R을 이용한 통계 기초와 데이터 분석

남현진

한성대학교

2020

강의 내용

- 점추정, 구간추청
- 귀무가설, 대립가설
- 실험군, 대조군
- 제1의 오류, 제 2의 오류
- 일표본 검정, 독립 표본 검정, 대응표본 검정
- 유의확률과 기각역

통계적 추정

"What! you have solved it already?"
"Well, that would be too much to say. I have discovered a suggestive fact, that is all"
- Dr.Watson and Sherlock Holmes, The sign of Four -

Definition

표본으로부터 통계량의 값을 구하여 그 값을 근거로 모수의 값이 얼마가 될 것이라고 추정하는 것을 통계적 추정이라고 한다.

Important theorem

- 점추정: 모수를 하나의 수치로 추정하는 것.
- 구간추정: 모수를 어떠한 범위 안의 값으로 추정하는 것.

점추정

Definition

점추정이란 모수를 추정하고자 모집단에서 임의로 추출된 n개 표본의 확률변수로 하나의 통계량을 만들고 주어진 표본으로부터 그 값을 계산하여 하나의 수치를 제시하는 것이다.

Important theorem

모수를 θ라고 하면 모수의 추정치는 모자를 씌운 형태로 $\hat{\theta}$라고 쓴다. 이 때 잘 알려진 간단한 점추정량들은 다음과 같다.

모수 θ	추정치 $\hat{\theta}$
μ	$\bar{x}=\frac{\sum(x)}{n}$
p	$\hat{p}=\frac{x}{n}$
σ^2	$s^2=\frac{\sum(x-\bar{x})^2}{n-1}$

점 추정

Examples

핸드폰을 생산하는 공장이 있다. 이 때 이 공장의 새로운 베터리 생산 공법이 적용되었는데, 새로운 베터리의 평균 지속 시간을 알고자 한다. 전체 핸드폰을 모두 실험해보는 것은 한계가 있기 때문에 표본을 10 개를 뽑아 지속 시간을 계산해보니 핸드폰 베터리 지속 시간은 다음과 같다.

$$x = (20, 22, 23, 21, 17, 23, 21, 21, 20, 16)$$

이 때 표본 x의 평균 $\bar{x}$ 는 (20+22+23+21+17+23+21+21+20+16)/10 = 20.4 이다. 따라서 베터리 공장의 새로운 베터리 공법을 사용해 생산한 베터리의 추정 지속시간은 20.4 시간이다.

구간추정

Definition

구간추정은 추정량의 분포를 이용하여 표본으로부터 모수 값을 포함하리라고 예상되는 구간을 제시하는 것이다. 이 때 제시되는 구간을 신뢰구간(confidence interval)이라 부른다.

Important theorem

모집단이 정규분포를 따르고 표준편차가 알려져 있는 경우, μ에 대한 $100(1/\alpha)\%$ 신뢰구간은 다음과 같이 정이할 수 있다.

$$P(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

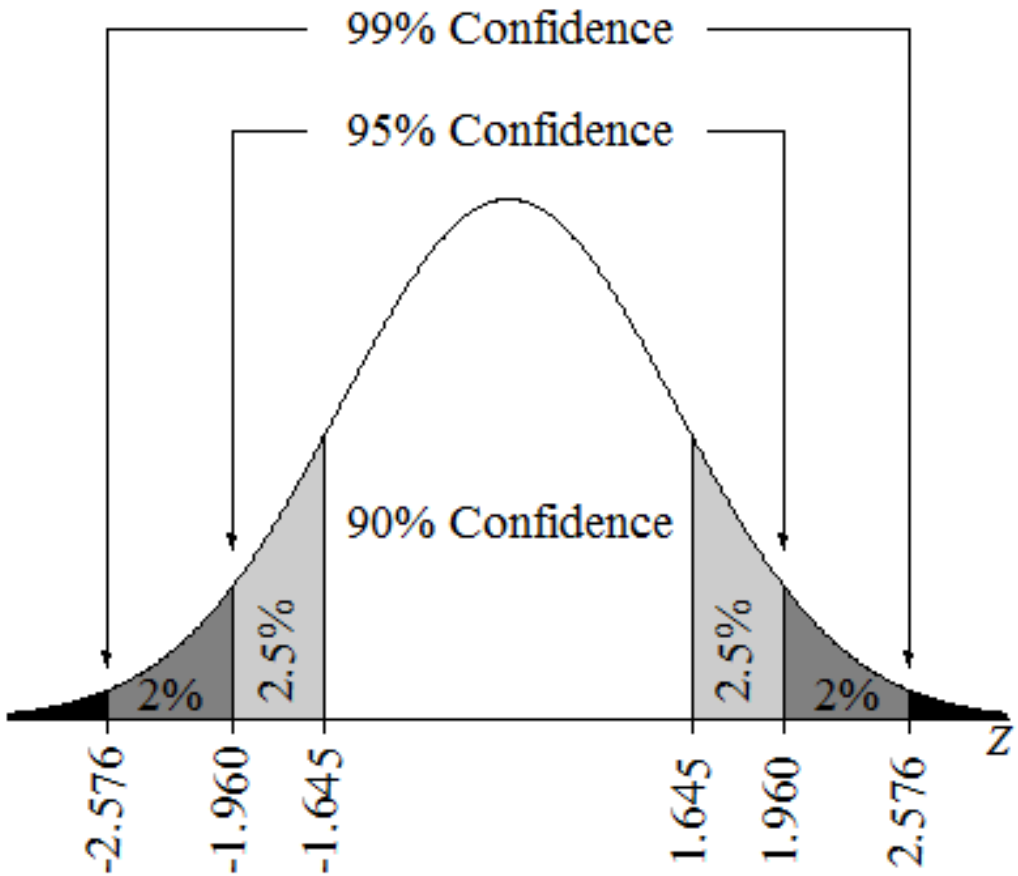
$$P(\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

따라서 μ에 대한 100(1-α)% 신뢰구간은 다음과 같다.

$$(\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

구간추정

α	신뢰구간	$Z_{\alpha/2}$
0.10	90%	1.645
0.05	95%	1.960
0.01	99%	2.576



구간추정

Examples

평균이 μ이고 표준편차가 $\sigma = 4$인 정규분포를 따르는 모집단에서 크기가 25인 표본을 추출하여 평균을 계산하였더니 $\bar{x} = 30$이었다. 이 때 μ에 대한 95% 신뢰구간을 구하면 다음과 같다.

$$P(\left|\frac{30-\mu}{4/\sqrt{25}}\right| < Z_{\alpha/2}) = 1 - 0.05$$

$$P(-Z_{0.025}\frac{4}{\sqrt{25}} < \bar{X} - \mu < Z_{0.025}\frac{4}{\sqrt{25}}) = 0.95$$

$$P(30 - Z_{0.025}\frac{4}{\sqrt{25}} < \mu < 30 + Z_{0.025}\frac{4}{\sqrt{25}}) = 0.95$$

따라서 μ에 대한 1095 신뢰구간은 다음과 같다.

$$(30 - Z_{0.025}\frac{4}{\sqrt{25}}, 30 + Z_{0.025}\frac{4}{\sqrt{25}})$$

$$(30 - 1.96\frac{4}{\sqrt{25}}, 30 + 1.96\frac{4}{\sqrt{25}})$$

$$(28.432, 31.568)$$

통계적 가설검정

Definition

모수에 대하여 어떠한 값을 가정하고 표본을 이용하여 그 가설이 합단한가를 결정하는 것을 가설 검정이라 한다.

통계적 가설검정

Definition

모집단의 미지의 모수에 대한 타당성을 확인하고자 하는 주장을 대립가설H_1이라 하고, 대립가설이 참이라는 확실한 근거가 없는 경우에 받아들이는 가설을 귀무가설H_0이라고 한다.

Important theorem

- 귀무가설은 반증의 대상이다.(효과가 없다, 차이가 없다, 다르지 않다.)
- 대립가설은 연구의 대상이다.(효과가 있다. 차이가 있다, 서로 다르다.)

Examples

- H_0: 새롭게 출시된 약의 효과가 없다.
- H_1: 새롭게 출시된 약의 효과가 있다.

실험군과 대조군

Definition

실험 결과를 도출하기 위해 인위적 또는 어떤 조작을 통해 환경 설정을 한 집단이 실험군이고 이와 달리 실험 결과가 제대로 도출되었는지의 여부를 판단하기 위해 어떤 조작이나 조건도 가하지 않은 집단을 대조군이라고 한다.

Important theorem

일반적으로 통계적 실험을 할 때에는 표본 집단에서 일정 수의 사람들을 대조군으로 분리시켜놓고 실험 전과 동일한 상태로 유지한 후 실험군에게만 변화를 주어 두 집단의 차이를 비교한다.

Examples

- 실험군 (Target Group): 새롭게 출시된 약을 먹은 그룹
- 대조군 (Control Group): 새롭게 출시된 약을 먹지 않은 그룹

귀무가설의 결과

Definition

가설 검정의 결과는 두 가지로 귀무가설 H_0이 참일 경우에는 귀무가설 H_0을 채택한다.
귀무가설 H_0이 거짓일 경우에는 귀무가설 H_0을 기각한다.

Important theorem

결과 \ 사실	H_0 : 참	H_0 : 거짓
H_0 : 채택	옳은 결정	제2종 오류
H_0 : 기각 (H_1 : 채택)	제1종 오류	옳은 결정

- 제 1종 오류란 귀무가설이 참인데도 불구하고 귀무가설을 기각하는 오류이다.
- 제 2종 오류란 귀무가설이 거짓인데도 불구하고 귀무가설을 채택하는 오류이다.

Examples

- 제 1종 오류: 약효가 없는 약을 약효가 있다고 결정을 내리는 오류이다.
- 제 2종 오류: 약효가 있는 약을 약효가 있다고 결정을 내리는 오류이다.

실험의 종류

Definition

- 일 표본 검정: 특정 집단의 평균이 어떤 숫자와 같은지 다른지를 비교
- 독립 표본 검정: 서로 다른 두개의 그룹 간의 평균 비교
- 대응 표본 검정: 하나의 집단에 대한 실험 전과 후의 비교

Examples

- 일 표본 검정: 새로운 베터리의 지속시간이 30시간이 넘는지 비교
- 독립 표본 검정: 남자와 여자 간 소득의 차이 비교
- 대응 표본 검정: 운동을 하기 전과 후의 몸무게 변화 비교

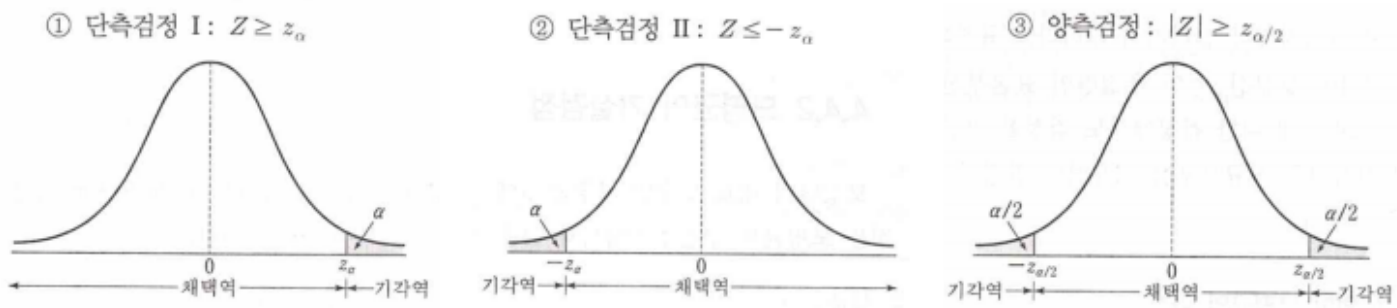
유의확률과 기각역

Definition

유의확률(p-value)이란 관측된 자료의 통계량이 귀무가설을 기각하는 방향으로 나타날 확률이다. p-value가 유의수준 α보다 크면 귀무가설을 채택하고, p-value가 유의수준 α보다 작으면 귀무가설을 기각한다.

Important theorem

검정통계량이 채택역에 속하면 귀무가설 H_0을 채택하고 검정통계량이 기각역에 포함되면 귀무가설 H_0을 기각한다.



귀무가설의 결과

Examples

다음과 같은 통계적 가설이 있다고 가정하자.

- H_0: 새롭게 출시된 약의 효과가 없다.
- H_1: 새롭게 출시된 약의 효과가 있다.

가설을 검증하기 위하여 실험을 해보니 새로운 약의 먹은 사람 중 병이 나은 사람의 비율은 80%이며 새로운 약을 먹지 않은 사람 중 병이 나은 사람의 비율은 72%이다. 이 때 가능한 결론은 다음과 같다.

- 새로운 약의 효과가 있다. 8% 차이는 약으로 인해 발생한 차이이다.
- 새로운 약의 효과가 없다. 8% 차이는 우연의 결과이다.

이 때 우연히 8% 이상의 약효 차이가 날 가능성을 유의확률(p-value)이라하며 유의확률이 유의수준 α보다 작으면 새로운 약의 효과가 있다고 결론을 내리며 귀무가설을 기각한다.

1. 일 표본 Z 검정

모분산 σ^2이 알려져 있는 경우 혹은 모집단의 모집단이 어떠한 분포를 따르고 표본의 크기가 충분히 큰 경우의($x \geq 30$) 모평균 μ에 대한 가설 검정은 다음과 같다.

Assumptions

- X의 분포가 $N(\mu, \sigma^2)$ 이거나 $n \geq 30$
- σ^2가 알려져 있는 경우

Important theorem

H_0	H_1	검정통계량	기각역
$\mu = \mu_0$	$\mu > \mu_0$	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$Z \geq z_\alpha$
$\mu = \mu_0$	$\mu < \mu_0$		$Z \leq -z_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$		$\lvert Z \rvert \geq z_{\alpha/2}$

2. 독립 표본 Z 검정

서로 독립인 두 모집단의 각각의 모분산 σ_1^1과 σ_1^2이 알려지거나 두 모집단이 어떠한 분포를 따르고 표본의 크기 n_1과 n_2이 충분히 큰 경우 ($n \geq 30$)의 두 모평균 μ_1와 μ_2에 대한 가설 검정은 다음과 같다.

Assumptions

- X_1의 분포가 $N(\mu_1, \sigma_1^2)$ 이거나 $n_1 \geq 30$
- X_2의 분포가 $N(\mu_2, \sigma_2^2)$ 이거나 $n_2 \geq 30$
- σ^1와 σ^2가 알려져 있을 때

Important theorem

H_0	H_1	검정통계량	기각역
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$	$Z \geq z_\alpha$
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$		$Z \leq -z_\alpha$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\lvert Z \rvert \geq z_{\alpha/2}$

3. 대응 표본 Z 검정

서로 대응인 두 모집단의 표본의 크기 n이 충분히 큰 경우 ($n \geq 30$)의 두 모평균 μ_1와 μ_2에 대한 가설 검정은 다음과 같다.

Assumptions

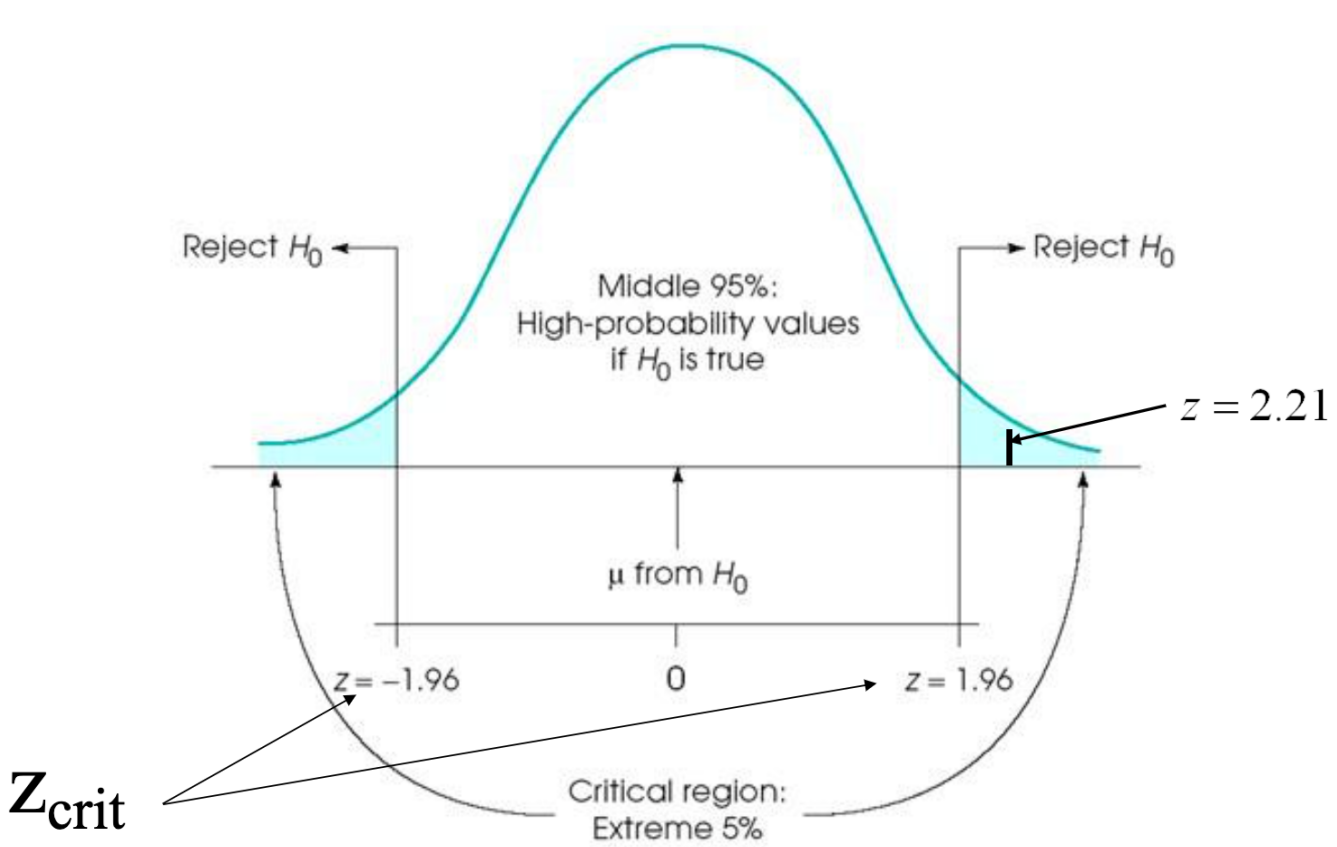
- X_1의 분포가 $N(\mu_1, \sigma_1^2)$ 이거나 $n_1 \geq 30$
- X_2의 분포가 $N(\mu_2, \sigma_2^2)$ 이거나 $n_2 \geq 30$
- σ^1와 σ^2가 알려져 있을 때
- X_1와 X_2가 종속적일 때

Important theorem

H_0	H_1	검정통계량	기각역
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 > 0$	$Z = \frac{\bar{D}}{S_D/\sqrt{n}}$	$Z \geq z_\alpha$
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 < 0$		$Z \leq -z_\alpha$
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$		$\lvert Z \rvert \geq z_{\alpha/2}$

이 때 $X_{i1} - X_{2i} = D_i$ 이고 $S_D^2 = \frac{(n-1)}{1}\Sigma_{i=1}^{n}(D_i - \bar{D})^2$

통계적 가설검정


Reject H_0
Middle 95%:
High-probability values
if H_0 is true
Reject H_0
$z = 2.21$
μ from H_0
$z = -1.96$
0
$z = 1.96$
Z_{crit}
Critical region:
Extreme 5%

통계적 가설검정

Examples

어떤 전자 회사의 기존의 베터리의 평균 수명은 100시간이며 정규분포를 따르고 있다고 알려져있다. 이 때 새로운 공법이 개발되었고 이를 사용하여 생산한 배터리는 기존의 배터리에 비해 평균 수명이 길어졌는가에 대한 실험을 해야한다. 이 주장을 확인하기 위해서 64개의 표본을 임의로 추출하여 수명을 측정한 결과 평균이 110시간이었으며 표준편차가 16시간이었다.

통계적 가설검정

위의 실험의 가설을 살펴보면 다음과 같다.

- $H_0 : \mu = 100$
- $H_1 : \mu > 100$

이 때 비교하고자 하는 값 μ_0 는 100임으로 일표본 Z검정에 해당한다. 가설 검정을 위한 표본의 분포는 $\bar{X} \sim N(110, 16^2)$ 이다. 검정 통계량을 계산해 보면
$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{110 - 100}{16/\sqrt{64}} = 5$이다. 이 때 유의수준 5% 에서 $z_\alpha = z_0.05 = 1.645$ 이다.
따라서 검정통계량 5 는 유의수준 1.645 보다 크기 때문에 귀무가설 H_0을 기각하고 대립가설 H_1을 채택한다. 따라서 전자 회사는 새로운 공법으로 인해 평균 수명이 길어졌다고 말할 수 있다.