

9.4 검정과 유의확률

검정(testing)은 데이터 뒤에 숨어있는 확률변수의 분포에 대한 가설이 맞는지 틀리는지 정량적으로 증명하는 작업이다. 예를 들어 다음과 같은 문제는 검정 방법론을 사용하여 접근할 수 있다.

예제

어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다. 이 동전은 조작되지 않은 공정한 동전이라고 할 수 있는가?

동전을 던져 앞면이 나오는 것을 베르누이분포 확률변수로 모형화하자. 모수를 추정하면 $\frac{12}{15} = 0.8$ 로 공정한 동전과는 거리가 멀어보인다. 하지만 이 모수는 추정값일뿐이므로 정말 이 동전이 조작된 동전이라고 할 수는 없다. 어떻게 하면 이 동전이 조작되었다는 혹은 조작되지 않았지만 어쩌다 우연히 저런 결과가 나왔다는 주장을 증명할 수 있을까?

예제

어떤 주식의 일주일 수익률은 다음과 같다.

-2.5%, -5%, 4.3%, -3.7%, -5.6%

이 주식은 장기적으로 수익을 가져다 줄 것인가 아니면 손실을 가져다 줄 것인가?

주식의 수익률을 정규분포 표본이라고 가정하자. 만약 해당 주식에 대한 정규분포의 기댓값이 양수라면 장기적으로 수익을 가져다주는 주식일 것이고 반대로 정규분포의 기댓값이 음수라면 장기적으로 손실을 가져다주는 주식일 것이다. 현재까지 나온 데이터를 사용하여 기댓값을 추정하면 약 -2.5%로 손실일 가능성이 있다. 하지만 이 모수 역시 단순한 추정치일 뿐이다. 어떻게 하면 주식이 음수의 기대수익률을가지고 있다는 것은 확인할 지 아니면 그렇지 않지만 우연히 저러 데이터가 나온 것인지 확인할 수 있을까?

가설과 검정

데이터를 특정한 확률분포를 가진 확률변수로 모형화하면 모수를 추정할 수 있다. 다음 작업으로는 데이터 뒤에 숨어있는 확률변수가 정말로 그 모숫값을 가졌는지 검증해보아야 한다. 다른 말로 하면 해당 확률변수가 그 모숫값을 가졌다는 주장을 논리적으로 증명해야 한다.

확률분포에 대한 어떤 주장을 **가설(hypothesis)**이라고 하며 H 로 표기한다. 이 가설을 증명하는 행위를 **통계적 가설검정(statistical hypothesis testing)** 줄여서 **검정(testing)**이라고 한다. 특히 확률분포의 모숫값이 특정한 값을 가진다는 가설을 검정하는 것을 **모수 검정(parameter testing)**이라고 한다.

귀무가설

검정 작업을 하려면

- 데이터가 어떤 확률변수의 표본이라고 가정한다.
- 데이터를 만드는 확률변수가 따르는 확률분포의 모수 θ 의 값이 어떤 특정한 실숫값 θ_0 으로 고정되어 있다고 가정한다.

확률분포의 모수에 대한 가설을 **귀무가설(null hypothesis)**이라고 하며 H_0 로 표기한다. 귀무가설은 확률분포를 특정한 상태로 고정시켜야 하므로 반드시 **등식(equality)**으로 표현되어야 한다. 특정한 실숫값 θ_0 는 우리가 증명하고자 하는 가설에 대한 기준값이 되는 상수를 사용한다.

$$H_0 : \theta = \theta_0 \quad (9.4.1)$$

예제

동전이 공정하다는 귀무가설은 다음과 같이 표현할 수 있다. 동전의 면은 두가지뿐이므로 베르누이 확률변수로 대표한다. 공정한 동전이라면 앞면이 나올 확률과 뒷면이 나올 확률이 같으므로 모수 μ 의 값이 0.5이다.

$$H_0 : \mu = 0.5 \quad (9.4.2)$$

예제

주식의 수익률에 대한 귀무가설은 다음과 같이 표현할 수 있다. 주식의 수익률은 정규분포로 대표할 수 있다. 주식이 장기적으로 수익이 나는 경우는 정규분포의 기대값 모수 μ 가 양수인 경우다. 반대로 주식이 장기적으로 손실을 보는 경우는 정규분포의 기대값 모수 μ 가 음수인 경우다. 이 두가지를 나누는 기준값은 0이 된다. 따라서 귀무가설은 다음과 같다.

$$H_0 : \mu = 0 \quad (9.4.3)$$

대립가설

귀무가설은 등식을 사용하여 표현한 어떤 기준 상태일 뿐이고 우리가 주장하려는 혹은 반박하려는 가설이 아닐 수도 있다. 이때 귀무가설과 같이 고려하는 가설이 **대립가설(alternative hypothesis)**이다. 대립가설은 기호로 H_a 로 표기한다.

일반적으로 생물학적 약품, 화학적 약품 등의 제품을 연구 개발할 때 연구 중인 새 제품이 기존의 제품 성능보다 더 큰 성능을 보여주거나 불량률 등이 더 낮아진 것을 보이는 것이 목표인 경우가 많기 때문에 대립가설을 **연구가설(research hypothesis)**이라고도 한다. 이 경우 기존 제품의 성능 혹은 목표 성능을 귀무가설로 놓고 진실임을 증명하고자 하는 가설을 대립가설로 놓는 경우가 많다.

(1) 모수 θ 가 어떤 특정한 값 θ_0 가 아니라는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta \neq \theta_0 \quad (9.4.4)$$

그런데 모수 θ 가 어떤 특정한 값 θ_0 보다 크거나 혹은 작다는 것을 증명하고 싶다면 어떻게 할까? 이때도 귀무가설은 등식이여야 한다. 귀무가설이 등식이 아니면 이후에 이야기할 검정통계량 분포를 구하는 것이 불가능하기 때문이다.

(2) 만약 θ 가 θ_0 보다 크다는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta > \theta_0 \quad (9.4.5)$$

(3) 만약 θ 가 θ_0 보다 작다는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta < \theta_0 \quad (9.4.6)$$

여기에서 주의할 점은 귀무가설과 대립가설이 반드시 서로 여집합(complement)의 관계에 있을 필요는 없다는 점이다. (2)번과 (3)번의 경우에는 귀무가설이 맞다면 우리가 증명하고자 하는 대립가설은 틀린 것이 된다. 우리의 주장 즉, 대립가설이 맞다고 증명하려면 귀무가설이 틀렸다는 것을 증명하되, 대립가설이 맞는 방향으로 귀무가설이 틀렸다는 것을 증명하면 된다.

예제

동전이 공정하지 않다고 증명하고 싶은 경우에는 귀무가설과 대립가설을 다음처럼 놓을 수 있다.

$$H_0 : \mu = 0.5, \quad H_a : \mu \neq 0.5 \quad (9.4.7)$$

이 주장을 증명하려면 귀무가설이 틀렸다는 증거가 있어야 한다.

예제

동전의 앞면이 뒷면보다 더 많이 나온다는 주장을 증명하고 싶은 경우에는 귀무가설과 대립가설을 다음처럼 놓을 수 있다.

$$H_0 : \mu = 0.5, \quad H_a : \mu > 0.5 \quad (9.4.8)$$

이 주장을 증명하려면 단순히 귀무가설이 틀렸다는 증거가 아니라 대립가설이 맞으면서 귀무가설이 틀렸다는 증거가 필요하다.

연습 문제 9.4.1

어떤 인터넷 쇼핑몰의 상품에 상품평이 있고 각 상품평이 '좋아요' 또는 '싫어요'다.

- (1) 이 상품이 좋은 상품인지 아닌지 어떤 확률분포로 모형화할 수 있는가?
- (2) 이 상품이 좋다는 주장을 하려면 귀무가설과 대립가설이 어떻게 되는가?

검정통계량

귀무가설이 맞거나 틀렸다는 것을 증명하려면 어떤 증거가 있어야 한다. 예를 들어보자.

- '어떤 병에 걸렸다'라는 가설을 증명하려면 환자의 혈액을 채취하여 혈액 내의 특정한 성분의 수치를 측정해야 한다고 가정하자. 이때 해당 수치가 바로 검정통계량이 된다.
- '어떤 학생이 우등 상장을 받을 수 있는 우등생이다'라는 가설을 증명하려면 시험 성적을 측정하면 된다. 이 시험 성적을 검정통계량이라고 부를 수 있다.

이 증거에 해당하는 숫자가 검정통계량이다. 보통 기호 t 로 나타낸다. **검정통계량(test statistics)**은 표본 데이터 집합을 입력으로 계산되는 함수의 값이다.

$$t = f(x_1, x_2, \dots, x_N) \quad (9.4.9)$$

검정통계량은 확률변수 X 의 표본에서 계산된 함수의 값이므로 어떤 값이 나올지 정확하게 예측할 수 없다. 따라서 검정통계량 t 도 **검정통계량 확률변수 T** 라는 새로운 확률변수의 표본으로 볼 수 있다.

예를 들어 '어떤 병에 걸렸다'는 가설을 혈액 성분 수치로부터 판단하려면 병에 걸린 환자의 성분 수치가 어떤 분포를 따르는지 알 수 있어야 한다. 현실에서는 실제로 병에 걸린 다수의 환자의 혈액 성분 수치를 사용하여 검정통계량 분포를 구한다. 또한 '어떤 학생이 우등생이다'라는 가설을 시험 성적으로부터 판단하려면 우등생인 모든 학생의 시험 성적에 대한 분포를 구해야 한다.

입력 데이터가 되는 확률변수 X 의 확률분포함수 $p_X(x)$ 와 검정통계량 수식 $f(x)$ 가 이미 결정되어 있기 때문에 검정통계량 확률변수 T 의 확률분포함수 $p_T(t)$ 도 수식으로 유도할 수 있다.

$$\left\{ \begin{array}{c} p_X(x) \\ f(x_1, x_2, \dots, x_N) \end{array} \right\} \rightarrow p_T(t) \quad (9.4.10)$$

다만 이 유도 과정이 수학적으로 아주 어려운 작업이다. 우리가 직접 검정통계량 확률분포를 수학적으로 계산하는 것은 쉽지 않고 시뮬레이션을 사용하거나 통계학자들이 몇몇 특정한 수식 $f(x)$ 에 대해 미리 구해놓은 검정통계량 확률분포만 사용한다. 유용한 검정통계량 분포를 어떤 통계학자가 증명하게 되면 관례적으로 그 통계학자의 이름을 따서 해당 검정통계량의 이름을 만든다.

일반적으로 많이 사용되는 검정통계량에는 다음과 같은 것들이 있다.

예제 : 베르누이분포 확률변수

모수 μ 를 가지는 베르누이분포 확률변수에 대해서는 전체 시도 횟수 N 번 중 성공한 횟수 n 자체를 검정통계량으로 쓸 수 있다. 이 검정통계량은 자유도 N 과 모수 μ 를 가지는 이항 분포를 따른다.

$$x \sim \text{Bern}(x; \mu) \rightarrow t = \sum_{i=1}^N x \sim \text{Bin}(t; N, \mu) \quad (9.4.11)$$

예제 : 분산 σ^2 값을 알고 있는 정규분포 확률변수

분산 σ^2 의 값을 알고 있는 정규분포 확률변수에 대해서는 다음과 같이 표본평균 m 을 분산 σ 로 정규화 (nomarlize)한 값을 검정통계량으로 쓴다. 이 검정통계량은 표준정규분포를 따른다. 이 검정통계량은 특별히 z 라고 부른다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \rightarrow z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(z; 0, 1) \quad (9.4.12)$$

여기에서 \bar{x} 은 표본평균

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (9.4.13)$$

예제 : 분산 σ^2 값을 모르는 정규분포 확률변수

이번에는 분산 σ^2 의 값을 모르는 정규분포 확률변수를 고려하자.

모수 μ 에 대한 검정을 할 때는 다음과 같이 표본평균 m 을 표본분산 s 로 정규화한 값을 검정통계량으로 쓴다. 이 검정통계량은 자유도가 $N - 1$ 인 표준스튜던트 t 분포를 따른다. N 은 데이터의 수다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \rightarrow t = \frac{m - \mu}{\frac{s}{\sqrt{N}}} \sim t(t; 0, 1, N - 1) \quad (9.4.14)$$

여기에서 m 은 표본평균

$$m = \frac{1}{N} \sum_{i=1}^N x_i \quad (9.4.15)$$

s^2 은 표본분산이다.

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - m)^2 \quad (9.4.16)$$

분산 σ^2 에 대한 검정을 할 때는 다음과 같이 표본분산을 정규화(normalize)한 값을 검정통계량으로 쓴다. 이 검정통계량은 자유도가 $N - 1$ 인 카이제곱분포를 따른다. N 은 데이터의 수다.

$$x \sim \mathcal{N}(\mu, \sigma^2) \rightarrow t = (N-1) \frac{s^2}{\sigma^2} \sim \chi^2(t; N-1) \quad (9.4.17)$$

연습 문제 9.4.2

어떤 인터넷 쇼핑몰의 상품에 상품평이 있고 각 상품평이 '좋아요' 또는 '싫어요'다.

- (1) 이 상품이 좋다는 주장을 검정으로 증명하려고 한다면 검정통계량은 무엇인가?
- (2) 검정통계량의 분포는 어떤 분포인가?

유의확률

이제 우리는 두가지 정보를 알고 있다.

- 검정통계량이 따르는 검정통계량 t 의 확률분포 $p_T(x)$ 를 알고 있다
- 실제 데이터에 구한 검정통계량의 값 t_0 , 즉 확률분포 $p_T(x)$ 의 표본 1개를 가지고 있다.

만약 우리가 최초에 가정한 귀무가설이 사실이라면 실제 데이터에서 구한 검정통계량의 값은 검정통계량 확률분포를 따르고 있으므로 기댓값이나 모드값 근처의 값이 나왔을 것이다. 반대로 우리가 가정한 귀무가설이 사실이 아니라면 실제 데이터에서 구한 검정통계량의 값은 검정통계량에서 나오기 어려운 값이 나왔을 것이다

In [1]:

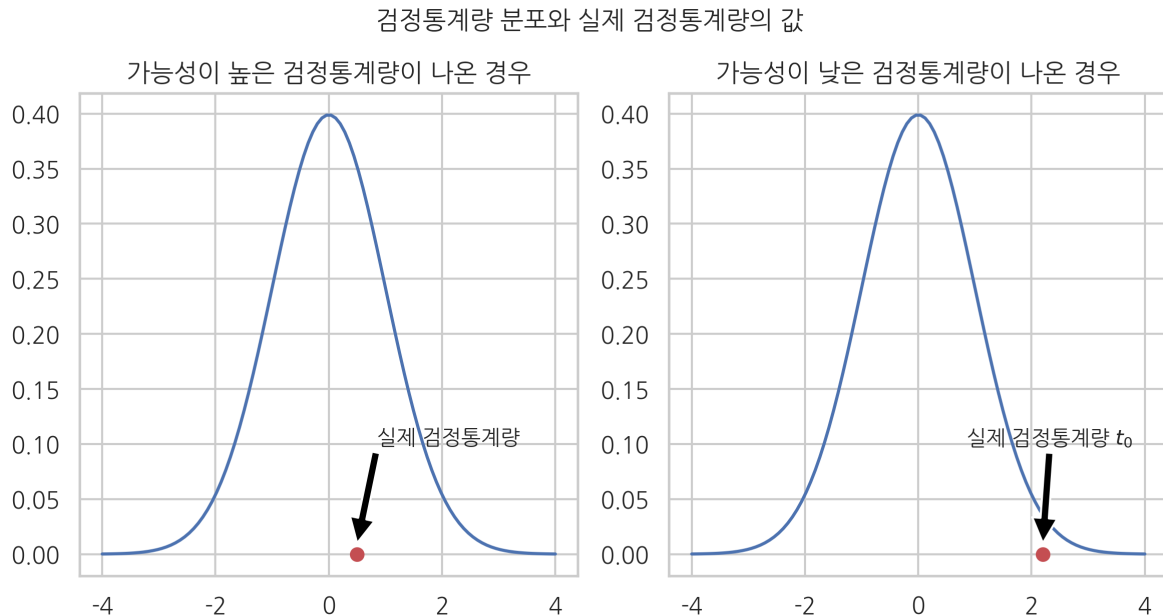
```
xx1 = np.linspace(-4, 4, 100)

black = {"facecolor": "black"}
plt.figure(figsize=(8, 4))

plt.subplot(121)
plt.title("가능성이 높은 검정통계량이 나온 경우")
plt.plot(xx1, sp.stats.norm.pdf(xx1))
plt.plot(0.5, 0, "ro")
plt.annotate('실제 검정통계량', xy=(0.5, 0.01), xytext=(0.85, 0.1), arrowprops=black)

plt.subplot(122)
plt.title("가능성이 낮은 검정통계량이 나온 경우")
plt.plot(xx1, sp.stats.norm.pdf(xx1))
plt.plot(2.2, 0, "ro")
plt.annotate('실제 검정통계량 $t_0$', xy=(2.2, 0.01), xytext=(0.85, 0.1), arrowprops=black)

plt.suptitle("검정통계량 분포와 실제 검정통계량의 값", y=1.05)
plt.tight_layout()
plt.show()
```



그러면 어떤 표본 데이터가 해당 확률분포에서 나오기 쉬운 값인지 나오기 어려운 값인지를 어떻게 숫자로 정량화할 수 있을까? 이 방법이 바로 유의확률(p-value)이다.

유의확률은 확률분포와 확률분포의 표본값 1개가 주어졌을 때 그 확률분포에서 해당 표본값 혹은 더 희귀한 (rare) 값이 나올 수 있는 확률로 정의한다.

유의확률의 값은 확률밀도함수에서 표본값을 기준으로 만들어진 양측 꼬리(tail)부분에 해당하는 영역의 면적이다.

누적확률분포함수 $F(x)$ 를 사용하여 다음처럼 계산할 수 있다.

$$\begin{aligned}\text{유의확률} &= 2 \min(P(t \leq t_0), P(t \geq t_0)) \\ &= 2 \min(F(t_0), 1 - F(t_0))\end{aligned}\tag{9.4.18}$$

이 식에서 t_0 는 현재 검정통계량의 값이다. 이 유의확률은 통계량분포의 양 끝단의 면적을 구하기 때문에 **양측검정 유의확률(two-sided test p-value)**이라고 한다.

만약 이산확률분포라면 등호가 성립하는 부분을 제외해야 하므로 다음처럼 구한다.

$$\text{유의확률} = 2 \min(F(t_0), 1 - F(t_0 - 1)) \quad (9.4.19)$$

In [2]:

```
xx1 = np.linspace(-4, 4, 100)

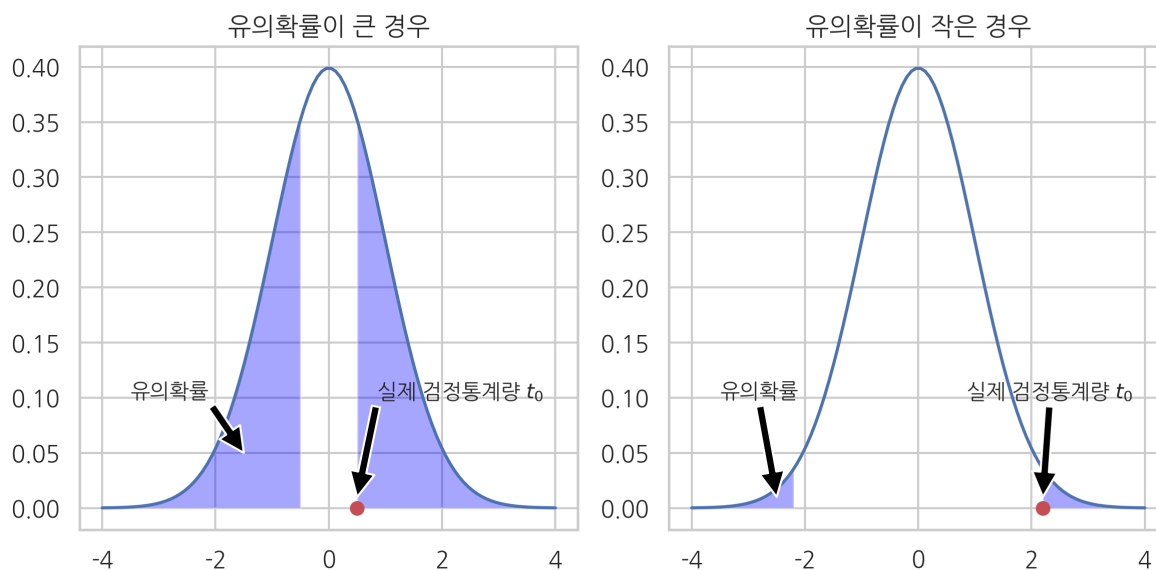
black = {"facecolor": "black"}
plt.figure(figsize=(8, 4))

plt.subplot(121)
plt.title("유의확률이 큰 경우")
plt.plot(xx1, sp.stats.norm.pdf(xx1))
plt.plot(0.5, 0, "ro")
plt.annotate('실제 검정통계량 $t_0$', xy=(0.5, 0.01), xytext=(0.85, 0.1), arrowprops=black)
xx2 = np.linspace(-4, -0.5, 100)
xx3 = np.linspace(0.5, 4, 100)
plt.fill_between(xx2, sp.stats.norm.pdf(xx2), facecolor='blue', alpha=0.35)
plt.fill_between(xx3, sp.stats.norm.pdf(xx3), facecolor='blue', alpha=0.35)
plt.annotate('유의확률', xy=(-1.5, 0.05), xytext=(-3.5, 0.1), arrowprops=black)

plt.subplot(122)
plt.title("유의확률이 작은 경우")
plt.plot(xx1, sp.stats.norm.pdf(xx1))
plt.plot(2.2, 0, "ro")
plt.annotate('실제 검정통계량 $t_0$', xy=(2.2, 0.01), xytext=(0.85, 0.1), arrowprops=black)
xx2 = np.linspace(-4, -2.2, 100)
xx3 = np.linspace(2.2, 4, 100)
plt.fill_between(xx2, sp.stats.norm.pdf(xx2), facecolor='blue', alpha=0.35)
plt.fill_between(xx3, sp.stats.norm.pdf(xx3), facecolor='blue', alpha=0.35)
plt.annotate('유의확률', xy=(-2.5, 0.01), xytext=(-3.5, 0.1), arrowprops=black)

plt.suptitle("검정통계량 분포와 실제 검정통계량 $t_0$의 값", y=1.05)
plt.tight_layout()
plt.show()
```

검정통계량 분포와 실제 검정통계량 t_0 의 값



검정의 관점에서 유의확률은

귀무가설이 맞음에도 불구하고 현재 검정통계량값과 같은 혹은 대립가설을 더 옹호하는 검정통계량값이 나올 확률

이라고 본다. 따라서 다음처럼 쓰기도 한다.

$$P(t \text{ for } H_a | H_0) \quad (9.4.20)$$

이 식에서 H_0 는 귀무가설이 진실인 사건을 뜻한다.

단측검정 유의확률

만약 증명하고자 하는 대립가설이 부등식인 경우에는 그 대립가설을 옹호하는 검정통계량값이 나올 확률을 구할 때 특정한 한 방향의 확률만을 구해야 한다. 이를 **단측검정(one-side test, single-tailed test)**이라고 한다.

모수 θ 가 특정한 값보다 크다는 것을 증명하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta > \theta_0 \quad (9.4.21)$$

반대로 θ 가 특정한 값보다 작다는 것을 검정하고 싶다면 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta < \theta_0 \quad (9.4.22)$$

단측검정의 유의확률은 다음과 같이 구한다. 여기에서는 모수가 클 때 검정통계량도 정비례해서 같이 커진다고 가정한다.

모수 θ 가 특정한 값보다 크다는 것을 증명하는 경우는 **우측검정(right-side test)** 유의확률을 사용한다. 우측검정 유의확률은 귀무가설이 맞음에도 불구하고 검정통계량이 현재 검정통계량과 같거나 더 큰 값이 나오는 확률이다.

$$\text{유의확률} = P(t \geq t_0 | H_0) = 1 - F(t_0) \quad (9.4.23)$$

만약 이산확률분포라면 등호가 성립하는 부분을 제외해야 하므로 다음처럼 구한다.

$$\text{유의확률} = 1 - F(t_0 - 1) \quad (9.4.24)$$

모수 θ 가 특정한 값보다 작다는 것을 증명하는 경우는 **좌측검정(left-tail test)** 유의확률을 사용한다. 좌측검정 유의확률은 귀무가설이 맞음에도 불구하고 실제로 나온 검정통계량과 같거나 더 작은 값이 나오는 확률이다.

$$\text{유의확률} = P(t \leq t_0 | H_0) = F(t_0) \quad (9.4.25)$$

좌측검정의 경우는 이산확률분포도 위 식과 같이 구한다.

예제

어떤 환자의 혈압이 고혈압이라는 것을 증명하고 싶을 때는 귀무가설과 대립가설을 다음과 같이 놓는다.

- 귀무가설: '혈압이 정상이다'
- 대립가설: '고혈압이다'

이 검정에서 혈압 검사 결과를 통계량 분포로, 해당 환자의 혈압을 검정통계량으로 사용하여 계산한 우측유의확률이 0.02%이 나왔다고 하자. 이는 정상인 중에서 혈압이 해당 환자의 혈압보다 더 높게 나온 사람은 0.02%뿐이었다는 뜻이다.

유의수준과 기각역

유의확률값이 아주 작다는 것은 귀무가설이 맞다는 가정하에 현재의 검정통계량값이 나올 가능성이 매우 적다는 의미다. 따라서 **유의확률값이 아주 작으면 귀무가설을 기각하고 대립가설을 채택할 수 있다.**

유의확률이 아주 작으면 즉, 귀무가설이 맞다는 가정하에 (귀무가설이 아닌 반대의) 대립가설을 옹호하는 현재의 실제 결과가 나올 가능성이 거의 없다면, **틀린 것은 가장 처음의 가정 즉, 귀무가설이다.** 따라서 귀무가설을 기각하고 대립가설을 채택한다.**

하지만 '아주 작다'는 판단을 위해서는 기준값이 필요하다. 계산된 유의확률값에 대해 귀무가설을 기각하는지 채택하는지를 결정할 수 있는 기준값을 **유의수준(level of significance)**라고 한다. 일반적으로 사용되는 유의수준은 1%, 5%, 10% 등이다. 따라서 위 문장은 다음과 같이 바꿀 수 있다.

유의확률이 유의수준보다 작으면 귀무가설을 기각하고 대립가설을 채택한다.

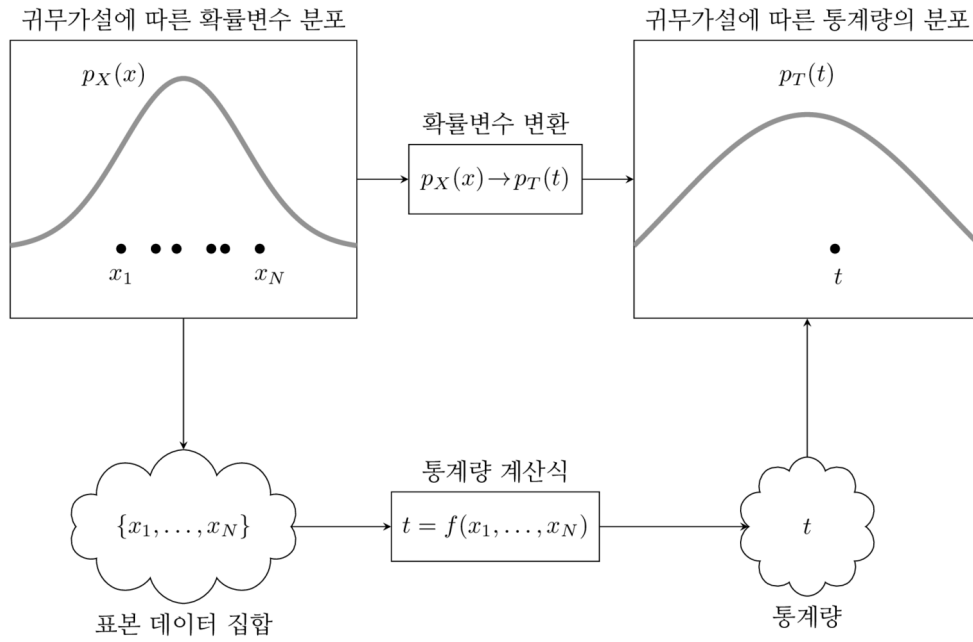
반대로 유의확률이 유의수준보다 크면 귀무가설을 기각하지 못하고 채택한다.

검정통계량이 나오면 확률밀도함수를 사용하여 유의확률을 계산할 수 있는 것처럼 반대로 특정한 유의확률값에 대해 해당하는 검정통계량을 계산할 수도 있다. 유의수준에 대해 계산된 검정통계량을 **기각역(critical value)**이라고 한다. 기각역을 알고 있다면 유의확률을 유의수준과 비교하는 것이 아니라 검정통계량을 직접 기각역과 비교하여 기각 여부를 판단할 수도 있다.

검정 방법론

검정의 기본적인 논리를 다시 정리하면 다음과 같다.

1. 데이터가 어떤 고정된 확률분포를 가지는 확률변수라고 가정한다. 예를 들어 동전은 베르누이분포를 따르는 확률변수의 표본이며 주식의 수익률은 정규분포를 따르는 확률변수의 표본이라고 가정한다.
2. 이 확률분포의 모수값이 특정한 값을 가진다고 가정한다. 이때 모수가 가지는 특정한 값은 우리가 검증하고자 하는 사실과 관련이 있어야 한다. 이러한 가정을 **귀무가설**이라고 한다. 예를 들어 동전이 공정한 동전이라고 주장하는 것은 베르누이 확률분포의 모수 θ 의 값이 0.5라고 가정하는 것과 같다. 주식이 손실을 보지 않는다는 것은 정규분포의 기댓값 모수 μ 가 0과 같거나 크다고 가정하는 것이다.
3. 만약 데이터가 주어진 귀무가설에 따른 표본이라면 이 표본 데이터를 특정한 수식에 따라 계산한 숫자는 귀무가설에서 유도한 특정 확률분포를 따르게 된다. 이 숫자를 **검정통계량**이라고 하며 검정통계량의 확률분포를 **검정통계분포**라고 한다. 검정통계분포의 종류 및 모수의 값은 처음에 정한 가설 및 수식에 의해 결정된다.
4. 주어진 귀무가설이 맞으면서도 표본 데이터에 의해서 실제로 계산된 검정통계량의 값과 같은 혹은 그보다 더 극단적인(extreme) 또는 더 희귀한(rare) 값이 나올 수 있는 확률을 계산한다. 이를 **유의확률**이라고 한다.
5. 만약 유의확률이 미리 정한 특정한 기준값보다 작은 경우를 생각하자. 이 기준값을 **유의수준**이라고 하는 데 보통 1% 혹은 5% 정도의 작은 값을 지정한다. 유의확률이 유의수준으로 정한 값보다도 작다는 말은 해당 검정통계분포에서 이 검정 통계치(혹은 더 극단적인 경우)가 나올 수 있는 확률이 아주 작다는 의미이므로 가장 근본이 되는 가설 즉, 귀무가설이 틀렸다는 의미다. 따라서 이 경우에는 귀무가설을 **기각**한다.
6. 만약 유의확률이 유의수준보다 크다면 해당 검정통계분포에서 이 검정 통계치가 나오는 것이 불가능하지만은 않다는 의미이므로 귀무가설을 기각할 수 없다. 따라서 이 경우에는 귀무가설을 **채택**한다.



예제

이제 서두에서 제기한 문제를 다시 풀어보자. 동전의 앞면이 나오는 것을 베르누이분포로 모형화한다. 판단하고 자하는 귀무가설은 베르누이분포 모수 μ 가 0.5라는 것이다.

$$H_0 : \mu = 0.5 \tag{9.4.26}$$

이 문제에 대한 검정통계량은 앞면이 나온 횟수가 된다.

$$t = \sum_{i=1}^N x_i = 12 \tag{9.4.27}$$

그리고 이 값은 전체 던진 횟수 $N = 15$ 인 이항 분포를 따른다.

In [3]:

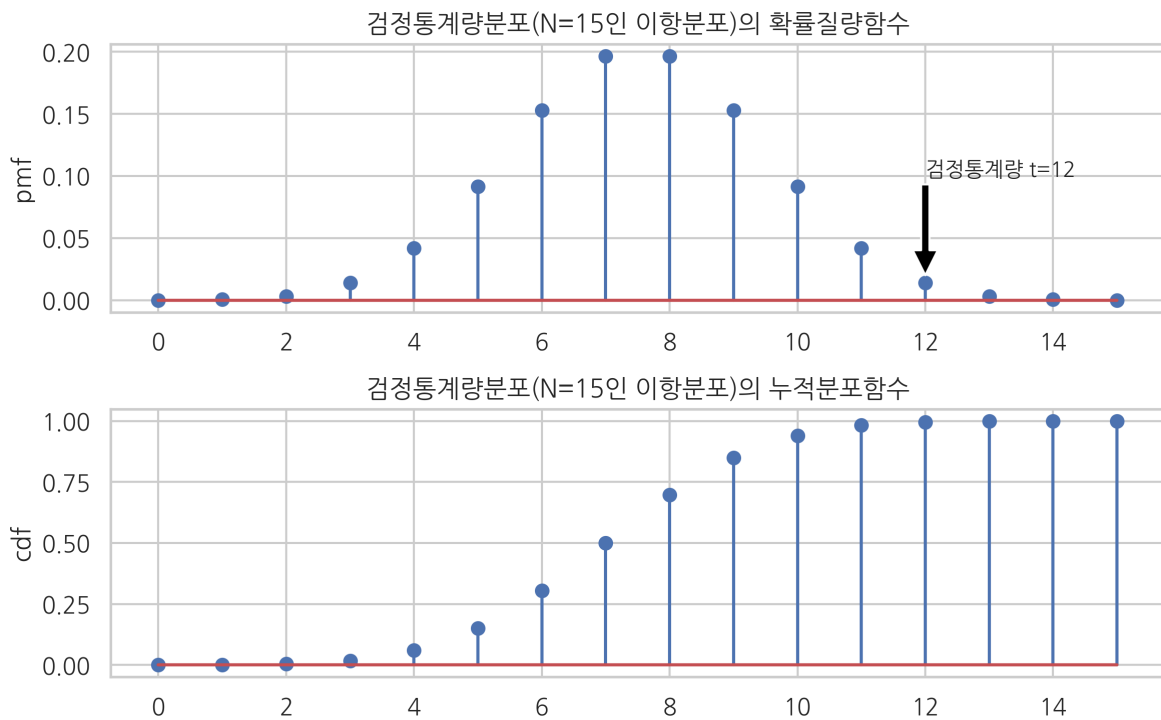
```
N = 15
mu = 0.5
rv = sp.stats.binom(N, mu)

xx = np.arange(N + 1)

plt.subplot(211)
plt.stem(xx, rv.pmf(xx))
plt.ylabel("pmf")
plt.title("검정통계량분포(N=15인 이항분포)의 확률질량함수")
black = {"facecolor": "black"}
plt.annotate('검정통계량 t=12', xy=(12, 0.02), xytext=(12, 0.1), arrowprops=black)

plt.subplot(212)
plt.stem(xx, rv.cdf(xx))
plt.ylabel("cdf")
plt.title("검정통계량분포(N=15인 이항분포)의 누적분포함수")

plt.tight_layout()
plt.show()
```



만약 단순히 동전이 공정하지 않다는 것을 보이고 싶다면 대립가설은 다음과 같다.

$$H_a : \mu \neq 0.5 \quad (9.4.28)$$

이때는 양측검정 유의확률을 계산해야 한다. 검정통계량 $t = 12$ 에 대한 양측검정 유의확률은 약 3.5%이다.

$$\begin{aligned} \text{유의확률} &= 2 \cdot \text{Bin}(n \geq 12; N = 15, \mu = 0.5) \\ &= 2 \cdot (1 - F(11; N = 15, \mu = 0.5)) \\ &= 0.03515625 \end{aligned} \quad (9.4.29)$$

In [4]:

```
2 * (1 - rv.cdf(12 - 1))
```

Out[4]:

0.03515625

이 값은 5%보다는 작고 1%보다는 크기 때문에 유의수준이 5%라면 귀무가설을 기각할 수 있으며 공정한 동전이 아니라고 말할 수 있다. 만약 유의수준이 1%라면 귀무가설을 기각할 수 없다. 즉, 공정한 동전이 아니라고 말할 수 없다.

만약 동전이 앞면이 더 많이 나온다는 것을 보이고 싶다면 대립가설은 다음과 같다.

$$H_a : \mu > 0.5 \quad (9.4.30)$$

μ 가 클수록 검정통계량도 커지므로 이때는 우측검정 유의확률을 계산해야 한다. 이 값은 약 1.8%다.

$$\begin{aligned} \text{유의확률} &= \text{Bin}(n \geq 12; N = 15, \mu = 0.5) \\ &= 1 - F(11; N = 15, \mu = 0.5) \\ &= 0.017578125 \end{aligned} \quad (9.4.31)$$

In [5]:

```
1 - rv.cdf(12 - 1)
```

Out[5]:

0.017578125

예제

수익률이 정규분포를 따른다고 가정하면 이 주식의 검정통계량은 다음과 같이 계산한다.

$$t = \frac{m}{\frac{s}{\sqrt{N}}} = -1.4025 \quad (9.4.32)$$

In [6]:

```
x = np.array([-0.025, -0.05, 0.043, -0.037, -0.056])
t = x.mean()/x.std(ddof=1)*np.sqrt(len(x))
t
```

Out[6]:

-1.4025921414082105

만약 이 주식이 장기적으로 손실을 낸다는 것을 보이고 싶다면 대립가설은 다음과 같다.

$$H_a : \mu < 0 \quad (9.4.33)$$

μ 가 클수록 검정통계량도 커지므로 이 때는 좌측검정 유의확률을 구해야다. 이 값은 약 11.67%다.

$$\begin{aligned} \text{유의확률} &= t(t \leq -1.4025; \nu = 4, \mu = 0) \\ &= F(-1.4025; \nu = 4, \mu = 0) \\ &= 0.1167 \end{aligned} \quad (9.4.34)$$

In [7]:

```
sp.stats.t(df=4).cdf(t)
```

Out[7]:

0.11669216509589829

In [8]:

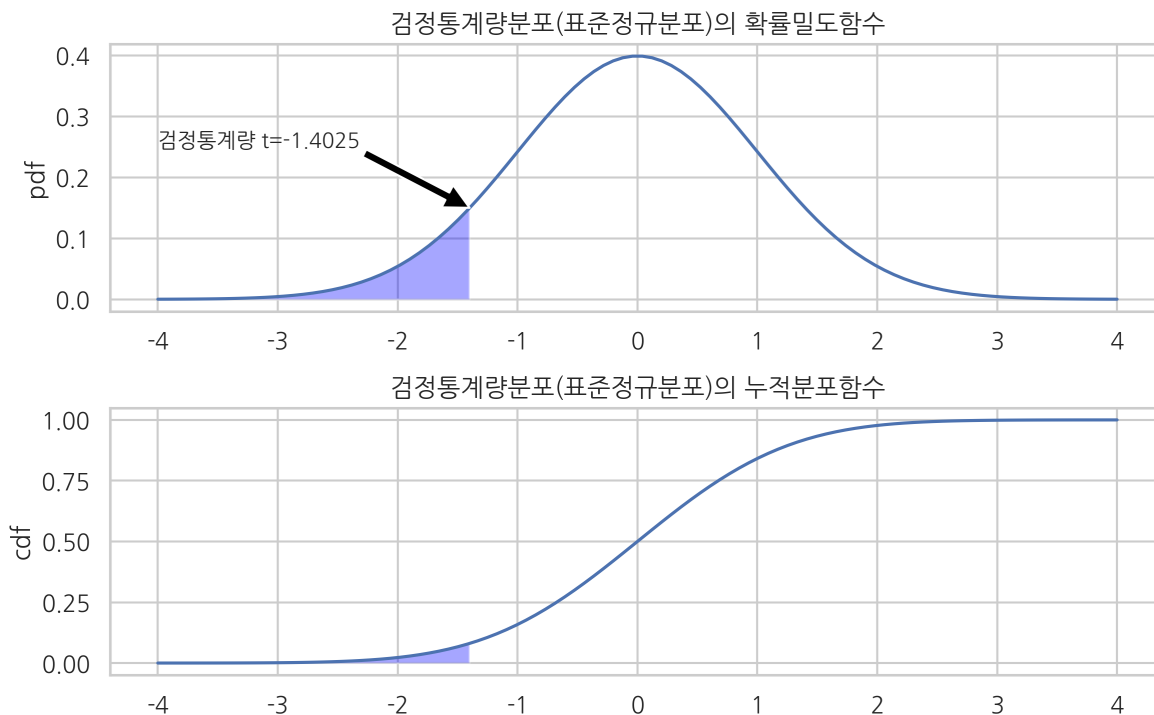
```
rv = sp.stats.norm()

xx = np.linspace(-4, 4, 100)

plt.subplot(211)
plt.plot(xx, rv.pdf(xx))
plt.ylabel("pdf")
plt.title("검정통계량분포(표준정규분포)의 확률밀도함수")
black = {"facecolor": "black"}
plt.annotate('검정통계량 t=-1.4025', xy=(-1.4, 0.15), xytext=(-4, 0.25), arrowprops=black)
xx2 = np.linspace(-4, -1.4025, 100)
plt.fill_between(xx2, rv.pdf(xx2), facecolor='blue', alpha=0.35)

plt.subplot(212)
plt.plot(xx, rv.cdf(xx))
plt.fill_between(xx2, rv.cdf(xx2), facecolor='blue', alpha=0.35)
plt.ylabel("cdf")
plt.title("검정통계량분포(표준정규분포)의 누적분포함수")

plt.tight_layout()
plt.show()
```



만약 유의수준이 10%라면 유의확률이 이보다 크기 때문에 귀무가설을 기각할 수 없다. 즉, 정규분포의 기댓값이 0보다 작다고 말할 수 없다. 이는 해당 주식이 장기적으로 손실을 보는 주식이라고 말할 수 있는 증거가 부족하다는 의미다.

연습 문제 9.4.3

어떤 인터넷 쇼핑몰의 상품 20개의 상품평이 있고 '좋아요'가 11개 또는 '싫어요'가 9개다. 이 상품이 좋다는 주장을 검정하라. 유의수준은 10%다.

