통계 분석

2019년 2학기

강봉주

통계적 추론

통계적 추론

[개요]

- 통계적 추론은 모집단으로부터 생성된 관측된 데이터로부터 모집단의 분포에 속성(가령, 중심위치나 산포)에 대한 가설을 검증(hypothesis testing) 하거나 추정(estimation) 하는 등의 과정을 의미한다.

통계적 추론

[추정]

- $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$
- θ: 모집단 모수(parameter)
- $u(X_1, \dots, X_n)$: 통계량(statistic),
- 추정량(estimator): 모집단 모수를 추정하기 위한 통계량
- 추정값, 추정치(estimate): 표본으로부터 계산된 추정량의 값

$$\theta, \bar{X}, \bar{x}$$

통계적 추론

[추정]

- 좋은 추정량?
- $E(\bar{X}) = \theta$: 비편향 추정량(unbiased estimator)

통계적 추론

[추정]

- 좋은 추정량?
- $X_1, \ldots, X_n \sim f(x;\theta) = \theta^x (1-\theta)^{1-x}, x = 0, 1$

$f(x_1;\theta) \cdots f(x_n;\theta)$

$= \theta^{x_1}(1-\theta)^{1-x_1} \cdots \theta^{x_n}(1-\theta)^{1-x_n}$

$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

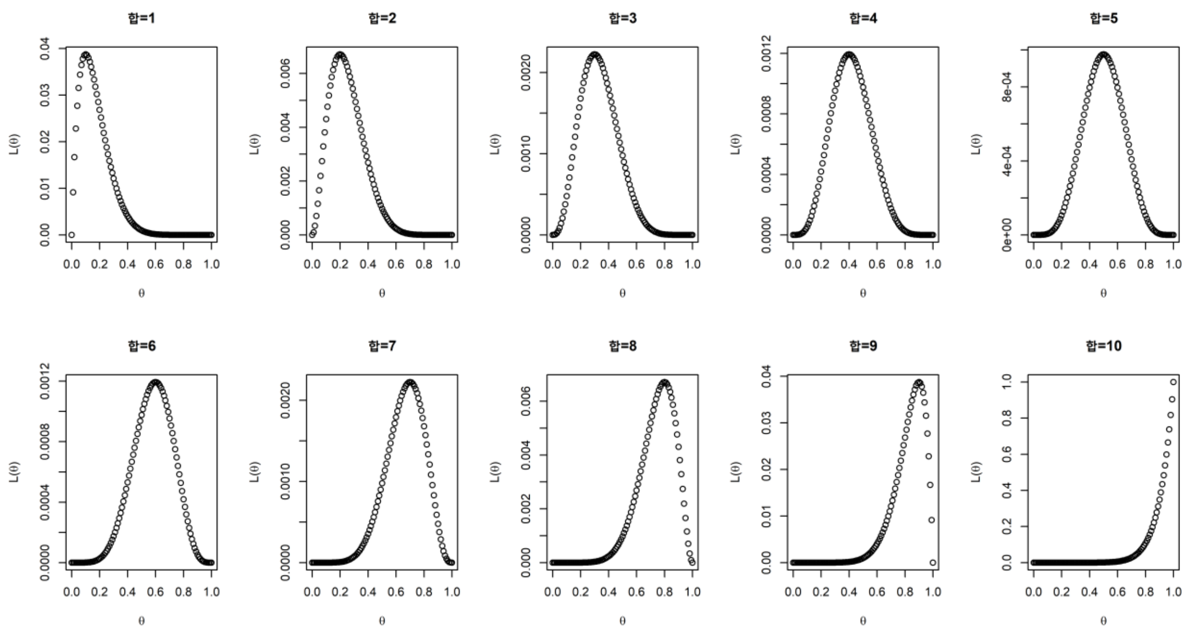
$L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$: 우도함수(likelihood function)

통계적 추론

[추정]

- 좋은 추정량?
- $L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$: 우도함수(likelihood function)

$n = 10, \sum x_i$값에 따른 우도함수의 변화



통계적 추론

[추정]

- 좋은 추정량?
- $L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$: 우도함수(likelihood function)
- 우도함수를 최대로 해주는 추정량(estimator)을
  최대우도추정량(maximum likelihood estimator: MLE)

$$\hat{\theta}^{MLE} = \frac{\sum X_i}{n}$$

통계적 추론

[추정]

- $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$: 표본 분산 통계량
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- $E\left(\frac{(n-1)S^2}{\sigma^2}\right) = n - 1 \Rightarrow \mathrm{E}\left(S^2\right) = \sigma^2$: 비편향 추정량
- 최대 우도 추정량: $\frac{n-1}{n}S^2$
- $E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2 \approx \sigma^2$: 일치 추정량(consistent)

통계적 추론

[추정]

- 점 추정값(point estimate): 하나의 값으로 θ를 추정
- 점 추정값의 신뢰도?
- $\bar{x} = 10$ 이 계산되었다 하더라도 실제 θ가 10인지를 어느 정도 신뢰할 수 있을까?

통계적 추론

[추정]

- 점 추정값(point estimate): 하나의 값으로 θ를 추정
- 점 추정값의 신뢰도?

$$\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right) \text{ 이므로 } \frac{\bar{X}-\theta}{\frac{\sigma^2}{\sqrt{n}}} \sim N(0,1)$$

$$\Pr\left(-2 < \frac{\bar{X}-\theta}{\frac{\sigma^2}{\sqrt{n}}} < 2\right) = 0.954$$

$$\Pr\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} < \theta < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954$$

통계적 추론

[추정]

- 점 추정값(point estimate): 하나의 값으로 θ를 추정
- 점 추정값의 신뢰도?
- 구간 $\left(\bar{X} - \frac{2\sigma}{\sqrt{n}},\ \bar{X} + \frac{2\sigma}{\sqrt{n}}\right)$: σ가 알려진 경우에 확률 구간
- 확률 구간이 모평균을 품을 확률: 95.4%
- $\left(\bar{x} - \frac{2\sigma}{\sqrt{n}},\ \bar{x} + \frac{2\sigma}{\sqrt{n}}\right)$: 계산된 표본 구간
- 모평균을 품을 신뢰수준(confidence level): 95.4%
- 구간 추정: 모수의 범위를 추정

통계적 추론

[추정]

- 구간 추정
- $|\bar{x} - \theta| < 2\frac{\sigma}{\sqrt{n}}$
- $2\frac{\sigma}{\sqrt{n}}$: 오차 한계
- 오차 한계는 신뢰 수준이 정해지고 나면 결정됨
- 표준 오차(standard error): 추정량의 표준편차
- $\sqrt{\mathrm{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$

통계적 추론

[추정]

- 모평균의 추정
- $X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$
- $100(1-\alpha)\%$ 신뢰수준에서 모평균에 대한 신뢰구간?

$$\Pr\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$z_{\frac{\alpha}{2}} = ?$$

통계적 추론

[추정]

- 모평균의 추정
- $X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$
- $100(1-\alpha)\%$ 신뢰수준에서 모평균에 대한 신뢰구간?

$$\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \qquad \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right)$$

신뢰수준이 99%인 신뢰구간

$$z_{\frac{\alpha}{2}} = 2.58$$

통계적 추론

[추정]

예제)

1) 신뢰수준이 90%, 95%, 99%에 해당하는 신뢰구간 분위수를 구하세요.

통계적 추론

[추정]

예제)

1) 신뢰수준이 90%, 95%, 99%에 해당하는 신뢰구간 분위수를 구하세요.

```
confidence_level=c(0.9, 0.95, 0.99)
alpha = 1-confidence_level
z_alpha_half <- qnorm(1-alpha/2)
print(z_alpha_half)
```

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$ 일 것이라는 추측(conjecture) 또는 주장: 통계적 가설
- 통계가설은 2가지
- $\theta \leq 75$: 영 가설(null hypothesis), H_0, 지우기 위한 가설(nullify)
- $\theta > 75$: 대립, 대안 가설(alternative hypothesis), H_1

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 어떻게 증명? 검증(test)?
- 예: 크기가 25인 임의 표본을 구성

$$C = \{(x_1, \ldots, x_{25}) | x_1 + \cdots + x_{25} > 25 \times 75\}$$

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- $C = \{(x_1, \ldots, x_{25}) | x_1 + \cdots + x_{25} > 25 \times 75\}$
- $\bar{x} > 75$ 이면 즉, $(x_1, \ldots, x_{25}) \in C$ 이면 영가설을 기각(reject)
- $\bar{x} \leq 75$ 이면 영가설을 기각하지 못하고 채택(accept)
- 올바른 방법인가?

통계적 추론

[통계적 가설]

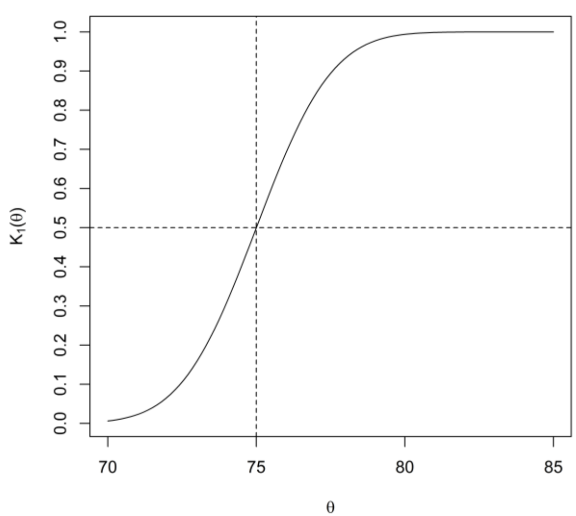
- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 영가설을 기각할 확률*: 계산된 통계량이 기각역에 포함될 확률*
- $\Pr\big((X_1, \ldots, X_n) \in C\big) = \Pr(\bar{X} > 75) =$

  $\Pr\left(\frac{\bar{X}-\theta}{10/\sqrt{25}} > \frac{75-\theta}{10/\sqrt{25}}\right) = 1 - \Phi\left(\frac{75-\theta}{2}\right)$: 검증력 함수(power function)
- 검증력 값(power): 하나의 θ 값에 따른 검증력 함수값

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- $K_1(\theta) = 1 - \Phi\left(\frac{75-\theta}{2}\right)$



$\theta = 75$이면 영가설을 기각할 확률이 0.5

영가설이 참(true)임에도 불구하고 이를 기각할 확률이 0.5인 것이다. 가령 건강한 환자에게 특정병에 걸렸다고 잘못 얘기할 확률이 0.5인 것과 동일하다. 뭔가 이러한 검증은 문제가 있다고 볼 수가 있다. 가급적 영가설이 참일 때 이를 기각할 확률을 줄이고 싶다.

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 기각역의 수정
- $\Pr\left((X_1, \ldots, X_n) \in C\right) = \Pr(\bar{X} > 78) =$

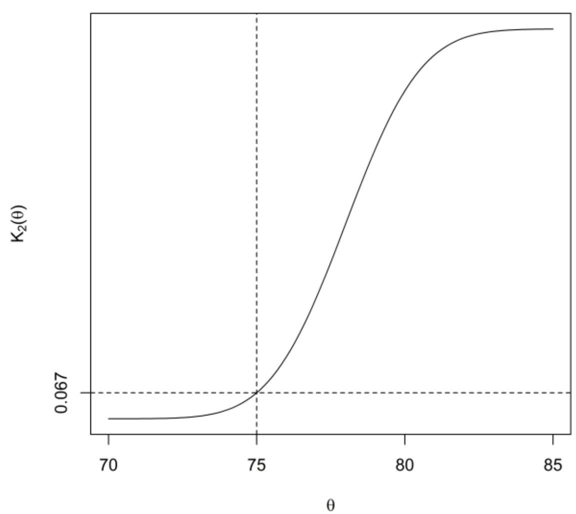
$$\Pr\left(\frac{\bar{X} - \theta}{10/\sqrt{25}} > \frac{78 - \theta}{10/\sqrt{25}}\right) = 1 - \Phi\left(\frac{78 - \theta}{2}\right)$$

- $K_2(\theta) = 1 - \Phi\left(\frac{78-\theta}{2}\right)$

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- $K_2(\theta) = 1 - \Phi\left(\frac{78-\theta}{2}\right)$



$\theta = 75$일 때 영가설을 기각할 확률이 0.067

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 영가설이 참일 때 이를 기각할 확률은 작게 하면서 대립가설이 참일 때 이를 채택할 확률을 크게 하는 검증
- 위를 만족하는 기각역을 찾아야 함
- 예를 들어서 $K_3(75) = 0.159$ 이고 $K_3(77) = 0.841$인 성질을 만족하는 기각역의 경계 값은?

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- $K_3(75) = 0.159$ 이고 $K_3(77) = 0.841$인 성질을 만족하는 기각역의 경계 값은?

$$K_3(\theta) = \Pr(\bar{X} > c) = 1 - \Phi\left(\frac{c - \theta}{\frac{10}{\sqrt{n}}}\right)$$

$$K_3(75) = 0.159 \rightarrow \ \Phi\left(\frac{c - 75}{\frac{10}{\sqrt{n}}}\right) = 1 - 0.159 = 0.841$$

$$K_3(77) = 0.841 \rightarrow \ \Phi\left(\frac{c - 77}{\frac{10}{\sqrt{n}}}\right) = 1 - 0.841 = 0.159$$

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- $K_3(75) = 0.159$ 이고 $K_3(77) = 0.841$인 성질을 만족하는 기각역의 경계 값은?

```
> qnorm(0.841)
[1] 0.9985763
> qnorm(0.159)
[1] -0.9985763
```

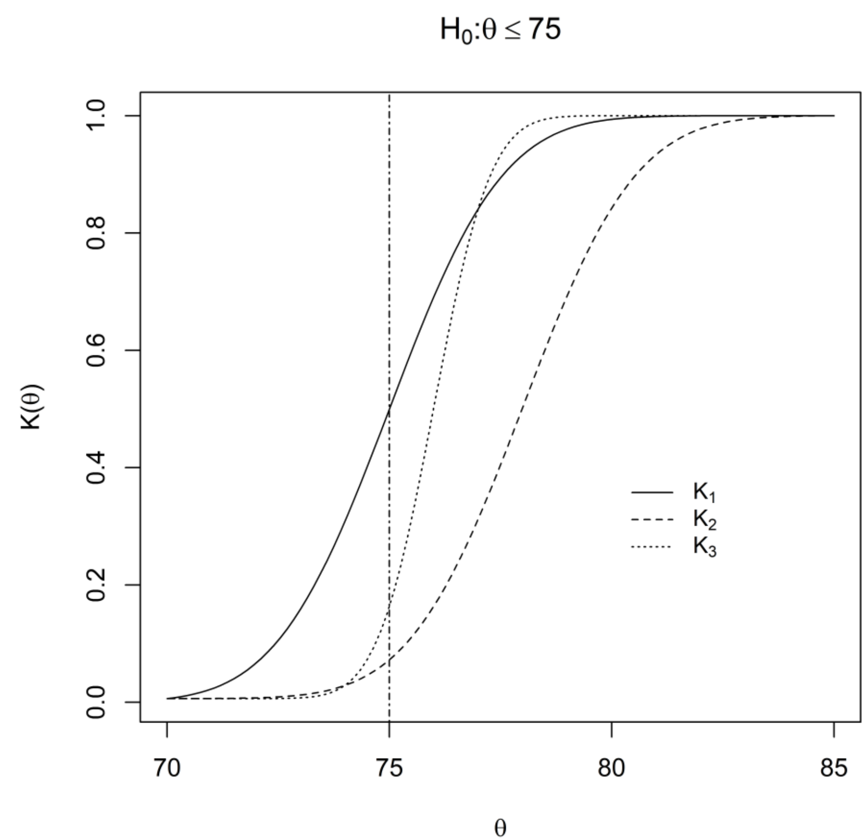
$\frac{c-75}{\frac{10}{\sqrt{n}}} = 1$ 이고 $\frac{c-77}{\frac{10}{\sqrt{n}}} = -1$ 이므로 대략 $n = 100, c = 76$

주어진 검증력 값을 확보하기 위해서는 최소 100개의 임의표본이 필요하며 이 때 $\bar{x} > 76$ 이면 주어진 가설을 원하는 성능으로 검증할 수 있다.

통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 검정력 함수 비교



통계적 추론

[통계적 가설]

- $X \sim N(\theta, 10^2)$
- $\theta > 75$
- 기각역(critical region): 영가설을 기각하기 위한 표본 공간의 부분 집합
- 유의 수준(significance level), 기각역의 크기(size): 영가설이 참일 때 검증력 함수의 최대값
- 각 검증력 함수의 유의 수준: $(K_1, K_2, K_3) = (0.5, 0.067, 0.159)$
- 영가설이 참임에도 불구하고 이를 기각할 확률(FPR)을 줄이고 영가설이 거짓인 경우에 이를 기각할 확률(TPR) 즉, 대립가설이 참일 때 이를 채택할 확률은 높이는 방향으로 기각역 정의

통계적 추론

[통계적 가설]

- 기각역의 정의: 유의 수준의 결정 후에 정의됨
- $X_1, \dots, X_{100} \sim (\mu = \theta, \sigma = 9.4)$
- $H_0: \theta = 75,\ H_1: \theta < 75$
- $\frac{\bar{X}-\theta}{9.4/\sqrt{100}} \sim N(0,1)$
- $K(\theta) = \Pr(\bar{X} < c) = \Phi\left(\frac{c-\theta}{9.4/\sqrt{100}}\right)$
- $K(75) = 0.05$
- $\frac{c-75}{9.4/\sqrt{100}} = -1.64$
- $c = 73.45$

```
> qnorm(0.05)
[1] -1.644854
> c <- 75 + qnorm(0.05)*9.4/sqrt(100)
> c
[1] 73.45384
```

통계적 추론

[통계적 가설]

- 기각역: $C = \{(x_1, \dots, x_{100}) \mid \bar{x} < 73.45\}$
- $\bar{x} = 73.5$: 영가설이 참일 때 기각할 수 없음
- $\Pr(\bar{X} < 73.5 | H_0: \theta = 75) = \Phi\left(\frac{73.5-75}{9.4/\sqrt{100}}\right) = 0.055$: 유의 확률(significance probability), P값(P value)
- P값:사전 정의된 유의수준보다 작으면 작을수록 즉, 희귀한 사건이 발생한 것이므로 영가설이 기각될 가능성이 높아진다.

통계적 추론

[통계적 가설]

- 가설 검증 절차

순서	내용	비고
1	모집단 분포에 대한 가설은 세운다.	
2	가설에 적합한 검정 통계량을 정의한다.	
3	검정통계량의 관측값을 계산한다.	
4	영가설이 참이라는 가정하에 관측값의 P값을 계산한다.	
5	P값과 유의수준과의 비교를 통하여 영가설과 대립가설의 기각 및 채택 여부를 결정한다.	기각역을 계산하여 검정통계량의 관측값과 비교할 수 있다.

통계적 추론

[통계적 가설]

예제)

한 보험회사에서 고객들의 평균보험료는 200,000원이고 표준편차는 100,000원이다. 최근에 추가판매 캠페인 등을 통하여 이 보험회사는 고객들의 평균보험료가 높아졌을 것이라고 판단하였다. 이를 알아보기 위하여 총 100명의 고객을 추출하여 표본평균 값을 계산해본 결과 220,000원이 나왔다. 통계적으로 유의미한 지를 알아보자.

통계적 추론

[통계적 가설]

- 가설 검증 절차

순서	내용	비고
1	모집단 분포에 대한 가설은 세운다.	$H_0: \theta = 200000,\ H_1: \theta > 200000$
2	가설에 적합한 검정 통계량을 정의한다.	$Z = \frac{\bar{X} - \theta_0}{\sigma_0/\sqrt{n}}$
3	검정통계량의 관측값을 계산한다.	$z = \frac{\bar{x} - \theta_0}{\sigma_0/\sqrt{n}} = \frac{220000 - 200000}{100000/\sqrt{100}} = 2$
4	영가설이 참이라는 가정하에 관측값의 P값을 계산한다.	$\Pr(Z > 2) = 1 - \Phi(2) = 0.023$
5	P값과 유의수준과의 비교를 통하여 영가설과 대립가설의 기각 및 채택 여부를 결정한다.	유의수준 0.05에서 통계적으로 유의하다(significant)