보를 통하여 국민감정을 자극하고 있다.

정부, 땅부자 통계왜곡 왜?

정부가 부동산 관련 통계를 발표하면서 정부 입맛에 맞춰 부풀리거나 조사대상·기준을 자의적으로 설정, 실상을 왜곡하고 있다는 지적을 받고 있다. 지난 15일 행정자치부는 "상위 1%가 전체 사유지의 51.5%, 상위 5%가 82.7%의 토지를 보유하고 있다"며 "땅을 1평이라도 소유하고 있는 사람은 전체의 28.7%(1397만명)"라고 발표했다.

이 자료에서 행자부는 대부분 토지가 가구주 명의로 되어 있음에도 불구하고 토지 소유자를 전체 가구수 대신 전체 인구로 나눠 계산했다. 인구수로 나누면 젖먹이를 포함한 1318만명의 미성년자가 모두 통계에 포함되기 때문에 토지를 갖고 있지 않은 사람들의 비율이 급격히 늘어난다. 〈표 참조〉

만약 전체 인구 대신 전체 가구수(6월 말 기준·1765만5000가구)를 이용해 토지소유율을 계산하면 현재 1평 이상 땅을 보유하고 있는 사람은 28.7%가 아니라 79.1%로 올라간다. 또 82.7%의 사유지는 상위 5%가 아니라 이보다 3배 많은 14%가 소유하고 있는 것으로 집계된다. 현실의 토지 집중 정도가 적어도 정부 발표보다는 덜하다는 얘기다.

과거 딱 한 차례 발표됐던 지난 1989년 정부의 토지소유율 보고서(토지공개념연구위원회 작성)에서도 인구수 아닌 가구수 통계를 이용해 토지소유율을 계산했었다.

이에 대해 행자부 관계자는 "인구수를 이용해 통계를 작성하는 데만 보름 이상 걸렸다"며 "물리적인 시간 제약 때문에 다양한 통계를 뽑을 수 없었다"고 해명했다. 하지만 6월 말 기준 전국 가구수 통계는 정부가 토지소유 통계를 내기 열흘 전인 지난 7월 4일자로 행자부 홈페이지에 떠있는 것으로 확인됐다. 이미 나와 있는 통계를 가져다 쓰기만 하면 됐다는 얘기다.

이에 앞서 국세청은 지난 1일 "2000년 이후 서울 강남권 아파트 취득자의 58.8%가 이미 집 2채 이상을 갖고 있었던 다주택 보유자"라며 강남 아파트값 상승의 원인을 투기적 수요로 돌렸다. 하지만 조사 대상이 된 9개 아파트 단지가 재건축 아파트나 대치동 선경아파트 등 평소 투기수요가 몰리는 곳들이어서 표본 선정이 편중돼 있다는 지적을 받고 있다. 부동산 업계 관계자는 "재건축 아파트는 큰 평수를 받을 수 있다는 기대감 때문에 원래 실수요보다 투자 목적 취득이 많은 곳"이라면서 "강남 아닌 다른 지역도 재건축 아파트 취득자 대부분은 다주택 보유자"라고 말했다.

한양대 나성린 교수는 "정부가 최근 발표하는 부동산 통계에는 부동산을 옥죄기 위해 국민 감정을 자극하려는 의도가 엿보인다"며 "정부가 통계를 쥐고 구미에 맞고 필요한 것만 제한적으로 발표하지 말고 객관적 검증이 가능하도록 통계의 전모를 공개해야 한다"고 말했다.

박종세 기자 (블로그)jspark.chosun.com
이진석 기자 (블로그)island.chosun.com

정부가 공개한 토지 보유 현황 통계의 문제점

쟁점	정부의 주장	일반적인 해석
땅 1평 이상 소유자	전체의 28.7%	79.1%
토지 51.5% 소유	상위 1%	상위 2.8%
토지 82.7% 소유	상위 5%	상위 14%
16년 전과 비교	상위 5% 토지소유 집중 심화 (65.2% ? 82.7%)	비슷 (65.2% ? 65% 안팎)

*행정자치부는 토지 소유자를 전체 인구로 나눠 계산한 반면, 전문가들은 토지를 대부분 가구주가 소유하고 있기 때문에 가구수로 나눠야 토지 집중이 과장되게 나타나는 오류를 막을 수 있다고 밝히고 있음.
표의 '일반적인 해석' 은 가구를 기준으로한 통계. 〈자료: 행정자치부, 국토연구원〉

(제공처 : 조선일보)

[그림 5.10] 정부의 잘못된 부동산통계를 질타한 신문 지면

통계의 오용은 의도적인 경우 우리 사회에 악영향을 끼치게 된다. 이러한 통계의 오용이 일어나면 통계적 오류(statistical fallacy)가 만들어진다. 통계의 오용의 형태에 대하여 Jaffe와 Spirer(1987)는 다음과 같이 5개의 범주로 나누었다.

1. 데이터와 관계된 전문지식의 부족
2. 낮은 데이터품질: 데이터의 결함(flaw), 편의(편향, bias), 잘못된 정의
3. 연구와 보고서의 준비 부족: 타당성이 결여된 실험계획, 그래프의 오용, 발견의 틀린 해석, 결과의 잘못된 표현
4. 부적절한 통계방법의 사용
5. 교묘한 데이터 감추기: 의도적인 자료의 삭제/첨가 및 조작

그래프의 오용에 대해서는 4장에서 언급하였고 허위상관(상관관계와 인과관계에 대한 혼동)에 대해서는 6장에서 언급되므로 5장에서는 '의도적인 자료의 삭제/첨가 및 조작'에 대해서만 간략히 살펴보자.

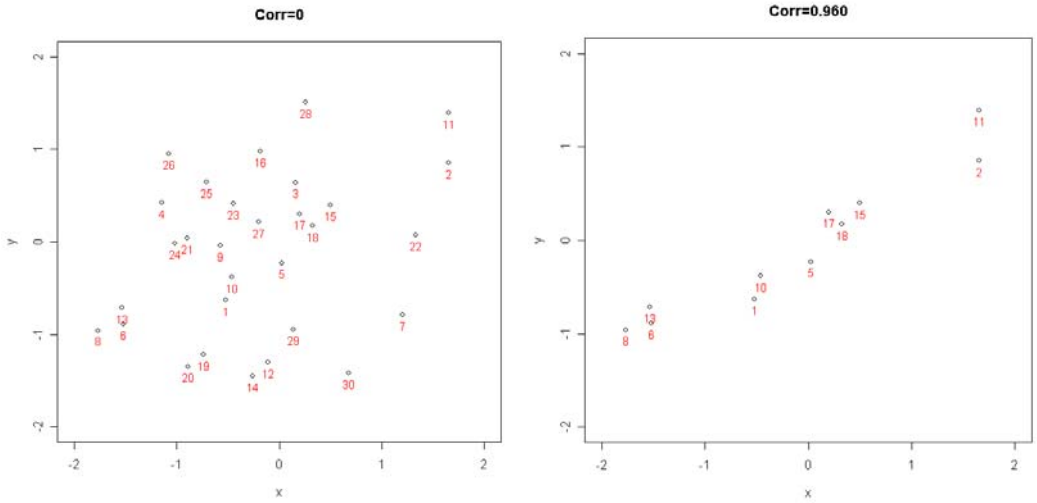
의도적인 자료의 삭제/첨가와 조작

우리가 실험이나 관찰 연구에서 원하지 않는 데이터를 버리거나 첨가하거나 조작하는 통계적 오용을 의도적으로 범할 수 있다. 예로서 신약의 약효를 입증할 만한 실험 결과만을 의도적으로 뽑아 이 표본들만 이용하여 통계분석을 행하고 효과가 있다는 주장을 할 수 있다.

예제 5.8 안전한 작업환경은 생산성 향상에 큰 기여를 한다고 한다. 육체노동자들에게 자기 능력을 벗어난 무거운 물체를 들어 올리게 하는 행위 같은 일을 시켜서는 안 된다. 18세에서 30세까지의 육체노동자들 중 랜덤하게 5명을 선정하여 1분에 4개의 무거운 물체를 들게 하여 그 중 가장 무겁게 들어 올린 무게(MAWL(maximum weight lifted))를 kg으로 조사하니 25.8, 36.6, 26.3, 21.8, 27.2가 나왔다. 통계분석을 해 보면 가장 무겁게 들어 올린 무게가 25kg보다 크다고 할 수 없다는 결론이 나온다. 자, 그러면 데이터를 조작하여 보자! 실험을 계속하여 실시하다 보면 MAWL가 크게 나오는 데이터가 발생한다. 이 데이터가 30.5kg이라 하자. 이 데이터를 앞의 데이터 중 제일 작은 값 21.8 대신 바꿔치기하여 25.8, 36.6, 26.3, 30.5, 27.2로 조작한 후 통계분석을 해 보면 가장 무겁게 들어 올린 무게가 25kg보다 크다고 할 수 있게 된다(왜 이런 결론이 나오는 지는 11장에서 배우게 될 것이다.). 랜덤화를 무시하고 데이터를 조작하면 통계적 의사결정을 자기 마음대로 주무르게 된다. ■

몇 개의 데이터를 조작하는 것이 아니라 원하지 않는 표본 전체를 통채로 버리고 원하는 표본이 될 때까지 실험이나 관측을 행하여 표본을 얻은 후 이것을 이용하면 통계의 오용을 의도적으로 일으킬 수도 있다.

예제 5.9 다음 [그림 5.11]의 오른쪽 그림은 30개 자료 쌍 (x,y)의 산점도를 나타낸다. 점 밑에 있는 숫자는 자료 쌍의 순서이다. 상관계수를 구하면 0이 된다(상관계수에 대한 자세한 설명은 6장에서 취급한다.). x와 y 사이에 상관관계가 있도록 조작하기 위하여 11개 데이터 (1, 2, 5, 6, 8, 10, 11, 13, 15, 17, 18)만을 의도적으로 뽑아 다음 [그림 5.11]의 왼쪽 그림처럼 왜곡된 산점도를 그린 후 상관계수를 구하면 0.960이 된다. 상관관계가 없는 자료에서 자료의 일부만을 선택하여 상관관계가 있는 것처럼 꾸밀 수가 있다. ■



[그림 5.11] 정상적인 산점도와 왜곡된 산점도

오용 방지책

우리는 자료를 보며 어떻게 숫자들이 잘못 쓰이고 있는 지를 유심히 살펴야 한다. 또한 숫자들이 정확히 무엇을 측정하였는지를 염두에 두고 자료를 보아야 한다. 그리고 다음과 같은 질문들에 대하여 답을 해본다.

1. 숫자들이 서로 일치하는가?
2. 계산이 틀리지 않았나?
3. 숫자가 완전한 사실을 반영하는가?
4. 숫자들이 믿을 만한가?

과학자들이 자신의 실험결과를 발표할 때 종종 데이터를 인위적으로 만들거나(data fabrication) 데이터를 변경한다(falsification). 임상실험 데이터의 경우 다음 [표 5.4]와 같은 패턴이 나타나면 데이터조작이 일어났을 가능성이 크다. 이러한 경우 다음 [표 5.4]와 같은 통계방법들을 이용하여 데이터 조작을 찾아낼 수 있다. 단변량인 경우는 간단한 그림(줄기-잎 그림, 히스토그램, 상자그림)이나 기술통계량들(산술평균, 중앙값, 최빈값, 분산, 표준편차, 사분위수범위 등)을 이용해서도 데이터조작의 패턴을 알아낼 수 있다.

구 분	패 턴	데이터조작 발견 통계기법
동시 단변량	특정 숫자 선호 반올림수 선호 너무 적거나 많은 특이값 너무 적거나 큰 분산 특이한 피크값 너무 한쪽으로 치우친 자료	기술통계량들 상자그림 히스토그램 줄기-잎 그림 편차검정
동시 다변량	다변량 내재값 다변량 특이값 레버리지 너무 약하거나 강한 상관계수	분할표/산점도 상관계수/회귀분석 쿡거리 마할라노비스 거리 집락분석 판별분석 체르노프얼굴 별(바늘)그림 호텔링의 T^2검정 대비검정
반복측정	보간(補間, 내삽內揷) 중첩 의도적으로 만들어진 패턴	자기상관계수 프로파일 다항대비 런검정
달력시간	랜덤화 위반 토요일, 휴일 믿기 어려운 증가 시간의 경향	잔차도 누적합(CUSUM) 관리도

[표 5.4] 데이터 조작 패턴과 데이터 조작 발견기법

자료조작을 찾아내는 방법들 중 하나로서 특이한 방법이 있는데 벤포드의 법칙(Benford's Law)을 첫 숫자(first digits)에 적용하여 보는 것이다. 첫 숫자란 각각의 숫자에서 유의한 첫 숫자, 예로 '351'이나 '0.0351'에서의 첫 숫자는 3이다. 인간이 생활을 하면서 발생시키는 숫자들의 첫 숫자가 각각 1, 2, 3, 4, 5, 6, 7, 8, 9가 될 확률은 균등분포처럼 1/9이 되지 못하고 다음 [표 5.5]처럼 제일 큰 것은 1일 때 0.301, 제일 작은 것은 9일 때 0.046이 되고 1에서 9순으로 발생 확률이 작아진다. 이를 이용하면 데이터조작을 찾아낼 수 있다.

첫 숫자 D	확률 P_D
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

[표 5.5] 벤포드의 법칙

학습요약

제 5장에서는 자료가 주어졌을 때 수치적인 요약된 값으로 자료의 특징을 알 수 있는 방법을 소개하였다. 중심경향적인 값은 평균 이외에도 어떤 종류가 있는지 그리고 평균과 차이점은 무엇인지, 그리고 주의할 점에 대해 설명하였다. 또한 자료의 산포도, 즉 흐트러져 있는 정도를 알 수 있는 분산, 표준편차 등에 대해서도 언급하였다. 그리고 이를 요약하여 정리하여 주는 상자그림을 소개하였다. 다양한 수치로 표시되는 값만 가지고도 얼마든지 자료의 모양새 및 특징을 볼 수 있었으나 사실 히스토그램과 같은 그림을 통해 이를 보완하여야 한다. 그러나 제 5장에서 소개한 내용은 주어진 변수가 하나인 경우에 사용이 가능한 것이고 변수가 두 개 이상으로 확장되는 경우는 다른 방법의 요약된 정보가 필요할 것이다.

5장 연습문제

5.1 [표 5.1]의 변수 중 주로 봉급변수에 대해서만 분석을 실시하였는데 나머지 변수에 대해서도 분석하라. 개개의 변수에 대해서만 분석도 중요하지만 지역간 의견의 차이, 성별 의견의 차이 등도 생각하여 볼 수 있다.

5.2 아래 자료는 도시 근로자 중 화이트 컬러 직장에 다니는 미혼 남성에게 물어본 조사 자료 중 일부를 무작위로 발췌하여 기록한 것이다. 각 변수에 대해 제 5장에서 나온 내용을 기초로 하여 요약하여 보라. <소비형태.xls>

	A	B	C	D
1	**100가구 소득 소비형태**			
2				
3	봉급	문화	스포츠	외식
4	54,600	1,020	990	1,510
5	57,500	1,100	460	1,180
6	53,300	900	780	1,590
7	43,500	570	860	1,750
8	57,200	900	1,390	2,120
9	63,400	820	1,880	3,090
10	58,500	1,340	710	1,540
11	55,600	1,250	680	1,800
12	61,300	1,190	1,220	2,330
13	61,100	640	1,480	2,670
14	77,200	900	820	2,850
15	58,800	710	1,080	2,200
16	62,900	1,240	1,230	2,430
17	61,900	1,270	1,000	2,110
18	76,500	1,180	690	1,820
19	50,300	810	1,490	2,100
20	45,900	840	730	920
21	61,900	1,290	1,050	2,480
22	56,700	780	970	1,930
23	43,300	910	1,120	1,720
24	63,000	560	1,570	1,990

5.3 피보나치수열은 다음과 같이 첫 번째와 두 번째 항 $a(1), a(2)$가 1이고, 세 번째 항 $a(3)$부터는 앞의 두 개의 항을 더한 결과가 그 항이 되는 수열을 말한다.

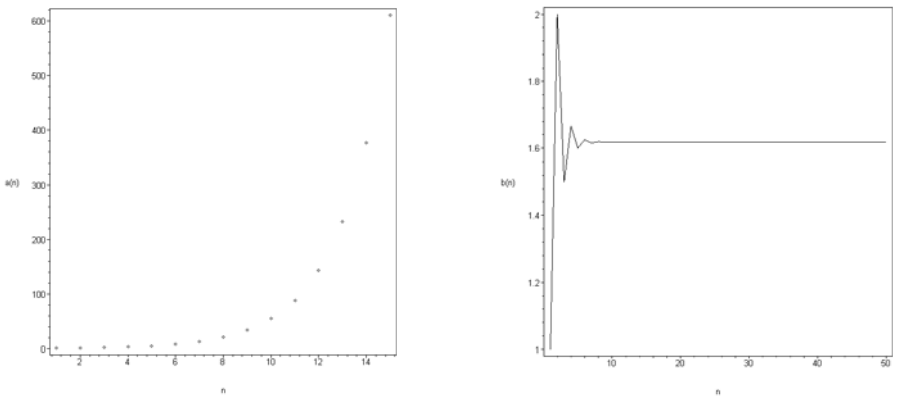
1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, …

이 피보나치수열의 각 항을 앞의 항으로 나눈 비 $b(n) = \frac{a(n+1)}{a(n)}$, $n = 1, 2, \ldots$을 계산하여 보면

1, 2, 1.5, 1.666667, 1.6, 1.625, 1.615385, 1.619048, 1.617647, 1.618182, …

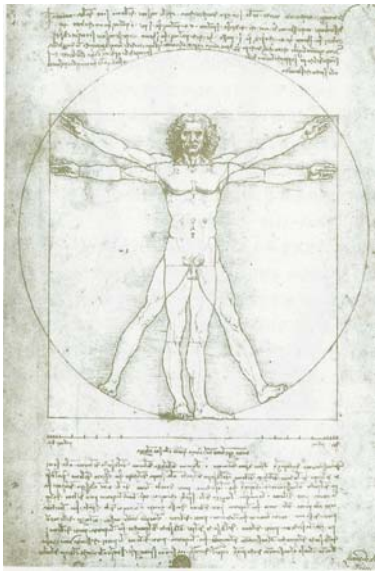
이 되고 이 비는 하나의 무리수인 특정값 $\phi = \frac{1+\sqrt{5}}{2} \approx 1.618034$로 수렴하게 되는 데 이 수렴값을 황금비(golden ratio), 황금분할(golden section), 신성비례(divine proportion)라

부른다. 다음 [그림 5.12]는 이 피보나치수열 $a(n)$의 15항까지 그린 그림과 황금비에 수렴하는 $b(n)$의 값을 나타내는 그림이다.



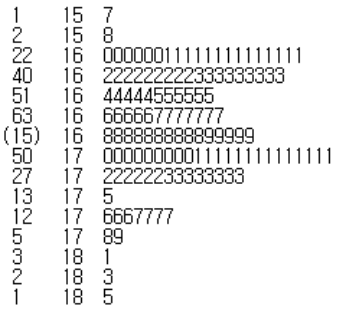
[그림 5.12] 피보나치수열 $a(n)$과 $b(n)$의 값

고대 서양의 예술가들은 그림, 조각상, 꽃병 등 각종 예술작품에 이 황금비를 적용하였고 이 황금비는 인체분석을 위한 이상적인 수단으로 여겼다(Atalay(2004)). 서양인의 미의식에서는 신장을 배꼽높이로 나눈 비(앞으로 이 비를 '키/배꼽높이 비'라 칭하겠다.)도 황금비가 된다고 여겼다. 레오나르도 다빈치는 '다빈치코드'라는 소설을 통해서도 더욱 회자되는 르네상스 시대의 예술가이자 과학자이다. 다음 [그림 5.13]은 이 레오나르도 다빈치가 그린 유명한 'Vitruvius의 인체비례'이다. 그림 속의 남자는 팔과 다리를 대자로 펼치면 원에 내접하고 배꼽이 원의 중심이 된다. 또한 신장과 두 팔을 벌린 길이가 같고 키/배꼽높이 비가 황금비가 된다.



[그림 5.13] Vitruvius의 인체비례

Atalay(2004)는 미국대학생들 21명(남자 10명, 여자 11명)을 대상으로 키/배꼽높이 비를 조사하여 평균이 1.618, 표준편차가 0.04가 됨을 밝혔다. 다음 [그림 5.14]는 2005년 3월 P대학교 학생 128명(남자 64명, 여자 64명)을 대상으로 키/배꼽높이 비를 조사하여 그린 줄기-잎 그림이다.



[그림 5.14] 128명의 키/배꼽높이 비를 나타내는 줄기-잎 그림(잎 단위: 0.01)

다음 [표 5.6]은 128명의 키/배꼽높이 비에 대한 기초통계량들을 나타낸 표이다.

평균	중앙값	표준편차	분산	Q_1	Q_3	최소값	최대값
1.6803	1.6802	0.0553	0.00306	1.6313	1.7157	1.5714	1.858

[표 5.6] 128명의 키/배꼽높이 비에 대한 기초통계량들

(1) 우리는 위의 [그림 5.14]를 통하여 무엇을 알 수 있는가?

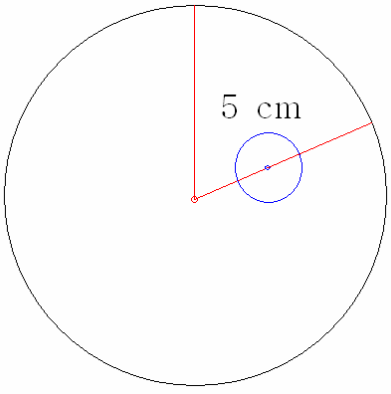
(2) [그림 5.14]와 [표 5.6]을 참조하여 미국대학생 키/배꼽높이 비와 한국대학생 키/배꼽높이 비를 비교하라.

5장 실습문제

5.1 사격장에서 공기총을 쏴 보거나 양궁장에서 화살을 쏴 보았던 경험들이 있을 것이다. 양궁장에서 화살을 쏠 때 이 문제를 변동의 문제로 보기 위하여 관심의 초점을 과녁의 동심원들의 중심(앞으로 '원중심'이라 하자.)과 꽂힌 화살촉 사이의 거리에 두고 보자. 화살을 100번 던져 보면 100개의 화살촉 자국이 원중심을 중심으로 흩어져 있을 것이다. 우리는 화살을 던질 때 원중심을 향하여 던지는데 결과는 원중심을 중심으로 흩어지게 되는 것이다. 이러한 현상이 바로 '변동'이라는 현상이다. 양궁장에서 화살을 쏜 후 원중심과 꽂힌 화살촉 사이의 거리를 재어 기록한 후 이 자료를 이용하여 줄기와 잎 그림이나 히스토그램을 그리면 오른쪽으로 치우친 비대칭분포를 이루는 것을 확인하게 될 것이다. 제조업체에서 제품을 만들 때 제품의 규격(specification)에서 목표값(target value)을 설정하고 제품이 이 목표값을 갖도록(예로 나사를 제조하는 제조업체에서 나사의 지름의 목표값이 1mm라고 하면 나사의 지름이 1mm보다 커도 안 되고 나사의 지름이 1mm보다 작아도 안 된다.) 노력을 하여도 목표값을 중심으로 흩어지게 된다. 제조업체에서는 이 제품의 변동을 어떻게 하면 최대한 줄일 것인가를 강구하게 되는 것이다. 서비스산업에서도 요원들의 서비스품질을 어떻게 균질화할 것인가, 즉 서비스품질의 변동을 어떻게 하면 줄일 것인가가 기업의 사활이 걸린 문제가 되곤 한다.

위의 예와 같은 과녁맞히기놀이보다 더 손쉬운 방법이 동전을 가지고 하는 놀이일 것이다. 동전을 이용하여 다음과 같은 실습을 행한다.

(1) 다음과 같은 반지름이 5cm인 원을 도화지에 그리고 이 원의 중심으로부터 1m 떨어진 거리에서 동전을 100번 던졌을 때 동전의 중심과 원의 중심과의 거리(0.1㎝까지) 및 각도(degree, 북쪽에서 시계방향 0.1°까지)를 재어 기록한다. 동전이 구르는 경우는 제외한다.



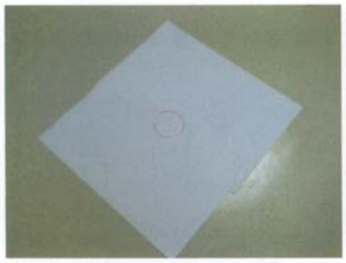
(2) 거리와 각도 자료 각각에 대하여 히스토그램과 줄기와 잎 그림을 원생들에게 그려보게 한다. 또한 동전 100개를 던진 것 중 원 안에 들어가는 경우의 비율을 구하게 한다.

(3) 원생 전체의 결과를 수집하여 분석하여 본다.

<목적> 1. '변동(variation)'을 체험하게 한다.
2. 대칭분포와 비대칭분포를 체험하게 한다.

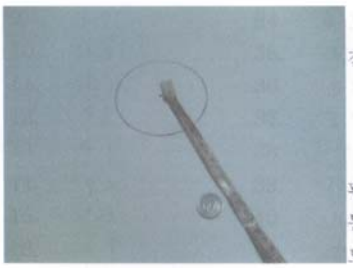
<참고> 다음 사진들은 위의 실험 과정을 사진으로 찍은 것들이다.

숙제 1) 100원짜리 동전을 100번 던져 각각의 동전의 중심과 원의 중심과의 거리를 재어 기록하기 --> 이 거리는 어떤 분포를 이룰까?



(먼저 큰 종이위에 지름 5cm인 원을 그려 넣었다.)

(그리고 1m 떨어진 지점을 표시하여 그곳에서 동전을 던졌다.)



(동전을 던지고 나면 줄자로 원의 중심과 동전의 중심과의 거리를 재어서 기록하였다.)

** 이번 숙제를 약간 폭신하다면 폭신한 장판이 깔린 방에서 해서 그런지 동전을 던질때 동전이 그리 많이 튀지 않고 착 달라붙어서 어느 한 경우도 저 흰종이 밖을 벗어나는 경우는 없었다.

어느 한 원생의 실험 결과(동전을 100번 던졌을 때 동전의 중심과 원의 중심과의 거리)를 보면 다음과 같았다.

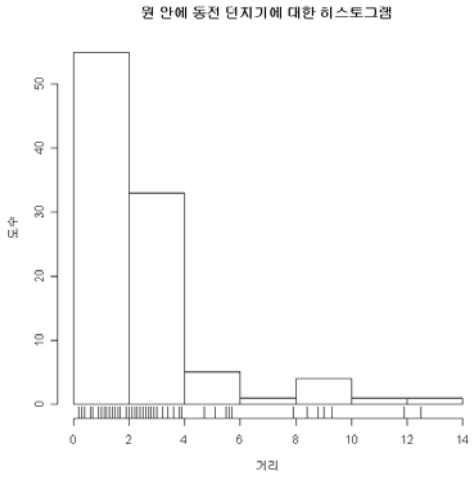
1.0, 1.4, 1.2, 1.4, 1.7, 2.2, 3.8, 7.9, 5.6, 5.1, 3.2, 2.5, 1.7, 1.4, 1.0, 0.7, 1.6, 2.2, 2.5, 9.0,
9.3, 3.9, 2.3, 2.4, 1.0, 1.1, 1.1, 1.4, 2.0, 0.7, 3.8, 2.6, 1.6, 1.2, 0.3, 0.4, 2.1, 1.9, 8.4, 5.5,
1.2, 1.1, 1.0, 1.5, 1.7, 2.9, 3.0, 0.6, 0.9, 11.9, 1.1, 1.5, 2.7, 2.9, 0.9, 1.3, 5.7, 3.6, 0.6, 0.4,
2.1, 2.5, 2.0, 1.6, 2.7, 3.6, 2.6, 0.9, 2.3, 3.0, 1.2, 1.0, 0.3, 1.5, 2.5, 3.2, 3.4, 2.6, 4.7, 8.8,
1.2, 1.0, 0.4, 3.0, 2.9, 2.6, 2.8, 1.0, 1.7, 1.4, 1.2, 0.7, 1.0, 1.9, 2.4, 1.2, 0.9, 0.4, 12.5, 0.2

이 자료에 대하여 줄기와 잎 그림을 그리면 다음과 같다. 동전 100개를 던진 것 중 원 안에 들어가는 경우의 비율은 0.89이고 동전의 중심과 원의 중심과의 거리는 오른쪽으로 치우친 비대칭분포를 이루는 것을 확인할 수 있다.

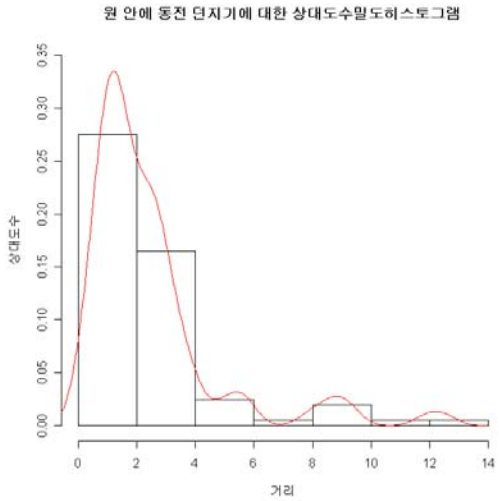
The decimal point is at the |

```
 0 | 2334444667779999000000001111222222234444455566677779 9
 2 | 00112233445555666677899900022466889
 4 | 71567
 6 | 9
 8 | 4803
10 | 9
12 | 5
```

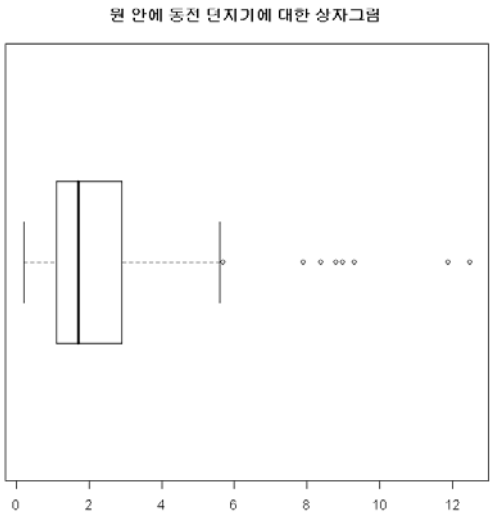
히스토그램을 그리면 다음과 같다. 오른 쪽으로 치우친 비대칭분포를 확인할 수 있다.



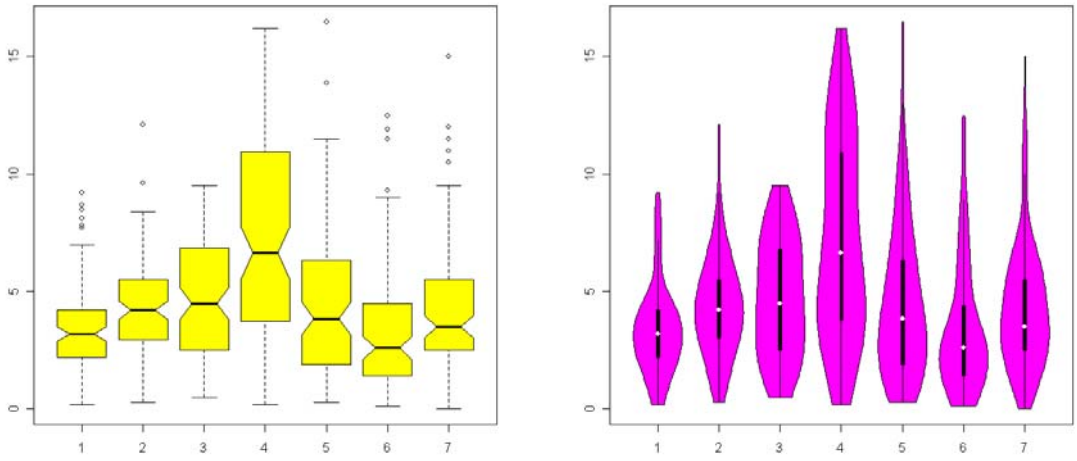
상대도수밀도히스토그램을 밀도함수추정량과 함께 그리면 다음과 같다.



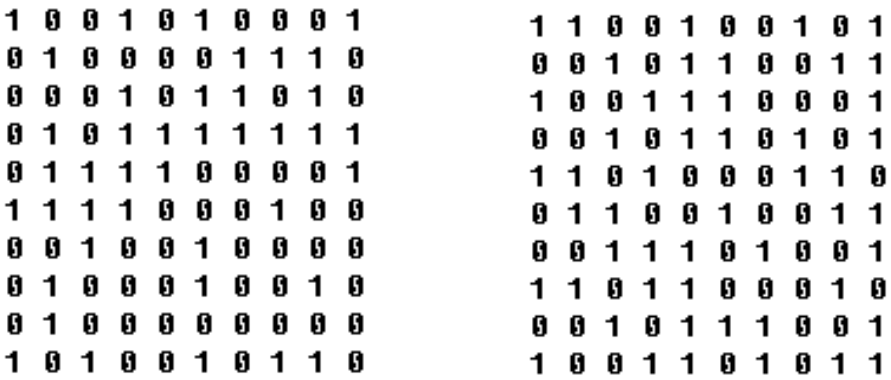
참고로 상자그림을 그리면 다음과 같다. 특이값(이상값, outlier)이 8개(5.7, 7.9, 8.4, 8.8, 9.0, 9.3, 11.9, 12.5)가 있음을 알 수 있다.



아래는 7명의 원생이 실험한 결과를 상자그림과 바이올린그림으로 나타낸 그림이다. 각 개인의 패턴이 다양함을 알 수 있다. 그러나 모두 큰 값 쪽으로 꼬리가 있는 비대칭분포를 이루고 최빈값이 2~3cm 내에 존재함을 알 수 있다.



5.2 다음 그림 중 하나는 100개의 동전을 실제로 던진 결과이고 하나는 100개의 동전을 실제로 던지지 않고 가짜로 머릿속에 떠오르는 데로 작성한 결과이다. 여기서 0은 숫자면, 1은 그림면이다. 어느 것이 가짜이고 어느 것이 진짜일까?



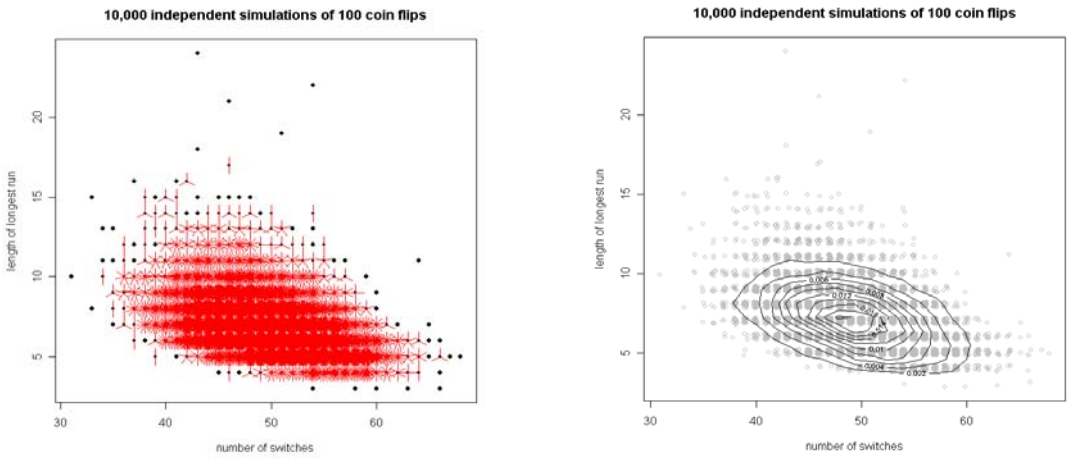
답은 왼쪽이 가짜이고 오른쪽이 진짜이다. 이 결과는 어느 교육원생이 실험한 결과이다. 우리는 다음과 같은 실습을 통하여 가짜를 알아내는 방법을 배울 수 있다(물론 100% 알아낼 수는 없다.).

교육원생 각자 다음과 같은 실습을 행한다.

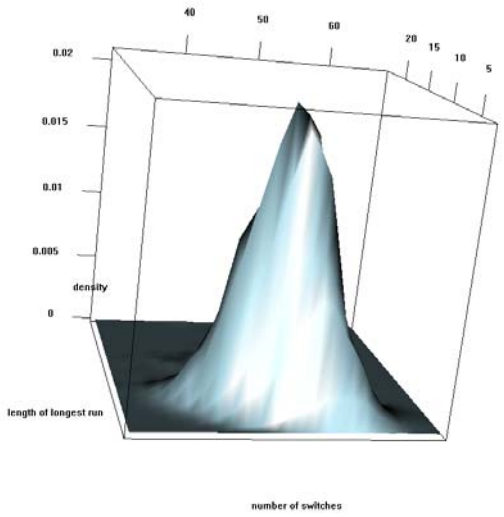
1. 100개의 동전을 실제로 던지지 않고 가짜로 머릿속에 떠오르는 대로 가짜로 작성하여라. (참고) 동전의 그림면(1)이 나올 확률은 숫자면(0)이 나올 확률과 같다.
2. 100개의 동전을 실제로 던져 나온 결과를 기록하여라.
3. 1과 2 각각에 대하여 그림면과 숫자면이 바뀌는 횟수(number of switches)와 그림면이나 숫자면이 연속적으로 나오는 연(run) 중 가장 긴 길이(length of largest run)를 세어 기록한다.
4. 전체의 결과를 수집하여 그림면과 숫자면이 바뀌는 횟수를 x축에, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이를 y축으로 하여 원생 각자의 결과를 좌표로 하여 진짜점은 파란색으로, 가짜점은 빨간색으로 하여 점을 그린 후 이 점들을 가짜점에서 시작하여 진짜점에서 끝나는 화살표시가 있는 점선으로 연결한다.
5. 어떤 패턴이 있는 가에 대하여 토론한다.

<목적> 통계의 조작이 발생할 때 이 조작을 발견할 수 있는 수단을 필히 강구하여야 한다는 사실을 강조하기 위하여 실습을 행한다.

<참고> 컴퓨터 시뮬레이션으로 100개의 동전을 던지는 모의실험을 10,000번을 행하여 각각에 대하여 그림면과 숫자면이 바뀌는 횟수(number of switches)를 x축에, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이(length of largest run)를 y축으로 하여 점을 찍은 결과가 다음 그림들이다. 왼쪽 그림은 해바라기그림이다. 많이 나타나는 점들이 빨갛게 물들어져 있어 집락을 형성하고 있다. 오른쪽 그림은 jittering을 행한 후 점들을 회색으로 찍은 후 그 위에 등고선도를 그렸다. 뾰쪽한 산 모양 형태를 이룸을 알 수 있다.

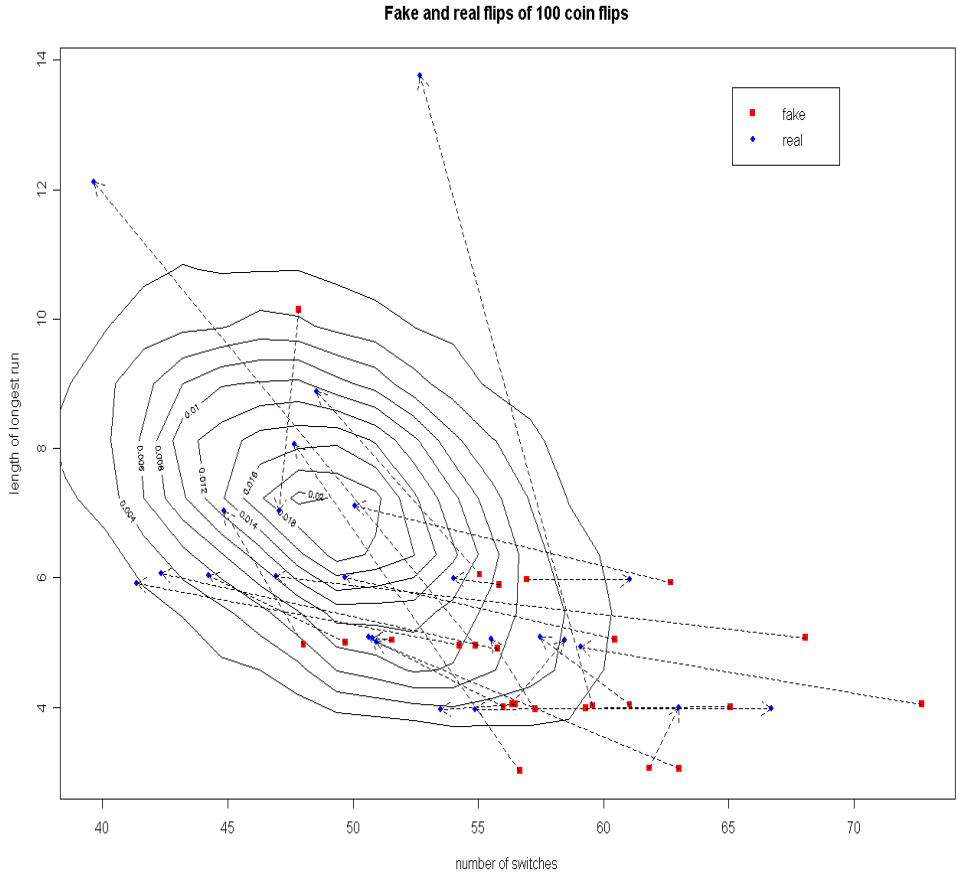


다음 그림은 위의 시뮬레이션 결과를 이용하여 밀도추정을 행하여 그린 3차원 그림이다. 앞의 등고선도와 비교하여 보라.



25명에게 앞에서와 같은 실습을 시킨 결과가 다음 그림이다. 교육원생 전체의 결과를 수집하여 그림면과 숫자면이 바뀌는 횟수(number of switches)를 x축에, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이(length of largest run)를 y축으로 하여 원생 각자의 결과를 좌표로 하여 진짜점은 파란색으로, 가짜점은 빨간색으로 하여 점을 그린 후 이 점들을 가짜점에서 시작하여 진짜점에서 끝나는 화살표시가 있는 점선으로 연결하였다. 대체적으로 그래프의 남동쪽에 있던 빨간점들(가짜 결과들)이 그래프의 북서쪽에 있는 파란점들(진짜 결과들)로 이동하고 있음을 알 수 있다. 왜 이런 패턴이 발생할까? 일반적으로 사람들은 그림면과 숫자면이 나오는 현상이 무작위하다(random)고 생각하기 때문에 그림면과 숫자면을 자주 바꾸게 되

어 그림면과 숫자면이 바뀌는 횟수가 많아지는 반면, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이는 짧게 된다. 그래서 빨간점들(가짜 결과들)이 그래프의 남동쪽에 주로 위치하게 된다. 그런데 실제 던져보면 우리가 언뜻 생각하는 것만큼 그림면과 숫자면이 상대적으로 자주 바뀌지 않기 때문에 그림면과 숫자면이 바뀌는 횟수가 상대적으로 적어지는 반면 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이는 상대적으로 길게 된다. 그래서 파란점들(진짜 결과들)이 그래프의 북서쪽에 주로 위치하게 된다. 물론 이러한 패턴을 따르지 않는 사람들(특이값)도 있다. 그러나 이러한 사람들의 숫자는 그리 많지 않다.



쉬어가기

1. 가중평균과 가중치

단순산술평균은 각 자료값의 상대적 중요성을 동일한 것으로 간주한다. 만약 각 자료값이 서로 다른 중요성을 가지고 있다면 각 자료값의 중요성을 반영한 가중산술평균을 사용하여야 한다. 가중산술평균은 각 자료의 값에 자료의 상대적 중요도(가중치)를 곱하여 합계한 값을 가중치의 합계로 나누어 구한다.

$$\text{가중산술평균} = = \frac{X_1 w_1 + X_2 w_2 + \cdots + X_n w_n}{w_1 + w_2 + \cdots + w_n}$$

여기서 w는 각 자료의 가중치를 의미한다.

예를 들어 A사와 B사라는 두 회사를 가정해 보자. A사의 직원은 50명, B사의 직원은 100명이라고 하고 A사 직원의 연간수입은 일률적으로 1,000만원, B사 직원의 연간수입은 일률적으로 2,000만원이라고 한다면 직원수를 가정한 가중평균은 다음과 같이 계산된다.

(1,000만원 × 50명 + 2,000만원 × 100명) ÷ (50명 + 100명) = 약 1천6백6십6만7천원

가중평균은 물가지수의 계산, 주가지수의 계산, 표본추출 등에 광범위하게 사용되는 개념으로 꼭 필요하다. 하지만 실제 사용하다 보면 어떤 가중치를 줄 것인가의 고민에 직면하게 되는 경우가 많다. 표본추출에서 이 가중의 개념을 사용하는 이유를 살펴보면 다음과 같다.

① 영역별로 다른 추출률(불균등 추출확률)에 대한 개선이 필요하기 때문이다. 예를 들어 경제활동인구조사의 경우 서울, 경기의 추출률은 기타지역의 추출률에 비해 낮으므로 조치가 필요하다.
② 무응답으로 인한 편향(bias)을 보정할 필요가 있다.
③ 추출틀의 부정확, 전체를 커버하지 못하는 등에 인한 편향을 줄여주는 효과가 있다.
④ 이미 알고 있는 모집단 분포를 이용한다면 사후층화 등을 통해 추정치 정도를 제고할 수 있다.

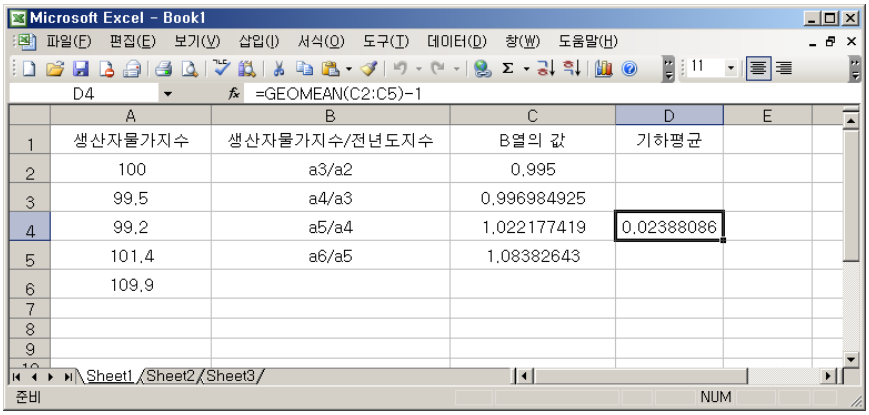
부연해서 설명하자면, 수집된 자료를 그냥 쓰지 않고 가중치 보정을 해야 하는 이유는 우연성 때문이다. 완벽한 표본기법을 쓰더라도 해당 표본이 전체 모집단 특성과 비교해 임의오차 범위내 차이가 있을 수 있다. 이 때 인구학적 특성으로 고려할 수 있는 것은 성, 인종, 지역, 나이, 교육, 소득 등이다. 예컨대 모집단 남녀비는 56대 44인데 표본은 48대 52라면 각 응답

자에게 가중치를 성별로 달리 주어 모집단 성비에 맞춘다는 것이다. 성별뿐 아니라 나이, 교육수준까지 모집단 비율로 표본비율을 맞추는 자료보정방법이 다양하게 사용되고 있다. 가중치를 주어야 하는 다른 세 가지 이유는 전화가 2대인 가구는 전화가 1대 뿐인 가구에 비해 2배이므로 가중치를 1/2로 주어야 옳다고 할 것이다. 두 번째는 가구원수가 많은 가구라면, 한 가구에서 한 명만 조사되므로 특정 1인이 표본에 선발될 확률이 작아진다. 그러므로 각 응답자의 가구원수에 비례하는 가중치 부여가 옳다. 마지막으로 투표의향이 응답자마다 달라서이다. 투표의향이 전혀 없는 경우 가중치는 0이 주어져야 한다.

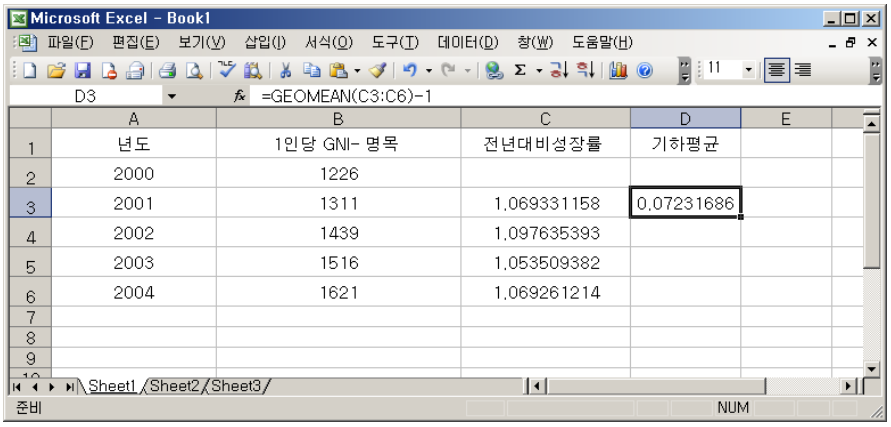
2. 기하평균

본문에서 다뤄진 대로 기하평균은 연평균 물가상승률, 경제성장률 등 일정기간에 걸친 평균 변화율을 산출하는데 주로 이용되고 있다. 여기서는 예만 살펴보기로 하겠다.
(1) 우리나라의 연평균 생산자물가상승률을 구해보자. 생산자물가지수가 2000년에 100.0, 2001년에 99.5, 2002년에 99.2, 2003년에 101.4 2004년에 109.9라고 가정해 보자.

여기서 이 지수를 대상으로 직접 기하평균을 계산하면 안된다. 지수는 전년도를 기준으로 계산한 값이기 때문이다. 즉 Excel로 =geomean(100.0, 99.5, 99.2, 101.4, 109.9)라고 하면 곤란하다. 정확한 계산식은 =geomean(99.5/100.0, 99.2/99.5, 101.4/99.2, 109.9/101.4) -1 로 계산해야 한다. 답은 0.023881, 약 2.4%라고 할 수 있다.

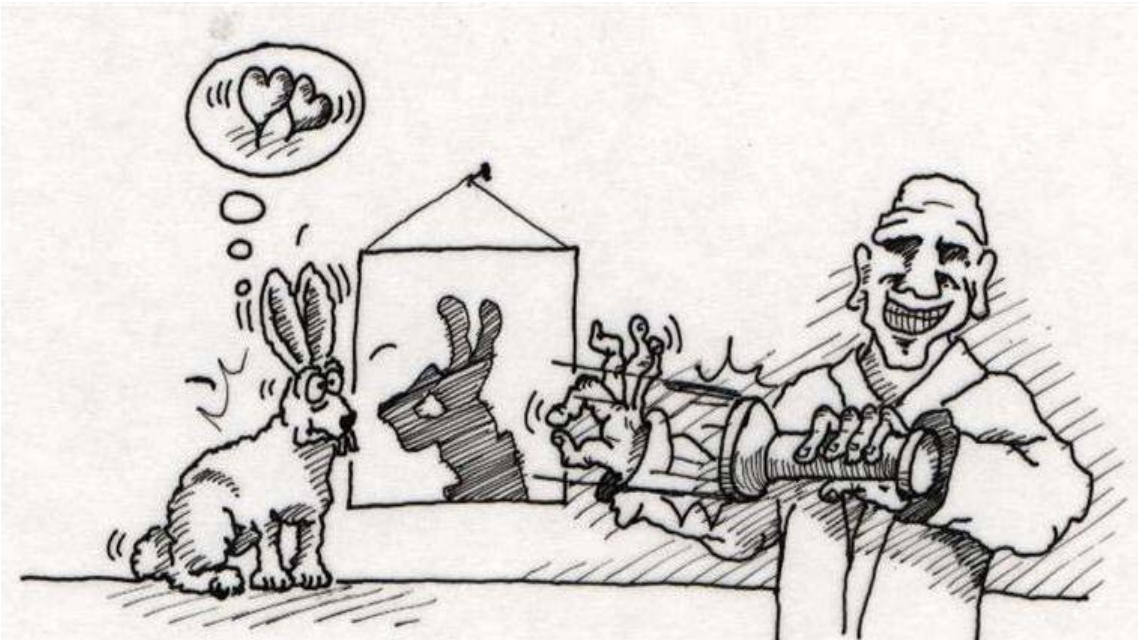


(2) 우리나라의 연평균 성장률을 구해보자. 1인당 GNI가 2000년에 1226, 2001년에 1311, 2002년에 1439, 2003년에 1516, 2004년에 1621 이라고 가정해 보자. 이때에도 전년도 대비 성장률을 계산하여 기하평균을 계산하는 것이 맞다. 만일 전년대비 성장률(%)이 직접 주어지는 경우에는 바로 계산하는 것이 맞다. 아래의 엑셀 결과를 참조하라.


Microsoft Excel - Book1
파일(F) 편집(E) 보기(V) 삽입(I) 서식(O) 도구(T) 데이터(D) 창(W) 도움말(H)
D3
=GEOMEAN(C3:C6)-1
A B C D E
1 년도 1인당 GNI- 명목 전년대비성장률 기하평균
2 2000 1226
3 2001 1311 1,069331158 0,07231686
4 2002 1439 1,097635393
5 2003 1516 1,053509382
6 2004 1621 1,069261214
7
8
9
Sheet1 Sheet2 Sheet3
준비
NUM

제 6 장

연관성은 왜 알아보는가?



차 례

학습목표

제 5장의 내용은 자료를 수치적으로 표현하는 데 있어 대상은 단일 변수였다. 변수 하나에 대해 중심경향적인 값과 산포도에 대해 집중적으로 알아보았지만 제 6장의 내용은 변수들 간의 관계를 측정하는 몇 가지 방법을 알아 볼 것이다. 공분산 및 상관계수를 통해 자료의 연관성을 알아보고 주의할 점에 대해서도 언급한다. 또한 분할표를 통해 자료를 표현하게 되면 분할되는 변수 간의 관계를 측정하는 방법 및 이를 도시화하는 방법 역시 필요하게 된다. 마지막으로 측정된 원래의 단위가 아닌 변환된 단위로 연관성을 표시하게 되면 손쉽게 모형화하는 경우를 알아본다.

6.1 관계성 측정

지금까지의 모든 수치적 요약방법은 모두 변수 하나에 국한되어 설명이 되었다. 그러나 변수가 2개 이상일 때 변수 간의 관계를 알아보고자 한다면 제일 대표적인 방법은 산점도이다. 산점도는 특히 자료의 개수가 클 때 유용하며 x-축에는 y-축의 변수를 설명하는 변수를 지정하는 것이 좋다. 그리고 산점도를 통해 다음과 같은 사항을 시각적으로 확인하여 본다.

- 두 변수 간의 관계의 형태 (선형, 비선형)
- 관계의 강도의 크기
- 관계의 방향(음, 양의 방향)
- 이상점 유무

그리고 이러한 산점도에 두 변수의 관계를 잘 보여줄 수 있는 선을 덧붙여 놓기도 한다. 이와 같은 방법으로 세 가지가 제안되는데 모두 기본적으로 선과 자료점 간의 거리가 최소화하도록 그려야 한다.

- 직선
- 곡선을 나타내는 이차나 다항함수 곡선
- 평활선

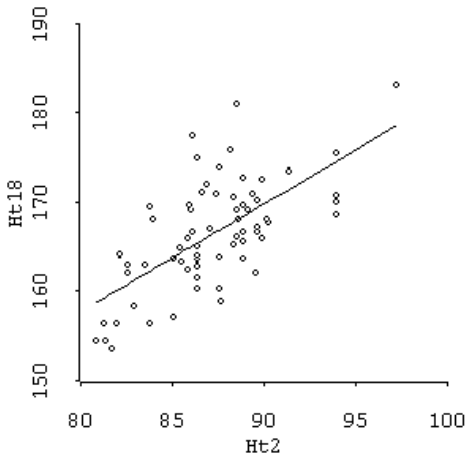
예제 6.1 [표 6.1]은 미국 버클리 대학에서 발표한 2세, 9세, 18세 때의 여성의 몸무게와 키의 자료이다. 몸무게는 WT2, WT9, WT18, 키는 HT2, HT9, HT18로 구분되어 있다. 예를 들어 WT2는 2세 때의 몸무게이고 HT2는 2세 때의 키이다. 몸무게는 kg으로, 키는 cm로 측정하였다.

<몸무게와 키.xls>

	A	B	C	D	E	F
1	wt2	wt9	wt18	ht2	ht9	ht18
2	13.6	32.5	56.9	87.7	133.4	158.9
3	11.3	27.8	49.9	90	134.8	166
4	17	44.4	55.3	89.6	141.5	162.2
5	13.2	40.5	65.9	90.3	137.1	167.8
6	13.3	29.9	62.3	89.4	136.1	170.9
7	11.3	22.8	47.4	85.5	130.6	164.9
8	11.6	30	57.3	90.2	136	168.1
9	11.6	24.3	50	82.2	128	164
10	12.4	29.9	58.8	85.6	132.4	163.3
11	17	44.5	80.2	97.3	152.5	183.2
12	12.2	31.8	59.9	87.1	138.4	167
13	15	32.1	56.3	88.9	135.2	163.8
14	14.5	39.2	67.9	87.6	142.3	174
15	10.2	23.7	52.9	82.6	129.1	163
16	12.2	26	58.5	87.1	133.2	167.1
17	12.8	36.3	73.2	84	136.3	168.1
18	13.6	29.9	54.7	83.6	133.1	163
19	10.9	22.2	44.1	81.4	123.2	154.6
20	13.1	34.4	70.5	89.7	135.8	170.3

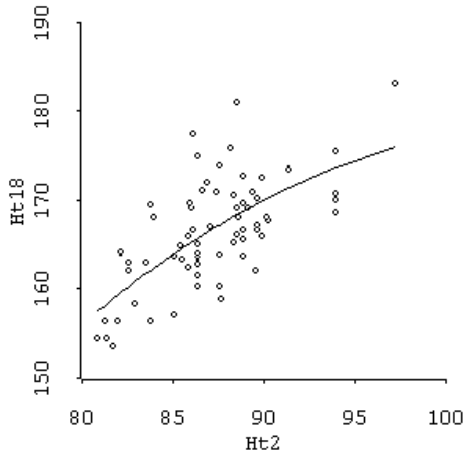
[표 6.1] 버클리대학 여성 몸무게 키 자료

[그림 6.1]은 산점도에 두 변수 간의 관계가 직선이라 짐작하고 산점도에 직선을 적합 시킨 그림이다. 자료의 x-축은 2세 때 키며 y-축은 18세 때 키를 나타낸 것이다. 선은 직선을 적합 시켰을 때 생기는 거리의 제곱의 합을 최소화하는 선을 그린 결과이다. 이런 추정방법을 최소 제곱법이라 하는데 이는 제 12장에서 자세히 배울 것이다. 산점도의 결과에 의하면 대부분의 자료는 선을 따라 양의 방향으로 움직임을 알 수 있다. <산점도.xls>



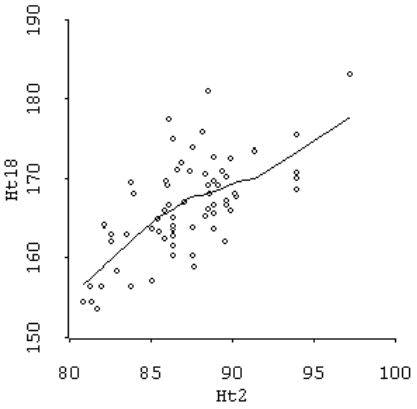
[그림 6.1] 직선

그러나 2세때 키가 비교적 크다 하더라도 18세에 도달 했을 때에는 키의 증가 속도가 다른 경우에 비해 둔화된다고 가정한다면 이차 함수 식을 이용하여 선을 적합하는 것이 좋을 것이다. [그림 6.2]를 참조하기 바란다. <산점도.xls>



[그림 6.2] 이차함수

그리고 만약 무슨 선을 그어야 할지 확신이 서지 못하는 경우는 평활선을 그려 추세를 알아볼 수 있다. 평활선을 그리는 방법은 책의 범위를 벗어난다. 그러나 많은 통계소프트웨어는 사용자가 전문적인 지식 없이도 선을 그릴 수 있는 인터페이스를 제공한다. [그림 6.3]에서는 추세선이 잠시 중간부분에서는 잠시 멈칫하다 다시 증가하는 형상을 보인다. ■ <산점도.xls>



[그림 6.3] 평활선

주의할 점은 자료가 통제된 실험계획에 의해 수집이 되지 않는 한, 변수 간의 인과 관계를 산점도가 알려주지는 않는다. 다음 절에서는 직선에 의한 관계식을 가정하였을 때 두 변수 간의 관계를 요약하는 방법에 알아보도록 하자.

6.2 공분산 및 상관계수

여기서 설명하는 개념은 공분산과 상관계수이다. 이는 산점도에 표시된 변수 관계의 크기와 부호를 정량화하기 위해서 만들어진 값들이다.

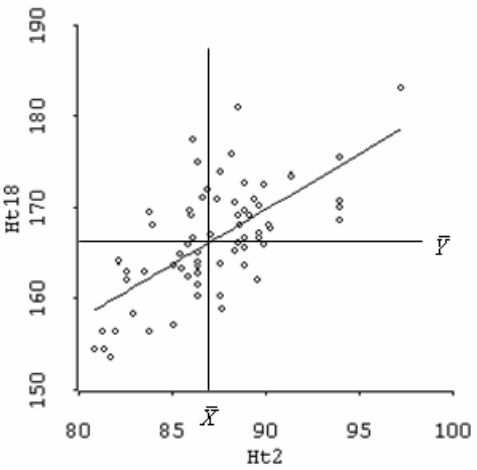
먼저 공분산을 정의하기 위해서는 두 변수의 값의 쌍을 만들어 볼 필요가 있다. 주어진 자료의 i-번째 관측값에서 한 변수의 관측값을 x_i , 그리고 다른 변수의 관측값을 y_i 라면 자료쌍은 (x_i, y_i) 이다. 즉, n개의 이런 쌍이라 만들어질 것이다. 이러한 쌍을 좌표 면에 표시한 것이 산점도이다.

변수 X와 Y의 공분산은 식 (6.1)에 의해 정의되는데 $Cov(X, Y)$ 라 표기한다.

공분산 : 공분산(covariance)은 식 (6.1)과 같은 정의에 의해 만들어진 값이다.

$$Cov(X, Y) = \frac{\sum_{i}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \quad (6.1)$$

위에서 소개한 실제 키의 자료를 이용하여 이러한 공분산을 이해해 보자. [그림 6.4]에서 2세 때의 키와 18세 때의 키의 산점도를 다시 그렸는데 x, y 축과 더불어 $\overline{x}$ 와 $\overline{y}$ 를 새로운 축으로 가운데 설정하였다.



[그림 6.4] 직선을 가정한 평면 4분할

$(x_i - \overline{x})(y_i - \overline{y})$ 의 쌍의 값은 (x_i, y_i) 쌍의 값이 어느 분면에 위치하고 있는 점이냐에 따라 그 부호가 바뀐다. 제 1사분면에 있다면 + 곱하기 + 의 부호인 양의 부호가 부여될 것이고 역시 제 4사분면에 위치하고 있다면 - 곱하기 - 가 되어 양의 부호일 것이다. 유사한 이유로 (x_i, y_i) 이 2사분면이나 3사분면에 위치하고 있었다면 음의 부호를 가질 것이다. 모든 n개의 $(x_i - \overline{x})(y_i - \overline{y})$ 의 값을 다 더하면 크기와 부호에 따라 양 혹은 음의 부호를 가지는 하나의 값을 가질 것이다. 이를 n-1로 나는 것이 공분산이다. 따라서 공분산 역시 평균, 분산과 마찬가지로 평균의 개념이 포함되어 있다. 엑셀에서는 변수 X와 Y의 공분산은 다음과 같은 명령문을

$$= cov(X,\ Y)$$

이용하면 된다. 그러나 공분산은 X의 원래 단위와 Y의 원래 단위의 곱형태이며 크기는 단위에 따라 값이 달라지기 때문에 직접 사용하기에는 문제가 있다. 따라서 이 단위에 상관없이 두 변수의 관계를 알아보는 방법은 없을까? 이것이 상관계수이다.

상관계수 : 상관계수(correlation coefficient)는 공분산을 각각의 변수의 표준편차로 나눈 값으로 정의된다. 따라서 분자와 분모의 단위가 같아지므로 상관계수는 단위가 없는 상태가 된다. 그리고 이렇게 정의된 상관계수는 -1과 1 사이에서(-1과 1을 포함해서) 값을 가진다. 변수가

어떤 단위가 주어진다 하더라도 모든 상관계수의 값은 -1과 1 사이에서 값을 가진다는 의미이다. 상관계수는 $Corr(X, Y)$ 로 표기한다. 식 (6.2)가 상관계수에 대한 정의이다.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\mathrm{sd(X)sd(Y)}} \tag{6.2}$$

여기서 sd는 표준편차를 의미한다. 엑셀에서는 상관계수의 명령문은 다음과 같이

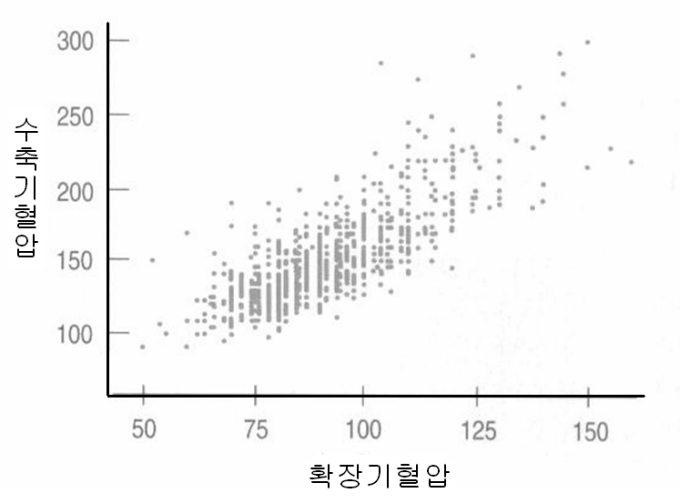
$$= correl(X,\ Y)$$

구현된다. 앞의 산점도에서 보았던 두 변수의 관계의 크기를 상관계수로 나타낸다면

$$Corr(X, Y) = \frac{Cov(X, Y)}{\mathrm{sd(X)sd(Y)}} = \frac{13.22923}{3.330523 \cdot 6.074886} = 0.663335$$

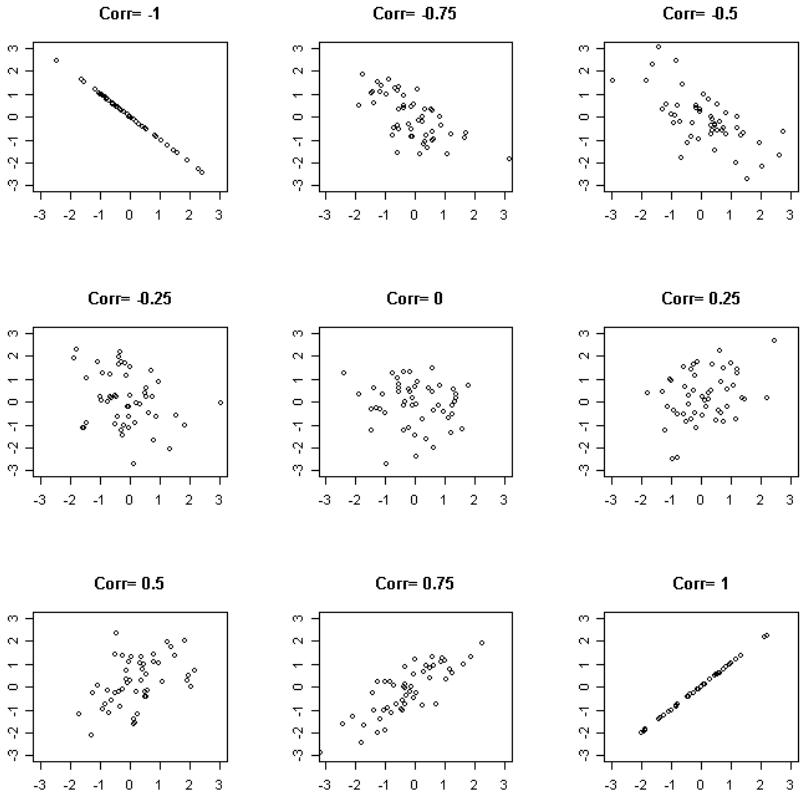
로 나타난다.

예제 6.2 다음 [그림 6.5]는 사람들의 확장기혈압(Diastolic BP(BP: Blood Pressure))과 수축기혈압(Systolic BP)에 대한 산점도이다. 표본상관계수를 구하면 0.792로 두 변수 사이의 직선관계가 강함을 알 수 있다. ■



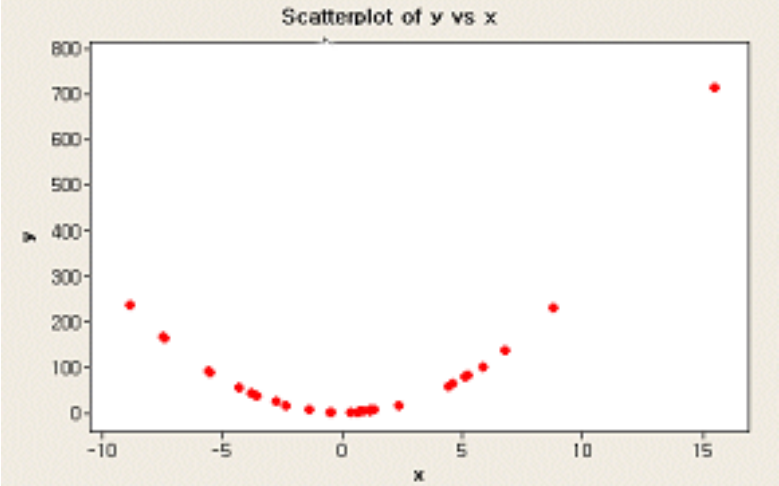
[그림 6.5] 확장기혈압과 수축기혈압에 대한 산점도

상관계수의 값이 얼마나 두 변수의 관계를 정도를 나타내 주는지 알기 위해서는 약간의 훈련이 필요하다. [그림 6.6]을 통해 두 변수 값의 관계 정도에 따른 상관계수의 값을 짐작하기 바란다.



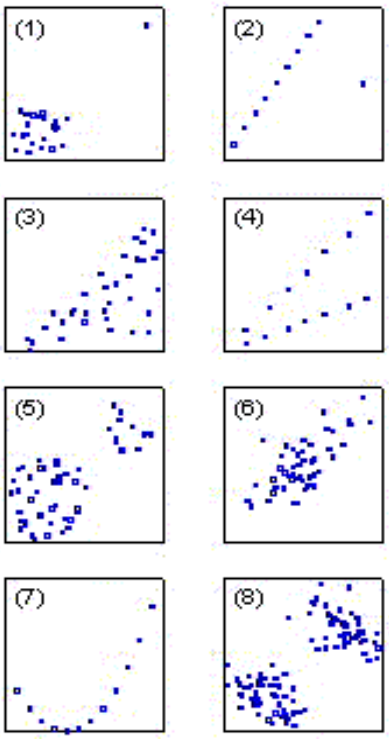
[그림 6.6] 상관계수의 변화

두 변수의 상관계수에서 이야기하는 관계는 선형의 관계를 의미한다. 즉, 상관계수는 선형의 관계의 정도를 나타내주는 값으로 이해하여야 한다. 그리고 모든 관측값이 기울기가 양인 직선 위에 있다면 상관계수는 +1 그리고 음의 기울기의 직선 위에 모든 관측값이 있다면 -1이 된다. 그리고 상관계수가 0 이라면 변수 간에는 기울기가 0이 아닌 직선의 관계는 존재하지 않는다는 의미이다. 그렇다고 변수 간에 아무런 관계가 존재하지 않는다는 의미는 아니다. 선형의 관계가 존재하지 않을 뿐 비선형의 관계가 존재할 가능성은 있다. [그림 6.7]의 관계는 분명 비선형의 관계가 두 변수 간에 존재하나 상관계수의 값은 0.17에 불과한 값이 나온다.



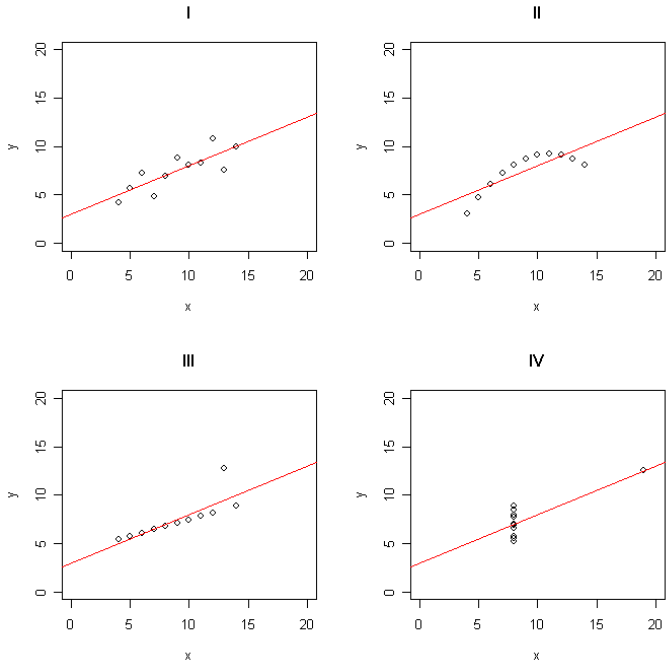
[그림 6.7] 비선형의 관계

또한 같은 상관계수 값이라 하더라도 자료의 관계의 형태에 따라 그 의미를 동일시해서는 안 된다. [그림 6.8]은 Cleveland, Kleiner, 그리고 Tukey의 책에서 인용한 그림인데 관계의 형태에 상관없이 상관계수는 0.7로 나온다. 이렇듯 상관계수는 산점도의 형태를 보지 않는 이상 두 변수와의 관계를 정확하게 반영하였다고 이야기 못한다.



[그림 6.8] 동일한 상관계수 = 0.7

또 다른 예를 보자. 다음 [그림 6.9]는 네 종류의 데이터셋에 대하여 산점도를 그린 것이다. 표본상관계수는 모두 0.82이고 추정된 회귀직선식(12장에서 배우게 된다.)은 $\hat{y}=3+0.5x$로 모두 같으나 점들의 패턴이 모두 다르다.



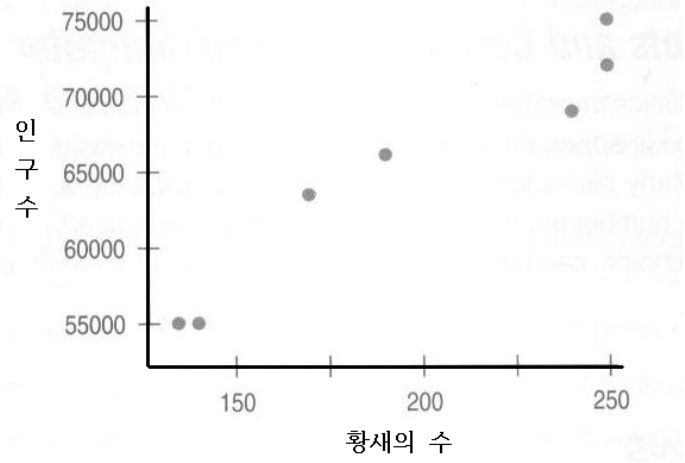
[그림 6.9] 4 종류의 데이터셋에 대한 산점도과 추정된 회귀직선식

III에서 특이값(x=13)을 빼면 표본상관계수가 0.82에서 1이 되고 IV에서 특이값(x=19, 이 점을 통상 영향력이 큰 자료값(influential observation)이라고 칭한다.)을 빼면 표본상관계수가 0.82에서 0이 된다. II에서 14와 20 사이에 완전곡선이 되게 점을 하나 잡으면 표본상관계수가 0.82에서 0이 된다. II와 IV에서 한 점의 첨가나 삭제로 인하여 표본상관계수의 급격한 변화(0.82에서 0)가 나타나는 것이다.

표본상관계수를 사용하는 데 있어서 가장 중요한 통계의 오용문제는 상관관계와 인과관계를 혼동시키는 문제이다. 표본상관계수는 두 변수 사이의 직선적인 관계를 측정하는 측도이지 두 변수 사이의 인과관계를 나타내는 측도는 아니다. 두 변수 A와 B 사이의 상관관계가 있을 때 우리는 다음과 같은 4가지 가능성이 있다.

1. A가 B를 일으킨다.
2. B가 A를 일으킨다.
3. 숨겨져 있는(lurking) 변수 C가 A와 B를 모두 일으킨다.
4. 이러한 상관관계는 순전히 우연에 의하여 일어났다.

해수욕장에서 아이스크림을 산 사람들의 숫자와 물에 빠져 구조를 받았거나 죽은 사람들의 숫자가 상관관계가 있다고 하여서 아이스크림을 사는 것이 물에 빠지는 사건을 일으킨다고 아무도 주장하지 않는다. 사실은 아이스크림을 산 사람들의 숫자와 물에 빠져 구조를 받았거나 죽은 사람들의 숫자에 다 같이 관계있는 변수가 해수욕장에서 온 사람들의 숫자이다. 다음 [그림 6.10]은 1930년대 독일의 Oldenburg시의 7년 동안의 황새 수와 시 인구수에 대한 산점도이다. 표본상관계수는 0.97이어서 황새 수와 시 인구수에 상관관계가 매우 강하다고 할 수 있다. 그렇다고 황새 수가 많아지면 인구수가 늘게 된다고 인과관계로 이야기하면 안 된다. 이러한 상관을 허위상관(spurious correlation)이라고 한다. '황새가 아기를 물어온다'는 신화 때문에 황새가 잘 깃드는 나무를 심었던 역사가 서양에서 있음을 상기하면 통계의 오용이 우리의 삶에 영향을 주고 있음을 다시 한 번 느끼게 된다. 그러면 황새 수와 시 인구수에 상관관계가 매우 강하게 나타난 이유는 무엇일까? 황새 수와 시 인구수에 다 같이 관계를 갖는 숨겨져 있는 변수(예로, 시의 발전성)가 있을 것이다.



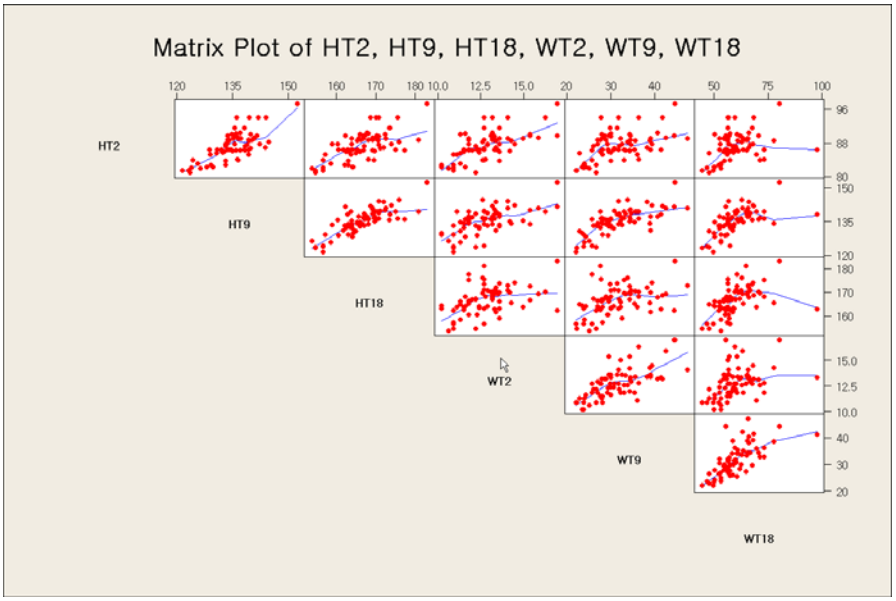
[그림 6.10] 허위상관

상관계수행렬 및 산점도행렬 : 변수가 두 개만 있는 경우 상관계수는 하나의 값이면 충분하지만 세 개 이상이면 많은 수의 상관계수가 존재한다. 이를 하나의 표로 모을 필요가 있는데 이러한 표를 상관계수행렬이라 부른다. [표 6.1]의 자료에서 모든 변수를 대상으로 상관계수행렬을 만들어 보면 [표 6.2]와 같다. 상관계수 행렬은 대칭으로 $Corr(X,Y) = Corr(Y,X)$ 이다. 따라서 그림에서 대각선 하단의 그림은 생략하였다. 그리고 변수 자신과의 상관계수는 당연히 1이 된다. 왜냐하면 $X = Y$ 가 되어 분자, 분모 모두 분산의 공식이 되기 때문이다.

WT2	1					
WT9	0.692539	1				
WT18	0.392152	0.692089	1			
HT2	0.64455	0.522928	0.363717	1		
HT9	0.607125	0.727612	0.609332	0.738356	1	
HT18	0.445099	0.42605	0.497935	0.663335	0.807808	1

[표 6.2] 상관계수 행렬

그리고 모든 변수 사이의 산점도를 한 그림으로 표현하면 [그림 6.11]과 같은 산점도 행렬이 된다. 이 그림을 통하여 모든 변수사이의 상관계수를 짐작하여 볼 수 있다. 각 산점도에 나타나는 곡선은 평활선이다.



[그림 6.11] 산점도행렬

6.3 분할표 및 모자이크그림

분할표에 대해서는 제 13장에서 자세하게 설명을 할 기회가 주어지겠지만 변수 간의 관계를 나타내는 데 있어 상관계수와 더불어 자주 쓰이는 방법이다. 물론 분할표는 분할되는 변수가 범주형인 경우에 위력을 발휘한다. 예를 들어 설명하여 보자.

예제 6.3 [표 6.3]은 라디오 청취자 2,300명을 분류한 결과이다. 범주형 변수로 쓰인 변수는 교육수준, 나이, 그리고 클래식 음악 청취 여부이다.

	교육수준			
	대졸이상		대졸미만	
	클래식 음악 청취 여부			
나이	예	아니오	예	아니오
40대 이상	210	190	170	730
40대 미만	194	406	110	290

[표 6.3] 분할표

여기서 교육 수준이나 나이에 의해 클래식 음악 청취 여부가 영향을 받는지를 알기 위해 경우에 따라서는 이를 동질적인 집단으로 나누어 자료를 순차적으로 분석을 하여야 한다. 먼저 나이다. 다음은 나이에 따른 분류이다.

40대 이상	40대 미만
56.5%	43.5%

[표 6.4] 나이에 의한 분류

이는 표본이 수집된 형태에 따라 결정되는 표이기도 하다. 나이그룹에 따른 교육 수준의 분할을 보게 되면

40대 이상		40대 미만	
대졸 이상	대졸 미만	대졸 이상	대졸미만
30.8%	69.2%	60.0%	40.0%

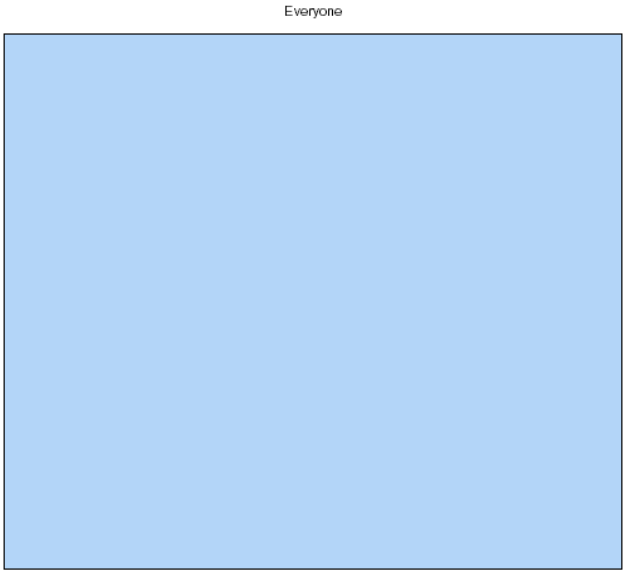
[표 6.5] 나이와 교육수준에 따른 청취자 분류

40대 미만의 표본이 교육수준이 높음을 알 수 있다. 그리고 이 4개의 범주에서 각각 클래식 음악 청취 여부를 확인하여 보면 [표 6.6]과 같다.

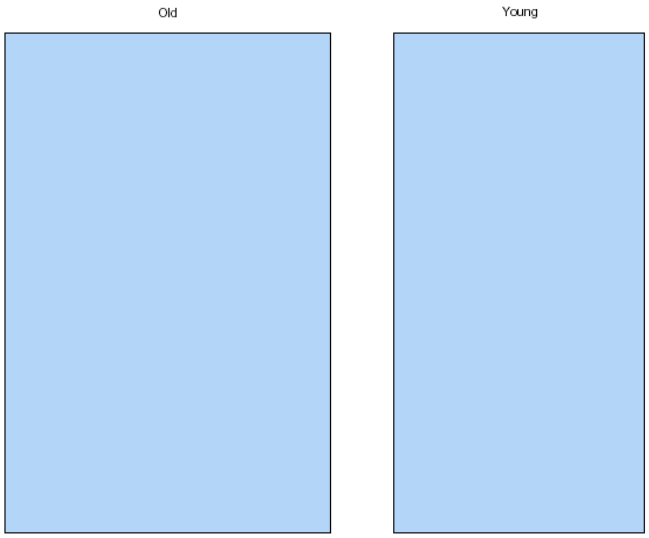
40대 이상		40대 미만	
대졸 이상	대졸 미만	대졸 이상	대졸미만
52.5%	18.9%	32.3%	27.5%

[표 6.6] 나이와 교육수준에 따른 클래식 음악 청취 여부

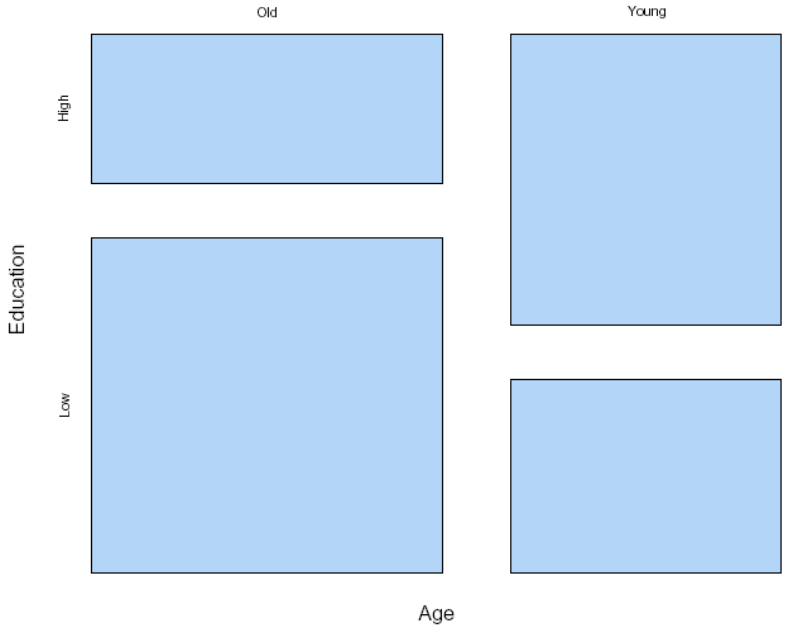
40대 미만의 그룹인 경우는 교육수준에 상관없이 클래식 음악 청취여부는 독립인 현상같이 보이나 40대 이상의 그룹에서는 교육수준이 낮으면 클래식 음악청취 비율이 낮음을 알 수 있다. 이러한 분석에 도움을 주는 그림이 모자이크(mosaic)그림이다. 전체 2,300명을 하나의 정사각형으로 본다면 순차적으로 분석이 됨과 동시에 해당하는 그림을 그려 분석을 보조하는 기법이다. 모자이크그림에서 면적은 빈도수에 비례하여 결정된다. 본 책의 모자이크 그림은 통계패키지 R을 사용하였음을 밝혀둔다. ■



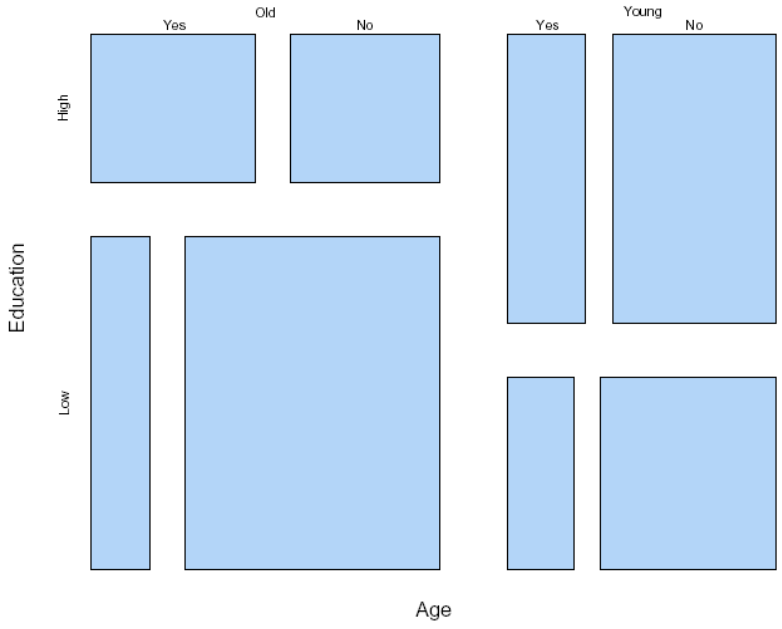
[그림 6.12] 전체 표본



[그림 6.13] 나이에 의한 분류



[그림 6.14] 나이 및 교육수준에 의한 분류



[그림 6.15] 클래식 음악 청취여부

그러나 사실 범주형 자료는 분석하기 쉽지만은 않다. 간혹 분류가 되는 범주형 변수를 고려하지 않는다면 엉뚱한 결과를 얻게 된다. 다음은 이를 위한 예제이다.

예제 6.4 이 예제는 미국 버클리대학의 여자들의 대학원 합격률이 남자에 비해 지나치게 낮다고 지원자 중 한명이 법원에 제소하는 과정에서 밝혀진 자료이다.

성	합격	불합격	합격률
남자	1,193	1,493	44.5%
여자	557	1,278	30.4%

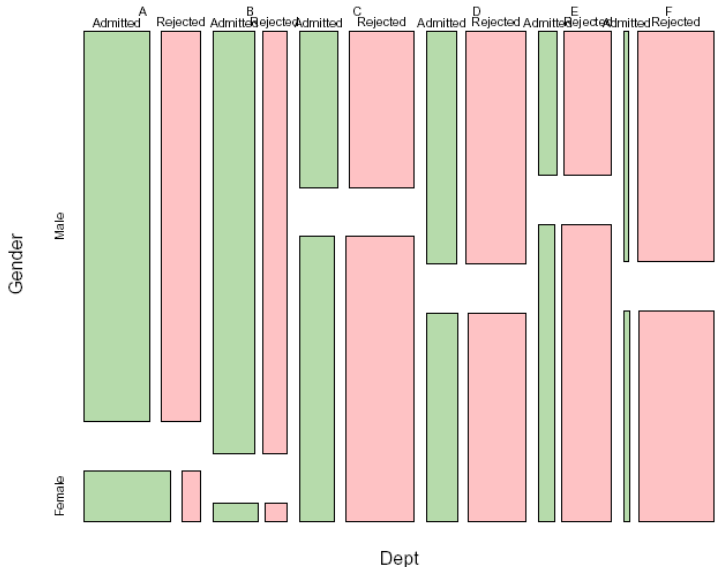
[표 6.7] 미국 버클리대학 합격률

[표 6.7]을 보게 되면 여학생 합격률이 남학생에 비해 현저하게 떨어진다는 사실을 안다. 그러나 이 합격률은 학과의 성격과 관계가 있을 것이다. 당연히 학과라는 범주형 변수를 고려하여 학과별 합격률에 어떤 차이가 있는지 알아 볼 필요가 있다. [표 6.8]이 이를 반영한 결과이다.

		남성	여성
학과 A	합격	512(62.1%)	89(82.4%)
	불합격	313	19
학과 B	합격	353(63.0%)	17(68.0%)
	불합격	207	8
학과 C	합격	120(36.9%)	202(34.1%)
	불합격	205	391
학과 D	합격	138(33.1%)	131(34.9%)
	불합격	279	244
학과 E	합격	53(27.7%)	94(23.9%)
	불합격	138	299
학과 F	합격	22(5.9%)	24(7.0%)
	불합격	351	317

[표 6.8] 학과에 따른 합격률(괄호 안은 합격률)

다음[그림 6.16]은 이를 위한 모자이크그림이다.



[그림 6.16] 모자이크 그림

이 그림을 보게 되면 여학생의 합격률은 남학생에 비해 결코 낮지 않으며 심지어 학과 A인 경우는 오히려 여학생에게 호의적인 사실을 알게 된다. ■

6.4 자료의 변환

지금까지는 자료의 분석은 주어진 단위를 가지고 이루어졌다. 그러나 변수의 단위는 우리가 정해 놓은 단위이지 고집할 필요는 없다. 단위를 바꾸어 변수 간의 관계가 더 잘 설명된다면 변수변환을 하지 않을 이유가 없다.

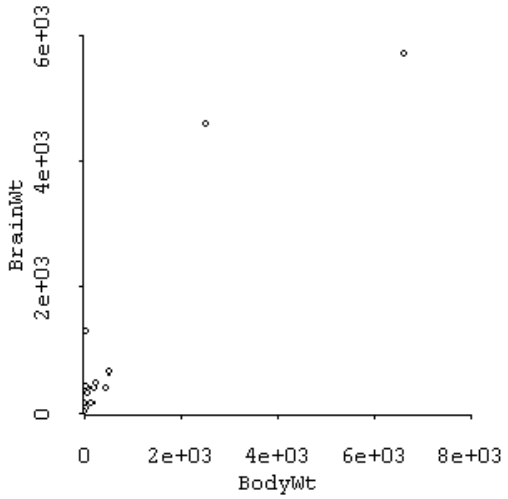
예제 6.5 [표 6.9]는 63종의 포유류 몸무게(킬로그램)와 뇌의 무게(그램)의 자료이다.

<포유류.xls>

	A	B	C
1	이름	몸무게	뇌무게
2	Arctic fox	3.385	44.5
3	Owl monkey	0.48	15.499
4	Beaver	1.35	8.1
5	Cow	464.983	423.012
6	Gray wolf	36.328	119.498
7	Goat	27.66	114.996
8	Roe deer	14.831	98.199
9	Guinea pig	1.04	5.5
10	Vervet	4.19	57.998
11	Chinchilla	0.425	6.4
12	Ground squirrel	0.101	4
13	Arctic ground squirrel	0.92	5.7
14	African giant pouched rat	1	6.6
15	Lesser short-tailed shrew	0.005	0.14
16	Star-nosed mole	0.06	1
17	Nine-banded armadillo	3.5	10.8
18	Tree hyrax	2	12.3
19	N. American opossum	1.7	6.3
20	Asian elephant	2547.07	4603.17
21	Big brown bat	0.023	0.3
22	Donkey	187.092	419.012
23	Horse	521.026	654.977
24	European hedgehog	0.785	3.5
25	Patas monkey	10	114.996
26	Cat	3.3	25.6

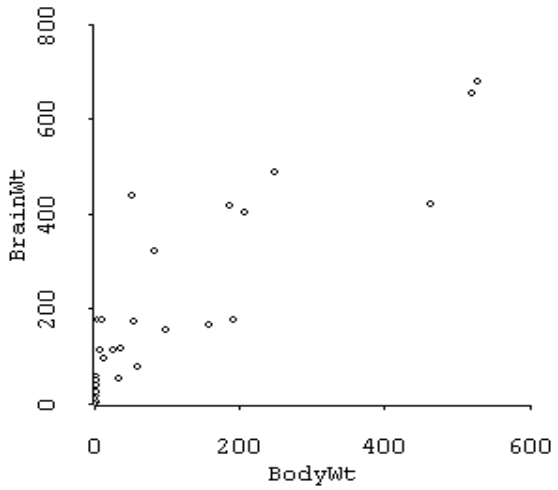
[표 6.9] 포유류 자료

그리고 [그림 6.17]은 두 변수의 산점도이다. 좌측하단에 자료가 너무 몰려있어 그림을 이해하는데 있어 어려움을 겪을 것이다.



[그림 6.17] 산점도

또한 뇌의 무게가 제일 큰 3개의 자료를 삭제하고 [그림 6.18]처럼 그림을 다시 그려도 여전히 두 변수와의 관계를 쉽게 설명하는 그림이 나오지 않는다.

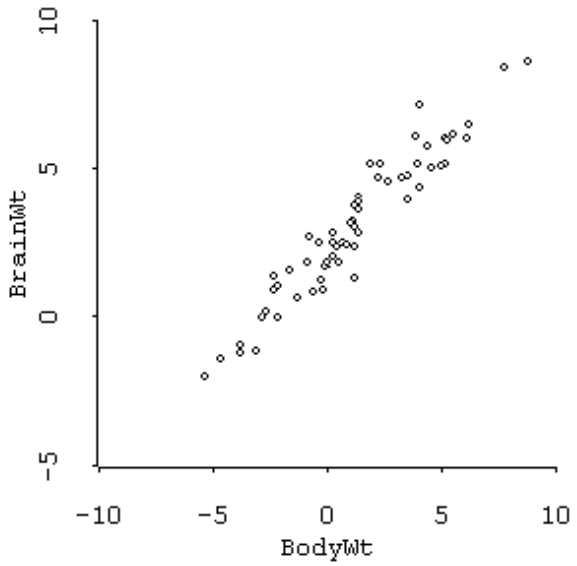


[그림 6.18] 3개의 이상점을 제거한 후 그린 산점도

이러한 이유는 수집된 자료 중 몸무게의 범위가 너무 광범위하게 퍼져 있기 때문이다. 0.01킬로에서 6,654킬로까지의 자료를 킬로그램 단위로 표시하는 것은 바람직하지 않다. 이때 필요한 개념이 변환이다. 변환에는 기본적으로 멱변환(power transformation)이 있다. 변수에 λ 제곱을 하여 변수변환을 하는 형태이다. 여기서 λ를 변환모수라 일컫는다. 예를 들어 λ가 0.5면 제곱근(square root)변환이다. 통상적으로 λ는 -1과 2사이의 범위 내에서 값을 구한다. 그러나 이러한 변환은 척도에 문제가 나타난다. 단위 변환을 하더라도 비교 목적인 경우에는 어느 λ가 더 나은 값인지를 판단하지 못하는 것이다. 이를 위해 만든 것이 척도(scaled) 멱변환이다. 주어진 λ에 대해 척도멱변환은 다음과 같이 정리된다.

$$Y^{\lambda} = (Y^{\lambda} - 1)/\lambda, \quad \text{if} \lambda \neq 0$$
$$Y^{\lambda} = \log(Y) \quad \text{if} \lambda \neq 0$$

단순한 멱변환과 달리 로그변환이 가능하다. 그러나 단순멱변환과 크게 다르지는 않다. 상수 1을 빼고 λ로 나누어 주는 것인데 이로 인해 변수변환에 영향을 주지 않기 때문이다. 다음 [그림 6.19]은 두 변수에 모두 로그변환을 취한 결과이다.



[그림 6.19] 로그변환 후의 산점도

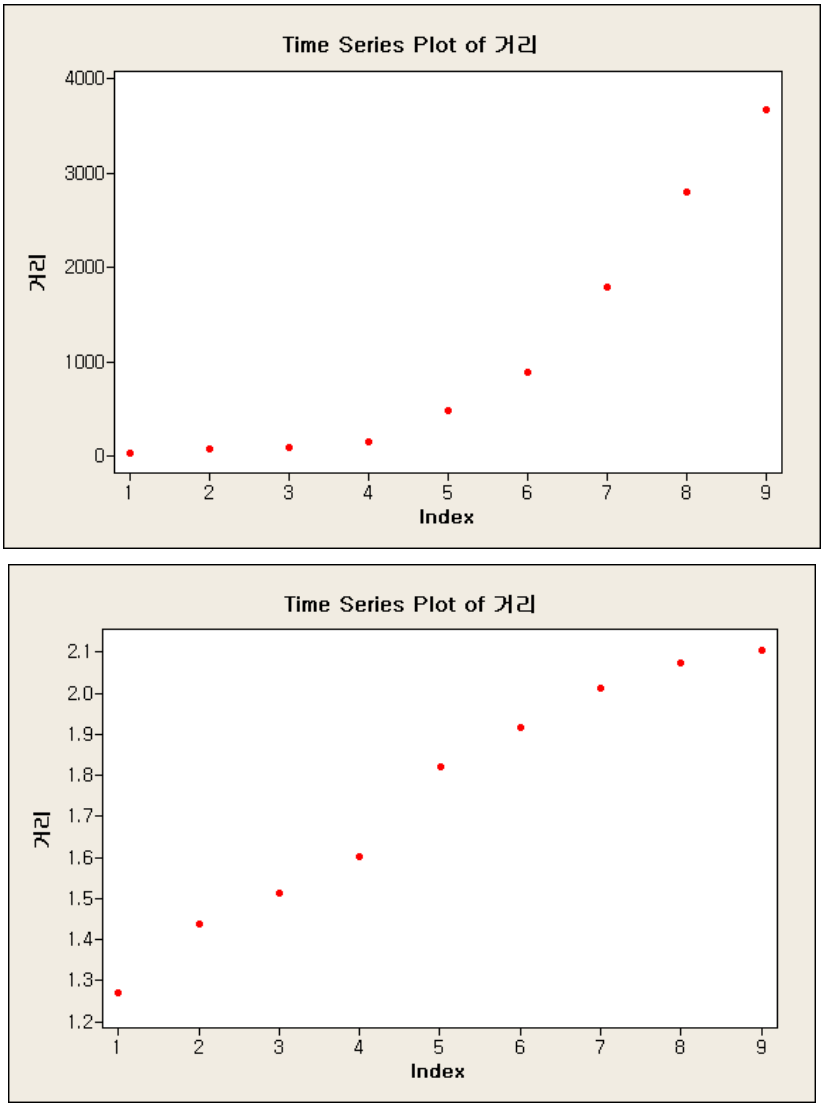
변수 변환을 한 결과 두 변수의 관계는 직선의 관계임이 밝혀진 셈이다. 수집된 단위에 고집하지 않고 단위의 척도를 바꿈으로써 관계를 보다 잘 이해한다면 바람직하지 않은가? 이 경우는 로그단위로 자료를 표현한 것이다. ■

예제 6.6 태양계에서 행성 간의 거리는 등거리가 아니고 태양으로부터 멀리 떨어지면 떨어질수록 길어진다. 다음 [표 6.10]은 태양으로부터 행성까지의 거리표이다.

행성	태양으로부터의 거리(백만 마일)
수성	35
금성	67.1
지구	92.9
화성	141.5
목성	483.4
토성	886.7
천왕성	1,782.7
해왕성	2,794.3
명왕성	3,666.1

[표 6.10] 행성까지의 거리

그리고 다음 [그림 6.20]은 행성을 x-축에서 순서대로 나열했을 때 태양으로 거리를 표시한 그림이다. 보다시피 직선으로 표시되지 않는다. 그러나 거리를 로그변환시킨 후의 그림은 사뭇 다름을 알 수 있다. 로그변환 후의 그림은 직선으로 표시된다. 단위 척도를 변환하는 것은 과학 분야에서는 통상적인 일이다. ■



[그림 6.20] 행성까지의 거리(로그변환 전과 후)

이상의 예와 같이 자료 분석은 수집된 단위에서만 이루어지는 것이 아니고 변환단위에서도 수시로 이루어진다. 어느 지역의 소득 수준을 다른 변수와 연계하여 분석하는 경우에는 소득을 원단위로 분석을 하는 것은 흔치 않다. 로그변환과 같은 적절한 변환을 한 후 분석을 하는 것이 더 바람직하다.

학습요약

제 6장에서는 변수들 간의 관계를 측정하는 몇 가지 방법을 알아보았다. 산점도를 그리고 공분산 및 상관계수를 통해 자료의 연관성을 알아보고 해석상 주의할 점에 대해서도 언급하였다. 또한 분할표를 통해 자료의 표현하여 분할되는 변수 간의 관계를 측정하는 방법 및 이를 도시화하는 방법인 모자이크그림도 역시 언급하였다. 마지막으로 측정된 원래의 단위가 아닌 변환된 단위로 연관성을 표시하게 되면 손쉽게 모형화하는 경우를 알아보았다. 변환된 단위가 분석을 더 용이하게 되는 점을 기억하기 바란다.

6장 연습문제

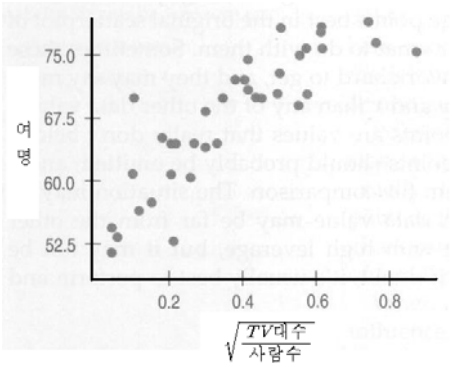
6.1 제 5장에서 나온 [표 5.1]의 변수들 중 봉급변수와 나이변수에 대해 산점도를 그리고 공분산, 상관계수를 구하라. 그리고 나이(범주형 변수), 성별이 환경시설에 대한 의견에 미치는 영향을 모자이크 그림을 통해 분석하여 보아라.

6.2 다음 자료는 도시 근로자 중 화이트 컬러 직장에 다니는 미혼 남성에게 물어본 조사 자료 중 일부를 무작위로 발췌하여 기록한 것이다. 각 변수 간의 산점도를 그리고 상관관계 등을 통해 변수 간의 관계를 설명하라. <소비형태.xls>

	A	B	C	D
1	**100가구 소득 소비형태**			
2				
3	봉급	문화	스포츠	외식
4	54,600	1,020	990	1,510
5	57,500	1,100	460	1,180
6	53,300	900	780	1,590
7	43,500	570	860	1,750
8	57,200	900	1,390	2,120
9	63,400	820	1,880	3,090
10	58,500	1,340	710	1,540
11	55,600	1,250	680	1,800
12	61,300	1,190	1,220	2,330
13	61,100	640	1,480	2,670
14	77,200	900	820	2,850
15	58,800	710	1,080	2,200
16	62,900	1,240	1,230	2,430
17	61,900	1,270	1,000	2,110
18	76,500	1,180	690	1,820
19	50,300	810	1,490	2,100
20	45,900	840	730	920
21	61,900	1,290	1,050	2,480
22	56,700	780	970	1,930
23	43,300	910	1,120	1,720
24	63,000	560	1,570	1,990

6.3 (허위상관) (1) 많은 사람들을 관측하니 대머리일수록 심장발작률이 높다는 사실을 확인하였다. 그렇다면 대머리가 심장발작의 원인이라고 할 수 있나?

(2) 다음 그림은 $\sqrt{\frac{TV대수}{사람수}}$ 와 여명(餘命) 사이의 산점도이다. 양의 상관관계가 강함을 알 수 있다. 그렇다면 $\sqrt{\frac{TV대수}{사람수}}$ 의 크기가 여명을 좌지우지하는가?



6장 실습문제

6.1 각자 자신의 키, 몸무게, 출근시 평균소요시간(분), 퇴근시 평균소요시간(분), 자기 주머니에 갖고 있는 동전의 개수, 자기 지갑에 있는 지폐의 개수, 1개의 동전을 10번 던졌을 때 그림면의 개수 및 첫 번째 그림면이 나왔을 때까지의 시행횟수를 조사한 후 교육원생 모두의 자료를 취합한다.

(1) 8개의 변수들에 대한 산점도행렬을 통계패키지로 구해보고 특징을 이야기해 보라.
(2) 8개의 변수들에 대한 상관계수 행렬을 통계패키지로 구하고 특징을 이야기해 보라.
(3) 1번 문제와 2번 문제를 종합적으로 비교하여 보라.

6.2 다음 자료는 자동차의 속도(시간당 거리)와 연비(리터당 거리)에 대한 자료이다.

속도	60	70	80	90	100
연비	12	14	16	14	12

(1) 속도와 연비에 대한 산점도를 그려보고 상관계수를 구하라.
(2) 결과를 토의해 보라.

6.3 다음 자료에 대한 산점도를 그려라.

x	1	2	3	4	10	10
y	1	3	3	5	1	11

(1) 계산기를 사용하여 상관계수를 구해보라.
(2) 강한 직선적 연관성에도 불구하고 상관계수가 대략 0.5 정도가 되는 이유를 토의해 보라.

쉬어가기

스피어만(Spearman) 상관계수

두 변수 X, Y에 대한 산점도를 검토해 본 결과 X가 Y와 함께 증가하거나 감소하는 추세를 보이지만 그 형태가 직선이 아닌 경우 두 변수의 연관성을 측정할 수 있는 것이 스피어만 상관계수이다. 즉 스피어만 상관계수는 변수 X와 Y의 단조적 연관성(monotonic association)을 측정하는 방법이다. 앞서 살펴본 직선형의 상관은 피어슨 상관계수이다.

스피어만 상관계수는 원자료를 순위자료로 변환하여 상관계수를 구한 것이므로 우리는 두 변수들이 순서형자료로 관측이 된 경우에도 이 방법을 활용할 수 있음을 알 수 있다. 사회과학에서 자주 등장하는 순서형 자료로는 예를 들어 어떤 제품의 선호도를 '매우 좋아한다', '좋아한다', '그저 그렇다', '싫어한다', '매우 싫어한다' 로 구분하여 그 중 하나로 대답된 것이 있을 수도 있다. 이같은 형태를 리커트 측도(Likert scales)라 하며 5가지 선호도는 1부터 5까지 순위로 표시한다. 자료들이 이처럼 순위로 표현되었거나 두 변수 사이에 선형관계성이 의심되는 경우, 스피어만 상관계수는 가장 적합한 상관계수이며, 또한 만족할만한 연관성의 정도를 제공해준다.

먼저 n 쌍의 관측값 $\{(x_1, y_1), \cdots, (x_n, y_n)\}$이 주어졌을 때, x_i 와 y_i 의 순위를 각각

$$R_i = x_1, x_2, \cdots, x_n \text{ 의 순위, } S_i = y_1, y_2, \cdots, y_n \text{ 의 순위, } i = 1, \cdots, n$$

으로 정의하자. 스피어만의 순위 상관계수는 피어슨의 상관계수 r_p 의 정의에서 x_i, y_i 대신에 각각의 순위인 R_i, S_i 를 대입한 통계량으로 다음과 같이 주어진다.

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

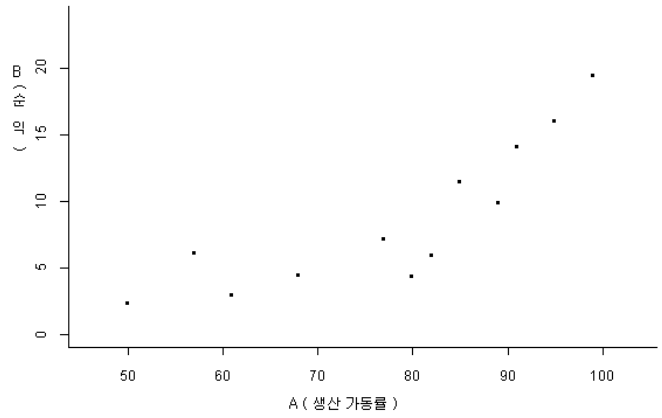
여기서 $\overline{R} = \overline{S} = \frac{(n+1)}{2}$이며 $(R_i, \cdots, R_n)$와 $(S_1, \cdots, S_n)$은 $(1, \cdots, n)$의 임의의 한 순열이다.

[예제] 다음 자료는 어떤 회사의 생산가동률(X)와 순익(Y)을 나타내고 있다.

[생산 가동률과 순익 자료]

X(생산가동률:%)	50	57	61	68	77	80	82	85	89	91	95	99
Y(순익 : 억원)	2.5	6.2	3.1	4.6	7.3	4.5	6.1	11.6	10.0	14.2	16.1	19.5

위 자료를 산점도로 나타내면 다음과 같다.



그림에서 알 수 있듯이 생산가동률과 순익은 서로 곡선형 단조관계를 갖고 있다. 따라서 두 변수의 연관성은 스피어만 상관계수로 측정할 수 있으며, 이를 위하여 원자료를 먼저 순위자료로 다음과 같이 변환한다.

	순 위 자 료											
X(R_i)	1	2	3	4	5	6	7	8	9	10	11	12
Y(S_i)	1	6	2	4	7	3	5	9	8	10	11	12

스피어만 상관계수 r_s는 위의 순위자료를 사용하여 계산되며,

$$r_s = \frac{\sum(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum(R_i - \overline{R})^2}\sqrt{\sum(S_i - \overline{S})^2}} = \frac{125}{\sqrt{143}\sqrt{143}} = 0.874$$

이다. 순익은 생산가동률이 높아감에 따라 증가한다. 두 변수간의 관계는 선형(직선)의 형태는 아니나 곡선형단조증가관계를 보여주며 그 정도는 약 87%임을 알 수 있다.

참고로 두 변수 간의 순위자료들 각각에서 동일순위가 없으면, d_i를 i번째 개체의 Y변수의 순위와 X변수의 순위 차이라 할 때 다음과 같이 간단하게 스피어만 상관계수를 계산할 수도 있다.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

제 3 부

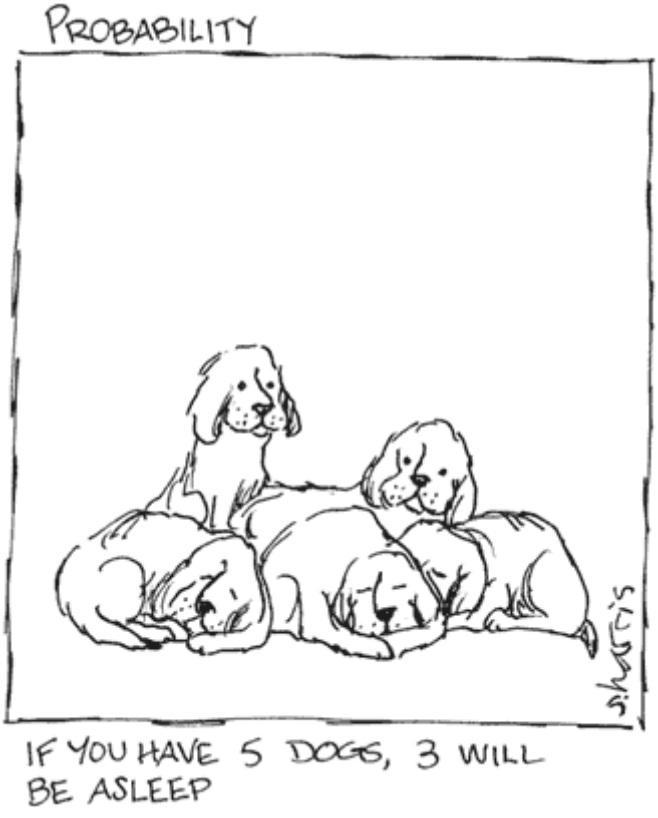
자료의 분석


If I'd known they wanted me to use all this info– I would never have asked for it!
Julie


통계
황금알을 낳는 숫자

제 7 장

통계학에 확률이 필요한 이유가 있다.



차 례

학습목표

확률(probability)이라는 단어는 대부분의 사람들에게는 엄청난 중압감을 준다. 확률과 관련된 공부를 하는 학생들 뿐 아니라 대학을 졸업하고 객관적이고 합리적인 사고체계를 가진 많은 사람들에게도 확률이란 단어는 넘어가기 힘든 산과 같은 존재이다. 그러나 확률은 사실 우리가 매일 접하는 단어 중 하나이다. 등교나 출근을 하기 전 라디오의 일기예보를 들으면 "오늘은 비가 약하게라도 올 확률이 약 40% 정도이다."든지 신문지상에서 "영국 오픈골프대회에서 타이거 우즈가 우승할 확률이 10% 정도 된다."든지 혹은 "월드컵에서 브라질이 우승할 승산(odds)이 4:1" 이라든지 하는 소식을 접한다. 이렇듯 알게 모르게 우리는 확률이라는 단어에 매우 친숙해 있다.

확률이란 단어가 우리 생활에 접근해 있듯이 많은 조직의 의사결정권자 혹은 경영관리자들도 매일 매일 확률의 필요성을 느끼고 있다. 왜냐하면 우리가 의사결정을 내리는데 있어 확률은 매우 중요한 역할을 담당하기 때문이다. 예를 들어 미래의 날씨가 예측되지 못하는 상황 하에서 아이스크림의 적정 제조량을 결정하기는 어렵지 않겠는가? 적정 제조량의 규모를 잘못 판단하게 되면 회사는 매우 큰 손실을 보게 되기 때문에 불확실성 하에서 의사결정은 신중을 기하여야 한다. 따라서 이 관리자는 기상청으로부터 일별, 혹은 월별 날씨정보를 구매한다던지 아니면 객관적인 데이터를 이용하여 좀 더 정확한 날씨에 대한 정보를 알고자 할 것이다.

이러한 불확실성을 다루기 위해서는 확률이 필요하다. 그러나 그 필요성은 알고 있지만 실제로 확률에 관련된 모형을 구축하거나 확률적인 사고를 정확하게 한다는 것은 저자들도 마찬가지이지만 인간이 가지고 있는 능력의 한계로 인해 많은 어려움을 겪는다.

본 장에서는 확률, 확률변수, 확률분포의 개념과 더불어 확률을 중심으로 응용 예제를 다룰 것이다. 그러나 확률규칙(rules of probability)은 쉽게 이해하기가 어려우므로 이러한 확률규칙 중에서 제일 핵심적인 내용만 추려 살펴보도록 하고 실제 확률적인 사고가 객관적인 시스템을 정립하는데 얼마나 도움을 주는지를 확인하여 보고자 한다.

7.1 확률을 가지고 무엇을 할 것인가?

14면체 주사위를 가지고 신라인은
술 문화를 발전시켰다? 안압지에서 출토.

- 확률은 어느 사건이 발생할 가능성을 측정하는 0과 1사이의 값이다.

어느 사건이 일어날 확률이 0이라는 이야기는 사건은 일어날 수 없다는 의미이고 확률이 1인 사건은 반드시 이 사건은 일어난다는 의미이다. 0보다는 크고 1보다 작은 확률을 가지는 사건은 불확실성을 내포하고 있다. 1에 가까운 확률을 가지고 있는 사건은 일어날 가능성이 많다는 뜻이다.

확률은 실험을 수많이 반복하여 이러한 사건이 몇 번이나 일어났는지 그 비율을 계산하여 구하여야 한다. 다음 예제들을 보고 확률이 계산되는 과정과 왜 확률이 의사결정의 기준이 되는지 보자.

- 조달청을 통하여 군부대에 납품되는 방독면 중에서 하나의 방독면을 선별하여 이상 유무를 가리는 행위는 실험이라 할 수 있다. 이러한 실험은 두 가지의 불확실성을 가지고 있는데 어느 특정한 부품이 선별되느냐 하는 것과 선별된 부품의 품질이 그것이다. 만약 1,000개의 방독면을 검사하여 990개가 양품의 판정을 받았다면 이 업체가 납품하는 방독면은 양품이 될 가능성이 99% 확률을 가지고 있다고 볼 수 있다. 조달청에서는 99% 이상의 양품이 나온 제품을 납품한 업체를 우수한 업체로 선정할 것이고, 양품 비율이 99% 미만의 업체는 불량한 업체로 판정을 하여 추후 납품업체 선정에서 탈락시킬 수 있는 기준을 마련할 수도 있을 것이다. 따라서 의사결정의 기준은 확률인 것이다.

다른 예제를 들어 보자.

- 인천에서 로스앤젤레스로 취항하는 항공사의 정시착륙비율이 80%라면 이 80%의 의미는 지

난 수많은 착륙기록을 검토한 결과 80%는 정시에 착륙을 하였다는 의미이다. 이러한 확률은 의사결정의 기준이며 목표가 설정될 수 있는 근거가 된다.

- 300개의 도자기로 만든 반도체(ceramic insulator) 중 294개가 열 충격(thermal shock)을 견디어 내었다면 어느 특정의 반도체가 열 충격(thermal shock)을 이겨낼 확률은 294/300 = 98%인 것이다. 이러한 확률은 우리가 기준으로 하고 있는 목표에 근접한 것인가? 만약 목표에 못 미친다고 판단된다면 공정관리에 더욱 더 신경을 많이 써야 한다.

많은 경우 우리가 흔히 생활에서 접하는 확률은 이러한 확률로서 **장기적 비율(long-run proportion)**에 의해 계산된 것이다.

이왕 말이 나온 김에 확률에 대해 구분을 하고 넘어가 보도록 하자. 일반적으로 확률은 두 가지 형태로 구분된다. **객관적(objective)**인 확률과 **주관적(subjective)**인 확률이다. 여기서 객관적인 확률은 위와 같이 장기적인 비율로 계산이 될 수 있는 확률을 의미한다. 물론 객관적인 확률이라 하더라도 모두 다 이렇게 구할 필요는 없다. 우리가 고등학교 교과서에 배운 동전 던지기 실험이 그것이다. 양쪽 면이 일정하게 균형을 이룬 동전을 던졌을 때 앞면이 나오는 확률을 만약 장기적인 도수관점에서 구한다면 동전을 무수히 많이 던져 앞면이 나오는 횟수를 이용해 비율로 구해야 하는데 그럴 이유는 없다. 왜냐하면 누구나 앞면이나 뒷면이 나올 가능성은 같다고 믿고 있기 때문에 확률은 50%라고 대답을 하면 된다. 물론 우리가 알고 있는 100원짜리 동전보다 옆면이 더 두꺼운 동전을 던졌을 때는 앞면과 뒷면뿐 아니라 옆면도 나올 수 있기 때문에 이 경우는 객관적 확률을 장기적인 비율에 의해 구해야 한다. 교과서에서 언급한 동전은 앞면과 뒷면만 있는 이상적인 동전을 두고 하는 말이다.

- 참고로 100원 짜리 동전 몇 개를 겹쳐 옆면의 길이가 앞면(혹은 뒷면)의 지름과 같은 길이가 되게끔 테이프로 붙인 다음 500회 정도 타일 바닥에 던져 보는 실험을 실시하게 되면 대략적으로 1/3 가량이 앞면, 1/3 가량이 뒷면, 그리고 나머지 1/3 가량이 옆면이 나올 것이다. 이는 저자가 경험한 지극히 경험적이고 장기적인 비율에 의한 확률 계산이다. 그렇다면 초기화면에 나온 14면체 주사위에서 1거3잔(一去三盞)이라는 면이 나와 술을 3배하는 확률은 얼마인지 독자들은 궁금하지 않은가?

그러나 이러한 객관적인 확률은 많은 공·사 조직에서 접하는 확률과는 거리가 있다. 장기적인 비율이란 개념을 도입하지 못하는 그런 의사결정 환경이 많기 때문이다.

- 어느 지방 자치단체가 사업을 의뢰한 기업에 대한 평가를 투자전문가에게 맡겼다고 하자. 투자전문가가 특정의 기업에 대해 평가를 할 때 “이 기업은 앞으로 10년간 시장에서 퇴출당할 가능성이 매우 높다”라고 한다면 이 투자 전문가가 이야기 하는 확률은 주관적인 확률이다. 왜냐하면 장기적인 비율로 확률을 구할 수 없기 때문이다. 기업의 퇴출은 단 일회라도 반복되지 않을 것이기 때문이다.

주관적 확률은 어떤 한 사람의 특정 사건이 일어날 가능성에 대한 평가(assessment) 혹은 의견(opinion)으로 규정지을 수 있다. 지자체의 회의석상에서 논의될 많은 확률은 바로 이러한 주관적 확률이다. 여기서 의견은 모든 주위의 정보를 동원한 제일 이성적인 값을 의미한다. 따라서 평가를 하는 사람에 따라 이 의견은 다를 수 있다. 평가하는 사람에 따라 확률은 같은 사건에 대해서도 평가자가 가지고 있는 감정, 정보 등에 의해 달라질 수 있다. 주관적이라는 단어를 쓰는 이유이다.

비가 올 주관적 확률은 얼마나 되는가?

주관적인 확률은 엄격한 의미에서 단 한번에 일어나는 사건에 적용된다. 그러나 대부분의 상황은 유일하게 결정되지는 않고 경험적인 직관을 많이 요구한다.

- 어느 회사가 신상품을 개발한다고 했을 때 다른 제품과는 다른 면도 있지만 이미 시판되는 기존제품과 공통점도 있을 것이다. 신상품의 유일한 특징도 중요하지만 유사제품을 판매했을 때 과거실적 등도 참고를 하여야 한다. 만약 유사제품의 성공률이 40%이였다면 이 정보는 신상품의 제품 성공률을 평가하는데 있어 출발점이 될 수 있다.

이 책에서 사용되는 대부분의 확률은 주관적인 확률에 가깝다. 여기서 확률은 일종의 경험에 의한 추측 값(educated guess)이라 보면 된다. 주관적 확률의 특성상 경우에 따라서는 민감도 분석이 필요할 때도 있을 것이다. 불확실한 상황 하에서의 의사결정은 이러한 도구들을 필요로 한다.

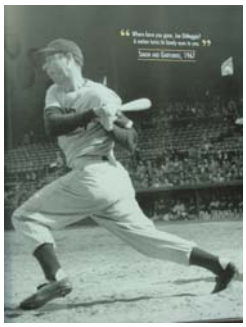
7.2 사건들의 독립은 무엇인가?

대학수학능력 시험에 제일 많이 출제되는 문제 중의 하나가 확률적 독립(probabilistic independence)의 개념을 물어 보는 문제이다. 독립은 그만큼 중요하다고 판단되는 개념으로 많은 응용가능성을 가지고 있다.

사건 A와 사건 B에 대해 이야기하여 보자.

- 두 사건이 독립이라는 의미는 사건 A(B)가 일어나는 확률을 계산하는데 있어 사건 B(A)가 일어났는지 여부는 아무런 도움을 주지 않는다는 뜻이다.
  ※ A(B)는 사건 B가 일어났을때 사건 A를 의미한다.

사건 A와 사건 B가 독립이라는 사실은 사건 A와 사건 B가 동시에 일어나는 확률은 개개의 확률로 곱을 하여 구할 수 있다는 의미이다.



확률로 계산한 결과 나는 피트 로즈를 좋아한다?

예제 7.1 피트 로즈와 조 디마지오는 어느 기록이 더 위대한가?

유명한 야구선수 중에 조 디마지오와 피트 로즈란 야구선수가 있었다. 조 디마지오는 55경기 그리고 피트 로즈는 44경기 연속 안타 행진을 벌인 바 있다. 누가 위대한가? 당연히 조디마지오가 평균 타율도 높고 피트 로즈보다 11경기 더 많은 경기에서 안타를 쳐 냈기 때문에 더 위대하다고 볼 수 있다.

만약 피트 로즈가 3할 타자라면 한 경기에 4번 타석에 들어온다면 한 경기에서 안타를 쳐낼 확률은

$$1-(.7)^4=0.76$$

이다. 여기서 안타가 나오는 사건은 서로 독립을 가정하였다. 한번에 안타를 치지 못할 확률이 0.7이기 때문에 4타석 모두 안타를 치지 못할 확률을 구하고 이를 1에서 빼면 적어도 하나의 안타를 한 경기에서 쳐낼 확률이 나온다. 따라서 44경기 연속해서 기록을 이어나갈 확률은

$$(.76)^{44}=0.000005666091173$$

로 매우 낮다. 물론 여기서도 독립을 가정한 것이다. 조 디마지오의 평균 타율은 피트 로즈보다 약간 높다. 만약 타율이 0.35라면 55게임 연속 안타의 확률은

$$((1-(0.65)^4)^{55} = 0.000020101779264$$

로 나온다. 따라서 피트 로즈의 기록이 더 위대하지는 않지만 더 어려웠을 것이다.

확률을 공부한 신문기자는 "조 디마지오의 기록은 더 위대하지만 피트 로즈의 기록은 더 어려운 것이다." 라고 기사를 써야 하지 않는가? ■

여기서 중요한 것은 사건들이 독립이냐 하는 것이다. 물론 독립이 아니어도 문제의 큰 틀을 유지한다면 별 문제가 없겠지만 말이다. 그러나 불행히도 수학적으로 '사건이 독립이다 아니다'라고 판정할 수가 없는 경우가 많다.

독립이라고 가정해도 무방한지 여부를 판단하기 위해서는 경험적 자료가 필요하다. 예를 들어보자.

- 복지부에서 남아 선호도를 조사하는 문제가 대두가 되었다. 사건 A를 첫 번째 아이가 남자일 사건이라 하고 사건 B를 두 번째 아이가 남자일 사건이라 하자. 두 사건은 독립이라고 이야기할 수 있을까?

남자아이를 낳고 연속해서 남자아이를 낳을 가능성이 두 번째에 여자아이를 낳을 가능성보다 높다고 믿고 있다면 '이 두 사건은 독립이 아니다'라고 답할 것이다. 그렇지 않고 첫 번째 아이가 남자아이 건 여자아이 건 상관없이 두 번째 아이가 남자일 가능성은 마찬가지로 같다고 말한다면 두 사건은 독립이 된다. 여기서 여자아이와 남자아이일 가능성이 같다는 뜻은 아니다. 수학적인 논리로는 이를 해결할 수는 없다.

이 논란을 잠재울 유일한 방법은 자료를 수집하는 방법이다. 적어도 2명의 아이를 가지고 있는 많은 가족을 조사하여야 한다. 만약 첫 번째 아이가 남자아이인 경우의 모든 가족들 중에서 55%가 두 번째 아이도 남자로 나오고 첫 번째 아이가 여자아이를 둔 가정들 중에서 45%가 두 번째 아이가 남자아이라면 두 사건은 독립이 아니라고 할 만한 확실한 증거를 확보하는 셈이 된다.

자 이제 야구 세계를 확률로 본격적으로 이해하여 보자.

예제 7.2 **야구의 7전 4선승제는 확률적인 규칙이다.**

사건들의 독립성을 가정하지 못하는 상황이라 하더라도 편의상 독립성을 가정하는 경우가 많다. 스포츠에서 나오는 문제를 통해 이해를 구해보도록 하자. 많은 나라의 프로야구에서 진정한 챔피언을 가리기 위해 7전 4선승 제도를 도입하고 있다. 즉 경기를 하여 4번 먼저 이기는 팀이 챔피언으로 결정되는 것이다. 야구를 좋아하는 많은 야구팬들이 알겠지만 시리즈는 4경기 만에 싱겁게 끝이 나는 경우도 있고 혈전을 치룬 다음 7경기에서 시리즈가 종료되는 경우도 있다.

여기서 확률의 의미를 생각하여 보자.

Red Sox' Comeback Lands Them in World Series

야구경기는 확률게임이다.

- 2004년 미국의 보스톤 레드삭스 야구팀 대 센트루이스 카디널스 야구팀이 월드 시리즈를 하기 전에 야구전문가들 20명에게 누가 우승할 것인가를 물어보았는데 11명은 보스톤 레드삭스를, 다른 9명은 센트루이스 카디널스가 우승한다고 하였다. 두 팀의 실력이 비슷하다고 본다면 한 경기를 하여 한 팀이 이길 확률은 동전을 던져 앞면이 나오는 확률과 같다. 즉, 1/2의 값을 배정하여도 아무도 이의를 달지 않겠지만 이 경우에는 보스톤 레드삭스가 이길 확률은 11명/20명 = 0.55가 되는 것이다. 전문가들은 보스톤 레드삭스가 아무래도 센트루이스 카디널스에 비해 강팀이라고 의견을 제시한 것이다.

그럼 **확률적인 사고를 가진 스포츠 신문기자**들은 어떤 식의 기사를 써야 할 것인가? 다음 경기의 결과가 이전 경기결과에 상관없이 결정지어 진다면, 즉 경기결과는 독립적으로 결정된다고 가정한다면 보스톤 레드삭스가 한 경기를 이길 확률은 0.55이다. 두 번 연속해서 이길 확률은 $0.55 \times 0.55 = 0.3025$가 된다. 이런 식으로 계산을 하면 4경기 연속해서 이길 확률은 0.09150이 된다. 반면 센트루이스 카디널스 역시 4경기 만에 시리즈를 이길 확률은 0.041006이다. 보스톤 레드삭스가 이기면 당연한 결과이고 센트루이스 카디널스가 이기면 예상이 빗나갔다고 사람들은 이야기 할 것이다.

이런 식의 유사한 계산을 하게 되면 [표 7.1]의 결과를 얻게 된다.

팀	성적	확률
보스톤 레드삭스	4-0	9.2%
보스톤 레드삭스	4-1	16.5%
보스톤 레드삭스	4-2	18.5%
보스톤 레드삭스	4-3	16.7%
센트루이스 카디널스	4-0	4.1%
센트루이스 카디널스	4-1	9%
센트루이스 카디널스	4-2	12.4%
센트루이스 카디널스	4-3	13.6%

[표 7.1] 야구경기 예상결과

먼저 더 좋은 실력을 가지고 있는 팀으로 평가된 보스톤 레드삭스가 시리즈를 이길 확률은 60.9% 이며 약한 팀으로 평가된 센트루이스 카디널스가 이길 확률은 39.1% 가 된다. 또한 이를 이용하여 4경기, 5경기, 6경기, 혹은 7경기에 시리즈가 종료될 확률을 구해보도록 하자. 먼저 4 경기 만에 종료될 확률은 9.2% + 4.1%인 약 13.3%가 된다. 비슷한 방법으로 합산하면 [표 7.2]와 같은 결과를 얻는다.

경기 수	확률
4경기	13.3%
5경기	25.5%
6경기	30.9%
7경기	30.3%

[표 7.2] 총 경기 수에 관한 예상 결과

마지막으로 평균적으로 끝나는 경기횟수를 구한다면 $4\times 0.228+5\times 0.289+6\times 0.2754+7\times 0.208=5.46$경기가 된다. 이러한 평균에 대한 계산은 후에 다시 한번 언급할 것이다. ■

이 시점에서 우리가 생각하여야 할 질문은 이러한 경기방식보다 더 나은 경기방식은 존재하지 않는가? 하는 것이다.

진정한 강자를 가리는 방법은 스포츠 경기마다 방식이 다르다. 그러나 강자를 가리는 방법에는 근본적으로 두 가지 원칙이 있다.

- 진정한 강자를 가려야 하며
- 다른 하나는 빠른 시간 안에 승부를 보아야 한다.

위의 야구의 경우 7전 4 선승제가 아니라 101경기 51선승제를 한다면 확실하게 실력이 더 있는 팀이 우승할 확률은 거의 1에 가깝게 높아진다. 물론 이러한 경기 방법은 불가능하다. 따

라서 제한된 시간 내에 진정한 강자를 가리는 방법을 찾아야 한다.

확률적인 사고는 이러한 경기규칙이 만들어 졌는지를 이해하는데 도움을 줄 뿐 아니라 새로운 경기방식을 만드는데도 유용하다. 잠시 야구 문제는 접어 두고 테니스의 듀스규칙을 이해하고 넘어가 보도록 하자.

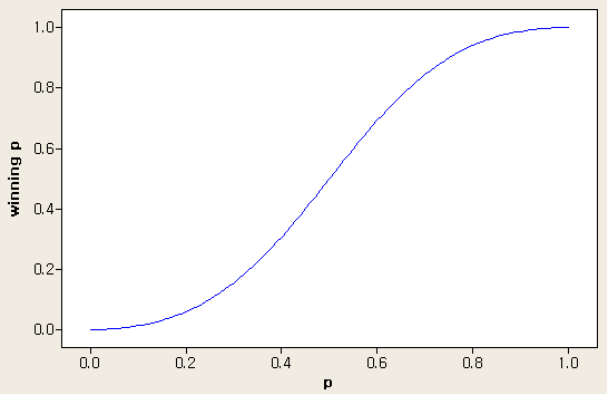
예제 7.3 **테니스의 규칙에도 확률이 들어가 있다.**

규칙을 알면 경기는 재미있다.

테니스 경기에서 한 선수가 서비스 권한을 가지고 있을 때 한 플레이를 이길 확률이 1/2이라면 듀스에서 그 경기를 이길 확률은 당연히 1/2이어야만 공정할 것이다. 그러나 1/2이 아닌 이보다 높은 0.55인 경우는 게임을 이길 확률은 0.55보다 더 높아진다. 이 확률은 두 번 연속 이길 확률에다가 두 경기 연속 질 확률을 더한 다음 두 번 연속 이길 확률이 차지하고 있는 비율을 구하면 된다. 왜냐하면 1:1이 되면 다시 원점에서 출발하기 때문이다. 즉,

$$0.55\times 0.55/(0.55\times 0.55+0.45\times 0.45)=0.59901$$

이 값이 듀스에서 게임을 이길 확률이 된다. 이를 [그림 7.1]에서 일반화 시켰다. x-축은 한 플레이를 이길 확률이고 y-축은 듀스를 이겨 게임을 따내는 확률이다. p가 0.5보다 크면 아래로 볼록한 함수이고 반대방향은 위로 볼록한 함수가 된다.



[그림 7.1] 듀스를 이길 확률

통상적으로 남자선수들이 서비스 권한을 가지고 있을 때 한 플레이를 이길 확률은 0.55 보다 훨씬 높다. 통상적으로 .70에 육박한다. 그렇다면 타이브레이크를 이길 확률은 0.844828로 껑충 뛴다. 그렇다고 해서 3게임차의 점수를 필요로 하는 타입의 타이브레이크 룰은 별로 재미가 없을 것이다. 시간이 길어질 뿐 아니라 서비스 권한을 가지고 있는 선수에게 절대적으로 너무 유리한 규칙이기 때문이다. 따라서 듀스에서의 2점 차의 규칙이 많은 관중에게 그래도 흥미를 유발하는 것이다.

서비스 권한을 가지고 있는 경우 점수를 얻을 성공확률은 탑 랭킹을 가지고 있는 선수인 경우는 매우 높다. 따라서 게임 스코어 6:6인 경우는 번갈아 가면서 경기를 하여야 이전 타이브레이크 규칙에서는 어느 한편이 게임을 이기는데(적어도 2점 차로 이겨야 함) 걸리는 시간은 매우 길어질 수가 있다. 스웨덴의 보리스와 미국의 맥켄로의 경기는 장장 6시간 30분이나 걸린 경기를 하였다. 따라서 테니스 협회는 경기 시간을 줄이기 위해 무한정 연장되었던 게임 브레이크를 마지막 세트를 제외한 나머지 세트에서는 선취 7점으로 묶여져 있는 규칙을 도입한 것이다. 물론 마지막 세트에서는 2경기 차로 마무리가 되게끔 하는 타이브레이크 규칙이 아직 존재한다. 진정한 강자가 누구인지를 보여주기 위함이다. 경기시간도 줄이고 진정한 승자도 가리는 경기 규칙을 국제 테니스 협회가 고안해 낸 것이다. ■

예제 7.4 **배구 규칙은 바뀌었다.**

알다시피 예전의 배구 규칙은 서비스 권한을 가지고 있을 때 공격이 성공이 되어야만 점수가 올라가지 않았는가? 그러나 이러한 경기방식은 경기에 소요되는 시간이 길어지는 단점이 있었던 관계로 세계배구연맹은 이러한 폐단을 극복하기 위해 서비스 권한에 상관없이 공격이 성공을 하면 점수가 올라가는 방식으로 바꾸어 새로운 경기규칙을 만든 것이다. 이런 달라진 규칙 하에서도 경기력이 우수한 팀이 이길 확률이 예전의 방식 하에서의 확률과 비슷하고 시간이 많이 단축이 된다면 당연히 관중들은 현재의 방식에 수긍을 할 것이다. 탁구 경기 또한 경기 규칙을 바꾸어 진정한 강자를 가리면서 경기시간을 단축시켜주는 규칙을 마련하여 시행하고 있지 않은가? ■

배구는 지난 규칙 때보다 재미있어졌다.

예제 7.2(계속) **새로운 월드시리즈 규칙을 만들어 보자.**

다시 야구 경기로 돌아 가보도록 하자. 현 7전 4 선승제 하에서 보스톤 레드삭스가 한 게임을 이길 확률을 .55 라고 하였을 때 시리즈를 이길 확률이 60.9%라 하였다. 그리고 평균 게임수는 5.78게임이다. 그럼 이보다 더 좋은 규칙은 만들 수는 없을까? ■

사고의 전환이 필요하다.

만약 시리즈가 3-0, 4-1, 4-2, 5-3, 5-4일 때 시리즈가 종료된다면 이러한 시리즈가 종래 우리가 쓰고 있는 7전 4선승제보다 좀 더 나은 시리즈 규칙이 아닐까 한번 생각해 보자.

수학자/통계학자들은 야구 규칙을 만드는 사람은 아니지만 이런 질문에 대한 명석한 답을 일정한 가정 하에서 할 수 있다. 독립성을 가정한다면 보스톤 레드삭스가 챔피언이 될 확률을 계산하여 보면 61.4%가 나오며, 평균 게임 수는 5.72가 나온다. 계산 방식은 위에서 했던 것과 유사하게 진행하면 된다. (물론 독자들은 이 문제의 계산에 대해 신경 쓰지 않아도 된다.) 확률도 높아지고 게임수도 줄어들고 하였지만 수치적으로 별로 차이가 나지 않는다. 새로운 방식이 월등히 좋아 보이지는 않는다. 그러나 한 게임을 이길 확률이 .60 혹은 .65로 상승하였을 경우에는 새로운 방식이 훨씬 좋다는 사실을 입증할 수 있다.

보스톤 레드삭스가 한 게임을 이길 확률이 .65이라면 굳이 3경기 혹은 4경기 이상 볼 이유가 있을까 한다. 실력차이가 난다면 일찍 시리즈를 종료할 필요도 생기는 것이다. 그러나 왜 야구는 다른 경기와 달리 우승자를 가리는 새로운 방식을 채택하지 않는 것일까? 여기에는 많은 이유가 있다.

제일 큰 이유는 전통일 것이다. 야구에서 전통은 무시하지 못하는 덕목이기 때문이다. 다른 이유는 아주 뛰어난 에이스 2명을 가진 팀이 우승할 확률은 매우 높아지며 시리즈가 8경기, 9

경기 째까지 간다면 투수진이 풍부한 팀이 우승할 가능성이 높아지기 때문이 아닐까 한다. 그러나 이미 플레이오프 시스템을 몇 번이나 바꾼 다른 경기, 예를 들면 아이스하키와 같은 경기에는 이런 룰이 적용될 법도 하지 않은가?

확률적 사고가 기준이 된다.

이상으로 확률을 이용하여 새로운 시스템을 개발하는 문제를 스포츠 분야에서 찾아보았다. 우리나라 야구에서 진정한 승자를 가리는 규칙이 어떤 방식이 제일 좋은가 하는 문제는 아마 한국야구위원회의 고유한 임무라고 본다. 뿐만 아니라 새로운 경기방식을 추구하는 스포츠의 많은 경기 운영자에게는 이러한 확률적 사고가 많은 도움을 주지 않을까 생각하여 본다.

예제 7.5 **서울시의 랜드마크는 무엇인가? -유머 한마디.**

어느 시골에 사는 노인 두 분이 전화로 내일 서울에서 보자는 약속을 하는데 그만 전화가 끊어져버렸다. 그럼에도 불구하고 이 두 분이 서울에 상경했다면 서로 만날 확률이 얼마나 되는가? 만약 이 두 노인 분이 프랑스에서 살고 있는 분이고 파리에 왔다면 그 확률은 달라지는가? 뉴욕은 어떠한가? 에펠탑이라는 유명한 랜드마크가 있는 도시에서는 이러한 확률이 서울보다 높을 가능성이 많다. 확률적인 사고를 랜드마크와 연결하여 우스개 유머를 하고 넘어 간다. 서울에 랜드마크가 있느냐라는 질문을 이렇게 우회하여 확률문제로 질문을 던져도 된다. ■

에펠탑에서 만나면 확률이 올라간다.

예제 7.6 **확률은 선거 전략에도 응용된다.**

미국 선거는 우리나라와 달리 한 주에서 승리하면 주가 가지고 있는 선거인단 투표권을 모두 가져가기 때문에 선거가 박빙인 경우에는 어느 주에 가서 마지막 유세를 하여야 하는지가 매우 중요한 이슈로 떠오른다. 미국 대통령인 부시가 후보시절 케리 후보와 선거전을 치루고 있을 때를 기억해 보자. 참고로 이 예제는 가상으로 꾸며 보았다. 케리 후보가 펜실베이니아 주를 이긴다면 부시는 오하이오 주와 플로리다 주를 반드시 이겨야 선거에서 승리할 수 있다.

부시가 오하이오 주에서 승리 할 확률은 30%, 그리고 플로리다 주를 이길 확률은 70%이다. 선거 전 단 한번의 유세 기회만 주어져 있다면 부시는 (혹은 케리 후보는) 어느 주를 가서 유세를 펼쳐야 하는가?

누구든 오하이오 주에서 마지막 선거 유세를 한다.

어느 후보라도 특정 주를 방문하게 되면 주에서 이길 확률은 10% 상승하고 방문하지 않으면 주에서 이길 확률은 10% 하강한다. 만약 두 후보가 같은 주를 방문하는 경우 확률은 변하지 않는다. 이럴 경우 부시가 두 주 모두에서 이길 확률은 0.3 × 0.7 = 0.21(21%)가 된다. 그러나 부시가 오하이오 주를 방문하고 케리 후보가 플로리다 주를 방문한다면 확률은 0.4 × 0.6 = 0.24(24%)가 되고 만약 부시가 플로리다 주를 방문하고 캐리 후보가 오하이오 주를 방문한다면 부시 후보가 두 주 모두 이길 확률은 0.2 × 0.8 = 0.16(16%)가 된다. 만약 여러분이 부시 후보의 선거 참모라면 어느 주에 가서 마지막 유세를 하라고 권고하겠는가? 케리 후보가 어느 지역을 방문하든지 부시 후보는 오하이오 주를 방문하여 마지막 선거 유세를 하여 두 주에서 이길 확률을 높여야 한다. 케리 후보도 마찬가지이다. 참고로 이런 전략을 배우는 학문 분야가 게임이론이다. ■

7.3 사건은 명확하게 명시하여야 한다.

알고자 하는 사건에 대한 정의를 명확하게 하지 않아 혼돈을 주는 수가 많다. 사실 이러한 문제는 확률을 구하는 문제보다 더 중요하다. 원하는 사건에 대한 정의가 명확하지 않는 이상 확률은 그 의미가 없기 때문이다. 예를 들어 보자.

예제 7.7 **복권번호가 같아도 그렇게 우연치는 않다.**

매사추세츠 주의 복권 당첨번호와 뉴햄프셔 주의 복권 번호는 4자리 숫자로 발표가 되는데 1987년 10월 10일 무작위로 뽑은 숫자가 우연치 않게 7923이란 숫자가 동시에 두 지역에서

나온 것이다. 이를 두고 온갖 미디어 업체에서 아주 희귀한 일이 벌어졌다고 이 이야기를 특종으로 다룬 적이 있다.

우리가 다루고자 하는 문제는 여기서 말하는 희귀한 일은 무엇을 의미하는가? 이다.

발표되는 숫자는 4자리 숫자이기 때문에 7923이란 복권 당첨 숫자가 나올 가능성은 1/10을 4번 곱하여 구한 1/10,000이다. 그리고 두 주의 복권은 독립적으로 판매되었고 숫자 역시 독립적으로 무작위로 뽑은 숫자를 발표하기 때문에 이러한 희귀한 사건이 나올 가능성은 1/100,000,000이 된다. 이는 대략적으로 일주일에 한번 복권 당첨을 발표한다면 이는 평균적으로 40만년 만에 한번 일어나는 사건이 된다.

그러나 이러한 확률은 7923이 아닌 다른 숫자라도 확률은 다를 바 없다. 8323과 같은 다른 숫자라도 확률은 1/100,000,000 이다.

- 과연 우리는 이러한 확률에 관심을 가지고 있는 것일까? 우리가 원하는 것은 두 복권의 번호가 같을 사건에 대한 확률이 아닌가? 그런데 이러한 경우는 숫자가 4자리이기 때문에 0=0부터 9,999=9,999까지 총 10,000번 일어나지 않는가? 따라서 확률은 10,000을 100,000,000으로 나누어야 한다.

$$10,000/100,000,000 = 1/10,000$$

이러한 사건은 일주일에 한 번씩 복권 당첨이 열린다면 40만년이 아니라 평균적으로 불과 40년 만에 나오는 사건인 것이다. ■

이와 같이 일반인들이 생각하는 확률과 통계학자가 계산한 확률에는 차이가 있을 수 있다. 물론 정확히 알고자 하는 사건이 무엇인지 정의를 하는 것은 어려운 작업이다.

위의 예는 우리가 구하고자 하는 사건에 대한 정의를 명확히 하고자 하는 차원에서 살펴보았는데 다음 절에서 소개할 조건부 확률도 일반인들이 많이 혼돈스러워 하는 개념 중의 하나이다.

7.4 조건부 확률은 무엇인가?

확률의 값은 현재 주어진 정보에 의해 결정이 된다. 그러나 새로운 정보가 도착하면 확률은 통상적으로 바뀌게 된다. 2007년 일본 요미우리 야구단은 센트럴리그에서 우승을 하였다. 2007년 8월말 선두를 달리고 있었다. 그런데 이승엽 선수가 부상으로 팀 전력에서 2007년 연말까지 이탈했다면 요미우리 야구단이 리그우승을 차지할 확률은 바뀌었을 것이다. 이렇듯 새로운 정보 하에서 확률을 개정하는 작업이 **조건부 확률(conditional probability)**의 개념이다.

예제 7.8 **조건부 확률은 무엇인가?**

어느 기업이 조달청에 제품을 공급한다고 보자. 현재 조달청에 제품을 7월말까지 공급하여야 하는 계약을 체결 중인데 7월말까지 공급이 가능할지는 불확실하다. 이는 다른 회사로부터 7월 중순까지 원자재를 공급받을 수 있는 가의 여부에 달려있기 때문이다. 현재는 7월 1일이다.

사건 A를 7월말까지 제품을 공급하는 사건, 그리고 사건 B를 7월 중순까지 원자재를 공급받는 사건이라 하자. 그리고 A^c, B^c는 각각 사건 A와 B가 일어나지 않는 사건을 의미한다.

사건을 말로 풀이하는 것보다는 기호 A, B로 표시하는 것이 설명을 하는데 편할 뿐 독자들의 어려움을 가중시키기 위해 만들지는 않았다.

현재로는 원자재를 공급받을 가능성은 3번 중에서 두 번 꼴이라고 판단된다. 따라서 사건 B가 일어날 확률은

$$P(B) = 2/3$$

이다.

공급을 받는다면 제품을 7월말까지 공급을 해 줄 가능성은 3/4라고 판단된다. 여기서 조건부의 개념이 들어간다. 이는 다음과 같이 수선을 사이에 두고 확률을 표시한다,

$$P(A|B) = 3/4$$

그렇다면 이 회사가 원자재를 공급받아 부품을 공급할 가능성은 어떻게 계산되는가? 이는 두 확률을 서로 곱하면 된다. 이러한 확률은 P(A and B) 로 표시한다.

$$P(A \text{ and } B) = P(A|B)P(B) = (3/4)(2/3) = 0.5$$

가 된다. 즉, 확률은 반반이다.

다른 몇 가지 흥미를 끌만한 사건을 더 살펴보도록 하자. 먼저 원자재를 7월 중순까지 받지 못할 확률이다.

$$P(B^c) = 1 - P(B) = 1/3$$

그리고 $P(A|B^C)$ 은 만약 원자재를 공급받지 못하면 제품을 공급하는 확률로서 1/5로 가정하자. 마찬 가지 요령으로 원자재가 제때 도착하지 않았지만 제품을 공급할 가능성을 구할 수 있다.

$$P(A \text{ and } B^c) = P(A|B^c)P(B^c) = (1/5)(1/3) = 0.0677$$

- 여기서 우리가 궁극적으로 원하는 사건은 A이다.

7월 1일 현재는 사건 B나 B^C중 어느 사건이 일어날지 모른다. 7월 15일이 지나면 사건 B에 대한 정보가 주어지므로 사건 A가 일어나는 확률은 둘 중의 하나가 된다.

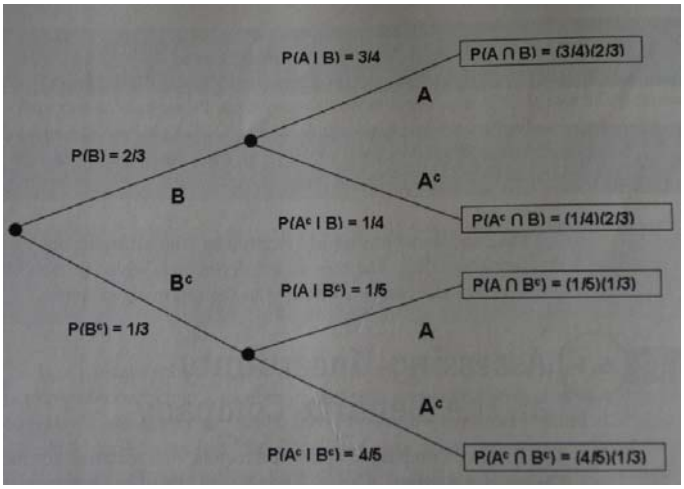
$$P(A|B) = 3/4 \text{나} \quad P(A|B^c) = 1/5 \text{이다.}$$

다행스러운 점은 어느 사건이 아직 일어날지는 모르지만 7월 15일이 지나면 사건 A에 대한 확률은 이미 구해진 정보에 의해 구할 수 있다는데 있다.

왜냐하면 사건 A가 일어나는 확률은 $A|B$ 와 $A|B^c$ 의 확률의 합으로 구할 수 있기 때문이다. 즉 사건 A가 일어난다면 사건 B 혹은 사건 B^c 와 같이 일어나야 하기 때문이다.

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^c) = 1/2 + 1/15 = 17/30 = 0.5667$$

7월 1일을 기준으로 우리가 가지고 있는 정보에 의하면 7월말까지 부품을 성공리에 조달청에 납품할 가능성은 30번 중에서 17번 정도이다. 이런 계산 과정을 [그림 7.2]와 같이 그려보면 좀 더 확연하게 알 수 있다.



[그림 7.2] 나무그림을 이용한 확률계산

조건부 확률은 7.2절에서 언급한 확률적 독립과 밀접한 개념이다. 사건 A의 확률은 사건 B가 일어났느냐 일어나지 않았느냐에 따라 다르다. ■

- **조건부확률과 독립성** 위의 예제에서 보았듯이 $P(A)$, $P(A|B)$, $P(A|B^C)$는 모두 다르게 나타났다. 그러나 이 세 가지 확률이 모두 다 같은 경우가 있는데 이럴 경우 사건 A와 사건 B는 독립이 되는 것이다.

예제 7.9 **청소년 흡연은 매우 위험하다.**

통계를 알면 광고 문안이 보인다.

- 청소년들에게 흡연의 위험성을 알리기 위해 청소년기에 흡연을 한 사람들은 대략적으로 3명 중의 한 명은 결국 성인이 되어서 흡연으로 인한 질병에 걸릴 위험이 있다고 텔레비전 매체에서 공익광고를 한다고 하자. 많은 일반인들은 이런 값이 어떤 의미이고 어떻게 계산되어 나왔는지 궁금해 할 것이다.

조사에 의하면 흡연을 한 청소년이라 하더라도 성인이 되어서 흡연습관을 버린 성인인 경우 흡연으로 인한 질병이 걸릴 확률이 10%이고 그렇지 못하고 흡연습관을 계속 가지고 있는 성인인 경우는 흡연으로 인한 질병에 걸릴 확률이 50%에 달한다고 의학계에서 보고된 바 있다고 하자.

청소년 때부터 담배를 계속 피는 사건을 A로 하고 청소년기 이후에 금연을 한 사건을 A의 여사건, A^c 로 하고 청소년기에 흡연습관이 있는 사람이 흡연으로 인한 질병에 걸릴 사건이 C 라 한다면 위에서 언급한 확률은 조건부 확률로 쉽게 이해가 되지 않는가?

즉, $P(C|A)$ 가 50% 이며 $P(C|A^c)$ 가 10%가 되는 것이다. 따라서 청소년기에 흡연을 할 경우 흡연으로 인한 질병에 걸릴 확률은 청소년기 후에 얼마나 금연을 하였느냐 하는 비율을 조사해서 구하면 된다.

$P(A)=.55$ 그리고 $P(A^c)$= .45 라고 조사가 되었다면 $P(C)$ 는

$$P(C|A)P(A)+P(C|A^c)P(A^c)= .50 \times .55 + .10 \times .45 = 0.320$$

이 된다.

약 3명 중의 한 명이 된다. 다만 이 문제에 있어서 문구가 **"3명 중 한 명"**이라는 문구를 삽입하는 것이 **"청소년기에 흡연을 하였다 하더라도 성인이 되어 금연을 하는 경우 질병에 걸릴 확률이 줄어든다."**는 문구를 쓰는 것보다 나을 것이다. 청소년들은 혹 "청소년기에 담배를 피웠다 하더라도 성인이 되어서 금연을 하면 흡연 관련 질병에 걸리지 않을 것이다." 라는 착각을 할 가능성이 매우 높기 때문이다. 그리고 3명 중 한 명은 매우 자극적인 문구이기 때문이다. 왜냐하면 흡연을 하는 친구들 3명은 얼마든지 찾을 수가 있기 때문에 자기도 포함될 수 있다는 생각이 들게 하는 것이다. ■

7.5 확률변수는 무엇이고 어떠한 모습을 가지고 있는가?

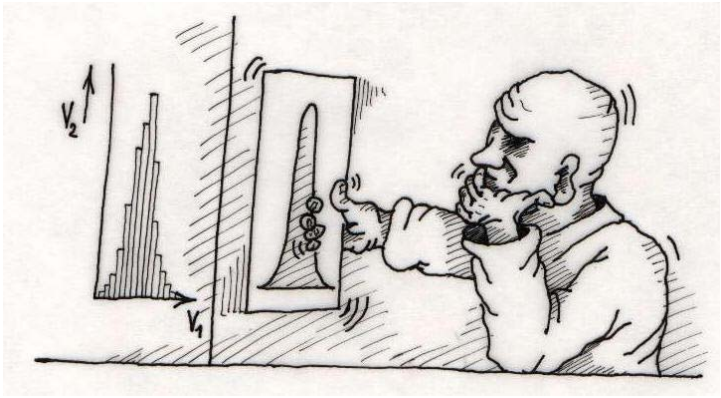
우리의 분석 대상이 되는 모든 개체들은 매번 다른 값을 가진다. 상수로 일컬어지는 대상들은 거의 없다고 해도 과언이 아니다.

이런 개념이 확률변수인데 확률변수(random variable 혹은 줄여서 변수, variable)는 두 가지 종류, 이산형과 연속형이 있다고 이미 언급하였다. 정해진 이외의 다른 값은 갖지 못하는 경우가 이산형이고 연속체의 구간에서 임의의 값도 가질 수 있는 특징을 가지고 있는 변수가 연속형이라 하였다.

그러나 이산형이라 하더라도 편의상 연속형으로 처리하는 경우가 있다.

- 예를 들어 PDP텔레비전과 같은 상품에 대한 수요는 기본적으로 이산형이다. 왜냐하면 PDP텔레비전이란 물건의 수요는 정수로 표시되어야 하기 때문이다. 그러나 수요라는 확률변수가 가질 수 있는 값이 2,000에서 6,000 사이의 모든 값이기 때문에 확률변수는 연속형으로 처리하는 것이 훨씬 바람직하다. 왜냐하면 2,000개부터 6,000개까지 모든 값을 나열하고 해당하는 확률을 부여한다는 것은 매우 거추장스러운 일이 되기 때문이다.

물론 문제의 단순성을 기한다면, 자료를 2,000개 미만, 3,000개 미만, 4,000개 미만, 5,000개 미만 그리고 5,000개 이상 5개의 범주로 나누어 변수를 단순한 구조의 이산형 형태로 만들어 볼 수도 있다. 자료의 성격을 정확하게 파악하고 분석의 편의성을 위해 자료를 가공하는 방법은 분석하는 사람들의 취향에 따라 다를 수 있지만 문제는 분석과정에서 가공으로 인한 정보의 손실이 없어야 하는 것이다.



연속형으로 자료를 처리하면 편하다.

위의 경우 임의적으로 범주의 수를 정한다던지 혹은 범주의 크기를 설정한다면 정보의 왜곡은 피할 수 없을 것이다. 따라서 이 경우에는 연속형으로 분석을 하는 것이 더 좋을 수가 있다.

- **그러나 변수가 연속형인 경우는 이산형으로 처리 할 이유는 없다.**

수학적으로는 이산형과 연속형 변수를 처리하는데 있어 차이가 있다. 이산형 변수를 다룰 때와 그 개념은 비슷하지만 연속형을 다루기 위해서는 미적분이 필요하다. 그러나 독자들은 걱정을 하지 않아도 된다. 왜냐하면 이산형 변수에 대한 개념만 가지고 있으면 연속형 변수를 이해하는데 충분하기 때문이다.

지금부터 이산형 변수와 변수의 생김생김을 알려주는 확률분포에 대해 알아보자. 이산형 확률변수 X가 있다고 보자. 참고로 확률변수는 X, Y, Z (알파벳의 마지막 문자들을 주로 이용)들을 이용하여 표기하는 것이 관례이다.

- 확률변수 X가 어떤 모양을 하고 있는지를 알기 위해서는 X가 가질 수 있는 값과 그에 해당하는 확률을 명기하여야 한다.

X가 가질 수 있는 k개의 값을 $x_1, x_2, \cdots, x_k$ 라 한다면 X가 특정한 값 x_i를 가지는 확률은 $P(X = x_i)$ 혹은 $p(x_i)$라 표기한다. 물론 확률은 0보다 커야 하고 합이 1이 되어야 한다.

- 모든 확률변수가 취하는 값들과 그와 관련된 확률의 나열을 **X의 확률분포**라 한다.

이러한 방법은 확률변수의 구조를 완벽하게 명시하지만 때때로 누적확률(cumulative probability)을 계산하는 것도 중요하다. 누적확률이란 확률변수가 어떤 특정한 값보다 작거나 같을 확률이다.

• 예를 들어 우체국에서 운용하는 특정 인터넷 사이트에서 물건을 주문하였을 때 물건이 인도될 때까지 걸리는 시간을 확률변수 X라 하자. 확률변수 X가 가질 수 있는 값이 1, 2, 3, 4일이고 확률은 다 같이 25%로 알려져 있다고 가정한다면 $P(X \leq 3)$과 같은 누적확률, 즉 물건을 주문했을 때 적어도 3일만에는 물건이 인도될 확률은 다음과 같이 계산된다.

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.25 + 0.25 + 0.25 = 0.75$$

확률분포와 누적확률분포로 확률변수를 표현하면 [표 7.3]과 같다.

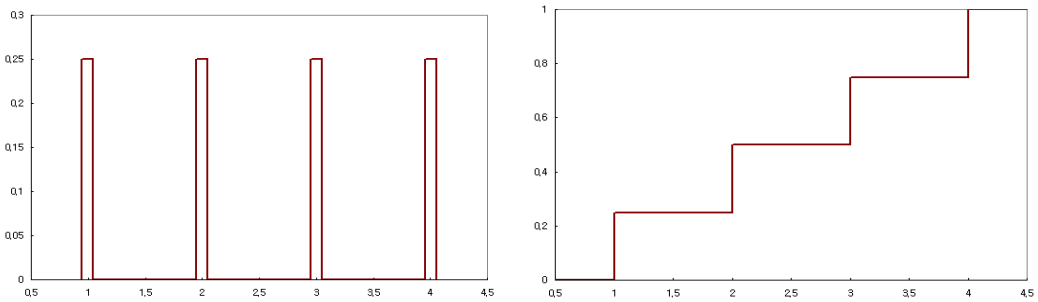
X(주문 인도기간) = x	확률분포 $(P(X = x))$	누적확률분포 $(P(X \leq x))$
1	0.25	0.25
2	0.25	0.50
3	0.25	0.75
4	0.25	1.00

[표 7.3] 확률분포 및 누적확률분포

누적확률분포 표만 주어진다 하더라도 우리가 원하는 특정의 확률 $P(X = x)$ 를 구할 수 있기 때문에 확률분포의 특징을 완벽하게 명기할 수 있다.

예를 들면 $P(X = 3) = P(X \leq 3) - P(X \leq 2)$ 로 구할 수 있지 않은가?

[그림 7.3]은 이러한 확률분포 표와 누적확률 표를 그림으로 표시한 것이다. 이들을 각각 확률분포그림 그리고 누적확률분포그림이라 부르면 된다.



[그림 7.3] 확률분포그림 및 누적확률분포그림

7.6 기댓값과 분산, 표준편차 등은 왜 필요한가?

복권은 매우 정직한 게임이다.

예제 7.10 **기댓값으로 복권의 가격을 책정하여 보자.**

여러분은 로토 복권을 산 적이 있는가? 미국에서는 복권을 산 사람들은 1부터 44까지의 숫자 7개를 임의로 선택하여 숫자가 다 일치하면 이런 숫자를 선택한 다른 사람들과 상금액을 나누어 가지는 방식이다. 현재까지 누적액은 280백만 달러이다. 그리고 지금까지 통계를 보면 다음과 같다.

한명의 당첨자가 나오지 않을 확률이 10%, 1명인 나올 확률이 21%, 2명, 26%, 3명, 21%, 4명 13%, 5명이 9%다.

그렇다면 당첨자는 평균적으로 총 걸려 있는 상금액의 몇 %(혹은 280백만 달러 X%)를 가지고 가는가? 그리고 복권에 당첨될 가능성을 p라 한다면 이 금액에 p를 곱해 복권 가치를 판단할 수 있지 않을까? ■

이제 기댓값에 대한 내용을 자세히 살펴보자. 확률변수의 값과 해당하는 확률을 나열함으로서 확률변수의 특징을 보고하는 것도 하나의 방법이겠지만 2~3개의 잘 선택된 숫자로 확률분포를 요약하는 것도 한가지 방법이다.

확률변수는 편의상 알파벳 문자 X를 선택한다.

● 그 첫째가 평균(mean)이다. 평균은 기댓값(expected value)이라고도 불리며 $\mathrm{E(X)}$로도 표기한다.

공·사 조직에서 작성하는 일반 보고서에서는 수학적인 용어를 쓰지 말고 평균이라는 단어를 써도 전혀 문제가 없다. 여기서는 단순히 수학적인 편의성 때문에 평균을 희랍문자로 표현하는 것이니 독자들은 두려워하지 말라.

평균은 기술적으로 이야기하면 확률로 가중치가 부여된 모든 가능한 값의 가중합(weighted sum)이다. 절차는 다음과 같다.

(1) 모든 x_i 에 $p(x_i)$ 를 곱하여 둔다.
(2) 이러한 $x_i \cdot p(x_i)$ 를 다 더하면 평균이 된다.

우리가 잘 알고 있는 평균이 드디어 식 (7.1)과 같이 탄생된 것이다.

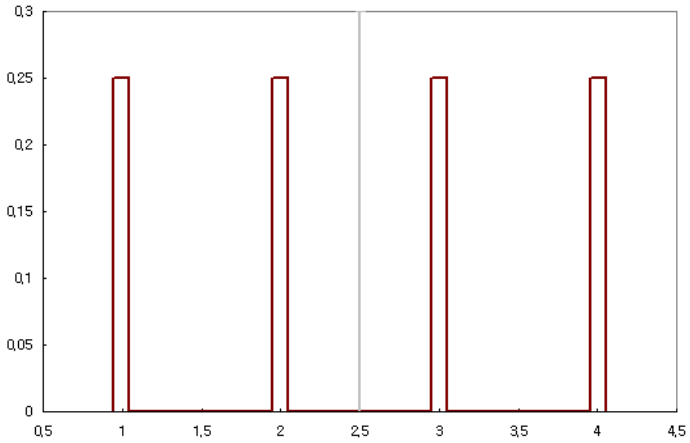
$$E(X) = \sum_{i=1}^{k} x_i \cdot p(x_i) \tag{7.1}$$

제 5장에서 표본 값들의 평균이 중심위치(central location)를 암시하였듯이 여기서 기댓값은 확률분포의 중심(center)을 의미한다.

● 예를 들면 위의 주문인도기간의 기댓값은 확률분포 그림에서 **엄지손가락을 갖다 대면 좌우로 평행을 이루는 점에 해당하는 값**을 의미한다. [그림 7.4]에서는 이러한 값이 2.5이다. 계산에 의해 그 평균을 구하면 다음과 같다.

x	P(X=x)	xP(X=x)
1	0.25	0.25
2	0.25	0.5
3	0.25	0.75
4	0.25	1
	평균	2.5

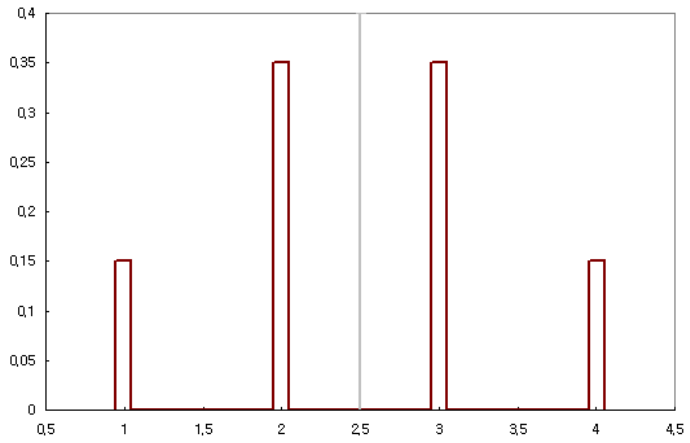
[표 7.4] 평균계산



[그림 7.4] 기댓값 1

- 평균은 확률분포를 잘 대표하여 말할 수 있는 한 숫자인 것만큼은 틀림없으나 평균 하나만 가지고는 분포를 잘 설명할 수가 없다.

  다음 확률분포 그림을 보도록 하자. [그림 7.5]는



[그림 7.5] 기댓값 2

위의 [그림 7.4]와 기댓값은 같지만 주문인도기간은 2일 혹은 3일 만에 이루어지는 확률이

$$P(X=2)+P(X=3)=0,35+0.35=0.70$$

70%를 차지한다.

• [그림 7.4]는 주문인도 기일에 부여되는 확률을 다 같다고 하였지만 [그림 7.5]에서는 2~3일 만에 주문이 인도되는 확률이 더 많다. 첫 번째 그림은 평균을 중심으로 일어날 수 있는 확률 값들이 흐트러져 있는 반면 두 번째 그림에서는 평균을 중심으로 확률 값들이 모여져 있는 모양새이기 때문에 이들의 차이점을 나타내 보고자 만든 것이 **분산**이며 **표준편차**이다.

확률분포의 변동(variability) 정도를 알기 위해 분산과 표준편차를 구해보자. 이에 따른 의미는 후에 언급하도록 한다. 먼저 분산을 구하기 위해서는 다음 절차를 따라야 한다.

(1) 확률변수 값에서 평균을 뺀다. 어떤 값은 음의 부호를, 어떤 값은 양의 부호를 가질 것이다. 평균과 정확하게 일치한다면 물론 0이다. 수학적인 표기로는 $x_i - E(X)$ 이고 이를 **편차**라 한다.

(2) 이러한 편차를 제곱을 한 후 해당하는 확률을 곱해준다. $(x_i - E(X))^2 P(x_i)$ 이다.

(3) 그리고 이를 모두 더하면 분산, $\sum_{i=1}^{k}(x_i - E(X))^2 p(x_i)$이 된다. 통계학에서 분산은 σ_X^2로 표기된다. 식 (7.2)이다.

$$\sigma_X^2 = \sum_{i=1}^{k}(x_i - E(X))^2 p(x_i) \qquad (7.2)$$

• 주의 깊은 독자들은 알겠지만 분산, $\sum_{i=1}^{k}(x_i - E(X))^2 p(x_i)$은 편차의 제곱을 확률이라는 가중치로 구한 일종의 평균의 공식이 된다. 즉, 편차의 제곱이 얼마나 큰지를 평균적으로 알아보는 개념이 분산이다.

• 제 5장에서도 언급되었듯이 분산은 단위가 원래 단위의 제곱이기 때문에 불편하다. 따라서 분산에 제곱근을 취한 표준편차를 사용하게 되면 원래의 단위로 돌아가기 때문에 편의상 분산보다는 표준편차를 많이 애용한다. 표준편차는 σ_X로 표기한다. 이상과 같은 분산과 표준편차를 주문인도기간 예제로 계산하여 보자. 아래 [표 7.5]와 [표 7.6] 계산에 의하면 두 번째 그림의 표준편차가 첫 번째 그림의 표준편차보다 약 0.2만큼 작게 나온다.

(1) [그림 7.4]

x	$P(X=x)$	$xP(X=x$	$(x-E(X)$	$(X-E(X))^2$	$(X-E(X))^2P(X=x)$
1	0.25	0.25	-1.5	2.25	0.5625
2	0.25	0.5	-0.5	0.25	0.0625
3	0.25	0.75	0.5	0.25	0.0625
4	0.25	1	1.5	2.25	0.5625
	$E(X)$	2.5		분산	1.25
				표준편차	1.118033989

[표 7.5] [그림 7.4] 확률분포에서 분산 및 표준편차 구하기

(2) [그림 7.5]

x	$P(X=x)$	$xP(X=x)$	$(x-E(X))$	$(X-E(X))^2$	$(X-E(X))^2\ P(X=x)$
1	0.15	0.15	-1.5	2.25	0.3375
2	0.35	0.7	-0.5	0.25	0.0875
3	0.35	1.05	0.5	0.25	0.0875
4	0.15	0.6	1.5	2.25	0.3375
	$E(X)$	2.5		분산	0.85
				표준편차	0.921954446

[표 7.6] [그림 7.5] 확률분포에서 분산 및 표준편차 구하기

- 처음 이런 개념을 접하는 독자들은 매우 혼돈스러울 것이다. 0.2의 차이로 이 두 그림의 변동을 구분한다는 이야기인가?

그러나 표준편차를 구할 때 평균에서 얼마나 떨어져 있냐를 측정하는 문제였기 때문에 평균값에 비해 편차의 크기가 어느 정도인지 이야기할 수 있다. 평균은 다 같이 2.5이기 때문에 두 번째 그림의 표준편차는 평균을 기준으로 하여 100% × 0.2/2.5 = 8% 정도의 편차가 작아졌다고 이야기할 수 있다.

주의: 기댓값 2.5일에 대한 의미는 2.5일의 주문인도기간이 제일 가능성이 있는 값이라는 의미도 아니고 인터넷으로 물건을 주문을 했을 때 물건이 2.5일 만에 온다고 기대하는 값도 아니다. 실제로도 2.5일 주문인도기간은 존재하지 않는다. 기댓값은 반복적으로 인터넷으로 물건을 주문했을 때 상황이 변하지 않는다면 주문인도기간은 평균은 2.5에 가까워지고 표준편차는 0.92일에 가까워진다는 의미이다.

7.7 기댓값은 주의를 하여야 한다.

평균으로 잰 위치로 이 취객의 생사를 판단한다면 이 취객은 살아남는다.

하나의 확률변수의 특징을 나타내 주는 첫 번째 숫자는 단연 평균이다. 그러나 평균은 많은 사람들에게 혼란을 야기하는 숫자이다. 분포의 특징을 평균 하나만 가지고는 설명하지 못한다는 단순한 의미만은 아니다. 의사결정을 하다 보면 평균을 많이 이용할 수밖에 없는데 이는 매우 위험한 수단이 될 수 있다는 점에서 기댓값이 가질 수 있는 오류에 대해 설명을 하고자 한다.

첫 번째 예제는 흥미를 유발하는 수학적인 문제이나 시사하는 바가 많아 포함시키니 독자들은 주의 깊게 보기 바란다.

예제 7.11 **형과 동생은 차별적으로 대우 받는다. 과학적으로 이야기 해보자.**

- 길이가 1인 엿가락이 있다고 보자. 엿가락을 임의로 반 토막을 낸 다음 항상 작은 토막은 동생이 가지고 큰 토막은 형이 먹는다고 가정하자. 그러면 동생이 먹는 토막은 기껏 해봐야 0.5를 넘어가지 못한다. 반면 토막이 발생되는 시점이 0에 가까운 값이라면 동생은 엿가락을 먹지 못할 것이다. 이런 관점에서 보면 동생이 먹는 엿가락의 길이로 정의되는 변수 X는 0과 0.5사이에서 결정되는 확률변수가 된다. 기댓값은 0.25가 될 것이다. 형이 가지는 엿가락의 길이는 당연히 0.75가 된다. 그렇다면 동생의 엿가락과 형의 엿가락의 길이의 비는 기댓값이 얼마나 되는지 답해보자. <기댓값.xls>

결론부터 이야기하면 길이의 비는 0.25/0.75 =1/3이 아니다. 왜냐하면 형이 가지는 엿가락의 길이가 길어지면 (짧아지면) 동생이 가지는 엿가락은 짧아지기(길어지기)때문이다. 수학적으

로 기댓값을 구하면 약 0.383이 나온다. 물론 이 과정은 독자들은 이해를 할 필요는 없다. 그러나 이 문제는 시사하는 바가 매우 많다. 다음과 같이 요약된다.

• 기댓값들을 대입하여 새로운 변수의 기댓값을 구하는 시도는 절대로 금물이다. ■

예제 7.12 **왜 수익성은 기대한 값보다 떨어지는가?**

• 어느 지역기업에서 수익성을 위한 사업을 하기 위해 새로운 지방 특산물을 개발하여 출하하려고 한다. 그러나 새로운 신상품이기 때문에 팀 미팅을 연 결과 다음과 같은 결론을 얻었다. 숫자는 계산의 단순성을 위하여 가공하였다. <수익성.xls>

(1) 매출개수는 평균적으로 8천개가 팔릴 것이다.
(2) 개당 평균판매가는 80원이다.
(3) 개당 평균 변동생산비용은 75원이다.
(4) 평균 고정비용은 3만원이 들어간다.

그렇다면 이러한 상황 하에서 평균을 가지고만 신상품을 팔았을 때 발생하는 이익의 기댓값은 얼마나 될까? 간단한 이익의 공식에 평균을 대입하여 값을 구한다면 다음과 같다.

$$\text{이익} = 8{,}000 \times (80-75) - 30{,}000 = 10{,}000\text{원}$$

이 상품을 팔았을 때 이익은 기댓값을 대입하여 구한 값 10,000원이라는 이익을 낼 가능성은 물론 없다. 주문인도기간의 예제에서 보았듯이 기댓값은 실제로 구현되어서 나오지 않는 값일 가능성이 매우 높기 때문이다.

이 경우도 마찬가지다. 가정 (2)에서 개당 평균 판매가가 80원이라 정한 것은 매출이 평균보다 많을 경우 다른 지자체에서 유사한 제품을 출하할 가능성이 높기 때문에 가격이 70원으로 떨어지고 그렇지 않다면 90원으로 가격을 유지한다는 의미였다면 80원은 일어나지 않는 숫자이다. 가격이 70원으로 떨어지면 이익이 음으로 돌아서고 90이면 이익은 양으로 나온다. 과연 기댓값을 집어넣고 사업성 분석을 하여야 하는가? [표 7.7]을 참조하라.

	베이스 모형	판매가 변동 :+10	판매가 변동 :-10
매출	8,000	8,000	8,000
가격	80	90	70
고정비	30,000	30,000	30,000
변동비	75	75	75
이익	10,000	90,000	-70,000

[표 7.7] 사업성분석

또한 변동비용과 매출은 독립적으로 움직이는 변수들이 아니라 서로 같은 방향으로 움직이지 않는가? 매출이 평균보다 많으면 경쟁자가 들어오고 가격은 하락하는 반면 물량이 많아짐으로 초과 작업시간이 늘어나고 변동비용이 늘어날 가능성이 많기 때문이다.

따라서 실제로 계산된 기댓값 10,000원은 실제 기댓값이 아니고 이보다 더 하락할 가능성이 많다. 위의 엿가락 예제에서도 보았듯이 이런 경우는 아무리 기댓값을 대입하더라도 평균이 나오지 않는다. 따라서 항목들의 기댓값을 이용하여 이익의 기댓값을 구할 수는 없다.

경영회의에서는 기댓값을 쓰지 말라?

그러나 현실은 어떠한가? 회의석상의 많은 관리자는 각 부서에서 올라오는 값을 모형 내에 그대로 대입(plug-in)시킴으로써 베이스 모형(base model)을 만들곤 하는데 **매우 잘못된** 관행이다. ■

예제 7.13 **프로젝트 기일은 왜 못 맞추는가?**

- 어떤 프로젝트를 수행하기 위해서는 여러 개의 중간 단계를 순차적, 혹은 독립적으로 수행하여야 하는 경우가 많다. 어떤 프로젝트가 10개의 중간 과정으로 구성이 되어 있다고 가정하고 개개의 항목을 수행하기 위해서는 100시간이라는 평균 종료시간을 가지고 있다고 하면 이 프로젝트를 100 × 10 =1,000시간 이내에 끝나기 위해서는 평균적으로 모든 항목들이 평균 이내로 일을 마무리해야만 가능할 것이다.

데드라인만 있을 뿐이다.

그러나 중간 단계항목들이 서로 독립적으로 따로 따로 진행이 되지 못할 경우는 1,000시간 내에 마무리가 될 가능성은 희박해진다. 그리고 한 단계가 끝난다고 해서 바로 다른 단계로 나아가지 못하고 기다려야 한다면 프로젝트가 마무리해야 하는 시간은 더 오래 걸릴 것이다. 통상적으로 전체 프로젝트가 완성되는데 걸리는 시간은 개별 단계 수행에 걸리는 시간 평균의 단순 합보다 더 길게 걸릴 확률이 짧게 걸릴 확률보다 훨씬 높아지는 이유이다. ■

7.8 확률변수의 함수도 이용한다.

주어진 확률변수 X에 대한 확률분포뿐만 아니라 변수 X의 함수로 나타나는 새로운 확률변수의 확률분포 역시 많은 경우 관심대상이 된다. 예를 들어보자.

예제 7.14 **주어진 확률변수가 관심사항이 아니다.**

- 어느 지방단체는 매년 10월쯤이면 지방의 명소를 찍은 고급 달력을 주문하여 유관기관 및 관광객들에게 유상으로 배포한다. 달력 하나당 주문 단가는 20,000원이고 판매단가는 45,000원이다. 만약 달력이 다 팔리지 않는다면 이듬해 1월 30일이 지나 7,500원에 제작자에게 반송을 할 수 있는 권한이 주어져 있다. 이 공공단체는 1월에 판매될 달력의 **수요**를 [표 7.8]과 같이 생각하고 있다. 그러면 이 공공단체가 달력을 2,000개 주문하여 생길 수 있는 **이익의 분포**를 생각하여 보자. <달력수요.xls>

수요(개)	확률
1,000	0.3
1,500	0.2
2,000	0.3
2,500	0.15
3,000	0.05

[표 7.8] 달력수요 확률분포

주문량과 수요의 크기에 따라 두 가지 경우의 시나리오가 나올 것이다.

하나는 주문을 2,000개로 했는데 2,000개의 모든 달력이 다 팔리는 경우다. 즉, 연초에 남는 달력이 없을 것이다. 그러나 다른 한 경우는 달력에 대한 수요가 예상보다 밑돌아 남는 경우이다. 예를 들면 1,000개의 달력만 팔린다면 1,000개의 달력은 7,500원만 받고 반송하여야 한다. 이 경우 이익은 들어오는 돈(판매수입금과 반품대금)에서 나가는 돈을 뺀 금액으로 결정이 된다. 계산하면

45,000원 × 1,000개 + 7,500원 × 1,000개 - 20,000원 × 2,000개 = 12,500,000원

이다. 이를 일반화 시킨 다음과 같은 공식을 적용하여 보면 다음과 같다.

이익 = 45,000원 × min(주문량, 수요) +7.500원 × max(0, 주문량-수요) -20,000원 × 주문량

향후 발생하는 수요에 따른 이익을 이와 같이 계산하여 정리하면 주문량 2,000개 일 때 이익 변수에 대한 확률분포표가 나온다. 주문량을 2,000개로 하였을 경우 기대이익은 3,500만원이 나오며 표준편차는 약 1,634만원이 된다. 이를 위한 모든 계산과정은 아래 [표 7.9]와 같다. <달력판매.xls>

주문개수	2,000				
수요	1,000	1,500	2,000	2,500	3,000
판매가	45,000	45,000	45,000	45,000	45,000
판매량	1,000	1,500	2,000	2,000	2,000
비판매량	1,000	500	0	0	0
구입가	20,000	20,000	20,000	20,000	20,000
환불가	7,500	7,500	7,500	7,500	7,500
수입	52,500,000	71,250,000	90,000,000	90,000,000	90,000,000
지출	40,000,000	40,000,000	40,000,000	40,000,000	40,000,000
이익	12,500,000	31,250,000	50,000,000	50,000,000	50,000,000
수요확률	0.30	0.20	0.30	0.15	0.05
이익 기대값	35,000,000				
(이익-기대값)	-22,500,000	-3,750,000	15,000,000	15,000,000	15,000,000
(이익-기대값)^2	506,250,000,000,000	14,062,500,000,000	225,000,000,000,000	225,000,000,000,000	225,000,000,000,000
이익분산	2.67188E+14				
표준편차	16345871.04				

[표 7.9] 이익에 대한 기댓값과 분산에 대한 계산과정

이익에 대해서만 요약 정리하면 [표 7.10]과 같다.

이익(만원)	확률
1,250	0.3
3,125	0.2
5,000	0.5

[표 7.10] 달력판매에 따른 이익 확률분포

주어진 수요 확률분포표에서 우리가 분석 대상이 되는 **이익이라는 유도된 확률변수**의 확률분포표를 구하는 경우는 매우 중요한 훈련 중의 하나가 된다.

- 아직 이 예제가 보여주는 의미가 남아 있다. 모형에서 입력된 주문량은 2,000개이지만 이는 우리가 원하는 주문량이 아닐 수 있다. 주문량의 크기를 1,000, 1,500, 2,000, 2,500, 3,000으로 변화시켜 이익의 기댓값이 어느 주문량에서 제일 크게 나오는지 의사결정할 수 있다. 이를 정리하면 [표 7.11]과 같다.

주문량	이익 평균(만원)	이익 표준편차(만원)
1,000	2,500	0
1,500	3,187	8,593
2,000	3,500	1,634
2,500	3,250	2,087
3,000	2,718	2,255

[표 7.11] 주문량에 따른 이익의 기댓값과 표준편차

- 우리는 주문량 2,000개에서 제일 높은 기댓값 3,500만원이 나온다는 사실을 확인 할 수 있다. 마찬가지 방법으로 표준편차를 최소화하는 주문량도 구할 수 있다. 물론 주문량이 1,000개 일 때는 어떤 달력 수요가 발생하든 간에 모든 달력을 판매할 수가 있기 때문에 표준편차가 0이 나오지만 이 경우를 제외하면 주문량 중에서는 주문량 2,000개에서 표준편차가 최소가 나온다. 그리고 주문량 2,000개일 때 이익의 기댓값이 제일 크게 나왔기 때문에 이 문제에서는 주문량 2,000개가 주어진 5가지 대안 중에서 제일 최적으로 보인다.

- **주문량을 정하는데 있어 수요의 평균 (1,725개)만큼 주문을 하겠는가? ■**

7.9 공분산과 상관계수란 무엇인가?

상관계수는 매우 흔하게 쓰는 용어이다.

지금까지는 하나의 확률변수의 분포에 대해 표현 방법과 요약 방법을 알아보았다. 확률변수 두 개가 주어져있을 때는 하나의 확률변수에 대한 설명 뿐 아니라 두 변수의 관계를 같이 설명하여야 하기 때문에 이를 위한 숫자가 필요하다. 공분산과 상관계수는 이미 제 6장에서 소개

되었지만 여기서는 변수 X와 Y의 결합확률의 개념으로 설명하는 정도의 차이이다.

• 공분산을 구하는 절차

(1) 먼저 확률변수 X가 취하는 x_i와 확률변수 Y가 취하는 y_i에서 각각의 평균을 뺀 다음 이를 곱한다. : $(x_i - E(X))(y_i - E(Y))$

(2) 여기에 해당하는 확률 $p(x_i, y_i)$ 를 곱한다. : $(x_i - E(X))(y_i - E(Y)) \times p(x_i, y_i)$

여기서 $p(x_i, y_i)$는 확률변수 X가 x_i의 값을, 확률변수 Y가 y_i의 값을 동시에 가지는 확률, $P(X = x_i, Y = y_i)$ 이다. 이러한 확률을 결합확률(joint probability)이라 한다.

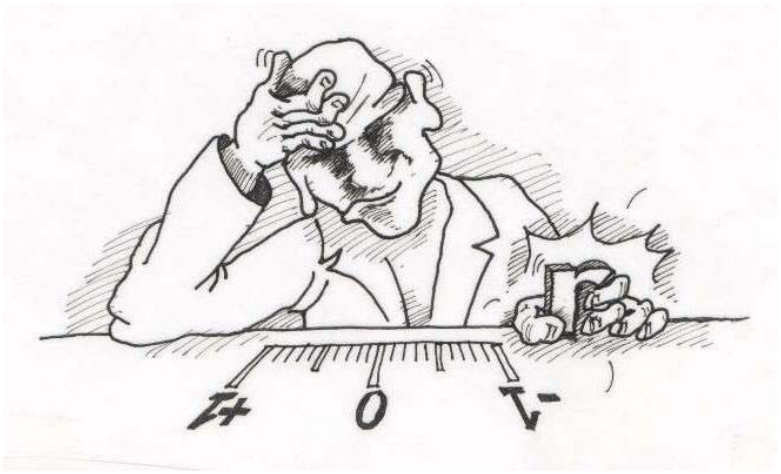
(3) (2)에서 구한 모든 값들을 더해 다음과 같이 공분산을 구한다.

$$\text{공분산} = \sigma_{X,Y} = \sum_{i=1}^{k} (x_i - E(X))(y_i - E(Y))p(x_i, y_i) \qquad (7.3)$$

• 상관계수는 $\frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$ 으로 구할 수 있다.

상관계수를 구하기 위해서는 공분산을 구해야 하고 공분산을 구하기 위해서는 결합확률을 구해야 하는데 결합확률을 명시한다는 것은 매우 힘든 작업이다. 따라서 이를 우회하는 방법을 소개하고자 한다.

많은 관리자가 애용하는 시나리오 접근방법은 결합확률을 명시하지 못할 때 사용하는 일반적인 방법이다. 이러한 시나리오 방법은 많은 조직에서 선호하는 방법이다. 예를 들어 설명하여 보자.



예제 7.15 시나리오 접근방법이란 무엇인가?

- 어느 지자체는 두 개의 프로젝트를 진행하려고 한다. 프로젝트의 수익률은 경기에 따라 달라지는데 경기는 단순하게 네 가지, 경기불황(depression), 경기후퇴(recession), 정상경기(normal), 경기호황(boom)으로만 구분한다. 그리고 이러한 경기가 나올 가능성은 전문가들의 의견을 종합하여 각각 0.05, 0.30, 0.50, 0.15로 설정하였다. 첫 번째 프로젝트와 두 번째 프로젝트의 수익률은 각각의 경기 시나리오별로 [표 7.12]에서와 같이 각각 다르게 설정되었다. <프로젝트1.xls>

경제여건	확률	첫 번째 프로젝트	두 번째 프로젝트
불황	0.05	-0.15	0.05
후퇴	0.30	0.05	0.20
정상	0.50	0.20	-0.12
호황	0.15	0.30	0.09

[표 7.12] 경기상황에 따른 두 프로젝트의 수익률

지자체는 이러한 두 프로젝트의 수익률의 상관계수를 요구하고 있다. 그리고 더 나아가 두 프로젝트의 포트폴리오(portfolio)를 구성하는 분석을 원하고 있다. 과연 여러분이 이 지자체에게 투자자문가로서 어떤 조언을 해 줄 수 있는가를 살펴보자.

- 첫 번째 프로젝트의 평균 수익률은 15.25%, 표준편차는 11%, 그리고 두 번째 프로젝트인 경우는 평균 수익률이 1.6%, 표준편차가 14.2%로 나왔다. 이에 대한 계산은 중복설명이 되므로 생략한다. 그리고 첫 번째 프로젝트의 수익률과 두 번째 수익률의 상관계수는 -0.495로 음의 관계가 나왔다. 물론 모든 계산과정은 엑셀로 실습이 될 것이다. <프로젝트2.xls>

	프로젝트1	프로젝트2	확률	프로젝트1 편차	프로젝트1 편차	두 편차의 곱	확률 X 두편차의 곱
	-0,15	0,05	0,05	-0,3025	0,034	-0,010285	-0,00051425
	0,05	0,2	0,3	-0,1025	0,184	-0,01886	-0,005658
	0,2	-0,12	0,5	0,0475	-0,136	-0,00646	-0,00323
	0,3	0,09	0,15	0,1475	0,074	0,010915	0,00163725
기대값	0,1525	0,016					
표준편차	0,11009	0,14242					
				공분산	확률 X 두 편차의 곱		-0,007765
				표준편차	공분산/각각의 표준편차		-0,495247859

[표 7.13] 공분산 및 상관계수 계산과정

이는 첫 번째 프로젝트의 수익률이 올라가면 두 번째 프로젝트의 수익률은 내려가고 반대로

첫 번째 프로젝트의 수익률이 내려가면 두 번째 프로젝트의 수익률이 올라갈 가능성이 발생하는데 상관계수 크기로는 -0.495가 된다는 이야기이다.

그러나 한편이 올라간다고 해서 다른 한편이 반드시 내려가고 반대로 한편이 내려간다고 다른 한편이 반드시 올라가는 그런 관계는 아니다. 만약 그렇다면 계산된 상관계수는 -0.495 보다 -1 에 가까운 값을 얻을 것이다. 물론 정확하게 두 수익률이 음의 선형의 관계에 있다면 -1 의 상관계수 값을 가질 것이다.

- 지자체가 1억원을 가지고 있고 60%는 첫 번째 프로젝트에, 40%를 두 번째 프로젝트에 투자한다면 이와 같은 포트폴리오의 수익률과 표준편차는 어떻게 나오는지 알아보자. 첫 번째 프로젝트의 1원당 수익을 X라 하고 두 번째 프로젝트의 1원당 수익은 Y라 하자. 그러면 포트폴리오의 수익률은 다음과 같아진다.

$$\text{포트폴리오의 수익금} = 6{,}000 \text{ 만원} \times X + 4{,}000 \text{ 만원} \times Y$$

경제가 불황일 경우 첫 번째 프로젝트는 -15%, 그리고 두 번째 프로젝트는 5%의 수익률을 얻는다. 따라서 포트폴리오의 1원 당 수익은 $0.6\times(-0.15) + 0.4\times(0.05) = -0.09 - 0.02 = -0.07$ 이 된다. 이를 1억원으로 곱하면 경제가 불황일 때는 700만원의 손해를 보는 것이다. 이러한 과정을 표로 작성하면 [표 7.14]와 같다. 단위는 만원 단위로 표시하였다.

경제상태	1 원당	전체(만원)
불황	-0.07	-700
후퇴	0.11	1,100
정상	0.072	720
호황	0.216	2,160

[표 7.14] 포트폴리오 수익금 분포

이를 바탕으로 포트폴리오의 수익금의 평균을 구하게 되면 약 979만원이 나오며 표준편차는 약 623만원이 나온다.

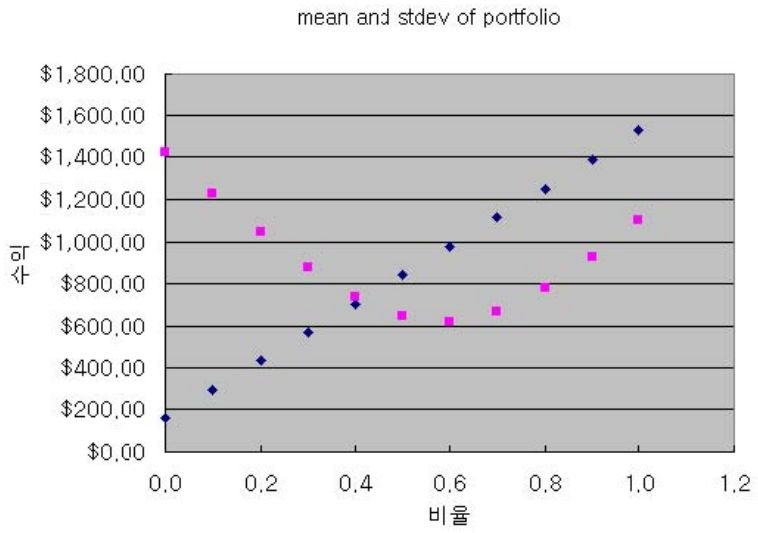
지금은 첫 번째 프로젝트에 대한 투자비율을 60%로 하였지만 이 비율을 0%에서 10% 단위로 늘려 나갈 때 포트폴리오의 수익금의 평균과 표준편차가 어떻게 변화하는지 [표 7.15]에 구하여 보았다.

첫 번째 프로젝트 비율	평균	표준편차
0.0	160.0	1,424.22
0.1	296.5	1,231.00
0.2	433.0	1,047.94
0.3	569.5	881.39
0.4	706.0	742.57
0.5	842.5	649.48
0.6	979.0	622.98
0.7	1,115.5	671.00
0.8	1,252.0	779.91
0.9	1,388.5	928.51
1.0	1,525.0	1,100.85

[표 7.15] 첫 번째 프로젝트가 포트폴리오에서 차지하는
비율의 변화에 따른 수익금의 평균과 표준편차

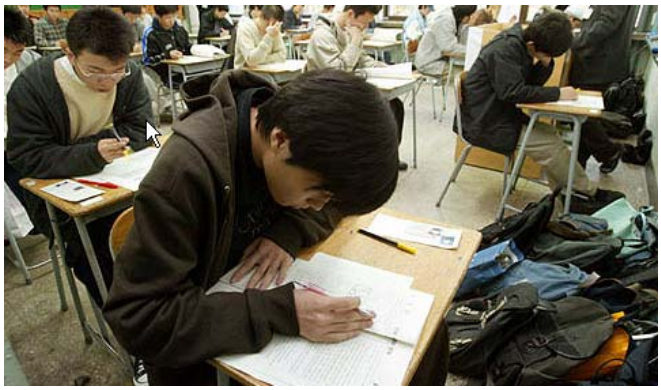
따라서 [표 7.15]를 보게 되면 첫 번째 프로젝트에 60%의 비율로 투자를 하게 되면 제일 표준편차가 작게 나옴을 알 수 있다. 만약 지방자치정부가 표준편차로 표시되는 위험을 제일 작게 하는 포트폴리오를 원한다면 첫 번째 프로젝트에 60%의 투자를 하고 나머지는 두 번째 프로젝트에 투자를 하여야 한다는 이야기다. 그러나 적어도 1,000만원 이상이 원하는 기대수익금이라면 투자비율을 60%에서 70%로 높여야 할 것이다.

그리고 [그림 7.6]을 참조하기 바란다. 포트폴리오의 기대수익금은 첫 번째 프로젝트에 대한 투자비율을 높이면 높일수록 선형으로 증가하지만 표준편차는 아래로 볼록한 모양을 하고 있다. ■



[그림 7.6] 포트폴리오 기댓값 및 표준편차 변화

시나리오 접근방법은 많은 관리자가 좋아하는 방법이다. 시나리오를 설정하고 각 시나리오별로 나올 수 있는 확률변수 값을 정하고 확률을 부여하는 방법이기 때문에 계산과정에서 결합확률이라는 개념을 도입할 필요가 없었다.



수학능력 시험만이 능력을 평가하는 잣대는 아니다.

예제 7.16 **수학능력 시험을 분석하자.**

- 그러나 이러한 시나리오 방법만이 두 변수의 확률분포를 표시할 수 있는 방법은 아니다. 좀 더 원시적인 방법으로 확률변수 X와 Y가 가질 수 있는 모든 값들을 인식하고 쌍으로 이루어지는 모든 값 (x, y)에 직접적으로 확률을 부여하는 방법이 있다. 여기서 부여되는 확률은 $P(X=x, Y=y)$, 혹은 줄여서 $p(x,y)$라 표기하며, 이 확률은 $X=x$ 그리고 $Y=y$가 동시에 일어나는 **결합확률**이라고 하였다.

그러나 이런 직접적인 방법으로 공분산 및 상관계수를 구하는 문제는 만만치 않다. 왜냐하면 모든 조합에 대한 결합확률을 구하기 어렵기 때문이다. 다만 주어진 표에서 확률을 상대 도수의 개념으로 변환하여 구할 수는 있다. [표 7.16]은 언어와 수리가 영역을 선택한 학생들의 등급별 분포이다.

	언어									
수리가	1	2	3	4	5	6	7	8	9	총합계
1	1,798	1,549	957	562	182	38	8	2	0	5,096
2	2,916	3,536	2,815	1,944	873	185	42	7	8	12,326
3	1,749	2,813	2,785	2,508	1,373	362	64	9	8	11,671
4	1,812	3,830	5,125	5,830	4,163	1,286	289	41	16	22,392
5	851	2,432	4,242	6,549	6,177	2,562	674	90	17	23,594
6	303	1,057	2,281	4,540	6,032	3,426	1,246	209	54	19,148
7	108	384	900	2,242	4,098	3,566	1,906	519	153	13,876
8	10	47	159	505	1,206	1,721	1,627	869	409	6,553
9	6	7	20	85	235	421	640	563	516	2,493
총합계	9,553	15,655	19,284	24,765	24,339	13,567	6,496	2,309	1,181	117,149

[표 7.16] 도수분포표

	언어									
외국어	1	2	3	4	5	6	7	8	9	총합계
1	0.015	0.013	0.008	0.005	0.002	0.000	0.001	0.000	0.000	0.044
2	0.025	0.030	0.024	0.017	0.019	0.002	0.000	0.000	0.000	0.105
3	0.015	0.024	0.0324	0.021	0.012	0.003	0.001	0.000	0.000	0.100
4	0.015	0.033	0.044	0.050	0.036	0.011	0.002	0.000	0.000	0.191
5	0.007	0.021	0.036	0.056	0.053	0.022	0.006	0.000	0.000	0.201
6	0.003	0.009	0.019	0.039	0.051	0.029	0.011	0.002	0.000	0.163
7	0.000	0.003	0.008	0.019	0.035	0.030	0.016	0.004	0.001	0.118
8	0.000	0.000	0.001	0.004	0.010	0.015	0.014	0.007	0.003	0.056
9	0.000	0.000	0.000	0.001	0.002	0.004	0.005	0.005	0.004	0.021
총합계	0.082	0.134	0.165	0.211	0.208	0.116	0.055	0.020	0.010	1.0

[표 7.17] 상대도수분포표

[표 7.17]에 나온 상대도수를 확률로 생각하고 수리 가형 성적과 외국어 성적을 분석하여 보면 [표 7.18]과 같다. 여기서는 순서척도인 등급점수를 비율척도나 구간척도로 가정하고 상관계수를 가정한 것이다.

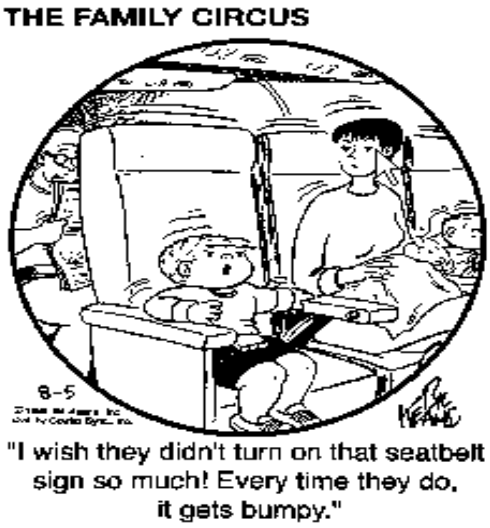
	수리가	외국어
평균	4.773	4.058
표준편차	1.914	1.780
공분산	2.101	
상관계수	0.617	

[표 7.18] 요약표

외국어는 수리 가형보다 평균 등급점수는 낮으나 표준편차로 계산되는 점수의 변동이 상대적으로 약간 적다. 또한 수리 가형과 외국어 점수 간의 상관계수는 약 62%로 나온다. ■

예제 7.17 **상관계수의 허와 실**

두 변수 간의 연관성을 알고자 할 때 우리는 자주 상관계수에 의존하는 습관이 있다. 그러나 상관계수가 높다고 해서 두 변수의 인과 관계까지 존재한다고 믿어서는 곤란하다. 특히 조사에 의해 수집된 자료인 경우는 더욱 더 그렇다. ■



이외에도 두 변수 간에 상관계수가 존재하지 않는 경우인데도 자칫 상관계수가 존재하는 것으로 착각하는 몇 가지 경우가 있어 소개한다.

- 두 변수가 독립이라 하더라도 두 변수와 상관관계가 있는 다른 변수로 나누게 되면 두 변수 사이에는 상관관계가 존재한다. 종종 다른 변수로 나누어 합이 100%가 되게 하는 경우 (영양소의 구성표 등) 변수 사이에는 음의 상관계수가 나타난다.

- 자료의 분포가 동질적이지 못할 때 상관계수가 나타난다. 신발의 크기가 작으면 임금의 수준이 낮다. 여자의 임금 수준은 남자보다 낮고 여자의 신발 크기는 작기 때문에 남녀의 자료를 같이 섞어 놓으면 이런 현상이 일어나는 것은 당연하다. 이럴 경우는 산점도를 그려 분포의 동질성을 확인하는 것이 좋다.

- 두 변수가 상관관계가 존재하는 것은 자료에 다른 변수가 빠져 있기 때문에 그렇다. 구두의 크기와 뼈의 칼슘 성분과는 상관관계가 존재한다. 그러나 아이들의 뼈에는 칼슘성분이 많지 않고 신발의 크기도 작다.

- 실험실에서 냉장고의 전원을 바꾸게 되면 두 시그널 장치에 높은 값을 갖는 이상점이 동시에 생길 가능성이 높다. 시그널을 유발하는 공동의 원인이 있어 높은 시그널을 두 자료가 같이 갖게 되는 경우 이러한 이상점은 두 시그널 사이에 높은 상관계수를 가져다준다.

학습요약

제 7장에서는 확률, 확률변수, 확률분포, 그리고 확률분포의 요약 척도(summary measures)등 매우 중요한 개념을 응용예제와 함께 배웠다. 확률과 부수되는 개념들은 불확실성하에서의 의사결정이라는 관점에서 매우 중요한 블록을 형성함을 알 수 있었다. 또한 이러한 내용은 다음 장의 내용과 직접적으로 연결되기 때문에 매우 중요하다. 확률의 기본적인 성질을 이용하여 시스템의 객관화 및 사고의 개선을 할 수 있다는 점과 특히 많은 의사결정에 자주 사용되는 기준으로 기댓값의 중요성 및 위험성을 적시할 수 있었다. 주위에서 일어나는 모든 확률적인 현상은 변수로 이해하며 변수끼리는 관계를 가지며 이를 위한 척도인 공분산, 상관계수도 배웠다. 또 시나리오 방법을 제안하였다. 결합확률은 직접적인 계산보다는 우회적인 계산이 훨씬 효과적이란 사실을 보았다. 제 7장의 내용은 통계학의 바다로 나아가는 첫 항해인 셈이다.

7장 연습문제

7.1 어느 지자체의 종업원 중 18%가 점심시간에 운동을 한다고 알려져 있다. 모든 종업원의 57%는 남자이다. 모든 종업원의 12%가 점심시간에 운동하는 남자로 밝혀졌다면

(1) 종업원 중 한 명을 임의로 표본추출 하였을 때 그 사람이 점심시간에 운동을 하는 여자일 가능성은?
(2) 종업원 중 한 명을 임의로 표본추출 하였을 때 그 사람이 점심시간에 운동을 하지 않는 여자일 가능성은?

7.2 어느 맥주공장은 두 가지 종류의 맥주를 판매하고 있다. 하나는 레귤러 맥주이며 다른 하나는 30% 열량이 덜 나가는 라이트 맥주이다. 홍보를 맡고 있는 관리자는 젊은 화이트칼러 직장인에게는 라이트맥주를 그리고 블루칼러 직장인에게는 레귤러 맥주를 홍보하는 전략이 적절한지 확인하고자 한다. 이를 위해 지역에 살고 있는 직장인 400명을 무작위로 추출하여 자료를 수집하였다. 여기서 Neutral은 레귤러맥주와 라이트 맥주 간의 차이가 없음을 나타낸다.

형태	Light	Neutral	Regular	합
Blue collar	111	17	122	250
White collar	79	12	59	150
합	190	29	181	400

이러한 전략이 적절한지 판단하고 그 이유를 설명하라. 필요한 확률은 무엇인지 파악하라.

7.3 도박장에서의 “house edge"는 어느 게임이든지 다음과 같이 정의된다. 게임을 하는 사람을 player라 하고 정의하면

$$\frac{\text{E(한 배팅에서 player의 잃어버리는 금액의 기댓값)}}{\text{player가 한 배팅에서 잃어버리는 금액의 크기}}$$

어느 게임에서 player가 확률 0.48을 가지고 10,000원을 따고 확률 0.52로 10,000원을 잃는다면 “house edge"를 계산하고 해석을 하여보라. 도박장은 얼마나 평균적으로 돈을 따는지 계산하라. 혹 룰렛 경기를 안다면 ”house edge"를 계산하여 보아라. 이기거나 지는 확률이 같은 축구경기에 돈을 걸어 이기면 10,000원을 받고 지면 11,000원을 잃어버리는 도박의 “house edge"는 어떠한가? 룰렛과 비교하여 어느 쪽이 더 ”house edge'가 높은가?

7.4 모 조직체의 사장의 비서가 3명의 임원과 내일 아침에 회의 스케줄을 잡으려고 한다. 그러나 세 명의 임원은 다른 선약 때문에 회의를 참석하지 못할 확률이 40%라 가정하면

(1) 회의에 참석하는 임원의 수를 확률변수 X라 한다면 X의 분포는?

(2) 모든 임원이 참석을 하여야만 회의가 진행이 된다면 회의가 열릴 확률은?

(3) 몇 명의 임원이 내일 참석할 수 있을 것으로 기대하는가?

7.5 어느 지자체가 판매하는 제품의 수요 함수는 다음과 같이 주어져 있다. 여기서 I는 1부터 40까지의 정수로 가격을 의미한다. 즉, $I=1,2,\cdots,40$ 이며 p값이 1부터 40까지 나올 가능성은 같다. 즉, $p(I=i)=0.025, i=1,2,\cdots,40$ 이라고 가정하자.

$$Q=200-5I$$

라 할 때

(1) 확률변수 I의 기댓값과 표준편차는? 그리고 어떻게 해석을 하여야 하는가?

(2) Q라는 유도된 확률변수의 기댓값과 표준편차는?

(3) 제품을 생산하는데 10원이 들고 생산된 제품을 모두 판매한다면 이 회사의 이익 함수는?

(4) 이익의 기댓값 및 표준편차를 구할 수 있는가?

7.6 다음은 내년도 경기 예측에 따른 두 백화점의 내년도 매출 예상액이다. 단위는 100억이다.

경기상태	확률	롯데	신세계
매우 좋음	0.15	10.5	8.6
중간	0.35	9.7	7.9
약함	0.25	8.2	7.5
매우 나쁨	0.25	7.5	7.0

(a) 두 백화점의 내년도 매출 예상액의 평균 및 표준편차를 이용하여 비교하라.

(b) 두 백화점 매출액의 공분산 및 상관계수를 구하라.

7.7 다음 표는 3층 건물에 들어 있는 두개의 엘리베이터가 동시에 정지되어 있는 경우 X는 첫 번째 엘리베이터의 층 위치를 나타내는 확률변수이며 Y는 두 번째 엘리베이터가 정지되어 있는 층의 위치를 나타내는 확률변수이다. 아르바이트 학생을 시켜 동시에 정지되어 있는 경우 층의 위치를 파악하여 결합확률을 작성하여 보았다.

X \ Y	1	2	3
1	0.25	0.08	0.14
2	0.07	0.10	0.07
3	0.16	0.08	0.05

(a) 같은 층에 정지되어 있지 않을 확률은?
(b) 두 번째 엘리베이터가 3층에 정지되어 있을 확률은?
(c) 첫 번째 엘리베이터가 1층에 정지되어 있지 않을 확률은?
(d) 첫 번째 엘리베이터가 1층에 없을 때 두 번째 엘리베이터가 1층에 있을 확률은?
(e) 1층에서 고객이 엘리베이터를 타기 위해 접근했을 때 엘리베이터가 1층에 정지되어 있지 않은 경우는 얼마나 되는가?
(f) 2층이나 3층에서 엘리베이터를 탄다면 (d)의 확률은?
(g) 여러분이 호텔 매니저라면 이러한 결과에 어떻게 반응할 것인가?
(h) 두 확률변수는 독립인가? 설명하라.

7.8 모 제지회사는 인도네시아 밀림 한쪽의 땅을 매입하여 제지공장의 펄프를 확보하려고 한다. 5월 현재 시점에서 가격은 22억원이다. 이 회사는 7월 초까지는 이 지역에서 나오는 목재를 원하지는 않지만 최고 경영진은 경쟁자가 그 사이에 이 땅을 매입하지 않을까 의심하는 바이다. 5월에 경쟁자가 이 땅을 매입할 가능성은 1/20이며 6월에는 1/10이다. 이 회사는 5월에 액션을 취하지 않아도 땅이 여전히 구입가능하다면 6월말이나 7월초에도 매입할 수는 있다.

이 땅을 지금 매입하지 않고 연기하는 주된 이유는 땅의 가격이 5월과 6월에 혹은 동시에 떨어질 가능성이 있기 때문이다. 아래 [표 7.19]와 [표 7.20]에 재무 분석가가 지가 변동에 대한 예측을 하였다. [표 7.19]는 5월 땅의 하락 가능성에 대한 표이며 [표 7.20]은 5월에 땅의 가격이 하락한다면 6월 땅의 가격하락에 관한 조건부 확률표이다.

지금 이 제지회사가 땅을 매입한다면 땅의 매입가격을 제외한 30억원 매출을 기대할 수 있다. 만약 땅을 매입하지 않는다면 다른 투자대안에서 약 6,500만원의 매출을 기대할 수 있다. 이 회사는 어떠한 결정을 내려야 하는지?

가격하락	Probability
0만원	0.5
6,000만원	0.3
1,2000만원	0.2

[표 7.19] 5월 가격하락 분포

5월 가격하락					
0원		6,000만원		12,000만원	
6월 가격하락	확률	6월 가격하락	확률	6월 가격하락	확률
0원	0.3	0	0.6	0	0.7
6,000만원	0.6	3,000만원	0.2	2,000만원	0.2
1,2000만원	0.1	6,000만원	0.2	4,000만원	0.1

[표 7.20] 6월 가격하락 분포

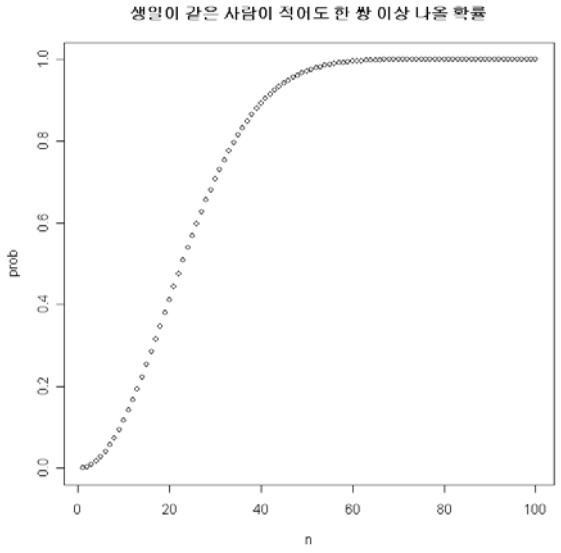
쉬어가기

1. 생일이 같을 확률은?

문제: 한 방에 모여 있는 사람의 수가 N이라 하자. 한 방의 사람들 중 서로 같은 생일을 갖게 될 확률을 구하여라(1년은 365일로 가정함).

(풀이) 우선 P[N명이 같은 생일을 갖지 않는다]를 구하여보자. 첫 번째 사람이 365일 중 하루를 선택할 확률은 $\frac{365}{365}$, 첫 번째 사람이 365일 중 하루를 선택하면 두 번째 사람은 나머지 364일 중 하나를 선택하여야 하므로 두 번째 사람이 하루를 선택할 확률은 $\frac{364}{365}$, 이런 식으로 생각하면 P[N명이 같은 생일을 갖지 않는다]=$\frac{365\times364\times363\times\ldots\times(365-N+1)}{365^N}$이 된다. 그러므로 우리가 구하여야 할 확률은 여사건의 확률공식을 이용하여 P[N명 중 최소한 두 명이 같은 생일을 갖는다]= 1-P[N명이 같은 생일을 갖지 않는다]=$1-\frac{365\times364\times363\times\ldots\times(365-N+1)}{365^N}$ 이 된다.

$N(N=1,2,\ldots,100)$에 따른 확률 P[N명 중 최소한 두 명이 같은 생일을 갖는다]의 변화를 그림으로 그리면 다음과 같다.



한 방에 23명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.507로 0.5를 넘게 되고 한 방에 40명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.891로 매우 큼을 알 수 있다. 이 문제는 365개의 면을 가지고 있는 주사위 40개를 동시에 던졌을 때

같은 숫자가 나오는 쌍이 하나라도 나오는 확률을 구하는 문제인 셈이다. 우리가 언뜻 생각할 때 1년은 365일이므로 최소한 두 명이 같은 생일을 갖게 되려면 365명 이상이 있어야 할 것 같지만 70명만 되어도 최소한 두 명이 같은 생일을 갖게 될 확률이 무려 0.999가 된다.

2. 14면 주사위

1) 목제주령구

목제주령구(木製酒令具)(7.1절 그림 참조)라는 것은 1975년 경주 안압지를 발굴하던 중 연못 바닥의 갯벌 속에서 발견된, 흑칠(黑漆)한 참나무로 만든 높이 4.8cm의 작은 14개의 면에 술자리에서의 여러 가지 벌칙을 적어놓은 주사위이다. 이 주사위에는 6개의 사각면과 8개의 육각면으로 구성되었다. 이 주사위는 통일 신라 시대에 귀족들이 술좌석 등 여러 사람이 모인 흥겨운 자리에서 놀이에 쓰였을 것으로 추측된다. 이 목제 주사위의 진품은 화재로 인하여 불타버렸으며, 현재는 그 모조품만이 국립경주박물관에 소장되어 있다.

사각면에 적힌 벌칙 문구와 육각면에 적힌 벌칙 문구는 다음과 같다.

- 사각면에 적힌 벌칙 문구

음진대소(飮盡大笑) : 술 마시고 크게 웃기, **삼잔일거(三盞一去)** : 술 석잔을 한 번에 마시고 한걸음걷기, **자창자음(自昌自飮)** : 혼자 노래 부르고 술 마시기, **금성작무(禁聲作舞)** : 소리 내지 않고 춤추기, **중인타비(衆人打鼻)** : 여러 사람들로부터 코 맞기, **유범공과(有犯空過)** : 여러 사람이 덤벼서 장난쳐도 참기

- 육각면에 적힌 벌칙 문구

추물막방(醜物莫放) : 더러워도 버리지 않기, **양잔즉방(兩盞則放)** : 술 두 잔을 빨리 마시고 다른 이에게 돌리기, **임의청가(任意請歌)** : 아무나 지목해 노래 청하기, **곡비즉진(曲臂則盡)** : 팔을 구부리고 술을 다 마시기, **농면공과(弄面孔過)** : 얼굴을 간지럽게 해도 참기, **자창괴래만(自昌怪來晩)** : '괴래만'이라는 노래 부르기, **월경일곡(月鏡一曲)** : '월경'이라는 노래 부르기, **공영시과(空詠詩過)** : 시 한수 읊기

14면 주사위는 정육면체의 8개 꼭지점을 잘라내어 만든다. 자를 때 한 꼭지점에 모이는 세 개의 각 변에 대하여 변의 중심점까지 자르면 8개의 **삼각면**과 6개의 사각면으로 구성된 14면 주사위가 된다. 허명회(1994a)는 14면 주사위에서 빈도적 확률을 구하는 문제를 얻는 데 기하적 단순 대칭성을 위하여 8개의 삼각면과 6개의 사각면으로 구성된 14면 주사위로 가정하였다. 14면 주사위에 외접하는 구 겉 표면에서의 구삼각형의 면적을 구하여 삼각면이 나오는 논리적 확률로서 Pr[삼각면이 나오는 사건]= 0.35096으로 계산하였다. 또한, 허명회(1994a)는 실제로 모형을 제작하여 14면 주사위를 2000번 굴려 삼각면이 481번 나와 빈도적 확률로서

Pr[굴렸을 때 삼각면이 나오는 사건]= 0.2405로 제시하였다. 한편, 허명회(1994b)는 14면 주사위를 1000번 던져 삼각면이 343번 나와 빈도적 확률로서 Pr[던졌을 때 삼각면이 나오는 사건]= 0.3430으로 제시하였다. 채경철과 이충석(1995)은 역학 에너지에 대한 가정을 통하여 베이즈 정리를 이용한 논리적 확률로서 Pr[굴렸을 때 삼각면이 나오는 사건]= 0.2376으로 계산하였다. 8개의 삼각면과 6개의 사각면으로 구성된 14면 주사위에서 우리는 다음과 같은 사실을 알 수 있다.

사실 1. 허명회(1994)가 제시한 논리적 확률은 14면 주사위를 던졌을 때의 빈도적 확률과 매우 유사하다.

사실 2. 채경철과 이충석(1995)이 제시한 논리적 확률은 14면 주사위를 굴렸을 때의 빈도적 확률과 매우 유사하다.

그런데, 고등학교 수학 수학 Ⅰ 교과서 1종(이강섭외 6인 공저, 2004)의 본문 중에 다음과 같은 구절이 나온다.

"경주 안압지에서 출토된 '목제주령구'라는 주사위는 보통 우리가 보아 온 6면체가 아니라 특이하게 14면체로 되어 있다. 이 주사위는 6개의 정사각면과 8개의 육각면으로 이루어져 있다. ...(중략)... 목제주령구의 각 면의 넓이를 계산하면, 정사각면의 넓이는 6.25 cm^2이고 육각면의 넓이는 6.265 cm^2이다. 그러므로 주사위를 굴렸을 때 각 면이 나오는 확률이 거의 같을 것으로 기대된다. 실제로 실험을 해 본 결과 각 면이 나타날 확률은 1/14로 볼 수 있다고 한다."

위의 저자들은 육각면이나 사각면이 나오는 고전적 확률로서 육각면이나 사각면의 넓이를 이용하여 Pr[굴렸을 때 육각면이 나오는 사건] = $(6.265\times 8)/(6.265\times 8+6.25\times 6)=0.5720\approx 8/14=0.5714$, Pr[굴렸을 때 사각면이 나오는 사건]=$6/14=0.4286$으로 계산할 수 있다고 주장하고 있다. 이러한 14면 주사위는 정육면체의 8개 꼭지점을 자를 때 한 꼭지점에 모이는 세 개의 각 변에 대하여 변의 중심점보다 더 깊이 자르면 8개의 **육각면**과 6개의 사각면으로 구성된 14면 주사위가 된다. 다음 절에서는 14면 주사위의 확률을 굴렸을 때와 던졌을 때로 나누어 시행하여 빈도적 확률을 각각 구하여 보고, 위의 주장에 대한 타당성을 검토하고 논리적 확률을 구하여 본다.

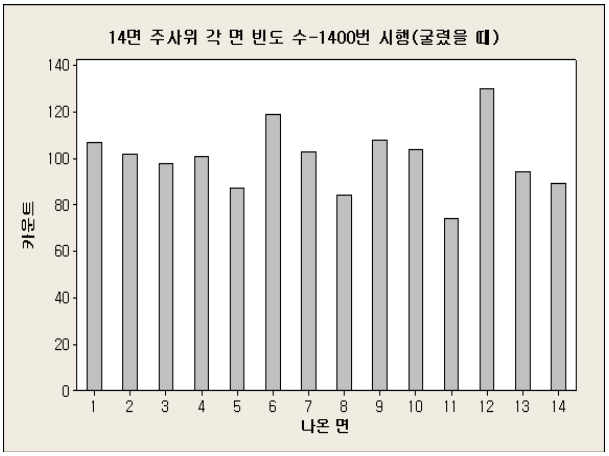
2) 14면 주사위의 확률

8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품(국립박물관 판매품)을 이용하여 1,400번 굴렸을 때의 결과는 다음 [표 7.21]과 같았다. 굴릴 때는 주의를 기울여 여러 차례 구를 만큼 세게 굴렸다.

면의 모양	각 면에 부여한 번호	각 면이 나온 횟수	합계		전체에 대한 각 면이 나온 비율(%)	전체에 대한 모양 별 비율(%)
육각면	1	107	801	1400	7.64	57.21
	2	102			7.29	
	3	98			7.00	
	4	101			7.21	
	5	87			6.21	
	6	119			8.50	
	7	103			7.36	
	8	84			6.00	
사각면	9	108	599		7.71	42.79
	10	104			7.43	
	11	74			5.29	
	12	130			9.29	
	13	94			6.71	
	14	89			6.36	

[표 7.21] 14면 주사위를 1,400번 굴렸을 때의 결과

[표 7.21]을 보면 Pr[굴렸을 때 육각면이 나오는 사건]=0.5721이어서 이강섭과 6인(2004)이 제안한 고전적 확률(육각면과 사각면의 면적을 이용한 확률) 8/14와 매우 유사하다. 이 14면 주사위를 1,400번 굴렸을 때 각 면이 나온 횟수를 그림으로 그리니 [그림 7.7]과 같았다. 100번을 중심으로 흩어져 나타나고 있다. 평균적으로 육각면이 101.1번, 사각면이 99.8번 나왔다. 거의 차이가 없고 각 면이 나타날 확률은 1/14로 볼 수 있다.



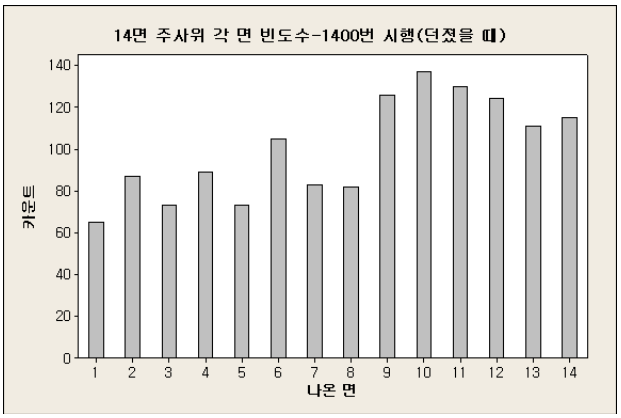
[그림 7.7] 14면 주사위를 1,400번 굴렸을 때 각 면이 나온 횟수

8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품을 이용하여 1,400번 던졌을 때의 결과는 다음 [표 7.22]와 같았다.

면의 모양	각 면에 부여한 번호	각 면이 나온 횟수	합계		전체에 대한 각 면이 나온 비율(%)	전체에 대한 모양 별 비율(%)
육각면	1	65	657	1400	4.64	46.93
	2	87			6.21	
	3	73			5.21	
	4	89			6.36	
	5	73			5.21	
	6	105			7.50	
	7	83			5.93	
	8	82			5.86	
사각면	9	126	743		9.00	53.07
	10	137			9.79	
	11	130			9.29	
	12	124			8.86	
	13	111			7.93	
	14	115			8.21	

[표 7.22] 14면 주사위를 1,400번 던졌을 때의 결과

[표 7.22]를 보면 Pr[던졌을 때 육각면이 나오는 사건]=0.4693이었다. 이 14면 주사위를 1,400번 던졌을 때 각 면이 나온 횟수를 그림으로 그리니 [그림 7.8]과 같았다. [그림 7.7]과는 다른 패턴을 보이고 있다. 평균적으로 육각면이 82.1번, 사각면이 123.8번 나왔다. 굴렸을 때와 달리 육각면이 사각면보다 오히려 더 적게 나오고 평균이 무려 약 40번 차이가 났다.



[그림 7.8] 14면 주사위를 1,400번 던졌을 때 각 면이 나온 횟수

두 종류의 14면 주사위에서 삼각면 또는 육각면이 나타나는 빈도적 확률을 비교하면 다음 [표 7.23]과 같다.

종류	굴렸을 때	던졌을 때
8개의 삼각면과 6개의 사각면	0.2405	0.3430
8개의 육각면과 6개의 사각면	0.5721	0.4693

[표 7.23] 두 종류의 14면 주사위에서 삼각면 또는 육각면이 나타나는 빈도적 확률

이강섭외 6인(2004)이 제안한 고전적 확률의 아이디어(육각면과 사각면의 면적을 이용한 확률 계산)를 8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품에 적용하기 위하여 사각면의 넓이와 육각면의 넓이를 구하여 보니 사각면의 넓이는 5.76 cm^2이고 육각면의 넓이는 5.684cm^2이었다. 이 14면 주사위는 다른 6면 주사위처럼 잘 구르게 하기 위하여 모서리를 약간 둥그스름하게 깍아 놓았다. 그래서 육각면과 사각면의 변의 길이를 잴 때 둥그스름한 부분의 한 가운데를 기준으로 하여 재었다. 이강섭외 6인(2004)의 계산으로는 사각면의 넓이는 6.25cm^2이고 육각면의 넓이는 6.265cm^2이어서 저자가 계산한 넓이와 다른 데 이는 육각면과 사각면의 변의 길이를 잴 때 이 둥그스름한 부분을 어떻게 고려하고 재었는가에 따른 차이라 생각된다. 저자의 계산에서는 사각면의 넓이가 약간 크나 이강섭외 6인(2004)의 계산에서는 육각면의 넓이가 약간 큰 것을 봐서 측량의 차이를 고려한다면 사각면의 넓이와 육각면의 넓이는 거의 같다고 볼 수 있다. 모조품이 아닌 진품의 경우 의도적으로 육각면과 사각면의 넓이가 같도록 제작되었을 가능성이 크다. 그러므로 저자의 계산을 이용한다면,

Pr[육각면이 나오는 사건]=$(5.684\times 80)/(5.684\times 8+5.76\times 6)=0.5682\approx 8/14=0.5714$가 나온다. 이 고전적 확률은 [표 7.21]에서 보는 것과 같이 이 14면 주사위를 굴렸을 때의 빈도적 확률과 매우 유사하다. 이강섭과 6인(2004)이 제안한 고전적 확률의 아이디어가 타당함을 알 수 있다. 그러나 이러한 아이디어가 물리적 이론이 뒷받침되지 않은 단순한 주장이어서 이러한 주장이 모든 14면 주사위에 적용될 수 있는지, 8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품에만 우연히 적용되었는지 따져 볼 필요가 있다.

이강섭과 6인(2004)이 제안한 고전적 확률의 아이디어(육각면과 사각면의 면적을 이용한 확률 계산)를 8개의 삼각면과 6개의 사각면으로 구성된 14면 주사위에 적용하여 보자. 삼각면의 한 변의 길이를 r이라 하면 삼각면의 넓이는 $(\sqrt{3}/4)r^2$ 이 되고 사각면의 넓이는 r^2이 된다. 그래서 삼각면의 넓이과 사각면의 넓이의 비는 $\sqrt{3}/4:1=0.4330:1$이 된다. 그러므로 Pr[삼각면이 나오는 사건]=$(0.4330\times 8)/(0.4330\times 8+1\times 6)=0.3660$ 이 나온다. 이 고전적 확률은 제시한 8개의 삼각면과 6개의 사각면으로 구성된 14면 주사위를 던졌을 때의 빈도적 확률과 비교적 유사하다.

8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품을 굴렸을 때와 던졌을 때를 합하여 살펴보면 육각면이 나오는 빈도적 확률은 Pr[육각면이 나오는 사건]=$(801+657)/2,800=0.5207$로서 육각면과 사각면이 비슷하게 나타남을 알 수 있

다. 반면, 8개의 삼각면과 6개의 사각면으로 구성된 14면 주사위를 굴렸을 때와 던졌을 때를 합하여 살펴보면 삼각면이 나오는 빈도적 확률은 Pr[삼각면이 나오는 사건]= $(481+343)/3,000=0.2747$ 로서 사각면이 삼각면보다 약 2.6배 많이 나옴을 알 수 있다.

8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위에서의 논리적 확률로서 채경철과 이충석(1995)이 제안한 역학 에너지에 대한 가정과 베이즈 정리를 이용한 논리적 확률을 이용하여 구하여 보면(구하는 과정은 생략하자.) 육각면이 나오는 사후확률은 $Pr[hexagon|S]=0.6588$이 된다. 여기서, S는 정지(stop)를 나타낸다. 실제 8개의 육각면과 6개의 정사각면으로 이루어져 있는 14면 주사위인 '목제주령구'의 모사품에서 육각면이 출현하였을 때(즉, 육각면이 바닥에 접하였을 때) 주사위의 중심에서 바닥까지의 길이는 $1.8cm$이었고 사각면의 경우는 $1.7cm$이었다. 육각면에서 인접한 다른 육각면으로 회전하는 과정에서의 중심의 최고 위치는 $2.2cm$이고 육각면에서 사각면으로 회전하는 과정에서의 중심의 최고 위치는 $2.0cm$이었다. 그러므로 육각면이 나오는 사후확률은 $Pr[hexagon|S]=0.6667$이었다. 여기서, S는 정지(stop)를 나타낸다. 위의 두 가지 논리적 확률은 [표 7.21]에서의 Pr[굴렸을 때 육각면이 나오는 사건]=0.5721과 다소 차이가 난다.

8개의 육각면과 6개의 사각면으로 구성된 14면 주사위에서 우리는 다음과 같은 사실을 알 수 있다.

사실 3. 육각면과 사각면의 면적이 비슷하다.

사실 4. 이강섭과 6인(2004)이 제안한 고전적 확률은 14면 주사위를 굴렸을 때의 빈도적 확률과 같다.

사실 5. 굴렸을 때와 던졌을 때를 합하여 살펴보면 육각면과 사각면이 비슷하게 나타난다.

위의 사실로 보건대 우리 선조들이 14면 주사위를 만들 때 6면 주사위와 같은 성질을 갖도록 절묘하게 만들었음을 알 수 있다.

참고문헌

[1] 이강섭, 허민, 김수환, 이정례, 임영훈, 왕규채, 송교식(2004). <수학 Ⅰ>, 지학사, 서울.

[2] 채경철, 이충석(1995). 14면 주사위 확률에 대한 역학적 고찰, 응용통계연구, 제 8권 2호, 179-185.

[3] 허명회(1994a). 14면 주사위의 확률, 응용통계연구, 제 7권 1호, 113-119.

[4] Huh, M. H.(1994b). Essays on probability from billiards and 14-face dice, *Proceedings of the 8th Japan and Korea Conference of Statistics*, 225-230.

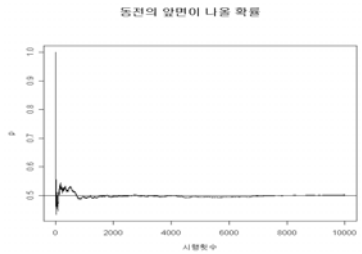
3. 통계적 확률에 대하여 알아보자!

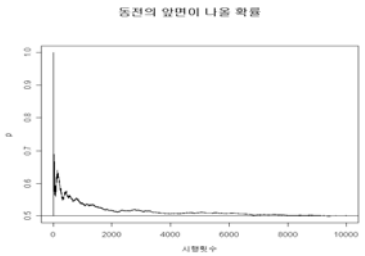
중고등학교 수학교과서에 나오는 통계적 확률의 예와 시행횟수는 다음 [표 7.24]와 같다. 수학적 확률로 구할 수 없는 예로서 병뚜껑을 던지는 시행에서 겉면이 나올 확률과 압정을 던지는 시행에서 침이 아래로 향하는 확률, 윷가락 한 개를 던지는 시행에서 앞면이 나오는(평평한 면이 위로 향할) 확률이 있다.

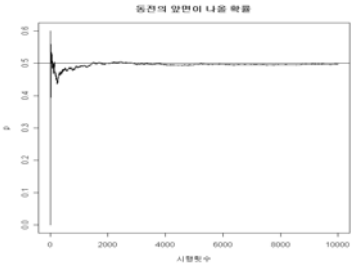
통계적 확률의 예와 시행횟수
동전(600), 태어난 아이의 남녀비(3,000)
동전(1500)
동전(100->횟수를 늘임), 주사위(2,000), 병뚜껑(500), 컴퓨터 모의실험
동전(800)
주사위(1명 30번->반 전체), 동전(1,000), 윷짝 한 개(20,000)
동전(400), 주사위(500)
주사위(1,000)
동전(2,000)
동전(1,000)
구슬주머니(흰색 4, 빨강 3, 녹색 2, 파랑 1)(50), 윷가락 한 개(900)
동전(500->1,000)
동전(1,000), 주사위(1,000)
뉴퐁의 바늘실험(100), 압정(1,000), 병마개(1,000), 동전(1,000), 주사위(1,000)
동전(100), 컴퓨터 모의실험
주사위(1,200)
동전(1,000), 주사위(2,000), 병뚜껑(1,000)

[표 7.24] 통계적 확률의 예

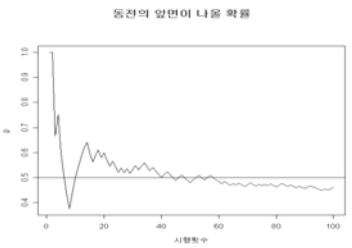
동전을 10,000번 던져 동전의 앞면이 나오는 확률을 세 차례 구하여 보니 다음과 같은 세 가지 그림이 나왔다. 우리는 세 가지 그림에서 각각 0.5에 수렴하는 것을 볼 수 있다. 그러나 수렴하는 패턴은 서로 다름을 알 수 있다.

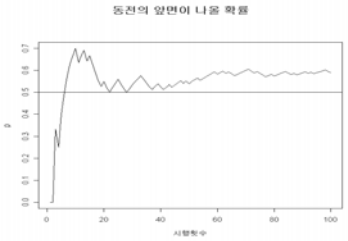


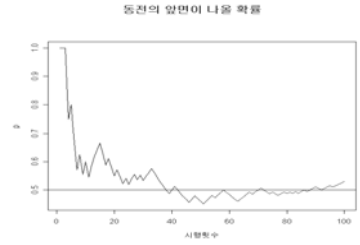




반면에 동전을 100번 던져 동전의 앞면이 나오는 확률을 세 차례 구하여보니 다음과 같은 세 가지 그림이 나왔다. 그림을 종합하여 보면 0.5에 수렴한다고 보기가 어렵다. 즉 시행횟수가 적으면 통계적 확률을 구하기가 어렵다는 것을 알 수 있다. 그러므로 통계적 확률을 보기 위해서는 컴퓨터를 이용한 시행횟수를 크게 하여 시뮬레이션을 행하여 볼 필요가 있다.

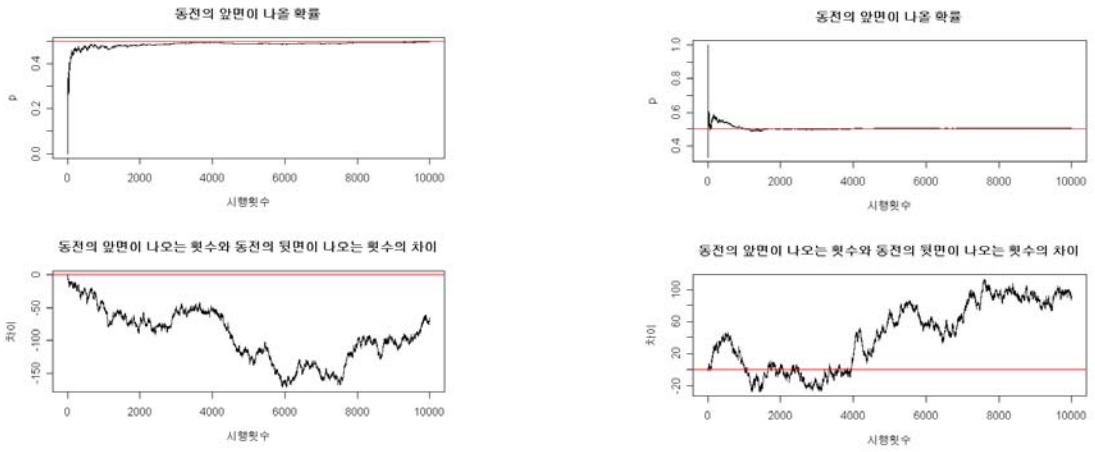




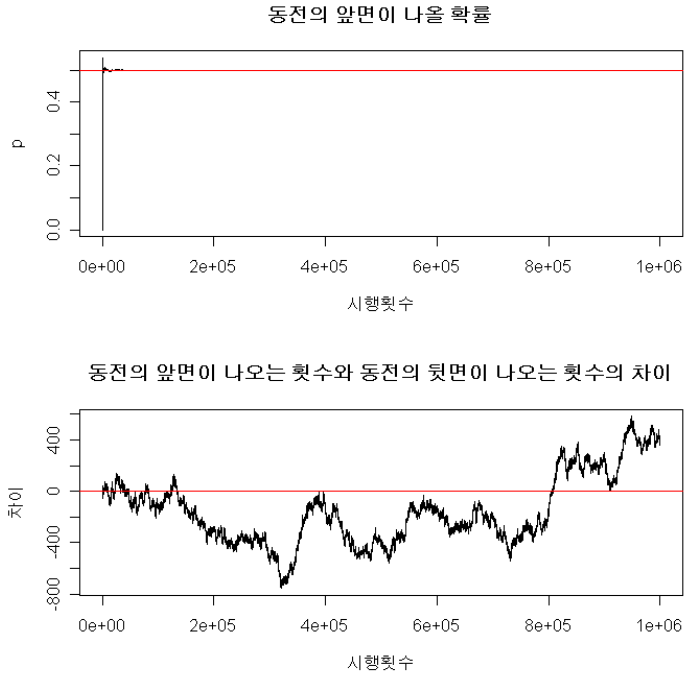


통계적 확률을 구할 때 주의할 또 한 가지 사항이 있다. 시행횟수가 커짐에 따라 동전의 앞면이 나오는 상대도수의 극한값이 0.5가 된다고 해서 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이가 0에 가까워진다는 것이 아니다. 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이는 커졌다 작아졌다 한다. 즉 이 차이는

랜덤하게 된다. 이를 다음과 같은 통계적 시뮬레이션으로 확인하여 보자.

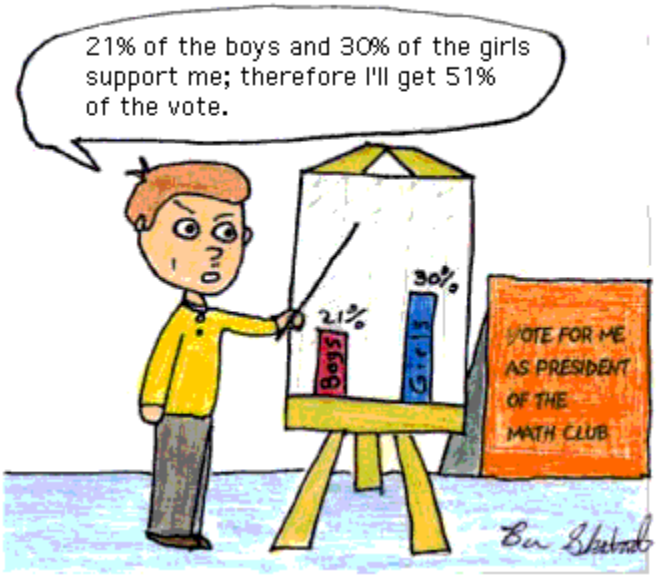


다음 그림은 시행횟수를 백만번으로 했을 때 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이를 나타내는 그림이다. 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이는 커졌다 작아졌다 하나 그 변동폭이 시행횟수가 만번일 때보다 큼을 알 수 있다.



제 8 장

분포를 알아야 숲이 보인다.



차 례

학습목표

제 7장에서는 확률분포에 대해 일반적인 정의를 하였다. 이번 장에서는 우리가 자주 접하는 몇 가지 대표적인 확률분포에 대해 이야기하도록 하자. 우선 확률분포는 이산형(discrete type)과 연속형(continuous type)으로 나누어 설명하는데 이산형의 확률분포 중에서는 베르누이 분포, 이항분포 등이 언급될 것이며 연속형 분포로는 균일분포, 삼각형분포, 정규분포 등이 언급될 것이다. 그리고 이에 관련된 예제를 들어 보이도록 한다. 이러한 확률분포는 통계학의 추론부문을 이해하는데 핵심적인 역할을 한다. 다소 딱딱한 부분도 있겠지만 주춧돌 역할을 하는 부분이니 독자들의 끈기가 필요하다.

8.1 이항분포란 무엇인가?

예제 8.1 **법원 판결에서도 통계학은 쓰인다.**

- 1964년 미국 로스앤젤레스에서 일어난 실제 사례이다. 이 사례를 통해 법원의 판결에도 통계학의 쓰임을 알아 볼 수 있을 것이다. 실제로 미국 스탠포드 대학교 법과대학의 교과목을 보게 되면 통계학 관련과목이 무려 4개이나 있다. 부당한 판결이 나오지 않게 과학적이고 논리적인 사고로 법의 정당성을 확보하고자 하는 것이다.

꽁지머리를 한 금발의 여인이 다른 여자의 핸드백을 소매치기하였다. 현장에서는 도망을 갔지만 카메라에 찍힌 필름을 검토한 결과 이 여자는 턱수염과 콧수염을 가진 흑인남자와 노란 승용차에 탄 것으로 확인되었다. 후에 경찰은 이러한 흑인과 접촉하고 있는 금발의 용의자를 체포하여 기소를 하였다. 다만 이 여자를 범인현장에 있었던 걸로 식별할 수 있는 증인은 없는 상태이다. 그러나 검사는 다음과 같은 확률에 의거하여 유죄를 주장하였다. 과연 검사의 주장을 받아들여 이 여자에게 유죄 판결을 내려야 하는지 생각하여 보자.

- **검사 주장**

로스앤젤레스에서 노란 차의 확률은 1/10, 콧수염을 기른 사람은 1/4, 꽁지머리를 한 여자는 1/10, 금발은 1/3, 턱수염을 기른 흑인남자는 1/10, 그리고 차 안에 다른 인종의 남녀 쌍이 있을 가능성은 1/1,000이다. 모든 확률을 곱하면

$$\frac{1}{10}\times\frac{1}{4}\times\frac{1}{10}\times\frac{1}{3}\times\frac{1}{10}\times\frac{1}{1{,}000}=\frac{1}{12{,}000{,}000}$$

으로 나와 이러한 쌍이 나올 가능성은 매우 희박하다.

그러나 이 시점에서 우리는 검사가 주장하는 이러한 확률만이 피의자의 유죄 여부를 결정할 숫자인가를 물어 보아야 한다. ■

이러한 문제에 대한 답을 하기 위해서는 이항분포에 대한 기본적인 지식을 가지고 있어야 한다. 이항분포의 선제 개념인 베르누이 분포부터 알아보자.

8.1.1 베르누이 분포란 무엇인가?

어떤 행위에 의해 결과가 두 가지로만 구분되어 나타나는 경우를 우리 주위에서 많이 볼 수 있다. 부품을 만들 때 부품이 불량품과 양품이 나오는 현상이라든지 사람이 태어날 때 남성과 여성으로 구분이 된다든지 사람이 죽고 사는 문제라든지 등이 모두 이런 범주에 속할 수 있다. 물론 여기서 현상을 구분해 주는 확률은 매번 시행될 때마다 변하지 않는다는 가정을 통상적으로 한다.

이러한 두 가지의 사건만 가져다주는 행위를 베르누이 시행(Bernoulli trial)이라 한다. 베르누이 시행을 베르누이 분포로 바꾸기 위해서는 어느 특정 사건은 1이라는 값을 가지게 하고 다른 사건은 0이라는 값을 가지게 할 필요가 있다. 1은 여기서 성공(success)을 의미하고 0은 실패(failure)를 의미한다.

따라서 어느 특정 사건이 나올 확률을 P(X= 1) = p라 한다면 P(X=0) = 1-p =q 는 특정사건이 나오지 않을 확률을 의미한다.

베르누이 분포 : 확률변수가 0과 1의 두 값만 가지고 확률변수의 확률분포가 식 (8.1)과 같이 정의되면

$$P(X=1)=p,\ P(X=0)=1-p \tag{8.1}$$

이 확률변수는 베르누이 분포를 따른다고 이야기한다.

따라서 베르누이 분포를 따르는 확률변수 X의 기댓값과 분산은 각각 [표 8.1]에서 구해진다.

x	$P(X=x)$	$xP(X=x)$	$(x-E(X))$	$(X-E(X))^2$	$(X-E(X))^2P(X=x)$
0	$1-p$	0	$-p$	p^2	$(1-p)p^2$
1	p	p	$1-p$	$(1-p)^2$	$p(1-p)^2$
	$E(X)$	p		분산	$p(1-p)$

[표 8.1] 베르누이 확률변수의 기댓값과 분산

그러나 베르누이 분포는 바로 이어지는 이항분포를 설명하기 위해 만든 것이기 때문에 베르누이 분포 자체는 큰 의미는 없지만 이런 예는 주위에서 많이 찾아 볼 수 있다.

예제 8.2 **한 장을 사는 것은 베르누이 시행이다.**

경매장에서 어느 말에 내기를 걸어 이기면 100원을 주고 지면 상금액이 없는 경우를 생각하여 보자. 위에서 언급한 베르누이 시행이 아닌가? 이 말이 경기에서 이길 확률이 1/10이라 한다면 기대 배당금은 얼마나 되는가? 그리고 이 가격보다 마권의 가격이 높으면 여러분은 마권 경기에 그다지 흥미를 느끼지 않을 것이다.

마권이 가지고 있는 가치에 대한 평가는 기댓값으로 정해진다.

$$1/10 \times 100 + 9/10 \times 0 = 10\text{원}$$

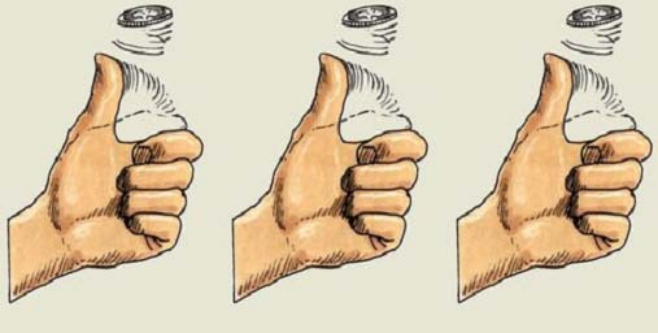
마권의 기댓값은 얼마나 되는가?

왜 이런 확률적인 논리 전개가 가능한가? 이런 종류의 마권을 한달 동안 250장을 샀다면 이 중 25개 경마에서 이겨 2,500원의 상금을 얻었기 때문에 마권 1장당 가격은 10원이 적절한 것이다. 이런 개념은 이항분포이다. ■

8.1.2 이항분포란 무엇인가?

- 던져서 앞면이 나올 확률이 1/2인 동전을 3번 던져 보자.

이는 결과가 둘 뿐인 베르누이 시행을 3번 하는 것이다. 매 시행 때마다 결과는 앞면 혹은 뒷면만 나오기 때문이다. 여기서 특정한 사건인 앞면이 나올 확률은 p=1/2이다. 즉, 동전을 던지는 베르누이 시행을 하여 앞면이 나오면 1의 값을 부여한다.



그러면 이 경우에는 3개의 동전을 던지므로 다음과 같은 8가지 결과가 나타날 것이다.

(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)

첫 번째 경우는 모두 앞면이 나오지 않는 경우이고 마지막 경우는 모두 앞면이 나온 것이다. 확률 변수 X를 앞면이 나온 횟수라 정의한다면 확률변수 X의 확률분포는 아래 [표 8.2]와 같다. 확률변수 X가 취할 수 있는 값은 0, 1, 2, 3 뿐이다.

x	$P(X=x)$
0	1/8
1	3/8
2	3/8
3	1/8

[표 8.2] n=3, p=0.5인 이항분포표

이럴 경우 확률변수 X는 '시행횟수 n이 3이고 성공률 p가 1/2인 이항분포를 따라 간다'고 말한다.

• 일반적인 n과 p에 대해서 확률을 구하는 문제가 남아 있는데 식 (8.2)과 같이 구한다.

$$P(X=x)=nCx\,p^{x}(1-p)^{n-x}, x=0,1,2,\cdots,n \qquad (8.2)$$

여기서 n은 시행횟수, 그리고 x 는 성공횟수를 의미한다. 예에서 n은 3이고 p는 0.5였기 때문에 $P(X=1)$은

$$P(X=1)=3\cdot 0.5^{1}(1-0.5)^{2} = 3\cdot(0.5)^{3} = 3/8$$

이 된다.

8개의 결과들 중에서 3개가 하나만 1이고 나머지는 0인 결과인 것이다. 하나의 결과만 보았다면 확률이 1/8이지만 이런 것들이 3개가 있지 않은가? 그러므로 경우의 수를 곱하여야 한다.

$nCx = \frac{n!}{x!(n-x)!}$ 은 n 개 중에서 x 개를 뽑는 경우의 수를 계산하는 방법이다.

이런 계산방법은 엑셀과 같은 스프레드시트 프로그램에 이미 내장되어 있다. 일반적인 n과 p에 대해서는 엑셀에서

=binomdist(k,n,p,cum)

을 이용하여 구하면 된다. 첫 번째 인자는 성공횟수이다. 마지막 인자 cum이 1의 값을 가지면 k보다 작거나 같은 확률을 보여주고 0의 값을 가지면 확률변수의 값이 k일 확률을 보여준다. n=1이면 베르누이 분포를 뜻할 것이다.

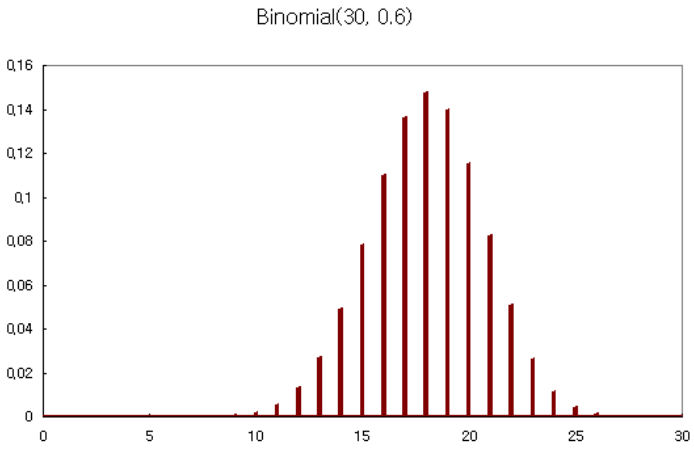
확률변수 X가 n과 p를 가지고 있는 이항분포를 우리는 통상 Binomial(n, p) 또는 B(n, p)로 표시한다. 이항분포의 모양은 n과 p에 따라 다른 형태를 가진다. 확률변수 X의 평균과 분산은 각각 식 (8.3)과 같다.

$$E(X) = np$$
$$Var(X) = np(1-p) \quad (8.3)$$

왜냐하면 X 는 베르누이 분포를 따르는 n개의 독립인 똑같은 확률변수의 합이기 때문에 베르누이분포에서 n개의 평균과 n개의 분산을 다 더하면 된다. 평균과 분산은 각각 p와 1-p로 알고 있는데 이러한 것이 n개 있는 것이다.

- 어느 손전등에 들어가는 한 개의 전지는 계속 켠 상태로 8시간이 지나도 계속 작동을 하면 성공이라 부르고 그렇지 않으면 실패라 하자. 그렇다면 성공일 확률은 0.6이다. 만약 이러한 손전등이 30개가 있다고 보자. n이 30이고 p가 0.6이므로 평균은 30×0.6=18, 표준편차는 $\sqrt{(30)(0.6)(0.4)} = 2.68$ 이 된다.

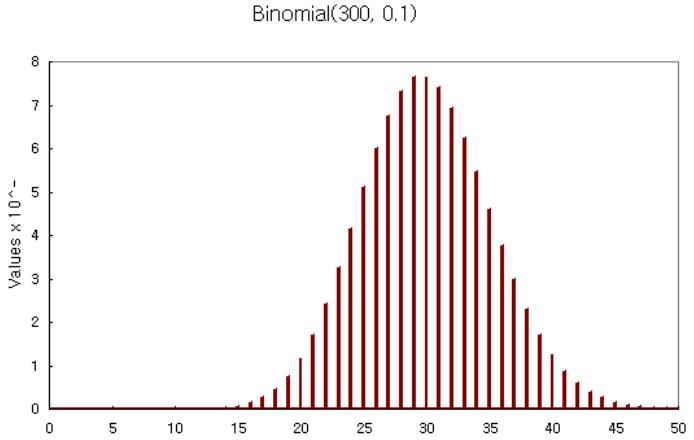
평균적으로 8시간 후에는 18개만 작동을 한다. 그리고 대략적으로 이야기하면 작동하는 손전등은 최소 18-3×2.68 =9.96개인 약 10개 정도가 될 것이다. 물론 운이 좋으면 18 + 3 × 2.68 = 26.04 개인 약 26개가 작동을 할 것이다. n=30, p=0.6인 이항분포 확률함수를 그려보면 [그림 8.1]과 같다.



[그림 8.1] n=30, p=0.6인 이항분포

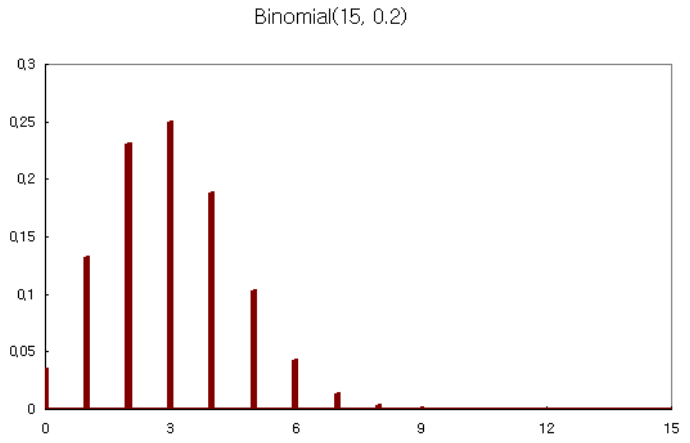
위의 그림은 기댓값 18(=30×0.6)을 중심으로 좌우 대칭의 모양새가 나타났다. 그러나 자세히 보면 정확한 좌우 대칭은 아님을 알 수 있다. 아직은 기댓값 18을 중심으로 우측이 좌측에 비해 질량이 조금 더 모여 있음을 알 수 있다. 구간 10에서 26에서 대부분의 성공횟수가 기록된다. 8시간 후에 30개의 전구가 모두 살아 있을 확률은 매우 적음을 그림으로 확인할 수 있다.

그러나 이러한 그림은 n이 충분히 커지면 p값에 상관없이 좌우대칭인 모양으로 바뀐다. 후에 언급할 정규분포와 밀접한 관계가 있음을 짐작하는 그림이다. [그림 8.2]는 n=300, p=0.1인 경우이다.

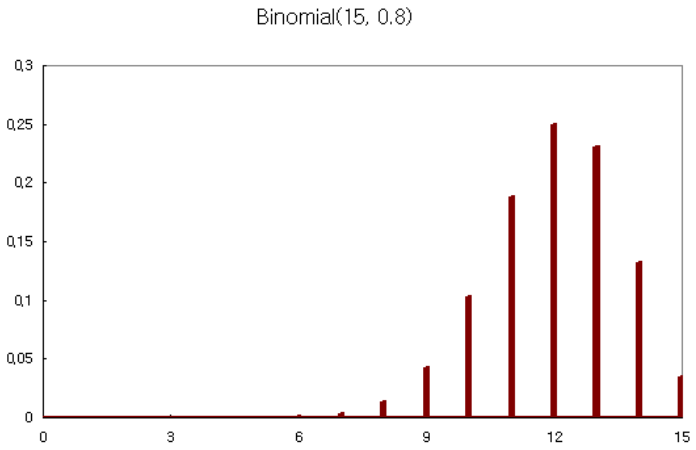


[그림 8.2] n=300, p=0.1

n이 작으면 이항분포는 n과 p에 따라 모양이 달라진다. p=1/2이면 정확하게 좌우대칭이다. 그러나 p가 1/2보다 크면 왼쪽으로 왜도가, 작으면 오른쪽으로 왜도가 발생한다. n=15이고 p=0.2인 경우와 n=15, p=0.8인 두 이항분포를 [그림 8.3]에 그려보았다.



n=15, p=0.2



n=15, p=0.8

[그림 8.3] p를 달리할 때 왜도 발생

이항분포는 독자들이 생각하는 것보다 매우 응용성이 뛰어난 분포이다. 단 이항분포는 독립인 베르누이 시행의 합으로 이해를 하여야 하므로 이러한 독립성에 대한 가정에 대해서는 주의를 기울여야 한다. 몇 가지 예를 들어 알아보도록 하자.

예제 8.3 역학조사의 기초

어느 의사가 전자파에 많이 노출된 (컴퓨터에 일정시간 이상 앉아 있는) 여성근로자 20명이 결혼 후 첫 아이를 낳았을 때 20명 중 3명만이 아들을 낳아 의문이 있다고 가정하자. 이러한 사건(아들이 3명 이하로 나올 사건)이 발생될 확률이 아주 적다면 우리는 당연히 전자파에 대한 유해성을 검정하여야 할 것이다. 이러한 문제가 역학검사의 기초이다.

확률을 구하여 보도록 하자. 여기서 아들을 낳을 확률은 편의상 1/2로 한다. 이 확률은 모든 여성에게 적용되는 확률이다. 그리고 여성근로자 20명은 독립적으로 아이를 낳기 때문에 이항분포를 적용하여 문제에서 요구하는 확률을 구하여 보면 다음과 같다.

$$P(X \leqq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

위의 확률은 0.0012884로 아주 작게 나온다. 엑셀 명령문에서

=binomdist(3,20,0.5,1)

로 구한다. 이러한 확률이 작은지 여부에 대한 의사결정은 물론 사람에 따라 다르지만 모두가 수긍할 만큼 매우 작게 나온다면 우리는 전자파가 영향을 미쳤다고 이야기할 수 있을 것이다. ■

예제 8.1(계속) 예제 8.1로 돌아가 보도록 하자.

로스앤젤레스에는 그 당시 2백만 쌍(couple)이 살고 있다고 추정되어진다. 따라서 위에서 언급한 쌍이 적어도 하나 이상 도시에 있다고 존재할 확률, $P(X \geq 1)$은

P($X \geq 1$ | n=2백만, p=1/12,000,000)

으로 계산하면 약 15.35%가 된다. 그러면 판사는 무슨 확률에 의거하여야 하는가? 잘 생각하여 보자. <판결.xls>

- 판사는 로스앤젤레스에 한 쌍 이상이 존재하는 조건하에서 두 쌍 이상이 존재할 조건부 확률, $\mathrm{P}(\mathrm{X} \geq 2 | \mathrm{X} \geq 1)$ 을 구해야 한다.

만약 이 확률이 작다면 이 쌍은 유죄의 혐의가 있다고 인정하겠지만 그렇지 않고 이 확률이 무시하지 못할 정도로 크다면 이 쌍에게 유죄 판결을 내리는 것은 신중하여야 할 것이다. 단순한 p의 값을 보면 아주 작지만 2백만이라는 쌍이 존재하는 이상 판사는 이러한 사실을 고려한 좀 더 의미가 있는 확률에 의존하여 판단하여야 한다.

이 확률은 약 8%가 나온다. 이는 물론 작기는 하지만 그렇다고 무시할 수 있는 숫자가 아니다.

- **실제 판결:** 판사는 8% 란 숫자가 이 사건의 판결을 유죄로 판정을 하기에는 너무 큰 숫자로 판단하였다. ■

예제 8.4 국가보훈처도 생명표를 참조한다.

2년에 한번씩 통계청에서는 생명표를 발간한다. [표 8.3]에는 2006년 기준 연령별(각 세별) 사망확률(전체, 성별), 생존자수 (전체, 성별), 기대여명(전체, 성별)이 기록되어 있다. 편의상 0-7세, 70-77세, 90-100세 이상만 표시하였다. <생명표.xls>

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	완전생명표(각세별)												
2													
3	각세별	2006											
4		* 사망확률(전체)	* 생존자(전체) (명)	* 정지인구(전체) (명)	* 기대여명(전체) (년)	사망확률(남자)	생존자(남자) (명)	정지인구(남자) (명)	기대여명(남자) (년)	사망확률(여자)	생존자(여자) (명)	정지인구(여자) (명)	기대여명(여자) (년)
5	0세	0.00454	100,000	99,619	79.18	0.00491	100,000	99,585	75.74	0.00415	100,000	99,656	82.36
6	1세	0.0004	99,546	99,526	78.54	0.00042	99,509	99,488	75.11	0.00039	99,585	99,566	81.71
7	2세	0.00031	99,506	99,490	77.57	0.00033	99,467	99,451	74.14	0.0003	99,547	99,532	80.74
8	3세	0.00025	99,474	99,462	76.59	0.00026	99,435	99,422	73.17	0.00023	99,517	99,505	79.76
9	4세	0.0002	99,450	99,440	75.61	0.00021	99,409	99,399	72.18	0.00019	99,494	99,484	78.78
10	5세	0.00018	99,430	99,421	74.63	0.00019	99,388	99,379	71.2	0.00017	99,474	99,466	77.8
11	6세	0.00017	99,412	99,403	73.64	0.00018	99,369	99,360	70.21	0.00016	99,457	99,449	76.81
12	7세	0.00015	99,395	99,387	72.65	0.00017	99,351	99,343	69.23	0.00014	99,442	99,435	75.82
13	70세	0.01949	81,129	80,338	14.59	0.02885	73,675	72,612	12.62	0.01235	88,643	88,095	15.89
14	71세	0.02148	79,547	78,693	13.87	0.03161	71,549	70,418	11.98	0.01404	87,548	86,933	15.08
15	72세	0.02361	77,839	76,920	13.16	0.03466	69,287	68,087	11.36	0.01587	86,319	85,634	14.29
16	73세	0.02608	76,001	75,010	12.47	0.03815	66,886	65,610	10.75	0.01807	84,949	84,181	13.51
17	74세	0.02914	74,019	72,940	11.79	0.04231	64,334	62,973	10.15	0.02091	83,413	82,541	12.75
18	75세	0.03286	71,862	70,681	11.13	0.04724	61,612	60,157	9.58	0.02438	81,669	80,673	12.01
19	76세	0.03721	69,500	68,207	10.49	0.05305	58,702	57,145	9.03	0.02835	79,678	78,548	11.3
20	77세	0.04205	66,914	65,507	9.88	0.05949	55,588	53,934	8.51	0.0327	77,419	76,153	10.62
21	90세	0.16348	19,664	18,056	4.36	0.18745	11,796	10,690	3.98	0.15583	26,156	24,118	4.49
22	91세	0.17662	16,449	14,996	4.11	0.19978	9,584	8,627	3.79	0.16965	22,080	20,207	4.22
23	92세	0.19003	13,544	12,257	3.88	0.21214	7,670	6,856	3.61	0.18373	18,334	16,650	3.98
24	93세	0.20362	10,970	9,853	3.68	0.22444	6,043	5,364	3.45	0.19792	14,966	13,485	3.77
25	94세	0.21729	8,736	7,787	3.49	0.23656	4,686	4,132	3.3	0.2121	12,004	10,731	3.57
26	95세	0.23093	6,838	6,048	3.32	0.24843	3,578	3,133	3.16	0.2261	9,458	8,389	3.4
27	96세	0.24441	5,259	4,616	3.16	0.25992	2,689	2,339	3.05	0.23975	7,319	6,442	3.24
28	97세	0.25761	3,974	3,462	3.03	0.27094	1,990	1,720	2.94	0.2529	5,565	4,861	3.11
29	98세	0.27042	2,950	2,551	2.9	0.28139	1,451	1,247	2.85	0.26537	4,157	3,606	2.99
30	99세	0.2827	2,152	1,848	2.79	0.29116	1,043	891	2.76	0.27699	3,054	2,631	2.9
31	100세이상	1	1,544	4,167	2.7	1	739	1,992	2.69	1	2,208	6,211	2.81

[표 8.3] 생명표

- 생명표를 편하게 읽도록 100,000명을 기준으로 하여 사망확률에 따른 명수를 기록하였다. 예를 들어 0세에서 인구가 100,000명이라고 한다면 0세 때의 사망확률이 0.00454 이므로 1세 때에는 100,000 × (1-0.00454) 인 99,546명이 있을 것이다.

이를 남녀로 구분하여 기록한 열도 확인할 수 있다. 그리고 기대여명은 각 연령별로 얼마나 수명이 남아 있는지 평균을 계산한 결과이다.

- 예를 들어 0세 때의 기대 수명은 남녀를 구분하지 않는다면 79.18세이다. 이 숫자가 신문에서 이야기하는 우리나라의 평균 수명이다. 통계청에서 발표하는 평균 수명은 아이가 태어난다면 평균적으로 앞으로 살아 갈 것인가 하는 숫자인 것이다. 물론 다른 연령별에서 앞으로 얼마나 내 수명이 평균적으로 남아 있는지 알아볼 수 있다. 이러한 평균은 조건부 기댓값의 개념이다. 표에 의하면 만약 현재 나이가 70세라면 앞으로 남은 수명은 14.59년이 된다.

이런 생명표는 수많은 정책 자료로 활용이 되는데 한 예를 보자.

- 국가보훈처에서는 국가보훈대상자를 직급별 나이별로 관리를 하고 있는데 현재 5급에 해당하는 국가 유공자가 5,000명이 있다고 보자. 그 중에서 70세의 나이를 가지고 있는 국가 유

공자가 1,000명이라면 내년에 71세의 5급 국가 유공자의 평균 명수는 1000 × (1 - 0.01949) = 약 981명이 되는 것이다. 그리고 2년 후에는 72세의 5급 국가 유공자의 평균 명수는 981 × (1-71세의 1년 후 사망확률)을 구하면 될 것이다. 여기서 71세의 1년 후 사망확률은 생명표에서 확인하면 0.02148로 70세 때의 사망확률보다 약간 높다.

물론 국가유공자는 일반인보다 평균수명이 짧을 가능성이 있기 때문에 이들이 세상을 떠난 나이와 일반인들의 평균을 비교해 적절한 확률을 배정하면 된다. 예를 들어 5년 정도 일찍 생을 마감하였다면 70세 때의 사망확률보다는 75세 때의 사망확률을 가지고 남아 있는 국가유공자의 수를 구하면 된다.

모든 직급에 대해 이와 같은 방법론을 적용하여 향후 10-20년간 국가유공자들의 인원을 파악하고 이에 적절한 예산 규모를 기획하는 것은 국가보훈처가 할 일이다. 물론 이 과정에서 생명표도 적절하게 예측을 하여야 한다. 왜냐하면 연령별 사망확률은 선진국으로 진입함에 따라 점점 감소하고 있기 때문이다.

독립을 요구하는 베르누이 시행(살고 죽음)이 많은 국가 유공자에게 일어나는 것이다. ■

예제 8.5 **생명표**

다음은 2004.12.12 중앙일보 기사에서 발췌한 내용이다. ■

"난 여든다섯이네."

"나는 아흔셋이야. 하지만 신경 쓰지 마시게. 요즘 누가 나이를 따지나."

10년 뒤면 이런 인사 차림이 흔해질 것 같다. 통계청이 발표한 '2002년 생명표'를 보면 평균 수명이 꽤 늘어나 장수를 복으로 치던 시절은 갔지 싶다. 이제는 시간에 따른 나이보다는 '생물학적 나이'가 더 중요해졌다. 얼마나 오래 살려져 있느냐보다 얼마나 건강하고 즐겁게 만족하며 사느냐가 사람들 관심사가 됐다.

우리보다 노인 문제를 앞서 겪은 서구에서는 '적극적 노화'라는 새 단어를 만들었다. 늙어가는 과정을 단순히 죽음을 늦추려는 안간힘이 아니라 정력적인 헌신과 성장의 기간으로 바꾸려는 노력을 일컫는 말이다. 일흔 된 할머니가 '사별한 자유발랄한 몸, 생물학적 나이 40세'라는 애교 섞인 '애인 구함' 광고를 내는 시대가 올 날도 멀지 않았다.

그렇다고 장밋빛 전망만 펼쳐지는 건 아니다. 수명은 늘었지만 정년은 안 늘어나는 괴리 때문이다. '노세 노세, 젊어서 노세'가 아무리 좋다 해도 빌어먹을 지경이 되는 장기 휴가는 오히려 고통이 될 수 있다. 방치된 노인이 공원이나 거리를 어슬렁거리고 그들을 '과잉 생존자'라고 비아냥거리는 사회는 제대로 굴러갈 수 없다.

엊그제 땅으로 돌아간 전우익(1925~2004) 선생은 장년이 넘어서는 경북 봉화군 상운면 구천리 낡은 옛 집에서 밭농사를 지으며 혼자 살았다. 그는 "가을의 낙엽에서는 버림, 청산을 결행하고 겨울의 얼어붙은 솔잎에서는 극한의 역경에서도 끝내 지켜야 할 것은 지키라는 것을 배웠다"고 썼다. 단풍과 지는 해가 산천을 아름답게 물들이는 것을 보면서 "인생의 마지막을 저렇게 멋지게 마치진 못할망정 추접하게 마치지는 말아야 하는데"라고도 적었다. 그가 쓴 편지모음 '혼자만 잘 살믄 무슨 재민겨'에는 이렇듯 홀로 익힌 '깊은 산속의 약초' 같은 이야기들이 그득 담겨 있다.

그는 삶을 제대로 이루는 일에 관심이 많았다. "제대로 이루어진다는 건 자연의 운행과 역사의 과제에 충실한 삶을 사는 건데, 세상의 흐름은 자연과 멀어지고 역사보다는 순간과 개인적인 삶으로 오그라드는 것 같습니다"라며 요즘 세상살이를 안타까워했다. 자신을 전우익이란 이름보다 무명씨를 뜻하는 '언눔'이라 불러달라던 그의 한마디가 겨울 바람처럼 서늘하게 우리 가슴을 베고 지나간다. '언눔네들, 오래만 살믄 무슨 재민겨'.

생명표가 사회복지에 미치는 단상에 대해

예제 8.6 펀드의 수익률광고는 얼마나 믿어야 하나?

뮤츄얼 펀드의 마케팅에는 항상 확률이 등장한다?

- 어느 뮤츄얼 펀드를 판매하는 회사가 "우리는 지난해(52주) 동안 37번이나 시장 인덱스(index)를 앞서는 수익률을 기록했다고 선전한다." 이러한 결과가 우연치 않게 나온 것인지 아니면 꾸준한 노력의 결과인지 알아보도록 하자. <뮤츄얼펀드.xls>

여기서 시장을 우연치 않게 앞질렀다는 기준으로는 n=52 , p=0.5인 이항분포를 사용하도록 한다. 왜냐하면 우연치 않게 시장을 이긴다는 것은 동전 던지는 실험의 결과로 앞면(혹은 뒷면)이 나오는 것과 같다고 볼 수 있기 때문이다.

이러한 이항분포를 따르는 확률변수가 37번 이상 시장보다 나은 수익률을 낼 확률은

$$\mathrm{P}(\mathrm{X} \geq 37|\mathrm{n}=52,\mathrm{p}=0.5) \;=\; 0.00159$$

가 나올 것이다. 단일 뮤츄얼 펀드가 37번 이상 시장을 이길 확률은 그렇게 높지 않다. 엑셀 명령문으로는 다음과 같이 입력하면 된다.

=1-binomdist(36,52, 0.5,1)

- 그러나 잠시 생각을 넓혀보자. 시장에는 이런 뮤츄얼 펀드가 굉장히 많다.

가령 400개 있다고 보자. 그러면 개개의 뮤츄얼 펀드가 52주 중에서 37번 시장을 이길 확률은 0.00159이지만 400 개중에서 그런 회사가 하나 이상 나올 확률은

$$1-\mathrm{P}(\mathrm{X}=0|\mathrm{n}=400,\mathrm{p}=0.005919)$$

무려 0.471이 나온다. 엑셀 명령문으로는 다음과 같이 입력하면 된다.

=1-binomdist(0, 400, 0.00159,1)

따라서 400개의 뮤츄얼 펀드가 시장에 있다면 적어도 하나의 뮤츄얼 펀드가 "37 out of 52" 보고서를 내는 것은 그렇게 놀랄 일도 아니다.

그러나 아래 [표 8.4]를 참고하면 뮤츄얼 펀드가 600개가 있는 경우 "40 out of 52" 보고서를 낼 하나의 회사라도 존재할 가능성은 불과 0.038뿐이다.

뮤츄얼 펀드개수	시장을 이긴 주(week)의 횟수				
	36	37	38	39	40
200	0.542	0.273	0.113	0.040	0.013
300	0.690	0.380	0.164	0.060	0.019
400	0.790	**0.471**	0.213	0.079	0.025
500	0.858	0.549	0.258	0.097	0.031
600	0.904	0.616	0.301	0.116	**0.038**

[표 8.4] 펀드개수와 이긴 주에 따른 확률변화

독자들은 이 경우 600개의 뮤츄얼 펀드의 예측능력은 같다고 가정을 하고 계산을 한 것임을 명심하기 바란다. ■

예제 8.7 **항공사는 어떤 방법으로 예약을 관리하는가?**

비행기의 예약 시스템을 이해하면 통계학의 쓰임새를 안다.

- 항공편을 예약해 놓고 나타나지 않는 경우를 no-show라 한다. 따라서 200좌석의 항공기를 운행한다 하더라도 200좌석 보다 많은 예약을 받는 것이 허용되어 있다. 지금까지 이 항공기를 이용하는 승객의 10%는 no-show 손님이다. 과다예약을 받으면 비행기 좌석이 채워질

가능성은 높아지나 200명 보다 많은 손님이 나타날 확률 또한 높아질 것이다.

만약 215개의 좌석이 예약되었다면 다음과 같은 확률이 어떻게 계산이 되는지 계산하여 보았다. 즉, 이 문제는 215개의 티켓을 가지고 있는 승객들을 n, 그리고 개개의 승객들이 공항에 나타날 확률 p를 0.90으로 보는 이항분포의 문제로 생각한다면 이러한 확률은 쉽게 구할 것이다. <항공사.xls>

(1) 205명 보다 많은 사람이 나타날 확률
(2) 200명 보다 많은 사람이 나타날 확률
(3) 적어도 195좌석이 채워질 확률
(4) 적어도 190좌석이 채워질 확률을 구해 보자.

4개의 확률은 공항에 나타날 승객의 수를 X 라고 한다면 다음과 같이 정리가 된다.

(1) $P(X > 205 | n = 215, p = 0.90)$
(2) $P(X > 200 | n = 215, p = 0.90)$
(3) $P(X \geq 195 | n = 215, p = 0.90)$
(4) $P(X \geq 190 | n = 215, p = 0.90)$

한편 이항분포의 독립성은 쉽게 만족이 되지만 p는 개개의 승객이 공항에 도착하는 날 교통 혼잡도에 따라 달라질 수 있으나 동일하다고 가정하는 것이다.

발행된 티켓을 215석으로 기준으로 하였을 때 위의 네 가지 확률은 각각 0.001, 0.050, 0.421, 0.820 이다. 그리고 발행된 티켓의 수를 206석에 233석까지 변화시켰을 때 이 4가지 확률이 어떻게 변하는지 민감도 분석을 시행하였다.

발행된 티켓 수	More than 205 show up	More than 200 show up	At least 195 seats filled	At least 190 seats filled
	0.001	0.050	0.421	0.820
206	0.000	0.000	0.012	0.171
209	0.000	0.001	0.064	0.384
212	0.000	0.009	0.201	0.628
215	0.001	0.050	0.421	0.820
218	0.013	0.166	0.659	0.931
221	0.064	0.370	0.839	0.978
224	0.194	0.607	0.939	0.995
227	0.406	0.802	0.981	0.999
230	0.639	0.920	0.995	1.000
233	0.822	0.974	0.999	1.000

[표 8.5] 항공회사 예약문제

이 모형은 관련된 비용 및 수익을 고려하여 얼마만큼 예약수가 최적인지를 결정하는 의사결정 문제로 발전시킬 수 있으며 수익률모형이라 한다. 호텔 예약시스템의 경우도 마찬가지이다. 왜냐하면 하룻밤이 지나면 공실은 수익을 창출하지 못하기 때문이다. 둘 다 공좌석 및 공실을 최소화하고 비용을 줄이는 문제이기 때문이다. 수익률모형에서 기본적으로 쓰이는 확률분포는 이항분포이다. ■

예제 8.8 **당선자 확정발표는 어느 시점이 좋은가?**

- 방송사는 당선자를 빨리 알고 싶어 하는 심리를 충족키 위해 당선 확정을 발표한다. 우리는 이를 확률적으로 이해할 필요가 있다. <선거.xls>

어느 선거에서 두 후보 A, B가 출마하였고 1,000명의 개표가 이루어 졌다고 가정하자. 그리고 현재까지 540명이 A 후보를 지지하였다고 하자. 궁극적으로 B 후보가 선거에서 이길 것으로 알려져 있다면 현재 이런 결과가 나올 가능성이 얼마나 되는지 알아보도록 하자. 만약 이런 가능성이 아주 낮다면 B 후보가 선거에서 이긴다는 가정은 하지 않아도 된다. 왜냐하면 후보 B가 이긴다는 가정 하에서 이런 결과는 거의 나오지 않기 때문이다.

A 후보를 지지하는 비율이 궁극적으로 49%일 때 이런 결과가 나올 가능성은 겨우

$$1-\mathrm{P}(\mathrm{X} \leq 539|\mathrm{n}=1000\mathrm{p}=0.49) \text{ =0.0009}$$

이 나온다. 엑셀명령문으로 다음과 같이 입력하면 된다.

=1-binomdist(539, 1000, 0.49,1)

아래 표에서는 A 후보를 지지하는(어느 경우라도 궁극적으로는 B 후보가 이김) 확률을 조금씩 높여 이러한 확률을 구해 보았다.

A 후보를 지지하는 확률	80표를 이길 확률
0.490	0.0009
0.492	0.0013
0.494	0.0020
0.496	0.0030
0.498	0.0043
0.499	0.0052

[표 8.6] 선거당선 발표에 필요한 확률

A 후보를 지지하는 비율이 49.9%(그래도 후보 B가 승리한다)라도 현재 80표(540표 -460표)차가 나올 가능성은 확률은 겨우 0.0052가 된다. 방송사는 후보 A가 승리하였다고 해도 무방하다. 겨우 1,000표를 개표하여도 방송사는 당선자 확정을 선언하여도 될 만큼 확률이 작아진다. 투표자의 수와 상관없이 불과 1,000표만 가지고 있어도 충분하다. 단 투표되는 상황이 변하지 않는다는 조건이 있지만 말이다. ■

이상에서는 우리가 제일 자주 접하는 이산형의 제일 중요한 분포인 이항분포를 알아보았다. 다음 절에서는 연속형 분포의 개념과 몇 가지 중요한 분포들에 대해 알아보자.

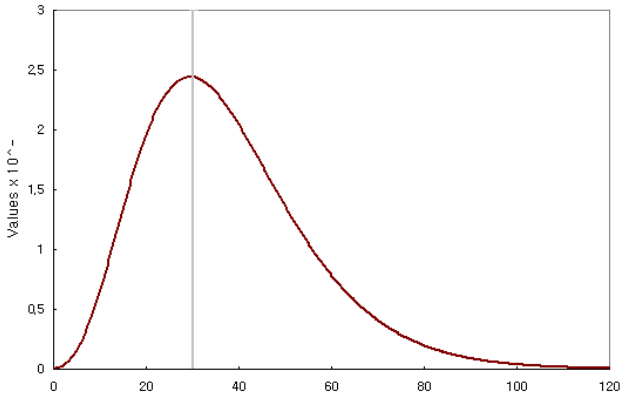
8.2 연속형분포

이산형 확률변수는 가질 수 있는 값이 정해져 있고 해당하는 확률로서 그 성격을 규정지을 수 있으나 연속형 확률변수는 그렇지 않다. 연속형 확률변수는 가질 수 있는 값의 리스트 대신에 변수가 가질 수 있는 값들의 연속체(continuum)를 명시하여야 한다. 예를 들어 0부터 100까지의 어떤 임의의 값도 가질 수 있는 특징을 어느 변수가 지녔다면 이러한 0부터 100까지의 모든 값들을 연속체라 한다. 우리는 0과 100까지의 구간에서 값들을 연속체로 처리하고 0과 100까지의 구간에 전체 크기가 1인 확률을 펼쳐(spread)주는 작업을 해야 한다. 이러한 펼쳐주는 작업을 가능케 하는 함수를 통계학에서는 확률밀도함수(probability density function)라 한다. 확률밀도함수는 일종의 히스토그램의 역할을 한다.

- 확률변수 X의 확률밀도함수 $f(X)$의 높이가 높을수록 x 가 나올 가능성이 높다.
- 확률밀도함수 $f(X)$와 수평축 사이의 면적은 1이 되어야 한다.
- 또한 $f(X)$는 확률변수 X의 모든 가능한 값에 대해서 음이 아니어야 한다.

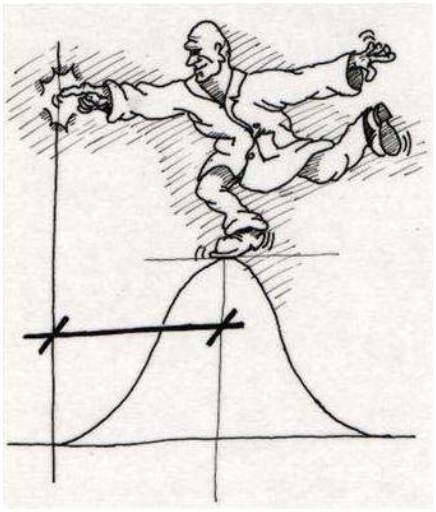
이해를 돕기 위해 하나의 확률밀도함수를 그려 보자! [그림 8.4]에서는 확률변수 값은 [0, 120] 구간에서 값을 가지고 30에 가까운 값들이 나올 가능성이 제일 높다. 그러나 확률밀도함수로부터 확률을 이야기하기 위해서는 미적분이 필요하다. 왜냐하면 확률변수 값이 20에서 60사이에 나타날 확률 등은 면적으로 구해야 하기 때문이다. 전체 면적을 1로 설정하면 미적

분으로 확률을 구할 수 있기 때문이다. 그러나 실제로 미적분의 도움을 받아 자료 분석을 하는 경우는 거의 없다. 따라서 미적분의 문제는 개념의 문제이지 현재 이 책을 읽는 독자들은 걱정을 하지 않아도 될 것이다.



[그림 8.4] 확률밀도함수 $f(x)$의 한 형태

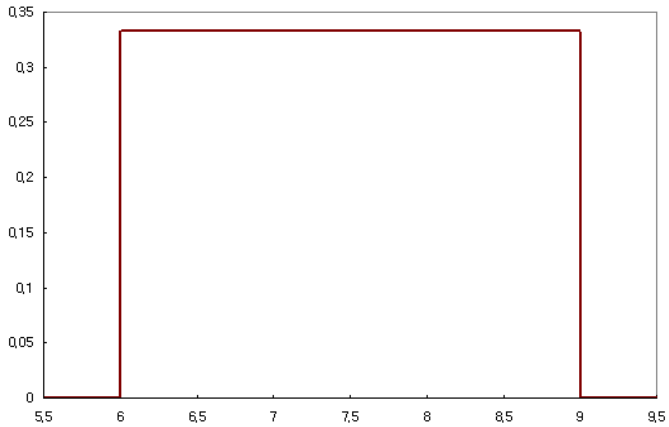
그리고 역시 평균과 표준편차와 같은 값들도 미적분을 통해서 구하지만 이를 구하는 것이 본 책의 목적이 아니므로 필요하면 구해진 공식을 인용하는 수준에서 쓰면 된다.



대칭분포이면 평균은 가운데 값이다.

8.2.1 균일분포란 무엇인가?

- 예를 들어 조달되는 볼트 부품 한 개당 생산비용에 대해 우리가 알고 있는 지식은 비용은 6원과 9원 사이에서 결정이 된다는 사실 뿐이고 6원과 9원 사이에서 어떤 값이 나올 가능성이 제일 높은지 전혀 사전지식이 없다면 [그림 8.5]와 같은 확률밀도함수를 만들 것이다.



[그림 8.5] 균일분포

이와 같이 주어진 구간에서 임의의 값을 가질 가능성이 같은 경우, 확률변수는 균일분포(uniform distribution)를 따른다고 한다. 구간이 [a,b]라면 기댓값과 분산은 식 (8.4)와 같이 계산된다.

$$E(X) = \frac{(a+b)}{2}$$
$$Var(X) = \frac{(b-a)^2}{12} \tag{8.4}$$

먼저 평균은 최대값과 최소값을 더해 2로 나누면 된다. 왜냐하면 균일분포는 대칭이므로 평균은 정 가운데 위치하기 때문이다. 그리고 분산은 이보다 조금 복잡하지만 최대값에서 최소값을 뺀 다음 제곱을 하고 12로 나누면 된다.

- 그림에서 $f(x)$ 의 높이는 1/3이 되어야 총 면적이 1이 된다. 그리고 평균, 분산 및 표준편차를 구해보면 각각

$$E(X) = \frac{(9+6)}{2} = 7.5$$
$$\sigma_X^2 = \frac{(9-6)^2}{12} = 3/4$$

$$\sigma_X = \sqrt{3/4} = \sqrt{3}/2 = 0.866$$

가 된다.

이와 같은 균일분포는 값이 나오는 구간의 범위를 알고 있으나 구간 내에서 값들이 가지는 행태에 대해서는 정보가 없을 때 사용하는 분포이다. 즉, 확률변수 값들이 가질 수 있는 최소값과 최대값에 대한 정보만 필요하다. 그러나 최소값 및 최대값 이외의 다른 정보도 가지고 있다면 아래 설명하는 삼각형 분포나 정규분포를 생각하여야 한다.

8.2.2 삼각형분포란 무엇인가?

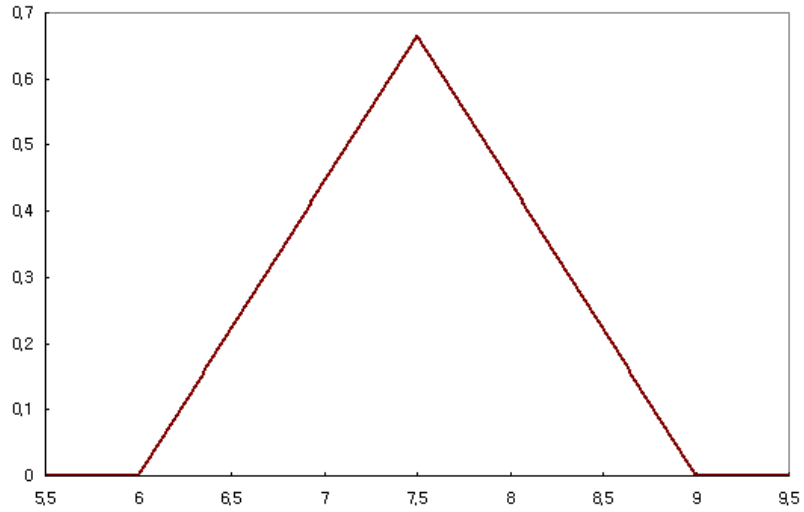
- 균일분포의 응용성은 매우 뛰어나나 대부분의 경영환경에서는 균일분포를 가정하기 위해 필요한 정보보다 더 많은 정보를 가지고 있다. 위에서 언급한 볼트 단위당 생산비용을 균일분포로 가정하는 것은 그렇게 흔하지 않다. 왜냐하면 그 구간에서 비용이 발생한다는 것만 알고 있을 뿐 구체적인 비용구조는 모른다고 설정하는 것은 현실적으로 맞지 않기 때문이다.

예제 8.9 **많은 사람들은 삼각형분포를 위한 모든 정보를 가지고 있다.**

많은 관리자에게 물어보면 다음과 같이 답하는 경우를 흔하게 볼 수 있을 것이다.

- "단위당 가격은 7원 50전일 가능성이 매우 높다. 그리고 가능성은 희박하지만 만의 하나 잘못된다면 단위당 가격은 9원까지도 올라갈 수 있고 반대로 어떤 경우에는 6원까지도 내려갈 수 있다."

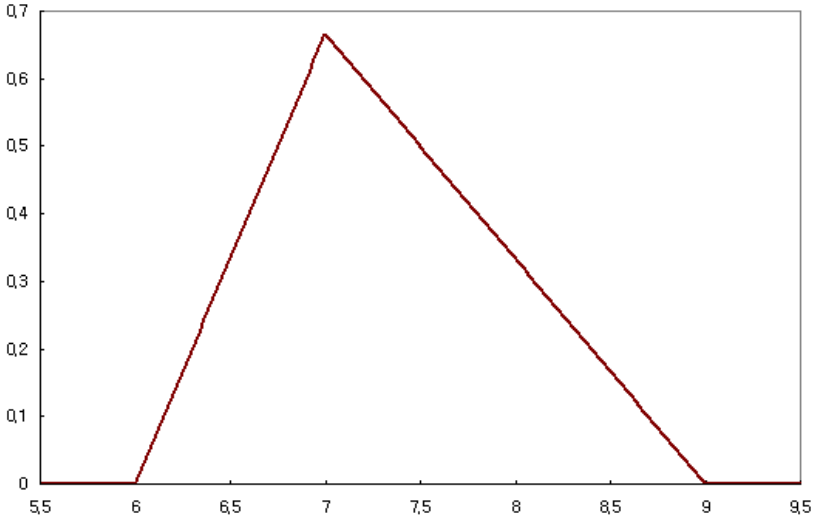
이런 경우 필요한 것이 삼각형분포(triangular distribution)이다. 그 가능성이 7원 50전에서 가장 높고 직선으로 양쪽 끝으로 감소한다면 [그림 8.6]과 같이 분포를 그릴 수 있다. ■



[그림 8.6] 단위당 생산비용

삼각형분포는 세 가지 값, 즉, 최소값(minimum), 최빈값(most likely), 최대값(maximum)만 가지고 있으면 그려낼 수 있는 분포이다. 균일분포와 달리 최빈값을 중심으로 최소값과 최대값의 구간에서 확률변수의 값이 나오는 형태를 뾰족한 산의 모양으로 표현할 수 있다.

위 그림에서는 좌우 대칭인 분포를 그렸지만 비대칭인 삼각형분포도 생각하여 볼 수 있다. [그림 8.7]처럼 우측으로 왜도가 발생한 삼각형 분포를 그려보았다. 대칭인 삼각형 분포와 달리 최빈값을 중심으로 기울기가 달라진다. 다양한 삼각형 모양을 단 3개의 값으로 그려낼 수가 있기 때문에 삼각형분포는 실질적으로 매우 유용하게 사용하는 분포이다.



[그림 8.7] 비대칭인 삼각형분포

통상적으로 삼각형 확률분포를 따르는 확률변수 X의 기댓값과 분산은 식 (8.5)와 같다.

$$E(X) = \frac{(a+b+c)}{3}$$
$$\sigma_X^2 = \frac{a^2+b^2+c^2-ab-ac-bc}{18} \quad (8.5)$$

여기서 a는 최소값, b는 최빈값, c는 최대값이다. [그림 8.7]에서 기댓값, 분산, 표준편차를 구해보면

$$E(X) = \frac{(6+7+9)}{3} = 7.33$$
$$\sigma_X^2 = \frac{6^2+7^2+9^2-6\text{x}7-6\text{x}9-7\text{x}9}{18} = 0.389$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{0.389} = 0.624$$

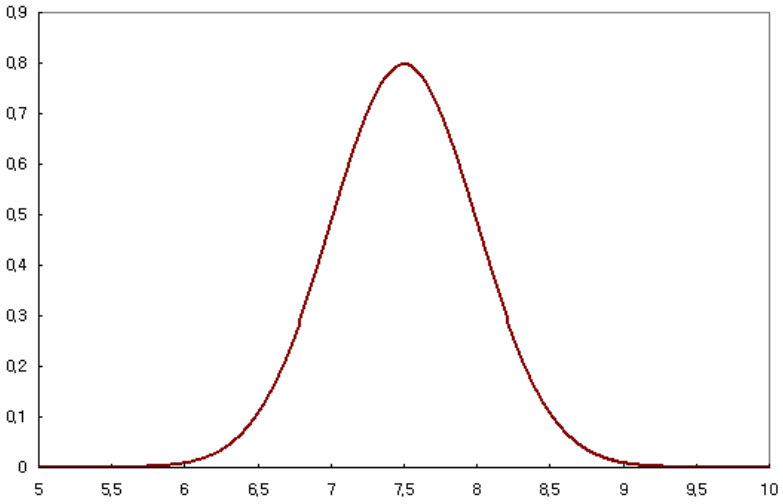
그러나 삼각형분포는 최빈값을 중심으로 최소값과 최대값으로 이어지는 선이 직선이다. 따라서 최빈값을 중심으로 많은 값들이 모여져 있는 상황 하에서는 사용하기가 어렵다. 그리고 모양새가 좌우대칭이라면 삼각형 분포보다 정규분포를 사용한다. 물론 정규분포를 사용하는 이유는 이러한 이유 때문만은 아니고 삼각형분포가 가지고 있지 못한 특성을 반영하기 때문이다.

8.2.3 정규분포란 무엇인가?

자연적인 현상은 정규분포로 모형화가 가능하다.

정규분포는 이론적인 이유로 사용을 하기도 하지만 자연적인 현상을 모형화하거나 현상을 이해하는데 자주 인용되는 분포이다. 어느 한 값을 중심으로 값의 빈도가 많이 나오며 그 값을 벗어나면 자연스럽게 대칭으로 빈도가 줄어드는 경우에는 더욱 그렇다. 우리들의 몸무게, 키 등이 대표적이다. 판매량, 주가 등도 이런 범주에 속한다.

- 위에서 언급한 볼트 제품의 단위당 생산비용이 평균 7원 50전이지만 값을 취할 수 있는 구간에서 좌우 대칭이고 가운데 최빈값을 중심으로 많은 값이 모여져 있다면 삼각형분포는 사용 목적에 맞지 않는다. 이런 경우는 [그림 8.8]과 같은 분포를 생각할 수 있을 것이다.



[그림 8.8] 단위당 생산비용

좌우대칭이긴 하지만 삼각형 분포와 달리 종(bell)모양을 하고 있는 분포이다. 정규분포는 아마 통계학에서 제일 많이 사용되는 분포일 것이다. 독일인 수학자 C. F. Gauss에 의하여 만들어졌다고 해서 가우스(Gaussian)분포라고도 한다.

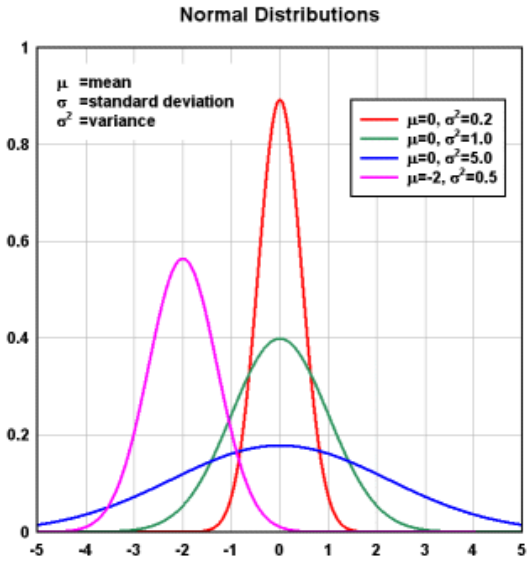
독일의 화폐에는 정규분포가 그려져 있다.

이론적으로 정규분포를 따르는 확률변수 값이 취하는 구간은 음의 무한대부터 양의 무한대까지 모든 실수 구간이다. 그러나 실제적으로는 정규분포를 따르는 확률변수는 좁은 구간에서 값이 발생한다. 그림에서 값이 발생하는 실질적인 구간은 6부터 9까지의 구간이다. 이러한 구간은 정규분포를 규정지어주는 다른 하나의 값, 즉 표준편차 σ에 의해 결정된다. σ의 의미를 먼저 그림에서 찾아보도록 하자.

- **정규분포의 특징:** 정규분포의 확률밀도함수의 변곡점은 중심에서 σ 떨어진 지점에서 발생한다. 이는 그림을 보게 되면 중앙의 곡선은 위로 볼록하다. 그러나 중앙에서 이탈하여 좌우측으로 이동하면 아래로 볼록한 곡선과 접하게 되는데 그 위치가 평균에서 표준편차 하나 만큼 떨어져 있는 곳이라는 뜻이다.

우리는 통상 평균이 μ이고 분산이 σ^2인 정규분포를 $N(\mu,\sigma^2)$ 또는 $N(\mu,\sigma)$로 표기한다.

평균과 분산 (μ,σ^2) 의 값에 따라 수많은 정규분포가 존재한다. 아래 [그림 8.9]를 보면 평균의 값에 따라 정규분포의 위치가 달라지고, 분산의 크기가 달라지면 모양새가 달라짐을 알 수 있다.



[그림 8.9] 평균과 표준편차가 다른 다양한 정규분포

이 중에서 특히 μ, σ 가 각각 0과 1인 경우를 표준정규분포라 한다. 우리가 궁극적으로 필요한 것은 이것 하나뿐이다. 위 그림에서는 y축의 0.4에서 봉우리가 형성된 분포이다. 잠시 후에 왜 하나만 필요한지 알 것이다. 그리고 이러한 표준정규분포를 따르는 확률변수를 특히 Z 라 부른다. 그리고 임의의 (μ, σ) 를 가지고 있는 정규분포를 따르는 확률변수 X와 확률변수 Z 와는 다음과 같은 관계가 성립한다.

• **표준화(standardization)** :

$$Z = \frac{X - \mu}{\sigma} \tag{8.6}$$

즉, X가 평균 μ, 분산 σ^2인 정규분포라면 표준화된 Z 변수의 기댓값은 0이고 분산은 1인 표준정규분포를 따라간다. 예를 들어 X의 값이 $X = \mu + \sigma$라면

$$Z = \frac{(\mu + \sigma - \mu)}{\sigma} = 1$$

로 변환이 가능하다. 만약 $X = 0$ 이면 $Z = 0$ 이고, $X = \mu + 2\sigma$ 라면 $Z = 2$가 된다.

• 어떤 평균 및 표준편차가 되어도 Z는 같은 값을 가진다. 어떠한 평균, 어떠한 표준편차라도 상관없기 때문에 이를 '표준화'라고 이름을 붙였다.

표준화는 매우 중요한 개념이다. 왜냐하면 각기 다른 평균과 표준편차를 가지고 있는 확률변수를 하나의 확률변수로 표준화하여 필요한 정보를 확보하기 때문이다.

예를 보자. 대학교의 한 강좌는 여러 강사에 의해 가르친다. 강사 개개인의 평가 방법에 따라 각반의 성적의 분포는 서로 다를 것이다. 그러나 원래의 점수에서 평균점수를 빼고 표준편차로 나누어주는 작업 처리를 한 표준화 된 점수의 분포는(각 반의 수강학생들의 능력은 비슷하게 분포가 되어 있다고 가정한다면) 매 반별로 차이가 없이 비슷해진다. 표준화된 점수는 양과 음으로 나누어진다. 예를 들어 표준점수가 2인 의미는 점수가 평균으로부터 표준편차 두 개만큼 위로 위치가 되어 있다는 뜻이다. 반대로 -1이라는 의미는 평균으로부터 표준편차 하나만큼 밑으로 점수가 위치되어 있다는 뜻이다. 대학수학능력시험에서 표준화 점수의 사용은 중요한 예가 될 수 있다.

• 표준정규분포표

때때로 확률, $P(Z \geq z)$ 을 구하기 위해 표를 작성할 필요가 있다. 왜냐하면 정규분포를 따르는 확률변수가 어떤 특정구간에서 값을 가지는 확률을 계산하기 위해서는 수치해석적인 방법으로 값을 구하여야 하기 때문이다. 이러한 불편함을 해소하기 위해 만들어진 표가 표준정규분포표(standard normal table)이다. 표를 이용하면 원하는 확률을 매우 편하게 구할 수 있다.

표준정규분포를 따르는 확률변수 Z는 이론적으로 음의 무한대부터 양의 무한대까지의 구간에서 값을 가진다. 그러나 평균이 0이고 표준편차가 1이므로 3.5를 넘어 가는 값을 가지는 경우는 거의 없다. 따라서 확률변수 Z가 값을 대부분 [-3.5, 3.5] 구간에서 가진다고 가정하자.

[표 8.7]이 표준정규분포표이다. 이와 같은 표를 엑셀로 만드는 실습을 하는 것이 바람직하다. 잠시 후에 나올 엑셀명령문을 이해한다면 그리 어렵지 않다. <표준정규분포표.xls>

	A	B	C	D	E	F	G	H	I	J	K
1	z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2	0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
3	0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
4	0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
5	0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
6	0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
7	0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
8	0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
9	0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
10	0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
11	0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
12	1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
13	1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
14	1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
15	1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
16	1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
17	1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
18	1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
19	1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
20	1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
21	1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
22	2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
23	2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
24	2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
25	2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
26	2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
27	2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
28	2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
29	2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
30	2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
31	2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
32	3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
33	3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
34	3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
35	3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
36	3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
37	3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002

[표 8.7] 표준정규분포표, $P(Z \geq z)$

열 A 에 있는 숫자는 z의 첫 번째 소수점 위치에 있는 수이면 행 1에 있는 수는 z의 두 번째 소수점 위치에 있는 수이다. 예를 들면 z=2.36는 셀 H25에 속한다. [표 8.5]에서 하이라이트 된 숫자 0.0091이 확률 $P(Z \geq 2.36)$ 에 해당하는 값이다. 몇 가지 확률을 더 찾아보자.

$$P(Z \geq 3) = 0.0013, \quad P(Z \geq 1.12) = 0.1314$$

이런 표는 일반적인 정규분포를 따르는 확률변수의 확률 값을 구하는데 매우 도움이 되곤 한다.

예제 8.10 신입사원 서류심사 시 TOEIC 점수의 커트라인은?

<<입사기준 토익 평균점수 676점>>

'사무직 692점', '기술직 673점'

신입채용에 필요한 토익성적은 평균 676점인 것으로 조사됐다. 또, 대부분 기업들이 신입채용시 영어능력을 평가하고 있는 것으로 나타났다. 취업·인사포털 인크루트가 주요 대기업 139개사를 대상으로 '채용시 영어 능력 평가현황'에 대해 설문조사한 결과, 입사 지원할 수 있는 기준 토익 점수는 평균 676점인 것으로 드러났다.
직군별로는 사무직 692점, 기술직 643점이다. 점수대별로 살펴보면 사무직군의 경우 700점대 요구하는 기업이 35.5%(27개사)로 가장 많았고 600점대를 28.9%(22개사), 800점 이상 26.3%(20개사), 500점대 9.2% (7개사) 순이었다. 기술직군의 경우도 700점대가 41.0%로 가장 많았으며, 600점대와 미만이 각각 25.6%였다. 800점 이상 고득점을 요하는 기업은 7.7%에 그쳤다.
또한 86.3%인 120개사는 신입사원 채용에 영어 평가를 진행한다고 응답했다. 평가방식은 '일정 점수 이상의 공인어학점수 제출'을 요구하는 기업이 56.8%(79개사)로 가장 많았고, '영어 면접'을 실시하는 기업도 45.3%(63개사)나 됐다. 자체 영어 필기시험을 보는 기업도 12.2% (17개사)였다. 이처럼 대부분 기업들이 채용시 영어 능력을 평가하고 있는 것은 실제 업무에서도 영어활용도가 높기 때문인 것으로 나타났다.
66.9%(93개사)가 '특정부서에서 영어 사용빈도가 높다'고 응답했으며 '전체부서에서 영어를 활용할 일이 많다'는 기업도 21.6%(30개사)나 되는 등 실제 업무에서 영어를 사용할 일이 많다는 기업이 총 88.5%(123개사)에 달했다.

<<주요기업 토익기준>> - 자료제공:인크루트

* '삼성전자'
- 730(인문계),620(연구기술개발직) 토익 자격기준으로만 활용,
영어 회화 면접
* 'LG전자'
- 620점 이상 토익 점수 반영, 영어프레젠테이션 및 토론면접 실시
* '한국전력공사'
- 750(사무직), 600(기술직) 토익 자격 기준으로 활용
* '두산그룹'
- 500점 이상 기존 700점에서 500점으로 하향 조정
* '팬택' 의무 제출 아님 영어토론면접 및 영어 단답식 면접 실시
* 'LG엔시스'
- 600점 이상 토익 자격기준으로만 활용,고득점인 경우 노력 및 성실성면에서 참고함
* '현대상선' 700점 이상 토익 자격 기준으로만 활용
* '국민은행' 700점 이상 -
* '대우정보시스템' 700점 이상 토익 자격 기준으로만 활용
* 'CJ시스템즈'
- 제출하되, 기준 점수 없음 점수화 하지 않음,다만, 동점자일 경우 토익 점수 참고자료로 활용
* '한화(화학)' 제출하되, 제한 점수 없음 토익 점수 평가에 반영
* '수출보험공사' 830점 이상 영어면접실시는 상황에 따라 유동적
* '효성' 의무적 제출 아님 반영 하되, 절대적 기준 아님.
* '대림산업'
- 800점(사무직), 700점(기술직) 서류 전형시 반영, 면접과정에 영어 면접 포함

2006년 인쿠르트 잡지에서 발췌한 각 기업의 TOEIC 입사 기준점 기사

- 어느 조직 구성원들의 TOEIC 성적 분포는 평균이 750점이고 표준편차가 50으로 알고 있다. 만약 인사팀에서 신입사원의 입사 때 필요한 최저 TOEIC 성적을 700으로 한다면 관리자는 너무 낮게 점수를 책정하지 않는지 여부를 알고자 할 것이다. <토익.xls>

따라서 우리가 알고자 하는 확률은 최저 기준점 700점이 현재 조직원들의 성적과 비교하여 얼마나 낮은지 여부일 것이다. 즉 다음과 같은 확률을 구할 필요가 있을 것이다.

$$P(X \leq 700) = P(\frac{X-\mu}{\sigma} \leq \frac{700-750}{50}) = P(Z \leq -1) = P(Z \geq 1) = 0.1587$$

여기서, $P(Z \leq -1) = P(Z \geq 1)$은 분포가 좌우 대칭이기 때문이다. 기존 조직원의 15.87%가 700점 이하인 점을 고려한다면 이러한 700점이란 기준점이 높은지, 아니면 너무 낮게 책정된지는 관리자의 판단이다. 그러나 만약 관리자가 이러한 15.87%의 비율이 너무 낮다고 판단되어 40%의 수준으로 올린다고 하였을 때는 기준 점수를 얼마로 하여야 하는지 알아보자. 이 문제는 X^*를 구하는 문제로서 다음과 같이 표현이 된다.

$$P(X \leq X^*) = P(\frac{X-\mu}{\sigma} \leq \frac{X^*-750}{50}) = P(Z \leq \frac{X^*-750}{50}) = P(Z \leq z^*) = 0.4$$

여기서 $\frac{X^*-750}{50} = z^*$ 이다. 그런데 $P(Z \leqq z^*) = P(Z \geqq -z^*) = 0.4$가 되게 하는 $-z^*$는 [표 8.7]에서 대략 0.255가 된다. $-z^*$가 0.25인 경우는 0.4013, 그리고 0.26인 경우는 0.3974이다. 따라서 중간 값인 0.255가 확률 0.4에 해당하는 $-z^*$ 값이 된다.
그러나 대칭인 z^*값을 찾아야 하므로 우리가 원하는 값은 -0.255이다. 따라서

$$\frac{X^*-750}{50} = -0.255 \Rightarrow X^* = 750 - 0.255*50 = 737.25$$

기준점을 약 37점 올리면 이는 기존 조직원의 40% 백분율에 해당하는 점수가 된다.
이렇듯 평균 및 분산의 크기에 상관없이 표준화를 하면 어느 종류의 확률이든지 구할 수 있는 것이 큰 특징이다.

- 사실 제 5장에서 언급된 경험법칙(rule of thumb)은 이 정규분포에서 나온 것이다. 봉우리가 하나인 산이나 종의 모양을 하고 있는 분포는 정규분포를 의미한다. 정규분포에서는 다음과 같이 확률계산이 된다. 표준편차가 평균으로부터 2만큼 떨어지면 자료의 95%를 커버하고 있지 않은가?

$$P(-1 < Z < 1) = 1 - 0.1587 \times 2 = 0.6826$$
$$P(-2 < Z < 2) = 1 - 0.0228 \times 2 = 0.9544$$

이러한 확률을 계산하여 주는 명령문은 전문적인 통계 소프트웨어 뿐 아니라 많은 사람들이 쓰는 엑셀과 같은 스프레드시트 프로그램에도 내장되어 있다. 사실 표준정규분포표는 기존의 모든 통계학 책에서는 부록으로 첨부되어 있던 표이다. 그러나 프로그램에서 임의의 평균과 표준편차의 정규분포라 하더라도 원하는 확률을 얻을 수 있음으로 사실은 필요 없는 존재가 되었다.

참고로 $P(X \leq x) = p$ 에서 x를 명시하면 확률 값을 제공하는 엑셀 명령문이

=normdist(x,mean, stdev, 1)

이다. 위의 예제에서 해당하는 x의 값이 700이므로 간단히

=normdist(700,750,50,1)

을 입력하면 같은 결과를 얻을 것이다. 마지막 입력 인자 1은 누적확률을 의미하는 숫자이다. 역으로 p 값을 명시하고 해당하는 x의 값을 제공받는 엑셀 명령문은

=norminv(p, mean, stdev)

이다. p 값을 0.40으로 명시를 하면, 즉,

=norminv(0.40, 750,50)

을 입력하면 정확한 x값인 737.3327란 숫자를 얻을 것이다. 위에서 어렵게 구한 값과 일치하지 않는가? ■

예제 8.11 포장백의 품질도 정규분포로 관리된다.

- 군대에 납품하는 물건을 담는 봉투는 1,000 제곱 평방미터당 섬유질이 20그램이 섞여 있는 종이로 만들어야 한다. 그러나 작업 과정 중에서 발생하는 임의성 때문에 포함되는 섬유질의 무게는 고르지 않다.

이 기계는 작업기계에 평균을 맞추어 주는 레버가 있어 어느 평균값으로도 맞추어 줄 수가 있다. 표준편차는 공정상태가 좋으면 0.1을 유지하고 공정상태가 불량하면 0.15까지 올라간다. 섬유질의 무게는 정규분포를 이룬다고 가정할 수 있다. 생산된 종이에 포함되는 섬유질의 무게가 19.8 파운드 미만이거나 20.3 파운드 이상이면 불량으로 처리된다. 현재 평균을 20으로 맞추고 공정상태가 우수할 때와 불량할 때의 불량률을 구하여 보자. <포장백.xls>
우리가 원하는 불량률은 수식으로는 각각 다음과 같다.

$$P(X \geq 20.3|\mu = 20, \sigma = 0.1) + P(X \leq 19.8|\mu = 20, \sigma = 0.1)$$
$$P(X \geq 20.3|\mu = 20, \sigma = 0.15) + P(X \leq 19.8|\mu = 20, \sigma = 0.15)$$

그리고 엑셀 명령문으로 표현한다면

=normdist(19.8, 20, 0.10,1) + 1-normdist(20.3, 20, 0.10, 1)
=normdist(19.8, 20, 0.15,1) + 1-normdist(20.3, 20, 0.15, 1)

와 같다. 각각 2.4%, 11.4%가 나온다. 표준편차가 0.15일 때 불량 판정 받을 가능성은 표준편차가 0.1일 때보다 4배 이상으로 나온다. 표준편차의 크기는 0.05정도의 차이이나 불량률은 4배가 나온다는 사실은 그만큼 표준편차의 관리가 중요하다는 의미이기도 하다.

참고로 [표 8.8]에서는 각 세팅되는 평균과 표준편차에서 나오는 불량률을 계산하였다. 잘못 조정된 평균과 큰 표준편차에서 크면 불량률이 높아짐을 알 수 있다. ■

18			표준편차					
19		0.024	0.1	0.11	0.12	0.13	0.14	0.15
20		19.7	0.841	0.818	0.798	0.779	0.762	0.748
21		19.8	0.500	0.500	0.500	0.500	0.500	0.500
22		19.9	0.159	0.182	0.203	0.222	0.240	0.256
23	평균	20	0.024	0.038	0.054	0.072	0.093	0.114
24		20.1	0.024	0.038	0.054	0.072	0.093	0.114
25		20.2	0.159	0.182	0.203	0.222	0.240	0.256
26		20.3	0.500	0.500	0.500	0.500	0.500	0.500

[표 8.8] 평균과 표준편차에 따른 불량률

예제 8.12 **우리가 내야 하는 세금도 변수이다.**

펀드에 따른 세금액은 펀드의 수익률에 따라 달라진다.

- 어느 개인이 모 펀드에 10,000(천원)을 투자하려고 한다. 이 펀드의 년 수익률 X는 평균이 10%이고 표준편차가 4%인 정규분포를 따른다고 알려져 있다. 즉, 연말에 투자자가 가지는 금액은 10,000(1+X)(천원) 이다. 그러나 이익분에 대해서는 33%의 세금을 공제하여야 한다. 이 개인은 다음과 같은 두 가지 사항에 관심을 둘 수 있다. <세금.xls>

(1) 400(천원) 이상 세금을 낼 확률
(2) 세후 이익의 90번째 백분위수.

세전 이익은 10,000X다. 따라서 10,000X의 33%를 세금으로 내야 한다. 즉 3,300X가 세금

이다. 첫 번째 질문은 3,300X가 400보다 클 확률을 물어 본 문제이다. 즉,

$$P(3{,}300X > 400) = P(X > 4/33)$$

이다. 엑셀에서 다음과 같은 명령문을

=1- normdist(4/33, 0.1, 0.04)

입력하면 된다. 29.8%가 나온다.

또한 두 번째는 세후이익 6,700X가 특정한 값보다 작을 확률이 90%가 되는지 그 특정한 값을 구하는 문제이다. 즉,

$$P(6{,}700X < x) = 0.90 \text{ 혹은 } P(X < x/6{,}700) = 0.90$$

을 구하여야 한다. 엑셀에서 다음과 같은 명령문에

=norminv(0.9, 0.1, 0.04)

의해 15.13%가 나온다. 그러면 x는 다음과 같이 결정된다.

x/6700 = 0.1513 => x = 6700× 0.1513 => x = 1,013

세후이익의 90번째 백분위수는 약 1,013(천원) 정도 나온다. ■

8.2.4 이항분포와 정규분포와의 관계

이항분포의 모양은 n이 충분히 크고 p가 0이나 1에 가깝지만 않으면 산의 모양을 하고 있다고 하였다.

• 이항분포를 정규분포로 근사 시킬 수 있는 점은 놀랄만한 사실이 아니다. 왜냐하면 이항분포를 따르는 확률변수의 구조는 n개의 베르누이를 따르는 확률변수의 합으로 구성되어 있기 때문에 n이 충분히 커지면 나중에 배우게 될 '중심극한 정리'란 힘에 의해 정규분포로 근사하는 것이다.

따라서 n이 충분히 크고 p가 0이나 1에 가깝지 않은 이항분포는 평균이 np 이고 표준편차가 $\sqrt{np(1-p)}$ 인 정규분포로 근사시킬 수 있는 것이다. 다만 이항분포는 이산형이고 정규분포는 연속형이기 때문에 약간의 조정만 해주면 된다.

예제 8.13 n=100 이고 p=0.6인 이항분포를 따르는 확률변수 X가 55보다 같거나 클 확률

$$P(X \geqq 55) = \sum_{x=55}^{100} 100Cx\, 0.6^x 0.4^{100-x}$$

을 구하는 문제를 정규분포로 근사하여 구해보자. 직접 이 확률을 구하면 0.86891이 나오지만 정규분포로 근사하면 다음과 같다. 거의 비슷하지 않은가?

$$P(X \geqq 55) = P(Z \geqq \frac{54.5-60}{\sqrt{100\text{x}0.6\text{x}0.4}}) = P(Z \geqq -1.12268) = 0.869214$$

엑셀에서는 다음과 같은 명령문으로 할 수 있다.

=1-normdist(54.5, 60, sqrt(100*0.6*0.4),1)

여기서 55를 쓰지 않고 54.5를 쓴 것은 이산형과 연속형의 차이를 조정하였기 때문이며 거의 비슷한 확률을 얻는다. 이러한 근사는 n이 충분히 크고 p가 0.5에 가까우면 더욱 좋다. 그러나 엑셀에는 웬만한 크기의 n에 대해서도 확률을 계산하는 명령문 구조가 있기 때문에 실제적으로는 근사의 필요성을 별로 느끼지는 못한다. 그러나 이론적으로는 함축적인 의미가 많이 있는 내용이다. ■

8.2.5. 미래로 갈수록 불확실성은 증대한다.

• 미래에 일어나는 사건은 불확실하다. 그리고 미래로 갈수록 이런 불확실성은 증대한다. 이런 말은 우리가 흔히 주위에서 듣는데 이런 개념을 한번 통계학 입장에서 실험을 하여 보자.

여러분의 미래는 보이는가?

예제 8.14 **미래로 갈수록 불확실하다.**

어느 회사는 앞으로 6년간 제품에 대한 판매량을 모형화하고자 한다. 지난 자료를 분석한 결과 내년도 판매량은 평균이 5,000개, 그리고 표준편차가 500개인 정규분포로 가정하였다. 이는 대략적으로 말하면 3,500개에서 6,500개까지 범위 내에서 판매가 이루어질 수 있다는 의미이다. 그리고 앞으로 두 번째 해의 판매량의 평균은 첫해 나온 실제 판매량으로 가정할 것이다. 예를 들어 첫해 실제 판매량이 5,500개이면 두 번째 해의 판매량의 평균이 5,500개가 된다는 의미이다. 이렇게 하면 첫 해 판매량과 두 번째 해의 판매량은 독립이 아니다. 즉, 첫해 판매량이 높으면 두 번째 해의 판매량 역시 높아질 가능성이 매우 높다.

엑셀에서 정규분포에서 값을 추출하는 명령문을 다음과 같이

=norminv(rand(),5000,500)

입력하자. 첫 해의 판매량은 이런 정규분포에 의해 결정이 될 것이다. 언급하였듯이 두 번째 판매량의 평균은 첫해 나온 실제 판매량을 가정할 것이다. 앞으로 6년 후의 판매량과 표준편차는 어떻게 나오는지 시뮬레이션을 하여 보자. 주어진 엑셀화면 셀 A2에는

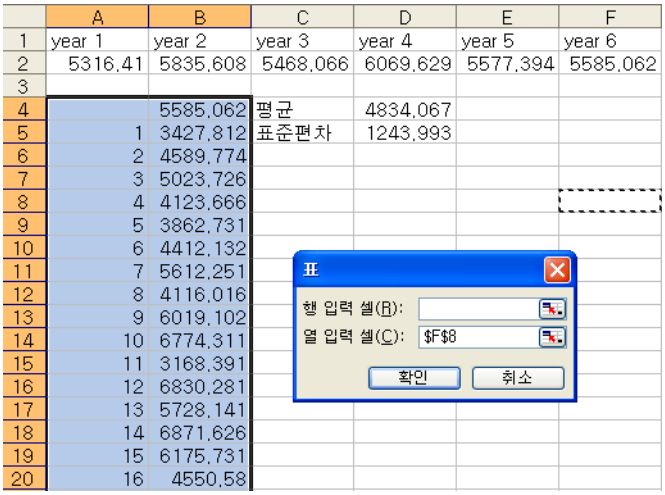
=norminv(rand(),5000, 500)

이 입력되어 있다. 그리고 셀 B2에는

=norminv(rand(), A2, 500)

을 입력하고 셀 C2부터 F2까지는 이를 복사시킨다. 그런 다음 셀 B4에 =F2를 링크시켜 놓는다. 그런 다음 이런 행위를 100번 하여 그 평균과 표준편차를 구하면 된다. 엑셀에서는 1부터

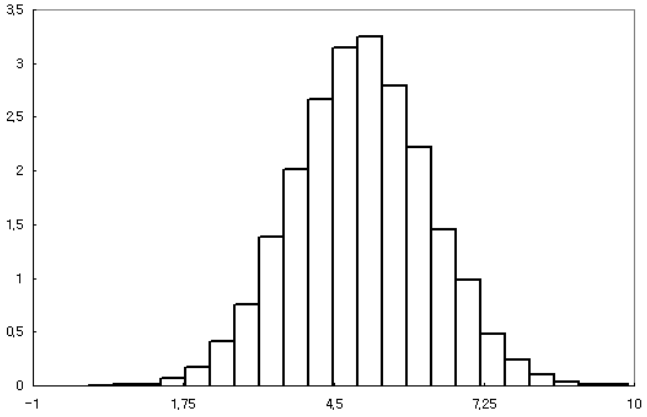
100까지의 숫자를 셀 A5:A104까지 입력한 후 범위 A4:B104를 블록으로 잡은 뒤 엑셀 메뉴 데이터>표에서 행 입력셀은 빈칸으로 놔두고 열 입력셀을 아무 빈 셀(그림에서는 F8)로 지정한 후 확인을 누르면 된다. 그러면 셀 B4부터 B104까지 값이 자동으로 채워질 것이다.



	A	B	C	D	E	F
1	year 1	year 2	year 3	year 4	year 5	year 6
2	5316.41	5835.608	5468.066	6069.629	5577.394	5585.062
3						
4		5585.062	평균	4834.067		
5	1	3427.812	표준편차	1243.993		
6	2	4589.774				
7	3	5023.726				
8	4	4123.666				
9	5	3862.731				
10	6	4412.132				
11	7	5612.251				
12	8	4116.016				
13	9	6019.102				
14	10	6774.311				
15	11	3168.391				
16	12	6830.281				
17	13	5728.141				
18	14	6871.626				
19	15	6175.731				
20	16	4550.58				

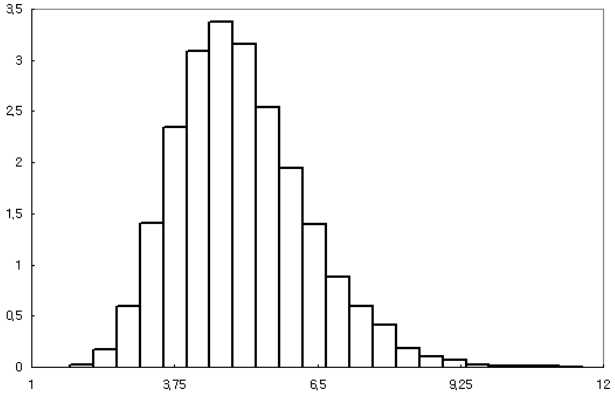
[표 8.9] 시뮬레이션

그런 다음 셀 D4와 셀 D5에 셀 B5:B104에 저장된 100개의 자료의 평균과 표준편차를 구하면 된다. 평균은 5,000과 거의 같은 값이 나오지만 표준편차의 값은 500이 아니라 1,243 값이 나왔다. 해가 6년 지난 후의 표준편차는 2배 이상으로 커진 상태가 된 것이다. 미래로 갈수록 불확실성이 커진다는 의미를 표준편차로 확인한 것이다. [그림 8.10]은 이런 반복시행의 횟수를 10,000개로 하여 구한 6년 후의 판매량에 대한 히스토그램이다. 표준편차의 크기가 늘어났을 뿐 좌우 대칭인 정규분포의 모양을 하고 있다.



[그림 8.10] 시뮬레이션 1 결과

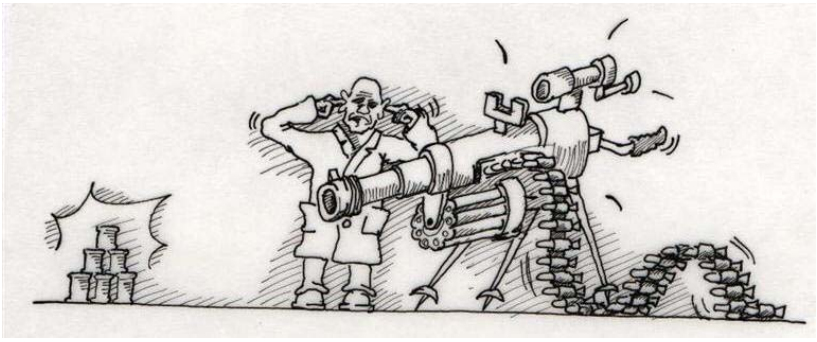
그러나 표준편차 역시 평균에 따라 그 크기를 달리 할 가능성이 높다. 따라서 표준편차는 평균이 5,000이면 그 1/10인 500이지만 평균이 6,000이면 그 500×(6,000/5,000)인 600으로 된다고 가정하자. 이럴 경우도 6년 후의 판매량의 모양이 좌우 대칭이 될까? 하는 문제이다. 결론부터 이야기하면 '아니다'이다. [그림 8.11]처럼 평균과 표준편차는 각각 4,977 및 1,207로 위의 경우와 유사하게 나왔으나 분포는 좌우대칭이 아닌 오른쪽으로 왜도가 발생한 모양을 하고 있다.



[그림 8.11] 시뮬레이션 2 결과

• 어느 경우나 표준편차의 크기는 첫해 우리가 가정한 500보다 훨씬 큰 값을 제공한다. ■

이러한 현상은 우리 주위에서 얼마든지 볼 수 있다. 제일 대표적인 것이 주식 시장에서의 주가흐름이다. 현재로부터 멀리 있는 증권의 가치는 예측하는데 어려움이 있는 것이다. 재무이론에 의하면 "주가는 랜덤워크 모형으로 로그정규분포의 모양을 하고 있다." 라고 이야기 하는데 이에 부합하는 현상을 설명한 것이다.



시뮬레이션기법은 매우 유용하게 쓰인다.

여기서 언급한 시뮬레이션 기법은 표본추출분포에서 매우 유용하게 쓰는 개념이니 숙지하기 바란다.

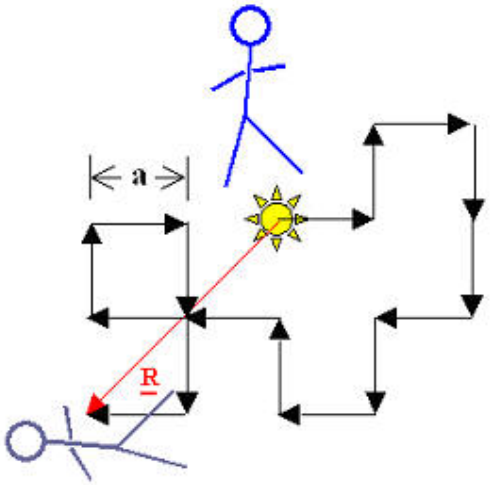
예제 8.15 왜 엄마는 아이에게 반드시 그 자리에 있어야 한다고 하는가?

아래 그림에서 출발점(해로 표시)에서 취객은 갈지자 걸음을 걷는다. 여기서 갈지자를 간략하게 하기 위해 현 위치에서 동서남북으로 다음 걸음을 옮길 확률은 각각 1/4로 가정하였다. 또한 걸음 폭은 일정하게 a로 표시하였다.

아래 그림을 보면 이 취객은 현재 위치에서 동쪽으로 첫 걸음을 옮겼다. 두 번째 걸음 역시 첫 번째 움직인 위치에서 동서남북으로 걸음을 옮길 것이다. 그림에 의하면 두 번째 발걸음은 북쪽으로 움직였다. 걸음을 걸을수록 이 취객의 위치는 점점 첫 위치에서 멀어지지 않는가?

16걸음을 걸은 후에 이 취객이 맨홀(R로 표시)에 빠질 가능성은 얼마인가? 가 이 그림에 대한 원래 질문이다. 그러나 이런 확률은 일반인들이 구하기 힘들다. 여기서 저자가 의도하는 바는 다만 걸음을 옮기면 옮길수록 이 취객은 찾기 더 힘들어지기 때문에 차라리 현 위치에서 술을 깨었다가 움직이는 것이 낫다는 뜻을 전달하기 위함이다.

놀이동산에서 혹 엄마를 잃어버리더라도 엄마는 반드시 움직이지 말고 그 자리에 있어야 한다는 말과 같다. ■



학습요약

제 8장에서는 확률분포의 개념을 우리가 일상에서 자주 접하는 현상 이해에 필요한 확률분포로 확대 적용하여 보았다. 이산형과 연속형으로 나누어 설명하고 이산형의 확률분포 중에서는 이항분포의 근간이 되는 베르누이 분포, 이항분포 등이 언급되었다. 특히 이항분포는 독립인 n개의 시행에 관련되어 만들어진 매우 유용한 분포이다. 이에 따른 다양한 응용예제는 본문에서 언급한 내용 이외에도 사례를 많이 찾을 수 있다. 연속형 분포로는 균일분포, 삼각형분포, 그리고 정규분포 등이 언급되었는데 이 중에서 특히 정규분포는 통계학을 대변한다고 할 정도로 중요한 분포이다. 자연적인 현상을 가정할 때도 정규분포가 이용되지만 앞으로 설명될 이론적인 이유 때문에도 많이 사용되는 분포이다. 한가지 덧붙일 말은 자료가 정규분포를 따르지 않는다고 해서 영어의 표현처럼 정상이 아니라고 생각하면 곤란하다. 분포의 모양이 정규분포라는 말이지 정상이란 말은 아니다. 표준정규분포표를 이해하는 과정을 통해 정규분포에 대한 훈련을 하는 것이 중요하다. 이어지는 제 9장에서 표본들의 특징에 대한 이해를 하는데 정규분포는 매우 중요한 역할을 한다. 여기서 소개한 모든 분포들은 엑셀에서 명령문이 존재한다. 독자들의 훈련이 필요한 시점이다. 또한 이산형인 이항분포와 정규분포의 근사성에 대해 언급하였다. 그리고 간단한 시뮬레이션을 통해 미래로 갈수록 불확실성이 증대한다는 개념을 설명하여 보았다.

8장 연습문제

8.1 이항분포의 실용적인 문제를 살펴보도록 하자. 시계를 유럽에 수출하는 모회사는 두 개의 공장을 가지고 있다. 시간당 생산능력은 각 500개이다, 그러나 생산라인 1은 생산라인 2에 비해 시설이 노후화되어 있어 불량률이 2%인 반면 공장 2는 불량률이 1%에 그친다. 다음 질문은 공장을 책임을 지고 있는 관리자가 인턴사원으로 있는 독자들에게 던진 질문이다. 이항분포의 가정을 하고 답하기 바란다.

(1) 주어진 시간에 k보다 큰 숫자의 불량시계를 만들어 내지 않을 확률이 99%되기 위한 제일 작은 k는 얼마인가?
(2) 수출 계약이 500개가 들어 왔다. 생산라인 2로부터 만들어 공급을 하기로 하였는데 소비자가 패키지를 열어 보았을 때 적어도 500개 이상의 양품의 시계가 들어 있을 확률이 99%가 되기 위해서는 몇 개를 포장해서 보내야 하겠는가?
(3) 다른 주문 1,000개가 들어 왔다. 두 생산라인에서 같은 수량의 시계로 공급하기로 하였다. 위와 마찬가지로 소비자가 패키지가 풀어 보았을 때 적어도 1,000개의 양품이 들어 있을 확률이 99%가 되어야 한다. 생산라인에서 몇 개의 제품을 생산하여야 하겠는가? 생산량은 같아야 한다. 그러나 총 생산된 시계의 합은 이항분포를 따르지 못한다. 따라서 이 문제는 시뮬레이션으로 풀어야 한다.
(4) 100개의 주문이 추가되었다. 양품인 경우는 개당 $500을 지불하기로 소비자가 약속을 하였다. 그렇다고 100개 이상의 양품의 시계를 보냈다 하더라도 수익은 늘지 않는다. 생산비용은 생산라인에 상관없이 개당 $450이다. 이 주문을 하나의 생산라인에서 만들어 보내려고 한다. 소비자가 패키지를 열어 보았을 때 양품의 개수가 100개가 안되면 그 개수만큼 돈은 지불이 된다. 반면 부족한 시계는 개당 $1,000씩 양품을 만들어 특급우편으로 반드시 보내야 한다. 이 시계에 대한 값은 소비자는 지불하지 않는다. 생산라인1에서 생산한다면 몇 개를 생산하여 보내야 하는가? 생산라인 2에서 생산을 한다면 몇 개를 생산하여 보내야 하는가?

8.2 150개의 회사에게 질문지를 보내 어느 항목에 대한 조사를 실시하고자 한다. 150개의 화사 중에서는 이러한 설문지에 응답을 하지 않을 경우가 많다. 즉, 무응답 비율은 45%이다. 이 조사기관은 적어도 90개 이상의 설문지를 회수하려고 한다. 그렇다면 90개 이상의 설문지를 회수할 가능성은 얼마나 되는가? 무응답 비율이 낮으면 낮을수록 이러한 확률은 올라갈 것이다. 엑셀의 데이터-테이블 기능을 이용하여 답을 구하는 것이 편리할 것이다.

8.3 (8.2 계속) 또한 이 회사는 일정기간이 지난 다음 무응답을 보인 회사에 전화를 걸어 재촉을 한다면 30%는 응답을 한다고 보자. 적어도 110장의 설문지를 회수할 가능성은 얼마나 되는가? 이 문제 역시 시뮬레이션의 도움이 필요하다.

8.4 어느 지자체에서 판매하고 있는 물건의 앞으로 3주간 수요는 각각 다음과 같이 파악된다.

제 1주 : 평균 50, 표준편차 10인 정규분포
제 2주 : 평균 45, 표준편차 5인 정규분포
제 3주 : 평균 65, 표준편차 15인 정규분포

각각의 주에서 발생하는 수요는 독립이다. 즉 앞으로 발생하는 3주 간의 수요는 이들의 합으로 구성된다. 즉 앞으로 3주간 발생하는 수요는 평균(50+45+65=155), 그리고 분산(100+25+165=290)인 정규분포를 따른다고 가정할 수 있다.

(1) 현재 재고는 180이다. 앞으로 3주간 제품을 공급받지 못한다면 이 지자체는 앞으로 3주간 재고부족이 일어날 가능성은 얼마인가?
(2) 재고부족이 일어나지 않을 확률이 적어도 98%이상이 되려면 이 지자체는 얼마만큼의 재고를 현재시점에서 더 확보하여야 하는가? 물론 앞으로 3주간 물품을 공급받지 않는다는 가정은 불변이다.

8.5 많은 운전자들은 고속도로를 운행하는데 있어 규정 속도를 지키지 않는다. 고속도로공사에서 조사한 바에 의하면 고속도로의 자동차 주행속도 평균은 모르지만 표준편차는 8킬로인 정규분포를 따른다고 알려져 있다.

(1) 약 40%의 운전자는 시속 100킬로 이상으로 운전을 하다면 평균은 얼마나 되는가?
(2) 약 25%의 운전자는 시속 80킬로 이하로 운전을 한다면 평균은 얼마나 되는가?
(3) 표준편차 역시 알려져 있지 않다고 보자. 모든 운전자의 40%는 시속 100킬로 이하로 운전을 하고 모든 운전자의 10%는 120킬로 이상으로 운전을 한다고 알려져 있다면 평균과 표준편차는 어떻게 계산이 되는가?

8.6 신용카드회사의 주 임무 중의 하나는 신용카드 사용금액에 대한 횟수문제이다. 일단의 그룹에 대한 분석을 해 본 결과 신용카드 사용금액에 대해 상환을 하지 않는 비율(부도율)이 7%이고 상환하지 않는 금액은 평균 35만원 그리고 표준편차 10만원의 정규분포를 따른다고 알려져 있다. 상환을 하지 않는 금액 중 20%는 추후에 추징이 가능하고 나머지 80%는 불가능하다.

(1) 이 그룹에 있는 신용카드 사용자가 부도를 내고 그 상금액이 20만원 이상 될 가능성은 얼마인가?
(2) 만약 이 그룹에 500명이 있다면 (A)에서 밝힌 조건을 만족하는 사람의 평균과 표준편차는 얼마인가?
(3) 적어도 500명의 사람 중에서 적어도 25명 이상이 (a)조건을 만족할 확률은 얼마인가?
(4) 부도된 금액 중에서 징수된 금액이 전혀 없다. 모든 부도금액은 손실로 처리가 되었다면 부도된 총 금액의 분포는 어떠한 모양새를 하고 있는가? 정규분포로 근사할 수 있는가?

8장 실습문제

8.1 각자 자신의 키, 몸무게, 출근시 평균소요시간(분), 퇴근시 평균소요시간(분), 핸드폰 번호(끝의 4자리), 자기 주머니에 갖고 있는 동전의 개수, 자기 지갑에 있는 지폐의 개수, 10개의 동전을 던졌을 때 그림면의 개수 및 첫 번째 그림면이 나왔을 때까지의 시행 횟수를 조사한 후 모두의 자료를 취합한다.

(1) 키의 분포는 어떤 분포를 이루는가? 정규분포를 이룬다고 볼 수 있나?

(2) 몸무게는 어떤 분포를 이루는가? 정규분포를 이룬다고 볼 수 있나?

(3) 출근시 평균소요시간(분)은 어떤 분포를 이루는가?

(4) 퇴근시 평균소요시간(분)은 어떤 분포를 이루는가?

(5) 핸드폰 번호(끝의 4자리)는 어떤 분포를 이루는가? 균일분포를 이룬다고 볼 수 있나?

(6) 자기 주머니에 갖고 있는 동전의 개수는 어떤 분포를 이루는가?

(7) 자기 지갑에 있는 지폐의 개수는 어떤 분포를 이루는가?

(8) 10개의 동전을 던졌을 때 그림면의 개수는 어떤 분포를 이루는가? 이항분포를 이룬다고 볼 수 있나?

(9) 10개의 동전을 던졌을 때 첫 번째 그림면이 나왔을 때까지의 시행횟수는 어떤 분포를 이루는가? 비대칭분포가 되는가?

8.2 다음 사이트에 들어가면 평균과 표준편차에 따라 정규분포를 그려준다. 표준편차가 1/2인 경우, 1인 경우, 2인 경우 그래프의 모양이 어떻게 다른가?

http://www.stattucino.com/berrie/dsl/index.html

제 9 장

표본을 추출할 때는 오차 계산이 가능해야 한다.

차 례

학습목표

통계학의 한 분야는 통계적 추론이다. 우리는 주어진 모집단에 대해 관심을 가지고 있는 사항에 대해 알고자 한다. 예를 들면 우리가 판매하는 치약을 사용하고 있는 소비자는 총 치약 시장 규모로 볼 때 몇 퍼센트를 차지하고 있는가? 어느 특정 쇼핑몰에서 쇼핑을 하는 고객들이 평균적으로 사용하고 있는 신용카드 사용금액은 얼마나 되는가? 특정 정부정책에 대해 일반 국민들은 얼마나 인지하고 있는가? 그러나 우리가 관심을 가지고 있는 모집단의 크기는 일반적으로 매우 크다. 따라서 모집단에 있는 모든 개체를 접촉하여 우리가 알고자 하는 사항을 구하는 것은 시간과 비용이 많이 드는 매우 번거로운 작업이 된다. 따라서 대부분의 경우 모집단에서 표본을 효율적으로 추출하여 표본에 담겨져 있는 정보를 이용하여 모집단의 성질에 대해 알고자 할 것이다. 이것이 통계적 추론이다. 좋은 추론을 얻기 위해서는 모집단을 닮은 좋은 표본이 있어야 함은 물론이다. 제 2장에서는 간단히 표본의 중요성에 대해 알아 본 바 있다. 이번에는 표본추출방법에 대한 자세한 소개를 함과 동시에 표본의 정보가 모집단의 성질을 추론하는데 어떻게 쓰일 수 있는지 이론적인 과정을 알아볼 것이다.

9.1 표본추출방법

우리는 이미 2장에서 모집단을 잘 대표하는 표본을 추출하여 모집단의 성질을 알고자 하는 개념을 소개한 바 있다. 먼저 앞서 배운 개념을 다시 보기로 하자.

모집단: 추론의 대상이 되는 모든 개체의 집합체다.

새 화장품을 출하한다고 하면, 모집단은 21세 이상으로서 백화점 및 쇼핑몰에서 쇼핑을 하는 여성 전체와 같이 정의가 될 수 있다. 일단 모집단이 정의되면 표본이 추출되는데 추론은 이렇게 정의된 모집단에 국한하여 사용됨은 물론이다. 그러나 모집단을 정의하는 것 자체가 어려운 경우가 많다.

예제 9.1 정확한 모집단을 정의하는 것은 사업성공의 지름길이다.

FM 인터넷 사이트

미국의 FM(Franklin Mint)회사는 희귀한 골동품과 같은 제품을 극소수의 소비자에게 우편 판매를 하는 회사이다. 따라서 이러한 극단의 소비자에게 우편으로 유인물을 보내 광고를 하여야 한다. 이 회사는 올바른 소비자에게 광고를 하는 것이 회사의 사활이 걸린 문제였다. 과대한

광고비는 이익 창출에 도움을 주지 않기 때문이다. 광고에 응답하는 비율이 평균적으로 200명 중 1명일 정도로 매우 낮은 이러한 시장인 경우는 더욱 그렇다. 최근까지만 해도 이 회사의 초기 전략은 주먹구구식이었는데 자료가 소비자 혹은 구매자에 대한 정보가 축적됨에 따라 특정의 물건을 구매하는 소비자의 모집단을 정의하는데 성공하였다. 즉, 구매 확률을

- 소비자의 구매이력
- 인구변수
- 주제, 소재, 아티스트, 후원자 등과 같은 물건의 속성

으로 연결하여 구매의사가 있는 소비자가 어디에 있는지 알아낸 것이다. 이러한 모형에서 소비자의 구매 반응을 조금이라도 높인다면 매출대비 우편물 발송에 따른 비용의 비율을 줄이는데 많은 효과를 볼 것이다. ■

표본추출틀(프레임) : 모집단에 있는 표본추출단위(sampling units)라 불리는 멤버의 리스트이다.

여기서 표본추출단위는 사람, 가구, 회사, 도시 등과 같이 정의될 수 있다.

여기에서는 모든 표본단위가 기록된 유한 N개의 모집단만을 대상으로 한다. 또한 이러한 프레임이 존재한다고 가정하겠으나 아무리 유한의 모집단이라 하더라도 전체프레임을 구하기는 매우 힘들다. 예를 들어 서울에 살고 있는 10대 비행청소년이라는 모집단의 프레임은 구하기가 힘들 것이다. 이런 경우는 부분적인 프레임으로부터 표본을 추출하여야 한다. 만약 선정한 부분적인 프레임이 모집단의 중요한 부분을 생략한다면 표본추출의 편의가 발생된다. 다른 예를 들어 보자. 서울시내에 있는 식당을 대상으로 표본 추출하려고 하는데 프레임을 광고 전화번호부에 등록되어 있는 식당 리스트로 해서 표본 추출한다면 광고를 하지 않는 많은 식당들이 빠져 있는 상태가 되므로 문제가 있을 수가 있다.

표본 추출은 기본적으로 두 가지가 있다. 하나는 **확률(probability)표본**이고 다른 하나는 **유의(judgement)표본**이다. 확률표본은 표본단위가 모집단으로부터 확률적인 메커니즘에 의해 구해지는 것을 의미하고 유의표본은 그런 확률적인 메커니즘이 전혀 없이 표본추출자의 판단에 의해 시행되는 점이다. 그러나 본 장에서는 유의표본은 논의대상에서 제외한다. 왜냐하면 우리가 유의표본의 정확성에 대해 논할 아무런 근거를 전혀 가지고 있지 않기 때문이다. 본 절에서는 여러 종류의 표본 추출방법이 존재하나 제일 대표적으로 많이 쓰이는 몇 가지 방법들을 소개하도록 하자.

9.1.1 단순임의추출법(simple random sampling)

제일 단순한 형태의 표본추출방법이다. 크기가 N인 유한 모집단에서 n개의 표본단위를 표본추출하는 경우를 생각하여 보자. 단순임의추출법이란 모집단의 각 표본추출단위가 표본으로 추출될 가능성이 동일한 표본추출방법이다.

예제 9.2 표본이 대표성을 갖는다는 통계적으로 무슨 의미인가?

예를 들어보자. 크기가 5인 5명의 구성원으로 만들어진 모집단이 있다. 이 모집단을

(a, b, c, d, e)

라 이름을 붙여보자. 여기서 n=2인 표본을 추출하도록 하자. 그러면 모든 가능한 표본은 다음과 같이 총 10개가 나올 것이다. 임의로 이를 나열하여 보자.

(a,b), (a,c), (a,d), (a,e), (b,c), (b,d), (b,e), (c,d), (c,e), (d,e)

이다. 단순임의추출이란 크기가 2인 개개의 표본이 선택될 확률이 1/10이란 뜻이다. 모집단의 한 멤버 예를 들면 b는 10개의 표본 중에서 4번 속해 있다. 그러므로 b가 표본에 속해있을 확률은 4/10=2/5가 된다. 일반화해서 이야기하면 N이 모집단의 크기, 그리고 n을 표본의 크기로 한다면 어느 멤버라도 표본에 포함될 가능성은 n/N 이다. ■

- 그러나 단순임의추출법은 제일 단순한 표본추출방법으로서 많은 통계학 예제에 쓰이나 실제 응용문제에서는 좀 더 효율적인 방법을 적용한다.

단순임의추출표본을 구성하는 방법은 몇 가지가 있다. 이러한 방법들은 모두 난수생성기(엑셀에서는 =rand()와 같은 명령문)를 구현하여야 한다.

첫 번째 방법은 간단하다. 난수생성기, =rand()를 작동한 다음 숫자가 0.465가 나왔다고 하자. 이 숫자는 0부터 1까지의 구간을 10개로 나눈다면 5번째 해당하는 구간에 속하는 숫자이다. 위의 경우 N=10이므로 구간을 10개로 나누면 된다. 따라서 선택되어지는 표본은 5번째인 (b, c)가 된다.

그러나 짐작하다시피 N이나 n이 커지면 이러한 방법은 효율적이지 못하다. 따라서 두 번째 방법을 제안한다. 이런 방법은 텍스트로 이해하는 것보다 실제적으로 구현하는 것이 개념을 이해하는데 좋다. 엑셀과 같은 스프레드시트 프로그램을 이용하도록 한다.

예제 9.3 단순임의추출법은 어떻게 구하는가?

[표 9.1]의 자료는 미국 중서부에 살고 있는 가구 중 일부를 발췌한 자료이다. 총 40가구의 연소득을 기록한 프레임이다. 이를 이해 편의상 모집단이라 정의하자. <단순임의추출법.xls>

- [표 9.1]에서 B10:B49에 있는 숫자들을 모집단으로 정의하고 여기서 크기가 10인 표본을 단순임의추출하려고 한다. 그리고 표본추출 후 표본의 평균과 표준편차를 모집단의 평균, 표준편차와 비교하려고 한다.

이를 위해서 먼저 셀C10에 =rand()를 입력하여 셀 C10:C49에 복사하여 둔다. 그런 다음 셀 A10:C49를 블록을 잡은 후 복사한다. 셀 F10에 커서를 위치한 후 마우스의 오른쪽 버튼을 눌러 붙여 넣기를 시행하는데 값만 지정한다. 그런 다음 난수값을 지정하여 소팅을 오름 순서대로 정해 자료를 정리한다. 그러면 그림에서 상자 안에 들어간 10개의 단위가 단순임의추출법에 추출된 값이다.

크기가 10인 또 다른 표본을 추출하려면 같은 방법으로 실행하면 된다. 이러한 절차는 엑셀의 매크로 기능을 이용하면 반복적인 수고를 덜어 낼 수 있다. 그러나 단순임의추출을 하기 위해서는 프레임만 필요함에도 불구하고 사실 실제로는 거의 쓰지 않는다. 왜냐하면 어느 표본추출단위이든지 표본에 포함될 가능성이 같아지므로 실제 얻어진 표본은 굉장히 넓은 지역적인 범위에 걸쳐 표본추출단위가 흐트러져 있을 가능성이 높다. 따라서 이런 표본을 조사하는 것은 비용이 많이 든다. 또한 조사가 이루어지기 전에 모든 표본추출단위가 확인되어야 하므로 이 역시 거의 불가능하다. ■

	A	B	C	D	E	F	G
4		평균	중앙값	표준편차			
5	모집단	$39,985	$38,500	$7,377			
6	표본	$42,270	$39,500	$10,894			
7							
8	**모집단**				임의추출표본		
9	가구	소득	난수		가구	소득	난수
10	1	$43,300	0.5896		27	$60,800	0.0017
11	2	$44,300	0.6131		4	$38,000	0.0508
12	3	$34,600	0.9757		12	$51,500	0.0951
13	4	$38,000	0.0508		40	$41,000	0.1578
14	5	$44,700	0.7127		21	$56,400	0.1882
15	6	$45,600	0.9472		36	$36,300	0.2232
16	7	$42,700	0.4662		25	$44,900	0.2375
17	8	$36,900	0.8081		10	$33,700	0.2921
18	9	$38,400	0.3191		32	$31,700	0.2976
19	10	$33,700	0.2921		37	$28,400	0.3189
20	11	$44,100	0.9807		9	$38,400	0.3191
21	12	$51,500	0.0951		16	$38,600	0.3263
22	13	$35,900	0.6537		14	$35,600	0.3501
23	14	$35,600	0.3501		29	$47,600	0.3571
24	15	$43,000	0.9690		23	$38,100	0.4064
25	16	$38,600	0.3263		22	$33,600	0.4338
26	17	$32,400	0.7891		7	$42,700	0.4662
27	18	$22,900	0.5765		28	$42,500	0.5038
28	19	$48,100	0.7836		18	$22,900	0.5765
29	20	$31,900	0.6748		31	$33,000	0.5782
30	21	$56,400	0.1882		1	$43,300	0.5896
31	22	$33,600	0.4338		2	$44,300	0.6131
32	23	$38,100	0.4064		13	$35,900	0.6537
33	24	$42,500	0.8548		20	$31,900	0.6748
34	25	$44,900	0.2375		5	$44,700	0.7127
35	26	$35,200	0.7878		33	$48,600	0.7184
36	27	$60,800	0.0017		35	$33,000	0.7372
37	28	$42,500	0.5038		19	$48,100	0.7836
38	29	$47,600	0.3571		26	$35,200	0.7878
39	30	$36,100	0.8836		17	$32,400	0.7891
40	31	$33,000	0.5782		8	$36,900	0.8081
41	32	$31,700	0.2976		39	$37,300	0.8087
42	33	$48,600	0.7184		24	$42,500	0.8548
43	34	$39,300	0.8665		34	$39,300	0.8665
44	35	$33,000	0.7372		38	$46,900	0.8676
45	36	$36,300	0.2232		30	$36,100	0.8836
46	37	$28,400	0.3189		6	$45,600	0.9472
47	38	$46,900	0.8676		15	$43,000	0.9690
48	39	$37,300	0.8087		3	$34,600	0.9757
49	40	$41,000	0.1578		11	$44,100	0.9807

[표 9.1] 단순임의추출법

이 자료는 후에 중심극한정리를 설명할 때 다시 언급하기로 한다.

9.1.2 계통표본추출법

• 55,000명이 등재된 전화번호부에서 250명의 표본을 추출하려고 한다고 보자.

계통표본추출법(systematic sampling)의 절차는 다음과 같다.

먼저 55,000/250=220개의 표본구간을 만들어 놓는다. 그런 다음 1부터 220까지의 사이 숫자 중에서 하나의 무작위 수를 뽑는다. 그 수가 131이라 하자. 그러면 131번째 이름을 첫 번째 구간에서 뽑고 그 이후에는 220번째 이름을 기계적으로 뽑아내는 방법이다. 이렇게 n=250명의 이름을 추출하는 방법이 계통표본추출법이다.

이렇게 구하는 표본은 단순임의추출과 달리 220개의 모든 가능한 표본만 존재한다. 그리고 220개의 표본이 선택될 가능성은 다 같다. 이러한 의미에서 보면 단순임의추출법과 그 맥락을 같이 한다.

계통표본조사인 경우는 프레임에 들어가 있는 표본추출단위의 순서와 조사 목적 간의 관계에 그 핵심이 있다.

예를 들어 만약 조사의 목적이 소득 수준을 파악하는 것이라면 전화번호부에 있는 이름의 순서와 소득과는 아무런 관계가 없기 때문에 그렇게 계통표본추출법은 그다지 매력이 되지 못한다. 그러나 프레임에 속해 있는 표본추출단위가 어떤 크기로 순서가 정해져 있다면 계통표본추출법은 더 좋은 방법이 될 수 있다. 예를 들어 어느 회사가 구매고객의 구매량의 크기 순서대로 고객의 프레임을 가지고 있다면 계통표본추출법은 단순임의추출법보다 구매고객을 좀 더 대표하는 표본을 구성할 수 있을 것이다.

그러나 표본추출단위의 명단 순서가 주기적인 요소가 섞여 있다면 조심하여야 한다. 예를 들어 표본 구간이 7이고 매일 거래량을 기록한 프레임에서 표본을 추출한다면 계속해서 월요일에 해당하는 표본추출단위를 추출할 가능성이 있기 때문이다. 그러나 이런 경우를 제외하면 계통표본추출법에 의해 구해지는 표본은 모집단을 잘 대표할 가능성이 매우 많다.

9.1.3 층화표본추출법

전체 모집단 내에 다양한 부모집단(subpopulation)이 인지가 되는 경우가 있다고 보자. 이러한 부모집단을 층화(stratification)라 하는데 경우에 따라서는 전체 모집단으로부터 단순임의추출 하기보다는 각 층화로부터 단순임의추출하는 방법이 더 나을 수 있다. 특히 층화간의 변동은 매우 큰 반면 층화 내에서는 변동의 폭이 그렇게 높지 않은 경우에는 층화표본추출법(stratified sampling)은 효과가 있다.

예를 들어 설명하여 보자. 텔레비전에 상품에 대한 광고를 하는 회사가 소비자의 반응을 알아보고자 한다. 먼저 어떤 층화가 타당할 것인가에 대한 질문을 하여야 한다. 성, 혹은 소득, 아니면 텔레비전을 본 시간 등이 타당한 층화가 될 것이다. 물론 어떤 층화를 선택할 것인가는

회사의 조사목적과 관련이 있을 것이다. 그러나 이런 층화의 기준을 구하는 것은 잠시 후에 언급할 예제 9.4에서 보다시피 매우 어려운 작업이다.

크기가 N개인 모집단을, 층의 개수가 I개이고 i번 째 층의 크기가 $N_i(i=1,2,\cdots,I)$ 인 층화로 구분하였다고 보자.

$$N=N_1+N_2+\cdots+N_I$$

그러면 각각의 층화로부터 크기가 n_i인 표본을 추출한다.

$$n=n_1+n_2+\cdots+n_I$$

여기서, 어떻게 n_i 를 정할 것인가에 대한 문제가 남아 있다.

제일 간편한 방법은 비율표본크기(proportional sample size)이다. 예를 들어 전체 모집단 중 15%가 첫 번째 층화에 속한다면 표본 크기 n의 15%에 해당하는 표본이 크기가 된다. 그러나 이런 방법은 층화 간의 변동을 무시하는 단점이 있기 때문에 별로 추천하지 않는다. 예를 들어 보자. A대학교는 학생들이 연간 교과서에 지불하는 평균 금액을 알기 위해 모집단을 학부, 석사, 박사 과정으로 층화를 구분하였다. 모집단에서 부모집단의 크기는 각각 20,000명, 4,000명, 1,000명이고 표본의 크기가 150명이라면 비례표본에 의하면 각각의 층화에서 120, 24, 6명의 학생을 추출하여야 한다. 그러나 층화에 있는 학생들이 지불하는 금액은 층화별로 그 표준편차가 다르다. 학부학생들은 표준편차가 작은 반면, 박사과정에 있는 학생들의 금액의 표준편차는 클 것이다. 따라서 이와 같은 경우는 첫 번째 층화에서 많은 수의 표본을 추출하기보다는 오히려 세 번째 층화에서 많은 수의 표본을 추출하여야 한다. 따라서 층화의 크기에 비례해서 표본의 크기를 할당하는 것은 별로 바람직하지 않다. 표준편차의 크기로 n_i 를 정하는 문제는 책의 범위를 벗어나므로 생략한다.

예제 9.4 세법에 따른 변화를 표본으로 알아본다.

OTA 인터넷 사이트

미국 의회의 세금관련 소위원회나 정부기관은 세법 개정에 따른 세수의 영향을 보고자 할 때 Office of Tax Analysis(OTA)란 기관에 자문을 의뢰한다. OTA는 표본 추출된 세납자를 대상으로 시뮬레이션을 실시한 다음 보고서를 작성하는데(흔히 보고서는 통상적으로 하루 만에 보고되어야 하는 신속성을 요구하는데) OTA는 이런 보고서를 매년 수천 건씩 실시하고 있다. 따라서 OTA는 현재 155,000명 수 만큼 유지하고 표본의 크기를 이런 이유로 반으로 줄이려고 한다. 물론 표본의 수를 줄이더라도 표본의 대표성은 있어야 한다.

참고로 표본에 들어 있는 세납자들의 속성은 192개로 구성이 되어 있는데 세전수입, 납부세액, 봉급 및 임금, 총 연금액 등이다. 하나의 방법은 층화를 이용하는 것이다. 155,000명을 층화로 나눈 다음 각 층화로부터 대표 표본을 추출하여 이들로 하여금 작은 표본을 형성하는 것이다.

그러나 무슨 기준으로 층화를 나눌 것인가? 층화를 세전수입과 같은 하나의 기준으로 한다면 다른 속성들은 무시가 되는 문제가 나타난다. 따라서 되도록이면 작은 수의 속성으로 층화를 나누는 방법이 중요한데 군집분석이라는 통계학 방법론을 이용하여 세전수입과, 납부세액으로 층화의 기준을 삼으면 세납자들이 자연스럽게 동질적인 층화를 형성한다고 판단하였다.

기존의 155,000명의 표본과 층화 표본추출로 표본의 크기를 반으로 줄인 75,000명의 표본은 각 속성별로 그렇게 차이가 나지 않는다. OTA는 1980년 이후부터는 작은 표본으로도 세법의 변화에 대한 영향을 신속하고 정확하게 보고할 수 있었다. ■

9.1.4 군집표본추출법

어느 지방 단체가 도시 가구의 특징을 조사한다고 보자. 표본추출단위는 가구가 될 것이다. 지금까지 논의한 표본추출법에 의해서도 가능하지만 좀 더 쉬운 방법으로 표본을 추출하는 방법이 있다. 직접적으로 가구를 표본추출단위로 하기보다는 먼저 '구(區)'라는 행정단위를 표본추출단위로 하여 단순임의추출 한다. 그런 다음 표본추출된 구에 들어간 모든 가구들을 대상으로 조사한다. 여기서 구라는 행정단위를 우리는 군집(cluster)이라 부른다. 이런 방법이 군집표본추출법(cluster sampling)이다. 단순임의추출법보다 조사를 행하는데 있어 훨씬 간편하다. 왜냐하면 흩어져있는 표본추출단위를 찾아다닐 노력을 하지 않아도 되기 때문이다. 그러나 한편으로는 이런 편리성으로 인한 단점도 있는데 같은 구에 살고 있는 표본추출단위의 응답은 거의 비슷하게 나올 가능성이 높기 때문이다. 오히려 여러 구에 걸쳐 있는 표본추출단위에서 응답을 구하는 것이 더 정확할 수 있기 때문이다.

따라서 위에서 언급한 1단계 군집표본추출법보다는 현실적으로는 보다 복잡한 표본설계를 한다. 전국적으로 표본추출을 하기 위해서 전국의 시·군 지역에서 일정 수를 단순임의추출(simple random sampling)한 다음 그 지역의 행정단위(cluster)를 단순임의추출하고 그런 다음 각 행정단위로부터 가구를 계통표본추출(systematic sampling)하는 방법을 예로 들 수가 있을 것이다. 이런 추출법을 통상적으로 다단계-군집 표본추출법(multistage sampling)이라 한다. 갤럽과 같은 조사기관이 전국적인 조사에서 많이 쓰는 기법이다. 대략 표본의 크기는 1,500명이다.

예제 9.5 **다단계 표본 추출법을 통한 표본은 미국의 경우 어떻게 구성되는가?**

미국지도에서 표본은 대충 이런 비율로 뽑힌다.

마케팅 조사기관은 매달 의뢰인에게 고객 표본 및 관련된 정보를 제공하여야 한다. 표본내의 고객들은 전국적으로 구성되어 있다. 이러한 표본의 정보는 의뢰인 회사에 매우 중요하다. 각 지역 담당자가 제공하는 서비스의 질을 평가하고 인사고과 기준으로 삼기 때문이다. 의뢰인이

가지고 있는 소비자 명단은 지역적으로 구성되어 있다. 미국 전역을 4개(북부, 중부, 남부, 서부)구역으로 나누고 각 구역을 몇 개의 지역(예를 들어 중부는 중서부, 남중부, 남서부)으로 나눈다. 그리고 이런 지역은 주별로 명시되는 구획(중서부는 아이오와, 미네소타, 미시간 등)으로 나누어진다. 그리고 마지막으로 구획은 몇 개의 행정지구로 구성이 되어 있다. 의뢰인은 마지막 단위인 행정지구에 들어 있는 소비자의 명단으로부터 일정수를 무작위로 표본 추출한 다음 이 명단을 목표하고 있는 소비자의 표본수와 함께 조사기관에 보낸다. 이 마케팅 조사기관은 이러한 리스트로부터 표본을 추출하여 관련된 정보를 의뢰인에게 제공하여야 한다.

예를 들어 의뢰인이 보내온 어느 특정 행정지구의 명단이 320명으로 구성되어 있다면 10명을 표본 추출하여 조사한 후 관련된 정보를 제공하여야 한다. 이 조사기관은 계통표본 추출방법을 이용하여 요구되는 10명의 5배 정도의 표본을 추출한다. 이 명단을 콜 센터로 보낸 다음 10명의 소비자가 연락이 닿을 수 있을 때까지 연락을 취하게 한다. 다단계 표본추출을 하지 않으면 이 의뢰인은 효율적인 고객관리를 할 수 없을 것이다. ■

9.2 통계량에는 오차가 있다.

- 모집단으로부터 임의추출된 표본들이 서로 독립이고 동일한 모집단분포에서 나왔다면 우리는 이러한 표본을 확률표본(random sample)이라 부른다.
- 통계량(statistic)이란 확률표본에 적용된 적절한 함수를 가리킨다. 예를 들면 모집단의 특성을 나타내는 값(이를 우리는 '모수(parameter)'라 한다.) 중 가장 중요한 모평균 μ에 대하여 다음과 같은 통계량을 정의할 수 있다. 이러한 통계량을 '표본평균'이라 부른다.

$$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{X_1+X_2+\ldots+X_n}{n}$$

여기서, $X_1, X_2, \ldots, X_n$은 확률표본을 가리킨다.

- 통계량을 구하는 목적은 표본에서 관측된 값으로부터 모집단의 모수를 추정하는데 있다. '추정'이란 추출된 표본으로부터 모집단의 성질을 규명한다는 뜻이다.

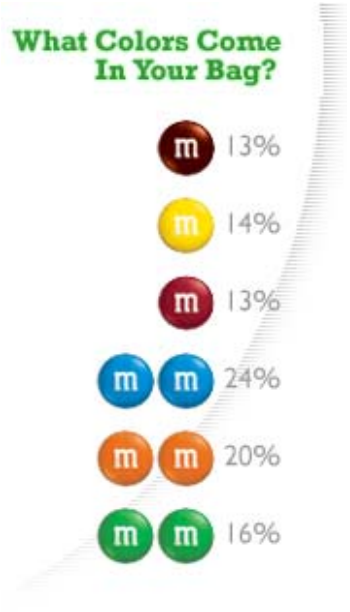
예를 들어 정부기관이 어느 특정 지역의 가구 평균소득 수준을 알고자 할 때는 그 지역에 있는 모든 가구의 소득의 평균을 의미하겠지만 그 평균소득은 그 지역에 살고 있는 가구들을 대표할 수 있는 표본을 추출하여 얻어지는 표본 평균을 가지고 추정을 한다.

추정을 하는 통계적인 절차는 어떤 표본을 추출하였느냐와 모집단의 어떤 성질에 대해 관심을 가지고 있느냐에 따라 달라진다. 그러나 단순임의추출법을 제외하면 그 절차가 복잡하다.

따라서 본 책에서는 단순임의추출법에 국한해서 논의하도록 한다. 그리고 거의 대부분은 모집단의 성질 중에서 평균 및 비율에 높은 관심을 기울일 것이다. 즉, 단순임의 추출된 표본으로부터 어떻게 모집단의 평균 및 비율을 추정할 것인가 하는 문제가 제일 집중적으로 다루어 질 것이다.

먼저 추정에서 발생하는 오차의 종류부터 살펴보도록 하자. 오차는 두 가지로 나누어진다. 표본(혹은 표본추출)오차(sampling error)와 비표본오차(non-sampling error)이다. 비표본오차는 표본오차가 아닌 모든 오차를 일컫는 용어이다. 표본오차는 확률적인 표본의 결과로 나타나는 오차이다.

- 표본오차는 표본에 의해서 나오는 추정값과 우리가 알고자 하는 모집단의 성질에 해당하는 값과의 차이이다.



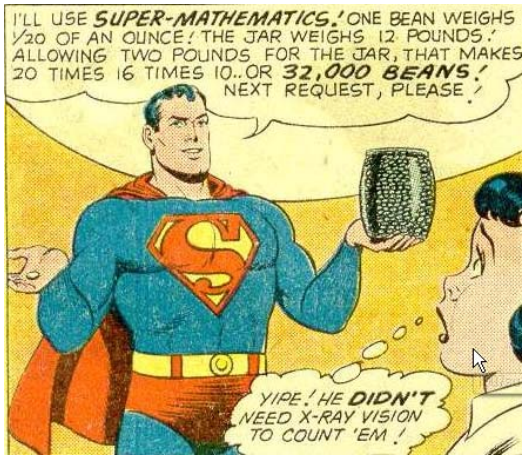
M&M 초콜릿의 색깔분포

예제 9.6 초콜릿 봉지에서도 표본오차를 읽는다.

위의 그림은 M&M을 만드는 회사의 인터넷 사이트에 올려져 있는 초콜릿 색깔의 분포이다. 숫자는 모집단의 자료라 간주한다. 슈퍼마켓에 가서 M&M을 한 봉지 사서 색깔에 대한 분포를 비교하여 보라. 차이가 난다면 이것이 표본오차인 것이다. 물론 색깔을 이런 분포로 하여 초콜릿을 만드는 것은 M&M을 사먹는 소비자 층을 중심으로 한 색깔 선호도 조사의 결과이다. 그래야만 소비자 만족도가 올라가 매출이 증대되기 때문이다. ■

예제 9.3(계속) **표본평균을 통해 표본오차를 이해하자.**

예제 9.3에서 언급된 모집단의 평균은 $39,985이다. 물론 이 값은 모집단을 전수조사하기 전에는 알려지지 않는 값이다. 만약 이 모집단의 평균을 표본평균으로 추정한다면 이러한 표본평균을 **점추정값**(point estimate)이라 한다.



우리는 너무 점추정값에 익숙하다.

- 일반적으로 우리가 알고자 하는 모집단의 성질을 모수라 한다면 점추정값은 관측된 표본자료에 의거해서 하나의 값으로 만든, 모수에 대한 “best guess"가 되는 값이다.

예제 9.3에서는 표본평균인 $42,270이 점추정값이다. [그림 9.1] 중 셀 B6에 계산되어 있다. 따라서 표본오차는 표본평균 값과 우리가 알고자 하는 모집단의 성질, 여기서는 평균값과의 차이다.

표본 오차 = $42,270-$39,985 = $2,285

표본오차가 양이므로 모집단의 평균을 $2,285만큼 과다추정(over-estimation)하였다. 음이 나왔다면 과소추정(under-estimation)이라 한다. ■

- 이를 일반화시키면 표본에서 얻어지는 표본평균을 $\overline{X}$ (통계량)라 하고 모집단의 성질 중 하나인 평균을 μ라 한다면 점추정값의 표본오차(sampling error)는 다음과 같이 정의된다.

$$\text{표본오차} = \overline{X} - \mu$$

표본오차는 어느 표본을 선택하느냐에 달려 있다. 왜냐하면 개개의 표본은 각각 다른 표본평균의 값을 가지고 있기 때문이다.

- 따라서 핵심은 통계량의 표본오차(sampling error)를 계량화할 방법이 없겠느냐 하는데 있다.

예를 들어 임의의 표본이 $2,285보다 큰 표본오차를 가져다줄 가능성은 얼마나 있을까 하는 것이다. 이와 같은 질문에 대한 답이 표본추출분포(sampling distribution)인 것이다. 위의 경우는 표본평균의 표본추출분포가 되는 것이다. 표본평균이 어떤 분포를 가지고 있는지에 대한 언급을 하면 위와 같은 질문에 답을 할 수 있다. 통계학에서는 제일 중요한 개념 중의 하나로 간주되고 있다. 잠시 후에 언급이 된다.

반면 비표본오차는 여러 가지 이유로 발생이 된다.

- 먼저 무응답 편의(non response bias)이다. 조사과정에서 표본의 일부가 응답을 하지 않아 발생된다. 문제는 무응답으로 인해 오차 오류가 발생하느냐 하는 것이다. 이러한 편의는 무응답자라도 응답을 했다면 "다른 응답자와 마찬가지일 것이다."라는 결론을 지을 수 있어서 그렇게 큰 문제는 없을 것이다. 그렇지만 이를 확인할 방법은 전혀 없다. 무응답 편의에 대해 추측을 하는 수밖에 없을 것이다.

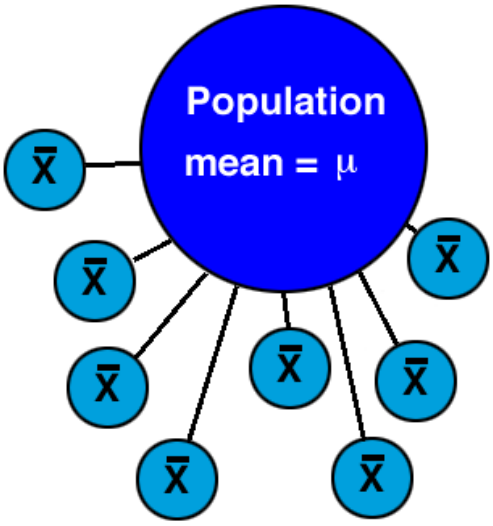
- 다른 비표본오차 하나는 진실하지 않은 응답이다. 질문이 매우 민감할 때 발생한다. 직접적으로 "당신은 마약을 한 경험이 있습니까?"라고 물어 볼 수 있는 상황은 그렇게 많지 않기 때문이다. 그렇지만 이러한 문제점은 무작위반응(random response) 기법 등을 이용하면 이는 어느 정도 극복이 된다. 이 기법에 대해서는 전문 도서를 참고해야 할 것이다.

- 다른 비표본오차의 하나는 측정오차이다. 이런 오차는 흔히 의도한 대로 설문이 잘 이루어지지 않을 때 나타난다. 조사자의 의도와는 달리 응답을 하는 경우가 된다.

앞으로 여기에서는 자료의 관측값에서 모든 비표본오차는 없다고 가정할 것이다.

9.3 표본추출분포란 무엇인가?

모집단으로부터 크기가 n인 모든 가능한 표본의 평균들로 만들어진 분포를 **표본평균의 표본추출분포**(sampling distribution of sample mean)라 한다.



표본평균들의 분포는 어떻게 생겼을까?

표본평균은 표본에 어떤 표본단위가 구성되었느냐에 따라 값이 작을 수도 있고 클 수도 있다. 그러나 대부분의 표본평균값들은 가운데 몰려 있을 것이다.

- 표본평균의 표본추출분포 역시 분포이므로 표본평균의 평균과 표준편차를 명시할 수 있다.

표본평균의 표본추출분포에서의 평균과 표준편차는 각각 식 (9.1), (9.2)와 같다. 표본평균의 표준편차, $\mathrm{stdev}(\overline{\mathrm{X}})$ 는 $SE(\overline{X})$ 라 표기하고 표준오차라 부르기도 한다.

$$E(\overline{X}) = \mu \tag{9.1}$$

$$\mathrm{stdev}(\overline{\mathrm{X}}) = \mathrm{SE}(\overline{\mathrm{X}}) = \sigma/\sqrt{\mathrm{n}} \tag{9.2}$$

식 (9.1)의 의미는 어느 특정 표본은 표본평균이 모집단의 평균을 과다 혹은 과소로 측정하나 평균적으로 보면 (on the average) 모집단의 평균과 일치한다는 뜻이다. 따라서 모집단의 평균이 알려져 있지 않은 경우는 과다추정 혹은 과소추정이라고 논할 이유가 없어지는 것이다.

그리고 식 (9.2)는 μ를 추정하는데 있어 $\overline{X}$의 정확성을 알려 주는 하나의 척도로 작으면 작을수록 $\overline{X}$의 값은 모집단의 평균 μ에 가깝다는 의미이다. 이는 모집단의 평균을 추정하는데 있어 표본평균의 정확성(accuracy)에 관련한 척도를 제공한다. 즉, 표준오차가 작으면 작을수록 표본평균은 모집단 평균에 근접한다는 의미이다.

그러나 σ를 알고 있는 상태는 매우 드물다. 따라서 σ대신에 표본표준편차를 대입하여 $SE(\overline{X})$의 근사값을 구해야 한다, 많은 경우 식 (9.3)을 그냥 $SE(\overline{X})$ 라 부른다.

$$\widehat{SE}(\overline{X}) = s/\sqrt{n} \qquad (9.3)$$

여기서, ^(hat라고 읽는다)의 의미는 근사값을 의미한다.

예제 9.3(계속)

표본평균이 $42,270이고 표준오차의 근사값은

$$\$10,894 \;/\; \sqrt{10} \;=\; \$3,445$$

이 나온다.

- 이에 대해 해석은 잠시 후에 설명을 자세히 하겠지만 "모집단의 평균으로부터 표본평균값이 2× $3,445보다 더 차이가 나지 않을 확률은 95% 정도 된다."로 해석이 가능하다. 이는 모든 가구의 소득의 평균은 $43,020 ± 2× $3,445구간 안에 있다는 점을 약 95% 정도 확신한다는 뜻이다.

- 독자들은 이 시점에서 s와 SE의 차이를 인식하는 것이 좋다. 표본의 표준편차는 자료의 변동을 측정하는 척도이고 표준오차는 모집단의 평균을 추정하는데 있어 표본평균의 정확성을 측정하는 방법이다. ■

마지막으로 한 가지를 언급하여야 한다. 통상적으로 모집단 크기인 N은 무한대로 가정해도 될 만큼 매우 크다, 그러나 위에서 언급한 예제 자료는 모집단의 크기가 유한인 N=40 이었다.

일반적으로 표본의 크기 n이 N에 대해서 5%를 넘어서게 되면 $SE(\overline{X}) = s/\sqrt{n}$ 은 식 (9.4)와 같이 조정하여 주는 것이 좋다. 이를 유한모집단수정(finite population correction factor: fpc) 이라 하다.

$$SE(\overline{X}) = \sqrt{\frac{N-n}{N-1}}\, s/\sqrt{n} \qquad (9.4)$$

식 (9.4)를 이용하면 표준오차는

$$\sqrt{\frac{40-10}{40-1}} \times \$3,445 = 0.877 \times \$3,445 = 3,021$$

이 나온다. 이 fpc는 항상 1보다 작다. 그리고 n이 커지면 줄어든다. 따라서 n이 커질수록 모집단 평균에 대한 추정의 정확성은 커진다.

그러나 N에 비해 n이 5% 미만인 경우는 이 fpc는 무시하고 표준오차를 구하는 것이 통상적인 규칙이다. 특별히 명시하지 않는 한 fpc는 무시하고 SE가 구해졌음을 이해하기 바란다.

9.4 중심극한 정리

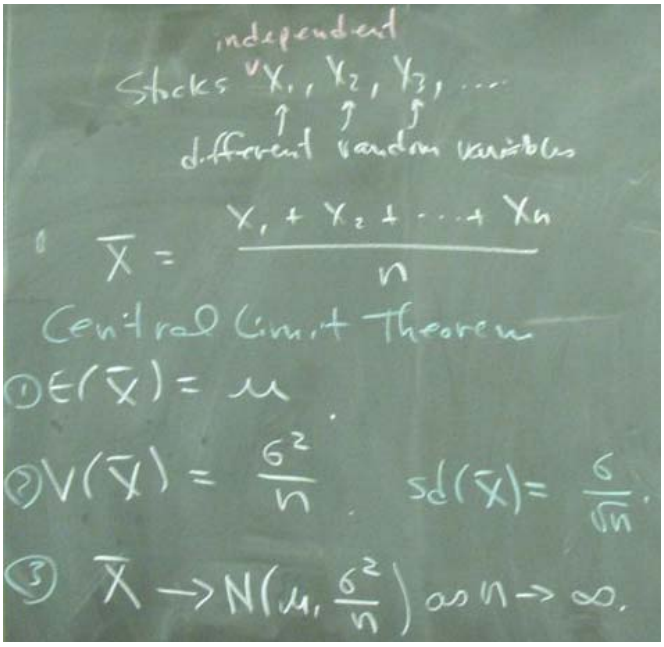
예제 9.7 **볼링장에서는 왜 점수가 평소보다 들쑥날쑥 하는가?**

여러분들은 친구들과 볼링장에 경기를 간다면 친구들과 내기를 하기 위해 여러분의 실력이 얼마나 되는지 친구들에게 알려 준다. 그러나 경기를 진행하면서 여러분들은 왜 오늘은 점수가 잘 안 나오지? 하는 질문을 스스로에게 묻곤 했을 것이다. 여러분은 "평균 실력이 130인데 오늘은 영 점수가 안 나오네?"하고 말이다.

볼링장에서도 중심극한 정리는 작동된다.

그리고 여러분의 친구들은 볼링실력이 진짜인지 의문시 할 것이다. 이런 의문을 풀어줄 개념이 중심극한 정리이다. 여러분의 개별적인 점수를 평균 볼링실력에 비교하는 것이다. 당일 게임의 수를 많이 하고 표본 평균을 낸다면 그 성적은 여러분의 평균 실력과 엇비슷할 것이다. 그러나 많아봐야 3게임 정도의 볼링을 즐긴다면 그 평균은 여러분의 평균 실력과는 많이 차이가 날 것이다. ■

지금까지는 표본평균의 평균과 표준오차(SE)에 대해서만 이야기하였을 뿐 표본평균의 모양새인 분포에 대해서는 언급하지 않았다. 본 절에서는 이에 대한 논의를 하여야 한다. 결론부터 이야기하면 이는 중심극한정리(central limit theorem)로 설명이 된다. 이 정리는 후에 나올 많은 개념들의 기본이 되는 매우 중요한 틀이 된다.



중심극한정리를 이해해야만 통계학이 보인다.

중심극한정리 : 평균이 μ이고 표준편차가 σ인 분포를 가지고 있는 모집단에서 표본 추출된 표본의 표본평균 $\overline{X}$의 표본추출분포는 표본의 크기 n이 충분히 크다면($n \geq 30$) 평균이 μ이고 표준편차가 $\sigma/\sqrt{n}$ 인 정규분포로 근사시킬 수 있다.

이미 우리는 표본의 크기에 상관없이 평균과 표준편차는 각각 μ, $\sigma/\sqrt{n}$ 임은 알고 있다. 그러나 중심극한정리는 여기서 더 나아가 표본의 크기 n이 충분히 크다면 표본평균의 표본추출분포는 정규분포로 근사할 수 있음을 뜻한다.

- 그럼 "어느 경우에 중심극한정리의 타당성을 보장받는 것일까?"

에 대한 질문을 할 필요가 있을 것이다.

일반적으로 문헌에서는 n>30인 경우라고 알려져 있다. 그러나 반드시 이러한 기준을 적용할 필요는 없다. 만약 모집단의 분포가 정규분포에서 매우 이탈된 모양을 하고 있다면 예를 들어 쌍봉 모양을 한다든지 왜도가 심한 경우는 표본평균의 분포가 정규분포를 하기 위해서는 표본의 크기가 30보다 훨씬 커야 한다. 반대로 모집단이 대칭형의 분포를 가지고 있다면 n이 30이 안되더라도 정규분포에 매우 근사한 분포를 하고 있을 것이다.

물론 모집단의 분포가 정규분포라면 중심극한 정리에 의존하지 않더라도 표본평균의 표본추출분포는 n의 크기에 상관없이 무조건 정규분포이다.

시뮬레이션 예제를 들어 이와 같은 중심극한 정리에 대해 알아보도록 하자.

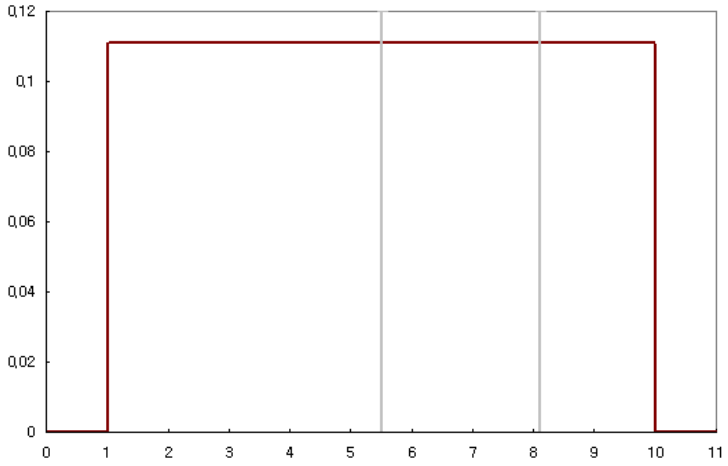
예제 9.8 **중심극한정리를 시뮬레이션 하여 보자.**

1부터 10까지 구간에서 난수를 임의로 2개 뽑아 난수의 평균을 상금으로 주는 게임을 가정하여 보자. 만약 난수 1.23과 4.77을 뽑았다면 상금

$$(1.23+4.77)/2 = 3.00$$

을 얻을 것이다. 이와 같은 게임을 무한정 반복했을 때 얻는 상금액의 분포는 어떤 모양을 하고 있을까? <중심극한정리.xls>

일단 우리가 가정하고 있는 모집단의 분포는 균일분포로서 [그림 9.1]과 같다. 평균보다 표준편차 한 단위 큰 값은 8.10에 위치하고 있다. 즉, 표준편차의 값은 8.10-5.50 =2.60이다.

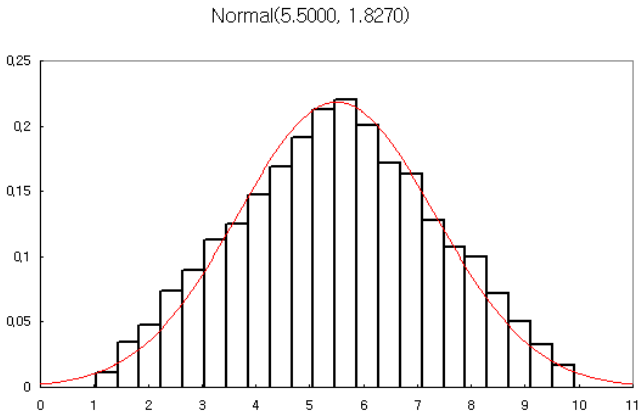


[그림 9.1] 균일분포(1,10)

1과 10구간에서 두 번 표본추출하여 보자. n=2 이다. 이러한 행위는 얼마든지 엑셀과 같은 프로그램에서 시험해 볼 수 있다. 표본평균의 분포를 이해하기 위해 잠시 엑셀의 명령문을 빌려 오도록 하자. 엑셀에서는 임의의 두 셀을 지정하여 각각의 셀에 다음과 같은

$$=1+rand()*(10-1)$$

명령문을 입력하면 두개의 난수를 구할 수 있다. 그런 다음 이 두 셀의 평균을 내어 상금액을 구하면 된다. 이러한 행위를 10,000번 반복하고 10,000개의 상금액을 히스토그램을 그리면 [그림 9.2]과 같은 삼각형 분포의 히스토그램이 나오게 된다.



[그림 9.2] n=2인 표본평균의 표본추출분포

이러한 분포의 평균, $E(\overline{X})$은 이론적으로 모집단의 평균 μ인

$$\frac{(1+10)}{2} = 5.5$$

가 나와야 하며 표준편차 $SE(\overline{X})$ 는

$$\frac{\sigma}{\sqrt{n}} = \frac{2.6}{\sqrt{2}} = 1.8385$$

로 나와야 한다. 여기서 2.6은 모집단의 표준편차,

$$\sigma = \sqrt{\frac{(10-1)^2}{12}}$$

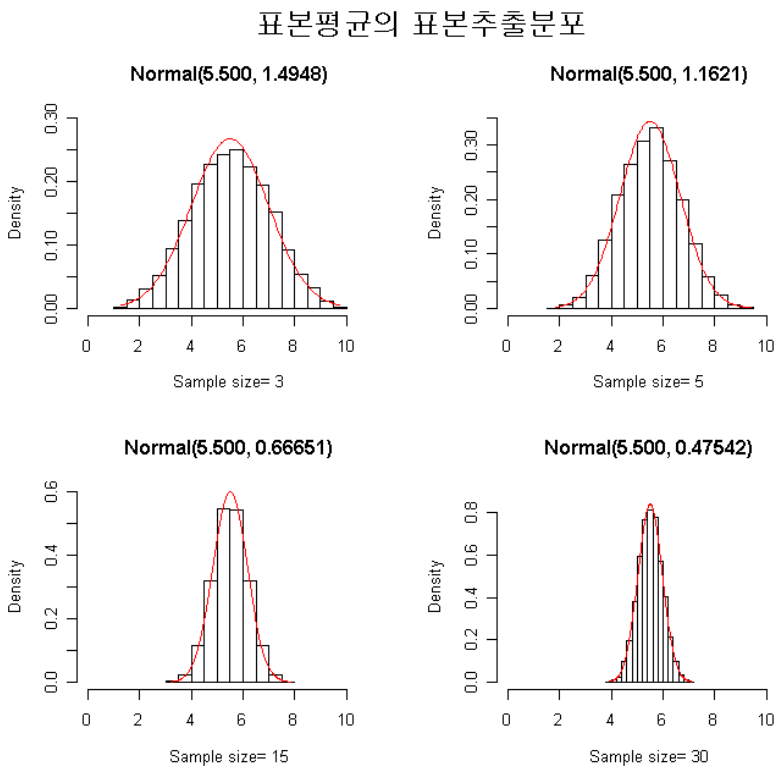
값이다. 그러나 특정의 시뮬레이션을 통해 나온 표본평균의 분포의 평균은 5.50, 그리고 표준편차는 1.827로 나왔다. 실제 이론적인 값과 약간 차이가 나는 것은 시뮬레이션의 결과이기 때문이다.

- [그림 9.2]를 보라. 놀라운 사실은 겨우 표본의 크기가 2인 경우에도 산의 모양을 하고 있지 않은가?

[그림 9.2]의 그림에서 곡선은 구해진 표본평균과 표준편차로 정규분포를 가정하고 그린 그림이다. Normal(5.5000, 1.8270)은 평균이 5.50이고 표준편차는 1.827인 정규분포를 가리킨다. 물론 n=2일 때에는 정규분포를 적합한 결과인 곡선과 모의실험의 결과인 히스토그램과는 아직 차이가 있다.

- 표본크기를 3, 5, 30으로 늘렸을 때는 어떤 현상이 벌어지는지 알아보자. 육안으로 본 결과 n=3인 경우라도 그림은 정규분포에 매우 근접되어 있음을 알 수 있다. 물론 n=30인 경우는 거의 정규분포와 일치함을 알 수 있다.
- 표본의 크기가 증가하면 표본평균의 값이 가질 수 있는 구간의 폭이 점점 가운데 몰려 있음을 알 수 있을 것이다. [그림 9.3]에서 n=5인 경우의 폭과 n=30일 경우의 폭을 비교하면 알 수 있을 것이다. 이는 $SE(\overline{X})$가 n이 커지면 커질수록 작아지기 때문이다.

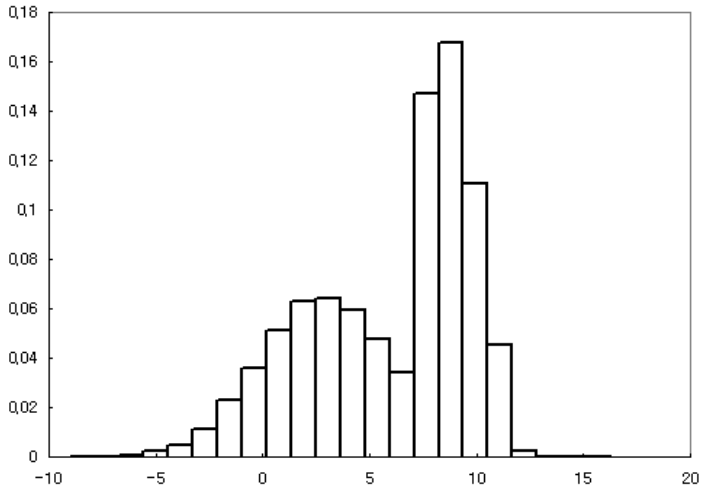
만약 이런 게임을 하여 상금을 탄다면 되도록이면 많은 수를 뽑아낸 다음 평균을 내는 것이 유리하다는 뜻이다. ■



[그림 9.3] n의 변화에 다른 표본평균의 표본추출분포의 변화

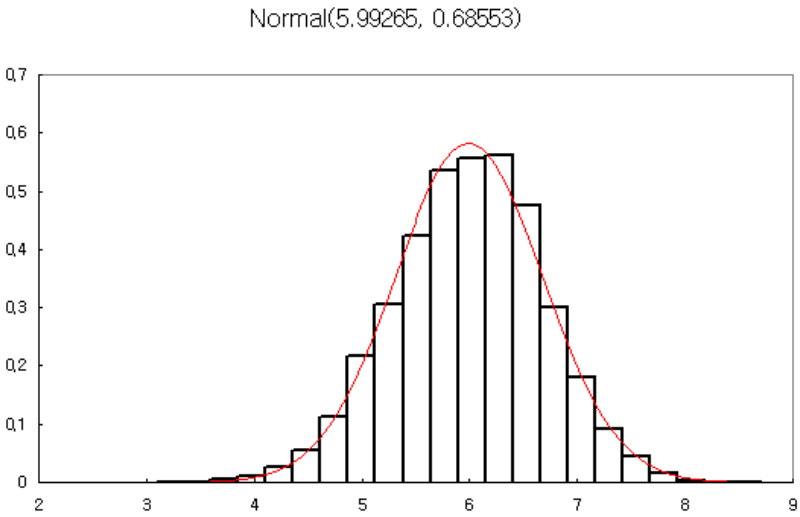
예제 9.9 **중심극한 정리는 모집단 분포의 모양새에 관계없다.**

모집단이 양봉인 [그림 9.4]와 같은 분포가 있다고 가정하자. 모집단의 평균은 6, 그리고 표준편차는 3.76이다.



[그림 9.4] 임의의 양봉분포

이러한 분포에서 30개의 변수 값을 추출하여 표본평균을 내면 표본평균의 분포는 과연 어떻게 되는지 구현하여 보았다. [그림 9.5]에서 보다시피 이 경우 역시 표본평균의 분포는 정규분포의 모양을 하고 있다.



[그림 9.5] n=30인 표본평균의 표본추출분포

모의시행 10,000번의 시뮬레이션 결과로는 표본평균의 표본추출분포 평균은 5.99265, 그리고 표준편차는 이론적인 값인 $3.76/\sqrt{30}$ = 0.686과 근사한 0.68553을 얻었다.

이상 결과로 보면 모집단의 분포가 심히 한쪽으로 왜도가 있지 않다면 표본크기 n=30에 육박하게 되면 표본평균의 표본추출분포는 정규분포로 근사해지는 경향을 실습할 수 있었다. ■

예제 9.3(계속) **실제자료에도 중심극한 정리는 통한다.**

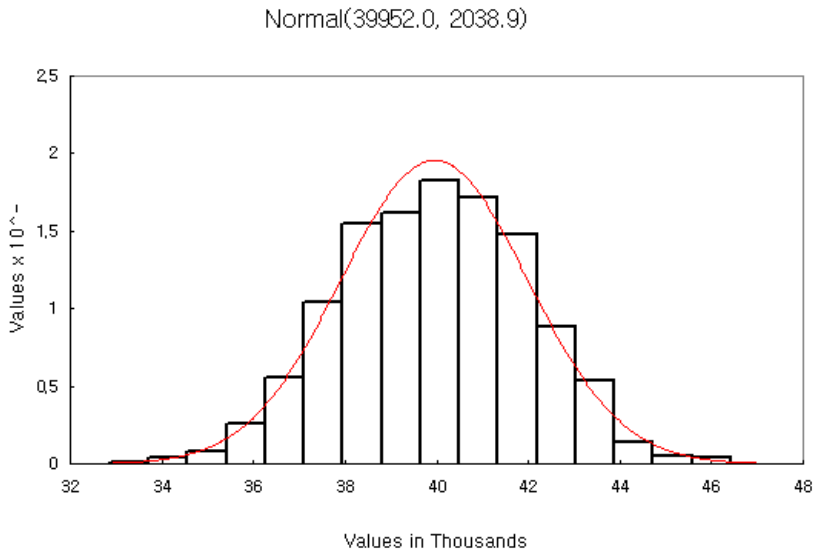
지금까지는 확률변수가 어느 특정 분포로부터 나온다고 가정하였지만 지금 주어진 자료가 모집단이라고 가정한다. 예제 9.3의 단순임의추출방법에서 사용한 자료를 가지고 중심극한정리를 실험해 보자.

- 모집단의 크기가 N=40 이고 표본의 크기는 10이다. 따라서 n/N의 비율이 5%를 넘어 가므로 finite population correction(fpc)을 적용하는 사례가 될 것이다.

다시 한번 단순임의추출방법을 정리하면 다음과 같다.

단순임의추출법: 먼저 C10:C49에 =rand()을 입력한다. 셀 A9:C49을 복사한 다음 붙여 넣기 위치인 셀 F9으로 이동한 다음 마우스의 오른쪽 버튼을 눌러 값만 복사한다. 그런 다음 열 G에 있는 난수 값을 오름차순으로 정렬한다. 그러면 첫 10개의 값이 우리가 찾는 표본이 된다.

엑셀이 가지고 있는 매크로 기능을 이용하여 반복적으로 10,000개의 표본을 추출하고 구해진 10,000개의 평균의 값을 히스토그램으로 [그림 9.6]에 그렸다. 즉, [그림 9.6]은 표본평균의 표본추출분포를 시뮬레이션하여 구한 결과이다.



[그림 9.6] 예제 9.3을 활용한 표본추출분포

평균은 모집단의 평균 $39,985와 비슷한 $39,952가 나왔으며 표준편차는 $2,038.9가 나왔다. 이는 다음과 같은 이론적인 값과 거의 일치함을 보인다.

$$SE(X) = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{30}{39}} \times \frac{7,377}{\sqrt{10}} = \$2,046.012$$

그리고 표본의 크기가 10에 불과하였지만 표본추출분포의 모양은 정규분포에 근접해 있음

을 알 수 있다. 물론 표준오차 공식 중 모집단의 표준편차는 모르는 값일 경우가 많다. 이런 경우는 표본표준편차의 값을 대입하여 표본오차의 값을 구하면 된다. ■

- 이상과 같은 내용을 정리하면 표본오차의 크기를 가늠할 수 있는 척도를 드디어 얻은 것이다. 즉, 임의의 표본이 가지고 있는 표본오차의 크기를 확률적으로 판단할 수 있는 근거가 마련되었다.

예제 9.10 **왜 많은 분포들은 정규분포인가? 중심극한정리가 알려주는 것들**

8장에서 정규분포를 설명할 때 정규분포는 자연적인 현상을 이해하는데 제일 적합한 분포라고 한 바 있다. 몸무게, 키, 온도 등이다. 이외에도 정규분포는 수학적인 이론에 근거해 많은 변수에도 적용이 가능하다고 이야기했는데 이는 중심극한정리를 염두에 두고 언급한 내용이었다. 어느 회사에서 팔리는 물건의 개수와 같은 변수를 생각하여 보자. 월 평균 판매량은 당연히 위에서 언급한 중심극한정리에 의해 정규분포를 따라가겠지만 하루에 팔리는 물건의 개수의 한 달간 합인 한 달간 판매량 같은 변수도 역시 중심극한정리에 의해 정규분포를 따라간다. 왜냐하면 월 평균판매량에 30을 곱하면 월 판매량이 나오지 않는가? 어느 변수에 상수를 곱한 변수 역시 같은 모양새를 하고 있는 것이기 때문이다. 월 판매량, 분기별 판매량 등은 중심 극한정리에 의해서 만들어지는 전형적인 정규분포이다. 다음과 같은 다른 형태의 중심극한정리 역시 의미가 있다.

중심극한정리 : 평균이 μ고 표준편차가 σ인 분포를 가지고 있는 모집단에서 표본 추출된 표본의 합 $\sum_{i}^{n} x_i$역시 표본추출분포는 표본의 크기 n이 충분히 크다면($n \geq 30$) 평균이 μ이고 표준편차가 $\sigma\sqrt{n}$ 인 정규분포로 근사시킬 수 있다.

- 우리가 취급하는 물건은 하루에 많으면 16개 그리고 적으면 4개 정도만 팔린다고 가정하자. 하루에 팔리는 물건은 산의 모양을 하고 있다면 우리는 평균 10개, 표준편차 2개 정도로 가정하고 있는 것이다. 왜냐하면 최대값에서 최소값을 뺀 범위를 6으로 나누면 대략적인 표준편차가 나온다.

$$\sigma \approx \frac{(16-4)}{6} = 2$$

그렇다면 한달에 팔리는 물건의 개수 X가 330개를 넘어갈 가능성은 얼마나 되는가? 이 문제는 바로 위에서 언급한 중심극한 정리를 이용하여 바로 구할 수 있다.

$$P(X \geq 330) = P(\frac{X-n\mu}{\sigma\sqrt{n}}) \geq \frac{33.0-300}{2\sqrt{30}}) = P(Z \geq 2.74) = 0.0031$$

이는 바꾸어 이야기 하면 하루에 팔리는 물건의 평균개수가 11개를 넘을 확률과 같다. 왜냐

하면 이는 한달에 팔리는 물건의 합의 개수와 같은 판매량이기 때문이다.

- 이상과 같이 평균이라는 개념이 들어간 확률변수, 혹은 합의 개념이 들어간 확률변수들은 정규분포를 따라갈 가능성이 매우 높은 것이다. ■

9.5 표본의 크기

결혼의 만족도를 조사하기 위해 부인을 늘려야 한다?

표본의 크기는 주어지는 문제가 아니라 기획단계에서 정해져야 하는 문제이기 때문에 쉽지 않다. 이미 논의했듯이 표본의 크기가 커지면 커질수록 표본오차(sampling error)는 작아진다. 그러므로 되도록 크기가 큰 표본을 얻고자 노력해야 한다. 그러나 반대로 작은 표본을 얻어야 하는 이유도 존재한다.

- 첫째가 표본을 추출하는데 들어가는 비용이다. 표본의 크기가 증가하면 비용 역시 늘어나기 때문에 설사 원하는 표본오차의 수준을 구하는데 표본의 크기가 500정도 들어간다면 예산의 문제로 인해 표본의 크기를 300 이하로 줄여야 할 필요가 있는 경우가 종종 있다.
- 두 번째는 자료 수집의 시의성(timeliness)이다. 예를 들어 돌아오는 주에 신상품에 대한 광고 여부를 결정하기 위해 표본조사를 한다면 빠른 시일 내에 조사를 마쳐야 함은 자명하다.
- 그리고 마지막으로 생갈 할 것은 무응답과 같은 문제로 발생하는 비표본오차들인데 표본의 크기가 커지면 이러한 오차는 늘어난다. 오히려 표본을 늘릴 때 사용되는 비용을 이런 비표본오차를 줄이고자 쓴다면 작은 표본으로 인한 표본오차의 문제를 오히려 상쇄하고 남을 수가 있다. 어쨌든 표본의 크기를 정하는 문제는 우리가 원하는 표본오차에 의해 결정된다.

표본평균을 가지고 모집단의 평균을 정하는 문제를 생각하여 보자. 이미 우리는 표본평균의 표준오차는 식 (9.5)와 같음을 알고 있다.

$$SE(\overline{X}) = \sigma/\sqrt{n} \tag{9.5}$$

그리고 중심극한 정리에 의해 n이 충분히 크면 표본오차(sampling error)의 크기는 표준오차의 2배를 넘지 않을 것이다. 이는 95% 확신할 수 있다.

식 (9.5)에서 표본크기의 문제는 $2SE(\overline{X})$ 를 우리가 받아들일 수 있을 정도로 작게 표본 n을 선택하는 문제가 된다. 이는 매우 간단하게 해결된다.

우리가 받아들일 수 있는(허용할 수 있는) 표본오차(sampling error)의 한계(bound, 이를 maximum probable absolute error라 부르기도 한다.) B를 선택하고

$$2SE(\overline{X}) = B$$

로 놓고 n에 대해 풀면 된다. 그러면 식 (9.6)과 같은 표본크기 n을 구하는 식을 얻을 수 있다.

$$n = \frac{4\sigma^2}{B^2} \tag{9.6}$$

식 (9.6)은 모집단에서 n크기의 표본을 무작위로 추출한다면 나오는 표본오차는 그 크기가 B보다 크지 않을 확률이 95%가 된다는 의미가 담겨져 있는 식이다.

식 (9.6)을 살펴보면 먼저 모집단의 분산 σ^2이 커진다면 많은 표본의 크기를 구하여야 한다. 즉, 모집단의 변동의 폭이 크면 클수록 많은 수의 표본추출을 하여야 한다. 반대로 작으면 표본의 크기가 클 필요는 없다. B와 n은 역의 관계에 있다. 우리가 원하는 B가 작으면 n은 커지고 반대인 경우는 n이 작아 질 것이다.

그러나 문제는 σ는 모집단의 성질이기 때문에 문제가 발생한다. 왜냐하면 알려진 경우가 많지 않기 때문이다. 따라서 σ를 대체하는 값을 찾아야 한다. 파일럿(pilot) 표본이 있다면 얻어진 표본 표준편차 s를 대입하면 되고 만약 이것도 존재하지 않는다면 보편적인 규칙(rule of thumb)에 의해 식 (9.7)으로 대체한다.

$$\sigma \approx \frac{\max - \min}{4 \text{ or } 6} \tag{9.7}$$

여기서 max는 최대값, min은 최소값을 의미한다.

왜냐하면 분산의 값은 모른다 하더라도 최소값과 최대값은 알려져 있는 경우가 많기 때문이다. 4로 나누어주면 좀 더 큰 표본을, 6으로 나누어주면 작은 표본을 얻기 때문에 보수적인 관점에서는 4로 나누어주는 관행이 타당할 것이다.

4 혹은 6으로 나누어주는 것은 평균을 중심으로 ± 2(혹은 3)의 표준편차로 이루어지는 구간 안에 95-99%의 자료가 포함된다고 보편적인 규칙에서 언급하였기 때문이다.

예제 9.11 **표본의 크기는 어떤 의미인지 알아보자.**

모집단이 $U(80,120)$ 인 균일분포를 가정하기로 하자. 모집단의 평균은 100이고 분산은 $(120-80)^2/12$ =133.33이다. B를 4로 원한다면 표본의 크기가 약 33이 된다.

$$n = 4\times\frac{133.33}{4^2} \approx 33$$

[표 9.2]에서는 모집단에서 33개의 확률변수를 추출하여 표본평균을 구한 다음 모집단의 평균과의 차이인 표본오차의 절대값을 구하고 이 값이 우리가 명시한 B의 값 2보다 작게 되는 경우가 95%의 확률을 보이는지 알아보았다.

이를 위해 엑셀 명령문을 이용하여 셀 A2에 =80 + rand()* (120-80)을 입력하고 셀 A2:A34까지 복사한다. 그리고 셀 C2에 표본평균을 계산하고 모집단의 평균과의 차이에 대한 절대값을 셀 E2에 기록한다. 그런 다음 =if(abs(e2<4,1,0) 을 통해 이 값이 B 보다 작으면 1 그리고 B 보다 크면 0으로 하여 셀 F2에 입력하였다. 셀 F2를 셀 D4에 연결한 다음 반복시행을 10,000번 한 후 평균을 보게 되면 평균이 약 95%가 나온다. 이제 독자들은 95%의 의미를 확인할 수 있는가? ■ <표본크기.xls>

	A	B	C	D	E	F	G
1	sample		xbar	mu			
2	115,8261		103,9648	100	3,964826	1	
3	96,68318						
4	94,01192			1		0,95	
5	99,56458		1	1			
6	112,1616		2	1			
7	107,2553		3	1			
8	82,93603		4	1			
9	83,22669		5	1			
10	97,76252		6	1		Press F9 to see the change	
11	94,17483		7	1			
12	99,49461		8	1			
13	105,0311		9	1			
14	118,9313		10	1			
15	112,4907		11	1			
16	99,25638		12	1			
17	85,36222		13	1			
18	110,7893		14	1			
19	109,9546		15	1			
20	112,756		16	1			
21	105,6174		17	1			
22	117,4179		18	1			
23	89,96024		19	1			
24	107,9629		20	1			
25	111,7217		21	1			
26	107,6661		22	1			
27	117,4206		23	1			
28	116,6226		24	1			
29	118,7317		25	0			
30	112,2187		26	1			
31	108,7906		27	1			
32	80,6052		28	1			
33	112,1024		29	1			
34	86,33213		30	1			

[표 9.2] 표본크기와 95% 관계

학습요약

통계학의 한 분야는 통계적 추론이다. 모집단의 성질을 잘 이해하기 위해서는 절대적으로 모집단을 잘 닮은 표본을 가지고 있어야 한다. 비용과 시간이라는 제약조건 하에서 이러한 표본을 추출하는 것은 매우 어렵다. 표본추출방법들 중, 대표적으로 쓰이는 단순임의추출법, 계통표본추출법, 층화추출법, 그리고 군집표본추출법을 알아보았다. 이러한 확률적 표본추출법에 의해 추출된 표본은 모두 표본오차를 가지고 있다. 표본의 성질 중 표본평균을 중심으로 이러한 개념을 알아보았으며 이 또한 확률변수임을 알아보았고 이에 대한 이해를 기댓값과 표준오차로 하였다. 표본평균의 분포는 표본의 크기가 커지면 커질수록 정규분포에 근사하는 성질, 즉 중심극한정리를 이해하는데 많은 지면을 할애하였다. 그만큼 중심극한 정리에 대한 이해는 매우 중요하다. 모집단의 모양새에 상관없이 표본의 크기가 충분히 크면 표본평균의 분포는 정규분포로 근사가 된다. 이러한 성질은 통계적 추정과 검정에서 매우 중요한 역할을 한다. 그리고 맨 마지막으로 표본의 크기를 정하는 문제를 다루고 표본오차의 개념을 시뮬레이션을 통해 확인하였다.

9장 연습문제

9.1 6면인 주사위를 여러 개 준비하기 바란다.

(1) n=9, 16, 36개의 주사위를 동시에 던져 나오는 눈의 평균을 기록하는 행위를 100번하여 평균에 대한 히스토그램을 그려보라. n이 커짐에 따른 차이점을 알 수 있는가?
(2) n=9, 16, 36개의 주사위를 동시에 던져 나오는 눈의 주사위의 합을 기록하는 행위를 100번하여 합에 대한 히스토그램을 그려보라. n이 커짐에 따른 차이점을 알 수 있는가?
(3) n=9, 16, 36개의 주사위를 동시에 던져 나오는 개개의 눈에 대해 4보다 같거나 큰 숫자가 나오면 1로 기록하고 그렇지 않으면 0으로 기록한다. 그런 다음 이런 n개의 수를 합한 다음 n으로 나눈다. 이런 행위를 100번하여 이러한 비율에 대한 히스토그램을 그려보라. n이 커짐에 따른 차이점을 알 수 있는가?

이상과 같은 실험은 물리적인 주사위를 던지지 않아도 엑셀로 가능하다. 엑셀 명령문 =int(1+ rand()*(7-1))을 사용하여 실험을 시행하면 된다.

이와 같은 훈련을 통해 평균, 합, 비율 등의 표본추출분포를 파악할 수 있어야 한다. 주사위의 개수를 늘리면 어떤 형태로 변하는지 짐작이 되는가?

9.2 모집단을 구성하고 있는 자료가 {1,2,3,4,5} 라고 가정해보자. 이들 자료의 모평균 μ=3이고 모표준편차 σ=1.4이다. 이 모집단으로부터 3개씩 복원추출하여 아래와 같이 표본을 구성해 보자. 5개의 숫자에서 3개씩을 뽑는 표본의 수는 5×5×5 로써 125개가 존재한다.

※ 복원추출이란 주머니에 5개의 숫자 {1,2,3,4,5} 가 있을 때, 이로부터 1이라는 숫자를 뽑은 후 다시 주머니에 넣고 다시 뽑을 수 있는 방법을 말한다.

표본번호	표본			평균
1	1	1	1	1
2	1	1	2	1.3333333
3	1	1	3	1.6666667
4	1	1	4	2
…	…	…	…	…
125	5	5	5	5

(1) 각각의 표본들로부터 계산된 125개 평균에 대해 히스토그램을 작성하라. 어떤 모양인가?
(2) 평균들의 평균을 계산한 값이 모평균 μ 의 값과 같은가?
(3) 평균들의 표준편차를 계산한 값이 $\sigma/\sqrt{n}$ 값과 같은가? (이를 표준오차라고 한다.)
(4) 125개의 표본들 가운데에서 나온 평균값들을 살펴보았을 때, 여러분들은 어느 정도(전체 몇 %)를 믿을 수 있겠는가?

쉬어가기

1. 비확률추출법

통계조사를 하는 모든 경우에 엄격한 의미의 확률추출법을 적용해야 되는 것은 아니고 조사의 편리성에 의해 비확률추출법이 사용되기도 한다. 예를 들어 정치여론조사나 신문사 여론조사의 경우 수치적인 정확도에 대해서는 관심이 없고 국민들의 정서나 여론만을 파악하고자 하는 경우가 있다. 또한 경우에 따라서는 연구하는 사람들이 어떤 생각을 하고 있는지를 사전에 살펴 연구범위를 결정하기 위해서 또는 함께 연관된 변수들이 어떤 것들이 있는가를 알아보기 위해서 사전연구를 계획하기도 한다. 실제로 길가는 사람들이나 수업을 받는 한 교실 학생들, 인터넷 접속자 등과 같이 주위에서 쉽게 접할 수 있는 사람들을 표본으로 정하여 유용한 결과를 얻을 수 있는데 이 경우를 비확률추출법(Non-Probability Sampling)이라고 한다.

비확률추출법이란 각 추출단위들이 표본에 추출될 확률을 객관적으로 나타낼 수 없는 표본추출법을 말한다. 비확률추출법은 확률추출법에 비해 훨씬 간편하고 경제적이라는 장점이 있지만 추정의 정확성을 평가할 수 없고, 표본추출에 조사자의 주관이 개입되어 표본자료로부터 분석된 결론을 모집단으로 일반화 할 수 없기 때문에 과학적인 조사방법으로 활용될 수 없다. 하지만 비확률추출법은 비용, 시간, 조사의 편리함 때문에 현실에서 자주 사용되고 있다.

일반적으로 비확률추출법은 모집단을 정확하게 규정지을 수 없는 경우, 표본오차가 큰 문제가 되지 않는 경우, 본 조사에 앞서 진행되는 새로운 개념에 대한 탐색적 연구 등에 사용하게 된다. 대표적인 비확률추출법으로 간편추출법(Convenience Sampling). 판단추출법(Judgement Sampling), 할당추출법(Quota Sampling), 눈덩이추출법(Snowball Sampling) 등이 있다.

① 간편추출법

아파트 분양신청을 마친 고객들을 대상으로 면접조사를 하는 경우 목표모집단은 분양신청자이고 확률표본을 얻기 위해서는 우선 모든 분양신청자에 대한 추출틀을 작성하는 것이 필요할 것이다. 그러나 이 경우에 추출틀의 작성이 거의 불가능하므로 조사자가 분양을 마친 사람들 중에서 일부를 주관적 판단에 따라 조사대상자로 선정하여 조사하게 된다.

이와 같이 응답자를 선정하는데 있어서 조사원 개인의 자의적 판단에 따라 간편한 방법으로 표본을 추출하는 것을 간편추출법이라고 한다. 그러나 이러한 표본추출법은 얻어진 표본이 목표모집단을 얼마나 잘 대표하는지 알 수 없고 얻어진 통계치에 대한 통계적 정확성을 평가할 수 없다.

예가 될 수 있는 조사에 다음과 같은 것들이 있다.

- 방송국행사에 자발적으로 참여한 사람을 대상으로 조사연구하는 경우
- 어떤 연구에서 특정 학교 학생들을 표본으로 선정하는 경우
- 시청 앞을 지나는 사람을 대상으로 새해의 설계를 설문하는 경우

- TV시사프로그램에서 특정사안에 대하여 ARS를 이용하여 여론조사를 하는 경우

② 판단추출법

판단추출법은 조사자가 주관적인 지식과 경험에 의해 모집단을 가장 잘 대표한다고 여기는 표본을 주관적으로 선정하는 방법이다. 교육과학기술부의 수능담당부서에서 우리나라 전체 학생들의 평균성적을 알아보기 위해서 전체 학생들의 성적을 대표한다고 생각되는 몇 학교를 나름대로 선택했다면 이것이 바로 판단표본이 된다. 이처럼 판단추출법을 쓰면 조사자의 주관적 판단에 의해서 표본이 추출되기 때문에 그 표본을 통해 얻은 추정치의 정확성에 대해서 평가할 수 없다.

일반적으로 판단추출법은 표본의 크기가 아주 작은 경우에 사용되고 표본의 크기가 커지면 확률추출법을 사용한다.

③ 할당추출법

할당추출법은 비확률추출법의 하나로 조사목적과 밀접하게 관련되어 있는 조사 대상자의 연령이나 성별과 같은 변수값에 따라서 모집단을 부분집단으로 구분하고, 모집단의 부분집단별 구성비율과 표본의 부분집단별 구성비율이 유사하도록 표본을 선정하는 방법이다. 가령 한 병원의 서비스 만족도를 조사하고자 한다면 기존의 자료에 의거하여 환자연령별, 과별, 성별 비율을 알아본 다음 그 비율에 따라 표본을 연령별, 과별, 성별로 할당하는 방법이다. 일단 각 속성별로 표본의 크기가 정해지고 나면 조사원은 정해진 크기대로 표본을 선정하는데 동일한 속성 내에서 누구를 표본으로 선택할 것이냐는 전적으로 조사원이 결정하도록 하는 방법이다.

할당추출은 비용이 적게 들고 손쉽기 때문에 단기간에 조사해야 하는 경우에 맞다. 이 방법은 조사목적과 관련되어 있고 일부 중요변수를 고려하여 표본을 추출하므로 두드러지게 나타나는 오차를 줄일 수 있지만 경우에 따라 심각한 오차가 발생할 수도 있다.

④ 눈덩이추출법

눈덩이 추출법은 접근이 어렵거나 추출틀의 작성이 곤란한 특정 집단에 대한 조사에서 사용되는 방법이다. 먼저 해당집단에 속하는 것으로 사전에 알고 있는 사람들을 대상으로 사람들을 소개받아서 조사를 진행하는 방법이다. 이와 같은 소개과정을 통하여 표본이 눈덩이처럼 점점 커지게 된다. 예를 들어 오토바이 폭주족들의 의식을 조사하거나 우리나라의 외국인 근로자를 대상으로 기업체에 대한 인식을 조사하는 경우라면 모집단 리스트의 작성이 쉽지 않아 눈덩이 추출법을 이용하면 비교적 쉽게 응답자를 찾을 수 있다.

2. 절사법과 응용절사법

① 절사법(cut-off method)

절사법은 모집단에서 관심이 있는 특성치의 분포가 한쪽으로 편중되어 있고 작은 규모에 대한 신뢰성 있는 표본틀이 없는 경우에 주로 사용하는 방법으로 전체 특성치 합에서 90%이상을 차지하지만 그 수는 작고 비중이 큰 것만을 조사하여 전체를 추정할 수 있게 하는 표본추출방법이다. 그러나 이러한 방법은 변화가 심한 모집단의 경우에는 주의를 요하고 또한 전체 특성치합의 90%이하일 경우에는 좋은 추정치의 결과를 기대하기 어렵다. 따라서 이를 보완하는 방법이 응용절사법이다.

② 응용절사법(modified cut-off method)

우리나라의 전국 사업체 분포를 예로 살펴보면, 규모(종사자수 등)별로 피라미드식 구조를 하고 있다고 볼 수 있다. 모집단의 특성에 절대적인 영향을 미치는 대규모 사업체는 소수인 반면, 영향의 정도가 미미한 소규모 사업체는 대부분을 차지한다. 따라서 전체 모수추정에 상당한 기여를 하는 대규모 사업체는 전수조사를 하고, 미미한 기여를 하는 소규모 사업체는 표본조사를 실시하는 응용절사법이 효율적이다. 응용절사법은 조사업무 부담과 비표본오차를 최소화하기 위해 표본규모를 최소화하는 절사점(cut-off point)을 찾아 전수조사와 표본조사 사업체수를 결정한다. 절사점이란 전수층과 표본층을 구분하여 주는 경계점이다.

3. The design is biased

체계적으로 한 쪽으로 치우친 통계적 연구는 그 설계가 편향(biased)되었다고 한다. 응답자가 자발적으로 참여하여 생성된 표본을 자발적 반응표본이라 한다. 뽑히기 쉬운 개체로 구성된 표본을 편의표본(convenience sample)이라고 한다. 자발적 반응표본과 편의표본은 편향되어있기 매우 쉽다.

제 10 장

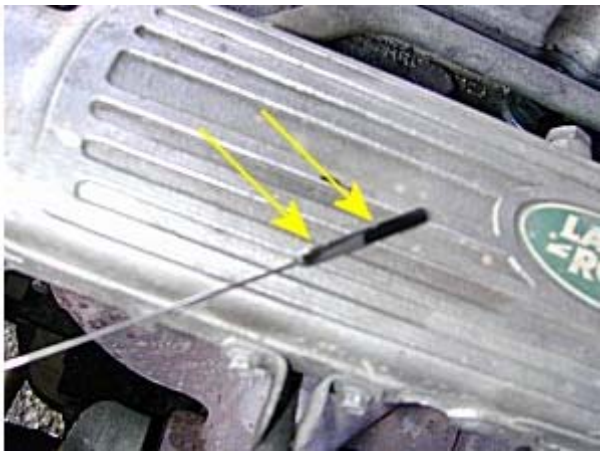
믿을 수 있는 구간이 필요하다.

차 례

학습목표

제 9장에서 우리는 점추정(point estimation)의 개념에 대해 알아보았다. 모집단 평균을 하나의 값으로 가지고 추정한다면 표본평균의 값으로 추정했다. 그러나 하나의 표본에서 나오는 한 개의 표본평균값만을 가지고 모집단의 평균을 추정한다는 것은 뭔가 부족한 면이 없지 않다. 왜냐하면 표본평균은 상수가 아니고 변수라는 점 때문이다. 이런 관점에서 나온 것이 신뢰구간 추정(confidence interval estimation)이다. 이는 모집단의 성질을 하나의 값으로 추정하지 않고 구간으로 추정하는 개념이다. 그러나 이는 새로운 내용의 또 다른 전개가 아니다. 이미 표본평균의 분포를 알고 있고 표준오차(SE)에 대한 정보를 가지고 있는 상태에서는 이를 구간추정으로 이야기 하는 것은 그렇게 어렵지 않다. 신뢰구간 작성에 대한 자세한 기술적인 내용은 본문에서 설명하겠지만 제 9장에서 배운 내용을 토대로 작성된다.

10.1 신뢰구간은 왜 필요한가?



엔진오일 게이지에도 신뢰구간이 설정되어 있다.

예제 10.1 **관리자는 신뢰구간을 좋아한다.**

신상품의 매출을 출하하려고 판매량에 대한 표본 조사한 결과 예상 판매량의 평균이 5,000개가 나오고 (3,000, 7,000)개의 신뢰구간이 나왔다고 보자.

어느 경영관리자는 평균을 기준으로 의사결정을 하는 경우도 있겠지만 신뢰구간의 하한값 혹은 상한값을 가지고 의사결정을 하는 경우도 있을 것이다. 하한값을 가지고 하는 경우는 보수적인 입장이며 상한값을 가지고 결정한다면 공격적인 의사결정이 될 것이다. ■

예제 10.2 **국세청도 신뢰구간을 좋아한다.**

다른 한 예로 국세청과 같은 세금 추징기관이 어느 기업의 소득에 대한 세금을 부과하기 위해 표본조사를 한 결과 (10억, 22억)이란 소득에 대한 신뢰구간을 구하였다고 보자.

중간 값을 가지고 세금을 부여하면 기업으로부터 강력한 반발을 가져올 것이다. 왜냐하면 반은 과다하게 세금을 부여하게 되는 결과를 초래하기 때문이다. 물론 해당기업이 법원에 소원을 신청하면 법원이 판단을 하여야 하겠지만 말이다. 그렇다고 보수적인 입장에서 하한값을 가지고 세금을 부여하게 되면 언론에서 과소하게 세금을 부여한다고 비판받을 것이다.

그렇다면 방법은 하나일 것이다. 정확한 기업소득을 추정하기 위해 표본 감사의 크기를 늘려야 한다. 그렇다면 당연히 하한값은 올라가고 세금징수기관의 입장에서도 언론과 기업의 비판을 피할 수 있을 것이다. 그러나 이는 비용에 관한 문제로 무한정 표본감사의 크기를 늘릴 수는 없다. 최적 표본의 크기를 정하는 문제가 대두가 된다. ■

- 신뢰구간을 작성하는 핵심적인 내용은 이미 우리가 배운 표본추출분포(sampling distribution)이다. 주어진 공식대로 값을 대입하여 기계적인 계산을 하기보다는 표본추출분포에 대한 이해를 통한 방법으로 설명하고자 한다.

- 일반적으로 모집단의 성질에 추론을 하고자 할 때는 표본평균과 같은 통계량(statistic)의 표본추출분포(sampling distribution)에 근거해 추론을 한다. 제 9장에서는 평균의 표본추출분포에 대해서만 논의하였지만 두 모집단의 평균의 차이, 두 모집단의 비율의 차이 등과 같은 다른 형태의 모집단의 성질에 대해서도 유사한 개념이 정립된다. 다만 이를 위해 좀 더 세밀한 설명이 필요할 뿐이다.

- 표본평균 $\overline{X}$는 n이 충분히 크면 모집단의 분포의 생김새에 상관없이 그 표본추출분포는 평균이 μ고 표준편차가 $\sigma/\sqrt{n}$ 인 정규분포로 근사한다고 하였다. 여기서 μ는 모집단의 평균, 그리고 σ는 모집단의 표준편차를 의미한다.

  이는 식 (10.1)과 같이 표준화된 확률변수 Z 는

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \qquad (10.1)$$

평균이 0이고 표준편차가 1인 정규분포로 근사한다는 이야기와 같다.

- 앞으로 우리는 이러한 사실에 근거해 μ에 대한 구간추정을 실시할 것이다. 그러나 여기에는 하나의 문제가 있다. 일반적으로 모집단의 표준편차 σ가 알려져 있는 경우는 드물다. 이러한 모수 σ 는 모집단의 성질이기는 하나 우리가 관심을 가지는 대상이 아니기 때문에 이와 같은 모수를 장애모수(nuisance parameter)라 부른다.

- 이러한 모수를 대체할 후보로는 표본표준편차 s가 있다. 그러나 σ를 표본표준편차로 대입하면 하나의 변동요인을 더 도입하는 관계로(즉, 표본마다 평균 뿐 아니라 표준편차도 달라진다.) 더 이상 표본추출분포는 정규분포를 근사하지 않는다. 대신 정규분포와 유사한 t-분포가 이를 대신한다.

t-분포

- 새로운 표본추출분포 t-분포에 대해 알아보도록 하자. 그리고 크기가 n인 표본으로부터 모집단의 평균 μ에 대한 추정을 할 것이다.

  모집단은 정규분포를 하고 있으며 알지 못하는 표준편차 σ를 가지고 있다. 지금부터는 σ대

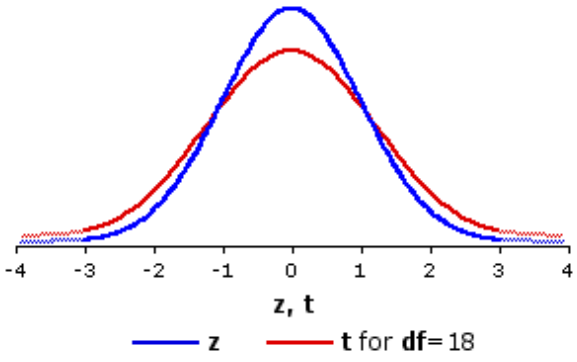
신 표본표준편차 s를 대입시킨 표준화된 확률변수에 기초하여 모집단의 평균에 대한 추정을 하고자 한다. 이와 같은 추정이 가능한 이유는 식 (10.2)와 같이 표준화된 변수

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \qquad (10.2)$$

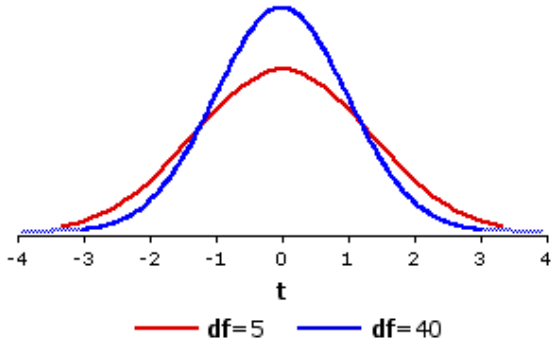
는 자유도 n-1인 t-분포를 따른다고 알려져 있기 때문이다. 여기서 자유도란 t-분포의 정확한 모양을 결정지어 주는 값이다. t-분포를 이야기 할 때마다 자유도는 반드시 명시를 하여야 한다, 표본의 크기가 n인 경우는 자유도는 n-1이 된다.

t-분포의 창시자 W. S. Gossett은 Student란 필명을 평생 썼다.

t-분포는 정규분포와 비슷한 모양을 하고 있다. 0을 평균으로 하고 종의 모양을 하고 있는데 정규분포보다 약간 흐트러진(spread out) 분포이다. 자유도가 작으면 작을수록 흐트러짐의 정도가 심하다. 그러나 자유도가 30을 넘어가면 정규분포와 거의 비슷한 모양을 하고 있다. [그림 10.1]은 자유도 18 일 때 t-분포를 정규분포와 비교한 그림이다. [그림 10.2]는 자유도가 5와 40일 때 t-분포의 비교이다. 물론 자유도가 30을 넘었기 때문에 자유도가 40인 경우는 표준정규분포와 모양새가 거의 같다.



[그림 10.1] 자유도 18인 t-분포와 표준정규분포 비교



[그림 10.2] 자유도 5와 40인 t-분포 비교

우리는 8장에서 엑셀의 =normdist() 이나 =norminv()의 명령문의 구조를 배운 바 있다. t-분포도 이와 비슷한 명령문이 엑셀에 존재한다. 그러나 정규분포와의 명령문 구조와는 약간 다르다.

=tdist(value, df, 1)

에서 value는 반드시 음이 아니어야 한다. df는 자유도이며, 마지막 인자는 1이면 =tdist는 분포 상에서 입력되는 value보다 오른 쪽에 있는 확률 값을 제공한다. 정규분포에서 해당하는 엑셀 명령문은 =1-normsdist(value)이다. 그리고 마지막 인자가 2이면

=tdist(value, df, 2)

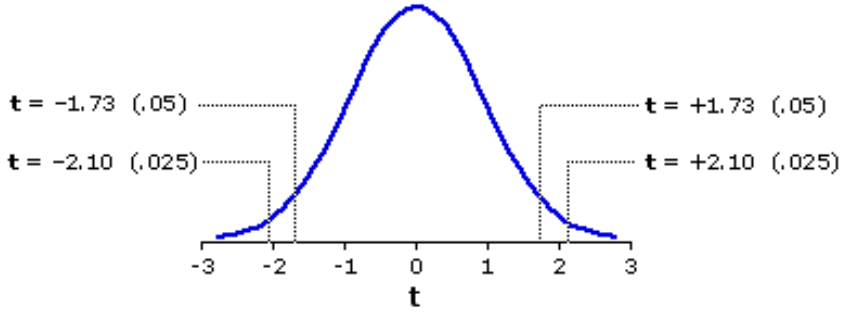
는 value보다 큰 확률뿐 아니라 value보다 작은 확률까지 합해서 구해진다. 따라서 이 명령문은 =tdist(value, df, 1)의 두 배 값을 제공한다.

다음과 같은 =tinv(prob, df) 명령문은

=tinv(0.1, df)

자유도가 df일때 분포의 양쪽 극단의 확률의 합이 10% 가 되는, 즉, 한쪽 끝의 확률이 5%가 되는 t-value를 제공하여 준다. 이는 =normsinv(1-prob/2) 에 해당하는 명령문이다.

- 보다시피 엑셀에서는 t분포와 표준정규분포의 명령문의 구조는 다르게 설정되었다. 왜 엑셀에서 다른 구조로 명령문을 만들었는지 모르지만 독자들이 불편을 느끼는 것은 어쩔 수 없는 상황이다. [표 10.1]에 몇 가지 예를 들어 놓았으니 독자들은 엑셀에서 명령문을 직접 확인하여 보는 것이 좋다. <t-값.xls>



	A	B	C	D
1	표본크기	19		
2	자유도	18		
3				
4	단측확률			Formula
5	t-값	2.100		
6	오른쪽 꼬리부분 확률	0.025		B6, =TDIST(B5,B2,1)
7				
8	양측확률			
9	t-값	1.730		
10	양쪽꼬리부분 확률	0.101		B10, =TDIST(B9,B2,2)
11				
12	tinv 계산			
13	오른쪽 꼬리부분 확률	0.025		
14	t-값	2.101		B14, =TINV(2*B13,B2)
15				
16	양쪽꼬리 확률	0.100		
17	t-값	1.734		B17, =TINV(B16,B2)

[표 10.1] 다양한 t-값 계산

- 모집단의 표준편차를 모를 때 사용하는 t-분포에 대해 알아보았다. 물론 t-분포가 이런 경우에만 쓰이는 것은 아니다. 두 평균의 차이에 대한 추론을 하는 경우에도 쓰인다. 물론 표준편차는 당연히 알려지지 않는 경우일 것이다.

- 또한 t-분포나 정규분포만이 표본추출분포의 전부는 아니다. 카이제곱분포와 F-분포 등이 있다. 이런 분포는 평균에 대한 추론을 하는 경우보다는 모집단의 분산에 대한 추론을 할 때 나타난다.

10.2 평균에 대한 신뢰구간

어떤 확률적인 표본추출법에 의해 만들어진 자료가 있다고 보자. 이러한 자료로부터 모집단의 성질인 모수에 대한 점추정을 하고 점추정에 대한 정확성(accuracy)을 가늠하기 위해 신뢰구간을 계산한다.

모집단 평균에 대한 신뢰구간을 구하고 이에 대한 해석을 시뮬레이션을 통해 알아보도록 하자.

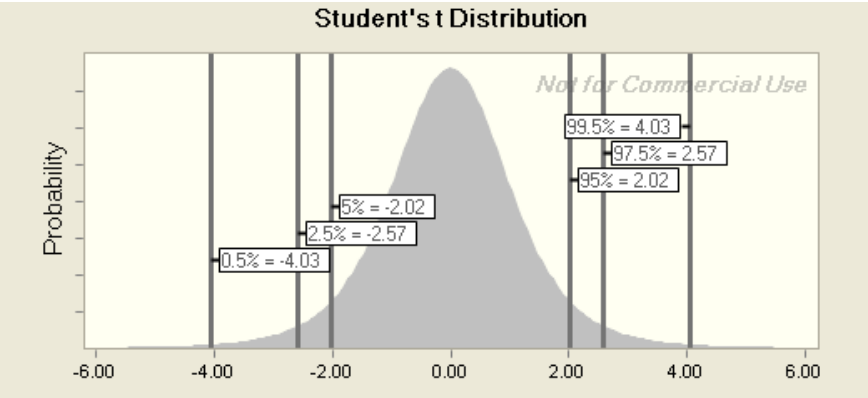
- 언제나 모집단의 평균에 대한 추정은 표본평균으로 한다. 먼저 μ에 대한 신뢰구간을 구하기 위해서는 신뢰수준(confidence level)을 정하여야 한다. 신뢰수준은 통상적으로 90%, 95%, 99% 중 하나의 값으로 정한다.
- 그런 다음 점추정값의 표본추출분포를 이용하여 주어진 신뢰수준을 달성할 만큼 점추정값의 양쪽으로 구해지는 표준오차의 폭(multiple)을 구한다. 만약 95%의 신뢰수준이라면 폭의 길이는 약 2정도 될 것이다. 이는 개념적으로 표본의 크기를 정하는 문제에 있어서 한계(B)를 $SE(\overline{X})$ 의 2배 가량의 크기로 정하는 것과 같다. 좀 더 정확하게 t-value의 크기이다.

t-value는 표본평균이 모집단의 평균과 얼마나 다른지 표준오차의 배수로 나타내는 숫자를 의미한다. 식 (10.3)이 μ에 대한 전형적인 신뢰구간의 형태이다.

$$\overline{X} \pm \text{ t-mutiple } \times SE(\overline{X}) \qquad (10.3)$$

여기서 $SE(\overline{X}) = s/\sqrt{n}$ 이다.

- 그럼 좀 더 정확한 t-multiple을 구하여 보자. α를 1-신뢰수준(소수점으로 나타냄) 이라 하자. 예를 들어 신뢰수준이 90%면 α는 0.10이다. 그러면 t-multiple은 $\alpha/2$의 확률을 자유도 n-1의 t-분포의 양쪽 끝에서 잘라내는 값이다. 자유도가 28이고 신뢰수준이 95%라면 [그림 10.3]에서 이 값은 2.57이 된다.



[그림 10.3] 자유도 28인 t-분포

따라서 평균에 대한 95% 신뢰구간은 다음과 같다.

$$\overline{X} \pm 2.57(s/\sqrt{n})$$

만약 신뢰수준이 90%라면 t-multiple은 1.70이 되어 평균에 대한 90% 신뢰구간은

$$\overline{X} \pm 2.02(s/\sqrt{n})$$

이 된다. 마지막으로 99%라면 t-multiple은 2.76이므로 평균에 대한 99% 신뢰구간은 다음과 같다.

$$\overline{X} \pm 4.03(s/\sqrt{n})$$

신뢰수준은 얼마나 크게 할 것인가?

신뢰구간 폭은 신뢰수준이 커지면 커질수록 늘어난다. 극단적인 경우가 되겠지만 신뢰수준이 100%라면 음의 무한대부터 양의 무한대 구간이 신뢰구간이 된다. 일반적으로 표본의 크기 n이 커지면 표준오차 $s/\sqrt{n}$ 가 줄어들므로 주어진 신뢰수준에 대해서는 신뢰구간의 폭이 통상적으로 줄어난다. 그러나 표본의 크기가 늘어나면 표본표준편차의 값 역시 늘어날 가능성이 높기 때문에 실제로는 반드시 줄어든다고 확신하지는 못한다.

예제 10.3 **점추정보다는 신뢰구간으로 의사결정하기가 더 좋다.**

[표 10.2]는 어느 지자체가 주관한 행사에 들어온 입장객 40명을 대상으로 만족도를 조사한 자료이다. 척도는 1점에서 10점까지로 하여 10이면 매우 만족이고 1이면 매우 불만족이다. 평균에 대한 신뢰구간을 공식을 이용하여 95% 신뢰수준을 구하였다. <만족도조사.xls>

	A	B	C	D	E	F	G
3	고객	만족도					
4	1	7					
5	2	5					
6	3	5			표본크기	40	
7	4	6			표본평균	6.250	
8	5	8			표본표준편차	1.597	
9	6	7					
10	7	6			평균에 대한 신뢰구간		
11	8	7			신뢰수준	95.0%	
12	9	10			표본평균	6.250	
13	10	7			표준오차	0.253	
14	11	9			자유도	39	
15	12	5			하한값	5.739	
16	13	5			상한값	6.761	
17	14	8					
18	15	8					
19	16	6			FORMULAS FROM RANGE F6:F8,F11:F16		
20	17	7			F6. =COUNT(A4:A43)		
21	18	8			F7. =AVERAGE(B4:B43)		
22	19	7			F8. =STDEV(B4:B43)		
23	20	5			F12. =AVERAGE(B4:B43)		
24	21	5			F13. =STDEV(B4:B43)/SQRT(COUNT(B4:B43))		
25	22	5			F14. =COUNT(B4:B43)-1		
26	23	5			F15. =F12-TINV(1-F11,F14)*F13		
27	24	5			F16. =F12+TINV(1-F11,F14)*F13		
41	38	9					
42	39	5					
43	40	4					

[표 10.2] 만족도 조사

95% 신뢰수준에서 신뢰구간은

$$\overline{X} \pm \text{ t-multiple } \times SE(\overline{X}) \;=\; 6.250 \pm 2.023 \cdot 1.597/\sqrt{40} = (5.739,\ 6.761)$$

여기서 t-multiple은 엑셀 명령문 =tinv(0.05, 39)으로 구하였다. 만약 만족도의 기준이 되는 점수가 7이였다면 이 지자체에서 실시한 사업에 대해 방문객이 느끼는 만족도는 만족스럽지 못한 것이다. 왜냐하면 신뢰구간의 상한값(6.761)이라도 7에 미치지 못하지 않은가? 또한 하한값 5.739는 기준에서 너무나 멀리 떨어져 있다. 다음 번 행사를 할 경우는 만족도의 불만이 어느 이유에서 나왔는지 좀 더 세밀한 분석을 통해 정책결정에 활용해야 한다. ■

다른 내용으로 넘어가기 전에 몇 가지 짚어야 할 사항이 있다.

- 40명의 손님이 전체 모집단을 대표하는 표본으로 추출하였다면 추정 그 자체에는 별 문제가 되지 않는다. 그러나 t-분포는 모집단이 정규분포라는 가정을 하였기 때문에 이러한 가정에 대한 언급을 하는 것이 그 순서일 것이다. 과연 1-10까지의 척도를 가지고 측정한 소비자의 만족도가 과연 정규분포를 가정해도 되는지 여부이다. 먼저 척도는 이산형의 구조를 가지고 있고 정규분포는 연속형이기 때문에 당연히 가정에 맞지 않는다. 그러나 이 경우는 그렇게 문제가 되지 않는다. 먼저 t-분포에 의거한 신뢰구간은 모집단의 분포가 정규분포가 아니라 하더라도 그와 유사한 모양의 형태를 가지고 있는 (좌우대칭) 분포라면 작성된 신뢰구간은 강건(robust) 한 성질을 가지고 있어 별로 문제가 되지 않기 때문이다. 또 하나의 이유는 표

본의 크기가 40이므로 중심극한 정리에서 이야기하는 일반적인 크기 30을 초과한다. 따라서 신뢰구간의 타당성은 부여받을 수 있다.

- 신뢰구간의 의미는 무엇인가? 평균이 신뢰구간 [5.739, 6.761]에 있을 것에 대한 신뢰수준을 95% 부여하였는데 이는 "평균이 이 구간에 있을 확률이 95%이다." 라는 의미는 아니다. 모집단의 평균은 이 신뢰구간에 있을 수도 있고 없을 수도 있다. 여기서 95% 신뢰라는 말은 신뢰구간을 구하는 절차에서 찾아 볼 수 있다. 즉, 같은 모집단에서 크기 n의 표본을 추출하여 평균에 대한 신뢰구간을 구하는 행위를 수많이 하였다고 보자. 95%의 의미는 그렇게 만들어진 신뢰구간의 95%는 참의 모집단의 평균을 포함하고 있다는 의미이다.

  엑셀을 이용하여 시뮬레이션 하여 보자.

- 평균이 100이고 표준편차가 20인 정규분포로부터 크기가 30인 표본을 무작위로 추출하여 표본평균과 표본 표준편차를 이용하여 1,000개의 신뢰구간을 구하였다. 이와 같은 절차는 이미 표본의 크기를 정하는 문제에서 본 바 있을 것이다. 1,000개의 신뢰구간 중에서 약 948개가 참의 평균 100을 포함하고 있었으며 나머지는 포함하고 있지 않았다. 시뮬레이션 결과이기 때문에 정확하게 950개(95% × 1,000번)를 적중시키지 못하였지만 근접한 948개가 평균을 포함하고 있었다. 이러한 의미가 95%의 신뢰수준이다. 즉 특정의 신뢰구간은 참의 평균을 포함하고 있든지 아니든지 하겠지만 이러한 행위를 반복하면 이중 95%에 해당하는 신뢰구간은 모집단의 평균을 포함하고 있다는 의미이다. <신뢰수준.xls>

	A	B	C	D	E	F	G	H	I	J
3	모집단평균	100			표본크기	30				
4	모집단표준편차	20			표본평균	102.128				
5					표본표준편차	18.636				
6		확률표본								
7		104.75		평균에 대한 신뢰구간						
8		127.19			신뢰수준	95.000%				
9		111.55			표본평균	102.128				
10		65.99			표준오차	3.402				
11		87.33			자유도	29				
12		95.02			하한값	95.169				
13		95.77			상한값	109.086				
14		98.36			모집단 평균 포함여부	1				
15		113.62								
16		112.67								
17		59.46		모집단 평균이 신뢰구간에 포함된 경우						
18		118.35			94.8%					
19		121.96		Data table used to replicate confidence intervals						
20		122.50		Rep	Mean captured?					
21		114.34			1					
22		97.44		1	1					
23		143.00		2	1		B7. =NORMINV(RAND(),$B$3,$B$4)			
24		89.35		3	1		F9. =COUNT(Data)			
25		99.18		4	1		F10. =AVERAGE(Data)			
26		95.86		5	1		F11. =STDEV(Data)			
27		82.70		6	1		F15. =F10			
28		96.12		7	1		F16. =F11/SQRT(F9)			
29		99.83		8	1		F17. =F9-1			
30		101.38		9	1		F18. =F15-TINV(1-F14,F17)*F16			
31		90.92		10	1		F19. =F15+TINV(1-F14,F17)*F16			
32		120.90		11	1		F20. =IF(AND(B3>=F18,B3<=F19),1,0)			
33		102.40		12	1					
34		130.45		13	1					
35		88.83		14	1					
36		76.57		15	1					

[표 10.3] 신뢰수준 의미

10.3 전체 합에 대한 신뢰구간

간혹 평균보다는 전체 합에 대한 관심을 가지는 경우가 있다. 예를 들면 회계분야에서는 계정당 받을 평균금액이 얼마나 되는지 보다는 전체 받을 금액이 얼마나 되는지에 더 관심을 가질 수 있다. 전체 합을 T라는 모수라 하고 $\hat{T}$을 T의 점추정량이라 하자.

모집단의 크기가 N이고 표본의 크기가 n이라면 식 (10.4)와 같이 $\hat{T}$을 구한다. 표본의 값을 다 더한 다음 전체 모집단의 크기 N에 대해 비율로 추정하는 것이다.

$$\hat{T}= \frac{N}{n} T_s = N\overline{X} \tag{10.4}$$

여기서 T_s 는 표본의 전체 합이다.

$\overline{X}$ 와 마찬가지로 $\hat{T}$도 표본추출분포를 가지고 있을 것이다. 그러나 식 (10.4)에서 N이 상수이므로 기본적으로 $\overline{X}$ 의 표본추출분포와 마찬가지 성질을 적용하면 된다. $\hat{T}$의 평균과 표준편차는 식 (10.5)과 같다.

$$\begin{aligned} E(\hat{T}) &= T \\ SE(\hat{T}) &= N\sigma/\sqrt{n} \end{aligned} \tag{10.5}$$

$\hat{T}$ 는 $\overline{X}$ 와 마찬가지로 기댓값을 구하면 모집단의 모수 T와 같다. 따라서 $\hat{T}$는 경우에 따라 모집단의 성질 T를 과다추정할 수도 있고 과소추정할 수도 있다. σ를 모를 경우는 표본의 표준편차를 대입하여 식 (10.6)과 같이 표준오차의 값을 구한다.

$$SE(\hat{T}) = Ns/\sqrt{n} = N \times \mathrm{SE}(\overline{\mathrm{X}}) \tag{10.6}$$

표본평균이나 평균의 표준오차에 N을 곱하면 바로 $\hat{T}$과 그의 표준오차를 구한다는 사실은 매우 편리하게 T에 대한 신뢰구간을 구하게 한다.

$$\hat{T} \pm \text{ t-multiple } \times SE(\hat{T})$$

혹은

$$N\overline{\mathrm{X}} \pm \text{ t-multiple } \times N \times \mathrm{SE}(\overline{\mathrm{X}})$$

전체 합에 대한 신뢰구간을 구하는 예제는 표본평균과 마찬가지 형식이므로 생략하기로 한다.

10.4 비율에 대한 신뢰구간

이번 여론은 현재 논란이 되고 있는 부동산 정책에 대한 조사 결과에서도 확인할 수 있다. 국민들은 향후 부동산 정책의 방향에 대해 '현재보다 규제를 강화해야 한다' 38.6%, '현재의 방향을 유지해야 한다' 11.6% 등으로 답해, 현재의 정책기조를 유지하거나 외려 강화해야 한다는 응답이 전체의 50%를 웃돌았다. 반면 '현재보다 규제를 완화하는 방향'이라는 응답은 44.9%였다. 연구소는 여당의 현 지지층과 전통적 지지층에서는 규제강화에 대한 응답이 훨씬 높게 나왔다고 밝혔다.
한편 이번 조사는 한국사회여론연구소(KSOI)가 디오피니언에 의뢰, 전국의 성인남녀 700명을 대상으로 13일 실시됐으며, 95% 신뢰수준에 오차범위는 $\pm 3.7\%$이다.

예제 10.4 **뉴스를 분석한다.**

신문지상에서 가끔 우리는 위와 같은 뉴스를 접한다. 위 기사를 읽으면 "대통령이 추진하고 있는 부동산 정책에 대해 국민의 50.2%는 동의하고 있으며 오차허용범위는 3.7%이다." 로 요약되는데 이런 뉴스를 접할 때 얼마나 많은 사람이 이해를 하고 정확한 해석을 할까? 하는 의문점은 심히 염려스러운 수준이다. 가끔 표본의 크기를 알려 주지 않는 신문기사라면 그리고 표본이 잘 표본추출되었는가 의심을 받는 경우라면 우리는 이런 발표에 대한 신뢰를 보낼 수가 없다. 이 발표를 심도 있게 분석해 보자.

이러한 뉴스는 모집단의 성질 중 비율, 즉 몇 명 중에서 몇 명이 정책에 대한 지지를 하였느냐에 대해 추정을 하는 경우다. 통계학에서는 이런 모집단의 성질인 모비율을 p라 표기한다.

p에 대한 신뢰구간은 모집단의 평균에 대한 신뢰구간과 같이 구하는 절차는 비슷하다.

- 점추정값을 구하고 점추정에 대한 표준오차를 구한 다음 신뢰수준에 맞는 multiple을 구하면 된다. 일반적인 형태는 다음과 같다.

$$\text{point estimate} \quad \pm \text{ multiple} \times \text{standard error}$$

예제 10.4에서 점추정값이 50.2%가 된다. 그리고 multiple $\times$ standard error는 3.7%이다. 그러므로 신뢰구간은 46.5%에서 53.9%이다. 만약 뉴스에서 밝힌 신뢰수준이 95%라면 대통령의 경제정책에 동의하는 비율은 46.5%에서 53.9%에 있을 것이라고 95% 신뢰한다는 의미이다.

여기서 신뢰수준은 이미 언급한 바와 같은 의미로 해석이 되어야 함은 물론이다. 자 이제 뉴스의 내용을 조금 더 과학적으로 분해하자.

만약 표본의 크기가 n일 때 $\hat{p}$ 라는 통계량을 다음과 같이 정의한다.

$$\hat{p} = \frac{\text{원하는 속성을 갖는 것의 개수}}{n}$$

이러한 $\hat{p}$를 표본비율(sample proportion)이라 한다. 700명을 조사했는데 352명이 동의를 했기 때문에 $\hat{p}$ =352/700 = 0.502가 된다. 이 $\hat{p}$ 가 모집단의 p 의 점추정 값이 된다.

n이 충분히 크다면 이러한

- $\hat{p}$ 의 표본추출분포(sampling distribution)의 평균은 p고 표준편차(standard deviation)는 $\sqrt{p(1-p)/n}$ 이 된다.
- 그러나 p를 모르기 때문에 표본비율 $\hat{p}$ 를 대입하여 식 (10.7)과 같이 $\hat{p}$ 의 표준오차를 구한다.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{10.7}$$

- 그리고 신뢰구간을 구하기 위해 사용되는 multiple은 z-값이다. 왜냐하면 표본의 크기를 크다고 가정을 하기 때문에 중심극한 정리에 의해서 이다. 따라서 식 (10.8)이 p에 대한 신뢰구간이다.

$$\hat{p} \pm \text{ z-multiple } \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{10.8}$$

따라서 위 뉴스에서 표본이 700이니 95% 신뢰구간은 다음과 같이 설정된다.

$$0.502 \pm 1.96\sqrt{\frac{0.502(1-0.502)}{700}}$$

오차 폭을 구해보면 0.037이 나온다. 그래서 앞의 기사에서 '95% 신뢰수준에 오차범위는 $\pm 3.7\%$'라는 표현을 쓴 것이다.

그리고 신뢰구간의 하한값이 p_L 그리고 상한값을 p_U 라 한다면 $np_L > 5, n(1-p_L) > 5$, $np_U > 5, n(1-p_U) > 5$ 을 만족하면 n은 충분히 크다고 할 수 있으며 z-multiple을 이용하여 신뢰구간을 쓰는데 타당성을 부여받는다. 그렇지 않으면 쓰는데 조심하여야 한다. 물론 위의 경우는 정당성을 부여 받는다. ■

예제 10.5 **신뢰구간은 점추정보다 더 많은 정보를 제공한다.**

예제 10.3에서 지자체 관리자는 만족도의 평균 자체보다는 6이상의 만족도를 보인 고객의 비율에 대해서도 관심을 가지고 있다고 보자. 신뢰수준은 95%라 하자. 40명 중 16명이 6이상의 만족도를 표시하였기 때문에 표본비율 $\hat{p}$ 는 16/40= 40%가 된다. 따라서 신뢰구간은 식 (10.6)에 의해

$$0.40 \pm 1.960\sqrt{\frac{0.40(1-0.40)}{40}}$$

인

[0.248, 0.552]

이 나온다. 또한 $p_L = 0.248,\ p_U = 0.552$ 를 이용하여 $np_L > 5, n(1-p_L) > 5, \quad np_U > 5, n(1-p_U) > 5$ 의 조건은 모두 만족하므로 신뢰구간의 타당성에 문제가 없다. ■ <신뢰구간.xls>

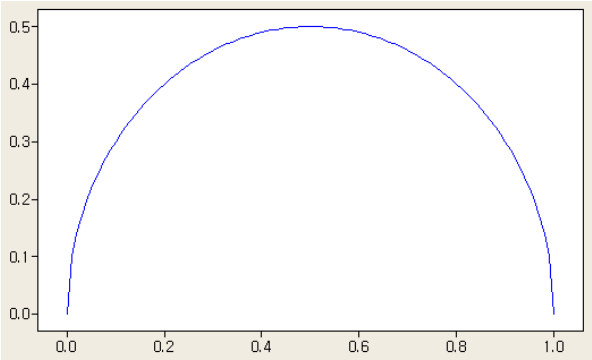
예제 10.6 **여론조사에서 언급하는 표본의 크기는 왜 1,000명일까?**

> **2) 지지율 53%와 47%는 실제 몇% 차**=그렇지 않다. 신뢰도가 95%이고 표본오차가 ±3.1%포인트라면 통계적 의미에서는 차이가 없다. 오차범위를 감안할 때 지지율 53%는 확률적으로 대략 '56%~50%'의 구간대에, 47%는 '50%~44%' 구간대에 분포함을 의미한다. 분명 두 지지율 구간대에 서로 겹치는 부분이 형성된다. 이 때문에 이를 1위, 2위로 표현하거나 '누가 앞선다'로 해석하는 건 엄밀하게 문제가 있다는 지적이다. '경합'으로 표현하는 게 옳으며 이보다 차이가 적다면 '박빙 접전' '백중세' 등으로 다뤄져야한다.
>
> 냉혹한 현실에서 '51% 대 49%'의 근소한 표차로 대통령 당락의 희비가 엇갈리는 것과 여론조사 세계는 다르다. 이런 점에서 양승찬 숙명여대 언론정보학부 교수는 과학적 조사방법에 입각한 조사와 함께 "결과에 대한 조사기관과 언론 등의 엄밀한 해석"을 강조한다. 여론조사 자체가 문제가 아니라 조사 설계부터 표집, 해석에 이르기까지의 단계에서 공정성, 정밀성이 담보될 경우 '신뢰할 수 있다'는 견해가 지배적이다.
>
> **3) 조사대상은 꼭 1000명 이상**=조사대상은 꼭 1000명 이상 돼야 하는가? 많이 하면 비교적 정확해지지만 그렇다고 4000명 조사결과에 비해 1000명이 부정확한 것은 절대 아니다. 보통 1000명일 경우 신뢰구간은 95% ±3.1%이고, 700명 안팎일 경우 95% ±3.7%이다. 표본오차가 그만큼 줄어든다는 사실은 그만큼 조사결과가 더 정확해짐을 의미한다.
>
> 문제는 조사비용 때문에 무한정 표본수를 늘리는 게 부담으로 작용한다는 단점이다. 결국 비용과 표본오차를 감안한 나름대로 합리적인 표본수가 1000명인 셈이다. 인구가 많다고 이에 비례해 표본수를 꼭 늘려야하는 것은 아니다. 미국의 전국 단위 여론조사도 1000명 조사가 표준이다. 객관성, 과학성을 확보하기 위해 역시 중요한 게 표본추출방법이다. '무작위 표집'이 시간, 비용 측면에서 현실적으로 적용하기 힘들기 때문에 보통 '인구비례 할당에 의한 층화 무작위 추출'방식을 이용한다.
>
> **2007.12.02 문화일보 기사 중 일부 발췌**

앞의 기사에서 밝힌 오차의 범위가 3%라고 명시하였다면 이는 95% 신뢰수준 하에서는

$$1.96\times\sqrt{\hat{\mathrm{p}}(1-\hat{\mathrm{p}})/\mathrm{n}} = 0.03$$

의 의미가 된다. 그런데 $\hat{p}$ 가 0과 1에 아주 가까운 수가 아니면 $\hat{p}(1-\hat{p})$ 값은 상당히 안정된 상태이다. p가 0에서부터 1로 변하면 $\sqrt{\hat{p}(1-\hat{p})}$ 이 어떻게 변하는지 보라. p의 값이 0.2와 0.8상에서 $\sqrt{\hat{p}(1-\hat{p})}$ 값은 0.4와 0.5 범위 내에 있지 않은가? 그리고 p=0.5일 때 최대값이 발생한다.



$\hat{p}$ 대 $\sqrt{\hat{p}(1-\hat{p})}$

그렇다면 보수적인 n(제일 표본의 크기가 큰)을 구하기 위해 $\hat{p}$ = 0.5를 대입하여 n에 대해서 풀어야 한다.

$$1.96\sqrt{(0.5)(0.5)/n} = 0.03$$

n $\approx$ 1,067을 얻는다. 3%의 신뢰구간 폭을 얻기 위해 필요한 표본의 크기는 95% 신뢰수준일 때 경우 약 1,000명 정도의 사람만 필요한 것이다. 이는 갤럽 같은 수많은 조사기관이 표본조사를 시행할 때 많은 수의 응답자를 필요로 하지는 않는 이유가 된다. 미국과 같은 넓은 나라도 보통 1,500명을 넘지 않는다. ■

예제 10.7 **회계분야도 표본론의 대상이다.**

비율에 대한 신뢰구간의 중요성은 회계학분야의 감리(auditing)에서도 찾을 수 있다. 감리사는 통칭 속성 표본(attribute sampling)을 이용하여 어떤 회계절차가 제대로 이행이 되었는지를 확인한다. 여기서 속성(attribute)은 회계절차가 정확하게 이행이 되었는지 안 되었는지 여부를 의미한다. 예를 들어 인보이스(invoice, 송장 送狀)가 회계직원에 의해 진행이 시작되었는지 여부, 허가된 가격으로 인보이스 가격이 처리가 되었는지 등이다. 회계처리가 제대로 안되어 있으면 오류로 분류한다. 감리사는 이러한 종류의 오류에 대해 항목별로 비율을 추정한다. 다만 감리사는 이러한 오류의 비율이 얼마나 커질 수 있는지 관심을 가지기 때문에 p_L 보다는 p_U 에 더 많은 관심을 기울인다. 따라서 신뢰구간도 양쪽 구간을 정하는 것이 아니라 p_U 만 설정한다. 식 (10.9)를 참조하기 바란다. <속성표본.xls>

$$p_U = \hat{p} + \text{ z-multiple } \times \sqrt{\hat{p}(1-\hat{p})/n} \qquad (10.9)$$

여기서 z-multiple은 α 의 값을 오른쪽 꼬리 부분에 전체 할당하여 만들어지는 값이다. 95%라면 z-값은 1.645가 된다. ■

그러나 이와 같은 공식은 이항분포를 정규분포로 근사할 때 나오는 공식이라 실제 감리사는 정규분포를 이용하지 않고 이항분포를 이용하여 p_U 값을 구하는 정확한(exact)방법을 즐겨 쓴다. 왜냐하면 오류의 비율이 매우 작게 나타나기 때문에 정규분포의 타당성을 찾지 못하기 때문이다. 예를 통해 이해하자.

예제 10.7(계속) 그러나 비율이 매우 낮으면 지금까지의 방법으로는 곤란하다.

95%의 신뢰수준 하에서 n=93일 경우 2개가 오류로 밝혀졌다. $\hat{p}$ =0.0215이고 정규분포를 이용하여 p_U 를 구하면

$$0.0215 + 1.645 \times \sqrt{0.0215(1-0.0215)/93} = 0.046$$

이 나온다. 그러나 $np_U = 93 \times 0.046 = 4.278 < 5$ 이므로 신뢰구간의 타당성이 부여되지 못한다.

만약 p_U 가 신뢰구간 상한값(upper confidence limit)을 만족하는 값이라면 이러한 다음의 식을 만족하여야 한다.

$$P(X \leq k) = \alpha$$

여기서 X는 모수 n과 p_U 를 가지고 있는 이항분포이다. 그리고 k는 관측된 오류의 개수이고 α 는 1-신뢰수준이다.

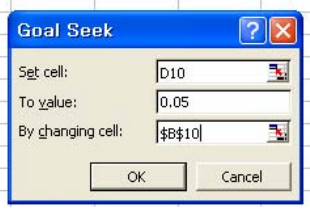
이러한 p_U 를 직접 찾을 수는 없고 엑셀의 목표값 찾기를 이용하여야 한다. [표 10.4]에서 셀 D10에

=binomdist (B4,B5, B10,1)

을 입력한다. 그런 다음 목표값 찾기에서 셀D10을 수식셀에 입력하고 0.05를 찾는 값에, 그리고 바꿀 셀에 B10을 지정하면 우리가 원하는 값을 셀 B10에서 얻을 수 있다. 이렇게 구한 값이 $p_U = 0.066$ 이다.

정규분포로 구한 값 0.046과는 많은 차이가 난다. 감리사는 오류의 비율이 6.6%보다 크지 않은 것에 대해 95% 신뢰한다고 이야기할 수 있을 것이다. ■

	A	B	C	D	E	F	G
2							
3	신뢰수준	95%		B7. =B4/B5			
4	오류의 개수	2		B13. =NORMSINV(B3)			
5	표본의 크기	93		B14. =B7+B13*SQRT(B7*(1-B7)/B5)			
6				D10. =BINOMDIST(B4,B5,B10,1)			
7	표본비율	0.0215		F10. =1-B3			
8							
9	exact 신뢰구간 상한값			목표값 설정			
10	상한값	0.066		0.050	=	0.05	
11							
12	대표본을 이용한 상한값						
13	z-multiple	1.645					
14	상한값	0.046					
15							
16							
17							
18							
19							
20							



[표 10.4] p가 낮으면 주의하라.

10.5 표준편차에 대한 신뢰구간

6시그마 운동의 첫걸음은 표준편차다.

예제 10.8 **식스시그마 운동은 식스표준편차운동으로 이름을 바꾸어야 한다.**

여기서는 표준편차에 대한 이야기를 하여 보자. 여러분 대부분은 식스시그마 운동에 대해 들어 보았을 것이다. 모 정유회사는 휘발유의 이름조차 '시그마식스'로 이름을 붙여 판매한 적이 있지 않았는가? 왜 각 기업에서는 식스시그마 운동이 조직의 사활이 걸린 것처럼 적극적인지 표준편차(영어로 표준편차는 '시그마'이므로 사실 식스시그마 운동은 식스표준편차로 명명을 하여야 한다.)의 개념에서 알아보기로 하자. ■

지금까지는 평균에 대한 신뢰구간을 구하는 방법이 제안되었다. 반면 표준편차는 단지 장애모수로서 표본평균의 표준오차를 추정하는데 필요로 하였다. 그러나 모표준편차와 같은 모집단

의 성질이 직접적으로 관심의 대상인 경우가 있다. 다만 모표준편차 σ 에 대한 신뢰구간을 구하는 방법은 평균에 대한 신뢰구간을 구하는 방법보다 복잡하다.

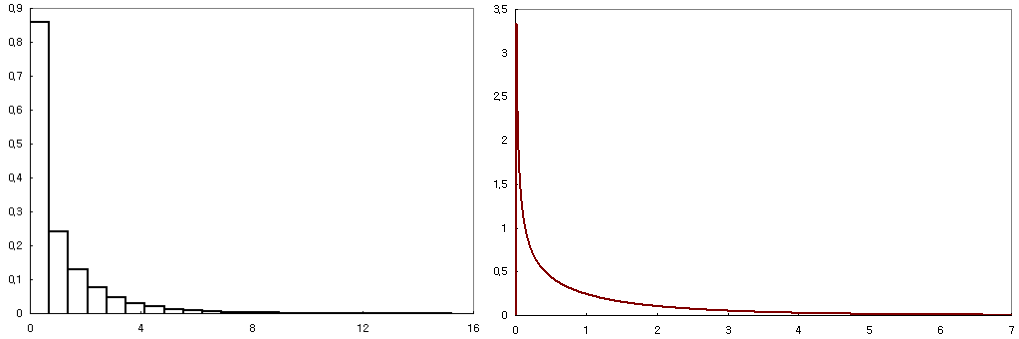
이를 위해 몇 가지 필요한 개념을 알아보아야 한다.

카이제곱분포 : 확률변수 Z_1 은 평균이 0이고 표준편차가 1인 표준정규분포를 따른다고 가정하자. 그러한 확률변수의 값을 제곱하는 새로운 확률변수

$$\chi^2 = Z_1^2$$

는 자유도가 1인 카이제곱분포를 따른다. 이러한 변수 χ^2 은 확률변수의 제곱으로 값이 정해지기 때문에 0보다 큰 값을 가질 것이고 0과 양의 무한대가 이 변수가 값을 가지는 구간이 된다. 간단한 시뮬레이션을 통해 자유도가 1인 카이제곱분포가 어떻게 생겼는지 보도록 하자.

[그림 10.4]의 왼쪽 그림은 셀 A1에 =norminv(rand(),0,1)^2을 입력한 후 이를 10,000번 시행한 결과의 히스토그램이다. 이와 같은 모양이 자유도 1인 카이제곱분포이다. 시뮬레이션의 결과이긴 하지만 시행횟수가 충분히 크기 때문에 [그림 10.4]의 오른쪽 그림인 이론적인 분포와 비슷하다.

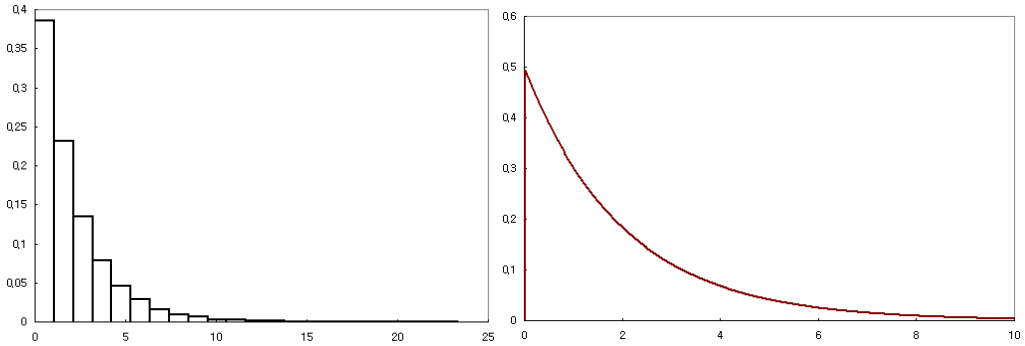


[그림 10.4] 자유도 1인 카이제곱 분포

그렇다면 표준정규분포를 따르는 확률변수 2개를 생성한 다음 제곱을 하여 더하면 어떤 결과가 나오는지 보도록 하자.

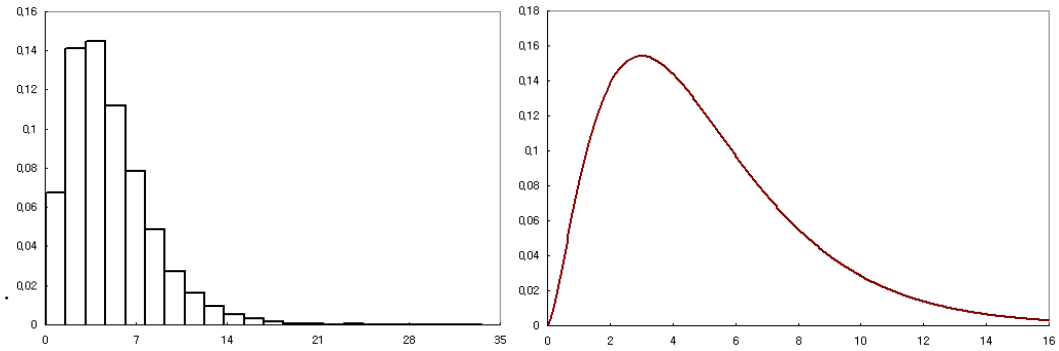
$$\chi^2 = Z_1^2 + Z_2^2$$

[그림 10.5]의 왼쪽 그림은 셀 A1과 셀 B1에 =norminv(rand(),0,1)^2을 입력한 후 셀 C1에 =sum(A1:B1)을 만들고 이를 100,000 번 시행한 결과의 히스토그램이다. 이와 같은 모양이 자유도 2인 카이제곱분포이다.

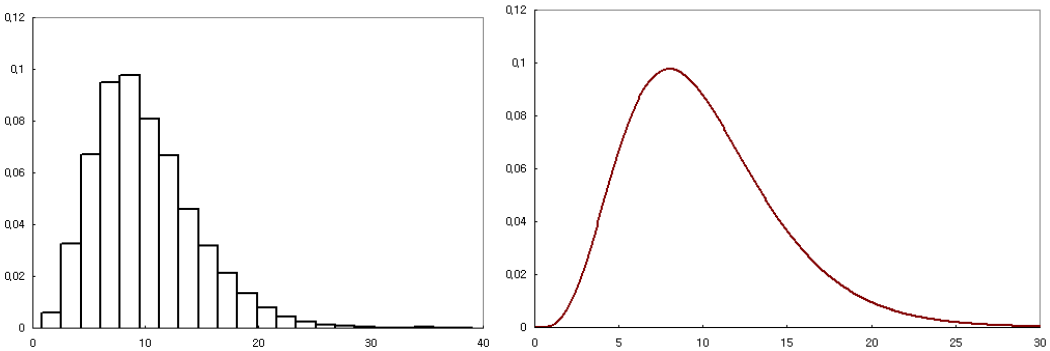


[그림 10.5] 자유도 2인 카이제곱 분포

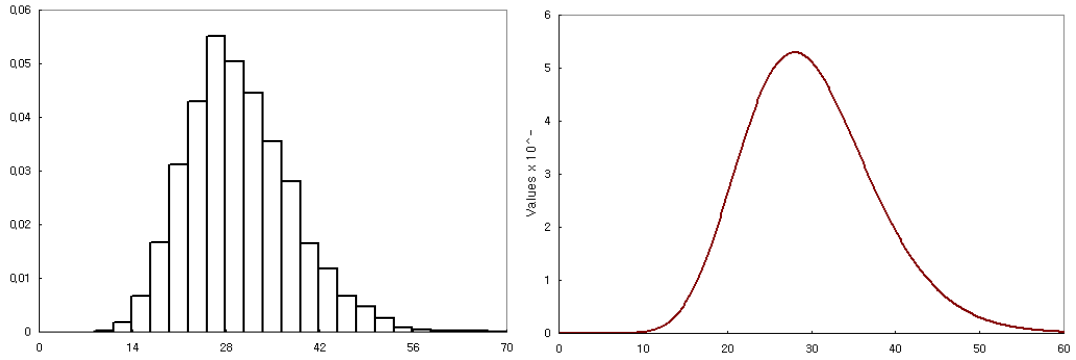
[그림 10.6]은 위로부터 각각 자유도가 5, 10, 그리고 30인 카이제곱분포의 모양이다.



자유도 5



자유도 10



자유도 30

[그림 10.6] 자유도에 따른 카이제곱 분포의 변화

[그림 10.6]의 결과를 보면 자유도가 커지면 커질수록 점점 카이제곱분포는 산의 모양을 하고 있는 것으로 나타난다. 이는 우리가 이미 배운 중심극한정리에 의해 자연스러운 현상이다. 모집단의 분포가 자유도가 1인 카이제곱분포라 하더라도 n이 충분히 크다면 합으로 표시되는 확률변수는 정규분포로 근사할 수 있다.

자유도가 df인 카이제곱분포의 평균과 표준편차는 각각 식 (10.10)과 식 (10.11)과 같다.

$$E(\chi^2) = df \tag{10.10}$$

$$\mathrm{stdev}(\chi^2) = \sqrt{2\times \mathrm{df}} \tag{10.11}$$

이런 사실에 비추어 보면 자유도가 60인 카이제곱분포는 평균이 60 그리고 표준편차 $\sqrt{2\times 60} = \sqrt{120}$ 인 10.96의 정규분포로 근사할 것이다.

자 이제 카이제곱분포 바탕으로 표본분산에 대한 표본추출분포를 알아보도록 하자.

표본의 크기가 n인 표본의 표본분산 s^2 은 식 (10.12)와 같이 자유도가 n-1인 카이제곱분포를 따른다고 알려져 있다.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2 \quad \text{with} \quad \mathrm{df} = \mathrm{n} - 1 \tag{10.12}$$

이에 대한 자세한 설명을 할 수는 없으나 시뮬레이션을 통해 개념을 정리하여 보자.

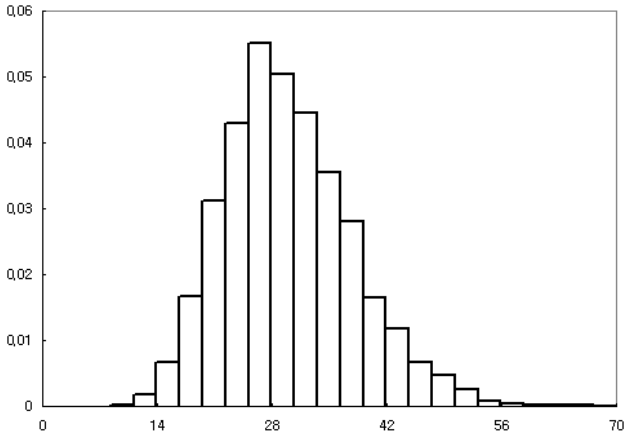
표본의 크기가 31이고 모집단이고 평균이 μ=100 이고 σ=10 인 정규분포라고 가정하자. 이러한 모집단으로부터 표본을 구성하고 표본에서 표본분산을 구하고 식 (10.12)에서 정의된 확률변수를 임의의 엑셀 셀 C1에 저장한다. 엑셀에서는 셀 A1:A31에

=norminv(rand(),100,10)

을 입력하고 임의의 셀 C1에

=30*var(A1:A31)/10^2

을 입력한다. 그리고 셀 C1의 값을 100,000번 모의시행한 후 히스토그램을 그리면 [그림 10.7]과 같다.



[그림 10.7] 식 (10.12)을 사용한 표본추출분포

시뮬레이션으로 나온 평균의 값은 30.01로 이론적인 값 30과 거의 일치하며 표준편차는 역시 7.7458로 이론적인 값 $\sqrt{2\times 30}$ = 7.7459와 거의 비슷하게 나온다. [그림 10.6]의 자유도 30인 카이제곱분포하고 거의 같은 그림이 나오지 않았는가? [그림 10.6]은 표준 정규분포를 따르는 변수 30개의 값을 제곱한 다음 그 합을 구하여 그림을 그린 것으로 자유도 30인 카이제곱 분포인 것을 기억하라.

[그림 10.7]의 분포가 우리가 원하는 분산에 대한 표본추출분포이다. 즉, 모집단이 정규분포를 하고 있으면 자유도가 n-1인 카이제곱분포를 참조하면 된다.

자유도에 따라 카이제곱분포를 따르는 확률변수가 결정되면 엑셀에서 해당하는 확률 값을 제공하는 명령문과 확률에 해당하는 카이제곱분포의 값을 제공하는 명령문을 실행하면 된다.

=chidist(v, df)
=chiinv(p, df)

여기서 v는 값을 의미하고 df는 자유도 그리고 p는 오른쪽 꼬리 부분에 해당하는 확률이다.

예를 들어 =chidist(24,30)는 0.772025로 나온다. 위의 자유도가 30인 카이제곱분포 그림에서 (물론 시뮬레이션에 의해 만든 분포이지만) 24보다 큰 확률의 크기를 짐작하여 볼 수 있을 것이다. 또한 역으로 =chiinv(0.5, 30)를 통해 이 분포의 중앙값을 구할 수 있다. 29.33603이

다. 이 값은 평균값 30보다 약간 작게 나오지만 그렇게 차이가 나지 않는다. 자유도 30인 경우에는 이미 중심극한 정리에 의해 정규분포로 근사가 가능하다는 이야기이다. 약간 오른쪽으로 왜도가 있을 뿐이다.

더 나아가 =chiinv(0.975, 30)는 16.70077, =chiinv(0.025, 30)는 46.97924를 얻는다. 따라서 $(n-1)s^2/\sigma^2$ 가 16.70077과 46.97924에 속하는 것에 대해 95% 신뢰할 수 있다.

$$16.70077 \leq \frac{(n-1)s^2}{\sigma^2} \leq 46.97924$$

이를 우리가 알고자 하는 σ^2에 대해 정리하면 다음과 같다.

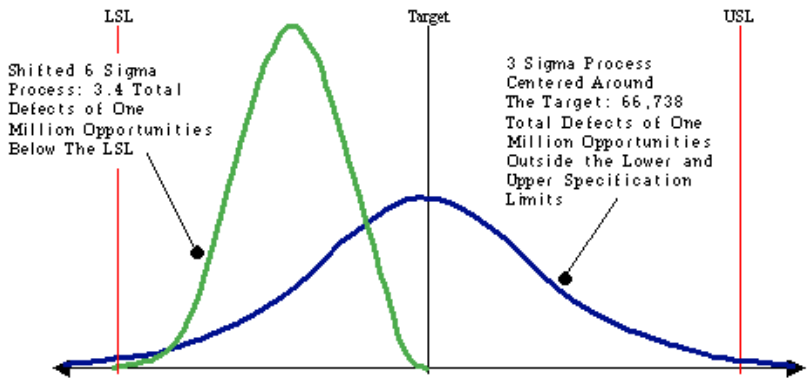
$$\frac{(n-1)s^2}{46.97924} \leq \sigma^2 \leq \frac{(n-1)s^2}{16.70077}$$

이것이 표본의 크기가 (n=31)일 때 모집단이 정규분포라 가정한다면 모집단의 성질 σ^2 에 대한 95% 신뢰구간이다. 그리고 모집단의 표준편차 σ 에 신뢰구간은 다음과 같이 구한다.

$$\sqrt{\frac{(n-1)s^2}{46.97924}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{16.70077}}$$

어려운 계산과정을 거쳐 나왔지만 우리는 아직 왜 이런 개념이 식스시그마 운동과 연관이 되는지 알지 못한다. 예제를 보는 순간 여러분들은 식스시그마 전문가의 길로 들어서는 것이다.

예제 10.9 **품질관리는 변동과의 싸움이다.**



식스시그마 운동은 변동과의 전쟁이다.

어느 기업에서 생산되는 부품의 지름의 지름은 10센티미터가 되어야 한다. 그러나 책임을 맡고 있는 생산관리자는 부품의 평균 뿐 아니라 변동에 대해서도 관심을 기울여야 한다. 전체적으로는 생산되는 부품의 평균은 10센티미터가 될지 모르지만 변동이 크게 되면 많은 부품은 규격에 맞지 않는 부품이 될 가능성이 높기 때문이다. 따라서 관리자는 50개의 부품을 표본추출하여 생산되는 부품에 평균과 변동에 대한 특징을 파악하려고 한다. 평균에 대한 신뢰구간은 [9.986, 10.005]로 나와 관리자는 목표로 삼고 있는 지름 10센티미터에 대해 제품의 평균 지름은 근접하다는 사실에 자신할 것이다. 그러나 과연 여기서 멈출 것인가?

이제 표준편차에 대한 신뢰구간을 구하여 본다. 표준편차는 0.034로 밝혀졌다. 표준편차에 대한 신뢰구간은 =chiinv(0,025, 49)= 70.222와 chiinv(0.975, 49)=30.555을 구한 다음 공식을 사용하면

$$\sqrt{\frac{49\times 0.034^2}{70.222}} \leqq \sigma \leqq \sqrt{\frac{49\times 0.034^2}{30.555}} \quad \Rightarrow \quad [0.029,\ 0.043]$$

이 나온다.

- 과연 표준편차에 대한 신뢰구간의 의미는 무엇일까? 부품의 지름이 목표로부터 0.065 센티미터를 벗어나면 이 부품은 쓰지 못한다고 하자. 그러면 얼마나 많은 부품을 쓰지 못하는 경우가 나오는지 계산하여 보자. <부품자료.xls>

[표 10.5] 계산에 의하면 참의 평균이 10센티미터이고 참의 표준편차가 표준편차의 신뢰구간의 상한값이라고 한다면 약 10.1%의 부품은 활용되지 못한다. 물론 부품의 지름은 정규분포에서 나온다고 가정한다. 왜 여러분은 참의 표준편차가 표준편차 신뢰구간의 상한값으로 잡아야 하는지 알겠는가? 얼마나 많은 부품을 쓰지 못하는 경우가 나오는지 계산하여 보면 엑셀에서

=normdist(10-0.065, 10, 0.043, 1) +(1-normdist(10+0.065, 10, 0.043, 1)

로 구하면 된다.

	A	B	C	D	E	F	G	H
1								
2	부품	지름			표본의 크기	50		
3	1	10.031			표본평균	9.996		
4	2	10.011			표본표준편차	0.034		
5	3	10.003						
6	4	10.025		평균에 대한 신뢰구간				
7	5	10.048			신뢰수준	95.0%		
8	6	10.014			표본평균	9.996		
9	7	10.030			표준오차	0.005		
10	8	10.008			자유도	49		
11	9	10.049			하한값	9.986		
12	10	9.995			상한값	10.005		
13	11	9.965						
14	12	10.003		표준편차에 대한 신뢰구간				
15	13	9.959			신뢰수준	95.0%		
16	14	10.013			표본표준편차	0.034		
17	15	10.012			자유도	49		
18	16	10.005			하한값	0.029		
19	17	9.921			상한값	0.043		
20	18	9.930						
21	19	9.990		사용할수 없는 부품비율				
22	20	9.948			허용되는 최대편차	0.065		
23	21	10.077			가정된 평균	10		
24	22	9.959			가정된 표준편차	0.043		
25	23	10.000			사용할수 없는 비율	0.131		
26	24	9.998						
27	25	9.983		F8. =COUNT(Data!Diameter)				
28	26	9.995		F9. =AVERAGE(Data!Diameter)				
29	27	9.917		F10. =STDEV(Data!Diameter)				
30	28	9.934		F14. =AVERAGE(Data!Diameter)				
31	29	10.044		F15. =STDEV(Data!Diameter)/SQRT(COUNT(Data!Diameter))				
32	30	10.023		F16. =COUNT(Data!Diameter)-1				
33	31	9.997		F17. =F14-TINV(1-F13,F16)*F15				
34	32	10.020		F18. =F14+TINV(1-F13,F16)*F15				
35	33	9.983		F22. =STDEV(Data!Diameter)				
36	34	9.998		F23. =COUNT(Data!Diameter)-1				
51	49	9.973						
52	50	9.970		셀 B4:B53 := Diameter 로 이름 지정				

[표 10.5] 부품 자료

[표 10.6]에서는 한 걸음 더 나아가 참의 평균과 참의 표준편차 값을 변했을 때 값이 어떻게 변하는지 보았다.

평균이 목표 값에 가까우면 가까울수록 그리고 표준편차 값은 작으면 작을수록 부품의 활용성은 높아진다. 그러나 그렇지 않은 경우는 15.1%까지 올라감에 주의하기 바란다.

• 즉, 품질관리는 변동과의 싸움인 것이다. ■

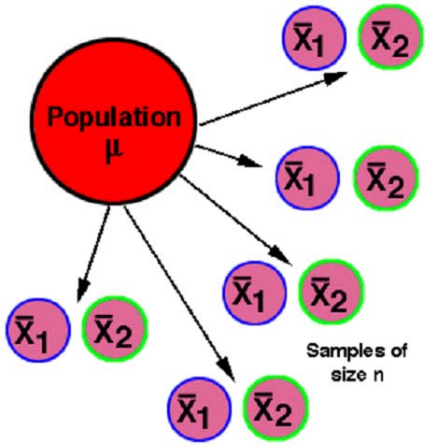
55	사용할수 없는 부품비율					
56				가정한 표준편차		
57			0.131	0.029	0.034	0.043
58		가정한 평균	9.986	0.043	0.077	0.151
59			9.996	0.026	0.058	0.132
60			10.005	0.027	0.059	0.133
61						

[표 10.6] 평균과 표준편차를 달리 했을 때의 불량률

10.6 두 평균의 차이에 대한 신뢰구간

추론의 중요한 응용분야 중 하나가 두 모집단의 평균을 비교하는 것이다. 다음과 같은 경우를 보도록 하자.

- 세일기간동안 쇼핑센터 옷가게에 오는 남자고객과 여자고객의 구매력의 차이에 대해 알고자 한다. 여자고객의 평균 구매력이 남자 고객에 비해 얼마나 차이가 나는가에 따라 판매 전략이 달라질 수 있기 때문이다.
- 항공사에서 도착시간에 관한 관리는 매우 중요하다. 자사의 출발지연시간은 다른 경쟁사의 출발지연시간과 평균적으로 얼마나 차이가 나는지는 매우 중요한 사안이다.
- 모 회사는 인턴사원을 모집하여 일종의 직무해결 능력 검사를 실시한 다음 실무 부서에 배치한다. 몇 개월 지나 다시 검사를 실시한 후 성과가 뛰어난 인턴에 대해서는 정식으로 채용을 하는 정책을 사용한다. 관리자는 3개월의 훈련과정 후에 얼마나 직무해결능력이 향상되었는지 알고자 할 것이다.



두 집단을 비교하자.

- 모두 두 집단의 평균 차이에 대한 경우이다. 그러나 처음 두 경우와 마지막 경우는 표본이 추출되는 과정이 다름을 알 수 있다. 처음 두개의 예제에서 각 집단에 들어가는 표본은 서로 다른 표본추출단위인 반면 마지막 예제는 각 집단에 들어간 단위가 같음을 알 수 있다.

표본의 크기를 $n(=n_1+n_2)$ 으로 한다면 첫 두 경우에서는 n_1 만큼은 첫 번째, 그리고 n_2 는 두 번째 표본의 크기로 하여 무작위(random)로 각 집단에서 표본을 추출하는 경우이고 세 번째의 경우는 n 명의 인턴사원에 대해 전과 후를 비교하는 것이기 때문에 각 집단에 들어가는 개체는 동일하다.

그러나 측정이 이루어진 횟수는 첫 두 번째 경우는 n_1+n_2 번이고 세 번째 경우는 $2n$ 이다. 그러므로 두 경우 모두 측정횟수는 기본적으로 같다고 보아야 한다. 첫 두 경우를 독립된 표본(independent sample)이라 부르고 세 번째 경우를 짝진(paired) 표본이라 구분하여 부른다. 이와 같이 분리하여 두 집단의 평균의 차이를 살펴보기로 한다.

10.6.1 독립된 표본

두 모집단의 평균을 각각 μ_1, μ_2 라하고 표준편차를 σ_1, σ_2 이라고 표기하자. 각각의 모집단에서 독립적으로 표본을 무작위로 추출하여 두 모집단의 평균의 차, $\mu_1-\mu_2$ 를 추정하고자 한다. 각각의 표본의 크기는 n_1, n_2 라 한다.

- 모집단의 평균의 차이에 대한 점추정은 자연스럽게 각각의 표본의 표본평균의 차이, $\overline{X_1}-\overline{X_2}$ 가 된다.
- 다음으로 이러한 점추정의 표본추출분포를 구하여야 하는데 이 역시 t-분포가 된다.

단 자유도는 n_1+n_2-2 가 된다. 각각의 표본의 자유도가 각각 n_1-1 , n_2-1이므로 이를 더하면 된다. 물론 여기서 두 모집단은 정규분포를 가정하여야 함은 물론이다. 따라서 $\mu_1-\mu_2$ 에 대한 신뢰구간은 식 (10.13)과 같다.

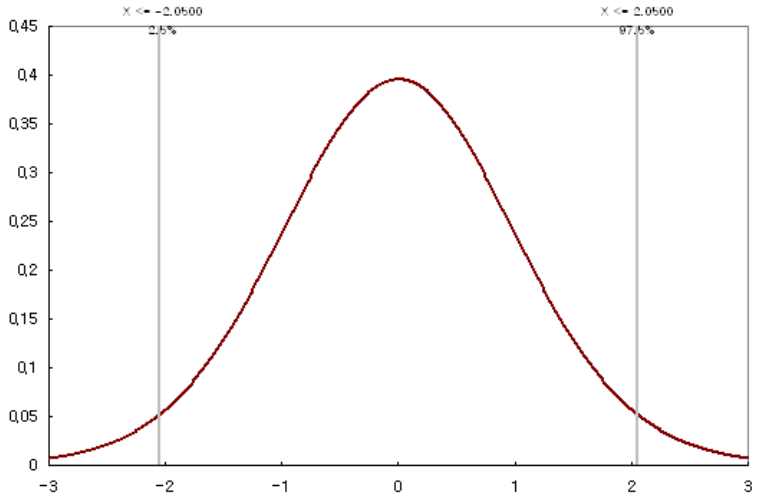
$$\overline{X_1}-\overline{X_2}\pm \text{ t-multiple } \times SE(\overline{X_1}-\overline{X_2}) \qquad (10.13)$$

t-multiple은 다른 예제들과 마찬가지로 엑셀에서 =tinv() 명령문을 이용하여 찾으면 된다. 자유도 n_1+n_2-2 인 t-분포에서 양쪽 끝의 확률을 잘라 내는 값이다.

예를 들어 $n_1=n_2=15$ 인 경우는 자유도가 (15-1)+(15-1)=28이다. 자유도가 28인 t-분포의 모양은 다음 [그림 10.8]과 같은데 신뢰수준이 95%라면 양쪽 끝에서 2.5% 씩을 잘라내는 값이 t-multiple이다. 엑셀에서는

=tinv(0.05, 28)

로 구한다. 값은 2.05이다.



[그림 10.8] 자유도가 28인 t-분포

$\overline{X_1} - \overline{X_2}$ 의 표준오차를 구하기 위해서는 일반적으로 모집단의 표준편차들은 같다고 가정하여야 한다. $\sigma_1 = \sigma_2$ 이다. 모집단의 표준편차들은 같다고 가정을 하기 때문에 다음 식 (10.14)처럼 각 표본에서 나오는 변동을 합하여 전체 자유도로 나누어주는 공동분산(pooled vzriance) s_P^2 으로부터 표준편차를 구하여 볼 수 있다.

$$s_p = \sqrt{s_p^2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n-1)s_2^2}{n_1 + n_2 - 2}} \tag{10.14}$$

여기서 s_1, s_2 는 각각의 표본표준편차를 의미하며 s_p 는 공동표준편차이다. s_p 의 p 는 공동(pooled)의 약자이다. 따라서 $\overline{X_1} - \overline{X_2}$ 의 표준오차는 식 (10.15)와 같다.

$$SE(\overline{X_1} - \overline{X_2}) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{10.15}$$

그러나 어떤 이유로 "두 모집단의 분산은 같다"라고 가정을 하지 못하는 경우가 있을 것이다. 그러면 $\overline{X_1} - \overline{X_2}$ 의 표준오차는 식 (10.16)과 같이 표기되어야 한다.

$$SE(\overline{X_1} - \overline{X_2}) = \sqrt{s_1^2/n_1 + s_2^2/n_2} \tag{10.16}$$

그러나 이 경우 적용되는 자유도는 더 이상 $n_1 + n_2 - 2$ 이 아니다. 그렇지만 n_1, n_2 의 크기가 충분히 큰 경우는 중심극한 정리가 적용이 되므로 크게 문제가 되지 않는다. 왜냐하면 중심극한 정리에 의해 t-multiple을 쓰기보다는 z-multiple을 써도 무방하기 때문이다. 그러나 작

은 경우는 식 (10.17)과 같은 자유도를 써야 한다.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad (10.17)$$

이렇게 구한 자유도는 일반적으로 정수가 되지 않는다. 따라서 정수에 제일 가까운 수를 자유도로 택한다.

예제 10.10 **두 하청업체를 비교하자.**

두 하청업체가 공급하는 모터 부품을 비교하기 위하여 각각 $n_1 =$ 29개, $n_2 =$ 23개의 표본을 추출하여 두 집단의 평균 차이에 대한 신뢰구간을 구하여 보았다. 단위는 모터가 고장을 일으킬 때까지 걸리는 시간으로 하였다. 두 업체가 제공하는 부품의 평균 차이가 있는지 없는지를 비교하는 게 목적이다. <모터자료.xls>

	A	B	C	D	E	F	G	H
1								
2	공급업자 #1	공급업자 #2						
3	1358	658				공급업자 #1	공급업자 #2	
4	793	404			표본크기	29	23	
5	587	735			표본평균	751.655	636.261	
6	608	457			표본표준편차	288.467	227.096	
7	472	431						
8	562	658		평균의	차이에 대한 신뢰구간	공동분산	이분산	
9	879	453			신뢰수준	95.0%		
10	575	488			표본평균의 차이	115.394		
11	1293	522			공동표준편차	263.233	NA	
12	1457	1247			차이에 대한 표준오차	73.4985	71.4963	
13	705	1095			자유도	50.0000	50.0000	
14	623	430			하한값	-32.2319	-28.2103	
15	725	726			상한값	263.0205	258.9989	
16	569	793						
17	424	498						
18	436	502						
19	1250	589						
20	493	975						
21	485	808						
22	462	456						
23	765	731						
24	854	491						
25	634	487						
26	1109							
27	800							
28	883							
29	522							
30	791							
31	684							

[표 10.7] 모터자료 신뢰구간

공동분산을 가정한 경우는 자유도가 50이고 신뢰구간은

$$[-32.2319,\ 263.0205]$$

가 나왔으며 그렇지 않은 경우도 자유도가 역시 50, 신뢰구간은

$$[-28.2103,\ 258.9989]$$

로 비슷한 결과를 얻었다. 두 신뢰구간 모두 0을 포함하고 있지 않은가?

- 신뢰구간이 0을 포함한다는 이야기는 두 하청업체의 평균 차이가 유의수준 95%에서는 없다는 의미이다. ■

10.6.2 짝진 표본

쌍인 경우 더 좋은 예술이 나온다.

예제 10.11 **짝진 표본은 독립된 표본보다 평균의 차이를 찾는데 좋다.**

다시 한번 예를 들어 짝 표본의 의미를 알아보도록 하자. 자동차라는 상품은 부부가 구매를 하는 경우에 배우자 한 명이 단독으로 구매를 결정하지는 않을 것이다. 부부가 자동차를 구매

하러 나온 경우 현재 세일즈맨의 판매 방법이 여자보다는 남자에게 더 호의적이라고 판단되는지 알고자 35쌍의 부부에게 판매 방법에 대해 1부터 10까지 척도로 물어 보았다. 1이면 호의적이지 않고 10이면 매우 호의적이다. [표 10.8]이 그 결과이다. <짝진표본자료.xls>

	A	B	C
1			
2	쌍(pair)	남편	아내
3	1	6	3
4	2	6	9
5	3	8	5
6	4	6	5
7	5	7	5
8	6	7	6
9	7	9	5
10	8	7	6
11	9	8	9
12	10	7	6
13	11	6	3
14	12	6	4
15	13	8	5
16	14	7	7
17	15	7	5
18	16	6	5
19	17	7	4
20	18	5	4
21	19	5	4
22	20	10	11
23	21	7	9
24	22	10	7
25	23	6	4
26	24	7	3
27	25	6	5
28	26	9	4
29	27	9	7
30	28	7	4
31	29	5	4
32	30	8	4
33	31	7	4
34	32	5	2
35	33	7	4
36	34	6	4
37	35	11	4

[표 10.8] 짝진 표본자료

이러한 자료를 짝진 표본이라 한다.

여기서는 남자와 여자가 독립적으로 표본추출이 된 경우가 아니다. 자연스럽게 짝이 이루어진 경우이다. 각 쌍에서 나오는 남자의 반응과 여자의 반응은 같은 환경 하에서 측정이 된 것이므로 독립을 기대하지 못하다. 실제로 두 표본의 상관계수 값은 0.398이 나온다. 이럴 경우는 이 표본을 독립인 두 개의 표본이라고 가정하고 절차를 취하게 되면 잘못된 추정을 하게 된다.

실제로 위의 자료를 독립인 두 표본처럼 두 모집단의 평균의 차이에 대한 신뢰구간을 구하면 95% 신뢰수준에서

[1.111, 2.775]

이 나온다. 이에 대한 계산은 독자에게 맡기겠다. 그러나 이는 두 집단 간에 양의 상관계수가 존재한다는 사실을 고려치 않은 잘못된 절차일 수밖에 없다.

정확한 절차는 먼저 각 쌍의 관측값의 차이를 계산하는 것부터 출발한다. 왜냐하면 각 쌍에 있는 남자와 여자의 관측값은 독립이 아닐지 모르나 이렇게 구한 30개의 관측값, 즉 $\triangle = \text{husband} - \text{wife}$ 는 독립이기 때문이다. 따라서 한 개의 표본으로 평균에 대한 신뢰구간을 구하는 절차에 의해 두 집단 간의 평균 차이를 구할 수 있다.

	A	B	C	D	E	F	G	H
1								
2	쌍	남편	아내	차이				
3	1	6	3	3			짝진 표본의 크기	35
4	2	6	9	-3			표본평균	1.943
5	3	8	5	3			표본표준편차	1.939
6	4	6	5	1				
7	5	7	5	2		평균의 차이에 대한 신뢰구간		
8	6	7	6	1			신뢰수준	95.0%
9	7	9	5	4			표본평균	1.943
10	8	7	6	1			표준오차	0.328
11	9	8	9	-1			자유도	34
12	10	7	6	1			하한값	1.277
13	11	6	3	3			상한값	2.609
14	12	6	4	2				
15	13	8	5	3				
16	14	7	7	0				
17	15	7	5	2		H3. =COUNT(C3:C37)		
18	16	6	5	1		H4. =AVERAGE(D3:D37)		
19	17	7	4	3		H5. =STDEV(D3:D37)		
20	18	5	4	1		H9. =AVERAGE(D3:D37)		
21	19	5	4	1		H10. =STDEV(D3:D37)/SQRT(H3		
22	20	10	11	-1		H11. =H3-1		
23	21	7	9	-2		H12. =H9-TINV(1-H8,H11)*H10		
24	22	10	7	3		H13. =H9+TINV(1-H8,H11)*H10		
25	23	6	4	2				
26	24	7	3	4				
27	25	6	5	1				
28	26	9	4	5				
29	27	9	7	2				
30	28	7	4	3				
31	29	5	4	1				
32	30	8	4	4				
33	31	7	4	3				
34	32	5	2	3				
35	33	7	4	3				
36	34	6	4	2				
37	35	11	4	7				

[표 10.9] 짝진 표본 신뢰구간

여기서 자유도는 쌍의 개수에서 1을 뺀 수 34이다. 95% 신뢰구간은

[1.277, 2.609]

이다. 두 독립된 표본을 가정하였을 때보다 폭이 적은 신뢰구간을 얻었다. 따라서 좀더 효율적인 추정을 하였다고 할 수 있을 것이다.

만약 35명의 남자에게 무작위로 반응을 측정하고 독립적으로 다른 35명의 여자를 무작위로 표본을 추출하여 두 독립된 표본에 의거 두 집단의 평균 차이를 구한다고 하자. 짝진표본에서 자유도는 적어지지만 (n_1+n_2-2 에서 $n-1$ 로) 반면에 $SE(\overline{X_1}-\overline{X_2})$ 의 값이 작아지므로 t-multiple이 커지는 것을 상쇄하고도 남는 경우이다. 짝진 표본조사가 독립표본에 비하여 효과적인 경우이다. ■

10.7 두 비율의 차이에 대한 신뢰구간

예제 10.12 **두 그룹의 반응 비율은 다른가?**

어느 기업체가 특정상품에 대해 세일을 기획하고 있다. 일단의 소비자에게 우편을 발송하여 이를 알리고자 한다. 그러나 소비자를 두 그룹으로 나누어 한 그룹에는 세일 가격에서 추가로 5% 할인할 수 있는 쿠폰을 끼워서 발송을 하고 다른 그룹은 쿠폰을 끼워서 발송하지 않았다. 지자체 관리자는 세일에 반응하는 정도가 차이 있는지를 알고자 한다. 여기서 반응은 소비자가 지자체로부터 실제로 세일 기간동안 특정제품을 구매한 것을 의미한다. 이렇듯 두 집단의 비율 차이가 관심의 대상이 되곤 한다. ■ <구매비율.xls>

p_1, p_2 를 두 모집단의 알려지지 않은 비율이라고 하자. 여기서 비율은 어떤 특정한 속성을 가지고 있는지를 따져주는 개념이 된다. 가지고 있으면 1, 그렇지 않으면 0으로 표기하기도 한다. 그리고 $\hat{p_1}, \hat{p_2}$ 는 각각의 표본 n_1개와 n_2개에서 계산한 표본비율이라고 하자. 두 모집단의 비율의 차이의 점추정값은 $\hat{p_1}-\hat{p_2}$ 로 대체한다. 그리고 하나의 표본에서 $\hat{p}$ 가 정규분포를 따라 갔듯이 $\hat{p_1}-\hat{p_2}$ 역시 정규분포를 따라간다고 이야기 할 수 있다. 그러므로 p_1-p_2 에 대한 신뢰구간은 식 (10.18)과 같다.

$$\hat{p_1}-\hat{p_2} \pm \text{ z-multiple } \times SE(\hat{p_1}-\hat{p_2}) \tag{10.18}$$

z-multiple은 예제 13.4의 경우와 마찬가지로 정하면 된다. 95%이면 1.96이다. $\hat{p}_1-\hat{p}_2$ 의 표준오차는 식 (10.19)와 같다.

$$SE(\hat{p_1}-\hat{p_2}) = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \tag{10.19}$$

두 표본은 독립적으로 추출되었기 때문에 각각의 표본의 분산을 합해 제곱근을 취하는 것이다.

쿠폰을 받은 고객의 150명 중에서 55명은 물건을 구매하였으며 그렇지 않은 고객 150명은

35명만 구매를 하였다.

	A	B	C	D	E
1					
2		구매	비구매	합	
3	큐폰 발송그룹	55	95	150	
4	큐폰 미발송그룹	35	115	150	
5					
6	구매비율				
7	큐폰발송그룹	0.3667			
8	큐폰 미발송그룹	0.2333			
9					
10	구매비율의 차이	0.1333			
11	차이의 표준오차	0.0524			
12					
13	신뢰수준	95%			
14	z-multiple	1.960			
15					
16	비율의 차이에 대한 신뢰구간				
17	하한값	0.0307			
18	상한값	0.2359			
19					
20					
21	D4. =SUM(B4:C4)				
22	D5. =SUM(B5:C5)				
23	B8. =B4/D4				
24	B9. =B5/D5				
25	B11. =B8-B9				
26	B12. =SQRT(B8*(1-B8)/D4+B9*(1-B9)/D5)				
27	B15. =NORMSINV(B14+(1-B14)/2)				
28	B18. =B11-B15*B12				
29	B19. =B11+B15*B12				

[표 10.10] 구매비율 차이 신뢰구간

두 비율의 차이에 대한 95% 신뢰구간은

[0.0307, 0.2359]

로 나타난다. 이 신뢰구간의 의미는 고객 100명을 기준으로 했을 때 쿠폰 발행으로 말미암아 쿠폰을 발급하지 않았다면 구매하지 않았을 3명 내지 23명까지의 새로운 구매고객을 유발할 수 있다는 뜻이다.

그러나 이 의미가 바로 이익의 큰 차이로 직결되는 사항은 아니다. 예를 들어 고객의 평균 구매액이 4만원이라 하자. 그리고 5천원이 이익이라면 쿠폰을 받지 않는 100명의 고객당 발생하는 이윤은 5천원×(0.2333)(100) = 126,650원인 반면 쿠폰을 받은 100명의 고객당 발생하는 이윤은 3천원×(0.3667)(100) = 110,010원이 된다. 왜냐하면 4만원의 5%인 2천원은 쿠폰으로 지불되기 때문이다. 따라서 쿠폰의 비율을 높여 구매고객을 더 유인하느냐 아니면 쿠폰의 비율을 낮추어도 이익을 낼 만큼 구매고객을 유인할 수 있는 정책을 개발하는 것은 전적으로 관리자의 몫이 된다.

예제 10.13 신문기사를 한번 짚고 넘어가자.

문 후보와 정 후보 지지율 오차범위 안으로 접어 들었다.

창조한국당 제주도당, 선대본부에 자발적인 봉사자 '속속'…고무적인 분위기 전해

문국현 대선 후보의 지지율이 대통합민주신당 정동영 대선 후보와의 격차가 오차범위 안으로 조혀졌다고 창조한국당 제주도당이 밝혔다.

또한 창조한국당 제주선대본부에 자발적인 봉사자가 속속 모여들고 있다며 고무적인 분위기를 전했다.

28일 창조한국당 제주도당은 보도자료를 통해 후보등록 마감일 직전 일부 언론이 실시한 여론조사 결과를 제시하며 "지난 24일 실시된 한겨레-리서치플러스(1000명, 95%신뢰수준에 ±3.1%P) 조사에서 문 후보는 지지율 8.0%로 대통합민주신당 정동영 후보(11.3%)와의 격차를 오차범위 안으로 좁혔다"며 "이는 1주일 전인 17일과 비교할 때 정 후보는 13.2%에서 하락한 반면, 문 후보는 1.4%포인트 상승한 것"이라고 밝혔다.

또한 도당은 "25일 실시된 조선일보-한국갤럽(1068명, 95%신뢰수준에 ±3.1%P) 조사에서도 문 후보는 1주일 전(6.6%)에 비해 소폭 오른 8.4% 지지율을 얻었다"며 "이와 같은 여론조사는 문 후보가 정 후보와 호남권을 제외한 전국에서 지지율 3위 경쟁을 벌이고 있다"고 분석했다.

더욱이 도당은 "호남권을 제외한 전국 지지율을 따져보면 두 사람이 박빙 또는 오차범위 내(한겨레 0.4%, 조선 2% 차)에서 각축전을 벌이고 있"며 "또 수도권에서는 '한겨레'조사에서 문 후보 9.4%, 정 후보 8.1%로 1.3%포인트 앞서고, '조선일보' 조사에선 10.5% 대 11.8%로 사실상 이미 '박빙 접전 구조'에 접어든 것으로 파악하고 있다"며 강조했다.

도당은 "본격적인 선거운동이 시작된 이후 문 후보 제주선대본 사무실에는 자발적 선거운동을 희망하는 시민들이 속속들이 모여들고 있다"며 "이들은 캠프측이 거리유세에 필요한 유세차량과 어깨띠 등이 준비되지 않아 거리유세 시기를 늦추자, 오히려 이에 대해 항의하는 등 문국현 후보 운동에 대한 강한 의지를 드러내 관계자들을 당황시키기도 한다"며 고무적인 분위기를 전했다.

한편, 제주선대본부는 28일 오후부터 거리유세에 착수하기로 하고, 자발적 운동원들을 중심으로 2~3개조를 편성 주요길목 홍보와 수목원, 시내 공원, 산책로 등을 중심으로 선거운동에 본격 착수한다는 계획이다.

2007.11.29 이슈 제주 기사 중 일부 발췌

위 기사에서 문국현 후보와 정동영 후보의 지지율 격차는 오차범위 0.031 내에서 차이가 없다는 기사이다. 과연 그런가?

$$(0.113 - 0.080) \pm 0.031 = 0.033 \pm 0.031 = (0.002, 0.064)$$

이 신뢰구간은 0을 포함하고 있지 않기 때문에 문국현 후보와 정동영 후보의 지지율 차이는 95% 신뢰수준에서 오차 범위를 벗어나고 있다. 따라서 오차 범위 안으로 접어들었다는 표현을

잘못하였다. 그럼 오차범위를 계산하는 과정에서 잘못한 것인가?

그러면 오차범위 0.031은 어디서 나온 숫자인가? 표본은 1,000명이다. 1,000명에게 문국현 후보와 정동영 후보의 지지에 대해 각각 물었기 때문에 n_1, n_2를 각 1,000명으로 가정하고 오차 허용범위를 구할 수는 없다. 만약 이렇게 구하면 오차범위는

$$1.96 \cdot SE(\widehat{p_1} - \widehat{p_2})$$

가 되고 계산하면 0.02584가 나온다. 이는 틀린 계산이다.

두 표본 t-검정과 이 문제는 확연히 다르다. 집단 간의 평균 비교가 아니고 1,000명에게 두 후보의 지지도를 물어보는 문제로 두 후보의 지지율을 따로 구해 독립된 표본으로 처리해서는 안 된다.

한 사람에게 지지하는 후보 한명을 선택하라고 하였기 때문에 이 표본은 엄격한 의미에서 짝진 표본이 되는 것이다. 평균의 차이는 여전히 0.033이지만 표본오차는 다음과 같이 계산하여야 한다. 정동영 후보와 문국현 후보 중 1명을 지지하면 1을 배정하고 그렇지 않으면 0을 배정하는 코딩을 하였다면 두 값의 차이로 짝진 표본의 최대 표준오차를 구하면 다음과 같다.

$$\sqrt{\frac{0.5 \times (1 - 0.5)}{1,000}} = 0.0153$$

따라서 신뢰구간은 앞에서 살펴본 것처럼,

$$0.033 \pm 1.96 \times 0.0153 = 0.033 \pm 0.031 = (0.002, 0.064)$$

이 된다.

두 지지율의 차이는 0.113-0.08 = 0.033인데 이는 분명히 주장하는 차의 범위인 0.031 보다 분명히 크다. ■

10.8 신뢰구간 폭의 통제

지금까지 논의된 신뢰구간 유형의 폭을 통제하는 방법에 대해 알아보도록 하자.

신뢰구간은 아래 세 가지 항목의 함수이다.

- 통계량
- 신뢰수준

• 표본의 크기이다.

많은 사람들은 “표본에 있는 자료는 무작위로 얻어졌다는 이유로 우리가 통제할 수 있는 것은 별로 없다.”라고 인식을 하는 경우가 많지만 사실은 그렇지 않다.

표본의 변동 폭을 줄일 수 있는 적절한 표본추출방법을 쓰면 이러한 문제를 어느 정도 해결할 수 있다. 이는 층화표본추출방법을 쓰는 이유 중의 하나이다. 그리고 짝진 표본의 예에서 보았듯이 두 독립된 표본추출방법보다 폭이 적은 신뢰구간을 구하는 방법을 보기도 하였다. 이렇듯 적절한 표본 및 실험계획법은 자료의 변동을 줄여 주어진 표본의 크기에서 효과적인 추론을 가능케 한다.

두 번째는 신뢰수준이다. 그러나 신뢰수준이 올라가면 폭은 커지고 내려가면 폭은 좁아지기 때문에 사실 신뢰수준을 가지고 폭을 통제하는 것은 아니다.

통상적인 수준이 95%로 못 박혀있는 경우가 많다. 신뢰구간의 폭을 통제할 수 있는 제일 강력한 수단은 물론 적절한 표본의 크기이다. 이는 신뢰구간의 한쪽 폭의 크기를 정해놓고 이를 달성하기 위한 표본의 크기를 정하는 문제가 된다.

10.8.1 평균 추정을 위한 표본크기

평균의 신뢰구간은 식 (10.20)과 같음은 이미 알고 있다.

$$\overline{X} \pm \text{ t-multiple } \times s/\sqrt{n} \tag{10.20}$$

신뢰구간의 한쪽 폭을 B로 명시한다면 이러한 크기의 신뢰구간을 달성하기 위한 표본의 크기를 구하기 위해서는 식 (10.21)을 풀어야 한다.

$$\text{t-multiple } \times s/\sqrt{n} = B \tag{10.21}$$

이를 n에 대해 풀면 식 (10.22)가 나온다. 이것이 표본의 크기를 구하는 식이다.

$$n = \left(\frac{\mathrm{t-mutiple} \times \mathrm{s}}{B} \right)^2 \tag{10.22}$$

그러나 표본의 크기가 주어져 있지 않은 상태에서 표본의 표준편차는 알 방법이 없다. 따라서 σ 의 적합한 추정값(σ_{est})을 대입하여야 한다. 그리고 t-multiple 역시 자유도를 모르는 상태에서는 z-multiple로 대체를 하여야 한다. 이는 n이 작지만 않다면 z-multiple은 t-multiple과 비슷한 값을 제공하기 때문이다. 따라서 식 (10.23)이 우리가 원하는 표본의 공식이다.

$$n = \left(\frac{\mathrm{z-multiple} \times \sigma_{\mathrm{est}}}{B} \right)^2 \tag{10.23}$$

나온 수를 제일 가까운 정수로 반올림하여 n을 결정한다.

10.8.2 기타 모수 추정을 위한 표본의 크기

모집단의 평균이 아닌 다른 모수, 즉 두 평균의 차이, 비율, 두 비율의 차이라 하더라도 절차는 다 같다. 신뢰구간의 폭을 B로 하고 이를 달성하기 위한 표본의 크기를 정하는 문제가 된다. 비율인 경우 표본의 크기는 식 (10.24)과 같이 나온다.

$$n = \left(\frac{\mathrm{z} - \mu\mathrm{ltiple}}{B}\right)^2 p_{est}(1 - p_{est}) \tag{10.24}$$

p_{est} 는 사용자가 대입하여야 하는데 이에 대한 추정값으로 보수적인 p 입장에서는 참의 p값에 상관없이 0.5를 대입하여 구하기도 한다. 그때 구한 표본의 수가 제일 크게 나오기 때문이다.

평균의 차이에 대한 표본의 크기는 식 (10.25)와 같다.

$$n = 2\left(\frac{\mathrm{z} - \mathrm{multiple} \times \sigma_{\mathrm{est}}}{B}\right)^2 \tag{10.25}$$

역시 σ_{est}는 사용자가 대입하여야 하는 σ에 대한 추정값으로 두 집단의 표준편차의 크기는 같다고 가정한다. 비율의 차이에 대한 표본의 크기는 식 (10.26)이다.

$$n = \left(\frac{\mathrm{z} - \mathrm{multiple}}{B}\right)^2 [p_{1est}(1 - p_{1est}) + p_{2est}(1 - p_{2est})] \tag{10.26}$$

p_{1est}, p_{2est} 는 사용자가 대입하여야 하는 p_1, p_2 에 대한 추정값으로 보수적인 입장에서는 참의 p값들에 상관없이 0.5를 대입하여 구하기도 한다.

이 중 두 비율의 차이에 대한 표본의 크기를 구하는 예제를 살펴보도록 하자. 동일한 물건을 생산하는 두 개의 공장이 있다고 하자. 공장별로 규격에 맞지 않는 비율이 차이가 나는지 알고자 한다. 그리고 두 공장에서 생산되는 물건 중 규격에 맞지 않는 물건의 비율은 3~5%라고 의심하고 있다. 99%의 신뢰구간의 한쪽 폭이 0.005(0.5%) 정도 가지기 위해 각 공장에서 표본을 얻는다면 얼마나 되는지 구해보자.

$$n = \left(\frac{2.576}{0.005}\right)^2 [0.05(0.95) + 0.05(0.95)] \approx 25,213$$

여기서 0.05를 대입한 것은 3%-5%의 값 중에서 제일 큰 표본의 크기를 얻기 때문이다. 그러나 25,213은 너무나 많은 숫자이다. 따라서 관리자가 가지고 있는 목적을 낮추어야 한다. 99%의 신뢰수준을 95%로 낮추든지 아니면 B=0.005의 크기를 0.025로 조정하여야 한다. 둘 다 조정하는 경우라 하더라도 n=584로 나온다. 두 비율의 차이에 대한 폭이 좁은 신뢰구간을

구하기 위해서는 여전히 크기가 큰 표본이 요구된다.

학습요약

표본자료로부터 모집단의 성질, 즉 모수를 추정하고자 할 때 제일 흔하게 보고하는 형태는 점추정 및 그에 수반하는 신뢰구간이다. 이러한 신뢰구간은 모집단의 성질이 어디에 위치하고 있는지 알려주는 아주 빠른 방법일뿐 아니라, 점추정이 가지고 있는 불확실성을 정량화하는 개념이기도 하다. 당연히 우리는 작은 폭의 신뢰구간을 원할 것이다. 신뢰구간의 폭은 자료의 변동성, 신뢰수준(통상적으로 95%), 그리고 표본의 크기에 의해 좌우됨을 알아보았다. 신뢰구간의 폭이 충분히 작게 미리 표본의 크기가 정해질 수 있다는 것을 이제 독자들은 알 것이다. 마지막으로 신뢰구간은 제 9장에서 본 내용과 여기서 소개한 t-분포에 의해 기계적으로 설정됨을 알 수 있었는데 여기 부여되는 신뢰수준의 의미를 시뮬레이션으로 알아보았다. 여기서 소개한 많은 형태의 신뢰구간은 서로 연관이 되어 있으며 많은 통계 소프트웨어에서 제공하기 때문에 독자들은 통계적인 개념에 집중하여 의사결정을 하면 될 것이다.

10장 연습문제

10.1 6면인 주사위를 여러 개를 준비하기 바란다.

(1) n=9, 16, 36개의 주사위를 동시에 던져 나오는 눈의 표본 평균 및 표준오차를 구하고 모집단의 평균에 대한 95% 신뢰구간을 구하라. 그리고 이런 행위를 100번 하였을 때 100개의 신뢰구간 중 모집단의 평균, 여기서는 3.5 × n을 포함하고 있는 개수가 몇 개인지 확인하여 보아라.

(2) n=9, 16, 36개의 주사위를 동시에 던져 나오는 눈의 표본 평균 및 표준오차를 구하고 이를 이용하여 주사위 전체 합에 대한 95% 신뢰구간을 구하라. 그리고 이런 행위를 100번 하였을 때 100개의 신뢰구간 중 모집단의 합, 여기서는 n(n+1)/2를 포함하고 있는 개수가 몇 개인지 확인하여 보아라.

(3) n=9, 16, 36개의 주사위를 동시에 던져 나오는 개개의 눈에 대해 4보다 같거나 큰 숫자가 나오면 1로 기록하고 그렇지 않으면 0으로 기록한다. 그런 다음 이런 n개의 수를 합을 한 다음 n으로 나눈다. 그렇다면 표본비율인 $\hat{p}$ 를 얻을 것이다. 이를 이용하여 모집단의 비율에 대한 95% 신뢰구간을 구하여 보아라. 이런 행위를 100번하여 100개의 신뢰구간 중 모집단의 비율, 여기서는 0.5를 포함하고 있는 개수가 몇 개인지 확인하여 보아라.

이상과 같은 실험은 물리적인 주사위를 던지지 않아도 엑셀로 가능하다. 엑셀 명령문 =int(1+ rand()*(7-1))을 사용하여 실험을 시행하면 된다.

이와 같은 훈련을 통해 평균, 합, 비율 등에 대한 신뢰구간의 의미를 파악하기 바란다.

10.2 6면의 주사위 9개와 16개를 준비하기 바란다. 영희는 9개의 주사위를 동시에 던지고 철수는 주사위 16개를 동시에 던진다.

(1) 영희와 철수는 각자 주사위를 동시에 던져 나오는 눈의 표본 평균 및 표준오차를 구한다. 이를 이용하여 두 모집단의 차이에 대한 95% 신뢰구간을 구하라. 그리고 이런 행위를 100번하였을 때 100개의 신뢰구간 중 모집단의 평균의 차이, 여기서는 0을 포함하고 있는 개수가 몇 개인지 확인하여 보아라.

(2) 영희와 철수는 각자 주사위를 동시에 던져 나오는 개개의 눈에 대해 4 보다 같거나 큰 숫자가 나오면 1로 기록하고 그렇지 않으면 0으로 기록한다. 그런 다음 이런 n개의 수를 합을 한 다음 n으로 나눈다. 그렇다면 두개의 표본비율인 $\widehat{p_1}, \widehat{p_2}$ 를 얻을 것이다. 이를 이용하여 모집단의 비율의 차이에 대한 95% 신뢰구간을 구하여 보아라. 이런 행위를 100번하여 100개의 신뢰구간 중 모집단의 비율, 여기서는 0을 포함하고 있는 개수가 몇 개인지 확인하여 보아라.

10.3 5장의 예제 5.1을 다시 언급하여 보자.

	A	B	C	D	E	F	G	H	I
1	서울 강동지역에 대한 환경시설유치에 대한 주민의견								
2									
3	age	sex	region	Children	salary('10000)	opinion		age_cat	sex_mod
4	61	F	강동	2	6,200	1		elderly	2
5	37	M	강동	2	5,200	5		middle-aged	1
6	32	F	강동	3	8,140	1		young	2
7	65	F	강동	2	4,960	1		elderly	2
8	40	M	강동	3	4,770	4		middle-aged	1
9	32	F	강동	1	5,990	4		young	2
10	38	F	강동	2	3,900	2		middle-aged	2
11	48	M	강동	1	6,150	2		middle-aged	1
12	40	M	강동	1	4,450	3		middle-aged	1
13	44	M	강동	2	4,520	3		middle-aged	1
14	57	F	강동	2	3,670	4		middle-aged	2
15	21	F	강동	2	5,430	2		young	2
16	49	M	강동	1	6,210	4		middle-aged	1
17	34	M	강동	0	7,800	3		young	1
18	38	M	강동	1	4,330	1		middle-aged	1
19	35	M	송파	1	6,540	5		middle-aged	1
20	35	M	송파	0	6,320	3		middle-aged	1
21	33	F	송파	3	4,630	5		young	2
22	45	M	송파	1	4,590	5		middle-aged	1
23	57	M	송파	1	4,810	4		middle-aged	1
24	38	F	송파	0	5,810	3		middle-aged	2
25	37	F	송파	2	5,600	1		middle-aged	2
26	42	F	송파	2	5,340	1		middle-aged	2
27	49	M	송파	0	4,320	5		middle-aged	1
28	52	M	송파	1	4,410	3		middle-aged	1
29	27	M	송파	3	4,540	2		young	1
30	40	M	송파	0	5,900	4		middle-aged	1
31	63	M	송파	2	5,390	1		elderly	1
32	48	F	송파	2	3,100	4		middle-aged	2
33	40	M	송파	0	3,770	1		middle-aged	1

(1) 이 자료를 강동 지역과 송파 지역으로 나누어 의견(opinion)의 차이에 대한 신뢰구간을 95% 신뢰수준에서 구하고 결과를 설명해 보자.

(2) 또한 이 자료를 여자와 남자로 나누어 의견의 차이에 대한 신뢰구간을 95% 신뢰수준에서 구하고 결과를 설명해 보라.

(3) 자녀의 수가 1이하인 경우와 그렇지 않은 경우로 나누어 의견의 차이에 대한 신뢰구간을 95% 신뢰수준에서 구하고 결과를 설명해 보라.

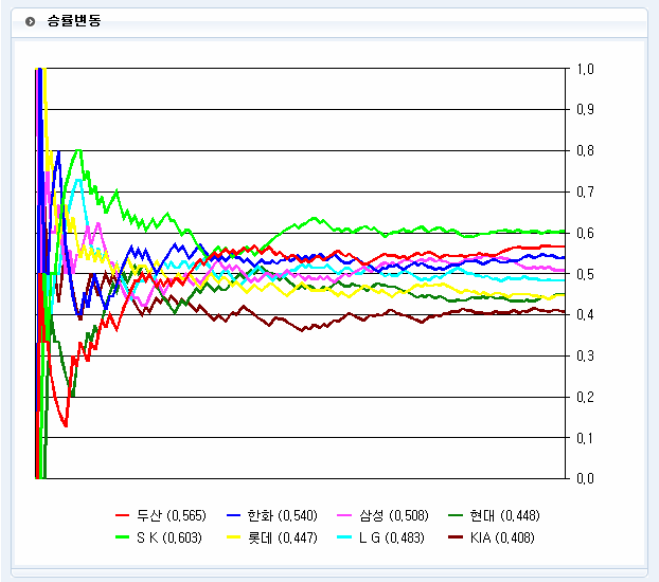
(4) 연령별 간에 의견에 대한 차이가 있는지 신뢰구간을 95% 신뢰수준에서 구하고 결과를 설명해 보라.

(5) 강동 지역과 송파 지역으로 나누어 4 이상의 의견(opinion)을 가지고 있는 비율의 차이가 있는지 신뢰구간을 95% 신뢰수준에서 구하고 결과를 설명해 보라.

쉬어가기

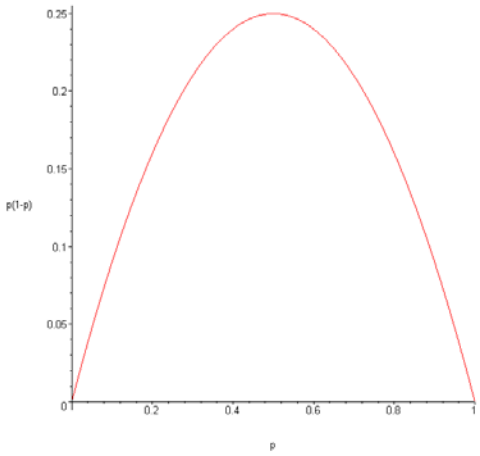
1. 표준오차 예제

다음 그림은 2007년 프로야구시즌이 끝난 후 8개 구단의 승률을 나타내고 있는 꺾은선 그래프이다. 그래프의 x축은 날짜, y축은 승률을 나타낸다. 각 구단의 승률이 시즌 초기에는 불안정하게 변동이 심하다가 시즌 후반기에 가면 안정적으로 특정값에 수렴하는 것을 볼 수 있다. 왜 이런 현상이 벌어질까?



(제공처: 스포츠조선)

각 팀에서 치룬 경기 수 중 이긴 경기수와 진 경기수를 합친 경기수를 n이라 하고 이긴 경기수를 w라 하면 승률은 $\hat{p}=\frac{w}{n}$이 된다. 예로 SK의 경우 이긴 경기수가 73, 진 경기수가 48이므로 승률은 $\frac{73}{73+48}\approx 0.603$이 된다. 이러한 승률의 표준오차(승률의 변동의 크기를 나타내는 값)는 $\sqrt{\frac{p(1-p)}{n}}$가 된다. 여기서 p는 모승률이다. $0\le p\le 1$이므로 $p(1-p)$의 최대값은 다음 그림에서 본 것과 같이 $p=\frac{1}{2}$에서 1/4이 되므로 승률의 표준오차의 최대값은 $\frac{1}{2\sqrt{n}}$이 되어 n이 점점 커지면 승률의 표준오차의 최대값은 점점 작아지게 된다. 그래서 각 구단의 승률이 시즌 초기에는 불안정하게 변동이 심하다가 시즌 후반기에 가면 안정적으로 특정값에 수렴하게 되는 것이다. 우리는 p를 모르므로 승률의 표준오차로서 (10.7)식인 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$을 주로 사용한다.



각 팀의 승/패/무와 승률 및 승률의 표준오차는 다음과 같다.

순위	팀	승	패	무	승률	승률의 표준오차
1	SK	73	48	5	0.603	0.044
2	두산	70	54	2	0.565	0.045
3	한화	67	57	2	0.540	0.045
4	삼성	62	60	4	0.508	0.045
5	LG	58	62	6	0.483	0.046
6	현대	56	69	1	0.448	0.044
7	롯데	55	68	3	0.447	0.045
8	KIA	51	74	1	0.408	0.044

각 팀 승률의 표준오차 값이 0.044에서 0.046까지 큰 차이가 없이 비슷하고 값이 작다. 그러므로 각 팀이 126 게임을 치룬 후의 승률을 우리는 참 승률에 대한 근사값으로 여기고 각 팀의 2007년 실력으로 순위를 정하는 것이다.

야구에서는 그 팀의 기록과 전력을 바탕으로 기대되는 다른 형태의 승률이 존재한다. 이러한 승률 중 하나인 '피타고라스 승률'은 야구통계학자 빌 제임스가 1980년대 초에 고안한 승률로서 '총득점의 제곱/(총득점의 제곱+총실점의 제곱)'의 값으로 그 팀의 기록과 전력을 바탕으로 기대되는 승률을 파악할 수 있다. 실제로 피타고라스 승률은 시즌 종료 시점에는 실제 팀 승률과 거의 근접하는지 조사하여 보자.

다음 표는 2007년 시즌이 끝난 후 각 팀의 승/패/무, 승률, 총득점, 총실점 및 피타고라스 승률은 다음과 같다. 승률에서의 순위와 피타고라스 승률에서의 순위가 다른 팀이 한 팀으로 롯데임을 알 수 있다. 롯데는 총득점과 총실점의 입장에서는 5위인 LG나 6위인 현대보다도 성적이 좋고 피타고라스 승률로서는 4위인 삼성에 필적한다. 그럼에도 불구하고 승률에서의 순위는 7위밖에 안 된다. 경제적인 야구를 못했다는 결론을 얻을 수 있다.

순위	팀	승	패	무	승률	총득점	총실점	피타고라스 승률
1	SK	73	48	5	0.603	603	465	0.627
2	두산	70	54	2	0.565	578	480	0.592
3	한화	67	57	2	0.540	534	481	0.552
4	삼성	62	60	4	0.508	497	509	0.488
5	LG	58	62	6	0.483	532	600	0.440
6	현대	56	69	1	0.448	530	615	0.426
7	롯데	55	68	3	0.447	533	554	0.481
8	KIA	51	74	1	0.408	499	602	0.407

표준오차는 추정량의 표준편차를 의미한다. 모집단의 모수(평균, 비율) 등을 추정하는데 있어 추정량(표본평균, 표본비율 등)이 얼마나 정확한가의 척도이다.

2. 상대표준오차

① 표준편차 : 분포의 중심위치를 나타내는 측도로서 확률변수의 평균에서 흩어진 정도를 나타내는 양
② 표본오차 : 전수가 아닌 표본으로 전체를 추정함으로써 발생되는 오차로 개념적인 오차
③ 표준오차 : 표본오차의 측도로서 동일 모집단에 대해 이론적으로 가능한 수많은 경우의 표본을 반복 추출했을 때 표집 평균값을 중심으로 하는 표준편차

통계청이 밝힌 2007년 1월 전체 실업률은 3.5%로 약 85만명에 이르며, 이에 대한 표준오차(SE)는 2.9만명으로 나타났다. (경제활동인구조사 中)

95% 신뢰도 하에서는 85 만명 ± (1.96 × 2.9 만명 = 5.68 만명)으로 79.1 만명에서 90.8 만명 내에 존재할 것으로 추정된다. 각각의 신뢰도에서는 다음과 같이 계산 가능하다.

· 90% 신뢰도에서 모집단 존재범위 : p ± 1.645 × SE
· 95% 신뢰도에서 모집단 존재범위 : p ± 1.96 × SE
· 99% 신뢰도에서 모집단 존재범위 : p ± 2.58 × SE

왜 신뢰수준은 보통 95%를 사용하는가? 만일 99%를 사용한다면 오차한계가 커짐으로 참값의 정도를 흐려 가치가 적어진다. 예를 들어 정확히 예측하고자 키가 1미터 이상 2미터 이하라고 한다면 정보라고 하기에는 곤란하다. 95%를 사용하는 것은 적절한 신뢰수준과 정보의 가치에 대한 적절한 타협점이라고 보면 된다. 다음과 같은 예를 살펴보자.

2006년 4/4분기 2인 이상 가구의 평균소득은 3,168천원이며, 평균 가계지출은 2,580천원으로 나타났다. 이 때 평균소득의 표준오차는 40천원이며, 평균 가계지출의 표준오차는 36천원으로 나타났다(가계조사 자료 中). 이 때 평균소득의 신뢰구간은

3,168천원 ± 80 (≒1.96*40) 천원

이고, 평균 가계지출의 신뢰구간은

2,580천원 ± 72 (≒1.96*36) 천원

이다. 표본오차, 엄격히 표준오차는 천원, 만원, %, 천명, 만명 등 여러 가지 단위를 가짐으로 해석하는데 어려움이 있다. 예로서 30년 전 자장면은 500원, 표준오차는 50원이었고, 현재의 자장면은 4,000원, 표준오차는 400원이라고 할 때 똑같은 오차를 가졌는데도 표준오차의 차가 커서 해석상 어려움이 있다. 이를 위해 상대표준오차(coefficient of variation) 개념이 도입되었다. 상대표준오차(CV)는 표준오차를 표준화하여 비교 가능하게 하는 지표로 표준오차를 추정치(평균치)로 나누어 계산하며 백분율로 표시한다(CV= (SE/추정치) × 100).

30년 전 자장면은 500원, 표준오차는 50원이었으며, 현재의 자장면은 4,000원, 표준오차는 400원이라고 가정해 보자.

30년 전 자장면 CV = 50/500*100=10%
현재 자장면 CV = 400/4000*100=10%

상대표준오차는 역(易)으로 추정값을 곱하면 표준오차를 얻을 수 있다는 장점이 있다. 따라서 상대표준오차 하나만 제시하더라도 정보를 잃지 않는다. 예를 들어, 실업자의 상대표준오차는 3.4%이고 추정값은 851천명인 경우에 851천명 * 0.034 = 29천명이 산출되어 실업자 851천명은 어느 정도의 오차를 가질 수 있는지 판단할 수 있는 기준이 될 수 있다. 우리나라에서는 일부만 제시하고 있지만, 일본의 경우에는 상대표준오차를 표준오차율로 표시하여 사용하고 있다. 그렇다면 상대표준오차(CV)는 어느 정도가 좋을까?

2007년 1월 실업자가 851천명, 상대표준오차가 3.4%이므로 표준오차는 29천명으로 나타났다고 하자. 만약 상대표준오차가 20.0%라면 표준오차는 851천명*0.2=170천명이 되어 95% 신뢰수준을 고려하면 851천명 ±334천명이 된다. 따라서 실업자는 517천명에서 1,185천명으로 구간이 넓게 되어 자료에 대한 신뢰도가 낮아지게 됨을 알 수 있다.

< 상대표준오차의 기준 : 캐나다 서베이 기준 >

- 0.00% ~ 4.99% : 매우 우수(Excellent)
- 5.00% ~ 9.99% : 우수(Very Good)
- 10.00% ~ 14.99% : 좋음(Good)
- 15.00% ~ 24.99% : 허용 가능(Acceptable)
- 25.00% ~ 34.99% : 주의사항과 함께 사용가능(Use with caution)
- 35.00% : 공표시 신뢰불가(Too unreliable to publish)

제 11 장

우연인가? 증거가 말한다.

차 례

학습목표

신뢰구간이 통계적 추론이라는 동전의 앞면이라면 가설검정은 동전의 뒷면에 해당되는 내용이 된다. 자료에 근거해 모집단의 성질에 대한 추론을 한다는 의미는 이 중 한 방법을 선택하는 것이다.

제 10장에서는 점추정을 구하고 점추정을 중심으로 신뢰구간을 구하였다. 이러한 방법은 자료 스스로 참의 모수(true parameter)가 어디에 있을지를 알려 주는 방법이라고 볼 수 있다. 그러나 대부분 실험자 혹은 조사자는 증명하고자 하는 이론 혹은 가설(hypothesis)을 가지고 있다. 예를 들면 “새로운 상품 디자인은 좀 더 많은 매출을 유발할 것이다.” 혹은 “새로운 의약품은 기존 약품에 비해 월등히 치료효과가 높다.”와 같은 믿음 등이다. 이런 경우는 조사자는 자료를 수집하여 가설을 뒤받침할 충분한 증거가 있는지 여부를 파악하고자 할 것이다.

신뢰구간과 가설검정은 거의 비슷한 결과를 보고한다. 다만 관점이 다를 뿐이다. 따라서 신뢰구간과 가설검정 중 어느 면을 더 선호하느냐 하는 것은 논란의 여지가 있다. 그러나 가설검정은 대부분의 통계패키지에 내장이 되어 있어 자동적으로 많은 결과가 출력된다. 따라서 논란의 여지 이전에 이에 대한 해석을 하여야 하는 것도 독자들의 몫이다.

11.1 가설검정을 위한 용어

어느 가설이 맞는가?

조사자가 증명하고자 하는 가설을 대립가설(alternative hypothesis), 혹은 연구가설(research hypothesis)이라 한다. 반대편에 있는 가설을 귀무가설(null hypothesis)이라 한다. 귀무가설은 현재 가지고 있는 믿음이다. 즉, 연구자가 반박(disprove)하고자 하는 가설이 된다. 예를 들면 "새로운 상품 디자인은 현재 디자인과 별로 다를 게 없다." 혹은 "새로운 약품은 시중에 나와 있는 약품과 치료효과 면에서 별로 다른 바 없다." 등이다. 증명의 부담은 거의 다 대립가설 쪽에 있다. 연구자는 대립가설을 지지할 충분한 증거를 제공하여야 하기 때문이다. 여기서 충분한 증거를 제공하여야 하는 것은 웬만해서는 귀무가설을 기각하고 대립가설을 채택하지 않기 때문이다. 예를 들어서 새로운 상품 디자인을 선보였는데 만약 기존 제품과 매출이 엇비슷하게 나온다면 새로운 상품 디자인을 시장에 선보이는 과정에서 나오는 모든 비용을 감수하여야 하기 때문이다.

여기서 충분한 증거는 후에 이야기하겠지만 가설검정의 결과가 통계적으로 유의할 경우를 의미한다.

예제 11.1 **개발된 신제품의 효능은 구제품에 비해 좋은가?**

Miracle Gro® "Pour & Feed" TerraCycle® 20oz Plant Food

신상품은 기존제품에 비해 제초 능력이 더 좋은가?

새로운 방법에 의해 만들어진 신상품(예를 들면 새로운 배합의 비료 등)이 기존 방법에 의해 만들어진 제품보다 더 성능이 있는지를 실험하기 위하여 한 달 동안 기존 제품을 주문하는 모든 고객에게 기존 제품과 더불어 새로운 방법의 제품을 하나 더 무료로 공급하게 하였다. 그런 다음 -10부터 10까지의 척도로 두 제품의 성능을 비교하는 설문조사를 실시하였다. -10이면 극단으로 이전의 제품이 성능이 좋지 않고 +10이면 극단으로 새로운 제품이 성능이 좋고 0이면 두 제품은 성능상 차이가 없다는 척도이다. 조사된 평균이 1.8이고 95% 신뢰구간이 0.3부터 3.3이라면 아마 관리자는 새로운 방법의 제품을 만들어 판매하고자 할 것이다. ■

● 귀무가설과 대립가설

이 두 제품의 성능 문제를 전체 제품을 사용한 소비자 집단의 평균을 가지고 판단을 한다고 하여 보자. $\mu \leq 0$ 라고 판단된다면 귀무가설이 맞고 $\mu > 0$ 라고 판단된다면 대립가설이 맞을 것이다. 귀무가설은 관례적으로 H_0 로 표기하고 대립가설은 H_a 또는 H_1이라고 표기한다. 이 경우에는

$$H_0 : \mu \leq 0 \ , \ H_a : \mu > 0$$

이 된다.

대립가설과 귀무가설은 모든 가능한 경우를 두 개의 중복되지 않는 집합으로 나누어 표현되며 반드시 하나가 참이어야 한다. 관리자는 표본자료를 이용하여 어느 가설이 참인지 여부를 가리고자 시도할 것이다. 가설검정은 일종의 의사결정이다. 표본의 결과에 따라 귀무가설을 채택을 하여야 할지 기각을 하여야 할지 결정하여야 한다. 귀무가설을 기각한다는 의미는 기존의 제품 대신 신상품으로 선회할 가능성을 염두에 두는 것이다.

• 오류의 종류

귀무가설을 채택을 하든지 기각을 하든지 무관하게 그 의사결정은 잘못된 의사결정이 될 수 있다. 왜냐하면 귀무가설이 맞는데 귀무가설을 기각한다든지 대립가설이 맞는데 귀무가설을 채택한다든지 하기 때문이다. 이런 종류의 오류들 중 전자를 제 1종 오류(type 1 error), 그리고 후자를 제 2종 오류(type 2 error)라 한다. 이는 [표 11.1]과 같다.

미지의 실제현상 / 의사결정	귀무가설 맞음	대립가설 맞음
귀무가설 기각	제 1종 오류	오류 없음
귀무가설 채택	오류 없음	제 2종 오류

[표 11.1] 제 1종 및 제 2종 오류

옛날 방식의 제품이 더 성능이 좋았음에도 불구하고 표본 자료에 의거 새로운 제품을 선택하였다면 관리자는 제 1종 오류를 범하는 경우이고 그렇지 않고 새로운 제품의 성능이 더 좋았음에도 불구하고 예전 방식의 제품을 고집하였다면 제 2종 오류를 범하는 경우가 된다.

따라서 보수적인 입장에서는 귀무가설을 기각할 충분한 증거가 나타나지 않는 한 귀무가설을 채택할 것이다. 그러나 귀무가설을 채택하는 순간에는 제 2종 오류를 범할 가능성이 있다. 관리자는 이 두 가지 종류의 오류에서 딜레마를 느낀다. 제 1종의 오류를 피하기 위해서는 관리자는 좀 더 강한 증거를 요구할 것이다. 만약 평균이 +1.5이고 95% 신뢰구간이 -0.3에서 3.3까지 나왔다면 이러한 증거는 새로운 제품을 채택할 만한 충분한 증거가 나왔다고 판단하지 않을 가능성이 높다. 왜냐하면 신뢰구간이 0을 포함하기 때문이다.

• 유의수준과 기각역

귀무가설을 기각하고 대립가설을 채택할 충분한 증거가 자료에 있는 지를 확인하는 작업이 가설검정의 형태가 되는데 충분하다는 의미가 무엇인지를 살펴보도록 하자.

여기에는 두 가지 방법이 있는데

첫 번째 방법은 분석자가 참을 수 있는 제 1종 오류의 크기를 명시하는 것이다. 이러한 제 1종 오류를 α라 표기하면 통상적으로 0.05, 0.01 혹은 0.10의 값이 많이 쓰인다. 이런 α값을 검정 유의수준(test significance level)이라 한다. 그러면 주어진 α에 대해 통계적으로 기각역을 결정한다. 표본증거가 이 기각역 안에 들어가면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다. 기각역은 제 1종 오류의 확률이 많아 봐야 α가 되게끔 정확하게 결정되어야 한다. 표본증거가 이 기각역 안에 들어가는 것을

"α 유의수준에서 통계적으로 유의하다."

라고 한다. 예를 들어 α=0.05 라면 “증거는 5% 유의 수준에서 통계적으로 유의하다.” 라고 말한다.

두 번째 방법은 α의 값을 명시하기보다 표본증거가 얼마나 유의하냐를 보고하는 것이다. 소위 p-값(p-value)이다. 예제 11.1에서 참의 평균이 $\mu = 0$이라고 하자. 물론 이 평균이 0인지 여부는 아무도 모른다. 그리고 표본의 평균이 +2.5가 나왔다고 하자. 이 시점에서 관리자의 옵션은 두 가지일 것이다. 하나는 관측된 표본 값은 매우 비정상적인 값이므로 귀무가설이 참이라고 결론을 내리는 것이고 다른 하나는 표본 값은 아주 정상적인 값이므로 귀무가설을 기각하고 새로운 제품을 선택하는 것이다. 표본의 p-값은 이러한 개념을 계량화시켜 준다.

• p-값

p-값은 "귀무가설이 맞다."는 가정 하에서 적어도 표본 값과 같이 극단적(extreme)인 표본을

얻을 확률을 의미한다. 여기서 극단이란 귀무가설을 기준으로 하는 것이다. 예를 들어 표본이 +3.5가 나왔다고 하자. 그러면 p-값은 귀무가설이 옳다는 가정 하에서 표본이 적어도 3.5와 같은 값을 얻을 가능성을 말하는 것이다.

신상품 예에서 p-값으로 0.03을 얻었다고 하자. 이는 100개 표본을 추출하였다면 두 제품의 성능이 동일하다는 가정 하에서는 3개 정도만 관측된 표본과 같이 대립가설을 옹호할 증거를 제공한다는 의미가 된다. 이런 결과를 바탕으로 귀무가설을 기각하고 대립가설을 채택할 것인가를 결정하여야 한다.

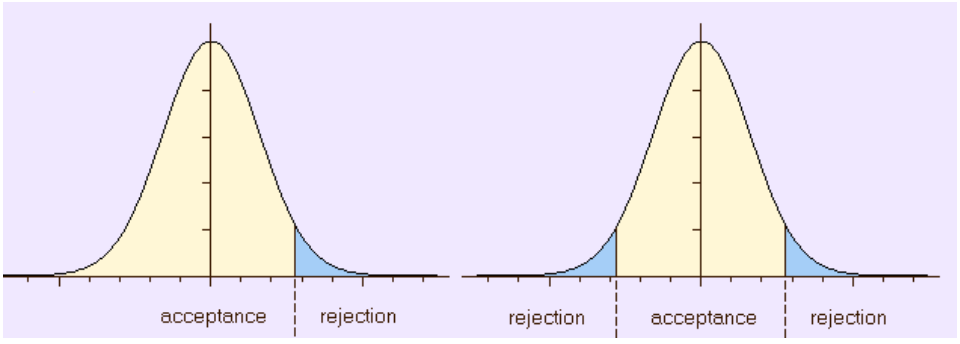
결과가 매우 비정상적인 값으로 판단한다면 귀무가설을 채택하고 그렇지 않다면 대립가설을 채택할 것이다. 이는 의사결정자의 몫이다. 그러나 일반적으로 p-값은 작으면 작을수록 대립가설을 선호할 충분한 증거가 있다는 의미가 된다.

그러나 p-값이 얼마나 적어야 적은 것인지에 대한 판단은 통계적인 문제가 아니다. p-값이 0.01보다 작으면 100개의 표본 중에서 1개 미만으로 대립가설이 틀렸음에도 불구하고 대립가설이 맞다고 할 증거를 제공하는 뜻이다.

p-값이 0.01 보다 적으면 대립가설을 선호할 아주 강력한 증거로 보고 0.01과 0.05 사이에 있으면 대립가설을 선호할 강력한 증거가 표본에 있다고 한다. 그리고 p-값이 0.05와 0.10 사이이면 중간 정도의 증거, 그리고 0.10보다 크면 약하거나 없다고 판정한다.

위에서 명기한 α와 p-값은 매우 밀접한 관계가 있다. p-값이 명시한 α보다 작으면 귀무가설을 기각하고 대립가설을 채택하고, 그렇지 않고 α보다 크면 당연히 대립가설을 기각하지 못한다. 그러나 α값은 사전에 명시를 하지만 p-값은 그럴 필요가 없다. 따라서 많은 소프트웨어의 결과물도 이런 p-값을 기준으로 만들어졌다.

- **단측가설과 양측가설**



양측가설에서는 기각역이 둘로 나누어진다.

대립가설의 형태는 증명을 하고자 하는 것에 따라 단측(one sided)과 양측(two sided)가설로 나누어진다. 예제 11.1에서는 대립가설이 단측으로 표시가 된다. 즉, "소비자의 판정이 평균적으로 0보다 크다"라는 사실을 증명하고자 하였기 때문이다. 따라서 귀무가설을 기각할 수 있는 표본은 표본평균의 값이 양인 표본만이다. 만약 관리자가 역으로 척도를 표시하였다면 귀

무가설을 기각할 수 있는 표본은 표본평균의 값이 음인 표본만이다.

양측검정은 귀무가설을 기각할 방향이 양방향인 경우를 의미한다. 만약 관리자가 두 방법 중 하나의 방법이 다른 방법에 비해 월등히 좋다고 여기고 두 개의 방법 중 하나를 포기하려고 한다고 했다면 귀무가설은

$$H_o : \mu = 0$$

이 되고, 대립가설은

$$H_a : \mu \neq 0$$

가 된다. 표본의 평균이 0보다 아주 작거나 아주 크다면 귀무가설을 기각하고 두 방법 중 하나를 포기할 것이다. 그러나 단측가설로 할 것인가 양측가설로 할 것인가를 판단하는 문제는 통계학의 몫이 아니고 증명하고자 하는 명제가 무엇인지에 달려있다.

- **가설검정과 신뢰구간의 관계**

신뢰구간과 가설검정은 동전의 앞면과 뒷면이다.

가설검정의 대립가설 형태가 양측가설로 표현될 때 가설검정과 신뢰구간은 관계가 있다. 즉, 5% 유의수준에서 귀무가설을 기각한다는 이야기는 95% 신뢰구간이 귀무가설에서 명시한 가설값을 포함하지 않는다는 의미와 동격이다.

$$H_0 : \mu = 0, \quad H_a : \mu \neq 0$$

인 가설검정에서 95% 신뢰구간이

$$[1.35,\ 3.42]$$

로 나왔다고 하자. 이 신뢰구간은 0을 포함하고 있지 않으므로 귀무가설은 5% 유의수준에서 기각이 된다. p-값은 0.05보다 작을 것이다. 만약 신뢰구간이

$$[-1.25,\ 2.31]$$

라면 0을 포함하고 있기 때문에 귀무가설은 기각이 되지 않을 것이다. 이 경우 p-값은 α 보다 크다.

- **실제적 유의와 통계적 유의**

어느 교육학자가 4,000명의 남학생과 4,000명의 여학생을 대상으로 특정 시험 결과에 대해 조사한 결과 남자가 521 점, 그리고 여자가 524 점을 맞았다고 발표를 했다고 보자. 그리고 p-값이 0.007이 나왔다. 과연 많은 사람들이 이러한 결과에 얼마나 관심을 기울일까? 표본의 크기가 커지면 얼마든지 성별로 평균의 차이가 있다는 사실은 얼마든지 통계적으로 유의하다고 밝힐 수 있다.

그러나 겨우 3점이 아닌가? 교육학자는 성별로 성적 (지적 능력)의 차이가 있다는 사실을 증명하고 싶어할 지 모르지만 많은 사람들은 별로 관심을 두지 않는다. 통계적으로 유의(statistically significant)할지 모르지만 실제적으로는 유의(practically significant)하지 않다. 3점이 아니라 30~40점이라면 모를까.

이 경우는 표본의 크기가 큰 경우지만 반대되는 경우도 보자. 예산부족으로 몇 안 되는 환자를 대상으로 신 약품의 치료율을 기존약품과 비교한 결과 p-값이 0.2가 나왔다고 보자. 이 경우는 통계적으로는 유의하지 못하지만 실제적으로는 매우 유의하다고 할 수 있다. 좀 더 많은 환자를 대상으로 실험을 하였다면 얼마든지 많은 수의 환자를 치유했을 실제적인 유의성을 확보할 수 있었기 때문이다. 이 경우는 작은 표본의 크기 때문에 실제적인 유의성을 확보하지 못하는 경우이다. 대부분의 통계학 책은 통계적인 유의성에 대해서만 언급하고 있다. 자료가 말하는(Data Saying) 통계학의 특성 중 하나이지만 반드시 현실과 결부해 생각해 보아야 한다.

11.2 모집단의 평균에 대한 가설검정

지금까지 일반적인 가설검정의 개념을 살펴보았다. 이제부터 모집단의 평균에 대한 가설검정을 시작으로 모수들에 대한 가설검정 절차를 차례로 알아보도록 한다.

표본이 작으면 t-분포가 필요하다.

신뢰구간의 제일 중요한 개념은 표본평균의 표본추출분포 이론이었다. 표본평균에서 참의 평균을 빼고 표준오차로 나누어 주면 그 결과는 자유도 n-1인 t-분포를 따른다는 사실이다.

가설검정 하에서는 참의 평균을 귀무가설 하에서의 가설값(hypothesized value)으로 사용한다. 귀무가설과 대립가설의 경계에 있는 값이다. 이러한 가설값은 귀무가설에 기초를 하기 때문에 μ_0라 표기한다.

먼저 가설검정을 실시하기 위해서는 다음과 같은 검정통계량(test-statistic)을 계산하여야 한다. 검정통계량이란 검정에 쓰이는 통계량을 지칭한다.

$$\mathrm{t-value} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \tag{11.1}$$

만약 귀무가설이 옳다면, 즉 $\mu = \mu_0$라면 이러한 검정통계량은 자유도가 n-1인 t-분포를 따른다. 그렇다면 p-값은 양측대립가설인 경우는 t-분포에서 양쪽으로 검정통계량의 값을 넘어설 확률이고 단측가설인 경우는 한쪽으로 검정통계량의 값을 넘어설 확률이다. 예를 들어 가설검정 절차를 살펴보도록 하자.

예제 11.1(계속) 신상품은 개발할 가치가 있는가?

신상품 예제에서 표본 평균 $\overline{X}$ =2.10이 나오고 표본표준편차 $s=$ 4.717이 나왔다고 하자. 표본의 크기는 $n=$ 40이다.

그러면 검정 통계량의 값은 다음과 같이 나온다.

$$\mathrm{t-value} = \frac{2.10 - 0}{4.717/\sqrt{40}} = 2.816$$

만약 대립가설이 형태가 $H_a : \mu > 0$ 이고 유의수준이 0.05라면 기각역은 t(0.05, 자유도=39)에 의해 결정되는데 엑셀에서

=tinv(2*0.05, 39)

로 구하면 1.6849가 나온다. 검정통계량 값이 이보다 크므로 귀무가설은 기각된다. 이런 방법이 첫 번째 방법이다.

두 번째 방법인 p-값을 구하여 보자. p- 값은 자유도 39인 t-분포에서 2.816보다 큰 확률이 된다. 엑셀에서

=tdist(2.816, 39, 1)

입력한 결과 0.004가 나온다. 0.004는 1,000번 중에서 겨우 4번 아닌가? 이러한 표본의 결과는 귀무가설이 옳다면 나올 가능성이 매우 희박하다는 의미로 받아들여진다. 따라서 귀무가설

은 기각되고 대립가설, 즉 새로운 방법의 제품이 더 성능이 좋다는 가설을 채택될 것이다. 쉽지 않은가? ■ <성능비교.xls>

	A	B	C	D	E	F	G
1							
2	고객	평점					
3	1	-7			표본의 크기	40	
4	2	7			표본평균	2.100	
5	3	-2			표본 표준편차	4.717	
6	4	4					
7	5	7		평균이	0보다 작은지에 대한 단측가설 검정		
8	6	6			귀무가설 값	0.000	
9	7	0			표본평균	2.100	
10	8	2			평균의 표준오차	0.746	
11	9	8			자유도	39	
12	10	2			t-검정 통계량	2.816	
13	11	3			p-값	0.004	
14	12	-4					
15	13	8					
16	14	-5		F6.	=COUNT(Rating)		
17	15	7		F7.	=AVERAGE(Rating)		
18	16	-5		F8.	=STDEV(Rating)		
19	17	-1		F12.	=AVERAGE(Rating)		
20	18	7		F13.	=STDEV(Rating)/SQRT(COUNT(Rating))		
21	19	3		F14.	=COUNT(Rating)-1		
22	20	4		F15.	=(F12-F11)/F13		
23	21	2		F16.	=IF(F15>0,TDIST(F15,F14,1),1-TDIST(-F15,F14,1))		
24	22	0					
25	23	2					
26	24	9		Range : Rating : B4.B43			
27	25	-5					
28	26	2					
42	40	-6					

[표 11.2] 신제품의 성능비교 자료 가설검정

만약 α값이 p-값보다 작다면 귀무가설은 기각된다. 이 경우(p=0.004)에는 통상적인 α =0.01, 0.05, 0.1에서 모두 귀무가설이 기각된다.

만약 대립가설의 형태가 양측가설 $H_a : \mu \neq 0$ 라면 기각역은 t(0.025, 자유도=39)에 의해 결정되는데 엑셀에서

=tinv(0.05, 39)

에 의해 2.022가 나온다. 검정통계량의 값이 기각역보다 크므로 귀무가설은 기각된다.
p-값은 0.007588이다. 이는 엑셀 명령문

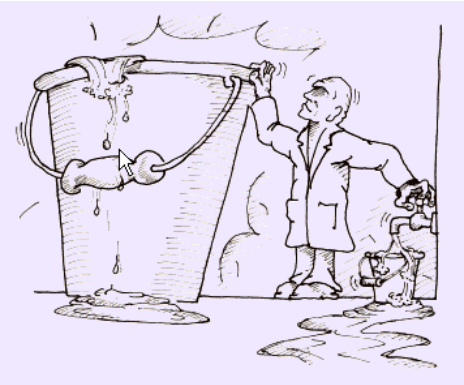
=tinv(abs(2.816), 9, 2)

으로 구한다. 따라서 관리자는 두 상품 중 하나는 포기할 것이다.
이상과 같은 내용을 기각역의 형태로 정리하면 [표 11.3]과 같다.

	one-tailed test		two-tailed test
hypothesis	$H_o : \mu_1 \geq \mu_o$ $H_1 : \mu_1 < \mu_o$	$H_o : \mu_1 \leq \mu_o$ $H_1 : \mu_1 > \mu_o$	$H_o : \mu_1 = \mu_o$ $H_1 : \mu_1 \neq \mu_o$
test statistic (t distribution)	$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$		
deg. of freedom	$n-1$		
rejection	reject H_o if $t < -t_\alpha$	reject H_o if $t > t_\alpha$	reject H_o if $\|t\| > t_{\alpha/2}$

[표 11.3] t를 이용한 평균에 대한 가설검정

그러나 표본의 크기가 충분히 크면 t-분포는 정규분포에 가까워지므로 굳이 t-분포에 의거해서 검정을 실시할 필요는 없다. z-분포를 이용하여 검정통계량을 만들어 [표 11.4]처럼 검정을 시행하면 된다. σ를 모르는 경우 s로 대신한다.



표본이 크면 굳이 t-분포를 쓸 필요 없다.

	one-tailed test		two-tailed test
hypothesis	$H_o : \mu_1 \geq \mu_o$ $H_1 : \mu_1 < \mu_o$	$H_o : \mu_1 \leq \mu_o$ $H_1 : \mu_1 > \mu_o$	$H_o : \mu_1 = \mu_o$ $H_1 : \mu_1 \neq \mu_o$
test statistic (normal distribution)	$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$		
deg. of freedom	—		
rejection	reject H_o if $z < -z_\alpha$	reject H_o if $z > z_\alpha$	reject H_o if $\|z\| > z_{\alpha/2}$

[표 11.4] z를 이용한 평균에 대한 가설검정

11.3 다른 모수에 대한 가설검정

다양한 모수에 대한 신뢰구간을 구했듯이 가설검정도 마찬가지이다. 신뢰구간에 적용된 같은 표본추출분포에 입각하여 절차가 진행이 된다. 어떤 경우라도 표본에서 검정통계량을 구하고 대립가설을 지지할 만한 충분한 증거가 충분한지 여부를 판단하기 위해 p-값을 구한다.

• 모집단 비율에 대한 가설검정

모집단의 비율 p에 대해 검정을 실시하기 위해서는 $\hat{p}$ 의 분포는 표본의 크기가 충분히 크다면 정규분포로 근사시킬 수 있다는 사실을 이용하면 된다. 구체적으로는 다음과 같이 표준화된 값

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$$

은 표준화된 정규확률변수(standard normal random variable), Z 와 근사하게 분포한다.
p_0 를 귀무가설과 대립가설 사이의 경계값이라 한다면 식 (11.2)가 검정통계량이 된다.

$$\text{z}-\text{value}=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \tag{11.2}$$

그러나 이 검정통계량은 표본의 크기가 큰 가정을 하였기 때문에 $np_0>5, n(1-p_0)>5$ 를 반드시 체크하여야 한다.

예제 11.2 **비율에 대한 가설 검정도 쉽다.**

은행 고객 관리를 맡고 있는 관리자는 매출의 30%을 점하고 있는 최상위 고객층의 서비스에 대한 불만율이 약 15% 된다는 사실에 매우 만족하지 못하고 있다. 따라서 새로운 절차에 의해 이 불만율을 반으로 줄이려고 한다. 고객의 만족도를 높이려고 30일 동안 400명의 손님을 대상으로 새로운 절차를 실시하였다. 표본의 결과는 400명 중 약 23명만이 불만족을 표시하였다면 관리자가 의도한 15%의 반인 7.5% 수준으로 낮추었는지 알아보고자 한다. [표 11.3]을 참조하기 바란다.

관리자의 의도대로 새로운 절차가 시행되었는지를 증명하고자 하는 문제이기 때문에 대립가설은 $H_a: p<0.075$ 가 된다. 따라서 $H_0: p\geq 0.075$ 이다. 표본에 의하면

$$\hat{p} = 23/400 = 0.0575$$

이다. 검정통계량인 z-value 는

$$\text{z-value} = \frac{0.0575 - 0.075}{\sqrt{0.075(1-0.075)/400}} = -1.329$$

로 나와

=normsdist(-1.329)

에 의해 z-value가 -1.329보다 왼쪽에 있을 확률인 p값은 0.092가 나온다.

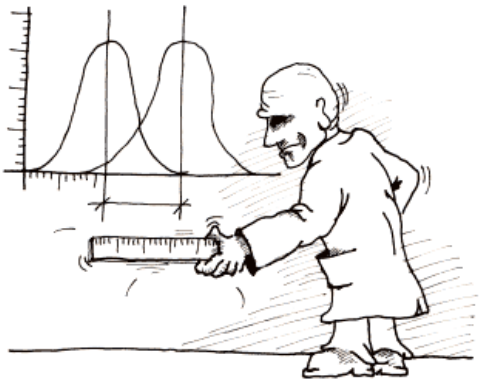
그리고 $np_0 = 400(0.075) = 30 > 5, n(1-p_0) = 400(0.925) > 5$ 로 나와 이 검정절차는 타당하다고 판단된다.

또한 신뢰구간이 [0.035, 0.080]으로 나와 가설값 0.075를 포함하므로 관리자는 소기의 목적을 달성하였다고 볼 수 있다. 신뢰구간은 가설검정과 독립적인 절차이다. 따라서 신뢰구간에 적용되는 SE를 구하기 위해서는 표본비율 값인 0.0575를 적용하여야 한다. 대부분의 경우에 표준오차의 계산은 거의 차이는 없으나, 계산과정은 다름을 유의하기 바란다. [표 11.5]를 참조하기 바란다. ■ <비율.xls>

	A	B
1		
2	새로운 절차에 의한 목표값	0.075
3		
4	30일 지난후에도 불만족인 고객의 수	23
5	표본의 크기	400
6	표본비 율	0.0575
7	표본비 율의 표준오차	0.01317
8		
9	z 검정 통계량	-1.329
10	양측 검정일때 p-값	0.092
11		
12	비율에 대한 신뢰구간	
13	신뢰수준	95%
14	표준도차	0.012
15	z-multiple	1.960
16	하한값:	0.035
17	상한값:	0.080
18		
19	B7. =B5/B6	
20	B8. =SQRT(B3*(1-B3)/B6)	
21	B10. =(B7-B3)/B8	
22	D11. =NORMSDIST(D10)	
23	B15. =SQRT(B7*(1-B7)/B6)	
24	B16. =NORMSINV(B14+(1-B14)/2)	
25	B17. =B7-B16*SQRT(B7*(1-B7)/B6)	
26	B18. =B7+B16*SQRT(B7*(1-B7)/B6)	

[표 11.5] 비율 가설검정

• **두 집단의 평균 차이에 대한 가설 검정**



두 평균을 비교하자.

두 모집단에서 추출된 표본이 독립이냐 아니면 짝진 표본이냐에 따라 검정통계량의 형태가 달라진다. 짝진 표본의 경우, $\bar{d}$가 n개의 짝에서 나온 차이에 대한 표본평균값이고, D_0가 귀무가설과 대립가설의 경계값, 그리고 s_D 가 차이의 표본표준편차라면 검정통계량은 식 (11.3)과 같다.

$$\text{t} - \text{value} = \frac{\bar{d} - D_o}{s_D/\sqrt{n}} \qquad (11.3)$$

만약 D_0가 차이의 참평균값이면 이 통계량은 자유도가 n-1인 t-분포를 따라간다. n이 충분한 큰 경우가 아니면 차이에 대한 모집단은 정규분포를 가정하여야 한다. 이를 정리하면 [표 11.6]과 같다.

	one-tailed test		two-tailed test
hypothesis	$H_o : \mu_D \geq D_o$ $H_1 : \mu_D < D_o$	$H_o : \mu_D \leq D_o$ $H_1 : \mu_D > D_o$	$H_o : \mu_D = D_o$ $H_1 : \mu_D \neq D_o$
test statistic (t distribution)	$t = \frac{\bar{d} - D_o}{s_D/\sqrt{n_D}}$		
deg. of freedom	$n_D - 1$		
rejection	reject H_o if $t < -t_\alpha$	reject H_o if $t > t_\alpha$	reject H_o if $\lvert t \rvert > t_{\alpha/2}$

[표 11.6] 짝진 표본 검정

만약 두개의 독립된 표본에서 나오는 결과로 두 집단의 평균의 차이에 대한 가설 검정을 실시할 때에는 표준편차가 같다고 가정을 하는 경우 식 (11.4)가 검정통계량이 된다.

$$\text{t-value} = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{s_p\sqrt{1/n_1 + 1/n_2}} \tag{11.4}$$

여기서 $\overline{X_1}$, $\overline{X_2}$ 는 두 표본평균이고 D_0는 통상적으로 두 평균의 차이가 없다는 귀무가설의 경우 0이다. n_1, n_2 는 각각의 표본의 크기를 의미한다. 그리고 s_p 는 공동 표준편차로서 식 (11.5)와 같다. 공동표준편차를 쓰는 이유는 두 집단의 분산이 같다고 가설검정을 하는 것이다. 이에 대한 이유는 잠시 후에 이야기 하도록 하자.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{11.5}$$

식 (11.4)의 검정통계량은 자유도가 n_1+n_2-2 인 t-분포를 따라간다. 표본의 크기가 크지 않다면 모집단은 정규분포라고 가정하여야 함은 물론이다. 이를 정리하면 [표 11.7]과 같다.

	one-tailed test		two-tailed test
hypothesis	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
test statistic (t distribution)	$t = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ $\quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$		
deg. of freedom	n_1+n_2-2		
rejection	reject H_0 if $t < -t_\alpha$	reject H_0 if $t > t_\alpha$	reject H_0 if $\lvert t\rvert > t_{\alpha/2}$

[표 11.7] 두 표본 평균 차이에 대한 검정

그러나 표본의 크기 n_1+n_2가 충분히 크면 굳이 t-분포를 이용할 필요가 없다. 식 (11.6)이 검정통계량이 된다. D_0는 통상적으로 두 평균의 차이가 없다는 귀무가설의 경우 0이다.

$$\text{z-value} = \frac{(\overline{X_1} - \overline{X_2}) - D_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{11.6}$$

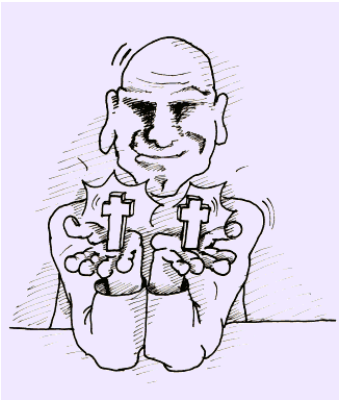
σ_1^2과 σ_2^2을 모르는 경우 s_1^2과 s_2^2으로 대신한다.

	one-tailed test		two-tailed test
hypothesis	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
test statistic (normal distribution)	$z = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$		
deg. of freedom	—		
rejection	reject H_0 if $z < -z_\alpha$	reject H_0 if $z > z_\alpha$	reject H_0 if $\|z\| > z_{\alpha/2}$

[표 11.8] 표본이 큰 경우 두 평균의 차이에 대한 검정

예제 11.3 **두 디자인을 평가하자.**

요즘의 많은 기업체들은 신제품을 출하할 때 용기에 많은 신경을 써 출하한다. 새로운 디자인을 선보이기 전에 소비자 180명에게 기존 디자인과 새로운 디자인에 대해 평가를 하여 보았다. 1-7까지의 척도로 숫자가 높으면 높을수록 평가가 좋다는 뜻이다. 두 평가 결과에 대한 상관관계는 0.740으로 높게 나온다. 한사람에게서 나온 두 개의 평가이기 때문에 짝진 표본을 구성하여 가설검정을 실시하였다.



짝진 표본은 매우 효율적이다.

검정통계량의 값은 -5.351이며 p-값은 0.000으로 나와 새로운 디자인은 완연히 다르다는 대립가설을 채택한다. 신뢰구간은

[-.801, -.277]

로 귀무가설의 가설값 0을 포함하고 있지 않다. [표 11.9]를 참조하기 바란다. ■<디자인.xls>

	A	B	C	D	E	F	G	H
1								
2	소비자	구디자인	신디자인	차이			표본의 크기	180
3	1	5	7	-2			표본평균	-0.539
4	2	7	7	0			표본표준편차	1.351
5	3	6	7	-1				
6	4	1	3	-2		차이에 대한 신뢰구간		
7	5	3	4	-1			신뢰수준	95.0%
8	6	7	7	0			표본평균	-0.539
9	7	5	7	-2			표준오차	0.101
10	8	6	7	-1			자유도	179
11	9	5	7	-2			하한값	-0.738
12	10	5	4	1			상한값	-0.340
13	11	1	3	-2				
14	12	2	1	1		양측가설검정		
15	13	6	6	0			가설값	0.000
16	14	4	5	-1			표본평균	-0.539
17	15	2	5	-3			표준오차	0.101
18	16	6	7	-1			자유도	179
19	17	4	5	-1			t-검정 통계량	-5.351
20	18	7	4	3			p-값	0.000
21	19	6	7	-1				
22	20	4	3	1		FORMULAS FROM RANGE H6:H8,H11:H16,H19:H24		
23	21	6	6	0		H6. =COUNT(D4:D183)		
24	22	3	3	0		H7. =AVERAGE(D4:D183)		
25	23	7	5	2		H8. =STDEV(D4:D183)		
26	24	4	5	-1		H12. =H7		
27	25	5	5	0		H13. =STDEV(D4:D183)/SQRT(COUNT(D4:D183))		
28	26	2	2	0		H14. =H6-1		
29	27	2	1	1		H15. =H12-TINV(1-H11,H14)*H13		
30	28	4	4	0		H16. =H12+TINV(1-H11,H14)*H13		
31	29	2	4	-2		H20. =AVERAGE(D4:D183)		
32	30	5	7	-2		H21. =H13		
33	31	7	5	2		H22. =H14		
34	32	4	4	0		H23. =(H20-H19)/H21		
35	33	2	4	-2		H24. =TDIST(ABS(H23),H22,2)		
36	34	3	1	2				

[표 11.9] 짝진 표본자료 가설검정절차

예제 11.4 **두 그룹으로 나누어 생산성을 비교하자.**

모 회사는 운동이 생산성에 미치는 영향을 조사하기 위해 구성원에게 운동을 적어도 세 번 규칙적으로 하는지 여부를 물어보았다. 1이라고 응답을 한 집단은 운동집단이고 그렇지 않은 집단은 0으로 기록하였다. 운동집단의 생산성이 더 월등하게 나온다면 이 회사는 공장에 운동시설을 설치하려고 마음먹고 있다. 생산성을 측정한 후 집단 간에 차이가 있는지 여부를 확인하였다.

생산성 비교는 자주 하는 업무다.

두 집단의 생산성이라는 변수의 차이에 대한 신뢰구간은

[-2.699, 1.862]

로 나온다. p-값은 0.359로 나와 유의수준 5%에서는 귀무가설을 기각하지 못한다. 이 결과는 모집단의 표준편차가 같다고 가정한 결과이다.

아직 두 모집단의 표준편차가 같은지 여부를 검정하는 절차는 아직 언급하지 않았다. 또한 설령 운동그룹과 비 운동그룹의 생산성의 차이가 있다 하더라도 이러한 결과는 관측 자료(observed data)에 의해 구한 것이기 때문에 생산성의 차이가 운동의 결과라고 이해하기는 힘들다. 단순히 운동을 하는 집단이 다른 집단에 비해 생산성이 높게 나왔다는 이상의 의미는 부여하기 힘들다. [표 11.10]을 참조하기 바란다. ■ <표본자료.xls>

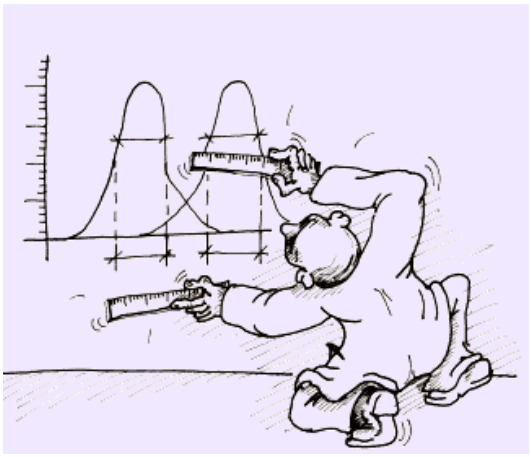
	A	B	C	D	E	F	G	H	I
1									
2	종업원	운동여부	생산성(rating)				비운동그룹	운동그룹	
3	2	1	19			표본크기	38	42	
4	4	1	14			표본평균	14.929	15.342	
5	5	1	20			표본표준편차	4.851	5.323	
6	6	1	12						
7	7	1	15						
8	8	1	16			신뢰수준	95.0%		
9	11	1	10			표본평균의 차이	-0.414		
10	12	1	24			공동 표준편차	5.105		
11	15	1	14			표준오차	1.143		
12	17	1	11			자유도	78		
13	19	1	14			하한값	-2.689		
14	20	1	11			상한값	1.862		
15	23	1	9						
16	24	1	20			가설(운동그룹의 생산성은 다른 그룹에 비해 높다.)			
17	27	1	22			가설값	0.000		
18	29	1	6			표본평균의 차이	-0.414		
19	30	1	9			공동 표준편차	5.105		
20	32	1	23			표준오차	1.143		
21	33	1	19			자유도	78		
22	34	1	18			t-검정 통계량	-0.362		
23	36	1	5			p-값	0.359		
24	37	1	5						
25	39	1	7		G7. =COUNT(Rating_0)				
26	40	1	21		H7. =COUNT(Rating_1)				
27	41	1	20		G8. =AVERAGE(Rating_0)				
28	42	1	10		H8. =AVERAGE(Rating_1)				
29	45	1	15		G9. =STDEV(Rating_0)				
30	47	1	18		H9. =STDEV(Rating_1)				
31	50	1	19		G13. =G8-H8				
32	51	1	11		G14. =SQRT(((G7-1)*G9^2+(H7-1)*H9^2)/G16)				
33	53	1	17		G15. =G14*SQRT(1/G7+1/H7)				
34	62	1	23		G16. =G7+H7-2				
35	67	1	15		G17. =G13-TINV(1-G12,G16)*G15				
36	72	1	20		G18. =G13+TINV(1-G12,G16)*G15				
37	75	1	19		G22. =G8-H8				
38	76	1	20		G23. =SQRT(((G7-1)*G9^2+(H7-1)*H9^2)/G25)				
39	78	1	19		G24. =G23*SQRT(1/G7+1/H7)				
40	80	1	13		G25. =G7+H7-2				
41	1	0	6		G26. =(G22-G21)/G24				
42	3	0	19		G27. =IF(G26<0,TDIST(-G26,G25,1),1-TDIST(G26,G25,1))				
43	9	0	16						
44	10	0	14		Rating_0. =$C$41:$C$82				
45	13	0	15		Rating_1. =$C$3:$C$40				

[표 11.10] 두 표본자료 가설 검정

11.4 분산의 동일성에 대한 가설검정

두 개의 독립된 표본에서 나오는 정보를 이용하여 두 모집단의 평균의 차이에 대한 신뢰구간이나 가설검정을 실시할 때 표준편차가 같다고 가정을 하는 것이 통상적인 관례이다.

물론 다르다고 가정을 하고 가설검정을 실시할 수 있으나 만약 같다면 굳이 후자의 방법을 취할 필요가 없다. 따라서 두 집단의 분산 혹은 표준편차가 같은지 여부를 검정하는 절차를 소개할 필요가 있다.



분산을 비교하자.

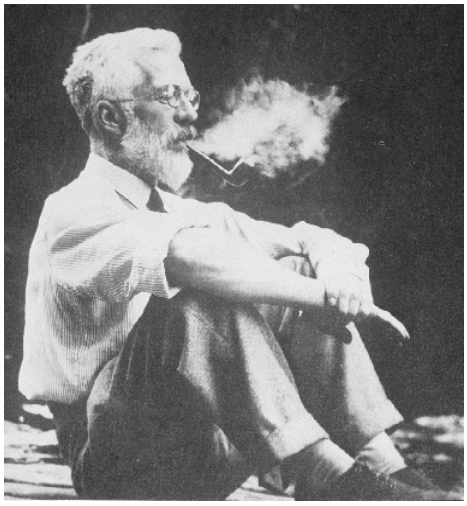
이미 우리는 표본분산 s^2 에 대한 표본추출분포를 본 바 있다. 모집단이 정규분포를 가정할 수 있다면 식 (11.7)이 성립한다.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2 \quad \text{with df} = \text{n} - 1 \tag{11.7}$$

두 개의 집단이므로 하나의 표본에서 나오는 표본분산을 s_1^2 , 다른 표본에서 나오는 표본분산을 s_2^2 라 한다. 둘 다 모집단이 정규분포를 한다고 가정한다면 식 (11.8)과 (11.9)의 두개의 확률변수는 서로 독립인 각각의 자유도가 $n_1 - 1, n_2 - 1$ 인 χ^2 분포가 된다.

$$\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2 \quad \text{with df} = \text{n}_1 - 1 \tag{11.8}$$

$$\frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2 \quad \text{with df} = \text{n}_2 - 1 \tag{11.9}$$



F-분포를 만든 R .A. Fisher

통계학의 이론에 의하면 이 두 확률변수를 각각 해당하는 자유도로 나누고 비율을 취한 새로운 확률변수는 식 (11.10)과 같은, 자유도 n_1-1, n_2-1 인 F-분포를 따라간다고 알려져 있다.

$$\frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_{2-1})s_2^2}{\sigma_2^2}/(n_2-1)} \sim F \quad \text{with} \quad n_1-1, n_2-1 \tag{11.10}$$

이러한 개념을 모의실험을 통하여 살펴보도록 한다. 엑셀에서

=norminv(rand(), 100,10)

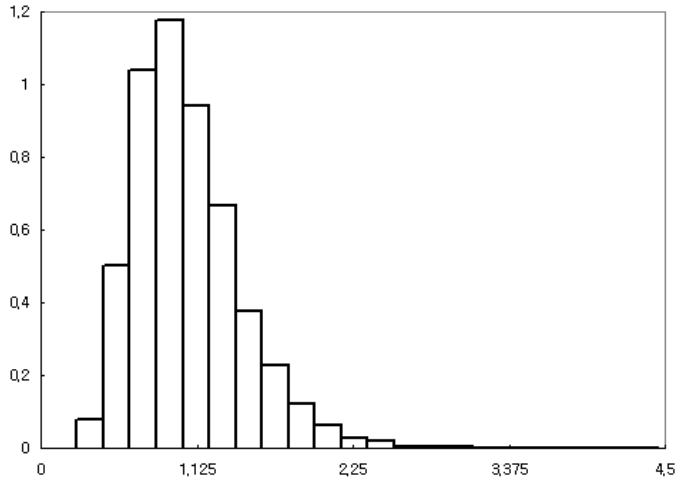
을 셀 A1에 입력한 다음 셀 A1:A30에 복사한다. 평균 100, 표준편차 10을 가지고 있는 정규분포에서 30개의 표본값을 추출한다는 의미이다. 역시 셀 B1에 평균 120, 표준편차 20을 가지고 있는 정규분포

=norminv(rand(), 120,20)

을 입력하고 셀 B1:B40에 복사한다. 그런 다음 F-확률변수의 정의에 의해 만든 값을 셀 C1에 입력한다.

=(var(a1:a30)/100)/(var(b1:b40)/400)

셀 C1을 1,000,000번 반복 시행하여 히스토그램을 그리면 [그림 11.1]을 얻는다.



[그림 11.1] F-분포(자유도, 29,39)

이 그림이 자유도 30-1=29, 40-1=39의 F-분포이다.

가설검정으로 돌아가자. 두 집단의 분산이 같다는 이야기는 "두 분산의 비가 1이다"라는 이야기와 같다.

$$\sigma_1^2/\sigma_2^2 = 1 \ <=> \ \sigma_1^2 = \sigma_2^2$$

따라서 귀무가설은

$$H_0 : \sigma_1^2 = \sigma_2^2 \ <=> \ H_0 : \sigma_1^2/\sigma_2^2 = 1$$

과 같다 . 대립가설은

$$H_a : \sigma_1^2/\sigma_2^2 \neq 1$$

가 된다.

따라서 이의 점추정값인 s_1^2/s_2^2 는 귀무가설 하에서는 자유도 n_1-1, n_2-1 인 F-분포를 하고 있다. 이것이 바로 검정통계량의 형태이다. 왜냐하면 식 (11.10)에서 정의한 확률변수 F는 귀무가설 하에서는 σ^2 가 서로 상쇄되기 때문이다.

엑셀에서는 =normdist() 와 비슷한

=fdist(v, df1, df2)

와 =norminv() 와 비슷한

=finv(p, df1, df2)

가 존재한다. 사용방법은 정규분포의 명령어 구조와 일치한다.

	one-tailed test		two-tailed test
hypothesis	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
test statistic (F distribution)	$F = \frac{s_2^2}{s_1^2}$	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$
deg. of freedom	$df_1 = n_1 - 1$ $\quad df_2 = n_2 - 1$		
rejection	reject H_0 if $F > F_\alpha$		reject H_0 if $F > F_{\alpha/2}$

[표 11.11] 두 분산의 비에 대한 검정

예제 11.5 **두 분산의 비는 어떤 분포를 하고 있는가?**

예제 11.4에서 첫 번째 표본의 표준편차가 4.851, 다른 집단의 표준편차가 5.323으로 나왔다. 따라서 검정통계량의 값은

$$(4.851)^2/(5.323)^2 = 0.830345$$

이 나온다. 그리고 자유도 37, 41인 F-분포에서 이보다 작은 값을 가질 확률은 0.284602가 나온다. 이것이 p-값이다. 엑셀에서

=1-fdist(0.830345, 37,41)

로 확인할 수 있을 것이다. p-값은 통상적인 유의수준인 5%보다 크게 나와 두 집단의 표준편차는 같다는 귀무가설을 기각하지 못한다. 따라서 이 경우에는 공동표준편차를 이용한 두 집단의 평균 차이에 대한 가설검정을 실시하는 것이 좋다.

만약 첫 번째 표준편차의 값이 두 번째 표준편차보다 크게 나왔다면 p-값은 F-확률변수 값이 검정통계량의 값보다 더 클 확률이 된다. ■

11.5 두 모집단의 비율 차이에 대한 가설검정

• 두 집단의 비율이 같은지 여부를 검정하는 것은 매우 흔한 가설검정의 한 형태이다.

p_1, p_2 를 각각의 모집단의 비율, 그리고 $\hat{p_1}$, $\hat{p_2}$ 를 크기가 n_1, n_2 인 각각의 표본에서 나오는 표본비율이라 한다면 두 표본비율의 차인 $\hat{p_1}-\hat{p_2}$ 에 의거 가설검정을 실시하기 위해서는 이에 대한 표준오차가 필요하다. 그런데

$$H_0: p_1 = p_2$$

하에서는 $\hat{p_1}-\hat{p_2}$ 의 표준오차는 식 (11.11)과 같음을 보일 수 있다.

$$SE(\hat{p_1}-\hat{p_2}) = \sqrt{\hat{p_C}(1-\hat{p_c})(1/n_1+1/n_2)} \qquad (11.11)$$

여기서 $\hat{p_c}$ 은 공동추정비율이다. 왜냐하면 귀무가설 하에서는 두 비율이 같다고 가정하였기 때문이다.

예를 들어 $\hat{p_1} = 20/85$ 이고 $\hat{p_2} = 34/115$ 이어서

$\hat{p_c}$ = (20+34)/(85+115) =54/200

이 된다. 표준오차가 주어져 있다면 표본의 크기가 충분히 크다는 가정 하에서는 식 (11.12)와 같은 검정통계량은 표준정규분포를 근사할 것이다.

$$\text{z-value} = \frac{\hat{p_1}-\hat{p_2}}{SE(\hat{p_1}-\hat{p_2})} \qquad (11.12)$$

예제 11.6 **새로운 경영방식이 타당한지 검정하자.**

모 회사는 여러 지역에 공장을 가지고 있다. 한 공장에 대해 새로운 경영방식을 실시한 후 6개월 후 이에 대한 효과가 남아 있는지 여부를 알기 위해 경영진은 종업원의 요구에 보다 신속한 대응을 하는가에 여부를 설문조사로 공장 전체에 대해 실시하였다. 만약 다른 공장과 비교하여 좀 더 나은 반응을 보였다면 이를 공장 전체로 확대하고자 한다. <비율차이.xls>

새로운 경영방식이 실시된 공장을 집단 1, 그렇지 않은 공장들을 집단 2라고 한다면 가설의 형태는

$$H_0: p_1 - p_2 \le 0, \quad H_a: p_1 - p_2 > 0$$

가 될 것이다. 여기서 $\widehat{p}_1, \widehat{p}_2$ 는 신속한 대응을 한다고 한 비율을 의미한다. 각각의 표본은 100, 300으로 한다. [표 11.12]를 참조하기 바란다.

	A	B	C
1			
2	신속히 대응을 하는가 여부	공장1	기타
3	예	39	93
4	아니오	61	207
5	Totals	100	300
6			
7	예라고 대답한 비율	0.39	0.31
8	공동 비율	0.33	
9			
10	비율의 차이	0.08	
11	표준오차	0.054	
12	검정통계량	1.473	
13	p-값	0.070	
14			
15	차이에 대한 신뢰구간		
16	신뢰수준	95%	
17	표준오차	0.056	
18	z-multiple	1.960	
19	하한값	-0.029	
20	상한값	0.189	
21			
22	B9. =B5/B7		
23	B10. =(B5+C5)/(B7+C7)		
24	B9. =B5/B7		
25	C9. =C5/C7		
26	B12. =B9-C9		
27	B13. =SQRT(B10*(1-B10)*(1/B7+1/C7))		
28	B14. =B12/B13		
29	B15. =1-NORMSDIST(B14)		
30	B19. =SQRT(B9*(1-B9)/B7+C9*(1-C9)/C7)		
31	B20. =NORMSINV(B18+(1-B18)/2)		
32	B21. =B12-B20*B19		
33	B22. =B12+B20*B19		

[표 11.12] 비율에 차이 가설검정

95% 신뢰수준의 신뢰구간은

[-0.029, 0.189]

로 나와 0을 포함하고 있다. 그리고 p-값은 0.070으로 나온다. 통상적인 유의수준 5%에서 귀무가설을 기각하지 못한다.

새로운 경영방식의 효과가 6개월이 지나면 소멸되어 다른 공장의 반응과 차이가 없다고 이야기 할 수 있을 것이다. ■

11.6 정규성 검정

검정에서 표본의 크기가 크지 않는 경우는 모집단의 분포가 정규분포여야 한다고 가정하였다.

물론 분포가 정규분포에서 약간 이탈을 하는 것은 문제가 되지 않는다. 왜냐하면 지금까지 논의한 검정통계량은 대부분 강건성(robustness)을 가지고 있기 때문이다. 이는 정규성이 아주 심하게 위배가 되지 않는 범위에서는 t-분포를 이용한 가설검정이나 신뢰구간을 구하는 것은 타당성을 어느 정도 범위에서는 인정받는다는 의미이다. 물론 분포의 정규성을 검정하는 통계적인 방법이 존재한다.

- 제일 쉬운 방법은 자료의 히스토그램을 그려 산의 모양을 하고 있는지 확인하는 작업이다. 그런 다음 정규분포라고 가정한다면 만들어져야 하는 히스토그램 사이에서 발생하는 차이가 얼마나 있는지를 확인하는 것이 정규성을 확인하는 절차이다.

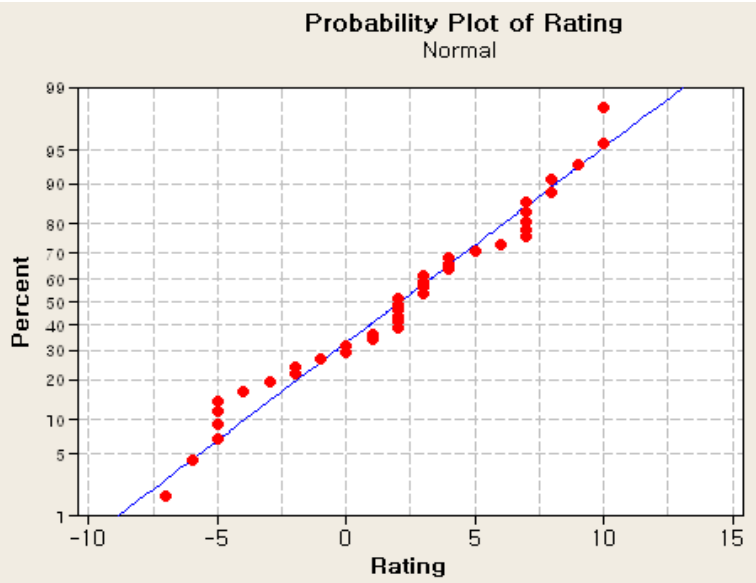
그러나 이러한 절차는 히스토그램을 만드는 데 문제가 있다. 알다시피 히스토그램은 간격의 크기에 따라 모양을 달리하는 단점을 가지고 있기 때문에 다음에 소개하는 정규확률그림이 보편적으로 많이 쓰인다.

- 정규확률그림은 자료가 정규분포에서 발생한 지 여부를 물어보는 대중적인 기법으로 절차가 약간 복잡하다. 대략적으로 설명하면 자료를 표준화한 다음 이 값(x-축)들과 자료의 평균 및 표준편차를 가정한 정규분포에서 나오는 이론적인 값(이 값은 y축 값으로 백분율로 만들어 작성)하고 산점도를 그리는 것이다.

자료가 정규분포에서 나왔다면 이 산점도는 당연히 직선 모양을 하고 있어야 한다. 따라서 산점도의 형태가 직선에서 벗어나면 정규성을 위배하였다고 할 수 있는 것이다. 직선의 오른쪽 위의 점들이 직선 위에 많이 모여 있는 경우는 정규성을 가정한 통상적인 경우보다 이상점이라 불릴 수 있는 값들이 자료에 많다는 뜻이 되기도 한다. 이러한 그림은 많은 통계 소프트웨어에서 제공한다.

예제 11.1(계속) 정규성 검정은 그림으로 하자.

예제 11.1 자료를 가지고 정규확률그림(normal probability plot)을 살펴보았다. [그림 11.2]는 -10에서 10까지의 척도를 가지고 만든 이산형의 자료이기 때문에 언급하였듯이 엄격한 의미에서 보면 정규분포를 따를 수 없지만 모든 점들이 직선을 중심으로 특별한 이탈 없이 몰려 있다는 사실을 확인할 수 있다. ■



[그림 11.2] 정규성확률그림

- 문제는 정규성이 위배가 된다고 판단되는 경우이다. 우리가 알고 있는 검정통계량은 강건성(robustness)을 가지고 있기 때문에 그렇게 큰 문제가 되지는 않으나 위배정도가 심한 경우는 변수변환 작업을 통하여 정규분포로 만들어 주는 작업을 하곤 한다. 이 문제는 책의 범위를 벗어나므로 생략키로 한다.

학습요약

가설검정은 신뢰구간 추정과 더불어 통계추론의 두 분야 중 하나이다. 어느 방법을 취할 것인지에 대한 판단은 관리자가 하겠지만 관점이 다를 뿐 크게 다른 점은 없다. 점추정의 표본추출분포와 표준오차를 가지고 출발하는 것이 공통된 사항이기 때문이다. 그러나 가설검정에 비해 신뢰구간은 더 많은 정보를 제공하는 것은 사실이다.

예를 들어 두 집단 간의 평균의 차이가 있는지 여부를 알기 위한다면 신뢰구간 추정이 더 효율적이다. 왜냐하면 신뢰구간 추정은 평균 차이가 0인지를 알려 줄 뿐 아니라 두 평균의 차이의 범위까지 알려 주기 때문이다. 많은 의사 결정은 비용에 대한 언급 없이 제 1종 오류와 제 2종 오류를 논할 수가 없으므로 통계적인 의사결정에서는 신뢰구간이 많이 애용되고 있는 것은 사실이다.

가설검정에서 제일 눈여겨 볼 값은 p-값으로 이는 분석가가 증명하고자 하는 가설인 대립가설을 지지할 충분한 증거가 있는지를 요약하는 값이다. 이러한 p-값은 웬만한 통계 소프트웨어에는 포함되어 있으며 작은 p-값은 대립가설을 지지하고 큰 p-값은 상대적으로 대립가설을 지지할 충분한 증거가 없다고 할 것이다.

또한 독자는 통계적인 유의성과 실제적인 유의성의 차이점을 이해하기 바란다. 많은 검정 절차를 소개하였으나 모두 서로 연관이 되어 있음은 이미 독자들은 눈치를 채었을 것이다. 관점의 차이일 뿐 신뢰구간의 내용과 별반 다르지 않음을 알기 바란다. 그리고 정규성 검정에 대한 논의를 하였는데 설령 정규성에 위배된다 하더라도 표본의 크기가 매우 작지 않다면 그렇게 문제가 되지 않는 것이 검정절차들이다. 잘못된 의사결정의 확률에 덜 민감한 매우 강건한 절차이기 때문이다.

11장 연습문제

11.1 6면인 주사위를 여러 개 준비하기 바란다.

(1) n=9, 16, 36개의 주사위를 동시에 던져 나오는 눈의 표본평균 및 표준오차를 구하고 모집단의 평균이 0이다 라는 귀무가설과 그렇지 않다는 대립가설을 5% 유의수준에서 검정을 실시하여라. 그리고 이런 행위를 100번하였을 때 100개의 가설검정 중 귀무가설을 기각하는 경우가 몇 번인지 확인하여 보아라.

(2) n=9, 16, 66개의 주사위를 동시에 던져 나오는 개개의 눈에 대해 4 보다 같거나 큰 숫자가 나오면 1로 기록하고 그렇지 않으면 0으로 기록한다. 그런 다음 이런 n개의 수를 합을 한 다음 n으로 나눈다. 그렇다면 표본비율인 $\hat{p}$ 를 얻을 것이다. 이를 이용하여 모집단의 비율이 0.5라는 귀무가설과 그렇지 않다는 대립가설에 대해 5% 유의수준에서 검정을 실시하여라. 이런 행위를 100번하여 100개의 가설검정 중 귀무가설을 기각하는 경우가 몇 번인지 확인하여 보아라.

이상과 같은 실험은 물리적인 주사위를 던지지 않아도 엑셀로 가능하다. 엑셀 명령문 =int(1+rand()*(7-1))을 사용하여 실험을 시행하면 된다.

11.2 6면의 주사위 9개와 16개를 두 세트 준비하기 바란다. 영희는 9개의 주사위를 동시에 던지고 철수는 주사위 16개를 동시에 던진다.

(1) 영희와 철수는 각자 주사위를 동시에 던져 나오는 주사위의 눈의 표본 평균 및 표준오차를 구하고 이를 이용하여 두 모집단의 차이는 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하여라. 그리고 이런 행위를 100번 하였을 때 100개의 가설검정 중 귀무가설을 기각하는 경우가 몇 번인지 확인하여 보아라.

(2) 영희와 철수는 각자 주사위를 동시에 던져 나오는 개개의 눈에 대해 4보다 같거나 큰 숫자가 나오면 1로 기록하고 그렇지 않으면 0으로 기록한다. 그런 다음 이런 n개의 수를 합을 한 다음 n으로 나눈다. 그렇다면 두개의 표본비율인 $\hat{p_1}, \hat{p_2}$ 를 얻을 것이다. 이를 이용하여 두 모집단의 비율의 차이는 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하여라. 그리고 이런 행위를 100번 하였을 때 100개의 가설검정 중 귀무가설을 기각하는 경우가 몇 번인지 확인하여 보아라.

11.3 (연습문제 10.3 참조) 다음은 2005년 4월 서울시에서 실시한 강동 지역 환경시설 유치에 대한 주민들 의견 중 일부를 발췌한 자료이다.

	A	B	C	D	E	F	G	H	I
1	서울 강동지역에 대한 환경시설유치에 대한 주민의견								
2									
3	age	sex	region	Children	salary('10000)	opinion		age_cat	sex_mod
4	61	F	강동	2	6,200	1		elderly	2
5	37	M	강동	2	5,200	5		middle-aged	1
6	32	F	강동	3	8,140	1		young	2
7	65	F	강동	2	4,960	1		elderly	2
8	40	M	강동	3	4,770	4		middle-aged	1
9	32	F	강동	1	5,990	4		young	2
10	38	F	강동	2	3,900	2		middle-aged	2
11	48	M	강동	1	6,150	2		middle-aged	1
12	40	M	강동	1	4,450	3		middle-aged	1
13	44	M	강동	2	4,520	3		middle-aged	1
14	57	F	강동	2	3,670	4		middle-aged	2
15	21	F	강동	2	5,430	2		young	2
16	49	M	강동	1	6,210	4		middle-aged	1
17	34	M	강동	0	7,800	3		young	1
18	38	M	강동	1	4,330	1		middle-aged	1
19	35	M	송파	1	6,540	5		middle-aged	1
20	35	M	송파	0	6,320	3		middle-aged	1
21	33	F	송파	3	4,630	5		young	2
22	45	M	송파	1	4,590	5		middle-aged	1
23	57	M	송파	1	4,810	4		middle-aged	1
24	38	F	송파	0	5,810	3		middle-aged	2
25	37	F	송파	2	5,600	1		middle-aged	2
26	42	F	송파	2	5,340	1		middle-aged	2
27	49	M	송파	0	4,320	5		middle-aged	1
28	52	M	송파	1	4,410	3		middle-aged	1
29	27	M	송파	3	4,540	2		young	1
30	40	M	송파	0	5,900	4		middle-aged	1
31	63	M	송파	2	5,390	1		elderly	1
32	48	F	송파	2	3,100	4		middle-aged	2
33	40	M	송파	0	3,770	1		middle-aged	1

(1) 이 자료를 강동 지역과 송파 지역으로 나누어 평균 의견(opinion)에 대해 차이가 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하라.

(2) 또한 이 자료를 여자와 남자로 나누어 평균 의견(opinion)에 대해 차이가 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하라.

(3) 자녀의 수가 1이하인 경우와 그렇지 않은 경우로 나누어 평균 의견(opinion)에 대해 차이가 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하라.

(4) 연령별 간에 평균 의견(opinion)에 대해 차이가 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하라.

(5) 강동 지역과 송파 지역으로 나누어 4이상의 의견(opinion)을 가지고 있는 비율의 차이가 없다는 귀무가설과 그렇지 않다는 대립가설에 대해 95% 신뢰수준에서 검정을 실시하라.

쉬어가기

1. 1종 오류와 2종 오류

가스가 새는데도 불구하고 경보가 울리지 않는다고 하자(제 2종 오류). 또한 가스가 새지 않는데도 불구하고 경보가 울린다고 하자(제 1종 오류). 1종 오류를 감소시키려면 경보기의 전원을 빼놓으면 될 것이다. 불행히도 1종 오류가 감소되면 2종 오류가 늘어날 것이다. 만일 2종 오류를 감소시키려면 탐지기의 감도를 높여야 하는데, 그러면 1종 오류인 잘못된 경보의 횟수가 늘어난다. 그럼 여기서 귀무가설 하에서 1종의 오류가 일어날 확률인 α는 무슨 의미일까? 오경보가 울릴 확률일 것이다. $1-\alpha$는 ? 우리가 들은 경보가 진짜라고 믿는 신뢰도의 척도라고 할 수 있다.

2. 통계적 유의성의 의미

어느 주장이 옳은가를 보이는 것이 통계를 이용한 많은 연구의 목적이다. 새로 개발한 약이 기존의 약 보다 임상 연구에서 효과가 있기를 바란다. 새로 연구된 교육법이 기존의 주입식 교육에 비해 효과가 있기를 바란다. 유의성 검정은 주어진 자료로부터 어떤 주장을 받아들일 수 있는지 평가하는 것이다.

이렇게 하기 위해 만약 그 주장이 옳지 않다면 무슨 일이 일어날 것인가에 대해 묻는다. 이것이 귀무가설(영가설)이다. 즉, “두 약 간에는 차이가 없다.” “교육법 간에는 차이가 없다.” 등으로 말이다. 유의성 검정은 단지 하나의 물음에만 답을 한다. “귀무가설이 참이 아닌 정도는 얼마인가?” p-값은 이 물음의 답이다. p-값은 귀무가설이 참이라면 우리가 가진 자료가 얼마나 믿기 어려운가에 대한 강도이다. 만일 p-값이 0.06이라면 귀무가설이 사실인 그 믿음의 정도가 6% 정도 된다는 것이다.

3. 검정이 꼭 필요한 현장

병아리 감별사 : 이 감별법의 탄생으로 암수에 따라 병아리의 운명을 일찌감치 가를 수 있다. 사료값 때문에 필요하겠지만 왠지 어렵겠다는 느낌이 든다. 숙련자들은 1시간에 1600마리를 구분한다고 하는데 단 0.4초 만에 1마리를 감별한다는 말이다. 100% 정확(암컷을 암컷으로, 수컷을 수컷으로)할까? 얼마나 정확하고 신속해야 “병아리 감별사”라고 부를 수 있을까? 실기시험에 합격하려면 어느 정도여야 할까? <중앙일보 기사에서 일부 발췌함>

제 4 부

자료분석 응용

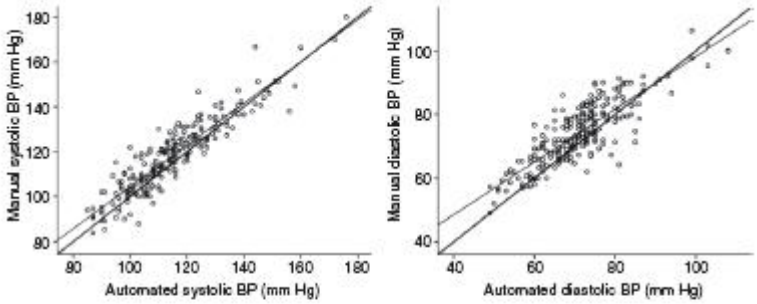


Figure 3 | A linear regression analysis was performed to examine the relationship between the automated and manual blood pressure (BP) readings with the automated systolic and diastolic BP as the independent variables. The r^2 values for systolic and diastolic BP are 0.84 and 0.70, respectively ($P < 0.001$).

제 12 장

인과관계의 추정과 예측

회귀분석의 정의 : 위키백과

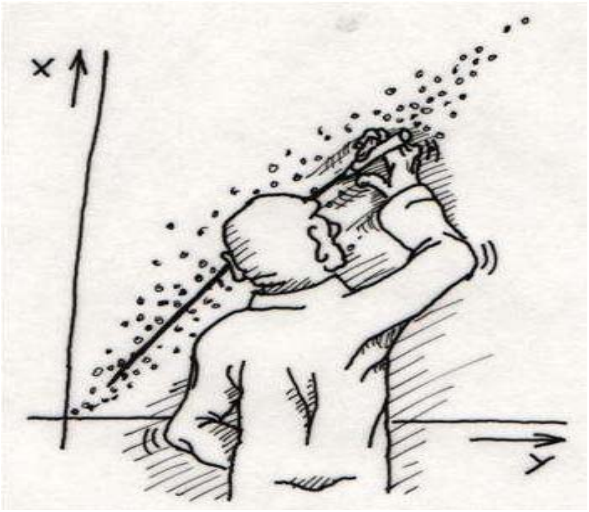
회귀분석(回歸分析, regression analysis)은 통계학에서 관찰된 연속형 변수들에 대해 독립변수와 종속변수 사이의 인과관계에 따른 수학적 모델인 선형적 관계식을 구하여 어떤 독립변수가 주어졌을 때 이에 따른 종속변수를 예측한다. 또한 이 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석 방법이다.

회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과관계의 모델링 등의 통계적 예측에 이용될 수 있다. 그러나 많은 경우 가정이 맞는지 아닌지 적절하게 밝혀지지 않은 채로 이용되어 그 결과가 오용되는 경우도 있다. 특히 통계소프트웨어의 발달로 분석이 용이해져서 결과를 쉽게 얻을 수 있지만 적절한 분석방법의 선택이였는지 또한 정확한 정보

분석인지 판단하는 것은 연구자에 달려 있다.

회귀(Regress)의 원래 의미는 옛날 상태로 돌아가는 것을 의미한다. 영국의 유전학자 프란시스 갈톤(Francis Galton)은 부모의 키와 아이들의 키 사이의 연관관계를 연구하면서 부모와 자녀의 키 사이에는 직선적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며 이를 분석하는 방법을 "회귀분석"이라고 하였다. 이러한 경험적 연구 후에 칼 피어슨(Karl Pearson)은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 수학적 전개를 정립하였다.

12.1 회귀분석이란 무엇인가?



회귀분석은 통계방법론 중에서 90% 이상을 차지한다.

- 과수의 수확량을 예측하는 기관에서는 매년 수확기에 과수를 재배한 농가 중 표본을 선택하여 표본 수확량을 조사한 후 이를 근거로 전체 과수의 수확량을 예측한다. 그러나 이 작업이 순수한 예측에서 끝나지 않고 다른 의사결정, 예를 들면, 대체과수의 수입 여부 등을 결정하여야 한다면 수확철의 표본조사는 시기적으로 늦은 감이 있다. 이러한 점에서 수확량을 직접 표본조사 하기보다 수확량을 결정하여 주는 요인들이 무엇인지를 파악함으로써 수확량을 예측하는 것이 보편화되어 있다. 한 예로 7월 말에는 포도나무 줄기에 열매가 달리기 시작하는데 이의 개수를 파악함으로써 수확철의 포도수확량을 미리 예측할 수 있다. 만약 두 변수 간에 관계식이 존재한다면 그 형태는 어떤 식으로 표현될까? 또 다른 예를 들어 보자.
- 고속도로공사는 자동차의 속도가 제동 거리에 미치는 영향을 파악하고 나아가 이를 바탕으로 여러 형태의 도로에서 적정 차간거리를 설정하기 위한 자료를 수집한다고 생각해 보기로 하자. 차의 제동 거리는 일반적으로 속도에 비례한다. 그러나 브레이크를 밟았을 때 걸리는 제동 거리와 그 때의 속도만 가지고 적정차간 거리를 이야기할 수 있을까? 아닐 것이다. 첫째로 도로의 상태 역시 제동거리에 영향을 미칠 것이다. 그러면 자료를 수집하고자 할 때의 도로 상황은 어떻게 설정하여야 하는가? 둘째로 동일한 도로 상황이 아니었으면 속도와 도로 상태의 교호작용은 제동 거리에 영향을 미치지 않을 것인가? 셋째로 계측되는 속도의 범위에 따라 제동거리가 변한다면 선형의 모형은 일정한 속도 범위를 벗어나면 타당하지 않을 것이다. 고속으로 주행하는 경우의 제동거리는 평상시의 제동거리에 비례하지 않고 더 급속하게 길 것이기 때문이다. 물론 비가 오면 더 길어질 것이다. 이상의 3가지 경우를 포함하는 경우의 실험은 어떻게 하여야 하고 수집된 자료를 어떠한 모형으로 분석해야 하는 것은 언뜻 보아 쉽지 않다. 그러나 이는 기본적으로 변수 간의 관계식을 설정하는 문제이다.

변수 간의 관계식을 이용하여 조사자나 분석가가 소기의 목적을 달성하는 것이 회귀분석이 가지고 있는 숨겨진 응용분야이다.

- 이러한 회귀분석은 시계열 분석에도 응용될 수 있다. 시간별로 모은 자료가 추세선을 가지고 있다고 판단되면 설령 자료가 계절적인 변동요소가 있다 하더라도 응용이 가능하다.

예제 12.1 **회귀분석 모형을 설정하여 보자.**

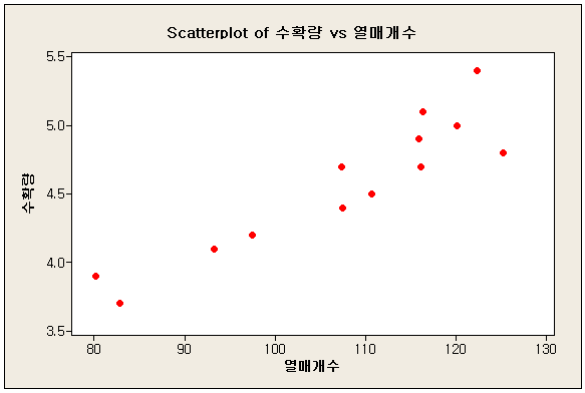
7월 말의 포도나무 모양

위에서 소개한 포도 작황 실제자료를 보자. 7월말 나무에 달린 열매 개수의 표본평균을 X, 포도수확량(톤/에이커)을 Y라 하여 1971년부터 1983년까지의 자료를 [표 12.1]에 기록하였다. [그림 12.1]은 해당하는 산점도이다. <포도수확량.xls>

	A	B	C
1	년도	수확량(Y)	열매개수(X)
2	1971	5.1	116.37
3	1972	3.7	82.77
4	1973	4.5	110.68
5	1974	4.2	97.5
6	1975	4.9	115.88
7	1976	3.9	80.19
8	1977	4.8	125.24
9	1978	4.7	116.15
10	1979	4.7	107.36
11	1980	4.1	93.31
12	1981	4.4	107.46
13	1982	5.4	122.3
14	1983	5	120.17

[표 12.1] 포도 수확량 자료

[그림 12.1]을 보면 수확량은 과수의 작황 정도의 지표인 열매 개수에 비례한다는 사실을 알 수 있다. 함수관계를 설정한다면 직선의 관계식을 생각하여 볼 것이다. 수확량을 반응변수 Y라 하고 열매개수를 설명변수 X라 하자.



[그림 12.1] 포도수확량 산점도

(X, Y)에 대한 관측값으로 (x_1, y_1), (x_2, y_2), $\cdots, (x_n, y_n)$ 이 주어졌다면 n개의 자료를 반영하는 모형은 식 (12.1)과 같이 쓰여질 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{12.1}$$

식 (12.1)에서 β_0와 β_1는 모수로서 회귀계수, ε_i는 오차라 불린다. ε_i 는 y_i와 $\beta_0 + \beta_1 x_i$의 차이를 나타내 주는데, 오차의 크기는 두 변수에 대한 관계식이 얼마나 정확하게 기술되었는지를 알 수 있기 때문에 오차의 추정값은 회귀분석에서 매우 중요한 의미를 갖는다.

아직 모형이 완성된 것은 아니다. 왜냐하면 오차에 대한 가정을 명기하지 않았기 때문이다.

- 오차의 평균은 0으로 가정하고, 발생과정에서 서로의 오차에 영향을 주지 않는다고 가정하면 임의의 서로 다른 두 오차 간의 공분산 $Cov(\varepsilon_i, \varepsilon_j)$는 0이 될 것이다. 그리고 모든 오차는 같은 크기의 분산 σ^2을 갖는다는 가정을 한다면 (12.1)은 식 (12.2)과 같이 확장하여 쓸 수 있다.

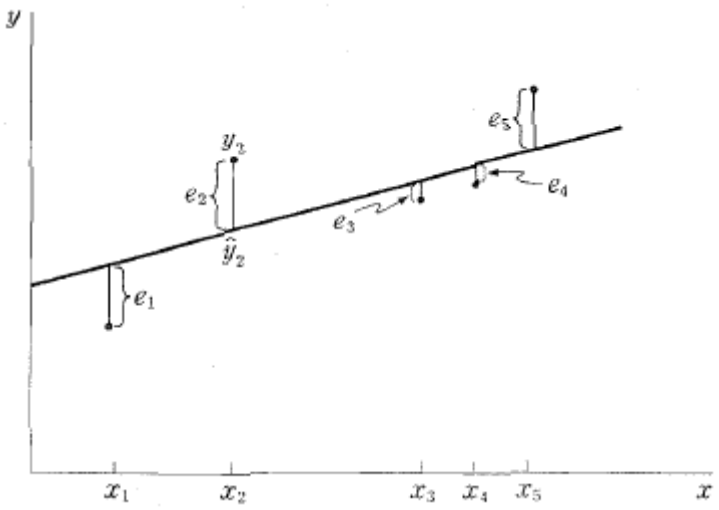
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \cdots, n.$$
$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, \quad Cov(\epsilon_i, \epsilon_j) = 0, i \neq j. \tag{12.2}$$

β_0은 $X = 0$ 일 때의 Y의 값으로 절편, β_1은 X가 한 단위 증가하였을 때의 Y의 변화량인 기울기로 불린다. 물론 β_0 및 β_1은 자료를 통하여 추정을 하여야 한다.

앞으로는 식 (12.2) 와 같은 모형을 '단순(선형)회귀모형(Simple (Linear) Regression Model)'이라 부르기로 한다.

12.2 최소제곱추정

본 절에서는 주어진 자료를 이용하여 모수들을 추정하는 방법 중 가장 많이 쓰이는 최소제곱법을 식 (12.2)로 설명하여 보자. [그림 12.2]에서와 같이 n개의 (x_i, y_i) 산점도에 직선을 그렸을 때, 선과 관측값 y_i의 사이에는 거리가 발생하는데 이를 잔차 e_i 라 한다. 어떤 잔차는 양의 값, 어떤 잔차는 음의 값을 가질 것이다.



[그림 12.2] 선 적합후의 잔차

- 최소제곱법이란 이러한 잔차의 제곱합이 최소값을 가질 수 있도록 모수를 추정하고 선을 긋는 방법을 뜻한다.

그러므로 이 방법에 의해 생기는 잔차제곱합은 다른 방법에 의해 생기는 어느 잔차제곱합과 비교하여도 제일 작은 값을 가질 것이다. 식 (12.3)은 이런 개념으로 탄생된 공식이다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad (12.3)$$

여기서 S_{xx} 는 $\Sigma(x_i - \bar{x})^2$, S_{xy} 는 $\Sigma(x_i - \bar{x})(y_i - \bar{y})$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$을 의미한다.

- ϵ_i의 추정값은 $\hat{\epsilon_i}$로 표시할 수 있으나 오차는 모수가 아니기 때문에 통상적으로 e_i로 표시한다. e_i 는 y_i와 $\hat{y_i}=\hat{\beta_0}+\hat{\beta_1}x_i$ 의 차이인 잔차로 자연스럽게 대체할 수 있다.
- σ^2의 추정은 모형에 의존하지 않으면서 추정을 할 수 있으면 가장 바람직하나, 이는 하나의 $X=x$ 값에 여럿의 반응변수의 값이 관측된 경우가 아니면 불가능하다. 이러한 경우가 아니면 σ^2의 추정은 모형의 적합도에 의존할 수 밖에 없다.

$$\hat{\sigma}^2=\frac{\sum(y_i-\hat{y}_i)^2}{n-2} \tag{12.4}$$

여기서 분자를 잔차제곱합(SSE, Error Sum of Squares 또는 Residual Sum of Squares)이라 부르고 분모는 자유도이다.

- $\hat{\sigma^2}$인 $SSE/(n-2)$는 평균의 의미를 담고 있기 때문에 평균제곱오차(MSE: Mean Squared Error)로 불린다. σ의 추정값은 MSE의 제곱근인 $\sqrt{MSE}$로 대체하고, s로 표기하기도 한다.

$$\hat{\sigma}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{MSE}=s \tag{12.5}$$

예제 12.1(계속) 최소제곱법을 이용하여 선을 그어보자.

최소제곱법으로 생긴 모수들의 추정값 및 회귀선의 특징을 자세하게 설명하기 전에 자료를 통한 모수추정의 계산과정을 살펴보기로 하자. 자료는 예제 12.1을 사용하였다.

	A	B	C	D	E	F	G	H	I	J
1	년도	수확량(Y)	열매개수(X)	y 편차	x 편차	DxE	ExE	예측값	잔차	잔차제곱
2	1971	5.1	116.37	0.530769231	9.033077	4.794479	81.59648	4.849838	0.250162	0.062581
3	1972	3.7	82.77	-0.86923077	-24.5669	21.35433	603.5337	3.806074	-0.106074	0.011252
4	1973	4.5	110.68	-0.06923077	3.343077	-0.23144	11.17616	4.673082	-0.173082	0.029957
5	1974	4.2	97.5	-0.36923077	-9.83692	3.632095	96.76506	4.263652	-0.063652	0.004052
6	1975	4.9	115.88	0.330769231	8.543077	2.825787	72.98416	4.834616	0.065384	0.004275
7	1976	3.9	80.19	-0.66923077	-27.1469	18.16756	736.9554	3.725927	0.174073	0.030301
8	1977	4.8	125.24	0.230769231	17.90308	4.131479	320.5202	5.125379	-0.325379	0.105872
9	1978	4.7	116.15	0.130769231	8.813077	1.152479	77.67032	4.843004	-0.143004	0.02045
10	1979	4.7	107.36	0.130769231	0.023077	0.003018	0.000533	4.569948	0.130052	0.016914
11	1980	4.1	93.31	-0.46923077	-14.0269	6.581864	196.7546	4.133493	-0.033493	0.001122
12	1981	4.4	107.46	-0.16923077	0.123077	-0.02083	0.015148	4.573054	-0.173054	0.029948
13	1982	5.4	122.3	0.830769231	14.96308	12.43086	223.8937	5.03405	0.36595	0.133919
14	1983	5	120.17	0.430769231	12.83308	5.528095	164.6879	4.967883	0.032117	0.001032
15						Sxy	Sxx			
16	평균	4.56923077	107.336923			80.34977	2586.553		잔차제곱합	0.451674
17										
18				FORMULAS FROM RANGE C18:C19						
19		기울기	0.03106442	C18. =F16/G16				sqrt(MSE)	0.202636	
20		절편	1.23487185	C19. =B16-C18*C16						

[표 12.2] 최소제곱법 예제

회귀계수추정을 위한 값들은 적절한 장소에 배치가 되었는데 이러한 값들을 이용한 회귀계수의 추정값은 다음과 같다.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{80.349}{2.947} = 27.258$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -2921.277$$

그리고 예측값, 잔차는 열 H와 I에, 잔차제곱합으로 구해지는 $\sqrt{MSE}$는 셀 I19에 계산되었다. ■

추정회귀선이 구해졌다면 다음과 같은 질문을 던질 수 있을 것이다.

- 선은 자료에 적합하게 그려졌는가?
- 회귀선에 포함되어 있는 가정들 즉, 오차간의 비상관성, 오차의 등분산 등은 어느 정도 만족을 하고 있는가? 만족을 하지 못하는 경우 치유할 수 있는 방안은 없을까?
- 회귀선이 예측의 목적으로 활용이 되는 경우에는 얼마만큼 믿을 수 있게 쓰는가?

이러한 질문에 대한 답을 하기 위해서는 추정량의 분포를 알 필요가 있으며 잔차를 이용한 오차의 가정에 대한 분석이 수반되어야 한다.

12.3 단순선형회귀분석에서의 통계적 추론

언급하였듯이 회귀선 추정에 필요한 연산을 마친 후에는 추정회귀선의 타당성을 판단하여야 한다. 즉, 수집된 자료에 의해 충분히 뒷받침이 될 만큼 Y와 X의 선형관계가 타당한가?라는 질문에 대하여 생각해 보아야 한다.

- β_1의 추정값 $\hat{\beta}_1$ 값이 $\hat{\beta}_1$의 표준오차(후에 설명)에 비해 상대적으로 크면 β_1는 0이 아닐 가능성이 높으므로 회귀선의 타당성이 보장받게 되는 반면, 다른 경우는 선의 타당성 여부를 의심받게 될 것이다. 이러한 논리는 표본평균에 대한 검정을 실시하는 경우에도 같은 방법으로 진행되어 왔다는 사실을 기억하면 좋을 것이다.
- 앞으로는 β_0 에 대해서는 언급하지 않을 것이다. 절편은 사실 기학학적인 의미만 있을 뿐 다른 의미는 없기 때문이다. 통상적으로 모든 회귀분석 모형에서는 유의한지에 상관없이 절편은 다 포함시킨다.

12.3.1 추정과 검정

12.2절에서 최소제곱법을 구하는 단계를 보면 오차의 분포에 대한 가정없이 전개되었으나 만약 오차가 식 (12.6)과 같이 기댓값 0이고 분산 σ^2인 정규분포를 따른다고 추가적인 가정을 한다면

$$\varepsilon_i \sim N(0, \sigma^2) \tag{12.6}$$

$\hat{\beta}_1$는 별도의 증명 없이 정규분포를 따르는 것을 짐작할 수 있다. 자세한 유도 과정은 생략한다. 그러나 독자들은 걱정하지 않아도 된다. 중요한 것은 표준오차의 개념이기 때문이다.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \tag{12.7}$$

여기서 문제가 되는 것은 σ^2을 모르는 경우이다. σ^2을 모르는 경우 표본의 크기가 작으면 오차에 대한 좀 더 제한적인 가정을 해야 한다. 다행히 자료가 정규분포에서 생성되었다고 가정하면 분산의 값을 모르는 경우라도 t-분포를 이용해 모수 β_0, β_1에 대한 신뢰구간을 구하거나 검정을 할 수 있다. 따라서 $\hat{\beta}_1$의 표준오차, $SE(\hat{\beta}_1)$가 필요한 것이다.

$$SE(\hat{\beta}_1) = \frac{\sqrt{MSE}}{\sqrt{S_{xx}}} \tag{12.8}$$

이와 같은 사실을 이용하여 우리가 관심을 갖는 모수들의 구간추정 및 가설검정을 하여 보자. 우리가 필요한 분포는 자유도 $n-2$ 를 갖는 t-분포이다.

- **기울기 구간추정**

β_1의 $100 \times (1-\alpha)\%$ 신뢰구간을 구하여 보면 식 (12.9)와 같다.

$$\left(\hat{\beta}_1 - t(\alpha/2, n-2)\ SE(\hat{\beta}_1),\quad \hat{\beta}_1 + t(\alpha/2, n-2)\ SE(\hat{\beta}_1) \right) \tag{12.9}$$

- **기울기 가설검정**

다음과 같은 가설검정

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

은 아래와 같이 수행 된다.

식 (12.10)에 제시된 검정통계값 t^*의 절대값이 임계값인 $t(\alpha/2, n-2)$와 비교하여 같거나 크게 나오면 귀무가설을 기각하고 작게 나오면 기각하지 않는다.

$$t^* = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} \tag{12.10}$$

여기서는 양측검정의 형태만 제시하였으나 단측가설검정의 형태도 물론 생각할 수 있다. 다만 이 경우는 임계값의 크기가 다를 뿐이다.

회귀분석 모형은 예측의 목적으로 많이 쓰이는데 이는 두 가지로 분류한다. 하나는 x값이 주어졌을 때 반응변수의 기댓값을 예측하는 것이고 다른 하나는 x값이 주어졌을 때 새로운 반응변수 값을 예측하는 것이다. 둘의 차이는 하나는 관심이 기댓값에 모여져 있고 하나는 새로운 관측값에 대한 예측에 모여져 있는 것이다.

• 설명변수의 값이 주어졌을 때의 반응변수의 기댓값에 대한 예측

이에 대한 점추정값은 다음과 같다.

$$\widehat{y_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \tag{12.11}$$

그리고 $SE(\widehat{y_0})$는 다음과 같다.

$$SE(\widehat{y_0}) = \sqrt{MSE}\ \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{12.12}$$

따라서 기댓값에 대한 100(1-α)% 신뢰구간은 다음과 구한다.

$$\left(\hat{y}_0 - t(\alpha/2, n-2)\, SE(\widehat{y_0}), \hat{y}_0 + t(\alpha/2, n-2)\, SE(\widehat{y_0})\right) \tag{12.13}$$

• 설명변수의 값이 주어졌을 때 반응변수의 새로운 관측값에 대한 예측

이는 주어진 반응변수에 대한 관측값의 기댓값에 관심을 두는 것이 아니라 개개의 미래 관측값에 대한 관심이다. 이에 대한 점추정값은 식 (12.11)과 마찬가지다.

$$\hat{y}_0{}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \tag{12.14}$$

그러나 $SE(\widehat{y_0^*})$는 약간 다르다. 왜냐하면 기댓값에 대한 관심이 아니라 새로운 관측값에 대한 예측이기 때문이다. 공식에 1이 추가된다.

$$SE(\ \hat{y}^*_0) = \sqrt{MSE}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} \tag{12.15}$$

따라서 새로운 관측값에 대한 100(1-α)% 신뢰구간은 다음과 같다.

$$\left(\ \hat{y}^*_0 - t(\alpha/2, n-2)SE\ (\ \hat{y}^*_0),\ \ \hat{y}^*_0 + t(\alpha/2, n-2)SE\ (\ \hat{y}^*_0)\right) \tag{12.16}$$

- 위의 식 (12.13)과 (12.16)을 살펴보면 신뢰구간은 x_0가 자료중앙인 $\bar{x}$에 가까이 있으면 폭이 좁고 x_0이 $\bar{x}$에서 멀어질수록 폭이 넓어짐을 알 수 있다.

이것은 무슨 의미인가? 중심에서 멀리 떨어지면 예측을 하더라도 표준오차의 크기가 커져 신뢰구간의 폭이 넓어져 신뢰성이 떨어진다는 의미이다.

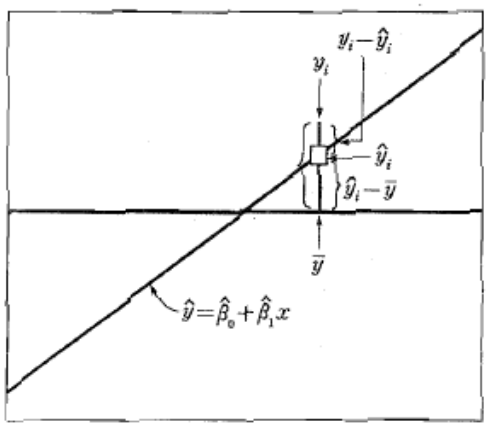
12.3.2 분산분석은 무엇인가?

다음의 두 모형을 생각해 보자.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1,2,\cdots,n \tag{12.17}$$

$$y_i = \beta_0 + \epsilon_i, i = 1,2,\cdots,n \tag{12.18}$$

모형 (12.17)에서의 y_i의 분산 σ^2의 추정값이 모형 (12.18)에서의 σ^2의 추정값보다 월등히 작으면 기울기 모수 β_1이 추가된 모형이 더 좋다고 볼 수 있고 더 나아가 설명변수와 독립변수사이에는 선형의 관계가 존재한다고 할 수 있다. 이러한 비교방법을 분산분석(ANOVA (Analysis of Variance))이라 한다.



[그림 12.3] SS의 분해

[그림 12.3]에서 $y_i - \bar{y}$ 는 $y_i - \hat{y}_i$ 및 $\hat{y}_i - \bar{y}$ 의 두 가지 요소로 분해가 되는데 이 요소를 제곱하여 합을 하면 다음과 같은 항등식이 발생된다.

$$\Sigma (y_i - \bar{y})^2 = \Sigma (y_i - \hat{y}_i)^2 + \Sigma (\hat{y}_i - \bar{y})^2 \qquad (12.19)$$

$$(SST = \ SSE + SSR)$$

식 (12.19)의 좌변은 식 (12.18)의 모형을 적합하였을 때의 SST(이를 통상 총제곱합(SST(Total Sum of Squares))이라 부른다.)이고 식의 우변 중에서 첫 번째 항은 식 (12.17)의 모형을 적합하였을 때 생기는 SSE이다.

- 따라서 SSR(Regression Sum of Squares)은 선을 그었을 때 생기는 SSE의 추가 감소량이다. 이를 회귀제곱합이라 한다. 결국, SSR이 얼마만큼 크냐에 따라 회귀선의 타당성 여부가 결정될 수 있는 것이다.

이것이 분산분석이다. 다음 [표 12.3]에 전형적인 분산분석표가 주어져 있는데 간단히 이야기하면

- 분산분석은 SSR의 크기를 평균으로 낸 값(MSR)과 SSE를 평균으로 낸 값(MSE)과 비교하는 것이다.

여기서 평균을 낸다하는 의미는 각각의 SS를 해당하는 자유도로 나누어 주는 작업을 의미한다. MSR, 즉 선을 적합 시켰을 때 추가로 설명되는 부분의 평균이 선을 적합 시켜도 설명되지 않는 부분의 평균, MSE에 비해 월등히 크다면 이것은 우연한 일치의 결과로 보지 않고 선의 효과라고 보는 것이다. 이러한 검정은 F-분포를 이용한다. 이에 대한 설명은 실제 자료 분석에서 구체적으로 보기로 하자. [표 12.3]에서 S_{yy}는 $\sum_{i=1}^{n}(y_i - \bar{y})^2$를 의미한다.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F^*
Regression	$SSR = \hat{\beta}_1 S_{xy}$	1	$SSR/1 = MSR$	MSR/MSE
Residual	$SSE = Syy - \hat{\beta}_1 S_{xy}$	$n-2$	$SSE/(n-2) = MSE$	
Total	$SST(= S_{yy})$	$n-1$		

[표 12.3] 분산분석표

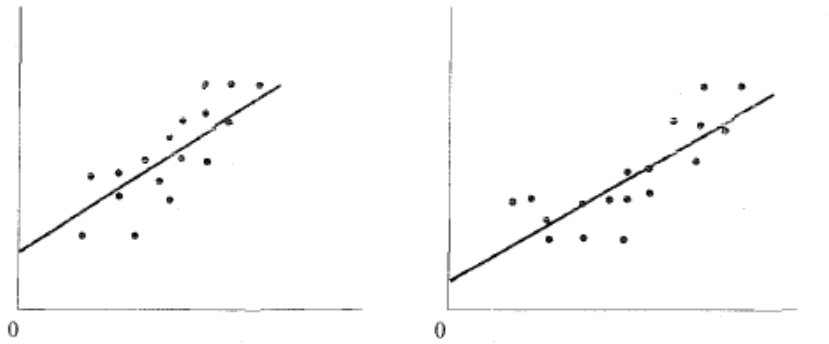
12.3.3 결정계수

- 결정계수 R^2는 다음과 같이 정의한다.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}} \tag{12.20}$$

결정계수는SSR의 S_{yy} 에 대한 비율 즉, 총변동 중에서 회귀모형에 의하여 설명되어지는 변동의 크기를 의미한다. SSE나 SSR은 S_{yy} 보다 작기 때문에 결정계수 값은 항상 0과 1사이에 있다. 1에 가까우면 선형관계가 강력함을 의미하고 0에 가까우면 강력하지 않음을 뜻한다.

결정계수에서 의미되는 관계는 비선형이 아니라 오직 선형의 관계를 의미하는 것이다. 그러나 아래 [그림 12.4]에서 보듯이 경우에 따라서는 왼쪽의 경우뿐 아니라 오른쪽의 경우와 같이 비선형의 관계라 할지라도 결정계수 값이 클 수가 있다.



[그림 12.4] 선형 및 비선형모형에서의 R^2

- **결정계수의 크기 여부**

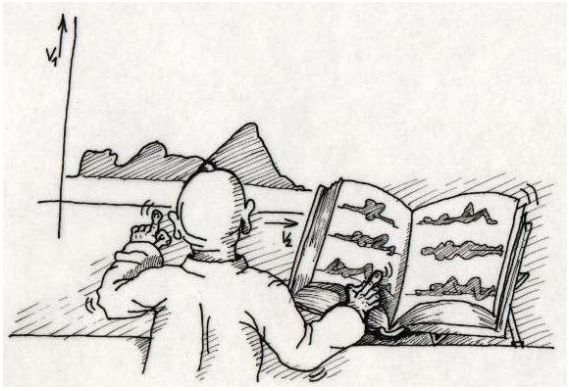
흔히, 많은 이들은 회귀분석결과를 이야기할 때 결정계수 값이 얼마만큼 크면 좋고, 혹은 절대적인 기준은 없느냐하는 질문을 한다.

화학적인 현상의 모형을 구축하기 위해서 화학자가 심적으로 느끼는 결정계수 값은 98%인 반면 실증주의적인 사회과학자들 사이에서는 결정계수 값은 $60-70\%$의 수준이면 만족한다고 믿는 경우가 있다. 그러나 이러한 기준은 절대적인 것이 아니라 다분히 상대적인 것이다. 또한 자료의 구성형태에 따라 값이 결정될 수 있다. 시계열자료를 이용한 회귀모형에서의 결정계수 값은 매우 큼을 경험상 알고 있는데, 이는 많은 시계열자료에서의 반응변수의 값은 평균값이기 때문이다. 즉, 시계열자료에서는 관측값들은 이미 그 평균에 의해서 변동이 줄어든 반면, 비시계열자료인 경우에는 그렇지 않기 때문이다.

응용과학분야에서는 결정계수의 값은 요약통계량이지 "성적" 통계량은 아닌 것이다.

12.4 예제로 본 단순선형회귀분석

예제 12.1(계속) 모든 계산은 통계소프트웨어로 이루어진다.



모형을 적합 시키자.

예제 12.1을 미니탭이란 통계 소프트웨어를 이용하여 모든 필요한 결과물을 구현하였다. [표 12.4]와 같은 결과물은 어느 소프트웨어나 거의 비슷하게 나온다. 열매개수를 설명변수로, 그리고 수확량을 반응변수로 하여 구한 결과물이다.

Regression Analysis: 수확량 versus 열매개수

The regression equation is
수확량 = 1.23 + 0.0311 열매개수

Predictor	Coef	SE Coef	T	P
Constant	1.2349	0.4313	2.86	0.015
열매개수	0.031064	0.003984	7.80	0.000

S = 0.202636 R-Sq = 84.7% R-Sq(adj) = 83.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.4960	2.4960	60.79	0.000
Residual Error	11	0.4517	0.0411		
Total	12	2.9477			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	5.1179	0.0901	(4.9197, 5.3162)	(4.6299, 5.6060)

Values of Predictors for New Observations

New Obs	열매개수
1	125

[표 12.4] 포도 수확량 자료 회귀분석 결과물

먼저 [표 12.4]에서 추정회귀식을 보면

$$\hat{y} = 1.23 + 0.0311X$$

이고 $\sqrt{MSE} = s = 0.202636$, 결정계수는 84.7%가 나왔다. 또한 분산분석에 필요한 제곱합들은

$$SST = 2.9477 \text{ with d.f. = 12}$$
$$SSE = 0.4517 \text{ with d.f. = 11}$$
$$SSR = 2.4960 \text{ with d.f. = 1}$$

로 계산되었고 해당하는 자유도로 나눈 평균제곱(MS)은

$$MSE = SSE/11 = 0.0411$$
$$MSR = 2.4960/1 = 2.4960$$

이며 이 두 평균제곱의 비가 결과물에서 보는 F-값이다.

F-값 = 2.4960/0.0411 = 60.79

이 F-값은 F(df1 = 1, df2= 11)인 분포를 따른다. 그리고

$$H_0: \quad y_i = \beta_0 + \varepsilon_i, i = 1, 2, \cdots, n$$
$$H_1: \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \cdots, n$$

의 가설검정 하에서 p-값은 0에 가깝기 때문에 귀무가설을 기각하기에 충분한 증거가 있다고 판단된다. 위의 가설검정은 다음과 같이 표현하기도 한다.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

그리고 올해 열매개수가 125개로 판단된다면 수확량 예측은

$$\hat{y} = 1.23 + 0.0311\,(125) = 5.1179$$

로 파악되며 95% 신뢰구간은 결과물에서 95% PI에서 볼 수 있다.

(4.6299, 5.6060)

또한 열매개수가 125개로 주어졌을 때 평균 수확량에 대한 신뢰구간은 95% CI에서 볼 수 있다.

(4.9197, 5.3162)

이는 각각 식 (12.16)와 (12.13)에 해당하는 신뢰구간이다.

- 또한 F-값은 t 값의 제곱으로 확인할 수 있다.

$$(t^*)^2 = (7.80)^2 = 60.79 = F^*$$

즉 단순회귀분석에서는 t를 사용하던지 분산분석의 F를 사용하던지 기울기의 검정(설명변수의 필요성)을 실시할 수 있다.

다중선형회귀모형과 단순선형회귀모형의 차이점은?

12.5 다중선형회귀분석

지금까지 소개한 단순선형회귀모형에서는 모형의 개발 및 해석상의 어려움은 그렇게 크다고 볼 수 없다. 그러나 설명변수의 수가 두 개 이상인 다중선형회귀모형에서는 변수 선택 및 계수의 해석은 상당히 신중을 기하여야 한다. 다중선형회귀모형의 여러 가지 성질을 알아보기 전에 이러한 현상이 어디에서 기원되는지 알아본다.

12.5.1 실험계획과 관측조사에 의한 자료수집

어떠한 이론에서 가설이 나왔다고 하자. 만약 이러한 가설의 타당성을 검정하기 위해 자료를 이용한다면 그 자체가 넓은 의미의 실험이라 할 수 있다. 이러한 의미에서는 응용분야의 연구가 전부 실험이라는 범주에 속할 수 있다.

그러나 가설을 검정하기 위해 수집한 자료가 어떻게 수집되었느냐에 따라 회귀모형에서의 계수에 대한 해석은 판이하게 달라진다. 다음 두 예제를 통해 이를 설명하여 보자.

예제 12.2 **통제될 수 있는 상황에서의 결과는 인과관계가 있다.**

곡물의 생산량 Y는 다음과 같은 변수들에 의해 결정이 된다고 알려져 있다.

$$Y = f(F, T, R, H, T) + \epsilon \tag{12.21}$$

여기서 $\mathrm{F(fertilizer)}$는 비료의 종류, $\mathrm{T(temperature)}$는 재배기간 동안의 평균온도, $\mathrm{R(rainfall)}$은 재배기간 동안의 강우량, $\mathrm{H(herbicide)}$는 제초제의 종류, $\mathrm{T(Technology)}$는 농업기술을 의미하는 변수이다. 물론 이외에도 많은 변수가 있을 수 있으나 여기서는 이 다섯 변수만 고려하기로 하자.

어느 회사에서 새롭게 개발한 비료가 기존의 비료와 비교하여 수확량의 효과가 얼마나 있는지 알아보고 싶어 한다. 이를 위해 재배면적을 둘로 나누어 하나는 기존의 비료, 다른 하나는 새로 개발된 비료를 투여하여 수확량을 비교하는 실험을 시행하였다면 이 두 재배지역은 강우량이나 온도가 될 수 있으면 같은 상황이 될 수 있게 통제되어야 할 것이다. 이런 통제된 상황이 아니면 변수간의 상호작용으로 인하여 비료의 수확량에 대한 순수효과에 대한 비교측정은 거의 불가능할 것이다. 예를 들면 비료 및 강우량과의 상호작용 등은 수확량에 영향을 미친다고 볼 수 있다. 뿐만 아니라 통제된 상황이 아닌 경우에서 나오는 결과만 가지고는 비료와 수확량의 인과관계를 설정하는 데 많은 논란이 있을 수 있다.

즉, 일반적으로 특정한 설명변수 F 에 의한 효과를 보고자 할 때는 식 (12.21)과 같은 모형보다는 식 (12.22)의 모형을 고려하여야 한다.

$$Y = f(F \mid T, R, H, T) + \epsilon \tag{12.22}$$

그러나 실제로는 이와 같이 비용과 시간이 많이 들어갈 수 있는 완벽한 통제된 상황 하에서 자료 수집을 하기보다는 수확량에 변동을 줄 수 있는 인자의 폭을 되도록 줄여가면서 실험계획을 할 것이다.

사회과학분야에서는 이러한 실험계획에 의한 자료는 기대하기가 어려울 뿐 아니라 설사 그렇다 할지라도 상당한 어려움이 자료를 수집하는 과정에서 발생한다. ■

예제 12.3 **사회과학 자료는 통제된 자료가 아니다.**

어느 교육 사회학자 교육수준 E에 따른 소득수준 Y의 변화를 연구를 하고자 할 때 다음의 단순한 선형모형을 가정하였다 하자.

$$Y = f(E) = \beta_0 + \beta_1 E + \epsilon \tag{12.23}$$

그러나 이러한 모형은 교육수준에 상관없이 젊었을 때보다는 나이가 들었을 때 소득이 높다는 논리를 반영하지 못한다. 그렇다고 위의 예와 같이 연령 A는 통제할 수 있는 성격의 변수가 아니기 때문에 교육수준에 따른 소득수준의 변화를 직접적으로 알기에는 많은 어려운 점이 따른다. 즉, 통제된 실험계획에 의한 자료 수집은 불가능하다.

많은 사회과학분야에서의 자료 수집은 이러한 문제점을 안고 있다. 이러한 경우는 연령과 같은 변수를 통제하기보다는 모형(넓은 의미에서의 실험)에서 이러한 변수에 의한 효과를 추정하고 이를 감안하여 교육수준이 소득수준에 미치는 효과를 검정해야 할 것이다. 이를 위해 아래와 같은 모형을 설정하였다고 하자.

$$Y = f(E, A) = \beta_0 + \beta_1 E + \beta_2 A + \epsilon \tag{12.24}$$

즉, β_1에 대한 해석은 A의 변수의 값이 주어진 상태에서 Y에 대한 E의 효과를 뜻한다. 이러한 의미에서 계수 β_1를 (편)회귀계수라 일컫는다. 여기서 말하는 (편)에는 다른 변수의 값은(즉 A는 A^*의 값으로서) 주어졌다는 의미가 들어가 있다.

$$\begin{aligned} Y = f(E, A^*) &= \beta_0 + \beta_1 E + \beta_2 A^* + \epsilon \\ &= (\beta_0 + \beta_2 A^*) + \beta_1 E + \epsilon \\ &= \beta_0^* + \beta_1 E + \epsilon \end{aligned} \tag{12.25}$$

여기서 β_0^* 는 $\beta_0 + \beta_2 A^*$ 을 의미한다.

물론 이러한 자료의 수집단계에 있어 A는 통제된 상황 하에서 수집된 것이 아니라 조사된 것이기 때문에 E와 Y의 (비록 실질적인 인과관계가 있다 하더라도) 인과모형을 설정할 수 없는 것은 당연하다. 앞으로 전개될 다중선형회귀모형에 대한 설명을 이러한 맥락에서 이해하여야 한다. ■

12.5.2 다중선형회귀모형

다중선형회귀모형에서는 하나의 반응변수에 대해 여러 개의 설명변수가 주어져 있다. 설명변수를 $X_1, X_2, X_3, \cdots, X_p$라 하면 아래 [표 12.5]와 같은 구조를 통하여 자료가 수집된다.

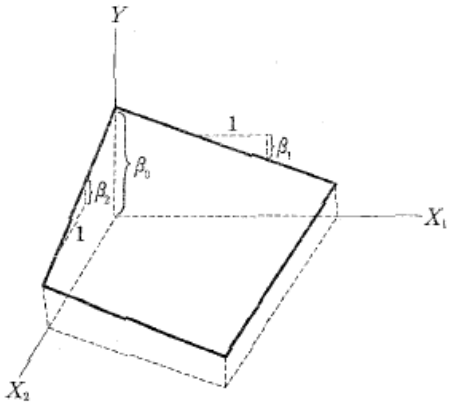
자료의 순서	변수값					
	y	X_1	X_2	X_3	...	X_p
1	y_1	x_{11}	x_{12}	x_{13}	...	x_{1p}
2	y_2	x_{21}	x_{22}	x_{23}	...	x_{2p}
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
n	y_n	x_{n1}	x_{n2}	x_{n3}	...	x_{np}

[표 12.5] 다중선형회귀모형을 위한 자료표

이러한 표현방법에 의하면 x_{ij}는 j번째 설명변수에서 i번째 자료의 값을 나타낸다. 다중회귀분석모형식은 다음과 같이 쓸 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, i = 1,2,\cdots,n \qquad (12.26)$$

p=2일 때에는 [그림 12.5]와 같이 3차원 공간에서 2차원 평면으로 설명할 수 있다.



[그림 12.5] p=2인 경우의 선형회귀반응표면

예제 12.4 **다중선형회귀모형은 무엇이 다른가?**

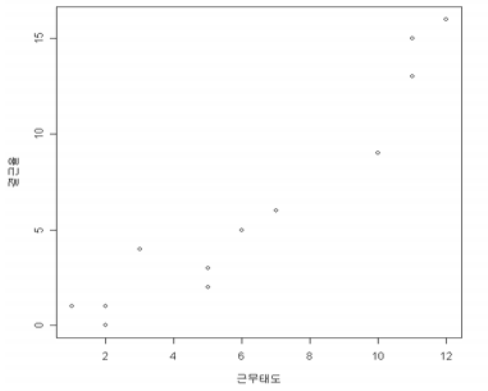
다음 [표 12.6] 자료는 어느 보험회사에서 일하는 12명의 남자사원에 대해 결근율을 조사한 결과이다. 설명변수 X_1은 회사원의 근무태도를 13개의 척도로 측정하여 얻은 결과이다. 점수가 낮을수록 근무태도가 좋다. 설명변수 X_2는 회사의 근무년수를 표시한 설명변수이고, 반응변수 Y의 값으로는 개개의 종업원이 결근한 날짜를 인사카드에서 발췌하여 기록하였다. 이와 같은 두 개의 설명변수를 이용하여 종업원의 결근율(absenteeism)을 분석하고자 한다.

<결근율.xls>

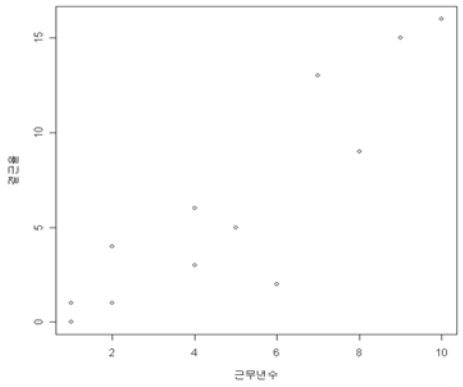
Y	X_1	X_2	Y	X_1	X_2
1	1	1	6	7	4
0	2	1	5	6	5
1	2	2	9	10	8
4	3	2	13	11	7
3	5	4	15	11	9
2	5	6	16	12	10

[표 12.6] 결근율 자료

다음 [그림 12.6]은 Y와 X_1의 산점도이고, [그림 12.7]은 Y와 X_2의 산점도이다.



[그림 12.6] 산점도(Y 대 X_1)



[그림 12.7] 산점도(Y 대 X_2)

그리고 각각 단순선형회귀모형을 구하면 다음과 같다.

$$\hat{y} = -2.31278 + 1.37004x_1 \quad R^2 = 0.9021$$
$$\hat{y} = -1.71708 + 1.62042x_2 \quad R^2 = 0.7935$$

근무년수가 한 해 늘수록 또는 근무태도의 척도가 한 단위 늘수록 각각 결근율은 1.62 및 1.37씩 증가한다고 볼 수 있다.

우리의 목적은 두 변수를 모두 이용하여 Y를 설명하고자 하는 것이다. 두 변수를 모두 포함하는 모형으로 다음과 같은 다중선형회귀모형을 가정하여 보자.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, i = 1, 2, \cdots, n \qquad (12.27)$$

두 변수를 모두 이용하는 경우 X_1과 X_2의 서로 중복되는 부분을 제거할 필요가 있을 것이다. 식 (12.27)의 다중선형회귀모형에서는 X_1이 설명하는 부분을 제외한, 나머지 부분을 X_2가

X_1과 중복되지 않고 얼마나 설명하느냐 하는 것이 그 핵심이라 할 수 있다. 그리고 X_2로 인한 추가설명력이 얼마만큼 유의하느냐를 결정하여야 한다.

다중선형회귀모형을 최소제곱법을 이용하여 적합하여 보면 다음과 같이 나온다.

$$\hat{y} = -2.2630 + 1.5497x_1 - 0.2385x_2$$

단순회귀에서 나온 기울기와 다중회귀에서 나온 두 기울기의 값을 비교하여 보자. 먼저 두 번째 회귀계수 값을 비교하여 보면

$\hat{\beta}_2 = 1.62042$: X_1을 무시한 계수의 값

$\hat{\beta}_2 = -0.2385$: X_1 에 대해 조정된 계수의 값

이는 X_1이 조정된 상태에서는 근무년수가 올라가면 갈수록 결근율은 떨어짐을 암시한다. 이는 1.62의 값보다 훨씬 상식적인 의미를 가질 수 있으나 이에 대한 판단은 아직 이르다. p-값이 보여주는 통계적인 유의성을 따져 주어야 하기 때문이다. 또한 이러한 음의 부호에 대한 결론은 X_1과 X_2사이에서 발생할 수도 있는 교호작용(interaction), 즉 근무년수가 많고 태도가 나쁘면 보통의 경우보다 상대적으로 결근율이 높을 수 있다는 사실을 감안하지 않았기 때문에 유보되어야 할 것이다. 게다가 이 자료의 경우에는 계수의 부호가 바뀔 정도로 X_1과 X_2의 상관관계가 있었기 때문에 이러한 교호작용을 추적한다는 것은 상당히 어렵다. 이러한 현상으로 말미암아 다중선형회귀모형에서의 계수의 값을 해석하는데 어려움이 많다는 사실을 기억하기 바란다. 이에 대한 논의는 잠시 후에 좀 더 자세히 다루기로 한다.

같은 방법으로 $\hat{\beta}_1$을 구하면 즉, 설명변수 X_2가 주어진 상태에서는 기울기가 1.5497임을 알 수 있는데 위의 $\hat{\beta}_2$의 경우와 달리 $\hat{\beta}_1$의 값은 별 차이를 보이지 않는다.

$\hat{\beta}_1 = 1.37004$: X_2를 무시한 계수의 값

$\hat{\beta}_1 = 1.5497$: X_2에 대해 조정된 계수의 값

이와 같이 다중선형회귀모형에서의 회귀계수 값은 단순회귀모형과 달리 해석이 항상 다른 변수 값이 이미 모형 안에 들어와 있다는 상황을 가정하고 해야 한다. ■

예제 12.4(계속) **다중선형회귀모형도 컴퓨터 분석을 하여보자.**

보험회사의 다중선형회귀모형에 대한 컴퓨터 결과를 요약하면 [표 12.7]과 같다.

회귀 분석: 결근율 대 근무태도, 근무년수

회귀 방정식은
결근율 = - 2.26 + 1.55 근무태도 - 0.239 근무년수

예측 변수	계수	SE 계수	T	P
상수	-2.263	1.096	-2.06	0.069
근무태도	1.5497	0.4805	3.23	0.010
근무년수	-0.2385	0.6064	-0.39	0.703

S = 1.94653 R-제곱 = 90.4% R-제곱(수정) = 88.2%

[표 12.7] 결근율 자료 회귀분석 결과물

먼저 다음과 같은 형태의 가설검정에 대한 F-검정을 실시하여 보자.

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: \text{적어도 한 개의 } \beta_i \neq 0, i = 1, 2.$$

[표 12.8]은 X_1과 X_2의 설명력을 다 합하여 계산한 분산분석표이다. 여기서의 F-통계량 값은 42.25 이고 이는 $F(2, 9)$ 분포를 따른다. p-값이 0에 가까운 아주 작은 값이 계산되어 나왔기 때문에 위의 귀무가설을 기각한다.

분산 분석

출처	DF	SS	MS	F	P
회귀	2	320.15	160.07	42.25	0.000
잔차 오차	9	34.10	3.79		
전체	11	354.25			

[표 12.8] 결근율 자료 분산분석표

그렇다면 두 번째 회귀계수의 값의 0의 여부를 묻는 다음과 같은

$$H_0: \beta_2 = 0 \ (\beta_1\text{의 값은 임의의 값})$$
$$H_1: \beta_2 \neq 0 \ (\beta_1\text{의 값은 임의의 값})$$

형태의 가설검정을 실시하면 t-통계량의 계산된 값, $t^*=-0.39$ 의 절대값은 유의수준 α= 5%에서의 임계값, $t(\alpha/2,9)$인 2.262보다 작으므로 귀무가설을 기각하지 못한다. 즉, 귀무가설을 기각하기에는 p-값이 너무 크다. 반면 β_1 계수에 대한 p-값은 매우 작게 나온다.

이 모형을 갖고 [표 12.9]처럼 $(X_1=2,\ X_2=5)$의 값을 갖는 직원의 결근율, y_0에 대한 95% 신뢰구간을 구하여 보자. $\widehat{y_0}\pm t(\alpha/2,11)SE(\widehat{y_0})$에 의하면

(-6.941, 6.229)

로 나타난다. 예측값이 음으로 나왔을 뿐 아니라 신뢰구간의 폭이 너무 커 다중선형회귀모형을 갖고 이와 같은 직원의 결근율을 예측하는 것은 무리가 있다. 이 경우는 오히려 X_2 변수가 빠진 단순선형회귀모형이 더 예측의 목적에 맞을 수가 있다.

새로운 관측치에 대한 예측치

새로운 관측치	적합치	SE 적합치	95% CI	95% PI
1	-0.356	2.164	(-5.253, 4.540)	(-6.941, 6.229)X

X는 예측 변수에서 특이치인 점입니다.

새로운 관측치에 대한 예측 변수의 값

새로운 관측치	근무태도	근무년수
1	2.00	5.00

[표 12.9] 결근율 자료 예측구간

분석을 좀 더 하여 보자. X_1과 X_2사이에서 발생할 수도 있는 교호작용을 고려한 다음과 같은 다중선형회귀분석모형을 생각하여 보자.

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_{12}x_{i1}x_{i2}+\epsilon_i, i=1,2,\cdots,n \qquad (12.28)$$

분산분석과 t-검정의 결과를 보면 다음 [표 12.10]과 같다. 추정된 회귀식은 다음과 같다.

$$\hat{y}=0.9990+0.8558x_1-1.1941x_2+0.1394x_1x_2$$

유의수준 5%에서 근무태도에 대한 회귀계수 $\widehat{\beta_1}$는 유의하지 않으나(유의수준 10%에서는 유의함.) 근무년수에 대한 회귀계수 $\widehat{\beta_2}$와 근무태도와 근무년수의 교호작용에 대한 회귀계수 $\widehat{\beta_{12}}$는 둘 다 유의하다. 근무태도와 근무년수의 교호작용을 고려하지 않은 모형보다 결정계수값도 크

다. $\widehat{\beta_{12}}$=0.1394이므로 근무년수가 많고 태도가 나쁘면 보통의 경우보다 상대적으로 결근율이 높을 수 있다는 사실을 확인할 수 있다.

회귀 분석: 결근율 대 근무태도, 근무년수, 근무태도*근무년수

회귀 방정식은
결근율 = 1.00 + 0.856 근무태도 - 1.19 근무년수 + 0.139 근무태도*근무년수

예측 변수	계수	SE 계수	T	P
상수	0.999	1.273	0.79	0.455
근무태도	0.8558	0.4000	2.14	0.065
근무년수	-1.1941	0.5185	-2.30	0.050
근무태도*근무년수	0.13939	0.04337	3.21	0.012

S = 1.36389 R-제곱 = 95.8% R-제곱(수정) = 94.2%

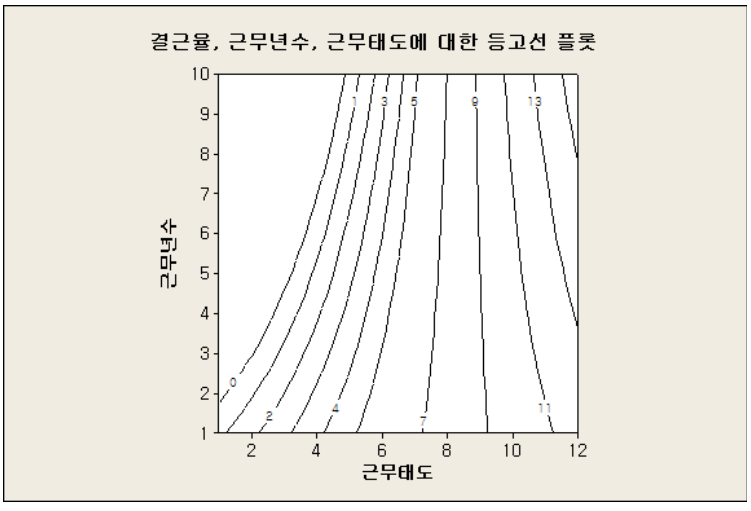
분산 분석

출처	DF	SS	MS	F	P
회귀	3	339.37	113.12	60.81	0.000
잔차 오차	8	14.88	1.86		
전체	11	354.25			

출처	DF	Seq SS
근무태도	1	319.56
근무년수	1	0.59
근무태도*근무년수	1	19.22

[표 12.10] 교호작용을 고려한 다중선형회귀모형

추정된 회귀식을 이용하여 등고선도를 그리면 다음과 같다. 근무태도와 근무년수 사이의 교호작용 효과를 확인할 수 있다. 결근율이 1 이하인 조건과 결근율이 13 이상인 조건을 비교하여 보아라. ■



[그림 12.8] 교호작용을 고려한 다중선형회귀모형을 이용한 등고선도

12.6 모형진단과 예제

지금까지는 자료에 모형을 적합시키는 관점에서 살펴보았지만 우리가 다루고 있는 모형과 부수되는 가정에는 어떤 문제점은 없는가를 살펴볼 필요가 있다. 주어진 모형과 가정이 맞지 않는다면 잘못된 모형을 적합시켜 나오는 결과물에 의한 추론은 무의미하기 때문이다. 본 절에서는 이러한 문제를 알아보는 것이다.

여기서 기본적이고 핵심적으로 사용되는 통계 값은 잔차이다. 잔차의 형태로부터 추정모형이 관측된 자료를 얼마나 닮았는지 알 수 있으며 우리가 세운 가정이 바람직한지 알 수 있다. 그러면 회귀모형에 부여되는 가정은 어떠한 것들이 있는가를 다시 알아보자.

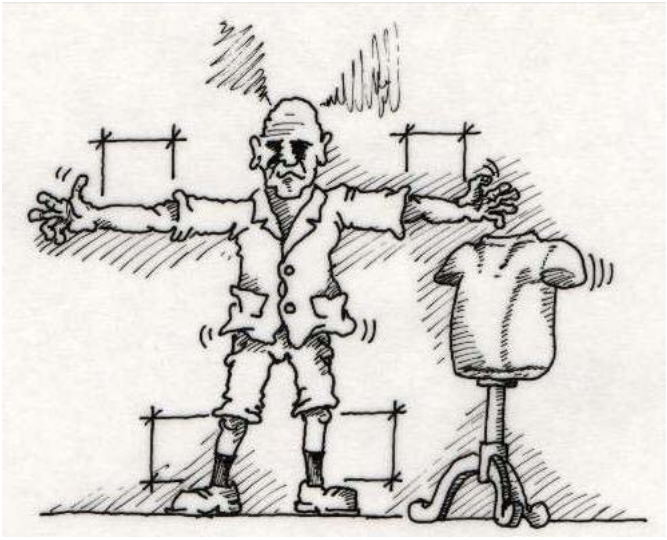
- 첫째, 오차의 등분산성(homogeneity)을 들 수 있다. 그러나 오차 분산의 크기는 반응변수나 한 두 개의 설명변수, 그리고 시간이나 장소 같은 요인에 의해서 달라지는 경우가 많기 때문에 이 가정이 만족되지 않을 수 있다.
- 둘째, 모형의 선형성(linearity)을 들 수 있다. 실제 문제는 선형보다 비선형인 관계로 설명되는 경우가 종종 있다.
- 셋째, 오차는 정규분포를 따른다는 가정을 들 수 있다. 그러나 회귀분석의 추정과 검정에 쓰이는 t-분포는 오차의 정규분포 가정이 약간 위배되더라도 덜 민감한, 강건한(robustness)특성을 갖고 있기 때문에 가정이 만족되지 않았을 때의 심각성은 위의 두 경우보다 덜하지만 정규성의 심각한 위배는 경우에 따라 문제가 될 수 있으므로 이에 대한 진단이 필요하다.

이러한 가정들이 타당한지 알 수 있는 가장 보편화된 방법은

- 표준화된 잔차를 y축으로 하고 $\hat{y}$ 을 x축으로 하는 산점도(이 그림을 잔차도(residual plot)라 부른다.)를 그려보는 것이다.

오차는 정규 분포를 따른다고 가정하였기 때문에 표준화잔차의 약 95%는 -2에서 +2 사이에 그리고 99.7%는 -3에서 +3 사이에 놓일 것이다.

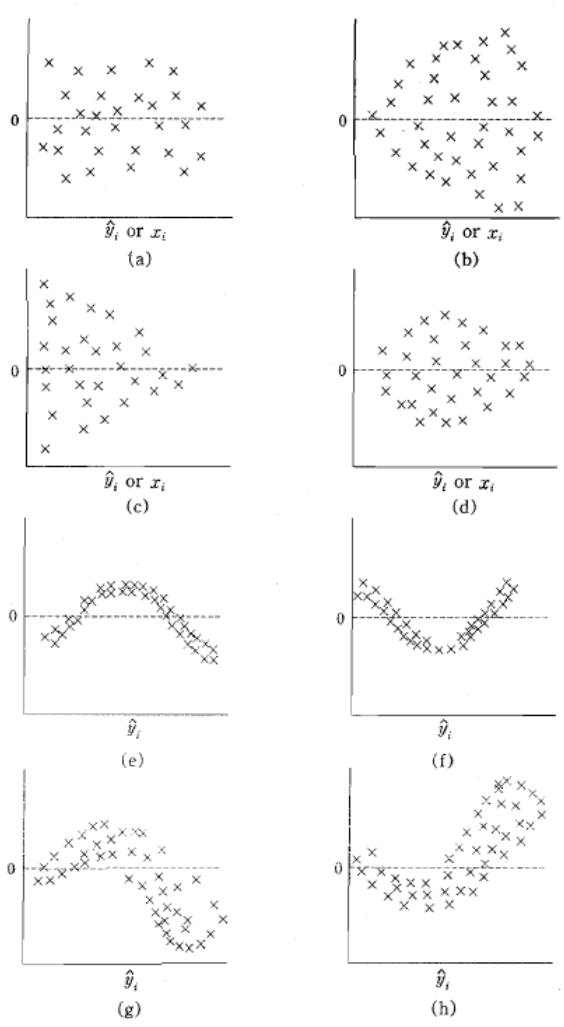
예를 들어 1,000개의 표준화잔차 중 -3에서 +3 밖에 놓이는 잔차의 수가 3개보다 훨씬 많은 경우 우리는 일단 모형의 적합성을 의심하게 된다. 그리고 모형의 부적합도 잔차도의 모양에서 판단이 가능하다.



모형은 몸에 맞는가?

이와 같이 표준화잔차를 이용하면 표준화된 기준을 제시할 수 있기 때문에 표준화되지 않은 잔차보다 더 유용하게 쓰인다. [그림 12.9]에 잔차산점도에 나타날 수 있는 일반적인 형태를 제시하였다.

[그림 12.9]에서 첫 번째 그림 (a)는 회귀모형이 제대로 적합되었을 때 보여지는 산점도이다. 그림 (b)나 (c)를 보자. $\hat{y}$ 이 증가함에 따라 y 축의 값, 잔차가 (b)에서는 증가, 그리고 (c)에서는 감소함을 보여주고 있다. 이와 같은 메가폰 모양의 산점도는 분산이 일정하지 않음을 보여주고 있는 것이다. (e)나 (f)에서는 $\hat{y}$ 가 증가함에 따라 y축의 값, 잔차가 곡선의 형태를 보여주는데 이는 회귀모형의 비선형성(non-linearity)을 보여주고 있는 것이다. (g)와 (h)를 보면 이 두 가지 즉, 일정치 않는 분산과 비선형성을 함께 나타내고 있다.



[그림 12.9] 잔차도 (r_i vs $\hat{y}_i$)의 여러가지 형태

예제 12.5 **잔차분석을 해야 회귀분석은 마무리가 된다.**

아스파라거스

[표 12.11]은 1987년 5월 6일부터 7월 2일까지 미국 보스턴지역의 아스파라거스 한 묶음의 가격과 아스파라거스의 질을 결정하여 주는 3가지 설명변수, GR, NS, DS값을 기록한 자료이다.

GR은 아스파라거스의 녹색부분을 인치로 잰 변수, NS는 묶음 안에 있는 줄기의 수를 기록한 변수, 그리고 DS는 줄기의 실제 지름을 재어 묶음에 있는 줄기의 4분위변동계수(Quartile coefficient of dispersion), $(Q_3 - Q_1)/2$의 값을 기록한 변수를 의미한다. 한 묶음의 무게는 보통 18온스로 기록되어 있어 줄기의 수가 많으면 아스파라거스 줄기의 지름은 작아질 것이다. 즉, NS는 줄기의 평균크기를 지칭하고 DS는 줄기의 변동을 지정한 변수이다.

이러한 자료는 회귀모형을 이용하면 아스파라거스 농가에서는 제품의 질을 결정하는 생산전략 중 가격 예측의 의사결정에 사용할 수 있을 것이다. <아스파라거스.xls>

GR	NS	DS	Price	GR	NS	DS	Price	GR	NS	DS	Price
600	45	33	45	625	12	7	121	600	14	8	100
300	30	25	55	500	11	8	112	950	22	20	126
550	12	8	97	575	12	7	82	650	14	17	126
600	15	9	100	850	14	17	121	550	12	8	91
500	18	20	82	600	12	9	90	600	12	8	93
525	36	25	64	350	28	14	46	625	15	17	100
600	12	14	93	575	18	9	82	550	19	20	79
650	12	14	105	550	12	8	98	650	18	27	100
550	19	20	69	900	26	25	146	500	24	20	81
500	10	14	96	700	15	27	106	475	12	7	97

자료원: Berndt(1991)

GR, DS는 100×인치, 가격은 100×$의 단위임.

[표 12.11] 아스파라거스 자료

[표 12.12]와 같은 다중선형회귀모형분석의 결과를 보면 R^2 값은 0.773이며 변수 DS에 대한 유의성이 상대적으로 다른 두 변수보다는 떨어짐을 알 수 있다. DS는 가격에 그렇게 큰 영향을 미치지 않는 변수임을 알 수 있다. [그림 12.10]처럼 DS변수를 제거하기 전에 다중선형회귀모형에 대한 진단을 실시하였다.

Regression Analysis: Price versus GR, NS, DS

The regression equation is
Price = 47.1 + 0.118 GR - 1.42 NS + 0.096 DS

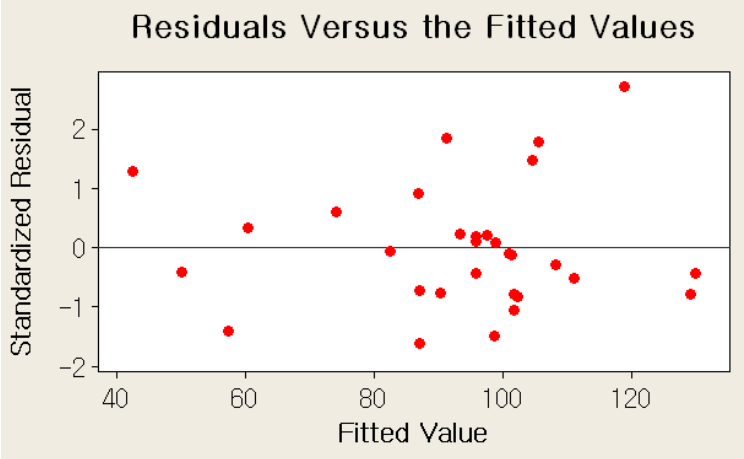
Predictor	Coef	SE Coef	T	P
Constant	47.10	11.60	4.06	0.000
GR	0.11824	0.01759	6.72	0.000
NS	-1.4184	0.4181	-3.39	0.002
DS	0.0957	0.4626	0.21	0.838

S = 11.6722 R-Sq = 77.3% R-Sq(adj) = 74.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	12033.1	4011.0	29.44	0.000
Residual Error	26	3542.2	136.2		
Total	29	15575.4			

[표 12.12] 아스파라거스 자료 회귀분석 결과



[그림 12.10] 잔차산점도 (DS변수 제거 전)

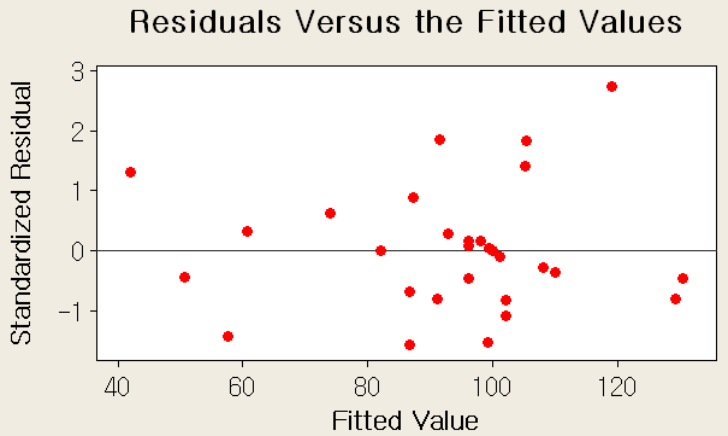
- 잔차산점도를 보면 0을 중심으로 모든 표준화된 잔차가 -3에서 +3 범위 안에서 골고루 퍼져 있음을 알 수 있다. 특별한 패턴이 없는 관계로 앞에서 언급한 세 가지 가정인 등분산성, 선형성, 정규성을 만족시켰다고 볼 수 있다.

또한 [표 12.13]처럼 DS 변수 제거 전 모형을 DS 변수를 제거한 모형과 비교하여도 회귀계수의 크기가 그렇게 변하지 않아 최종모형으로는 DS변수를 제거한 모형을 선택하여도 좋다.

```
The regression equation is
Price = 46.6 + 0.120 GR - 1.35 NS

Predictor      Coef  SE Coef      T      P
Constant      46.58    11.12   4.19  0.000
GR          0.11962  0.01599   7.48  0.000
NS          -1.3515   0.2602  -5.19  0.000

S = 11.4634   R-Sq = 77.2%   R-Sq(adj) = 75.5%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  12027.3  6013.7  45.76  0.000
Residual Error  27   3548.1   131.4
Total           29  15575.4
```

[표 12.13] DS 변수 제거한 회귀분석 결과물

[그림 12.11]과 같은 최종모형에 대한 잔차 산점도 역시 이상 징후를 발견하지 못했다.



[그림 12.11] 잔차 산점도 (DS변수 제거 후)

- 한 농가에서 평균적으로 5.8인치의 녹색부분을 갖고 있는 줄기의 개수가 20개인 아스파라거스를 출하한다고 가정을 하였을 때, GR 및 NS의 변수만 들어간 예측모형을 사용하여 이 농가가 얻는 가격의 기댓값에 대한 95% 신뢰구간을 구하여 보자. 기댓값에 대한 예측값은 $\hat{y}=88.930$이고 95% 예측 신뢰구간은 (64.99, 112.87)으로 나온다. [표 12.14]의 결과물을 참조하기 바란다.

```
New
Obs    Fit  SE Fit      95% CI            95% PI
  1  88.93    2.18  (84.46, 93.40)  (64.99, 112.87)

Values of Predictors for New Observations

New
Obs   GR   NS
  1  580  20.0
```

[표 12.14] 예측구간

예제 12.6 **다중선형회귀분석은 잔차분석을 통하여 변환이 가능하다.**

흑체리나무의 부피는 회귀분석을 이용하여 구한다.

31개의 쓰러진 흑체리 나무의 표본을 조사하여 나무의 높이, 지표면 4.5피트에서의 나무의 지름, 및 부피를 조사하였다. 이러한 자료를 근거로 성장속도에 따른 나무의 부피를 예측하는데 분석의 목적이 있다. <흑체리나무.xls>

V: 나무의 부피 (단위: 입방미터)

D: 지표면 4.5 피트에서의 나무 지름 (단위: 인치)

H: 나무의 높이 (단위: 피트)

[표 12.15]는 V를 반응변수로 하고 D와 H를 설명변수로 한 회귀모형 분석결과이다.

Regression Analysis: V versus D, H

The regression equation is
V = - 58.0 + 4.71 D + 0.339 H

Predictor	Coef	SE Coef	T	P
Constant	-57.988	8.638	-6.71	0.000
D	4.7082	0.2643	17.82	0.000
H	0.3393	0.1302	2.61	0.014

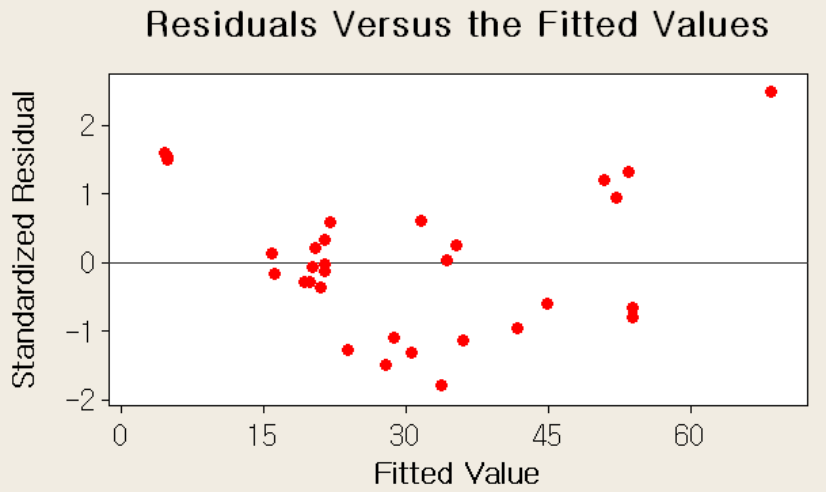
S = 3.88183 R-Sq = 94.8% R-Sq(adj) = 94.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7684.2	3842.1	254.97	0.000
Residual Error	28	421.9	15.1		
Total	30	8106.1			

[표 12.15] 흑체리나무 자료 회귀분석 결과물

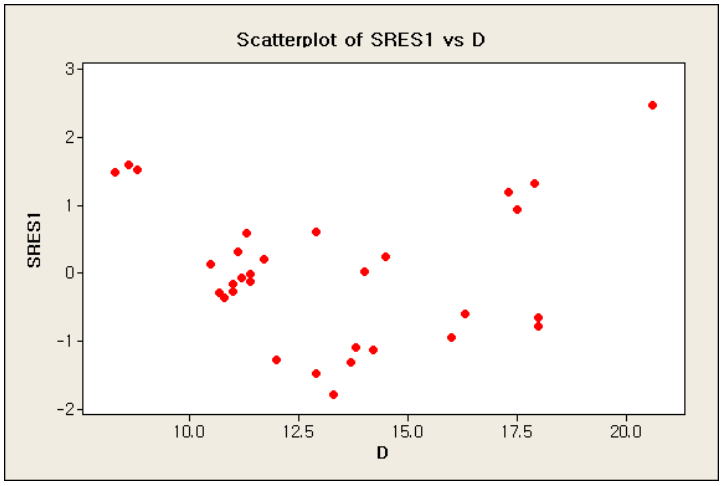
모든 설명변수의 p-값이 매우 작고 결정계수 역시 높아 성공적으로 분석이 된 것 같으나 [그림 12.12]의 잔차 산점도를 분석하여 보면 이는 성급한 결론임을 알 수 있다.



[그림 12.12] 흑체리나무 자료 잔차도

따라서 이와 같은 비선형이 어디에서 유발이 되었는지 파악할 필요가 있다.

부피는 원래 지름의 제곱에 높이를 곱한 식의 형태로 구해져야 한다. 그러므로 의심되는 설명변수는 D변수이다. [그림 12.13]은 설명변수 D를 x-축으로 하여 잔차도를 그려보았다. 짐작하였듯이 변수 D가 비선형성을 유발하는 변수임이 밝혀졌다. 따라서 변수 D는 직접적으로 모형에 진입시키는 것보다는 제곱을 취한 다음 모형을 개발하는 것이 순서일 것이다.



[그림 12.13] 비선형의 잔차산점도

[표 12.16]은 변수 D를 제곱한 다음, 모형에 포함시킨 후의 회귀분석 결과이다.

Regression Analysis: V versus D^2, H

The regression equation is
V = - 27.5 + 0.168 D^2 + 0.349 H

Predictor	Coef	SE Coef	T	P
Constant	-27.512	6.558	-4.20	0.000
D^2	0.168458	0.006679	25.22	0.000
H	0.34881	0.09315	3.74	0.001

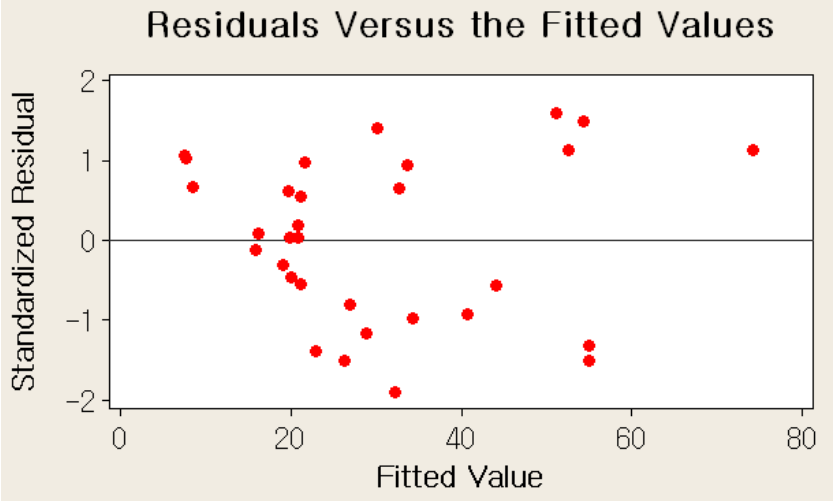
S = 2.79946 R-Sq = 97.3% R-Sq(adj) = 97.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7886.6	3943.3	503.17	0.000
Residual Error	28	219.4	7.8		
Total	30	8106.1			

[표 12.16] D 변수 제곱후의 회귀분석 결과물

이전 모형에 비해 결정계수 값이 상승하였으며 $\sqrt{MSE}$는 감소한 사실을 확인할 수 있다. 역시 [그림 12.14]의 잔차도 역시 이전 그림에 비해서는 가정에 충실한 결과를 가져다주었다.



[그림 12.14] D 변수 제곱 후의 잔차산점도

- 지금까지의 회귀분석 내용 이외에도 모형 개발에는 난제가 많이 남아 있다. 변수 사이에는 항상 중복된 내용이 존재하는데 이럴 경우 어떤 식으로 진단하고 처리할 것인가? 그리고 수많은 설명변수 중에서 반응변수를 제일 잘 설명하는 변수들은 어떻게 선별하는 것이 좋은가? 하는 이슈들이다. 이와 같은 이슈들은 이 책이 의도하고 있는 범위를 벗어나는 것 같아 회귀분석은 여기서 진도를 중단하기로 한다. 그러나 회귀분석은 통계분석의 90% 이상을 차지하고 있을 정도로 매우 중요한 기법이다. 따라서 관심 있는 독자들은 다음 단계의 회귀분석 책을 참조하기로 하고 대신 회귀분석을 이용하여 시계열 분석을 간단히 다뤄보기로 하자.

변수 선택 등 회귀분석의 내용은 아직 많이 남아 있다.

12.7 회귀분석을 이용한 시계열분석

회귀분석으로 추세선과 계절적인 요인이 들어간 시계열 자료를 분석하는 기법을 설명하자. 이는 어느 시계열분석 방법론보다 강력한 도구를 제공한다.

예제 12.7 **시계열 자료도 회귀분석으로 분석이 가능하다.**

[표 12.17]은 1998년부터 2002년까지 수집된 모 회사의 분기별 매출액이다. <매출액.xls>

	A	B	C	E
1			Time	Actual
2	Year	Qtr	Period	Sales
3	1998	1	1	$684.2
4		2	2	$584.1
5		3	3	$765.4
6		4	4	$892.3
7	1999	1	5	$885.4
8		2	6	$677.0
9		3	7	$1,006.6
10		4	8	$1,122.1
11	2000	1	9	$1,163.4
12		2	10	$993.2
13		3	11	$1,312.5
14		4	12	$1,545.3
15	2001	1	13	$1,596.2
16		2	14	$1,260.4
17		3	15	$1,735.2
18		4	16	$2,029.7
19	2002	1	17	$2,107.8
20		2	18	$1,650.3
21		3	19	$2,304.4
22		4	20	$2,639.4

[표 12.17] 분기별 매출 자료

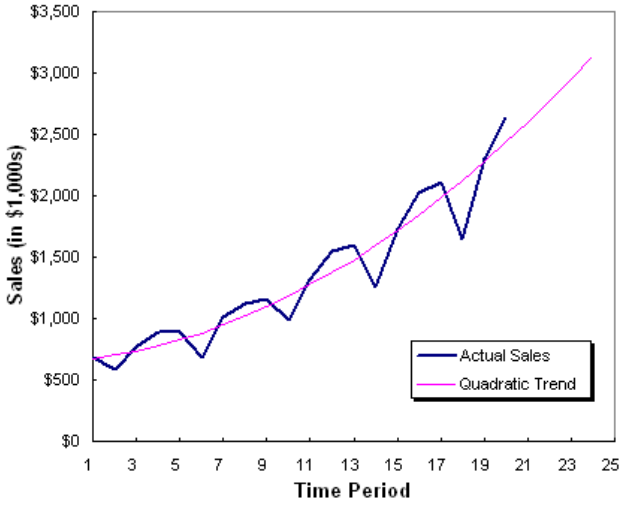
자료를 자세하게 살펴보면 직선의 추세선보다는 곡선의 추세선이 있으므로 시간 t 의 선형모형보다는 이차형식의 모형으로 만들어 추세선을 구하여 보았다. 이는 다음과 같은 모형을 적합 시키는 의미인데

$$y = \beta_o + \beta_1 t + \beta_{11} t^2 \tag{12.29}$$

시간은 설명변수로 열 C에 저장되어 있다. 그리고 열 D에 시간 설명변수 t를 제곱한 다음 회귀모형을 추정하면 다음과 같이 나온다.

$$\hat{y_t} = 653.67 + 16.671t + 3.617t^2$$

이를 이용하여 2003년을 포함한 각 분기별 예측값을 구하여 열 F에 저장한 것이다. 즉 2003년의 예측값은 t에 $t = 21, 22, 23, 24$을 대입하여 구한 값이다.



[그림 12.15] 이차형식으로 본 추세선

[그림 12.15]에서 곡선은 $\hat{y}_t = 653.67 + 16.671t + 3.617t^2$을 적합 시켜 나온 선이다. [표 12.18]은 이를 정리한 표이다.

	A	B	C	D	E	F
1			Time		Actual	Quadratic
2	Year	Qtr	Period	Time^2	Sales	Trend
3	1998	1	1	1	$684.2	$674.0
4		2	2	4	$584.1	$701.5
5		3	3	9	$765.4	$736.2
6		4	4	16	$892.3	$778.2
7	1999	1	5	25	$885.4	$827.4
8		2	6	36	$677.0	$883.9
9		3	7	49	$1,006.6	$947.6
10		4	8	64	$1,122.1	$1,018.5
11	2000	1	9	81	$1,163.4	$1,096.7
12		2	10	100	$993.2	$1,182.1
13		3	11	121	$1,312.5	$1,274.7
14		4	12	144	$1,545.3	$1,374.6
15	2001	1	13	169	$1,596.2	$1,481.6
16		2	14	196	$1,260.4	$1,596.0
17		3	15	225	$1,735.2	$1,717.5
18		4	16	256	$2,029.7	$1,846.3
19	2002	1	17	289	$2,107.8	$1,982.4
20		2	18	324	$1,650.3	$2,125.6
21		3	19	361	$2,304.4	$2,276.1
22		4	20	400	$2,639.4	$2,433.8
23	2003	1	21	441	--	$2,598.8
24		2	22	484	--	$2,771.0
25		3	23	529	--	$2,950.4
26		4	24	576	--	$3,137.1

[표 12.18] 이차모형 추세선

계절적인 변동을 고려하기 위해서 원래의 관측값과 예측값의 비(ratio)를 구하여야 한다. 예를 들어 1998년 1분기는

684.2/674 =102%

로 계산한다. 그런 다음 각 1분기별 평균을 내어 계절인덱스를 만들어야 한다. 예를 들면 1분기 계절인덱스는

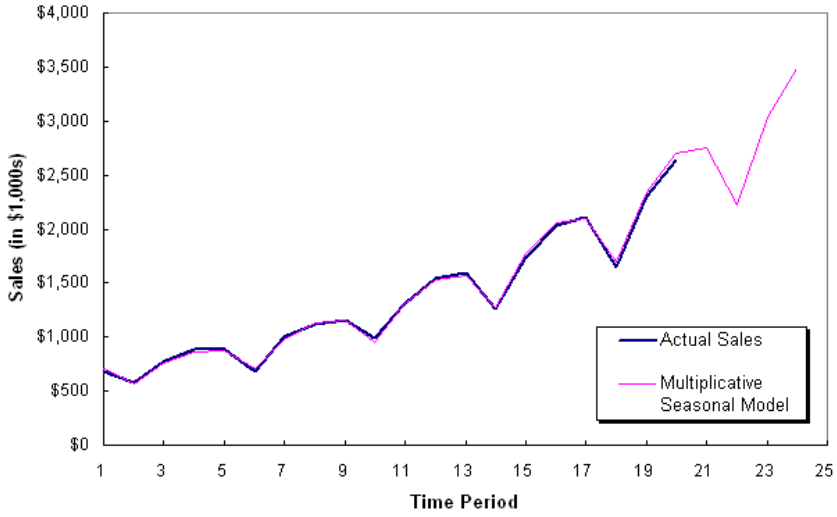
(102+107+106+108+106)/ 5 = 105.7

그런 다음 추세선으로 예측한 내년도 예측값에 이를 곱하여 계절적인 요인이 감안된 예측값을 구한다. [표 12.19]를 참조하라.

	A	B	C	D	E	F	G	H	I	J	K
1			Time		Actual	Quadratic	Actual as a	Seasonal			Seasonal
2	Year	Qtr	Period	Time^2	Sales	Trend	% of Trend	Forecast		Qtr	Index
3	1998	1	1	1	$684.2	$674.0	102%	$712.6		1	105.7%
4		2	2	4	$584.1	$701.5	83%	$561.9		2	80.1%
5		3	3	9	$765.4	$736.2	104%	$758.9		3	103.1%
6		4	4	16	$892.3	$778.2	115%	$864.8		4	111.1%
7	1999	1	5	25	$885.4	$827.4	107%	$874.9			
8		2	6	36	$677.0	$883.9	77%	$708.0			
9		3	7	49	$1,006.6	$947.6	106%	$976.8			
10		4	8	64	$1,122.1	$1,018.5	110%	$1,131.8			
11	2000	1	9	81	$1,163.4	$1,096.7	106%	$1,159.6			
12		2	10	100	$993.2	$1,182.1	84%	$946.8			
13		3	11	121	$1,312.5	$1,274.7	103%	$1,314.0			
14		4	12	144	$1,545.3	$1,374.6	112%	$1,527.5			
15	2001	1	13	169	$1,596.2	$1,481.6	108%	$1,566.6			
16		2	14	196	$1,260.4	$1,596.0	79%	$1,278.4			
17		3	15	225	$1,735.2	$1,717.5	101%	$1,770.5			
18		4	16	256	$2,029.7	$1,846.3	110%	$2,051.7			
19	2002	1	17	289	$2,107.8	$1,982.4	106%	$2,096.0			
20		2	18	324	$1,650.3	$2,125.6	78%	$1,702.6			
21		3	19	361	$2,304.4	$2,276.1	101%	$2,346.3			
22		4	20	400	$2,639.4	$2,433.8	108%	$2,704.6			
23	2003	1	21	441	--	$2,598.8	--	$2,747.8			
24		2	22	484	--	$2,771.0	--	$2,219.6			
25		3	23	529	--	$2,950.4	--	$3,041.4			
26		4	24	576	--	$3,137.1	--	$3,486.1			

[표 12.19] 계절인덱스 계산

[그림 12.16]은 이차형식의 회귀모형을 이용한 예측값과 관측값의 비교 그림이다. 굳이 복잡한 모형을 가지고 분석하지 않아도 되지 않는가?



[그림 12.16] 이차모형을 이용한 예측

학습요약

제 12장에서 회귀분석에 대해 알아보았다. 회귀분석은 응용되는 통계 분야 중에서 90% 이상을 차지할 정도로 많은 응용성을 가지고 있다. 그러나 다중선형회귀모형을 개발하는 것은 생각보다 매우 어려운 작업이다. 컴퓨터의 결과물을 정확하게 읽어낼 수 있는 능력 뿐 아니라 분석자가 목적하고 있는 것이 무엇인지에 따라 모형개발이 달라진다. 모형이 가지고 있는 가정들은 충분히 적합한지 여부를 따져주어야 하기 때문이다.

최소제곱법의 의미와 분산분석을 통해 모형에 대한 가설검정을 실시하는 과정을 보았다. 또한 회귀계수 및 새로운 관측값에 대한 신뢰구간 및 예측구간도 살펴보았다. 모형이 가지고 있는 기본적인 가정, 즉 정규성, 독립성, 그리고 등분산성은 잔차분석을 통하여 검토하였다.

마지막으로 회귀분석의 추세선을 이용한 시계열 분석도 가능하다는 사실도 언급하였다.

12장 연습문제

Country	GNPCapita	PopGrowth	Calorie	LifeExp	Fertility
Antigua_&_Barbud	4595	0.5%	2222	74	1.9
Argentina	2369	1.4%	3118	71	2.8
Bahamas	11514	1.9%	2678	69	2.2
Bangladesh	199	2.6%	1925	52	4.8
Belgium	15444	0.1%	3942	76	1.6
Belieze	1974	2.8%	2649	68	4.7
Benin	362	3.2%	2145	51	6.3
Boliva	619	2.8%	2086	54	5.9
Botswana	2042	3.4%	2269	68	4.7
Brazil	2682	2.2%	2709	66	3.2
BurKina_Faso	328	2.6%	2061	48	6.5
Cameroon	941	3.2%	2161	57	6.5
Canada	20449	0.9%	3447	77	1.7
Central_African	393	2.7%	1980	51	5.8
China	367	1.5%	2632	70	2.5
Columbia	1242	2.0%	2561	69	2.9
Comoros	478	3.7%	2046	55	6.8
Congo	1008	3.4%	2512	54	6.6
Costa_Rica	1907	2.4%	2782	75	3
Czechoslovakia	3139	0.3%	3564	72	2
Dominica	1951	1.2%	2877	75	2.8
Dominican_Rep.	819	2.3%	2357	67	3.5
Egypt	603	2.5%	3213	60	4.1
El_Salvador	1097	1.5%	2415	64	4.6
Ethiopia	118	3.1%	1658	48	7.5
Fiji	1770	1.7%	2763	68	3
France	19481	0.4%	3310	77	1.8
Gambia	262	3.3%	2360	44	6.5
Germany	18256	0.0%	3594	75	1.5
Greece	5996	0.4%	3699	77	1.5
Guinea	482	2.5%	2042	43	6.5
Guinea-Bissau	179	1.9%	2690	40	6
Guyana	367	0.5%	2373	64	2.8
Iran	5167	3.5%	3100	63	6
Israel	10972	1.7%	3138	76	2.8
Jamaica	1509	1.2%	2572	73	2.4
Jordan	1244	3.6%	2907	68	6.3
Kiribati	771	1.9%	2952	55	4.2
Korea_Rep.	5454	1.1%	2878	70	1.8
Lao_PDR	209	2.8%	2637	50	6.7
Lesotho	470	2.7%	2307	57	5.6
Luxembourg	28770	0.4%	3942	75	1.5
Madagascar	233	2.9%	2101	51	6.5
Malawi	195	3.4%	2009	48	7.6
Malaysia	2339	2.6%	2686	70	3.6
Mali	271	2.5%	2181	48	7
Malta	6635	-0.5%	3318	73	2.1

Mauritania	501	2.4%	2528	47	6.8
Mexico	2490	2.0%	3135	70	3.3
Morocco	948	2.7%	2820	62	4.7
Nepal	172	2.6%	2078	52	5.7
Netherlands	17333	0.5%	3354	77	1.5
New_Zealand	12683	0.8%	3459	75	2
Niger	309	3.4%	2340	46	7.1
Panama	1825	2.1%	2468	73	2.8
Papua_New_Guinea	861	2.5%	2236	55	5.1
Paraguay	1112	3.2%	2816	67	4.6
Portugal	4887	0.6%	3382	75	1.6
Romania	1636	0.4%	3357	71	2.1
Sao_Tome	382	2.8%	2657	66	5.1
Senegal	708	3.0%	1989	49	6.5
Seychelles	4676	0.7%	2146	71	2.8
Sierra_Leone	237	2.4%	1806	42	6.5
Solomon_Islands	577	3.5%	2115	65	6.5
Sri_Lanka	469	1.5%	2319	71	2.4
St._Kitts_&_Nevi	3325	-1.2%	2801	70	2.6
St._Vincent	1614	1.0%	2818	70	2.6
Suriname	3054	2.5%	2809	68	3.4
Sweden	23678	0.3%	3007	78	2
Switzerland	32786	0.5%	3547	78	1.6
Tanzania	113	3.1%	2151	50	6.6
Togo	405	3.5%	2133	54	6.6
Tonga	1010	0.5%	2980	67	4
Trinidad	3475	1.7%	2960	72	2.8
Uganda	220	3.2%	2013	49	7.3
United_Kingdom	16074	0.2%	3252	76	1.8
Venezuela	2562	2.8%	2547	70	3.5
Western_Somoa	733	0.6%	2477	66	4.7
Zambia	418	3.7%	2026	54	6.7
Zimbabwe	644	3.4%	2232	64	4.9

12.1 이 자료는 세계 각국에 대해 수집한 경제지표 자료이다.

* country : 나라이름 * GNPCapita : GNP per capita(달러)
* POPGrowth : 1980-1990년까지 퍼센티지 년 인구변동
* Calorie : daily per capita calories
* LifeExp : 신생아를 기준으로 한 기대수명
* Fertility : 여자 일인당 출생율

다음 세 가지 회귀분석모형을 수립하라.

(1) 반응변수 :LefeExp, 설명변수 : Calorie, Fertility
(2) 반응변수 :LefeExp, 설명변수 : GNPCapita, PopGrowth
(3) 반응변수 :GNPCapita, 설명변수 : PopGrowth, Calorie, Fertility

12.2 다음은 미국의 Variety 란 잡지가 발표한 자료로 1980년을 기준(100)으로 하여 1987년부터 1996년까지 박스오피스 매출을 기록한 자료이다. 따라서 단위는 달러기록이 아니다. 제 12장에서 배운 시계열 분석 자료를 이용하여 1997년 Box Office 매출을 예측하라.

월	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
1	95	104	101	88	132	125	111	127	119	147
2	94	100	96	110	109	118	123	129	147	146
3	98	99	82	129	101	121	121	132	164	133
4	96	88	84	113	111	140	139	108	135	148
5	95	89	85	114	140	141	119	115	124	141
6	115	108	124	169	179	201	156	149	168	191
7	107	109	134	131	145	152	15	155	159	178
8	104	101	109	139	140	138	136	129	137	156
9	96	106	121	120	120	137	105	117	149	119
10	112	102	111	115	129	138	132	166	159	138
11	98	78	101	116	118	144	123	152	175	175
12	102	111	112	128	139	148	164	173	195	188

이 자료는 추세선과 계절적인 변동요인이 있는 자료이다. 적절한 회귀분석을 이용하여 구하라.

12.3 (가변수의 회귀분석) 어느 회사의 인사관리담당자는 졸업생의 초봉을 결정하는 요인으로 대학교의 평점과 전공이라고 믿고 있다. 초봉을 설명하는 적절한 모형을 설정하여 그 모형의 타당성에 대해서 조사하고 설명해보라. 전공이 통계학이고 평점이 3.0인 사람이 기대하는 초봉은 얼마인가?

평점	전공	초봉
2.95	경영학	21,500
3.20	경영학	23,000
3.40	경영학	24,100
2.85	통계학	24,000
3.10	통계학	27,000
2.85	통계학	27,800
2.75	경영학	20,500
3.10	경영학	22,200
3.15	경영학	21,800

쉬어가기

지수란 무엇인가?

서로 측정단위가 다른 통계들을 현실적으로 어떻게 정리하는지를 알아 볼 필요가 있는데 그 중 하나가 지수이다. 정부 또는 민간기관에서는『태풍 피해로 과일 및 야채의 가격이 올라 물가가 7% 상승했다』든가 『수출 호조로 생산이 12% 증가하여 경기가 상승하고 있다』와 같은 보도를 하고 있다. 이와 같이 보도되는 경제지표의 형태에는 지수가 많이 쓰인다. 지수는 통계를 가공하는 것인데, 작성하는 목적은 비교가 용이(容易)하고, 서로 다른 단위로 측정한 것을 동일한 단위로 집계하기 위한 것이다. 간략하게 정의하면 같은 종류의 통계수치에 대한 대소 관계에 대한 형태를 비율의 형태로 표시한 것이다. 통상 비교의 기준이 되는 시점의 수치를 100으로 하여 산출하고 있다. 지수는 장소적 비교나 시간적 비교에도 사용할 수 있다.

예를 들어 자동차의 기준년도 생산대수가 200만대이고 비교년에는 300만대를 생산하였다면 자동차의 생산지수를 (300/200)×100=150.0으로 나타낼 수 있다. 이 수치의 의미는 비교년도의 자동차 생산이 기준년도에 비해 50.0%가 증가하였다는 것이다. 이러한 지수를 계산하는 데는 기준시점, 가중치(weight) 그리고 산식이 필요하게 되는데 이를 지수의 세 요소라고 말한다.

(1) 지수의 기준시점

지수를 작성하기 위해서는 기준이 되는 일정기간의 시점(기준시)이 필요하다. 지수의 기준시는 경제적으로 안정된 시점이 바람직하다고 할 수 있다. 고정기준지수에서는 그 지수에 대한 전 기간에 걸쳐 동일한 가중치가 사용되기 때문에 특수한 시점의 가중치는 총합지수에 괴리된 영향을 주게 된다. 따라서 한 시점을 기준시로 하는 것이 아니라 수년간의 평균을 기준값(기준시)으로 함으로써 괴리가 적은 지수를 만들 수도 있다.

(2) 가중치

총합지수를 계산하는 데에는 구성품목에 대해 개별적으로 계산된 지수를 단순평균하는 방법이 있을 수 있는데 이렇게 구해진 단순평균지수는 구성품목의 중요도를 무시한 산술평균에 의한 방법이 되기 때문에 계산된 총합지수는 적절한 수준을 나타낸 것이라고 말할 수 없다. 이에 대해 각 품목의 중요도(비중) 차이를 지수 작성에 반영한 지수를 가중평균지수라고 말하는데 이 때 고려된 각 개별지수의 중요도가 바로 가중치이다. 예를 들어 물가지수의 경우는 상품거래금액 또는 소비지출금액을, 생산지수의 경우는 부가가치액 또는 생산액을 가중치로 하고 있다.

(3) 지수의 산식

지수(가중평균지수)의 산식에는 라스파이레스 (Laspeyres)식, 파쉐(Paasche)식, 피셔 (Fisher)식 등이 있다.

라스파이레스식은 1864년 독일의 통계학자인 라스파이레스 가 창안한 산식으로 기준시 (T=0)와 같은 생산금액의 상품을 비교시 (T=t)의 상품을 생산했을 때의 생산수량을 기준시의 생산수량으로 나누어 그 변화의 정도로써 생산수준을 측정하는 것이다. 기준시 생산금액이 가중치로 고정되게 되는데 이는 상품 생산금액이 크게 변화하지 않는다는 것을 전제로 하고 있는 것이다.

$$\text{라스파이레스식(L식)}\ Q_{t,i}^{L} = \frac{\sum P_{0,i} Q_{t,i}}{\sum P_{0,i} Q_{0,i}}$$

여기서, $Q_{0,i}$: 기준시 생산수량, $Q_{0,i}$: 비교시 생산수량, $P_{0,i}$: 기준시 생산금액

파쉐식은 독일의 통계학자인 파쉐 (Herman Paasche)에 의해 창안된 산식으로 라스파이레스식과는 반대로 비교시의 생산금액을 가중치로 사용된다. 이 경우 가중치가 매년 바뀌어 현실의 생산동향을 정확히 반영할 수 있는 것처럼 보이나 비교시의 대상품목과 가중치를 매번 조사해야 하는 번거로움이 따르기 때문에 실제로는 실용성이 거의 없다고 볼 수 있다. 따라서 파쉐식은 소비자기호 및 산업구조의 변화 등으로 제품 생산금액의 구성이 시간의 변화에 따라 크게 변화되고 있을 때만 사용된다.

$$\text{파쉐식(P식)}\ Q_{t,i}^{P} = \frac{\sum P_{t,i} Q_{t,i}}{\sum P_{t,i} Q_{0,i}}$$

여기서, $Q_{0,i}$: 기준시 생산수량, $Q_{0,i}$: 비교시 생산수량, $P_{t,i}$: 비교시 생산금액

피셔식은 미국의 경제학자 피셔 (Irving Fisher)가 제안한 산식으로 라스파이레스식과 파쉐식을 기하평균한 것으로서 이론적으로는 가장 완벽한 지수이며 이상적 지수 (ideal index)라고도 부른다.

$$\text{피셔식 (F식)}\quad Q_{t,i}^{F} = \sqrt{Q_{t,i}^{L} \times Q_{t,i}^{P}}$$

(4) 연환지수

5년간 각 품목의 생산이 어떠한 변동을 해왔는가에 대해 <표 1>을 살펴보면, t=-1년과 t=3년을 비교하면 무연탄은 감소하고, 소주 등 4가지 품목은 증가했다. 소주 등 증가한 4가지 품목은 4년간에 있어 증가했지만, 각 연도의 움직임이 동일하지 않다. 무연탄의 감소추세도 연도별로 일정하지 않다.

<표 1> 품목별 생산실적 추이

년도	무연탄(M/T)	소주(kℓ)	카페트(천㎡)	휴대폰(대)	승용차(대)
t=-1	4,137,315	1,070,786	12,736	47,962,065	1,465,666
기준년(t=0)	4,151,194	945,544	13,613	66,910,145	1,561,510
t=1	3,814,161	1,080,870	12,546	89,982,724	1,457,602
t=2	3,332,000	1,067,206	14,761	107,031,100	1,520,374
t=3	3,312,319	1,125,568	13,191	133,907,276	1,681,531

품목별로 움직임을 살펴보기 위하여 <표 2>와 같이 전년에 대한 비율(전년비)을 계산하자. 무연탄은 t=2년에 증가한 것을 제외하고 다른 연도에는 감소하는 것으로 나타나고 있어 전반적으로 감소 추세를 보이고 있다. 소주와 카펫은 매년 증감을 반복하고 있으며, 휴대폰은 지속적으로 증가하는 추이를 보이고 있다. 승용차는 t=0년 증가하고 t=1년에 감소한 후 t=2년부터 2년 연속 증가하는 추이를 보이고 있다. 이와 같이 전년비를 구하는 것도 지수의 한 종류라 할 수 있다. 즉 전년비나 전월비와 같이 관찰하려하는 시점과 그 직전시점의 통계수치의 비율을 취해 시계열로 나타내는 것을 연환지수(연환비율)라고 한다.

<표 2> 품목별 전년 비(연환지수)의 추이

년도	무연탄(M/T)	소주(kℓ)	카펫(천㎡)	휴대폰(대)	승용차(대)
t=-1					
기준년(t=0)	100.3	88.3	106.9	139.5	106.5
t=1	91.9	114.3	92.2	134.5	93.3
t=2	87.4	98.7	117.7	118.9	104.3
t=3	99.4	105.5	89.4	125.1	110.6

연환지수는 관찰시점 직전의 시점을 기준으로 해서 작성되는데 일반적인 지수는 관찰하려하는 기간의 모든 시점에 관해 동일한 시점의 수치를 기준으로 해서 작성된다. 이것을 고정기준지수라고 한다. 기준으로 되는 시점을 기준시, 비교되는 시점을 비교시라고 한다. <표 1>에 대해 t=0년을 기준으로 한 품목별 비율을 <표 3>과 같이 계산해 보자. 이것을 품목별 개별지수라 하고 100을 곱해서 나타낸다.

<표 3> 품목별 생산지수

년도	무연탄(M/T)	소주(kℓ)	카페트(천㎡)	휴대폰(천대)	승용차(대)
t=-1	99.7	113.2	93.6	71.6	93.9
기준년(t=0)	100.0	100.0	100.0	100.0	100.0
t=1	91.9	114.3	92.2	134.5	93.3
t=2	80.3	112.9	108.4	160.0	97.4
t=3	79.8	119.0	96.9	200.1	107.7

(5) 금액지수

앞에서 품목별 생산의 동향은 알 수 있었는데, 전체의 생산활동이 어떻게 변화했는가에 대해 생각해 보자. 이를 위해서는 각 품목의 생산실적을 어떠한 방법으로든 집계해서 전체의 변동을 나타내는 수치로 작성할 필요가 있다. 총지수를 어떠한 목적으로 어떻게 작성하는가가 제일 중요한 문제이며, 개별지수는 그 목적 및 합산의 방법에 대응해서 선택되어지게 된다.

<표 1>를 이용하여 총지수 작성과정을 살펴보자. 이들 5품목은 각각 측정단위가 다르기 때문에 단순하게 합계하는 것은 의미가 없다. 단위가 다른 통계수치를 집계하는 데는 공통적인 측정단위로 고쳐야만 한다. 그 가장 일반적인 방식이 생산금액에 의한 방법이다. t=0년과 t=2년의 각 품목의 단위당 가격이 <표 4>와 같다고 하자.

<표 4> 품목별 단가추이

	무연탄 (천원/T)	소주 (천원/kℓ)	카페트 (천원/천㎡)	휴대폰 (천원/대)	승용차 (천원/대)
t=0	69	863	12,377	177	8,319
t=2	70	927	13,021	209	9,540

각 연도의 생산금액을 구해 그 비율을 취하면 금액지수가 된다. 이 지수의 변동에는 생산의 양적변동과 가격변동이 모두 포함되어 있다.

$$\frac{t=2\text{년무연탄생산개수}\times\text{단가}(t=2\text{년무연탄의생산금액})+\cdots}{t=0\text{년무연탄생산개수}\times\text{단가}(t=0\text{년무연탄의생산금액})+\cdots}\times 100=\text{금액지수}$$

$$\frac{3,332,000\times 70+1,067,206\times 927+14,761\times 13,021+107,031,100\times 209+1,520,374\times 9,540}{4,151,194\times 69+\quad 945,544\times 863+13,613\times 12,377+66,910,145\times 177+1,561,510\times 8,319}\times 100$$

$$=\frac{38,263,771,255}{26,192,198,657}\times 100=146.4$$

(6) 수량지수

우리가 실질적인 경제활동을 관찰하려면 명목금액에서 가격의 영향을 제거해서 양적인 변동만을 보아야 한다. 개별 품목의 경우라면 문제는 간단하다. 무연탄의 t=0년과 t=2년의 생산금액에 대한 비율은

$$\frac{3,332,000(M/t)\times 70\text{천원}}{4,151,194(M/t)\times 69\text{천원}}\times 100=\frac{2,347\text{억원}}{2,854\text{억원}}\times 100=82.2$$

로 되는데 양적인 변동만을 보고 싶을 경우에는 생산수량 그 자체의 움직임을 보면 되기 때문에 수량비율은

$$\frac{3,332,000(M/t)}{4,151,194(M/t)}\times 100=80.2$$

이다. 즉, <표 3>의 품목별 생산지수와 같다. 그러나 총지수에 대한 가격변동을 제거하려면 일정한 규칙이 필요하다. 가격변동의 영향이 없는, 실질적인 수량만의 변동을 표현하는 지수를 수량지수라고 한다. 총합한 수량지수를 작성하는 가장 간단한 방법은 원래의 품목별 실적계열에 대한 단위가 다른 것을 무시하고 합계해서 그 비율을 계산하는 것이다. <표 1>를 기초로 이 방식에 의해 t=0년을 기준시로 하여 t=2년의 지수를 작성해 보자.

$$\frac{3,332,000+1,067,206+14,761+107,031,100+1,520,374}{4,151,194+\quad 945,544+13,613+66,910,145+1,561,510}\times 100$$

$$=\frac{112,965,441}{73,582,006}\times 100=153.5$$

이 지수에 의하면 t=0년에 비해 t=2년은 품목에 따라서는 가격의 상승이나 하락이 있었지만 전체적인 생산량의 수준은 153.5% 증가한 것으로 된다. 그러나 이 방법은 원래부터 의미 없는 수치를 합계한 것이기 때문에 적절한 지수라고는 말할 수 없다. 만약 휴대폰의 단위를 절상해서 각각 66,910,145대에서 66,910천대로 107,031,1000대에서 107,031천대로 변경해서 계산해 보자.

$$\frac{3,332,000+1,067,206+14,761+107,031+1,520,374}{4,151,194+\quad 945,544+13,613+66,910+1,561,510}\times 100$$

$$=\frac{6,041,372}{6,738,771}\times 100=89.7$$

이 결과는 t=2년의 생산활동이 t=0년에 비해 89.7%의 수준으로 떨어져 먼저 계산한 결과와 모순된다. 휴대폰의 생산활동이 전혀 바뀌지 않음에도 불구하고 단지 단위의 변경에 의해 총합지수가 크게 변하게 되면 이 총합방법은 선택할 수가 없다. 카펫에 대해서도 천㎡에서 백만㎡로 절상하거나 ㎡로 절하해도 지수가 바뀌어 불합리하다.

다음의 총합방법은 앞서 개별지수를 계산해서 <표 3>과 같은 형태로 만든 후 단순산술평균을 하는 방법이다.

$$\frac{80.3+112.9+108.4+160.0+\ 97.4}{100.0+100.0+100.0+100.0+100.0}\times 100=\frac{558.9}{500.0}\times 100=111.8$$

이 방식은 단위가 대에서 천대로 변경됨에 의해 수치가 변화하지는 않는다. 그러나 이 방식에도 문제가 있다. 이 방식에 의하면 기준시에 대한 각 품목의 생산량의 비율 즉, 무연탄 415만1천M/t, 소주 94만5천㎘, 카펫 1,361만㎡, 휴대폰 6,691만대, 승용차 156만대 각각이 전체에 주는 영향도가 같았을 때를 가정한 것에 지나지 않는다. 예를 들면 기준시에 비해 무연탄이 415,119M/t 늘고, 승용차가 156,151대 늘어나서 두 품목의 증가율이 각각 10%씩 상승하였다면 총지수의 상승률에 동일하게 영향을 준다는 것이다. 그러나 무연탄 415,119M/t과 승용차

156,151대가 총합(총지수)의 상승률에 같은 영향을 미친다는 가정은 객관적인 근거로 미흡하다.

기준시의 품목별 가격을 알 수 있기 때문에 이것을 비교시의 품목별 생산실적에 곱해 금액의 형태로 고치고 이것과 기준시의 금액과의 비율을 취하면 가격변동의 영향을 받지 않는 총지수의 계산이 가능한 것은 아닐까? 이것은 매우 당연한 문제의식이다. **그리고 실제 대부분의 지수는 이 방법에 의해 작성되고 있다.**

다시 <표 1> 및 <표 4>를 이용하여 이것을 계산해 보자.

$$\frac{3,332,000\times 69+1,067,206\times 864+14,761\times 12,377+107,031,100\times 177+1,520,374\times 8,319}{4,151,194\times 69+\ \ 945,544\times 864+13,613\times 12,377+66,910,145\times 177+1,561,510\times 8,319}\times 100$$

$$=\frac{32,966,652,356}{26,192,198,657}\times 100=126.2$$

한편, 비교시의 품목별 가격이 기준 시에도 같다라고 생각되면

$$\frac{3,332,000\times 70+1,067,206\times 927+14,761\times 13,021+107,031,100\times 209+1,520,374\times 9,540}{4,151,194\times 70+\ \ 945,544\times 927+13,613\times 13,021+66,910,145\times 209+1,561,510\times 9,540}\times 100$$

$$=\frac{38,263,771,255}{30,211,133,546}\times 100=126.7$$

와 같이 된다.

제 13 장

범주형 자료를 분석하여 보자.

Table 3. Disease free survival for patients and treatments characteristics (univariate analysis)

Factors		Years overall survival(%)			P value
		1YR	3YR	5YR	
Child classification					
A	476	63	42	35	0.901
B	21	69	40	30	
ICGR15					
<20	423	63	41	36	0.362
≥20	21	69	40	30	
AFP					
<20	146	78	53	45	0.000*
≥20	321	56	35	28	
HbsAg					
(-)	111	71	49	40	0.018
(+)	376	60	38	31	
Anti-HCV					
(-)	351	64	43	36	0.517
(+)	41	60	31	27	
Op time(min)					
<300	349	65	43	37	0.018
≥300	144	58	34	27	
Transfusion					
(-)	362	67	45	38	0.000*
(+)	131	53	30	24	
Resection margine (cm)					
<1	191	55	36	29	0.088
≥1	291	68	44	37	

*Statistically significant in multivariable analysis.

차 례

학습목표

주어진 자료가 범주형으로 수집된 경우는 자료를 표현하는 방법이나 자료분석하는 방법이 수치자료인 경우와는 다르다. 13장에서는 자료가 범주형자료인 경우 어떻게 자료를 표현하고 자료분석을 행하는 지에 대하여 배워보자.

주어진 자료가 범주형자료인 경우 범주와 그 범주에 대응되는 도수로 주어짐으로 수치형자료일 때 적용하는 통계분석방법을 적용할 수가 없다. 우리는 범주형자료를 표현하는 방법이나 자료분석하는 방법을 따로 배워야 할 필요가 있다.

13.1 적합도검정

범주형자료에서 범주의 개수가 k개이고 $i(i=1,2,...,k)$범주에 대응되는 관측도수가 O_i라 할 때 우리는 다음과 같은 검정(적합도검정, goodness-of-fit test)을 시행하기 원한다(유의수준: α%).

귀무가설 H_0: 범주형자료에서 각 범주가 특정 비율을 따라간다.
대립가설 H_1: 범주형자료에서 각 범주가 특정 비율을 따라가지 않는다.

검정규칙은 다음과 같다.

검정통계량 $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$이 $\chi^2(k-1;\alpha)$보다 크면 H_0를 기각한다.

여기서, k는 범주의 개수, O_i는 $i(i=1,2,...,k)$범주에 대응되는 관측도수이고, E_i는 기대도수이다. $\sum_{i=1}^{k}$은 k개의 모든 범주에 대하여 더한다는 것이다. $\chi^2(k-1;\alpha)$는 자유도가 $k-1$인 카이제곱분포에서의 상위 α% 백분위수이다. 귀무가설이 맞다는 가정 하에서는 표본의 크기가 클 때 검정통계량 $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$이 자유도가 $k-1$인 카이제곱분포를 이룬다.

예제 13.1 멘델의 법칙에서 두 쌍(우성인자: A와 B, 열성인자: a와 b)의 대립형질 유전은 잡종 제 2세대로 내려가면 AB: Ab: aB: ab=9: 3: 3: 1의 비율을 이룬다고 알려져 있다. 이를 확인하기 위하여 실험을 실시한 후 잡종 제 2세대 556개에 대하여 조사하여 보니 다음 [표 13.1]과 같았다.

타입	AB	Ab	aB	ab	합계
관측도수	315	101	108	32	556

[표 13.1] 잡종 제 2세대 결과

AB: Ab: aB: ab=9: 3: 3: 1의 비율을 이룬다고 했으니 우리가 행하고자 하는 검정(유의수준: 5%)은 다음과 같다.

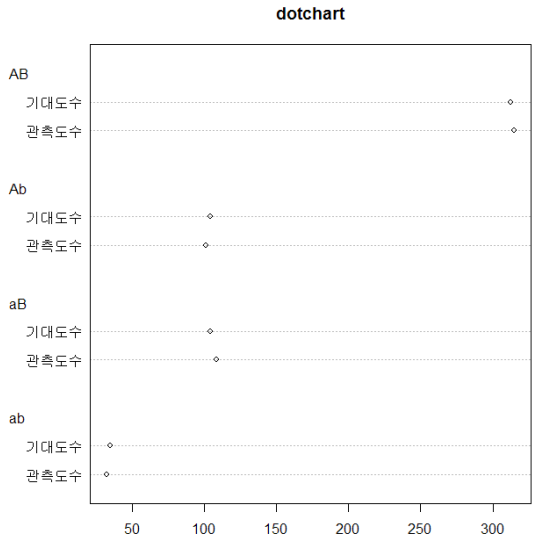
귀무가설 H_0: $p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$

대립가설 H_1: H_0는 아니다.

여기서, $p_i (i = 1,2,3,4)$는 각각 잡종 제 2세대에서 AB, Ab, aB, ab의 모비율을 가리킨다. 각 타입별로 기대도수를 계산하면 다음과 같다.

$$E_1 = 556 \times \frac{9}{16} = 312.75,\ E_2 = 556 \times \frac{3}{16} = 104.25,$$
$$E_3 = 556 \times \frac{3}{16} = 104.25,\ E_4 = 556 \times \frac{1}{16} = 34.75$$

관측도수와 기대도수를 점차트로 표시하여 보면 [그림 13.1]과 같다. 각 범주에서 관측도수와 기대도수와의 차이가 거의 없음을 알 수 있다.



[그림 13.1] 점차트

자, 이제는 검정을 해 보자.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.25} = 0.47$$

이 되고 자유도는 $4-1=3$가 된다. $\chi^2 = 0.47 < \chi^2(3;0.05) = 7.81$이므로 귀무가설 H_0를 기각할 수 없다(유의수준: 5%). 즉, 멘델의 법칙에서 두 쌍(우성인자: A와 B, 열성인자: a와 b)의 대립형질의 유전은 잡종 제 2세대로 내려가면 AB: Ab: aB: ab=9: 3: 3: 1의 비율을 이루지 못한다는 강력한 증거가 없다. ■

13.2 분할표의 분석

요즈음 국제 광물가격 상승의 충격이 대단하다. 채산성이 맞지 않아 폐광되었던 국내 광산들이 다시 주목을 받는 실정이다. 다음 기사는 서울경제신문에 실린 기사이다.

"불꺼진 광산도 다시 보자"
국제 광물가격 급등으로 국내 구리鑛등 주목
加업체선 영월 텅스텐광산에 5억弗 투자키로

가격경쟁에서 밀려 천덕꾸러기로 전락했던 국내 광산들이 다시 주목을 받고 있다. 휴광 상태인 상동광산에 캐나다 업체가 5억달러를 투자하기로 하는가 하면 해외에서 자원개발에 치중했던 산업자원부는 국내 금·동·몰리브덴 등의 광산을 본격적으로 개발할 예정이다.
12일 서울 신대방동 광업진흥공사에서 열린 제6회 광물자원 투자포럼에서 김정관 산자부 에너지자원개발본부장은 "금속 가격의 상승으로 기존 개발 광종인 금·철 외에 동·몰리브덴·중석도 개발잠재력이 있다"며 취약한 국내 부존 금속광의 개발 확대 필요성을 지적했다.
지난 1990년대 이후 광산물 수요가 5년마다 2배 정도로 급격하게 늘면서 금속광의 경우 전체 수요량의 99.3%를 수입에 의존하고 있는 실정이다. 하지만 부존량이 부족하고 가격경쟁에서 밀리면서 국내 광산 개발은 사실상 중단됐다. 김 본부장은 "전세계 광물 가격이 급등했고 광물자원의 확보 경쟁으로 국내 광업이 재조명되고 있다"고 말했다. 또 일부 부존자원의 경우 연간 1,000억원 이상의 개발가치가 있다는 설명이다. 김 본부장은 "국내에서 금·철 외에 동·몰리브덴·중석이 개발잠재력이 있다"며 "이들 광종의 8개 광산을 신규 개발할 경우 연간 1,135억원, 10년간 1조1,000억원의 생산을 기대할 수 있다"고 말했다.
광물개발을 위해 산림청도 지원에 나선다. 산림청은 이날 포럼에 참석, 광물탐사와 채굴을 위한 산지 전용 및 채광계획 인허가 절차를 소개하면서 "산지의 보전과 개발이라는 두 가지 명제가 조화를 이룰 수 있도록 정책에 적극 반영하겠다"고 밝혔다.
국내 광산에 대한 외국의 자원탐사 전문기업의 투자도 유치됐다. 캐나다의 자원탐사 전문기업 OTL사는 강원도 영월 상동광산의 텅스텐과 몰리브덴광 매장량의 잠재가치를 600억달러로 평가하면서 "상동광산에 오는 2010년까지 5억달러를 투자할 계획"이라고 밝혔다. 또 국내 광업회사와의 제휴 의사도 피력했다. 한때 국내 유수의 텅스텐광이었던 상동광산은 1990년대 중국의 가격공세에 밀려 현재는 휴광 상태다.
산자부의 한 관계자는 이에 대해 "OTL사가 평가한 가치는 아직 추정치여서 좀더 조사와 검토가 필요하다"고 설명했다.

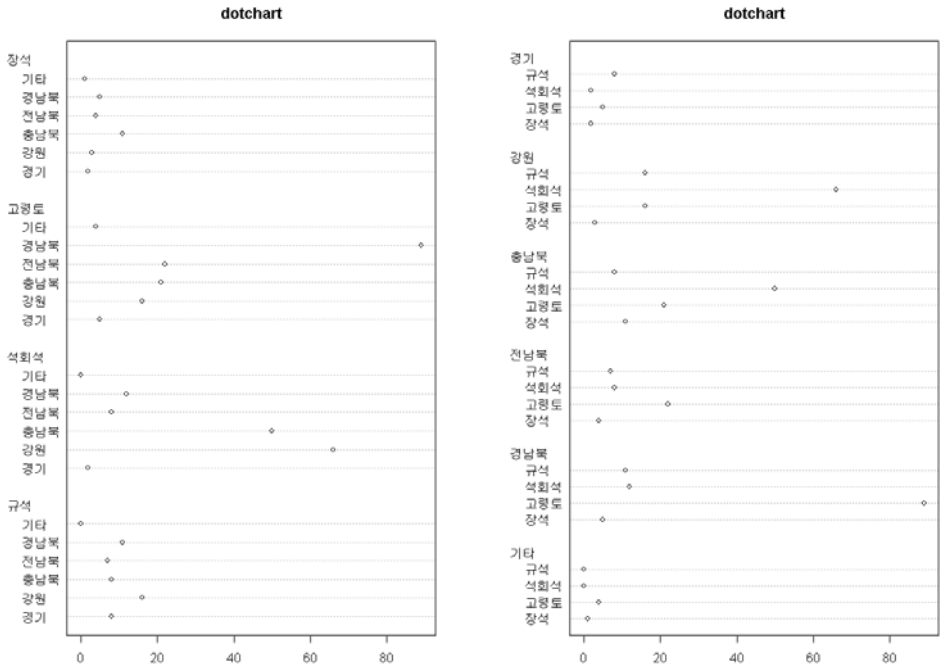
다음 [표 13.2]는 국가통계포털(www.kosis.kr)에 있는 '도별 광종별 생산 광산수'를 이용하여 재구성한 표이다. 우리나라 광산은 주로 비금속 광산이어서 비금속 중 장석, 고령토, 석회

석, 규석 4종류를 선택하고 9개의 도를 6개로 묶었다. 우리는 이런 표를 분할표(contingency table)라 부른다. '지역'이라는 특성(변수)와 '비금속 종류'라는 특성(변수)는 모두 범주형 자료가 된다. 이 두 변수 모두 명목척도가 된다. 우리는 이러한 분할표를 '주변 합계가 고정되지 않은 분할표'라고 부른다. 이 분할표에서 우리가 알고 싶은 것은 "지역과 비금속 종류 사이에 연관성이 있느냐?"는 것이다.

종류 \ 지역	경기	강원	충남북	전남북	경남북	기타	합계
장석	2	3	11	4	5	1	26
고령토	5	16	21	22	89	4	157
석회석	2	66	50	8	12	0	138
규석	8	16	8	7	11	0	50
합계	17	101	90	41	117	5	371

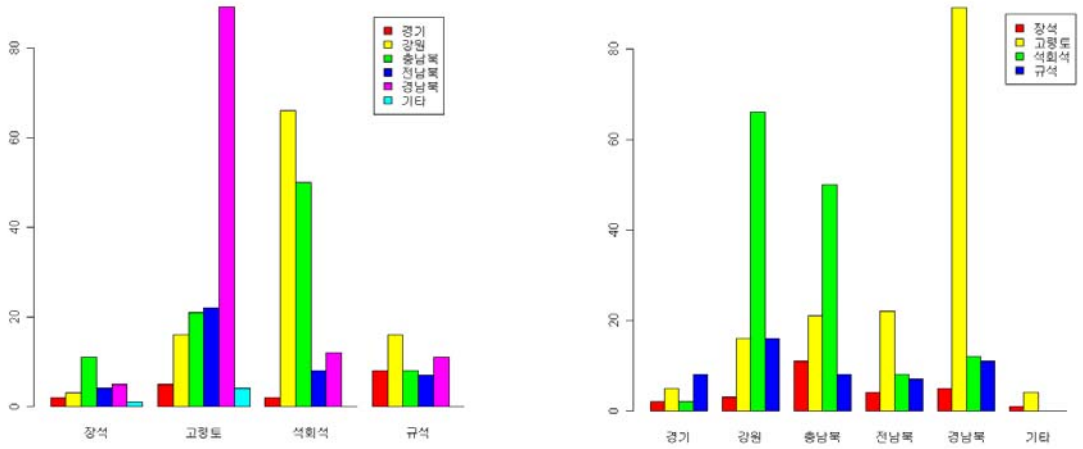
[표 13.2] 도별 광종별 생산 광산수에 대한 분할표

자, 앞에서 제시한 분할표에서 "지역과 비금속 종류 사이에 연관성이 있느냐?"에 대하여 자료분석을 해 보자! 탐색적 단계로서 앞의 분할표를 이용하여 점차트(dotchart)를 그리면 다음 [그림 13.2]와 같다. 지역과 비금속 종류 사이에 연관성이 있음을 알 수 있다. 고령토 광산은 경남북에, 석회석 광산은 강원과 충남북에 많이 소재하고 있음을 알 수 있다.



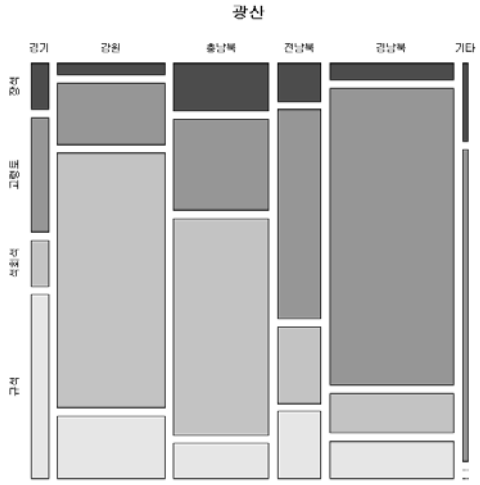
[그림 13.2] 점차트

앞의 분할표를 이용하여 병렬막대그래프를 그리면 다음 [그림 13.3]과 같다. 지역과 비금속 종류 사이에 연관성이 있음을 알 수 있다. 고령토 광산은 경남북에, 석회석 광산은 강원과 충남북에 많이 소재하고 있음을 알 수 있다.



[그림 13.3] 병렬막대그래프

다음 [그림 13.4]는 앞의 분할표에 대한 모자이크그림이다. 이러한 모자이크그림은 R이라는 공개용(GPL, General Public License) 통계패키지(무료 통계패키지)를 사용하여 그렸다. 이 모자이크그림에서 각 상자는 분할표 상의 각 칸(cell)에 대응되는데 각 상자의 가로 길이는 지역에 대한 주변표(marginal table), 즉 지역에 대한 합계에 비례하여 그리고 각 상자의 세로 길이는 광산 종류에 대한 주변표, 즉 광산 종류에 대한 합계에 비례하여 그린다. 예로 왼쪽 가장 아래 상자(경기지역 규석)의 가로길이는 17, 세로길이는 50에 대응된다. 각 상자의 가로 길이를 지역에 대한 합계에 비례하여 그리고 각 상자의 세로 길이를 광산 종류에 대한 합계에 비례하여 그리면 각 상자의 크기는 분할표 상의 각 칸의 빈도수에 비례하게 된다. 상자가 제일 큰 것은 경남북/고령토이고 그 다음 큰 것이 강원/석회석, 충남북/석회석 순이다. 이 상자를 통하여 강원고령토 광산은 경남북에, 석회석 광산은 강원과 충남북에 많이 소재하고 있음을 알 수 있다. 지역과 비금속 종류 사이에 연관성이 있음을 알 수 있다. 지역과 비금속 종류 사이에 연관성이 없다면 상자 사이의 틈이 동서남북 일직선으로 놓일 것이다.



[그림 13.4] 모자이크그림

좀 더 나아가 보자! 앞의 분할표에서 지역과 비금속 종류 사이에 연관성이 있는 지를 밝히기 위하여 확증적 단계로서 검정을 행하여 보자. 이 분할표는 주변 합계가 고정되지 않은 분할표이므로 이러한 분할표에서의 검정을 우리는 '독립성검정(Test of Independence)'이라 부른다. 독립성검정(카이제곱검정)은 다음과 같이 표현할 수 있다(유의수준: α%).

귀무가설 H_0: 지역과 비금속 종류는 서로 독립이다.(지역과 비금속 종류 사이에 연관성이 없다.)
대립가설 H_1: 지역과 비금속 종류는 서로 종속이다.(지역과 비금속 종류 사이에 연관성이 있다.)

검정규칙은 다음과 같다.

검정통계량 $\chi^2 = \sum_{\text{모든 칸}} \frac{(O-E)^2}{E}$이 $\chi^2((l-1)(m-1);\alpha)$보다 크면 H_0를 기각한다.

여기서, l과 m은 각 범주형 변수에서 범주의 개수, O는 분할표 상의 관측도수이고, E는 지역과 비금속 종류는 서로 독립이라 가정할 때의 기대도수로 $E=\frac{\text{칸이 속한 열의 합} \times \text{칸이 속한 행의 합}}{\text{전체 합}}$이고, $\sum_{\text{모든 칸}}$은 모든 칸에 대하여 더한다는 것이다. $\chi^2((l-1)(m-1);\alpha)$는 자유도가 $(l-1)(m-1)$인 카이제곱분포에서의 상위 α% 백분위수이다. 귀무가설이 맞다는 가정 하에서는 표본의 크기가 클 때 검정통계량 $\chi^2 = \sum_{\text{모든 칸}} \frac{(O-E)^2}{E}$이 자유도가 $(l-1)(m-1)$인 카이제곱분포를 이룬다는 것이 밝혀져 있다. 지역이 '경기'이고 비금속 종류가 '장석'인 칸에 대하여 E를 계산하면 $E=\frac{26\times 17}{371}=1.19$이다. 24개의 칸 각각에 대하여 $\frac{(O-E)^2}{E}$을 계산한 후 모두 더하여 χ^2을 계산

하면

$$\chi^2 = \sum_{\text{모든 칸}} \frac{(O-E)^2}{E} = \frac{\left(2-\frac{17\times 26}{371}\right)^2}{\frac{17\times 26}{371}} + \frac{\left(3-\frac{101\times 26}{371}\right)^2}{\frac{101\times 26}{371}} + \ldots + \frac{\left(0-\frac{5\times 50}{371}\right)^2}{\frac{5\times 50}{371}} = 148.09$$

이 되고 자유도는 $(l-1)(m-1)=(6-1)(4-1)=15$가 된다. $\chi^2 = 148.9 > \chi^2(15;0.05) = 25.00$이므로 귀무가설 H_0를 기각한다(유의수준: 5%). 즉 지역과 비금속 종류 사이에 연관성이 있다고 할 수 있다. p-값 < 2.2×10^{-6}을 통하여서도 이러한 사실을 알 수 있다. 다음 표는 기대도수를 나타낸 표이다.

	경기	강원	충남북	전남북	경남북	기타
장석	1.191375	7.078167	6.307278	2.873315	8.199461	0.3504043
고령토	7.194070	42.741240	38.086253	17.350404	49.512129	2.1159030
석회석	6.323450	37.568733	33.477089	15.250674	43.520216	1.8598383
규석	2.291105	13.611860	12.129380	5.525606	15.768194	0.6738544

다음 표는 표준화잔차(standardized residual, 피어슨잔차(Pearson's residual)라고도 함) $\frac{O-E}{\sqrt{E}}$를 나타낸 표이다.

	경기	강원	충남북	전남북	경남북	기타
장석	0.7408379	-1.5328675	1.868546	0.6646771	-1.117337	1.0973835
고령토	-0.8180186	-4.0903266	-2.768616	1.1162472	5.611874	1.2952553
석회석	-1.7193069	4.6385574	2.855701	-1.8566663	-4.777973	-1.3637589
규석	3.7716340	0.6472932	-1.185675	0.6272251	-1.200779	-0.8208864

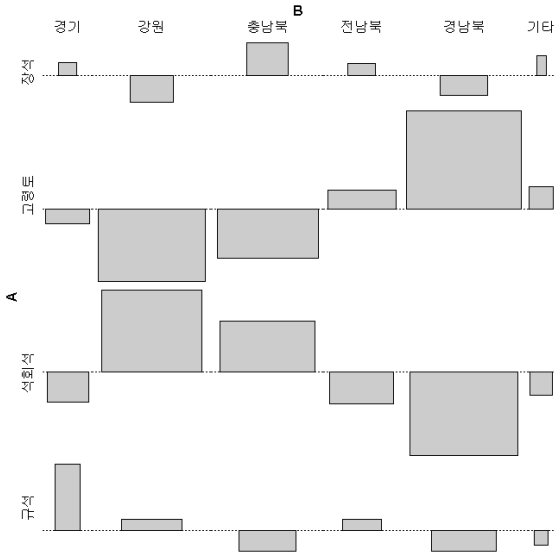
절대값이 큰 양의 표준화잔차를 갖고 있는 셀은 크기순으로 5.612(경남북, 고령토) > 4.639(강원, 석회석) > 3.772(경기, 규석) > 2.856(충남북, 석회석) 등이고 절대값이 큰 음의 표준화잔차를 갖고 있는 셀은 크기순으로 -4.090(강원, 고령토) < -4.778(경남북, 석회석) < -2.769(충남북, 고령토) 등이다. 이 표준화잔차를 이용하면 다음과 같은 결론을 이끌어낼 수 있다.

1. 강원도에서는 고령토가 관측도수가 기대도수보다 상대적으로 적고 석회석은 관측도수가 기대도수보다 상대적으로 많다.
2. 경상남북도에서는 고령토가 관측도수가 기대도수보다 상대적으로 많고 석회석은 관측도수가 기대도수보다 상대적으로 적다.
3. 충청남북도는 강원도나 경상남북도보다는 약하나 고령토가 관측도수가 기대도수보다 상대적

으로 적고 석회석은 관측도수가 기대도수보다 상대적으로 많다.

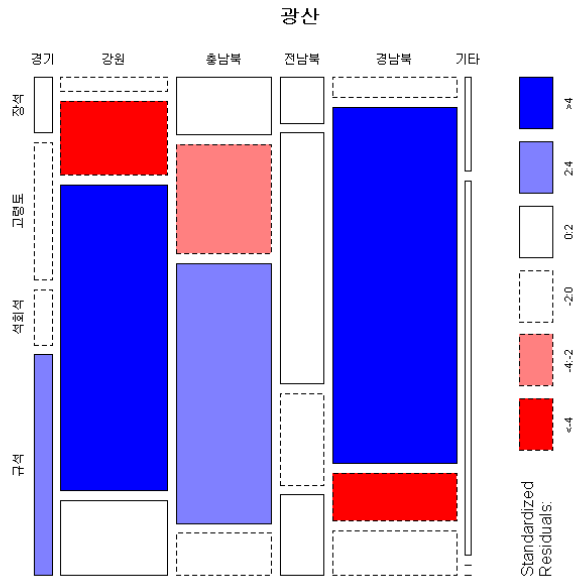
4. 대체적으로 이 세 지역(강원도, 경상남북도, 충청남북도)에서의 패턴 때문에 지역과 비금속 종류 사이에 연관성이 있게 되는 것이다.

다음 [그림 13.5]는 연관도(association plot)인데 상자의 높이는 피어슨잔차의 크기를 나타내고, 폭은 $\sqrt{E}$를 나타내고, 면적은 $O-E$에 비례한다. 피어슨잔차가 음수이면 직선 아래에, 피어슨 잔차가 양수이면 직선 아래에 표시하였다. 상자의 높이와 폭, 상자의 높이 부호 그리고 상자의 크기를 통하여 우리는 앞에서와 같은 결론을 이끌어 낼 수 있다.



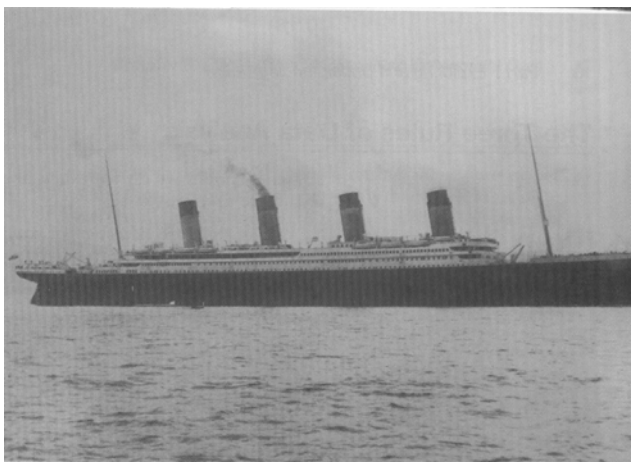
[그림 13.5] 연관도

다음 [그림 13.6]과 같은 모자이크그림에서 점선은 피어슨잔차가 음수이면 점선 상자로, 피어슨 잔차가 양수이면 실선 상자로 표시하였다. 피어슨잔차가 음수이면서 절대값이 클수록 진한 빨간색으로 표시하고 피어슨잔차가 양수이면서 절대값이 클수록 진한 파란색으로 표시하였다. 강원도에서는 고령토가 관측도수가 기대도수보다 상대적으로 적고(진한 빨간색) 석회석은 관측도수가 기대도수보다 상대적으로 많다(진한 파란색). 반면에 경상남북도에서는 고령토가 관측도수가 기대도수보다 상대적으로 많고(진한 파란색) 석회석은 관측도수가 기대도수보다 상대적으로 적다(진한 빨간색). 충청남북도는 강원도나 경상남북도보다는 약하나 고령토가 관측도수가 기대도수보다 상대적으로 적고(옅은 빨간색) 석회석은 관측도수가 기대도수보다 상대적으로 많다(옅은 파란색). 대체적으로 이 세 지역(강원도, 경상남북도, 충청남북도)에서의 패턴 때문에 지역과 비금속 종류 사이에 연관성이 있게 되는 것이다.



[그림 13.6] 모자이크그림

예제 13.2 1912년 영국을 떠나 미국으로 처녀항해에 나서던 호화유람선 타이타닉호가 북극해의 빙산과 충돌하여 침몰한 대형 해양참사는 지금까지 세인들의 화제가 되고 있다. 이 사건은 여러 번 영화로 만들어졌고 가장 최근에는(10년이 넘는 옛날이야기이지만) 1997년 제임스 카메론 감독이 영화로 만들었다. 영화 자체가 엄청난 스케일에다가 극본도 뛰어났고 남자주인공이 디카프리오여서 많은 화제를 불러 일으켰고 이 영화는 1997~8년도 세계적인 불황기에도 18억 달러라는, 전대미문의 금액을 번 영화로 기록되어 있다. 다음 [표 13.3]은 총 2201명의 승객과 승무원에 대하여 4개의 범주형 변수(등급(1, 2, 3등석, 승무원), 성별(남자, 여자), 나이(아이, 어른), 생존여부(사망,생존)로 나누어 작성한 4차원 분할표이다.



생존여부 / 나이 / 등급 / 성별	사망				생존			
	아이		어른		아이		어른	
	남자	여자	남자	여자	남자	여자	남자	여자
1등석	0	0	118	4	5	1	57	140
2등석	0	0	154	13	11	13	14	80
3등석	35	17	387	89	13	14	75	76
승무원	0	0	670	3	0	0	192	20

[표 13.3] 타이타닉호 참사에 대한 분할표

이러한 4차원 분할표를 통계분석하는 방법이 있으나 본 저서의 범위를 넘어감으로 가능한 범위 내에서 분석하여 보자. 우리의 관심은 다음과 같은 것들이다.

1. 4개의 범주형변수들 사이에 연관성이 있는가?
2. 연관성이 있다면 서양인들이 미덕으로 내세우고 있는 'women and children policy'(여자와 아이를 우선적으로 배려하는 미덕)가 지켜졌는가?
3. 좀 어폐가 있는 표현이지만 '유전무죄, 무전유죄'가 여기에도 적용되는가? 즉, 1, 2, 3등석 승객 모두 공평하게 사망하거나 생존하였는가?

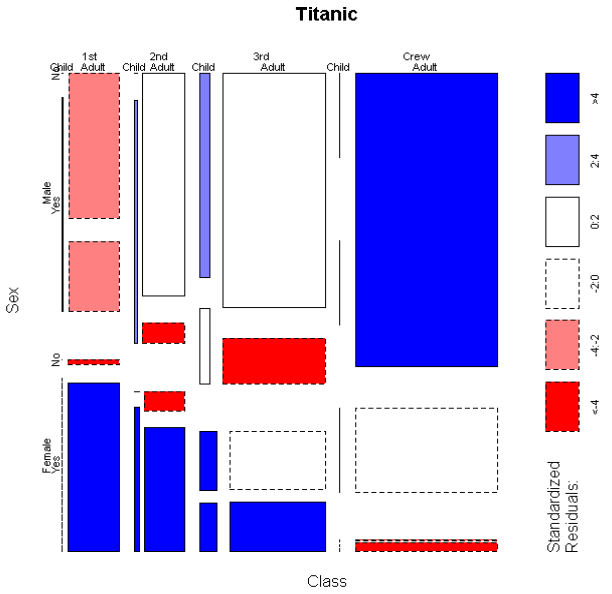
먼저 [표 13.4]와 같은 사망률(100명당 사망자수)표에서 사망률을 비교하여 보자.

등급	성별		나이	
	남자	여자	아이	어른
1등석	65	3	0	38
2등석	87	12	0	64
3등석	83	54	66	76
승무원	78	13	-	76

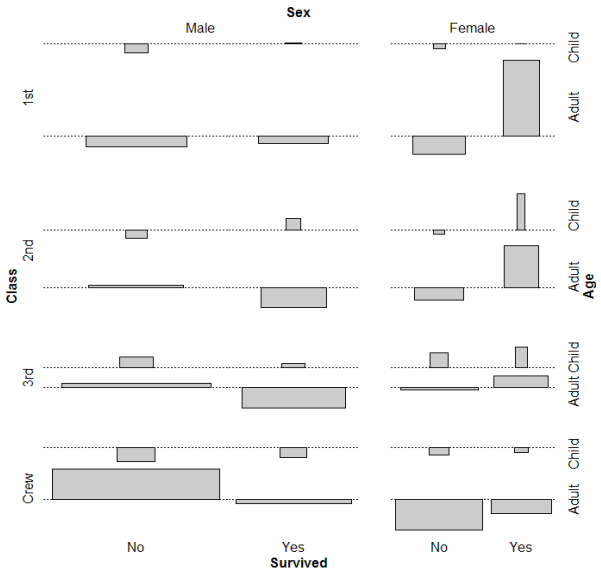
[표 13.4] 사망률표

남자의 사망률에서는 경제적 지위를 나타내는 등급에 따라 차이가 별로 크게 나지 않는다. 그러나 어른의 사망률을 들여다보자. 어른의 사망률에서는 3등석이 1등석의 2배의 차이가 난다. 그래도 이것은 약과다. 여자의 사망률을 보자. 여자의 사망률에서는 3등석이 2등석의 4.5배, 1등석과는 무려 18배의 차이가 난다. 아이의 사망률에서는 더 가관이다. 아이의 사망률에서는 1등석과 2등석의 사망률은 0인 반면 3등석은 무려 66이다. 여자와 아이를 우선적으로 배려하는 미덕도 1등석이나 2등석에나 적용되는 미덕인 셈이다. 1, 2, 3등석 승객 모두 공평하게 사망하거나 생존하지 않았음을 짐작할 수 있다. 이 [표 13.3]과 같은 사망률표만 봐도 4개의 범주형변수들 사이에 연관성이 있음을 짐작할 수 있다.

앞에서와 같은 언급을 그림으로 확인하여 보자. 다음 [그림 13.7]과 [그림 13.8]은 [표 13.3]을 이용하여 각각 구한 모자이크그림과 연관도이다. 이 그림들을 통하여서도 앞에서와 같은 언급(여자와 아이를 우선적으로 배려하는 미덕도 1등석이나 2등석에나 적용되는 미덕이다. 1, 2, 3등석 승객 모두 공평하게 사망하거나 생존하지 않았다.)을 확인할 수 있다.

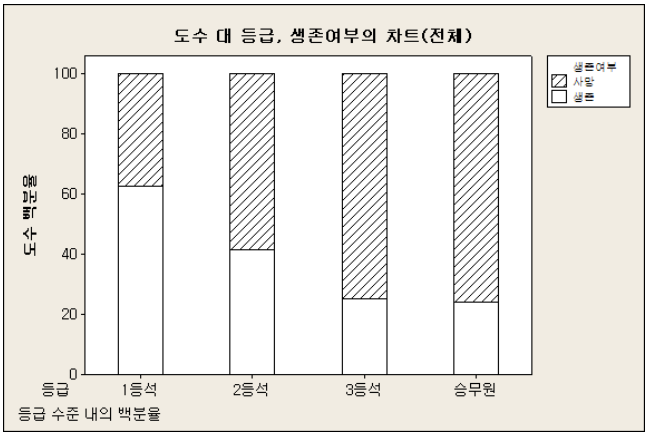


[그림 13.7] 타이타닉호 참사에 대한 모자이크그림

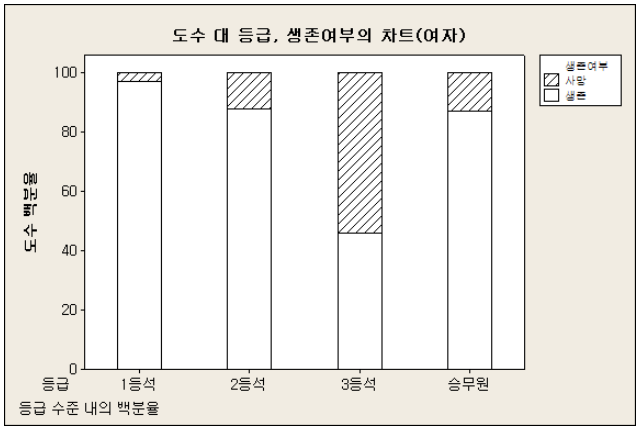


[그림 13.8] 타이타닉호 참사에 대한 연관도

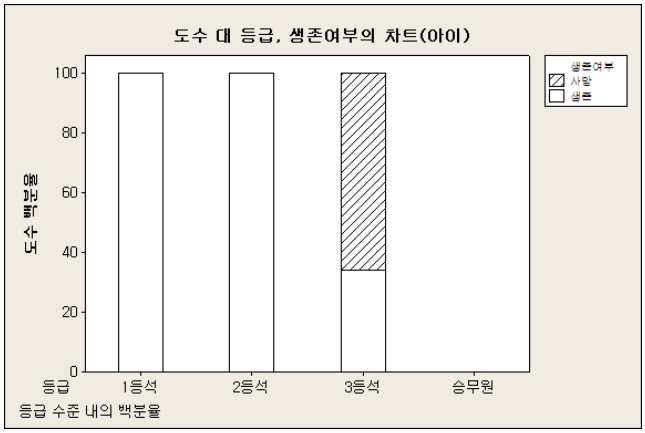
그러면 앞에서와 같은 언급(여자와 아이를 우선적으로 배려하는 미덕도 1등석이나 2등석에나 적용되는 미덕이다. 1, 2, 3등석 승객 모두 공평하게 사망하거나 생존하지 않았다.)을 더 간단한 막대그래프로 확인하여 보자! [그림 13.9]에서 [그림 13.11]까지의 그림들을 통하여 확인할 수 있다. ■



[그림 13.9] 등급에 따른 막대그래프(전체)



[그림 13.10] 등급에 따른 막대그래프(여자)



[그림 13.11] 등급에 따른 막대그래프(아이)

다음은 2007.01.30 조선일보 기사이다.

말 많았던 '일해공원' 명칭 확정-합천군 주민 설문 거쳐

경남 합천군은 그 동안 논란이 일었던 '새천년 생명의 숲'공원 명칭을 지역 출신인 전두환(全斗煥) 전 대통령의 아호를 따 '일해(日海)공원'으로 29일 확정했다.
합천군청 실·과장 등 20명으로 구성된 합천군정조정위원회(위원장 정희식 부군수)는 주민 설문조사에서 가장 많은 표를 얻었고, 군 의원 11명 중 9명이 '일해공원' 지지 입장을 밝힌 점을 고려해 이같이 결정했다고 밝혔다.
'일해' '황강' '군민' '죽죽'등 네 가지 명칭을 놓고 지난해 12월 합천군과 읍·면의 유관 기관장, 단체장, 새마을지도자, 이장 등 1364명을 대상으로 설문조사 한 결과 회수된 591장(43%)의 설문지 중 '일해'가 302표로 가장 많았다.
합천군의 '일해공원' 결정에는 "지명도가 있어 관광객 유치 등 지역 발전에 도움된다"는 판단도 한 몫 했다.
그러나 지난 18일 서울 연희동 전(全) 전 대통령 자택 앞 시위 등을 통해 일해공원 명칭에 반대해온 '일해공원 반대 경남대책위원회'와 '새천년 생명의 숲 지키기 합천군민 모임' 등은 이번 결정에 반발, 반대 투쟁을 더 적극적으로 벌인다는 방침이어서 갈등은 한동안 계속될 전망이다.
합천읍 황강변 '새천년 생명의 숲'은 2000년을 맞아 합천군이 도비(道費) 지원을 받아 진행한 '밀레니엄 기념사업'이다. 황강변 5만3000여㎡에 68억원을 들여 만든 인공 공원으로 다양한 수목과 잔디밭을 배경으로 550석 규모 야외 공연장, 산책로, 체육시설 등이 설치돼 있다.

이 기사에서 문제가 되는 것은 다음과 같은 두 가지이다.

1. 설문조사 대상자가 군내 대표자들(합천군과 읍·면의 유관 기관장, 단체장, 새마을지도자, 이장)인데 이들이 합천군민의 의사를 정확히 반영하고 있는가? 이들은 자치단체와 밀접한 관계를 갖고 있는 이해당사자들이다. 설문조사에 자치단체에 우호적인 그룹인 군내 대표자그룹과 일반군민그룹으로 나누어 설문조사를 시행했어야 하지 않은가?
2. 응답하지 않은 사람들 773명(1,364명의 57%)을 제외하고 응답한 사람들 591명 중 302명이 '일해' 명칭에 찬성하여 찬성률이 51.1% 과반수찬성이므로 정당성이 있다고 주장하고 있다. 이러한 의사결정이 좀 이상하지 않는가? 1,364명 중 591명이 응답하였으므로 응답률은 43%이다. 1,364명 중 302명(22%)만이 자치단체안에 찬성, 289명(21%) 반대, 773명(57%) 무응답이다. 자치단체에 유리한 쪽으로 몰고 간 사건이라 할 수 있다.

예로 설문조사 대상자를 군내 대표자그룹 1,000명과 일반군민그룹 1,000명으로 나누어 설문조사를 시행했다면 다음 [표 13.5]와 같은 분할표가 작성되었을 것이다. 우리는 이러한 분할표를 '한 쪽 주변합계가 고정된 분할표'라고 부른다. [표 13.2]('주변 합계가 고정되지 않은 분할표')와 비교하여 보아라. 차이점이 무엇인지 알겠는가?

명칭 / 그룹	일해 황강 군민 죽죽	합계
군내 대표자그룹		1,000
일반군민그룹		1,000
합계		2,000

[표 13.5] 가상설문조사를 통하여 얻어지는 분할표

이 분할표는 한 쪽 주변합계가 고정(각 1,000명씩)된 분할표이므로 이러한 분할표에서의 검정을 우리는 '동질성검정(Test of Homogeneity)'이라 부른다. 동질성검정(카이제곱검정)은 다음과 같이 표현할 수 있다.(유의수준: α%)

귀무가설 H_0: 그룹별로 명칭 선택 비율에 차이가 없다.

대립가설 H_1: 그룹별로 명칭 선택 비율에 차이가 있다.

검정규칙은 다음과 같다.

검정통계량 $\chi^2 = \sum_{\text{모든 칸}} \frac{(O-E)^2}{E}$이 $\chi^2((l-1)(m-1);\alpha)$보다 크면 H_0를 기각한다.

독립성 검정과 절차가 동일함을 확인할 수 있을 것이다.

예제 13.3 어느 지방자치단체의 시책에 대하여 여론조사를 실시하였다. 거주지역에 따라 세 그룹(부차모집단이라고 부르기도 함)으로 나누어 각각 200, 200, 100명씩 조사하여 다음 [표 13.6]과 같은 분할표를 얻었다.

찬성여부 / 지역	찬성	반대	합계
도시지역	143	57	200
도시근교지역	98	102	200
농촌지역	13	87	100
합계	254	246	500

[표 13.6] 여론조사를 통하여 얻어진 분할표

우선 다음 [표 13.7]의 왼쪽 표(백분율표)에서 행백분율, 열백분율, 전체백분율을 보자. 각 셀에서 첫 번째 값이 관측도수, 두 번째 값이 행백분율, 세 번째 값이 열백분율, 네 번째 값이

전체백분율이다. 한 예로 '농촌, 반대' 셀에서 87명이 관측도수이고 '농촌' 행에서 '반대'가 87명, '찬성'이 13명이니 '농촌, 반대' 셀의 행백분율은 87/100*100=87(%)가 된다. '반대' 열에서 '농촌'이 87명, '도시'가 57명, '도시근교'가 102명이니 '농촌, 반대' 셀의 열백분율은 87/246*100=35.37(%)가 된다. 또한, 전체 합계가 500명이니 '농촌, 반대' 셀의 전체백분율은 87/500*100=17.40(%)가 된다. 행백분율을 살펴보면 '농촌'에서 '반대'가 87%, '찬성'이 13%인 반면 '도시'에서는 '반대'가 28.5%, '찬성'이 71.5%이어서 '도시'와 '농촌'은 서로 '농촌'과 반대현상을 나타내고 있다. '도시근교'는 '반대'가 51%, '찬성'이 49%로 백중세이다. 다음 [표 13.7]의 오른 쪽 표(잔차표)에서 잔차와 표준화잔차를 보자. '농촌, 반대' 셀에서 87명이 관측도수이고 49.2명(=$\frac{246\times100}{500}$)이 기대도수이다. 그러므로 잔차는 37.8(=87-49.2)이고 표준화잔차는 5.389(=$\frac{87-49.2}{\sqrt{49.2}}$)가 된다. 표준화잔차를 살펴보면 '농촌, 반대' 셀에서 5.389, '농촌, 찬성' 셀에서 -5.303인 반면 '도시, 반대' 셀에서 -4.174, '도시, 찬성' 셀에서 4.107이어서 '도시'와 '농촌'은 서로 반대현상을 나타내고 있다. '도시근교'는 '도시근교, 반대' 셀에서 0.363, '도시근교, 찬성' 셀에서 -0.357이어서 백중세이나 두 개의 값 모두 절대값이 작다. 결론적으로 거주지역에 따라 세 그룹 사이의 찬성비율이 다르다는 것을 짐작할 수 있는데 이는 주로 '농촌' 지역과 '도시'지역의 찬성 패턴에서 서로 반대현상이 나타나고 있다는 데서 비롯된다. 우리는 이렇게 백분율표와 잔차표를 보더라도 거주지역에 따라 세 그룹 사이의 찬성비율이 다르다는 것을 짐작할 수 있다.

제표 통계량: 지역, 찬성여부

도수에서 빈도 사용

행: 지역 열: 찬성여부

	반대	찬성	모두
농촌	87	13	100
	87.00	13.00	100.00
	35.37	5.12	20.00
	17.40	2.60	20.00
도시	57	143	200
	28.50	71.50	100.00
	23.17	56.30	40.00
	11.40	28.60	40.00
도시근교	102	98	200
	51.00	49.00	100.00
	41.46	38.58	40.00
	20.40	19.60	40.00
모두	246	254	500
	49.20	50.80	100.00
	100.00	100.00	100.00
	49.20	50.80	100.00

셀 내용: 카운트
행의 %
열의 %
총계의 %

제표 통계량: 지역, 찬성여부

도수에서 빈도 사용

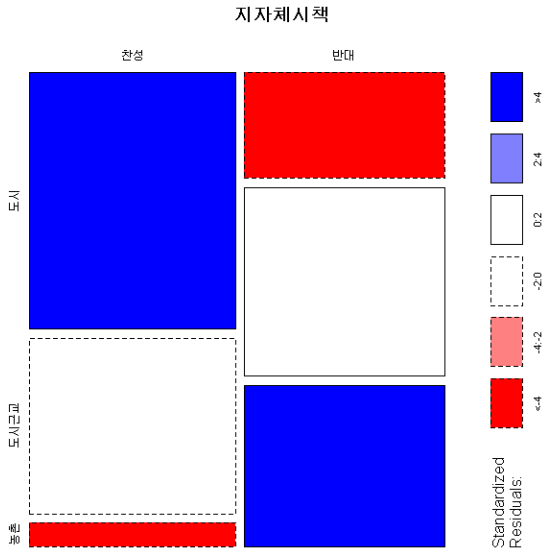
행: 지역 열: 찬성여부

	반대	찬성	모두
농촌	87	13	100
	49.2	50.8	100.0
	37.80	-37.80	*
	5.389	-5.303	*
도시	57	143	200
	98.4	101.6	200.0
	-41.40	41.40	*
	-4.174	4.107	*
도시근교	102	98	200
	98.4	101.6	200.0
	3.60	-3.60	*
	0.363	-0.357	*
모두	246	254	500
	246.0	254.0	500.0
	*	*	*
	*	*	*

셀 내용: 카운트
기대 카운트
잔차
표준화 잔차

[표 13.7] 백분율표와 잔차표

다음 [그림 13.12]은 모자이크그림이다. 점선은 표준화잔차가 음수이면 점선 상자로, 피어슨 잔차가 양수이면 실선 상자로 표시하고 표준화잔차가 음수이면서 절대값이 클수록 진한 빨간색으로 표시하고 표준화잔차가 양수이면서 절대값이 클수록 진한 파란색으로 표시하였다. '도시, 찬성' 셀은 진한 파란색 실선상자, '도시, 반대' 셀은 진한 빨간색 점선상자인 반면 '농촌, 찬성' 셀은 진한 빨간색 점선상자, '농촌, 반대' 셀은 진한 파란색 실선상자이어서 '도시'와 '농촌'은 서로 반대현상을 나타내고 있다. '도시근교'는 '도시근교, 찬성' 셀과 '도시근교, 반대 ' 셀 모두 흰색이어서 두 개의 표준화잔차값 모두 절대값이 작다. 결론적으로 거주지역에 따라 세 그룹 사이의 찬성비율이 다르다는 것을 짐작할 수 있는데 이는 주로 '농촌' 지역과 '도시'지역의 찬성 패턴에서 서로 반대현상이 나타나고 있다는데서 비롯된다.



[그림 13.12] 지자체 시책 여론조사에 대한 모자이크그림

거주지역에 따라 세 그룹 사이의 찬성비율이 다른지를 알기 위하여 6개의 칸 각각에 대하여 $\frac{(O-E)^2}{E}$을 계산한 후 모두 더하여 χ^2을 계산하면

$$\chi^2 = \sum_{\text{모든 칸}} \frac{(O-E)^2}{E} = \frac{\left(143-\frac{254\times 200}{500}\right)^2}{\frac{254\times 200}{500}} + \frac{\left(57-\frac{246\times 200}{500}\right)^2}{\frac{246\times 200}{500}} + \frac{\left(98-\frac{254\times 200}{500}\right)^2}{\frac{254\times 200}{500}}$$

$$= \frac{\left(102-\frac{246\times 200}{500}\right)^2}{\frac{246\times 200}{500}} + \frac{\left(13-\frac{254\times 100}{500}\right)^2}{\frac{254\times 100}{500}} + \frac{\left(87-\frac{246\times 100}{500}\right)^2}{\frac{246\times 100}{500}}$$

$$= 16.87 + 17.42 + 0.13 + 0.13 + 28.13 + 29.04 = 91.72$$

이 되고 자유도는 $(l-1)(m-1)=(3-1)(2-1)=2$가 된다. $\chi^2=91.72>\chi^2(2;0.05)=5.99$이므로 귀무가설 H_0를 기각한다(유의수준: α%). 즉 거주지역에 따라 세 그룹 사이의 찬성비율이 다르다고 할 수 있다. p-값 < 2.2×10^{-16}을 통하여서도 이러한 사실을 알 수 있다. ■

심슨의 역설(Simpson's Paradox)

분할표를 전체적으로 볼 때와 부분적으로 볼 때 서로 다른 결과를 나타내는 경우가 종종 있는 데 이를 심슨의 역설(Simpson's paradox)이라고 한다. 이 심슨의 역설에 대해서는 예제 6.4에서 살펴본 바가 있다. 비슷한 예제를 다시 살펴보자.

예제 13.4 다음 [표 13.8]은 어느 대학교의 4개의 학부(이 대학교는 4개의 학부로 구성되어 있음.)에 지원한 학생 수와 합격한 학생 수를 남녀별로 정리한 분할표이다. 다음 문제를 풀어보자!

	합격한 남학생	합격한 여학생
학부 A	511(825)	89(108)
B	352(560)	17(25)
C	137(407)	132(375)
D	22(373)	24(341)
합계	1,022(2,165)	262(849)

(괄호 안은 지원자수)

[표 13.8] 지원자수와 합격자수를 남녀별로 정리한 분할표

(a) 전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그려 보라. 무엇을 알 수 있는가?
(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그려 보라. 무엇을 알 수 있는가?
(c) (a)와 (b)를 종합하여 결론을 내려 보라.
(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그려 비교하라.

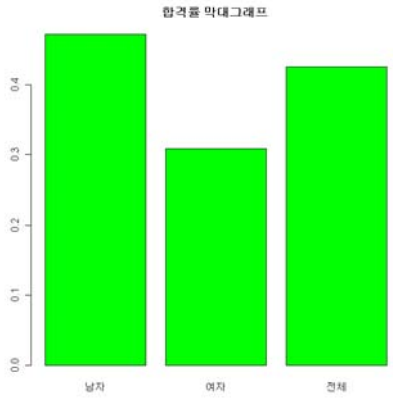
(풀이) (a)
• 우선 분할표를 다음과 같이 작성하여 보자.

	합격 남학생	지원 남학생	합격 여학생	지원 여학생
학부 A	511	825	89	108
학부 B	352	560	17	25
학부 C	137	407	132	375
학부 D	22	373	24	341
전체	1022	2165	262	849

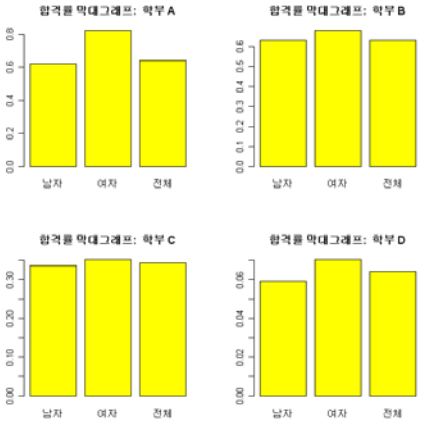
• 분할표를 이용하여 합격률표를 다음과 같이 작성한다.

	prop.man.entrance	prop.woman.entrance	prop.total.entrance
학부 A	0.61939394	0.82407407	0.64308682
학부 B	0.62857143	0.68000000	0.63076923
학부 C	0.33660934	0.35200000	0.34398977
학부 D	0.05898123	0.07038123	0.06442577
전체	0.47205543	0.30859835	0.42601194

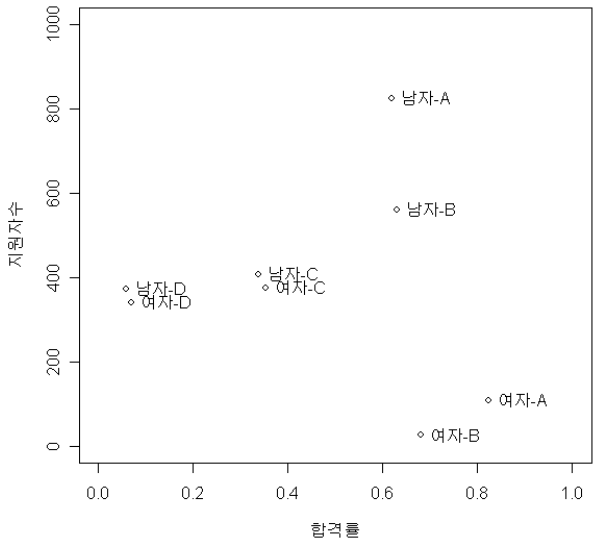
• 전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음과 같다. 전체 남자의 합격률이 여자의 합격률보다 약 16% 큼을 알 수 있다.



(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음과 같다. 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다.



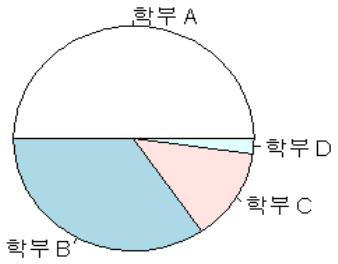
(c) 전체적으로는 남자의 합격률이 여자의 합격률보다 크나 각 학부별로는 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다. 이러한 현상을 심슨의 역설이라고 한다. 이러한 현상이 발생한 이유는 무엇일까? 이 이유를 알기 위하여 다음과 같은 산점도를 살펴보자.



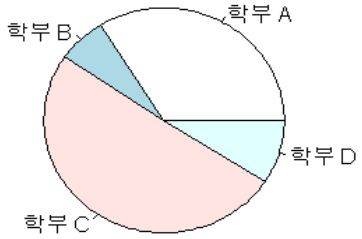
여자의 경우 학부 C와 D에서 상대적으로 지원자수는 많으나 합격률이 낮은 반면 학부 A와 B에서는 합격률이 높으나 상대적으로 지원자수가 적다. 즉, 여자들은 합격률이 낮은 학부에 많이 지원한 반면 합격률이 높은 학부에는 적게 지원하였다. 반면 남자의 경우는 학부 C와 D에서 상대적으로 지원자수가 적고 합격률도 낮은 반면 학부 A와 B에서는 상대적으로 지원자수가 많고 합격률도 높다. 즉, 남자들은 여자들과 달리 합격률이 낮은 학부에 적게 지원한 반면 합격률이 높은 학부에는 많이 지원하였다. 이런 연유로 전체적으로는 남자의 합격률이 여자의 합격률보다 크나 각 학부별로는 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 크게 된 것이다. 우리가 관심이 있는 남녀 간의 전체 합격률의 차이는 남녀 간의 차별 때문이 아니라 남녀 간의 학부별 지원 선호도 차이로 인하여 발생하였음을 알 수 있다. '남녀 간의 학부별 지원 선호도 차이'라는 중요한 변수를 고려하지 않고 '남녀'라는 성별의 차이로만 합격률을 보면 잘못된 결론을 내릴 수가 있다.

(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그리면 다음과 같다. 합격한 남학생과 합격한 여학생의 학부 분포율이 아주 다름을 알 수 있다. 합격한 남학생은 A(50.0%) -> B(34.4%) -> C(13.4%) -> D(2.2%) 순이나 합격한 여학생은 C(50.0%) -> A(34.0%) -> D(9.1%) -> B(6.5%) 순이다. ■

합격남학생 원형그래프



합격여학생 원형그래프



학습요약

자료가 범주형자료인 경우 수치자료와는 다른 자료표현방법이나 자료분석방법을 사용한다. 자료가 분할표로 주어졌을 때 변수 사이의 연관성을 알아보기 위하여 우리는 모자이크그림을 유용하게 사용할 수가 있다. 또한, 자료분석을 위해서는 카이제곱을 이용하여 검정을 행할 수 있다.

13장 연습문제

13.1 어떤 난수표에서 한 페이지를 선택하니 800개의 난수가 있었다. 800개의 난수를 숫자별(1~9)로 나누어 보니 다음 표와 같았다. 9개의 숫자가 골고루 있다고 할 수 있는 지를 그림을 그려 살펴보고 카이제곱검정을 행하라(유의수준: 5%). 어떤 결론을 내릴 수 있는가?

숫자	0	1	2	3	4	5	6	7	8	9	합계
도수	85	77	83	90	69	79	80	76	84	77	800

13.2 다음 표는 스포츠조선(sports.chosun.com/sports/baseball/ranking/team.htm)에 나타난 2007년 프로야구 최종순위표이다.

» 최종순위 «

팀순위

순위	팀	승	패	무	승률	연속	승차	실책	득점	실점	홈런	도루	타율	방어율
1	SK	73	48	5	.603	1승	-	88	603	465	112	136	.264	3.24
2	두산	70	54	2	.565	1패	4.5	73	578	480	78	161	.263	3.44
3	한화	67	57	2	.540	1승	7.5	76	534	481	104	48	.254	3.54
4	삼성	62	60	4	.508	2패	11.5	87	497	509	86	101	.254	3.71
5	LG	58	62	6	.483	1패	14.5	94	532	600	78	130	.268	4.33
6	현대	56	69	1	.448	2승	19.0	96	530	615	96	51	.271	4.41
7	롯데	55	68	3	.447	2승	19.0	82	533	554	76	67	.270	4.12
8	KIA	51	74	1	.408	1패	24.0	77	499	602	73	70	.257	4.49

(제공처: 스포츠조선)

(1) 8개 팀을 하나의 범주형 변수(팀)로 정의하고 실책, 득점, 실점, 홈런, 도루 횟수를 묶어 하나의 범주형 변수(게임내용)로 정의한 후 이를 이용하여 분할표를 작성하라.
(2) 점차트, 병렬막대그림, 모자이크그림을 그리고 특징을 말하라.
(3) 팀과 게임내용 사이에 연관성이 있는 지 카이제곱검정을 행하라(유의수준: 5%). 어떤 결론을 내릴 수 있는가?

13.3 부품을 조사한 결과 부품의 품질과 생산된 시간대에 따라 다음과 같은 분할표를 얻었다. 유의수준 5%에서 생산된 시간대와 부품의 품질이 독립인지 검정하라. 그리고 결과를 설명해보라.

생산된 시간대	품질	
	양품	불량품
09-17	368	32
01-12	285	15
09-1	176	24

13.4 철도청에서는 고객들이 이용하는 목적과 이용하는 기차의 종류와의 관계를 알아보기 위하여 자료를 수집하였다. 이용 목적은 크게 개인적인 목적과 업무상 목적으로 구분하였으며 기차의 종류는 무궁화, 새마을, KTX로 구분하였다. 유의수준 5%에서 이용 목적과 이용하는 기차의 종류가 독립인지 검정하라. 어떤 결론을 내릴 수 있는가?

기차의 종류	이용 목적	
	개인적인 용무	업무상 이용
무궁화	29	22
새마을	95	121
KTX	75	135

13.5 3종류의 캔을 만드는 제조업체에서는 3종류의 캔을 각각 3개의 다른 생산라인에서 생산해 낸다. 캔에 대하여 품질관리기사는 5종류의 결함을 지적하고 있다. 다음 표는 1, 2, 3생산라인에서 만든 캔 중 결함이 있는 캔을 각각 150개, 125개, 100개 뽑아 어떤 결함이 있는 지를 조사하여 작성한 분할표이다.

결함 / 생산라인	흠집	갈라짐	손잡이불량	손잡이분실	기타	합계
1	34	65	17	21	13	150
2	23	52	25	19	6	125
3	32	28	16	14	10	100
합계	89	145	58	54	29	375

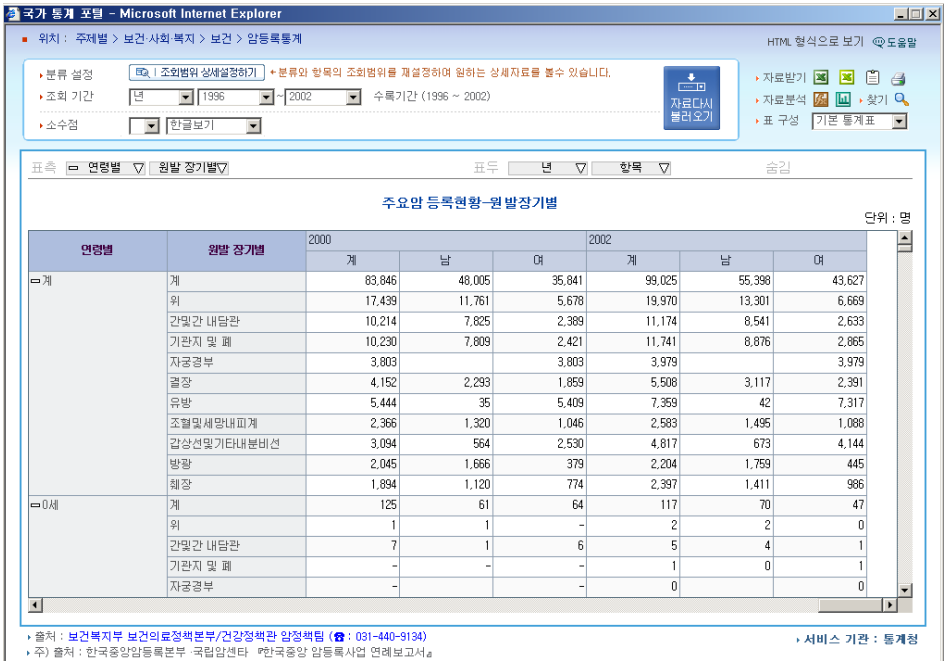
(1) 분할표를 이용하여 행백분율, 열백분율, 전체백분율을 구하여 보고 잔차와 표준화잔차를 구하여 보라. 무엇을 알 수 있나?

(2) 점차트, 병렬막대그림, 모자이크그림을 그리고 특징을 말하라.

(3) 캔의 종류에 따라 결함의 패턴이 차이가 나는지 카이제곱검정을 행하라(유의수준: 5%). 어떤 결론을 내릴 수 있는가?

13장 실습문제

국가통계포털(www.kosis.kr)에서 국내통계>주제별통계>보건사회복지>보건>암등록통계 메뉴 중 주요암 등록현황-원발장기별 기본조회를 선택하면 다음과 같은 화면이 나온다. 자료는 연령별(5세 간격(0세, 1-4세, 5-9세, 10-14세, ..., 80-84세, 85세 이상)), 장기별(위, 간 및 간 내담관, 기관지 및 폐 등), 남녀별(남, 여), 년도별(1996년-2002년)로 구분되어 있다. 즉, 4차원 분할표가 주어지는 셈이다.



국가 통계 포털 - Microsoft Internet Explorer

위치 : 주제별 > 보건·사회 복지 > 보건 > 암등록통계

HTML 형식으로 보기 | 도움말

분류 설정 | 조회범위 상세설정하기 | 분류와 항목의 조회범위를 재설정하여 원하는 상세자료를 볼수 있습니다.

조회 기간 년 1996 ~ 2002 수록기간 (1996 ~ 2002)

소수점 한글보기

자료다시 불러오기

자료받기 | 자료분석 | 찾기 | 표 구성 기본 통계표

표측 연령별 원발 장기별 | 표두 년 항목 | 숨김

주요암 등록현황-원발장기별

단위 : 명

연령별	원발 장기별	2000			2002		
		계	남	여	계	남	여
계	계	83,846	48,005	35,841	99,025	55,398	43,627
	위	17,439	11,761	5,678	19,970	13,301	6,669
	간및간 내담관	10,214	7,825	2,389	11,174	8,541	2,633
	기관지 및 폐	10,230	7,809	2,421	11,741	8,876	2,865
	자궁경부	3,803		3,803	3,979		3,979
	결장	4,152	2,293	1,859	5,508	3,117	2,391
	유방	5,444	35	5,409	7,359	42	7,317
	조혈및세망내피계	2,366	1,320	1,046	2,583	1,495	1,088
	갑상선및기타내분비선	3,094	564	2,530	4,817	673	4,144
	방광	2,045	1,666	379	2,204	1,759	445
	췌장	1,894	1,120	774	2,397	1,411	986
0세	계	125	61	64	117	70	47
	위	1	1	-	2	2	0
	간및간 내담관	7	1	6	5	4	1
	기관지 및 폐	-	-	-	1	0	1
	자궁경부	-		-	0		0

출처 : 보건복지부 보건의료정책본부/건강정책관 암정책팀 (☎ : 031-440-9134)

주) 출처 : 한국중앙암등록본부 국립암센타 『한국중앙 암등록사업 연례보고서』

서비스 기관 : 통계청

(1) 각 년도별로 3차원 분할표(연령별, 장기별, 남녀별)를 작성하여 보라.

(2) 각 년도별로 작성된 (1)의 분할표를 이용하여 점차트, (병렬)막대그림, 모자이크그림 등의 그림을 이용하여 분할표의 특징을 밝혀라.

(3) 각 년도별로 나타나는 자료의 특징이 년도가 바뀜에 따라 어떤 변화가 일어나는가?

(4) 각 년도별로 2차원 분할표(장기별-연령별, 장기별-남녀별)를 작성한 후 점차트, (병렬)막대그림, 모자이크그림 등의 그림을 이용하여 분할표의 특징을 밝혀라.

(5) 각 년도별로 2차원 분할표(장기별-연령별, 장기별-남녀별)를 작성한 후 카이제곱검정을 행하여 보아라. 우리는 무엇을 알 수 있나?

참고문헌

Cleveland, W. S.(1985). *The Elements of Graphing Data*, 1994 2nd edition, Wardsworth.

Cleveland, W. S.(1990). A model for graphical perception, *1990 Proceedings of the section on Statistical Graphics, American Statistical Association*, 1-24.

Cleveland, W. S. and McGill, R.(1984). Graphical perception: theory, experimentation, and application to the development of graphical methods, *Journal of the American Statistical Association*, **79**, 531-554.

Cleveland, W. S. and McGill, R.(1986). An experiment in graphical perception, *International Journal of Man-Machine Studies*, **25**, 491-500.

Cleveland, W. S. and McGill, R.(1987). Graphical perception: the visual decording of quantitative information on graphical displays of data, *Journal of the Royal Statistical Society*, **A150**, 192-229.

Gal, I.(2002). Adults' statistical literacy: meanings, components, responsibilities, *International Statistical Review*, **70**, 1-51.

Jaffe, A. J. and Spirer, H. F.(1987). *Misused Statistics*, Marcel Dekker, Inc.

Mahon, B. H.(1977). Statistics and decisions: the importance of communication and the power of graphical presentation, *Journal of the Royal Statistical Society*, A140, 298-307.

Playfair, W.(1801). *Commercial and Political Atlas and Statistical Breviary*. 2005 edition by wainer, H. and Spence, I., Cambridge University Press.

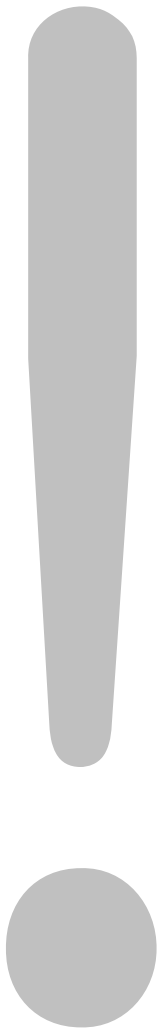
Rao, C. R.(2003). *혼돈과 질서의 만남*, 이재창 옮김, 나남출판사.

Schmid, C. F.(1992). *Statistical Graphics: Design Principles and Practices*, Krieger Pub. Co.

Tufte, E. R.(1990). *Envisioning Information*, Graphics Press.

Tufte, E. R.(1997). *Visual Explanations*, Graphics Press.

Tufte, E. R.(2001). *The Visual Display of Quantitative Information*, 2nd edition, Graphics Press.



저자소개

김영일
서울대학교 고고/인류학과
미국 Minnesota대학교 경영학박사
중앙대학교 정보시스템학과 교수

장대홍
서울대학교 계산통계학과
서울대학교 통계학박사
부경대학교 수리과학부 통계학전공 교수

이태림
서울대학교 계산통계학과
중앙대학교 통계학박사
방송통신대학교 정보통계학과 교수

강명희
이화여자대학교 교육공학과
미국 Indiana대학교 교육공학박사
이화여자대학교 교육공학과 교수