제5장: 통계적 추론의 기초

Foundations for Inference

Contents

통계적 추론은 주로 모수 추정의 불확실성을 이해하고 정량화하는 것에 관심을 둔다. 설정에 따라 공식과 세부 사항은 달라지지만, 추론의 기초는 통계학 전반에서 동일하다.

우리는 친숙한 주제로 시작한다: 표본비율을 사용하여 모집단비율을 추정하는 아이디어. 다음으로, **신뢰구간(confidence interval)**이라 불리는 것을 만드는데, 이는 실제 모집단 값을 찾을 수 있는 그럴듯한 값들의 범위다. 마지막으로, **가설검정(hypothesis testing)**프레임워크를 소개하는데, 이는 후보가 투표 인구의 과반수 지지를 받는지와 같은 모집단에 대한 주장을 공식적으로 평가할 수 있게 해준다.

5.1 점추정과 표본 변동성

Pew Research와 같은 회사들은 정치, 과학적 이해, 브랜드 인지도 등 많은 주제에 대한 여론이나 지식의 상태를 이해하기 위해 자주 여론조사를 실시한다. 여론조사를 하는 궁극적인 목표는 일반적으로 응답을 사용하여 더 넓은 모집단의 의견이나 지식을 추정하는 것이다.

5.1.1 점추정과 오차

어떤 여론조사에서 미국 대통령의 지지율이 45%라고 했다고 가정하자. 우리는 45%를 전체 모집단에서 응답을 수집했을 때 볼 수 있는 지지율의 **점추정값(point estimate)**으로 간주한다. 이 전체 모집단 응답 비율을 일반적으로 관심 **모수(parameter)**라고 한다. 모수가 비율일 때, 종종 p로 표기하고, 표본비율을 $\hat{p}$("p-hat"으로 발음)로 표기한다.

우리가 두 번째 여론조사를 실시했다면 같은 결과를 얻었을까? 아마 아닐 것이다. 두 번째 여론조사에서 약간 다른 사람들이 무작위로 선택되었을 것이고, 그러면 그 여론조사도 약간 다른 점추정값을 가졌을 것이다. 이러한 표본마다의 차이를 **표본오차(sampling error)**라 한다.

표본오차와 함께, 우리가 주의해야 할 또 다른 유형의 오차가 있다: **편향(bias)**. **편향**은 추정값이 모수를 체계적으로 과대추정하거나 과소추정하도록 만들 때 발생한다. 편향이 발생하는 한 가지 이유는 표본이 전체 모집단을 대표하지 않을 때다 - 이것을 **표본추출 편향(sampling bias)**이라 한다.

5.1.2 점추정의 변동성 이해

Pew Research는 태양 에너지 역할 확대를 지지하는 미국 성인의 비율이 약 $p = 0.88$, 즉 88%라고 추정한다. 만약 1000명의 미국 성인을 다시 표본 추출하고 그들에게도 같은 질문을 한다면, 표본 비율이 얼마나 달라질 것으로 예상할까? 이 맥락에서 표본비율 $\hat{p}$의 변동성을 이해하기 위한 시뮬레이션을 실행할 수 있다. 우리는 다음과 같이 시뮬레이션을 구성할 수 있다:

1. 2018년에 약 2억 5천만 명의 미국 성인이 있었다. 2억 5천만 장의 종이에 88%는 "지지"라고 쓰고 나머지 12%는 "반대"라고 쓴다.
2. 종이 조각들을 섞고 1000명의 미국 성인 표본을 나타내기 위해 1000장을 꺼낸다.
3. "지지"라고 말하는 표본의 비율을 계산한다.

이 시뮬레이션을 2억 5천만 장의 종이로 실행하는 것은 시간이 많이 걸리고 비용이 많이 들지만, 컴퓨터 코드를 사용하여 시뮬레이션할 수 있다. 이 시뮬레이션에서 표본은 $\hat{p}_1 = 0.894$의 점추정값을 주었다. 우리는 시뮬레이션의 모집단비율이 $p = 0.88$임을 알고 있으므로, 추정값의 오차가 $0.894 - 0.88 = +0.014$임을 안다.

한 번의 시뮬레이션으로는 시뮬레이션에서 예상할 수 있는 추정값의 분포를 잘 파악하기에 충분하지 않으므로, 더 많은 시뮬레이션을 실행해야 한다. 두 번째 시뮬레이션에서 우리는 $\hat{p}_2 = 0.885$를 얻었고, 이는 $+0.005$의 오차를 가진다. 또 다른 것에서 $\hat{p}_3 = 0.878$로 -0.002의 오차. 그리고 또 다른 것에서 $\hat{p}_4 = 0.859$의 추정값으로 -0.021의 오차. 컴퓨터의 도움으로 시뮬레이션을 10,000번 실행하고 10,000번의 모든 시뮬레이션 결과의 히스토그램을 그림 5.2에 만들었다. 이 표본비율들의 분포를 **표본추출 분포(sampling distribution)**라 한다. 이 표본추출 분포를 다음과 같이 특성화할 수 있다:

중심(Center). 분포의 중심은 $\bar{x}_{\hat{p}} = 0.880$으로, 모수와 같다. 시뮬레이션이 모집단의 단순무작위표본을 모방했음에 주목하라. 이는 표본추출 편향을 피하는 데 도움이 되는 간단한 표본추출 전략이다.

퍼짐(Spread). 분포의 표준편차는 $s_{\hat{p}} = 0.010$이다. 표본추출 분포나 점추정값의 변동성에 대해 이야기할 때, 우리는 일반적으로 표준편차보다 **표준오차(standard error)**라는 용어를 사용하고, 표본비율과 관련된 표준오차에 대해 $SE_{\hat{p}}$라는 표기법을 사용한다.

모양(Shape). 분포는 대칭이고 종 모양이며, 정규분포와 닮았다.

이러한 발견들은 고무적이다! 모집단비율이 $p = 0.88$이고 표본 크기가 $n = 1000$일 때, 표본비율 $\hat{p}$은 모집단비율의 꽤 좋은 추정값을 주는 경향이 있다. 히스토그램이 정규분포와 닮았다는 흥미로운 관찰도 있다.

표본추출 분포는 관찰되지 않지만 염두에 둔다

실제 세계 응용에서, 우리는 실제로 표본추출 분포를 관찰하지 않지만, 점추정값이 그러한 가상의 분포에서 온다고 항상 생각하는 것이 유용하다. 표본추출 분포를 이해하면 우리가 실제로 관찰하는 점추정값을 특성화하고 이해하는 데 도움이 된다.

예제 5.1

문제: 만약 훨씬 작은 표본 크기 $n = 50$을 사용했다면, $\hat{p}$의 표준오차가 $n = 1000$을 사용했을 때보다 클까 작을까?

풀이: 직관적으로, 더 많은 데이터가 더 적은 데이터보다 나은 것 같고, 일반적으로 그것이 맞다! $p = 0.88$이고 $n = 50$일 때의 전형적인 오차는 $n = 1000$일 때 예상하는 오차보다 **클 것이다**.

예제 5.1은 우리가 계속해서 보게 될 중요한 성질을 강조한다: **더 큰 표본은 더 작은 표본보다 더 정확한 점추정값을 제공하는 경향이 있다**.

5.1.3 중심극한정리

그림 5.2의 분포는 정규분포처럼 보인다. 이것은 이상 현상이 아니다; 이것은 **중심극한정리(Central Limit Theorem)**라 불리는 일반적인 원리의 결과다.

중심극한정리와 성공-실패 조건

관측값들이 독립이고 표본 크기가 충분히 크면, 표본비율 $\hat{p}$은 다음의 평균과 표준오차를 가진 정규분포를 따르는 경향이 있다:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

중심극한정리가 성립하려면, 표본 크기는 $np \geq 10$이고 $n(1-p) \geq 10$일 때 충분히 크다고 여겨지며, 이를 **성공-실패 조건(success-failure condition)**이라 한다.

중심극한정리는 매우 중요하며, 통계학의 많은 부분에 대한 기초를 제공한다. 중심극한정리를 적용하기 시작할 때, 두 가지 기술적 조건에 유의하라: 관측값들이 독립이어야 하고, 표본 크기가 $np \geq 10$이고 $n(1-p) \geq 10$이 되도록 충분히 커야 한다.

예제 5.2

문제: 앞서 $p = 0.88$이고 $n = 1000$일 때 시뮬레이션된 데이터를 사용하여 $\hat{p}$의 평균과 표준오차를 추정했다. 중심극한정리가 적용되고 표본추출 분포가 근사적으로 정규분포임을 확인하라.

풀이:

독립성. 각 표본비율 $\hat{p}$에 대해 $n = 1000$개의 관측값이 있고, 그 관측값들 각각은 독립적인 추출이다. 관측값들이 독립으로 간주되는 가장 일반적인 방법은 그들이 단순무작위표본에서 왔을 때다.

성공-실패 조건. 계산된 두 값이 10보다 큰지 확인하여 표본 크기가 충분히 큰지 확인할 수 있다:

$$np = 1000 \times 0.88 = 880 \geq 10$$
$$n(1-p) = 1000 \times (1 - 0.88) = 120 \geq 10$$

독립성과 성공-실패 조건이 모두 만족되므로, 중심극한정리가 적용되고, $\hat{p}$을 정규분포를 사용하여 모델링하는 것이 합리적이다.

표본 관측값이 독립인지 확인하는 방법

- 실험의 피험자들은 처리 그룹에 무작위 배정을 받으면 독립으로 간주된다.
- 관측값들이 단순무작위표본에서 왔다면, 그들은 독립이다.
- 표본이 겉보기에 무작위인 과정에서 왔다면(예: 조립 라인에서의 가끔 발생하는 오류), 독립성을 확인하는 것이 더 어렵다. 이 경우, 최선의 판단을 사용하라.

모집단에서 온 표본에 대해 때때로 추가되는 조건은 모집단의 10%보다 크지 않아야 한다는 것이다. 표본이 모집단 크기의 10%를 초과하면, 우리가 논의하는 방법은 더 고급 방법을 사용했을 때 얻을 수 있는 것보다 표본추출 오차를 약간 과대추정하는 경향이 있다. 이것은 매우 드물게 문제가 되고, 문제가 될 때 우리의 방법은 보수적인 경향이 있으므로, 이 추가 확인을 선택사항으로 간주한다.

예제 5.3

문제: 중심극한정리에 따라 $p = 0.88$이고 $n = 1000$일 때 $\hat{p}$의 이론적 평균과 표준오차를 계산하라.

풀이: $\hat{p}$들의 평균은 단순히 모집단비율이다: $\mu_{\hat{p}} = 0.88$.

$\hat{p}$의 표준오차 계산은 다음 공식을 사용한다:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(1-0.88)}{1000}} = 0.010$$

Python 코드:

```
import numpy as np

p = 0.88
n = 1000

mu_p_hat = p
se_p_hat = np.sqrt(p * (1 - p) / n)

print(f"  평균:  = {mu_p_hat}")
print(f"  :  = {se_p_hat:.4f}")

# 출력:
# 이론적  평균:  = 0.88
# 표준오차:  = 0.0103
```

예제 5.4

문제: 표본비율 $\hat{p}$이 모집단 값 $p = 0.88$의 ± 0.02 (2%) 이내에 있을 빈도를 추정하라. 예제 5.2와 5.3에 기초하여, 분포가 근사적으로 $N(\mu_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010)$임을 안다.

풀이: 4.1절에서 많이 연습한 후, 이 정규분포 예제가 익숙하게 느껴지기를 바란다! 우리는 0.86과 0.90 사이의 $\hat{p}$들의 비율을 이해하고 싶다.

$\mu_{\hat{p}} = 0.88$이고 $SE_{\hat{p}} = 0.010$으로, 왼쪽과 오른쪽 절단점 모두에 대해 Z-점수를 계산할 수 있다:

$$Z_{0.86} = \frac{0.86 - 0.88}{0.010} = -2 \qquad Z_{0.90} = \frac{0.90 - 0.88}{0.010} = 2$$

통계 소프트웨어, 그래프 계산기, 또는 표를 사용하여 꼬리 면적을 찾을 수 있고, 어떤 경우든 각각 0.0228임을 발견할 것이다. 총 꼬리 면적은 $2 \times 0.0228 = 0.0456$이고, 이는 음영 면적 0.9544를 남긴다. 즉, 그림 5.2의 표본추출 분포의 약 **95.44%**가 모집단비율 $p = 0.88$의 ± 0.02 이내에 있다.

Python 코드:

```
from scipy import stats
import numpy as np

mu = 0.88
se = 0.010

# Z-점수  계산
z_lower = (0.86 - mu) / se
z_upper = (0.90 - mu) / se

# 꼬리  면적
tail_lower = stats.norm.cdf(z_lower)
```

```
tail_upper = 1 - stats.norm.cdf(z_upper)

# 중앙 면적
central_area = 1 - (tail_lower + tail_upper)

print(f"Z_0.86 = {z_lower}, Z_0.90 = {z_upper}")
print(f"  면적: {tail_lower:.4f} + {tail_upper:.4f} = {tail_lower + tail_upper:.4f}")
print(f"p̂ ±0.02 이내에 있을 확률: {central_area:.4f}")

# 출력:
# Z_0.86 = -2.0, Z_0.90 = 2.0
# 꼬리 면적: 0.0228 + 0.0228 = 0.0456
# p̂이 ±0.02 이내에 있을 확률: 0.9544
```

Guided Practice 5.5

문제: 예제 5.1에서 더 작은 표본이 덜 신뢰할 수 있는 추정값을 생성하는 경향이 있음을 논의했다. 이 직관이 $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ 공식에 어떻게 반영되는지 설명하라.

풀이: 표본 크기 n이 분수의 분모(아래쪽)에 있으므로, 더 큰 표본 크기는 계산될 때 전체 표현식이 더 작아지는 경향이 있다는 것을 의미한다. 즉, **더 큰 표본 크기는 더 작은 표준오차에 해당한다**.

5.1.4 중심극한정리를 실세계 설정에 적용하기

우리는 모집단의 모든 개인에 대한 비싼 여론조사를 실시하지 않는 한 실제로 모집단비율을 알지 못한다. 앞서 $p = 0.88$의 값은 Pew Research가 1000명의 미국 성인에 대해 실시한 여론조사에 기반했고, $\hat{p} = 0.887$이 태양 에너지 확대를 지지하는 것으로 발견되었다. 연구자들은 궁금해했을 것이다: 여론조사의 표본비율이 근사적으로 정규분포를 따르는가? 중심극한정리의 조건을 확인할 수 있다:

독립성. 여론조사는 미국 성인의 단순무작위표본이며, 이는 관측값들이 독립임을 의미한다.

성공-실패 조건. 이 조건을 확인하려면 np와 $n(1-p)$가 모두 10보다 큰지 확인하기 위해 모집단비율 p가 필요하다. 그러나 우리는 실제로 p를 알지 못한다 - 이것이 정확히 여론조사원들이 표본을 추출하는 이유다! 이런 경우, 우리는 종종 성공-실패 조건을 확인하는 차선책으로 $\hat{p}$을 사용한다:

$$n\hat{p} = 1000 \times 0.887 = 887 \qquad n(1-\hat{p}) = 1000 \times (1 - 0.887) = 113$$

표본비율 $\hat{p}$은 이 확인 중에 p의 합리적인 대용물로 작용하고, 이 경우 각 값은 최소 10을 훨씬 넘는다.

p 대신 $\hat{p}$을 사용하는 이 대체 근사는 표본비율의 표준오차를 계산할 때도 유용하다:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.887(1-0.887)}{1000}} = 0.010$$

이 대체 기법은 때때로 **"플러그인 원리(plug-in principle)"**라 불린다. 이 경우, $SE_{\hat{p}}$는 앞서 0.88로 계산을 완료했을 때와 비교하여 소수점 세 자리만 사용해서는 감지될 만큼 변하지 않았다. 계산된 표준오차는 한 표본에서 다른 표본으로 약간 다른 비율을 관찰하더라도 합리적으로 안정적인 경향이 있다.

5.1.5 중심극한정리에 대한 추가 세부사항

이 장에서 지금까지 많은 예제에서 중심극한정리를 적용해왔다:

중심극한정리 요약

관측값들이 독립이고 표본 크기가 충분히 크면, $\hat{p}$의 분포는 정규분포를 닮는다:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

표본 크기는 $np \geq 10$이고 $n(1-p) \geq 10$일 때 충분히 크다고 간주된다.

이 절에서는 성공-실패 조건을 탐구하고 중심극한정리를 더 잘 이해하려고 한다.

대답할 흥미로운 질문은, $np < 10$이거나 $n(1-p) < 10$이면 어떻게 되는가? 5.1.2절에서 했듯이, 실제 비율이 예를 들어 $p = 0.25$인 다른 크기의 표본을 추출하는 것을 시뮬레이션할 수 있다. 여기 크기 10인 표본이 있다:

아니오, 아니오, 예, 예, 아니오, 아니오, 아니오, 아니오, 아니오, 아니오

이 표본에서 우리는 예의 표본비율 $\hat{p} = \frac{2}{10} = 0.2$를 관찰한다. $n = 10$이고 $p = 0.25$일 때 $\hat{p}$의 표본추출 분포를 이해하기 위해 많은 그러한 비율들을 시뮬레이션할 수 있는데, 이를 그림 5.3에서 같은 평균과 변동성을 가진 정규분포와 함께 그렸다. 이 분포들은 여러 가지 중요한 차이점이 있다:

	단봉?	부드러움?	대칭?
정규분포: $N(0.25, 0.14)$	예	예	예
$n = 10, p = 0.25$	예	아니오	아니오

$n = 10$이고 $p = 0.25$일 때 성공-실패 조건이 만족되지 않았음에 주목하라:

$$np = 10 \times 0.25 = 2.5 \qquad n(1-p) = 10 \times 0.75 = 7.5$$

우리는 여러 추가 시뮬레이션을 완료하여 몇 가지 추세를 볼 수 있다:

1. np 또는 $n(1-p)$가 작으면, 분포는 더 이산적이다, 즉 연속적이지 않다.
2. np 또는 $n(1-p)$가 10보다 작으면, 분포의 비대칭이 더 주목할 만하다.
3. np와 $n(1-p)$가 둘 다 클수록, 분포는 더 정규분포에 가깝다.
4. np와 $n(1-p)$가 둘 다 매우 크면, 분포의 이산성이 거의 드러나지 않고, 분포는 정규분포처럼 훨씬 더 보인다.

분포의 평균과 표준오차가 어떻게 변하는지도 주목하라:

1. 분포의 중심은 항상 시뮬레이션을 생성하는 데 사용된 모집단비율 p에 있다. $\hat{p}$의 표본추출 분포가 항상 모집단 모수 p에 중심을 두고 있기 때문에, 데이터가 독립이고 그러한 모집단에서 추출되었을 때 표본비율 $\hat{p}$은 **불편(unbiased)**이다.
2. 특정 모집단비율 p에 대해, 표본추출 분포의 변동성은 표본 크기 n이 커질수록 감소한다. 이것은 직관과 일치할 것이다: 더 큰 표본 크기에 기반한 추정은 더 정확한 경향이 있다.
3. 특정 표본 크기에 대해, 변동성은 $p = 0.5$일 때 가장 클 것이다. 이것은 표준오차 공식 $SE = \sqrt{\frac{p(1-p)}{n}}$에서 비율 p의 역할을 반영한다. 표준오차는 $p = 0.5$일 때 가장 크다.

$\hat{p}$의 분포는 $\hat{p}$이 항상 이산적인 값(x/n)을 취하기 때문에 완벽하게 정규분포처럼 보이지는 않을 것이다. 이것은 항상 정도의 문제이며, 우리는 이 책에서 np와 $n(1-p)$에 대해 최소 10인 표준 성공-실패 조건을 지침으로 사용할 것이다.

5.1.6 다른 통계량으로 프레임워크 확장하기

표본 통계량을 사용하여 모수를 추정하는 전략은 매우 일반적이며, 비율 외의 다른 통계량에도 적용할 수 있는 전략이다. 예를 들어, 특정 대학 졸업생의 평균 급여를 추정하고 싶다면, 최근 졸업생의 무작위 표본을 조사할 수 있다; 그 예에서 우리는 표본 평균 $\bar{x}$를 사용하여 모든 졸업생에 대한 모집단 평균 μ를 추정할 것이다. 또 다른 예로, 두 웹사이트의 제품 가격 차이를 추정하고 싶다면, 두 사이트 모두에서 구입할 수 있는 제품의 무작위 표본을 추출하고, 각각의 가격을 확인한 다음, 평균 차이를 계산할 수 있다; 이 전략은 확실히 점추정값을 통해 실제 차이에 대한 어떤 아이디어를 줄 것이다.

이 장은 단일 비율 맥락을 강조하지만, 이 책 전체에서 이러한 방법이 적용될 많은 다른 맥락을 만나게 될 것이다. 세부 사항이 조금 바뀌더라도 원리와 일반적인 아이디어는 같다.

새로운 시각: 표본추출 분포의 힘

통계학의 가장 아름다운 점 중 하나는 **표본추출 분포**라는 개념이다. 우리는 실제로 표본추출 분포를 관찰하지 않지만, 그것이 어떤 모양인지 이론적으로 알고 있다. 이 지식 덕분에:

- 단 한 번의 표본에서 모집단에 대한 추론을 할 수 있다
- 추정값의 불확실성을 정량화할 수 있다
- 다른 통계학자들이 같은 방법을 사용하면 장기적으로 얼마나 자주 맞을지 알 수 있다

이것은 마치 주사위를 한 번만 던지면서도 주사위가 공정한지 판단할 수 있는 것과 같다 - 우리는 공정한 주사위의 이론적 분포를 알고 있기 때문이다.

5.1절 연습문제

연습문제 5.1 (모수 식별하기, Part I)

다음 각 상황에서 관심 모수가 평균인지 비율인지 말하라. 개별 응답이 수치형인지 범주형인지 검토하는 것이 도움이 될 수 있다.

(a) 조사에서 100명의 대학생에게 일주일에 인터넷에서 몇 시간을 보내는지 질문한다. → **평균** (수치형 응답)

(b) 조사에서 100명의 대학생에게 "인터넷에서 보내는 시간의 몇 퍼센트가 과제 작업의 일부인가?"라고 질문한다. → **평균** (백분율은 수치형)

(c) 조사에서 100명의 대학생에게 논문에서 위키피디아의 정보를 인용했는지 여부를 질문한다. → **비율** (예/아니오 범주형)

(d) 조사에서 100명의 대학생에게 총 주간 지출의 몇 퍼센트가 알코올 음료에 쓰이는지 질문한다. → **평균** (백분율은 수치형)

(e) 100명의 최근 대학 졸업생 표본에서 85%가 졸업일로부터 1년 이내에 직장을 구할 것으로 기대한다고 발견된다. → **비율** (범주형 응답의 비율)

연습문제 5.3 (품질 관리)

한 공장에서 매일 아침 시리얼 상자를 채운다. 상자는 평균 20온스, 표준편차 2온스의 정규분포를 따르는 것으로 추정된다.

(a) 한 상자가 21온스보다 무거울 확률은?

풀이:

$$Z = \frac{21 - 20}{2} = 0.5$$

$$P(X > 21) = P(Z > 0.5) = 1 - \Phi(0.5) = 1 - 0.6915 = 0.3085$$

약 **30.85%**

(b) 25개 상자의 무작위 표본 평균이 21온스보다 무거울 확률은?

풀이:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{25}} = 0.4$$

$$Z = \frac{21 - 20}{0.4} = 2.5$$

$$P(\bar{X} > 21) = P(Z > 2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062$$

약 **0.62%**

Python 코드:

```
from scipy import stats
import numpy as np

mu = 20
sigma = 2

# (a) 한 상자가 21온스보다 무거울 확률
z_single = (21 - mu) / sigma
prob_single = 1 - stats.norm.cdf(z_single)
print(f"(a) P(X > 21) = {prob_single:.4f}")

# (b) 25개 상자 평균이 21온스보다 무거울 확률
n = 25
se = sigma / np.sqrt(n)
z_sample = (21 - mu) / se
```

```python
prob_sample = 1 - stats.norm.cdf(z_sample)
print(f"(b) P(X > 21) = {prob_sample:.4f}")

# 출력:
# (a) P(X > 21) = 0.3085
# (b) P(X > 21) = 0.0062
```

연습문제 5.5 (물 표본)

한 비영리 단체가 식수에서 납 수치가 높은 가구의 비율을 이해하고 싶어한다. 그들은 가구의 최소 5%는 납 수치가 높을 것으로 예상하지만 약 30% 이상은 아닐 것이다. 그들은 800가구를 무작위로 표본 추출하고 주인과 협력하여 물 표본을 회수하고, 납 수치가 높은 가구의 비율을 계산한다. 그들은 이것을 1000번 반복하고 표본비율의 분포를 만든다.

(a) 이 분포를 무엇이라 부르는가? → **표본추출 분포(sampling distribution)**

(b) 이 분포의 모양이 대칭, 오른쪽 비대칭, 왼쪽 비대칭 중 어떤 것일 것으로 예상하는가? → 비율이 0.08 (8%)에 분포되어 있고 이는 0.5에서 멀리 떨어져 있으므로, 분포는 약간 **오른쪽으로 비대칭**일 수 있다. 그러나 $np = 800 \times 0.08 = 64 \geq 10$이고 $n(1-p) = 736 \geq 10$이므로 꽤 대칭에 가까울 것이다.

(c) 비율이 약 8%에 분포되어 있다면, 분포의 변동성은 얼마인가?

$$SE = \sqrt{\frac{0.08 \times 0.92}{800}} = \sqrt{0.000092} = 0.0096$$

(d) (c)에서 계산한 값의 공식적인 이름은? → **표준오차(standard error)**

(e) 연구자들의 예산이 감소하여 표본당 250개의 관측값만 수집할 수 있게 되었지만, 여전히 1000개의 표본을 수집할 수 있다. 새 분포의 변동성은 각 표본이 800개의 관측값을 포함했을 때의 분포의 변동성과 어떻게 비교될까?

$$SE_{250} = \sqrt{\frac{0.08 \times 0.92}{250}} = 0.0172$$

표본 크기가 감소하면 변동성이 **증가한다**.

Python 코드:

```python
import numpy as np

p = 0.08

# (c) n = 800
n1 = 800
se1 = np.sqrt(p * (1 - p) / n1)
print(f"(c) SE (n=800) = {se1:.4f}")

# (e) n = 250
n2 = 250
se2 = np.sqrt(p * (1 - p) / n2)
print(f"(e) SE (n=250) = {se2:.4f}")
print(f"   비율: {se2/se1:.2f}배  증가")

# 출력:
# (c) SE (n=800) = 0.0096
# (e) SE (n=250) = 0.0172
# 변동성  비율: 1.79배  증가
```

5.2 비율에 대한 신뢰구간

표본비율 $\hat{p}$은 모집단비율 p에 대한 단일의 그럴듯한 값을 제공한다. 그러나 표본비율은 완벽하지 않으며 그와 관련된 표준오차가 있을 것이다. 모집단비율에 대한 추정값을 말할 때, 단지 점추정값만 제공하는 것보다 그럴듯한 값들의 범위를 제공하는 것이 더 좋은 관행이다.

5.2.1 모집단 모수 포착하기

점추정값만 사용하는 것은 흐린 호수에서 작살로 물고기를 잡으려는 것과 같다. 물고기를 본 곳에 작살을 던질 수 있지만, 아마도 빗나갈 것이다. 반면에, 그 지역에 그물을 던지면 물고기를 잡을 좋은 기회가 있다. **신뢰구간(confidence interval)**은 그물로 낚시하는 것과 같으며, 모집단 모수를 찾을 가능성이 있는 그럴듯한 값들의 범위를 나타낸다.

점추정값 $\hat{p}$을 보고하면, 정확한 모집단비율을 맞추지 못할 것이다. 반면에, 신뢰구간을 나타내는 그럴듯한 값들의 범위를 보고하면, 모수를 포착할 좋은 기회가 있다.

Guided Practice 5.6

문제: 신뢰구간에서 모집단비율을 포착하는 것을 매우 확실히 하고 싶다면, 더 넓은 구간을 사용해야 할까 더 좁은 구간을 사용해야 할까?

풀이: 물고기를 더 확실히 잡고 싶다면, 더 넓은 그물을 사용해야 한다. 마찬가지로, 모수를 포착하는 것을 더 확실히 하고 싶다면 **더 넓은 신뢰구간**을 사용한다.

5.2.2 95% 신뢰구간 구성하기

우리의 표본비율 $\hat{p}$은 모집단비율의 가장 그럴듯한 값이므로, 이 점추정값 주변에 신뢰구간을 구축하는 것이 합리적이다. 표준오차는 신뢰구간을 얼마나 크게 만들어야 하는지에 대한 지침을 제공한다.

표준오차는 점추정값의 표준편차를 나타내고, 중심극한정리 조건이 만족될 때, 점추정값은 정규분포를 밀접하게 따른다. 정규분포에서 데이터의 95%는 평균의 1.96 표준편차 이내에 있다. 이 원리를 사용하여, 구간이 모집단비율을 포착한다고 95% 확신하기 위해 표본비율에서 1.96 표준오차까지 확장되는 신뢰구간을 구성할 수 있다:

$$\text{점추정값} \pm 1.96 \times SE$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

그런데 "95% 확신"은 무엇을 의미하는가? 많은 표본을 추출하고 각각에서 95% 신뢰구간을 구축했다고 가정하자. 그러면 그 구간의 약 95%가 모수 p를 포함할 것이다. 그림 5.6은 5.1.2절의 시뮬레이션에서 25개의 표본으로부터 25개의 구간을 만드는 과정을 보여주는데, 결과로 나온 신뢰구간 중 24개는 시뮬레이션의 모집단비율 $p = 0.88$을 포함하고, 1개의 구간은 포함하지 않는다.

예제 5.7

문제: 그림 5.6에서 하나의 구간이 $p = 0.88$을 포함하지 않는다. 이것은 시뮬레이션에 사용된 모집단비율이 $p = 0.88$이 아닐 수 있음을 의미하는가?

풀이: 일부 관측값들이 자연스럽게 평균에서 1.96 표준편차 이상 떨어져 발생하는 것처럼, 일부 점추정값들은 관심 모수에서 1.96 표준오차 이상 떨어져 있을 것이다. 신뢰구간은 그럴듯한 값들의 범위만을 제공한다. 데이터에 기반하여 다른 값들이 그럴듯하지 않다고 말할 수 있지만, 이것이 그들이 불가능하다는 것을 의미하지는 않는다.

모수에 대한 95% 신뢰구간

점추정값의 분포가 중심극한정리의 조건을 충족하여 정규분포를 밀접하게 따를 때, 95% 신뢰구간을 다음과 같이 구성할 수 있다:

$$\text{점추정값} \pm 1.96 \times SE$$

예제 5.8

문제: 5.1절에서 우리는 1000명의 미국 성인 무작위 표본 중 88.7%가 태양열 발전의 역할 확대를 지지한다는 Pew Research 여론조사에 대해 배웠다. 모집단비율에 대한 95% 신뢰구간을 계산하고 해석하라.

풀이: 우리는 앞서 $\hat{p}$이 정규분포를 따르고 표준오차가 $SE_{\hat{p}} = 0.010$임을 확인했다. 95% 신뢰구간을 계산하려면, 점추정값 $\hat{p} = 0.887$과 표준오차를 95% 신뢰구간 공식에 대입한다:

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \rightarrow 0.887 \pm 1.96 \times 0.010 \rightarrow (0.8674, 0.9066)$$

태양열 발전 확대를 지지하는 미국 성인의 실제 비율이 **86.7%에서 90.7% 사이**에 있다고 95% 확신한다. (신뢰구간을 보고할 때 가장 가까운 퍼센트 포인트나 0.1 퍼센트 포인트로 반올림하는 것이 일반적이다.)

Python 코드:

```python
import numpy as np

p_hat = 0.887
n = 1000
z_star = 1.96

se = np.sqrt(p_hat * (1 - p_hat) / n)
margin_of_error = z_star * se

lower = p_hat - margin_of_error
upper = p_hat + margin_of_error

print(f"   : {p_hat}")
print(f"   : {se:.4f}")
print(f"   : {margin_of_error:.4f}")
print(f"95% 신뢰구간: ({lower:.4f}, {upper:.4f})")
print(f"95% 신뢰구간: ({lower*100:.1f}%, {upper*100:.1f}%)")

# 출력:
# 점추정값: 0.887
# 표준오차: 0.0100
# 오차한계: 0.0196
# 95% 신뢰구간: (0.8674, 0.9066)
# 95% 신뢰구간: (86.7%, 90.7%)
```

5.2.3 신뢰수준 변경하기

신뢰수준이 95%보다 높은 신뢰구간, 예를 들어 99%의 신뢰수준을 고려하고 싶다고 가정하자. 물고기를 잡으려는 비유를 다시 생각해보자: 물고기를 잡을 것을 더 확신하고 싶다면, 더 넓은 그물을 사용해야 한다. 99% 신뢰수준을 만들려면 95% 구간도 넓혀야 한다. 반면에, 90%와 같이 더 낮은 신뢰의 구간을 원한다면, 원래의 95% 구간보다 약간 좁은 구간을 사용할 수 있다.

95% 신뢰구간 구조는 다른 신뢰수준의 구간을 만드는 방법에 대한 지침을 제공한다. 정규분포를 따르는 점추정값에 대한 일반적인 95% 신뢰구간은:

$$\text{점추정값} \pm 1.96 \times SE$$

이 구간에는 세 가지 구성요소가 있다: 점추정값, “1.96”, 그리고 표준오차. 1.96 × SE의 선택은 추정값이 약 95%의 시간 동안 모수의 1.96 표준오차 이내에 있기 때문에 데이터의 95%를 포착하는 것에 기반한다. 1.96의 선택은 95% 신뢰수준에 해당한다.

Guided Practice 5.9

문제: X가 정규분포를 따르는 확률변수라면, X의 값이 평균의 2.58 표준편차 이내에 있을 확률은 얼마인가?

풀이: 이것은 Z-점수가 -2.58보다 크지만 2.58보다 작을 빈도를 묻는 것과 동등하다. 이 확률을 결정하기 위해, 통계 소프트웨어, 계산기, 또는 표를 사용하여 정규분포에서 -2.58과 2.58을 찾을 수 있다: 0.0049와 0.9951. 따라서, 관측되지 않은 정규 확률변수 X가 μ의 2.58 표준편차 이내에 있을 확률은 $0.9951 - 0.0049 \approx$ **0.99** 또는 **99%**다.

Guided Practice 5.9는 정규 확률변수가 99%의 시간 동안 평균의 2.58 표준편차 이내에 있을 것임을 강조한다. 99% 신뢰구간을 만들려면, 95% 신뢰구간 공식의 1.96을 2.58로 변경한다. 즉, 99% 신뢰구간의 공식은:

$$\text{점추정값} \pm 2.58 \times SE$$

임의의 신뢰수준을 사용한 신뢰구간

점추정값이 표준오차 SE를 가진 정규 모델을 밀접하게 따른다면, 모집단 모수에 대한 신뢰구간은:

$$\text{점추정값} \pm z^* \times SE$$

여기서 z^*는 선택된 신뢰수준에 해당한다.

그림 5.7은 신뢰수준에 기반하여 z^*를 식별하는 방법을 보여준다. 표준 정규분포 $N(0, 1)$에서 $-z^*$와 z^* 사이의 면적이 신뢰수준에 해당하도록 z^*를 선택한다.

신뢰수준	z^*	꼬리 면적 (각각)
90%	1.645	5%
95%	1.96	2.5%
99%	2.576	0.5%

오차한계

신뢰구간에서 $z^* \times SE$를 **오차한계(margin of error)**라 한다.

예제 5.10

문제: 예제 5.8의 데이터를 사용하여 태양열 발전 사용 확대를 지지하는 미국 성인 비율에 대한 90% 신뢰구간을 만들어라. 정규성 조건은 이미 확인했다.

풀이: 먼저 표준 정규분포 $N(\mu = 0, \sigma = 1)$에서 $-z^*$와 z^* 사이에 분포의 90%가 떨어지도록 하는 z^*를 찾는다. 그래프 계산기, 통계 소프트웨어, 또는 상위 꼬리 5%(다른 5%는 하위 꼬리에 있음)를 찾아 확률 표를 사용하여 이것을 할 수 있다: $z^* = 1.65$. 그러면 90% 신뢰구간은 다음과 같이 계산할 수 있다:

$$\hat{p} \pm 1.6449 \times SE_{\hat{p}} \to 0.887 \pm 1.65 \times 0.0100 \to (0.8705, 0.9034)$$

즉, 2018년에 미국 성인의 **87.1%에서 90.3%**가 태양열 발전 확대를 지지했다고 90% 확신한다.

Python 코드:

```python
import numpy as np
from scipy import stats

p_hat = 0.887
se = 0.010

# 90% 신뢰구간
z_90 = stats.norm.ppf(0.95)  # 상위 5% 점
```

```
ci_90 = (p_hat - z_90 * se, p_hat + z_90 * se)

# 95% 신뢰구간
z_95 = stats.norm.ppf(0.975)  # 상위 2.5% 점
ci_95 = (p_hat - z_95 * se, p_hat + z_95 * se)

# 99% 신뢰구간
z_99 = stats.norm.ppf(0.995)  # 상위 0.5% 점
ci_99 = (p_hat - z_99 * se, p_hat + z_99 * se)

print(f"90% CI: z* = {z_90:.3f}, ({ci_90[0]:.4f}, {ci_90[1]:.4f})")
print(f"95% CI: z* = {z_95:.3f}, ({ci_95[0]:.4f}, {ci_95[1]:.4f})")
print(f"99% CI: z* = {z_99:.3f}, ({ci_99[0]:.4f}, {ci_99[1]:.4f})")

# 출력:
# 90% CI: z* = 1.645, (0.8706, 0.9034)
# 95% CI: z* = 1.960, (0.8674, 0.9066)
# 99% CI: z* = 2.576, (0.8612, 0.9128)
```

단일 비율에 대한 신뢰구간 구성 단계

단일 비율 신뢰구간이 응용에 도움이 될 것으로 결정했다면, 구간을 구성하는 네 가지 단계가 있다:
준비(Prepare). $\hat{p}$와 n을 식별하고, 어떤 신뢰수준을 사용할지 결정한다.
확인(Check). $\hat{p}$이 거의 정규분포임을 보장하는 조건을 확인한다. 단일 비율 신뢰구간에서, 성공-실패 조건을 확인할 때 p 대신 $\hat{p}$을 사용한다.
계산(Calculate). 조건이 충족되면, $\hat{p}$을 사용하여 SE를 계산하고, z^*를 찾고, 구간을 구성한다.
결론(Conclude). 문제의 맥락에서 신뢰구간을 해석한다.

5.2.4 추가 사례 연구

2014년 10월 23일 뉴욕시에서, 최근 기니에서 에볼라 환자를 치료하던 의사가 약간의 열로 병원에 갔고 이후 에볼라 진단을 받았다. 얼마 지나지 않아, NBC 4 New York/The Wall Street Journal/Marist 여론조사에서 뉴욕 주민의 82%가 "에볼라 환자와 접촉한 사람에 대한 의무적인 21일 격리"를 지지하는 것으로 나타났다. 이 여론조사는 2014년 10월 26일에서 28일 사이에 1,042명의 뉴욕 성인의 응답을 포함했다.

예제 5.11

문제: 이 경우 점추정값은 무엇이고, 그 점추정값을 정규분포를 사용하여 모델링하는 것이 합리적인가?

풀이: 크기 $n = 1042$의 표본에 기반한 점추정값은 $\hat{p} = 0.82$다. $\hat{p}$이 정규분포를 사용하여 합리적으로 모델링될 수 있는지 확인하기 위해, 독립성(여론조사는 단순무작위표본에 기반함)과 성공-실패 조건($1042 \times \hat{p} \approx 854$와 $1042 \times (1-\hat{p}) \approx 188$, 둘 다 10보다 쉽게 큼)을 확인한다. 조건이 충족되면, $\hat{p}$의 표본추출 분포가 정규분포를 사용하여 합리적으로 모델링될 수 있음을 확신한다.

예제 5.12

문제: 에볼라 조사에서 $\hat{p} = 0.82$의 표준오차를 추정하라.

풀이: $p \approx \hat{p} = 0.82$의 대체 근사를 사용하여 표준오차를 계산한다:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.82(1-0.82)}{1042}} = 0.012$$

예제 5.13

문제: 에볼라 환자와 접촉한 사람에 대한 격리를 지지하는 뉴욕 성인의 비율 p에 대한 95% 신뢰구간을 구성하라.

풀이: 예제 5.12의 표준오차 $SE = 0.012$, 점추정값 0.82, 그리고 95% 신뢰수준에 대해 $z^* = 1.96$을 사용하여, 신뢰구간은:

$$\text{점추정값} \pm z^* \times SE \rightarrow 0.82 \pm 1.96 \times 0.012 \rightarrow (0.796, 0.844)$$

에볼라 환자와 접촉한 사람에 대한 격리를 지지하는 2014년 10월 뉴욕 성인의 비율이 **0.796과 0.844 사이**에 있다고 95% 확신한다.

Python 코드:

```python
import numpy as np

p_hat = 0.82
n = 1042
z_star = 1.96

se = np.sqrt(p_hat * (1 - p_hat) / n)
lower = p_hat - z_star * se
upper = p_hat + z_star * se

print(f"  : {p_hat}")
print(f"  크기: {n}")
print(f"  : {se:.4f}")
print(f"95% CI: ({lower:.3f}, {upper:.3f})")

# 출력:
# 점추정값: 0.82
# 표본  크기: 1042
# 표준오차: 0.0119
# 95% CI: (0.797, 0.843)
```

Guided Practice 5.14

문제: 예제 5.13의 신뢰구간에 대해 다음 두 질문에 답하라: (a) 이 맥락에서 95% 확신은 무엇을 의미하는가? (b) 오늘날 뉴욕 주민들의 의견에 대해 이 신뢰구간이 여전히 유효하다고 생각하는가?

풀이: (a) 많은 그러한 표본을 추출하고 각각에 대해 95% 신뢰구간을 계산했다면, 그 구간의 약 95%가 에볼라 환자와 접촉한 사람에 대한 격리를 지지하는 뉴욕 성인의 실제 비율을 포함할 것이다.

(b) 반드시 그렇지는 않다. 여론조사는 거대한 공중 보건 우려가 있던 시기에 실시되었다. 이제 사람들이 한 발 물러설 시간이 있었으므로, 그들은 의견을 바꿨을 수 있다. 그러한 격리 기간을 지지하는 뉴욕 성인의 현재 비율에 대한 추정값을 얻으려면 새로운 여론조사를 실시해야 할 것이다.

Guided Practice 5.15

문제: Pew Research의 태양 에너지 여론조사에서 다른 형태의 에너지에 대해서도 질문했고, 1000명의 응답자 중 84.8%가 풍력 터빈 사용 확대를 지지했다. (a) 풍력 터빈 확대를 지지하는 미국 성인의 비율을 정규분포를 사용하여 모델링하는 것이 합리적인가? (b) 전력 생산을 위한 풍력 터빈 사용 확대에 대한 미국인 지지 수준에 대한 99% 신뢰구간을 만들어라.

풀이: (a) 조사는 무작위 표본이고 두 수 모두 ≥ 10이다 ($1000 \times 0.848 = 848$과 $1000 \times 0.152 = 152$), 따라서 독립성과 성공-실패 조건이 만족되고, $\hat{p} = 0.848$은 정규분포를 사용하여 모델링될 수 있다.

(b) Guided Practice 5.15가 $\hat{p}$이 정규분포를 밀접하게 따름을 확인했으므로, CI 공식을 사용할 수 있다:

$$SE_{\hat{p}} = \sqrt{\frac{0.848(1-0.848)}{1000}} = 0.0114$$

99% 신뢰구간: $0.848 \pm 2.58 \times 0.0114 \rightarrow (0.8186, 0.8774)$

2018년에 풍력 터빈 사용 확대를 지지하는 미국 성인의 비율이 **81.9%와 87.7% 사이**라고 99% 확신한다.

5.2.5 신뢰구간 해석하기

각 예제에서 우리는 신뢰구간을 데이터의 맥락에 넣고 다소 형식적인 언어를 사용하여 설명했다:

태양열. 2018년에 미국 성인의 87.1%에서 90.4%가 태양열 발전 확대를 지지한다고 90% 확신한다.

에볼라. 에볼라 환자와 접촉한 사람에 대한 격리를 지지하는 2014년 10월 뉴욕 성인의 비율이 0.796과 0.844 사이라고 95% 확신한다.

풍력 터빈. 2018년에 풍력 터빈 사용 확대를 지지하는 미국 성인의 비율이 81.9%와 87.7% 사이라고 99% 확신한다.

먼저, 진술이 항상 **모집단 모수**에 대한 것임을 주목하라. 에너지 여론조사에서는 모든 미국 성인을 고려하고 격리 여론조사에서는 모든 뉴욕 성인을 고려한다.

우리는 또 다른 흔한 실수를 피했다: **잘못된 언어**는 신뢰구간이 특정 확률로 모집단 모수를 포착한다고 설명하려 할 수 있다. 확률 해석을 하는 것은 흔한 오류다: 확률로 생각하는 것이 유용할 수 있지만, 신뢰수준은 모수가 주어진 구간에 있을 가능성이 얼마나 그럴듯한지만 정량화한다.

신뢰구간의 또 다른 중요한 고려사항은 그들이 **모집단 모수에 대해서만**이라는 것이다. 신뢰구간은 개별 관측값이나 점추정값에 대해 아무것도 말하지 않는다. 신뢰구간은 모집단 모수에 대한 그럴듯한 범위만을 제공한다.

마지막으로, 우리가 논의한 방법은 **표본오차에만 적용되고, 편향에는 적용되지 않음**을 명심하라. 데이터 세트가 모집단 모수를 체계적으로 과소추정(또는 과대추정)하는 경향이 있는 방식으로 수집되었다면, 우리가 논의한 기법은 그 문제를 해결하지 않을 것이다. 대신, 우리는 편향에 대항하기 위해 데이터 과학자들이 사용하는 일반적인 관행인 신중한 데이터 수집 절차에 의존한다.

Guided Practice 5.16

문제: 태양 에너지 조사에 대한 90% 신뢰구간: 87.1%에서 90.4%를 고려하라. 조사를 다시 실시한다면, 새 조사의 비율이 87.1%에서 90.4% 사이에 있을 것이라고 90% 확신한다고 말할 수 있는가?

풀이: 아니오, 신뢰구간은 미래의 점추정값이 아니라 **모수**에 대한 그럴듯한 값의 범위만을 제공한다.

새로운 시각: 신뢰구간의 참된 의미

신뢰구간을 이해하는 가장 좋은 방법은 다음과 같다:

- 신뢰수준(95%)은 **방법**에 대한 것이지, 특정 구간에 대한 것이 아니다
- "이 방법으로 100번 구간을 만들면, 약 95번은 참값을 포함할 것이다"
- 특정 구간에 대해 "참값이 이 구간 안에 있을 확률이 95%다"라고 말하는 것은 **잘못된** 해석이다
- 왜냐하면 모수는 고정된 값이고, 특정 구간도 이미 계산된 고정된 범위이기 때문이다
- 모수는 구간 안에 있거나 없다 - 확률의 문제가 아니다!

5.2절 연습문제

연습문제 5.7 (만성 질환, Part I)

2013년에 Pew Research Foundation은 "미국 성인의 45%가 하나 이상의 만성 질환과 함께 살고 있다고 보고한다"고 보고했다. 그러나 이 값은 표본에 기반한 것이므로, 그 자체로는 관심 모집단 모수에 대한 완벽한 추정값이 아닐 수 있다. 연구는 약 1.2%의 표준오차를 보고했고, 이 설정에서 정규 모델이 합리적으로 사용될 수 있다. 하나 이상의 만성 질환과 함께 사는 미국 성인 비율에 대한 95% 신뢰구간을 만들어라. 또한 연구의 맥락에서 신뢰구간을 해석하라.

풀이:

$$\text{점추정값} \pm z^* \times SE = 0.45 \pm 1.96 \times 0.012$$
$$= 0.45 \pm 0.0235 = (0.4265, 0.4735)$$

하나 이상의 만성 질환과 함께 사는 미국 성인의 비율이 **42.6%에서 47.4% 사이**라고 95% 확신한다.

Python 코드:

```python
import numpy as np

p_hat = 0.45
se = 0.012
z_star = 1.96

margin = z_star * se
lower = p_hat - margin
upper = p_hat + margin

print(f"95% CI: ({lower:.3f}, {upper:.3f})")
print(f"95% CI: ({lower*100:.1f}%, {upper*100:.1f}%)")

# 출력:
# 95% CI: (0.426, 0.474)
# 95% CI: (42.6%, 47.4%)
```

연습문제 5.9 (만성 질환, Part II)

연습문제 5.7의 정보를 사용하여, 다음 각 진술이 참인지 거짓인지 식별하라.

(a) 연습문제 5.7의 신뢰구간이 만성 질환을 앓고 있는 미국 성인의 실제 비율을 포함한다고 확실히 말할 수 있다.

거짓. 신뢰구간은 그럴듯한 값들의 범위를 제공하며, 때때로 참값을 놓친다. 95% 신뢰구간은 약 5%의 시간 동안 "놓친다".

(b) 이 연구를 1,000번 반복하고 각 연구에 대해 95% 신뢰구간을 구성한다면, 그 구간의 약 950개가 만성 질환을 앓고 있는 미국 성인의 실제 비율을 포함할 것이다.

참. 설명이 참인 모집단 값에 초점을 맞추고 있음에 주목하라.

(c) 여론조사는 만성 질환을 앓고 있는 미국 성인의 비율이 50% 미만이라는 통계적으로 유의한 증거를 ($\alpha = 0.05$ 수준에서) 제공한다.

참. 연습문제 5.7에서 계산한 95% 신뢰구간을 살펴보면, 50%가 이 구간에 포함되지 않음을 알 수 있다. 이것은 가설검정에서 비율이 0.5라는 귀무가설을 기각할 것임을 의미한다.

(d) 표준오차가 1.2%이므로, 연구의 사람들 중 1.2%만이 답변에 대한 불확실성을 전달했다.

거짓. 표준오차는 무작위성으로 인한 자연적인 변동으로부터의 전체 추정값의 불확실성을 설명하며, 개인의 응답에 해당하는 불확실성이 아니다.

연습문제 5.11 (ER에서 대기)

응급실에서의 평균 대기 시간을 추정하기 위해 연구가 설계되었다. 64명 환자의 무작위 표본에서 평균 대기 시간은 137.5분이며, 그 추정값의 95% 신뢰구간은 (128, 147)분이었다.

다음 진술이 참인지 거짓인지 결정하라:

(a) 이 신뢰구간은 표본 평균이 128분과 147분 사이에 있음을 의미한다.

거짓. 점추정값은 항상 신뢰구간 안에 있으며, 이것은 점추정값과 신뢰구간의 무의미한 사용이다.

(b) 이 데이터의 95% 신뢰구간의 오차한계는 약 9.5분이다.

참. 오차한계 = $(147 - 128)/2 = 9.5$

(c) 무작위 표본의 95%는 표본 평균이 128분과 147분 사이에 있다.

거짓. 신뢰구간은 표본 평균에 대한 것이 아니다.

(d) 99% 신뢰구간은 추정에 더 확실해야 하므로 95% 신뢰구간보다 좁을 것이다.

거짓. 모수를 포착하는 것에 더 확신하려면 더 넓은 구간이 필요하다.

(e) 오차한계는 9.5이고 표본 평균은 137.5이다.

참. 선택적 설명: 정규 모델이 표본 평균을 모델링하는 데 사용되었기 때문에 이것은 참이다. 오차한계는 구간 너비의 절반이고, 표본 평균은 구간의 중점이다.

(f) 95% 신뢰수준에서 오차한계를 지금의 절반으로 줄이려면 표본 크기를 두 배로 해야 할 것이다.

거짓. 표준오차 계산에서 표본 크기의 제곱근으로 나눈다. SE(또는 오차한계)를 절반으로 줄이려면 초기 표본의 $2^2 = 4$배의 사람들을 표본 추출해야 할 것이다.

5.3 비율에 대한 가설검정

다음 질문은 Hans Rosling, Anna Rosling Ronnlund, Ola Rosling이 쓴 Factfulness라는 책에서 나온 것이다:

오늘날 세계의 1세 아이들 중 몇 퍼센트가 어떤 질병에 대해 예방접종을 받았는가? a. 20%, b. 50%, c. 80%

정답을 적어 두고, 준비가 되면 각주에서 답을 찾아보라. (정답: c. 80%)

이 절에서 우리는 대학 교육을 받은 사람들이 이 질문과 다른 세계 보건 질문에서 어떻게 수행하는지 탐구하면서 경쟁하는 아이디어와 주장을 엄격하게 평가하는 데 사용되는 프레임워크인 가설검정(hypothesis tests)에 대해 배울 것이다.

5.3.1 가설검정 프레임워크

2014년 Pew Research 조사에서 50명의 대학 교육을 받은 미국 성인에게 다음 세 가지 신생아 백신 접종 일정 중 가장 정확한 것을 고르라고 했다:

- 부모 선택에 따라 시간에 걸쳐 펼쳐야 한다 (응답자의 12% 선택)
- 의사가 권장하는 대로 빨리 맞아야 한다 (응답자의 24% 선택)
- 건강한 아기에게는 필요하지 않다 (응답자의 64% 선택)

실제로 정답은 "의사가 권장하는 대로 빨리 맞아야 한다"였고, 응답자의 24%만이 이 질문에 정답을 맞혔다. 이것이 의미하는 바는, 비록 대학 교육을 받은 성인들이지만 세 가지 선택지 중 하나를 무작위로 고른 것(33.3%)보다 더 잘하지 못했다는 것인가?

이 질문에 대답하려면 두 가지 경쟁하는 가설을 세울 수 있다:

- H_0 (귀무가설): 대학 교육을 받은 성인들은 무작위 추측으로 예상되는 것보다 더 잘하지 않았다. 실제 비율은 33.3%: $p = 0.333$
- H_A (대립가설): 대학 교육을 받은 성인들의 비율은 무작위 추측과 다르다: $p \neq 0.333$

이 설정은 법적 재판과 유사하다. 피고는 "무죄가 입증될 때까지 유죄가 아닌" 것으로 간주된다. 무죄 추정을 얻기 위해 피고는 자신의 무죄를 증명할 필요가 없다; 무죄는 피고에게 주어진다. 검찰 측은 유죄 판결을 얻기 위해 "합리적 의심을 넘어서" 피고의 유죄를 입증해야 한다. 마찬가지로, 가설검정에서 우리는 특별한 이유 없이 귀무가설 H_0를 기각하지 않는다. 귀무가설은 무죄 추정처럼 합리적인 의심의 여지없이 데이터가 충분한 증거를 제공할 때만 기각될 수 있다.

귀무가설과 대립가설

귀무가설(H_0): 종종 회의적인 시각 또는 아무 일도 일어나지 않았다는 주장을 나타낸다.
대립가설(H_A): 무언가 새로운 것이 일어나고 있다는 대안적 주장을 나타낸다.

예제 5.18

문제: 대학 교육을 받은 성인과 백신 접종에 대한 맥락에서 귀무가설과 대립가설을 말로 작성하라.

풀이:

H_0: 대학 교육을 받은 성인들이 올바른 백신 접종 일정을 식별할 확률은 무작위 추측(1/3)과 같다.

H_A: 대학 교육을 받은 성인들이 올바른 백신 접종 일정을 식별할 확률은 무작위 추측과 다르다.

예제 5.19

문제: 대학 교육을 받은 성인들이 무작위 추측보다 못한다고 생각되면, 왜 $H_A : p < 0.333$으로 대립가설을 설정하지 않았는가?

풀이: 대학 교육을 받은 성인들이 무작위 추측보다 못할 것이라는 주장은 검정을 시작하기 전에 만들어진 편향된 주장일 것이다. 이 맥락에서 귀무가설을 기각하고 비율이 무작위 추측과 다르다는 결론에 도달하면 그것만으로도 충분히 흥미로울 것이다. 단측 가설검정의 세부 사항은 5.3.7절에서 다룬다.

5.3.2 신뢰구간을 사용한 가설검정

표본에서 정답을 맞힌 비율은 $\hat{p} = 12/50 = 0.24$이고, 귀무가설 값은 $p_0 = 1/3 \approx 0.333$이다. 우리는 5.1절에서 표본마다 변동이 있다는 것을 배웠고, 표본비율 $\hat{p}$이 정확히 p와 같을 것 같지 않지만, p에 대한 결론을 내리고 싶다. 우리에게는 걱정거리가 있다: 이 24%와 33.3%의 차이는 단순히 우연에 의한 것인가, 아니면 데이터가 모집단비율이 33.3%와 다르다는 강력한 증거를 제공하는가?

5.2절에서 우리는 신뢰구간을 사용하여 추정의 불확실성을 정량화하는 방법을 배웠다. 변동성을 측정하는 같은 방법이 가설검정에 유용할 수 있다.

예제 5.20

문제: 표본 데이터를 사용하여 p에 대한 신뢰구간을 구성하는 것이 합리적인지 확인하고, 그렇다면 95% 신뢰구간을 구성하라.

풀이: 조건은 $\hat{p}$이 근사적으로 정규분포가 되도록 충족된다: 데이터는 단순무작위표본에서 왔으므로(독립성 만족), $n\hat{p} = 12$와 $n(1-\hat{p}) = 38$이 모두 최소 10이다(성공-실패 조건).

신뢰구간을 구성하려면 점추정값($\hat{p} = 0.24$), 95% 신뢰수준에 대한 임계값($z^* = 1.96$), 그리고 $\hat{p}$의 표준오차($SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.060$)를 식별해야 한다. 이 조각들로 p에 대한 신뢰구간을 구성할 수 있다:

$$\hat{p} \pm z^* \times SE_{\hat{p}}$$
$$0.24 \pm 1.96 \times 0.060$$
$$(0.122, 0.358)$$

우리는 이 특정 영아 백신 접종 질문에 올바르게 대답하는 모든 대학 교육을 받은 성인의 비율이 12.2%에서 35.8% 사이에 있다고 95% 확신한다.

가설검정의 귀무값 $p_0 = 0.333$이 신뢰구간의 그럴듯한 값 범위 내에 있으므로, 귀무값이 그럴듯하지 않다고 말할 수 없다. 즉, 데이터는 대학 교육을 받은 성인의 성과가 무작위 추측과 다르다는 충분한 증거를 제공하지 않으며, **귀무가설 H_0를 기각하지 않는다**.

Python 코드:

```python
import numpy as np

p_hat = 0.24
n = 50
p0 = 1/3  # 귀무값

se = np.sqrt(p_hat * (1 - p_hat) / n)
z_star = 1.96

lower = p_hat - z_star * se
upper = p_hat + z_star * se

print(f"  : {p_hat}")
print(f"  : {se:.4f}")
print(f"95% CI: ({lower:.3f}, {upper:.3f})")
print(f"  {p0:.3f}이  구간  내에  있는가? {lower <= p0 <= upper}")

# 출력:
# 표본비율: 0.24
# 표준오차: 0.0604
# 95% CI: (0.122, 0.358)
# 귀무값  0.333이  구간  내에  있는가? True
```

예제 5.21

문제: 대학 교육을 받은 성인들이 영아 백신 접종 질문에서 단순히 추측했다고 결론지을 수 없는 이유를 설명하라.

풀이: H_0를 기각하지 못했지만, 이것이 귀무가설이 반드시 참이라는 것을 의미하지는 않는다. 아마도 실제 차이가 있었지만, 비교적 작은 50명의 표본으로는 이를 감지할 수 없었을 수 있다.

통계학에서 이중 부정이 때때로 사용된다

많은 통계적 설명에서 우리는 이중 부정을 사용한다. 예를 들어, 귀무가설이 그럴듯하지 않다고 말할 수 없다 또는 귀무가설을 기각하지 못했다라고 말할 수 있다. 이중 부정은 우리가 어떤 입장을 기각하지 않지만, 그것이 맞다고도 말하지 않는다는 것을 전달하기 위해 사용된다.

5.3.3 의사결정 오류

가설검정은 완벽하지 않다: 데이터에 기반한 통계적 가설검정에서 잘못된 결정을 내릴 수 있다. 예를 들어, 법원 시스템에서 무고한 사람들이 때때로 부당하게 유죄 판결을 받고 유죄인 사람들이 때때로 무죄로 풀려난다. 통계적 가설검정의 한 가지 핵심 구별점은 우리가 결론에서 얼마나 자주 오류를 범하는지 확률적으로 정량화하는 데 필요한 도구를 가지고 있다는 것이다.

두 가지 경쟁하는 가설이 있음을 상기하라: 귀무가설과 대립가설. 가설검정에서 우리는 어느 것이 참일 수 있는지에 대한 진술을 하지만, 잘못 선택할 수 있다. 네 가지 가능한 시나리오가 있으며, 이는 그림 5.8에 요약되어 있다.

	H_0 참	H_A 참
H_0 **기각하지 않음**	올바른 결정	제2종 오류
H_0 **기각**	제1종 오류	올바른 결정

제1종 오류(Type 1 Error)는 H_0가 실제로 참일 때 귀무가설을 기각하는 것이다.

제2종 오류(Type 2 Error)는 대립가설이 실제로 참일 때 귀무가설을 기각하지 못하는 것이다.

Guided Practice 5.25

문제: 미국 법원에서 피고는 무죄(H_0)이거나 유죄(H_A)다. 이 맥락에서 제1종 오류는 무엇을 나타내는가? 제2종 오류는 무엇을 나타내는가?

풀이:

제1종 오류: 법원이 제1종 오류를 범하면, 이는 피고가 무죄(H_0 참)이지만 부당하게 유죄 판결을 받았음을 의미한다. 제1종 오류는 귀무가설을 기각했을 때만 가능하다.

제2종 오류: 법원이 H_0를 기각하지 못했지만(즉, 사람을 유죄 판결하지 못했지만) 그녀가 실제로 유죄(H_A 참)였을 때를 의미한다. 제2종 오류는 귀무가설을 기각하지 못했을 때만 가능하다.

예제 5.26

문제: 미국 법원에서 제1종 오류율을 어떻게 줄일 수 있는가? 이것이 제2종 오류율에 어떤 영향을 미치겠는가?

풀이: 제1종 오류율을 낮추려면, 유죄 판결 기준을 "합리적 의심을 넘어서"에서 "생각할 수 있는 의심을 넘어서"로 높여 더 적은 사람이 부당하게 유죄 판결을 받도록 할 수 있다. 그러나 이것은 또한 실제로 유죄인 사람들을 유죄 판결하는 것을 더 어렵게 만들어, 더 많은 제2종 오류를 범하게 될 것이다.

Guided Practice 5.27

문제: 미국 법원에서 제2종 오류율을 어떻게 줄일 수 있는가? 이것이 제1종 오류율에 어떤 영향을 미치겠는가?

풀이: 제2종 오류율을 낮추려면, 더 많은 유죄인 사람들을 유죄 판결하고 싶다. 유죄 판결 기준을 "합리적 의심을 넘어서"에서 "약간의 의심을 넘어서"로 낮출 수 있다. 유죄에 대한 기준을 낮추면 더 많은 부당한 유죄 판결도 발생하여 제1종 오류율이 상승한다.

Guided Practice 5.25-5.27은 중요한 교훈을 제공한다: **한 유형의 오류를 덜 범하도록 줄이면, 일반적으로 다른 유형의 오류를 더 범하게 된다**.

가설검정은 귀무가설을 기각하거나 기각하지 못하는 것을 중심으로 구축된다. 즉, 강력한 증거가 없으면 H_0를 기각하지 않는다. 그런데 정확히 강력한 증거는 무엇을 의미하는가? 일반적인 경험 법칙으로, 귀무가설이 실제로 참인 경우, 우리는 5%의 시간 이상으로 H_0를 잘못 기각하고 싶지 않다. 이것은 **유의수준(significance level)**0.05에 해당한다. 즉, 귀무가설이 참이면, 유의수준은 데이터가 H_0를 잘못 기각하도록 이끄는 빈도를 나타낸다. 유의수준은 종종 α(그리스 문자 알파)를 사용하여 쓴다: $\alpha = 0.05$. 다른 유의수준의 적절성은 5.3.5절에서 논의한다.

95% 신뢰구간을 사용하여 가설검정을 평가하고 귀무가설이 참인 경우, 점추정값이 모집단 모수에서 최소 1.96 표준오차 떨어져 있을 때마다 오류를 범할 것이다. 이것은 약 5%의 시간(각 꼬리에 2.5%)에 발생한다. 마찬가지로, 99% 신뢰구간을 사용하여 가설을 평가하는 것은 유의수준 $\alpha = 0.01$과 동등하다.

5.3.4 p-값을 사용한 공식적인 검정

p-값(p-value)은 귀무가설에 대한 증거와 대립가설에 유리한 증거의 강도를 정량화하는 방법이다. 통계적 가설검정은 일반적으로 신뢰구간에 기반한 결정보다 p-값 방법을 사용한다.

p-값 정의

p-값은 귀무가설이 참이라면, 현재 데이터 세트만큼 대립가설에 유리하거나 더 유리한 데이터를 관찰할 확률이다. 우리는 일반적으로 데이터의 요약 통계량(이 절에서는 표본비율)을 사용하여 p-값을 계산하고 가설을 평가한다.

예제 5.28

문제: Pew Research는 1000명의 미국 성인 무작위 표본에게 에너지 생산을 위한 석탄 사용 확대를 지지하는지 질문했다. 표본에서 37%만이 석탄 에너지 역할 확대를 지지했다. 이것이 미국 성인의 과반수가 석탄 에너지 확대에 찬성하지 않는다는 강력한 증거를 제공하는가?

풀이:

1단계 - 가설 설정: - $H_0 : p = 0.5$ (지지와 반대가 동등) - $H_A : p \neq 0.5$ (지지가 50%와 다름)

2단계 - 조건 확인:

독립성: 표본은 무작위 표본이므로 독립이다.

성공-실패 조건: 단일 비율 가설검정에서 이 조건은 귀무 비율로 확인한다: $np_0 = n(1-p_0) = 1000 \times 0.5 = 500 \geq 10$.

3단계 - 표준오차 계산 (귀무값 사용):

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5 \times 0.5}{1000}} = 0.0158$$

4단계 - 검정통계량(Z-점수) 계산:

$$Z = \frac{\text{점추정값} - \text{귀무값}}{SE} = \frac{0.37 - 0.50}{0.0158} = -8.23$$

5단계 - p-값 계산:

우리의 점추정값이 귀무값과 정확히 같다면, 검정통계량은 $Z = 0$이 될 것이다. Z-점수가 0에서 더 멀어질수록 귀무가설에 대한 증거가 더 강해진다. $Z = -8.23$의 테스트 통계량을 가졌으므로, 꼬리 면적을 계산한다. 양측 검정이므로 양쪽 꼬리 면적을 더한다.

p-값 $\approx 2 \times P(Z < -8.23) \approx 0$ (사실상 0에 가까움)

6단계 - 결론: p-값이 유의수준 $\alpha = 0.05$보다 훨씬 작으므로, **귀무가설을 기각한다**. 여론조사는 미국 성인의 과반수가 석탄 에너지 확대를 지지하지 않는다는 설득력 있는 증거를 제공한다.

Python 코드:

```python
import numpy as np
from scipy import stats

p_hat = 0.37
p0 = 0.5
n = 1000
alpha = 0.05

# 귀무가설 하에서 표준오차
se = np.sqrt(p0 * (1 - p0) / n)
print(f"  : {se:.4f}")

# Z-점수
z = (p_hat - p0) / se
print(f"Z-점수: {z:.2f}")

# p-값 (양측)
p_value = 2 * stats.norm.cdf(z)  # z가 음수이므로 하단 꼬리 사용
print(f"p-값: {p_value:.2e}")

# 결론
if p_value < alpha:
    print(f"p-값 ({p_value:.2e}) <  ({alpha}): H 기각")
else:
    print(f"p-값 ({p_value:.2e}) >=  ({alpha}): H 기각 실패")

# 출력:
# 표준오차: 0.0158
# Z-점수: -8.23
# p-값: 1.85e-16
# p-값 (1.85e-16) <  (0.05): H 기각
```

p-값을 사용한 결론 도출

p-값 $< \alpha$**일 때**: H_0를 기각하고, 데이터가 대립가설을 지지하는 강력한 증거를 제공한다고 보고한다.
p-값 $\geq \alpha$**일 때**: H_0를 기각하지 않고, 귀무가설을 기각할 충분한 증거가 없다고 보고한다.
어느 경우든, 데이터의 맥락에서 결론을 설명하는 것이 중요하다.

Guided Practice 5.32

문제: 미국인의 다수가 핵무기 감축을 지지하는가 반대하는가? 이 질문을 평가하기 위한 가설을 세워라.

풀이: 다수가 지지하는지 반대하는지, 또는 궁극적으로 차이가 없는지 이해하고 싶다. p가 핵무기 감축을 지지하는 미국인의 비율이라면: - $H_0 : p = 0.50$ - $H_A : p \neq 0.50$

예제 5.33

문제: 2013년 3월 1028명의 미국 성인 단순무작위표본은 56%가 핵무기 감축을 지지함을 보여준다. 이것이 5% 유의수준에서 미국인의 다수가 핵무기 감축을 지지한다는 설득력 있는 증거를 제공하는가?

풀이:

먼저 조건을 확인한다:

독립성: 여론조사는 미국 성인의 단순무작위표본이었으므로 관측값들이 독립이다.

성공-실패 조건: 단일 비율 가설검정에서 이 조건은 귀무 비율로 확인한다: $np_0 = n(1-p_0) = 1028 \times 0.5 = 514 \geq 10$.

이 조건들이 확인되면, 정규 모델을 사용하여 $\hat{p}$을 모델링할 수 있다.

다음으로 표준오차를 계산할 수 있다. 단일 비율에 대한 가설검정이므로 귀무값 p_0가 다시 사용된다.

$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.0156$$

정규 모델에 기반하여 검정통계량은 점추정값의 Z-점수로 계산될 수 있다:

$$Z = \frac{\text{점추정값} - \text{귀무값}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.85$$

상위 꼬리 면적은 약 0.0001이고, 이 꼬리 면적을 두 배로 하여 p-값을 얻는다: 0.0002.

p-값이 0.05보다 작으므로 H_0를 기각한다. 여론조사는 **2013년 3월에 미국인의 다수가 핵무기 감축 노력을 지지했다는 설득력 있는 증거를 제공한다**.

Python 코드:

```python
import numpy as np
from scipy import stats

p_hat = 0.56
p0 = 0.5
n = 1028

se = np.sqrt(p0 * (1 - p0) / n)
z = (p_hat - p0) / se
p_value = 2 * (1 - stats.norm.cdf(abs(z)))

print(f"SE = {se:.4f}")
print(f"Z = {z:.2f}")
print(f"p-값  = {p_value:.5f}")
print(f" : {'H 기각' if p_value < 0.05 else 'H 기각 실패'}")

# 출력:
# SE = 0.0156
# Z = 3.85
# p-값  = 0.00012
# 결론: H 기각
```

단일 비율에 대한 가설검정 단계

단일 비율 가설검정이 올바른 절차라고 결정했다면, 검정을 완료하는 네 가지 단계가 있다:
준비(Prepare). 관심 모수를 식별하고, 가설을 나열하고, 유의수준을 식별하고, $\hat{p}$와 n을 식별한다.
확인(Check). H_0 하에서 $\hat{p}$이 거의 정규분포임을 보장하는 조건을 확인한다. 단일 비율 가설검정에서, 성공-실패 조건을 확인하기 위해 귀무값을 사용한다.
계산(Calculate). 조건이 충족되면, 다시 p_0을 사용하여 표준오차를 계산하고, Z-점수를 계산하고, p-값을 식별한다.
결론(Conclude). p-값을 α와 비교하여 가설검정을 평가하고, 문제의 맥락에서 결론을 제공한다.

5.3.5 유의수준 선택하기

검정에 대한 유의수준을 선택하는 것은 많은 맥락에서 중요하며, 전통적인 수준은 $\alpha = 0.05$이다. 그러나 적용 분야에 따라 유의수준을 조정하는 것이 도움이 될 수 있다. 검정에서 도달한 결론의 결과에 따라 0.05보다 작거나 큰 수준을 선택할 수 있다.

제1종 오류가 위험하거나 특히 비용이 많이 드는 경우, 작은 유의수준(예: 0.01)을 선택해야 한다. 이 시나리오에서는 귀무가설을 기각하는 것에 대해 매우 신중하고 싶으므로, H_0를 기각하기 전에 H_A를 매우 강력하게 지지하는 증거를 요구한다.

제2종 오류가 제1종 오류보다 상대적으로 더 위험하거나 훨씬 더 비용이 많이 드는 경우, 더 높은 유의수준(예: 0.10)을 선택할 수 있다. 여기서는 대립가설이 실제로 참일 때 H_0를 기각하지 못하는 것에 대해 신중하고 싶다.

예제 5.34

문제: 자동차 제조업체가 차량 도어 힌지를 만드는 새로운 고품질 장비로 전환을 고려하고 있다. 이 새 기계가 0.2% 미만의 비율로 결함이 있는 힌지를 생산하면 장기적으로 비용을 절약할 것이다. 그러나 힌지 결함 비율이 0.2%보다 높으면 투자 수익률이 충분하지 않아 손실을 볼 것이다. 이러한 가설검정에서 유의수준을 수정할 좋은 이유가 있는가?

풀이: 귀무가설은 결함 힌지 비율이 0.2%이고, 대립가설은 비율이 0.2%와 다르다는 것이다. 이 결정은 자동차와 회사에 한계적인 영향을 미치는 많은 결정 중 하나일 뿐이다. 제1종 오류나 제2종 오류 모두 위험하거나 (상대적으로) 훨씬 더 비용이 많이 들지 않으므로 **유의수준 0.05가 합리적**으로 보인다.

예제 5.35

문제: 같은 자동차 제조업체가 도어 힌지가 아닌 안전과 관련된 부품에 대해 약간 더 비싼 공급업체를 고려하고 있다. 이 안전 구성요소의 내구성이 현재 공급업체보다 좋은 것으로 나타나면 제조업체를 전환할 것이다. 이러한 평가에서 유의수준을 수정할 좋은 이유가 있는가?

풀이: 귀무가설은 공급업체의 부품이 동등하게 신뢰할 수 있다는 것이다. 안전이 관련되어 있기 때문에, 자동차 회사는 증거가 적당히 강하더라도 약간 더 비싼 제조업체로 전환하기를(H_0 기각) 열망해야 한다. 약간 **더 큰 유의수준(예: $\alpha = 0.10$)**이 적절할 수 있다.

Guided Practice 5.36

문제: 기계 내부의 부품은 교체하는 데 매우 비싸다. 그러나 기계는 이 부품이 고장 나더라도 보통 제대로 작동하므로, 측정 시리즈에 기반하여 고장났다고 극도로 확신할 때만 부품을 교체한다. 이 검정에 대한 적절한 가설을 (일반 언어로) 식별하고 적절한 유의수준을 제안하라.

풀이: 여기서 귀무가설은 부품이 고장 나지 않았다는 것이고, 대립가설은 고장났다는 것이다. H_0를 기각할 충분한 증거가 없으면 부품을 교체하지 않을 것이다. 고장났을 때 부품을 고치지 못하는 것(H_0 거짓, H_A 참)은 그다지 문제가 되지 않는 것 같고, 부품 교체는 비싸다. 따라서 부품을 교체하기 전에 H_0에 대한 매우 강력한 증거를 요구해야 한다. **작은 유의수준(예: $\alpha = 0.01$)**을 선택한다.

왜 0.05가 기본값인가?

$\alpha = 0.05$ 임계값이 가장 일반적이다. 하지만 왜 그럴까? 표준 수준이 더 작아야 하거나 더 커야 할 수도 있다. 약간 당혹스럽다면, 여분의 비판적인 눈으로 읽고 있는 것이다 - 잘하고 있다! 0.05가 왜인지 명확히 하는 데 도움이 되는 5분짜리 과제를 만들었다: www.openintro.org/why05

5.3.6 통계적 유의성 대 실질적 유의성

표본 크기가 더 커지면, 점추정값은 더 정확해지고 평균과 귀무값 사이의 실제 차이는 감지하고 인식하기 더 쉬워진다. 충분히 큰 표본을 취하면 매우 작은 차이도 감지될 수 있다. 때때로 연구자들은 매우 큰 표본을 추출하여 가장 작은 차이도 감지되고, 실질적 가치가 없는 차이도 감지된다. 그러한 경우, 우리는 여전히 차이가 **통계적으로 유의하다**고 말하지만, **실질적으로 유의하지는 않다**. 예를 들어, 온라인 실험은 영화 리뷰 웹사이트에 추가 광고를 배치하면 TV 쇼 시청률이 통계적으로 유의하게 0.001% 증가한다고 식별할 수 있지만, 이 증가는 실질적 가치가 없을 수 있다.

새로운 시각: 빅 데이터 시대의 주의점

데이터 과학에서 "빅 데이터"가 점점 보편화됨에 따라, 통계적 유의성과 실질적 유의성의 구분은 매우 중요해졌다:

- 수백만 명의 데이터를 가지고 있으면 거의 모든 작은 차이도 통계적으로 유의미해질 수 있다
- "통계적으로 유의미하다"와 "실제로 중요하다"는 매우 다른 질문이다
- 항상 **효과 크기(effect size)**를 검토하고 그 크기가 실제 응용에서 의미가 있는지 질문해야 한다
- p-값만 보고하지 말고, 신뢰구간이나 효과 크기도 함께 보고하는 것이 좋은 관행이다

5.3.7 단측 가설검정 (특별 주제)

지금까지 우리는 **양측 가설검정(two-sided hypothesis tests)**만 고려했는데, 여기서 우리는 p가 어떤 귀무값 p_0 보다 위인지 아래인지 감지하는 데 관심이 있다. **단측 가설검정(one-sided hypothesis test)**이라 불리는 두 번째 유형의 가설검정이 있다. 단측 가설검정에서 가설은 다음 형태 중 하나를 취한다:

1. 모집단 모수가 어떤 값 p_0 보다 작은지 감지하는 데만 가치가 있다. 이 경우, 대립가설은 어떤 귀무값 p_0 에 대해 $p < p_0$ 로 작성된다.
2. 모집단 모수가 어떤 값 p_0 보다 큰지 감지하는 데만 가치가 있다. 이 경우, 대립가설은 $p > p_0$ 로 작성된다.

대립가설의 형태를 조정하지만, 단측 가설검정의 경우에도 등호를 사용하여 귀무가설을 계속 작성한다.

전체 가설검정 절차에서 **단측 가설검정과 양측 가설검정을 평가하는 데 단 하나의 차이**가 있다: p-값을 계산하는 방법. 단측 가설검정에서, 우리는 대립가설 방향으로만 꼬리 면적으로 p-값을 계산한다, 즉 단일 꼬리 면적으로 표현된다.

단측 검정을 언제 사용할까? 매우 드물다. 단측 검정을 사용하는 것을 고려한다면, 다음 질문에 신중하게 답하라:

데이터가 내 대립가설의 반대 방향으로 명확하게 간다면, 나 또는 다른 사람들은 무엇을 결론지을까?

단측 검정의 반대 방향으로 가는 데이터에 대해 결론을 내리는 데 어떤 가치를 찾는다면, 양측 가설검정이 실제로 사용되어야 한다. 이러한 고려 사항은 미묘할 수 있으므로 주의를 기울여라. 우리는 이 책의 나머지 부분에서 양측 검정만 적용할 것이다.

예제 5.38

문제: 왜 데이터 방향으로 가는 단측 검정을 단순히 실행할 수 없는가?

풀이: 우리는 제1종 오류를 통제하는 신중한 프레임워크를 구축해왔는데, 이것은 가설검정에서 유의수준 α이다. 간단하게 하기 위해 아래에서 $\alpha = 0.05$를 사용할 것이다.

데이터를 본 후에 단측 검정을 선택할 수 있다고 상상해보자. 무엇이 잘못될까?

- $\hat{p}$이 귀무값보다 작으면, $p < p_0$ 인 단측 검정은 귀무 분포의 하위 5% 꼬리에 있는 관측값이 H_0 를 기각하도록 이끌 것을 의미한다.
- $\hat{p}$이 귀무값보다 크면, $p > p_0$ 인 단측 검정은 귀무 분포의 상위 5% 꼬리에 있는 관측값이 H_0 를 기각하도록 이끌 것을 의미한다.

그러면 H_0가 참이면, 두 꼬리 중 하나에 있을 확률이 10%이므로, 우리의 검정 오류는 실제로 $\alpha = 0.10$이고 0.05가 아니다. 즉, 단측 검정을 언제 사용할지 신중하지 않으면 우리가 열심히 개발하고 활용하려는 방법을 효과적으로 훼손한다.

5.3절 연습문제

연습문제 5.21 (최저임금, Part I)

미국 성인의 다수가 최저임금 인상이 경제에 도움이 될 것이라고 믿는가, 아니면 그렇지 않은 다수가 있는가? Rasmussen Reports의 1,000명 미국 성인 조사에서 42%가 경제에 도움이 될 것이라고 믿는 것으로 나타났다. 적절한 가설검정을 수행하여 연구 질문에 답하라.

풀이:

(i) 가설 설정: - $H_0 : p = 0.5$ (지지와 반대가 동등) - $H_A : p \neq 0.5$

유의수준 $\alpha = 0.05$ 사용

(ii) 조건 확인: - 독립성: 단순무작위표본이므로 독립 획득 - 성공-실패: $0.5 \times 1000 = 500$이 각 그룹에 대해 최소 10

(iii) 계산:

$$SE = \sqrt{\frac{0.5(1-0.5)}{1000}} = 0.0158$$

$$Z = \frac{0.42 - 0.5}{0.0158} = -5.06$$

단측 면적 약 0.0000003, p-값 = 이 단측 면적의 두 배 = **0.0000006**

(iv) 결론: p-값이 $\alpha = 0.05$보다 작으므로 귀무가설을 기각하고, 최저임금 인상이 경제에 도움이 될 것이라고 믿는 미국 성인의 비율이 50%가 아니라고 결론짓는다. 관찰값이 50%보다 낮고 귀무가설을 기각했으므로, 이 믿음은 **미국 성인의 50% 미만**이 가지고 있다고 결론지을 수 있다.

Python 코드:

```python
import numpy as np
from scipy import stats

p_hat = 0.42
p0 = 0.5
n = 1000

se = np.sqrt(p0 * (1 - p0) / n)
z = (p_hat - p0) / se
p_value = 2 * stats.norm.cdf(z)  # z가 음수

print(f"SE = {se:.4f}")
print(f"Z = {z:.2f}")
print(f"p-값 = {p_value:.2e}")

# 출력:
# SE = 0.0158
# Z = -5.06
# p-값 = 4.19e-07
```

연습문제 5.23 (역으로 계산하기)

다음 가설이 주어졌다: $H_0 : p = 0.3, H_A : p \neq 0.3$

표본 크기가 90임을 알고 있다. 추론에 필요한 모든 조건이 만족된다고 가정한다. **p-값이 0.05가 되는 표본비율**은 얼마인가?

풀이:

p-값이 0.05이면, 검정통계량은 $Z = 1.96$ 또는 $Z = -1.96$이다.

표준오차: $SE = \sqrt{\frac{0.3 \times 0.7}{90}} = 0.048$

$Z = 1.96$인 경우:

$$1.96 = \frac{\hat{p} - 0.3}{0.048}$$

$$\hat{p} = 0.3 + 1.96 \times 0.048 = 0.394$$

$Z = -1.96$인 경우:

$$-1.96 = \frac{\hat{p} - 0.3}{0.048}$$

$$\hat{p} = 0.3 - 1.96 \times 0.048 = 0.206$$

답: $\hat{p} = 0.394$ 또는 $\hat{p} = 0.206$

연습문제 5.25 (섬유근육통 검정)

Diana라는 환자가 만성 통증 증후군인 섬유근육통 진단을 받고 항우울제를 처방받았다. 회의적인 그녀는 처음에 항우울제가 증상에 도움이 될 것이라고 믿지 않았다. 그러나 약을 복용한 지 몇 달 후 증상이 실제로 나아지고 있다고 느끼기 때문에 항우울제가 효과가 있다고 결정한다.

(a) 항우울제를 복용하기 시작했을 때 Diana의 회의적인 입장에 대한 가설을 작성하라.

H_0: 항우울제는 섬유근육통 증상에 영향을 미치지 않는다. H_A: 항우울제는 섬유근육통 증상에 영향을 미친다 (도움이 되거나 해가 됨).

(b) 이 맥락에서 제1종 오류는 무엇인가?

항우울제가 실제로 효과가 없는데, Diana가(또는 실험이) 효과가 있다고 결론짓는 것.

(c) 이 맥락에서 제2종 오류는 무엇인가?

항우울제가 실제로 효과가 있는데, Diana가(또는 실험이) 효과가 없다고 결론짓는 것.

연습문제 5.32 (근시)

근시는 모든 어린이의 약 8%에게 영향을 미친다고 알려져 있다. 194명의 어린이 무작위 표본에서 21명이 근시였다. 이 데이터가 8% 값이 정확하지 않다는 증거를 제공하는지 가설검정을 수행하라.

풀이:

가설: - $H_0 : p = 0.08$ - $H_A : p \neq 0.08$

조건 확인: - 독립성: 무작위 표본 - 성공-실패: $np_0 = 194 \times 0.08 = 15.5 \geq 10, n(1 - p_0) = 178.5 \geq 10$

계산:

$$\hat{p} = \frac{21}{194} = 0.108$$

$$SE = \sqrt{\frac{0.08 \times 0.92}{194}} = 0.0195$$

$$Z = \frac{0.108 - 0.08}{0.0195} = 1.44$$

$$\text{p-값} = 2 \times P(Z > 1.44) = 2 \times 0.0749 = 0.150$$

결론: p-값 = 0.150 > 0.05이므로 H_0를 기각하지 않는다. 8%가 정확하지 않다는 **충분한 증거가 없다**.

Python 코드:

```python
import numpy as np
from scipy import stats

p_hat = 21 / 194
p0 = 0.08
n = 194

se = np.sqrt(p0 * (1 - p0) / n)
z = (p_hat - p0) / se
p_value = 2 * (1 - stats.norm.cdf(abs(z)))

print(f"  : {p_hat:.3f}")
print(f"SE = {se:.4f}")
print(f"Z = {z:.2f}")
print(f"p-값  = {p_value:.3f}")
print(f" : {'H 기각' if p_value < 0.05 else 'H 기각  실패'}")

# 출력:
# 표본비율: 0.108
# SE = 0.0195
# Z = 1.44
# p-값  = 0.150
# 결론: H 기각  실패
```

연습문제 5.35 (실질적 유의성 vs 통계적 유의성)

다음 진술이 참인지 거짓인지 결정하고 추론을 설명하라: “표본 크기가 크면, 귀무값과 관찰된 점추정 사이의 작은 차이도 통계적으로 유의미할 수 있다.”

풀이: **참**이다.

표본 크기 n이 증가하면 표준오차 $SE = \sqrt{p(1-p)/n}$이 감소한다. 표준오차가 작아지면 같은 크기의 차이에 대해 Z-점수가 더 커지고, 따라서 p-값이 작아진다. 결과적으로 실질적으로 의미 없는 작은 차이도 통계적으로 유의미해질 수 있다.

이것이 통계적 유의성과 실질적 유의성을 구분해야 하는 이유다.