

# 제 1 장

## 데이터 소개

과학자들은 엄격한 방법과 주의 깊은 관찰을 통해서 질문에 답을 탐구한다. 이러한 관찰 – 연구수첩, 설문조사, 실험을 통해 수집 – 결과가 통계 조사에서 중추적인 역할을 구성하고 **데이터 (data)**라고 부른다. 통계는 자료를 수집하고, 분석하고, 데이터로부터 결론을 도출하는 최상의 방법을 연구하는 학문이다. 일반적인 조사 과정 맥락에서 통계를 두는 것이 도움이 된다:

1. 질문 혹은 문제를 식별한다.
2. 주제와 관련된 데이터를 수집한다.
3. 데이터를 분석한다.
4. 결론을 구성한다.

학문으로 통계는 2-4개 단계를 객관적이고, 엄격하며, 효율적으로 만드는데 집중한다. 즉, 통계는 주요 구성요소가 세 가지다: 데이터를 얼마나 잘 수집하는가? 데이터를 어떻게 분석하는가? 그리고, 분석으로부터 무엇을 추론할 것인가?

과학자가 조사하는 주제는 과학자가 끌임없이 질문하는 질문만큼 다양하다. 하지만, 이러한 조사의 상당부분은 일부 자료수집기법, 분석 도구, 통계적 추론의 근본개념으로 해결될 수 있다. 이번 장에서 간략히 들여다보고 다른 주제는 이 책 나머지 부분에서 다뤄질 예정이다. 각각에 대한 기본 원칙을 소개하고 기본 도구 일부를 학습해 나갈 것이다. 다른 분야 응용사례를 마주할 것인데, 일부는 과학과 꼭 연관되어 있지는 않지만, 그럼에도 불구하고, 통계조사를 통해 혜택을 받을 수 있다.

### 1 사례 연구: 뇌졸중을 방지에 사용되는 스텐트(Stent)

1 절에 통계학의 전통적 도전과제가 소개되어 있다: 치료효능 평가. 이번 절에서 사용되는 용어는 본 교과서 후반부에서는 모두 수정될 것이다. 현시점에서 계획은 통계가 실무에서 수행하는 역할에 대해 느낌만 잡기 바란다.

이번 절에서는 뇌졸증 위험 환자를 치료하는데 사용되는 스텐트의 유효성을 연구하는 실험을 고려한다.<sup>1</sup> 스텐트(stent)는 혈관내부에 장착하는 장치로 심장 쇼크 이후 환자회복을 돋고, 추가적인 심장마비 혹은 경색 위험을 줄여준다. 많은 의사들이 뇌졸증 위험이 있는 환자에게도 비슷한 효능이 있을 것이라는 희망을 가졌다. 연구자가 답하고자 하는 원칙적인 질문을 적으면서 시작해 보자:

스텐트를 사용하면 뇌졸증 위험이 줄어들까?

상기 질문을 제기한 연구자가 위험상태에 있는 451명 환자 데이터를 수집했다. 각자 스스로 자원한 환자는 무작위로 두 그룹중에 하나에 배정됐다:

**치료 집단(Treatment group).** 치료 집단에 환자는 스텐트와 의료관리를 받았다. 의료관리에는 처방, 위험 인자 관리, 생활습관 교정이 포함됐다.

**대조 집단(Control group).** 대조집단에 환자는 치료 집단과 동일한 의료관리를 받았지만, 스텐트는 받지 못했다.

무작위로 연구원이 224명은 치료집단에, 227명은 대조집단에 배정했다. 본 연구에서, 대조집단은 치료집단에 스텐트 의료치료 효과를 측정할 수 있는 참조값을 제시하게 된다.

연구원이 두 시점에서 스텐트 효과를 조사했다: 실험 참가 후 30일 후와 365일 후. 환자 5명 결과가 테이블 1.1에 요약되어 있다. 환자 결과치는 “뇌졸증(stroke)”과 “증상없음(no event)”으로 기록되는데, 환자가 각 측정시점 말미에 뇌졸증 존재유무를 나타낸다.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
:	:	:	
450	control	no event	no event
451	control	no event	no event

표 1.1: 스텐트 연구에서 환자 5명에 대한 결과.

최초 연구 질문에 답을 찾는데 있어, 개별 환자별로 데이터를 살펴보는 것은 길고도 성가신 길을 걷게 된다. 대신에, 통계 자료분석을 수행하면 모든 데이터를 한번에 고려할 수 있다. 표 1.2에 원데이터를 좀더 도움이 되는 방식으로 요약했다. 해당 표를 통해서, 전체 조사과정에서 어떤 일이 발생했는지 빠르게 볼 수 있다. 예를 들어, 30일 이내 뇌졸증 발생 환자수를 치료집단에서 식별하기 위해서, 표 왼쪽에서 치료와 뇌졸증이 교차점을 살펴본다: 33.

<sup>1</sup>Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. <http://www.nejm.org/doi/full/10.1056/NEJMoa1105335>. 본 조사연구를 보도하는 NY 타임즈 기사: <http://www.nytimes.com/2011/09/08/health/research/08stent.html>.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

표 1.2: 스텐트 조사연구를 위한 기술 통계량.

④ **Guided Practice 1.1** 치료집단 환자 224명 중에서, 환자 45명에서 첫해년도 말에 뇌졸증이 발생했다. 두 숫자를 사용해서, 첫해년도 말에 치료집단에서 뇌졸증이 발생한 환자비율을 계산하시오. (유의사항: 모든 연습문제 해답은 각주에 나와 있다.)<sup>2</sup>

표에서 요약통계를 계산할 수 있다. **요약통계량(summary statistic)**는 대량의 데이터를 요약하는 단일 숫자다.<sup>3</sup> 예를 들어, 1년 후 주된 조사결과는 요약통계량 두개로 기술될 수 있다: 치료집단과 대조집단에서 뇌졸증 발병 비율.

치료(스텐트) 집단에서 뇌졸증 발병비율:  $45/224 = 0.20 = 20\%$ .

대조집단에서 뇌졸증 발병비율:  $28/227 = 0.12 = 12\%$ .

상기 두 요약통계량은 집단간 차이를 살펴보는데 유용하고, 놀라운 점이 있다: 뇌졸증이 치료집단에서 8% 환자가 더 많다! 두가지 사유로 이것이 중요하다. 첫째, 의사가 예상한 것과 반대로, 처음에 스텐트 사용이 뇌졸증 발병율을 감소시킬 것으로 예상했다. 둘째로, 통계적 질문이다: 집단간 “실제” 차이를 데이터가 나타내고 있을까?

두번째 질문은 미묘하다. 동전을 100번 던진다고 가정하자. 동전을 던질 때 앞면이 나올 확률은 50%지만, 아마도 정확하게 50번 앞면이 나오는 것을 목격하지는 못할 것이다. 이런 유형의 변동성은 거의 모든 데이터 생성 과정(data generating process)의 일부다. 스텐트 조사에서 8% 차이가 이러한 자연 변이 때문일 수도 있다. 하지만, 더 큰 차이를 관측하면 할수록 (특정 표본크기에 대해서), 차이가 우연 때문이라고 믿기는 힘들어 진다. 그래서, 정말 질문하는 것은 다음과 같다: 우연히 발생했다는 관념을 거부할만큼 차이가 현격히 큰가요?

상기 질문을 전반적으로 다룰 수 있는 통계적 도구를 아직 갖추지 못했지만, 출간된 분석 결론을 통해 이해할 수는 있다: 뇌졸증 환자 조사에서 스텐트 위험성에 대한 강력한 증거가 있다.

**주의사항:** 상기 조사결과를 모든 환자와 모든 스텐트에 일반화하지 마라. 상기 조사는 매우 특수한 성격을 갖는 환자만 살펴봤다. 조사에 참가한 환자는 자발적으로 참여했으며 모든 뇌졸증 환자에 대표성을 갖지는 못한다. 게다가, 스텐트는 많은 유형이 존재하며, 이번 조사에서 자가확장기능이 있는 윙스팬 스텐트(Boston Scientific)만 참조했다. 하지만, 이번 조사를 통해서 중요한 교훈을 얻었다: 놀라움에 눈을 바짝 떠야만 된다.

<sup>2</sup>환자 224명 중에서 365일 이내 뇌졸증이 발생한 환자 비율은  $45/224 = 0.20$  이다.

<sup>3</sup> 공식적으로 표현하면, 요약통계량은 데이터로부터 계산된 값(value)이다. 일부 요약통계량은 다른 요약통계량 보다 더 유용하다.

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
:	:	:	:	:	:
50	no	15,829	242	html	small

표 1.3: Four rows from the `email150` data matrix.

변수	설명
<code>spam</code>	메시지가 스팸인지 아닌지 명세한다
<code>num_char</code>	전자우편 문자 수
<code>line_breaks</code>	전자우편 줄바꿈 수 (텍스트가 너무 길어 행갈이(text wrapping) 되는 것은 제외)
<code>format</code>	전자우편이 굽게, 표, 링크 같은 특수 서식을 여부를 통해서 메시지가 HTML 서식으로 작성되었는지 표기
<code>number</code>	전자우편에 숫자가 하나도 없는지, (백만 이하) 작은 숫자, 혹은 아주 큰 숫자가 담겨있는지 표기

표 1.4: `email150` 데이터셋에 대한 변수와 변수설명.

## 2 데이터 기초

대부분 분석에서 첫번째 단계는 효과적인 자료 기술과 표현이다. 이번 절에서 데이터 구조화와 더불어 책전반에서 사용될 몇가지 기법을 소개한다.

### 2.1 관측점, 변수, 데이터 행렬

표 1.3에 1, 2, 3, ... 50 행을 갖는 데이터셋이 나타나 있다. 2012년 초반에 수집된 50개 전자우편에 관한 것이다. 관측점은 `email150` 데이터셋으로 불리고, 7 절에서 자세히 살펴볼 좀더 큰 데이터에서 표본추출된 것이다.

표에 각 행은 단일 전자우편 혹은 **사례(case)**를 표현한다.<sup>4</sup> 칼럼(열, column)은 각 전자우편에 대한 **변수(variable)**라고 불리며, 특성을 표현한다. 예를 들어, 첫번째 행은 1번 전자우편으로, 스팸(spam)이 아니고, 21,705 문자, 551 줄바꿈, HTML 형식, 적은 숫자(small)만 담겨있는 것을 표현한다.

실무에서, 데이터 중요한 면을 이해했는지 확실히 하도록 질문을 명확히 하는 것이 중요하다. 예를 들어, 각 변수가 의미하는 것과 측정단위가 무엇인지 확실히 하는 것이 항상 중요하다. 전자우편 5개 변수에 대한 기술이 표 1.4에 주어졌다.

1.3 표에 데이터가 **데이터 행렬(data matrix)**을 표현하는데 데이터를 구조화하는 일반적인 방법이다. 데이터 행렬 각 행은 단일 사례에 대응되고, 각 칼럼은 변수에 대응된다. 1 절에 소개된 뇌졸증 조사에 대한 데이터 행렬이 1.1 on page 2 표로 나와있다. 이 표에서 사례는 환자가 되고 각 환자별로 상태를 기록한 변수가 3개 나와있다.

<sup>4</sup>사례(case)는 때때로 **관측 단위(unit of observation)**라고도 불린다.

데이터 행렬은 데이터를 기록하고 저장하는 편리한 방법이다. 만약 또 다른 환자 혹은 사례가 데이터셋에 추가되면, 쉽게 부가적으로 행을 추가할 수 있다. 유사하게, 또 다른 칼럼도 신규 변수로 추가될 수 있다.

- ④ **Guided Practice 1.2** 미국 3,143개 군에 대한 정보를 요약한 county 공개데이터를 살펴보자. 이 데이터셋에는 각 군에 대한 정보가 포함되어 있다: 군명칭, 어느 주에 속해 있는지, 2000년과 2010년 인구, 인당 주정부지출, 빈곤율, 그리고, 5가지 추가 특성정보. 해당 정보가 데이터 행렬로 어떻게 구조화 될까요? 조언: 교과서 연습문제 해답은 주석을 참조한다.<sup>5</sup>

county 데이터 첫 일곱 행이 표 1.5에 나와 있다. 변수정보는 표 1.6에 요약되어 있다. 상기 데이터는 US 인구조사 웹사이트에서 수집했다.<sup>6</sup>

---

<sup>5</sup>각 군을 사례로 간주할 수 있고, 각 사례별로 11개 정보가 들어있다. 3,143 행과 11 칼럼으로 구성된 표에 데이터가 담겨있고, 각 행이 군을 표현하고, 각 칼럼이 특정 정보를 표현한다.

<sup>6</sup>[quickfacts.census.gov/qfd/index.html](http://quickfacts.census.gov/qfd/index.html)

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6,068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6,140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8,752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7,122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5,131	13.4	82.0	3.7	21070	45549	none
:	:	:	:	:	:	:	:	:	:	:	:
3142	Washakie	WY	8289	8533	8,714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6,695	7.9	77.9	6.5	28463	53853	none

표 1.5: Seven rows from the county data set.

변수	변수설명
name	군(County) 명칭
state	군이 포함된 주(State) (컬럼비아 특별구도 포함)
pop2000	2000년 인구
pop2010	2010년 인구
fed_spend	인구 1명당 연방정부 지출 민고율 비율 (%)
poverty	자기소유집을 보유한 혹은 자가 소유자와 동거(예를 들어, 자가소유 부모와 동거하는 자녀)하는 인구 비율 (%)
homeownership	다층구조 건물(예, 아파트)에 살고 있는 비율 (%)
multiunit	인구 1인당 소득
income	군별 중위 가구 소득. 여기서 가구 소득은 15세 이상 근로자 총소득과 같다.
med_income	2011년 말 기준 금연금지구역 실시 유형으로 다음 세 가지 값이 갖는다: 없음(none),
smoking_ban	부분(partial), 전면(comprehensive). 전면(comprehensive) 금지는 식당, 술집, 군로장소에서 흡연이 허락되지 않는다는 것을 뜻한다. 부분(partial) 금지는 흡연이 세곳중에서 최소 한곳에서 금지된다는 것을 의미한다.

표 1.6: county 데이터셋에 대한 변수와 변수정보.

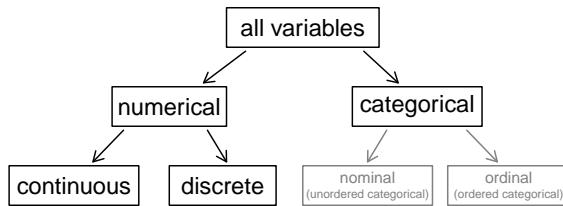


그림 1.7: 각 변수 유형별 변수 분해

## 2.2 변수 유형

county 데이터셋에 fed\_spend, pop2010, state, smoking\_ban 변수를 조사한다. 각 변수는 본질적으로 서로 다르지만 특정한 성질은 공유하다.

먼저, fed\_spend을 생각해보자. 폭넓은 숫자값을 취할 수 있고, 더하고 빼고 평균을 계산할 수 있어 **숫자형(numerical)** 변수로 불린다. 다른 한편으로, 지역 전화번호를 표현하는 변수를 숫자형으로 분류하지는 않는데 이유는 평균, 합계, 차이가 특별한 의미가 없기 때문이다.

fed\_spend와 약간 달라 보이지만, pop2010 변수도 숫자형이다. 인구수 변수는 음이 아닌 정수(0, 1, 2, ...)만 갖을 수 있다. 이러한 이유로, 인구변수는 **이산형(discrete)**이라고 불린다. 왜냐하면 점프되는 숫자값만 갖기 때문이다. 반대로, 정부지출변수는 **연속형(continuous)**이라고 불린다.

워싱턴 DC를 고려하면 state 변수는 51개 값까지만 갖게된다: AL, ..., and WY. 반응값 자체로 범주형이기 때문에, state는 **범주형(categorical)** 변수라고 불린다. 가능한 값을 변수 수준(levels)이라고 부른다.

마지막으로... smoking\_ban 변수를 생각해 보자. 이 변수는 군별로 금연 유형을 기술하는데, 없음(none), 부분금지(partial), 전면금지(comprehensive) 값을 갖는다. 하이브리드 변수로 볼 수 있다: 범주형 변수지만, 각 수준은 자연적인 순위를 갖는다. 이러한 특성을 갖는 변수를 **서수형(ordinal)** 변수라고 부른다. 분석을 단순화하기 위해서, 이 책에서 어떤 서수형 변수도 범주형 변수로 처리한다.

- **Example 1.3** 통계학 과목 수강한 학생에 관한 데이터가 수집되었다. 각 학생에 대해 변수 세개가 기록되어 있다: 형제자매 숫자, 학생 키, 이전에 통계학 과목 수강여부. 각 변수를 연속 숫자형, 이산 숫자형, 범주형으로 구분하라.

---

형제자매 숫자와 학생 키는 숫자형 변수를 나타낸다. 왜냐하면, 형제 자매 숫자는 셉(count)으로 이산형이다. 신장은 연속적으로 변해서 연속형 숫자 변수다. 마지막 변수는 학생을 두 범주로 구분한다 – 통계학 과목을 수강했던 학생과 그렇지 않은 학생 – 이 과정을 통해서 변수가 범주형이 된다.

- **Guided Practice 1.4** 1에 있는 스텝츠 조사로부터 group과 (30일) outcome 변수를

고려해보자. 이 변수는 숫자형 변수일까요, 범주형 변수일까요?<sup>7</sup>

### 2.3 변수 사이 관계

연구자들은 두개 혹은 그 이상 변수 사이 관계를 찾으려는 동기로 분석을 시작한다. 사회 과학자는 다음 질문에 대답하고 싶을 것이다:

- (1) 높은 빈곤율을 갖는 시군에서 연방정부지출은 평균적으로 더 높은가 혹은 낮은가?
- (2) 만약 자가소유가 특정 시군에서 전국평균보다 낮다면, 해당 시군에 다가구 비율이 전국 평균을 상회할까 혹은 하회할까?
- (3) 더 높은 평균소득을 올리는 시군은 어딜까: 금연정책을 시행하는 시군 혹은 그렇지 않은 시군일까?

이러한 질문에 대답하기 위해서, 표 1.5에 나타난 county 데이터셋 같은 데이터가 수집되어 한다. 요약 통계량을 조사자를 통해 시군별로 상기 3개 질문에 대한 통찰을 얻을 수 있다. 추가적으로, 그래프를 사용해서 데이터를 시각적으로 요약할 수 있고, 상기 질문에 답하는데도 유용하다.

산점도는 두 숫자형 변수 사이에 관계를 조사하는데 사용되는 일종의 그래프다. 그림 1.8은 fed\_spend 변수와 poverty 변수를 비교한다. 플롯에 각 점은 단일 시군을 나타낸다. 예를 들어, 부각된 점은 county 데이터셋에서 1088 군에 상응한다: 켄터키주 오우슬리 군(Owsley County, Kentucky)으로 빈곤율이 41.5%이고 연방정부지출은 일인당 \$21.50 달러다. 산점도를 통해서 두 변수 사이에 관계를 유추할 수 있다: 높은 빈곤율을 갖는 시군이 약간 더 높은 연방정부지출을 받는 경향이 있다. 다양한 생각을 내서 이런 관계가 존재하는 이유는 무엇이고 각 아이디어를 조사해서 어느 것이 가장 합리적인 설명인지 판단해보자.

**○ Guided Practice 1.5** 표 1.4 on page 4에 기술된 email150 데이터셋에 있는 변수를 조사하라. 본인 관심을 사로잡는 변수 사이 관계에 대해서 질문 2개를 만들어보세요.<sup>8</sup>

fed\_spend 와 poverty 변수는 연관(associated)되었다고 하는데 플롯을 통해 식별할 수 있는 패턴이 보이기 때문이다. 두 변수가 서로에 어떤 연관을 보일 때, 두 변수를 **연관(associated)** 변수라고 부른다. 연관된 변수는 또한 **종속(dependent)** 변수라고 불리기도 하고 역도 또한 같다.

**● Example 1.6** 상기 예제는 자가소유와 다가구 건축물(예를 들면, 콘도, 아파트) 비율 관계를 조사해서 그림 1.9에 산점도를 사용해서 시각화했다. 두 변수는 연관되었는가?

---

<sup>7</sup>각 변수마다 단지 두가지 가능한 값이 있고, 각 경우에 대해 범주를 기술한다. 그래서 각각은 범주형 변수가 된다.

<sup>8</sup>두 예제 질문: (1) 전자우편에 줄바꿈이 많으면 전자우편에 문자가 많은 경향이 있다고 직관적으로 볼 수 있다: 이것이 사실일까? (2) 전자우편 형식이 일반 텍스트면 스팸 메시지 혹은 HTML 이면 스팸메시지라는 연관관계가 있을까?

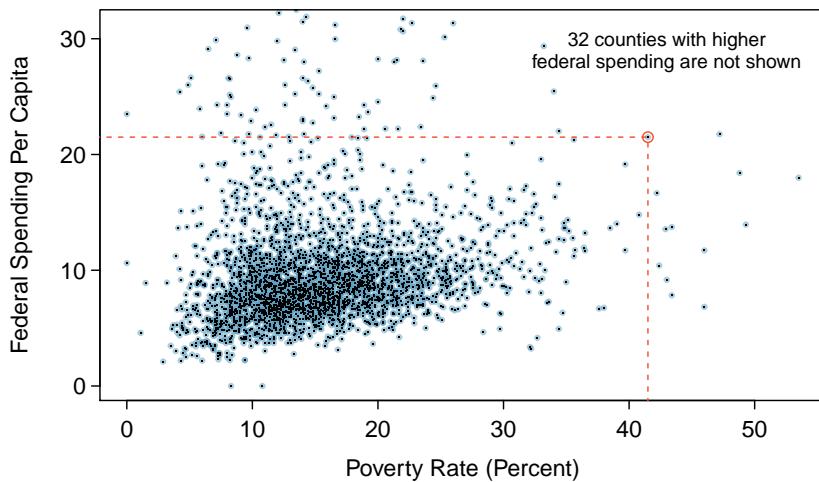


그림 1.8: poverty 빈곤율에 대한 fed\_spend 연방정부지출을 보여주는 산점도. 41.5% 빈곤율과 1인당 \$21.50 연방정부지출 정보를 갖는 켄터키주 오우슬리군이 부각되어 있다.

다가구 건축물에서 구성단위 분율이 크면 클수록 자가소유비율이 낮아지는 것으로 나타난다. 두 변수 사이에 일정 관계가 있기 때문에, 두 변수는 연관되어 있다.

그림 1.9에 우하향 경향이 보이기 때문에 – 다가구 건축물 대비 비율이 높은 시군이 더 낮은 자가소유와 연관되어 있다 – 두 변수는 음의 연관성을 갖는다고 한다. 양의 연관은 그림 1.8에 poverty 와 fed\_spend 변수 관계에 나타나 있다. 여기서 더 높은 빈곤율을 갖는 시군이 일인당 더 많은 정부지출지원을 받는 경향이 있다.

만약 두 변수가 연관되지 않는다면, 두 변수는 독립(independent)적이라고 한다. 즉, 두 변수 사이에 명백한 관계가 없다면, 두 변수는 독립이다.

### 연관되거나 독립적이거나 하지만 둘다는 안된다

변수 한쌍은 어떤 방식으로 (연관) 관련이 있거나 (독립적) 없다. 어떤 변수 쌍도 연관되고 독립적이지는 않다.

## 3 데이터 수집 원칙 개요

과학적인 연구를 수행하는 첫번째 단계는 조사할 연구주제 혹은 문제을 식별하는 것이다. 명확히 설정된 연구문제는 어떤 주제 혹은 사례가 연구되어야 하고, 어떤 변수가 중요한지 식별하는데 도움을 준다. 또한 데이터를 어떻게(*how*) 수집할지도 중요하다. 그렇게 함으로써 데이터에 신뢰성이 높아지고 연구 목적을 달성하는데 도움이 된다.

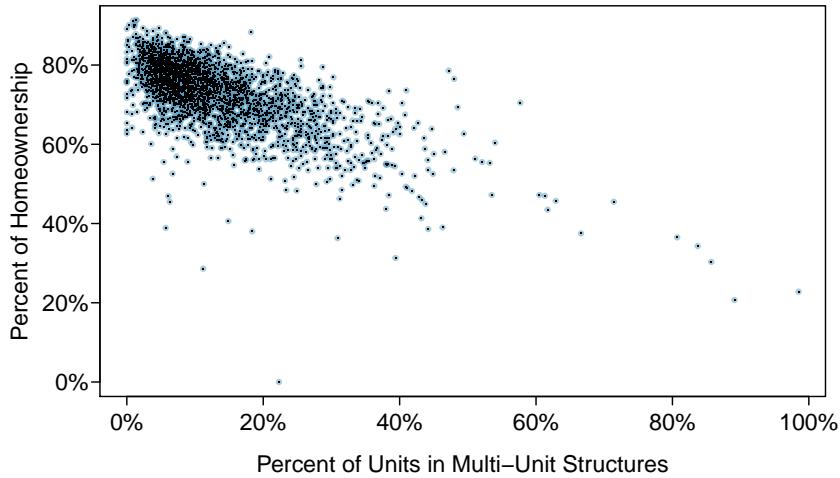


그림 1.9: 전체 3,143 시군에 대한 다가구 건축물 구성 비율과 자가소유비율 사이의 산점도. 흥미있는 독자는 추가적으로 세번째 변수(군 인구)를 추가한 플롯 이미지를 [www.openintro.org/stat/down/MHP.png](http://www.openintro.org/stat/down/MHP.png)에 확인할 수 있다.

### 3.1 모집단과 표본

다음 세가지 연구문제를 생각해보자:

1. 대서양 황새치(swordfish)에 포함된 평균 수은 함량은 얼마인가?
2. 지난 5년간 듀크대학생이 학위를 얻는데 평균적으로 걸리는 기간은 얼마인가?
3. 신약은 심각한 심장질환을 가진 환자의 사망률을 낮추는가?

각 연구문제는 목표 **모집단**과 관련되어 있다. 첫번째 문제는 목표 모집단이 대서양에 있는 모든 황새치가 되고 각 물고기는 한가지 사례를 대표한다. 종종 모집단에 있는 모든 사례에 대한 데이터를 수집하는 것은 비용이 너무 많이 듈다. 대신에, 표본을 수집한다. **표본**은 사례의 일부이며 종종 모집단의 작은 일부분이다. 예를 들어, 모집단에서 청새치 60 마리 (혹은 다른 숫자)가 선택되면 표본 데이터를 사용해서 모집단 평균을 추정하고 연구문제에 답을 구한다.

**◎ Guided Practice 1.7** 상기 두번째 세번째 문제 대해서, 목표 모집단과 개별 사례가 나타내는 것을 식별하세요.<sup>9</sup>

### 3.2 일화적 증거(Anecdotal evidence)

상기 세가지 연구문제에 대한 가능한 응답을 생각해보자:

---

<sup>9</sup>(2) 두번째 질문은 단지 학위를 마친 대학생에게만 관련됨에 주목한다; 학위를 끝마치지 못한 대학생 정보를 사용해서 평균을 계산할 수 없다. 그래서 단지 지난 5년간 졸업한 대학생만이 고려중인 모집단에 사례를 대표하게 된다. 그런 대학생만 개별 사례를 대표하게 된다. (3) 심각한 심장 질환을 가진 환자가 사례를 대표하게 된다. 모집단에는 심각한 심장 질환을 갖는 모든 환자가 포함된다.



그림 1.10: 2010년 2월 일부 미디어 전문가가 나서서 대형 눈폭풍 하나를 지구 온난화에 대한 적법한 증거로 인용했다. 코메디언 존 스튜어트(Jon Stewart)는 “특정 군에 한 지역에 폭풍 하나.”라는 점을 지적했다.

1. 청새치를 먹어서 수온에 중독된 뉴스에 한 남자가 나왔다. 그래서 청새치에 평균 수온 함량은 치명적으로 매우 높음에 틀림없다.
2. 저자가 듀크 대학을 졸업하는데 7년 이상 걸린 두 대학원생을 만났다. 그래서 다른 대학 보다 듀크 대학을 졸업하는데 더 오래 걸림이 틀림없다.
3. 내 친구 아버지가 심장마비가 생겼고, 새로운 심장질환 약을 처방받은 후에 돌아가셨다. 그래서 신약은 약효가 없음에 틀림없다.

결론 각각은 데이터에 기반하고 있다. 하지만, 두가지 문제점이 있다. 첫째, 데이터가 단지 하나 혹은 두가지 사례만 대표한다. 둘째, 그리고 좀더 중요하게 해당 사례가 모집단을 실제 대표하는지 명확하지가 않다. 이러한 무계획적인 방식으로 수집된 데이터를 **일화적 증거**라고 부른다.

#### 일화적 증거

무계획적인 방식으로 수집된 데이터에 주의하라. 그러한 증거는 사실이고 입증할 수 있지만, 흔치 않은 사례만 대표할 수 있다.

일화적 증거는 통상 흔치 않는 사례로 구성되는데, 두드러진 특성에 기반해서 회상되는 것이다. 예를 들어, 4년만에 졸업한 학생 6명보다 7년 걸려 졸업한 두명을 좀더 기억할 듯 하다. 가장 특이한 사례를 살펴보는 대신에, 모집단을 대표하는 많은 표본 사례를 조사해야 한다.

### 3.3 모집단으로부터 표본추출

대학생 표본을 수집해서 지난 5년간 듀크 대학생에 대한 졸업기간을 추정해보자. 지난 5년간 모든 졸업생이 모집단을 대표하고, 검토를 위해서 선정된 대학생을 총괄해서 표본이라고 부른다. 일반적으로 항상 모집단에서 표본을 무작위로(*randomly*) 선택하려고 한다. 가장 기본적인 임의 표본 선정 유형은 추첨 방식과 동등하다. 예를 들어, 대학생을 선택할 때 추첨표에 각 학생 이름을 적고 100장을 뽑는다. 뽑힌 이름이 무작위로 추출한 학생 100명 표본을 대표한다.

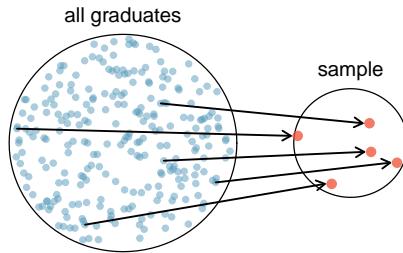


그림 1.11: 상기 도표에서, 졸업생 5명이 모집단에서 무작위로 뽑혀 표본에 포함된다.

표본을 왜 무작위로 뽑을까요? 손으로 표본을 뽑으면 안되나요? 다음 시나리오를 생각해 보자.

- **Example 1.8** 우연히 영양학을 전공하는 학생에게 연구조사를 위해서 졸업생 몇 명을 선정해 달라고 요청한다고 가정하자. 본인이 판단하기에 어떤 학생 유형을 선정할 것 같은가? 이러한 표본이 모든 졸업생을 대표할 것으로 생각합니까?

---

아마도, 건강과 연관된 전공분야에서 균형이 맞지 않는 숫자의 졸업생을 선택할 것이다. 혹은 아마도 모집단을 잘 대표하게 표본을 추출할 수도 있다. 설사 해당 편의(*biased*)가 의도적이지 않거나 식별하기 어렵더라도 수작업으로 표본을 추출할 때, 편의(*biased*)된 표본을 추출할 위험을 감수하게 된다.

만약 누군가 표본에 어떤 졸업생이 포함되어야 하는지 정확하게 선택해서 뽑을 수 있다면, 표본이 해당 사람의 관심에 치우치는 것도 완전히 가능한데, 이것은 완전히 고의가 아닐 수 있다. 이렇게 하면 표본에 편의가 들어오게 된다. 무작위 표본추출은 이러한 문제를 푸는데 도움이 된다. 가장 기본적인 임의 표본은 단순 임의 표본으로 불리며 사례를 선택하는데 추첨 표를 사용하는 것과 동등하다. 모집단 각 사례는 표본에 포함될 동일한 확률을 갖고 표본에 있는 사례간에는 어떤 암묵적인 연관관계도 없다는 것을 의미한다.

단순임의 표본을 취하는 행위가 편의를 최소화하는데 도움을 주지만, 편의는 다른 방식으로 불쑥 나타난다. 예를 들어 설문조사를 위해 사람을 임의로 뽑을 때 조차도, 무응답이 높다면 주의를 기울여야 한다. 예를 들어, 설문조사에 무작위로 뽑힌 30% 사라만 응답했다면, 결과가 전체 모집단에 대표성을 갖는지 명확하지 않다. 이러한 무응답 편의가 결과를 왜곡할 수 있다.

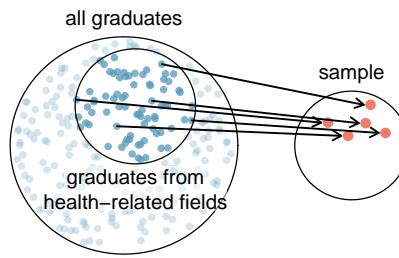


그림 1.12: 동일하게 모든 졸업생으로부터 표본을 추출하는 대신에, 영양학 전공 학생은 우연히 건강 관련 전공자를 불균형하게 표본으로 추출할 수 있다.

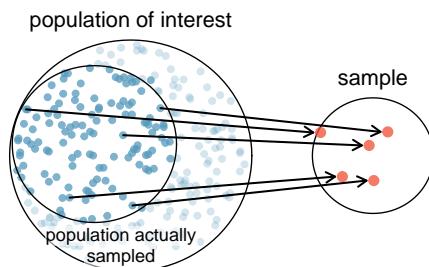


그림 1.13: 무응답 가능성 때문에, 설문조사는 모집단 내부에서 특정 집단에만 도달하게 된다. 이 문제를 완전히 고치는 것은 어렵고, 종종 불가능하다.

또 다른 흔히 빠지는 함정이 **편의 표본**으로 쉽게 접근할 수 있는 개인이 표본에 좀 더 포함될 듯 하다. 예를 들어, 만약 정치 설문조사가 브롱크스 행인을 대상으로 수행된다면, 뉴욕시 전체를 대표할 것 같지는 않다. 편의 표본이 어떤 하위-모집단을 대표하는지 분간하는 것이 종종 어렵다.

④ **Guided Practice 1.9** 웹사이트를 통해서 쉽게 제품, 판매자, 회사에 대한 평가점수에 접근할 수 있다. 이러한 평점은 평가점수를 제공하려고 노력한 사람들에 기반해서 매겨졌다. 만약 제품에 대한 50% 온라인 사용후기가 부정적이라면, 이러한 사실이 구매자 50%가 해당 제품에 불만족한다고 보십니까? <sup>10</sup>

### 3.4 설명 변수와 반응 변수

county 데이터셋에 대해 8 페이지에 나온 다음 질문을 생각해 보자.

- (1) 높은 빈곤율을 갖는 시군에서 연방정부지출은 평균적으로 더 높은가 혹은 낮은가?

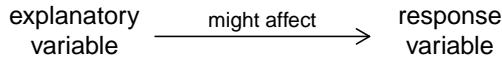
---

<sup>10</sup> 대답은 다양할 수 있다. 우리 스스로 일화적 경험으로부터, 기대한대로 동작하는 제품에 대해서 격찬하기보다는 기대에 떨어지는 제품에 대해서 못마땅하게 과장되게 사람들이 말하는 경향이 있다고 믿는다. 이러한 이유로, 아마존 같은 사이트에 올라온 제품 평점에 음의 편향이 있다고 의심할만 하다. 하지만, 본인의 경험이 대표성을 갖지 않을 수 있기 때문에, 마음을 항상 열어놓자.

만약 빈곤이 해당 군 지출에 영향을 미친다고 의심하면, 빈곤은 **설명** 변수가 되고 연방정부지출은 관계에 있어 **반응** 변수가 된다.<sup>11</sup> 만약 변수가 많다면, 이들 중 일부를 설명변수로 간주하는 것도 가능할 수 있다.

### TIP: 설명 변수와 반응 변수

변수 한쌍에서 설명 변수를 식별하기 위해서, 둘 중에 어느 것이 다른 것에 영향을 주는지 식별하고 적절한 분석을 계획하라.



### Caution: 연관성이 인과성을 함축하지는 않는다

설명변수와 반응변수로 표식을 했다는 것이 두 변수 사이 관계가 실제로 인과성이 있다는 것을 담보하지는 않는다. 설사 두 변수 사이에 연관이 확인되어도 그렇다. 어느 변수가 다른 변수에 영향을 미치는지 용의점을 추적하는데만 표식을 사용한다.

일부 설명변수와 반응변수가 없는 경우가 있다. 페이지 8에서 다음 질문을 생각해보자:

- (2) 만약 특정 군에서 자가소유가 전국 평균보다 낮다면, 다가구 건축물 비율이 전국 평균을 상회할까요 아니면 하회할까요?

어느 변수가 설명변수이며, 반응변수인지 판단하기가 어렵다. 즉, 방향성이 애매모호하다. 그래서 설명 혹은 반응 표식을 여기서는 제시할 수는 없다.

## 3.5 관측연구와 실험 소개

데이터 수집에 두 가지 유형이 있다: 관측연구와 실험.

데이터가 생성되는 방식에 직접적인 지장을 받지 않는 방식으로 데이터를 연구원이 수집할 때 **관측연구**를 수행한다. 예를 들어, 왜 특정 질병이 진행되는지 조사하기 위해서, 과학연구원은 설문조사를 경유해 정보를 수집하고, 진료 및 회사 기록을 검토하고, 유사성 있는 다수 **코호트** 환자를 추적한다. 이러한 상황에서 과학연구원은 단지 발생된 데이터를 관측한다. 일반적으로, 관측연구는 변수사이에 자연적으로 발생하는 연관 증거를 제공할 수 있지만, 그 자체로 인과관계를 제시할 수는 없다.

연구원이 인과관계 가능성은 조사하려면, **실험**을 수행한다. 대체로 설명변수와 반응변수 모두 있다. 예를 들어, 약을 투여하는 것이 다음해에 심장마비 환자의 사망율을 줄일 수 있는지 의심할 수 있다. 설명변수와 반응변수 사이에 인과관계가 정말 있는지 검사하기 위해서, 연구원은 개인 표본을 수집하고 그룹으로 나눈다. 각 그룹에 개인은 처리군(treatment)에 배정된다.

<sup>11</sup> 종종 설명 변수는 독립 변수로 불리고, 반응 변수는 종속 변수로 불린다. 하지만, 이렇게 되면 혼동스러울 수 있는데 이유는 변수 한쌍이 독립적 혹은 종속적일 수 있기 때문이다. 이러한 사유로 이런 단어를 회피한다.

개인이 무작위로 그룹에 배정될 때, 해당 실험을 **무작위 실험**이라고 부른다. 예를 들어, 의약 시험에 각 심장마비 환자를 아마도 동전던지기를 통해서 두 그룹중에 한 곳에 임의로 배정할 수 있다: 첫번째 그룹은 **위약** (placebo, 거짓처방)을 처방받고, 두번째 그룹은 약물을 처방받는다. 실험의 또다른 예제로 1에 사례연구를 참조한다. 하지만 해당 조사는 위약을 사용하지는 않는다.

**TIP: 연관 ≠ 인과**

일반적으로 연관(association)은 인과(causation)를 함축하지는 않는다. 그리고 인과는 무작위 실험으로만 추론될 수 있다.

## 4 관측연구와 표본추출 전략



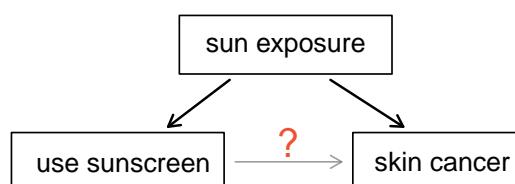
### 4.1 관측연구

일반적으로, 관측연구에 있어 데이터는 일어난 것을 모니터링하면서만 수집된다. 반면에, 연구 주요 설명변수를 연구원이 각 대상마다 배정할 것을 실험에서는 요구한다.

실험에 근거해서 인과결론을 내리는 것은 합리적이다. 하지만, 관측연구에 근거해서 동일한 인과 결론을 내리는 것은 겉보기와 달리 위험하고 추천되지 않는다. 그래서, 관측연구는 일반적으로 연관만 보이는데 충분하다.

- **Guided Practice 1.10** 관측연구가 선크림 사용과 피부암을 추적했다고 가정하자. 그리고 선크림을 더 많이 사용하면 할수록, 사람들이 더 피부암에 걸릴 것 같다는 것이 밝혀졌다. 이것이 선크림이 피부암에 원인이 된다는 것을 의미할까? <sup>12</sup>

일부 이전 연구를 통해 선크림을 사용하면 피부암 위험을 실제로 줄여준다고 알려졌다. 그래서 아마도 선크림 사용과 피부암 사이에 가상의 연관(hypothetical association)을 설명하는 또다른 변수가 있을 수 있다. 빠진 중요한 정보 조각이 일광노출이다. 누군가 하루종일 태양에 노출된다면, 선크림을 좀 더 사용할 것 같고 더 피부암에 걸릴 것 같다. 일광노출이 단순 조사에서 고려되지 못했다.



<sup>12</sup>아니다. 설명으로 문단과 이어진 예제를 살펴보라.

일광노출을 **교락변수**라고 부르고,<sup>13</sup> 설명변수와 반응변수 모두와 상관된 변수다. 관측 연구로부터 도출된 인과결론을 정당화하는 방법은 교락변수를 철저히 찾는 것이지만, 모든 교락변수가 측정되거나 조사될 수 있다는 보장은 없다.

동일한 방식으로, county 데이터셋은 교락변수를 갖는 관측연구다. 그리고 인과결론을 도출하는데 데이터가 쉽게 사용될 수는 없다.

**○ Guided Practice 1.11** 그림 1.9에 자가소유비율과 다가구 건축물 비율 사이에 부의 연관성이 보인다. 하지만, 두 변수 사이에 인과관계가 존재한다고 결론 내리는 것은 합리적이지 못하다. 그림 1.9에 나타난 시각적 관계를 설명하는 하나 혹은 그 이상 다른 변수를 제시하세요.<sup>14</sup>

관측연구는 두가지 형태가 있다: 전향적 연구(prospective studies)와 후향적 연구(retrospective studies). 전향적 연구는 개인을 식별하고 사건이 진행됨에 따라 정보를 수집해 나간다. 예를 들어, 의학 연구원은 수년간에 걸쳐 한 집단의 유사한 개인을 식별하고 추적하여 암위험에 있어 행동의 가능한 영향을 평가한다. 이러한 유형을 갖는 연구 한 사례가 간호사 건강연구(The Nurses' Health Study)로 1976년 시작해서 1989년까지 진행되었다.<sup>15</sup> 이 전향 연구는 등록된 간호사를 모집하고 나서 설문지를 사용해서 데이터를 수집했다. 후향적 연구는 사건이 발생한 후에 데이터를 수집한다. 즉, 연구원이 의료기록을 통해서 지난 사건을 검토할 수 있다. county 같은 일부 데이터는 전향적이며 후향적 방식으로 시집된 변수를 모두 포함하고 있다. 지방정부는 전향적으로 사건이 전개되어가면 일부 변수를 수집(예를 들어, 소매)하는 반면에, 연방정부는 후향적으로 2010년 인구총조사 기간동안 다른 변수를 수집했다(예를 들어, 군 인구).

## 4.2 네가지 표집 방법 (특별 주제)

거의 모든 통계적 방법은 암묵적 임의성(implied randomness) 개념에 기초하고 있다. 만약 관측된 데이터가 모집단으로부터 임의 구조 아래에서 수집되지 않는다면, 이런 통계적 방법—추정값과 추정값과 연관된 오차—은 신뢰성이 없다. 여기서 네가지 임의 표본추출 기법을 고려한다: 단순표집(simple), 충화표집(stratified), 군집표집(cluster), 다단계표집(multistage). 그림 1.14와 1.15에 네가지 기법을 시각적으로 표현했다.

단순임의표집은 아마도 가장 직관적인 임의추출방법이다. 메이저 리그 야구(Major League Baseball, MLB) 선수 연봉을 생각해보자. 여기서 선수 각각은 리그에 속한 30개팀 중 한팀의 일원이다. 야구선수 120명을 임의 표본으로 뽑아 2010 시즌 연봉 정보를 수집하려면, 제비쪽지에 해당 시즌 야구선수 828명 명단을 적고, 뽑기함에 제비쪽지를 넣고, 선수명단이 잘 섞일 때까지 뽑기함을 흔들고 나서 표본 야구선수 120명을 뽑을 때까지 제비뽑기를 진행한다. 일

<sup>13</sup>또한 잠복변수(lurking variable), 교락요인(confounding factor), 혹은 교락요인(confounder)이라고 부른다.

<sup>14</sup>정답은 다양하다. 인구밀도가 중요할 수도 있다. 만약 특정 군이 매우 밀도가 높다면, 주민 상당수가 다가구 건축물에 살아야만 된다. 부가적으로 높은 밀도는 부동산 가치를 높여서 다수 주민이 자가소유건물을 갖지 못하게 할 것 같다.

<sup>15</sup>[www.channing.harvard.edu/nhs](http://www.channing.harvard.edu/nhs)

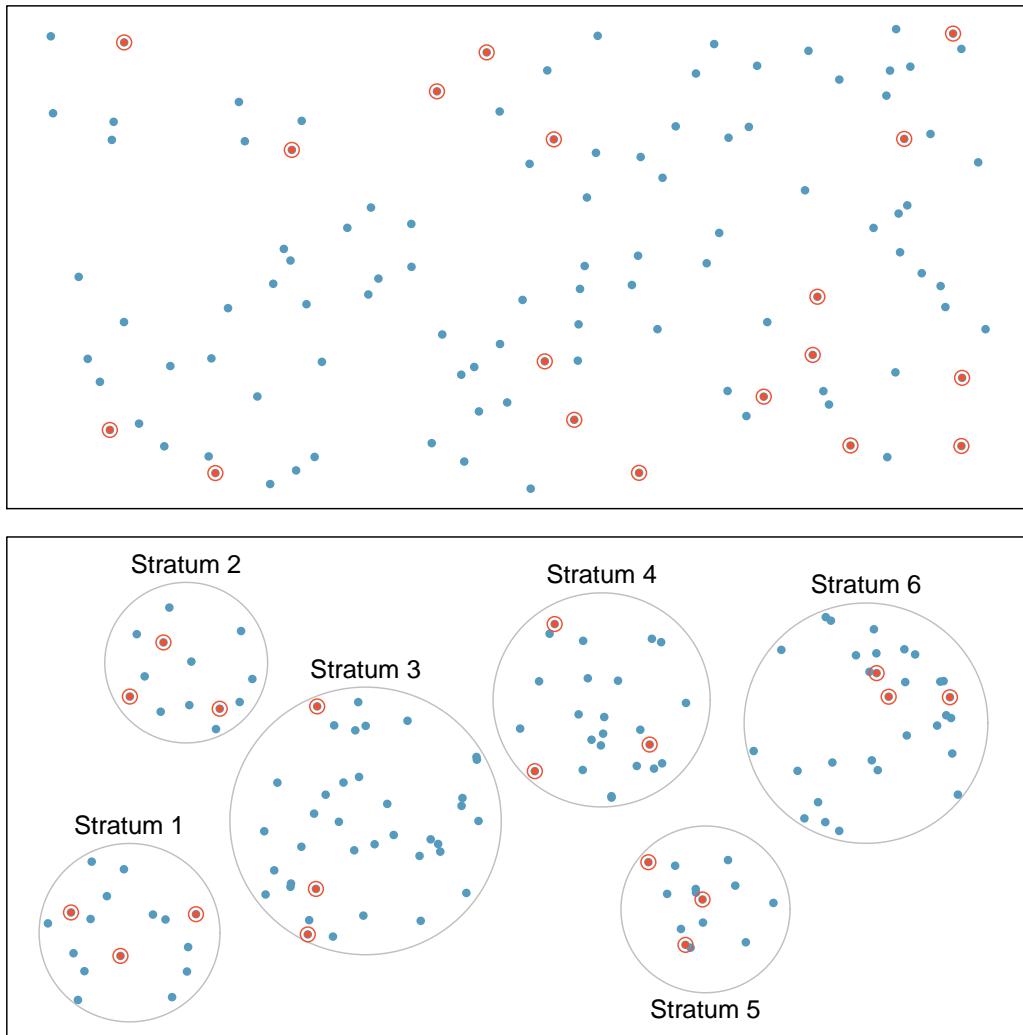


그림 1.14: 단순임의표집과 층화표집 예제. 상단 그림에서는 단순임의표집을 사용해서 18개 사례를 임의로 뽑았다. 하단 그림에서는 층화표집이 사용되었다: 사례를 층(strata)으로 그룹으로 만들고 나서 각 층마다 내부적으로 단순임의표집을 사용했다.

반적으로, 만약 모집단에서 사례 각각이 최종 표본에 포함될 동일한 확률을 갖고 특정 사례가 표본에 포함된다는 것을 아는 것이 어떤 다른 사례가 포함될지에 관한 유용한 정보를 제공하지 않는다면, 해당 표본을 “단순임의표본(simple random sample)”이라고 부른다.

**총화표집**은 분할정복(divide and conquer) 표집 전략이다. 모집단을 총으로 불리는 집단으로 분할한다. 총은 유사한 사례를 함께 무리짓고 나서, 두번째 표집 방법, 일반적으로 단순임의표집을 각 총에 사용한다. 프로야구 연봉 예제에서, 팀이 총을 대표할 수 있다. 왜냐하면 일부 팀이 돈을 많이 갖고 있기 때문이다(최대 4배). 그러면, 무작위로 각 팀마다 선수 4명을 임의로 추출하여 전체 120명을 뽑을 수 있다.

총화표집이 특히 유용할 때는 각 총에 사례가 관심있는 결과에 비추어 매우 유사할 때다. 단점은 총화표본 데이터를 분석하는 것이 단순임의표본에서 나온 데이터를 분석하는 것보다 더 복잡한 작업이라는 것이다. 총화표집을 사용해서 수집된 데이터를 분석하기 위해서는 이 책에서 소개된 분석 방법을 확장될 필요가 있다.

### ● Example 1.12 각 총 내부에 사례를 매우 유사하게 만드는 것이 왜 좋을까?

---

만약 사례가 매우 유사하다면, 총내 하위모집단에 대해서 좀더 안정된 추정값을 얻을 수 있다. 각 하위모집단에 대한 이러한 향상된 추정값을 통해 전체 모집단에 대한 신뢰성 있는 추정값을 만들어 낼 수 있다.

**군집표본**에서, 모집단을 군집으로 불리는 다수 집단으로 쪼갠다. 그리고 나서, 정행진 군집수를 표집하고 표본에는 각 군집으로부터 모든 관측점을 포함한다. **다단계 표본**은 군집표본과 같지만, 각 군집에 모든 관측점을 선택하기 보다는 선택된 각 군집내부에서 임의 표본을 추출한다.

때때로, 군집표집과 다단계표집이 대안 표집기법보다 더 경제적일 수 있다. 또한, 총화표집과 달리, 군집 내부에 사례마다 변동성이 상당하지만, 근접 자체는 다른 것과 그다지 차이가 없을 때 이러한 접근법이 가장 도움이 된다. 예를 들어, 만약 이웃이 군집을 대표한다면, 군집 혹은 다단계 표집은 이웃이 매우 다양할 때 가장 잘 동작한다. 설사 이책에 방법을 확장할 수는 있지만, 이러한 방법의 단점은 일반적으로 좀더 고급 분석기법이 요구된다는 점이다.

### ● Example 1.13 인도네시아 농촌 열대 밀림지역에 말라리아 발병율을 추정하는데 관심이 있다고 가정하자. 해당 인도네시아 정글 지역에 마을이 30개가 있는데, 각 마을을 옆 마을과 유사하다는 것을 알게되었다. 목적은 말라리아에 대해 150 명을 테스트하는 것이다. 어떠한 표집방법이 적용되어야 하는가?

---

단순임의표본은 모든 30개 마을에서 개인 표본을 뽑을 듯 한데 데이터 수집에 비용이 엄청 많이 들 수 있다. 총화표집은 도전과제가 될 수 있는데 유사한 개인을 총화하는 방법이 명확하지 않다. 하지만, 군집표집 혹은 다단계표집은 매우 좋은 아이디어처럼 보인다. 만약 다단계표집을 사용하기로 결정한다면, 무작위로 마을 반을 선택하고 나서 각 마을마다 10명을 무작으로 선택한다. 이렇게 하면 아마도 단순임의표본과 비교해서 데이터 수집 비용이 상당히 줄 수 있으면서도 여전히 신뢰성 있는 정보를 줄 것이다.

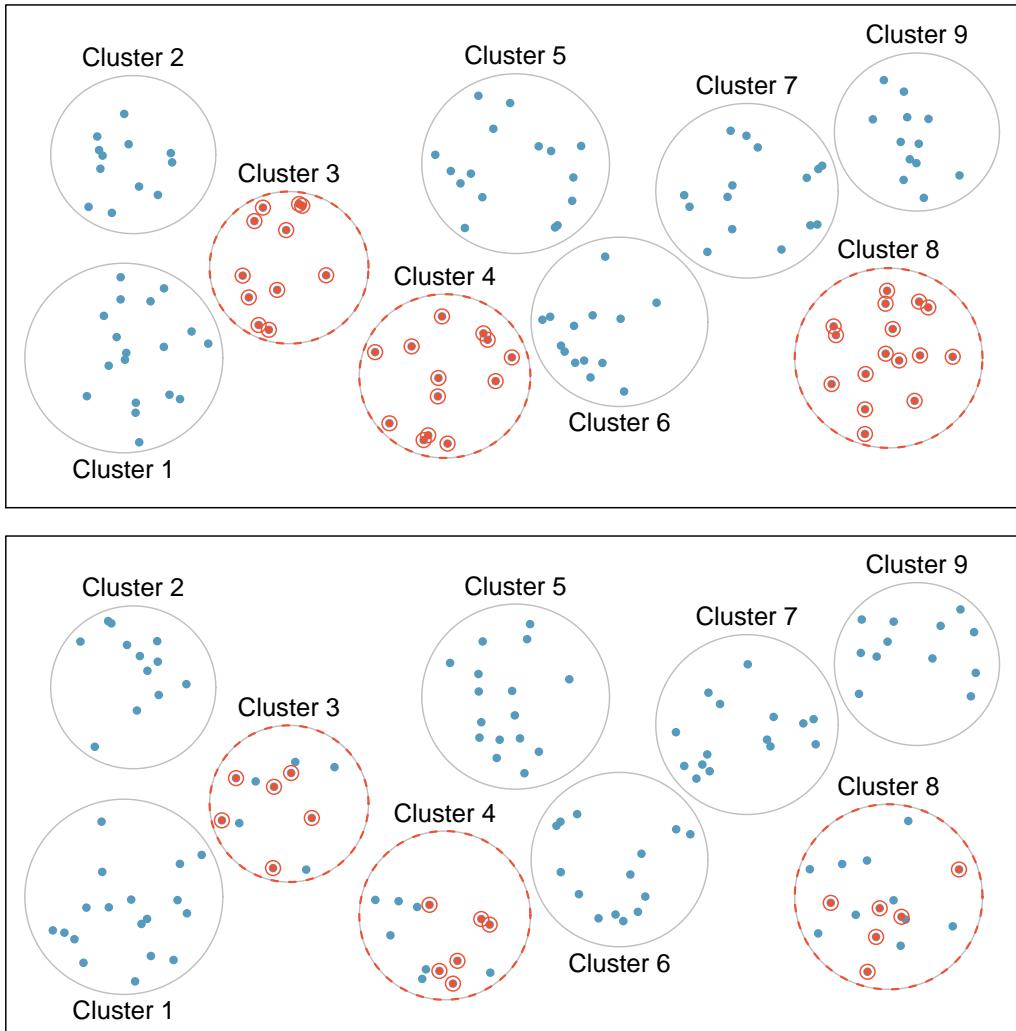


그림 1.15: 군집표집과 다단계표집 예제. 상단 그림에서 군집표집이 사용되었다. 여기서 데이터가 9개 군집으로 나누고, 이 군집중에서 세개를 표본으로 선택했고, 세 군집에 있는 모든 관측점이 최종표본에 포함됐다. 아래 그림에서 다단계표집이 사용되었다. 선택된 군집중에서 최종 표본에 포함되려면 각 군집의 일부를 무작위로 선택한다는 면에서 군집표집과 다르다.

## 5 실험

연구원이 처리(treatment)를 사례에 배정하는 연구를 **실험**이라고 부른다. 이러한 배정에 임의화(randomization)가 포함될 때, 예를 들어 동전을 던져 환자가 어떤 처리를 받을지 결정하는 것 같은, **임의화 실험(randomized experiment)**이라고 부른다. 두 변수 사이에 인과관계를 보여주려고 할 때 임의화 실험은 근본적으로 중요하다.

### 5.1 실험계획의 원칙

임의화 실험은 일반적으로 네 가지 원칙 아래 구성된다.

**제어(Controlling).** 연구원이 처리를 사례에 배정하고 집단에 다른 어떤 차이도 제어하기 위해서 최선을 다한다. 예를 들어, 환자가 알약형태 약을 복용할 때, 일부 환자는 약을 물 조금과 함께 복용하고, 다른 환자는 약을 가득찬 물 한잔과 함께 복용할 수 있다. 물 소비 효과를 제어하기 위해서, 의사가 모든 환자에게 약과 12 온스 물 한잔만 복용하도록 할 수 있다.

**임의화(Randomization).** 제어할 수 없는 변수를 설명하기 위해서, 연구원이 처리집단에 환자를 임의화한다. 예를 들어, 일부 환자는 다른 환자보다 식습관 덕분에 질병에 더 걸리기 쉬울 수 있다. 환자를 처리집단과 대조집단으로 임의화하면 그러한 차이를 균등하게 할 수 있고 우연적 편향이 연구에 들어오지 못하게도 한다.

**반복(Replication).** 연구원이 더 많은 사례를 관측하면 할수록, 반응변수에 설명변수 효과를 좀더 정확하게 추정할 수 있다. 단일 연구에서 충분히 큰 표본을 수집해서 반복한다. 추가적으로, 한 무리 과학자가 전체 연구를 반복해서 이전 연구결과를 확인한다.

**블로킹(Blocking).** 종종 연구원이 처리가 아닌 다른 변수가 반응에 영향을 준다고 알고 있거나 의구심을 갖고 있을 수 있다. 이러한 상황에서, 먼저 개인을 이 변수에 기반해서 블록으로 무리짓고 나서 처리 집단에 각 블록내부에서 사례를 확률화한다. 이 전략은 종종 블로킹으로 불린다. 예를 들어, 만약 심장마비에 대한 약효를 살펴본다면, 먼저 연구에 환자를 저위험 블록과 고위험 블록으로 나누고 나서 그림 1.16에 나와있듯이, 무작위로 각 블록에서 환자 절반을 대조집단에 다른 환자 절반을 치료집단에 배정한다. 이 전략은 각 처리집단에 동일한 숫자의 저위험과 고위험 환자를 갖도록 확실히 한다.

첫 세 가지 실험계획 원칙을 어떤 연구에도 포함하는 것이 중요하다. 이 책에서는 그런 실험에서 나온 데이터를 분석하는데 적용 가능한 방법을 기술한다. 블로킹은 다소 좀더 고급 기법으로 이 책에 나온 통계적 방법을 확장해서 블로킹을 사용해 수집된 자료를 분석한다.

### 5.2 인간 실험에 편향 줄이기

임의화 실험이 데이터 수집에 대한 황금률이지만, 모든 사례에 불편향 측면을 인관관계에 보장하지는 못한다. 인간 실험이 완벽한 예제로서 편향이 의도하지 않았는데 일어날 수 있다.

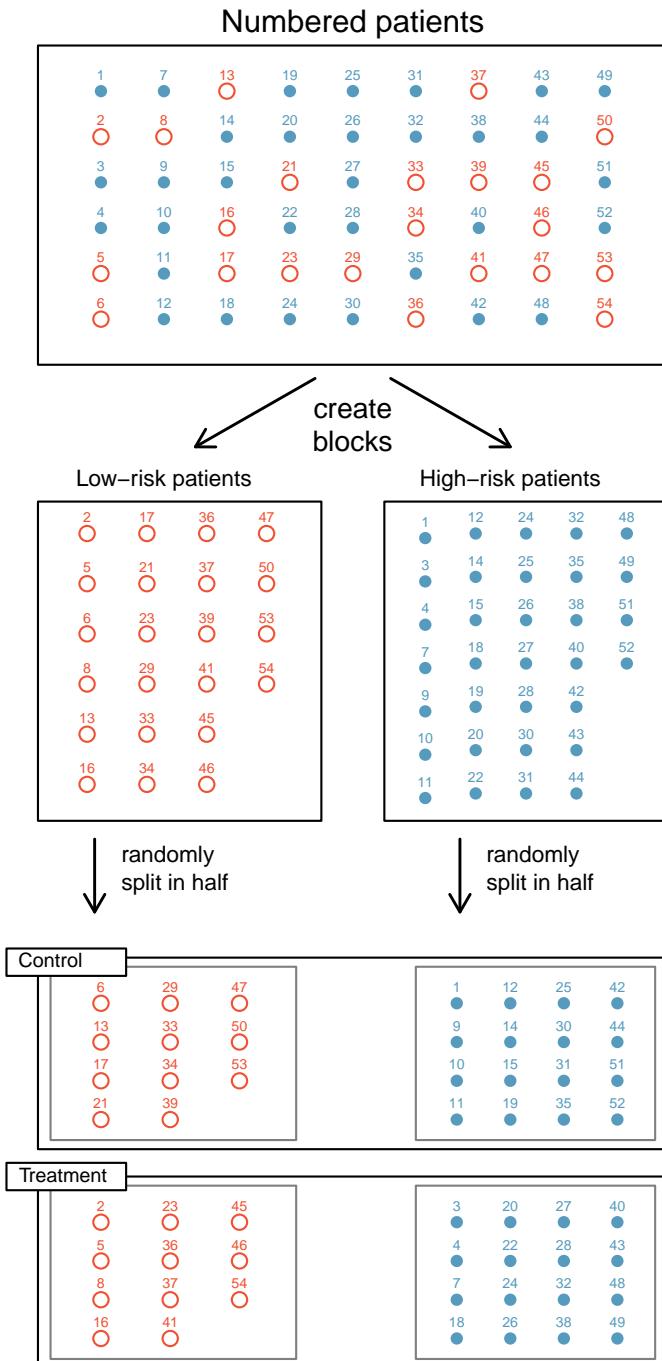


그림 1.16: 환자위험을 묘사하는 변수를 사용한 블로킹. 환자가 먼저 저위험 블록과 고위험 블록을 나누고 나서, 각 블록은 임의화를 사용해서 공평하게 처리집단으로 분리된다. 이 전략은 고위험과 저위험 범주 모두로부터 각 처리 집단에 동일한 환자 대표성을 갖도록 확실히 한다.

심장마비 환자를 처리하는데 신약을 사용한 연구를 다시 생각해보자.<sup>16</sup> 특히 연구원은 신약이 환자 사망률을 줄였는지 알고자 한다.

해당 연구원은 임의화 실험을 설계하는데 신약 효과에 관한 인과관계를 도출하고자 하기 때문이다. 연구 지원자<sup>17</sup>를 두 연구집단에 임의로 나누어 배치한다. 한 집단, **처리집단**은 신약을 처방받는다. 다른 집단, **대조집단**은 어떤 신약 처방도 받지 않는다.

연구에서 참여자 위치에 본인을 투영해보라. 만약 본인이 처리집단에 있다면, 도움이 기대되는 멋진 신약을 처방받을 것이다. 다른 한편으로, 또 다른 집단에 있는 참여자는 신약을 받지 못하고 놀면서 기다리는데 다만 실험 참여가 사망율을 증가시키지 않기를 기대할 뿐이다. 이러한 측면은 실제로 두가지 효과가 있다는 것을 제시한다: 관심을 둔 한 가지가 신약 효과이며, 두번째는 감정적 효과로 정량화하기가 어렵다.

연구원은 대체로 감정적 효과에는 관심이 없는데 연구에 편향을 준다. 이러한 문제를 우회하려고, 연구원은 참여자가 어느 집단에 속해있는지 알려주지 않는다. 연구원이 처리에 관한 정보를 환자에게 알려주지 않을 때, 연구를 **눈가립(blind)** 연구라고 한다. 하지만 문제가 하나 있다: 만약 환자가 처리를 받지 못하면, 환자를 본인이 대조집단에 있다고 알게된다. 이러한 문제에 해법이 거짓 처방을 대조집단 환자에 주는 것이다. 거짓 처방을 **위약**이라고 부르고, 효과적인 위약이 연구를 진정하게 눈가립으로 만드는 열쇠다. 전통적인 위약 사례가 실제 처방약처럼 만든 설탕약이다. 종종 위약을 통해 환자에서 적지만 실제 호전된 결과가 나오기도 한다. 이러한 효과를 **위약 효과**라는 말로 부른다.

환자만이 유일한 눈가립되어야 되는 사람은 아니다: 의사와 연구원도 우연히 연구에 편향을 줄 수 있다. 환자에 실제 처리가 주어지고 있다는 사실을 의사가 알게될 때, 의사가 무심코 위약이 처방된 환자보다 더 관심과 주의를 둘 수 있다. 일부 사례에서 측정가능한 효과를 갖는 것으로 확인된 이런 편향이 생기지 않도록 보호하기 위해서, **이중눈가립(double blind)** 방식을 사용한다. 환자와 상호작용하는 의사 혹은 연구원은 환자와 마찬가지로 처리를 받았는지 받지 않았는지 인식하지 못한다.<sup>18</sup>

 **Guided Practice 1.14** 1 절에 연구로 되돌아가자. 연구원이 스텐트가 위험에 처한 환자의 심근경색을 줄이는데 효과가 있는지 테스트한다. 이것은 실험인가요? 이 연구는 눈가립 방식인가요? 이 연구는 이중눈가립 방식인가요?<sup>19</sup>

## 6 숫자형 데이터 살펴보기

이번 절에서 숫자형 변수를 탐색하고 요약하는 기법을 소개한다. 2 절에서 email50과 county 데이터셋이 예제로 풍부한 기회를 제공했다. 숫자형 변수 결과는 숫자로 기본적인 산술연산을

<sup>16</sup>Anturane Reinfarction Trial Research Group. 1980. Sulfapyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

<sup>17</sup>인체 실험대상은 종종 **환자**, **지원자**, **연구 참여자**라고 불린다.

<sup>18</sup>항상 연구에 관련된 일부 연구원은 어느 환자가 어느 처리를 받고 있는지 알고 있다. 하지만, 이들 연구원은 연구 환자와 상호작용하지 못하고 누가 어느 처리를 받고 있는지 눈가립된 의료전문가에게 말하지 않는다.

<sup>19</sup>연구원이 환자를 처리집단에 배정했다. 그래서 해당 연구는 실험이다. 하지만, 환자는 무슨 처리를 받고 있는지 분간할 수 있다. 그래서 해당 연구는 눈가립이 아니다. 해당 연구는 이중눈가립이 될 수 없는데 연구가 눈가립 연구도 아니기 때문이다.

수행할 수 있다는 것을 상기하라. 예를 들어, `pop2010` 변수는 2010년 시군 인구를 대표하는데 숫자형이여서 두 시군 사이에 인구 차이와 비율을 분별력을 갖고 논의할 수 있다. 다른 한편으로, 지역코드와 우편번호는 숫자형이라기 보다는 범주형 변수다.

## 6.1 쌍체자료(paired data) 산점도

산점도는 두 숫자형 변수에 대한 데이터 사례별 뷔를 제공한다. 그림 1.8 on page 9에서 산점도를 사용해서 연방지출비용과 빙곤율이 `county` 데이터에서 연관되어 있는지 조사했다. 또 다른 산점도는 그림 1.17에 나와 있는데, `email150` 데이터셋에 전자우편에 줄바꿈(`line_breaks`)과 문자 숫자(`num_char`)를 비교했다. 임의 산점도에서 각 점은 사례 하나를 표현한다. `email150` 데이터셋에 50 개 사례가 있기 때문에, 그림 1.17에 점이 50개 있다.

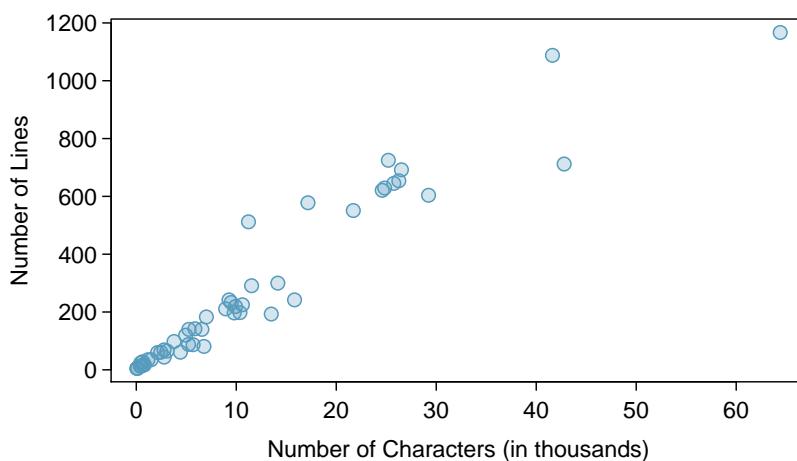


그림 1.17: `email150` 데이터에 `line_breaks` 와 `num_char`에 대한 산점도.

이러한 관점에서 문자숫자를 살펴보면, 이 문단은 문자 363개로 구성되었다. 그림 1.17을 살펴보면, 일부 전자우편은 말이 안될 정도로 장황한 것처럼 보인다. 좀 더 조사를 하면, 장문 전자우편 대부분은 실제로 HTML 형식을 사용하는 것을 알 수 있다. 이것이 의미하는 바는 해당 전자우편 문자 대부분이 텍스트로 정보를 제공하기 보다는 전자우편 서식에 사용되었다는 것이다.

- **Guided Practice 1.15** 산점도가 데이터에 관한 무엇을 밝혀냈고, 어떻게 유용하게 사용될 수 있을까?<sup>20</sup>

- **Example 1.16** 변수 두개와 54개 자동차로 구성된 신차 데이터셋을 고려해보자: 자동차 가격과 자동차 중량<sup>21</sup> 자동차 가격과 중량에 대한 산점도가 그림 1.18에 나타나 있다. 이를 변수 사이 관계에 감해서 무엇을 말할 수 있을까?

<sup>20</sup>해답은 다양하다. 산점도는 선속하게 변수와 관련된 연관성을 식별하는데 도움이 된다. 즉, 연관이 단순한 추세 형태로 오는지 혹은 좀 더 복잡한 관계로 나타나는지 알 수 있다.

<sup>21</sup>출처: [www.amstat.org/publications/jse/v1n1/datasets.lock.html](http://www.amstat.org/publications/jse/v1n1/datasets.lock.html) 데이터 일부

점선으로 부각되듯이 관계는 명백히 비선형이다. 이점이 지금까지 봤던 이전 산점도와 다른 면이다. 그림 1.8 on page 9와 그림 1.17은 매우 선형적 관계를 보여주고 있다.

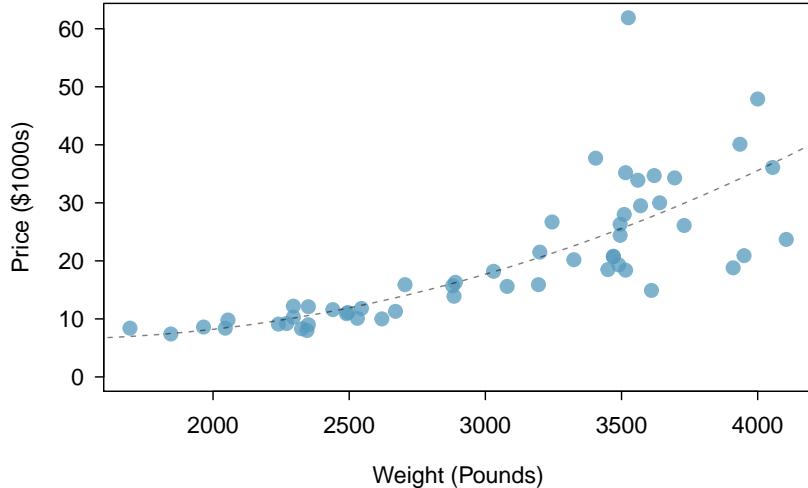


그림 1.18: 자동차 54개에 대한 price 대비 weight 산점도.

④ Guided Practice 1.17 산점도에서 말발굽 모양을 갖는 두변수를 기술하시요.<sup>22</sup>

## 6.2 점그림(Dot plots)과 평균

종종 두변수가 한 변수에 너무 많을 수 있다: 단지 한 변수만 관심이 있을 수 있다. 이러한 경우에 점그림이 가장 기본적인 시각적 출력력을 제공한다. 점그림(dot plot)은 일변수 산점도다; 50개 전자우편에서 나온 문자 숫자를 사용한 예제가 그림 1.19에 나타나 있다.

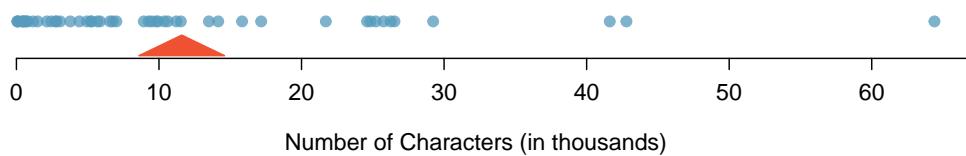


그림 1.19: email150 데이터셋에 대한 num\_char를 점그림으로 표현.

평균(평균)은 종종 평균(average)으로도 불리는데 데이터 분포의 중심을 측정하는 일반적인 방식이다. 50개 전자우편에 평균 문자숫자를 찾아내려면, 문자 갯수를 모두 더하고 전자우편 갯수로 나눈다. 계산 편의를 위해서, 문자 숫자는 첫단위로 표시되고 소수점 첫자리에서 반올림한다.

<sup>22</sup> 수직축은 뭔가 “좋은” 것을 표현하고 수평축은 단지 적당히 좋은 것을 표현한다. 건강과 물 소비가 이러한 기술에 적합한데 물을 너무 과도하게 소비하게 되면 독성을 갖는다.

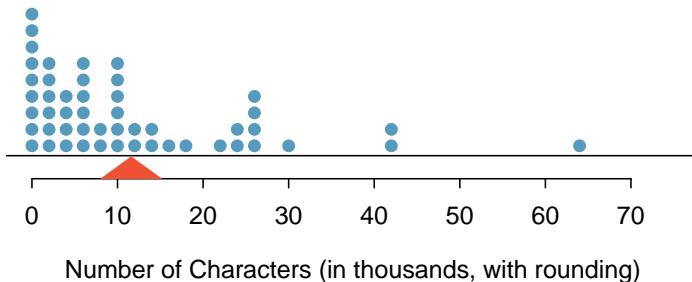


그림 1.20: email150 데이터셋에 대한 변수 num\_char의 스택 점그림(stacked dot plot). 상기 점그림에서 값을 가장 가까운 12,000에 반올림했다.

$$\bar{x} = \frac{21.7 + 7.0 + \dots + 15.8}{50} = 11.6 \quad (1.18)$$

표본 평균은  $\bar{x}$ 로 표기한다. 문자  $x$ 는 관심있는 변수, num\_char에 대한 일반 자리차지용도(placeholder)로 사용된다. 그리고,  $x$  위에 막대는 50개 전자우편 문자 평균 숫자가 11,600임을 전달한다. 평균을 분포의 균형점으로 간주하는 것이 유용하다. 표본 평균은 그림 1.19와 1.20에 삼각형으로 표현되어 있다.

$\bar{x}$   
표본  
평균

### 평균

숫자 변수의 표본 평균은 전체 관측값을 관측수로 나눠서 계산한다:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (1.19)$$

$x_1, x_2, \dots, x_n$ 은  $n$ 개 관측값을 표현한다.

$n$   
표본 크기

### ◎ Guided Practice 1.20

상기 방정식 (1.18) 와 (1.19) 자세히 보자.  $x_1$ 은 무엇에 대응되는가?  $x_2$ 는?  $x_i$ 가 무엇을 표현하는지에 대해 일반적 의미를 추론할 수 있는가? <sup>23</sup>

### ◎ Guided Practice 1.21 전자우편 표본에서 $n$ 은 얼마인가요?<sup>24</sup>

email150 데이터셋은 1월에서 3월 사이에 받은 전자우편 모집단에서 나온 표본을 나타낸다. 표본평균과 동일한 방식으로 모집단에 대한 평균을 계산할 수 있다. 하지만, 모집단 평균은 특별한 기호를 갖는다:  $\mu$ .  $\mu$  기호는 그리스 문자 *mu*(뮤)로 모집단에 모든 관측점 평균을 표

$\mu$   
모집단  
평균

<sup>23</sup>  $x_1$ 은 표본에서 첫번째 전자우편 문자갯수에 대응된다(21.7, 천단위),  $x_2$ 는 두번째 전자우편 문자갯수에 대응

현한다. 종종 아래첨자  $x$  처럼 사용해서 모집단 평균 어떤 변수를 지칭하는지 표현한다, 즉,  $\mu_x$ .

- **Example 1.22** 전체 전자우편에 걸친 평균 문자갯수를 표본 데이터를 사용해서 추정 할 수 있다. 50개 전자우편 표본에 기초해서, email 데이터셋의 전체 전자우편에 대한 평균 문자갯수,  $\mu_x$ 에 대한 합리적인 추정값은 얼마일까?(email로부터 표본 email150 데이터셋이 생성된 것을 상기하라.)

---

표본 평균, 11,600이 합리적인  $\mu_x$ 의 추정값을 제공할 수 있다. 이 숫자가 완벽하지는 않지만, 모집단 평균의 점추정값(point estimate)을 제공한다. ?? 장과 그 후에, 점추정값 정확성을 특정짓는 도구를 개발할 것이고, 좀더 큰 표본에 기반한 점추정값이 좀더 작은 표본에 기반한 것보다 좀더 정확한 경향이 있다는 것을 알게 된다.

- **Example 1.23** 미국의 일인당 평균소득을 계산해보자. 이를 위해서, 먼저 county 데이터셋에서 3,143개 시군의 일인당 소득 평균을 생각할 수 있다. 더 나은 접근법은 무엇일까?

---

county 데이터셋은 각 시군이 사실 많은 개별 주민을 대표한다는 점에서 특별하다. 만약 income 변수로 단순히 평균을 계산한다면, 계산과정에서 주민 5,000명과 주민 5,000,000명 갖는 군을 동일하게 처리하는 것이다. 대신에, 각 시군에 전체 소득을 계산하고, 모든 시군의 합을 더해서 총합을 계산하고 나서, 전체 시군 사람숫자로 나눠야 된다. 만약 county 데이터로 이런 단계를 완료하면, 미국 일인당 소득은 \$27,348.43이 된다. 만약 시군에 걸쳐 일인당 소득의 단순 평균을 계산하면, 결과는 단지 \$22,504.70만 된다!

예제 1.23에서 소위 **가중평균**을 사용했는데 이 교과서의 주요 주제는 아니다. 하지만, 관심있는 독자를 위해서 가중평균에 대한 온라인 보충정보를 제공한다:

[www.openintro.org/stat/down/supp/wtdmean.pdf](http://www.openintro.org/stat/down/supp/wtdmean.pdf)

### 6.3 히스토그램과 형상

점그림은 각 관측점에 대한 정확한 값을 보여준다. 작은 데이터셋에는 유용하지만, 표본 크기가 커지면 가독성이 떨어진다. 각 관측점 값을 보여주는 대신에, 값이 구간(bin)에 속한 것으로 생각하는게 더 낫다. 예를 들어, email150 데이터셋으로 문자갯수가 0에서 5,000인 경우 사례를 갯수로 세고, 5,000에서 10,000인 경우 갯수로 세고 등등 표를 생성한다. 구간 경계에 떨어지는 관측점(즉, 5,000)은 아래쪽 구간에 할당한다. 이렇게 도표로 작성한 된 것이 표 1.21이다. 이런 구간 갯수를 막대로 소위 **히스토그램**으로 불리는 그림 1.22에 그렸는데, 그림 1.20에 표현된 스택점그림과 닮았다.

---

된다 (7.0, 천단위), 그리고  $x_i$ 는 데이터셋에서  $i^{th}$  번째 전자우편 문자갯수에 대응된다.  
<sup>24</sup> 표본크기는  $n = 50$ 이다.

문자 (단위: 천)	0-5	5-10	10-15	15-20	20-25	25-30	...	55-60	60-65
갯수	19	12	6	2	3	5	...	0	1

표 1.21: num\_char 데이터에 대한 구간별 갯수.

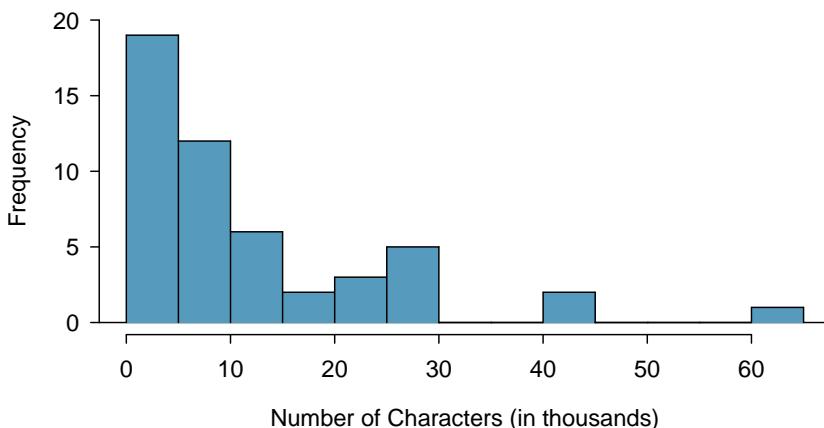


그림 1.22: 변수 num\_char의 히스토그램. 분포가 매우 강하게 우측으로 기울었다.

히스토그램은 **데이터 밀도**(data density)를 볼 수 있게 한다. 더 높은 막대는 데이터가 상대적으로 더 흔하게 위치함을 나타낸다. 예를 들어, 데이터셋에서 적어도 20,000개 문자를 갖는 전자우편보다 20,000개 문자보다 더 적은 문자를 갖는 전자우편이 훨씬 더 많다. 막대는 데이터 밀도가 문자 갯수에 상대적으로 어떻게 변하는지 보기 쉽게 한다.

히스토그램은 특히 데이터 분포 형상을 기술하는데 편리하다. 그림 1.22에는 전자우편 대부분이 상대적으로 적은 문자를 갖는다는 것을 보여주지만, 일부 전자우편은 매우 많은 문자를 갖는다는 것이 보여진다. 이러한 방식으로 데이터가 오른쪽으로 잣아들고 우측으로 좀더 긴 꼬리를 갖을 때, 형상이 **우측으로 기울었다**고 한다.<sup>25</sup>

역의 특성을 갖는 데이터셋 – 좌측으로 길고, 얇은 꼬리 –은 **좌측으로 기울었다**고 한다. 또한 그러한 분포는 좌측으로 긴 꼬리를 갖는다고 한다. 양쪽 방향으로 대략 동일하게 잣아드는 경향을 보이는 데이터셋을 **대칭**이라고 부른다.

### 기울을 식별하는 긴꼬리

데이터가 한방향으로 잣아들 때, 분포가 **긴꼬리**(long tail)를 가졌다고 한다. 만약 분포가 좌측으로 긴꼬리를 갖으면, 좌측으로 기울었다. 만약 분포가 우측으로 긴꼬리를 갖는다면, 우측으로 기울었다.

<sup>25</sup>우측으로 기울 데이터를 기술하는 다른 방식이 있다: 오른쪽으로 기울, 고가쪽으로 기울, or 양수쪽으로 기울.

④ Guided Practice 1.24 그림 1.19 와 1.20에 점그림을 살펴보라. 데이터에서 기울을 볼 수 있는가? 히스토그램에서 기울을 보기 더 쉬운가 혹은 점그림에서 더 쉬운가?<sup>26</sup>

④ Guided Practice 1.25 (표식으로 표현되어) 평균을 제외하고, 히스토그램에서 볼 수 없지만 점그림에서 무엇을 볼 수 있나요?<sup>27</sup>

분포가 기울었는지 대칭인지 살펴보는 것에 추가해서, 히스토그램을 사용해서 모드를 식별할 수 있다. 모드(mode)는 분포에 눈에 띠는 봉우리로 표현된다.<sup>28</sup> 변수 num\_char에 대한 히스토그램에는 단지 눈에 띠는 봉우리가 하나만 있다.

그림 1.23에 있는 히스토그램에는 하나, 둘, 혹은 세개 눈에 띠는 봉우리가 있다. 이러한 분포를 각각 단봉(unimodal), 이봉(bimodal), 다봉(multimodal)으로 부른다. 2개 이상 눈에 띠는 봉우리를 갖는 분포는 다봉(multimodal)으로 부른다. 단봉 분포에서 눈에 띠는 봉우리가 하나만 있는 것에 주목한다. 여기서 단지 관측점 몇개로만 인접 구간과 차이나기 때문에 두번째 눈에 띠는 봉우리는 별도 봉우리로 간주되지 않았다.

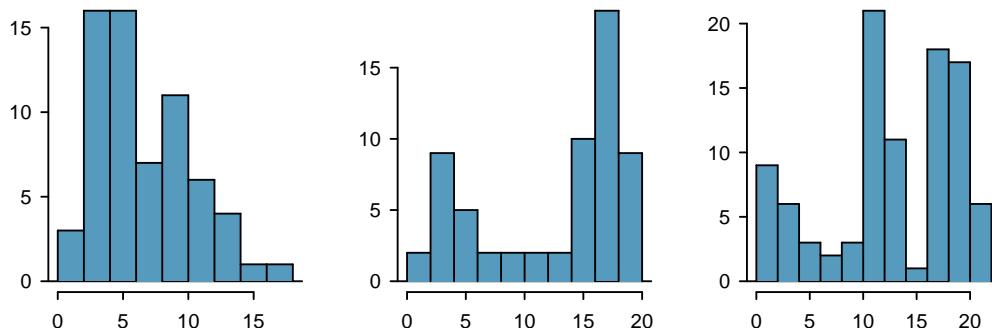


그림 1.23: 단지 눈에 띠는 봉우리만 개수함. (좌측에서 우측으로) 분포가 단봉, 이봉, 다봉임을 나타낸다.

④ Guided Practice 1.26 그림 1.22에는 문자 갯수에 단지 눈에 띠는 모드가 하나만 드러났다. 해당 분포는 단봉, 이봉, 혹은 다봉인가?<sup>29</sup>

④ Guided Practice 1.27 초등학교 3학년 어린 학생과 성인 선생님 키를 졌다. 생성된 신장 데이터셋에서 몇개 모드를 기대합니까?<sup>30</sup>

<sup>26</sup> 기울(왜도)는 그림 세개 모두에서 볼 수 있다. 하지만, 평면 점그림(flat dot plot)이 가장 덜 유용하다. 스택점그림과 히스토그램이 기울을 식별하는데 도움이 되는 시각화 도구가 된다.

<sup>27</sup> 각 개별 문자우편에 대한 문자 갯수.

<sup>28</sup> 일반적으로 통계에서 사용되지 않는 모드의 또 다른 정의는 가장 많은 발생값이다. 데이터셋에 같은 값을 갖는 관측점이 없는 것이 일반적으로 많은 실제 데이터셋에 대해 이런 정의는 쓸모없게 된다.

<sup>29</sup> 단봉(unimodal)이다. uni는 1을 상징함을 기억하라 (외발자전거 unicycles을 생각한다). 유사하게  $b_i$ 는 2를 상징한다 (두발 자전거 bicycles를 생각한다). (이러한 비유를 완성하도록 다발 자전거 multicycle가 발명되길 희망한다.)

<sup>30</sup> 데이터셋에 눈에 보이는 두개 키집단이 있을 수 있다: 어린 학생과 성인 선생님. 즉, 데이터는 아마도 이봉이다.

**TIP: 모드(mode) 찾기**

모드를 찾을 것이 분포에 내재하는 모드 숫자에 관해 명확하고 올바른 정답을 찾는 것은 아니다. 이러한 사유로 이 책에서 눈에 띄는(prominent)이라는 표현은 엄격하게 정의되지 않는다. 이러한 조사에서 중요한 점은 데이터가 어떻게 구조화되었는지 데이터를 더 잘 이해하는 것이다.

## 6.4 분산과 표준편차

데이터셋 중심을 기술하는 방법으로 평균을 소개했지만, 데이터에 있어 변동성도 중요하다. 여기서, 변동성을 측정하는 두 측도를 소개한다: 분산과 표준편차. 설사 손으로 공식에 따라 계산하는 것은 다소 지루할 수 있지만, 둘 모두 자료분석에서 매우 유용하다. 표준편자는 둘 중에서 이해하기 더 쉽고, 일반 관측점이 평균으로부터 얼마나 떨어져 있는지 대략 기술해준다.

평균으로부터 관측점의 거리를 편차(deviation)라고 부른다. `num_char` 변수에 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 와 50<sup>th</sup> 번째 관측점에 대한 편차가 다음에 나와 있다. 계산 편의를 위해서, 문자 숫자는 천단위로 표시되고 소수 첫자리에서 반올림했다.

$$x_1 - \bar{x} = 21.7 - 11.6 = 10.1$$

$$x_2 - \bar{x} = 7.0 - 11.6 = -4.6$$

$$x_3 - \bar{x} = 0.6 - 11.6 = -11.0$$

$$\vdots$$

$$x_{50} - \bar{x} = 15.8 - 11.6 = 4.2$$

만약 편차를 제곱하고 나서 평균을 계산하면, 결과는  $s^2$

$$s^2$$

표본 분산

$$\begin{aligned}s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44\end{aligned}$$

분산을 계산할 때,  $n$ 으로 나누는 대신에  $n - 1$ 로 나눈다: 이 교재에 사용된 이러한 수학적 뉴앙스에 대해서 걱정할 필요는 없다. 편차를 제곱하면 두가지를 수행함에 주목한다. 먼저,  $10.1^2$ ,  $(-4.6)^2$ ,  $(-11.0)^2$ , 와  $4.2^2$ 을 비교하면 알 수 있듯이, 큰 값을 훨씬 더 크게한다. 둘째로, 모든 음의 부호를 제거한다.

**표준편차**는 분산에 제곱근을 씌운 것으로 정의된다:

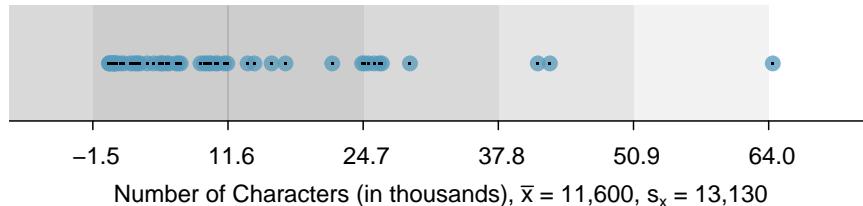


그림 1.24: num\_char 데이터에서 전자우편 50개 중에서 41개(82%)가 평균의 1 표준편차 내에 있고, 50개 중 47개(94%)는 2 표준편차 내에 위치한다. 대체로 약 데이터의 70%는 1 표준편차 내에 있고, 95%는 2 표준편차 내에 있다. 하지만 이러한 경험 규칙은 상기 예제에서 볼 수 있듯이 기울어진 데이터에 대해서는 정확도가 떨어진다.

 $s$ 

표본 표준편차

$$s = \sqrt{172.44} = 13.13$$

전자우편에 문자 갯수 표준편자는 약 13.13 천개다. 첨자  $x$ 는 분산과 표준편차에 추가될 수 있다. 즉,  $x_1, x_2, \dots, x_n$  으로 표현되는 관측점의 분산과 표준편차를 상기시키도록  $s_x^2$  와  $s_x$  처럼 나타낸다. 분산과 표준편차가 어느 데이터를 참조하는지 명확할 때는 첨자  $x$ 를 대체로 생략한다.

### 분산과 표준편차

분산은 대략적으로 평균으로부터 평균 제곱거리다. 표준편자는 분산의 제곱근이다. 데이터가 평균에 얼마나 가까울지 고려할 때는 표준편자가 유용하다.

 $\sigma^2$ 

모분산

 $\sigma$ 

표표준편차

### TIP: 표준편자는 변동성을 기술한다

공식보다는 변동성을 기술하는 도구로서 표준편차 개념적 의미에 집중하라. 대체로 데이터 70%는 평균의 1 표준편차 내에 있고 약 95%는 2 표준편차 내에 있게 된다. 하지만, 그림 1.24 와 1.25에 나타나듯이, 이런 백분율이 엄격한 규칙은 아니다.

- ④ Guided Practice 1.28 페이지 27 쪽에서 분포형상의 개념을 소개했다. 분포형상의 좋은 기술은 모드(modality)와 분포가 대칭인지, 한쪽으로 기울어졌는지에 대한 정보를

<sup>31</sup>유일한 차이점은 모집단 분산은  $n - 1$  대신에  $n$ 으로 나눈다는 것이다.

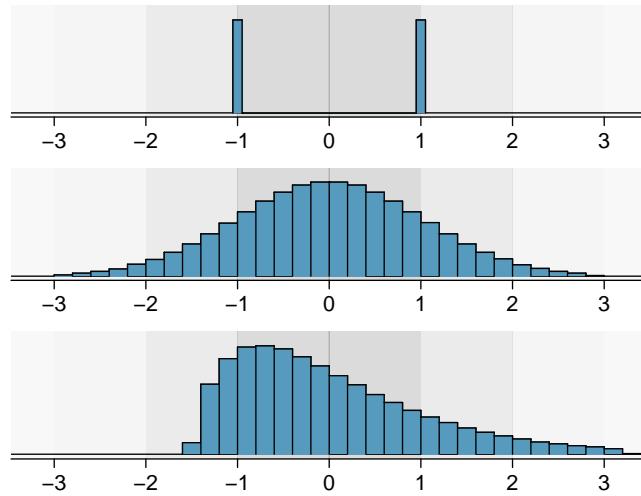


그림 1.25: 동일한 평균  $\mu = 0$  와 표준편차  $\sigma = 1$  을 갖는 매우 다른 세가지 모집단 분포.

포함해야 한다. 예제로 그림 1.25을 사용해서 이러한 기술이 왜 중요한지 설명하시요.<sup>32</sup>

- **Example 1.29** 그림 1.22 on page 27에 히스토그램을 사용해서 변수 num\_char의 분포를 기술하시요. 기술할 때, 중심, 변동성, 분포 형상을 사용해야 하고 문맥적으로 유의미하게 배치되어야 된다: 전자우편 문자갯수. 또한 특별히 유별란 사례에 대해서도 언급한다.

---

전자우편 문자갯수 분포는 단봉이고 매우 높은쪽(high-end)으로 기울어져 있다. 많은 갯수가 평균 11,600 근처에 위치하고, 대부분은 평균의 1 표준편차 (13,130) 내에 포함된다. 예외적으로 장문 전자우편이 하나 있는데 약 65,000개 문자로 되어 있다.

실무에서, 분산과 표준편차는 목적을 달성하기 위한 수단으로 사용된다. 여기서 목적은 표본 통계량과 연관된 불확실성을 정확하게 추정할 수 있는 것이다. 예를 들어, ?? 장에서 분산과 표준편차를 사용해서 표본평균이 모평균에 얼마나 근접한지 평가한다.

## 6.5 상자그림, 사분위수, 중위수

상자그림(boxplot)은 특이한 관측점도 도면에 나타낼 수 있지만, 통계량 5개를 사용해서 데이터셋을 요약한다. email150 데이터셋에서 num\_char 변수를 상자그림과 함께 수직 점그림으로 그림 1.26에 표현했다.

---

<sup>32</sup> 그림 1.25은 매우 다르게 보이는 분포 세개를 보여주고 있지만, 모두 동일한 평균, 분산, 표준편차를 갖는다. 모드(modality)를 사용해서, 첫번째 플롯(이봉)과 나머지 두 플롯(단봉)을 구별할 수 있다. 기울(왜도)을 사용해서, 마지막 플롯(우측 기울)과 첫 두 플롯을 구별할 수 있다. 히스토그램 같은 그림이 좀더 완전한 이야기를 전해 주는 동안에, 모드(modality)와 형상(대칭/기울)을 사용해서 분포에 관한 기본 정보를 특징으로 나타낼 수 있다.

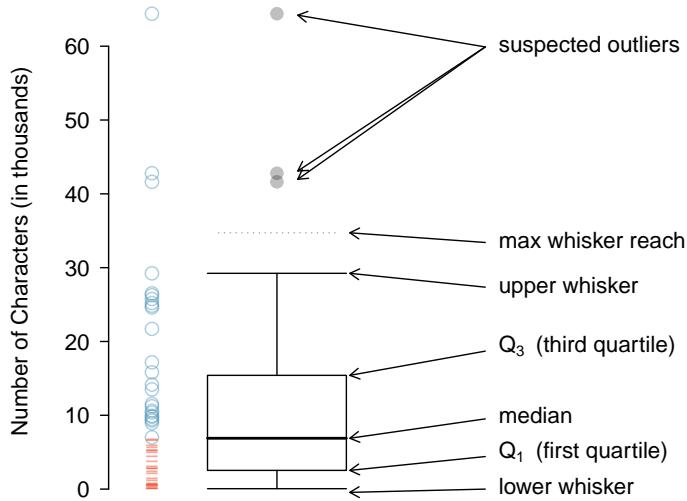


그림 1.26: 전자우편 50개에서 문자갯수에 대한 수직 점그림과 옆에 표식을 붙인 상자그림. 중위수 (6,890)는 데이터를 상위 50%와 하위 50%로 나누는데, 점그림에서 수평대쉬와 열린원으로 각각 표식했다.

상자그림을 만들 때 첫번째 단계가 **중위수**를 표시하는 검은선을 그리는데, 데이터를 반으로 나누는 역할을 한다. 그림 1.26에 중위수 아래 데이터 50%가 자리잡고(대쉬), 중위수 위에 50% 데이터가 포함된다(열린 원). 데이터셋 (짝수)에 50개 문자갯수가 있어서 데이터가 완벽하게 25개씩 두 집단으로 쪼개진다. 이런 경우 중위수는 50<sup>th</sup>번째 백분위에 가장 가까운 두 관측점을 평균낸 것이 된다:  $(6,768 + 7,012)/2 = 6,890$ . 관측점이 홀수인 경우에, 데이터를 둘로 쪼개는 관측점은 단 한개만 있게 되고, 이런 경우 해당 관측점이 중위수가 된다 (평균을 낼 필요가 없음).

### 중위수: 중간에 위치한 숫자

만약 데이터를 가장 작은 것부터 가장 큰 순으로 정렬한다면, **중위수**는 정중앙에 위치한 관측점이다. 만약 관측점 숫자가 짝수이면, 중앙에 값이 두개 있어서 중위수는 평균을 내서 얻게 된다.

상자그림을 만드는 두번째 단계는 사각형을 그려서 가운데 데이터 50%를 나타낸다. 그림 1.26에 수직으로 나타난 상자 길이를 **사분위 범위** (interquartile range, IQR)라고 부른다. 표준편차처럼 데이터 변동성 측도다. 데이터 변동성이 크면 클수록, 표준편차와 IQR은 더 커진다. 상자의 두 경계값을 **제1사분위수** (25<sup>th</sup> 번째 백분위, 즉 데이터 25% 가 이 값에 포함된다.) 그리고 **제3사분위수** (75<sup>th</sup> 번째 백분위수). 각각  $Q_1$ ,  $Q_3$ 로 종종 표시한다.

### 사분위 범위 (IQR)

IQR는 상자그림에서 상자 길이다. 다음과 같이 계산된다.

$$IQR = Q_3 - Q_1$$

$Q_1$  과  $Q_3$ 는 25<sup>th</sup>번째, 75<sup>th</sup>번째 백분위수다.

- Guided Practice 1.30  $Q_1$ 과 중위수 사이에 몇 퍼센트 데이터가 포함되는가? 중위 수와  $Q_3$  사이에 몇 퍼센트 데이터가 포함되는가?<sup>33</sup>

상자로부터 쭉 연장해서, **상자수염**(whiskers)은 상자 밖에 있는 데이터를 잡아내는데 사용된다. 하지만, 범위는 결코  $1.5 \times IQR$ 을 넘어가지 못한다.<sup>34</sup> 상장수염은 범위내 모든 것을 포함한다. 그림 1.26에서 위쪽 상자수염은 마지막 세점까지 연장되지 못했다. 세점은  $Q_3 + 1.5 \times IQR$  범위를 벗어나서 상자수염은 단지 마지막 점 아래까지만 연장되었다. 아래쪽 상자수염은 가장 낮은 값, 33에서 멈췄다. 왜냐하면, 더이상 도달할 데이터가 없기 때문이다; 아래쪽 상자수염 한계가 그림에서 나타나지 않았는데 이유는 상자그림은  $Q_1 - 1.5 \times IQR$  밑으로 연장되지 않기 때문이다. 이러한 점에서 상자는 상자그림의 몸통과 같고, 상자수염은 나머지 데이터에 도달하려고 애쓰는 팔과 같다.

상자수염 밖에 놓여있는 관측점은 점으로 표시되었다. 이러한 점을 표시하는 목적은 – 상자수염을 최소값과 최대값으로 연장하는 대신에 – 나머지 데이터로부터 특이하게 멀리 떨어져 보이는 관측점을 식별하게 돋는 것이다. 특이하게 멀리 떨어진 관측점을 **이상점**(outlier)이라고 부른다. 상기 사례에서, 이상점으로 문자 갯수 41,623, 42,793, 64,401 을 갖는 전자우편을 분류하는 것이 합리적인데 이유는 수치적으로 대부분의 데이터와 멀리 떨어져 있기 때문이다.

### 이상점은 맨 끝에 있다

이상점은 나머지 데이터와 비교해서 상대적으로 비교적 맨 끝에 있어 보이는 관측점이다.

<sup>33</sup>  $Q_1$  과  $Q_3$ 는 중간 데이터 50%를 잡아내고, 중위수가 데이터 중간을 가르기 때문에 25% 데이터가  $Q_1$ 과 중위수 사이에 포함되고, 또 다른 25% 데이터가 중위수와  $Q_3$  사이에 포함된다.

<sup>34</sup> 정확하게 1.5를 선택한 것이 작위적이지만, 상자그림에서 가장 일반적으로 사용된다.

**TIP: 이상점을 찾는 것이 왜 중요한가**

가능한 이상점에 대한 데이터 조사는 다음을 포함하는 유용한 많은 목적을 수행한다

1. 분포에 강한 기울을 식별한다.
2. 데이터 수집 혹은 입력 오류 식별. 예를 들어, 문자 64,401 개를 갖는다고 주장하는 전자우편을 다시 재조사해서, 해당 값이 정확한지 점검한다.
3. 데이터에 내재한 흥미로운 특성에 대한 직관을 제공

**◎ Guided Practice 1.31**

이상점으로 의심된 관측점 64,401은 정확한 관측점으로 판명났다. 이러한 관측점은 전자우편 문자갯수의 본질에 관해 제시하는 것이 무엇일까요?<sup>35</sup>

**◎ Guided Practice 1.32** 그림 1.26을 사용해서, email150 데이터셋에서 num\_char 변수에 대한 다음 값을 추정하세요: (a)  $Q_1$ , (b)  $Q_3$ , (c) IQR.<sup>36</sup>**🎥 Calculator videos**

Videos covering how to create statistical summaries and box plots using TI and Casio graphing calculators are available at [openintro.org/videos](http://openintro.org/videos).

**6.6 강건 통계량(Robust statistics)**

num\_char 데이터셋의 표본 통계량은 64,401 관측점에 의해서 어떻게 영향을 받는가? 만약 전자우편이 관측되지 않았다면 상황이 어떻게 됐을까? 만약 64,401 관측점이 더 큰 가령 150,000 이었다면 요약통계량은 어떻게 됐을까? 그림 1.27에 원데이터와 함께 이러한 시나리오를 도표에 그려놨다. 표본 통계량도 이러한 시나리오 맞춰 표 1.28에 계산했다.

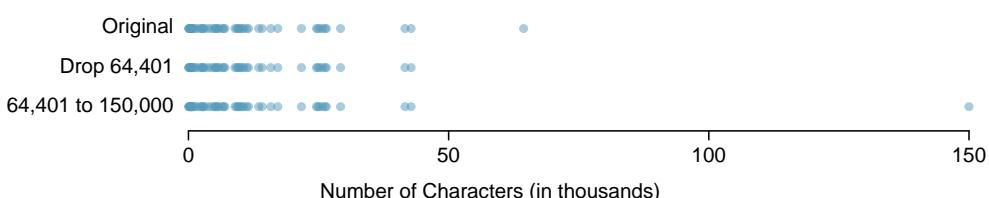


그림 1.27: 원 문자갯수 데이터와 두 변형된 데이터셋에 대한 점그림.

<sup>35</sup>가끔 매우 긴 전자우편이 있을 수도 있다.

<sup>36</sup>시각적 추정값은 사람마다 조금씩 다를 수 있다:  $Q_1 = 3,000$ ,  $Q_3 = 15,000$ , IQR =  $Q_3 - Q_1 = 12,000$ . (실제값:  $Q_1 = 2,536$ ,  $Q_3 = 15,411$ , IQR = 12,875.)

시나리오	강건		강건하지 않음	
	중위수	IQR	$\bar{x}$	$s$
원 num_char 데이터	6,890	12,875	11,600	13,130
관측점 66,924 누락	6,768	11,702	10,521	10,798
66,924을 150,000로 취환	6,890	12,875	13,310	22,434

표 1.28: 극단 관측점이 존재할 때, 중위수, IQR, 평균( $\bar{x}$ ), 표준편차( $s$ )가 어떻게 변화하는지 비교.

- ④ **Guided Practice 1.33** (a) 극단 관측점에 의해서 평균과 중위수 중 어느 것이 더 영향을 받는가? 표 1.28가 도움이 될 수 있다. (b) 표준편차와 IQR 중 어느 것이 극단 관측점에 의해 영향을 받는가?<sup>37</sup>

중위수와 IQR을 **강건 통계량**(robust estimates)으로 부르는데 극단 관측점이 강건 통계량에 거의 영향을 주지 않기 때문이다. 평균과 표준편차가 극단 관측점 변화에 영향을 훨씬 더 많이 받는다.

- **Example 1.34** 중위수와 IQR은 표 1.28의 세가지 시나리오에서 그다지 많이 변경되지 않는다. 왜 그럴까요?

중위수와 IQR은  $Q_1$ , 중위수,  $Q_3$  근방 숫자에만 민감하다. 해당 지역 값은 상대적으로 안정되어 있기 때문에 – 관측점 사이에 큰 점프는 없다 – 중위수와 IQR 추정값도 또한 꽤 안정되어 있다.

- ④ **Guided Practice 1.35** 자동차 가격 분포는 고급차 몇종과 경주차가 오른쪽 꼬리에 머물고 있어서 우측으로 기운 경향이 있다. 만약 신차를 검색하고서 가격에 주의를 기울인다면, 일반적인 차를 구입하려고 시장에 나왔다고 가정하고서 자동차 평균 혹은 중위수 가격 중 어디에 더 관심을 둘어야 하나요?<sup>38</sup>

## 6.7 데이터 변환 (특별 주제)

데이터가 매우 강하게 기울어졌을 때, 종종 데이터를 변환하면 모형화하기 쉽다. 2010년도 메이저 리그 야구선수 연봉 정보로부터 연봉을 히스토그램으로 그려보자. 그림 1.29(a)에 결과가 나와 있다.

- **Example 1.36** MLB 야구선수 연봉 히스토그램은 데이터가 극단적으로 기울어져 있고 약 \$1 백만달러(중위수로 측정됨)에 몰려 있다는 점에서 유용하다. 이 히스토그램에 관해서 어떤 점은 유용하지 않은가?

<sup>37</sup>(a) 평균이 더 영향을 받음. (b) 표준편차가 더 영향을 받음. 전체적 설명은 다음에 오는 실습지도 1.33에 나와 있다.

<sup>38</sup>“일반적인 자동차(regular car)” 구매자는 중위수 가격에 관심을 둬야 한다. 고급 자동차 판매가 평균 가격을 심하게 부풀릴 수 있지만, 중위수는 고급차 판매 영향에 더 강건한다.

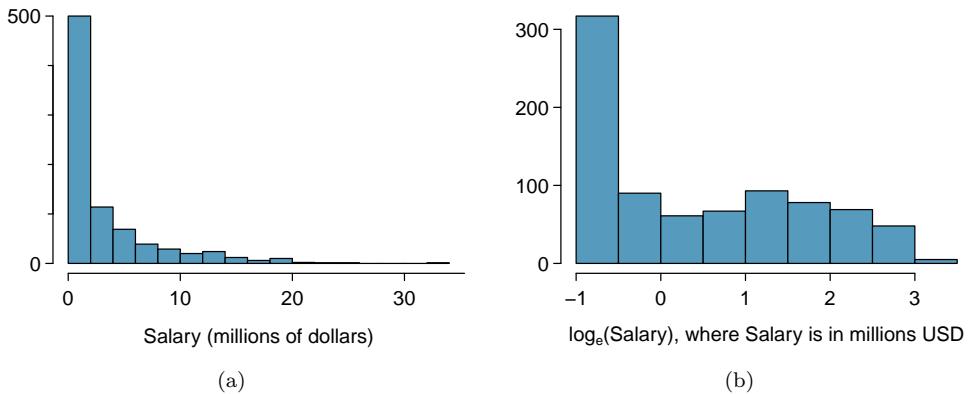


그림 1.29: (a) 2010년도 MLB 야구선수 연봉 히스토그램 (단위 백만달러).(b) 2010년도 MLB 야구선수 로그변환한 연봉 히스토그램.

데이터 대부분이 히스토그램 한 구간으로 수집되어 있고, 데이터가 매우 강하게 기울어져 있어서 데이터의 많은 자세한 정보가 뭉개져 있다.

모든 관측점이 양수이고, 데이터 상당수가 0 근처에 몰려있을 때 (데이터셋에 상대적으로 더 많은 값), 적용되는 표준 변환이 몇가지 있다. **변환**(transformation)은 함수를 사용해서 데이터를 재조정하는 것이다. 예를 들어, 야구선수 연봉에 자연로그<sup>39</sup>를 취한 그림이 그림 1.29(b)에 있는 새로운 히스토그램이다. 통계적 모형을 적용할 때 변환된 데이터는 종종 더 작업하기 쉽다. 왜냐하면, 변환된 데이터는 훨씬 덜 기울어져 있고 이상점이 일반적으로 극단에 덜 위치한다.

산점도에서 변환을 변수 하나 혹은 둘 모두에 적용할 수 있다. 그림 1.30(a)에 변수 `line_breaks` 와 `num_char`의 산점도가 나와 있다. 이것은 앞에서 살펴본 그림 1.17이다. 변수 사이에 양의 연관성을 볼 수 있고 많은 관측점이 0 근처에 군집지어 있다. ?? 장에서 직선을 사용해서 데이터를 모형화한다. 하지만, 현재 상태로 데이터를 모형화가 잘 안될 것을 알게 된다. 그림 1.30(b)에 산점도가 나와 있는데 변수 `line_breaks` 와 `num_char`을 로그 (밑  $e$ ) 변환을 사용해서 변환했다. 플롯 각각에 양의 연관성이 있지만, 변환된 데이터가 더 꾸준한 경향을 보이는데 변환되지 않은 데이터보다 모형화하기 더 쉽다.

또한, 로그 말고 다른 변환도 유용하다. 예를 들어 제곱근( $\sqrt{\text{원 관측점}}$ )과 역( $\frac{1}{\text{원 관측점}}$ ) 변환을 통계학자가 사용한다. 데이터를 변환하는 공통된 목적은 데이터 구조를 다르게 보고, 왜도를 줄이고, 모형화를 돋거나 산점도에 나와 있는 비선형 관계를 직선화하는데 있다.

## 6.8 데이터 매핑 (특별 주제)

`county` 데이터셋에는 많은 숫자형 변수가 있어서 점그림, 산점도, 상자그림을 사용해서 플롯을 그릴 수 있지만, 이러한 그림은 데이터 본질을 놓치고 있다. 그보다는 지리정보 데이터를

<sup>39</sup>통계학자는 종종 자연로그를 log로 사용한다. ln로 자연로그를 적는데 더 익숙할 수도 있다.

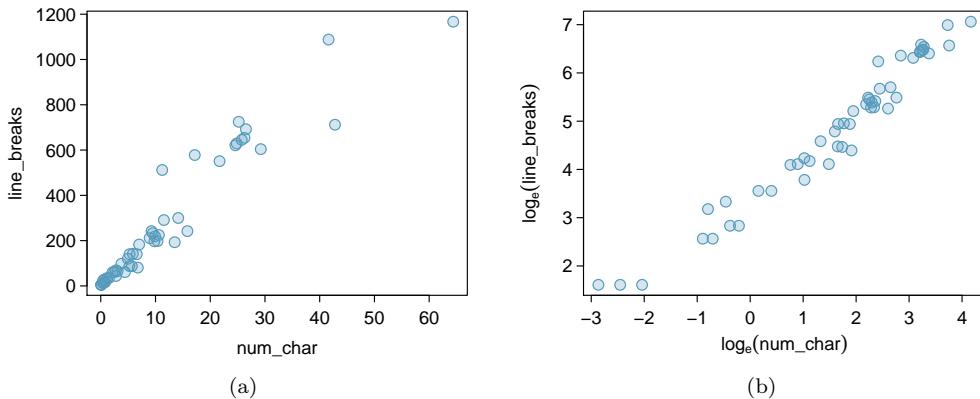
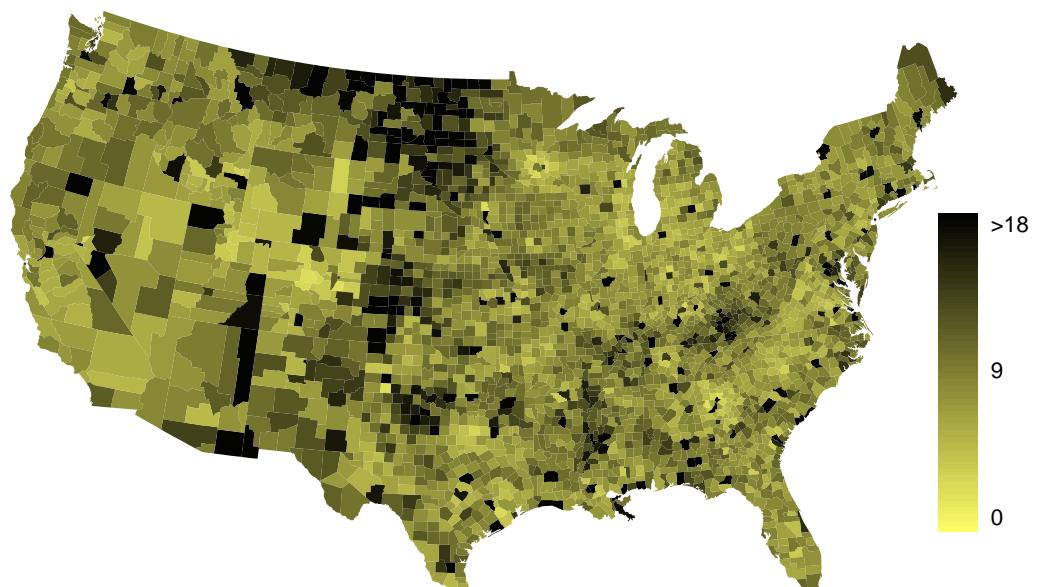
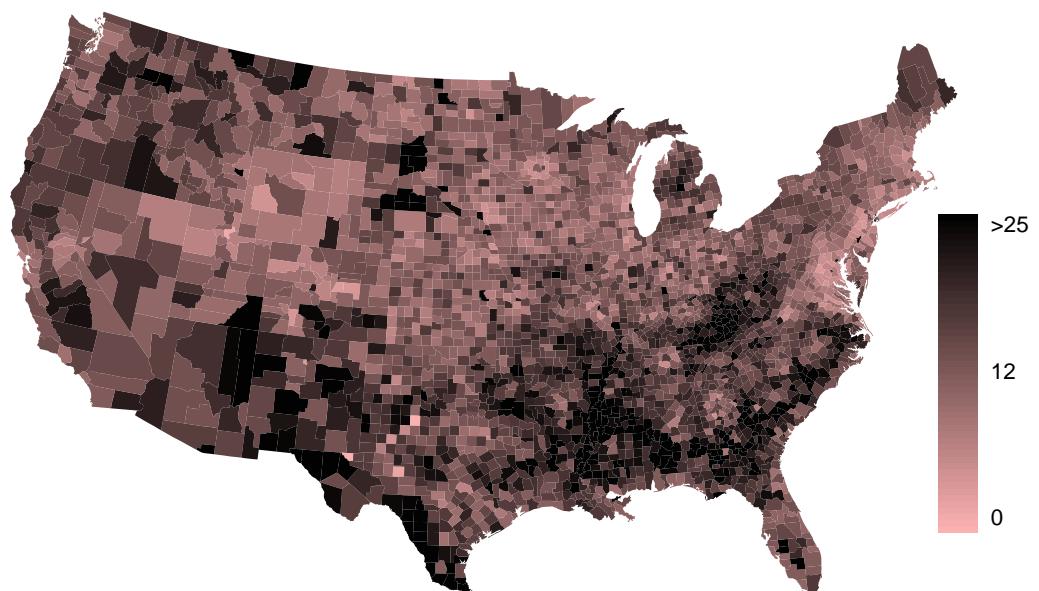


그림 1.30: (a) 전자우편 50개에 대한 `num_char` 변수에 대한 `line_breaks` 변수 산점도. (b) 동일한 데이터의 산점도, 하지만 각 변수를 로그 변환했다.

접했을 때, **강도 그림**(intensity map)을 사용해서 데이터를 매핑한다. 여기서 색깔을 사용해서 변수의 높고 낮은 값을 보여준다. 그림 1.31 와 1.32는 일인당 연방정부지출(`fed_spend`)과 퍼센트로 빈곤율(`poverty`), 퍼센트로 자가소유비율(`homeownership`), 중위수 가계소득(`med_income`)을 보여주고 있다. 컬러키(color key)는 어떤 색이 어떤 값에 대응되는지 나타낸다. 강도 지도는 일반적으로 특정 군에 정확한 값을 얻는데는 그다지 도움이 되지 않지만, 지리적 경향성을 살펴보고 흥미로운 연구문제를 뽑아내는데는 유용한 것에 주목한다.



(a)



(b)

그림 1.31: (a) 연방정부지출 지도 (단위: 인당 달러). (b) 빈곤율 강도 지도 (퍼센트).

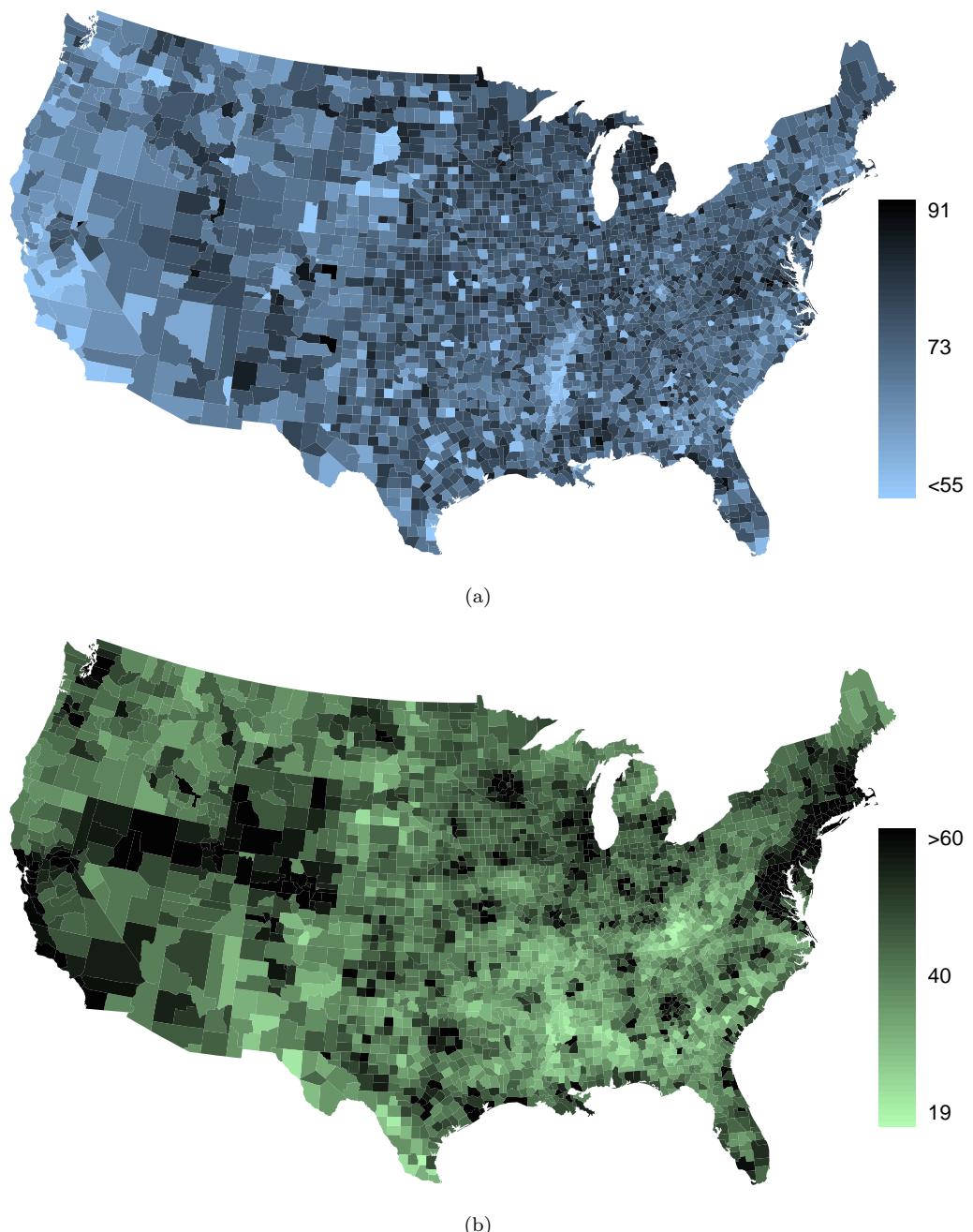


그림 1.32: (a) 자가소유비율 강도 지도 (퍼센트). (b) 중위수 가계소득 강도 지도 (\$1000 달러).

- **Example 1.37** `fed_spend` 와 `poverty` 강도 지도에서 어떤 흥미로운 특징이 명확히 드러나는가?

---

정부지출금 강도 지도는 다코타주와 캐나다 국경을 접한 중서부 지역을 따라 상당한 정부지출을 보여주고 있는데, 이 지역에 오일붐과 관련되어 있다. 또 다른 정부지출금 지역이 일부 있는데 예를 들어 동부 유타와 아리조나에 수직으로 난 조각과 콜로라도, 네브라스카, 캔사스가 만나는 지역이다. 다른 인접지역과 비교해서 상당한 정부지출금을 갖는 특정한 패턴이 없는 시군도 있다. 만약 일인당 \$18 달러로 연방정부지출금 상한을 두지 않았다면, 실제로 상당히 높은 연방정부지출금을 일부 시군에서 사용하는 반면에 인근 시군에서는 거의 연방정부지출금을 사용하지 않는 것을 볼 수 있다. 이러한 지출이 많은 시군에는 군부대, 대형 정부계약 맺은 회사, 혹은 많은 공무원을 둔 정부기관시설이 포함될 수 있다.

빈곤율은 분명히 몇몇 장소에서 더 높다. 주목할 점은 텍사스 남서쪽 경계지역에서처럼, 최남동부 지역에 높은 빈곤율을 보인다. 높은 연방정부지출에서도 표기되었듯이, 동부 유타와 아리조나 수직으로 난 조각지역도 또한 높은 빈곤율을 보인다 (일반적으로 두변수 사이에 거의 대응관계가 보이지 않지만). 높은 빈곤율은 뉴올리昂스 약간 북쪽 미시시피 범람 지역과 컨터키와 웨스트 버지니아 상당지역이 눈에 띄게 나타난다.

- **Guided Practice 1.38** 그림 1.32(b)에서 `med_income` 강도 지도에서 어떤 흥미로운 특징이 눈에 확실히 보이는가?<sup>40</sup>

---

<sup>40</sup>주목: 정답은 다양할 수 있다. 높은 수입과 대도시 지역 사이에 강한 상응이 있다. 친숙한 대형 도시를 찾고 지도에 위치한 곳이 어두운 색으로 표시되어 있는지 확인해보라.

## 7 범주형 데이터 고려하기

숫자형 데이터와 마찬가지로, 범주형 데이터도 구조화해서 분석될 수 있다. 이번 절에는 이 책 전반적으로 사용되는 범주형 데이터에 대한 기본 도구와 표가 소개된다. `email150` 데이터셋은 `email`로 불리는 좀더 커다란 전자우편 데이터에서 추출한 표본이다. 좀더 커다란 데이터셋에는 3,921개 전자우편에 대한 정보가 포함되어 있다. 이번 절에서 전자우편에 숫자의 존재, 크거나 작은 사실이 스팸과 정상 전자우편을 분류하는데 유용한 값을 제공하는지 조사한다.

### 7.1 분할표(Contingency tables)와 막대그림

표 1.33에는 요약된 두 변수가 있다: `spam`, `number`. `number` 변수는 범주형 변수로 전자우편이 숫자가 없거나, 작은 숫자 (백만 이하), 최소 큰 숫자 하나 (백만 혹은 그이상)를 갖는지 기술하는 범주형 변수임을 상기한다. 이러한 방식으로 두 범주형 변수에 대한 데이터를 요약하는 표를 **분할표**(contingency table)라고 부른다. 표에 각 값은 특정 변수결과 조합이 몇번 발생했는지 나타낸다. 예를 들어, 값 149는 데이터셋에서 스팸이며 전자우편에 숫자가 하나도 없는 전자우편 갯수에 상응한다. 행과 열 총계도 포함되었다. **행 총계**는 각 행에 걸친 전체 갯수에 대한 정보를 제공한다 (즉,  $149 + 168 + 50 = 367$ ). **열 총계**는 각 열에 걸친 전체 갯수가 된다.

단일 변수에 대한 표는 **빈도표**(frequency table)라고 부른다. 표 1.34는 `number` 변수에 대한 빈도표다. 만약 갯수를 퍼센트 혹은 비율로 교체하면 표는 **상대도수표**(relative frequency table)로 불린다.

		number			
		none	small	big	Total
spam	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

표 1.33: `spam` 와 `number`에 대한 분할표.

	none	small	big	Total
	549	2827	545	3921

표 1.34: `number` 변수에 대한 빈도표.

막대그림이 단일 범주형 변수를 시각화하는 일반적인 방식이다. 그림 1.35 왼쪽 패널에 `number` 변수에 대한 **막대그림**이 나와 있다. 오른쪽 패널에는 갯수를 비율로 변환해서 (즉, `none`에 대해  $549/3921 = 0.140$  ), 각 수준별(즉, 각 범주별)로 관측점의 비율을 보여주고 있다.

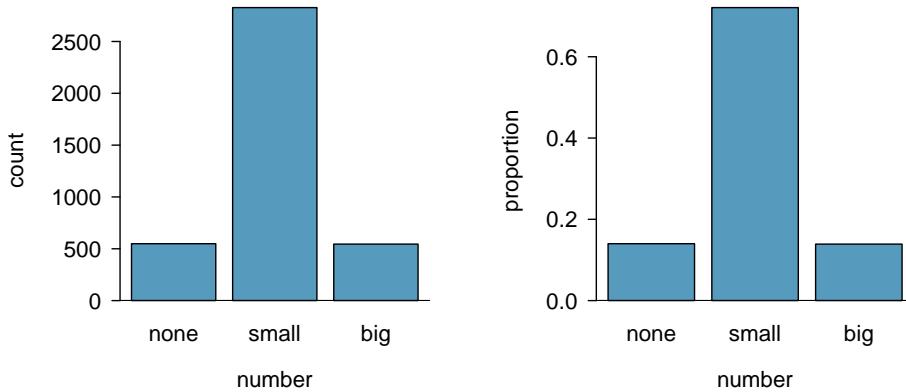


그림 1.35: `number`에 대한 막대그림 두개. 왼쪽 패널은 갯수, 오른쪽 패널은 각 집단에 대한 비율을 보여준다.

	none	small	big	소계
spam	$149/367 = 0.406$	$168/367 = 0.458$	$50/367 = 0.136$	1.000
not spam	$400/3554 = 0.113$	$2657/3554 = 0.748$	$495/3554 = 0.139$	1.000
소계	$549/3921 = 0.140$	$2827/3921 = 0.721$	$545/3921 = 0.139$	1.000

표 1.36: 변수 `spam` 과 `number`에 대한 행비율 정보를 갖는 분할표.

## 7.2 행과 열 비율

표 1.36에는 표 1.33에 대한 행비율 정보가 나와있다. 행비율(row proportions)은 갯수를 각 행 소계로 나누어서 계산된다. `spam` 과 `none`이 교차하는 값 149가  $149/367 = 0.406$ 으로 대체되었다, 즉 149를 행 소계 367로 나누었다. 그래서 0.406은 무엇을 나타내는가? 이 값은 표본에서 어떤 숫자도 갖지 않는 스팸 전자우편 비율에 상응된다.

열 비율의 분할표도 비슷한 방식으로 계산된다, 여기서 열 비율(column proportion)은 갯수를 상응하는 열 소계로 나눠서 계산된다. 표 1.37가 그런 표를 보여주고 있다. 여기서 값 0.271은 숫자가 없는 전자우편의 27.1%가 스팸임을 나타낸다. 스팸율이 작은 숫자(5.9%)만 갖거나 큰 숫자(9.2%)를 갖는 전자우편과 비교해서 훨씬 더 높다. 이러한 스팸율이 `number` 세가지 수준(`none`, `small`, `big`)마다 다르기 때문에, `spam` 과 `number` 변수가 연관되었다는 증거가 된다.

	none	small	big	소계
spam	$149/549 = 0.271$	$168/2827 = 0.059$	$50/545 = 0.092$	$367/3921 = 0.094$
not spam	$400/549 = 0.729$	$2659/2827 = 0.941$	$495/545 = 0.908$	$3684/3921 = 0.906$
소계	1.000	1.000	1.000	1.000

표 1.37: 변수 `spam` 과 `number`에 대한 열비율 정보를 갖는 분할표.

행비율을 사용해서 표 1.36에서 `spam` 과 `number` 변수 사이에 관계를 점검할 수 있다. 행비율을 비교할 때, 열 아래로 내려다 보고 숫자가 없는, 작은 숫자, 큰 숫자를 갖는 전자우편

비율이 `spam` 과 `not spam`에 따라 달라지는지 살펴본다.

- Ⓐ **Guided Practice 1.39** 표 1.36에서 0.458은 무엇을 나타내는가? 표 1.37에서 0.059는 무엇을 나타내는가?<sup>41</sup>

- Ⓑ **Guided Practice 1.40** 표 1.36에서 `not spam` 과 `big`이 교차하는 값 0.139는 무엇인가? 표 1.37에서 0.908은 무엇을 나타내는가?<sup>42</sup>

- Ⓒ **Example 1.41** 데이터 과학자는 통계량을 사용해서 인입되는 전자우편 메시지에서 스팸을 걸러낸다. 전자우편의 특정 문자에 주목함으로써 전자우편 일부를 스팸 혹은 정상 메일로 높은 정확도를 갖고 분류할 수도 있다. 이러한 특성 중 하나가 전자우편이 숫자가 없거나, 작은 단위 숫자, 혹은 큰 단위 숫자를 포함하는지 여부다. 또 다른 특성은 전자우편이 HTML 콘텐츠를 포함했느냐 여부다. `email` 데이터셋으로부터 `spam`과 `format` 변수의 분할표가 표 1.38에 나와있다. HTML 전자우편은 예를 들어 짧은 텍스트 같은 특수한 서식 기능을 갖는 전자우편이다. 표 1.38에서 스팸 혹은 정상 메일을 분류(행비율 혹은 열비율)하는데 어느 쪽이 더 도움이 될까?

이 문제에 관련있는 사람은 각 전자우편 서식 변화에 스팸비율이 얼마나 변하는지 관심 있다. 이것은 열비율에 대응된다: 일반 텍스트 스팸비율과 HTML 메일에 스팸비율.

만약 열비율을 만들어 내면, HTML 전자우편( $158/2726 = 5.8\%$ )보다 일반 텍스트 전자우편( $209/1195 = 17.5\%$ )의 분율이 더 높음을 볼 수 있다. 하지만, 이 정보를 많은 다른 특성, 예를 들어 `number`와 다른 변수와 주의깊이 결합하게 될 때, 전자우편을 스팸과 일반메일로 분류할 가능성이 높게 된다. ?? 장에서 다룰 주제다.

	텍스트	HTML	총계
스팸	209	158	367
일반 메일	986	2568	3554
총계	1195	2726	3921

표 1.38: `spam` 와 `format` 변수에 대한 분할표.

행비율과 열비율이 동치되지 않다는 것을 예제 1.41가 지적하고 있다. 한 형태로 표를 정하기 전에, 각각을 고려해서 가장 유용한 표가 구축되었는지 확실히 하는 것이 중요하다.

- Ⓐ **Guided Practice 1.42** 다시 표 1.36 와 1.37을 살펴보자. `number` 변수를 사용해서 스팸 전자우편을 식별하려는 사람에게 어느 표가 더 유용할까?<sup>43</sup>

<sup>41</sup>0.458은 작은 숫자를 갖는 스팸 전자우편 분율을 나타낸다. 0.059는 스팸인 작은 숫자를 갖는 전자우편 분율을 나타낸다.

<sup>42</sup> 0.139는 큰 숫자를 갖는 스팸이 아닌 전자우편 분율을 나타낸다. 0.908은 스팸이 아닌 큰 숫자를 갖는 전자우편 분율이다.

<sup>43</sup> 표 1.37에 열비율 정보가 아마도 가장 유용할 것이다. 왜냐하면 작은 숫자를 갖는 전자우편의 약 5.9%(상대적으로 드문 경우)가 스팸 전자우편이라는 것을 더 쉽게 볼 수 있게 한다. 추가로 숫자가 없는 전자우편의 약 27.1%가 스팸이고, 큰 숫자를 갖는 전자우편의 9.2%가 스팸이라는 것도 알 수 있다.

### 7.3 조각 막대그래프와 모자이크 그래프

행비율 혹은 열비율을 갖는 분할표는 특히 두 범주형 변수가 어떻게 연관되었는지 조사할 때 유용하다. 조각 막대그래프와 모자이크 그래프는 이러한 분할표 정보를 시각화하는 방법을 제공한다.

**조각 막대그래프**는 분할표 정보의 시각적 표현이다. 예를 들어, 표 1.37을 나타내는 조각 막대그래프가 그림 1.39(a)에 나와 있다. 여기서 varnumber 변수를 사용해서 막대그래프를 먼저 생성했다. 그리고 나서 spam 변수 수준에 따라 집단별로 나누었다. 표 1.37의 열비율 정보가 그림 1.39(b)에 표준화된 조각 막대그래프로 옮겨졌다. number 변수 각 수준별로 스팸 전자우편 분율을 유용하게 시각화했다.

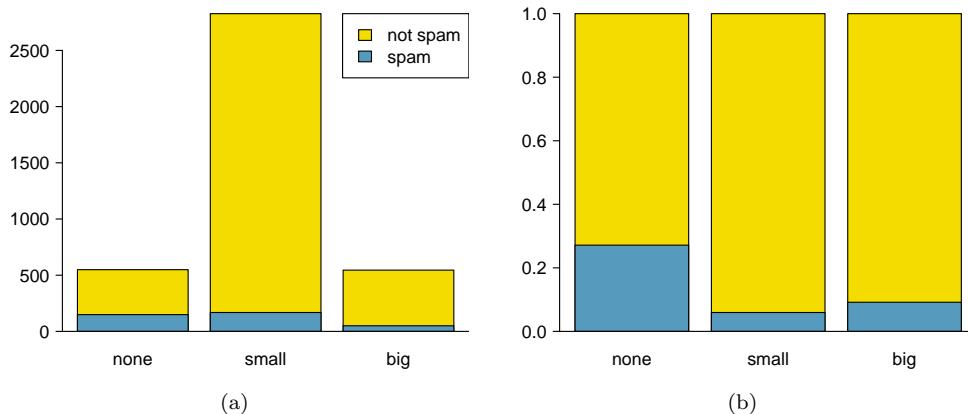


그림 1.39: (a) 전자우편에서 발견된 숫자에 대한 조각 막대그래프로, 갯수는 spam 여부에 따라 더 나누었다. (b) 그림 (a)의 표준화된 버전.

#### ● Example 1.43 양쪽 조각 막대그래프를 면밀히 살펴보라. 어느 것이 더 유용한가?

그림 1.39(a)에 더 많은 정보가 포함되어 있지만, 그림 1.39(b)은 정보를 좀더 명확히 제시한다. 두번째 그래프는 숫자가 없는 전자우편이 상대적으로 스팸 전자우편율이 더 높다는 것을 명확히 한다 – 약 27%. 다른 한편으로, 작은 숫자나 큰 숫자를 갖는 전자우편은 10% 미만이 스팸이다.

스팸 비율이 그림 1.39(b)에서 나타나듯이 집단에 따라 변하기 때문에, 변수가 종속된다고 결론 내릴 수 있는데, 분할표 비율을 사용해서 분간할 수 있었던 것이다. none 와 big 집단은 small 집단과 비교해서 상대적으로 관측점이 적기 때문에, 연관을 그림 1.39(a)에서 알아채기가 더 어렵다.

일부 다른 경우에, 표준화되지 않은 조각 막대그래프가 중요한 정보를 의사소통하는데 더 유용하다. 특정 조각 막대그래프를 결정하기 전에, 표준화된 형태와 표준화되지 않은 형태를 모두 생성하고 어느 것이 해당 데이터 특징을 의사소통하는데 더 효과적인 판단하라.

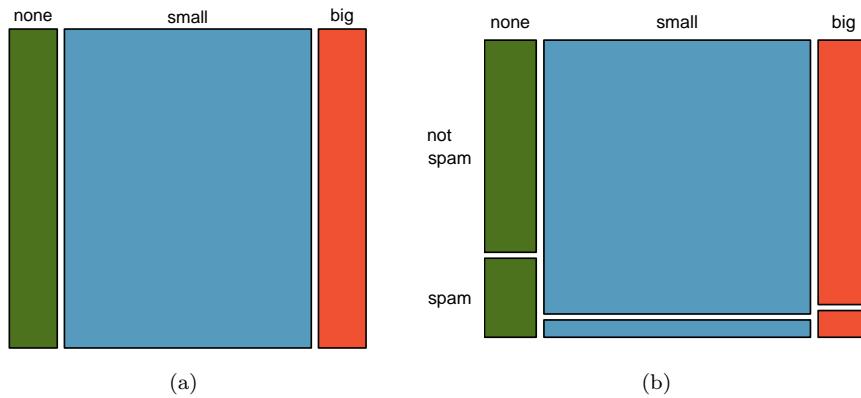


그림 1.40: `number` 변수에 대한 일변량 모자이크 그래프와 `number` 와 `spam` 변수에 대한 이변량 모자이크 그래프.

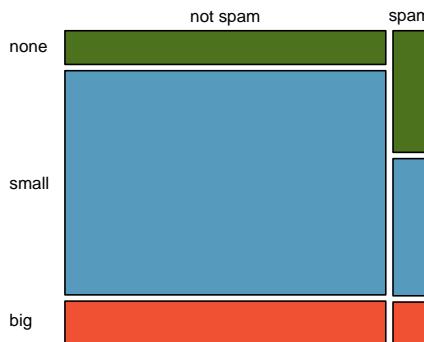


그림 1.41: `spam` 과 `not spam`으로 전자우편을 나눈 다음에 `number` 변수로 전자우편을 분류한 모자이크 그래프.

**모자이크 그래프(mosaic plot)**는 변수 하나에 대한 막대그래프 혹은 두 변수에 대한 조각 막대그래프와 유사하게 분할표 정보를 시각적으로 표현합니다. 그림 1.40(a)에 `number` 변수에 대한 모자이크 그래프가 나와 있다. 열 각각은 `number` 변수의 수준을 나타내고, 열 넓이는 각 숫자 유형에 대한 전자우편 비율에 대응된다. 예를 들어, 작은 숫자를 갖는 전자우편보다 숫자가 없는 전자우편이 더 적다. 그래서 숫자가 없는 전자우편 열이 더 가늘다. 일반적으로, 모자이크 그래프는 상자 면적을 사용해서 상자가 표현하는 관측점 갯수를 나타낸다.

일변량 모자이크 그래프를 `spam` 변수를 사용해서 그림 1.40(b)에서 보듯이 더 쪼갠다. 각 범주별로 스팸 전자우편 분율에 따라 비례해서 각 열을 쪼갠다. 예를 들어, 단지 적은 숫자만 갖는 전자우편을 대표하는 두번째 열을 스팸(아래)과 정상(위) 전자우편으로 나눈다. 또 다른 예제로, 세번째 열 하단은 큰 숫자를 갖는 스팸 전자우편을 나타내고, 세번째 열의 위쪽 부분은 큰 숫자를 갖는 정상 우편을 나타낸다. 모자이크 그래프를 사용해서 `spam` 과 `number` 변수가 연관된 것을 볼 수 있다. 왜냐하면, 일부 칼럼이 다른 칼럼보다 수직 위치가 다르게 나뉘기 때문이다. 이것은 표준화된 조각 막대그래프에서 연관성을 검사하는데 사용한 동일한

기법이다.

유사한 방식으로, 그림 1.41에 도시된 바와 같이, 표 1.33의 행비율을 나타내는 모자이크 그래프를 그릴 수 있다. 하지만, `number` 변수의 각 범주마다 스팸 분율을 고려하는 것이 이번 응용에는 더 많은 통찰력을 전달하기 때문에 저자는 그림 1.40(b) 표현을 선호한다.

## 7.4 이 책에서 보게 되는 유일한 파이그림(Pie Chart)

파이그림이 잘 알려졌지만, 일반적으로 데이터 분석에서 다른 그림만큼 유용하지는 않다. **파이그림(pie chart)**이 막대그림과 함께 그림 1.42에 도시되어 있다. 일반적으로 막대그림보다 파이그림에 집단 크기를 비교하기가 더 어렵다. 특히, 범주가 거의 동일한 갯수나 비율을 갖을 때 더욱 그렇다. `none` 과 `big` 범주의 경우에 차이가 아주 미세해서 어떤 그림으로도 집단 크기에 차이를 구별할 수 거의 없다!

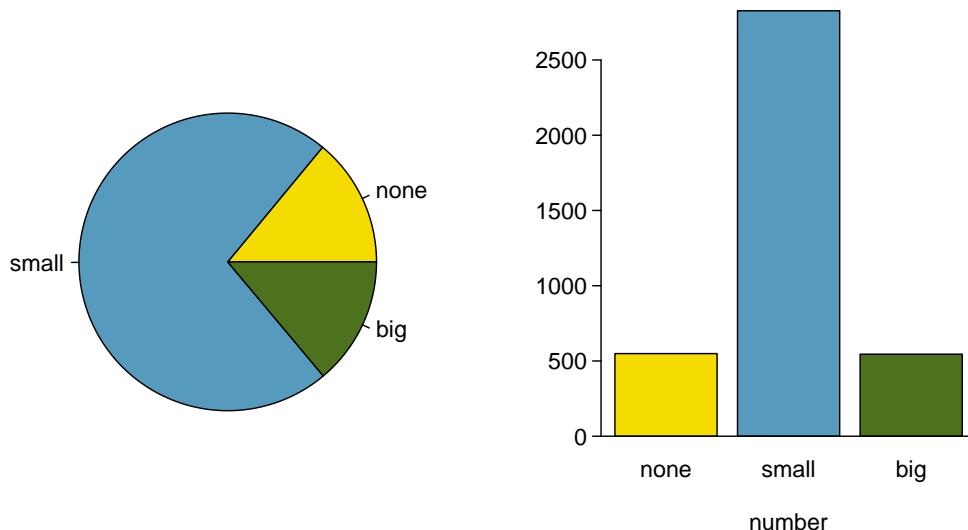


그림 1.42: `email` 데이터셋에 대한 `number` 변수를 파이그림과 막대그림으로 표현.

## 7.5 집단별로 수치 데이터를 비교하기

집단별로 수치 데이터를 검사해서, 좀 더 흥미로운 조사내용 중 일부를 자세히 바라볼 수 있다. 여기에 요구되는 방법은 정말 새로운 것은 아니다. 필요한 전부는 각 집단별로 수치그림을 그리는 것이다. 여기서 두가지 편리한 방법을 소개한다: 병행 상자그림(side-by-side box plot)과 공히스토그램(hollow histogram).

`county` 데이터셋을 다시 살펴보고 2000년부터 2010년까지 인구증가된 시군과 그렇지 않는 시군에 대한 중위수 가구소득을 비교한다. 여기서 인과적인 연결을 맺고 싶지만, 관측 데이터로 인과관계를 맺는 해석이 정당화될 수 없다는 사실을 기억하라.

인구 증가						인구 무증가		
41.2	33.1	30.4	37.3	79.1	34.5	40.3	33.5	34.8
22.9	39.9	31.4	45.1	50.6	59.4	29.5	31.8	41.3
47.9	36.4	42.2	43.2	31.8	36.9	28	39.1	42.8
50.1	27.3	37.5	53.5	26.1	57.2	38.1	39.5	22.3
57.4	42.6	40.6	48.8	28.1	29.4	43.3	37.5	47.1
43.8	26	33.8	35.7	38.5	42.3	43.7	36.7	36
41.3	40.5	68.3	31	46.7	30.5	35.8	38.7	39.8
68.3	48.3	38.7	62	37.6	32.2	46	42.3	48.2
42.6	53.6	50.7	35.1	30.6	56.8	38.6	31.9	31.1
66.4	41.4	34.3	38.9	37.3	41.7	37.6	29.3	30.1
51.9	83.3	46.3	48.4	40.8	42.6	57.5	32.6	31.1
44.5	34	48.7	45.2	34.7	32.2	46.2	26.5	40.1
39.4	38.6	40	57.3	45.2	33.1	38.4	46.7	25.9
43.8	71.7	45.1	32.2	63.3	54.7	36.4	41.5	45.7
71.3	36.3	36.4	41	37	66.7	39.7	37	37.7
50.2	45.8	45.7	60.2	53.1		21.4	29.3	50.1
35.8	40.4	51.5	66.4	36.1		43.6	39.8	

표 1.43: 상기 표에서, 2000-2010에 걸쳐 인구가 증가한 임의 표본 100개 시군에 대한 가계소득 중위수 (\$1000 달러)가 원편에, 인구가 증가하지 않은 임의 표본 50개 시군에 대한 가계소득 중위수가 오른편에 나타나 있다.

2000년부터 2010년 사이 인구가 증가한 시군이 2,041 개 있고, 1,099 개 시군은 증가가 없다 (하나를 제외하고 모두 감소). 일부 원데이터에 대한 감을 잡을 수 있도록, 첫번째 집단에서 임의 표본으로 100개 시군을 뽑고, 두번째 집단에서 50개 시군을 뽑아 표 1.43로 나타냈다.

**평행 상자그림**(side-by-side box plot)은 집단을 비교하는데 사용되는 전통적인 도구다. 그림 1.44 원편에 예제가 나와 있는데, 상자그림이 두개 있는데 각 상자그림이 각 집단을 나타낸다. 하나의 원도우에 동일한 척도로 두 상자그림을 위치시킨다.

또 다른 유용한 그리기 방법으로 **공히스토그램**(hollow histograms)이 있다. 이를 사용해서 집단별 수치 데이터를 비교한다. 그림 1.44 오른쪽에 도시되어 있듯이, 동일한 그림 위에 각 집단의 히스토그램 윤곽으로 표현된다.

④ **Guided Practice 1.44** 그림 1.44을 사용해서 두 집단 간 시군 소득을 비교하시오. 각 집단별 대략적 중심에 대해 알아낸 것이 무엇인가요? 집단간 변동성에 관해서 알아낸 것은 무엇인가요? 형상이 상대적으로 집단간에 일관성이 있나요? 각 집단에 대해 얼마나 많은 눈에 띄는 모드가 있나요?<sup>44</sup>

⑤ **Guided Practice 1.45** 그림 1.44에서 각 그림의 어떤 구성요소가 가장 유용하다고

<sup>44</sup> 정답은 약간 다를 수 있다. 인구증가된 시군(약 \$45,000 중위수 소득)이 증가가 없는 시군(약 \$40,000 중위수 소득)과 비교해서 더 높은 소득을 갖는 경향이 있다. 변동성은 인구증가 집단에서 다소 더 크다. 변동성은 IQR에서 확인한데 증가 집단에서 약 50% 더 크다. 두 분포 모두 우측으로 일부 기울어진 것과 일봉을 나타내고 있다. 증가 없는 집단에서 약 \$60,000 달러 근처에 위치에 맞지 않아 보이는 두번째 작은 융기가 있고, 공히스토그램에서도 시각적으로 확인된다. (데이터셋을 살펴보면, 15 시군중에 8개 시군이 알라스카주와 텍사스주에 있다는 것을 알 수 있다.) 이러한 큰규모 데이터셋을 사용할 때 많은 관측점이 상자수염 범위를 넘어 있을 것으로 예상하지만, 많은 관측점이 각 집단에서 중위수보다 훨씬 멀리 떨어 있음을 상자그림이 나타낸다.

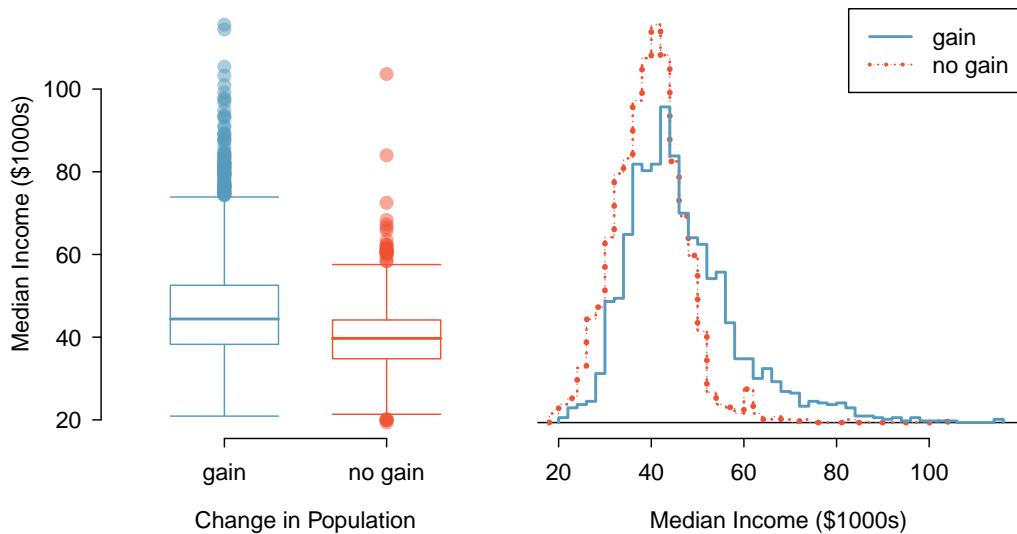


그림 1.44: 변수 `med_income` 가계소득 중위수에 대한 평행 상자그림(왼쪽)과 공히스토그램(오른쪽). 여기서 2000년부터 2010년 사이 인구가 증가 혹은 감소에 따라 시군을 쪼갰다. 소득 데이터는 2006년과 2010년 사이 수집됐다.

생각합니까?<sup>45</sup>

## 8 사례 연구: 성 차별 🎥 (특별 주제)

- **Example 1.46** 교수님이 교실 학생을 두 집단으로 쪼갠다고 가정하자: 왼편에 있는 학생과 오른편에 있는 학생. 각각  $\hat{p}_L$  과  $\hat{p}_R$  가 애플 제품을 소유하고 있는 학생비율을 나타낸다면, 만약  $\hat{p}_L$  이  $\hat{p}_R$  ?

---

와 정확하게 같다면 놀랄 일인가? 비율이 아마도 두 집단 모두 가까울 것이지만, 정확하게 같은 것은 매우 이례적일 것이다. 확률적 이유로 아마도 작은 차이를 관측할 것이다.

- **Guided Practice 1.47** 교실에 학생이 앉아 있는 위치가 왼편이거나 오른편이거나에 따라 애플 제품 소유 여부가 연관되어 있다고 생각하지 않는다면, 이 두 변수 사이 관계에 대해서 어떤 가정을 할까요?<sup>46</sup>

<sup>45</sup> 정답은 다를 수 있다. 병행 상자그림은 중심과 편차를 비교하는데 특히 유용한 반면, 공히스토그램은 분포 형상, 기울기, 집단 이상(anomaly)을 확인하는데 좀 더 유용하다.

<sup>46</sup> 두 변수는 독립적이라는 가정을 하고 있다.

## 8.1 데이터 내부 변동성

은행 내부 인사결정 맥락에서 1970년대 성차별을 조사하는 연구를 고려해보자.<sup>47</sup> 답하고자 하는 연구 질문은 다음과 같다. “여성이 승진 심사에서 불공정하게 남성 매니저에 의해서 차별을 받는가?”

본 연구에 참가자는 1972년 North Carolina 대학 경영자 과정에 참석한 남성 은행 감독자 48명이다. 은행 인사담당 임원 역할을 수행하도록 했다. 인사정보를 제공해서 해당 직원을 은행지점 관리자로 승진시켜야 하는지 판단하도록 했다. 참가자에게 주어진 파일의 절반은 대상자가 남성, 다른 절반은 여성이라는 점만 제외하면 동일했다. 해당 파일은 무작위로 실험 개체, 즉 연구참여자에게 배정되었다.

- Guided Practice 1.48 본연구는 관측연구인가 혹은 실험인가? 연구 유형이 결과로부터 추론한 것에 대해 어떤 함의를 갖는가?<sup>48</sup>

각 감독자마다 배정된 파일에 관련된 성별과 승진결정을 기록했다. 표 1.45에 요약된 연구결과를 사용해서, 여성이 불공정하게 승진결정에 차별을 받았는지 평가한다. 본 연구에서, 남성보다 여성 중 더 적은 비율만이 승진했다 (0.583 vs. 0.875). 하지만, 이러한 차이가 여성이 불공정하게 차별을 받았다는 설득력있는 증거가 되는지는 불명확하다.

		decision		합계
		승진	탈락	
gender	남성	21	3	24
	여성	14	10	24
	총계	35	13	48

표 1.45: 성차별 연구 요약 결과.

- Example 1.49 종종 통계학자를 호출해서 증거의 강도를 평가한다. 해당 연구에서 남성과 여성에 대한 승진율을 살펴볼 때, 데이터가 실제 차이에 대한 설득력있는 증거를 제시하는지 판단하려고 할 때 무엇이 떠오르나요?

관측된 승진율 (여성 58.3% vs 남성 87.5%) 정보가 승진결정에 여성에 대한 차별이 있을 수 있다는 것을 제시한다. 하지만, 관측된 차이가 차별을 나타내는지 혹은 우연적 요인에 의한 것인지에 대해서는 확신할 수 없다. 일반적으로, 표본 데이터에 일부 변동성이 있다. 그리고, 설사 승진율이 성별에 무관하다는 사실이 있음에도 불구하고, 표본 비율이 정확하게 같을 것으로 예상하지는 못한다.

예제 1.49를 통해 표본에서 관측된 결과가 밑에 있는 모집단 변수간 참된 관계를 완벽하게 반영하지 못할 수도 있다는 점이 상기된다. 표 1.45에는 남성집단보다 여성집단의 승진 건수가

<sup>47</sup>Rosen B and Jerdee T. 1974. Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology 59(1):9-14.

<sup>48</sup>무작위로 남성서류와 여성 서류를 개체에게 배정했다는 점에서 해당 연구는 실험이다. 실험이기 때문에, 승진 결정과 대상자의 성별 사이에 인과관계를 평가하는데 결과를 사용할 수 있다.

7개 적다고 나와있다 (승진율 차이가 29.2%,  $(\frac{21}{24} - \frac{14}{24} = 0.292)$ ). 이 차이가 상당하지만, 연구에 사용된 표본 크기가 작아서, 관측된 차이가 성차별을 나타내는지 혹은 단순히 확률적인 요인 때문인지 불명확하다. 두 경쟁하는 주장을  $H_0$  와  $H_A$  로 표식한다:

$H_0$ : **독립 모형.** 변수 gender 와 decision는 독립적이다. 두 변수는 관계가 없고, 승진된 여성비율과 남성비율 사이 관측된 차이, 29.2%는 우연 때문이다.

$H_A$ : **대안 모형.** 변수 gender 와 decision는 독립적이 아니다. 승진율 29.2%의 차이는 우연 때문이 아니고, 동일한 자격을 갖춘 여성이 남성보다 덜 승진될 것 같다.

변수 gender 와 decision가 연관되지 않았다고 말하는 독립 모형이 참이라면 어떤 의미 일까요? 이것이 의미하는 바는 은행원 각자는 서류철에 표기된 성별에 상관없이 승진 대상자를 승진하는 결정을 한다는 것이다. 즉, 승진 퍼센트에 나타난 차이는 파일이 무작위로 은행원에게 나눠졌고, 임의화(randomization) 과정을 통해서 상대적으로 큰 차이, 29.2%가 발생했다.

대안 모형을 고려해 보자: 은행원은 인사서류철에 어느 성별이 목록에 올라갔느냐에 영향을 받는다. 만약 이것이 사실이고, 특히 이러한 영향이 상당하다면, 남성 승진 대상자와 여성 승진 대상자의 승진율에 있어 일부 차이를 볼 것으로 예상된다. 만약 이러한 성 편향이 여성에 불리하게 작용한다면, 남성 인사 서류철에 비해 여성 인사 서류철에서 더 적은 승진 분율이 예상된다.

데이터가  $H_0$ 와 상당히 충돌되서 독립 모형을 합리적으로 받아들일 수 없는지를 평가함으로써, 두 경쟁하는 주장 중에서 하나를 선택한다. 만약 이것이 그런 경우이고, 데이터가  $H_A$  를 지지한다면, 독립 개념을 받아들이지 않고, 차별이 존재했다고 결론낸다.

## 8.2 연구 모의실험

표 1.45에 은행 감독자 35명이 승진 대상으로 추천되었고, 13명은 추천되지 않았다는 것이 나와있다. 이제, 은행원의 결정은 성별에 독립적이라고 가정하자. 그리고 나서, 만약 서류철을 다른 방식으로 임의적으로 배정해서 다시 실험을 수행한다면, 승진율 차이는 단지 확률적 변동 때문일 것이다. 실지로 이러한 임의화(randomization)를 수행할 수 있다. 즉, 만약 은행원 결정이 성별에 독립적이라면 무슨 일이 일어날지를 모의실험한다. 하지만, 서류철은 다르게 분배한다.

해당 모의실험(simulation)에서, 인사서류철 48개를 철저하게 뒤섞고, 각각 24개 male\_sim 와 24개 female\_sim로 표식해서 이를 두 더미에 분배해서 넣는다. 첫번째 더미에 35개 서류철이 분배되었는데, 승진을 추천한 지점 감독관 35명을 나타낸다. 두번째 더미는 서류철 13개가 있는데 승진에 추천되지 못한 대상자 13명을 나타낸다. 그리고 나서, 원데이터에 수행했던 것처럼, 결과를 표로 그려서, male\_sim 과 female\_sim에 대한 승진 분율을 결정한다. 이번 실험에서 서류철을 임의화한 것은 승진 결정과 독립으로, 두 분율에 어떤 차이도 완전히 우연 때문임을 의미한다. 표 1.46에 이런 모의실험 결과가 도시되어 있다.

		decision		총계
		승진	탈락	
gender_sim	male_sim	18	6	24
	female_sim	17	7	24
	총계	35	13	48

표 1.46: 모의실험 결과, 여기서 변수 male\_sim 와 female\_sim의 승진율에 대한 어떤 차이도 순전히 우연 때문이다.

⦿ Guided Practice 1.50 표 1.46에 두 모의실험 집단 사이에 승진율 차이는 얼마인가?

실제 집단에서 관측된 29.2%와 이 결과가 어떻게 비교되는가?<sup>49</sup>

### 8.3 독립성 검사

Guided Practice 1.50에서 독립모형을 놓고 가능한 차이(우연 때문에 차이)를 계산했다. 첫번째 모의실험에서 물리적으로 파일을 배분했지만, 컴퓨터를 사용해서 이러한 모의실험을 수행하는 것이 더 효율적이다. 컴퓨터로 모의실험을 반복해서, 우연으로 인한 또다른 차이값을 얻는다: -0.042. 또다른 모의실험은 0.208. 우연에서만 얻은 차이 분포를 나타낸다는 생각을 들때까지 충분히 모의실험을 반복한다. 그림 1.47에 모의실험 100번에서 나온 차이를 도식했다. 여기서 각 점은 승진에 추천된 남성과 여성 서류철 비율의 모의실험 차이를 나타낸다.

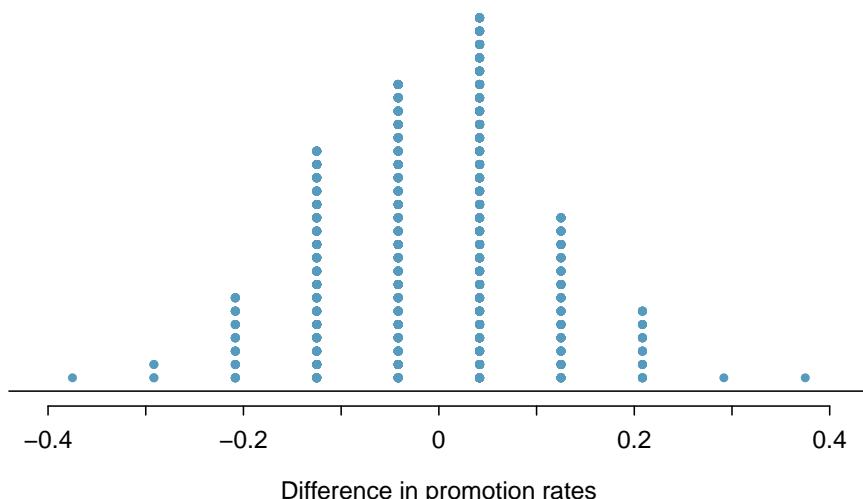


그림 1.47: 독립모형  $H_0$  아래서 생산된 모의실험 100개로부터 차이를 스택 점 그림으로 표현. 여기서 gender\_sim 와 decision는 독립이다. 모의실험 100 번 중 2번은 적어도 29.2% 차이를 보이는데, 연구에서 관측된 차이다.

상기 모의실험 차이 분포는 0 주위에 중심으로 몰려있음에 주목한다. 독립 모형이 참이고, 이러한 조건에서 차이는 0에 확률적 변동이 있다고 가정하고 차이를 모의실험했다. 일반적으로

<sup>49</sup>  $18/24 - 17/24 = 0.042$  즉, 약 4.2% 정도 남성에게 이익이 된다. 우연 때문에 이러한 차이는 실제 집단에서 관측된 차이보다 훨씬 더 작다.

차이가 정확하게 0이 되면 놀라게 된다: 종종 우연으로 차이는 0보다 크고, 다른 때는 0보다 작다.

- **Example 1.51** 표 1.47에 따르면 적어도 29.2% (0.292) 차이를 얼마나 자주 관측할까요? 자주, 때때로, 거의 없거나, 전혀 없거나?

---

그림 1.47에 따르면, 우연으로 인해 적어도 29.2% 차이는 약 2% 정도 발생할 것으로 보인다. 이러한 낮은 확률은 드문 사례임을 나타낸다.

29.2% 차이가 드문 사례라는 것은 연구 결과에 대한 두 가지 가능한 해석을 제시한다:

$H_0$  **독립 모형.** 성별은 승진결정에 영향을 주지 않고, 드물게만 발생하는 차이를 관측했다.

$H_A$  **대안 모형.** 성별은 승진결정에 영향이 있고, 관측한 것은 실제로 동일한 자격이 갖춘 여성이 승진결정에 차별을 받았기 때문이고, 이것이 29.2% 큰 차이를 설명한다.

모의실험에 근거해서 두 가지 선택옵션이 있다. (1) 연구결과 독립모형에 반하는 강력한 증거가 제시되지 못했다고 결정한다. 즉, 성차별이 있었다고 결정할 정도로 충분히 강력한 증거를 발견하지 못했다. (2)  $H_0$ 를 기각할 충분히 강력한 증거가 있다고 결정내고 성차별이 존재했다고 주장한다. 공식적인 연구를 수행할 때, 일반적으로 단지 드문 사례가 관측되었다는 개념은 기각된다.<sup>50</sup> 그래서, 이번 사례에서 대안 모형에 대해 독립 모형을 기각한다. 즉, 감독관에서 의한 여성에 대한 성차별의 강력한 증거를 데이터가 제시하고 있다고 결론낸다.

통계의 한 분야, 통계적 추론은 이러한 차이가 우연 때문인지를 평가하는 것이다. 통계적 추론에서, 통계학자는 어느 모형이 주어진 데이터에 가장 합리적인지를 평가한다. 드문 사례처럼, 오차가 발생하고, 틀린 모형을 선택할 수도 있다. 항상 올바르게 선택하지는 못하지만, 통계적 추론은 이러한 오차가 얼마나 자주 발생하는지 제어하고 평가할 수 있는 도구다. ??장에 모형선택 문제에 정형화된 안내가 나와 있다. 이러한 논의를 엄격하게 진행하는데 필요한 기초 확률과 이론을 다음 두 장에 걸쳐 다룰 것이다.

---

<sup>50</sup> 일반적으로 이러한 추론은 일화적 관측에 확장되지는 않는다. 우리들 각각은 매일 놀랍도록 드문 사건을 목격한다. 즉, 아마도 예측할 수 없는 사건. 하지만, 일화적 증거를 엄격하지 않은 환경에서는 거의 모든 것이 드문 사건처럼 보인다. 그래서 일상 생활에서 드문 사건을 찾는 생각은 신뢰할 수 없다. 예를 들어, 복권을 살펴보자: 역사상 가장 큰 당첨금(2012년 3월 30일)은 메가 번호 (23)을 갖는 (2, 4, 23, 38, 46) 번호로 1.76 억개 중 1의 확률이다. 하지만, 이 번호가 나왔다! 하지만, 무슨 번호가 나오던지, 번호 모두 놀랍도록 같은 드문 가능성을 갖는다. 즉, 관측한 어떤 번호 조합도 궁극적으로 믿을 수 없을 정도로 드문 경우다. 이러한 상황이 전형적인 일상생활이다: 그 자체로 가능한 각 사건은 믿을 수 없을 정도로 드물다. 하지만, 만약 모든 대안을 고려해보면, 이러한 결과도 또한 믿을 수 없을 정도로 드물다. 이러한 일화적 증거를 잘못 해석하지 않도록 주의해야 한다.

## 9 Exercises

### 9.1 Case study: using stents to prevent strokes

**1.1 Migraine and acupuncture.** A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.<sup>51</sup>

		Pain free		Total
		Yes	No	
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

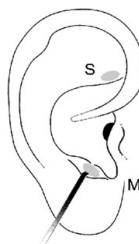


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

- (a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture? What percent in the control group?
- (b) At first glance, does acupuncture appear to be an effective treatment for migraines? Explain your reasoning.
- (c) Do the data provide convincing evidence that there is a real pain reduction for those patients in the treatment group? Or do you think that the observed difference might just be due to chance?

**1.2 Sinusitis and antibiotics.** Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period patients were asked if they experienced significant improvement in symptoms. The distribution of responses are summarized below.<sup>52</sup>

<sup>51</sup>G. Allais et al. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

<sup>52</sup>J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

		<i>Self-reported significant improvement in symptoms</i>		
		Yes	No	Total
<i>Group</i>	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

- (a) What percent of patients in the treatment group experienced a significant improvement in symptoms? What percent in the control group?
- (b) Based on your findings in part (a), which treatment appears to be more effective for sinusitis?
- (c) Do the data provide convincing evidence that there is a difference in the improvement rates of sinusitis symptoms? Or do you think that the observed difference might just be due to chance?

## 9.2 Data basics

**1.3 Air pollution and birth outcomes, study components.** Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter ( $PM_{10}$ ) in  $\mu g/m^3$ . Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient  $PM_{10}$  and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.<sup>53</sup>. Identify

- (a) the cases,
- (b) the variables and their types, and
- (c) the main research question

in this study.

**1.4 Buteyko method, study components.** The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.<sup>54</sup>. Identify

<sup>53</sup>B. Ritz et al. “Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993”. In: *Epidemiology* 11.5 (2000), pp. 502–511.

<sup>54</sup>J. McGowan. “Health Education: Does the Buteyko Institute Method make a difference?” In: *Thorax* 58 (2003).

- (a) the cases,
- (b) the variables and their types, and
- (c) the main research question

in this study.

**1.5 Cheaters, study components.** Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls.<sup>55</sup> Identify

- (a) the cases,
- (b) the variables and their types, and
- (c) the main research question

in this study.

---

<sup>55</sup>Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.

**1.6 Stealers, study components.** In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others.<sup>56</sup> Identify

- (a) the cases,
- (b) the variables and their types, and
- (c) the main research question

in this study.

**1.7 Fisher's irises.** Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.<sup>57</sup>

- (a) How many cases were included in the data?
- (b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.
- (c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).



Photo by Ryan Claussen  
(<http://flic.kr/p/6QTcuX>)  
CC BY-SA 2.0 license

**1.8 Smoking habits of UK residents.** A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that “£” stands for British Pounds Sterling, “cig” stands for cigarettes, and “N/A” refers to a missing component of the data.<sup>58</sup>

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
:	:	:	:	:	:	:	:
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

- (a) What does each row of the data matrix represent?

<sup>56</sup>P.K. Piff et al. “Higher social class predicts increased unethical behavior”. In: *Proceedings of the National Academy of Sciences* (2012).

<sup>57</sup>R.A Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. In: *Annals of Eugenics* 7 (1936), pp. 179–188.

<sup>58</sup>National STEM Centre, Large Datasets from stats4schools.

- (b) How many participants were included in the survey?
- (c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

### 9.3 Overview of data collection principles

**1.9 Air pollution and birth outcomes, scope of inference.** Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth.

- (a) Identify the population of interest and the sample in this study.
- (b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

**1.10 Cheaters, scope of inference.** Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

- (a) Identify the population of interest and the sample in this study.
- (b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

**1.11 Buteyko method, scope of inference.** Exercise 1.4 introduces a study on using the Buteyko shallow breathing technique to reduce asthma symptoms and improve quality of life. As part of this study 600 asthma patients aged 18-69 who relied on medication for asthma treatment were recruited and randomly assigned to two groups: one practiced the Buteyko method and the other did not. Those in the Buteyko group experienced, on average, a significant reduction in asthma symptoms and an improvement in quality of life.

- (a) Identify the population of interest and the sample in this study.
- (b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

**1.12 Stealers, scope of inference.** Exercise 1.6 introduces a study on the relationship between socio-economic class and unethical behavior. As part of this study 129 University of California Berkeley undergraduates were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others.

- (a) Identify the population of interest and the sample in this study.
- (b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

**1.13 Relaxing after work.** The 2010 General Social Survey asked the question, “After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?” to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.

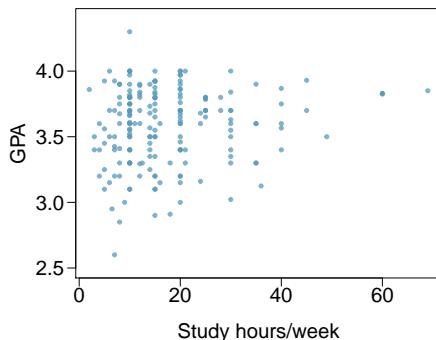
- (a) An American in the sample.
- (b) Number of hours spent relaxing after an average work day.
- (c) 1.65.
- (d) Average number of hours all Americans spend relaxing after an average work day.

**1.14 Cats on YouTube.** Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.

- (a) Percentage of all videos on YouTube that are cat videos.
- (b) 2%.
- (c) A video in your sample.
- (d) Whether or not a video is a cat video.

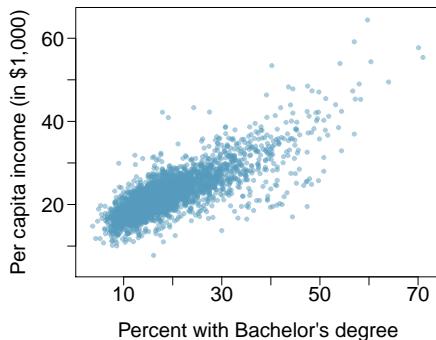
**1.15 GPA and study hours.** A survey was conducted on 193 Duke University undergraduates who took an introductory statistics course in 2012. Among many other questions, this survey asked them about their GPA, which can range between 0 and 4 points, and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.

- (a) What is the explanatory variable and what is the response variable?
- (b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
- (c) Is this an experiment or an observational study?
- (d) Can we conclude that studying longer hours leads to higher GPAs?



**1.16 Income and education in US counties.** The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor’s degree in 3,143 counties in the US in 2010.

- (a) What are the explanatory and response variables?
- (b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
- (c) Can we conclude that having a bachelor's degree increases one's income?



## 9.4 Observational studies and sampling strategies

**1.17 Course satisfaction across sections.** A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

- (a) What type of study is this?
- (b) Suggest a sampling strategy for carrying out this study.

**1.18 Housing proposal across dorms.** On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

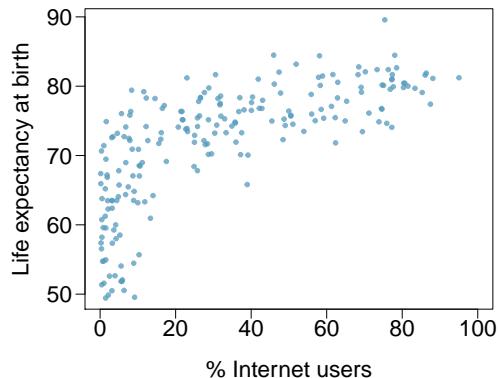
- (a) What type of study is this?
- (b) Suggest a sampling strategy for carrying out this study.

**1.19 Internet use and life expectancy.** The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.<sup>59</sup>

---

<sup>59</sup>CIA Factbook, Country Comparisons, 2014.

- (a) Describe the relationship between life expectancy and percentage of internet users.
- (b) What type of study is this?
- (c) State a possible confounding variable that might explain this relationship and describe its potential effect.



**1.20 Stressed out, Part I.** A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

- (a) What type of study is this?
- (b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?
- (c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

**1.21 Evaluate sampling methods.** A university wants to determine what fraction of its undergraduate student body support a new \$25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

- (a) Survey a simple random sample of 500 students.
- (b) Stratify students by their field of study, then sample 10% of students from each stratum.
- (c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

**1.22 Random digit dialing.** The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

**1.23 Haters are gonna hate, study confirms.** A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled men and women to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their dispositional attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively also tended to react negatively to it. Researcher concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."<sup>60</sup>

- (a) What are the cases?
- (b) What is (are) the response variable(s) in this study?
- (c) What is (are) the explanatory variable(s) in this study?
- (d) Does the study employ random sampling?
- (e) Is this an observational study or an experiment? Explain your reasoning.
- (f) Can we establish a causal link between the explanatory and response variables?
- (g) Can the results of the study be generalized to the population at large?

**1.24 Family size.** Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

**1.25 Flawed reasoning.** Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

- (a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

<sup>60</sup>Justin Hepler and Dolores Albarracín. "Attitudes without objects - Evidence for a dispositional attitude, its measurement, and its consequences". In: *Journal of personality and social psychology* 104.6 (2013), p. 1060.

- (b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

- (c) A orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

**1.26 City council survey.** A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Identify the sampling methods described below, and comment on whether or not you think they would be effective in this setting.

- (a) Randomly sample 50 households from the city.
- (b) Divide the city into neighborhoods, and sample 20 households from each neighborhood.
- (c) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and sample all households from those neighborhoods.
- (d) Divide the city into neighborhoods, randomly sample 10 neighborhoods, and then randomly sample 20 households from those neighborhoods.
- (e) Sample the 200 households closest to the city council offices.

**1.27 Sampling strategies.** A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

- (a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
- (b) He gives out the survey only to his friends, making sure each one of them fills out the survey.
- (c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.
- (d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

**1.28 Reading the paper.** Below are excerpts from two articles published in the *NY Times*:

- (a) An article titled *Risks: Smokers Found More Prone to Dementia* states the following:<sup>61</sup>  
“Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer’s disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs.”

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

- (b) Another article titled *The School Bully Is Sleepy* states the following:<sup>62</sup>  
“The University of Michigan study, collected survey data from parents on each child’s sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students

<sup>61</sup>R.C. Rabin. “Risks: Smokers Found More Prone to Dementia”. In: *New York Times* (2010).

<sup>62</sup>T. Parker-Pope. “The School Bully Is Sleepy”. In: *New York Times* (2011).

studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders.”

A friend of yours who read the article says, “The study shows that sleep disorders lead to bullying in school children.” Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

**1.29 Shyness on Facebook.** Given the anonymity afforded to individuals in online interactions, researchers hypothesized that shy individuals might have more favorable attitudes toward Facebook, and that shyness might be positively correlated with time spent on Facebook. They also hypothesized that shy individuals might have fewer Facebook “friends” as they tend to have fewer friends than non-shy individuals have in the offline world. 103 undergraduate students at an Ontario university were surveyed via online questionnaires. The study states “Participants were recruited through the university’s psychology participation pool. After indicating an interest in the study, participants were sent an e-mail containing the study’s URL.” Are the results of this study generalizable to the population of all Facebook users?<sup>63</sup>

## 9.5 Experiments

**1.30 Stressed out, Part II.** In a study evaluating the relationship between stress and muscle cramps half the subjects are randomly assigned to be exposed to increased stressed by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

- (a) What type of study is this?
- (b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

**1.31 Light and exam performance.** A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

- (a) What is the response variable?
- (b) What is the explanatory variable? What are its levels?
- (c) What is the blocking variable? What are its levels?

**1.32 Vitamin supplements.** In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.<sup>64</sup>

<sup>63</sup>E.S. Orr et al. “The influence of shyness on the use of Facebook in an undergraduate sample”. In: *CyberPsychology & Behavior* 12.3 (2009), pp. 337–340.

<sup>64</sup>C. Audera et al. “Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial”. In: *Medical Journal of Australia* 175.7 (2001), pp. 359–362.

- (a) Was this an experiment or an observational study? Why?

(b) What are the explanatory and response variables in this study?

(c) Were the patients blinded to their treatment?

(d) Was this study double-blind?

(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

**1.33 Light, noise, and exam performance.** A study is designed to test the effect of light level and noise level on exam performance of students. The researcher believes that light and noise levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The light treatments considered are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). The noise treatments considered are no noise, construction noise, and human chatter noise.

- (a) What type of study is this?
- (b) How many factors are considered in this study? Identify them, and describe their levels.
- (c) What is the role of the sex variable in this study?

**1.34 Music and learning.** You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

**1.35 Soda preference.** You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

**1.36 Exercise and mental health.** A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

- (a) What type of study is this?
- (b) What are the treatment and control groups in this study?
- (c) Does this study make use of blocking? If so, what is the blocking variable?
- (d) Does this study make use of blinding?
- (e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
- (f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

**1.37 Chia seeds and weight loss.** Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant

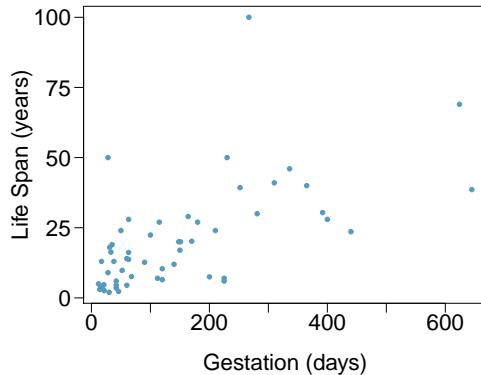
difference between the groups in appetite or weight loss.<sup>65</sup>

- (a) What type of study is this?
- (b) What are the experimental and control treatments in this study?
- (c) Has blocking been used in this study? If so, what is the blocking variable?
- (d) Has blinding been used in this study?
- (e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

## 9.6 Examining numerical data

**1.38 Mammal life spans.** Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.<sup>66</sup>

- (a) What type of an association is apparent between life span and length of gestation?
- (b) What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?
- (c) Are life span and length of gestation independent? Explain your reasoning.



**1.39 Associations.** Indicate which of the plots show a

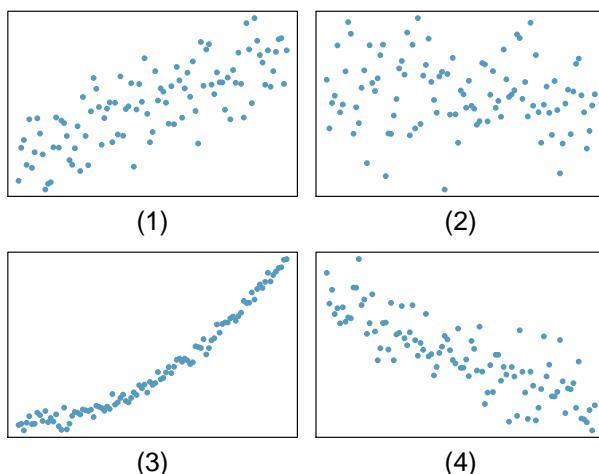
---

<sup>65</sup>D.C. Nieman et al. “Chia seed does not promote weight loss or alter disease risk factors in overweight adults”. In: *Nutrition Research* 29.6 (2009), pp. 414–418.

<sup>66</sup>T. Allison and D.V. Cicchetti. “Sleep in mammals: ecological and constitutional correlates”. In: *Arch. Hydrobiol* 75 (1975), p. 442.

- (a) positive association
- (b) negative association
- (c) no association

Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.



**1.40 Office productivity.** Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

**1.41 Reproducing bacteria.** Suppose that there is only sufficient space and nutrients to support one million bacterial cells in a petri dish. You place a few bacterial cells in this petri dish, allow them to reproduce freely, and record the number of bacterial cells in the dish over time. Sketch a plot representing the relationship between number of bacterial cells and time.

**1.42 Sleeping in college.** A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

**1.43 Parameters and statistics.** Identify which value represents the sample mean and which value represents the claimed population mean.

- (a) American households spent an average of about \$52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was \$58 per household.
- (b) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.

**1.44 Make-up exam.** In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

- (a) Does the new student's score increase or decrease the average score?
- (b) What is the new average?
- (c) Does the new student's score increase or decrease the standard deviation of the scores?

**1.45 Days off at a mining plant.** Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

**1.46 Medians and IQRs.** For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

- |                       |                            |
|-----------------------|----------------------------|
| (a) (1) 3, 5, 6, 7, 9 | (c) (1) 1, 2, 3, 4, 5      |
| (2) 3, 5, 6, 7, 20    | (2) 6, 7, 8, 9, 10         |
| (b) (1) 3, 5, 6, 7, 9 | (d) (1) 0, 10, 50, 60, 100 |
| (2) 3, 5, 8, 7, 9     | (2) 0, 100, 500, 600, 1000 |

**1.47 Means and SDs.** For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

- |                                       |                                 |
|---------------------------------------|---------------------------------|
| (a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13 | (c) (1) 0, 2, 4, 6, 8, 10       |
| (2) 3, 5, 5, 5, 8, 11, 11, 11, 20     | (2) 20, 22, 24, 26, 28, 30      |
| (b) (1) -20, 0, 0, 0, 15, 25, 30, 30  | (d) (1) 100, 200, 300, 400, 500 |
| (2) -40, 0, 0, 0, 15, 25, 30, 30      | (2) 0, 50, 300, 550, 600        |

**1.48 Stats scores.** Below are the final exam scores of twenty introductory statistics students.

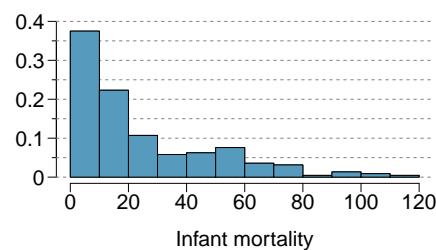
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

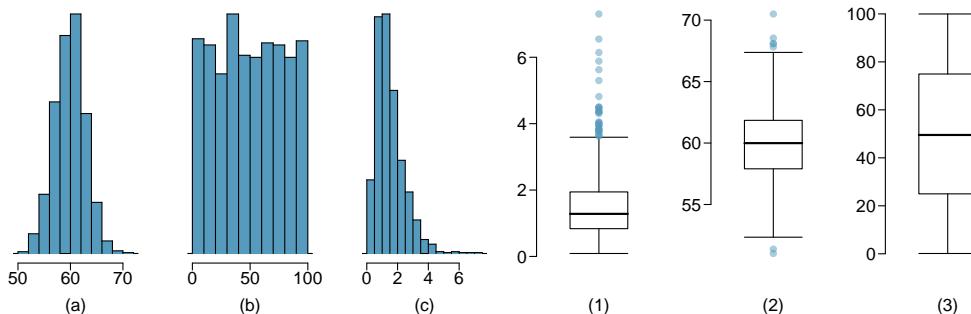
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

**1.49 Infant mortality.** The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.<sup>67</sup>

- (a) Estimate Q1, the median, and Q3 from the histogram.
- (b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

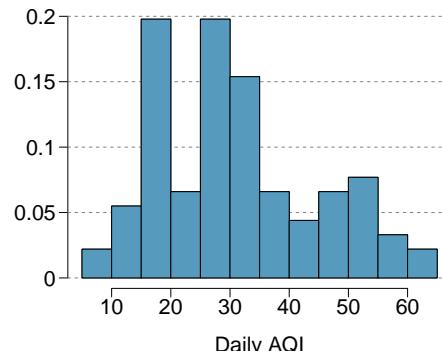


**1.50 Mix-and-match.** Describe the distribution in the histograms below and match them to the box plots.



**1.51 Air quality.** Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.<sup>68</sup>

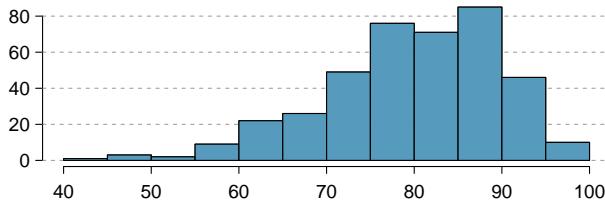
- (a) Estimate the median AQI value of this sample.
- (b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.
- (c) Estimate Q1, Q3, and IQR for the distribution.
- (d) Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.



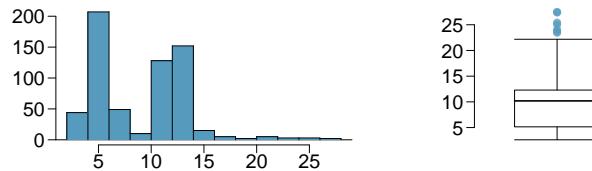
<sup>67</sup>CIA Factbook, Country Comparisons, 2014.

<sup>68</sup>US Environmental Protection Agency, AirData, 2011.

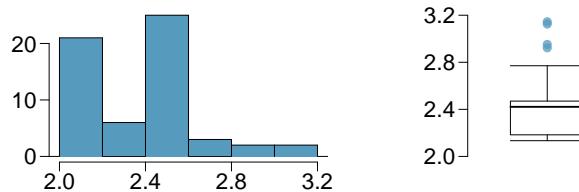
**1.52 Median vs. mean.** Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.



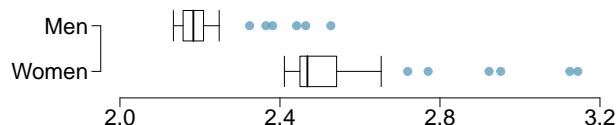
**1.53 Histograms vs. box plots.** Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?



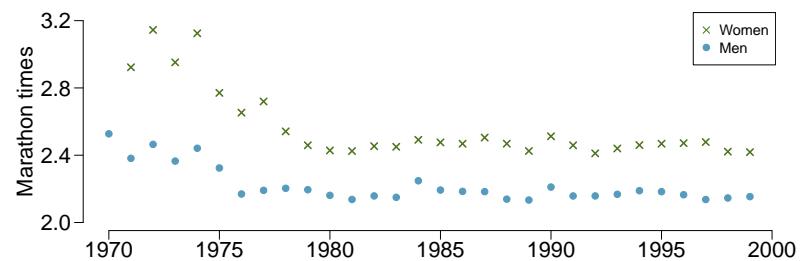
**1.54 Marathon winners.** The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



- (a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?
- (b) What may be the reason for the bimodal distribution? Explain.
- (c) Compare the distribution of marathon times for men and women based on the box plot shown below.



- (d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



**1.55 Distributions and appropriate statistics, Part I.** For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

- (a) Number of pets per household.
- (b) Distance to work, i.e. number of miles between work and home.
- (c) Heights of adult males.

**1.56 Distributions and appropriate statistics, Part II .** For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

- (a) Housing prices in a country where 25% of the houses cost below \$350,000, 50% of the houses cost below \$450,000, 75% of the houses cost below \$1,000,000 and there are a meaningful number of houses that cost more than \$6,000,000.
- (b) Housing prices in a country where 25% of the houses cost below \$300,000, 50% of the houses cost below \$600,000, 75% of the houses cost below \$900,000 and very few houses that cost more than \$1,200,000.
- (c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.
- (d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

**1.57 TV watchers.** Students in an AP Statistics class were asked how many hours of television they watch per week (including online streaming). This sample yielded an average of 4.71 hours, with a standard deviation of 4.18 hours. Is the distribution of number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

**1.58 Exam scores.** The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

**1.59 Facebook friends.** Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users?<sup>69</sup>

**1.60 A new statistic.** The statistic  $\frac{\bar{x}}{\text{median}}$  can be used as a measure of skewness. Suppose we

---

<sup>69</sup>Lars Backstrom. "Anatomy of Facebook". In: *Facebook Data Team's Notes* (2011).

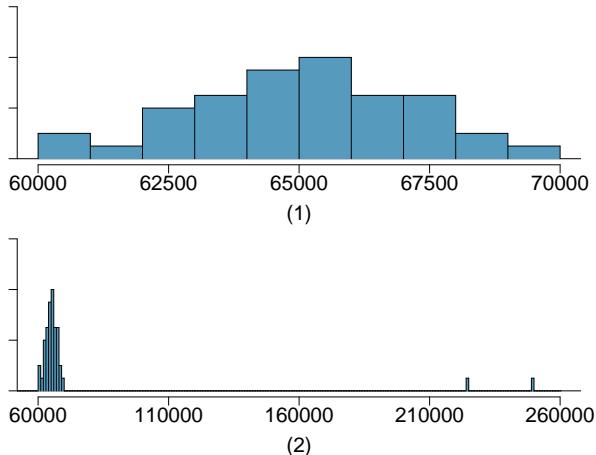
have a distribution where all observations are greater than 0,  $x_i > 0$ . What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a)  $\frac{\bar{x}}{median} = 1$

(b)  $\frac{\bar{x}}{median} < 1$

(c)  $\frac{\bar{x}}{median} > 1$

**1.61 Income at the coffee shop.** The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making \$225,000 and the other \$250,000. The second histogram shows the new income distribution. Summary statistics are also provided.

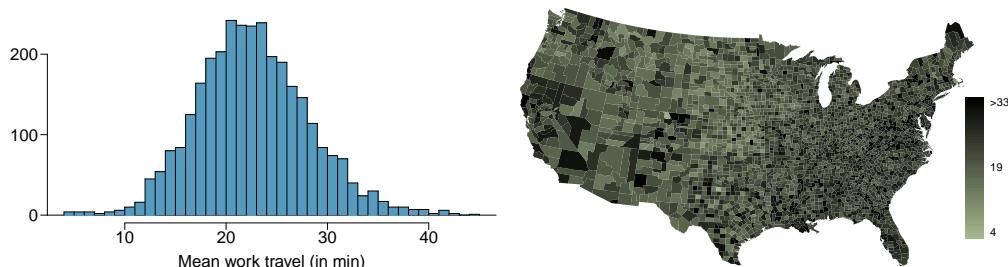


	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

- (a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?
- (b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

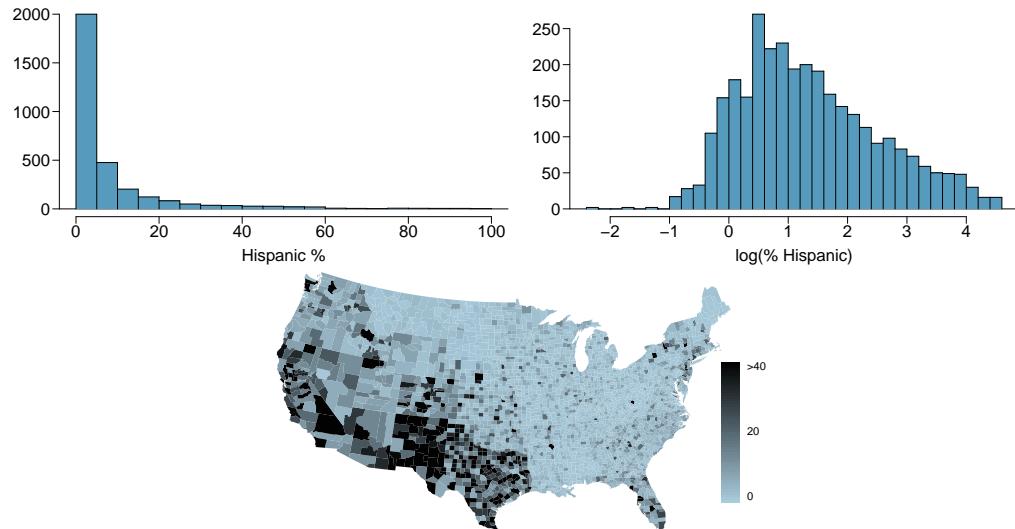
**1.62 Midrange.** The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

**1.63 Commute times.** The US census collects data on time it takes Americans to commute to work, among many other variables. The histogram below shows the distribution of average commute times in 3,143 US counties in 2010. Also shown below is a spatial intensity map of the same data.



- (a) Describe the numerical distribution and comment on whether or not a log transformation may be advisable for these data.
- (b) Describe the spatial distribution of commuting times using the map below.

**1.64 Hispanic population.** The US census collects data on race and ethnicity of Americans, among many other variables. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010. Also shown is a histogram of logs of these values.



- (a) Describe the numerical distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.
- (b) What features of the distribution of the Hispanic population in US counties are apparent in the map but not in the histogram? What features are apparent in the histogram but not the map?
- (c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

## 9.7 Considering categorical data

**1.65 Antibiotic use in children.** The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.



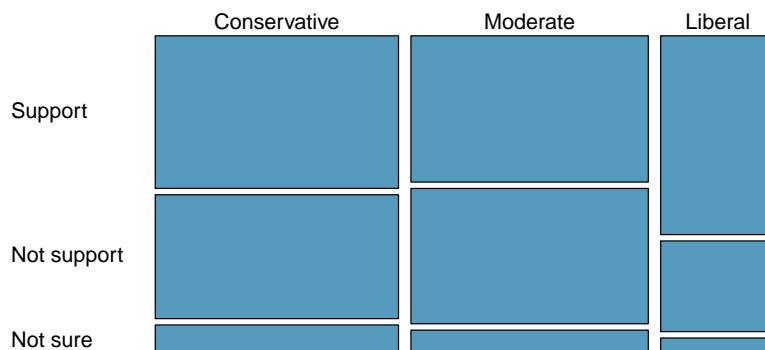
- (a) What features are apparent in the bar plot but not in the pie chart?
- (b) What features are apparent in the pie chart but not in the bar plot?
- (c) Which graph would you prefer to use for displaying these categorical data?

**1.66 Views on immigration.** 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.<sup>70</sup>

		Political ideology			Total
		Conservative	Moderate	Liberal	
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910

- (a) What percent of these Tampa, FL voters identify themselves as conservatives?
- (b) What percent of these Tampa, FL voters are in favor of the citizenship option?
- (c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?
- (d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?
- (e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

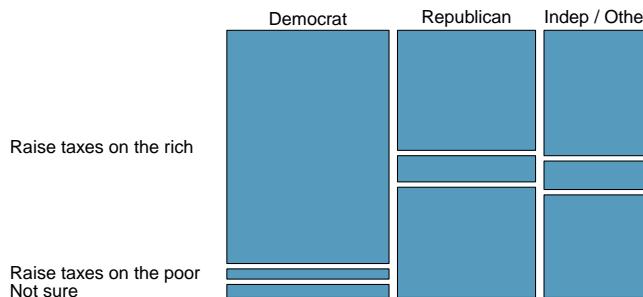
**1.67 Views on the DREAM Act.** A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, do views on the DREAM Act and political ideology appear to be independent? Explain your reasoning.<sup>71</sup>



<sup>70</sup>SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

<sup>71</sup>SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

**1.68 Raise taxes.** A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning.<sup>72</sup>



## 9.8 Case study: gender discrimination

**1.69 Side effects of Avandia.** Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.<sup>73</sup>

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
Total		7,979	219,592	227,571

- (a) Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.
- i. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.
  - ii. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was  $(2,593 / 67,593 = 0.038)$

<sup>72</sup>Public Policy Polling, Americans on College Degrees, Classic Literature, the Seasons, and More, data collected Feb 20-22, 2015.

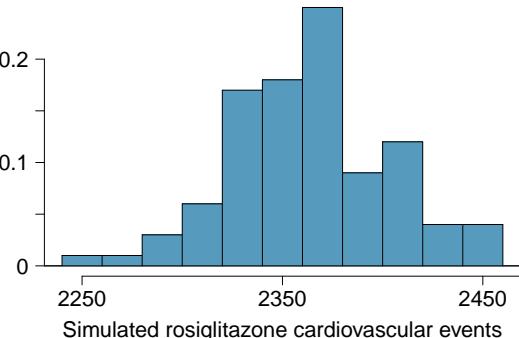
<sup>73</sup>D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. ISSN: 0098-7484.

3.8% for patients on this treatment, while it was only  $(5,386 / 159,978 = 0.034)$  3.4% for patients on pioglitazone.

- iii. The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.
- iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

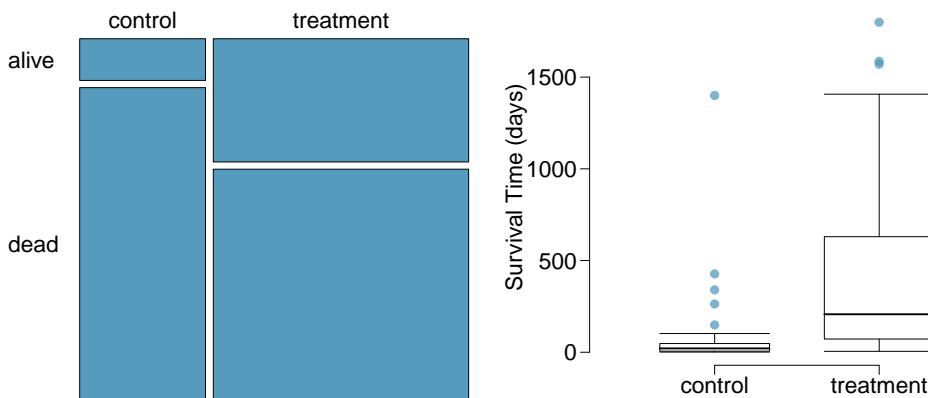
**(See the next page for additional parts to this question.)**

- (b) What proportion of all patients had cardiovascular problems?
  - (c) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?
  - (d) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Use the relative frequency histogram of these counts to answer (i)-(iii).
- i. What are the claims being tested?
  - ii. Compared to the number calculated in part 0.2  
(b), which would provide more support for the alternative hypothesis, *more* or *fewer* patients with cardiovascular problems in the rosiglitazone group?
  - iii. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?



**1.70 Heart transplants.** The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study.<sup>74</sup>

<sup>74</sup>B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.



- (a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

**(See the next page for additional parts to this question.)**

- (b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

- (c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

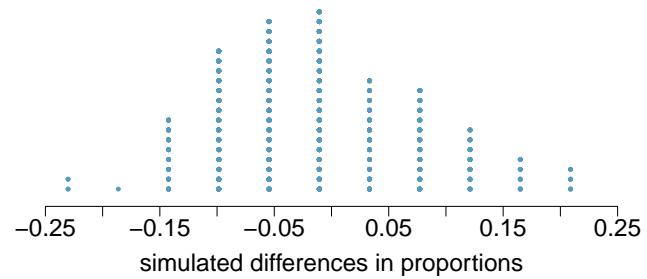
- (d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

- i. What are the claims being tested?

- ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write *alive* on \_\_\_\_\_ cards representing patients who were alive at the end of the study, and *dead* on \_\_\_\_\_ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size \_\_\_\_\_ representing treatment, and another group of size \_\_\_\_\_ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at \_\_\_\_\_. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are \_\_\_\_\_. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

- iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?



# 찾아보기

- control group, 2  
data  
    cars, 24  
df, degrees of freedom (df)를 참고  
experiment, 20  
IQR, 32  
modality  
    다봉, 28  
    단봉, 28  
    이봉, 28  
negative association, 9  
probability sample, sample을 참고  
     $Q_1$ , 32  
     $Q_3$ , 32  
retrospective studies, 16  
 $s$ , 30  
scatterplot, 8  
simulation, 50  
treatment group, 2  
variability, 29  
가중평균, 26  
강건 통계량, 35  
강도 그림, 37  
강도 지도, 36–40  
공히스토그램, 47  
관측 단위(unit of observation), 4  
관측연구, 14  
교락변수, 16  
교락요인, 16  
그리스 문자  
    mu ( $\mu$ ), 25  
    sigma ( $\sigma$ ), 30  
기울  
    꼬리, 27  
    대칭, 27  
    예제: 강한, 34  
    예제: 극단, 35  
    예제: 매우 강함, 27  
    우측 기울, 27  
    좌측 기울, 27  
    긴꼬리, 27  
    꼬리, 27  
눈가림, 22  
단순 임의 표본, 12  
대조집단, 22  
대칭, 27  
대표성, 12  
데이터  
    cars, 23  
    county, 46–48  
    discrimination, 48–52  
    email, 41–46  
    email50, 23–36  
    군, 14  
    시군, 9, 13–40

- 데이터 밀도, 27  
 데이터 행렬(data matrix), 4  
 데이터(data), 1  
     county, 5  
     email50, 4  
     뇌졸중, 1–3, 7  
 독립, 14  
 독립(independent), 9  
 막대그래프  
     조각 막대그래프, 44  
 막대그림, 41  
 모드, 28  
 모의실험, 50  
 모자이크 그래프, 45  
 모집단, 10, 9–13  
 무응답, 12  
 무응답 편의, 12  
 무작위 실험, 15  
 반복, 20  
 반응, 14  
 백분위, 32  
 범주형(categorical), 7  
 변수(variable), 4  
 변환, 36  
 분포, 24  
 분할표, 41  
     열 비율, 42  
     열 총계, 41  
     행 총계, 41  
     행비율, 42  
 블로킹, 20  
 블록, 20  
 빈도표, 41  
 사례(case), 4  
 사분위 범위, 32, 33  
 사분위수  
     제1사분위수, 32  
     제3사분위수, 32  
 산점도, 23  
 상대도수표, 41  
 상자그림, 31  
 평행 상자그림, 47  
 상자수염, 33  
 서수형(ordinal), 7  
 설명, 14  
 수준(levels), 7  
 숫자형(numerical), 7  
 실험, 14  
 양의 연관, 9  
 연관(associated), 8  
 연구 참여자, 22  
 연속형(continuous), 7  
 열 총계, 41  
 요약통계량(summary statistic), 3  
 위약, 15, 22  
 위약 효과, 22  
 이산형(discrete), 7  
 이상점, 33  
 이중눈가림, 22  
 일화적 증거, 11  
 임의화, 50  
 임의화 실험(randomized experiment), 20  
 잠복변수, 16  
 전향적 연구, 16  
 점그림, 24  
 제1사분위수, 32  
 제3사분위수, 32  
 제어, 20  
 종속, 14  
 종속(dependent), 8  
 중위수, 32  
 지원자, 22  
 처리집단, 22  
 총, 18  
 코호트, 14  
 파이그림, 46  
 편의, 12  
 편의 표본, 13  
 편차, 29  
 평균, 24

- 가중평균, 26
- 평균(average), 24
- 평행 상자그림, 47
- 표본, 10, 9–13
  - 군집, 18
  - 군집 표집, 19
  - 군집표본, 18
  - 다단계 표본, 18
  - 다단계 표집, 19
  - 단순임의보집, 16
  - 단순임의표집, 17
  - 무응답, 12
  - 무응답 편의, 12
  - 임의 표본, 12–13
- 총, 18
- 총화표집, 17, 18
- 편의 표본, 13
- 표본 통계량, 34
- 표준편차, 29
- 행 총계, 41
- 환자, 22
- 히스토그램, 26