제10장 가설검정

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-09-13

통계적 가설검정 (hypothesis test)

- 모수 또는 모집단 분포에 대한 가정 (가설)을 세우고, 표본을 기초로 가정의 참 거짓을 판단하는 방법
- 가설의 종류
  1. 대립가설(alternative hypothesis): 통상적으로 연구자가 입증하려는 가설로, 표본을 토대로 확실한 근거가 있어야 받아들임
  2. 귀무가설(null hypothesis): 대립가설과 상반되는 가설
  3. 보통 귀무가설을 H_0, 대립가설을 H_1으로 표시함
- 검정통계량 (test statistic): 가설검정을 하기 위해 이용하는 통계량
- 유의수준 (significance level): 귀무가설이 참인데도 불구하고 귀무가설을 기각할 확률 (α로 표시함)
- 기각역 (critical region, rejection region): 귀무가설을 기각시키기 위한 검정통계량이 포함되는 관측값의 영역
- P값 (유의확률; P-value): 귀무가설이 참이라는 전제하에서 검정통계량이 관측값을 벗어날 확률

$$P(T > t | H_0 \ is \ true)$$

귀무가설과 대립가설의 예

- 모평균(μ)에 대한 가설 검정
  1. 귀무가설: 모평균(μ)에 대한 통상적인 주장
  2. 대립가설: 실험자, 연구자가 입증하기를 원하는 내용

$$H_0 : \mu = 170, \text{ v.s. } H_1 : \mu > 170$$

가설검정에서 오류

	H_0 is TRUE	H_0 is FALSE
Reject H_0	Type I Error	True Positive
Accept H_0	True Negative	Type II error

- 제1종오류 (Type I Error): 귀무가설이 참인데도 불구하고 귀무가설을 기각할 오류
- 유의수준 (significance level): Type I Error를 범할 확률의 최대값

$$\alpha = P(Reject\ H_0 | H_0\ is\ TRUE)$$

- 제2종 오류 (Type II Error): 귀무가설이 거짓인데도 불구하고 귀무가설을 채택하는 오류
- 검정력 (Power) : 귀무가설이 거짓일 때, 귀무가설을 기각할 확률

$$1 - \beta = P(Reject\ H_0 | H_0\ is\ FALSE)$$

t검정

1. 일표본 t검정
2. 이표본 t검정
3. 대응비교

일표본 t검정

- 표본자료: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- 모평균이 μ_0 (특정값)인지를 관측된 표본을 이용하여 검증
- 가설: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- 검정통계량 :
  1. 모집단 분산(σ^2)이 알려진 경우

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ :표준정규분포}$$

  2. 모집단 분산을 모르는 경우: 자유도(degree of freedom)가 $n-1$인 t분포

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1),$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- 기각역: 유의수준 α인 양측검정에서

$$|T| > t_{\alpha/2, n-1}$$

- p-value(유의 확률): 검정통계량의 관측값이 t_0일 때,

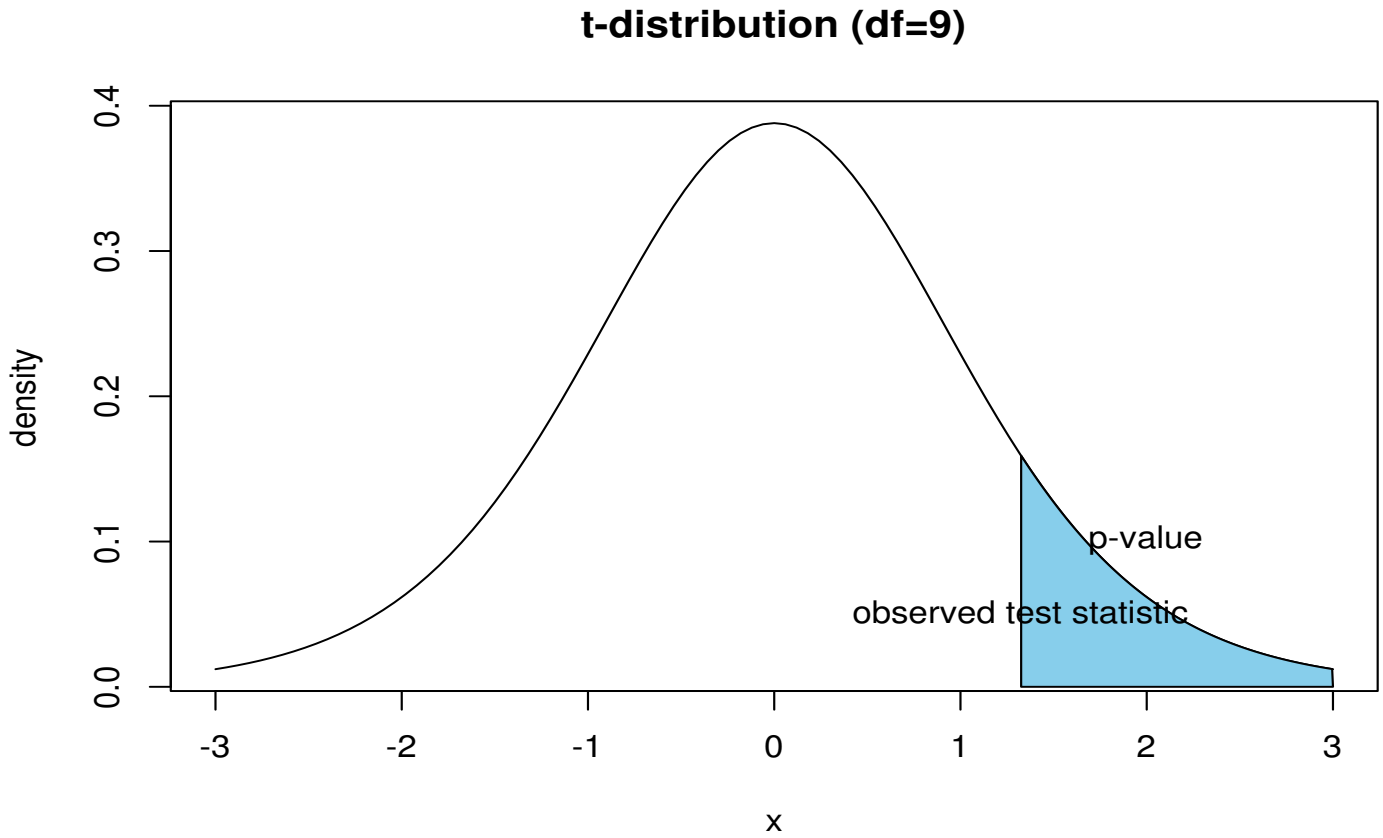
예제: 일표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 10명의 환자에게 수면제를 투여 후, 초과 수면시간 측정
- 가설 : 초과 수면시간(μ)이 0보다 큰가? ($H_0 : \mu = 0$ vs, $H_1 : \mu > 0$)

ID	extra
1	0.7
2	-1.6
3	-0.2
4	-1.2
5	-0.1
6	3.4
7	3.7
8	0.8
9	0.0
10	2.0

표본평균	표준오차	t값	자유도	p값
0.75	0.5657345	1.32571	9	0.1087989

예제: 일표본 t검정 (conti.)



이표본(독립표본) t검정 (Two Sample t-test) : 두 모집단의 평균의 차이 검정

- 표본자료

$$\begin{aligned} X_1, \ldots, X_m &\sim N(\mu_1, \sigma_1^2) \\ Y_1, \ldots, Y_n &\sim N(\mu_2, \sigma_2^2) \end{aligned}$$

- 가설 : $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
- 검정통계량:

$$T = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}, \text{ where } s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

- 검정통계량의 분포
  1. 분산이 서로 같은 경우 : $t(n+m-2)$
  2. 다른 경우: Welch t 검정이며 자유도는

$$df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/m)^2/(m-1) + (s_y^2/n)^2/(n-1)}$$

예제: 독립표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 수면제 A, B를 각 10명의 환자에게 투여 후, 초과 수면시간 측정
- 가설 : A의 초과 수면시간(μ_1) 보다 B(μ_2)가 큰가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

n	A	B
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

	A평균	B평균	t값	자유도	p값
이분산	0.75	2.33	-1.860813	17.77647	0.0396971
등분산	NA	NA	-1.860813	18.00000	0.0395934

예제: 독립표본 t검정 (conti.)

- 등분산 검정
- 가설 : $H_0 : \sigma_1^2/sigma_2^2 = 1$ vs. H_1: not H_0
- 검정통계량 : $F = s_x^2/s_y^2 \sim F(m-1, n-1)$

F	df1	df2		p값
0.7983426	9	9		0.7427199

- 분산이 다르지 않음 : 등분산 가정의 t검정 결과를 이용함

대응비교(짝비교, Paired t-test)

- 동일한 대상에 대하여 서로 다른 처리를 한 후 처리 효과의 차이를 비교할 때
- 표본자료:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \Sigma\right)$$

Then

$$D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n \sim N(\mu_1 - \mu_2, \sigma_D^2)$$

- 검정통계량:

$$T = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t(n-1), s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

예제: 대응비교 t검정

- 운동선수 트레이닝 방법 : 100m 단거리 육상선수 10명에게 새로운 훈련방법을 도입하여 전/후 효과를 측정
- 가설 : 훈련 후 (μ_2)가 훈련 전 (μ_1)보다 효과가 있는가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

id	pre	post	d
1	12.9	12.7	-0.2
2	13.5	13.6	0.1
3	12.8	12.0	-0.8
4	15.6	15.2	-0.4
5	17.2	16.8	-0.4
6	19.2	20.0	0.8
7	12.6	12.0	-0.6
8	15.3	15.9	0.6
9	14.4	16.0	1.6
10	11.3	11.1	-0.2

사전	사후	사후-사전	t값	자유도	p값
14.48	14.53	0.05	0.2133085	9	0.41792

비율검정

- one sample 비율검정

$$X_1, \ldots, X_n \sim \text{iid } Ber(p) \Rightarrow X = \sum X_i \sim Bin(n, p)$$

- Hyperthesis:

$$H_0 : p = p_0, \text{vs } H_1 : \text{ not } H_0$$

- Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1) \text{ under } H_0$$

- 예제 : 대학생의 흡연율이 40%보다 작은가?
  - 가설: $H_0 : p = p_0, \text{vs } H_1 : \text{ not } H_0$
  - 결과

대상수	흡연자	흡연율	z값	p값
100	30	0.3	2.041241	0.0206134

비율검정 (이표본 비교)

- Data

	group 1	group 2
success	X_1	X_2
trials	n_1	n_2
success probability	$\hat{p}_1 = X_1/n_1$	$\hat{p}_2 = X_2/n_2$

- Hyperthesis:

$$H_0 : p_1 = p_2, \text{vs } H_1 : \text{ not } H_0$$

- Test statistic:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

ex: 비율검정 (이표본 비교)

- 특정 바이러스에 대한 남, 여 항체보유율을 비교,

	남	여	합계
대상자	100	150	250
항체보유자	24	64	58

- 가설: $H_0 : p_1 = p_2$ vs $H_1 : p_1 < p_2$

대상	수 항체	보유자 항체	보유율	z값	p값
남자	100	24	0.2400000	3.027496	0.0012329
여자	150	64	0.4266667	NA	NA

- 여자의 항체보유율이 남자에 비해 높다고 판단