9.4 검정과 유의확률

검정 (testing)

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

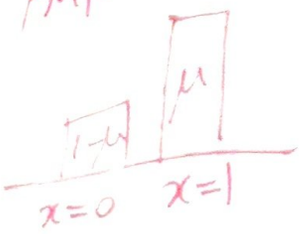
(<예제>
어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다.
이 동전은 조작되지 않은 공정한 동전이라고 할 수 있는가?)

베르누이 확률분포 모형 / 모수추정 → $\frac{12}{15} = 0.8 \neq$ 공정하지 않음

BUT 추정값일 뿐!

"어떻게 해야 동전의 조작여부를 증명할 수 있을까?"

PMF



$\mu = \frac{12}{15} = \frac{4}{5}$

"얼마나 가능한가?"

[예제]
어떤 주식의 월별 수익률은 다음과 같다.

$\boxed{-2.5\%,\ -5\%,\ 4.3\%,\ -3.7\%,\ -5.6\%}$

이 주식을 장기적으로 수익 OR 손실?

→ 정규분포 모델이라고 할 때, 기대수익률이 음수일까 아니면 단순한 우연일까?

가설과 검정

데이터를 특정 확률분포를 가진 확률변수로 모형화하면 모수를 추정할 수 있다.

다음 작업으로는 데이터 뒤에 숨어있는 확률변수가 정말로 그 모수값을 가졌는지 검증해보아야 한다.

= "해당 확률변수는 정말 그 모수값을 가졌다는 주장을 논리적으로 증명해야 함"

확률분포에 대한 어떤 가설 (hypothesis), H

이 가설을 증명하는 행위 = 통계적 가설검정 (statistical hypothesis testing)

≒ 검정 (testing)

> * 특히, 확률분포의 모수값이 특정한 값을 가진다는 가설을 검정하는 것을 모수검정 (parameter testing) 이라고 한다.

회귀분석 작업 어려운 = 검정!

① 귀무가설.

검정 작업을 하려면,

- 데이터가 어떤 확률변수의 표본이라고 가정한다.
- 데이터가 만드는 확률변수가 따르는 확률분포의 모수 θ의 값이 어떤 특정한 모수값 θ_0으로 고정되어 있다고 가정한다.

귀무가설 (null hypothesis) — 확률분포 모수에 대한 가정 = H_0 = 가정!

$$\boxed{H_0 : \theta = \theta_0}$$

등식관계 상수 (상수)

(예제 1)

동전이 공정하다는 귀무가설

$$H_0 : \mu = 0.5$$

(예제 2)

주식 수익률이 장기적으로 + 인가 − 인가에 대한 귀무가설

$$H_0 : \mu = 0 \quad (\text{기준점})$$

ⓜ 대립가설

귀무가설(H_0) = 아닌 가능 상태

대립가설 (alternative hypothesis) = H_a = $\boxed{\text{주장 또는 반박}}$

≒ 연구가설 (research hypothesis)

ⓔ 기존 제품의 성능 / 목표성능 = H_0

진실임을 증명하고자 하는 가설 = H_a

(예제 1)

모수 θ가 어떤 특정한 값 θ_0이 아니라는 것을 증명하고 싶다면

$$H_0 : \theta = \theta_0, \quad \left[\begin{array}{l} H_a : \theta \neq \theta_0 \\ H_a : \theta > \theta_0 \\ H_a : \theta < \theta_0 \end{array}\right]$$

* H_0과 H_a는 반드시 여집합 관계 (complement)이 있을 필요는 없다!

(예제)

동전이 공정하지 않다고 주장.

$$H_0 : \mu = 0.5, \quad H_a : \mu \neq 0.5$$

증명하기 위해서,

H_a가 참이면서 (&) H_0가 틀렸다는 증거가 필요함.

③ 검정통계량

귀무가설이 맞거나 틀렸거나 하는 것을 증명하려면 어떤 증거가 있어야 한다.

예를 들어보자.

- '어떤 병에 걸렸다' 가설 증명
  - ↳ 혈액 내의 특정 성분의 수치) = 검정통계량.
- '어느 학생이 우등 상장을 받을 수 있는 우등생' 가설 증명
  - ↳ 시험 성적 = 검정통계량

검정통계량 (test statistics)

= 표본데이터 집합을 입력으로 계산되는 함수의 값

$$t = f(x_1, x_2, \ldots, x_N)$$

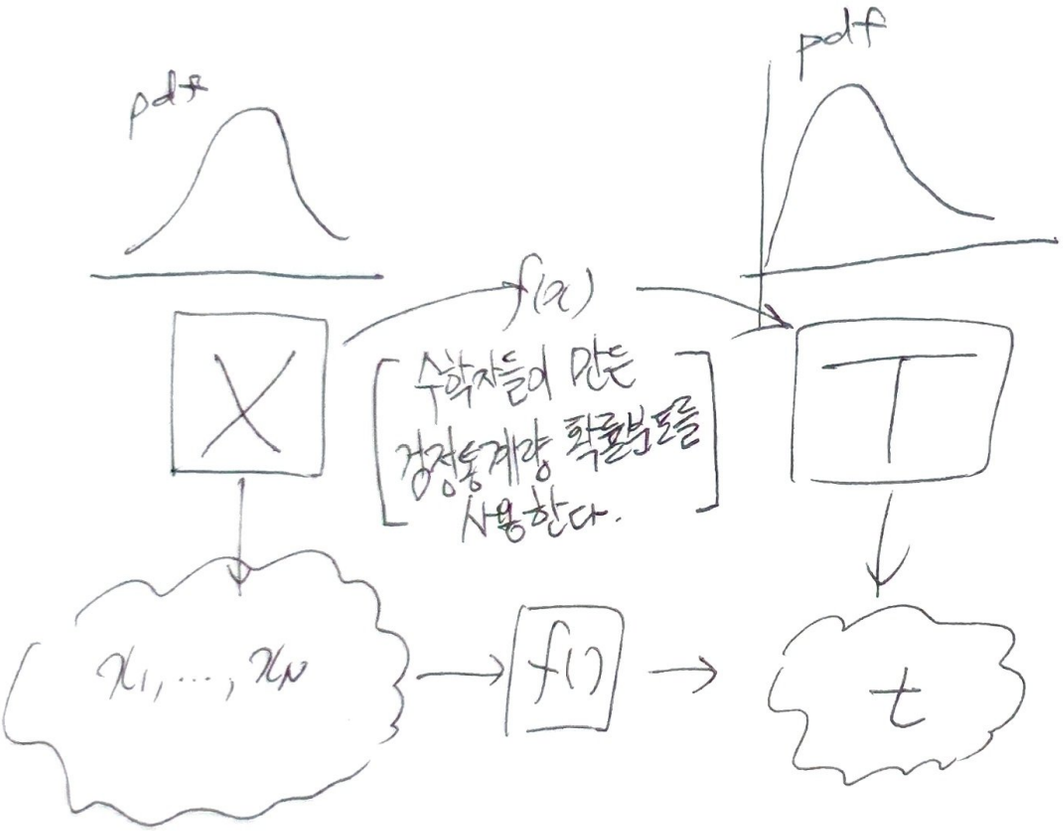
↳ 확률변수 X의 표본으로 계산하는 함수값

↳ 확정된 값이 아니다

↳ 검정통계량 확률변수 T의 표본으로 볼 수 있다.

입력데이터인 확률변수 X의 확률분포함수 $p_X(x)$와
검정통계량 수식 $f(x)$가 이미 결정되기 때문에,
검정통계량 확률변수 T의 확률분포함수 $p_T(t)$도 수식 유도가 가능하다.

$$\left\{ \begin{array}{l} p_X(x) \\ f(x_1, x_2, \ldots, x_N) \end{array} \right\} \longrightarrow p_T(t)$$



(예제) 베르누이분포 확률변수.

모수 μ를 가지는 베르누이분포 확률변수에 대해서는 전체 시도 횟수 N번 중 성공한 횟수 n자체를 검정통계량으로 뽑을 수 있다.

· 이 검정통계량은 자유도 N과 모수 μ를 가지는 이항분포를 따른다.

$$\boxed{x \sim \mathrm{Ber}(x;\mu) \longrightarrow t = \sum_{i=1}^{N} x \sim \mathrm{Bin}(t; N, \mu)}$$

(예제) 분산 σ^2 값을 알고 있는 정규분포 확률변수

분산 σ^2의 값을 알고 있는 정규분포 확률변수에 대해서는, 다음과 같이 표본평균 m을 분산 σ^2로 정규화 (normalize) 한 값을 검정통계량으로 쓴다.

* 이 검정통계량은 표준정규분포를 따른다.
* 이 검정통계량은 특별히 z 라고 부른다.

normalize

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

$$\text{for } \bar{x} = \text{표본평균} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(예제) 분산 σ^2 값을 모르는 정규분포 확률변수

이번에는 분산 σ^2의 값을 모르는 정규분포 확률변수를 고려하자.

모수 μ에 대한 검정을 할 때는 다음과 같이 표본평균 m을 표본분산 s로 정규화한 값을 검정통계량으로 쓴다. 이 검정통계량은 자유도가 $N-1$인 스튜던트 t 분포를 따른다. (N = 데이터의 수)

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

$$\text{for } m = \text{표본평균} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad s^2 = \text{표본분산} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

* 분산 σ^2에 대한 검정을 할 때는,

다음과 같이 표본분산을 정규화 (normalization) 한 값을 검정통계량으로 쓴다.

- 이 검정통계량은 자유도 $N-1$인 카이제곱분포를 따른다.
  (N = 데이터의 수)

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

유의확률

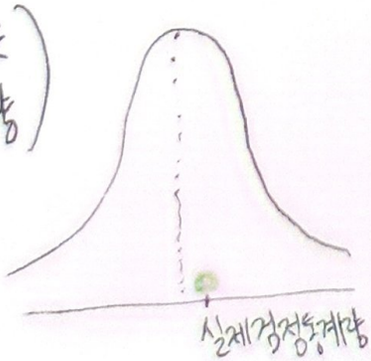
〈우리가 알게된 두가지 정보〉

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

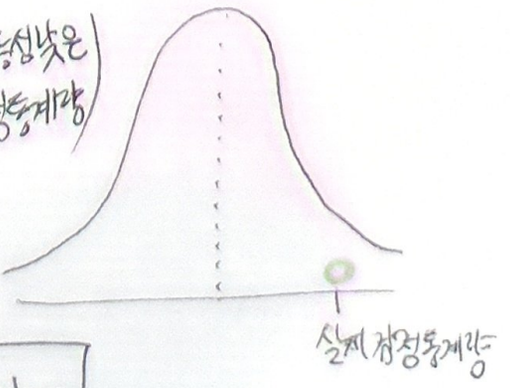
① H_0 == True,
실제 데이터에서 구한 검정통계량 값은 검정통계량 확률분포를 따르므로, 기댓값이나 모드값 근처의 값이 나올것

② H_0 != True,
~~우리가 가정한~~ 실제 데이터에서 구한 검정통계량 값은 검정통계량 분포에서 나오기 힘든 값일 것

(* 가능성 높은 검정통계량)

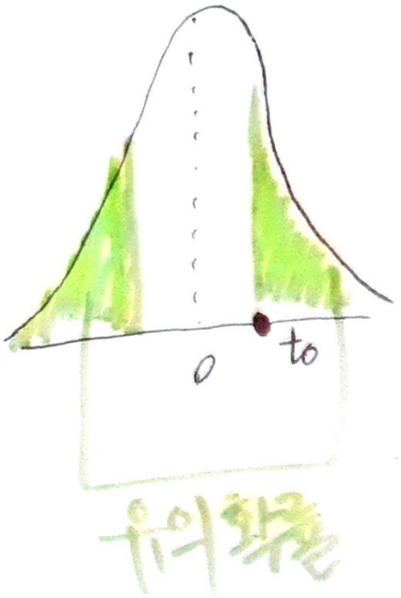


(* 가능성 낮은 검정통계량)

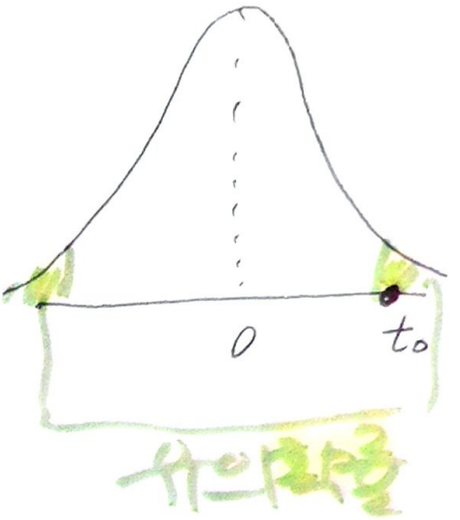


분포에서 나오기 쉽다, 어렵다 → $\boxed{\text{p-value}}$

* 유의확률이 큰 경우



* 유의확률이 작은 경우



유의확률 = 중심으로부터 떨어진 거리 개념

= 귀무가설이 맞음에도 불구하고, 현재 검정통계량과 같은 혹은 대립가설을 더 옹호하는 검정통계량이 나올 확률.

$= P(t \text{ for } H_a \mid H_0)$

귀무가설이 진실인 사건

단측검정 유의확률

증명하고자 하는 대립가설이 부등식인 경우,
2 대립가설을 옹호하는 검정통계량이 나올 확률을 구할 때
특정한 한 방향의 확률만을 구해야 한다.
⇒ 단측검정 (one-side test, single-tailed test)

모수 θ가 특정한 값보다 크다는 것을 증명하고 싶다면,
귀무가설과 대립가설은 다음과 같다.

$$H_0 : \theta = \theta_0, \quad H_a : \theta > \theta_0$$

반대로 모수 θ가 특정한 값보다 작다는 것을 증명하고 싶다면,

$$H : \theta = \theta_0, \quad H_a : \theta < \theta_0$$

<< 단측검정 유의확률 구하는 법 >>

(모수가 클 때, 검정통계량도 정비례해서 같이 커진다고 가정)

① 모수 θ가 ~~특정한 값보다~~ 특정한 값보다 크다는 것을 증명하는 경우

우측검정 (right-side) 유의확률 사용.

→ 귀무가설이 맞음에도 불구하고, 검정통계량이 현재 검정통계량과 같거나 더 큰 값이 나오는 확률.

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

만약 이산확률분포라면, 등호가 성립하는 부분을 제외해야 하므로 다음처럼 구한다.

$$\text{유의확률} = 1 - F(t_0 - 1)$$

* 좌측검정의 경우는 이산확률분포도 위 식과 같이 구한다.

[유의확률]

확률분포와 확률분포의 표본값 1개가 주어졌을 때,

2 확률분포에서 해당 표본값 혹은 더 희귀한 (rare) 값이 나올 수 있는 확률

- 확률밀도함수에서 표본값을 기준으로 만들어진 양측 꼬리 (tail) 부분에 해당하는 면적

누적확률분포함수 $F(x)$를 이용하여 다음과 같이 계산할 수 있다.

$$\text{유의확률} = 2\min(P(t \le t_0), P(t \ge t_0))$$

$$= 2\min(F(t_0), 1 - F(t_0))$$

$\boxed{t_0 = \text{현재 검정통계량의 값}}$ 이며, 위 유의확률은 통계량분포의 양 끝단의 면적을 포함하기 때문에 양측검정 유의확률 (two-sided test p-value) 이라고 한다.

< * 만약 이산확률분포라면 등호를 제거하여 계산 >

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

→ "유의확률이 작다면, 귀무가설은 틀렸다."

▨ 유의수준과 기각역

유의확률값이 아주 작다 ⇒ 귀무가설이 맞다는 가정하에
현재 검정통계량이 나올 가능성이 매우 적다.

⇒ 귀무가설이 틀렸다

⇒ 귀무가설을 기각하고 대립가설을 채택한다.

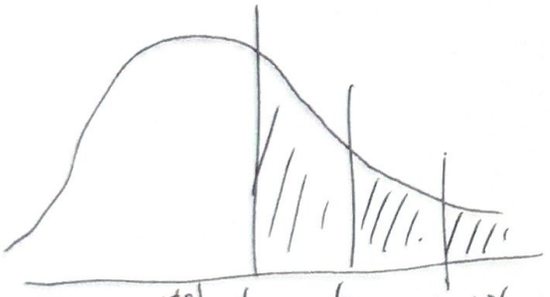
'아주 작다'의 기준 = 유의수준 (level of significance)

* 유의확률이 유의수준보다 낮으면 귀무가설을 기각하고 대립가설을 채택한다.

(일반적인 유의수준) 1%, 5%, 10%

* 유의확률값에 대해 해당하는 검정통계량 = 기각역 (critical value)
기각역을 알고 있다면, 검정통계량을 직접 비교하여 귀무가설 기각 여부를 판단 가능.

* 수식이 너무 복잡해지면, 검정통계량을 계산한 이후에도 p-value를 못구할 수 있다.
cdf가 너무 복잡해서 구할수 없기때문! 대신에 통계학자들이 미리 계산한 값으로 범위유추.



면적 $t_{10\%}$ $t_{5\%}$ $t_{1\%}$ = 10.
(위치의 검정통계량.)

$t=3$이면, p-value를 몰라도, $t_{10\%} \sim t_{5\%}$
사이에 위치한다는 것은 알수 있다.

검정 방법론 정리.

1. 데이터가 어떤 고정된 확률분포를 따르는 확률변수라고 가정한다.

2. 이 확률분포의 모수값이 특정한 값을 가진다고 가정한다.
   - '특정한 값'은 우리가 검증하고자 하는 사실과 관련이 있어야 한다.

3. 데이터가 주어진 귀무가설에 따른 표본이라면, 이 표본 데이터를 특정한 수식에 따라 계산한 숫자는 귀무가설에서 유도한 특정 확률분포를 따르게 된다. 이 숫자를 검정통계량이라 하며, 검정통계량의 확률분포를 검정통계분포라고 한다.

4. 주어진 귀무가설이 맞음에도 불구하고, 표본 데이터에 의해서 실제로 계산된 검정통계량의 값과 같은, 혹은 그보다 더 극단적인 (extreme) 또는 더 희귀한 (rare) 값이 나올 수 있는 확률을 계산한다. 이를 유의확률이라고 한다.

5. 유의확률이 높다 / 낮다를 판단하는 기준값, 즉 유의수준을 정한다.
   - 1%, 5% 등
   - 지정 유의수준보다 유의확률이 낮다
     ⇒ 더 극단적인 검정통계치가 나올 확률이 희박하다
     ⇒ 귀무가설이 틀렸으로 기각, 대립가설을 채택한다

(예제)

어떤 주식의 일주일 수익률은 다음과 같다.

-2.5%, -5%, +1.3%, -3.7%, -5.6%

장기적으로 수익 OR 손실?

풀이

수익률이 정규분포를 따른다고 가정하면,
이 주식의 검정통계량은 다음과 같이 계산한다.

$$t = \frac{m}{\frac{s}{\sqrt{N}}} = -1.4025 \sim \text{스튜던트 } t \text{ 분포}$$

장기적으로 손실이 난다는 것을 보이려 한다면, 귀무가설은

$$H_0 : \mu < 0$$

μ가 클수록 검정통계량도 커지므로, 이 때는 좌측검정 유의확률을 구한다.
이 값은 약 11.67%이다.

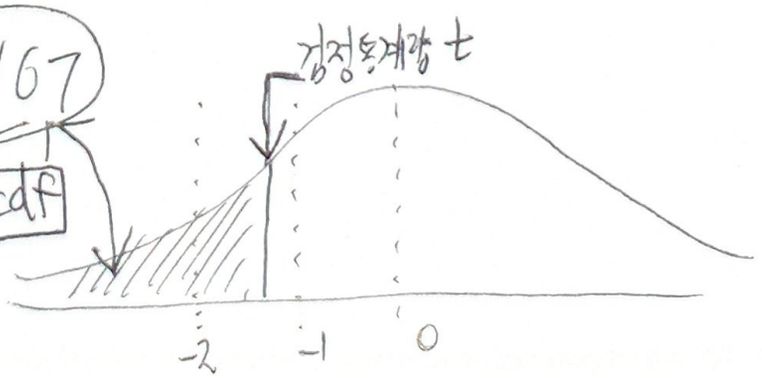
$$\text{유의확률} = t(t \le -1.4025;\ \nu = 4,\ \mu = 0)$$

($\mu < 0$이 참일때, 더 극단적인 값이 나올 확률)

$$= F(-1.4025;\ \nu = 4,\ \mu = 0)$$

$$= 0.1167$$

cdf



$\boxed{t\text{ 통계량이 충분히 크다}}$

- 귀무가설을 기각할 수 없다
- 주식 수익률 정규분포의 기댓값이 0보다 작다고 ($\mu < 0$) 말할 수 없다. = "판단할 수 없다"

Ⅲ 카이제곱검정

베르누이분포 모수에 대해 검정

⇝ 베르누이분포 표본의 합이라는 통계량이 이항분포를 따른다는 성질을 이용하면 된다고 배웠다.

* 그러면, 카테고리분포의 모수에 대해 검정을 하려면 카테고리분포 표본의 합이라는 통계량이 다항분포를 따른다는 성질을 이용할 수 있을까?

⇒ 통계량으로 카테고리 분포의 합은 이용할 수 없다.

∵ 통계량은 스칼라가 아닌 벡터값이 나오기 때문.

따라서, 카테고리분포의 모수 검정 문제에서는 카이제곱검정(chi-squared test)을 사용한다.

$\boxed{\text{카이제곱검정}}$ = 범주형 확률분포의 모수 ($\mu = (\mu_1, \dots, \mu_K)$)에 대한 가설을 조사하는 검정방법

= 적합도검정 (goodness of fit test)

"원래 범주형 값 k가 나와야 할 횟수의 기댓값 m_k 와 실제 나온 횟수 x_k의 차이를 이용하여 다음처럼 검정통계량을 구한다."

$$\sum_{k=1}^{K} \frac{(x_k - m_k)^2}{m_k}$$

- scipy.stats.chisquare(f_obs, f_exp=None)
  - f_obs : 데이터 행렬
  - f_exp : 기댓값 행렬 (default $\mu = (\frac{1}{K}, \frac{1}{K}, \cdots \frac{1}{K})$)

카이제곱 독립검정

카이제곱검정은 어떤 범주형 확률변수 X가 다른 범주형 확률변수 Y와 독립인지 상관관계를 가지는가를 검증하는 데 사용할 수 있다.

- 카이제곱 검정을 독립 확인 목적으로 사용하면 "카이제곱 독립검정" 이라고 부른다.

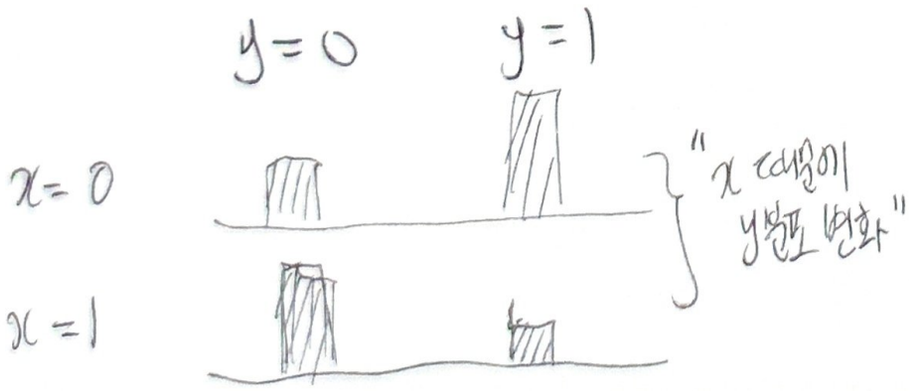
확률변수 X와 Y가 독립이라면, 결합 확률질량함수 $P(x,y)$는 각 확률변수 X와 Y의 주변 확률질량함수 $P(x)$, $P(y)$의 곱이다. 예를 들어 다음과 같은 확률분포를 가진다면 X, Y는 독립이다.

	$P(Y=0)=0.3$	$P(Y=1)=0.7$
$P(X=0)=0.4$	0.12	0.28
$P(X=1)=0.6$	0.18	0.42

분할표 (contingency table)

→ 확률변수의 도수를 측정하여, 2 횟수를 표로 나타낸 것

✓ 범주형 확률변수 간의 상관관계?



예를 들어, 50개의 표본을 측정한 분할표가 다음과 같다면
확률변수 X와 Y가 독립이라고 주장할 수 있을 것이다.

	Y=0	Y=1
X=0	6	14
X=1	9	21

×1.5 ×1.5 ×$\frac{2}{3}$ ×$\frac{2}{3}$

X가 Y의 분포에 영향 X

Y가 X의 분포에 영향 X

그런데 만약 분할표가 다음과 같다면, 독립일까 독립이 아닐까?

① 원래 독립인데 표본오차에 의해 약간 차이가 발생했다
② 원래 독립이 아니다

	Y=0	Y=1
X=0	5	15
X=1	10	20

⇒ 카이제곱 독립검정을 사용하여 판단

두 확률변수가 독립이라면,
X=0일 때의 Y분포와 X=1일 때의 Y분포가 같아야 한다.
따라서, 두 경우의 표본집합이 같은 확률분포로부터 나왔다는 것을 귀무가설로 하는 카이제곱검정을 하여 채택된다면, 두 확률변수는 독립이다.
└ 만약 기각된다면 두 확률변수는 상관관계가 있다.

* scipy의 chi2_contingency() 명령

→ 카이제곱 독립검정 수행

→ X에 따른 각각의 Y 분포가 2차원 표(contingency table) 형태로 주어지면, 독립인 경우와 분포와 실제 y 표본분포의 차이를 검정통계량으로 계산한다.

— 이 값이 충분히 크다면, X-Y는 상관관계가 있다

— chi2_contingency() 결과는 tuple로 반환

① 검정통계량

② 유의확률

(통계량이 작다면 상관관계 X = 독립이다.)

```
obs = np.array([[5,15], [10,20]])
sp.stats.chi2_contingency(obs)
```

(0.09920634…, 0.752784…, …)

검정통계량

유의확률 (귀무가설 = 같은 확률분포)

↓

[5,15]와 [10,20]은 상관관계가 있다고 말할 수 없다.

∴ 귀무가설일 가능성이 충분히 크다.

단일표본 t 검정 (One-sample t-test)

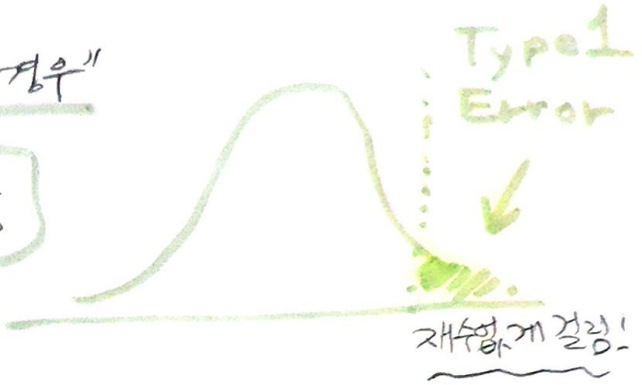
분산 σ^2의 값을 정확히 알고 있는 정규분포의 표본에 대해 기댓값을 조사하는 방법.

- 거의 사용되지 않음. (데이터 분산을 정확히 아는 경우 X)

* 1종 오류 (Type 1 Error)

데이터의 숫자가 너무 작거나 하는 이유로,
"귀무가설이 진실임에도 불구하고 기각되는 경우"

$\boxed{\text{1종 오류의 발생 확률} = \text{유의수준}}$



→ 검정통계량으로
스튜던트 t 분포를 가진 t 통계량을 사용한다.

$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{N}}}$$

scipy.stats.ttest_1samp()

독립표본 t 검정 (independent two-sample t-test)

두 개의 독립적인 정규분포에서 나온 N_1, N_2 개의 데이터 셋을 사용하여, 두 정규분포의 기댓값이 동일한지를 살핀다.

검정통계량 (두 정규분포의 분산이 같음)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

표본표준편차 s는 각각의 표본분산 S_1^2, S_2^2로부터 다음과 같이 구한다.

$$s = \sqrt{\frac{(N_1-1)S_1^2 + (N_2-1)S_2^2}{N_1+N_2-2}}$$

이 통계량은 자유도가 $\underline{N_1+N_2-2}$인 스튜던트 분포를 따른다.

두 정규분포의 분산이 다른 경우에는, 검정통계량으로 아래 식을 사용한다.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

위 값은 자유도(degree of freedom)가 아래와 같은 스튜던트 t 분포를 따른다.

$$d.o.f = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{(S_1/N_1)^2}{N_1-1} + \frac{(S_2^2/N_2)^2}{N_2-1}}$$

독립표본 t 검정은 scipy.stats 서브패키지의 ttest_ind 명령을 사용하여 계산한다.

↳ 독립표본 t 검정은 두 정규분포의 ① 분산값이 같은 경우와 ② 같지 않은 경우 사용하는 검정통계량이 다르기 때문에, equal_var 인수를 사용하여 이를 지정해 주어야 한다.

- 두 분포의 분산이 같은지 다른지는 다음에 배울 등분산검정(equal-variance test) 실시
- 잘 모르겠으면 equal_var = False

* 2종 오류 (Type 2 Error)

귀무가설이 실제로 거짓임에도 불구하고, 귀무가설이 진실이라는 가정을 기각할 수 없음

- 데이터 수가 증가하면 이러한 오류는 감소한다.

대응표본 t 검정 (Paired two-sample t-test)

대응표본 t검정 (Paired two-sample t-test)

- 독립표본 t검정을 두 집단의 표본이 1:1 대응하는 경우에 대하여 수정한 것.
  → 독립표본 t검정과 마찬가지로, 두 정규분포의 기댓값이 같은지 확인하기 위한 검정.

통계량 계산 (대응하는 표본값의 차이: $x_d = x_{i,1} - x_{i,2}$)

difference

$$t = \frac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{N}}}$$

ex 한 학생이 데이터사이언스 수업을 듣기 전과 들은 후의 시험성적 — pair

scipy.stats.ttest_rel(a, b)

└ p-value 0.004 일때,
유의수준 1% 에서 귀무가설을 기각할 수 없다
= 서로 다른 분포 ($a \neq b$)

등분산검정

앞서 살펴본 ttest_inds (독립표본 t검정)을 사용하려면,
두 데이터 집합의 분산이 같은지 먼저 알아내야 한다.

등분산검정(Equal-variance test)은 두 정규분포로부터 생성된
두 개의 데이터 집합으로부터, 두 정규분포의 분산 모수가 같은지
확인하는 검정이다.

- 바틀렛 (bartlett) — sp.stats.barlett
- 플리그너 (fligner) — sp.stats.fligner
- 레빈 (levene) — sp.stats.levene

정규성 검정

회귀분석 등에는 확률분포가 가우시안 정규분포를 따르는지 아닌지를 확인하는 것이 중요.

→ 정규성검정 (normality test) ... 귀무가설 = 정규분포 O

→ 다양한 검정 방법들이 개발되어 있음.

scipy	statsModels (sm)
콜모고로프-스미르노프 scipy.stats.ks_2samp	콜모고로프-스미르노프 sm.stats.diagnostic.kstest_normal
샤피로-윌크 scipy.stats.shapiro	옴니버스 sm.stats.stattools.omni_normtest
앤더슨-달링 scipy.stats.anderson	자크-베라 sm.stats.stattools.jarque_bera
다고스티노 K제곱 scipy.stats.mstats.normaltest	릴리포스 sm.stats.~~stattools~~.diagnostic.lilliefors