발 간 등 록 번 호
11-1240162-000021-01

통계기초 및 활용

http://sti.kostat.go.kr

통계교육원
STATISTICAL TRAINING INSTITUTE

통계기초 및 활용

1부

통계학으로 통계 읽기

1부. 통계학으로 통계 읽기

목차

통계학으로 통계 읽기 과목의 개요

학습 목표

공식통계를 작성하기 위하여 자료를 생성, 정제, 기술, 요약, 정리 하는데 활용되는 통계학을 학습하여 보다 정확하고 타당한 자료를 생성하고, 그 자료로 부터 작성되는 표와 그래프를 통하여 자료에 대한 검토와 그 자료가 생산된 모집단의 관심 있는 특성과, 현상을 왜곡 없이 이해할 수 있는 능력을 제고하는 것을 목표로 한다.

선수학습

자료 생성 과정이나 통계작성 과정에 관한 진행되는 교육을 받은 수준이면 좋다.

주요 용어

통계, 통계학, 모집단, 모수, 표본, 통계량, 자료의 척도, 통계표, 그래프, 상관계수, 다섯숫자요약, 이상치

학습과목의 내용요약

이 과목에서는 연수대상자들에게 이미 인각된 통계와 통계학의 모습을 토의를 통해서 드러나게 하면서 두 분야를 연결시킨다. 국가통계의 기획부터 활용단계까지 통계학과 관련된 내용이 포함되어 있는 것을 실례로 보이면서 느끼도록 하여 교과목 전반에 대한 학습동기를 제고하려고 한다.

2장에서는 모집단과 표본, 전수조사 자료, 표본조사 자료, 모수와 통계량, 실험자료, 행정자료 등을 구별하도록 하여 연수생 자신이 관심을 갖고 있는 통계의 대상 모집단과 모수를 확인하도록 하고 관련 자료의 수집 방법에 대해서 알게 하여 3장 이후 학습하게 되는 기술통계 부분에 자료 정리의 목표를 놓치지 않도록 한다.

3장에서는 자료수집 단계에서 요구되는 필수개념인 개체와 변수, 그리고 변수의 종류를 설명하고 사례로서 엑셀과 SPSS에서 나타나는 개체와 변수를 설명하여 구별하도록 한다.

4장에서는 특별히 몇 개의 숫자로서 자료를 요약하여 자료의 특성을 전달하는 것에 주안점을 두고 토의한다. 5장에서는 그래프를 이용한 자료의 기술을 다루었고 6장과 7장에서는 다변량 변수 간 관계를 찾아보는 방법으로 연속형 변수에서는 상관계수, 회귀선, 이산형 변수에서는 교차표 등을 제시하고 교락개념과 Simpson's paradox를 소개하였다.

8장에서는 자료가 수집(획득)되어서 본격적인 분석을 시작하기 전에 반드시 거쳐야 할 사전단계인 이상치탐색에 대하여 토의한다.

9장에서는 다소 어렵지만 실용적이라고 판단하여 가중치 주제를 다룬다. 단 국가통계분야에서 가장 잘 언급되는 3가지에 국한하여 개념적인 수준에서 설명한다. AHP 기법과 표본론에서 사용하는 사후가중치와 물가지수에서 사용하는 가중치를 소개하였다.

제 1 장

통계와 통계학

1-1. 통계의 정의

학습목표

- **일반적으로 알고 있는 통계에 대한 생각들을 학습자들이 서로 공유하도록 하고 통계학과 차이를 이해한다.**

1 우리 안에 있는 "통계"의 의미를 생각해보자.

1. 통계와 연관된 단어들

우리는 이미 초등학교 시절 아니면 그 이전부터 통계라는 이름의 교과목을 이수해 왔다.

전국의 공무원, 교사 등 상당히 다양한 영역에서 일하는 사람들로부터 통계 키워드와 익숙한 통계를 조사하여 자료를 수집해 보았는데 그 내용들은 다음과 같이 분류할 수가 있었다.

(1) 학교 교육 관련

- 표와 그래프, 신뢰구간
- 가설 검정, 분석, 비교, 평균

(2) 활용 분야 관련

- 인구, 물가, 조사, 통계청

(3) 신뢰 관련

• 불신, 엉터리

(4) 두려움 관련

• 배우기 어렵다, 복잡하다

2. 익숙한 통계들

(1) 생활 관련

• 물가지수, 인구통계, 실업률 통계, 베이비 붐 세대 통계

(2) 사회 관련

• 부동산 가격 통계, 교육 통계, 청소년 자살 통계, 체력 통계

(3) 경제 관련

• 환율, 증권

(4) 보건 관련

• 건강 보험 관련 통계

❷ 통계의 정의

1. 통계법에서 정의한 통계

우리나라 통계법에서 정의하는 '통계'는 통계작성기관이 정부정책의 수립·평가 또는 경제·사회현상의 연구·분석 등에 활용할 목적으로 산업·물가·인구·주택·문화·환경 등 특정 집단이나 대상 등에 관하여 직접 또는 다른 기관이나 법인, 단체 등에 위임 위탁하여 작성하는 수량적 정보(통계법 제3조 제1호)로서 통계는 주로 숫자(numbers)를 의미하는 것으로 되어있다.

2. UN에서 통용되는 통계

최봉호(2015)는 그의 저서에서 UN에서 만들어진 통계영역의 표준분류 체계를 소개하였다. 그는 분류체계가 만들어진 배경을 '통계학'과 '통계'에 대해 사람들마다 생각하는 바가 다른데서 발생하는 논란 때문에 각국의 통계기관장들을 중심으로 통계의 영역에 대한 표준분류를 제정하려는 노력이라고 설명하였다. 이 통계영역 표준분류는 UN 통계위원회의 최종적 승인을 받지는 않았지만, 국제기구들은 이를 여러 분야에서 많이 활용하고 있다.

그는 이 책에서 '통계영역 표준분류' 체계를 1) 인구 및 사회 분야의 통계, 2) 경제 분야의 통계, 3) 자원 및 환경 분야의 통계, 4) 통계방법론 분야, 5) 통계조직 전략 및 관리 분야 등 크게 다섯 분야로 나누어 소개하였다. 이들 분야를 좀 더 세분화해 보면 인구 및 사회분야의 경우 인구동태통계, 인구센서스 및 특수집단별 인구통계, 주택통계, 노동통계, 교육/직업훈련 통계, 문화통계, 가구 소득/지출/분배 통계, 사회보장통계, 보건통계, 여성통계, 기타(범죄통계 등)로 나누어진다. 경제분야의 통계는 국민계정, 농업/임업/어업, 산업, 에너지, 도/소매업(유통통계), 국제무역(상품/서비스), 교통, 정보통신, 관광, 기타서비스, 금융/보험, 재정, 국제수지, 물가, 과학/기술/특허통계, 기타 경제통계가 있다. 자원 및 환경 분야의 통계로는 자원/환경통계, 자원/환경계정, 기상통계가 있으며, 통계 방법론 분야는 메타 데이터, 각종 표준분류, 자료 소스, 데이터 내검/연계, 데이터 보급/관리, 데이터 비밀보호/제공, 데이터 분석으로 나누어진다. 마지막으로 조직 전략 및 관리 분야는 통계조직 구축 및 원칙 설정, 통계활동의 조정, 통계품질관리, 성과의 측정, 인적자원 개발/관리, 외국/국제기구와의 협력, 개발도상국 기술협력/역량강화 지원 활동으로 세분화된다.

1-2. 통계학의 정의

학습목표

- **학습자들이 가지고 있는 통계학에 대한 선입견, 편견, 두려움 등을 깨닫게 하며 통계학의 주관심사가 변이, 다양성이라는 점을 주지한다.**

1 학교수학에서 느낀 "확률과 통계"

즐거웠던 초등학교 시절의 표와 그래프, 그리고 중·고교 시절에 힘들게 공부했던 경우의 수, 확률, 정규분포 등 우리는 이미 학교 교육에서 비록 즐겁지만은 않았지만 상당 부분의 통계에 접촉한 바가 있다. 이 절에서 우리가 접촉했던 통계가 어떠한 것들이었는지를 편안한 마음으로 돌아보면서 자신이 언제 어느 부분에서 통계에 대해 두려움이 있었는지, 긍정적 호기심이 있었는지를 기억하고 이 인식을 바탕으로 통계학을 정의해보고자 한다. 현재는 통계가 수학의 한 분야로 있기 때문에 2009년도에 개정된 수학과 교육과정을 검토해야 한다. 강현영(2015)이 "통계 교육 활성화를 위한 수학 교육과정 개선 방안 연구"에서 검토한 바에 의하면 초등학교 1학년에서 고등학교까지 확률과 통계 영역을 다음과 같이 설명하고 있다.

초등학교 1-2학년군과 3-4학년군에서는 통계와 관련하여 기본적인 그래프를 다루도록 한다. 초등학교 5-6학년 군에서 확률은 '가능성'으로 간단하게 도입되지만 이후 내용의 대부분은 이전 학년군에서와 마찬가지로 그래프와 관련된 내용이 주류를 이룬다. 초등학교 확률과 통계 교육과정에서는 주로 자료의 표현 부분을 다루는데 최근 들어 통계청의 통계활용대회 등에 힘입어 통계적 문제해결 과정의 '자료 분석' 부분도 많이 다룬다.

중학교 1학년에서도 도수분포표와 히스토그램, 도수분포다각형, 줄기와 잎 그림 등 역시 자료의 표현과 관련된 내용이 중점적이다. 중학교 2학년에서는 초등학교 5-6 학년군에서 가능성과 관련하여 비형식적으로 도입된 확률이 경우의 수, 확률의 계산 등을 통하여 이론적으로 다루어진다. 중학교 3학년에서는 '변이'와 관련된 내용이 지도되는데 평균, 중앙값, 최빈값 등의 대푯값과 범위, 편차, 분산, 표준편차 등의 산포도가 주요 내용이다.

고등학교에서 확률과 통계는 별도의 선택 교과로 편성되어 고등학교 1학년 과정을 이수한 학생이면 학습할 수 있도록 하였다. 그 중 여론조사 발표에 반드시 나타나는 신뢰수준, 표본오차, 허용오차나 수능 성적표에 나오는

표준점수, 정규분포 등의 내용은 현업이나 현실에서 많이 사용되고 있다.

그러나 고등학교 과정에서 배우는 확률과 통계 교과는 본격적인 확률과 통계로 들어가기에 앞서서 순열과 조합에 많은 시간을 할애한다. 그러나 이러한 내용은 통계적 문제 해결 과정인 '문제제기', '자료수집', '자료분석', '결과해석'의 단계에서 필요한 통계적 사고방식을 키우는 것과는 관련성이 약하고 현실의 통계적 문제 해결을 위한 도구로 많이 활용되지는 않는다. 또한 확률과 통계 단원의 전개방식이 다양한 자료를 다루어보기 전에 학문적인 통계 용어의 정의로부터 시작이 되어 학생들이 통계가 어렵고 복잡한 것이라는 편견이나 선입견을 가지게 하기도 한다.

❷ 표와 그래프를 작성하는 이유

"왜 표와 그래프를 그려야 하는가?" 또는 "어떤 상황에서 표와 그래프를 그리고 싶은 마음이 생겼나?"를 생각해보자. 여러 사람들과 대화해 본 바로는 분명한 답을 가지고 있지 않은 사람이 많이 있었다. 또 많은 사람들은 "자료를 효과적으로 빨리 전달하기 위하여 표와 그래프를 작성한다"고 대답하였다. 그렇다면 이 대답은 우리 앞에 있는 자료가 효과적으로 빨리 전달하기 어려운 상태에 있다고 할 수 있다. 그 상태는 어떤 상태라고 표현할 수 있을까? 바로 이 상태가 통계학을 필요로 하는 상태라고 할 수 있을 것이다. 그래서 우리는 다음의 주제를 생각해본다.

❸ 통계학의 주 관심대상

1. 학문분야와 관심대상

통계학을 이야기하기 전에 먼저 대학에 있는 많은 전공학과를 생각해 보면서 이들 분야를 전공하지 않은 사람으로서 이들 학과에서 무엇을 주 관심 대상으로 하고 있는지 상식적으로 이야기 해 보자.

대표적으로 통계학과가 많이 소속되어 있는 경상·경영계열의 학과들부터 생각해보자. 경제학과는 우리 사회의 경제생활과 경제구조를 주 관심대상으로 하고, 회계학과는 기업의 회계 상태에 관심을 두고, 무역학과는 무역에, 부동산학과는 부동산에 관심을, 경영학과는 기업의 경영을 주 관심대상으로 한다. 한편 국립대학에 있는 통계학과가 소속된 자연과학대학

의 전공학과를 보면 생물학과는 생물을, 화학과는 화학을, 물리학과는 물리 현상을, 천문우주학과는 천문우주를 주 관심대상으로 한다. 이렇게 대학에 설치되어 있는 대부분의 전공학과의 주 관심대상이 무엇인지는 그 학과를 전공하지 않은 사람들뿐 아니라 중학생들도 상식적으로 쉽게 말할 수 있다.

그렇다면 통계학과는 무엇을 주 관심대상으로 하고 있는가를 생각해보자. 이런 질문에 대해서 많은 사람이 통계라고 답한다. 그런데 막상 통계학과가 주목하는 대상은 통계가 아니라고도 할 수 있다. 왜냐하면 복지통계를 주 관심대상으로 하는 학과는 어느 학과일까? 사회복지학과가 아니겠는가? 경제통계를 주 관심대상으로 하는 학과도 통계학과보다는 경제학과일 것이다. 통계의 꽃이라고도 할 수 있는 인구통계는 인구학과 또는 사회학과의 주 관심대상일 것이다. 또 다른 사람들은 통계학과의 주 관심대상을 "조사", "자료"라고도 하는데 조사나 자료를 실제로 중요하게 수행하는 사람들은 통계학 전공자들보다 조사내용이나 자료의 출처와 관련된 분야(경제, 사회, 환경, 산업분야 등)를 전공한 사람들이다. 이렇게 생각해보면 조사나 자료도 통계학과의 주 관심대상이라고 하기에는 무엇인가 부족하다.

2. 통계학의 주 관심(연구)대상

앞에서 이야기 한 바와 같이 통계학의 주 관심대상은 통계라고도 할 수 있지만 또 통계라고하기에는 만족스럽지가 않다. 다른 말로 하면 통계학이 가장 중요하게 관심을 갖고 있는 대상에 대하여 우리는 많이 생각해보지 않았을지도 모른다. 이런 관점을 가지고 "그래프를 왜 그리는가?"에 관한 이야기를 다시 기억해보자. 그래프를 그리고 싶은 마음이 생기게 만든 상태가 바로 통계학의 대상이라고 이야기 할 수 있다고 하였다. 현대통계학의 초석을 만들었다고 할 수 있는 Fisher의 생각을 이기원(2001)은 다음과 같이 정리하였다.

통계학이 무엇을 연구대상으로 삼고 있는가에 대한 피셔(Ronald Aylmer Fisher, 1890-1962)의 생각을 원문대로 살펴보면 다음과 같다.

"The science of statistics is essentially a branch of Applied Mathematics,

and may be regarded as mathematics applied to observational data. … Statistics may be regarded (i) as the study of population, (ii) as the study of variation, (iii) as the study of methods of the reduction of data."

즉, 통계학이란 관찰 자료에 수학적 원리를 적용하는 응용수학의 한 분야로서 모집단(population), 변분(variation), 자료축약방법(methods of data reduction)을 연구대상으로 하는 학문이라고 설명하고 있다. 첫 문장에 등장하는 observational data는 관찰로 얻은 자료만을 의미한다기보다는 넓은 의미로 실험계획이나 조사연구를 통하여 얻은 자료 모두를 포함한다고 보아야 한다.

Fisher는 통계학의 연구대상을 모집단과 다양성(variation, 변동, 변이, 변분, 불확실성), 자료축약방법이라고 말하였다. 그런데 모집단을 연구대상으로 하는 사람은 그 모집단의 특성에 관심을 가지고 있는 분야를 전공한 사람들이기 때문에 통계학의 대상이라고 하기에는 무리가 있다. 그럼에도 불구하고 Fisher가 모집단을 통계학의 연구대상이라고 한 것은 정확히 말하면 모집단이 가지고 있는 다양성(실험에서는 불확실성)이며 또 자료축약방법을 연구하는 이유도 자료 속에 나타난 다양성 때문일 것이다. 따라서 통계학의 대상은 모집단이나 현상에 포함된 다양성(variation, 변동, 변이, 변분, 불확실성)이라고 할 수 있겠다. 이러한 관점에서 보면 그래프를 그리고 싶도록 만든 상태 역시 자료가 내포하고 있는 다양성이라고 표현할 수 있겠다.

4 통계학의 다양한 정의

통계에 대한 정의는 그야말로 다양하다. 허명회(2015)는 통계의 정의가 각인각색으로 다양하다고 표현하며 자신이 공감하는 세 가지를 제시하였는데 적절하게 표현되었다고 생각한다.

- 통계학은 데이터의 수집, 분석, 그리고 추론에 관한 과학과 기술이다.

  [Data Science and Technology]

- 통계학은 자연과학과 사회과학을 망라한 모든 경험과학의 언어이다.

  [Language of All Empirical Science]

- 통계학은 복잡계의 기술과 이해를 위한 정량적 방법론이다.

  [Quantitative Methodology for Complex System]

한편 많은 통계학 교재에 나와 있는 정의를 하나 더 소개하면 미국통계협회(ASA) 회장을 역임한 Jon Kettenring이 미국통계협회 홈페이지에서 정의한 것이다. 그는 통계학이란 조사와 실험에 대한 설계, 데이터의 수집, 처리, 분석, 결과의 해석을 행하는 과학이라고 정의하고 있다.

이들 모두를 포함하여 또 다르게 표현하면 통계학은 데이터를 생산하고 이해하는 논리와 방법들을 제공하는 학문이라고 하겠다.

1-3. 통계속의 통계학

학습목표

• 통계를 기획, 작성, 활용하는 전 과정에 통계학이 사용되고 있는 것을 인지한다.

❶ 통계기획단계에서 사용되는 통계학

1. 사회조사에서 사용되는 통계학

사회조사는 통계청이 국민의 삶의 질과 관련된 사회적 관심사와 주관적 의식에 관한 사항을 조사하여 삶의 수준과 사회적 변동을 파악하고 이를 사회개발 정책의 기초 자료로 제공하기 위하여 실시하는 조사인데, 2014년에는 「보건」, 「교육」, 「안전」, 「가족」, 「환경」 부문에 대한 조사를 하였다. 다음은 2014년 보고서에 수록된 조사개요 부분이다.

> 이 조사의 조사 대상 시점은 2014년 5월 15일이고, 조사 기간은 2014년 5월 15일부터 5월 30일까지 16일간 이었다. 조사대상은 전국 17,664 표본가구내의 만 13세 이상 가구원 37,000여명 이었다. 표본가구를 추출하는 과정은 시도별로 독립적 추정이 가능하도록 서울, 부산, 대구, 인천, 광주, 대전, 울산의 7대 도시 및 9개도의 동부, 읍면부 등 모두 25개 지역으로 층화하고, 조사구별로 주택유형, 농가비율, 유배우 비율, 1인 가구 비율, 60세 이상 인구 비율, 자가 비율 및 행정구역 등 지역별 분류순서를 정하여 조사구명부를 정렬하고 이를 층별로 가구 수(MOS)를 기준으로 확률비례추출방법(PPS: probability proportional to size)을 이용하여 조사구를 추출하였으며, 표본조사구의 가구에 일련번호를 부여한 후 랜덤으로 최초가구를 설정하여 그 가구를 포함해서 연속하여 12가구를 조사하는 방법으로 표본가구를 최종 확정하였다.

위의 조사개요 부분을 이해하기 위해서는 다음과 같은 통계학적 질문을 할 수 있을 것이다.

질문 1) 독립적 추정이 가능하다는 말이 무슨 뜻인가?

질문 2) 25개 지역으로 층화하였다는 의미는 무엇인가?

질문 3) 확률비례추출방법(PPS)의 의미는 무엇인가?

2. 경제활동인구조사

경제활동인구조사의 목적은 국민의 경제활동 즉, 국민의 취업, 실업 등과 같은 경제적 특성을 조사하여 거시경제 분석과 인력자원의 개발 정책 수립에 필요한 기초 자료인 노동공급, 고용구조, 가용노동시간 및 인력자원의 활용정도를 제공하고 정부의 고용정책입안 및 평가에 필요한 기초 자료를 제공하는데 있다.

… 또한 경기변화를 제대로 반영할 수 있도록 계절조정 실업률을 작성함으로써, 자료이용의 효율화를 기하면서 국제비교를 더욱 용이하게 하였다.

… 1999년 7월에는 1995년 인구주택총조사 결과를 기준으로 작성된 추계인구를 기초로 1991년 1월 이후 자료에 대해 시계열을 보정함으로써 고용통계의 현실반영도를 제고하였으며, 실업자의 구직기간을 4주간으로 확장한 구직기간 4주기준 실업통계를 작성 · 발표하였다.

… 그리고 응답자의 부담경감을 위해 전국적으로 연동표본을 도입, 적용하였다.

… 2008년 1월에는 응답자 부담 경감, 다양한 구직경로 파악 등을 위해 전문가 의견수렴 등을 거쳐 조사항목을 일부 축소 및 수정하였으며, 인터넷(CASI)조사 도입으로 응답편의를 도모하였다.

2009년 7월에는 전화면접(CATI)조사를 도입하는 등 다양한 조사방법 적용으로 조사와 응답 편의를 도모하였으며 2014년 2월에는 2013년 계열을 추가하여 계절조정인자를 재작성함에 따라 1999년 6월 이후 계절 조정계열 보정을 하였다.

위의 조사 개요를 이해하기 위하여 다음과 같은 질문들을 할 수 있을 것이다.

질문 1) 계절조정 실업률, 계절조정인자는 무엇을 의미하는가?

질문 2) 시계열을 보정함으로써 고용통계의 현실반영도를 제고한다는 말의 의미는 무엇인가?

질문 3) 연동표본은 무엇을 의미하는가?

질문 4) 인터넷(CASI)조사와 전화면접(CATI)는 무엇인가?

2 통계작성단계에서 사용되는 통계학

1. 통계표 속의 용어 이해

다음은 15세 이상 인구 및 경제활동인구를 나타내는 표이다.

<표 1-1>
경제활동인구조사
통계표 (일부)

< 15세 이상 인구 및 경제활동인구 >
< Population aged 15 and over and Economically Active Population >

(단위: 천명, %, 전년동월대비) **(Unit : Thousand persons, %, the same month last year)**

	2014. 8 2014. 8		2015. 7 2015. 7		증감 증감	증감률 증감률	2015. 8 2015. 8		증감 증감	증감률 증감률
		구성비		구성비				구성비		
• 15세이상인구 (Population aged 15 & over)	42,571	100.0	43,055	100.0	527	1.2	43,086	100.0	515	1.2
남자(Male)	20,827	48.9	21,079	49.0	277	1.3	21,098	49.0	271	1.3
여자(Female)	21,745	51.1	21,976	51.0	250	1.2	21,988	51.0	244	1.1
- 경제활동인구 (Econo. active pop.)	26,775	100.0	27,303	100.0	413	1.5	27,064	100.0	290	1.1
남자(Male)	15,497	57.9	15,736	57.6	189	1.2	15,607	57.7	110	0.7
여자(Female)	11,278	42.1	11,568	42.4	224	2.0	11,457	42.3	179	1.6
- 비경제활동인구 (Econo. inactive pop.)	15,797	100.0	15,751	100.0	114	0.7	16,022	100.0	225	1.4
남자(Male)	5,330	33.7	5,343	33.9	88	1.7	5,490	34.3	161	3.0
여자(Female)	10,467	66.3	10,408	66.1	26	0.2	10,532	65.7	64	0.6

이 표에 포함된 통계학적 요소를 찾아보자. 정확한 통계표를 작성하기 위해서는 구성비, 증감, 증감률, 전년동월대비 등의 개념을 이해하여야 한다.

2. 자료처리단계에서 사용되는 용어 이해

다음은 2013년 체력실태조사보고서의 일부이다.

가. 불량 자료의 제거

- 1차 : 검사감독관의 평가에 의한 자료 검색
- 2차 : 기록 불량 자료 제거
- 3차 : Coding 불량 자료 제거
- 4차 : 입력 자료 전산 프로그램에 의한 제거

나. 기술통계 분석

- 성별, 연령별, 지역별 평균 및 표준편차 제시
- 성별, 연령별, 백분위 점수 제시
- 막대 및 꺾은선 그래프를 사용한 도식화
- 연령 간 차이 분석
- 연도별 변화 추이 분석

다. 체력 항목별 기준치 설정

- 성별, 연령별 5단계 국민체력 평가 기준 설정
  - 각 항목의 기준치 설정 작업을 위하여 각 항목의 정규분포를 K-S(Kolmogorov-Smirnov)검정을 통해 분석하였으며, 기준치는 각 항목별 percentile을 이용하여 설정
  - 5단계 평가기준은 89년부터 실시해온 방식을 유지하여 연도별 비교가 가능하도록 각 구간의 분포는 10%, 22%, 36%, 22%, 10%로 설정
- 전문가회의를 통하여 성별, 연령별 건강체력 기준치를 30백분위수로 설정 기존 기준치 유지
- 자신의 수준을 평가할 수 있도록 측정 항목별 백분위 분석(부록 1 참고)

위의 내용에서 볼 수 있듯이 기록 불량 자료 제거, coding, 평균 및 표준편차, 백분위점수, 막대 및 꺾은선 그래프, 연도별 변화 추이, 정규분포, K-S 검정, percentile, 30백분위수 등을 알아야 한다.

❸ 통계활용단계에서 사용되는 통계학

1. 사회조사 통계표 활용

다음은 2014년 사회조사 보고서의「보건」,「교육」,「안전」,「가족」,「환경」부문에 대한 조사결과 중 건강평가에 대한 설문문항과 분석결과이다. 설문문항은 "귀하의 전반적인 건강 상태는 어떠하십니까?"이며, 응답항목은 "①매우 좋다/ ②좋은 편이다/ ③보통이다/ ④나쁜 편이다/ ⑤매우 나쁘다"의 다섯가지로 되어 있다. 분석결과는 자신의 건강상태를 좋다고 평가한, 즉"①매우 좋다"와 "②좋은 편이다"로 응답한 사람의 비율이다.

[그림 1-1]
2014 사회조사 설문지
(일부)

건 강 평 가

7 귀하의 전반적인 건강 상태는 어떠하십니까?

1 매우 좋다
2 좋은 편이다
3 보통이다
4 나쁜 편이다
5 매우 나쁘다

[그림 1-2]
2014 사회조사 보고서
(일부)

1. 건 강 평 가

(단위: %)

	좋 다 Good				
	추정치 Estimate	표 준 오 차 S.E.	상 대 표준오차 C.V.	95% 신뢰구간(C.I.) 하 한 Lower	95% 신뢰구간(C.I.) 상 한 Upper
전 국 Whole country	48.7	0.5	1.0	47.7	49.6
동·읍면부 Total for dongs, eups& myeons					
도 시(동 부) Total for dongs	49.7	0.5	1.0	48.7	50.7
농어촌(읍면부) Total for eups & myeons	43.7	1.1	2.4	41.7	45.8
성 별 Gender					
남 자 Male	52.7	0.6	1.1	51.6	53.8
여 자 Female	44.7	0.5	1.2	43.7	45.8
연 령 Age					
13 ~ 19 세 years	74.7	0.9	1.1	73.1	76.4
20 ~ 29 세 years	67.1	1.0	1.4	65.2	69.0
30 ~ 39 세 years	56.2	1.0	1.7	54.3	58.0
40 ~ 49 세 years	47.3	0.8	1.7	45.7	48.8
50 ~ 59 세 years	41.2	0.8	2.0	39.6	42.8
60 세 이 상 years and over	23.9	0.7	2.8	22.6	25.2
65 세 이 상 years and over	20.9	0.7	3.4	19.5	22.3
교 육 정 도 Educational attainment					
초 졸 이 하 Elementary school graduate & under	30.9	0.7	2.4	29.4	32.3
중 졸 Middle school graduate	46.2	0.9	2.0	44.4	48.1
고 졸 High school graduate	48.6	0.6	1.3	47.3	49.8
대 졸 이 상 College or university graduate & over	58.0	0.7	1.2	56.7	59.3
직 업 Occupation					
전 문 관 리 Pro., Tech., Managers	58.0	1.0	1.7	56.0	60.0
사 무 Clerks	58.0	1.1	1.9	55.8	60.2
서 비 스 판 매 Service & Sales workers	49.7	1.0	2.0	47.7	51.6
농 · 어 업 Agri., Fishery workers	34.9	1.4	4.1	32.1	37.7
기 능 노 무 Craft, Operator, Assemblers	45.5	0.9	1.9	43.8	47.2
가 구 의 월평균 소득 Average monthly household income					
100만원미만 Less than 1000 thousand won	26.3	0.9	3.5	24.5	28.1
100 ~ 200 1000~2000 thousand won	42.8	0.8	1.8	41.2	44.3
200 ~ 300 2000~3000 thousand won	50.5	0.8	1.6	48.9	52.1
300 ~ 400 3000~4000 thousand won	54.8	1.0	1.8	53.0	56.7
400 ~ 500 4000~5000 thousand won	55.7	1.2	2.1	53.3	58.0
500 ~ 600 5000~6000 thousand won	56.9	1.4	2.5	54.0	59.7
600만원이상 6000 thousand won and over	64.1	1.1	1.8	61.9	66.3
서 울 Seoul	49.8	1.2	2.4	47.5	52.1
부 산 Busan	49.7	1.7	3.4	46.3	53.0
대 구 Daegu	43.4	1.9	4.3	39.7	47.0
인 천 Incheon	51.9	1.6	3.1	48.7	55.1
광 주 Gwangju	50.9	1.9	3.8	47.1	54.7
대 전 Daejeon	54.1	1.7	3.2	50.7	57.5
울 산 Ulsan	52.1	2.1	4.0	48.0	56.2
경 기 Gyeonggi	48.0	1.2	2.4	45.7	50.3
강 원 Gangwon	48.2	1.8	3.8	44.6	51.8
충 북 Chungbuk	44.6	1.8	4.1	41.0	48.2
충 남 Chungnam	50.0	1.9	3.8	46.3	53.7
전 북 Jeonbuk	50.1	2.0	4.0	46.1	54.0
전 남 Jeonnam	46.0	1.8	3.9	42.4	49.5
경 북 Gyeongbuk	43.1	2.1	5.0	38.9	47.3
경 남 Gyeongnam	50.3	1.6	3.1	47.2	53.4
제 주 Jeju	42.1	1.6	3.8	39.0	45.3

위의 표의 수치를 보고 다음과 같은 질문을 대답할 수 있다면 여러 가지 정책을 제안할 수 있을 것이다.

질문 1) 교육정도에 따라 자신의 건강평가가 통계적으로 유의하게 다르다고 할 수 있는가?

질문 2) 서울시민이 자신의 건강상태를 양호하다고 생각하는 비율이 49.8%인 반면 제주도민이 자신의 건강상태를 양호하다고 생각하는 비율은 42.1%이다. 당신은 이 차이를 굉장히 큰 차이라고 판단하는가?

2. 생활시간조사 통계표 활용

생활시간조사는 우리나라 국민들이 하루 24시간 동안 어떤 행동을 언제 얼마나 하는가를 조사하여 국민의 평균적인 생활방식과 삶의 질을 파악하는 기초 자료로 제공하고자 1999년부터 5년 주기로 통계청에서 실시하고 있는 국가통계로 각종 노동, 복지, 문화, 교통 관련 정책수립이나 학문연구의 기초자료로 활용된다. 다음은 2009년과 2014년 생활시간조사 결과 중 일부로 해당연도별 연령대에 따른 독서시간을 나타낸다.

<표 1-2>
생활시간조사 결과 (일부)

단위(시간:분, %)

	평일						토요일						일요일					
	전 체 평균시간		행위자 비 율		행위자 평균시간		전 체 평균시간		행위자 비 율		행위자 평균시간		전 체 평균시간		행위자 비 율		행위자 평균시간	
	2009	2014	2009	2014	2009	2014	2009	2014	2009	2014	2009	2014	2009	2014	2009	2014	2009	2014
전체	0:07	0:06	11.3	9.7	0:59	1:05	0:09	0:08	13.2	10.2	1:10	1:16	0:10	0:09	14.1	10.9	1:11	1:18
10대	0:10	0:09	22.0	18.2	0:46	0:49	0:19	0:16	29.0	22.5	1:04	1:12	0:19	0:17	28.6	24.3	1:06	1:12
20대	0:09	0:09	13.5	12.8	1:10	1:12	0:11	0:09	14.6	12.1	1:18	1:10	0:13	0:11	16.3	13.8	1:23	1:19
30대	0:07	0:06	11.1	8.9	1:00	1:07	0:09	0:07	12.6	8.9	1:09	1:16	0:09	0:09	14.3	10.7	1:05	1:24
40대	0:06	0:07	9.7	9.9	1:01	1:07	0:08	0:08	10.8	10.2	1:11	1:20	0:09	0:08	12.2	10.3	1:13	1:22
50대	0:05	0:04	7.4	5.7	1:02	1:09	0:06	0:05	8.3	6.3	1:11	1:23	0:06	0:05	8.4	6.1	1:11	1:21
60세 이상	0:04	0:05	5.0	5.9	1:12	1:18	0:05	0:05	5.9	5.7	1:23	1:20	0:04	0:04	5.6	5.1	1:09	1:18
남자	0:06	0:07	10.5	9.5	1:00	1:09	0:10	0:08	13.5	9.8	1:13	1:23	0:10	0:09	13.6	10.8	1:12	1:23
여자	0:07	0:06	12.0	9.9	0:58	1:01	0:09	0:07	13.0	10.4	1:08	1:10	0:10	0:08	14.5	10.9	1:09	1:14

주) 1. 일반 책 및 만화책 읽기(신문, 잡지류는 제외). 소설, 전기, 교육서적 등을 읽는 행동 포함

위의 표를 보고 다음과 같은 질문들을 생각해볼 수 있을 것이다.

질문 1) 우리나라에서 2009년에서 2014년 5년 동안 독서를 하는 사람의 비율이 늘어난 연령대는 어느 연령대라고 생각하는가?

질문 2) 우리나라에서 독서를 하는 60대 이상의 비율은 얼마라고 생각하는가?

참고자료

- 강현영 (2015), 통계 교육 활성화를 위한 수학 교육과정 개선 방안 연구, 한국과학창의재단.
- 미국통계협회, http://www.amstat.org/
- 이기원 (2001), 인터넷 시대의 생활 속의 통계학, 교우사.
- 최봉호 (2015), 현장 중심의 통계조사의 이해, 자유아카데미.
- 허명회 (2011), 법과 통계학, 한나래 아카데미.
- UN(2009) Classification of International Statistical Activities
- 통계청 (2015), 경제활동인구조사 보고서.
- 통계청 (2014), 사회조사 보고서.
- 통계청 (2014), 생활시간조사 보고서.
- 문화체육관광부 (2013), 체력실태조사 보고서.

제 2 장 자료 수집하기

2-1. 모집단과 모수

학습목표

- **다양성이나 불확실성을 내포하고 있는 상황에 대하여 연구할 가치가 있다고 판단하게 되면 좋은 정보를 제공할만한 자료를 구할 수 있는지를 알아보게 된다. 이 절에서는 먼저 연구할 가치가 있다고 판단한 대상을 구체적으로 정의하는 용어인 모집단과 그 모집단의 관심 특성인 모수를 이해한다.**

1 모집단과 모수의 정의

1. 모집단

모집단은 관심대상이 되는 다양성이나 불확실성이 내재된 집단에 속한 모든 개체의 집합으로 정의한다. 그런데 때로는 집단에 속한 개체의 집합으로 정의하지 않고 연구대상이 되는 가능한 관측값이나 측정값의 집합을 모집단(population)이라고 정의하기도 한다. 이 두 경우는 관측대상이 되는 개체를 원소로 보는지 그 개체로부터 측정된 값을 원소로 보는지 하는 관점의 차이이므로 문맥을 통해서 결정하면 된다. 예를 든다면 서울시민의 행복점수를 조사한다고 하자. 이 경우 모집단은 서울 시민의 자격을 가진 – 정확히 하면 특정시점에 주민등록상 서울 시민인 – 사람들이 될 것이다. 그런데 교재에 따라서는 시민들의 행복점수들을 모집단이라고 기술한 경우도 있다는 것이다.

모집단은 구성요소의 특징에 따라 유한모집단과 무한모집단으로 나누게 되는데 유한모집단(finite population)은 모집단의 크기가 유한한 경우이고 무한모집단(infinite population)은 모집단의 크기가 무한한 경우이다.

그런데 현실에 존재하는 관심대상이 되는 대부분의 모집단은 유한모집단이다. 한편 현실에 존재하기 보다는 개념적으로만 존재하는 모집단은 무한모집단으로 간주한다. 예컨대 방사능 노출자의 생존수명에 대한 연구를 한다고 하자. 이 경우에 연구대상이 되는 모집단은 방사능에 노출된 모든 사람이 된다. 이 모집단은 현실적으로 현재 시점에만 국한된 것이 아니라 과거와 미래의 방사능 노출자도 포함하여 연구자의 개념 속에 그려지는 모집단으로 모집단에 속한 개체의 수는 무한이다. 다음 절에서 다루겠지만 이와 같은 무한모집단은 실험을 통하여 자료를 얻게 되는 상황에서 주로 나타난다.

모집단은 실제 현장에서는 목표(표적)모집단 또는 연구모집단이라는 용어와 조사모집단이라는 용어로 구별하여 정의한다. 목표모집단이란 연구(조사)목적에 따라 개념적이고 이론적인 모집단을 의미한다. 그런데 실제 조사를 위해서는 목표모집단을 정확히 파악할 수 없는 경우가 많다. 왜냐하면 이 모집단에는 실제로 조사하기 위하여 접촉이 가능하지 않은 개체들이 포함되어 있기 때문이다. 따라서 개념상 정의된 모집단은 현실적으로 수정되어야 한다. 이와 같이 표본으로 추출이 가능한 개체들로만 구성된 모집단을 조사모집단 또는 조사가능모집단이라고 한다.

2. 모수

모수는 모집단의 특성을 나타내는 값으로 대부분은 숫자이지만 아주 드물게 문자일 수도 있다. 예컨대 어느 지역의 종교는 무엇인가? 라는 질문에는 그 지역에 사는 주민들이 가장 많이 택한 종교를 파악하려고 하는 의도가 있으므로 이 경우에 모수는 가장 빈도수가 높은 종교인 “특정 종교 이름”이 된다. 모수는 고정된 값인데 대부분의 경우에는 현실적으로 그 값을 알지는 못한다. 따라서 통계적 추론을 요구한다. 즉 추정을 하거나 가정을 하게 된다.

예를 들자면 “가구 월평균 소득”에 관심을 갖고 조사를 했다면 조사대상 가구마다 다양한 값을 응답할 것이다. 이 경우에 모집단에 속한 가구의 소득에 관한 관심사는 구체적으로 무엇인가? 연구자들은 지역별 평균 소득, 지역간 소득 격차, 지역별 하위 10%에 있는 가구주 소득 등에 관심을 갖게 될 것이다. 이러한 구체적인 관심사를 모수라고 한다.

❷ 사례검토

위에서 정의한 모집단과 모수를 다음의 국가통계를 통해서 파악해 보자.

1. 사교육비통계

우리나라의 초·중·고 학생들의 사교육비 실태를 체계적으로 조사하여 공신력 있는 통계를 정기적으로 작성·제공하여 사교육비 경감대책 및 공교육 내실화 등 교육정책 수립의 기초자료를 제공하는 것을 조사목적으로 한 사교육비 통계의 경우 모집단과 모수를 생각해보자.

(1) 모집단

이 통계의 조사개요의 표본설계 부분에서 모집단을 다음과 같이 정의하였다.

1. 모집단 정의

- 조사의 모집단은 전국의 초 · 중 · 고 재학생과 그 학부모이며, 가구를 통한 접근은 해당 가구에 초 · 중 · 고 재학생이 있는지 여부를 정확히 파악하기가 어렵고 비용이 많이 들기 때문에 비효율적이므로 학교를 통하여 조사

2. 표본추출틀 작성

- 표본추출틀은 2013년 4월 기준 교육통계정보센터의 학교 DB 상의 학교를 사용. 단, 조사의 현실성을 고려하여 아래에 해당하는 학교 및 학급은 제외하여 학년별 조사모집단(survey population)을 설정

〈 표본추출틀 제외 조건 〉

폐교와 휴교 제외 : 62개교
도서지역에 소재하는 학교 제외 : 756개교
학급당 평균 학생수가 10미만인 경우 제외 : 1,146개교
학교급별 해당 학년의 학급수가 0인 학교는 제외

- 2014년 표본추출틀 : 9,817학교, 226,276학급, 학생 6,394천명

위의 표에서 보는 바와 같이 모집단은 “전국의 초 · 중 · 고 재학생과 그 학부모”로 정의하였지만 가구를 통한 접근의 경우 해당 가구에 초 · 중 · 고 재학생이 있는지 여부를 정확히 파악하기가 어렵고 비용이 많이 들기 때문에 학교를 조사하기로 하였다. 이에 따라 2013년 4월 기준 교육통계정보센터의 학교 DB 상의 학교 중 조사현실성을 고려하여 표의 <표본추출틀 제외조건>에 해당되는 학교를 제외한 학년별 조사모집단을 설정하였다. 이러한 이유로 ‘전국의 초 · 중 · 고 재학생과 그 학부모’로 정의된 모집단을 연구모집단이라 하고 ‘2013년 4월 기준 교육통계정보센터의 학교 DB 상의 학교(제외조건에 포함되지 않은 학교)’를 조사모집단이라고 하였다. 이와 같이 많은 경우 모집단을 두 개로 정의한다.

여기서 표본추출틀(sampling frame)은 표본을 추출하기 위한 표본추출단위들의 목록을 말한다. 표본추출틀을 이용하여 표본을 뽑을 수 있으므로 표본추출틀은 표본조사에 있어서는 꼭 필요한 존재이다. 그러나 위의 예에서 만일 2013년 4월 기준 교육통계정보센터의 학교 DB가 없다면 그 이전에 작성된 DB를 사용해야 할 수도 있다. 이 경우 원래는 포함되어야 할 새로 개교된 학교가 제외되거나, 제외되었어야 할 폐교된 학교가 포함될 수도 있다. 이와 같은 경우처럼 표본추출틀은 수시로 갱신되는 것이 쉽지 않기 때문에 완전하다고 볼 수는 없다. 그러나 표본추출틀이 모집단을 잘 대표한다면 표본조사는 좋은 결과를 얻을 수 있다.

(2) 모수

모수에 대한 기본적인 정보도 많은 경우 조사개요에 나타나는데, 본 사교육비 조사에서는 다음과 같은 내용(일부)을 확인할 수 있다.

조사범위

1. 사교육비

- 주요 과목별로 학원비, 개인 및 그룹과외비, 학습지, 인터넷 및 통신강의(EBS제외) 과외비(교재비 포함)
- 일반교과 및 논술 관련 사교육비 (국어, 영어, 수학, 사회(역사/도덕 포함), 과학, 논술 등, 제2외국어 · 한문 등)

• 예 · 체능 및 취미 · 교양 관련 사교육비 (음악, 체육, 미술, 취미 · 교양)

• 취업 목적 관련 사교육비 (공업계, 상업계, 농업계, 전산계, 기타분야 등)

2. 방과후학교 비용, EBS 교재비 및 어학연수비

• 방과후학교 활동비, EBS 관련 교육비(교재구입비)는 사교육비 경감효과의 정확한 분석을 위해 학교 밖에서 이루어지는 사교육비와 분리조사

2. 2014 서울서베이

서울시의 특성을 파악할 수 있는 통계자료를 생성하고, 각 지역특성을 반영한 시정운영에 활용하고자 2003년부터 매년 20,000가구를 대상으로 실시하고 있는 서울서베이 모집단과 모수를 생각해 보자. 서울서베이는 가구조사, 외국인조사, 국내사업체조사, 외국인투자 사업체조사로 나누어지는데 그 중 가구조사에 대하여만 살펴보기로 한다.

(1) 모집단

2014 서울서베이의 조사개요에는 다음의 내용이 포함되어 있다.

위의 내용에 나타난 것처럼 모집단은 "2014년 10월 서울시 거주 가구 및 만 15세 이상 가구 구성원"으로 정의하고 주민등록DB와 과세대장DB를 연계하여 구/동별, 주택유형별 세대수를 파악하여 구체화 하였다.

1. 조사설계

1) 가구조사

모집단 – 2014년 10월 서울시 거주 가구 및 만 15세 이상 가구 구성원

2. 표본설계

1) 가구조사

◎ 모집단 분석

• 가구조사의 목표 모집단을 다음과 같이 정의함

• 주민등록DB와 과세대장DB를 연계하여 구/동별, 주택유형별 세대수를 파악함

• '건물/기타'의 '근린생활시설'등을 '단독주택/다가구주택, 아파트, 다세대주택, 연립주택/기타' 및 '기타건물'로 재분류함

- 세대주가 서울시에 거주하는 세대

- 주택유형이 '단독주택/다가구주택', '아파트', '다세대', '연립주택/거주용 건물'인 세대

- 최종 표본추출 단위는 세대임

<표 1-1> 2010 인구주택총조사의 서울시 가구수

가구수	단독주택 (공동주택)	아파트	다세대주택	연립주택	비거주용 건물내 주택	주택이외의 거처
3,459,093	1,304,509	1,439,259	442,458	140,566	48,052	84,249
100%	37.7%	41.6%	12.8%	4.1%	1.4%	2.4%

<표 1-2> 주민등록DB+과세대장DB의 서울시 세대수

세대수	단독주택	다가구주택	아파트	다세대주택	연립주택 거주용 건물	기타 비거주용 건물
3,947,660	429,732	1,099,277	1,466,721	533,632	236,662	171,672
100%	10.9%	27.9%	37.2%	13.5%	6.0%	4.3%

<표 1-3> 주택유형별 모집단 특성

세대수	단독주택 다가구주택	아파트	다세대주택	연립주택 거주용 건물
3,776,024	1,529,009	1,466,721	533,632	236,662
100%	40.6%	38.9%	14.2%	6.3%

- 2014년 8월 기준

- '연립주택/주거용 건물'에는 '주거용 오피스텔', '다중주택' 등이 일부 포함되어 있음

- '기타/비거주용 건물'에는 '근린생활시설', '사무실', '교육연구시설', '종교시설', '사무용 오피스텔', '기숙사', '주택용도 불분명' 등이 포함되어 있음

• 모집단크기는 3,776,024세대이며, 주택유형별 비율은 단독주택/다가구주택 40.6%, 다세대주택 14.2%, 연립주택/거주용건물 6.3%임

(2) 모수

제2절 조사 내용

- 격년주기 순환조사 실시에 따른 모듈설계
  - 모듈 A(2013) : 여성가족, 복지, 재난/안전, 정보참여, 가치의식
  - 모듈 B(2014) : 관광/여가, 문화, 환경, 교통, 교육
- BSC지표(보육/문화환경/교통/보행만족도) 등 정책지표는 매년 조사

제3절 조사 설계

4. 모수 추정

- 용어정의
- 모평균에 대한 추정
- 모비율에 대한 추정
- 분산의 추정
- 표준오차, 오차한계(표본오차)의 추정

조사내용을 보면 2014년 서울서베이의 모수는 매년 조사되는 BSC지표(보육/문화환경/교통/보행만족도)와 모듈B의 관광/여가, 문화, 환경, 교통, 교육 관련 사항에 대한 모수 평균, 비율, 분산, 표준오차 등을 추정하고 있는 것을 알 수 있다.

2-2. 자료수집 방법

학습목표

- **연구할 가치가 있다고 판단한 모집단과 그 모집단의 특성인 모수가 정의되면 관련된 유용한 정보를 얻기 위한 자료를 수집하기 위한 방안을 검토하게 된다. 이 절에서는 자료를 수집하는 방법들로 관측(관찰 또는 측정), 실험 그리고 이미 존재하고 있는 행정자료를 이용하여 자료를 수집하는 방법을 이해한다.**

1 관측을 통한 자료수집 방법

1. 전수조사

모집단 전체를 조사하는 것을 전수조사(census)라고 하는 데 시간이나 비용 등 현실적인 어려움 때문에 주로 국가기관에서 표본추출틀 생성과 같은 특별한 목적을 갖고 실시한다. 박진우(2006)는 성경 민수기에 보면 B. C. 1500년경 이스라엘민족이 이집트를 탈출하여 나라를 세우기 위하여 팔레스타인으로 들어가기 전에 두 번에 걸쳐 인구조사를 행하였다고 하는 기록이 있고 성경 누가복음에 보면 예수가 마굿간에서 태어나는 사건이 호구조사와 관련되어 있다고 기술하고 있다. 그러나 세계적으로 보면 우리가 알고 있는 정규적인 인구조사는 17세기에 가서야 비로소 시작되었다. 우리나라에서 하는 가장 큰 전수조사는 통계청에서 실시하고 있는 인구 및 주택센서스이다. 이 센서스는 1925년 10월 1일 간이국세조사라는 이름으로 인구조사를 행하였다. 그리고 5년 후 1930. 10. 1 조선국세조사라는 이름으로 최초로 직업 등 경제활동 사항을 포함하여 시행되었고 다시 5년 후 1935. 10. 1 상주지 항목이 추가 된 조사를 하면서 44, 49, 66 년을 제외하고는 모두 0년과 5년에 발전적으로 시행되었고 가장 최근에는 2010년에 시행된 바 있다. 그런데 이 전수조사는 2015년부터는 조사시점에 존재하는 행정자료(등록자료)를 활용하여 통계를 작성하는 등록센서스로 변경되어서 더 이상 전수조사는 하지 않게 되었다.

또 다른 큰 전수조사로는 2011년에 최초 시행된 경제총조사이다. 이 조사는 국가 전체 산업에 대해 하나의 확정된 조사기준과 방법에 의하여 산업구조와 분포, 경영실태 등에 관한 사항을 종합적으로 파악하고 정부의 경제 및 산업별 정책 수립과 기업의 경영계획 수립·평가의 기초자료를 제공하는 것을 목적으로 한다. 국내에서 산업 활동을 하는 종사자 1인 이상인 모든 사업체를 조사대상으로 하는 조사로서 조사주기는 매 5년(0자 및 5

자로 끝나는 연도를 기준으로 실시)이다.

또 하나의 국가적인 전수조사는 농림어업총조사이다. 이 조사는 농림어가 정의에 해당하는 전국의 모든 농가, 임가, 어가의 총수는 물론 개별 특성까지 파악하여 농림어업 정책 및 농산어촌 지역개발계획 수립 및 평가, 각종 학술연구 자료와 표본조사의 표본 틀로 활용하기 위해 실시하는 전국적 규모의 통계조사이다.

국가와 지방자치단체의 정책수립에 필요한 기초자료를 제공하고, 사업체를 조사대상으로 하는 통계조사의 모집단을 제공하기 위하여 매년 300만 개가 넘는 우리나라의 모든 사업체를 대상으로 조사하는 전국사업체조사도 있다. 전수조사가 꼭 대규모 조사인 것은 아니다. 세부산업분야에서 실시하는 전수조사에는 60여개의 원양어업을 주 업종으로 하는 기업을 대상으로 하는 원양산업통계조사도 있다.

2. 표본조사

조재근(2010)이 '통계로 읽는 사회와 경제'에서 조사통계의 역사에 대하여 기술한 것을 요약하여 정리하면 다음과 같다.

조사통계의 역사에서는 1800년경 영국, 프랑스, 미국 등 서구 여러 나라가 거의 동시에 근대적인 총인구조사, 즉 센서스를 실시한 시기로 기록하고 있다. 여러 국가의 정부는 5년이나 10년 주기로 정기적인 센서스를 실시하였는데 초기에는 인구를 추정하는 집계수준이었다. 그리고 같은 시기에 민간단체가 주축이 되어 사회통계를 생산하고 있던 것으로 기록되고 있다. 이러한 움직임은 1830년쯤에 절정에 이르게 되는데 이 시대 통계의 역사를 연구한 학자들은 1830년대부터 약 20여 년간을 "the age of enthusiasm in statistics"라고 부른다. 즉 당시 사람들이 보여준 통계에 대한 놀라운 열정을 일컬어 그렇게 부른 것이다. 그런데 그 "열광의 시대"를 주도적으로 만들어간 사람들은 수학자들이나 교수와 같은 학자들이 아니었다.

이와 같이 활발하게 진행된 민간주도의 조사활동은 1850년을 넘어서면서부터 급격히 위축되기 시작하였는데 그 대표적인 이유는 산업화와 프랑스혁명에 의한 도시화가 조사지역을 급격히 확대시켰기 때문이다. 다시 말하면 전수조사를 하기에는 경제적, 시간적으로 절대적인 한계에 봉착했기 때문이다. 이와 같은 상황에서 대안으로 표본조사가 제기되었지만

당시의 조사관계자들의 개념에서는 쉽게 받아들이기 어려웠다. 거의 1900년(1896년)에 비로소 표본조사가 공론화되었고, 일찍이 연구가 진행되었던 확률론과 접목되면서 표본의 대표성과 오차의 계량화 등의 어려움을 해결하며 1925년경에 표본조사가 본격적인 궤도에 오르게 되었다.

❷ 실험을 통한 방법

국가통계에서는 많이 나타나지는 않지만 자료수집의 대표적인 방법 중에 하나는 실험에 의한 자료수집이다. 실험은 연구대상 집단에서 표본으로 추출된 개체들의 환경을 통제하여 자료를 수집하는 방법으로 정의하는데 연구자가 의도적으로 조작한 변수의 영향이 어떤 결과로 나타나는가를 파악하기 위한 방법이다. 앞에서 토의한 조사가 연구모집단의 특성이나 조사항목(변수) 간의 관계와 상호영향 여부를 파악하는 데까지만 관심을 갖는 것이라면 실험은 변수 간에 인과관계를 밝히는 것을 궁극적 목적으로 한다고 할 수 있다. 이러한 방법은 공정의 온도와 속도가 제품의 완성도에 영향을 미치는가를 확인하기 위하여 온도와 속도를 다르게 하여 제품의 완성도를 평가하거나, 새로운 치료제를 개발하여 효능을 검정하기 위하여 임상시험을 수행하는 것, 화학물질의 수서 생물에 대한 영향을 평가하기 위해 어류에 대하여 몇 가지 시험물질에 노출시킨 후 체중을 조사하는 것 등을 예로 들 수 있다.

❸ 행정자료를 활용하는 방법

행정자료를 활용하는 방법은 행정(보고) 목적으로 축적된 자료들을 수집 또는 제공받는 방법으로 최봉호(2015)는 행정자료를 활용하여 작성된 통계 사례를 다음과 같이 정리하였다.

- 통계청 작성통계: 호적신고 및 주민등록신고 자료로부터 인구동태통계 및 인구이동통계
- 관세청: 무역통계
- 법무부 출입국관리국: 출입국통계도
- 국세청: 징세실적계, 사업자 등록신고 자료로부터 사업체 생멸통계
- 경찰청: 교통사고통계

- 고용노동부: 산업재해통계

- 소방관서: 화재 및 구급·구조 통계

- 외국: 스웨덴, 노르웨이, 덴마크 등 북유럽 국가는 통계청에서 만들어 내는 통계의 80% 이상이 이 방법으로 작성됨

1. 인구등록센서스

2015년에 처음으로 실시된 인구등록센서스가 대표적인 최신사례라고 할 수 있다. 인구등록센서스는 주민등록부, 건축물대장 등 인구, 가구, 주택과 관련된 가용한 행정자료를 이용하여 현장조사 없이 통계를 생산하는 새로운 인구주택 총조사 방식으로 13개 기관의 24종 행정자료를 연계하여 하는 조사이다. 이 조사의 기준시점은 11월 1일 0시로 한다. 이 조사를 통하여 인구분야에서는 성명, 성별, 나이, 가구주와의 관계, 국적, 입국연월, 본관 등 7개 항목과 가구 분야에서는 가구구분 한 개 그리고 주택분야에서 거처의 종류, 주거용 연면적, 대지면적, 건축연도 등 4개를 조사하여 총 12개의 기본항목에 대한 통계를 제공한다.

4 통계를 이용하는 방법

이미 분석·가공된 통계를 이용하여 작성하는 것인데 이렇게 작성된 통계를 가공통계라고 한다. 대표적인 가공통계는 여러 경제관련 기관이 공표한 거시경제지표를 이용하여 매월 공표되는 경기종합지수가 있다. 이 외에도 통계청의 인구추계, 한국은행의 국민계정(national account) 통계 등이 있다.

2-3. 표본과 통계량

학습목표

- **모집단에 속한 모든 개체에 대하여 자료를 수집하지 않고 이들의 일부를 선택하여 자료를 수집하고 이를 분석하여 모집단의 성질을 유추하려는 표본조사(실험)를 통계량과 함께 이해한다.**

1 표본

모집단과 표본을 구별하는 것은 대단히 중요하다. 자료수집 방법에서 표본조사를 소개한 적이 있지만 표본조사는 통계학의 큰 열매 중에 하나라고 할 만큼 통계학과 연관이 많은 분야이고 최근 다소 줄어드는 추세이지만 그래도 계속적으로 중요한 역할을 하고 있으므로 이 절에서 다시 한 번 좀 더 깊이 다루어 본다. 현실에서는 많은 경우 모집단 전체를 조사하는 전수조사를 수행할 수 없는 다음과 같은 상황에 처하게 된다.

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

이러한 경우에는 전수조사를 할 수 없거나 전수조사의 결과를 신뢰할 수 없게 된다. 이 때 하나의 대안으로 모집단을 가장 잘 대표할 수 있는 일부를 뽑아 조사하거나 측정하게 되는데 이 조사대상이 되는 모집단의 일부를 표본이라고 한다. 즉 표본은 실제 조사되거나 측정되는 모집단의 일부이다. 이 표본으로 추출된 개체들은 어떤 의도가 있어서 선택된 것이 아니라 전체 집단을 대표하는 개체들이라고 인정할 수 있어야 한다. 이를 근거로 집단 전체를 조사한 것과 같은(가까운) 결과를 얻을 것이라고 기대하게 된다.

현재 작성되고 있는 상당히 많은 국가통계가 표본조사이고 국가통계는 아니지만 언론매체에서 많이 보게 되는 여론조사(opinion survey)결과도

표본조사를 통해 이루어지고 있다. 많은 기업체는 자신의 사업 전략을 위하여 시장조사(market research, marketing survey)를 하고 있는데 이 역시 모두 표본조사이다. 잘 알려진 표본조사로 이루어지는 국가통계는 가계동향조사, 경제활동인구조사, 사업체노동력조사, 독서실태조사, 사교육비조사, 도소매업조사, 노인실태조사, 가족실태조사, 환경산업실태조사 등은 표본조사로 이루어지는 잘 알려진 국가통계들이다. 이 중에서 국가적으로 가장 중요하고 대표적인 표본조사인 가계동향조사를 살펴보자. 이 조사는 가구의 생활수준실태와 그 변동사항을 파악하기 위해서 가계의 수입과 지출을 조사하여 국민소비수준 변화를 측정 및 분석하고 소비자물가지수를 산출하는데 필요한 가중치를 비롯하여 각종 경제, 사회정책에 필요한 자료를 제공하는 조사이다. 우리사회의 심각한 문제 중 하나인 세입자의 주거대책비 산정 및 국민주택 공급대상의 기준설정 자료로도 활용된다. 가계동향조사의 개요를 보면 조사대상은 전국에 거주하는 일반가구로서 이 중 약 9,000가구를 표본으로 추출하여 가구주와의 관계, 성별, 교육정도 등 가구실태 관련 항목과 가구주소득, 주거비, 교육비 등 가계수지 관련 항목을 조사한다.

❷ 통계치와 통계량

2014년 사회조사 결과에서 발표한 "13세 이상 인구 중 48.7%가 전반적으로 자신의 건강상태에 대해「좋다」고 생각함"이라는 주장은 13세 이상 우리나라 국민에 대한 입장을 표현한 것이다. 그러나 사실 이 수치는 약 37,000명으로 구성된 표본에서 얻은 수치로서, 이를 이용하여 모집단 - 13세 이상 우리나라 국민 - 이 자신의 건강상태에 대해「좋다」고 생각하는 비율인 모수를 추정하는 것이다. 이와 같이 표본으로부터 얻은 사실로 전체 모집단을 추정하는 것이 통계학의 핵심역할이라고 할 수 있다. 위의 예에서 보듯이 우리는 문장에 나오는 수치를 볼 때 그 수치가 표본으로부터 산출된 값을 전달하고 있는지 아니면 모집단의 특성인 모수를 의미하는 것인지를 잘 구별해야 한다.

통상적으로 통계라고 하면 많은 사람들은 관측이나 실험 등을 통하여 얻은 자료로부터 산출된 숫자라고 말한다. 이 예에서는 약 37,000명으로 구성된 표본에서 얻은 수치 48.7%가 바로 통계이다. 우리가 익숙한 통계들로 실업률, 만족도, 평균 행복점수 등의 수치를 토의한 바 있다. 그런데 통계학에서

는 이와 같은 구체적인 값(숫자)들을 통계라고 하지 않고 통계치라고 한다.

반면에 통계학에서는 이와 같은 통계와 비슷한 의미를 갖으면서 또 좀 다른 용어가 있는데 그것이 통계량이다. 통계량을 일반적으로 표본을 설명하는 숫자라고 정의하기도 하는데 이 역시 명확한 표현은 아니다. 이 용어는 48.7%와 같이 실제 표본을 관측한 값으로부터 산출된 수치(값)인 통계치와 구별된다.

그렇다면 통계량은 무슨 뜻인가?

이것을 이해하기 위해서는 대표성 있는 확률표본을 통해서 얻어진 표본으로부터 산출되는 값은 상수일 수가 없다는 사실을 주목해야 한다. 예컨대 사회조사 예를 보자. 약 37,000명의 가구원을 조사하여 얻은 비율 48.7%라고 하였다. 그런데 만약 다른 가구들이 표본으로 추출되어 다른 가구원들이 조사되었다면 48.7%가 아닌 다른 값이 나왔을 것이다. 2014년 사회조사에서 얻은 표본비율 48.7%는 수많은 가능한 값들 중에 하나일 뿐이다.

바로 이러한 상황을 고려하면서 표본의 비율을 말할 때 우리는 통계량 표본비율이라고 한다. 따라서 통계량은 엄격히 말하면 숫자가 아니고 표본으로 추출되는 개체들로부터 수집되는 값을 이용하여 산출될 가능한 모든 표본비율을 고려할 때 사용하는 용어이다.

다시 한 번 정리해보면 통계치는 표본으로 추출된 개체들로부터 관측된 값을 이용하여 계산된 숫자이고 통계량은 그 값이 모집단의 어떤 개체들이 표본으로 추출되는가에 따라 변하는 것을 고려하는 변수라는 것이다. 그러므로 확률표본으로부터 얻을 수 있는 통계량은 확률분포를 따르게 된다. 다음의 예를 들어 보자.

서울시민의 행복점수이야기를 해 보자. 서울서베이의 경우 “귀하는 지금 얼마나 행복하십니까? 가장 행복한 상태를 100점으로, 가장 불행한 상태를 0점으로 하여, 요즘 귀하의 행복 정도를 점수로 말씀해 주십시오”라는 문항을 통하여 서울시민들의 행복점수를 조사하였다. 2014년 서울서베이 조사 결과보고서에 따르면, 15세 이상 서울시민 45,496명을 대상으로 조사한 2014년 서울시민의 행복점수는 72.0점이 나왔다. 이 숫자는 통계치이다. 그런데 이 값을 얻기 위해서 연구자들은 서울시민을 대표할 수 있는 표본을 추출하기 위하여 적절한 표본설계를 하고 이 설계에 따라 표본으로 추출된 개체들로부터 수집된 자료를 가지고 표본평균도 구하고 표본

비율도 구할 계획을 세웠을 것이다. 그런데 이때까지는 이들은 이 값들(평균, 비율)이 얼마가 될지는 모른다. 왜냐하면 이 값은 어느 시민이 표본으로 추출되느냐에 따라 달라지기 때문이다. 바로 이 상태에서 생각하는 것이 통계량이다.

대표적인 통계량은 다음과 같은 것으로 숫자가 아니다.

- 표본을 추출하여 관측되는 값들의 평균인 표본평균
- 표본을 추출하여 관측되는 값들의 중앙값(중위수)인 표본중위수
- 표본을 추출하여 관측되는 관심 사건의 비율인 표본비율
- 표본을 추출하여 관측되는 값들의 하위 10%에 위치한 값인 표본 하위 10%
- 표본을 추출하여 관측되는 값들의 최댓값과 최솟값의 차이인 표본범위

통계량의 개념은 대단히 중요하기 때문에 위의 설명을 좀 다르게 표현해 보자. 표본을 추출하여 모집단에 대한 정보를 얻으려고(모집단의 모수를 추정하려고) 하는 사람은 연구단계에서부터 그 모수에 대응하는 통계량을 생각하게 된다. 그리고 표본을 추출한 후 통계치를 얻은 후에도 연구자의 마음속에는 다른 개체들이 표본으로 추출되었을 경우에 산출될 수 있었을 통계치(이 때는 이를 통계치 대신에 추정을 목적으로 대하기 때문에 추정치라고 부른다)들을 머리로 생각해야하기 때문에 역시 통계량(이 때는 이를 통계량 대신에 추정을 목적으로 대하기 때문에 추정량이라고 부른다)을 생각하고 있다고 말할 수 있다.

참고자료

- 김수택 · 김영원 · 류제복 · 박진우 · 변종석 · 이기성 · 이해용 · 이흥철 · 최경호 · 한근식 · 홍기학(2002), 조사방법의 이해, 교우사.
- 박진우(2006), 통계학의 길잡이, 교우사.
- 조재근(2010), 통계로 읽는 사회와 경제, 교우사.
- 최봉호(2015), 현장 중심의 통계조사 이해, 자유아카데미.
- 통계청(2014), 사교육비통계조사 보고서.
- 통계청(2014), 사회조사 보고서.
- 서울특별시(2014), 서울서베이.

제 3 장

자료의 특성 확인하기

3-1. 개체와 변수 그리고 자료

학습목표

- **통계의 작성, 활용에서 자료의 생성 과정이 중요한 과정 중의 하나인 것과 자료생성의 중요한 두 요소가 개체와 변수라는 것을 이해한다.**

1 개체와 변수 그리고 자료

개체는 자료를 수집하는 대상이다. 국가통계에서 대표적인 개체는 개인이나 기업이 된다. 그러나 때로는 개인이 아닌 개인의 집합인 가구가 될 수도 있고 가구들의 모임인 조사구가 자료의 수집단위가 될 수도 있다. 한편 변수는 개체의 특징을 나타낸다. 따라서 변수는 개체마다 고유의 값을 갖게 된다. 변수를 잘못 정의하거나 오해할 수 있는 의미로 설명하면 원하는 정보를 얻을 수 없게 되거나 상황을 잘 설명하지 못하는 자료를 만들 수도 있다. 그리고 자료는 조사대상인 각 개체로부터 각 변수에 대하여 수집된 값들을 의미한다. 따라서 자료라고 하면 한 개체의 한 변수의 값을 말 할 때도 있고 여러 개체의 여러 변수의 값들을 말 할 때도 있어서 이 역시 문맥을 통하여 확인해야 한다. 경우에 따라서는 여러 개체의 여러 변수의 값들을 집합으로 전달 할 경우에는 데이터 세트(data set) 라고 하고 변수가 갖는 값을 자료라고 하여 구별 하는 경우도 있는데 이 역시 교재마다 다르게 표현하고 있으므로 문맥을 통하여 이해하는 것이 좋겠다.

예를 들어, 가계동향조사에서 수집된 자료를 보자. 하나의 개체는 표본으로 추출된 가구인데 다음과 같이 다섯 가구의 일부 자료를 보자.

처음변수는 이들 가구의 일련번호이고 이 외에 네 개의 변수가 사용되었

<표 3-1>
가계동향조사 자료
(일부) (1)

가구 일련번호	가구주성별	가구원수	세대구분	소득(천원)
1732	2	2	1세대	1,875,460
3073	2	1	1세대	4,069,740
906	1	4	2세대	6,452,620
60	1	4	2세대	8,336,790
101	1	2	1세대	2,553,000

다. 첫 번째 변수는 가구주의 성별을 나타내고 두 번째 변수는 가구원 수를 나타낸다. 세 번째 변수는 가구의 세대구분을 나타내고 마지막으로 네 번째 변수는 가구의 연간소득을 나타내는 변수이다.

통계는 숫자를 다루지만 그렇다고 모든 변수가 숫자만을 값으로 갖지는 않는다. 또한 숫자를 값으로 갖는 것처럼 보이지만 그 숫자는 단순히 기호와 같은 역할만 하는 경우도 많다. 예를 들자면 위의 자료에서 성별변수를 '남', '여'의 값을 갖도록 하면 자료의 모습은 다음과 같이 된다.

또한 수집된 자료를 이용하여 새로운 변수를 만들어서 자료를 수집된 시

<표 3-2>
가계동향조사 자료
(일부) (2)

가구 일련번호	가구주성별	가구원수	세대구분	소득(천원)
1732	여	2	1세대	1,875,460
3073	여	1	1세대	4,069,740
906	남	4	2세대	6,452,620
60	남	4	2세대	8,336,790
101	남	2	1세대	2,553,000

점의 크기로부터 확장할 수도 있다. 다음은 가계동향조사의 조사항목 중 멥쌀, 찹쌀, 맥류, 두류, 기타곡물을 합하여 '곡물전체'라는 새로운 변수를 만들어 변수를 추가시킨 자료이다.

<표 3-3>
가계동향조사 자료
(일부) (3)

일련번호	멥쌀	찹쌀	맥류	두류	기타곡물
425	132,000	0	0	0	0
1619	48,000	0	20,000	15,000	0
1734	0	0	0	6,000	0
1585	0	16,000	6,000	0	0
1141	144,600	27,800	0	17,800	6,500

<표 3-4>
가계동향조사 자료 (일부) (4)

일련번호	멥쌀	찹쌀	맥류	두류	기타곡물	곡물전체
425	132,000	0	0	0	0	132,000
1619	48,000	0	20,000	15,000	0	83,000
1734	0	0	0	6,000	0	6,000
1585	0	16,000	6,000	0	0	22,000
1141	144,600	27,800	0	17,800	6,500	196,700

가계동향조사의 또 하나의 특징은 이 조사는 패널조사로서 한 개체로부터 동일변수를 매년 조사하기 때문에 동일변수의 관측값이 매년 나타나서 다음과 같은 형태의 자료로 나타낼 수도 있다.

<표 3-5>
가계동향조사 소득 (일부)

일련번호	2014년 소득	2015년 소득
1	2,964,170	4,859,150
2	514,970	615,870
3	325,000	1,270,000
4	3,016,147	2,195,027
7	4,536,700	4,275,630
8	1,500,000	2,000,000
9	603,100	899,000
10	3,687,000	3,687,000
15	4,052,590	3,949,930
20	1,720,850	1,564,780

3-2. 변수의 종류

학습목표

- **개체의 특성으로 정의된 변수는 특성을 파악하는 방법에 따라 달라지게 되는 것을 주목하게 하고 그 종류를 식별한다.**

1 변수의 종류

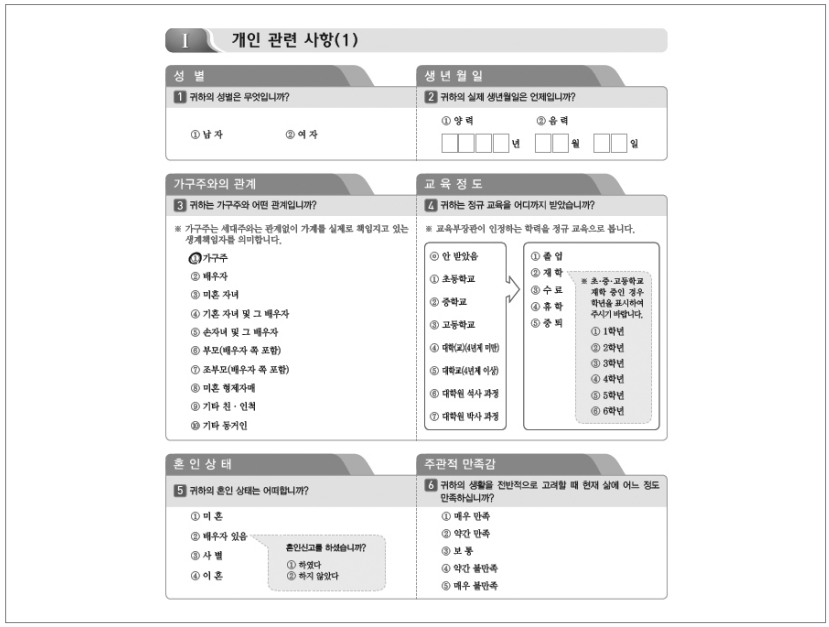
성별을 나타내는 변수는 '남자', '여자'의 값을 갖고 자녀의 수를 나타내는 변수는 0, 1, 2, 3 등의 정수 값을 갖고 사망시간 변수는 실수의 값을 갖는다. 이와 같이 변수는 그 변수가 갖는 값인 자료의 성격에 따라 다음과 같이 4가지의 척도(scale)로 구분된다.

1. 명명척도(명목척도, nominal scale)

가장 원시적인 변수로서 자료의 정보가 가장 적은 자료이다. 서열을 알 수 없는 각 범주(category)에 대응되는 도수(frequency)만 주어지는 자료이다.

아래의 2014년 사회조사표(일부)에서 보면 '성별'이나 '가구주와의 관계', '혼인상태' 등의 변수가 명명척도의 예이다. 여기에서 성별의 남, 녀를 의미하는 숫자 '1'과 '2'는 숫자로서의 의미를 갖지는 않는다.

[그림 3-1]
사회조사 조사표
(일부)



2. 순서척도(서열척도, 순위척도, ordinal scale, rank scale)

명명척도에 순위라는 정보를 더 갖는 자료이다. 범주 사이의 서열이 존재하는 자료이다.

위의 2014년 사회조사표의 6번 문항 '주관적 만족감'은 명목형으로 되어 있지만 그 값이 순서를 가지고 있으므로 순서척도의 예이다. 또한 아래의 2014년 서울서베이 조사표(일부)에서의 도시위험도처럼 '매우 위험하다'에서부터 '전혀 위험하지 않다'까지를 나타내는 5에서 1의 숫자는 순서의 의미를 포함하고 있다.

[그림 3-2]
서울서베이 조사표 (일부)

■ 서울 시민 고향 인식도

문1. 귀하는 서울에 거주하시면서 서울이 *고향같이 같이* 느끼시는지요?

① 서울에서 태어나서 서울이 고향같이 느껴진다
② 서울에서 태어났으나 서울이 고향 같지 않다
③ 서울에서 태어나지는 않았지만 살다보니 서울이 고향 같이 느껴진다
④ 서울에서 태어나지 않았고 서울이 고향으로 느껴지지도 않는다

■ 서울축제에 대한 인지도 및 참여자 만족도

문2. 귀하는 서울시 및 구청에서 개최하고 있는 *다양한 축제*에 대해 얼마나 관심이 있습니까?
(예 : Hi Seoul 페스티벌, 청계천축제, 불꽃축제 등)

⑤ 매우 관심있다 ④ 약간 관심있다 ③ 보통이다
② 약간 관심없다 ① 전혀 관심없다

문2-1. 귀하는 서울에서 *개최된 축제*에 참여한 경험이 있으신지요?

① 있다 ⇒ [문2-2로] ② 없다 ⇒ [문3으로]

문2-2. 귀하는 *축제*에 참여하신 후 어느 정도 *만족*하셨는지요?

⑤ 매우 만족 ④ 약간 만족 ③ 보통이다
② 약간 불만족 ① 매우 불만족

■ 사회자본

문3. 귀하는 지난 1년 동안 다음의 모임 또는 단체활동에 참여한 경험이 있습니까? 참여 경험이 있는 것을 모두 표시해 주십시오.

① 친목회/친목계 ② 동창회/동창모임
③ 지역모임/향우회/종친회 ④ 인터넷 커뮤니티
⑤ 동호회 ⑥ 자원봉사단체
⑦ 시민단체 ⑧ 노조 및 직능 단체
⑨ 정당 ⑩ 종교단체
⑪ 기타(구체적으로 : _____)
⑫ 어느 모임이나 단체에도 참여한 적이 없다

■ 행복지수

문4. 귀하는 요즘 스스로 행복하다고 생각하십니까? *가장 행복한 상태*를 10점으로, *가장 불행한 상태*를 0점으로 하여 각 영역별 자신의 행복점수를 표시해 주십시오.

보통
가장 불행한 상태 0 1 2 3 4 5 6 7 8 9 10 가장 행복한 상태

1) 자신의 건강상태 ..... (___)점
2) 자신의 재정상태 ..... (___)점
3) 주위 친지, 친구와의 관계 ..... (___)점
4) 가정생활 ..... (___)점
5) 사회생활(직장 학교, 종교, 취미, 계모임 등) ..... (___)점

문4-1. 귀하는 지금 얼마나 행복하십니까? *가장 행복한 상태*를 100점으로, *가장 불행한 상태*를 0점으로 하여, 요즘 귀하의 행복 정도를 점수로 답변해 주십시오.

100점 중 [_____]점

■ 도시위험도

문5. 귀하는 서울에 거주하면서 다음과 같은 영역에 대해 어느 정도 *위험*하다고 생각하십니까? 각 영역에 대해 *귀하의 생각*을 평가해 주십시오.

매우 위험하다	약간 위험하다	보통이다	별로 위험하지 않다	전혀 위험하지 않다
5	4	3	2	1

1) 화재나 홍수, 산사태 등의 재해로 인한 피해	
2) 밤늦게 걸어 다닐 경우	
3) 강도, 소매치기, 성추행 등 다양한 범죄 피해	
4) 건물, 엘리베이터 추락, 다리붕괴 등 여러 유형의 건축물 사고	

■ 현대사회의 위험 요인에 따른 피해 정도

문6. 아래의 표에 나열되어 있는 위험물이 발생한다면, 이로 인해 예상되는 피해가 얼마나 크다고 생각하십니까? 다음 중 귀하께서 생각하시기에 예상되는 피해 정도와 일치하는 곳에 응답하여 주십시오.

매우 피해가 크다	대체로 피해가 큰 편이다	보통이다	별로 피해가 없는 편이다	전혀 피해가 없다
5	4	3	2	1

1) 자연재해(태풍, 지진, 홍수 등)	
2) 교통사고	
3) 실업	
4) 테러가물, 방사능 사고, 피해	
5) 전염병(사스, 결핵, 콜레라, 장티푸스, 에볼라 등)	
6) 부정부패	
7) 폭력 범죄(성폭력, 학교폭력, 강도, 유괴, 폭행 및 살해 등)	
8) 사회 갈등	
9) 경제위기(금융위기 등)	
10) 화재	
11) 빈부격차 확대	
12) 컴퓨터 바이러스, 사이버범죄로 인한 혼란 (개인정보유출과 온라인 사기 등)	
13) 성인병(암, 고혈압, 당뇨 등)	

문7. 귀하는 10년 전과 비교할 때 서울 시민이 오늘날 경험하는 위험의 정도가 어떻게 변했다고 생각하십니까?

① 위험이 매우 커졌다 ② 위험이 상당히 커졌다
③ 10년 전과 비슷하다 ④ 위험이 약간 줄었다
⑤ 위험이 많이 줄었다

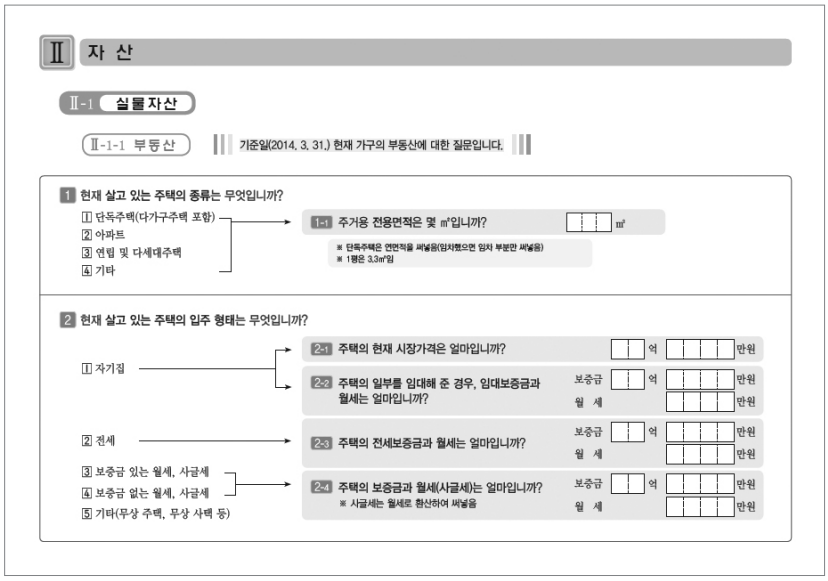
3. 등간척도(interval scale)

순위척도에 차이(difference)라는 정보가 부여된 자료이다. 그러나 절대영점이 없기 때문에 비율이 지켜지지 않는다. 기상통계에서 일별 최고온도 변수가 등간척도이다.

4. 비율척도(ratio scale)

등간척도에 절대영점이 부여된 자료이다. 아래의 가계금융 · 복지조사 2-1번 문항의 주택의 현재 시장가격을 "OOOO만원"으로 조사하는 것이 비율척도의 예이다.

[그림 3-3]
가계금융 · 복지조사 조사표 (일부)


Ⅱ 자 산

Ⅱ-1 실물자산

Ⅱ-1-1 부동산 ‖ 기준일(2014. 3. 31.) 현재 가구의 부동산에 대한 질문입니다. ‖

1 현재 살고 있는 주택의 종류는 무엇입니까?
① 단독주택(다가구주택 포함)
② 아파트
③ 연립 및 다세대주택
④ 기타
→ 1-1 주거용 전용면적은 몇 ㎡입니까? ㎡
※ 단독주택은 연면적을 써넣음(임차했으면 임차 부분만 써넣음)
※ 1평은 3.3㎡임

2 현재 살고 있는 주택의 입주 형태는 무엇입니까?
① 자기집 →
2-1 주택의 현재 시장가격은 얼마입니까? 억 만원
2-2 주택의 일부를 임대해 준 경우, 임대보증금과 월세는 얼마입니까? 보증금 억 만원 / 월 세 만원
② 전세 →
2-3 주택의 전세보증금과 월세는 얼마입니까? 보증금 억 만원 / 월 세 만원
③ 보증금 있는 월세, 사글세
④ 보증금 없는 월세, 사글세
⑤ 기타(무상 주택, 무상 사택 등)
→ 2-4 주택의 보증금과 월세(사글세)는 얼마입니까? 보증금 억 만원 / 월 세 만원
※ 사글세는 월세로 환산하여 써넣음

변수의 척도에 따라 이들 변수가 갖는 값인 자료를 분류하게 된다.

먼저 명명척도와 순위척도를 묶어 범주형 자료(categorical data) 또는 질적 자료(qualitative data)라고 부르고 등간척도와 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 이산형 자료(discrete data)로 구분하기도 한다. 자료가 어떤 종류이냐에 따라 자료에 대한 통계분석 방법이 다르게 됨으로 우리는 자료의 성격을 잘 파악하여야 한다. 그런데 이 분류관점을 현실 자료에 적용할 때에는 융통성을 갖는 것이 필요하다. 예컨대 사회조사 자료에서 문항 '귀하의 전반적인 건강 상태는 어떠하십니까?'에 대하여 자신의 건강 상태 변수는 '매우 좋다'부터 '좋은 편이다', '보통이다', '나쁜 편이다', '매우 나쁘다'까지가 '1'에서 '5'의 값을 갖는 순위척도로서 엄격한 의미에서 질적 자료이다. 그러나 때에 따라서는 이 값들을 마치 등간격을 가지는 이산형 변수로 보아 마치 5점 만점의 정수 값을 갖는 만족점수 변수처럼 취급하기도 한다.

척도와는 다른 관점에서 변수를 구별할 수 있다. 분석적인 측면에서 변수를 나눌 때는 독립변수와 종속변수로 구분하는데, 독립변수(independent variable)는 다른 변수에 영향을 주는 변수를 의미하고 다른 변수의 영향을 받는 변수를 종속변수(dependent variable)라고 한다. 이 구별은 대부분 연구자의 연구목적에 의해서 결정된다.

3-3. 사례 분석

학습목표

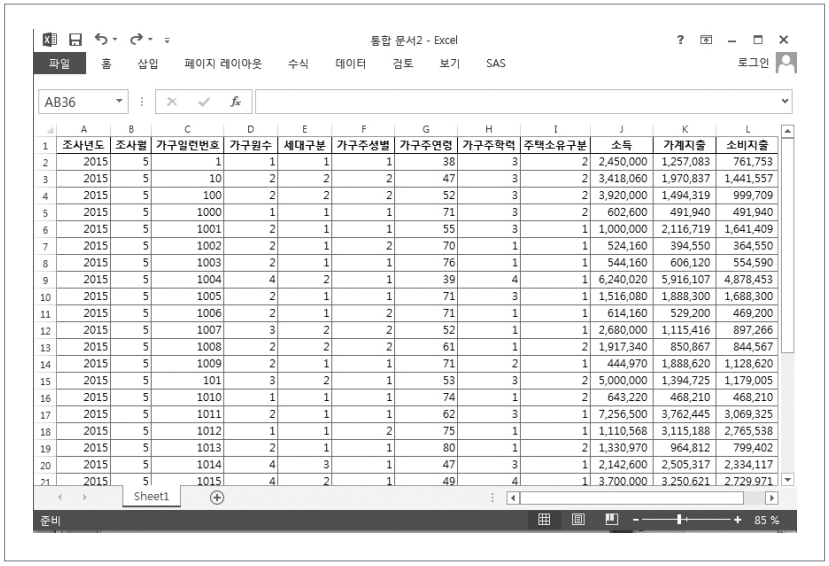
- **지금까지 조사단위인 개체와 변수 그리고 자료를 정의하였다. 이 절에서는 자료들을 통하여 개체와 변수, 자료의 척도, 성격 등을 자료가 입력된 도구와 연계하여 파악하는 능력을 제고한다.**

일반적으로 분석도구에서는 행에는 동일 개체로부터 측정되는 특성인 모든 변수에 해당되는 측정값을 입력하고, 열에는 동일 변수에 대한 모든 개체에 대한 측정값을 입력한다. 대표적인 도구인 엑셀과 SPSS의 자료입력창을 보면서 확인해보자.

1 엑셀 자료

엑셀에서는 앞에서 기술한 일반적인 입력방식을 따르지 않고 이용자의 의지에 따라 열과 행이 개체와 변수에 자유롭게 입력되고 분석될 수 있다. 다음은 2015 가계동향조사 자료의 일부를 엑셀에 입력한 것이다.

[그림 3-4]
엑셀 자료 입력 (1)



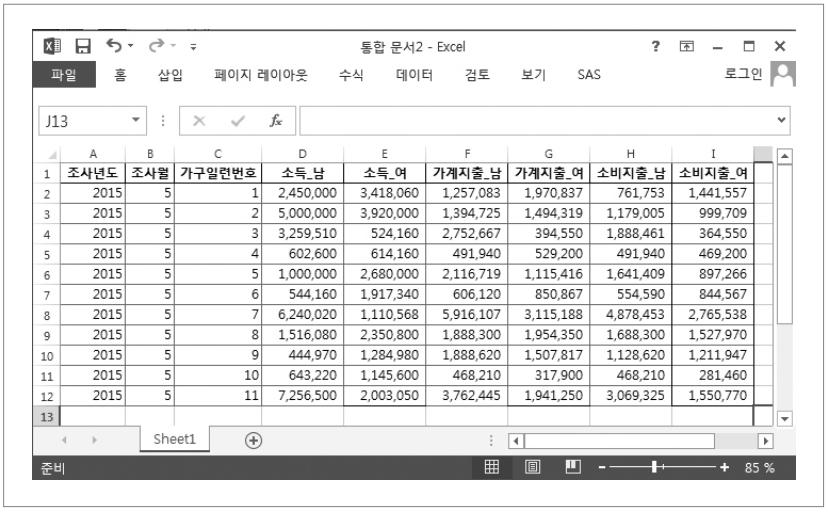
	A	B	C	D	E	F	G	H	I	J	K	L
1	조사년도	조사월	가구일련번호	가구원수	세대구분	가구주성별	가구주연령	가구주학력	주택소유구분	소득	가계지출	소비지출
2	2015	5	1	1	1	1	38	3	2	2,450,000	1,257,083	761,753
3	2015	5	10	2	2	2	47	3	2	3,418,060	1,970,837	1,441,557
4	2015	5	100	2	2	2	52	3	2	3,920,000	1,494,319	999,709
5	2015	5	1000	1	1	1	71	3	2	602,600	491,940	491,940
6	2015	5	1001	2	1	1	55	3	1	1,000,000	2,116,719	1,641,409
7	2015	5	1002	2	1	2	70	1	1	524,160	394,550	364,550
8	2015	5	1003	2	1	1	76	1	1	544,160	606,120	554,590
9	2015	5	1004	4	2	1	39	4	1	6,240,020	5,916,107	4,878,453
10	2015	5	1005	2	1	1	71	3	1	1,516,080	1,888,300	1,688,300
11	2015	5	1006	2	1	2	71	1	1	614,160	529,200	469,200
12	2015	5	1007	3	2	2	52	1	1	2,680,000	1,115,416	897,266
13	2015	5	1008	2	2	2	61	1	2	1,917,340	850,867	844,567
14	2015	5	1009	2	1	1	71	2	1	444,970	1,888,620	1,128,620
15	2015	5	101	3	2	1	53	3	2	5,000,000	1,394,725	1,179,005
16	2015	5	1010	1	1	1	74	1	2	643,220	468,210	468,210
17	2015	5	1011	2	1	1	62	3	1	7,256,500	3,762,445	3,069,325
18	2015	5	1012	1	1	2	75	1	1	1,110,568	3,115,188	2,765,538
19	2015	5	1013	2	1	1	80	1	2	1,330,970	964,812	799,402
20	2015	5	1014	4	3	1	47	3	1	2,142,600	2,505,317	2,334,117
21	2015	5	1015	4	2	1	49	4	1	3,700,000	3,250,621	2,729,971

입력된 데이터를 보면 행은 조사하고자 하는 대상인 하나의 가구 즉, 개체를 나타내며 열은 각각의 개체에 대하여 측정하고자 하는 관심 있는 특성 즉, 변수를 나타낸다. 엑셀의 시트는 기본적으로 열 번호와 행 번호가 주

어져 있으며 만일 변수명을 입력하고자 하는 경우에는 데이터 시트에서 한 행을 선택하여 입력하여야 한다. 따라서 변수명 없이 데이터를 입력할 수도 있고 두 행 이상을 이용하여 입력할 수도 있다.

그러나 엑셀의 경우는 다음과 같이 일반적인 입력방식 대신 이용자의 의지에 따라 열과 행에 개체와 변수를 자유롭게 입력될 수 있다.

[그림 3-5]
엑셀 자료 입력 (2)



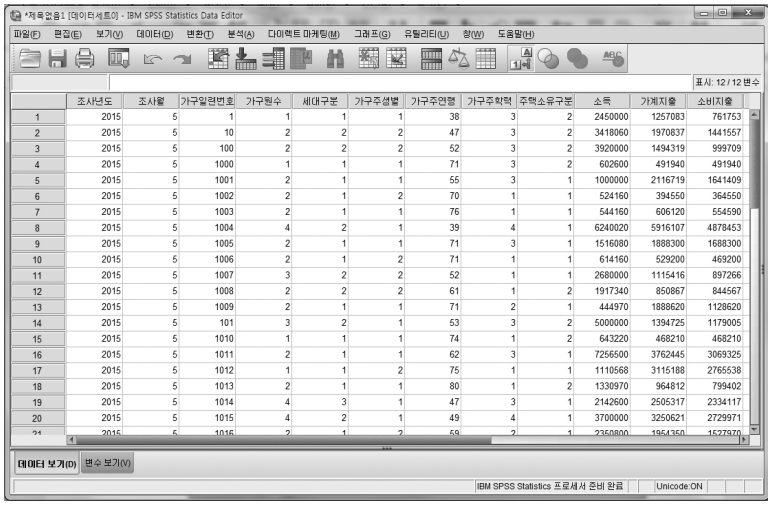
	A	B	C	D	E	F	G	H	I
1	조사년도	조사월	가구일련번호	소득_남	소득_여	가계지출_남	가계지출_여	소비지출_남	소비지출_여
2	2015	5	1	2,450,000	3,418,060	1,257,083	1,970,837	761,753	1,441,557
3	2015	5	2	5,000,000	3,920,000	1,394,725	1,494,319	1,179,005	999,709
4	2015	5	3	3,259,510	524,160	2,752,667	394,550	1,888,461	364,550
5	2015	5	4	602,600	614,160	491,940	529,200	491,940	469,200
6	2015	5	5	1,000,000	2,680,000	2,116,719	1,115,416	1,641,409	897,266
7	2015	5	6	544,160	1,917,340	606,120	850,867	554,590	844,567
8	2015	5	7	6,240,020	1,110,568	5,916,107	3,115,188	4,878,453	2,765,538
9	2015	5	8	1,516,080	2,350,800	1,888,300	1,954,350	1,688,300	1,527,970
10	2015	5	9	444,970	1,284,980	1,888,620	1,507,817	1,128,620	1,211,947
11	2015	5	10	643,220	1,145,600	468,210	317,900	468,210	281,460
12	2015	5	11	7,256,500	2,003,050	3,762,445	1,941,250	3,069,325	1,550,770
13									

소득의 경우 하나의 변수이지만 남자와 여자를 따로 구분하여 각 열로 입력할 수 있다. 상황에 따라서는 자료를 위와 같이 입력하는 것이 보기에 용이할 수 있다. 그러나 이 경우에는 행이 하나의 개체를 의미하는 것이 아니며 소득이라는 하나의 변수가 한 열에 입력되지 않으므로 분석 시 주의해야 한다.

2 SPSS 자료

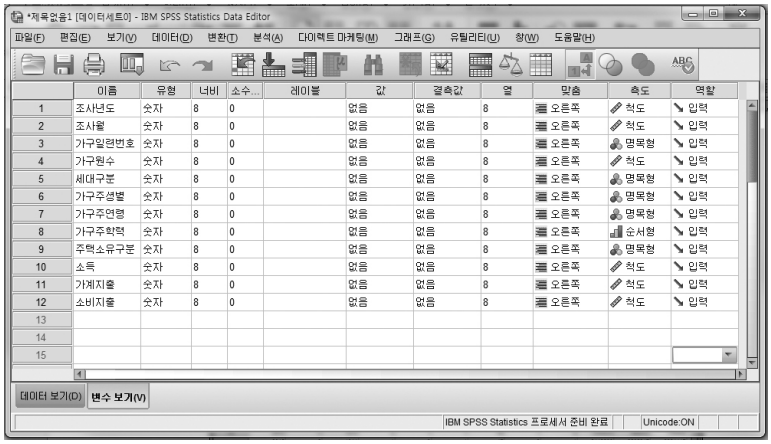
SPSS의 경우에는 행에는 동일 개체로부터 측정되는 특성인 모든 변수에 해당되는 측정값을 입력하고, 열에는 동일 변수에 대한 모든 개체의 측정값을 입력하는 일반적인 방식을 그대로 따라야 한다. 다음은 2015 가계동향조사 자료의 일부를 SPSS에 입력한 것이다.

[그림 3-6]
SPSS 자료 입력 (1)



SPSS는 일반적인 입력방식을 따르므로 행은 개체를 의미하고 열은 변수를 의미한다. 따라서 데이터 시트에 입력된 변수들에 대하여 변수명이나 변수의 특징은 다음과 같이 '변수보기'라는 별도의 시트에서 입력하도록 되어있다.

[그림 3-7]
SPSS 자료 입력 (2)



참고자료

- 김수택 · 김영원 · 류제복 · 박진우 · 변종석 · 이기성 · 이해용 · 이흥철 · 최경호 · 한근식 · 홍기학(2002), 조사방법의 이해, 교우사.
- 박진우(2006), 통계학의 길잡이, 교우사.
- 성내경(1995), 정보시대 그리고 통계, 이화여자대학교 출판부.
- 통계청(2014), 가계금융·복지조사 조사표.
- 통계청(2014), 사회조사 조사표.
- 통계청(2015), 가계동향조사.
- 서울특별시(2014), 서울서베이.

제 4 장

분포를 숫자로 파악하기

4-1. 학창시절 회고

학습목표

- **본 절에서는 자료의 요약과 그 결과에 대한 해석이 대부분 학교 교육현장에서 들어왔던 내용이었음을 대화체를 통하여 확인하여 토의 주제가 결코 새롭고 어려운 주제가 아니라는 인식을 심어주면서 요약 통계치를 이해한다.**

분포를 숫자로 파악하기란 다른 말로 하면 우리나라 고등학교의 확률과 통계 교재로부터 대학에서 사용하는 거의 모든 통계학 교재에서 다루는 자료요약을 의미한다. 사실 이 내용은 어렸을 때부터 너무도 많이 들어왔고, 수학시험의 단골 문제이기도 했고 그리고 사실은 잘 모르면서도 또 이야기를 하면 듣는 것에 저항 심리가 생기는 그런 주제이다. 따라서 다음과 같은 대화를 읽으면서 학창시절을 느껴보면 좋겠다.

교사 : 이번에는 자료의 요약에 대해서 이야기하도록 할까요? 어제 옛 친구 한 명을 30년 만에 만났는데 키가 180cm에 체중은 약 60kg쯤 되는 것 같았어요. 여러분들은 그 친구의 체형이 어떠한지 감이 잡히나요?

철수 : 키가 좀 큰 보통 체격이겠네요.

영희 : 아니, 상당히 마른 체형이겠는데요.

철수 : 맞아, 그렇겠네요. 키가 180cm이면 보통 체중은 75kg은 되어야지요.

교사 : 그래, 상당히 살이 빠진 것 같았어요. 하여간 180, 60이라는 두 숫자로 대략 한 사람의 체형에 대한 감을 잡을 수 있었으니까… 우리는 키와 체중을 한 사람 체형의 대푯값이라고 할 수 있겠네요.

이처럼 앞으로 할 이야기는 수집된 자료를 몇 개의 숫자(요약값)를 통해 자료의 전체적인 감, 즉 분포를 잡아보려고 하는 것이에요. 보통 키, 체중으로 체형에 대해서 감을 잡듯이 자료에 대해서는 어떤 요약값이 잘 사용되는지 대답할 수 있겠어요?

영희 : 평균을 쓰지 않나요?

철수 : 표준편차나 범위도 있었던 것 같습니다. 공식은 잊었지만….

영희 : 표준편차를 다 배운 것 같아서 또 묻기가 좀 그렇지만 저는 잘 모르겠어요. 수식도 좀 복잡하구요.

교사 : 영희가 많이 성장했구나. 여러 사람들 앞에서 창피하게 느낄 수 있는 것을 당당하게 질문하는 것을 보니 말이야. 사실 표준편차란 것이 쉽지는 않은 개념이에요. 많은 사람들이 공식은 알고 있을지 몰라도 그 지식을 사용하는 사람은 그렇게 많지가 않은 것 같아요.

질문이 나왔으니 먼저 표준편차라는 지식이 어느 상황에서 유용하게 사용되는지를 공부해 봅시다. 다음은 두 집단 A, B에서 각각 추출한 크기가 4인 임의표본의 관찰값이에요. 이 값들을 보며 각 집단의 특성과 두 집단의 차이를 토의해 봅시다.

- A집단의 표본 관찰값 : 1 2 8 9
- B집단의 표본 관찰값 : 3 4 6 7

철수 : 두 집단 모두 평균은 비슷하겠네요, 표본 관찰값의 평균은 5로 같으니까요. 그런데 A집단은 B집단에 비해서 양극으로 나누어져 있을 것 같아요. 최댓값과 최솟값의 차이가 9-1=8로써 B집단의 7-3=4보다 두 배나 크니까요.

교사 : 훌륭한 논평이구나. 지금 사용한 관점이 '범위'라는 것으로 자료의 퍼짐, 즉 산포도를 나타내는 하나의 잣대(측도)이다.

영희 : 그것은 저도 아는데… 표준편차는 뭐지요?

교사 : 영희는 표준편차가 몹시 궁금한 모양이구나. 그러면 영희는 두 집단 중에 어느 집단이 더 넓게 퍼져 있다고 생각하니?

영희 : A집단이요.

교사 : 왜 그렇게 생각하니?

영희 : 철수가 범위를 검토하지 않았나요?

교사 : 다른 관점을 생각해 보는 건 어떻겠니?

영희 : 일반적으로 퍼짐이라는 말은 중심에서 얼마나 떨어졌는가를 말하는 것이니까 이

렇게 보면 어떨까요. A집단은 평균 5로부터 4만큼 떨어진 개체(관찰값이 1과 9)가 두 개, 또 평균에서 3만큼 떨어진 개체(관찰값이 2와 8)가 두개 있으니까 4개의 관찰값은 평균 5에서 평균적으로

$$\frac{|1-5|+|2-5|+|8-5|+|9-5|}{4}=3.5$$

떨어져 있다고 보이고 B집단은

$$\frac{|3-5|+|4-5|+|6-5|+|7-5|}{4}=1.5$$

떨어져 있다고 보면 되겠네요.

교사 : 좋은 아이디어네요. 지금 영희가 제안한 관점으로 계산된 값 3.5, 1.5를 평균절대편차라고 한다. 즉 각 개체가 평균에서 떨어져 있는 거리의 평균이라고 보면 되요.

영희 : 표준편차가 아니에요?

교사 : 지금 영희가 계산한 것은 표준편차가 아니고 평균절대편차이다.

영희 : 그럼 표준편차는 무엇인가요?

교사 : 평균편차와 비슷한 값인데 설명은 좀 복잡해요. 표준편차는 분산이라는 값의 제곱근을 말하는데 그 값의 해석은 조금 전에 영희가 구한 평균편차처럼 하면 큰 무리는 없어요. 그렇다면 분산은 어떤 값이냐 하면 '각 관찰값과 평균값(5)과의 차이의 제곱'의 평균을 뜻해요. 더 정확히 말하면 '각 관찰값과 평균값(5)과의 차이의 제곱을 다 합한 값을 자료의 수에서 하나를 뺀 값으로 나눈 값'이에요. 구체적으로 써 보면

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

과 같이 됩니다. (여기서 분모가 $n-1$ 이 된 이유는 이 자료가 표본자료이기 때문인데 이해가 안 되더라도 받아들이고 이야기를 진행합시다.) 표본표준편차는

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

이 되지요. 우리가 이야기 하고 있던 자료에서 확인해 볼까요?

• A집단 :

표본분산 $\Rightarrow$ $\frac{(1-5)^2+(2-5)^2+(8-5)^2+(9-5)^2}{4-1}=16.67$

표본표준편차 ⇒ $\sqrt{16.67} = 4.08$

• B집단 :

표본분산 ⇒ $\dfrac{(3-5)^2+(4-5)^2+(6-5)^2+(7-5)^2}{4-1} = 3.33$

표본표준편차 ⇒ $\sqrt{3.33} = 1.82$

A집단의 표준편차 4.08은 비록 앞에서 구한 평균절대편차 3.5와 좀 다르지만 해석할 때는 A집단의 각 개체가 평균에서 떨어진 거리의 평균인 것처럼 해석하면서 A집단의 퍼짐에 대한 감을 잡기 바랍니다.

표준편차 얘기는 다시 정리하기로 하고… 영희 학생 때문에 토의가 너무 수리적으로 흐른 것 같은데 학창시절에 접했던 좀 더 직관적인 생각들을 기억해 봅시다.

토의를 쉽게 하기 위하여 다시 작은 크기의 자료를 예로 제시해 볼게요. 다음은 크기가 30인, 가구원수가 4명인 도시근로자가구의 한달 생활비 자료입니다.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	400

여러분은 이 자료에서 평균과 표준편차 말고 또 무슨 값이 궁금한가요?

영희 : 최댓값, 최솟값이요. 도대체 많이 쓰는 사람은 얼마나 쓰나 궁금하니까요.

철수 : 중앙값이요. 딱 중간에 있는 사람…

영희 : 상류층들의 생활비요.

교사 : 상류층이라 함은 누구를 말하나요?

철수 : 상위 25% 이상 어때요?

영희 : 말이 되네. 그럼 하위 25%도 구하면 하류층의 경계선이 되겠네.

교사 : 두 사람이 너무 빨리 의견을 주어서 일단 정리를 할 필요가 있겠군요.

지금 두 사람이 말한 것은 아래와 같은 다섯 개의 값이 맞지요?

최솟값(min), 하위 25%값(Q1), 중앙값(Q2; M), 상위25%값(Q3), 최댓값(max)

위의 자료에서 다섯 숫자를 찾아보면 최솟값=100, Q1=130, Q2=150, Q3=200, 최댓값=400이 되겠네요. 맞지요? 대단히 좋은 값들이네요.

그런데 이 다섯숫자로 우리가 수집한 30가구의 생활비가 대개 어떠할지 설명해 볼 수 있겠어요?

영희 : 상위 25% 이상 가구의 생활비는 상당히 차이가 크네요, 400-200=200만원. 하위 25% 이하에 속한 가구들은 130에서 100사이인데 비해서요. 또 전체 생활비의 범위가 400-100=300만원인데 그 중 상위 25% 이하의 가구의 생활비는 100만원에서 200만원 사이에 있고, 상위 25%는 200만원에서 400만원 사이에 있네요. 그럼 대부분의 사람들의 생활비는 다 그만그만한데 상위 25%에 속한 사람들의 생활비는 그렇지 않은 것 같아요.

철수 : 보통 사람들 50%는 130만원에서 200만원 사이를 쓰는군요.

영희 : 보통사람이 누구인데?

철수 : 위쪽과 아래쪽 25%를 제외한 가운데에 있는 절반의 사람들을 보통사람이라고 할 수 있지 않을까?

교사 : 그럴 수도 있지요. 그래서 가운데 있는 50%의 범위, 즉 Q3-Q1을 다른 말로는 IQR(inter quartile range)라고 하지요.

철수 : 그런데 선생님, 중앙값을 어떻게 150이라고 하셨어요? 자료에는 150이 없는데…

영희 : 저도 고민을 했는데 생각해 보니 30명의 중앙은 15.5번째 잖아요. 그런데 그런 가구는 없으니 중앙값은 아래처럼 30명 중에서 *표시가 된 15번째 149만원과 16번째 151만원을 쓰는 사람들의 평균값으로 한 것 아닌가요?

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149*	151*	164	168	172	176
180	184	200	205	219	225	235	245	255	400

교사 : 정확히 추측했네요. 30명의 딱 중간이 없으니까 15번째 사람과 16번째 가구를 모두 중앙이라고 하고 이 가구의 생활비의 평균으로 중앙값을 정했지요. 그런데 지금까지 우리가 토의한 것이 통계학에서는 '다섯숫자 요약(five number summary)' 이라는 용어로 중요하게 생각되는 개념이에요. 여러분은 그런 내용을 배우지 않았는데도 직관적으로 알고 있었던 것 같아요. 아주 좋은 토의시간이었습니다.

4-2. 중심경향 측정을 위한 수치적 요약

학습목표

- **중심경향에 관련되어 많이 쓰이는 요약 통계치인 평균, 중앙값, 최빈값을 학습하는데 특히 평균의 여러 종류를 소개하여 이를 활용한다.**

수치적인 요약(numerical summary)에 관련된 방법들은 거의 대부분 한 개의 변수에 대한 분포의 특징을 설명하는 목적을 가지고 있다.

먼저 중심경향(central tendency)에 관련되어 많이 쓰이는 요약 수단을 보도록 하자. 여기에는 기본적으로 산술평균(mean), 중앙값(median), 그리고 최빈값(mode)의 3가지가 있다. 물론 이외에도 몇 가지가 더 있지만 이들은 이를 설명한 후에 다루도록 한다.

1 평균

1. 산술평균

산술평균(arithmetic mean)은 키와 체중과 같은 연속형 변수나 간혹 만족도 조사에서 5점 척도(① 매우 만족, ② 만족, ③ 보통, ④ 불만족, ⑤ 매우 불만족)로 조사된 자료를 요약할 때 사용된다. 자료가 모집단 전체인 경우는 모집단평균(population mean)이라 하고 자료가 모집단에서 추출된 표본이라면 표본평균(sample mean)이라 한다. 평균의 공식은 다음과 같다.

$$\text{표본평균} = \frac{\sum_{i=1}^{n} x_i}{n}$$

여기서 x_i는 i번째 관측값, 그리고 n은 관측값의 개수이다. 식에서 보는 것처럼 평균은 단순히 모든 관측값을 더한 다음 관측값의 개수로 나누어 주는 것이다.

4.1절에서 나온 도시근로자가구의 한달 생활비 자료의 예를 통하여 평균을 확인해보자.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	400

위의 표에서 평균은

$$\frac{(100+107+\ \cdots\ +255+400)}{30}=170$$

이 된다.

그런데 만일, 자료에서 마지막 관측값인 400만원이 2,500만원이었다면 평균은 어떻게 될까?

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	2,500*

이 자료의 평균은

$$\frac{(100+107+\ \cdots\ +255+2500)}{30}=240$$

이 된다. 평균이 자료의 중심경향을 나타내는 값이라고 했지만 현재 자료에서 240만원의 위치는 전체 자료 30개 중에서 4번째로 큰 값보다 크다. 이처럼 주어진 자료의 대부분이 240만원 미만이라는 점을 고려한다면, 240만원이라는 평균은 자료의 중심경향을 나타낸다고 보기 어렵다. 즉 자료에 아주 크거나 작은 관측값 혹은 관측값들이 섞여 있는 경우, 다른 말로 자료의 대칭성을 알 수 없는 경우 평균은 그 값의 영향을 받아 커지거나 작아지거나 하며 이에 따라 자료의 중심을 나타내는 대푯값으로서의 역할을 제대로 할 수 없다.

그렇다면 이런 평균과 달리 자료중 특이한(이상한) 관찰값에 의한 영향을 가급적 받지 않는 중심경향을 나타내는 요약값은 무엇이 있을까? 잠시 후에 토의해보자.

2. 기하평균

기하평균(geometric mean)은 다음과 같이 정의되는 값으로 인구의 변동률이나 물가변동률과 같은 비율에 대한 대푯값으로 주로 사용된다.

$$\text{기하평균} = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

기하평균의 예로 통계청에서 제공하는 2000년부터 2015년까지 대한민국의 인구 자료를 이용하여 2015년 이후 인구를 추계하는 과정을 한 가지 소개한다.

<표 4-1> 연도별 추계인구

연도	추계인구	비고
2000	47,008,111	
2001	47,357,362	
2002	47,622,179	
2003	47,859,311	
2004	48,039,415	
2005	48,138,077	
2006	48,371,946	
2007	48,597,652	
2008	48,948,698	
2009	49,182,038	
2010	49,410,366	
2011	49,779,440	
2012	50,004,441	
2013	50,219,669	
2014	50,423,955	
2015	50,617,045	
연평균 (2000~2015) 지수평균	0.0049	$=\sqrt[15]{\frac{50,617,045}{47,008,111}}-1$
2016	50,867,265	$=50,617,045+(0.0049\times 50,617,045)$
2017	51,052,536	$=50,867,265+(0.0047\times 50,867,265)$
2018	51,228,516	〃
2019	51,393,499	〃
2020	51,547,271	〃

이 자료에서 연도별 인구 증가율에 대한 평균은 시간에 따라 증가율이 변함으로 산술평균이 아닌 아래와 같이 기하평균으로 구해야 한다.

$$\text{기하평균} = \sqrt[15]{\frac{47,357,362}{47,008,111} \times \frac{47,622,179}{47,357,362} \times \cdots \times \frac{50,617,045}{50,423,955}} - 1 = 0.0049$$

그러면 연도별 추계인구에 대한 평균 증가율은 기하평균으로 0.49%이다.

3. 조화평균

조화평균(harmonic mean)은 거리가 일정한 경우에 각 구간에서의 평균속도를 이용하여 전 구간에서의 평균속도를 구한다든가 각 지불금액이 일정한 경우의 평균가격을 얻는 경우에 사용할 수 있는 대푯값으로 다음과 같이 정의된다.

$$\text{조화평균} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

조화평균의 예로 다음의 사례를 살펴보자. 아래의 표는 어느 기관이 3개월 동안 매달 10,000원어치씩 구매하는 제품 A의 월별 구매 개수와 단가를 나타낸 것이다.

<표 4-2>
제품 구매 자료

	금액	A제품 구매 개수	단가
1월	10,000원	10개	1,000원
2월	10,000원	20개	500원
3월	10,000원	50개	200원

A제품의 3개월간의 단가는 1000원, 500원, 200원이므로 평균 개당 구매가격을 산술평균으로 한다면 $\frac{1,000+500+200}{3} = 566.7$ 원이 되지만 이 기관에서 3개월 동안 구매한 개수와 총 지불액을 고려하면 지난 3개월간 30,000원으로 80개를 구매하였으므로 개당 375원에 구매한 것이 된다. 이런 경우는 산술평균이 아니라 조화평균 $\frac{3}{\frac{1}{1000}+\frac{1}{500}+\frac{1}{200}} = 375$원을 사용하는 것이 적절하다.

❷ 중앙값

앞의 4.1절에서 학생들과 교사의 대화에서 나온 중앙값을 떠올려보자. 중앙값(median)은 자료를 작은 값에서부터 큰 값까지 순서대로 나열했을 때 중앙에 있는 값을 말한다. 즉 자료의 절반은 중앙값보다 작고 나머지 절반은 중앙값보다 크다. 자료의 개수가 홀수라면 중앙값은 $\frac{n+1}{2}$ 번째 관측값이 되며 자료의 개수가 짝수이면 $\frac{n}{2}$ 번째 관측값과 $\frac{n+2}{2}$ 번째 관측값의 평균이 된다.

같은 예에서 중앙값을 구해보자.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	400

자료의 개수가 30으로 짝수이므로 중앙값은 15번째 관측값과 16번째 관측값의 평균인 $\frac{(149+151)}{2}=150$ 만원이 된다. 즉 자료의 절반인 15개 자료는 150만원보다 작고 나머지 15개 자료는 150만원보다 크다.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	2,500*

마지막 관측값이 2,500만원으로 변경된 자료의 경우 중앙값은 $\frac{(149+151)}{2}=150$ 만원이 된다. 이 값은 자료의 마지막 관측값이 400만원일 때와 동일한 값이며, 여전히 자료의 절반은 150만원보다 작고 나머지 절반은 150만원보다 크다.

동일한 자료에서 한 관측값이 400만원에서 2,500만원으로 변화할 때 평균은 170만원에서 240만원으로 그 값의 영향을 크게 받았지만, 중앙값은 그 값의 영향을 받지 않고 150만원을 유지하는 것을 확인할 수 있다. 이처럼 아주 크거나 작은 관측값(들)이 섞여 있는 경우에는 평균보다는 중앙값이 대표로서의 역할에 더 적절하다고 할 수 있다. 그러나 자료의 하나하나의 값에 민감하지 않다는 말은 다른 말로 하면 하나하나의 값 자체는 무시된다는 의미이기도 하다. 즉 중앙값에서 자료의 마지막 관측값이 400만원인지 2,500만원인지의 정보는 활용되지 않는다.

평균과 중앙값의 차이를 요약해 보자. 평균은 자료에 굉장히 크거나 작은 값들이 있는 경우 이 값들에 영향을 받아 민감하게 움직이지만 중앙값은 그렇지 않다. 이를 저항성이 강하다고 표현하기도 한다.

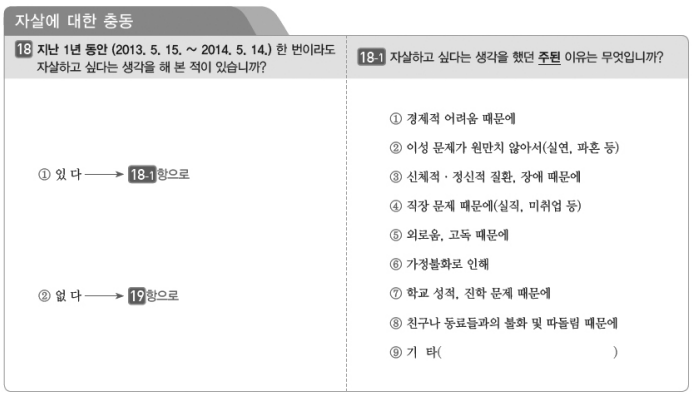
자료가 대칭인 경우는 평균과 중앙값이 비슷한 값을 갖는다. 그러나 자료에 굉장히 큰 값이 있거나 큰 값들이 많이 포함된(오른쪽으로 꼬리가 길다고 표현하기도 한다) 경우 평균은 그 값의 영향을 받아 커지는 반면 중앙값은 영향을 받지 않으므로 평균이 중앙값보다 크다. 반대로 자료에 굉장히 작은 값이 있거나 작은 값들이 많이 포함된 왼쪽으로 꼬리가 긴 경우에는 평균보다 중앙값이 크다. 이처럼 평균과 중앙값을 선택할 때에는 대칭 여부가 고려되어야 한다. 위의 예처럼 돈과 관련이 있는 변수들의 분포는 대부분이 비대칭, 특히 오른쪽으로 꼬리가 길며(큰 관측값이 포함된) 이런 경우에는 일반적으로 중앙값을 사용한다.

❸ 최빈값

최빈값(mode)은 자료에서 가장 많이 나타나는 관측값을 말한다. 따라서 빈도를 셀 수 있는 범주형 변수의 경우에 의미가 있으며, 같은 값이 한 두 번 나타나게 되는 연속형 변수의 경우에는 의미가 없다.

대표적인 예는 2014년 사회조사 결과에서 볼 수 있다. 2014 사회조사 18-1번 문항 '자살하고 싶다는 생각을 했던 주된 이유는 무엇인가요?'라는 질문은 아래의 질문지와 같이 구성되어 있다.

[그림 4-1]
사회조사 조사표
(일부)


자살에 대한 충동

18 지난 1년 동안 (2013. 5. 15. ~ 2014. 5. 14.) 한 번이라도 자살하고 싶다는 생각을 해 본 적이 있습니까?

① 있 다 ⟶ 18-1항으로

② 없 다 ⟶ 19항으로

18-1 자살하고 싶다는 생각을 했던 주된 이유는 무엇입니까?

① 경제적 어려움 때문에
② 이성 문제가 원만치 않아서(실연, 파혼 등)
③ 신체적 · 정신적 질환, 장애 때문에
④ 직장 문제 때문에(실직, 미취업 등)
⑤ 외로움, 고독 때문에
⑥ 가정불화로 인해
⑦ 학교 성적, 진학 문제 때문에
⑧ 친구나 동료들과의 불화 및 따돌림 때문에
⑨ 기 타( )

이 질문의 응답항목은 '1'에서 '9'까지의 숫자가 숫자로서의 의미를 가지는 것이 아니다. 따라서 응답내용을 대표하는 값으로는 연속형 변수에서처럼 평균이나 중앙값이 아닌 응답자들이 가장 많이 응답한 항목 즉 최빈값이어야 의미가 있다.

위의 예에서 보듯이 주어진 변수가 범주형(명목형)인 경우 이 변수를 대표하는 가장 적절한 값은 최빈값이라고 할 수 있다.

지금까지 토의한 내용을 정리하면 다음과 같다. 자료의 중심경향을 요약하는 평균과 중앙값, 최빈값을 3-2절에서 토의했던 변수의 종류와 함께 생각해 보자. 평균은 관측값들을 모두 이용하므로 비율척도이거나 등간척도인 경우 가장 적합한 중심경향을 대표하는 값이 된다. 그러나 자료가 대칭이 아니거나 자료에 매우 큰 값이나 작은 값이 섞여 있는 경우, 이 값들에 민감하게 영향을 받으므로 대표하는 값으로서의 역할을 적절히 하지 못한다. 중앙값은 비율척도나 등간척도, 순서척도에 모두 사용할 수 있다. 최빈값은 명목척도에 가장 적절하며 순서척도인 경우에도 사용할 수 있다.

4-3. 변동측정을 위한 도구

학습목표

- **1절에서 다루었던 자료의 퍼짐, 변동을 측정하는 도구에 대한 대화내용을 정리하는 방식으로 중고교시절을 회상시켜서 친숙함을 느끼고 개념을 이해한다.**

1 분산 및 표준편차

자료의 퍼짐, 변동을 측정하는 도구 중 가장 일반적으로 사용되는 것이 분산(variance)과 표준편차(standard deviation)이다. 분산은 각각의 관측값들이 평균과 얼마나 떨어져 있는지, 즉 편차를 구하여 제곱한 후 이들의 평균을 구하는 것이다. 분산을 구하는 방법은 두 가지가 있는데, 자료가 모집단인 경우와 표본인 경우에 따라 분모가 달라진다. 1절에서 나온 분산을 구하는 식

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

은 크기가 n인 표본인 경우이고, 크기가 N인 모집단인 경우는

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

로 구한다(여기서, μ는 모집단의 평균이다).

그러나 위 식에서 알 수 있듯이 분산은 관측값과 평균의 차이를 뜻하는 편차가 아닌 편차의 제곱을 사용하는데, 여기서 문제가 생긴다. 편차는 우리가 생각하는 각 관측값들과 같은 단위인 반면 분산은 단위의 제곱이 되는 것이다. 이를 해결하기 위하여 분산에 제곱근을 취하여 사용하는데 이를 표준편차라고 부른다.

$$\text{표본 표준편차} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$$

표준편차는 자료의 관측값들과 같은 단위를 가지며, 평균편차와 같이 자료가 평균으로부터 얼마나 떨어져 있는가의 평균이라고 해석할 수 있다. 즉 표준편차 혹은 분산이 작으면 자료는 평균을 중심으로 몰려 있다는 의

미이고, 표준편차 혹은 분산이 크면 자료는 평균과 멀리 흩어져 분포한다는 것을 알게 된다.

앞의 예를 이용하여 분산과 표준편차를 구해보자.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	400

주어진 자료에서 분산과 표준편차는 각각 다음과 같다.

$$\text{표본 분산} = \frac{(100-170)^2+(107-170)^2+\ \cdots\ +(400-170)^2)}{30-1} = 3737$$

$$\text{표본 표준편차} = \sqrt{3737} = 61$$

즉, 4인가구인 도시근로자의 생활비는 평균 170만원으로부터 평균적으로 61만원만큼 떨어져 있다고 할 수 있다. (분산은 단위가 만원의 제곱이므로 직관으로 해석하기에는 무리가 있다.) 평균 170만원에서부터 1배의 표준편차 이내에 흩어져 있는 값은 109만원에서 231만원 사이이다. 주어진 자료의 80%가 이 범위 내에 있으므로 표준편차가 자료의 변동을 제대로 설명하고 있다고 할 수 있다.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	2,500*

마지막 관측값이 400만원에서 2,500만원으로 변경된 자료의 경우 분산과 표준편차를 구해보자.

$$\text{표본 분산} = \frac{(100-240)^2+(107-240)^2+\ \cdots\ +(2500-240)^2)}{30-1} = 184047$$

$$\text{표본 표준편차} = \sqrt{184047} = 429$$

평균에서와 같이 2,500만원이라는 전체 자료에 비해 굉장히 큰 값이 포함되자 표준편차가 61만원에서 429만원으로 늘어난 것을 볼 수 있다. 마지

막 관측값 2,500만원을 제외하고는 모든 자료가 100만원에서 255만원 사이, 즉 155만원의 범위 내에 있음을 생각한다면 표준편차가 429만원이라는 것은 자료의 변동을 잘 설명한다고 보기 어렵다. 평균과 같이 분산과 표준편차도 자료에 이상하게 크거나 작은 값이 있다면 자료의 변동을 측정하는 값으로서의 역할을 제대로 한다고 보기 어렵다. 이런 경우에는 중앙값의 경우처럼 하나하나의 값에 덜 민감한 변동 측정도구를 생각해 보아야 할 것이다.

추가로, 다음과 같은 규칙(rule of thumb)을 알아두면 표준편차에 대한 감을 잡는데 도움이 될 것이다.

- 평균을 중심으로 1배의 표준편차구간, $(\bar{x}-1\times SD,\ \bar{x}+1\times SD)$에는 자료의 약 68%가 포함되어 있다.
- 평균을 중심으로 2배의 표준편차구간, $(\bar{x}-2\times SD,\ \bar{x}+2\times SD)$에는 자료의 약 95%가 포함되어 있다.
- 평균을 중심으로 3배의 표준편차구간, $(\bar{x}-3\times SD,\ \bar{x}+3\times SD)$에는 자료의 약 99%가 포함되어 있다.

그런데 이 규칙은 자료의 분포가(다음 장에서 배우게 되는 히스토그램의 모습이) 종이나 산 모양 등의 대칭의 형태일 때만 특히 유효하다. 주어진 예에서 확인해 보면 1배 표준편차의 범위 내에 자료의 80%가 있고 2배 표준편차의 범위 내에 자료의 97%가 있음을 알 수 있다. 한달생활비 자료가 오른쪽으로 약간 비대칭이지만 어느 정도 적용가능하다고 할 수 있을 것이다.

❷ 변동계수

자료의 값의 변화에 덜 민감한 측정도구를 생각하기 전에 한 가지 이야기를 더 살펴보자. 계속된 예를 가지고 이야기를 진행하도록 하자. 자료가 4인가구 도시근로자 가구의 한달 생활비가 아닌 어떤 대학 동아리 학생들의 한달 용돈이라고 해 보자. 학생들의 용돈은 가구의 한달 생활비와는 규모가 다르므로 10으로 나누어 10만원에서 40만원의 값을 갖는 자료로 바꾸었다.

10.0	10.7	11.0	11.2	11.8	12.2	12.4	13.0	13.2	13.6
14.0	14.4	14.8	14.9	14.9	15.1	16.4	16.8	17.2	17.6
18.0	18.4	20.0	20.5	21.9	22.5	23.5	24.5	25.5	40.0

자료에서 분산과 표준편차를 구해보자. 10으로 나누어졌지만 값은 동일하므로 값은 다음과 같다.

$$\text{표본 분산} = \frac{(10.0-17.0)^2+(10.7-17.0)^2+\cdots+(40.0-17.0)^2)}{30-1} = 37.37$$

$$\text{표본 표준편차} = \sqrt{37.37} = 6.1$$

앞의 경우에서처럼 동아리 학생들의 한달 용돈은 평균17만원을 중심으로 6.1만원만큼 흩어져 있다고 할 수 있다.

여기서 한 가지 질문을 해보자. 그렇다면, 4인가구 도시근로자 가구의 한달 생활비와 대학 동아리 학생들의 한달 용돈의 변동은 서로 다르다고 할 수 있는가? 도시근로자 가구의 한달 생활비의 표준편차가 61만원이고 동아리 학생들의 한달 용돈의 표준편차가 6.1만원이므로 가구의 생활비의 변동이 더 크다고 할 수 있을까? 10으로 나누기는 했지만 동일한 자료를 사용했음을 기억한다면 직접적으로 표준편차의 크기로 자료의 변동을 비교하는 것은 의미가 없다고 할 수 있다. 가구에서 학생으로 바뀌면서 평균이 170만원에서 17만원이 된 것처럼 전체적인 금액의 규모가 작아져 표준편차의 크기도 따라서 작아졌을 뿐 자료의 퍼져있는 정도는 동일하다. 따라서 이를 고려한 변동 측정도구를 생각해보게 되는데 그 값이 변동계수(coefficient of variation; CV)이다. 변동계수는 다음과 같이 구해진다.

$$CV = \frac{\text{표준편차}}{\text{평균}}$$

식에서 알 수 있듯이 변동계수는 평균에 비하여 표준편차가 얼마나 큰 가를 나타내는 값이다. 평균을 자료의 중심을 나타내는 대푯값으로 본다면 자료가 전체적으로 어떤 값들을 갖는가를 고려한 변동 측정도구라고 할 수 있다. 또한 표준편차와 평균이 같은 단위를 갖고 있으므로 변동계수는 단위가 사라지며 이는 서로 다른 단위를 가지는 자료들의 변동을 직접적으로 비교할 수 있게 해 준다.

도시근로자 가구의 한달 생활비와 동아리 학생들의 한달 용돈에서 변동계수를 구해보면 다음과 같다.

$$CV_{\text{한달생활비}} = \frac{61}{170} = 0.36$$

$$CV_{\text{한달용돈}} = \frac{6.1}{17} = 0.36$$

두 값이 같은 것을 볼 수 있다. 즉 평균을 감안한다면 도시근로자 가구의 한달 생활비와 동아리 학생의 한달 용돈의 변동은 같다고 할 수 있다.

❸ 다섯숫자 요약

표준편차가 평균과 같이 굉장히 크거나 작은 값들에 영향을 받는다는 점을 고려하여 극단값들에 덜 영향을 받는 변동 측정도구를 생각해보자. 1절의 학생들과 교사의 대화에서 범위에 대한 대화 중에 철수가 '보통사람들'이라고 표현한 가운데 50%의 값들의 범위에 대한 이야기가 있었다. 자료의 50%가 분포하는 범위 다른 말로 사분위범위(IQR)를 자료의 변동 측정도구로 사용할 수 있을 것이다. 이를 위해서 우리에게 익숙하기도 한 다섯숫자들을 먼저 살펴보자.

> 최솟값(min), 하위 25%값(Q1), 중앙값(Q2; M), 상위25%값(Q3), 최댓값(max)

최솟값(minimum)은 가장 작은 값, 자료에서 그 아래로 더 이상 값이 존재하지 않는 값을 뜻하고 최댓값(maximum)은 자료에서 가장 큰 값을 뜻한다. 위의 다섯숫자 중에서 Q1, Q2(M), Q3를 사분위수(quartile)라고 하는데 이는 자료를 순서대로 나열하여 4등분하는 점들을 말한다. 제1사분위수(Q1)는 하위 25%의 자료가 있는 첫 번째 4등분점을 의미하며, 중앙값(Q2; M)은 자료를 반(50%)으로 가르는 한 가운데의 값이다. 제3사분위수(Q3)는 하위 75%, 즉 상위 25%가 있는 세 번째 4등분점이다. 따라서 사분위범위는 $IQR = Q3 - Q1$ 으로 계산할 수 있으며 이 범위 내에 자료의 50%가 존재한다.

계속된 한달 생활비의 자료를 통하여 확인해 보자.

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	400

다섯숫자 요약과 범위, 그리고 사분위범위는 다음과 같다.

$$(\min,\ Q1,\ M,\ Q3,\ \max) = (100,\ 130.5,\ 150,\ 196,\ 400)$$

범위는 최댓값(max)에서 최솟값(min)을 뺀 값, 즉 400-100=300이 되는 반면 사분위수범위 IQR 은 196-130.5=65.5가 된다.

여기서 다시 한 번 $Q1$과 $Q3$를 구하는 과정을 보자. 먼저 중위수는

$\frac{(30+1)}{2} = 15.5$ 번째 값이 되며, 이 값을 깊이(d)라고 부른다. $Q1$과 $Q3$의 깊이는 다음과 같이 구해진다.

$$d(Q1) = \frac{d(M)+1}{2} = \frac{15.5+1}{2} = 8.25$$

따라서 작은 쪽에서부터 8번째 값과 9번째 값의 4분의 1지점이 $Q1$이 되고, 큰 쪽에서부터 8번째 값과 9번째 값의 4분의 1지점이 $Q3$가 된다.

$$Q1 = 130 + \left\{(132-130)\times\frac{1}{4}\right\} = 130.5$$

$$Q3 = 200 - \left\{(200-184)\times\frac{1}{4}\right\} = 196$$

한편 2,500이라는 특이치가 들어간 상태에 이들 요약값들을 구하면,

100	107	110	112	118	122	124	130	132	136
140	144	148	149	149	151	164	168	172	176
180	184	200	205	219	225	235	245	255	2,500*

다섯숫자요약은

$$(\min,\ Q1,\ M,\ Q3,\ \max)^* = (100,\ 130.5,\ 150,\ 196,\ 2{,}500)$$

이다. 범위는 2,500-100=2,400이므로 300에 비하면 엄청나게 커지지만, 사

분위범위는 196-130.5=65.5로 동일함을 알 수 있다. 자료에 극단값(들)이 포함되어 있는 경우는 범위보다는 이 값들에 영향을 덜 받는 사분위범위가 변동을 측정하는 적절한 요약값이라고 할 수 있다.

참고자료

- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석, 경문사.
- 통계교육원(2008), 통계와 정책 표준교재.
- 허명회 · 문승호(2003), 탐색적 자료분석(EDA), 자유아카데미.

제 5 장 그래프로 자료 보기

5-1. 통계그래프

학습목표

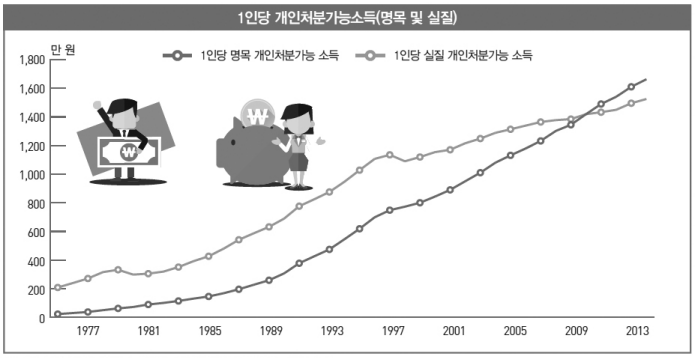
- **각종 보고서나 언론매체에 나와 있는 통계 그래프를 보며 해석하는 토의를 통하여 해석능력을 제고한다.**

1 보고서 속의 그래프

통계보고서나 발표 자료, 언론매체 등에는 많은 그래프들이 나타나는데 이들 대부분은 통계를 나타내거나 자료의 분포를 표현하는 것이다. 보고서 안에서 발견되는 대표적인 그래프들을 보며 토의를 진행하자.

1. 첫 번째 그래프는 통계청에서 발간한 “통계로 본 광복 70년 한국사회의 변화”에서 소득, 소비, 물가 항목을 나타낸 그래프이다.

[그림 5-1]
통계로 본
광복 70년 한국사회의
변화 그래프 (1)

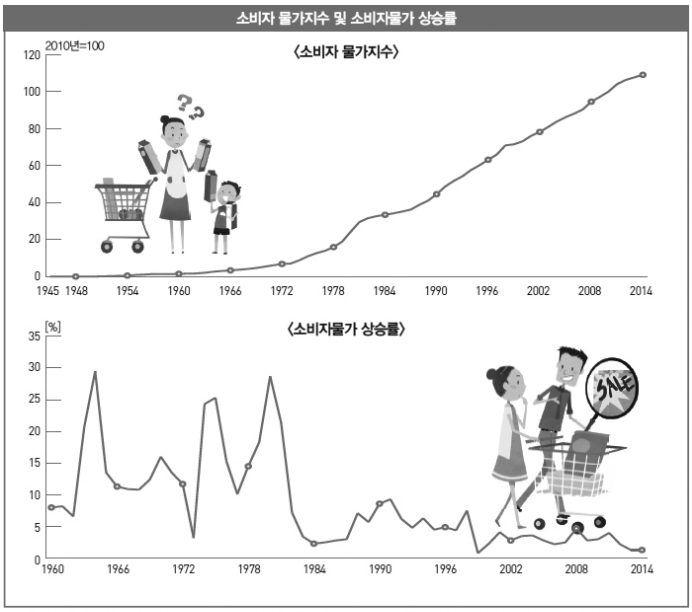


주 : 1인당 실질개인처분가능소득은 소비자 물가지수(2010년=100) 적용.
출처 / 한국은행 〈국민계정〉, ECOS(ecos.bok.or.kr).

이 그래프는 1975년부터 2014년까지의 1인당 개인처분가능소득을 점으로 표시하고 각 점들을 선으로 이어 그린 것이다. 그래프에서 볼 수 있듯이 1975년부터 2014년까지 1인당 개인처분가능 소득은 전체적으로 증가하고 있음을 한 눈에 알 수 있다. 이러한 그래프를 꺾은선 그래프라고 한다. 꺾은선 그래프는 이처럼 연속형 변수의 시간에 따른 변화를 확인할 때 유용하게 사용된다. 우리의 관심이 시간에 따른 변화일 경우 꺾은선 그래프를 통하여 전체적인 추세, 상승과 하락, 시간에 따라 반복되는 변동 등의 규칙적인 형태 등을 알아볼 수 있다. 위의 그래프는 1인당 개인처분가능 소득을 명목과 실질로 나누었으며, 명목 개인처분가능 소득은 짙은 색의 꺾은선으로, 실질 개인처분가능 소득은 옅은 색의 꺾은선으로 구분하여 표시하였다. 꺾은선의 기울기를 살펴보면 1인당 실질 개인처분가능 소득의 경우 1997년 이전의 증가추세가 그 이후의 증가추세에 비하여 가파르게 증가했음도 알 수 있다. 이처럼 꺾은선 그래프는 기울기를 통하여 증가추세를 확인하게 되므로 세로축의 중간을 절단하거나 각 축의 눈금 크기를 조정하는 것은 기울기를 가파르거나 완만하게 보이게 만들 위험성이 있으므로 각 축의 눈금을 유의해서 살펴보는 것이 필요하다. 최근에는 1인당 실질 개인처분가능 소득이 명목 개인처분가능 소득보다 높아졌음도 확인할 수 있다. 이러한 경향은 표에서 숫자를 통하여 확인하는 것보다 짧은 시간에 한 눈에 들어오게 된다.

2. 두 번째 그래프는 동일한 발간물의 소비자 물가지수와 소비자물가 상승률을 나타낸 그래프이다.

[그림 5-2]
통계로 본
광복 70년 한국사회의
변화 그래프 (2)



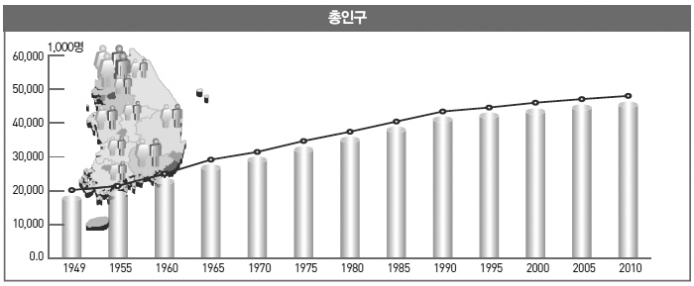
출처 / 통계청, 소비자물가조사, 1945~1964년은 한국은행 경제통계국(2005) 〈숫자로 보는 광복 60년〉 p. 83의 2000년=100 수치를 2010=100 기준으로 재산정한 값임.

번호	1950	1960	1970	1980	1990	2000	2010	2014
소비자물가지수	0.03	1.48	5.38	24.27	44.54	73.10	100	109.04
소비자물가 상승률 (%)	167.5	8.0	16.0	28.7	8.6	2.3	3.0	1.3

이 그래프는 위의 그래프와 달리 그래프의 면을 분할하여 두 개의 꺾은선 그래프를 보여주고 있다. 위쪽의 그래프를 보면 소비자 물가 지수는 계속 증가하고 있음을 알 수 있다. 또 아래의 그래프를 통하여 소비자물가 상승률은 상승과 하락의 변동이 있으며 특히 1984년 이전에 변동의 폭이 크고 그 이후에는 변동의 폭이 작아졌음을 확인할 수 있다. 따라서 소비자 물가 지수는 시간이 흐름에 따라 증가하고 있지만 그 상승률은 변동의 폭이 시간이 흐름에 따라 줄어들고 있음을 함께 확인하게 해 준다.

3. 세 번째 그래프는 동일한 발간물에서 총인구의 변화를 나타낸 그래프이다.

[그림 5-3]
통계로 본
광복 70년 한국사회의
변화 그래프 (3)

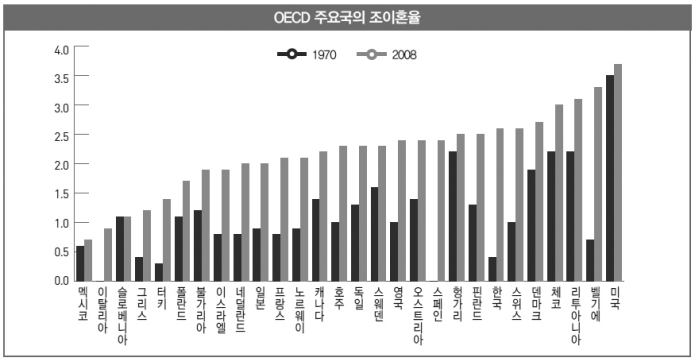


주 : 1) 인구는 내국인 인구를 기준으로 함.
2) 전체 인구에는 '연령 미상'이 포함되어 있어 연령집단별 인구의 합계와 차이가 있음.
3) 1949년 자료의 경우 64세 인구까지만 자료가 제공되어, 만 60~69세 인구는 60~64세 인구임.
4) 1960년 자료의 경우 자료의 제약으로 인해 연령집단별 인구에서는 외국인 인구가 포함된 수치를 이용하였음.
5) 1970년에는 해외주재공관원 및 그 가족(약 1,200인)을 제외함.
출처 / 통계청, 「인구주택총조사」, 각 연도.

그래프를 보면 꺾은선과 함께 막대그래프가 그려져 있다. 막대그래프는 주로 양적인 변수를 질적인 그룹에서 비교하도록 표현하거나 질적인 변수들의 빈도 등을 나타내는데 사용된다. 막대의 높이가 높으면 값이 크거나 빈도가 많다는 것을 의미하며 이렇게 막대의 높이를 통하여 여러 양들을 비교할 수 있게 된다. 이는 시각적으로 감지하게 되는 것이므로 막대의 높이뿐만 아니라 막대의 면적에도 영향을 받게 된다. 그러므로 막대의 너비를 동일하지 않게 표현하거나 색을 다르게 표현하는 것은 왜곡의 위험이 있으므로 주의해야 한다.

4. 네 번째 그래프는 동일 발간물의 OECD 주요국의 조이혼율을 나타낸 그래프이다.

[그림 5-4]
통계로 본
광복 70년 한국사회의
변화 그래프 (4)

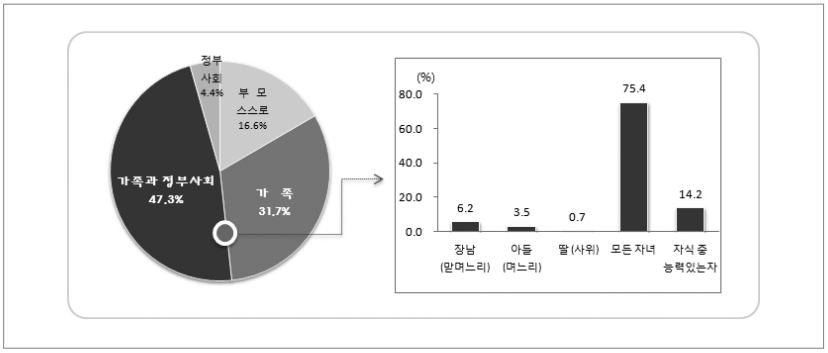


주 : 1) 조이혼율은 인구 1,000명당 이혼 건수임.
2) 2008년 수치 중 영국, 호주, 캐나다, 일본, 그리스, 멕시코 등은 2007년 자료이고, 미국, 프랑스, 이스라엘 등은 2006년 자료임.
3) 국제기구에서 작성한 나라별 지표는 국제비교를 위해 특정 가정에 따라 조정된 자료로서, 주요 · 보조지표의 수치, 수록기간, 단위 등이 다를 수 있음.
출처 / OECD 「http://www.oecd.org/」, 2014 .06; 「OECD Family Database 2014」.

이 그래프는 각 나라별 조이혼율을 막대의 크기로 표현하여 조이혼율이 나라별로 어떻게 다른지를 나타내었다. 또 1970년의 조이혼율은 짙은 색으로 2008년의 조이혼율은 옅은 색으로 구분하여 각 나라에서 조이혼율이 시간이 지남에 따라 어떻게 변화하였는지를 함께 표현하였다. 또한 2008년도의 조이혼율을 낮은 나라부터 높은 나라까지 순서대로 나열함으로써 1970년대와 2008년의 양상이 다름을 알 수 있게 해 준다.

5. 다섯 번째 그래프는 2014년 사회조사의 부모 부양에 대한 견해를 나타낸 그래프이다.

[그림 5-5]
2014 사회조사 보고서 – 부모부양에 대한 견해 그래프 (1)



이 그래프는 부모의 노후 생계를 주로 누가 돌보아야 하는지에 대한 응답 항목 '①스스로 해결/ ②가족/ ③가족과 정부·사회/ ④정부·사회/ ⑤기타' 중 각 항목을 응답한 비율을 원그래프로 표현하였다. 가족이 포함된 응답 항목에 대부분의 사람들이 응답하였음을 쉽게 확인할 수 있다. 원그래프는 원을 조각으로 나누어 표현하므로 각도를 이용하여 그 크기 혹은 비율을 표현한다. 따라서 원이 여러 조각으로 나누어지는 경우는 크기 비교가 용이하지 않으므로 원그래프는 유용하지 않다. 원그래프는 막대그래프에 비해서 크기를 명확히 비교하기가 어렵기 때문에 위의 그래프에서 보듯이 백분율을 함께 표현하여 준다. 또 이 그래프의 오른쪽은 가족이 포함된 응답을 한 경우 가족 중 누가 부모님의 노후생계를 돌보아야 하는지의 추가질문에 대한 항목별 응답 비율을 막대그래프로 나타내었다. 여기서 막대의 높이는 백분율을 나타내며 막대의 위쪽에 각 백분율을 함께 표현하여 주었다. 이처럼 막대그래프와 원그래프는 범주를 갖는 변수에 대하여 표현하고자 할 때 특히 유용하다.

5-2. 히스토그램

학습목표

- **연속형 자료의 시각화로서 히스토그램 및 줄기와 잎 작성법을 전달하여 작성 능력을 제고한다.**

1 히스토그램

토의를 시작하기 전에 이미 익숙한 용어들이지만 다시 한 번 다음 용어들을 정의해 보자.

- 도수: 자료 중에서 특정한 값이나 특정한 구간내의 값이 나타난 횟수 (자료 중에서 특정한 값이나 특정한 구간내의 값을 갖는 개체의 수)
- 도수분포표: 서로 다른 관찰값이나 서로 다른 관찰값의 범주와 그에 대응하는 빈도수를 기술한 표
- 히스토그램(histogram): 도수분포표를 이용하여 자료의 분포를 나타낸 그래프 (통계학에서는 질적 자료의 막대그래프와 구분하여 양적 자료를 그린 막대그래프를 히스토그램이라고 부른다.)

1. 도수분포표

(1) 도수분포표

정책지지 관련 자료에서는 다음과 같은 질적 자료에 대한 도수분포표가 나온다.

<표 5-1> 도수분포표

찬 성	585
반 대	415
합 계	1,000

그런데 20명의 미혼 동료들의 한 달 생활비를 요약한 다음의 내용을 보자.

150	200	250	250	150	200	150	200	150	300
200	300	200	200	150	200	300	250	200	250

① 이 자료는 양적자료이고, 생활비는 연속형 변수이다.

② 그럼에도 불구하고 서로 다른 종류의 숫자값이 4가지(150, 200, 250, 300)뿐이므로 구간을 나눌 필요는 없다고 판단된다.

③ 그래서 질적 자료(범주형 자료) 같이 도수분포표를 작성하였다.

<표 5-2>
한달 생활비
도수분포표

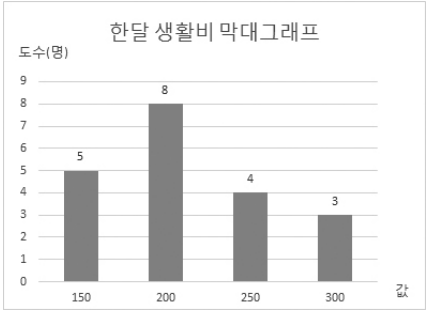
값	도수	누적도수	상대도수	누적상대도수
150	5	5	0.25	0.25
200	8	13	0.4	0.65
250	4	17	0.2	0.85
300	3	20	0.15	1
합 계	20		1	

표를 살펴보면 200에서의 도수 8은 20명중 8명이 200만원을 생활비로 쓴다는 것이고, 누적도수 13은 이 자료에서 200만원 이하의 생활비를 쓰는 응답자의 수를 의미한다. 200에서의 상대도수 0.4는 이 자료에서 200만원 생활비를 쓰는 응답자의 비율을 나타낸다.

$$200\text{만원의 상대도수} = \frac{8}{20} = 0.4$$

④ 자연스럽게 이 도수분포표를 수평축을 관찰값, 수직축을 도수로 하여 막대그래프를 그리면 다음과 같다.

[그림 5-7]
한달 생활비
막대그래프



(2) 계급구간의 필요성

이번에는 크기가 100인 표본으로부터 조사된 다음의 생활만족도 자료를 요약해보자.

35	32	31	30	25	41	7	17	18	20
8	9	13	14	62	63	63	65	67	66
69	75	15	15	22	23	24	33	34	35
35	36	37	26	27	39	39	38	29	29
27	45	46	47	45	81	80	79	14	17
82	8	43	44	49	48	47	49	44	42
49	50	50	51	52	53	52	49	48	49
54	53	56	56	59	59	60	61	76	77
78	79	80	22	83	84	35	47	85	83
66	77	75	76	77	60	56	52	92	93

① 이 자료는 양적자료이고 만족도는 연속형 변수이다.

② 서로 다른 종류의 숫자 값이 상당히 많으므로 구간을 나누는 것이 좋겠다.

③ 적당한 계급(계급구간이라고도 한다)을 정하자.

이 자료는 최솟값이 7이고 최댓값이 93이므로, 5에서 95까지로 하고, 6구간 정도로 하면 계급구간의 폭이 15(=(95-5)/6)가 된다.

여기서 왜 구간의 개수를 6개로 했는지, 8개로 하면 안 되는지 등의 질문이 나올 수 있다. 물론 8개로 해도 괜찮다. 구간의 수를 결정하는 특별한 규칙은 없다.

6개로 하게 된 생각은 이런 것이었다.

- 자료의 크기가 100이고 최댓값과 최솟값을 포함하여 구간을 5~95로 잡았기 때문에 구간의 개수를 6개로 하면 구간의 폭이 15가 되어 해석하기가 좋을 것이다.
- 자료의 크기가 100이므로 구간의 개수를 6개로 하면 각 구간별로 해당하는 도수가 평균 16정도가 된다.

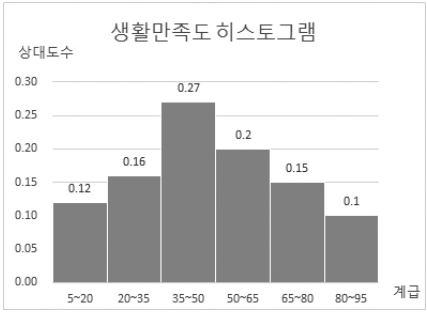
• 구간의 개수를 6개로 정한 것은 일단 해보는 것이고 6개, 7개, 8개의 각 경우를 비교하면서 적절한 구간의 개수를 정하려고 했다. 이 때 '적절한'이란 전체적인 자료의 특징을 잘 보여주며 도수가 0에 가까운 구간이 나타나지 않는 상태를 의미한다.

그럼 일단 구간의 개수를 6개로 하여 도수분포표와 히스토그램을 작성해 보자.

<표 5-3> 생활만족도 도수분포표 (1)

계급(이상~미만)	도수	누적도수	상대도수	누적상대도수
5~20	12	12	0.12	0.12
20~35	16	28	0.16	0.28
35~50	27	55	0.27	0.55
50~65	20	75	0.20	0.75
65~80	15	90	0.15	0.90
80~95	10	100	0.10	1.00
합 계	100		1	

[그림 5-8] 생활만족도 히스토그램 (1)



2. 히스토그램

위의 생활만족도 히스토그램을 보고 무슨 말을 할 수 있을까?

① 전체적인 모양은 산 모양으로 대체로 대칭적이다.

② 35에서 50사이에 값들이 가장 많다.

③ 자료의 절반을 나누는 중앙값은 35에서 50사이에 있을 것 같다.

여기서 중앙값을 아는 방법은 막대에 해당되는 수직축의 값을 더해서 알 수 있다. 첫 번째 막대의 수직축 값과 두 번째 막대의 수직축 값, 그리고 세 번째 막대의 수직축 값의 합은 0.12+0.16+0.27=0.55로 합이 0.5를 조금 넘게 된다. 그러니까 30에서 55사이의 막대에 중앙값이 있다는 것을 알 수 있다.

히스토그램을 바라볼 때 중요한 관점은 거의 다 언급된 것 같다. 주요관점만 정리해 보자.

① 전체적인 형태로서 모양, 대칭성 여부, 봉우리 개수

② 중심의 위치: 중앙값이 즉 50%되는 점이 어느 정도에 있는지

③ 다름(퍼짐)의 정도: 작은 쪽과 큰 쪽의 대체적인 범위

④ 이상점의 유무

여기서 이상점의 유무라는 것은 말 그대로 이상점이 나타나는지를 본다는 것이다. 생활만족도 자료를 나타내는 히스토그램으로부터는 특이하게 보이는 관찰치가 나타나지는 않는다. 그런데 다음의 가상적인 자료를 한 번 보자. 히스토그램의 복습 겸 한번 그려보도록 하자.

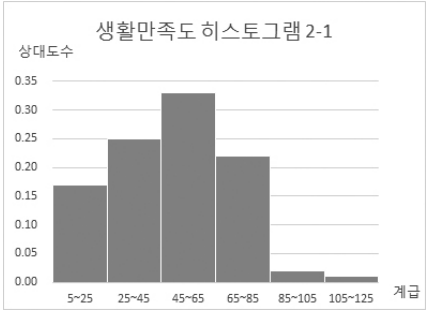
35	32	31	30	25	41	7	17	18	20
76	9	13	14	62	63	63	65	67	66
69	75	15	15	22	23	24	33	34	35
35	36	37	26	27	39	39	38	29	29
27	45	46	47	45	81	80	79	14	17
82	8	43	44	49	48	47	49	44	42
49	50	50	51	52	53	52	49	48	49
54	53	56	56	59	59	60	61	76	77
78	79	80	22	83	84	35	47	85	83
66	77	75	8	77	60	56	52	92	123

최솟값이 7이고 최댓값이 123이므로 5에서 125까지 6구간으로 나눈다면 계급구간의 폭은 20(=(125-5)/6)으로 하면 되겠다. 도수분포표 및 히스토그램을 작성하면 다음과 같다.

<표 5-4> 생활만족도 도수분포표 (2)

계급(이상~미만)	도수	누적도수	상대도수	상대누적도수
5~25	17	17	0.17	0.17
25~45	25	42	0.25	0.42
45~65	33	75	0.33	0.75
65~85	22	97	0.22	0.97
85~105	2	99	0.02	0.99
105~125	1	100	0.01	1.00
합 계	100		1	

[그림 5-9] 생활만족도 히스토그램 (2)



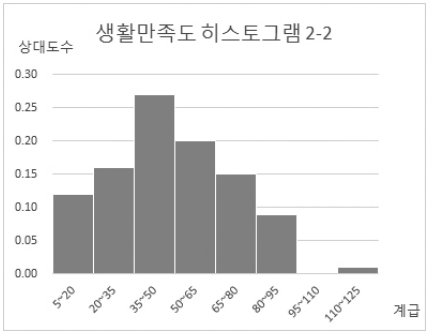
도수분포표나 히스토그램에서 105~125 사이에 하나의 개체가 나타난다. 전체 자료 중에서 다소 튀는 값인 것 같다. 이것을 이상점(outlier)이라고 한다. 따라서 이 구간에 속한 값 123에 대해서 검토해 볼 필요가 있다. 구체적으로 말하면 혹시 입력에 오류가 있었는지, 기록에 문제가 있었는지, 부실한 관찰값인지, 이와 같이 큰 값을 갖는 개체가 실제로 있는지 등을 검토하게 된다.

동일한 자료로 계급구간을 달리하여 도수분포표와 히스토그램을 그려보자.

<표 5-5>
생활만족도
도수분포표 (3)

계급(이상~미만)	도수	누적도수	상대도수	상대누적도수
5~20	12	12	0.12	0.12
20~35	16	28	0.16	0.28
35~50	27	55	0.27	0.55
50~65	20	75	0.20	0.75
65~80	15	90	0.15	0.90
80~95	9	99	0.09	0.99
95~110	0	99	0.00	0.99
110~125	1	100	0.01	1.00
합 계	100		1	

[그림 5-10]
생활만족도
히스토그램 (3)

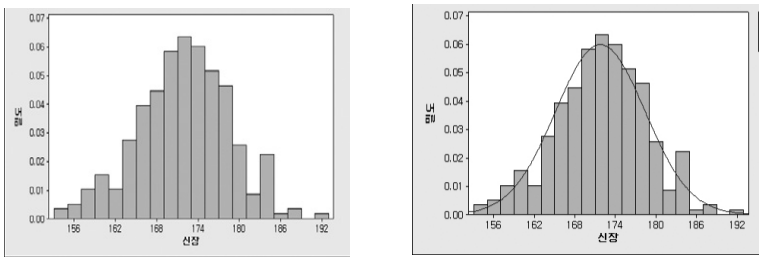


이렇게 구간을 나누니까 이상점이 더 확연히 나타난다. 히스토그램을 작성하는 1차 목적은 수집된 자료를 효과적으로 기술하는 것이지만 히스토그램을 설명, 해석하는 단계에서도 이상점이 있는지를 조사하려는 마음을 계속 가져야 한다.

3. 히스토그램 사례

(1) 2013 체력실태조사 신장(성인 남자)에 대한 히스토그램

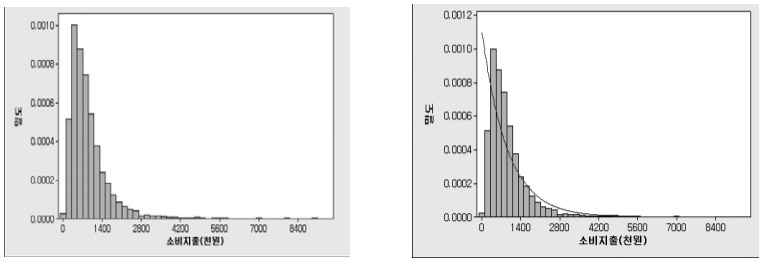
[그림 5-11]
체력실태조사 신장
히스토그램



위의 히스토그램을 보면 성인 남자의 신장은 가운데가 솟아 있는 대칭의 모습을 보이며, 봉우리는 하나로 나타난다. 가운데를 중심으로 큰 쪽과 작은 쪽이 퍼져있는 정도는 비슷하게 나타나며 190이상의 값이 다른 값들과 떨어져있지만 이상점으로 보이지는 않는다. 10장에서 다루겠지만, 신장의 히스토그램은 전체적으로 오른쪽 그림에 그려진 곡선으로 표현할 수 있다.

(2) 2015 가계동향조사 소비지출에 대한 히스토그램

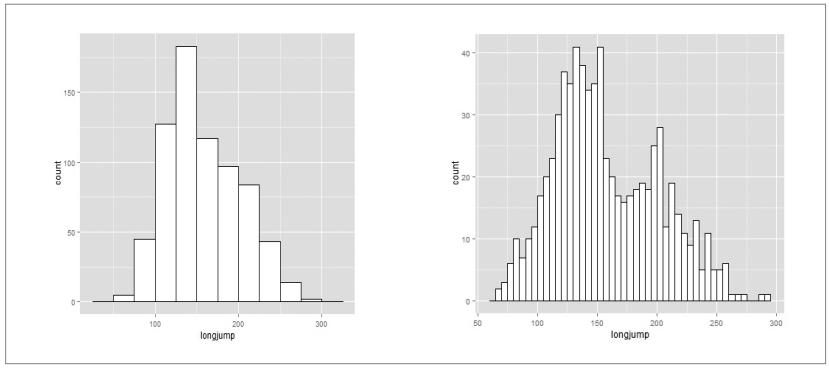
[그림 5-12]
가계동향조사
소비지출 히스토그램



이 히스토그램은 신장의 히스토그램과 달리 왼쪽이 솟아 있고 대칭이 아니다. 중심은 왼쪽에 위치하고 있으며 오른쪽으로 길게 뻗어 있다. 즉, 중심보다 큰 값들이 퍼져있는 정도가 작은 값들이 퍼져있는 정도보다 크다. 봉우리는 하나이다.

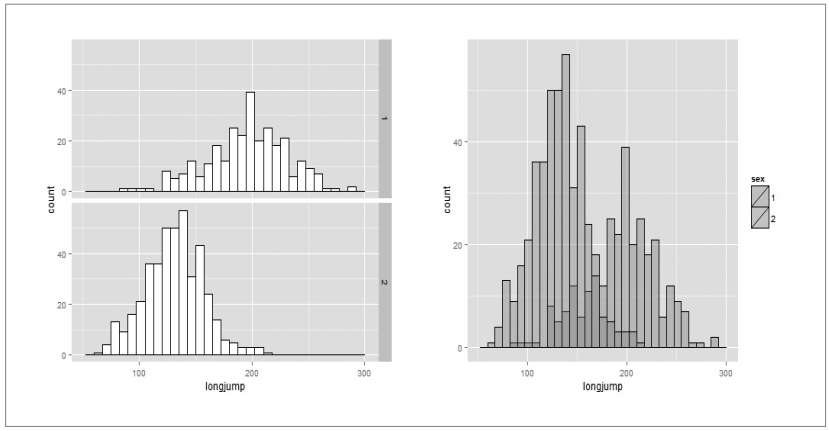
(3) 2013 체력실태조사 제자리멀리뛰기 히스토그램

[그림 5-13]
체력실태조사
제자리 멀리뛰기
히스토그램 (1)



[그림 5-13]에서 왼쪽의 히스토그램을 보면 [그림 5-11]에 있는 신장 히스토그램과 비슷한 모습으로 보인다. 봉우리는 하나이며 약간 왼쪽으로 치우쳐 있긴 하지만 대체적으로 대칭의 모습을 보인다. 그러나 같은 자료를 계급의 구간을 더 좁게 하여 그린 [그림 5-13]의 오른쪽 히스토그램을 보면, 왼쪽의 히스토그램과 달리 봉우리가 두 개로 나타나는 것을 알 수 있다. 이런 경우에는 혹시 자료가 이질적인 집단으로 이루어져 있는 것이 아닌지 생각해 볼 수 있다. 실제 이 자료는 남녀의 제자리멀리뛰기 기록을 함께 그린 것으로 남녀를 구분하여 히스토그램을 다시 그리면 [그림 5-14]와 같이 두 집단으로 나누어지는 것을 볼 수 있다. 동일한 자료이더라도 계급의 구간이 달라지면 히스토그램에서 얻을 수 있는 정보가 달라질 수 있다.

[그림 5-14]
체력실태조사
제자리 멀리뛰기
히스토그램 (2)



❷ 줄기와 잎 그림

히스토그램이 표현하는 내용을 조금 더 상세하게 표현하는 방법이 있는데 특히 자료의 개수가 적은 경우에 유용한 줄기와 잎 그림이다.

줄기와 잎 그림을 그리는 방법은 다음과 같다.

① 각 관측값을 마지막 자리 수를 제외한 모든 자리 수로 이루어진 줄기들로 분리하고, 마지막 자리 수는 잎으로 한다. 줄기는 필요한 경우에는 많은 자리 수를 가질 수 있으나 각 잎은 오직 하나의 자리 수만 갖는다.

② 위부터 가장 작은 줄기들을 세로 열로 쓰고 이 열의 오른쪽에 수직선을 긋는다.

③ 각 줄기에서 잎들을 왼쪽에서 오른쪽으로 증가하는 순서대로 나열한다.

아래의 그림은 2013 체력실태조사에서 임의로 추출한 성인남자 60명의 체중에 대한 줄기와 잎 그림과 히스토그램이다.

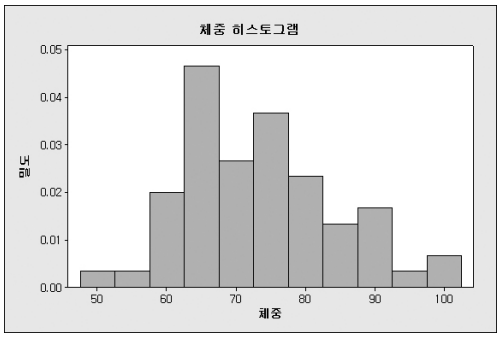
[그림 5-15]
체력실태조사 체중
줄기와 잎 그림 (1)

줄기-잎 그림: 체중

줄기-잎 그림: 체중 N = 60
잎 단위 = 1.0

```
 1   5   2
 4   5   689
 12  6   01123334
 28  6   55555666677778999
(8)  7   02222233
 24  7   666777778999
 13  8   0234
 9   8   77889
 4   9   04
 2   9   8
 1   10  0
```

[그림 5-16]
체력실태조사 체중
히스토그램 (1)



줄기와 잎 그림은 히스토그램을 옆으로 누인 것과 같다. 히스토그램을 그리기위해 선택된 계급들은 줄기그림의 줄기와 같다. 히스토그램처럼 줄기와 잎 그림을 통해서도 전체적인 형태 그리고 이상점들을 찾아야 한다. 자료의 자릿수가 많은 경우에는 반올림하여 좀 더 유연하게 반올림한 마지막 자리 수를 잎으로 할 수 있다.

줄기와 잎 그림의 가장 큰 장점은 실제 관측값이 나타난다는 것이다. 히스토그램에서는 그 값을 알 수 없으나 줄기와 잎 그림을 통해서는 알 수 있다. 또한 줄기와 잎 그림은 히스토그램보다 더 빨리 작성될 수 있다.

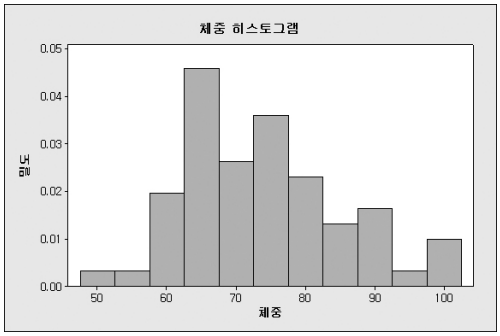
줄기와 잎 그림은 한자리 또는 여러 자리 숫자를 줄기로 사용할 수 있으며 이것은 자동적으로 계급들이 되기 때문에 분포를 잘 나타내지 못할 수도 있다. 아래의 그림처럼 관측값들이 많은 큰 자료의 경우에는 줄기가 너무 많은 잎들을 갖기 때문에 줄기와 잎 그림을 사용하기 어렵다.

[그림 5-17]
체력실태조사 체중 줄기와 잎 그림 (2)

```
줄기-잎 그림: 체중  N  = 3674
잎 단위 = 1.0

  8     3   66689999
 50     4   00011111122222222223333333333444444444444
298     4   55555555555555555555555555555566666666666666666666666666666666666666+
866     5   00000000000000000000000000000000000000000000000000000000000000000000+
1474    5   55555555555555555555555555555555555555555555555555555555555555555555+
(583)   6   00000000000000000000000000000000000000000000000000000000000000000000+
1617    6   55555555555555555555555555555555555555555555555555555555555555555555+
1080    7   00000000000000000000000000000000000000000000000000000000000000000000+
657     7   55555555555555555555555555555555555555555555555555555555555555555555+
336     8   00000000000000000000000000000000000000000000000000000011111111111111+
170     8   55555555555555555555566666666666666677777777777777777788888888888+
85      9   00000000000000000111111111122222223333344444
39      9   555555555566777777888889
14      10  00000001
6       10  557
3       11
3       11  56
1       12  0
```

[그림 5-18]
체력실태조사 체중 히스토그램 (2)



5-3. 상자그림

학습목표

• 상자그림 작성법을 상세히 설명하여 상자그림의 해석능력을 제고한다.

앞 절의 히스토그램을 꼭 그리지 않더라도 우리가 4장에서 학습한 평균, 중앙값, 분산, 사분위범위, 다섯숫자요약 등의 숫자를 통해서도 자료가 어떤 모양새를 하고 있는지 어느 정도는 알아볼 수 있다. 그러나 이런 숫자들도 그림으로 일목요연하게 표현한다면 더 쉽게 자료를 이해할 수 있을 것이며, 그 방법이 바로 미국의 통계학자 J. Tukey가 제안한 상자그림(box and whisker plot)이다.

일단 상자그림을 그리는 방법부터 알아보자.

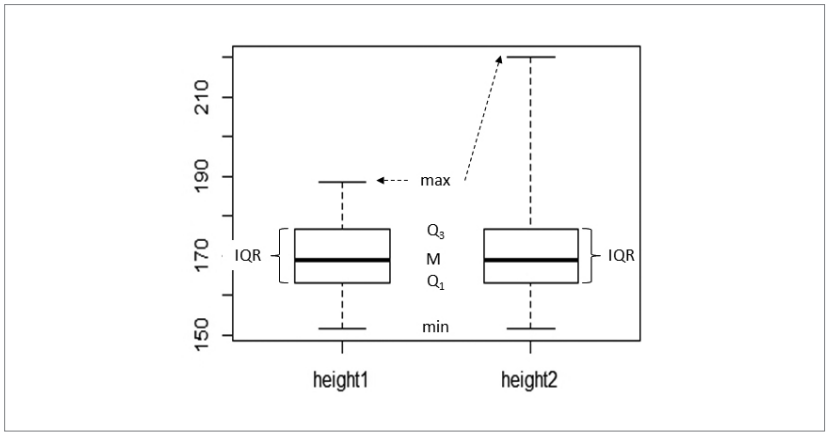
① 상자그림을 그리기 위해서는 먼저 (min, *Q1*, *M*, *Q3*,max)의 다섯숫자가 필요하다.

② 다섯숫자들을 구하면 그 중 제1사분위수(Q_1)과 제3사분위수(Q_3)를 이용하여 상자를 그린다. 상자의 왼쪽 끝이 Q_1이 되고 상자의 오른쪽 끝이 Q_3가 된다. 그러면 상자의 길이가 바로 *IQR*이 되는 것을 알 수 있다.

③ 중앙값(*M*)을 상자 안에 "+"나 선으로 그어 표시한다. 필요하다면 평균은 그 안에 점이나 동그라미 등의 표식으로 따로 표현하면 두 값을 비교하는데 도움이 될 수 있다.

④ 최솟값(min)과 최댓값(max)을 표시하고 상자의 끝부분과 선으로 잇는다.

[그림 5-19]
상자그림 (1)



최솟값과 최댓값이 너무 크거나 작을 경우에는 ④대신 상자와 선으로 연결하지 않고 따로 표시하여 상자그림을 그릴 수도 있다. 너무 크거나 작다의 기준은 상자의 끝 즉, Q_1과 Q_3에서부터 $1.5\times IQR$을 벗어나는가이다. (이는 정규분포의 경우 $\mu \pm 1.5\times IQR$을 벗어나는 바깥의 확률은 약 0.7%이다.) 이러한 상자그림은 다음과 같이 그린다.

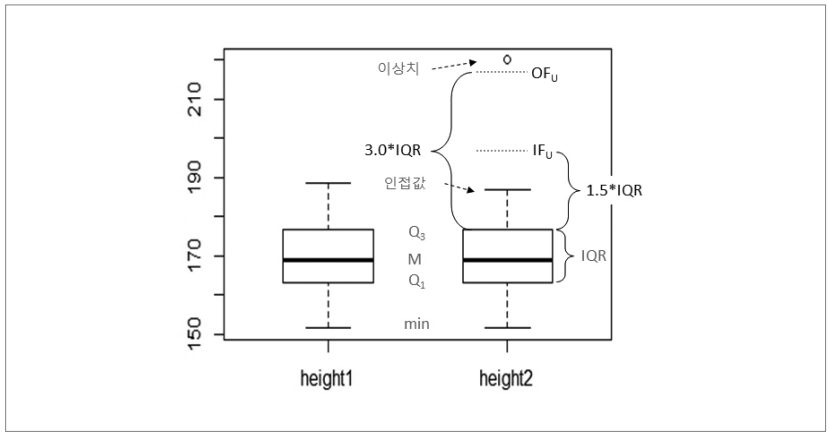
①~③은 동일

④ Q_1과 Q_3에서부터 $1.5\times IQR$ 떨어진 지점을 안 울타리(inner fence; IF)로, $3.0\times IQR$떨어진 지점을 바깥 울타리(outer fence; OF)로 표시한다.

⑤ 최솟값과 최댓값이 안 울타리를 벗어나면 이상치 혹은 특이점으로 본다. 이런 점들은 선으로 긋지 않고 따로 표시하는데 안 울타리에서 바깥 울타리 사이에 있는 경우에는 '*'로, 바깥 울타리를 벗어나는 경우에는 'o'로 표시한다.

⑥ 최솟값과 최댓값이 울타리를 벗어나서 따로 표시하게 되는 경우에는 안 울타리 내에 있으면서 가장 큰 값을 다시 찾아서 상자의 끝과 연결하여 준다. 이 값은 인접값(adjacent value; AV)이라고 한다.

[그림 5-20]
상자그림 (2)



종종 한 변수만이 아니라 여러 변수들을 함께 비교하고 싶거나, 남녀 등의 구분으로 비교하고자 할 때에는 상자그림을 나란히 그려볼 수 있다. 이때는 한 변수에 대한 분포의 모습뿐만 아니라 보다 전체적인 비교도 가능하게 해 준다.

참고자료

- 허명회(1993), 탐색적 방법에 의한 통계자료 분석론, 자유아카데미.
- 허명회 · 문승호 (2003), 탐색적 자료분석(EDA), 자유아카데미.
- 문화체육관광부(2013), 체력실태조사.
- 통계청(2014), 사회조사 보고서.
- 통계청(2015), 통계로 본 광복 70년 한국사회의 변화.

제 6 장

연속형 자료에서 관계 찾기

6-1. 다변량 자료의 특징

학습목표

- **지금까지 개체로부터 한 가지 특성에 관한 변수의 자료를 다루었는데 이 절에서는 동일 개체로부터 두 개 이상의 특성에 관한 변수들의 자료를 수집하였을 때 처리하는 방법을 이해한다.**

1 다변량 자료의 특징

다변량 자료는 동일개체로부터 두 개 이상의 변수에 관하여 수집된 자료를 말하는데 다음의 예를 통하여 다변량 자료의 분석과정에서 그 특징을 찾아보자.

아래의 자료는 문화체육관광부가 공표하는 2013년 국민체력실태조사 자료의 일부로 서울에 거주하는 자신의 건강상태를 건강하다고 평가한 20대와 30대 23명의 신장과 제자리멀리뛰기 항목에 대한 자료이다.

<표 6-1>
신장과
제자리 멀리뛰기
자료 1

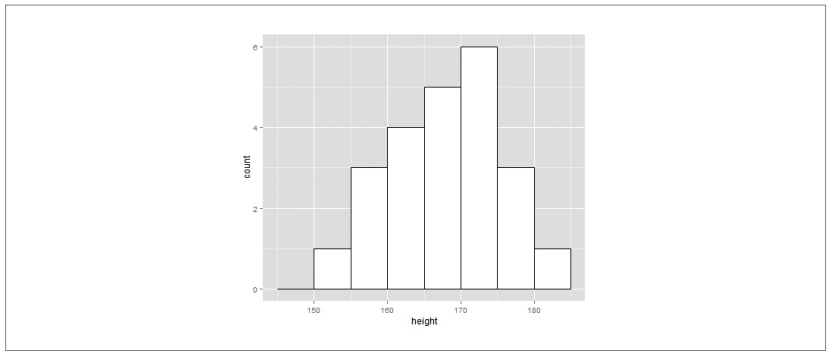
ID	신장(㎝)	제자리 멀리뛰기(㎝)	ID	신장(㎝)	제자리 멀리뛰기(㎝)
1	161.2	209	13	161.5	170
2	157.4	99	14	166.0	155
3	173.0	197	15	177.3	202
4	173.0	194	16	173.7	219
5	152.9	153	17	169.1	210
6	176.5	163	18	170.2	197
7	167.5	130	19	157.0	150
8	175.4	240	20	160.8	150
9	173.2	240	21	161.0	160
10	168.8	210	22	174.3	230
11	156.5	135	23	180.0	230
12	166.1	150			

이 자료로부터 신장 자료에 4장과 5장에서 다룬 방법을 적용하여 일변량 분석을 하면 다음과 같은 결과와 그래프를 얻게 된다.

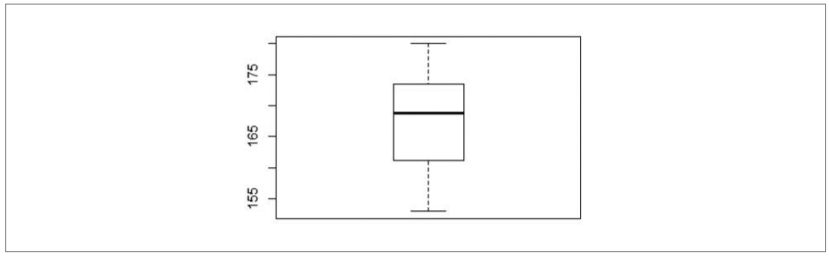
<표 6-2>
신장 일변량 분석 결과

평균	167.5
표준편차	7.68
최솟값	152.9
제1사분위수(Q_1)	161.1
중앙값	168.8
제1사분위수(Q_3)	173.5
최댓값	180.0
사분위범위(IQR)	12.4

[그림 6-1]
신장 히스토그램 (1)



[그림 6-2]
신장 상자그림 (1)



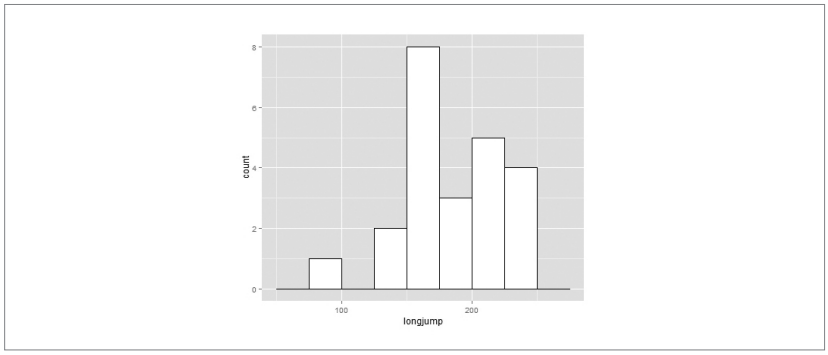
일변량 분석결과와 그래프를 보면 신장은 152.9에서 180.0까지 있다. 평균은 167.5로 중앙값 168.8보다 작으며 그래프를 보면 약간 왼쪽으로 꼬리가 길게 퍼져있는 것을 볼 수 있다. 표준편차는 7.68이고 사분위범위는 12.4이고 히스토그램과 상자그림을 살펴보면 이상치로 의심되는 점은 보이지 않는다.

제자리멀리뛰기 자료에 대하여도 동일한 방법으로 일변량 분석을 하면 다음과 같은 결과와 그래프를 얻게 된다.

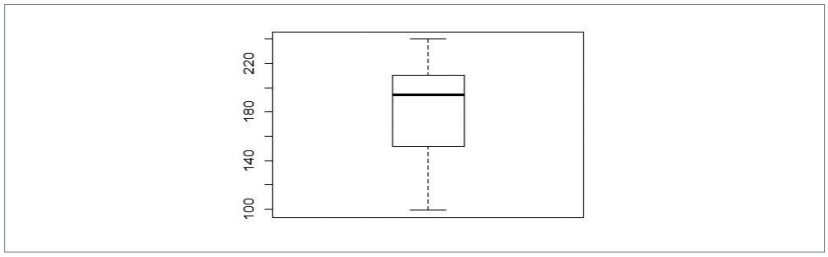
<표 6-3>
제자리멀리뛰기
일변량 분석 결과

평균	182.3
표준편차	38.98
최솟값	99.0
제1사분위수(Q_1)	151.5
중앙값	194.0
제1사분위수(Q_3)	210.0
최댓값	240.0
사분위범위(IQR)	58.5

[그림 6-3]
제자리멀리뛰기
히스토그램 (1)



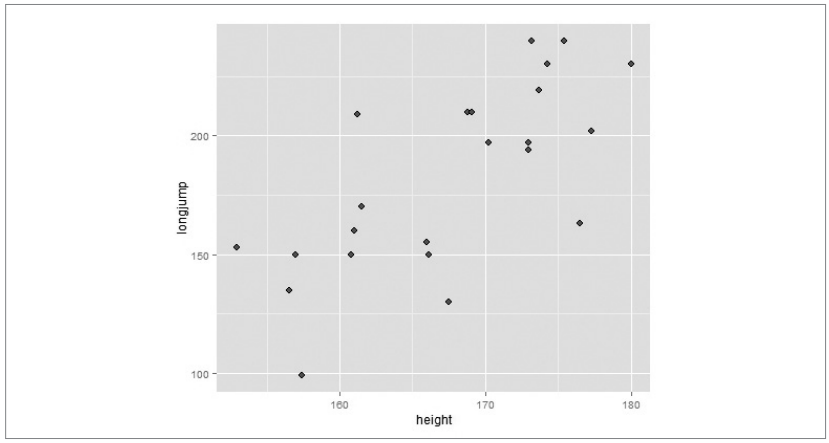
[그림 6-4]
제자리멀리뛰기
상자그림 (1)



일변량 분석결과와 그래프를 보면 제자리멀리뛰기 기록은 99.0에서 240.0까지 퍼져 있다. 평균은 182.3으로 중앙값 194.0보다 작으며 표준편차는 38.98이고 사분위범위는 58.5이다. 히스토그램을 보면 약간 왼쪽으로 꼬리가 길게 퍼져있으며 최솟값인 99.0은 다른 자료와는 조금 떨어져 따로 있는 것을 볼 수 있다. 상자그림을 보면 최솟값 99.0은 이상치로 표시되지는 않았다.

이와 같이 신장과 제자리멀리뛰기 자료로부터 서울에 거주하는 건강상태를 건강하다고 평가한 20대와 30대의 신장과 멀리뛰기 기록에 대한 정보를 각각 얻었다. 그 이외에 다른 유용한 정보는 없겠는가? 키가 크면 멀리뛰기를 잘하지 않을까 하는 생각이 들 수 있다. 당신은 어떻게 생각하는가? 이 생각을 자료로부터 점검하기 위하여 아래와 같은 그래프를 그려본다. 이러한 그래프를 산점도(scatter plot)라고 한다. 산점도는 두 연속형 변수의 관계를 살펴볼 때 유용한 그래프로서 한 변수의 값을 가로축에, 다른 변수를 세로축에 함께 표현하여 두 변수가 서로 어떤 관련성을 보이는지 확인할 수 있게 해 준다.

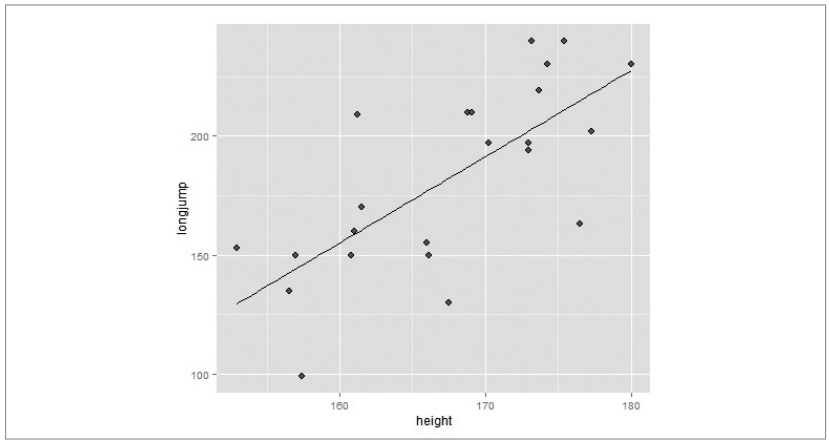
[그림 6-5]
신장과
제자리멀리뛰기
산점도 (1)



이 그래프에서 보면 전반적으로는 신장이 클수록 제자리멀리뛰기 기록도 함께 늘어나는 경향을 볼 수 있다. 그런데 이러한 경향의 배후에는 무시할 수 없는 다양성이 있다. 예를 들자면 신장은 비슷한데도 불구하고 멀리뛰기 기록에서는 차이가 나타나는데 160에서 165사이의 점들을 살펴보면 작게는 150에서 200이 넘는 값까지 다양하게 나타나는 것이 보인다.

비록 신장이 175 이상인데도 멀리뛰기 기록이 175 정도인 사람도 있고 신장이 160 정도인데도 멀리뛰기 기록이 200보다 멀리 뛰는 사람도 있지만, 그럼에도 불구하고 전반적으로 나타나는 경향을 표현해 본다면 아래의 그래프처럼 직선으로 나타낼 수 있을 것이다. 이 직선을 찾아내는 과정을 통계학에서는 회귀식 찾기라고 한다.

[그림 6-6]
신장과
제자리멀리뛰기
산점도 (2)



이번에는 위의 자료를 인위적으로 변형하여 다음과 같은 자료를 만들어 보았다.

<표 6-4> 신장과 제자리 멀리뛰기 자료 2

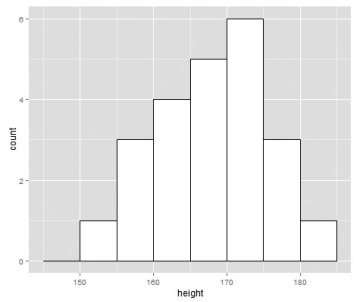
ID	신장(㎝)	제자리 멀리뛰기(㎝)	ID	신장(㎝)	제자리 멀리뛰기(㎝)
1	161.2	210	13	161.5	209
2	157.4	230	14	166.0	202
3	173.0	160	15	177.3	130
4	173.0	155	16	173.7	150
5	152.9	240	17	169.1	170
6	176.5	135	18	170.2	163
7	167.5	197	19	157.0	230
8	175.4	150	20	160.8	219
9	173.2	153	21	161.0	210
10	168.8	194	22	174.3	150
11	156.5	240	23	180.0	99
12	166.1	197			

이 자료로부터 체중과 제자리멀리뛰기 변수에 대하여 일변량 분석을 한 결과와 그래프는 <표 6-1>의 자료로 얻은 결과들과 동일하다.

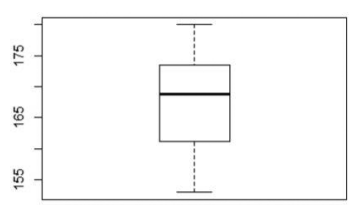
<표 6-5> 신장과 제자리멀리뛰기 일변량 분석 결과

	신장	제자리멀리뛰기
평균	167.5	182.3
표준편차	7.68	38.98
최솟값	152.9	99.0
제1사분위수(Q_1)	161.1	151.5
중앙값	168.8	194.0
제1사분위수(Q_3)	173.5	210.0
최댓값	180.0	240.0
사분위범위(IQR)	12.4	58.5

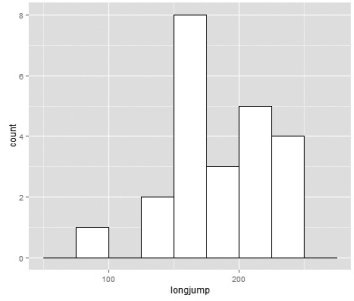
[그림 6-7]
신장 히스토그램 (2)



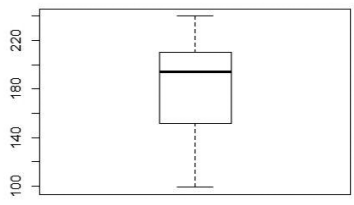
[그림 6-8]
신장 히스토그램 (2)



[그림 6-9]
제자리멀리뛰기
히스토그램 (2)

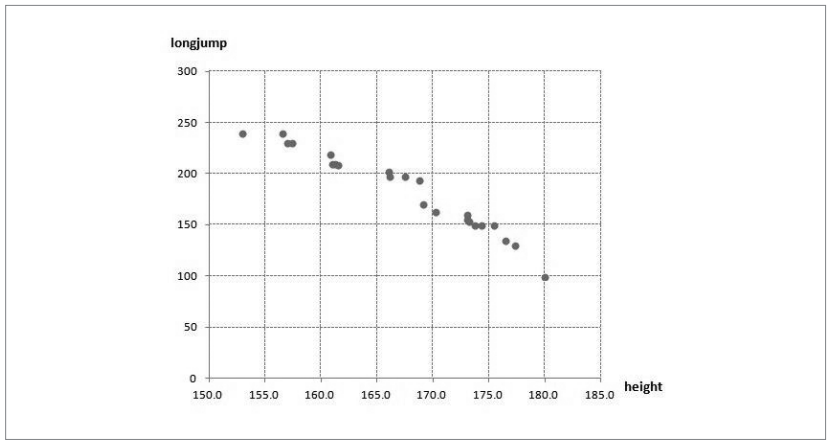


[그림 6-10]
제자리멀리뛰기
상자그림 (2)



그런데 이 자료로 산점도를 그려보면 아래와 같은 모습이 나타난다.

[그림 6-11]
신장과
제자리멀리뛰기
산점도 (3)



이는 앞의 자료로부터 얻었던 산점도와는 반대의 경향이 나타나서 신장이 클수록 오히려 제자리멀리뛰기의 기록은 낮아지는 것을 볼 수 있다. 또 그 경향은 앞의 자료에 비해서 굉장히 강하게 나타나서 신장이 비슷한 사람들은 거의 비슷한 제자리멀리뛰기 기록을 보이고 있다.

이와 같이 두 연속형 변수로 이루어진 자료를 일변량 자료로만 분석해서는 변수 간에 내재된 중요한 특징(정보)을 놓칠 수 있어서 산점도와 같이 변수 간의 관계를 주목하는 기법을 활용하는 것이 필요하다.

6-2. 상관계수 알아보기

학습목표

- **산점도를 통해서 나타난 두 변수 간의 관계의 강도를 수치화시킬 수 있는 방법으로 상관계수를 소개하고 계산과정을 학습하여 상관계수를 왜곡 없이 해석한다.**

1 선형관계의 정도

이 절에서는 하나의 관측 대상에 대하여 두 가지 양적 변수를 측정하여 얻은 이변량자료(bivariate data)를 취급한다. 이들 자료로부터 변수들간에 관련성의 정도(관계의 강도)를 계량화시키려고 한다. 다시 말하면 이들 변수간에 관련성이 존재한다면 그 관련성의 정도를 어떻게 효과적으로 표현할 수 있는가? 그리고 한 변수의 값으로 다른 변수의 값을 예측하기 위하여 어떤 방법을 써야 할 것인가? 이에 관한 해답을 두 변수의 관계 중 가장 간단한 관계인 선형 관계를 통하여 찾아보기로 하자. 선형상관계수는 이와 같이 두 변수의 선형 관계가 얼마나 강한지를 보여주는 대표적인 척도이다.

2 선형 상관계수

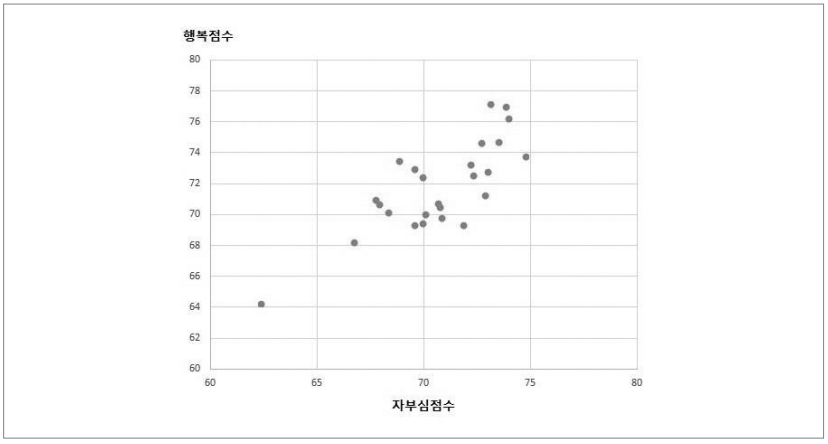
서울시민의 자부심점수와 행복점수로부터 이야기를 시작해 보자. 다음의 자료는 2014년 서울서베이에서 가구주에게 서울시민으로서의 자부심과 지금 얼마나 행복한가를 100점 만점의 점수로 조사한 결과를 25개 행정구역(구)별로 평균을 구하여 나타낸 표이다.

<표 6-6>
자부심점수와
행복점수

행정구역	자부심점수	행복점수	행정구역	자부심점수	행복점수
종로구	72.8	71.3	양천구	73.0	72.8
중구	71.8	69.3	강서구	67.9	70.7
용산구	70.6	70.8	구로구	68.8	73.5
성동구	69.9	72.4	금천구	62.3	64.3
광진구	70.7	70.5	영등포	72.3	72.5
동대문구	69.9	69.5	동작구	72.2	73.2
중랑구	74.7	73.8	관악구	66.7	68.2
성북구	67.7	71.0	서초구	73.9	76.3
강북구	70.1	70.0	강남구	73.8	77.0
도봉구	70.8	69.8	송파구	69.5	73.0
노원구	68.3	70.2	강동구	73.1	77.1
은평구	69.5	69.4			
서대문구	72.7	74.6	전체평균	70.7	71.8
마포구	73.5	74.7	표준편차	2.78	2.93

앞 절에서는 이렇게 두 연속형 변수가 함께 있는 이변량(다변량) 자료의 경우에 산점도를 그려서 각 변수별 특징 외에 두 변수의 관계나 경향성을 살펴보았다. 주어진 자료의 산점도를 그려보면 다음과 같다.

[그림 6-12]
자부심점수와
행복점수 산점도 (1)



전반적으로 자부심점수가 높은 구가 행복점수도 높은 경향이 있음을 볼 수 있다. 물론 같은 자부심점수를 가지더라도 행복점수는 다양할 수 있지만 이렇게 전반적으로 한 변수가 증가할 때 다른 변수도 함께 증가하는 경

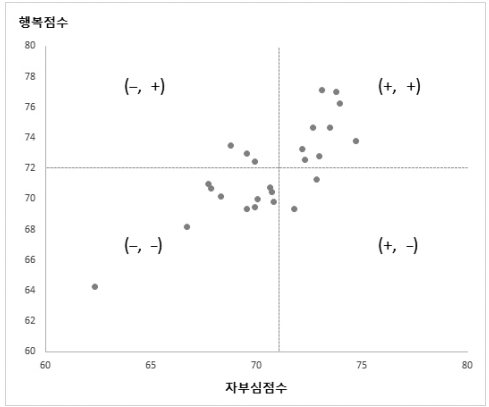
향을 양의 상관관계 혹은 정적 상관관계라고 부른다. 반대로 한 변수가 증가할 때 다른 변수가 감소하는 관계를 음의 상관관계 혹은 부적 상관관계라고 부른다. 앞 절에서 보았던 인위적 자료가 바로 음의 상관관계이다.

[그림 6-13]
두 변수간의 관련성 유형

산점도를 통하여 자부심점수와 행복점수가 전반적으로 선형의 관계 특히 양의 상관관계를 보이면서 흩어져 있는 관계의 형태를 파악하였다면 이번에는 그 관련성의 정도가 어느 정도 강한지 측정할 수 있는 척도를 생각해보자.

자부심점수를 x, 행복점수를 y라고 하고 자부심점수의 평균을 $\overline{x}$, 행복점수의 평균을 $\overline{y}$라 하자. 산점도에 아래와 같이 $\overline{x}$ 와 $\overline{y}$ 의 선을 그으면 전체 공간이 4개의 공간으로 나누어진다.

[그림 6-14]
자부심점수와 행복점수 산점도 (2)



1사분면에 있는 자료들은 자부심점수의 값과 행복점수의 값이 둘 다 평균보다 크고, 2사분면에 있는 자료들은 자부심점수는 평균보다 작고 행복점수는 평균보다 크다. 3사분면에 있는 자료들은 자부심점수와 행복점수의

값이 둘 다 평균보다 작고, 4사분면에 있는 자료들은 자부심점수는 평균보다 크고 행복점수는 평균보다 작다.

이를 정리해보면 1사분면과 3사분면에 있는 자료들은 둘 다 평균보다 크거나 평균보다 작아서 두 변수의 움직임이 같은 자료들이고, 2사분면과 4사분면에 있는 자료들은 한 변수가 평균보다 크면 다른 변수는 평균보다 작아서 두 변수의 움직임이 반대인 자료들이다. 따라서 두 변수가 양의 관계를 갖는다면 자부심점수와 행복점수와 같이 자료들은 주로 1사분면과 3사분면에 있을 것이고, 두 변수가 음의 관계를 갖는다면 자료들은 주로 2사분면과 4사분면에 있을 것이다.

이제는 각 변수에 대한 편차 곱인 $(x_i-\overline{x})(y_i-\overline{y})$을 생각해 보자. 1사분면과 3사분면에 있는 자료들은 두 변수의 편차의 부호가 같으므로 편차의 곱은 양(+)이 될 것이고, 2사분면과 4사분면에 있는 자료들은 두 변수의 편차의 부호가 다르므로 편차의 곱은 음(-)이 될 것이다.

따라서 두 변수가 양의 상관관계를 갖는다면 자료가 주로 1사분면과 3사분면에 많으므로 편차들의 곱 또한 양의 값이 많아지게 되고, 편차곱의 평균도 양의 값을 가질 것이다. 반면에 두 변수사이에 음의 상관관계가 존재한다면 주로 자료가 2사분면과 4사분면에 많으므로 편차들의 곱이 음인 것이 많아 편차곱의 평균도 음의 값을 가질 것이다. 이렇게 구한 값을 공분산(covariance)이라고 하고 다음과 같이 계산된다.

$$\text{표본 공분산} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

여기서 분모를 $(n-1)$로 한 것에 관심이 있는 독자들은 부록을 참고하기 바란다.

주어진 자료로부터 자부심점수(x)와 행복점수(y)의 공분산을 구해보면 다음과 같다.

$$\text{공분산} = \frac{(72.8-70.7)(71.3-71.8)+\cdots+(73.1-70.7)(77.1-71.8)}{25-1} = 6.17$$

이제 이렇게 구한 공분산이 두 변수의 관련성을 잘 측정할 수 있으리라 생각되지만 이 값은 편차의 문제, 측정 단위의 문제 등으로 불완전한 척도이다. (이유는 부록에 기술)

보다 완전한 척도를 생성하기 위해서는 공분산을 각 변수의 표준편차로 나누어주어 단위에 따른 변화나 단순히 편차가 커서 공분산이 커지는 문제점을 해결한 것이 상관계수(correlation coefficient)이다. 상관계수 r이라고 하며 다음과 같이 계산된다.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

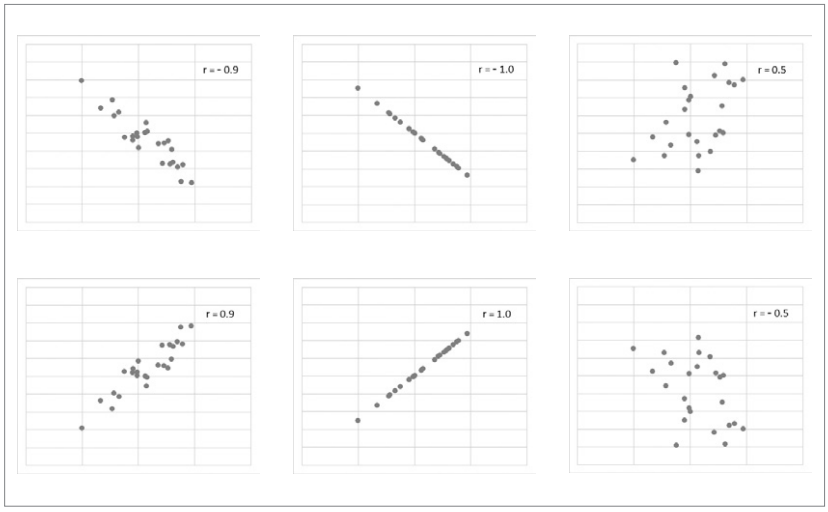
주어진 자부심점수와 행복점수의 상관계수를 구해보자.

$$r = \frac{6.17}{2.78 \times 2.93} = 0.79$$

자부심점수와 행복점수 사이의 상관계수 값 0.79의 의미를 올바르게 해석하기 위해서는 상관계수의 기본적인 성질을 알아야 할 필요가 있다. 중요한 성질들 몇 가지를 함께 생각해보자.

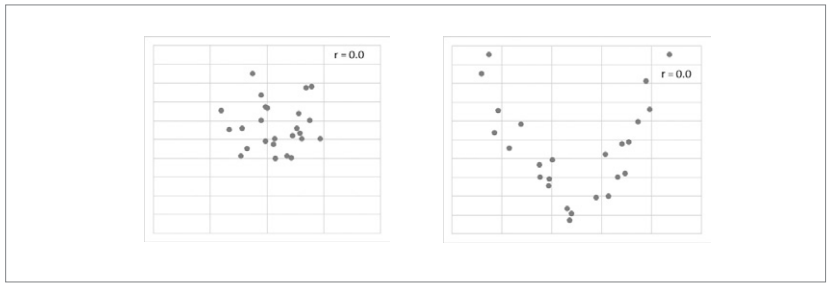
상관계수 r은 -1에서 +1사이의 값을 가진다. $r = 1$일 때를 완전 양의 상관관계라 하고 $r = -1$일 때를 완전 음의 상관관계라고 하며 이 때 관측값들은 모두 직선상에 놓이게 된다. 아래의 그림처럼 두 변수 사이의 관계가 강할수록 $|r|$의 값은 1에 가까워지고, 두 변수 사이의 관계가 약할수록 $|r|$의 값은 0에 가까워진다. 즉 상관계수는 두 변수 간의 관련성이 얼마나 큰가를 나타내주는 값으로 $|r|$이 1에 가까울수록 강한 관련성을 가진다고 할 수 있다.

[그림 6-15]
상관계수 크기에 따른 산점도 (1)



그러나 $|r|$이 0에 가깝다는 것이 곧 두 변수 사이의 관계가 없다는 의미인 것은 아니다. 여기에는 두 가지 경우가 있다. 하나는 정말 두 변수 사이의 관련성이 없거나 혹은 두 변수의 관계가 선형적인 관계가 아닌 경우이다. 예를 들어 다음의 그래프를 살펴보면 두 변수 간에 분명한 곡선의 관련성을 가지고 있더라도 상관계수 값은 0으로 나타날 수 있다. 상관계수는 선형의 관계가 얼마나 강한지를 알 수 있게 해 주는 값이므로 선형의 관계가 없으면 상관계수는 0이 되지만 이것을 곧바로 두 변수 간에 관련성이 없다고 해석하는 것은 위험하다. 따라서 상관계수를 확인하기 전에 산점도를 통하여 두 변수 간의 관계를 먼저 검토해보는 것이 필요하다. 만일 두 변수 간에 곡선의 관계가 존재한다면 변수변환 등을 통해 선형의 관계를 찾아낸 후 상관계수를 구할 수 있다.

[그림 6-16]
상관계수 크기에 따른 산점도 (2)



이제 자부심점수와 행복점수의 관계를 산점도와 상관계수를 함께 활용하여 표현해보자. 자부심점수와 행복점수는 자부심점수가 커질수록 행복점수도 커지는 양의 상관관계를 보이며, 둘 사이의 상관계수는 0.79로 자부심점수와 행복점수는 강한 선형의 관계를 가진다고 해석할 수 있을 것이다.

6-3. 회귀선 이해하기 (심화)

학습목표

- **2절에서는 두 변수 간의 관계를 산점도를 통해서 나타난 관련성의 정도를 수치화시킬 수 있는 방법으로 상관계수를 학습했는데 이 절에서는 경향성 특징을 찾아내는 방법으로 회귀선을 소개하여 회귀분석기법의 기본 개념을 습득하게 하고 보고서 등을 더 깊이 이해한다.**

1 경향성 주목하기

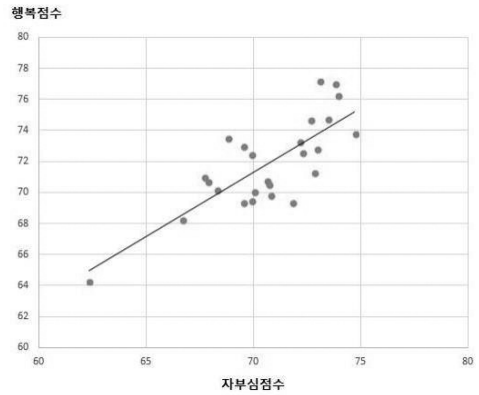
먼저 선형관계를 수식으로 표현하는 예를 생각해 보자. 1절에서 다루었던 신장과 제자리멀리뛰기의 자료로부터 신장과 제자리멀리뛰기의 경향성의 특징을 찾아보자. 산점도를 통해서 전체적으로 신장이 증가하면 멀리뛰기 기록도 증가하는 경향을 발견했는데, 이번에는 보다 구체적으로 신장이 1cm 더 큰 사람의 제자리멀리뛰기 기록은 얼마나 증가하는가라는 질문을 던져보자. 또 2절에서 다루었던 2014년 서울서베이의 자부심점수와 행복점수 자료로부터 자부심점수가 10점 올라가면 행복점수는 어느 정도 높아지는가의 문제도 생각해 볼 수 있을 것이다.

이 때 사용할 수 있는 가장 대표적인 방법은 산점도에서 보았던 직선을 이용하는 것이다. 우리는 앞 절을 통하여 산점도와 상관계수를 계산해 봄으로써 두 변수 간에 강한 선형관계가 존재한다면 관측점들이 어떤 직선에 가까이 분포하게 되는 것을 보았다. 이 직선을 이용하여 한 변수의 값으로 다른 변수의 값을 예측할 수 있을 것이다. 이렇게 두 변수의 선형 관계를 잘 나타내 줄 수 있는 직선을 회귀직선이라 한다. 직선뿐만 아니라 여러 형태의 회귀선의 방정식을 주어진 자료로부터 구하고 그것을 이용하여 예측하는 통계적 방법을 회귀분석(regression analysis)이라고 한다. 이러한 기법은 가계동향조사의 소득과 지출의 관계 등에서 활용된다.

2 경향성의 수치화

2절에서 다루었던 자부심점수(x)와 행복점수(y)로 회귀식을 직접 구해 보자.

[그림 6-17]
자부심점수와
행복점수 산점도 (3)



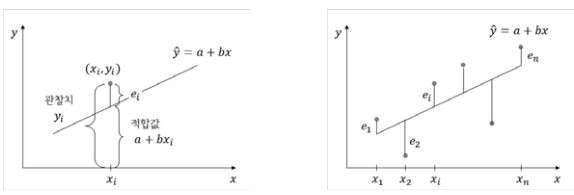
관측값들이 직선상에 있는 것이 아니므로 우리는 이 관측값들의 선형적인 경향을 가장 잘 반영하는 직선을 찾아야 한다. 자료에 "가장 잘 들어맞는 직선"을 찾는 방법은 A. M. Legendre(1752－1833)에 의해 개발된 최소제곱법(method of least squares)이라는 방법을 이용하는 것이 가장 일반적이다. 최소제곱법의 개념을 잠시 살펴보자.

n개의 관측점 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$에 가장 잘 맞는 직선을 $\hat{y} = a + bx$ 라고 하자. 이를 적합방정식이라 부른다. 말하자면 $\hat{y_i} = a + bx_i$는 독립변수 x의 값이 x_i 일 때에 종속변수 y 의 적합값(예측값)을 나타내고, y_i는 그에 대응하는 y의 관측값을 나타낸다. $x_1, \cdots, x_n$에 대응하는 적합방정식을 나타내는 직선 위의 점들은 다음과 같다.

$$(x_1, \hat{y_1}), \cdots, (x_n, \hat{y_n}) \Rightarrow (x_1, a+bx_1), \cdots, (x_n, a+bx_n)$$

따라서 $y_1 - \hat{y_1}, \cdots, y_n - \hat{y_n}$ 즉, $y_1 - (a+bx_1), \cdots, y_n - (a+bx_n)$ 은 관측값과 그에 대응하는 적합값과의 편차이고 이를 $e_1, \cdots, e_n$ 으로 표시한다. 그림에서와 같이 $e_i = y_i - \hat{y_i} = y_i - (a+bx_i)$는 관측값 y_i 와 적합방정식 직선과의 수직 거리를 표시한다.

[그림 6-18]
y 의 적합값 $a+bx_i$
와 관측값 y_i간의 편차



최소제곱법에서 말하는 "가장 잘 적합되는" 방정식은 적합값과 관측값 간의 편차를 최소로 만드는 방정식이다. 그러나 모든 편차의 합은 항상 0이 되므로 편차의 제곱의 합을 최소로 하는 적합방정식을 사용하는데 이것이 최소제곱법의 원리이다.

관측값 y_i와 적합값 y_i와의 편차제곱합이 최소인 직선을 구하여야 하므로 편차의 제곱합인 $\sum_{i=1}^{n}\{y_i-(a+bx_i)\}^2$ 을 최소화하는 상수 a와 b를 구하면 된다. 편미분을 이용하면 직선의 절편인 a와 기울기인 b는 다음과 같이 구해진다.

$$\text{회귀선의 기울기 } b = r \times \frac{y\text{의 표준편차}}{x\text{의 표준편차}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\text{회귀선의 절편 } a = \bar{y} - b\bar{x}$$

❸ 회귀식 해석하기

1. 서울서베이 자부심점수와 행복점수

이제 자부심점수와 행복점수를 이용하여 회귀방정식을 구해보자. 자부심점수와 행복점수의 평균과 표준편차, 상관계수는 다음과 같다.

<표 6-7> 자부심점수와 행복점수 평균, 표준편차, 상관계수

	자부심점수	행복점수
평균	70.7	71.8
표준편차	2.78	2.93
상관계수	0.79	

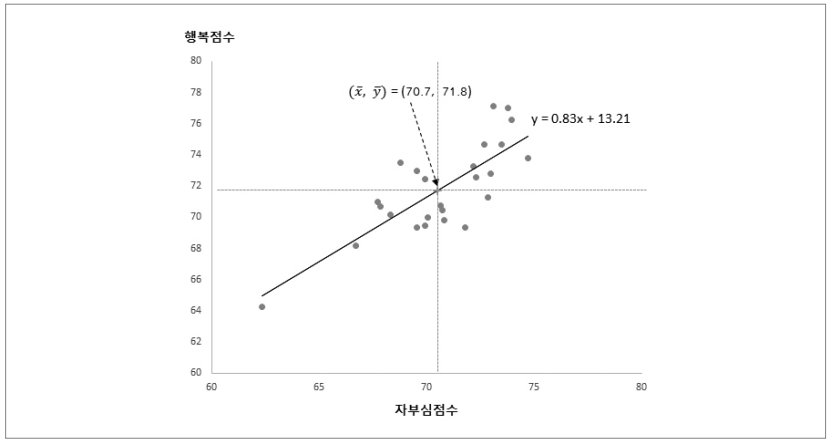
자부심점수와 행복점수 회귀식의 절편 a와 기울기 b를 구하면 다음과 같다.

$$b = 0.79 \times \frac{2.93}{2.78} = 0.83$$

$$a = 71.8 - 0.83 \times 70.7 = 13.21$$

이를 통하여 얻어진 회귀식은 $\hat{y} = 13.21 + 0.83x$이다. 기울기가 0.83이라는 것은 자부심점수가 1점 높아지면 행복점수가 0.83점씩 높아진다는 것을 말한다. 지금 구한 회귀직선을 그래프에 그려보면 다음과 같으며 이 회귀직선은 항상 점 $(\overline{x}, \overline{y})$를 통과하게 된다.

[그림 6-19]
자부심점수와 행복점수 산점도 (4)



2. 문화체육관광부 체력실태조사 활용(신장과 제자리멀리뛰기)

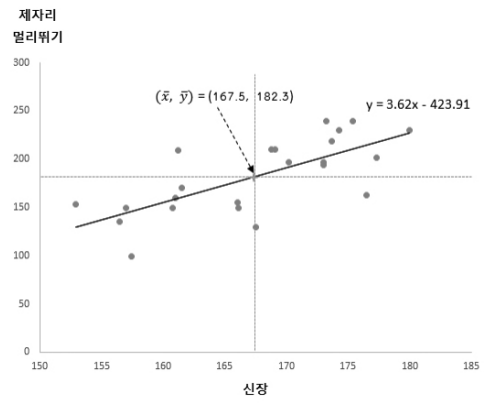
1절에서 다루었던 신장과 제자리멀리뛰기를 이용하여 회귀방정식을 구해보자. 신장과 제자리멀리뛰기의 평균과 표준편차, 상관계수는 다음과 같다.

<표 6-8>
신장과 제자리멀리뛰기 평균, 표준편차, 상관계수

	신장	제자리멀리뛰기
평균	167.5	182.3
표준편차	7.7	39.0
상관계수	0.71	

신장과 제자리멀리뛰기의 회귀식은 $\hat{y} = -423.91 + 3.62x$이다. 기울기가 3.62라는 것은 신장이 1cm 커지면 제자리멀리뛰기 기록이 3.62cm만큼 커진다는 것을 말한다. 이 회귀직선을 그래프에 그려보면 다음과 같다.

[그림 6-20]
신장과
제자리멀리뛰기
산점도



3. 가계동향조사 소득과 지출

다음은 가계동향조사의 2015년 5월 소득과 지출에 관한 자료의 일부이다. 소득과 가계지출에 관하여 회귀식을 구해보자.

<표 6-9>
소득과 가계지출
(단위: 천원)

가구번호	소득	가계지출	가구번호	소득	가계지출
1	2,450	1,257	16	7,257	3,762
2	3,418	1,971	17	1,111	3,115
3	3,920	1,494	18	1,331	965
4	603	492	19	2,143	2,505
5	1,000	2,117	20	3,700	3,251
6	524	395	21	2,351	1,954
7	544	606	22	3,582	2,915
8	6,240	5,916	23	585	397
9	1,516	1,888	24	3,260	2,753
10	614	529	25	1,285	1,508
11	2,680	1,115	26	1,146	318
12	1,917	851	27	2,003	1,941
13	445	1,889	28	6,503	6,256
14	5,000	1,395	29	2,850	2,203
15	643	468	30	1,782	2,163

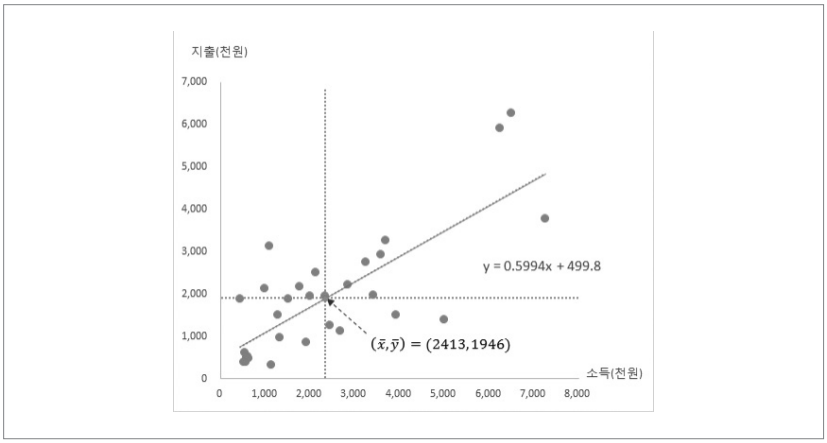
소득과 가계지출의 평균과 표준편차, 상관계수는 다음과 같다.

<표 6-10>
소득과 지출 평균, 표준편차, 상관계수

	소득	가계지출
평균	2,413	1,946
표준편차	1,877	1,466
상관계수	0.77	

이를 이용하여 회귀식을 구하면 $\hat{y}=499.8+0.60x$이다. 여기서 기울기 0.60은 소득이 천원 올라가면 지출이 600원 만큼 올라간다는 것이다. 이 회귀직선을 그래프로 나타내면 다음과 같다.

[그림 6-21]
소득과 지출 산점도



참고자료

- 김주한 · 김홍기 · 박래현 · 박석윤 · 배종호 · 이낙영 · 이석훈 · 이민구 · 이주호(2009), 통계학 입문, 정익사.
- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- 문화체육관광부(2013), 체력실태조사.
- 서울특별시(2014), 서울서베이.
- 통계청(2015), 가계동향조사.

제 7 장

범주형 자료에서 관계 찾기

7-1. 교차표 활용하기

학습목표

- 6장에서는 연속형 자료의 관계의 강도와 관계의 특징 관점에서 토의하였다. 이어서 이 장에서는 범주형 자료에 내포되어 있는 관계를 나타내는 교차표에 대해서 학습한다.

1 관계의 의미

다음과 같은 민감한 주제에 대한 설문조사로부터 토의를 시작해 보자.

주제 1) 종교유무와 안락사 지지 간에는 관계가 있다고 보십니까?

① 관계있다
② 관계없다
③ 모르겠다

주제 2) 남녀 간에 EQ(감성지수)는 관계가 있을까요?

① 관계있다
② 관계없다
③ 모르겠다

주제 3) 목둘레와 허리둘레는 서로 관계가 있나요?

① 관계있다
② 관계없다
③ 모르겠다

위의 3개의 주제에서 "서로 관계가 있다"는 말의 의미는 무엇인가 살펴보자.

주제 1) 종교가 있는 사람의 안락사 지지율과 종교가 없는 사람의 지지율에 차이가 있다.

주제 2) 남자의 EQ가 여자의 EQ보다 높은가? 남자의 평균 EQ보다 여자의 평균 EQ가 높은가?

주제 3) 목둘레가 크면 허리둘레도 크다. 혹은 목둘레는 허리둘레의 반쯤 된다.

2 관심 변수의 특성

3개의 주제에 포함된 변수의 특징을 살펴보자.

주제 1) 종교유무(질적변수) / 안락사지지여부(질적변수)

주제 2) 성별(질적변수) / EQ(양적변수)

주제 3) 목둘레(양적변수) / 허리둘레(양적변수)

분할표는 변수의 종류가 주제 1)과 같이 질적 변수(종교유무)와 질적 변수(안락사지지여부)일 때 두 변수 간의 관계를 표현하는 도구로 아래와 같이 표현할 수 있다.

<표 7-1> 종교와 안락사 지지여부 교차표 (1)

	안락사 지지	안락사 반대	합계
종교 유	26	4	30
종교 무	11	9	20
합계	37	13	50

이를 통하여 우리가 알 수 있는 것들은 안락사를 지지하는 사람은 전체 50명 중 37명이며 반대하는 사람은 13명으로 지지하는 사람이 더 많다는 것이다. 또 종교가 있는 30명의 사람들 중 안락사를 지지하는 사람은 26명, 종교가 없는 20명의 사람들 중 안락사를 지지하는 사람은 11명이다. 주어진 표는 총 사람 수가 50명이지만 인원이 늘어나는 경우에는 숫자만으로 파악하기에는 어려움이 있다. 교차표를 작성할 때에는 아래와 같이 백분율을 함께 표현하여 주면 좋다.

<표 7-2>
종교와 안락사
지지여부 교차표 (2)

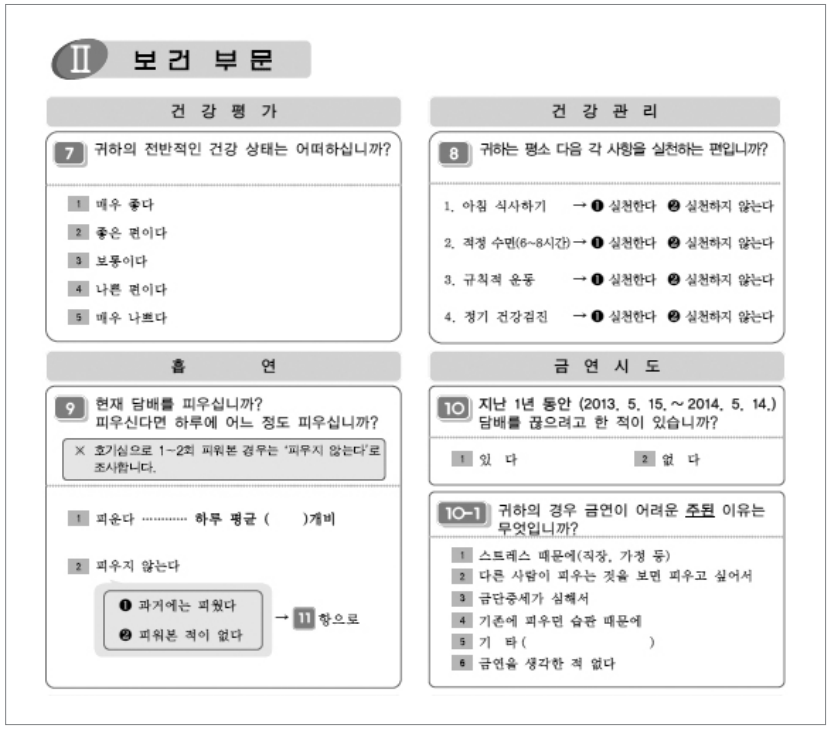
	안락사 지지	안락사 반대	합계
종교 유	26	4	30
(행백분율)	(86.7%)	(13.3%)	(100.0%)
(열백분율)	(70.3%)	(30.8%)	(60.0%)
(전체백분율)	(52.0%)	(8.0%)	(60.0%)
종교 무	11	9	20
(행백분율)	(55.0%)	(45.0%)	(100.0%)
(열백분율)	(29.7%)	(69.2%)	(40.0%)
(전체백분율)	(22.0%)	(18.0%)	(40.0%)
합계	37	13	50
(행백분율)	(74.0%)	(26.0%)	
(열백분율)	(100.0%)	(100.0%)	
(전체백분율)	(74.0%)	(26.0%)	

행백분율은 행을 기준으로 계산된 것을 말한다. 즉 종교가 있는 사람 30명 중 26명인 86.7%가 안락사를 지지한다는 것이며, 종교가 없는 사람 20명 중에서는 11명인 55.0%가 안락사를 지지한다는 것이다. 전체 50명 중에서는 74.0%가 안락사를 지지한다. 열백분율은 열을 기준으로 계산된다. 전체 50명 중 60.0%는 종교가 있는 사람들이다. 안락사를 지지하는 37명 중 26명인 70.3%는 종교가 있고 안락사를 반대하는 사람들 중에서는 30.8%가 종교를 가지고 있다. 전체백분율은 전체 50명에 대한 비율로 종교가 있으면서 안락사를 반대하는 사람은 전체의 8.0%라는 것을 알 수 있다.

❸ 사례

다음은 통계청에서 실시한 2014년 사회조사 중 보건부문에 대한 조사표의 일부와 사회조사에 표본으로 추출된 응답자 중 대전에 거주하는 가구에 속한 사람들의 건강평가와 건강관리 중 아침식사 실천여부, 적정수면 실천여부, 규칙적 운동 실천여부에 대한 응답결과를 나타낸 교차표이다.

[그림 7-1]
사회조사 조사표
(일부)



Ⅱ 보 건 부 문

건 강 평 가

7 귀하의 전반적인 건강 상태는 어떠하십니까?

1 매우 좋다
2 좋은 편이다
3 보통이다
4 나쁜 편이다
5 매우 나쁘다

건 강 관 리

8 귀하는 평소 다음 각 사항을 실천하는 편입니까?

1. 아침 식사하기 → ❶ 실천한다 ❷ 실천하지 않는다
2. 적정 수면(6~8시간) → ❶ 실천한다 ❷ 실천하지 않는다
3. 규칙적 운동 → ❶ 실천한다 ❷ 실천하지 않는다
4. 정기 건강검진 → ❶ 실천한다 ❷ 실천하지 않는다

흡 연

9 현재 담배를 피우십니까?
피우신다면 하루에 어느 정도 피우십니까?

× 호기심으로 1~2회 피워본 경우는 '피우지 않는다'로 조사합니다.

1 피운다 ····· 하루 평균 ( )개비

2 피우지 않는다

❶ 과거에는 피웠다
❷ 피워본 적이 없다

→ 11 항으로

금 연 시 도

10 지난 1년 동안 (2013. 5. 15. ~ 2014. 5. 14.) 담배를 끊으려고 한 적이 있습니까?

1 있 다 2 없 다

10-1 귀하의 경우 금연이 어려운 주된 이유는 무엇입니까?

1 스트레스 때문에(직장, 가정 등)
2 다른 사람이 피우는 것을 보면 피우고 싶어서
3 금단증세가 심해서
4 기존에 피우던 습관 때문에
5 기 타 ( )
6 금연을 생각한 적 없다

<표 7-3>
아침식사와 건강상태 교차표

	건강상태					합계
	매우좋다	좋은편이다	보통이다	나쁜편이다	매우나쁘다	합계
아침식사 실천	208	505	575	115	39	1,442
(행백분율)	(14.4)	(35.0)	(39.9)	(8.0)	(2.7)	(100.0)
(열백분율)	(77.6)	(73.0)	(75.0)	(62.5)	(62.9)	(73.1)
(전체백분율)	(10.5)	(25.6)	(29.1)	(5.8)	(2.0)	(73.1)
아침식사 미실천	60	187	192	69	23	531
(행백분율)	(11.3)	(35.2)	(36.2)	(13.0)	(4.3)	(100.0)
(열백분율)	(22.4)	(27.0)	(25.0)	(37.5)	(37.1)	(26.9)
(전체백분율)	(3.0)	(9.5)	(9.7)	(3.5)	(1.2)	(26.9)
합계	268	692	767	184	62	1,973
(행백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	
(열백분율)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	
(전체백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	

전체 1,973명 중 아침식사를 하는 사람의 비율이 73.1%로 더 높으며, 건강에 대한 평가는 좋다(매우 좋다와 좋은 편이다 포함)고 평가한 사람의 비율이 48.7%로 나타났다. 한편 아침식사를 하는 사람들 중에서 49.4%가, 아침식사를 하지 않는 사람 중에서는 46.5%로 양측 모두 비슷한 비율로 자신의 건강상태를 좋다고 평가하고 있다. (여기서 비슷하다는 말에 대하여 어떻게 생각하는가? 고민해보자.)

<표 7-4>
적정수면과 건강상태 교차표

	건강상태					합계
	매우좋다	좋은편이다	보통이다	나쁜편이다	매우나쁘다	합계
적정수면 실천	218	587	594	127	37	1,563
(행백분율)	(13.9)	(37.6)	(38.0)	(8.1)	(2.4)	(100.0)
(열백분율)	(81.3)	(84.8)	(77.4)	(69.0)	(59.7)	(79.2)
(전체백분율)	(11.0)	(29.8)	(30.1)	(6.4)	(1.9)	(79.2)
적정수면 미실천	50	105	173	57	25	410
(행백분율)	(12.2)	(25.6)	(42.2)	(13.9)	(6.1)	(100.0)
(열백분율)	(18.7)	(15.2)	(22.6)	(31.0)	(40.3)	(20.8)
(전체백분율)	(2.5)	(5.3)	(8.8)	(2.9)	(1.3)	(20.8)
합계	268	692	767	184	62	1,973
(행백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	
(열백분율)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	
(전체백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	

적정수면 실천여부에 관하여는 적정수면을 취하는 사람이 79.2%이고 적정수면을 취하지 못하는 사람이 20.8%이다. 건강상태를 좋다고 평가한 사람은 적정수면을 취하는 사람 중에서는 51.5%이고 적정수면을 취하지 못하는 사람 중에서는 37.8%로 13.7%포인트 차이가 난다. 그런데 여기서 고민할 문제가 하나 있다. 이 차이를 큰 차이라고 할 수 있을지는 주관적인 문제가 된다. 이 문제는 시간이 되면 함께 토의하자.

<표 7-5>
규칙적 운동과
건강상태 교차표

	건강상태					합계
	매우좋다	좋은편이다	보통이다	나쁜편이다	매우나쁘다	합계
규칙적운동 실천	134	274	300	48	15	771
(행백분율)	(17.4)	(35.5)	(38.9)	(6.2)	(1.9)	(100.0)
(열백분율)	(50.0)	(39.6)	(39.1)	(26.1)	(24.2)	(39.1)
(전체백분율)	(6.8)	(13.9)	(15.2)	(2.4)	(0.8)	(39.1)
규칙적운동 미실천	134	418	467	136	47	1,202
(행백분율)	(11.1)	(34.8)	(38.9)	(11.3)	(3.9)	(100.0)
(열백분율)	(50.0)	(60.4)	(60.9)	(73.9)	(75.8)	(60.9)
(전체백분율)	(6.8)	(21.2)	(23.7)	(6.9)	(2.4)	(60.9)
합계	268	692	767	184	62	1,973
(행백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	
(열백분율)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	
(전체백분율)	(13.6)	(35.1)	(38.9)	(9.3)	(3.1)	

규칙적 운동 실천여부에 관하여는 규칙적 운동을 실천하는 사람이 39.1%, 실천하지 못하는 사람이 60.9%이다.

이번에는 동일한 자료로 아침식사 여부와 적정수면 여부와의 관계 그리고 아침식사 여부와 규칙적 운동 실천 여부의 관계를 살펴보자.

<표 7-6>
실습사례 (1)

	적정수면 실천	적정수면 미실천	총합계
아침식사 실천	1,204	238	1,442
(행백분율)	(83.5)	(16.5)	(100.0)
(열백분율)	(77.0)	(58.0)	(73.1)
(전체백분율)	(61.0)	(12.1)	(73.1)
아침식사 미실천	359	172	531
(행백분율)	(67.6)	(32.4)	(100.0)
(열백분율)	(23.0)	(42.0)	(26.9)
(전체백분율)	(18.2)	(8.7)	(26.9)
합계	1,563	410	1,973
(행백분율)	(79.2)	(20.8)	
(열백분율)	(100.0)	(100.0)	
(전체백분율)	(79.2)	(20.8)	

<표 7-7>
실습사례 (2)

	규칙적운동 실천	규칙적운동 미실천	합계
아침식사 실천	650	792	1,442
(행백분율)	(45.1)	(54.9)	(100.0)
(열백분율)	(84.3)	(65.9)	(73.1)
(전체백분율)	(32.9)	(40.1)	(73.1)
아침식사 미실천	121	410	531
(행백분율)	(22.8)	(77.2)	(100.0)
(열백분율)	(15.7)	(34.1)	(26.9)
(전체백분율)	(6.1)	(20.8)	(26.9)
합계	771	1,202	1,973
(행백분율)	(39.1)	(60.9)	
(열백분율)	(100.0)	(100.0)	
(전체백분율)	(39.1)	(60.9)	

7-2. 교락변수

학습목표

- **예를 통해서 교락변수의 경향에 대하여 연수생이 느낄 수 있도록 하여 교차표 해석시 주의를 환기시켜서 현업에 적용한다.**

1 DIET 프로그램과 체중감량 관계분석

기본 DIET 프로그램에 DIET A를 추가 실시하는 것이 체중감량의 성패에 영향을 주는가를 살펴보기 위하여, 100명의 근무자들에게서 다음과 같은 자료를 얻었다고 생각해보자.

<표 7-8> DIET A와 감량성패 자료 (1)

DIET A	감량성패	도수
실시	성공	18
실시	실패	12
미실시	성공	6
미실시	실패	4
실시	성공	2
실시	실패	18
미실시	성공	4
미실시	실패	36

DIET A와 체중감량 성패에 관한 교차표를 작성하면 다음과 같다.

<표 7-9> DIET A와 감량성패 교차표

	감량성공	감량실패	합계
DIET A 미실시	10 (20.0)	40 (80.0)	50 (100.0)
DIET A 실시	20 (40.0)	30 (60.0)	50 (100.0)
합계	30 (30.0)	70 (70.0)	100 (100.0)

DIET A를 실시하였을 경우의 체중감량 성공률은 40.0%이고 DIET A를 실시하지 않았을 경우의 성공률은 20.0%이므로 DIET A의 실시와 체중감량의 성패는 서로 관계가 있다는 결론을 내릴 수 있다.

❷ 근무패턴 항목의 출현

그러나 만일 조사항목에 근무패턴도 포함되어 있어서 다음과 같은 자료가 있었다고 하자. 연구자가 주어진 자료에 나타나 있는 근무패턴이 체중감량에 영향을 줄 수도 있다고 생각하여 근무패턴을 포함한 교차표를 다시 작성한다면 다음의 결과를 얻을 수 있다.

<표 7-10>
DIET A와 감량성패 자료 (2)

DIET A	감량성패	근무패턴	도수
실시	성공	기립근무	18
실시	실패	기립근무	12
미실시	성공	기립근무	6
미실시	실패	기립근무	4
실시	성공	착석근무	2
실시	실패	착석근무	18
미실시	성공	착석근무	4
미실시	실패	착석근무	36

<표 7-11>
근무패턴에 따른 DIET A와 감량성패 교차표

		감량성공	감량실패	합계
기립 근무	DIET A 미실시	6 (60.0)	4 (40.0)	10 (100.0)
기립 근무	DIET A 실시	18 (60.0)	12 (40.0)	30 (100.0)
기립 근무	합계	24 (60.0)	16 (40.0)	40 (100.0)
착석 근무	DIET A 미실시	4 (10.0)	36 (90.0)	40 (100.0)
착석 근무	DIET A 실시	2 (10.0)	18 (90.0)	20 (100.0)
착석 근무	합계	6 (10.0)	54 (90.0)	60 (100.0)

표를 보면 기립근무인 경우에는 체중감량에 성공한 비율이 DIET A의 실시여부와 상관없이 60.0%이고, 착석근무인 경우에는 DIET A의 실시여부와 상관없이 10.0%가 체중감량에 성공하였음을 알 수 있다. 즉 DIET A와 상관없이 근무패턴이 체중감량 성패에 영향을 미치고 있다는 이야기가 된다.

❸ DIET A와 근무패턴과의 관계

그렇다면 첫 번째 교차표에서 얻어진 DIET A의 실시여부와 체중감량 성패와의 관계는 어떻게 나타난 것일까? 다음의 근무패턴과 DIET A에 관한 표를 보면서 그 이유를 생각해 보기로 하자.

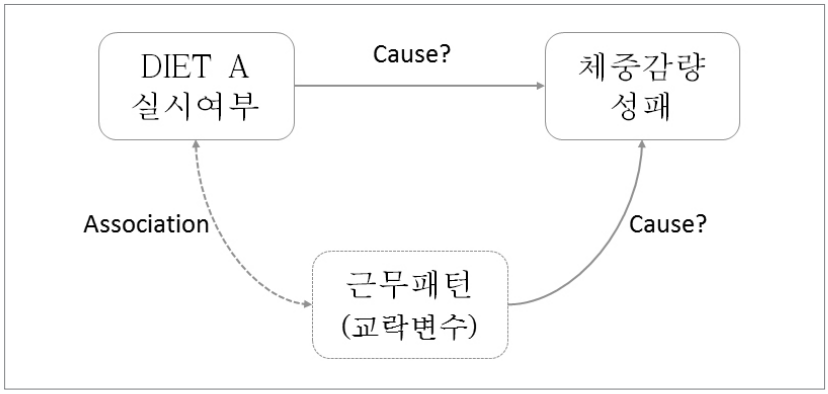
<표 7-12> DIET A와 근무패턴 교차표

	기립근무	착석근무	합계
DIET A 미실시	10 (20.0)	40 (80.0)	50 (100.0)
DIET A 실시	30 (60.0)	20 (40.0)	50 (100.0)
합계	40 (40.0)	60 (60.0)	100 (100.0)

DIET A를 실시하지 않은 사람 중에는 착석근무자가 많고, DIET A를 실시하는 사람 중에는 기립근무자가 많다는 것을 알 수 있다. 따라서 이 둘 사이의 관련성이 결과에 영향을 미치고 있는 것이다. 결국 DIET A의 실시여부와 체중감량의 성패는 체중감량의 성패에 영향을 미치는 또 다른 변수인 근무형태로 인하여 서로 관련이 없다는 결론을 내리게 된다.

이처럼 결과에 영향을 주는 관심 있는 변수의 효과를 알아보고자 할 때, 이 변수 외에 결과에 영향을 주는 다른 변수를 교락변수(confounding variable) 혹은 교락요인(confounding factor)이라고 한다. 주어진 자료에서는 근무패턴이 교락변수이다.

[그림 7-2]
교락변수



그러나 교락변수는 원래 관심을 가지고 있던 변수와 관련성이 없다면 결과에 영향을 주지 않는다. 위의 예에서 근무패턴과 DIET A의 실시여부가 서로 관련이 없다면, DIET A와 체중감량 성패의 비교에는 근무패턴이 영향을 주지 않는다.

4 국가통계 사례

다음은 2005년도 한국의 월평균 납세액 자료의 일부이다.

<표 7-13>
2005년 월평균
납세액 자료 (일부)

구분		전가구	근로자 가구	근로자 외 가구
의회의원, 고위임직원 및 관리자	조세(천원)	171	264	135
의회의원, 고위임직원 및 관리자	(가구비중)	(100)	(30)	(70)
전문가 전문가	조세(천원)	223	243	126
전문가 전문가	(가구비중)	(100)	(84)	(16)

출처: 통계청(http://nso.go.kr)

이 자료를 근거로 하여 어느 기관에서 다음과 같은 주장을 하였다.

> 통계청의 직업그룹 분류에 따르면 의회의원, 고위임직원 및 관리자는 전체 9개 직업군 가운데 최상위에 해당된다고 한다. 여기에는 입법부, 사법부, 행정부의 1급 이상 공무원, 기업체의 고위직 임직원, 국회의원, 지방의원, 구청장 그리고 부시장급 이상의 지자체 고위직 등이 속한다. 그런데 이들 그룹이 올들어 9월까지 납부한 세금은 월평균 17만 1,201원으로 전문가 그룹이 낸 월평균 세금 22만 2,827원의 76.8%에 불과하다.

위에 나타난 정보를 토의해 보자.

① 이 기관에서는 통계청 직업그룹 분류에 따른 최상위 집단과 전문가 집단의 월납세액을 비교하고 있다. 그런데 이 표를 해석하는 과정에서 '의회의원, 고위임직원 및 관리자'를 최상위 그룹으로 일컫는 것은 적절하지 않다. 당시 통계청이 해명자료를 통해 밝혔듯「최상위 계층」은 소득수준에 따라 구분할 때의 범주로서 직업분류에 따라 구분했을 때의 범주가 아닌 것이다.

② 의회의원, 고위임직원 및 관리자 그룹의 월평균 납세액은 171,000원이고, 전문가 그룹의 월평균 납세액은 223,000원이다.

③ 의회의원, 고위임직원 및 관리자 그룹의 30%는 근로자가구인 반면, 전문가 그룹의 경우 84%가 근로자이다. 즉 의회의원, 고위임직원 및 관리자 그룹의 70%가 근로자가 아닌 반면, 전문가 그룹은 대부분이 근로자가구이다.

④ 근로자 가구의 월평균 납세액만 비교하면 전체 가구에서 나타난 것과는 달리 의회의원, 고위임직원 및 관리자 그룹이 $\frac{264}{243}-1=8.9\%$ 더 많이 냈다.

⑤ 근로자가 아닌 가구에서도 마찬가지이다. 의회의원, 고위임직원 및 관리자 그룹이 $\frac{135}{126}-1=7.1\%$ 더 많이 냈다.

이 절을 통하여 배운 내용을 바탕으로 모순된 사실을 설명해 보자.

이 기관이 표로부터 얻고자 하는 주요 정보는 두 집단(의회의원, 고위임직원 등으로 이루어진 관리자 그룹과 전문가 그룹)의 비교결과이다. 그런데

두 집단을 비교하기 위해서 가장 중요한 것은, 두 집단을 구별하는 특성(여기서는 '의회의원, 고위임직원 및 관리자'와 '전문가'라는 직업)을 제외하고는 납세액과 관련될 가능성이 있는 나머지 어떤 특성에서도 두 집단이 비슷해야 한다는 점이다. 그러나 이 사례에서는 두 집단의 근로자 가구 비율이 서로 크게 다르다. 따라서 근로자 가구 여부를 고려하지 않았을 때와 근로자 가구 여부를 고려하였을 때 의회의원, 고위임직원 및 관리자 그룹과 전문가 그룹의 월평균 납세액에 대한 결과가 달라짐을 확인할 수 있다. 앞서 전문가 그룹이 세금을 더 많이 내고 있는 것처럼 보인 것은 근로자 가구 여부를 고려하지 않았기 때문이고, 이를 고려하면 의회의원, 고위임직원 및 관리자 그룹이 전문가 그룹보다 월평균 세금을 더 많이 내는 것을 알 수 있다. 즉 이 사례에서는 근로자 가구여부가 교락변수로 작용하였다.

7-3. Simpson's paradox

학습목표

- **교락변수에 대한 느낌을 Simpson's paradox를 통하여 한 번 더 확인하여 교차표 작성 시 유념한다.**

1 조사자료 분석

1. 조사표

어떤 지자체 기획실에서는 수질 환경보호를 위한 샴푸 사용에 관한 특정 XX정책에 대한 지역별 홍보 전략을 세우기 위하여 먼저 지지율 조사를 실시하였다. 다음은 조사표의 일부이다.

1. 당신의 거주 지역은?

① A ② B

2. XX정책을 지지합니까?

① 예 ② 아니오

(인구특성 문항)

1. 성별은? ① 남자 ② 여자

2. 연령대는? ① 20대 ② 30대

지역별 응답자의 통계는 다음과 같다.

<표 7-14> 거주지역 빈도표

거주지역	빈도	백분율(%)
A	400	50%
B	400	50%

2. 지역간 지지율 비교

다음은 수집된 자료를 1차 분석한 결과로서 다음과 같은 표를 얻었다.

<표 7-15>
거주지역과 지지여부
교차표

	찬성	반대	합계
A	200	200	400
B	160	240	400
합계	360	440	800

A지역 지지율이 200/400=0.5(50.0%)이므로 B지역 지지율 160/400=0.4(40.0%)에 비해 10%포인트 높게 나타났다. 결과적으로 B지역이 낮은 지지율이 나왔으므로 이 지역에 대한 홍보를 강화하기로 하였다.

3. 남녀별 분리 필요 발견

추가적으로 남녀 간의 지지율을 비교하기 위하여 다음과 같은 표도 작성하였다.

<표 7-16>
성별과 지지여부
교차표

	찬성	반대	합계
남자	250	150	400
여자	110	290	400
합계	360	440	800

결과는 남자 지지율 62.5%, 여자 지지율 27.5%로서 여자들의 지지율이 남자에 비하여 상당히 낮게 나타났다. 따라서 할 수 있다면 지역 B에 홍보를 강화하되 여자를 주 대상으로 하는 방식의 홍보 전략을 기획하기로 하였다.

그런데 발표 자료를 작성하기 위하여 인구특성 변수에 따라 정리한 자료 중 다음의 표를 작성하였다.

<표 7-17>
성별과 거주지역
교차표

	남자	여자	합계
지역 A	300	100	400
지역 B	100	300	400
합계	400	400	800

4. 교락변수 출현

위의 표는 응답자 분포를 지역별, 성별로 정리한 것으로 산업특성상 지역 A에 남자들이 많고, 지역 B에는 여자들이 많은 인구실태를 나타낸 표이다. 여기서 K씨에게 불현듯 떠오른 생각은 무엇일까요?

지역에 따라 남녀 인구차이가 크다.

+

남녀 간에 지지율 차이가 크다.

⇓

지역 간 차이는 남녀 인구차이 때문에 생긴 문제는 아닐까?

그래서 K씨는 자료를 남녀로 분할하여 남녀별로 지역 간 지지율 차이를 분석해 보았다.

남자 응답자들로 분할표를 작성하면 다음과 같다.

<표 7-18> 남자의 거주지역과 지지여부 교차표

	찬성	반대	합계
A	180	120	300
B	70	30	100
합계	250	150	400

여자 응답자들로 분할표를 작성하면 다음과 같다.

<표 7-19> 여자의 거주지역과 지지여부 교차표

	찬성	반대	합계
A	20	80	100
B	90	210	300
합계	110	290	400

남자의 경우를 보면 A지역은 180/300=0.6으로 60%, B지역은 70/100=0.7로 70%가 되어 오히려 A지역이 낮게 나왔다. 또한 여자의 경우도 A지역 20% B지역 30%로 A지역이 낮게 나왔으므로 남녀모두 A지역의 지지율이 낮으므로 A지역에 대한 홍보를 강화하는 것이 바람직하다고 나타났다.

이 결과는 A, B 두 지역을 전체로 비교한 결과인 "B지역의 지지율이 더 낮다"는 것과 상반 된 것으로 모순이 된다. 왜 이런 결과가 나왔으며 우리는 어떻게 대처해야 하는가? 어떻게 XX정책에 대한 홍보 전략을 세울 것인가?

「홍보물은 여성을 대상으로 작성하고 지역은 여성이 많은 지역B를 우선으로 하는 것이 바람직」

5. 연령대별 분리 확인

K씨는 이렇게 모순된 결과를 보고 연령에 대해서도 검토를 해보았는데 그 결과는 다음과 같았다.

지역과 연령대로 작성된 분할표는 다음과 같다.

<표 7-20> 거주지역과 연령대 교차표

	20대	30대	합계
지역 A	210	190	400
지역 B	200	200	400
합계	410	390	800

20대의 지역과 지지여부로 작성된 분할표는 다음과 같다.

<표 7-21> 20대의 거주지역과 지지여부 교차표

20대	찬성	반대	합계
지역 A	105	105	210
지역 B	80	120	200
합계	185	225	410

30대의 지역과 지지여부로 작성된 분할표는 다음과 같다.

<표 7-22> 30대의 거주지역과 지지여부 교차표

30대	찬성	반대	합계
지역 A	94	96	190
지역 B	81	119	200
합계	175	215	390

다행히 연령대별로 작성한 분할표의 결과는 전체 지역간 지지율 비교 결과와 거의 같기 때문에 홍보 전략에서 연령은 고려하지 않아도 좋겠다는 결론을 얻게 되었다.

참고자료

- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석, 경문사.
- 허명회(2006), 통계적 사고, 교우사.
- 허명회(2011), 법과 통계학, 한나래 아카데미.
- 통계교육원(2008), 통계와 정책 표준교재.
- 통계청(2014), 사회조사 조사표.
- 통계청(2014), 사회조사.
- http://media.daum.net/economic/stock/others/newsview?newsid=20051114154616535

제 8 장

이상치 자료 탐색하기

8-1. 우연한 변동인가 이상치인가

학습목표

- **이상치를 정의하고 이상치를 규정하게 되는 두 가지 관점과 유형을 배우고 이상치가 제공하는 정보를 토의하여 이상치에 적절히 대응한다.**

1 모집단의 분포 관점에서 이상치

Barnett & Lewis(1994)는 이상치란 자료의 다른 관찰치들과 일관성이 없는 것으로 나타나는 관찰치나 관찰치의 집합이라고 정의한다. 비록 그들은 이렇게 간단하게 정의하였지만 이상치의 개념, 식별 등이 결코 단순하지 않다고 강조한다. 이상치를 모집단의 분포의 관점(즉 개체들간 차이는 우연한 차이라고 간주하는 입장에서)에서 우연한 차이라고 볼 수 없는 이상치, 구조적으로 데이터를 볼 때 구조에서 벗어난 이상치 등으로 나누어 설명한다. 여기는 먼저 모집단 분포 관점에서 이상치를 설명한다.

모집단 분포 관점에서 이상치(outliers)는 극단치(extreme observations) 그리고 이질자료(contaminants)와는 다르다.

예컨대, $x_1, x_2, \cdots, x_n$을 어떤 모집단에서 추출된 크기가 n인 표본의 관찰값이라고 하고 이들을 순서대로 나열한 값을 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$이라고 표현하자. 여기서 $x_{(n)}$은 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$중 가장 큰 값을 나타내고 $x_{(1)}$은 가장 작은 값을 나타낸다.

이러한 $x_{(1)}$과 $x_{(n)}$을 극단치라고 한다. 이 때 우리가 이 극단치 $x_{(1)}$과 $x_{(n)}$을 이상치 라고 식별 하느냐 하지 않느냐는 이들이 모집단에 속하는 개체로서 관찰되었다고 보느냐 보지 않느냐 하는 주관적인 판단과 연결되어 있

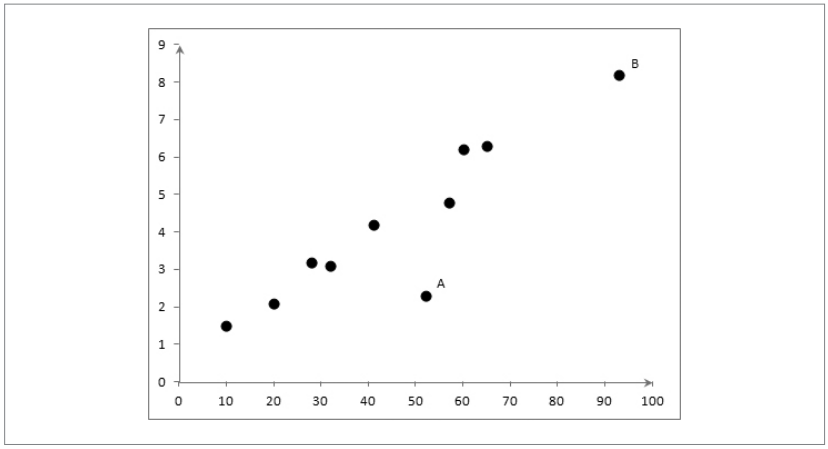
다. 그렇지만 모든 이상치는 극단치이거나 최소한 상대적인 극단치가 된다.

그런데 때때로 모집단에 속하지 않는 개체가 표본에 실수(?)로 포함되어 그 개체의 관찰값이 표본의 관찰값에 포함되는 경우가 있다. 이 때 우리는 이것을 이질자료(contaminants)라고 부르는데 이 값은 때로 극단치가 될 수도 있다. 한편 이질자료(contaminants)는 극단치와 같이 이상치가 될 수도 있고 또 아닐 수도 있다. 또한 이상치는 이질자료(contaminants)일수도 아닐 수도 있다. 그러나 우리가 어떤 관찰치가 이질자료(contaminants)인지 아닌지는 알 수가 없으므로 이상치나 극단치를 보면서 이질자료(contaminants) 여부를 신중히 판단해야 한다.

2 구조적 상황에서 발생하는 이상치

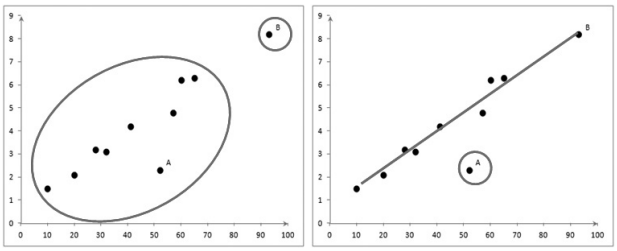
구조적 상황이란 우리가 갖는 관심이 한 집단이나 하나의 현상에 대한 관심이 아니라 이 집단이나 현상에 포함된 변수들 간의 내재적 관계에 관심을 갖는 상황을 의미한다. 아래의 그림은 사회조사에서 조사된 행복점수(100점 만점)와 삶의 질 점수(10점 만점)의 산점도이다.

[그림 8-1]
행복점수와
삶의질점수 산점도



이러한 경우 가장 먼저 이상치로 보이는 것은 무엇인가? 사람에 따라서 A를 또는 B를 이상치로 식별할 것이다.

[그림 8-2]
행복점수와
삶의질점수 산점도 –
이상치



우리가 앞에서 다룬 모집단의 분포 관점에서 본다면 B는 두 점수 모두 극단치로 보이기 때문에 B가 이상치라 할 수 있을 것이다.

그러나 또 다른 관점에서 본다면 A가 이상치가 된다. 이 때 관점은 무엇인가? 그것은 두 개의 특성이 갖는 선형관계라는 구조적 관계의 관점이다. 자료 A는 비록 각 변수의 값에서는 정상적으로 보이지만 이 구조적 관계 관점에서 이상치가 된다. 반면, 이러한 구조적 관점에서 B는 이상치라기보다는 이 구조(선형관계)에 영향을 크게 주는 관찰치로 간주하는 입장을 갖는다. (B의 조그만 변화에도 관계식의 절편은 요동치게 된다.)

❸ 이상치 유형

1. 일변량과 다변량 이상치

일변량 이상치는 하나의 변수와 관련된 이상치이고 다변량 이상치는 여러 개의 변수와 관련된 이상치이다. 어떤 개체의 다변량 관측치를 보면 각 개별 변수에 대해서는 이상치가 아니지만 다른 변수들과의 조합에서는 이상치로 나타날 수도 있다.

2. 전역적, 맥락적, 집단적 이상치

전역적(global) 이상치는 다른 값들과 유의하게 떨어져 있는 이상치로서 가장 일반적인 개념에서 말하는 이상치이다. 맥락적(contextual) 이상치는 조건부 이상치라고도 하는데 주어진 조건을 고려할 때 큰 편차를 보이는 이상치를 말한다. 예컨대 37도라는 온도가 측정되었을 때 이 온도가 관찰된 시점이 여름인 경우는 이상치가 아니지만 겨울인 경우에는 이상치가 된다. 즉 계절이라는 맥락에 따라 나타나는 이상치를 맥락적 이상치라고 한다. 그리고 집단적(collective) 이상치는 개별 값으로서는 이상치가

아니더라도 전체 자료를 고려할 때 집단적으로 구별되는 이상치로서 공간통계 등에서 주로 나타나는데 어느 부락에서 각 가구당 생산량이 조사되었는데 지리적으로 몰려있는 5가구의 생산량이 다른 가구들과 크게 차이가 많이 나는 그런 경우에 정의되는 이상치를 의미한다.

4 이상치에 포함된 정보

이상치는 통계분석의 결과에 적절하지 않은 영향을 끼치기 때문에 특히 관심의 대상이 되어 왔다. 예를 들어 이상한 관찰치들을 포함한 분석결과에서 평균은 좌나 우로 치우칠 수 있고 상관계수는 높거나 낮게 나타날 수 있으며 대부분의 점들을 대표하지 못하는 회귀식을 얻을 수도 있다. 이런 영향력 있는 값들은 구별되어야 하고 그 값들에 대한 의사결정이 필요하다.

그런데 이와 같은 이상치는 자료수집단계에서부터 분석에 이르기까지 전체 과정에 대하여 다음과 같은 특별한 정보를 제공한다.

- 실험과정의 위치(평균)와 스케일(변동)의 변화에 의해 발생하는 이상치는 실험과정의 변화를 의미할 수 있다.
- 때로 기록오차나 측정오차가 이상치를 발생시킬 수 있다.
- 분포에 대한 잘못된 가정에 의해서도 이상치가 만들어진다.
- 이상치는 잘못된 자료를 의미할 수 있다. 예를 들어 자료를 잘못 입력했거나 실험이 정확하게 수행되지 못했다는 것을 의미할 수 있다. 만일 이상한 값이 사실상 잘못에 기인한 것이라면 이상치는 수정되거나 분석에서 제외되어야 한다.
- 어떤 경우에는 이상한 값이 잘못된 자료인지 아닌지 결정이 가능하지 않을 수 있다. 이상치는 임의 변동에 의해 나타날 수 있고 또는 과학적으로 흥미 있는 무언가를 내포하고 있을 수 있다.

따라서 이상치로 의심되는 관찰치가 발견이 되었다고 해서 이상한 관찰치를 단순히 지워서는 안 된다. 만일 자료가 중요한 의미를 갖는 이상치를 포함한다면 우리는 통계적 기법의 사용을 고려해야 할 필요가 있다.

8-2. 그래프로 이상치 찾기

학습목표

- **이상치 식별방법 중 가장 쉬운 그래프를 통하여 식별하는 방법을 토의하며 이상치 식별법에 익숙해진다.**

1 일변량 자료의 이상치

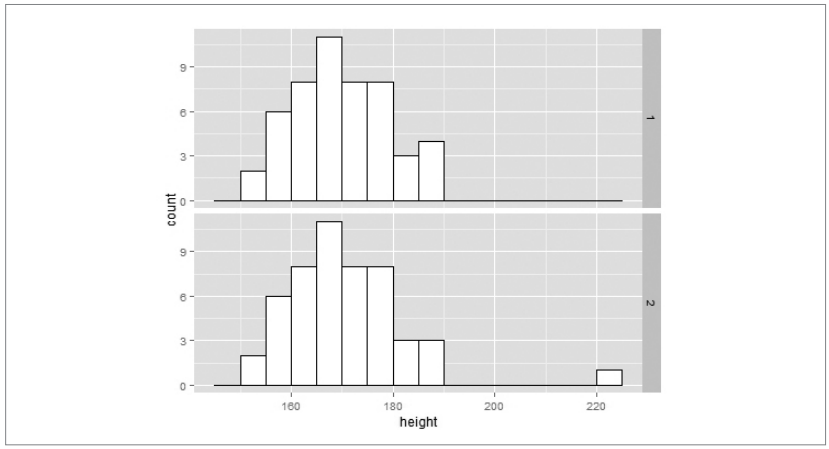
1. 히스토그램

다음은 50명의 키에 대한 가상의 2개 자료와 자료별 히스토그램이다.

<표 8-1> 신장자료

자료 1				자료 2			
id	height	id	height	id	height	id	height
1	156.9	26	157.4	1	156.9	26	157.4
2	165.7	27	154.3	2	165.7	27	154.3
3	163.2	28	162.9	3	163.2	28	162.9
4	188.5	29	177.1	4	220.0	29	177.1
5	181.6	30	181.7	5	181.6	30	181.7
6	157.9	31	165.7	6	157.9	31	165.7
7	169.5	32	169.2	7	169.5	32	169.2
8	176.6	33	164.8	8	176.6	33	164.8
9	171.4	34	168.0	9	171.4	34	168.0
10	174.3	35	177.6	10	174.3	35	177.6
11	159.6	36	166.2	11	159.6	36	166.2
12	168.8	37	179.4	12	168.8	37	179.4
13	159.9	38	184.3	13	159.9	38	184.3
14	158.3	39	162.1	14	158.3	39	162.1
15	177.3	40	173.2	15	177.3	40	173.2
16	167.9	41	186.3	16	167.9	41	186.3
17	172.0	42	172.7	17	172.0	42	172.7
18	172.4	43	166.7	18	172.4	43	166.7
19	151.5	44	170.2	19	151.5	44	170.2
20	179.2	45	166.0	20	179.2	45	166.0
21	186.0	46	187.1	21	186.0	46	187.1
22	163.2	47	175.7	22	163.2	47	175.7
23	166.9	48	164.2	23	166.9	48	164.2
24	161.1	49	173.2	24	161.1	49	173.2
25	175.6	50	163.4	25	175.6	50	163.4

[그림 8-3]
신장자료 히스토그램

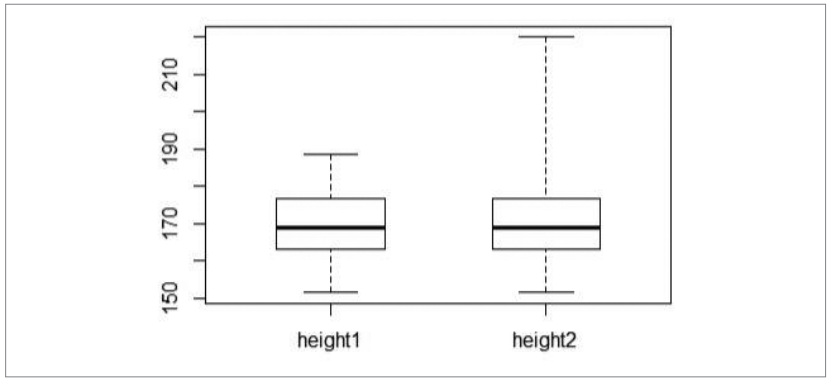


히스토그램을 보면 자료 1과 달리 자료 2의 경우는 오른쪽 끝으로 동떨어져 분포하는 값이 있음을 알 수 있다. 자료를 살펴보면 자료 2의 ID 4번은 키가 220cm인 것을 알 수 있다. 이 값을 극단치로 볼 것인지 이상치로 볼 것인지는 다시 살펴보아야 할 문제이지만, 이렇게 히스토그램에서 다른 자료들과 떨어진 값이 있는 경우에는 이상치나 극단치가 아닌지 의심해 볼 수 있다.

2. 상자그림

다음은 위의 자료를 상자그림으로 나타낸 것이다.

[그림 8-4]
신장자료 상자그림 (1)



5장에서 살펴보았듯이 상자그림에서 상자의 길이는 사분위범위(*IQR*) 즉 가운데 50%의 자료가 퍼져있는 범위이다. 상자에 이어진 선(수염)은 각각 상위 25%와 하위 25%가 퍼져있는 범위를 나타내는데, 이 길이가 상자에

비하여 너무 길면 한 쪽에 멀리 떨어진 값이 있을지 모른다는 생각을 하게 된다. 자료 2와 같이 위쪽 수염의 길이가 상자에 비하여 길게 나타나면 위쪽으로 극단치나 이상치가 하나 이상 있으리라고 예상할 수 있다. 그러나 그 값이 하나인지 여러 개인지 등을 구체적으로 살펴보기 어려우므로 다음과 같이 안 울타리와 바깥 울타리를 구하고, 이를 벗어나는 값은 따로 표시해 줄 수도 있다.

$$\text{아래 안 울타리} \quad IF_L = Q_1 - 1.5 * IQR$$

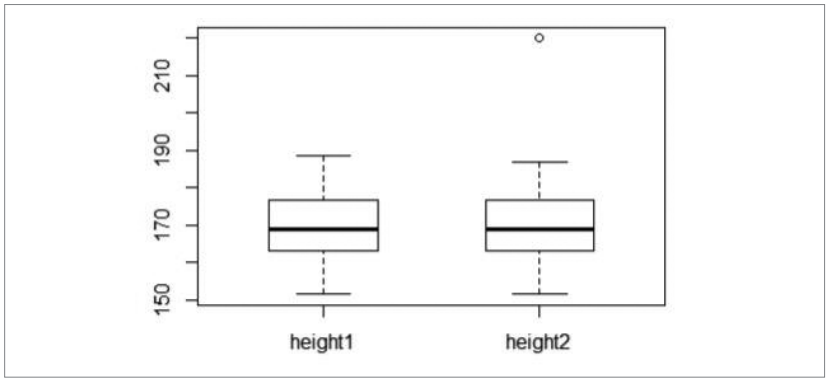
$$\text{아래 바깥 울타리} \quad OF_L = Q_1 - 3.0 * IQR$$

$$\text{위 안 울타리} \quad IF_U = Q_3 + 1.5 * IQR$$

$$\text{위 바깥 울타리} \quad OF_U = Q_3 + 3.0 * IQR$$

아래의 그림이 안 울타리와 바깥 울타리를 이용한 상자그림이다.

[그림 8-5]
신장자료 상자그림 (2)



"o"로 표시된 점이 안 울타리를 벗어나는 점이며 ID를 함께 표시할 수도 있다. 자료 1의 상자그림은 이런 값이 없고 자료 2의 경우에는 하나 있는 것을 알 수 있다. 앞의 상자그림보다 이상치나 극단치로 의심되는 값들을 좀 더 구체적으로 검토해 볼 수 있다.

2 이변량 자료의 이상치

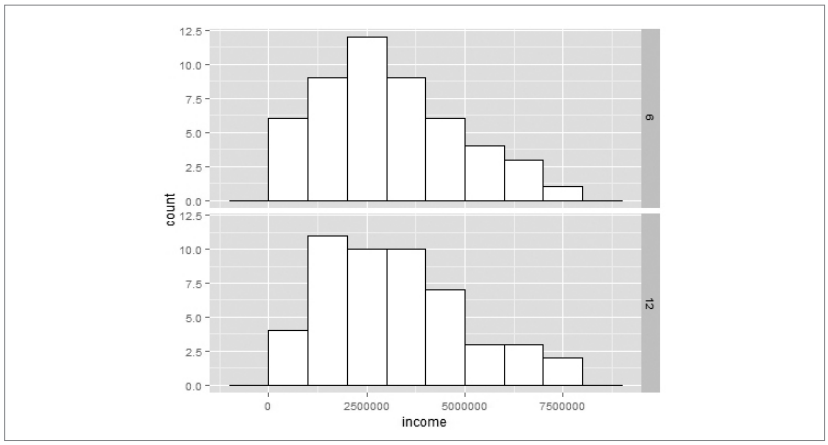
다음은 50가구의 2014년 6월 소득과 2014년 12월 소득에 대한 가상의 자료이다.

<표 8-2>
소득자료

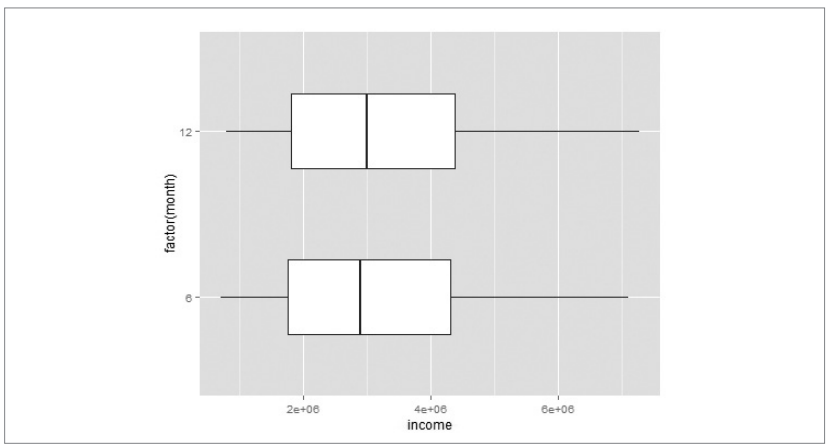
ID	incom 201406	income 201412	ID	income 201406	income 201412
1	1,720,850	1,564,780	26	3,490,700	3,501,700
2	3,950,060	3,675,360	27	5,830,000	6,236,340
3	899,100	1,000,000	28	2,269,420	2,269,420
4	3,375,500	3,497,709	29	3,954,800	3,954,800
5	780,000	4,100,000	30	1,870,040	1,774,290
6	1,500,000	1,500,000	31	759,960	782,792
7	4,957,770	4,796,667	32	2,881,262	2,617,740
8	3,064,370	3,077,910	33	2,393,290	2,293,290
9	4,350,000	4,000,000	34	1,299,100	1,316,000
10	6,281,569	6,652,212	35	2,220,000	2,220,000
11	700,000	800,000	36	2,240,000	2,200,000
12	5,800,000	5,600,000	37	2,926,930	3,148,420
13	1,805,110	1,914,610	38	3,928,050	3,750,000
14	1,737,980	1,730,030	39	6,500,000	6,500,000
15	6,929,920	7,286,520	40	3,416,760	3,600,000
16	2,413,500	2,413,500	41	1,123,700	1,224,600
17	4,381,030	4,676,450	42	2,080,000	2,170,020
18	981,830	981,830	43	3,000,000	3,300,000
19	7,107,160	7,229,660	44	4,368,144	4,470,300
20	1,049,000	1,100,000	45	3,221,980	3,221,980
21	847,100	936,000	46	2,688,940	2,415,607
22	2,975,470	2,933,070	47	4,200,000	4,470,000
23	1,336,800	1,336,800	48	5,670,000	5,677,140
24	2,386,500	2,350,000	49	4,358,760	4,492,900
25	5,598,080	5,608,440	50	2,046,900	1,940,000

아래는 6월 소득과 12월 소득 각각을 히스토그램과 상자그림으로 나타낸 것이다.

[그림 8-6]
소득자료 히스토그램



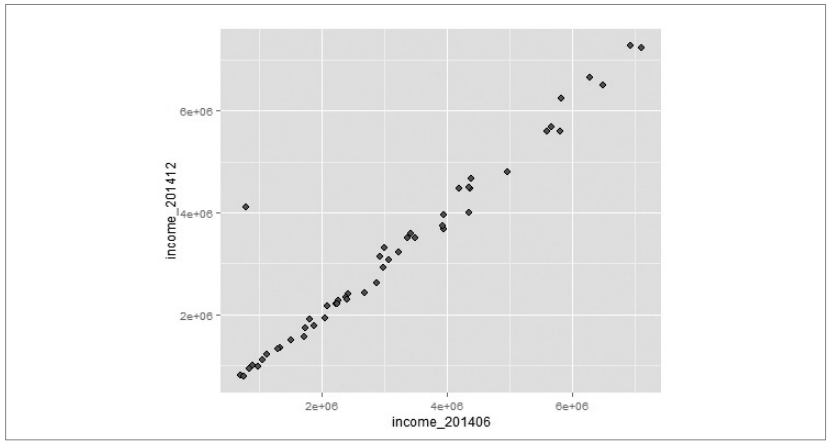
[그림 8-7]
소득자료 상자그림



앞 절의 내용을 다시 확인하여 보자. 히스토그램을 보면 약간 오른쪽으로 길게 뻗어서 분포하는 모양을 보이기는 하지만 따로 떨어져 있는 값은 보이지 않는다. 상자그림에서도 따로 "o"등으로 표시되는 점이 없으며 특별히 이상하다고 의심되는 값은 보이지 않는다.

그러나 위의 소득 자료는 한 가구가 두 개의 변수를 가지는 이변량 자료이다. 6장에서 살펴보았듯이 이러한 두 연속형 변수로 이루어진 자료의 경우 산점도를 통하여 두 변수간의 관계에 주목할 필요가 있다. 소득 자료로 산점도를 그려보면 아래와 같다.

[그림 8-8]
소득자료 산점도



산점도를 보면 대부분의 자료가 직선의 형태를 띠는데 반하여 한 점이 따로 떨어져 있는 것을 볼 수 있다. 가로축이 6월 소득이고 세로축이 12월 소득이므로 대부분이 6월 소득과 12월 소득이 비슷하게 나타났는데, 이 값은 6월의 소득에 비하여 12월이 굉장히(??) 크다는 것을 알 수 있다. 이러한 경우 이 값은 1절에서 확인한 것 같이 선형관계라는 구조적 관점에서 이상치로 의심해 볼 수 있을 것이다.

8-3. 통계치를 이용한 이상치 판정(심화)

학습목표

• 이 절에서는 통계적 기법을 이용하여 이상치를 식별하는 방법을 배운다.

1 이상치를 감추는 효과

앞에서 언급한 바와 같이 자료의 적용방법이나 자료의 크기에 따라 이상치를 식별하는 방법들은 다양하다. 어떤 이상치 검증법들은 하나의 이상치를 감지해내는 것인 반면, 또 어떤 것들은 여러 개의 이상치를 동시에 식별하는 것이다. 하나의 이상치를 알아내는 검증법을 여러 개의 이상치를 알아내는데 적용하는 것은 적절하지 않을 수 있다. 게다가 다수의 이상치를 식별하는 몇 가지 검증법 중에는 구체적으로 의심되는 이상치의 숫자를 지정하기도 한다.

그런데 이상치를 식별하는 과정에서 자료에 따라 이상치 식별이 어렵게 될 때가 있는데 이상치 탐색을 어렵게 하는 효과를 가면효과(masking effect)와 나눔효과(swapping effect)라고 한다. 다음 자료의 예를 통해서 이해해 보자. 다음의 표는 두 개의 순서대로 나열한 순위자료이다.

<표 8-3> 순서화된 자료

순서	(1)	(2)	(3)	(4)	(5)	(6)	(7)
자료A	3	4	7	8	10	13	950
자료B	3	4	7	8	10	949	950

이상치를 식별하기 위한 직관적인 방법 중 하나인 범위(range)에 대한 이웃과의 차이의 비를 생각해보자. 즉, $\frac{X_{(7)}-X_{(6)}}{X_{(7)}-X_{(1)}}$ 의 값이 크면 클수록 이상치라고 생각한다.

그러면 자료 A에서 최대값 $x_{(7)}=950$ 의 경우 범위에 대한 이웃과의 차이의 비는 $\frac{950-13}{950-3}=\frac{937}{947}$로서 우리가 육안으로 확인 할 수 있듯이 이상치라고 할 수 있다. 반면에 자료 B에서는 이 값이 $\frac{950-949}{950-3}=0.001$이다. 이는 거의 0에 가까운 값으로 이상치가 아니라고 할 가능성이 높고, 따라서 이 자료에는 이상치가 없다고 판단을 할 가능성도 높아진다. 이렇게 된 이

유는 최댓값의 바로 이웃값 $x_{(6)}$이 최댓값과 비슷한 949로 또 다른 이상치이기 때문이다. 이러한 상황을 " $x_{(6)}$ 이 $x_{(7)}$에 가면을 만들었다($x_{(6)}$ has masked $x_{(7)}$)"고 한다.

반면에 나눔효과란 두 개 이상의 이상치를 한 번에 식별할 때 발생하는 것으로서 만약 이상치를 두 개 찾는다고 하는 상황을 생각하면 자료 B에서는 당연히 $x_{(6)}$과 $x_{(7)}$을 이상치라고 할 것이다. 그런데 자료 A에서는 $x_{(7)}$= 950이 이상치이고 두 개의 이상치를 찾는 중이므로 $x_{(6)}$도 이상치로 식별되는 오류가 발생한다. 이런 상황을 " $x_{(7)}$ 이 $x_{(6)}$과 이상치 효과를 나누었다($x_{(7)}$ has swapped $x_{(6)}$)"고 한다.

위의 예에서 가면효과를 제거하는 방법 중 하나는 자료 B와 같은 상황이 발생할 때는 최댓값 $x_{(7)}$부터 식별판단을 하는 것이 아니라 이웃인 $x_{(6)}$부터 식별을 시작하는 것이다. $x_{(6)}$을 이상치라고 판단하면 $x_{(7)}$은 더 이상의 식별작업 없이 이상치라고 판단되기 때문에 $x_{(6)}$ 의 가면효과를 제거할 수 있게 된다.

❷ 모집단 분포 관점에서 이상치의 식별

1. 사분위수 범위 이용

사분위수 범위를 이용한 이상치의 판정은 그래프를 이용한 방법에서 설명이 되었다. 여기서는 복습의 차원에서 아래의 자료로부터 사분위수 범위를 이용하여 이상치를 식별해본다.

<표 8-4>
자료

ID	x	ID	x	ID	x
1	99.08	13	107.43	25	112.69
2	99.41	14	107.98	26	114.33
3	100.77	15	108.15	27	114.51
4	101.55	16	108.17	28	115.48
5	101.94	17	108.37	29	115.99
6	103.61	18	108.50	30	116.38
7	104.57	19	108.83	31	116.71
8	105.11	20	109.07	32	118.31
9	106.13	21	109.57	33	119.59
10	106.55	22	109.58	34	119.86
11	106.73	23	110.67	35	121.88
12	107.16	24	111.22	36	131.87

사분위범위를 구해보면 다음과 같다.

$$IQR = Q_3 - Q_1 = 114.75 - 106.45 = 8.31$$

여기에 1.5를 곱하면 안울타리가 되고 3.0을 곱하면 바깥울타리가 된다. 이 값을 벗어나면 이상치로 생각해볼 수 있다.

안울타리의 범위는, $1.5 \times IQR = 12.46$이므로 (93.98, 127.21)이 되고 마지막 관측값인 131.87은 안울타리를 벗어나는 이상치로 볼 수 있다. 바깥울타리는 $3.0 \times IQR = 24.92$이므로 (81.52, 139.68)가 되며 바깥울타리를 벗어나는 이상치는 없다.

2. 절사평균(trimmed means)

절사평균이란 가장 높고 가장 낮은 특정 퍼센트만큼을 버리고 난 후 남아있는 값들로 평균을 계산하는 것이다. 절사평균은 이상치에 상대적으로 저항성이 크다는 장점이 있다. 자료에 이상치가 나타날 때 절사평균은 이상치에 덜 민감한 모평균의 추정치가 된다. 만일 상위 그리고 하위 5%의 자료가 제거되면 그것은 10% 절사평균이 된다. 만일 절사평균과 절사되지 않은 평균을 비교하면 이상치를 식별할 수 있으며 또 이상치의 효과 등을 알 수 있게 된다. 위의 예에서 전체 평균과 10% 절사평균의 차이를 구해보자.

평균은 110.22이고, 자료를 순서대로 나열하여 양 끝 2개(5%)를 제외한 10% 절사평균은 109.86이 되고 두 평균의 차이는 0.36이다.

3. 표준점수(Z-score)

11장에서 다루겠지만 고교과정에서 배웠던 정규분포를 따른다고 생각하는 현상에서 얻은 자료가 $x_1, x_2, \cdots, x_n$라고 할 때 표준점수는 다음과 같이 정의되는 값을 말한다.

$$Z_i = \frac{x_i - \bar{x}}{s},\ \bar{x}\text{는 표본평균},\ s\text{는 표본 표준편차}$$

이 값은 관찰치가 평균으로부터 표준편차의 몇 배만큼 떨어져 있는가를 나타내는 것으로 정형화된 규칙은 아니지만 표본크기가 크지 않은 경우(80보다 작은 경우)는 표준점수가 ± 2.5를 넘으면 이상치로 간주하고, 표본의 크기가 큰 경우(80보다 큰 경우)는 ± 3.0을 넘으면 이상치로 간주하는 관점도 있다.

표준점수가 이상치를 판별하는데 도움을 주기는 하지만 표본의 크기가 작은 경우에는 잘못 해석될 수 있다. 수리통계학적 이론에 의하면 크기가 n인 표본에서 표준점수의 최댓값은 $\frac{n-1}{\sqrt{n}}$이 된다. 따라서 표본의 크기가 작은 경우에는 이상치를 판별하기 어렵다. 예를 들어, 표본의 크기가 6인 경우 표준점수의 최댓값이 약 2.04($=\frac{6-1}{\sqrt{6}}$)이므로 자료 내의 어떤 값도 이상치라고 판단할 수 없게 된다. 따라서 그럴 경우에는 중앙값을 이용한 다음의 수정된 표준점수를 이용한다. 이 값이 3.5보다 큰 경우 이상치로 간주한다.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD},\ \tilde{x}\text{는 표본중앙값}$$

여기서 MAD(median absolute deviation)는 중앙값을 이용하여 절대편차를 구한 후, 다시 이 편차들의 중앙값을 구하여 얻어진다.

$$MAD = median\left(|x_i - \tilde{x}|\right)$$

주어진 자료로부터 표준점수와 수정된 표준점수를 구해보자.

<표 8-5>
자료 –
표준점수와 수정된
표준점수

ID	x	z-score	M_i	ID	x	z-score	M_i
1	99.08	-1.60	-1.59	19	108.83	-0.20	0.03
2	99.41	-1.55	-1.54	20	109.07	-0.16	0.07
3	100.77	-1.36	-1.31	21	109.57	-0.09	0.15
4	101.55	-1.25	-1.18	22	109.58	-0.09	0.15
5	101.94	-1.19	-1.12	23	110.67	0.07	0.33
6	103.61	-0.95	-0.84	24	111.22	0.14	0.42
7	104.57	-0.81	-0.68	25	112.69	0.36	0.67
8	105.11	-0.73	-0.59	26	114.33	0.59	0.94
9	106.13	-0.59	-0.42	27	114.51	0.62	0.97
10	106.55	-0.53	-0.35	28	115.48	0.76	1.13
11	106.73	-0.50	-0.32	29	115.99	0.83	1.22
12	107.16	-0.44	-0.25	30	116.38	0.89	1.28
13	107.43	-0.40	-0.21	31	116.71	0.93	1.34
14	107.98	-0.32	-0.11	32	118.31	1.16	1.60
15	108.15	-0.30	-0.09	33	119.59	1.35	1.82
16	108.17	-0.29	-0.08	34	119.86	1.39	1.86
17	108.37	-0.27	-0.05	35	121.88	1.68	2.20
18	108.5	-0.25	-0.03	36	131.87	3.11	3.86

마지막 관측값인 131.87은 표준점수가 3.11이고 수정된 표준점수는 3.86이므로 두 가지 방법 모두에서 이상치로 간주된다.

참고자료

- V. Barnett and T. Lewis (1994), Outliers in Statistical Data, John Wiley & Sons.
- http://www.munhwa.com/news/view.html?no=2011032501070123082002
- http://www.hani.co.kr/arti/politics/assembly/549296.html

제 9 장 가중치 알아보기

9-1. 조사항목 가중치 설정

학습목표

- **설문조사의 문항에 대한 응답결과를 종합할 때 발생하는 각 문항의 중요도를 반영하는 가중치를 결정하는 문제, 국가통계의 작성이나 활용에서 나타나는 가중치의 기본개념을 이해한다.**

다음은 어느 지역 정책에 대한 주민의 만족도 조사문항과 4명의 응답자 A, B, C, D의 응답결과이다.

1. 경제생활에 도움이 되십니까?

① 전혀 그렇지 않다 ② 그렇지 않다 ③ 보통이다
④ 그렇다 ⑤ 아주 그렇다

2. 문화생활에 도움이 되십니까?

① 전혀 그렇지 않다 ② 그렇지 않다 ③ 보통이다
④ 그렇다 ⑤ 아주 그렇다

3. 자녀교육에 도움이 되십니까?

① 전혀 그렇지 않다 ② 그렇지 않다 ③ 보통이다
④ 그렇다 ⑤ 아주 그렇다

<표 9-1>
만족도 응답결과 (1)

	문항1 (경제생활)	문항2 (문화생활)	문항3 (자녀교육)	합계
A	1	3	5	9
B	3	3	3	9
C	5	3	1	9
D	2	5	2	9

4명의 총합은 모두 9로 평균 3점의 만족도를 보여주고 있다. 그런데 이러한 분석에서 경제생활, 문화생활, 자녀교육에 대한 만족도는 모두 동일하게 취급된다. 즉, 경제생활 만족도 5점과 문화생활 만족도의 5점을 같은 것으로 생각한다는 것이다. 그러나 조사목적에 따라서는 경제생활 만족도를 문화생활 만족도보다 더 중요하게 생각하고 자녀교육 만족도를 경제생활 만족도보다 더 중요하게 생각해야할 경우도 있다. 만약 경제생활을 문화생활보다 2배 더 중요하게 생각하고, 자녀교육이 문화생활보다 4배, 경제생활보다 2배 더 중요하다고 생각한다면 위의 조사 자료는 다음과 같이 바뀐다.

<표 9-2>
만족도 응답결과 (2)

	2×문항1 (경제생활)	1×문항2 (문화생활)	4×문항3 (자녀교육)	합계
A	2	3	20	25
B	6	3	12	21
C	10	3	4	17
D	4	5	8	17

따라서 지역 정책에 대한 만족도는 응답자 중 A가 가장 높고, 그 다음이 B 그리고 C, D는 같은 값이 된다. 보다 현실적인 또 다른 예는 마케팅 영역에서 찾을 수 있다. 우리가 음식점을 선택할 때 맛도 좋고, 가격도 좋고, 분위기도 좋은 곳을 찾지만 이렇게 세 가지를 모두 만족시키는 곳은 여간해서 찾기 어려운 법이다. 이러한 경우에 맛, 가격, 분위기라는 요소 중에서 어느 요소를 더 중요하게 생각하느냐에 따라서 결정하게 된다. 이러한 상황에서 일반적으로 사용하는 방법이 가중치를 이용하는 것인데 가중치를 구하는 것은 쉬운 방법이 아니다. 그 이유는 우선 실제 복잡한 상황에서는 가중치를 적절히 정하기가 어렵고, 둘째는 그 가중치가 의사결정판단을

정확히 반영해 주는지에 대한 객관적 근거가 없거나 시간에 따라 변할 수 있다는 것이다.

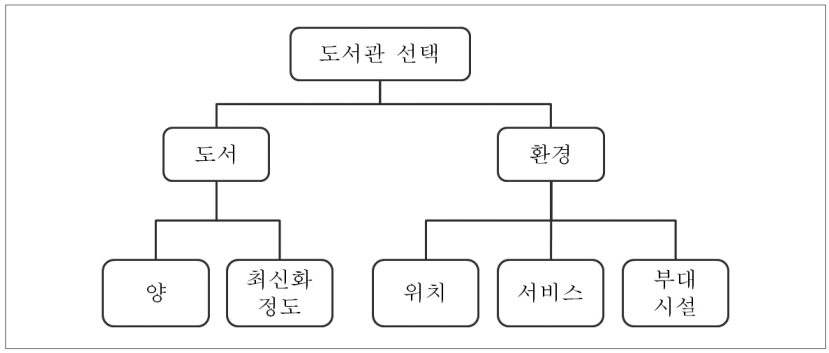
이러한 가중치 도출방법으로는 5점 척도, 7점 척도 등으로 알려진 척도표시법과 합계 100점을 가지고 각 속성에 중요도에 따라 분배하여 중요한 속성에 많은 점수를 부여하는 점수할당법이 있는데 이 방법은 누구나 쉽게 사용할 수 있는 방법이다. 한편 다소 복잡한 방법으로는 다중회귀분석법, 교환분석법, 그리고 Analytic Hierarchy Process(AHP)라고 불리는 '계층적 분석 과정/방법'이라는 것이 있는데 이는 수리적인 훈련이 요구되는 방법들이다. 이들 중 본 절에서는 직관적 이해가 가능하다고 판단되는 AHP의 각 단계와 특징을 다음의 예를 통해서 알아보자.

AHP 사례) 도서관 만족도에 미치는 요인들의 우선순위는?

도서관을 선택을 할 때 사람들이 생각하는 좋은 도서관은 어떤 것일까?

도서관 선호에 관한 조사결과를 보면 그 속성들은 다양하다. 어떤 연구자가 이를 다음과 같은 간단한 계층구조로 정리하였다.

[그림 9-1]
도서관 선호 계층구조



우선 계층1의 두 가지 속성(도서, 환경)을 비교하여 도서관을 선택하는데 어느 것이 더 중요한가를 판단하게 된다. 이 때 도서를 환경보다 3배 더 중요하게 생각한다고 하면 이를 속성비중행렬로 다음과 같이 표현한다.

$$\begin{array}{c} \\ \text{도서} \\ \text{환경} \end{array} \begin{array}{c} \begin{array}{cc} \text{도서} & \text{환경} \end{array} \\ \begin{bmatrix} 1 & 3 \\ 1/3 & 1 \end{bmatrix} \end{array}$$

각 속성의 가중치는 각 열의 수치를 그 열의 합계로 나누어 준 후 행별로 평균값을 구한다.

	도서	환경
도서	1	3
환경	1/3	1
합	4/3	4

계층 1의 가중치

	도서	환경	행평균
도서	3/4	3/4	3/4
환경	1/4	1/4	1/4
합	1	1	1

계층1의 가중치가 구해졌으므로 계층2로 내려간다. 계층2의 속성 양과 최신화정도가 여기서는 계층1에 기여하는 정도를 구하는 것이 된다. 다르게 말하면 양과 최신화정도 속성 중에서 어느 것이 더 도서 속성에 기여하는가 하는 질문을 하고 있는 것이다. 도서를 결정하는데 있어서 최신화정도가 양보다 3배 더 중요하게 생각한다면 도서의 가중치가 3/4이므로 이를 3:1로 분배하여 속성비중행렬은

최신화정도	1	3
양	1/3	1
	4/3	4

$$\text{최신화정도의 가중치} : \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$$

$$\text{양의 가중치} : \frac{3}{4} \times \frac{1}{4} = \frac{1}{16}$$

이 된다. 환경의 경우, 계층 2의 속성이 세 가지이므로 쌍별 비교를 해야 하는데 그 결과가 다음과 같다고 하자

위치 : 서비스 = 3:1

위치 : 부대시설 = 1:3

서비스 : 부대시설 = 1:2

여기서, 속성비중행렬은

	위치	서비스	부대시설
위치	1	3	1/3
서비스	1/3	1	1/2
부대시설	3	2	1
합계	13/3	6	11/6

계층 1에서 환경의 가중치는 1/4이었으므로, 계층 2에서의 가중치는

	위치	서비스	부대시설		행평균
위치	0.2308	0.5000	0.1818		0.3042
서비스	0.0769	0.1667	0.2727	⇒	0.1721
부대시설	0.6923	0.3333	0.5454		0.5237
합	1	1	1		1

$$\text{위치의 가중치} : \frac{1}{4} \times 0.3040 = 0.0760$$

$$\text{서비스의 가중치} : \frac{1}{4} \times 0.1721 = 0.0430$$

$$\text{부대시설의 가중치} : \frac{1}{4} \times 0.5237 = 0.1309$$

따라서 도서관 선택에 영향을 미치는 요인으로 선택된 양, 최신화정도, 위치, 서비스, 부대시설에 대한 가중치는 다음과 같아서

양	최신화정도	위치	서비스	부대시설
0.5625	0.0625	0.0760	0.0430	0.1309

양을 다른 어떤 것보다도 중요하게 생각하고(0.5625) 그 다음으로는 부대시설을 중요하게 생각하는 것(0.1309)으로 나타났다.

9-2. 표본개체 가중치 결정

학습목표

• 표본 추출 과정에서 추출된 개체에게 부여하는 가중치를 결정하는 문제, 표본 설계나 표본자료에 나타나는 가중치라는 용어에 대한 이해를 깊이 한다.

1 표본개체에 대한 가중치 논란

다음은 2015년 중앙선거여론조사공정심의위원회의 '제20대 국회의원선거 「선거여론조사기준개정」을 위한 공청회'에서 나온 토론내용의 일부로서 선거관리위원회는 현재 공표되는 여론조사의 공정성 문제의 하나를 가중치 문제로 보고 있다.

가중값 배율 관련 개정

선거여론조사를 통해서 얻게 되는 추정결과는 조사된 데이터에 가중값을 적용하여 산출하게 된다. 가중값은 선거여론조사에서 각 응답자가 해당 성별·연령대별·지역별 특성에 따라 모집단을 대표하는 정도를 의미한다. 일반적으로 가중값 조정 혹은 표본 할당을 위해 사용되는 인구 특성 변수는 거주 지역, 성 그리고 연령 그룹이며, 이 경우 최종 추정값을 계산하기 위하여 사용된 가중값의 최대값 혹은 최소값이 매우 크거나 작은 경우는 추출된 표본의 성, 연령 그리고 지역 분포가 모집단과 매우 다를 수 있음을 의미한다.

<표 2>는 제6회 지방선거 여론조사에서 사용된 가중값 분포 현황으로 연령대별 가중값의 변동이 매우 심하게 나타나는 것을 확인할 수 있다(성별과 지역별 가중배율은 대체로 0.5~2.0이었음). 비록 가중값 조정을 통한 모수 추정의 정확도를 높이고 있으나, 과도한 가중배율은 추정의 정확도를 크게 떨어뜨릴 수 있고, 유권자에게는 그 자체만으로 여론조사에 대한 심각한 불신을 초래할 수 있다. 이 같은 이유로 과도한 가중값 배율에 대해서는 최소한의 제한이 필요하다고 판단되며, 처음 수치화하는 것임을 감안하여 결과분석자료 응답자 특성표상에서 성 연령 지역별로 가중값 배율 제한기준을 두는 것이 타당할 것이다.

다음은 언론매체에 여론조사결과를 발표할 때 발표내용 말미에서 쉽게 발견하는 문구이다.

이번 조사는 11월 3일과 4일 이틀간 전국 19세 이상 성인 1,000명을 대상으로 휴대전화(50%)와 유선전화(50%) 임의전화걸기(RDD) 자동응답 방식으로 진행했고, 행정자치부 주민등록 인구통계 기준 성, 연령, 권역별 인구비례에 따른 가중치 부여를 통해 통계 보정했다. 응답률은 5.1%, 표본오차는 95% 신뢰수준에서 ±3.1%p이다.

http://www.hg-times.com/news/articleView.html?idxno=100956

❷ 기본 가중치의 정의

표본조사에서 확률 p_i로 추출된 표본의 개체 i는 모집단에 있는 개체들을 $1/p_i$만큼 대표한다. 이 때 추출된 개체 i의 가중값을 $w_i = 1/p_i$라 하고 이를 기본가중치라고 부른다. 예를 들면 추출확률 1/10로 추출된 표본의 개체는 모집단의 개체 10개를 대표한다.

❸ 가중치가 필요한 경우

1. 이질적 표본추출확률 보정

대기업과 중소기업의 두 개의 부모집단으로 나누어지는 모집단에 대해 월평균임금을 추정하는 경우를 살펴보자.

<표 9-3>
월평균임금

	부모집단		전체
	대기업	중소기업	전체
모집단크기	50	1,950	2,000
표본크기	5	95	100
월평균임금(만원)	1,500	300	
가중치	10	20.5	

가중치를 적용하지 않은 평균 월평균임금은 다음과 같다.

$$\frac{1{,}500 \times 5 + 300 \times 95}{100} = 360(\text{만원})$$

그러나 각 부모집단별로 표본의 추출확률이 동일하지 않으므로 가중치를 이용하여 조정하면 다음과 같이 구할 수 있다.

대기업의 가중치는 추출확률이 5/50=0.1이므로 1/0.1=10이고 중소기업의 추출확률은 100/1,950=0.049 가중치는 1/0.049=20.5이다. 따라서 가중치를 고려한 평균 월평균임금은 다음과 같다.

$$\frac{1{,}500\times 5\times 10+300\times 95\times 20.5}{2{,}000}=330(\text{만원})$$

2. 무응답 보정

(1) 무응답을 보정할 때

박진우(2006)에서 사용한 자료를 변형하여 설명해본다. N=10인 가구로 이루어진 모집단을 가정하여 생각해보자. 다음은 이 모집단의 소득 자료이다.

<표 9-4>
소득자료 – 모집단

가구번호	소득
1	245
2	53
3	317
4	82
5	148
6	242
7	600
8	145
9	244
10	352

모집단에 대하여 임의로 $n=5$인 가구를 표본으로 추출하여 소득을 조사하여 다음의 결과를 얻었다고 하자. 소득이 표시되지 않은 6번 가구는 무응답을 나타낸다.

<표 9-5>
소득자료 – 표본 (1)

가구번호	기본가중치	소득
1	10/5	245
2	10/5	53
5	10/5	148
6	10/5	.
10	10/5	352

모집단의 소득 총액을 추정하고 싶다면 $\frac{10}{5}(245+53+148+?+352)$을 계산해야 하지만 6번 가구의 소득이 무응답이므로 구할 수 없다. 만일 무응답이 임의로 발생한 것이라면 응답된 자료만을 이용하여 다음과 같이 구할 수 있을 것이다.

<표 9-6>
소득자료 – 표본 (2)

가구번호	기본가중치	소득
1	10/4	245
2	10/4	53
5	10/4	148
6		.
10	10/4	352

소득 총액은 $\frac{10}{4}(245+53+148+352)=1{,}995$가 된다. 가중치가 $\frac{10}{5}$에서 $\frac{10}{4}$로 변경된 것을 볼 수 있다. 이렇게 무응답으로 인한 효과를 고려하여 가중치로 조정하고 총계나 평균 등을 추정할 수 있다.

(2) 보조정보가 있는 경우

위의 모집단에 다음과 같이 가구주의 학력에 대한 변수가 추가되었다고 생각하여 보자. 모수 총소득은 2,336만원이다.

<표 9-7>
보조정보 포함
소득자료 – 모집단

가구번호	소득	가구주학력
1	245	A
2	53	B
3	317	A
4	82	A
5	148	B
6	242	B
7	600	B
8	53	A
9	244	B
10	352	B

$n=6$인 표본을 임의로 추출하여 다음의 자료를 얻었다고 하자. 5번, 6번, 8번 가구의 소득 항목은 무응답이다.

<표 9-8>
보조정보 포함
소득자료 – 표본 (1)

가구번호	기본가중치	소득	가구주학력	가중치
1	10/6	245	A	4/3
2	10/6	53	B	6/3
5	10/6	·	B	6/3
6	10/6	·	A	4/3
8	10/6	·	A	4/3
10	10/6	352	B	6/3

가구주의 학력에 대한 정보가 있으므로 이 정보를 고려한 가중치를 이용하여 총소득을 추정하면 $\frac{4}{3}(245+?+?)+\frac{6}{3}(53+?+352)$이 된다. 그러나 무응답이 있으므로 아래와 같이 가중치를 조정하여 보자.

<표 9-9>
보조정보 포함
소득자료 – 표본 (2)

가구번호	기본가중치	소득(만원)	가구주학력	가중치
1	10/3	245	A	4
2	10/3	53	B	6/2
5		·	B	
6		·	A	
8		·	A	
10	10/3	352	B	6/2

가구주학력이 A인 경우 가중치가 4/3에서 4로 조정되었으며, B의 경우는 6/3에서 6/2로 조정된 것을 볼 수 있다. 이 가중치를 이용하여 총소득을 추정하면 다음과 같은 값을 얻을 수 있다.

$$4(245)+\frac{6}{2}(53+352)=2{,}195$$

따라서 모수가 2,336만원이므로 오차는 2,336 – 2,195=141만원이다.

3. 사후층화로 특정 변수의 분포 조정

240가구에서 6가구를 임의로 추출하고 다시 표본가구에서 임의로 1명의 성인 *b*을 추출하여 삶의 질 점수 y_i를 측정한다. 해당 표본이 취업자이면 $z_i=1$이고 미취업자이면 $z_i=0$이다.

이 때 가구와 성인이 동시에 표본으로 추출될 확률은 다음과 같다.

가구 i가 추출될 확률은 $\frac{6}{240}$ 이고, 가구 i에 있는 성인의 수를 A_i라고 하면 가구 i에서 성인 *b*가 추출될 확률은 $\frac{1}{A_i}$이 된다. 그러므로 가구 i가 추출되고 또 성인 *b*가 추출될 확률 p_i는

$$p_i=\frac{6}{240}\times\frac{1}{A_i}$$

이 된다. 따라서 가구 i에서 추출된 한 명의 성인에게 부여되는 가중치 w_i는

$$w_i=\frac{1}{p_i}$$

이 된다. 다음의 자료에 적용해 보자.

표본가구와 그에 따른 각 값들이 다음 표와 같을 때, 평균 삶의 질 점수와 평균 취업자 수를 추정해 보자.

<표 9-10> 표본가구와 성인수 및 가중치

가구 (i)	기본 가중치	성인 수 (B_i)	가중치 (w_i)	점수 (y_i)	취업여부 (z_i)	$w_i \times y_i$	$w_i \times z_i$
1	40	2	80	7	1	560	80
2	40	1	40	9	1	360	40
3	40	3	120	6	0	720	0
4	40	2	80	8	1	640	80
5	40	3	120	4	0	480	0
6	40	4	160	3	0	480	0
합계		15	600	37	3	3,240	200

각 가구의 가중치를 고려하지 않은 평균 삶의 질 점수와 평균 취업자 수의 추정치는 다음과 같다.

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{37}{6} = 6.2, \qquad \bar{x} = \frac{1}{n}\sum z_i = \frac{3}{6} = 0.5$$

한편 각 가구의 가중치를 고려한 평균 삶의 질 점수와 평균 취업자 수의 추정치는

$$\bar{y} = \frac{\sum w_i y_i}{\sum w_i} = \frac{3{,}240}{600} = 5.4, \qquad \bar{x} = \frac{\sum w_i z_i}{\sum w_i} = \frac{200}{600} = 0.33$$

이 된다. 따라서 차이가 $0.5 - 0.33 = 0.17$이 되어 가중치 고려여부가 상당한 차이를 보인다.

9-3. 물가지수와 가중치

학습목표

- **물가지수를 산출하는 과정에서 품목별 가중치의 결정문제 등을 토의하여 가장 잘 알려진 통계 중 하나인 물가에 대한 이해를 깊게 하는 것과 가중치 평균에 대한 것을 이해한다.**

1 물가지수 산출법

물가란 일반적으로 우리의 경제생활 과정에서 거래되는 제반 상품과 서비스의 가격을 평균한 종합적인 가격수준을 말하는데 이러한 종합적인 개념으로서 물가의 움직임을 구체적으로 측정하여 작성되는 통계이다. 따라서 물가(통계)를 작성할 때는 수많은 개별상품과 서비스의 가격을 거래량(구입하는 쪽에서 보면 지출액)에 따라 가중평균하게 되는데 이때 사용되는 품목별 비중을 가중치라 한다. 즉 물가에서 가중치는 전체 집단에서 개별 구성요소가 차지하는 비중이나 중요도를 수치로 나타낸 값을 말한다.

이석훈 등(2008)이 통계와 정책에서 사용한 다음의 표를 통하여 물가지수를 이해해 보자.

<표 9-11> 품목별지수 및 가중치

품목	기준년도 가격	기준년도 거래량	현재가격	품목별지수	가중치
품목 A	1000/kg	10kg	1000/kg	100	10,000
품목 B	100/개	5개	200/개	200	500
품목 C	200/l	5l	600/l	300	1,000

품목별지수는 현재가격을 기준년도 가격으로 나눈 값에 100을 곱한 것으로 품목 B에서 품목별지수는 $\frac{200/개}{100/개} \times 100 = 200$이 된다. 가중치는 소비지출금액을 말하는 것으로 품목 B의 가중치는 $100/개 \times 5개 = 500$ 으로 계산된다.

만일 물가지수를 가중치를 고려하지 않고 품목별 지수들의 산술평균으로 구하면 (100+200+300)/3=200이므로 기준년도에 비하여 물가가 2배 상승한 셈이 된다. 그러나 각 구성품목의 상대적 중요도를 고려한 가중치를 이

용하여 평균을 구하면 다음과 같다.

$$\frac{(100\times 10{,}000+200\times 500+300\times 10{,}000)}{(10{,}000+500+10{,}000)}=122$$

즉 물가수준이 기준년도에 비하여 1.22배 상승했다고 할 수 있다.

❷ 통계물가와 체감물가

다음은 2015년 언론에 나온 "0%대 저물가? 도대체 어느 나라 통계입니까! 뿔난 주부들 가장 큰 이유는 '가중치'"라는 제목의 기사(http://www.fnnews.com/news/201510041724005963)를 요약한 것으로 통계청 품목마다 가중치가 다른데 2012년 기준이라 괴리감이 큰 것이고 또 하나는 '주관적 느낌'으로 집집마다 소비품목 모두 달라 자주 사는 상품 가격에 더 민감하기 때문이라는 설명이 붙어 있다. 이 기사는 소비자 물가지수에 사용되는 가중치를 다음과 같이 설명하였다.

> 4일 통계청에 따르면 현재 소비자물가를 구성하는 품목은 481개다. 목돈을 지출해야 하는 전·월세를 비롯해 매달 내는 도시가스, 수시로 구입하는 쇠고기·돼지고기, 10년에 한번 살까 말까 한 자동차, TV, 냉장고 등이 모두 여기에 포함된다. 481개 품목은 2010년 기준으로 결정됐다. 또 소비자물가에는 '가중치'라는 개념이 있다. 가중치는 각 가정의 가계부, 즉 매달 어느 항목에 얼마를 지출하는지 살펴보는 '가계동향조사'를 토대로 산출한다. 많이 지출하는 품목에는 그만큼 가중치를 줘 물가변동 시 더 많은 영향을 미치도록 한 것이다. 481개 품목의 가중치 합은 총 1000이다. 부문별로는 주택·수도·전기 및 연료가 173으로 가장 많고 식료품 및 비주류음료(139), 음식·숙박(121.6), 교통(111.4)순이다.
>
> 이 가중치는 당초 '0'과 '5'가 들어가는 해마다 변경했지만 생활패턴이 빠르게 변하면서 5년 사이 한 차례 더 변경하고 있다. 현재 가중치는 2012년 것으로 올해를 기준으로 내년에 추가로 바꿀 예정이다.

이어서 통계와 실제 느끼는 물가가 차이가 나는 이유가 물가통계 품목과 가중치 때문임을 다음의 수치를 사용하여 설명하고 있다.

통계청의 9월 조사에 따르면 1년 전과 비교해 담배는 국산이 83.7%, 수입은 67.9%나 급등했다. 또 양파는 84.7%, 마늘은 30.2% 올랐다. 같은 기간 전철료(15.2%)와 학교급식비(10.2%)도 상승했다. 전세는 3.9% 뛰었다.

국산 담배(4.8), 수입 담배(2.9), 양파(0.8), 마늘(1.4), 전철료(3.5), 학교급식비(5.4), 전세(62) 등의 품목이 가중치 1000 중 차지하는 비중은 극히 미미하다. 내가 피우는 국산 담배의 값이 2배 가까이 올랐는데 물가통계에서 차지하는 비중은 고작 1000 중 4.8가량인 것이다.

또 통상 가구당 한 대의 차량을 보유하고 있지만 물가통계에는 경(1.4)·소형(2.7)·중형(4)·대형(5.2) 승용차를 비롯해 다목적승용차(1.4), 수입 승용차(3.3) 가격이 모두 영향을 주고 가중치도 다 다르다. 내가 타고 있는 소형차 가격이 올랐는데 수입 승용차 값이 개별소비세 등의 영향으로 더 많이 하락했다면 물가는 내려가는 식이다.

이와 같이 품목별 가중치는 481개 품목의 가격과 거래량과 연계하여 결정되기 때문에 자신의 소비지출 형태와 차이가 있을 수밖에 없다는 점을 잘 이해할 수 있도록 설명하였다.

참고자료

- 김수택 · 김영원 · 류제복 · 박진우 · 변종석 · 이기성 · 이해용 · 이흥철 · 최경호 · 한근식 · 홍기학(2002), 조사방법의 이해, 교우사.
- 박진우(2006), 통계학의 길잡이, 교우사.
- 성내경(2012), 표본조사 방법론, 자유아카데미.
- 통계교육원(2008), 통계와 정책 표준교재.
- http://www.hg-times.com/news/articleView.html?idxno=100956
- http://www.fnnews.com/news/201510041724005963
- http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1060

연구과제 또는 연습문제

■ 단원 1.

• 강의 중에 제기된 통계와 연관된 단어들을 분류기준을 세워서 분류하는 과제

• 통계학에 대한 다양한 정의를 정리하기

• 자신이 관심 갖는 통계에 포함된 통계학적 요소를 찾고 이 중에서 중요하다고 판단된 요소 탐색하기

■ 단원 2.

• 연수생들이 전체적으로 관심 갖으리라 판단되는 통계나 통계의 조사개요를 2개 정도 제시하고 모집단, 모수, 표본, 통계치, 자료수집방법을 기술하도록 한다.

■ 단원 3.

• 엑셀 자료 출력물을 주고 개체에 대한 설명, 변수명 기술, 변수별 자료의 특성을 기술하도록 한다.

• 연수생이 관심이 있는 조사표를 제시하고 개체에 대한 설명, 변수명 기술, 변수별 자료의 특성을 기술하도록 한다.

■ **단원 4.**

• **다음은 성인 남성 23명의 신장에 관한 자료이다. 이 자료를 다음과 같은 숫자들로 요약해보고 자료의 특성을 기술하라.**

ID	신장	ID	신장
1	161.2	13	161.5
2	157.4	14	166.0
3	173.0	15	177.3
4	173.0	16	173.7
5	152.9	17	169.1
6	176.5	18	170.2
7	167.5	19	157.0
8	175.4	20	160.8
9	173.2	21	161.0
10	168.8	22	174.3
11	156.5	23	180.0
12	166.1		

→

신장자료 요약
평균
표준편차
최솟값
제1사분위수(Q_1)
중앙값
제1사분위수(Q_3)
최댓값
사분위범위(IQR)

연구과제 또는 연습문제

■ **단원 5.**

• **다음은 문화체육관광부의 2013 체력실태조사 일부이다. 이 자료를 이용하여 변수별로 적절한 그래프를 작성하고 내용을 발표, 기술해보라.**

ID	나이	성별	소득	건강상태	신장	체중
1	27	여	4	2	163.0	57.3
2	40	여	3	2	153.2	43.0
3	57	여	4	3	154.3	53.9
4	37	남	2	2	175.5	76.7
5	42	남	3	3	161.2	64.3
6	55	여	4	2	153.5	62.0
7	46	여	2	2	170.9	62.0
8	34	여	3	2	166.0	57.8
9	38	여	4	1	166.0	64.8
10	53	여	3	2	156.1	58.4
11	53	여	2	2	151.5	58.4
12	51	여	3	2	156.1	57.7
13	34	남	3	3	175.7	64.6
14	58	남	3	2	180.9	89.2
15	63	남	3	2	165.1	66.9
16	61	여	1	3	154.7	53.2
17	47	여	4	1	148.8	45.7
18	52	남	3	1	170.7	84.3
19	42	남	3	2	178.0	67.9
20	21	여	1	2	167.0	47.0
21	64	여	1	2	158.1	57.2
22	48	여	4	2	156.0	55.0
23	33	남	1	2	177.4	61.9
24	46	여	4	3	156.2	58.5
25	63	남	4	2	174.9	69.4

• **소득 구분**

① 70만원 미만 / ② 70만원~ 203만원 미만 / ③ 203만원~350미만 / ④ 350만원 이상

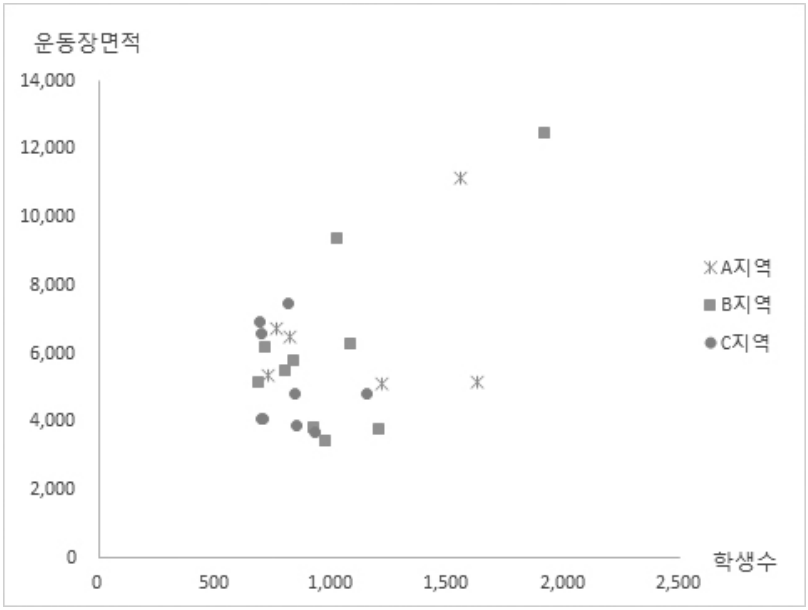
• **건강상태 구분**

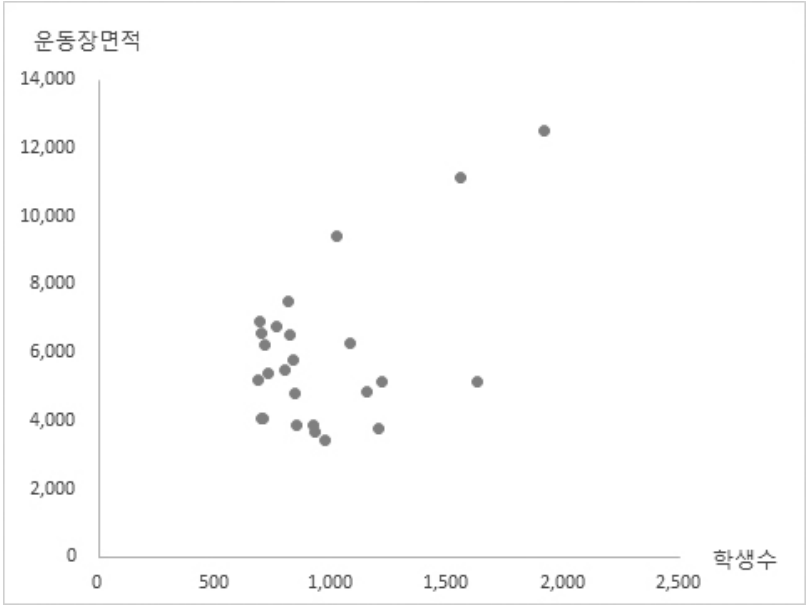
① 건강하지않다 / ② 보통이다 / ③ 건강하다

■ 단원 6.

• 다음은 중학교 학생 수와 운동장 면적에 관한 자료이다. 이 두 변수 간의 관계를 알기 위한 다음의 산점도를 보고 발견되는 정보를 기술하시오.

ID	지역	학생수	운동장 면적	ID	지역	학생수	운동장 면적
1	A	723	5,391	14	B	1,079	6,290
2	A	761	6,764	15	B	1,199	3,776
3	A	817	6,500	16	B	1,917	12,508
4	A	1,216	5,136	17	C	693	6,930
5	A	1,552	11,154	18	C	694	6,579
6	A	1,628	5,160	19	C	696	4,070
7	B	683	5,195	20	C	702	4,072
8	B	710	6,220	21	C	809	7,500
9	B	797	5,508	22	C	840	4,807
10	B	834	5,798	23	C	852	3,873
11	B	922	3,852	24	C	929	3,687
12	B	969	3,445	25	C	1,153	4,829
13	B	1,022	9,402				


운동장면적
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0
500
1,000
1,500
2,000
2,500
학생수
A지역
B지역
C지역


운동장면적
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0
500
1,000
1,500
2,000
2,500
학생수

■ **단원 7.**

• **다음의 버클리대 입학관련 문제를 통하여 Simpson's Paradox의 현상을 다시 한 번 확인해보라.**

(자료 출처: 허명회(2006), 통계적 사고, 교우사)

1) 다음의 표는 1973년 성별에 따른 버클리대 대학원 입학 자료이다. 이로 인하여 버클리 대학은 여성운동가로부터 맹렬한 비난을 받았다.

	합격	불합격	합계
남학생	1,400 (52%)	1,291 (48%)	2,691 (100%)
여학생	772 (42%)	1,063 (58%)	1,835 (100%)
전체	2,172 (48%)	2,354 (52%)	4,526 (100%)

2) 성차별 비판에 대응하여 버클리 대학이 제시한 분야별 합격률 표는 다음과 같다.

	합격	불합격	합계
남학생	1,400 (52%)	1,291 (48%)	2,691 (100%)
분야 A	512 (62%)	313 (38%)	825 (100%)
B	353 (63%)	207 (37%)	560 (100%)
C	120 (37%)	205 (63%)	325 (100%)
D	138 (33%)	279 (67%)	417 (100%)
E	53 (28%)	138 (72%)	191 (100%)
F	224 (60%)	149 (40%)	373 (100%)
여학생	772 (42%)	1,063 (58%)	1,835 (100%)
분야 A	89 (82%)	19 (18%)	108 (100%)
B	17 (68%)	8 (32%)	25 (100%)
C	202 (34%)	391 (63%)	593 (100%)
D	131 (35%)	244 (67%)	375 (100%)
E	94 (24%)	299 (76%)	393 (100%)
F	239 (70%)	102 (30%)	341 (100%)
전체	2,172 (48%)	2,354 (52%)	4,526 (100%)

■ **단원 8.**

• **다음은 정부공직자윤리위원회가 2008년 5월 7일에 고위공직자의 재산등록 현황에 근거해 공개한 고위공직자 73명에 대한 재산가액 현황 자료이다. 그래프를 이용한 방법과 사분위수, 절사평균, 표준화 점수 등의 통계치를 이용하여 이상치를 판정하여보라.**

고위공직자 재산가액 현황 (단위: 천원)

ID	재산가액	ID	재산가액	ID	재산가액	ID	재산가액	ID	재산가액
1	2,124,649	16	903,648	31	9,731,559	46	812,671	61	416,519
2	501,308	17	1,740,529	32	1,325,317	47	1,181,001	62	832,739
3	999,548	18	3,179,364	33	1,005,737	48	1,106,685	63	1,096,834
4	1,127,537	19	255,733	34	740,000	49	1,334,014	64	774,981
5	189,739	20	1,091,848	35	2,542,533	50	1,266,561	65	3,860,932
6	709,312	21	2,219,388	36	7,248,971	51	2,003,161	66	1,315,029
7	1,453,984	22	256,353	37	2,599,160	52	650,177	67	1,261,680
8	2,147,767	23	184,260	38	2,467,329	53	6,250,937	68	1,189,092
9	312,234	24	884,650	39	698,023	54	1,462,305	69	2,194,494
10	1,624,862	25	1,479,604	40	795,513	55	803,606	70	715,194
11	4,149,142	26	2,077,355	41	772,946	56	1,533,240	71	1,485,640
12	4,077,191	27	323,034	42	3,493,172	57	923,340	72	630,869
13	4,751,041	28	1,814,765	43	2,845,854	58	800,223	73	221,283
14	701,174	29	507,006	44	1,497,949	59	493,956		
15	5,932,923	30	567,876	45	5,429,133	60	890,108		

■ **단원 9.**

• **다음의 표를 이용하여 각 품목의 품목별지수와 가중치를 구하고 물가가 기준년도에 비하여 얼마나 상승 또는 하락했는지 살펴보라.**

품목	기준년도 가격	기준년도 거래량	현재가격	품목별지수	가중치
품목 A	400/kg	4kg	800/kg		
품목 B	50/개	10개	50/개		
품목 C	1000/l	20l	500/l		

• **다음의 10개 가구로 이루어진 모집단을 가정한 자료이다.**

가구번호	소득	가구주학력	가구번호	소득	가구주학력
1	2,964,170	A	6	1,500,000	B
2	514,970	B	7	603,100	B
3	325,000	A	8	3,687,000	B
4	3,016,147	A	9	4,052,590	A
5	4,536,700	A	10	1,720,850	A

• **이 모집단에서 임의로 크기 6인 표본을 추출하여 다음과 같은 자료를 얻었다. 가구번호 4, 6, 10번의 소득은 무응답이다. 소득총액을 추정하여 보고 모수와 비교하라.**

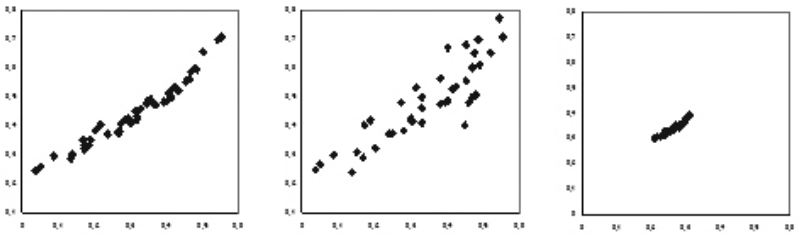
가구번호	소득	가구주학력
1	2,964,170	A
2	514,970	B
4	·	A
6	·	B
8	3,687,000	B
10	·	A

■ **단원 6.**

• 공분산을 구할 때 분모에 $(n-1)$을 사용하는 이유

표본으로부터 공분산을 계산할 때 n 대신 $n-1$ 로 나눈 이유는 다음과 같다. 각 변수의 편차(각 값에서 평균을 뺀 값)의 합은 0이 된다. 따라서 공분산을 계산할 때 $\overline{X}$와 $\overline{Y}$를 알기 때문에 전체 자료 n개 중 하나의 자료는 다른 자료로부터 계산이 되므로 n개의 자료를 사용하여 공분산을 계산하지만 실제로 정보로 느껴지는 자료의 양이 $n-1$이므로 n대신 $n-1$로 나누어 주는 것이다. 표본분산을 계산할 때 n대신 $n-1$로 나누는 것과 같은 이유이다.

• 공분산이 불완전한 척도인 이유



A B C

위의 세 가지 경우에 대하여 두 변수간의 관계의 강도와 공분산의 크기를 비교해보자. A와 B를 살펴보면 A가 더 강한 선형의 관계를 가지고 있으며(직선에 더 가까운), 이 때 공분산은 네 개의 사분면에 좀 더 퍼져 있는 B보다 A가 더 크다. C의 경우는 A를 단위만 바꾸어 표현한 것이므로 관계의 강도는 A와 C가 동일하지만, 편차의 곱의 평균인 공분산은 편차가 크게 나타나는 A가 더 크게 나타난다. 마지막으로 B와 C의 경우를 보면, A와 C가 동일한 관계의 강도를 가지고 있으므로 A와 B의 비교에서처럼 C가 B보다 관계의 강도가 크지만, 공분산은 반대로 B가 더 크게 계산된다.

이처럼 A와 B의 비교 시에는 문제가 없는데, A와 C, B와 C를 비교할 때에는 관련성과 공분산간에 일관된 방향성이 존재하지 않는다. 왜 이런 문제점이 있을까? 공분산은 두 변수의 편차 곱에 대한 평균이므로 두 변수간의 관련성은 적더라도 편차가 크면 공분산은 크게 나올 수 있다(B와 C의 비교 시 생기는 문제점). 또 똑같은 자료도 단위를 바꾸면 공분산은 달라진다. 예를 들어 키와 체중을 m과 kg으로 측정했을 때 보다 cm와 g으로 측정하면 편차가 각각 100배와 1,000배

로 커져서 똑같은 자료임에도 단위의 변화로 인해 공분산은 100,000배로 커지게 된다. 이렇듯 공분산은 각 변수의 편차 곱을 자료로 하기 때문에 변수간의 관련의 정도뿐만 아니라 편차의 크기에도 영향을 받는다.

2부

통계학으로
통계 활용하기

2부. 통계학으로 통계 활용하기

목차

통계학으로 통계 활용하기 과목의 개요

학습 목표

통계학으로 통계 활용하기는 통계학의 기본 지식을 이용하여 작성, 공표된 통계를 보다 넓고, 깊게 활용할 수 있도록 하는 학습 목표를 갖고 유용한 도구로 사용될 수 있는 통계적 기법을 가급적 직관적으로 이해하도록 하는 하위목표를 갖는다.

선수학습

"통계학으로 통계 읽기"의 내용이나 이에 준하는 기술통계학 수준

주요 용어

확률분포, 모형, 정규분포, 표준점수, 표본추출분포, 표준오차, 표본오차, 신뢰수준, 신뢰구간, 상대표준오차

학습과목의 내용요약

10장에서는 먼저 통계에 의해서 파악하려고 한 모집단을 확률 분포로 표현하는 방법을 모형의 개념과 함께 직관적으로 설명하려고 하였으나 다소 어려움을 느낄 수 있으리라 생각되기 때문에 건너뛰어도 좋다고 생각한다. 따라서 이 장을 건너뛸 경우에는 11장 도입에서 직관적인 개념에 대한 토의가 진행되어야 한다고 생각한다. 11장에서 현실에서 상식수준으로 받아들이고 있는 정규분포를 구체적으로 소개하고 이를 이용하여 표준점수와 이 점수를 활용하는 방법을 토의하였다.

12장에서는 모집단에 속한 임의의 한 개체로부터 얻은 관측값의 상대적 크기를 표준점수를 이용하여 백분율로 수량화시키는 것을 학습한다. 또한 관측값의 상대적 크기에 따라서 갖는 느낌과 통계적 가설검정 이론의 개념과 배경이 연결되도록 토의를 유도한다.

13장에서는 12장에서 학습된 내용의 사례로서 통계청에서 제공하는 "통계로 보는 자화상"에 나타나는 수치의 의미를 학습한다. 체력조사자료, 소득

조사자료, SNS 이용횟수 자료 등을 통하여 모집단을 모형화해보고 이를 바탕으로 특정개체의 관측값의 크기에 대한 평가를 하는 훈련을 한다.

14장에서는 표본추출분포의 개념을 소개한다. 먼저 이 개념의 필요성을 현실적으로 설명하여 직관적으로 이해하도록 하고 표본평균과 표본비율의 표본추출분포를 숙지하도록 한다. 15장에서는 표본추출분포의 활용으로 유한한 크기의 특정집단의 평균과 비율이 주어졌을 때 이 값들의 크기에 대한 평가를 하는 과정을 토의하고 이 기법의 활용사례를 제시한다.

16장에서는 신뢰구간을 현실에서 접한 경험을 소개하고 측정과 오차의 존재를 현실적으로 느끼게 한다. 그리고 신뢰수준과 표본에서 얻은 추정치를 해석하는 태도와 연결시켜서 신뢰구간을 직관적으로 이해할 수 있도록 이야기 방식으로 설명한다. 17장에서는 16장에서 신뢰구간의 핵심개념이 표본오차인 것을 전달하면서 표본오차가 표본크기 결정에도 중요한 역할을 하는 것과 상대표준오차에 대한 설명을 추가한다.

18장은 통계적 가설검정의 직관적 이해를 통하여 가설검정 관련 과제를 수행하여야 하는 연수생들에게 개념을 1시간에 직관적으로 이해할 수 있도록 하여 추후 가설검정을 공부하는데 도움이 되도록 한다.

제 10 장 모집단을 표현하기

10-1. 확률의 의미

학습목표

- **학교교육에서 배운 확률의 의미를 기억해보고 확률을 주관적 모형으로 보는 견해를 토의하면서 확률분포 개념을 이해한다.**

1 학교교육에서 확률의 정의

천재교육이 발행한 고등학교 확률과 통계 교과서에는 확률을 아래와 같이 정의하고 있다. (2014 발행, 2009 개정 교육과정)

"한 개의 주사위를 던지는 시행에서 나오는 눈의 수가 무엇인지 정확하게 예측할 수는 없지만 나오는 눈의 수는 1, 2, 3, 4, 5, 6 중에서 어느 하나이다. 따라서 각 면이 나올 가능성이 모두 같은 주사위라면 각 눈의 수가 나올 가능성은 모두 $\frac{1}{6}$ 이라고 할 수 있다.

이와 같이 어떤 시행에서 사건 A가 일어날 가능성을 0에서 1까지의 실수의 값으로 나타낸 것을 확률이라 하고, 기호 $P(A)$로 나타낸다. 일반적으로 어떤 시행에서 원소가 유한개인 표본공간 S에 대하여 각 근원사건이 일어날 가능성이 모두 같은 정도로 기대될 때, 사건 A가 일어날 확률 $P(A)$는 $P(A)=\frac{n(A)}{n(S)}$로 정의 하고, 이를 사건 A가 일어날 수학적 확률이라고 한다.

수학적 확률은 각 근원사건이 일어날 가능성이 모두 같은 정도로 기대된

다는 전제 아래에서 정의한다. 그런데 우리 주변의 여러 가지 현상 중에는 각 근원사건이 일어날 가능성이 모두 같은 정도로 기대된다고 생각하기 어려운 경우가 흔히 있다.

예를 들어, 어떤 농구 선수가 자유투를 던질 때, 성공할 가능성은 항상 같은 정도로 기대되지 않는다. 이와 같은 경우에는 시행을 여러 번 반복하여 얻은 상대도수를 통하여 어떤 사건이 일어날 가능성을 알아볼 수 있다.

일반적으로 동일한 시행을 n 번 반복하여 사건 A가 r_n 번 일어난다고 할 때, n 이 한없이 커짐에 따라 상대도수 $\frac{r_n}{n}$ 이 일정한 값 p 에 가까워지면 이 p 를 사건 A가 일어날 통계적 확률이라고 한다.

한편, 어떤 시행에서 사건 A가 일어날 수학적 확률이 p 일 때, 그 시행을 n 번 반복하여 사건 A가 일어나는 상대도수는 n 이 한없이 커짐에 따라 p 에 가까워짐이 알려져 있다. 따라서 통계적 확률과 수학적 확률은 같다는 것을 알 수 있다."

이러한 정의는 대학교에서 사용하는 통계학입문 수준의 강의 교재에 나와 있는 정의와 비슷한 수준의 내용으로 더 이상 추가할 내용이 없을 만큼 완성도가 높다. 그러나 고등학생의 상당수가 이러한 정의가 마음에 와 닿기는 어렵다고 본다.

❷ 상대도수밀도

3장에서 토의한 방법에 의하여 수집된 데이터로부터 4장과 5장에서 모집단의 모수 및 분포에 대한 여러 가지 정보를 얻는 기법을 배웠다. 위에서 토의한 바와 같이 학교교육과정에서 제시된 확률은 범주형 자료의 상대도수의 극한(표본크기가 커질수록 나타나는)을 이용하여 정의되었다. 이번에는 연속형 자료로 인식되는 다양성을 내포한 현상의 모습(분포)을 모형적으로 인식하는 도구인 확률분포함수에 대하여 이야기해보자. 연속형 자료로부터 얻어진 히스토그램과 확률개념의 결합을 설명하는 다음의 예제 자료로부터 토의를 시작하자. 이 예제는 김주한 외(2009)의 통계학입문에 있는 내용을 변형한 것이다.

다음의 <표 10-1>은 특정기관에서 어느 월요일 오전 8시30분부터 12시까지 방문하는 민원인 200명으로부터 수집된 입장시간 자료를 이용하여 작

성한 상대도수분포표이다. 여기서 시간 t는 업무개시 시각인 8시 30분을 0으로 하여 오전업무 종료시각인 12시를 3.5로 한다.

<표 10-1> 도착시간 상대도수분포표

도착 시간	상대도수	상대도수밀도
0.0≤ t <0.5	0.10	0.20
0.5≤ t <1.0	0.35	0.70
1.0≤ t <1.5	0.20	0.40
1.5≤ t <2.0	0.18	0.36
2.0≤ t <2.5	0.10	0.20
2.5≤ t <3.0	0.05	0.10
3.0≤ t <3.5	0.02	0.04

도수밀도 대신에 상대도수밀도를 사용하여 분포의 히스토그램을 그려보자. 한 계급의 상대도수밀도(relative frequency density)는 그 계급의 상대도수를 구간의 크기로 나누어서 구한다.

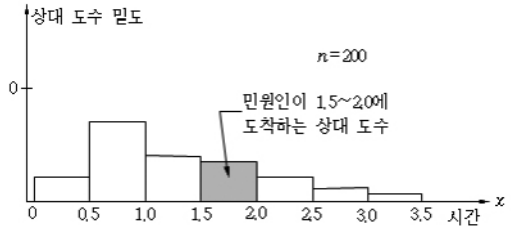
$$\text{계급의 상대도수밀도} = \frac{\text{계급의 상대도수}}{\text{계급의 크기}}$$

위의 예에서 계급의 크기는 0.5이므로 세 번째 열의 상대도수밀도는 두 번째 열의 값들을 0.5로 나누어 구한다. 그리고 일반적인 방법으로 계급값을 밑변으로 하고 상대도수밀도를 높이로 하는 직사각형을 그린다. 이렇게 얻어지는 히스토그램을 상대도수밀도 히스토그램이라고 한다.

$$\begin{aligned}\text{직사각형의 넓이} &= \text{밑변} \times \text{높이} \\ &= \text{계급의 크기} \times \text{상대도수밀도} \\ &= \text{계급의 크기} \times \frac{\text{계급의 상대도수}}{\text{계급의 크기}} \\ &= \text{계급의 상대도수}\end{aligned}$$

이와 같이 기하학적인 의미에서 어떤 계급의 직사각형의 면적은 그 계급내의 상대도수를 나타낸다. 예컨대, 1.5에서 2.0사이에 도착하는 민원인의 상대도수는 아래 그림의 빗금친 면적과 같으며, 상대도수의 합이 1이므로 상대도수밀도 히스토그램의 면적의 합은 항상 1이다.

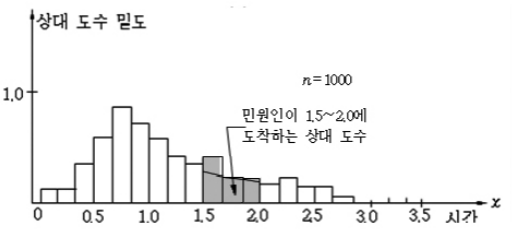
[그림 10-1]
도착시간
상대도수밀도
히스토그램 (1)



❸ 확률밀도 함수

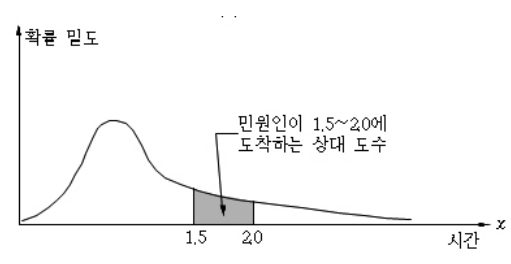
표본이 크면 계급구간의 크기를 더 작게 할 수 있고, 아래 그림에서 볼 수 있듯이 더 정교한 히스토그램을 얻게 된다. 1,000개의 관측치로부터 얻은 [그림 10-2]의 히스토그램에서 민원인이 1.5에서 2.0에 도착하는 상대도수는 3개의 직사각형의 빗금친 면적의 합과 같다.

[그림 10-2]
도착시간
상대도수밀도
히스토그램 (2)



표본 크기를 계속 늘리고, 계급을 좀 더 세분하여 히스토그램을 작성해 나가면 [그림 10-3]과 같은 이상적인 부드러운 곡선에 가까이 갈 것이라고 생각할 수 있을 것이다. 이러한 곡선을 확률밀도곡선(probability density curve)이라 하며, 이 곡선에 대한 함수를 확률밀도함수(probability density function)라고 부른다. 그림의 빗금친 부분의 면적은 민원인이 1.5~2.0 시간에 도착할 확률을 나타내고, 수학적으로는 확률밀도함수를 정적분하여 구할 수 있다.

[그림 10-3]
도착시간
확률밀도함수



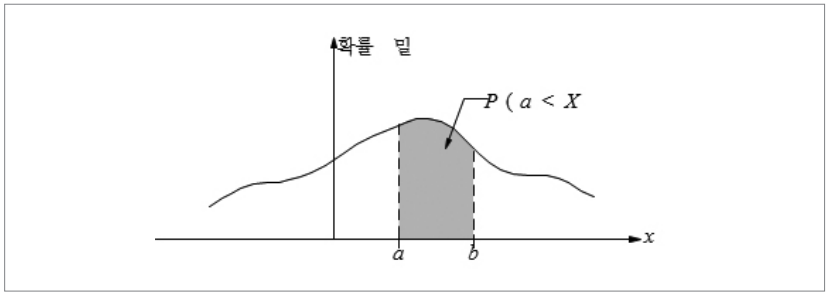
확률 밀도함수의 특징을 요약하면 다음과 같다.

(1) 상대도수밀도 히스토그램의 전체 면적이 1이므로 확률밀도곡선 아래의 면적은 항상 1이어야 한다.

(2) 확률밀도함수의 그래프는 항상 수평축 위에 있다.

(3) 확률밀도함수의 그래프는 어떤 구간에서 그래프 아래의 면적이 확률의 의미를 갖으며, 어떤 점에서의 높이는 확률의 의미를 갖지 못한다. 즉, 두 실수 a, b가 있을 때 확률변수가 a와 b 사이의 값을 취할 확률 $P(a < X < b)$는 아래의 그림과 같이 a와 b 사이의 확률밀도곡선 아래의 면적으로 주어진다.

[그림 10-4]
확률밀도함수



(4) 연속확률변수에서 확률변수가 어떤 특정한 값을 가질 확률은 0이다. 왜냐하면 모든 a에 대하여 $P(X=a) = P(a \leqq X \leqq a)$ 이므로 a에서 a 사이의 곡선 아래의 면적은 0이기 때문이다.

(5) 연속확률변수에서는 $P(X=a) = 0$이기 때문에 구간의 끝점에서 등호를 붙이거나 떼거나 영향이 없다. 따라서 다음이 성립된다.

$$P(c < X < d) = P(c < X \leqq d) = P(c \leqq X < d) = P(c \leqq X \leqq d)$$

여기서 주의할 점은 이산확률변수일 경우에는 연속확률변수의 경우와는 달리 끝점이 포함될 때와 포함되지 않을 때의 확률이 다르기 때문에 끝점의 포함 여부가 중요하다.

4 "가까워진다…"

상대도수가 가까워지는 값을 확률로 하거나 상대도수밀도가 가까워지는

것을 확률밀도함수라고 하며 그 구간의 면적을 확률로 한 두 개의 정의 모두 "가까워진 것"이라는 표현을 사용하고 있다. 바로 이 점이 확률을 사실이 아니라 하나의 관점(이를 모형이라 한다), 입장, 가정으로 받아들이는 이유이다.

예제)

아래의 자료는 특정 기관의 민원인 DB에서 표본으로 추출된 30명의 민원인으로부터 얻은 20점 만점의 만족도 조사 결과이다. 자료의 요약값은 평균이 15.82이고 분산은 0.06으로 표준편차는 0.25가 된다.

<표 10-2>
민원인 만족도

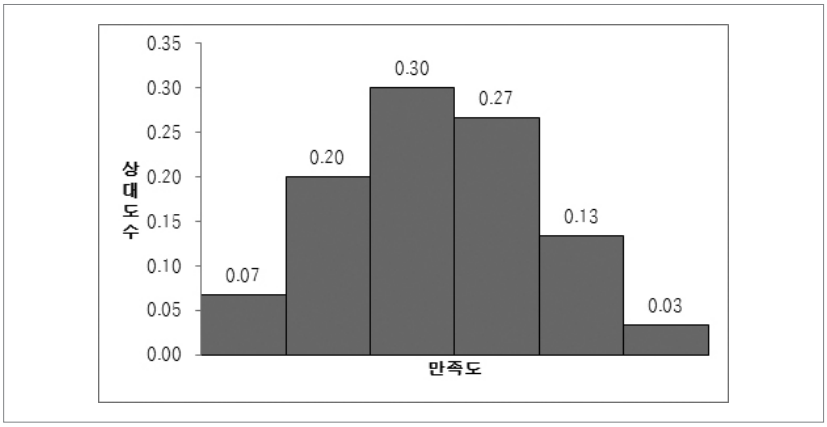
15.8	15.6	16.2	16.0	15.8	15.9
15.5	15.9	15.8	15.6	16.1	15.4
15.6	15.8	16.0	15.5	15.7	16.0
16.0	16.2	15.7	16.3	15.8	16.2
15.9	15.7	15.7	15.3	15.9	15.6

다음은 이 자료를 계급 구간의 폭을 0.2로 하여 만든 도수분포표와 이 표를 이용하여 만든 y축이 각각 상대도수, 상대도수밀도인 히스토그램이다.

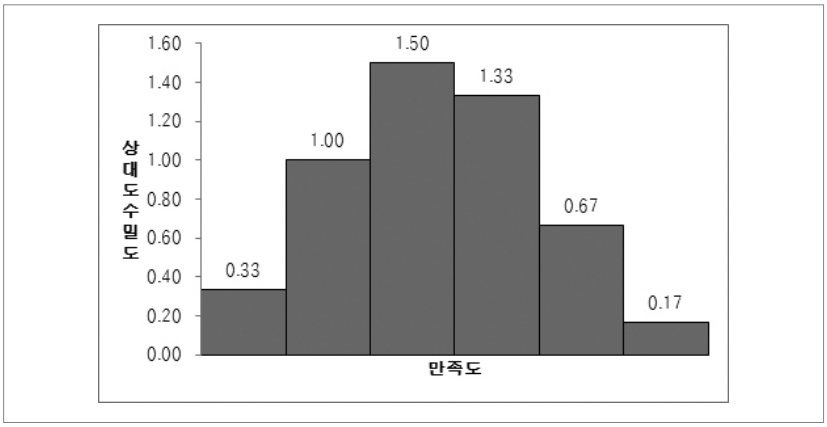
<표 10-3>
민원인 만족도
상대도수분포표

계급	도수	상대도수	상대도수밀도
15.3 - 15.5	2	2/30	(2/30) / (15.5-15.3)
15.5 - 15.7	6	6/30	(6/30) / (15.7-15.5)
15.7 - 15.9	9	9/30	(9/30) / (15.9-15.7)
15.9 - 16.1	8	8/30	(8/30) / (16.1-15.9)
16.1 - 16.3	4	4/30	(4/30) / (16.3-16.1)
16.3 - 16.5	1	1/30	(1/30) / (16.5-16.3)
계	30	1	

[그림 10-5]
민원인 만족도
상대도수 히스토그램



[그림 10-6]
민원인 만족도
상대도수밀도
히스토그램



위의 자료를 보고 다음의 질문을 생각해보자.

1) 위에 30개의 자료를 통하여 경험적으로 인식된 모집단의 특성을 어떻게 표현하겠는가?

2) 이 자료를 근거로 할 때, 민원인들의 만족도에 대하여 어떤 생각을 할 수 있겠는가?

3) 두 개의 히스토그램을 보면서 각 히스토그램의 특징을 이야기해 보자.

10-2. 확률분포 모형

학습목표

- **이산형 확률변수 자료의 막대그래프, 연속형 확률변수에 대하여 수집된 자료의 히스토그램에 적용할 수 있는 확률분포 모형들을 이해한다.**

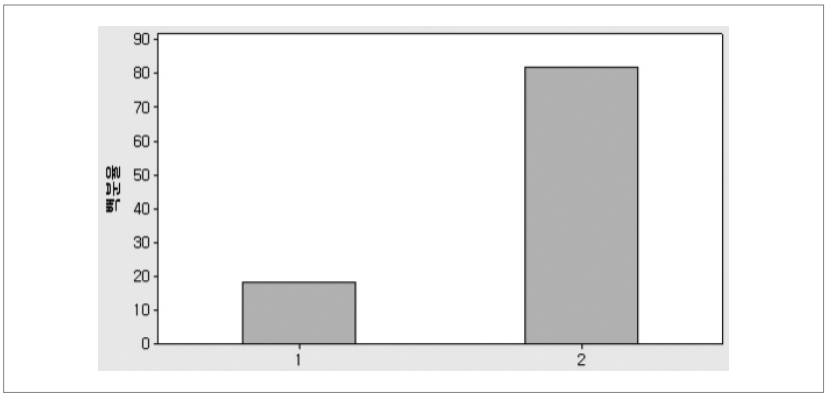
이번 절에서는 우리가 자주 듣게 되는 대표적인 확률분포를 알아보자. 확률분포는 변수의 특성에 따라 이산형(범주형) 확률분포와 연속형 확률분포로 나누어진다.

1 이산형(범주형) 확률분포

1. 베르누이분포

다음의 그래프는 관심 있는 결과가 1과 2의 두 가지인 경우의 막대그래프로 세로축은 상대도수가 백분율로 표현되었다. 1은 18.14%이고 2는 81.86%이다.

[그림 10-7] 베르누이분포 막대그래프



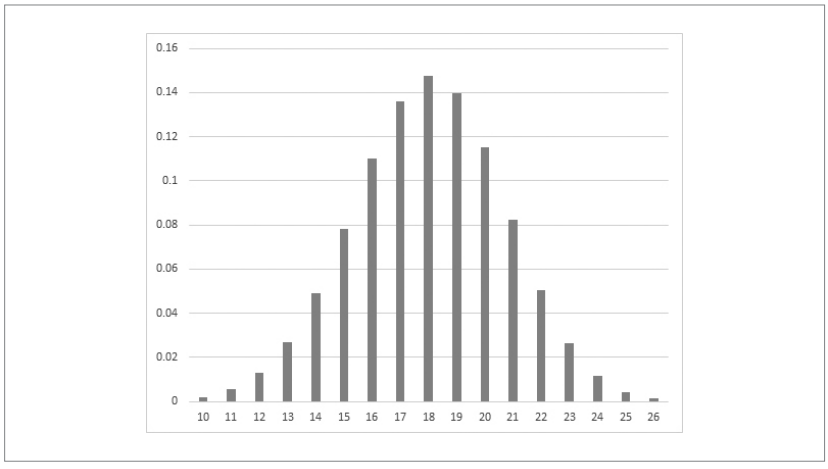
위의 그래프처럼 성별, 흡연여부, 불량여부 등과 같이 결과가 두 가지만 나타나는 경우에 관심 있는 사건(막대그래프 가로축의 '1'의 사건)을 1로, 다른 사건(막대그래프 가로축의 '2'의 사건)은 0으로 표현한다면 사건의 결과를 나타내는 확률변수는 0과 1 두 값만 갖는다. 이러한 확률변수를 X라고 하면 X는 베르누이분포를 따른다고 한다. 이 때 1이라는 사건이 일어날 확률을 $p = P(X=1)$라 하면 0의 사건이 일어날 확률은 $1-p$가 되어 확률분포함수가 표현된다.

2. 이항분포

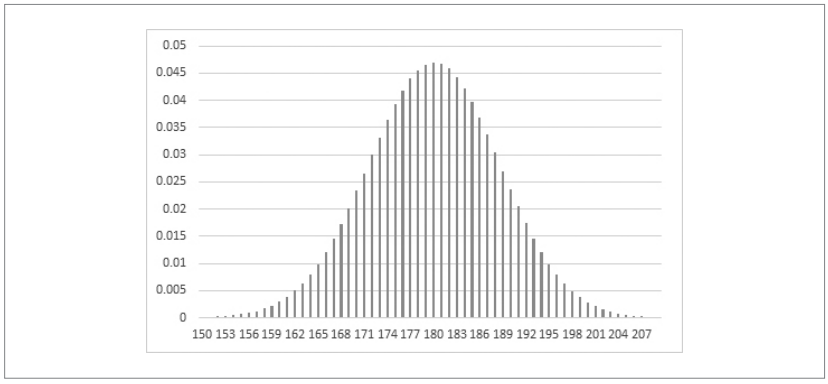
결과가 둘뿐인 시행에서 관심 있는 사건이 일어날 확률을 p라고 하자. 이러한 시행을 n번 반복할 때, 관심 있는 사건이 일어난 횟수를 나타내는 확률변수는 0에서 n 까지의 값을 가질 수 있다. 이 확률변수를 X라고 하면 X는 이항분포를 따른다고 한다. 예컨대 우리가 많이 다루었던 2014년 사회조사 보고서에는 60대의 23.9%가 자기건강에 대하여 긍정적('매우 좋다'와 '좋은 편이다')으로 평가하고 있다. 60대로 이루어진 어느 동아리 회원들 20명 중에는 몇 명이나 자기건강에 대하여 긍정적으로 평가하는가의 문제에서 긍정적으로 평가하는 사람의 수는 이항분포를 따른다고 할 수 있다.

아래의 그래프는 $p=0.6$일 때 n이 30과 300인 경우 관심사건의 발생횟수 X의 확률을 나타낸다.

[그림 10-8]
이항분포 막대그래프
($n=30,\ p=0.6$)

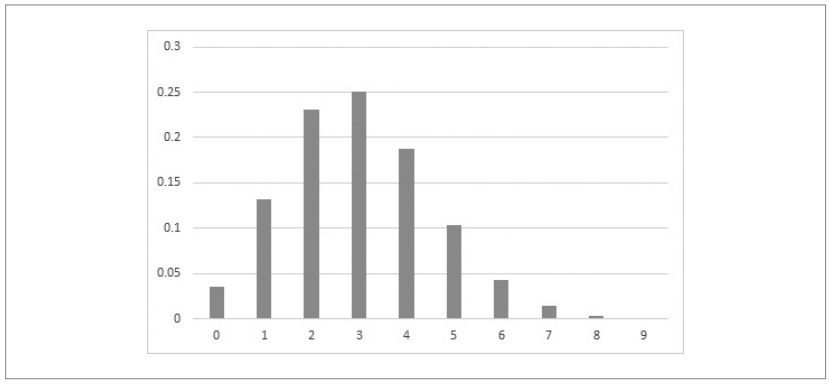


[그림 10-9]
이항분포 막대그래프
($n=300,\ p=0.6$)

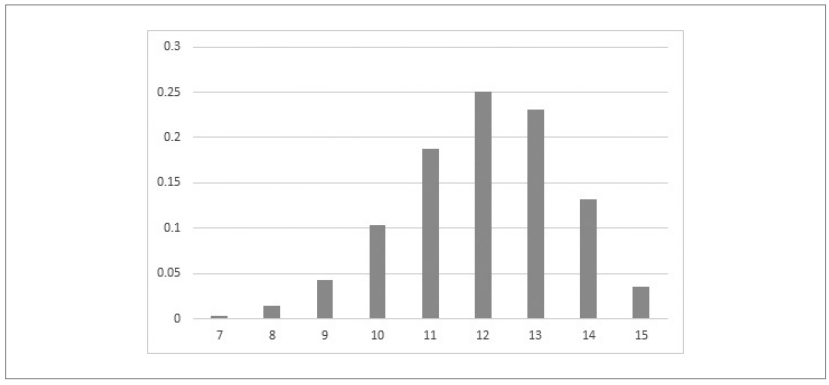


다음은 $n=15$로 동일하나 p가 0.2일 때와 0.8일 때 각 발생횟수의 확률이 어떻게 나타나는지를 보여준다.

[그림 10-10]
이항분포 막대그래프
($n=15,\ p=0.2$)



[그림 10-11]
이항분포 막대그래프
($n=15,\ p=0.8$)



3. 포아송분포

단위시간이나 단위면적, 단위공간에서 발생하는 사건의 수를 나타내는 확률변수가 다음의 조건을 만족하면 포아송분포를 따른다고 말한다.

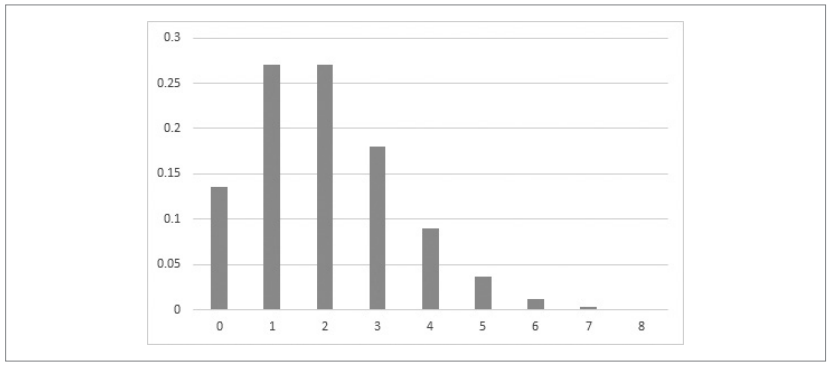
(1) 주어진 시간 혹은 공간에서 일어나는 사건의 횟수는 중복되지 않는 다른 시간 혹은 공간에서 일어나는 사건의 횟수와 서로 독립이다.

(2) 사건이 발생하는 확률은 시간 혹은 면적, 공간에 비례하며 이는 전 구간을 통해 일정하다. 하루에 10건의 사건이 발생한다면 이틀 동안에는 20건의 사건이 발생한다는 뜻이다.

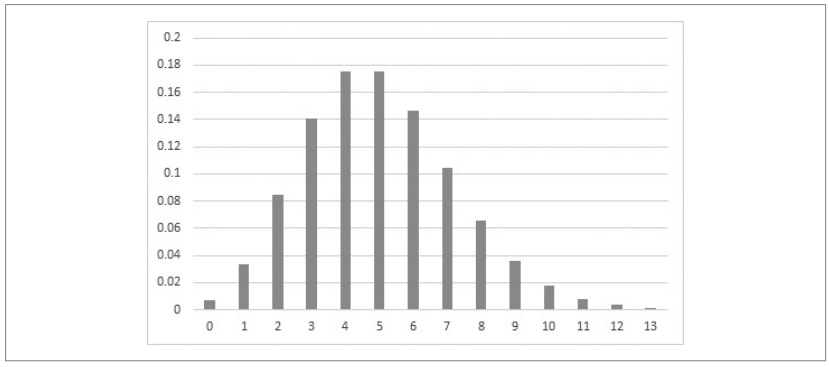
(3) 짧은 시간 혹은 작은 면적, 공간에서 사건이 두 번 이상 발생할 확률은 0에 가깝다. 하루에 10건 일어나는 사건의 경우 시간을 아주 짧게 쪼개어 1분에 일어나는 사건의 수를 생각하여 본다면 두 건 이상 일어날 확률은 무시할 수 있을 만큼 작다.

포아송분포를 따르는 예로는 산불건수나 자살자수통계 등이 있다. 중앙자살예방센터에 따르면 2014년 10대의 10만명당 자살자 수는 약 5명이다(http://www.spckorea.or.kr/index.php). 2016년에도 2014년과 동일한 상황이 계속된다면 10대의 10만명당 자살자수가 7명 이상이 될 가능성은 얼마나 되는가를 살펴보는 문제에서 10대의 10만명당 자살자수는 포아송분포를 따르는 것으로 볼 수 있다. 이 확률변수를 X라 하고 단위시간당 발생하는 사건의 수의 평균을 λ라고 하자. 다음은 λ값이 2, 5, 10인 경우 각 발생건수에 해당하는 확률을 나타내는 그래프이다.

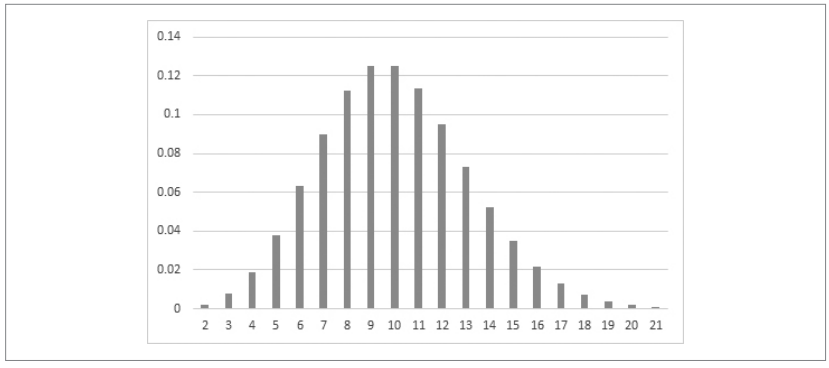
[그림 10-12]
포아송분포
막대그래프 ($\lambda = 2$)



[그림 10-13]
포아송분포
막대그래프 ($\lambda = 5$)



[그림 10-14]
포아송분포
막대그래프 ($\lambda = 10$)



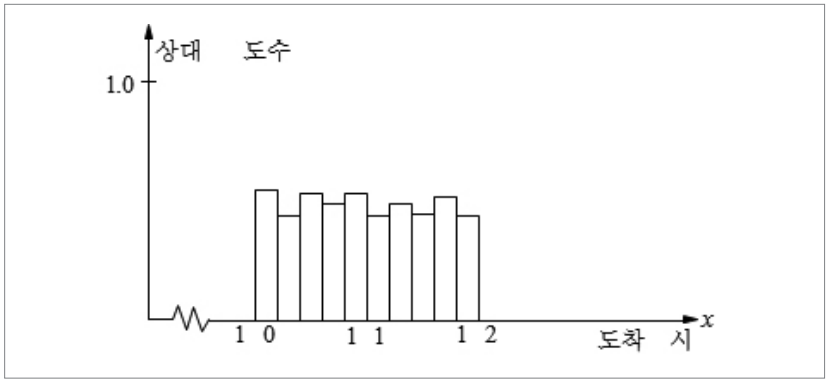
❷ 연속형 확률분포

이산형 확률변수는 가질 수 있는 값이 정해져 있고 확률분포가 가정되면 그 값을 가질 확률을 직접 구할 수 있다. 그러나 연속형 확률변수는 가질 수 있는 값 대신에 그 값을 포함하고 있는 어떤 연속적인 구간을 생각하며 확률은 그 구간의 면적으로 주어진다.

1. 균일분포

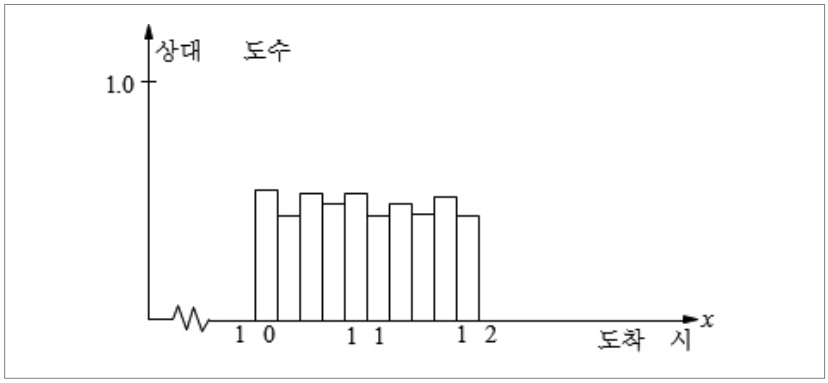
특정 기관이 민원실의 근무자 수를 결정하기 위하여 오전 10시에서 12시 사이에 도착하는 기관 방문자들의 도착 시간을 한달간 조사하였다. 총 500명의 도착시간 자료를 1/5시간(12분) 단위로 기록하여 다음과 같은 상대도수밀도 히스토그램을 얻었다. 약간의 차이는 있지만 단위시간당 도착하는 방문자의 상대도수밀도 히스토그램은 주어진 시간에서 거의 동일한 모습을 보인다. 즉 방문자가 도착하는 시간은 10시에서 12시 사이의 고르게 분포하는 것으로 나타났다.

[그림 10-15]
균일분포 히스토그램 – 도착시간



이 히스토그램을 모형화 한다면 주어진 구간 내에서 가로축이 10시와 12시 사이이고 면적이 1인 직사각형 형태를 갖는 함수를 생각할 수 있다. 이런 관점을 가졌을 때 우리는 "방문자의 도착시간을 균일분포를 따른다고 가정(모형화)한다"고 말한다. 측정값이 나타나는 구간을 (a, b)라고 하면 다음과 같은 그래프로 나타낼 수 있다.

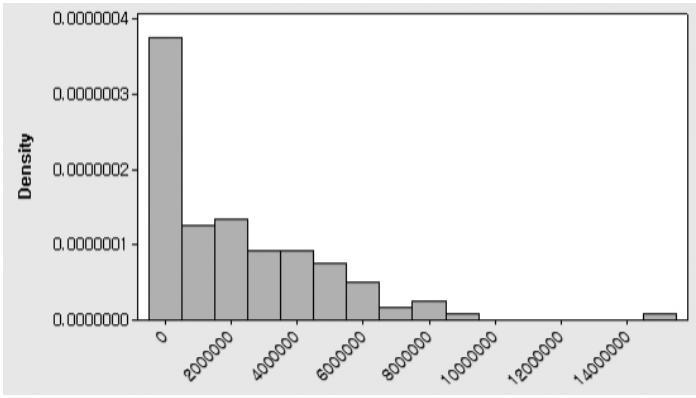
[그림 10-16]
균일분포
확률밀도함수



2. 지수분포

포아송분포를 다시 떠올려보자. 포아송분포가 단위시간에 발생하는 사건의 수를 나타낸다면, 이번에는 그 사건이 발생할 때까지 걸리는 시간으로 관점을 바꾸어보자. 사건이 발생하는 때까지 걸리는 시간은 0보다 큰 모든 값들을 가질 수 있다. 이 확률변수를 X라고 하자. X는 무한히 큰 값을 가질 수는 있지만, 값이 크면 클수록 그 값을 가질 가능성은 점점 줄어들 것이다. 단위시간당 발생하는 평균 사건수가 λ라면 사건과 사건 사이에 걸리는 시간의 평균은 $\frac{1}{\lambda}$이 되고 그래프로 표현하면 다음과 같이 나타낼 수 있다. 대표적인 예는 지출이나 민원인 응대시간 등의 자료에 적용가능하다.

[그림 10-17]
지수분포 히스토그램

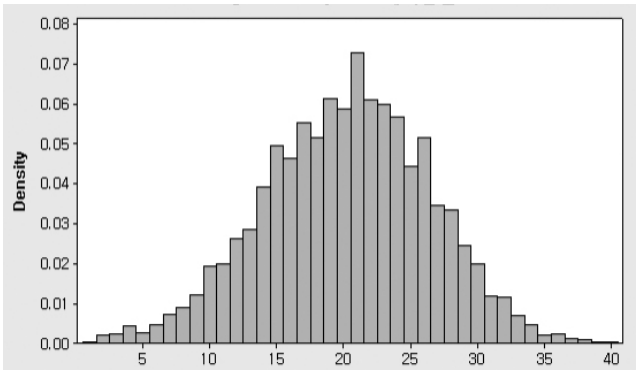


[그림 10-18]
지수분포
확률밀도함수
($\lambda = 2$)

3. 정규분포

다음의 히스토그램은 가장 높이 솟아있는 가운데를 중심으로 양끝으로 가면서 비슷한 모습으로 낮아지는 형태이다. 즉, 자료의 평균을 중심으로 좌우 대칭으로 양 끝으로 가면서 부드럽게 떨어지는 종 모양과 비슷하다.

[그림 10-19]
정규분포 히스토그램

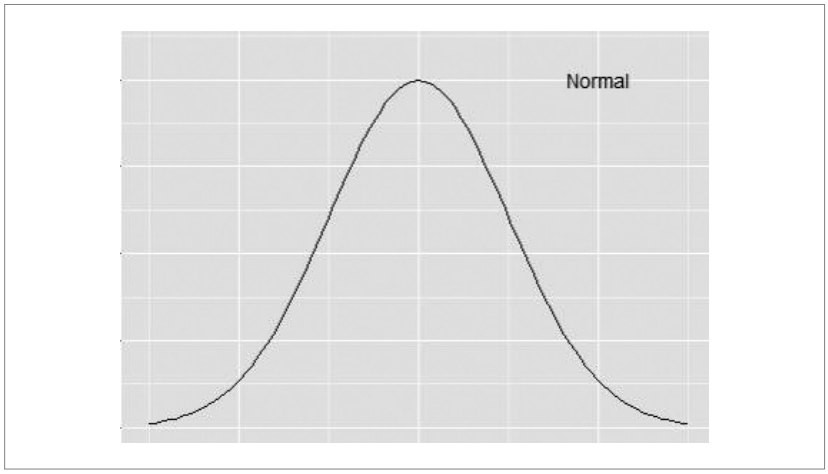


이러한 히스토그램의 모습은 주위에서 쉽게 떠올려볼 수 있다. 과자를 살 때 포장에 적혀있는 내용량을 실제로 재 본다면, 대부분의 값들이 적혀진 내용량과 비슷할 것이고 가끔 그보다 조금 작거나 큰 값들도 나타날 것이며 드물지만 아주 작거나 큰 값도 나타날 수 있을 것이다.

이러한 히스토그램을 부드러운 곡선으로 연결하여 보면 가운데가 두텁고 양쪽 끝으로 갈수록 가늘고 길어지는 아래의 그림과 같은 형태를 그려볼 수 있는데 이러한 모습으로 표현되는 확률분포를 정규분포라고 한다.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

[그림 10-20]
정규분포
확률밀도함수



10-3. 모집단의 표현과 이해

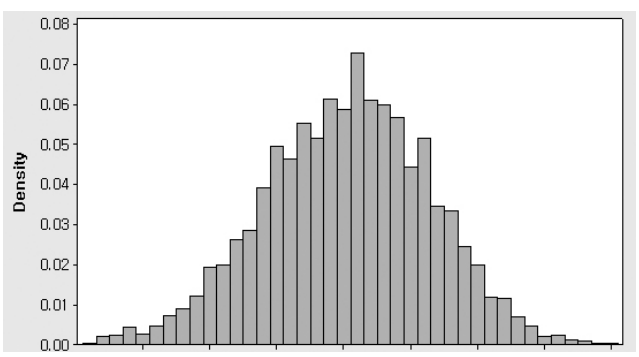
학습목표

- **국가통계에서 관심을 갖는 모집단의 특성을 확률분포모형으로 표현하는 방식을 배운다.**

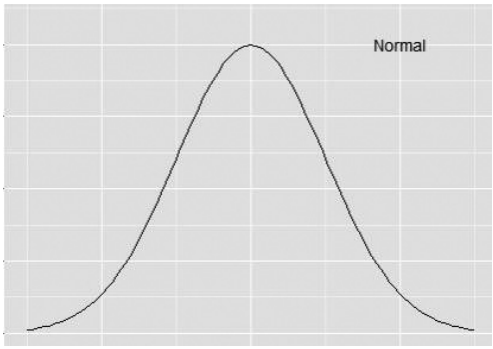
만약 한국인의 삶의 질이 어떠한지를 알고자 하여 20,000가구의 가구원들을 표본조사를 하여 얻은 한 달 평균 음주량, 자신의 건강에 대한 평가에 관한 자료를 얻었다고 하자. 우리는 이 자료로부터 우리나라 사람들의 음주량과 자신의 건강에 대한 평가에 내재된 다양성을 인식하고 표현해 보려고 할 것이다. 그러나 이러한 시도는 20,000가구의 자료와 같은 유한한 경험(표본)의 결과를 근거로 모든 국민의 일반적인 현상(모집단)을 표현하는 것이므로 주관적 관점, 입장, 가정이 요구된다.

예를 들어서 설명해보자. 어떤 모집단에서 연속형 관심변수(확률변수)에 대해 수집된 자료로부터 그려진, y축이 상대도수밀도인 히스토그램의 모습이 [그림 10-21]과 같은 형태였다고 하자.

[그림 10-21] 상대도수밀도 히스토그램



[그림 10-22] 정규분포 곡선



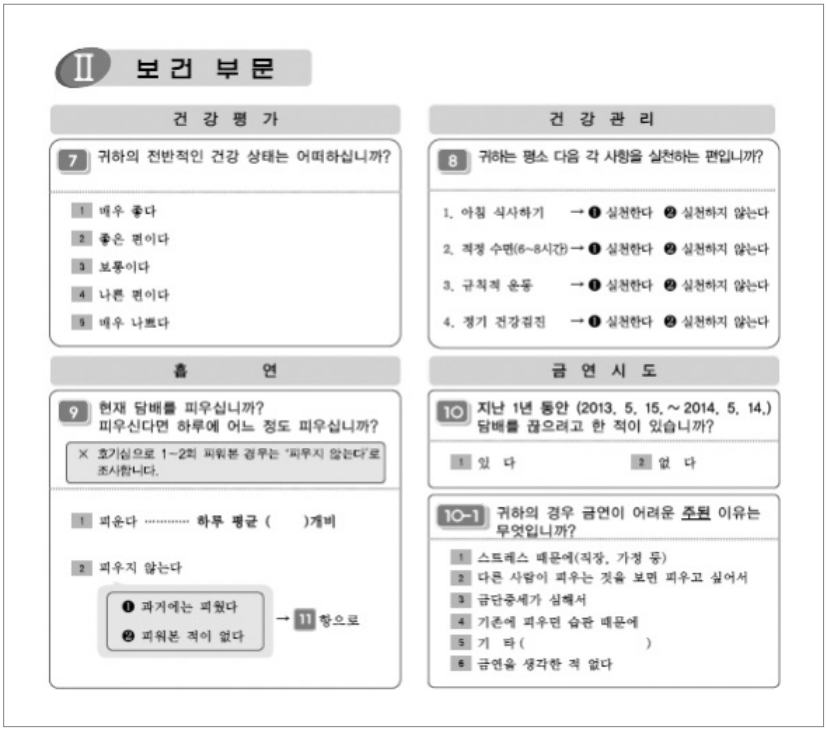
이 자료로부터 그려진 히스토그램을 A씨가 [그림 10-22]와 같은 종모양의 정규분포 곡선에 가깝다(비슷하다)고 한다면 A씨는 이 자료의 분포를 정규분포로 가정하는 입장을 갖고 있는 것이다. 한편 B씨는 A씨와 달리 정규분포곡선으로 보기에는 대칭이 안 된다고 한다면 B씨는 이 자료를 정규분포로 가정하지 않겠다는 입장을 갖고 있는 것이다.

다음의 자료들을 살펴보면서 우리가 앞 절에서 배운 분포로 자료의 모집단을 표현하고 모집단에 대한 입장을 세워보자.

1 2014 사회조사의 흡연여부

다음은 2014 사회조사의 설문문항 중 일부이다.

[그림 10-23]
사회조사 조사표
(일부)



Ⅱ 보 건 부 문

건 강 평 가

7 귀하의 전반적인 건강 상태는 어떠하십니까?

1 매우 좋다
2 좋은 편이다
3 보통이다
4 나쁜 편이다
5 매우 나쁘다

건 강 관 리

8 귀하는 평소 다음 각 사항을 실천하는 편입니까?

1. 아침 식사하기 → ❶ 실천한다 ❷ 실천하지 않는다
2. 적정 수면(6~8시간) → ❶ 실천한다 ❷ 실천하지 않는다
3. 규칙적 운동 → ❶ 실천한다 ❷ 실천하지 않는다
4. 정기 건강검진 → ❶ 실천한다 ❷ 실천하지 않는다

흡 연

9 현재 담배를 피우십니까?
피우신다면 하루에 어느 정도 피우십니까?

× 호기심으로 1~2회 피워본 경우는 '피우지 않는다'로 조사합니다.

1 피운다 ····· 하루 평균 ( )개비
2 피우지 않는다
❶ 과거에는 피웠다
❷ 피워본 적이 없다
→ 11 항으로

금 연 시 도

10 지난 1년 동안 (2013. 5. 15. ~ 2014. 5. 14.) 담배를 끊으려고 한 적이 있습니까?

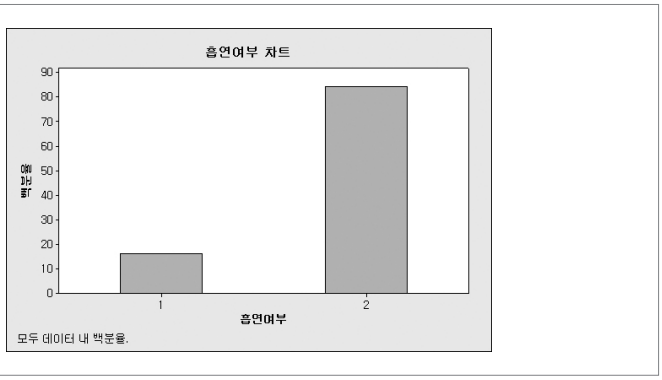
1 있 다 2 없 다

10-1 귀하의 경우 금연이 어려운 주된 이유는 무엇입니까?

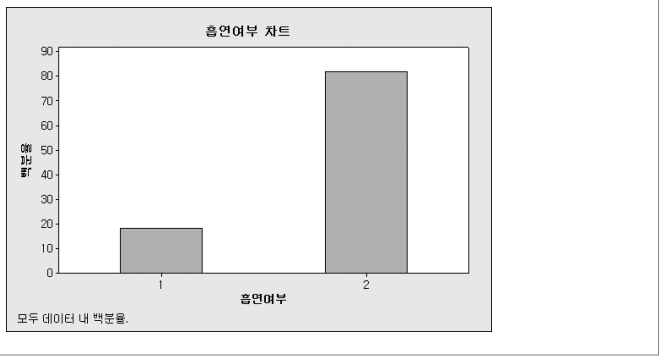
1 스트레스 때문에(직장, 가정 등)
2 다른 사람이 피우는 것을 보면 피우고 싶어서
3 금단증세가 심해서
4 기존에 피우던 습관 때문에
5 기 타 ( )
6 금연을 생각한 적 없다

위의 설문문항 중 흡연여부에 관한 9번 문항에 대하여 대전지역에서 표본으로 추출된 가구의 가구주에 대한 막대그래프를 작성하였다. 막대그래프의 가로축은 흡연여부를 나타내며, 세로축은 상대도수를 백분율로 표현한 것이다. [그림 10-24]는 자료 중에서 일부(100명)를 임의로 추출하여 나타낸 막대그래프이고 [그림 10-25]는 대전지역에서 표본으로 뽑힌 전체 가구의 가구주(1,973)에 대한 막대그래프이다.

[그림 10-24]
흡연여부 막대그래프
($n=100$)



[그림 10-25]
흡연여부 막대그래프
($n=1,973$)

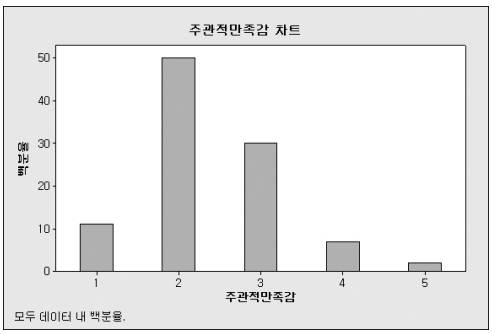


이와 같은 그래프는 두 가지 결과 즉, 흡연과 비흡연의 값만을 가지는 변수의 표현이다. 대전지역 가구의 가구주 1,973명을 모집단이라고 가정해보자. 여기서 흡연이 관심 있는 결과라면 흡연여부를 흡연인 경우 1, 비흡연인 경우 0을 가지는 확률변수 X라고 할 수 있다. 그러면 흡연여부는 주어진 자료에서 흡연율 0.1814를 p로 가지는 베르누이분포로 표현할 수 있다.

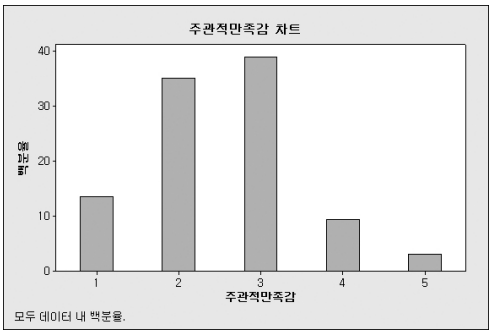
이처럼 흡연여부, 성별, 찬반, 지지여부 등과 같이 두 가지 결과로 나타나는 변수들은 베르누이분포로 표현할 수 있다.

아래의 막대그래프는 가로축은 7번 문항 건강상태에 대한 주관적 평가(① 매우 좋다, ② 좋은 편이다, ③ 보통이다, ④ 나쁜 편이다, ⑤ 매우 나쁘다), 세로축은 해당 응답에 대한 상대도수의 백분율을 나타낸 것이다. [그림 10-26]은 자료 중에서 일부(100명)를 임의로 추출하여 나타낸 막대그래프이고 [그림 10-27]은 대전지역에서 표본으로 뽑힌 가구의 가구주 전체(1,973)에 대한 막대그래프이다. [그림 10-26]과 그림 [10-27]을 비교해보면 표본크기가 $n=1,973$으로 커졌을 때 주관적 평가 ①, ②, ③, ④, ⑤의 비율이 달라지는 것을 볼 수 있다. 이때 우리는 n이 커지면서 나타나는 각 점수의 비율을 각 점수의 모비율로 추정하게 된다.

[그림 10-26]
주관적 만족감
막대그래프
($n = 100$)



[그림 10-27]
주관적 만족감
막대그래프
($n = 1{,}973$)

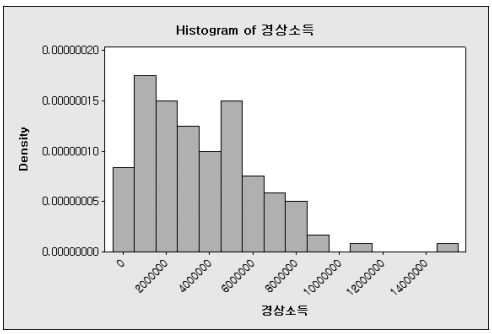


이러한 경우 임의의 개체는 5개의 문항 중 하나를 택하게 되며 응답결과는 5개 중의 하나가 된다. 그 중 '① 매우 좋다'로 평가한 가구주의 수에 대하여 관심이 있다면 이때 적용되는 분포가 이항분포이다.

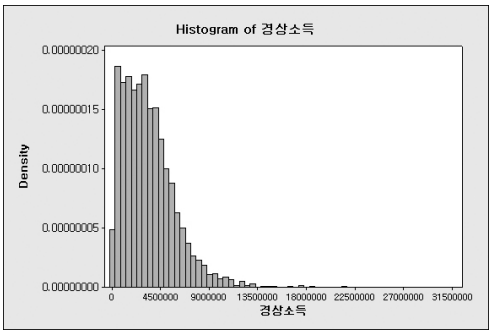
❷ 2015 가계동향조사의 경상소득

다음은 2015 가계동향조사 자료 중 4월의 경상소득에 대한 히스토그램이다. [그림 10-28]의 히스토그램은 자료 중 120가구를 임의로 추출하여 그린 것이고 [그림 10-29]는 표본으로 뽑힌 전체 6,455가구의 히스토그램이다.

[그림 10-28]
경상소득 히스토그램
($n = 120$)



[그림 10-29]
경상소득 히스토그램
($n = 6,455$)



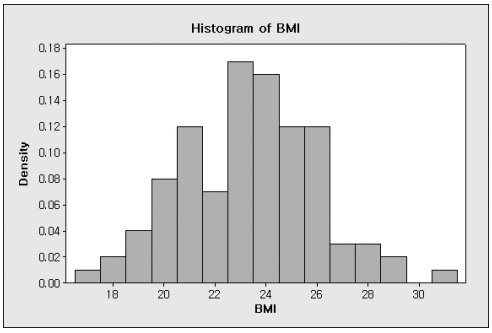
위의 두 히스토그램은 왼쪽의 작은 값들을 가질 가능성이 높으며 오른쪽의 큰 값으로 갈수록 작아지는 형태를 그리고 있다. 이와 같이 왼쪽이 높게 솟아있으며 오른쪽으로 길게 늘어지는 곡선의 형태는 앞절의 지수분포와 유사하다고 할 수 있다.

따라서 이 히스토그램을 지수분포와 유사하게 본다면 이 사람은 경상소득은 지수분포를 따른다고 가정하게 된다.

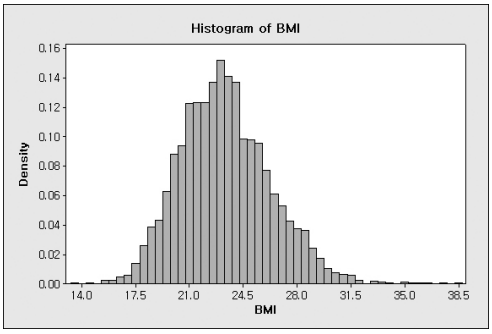
❸ 2013 국민체력실태조사의 BMI

다음은 2013 국민체력실태조사에서 BMI 값에 대한 히스토그램이다. [그림 10-30]은 이 중 임의로 100명을 추출하여 그린 히스토그램이고 [그림 10-31]은 표본으로 뽑힌 전체 3,674명의 히스토그램이다.

[그림 10-30]
BMI 히스토그램
($n = 100$)



[그림 10-31]
BMI 히스토그램
($n = 3,674$)



위의 히스토그램을 보면서 가운데가 가장 높이 솟아있으며 양쪽 끝으로 갈수록 낮아지는 종모양의 곡선을 떠올릴 수 있다. 히스토그램은 그 끝이 무한대인 모든 실수구간이 아닌 좀 더 좁은 구간 안에서 그려지긴 하지만 이러한 곡선의 형태는 앞 절에서 배운 정규분포의 확률밀도함수와 비슷하다고 할 수 있다. 이럴 경우 BMI는 정규분포를 따른다고 가정할 수 있다.

정규분포로 표현할 수 있는 예는 많은데 키나 체중, 연간 강우량, 부품이나 용량의 오차 등이 정규분포로 잘 설명되는 예들이다.

참고자료

- 김주한 · 김홍기 · 박래현 · 박석윤 · 배종호 · 이낙영 · 이석훈 · 이민구 · 이주호(2009), 통계학 입문, 정익사.
- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- 이준열 · 최부림 · 김동재 · 한대희 · 전용주 · 장희숙 · 조석연 · 조성철 · 황선미 · 박성준(2014), 고등학교 확률과 통계, 천재교육.
- 문화체육관광부(2013), 체력실태조사.
- 통계청(2014), 사회조사.
- 통계청(2014), 사회조사 조사표.
- 통계청(2015), 가계동향조사.
- 중앙자살예방센터, http://www.spckorea.or.kr/index.php.

제 11 장

표준점수 이해하기

11-1. 표준점수 사례 검토

학습목표

- **표준점수를 가장 중요하게 사용하는 사례로 수능성적표를 자세히 검토하여 표준점수 개념을 이해하는 것을 목표로 한다.**

1 표준점수 활용분야(수능성적)

다음은 한국교육과정평가원 홈페이지를 통해 제공하는 "2016학년도 대학수학능력시험 Q&A 자료집"에서 제시한 2016학년도 수능 성적표의 예시 양식이다. 수능 성적표에는 수험번호와 성명, 응시한 영역과 유형, 그리고 표준점수, 백분위 등급이 표시된다.

(http://suneung.re.kr/sub/info.do?m=0401&s=suneung)

[그림 11-1]
수능성적표(예시)

〈2016학년도 대학수학능력시험 성적통지표(예시)〉

수험번호	성 명	생년월일	성별	출신고교 (반 또는 졸업년도)		
12345678	홍 길 동	97.09.05.	남	한국고등학교 (9)		
구 분	국어 영역	수학 영역	영어 영역	사회탐구 영역		제2외국어/한문 영역
	B형	A형		생활과 윤리	사회 · 문화	일본어 I
표준점수	131	137	141	53	64	69
백 분 위	93	95	97	75	93	95
등 급	2	2	1	4	2	2

2015. 12. 2.

한국교육과정평가원장

1. 수능성적표에 나타난 주요용어에 관한 설명을 살펴보자.

[주요 용어 설명]

■ 표준점수 : 원점수(정답한 문항에 부여된 배점을 합한 점수)의 분포를 영역 또는 선택과목별로 정해진 평균과 표준편차를 갖도록 변환한 분포상에서 수험생이 획득한 원점수가 어느 위치에 해당하는가를 나타낸 점수

$$\text{표준점수} = 20(\text{또는 } 10) \times \left(\frac{\text{수험생의 원점수} - \text{수험생이 속한 집단의 원점수 평균}}{\text{수험생이 속한 집단의 원점수 표준편차}} \right) + 100(\text{또는 } 50)$$

- 국어, 수학, 영어, 직업탐구 영역의 표준점수는 평균 100, 표준편차 20으로 함.
- 사회/과학탐구 영역과 제2외국어/한문 영역의 표준점수는 과목당 평균 50, 표준편차 10으로 함.
- 표준점수는 소수 첫째 자리에서 반올림한 정수로 표기함.

■ 백분위 : 수험생이 받은 표준점수보다 낮은 표준점수를 받은 수험생 집단의 비율을 백분율로 나타낸 점수.

- 백분위는 정수로 표기된 표준점수에 근거하여 산출되며, 소수 첫째 자리에서 반올림한 정수로 표기함.

■ 등급 : 점수로 표기된 표준점수의 분포를 9구간으로 나누어 결정함.

- 등급 구분 점수에 놓여 있는 동점자에게는 해당되는 등급 중 상위 등급을 부여함.

(1) 표준 점수의 계산

다시 정리하면, 먼저 영역/과목별로 다음의 공식에 의하여 A를 구한다.

$$A = \frac{(\text{수험생의 원점수} - \text{수험생이 속한 집단의 평균})}{\text{수험생이 속한 집단의 원점수 표준편차}}$$

수능에서는 이 A를 다시 조정하여 표준점수라고 표현하지만 통계학에서는 이 A를 표준점수 또는 교재에 따라서는 Z점수라고 한다. 공식에서 알 수 있듯이 표준점수와 집단의 평균, 집단의 표준편차를 알면 그 사람의 원점수도 알게 된다.

$$\text{원점수} = \text{평균} + A \times \text{표준편차}$$

우리나라 수능에서는 A에 해당과목에 대한 평균과 표준편차를 조정하여 학생들의 조정된 점수(이것을 수능에서는 표준점수라고 한다)를 수능성적표의 표준점수로 사용한다.

$$\text{수능의표준점수} = A \times \text{해당과목의조정평균편차} + \text{해당과목의조정평균}$$

수능에서는 다음과 같이 조정한다.

- 국어/영어/수학 과목은 평균 100, 표준편차 20을 적용

  $\text{표준점수} = A \times 20 + 100$

- 탐구/제2외국어 과목은 평균 50, 표준편차 10을 적용

  $\text{표준점수} = A \times 10 + 50$

수능의 원점수와 표준점수의 차이를 다시 한 번 설명하면, 원점수는 시험 총점과 대비하여 몇 점을 받았는가의 정보만 나타내므로 상대적인 비교나 개인의 점수에 대한 영역 간의 비교는 어렵다. 그러나 표준점수는 집단의 특성을 고려하여 각 개인의 원점수를 의미 있는 해석이 가능하도록 바꾼 점수로 각 영역의 평균과 표준편차가 어떻게 나왔느냐를 고려하여 원점수를 평균과 표준편차가 일정한 값을 갖도록 조정한 점수이다. 즉, 표준점수는 난이도 차이를 조정한 점수로 이해할 수 있다.

앞에서 언급했지만 다시 한 번 강조하는데 통계학에서는 표준점수를 위의 설명에서 사용된 A값이라고 정의하는 교재가 많고 이를 Z값, Z점수라고도 부른다.

11-2. 정규분포

학습목표

• 정규분포에 대한 전반적인 지식을 습득하는 것을 목표로 한다.

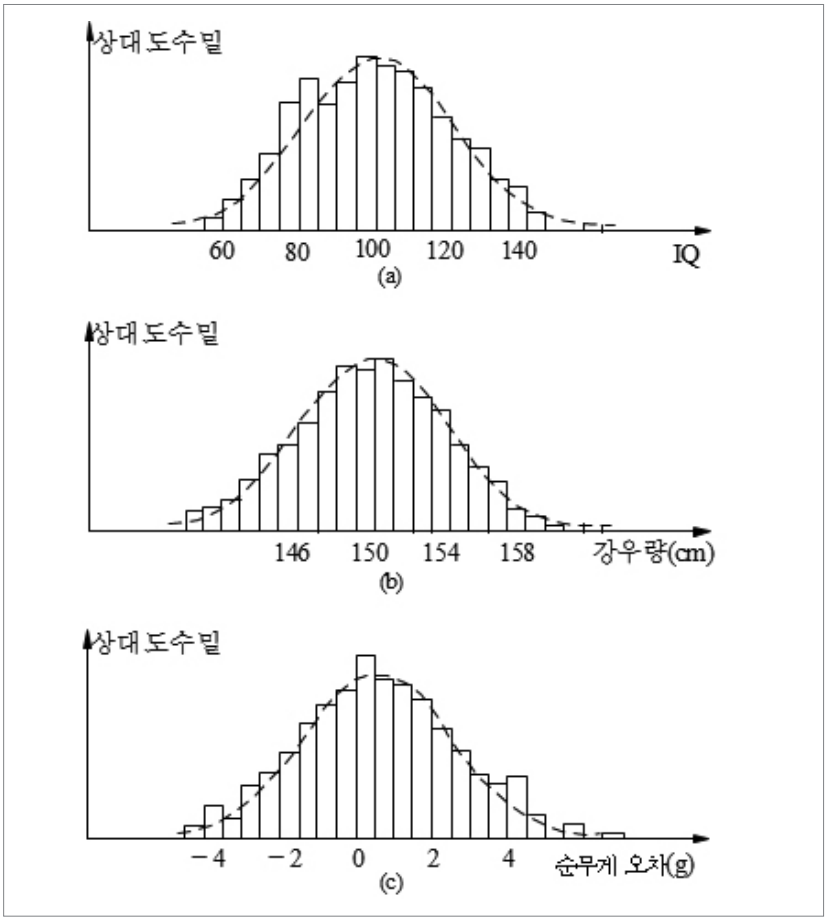
1 정규분포

10장에서 토의한 바와 같이 상대도수밀도 히스토그램의 모습이 좌우대칭의 종모양인 경우 이를 정규곡선(normal curve)이라고 하며, 그 확률분포를 정규분포(normal distribution)를 따른다고 말한다고 하였다. 정규분포는 수학자 F.Gauss(1777~1855)가 각종의 물리학 실험을 수행할 때 수반되는 계측 오차를 가지고 y축이 상대도수밀도인 히스토그램을 작성하면서 가우스 분포(Gauss distribution)라는 확률분포를 제시한 이래, 많은 학문 분야에서 이 분포를 기본 확률모형 또는 근사적인 확률모형으로 채택하였다. 특히 통계학의 초기 발전 단계에 있어서는 모든 자료의 히스토그램이 이 분포의 곡선 형태와 비슷하지 않으면 자료수집과정이 잘못된 것이라고 믿었던 적도 있었다. 실제적으로 많은 계측 오차 현상들의 확률분포는 종 모양의 부드러운 곡선으로 표현할 수 있다.

대표적으로 대학생들의 지능 지수나 연간 강우량, 특정제품의 순무게 오차가 정규분포에 의해서 잘 설명되는 예들이다.

10장에서 토의한 내용을 다시 한 번 이야기 해 보자. 아래 3개의 그림을 보면 상대도수밀도 히스토그램을 부드러운 곡선으로 연결하면 양쪽 끝이 가늘고 길며, 가운데 부분이 두터운 좌우 대칭인 부드러운 곡선으로 근사됨을 볼 수 있다.

[그림 11-2]
정규분포
상대도수밀도
히스토그램 예

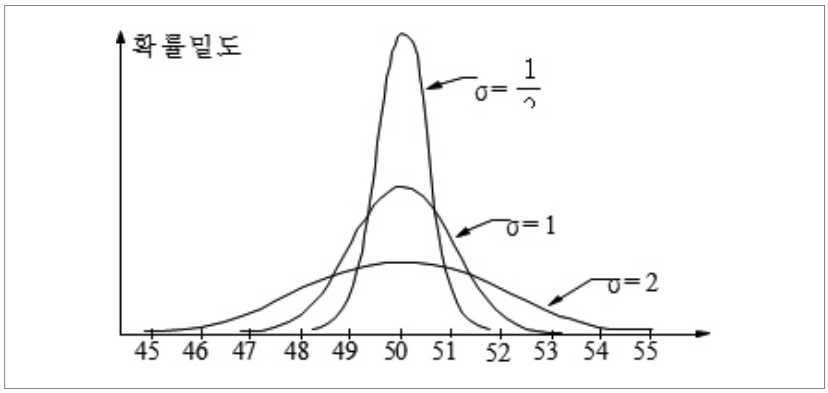


가우스(Gauss)가 제시한 수학적 모형으로서 정규분포의 특징은 다음과 같다.

- 평균과 표준편차에 의해 그 형태가 결정되는데, 평균(μ)은 실수값을 갖고 표준편차(σ)는 양수값을 갖는다.
- 정규분포의 함수값은 중심에서 최대이며, 평균을 중심으로 대칭이다.
- 정규곡선의 양쪽 끝인 꼬리 부분은 중심에서 양쪽으로 무한히 연장되며, 수평축에 한없이 가깝게 갈 수는 있지만, 아무리 멀리 가더라도 수평축에 닿지는 않는다. 즉, 수평축을 점근선으로 갖는다.
- 평균(μ)은 곡선의 중심 위치를 결정하고, 표준편차(σ)는 그 곡선의 퍼진 정도를 결정한다.

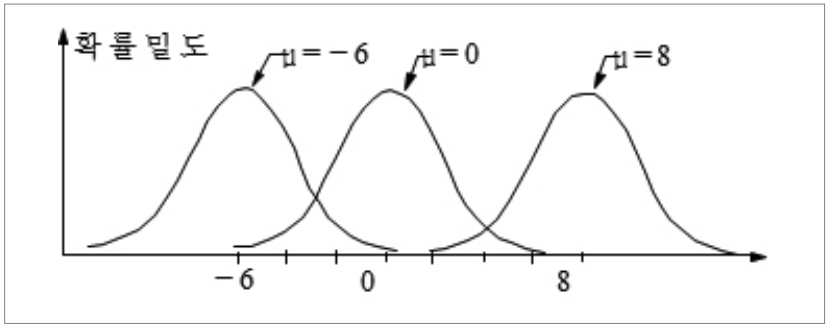
아래 그림의 곡선들은 평균이 모두 50으로 같지만, σ=1/2, σ=1 및 σ=2의 각기 다른 표준편차를 갖는다. 표준편차가 클수록 곡선이 평평함을 알 수 있다. 예컨대, σ=1/2인 곡선은 σ=2인 곡선보다 더 뾰족하다.

[그림 11-3]
μ=50이고
σ=1/2, 1, 2인
정규분포



다음 그림은 표준편차는 σ=2로 모두 같지만 평균이 μ=-6, μ=0, μ=8로 각기 다른 정규곡선들을 보여준다. σ가 서로 같으므로 곡선들은 퍼짐의 정도, 즉 모양은 같지만 μ=8인 곡선은 μ=-6인 곡선보다 오른쪽에 위치해 있다.

[그림 11-4]
σ=2이고
μ=−6, 0, 8인
정규분포



이렇게 표현되는 그래프들에 대응되는 함수식은 평균(μ)과 표준편차(σ)를 사용하여 다음과 같이 나타낼 수 있다.

평균 μ와 유한한 분산 σ^2을 갖는 정규분포의 확률밀도함수는 아래와 같다.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\qquad -\infty<x<\infty$$

여기에서 π=3.14159…이며, e =2.71828…이다. 이러한 정규분포를 나타내는 기호로 일반적으로 $N(\mu, \sigma^2)$을 쓴다.

X가 정규분포 $N(\mu, \sigma^2)$을 따를 때 평균 μ로부터 κ배의 표준편차 이내에 있을 확률은 다음과 같다.

<표 11-1> 정규분포 구간에 따른 확률

κ	구간	구간의 확률
1	$\mu-1\sigma$ 에서 $\mu+1\sigma$	0.6826
2	$\mu-2\sigma$ 에서 $\mu+2\sigma$	0.9544
3	$\mu-3\sigma$ 에서 $\mu+3\sigma$	0.9974

예를 통하여 이것을 활용해 보자. 특정 시험을 주관한 기관이 학생들의 성적이 평균 70점에 표준편차가 10인 정규분포를 따르는 것으로 나타났다고 발표하였다. 그동안 여러 사람들을 접해본 결과 많은 사람들은 이 정보로부터 다음과 같은 것들을 추측하는 것을 발견하였다.

- 70점 근처의 학생들이 많다.
- 70점을 중심으로 학생들의 성적이 대칭적이다. 그러니까 70점 이상인 학생들이 50%정도 되고 70점 이하인 학생들은 50%정도 된다.
- 대부분의 학생들이 70점에서 1배 표준편차인 60점과 80점 사이에 있다.

그런데 우리가 위의 확률에 대한 표를 참고한다면 우리는 다음과 같은 더 많은 정보를 얻을 수 있다.

- 70점 근처의 학생들이 많고 70점에서 멀어질수록 적다. 70점을 중심으로 1배 표준편차인 60점과 80점 사이에 전체 학생의 68.26%가 있고, 2배 표준편차인 50점과 90점 사이에 95.44%가 있다.
- 90점 넘는 아주 우수한 학생들은 전체의 2.28%가 있다.

(2배 표준편차인 $(\mu-2\sigma, \mu+2\sigma)$사이에 95.44%가 있으므로 그 구간을 벗어난 $(-\infty, \mu-2\sigma)$와 $(\mu+2\sigma, \infty)$사이에 4.56%가 있다. 정규분포는 좌우대칭이므로 $(\mu+2\sigma, \infty)$, 즉 90점 이상인 학생은 4.56%를 반으로 나눈 2.28%가 된다.)

- 또 50점 이하의 학업에 어려움을 가지고 있는 학생들도 2.28%가 있다.

(위와 동일하게 $(-\infty, \mu-2\sigma)$, 즉 50점 이하 역시 2.28%가 된다.)

그렇다면 다음의 질문은 어떻게 답할 수 있을지 생각해보자.

- 상위 1%는 몇 점 이상일까?

- 만약 80점 이상에게 장학금을 준다면 얼마나 많은 학생들이 장학금을 받게 될까?
- 50점 이하를 받은 학생들은 얼마나 못했다고 할 수 있는가?
- 그러면 55점에서 80점까지 받은 학생들은 얼마나 될까?

위의 질문은 k배 표준편차에 관한 확률을 나타낸 표만으로는 알 수 없다. 이제 표준점수를 활용해야 한다.

11-3. 표준점수

학습목표

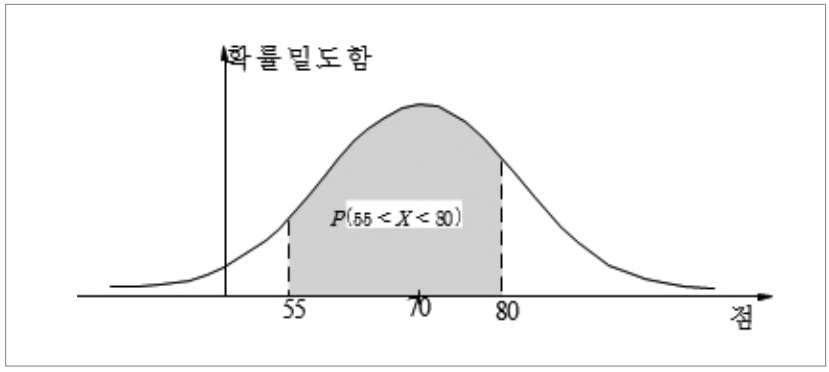
- **정규분포를 가정한 모집단에서 개체의 상대적 위치를 찾아내는 방법을 습득한다.**

1 표준점수

앞 절의 마지막 질문 "55점에서 80점까지 받은 학생들은 얼마나 될까?"에서부터 이야기를 시작해보자.

정규분포는 연속확률분포이므로 확률은 정규곡선 밑의 면적으로 주어진다. 그러므로 정규분포를 갖는 확률변수 X가 두 수 a와 b사이에 있을 확률은 곡선 아래의 면적을 구하여야 한다. 주어진 질문에서 평균 70, 표준편차 10인 정규분포에서 55와 80 사이의 확률은 아래 그림에서 빗금친 부분의 면적과 같다.

[그림 11-5] 정규분포 확률밀도함수 (μ=70, σ=10)



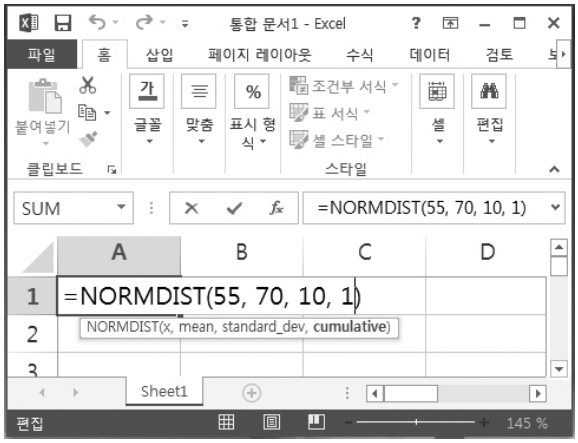
그러나 곡선의 모양에서 보는 바와 같이 우리는 그 면적을 기하학적으로 간단히 구할 수 없다. 대신 정적분을 이용해야 하지만 이는 매우 복잡하다. 면적을 구하기위한 수학적인 적분이 공식으로는 불가능하기 때문에 컴퓨터를 통해서 근사적으로 계산할 수밖에 없는데, 이전에는 컴퓨터를 통하여 이러한 확률을 계산하는 작업을 하는 것은 쉬운 일이 아니었다. 그래서 능력 있는 사람들이 고민을 시작하였고, 그 결과 정규분포는 평균에 따라 분포가 자리하는 중심이 결정되고 표준편차에 비례해서 분포의 폭이 커진다는 특성으로부터 한 가지 정규분포의 확률만 알고 있으면 어떤 정규분포의 확률도 계산할 수 있다는 사실을 알게 되었다. 그 사실을 이용

하여 그들은 평균이 0이고 분산이 1인 정규분포에 대해 X가 –3보다 큰 임의의 x보다 작을 확률을 계산하여 표로 만들어 놓았는데 이것이 고등학교 교과서에서 본 정규분포표라는 것이다.

지금은 엑셀을 통하여 어떤 정규분포의 확률이든 바로 구할 수 있어서 정규분포표의 사용이 많지는 않다.

엑셀의 $=normdist(x, mean, sd, cumulative)$라는 함수를 이용하면 쉽게 구할 수 있다. 예를 들어 엑셀에서 아래와 같이 입력하면 평균 70, 표준편차 10인 정규분포에서 55점 이하일 확률, 즉 $P(X<55)$를 구해준다.

[그림 11-6]
엑셀 정규분포 확률 계산 (1)

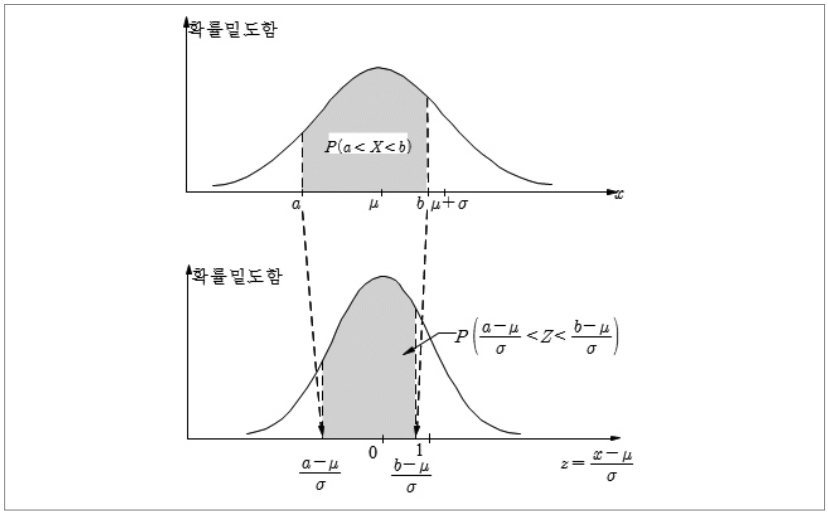


그리고 평균이 0이고 분산이 1인 정규분포를 표준정규분포(standard normal distribution)라고 부르고 통상 Z라고 표현하는데 임의의 정규분포를 따르는 변수(확률변수)와의 관계는 다음과 같다.

$$Z=\frac{X-\mu}{\sigma}$$

이 식을 다시 설명하면 평균이 70이고 표준편차가 10인 정규분포를 따른다고 가정한 집단에 속한 임의의 한 개체의 측정값 X에서 평균인 70을 빼고 표준편차인 10으로 나눈 값은 평균이 0이고 표준편차가 1인 형태의 정규분포, 즉 표준정규분포를 따르게 된다. 그리고 이러한 과정을 "평균이 70이고 표준편차가 10인 정규분포를 따른다고 가정한 집단에 속한 임의의 한 개체를 표준화(standardization)했다"고 한다. 또는 "평균이 70이고 표준편차가 10인 정규분포를 표준화했다"고 한다.

[그림 11-7]
정규분포의 표준화



표준화 기법이란 이렇게 개체의 관찰값과 평균의 차이를 표준편차로 나눈 값을 이 개체의 표준화 값(표준점수)으로 정의하는 개념이다. 따라서 표준점수는 본래의 관찰값이 신장에 대한 자료이든 체중에 관한 자료이든 무엇이 되었든지 평균이 0이고 표준편차가 1이 되는 값들이 된다.

대부분의 통계학 책에 부록으로 붙어있는 표준정규분포표가 이렇게 구한 것이다. 예를 든다면 평균이 70, 표준편차가 10인 정규분포를 따르는 확률변수 X가 55이상일 확률을 구하라는 말은 표준정규분포에서 표준점수 Z가 55의 표준점수인 $\frac{55-70}{10}=-1.5$ 이상인 확률을 구하는 것이 되고, 그 값은 정규분포표를 이용하거나 엑셀을 사용하여 구할 수 있다.

따라서 다시 질문으로 돌아가면 학생들의 점수가 평균 70, 표준편차 10인 정규분포를 따를 때, 55점에서 80점까지 받은 학생이 얼마나 되는가의 질문은 표준점수 Z가 $\frac{55-70}{10}=-1.5$에서 $\frac{80-70}{10}=1$ 사이에 있을 확률이 얼마나 되는가를 이용하여 구할 수 있고 그 값은 다음과 같다.

$$P(55<X<80)=P(-1.5<Z<1)=P(Z<1)-P(Z<-1.5)$$
$$=0.8413-0.0668=0.7745$$

즉 55점에서 80점까지 받은 학생은 전체의 77.45%이다.

또 다른 질문인 80점 이상을 받은 학생에게 장학금을 준다면 얼마나 많은 학생이 장학금을 받을 수 있는가의 질문도 이제 답할 수 있을 것이다. 아래와 같이 계산해 보면 15.87%가 장학금을 받을 수 있다는 것을 알게 된다.

$$P(X>80)=P(Z>1)=1-P(Z<1)=1-0.8413=0.1587$$

앞 절에서 정규분포가 k배 표준편차 내에 있을 확률을 다음과 같이 언급한 바 있다.

$$P(\mu-1\sigma<X<\mu+1\sigma)=0.6826$$

$$P(\mu-2\sigma<X<\mu+2\sigma)=0.9544$$

$$P(\mu-3\sigma<X<\mu+3\sigma)=0.9974$$

따라서 Z는 다음과 같은 특징을 갖게 된다.

$$P(-1<Z<1)=0.6826$$

$$P(-2<Z<2)=0.9544$$

$$P(-3<Z<3)=0.9974$$

표준점수 Z는 자료값과 평균과의 차이가 그 집단의 표준편차의 몇 배에 해당하는지를 알려주므로 우리는 어떤 값에 해당하는 표준점수를 통해 그 값이 집단에서 갖는 상대적 위치를 알 수 있다.

다시 또 다른 질문 "50점 이하를 받은 학생들은 얼마나 못한 것일까"를 생각해보자. 50점이라는 점수만으로는 이 학생의 성적을 판단할 수가 없는데, 왜냐하면 다른 아이들의 성적이 어떤가에 따라 이 점수가 낮은 점수일 수도 있고 높은 점수일 수도 있기 때문이다.

그렇다면 어떤 정보가 더 필요한가? 평균보다 얼마나 낮은가를 알면 될까? 이런 생각을 해보자. 어떤 학생이 중간고사에서는 평균보다 20점을 잘 봤고 기말고사에서는 평균보다 10점을 잘 봤다면, 어떤 시험에서 더 좋은 성적을 얻은 것인가? 평균과의 차이만으로 이 학생의 성적을 판단할 수 있는가? 만일 중간고사 성적의 표준편차가 20점이고, 기말고사 성적의 표준편차가 5점이라면 어떻겠는가? 중간고사에서의 20점 차이는 중간고

사 성적 표준편차의 1배에 해당하는 점수인 반면 기말에서의 10점 차이는 기말성적 표준편차의 2배에 해당하는 점수로 중간고사의 20점 차이가 기말고사의 10점 차이보다 더 작은 차이가 된다. 두 점수의 표준점수를 구해보면 중간고사 성적의 Z 값은 1이고, 기말고사 성적의 Z값은 2로, 중간고사 성적은 상위 15.87%에 해당하고 기말고사 성적은 상위 2.28%에 해당한다. 따라서 평균보다 10점이 높은 기말고사가 평균보다 20점이 높은 중간고사보다 훨씬 더 시험을 잘 본 것이 된다. 즉, 표준편차에 따라 각 점수의 상대적 위치가 달라지고 이는 표준점수를 이용하면 확연하게 드러난다.

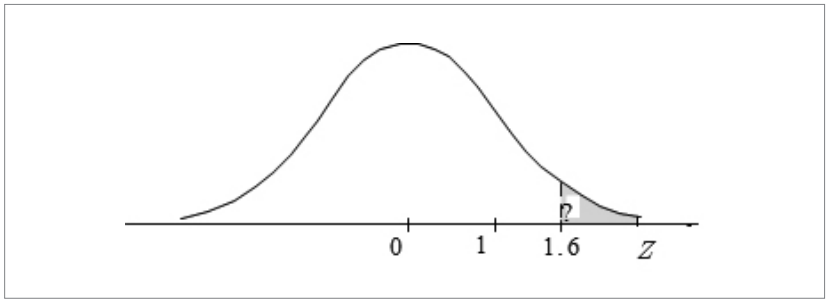
다시 50점 이하를 받은 학생들의 이야기로 돌아가 보자. 이 학생들이 얼마나 못한 것인지를 알기 위해서는 50점 이하인 학생들이 얼마나 있는지를 통하여 알 수 있고, 이 값은 표준점수를 이용하여 구할 수 있다. 50점의 표준점수는 $\frac{50-70}{10}=-2$이므로 50점 이하인 학생은 전체의 2.28%가 된다. 이 학생들은 하위 2.28%에 위치한다.

이렇게 표준점수는 자료의 상대적 위치를 알려주어 단위는 같으나 자료의 중심이나 퍼진 정도가 다른 자료나, 무게와 부피처럼 단위가 다른 자료에서도 비교가 가능하도록 해 준다. 즉, 표준점수 $Z=\frac{X-\mu}{\sigma}=k$ 는 임의의 개체의 값 X와 그가 속한 집단의 평균(μ)과의 차이가 표준편차(σ)의 k배임을 나타낸다.

몇 가지 Z 값들을 더 살펴보자.

- $Z=k=1.645$ 즉, 확률변수가 1.645배 표준편차 이상에 속하는 확률은 얼마나 될까?

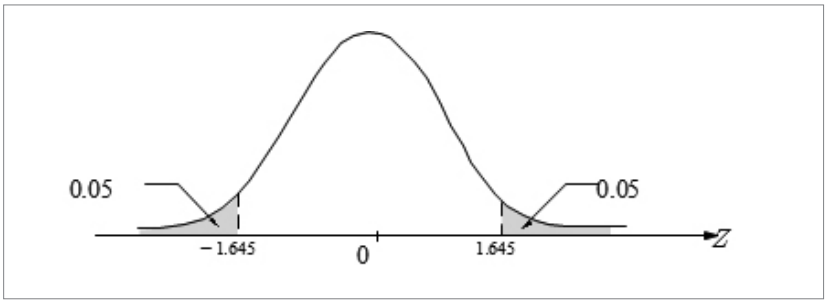
[그림 11-8]
표준정규분포
확률밀도함수 (1)



$P(Z>1.645)=1-P(Z<1.645)=1-0.95=0.05$이므로 1.645배 표준편차 이상인 값들은 5%가 된다.

- Z가 1.96배 표준편차 이상에 속하는 확률은 얼마나 될까? 동일하게 계산해보면 $P(Z>1.96)=1-P(Z<1.96)=1-0.975=0.025$가 된다.
- 정규분포는 좌우대칭이므로 Z가 1.645배 표준편차 바깥에 있을 확률은 10%가 되고, 동일하게 1.96배 표준편차 바깥에 있을 확률은 5%가 된다.

[그림 11-9]
표준정규분포
확률밀도함수 (2)



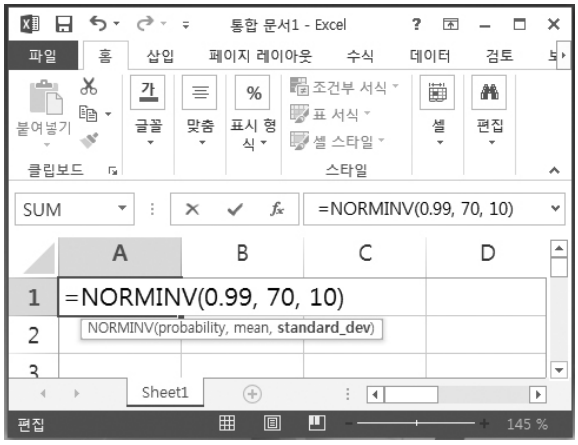
이러한 값들은 자주 등장하게 된다. 익숙해지면 도움이 될 것이다.

그러면 이제 마지막으로 "상위 1%는 몇 점 이상일까?"라는 질문으로 넘어가 보자. 지금까지 우리가 $P(Z>z)=p$ 에서 p를 구하는데 초점을 두었다면 이제는 p를 가지는 Z 값을 구하면 된다. 그리고 표준점수를 다음과 같이 뒤집으면 원래 구하고 싶은 점수를 구할 수 있다.

$$Z=\frac{X-\mu}{\sigma} \quad \Rightarrow \quad Z\sigma=X-\mu \quad \Rightarrow \quad X=\mu+Z\sigma$$

엑셀의 $=norminv(p, mean, sd)$ 라는 함수를 이용할 수 있는데 여기서 p는 누적확률 $P(Z<z)=p$ 의 p 값을 나타내므로 상위 1%를 구하려면 p에 0.99를 넣어줘야 한다.

[그림 11-10]
엑셀 정규분포 확률 계산 (2)



$P(Z>z)=1-P(Z<z)=1-0.99=0.01$ 인 z는 2.33이고 이 값을 표준화시키기 전으로 돌리면 $\mu+Z\sigma=70+2.33\times10=93.3$ 이 된다. 엑셀을 통하여 직접 구하면 $P(X>x)=1-P(X<x)=1-0.99=0.01$인 값은 93.3점으로 동일하게 계산된다. 따라서 전체 학생들 중에서 상위 1%에 속하는 학생들은 93.3점 이상이라는 것을 알 수 있다.

참고자료

- 김주한 · 김홍기 · 박래현 · 박석윤 · 배종호 · 이낙영 · 이석훈 · 이민구 · 이주호(2009), 통계학 입문, 정익사.
- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- http://suneung.re.kr/sub/info.do?m=0401&s=suneung
- https://ko.wikipedia.org/wiki/%ED%91%9C%EC%A4%80_%EC%A0%90%EC%88%98

제 12 장

개체의 자료 평가하기

12-1. '매우', '조금', '보통'의 정량화

학습목표

- **숫자를 보았을 때 그 크기에 대하여 우리가 갖는 느낌을 표현하는 주관적 부사인 '매우', '조금' 등에 대하여 정량화 할 필요성의 인지를 목표로 한다.**

1 키가 185cm인 사람은 얼마나 큰가?

이런 질문 앞에서 사람들은 대부분의 경우 세 가지 유형으로 나누어진다.

- 제1유형: 자신의 키와 비교한다. 예컨대 자신의 키가 175cm인 사람은 자기보다 10cm 크므로 많이 크다고 말하기도 하고 자신의 키가 180cm인 사람은 조금 크다고 말하기도 한다. 이런 사람들은 '매우', '조금'의 부사를 자신을 기준으로 결정하는 유형이지만 이들에게 왜 '매우'인지 질문하면 대답하지 못하는 경우가 많다.
- 제2유형: 성인의 평균 키와 비교한다. 성인의 평균 키를 175cm라고 생각하는 어떤 사람들은 평균보다 10cm크기 때문에 조금 크다고 말하기도 하고, 또 어떤 사람들은 동일한 평균키 175cm를 인정하면서도 매우 크다고 말하기도 한다. 같은 대상에 대하여 한 기관 안에서 이렇게 다른 생각을 하는 경우에는 의사결정과정에서 적지 않은 갈등을 경험하게 될 가능성이 높다.
- 제3유형: 185cm인 사람이 누구인지를 따져보는 유형이다. 남자인지 여자인지, 연령대가 어떻게 되는지를 먼저 확인하려 한다. 만약 185cm인 사람이 50대 성인 남성이라고 한다면 이 유형의 사람들은 50대 성인 남성의 평균 키를 생각하고 또 20대 성인 남성이라고 하면 20대 성인 남성

의 평균 키를 생각한다. 역시 앞의 유형들과 비슷하게 주관적인 결론에서 차이가 나타날 여지가 있다.

❷ 1인 가구로서 소득이 120만원인 사람은 소득이 얼마나 많은가?

이 질문에 대해서도 대부분 동일하게 반응한다. 사람들은 자기 자신과 비교하거나 또는 평균과 비교한다. 혹은 소득이 성별 혹은 연령대에 따라 다르다는 점을 고려하여 이 특정 1인 가구가 속해 있는 인구집단의 평균 소득과 비교할 것이다. 그러나 이 소득에 대한 표현에 있어서 '매우'와 '조금'을 붙이는 문제는 역시 앞의 신장에 관련된 토의에서 나타난 바와 같이 주관적인 입장이 포함되게 된다. 이러한 차이가 나타나는 경우 서로 간에 다른 평가가 내려지고 종종 의사결정에 갈등을 일으키게 된다.

❸ SNS 하루 평균 소통자수가 50명인 사람은 얼마나 많이 소통하는가?

어떤 사람이 SNS로 하루에 소통하는 평균 사람 수가 50명이라면 이 사람의 SNS 소통자수는 대단히 많다고 할 수 있을까?

이 문제 역시 많은 사람들은 자신의 하루 평균 SNS 소통자수나 평균 SNS 소통자수와 비교하려고 한다. 그러나 이 경우에는 사람들의 평균 SNS 대상자 수를 구하는 것은 쉬운 일이 아니어서 평균과 비교하는 것도 어렵다. 설령 평균을 알아서(예컨대 20명), 이 사람의 평균 소통자수 50명과 평균 20명의 차이인 30명을 얻었다고 해보자. 그러나 이 30명이라는 차이를 '매우' 큰 차이라고 해야 할 것인지 '조금' 차이 난 것이라고 해야 할 지에 대해서는 자신의 주관적인 부사를 사용할 수밖에 없다.

❹ 30분 정도 가사노동을 하는 남편은 얼마나 가정적인 남편인가?

우리 사회에서 아직까지 자주 제기되는 가정문제 중의 하나는 남성의 가사분담이 여성에 비하여 작다는 것이다. 빠른 속도로 변화되고 있는 현상이긴 하지만 아직도 적지 않게 힘듦을 호소하는 가정들이 있다. 다음의 질

문은 우리 주위에서 나타나는 흔한 이야기 거리일 것이다.

한 남성이 평일 가사노동에 참여하는 시간이 30분 정도라면 이 남성은 상당히 가정적이라고 할 수 있을까? 어떤 사람은 자신의 남편 혹은 자신과 비교하여 이 사람이 매우 가정적이라고 평가할지도 모르고 조금 가정적이라고 평가할지도 모른다. 또 통계청의 생활시간조사에 나타난 남편의 평균 가사노동 참여시간과 비교하여 평가하거나 혹은 지역이나 대도시 등을 따져볼 수도 있다. 그러나 여기에 붙이는 '매우', '조금'의 부사는 모두 주관적인 판단에 의해 결정되는 것이라는 점은 앞의 세 가지 문제와 동일하다.

이 장에서는 이러한 차이, '매우' 혹은 '조금'이라는 부사를 사용하는 주관적인 결정의 차이를 어떻게 해결할 수 있을지에 대하여 토의해보고자 한다.

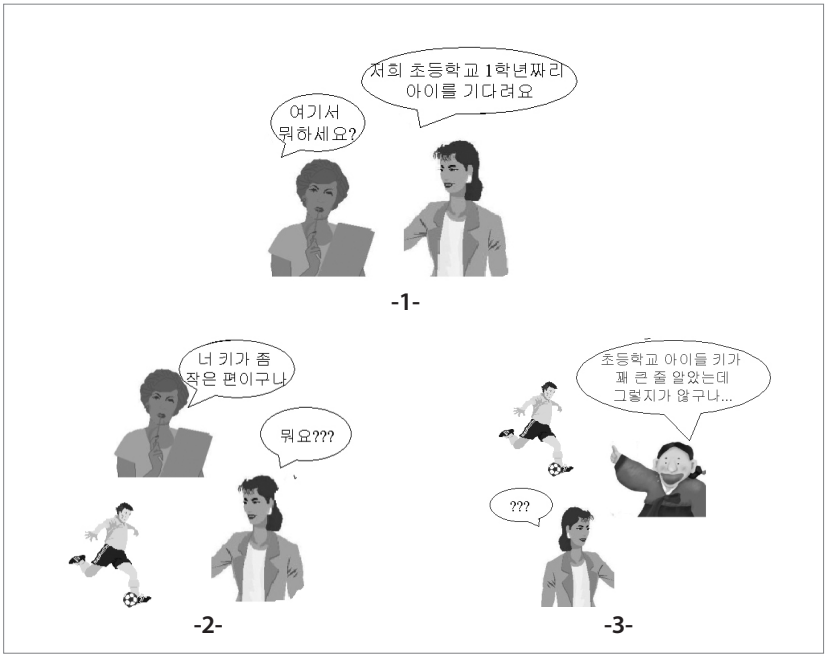
12-2. 모집단의 확률분포 가정

학습목표

- **개체로부터 측정된 값의 크기에 대한 상대적 평가를 하려면 그 개체가 포함된 모집단의 분포를 가정해야 한다는 것을 깨닫는다.**

앞의 부사의 문제에 대한 토의를 위하여 다음 두 여성의 대화에서부터 시작해보자. 먼저 그들의 대화를 살펴보자.

[그림 12-1]
두 여성의 대화



1 대화의 직관적 분석

대화에 나타난 두 여성의 차이가 생각의 차이인지 성격의 차이인지부터 따져보도록 하자.

1. 등장인물과 대화록

여인 A : 자신의 초등학교 1학년 아들을 기다리고 있다.
여인 B : "너 키가 좀 작은 편이구나!"라고 반응
여인 C : "초등학교 아이들 키가 꽤 큰 줄 알았는데 그렇지가 않구나!"라고 반응
초등학생 : 키가 120cm인 초등학교 1학년 학생으로 여인 A의 아들이다.

2. 등장인물 성격

동일한 초등학생의 키에 대하여 여인 B와 여인 C는 서로 다른 반응을 보이고 있다. 왜 한 초등학생의 키에 대하여 두 여인은 다른 평가를 하게 되는 것일까?

여인 B가 여인 C에 비하여 성격이 강하다고 생각되지 않는가? 여인 B는 자기의 기준(주관, 입장, 생각)에 따라서 앞에 있는 초등학생의 키(관찰, 경험)를 평가했다. 반면에 여인 C는 자신의 기준이 있기는 하지만, 자기 앞에 있는 임의의 한 초등학교 1학년 학생(관찰, 경험)을 초등학교 1학년 학생들 중 보통 키의 학생, 다시 말하면 중간 정도의 키를 갖고 있는 학생으로 생각(간주)하면서 자신이 가지고 있던 기준이 너무 큰 것은 아니었는지 검토하는 신중함(?) 혹은 자신 없음(?)을 보이고 있다.

3. 등장인물의 대화 분석

두 여인 모두 앞에 나타난 초등학생의 키 120cm를 작은 키로 평가했다는 점에서는 비슷한 생각을 하고 있다고 볼 수 있다. 다시 말하면 두 사람의 초등학교 1학년 학생의 키에 대한 기준은(120cm보다 크다는 의미에서) 비슷하다.

지금까지 살펴본 대화의 직관적인 분석에 대하여 어떤 생각이 드는지 정리해보자. 자신은 여인 B와 여인 C 중에서 어느 쪽에 더 가깝다고 생각하는가? 또 특별히 여인 C가 이 초등학생의 키를 초등학교 1학년 학생들의 보통 키로 간주한 태도에 대해 생각해 보면서 다음으로 넘어가자.

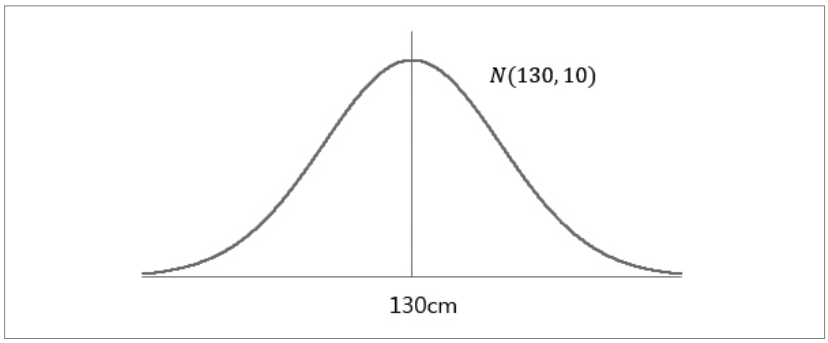
❷ 대화의 논리적 분석

여인 B와 여인 C 모두 초등학생을 보았을 때, '작다'나 '크지 않다'처럼 비교를 나타내는 단어를 사용한 것으로 보아 두 사람 모두 여기 나타난 초등학생과 비교하는 대상이 있었다고 생각된다. 여인 B와 여인 C가 떠올린 비교 대상은 무엇이었을까?

1. 초등학교 1학년 학생의 키에 대하여 두 사람은 10장에서 살펴본 것처럼 모집단(실제)이 정규분포를 따른다고 생각하였다고 하자. 그

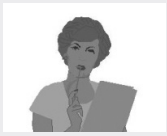
리고 두 사람 모두 평균을 그들이 만난 초등학생의 키 120cm보다 큰 130cm로, 표준편차는 10cm로 보았다고 가정해 보자.

[그림 12-2]
초등학생 평균키의 분포



이렇게 생각하고 있던 여인들은 아이를 보고 난 후 어떤 생각을 하게 되었을까?

2. 아이를 본 후 여인 B의 생각



"너 키가 좀 작은 편이구나."
⇒ 이 아이보다 작은 초등학교 1학년 학생의 수가 적다.
⇒ "좀" = $P(X<120)$

여인 B는 자신이 갖고 있는 초등학교 1학년 학생들의 키에 대한 입장에서 120cm인 학생을 만난 것이다. 따라서 여인 B에게 이 학생은 전체 학생들 중에서 작은 편(평균 130cm에서 한 배의 표준편차만큼 작은, 즉 표준화 점수가 -1인)이라고 생각하게 되었고 그래서 그냥 가볍게 '좀 작은 편'이라고 한 것이다. 여기서 '좀 작다'라는 말을 통계적으로 다시 표현해보면, 초등학교 1학년 학생들 중에서 이 학생(120cm)보다 더 작은 학생은 조금 있다는 뜻이라고도 할 수 있다. 그러므로 여인 B가 언급한 '좀'이라는 부사는 다음과 같이 계량화할 수 있다.

$$'좀' = P(X < 120 | X\text{는 평균 } 130,\ \text{표준편차 } 10\text{인 정규분포})$$

$$= P\left(\frac{X-130}{10} < \frac{120-130}{10}\right) = P(Z<-1) = 0.1587$$

"너 키가 좀 작은 편이구나."

⇒ 이 아이보다 작은 초등학교 1학년 학생의 수가 적다.

⇒ "좀" = $P(X<120)$

⇒ "이 학생은 키에 있어서는 작은 쪽에서 15.87% 정도 되는 위치에 있다."

3. 아이를 본 후 C 씨의 생각

"초등학교 아이들 키가 꽤 큰 줄 알았는데 그렇지가 않구나!"

⇒ 이 아이를 보고 자신의 생각을 판단. 즉, 초등학교 학생들의 키의 평균이 130cm라는 생각에 대해 판단.

이제 여인 C의 생각의 흐름을 살펴보도록 하자. 여인 C도 여인 B와 마찬가지로 초등학교 1학년 학생들의 키의 분포를 $N(130, 10^2)$(평균이 130, 표준편차가 10인 정규분포)로 생각하고 있었다. 그런데 여인 C는 키가 120cm인 학생을 보고 자신이 만난 이 학생을 초등학교 1학년 학생들의 평균 키인 학생을 만났다고 간주하면서, 자신이 초등학교 1학년 학생들의 키의 분포를 너무 큰 쪽으로 생각하고 있었다고 판단한 것을 알 수 있다.

(참고) 여인 C의 생각에 대한 또 하나의 이야기

여인 C의 생각에서 한 가지를 주목해 보자. 여인 C는 한 임의의 초등학교 1학년 학생의 키(120cm)를 보고 이 학생의 키를 초등학교 1학년 학생들의 평균키로 간주했다. 사실 그 학생은 초등학교 1학년 중에서 굉장히 작은 학생이었을지도 모르고 반대로 굉장히 큰 학생이었을지도 모른다. 그런데 굳이 그 학생을 '보통' 키의 학생으로 받아들이고 있는 것을 알 수 있다. 이 점에 관하여는 여러 가지 의견이 있을 수 있지만 일단 여인 C의 태도가 편견이나 선입견이 없는 일반적이고 보편적인 태도라고 볼 수 있다. 왜냐하면 기본적으로 우리는 보통의, 아주 일상적인 현상이 우리 앞에 나타날 것이라고(경험될 것이라고) 생각한다. 이러한 흐름에서 본다면 여인 C의 태도는 크게 잘못되었다고 볼 수는 없다.

그러나 그럼에도 불구하고, 한명의 학생을 보고 그 학생을 초등학교 1학년 학생들의 '보통' 키로 간주하는 것은 너무 위험한 것이 아닌가 하는 생각을 완전히 떨쳐버릴 수는 없다. 어느 날 임의의 한 사람을 만났을 때 이 사람을

보통 사람으로 간주하고 대해야 하는지 아니면 조금 경계하며 신중히 대해야 하는지를 선택하는 상황에서는 후자의 입장도 강하게 지지받고 있기 때문이다. 물론 여인 C는 고집스럽거나 자기주장이 강한 그런 여인은 아닌 것 같다. 순종형이라고 표현할 수도 있다. 하지만 또 한편으로는 관찰된 사실, 경험 등에 의해 자신의 생각 혹은 주관을 너무 쉽게 바꾸려는 경향을 가진 자신 없는, 자기 확신이 약한 그런 사람이라고도 할 수 있다.

❸ 통계적 모형화 과정

지금까지 우리는 "너 키가 좀 작은 편이구나.", "초등학교 아이들 키가 꽤 큰 줄 알았는데 그렇지가 않구나."라는 표현을 직관적으로 또 논리적으로 생각해 보았다. 그 생각의 순서를 단계적으로 기술해보면 다음과 같다.

- 단계 1 : 대화 장면만을 보았을 때는 통계적인 개념이 개입되었다고 보지 못했다.
- 단계 2 : '크다', '작다'라는 표현의 배후에 있는 비교 개념을 생각해 보았다.
- 단계 3 : 이 학생이 비교되고 있는 대상을 찾게 되었고 그 대상을 자신의 자녀의 키나 또는 초등학교 1학년 학생의 평균키로 생각하였다.
- 단계 4 : 이 학생의 키 120cm와 다른 또 하나의 숫자로써 초등학교 1학년 학생들의 평균 키를 떠올렸으나 두 숫자만의 비교로는 문제가 너무 단순화된다.
- 단계 5 : (통계적 사고방식) '크다', '작다'에 붙인 부사 '좀'의 개념을 이 학생이 속한 초등학교 1학년 학생 전체 중에서 이 학생의 키인 120cm가 어느 정도에 위치하는지를 통하여 파악한다.
- 단계 6 : (통계적 모형화) 이 학생이 속하는 초등학교 1학년 학생 전체의 키에 대한 자신의 입장을 정리, 표현하였다. 히스토그램을 통한 확률변수와 확률분포 개념을 도입한다. (X가 $N(130, 10^2)$을 따른다.)
- 단계 7-1 : 120cm인 학생의 키가 초등학교 1학년 학생집단 내에서 어디에 위치하는지를 누적확률로 계산하고, '좀 작은 편이구나'의 '좀'의 정도를 수량화하였다. (여인 B)
- 단계 7-2 : 120cm인 학생의 키를 초등학교 1학년 학생의 보통 키로 간주하면서 자신의 모형 $N(130, 10^2)$을 수정하는 것을 고려하였다. (여인 C)

12-3. 굉장히 큰 값이 나온다면… (심화)

학습목표

- **2절에서는 주관적 부사인 '굉장히', '매우', '조금' 등에 대하여 정량적 수치를 부여하였는데 이 절에서는 '굉장히'라는 부사를 어떻게 붙이는지 그때 생각의 흐름에 대해 주목하면서 활용의 방향을 인지한다.**

1 극단적인(희귀한) 사건 앞에서

지금까지 이야기 나눈 이 학생이 초등학교 4학년이 되었다고 가정하며 토의를 시작해보자. 한국교육개발원에서 발간된 교육통계연보에는 초등학생들의 평균키가 나와 있고 2013년 연보에 따르면 초등학교 4학년 남학생들의 평균키는 137.1cm이다. 앞의 초등학생도 키가 자라서 이제는 167.1cm가 되었다. 이 학생의 키에 대하여 다시 반응해보자.

이 학생의 키에 대해 이야기하려면 이 학생이 속한 집단의 키가 어떤 모습인지를 생각해보아야 한다. 키는 정규분포로 가정할 수 있다고 하였고 통계연보에서 초등학교 4학년 남학생들의 평균키가 137.1cm라고 하였으므로 그것을 그대로 받아들이기로 한다. 키의 표준편차는 10cm라고 알려져 있다고 가정하자. 다시 말하면 우리는 초등학교 4학년 남학생들의 키가 평균이 137.1이고 표준편차가 10인 정규분포를 따른다는 입장을 세운 것이다. 즉, 초등학교 4학년 남학생들의 키를 X로 표현한다면 $X \sim N(137.1, 10^2)$이라고 할 수 있다.

11장의 표준점수와 앞 절의 내용을 떠올려본다면 이 학생의 키 167.1cm의 위치를 찾아보는 것이 좋겠다. 167.1cm는 평균보다 크므로 이 학생의 키보다 큰 학생들이 얼마나 있는가를 살펴보자.

$$P(X > 167.1) = P\left(\frac{X-167.1}{10} > \frac{167.1-137.1}{10}\right) = P(Z > 3) = 0.0013$$

이 학생의 키보다 큰 학생들은 0.13% 있다. 즉, 이 학생은 상위 0.13%에 속하는 학생인 것이다.

상위 0.13%에 대해서 어떤 생각이 드는가? 여기에서부터 우리의 생각은 다시 차이가 생긴다. 키가 상위 0.13%인 학생을 보았을 때, 어떤 사람은

'굉장히' 크다며 놀라는 사람이 있는가 하면 어떤 사람들은 '보통'이라며 흔히 볼 수 있는 것처럼 반응한다. 후자의 경우 상위 0.13%보다 더 적은 가능성을 가진 사람을 만나야 놀라는 사람이라 할 수 있다. 그렇다면 원활한 의사소통을 위해 이제부터는 각자 자신의 기준을 정하고 이야기를 진행하자. 일단 상위 5%에 속하는 사건을 만났을 경우 '굉장히'라는 부사를 붙이기로 한다. 그러한 기준에서는 이 학생을 '굉장히' 큰 학생이라고 말하게 된다. 그렇다면 당신 앞에 '굉장히' 큰 키의 초등학생이 나타났다. 당신은 어떤 생각을 하게 될까?

'굉장히'라는 부사를 붙인 것은 이런 일이 일어나리라 예상하지 못했다고도 할 수 있다. 그건 놀라운 일이라고도 할 수 있고 이상한 일이라고도 할 수 있다. 그러나 무엇에 대하여 놀랍고 이상한 것일까? 이 학생의 키를 잘못 잰 것은 아닐까? 등의 여러 가지 생각이 따라올 것이다. 만약 이 학생이 147cm 정도였다면 그렇게까지 놀라거나 측정이 잘못된 것인가 등등은 생각하지 않을 것이다. 그렇지만 167.1cm에는 '왜?'라는 질문과 함께 이유를 또는 설명을 붙이고 싶은 마음이 든다. 이러한 생각이 들게 되는 것은(혹자는 동의가 안 될지도 모르지만) 우리는 우리 앞에 가능성이 높은 사건이 일어날 것이라고 믿고 살기 때문일 것이다. 따라서 초등학교 4학년 학생 중 다수를 차지하는 137cm 근처 학생이 나타나면 '왜?'라는 질문이 없지만 167cm와 같이 1,000명에 한 명이나 만날까 하는 학생을 보면 여러 생각이 나게 되는 것이다.

참고자료

- 김응환 · 이석훈(2015), 통계와 확률 교육.
- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- 통계교육원(2015), 통계기초 및 활용.

제 13 장 평가 사례 검토하기

13-1. 내 키는 얼마나 작을까?

학습목표

- **이 장에서는 10장과 11장, 12장에서 학습한 모집단을 정규분포로 가정하는 방법을 이용하여 주어진 상황에서 요구되는 개체의 관찰값에 대한 평가 방법을 학습하여 측정값에 대한 평가능력을 제고한다.**

1 정규분포를 따르는 모집단인 경우

1. 문제제기 1

초등학교 6학년인 수민이는 얼마 전 체격검사를 했는데 검사결과 키가 140cm라는 것을 알게 되었다. 수민이는 평소에도 키가 작다는 이야기를 많이 들었고 자기 스스로도 또래 아이들보다 키가 작다고 생각하고 있던 중이라 자신의 키가 과연 우리나라 초등학교 6학년 남학생들 중에 얼마나 작은지 궁금해졌다. 과연 수민이의 키는 얼마나 작을까?

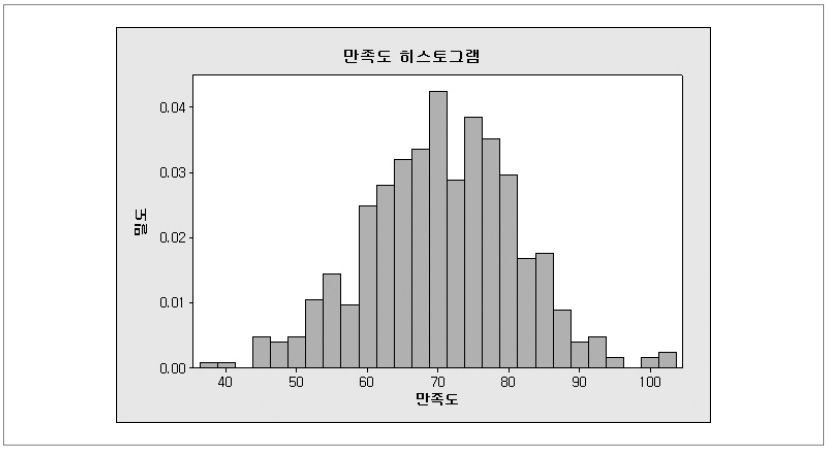
(1) 수민이의 키가 얼마나 작은지는 어디에서 비교해야 할까?

수민이의 키가 얼마나 작은지 알아보는데 20대 성인들의 키와 비교할 수는 없다. 따라서 수민이의 키가 어떤 집단에 속한 것인지 즉, 누구와 비교해야 하는지를 먼저 알아야 한다. 그런 다음 그 집단에 속한 사람들 중에서 얼마나 키가 작은지를 살펴보아야 한다. 그래서 수민이는 자신과 같은 우리나라 초등학교 6학년 남학생들이라는 집단을 떠올렸고 자신의 키가 이 초등학교 6학년 남학생들의 집단에서 차지하는 위치를 알아보는 것이 좋겠다고 생각했다.

(2) 수민이의 키가 속한 집단의 분포는?

우리나라 초등학교 6학년 남학생들의 키는 매우 다양한 값을 가지고 있다. 수민이네 반만 보아도 자신처럼 작은 키도 있고, 어떤 아이들은 아주 크기도 하다. 수민이는 '초등학교 6학년 남학생들의 키'의 다양성을 알아야한다는 생각이 들었고 앞의 12장이 떠올라 초등학교 6학년 남학생들의 키에 대한 히스토그램을 그려 살펴보려 했지만 자료를 찾을 수 없었다. 그러나 평균은 다르더라도 초등학교 6학년들의 키와 성인들의 키의 모양은 비슷할 것이라는 생각이 들어 우선 다음과 같은 국민체력실태조사의 성인 남성 2,779명의 키에 대한 히스토그램을 살펴보았다.

[그림 13-1]
성인 남성 키의 히스토그램



히스토그램의 형태를 통해 키의 분포는 일반적으로 정규분포로 볼 수 있겠다는 생각이 들었고 초등학교 6학년의 키에 대해서도 정규분포를 따른다고 가정하기로 하였다. 그 다음 초등학교 6학년 남학생들 키의 평균과 표준편차를 알기위하여 국가통계포털(KOSIS)에서 다음과 같은 자료를 수집했다.

<표 13-1>
초등학교 학년별
남학생 키의 평균과
표준편차

학년	통계	2015년 남자
초등학교1	평균	122.4
	표준편차	5.1
초등학교2	평균	128.6
	표준편차	5.39
초등학교3	평균	134
	표준편차	5.53
초등학교4	평균	139.1
	표준편차	5.67
초등학교5	평균	143.5
	표준편차	5.98
초등학교6	평균	150
	표준편차	7.15

http://kosis.kr/statHtml/statHtml.do?orgId=113&tblId=DT_113_STBL_1012274&vw_cd=MT_ZTITLE&list_id=113_11304_2010_004_007&seqNo=&lang_mode=ko&language=kor&obj_var_id=&itm_id=&conn_path=E1

수민이는 [그림 13-1]의 히스토그램과 <표 13-1>의 자료를 통하여 초등학교 6학년 남학생들의 키를 평균이 150cm에 표준편차가 7.15인 정규분포로 보기로 하였다. 이제 수민이의 키 140cm와 평균과의 차이 10cm가 얼마나 작은 것인지를 11장에서 배운 표준점수를 통하여 알아보자.

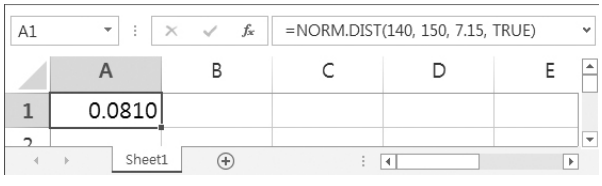
(3) 표준점수 계산

수민이의 키 140cm의 표준점수를 구하면

$$Z = \frac{140 - 150}{7.15} = -1.40$$

이 나온다. 이 값이 하위 몇 %인지 엑셀로 계산해보자(실습시간에 학습할 예정).

[그림 13-2]
엑셀 정규분포
확률 계산



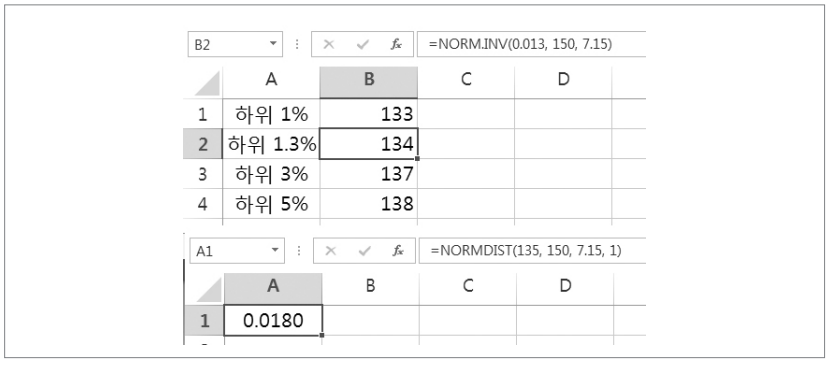
수민이의 키는 하위 8.10%에 속하는 값임을 알 수 있다. 따라서 수민이는 자신의 키가 작은 쪽에서 10% 이내에 있는 키라는 것을 알게 되었다.

2. 문제 제기 2

수민이의 친구 홍만이는 수민이보다 키가 더 작다. 이번 체격검사에서 홍만이의 키는 135cm로 반에서 가장 작았다. 홍만이의 어머니는 체격검사 결과를 듣고 전에 읽었던 'FDA, 건강한 아동에도 성장호르몬 사용 승인'이라는 기사를 생각하며 전문가와 상담을 받아야 하는 것이 아닐까 고민하게 되었다. 그러나 아직 성장호르몬의 안전성이 입증된 것이 아니고 또 키가 하위 1.2%에 속하는 아이들에 대해서만 성장호르몬을 사용할 수 있다고 했기 때문에 좀 더 신중해야 한다는 생각이 들었고, 일단 기사에서 말한 것처럼 홍만이의 키가 과연 하위 1.2% 안에 드는지를 먼저 알아보기로 하였다. (http://media.daum.net/foreign/others/newsview?newsid=20030728101820761)

홍만이의 어머니는 수민이가 구한 자료를 바탕으로 초등학교 남학생들의 키는 평균이 150cm이고, 표준편차가 7.15cm 정규분포를 따른다고 보고 여기서 하위 1.2%에 속하는 값을 계산해보기로 하였다. 홍만이의 어머니는 10장에서 배운 엑셀의 $=normdist(x, mean, sd, cumulative)$와 $=norminv(p, mean, sd)$라는 함수를 이용하기로 하였다.

[그림 13-3]
엑셀 정규분포 계산



[그림 13-3]의 결과를 보면 133.9cm가 하위 1.2%이고 홍만이의 키 135cm는 하위 1.8%로 홍만이가 반에서 가장 작기는 하지만 성장호르몬을 사용할 수 있는 대상에 속하지 않는다는 사실을 알았다. 홍만이의 어머니는 많은 생각 끝에 홍만이의 성장을 조금 더 느긋한 마음으로 바라보기로 결정하였다.

13-2. 내 소비 지출액은 얼마나 많을까?

학습목표

- **이 장에서는 10장과 11장, 12장에서 학습한 모집단을 지수분포로 가정하는 방법을 이용하여 주어진 상황에서 요구되는 개체의 관찰값에 대한 평가 방법을 학습하여 측정값에 대한 평가능력을 제고한다.**

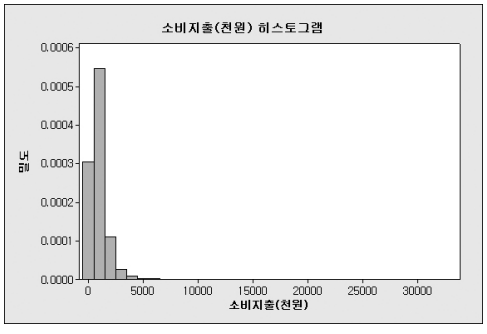
1 지수분포를 따르는 모집단인 경우

임씨는 작은 회사에서 근무 중인 1인 가구이다. 그는 직장동료들과 점심식사를 하던 중 한 달에 얼마나 지출(소비지출)하고 있는지를 이야기 하였고, 과연 자신은 얼마나 많이 소비지출을 하는지 궁금해지기 시작했다. (여기서 소비지출은 생계 및 생활을 위해 소비하는 내구재, 비내구재, 준내구재의 상품 및 서비스의 구입에 대한 대가로 지출하는 비용을 말한다. – 통계용어 지표이해 제6장 가계통계 부문

(http://kostat.go.kr/portal/korea/kor_ki/2/6/index.board?bmode=read&bSeq=&aSeq=198915&pageNo=1&rowNum=10&navCount=10&currPg=&sTarget=title&sTxt=)

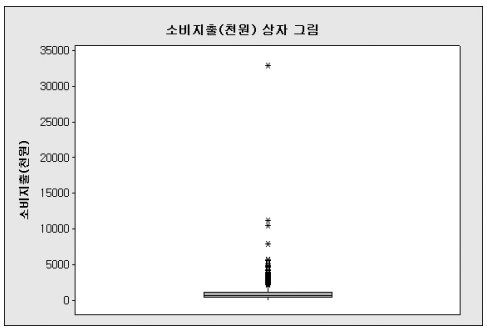
임씨는 지난달에 쓴 가계부를 찾아서 2015년 5월 총 지출액을 정리하였고 그 중 소비지출액으로 월 1,158,000원을 사용하였음을 알게 되었다. 임씨는 소비지출을 얼마나 많이 하는 것일까?

임씨는 자신이 5월에 지출한 소비지출액이 얼마나 많은 것인지를 알기 위하여 통계청 가계동향조사의 2015년 5월 소비지출액 자료에서 가구원수가 1인인 경우 ($n = 1{,}307$)를 이용하여 다음과 같은 히스토그램과 상자그림을 얻었다.



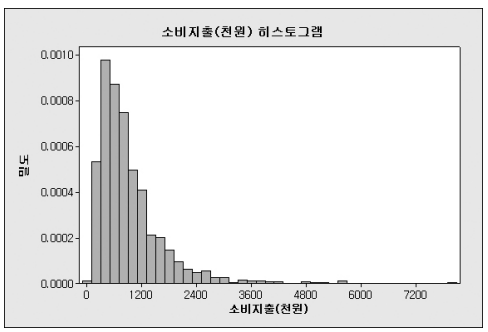
[그림 13-4] 소비지출 히스토그램 ($n = 1{,}307$)

[그림 13-5]
소비지출 상자그림
($n = 1{,}307$)



그래프를 확인한 결과 이상하게 큰 값이 존재한다는 생각이 들었다. 임씨는 8장에서 학습한 내용을 기억하며 월 소비지출액이 1,000만원을 넘는 3개의 값을 제외하고($n = 1{,}304$) 다시 히스토그램을 그려보았다.

[그림 13-6]
소비지출 히스토그램
($n = 1{,}304$)



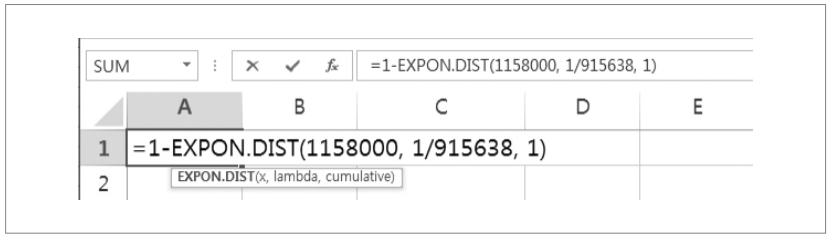
임씨는 이 히스토그램을 보고 1인가구의 월 소비지출액은 작은 값들은 오밀조밀하고 큰 값들은 넓게 퍼져있는 모습이며 이는 정규분포와는 다른 모습이라는 것을 알았다. 임씨는 이 모습을 10장에서 배운 지수분포의 형태와 비슷하다고 생각하고 1인가구의 월 소비지출액을 지수분포를 따른다고 가정하기로 하였다.

2015년 가계동향조사의 1인가구 자료를 이용하여 1인가구의 5월 소비지출액 평균 915,638원과 표준편차 740,497원을 계산하였다. 그러나 4장에서 대칭이 아닌 분포의 경우 중앙값과 사분위범위를 함께 확인하라는 내용을 참고하여 중앙값 725,495원과 사분위범위 726,258원도 함께 구하였다.

임씨는 자신의 5월 소비지출액이 1,158,000원이므로 자신이 평균보다 242,363원 많이 지출하고 있으며, 중앙값보다는 432,505원 많이 지출하고

있음을 알았다. 그는 자신이 과연 얼마나 많이 지출하는가를 자신의 소비 지출액이 상위 몇 %인지를 통하여 알아보기로 하고 엑셀(실습시간에 학습할 예정)로 다음과 같이 계산하여 0.2823의 값을 얻었다.

[그림 13-7] 엑셀 지수분포 확률 계산



이를 통하여 임씨의 5월 소비지출액은 상위 28.23%에 속하는 값이며 이는 1인 가구들의 5월 소비지출액과 비교하였을 때 많이 지출하는 편이기는 하지만 '굉장히'라는 말을 붙일 만큼 많이 지출하는 것은 아니라고 결론 내렸다.

13-3. 내 SNS 이용횟수는 얼마나 많을까?

학습목표

- **이 장에서는 10장과 11장, 12장에서 학습한 모집단을 포아송분포로 가정하는 방법을 이용하여 주어진 상황에서 요구되는 개체의 관찰값에 대한 평가 방법을 학습하여 측정값에 대한 평가능력을 제고한다.**

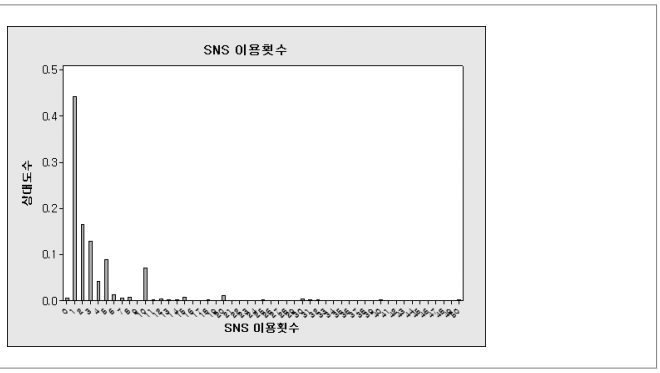
1 포아송분포를 따르는 모집단인 경우

임씨는 최근 "스마트폰 중독, 주요 원인은 SNS 통한 '대인관계 중독'"이라는 제목의 한 신문기사를 읽고 자신도 요즘 SNS로 맺고 있는 대인관계와 이용횟수가 늘고 있다는 생각이 들어 자신이 과연 SNS를 얼마나 많이 이용하고 있는지 확인하기로 하였다.

(http://news.chosun.com/site/data/html_dir/2015/04/13/2015041302766.html)

임씨는 지난 한 달간 자신의 SNS 이용횟수를 기록하고 일일 평균 SNS 이용횟수 7회를 얻었다. 임씨는 국가통계포털(KOSIS)에서 제공하고 있는 2011년 인터넷중독실태조사의 SNS 이용횟수 통계표를 사용하여 1일 SNS 이용횟수의 평균 3.35회를 구하고 아래의 막대그래프 ($n = 1{,}772$)를 작성하였다.

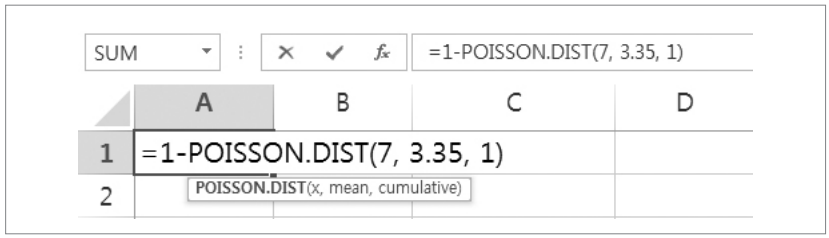
[그림 13-8] SNS 이용횟수 막대그래프 ($n = 1{,}772$)



임씨는 이 그래프를 통하여 이용횟수는 대칭이 아니라는 것을 볼 수 있었다. 이용횟수의 막대그래프는 앞 절의 소비지출액의 히스토그램과 비슷한 형태이지만 이용횟수가 이산형 변수라는 것을 생각하여 일일 SNS 이용횟수를 10장에서 배운 포아송분포를 따른다고 가정하기로 하였다.

임씨의 이용횟수 7회는 일일 SNS 이용횟수 평균 3.35회보다 더 많으므로 자신이 얼마나 많이 이용하는가를 확인하기 위하여 상위 몇 %에 해당하는지를 엑셀로 계산(실습시간에 학습할 예정)해 보기로 하였다.

[그림 13-9]
엑셀 포아송분포 확률 계산



위의 함수식을 이용하여 계산한 결과 임씨의 일일 SNS 이용횟수는 상위 2.14%에 해당하는 값이며 자신이 SNS를 '굉장히' 많이 이용하고 있음을 알았다.

참고자료

- 문화체육관광부(2013), 체력실태조사.
- 미래창조과학부(2011), 인터넷중독실태조사.
- 통계청(2015), 가계동향조사.
- 국가통계포털, http://kosis.kr/
- 중앙자살예방센터, http://www.spckorea.or.kr/index.php
- http://media.daum.net/foreign/others/newsview?newsid=20030728101820761
- http://kostat.go.kr/portal/korea/kor_ki/2/6/index.board?bmode=read&bSeq=&aSeq=198915&pageNo=1&rowNum=10&navCount=10&currPg=&sTarget=title&sTxt=
- http://news.chosun.com/site/data/html_dir/2015/04/13/2015041302766.html

제 14 장

표본추출분포 알아보기

14-1. 표본추출 분포의 필요성

학습목표

- **모집단에 대한 정보를 가지고 수집된 자료에서 산출된 통계치(표본평균, 표본비율 등)를 평가하기 위하여 표본추출분포가 필요하다는 것을 인식한다.**

다음의 글은 A씨가 어느 임원회의를 참석하고 느낀 느낌을 전한 내용이다. A씨의 글에서부터 토의를 시작해보자.

A씨는 어제 어떤 환경운동단체의 남자 임원회의에 초청되어 강의를 하였는데 그 곳에서 만난 25명의 남자 임원들의 키가 모두들 크다는 느낌이 들었다. 우선 그들의 키를 조사하여 통계 연수부에서 연수받은 대로 평균과 표준편차를 구했더니 177cm, 6cm가 계산되었다.

그런데 A씨는 얼마 전 문화체육관광부에서 발표한 가장 최근의 체력실태조사를 분석한 지인으로부터 한국 성인 남성의 키는 평균이 175cm, 표준편차가 5cm라고 들은 것을 기억하면서 불과 2cm정도 크다는 사실에 참 의아하다는 생각이 들었다. 왜냐하면 자신이 그 회의에 참석했을 때 느꼈던 느낌은 임원들의 키가 전체적으로 상당히 크다는 생각, 그러니까 보통 키보다 2cm 큰 것이 아니라 그보다 훨씬 크다는 느낌을 가졌기 때문이다.

1 다각적인 관점

1. 확률분포모형의 필요성

위 글이 통계적으로 검토되어야 한다고 생각하는가? 그렇다면 그 이유는 무엇일까? 다른 말로 하면 이 글의 배경에 기본적으로 깔려있는 불확실성을 내포한 현상 또는 집단을 찾아보자는 것이다. 이런 문제를 이미 여러 번 토의했기 때문에 쉽게 파악할 수 있을 것이다. 이 이야기의 핵심은 사람들의 키가 '다양하다'는 것이다. 따라서 우리는 11장과 12장에서 토의한대로 먼저 한국 성인 남성의 키에 대한 확률분포모형을 생각해야 한다. 키에 관해서는 대부분이 정규분포를 가정하는 것에 대하여 동의하리라 생각한다. A씨는 체력실태조사 결과인 평균 175cm와 표준편차 5cm를 알고 있으므로 성인 남성의 키의 분포는 평균 175에 표준편차 5의 정규분포로 가정하였을 것이다.

2. 성인 남성의 키에 대한 A씨의 이론적 결론

A씨가 한국 성인 남성의 키를 평균 175, 표준편차 5인 정규분포로 가정했기 때문에 키가 180cm인 성인 남성에 대해서는 11장과 12장에서 배운 것처럼 표준점수가

$$\frac{180-175}{5}=1$$

이 되어 상위 15.87%라는 것을 알 수 있고, 키가 185cm인 성인 남성은 표준점수가

$$\frac{185-175}{5}=2$$

로 상위 2.28%라고 생각했을 것이다. 마찬가지로 177cm인 성인 남성에 대해서도 표준점수를 구해 보면

$$\frac{177-175}{5}=0.4$$

가 되고 이 값이 얼마나 큰가는(엑셀을 이용해야겠지만) 상위 34.46%정도라는 것을 확인할 수 있을 것이다.

따라서 어제 만난 환경운동단체의 남자 임원 25명의 평균 키가 177cm라는 이야기를 듣고 이들 역시 상위 34.46%로 보통이거나 조금 큰 키라는 생각을 했을 것이다.

3. 25명 임원들의 키에 대한 A씨의 느낌

A씨는 그의 글에서 25명의 남자 임원들의 키가 전반적으로 크다는 느낌을 받았다고 했다. 그의 이 느낌은 11장과 12장에서 배운 대로 도출한 이론적 결론인 상위 34.46%보다 더 크다고 느껴졌다는 것이다. 여기서 그가 '전반적'이라고 말한 것을 보면 25명의 임원들 중에는 키가 작은 사람들도 더러 있다는 것으로 해석할 수 있다.

4. 이론적 결론과 느낌의 차이에 의한 혼란

한국 성인 남성 전체의 평균 신장 175cm와 임원 25명의 평균 신장 177cm의 차이 2cm를 별로 큰 차이라고 보지 않는 A씨의 이론적 결론에 대하여 당신은 어떻게 생각하는가? 당신도 A씨와 같은 결론을 얻었는가? 11장과 12장을 잘 소화한 사람이라면 아마도 A씨와 같은 결론을 얻었으리라고 추측된다.

그럼 어느 것이 더 실체에 가까운 결론일까? 이 25명의 임원들의 평균 키 177cm는 A씨의 이론적 결론대로 조금 큰 것일까? 아니면 A씨의 느낌대로 많이 큰 것일까?

5. 개체의 키와 평균 키의 동일시

그런데 A씨의 생각의 흐름을 다시 살펴보면, 한 개인의 키나 25명의 평균 키나 모두 동일한 표준점수를 갖고 있다. 왜냐하면 키 177cm인 한 사람의 표준점수를 구하기 위하여 분모에 사용된 개개인의 키의 표준편차가 25명의 평균 키 177cm의 표준점수를 구하기 위한 분모에도 동일하게 사용되었기 때문이다. 이것은 좀 이상하지 않은가? 개인의 키의 표준점수를 계산할 때 분모에 개인의 키에 대한 표준편차를 사용하였다면, 25명 평균 키의 표준점수를 계산하기 위한 분모에는 25명 평균 키의 표준편차를 사용하는 것이 맞지 않을까?

6. 평균 키의 표준편차의 유도

그러면 "평균 키의 표준편차"는 어떻게 알 수 있을까? 성인 남성 개개인의 키의 표준편차는 이들 25명 개개인의 키를 통해서도 계산될 수 있고 또 체

력실태조사의 결과에서도 알아낼 수 있다. 왜냐하면 개인들의 키가 다양한 것은 우리의 일상생활에서 만나는 여러 사람들로부터 자연스럽게 확인되는 내용이기 때문이다. 그런데 평균 키는 현재 177cm로 확인된 이 25명의 평균 값 하나밖에는 없다. "평균 키의 표준편차"라는 말이 의미를 가지려면 평균 키의 산포 정도(다양함의 정도)를 알아야 한다. 이 다양성에 관한 논리가 현대 통계학의 핵심적 내용 중의 하나이다.

7. 임의의 성인 남성 25명의 평균 키의 다양성(변이)

임의의 성인 남성 25명의 평균 키의 다양함을 생각한다는 말은 이들 임원 25명이 아니라, 만약 우리가 일반적인 성인 남성 25명을 단순임의추출법(Box에 이름을 써넣고 흔들어 뽑는 방법)으로 추출하여 그들의 키를 측정한다면 이들 25명의 평균 키는 얼마쯤 되는지 가상으로 생각해 본다는 것이다. 25명의 평균 키는 물론 다양할 것이다. 그런데 어떻게 다양할까? 그것을 알기 위하여 25명의 키의 표본평균의 분포를 알아낼 필요가 있는 것이다.

❷ 관심의 정량화

통계적 관점에서 제기한 질문은 평균 키에서 2cm의 차이를 큰 차이로 봐야 하는지, 즉 25명의 평균 키 177cm를 얼마나 큰 값이라고 할 수 있는가이다.

잠깐 12장으로 돌아가서 "키가 120cm인 학생을 작다고 할 수 있느냐?"를 "신장이 120cm인 학생보다 작은 학생들이 얼마나 되느냐?"로 바꾸었던 것을 떠올려 보자. 이 아이디어를 단순하게 접목한다면 "(확률표본 성인 남성 25명의 평균 키) 177cm는 큰 값인가?"라는 질문을 "177cm보다 큰 성인 남성들이 얼마나 되는가?" 라는 질문으로 만들 수 있다. 그러나 불행하게도 이런 단순한 접목은 말이 되지 않는다는 것을 곧 알 수 있다. 무엇이 말이 안 되는가 하면 확률표본 25명의 평균 키 177cm를 임의의 한 성인의 키와 비교하고 있다는 점이다. A씨는 지금 개인과 개인의 비교를 하는 것이 아니라, 어제 만난 25명의 임원들의 평균 키를 그가 다른 곳(어떤 25여명이 있는 사무실이거나 혹은 어떤 단체)에서 보았던 25명의 집단의 키와 비교하는 것이기 때문에 위의 질문은 다음과 같이 바뀌어야 한다.

"임의로 추출된 25명의 성인 남성의 평균 키가 177cm보다 클 가능성은 얼마나 될까?"

위의 질문을 답하기 위하여 필요한 것이 바로 '평균 키의 표준편차'이다.

다른 곳에서 만날 수 있는 임의의 성인 남성 25명의 평균 키가 어떠한가의 문제로 넘어가기 전에, 먼저 A씨의 생각의 흐름을 다시 한 번 분석해보자.

❸ 중대한 오류

먼저 A씨가 의아해 했던 이유를 추측해보자. A씨(아마도 대다수의 여러분들도)는 25명의 임원 한사람, 한사람의 키를 둘러보면서 전체적으로 키가 크다는 느낌을 받았고, 이들 25명의 평균 키 177cm를 이들 25명 집단의 대푯값으로 생각했다. 그런데 A씨는 이 대푯값 177cm를 그가 보통 만나는 사람들 개인의 키 – 평균이 175cm이고 표준편차가 5cm인 성인 남성 집단에 속한 임의의 한 사람의 키 – 와 비교하였기 때문에 대푯값으로의 평균 신장 177cm가 그저 보통 키로 느껴졌던 것이다.

평균이 175cm이고 표준편차가 5cm인 집단에서 177cm는 보통 키라고 할 수 있다. 왜냐하면 X를 임의의 한 사람의 키라고 할 때,

$$P(X>177)=P\left(\frac{X-175}{5}>\frac{177-175}{5}\right)=P(Z>0.4)=0.3446$$

이 된다. 즉, 177cm보다 큰 사람이 34.46%이고 작은 사람이 65.54%로 보통 키라고 할 수 있다. 그러나 A씨가 만난 177cm는 한 사람이 아닌 25명의 평균이다.

A씨가(그리고 아마도 여러분 대부분도) 이와 같이 착각하게 된 주된 이유는 A씨가 평소에 임의로 만난 성인 남성 25명의 평균 키가 얼마 정도인지에 관하여는 전혀 관심도, 생각도 해보지 않았기 때문이다. 다시 말하면 A씨가 환경운동단체 임원들 25명의 평균 키를 구한 것은 대단히 특별한 경우일 뿐, 평균 키 177cm와 비교해 볼 만한 다른 임의의 성인 남성 25명 집단의 평균 키는 본 적이 없다. 그렇기 때문에 순간적으로 그가 평상시에 보아왔던(생각해왔던) 사람들 개개인의 키와 비교하는 오류를 범한 것이다. 여기서 대단히 중요한 발견은 "경험한 적이 별로 없다"는 부분이다.

개인(여기서는 성인 남성)의 키의 분포는 평상시에 우리가 주위 사람들의 키로부터 히스토그램을 그려보고, 평균과 표준편차 등을 구해보기도 했고 또 통계연보나 신문 등에서 자주 보아온 것이기 때문에 평균이 175cm, 표준편차 5cm인 정규분포를 따른다고 하면 자연스럽게 받아들이게 된다. 하지만 이들 성인 남성으로부터 임의의 25명을 추출하여 그들의 평균 키를 구한다면 그 값들이 얼마인가에 대한 대답은 우리가 별로 본 적이 없기 때문에 참 막연해진다. 누가 25명 중에 한 명으로 뽑히느냐에 따라서 평균 키는 매번 다른 값을 가지게 되는데, 평균 키를 구해본 경험이 우리 안에 거의 없기 때문에 히스토그램의 모습도 상상이 되지 않고 이 값(평균 키)들의 대략적인 윤곽도 잡히지 않는 것이다.

이러한 이유로 해서 A씨의 오류 – 임원 25명의 평균 키 177cm를 그가 평소에 보아오던 177cm되는 한 개인과 같이 간주하여서 이들 25명의 평균 키를 보통 키로 결론지은 오류 – 를 많은 사람들도 하게 되는 것이다. 대단히 중요하기 때문에 반복, 반복하여 설명하고 있음을 이해하기 바란다. 다시 한 번 강조하지만 한 개인, 한 개체에 관하여는 경험이 많지만, 이들 개체가 다수 모인 집단의 평균에 관하여 우리는 거의 경험이 없다는 사실이다.

14-2. 표본평균의 표본추출 분포

학습목표

- **표본평균의 표본추출분포와 표준오차를 소개하여 모평균의 신뢰구간 등을 이해한다.**

1 표본평균의 표본추출분포

1절 마지막 부분에서 우리는 누가 표본으로 추출되느냐에 따라서 표본평균이 달라진다는 사실을 알았다. 이 사실과 25명의 임원들의 평균 키 177cm를 어떻게 결합하면 이 값의 크기를 평가할 수 있는데 이것은 이미 고등학교 교육과정에서도 중요하게 제시되었던 정리로 다음과 같다.

[정리 1] 정규분포와 관련된 정리

모집단이 평균이 μ이고 표준편차가 σ인 정규분포를 따를 때, 이들 모집단으로부터 크기가 n인 확률표본을 추출하여 이들 n개의 평균을 구하면 이 표본평균은 평균이 모집단의 평균 μ와 같고 표준편차는 모집단의 표준편차 σ를 $\sqrt{n}$으로 나눈, $\frac{\sigma}{\sqrt{n}}$ 인 정규분포를 따른다.

[정리 2] 중심극한정리

모집단이 평균이 μ이고 표준편차가 σ인 어떤 분포를 따를 때(정규 분포가 아니다), 이들 모집단으로부터 크기가 n 인 확률표본을 추출하여 이들 n개의 평균을 구하면 이 표본평균은 n이 크면 클수록 평균이 모집단의 평균 μ와 같고 표준편차는 모집단의 표준편차 σ를 으로 $\sqrt{n}$ 나눈, $\frac{\sigma}{\sqrt{n}}$ 인 정규분포에 가까워진다.

우리의 직관과 경험이 미치지 않는 세계(우리가 보지 않은 세계)를 위의 두 개의 정리가 해결하기 때문에 두 개의 정리가 중요하고 유명한 것이다.

다시 돌아가서 A씨의 오류를 해결하도록 하자. A씨는 성인 남성의 키의 분포를 평균이 175cm, 표준편차가 5cm인 정규분포를 따른다고 보았다. 이를 $N(175, 5^2)$으로 표기한다. 그리고 A씨는 환경운동단체의 임원들 25명에게서 평균 키 177cm, 표준편차 6cm를 얻었다.

이제 A씨는 [정리 1]로부터 임의의 성인 남성 25명을 추출하여 그들의 평균 키를 구하면 이들 평균 키의 값이 다양하게 계산될 것인데, 이 평균 키들의 평균은 모집단의 키의 평균과 같은 175cm이고, 표준편차는 $\frac{\text{모집단의 표준편차}}{\sqrt{\text{표본의 크기}}}=\frac{\sigma}{\sqrt{n}}=\frac{5}{\sqrt{25}}=1$ 인 정규분포를 따른다는 사실을 알았다(이것은 경험으로부터가 아니라 이론적으로 안 것이다). 다시 말하면 임의의 성인 남성 25명을 뽑아서 그들의 평균 키를 구하면 그 값은 175cm근처일 가능성이 높고 간혹 176cm나 174cm, 또 아주 간혹 177cm나 173cm도 나올 수 있겠다는 뜻이다.

이를 확률적으로 표현하면 다음과 같다. 여기서 $\overline{X}$는 임의의 성인 남성 25명의 평균키를 나타낸다.

$$P(174<\overline{X}<176)=P\left(\frac{174-175}{5/\sqrt{25}}<Z<\frac{176-175}{5/\sqrt{25}}\right)$$

$$=P(-1<Z<1)=0.3413\times 2=0.6826$$

$$P(173<\overline{X}<177)=P\left(\frac{173-175}{5/\sqrt{25}}<Z<\frac{177-175}{5/\sqrt{25}}\right)$$

$$=P(-2<Z<2)=0.4772\times 2=0.9544$$

$$P(\overline{X}>174)=P\left(\frac{\overline{X}-175}{5/\sqrt{25}}>\frac{174-175}{5/\sqrt{25}}\right)=P(Z>2)=0.0228$$

이제 $\overline{X}$가 177cm보다 클 확률이 0.0228로 나왔다. 서술적으로 말하면 A씨가 임의로 25명의 성인 남성을 추출해서 평균 키를 구할 경우(해보지는 않았다), 이 값이 A씨가 환경단체 임원 25명으로부터 얻은 평균 키 177cm보다 클 가능성은 2.28%라는 뜻이다. 따라서 A씨가 얻은 177cm란 값은 평균 키로서는 상당히 큰 값으로 A씨가 회의에 참석했을 때 느꼈던 "이 사람들 키가 참 크구나."라는 느낌과 일치하게 되는 것이다.

❷ 표본평균의 표준오차

다시 한 번 점검해 볼 것이 있다. 지금까지 진행해 온 내용을 통해서 깨달은 우리의 오류는 한 개인의 다름과 25명 집단의 평균의 다름을 동일한 기준으로 평가하려 했다는 것이다. 한 개인의 키가 얼마나 다른가에 대해 이

야기 할 때는 한 사람 한 사람의 다름을 표현하는 개인의 표준편차를 이용해야 하고, 25명의 평균에 대하여 이야기 할 때에는 25명의 평균들의 다름을 나타내는 25명 집단의 평균들의 표준편차를 사용해야 한다는 것이다. 이러한 차이 때문에 표본평균의 다름을 나타내는 표준편차를 표본평균의 표준오차(standard error)라는 이름으로 부른다.

표본평균의 표준오차는 모집단의 표준편차인 모표준편차를 표본의 크기의 제곱근으로 나누어 다음과 같이 구한다.

$$\frac{\text{모표준편차}}{\sqrt{\text{표본의 크기}}}$$

계산 방법이 너무 쉬워서 많은 사람들이 잘 알고 있다고 생각하지만, 그 의미는 계산만큼 쉽지 않음을 이절을 통하여 발견하였을 것이다.

❸ 표본평균의 표본추출분포의 활용

다음의 '간이보고서 1'을 통하여 표본평균의 표본추출분포에 대해 토의해 보자.

(간이보고서 1)

목표 직경 10cm인 볼 베어링을 생산하는 업체의 제품 직경은 평균 10.0cm, 표준편차 0.10cm인 분포를 따른다고 알려져 있다. 품질검사팀은 금일 생산 제품 중 임의로 추출한 100개의 제품으로부터 평균 직경 9.97cm, 표준편차 0.09cm를 얻었다. 품질검사팀은 평균 직경 9.97cm는 목표 직경 10.0cm와 한 배의 표준편차인 0.10cm를 고려할 때 거의 목표 직경에 근사한 값이라고 판단하여 금일 생산제품 전량에 대하여 합격판정을 내리고 출하를 결정하였다.

볼 베어링 생산업체의 상황은 다음과 같이 요약할 수 있다.

- 볼 베어링 생산업체는 자사 제품의 베어링의 평균직경이 10.0cm, 표준편차가 0.10cm로 생각하고 있다(이들은 분포의 형태가 정규분포라고 까지는 입장을 갖고 있지 않은 듯하다).
- 품질 검사팀은 하루 생산품 중 임의로 100개를 추출하여 직경을 측정하였고, 평균 9.97cm와 표준편차 0.09cm를 얻었다고 한다.

• 품질 검사팀은 10.0cm와 9.97cm의 차이가 0.03cm이므로 이 값은 표준편차 0.10cm에 비해서 상당히 작다고 판단하여 합격판정을 하였다.

이제 여러분은 여기에 큰 오류가 있음을 발견하였을 것이다.

품질 검사팀은 100개의 평균 직경이 얼마나 작은가를 판정하려고 한다. 그런데 여기서 그 다름의 기준으로 제품 하나하나의 직경이 서로 얼마나 다른가를 나타내는 표준편차(0.10cm)를 그 기준으로 삼아 판정하는 오류를 저질렀다. 우리가 앞 절을 통하여 배운 바에 의하면 100개의 평균 직경이 얼마나 작은가(혹은 큰가)를 판정하려면 100개의 평균 직경들의 다름을 생각해야 한다. 그리고 그 다름은 표준편차 대신 표본평균의 표준오차라는 이름으로 부르기로 하였다.

따라서 품질 검사팀은 표본의 평균 직경이 어떻게 다른가를 알아야 하는데 위의 상황에서 보면 볼 베어링 직경이 정규분포를 따른다고 가정하지 않았으므로 1절의 [정리 2] '중심극한정리'를 사용하여야 할 것이다.

중심극한정리를 통하여 임의의 볼 베어링 100개의 평균 직경은 평균이 10.0cm이고 표준오차가 $\frac{0.10}{\sqrt{100}} = 0.01$㎝ 인 정규분포에 가까운 분포를 따른다는 사실을 알 수 있다. 그러므로 품질 검사팀이 관찰한 9.97cm는 상위 0.13%에 속하는 값임을 다음의 계산으로 알 수 있다.

$$P(\overline{X} < 9.97) = P\left(\frac{\overline{X} - 10.0}{0.10/\sqrt{100}} < \frac{9.97 - 10.0}{0.10/\sqrt{100}}\right)$$

$$= P\left(Z < \frac{-0.03}{0.01}\right) = P(Z < -3) = 0.0013$$

즉, 업체가 믿고 있는 생산규격(10.0cm)에서 볼 때 9.97cm라는 평균값은 100개의 평균값으로서는 거의 불가능할 정도의(가능성이 0.13%인) 작은 값이라는 것을 알아야 했던 것이다. 따라서 품질 검사팀은 자신들이 추출한 100개의 제품 중에는 정상적인 공정에서는 거의 발견할 수 없을 정도의 짧은 직경을 갖는 제품들이 다수 포함되어 있을 가능성이 있으므로 금일 생산된 제품의 생산 공정에 심각한 문제가 있을 수 있다는 사실을 지적하여 불합격 판정을 했어야 했다.

다음의 '간이보고서 2'를 추가로 검토해보자.

(간이보고서 2)

OO기관은 지난해까지의 조사결과를 바탕으로 기관에 대한 민원인 만족도가 100점 만점에 평균 70점, 표준편차가 4점인 정규분포를 따른다고 생각하고 있다. OO기관은 민원인 만족도를 제고하고자 1억 이상이 투입된 환경개선운동 및 인센티브제도 등 대대적인 혁신운동을 전개하였다. 1년간 제고노력 후에 "기관을 방문한 민원인 중 임의로 추출된 400명의 민원인으로부터 조사된 만족도 결과는 평균이 70.4점이고 표준편차는 3.7점"인 것으로 확인되었다.

이 결과를 통보받은 기관장은 겨우 0.4점이 높아진 것에 크게 실망하고 전직원 회의에서 격한 어조의 훈시와 함께 직원들을 크게 꾸짖었다.

이 내용에서 통계적인 부분을 다시 정리하면 다음과 같다.

- OO기관은 지난해까지의 조사결과를 바탕으로 기관에 대한 민원인 만족도가 100점 만점에 평균 70점, 표준편차가 4점인 정규분포를 따른다고 생각하고 있다.
- 기관을 방문한 민원인 중 임의로 추출된 400명으로부터 조사된 만족도의 평균은 70.4점이고, 표준편차는 3.7점으로 나타났다.
- 기관장은 만족도 평균이 겨우 0.4점이 높아진 것에 크게 실망하고 격한 어조의 훈시와 함께 직원들을 크게 꾸짖었다.

OO기관의 만족도는 평균이 70점이고 표준편차 4점인 정규분포를 따르므로 임의의 400명의 민원인으로부터 얻은 표본평균의 표본추출분포는 평균이 70점이고 표준오차가 $\frac{4}{\sqrt{400}}=0.2$인 정규분포라는 것을 알 수 있다. 따라서 조사된 만족도의 평균 70.4점이 얼마나 높은 점수인가를 알아보면 다음과 같다.

$$P(\overline{X}>70.4)=P\left(\frac{\overline{X}-70}{0.2}>\frac{70.4-70}{0.2}\right)=P(Z>2)=0.0228$$

혁신운동을 한 후 얻은 400명의 만족도 평균 70.4는 만족도의 평균이 기존에 알려져 있던 70점인 상황에서는 상위 2.28%에 속하는 값으로 '겨우'가

아니라 '굉장히'(기준이 필요하지만) 높은 값으로 여간해서는 나오기 어려운 값이다. 그렇다면 이렇게 나오기 어려운 값이 나왔다는 것은 무엇을 의미하는가? 이 400명 속에 특별히 긍정적인 사람들이 많이 포함되어 있다는 말인가? 그렇다면 임의로 추출하였다는 말이 거짓일 가능성을 의심해야 한다. 만약 이 가능성을 배제할 수 있다면 어떤 결론을 내릴 수 있겠는가? "민원인의 만족도가 70점보다 높아졌다"라고 생각해볼 수 있지 않겠는가? 따라서 기관장이 표본평균의 표준오차와 표본평균의 표본추출분포를 정확히 알았다면 직원들에게 격한 어조의 훈시와 꾸짖음이 아니라 칭찬과 보상의 적절한 반응을 보일 수 있었을 것이다.

14-3. 표본비율의 표본추출 분포(심화)

학습목표

- **표본비율의 표본추출분포와 표준오차를 소개하여 모비율의 신뢰구간 등을 이해한다.**

1 표본비율의 표본추출분포

앞 절에서 우리는 표본평균을 보았을 때 표본평균들의 다양성에 주목해야 한다는 것을 알게 되었고, 이 다양성을 표본평균의 표준오차 그리고 표본평균의 표본추출분포라는 이름으로 불렀다. 이 절에서는 표본비율의 다양성, 즉 표본비율의 표준오차와 표본비율의 표본추출분포에 대하여 토의해보자.

표본비율의 표본추출분포에 대한 토의를 위하여 앞의 볼 베어링 생산업체의 이야기를 조금 바꾸어 생각해보자.

> 볼 베어링을 생산하는 A업체의 제품은 불량률이 0.1(10%)로 알려져 있다. 품질검사팀은 최근 A업체 제품의 불량률이 증가했다는 제보를 받고 A업체의 불량률을 평가하기로 하였다. 금일 생산제품 중 임의로 추출한 100개의 제품을 검사한 결과 불량인 제품은 16개인 것으로 나타났다. 품질검사팀은 A업체의 볼 베어링 제품에 대하여 불량률이 10%라고 보아도 되겠는가?

이 이야기에서 품질검사팀이 관심을 가지고 있는 것은 제품의 불량률이며 이를 확인하기 위하여 임의로 100개의 제품을 추출한 표본에서 불량률을 얻었다. 이 표본 불량률은 16%이다. 이쯤에서 떠오르는 질문이 있는가? "이 표본불량률 16%는 얼마나 큰 것일까? 과연 불량률이 높아졌다고 할 만큼 큰 값인가?" 이 질문은 다음과 같이 표현할 수 있다. "이 불량률 16%보다 큰 값이 나올 가능성은 얼마나 될까?" 이 질문에 대답하기 위하여 우리는 표본불량률(표본비율)의 표준오차와 표본불량률의 표본추출분포를 알아야 한다. 그렇다면, 표본비율의 표본추출분포는 어떻게 알 수 있을까?

표본비율을 이렇게 표현해보자. X가 제품의 불량여부를 나타내는 확률변수라고 하자. 즉, 제품이 불량이면 X는 1이고 불량이 아니면 X는 0이다. 그

러면 표본으로 뽑힌 n개의 제품 중에서 불량인 제품의 개수는 X가 1인 값을 모두 더한 것이 되고, X는 0과 1로 이루어져 있으므로 X의 합

$$X_1 + X_2 + \cdots + X_n$$

으로도 표현할 수 있다. 그러면 전체 제품 중 불량인 것의 비율을 나타내는 표본불량률은 다음과 같이 구할 수 있다.

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

이 식은 X값들을 전체 표본의 크기 n으로 나누는 평균을 구하는 식과 같다. 그렇다면 표본불량률 즉, 표본비율은 표본평균과 동일한 의미로 해석할 수 있게 된다. 따라서 모집단의 비율(불량률)을 p라 하고 표준편차를 σ라고 하면, 크기가 n인 표본의 표본비율(표본불량률) 곧, 표본평균 $\overline{X}$의 표본추출분포는 근사적으로 평균이 p 이고 표준오차가 $\frac{\text{모표준편차}}{\sqrt{n}}$인 정규분포가 된다.

❷ 표본비율의 표준오차

주어진 볼 베어링에 관한 이야기로 다시 돌아가면, 표본에서 불량품의 개수는 $x_1 + x_2 + \cdots + x_{100} = 16$개이고, 이 값을 전체 표본의 크기 $n = 100$로 나눈 다음의 평균이 곧 표본불량률이 된다.

$$\text{표본불량률} = \overline{x} = \frac{\sum_{i=1}^{100} x_i}{100} = \frac{16}{100} = 0.16$$

모집단의 불량률 p는 현재 0.1로 알려져 있다. 그러므로 표본불량률은 p보다 0.06만큼 더 크다. 그러면 으로도 표현할 수 있다. 그러면 전체 제품 중 불량인 것의 비율을 나타내는 표본불량률은 다음과 같이 구할 수 있다.

여기서 다시 한 번 말하지만, 우리는 지금 한 개체가 아닌 100개 베어링의 불량률에 관심을 가지고 있다. 따라서 크기가 100인 표본비율의 다름을 평가하기 위해서는 크기가 100인 표본비율들의 표준오차를 사용해야 한다. 앞 절에서도 말하였지만 표준오차는 다음과 같이 구한다.

$$\frac{\text{모표준편차}}{\sqrt{\text{표본의 크기}}}$$

앞에서 표본으로 뽑힌 개체의 불량여부를 X로 표현하였고 X는 '1'과 '0' 두 가지 값만을 가질 수 있다. 10장에서 배운 것을 기억한다면 X는 1의 확률이 p인 베르누이 분포를 따르며 베르누이 분포의 표준편차는 $\sqrt{p(1-p)}$로 알려져 있다. 따라서 표본비율의 표준오차는

$$\text{표본비율의 표준오차} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

가 된다. 주어진 예에서 표본불량률의 표준오차는

$$\frac{\sqrt{0.1(1-0.1)}}{\sqrt{100}} = 0.03$$

이다. 이 값을 이용하여 주어진 표본불량률 0.16의 표준점수를 구하면

$$Z = \frac{0.16-0.1}{0.03} = 2$$

가 된다. 이는 상위 2.28%에 속하는 값으로 이 제품의 불량률이 10%라면 나타날 가능성이 작은 굉장히 큰 값이라고 할 수 있다. 따라서 품질검사팀은 이 업체의 제품 불량률이 10%보다 높아졌을 것이라 판정하고 제조공정을 다시 살펴보아야 할 것이다.

참고자료

- 김응환 · 이석훈(2015), 통계와 확률 교육.
- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- 문화체육관광부(2013), 체력실태조사.
- 통계청(2014), 사회조사 보고서.
- 청소년건강행태 온라인조사.

제 15 장

표본추출분포 활용하기

15-1. 우리 팀원들은 행복한가?

학습목표

- **모집단을 정규분포로 가정할 수 있는 상황에서 표본추출분포를 이용하여 개인의 자화상이 아니라 우리들 집단의 모습을 보는 방법을 학습하여 통계적 가설검정의 개념을 느껴본다.**

1 정규분포를 따르는 모집단의 경우

다음은 K씨가 속한 A라는 동호회 회원들의 행복점수에 관한 토의 내용이다. K씨는 자신의 동호회가 자신을 많이 행복하게 했다고 생각하면서 자기 동료회원들도 같은 생각을 하는지 알고 싶어서 서울시(2014)가 조사한 설문지의 문항을 이용하여 36명의 동호회 회원들의 행복점수를 조사하였다. 조사결과 동호회 회원들의 평균은 74점이고 표준편차는 8점으로 나타났다.

1. 토의주제 1 – 동호회 회원들이 행복한가를 무엇으로 판단하는가?

K씨는 동호회 회원들의 행복 정도를 회원들의 평균 행복점수로 판단할 수 있다고 생각하였고 또한 회원들의 표준편차를 통하여 동호회 회원들이 얼마나 비슷한 수준에서 행복한 삶을 사는지도 알 수 있을 것이라고 생각했다.

2. 토의주제 2 – 동호회 회원들의 평균 행복점수가 높은지를 알기 위해서는 누구와 비교해야 하는가?

K씨는 가장 쉬운 방법은 다른 동호회 회원들을 조사하는 것으로 생각했는데 조사한다는 것이 현실적으로 결코 쉽지 않다는 것을 알았다. 그래서 K씨는 서울 시민들의 행복점수에 관심을 가지고 그가 읽어본 적이 있는 2014 서울서베이 보고서를 찾아보았다. 다행히 보고서에서 아래의 표를 발견하였다.

[그림 15-1]
서울서베이 보고서 – 행복점수

부록1. (25)주요 변수별 표준 오차 : 행복점수(100점)

전체	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
전체	72.0263	0.05844	71.91180	72.14089	12.46513

주택유형	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
단독주택	71.5300	0.09551	71.34603	71.71398	12.70662
아파트	72.6609	0.09027	72.48049	72.84137	12.37263
다세대 주택	71.5279	0.15478	71.22470	71.83106	12.25527
연립/기타	72.0226	0.22516	71.58819	72.45691	11.78920

구	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
종로구	71.2896	0.36846	70.72136	71.85781	9.64147
중 구	69.3490	0.46516	68.70564	69.99243	10.71260
용산구	70.7542	0.36417	70.12941	71.37900	11.54639
성동구	72.4267	0.32674	71.84355	73.00984	12.06710
광진구	70.4823	0.32274	69.84631	71.11828	13.55077
동대문	69.4651	0.29031	68.88663	70.04363	11.89062
중랑구	73.7840	0.25514	73.26532	74.30259	11.50360
성북구	71.0006	0.27540	70.41285	71.58839	12.88095
강북구	70.0191	0.34223	69.37792	70.66029	13.68504
도봉구	69.8225	0.25681	69.33783	70.30712	10.54884
노원구	70.1759	0.23608	69.66420	70.68758	11.86353
은평구	69.3567	0.24833	68.84856	69.86486	11.88936
서대문	74.6429	0.28808	74.12481	75.16090	10.91372
마포구	74.6946	0.26459	74.19503	75.19409	10.89963
양천구	72.8077	0.20926	72.39255	73.22288	9.56238
강서구	70.6942	0.30630	70.04086	71.34760	15.86489
구로구	73.4708	0.25980	72.97634	73.96526	11.29621
금천구	64.2755	0.54613	63.32865	65.22227	18.28013
영등포	72.5485	0.24608	72.05706	73.03999	10.24390
동작구	73.2403	0.28751	72.66787	73.81267	12.70627
관악구	68.2057	0.29788	67.60205	68.80939	14.76766
서초구	76.2605	0.22897	75.79545	76.72550	9.96757
강남구	77.0154	0.24909	76.51958	77.51125	11.93556
송파구	72.9937	0.17495	72.62173	73.36568	9.07401
강동구	77.1461	0.25019	76.67181	77.62040	11.13134

K씨는 이 표에서 서울시민을 대표하는 약 20,000가구를 표본으로 추출하여 얻은 평균 행복점수가 100점 만점에 약 72(72.0263)점이고 표준편차가 약 12(12.46213)점인 것을 알았다. 한편 주거하는 주택유형에 따라서는 아파트에 거주하는 시민들의 행복점수가 72.7점 정도로 다른 유형의 주택에 거주하는 시민들의 행복점수보다 높은 것을 알았고, 지역(구)에 따라서는

생각했던 것보다는 차이가 많이 나는 것을 발견하였다. 한편, K씨의 동호회 회원들은 아파트와 다른 주택유형에 사는 사람들이 모두 포함되어 있고 또 여러 지역에 거주하고 있기 때문에 서울시 전체 시민들과 비교해야 한다고 생각하였다. 그런데 K씨는 서울시민의 행복점수는 특별한 경우가 아니라면 좌우대칭의 정규분포를 따를 것이라고 판단하고 행복점수의 분포를 평균이 72점이고 표준편차가 12점인 정규분포로 가정하였다.

3. 토의주제 3 – 동호회 평균점수와 서울시민 전체 평균점수를 어떻게 비교해야 하는가?

이 주제는 14장의 핵심내용으로 서울시민 중 임의의 36명을 표본으로 추출하였을 때 그 표본이 갖는 평균값들의 분포를 생각하고 이 분포를 기준으로 동호회의 평균점수의 크기에 대한 상대적 평가를 하기로 하였다. 그래서 K씨는 임의의 36명의 평균 행복점수의 표본추출분포를 고심 끝에 구하였다.

임의의 36명의 행복점수에 대한 표본평균의 표본추출분포는 14장의 [정리 1]에 따라 평균이 전체 서울시민의 행복점수 평균인 72점과 같고 표본평균의 표준오차는 $\frac{12}{\sqrt{36}}=2$점인 정규분포이다.

따라서 K씨의 동호회 36명 회원들의 행복점수 74점의 표준점수는 $\frac{74-72}{2}=1$로 상위 15.87%에 해당하는 값이 된다. K씨는 자신이 속한 동호회 회원들의 행복점수는 상위 약 16%에 해당하며 굉장히 높다고 까지는 못하겠지만 그래도 높은 편이라는 결론을 내렸다.

15-2. 우리 모임은 소비지출을 많이 하는가?

학습목표

- **모집단을 지수분포로 가정할 수 있는 상황에서 표본추출분포를 이용하여 개인의 자화상이 아니라 우리들 집단의 모습을 보는 방법을 학습하여 통계적 가설검정의 개념을 느껴본다.**

1 정규분포를 따르지 않는 모집단

13-2절의 임씨와 같은 독신들이 형성한 '녹색실천모임' 회원 16명은 자신들의 소비규모를 알기 위하여 월 소비지출액을 조사하여 다음과 같은 자료를 얻었다.

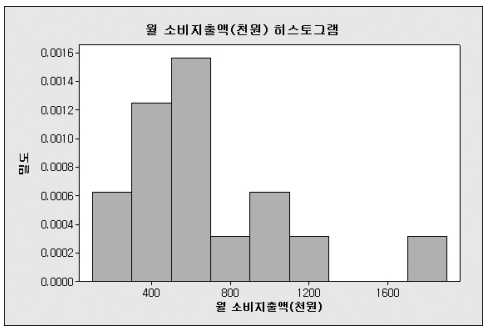
<표 15-1> 월 소비지출액 – 녹색실천모임

ID	월 소비지출액	ID	월 소비지출액
1	1,047,627	11	959,087
2	636,420	12	568,548
3	1,153,536	13	799,760
4	329,211	14	406,397
5	354,520	15	1,807,759
6	555,333	16	168,970
7	382,867	평균	677,104
8	298,360	표준편차	412,519
9	667,732	중앙값	602,484
10	697,540	표준편차	463,812

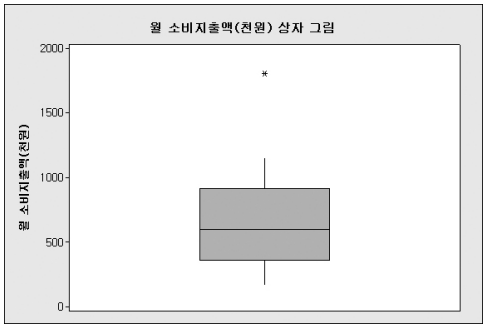
1. 토의주제 1 – 이 모임 16명의 소비지출의 분포를 알아보자.

임씨는 '녹색실천모임' 16명의 평균 월 소비지출액이 과연 많은 편인지 궁금해졌다. 먼저 자료를 이용하여 평균 677,104원과 표준편차 412,519원을 얻고, 상자그림과 히스토그램을 그려보았다.

[그림 15-2]
'녹색실천모임'
월 소비지출액
히스토그램
($n = 16$)



[그림 15-3]
'녹색실천모임'
월 소비지출액
상자그림
($n = 16$)

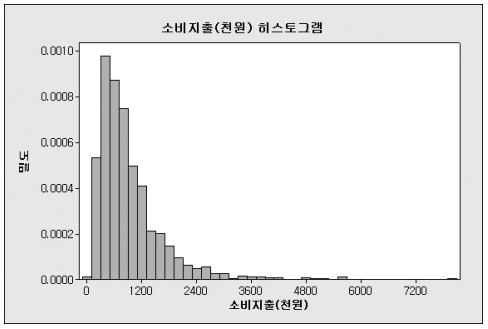


상자그림과 히스토그램을 통해 살펴보면 소비지출 중에 큰 값이 하나 포함되어 있는 것을 알 수 있고, 분포형태는 왼쪽으로 기울어진 형태를 보이고 있다.

2. 토의주제 2 – 우리나라 1인가구들의 소비지출액은?

한편 13장에서 1,304명으로부터 조사한 가계동향조사의 조사자료를 이용하여 작성한 히스토그램을 기억해보면 다음과 같다.

[그림 15-4]
월 소비지출액
히스토그램
($n = 1,304$)



임씨는 이 히스토그램을 통해 월 소비지출이 평균 915,638원인 지수분포를 따른다고 가정하였다.

3. 토의주제 3 – 임의의 1인가구 16명의 평균 월 소비지출액은 얼마나 될까?

비록 1인가구 개개인의 월 소비지출액은 지수분포를 따르지만, 16명의 평균 월 소비지출액은 14장에서 소개한 [정리 2] 중심극한정리에 의하여 평균이 915,638원이고 표준편차가

$$\frac{\sigma}{\sqrt{n}}=\frac{915,638}{\frac{915,638}{\sqrt{16}}}=228,910\text{원}$$

인 근사적인 정규분포를 따른다(지수분포는 평균과 표준편차가 같은데 여기서는 받아들이자).

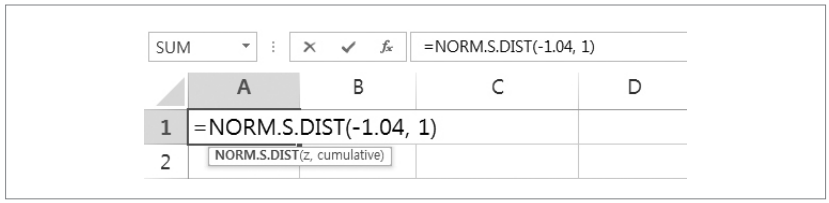
4. 토의주제 4 – '녹색실천모임'의 소비규모는 어느 정도인가?

'녹색실천모임'의 평균 월 소비지출액은 677,104원으로 가계동향조사에서 얻은 평균 915,638원보다 238,534원 작다. 이 값이 얼마나 작은가를 알아보기 위해서 이 표본평균값이 하위 몇 %에 속하는지 알아보자. 우선 '녹색실천모임' 16명의 평균 월 소비지출액에 대한 표준점수를 구하면

$$Z=\frac{677,104-915,638}{\frac{228,910}{\sqrt{16}}}=-1.04$$

가 된다. 다음과 같은 엑셀의 함수를 이용하여 이 값이 하위 몇 %인지를 계산해 보면 0.1492라는 값을 얻을 수 있다.

[그림 15-5]
엑셀 표준정규분포 확률 계산 (1)



임씨는 16명의 독신회원으로 이루어진 '녹색실천모임'의 평균 월 소비지출액은 하위 14.92%로 이들의 월 소비지출액은 적은 편이라고 할 수 있다.

5. 토의주제 5 – '독신행복모임'의 소비규모는 어느 정도인가?

한편 임씨의 친구 독신 K씨가 속한 '독신행복모임'에서도 같은 조사를 하였는데 이들 회원 25명의 조사결과는 평균이 1,299,586원으로 평균 915,638원보다 383,948원 많이 지출하는 것으로 나타났다. '독신행복모임'은 25명이므로 임의의 25명의 평균 월 소비지출액은 중심극한정리에 의하여 평균이 915,638원이고 표준편차가

$$\frac{\sigma}{\sqrt{n}}=\frac{915,638}{\frac{915,638}{\sqrt{25}}}=183,128\text{원}$$

인 근사적인 정규분포를 따른다. 따라서 '독신행복모임'의 평균 월 소비지출액의 표준점수를 구해보면

$$Z=\frac{1,299,586-915,638}{\frac{228,910}{\sqrt{25}}}=2.10$$

이 되고 이 값은 상위 1.8%에 해당하는 값이다. 따라서 '독신행복모임'의 소비규모는 '녹색실천모임'의 규모와는 반대로 월 소비지출이 상위 1.8%로 '상당히' 큰 모임인 것으로 파악되었다.

15-3. 우리 모임은 SNS를 많이 이용하는가?

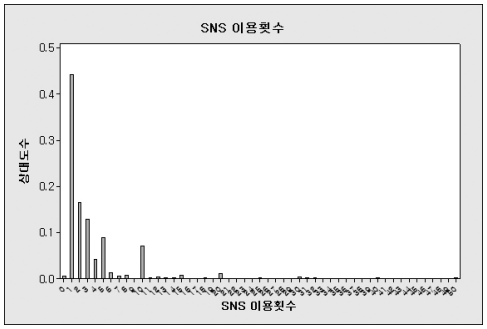
학습목표

- **모집단을 포아송분포로 가정할 수 있는 상황에서 표본추출분포를 이용하여 개인의 자화상이 아니라 우리들 집단의 모습을 보는 방법을 학습하여 통계적 가설검정의 개념을 느껴본다.**

1 포아송분포를 따르는 모집단

임씨는 2011년 인터넷중독실태조사의 SNS 이용횟수 통계표를 이용하여 다음과 같은 막대그래프를 그리고 평균 3.35회라는 것을 파악하였고, 13-3절을 통하여 자신의 일일 SNS 이용횟수 7회는 상위 2.14%에 속하는 값이라는 것을 알았다. 한편 통계적 사고에 심취된 임씨는 자신이 참여하는 종교모임에 소속된 회원들의 SNS 이용횟수는 어떠할지 궁금하여 25명을 대상으로 조사한 결과, 평균 4회로 3.35회보다 0.65회 더 많이 이용하고 있는 것으로 나타났다. 이 '종교모임'에 속한 사람들의 SNS 이용횟수는 굉장히 많다고 할 수 있을까?

[그림 15-6] SNS 이용횟수 ($n = 1,772$)



13-3절에서와 같이 이 막대그래프의 형태를 통해 일일 SNS 이용횟수는 포아송분포를 따른다고 가정한다. 그러나 지금 임씨가 관심을 가지고 있는 것은 종교모임 25명의 평균 일일 SNS 이용횟수이므로 우리는 개개인의 이용횟수가 아닌 임의의 25명의 평균 SNS 이용횟수(표본평균)의 분포를 생각해야 한다. 14장의 중심극한정리를 떠올려 보면 표본평균의 표본추출분포는 모집단이 포아송분포라고 해도(정규분포를 따르지 않아도) 근사적으로 정규분포를 따른다는 것을 알 수 있다.

따라서 임의의 25명의 일일 SNS 이용횟수의 표본추출분포는 근사적으로 평균이 3.35이고 분산이

$$\frac{\sigma}{\sqrt{n}}=\frac{\sqrt{3.35}}{\sqrt{25}}=0.37$$

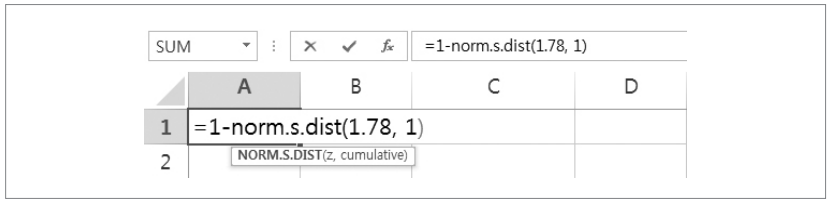
인 정규분포를 따른다(포아송분포는 평균과 분산이 같다).

따라서 임씨가 속한 종교모임의 평균 SNS 이용횟수 4회에 대한 표준점수를 구하면

$$Z=\frac{4-3.35}{\frac{\sqrt{3.35}}{\sqrt{25}}}=1.78$$

이 된다. 다음과 같은 엑셀의 함수를 이용하면 0.0375라는 값을 얻을 수 있다.

[그림 15-7]
엑셀 표준정규분포 확률 계산 (2)



임씨가 속한 종교모임의 평균 일일 SNS 이용횟수 4회는 상위 3.75%에 해당하는 값으로 이 종교모임은 SNS를 굉장히(상당히, 매우 등 주관적으로 결정하겠지만) 많이 이용한다고 할 수 있다.

참고자료

- 서울특별시(2014), 서울서베이.
- 통계청(2015), 가계동향조사.
- 한국인터넷진흥원(2011), 인터넷중독실태조사.

제 16 장

신뢰구간 이해하기

16-1. 신뢰구간 만나보기

학습목표

- **신뢰구간이 이미 고등학교 교과과정에서부터 나왔던 용어, 개념인 것과 생활 주변에서 자주 접하는 것인데도 이해를 잘 못하고 있음을 자각한다.**

1 학교교육 현장

다음의 <표 16-2>는 이준열 등(2014)이 집필한 2009 개정 교육과정 고등학교 "확률과 통계"교과서 내용 중 신뢰구간에 대한 것이다. 여러분이 기억하고 있는 것보다 더 어려운 내용이 이미 고교 교육과정에서 다루어지고 있다는 사실에 다소 당황할 수도 있다고 생각이 된다. 따라서 신뢰구간과 관련된 용어를 우리가 생활에서 얼마나 자주 접하고 있는지 살펴보기 위하여 모든 사람들이 쉽게 접하는 언론매체의 사례를 고등학교 교과서 내용을 보기 전에 먼저 제시한다. 다음의 기사 일부를 보자. <표 16-1>은 국내 여객선 안심수준을 측정한 것인데 1,000명을 조사한 결과 36.7점의 심각성을 전하면서 "신뢰수준 95.0%에서 표본오차 ±3.10%p"를 제시하고 있다. 또 다른 사례에서는 서울의대가 의뢰한 조사로 사전 연명의료 결정제도에 대한 조사결과 80.2% 가 찬성하고 응답자의 95.5%가 호스피스가 필요하다고 응답한 결과를 "표본오차는 95% 신뢰수준에서 ± 4.4% 포인트다"라는 정보와 함께 수록된 연합뉴스 기사의 일부이다. 두 사례의 차이점은 하나는 백점만점의 점수를 조사하였고 다른 하나는 찬성률을 조사했다는 것이다.

<표 16-1>
언론매체에 나타난 신뢰수준과 표본오차 사례

세월호 사건 이후 국내 여객선 안심수준 '매우 심각'

사후 처리(후속조치와 책임소재 규명 등)에 대한 염려와 불신이 매우 높은 상황

[전국뉴스 한용덕 기자] 세월호 사건 이후 국내 여객선 안전에 대한 사회적 관심과 안전을 위한 다차원적 노력이 이루어지고 …이에 성균관대 SSK 위험커뮤니케이션연구단(단장 송해룡 교수, 이하 위험컴연구단)과 (주)포커스컴퍼니(대표이사 최정숙)에서는 여객선 이용에 대한 국민의 안심수준을 측정하고자, 국내에 거주하는 만 20세 이상의 성인남녀 1,000명을 대상으로 설문조사를 진행했다. (신뢰수준 95.0%에서 표본오차 ±3.10%p) …

국내 여객선 이용에 대한 국민의 안심수준은 36.7점, 매우 심각한 수준 국내 여객선의 안심지수는 100점 만점을 기준, 36.7점으로 매우 심각한 수준인 것으로 나타났다. … 성별로는 '여성'(33.4점)이 '남성'(39.9점)보다 여객선 안심지수가 낮게 나타났고, 연령별로는 '30대'(35.7점)가 '20대'(37.3점), '40대'(36.1점), '50대'(37.8점)에 비해 상대적으로 낮은 안심수준을 보인 것으로 나타났다. …

http://www.jeonguknews.co.kr/news/articleView.html?idxno=14995

국민 96% "말기 환자, 호스피스 필요"

…

서울대의대 조사…사전 연명의료 결정도 80.2% 찬성

(서울=연합뉴스) 김길원 기자 = 우리나라 국민 10명 중 9명 이상은 말기 환자에게 '호스피스'가 필요하다고 생각하는 것으로 나타났다

…

패널은 전국 단위의 대표성 있는 30만 명으로 구성됐으며, 조사 대상자 선정에는 지역, 성별, 연령 등에 따른 할당 추출 방식이 사용됐다. 표본오차는 95% 신뢰수준에서 ± 4.4% 포인트다.

조사결과를 보면 응답자의 95.5%가 호스피스가 필요하다고 답했다. 환자의 증상 호전 없이 임종을 연장하는 의학적 시술을 보류하거나 중단하는 '연명의료결정'에 대해서도 응답자의 80.2%가 필요하다는 의견을 내놨다. …

http://www.yonhapnews.co.kr/bulletin/2015/11/26/0200000000AKR20151126050900017.HTML?input=1195m

<표 16-2> 고등학교 교과서에 포함된 신뢰구간 내용

표본평균을 이용하여 모평균을 추정하는 방법에 대하여 알아보자.

모평균 m이 알려져 있지 않고, 모분산 σ^2이 알려져 있는 정규분포를 따르는 모집단에서 임의추출한 크기 n인 표본을 $X_1, X_2, \cdots, X_n$이라고 할 때, 표본평균 $\overline{X}$는 정규분포 $N\left(m, \frac{\sigma^2}{n}\right)$을 따른다.

이때, $\overline{X}$를 표준화한 …

$P(-1.96 \leqq Z \leqq 1.96) = 0.95$ 이므로

$$P\left(-1.96 \leqq \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \leqq 1.96\right) = 0.95$$

$P\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$ 이다. 따라서 모평균 m이

$\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X}+1.96\frac{\sigma}{\sqrt{n}}$ …… ①

일 확률이 0.95임을 알 수 있다.

이 때 ①을 신뢰도 95%인 모평균 m의 신뢰구간이라고 한다.

…

모집단에서 임의추출한 크기 n인 표본 $X_1, X_2, \cdots, X_n$의 관측값을 $x_1, x_2, \cdots, x_n$이라고 하면, 이 관측값의 평균 $\overline{x}$는 표본평균 $\overline{X}$의 하나의 값이므로

$\overline{x}-1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{x}+1.96\frac{\sigma}{\sqrt{n}}$

도 신뢰도 95%인 모평균 m의 신뢰구간이라고 부르기로 한다.

여기서 관측값의 평균 $\overline{x}$는 추출되는 표본에 따라 그 값이 달라지므로 그 신뢰구간도 달라진다. 따라서 '신뢰도 95%인 모평균 m의 신뢰구간'의 의미는 크기 n인 표본의 추출을 무한히 계속하여 모평균 m의 신뢰구간을 만들 때, 이 신뢰구간 중 95%가 모평균 m을 포함한다는 뜻이다.

<표 16-2>를 보면 14장에서 토의한 표본평균의 표본추출분포와 표준오차 등이 이미 이 내용에 앞서 설명되었던 것을 알 수 있다. 다른 말로 하면 본 과정에서 다루고 있는 대부분의 내용이 사실은 고교 교육과정에서 다루었던 내용이었다는 것이다. 이를 생각해본다면 본 과정의 내용들은 일반 시민으로서의 생활을 하는데 필수적인 지식이라는 점을 인정할 수밖에 없다.

❷ 국가통계보고서

다음은 앞에서 여러 번 소개한 서울서베이 보고서에 포함된 100점 만점의 행복점수에 관한 조사결과이다. 서울시민 전체의 평균 행복점수가 약 72.03을 공표하면서 그에 대응하는 주요 통계치들을 신뢰구간과 함께 제시하고 있다. 여기서 'Mean'은 평균이고 'Std Error of Mean'은 표본평균의 표준오차, 그리고 '95% CL for Mean'이 95% 신뢰수준에서의 신뢰구간을, 마지막으로 'Std Deviation'은 표준편차이다.

[그림 16-1]
서울서베이 보고서 – 행복점수 신뢰구간

부록1. (25)주요 변수별 표준 오차 : 행복점수(100점)

전체	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
전체	72.0263	0.05844	71.91180	72.14089	12.46513

주택유형	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
단독주택	71.5300	0.09551	71.34603	71.71398	12.70662
아파트	72.6609	0.09027	72.48049	72.84137	12.37263
다세대 주택	71.5279	0.15478	71.22470	71.83106	12.25527
연립/기타	72.0226	0.22516	71.58819	72.45691	11.78920

구	Mean	Std Error of Mean	95% CL for Mean(하한, 상한)		Std. Deviation
종로구	71.2896	0.36846	70.72136	71.85781	9.64147
중 구	69.3490	0.46516	68.70564	69.99243	10.71260
용산구	70.7542	0.36417	70.12941	71.37900	11.54639
성동구	72.4267	0.32674	71.84355	73.00984	12.06710
광진구	70.4823	0.32274	69.84631	71.11828	13.55077
동대문	69.4651	0.29031	68.88663	70.04363	11.89062
중랑구	73.7840	0.25514	73.26532	74.30259	11.50360
성북구	71.0006	0.27540	70.41285	71.58839	12.88095
강북구	70.0191	0.34223	69.37792	70.66029	13.68504
도봉구	69.8225	0.25681	69.33783	70.30712	10.54884
노원구	70.1759	0.22608	69.66420	70.68758	11.86353
은평구	69.3567	0.24833	68.84856	69.86486	11.88936
서대문	74.6429	0.28808	74.12481	75.16090	10.91372
마포구	74.6946	0.26459	74.19503	75.19409	10.89963
양천구	72.8077	0.20926	72.39255	73.22288	9.56238
강서구	70.6942	0.30630	70.04086	71.34760	15.86489
구로구	73.4708	0.25980	72.97634	73.96526	11.29621
금천구	64.2755	0.54613	63.32865	65.22227	18.28013
영등포	72.5485	0.24608	72.05706	73.03999	10.24390
동작구	73.2403	0.28751	72.66787	73.81267	12.70627
관악구	68.2057	0.29788	67.60205	68.80939	14.76766
서초구	76.2605	0.22897	75.79545	76.72550	9.96757
강남구	77.0154	0.24909	76.51958	77.51125	11.93556
송파구	72.9937	0.17495	72.62173	73.36568	9.07401
강동구	77.1461	0.25019	76.67181	77.62040	11.13134

인터넷의 위키백과에서는 신뢰구간을 다음과 같이 쉽게 설명하고 있다.

> 통계학에서 신뢰구간(confidence interval)은 모수가 어느 범위 안에 있는지를 확률적으로 보여주는 방법이다.
>
> 신뢰구간은 보통 표본에서 산출된 통계와 함께 제공된다. 예를 들어, "신뢰수준 95%에서 투표자의 35%~45%가 A후보를 지지하고 있다."라고 할 때 95%는 신뢰수준이고 35%~45%는 신뢰구간이며 모수는 A후보의 지지율이다.
>
> https://ko.wikipedia.org/wiki/%EC%8B%A0%EB%A2%B0%EA%B5%AC%EA%B0%84

이상과 같이 신뢰수준, 신뢰구간은 우리 생활에 아주 가까이 와 있는데도 불구하고 통계작성에 관여하는 사람들마저도 어려운 전문적 개념으로 치부하여 멀리하는 현상이 있음을 부인할 수 없다. 따라서 다음 절에서는 신뢰수준과 신뢰구간의 핵심개념에 대하여 토의하고자 한다.

16-2. 추정치와 오차

학습목표

- **모집단의 모수를 추정하는 과정에서 발생하는 표본오차의 배경개념을 전달하여 추정치와 오차를 연결하는 안목을 강화한다.**

1 모비율은 얼마인가?

다음은 어느 자치단체에서 일어났던 사건을 변형하여 우리의 토의 주제로 만들어 놓은 가상 상황이다.

> 유권자 100만인 A 지방자치단체의 정책팀에서는 자체 팀에서 수립한 정책의 주민 지지도가 과반수가 되리라고 판단하고, 이를 검정하기 위하여 지역 유권자(성인)중 400명을 임의로 추출하여 그들에게 지지여부를 조사한 결과 210명이 지지한 것으로 나타났다. 이에 정책팀에서는 주민의 과반수가 지지하는 것으로 간주하여 정책의 집행단계로 넘어가려고 하고 있다.
>
> 한편 이 정책의 집행을 반대하는 시민단체에서는 조사 응답자의 수가 400명 뿐이기 때문에 주민의 과반수가 지지했다고 볼 수 없다는 반박문을 내면서 다른 400명을 추출해서 다시 조사 하겠다고 한다.

대부분의 사람들은 이 조사결과 나타난 지지율 p 값을

$$\frac{210}{400}=0.525$$

로 이 지역 주민의 지지율을 추정하는 것을 자연스럽게 받아들일 것이다. 그러나 적지 않은 사람들은

"정말 전체 100만 주민의 지지율이 52.5%일까?"

"얼마나 차이가 날까?"

라는 질문을 하게 될 것이다.

이 질문 앞에서 사람들은 대부분 다음과 같은 3가지 대응을 하게 된다.

▶유형 1: 우리가 얻은 추정치 0.525(52.5%)와 100만 주민의 참 지지율 p를 연계하는 것은 말이 안 된다고 보는 유형으로 지금까지 토의한 여러 지식과는 상관없이 100만 명 중에 400명을 뽑아서 그들의 지지율로 100만 명의 지지율을 추측하는 것을 전혀 믿지 않는 사람들이다. 지금은 우리 주위에서 이런 유형을 보이는 사람들은 거의 없을 것이라고 생각한다. 그렇지만 조사결과가 자기가 기대한 값과 차이가 많이 나는 상황이 벌어지면 생각이 달라진다. 엉터리다, 말도 안 된다 등의 반응이 나타난다.

▶유형 2: 이 유형을 보이는 사람들의 특성은 유형 1의 사람들과 전혀 반대로서 14장에서 토의한 표본비율의 표본추출분포가 중심극한정리에 의해서 전체주민의 지지율 p를 평균으로 하는 정규분포에 가까운 분포를 보일 것이라는 사실을 받아들이는 사람들이다. 그래서 이들은 임의표본으로 뽑힌 400명의 지지율은 참 지지율 p근처에서 나올 확률이 높기 때문에 표본비율 52.5%가 참지지율 근처의 값일 것이라고 확신을 갖는 사람들이다.

▶유형 3: 이 유형은 유형 1과 유형 2의 중간쯤 되는 유형이다. 이 유형의 사람들 생각은 이러하다.

- 표본을 추출할 때는 여하히 과학적 방법으로 전체주민을 잘 대표하는 표본이 되도록 노력하였으리라 믿는다.
- 그러나 혹시 추출된 표본에 우연히도 반대하는 사람들이 다소 많이 포함되어 있을 가능성을 완전히 배제할 수 없다고 생각한다.
- 그래서 이들은 전체 주민의 지지율은 52.5%보다 클 수도 있다고 생각한다(52.5%는 모비율 p보다 작음).
- 그런데 여기서 한 가지를 더 생각한다. 그렇다고 추출된 표본에 이 지역의 모든 반대자들이 다 들어가고 표본으로 추출되지 않은 모든 사람들은 다 지지자일 것이라고까지 생각하는 극단적인 사람들은 아니다. 따라서 표본비율 52.5%가 가능한 표본 비율 값들 중에서 제일 작은 값이라고는 생각하지 않는 사람들이다.
- 그리고 또 반대로도 생각해 보는 균형 감각이 있는 사람들로서 혹시 추출된 표본에 우연히도 지지하는 사람들이 다소 많이 포함되어 있을 가능성을 완전히 배제할 수 없다고 생각한다.
- 그래서 이들은 전체 주민의 지지율은 52.5%보다 작을 수도 있다고 생

각한다(52.5%는 모비율 p보다 큰 값임).

- 그렇다고 해서 추출된 표본에 이 지역의 모든 지지자들이 다 들어가고 표본으로 추출되지 않은 모든 사람들은 다 반대자일 것이라고까지 생각하는 극단적인 사람들은 아니다. 따라서 52.5%는 가능한 표본비율들 중에서 제일 큰 값이라고까지는 생각하지 않는 사람들이다.
- 결론적으로 추출된 표본으로부터 얻은 52.5%는 전체 주민의 지지율 p와는 다를 수 있다는 것을 인정하되 극단적으로 차이가 엄청나게 난다고 생각하지는 않는 유형이다.

위의 3가지 유형 중에서 여러분은 어느 유형의 반응을 하고 있는지 생각해보기 바란다.

2 표본비율의 표준오차 필요

위에서 제시한 3개의 유형은 모두 다 표본비율의 불확실성을 이야기하는 것으로 표본으로 추출된 400명이 어떠한 개체들로 구성되는지에 따라 표본비율과 참 지지율 는 차이가 조금 날 수도 있고 많이 날 수도 있어서 이에 대한 입장의 차이를 나타낸다. 결국 추출된 400명을 구성한 개체들이 누구냐에 따라서 추정치가 달라지기 때문에 생긴 문제인 것이다.

임의표본으로 추출되는 400명의 지지율은 어떠한 값들을 갖는지를 알아야한다. 즉 앞에서 제시한 질문으로 돌아가자.

"정말 전체 100만 주민의 지지율이 52.5%일까?"

"얼마나 차이가 날까?"

을 해결하려면 임의표본으로 추출되는 400명의 표본비율의 표본추출분포를 구하여야 한다. 그런데 이 분포는 중심극한정리에 의해서 평균이 p이고 표준오차가 $\frac{\sigma}{\sqrt{n}}$로 여기서 σ는 확률변수 X의 표준편차인 $\sqrt{p(1-p)}$가 되는 것을 배운 바 있다. 그런데, 여기서 p를 모르기 때문에 p의 추정값을 사용해보자.

이를 기술하면 표본비율의 표준오차는

$$\frac{\sqrt{\text{전체 주민 지지율}(1-\text{전체 주민 지지율})}}{\sqrt{\text{표본의 크기}}}$$

이다. 여기서 전체주민 지지율 대신 추출된 표본에서 얻은 표본 지지율 0.525로 바꾸면

$$\frac{\sqrt{0.525(1-0.525)}}{\sqrt{400}}=\frac{0.4994}{20}=0.025$$

가 되는데 이 값을 표본비율의 표준오차의 추정치로 받아들이자.

> 표본비율의 표준오차는 지금 표본으로 뽑힌 400명이 아닌 다른 400명을 뽑았을 때 얻을 수 있는 표본비율 값들이 얼마나 다양한지(퍼져있는지)를 나타내는 값으로 표본비율의 표준편차를 의미한다.

❸ 표본비율은 어떠한 값들이 나올까?

여기서 잠깐 12장에서 정규분포를 토의할 때 학습한 내용 한 가지를 복습해 보자. 평균이 0.5이고 표준편차가 0.025인 정규분포를 따르는 모집단에서 측정 되었으리라 생각하는 측정값 20개를 써보자.

다음의 수치를 검토해 보고 이상한 수치가 있다면 지적해 보자.

0.458	0.505	0.530	0.548	0.524
0.515	0.518	0.540	0.501	0.539
0.480	0.497	0.486	0.511	0.505
0.514	0.536	0.502	0.500	0.477

전체적으로 정규분포를 따르는 집단에서 추출된 느낌이 든다. 왜냐하면 20개 중에서 70%인 14개가 평균 0.5에서 한배의 표준편차 이내인 (0.475, 0.525) 사이에 있고 나머지는 2배의 표준편차 안에 있는 수치들로 구성되어 있기 때문이다.

그렇다면 평균이 0.5가 아니라 우리가 모르는 p 라고 한다면 어떤 값들이 나올 것이라고 생각되는가?

위의 수치를 사용한다면 p에서 0.025 떨어진 ($p-0.025$, $p+0.025$) 사이에

서 70%쯤 나오고 2배 표준편차 떨어진 ($p-0.05,\ p+0.05$) 사이에서 95% 쯤 나올 것이다.

그렇다면 우리가 토의 중인 400명의 표본 지지율 0.525는 위에서 생각해 $\hat{p}$ 본 가능한 표본비율 중의 하나인데 어디쯤에서 나온 값일까? 그것은 아마도 참값 p 에서 크게는 0.05만큼 떨어진 값일 수도 있고 아니면 p에서 아주 가까운 값일 수도 있을 것이다.

이 절에서는 여기까지 하는 것으로 하고 다음절로 넘어가자.이 절에서는 여기까지 하는 것으로 하고 다음절로 넘어가자.

16-3. 신뢰수준과 신뢰구간

학습목표

- **신뢰수준에 대한 느낌과 신뢰구간을 함께 생각하도록 하여 배경 개념을 이해한다.**

1 "어느 정도 떨어질 수 있다"의 "어느 정도"는?

표본비율의 표준오차 0.025를 구했는데 이 값과

"정말 전체 100만 주민의 지지율이 52.5%일까?"

"얼마나 차이가 날까?"

라는 질문과의 관계를 토의해보자.

먼저 우리는 유형 3의 입장을 택하기로 한다. 왜냐하면 유형 1은 너무 큰 오차까지 생각하여 조사결과에 대한 신뢰가 전혀 없는 사람들이고, 유형 2는 오차가 없는 것으로 스스로 판단하고 주장하는 사람의 유형이다. 반면 유형 3의 사람들은 실제로 얻은 표본비율 0.525는 참 주민지지율 p와 아주 가까운 값일 수도 있지만 어느 정도 떨어진 값일 수도 있다는 것을 함께 생각하겠다는 것이다. 그런데 유형 3에 속하는 사람은 추정치를 발표할 때 자기가 생각하는 어느 정도 떨어질 수 있다는 자신의 생각을 함께 표현해야 한다.

즉 자신이 가지고 있는 극단(어느 정도를 넘어서는)에 대한 기준과 함께 말해야 하는 부담이 있다. 먼저 결론을 말한다면 여러 사람들이 받아들이는 기준(이유는 없다)은 양끝 2.5%, 즉 상위 2.5%와 하위 2.5%이다. 이 생각을 다시 쓰면,

1. 혹시 추출된 표본에 반대자들이 많이 포함되어서 52.5%가 보다 작긴 작다고 해도 조사된 표본비율이 가능한 모든 표본비율 중에서 하위 2.5%를 나타내는 1.96배의 표준오차(1.96*0.025=0.049)이상 작지는 않을 것이라고 믿는다는 입장이다.
2. 반대로 생각할 때는 52.2%가 p보다 크긴 크다고 해도 조사된 표본비율이 가능한 모든 표본비율 중에서 상위 2.5%를 나타내는 1.96배의 표준오차, 0.049이상 크지는 않을 것이라고 믿는다는 입장이다.

두 생각을 합하면 p는 0.476(=0.525-0.049)보다는 크고 0.574(=0.525+

0.049)보다는 작다고 믿겠다는 것이다. 이 때 이 사람의 믿음의 정도를 "어느 정도"를 생각할 때 사용한 극단의 기준 상하위 2.5%로부터 계산된 95%를 이용하여, 신뢰수준 95%라고 표현한다. 다시 말하면 모비율 p는 우리가 표본으로부터 얻은 52.5%와 차이가 나긴 날 것인데 그 차이는 최대 1.96*2.5%포인트 보다는 작을 것이라고 95% 믿는다는 것이다.

❷ 신뢰구간에 대한 전통적 해석

신뢰구간에 대하여 1절에서 소개한 고등학교 확률과 통계 교과서에서 설명한 것과 같은 전통적 방법을 살펴보기 위해 모비율이 아닌 모평균의 신뢰구간을 생각해 보자. 동일 부피를 가진 특정물질의 무게의 평균 μ에 대하여, 크기가 64인 표본을 추출하여 얻은 표본평균으로 모평균 μ를 추정하려는 연구자가 있다고 하자. 그는 이물질의 무게의 표준편차 는 0.8kg라고 알려져 있는 것을 받아들였다. 그러나 이 물질의 무게가 정규분포를 따른다고 볼 수 있는지는 아직 확인되지 않았다. 어느 특정 물질의 평균 무게에 대한 95% 신뢰구간을 만들어 보자.

과거 경험에 의한 σ(=0.8kg)와 64개의 무게를 관찰한 결과를 이용하여 신뢰구간을 구해 보자.

여기서는 표본비율이 아니라 표본평균을 주목한다. 표본평균의 표본추출분포는 물질의 무게가 정규분포를 따르지 않더라도 중심극한정리에 의해서 평균이 μ, 표준오차는 $\frac{\sigma}{\sqrt{n}}=\frac{0.8}{\sqrt{64}}=0.1$인 정규분포에 가까워진다. 따라서 크기가 64인 표본의 평균 $\overline{X}$가

$$P\left(-1.96<\frac{\overline{X}-\mu}{0.1}<1.96\right)$$

을 만족시킬 확률은 0.95이다. 미 말의 의미는 크기가 64인 표본의 평균 $\overline{X}$ 중 95%는 $(\overline{X}-1.96*0.1<\mu<\overline{X}+1.96*0.1)$을 만족한다는 뜻이다. 이를 또 다시 쓰면 크기가 64인 표본의 평균 $\overline{X}$중 95%의 표본평균은 다음의 구간 $(\overline{X}-0.196,\ \overline{X}+0.196)$안에 모평균 μ를 포함한다. 즉 어떤 사람이 이 물질 집단에서 크기가 64인 표본을 100회 뽑았다고 상상해보자(실제 상황에서는 이런 일은 결코 일어나지 않는다). 그러면 100개의 표본평균이 얻어질 것이고 그 값이 50kg, 40kg, …, 55kg이었다고 생각해보자. 그러

면 각각의 표본평균으로부터 μ를 포함할 것이라 생각하는 구간이 100개 산출된다. 이론적으로는 이렇게 산출된 100개의 구간 중에서 95% 정도의 구간은 모평균 μ를 포함한다는 뜻이다.

실제 상황을 생각해보자. 64개로 구성된 하나의 표본으로부터 계산된 평균 무게가 50kg이라면 표준오차 0.1을 대입한 구간은 (50－0.196, 50＋0.196), 즉 (49.804, 50.196)이 된다. 이 구간을 신뢰수준 95%에서의 신뢰구간이라고 하는데 그 의미는 이 구간에 μ가 포함되었다고 믿어도(기대하여도) 된다는 뜻이다. 그러나 이 구간이 μ를 실제로 포함하고 있는지는 아무도 모른다.

64개의 평균 무게가 40kg이었다면 우리는 신뢰수준 95%에서 μ가 신뢰구간 (39.804, 40.196)에 포함되었다고 믿어도 된다고 발표하게 된다. 여기서도 이 구간이 실제로 μ를 포함하고 있는지는 알 수 없다.

그렇다면 이때 믿음의 근거는 무엇일까?

믿음의 근거는 크기가 64인 수많은 표본의 평균들 중 95%의 표본평균은 이 표본평균으로부터 계산되는 구간($\overline{X}-0.196,\ \overline{X}+0.196$)에 모평균 μ가 포함된다는 이론적 사실이다.

그러나 어느 표본평균으로 계산되는 구간이 μ를 포함할지는 모른다.

우리가 제시하는 구간은 수많은 구간 중에 하나일 뿐이다.

따라서 우리가 제시한 구간에 μ가 포함되었을지 아닐지는 모른다.

다만 95%의 믿음을 가지고 의사결정을 한다는 뜻이다.

참고자료

- 이부일 · 신지은 · 박영옥 · 이석훈(2007), 엑셀을 활용한 통계자료분석 – 기초편, 경문사.
- 이준열 · 최부림 · 김동재 · 한대희 · 전용주 · 장희숙 · 조석연 · 조성철 · 황선미 · 박성준(2014), 고등학교 확률과 통계, 천재교육.
- 이석훈(2006), 통계적 사고방식, 통계교육원.
- 이석훈(2015), 통계기초 및 활용교재, 통계교육원.
- 서울특별시(2014), 서울서베이.
- http://www.jeonguknews.co.kr/news/articleView.html?idxno=14995
- http://www.yonhapnews.co.kr/bulletin/2015/11/26/0200000000AKR20151126050900017.HTML?input=1195m
- https://ko.wikipedia.org/wiki/%EC%8B%A0%EB%A2%B0%EA%B5%AC%EA%B0%84

제 17 장

표본오차 활용하기

17-1. 표본오차 개념 이해하기

학습목표

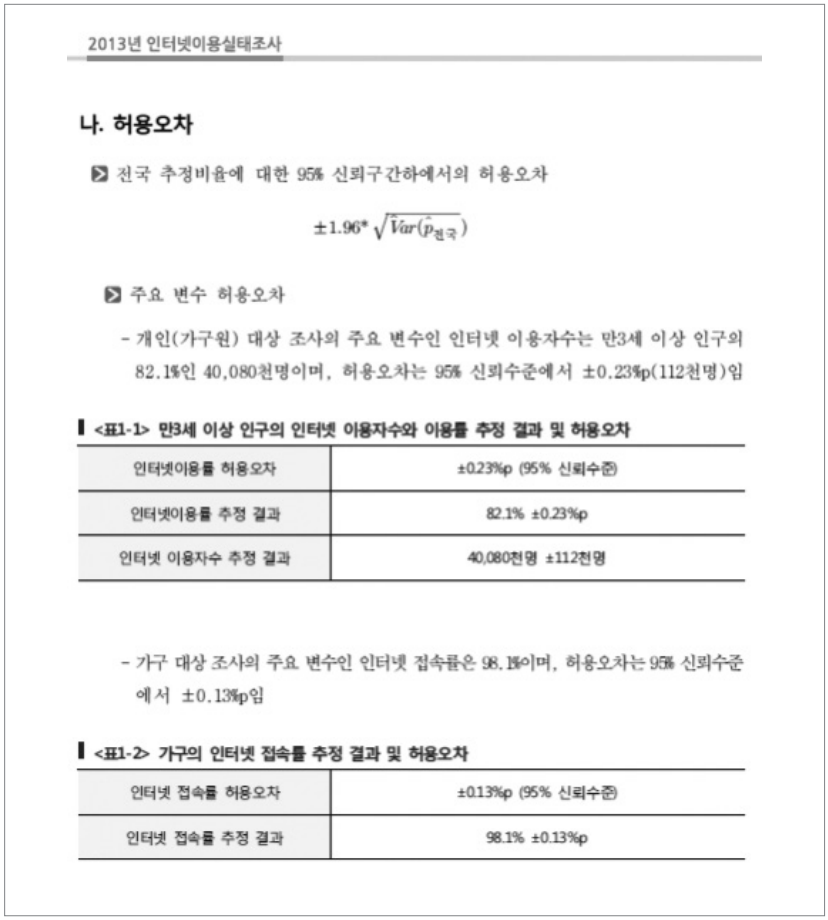
- **통계학을 전혀 모르는 연수생들은 언론매체에서 대단히 자주 접하는 표본오차라는 용어에 대한 두려움을 없애고, 허용오차, 오차의 한계 등을 접했던 연수생들 중 용어에 대한 혼란이 있었던 사람들은 개념을 정리한다.**

1 추정치의 정확성에 관한 정보

우리가 일반적으로 발견하는 표본조사 결과의 발표 보고서나 언론의 기사내용에는 다음과 같은 내용이 있다.

사례 1) 2013년 인터넷이용실태조사 조사개요

[그림 17-1]
인터넷 이용실태조사 개요 (일부)



2013년 인터넷이용실태조사

나. 허용오차

▶ 전국 추정비율에 대한 95% 신뢰구간하에서의 허용오차

$$\pm 1.96^{*}\sqrt{Var(\hat{p}_{\text{전국}})}$$

▶ 주요 변수 허용오차

- 개인(가구원) 대상 조사의 주요 변수인 인터넷 이용자수는 만3세 이상 인구의 82.1%인 40,080천명이며, 허용오차는 95% 신뢰수준에서 ±0.23%p(112천명)임

<표1-1> 만3세 이상 인구의 인터넷 이용자수와 이용률 추정 결과 및 허용오차

인터넷이용률 허용오차	±0.23%p (95% 신뢰수준)
인터넷이용률 추정 결과	82.1% ±0.23%p
인터넷 이용자수 추정 결과	40,080천명 ±112천명

- 가구 대상 조사의 주요 변수인 인터넷 접속률은 98.1%이며, 허용오차는 95% 신뢰수준에서 ±0.13%p임

<표1-2> 가구의 인터넷 접속률 추정 결과 및 허용오차

인터넷 접속률 허용오차	±0.13%p (95% 신뢰수준)
인터넷 접속률 추정 결과	98.1% ±0.13%p

사례 2) 언론매체

… 또 지난 10~11일 여론조사 전문기관 (주)에스티아이와 <미디어오늘>이 전국 성인 1000명을 대상으로 여론조사를 실시한 결과, 정당지지율은 ○○○당 38.5%, □□□당 26.3%로 ○○○당 지지율이 13.2%포인트나 높게 나왔다. △△△당은 6.0%다.

이 조사의 표본오차는 95% 신뢰수준에서 오차범위 ±3.1%p, 응답률은 4.4%다.

여론조사 기관 '한국갤럽'이 지난 10~12일 사흘간 전국 성인 1012명을 대상으로 실시한 여론조사 결과 역시 비슷했다. … 이 조사의 표본오차는 ±3.1%p(95% 신뢰수준), 응답률은 20%(총 통화 5,069명 중 1,012명 응답 완료)다.

(http://www.siminilbo.co.kr/news/articleView.html?idxno=423147)

위의 사례에서 허용오차, 표본오차, 신뢰수준, 신뢰구간, 오차범위 등의 용어가 나타난다. 이 절에서는 이러한 용어의 배경개념을 토의한다. 유형이건 무형이건 무엇이건 생산하여 제품을 제공하는 사람은 그 제품에 대한 품질 –정확성, 신뢰성, 유용성 등 – 에 대한 정보를 함께 제공하여야 한다. 표본조사를 통하여 관심 있는 모집단의 특성(관심 모수)에 대하여 추정을 하려는 사람이 가장 중요하게 생각해야 하는 것은 자신이 제시하게 되는 추정치의 정확성이다. 다시 말하면 제시하는 추정치가 얼마나 모수에 가까운 값인가에 대한 정보를 동시에 전해야 한다.

그러나 이 정확성에 관한 정보를 얻는다는 것은 대단히 어렵다. 사실은 불가능하다. 그 이유는 다음과 같이 생각해 볼 수 있다.

첫째는 모수를 제시하는 자나 제시 받는 자나 모수의 참값을 아무도 모르기 때문이다.

둘째는 발표하게 될 추정치는 모든 개체들을 측정하여 얻은 것이 아니고 표본으로 추출된 모집단의 일부 개체들의 측정치로부터 얻어진 것이기 때문에 어떤 개체들이 표본으로 추출되느냐에 따라서 달라질 수밖에 없다. 그래서 어느 경우에는 추정치와 모수가 상당히 일치할 경우도 있을 것이지만 우리는 상황적으로 추정치와 모수는 차이가 있을 것이라고 생각하는 것이다. 이 차이를 조사방법론에서는 표본오차, 표본추출오차, 표집오차라고 하는데 이 오차는 표본을 추출하여 조사하기 때문에 불가피한 오차이며, 표본추출방법, 추정방법 등과 밀접한 관련이 있다.

셋째는 조사과정에서 발생할 수 있는 많은 문제들이 측정치에 영향을 준다는 것이다. 확률추출을 제외한 조사, 집계 등 조사실시의 모든 과정에서도 작지 않은 오차가 발생한다는 것이다.

예를 들어 표본추출틀(임의 표본을 추출하기 위한 모집단 개체들의 목록) 또는 모집단 명부의 부정확성, 조사기획 단계에서의 오차, 무응답오차(면접접근 불능, 응답자 면접불능, 응답자 비협조, 응답거부 등 무응답 발생), 자료수집 단계에서의 오차 등에 의하여 발생한다. 참고로 이러한 오차를 비표본오차, 비표집오차라고 부른다.

❷ 표본오차의 표현

사례 2)에서 볼 수 있듯이 언론매체, 신문방송 등을 통하여 제시되는 조사

결과 관련 기사에서는 "표본오차는 95% 신뢰수준에서 오차범위 ±3.1%p …", "이 조사의 표본오차는 ±3.1%p(95% 신뢰수준)"라고 발표하고 있다. 이 발표문에서는 크기가 1,012명인 표본의 표본오차를 16장에서 학습한 두 가지 개념 95% 신뢰수준과 신뢰구간을 계산하는 신뢰구간의 거리(폭)를 이용하였다.

1. 3.1%포인트의 유도

3.1%포인트는 다음과 같이 계산된다.

1) 신뢰수준 95%에 대응하는 표준정규분포의 상위 2.5%값인 1.96을 구한다.

2) 모비율 대신 0.5를 사용하여(2절에서 설명함) 다음과 같이 표본비율의 표준오차의 추정치 0.0518을 구한다.

$$\text{표준오차} = \frac{\sqrt{0.5(1-0.5)}}{\sqrt{\text{표본의 크기}}} = \frac{0.5}{\sqrt{1000}} = 0.0158$$

3) 1)과 2)에서 구한 두 개의 값을 곱한다.

$$1.96 \times \text{표준오차}(0.0158) = 0.0309 \ (\text{약 } 3.1\%)$$

2. 허용오차(오차의 한계)

위에서 설명한 바와 같이 3.1%포인트를 엄격히 말하면 "신뢰수준 95%로 진술하는 사람이 자신의 추정치가 참 값과 가장 멀리 떨어진 정도가 3.1%포인트이다"라는 뜻이다. 이러한 맥락에서 3.1%포인트를 신뢰수준 95%에서 허용오차 또는 오차의 한계라고 한다. 그리고 표본조사에서 비롯된 불가피한 오차인 표본오차는 모수의 추정치에 대하여 신뢰수준과 허용오차로서 정량화되는 것이다.

3. 예제

1) 표본크기가 100인 표본으로부터 모집단의 모비율을 추정하고자 한다. 이때의 표본오차를 제시하라.

제시 1) 신뢰수준 95%에 해당되는 상위 2.5% 값은 1.96이고 크기가 100인 표본비율의 표준오차 추정치는 다음과 같다.

$$\frac{\sqrt{(0.5)(1-0.5)}}{\sqrt{100}} = 0.05$$

이 두 개의 값을 이용하여 허용오차를 계산하면

$$1.96 * 0.05 = 0.098$$

이 된다. 따라서 신뢰수준 95%에서 허용오차는 9.8%포인트(0.098)이다.

제시 2) 신뢰수준 90%에 해당되는 상위 5% 값은 1.645이므로 신뢰수준 90%에서 허용오차는 8.2%포인트(1.645*0.05=0.082)이다.

제시 3) 신뢰수준 68%, 상위 16%의 값은 1이 되므로 신뢰수준 68%에서 허용오차는 5%포인트(1.0*0.05=0.05)이다.

17-2. 허용오차와 표본크기

학습목표

- **이 절에서는 표본비율의 표준오차가 추정치와 상관없이 주어진다는 점을 이용하여 사전에 정해진 신뢰수준과 허용오차로부터 표본크기가 산출되는 과정을 토의하고 표본크기 결정에 요구되는 정보와 그 결정방법을 습득한다.**

❶ 표본크기는 클수록 좋다

1. '좋다'의 의미

1절에서 토의한 것처럼 표본조사를 하기로 결정한 사람에게 가장 중요한 목표는 추정치를 얼마나 모수에 가깝게 그리고 얼마나 신속하고 경제적으로 구하느냐이다. 그러나 이 목표를 달성하는 것은 대단히 어렵다. 보다 정확히 하려면 표본의 크기를 늘릴수록(전수조사를 하면 표본오차는 0이다) 좋은데 그렇게 하면 조사시간은 길어지고 비용은 늘어나게 되기 때문에 이를 조정하는 작업이 필요하다.

2. 조정 작업에 필요한 정보

조정 작업은 정확성을 나타내는 신뢰수준과 허용오차를 통하여 이루어진다. 신뢰수준이 높으면 높을수록 허용오차가 커지게 되고 허용오차가 커지면 커질수록 추정치가 참값과 차이가 커지는 경우를 허용하기 때문에 참값에 대한 범위가 넓어져서 의사결정에 불확실성이 커지게 된다. 예컨대, 1절의 예에서 표본비율이 70%였다면 68% 신뢰수준에서의 폭은 5%이고, 90% 신뢰수준에서의 폭은 8.2%, 95% 신뢰수준에서의 폭은 9.8%가 되어 구간의 폭이 점점 더 커지게 된다.

❷ 표본크기 결정

1. 표본크기 계산

지금까지 표본비율의 표준오차는

$$\text{표본비율의 표준오차} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

으로 배웠다. 그런데 이 식을 사용하려면 모집단의 비율(p)을 알아야 하

거나 표본으로부터 표본비율을 알아야만 한다. 그렇지만 이 절에서는 아직 표본을 추출하지 않은 상태에서 표본크기를 구하려는 시점이기 때문에 위 식을 사용할 수 없다. 그래서 표본비율의 표준오차를 계산할 때 추정된 표본비율대신 0.5를 사용한다. 신뢰수준 95%에서 허용오차는 추정치와 상관없이 표본의 크기가 n이라면,

$$1.96^* \frac{\sqrt{(0.5)(1-0.5)}}{\sqrt{n}} = 1.96^* \frac{0.5}{\sqrt{n}}$$

이다.

따라서 허용오차를 B라고 하면

$$n = (1.96)^{2*} \frac{(0.5)^2}{B}$$

가 되어 n과 B의 관계식이 생긴다. 이 관계식을 정리해보면 다음과 같다.

<표 17-1> 표본비율로 모비율을 추정할 때 95% 허용오차

신뢰수준 95% 허용오차 1.96*표준오차(B)	표본의 크기 (n)
0.049	400
0.031	1000
0.025	1600
0.021	2200
0.019	2800

그런데 이 n과 B의 관계식의 특징을 한 가지 발견할 수 있다.

표본의 크기가 커지는데 비해서 허용오차는 그렇게 많이 작아지지 않는다. 구체적으로 보면, 표본크기는 400에서 1,000으로 600이 늘었을 때 허용오차는0.049-0.031=0.018 정도 줄었다. 그러나 표본의 크기가 1,000에서 1,600으로 600이 늘었을 때에는 400에서 1,000으로 늘어났을 때와 비교하여 볼 때 허용오차는 0.031-0.025=0.006 밖에 줄지 않았다. 이 현상은 표본크기가 2,200 또 2,800으로 늘어나면서 더욱 두드러진다.

표본크기가 1,600에서 2,200으로 600이 늘었을 때 허용오차의 감소는 0.025-0.021=0.004이고 표본크기가 2,200에서 2,800으로 600이 늘었을 때의 허용오차 감소는 0.021-0.019=0.002이다.

결국 똑같이 600씩 늘어났지만 허용오차가 줄어드는 정도는 다르다는 뜻이다. 처음에는 약 2%p 줄었지만 그 다음부터는 0.6%p, 0.4%p, 0.2%p정도 줄어들었다. 이러한 특징은 표본의 크기를 600명 늘렸을 때 경비는 상당량 증가하는데 반해서 유익의 정도는 다르다고도 표현할 수 있다. 즉 400에서 600을 늘린 것에 비하면 1,000에서 600을 늘린 것은 별로 유익하다고 할 수 없다. 그래서 일반적인 여론조사가 1,000명, 또는 아주 중요한 조사일 때는 2,000여명을 조사하게 되는 것이다. 그 이상으로 표본크기를 늘린다고 해도 경비가 늘어나는데 비하여 허용오차는 별로 줄어들지 않는다.

2. 전체 모집단 크기와의 관계

여기서 궁금할 수 있는 질문을 하나 생각해보자.

그렇다면 전체 주민 150만명은 표본오차에 영향을 끼치지 않는 것일까? 다시 말하면 전체주민 즉, 모집단의 크기는 표본오차와 관계가 없는 것일까? 엄격하게 말하면 관계가 있다. 지금까지 제시한 표본비율의 표준오차의 공식은 이론적으로 모집단의 크기가 무한이라고 가정하고 산출된 것이다. 그러나 모집단의 크기가 크기는 크겠지만 대부분의 경우 모집단의 크기는 유한이다. 따라서 앞의 표본비율의 표준오차의 식에 이를 고려한 내용이 추가되어야 한다. 구체적으로 말하면 식에 어떤 값을 곱해주어야 하는데 그 값을 알기 위해서는 두 가지 개념이 필요하다. 하나는 전체 모집단의 크기(N)이고, 또 다른 하나는 표본의 크기(n)이다. 전체 모집단에서 표본으로 추출되는 비율은 다음과 같이 나타낼 수 있다.

$$\text{추출율}(f) = \frac{n}{N}$$

이 값을 이용하여 유한 모집단 수정계수(fpc : finite population connection)를 다음과 같이 구한다.

$$\frac{N-n}{\text{전체}(N)} = (1-f)$$

수정계수 fpc가 구해지면 이 값의 제곱근 $\sqrt{fpc}$를 우리가 지금까지 사용한 표준오차에 곱해주면 된다. 여기서는

$$fpc = \frac{N-n}{N} = 1-f = 0.99933$$

가 되어 결국 $\sqrt{fpc} = 0.99966$이 되는데 이 값은 거의 1에 가깝게 되어 우

리가 사용했던 표준오차에 곱하여도 거의 영향을 주지 않게 된다. 따라서 모집단이 어느 정도 커서 $\sqrt{fpc}$가 1에 가까우면 모집단의 크기는 무시해도 좋다는 말이 된다.

17-3. 상대표준오차(RSE) 이해하기

학습목표

- 대부분의 조사통계보고서에 나타나는 상대표준오차의 정확한 의미를 파악하고 특히 CV와 혼용하는 과정에서 모집단의 CV와 혼돈되는 것을 명확히 한다.

1 상대표준오차의 필요성

1. 상대표준오차 만나기

상대표준오차는 다음과 같이 조사통계 보고서의 품질을 나타내는 핵심적인 정보로서 다음과 같이 여러 보고서에서 발견할 수 있다.

(1) 생활시간조사보고서 - 10세 이상 성별 평균시간의 표본오차(요일평균, 평일)

요일 평균의 추정치와 추정치의 표준오차 그리고 상대표준오차가 수록되어 있다.

<표 17-2> 생활시간조사 보고서 – 10세 이상 성별 평균시간의 표본오차

행동분류		요일평균(Average for 7-Day Week)								
		전 체			남 자			여 자		
		추정치	표준오차	상대표준오차	추정치	표준오차	상대표준오차	추정치	표준오차	상대표준오차
개인유지		11:14	1.1	0.2	11:11	1.4	0.2	11:16	1.2	0.2
A1	수면	7:59	0.9	0.2	7:59	1.0	0.2	7:59	1.1	0.2
A120	수면	7:58	0.9	0.2	7:58	1.0	0.2	7:58	1.1	0.2
A140	잠 못 이룸	0:01	0.1	6.4	0:01	0.1	7.9	0:02	0.1	6.9
A2	식사 및 간식	1:56	0.6	0.5	1:59	0.7	0.6	1:54	0.6	0.6
A220	식사	1:26	0.5	0.5	1:27	0.5	0.6	1:24	0.5	0.6
A240	간식 · 음료	0:31	0.4	1.2	0:31	0.4	1.3	0:30	0.4	1.3
A3	개인 건강관리	0:05	0.2	4.0	0:04	0.3	6.8	0:07	0.3	4.1
A320	자기치료	0:01	0.1	5.5	0:01	0.1	7.5	0:02	0.1	6.0
A340	아파서 쉼	0:01	0.1	12.9	0:01	0.2	23.0	0:01	0.2	13.9
A360	의료서비스 받기	0:03	0.1	3.8	0:02	0.1	5.7	0:04	0.2	4.7

(2) 산림청 임산물생산비조사

생활시간조사 보고서에서는 추정치도 수록하였는데 아래의 표에는 지역별로 표준오차와 상대표준오차만을 수록하였다.

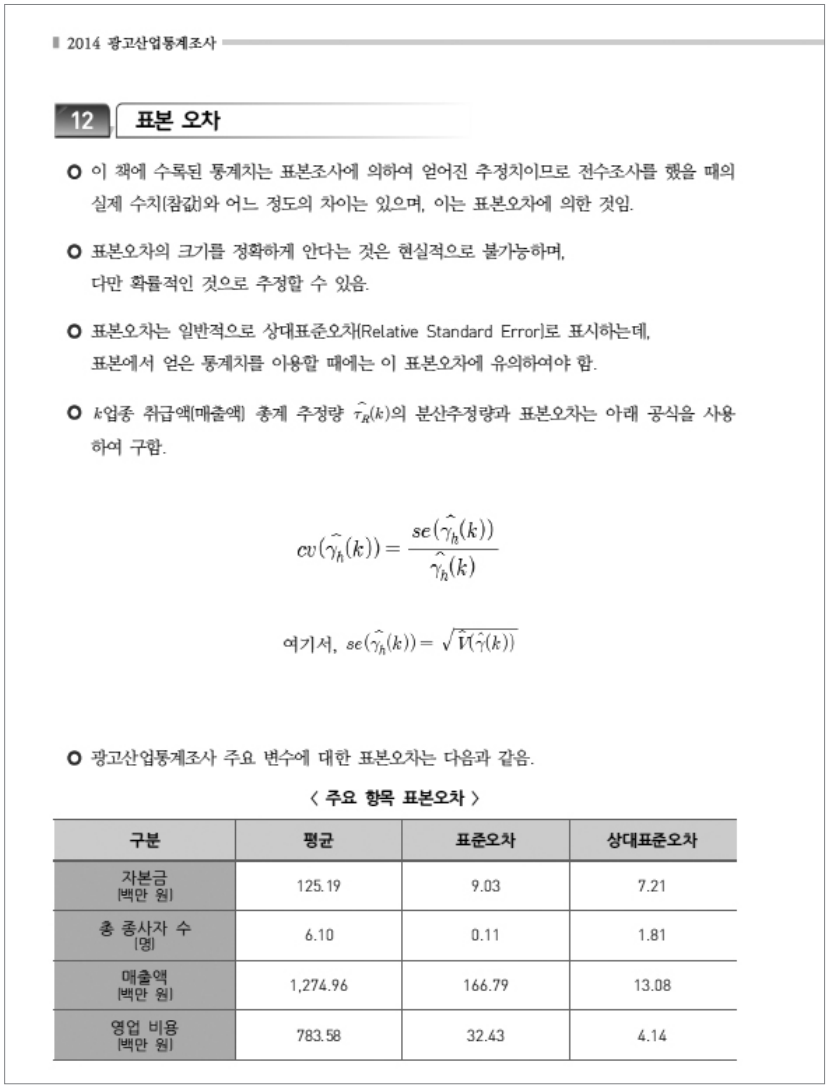
<표 17-3> 산림청 임산물생산비조사 보고서 – 지역별 생산비 표준오차, 상대표준오차

생산비 목별(1)	생산비 목별(2)	2014									
		전국평균		충청남도		전라남도		경상남도		기타	
		표준오차 (원)	상대표준오차 (%)	표준오차 (원)	상대표준오차 (%)	표준오차 (원)	상대표준오차 (%)	표준오차 (원)	상대표준오차 (%)	표준오차 (원)	상대표준오차 (%)
생산비 합계	소계	128,940	3.0	223,128	4.3	312,509	8.5	198,806	6.0	374,928	7.7
직접 생산비	소계	117,952	3.1	204,064	4.4	282,312	8.6	183,363	6.3	343,463	8.0
	비료비	32,704	9.5	72,720	14.9	43,817	17.9	26,807	13.1	77,771	19.9
	농약비	4,660	9.8	10,040	12.0	5,894	48.8	2,240	21.8	19,876	22.2
	수도광열비	4,712	5.4	8,011	7.8	12,200	13.3	7,229	11.3	13,280	13.2
	소농구비	777	10.8	1,611	17.2	1,050	20.0	961	13.9	2,103	99.0
	수리비	5,048	17.2	11,283	24.1	9,433	33.9	3,742	25.3	2,595	43.6
	감가상각비	8,731	5.7	14,971	8.9	22,720	13.2	11,365	9.4	34,200	20.4
	노동비	93,382	3.4	156,272	4.8	222,549	9.7	152,729	7.1	282,687	9.1
	자동차비	7,097	9.3	9,397	12.0	27,934	25.0	10,318	17.7	12,893	20.6
	기타재료비	3,447	31.0	8,171	48.7	3,548	27.3	198	56.1	6,330	30.9
	기타비용	12,820	20.8	26,577	34.3	29,605	50.2	11,233	34.8	28,909	28.4
간접 생산비	소계	16,470	3.4	30,348	5.1	35,911	8.5	23,883	6.5	46,232	8.2
	자가토지용역비	10,058	6.5	22,405	12.1	11,625	9.8	7,892	5.9	27,783	18.3
	임차토지용역시	8,310	9.9	18,226	13.2	8,351	28.6	7,221	22.2	26,317	20.5
	유동자본용역비	3,991	3.4	6,946	4.8	9,439	9.7	6,183	7.1	11,579	8.6
	고정자본용역비	11,276	8.4	17,873	13.6	29,661	16.9	19,519	17.7	28,610	19.0

(3) 광고산업통계조사

이 조사보고서의 조사개요 부분에서는 상대표준오차를 CV와 혼용하여 표시하였다.

[그림 17-2]
광고산업 통계조사
보고서 조사개요
(일부)



2014 광고산업통계조사

12 표본 오차

- 이 책에 수록된 통계치는 표본조사에 의하여 얻어진 추정치이므로 전수조사를 했을 때의 실제 수치(참값)와 어느 정도의 차이는 있으며, 이는 표본오차에 의한 것임.
- 표본오차의 크기를 정확하게 안다는 것은 현실적으로 불가능하며, 다만 확률적인 것으로 추정할 수 있음.
- 표본오차는 일반적으로 상대표준오차(Relative Standard Error)로 표시하는데, 표본에서 얻은 통계치를 이용할 때에는 이 표본오차에 유의하여야 함.
- k업종 취급액(매출액) 총계 추정량 $\hat{\tau}_R(k)$의 분산추정량과 표본오차는 아래 공식을 사용하여 구함.

$$cv(\hat{\gamma}_h(k)) = \frac{se(\hat{\gamma}_h(k))}{\hat{\gamma}_h(k)}$$

여기서, $se(\hat{\gamma}_h(k)) = \sqrt{\hat{V}(\hat{\gamma}(k))}$

- 광고산업통계조사 주요 변수에 대한 표본오차는 다음과 같음.

〈 주요 항목 표본오차 〉

구분	평균	표준오차	상대표준오차
자본금 (백만 원)	125.19	9.03	7.21
총 종사자 수 (명)	6.10	0.11	1.81
매출액 (백만 원)	1,274.96	166.79	13.08
영업 비용 (백만 원)	783.58	32.43	4.14

2. 상대표준오차 정의

상대표준오차를 이론적으로 정의하면 추정량의 표준오차를 추정량의 기댓값으로 나눈 값에 100을 곱하여 퍼센트로 표현한다.

$$\text{상대표준오차} = \frac{\text{추정량의 표준오차}}{\text{추정량의 기댓값}} \times 100$$

그런데 이 값은 실제로 이론적으로 계산되는 값들이기 때문에 현실적으로는 구하기 어렵고 데이터로부터 이 값들을 추정한 상대표준오차의 추정치를 다음과 같이 정의하고 이 값을 일반적으로 상대표준오차라고 발표한다.

$$\text{상대표준오차 추정치} = \frac{\text{추정량의 표준오차의 추정치}}{\text{추정치}} \times 100$$

예를 들면 표본크기가 900인 만족도조사에서 평균만족도 $\bar{x}$가 60이고 표준편차 s가 9였다면 전체 모집단의 만족도 평균의 추정치는 60이 되고 추정량의 표준오차의 추정치는

$$\frac{s}{\sqrt{n}} = \frac{9}{\sqrt{900}} = 0.3$$

이므로 상대표준오차의 추정치는 $\frac{0.3}{60} = 0.05$ 가 된다. 이를 퍼센트로 나타내면

$$\frac{0.3}{60} \times 100 = 5(\%)$$

가 된다.

3. 상대표준오차 유용성

상대표준오차는 단위가 없기 때문에 단위가 다른 두 추정치의 변동을 비교하거나 추정치가 크게 차이가 나는 상황에서 추정치의 표준오차의 비교할 때 유용하다. 예컨대 앞의 생활시간조사 통계표와 임산물생산비조사의 통계표 사례에서 보는 바와 같이 통계를 공표할 때 다수의 추정치가 제시된다. 그런데 이때 이들 추정치들은 단위도 다를 수 있고 크기도 다른 경우가 많다. 따라서 그 추정치의 표준오차만 제시하기 보다는 상대표준오차도 함께 제시하는 것이 변동에 관하여 보다 현실적인 비교를 할 수 있다.

2 변동계수와 상대표준오차

1. 혼용하는 경우

통계학입문에서는 변동계수(CV: coefficient of variation)를 표준편차를 평균으로 나눈 값으로 다음과 같이 정의하면서 이를 특별히 모집단변동계수라고 부른다.

$$cv = \frac{\sigma}{\mu}$$

그리고 크기가 n인 표본의 평균 $\bar{y}$와 표준편차 s가 주어지면 표본변동계수는 다음과 같이 주어진다.

$$\hat{cv} = \frac{s}{\bar{y}}$$

그런데 박홍래 등(2008)의 표본조사방법론 교재에서는 상대표준오차를 「추정량의 변동계수」라는 용어로 사용하는 경우가 있다. 이 용어를 사용하는 교재에서는 모수 θ에 대한 추정량 $\hat{\theta}$의 변동계수를 $cv(\hat{\theta})$의 기호로 표시하며 상대표준오차의 추정치로 정의하고 추정량의 변동계수라고 부른다.

따라서 변동계수라고 하면 사람에 따라서 CV인지 추정량의 변동계수(상대표준오차) $cv(\hat{\theta})$ 인지를 혼돈할 위험성이 생긴다. 특히 현장에서는 두 값 모두 CV라는 용어를 많이 사용한다.

$\frac{s}{\bar{y}}$ 로서 표본변동계수 CV일 수도 있고,

$\frac{(s/\sqrt{n})}{\bar{y}}$ 로서 상대표준오차 CV일 수도 있다.

아래의 중소제조업 직종별 임금조사 보고서(2015.11, 중소기업중앙회)에서는 '상대표준오차 개념인 변동계수'라는 표현으로 모집단 변동계수와의 혼돈을 피하려고 노력하고 있다.

[그림 17-3]
중소제조업 직종별
임금조사 보고서
(일부)

2 표본설계

1. 추출단위(Sampling Unit) : 기업체(Enterprise)

2. 표본틀(Sampling Frame)

○ 통계청의 2013년 기준 광업·제조업통계조사 결과중 종사자 20~299인 중소제조업체를 산업 중류별(22개), 종사자 규모별(3개)로 구분하여 표본 틀 구성

3. 층 화

○ 산업 중분류별(22개)로 1차 층화한 후 이를 다시 종사자규모별(3개)로 2차 층화

① 산업중분류(업종별)	
10. 식료품	23. 비금속 광물제품
11. 음료	24. 제1차 금속
13. 섬유제품	25. 금속가공제품
14. 의복, 의복액세서리 및 모피제품	26. 전자부품, 컴퓨터, 영상, 음향 및 통신장비
15. 가죽, 가방 및 신발	27. 의료, 정밀, 광학기기 및 시계
16. 목재 및 나무제품	28. 전기장비
17. 펄프, 종이 및 종이제품	29. 기타 기계 및 장비
18. 인쇄 및 기록매체 복제	30. 자동차 및 트레일러
20. 화학물질 및 화학제품	31. 기타 운송장비
21. 의료용 물질 및 의약품	32. 가구
22. 고무제품 및 플라스틱 제품	33. 기타 제품
	※ 단, 12. 담배제조업, 19. 코크스, 연탄 및 석유정제품 제외
② 종사자규모	
1규모 (20~49인) 2규모 (50~99인) 3규모 (100~299인)	

4. 표본크기 결정

○ 업종별, 종사자규모별 추정에 대한 표본오차 관리를 위해 업종 · 종사자규모별 목표오차 설정하여 1차 층의 표본크기를 결정

○ 추정량의 정도는 각 업종별 모집단 사업체수 규모에 따라 표본오차의 절대량으로 목표오차를 설정하기 보다는 상대표본오차 개념인 변동계수(Coefficient of Variation: CV)를 사용하여 각 층의 목표오차를 설정

[그림 17-3]
중소제조업 직종별 임금조사 보고서 (일부) (계속)

○ 각 층의 표본크기는 모집단의 중요변수인 기업체 평균급여액을 고려하여 결정

○ 목표오차는 각 산업중분류의 기업체 구성비에 따라서 다르게 적용하며, 산업별 구성비가 높을수록 목표오차를 작게 하여 비중이 높은 산업의 오차를 최소화

〈 산업중분류별 목표오차 〉

(단위: 개, %)

업종	기업체수	목표오차
11. 음료	81	7.5%
21. 의료용 물질 및 의약품	251	
14. 의복, 의복액세서리 및 모피제품	845	7.0%
15. 가죽, 가방 및 신발	301	
16. 목재 및 나무제품	222	
18. 인쇄 및 기록매체 복제	426	
32. 가구	488	
33. 기타제품	373	
10. 식료품	1,852	6.3%
13. 섬유제품	1,343	
17. 펄프, 종이 및 종이제품	690	
20. 화학물질 및 화학제품	1,126	
22. 고무제품 및 플라스틱 제품	2,623	
23. 비금속 광물제품	939	
24. 제1차 금속	1,350	
25. 금속가공제품	3,680	
26. 전자부품, 컴퓨터, 영상, 음향 및 통신장비	2,078	
27. 의료, 정밀, 광학기기 및 시계	950	
28. 전기장비	1,819	
29. 기타 기계 및 장비	4,188	
30. 자동차 및 트레일러	2,401	
31. 기타 운송장비	1,060	

○ 업종·종사자 규모별 표본수 결정 공식

$$n=\left(\frac{C}{C_{\bar{y}}}\right)^2 / \left[1+\frac{1}{N}\left(\frac{C}{C_{\bar{y}}}\right)^2\right]$$

여기서, $C=S/\overline{Y}$: 모집단 변동계수(평균급여액)

$C_{\bar{y}}$: 목표오차

N : 업종별·종사자규모별 모집단 업체수

❸ 맥락적 이해

현재는 점점 더 추정량의 변동계수를 상대표준오차라는 용어로 사용하는 경향이 나타나고 있지만 각종 보고서에서 두 용어가 혼용되고 있기 때문에 맥락적으로 신중하게 구별하는 것이 필요하다고 생각한다.

참고자료

- 박홍래(2008), 통계조사론, 영지문화사.
- 성내경(2012), 표본조사 방법론, 자유아카데미.
- 문화체육관광부(2014), 광고산업통계조사 보고서.
- 산림청(2014), 임산물생산비조사 보고서.
- 중소기업중앙회(2014), 중소제조업 직종별 임금조사보고서.
- 통계청(2015), 생활시간조사 보고서.
- 한국인터넷진흥원(2013), 인터넷이용실태조사 보고서.
- http://www.siminilbo.co.kr/news/articleView.html?idxno=423147
- http://www.yonhapnews.co.kr/bulletin/2015/11/26/0200000000AKR20151126050900017.HTML?input=1195m
- http://www.yonhapnews.co.kr/bulletin/2015/12/03/0200000000AKR20151203202700033.HTML?input=1195m

제 18 장

우연에 대하여 생각하기

18-1. 우연사건

학습목표

- 「우연」이라는 단어에 대한 자신의 생각들을 토의하며 가능성의 정도, 확률의 의미를 느끼는 것을 목표로 한다.

1 우리의 만남은 「우연 사건」인가?

우리가 지금 이 시간에 이곳에서 만나게 된 것은 우연한 일인가? 다른 말로 하면 이 만남이라는 사건을 우연적으로 발생한 사건(이하 "우연사건"으로 표현) 이라고 할 수 있겠는가?

지금 이 장소에, 이 시간에 우리가 함께 있는 것은 우연한 일인가? 아니면 억겁년 전에 맺어놓은 신의 계획 가운데에 있는 것인가?

나도 모르는 한 사람을, 하루에 각기 다른 장소에서 네 번을 만났다면 이는 우연인가? 아니면 그가 나를 쫓아오고 있는 것인가?

공정한(어느 쪽으로도 휘어지지 않은) 동전을 던져보면 어떤 때는 앞면이 나오고 어떤 때는 뒷면이 나온다. 내가 동전을 한번 던졌는데 앞면이 나왔다고 하자. 앞면이 나온 이유가 뭘까? 내 무의식의 어떤 능력이 개입한 결과일까?(나는 의도적으로 앞면이 나오도록 하는 어떠한 조작을 하거나 사전 훈련을 받은 바가 없다.) 또 다시 던져도 어느 면이 나올지는 알 수 없다. 이렇게 10번을 던졌다고 하자. 10번 모두 앞면이 나왔다면 이것은 우연한 결과인가? 아니면 계획된 어떤 조작의 결과인가? 그런데 만약 10번 던져서 앞면이 6번 나왔다면 이것은 어떠한가?

조금 달리 상황을 보자. 위와 똑같이 동전을 던지는데 이번에는 깊이 심호흡을 하고 앞면이 나오길 간절히 바라는 마음으로 던졌는데 앞면이 나왔다고 하자. 이 때 앞면이 나온 것도 우연이라고 할 수 있을까? 아니면 「간절한 바람」이 어떤 힘으로 작용한 결과일까?

이와 같이 「우연」이라는 단어는 우리가 사는 삶의 경험 한 가운데에서 우리와 함께 하고 있다. 어떤 사건을 경험하는 순간에 우리는 질문한다. 이 사건이 발생한 것은 우연한 것인가? 아니면 어떤 설명이 가능한 것인가?

❷ 「우연 사건」의 공통점

지금까지 「우연」인가라는 질문이 어떤 상황에서 어떻게 떠오르는지 생각해 보았다. 그렇다면 이러한 「우연 사건」이 갖고 있는 공통점은 무엇인가?

공통점 1) 불확실성이 내포된 상황이다. 많은 사람을 만날 수 있었고, 동전은 앞면도 뒷면도 나올 수 있고, 10번 던졌을 때에도 앞면이 한 번도 안 나올 수도 있었고, 한 번, 두 번, …, 열 번 모두 나올 수도 있었다. 다양함, 다름을 포함하고 있는 상황이다. 다른 말로 하면 어떤 결과가 나올지, 어떤 사건을 경험하게 될지 예측할 수 없는 상황이다.

공통점 2) 더 이상 자신의 이성적 수준에 맞는(이해할 수 있는) 설명을 찾을 수 없다고 또는 찾을 필요가 없다고 생각하는 상황이다. 특별한 예외가 있을 수 있고 사람마다 다소간에 차이가 있겠지만 일반적으로 사람들은 자신이 경험하는 것을 설명하고 싶어 한다. 그러나 설명이 되지 않는 경우가 너무도 많다. 당신은 어떠한 사람인가? 설명이 안 되면 많이 괴로워하는 사람인가? 아니면 그냥 편한 사람인가? 당신이 어떠한 사람이든 설명이 되지 않는 상황에서는 「우연 사건」으로 결론 내릴 수밖에 없게 된다.

❸ 「우연 사건」 앞에 선 우리의 심경

그렇다면 불확실성과 설명 불가능성(또는 설명 불필요성) 아래에서 우리에게 발생한 사건을 우리가 「우연 사건」이라고 생각할 때의 우리의 심경은 어떠한가?

달리 말하면 어떠한 심경일 때 우리는「우연 사건」으로 생각하게 되는가?

첫째는 설명이 되지 않아도 마음이 편안한 상황이다.「우연」이라는 단어 자체가 떠오르지도 않는 그런 자연스러운 상황이 실은 우연으로 간주하고 있는 상황이다.

길을 가다가 돌부리에 걸려 잠깐 휘청거렸을 때 우리는 우연히 돌부리에 걸린 것이라고 생각할 것이다. 또한 오랜만에 길에서 가까웠던 친구를 만났을 때에도 우리는 "우연히 옛친구를 만났다"고 편안한 마음으로 말할 수 있을 것이다. 공정한 동전을 10번 던지는 게임에서 앞면이 6번 나왔을 때 우리는 "어! 생각보다 (한번) 더 나왔네. 던지다 보면 그럴 수 있지"라고 말하면서 편안한 마음으로 6번이 나온 것을 우연한 결과로 생각한다. 어느 중소기업이 불량률이 10%라고 하는 부품을 100개 구입했는데 불량이 11개가 나왔다고 할 때 이 회사에서는 "어! 생각보다 하나 더 나왔네 (우연히) 그럴 수 있지"하면서 심각하게 생각하지 않고 그 상황을 그대로 받아들일 수도 있다.

둘째는 설명이 되지 않아서 마음이 불편하지만 그렇다고 달리 설명할 길이 없는 상황이다. 이런 경우는「우연히 일어났다」고 말하지만 마음이 불편하다.

길을 가다가 반대편에서 오는 사람과 심하게 부딪쳤는데 오전, 오후 계속 서너 번을 반복해서 그런 일이 일어난다면 어떨까? "내가 정신이 없나?" 혹은 "재수가 없네! 오늘 왜 이러지?" 등의 말과 함께「우연」으로 생각하려고 한다. 오랜만에 길에서 그다지 가깝지는 않았던 친구를 만났을 때에도 우리는 "우연히 옛 친구를 만났다"고 말하지만 만약 이 친구를 3일 연속 당신 사무실 근처에서 만났고, 그 친구는 이 만남을 우연으로 생각하면서 즐거워하는 표정을 짓고 있을 때 당신의 마음은 어떠한가? 그렇다고 그 친구의 말을 믿지 않고 부정적으로(무슨 곤란한 부탁을 하려고 의도적으로 접근했다… 등) 단정할 수는 없지 않은가? 필자 같으면 조금 찜찜하지만 일단은 우연사건으로 받아들일 것이다.

18-2. 우연의 정도

학습목표

- **일상생활에서 일어나는 사건에서 "우연"이라고 말하고 싶은 정도를 느끼는 것을 목표로 한다.**

1 우연이라고 말하고 싶은 정도

지금까지 우리는 우리가 자주 접하는 상황 속에서 마주치는「우연」에 관한 세 가지 면을 이야기했다.

첫째는 우연사건으로 보느냐, 마느냐는 것은 사건을 경험한 당사자의 주관적인 결정이라는 것이다.

둘째는「우연」이라는 단어와 관련된 상황의 공통점을 생각해 보았다. 그것은 불확실성, 다양성이 있는 상황이고, 또한 경험된 사건을 설명할 수 없다고 생각하거나 설명의 필요성이 느껴지지 않는 상황이다.

셋째는「우연」이라고 결론을 내리게 되는 상황에서의 우리의 심경을 생각해 보았다. 한 가지는 설명이 되지 않아도 굳이 "왜 그럴까?"라는 생각 없이「우연」이라고 결론을 내리는 상황이고, 또 다른 한 가지는 설명을 하고 싶어도 설명을 할 수가 없어서 마음이 불편한 상태에서「우연」이라고 결론을 내리는 상황이다.

이 세 가지는 결국 우리는 어떤 사건에 대해서는 쉽게 우연이라고 말하고, 또 다른 어떤 사건에 대해서는 우연이라고 말하기를 싫어하고 있는 것이다. 따라서 우리는「우연이라고 말하고 싶은 정도(degree of randomness)」를 생각해본다. 그러므로 전자는「우연이라고 말하고 싶은 정도」가 큰 상황이고, 후자는「우연이라고 말하고 싶은 정도」가 작은 상황이다.

지금까지 다소 지루하게「우연」이라는 얘기를 했는데 확실한 것은 우리는 살아가면서 의식적이건 무의식적이건 경험하는 모든 사건에「우연이라고 말하고 싶은 정도」를 부여하고 있다는 것이다. 어떤 사람은 거의 대부분의 사건들을 우연의 결과라고 보기도 하고, 반대로 종교인 등과 같은 사람들은 모든 삶에서 경험되는 사건이 우연이 아닌 필연의 결과라고 보기도 한다. 또 다른 말로 하면 어떤 사건이 우연 사건이라고 하는 것은 그 사건 고유의 특성이 아니라 그 사건을 경험하고 있는 당사자의 입장이다. 여러 가능한 경우들(불확실성)이 있었는데도 그 중 하나를 경험했을 때 그

당사자가 어떤 설명을 하고 있다면 우연 사건이 아닌 것이고, 그 어떤 설명도 받아들이지 않는다면 우연 사건 쪽으로 생각을 정리한다고 할 수 있다.

자녀를 양육하는 부모의 입장에서는 자녀들의 이러한 성향을 잘 살펴보면 적성 파악이나 진로 지도에 좋을 것이다. 발생한 사건에 대하여 「우연이라고 말하고 싶은 정도」가 작은 자녀는 결국 설명을 하고자 하는 욕망이 강한 아이일 것이고, 따라서 논리적 훈련을 잘 시킨다면 학자, 연구자의 길을 갈 수 있을 것이다. 반면에 발생한 사건에 대하여 「우연이라고 말하고 싶은 정도」가 큰 자녀인 경우는 느긋하고 설명이 안 되어도 순리에 따르는 성향이 많은 아이일 것이므로 가능한 한 단순한 일을 하는 것이 더 좋을 듯도 하다. 그렇다면 당신은 어떠한가? 당신의 동료는 어떠한 성향의 사람인가? 다음 절에서 결론적으로 당신과 당신의 동료 중 누가 더 이 정도가 강한지를 알아보면서 「우연」이라는 개념의 한 단면을 이야기하며 마치려고 한다.

❷ 「우연 사건」의 특징

어떤 사람이 소경으로 태어난 사건, 어떤 사람의 IQ가 100인 사건, 벼락을 맞는 사건, 어느 날 돌부리에 넘어진 사건, 로또를 시도했는데 '꽝'으로 판명된 사건, 주사위를 던졌는데 3이 나온 사건, 윷놀이를 하는데 5번 연속 모가 나오고 마지막에 걸이 나와서 끝나는 사건, 결혼한 아들이 1남 1녀를 둔 사건, 이 개별적 사건들 하나하나는 모두 불확실한(여러 가능성이 있는) 가운데 발생한 「우연 사건」 이라고 할 수 밖에 없다. 그러나 장기적으로 이들 사건의 발생 빈도를 조사해 보면 특이하게도 그 결과는 어떤 규칙성을 갖는다는 것이다. 복권을 계속 시도하면 대부분 '꽝'이고 아주 어쩌다가(희귀하게) 당첨이 된다. 자녀가 두 명인 많은 가정을 조사해 보면 약 50%가 1남 1녀이다. 지금 이 순간 태어나는 많은 아이들 중에 불행하게도 소경으로 나오는 아이의 비율은 대단히 낮다. 이런 것들은 인생을 살아 본 사람들은 경험적으로 알 수 있다. 보다 구체적인 예로는 공정한 동전을 10회 던졌을 때 우연히 10회 모두 앞면이 나올 수도 있지만, 10회씩 동전을 던져서 앞면이 몇 번 나오는지 관찰하는 일을 1,000번 반복해 보면 10회 모두 앞면이 나오는 경우는 한 번 정도이다. 반면에 우연히 앞면이 6번 나오는 경우는 1,000번 중 200여 번 정도가 된다. 물론 여기서 많은 사람들은 '1번'이라든가 '200번'처럼 정확한 값을 쉽게 알 수는 없지만 느낌으로는

10회 모두 앞면이 나오는 것은 굉장히 드물고, 10회 중 6회가 앞면이 나오게 되는 것은 1,000번 중 여러 번 일 것이라는 것까지는 쉽게 알 수 있다. 이와 같이「우연 사건」이 장기적으로 관찰(경험)되면 어떤 규칙적인 성질을 나타낸다는 특징을 간파한 학자들은 상당히 오랜 기간(약 3~400년)에 걸쳐서 확률이라는 개념을 발전시켜 왔다.

우리는 확률을 수학과목에서 다루는 상당히 어렵고, 수학을 잘 하는 사람들이나 해결할 수 있는 분야로 생각하는데 사실 이 개념의 근원은 우리의 삶과 떨어질 수 없는「우연 사건」또는「우연이라고 말하고 싶은 정도가 큰 사건」과 관계된다. 따라서 확률이라는 용어가 나오는 보고서나 글을 접하면 어려운 주제로 인식할 이유가 없다.「우연」이라는 말과 관련된 자신의 생각을 조심스럽게 정리해 보면 어렵지 않게 이해가 되는 내용이다. 물론 위에서 살펴본 동전을 10회 던졌을 때 나오는 앞면의 횟수와 같은 것은 쉽지 않다. 그러나 이런 것은 엑셀 등과 같은 컴퓨터 소프트웨어가 잘 계산해 주기 때문에 걱정하지 않아도 된다.

18-3. 우연의 정도의 계량화

학습목표

- **13장, 15장에서 다룬 얼마나 극단적이냐(얼마나 크냐, 얼마나 작냐, 얼마나 다르냐)에 대한 토의 내용을 토대로 우연의 정도를 계량화하는 논리에 접촉하여 귀납적 사고와 만난다.**

1 우연사건의 확률

앞에서 확률을「우연 사건」의 장기적 결과의 표현이라고 정의하였다. 다음과 같은 예를 들어보자. 공정한 동전(즉 앞면이 나올 확률이 1/2, 더 쉽게 말하면 동전을 많이 던져 약 반 정도가 앞면이 나오는 그런 동전)을 10회 던진다고 해보자. 엑셀 등에서 다음과 같은 확률을 쉽게 구할 수 있고, 또한 우리 마음에서는 우연이라고 말하고 싶은 정도의 순위가 다음과 같을 것이다.

<표 18-1> 동전던지기 확률

앞면의 수	순위	확률
0	6	0.00098
1	5	0.00977
2	4	0.04395
3	3	0.11719
4	2	0.20508
5	1	0.24609
6	2	0.20508
7	3	0.11719
8	4	0.04395
9	5	0.00977
10	6	0.00098

여기에서 우리는 우리가「우연이라고 말하고 싶은 정도」와 확률이 연결되어 있음을 본다. 공정한 동전을 10회 던졌을 때 우연히 10회 모두 앞면이 나올 수 있다는 것은 인정하지만 막상 10번 모두 앞면이 나왔다면 당신은 이것을「우연 사건」이라고 보고 싶지 않을 것이다. 동전이 반듯한 것이 아니라 휘어진 것 아닌가? 조작된 것 아닌가? 등등 여러 생각이 머릿속을 스칠 것이다. 즉, 우연이라고 하고 싶은 정도가 아주 작아졌음을 느낀다. 앞

에 계산해 놓은 확률을 보면 공정한 동전을 10회 던졌을 때 10회 모두 앞면이 나올 확률은 약 0.001로 아주 작다. 이제부터는 이 정도를「우연의 정도」라고 부르기로 한다. 반면에 10회 시행에서 앞면이 6회 나오면 별다른 생각이 들지 않을 것이다. 왜 5회가 아니라 6회냐고 묻는다면 편하게 "던지다 보면"또는 "우연히"라고 답할 것이다. 즉 우연의 정도가 큰 것이다.

❷ 우연의 정도의 계량화

위에서 구한 확률을 이용하여 우연의 정도를 다음과 같이 계량화해보자. 예컨대 앞면이 6회 나왔다면 이 경우의 우연의 정도는 다음과 같이 6회 이상의 모든 경우가 나타나는 확률의 합으로 정의해보자.

0.205(6회) + 0.117(7회) + 0.044(8회) + 0.010(9회) + 0.001(10회) = 0.377

이 의미는 앞면이 6회가 나왔을 때 5회를 기대했던 사람으로서 1회가 더 나왔으므로 우리는 6회와 그 이상의 큰(작은) 값이 나오는 사건에 대한 우리 마음을 우연의 정도(우연이라고 말하고 싶어 하는 정도)라고 하면서 0.377로 수치화한다는 것이다.

❸「우연의 정도」의 계량화의 활용

우리 기관에서 특정 사무용품을 구매하면서 계약서에 불량률 5%를 받아들이기로 하였다고 하자. 200개를 구매하였는데 불량품이 11개가 나왔다. 처음에 생각했던 것보다 1개가 더 나왔는데 이것을 납품과정에서 어쩌다 일어날 수 있는「우연 사건」이라고 해야 하나? 아니면 의도적으로 불량품을 많이 납품한 것으로 보아야 하나? 만약 불량품이 15개였다면 당신은 이것을 어떻게 보겠는가? 이와 같은 관찰을 통해서 우리는「우연의 정도」를 다음과 같이 정의해 볼 수 있다.

사건 A의「우연의 정도」는 사건 A나 사건 A보다 더 희귀한(극단적) 사건이 일어날 확률(가능성)이다.

가상이지만 이런 예를 생각해 보자. 우리가 만드는 보고서 초안에는 5%의 오타가 포함된다고 하자. 당신이 작성한 보고서 초안에서 200자 중 오타자가 11자가 나왔다면 일반적으로 기대하는 것보다 우연히 하나가 더 많이 나온 것인가? 아니면 당신의 입력 태도가 불량한가? 15자가 나왔다면

어떠한가? 입력한 사람을 평가할 때 15자의 오타가 발생한 것을 입력하다가 나타난 우연 사건으로 보느냐 아니면 불량사건으로 보느냐? 이런 류의 의사결정은 기관 내에서 자주 발생한다. 이럴 때 결정을 내리는 사람들은 막연히 "우연이라고 할 수도 있겠네!" 하면서 너그러울 수도 있고 "우연이 아니지! 이것은 일하고자 하는 마음이 없는 거야" 등의 마음을 가질 수도 있다. 그러나 이러한 주관적인 느낌만으로는 조직의 공정성, 투명성, 일관성 등이 확보되기 어렵다.

또 다른 예를 보자. 지금 당신이 속한 기관에서 제시하는 정책에 대한 주민들의 지지율이 최소 80% 이상이라고 믿고 있다고 하자. 지역 주민 400명을 잘 설계된 표본추출방식으로 추출하여 조사한 결과 280명(70%)이 지지한 것으로 나타났다. 결과에 대해서 어떻게 생각하는가? 실제 지지율이 생각했던 것보다 '조금' 낮은 것인가? 아니면 '상당히' 낮은 것인가? 실제로 전 주민의 80%가 지지한다고 하더라도 표본으로 선출된 400명 가운데 우연히 반대자들이 많이 포함되었기 때문에 400명으로부터 추정된 지지율이 70%가 된 것은 아닐까? 그럴 수도 있겠지만 아무리 우연히 반대자들이 많이 포함될 수 있다고 하더라도 우연이라고 하기에는 반대자들이 너무 많은 것은 아닌가? 정책분석팀장은 이를 우연사건으로 간주하고 주민의 지지율을 80%라고 믿고 정책을 추진해 나간다고 하자. 당신이 정책분석팀의 팀원으로 있다면 팀장의 관점을 따를 것인가? 팀장과 당신 생각이 다르다면 당신은 어떤 논리로 당신의 의견을 개진할 것인가? 팀장을 따른다면 팀장이 바뀌었을 때 이 팀의 의사결정 방법은 어떻게 될 것인가? 바로 이 지점에서 우리는 토론과 의사결정을 위하여 「우연의 정도」에 대한 계량화가 절실히 필요해진다. 그래서 실제로 80%의 주민이 지지한다고 할 때 그 중 대표성 있는 표본 400명 속에 우연히 30%의 반대자(70%의 지지자)들이 포함될 수 있는 「우연의 정도」가 작으면 작을수록 겸허히 주민들의 80%가 정책을 지지할 것이라고 생각(믿음)한 정책분석팀의 생각(믿음)을 재검토해 보는데 무게를 실어야 할 것이다.

참고자료

• 이석훈(2006), 통계적 사고방식, 통계교육원.

■ **단원 10.**

• 다음의 자료에 적용할 수 있는 확률분포 모형을 제안하시오.

1) 임의의 25개 가구의 한달 소득자료

ID	소득	ID	소득	ID	소득	ID	소득	ID	소득
1	1,720,850	6	1,500,000	11	700,000	16	2,413,500	21	847,100
2	2,080,000	7	4,957,770	12	2,220,000	17	4,381,030	22	2,975,470
3	899,100	8	3,064,370	13	1,805,110	18	981,830	23	1,336,800
4	3,375,500	9	4,350,000	14	1,737,980	19	7,107,160	24	2,386,500
5	780,000	10	1,870,040	15	2,240,000	20	1,049,000	25	5,598,080

2) 임의의 25명이 좋아하는 색(무지개 색 7가지)

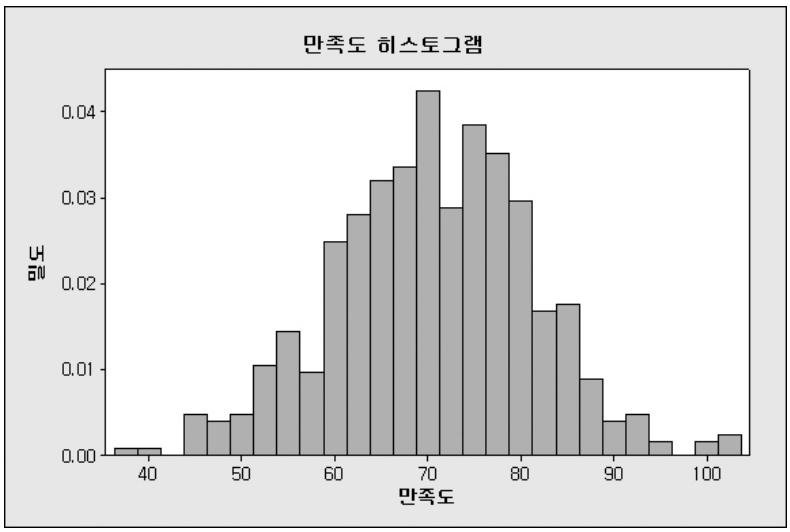
ID	색	ID	색	ID	색	ID	색	ID	색
1	3	6	7	11	1	16	1	21	1
2	5	7	2	12	7	17	6	22	4
3	5	8	4	13	7	18	7	23	2
4	3	9	4	14	4	19	2	24	5
5	3	10	2	15	6	20	1	25	2

3) 임의의 25개 사업체의 연간 안전사고 발생 건수

ID	사고건수	ID	사고건수	ID	사고건수	ID	사고건수	ID	사고건수
1	3	6	6	11	5	16	3	21	5
2	4	7	5	12	7	17	3	22	5
3	4	8	3	13	6	18	8	23	5
4	4	9	3	14	1	19	9	24	3
5	4	10	2	15	5	20	7	25	7

■ **단원 11.**

• **A학교 학생 500명의 학교생활 만족도는 평균 70점에 표준편차 10점이고 히스토그램은 다음과 같이 나타났다.**



1) 만족도 점수가 80점 이상인 학생은 얼마나 될까?

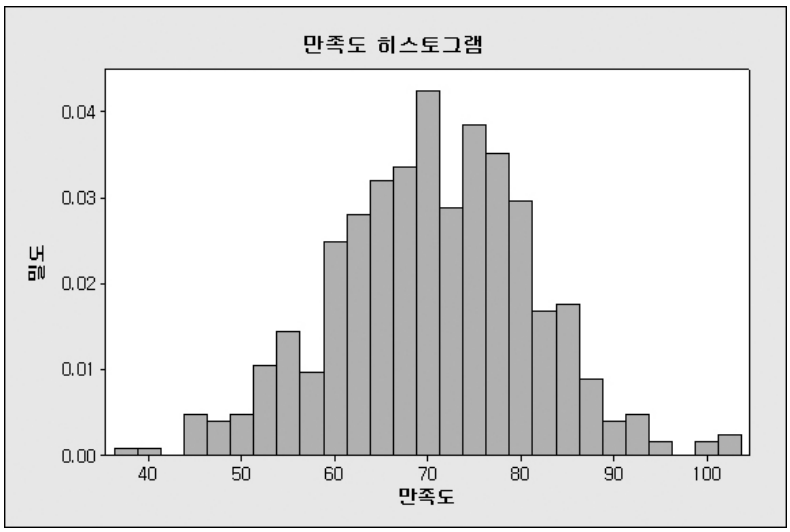
2) 만족도 점수가 90점 이상인 학생은 얼마나 될까?

3) 만족도 점수가 60점에서 85점 사이인 학생은 얼마나 될까?

4) 만족도 점수가 40점 미만이거나 95점 이상인 학생은 얼마나 될까?

■ 단원 12.

• A학교 학생 500명의 학교생활 만족도는 평균 70점에 표준편차 10점이고 히스토그램은 다음과 같이 나타났다. 이 정보로부터 만족도가 80점인 학생은 얼마나 높다고 할 수 있을까? 만족도가 90점인 학생은 얼마나 높다고 할 수 있을까?



• 초등학교 5학년 남학생의 키가 145cm라고 하면 이 학생의 키는 또래 학생들과 비교하여 얼마나 크다고(작다고) 할 수 있을지를 판단하기 위한 과정을 기술하시오.

• 어떤 1인 가구의 소득이 100만원이라면 이 소득은 얼마나 적은지를 정량화해 보시오.

■ 단원 13.

• 통계청 자화상을 보며 관심 있는 주제에 대하여 모집단 분포를 가정해보고 자신에 대한 수치를 확인해본다.

• 13-1의 경우 초등학교 6학년 여학생의 키가 홍만이의 키와 같은 135cm였다고 하면 당신은 성장호르몬에 대하여 어떤 입장을 취하겠는가?

• 어린왕자의 상자 안에 있는 양들은 그 길이가 평균 70cm에 표준편차 3cm인 정규분포를 따른다고 알려져 있다. 이 상자 안에서 길이가 67cm인 양을 보았다면 이 양을 본 당신은 어떤 느낌이 드는가? 만약 이 양의 길이가 64cm라면 어떤 느낌이 드는가?

• 중앙자살예방센터에 따르면 2014년 10대의 자살자 수는 10만명당 약 5명으로 알려져 있다(http://www.spckorea.or.kr/index.php). 2016년에도 이와 같은 상황이 계속된다면 인구 10만명당 자살자 수가 7명 이상이 될 가능성은 얼마나 될까?

■ 단원 14.

• 2013 국민체력실태조사 보고서에 따르면 성인 남자 30~34세의 평균은 74.9kg이고 표준편차가 약 10.2kg으로 나타나 있다. 30-34세의 임의의 성인 남자 9명의 평균체중이 68.1kg이고 표준편차가 11.5kg이라면 이 9명의 남성의 평균체중은 얼마나 작다고 하겠는가?

• 제11차(2015년) 청소년건강행태 온라인조사 결과에 따르면 우리나라 청소년 흡연율은 약 10%로 보도되었다. OO학교에서 전교 900명의 학생을 대상으로 흡연여부에 대해 조사한 결과 117명이 흡연을 하고 있는 것으로 나타났다면, 이 학교 학생들은 흡연율이 얼마나 높다고 하겠는가?

■ 단원 15.

• 어린왕자의 상자 안에 있는 양들은 그 길이가 평균 70cm에 표준편차 3cm인 정규분포를 따른다고 알려져 있다. 당신이 이 상자 안을 들여다보았을 때 9마리의 양을 보았는데, 그 양들의 길이는 평균 67cm이고 표준편차는 4cm이다. 이 9마리의 양들을 본 당신은 어떤 느낌이 드는가?

• 13-3절의 경우 산악동호회 회원인 A씨는 같은 동호회 회원 36명에게 평균 SNS 이용횟수를 조사한 결과, 평균 2.5회로 3.35회보다 0.85회 더 적게 이용하고 있는 것으로 나타났다. 이 산악동호회에 속한 사람들의 SNS 이용횟수는 얼마나 적은가?

연구과제 또는 연습문제

■ 단원 16.

• A 기관은 금년도 민원인 만족도를 추정하기 위하여 금년에 민원실을 방문한 민원인 100명을 임의로 추출하여 우편조사를 실시한 결과 평균 만족도가 100점 만점에 70점, 표준편차가 10점이 나왔다. 이 값으로부터 이 기관의 만족도 평균을 추정하라.

• B 후보에 대한 지지율을 알아보기 위하여 OO 여론조사기관이 유권자 1000명을 대상으로 조사한 결과 650명이 지지한 것으로 나타났다. B 후보에 대한 지지율을 추정해보라.

■ 단원 17.

• A 기관은 금년도 민원인 만족도를 추정하기 위하여 금년에 민원실을 방문한 민원인 100명을 임의로 추출하여 우편조사를 실시한 결과 평균 만족도가 100점 만점에 70점, 표준편차가 10점이 나왔다. 이 값으로부터 이 기관의 만족도의 변동계수(CV)를 구하라. 그리고 이 기관의 만족도 평균의 95% 신뢰구간과 RSE를 구하라.

• 아래는 2015년 12월 3일자 "20년간 주1회 이상 음주율 … 남성 줄고 여성 늘었다<한국갤럽>" 기사 내용의 일부이다.

> 지난 20년간 주 1회 이상 음주하는 비율이 남성은 줄고 여성은 크게 늘어난 것으로 조사됐다. 음주자들이 선호하는 술은 맥주에서 소주로 옮겨갔다. 한국갤럽은 지난달 10~12일 만 19세 이상 남녀 1천12명을 대상으로 평소 음주 빈도 등에 관한 설문조사(신뢰수준 95%, 표본오차 ±3.1%포인트) 결과를 3일 발표했다. 남성의 주 1회 이상 음주 비율은 1994년 58%에서 2015년 52%로 다소 줄었다. 반면 여성은 1994년 8%에서 2015년 18%로 눈에 띄게 늘었다. 성인 남녀 통틀어 주 1회 이상 음주 비율은 35%, 셋 중 한 명꼴이었다. …
>
> http://www.yonhapnews.co.kr/bulletin/2015/12/03/0200000000AKR20151203202700033.HTML?input=1195m

• 위의 기사에 나타난 표본오차에 대한 내용을 기술하고 3.1%포인트를 산출하는 과정을 기술해보라.

■ 단원 18.

- 성인의 키가 $N(\mu = 170, \sigma^2 = 10^2)$ 으로 알고 있는데 크기가 25명인 표본의 평균이 176cm로 나왔다. 이 결과는 모집단에서 임의로 추출된 표본에서 우연히 나타난 결과라고 생각하는가?
- 지지율이 70%라고 알고 있던 사람이 1000명을 조사해본 결과 650명이 지지한 것으로 나타났다면 이 결과가 임의로 추출된 1000명으로부터 우연히 나타난 결과라고 보겠는가? 아니면 70%라고 알고 있던 지지율이 다소 과장된 것이라고 생각하겠는가?

3부

통계분석 도구 활용하기

3부. 통계분석 도구 활용하기

목차

통계분석 도구 활용하기 과목의 개요

학습 목표

- 엑셀(excel) 프로그램을 이용하여 자료에 대한 기초적인 분석을 할 수 있다.
- 다양한 형태의 표와 그래프를 작성할 수 있다.
- 기술통계량을 구할 수 있다.
- 모집단의 분포인 정규분포, 이항분포, 포아송분포에서 확률을 구할 수 있다.

선수학습

없음

주요 용어

빈도, 백분율, 빈도표, 교차표, 막대그래프, 원그래프, 히스토그램, 분산형 그래프, 기술통계량, 정규분포, 이항분포, 포아송분포

학습과목의 내용요약

일변량인 범주형 자료(또는 질적 자료)와 양적 자료에 대한 특징을 파악하기 위해서 엑셀을 이용하여 표(빈도, 백분율), 그래프(막대그래프, 원그래프, 히스토그램), 기술통계량을 작성하고, 두 개의 범주형 자료 간의 현황을 알아볼 수 있는 교차표, 두 개의 양적 자료 간의 관련성을 알아보기 위한 산점도 작성하는 방법을 다룬다.

제 1 장 표 작성하기

1-1. 빈도표

학습목표

- **수집된 범주형 자료(categorical data) 또는 질적 자료(qualitative data)에 있는 중요한 특징을 알아보기 위해서 엑셀(excel)을 이용하여 빈도표를 작성하는 것을 학습한다.**

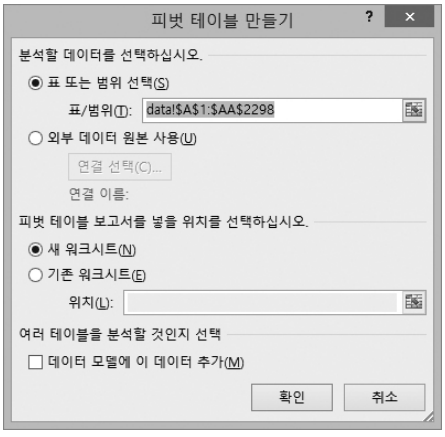
1 빈도표 작성하기

1. 빈도 구하기

[단계 1] A1셀을 선택한다.

[단계 2] 메뉴에서『삽입』→『피벗 테이블』→『피벗 테이블』을 선택한다. 그러면 아래의 화면처럼 분석할 데이터의 범위를 자동으로 선택된다.

[그림 1-1]
피벗 테이블 만들기



[단계 3]『확인』 버튼을 클릭한다. 그러면 아래의 그림처럼 피벗 테이블을 작업할 수 있는 새로운 시트가 생성된다.

[그림 1-2]
피벗 테이블 화면

[단계 4]『피벗 테이블 필드 목록』에 있는 『성별』을 드래그(drag)해서 『행 레이블』에 한 번, 『값』에 두 번 넣는다.

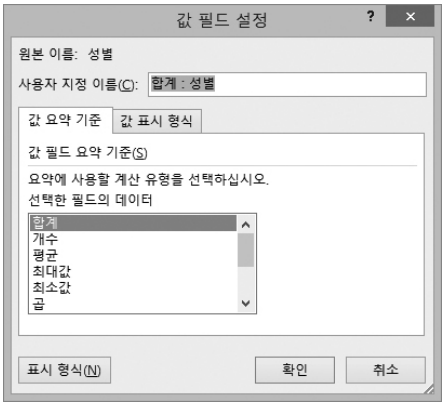
[그림 1-3]
피벗 테이블 작업(1)

[단계 5]『값』에 있는 첫 번째『합계 : 성별』을 선택하고, 마우스 오른쪽을 클릭한다.

[그림 1-4]
값필드 설정

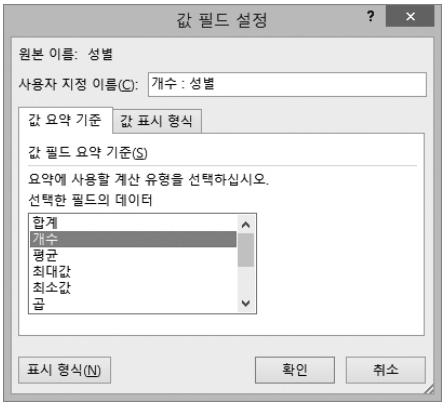
[단계 6] 팝업 메뉴(pop-up menu)에서『값 필드 설정』을 클릭한다.

[그림 1-5]
값필드 설정(2)

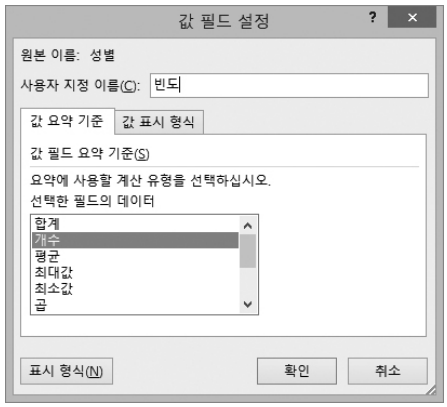


[단계 7]『값 필드 설정』화면에서『합계』를 『개수』로 수정하고,『사용자 지정 이름』의『합계 : 성별』을『빈도』로 수정한다.

[그림 1-6]
값필드 설정(3)



[그림 1-7]
값필드 설정(4)



[단계 8] 『확인』버튼을 클릭한다. 그러면 성별에 대한 빈도의 현황이 나타난다.

[그림 1-8]
피벗 테이블 : 빈도

2. 백분율(percent) 구하기

[단계 1]『값』에 있는 두 번째『합계 : 성별』을 선택하고, 마우스 오른쪽을 클릭한다.

[단계 2] 팝업 메뉴(pop-up menu)에서『값 필드 설정』을 클릭한다.

[단계 3]『값 필드 설정』화면에서 다음과 같이 수정한다.

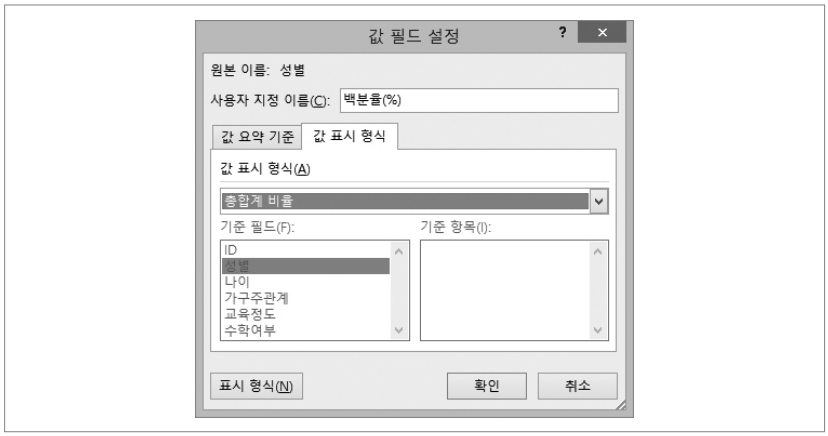
선택한 필드의 데이터 :『합계』→『개수』

사용자 지정 이름 :『합계 : 성별2』→『백분율(%)』

값 표시 형식 탭을 선택

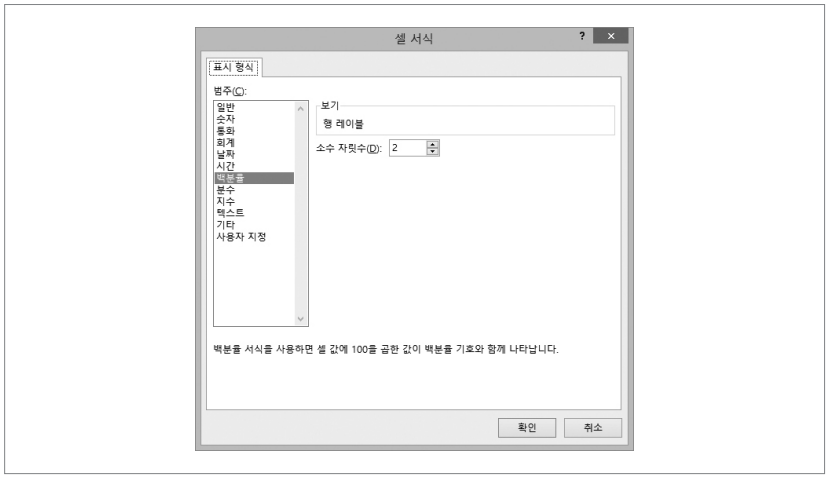
[단계 4]『값 표시 형식』의『계산 없음』을『총합계 비율』로 수정한다.

[그림 1-9]
값 필드 설정(3)



[단계 5]『표시 형식』버튼을 클릭한다.

[그림 1-10]
셀 서식



[단계 6] 범주를『백분율』로 선택하고, 소수 자릿수의『2』를『1』로 수정한다.

[단계 7]『확인』버튼을 클릭한다.

[단계 8]『확인』버튼을 클릭한다. 그러면 아래의 그림처럼 백분율이 나타난다.

[그림 1-11]
성별에 대한 빈도표

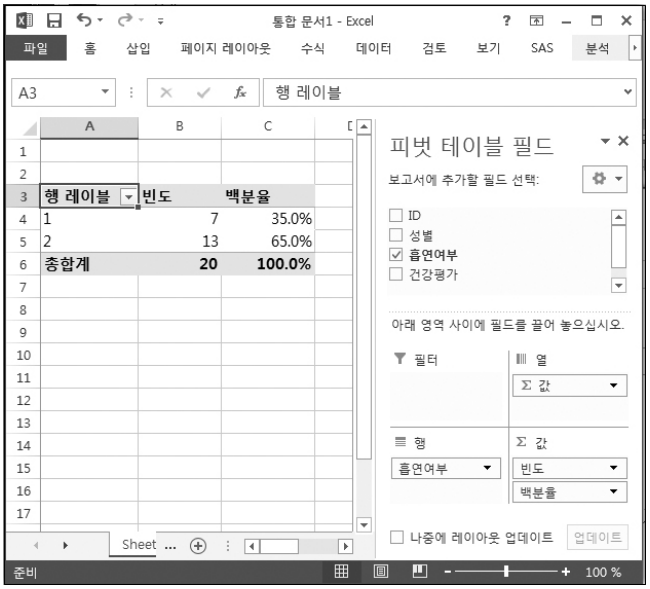
❷ 다음은 2014 사회조사 자료의 일부이다. 이를 이용하여 빈도표를 작성해보자.

<표 1-1>
2014 사회조사 자료
(일부)

ID	성별	흡연여부	건강평가
1	2	2	2
2	2	2	3
3	1	2	3
4	2	2	3
5	2	1	2
6	1	1	1
7	1	1	2
8	1	1	1
9	2	2	3
10	1	2	2
11	2	2	3
12	2	2	2
13	1	2	4
14	1	1	2
15	1	1	2
16	1	2	2
17	2	2	3
18	1	2	2
19	2	2	3
20	1	1	2

1. 흡연여부 빈도표

[그림 1-12]
흡연여부에 대한 빈도표



[단계 1] 데이터를 블록잡고 메뉴에서 『삽입』 → 『피벗 테이블』 → 『피벗 테이블』을 선택한다.

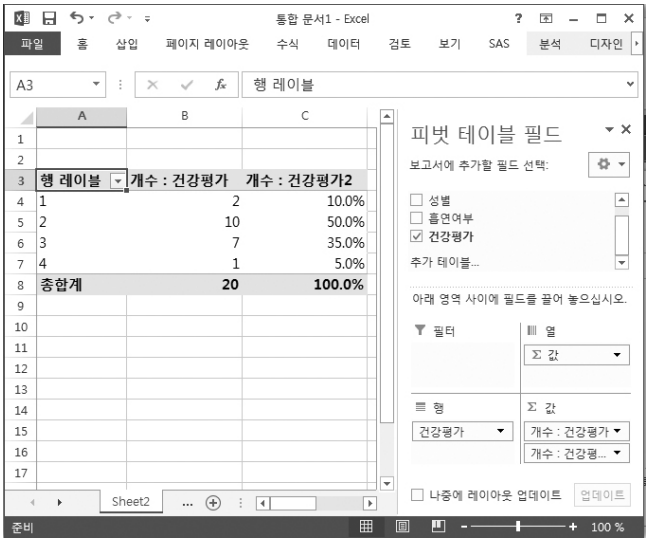
[단계 2] 『피벗 테이블 필드 목록』 에 있는 『흡연여부』를 드래그(drag)해서 『행 』과 『값』에 넣는다.

[단계 3] 『값 필드 설정』화면에서 『합계』를 『개수』로 수정하고, 『사용자 지정 이름』의 『합계 : 흡연여부』를 『개수』로 수정한다.

[단계 4] 백분율을 함께 표시하고 싶은 경우 『흡연여부』를 드래그(drag)해서 『값』에 한 번 더 넣고, 『값 표시 형식』을 『개수』로 변경한 후, 『값 표시 형식』의 『계산 없음』을 『총합계 비율』로 수정하고 『표시 형식』을 『백분율』로 수정한다.

2. 주관적 건강평가 빈도표

[그림 1-13]
주관적 건강평가에 대한 빈도표



행 레이블	개수 : 건강평가	개수 : 건강평가2
1	2	10.0%
2	10	50.0%
3	7	35.0%
4	1	5.0%
총합계	20	100.0%

[단계] 동일한 방법으로 작성한다.

1-2. 교차표

학습목표

- 두 개의 범주형 자료(또는 질적 자료) 간의 관련성이 있는 지를 파악하기 위해서 엑셀의 피벗 테이블을 이용하여 교차표를 작성하는 것을 학습한다.

1 아침식사여부에 따른 건강상태에 차이가 있는지를 알아보기 위해서 교차표를 작성한다.

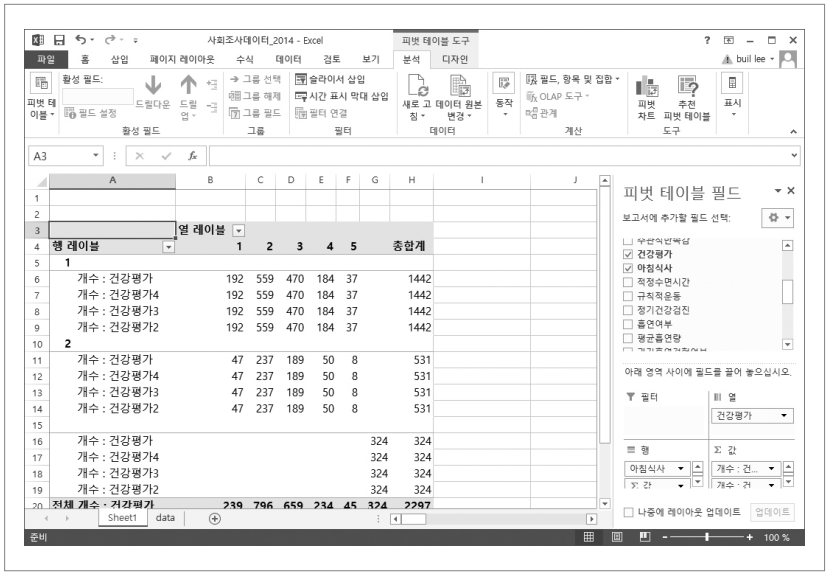
[단계 1] A1셀을 선택한다.

[단계 2] 메뉴에서『삽입』→『피벗 테이블』→『피벗 테이블』을 선택하고,『확인』버튼을 클릭한다.

[단계 3]『피벗 테이블 필드 목록』에 있는『아침식사』를 드래그(drag)해서『행 레이블』에 한 번,『건강상태』를 드래그(drag)해서『열 레이블』에 한 번,『건강상태』를 드래그(drag)해서『값』에 네 번 넣는다.

[단계 4]『열 레이블』에 있는『값』를 드래그(drag)해서『행 레이블』의『아침식사』의 아래쪽에 드래그(drag)해서 넣는다.

[그림 1-14] 아침식사와 건강상태의 교차표(1)



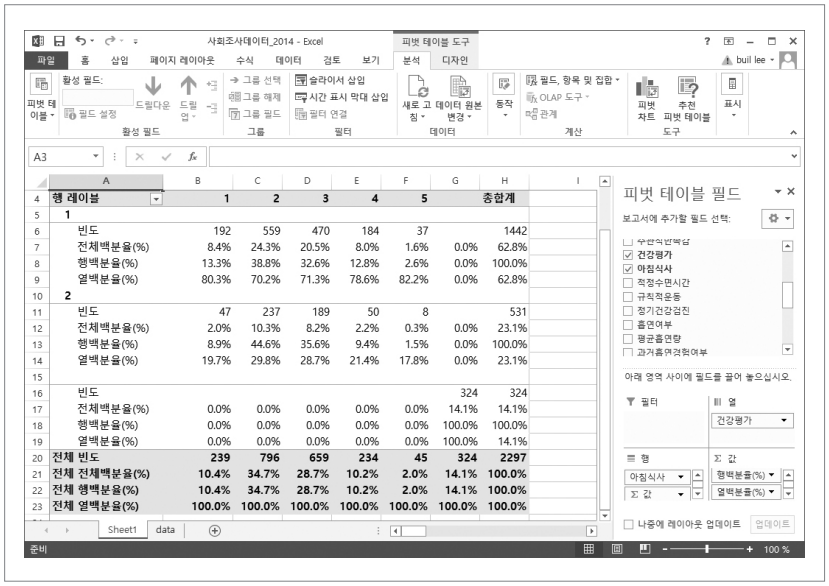
[단계 5] 값에 있는 첫 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 빈도로 수정한다.

[단계 6] 값에 있는 두 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 전체백분율(%), 값 표시형식에서 계산없음을 총합계비율로 수정하고, 표시형식에서 백분율의 소수자리수를 1로 수정한다.

[단계 7] 값에 있는 세 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 행백분율(%), 값 표시형식에서 계산없음을 행합계비율로 수정하고, 표시형식에서 백분율의 소수자리수를 1로 수정한다.

[단계 8] 값에 있는 네 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 열백분율(%), 값 표시형식에서 계산없음을 열합계비율로 수정하고, 표시형식에서 백분율의 소수자리수를 1로 수정한다.

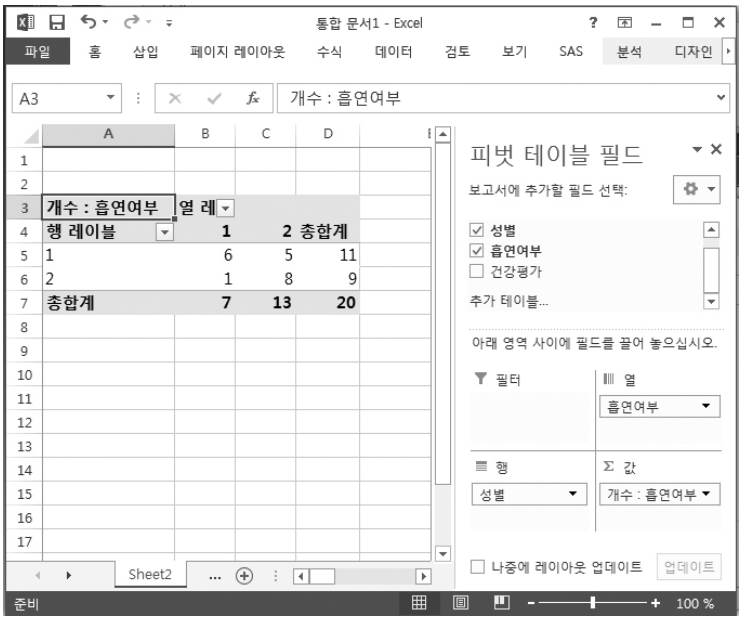
[그림 1-15]
아침식사와
건강상태의 교차표(2)



❷ [표 1-1]을 이용하여 성별에 따른 흡연여부와 주관적 건강평가의 교차표를 작성해보자.

1. 성별에 따른 흡연여부 교차표

[그림 1-16] 성별과 흡연여부의 교차표



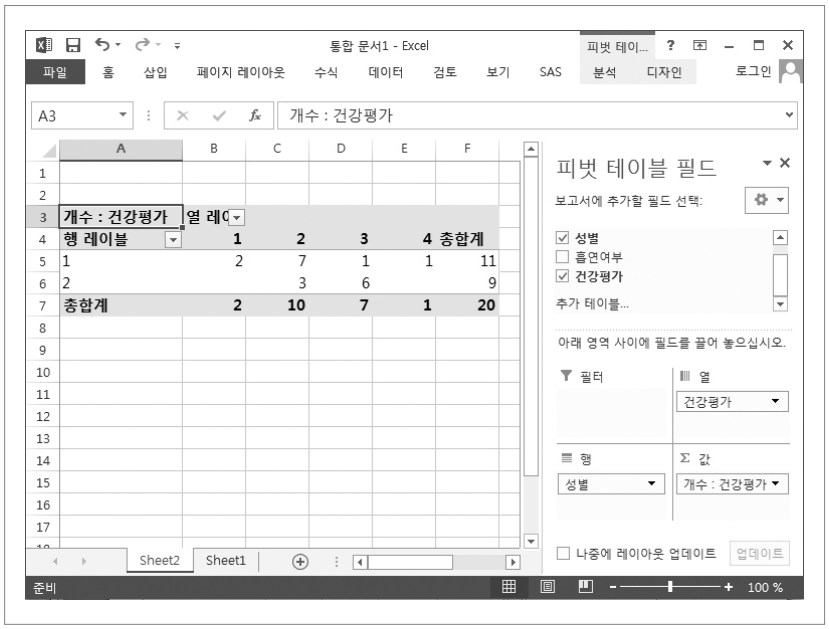
[단계 1] 데이터를 블록잡고 피벗테이블을 삽입한다.

[단계 2] 『피벗 테이블 필드 목록』에 있는 『성별』를 드래그(drag)해서 『행 레이블』에 넣고 『흡연여부』를 드래그(drag)해서 『열 레이블』과 『값』에 넣는다.

[단계 3] 『값 필드 설정』화면에서 『합계』를 『개수』로 수정하고, 필요한 경우 『사용자 지정 이름』의 『합계 : 흡연여부』를 『개수』로 수정한다.

2. 성별에 따른 주관적 건강평가 교차표

[그림 1-17]
성별과 건강평가의
교차표



[단계] 동일한 방법으로 작성하고 필요한 경우『값』에 건강평가를 추가하여 백분율을 함께 제시한다.

1-3. 다차원 교차표

학습목표

• 엑셀의 피벗 테이블을 이용하여 다차원 교차표를 작성하는 것을 학습한다.

1 성별과 교육정도에 따른 아침식사여부에 차이가 있는지를 알아보기 위해서 다차원 교차표를 작성한다.

[단계 1] A1셀을 선택한다.

[단계 2] 메뉴에서『삽입』→『피벗 테이블』→『피벗 테이블』을 선택하고,『확인』버튼을 클릭한다.

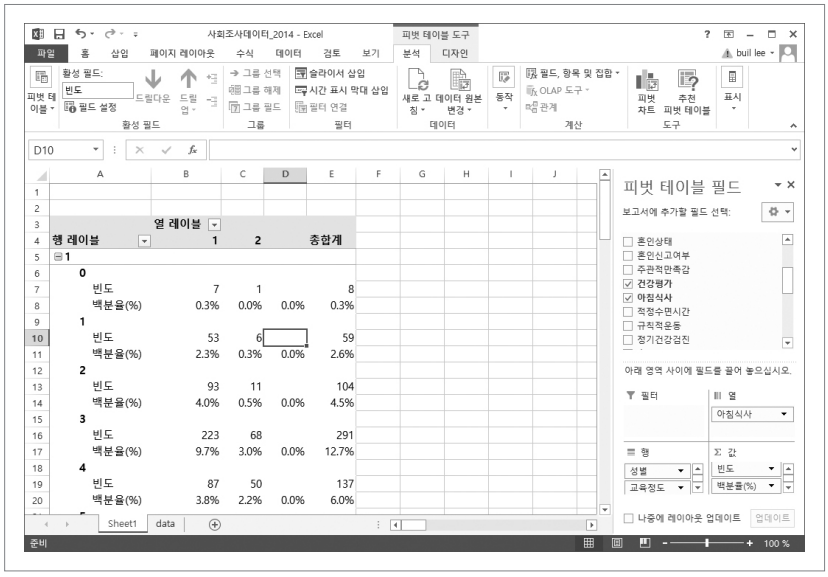
[단계 3]『피벗 테이블 필드 목록』에 있는『성별』을 드래그(drag)해서『행 레이블』에 한 번,『교육정도』를 드래그(drag)해서『행 레이블』에 있는 성별 아래쪽에 한 번,『아침식사』를 드래그(drag)해서『열』에 한 번,『성별』을 드래그(drag)해서『값』두 번 넣는다.

[단계 4]『열 레이블』에 있는『값』을 드래그(drag)해서『행 레이블』의『교육정도』의 아래쪽에 드래그(drag)해서 넣는다.

[단계 5] 값에 있는 첫 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 빈도로 수정한다.

[단계 6] 값에 있는 두 번째를 선택하고, 값 필드 설정에서 함수를 개수, 사용자 지정 이름은 백분율(%), 값 표시형식에서 계산없음을 총합계비율로 수정하고, 표시형식에서 백분율의 소수자리수를 1로 수정한다.

[그림 1-18]
엑셀을 이용한 다차원 교차표

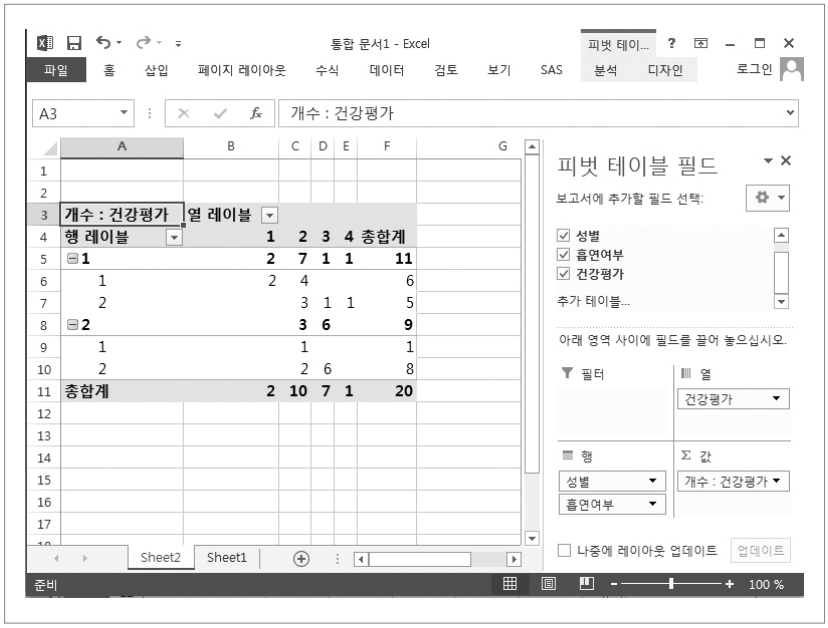


	열 레이블			
행 레이블	1	2		총합계
1				
0				
빈도	7	1		8
백분율(%)	0.3%	0.0%	0.0%	0.3%
1				
빈도	53	6		59
백분율(%)	2.3%	0.3%	0.0%	2.6%
2				
빈도	93	11		104
백분율(%)	4.0%	0.5%	0.0%	4.5%
3				
빈도	223	68		291
백분율(%)	9.7%	3.0%	0.0%	12.7%
4				
빈도	87	50		137
백분율(%)	3.8%	2.2%	0.0%	6.0%

❷ [표 1–1]을 이용하여 성별과 흡연여부에 따른 주관적 건강평가의 교차표를 작성해보자.

1. 성별과 흡연여부에 따른 주관적 건강평가의 다차원 교차표

[그림 1-19]
성별과 흡연여부에 따른 건강평가의 교차표



개수 : 건강평가	열 레이블				
행 레이블	1	2	3	4	총합계
1	2	7	1	1	11
1	2	4			6
2		3	1	1	5
2		3	6		9
1		1			1
2		2	6		8
총합계	2	10	7	1	20

[단계 1] 데이터를 블록잡고 피벗테이블을 삽입한다.

[단계 2] 『피벗 테이블 필드 목록』에 있는 『성별』과 『흡연여부』를 드래그 해서『행 레이블』에 넣는다. 『건강평가』를 드래그(drag)해서 『열』과 『값』에 넣는다.

[단계 3] 값 필드 설정에서 함수를 개수로 수정한다.

제 2 장

그래프 작성하기

2-1. 질적 자료의 그래프

학습목표

- **일변량의 질적 자료(또는 범주형 자료)의 특징을 시각적으로 표현할 수 있는 막대그래프와 원그래프를 엑셀로 작성하는 것을 학습한다.**

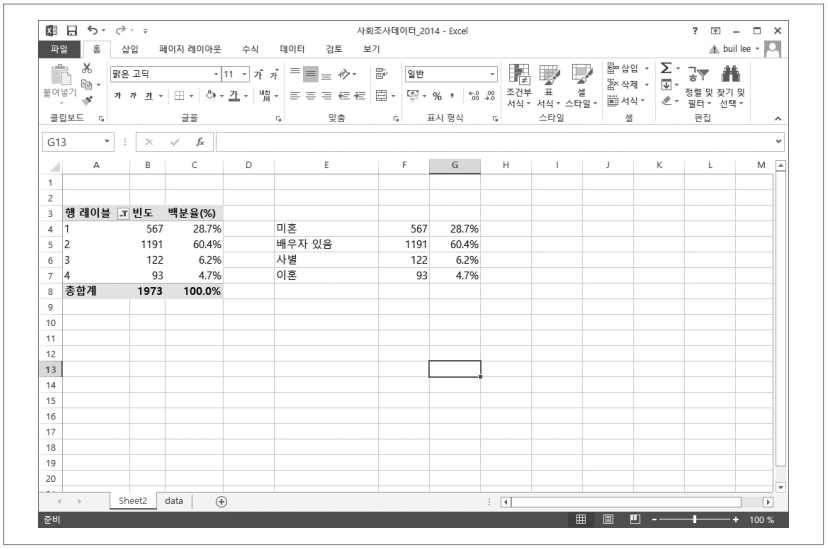
1 혼인상태의 자료를 막대그래프와 원그래프로 작성한다.

1. 막대그래프

[단계 1] 피벗 테이블을 이용하여 혼인상태에 대한 빈도와 백분율을 구한다.

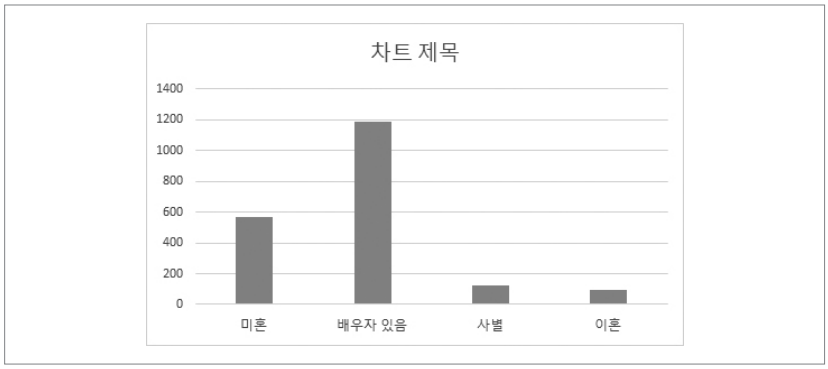
[단계 2] 피벗 테이블을 결과 중에서 일부 내용을 복사하여 빈 셀에 붙여넣기를 하고, 값에 대한 설명을 한글로 작성한다.

[그림 2-1] 혼인상태에 대한 빈도표



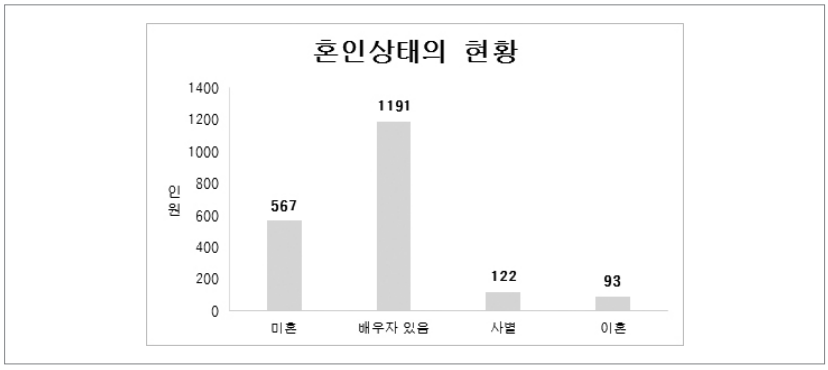
[단계 3] E4셀부터 F7셀까지 블록을 잡고, 메뉴의『삽입』→『차트』→『2차원 세로 막대형』→『묶은 세로 막대형』을 선택한다.

[그림 2-2]
혼인상태에 대한
막대그래프(1)



[단계 4] 차트도구에 있는 서식, 레이아웃, 디자인 등에 있는 기능을 이용하여 막대그래프를 편집한다.

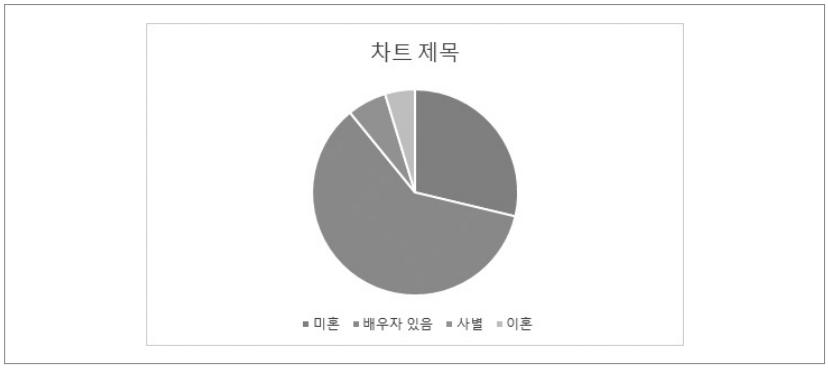
[그림 2-3]
혼인상태에 대한
막대그래프(2)



2. 원그래프

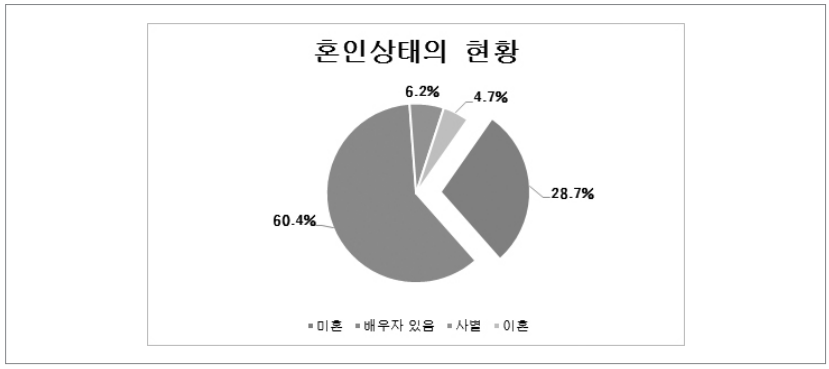
[단계 1] E4:E7, G4:G7 셀을 블록을 잡고, 메뉴의『삽입』→『원형』→『2차원 원형』→『원형』을 선택한다.

[그림 2-4]
혼인상태에 대한
원그래프(1)



[단계 2] 차트도구의 서식, 레이아웃, 디자인 메뉴를 이용하여 원그래프를 편집한다.

[그림 2-5]
혼인상태에 대한 원그래프(2)



2-2. 양적 자료의 그래프

학습목표

- **일변량의 양적 자료의 특징을 알아보기 위해서 작성하는 히스토그램을 엑셀을 이용하여 작성하는 것을 학습한다.**

1 히스토그램

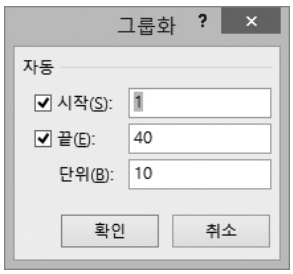
엑셀로 히스토그램을 작성하기 위해서는 평균흡연량의 자료를 구간으로 만들고, 구간의 빈도가 구해져 있어야 한다.

[단계 1] 피벗 테이블을 이용하여 평균흡연량에 대한 빈도를 구한다.

[그림 2-6]
평균흡연량의 빈도(1)

[단계 2] 평균흡연량을 구간으로 변경하기 위해서 완성된 피벗 테이블을 선택하고, 메뉴의『분석』→『그룹필드』를 선택한다.

[그림 2-7] 그룹화

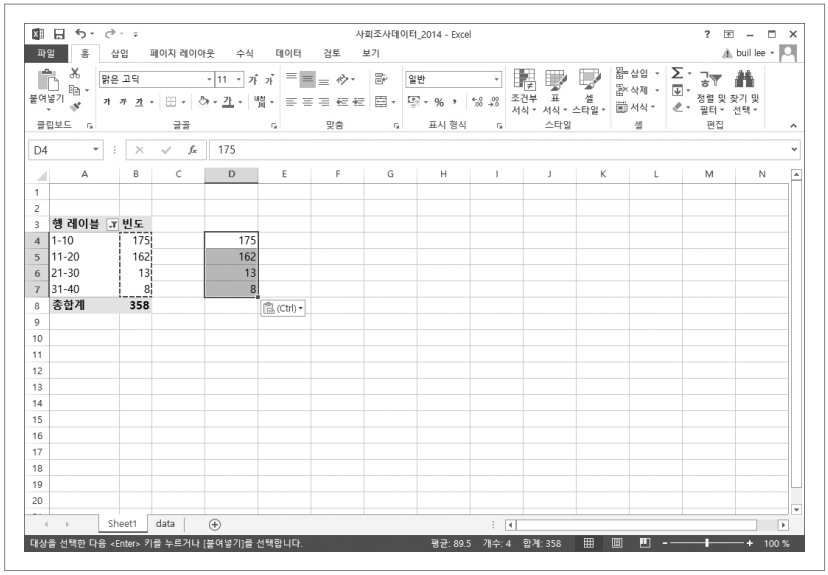


[단계 3] 그룹화 화면은 평균흡연량의 최소값과 최대값을 알려준다. 사용자가 원하는 구간을 위해서 시작, 끝, 단위의 값을 수정한다. 여기서는 그래로 사용하며, 확인 버튼을 선택한다.

[그림 2-8]
평균흡연량의 빈도(2)

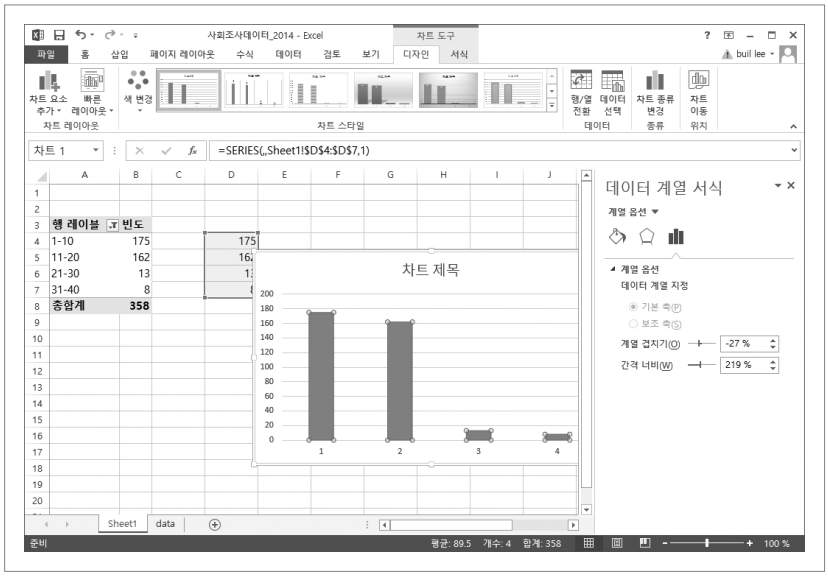
[단계 4] 빈도를 복사하여 D4셀에 넣는다.

[그림 2-9]
평균흡연량의 빈도(3)



[단계 5] 복사된 빈도를 블록잡고, 메뉴의 『삽입』→『차트』→『세로 막대형』→『2차원 세로 막대형』→『묶은 세로 막대형』을 선택한다.

[그림 2-10]
평균흡연량에 대한
히스토그램(1)

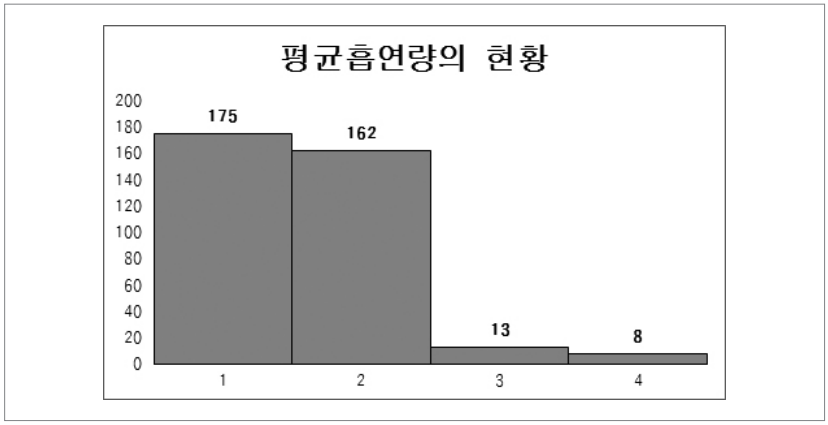


[단계 6] 막대를 선택하고, 마우스 오른쪽을 선택하여『데이터 레이블 서식』을 선택한다.

[단계 7] 간격 너비 219%를 0%로 수정한다.

[단계 8] 차트도구의 서식, 레이아웃, 디자인 메뉴를 이용하여 히스토그램을 편집한다.

[그림 2-11]
평균흡연량에 대한 히스토그램(2)



2-3. 이변량 자료의 그래프

학습목표

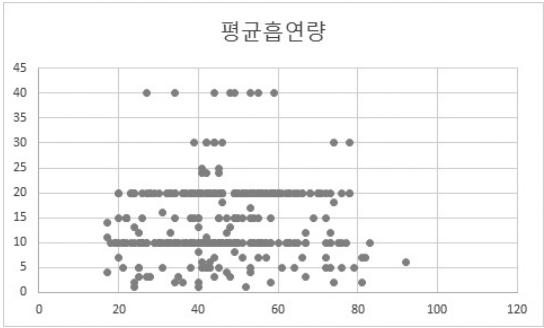
- **두 개의 양적 자료 간의 관계를 알아보기 위해서 작성하는 산점도를 엑셀을 이용하여 작성하는 것을 학습한다.**

1 나이와 평균흡연량에 대한 산점도를 작성한다.

[단계 1] 나이와 평균흡연량이 있는 C열과 Q열을 블록 잡는다.

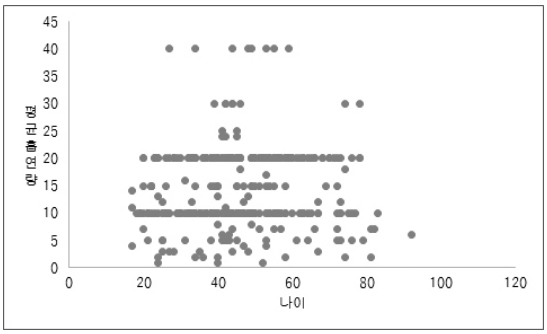
[단계 2] 메뉴의『삽입』→『차트』→『분산형』→『분산형』을 선택한다.

[그림 2-12] 엑셀을 이용한 산점도(1)



[단계 3] 차트도구의 서식, 레이아웃, 디자인 메뉴를 이용하여 산점도를 편집한다.

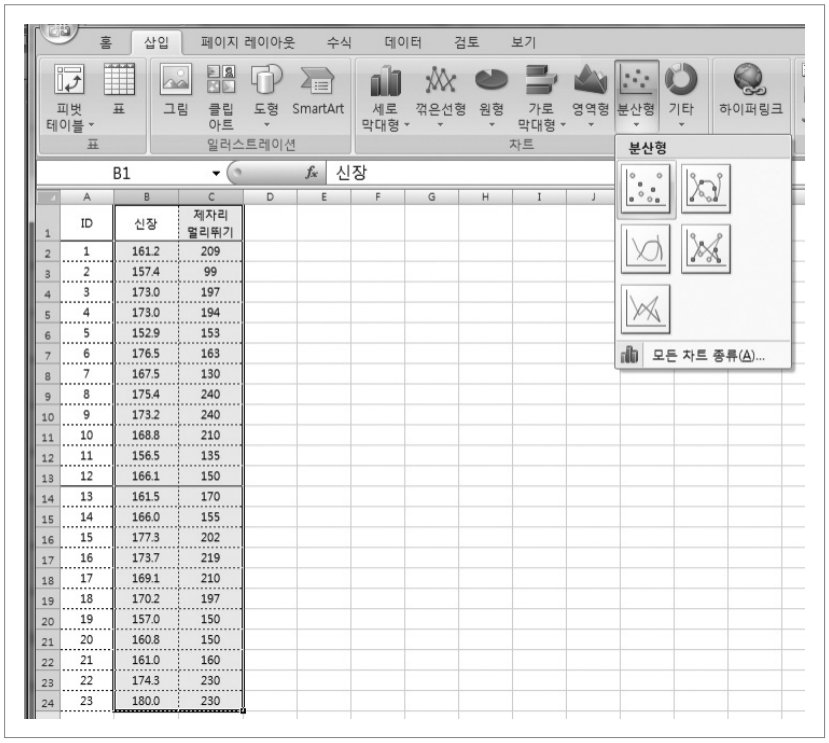
[그림 2-13] 엑셀을 이용한 산점도(2)



❷ 6-1절의 다변량 자료의 특징에서 다루었던 2013년 국민체력실태조사의 일부인 23명의 신장과 제자리멀리뛰기 자료를 이용하여 산점도를 그려보자.

[단계 1] 신장과 제자리멀리뛰기 데이터가 있는 B열과 C열을 블록 잡는다.

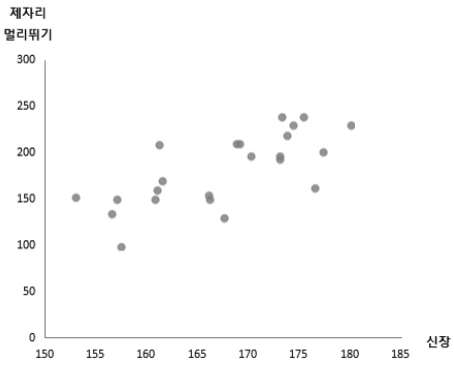
[그림 2-14] 데이터 선택 및 삽입도구



	A	B	C
1	ID	신장	제자리 멀리뛰기
2	1	161.2	209
3	2	157.4	99
4	3	173.0	197
5	4	173.0	194
6	5	152.9	153
7	6	176.5	163
8	7	167.5	130
9	8	175.4	240
10	9	173.2	240
11	10	168.8	210
12	11	156.5	135
13	12	166.1	150
14	13	161.5	170
15	14	166.0	155
16	15	177.3	202
17	16	173.7	219
18	17	169.1	210
19	18	170.2	197
20	19	157.0	150
21	20	160.8	150
22	21	161.0	160
23	22	174.3	230
24	23	180.0	230

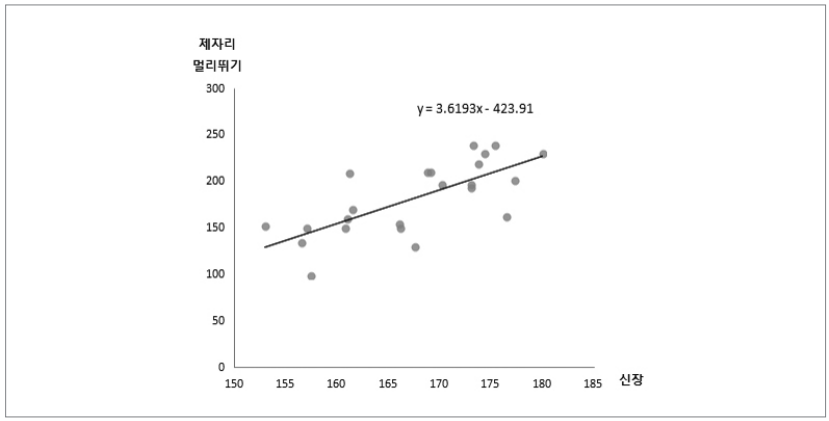
[단계 2] 메뉴의『삽입』→『차트』→『분산형』→『분산형』을 선택하고 차트 도구를 이용하여 편집한다.

[그림 2-15] 엑셀을 이용한 산점도(3)



[단계 3] 메뉴의『차트도구』→『레이아웃』→『추세선』→『선형추세선』을 선택하면 회귀선을 함께 그릴 수 있다. 수식을 산점도에 삽입하려면『차트도구』→『레이아웃』→『추세선』→『기타 추세선 옵션』에서 수식을 차트에 표시를 체크해준다.

[그림 2-16]
엑셀을 이용한 산점도(4)



제 3 장

기술통계량

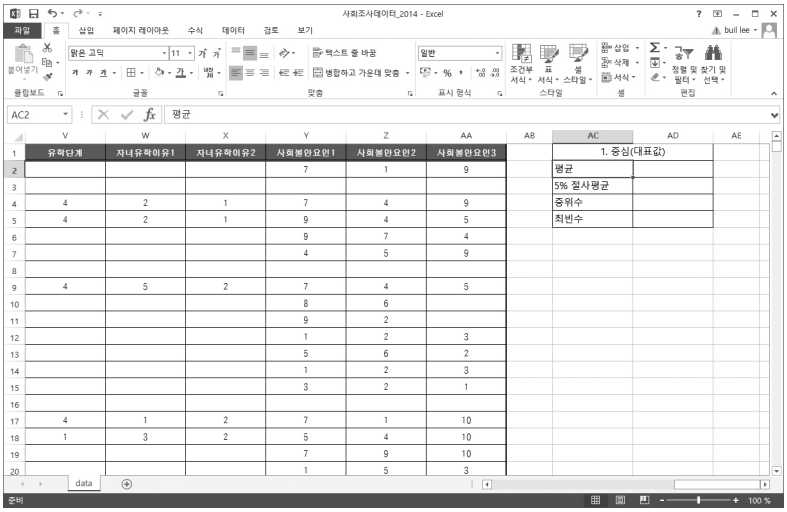
3-1. 중심 (대표값)

학습목표

- 일변량의 양적 자료의 특징 중 중심(대표값)을 알려주는 평균, 절사평균, 중위수, 최빈수를 엑셀로 구하는 방법을 학습한다.

❶ 평균흡연량의 자료에 대한 중심(대표값)을 구한다.

[그림 3-1] 기술통계량의 중심 (대표값)



	V	W	X	Y	Z	AA
1	유학단계	자녀유학이유1	자녀유학이유2	사회불안요인1	사회불안요인2	사회불안요인3
2				7	1	9
3						
4	4	2	1	7	4	9
5	4	2	1	9	4	5
6				9	7	4
7				4	5	9
8						
9	4	5	2	7	4	5
10				8	6	
11				9	2	
12				1	2	3
13				5	6	2
14				1	2	3
15				3	2	1
16						
17	4	1	2	7	1	10
18	1	3	2	5	4	10
19				7	9	10
20				1	5	3

1. 평균

[단계 1] AD2 셀에『=average(q2:q2298)』을 입력하고, 엔터를 친다.

[그림 3-2] 평균(1)

	AC	AD
1. 중심(대표값)		
평균		14.42178771
5% 절사평균		
중위수		
최빈수		

[단계 2] 셀서식을 이용하여 평균의 소수점 자리수를 조정한다. 여기서는 소수점 둘째자리까지 표현한다(참고로 소수점 셋째 자리에서 반올림된 결과가 된다).

[그림 3-3] 평균(2)

	AC	AD
1. 중심(대표값)		
평균		14.42
5% 절사평균		
중위수		
최빈수		

2. 5% 절사평균

[단계 1] AD3 셀에『=trimmean(q2:q2298, 0.05)』을 입력하고, 엔터를 친다. 셀서식을 이용하여 소수점의 자리수를 둘째자리까지 표현한다.

[그림 3-4] 절사평균

	AC	AD
1. 중심(대표값)		
평균		14.42
5% 절사평균		14.12
중위수		
최빈수		

3. 중위수

[단계 1] AD4 셀에『=median(q2:q2298)』을 입력하고, 엔터를 친다.

[그림 3-5] 중위수

	AC	AD
1. 중심(대표값)		
평균		14.42
5% 절사평균		14.12
중위수		12
최빈수		

4. 최빈수

[단계 1] AD5 셀에『=mode(q2:q2298)』을 입력하고, 엔터를 친다.

[그림 3-6] 최빈수

AC	AD
1. 중심(대표값)	
평균	14.42
5% 절사평균	14.12
중위수	12
최빈수	20

5. 4장의 한달 생활비 자료를 이용하여 평균과 중앙값을 구해보자.

[단계 1] 데이터가 있는 영역을 블록잡고『=average(B2:B31)』을 입력하면 평균이 계산되고,『=median(B2:B31)』을 입력하여 중앙값을 계산할 수 있다.

[그림 3-7] 엑셀을 이용한 평균과 중앙값

	A	B	C	D	E
1	id	생활비1	생활비2		
2	1	100	100		
3	2	107	107		
4	3	110	110		
5	4	112	112		
6	5	118	118		
7	6	122	122		
8	7	124	124		
9	8	130	130		
10	9	132	132		
11	10	136	136		
12	11	140	140		
13	12	144	144		
14	13	148	148		
15	14	149	149		
16	15	149	149		
17	16	151	151		
18	17	164	164		
19	18	168	168		
20	19	172	172		
21	20	176	176		
22	21	180	180		
23	22	184	184		
24	23	200	200		
25	24	205	205		
26	25	219	219		
27	26	225	225		
28	27	235	235		
29	28	245	245		
30	29	255	255		
31	30	400	2500		
32	평균	=AVERAGE(B2:B31)			
33		AVERAGE(number1, [number2], ...)			
34	중앙값	=MEDIAN(B2:B31)			
35		MEDIAN(number1, [number2], ...)			

3-2. 퍼짐(산포)

학습목표

- **일변량의 양적 자료의 특징 중 퍼짐(산포)을 알려주는 범위, 사분위범위, 분산, 표준편차를 엑셀로 구하는 방법을 학습한다.**

❶ 평균흡연량의 자료에 대한 퍼짐(산포)을 구한다.

[그림 3-8] 기술통계량의 퍼짐(산포)

1. 범위

[단계 1] AD9 셀에『=max(q2:q2298) - min(q2:q2298)』을 입력하고, 엔터를 친다. 참고로『max(q2:q2298)』는 최댓값을 구해주는 함수이며,『min(q2:q2298)』은 최솟값을 구해준다. 두 값의 차이가 범위가 된다.

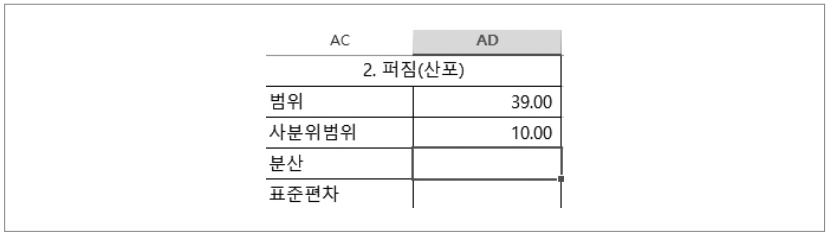
[그림 3-9] 범위

AC	AD
2. 퍼짐(산포)	
범위	39.00
사분위범위	
분산	
표준편차	

2. 사분위범위

[단계 1] AD10 셀에『=quartile(q2:q2298, 3) - quartile(q2:q2298, 1)』을 입력하고, 엔터를 친다. 참고로『quartile(q2:q2298, 3) 』은 제3사분위수를 구해주는 함수로 마지막에 입력된 '3'은 몇 번째 사분위수를 구할 것인지를 알려주는 인수이다. 따라서『quartile(q2:q2298, 1) 』은 제1사분위수을 구해주며 만일『quartile(q2:q2298, 2)』라고 입력하면 제2사분위수인 중앙값을 구해준다. 제3사분위수와 제1사분위수의 차이가 사분위범위가 된다.

[그림 3-10] 사분위범위

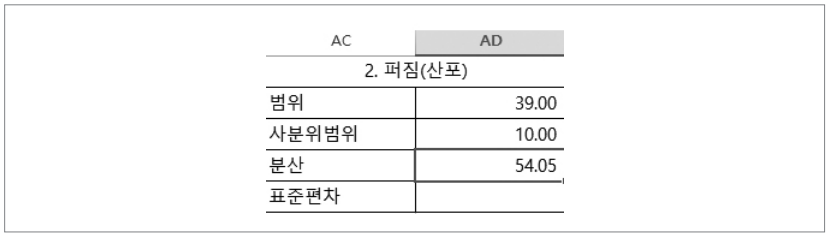


AC	AD
2. 퍼짐(산포)	
범위	39.00
사분위범위	10.00
분산	
표준편차	

3. 분산

[단계 1] AD11 셀에『=var(q2:q2298)』을 입력하고, 엔터를 친다. 셀서식을 이용하여 소수점의 자리수를 둘째자리까지만 표현한다. 참고로 var() 함수는 표본분산을 구해주며, 모분산을 구할 경우에는 varp() 함수를 사용하면 된다.

[그림 3-11] 분산



AC	AD
2. 퍼짐(산포)	
범위	39.00
사분위범위	10.00
분산	54.05
표준편차	

4. 표준편차

[단계 1] AD12 셀에『=stdev(q2:q2298)』을 입력하고, 엔터를 친다. 셀서식을 이용하여 소수점의 자리수를 둘째자리까지만 표현한다. 참고로 stdev() 함수는 표본의 표준편차를 구해주며, 모집단의 표준편차를 구할 경우에는 stdevp() 함수를 사용하면 된다.

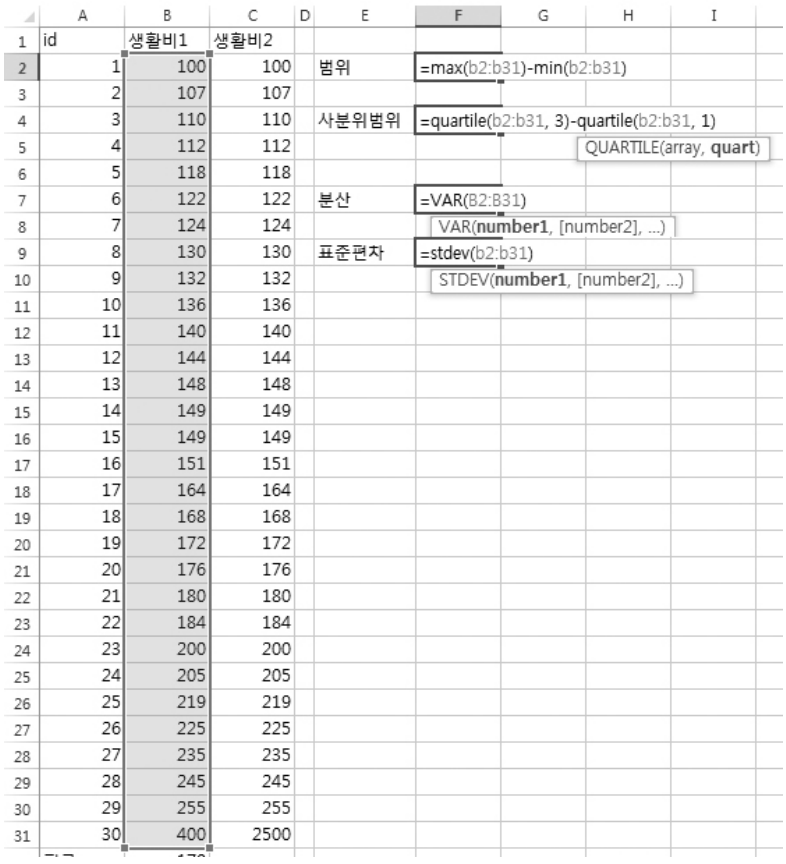
[그림 3-12]
표준편차

AC	AD
2. 퍼짐(산포)	
범위	39.00
사분위범위	10.00
분산	54.05
표준편차	7.35

5. 4장의 한달 생활비 자료를 이용하여 범위와 사분위범위, 분산과 표준편차를 구해보자.

[단계 1] 데이터가 있는 영역을 블록잡고『=max(B2:B31) - min(B2:B31)』을 입력하면 범위가 계산되고,『=quartile(B2:B31, 3) - quartile(B2:B31, 1)』을 입력하여 사분위범위를 계산할 수 있다. 분산은『=var(B2:B31)』, 표준편차는『=stdev(B2:B31)』로 구할 수 있다.

[그림 3-13]
엑셀을 이용한 범위, 사분위범위, 분산, 표준편차



	A	B	C	D	E	F
1	id	생활비1	생활비2			
2	1	100	100		범위	
3	2	107	107			
4	3	110	110		사분위범위	
5	4	112	112			
6	5	118	118			
7	6	122	122		분산	
8	7	124	124			
9	8	130	130		표준편차	
10	9	132	132			
11	10	136	136			
12	11	140	140			
13	12	144	144			
14	13	148	148			
15	14	149	149			
16	15	149	149			
17	16	151	151			
18	17	164	164			
19	18	168	168			
20	19	172	172			
21	20	176	176			
22	21	180	180			
23	22	184	184			
24	23	200	200			
25	24	205	205			
26	25	219	219			
27	26	225	225			
28	27	235	235			
29	28	245	245			
30	29	255	255			
31	30	400	2500			

3-3. 분포의 모양

학습목표

- **일변량의 양적 자료의 특징 중 분포의 모양을 알려주는 왜도, 첨도를 엑셀로 구하는 방법을 학습한다.**
- **엑셀의 추가기능 중 데이터 분석을 설치하고. 데이터 분석 중에서 기술 통계법을 이용하여 일반량의 양적 자료에 대한 기술통계량을 구하는 것을 학습한다.**
- **피벗 테이블을 이용하여 일변량의 양적 자료에 대한 기술통계량을 구하는 것을 학습한다.**

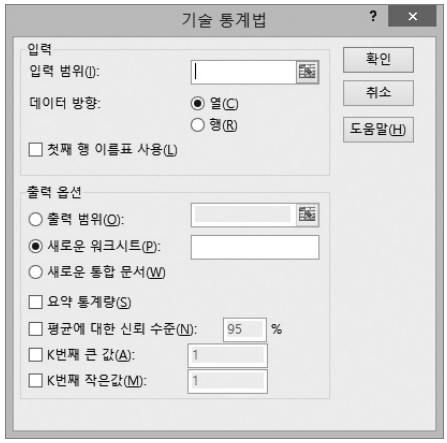
❶ 평균흡연량의 자료에 대한 기술통계량을 구한다.

1. 데이터분석의 기술통계법을 이용하여 기술통계량 구하기

[단계 1] 엑셀 메뉴의『파일』→『옵션』→『추가 기능』→『이동』→『분석 도구 체크』→『확인』버튼을 클릭하여 데이터 분석 기능을 설치한다.

[단계 2] 엑셀 메뉴의『데이터』→『데이터 분석』→『기술 통계법』을 선택한다.

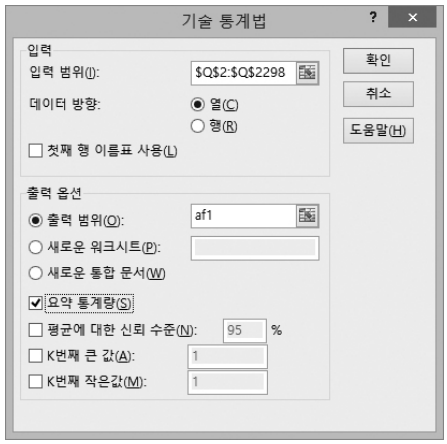
[그림 3-14]
기술통계법(1)



[단계 3] 기술 통계법 화면에 다음과 같은 내용을 입력한다.

- 입력 범위 : q2:q2298
- 출력범위 : af1
- 요약 통계량 : 체크

[그림 3-15]
기술통계법(2)



[단계 4] 확인 버튼을 클릭한다.

[그림 3-16]
기술통계법(3)

AF	AG
Column1	
평균	14.42179
표준 오차	0.388553
중앙값	12
최빈값	20
표준 편차	7.351768
분산	54.04849
첨도	1.58252
왜도	0.874229
범위	39
최소값	1
최대값	40
합	5163
관측수	358

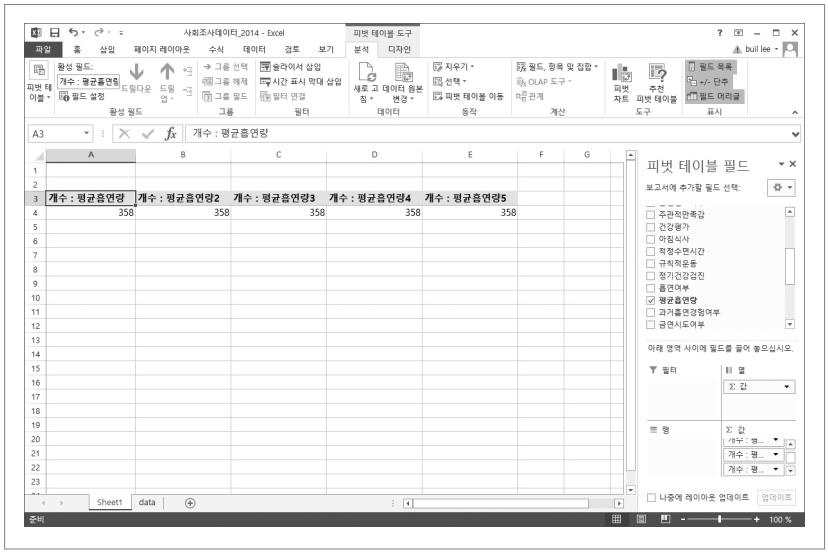
참고로 여러 개의 양적 자료에 대해서 데이터 분석의 기술통계법을 이용하려면, 양적 자료들을 연속적으로 있도록 한 다음에 사용하면 한 번에 여러 개의 양적 자료에 대한 기술통계량을 구할 수 있다.

2. 피벗 테이블을 이용한 기술통계량 구하기

[단계 1] 데이터 중에서 임의로 하나를 선택한 다음에, 엑셀 메뉴의『삽입』→『피벗 테이블』을 선택하고, 확인 버튼을 클릭한다.

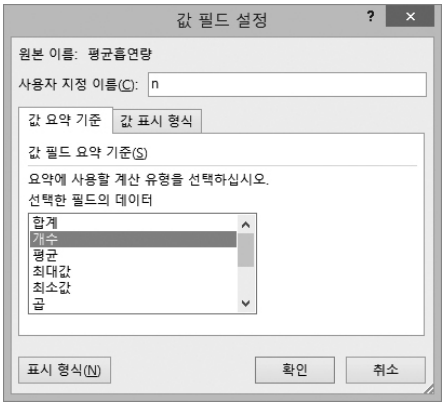
[단계 2] 피벗 테이블 필드에 있는『평균흡연량』을 드래그하여 값에 다섯 번 가져다 넣는다.

[그림 3-17]
피벗 테이블(1)



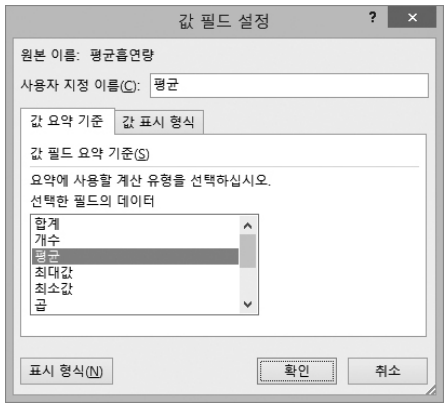
[단계 3] 값에 있는 첫 번째『평균흡연량』을 선택하여 값 필드설정 화면에서 함수는『개수』, 사용자 지정 이름은『n』을 입력하고, 확인 버튼을 클릭한다.

[그림 3-18]
피벗 테이블 : 빈도



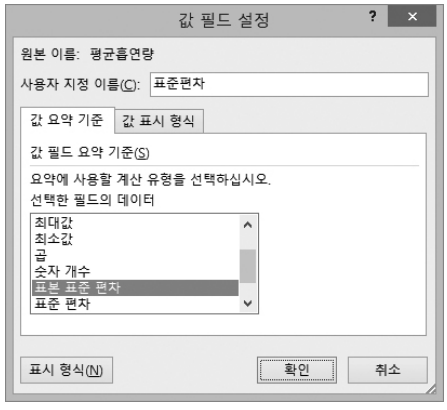
[단계 4] 값에 있는 두 번째『평균흡연량』을 선택하여 값 필드설정 화면에서 함수는『평균』, 사용자 지정 이름은『평균』을 입력하고, 표시 형식에서 소수점 자리수를 2로 수정한 다음에 확인 버튼을 클릭한다.

[그림 3-19]
피벗 테이블 : 평균



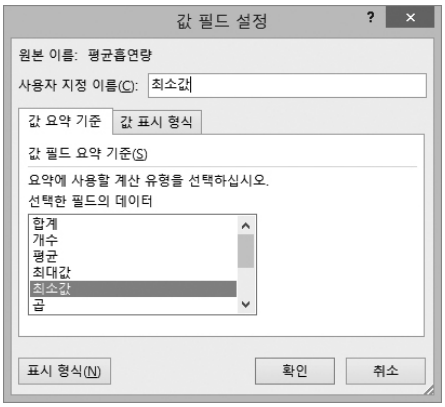
[단계 5] 값에 있는 세 번째『평균흡연량』을 선택하여 값 필드설정 화면에서 함수는『표본 표준편차』, 사용자 지정 이름은『표준편차』를 입력하고, 표시 형식에서 소수점 자리수를 2로 수정한 다음에 확인 버튼을 클릭한다.

[그림 3-20]
피벗 테이블 : 표준편차



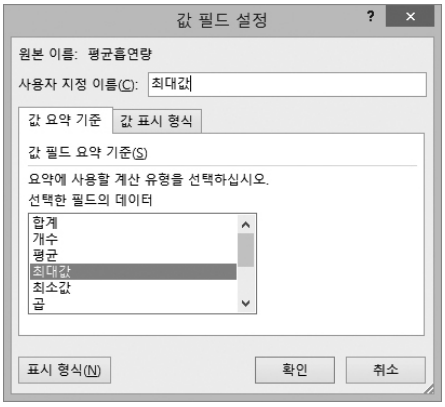
[단계 6] 값에 있는 네 번째『평균흡연량』을 선택하여 값 필드설정 화면에서 함수는『최소값』, 사용자 지정 이름은『최소값』을 입력하고 확인 버튼을 클릭한다.

[그림 3-21]
피벗 테이블 : 최소값



[단계 7] 값에 있는 다섯 번째『평균흡연량』을 선택하여 값 필드설정 화면에서 함수는『최대값』, 사용자 지정 이름은『최대값』을 입력하고 확인 버튼을 클릭한다.

[그림 3-22]
피벗 테이블 : 최대값

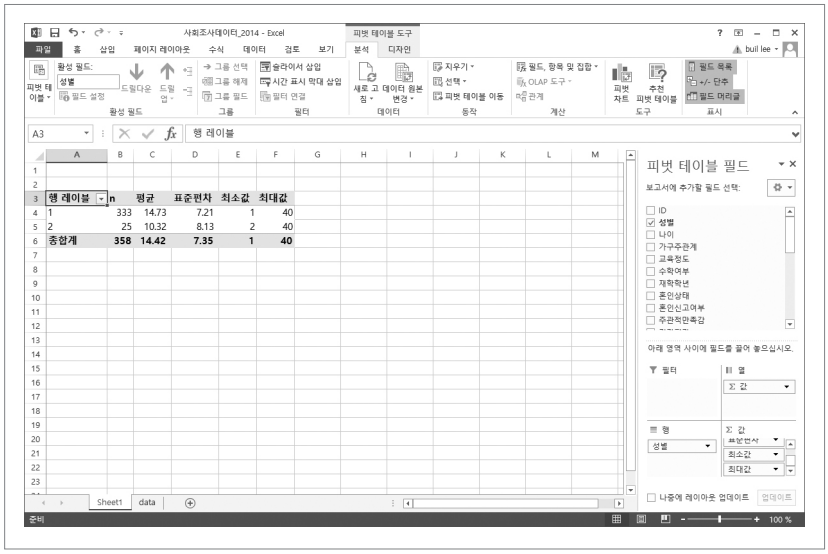


[그림 3-23]
피벗 테이블 :
기술통계량

	A	B	C	D	E
1					
2					
3	**n**	**평균**	**표준편차**	**최소값**	**최대값**
4	358	14.42	7.35	1	40

피벗 테이블은 절사평균, 중위수, 범위, 왜도, 첨도를 구할 수 없다. 하지만 피벗 테이블의 강점은 집단별로 기술통계량을 구할 때에는 매우 유용하다. 집단에 해당하는 질적 자료를 드래그 하여 행이나 열에 가져다 넣으면 된다. 여기서는 성별에 대한 평균흡연량의 기술통계량을 구해본다.

[그림 3-24]
피벗 테이블 : 집단별 기술통계량



제 4 장 모집단의 분포

4-1. 정규분포

학습목표

- 엑셀을 이용하여 정규분포에서의 확률을 구하는 방법을 학습한다.
- 엑셀을 이용하여 정규분포에서의 확률변수를 구하는 방법을 학습한다.

1 모집단의 평균이 170cm이고, 표준편차가 10cm인 정규분포를 따른다고 할 때 다음을 구하시오.

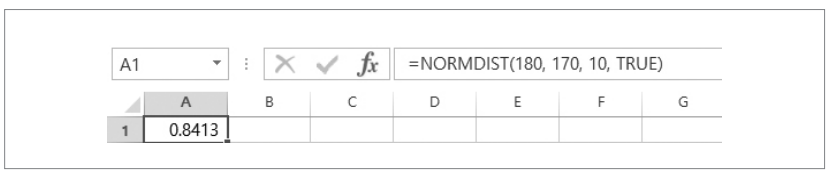
<표 4-1> normdist( ) 함수

인수	설명
x	정규분포에서 구하려고 하는 값
mean	정규분포에서의 모평균
standard_dev	정규분포에서의 모표준편차
cumulative	함수의 형태를 결정하는 논리값으로서, cumulative가 TRUE이면 누적 분포 함수가 반환되고 FALSE이면 확률 질량 함수가 반환

1. 180cm 이하는 전체 중에서 얼마나 있는가?

[단계 1] 빈 셀에『=normdist(180, 170, 10, TRUE)』를 입력한다.

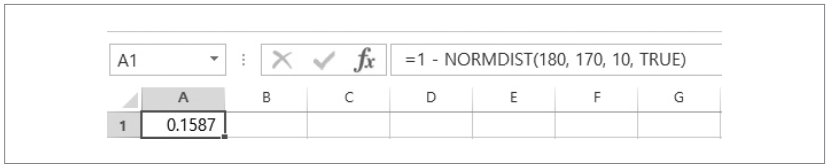
[그림 4-1] 정규분포(1)



2. 180cm 이상은 전체 중에서 얼마나 있는가?

[단계 1] 빈 셀에『=1 - normdist(180, 170, 10, TRUE)』를 입력한다.

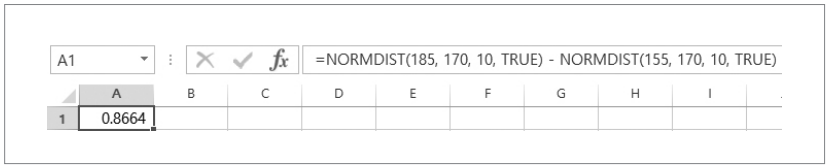
[그림 4-2]
정규분포(2)



3. 155cm ~ 185cm 사이는 전체 중에서 얼마나 있는가?

[단계 1] 빈 셀에『=normdist(185, 170, 10, TRUE) - normdist(155, 170, 10, TRUE)』를 입력한다.

[그림 4-3]
정규분포(3)



❷ 모집단의 평균이 170cm이고, 표준편차가 10cm인 정규분포를 따른다고 할 때 다음을 구하시오.

<표 4-2>
norminv( ) 함수

인수	설명
probability	정규분포에서의 확률
mean	정규분포에서의 모평균
standard_dev	정규분포에서의 모표준편차

1. 상위 5%에 해당하는 키는 얼마인가?

[단계 1] 빈 셀에『=norminv(0.95, 170, 10)』를 입력한다.

[그림 4-4]
정규분포(4)

2. 상위 1%에 해당하는 키는 얼마인가?

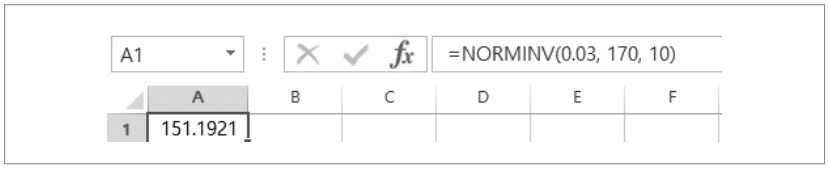
[단계 1] 빈 셀에『=norminv(0.99, 170, 10)』를 입력한다.

[그림 4-5]
정규분포(5)

3. 하위 3%에 해당하는 키는 얼마인가?

[단계 1] 빈 셀에『=norminv(0.03, 170, 10)』를 입력한다.

[그림 4-6]
정규분포(6)

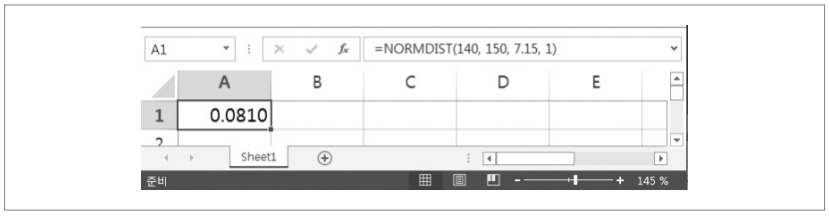


4. 13장과 15장에서 토의하였던 내용을 엑셀로 구해보자.

(1) 초등학교 6학년 남학생의 신장이 평균 150cm에 표준편차 7.15cm라면 140cm인 수민이의 키는 하위 몇 %인가?

→『=normdist(140, 150, 7.15, TRUE)』를 입력하면 0.0810이 계산되며 하위 8.1%에 속하는 값임을 알 수 있다.

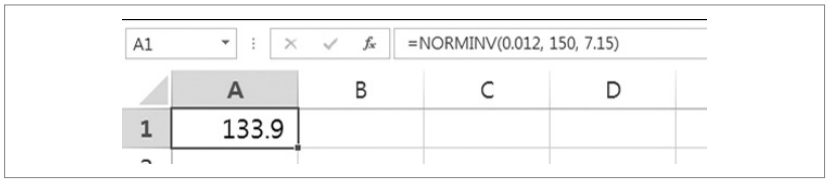
[그림 4-7]
정규분포(7)



(2) 초등학교 6학년 남학생의 신장이 평균 150cm에 표준편차 7.15cm라면 하위 1.2%에 해당하는 값은 얼마인가?

→『=norminv(0.012, 150, 7.15)』를 입력하면 이 계산되며 133.9cm가 하위 1.2%의 값임을 알 수 있다.

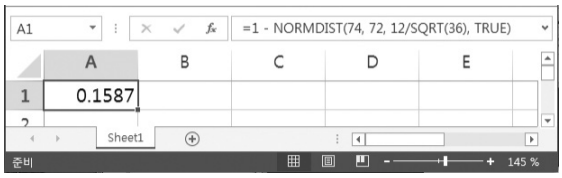
[그림 4-8]
정규분포(8)



(3) 서울시민들의 행복점수가 평균 72점에 표준편차 12점인 정규분포라면, 서울에 살고 있는 K씨가 속한 동호회 회원 36명의 행복점수 74점은 상위 몇 %인가?

→『=1 - normdist(74, 72, 12/sqrt(36), TRUE)』를 입력하면 0.1587이 계산되며 상위 15.87%에 속하는 값임을 알 수 있다.

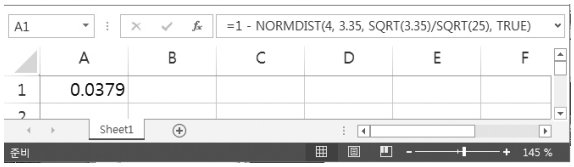
[그림 4-9]
정규분포(9)



(4) SNS 이용횟수가 평균 3.35회인 포아송분포를 따른다면 임씨가 속한 종교모임 회원 25명의 평균 SNS 이용횟수 4회는 상위 몇 %인가? (SNS 이용횟수가 포아송분포라고 해도 표본평균의 분포는 중심극한정리에 의하여 근사적으로 정규분포를 따른다는 것을 기억해야 한다.)

→『=1 - normdist(4, 3.35, sqrt(3.35)/sqrt(25), TRUE)』를 입력하면 0.0379로 계산되며 상위 3.79%에 속하는 값임을 알 수 있다.

[그림 4-10]
정규분포(10)



4-2. 이항분포

학습목표

• 엑셀을 이용하여 이항분포에서의 확률을 구하는 방법을 학습한다.

1 미국 NBA의 전설인 마이클 조던의 자유투 성공률은 80%라고 한다. 마이클 조던이 자유투를 10번 한다고 할 때 다음을 구하시오.

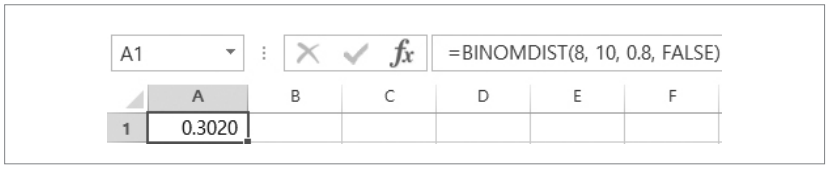
<표 4-3> binomdist( ) 함수

인수	설명
number_s	시행에서의 성공 횟수
trials	독립 시행 횟수
probability_s	각 시행에서 성공할 확률
cumulative	함수의 형태를 결정하는 논리값으로서, cumulative가 TRUE이면 시행이 number_s회 이하로 성공할 확률을 나타내는 누적 분포 함수가 반환되고 FALSE이면 시행이 정확히 number_s회 성공할 확률을 나타내는 확률 질량 함수가 반환

1. 8번 성공할 확률은 얼마인가?

[단계 1] 빈 셀에『=binomdist(8, 10, 0.8, FALSE)』를 입력한다.

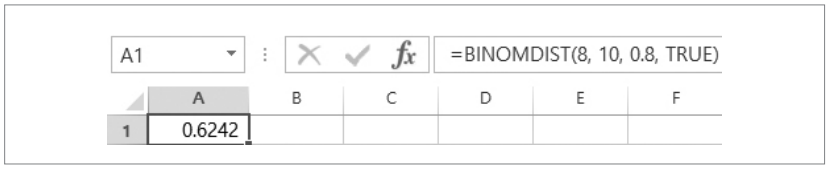
[그림 4-11] 이항분포(1)



2. 8번 이하로 성공할 확률은 얼마인가?

[단계 1] 빈 셀에『=binomdist(8, 10, 0.8, TRUE)』를 입력한다.

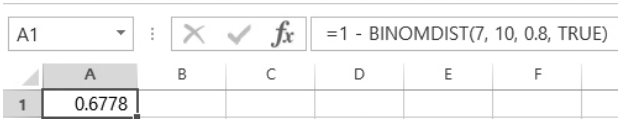
[그림 4-12] 이항분포(2)



3. 8번 이상으로 성공할 확률은 얼마인가?

[단계 1] 빈 셀에『=1 - binomdist(7, 10, 0.8, TRUE)』를 입력한다.

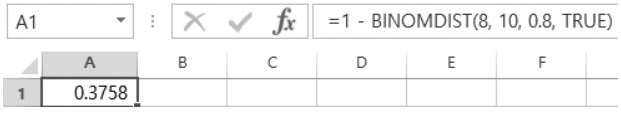
[그림 4-13]
이항분포(3)



4. 8번 초과로 성공할 확률은 얼마인가?

[단계 1] 빈 셀에『=1 - binomdist(8, 10, 0.8, TRUE)』를 입력한다.

[그림 4-14]
이항분포(4)

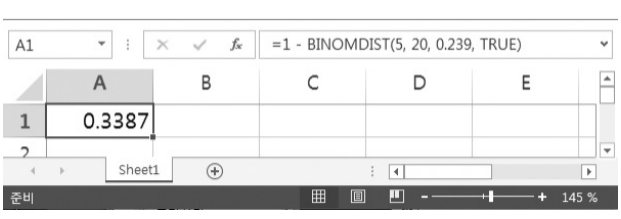


5. 10장의 이항분포 예에서 60대의 23.9%가 자기건강에 대하여 긍정적으로 평가하고 있다면 어느 동아리 회원 20명 중에서

(1) 6명 이상이 긍정적으로 평가할 확률은 얼마나 될까?

→『=1 - binomdist(5, 20, 0.239, TRUE)』를 입력하면 0.3387이 계산된다.

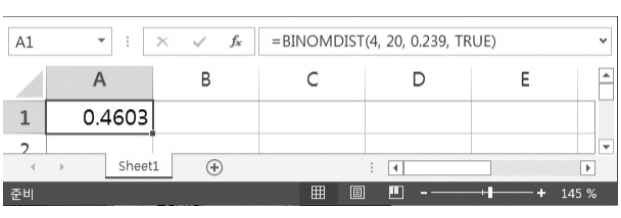
[그림 4-15]
이항분포(5)



(2) 4명 이하가 긍정적으로 평가할 확률은 얼마나 될까?

→『=binomdist(4, 20, 0.239, TRUE)』를 입력하면 0.4603이 계산된다.

[그림 4-16]
이항분포(6)



4-3. 포아송분포

학습목표

• 엑셀을 이용하여 포아송분포에서의 확률을 구하는 방법을 학습한다.

1 한국의 특허 출원건수는 연도별 평균 9건이라고 할 때, 다음을 구하시오.

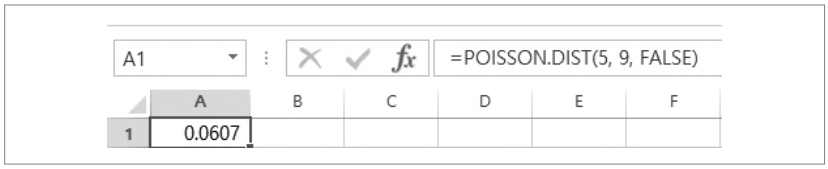
<표 4-4> poisson.dist( ) 함수

인수	설명
x	사건의 수
trials	기대값
cumulative	반환되는 확률 분포의 형태를 결정하는 논리값으로서, cumulative가 TRUE이면 임의의 사건이 발생하는 횟수가 0에서 x 사이에 있는 누적 포아송 확률 분포 함수가 반환되고, FALSE이면 사건이 정확히 x회 발생하는 포아송 확률 질량 함수가 반환

1. 연간 5건의 특허 출원건수가 발생할 확률은?

[단계 1] 빈 셀에『=poisson.dist(5, 9, FALSE)』를 입력한다.

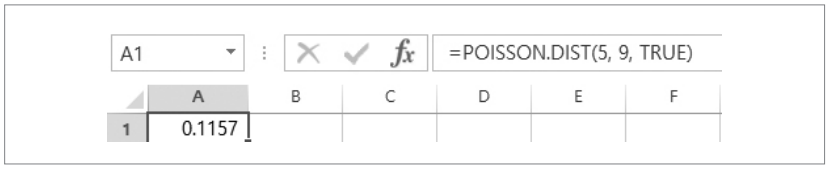
[그림 4-17] 포아송분포(1)



2. 연간 5건 이하의 특허 출원건수가 발생할 확률은?

[단계 1] 빈 셀에『=poisson.dist(5, 9, TRUE)』를 입력한다.

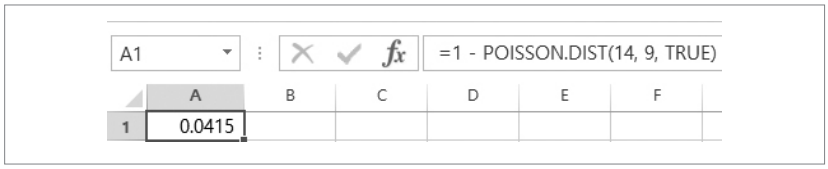
[그림 4-18] 포아송분포(2)



3. 연간 15건 이상의 특허 출원건수가 발생할 확률은?

[단계 1] 빈 셀에『=1 - poisson.dist(14, 9, TRUE)』를 입력한다.

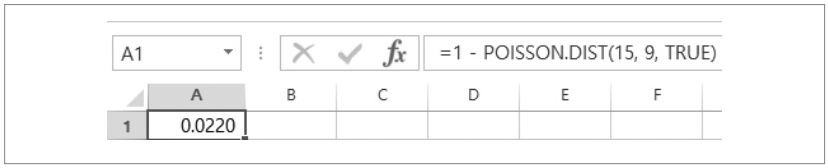
[그림 4-19]
포아송분포(3)



4. 연간 15건 초과의 특허 출원건수가 발생할 확률은?

[단계 1] 빈 셀에『=1 - poisson.dist(15, 9, TRUE)』를 입력한다.

[그림 4-20]
포아송분포(4)

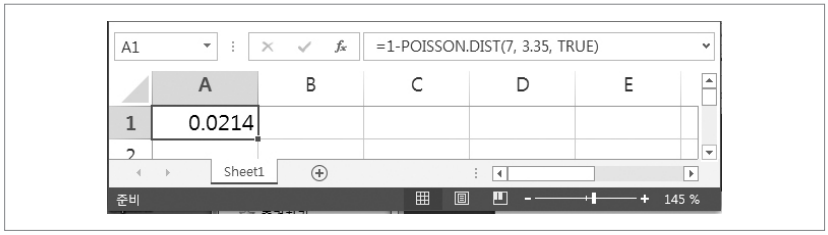


5. 10장과 13장에서 토의하였던 내용을 엑셀로 구해보자.

(1) SNS 이용횟수가 평균 3.35회인 포아송분포를 따른다고 할 때 SNS를 평균 8회 이용하는 임씨는 상위 몇 %에 속하는가?

→『=1 – poisson.dist(7, 3.35, TRUE)』를 입력하면 0.0214로 계산되며 상위 2.14%에 속하는 값임을 알 수 있다.

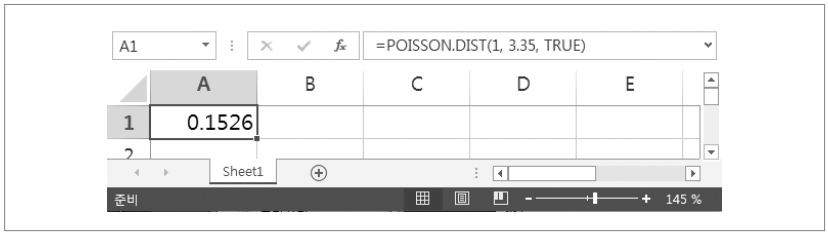
[그림 4-21]
포아송분포(5)



(2) 만일 임씨의 친구가 평균 1회 이용한다면 임씨의 친구는 하위 몇 %에 속하는가?

→『=poisson.dist(1, 3.35, TRUE)』를 입력하면 0.1526로 계산되며 하위 15.26%에 속한다.

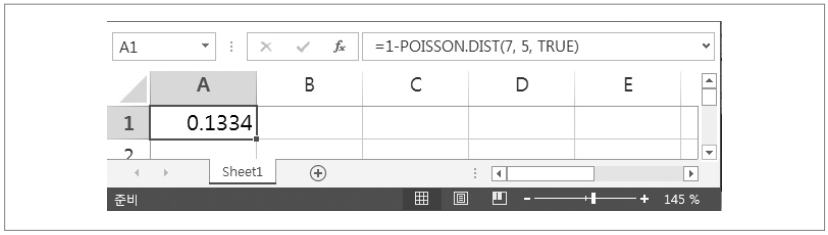
[그림 4-22]
포아송분포(6)



(3) 2014년 10대의 10만명당 자살자 수는 약 5명이다(http://www.spckorea.or.kr/ index.php). 2016년에도 2014년과 동일한 상황이 계속된다면 10대의 10만명당 자살자수가 8명 이상 될 가능성은 얼마나 되는가?

→『=1 – poisson.dist(7, 5, TRUE)』를 입력하면 0.1334로 계산되며 상위 13.34%에 속하는 값임을 알 수 있다.

[그림 4-23]
포아송분포(7)



제 5 장 종합 실습

5-1. 종합 실습

학습목표

- **4시간 동안 학습했던 일변량의 범주형 자료와 양적 자료에 대한 기초적인 통계분석을 복습한다.**
- **복습을 통하여 학습내용을 숙지하도록 한다.**
- **수업생의 실무 데이터를 가지고 학습해 보도록 한다.**

1. 성별, 교육기대정도, 이전수강여부에 대한 빈도표를 다음과 같이 작성하시오.

인구특성	구분	빈도	백분율(%)
성별	남자		
	여자		
	합계		
교육기대정도	매우 아니다		
	아니다		
	보통이다		
	그렇다		
	매우 그렇다		
	합계		
이전수강여부	아니오		
	예		
	합계		

2. 성별, 교육기대정도, 이전수강여부에 대한 빈도를 이용하여 막대그래프를 작성하시오.

3. 평가결과에 대한 빈도표를 다음과 같이 작성하시오.

구분	빈도	백분율(%)
20점 미만		
20점 이상 ~ 40점 미만		
40점 이상 ~ 60점 미만		
60점 이상 ~ 80점 미만		
80점 이상		
합계		

4. 평가결과에 대한 빈도표의 빈도를 이용하여 히스토그램을 작성하시오.

5. 환경만족도와 평가결과, 내용만족도와 평가결과에 대한 산점도를 작성하시오.

6. 평가결과에 대한 기술통계량을 구하시오.

연구과제 또는 연습문제

■ 단원 1.

• 교육정도, 혼인상태, 주관적만족감에 대한 빈도표를 작성하시오.

• 성별과 혼인상태에 대한 교차표를 작성하시오(단 교차표에는 빈도와 행백분율만 나타나도록 합니다).

• 흡연여부와 음주여부에 따라 규칙적운동의 현황을 살펴보려고 한다. 흡연여부, 음주여부, 규칙적운동에 대한 다차원 교차표를 작성하시오.

■ 단원 2.

• 수학여부에 대한 빈도표를 작성하시오.

• 수학여부에 대한 빈도를 이용하여 막대그래프를 작성하시오.

• 수학여부에 대한 백분율을 이용하여 원그래프를 작성하시오.

• 나이에 20대 미만, 30대, 40대, 50대 이상으로 구분하여 빈도표를 작성하시오.

• 위 문제의 빈도를 가지고 히스토그램을 작성하시오.

■ **단원 3.**

• 나이에 대한 빈도표를 작성하시오.

구분	빈도	백분율(%)
20대 미만		
20대		
30대		
40대		
50대 이상		
합계		

• 나이에 대한 기술통계량을 다음과 같이 작성하시오.

n	평균	중위수	중위수	최빈수	표준편차	왜도	첨도

• 성별에 대한 나이의 기술통계량을 다음과 같이 작성하시오.

성별	n	평균	표준편차	최소값	최대값
남자					
여자					

■ **단원 4.**

• 2015년도 공무원 전체의 월소득의 평균이 467만원, 표준편차가 33만원인 정규분포를 따른다고 할 때, 다음을 구하시오.

(1) 월소득이 600만원 이상은 몇 %가 존재하는가?

(2) 상위 0.1%에 해당하는 월소득은 얼마인가?

• A라는 병아리 감별사의 암컷과 숫컷에 대한 감별률(성공률)이 85%라고 하며, 새로운 50마리의 병아리에 대해서 감별을 하려고 할 때, 다음을 구하시오.

(1) 45마리 이상을 잘 감별할 확률은 얼마인가?

(2) 5마리 이하로 감별할 확률은 얼마인가?

• 2014년도의 자동차 1만대당 사망자수의 평균이 2명이라고 한다.

(1) 자동차 1만대당 5명 이상이 사망할 확률은 얼마인가?

(2) 자동차 1만대당 1명 이하로 사망할 확률은 얼마인가?

참고자료

- 엑셀을 활용한 통계자료분석(기초편)(2012), 경문사.
- 월스트리트저널 인포그래픽 가이드(2014), 인사이트(insight).
- EXCEL 활용한 통계학(2015), One(원).

교재 개발진

■ **교재개발 책임연구원**

김용환 통계진흥원 부장

■ **교재 집필진**

이석훈 충남대학교 교수

이부일 충남대학교 강사

■ **교재 검토위원**

변효섭 통계교육원 명예교수

최봉호 통계교육원 명예교수

안형진 고려대학교 교수

이광진 목원대학교 교수

이기성 우석대학교 교수

백지선 통계개발원 사무관

심규호 통계개발원 주무관

통계기초 및 활용

발 행 | 2015년 12월 23일
인 쇄 | 2015년 12월 31일
발행인 | 통계교육원장 박성동
발행처 | 통계교육원
기 획 | 김병우 · 최병연 · 이정만

주 소 | (35220) 대전광역시 서구 한밭대로 713(월평동) 통계센터 5층 통계교육원
전 화 | 042 - 366 - 6232
홈페이지 | http://sti.kostat.go.kr/

발간등록번호 11-1240162-000021-01

통계교육원
STATISTICAL TRAINING INSTITUTE
(35220) 대전광역시 서구 한밭대로 713(월평동)
통계센터 5층 통계교육원