Chapter 1

기초 통계 개념

1.1 왜 통계학인가?

통계학은 데이터의 수집, 처리, 요약, 분석, 해석을 다루는 학문이다. 반면에 과학자와 공학자는 신제품 개발, 자료 및 노동의 효율적인 활용, 생산 문제 해결, 품질 및 신뢰성 향상 그리고 기초 연구와 같은 다양한 사안을 다룬다. 통계학의 유용성은 위와 같은 문제들을 해결하는 도구로 다음과 같은 구체적인 사례를 해결하는 데 가장 좋은 방법으로 쓰인다.

예제 1.1-1

과학 및 공학에서 나타나는 구체적인 사례연구는 다음과 같다.

1. 금속의 열팽창계수 추정
2. 국제선 공항의 우박 및 안개 억제를 위한 두 가지 인공강우 방법 비교
3. 두 가지 이상의 시멘트 제조방법에 다른 압축 강도 비교
4. 네 가지의 얼룩을 제거하는 3개의 청소용품 효율성 비교
5. 작용응력에 기반을 둔 보의 파괴 시간 예측
6. 주간 교통사고율을 줄일 수 있는 교통 규제방법의 효율성 평가
7. 제조업체가 주장하는 제품의 품질 시험
8. 대기업의 급여 인상과 직원들의 생산성 간 관계 연구
9. 태양열 에너지 자원 확대에 찬성하는 18세 이상 미국 시민의 비율 추정
10. 특정 호수 물의 납 성분이 안전 제한치 이하로 함유하고 있는지 검사 ■

위와 같은 사례에서 통계학이 필요한 이유는 **변동성**(variability) 때문이다. 만약 같은 방법으로 제조된 모든 시멘트가 모두 같은 압축 강도를 나타낸다면, 사례연구 3에서 나타난 각각 다른 방법으로 제조된 압축 강도의 비교를 위해 통계학을 이용할 필요가 없을 것이다. 즉, 각 방법으로 제조된 하나의 시멘트 표본의 압축 강도만을 비교하면 충분할 것이다. 하지만 같은 방법으로 제조된 여러 시멘트 표본의 강도는 일반적으로 같지 않다.

그림 1-1
32개 압축 강도 측정결과의 히스토그램

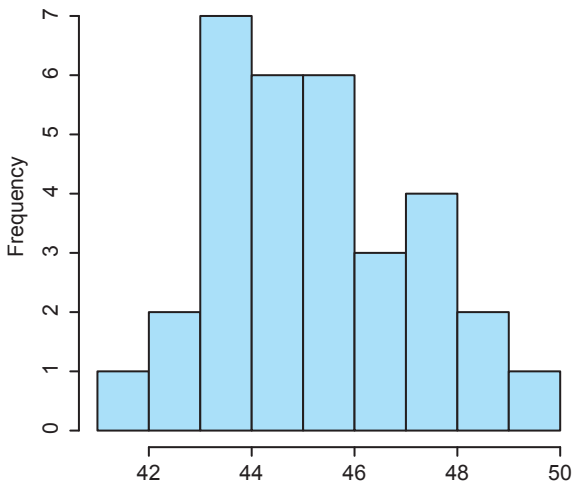


그림 1-1은 32개 압축 강도 측정결과[1]의 히스토그램을 나타낸다. (1.5절의 히스토그램에 대한 논의 참고) 비슷한 사례로, 만약 모든 보가 주어진 응력 수준에서 똑같은 시간에 파괴된다면, 사례연구 5의 예측 문제는 통계학이 필요하지 않을 것이다. 예제 1.1-1에서 주어진 모든 사례연구는 이와 비슷하게 설명할 수 있다.

변동성으로 인해 발생하는 이러한 복잡성을 파악하는 것은 사례연구 3번의 문제점을 알아내는 것으로부터 시작하는데, 이는 언급한 바와 같이 애매모호하다. 만약 정말 똑같이 제조된 혼합된 시멘트 간의 강도가 다르게 나타난다면, 각기 다른 시멘트 혼합물의 강도를 비교하는 것은 어떤 의미가 있을까? 보다 정확한 문제점 분석을 위해 각각 다른 시멘트 혼합물의 **평균** (혹은 **중앙값**) 강도를 비교할 수 있다. 비슷한 예로 사례연구 1의 추정 문제는 평균 (혹은 **중앙값**) 열팽창을 언급하여 문제를 더 정확하게 명시하고 있다.

변동성으로 인해 우리에게 친숙한 단어인 **평균**과 **중앙값**은 통계학에서 전문적인 의미를 가지며 모집단과 표본의 개념을 통해 명백한 의미를 지니게 된다. 다음 절에서 **모집단**과 **표본**의 개념을 상세히 논의한다.

1.2 모집단과 표본

예제 1.1-1의 다양한 사례가 나타내는 바와 같이 통계학은 해당 연구가 특정 **모집단**(population) 혹은 모집단의 특정 구성원(개체 혹은 주제)의 특징(들)에 대한 연구를 포함하고 있을 때는 언제나 적절한 학문이다. 통계학에서 모집단은 같은 처리 혹은 방법을 사용하는 특정한 연구에 관련된 모든 개체 혹은 주제의 집합을 나타낼 때 사용된다. 모집단을 구성하는 것을 **모집단 단위** (population units)라 한다.

1 MPa(메가파스칼 단위) 단위의 압축 강도, 지름 15cm, 길이 30cm의 시험용 실린더, 물/시멘트 비율 0.4, 제조일로부터 28일 후 측정

예제 1.2-1

(a) 예제 1.1-1, 사례연구 1에서 조사하고자 하는 특성은 특정 금속의 모든 표본의 모집단에 있는 금속의 열팽창이다.

(b) 예제 1.1-1, 사례연구 3에서 시멘트 혼합물의 종류에 따라 두 가지 이상의 모집단이 있고, 조사하고자 하는 특성은 압축 강도다. 모집단 단위는 시멘트 제조다.

(c) 예제 1.1-1, 사례연구 5에서 알고자 하는 특성은 특정 작용응력에서 보의 파괴 시간이다. 이 연구의 각 작용응력은 이에 적용될 모든 보를 포함하는 분리된 모집단에 해당된다.

(d) 예제 1.1-1, 사례연구 8에서 우리는 급여 인상과 생산성, 두 특성을 가지고 있다. 대기업의 직원들로 이루어진 모집단이 각 특성에 대한 연구대상이다. ■

예제 1.2-1의 (c)는 모든 모집단이 동일한 종류의 보에 대한 실험결과이고, 각각 다르게 적용되는 작용응력에 따라 모집단이 구별된다. 이와 비슷하게 예제 1.1-1의 사례연구 2의 두 모집단은 같은 종류의 구름에 적용되는 두 가지 다른 인공강우법에 의해 구분된다.

이전 절에서 언급한 바와 같이 동일한 모집단의 구성요소에 따라 알고자 하는 특성이 달라진다. 이는 모집단의 **내재변동성**(inherent variability) 혹은 **고유변동성**(intrinsic variability)이라 불린다. 고유변동성의 결과는 전체 혹은 **모집단 수준**의 특성(들)을 이해하기 위해서는 **전수조사**(census)가 필요하다. 즉, 모집단의 모든 구성요소를 조사하는 것이다. 예를 들어 대기업 직원 모집단의 급여와 생산성의 관계를 완벽하게 이해하기 위해서는 특정 대기업 소속 모든 직원의 급여와 생산성에 관한 정보의 수집이 필요하다. 하지만 전수조사는 소요 시간과 비용 문제로 일반적으로 시행하지 않는다.

예제 1.2-2

(a) 소요 비용과 시간의 측면에서 보았을 때, 태양열 에너지 자원에 찬성하는 시민의 비율을 알기 위해 18세 이상 모든 미국 시민을 대상으로 전수조사를 하는 것은 실리적이지 않다.

(b) 소요 비용과 시간의 측면에서 보았을 때, 호수의 납 함유량을 알고자 호수의 모든 물을 대상으로 분석하는 것은 실리적이지 않다. ■

게다가 전수조사는 종종 실현 가능하지 않을 때가 있는데, 모집단의 모든 구성요소를 조사하는 건 불가능하다는 의미로, 이는 모집단이 **가설**이거나 **개념적**이기 때문이다.

예제 1.2-3

(a) 만약 연구의 목적이 제품의 품질이라면(예제 1.1-1, 사례연구 4와 7) 관련 모집단은 현재 가능한 제품뿐만 아니라 미래에 생산될 제품의 품질 또한 알아야 한다. 그러므로 관련 모집단은 가설적이다.

(b) 주간 교통사고율을 줄일 수 있는 연구에서(예제 1.1-1, 사례연구 6), 관련 모집단은 일주일 동안의 기록 기간을 살피는 것뿐만 아니라 향후 일주일의 기간까지 포함하고 있다. 그러므로 관련 모집단은 가설적이다. ■

전수조사 실시가 실리적이지 않거나 실현 가능하지 않는 연구에서는(거의 모든 경우가 그러하다), 모집단 수준의 조사 대상의 특성을 알기 위해 모집단으로부터 **표본**을 추출하여 이에 대한 해답을 얻는다. 표본추출은 수많은 모집단 단위로부터 표본을 선택하는 과정을 의미하며 그 특성(들)을 기록한다. 예를 들어 태양열 에너지 자원에 찬성하는 18세 이상 미국 시민의 비율은 시민들의 표본으로부터 확인한다. 비슷한 사례로 특정 호수의 납 함유량이 안전 수준하에 있는지 확인하기 위해서는 물의 표본을 이용해야 한다. 만약 표본이 모집단으로부터 적절히 선택된다면 우리가 관심을 갖는 특성의 표본 특성은 모집단 특성과 유사하게 나타난다.

예제 1.2-4

(a) 태양열 에너지 사용에 찬성하는 미국 시민 **표본비율**(sample proportion)(즉 선택된 표본의 비율)은 **모비율**(population proportion)에 거의 가깝다. (하지만 일반적으로는 다르다.) (표본비율과 모비율의 정확한 정의는 1.6.1절에서 볼 수 있다.)

(b) 물 표본의 평균 납 농도(**표본평균**)는 전체 호수의 평균 농도(**모평균**)와 거의 비슷하다. (하지만 일반적으로는 다르다.) (표본평균과 모평균에 대한 정확한 정의는 1.6.2절에서 볼 수 있다.)

(c) 직원 표본에서 나타나는 급여와 생산성의 관계는 대기업 소속 전체 직원 모집단의 관계와 거의 비슷하다. (하지만 일반적으로는 다르다.) ■

예제 1.2-5

비교적 측정하기 쉬운 곰의 가슴둘레는 종종 측정하기 어려운 곰의 무게를 추정하는 데 쓰인다. 그림 1-2에서 특정 숲 지역에서 서식하는 곰 50마리의 가슴둘레와 무게를 x로 표기하였다. 파란 원은 표본 크기가 10인 표본 곰 집단의 가슴둘레와 무게 측정결과를 나타내고 있다.[2] 검은 선은 모집단인 곰 50마리의 가슴둘레와 무게의 선형관계를 대략적으로 표현하였고, 파란 선은 표본 곰 집단의 선형관계를 나타낸다.[3] 표본이 나타내는 가슴둘레와 무게의 관계는 모집단은 비슷해 보이지만, 정확히 일치하지는 않는다. ■

또한 알고자 하는 표본의 특성은 각 표본별로 다르게 나타난다. 이는 모집단으로부터 어떤 표본이 추출되느냐에 따라 나타나는 내재변동성의 또다른 결과다. 예를 들어 표본 크기가 20인 태양열 에너지 확대에 찬성하는 미국 시민의 수는 다른 표본 크기 20인 미국 시민의 표본과 다르게 대응하는 수가 나타날 것이다(아마도 분명히 다를 것이다). (1.6.2절 참고) **표본 변동성**(sampling variability)은 이렇게 찾고자 하는 특성이 표본별로 차이점이 있을 때 설명하는 용어이다.

2 해당 표본은 1.3절의 단순임의추출법으로부터 산출되었다.
3 해당 선들은 6장의 최소제곱법으로부터 적합되었다.

그림 1-2

흑곰의 가슴둘레(인치)와 무게(파운드) 간의 모집단과 표본 관계

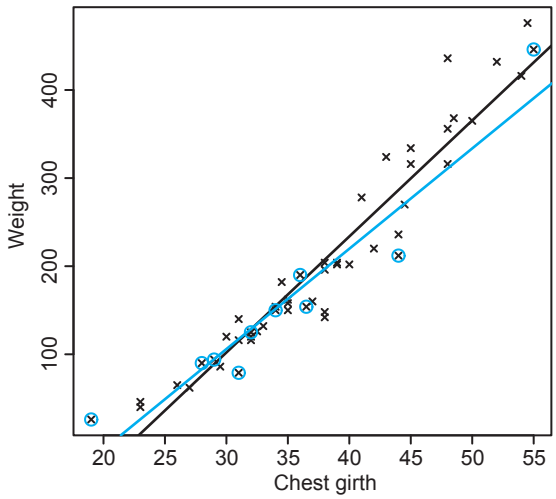
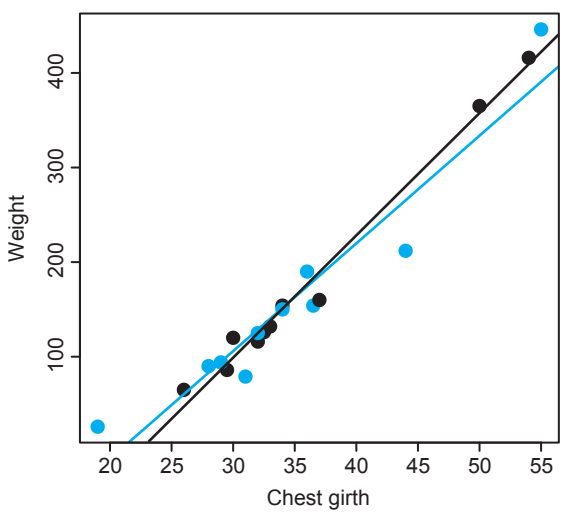


그림 1-3

표본 크기 10의 두 표본에서 흑곰의 가슴둘레와 무게 간 관계의 변동성



예제 1.2-6

표본 변동성이 나타내는 바와 같이 예제 1.2-5의 흑곰 모집단 50마리에서 표본 크기 10의 두 번째 표본이 추출되었다. 그림 1-3은 파란 점의 원표본 가슴둘레와 무게 측정을 나타내는 반면에 두 번째 표본은 검은 점으로 표현되어 있다. 표본 변동성은 다소 다른 가슴둘레와 무게 관계를 나타내는 파란 선과 검은 선으로 확인할 수 있다. 하지만 두 선 모두 모집단 관계와 비슷하다. ■

모든 과학적 탐구는 모집단 수준의 특성을 찾아내고자 함이 목표라는 사실을 결코 잊어서는 안 된다. 특히 예제 1.1-1에 언급된 모든 사례연구의 문제점은 모집단 수준의 특성과 관련 있다. 그러므로 1.1절에서 언급하여 이미 우리에게 친숙한 **평균**이라는 단어의 전문적 의미는 모평균이다. 정확한 정의를 알기 위해서는 1.6.2절을 참고하라.

모집단 수준의 특성을 나타내고자 하는 양은 **모수**(population parameter)라고 한다. 예제

1.2-4의 사례는 모평균과 모비율을 포함하고 있다. 이러한 사례와 모수와 관련된 추가 사례는 1.6절과 1.7절에 설명되어 있다. 추후에 논의할 보다 많은 사례는 급여 인상과 생산성 혹은 가슴둘레와 무게와 같은 두 특성 간 **상관 계수**를 포함한다. 대응 표본의 특성은 **통계량**(statistics)으로 불리는데 이는 **스포츠 통계량**으로 인해 이미 친숙한 단어다. 표본평균, 표본비율 그리고 추가적인 통계량은 1.6절과 1.7절에 설명되어 있고, 추후의 장에서는 더 많은 통계량이 소개된다.

표본은 모집단을 언뜻 나타내 주는 창문으로 생각할 수 있다. 하지만 표본 변동성 때문에 표본은 모집단 특성에 대한 정확한 정보를 산출하지 못한다. 이는 이전 단락에서 새롭게 소개한 용어를 이용하여 다음과 같이 말할 수 있다. 통계량은 대응하는 모수와 거의 비슷하지만 일반적으로 동일하지는 않다.

단지 표본 정보만을 활용할 수 있기 때문에 모수는 알려지지 않는다. **통계적 추론**(statistical inference)은 통계학의 한 분야로 표본에 포함된 모수의 정보를 추정하는 불확실한 문제를 다룬다. 통계적 추론은 모집단에 대한 정확한 정보가 없을 때 다음과 같은 방식으로 관리자의 의사결정을 돕는다.

- 어떤 통계량이 모수에 가장 가까운 정확성을 나타내는지 평가한다.
- 잘못된 결정 혹은 부정확한 예측의 확률에 대한 평가를 제공한다.

예를 들어 시 공무원은 새로운 산업 시설로 인한 평균 공기 오염이 규정 제한을 넘어서는지 알고 싶을 것이다. 이에 공기 표본을 추출하였고 각 표본의 공기 오염을 측정하였다. 그 후 종합(즉 모집단 수준) 평균 공기 오염이 시정 조치를 취할 정도로 상승한다면 공기 오염 측정의 표본평균을 이용하여 이를 반드시 결정해야 한다. 정확한 지식이 없는 경우에 표본평균의 공기 오염이 허용 한계를 넘는 것으로 나타났으나 실제로는 넘지 않을 때, 혹은 반대로 넘지 않는 것으로 나타났으나 실제로는 넘을 때, 이러한 경우로 인해 시 공무원들이 시정조치를 명령하는 데 있어 리스크가 있다.

통계적 추론은 주로 알고자 하는 모수의 **추정**(estimation)[**점**(point)**추정**과 **구간**(interval)**추정**]의 형태를 나타내며, 알고자 하는 모수값의 다양한 **가설**(hypotheses)을 **검증**(testing)한다. 예를 들어 추정은 금속 열팽창의 평균 계수를 추정하는 분석에 사용된다(예제 1.1-1, 사례연구 1). 반면에 품질에 관한 제조사의 주장을 검증하는 것은 가설 검증을 중요 요소로 수반한다(예제 1.1-1, 사례연구 7). 마지막으로 통계적 추론의 원칙은 **예측**(prediction) 문제에서도 사용한다. 예측 문제의 예로는 노출되는 응력에 기반을 둔 특정 보의 고장 시간을 예측하고자 하는 경우가 있다(예제 1.1-1, 사례연구 5). 이 책에서 소개하는 대부분의 통계적 방법은 통계적 추론의 범위에 있다.

연습문제

1. 자동차 제조회사에서는 전년도에 자동차를 구매한 고객의 만족도를 산출하고자 한다.
(a) 포함된 모집단을 설명하라.
(b) 모집단이 가정을 포함하고 있는가?

2. 바이오 연료를 생산하는 데 쓰이는 세 가지 종류의 곡물 수확량을 비교하기 위해 포장시험을 실시했다. 각 종류를 임의로 선정된 10개의 땅에 심었고 수확량은 수확시기에 산출될 것이다.
(a) 포함된 모집단을 설명하라.
(b) 알고자 하는 특성은 무엇인가?
(c) 표본(들)을 설명하라.

3. 자동차 생산 라인에는 1일 2교대로 근무를 선다. 첫 번째 교대는 총생산의 2/3를 차지한다. 품질관리 엔지니어는 각 2교대에서 생산량 중 대당 평균 부적합 수를 비교하고자 한다.
(a) 포함된 모집단을 설명하라.
(b) 모집단(들)이 가설을 포함하고 있는가?
(c) 알고자 하는 특성은 무엇인가?

4. 소비자 잡지의 기사 중 "비행기 내 공기는 얼마나 안전한가?"라는 제목의 보고서를 통해 부패도로 수치화되는 공기 청정도를 175개의 국내선 항공기를 대상으로 측정하였다.
(a) 모집단의 특성을 찾아라.
(b) 표본을 확인하라.
(c) 알고자 하는 특성은 무엇인가?

5. 공학도를 위한 통계학 수업 코스에서 활용하는 컴퓨터 활동의 교육적 편익을 확인하고자 하는 노력의 일환으로, 한 절(section)에서는 기존 방법으로 가르치는 반면 다른 절에서는 컴퓨터를 활용하여 가르친다. 학기 후반부에 같은 시험을 치른 학생들의 시험 결과가 기록되었다. 불필요한 변동성을 없애고자 두 절 모두 같은 교수가 가르쳤다.
(a) 연구에서 하나 혹은 두 모집단이 포함되어 있는가?
(b) 포함된 모집단을 설명하라.
(c) 모집단(들)은 가설을 포함하고 있는가?
(d) 연구에서 표본(들)은 무엇인가?

1.3 표본추출의 개념

1.3.1 대표표본

모집단의 표본 정보에 적용하는 적절한 외삽법, 즉 유효한 통계적 추론을 위해서는 표본이 모집단에 **대표성**(representative)을 띠어야 한다. 예를 들어 미국 시민 모집단에서 정유 산업에 종사하는 사람들을 포함하는 표본 정보에 외삽법을 적용한다면 불가피하게 태양 에너지 활용에 대한 퍼져 있는 여론에 대해 잘못된 판단을 내리게 될 것이다.

대표성을 띠지 않는 표본을 활용하면 어떻게 잘못될 수 있는지를 알려 주는 유명한(혹은 유명하지 않은) 사례로 1936년 「Literary Digest」의 여론 조사가 있다. 「Literary Digest」 잡지는 미국 대통령 선거의 결과를 매우 성공적으로 예측하는 곳으로 유명했다. 하지만 이 잡지는 1936년, 공화당 알프 랜던이 당시 재임 중이었던 프랭클린 루스벨트를 3대2로 이길 것으로 예측했으나 결과는 정반대였다. 이러한 어리석은 실수는 1.3.4절에서 논의할 대표성을 띠지 않는 비대표적 표본을 이용한 결과였다. 비록 230만 명(천만 개의 질문지 중)이 답변한 결과를 토대로 예

측했지만 「Literary Digest」가 틀렸다는 것은 주목할 만하다. 하지만 갤럽은 5만 명만을 대상으로 실시한 선거 결과를 토대로 정확하게 예측하였다.

대표표본의 개념은 직관적이지만 단순히 표본을 보는 것만으로는 대표표본인지 아닌지 정확하게 밝히기가 어렵다. 그러므로 우리는 간접적인 정의를 내리는데, 만약 표본이 유효한 통계적 추론을 이끌어 낸다면 표본이 대표성을 띤다고 말한다. 표본이 대표성을 띠고 있다는 유일한 확신은 표본을 추출하는 방법에서 알 수 있다. 표본추출 방법은 다음에서 논의된다.

1.3.2 단순임의추출 및 층화추출

가장 간단하게 대표표본을 추출하는 방법은 단순임의추출(simple random sampling)이다. 만약 추출 과정에서 크기 n의 모든 표본이 같은 추출 확률을 갖고 있다면, 모집단에서 추출된 크기 n의 표본은 단순임의표본이다. 특히 모집단의 모든 구성요소는 같은 확률로 표본에 포함될 가능성을 갖고 있다.

N 단위를 포함하고 있는 한정된 모집단으로부터 크기 n의 단순임의표본을 추출하는 일반적인 방법은 모집단 단위에 1부터 N까지 번호를 매기고 **난수생성기**(random number generator)를 사용하여 n개의 표본을 임의로 추출한 후 추출된 n에 일치하는 표본을 형성한다. 단순임의표본을 추출하기 위한 난수생성기는 각 종이에 1부터 N까지의 수를 쓰고, 모든 종이를 상자에 넣은 후 완벽하게 섞고 임의로 종이 한 장씩 뽑아서 기록하는 과정을 활발하게 한다. 이 과정은 1부터 N의 수로부터 n개의 특정 개수가 추출될 때까지 반복된다(뽑은 숫자는 다시 상자에 넣지 않는다).

예제 1.3-1

Sixty KitchenAid 사의 믹서는 매일 생산된다. 생산 후 배송 전에 불량품의 가능성을 확인하고자 12개의 표본을 단순임의추출한다.

(a) 하루 60개의 믹서 생산량으로부터 12개의 믹서를 단순임의추출하는 과정을 설명하라.

(b) 위 (a)의 과정을 R을 활용하여 실행하라.

해답

가장 먼저 우리는 각 믹서에 1번부터 60번까지 번호를 매긴다. 그다음 모두 같은 쪽지에 1부터 60까지의 번호를 각각 쓴 후 60개의 쪽지를 모두 박스에 넣어 완벽히 흔든다. 그 후 12개의 쪽지를 하나씩 꺼낸다. 12개의 쪽지에 쓰여 있는 번호가 매일 60개의 생산량 중 표본 크기 $n = 12$의 훌륭한 표본이 된다. 이 과정은 다음과 같은 명령어를 통해 R에서 실행할 수 있다.

단순임의추출법 : R

```
y=sample(seq(1,60), size=12)
```
(1.3.1)

y = 을 제외한 명령어, 즉 *sample(seq(1, 60), size = 12)*가 R 콘솔에 입력될 경우 12개 임의의

수가 나타날 것이다. 위에서 나타난 명령어는 y에 저장되고 그 후 'y'를 입력할 경우 12개의 임의의 수를 볼 수 있다. 따라서 추출된 12개의 임의의 수는 6, 8, 57, 53, 31, 35, 2, 4, 16, 7, 49, 41이다. ■

위에 소개한 기법은 무한(infinite) 모집단에서는 사용할 수가 없다. 하지만 잘 정의된 지침에 따라 실시한 측정은 단순임의추출이 갖는 본질적 속성을 확인해 준다. 예를 들어 시멘트 혼합물에 대한 압축 강도의 비교에서 지침은 측정에 의한 표본이 대표성을 보장할 수 있도록 시멘트 혼합의 준비와 측정과정을 확립해 준다.

이미 언급한 바와 같이 단순임의추출은 모든 모집단 단위가 같은 확률로 표본추출되는 것을 보장한다. 하지만 모집단 단위가 같은 확률로 표본추출되는 것은 표본추출 과정이 단순히 임의로 된다는 것을 보장하지는 않는다. 이는 다음 예제에서 설명한다.

예제 1.3-2

50명의 남학생과 50명의 여학생으로 구성된 100명의 학부생 집단에서 10개의 대표표본을 선택하기 위해 다음과 같은 추출방법을 시행한다. (a) 남학생 집단에 1부터 50까지의 숫자를 부여하고 난수생성기를 통해 5명을 뽑는다. (b) 같은 과정을 여학생에게 시행한다. 이 방법은 10명의 학생을 단순임의추출하는가?

해답

첫 번째 설명한 표본추출방법은 모든 학생이 표본으로 추출될 같은 확률(열 중 하나)을 갖고 있다고 보장한다. 하지만 이 방법은 남학생과 여학생의 수가 같은 표본 이외의 표본은 제외한다. 예를 들어 4명의 남학생과 6명의 여학생을 포함하는 표본은 제외되었고, 표본으로 선택될 가능성도 전혀 없다. 그러므로 단순임의추출의 조건, 다시 말해 크기 10의 표본이 선택될 수 있는 공통 확률은 무너졌다. 이는 앞서 설명한 방법으로는 단순임의표본을 구할 수 없다. ■

예제 1.3-2의 표본추출방법은 **층화표본추출**(stratified sampling)이라 불리는 방법의 예다. 층화표본추출은 분석 모집단이 **층**(strata), 즉 잘 정리된 하위 집단 혹은 하위 모집단을 포함하고 있으면 언제든지 사용될 수 있다. 층의 예로 민족, 자동차 종류, 장비의 수명, 각기 다른 연구실로 보낸 분석 대상 표본 물 등이 있다. 기본적으로 층화표본은 각 층으로부터의 단순임의표본을 포함한다. 층 내에서 표본 크기를 구하는 일반적인 방법은 각 층의 표본이 대표성을 띠며 이는 모집단의 대표성을 띠게 된다. 이 방법은 **비례 할당**(proportionate allocation) 방법으로 예제 1.3-2에서 사용되었다. 층화표본 또한 대표성의 띠는데, 이는 유효한 통계적 추론을 허용한다. 실제로 만약 같은 층에 속하는 모집단 단위는 다른 층에 속하는 모집단 단위보다 더 동질성(즉 비슷함)을 띠는 경향이 있다. 층화표본은 전체 모집단에 대해 좀 더 정확한 정보를 제공하기 때문에 더 선호하는 방법이다.

1.3.3 복원 및 비복원 표본추출

한정된 모집단으로부터 표본추출을 할 경우 **복원**(with replacement) 혹은 **비복원**(without replacement) 표본추출을 할 수 있다. 복원추출은 개체가 선택되고 그 특징을 기록한 후 모집단으로 다시 복원되어 추후에 다시 선택될 수도 있는 추출방법을 뜻한다. 동전 던지기가 모집단 복원추출의 하나의 예다 {앞면, 뒷면}. 비복원추출의 경우 각 단위는 딱 한 번만 표본에 포함될 수 있다. 그러므로 단순임의추출은 비복원추출 방법이다.

복원추출된 표본의 특성은 분석하기가 쉬운데 이는 선택된 각 단위가 N 단위의 같은(본래의) 모집단으로부터 추출되었기 때문이다(비복원추출의 경우 두 번째 추출이 $N-1$ 단위로 줄어든 모집단에서 추출되고, 세 번째 추출은 $N-2$ 단위의 모집단에서 추출된다). 반면에 모집단 단위를 한 번 이상 포함하는 것(복원추출에서 가능)은 표본의 대표성을 향상시키지 않는다. 그러므로 복원추출의 개념적 편의성(conceptual convenience)은 그에 따른 손실이 있으며, 이와 같은 이유로 보통 잘 시행하지 않는다. 하지만 모집단 크기가 표본 크기보다 훨씬 클 때 손실은 그 규모가 작아지기 때문에 무시해도 될 정도이다. 그러한 경우 우리는 단순임의추출(즉, 비복원추출)로 구한 표본이 복원추출로 구한 표본과 같은 특성을 나타낸다고 주장할 수 있다.

복원추출의 주요 활용 분야는 **부트스트랩**(bootstrap)이라는 통계적 방법에서 찾아볼 수 있다. 하지만 보통 이런 유용하고 널리 사용되는 통계적 추론 도구는 이 책과 같은 입문서에서는 다루지 않는다.

1.3.4 비대표 표본추출

비대표표본은 표본추출 계획에서 분석 대상 모집단이 표본에서 제외되었거나 체계적인 대표성이 불충분할 때 나타난다.

전형적인 비대표표본은 소위 **자체 선택** 및 **편의성** 표본으로 불린다. 자체 선택 표본의 예로, 구독자에게 답신용 엽서를 보내는 잡지를 보자. 그리고 '80%의 구독자가 디지털카메라 기능이 있는 휴대폰을 구매해 왔다'는 내용을 담은 답신 정보를 구한다. 이 경우 새로운 기능을 좋아하는 구독자가 그들의 구매활동을 나타내기 위한 답장을 했을 가능성이 높다. 그러므로 답신 엽서의 표본에는 전체 구독자 중 디지털카메라 기능이 있는 휴대폰을 구매한 사람의 비율이 높을 가능성이 훨씬 높다. 편의성 표본의 예로 당신이 다니는 대학교 학생들의 표본을 통계학 수업을 듣는 학생들로 구했을 때, 이 표본추출 계획은 통계학 수업을 듣지 않아도 되는 전공 학생들을 제외하게 된다. 게다가 통계학 수업을 듣는 대부분의 학생은 2학년 혹은 3학년 학생이기 때문에 1학년과 4학년 학생을 대표한다고 할 수 없다.

아마 표본추출 실수의 가장 유명한 역사적 사례는 1936년 「Literary Digest」 잡지의 선거 전 여론조사일 것이다. 「Literary Digest」는 여론조사를 위해 잡지구독자, 자동차 소유주 그리고 전화번호부로부터 천만 명의 표본을 추출했다. 1936년에 전화기 혹은 자동차를 소유하거나 잡지를 구독하는 사람들은 보통 민주당을 좋아하지 않는 부유한 사람들이 대부분이었다. 그렇기 때문에 이 여론조사의 표본은 모집단을 제대로 반영하지 못한 편의성 표본이었다. 게다가 답신

용 엽서를 보낸 1,000만 명 중 오직 230만 명만이 답장을 보냈다. 이는 명백하게 선거에 대해 강한 생각을 하고 변화가 필요하다 생각한 사람들이 보냈을 것이다. 그러므로 「Literary Digest」의 표본은 자체 선택되었으며 표본의 편의성을 띠게 되었다(갤럽이 다른 시기에 다른 실수[1948년 듀이-트루먼 선거]를 했음에도 살아남았지만 「Literary Digest」는 파산했다).

선택 편향이라는 단어는 체계적인 제외 혹은 분석 대상 모집단 일부 부분의 비대표성을 뜻한다. 표본이 자체 선택과 편의성을 내재하고 있는 선택 편향은 비대표표본의 전형적인 원인이다. 단순임의추출법과 층화추출법은 선택 편향을 방지한다. 선택 편향을 피하기 위한 다른 표본추출방법도 물론 존재하고, 또 특정 상황에서 그 방법들은 비용이 적거나 시행하기 쉬울 수도 있다. 하지만 이 책에서 우리는 주로 표본을 단순임의표본으로 가정하고 층화추출법은 가끔 언급할 것이다.

연습문제

1. 1.2절의 연습문제 5번의 연구를 설계하는 연구자는 컴퓨터 활용의 교육적 효과를 확인하기 위해 둘 중 하나의 선택을 할 수 있다. (i) 학생들이 두 수업 중 어떤 수업이 컴퓨터를 활용하여 가르치는지 알게 하여 학생들이 이에 근거한 선택을 하게 한다. 또는 (ii) 어떤 교육방법이 이루어지는지 학생들이 전혀 모르게 한다. 어떤 선택이 단순임의추출법에 더 가까운 결과를 나타내는가?

2. 홈시어터 시스템의 범용 리모컨은 3군데 특정 지역에서 제조된다. 전체 생산량의 20%가 *A* 공장에서 제조되고, 50%는 *B* 공장, 나머지 30%는 *C* 공장에서 제조된다. 품질관리팀(QCT)은 100개의 표본을 단순임의추출하여 최근 보고된 메뉴 기능 문제가 해결되었는지 살펴보려고 한다. QCT는 *A* 공장에서 20개, *B* 공장에서 50개, *C* 공장에서 30개의 단순임의추출된 표본 리모컨을 QC 점검 시설로 보낼 것을 주문한다.

(a) 최근 생산된 리모컨 중 크기 100의 리모컨 표본은 단순임의추출된 것인가?

(b) (a)에서 답변한 내용을 설명하라. 만약 단순임의추출된 것이 아니라면, 위의 표본추출방법은 어떤 방법으로 시행된 것인가?

3. 현재 논문 작업을 하고 있는 한 토목공학 학생은 대학가의 운전자들이 안전벨트를 정기적으로 착용하는 비율을 알고자 설문조사를 계획하고 있다. 그는 현재 세 과목을 수강 중인데, 해당 과목을 수강하는 학생들을 인터뷰하기로 결정했다.

(a) 대상 모집단은 무엇인가?

(b) 저 학생과 같은 수업을 듣는 학생들은 모집단으로부터 단순임의추출된 표본인가?

(c) 이와 같이 선택된 학생들로 이루어진 표본을 무엇으로 불러야 하는가?

(d) 이 표본비율이 실제 정기적으로 안전벨트를 착용하는 모든 운전자를 나타내는 모집단을 과대추정(overestimate) 또는 과소추정(underestimate)하는 것으로 생각되는가?

4. 2007년 1월 9일에 열린 맥월드 콘퍼런스 및 엑스포에서 스티브 잡스는 새로운 제품인 아이폰을 공개했다. 소비자 잡지의 기술 컨설턴트는 아이폰 70개 중 15개 기기를 수집해 기능 검사를 하고자 한다. 아이폰 70개 중 15개를 단순임의추출하는 방법을 설명하라. R을 이용하여 15개의 표본을 추출하라. 분석에 사용된 R 명령어와 수집한 표본을 제시하라.

5. 유통업자는 방금 주요 파이프 제조회사로부터 90개의 파이프를 수송받았다. 유통업자는 불량품 검사를 위해 5개의 표본을 추출하여 검사하고자 한다. 90개의 파이프

중 5개를 단순임의추출하는 방법을 설명하라. 분석에 사용된 R 명령어와 수집한 표본을 제시하라.

6. 어떤 서비스 기관에서는 작년 한 해 동안 고객들이 서비스 품질을 어떻게 생각했는지 평가하고자 한다. 컴퓨터를 이용해 지난 12개월 동안 1,000명의 고객 기록을 확인하였고, 그중 100명의 표본을 선택하여 설문조사를 하기로 결정했다.

(a) 작년 1,000명의 고객 모집단에서 100명을 단순임의추출하는 과정을 설명하라.

(b) 1,000명의 고객 모집단 중 800명은 백인이고, 150명은 흑인, 50명은 라틴 아메리카 계열의 미국인이다. 모집단에서 대표성을 띠는 표본 100명을 추출하기 위한 다른 방법을 설명하라.

(c) (a)와 (b)의 표본추출 과정을 시행하기 위한 R 명령어를 제시하라.

7. 자동차 제조업체는 지난해 판매한 자동차에 대한 고객 만족도를 조사하고자 한다. 제조업체는 3개의 차종을 제조한다. 단순임의표본추출을 위한 두 가지 방법을 제시하고 설명하라.

8. 어떤 제품은 A와 B, 두 제조설비에서 생산된다. B 설비는 더 최신 설비로 전체 생산량의 70%를 생산한다. 품질관리 기술자는 1시간 동안 생산된 제품 중 50개의 표본을 단순임의추출을 통해 수집하고자 한다. 동전 던지기를 하여 앞면이 나올 때 기술자는 A 설비의 제품을 임의로 추출하고, 뒷면이 나올 때 기술자는 B 설비에서 생산된 제품을 임의추출한다. 이러한 방법은 단순임의추출이라 할 수 있는가? 정답을 제시하고 설명하라.

9. 자동차 조립라인은 하루에 2교대를 실시한다. 첫 번째 교대 근무자들은 전체 생산량의 2/3를 생산한다. 품질관리 기술자는 자동차의 부적합사항의 수를 모니터한다. 첫 번째 교대 근무자들이 생산한 자동차에서는 6대의 자동차를, 두 번째에서는 3대의 자동차를 단순임의추출하고 부적합사항의 수를 기록한다. 이러한 방법을 통해 추출된 9대의 자동차는 하루 자동차 생산량으로부터 단순임의추출된 것인가? 정답을 제시하고 설명하라.

1.4 확률변수와 통계적 모집단

1.1절에서 소개한 모든 연구사례에서 알고자 하는 특성은 측정이 가능하며 수로 표현이 될 수 있다는 점에서 **양적**(quantitative)이다. 비록 양적 특성이 더 일반적이지만, 질적(qualitative) 특성을 포함하여 **범주형**(categorical) 특성을 지니기도 한다. 질적 특성의 두 예로 성별과 자동차 종류가 있는 반면에 의견에 대한 강도(strength)는 순서형(ordinal)이다. 통계적 과정은 수치형 데이터 집합에 적용되기 때문에 숫자는 범주형 특징을 나타내게 되어 있다. 예를 들어 −1의 경우 연구 대상이 남성이고, +1은 여성을 의미하도록 쓰일 수 있다.

모든 종류의 특성이 수로 표현되는 것을 **변수**(variable)라 한다. 범주형 변수는 **이산형**(discrete) 변수의 일종이다. 양적 변수 또한 이산형일 수 있다. 예를 들어 어떠한 사업상 제의에 찬성하는 수와 같이 개수를 나타내는 모든 변수는 이산형이다. 길이, 강도, 무게, 고장 시간 등 연속형 척도에서의 측정치를 나타내는 양적 변수는 **연속형**(continuous) 변수의 예다. 마지막으로 변수는 각 모집단 단위에 대해 측정하거나 기록하는 특성이 하나냐, 둘이냐, 또는 그 이상이냐에 따라 **단변량**(univariate), **이변량**(bivariate) 혹은 **다변량**(multivariate) 변수로 구분한다.

예제 1.4-1

(a) 급여 인상과 생산성의 연관성을 확인하고자 하는 연구에서 각 모집단 단위(생산성과 급여 인상)의 두 특징을 기록하는 경우, 이변량 변수가 된다.

(b) 18세 이상 미국 시민을 대상으로 태양열 에너지에 관한 입장을 조사하는 설문조사가 있다. 만약 연구에서 추가적으로 연령별 그룹에 따른 입장 변화를 확인할 목적으로 표본의 각 설문 당사자 나이까지 함께 기록하면 이변량 변수가 된다. 그리고 추가적으로 성별에 따른 입장 변화를 함께 파악하기 위해 표본의 각 설문 당사자 성별까지 함께 기록하면, 다변량 변수가 된다.

(c) 호수에 안전 수치 이상의 납 함유 여부를 알기 위해 물 표본의 납 농도를 측정하는 연구가 있다. 만약 다른 오염 물질도 함께 파악하고자 한다면, 각 물 표본의 다른 오염 물질 농도도 함께 측정하게 되며, 이는 다변량 변수가 된다. ■

고유 변동성으로 인해 모집단 개체 속에서 변수의 값은 변화한다. 하나의 모집단 개체가 모집단으로부터 임의추출된다고 가정하면 그 개체의 값은 사전에 알 수 없다. 임의로 추출된 모집단 단위의 변숫값은 X와 같이 대문자로 나타낸다. X가 사전에 알려지지 않은 사실은 X가 **확률변수**(random variable)라는 타당성을 보여 준다.

> 확률변수 X는 표본추출되는 모집단 단위의 변수의 값을 나타낸다.

확률변수가 추출된 모집단은 확률변수의 **기본 모집단**(underlying population)이라 불린다. 이러한 용어는 모든 연구가 두 가지 혹은 그 이상의 방법이나 제품의 성능을 비교할 때와 같이 여러 모집단을 분석하는 연구에서 특히 유용하다. 예제 1.1-1의 사례연구 3을 예제로 참고하라.

마지막으로 모집단의 단위를 분석할 변수의 전체 수집한 값을 부를 용어가 필요하다. 언급한 바와는 다르게 모집단의 각 단위가 분석 중인 변수의 값으로 표시되어 있다 하고 표시된 모든 값을 수집한다. 수집한 값은 **통계적 모집단**(statistical population)이라 부른다. 만약 두 모집단(혹은 그 이상) 단위가 같은 변수의 값을 갖고 있다면, 이 값은 통계적 모집단에서 두 번(혹은 그 이상) 나타난다.

예제 1.4-2

18세 이상 미국 시민을 대상으로 태양열 에너지에 관한 입장을 조사하는 설문조사가 있다. 시민의 의견은 0부터 10까지의 범위로 매기고, 각 모집단의 구성원이 그들 의견의 값에 따라 표시된다고 하자. 통계적 모집단은 0에 해당하는 사람이 많을수록 많은 0을 포함하고, 1에 해당하는 사람이 많을수록 많은 1을 포함하는 식으로 나아간다. ■

'모집단'이라는 단어는 모집단 단위 혹은 통계적 모집단을 지칭할 때 쓰인다. 내용 또는 설명은 어떤 사례인지 명확하게 할 것이다.

위의 논의를 통해 확률변수는 (통계적) 모집단에서 임의로 추출한 수치형 결과로 소개되었다. 보다 일반적으로 확률변수의 개념은 임의의 수치형 결과를 생성하는 모든 행위 혹은 과정의 결과에 적용된다. 예를 들어 단순임의표본의 산술평균을 구하는 과정은 임의의 수치형 결과, 즉 확률변수를 생성한다(자세한 사항은 1.6절을 참고하라).

연습문제

1. 양철판 500개의 모집단에서 스크래치가 0개, 1개, 2개가 있는 철판의 수는 각각 $N_0 = 190$, $N_1 = 160$, $N_2 = 150$이다.

(a) 분석 대상 변수와 통계적 모집단은 무엇인가?

(b) 분석 대상 변수는 질적 변수인가, 양적 변수인가?

(c) 분석 대상 변수는 일변량인가, 이변량인가, 다변량인가?

2. 모집단과 함께 각 모집단 단위의 변수/특징을 측정한 다음 예제를 보라.

(a) 현재 펜실베이니아 주립대학교에 등록한 모든 학부생. 변수 : 전공

(b) 모든 학교 내 식당. 변수 : 좌석 수

(c) 펜실베이니아 주립대학교 도서관의 모든 책. 변수 : 대출 빈도

(d) 주어진 달에 생산한 모든 강재 실린더. 변수 : 지름

상기 각 예제의 통계적 모집단을 설명하고 분석 대상 변수가 질적인지 양적인지 설명하라. 그리고 모집단 변수에 측정된 또 다른 변수를 찾아라.

3. 오스트리아 그라츠에 위치한 BMW 자동차 최종 조립라인에 독일과 프랑스로부터 각각 자동차 엔진과 변속기가 수송된다. 품질관리자는 N개의 조사 대상 자동차로부터 n개의 단순임의표본을 추출하여 각 n개 자동차의 엔진과 변속기의 총부적합사항 수를 기록하고자 한다.

(a) 분석 대상 변수는 일변량인가, 이변량인가, 다변량인가?

(b) 분석 대상 변수는 질적인가, 양적인가?

(c) 통계적 모집단을 설명하라.

(d) 엔진과 변속기의 부적합사항 수가 따로 기록된다고 할 때 새로운 변수는 일변량인가, 이변량인가, 다변량인가?

4. 1.2절의 연습문제 4번에서는 한 소비자 잡지의 기사가 국내선 항공기 175대를 대상으로 부패도로 수치화되는 공기 청정도를 조사하였다.

(a) 분석 대상 변수와 통계적 모집단을 설명하라.

(b) 분석 대상 변수는 질적인가, 양적인가?

(c) 분석 대상 변수는 일변량인가, 다변량인가?

5. 3개 차종의 자동차를 제조하는 한 자동차 제조업체는 지난해 판매한 자동차에 대한 고객 만족도를 조사하고자 한다. 각 고객에게 작년 구매한 자동차 차종을 물어본 후 만족도를 1부터 6까지의 범위로 점수를 매긴다.

(a) 기록된 변수와 통계적 모집단을 설명하라.

(b) 분석 대상 변수는 이변량인가?

(c) 분석 대상 변수는 양적인가, 범주형인가?

1.5 데이터 시각화를 위한 기본적인 그래픽스

이 절에서는 데이터 표현과 시각화에 쓰이는 가장 일반적인 그래픽스를 나타낸다. 추가적인 그래픽스는 책이 진행됨에 따라 소개된다.