

## 제7장 대표본에서의 추론

### 7.1 서론

#### (1) 통계적 추론 (statistical inference)

: 표본자료를 분석함으로써 모수에 대한 결론을 도출하는 과정

#### (2) 모수에 대한 추론

: 전체 모집단을 조사하는 대신 그 일부분인 표본자료를 이용하기 때문에 모수에 대한 추론은 언제나 불확실성을 내포하게 된다.

- 통계적 추론은 이 불확실성을 규정할 수 있어야 한다.
- 확률의 개념과 통계량의 표본분포는 중요한 역할을 한다.

#### (3) 조사목적에 따른 추론의 유형

##### ① 모수의 추정 (estimation of parameters)

: 미지의 참값에 대한 추측값 혹은 추정값을 얻되 그 정밀도를 함께 구하는 것이 목적인 경우

- 점추정 (point estimation)
- 구간추정 (interval estimation)

##### ② 통계적 가설검정 (testing of statistical hypotheses)

: 표본자료가 모수의 참값에 대한 조사자의 추측을 뒷받침하는지 또는 반증하는지를 결정하는 것이 목적인 경우

### [예제 7.2]

정부기관의 조사 목적: 어떤 도시의 실업률  $p$  조사

( $p$ 의 참값은 알려져 있지 않음)

이 도시 노동인구 중 임의추출된 500명 중 41명이 실업자

$$\rightarrow \text{표본실업률: } \hat{p} = \frac{41}{500} = 0.082$$

표본실업률  $\hat{p}$ 을 통해  $p$ 에 대한 정보를 얻을 수 있지만 이 표본실업률  $\hat{p}$ 에는 오차가 존재한다. 이 오차의 한계를 추정하고  $p$ 가 속할 것으로 기대되는 적절한 구간을 구할 수 있고,  $p$ 가 이미 알려진 전국의 실업률보다 낮다는 가설을 검정할 수도 있다.

## 7.2 모평균의 점추정

(1) 점추정량 (point estimator): 모수를 추정하기 위하여 사용되는 통계량

(2) 표준오차 (standard error): 추정량의 표준편차

(3) 예

- 모수: 모평균  $\mu$
- 자료:  $X_1, \dots, X_n$  (표본크기가  $n$ 인 확률표본)
- 점추정량 (point estimator): 표본평균  $\bar{X}$
- 표준오차 (standard error: 추정량의 표준편차)

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad \text{여기서 } \sigma \text{는 미지의 값이므로 } s \text{로 대체}$$

$$- 100(1-\alpha)\% \text{ 오차한계(margin of error)} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

만약  $n$ 이 충분히 크다면

### [예제 7.3]

어린 소나무의 성장을 연구하기 위하여 1년생 붉은 소나무 묘목 40그루의 키를 조사하였다. 이 자료의 기술통계치들은 다음과 같다.

표본평균  $\bar{x} = 1.715$ , 표본표준편차  $s = 0.475$

표본중앙값 = 1.6,  $Q_1 = 1.5$ ,  $Q_3 = 2.0$

이 자료를 이용하여 묘목 키의 모평균을 추정하고 95% 오차한계를 구하라.

2.6	1.9	1.8	1.6	1.4	2.2	1.2	1.6
1.6	1.5	1.4	1.6	2.3	1.5	1.1	1.6
2.0	1.5	1.7	1.5	1.6	2.1	2.8	1.0
1.2	1.2	1.8	1.7	0.8	1.5	2.0	2.2
1.5	1.6	2.2	2.1	3.1	1.7	1.7	1.2

【풀이】

$$\bar{x} = \frac{\sum_{i=1}^{40} x_i}{40} = 1.715, \quad s = \sqrt{\frac{\sum_{i=1}^{40} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{40} (x_i - 1.715)^2}{39}} = \sqrt{0.2254} = 0.475$$

95% 오차한계:  $1 - \alpha = 0.95, \alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025, z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

$$\frac{1.96s}{\sqrt{n}} = \frac{1.96 \times 0.475}{\sqrt{40}} = 0.15(cm)$$

□ 표본크기의 결정(the determination of sample size  $n$ )

$$\boxed{\begin{aligned} z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= d \quad \text{여기서 } d : \text{오차한계} \\ n &= \left[ \frac{z_{\alpha/2} \sigma}{d} \right]^2 \end{aligned}}$$

□  $100(1 - \alpha)\%$ 의 신뢰도를 가지고 추정오차  $|\bar{X} - \mu|$ 가  $d$ 를 넘어서지 않기 위하여 요구되는 표본의 크기는  $n \geq \left[ \frac{z_{\alpha/2} \sigma}{d} \right]^2$ 를 만족하는 최소의 정수  $n$ 이다.

[예제 7.4]

호수 물 1리터당 포함된 평균 인산염 함유량을 추정하려고 한다. 지금까지의 연구에 의하면 표준편차는 상당히 안정적인  $\sigma = 4$ 라고 알려져 있다. 90%의 신뢰도로 추정오차가 0.8을 넘지 않기 위해서는 표본의 크기가 얼마나 되어야 하는가?

【풀이】

$$\sigma = 4, 1 - \alpha = 0.9 \rightarrow \alpha = 0.1, \frac{\alpha}{2} = 0.05, z_{0.05} = 1.645, d = 0.8$$

$$n \geq \left[ \frac{z_{\alpha/2} \sigma}{d} \right]^2 = \left[ \frac{1.645 \times 4}{0.8} \right]^2 = 67.65$$

### 7.3 모평균 $\mu$ 의 신뢰구간

(1) 모평균  $\mu$ 의  $100(1-\alpha)\%$  신뢰구간

모집단이 정규분포를 따르고  $\sigma$ 가 알려져 있는 경우

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

#### [예제 7.5]

평균  $\mu$ , 표준편차  $\sigma=8$ 인 정규모집단으로부터 크기  $n=25$ 인 표본을 추출한 후 표본평균을 계산한 결과  $\bar{x}=42.7$ 이었다.  $\mu$ 의 95% 신뢰구간을 구하라.

#### 【풀이】

모집단의 분포: 정규분포  $\rightarrow \bar{X}$ 의 분포: 정규분포

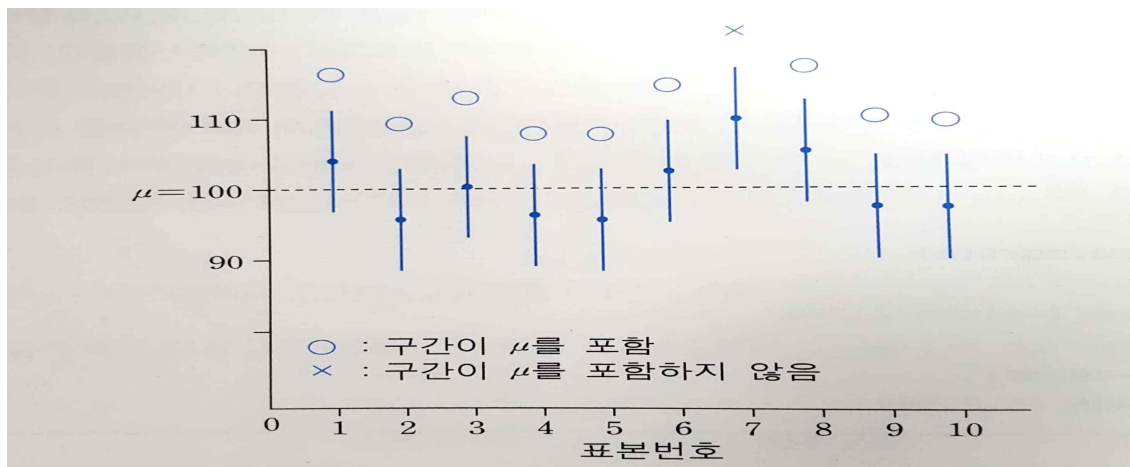
표본평균의 관측값  $\bar{x}=42.7$ 을 이용하여  $\mu$ 의 95% 신뢰구간을 구하면

$$\left( 42.7 - 1.96 \frac{8}{\sqrt{25}}, 42.7 + 1.96 \frac{8}{\sqrt{25}} \right) = (39.6, 45.8)$$

(2) 신뢰구간을 해석하는 방법

$$P\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95$$

$\Rightarrow$  동일한 모집단으로부터 반복 추출된 신뢰구간들의 95%가  $\mu$ 를 포함하고 있다.



(3) 대표본에서  $\mu$ 의  $100(1-\alpha)\%$  신뢰구간

$n$ 이 충분히 크고  $\sigma$ 가 알려져 있지 않는 경우

$$\left( \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

### [예제 7.6]

어떤 대도시의 식당 종업원들의 주당 평균 수입을 추정하기 위하여 75명의 식당 종업원들을 표본추출하여 주급을 조사하였다. 표본평균과 표본표준편차를 계산한 결과  $\bar{x}=150,000$ 원,  $s=10,000$ 원이었다. 주당 평균 소득의 (1) 90% (2) 80% 신뢰구간을 구하라.

### 【풀이】

$n=75(>30) \rightarrow$  표본평균  $\bar{X}$ 의 분포를 정규분포로 근사시킬 수 있다.

$$(1) 1-\alpha=0.9, \alpha=0.1 \rightarrow \frac{\alpha}{2}=0.05, z_{\frac{\alpha}{2}}=z_{0.05}=1.645$$

모평균  $\mu$ 의 90% 신뢰구간은

$$\begin{aligned} & \left( \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right) \\ &= \left( 150,000 - 1.645 \frac{10,000}{\sqrt{75}}, 150,000 + 1.645 \frac{10,000}{\sqrt{75}} \right) \\ &= (150,000 - 1,899.482, 150,000 + 1,899.482) \end{aligned}$$

$$(2) 1-\alpha=0.8, \alpha=0.2 \rightarrow \frac{\alpha}{2}=0.1, z_{\frac{\alpha}{2}}=z_{0.1}=1.28$$

모평균  $\mu$ 의 80% 신뢰구간은

$$\begin{aligned} & \left( \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right) \\ &= \left( 150,000 - 1.28 \frac{10,000}{\sqrt{75}}, 150,000 + 1.28 \frac{10,000}{\sqrt{75}} \right) \\ &= (150,000 - 1,478.017, 150,000 + 1,478.017) \end{aligned}$$

$\Rightarrow$  (1)과 (2) 비교: 80% 신뢰구간이 90% 신뢰구간보다 짧음을 알 수 있다. 짧은 구간의  $\mu$ 의 위치를 더욱 더 정밀하게 추정하는 것 같지만, 장기적 관점에서 모수가 참값을 포함하게 되는 상대도수는 낮아지게 된다.

## 7.4 모평균의 가설검정

※ 문제: 새로 생산되는 계란은 콜레스테롤이 적은가?

어떤 양계학자는 건강식품 개발의 일환으로 콜레스테롤의 양이 적은 계란의 생산을 위한 새로운 닭 사료 개발을 연구 중이다. 기존의 표준사료를 사용했을 때 1등급 계란의 콜레스테롤의 양은 평균  $\mu = 270\text{mg}$ , 표준편차  $\sigma = 24\text{mg}$ 을 가진다고 알려져 있다. 이 양계학자는 새로 개발된 사료를 사용하면 평균 콜레스테롤을 줄일 수 있을 것이라고 믿고 있다. 이를 증명하기 위하여, 새로운 사료를 먹인 닭으로부터 생산된 계란 중 38개를 임의추출한 다음 표본평균  $\bar{X}$ 를 계산한다고 하자. 실제 모평균이 270mg보다 작다는 주장을 통계적으로 증명하기 위해서 이 결과를 어떻게 이용해야 하는가?

### (1) 가설 설정

①  $H_0$ (null hypothesis; 귀무가설):  $H_0 : \mu = 270$

②  $H_1$ (alternative hypothesis; 대립가설)

▸ one-sided  $H_1$  (단측대립가설):  $H_1 : \mu < 270$

cf)two-sided  $H_1$  (양측대립가설):  $H_1 : \mu \neq 270$

### (2) 검정통계량과 기각역의 형태 결정

#### ① 검정통계량

-  $n$ 이 충분히 크다면 모집단의 분포에 상관없이, 표본평균  $\bar{X}$ 는 평균  $\mu$ , 표준편차  $\frac{\sigma}{\sqrt{n}}$ 인 정규분포를 따른다. (이 경우  $\mu = 270$ ,  $n = 38$ ,  $\sigma = 24$ )

- 표본평균  $\bar{X}$ 는 평균  $\mu = 270$ , 표준편차  $\frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{38}}$ 인 정규분포를 따른다.

- 표준화된 검정통계량  $Z \sim N(0, 1)$ 을 사용할 수 있다.  $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - 270}{24 / \sqrt{38}}$

#### ② 기각역의 형태

(3) 주어진 유의수준  $\alpha$ 에 대하여 기각역 결정

① 유의수준(significance level)  $\alpha$ :  $\alpha = 0.05$

② 기각치(critical value) and 검정기준(criteria for test)

만약  $\bar{X} \leq c$ 이면  $H_0$ 을 기각할 수 있다.

만약  $\bar{X} > c$ 이면  $H_0$ 을 기각할 수 없다. 여기서  $c$ 는 기각치(critical value)

• 기각치 계산에  $\bar{X}$ 를 사용

$$P[Z \leq -1.645] = 0.05$$

$$c = 270 - 1.645 \frac{24}{\sqrt{38}} = 270 - 6.4 = 263.60$$

• 기각치 계산에  $Z$ 를 사용

$$P[Z \leq -1.645] = 0.05 \Rightarrow c = -1.645$$

③ 기각역(rejection region):  $R: \bar{X} \leq 263.6$  또는  $R: Z \leq -1.645$

(4) 자료를 가지고 검정통계량 계산: data  $\bar{x} = 261, s = 22$

• 검정통계량 계산에  $\bar{x}$  사용

$R: \bar{X} \leq 263.6, \bar{x} = 261 \Rightarrow H_0$ 를 기각할 수 있다.

• 검정통계량 계산에  $z$  사용

$$z = \frac{261 - 270}{22 / \sqrt{38}} = -2.52$$

$R: Z \leq -1.645, z = -2.52 \Rightarrow H_0$ 를 기각할 수 있다.

(5) 결론 도출하기:  $H_0$ 를 기각할 수 있다.

$\Rightarrow$  이 자료는 새로운 자료를 사용할 때 평균 콜레스테롤 함유량이 줄어든다는 양계학자의 주장을 강력히 뒷받침하고 있다.

※  $H_1$ 의 형태와 상응하는 기각역

$H_1$	기각역 (rejection region)
$H_1: \mu > \mu_0$	$R: Z \geq z_\alpha$
$H_1: \mu < \mu_0$	$R: Z \leq -z_\alpha$
$H_1: \mu \neq \mu_0$	$R:  Z  \geq z_{\alpha/2}$

※ 제1종 오류(type I error) vs. 제2종 오류(type II error)

표본에 의한 결정	미지의 실제 상황	
	$H_0$ 이 참 ( $\mu = 270$ )	$H_1$ 이 참 ( $\mu < 270$ )
$H_0$ 기각	$H_0$ 을 잘못 기각함 (제 I 종 오류)	옳은 결정
$H_0$ 채택	옳은 결정	$H_0$ 를 잘못 채택함 (제 II 종 오류)

① 개념

- 제 I 종 오류:  $H_0$ 가 참인데  $H_0$ 를 기각함
- 제 II 종 오류:  $H_1$ 이 참인데  $H_0$ 를 기각하지 않음
- $\alpha$ : 제 I 종 오류를 범할 확률  $\rightarrow$  유의수준(significance level)
- $\beta$ : 제 II 종 오류를 범할 확률

② 예

- 제 I 종 오류:  $H_0(\mu = 270)$ 가 참인데  $H_0$ 를 기각하는 경우
  - $\alpha = P[\bar{x} \leq \bar{x}_c]$
  - $\alpha = P[z \leq c]$
- 제 II 종 오류:  $H_1(\mu < 270)$ 가 참인데  $H_0(\mu = 270)$ 를 기각하지 않은 경우
  - $\beta = P[\bar{x} > \bar{x}_c]$
  - $\beta = P[z > c]$



※  $p$ -값( $p$ -value) [유의확률(significance probability)]

① 정의

- $H_0$ 를 기각할 수 있는 최소한의 유의수준
- 대립가설을 뒷받침하는 정도

② 예

$$\square Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - 270}{24 / \sqrt{38}}$$

$$\square \bar{x} = 261, s = 22$$

$$\square z = \frac{261 - 270}{22 / \sqrt{38}} = -2.52$$

$$\square p = P[Z \leq z] = P[Z \leq -2.52] = 0.0059$$

[예제 7.8]

새로운 다이어트 프로그램의 회원모집광고에서는 프로그램 참가자들이 5주 내에 평균 10kg보다 많은 체중감량을 할 수 있다고 선전하고 있다. 56명의 참가자들의 5주 동안의 감량자료로부터 표본평균과 표본표준편차를 계산한 결과 각각  $\bar{x} = 10.7\text{kg}$ ,  $s = 4.6\text{kg}$ 이었다. 이 자료에 의하면 광고의 선전문구가 사실이라고 할 수 있는가? 유의수준  $\alpha = 0.05$ 에서 검정하라. 또한  $p$ -값을 계산하고 그 결과를 해석하라.

【풀이】

① 가설 설정

- $\square H_0$ (null hypothesis; 귀무가설):  $H_0 : \mu = 10$
- $\square H_1$ (alternative hypothesis; 대립가설):  $H_1 : \mu > 10$

② 주어진 유의수준  $\alpha$ 에 대하여 기각역 결정

- $\square$  유의수준(significance level)  $\alpha$ :  $\alpha = 0.05$
- $\square$  기각치(critical value):  $P[Z \leq 1.645] = 0.95 \Rightarrow c = 1.645$
- $\square$  기각역(rejection region):  $R: Z \geq 1.645$

③ 자료를 사용하여 검정통계량을 계산

□ 표본 통계량:  $\bar{x}=10.7$ ,  $s=4.6$

□ 유의수준(significance level)  $\alpha$ :  $\alpha=0.05$

□  $z = \frac{10.7 - 10}{4.6 / \sqrt{56}} = 1.139$

□ 기각치(critical value  $c$ ):  $P[Z \leq 1.645] = 0.95 \Rightarrow c = 1.645$

□ 결정 기준

만약  $z \geq c$ 이면,  $H_0$ 를 기각할 수 있다.

만약  $z < c$ 이면,  $H_0$ 를 기각할 수 없다. 여기서  $c$ 는 기각치

□ 기각역 (rejection region):  $R: Z \geq 1.645$

□ 결론

$z=1.139$ ,  $c=1.645 \Rightarrow z < c \Rightarrow H_0$ 를 기각할 수 없다.

$\Rightarrow$  유의수준  $\alpha=0.05$ 에서 광고에서의 주장  $\mu > 10$ 은 입증되지 않는다.

④  $p$ -값 ( $p$ -value) [유의확률(significance probability)]

$p = P[Z \geq z] = P[Z \geq 1.139] = 0.1274$

$\Rightarrow 0.1274$ 가  $H_0$ 을 기각할 수 있는 최소의 유의수준  $\alpha$ 이다. 일반적으로 (무시되어도 좋을 만큼) 작다고 할 수 없으므로, 이 자료는  $H_0$ 를 기각할 만큼 강력한 근거를 제공하지 못한다.

## 7.5 모비율에 관한 추론

### [예제 7.2]

정부기관에서 어떤 도시의 실업률을 조사하고자 한다. 이 도시의 노동인구 중 작은 비율만큼만 표본추출하여 실업자의 수를 세면 이 조사는 신속하고 효과적으로 행해질 수 있을 것이다. 임의로 추출된 500명의 사람들 중 41명이 실업자인 것으로 밝혀졌다고 하자. 표본실업률은  $\hat{p} = \frac{41}{500} = 0.082$ 가 된다. 표본실업률  $\hat{p} = 0.082$ 를 통하여  $p$ 에 대한 정보를 얻을 수 있지만 이 표본실업률은 모집단 중 일부만의 실업률이기 때문에 오차가 있을 수 밖에 없다. 이 오차의 한계를 추정하고  $p$ 가 속할 것으로 기대되는 적절한 구간을 구할 수도 있고, 또한  $p$ 가 이미 알려진 전국의 실업률보다 낮다는 가설을 검정할 수도 있다.

#### (1) 표본비율

모집단으로부터  $n$ 개의 원소를 임의로 추출하여, 표본 중 어떤 속성을 가진 것의 개수를  $X$ 라고 하자. 이때 모비율  $p$ 의 가장 자연스러운 추정량은 표본비율  $\hat{p} = \frac{X}{n}$ 가 된다.

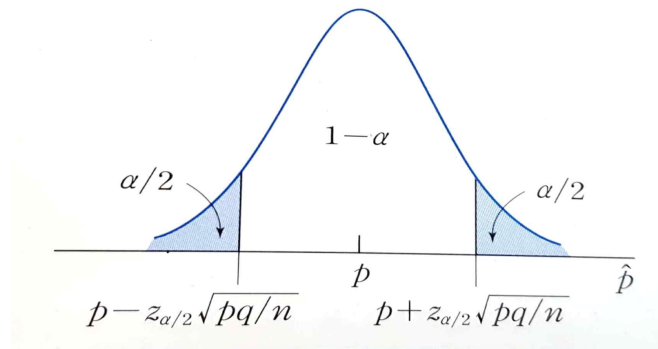
#### (2) 이항분포의 정규근사

$n$ 이 클 때 이항확률변수  $X$ 는 근사적으로 평균  $np$ , 표준편차  $\sqrt{npq}$ 인 정규분포를 따른다. 즉  $Z = \frac{X - np}{\sqrt{npq}}$ 는 근사적으로 표준정규분포를 따른다.

(3) 모비율의 정규근사

$$Z = \frac{(X - np)/n}{\sqrt{npq}/n} = \frac{(X/n) - p}{\sqrt{pq/n}} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

$\hat{p}$ 이 근사적으로 평균  $p$ , 표준편차  $\sqrt{pq/n}$ 인 정규분포를 따른다.



(4) 모비율의 점추정(point estimation)

① 모수: 모비율  $p$

② 자료:  $X$ 는 크기가  $n$ 인 확률표본 중 특정한 속성을 갖는 개수

③ 추정량:  $\hat{p} = \frac{X}{n}$ ,  $S.E.(\hat{p}) = \sqrt{\frac{pq}{n}}$ ,  $S.E.(\hat{p})$ 의 추정량 =  $\sqrt{\frac{\hat{p}\hat{q}}{n}}$

④  $n$ 이 클 때,  $100(1-\alpha)\%$ 의 근사오차한계는  $z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}$

**[예제 7.10]**

매달 특선 품목을 판매하는 어떤 인터넷 판매회사에서는 새로운 품목을 시판하고자 한다. 9,000명이 넘는 가입 회원 중 250명을 임의로 추출하여 견본을 발송한 결과 70명이 구입의사가 있다고 한다. 전체 가입 회원 중 이 상품을 살 것으로 기대되는 회원의 모비율  $p$ 에 대한 점추정값과 95.4% 오차한계를 구하라.

**【풀이】**

표본의 크기: 전체 가입회원의 일부분

→ 구입의사가 있는 사람의 수: 이항확률변수

$$n = 250, X = 70, \hat{p} = \frac{70}{250} = 0.28$$

$$S.E.(\hat{p}) \text{의 추정값} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.28 \times 0.72}{250}} = 0.028$$

$$95.4\% \text{ 오차한계} = 2 \times 0.028 = 0.056$$

(5) 대표본에서  $p$ 의 신뢰구간

$$n \text{이 클 때, } p \text{의 } 100(1-\alpha)\% \text{ 신뢰구간: } (\hat{p} - z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n})$$

**[예제 7.11]**

[예제 7.2]의 자료에서 어떤 도시의 노동인구 중 500명을 임의추출하였을 때 41명이 실업자였다. 이 도시의 실업률에 대한 95% 신뢰구간을 구하라.

**【풀이】**

$n = 500 (> 30) \rightarrow$  표본비율  $\hat{p}$ 의 정규근사가 이루어진다.

$$1 - \alpha = 0.95, \frac{\alpha}{2} = 0.025, z_{0.025} = 1.96$$

$$\hat{p} \text{의 관측값: } \hat{p} = \frac{41}{500} = 0.082 \rightarrow \hat{q} = 1 - \hat{p} = 1 - 0.082 = 0.918$$

$$z_{0.025} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{0.082 \times 0.918}{500}} = 1.96 \times 0.012 = 0.024$$

이 도시의 실업률에 대한 95% 신뢰구간:  $0.082 \pm 0.024 = (0.058, 0.106)$

이 절차에 의해 많은 신뢰구간을 구하면 그 중 95% 정도가 모수  $p$ 를 포함하므로, 실업률이 5.8%와 10.6% 사이에 있다고 95% 신뢰할 수 있다.

(6) 표본크기의 결정

$$\begin{aligned} z_{\alpha/2} \sqrt{\frac{pq}{n}} &= d \quad \text{여기서 } d : \text{오차한계} \\ n &= pq \left[ \frac{z_{\alpha/2}}{d} \right]^2 \end{aligned}$$

- ① 만약  $p$ 의 값이 대략  $p^*$  근처임이 알려져 있다면  $n$ 은  $n = p^*(1-p^*) \left[ \frac{z_{\alpha/2}}{d} \right]^2$
- ②  $p$ 에 대한 사전정보가 전혀 없을 때는  $pq$  대신 그것이 취할 수 있는 최대값인  $\frac{1}{4}$ 을 사용하여  $n$ 은  $n = \frac{1}{4} \left[ \frac{z_{\alpha/2}}{d} \right]^2$  으로부터 결정된다.

[예제 7.12]

보건복지부에서 약시율을 추정하기 위한 공중보건조사를 실시한다고 한다. 다음의 각 경우에 대하여, 추정오차가 0.05 이내임을 98% 확신하기 위해서는 최소 몇 명에 대하여 검사를 하여야 하는가?

허용오차:  $d = 0.05$ ,  $1 - \alpha = 0.98$ ,  $\frac{\alpha}{2} = 0.01$ ,  $z_{0.01} = 2.33$

(1) 약시율  $p$ 에 대한 아무런 정보가 없는 경우

【풀이】

$$p \text{가 전혀 알려져 있지 않으므로 } \frac{1}{4} \left[ \frac{2.33}{0.05} \right]^2 = 542.89$$

→ 표본의 크기는 543이면 충분

(2)  $p$ 가 약 0.3이라고 알려져 있는 경우

【풀이】

$$p^* = 0.3 \text{이면 필요한 표본의 크기는 } n = (0.3 \times 0.7) \left[ \frac{2.33}{0.05} \right]^2 \approx 456$$

(7) 대표본에서  $p$ 의 가설검정

① 가설 설정:  $H_0: p = p_0$        $H_1: p \neq p_0$

② 검정통계량

시행횟수  $n$ 이 충분히 크면 표본비율  $\hat{p} = \frac{X}{n}$ 은 근사적으로 정규분포를 따른다.

귀무가설 하에서  $p$ 는 주어진 값  $p_0$ 를 가지며  $\hat{p}$ 의 분포는 근사적으로  $N(p_0, p_0q_0/n)$

를 따르고, 표준화된 검정통계량  $Z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}}$ 은 (근사적으로)  $N(0,1)$ 를 따른다.

③ 주어진 유의수준  $\alpha$ 에 대하여 기각역 설정:  $R: |z| \geq z_{\alpha/2}$

④ 결론

**[예제 7.13]**

5년 전의 인구조사에 따르면 어떤 도시의 세대 중 20%가 빈곤층으로 밝혀졌다. 이 비율이 변하였는지를 알아보기 위하여, 400세대를 임의추출하여 조사한 결과 70세대가 빈곤층이었다. 이 표본 자료에 의하면 현재 이 도시의 빈곤층 비율이 5년 전과 달라졌다고 할 수 있는가?

**【풀이】**

① 가설 설정:  $H_0: p = 0.2$        $H_1: p \neq 0.2$

② 검정통계량

$$Z = \frac{(70/400) - 0.2}{\sqrt{0.2 \times 0.8/400}} = -1.25$$

③ 유의수준  $\alpha = 0.05$ 에 대하여 기각역 설정:  $R: |z| \geq z_{\alpha/2} = z_{0.05/2} = 1.96$

④ 결론

검정통계치  $-1.25$ 는 기각역에 포함되지 않으므로 유의수준  $\alpha = 0.05$ 에서 귀무가설을 기각할 수 없다. 따라서 빈곤층의 비율이 달라졌다는 강력한 증거를 찾을 수 없다.

※ 관측된  $z$ 값에 대한 유의확률( $p$ -value)

$$p\text{-값} = P[|Z| \geq 1.25] = P[Z \leq -1.25] + P[Z \geq 1.25] = 2 \times 0.1056 = 0.2112$$

귀무가설을 기각하기 위해서는 유의수준  $\alpha$ 를 0.21보다 크게 해야 한다. 따라서  $H_0$ 를 기각할 수 있는 증거는 매우 약하다.