제 9장. 표집분포 (Sampling distribution)

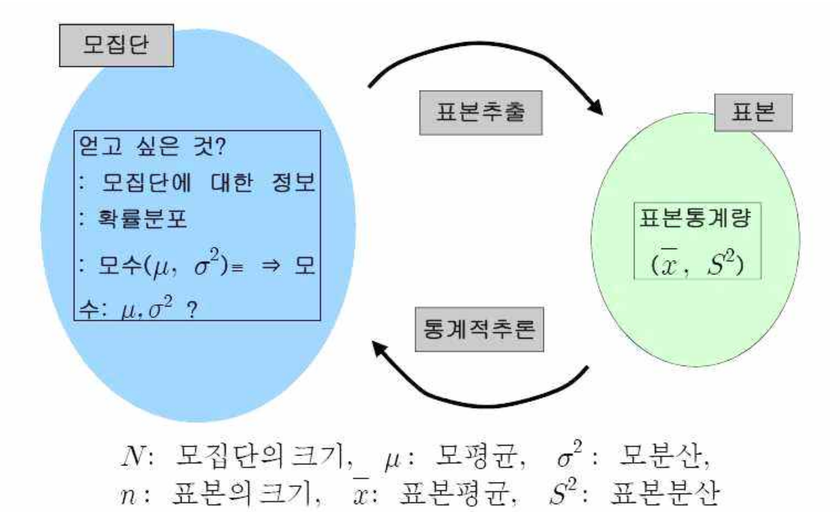
- 주어진 표본으로부터 모집단의 성격을 알아내고자 하는 추론(inference)이 통계학의 가장 핵심부분이다.

모수 : 수치로 표현되는 모집단의 특성 / 확률밀도함수(모집단의 모양) 의 특성을 결정 but 일반적으로 모르는 값

통계량 : 표본의 관측값들에 의하여 결정되는 양
* 통계량의 값은 모수의 참값과는 차이가 있다.
* 통계량의 값은 그 당시 추출된 표본에 영향 받음.
* 통계량의 값은 표본에 따라 값이 변한다.

-notation) N : 모집단의 크기, μ : 모평균, σ^2 : 모분산, n : 표본의 크기, $\overline{x}$: 표본평균, S^2 : 표본분산

-모집단과 표본의 관계



Example) 신도시를 계획하려는 사람이 근로자드르이 집에서 직장까지의 평균 통근거리가 얼마나 되는지 알아보려고 한다고 하자.

-관심의 대상 : 전체 모집단의 평균 통근거리 (모평균 : μ)
-80명의 근로자를 표본추출하여 이들로부터 통근거리를 조사하면
표본평균 : $\overline{x} = 12km$
-또 다른 80명의 근로자를 조사한다면 표본평균값이 달라짐.
→ 한 번의 표본추출로 근로자들의 집에서 직장까지의 평균 통근거리는 $12km$ 라고 추

표집(표본)분포(Sampling distribution) 한 모집단에서 동일한 크기 n 으로 뽑을 수 있는 모든 가능한 표본에서 통계량(statistic) 을 계산할 때, 이 통계량의 확률분포를 의미한다.

→모집단의 분포에 영향을 받고, 표본 크기에도 영향을 받는다.

1. 표본평균의 표집분포

표본평균의 성질

모평균이 μ 이고 모분산이 σ^2 인 모집단에서 크기 n 의 확률표본 $X_1, X_2, \cdots, X_n$을 얻었을 때 표본평균 $\overline{X}$ 의 기댓값과 분산은 다음과 같다.

$$E(\overline{X}) = \mu,\ \ Var(\overline{X}) = \frac{\sigma^2}{n},\ \ sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

→ 표본평균의 평균(기댓값)$E(\overline{X})$=모평균 μ

→ 표본평균의 표준편차(표준오차)에 대하여
- n의 값이 커질수록 표준오차가 작아진다.
- 표준오차가 작다 ⇔ 어떤 표본이 뽑히느냐에 관계없이 표본평균이 크게 변동하지 않는다.
- ⇔ 모평균에 대한 정확성이 높은 추측값

모집단이 정규분포인 경우

정규분포 $N(\mu, \sigma^2)$을 따르는 모집단에서 크기 n 인 표본의 표본평균 $\overline{X}$ 의 분포는 다음과 같다.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

→ 모집단이 정규분포를 따를 때, 표본의 평균은 표본의 개수와 상관없이 항상 정규분포를 따른다.

모집단이 정규분포가 아닌 경우

Theorem 중심극한정리 (Central limit theorem)

평균이 μ 이고 분산이 σ^2 인 임의의 모집단에서 표본의 크기 n이 충분히 크면 $(30 \le n)$,

표본평균 $\overline{X}$ 는 근사적으로 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다. 이를 표준화하여 나타내면 다음과 같다.

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Example) 어느 대학교 입학생 1000명의 평균 수능성적은 315점, 분산은 400점이다. 크기 50의 표본을 추출하였을 때 표본평균의 기댓값과 분산을 구하라.

Example) A대학의 건물 엘레베이터에는 다음과 같은 문구가 적혀있다. “정격 하중 1.350kg”, “정원 20명”. 이 때 20명의 몸무게가 과연 1.350kg을 넘을 확률이 얼마나 되는가? A대학 구성원들의 평균 몸무게는 64kg 이고 표준편차는 10kg인 정규분포를 따른다고 가정하자.

Example) 평균이 82이고 분산이 144인 모집단에 대하여 생각하면

1) 크기 64인 표본이 임의로 추출되었다고 할 때 표본평균이 80.8에서 83.2사이에 있을 확률?

2) 크기 100일 때 표본평균이 80.8에서 83.2사이에 있을 확률?

2. 표본비율의 표집분포

- 성공확률이 p 인 베르누이 시행으로부터 얻은 $X_{1,}X_2,\cdots,X_n$의 합계를 X라 하면
  (즉, 베르누이 시행의 결과는 성공과 실패뿐이고, 성공을 1, 실패를 0 이라 하면)
  표본의 합 $X=\sum_{i=1}^{n}X_i$ 는 성공횟수가 되어 이항분포를 따른다. $X \sim B(n,p)$

표본비율의 표집분포
p : 모비율(population proportion) , $\overline{p}$: 표본비율(sample propotion)
$\hat{p}=\dfrac{\sum_{i=1}^{n}X_i}{n}=\dfrac{X}{n}$

→ 표본비율의 평균 : $\mu_{\hat{p}}=E(\hat{p})=E\left[\dfrac{X}{n}\right]=\dfrac{1}{n}np=p$

→ 표본비율의 분산 : $\sigma_{\hat{p}}^2=Var(\hat{p})=Var\left[\dfrac{X}{n}\right]=\dfrac{1}{n^2}np(1-p)=\dfrac{p(1-p)}{n}$

표본크기가 큰 경우
표본크기 n이 충분히 큰 경우, 표본비율 $\hat{p}$ 는 근사적으로 평균이 p이고 표준편차가 $\sqrt{\dfrac{p(1-p)}{n}}$ 인 정규분포를 따른다.
$\dfrac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx N(0,1)$

Example) 어느 대학의 사회계열 학생들을 대상으로 하는 통계학 입문 강의는 70명씩 10개반으로 나누어 진행되고 있다. 사회계열 학생들은 각자 수강하려는 모든 과목들의 시간표를 고려하여 다른 교양과목이나 전공과목들과 시간이 겹치지 않도록 10개의 반 중 어느 하나를 수강해야 한다. 어느 특정 통계학입문반의 여학생 수가 과반이상이 될 확률을 구하여라. 사회계열 전체 여학생의 비율은 0.54이다.

sol) $p=0.54$, $\hat{p} \sim N\left(0.54,\dfrac{0.54\times0.46}{70}\right)=N(0.54,0.00355)$

$P(\hat{p}\leq 0.5)=P\left(Z\geq\dfrac{0.5-0.54}{\sqrt{0.00355}}\right)=P(Z\geq-0.67)=0.7486$

Example) 어떤 작업을 완료하는데 걸리는 시간은 평균이 30분, 표준편차가 9분인 정규분포를 따른다. 25명의 작업자를 임의로 추출했을 때 평균작업시간이 28분에서 33분 사이일 확률을 구하시오.

Example) 한 교통에 관한 연구에서 승용차 한 대당 탑승객의 수는 평균 1.75명이고 표준편차는 0.65명이었다. 승용차 50대를 추출하여 조사하였을 때 평균 탑승객의 수가 2명 이상일 확률을 구하시오.

Example) 어느 대학교의 총학생회장 선거에서의 투표율은 60%라고 하자. 이 대학교의 투표율을 알아보기 위해서 78명의 학생을 대상으로 조사한 결과 투표율이 0.5와 0.7사이에 있을 확률을 구하시오.

제 10 장. 통계적 추론 (Statistical Inference)

-통계적 추론이란 표본이 갖고 있는 정보를 분석하여 모수에 관한 결론을 유도하고, 모수에 대한 가설의 옳고 그름을 판단하는 것을 말한다.

$$* \text{ 통계적 추론 } = \begin{cases} (1)\text{모수의 추정} \\ (2)\text{모수에 대한 가설검정} \end{cases}$$

1. 모평균에 대한 추정

추정 (Estimation)
표본으로부터 정보를 이용하여 모집단의 성질을 나타내는 수치인 모수를 예측하는 것
즉, N개의 원소로 구성된 모집단에서 n개의 표본을 추출한 후, 이를 토대로 모집단의 모수를 예측하는 과정이다

→1) 모평균의 점 추정 (point estimation) : μ를 하나의 값으로 추정 ($\hat{\mu}=38$)
→2) 모평균의 구간 추정 (interval estimation) : μ를 포함할만한 적당한 구간을 정한다. ($36<\mu<40$)

점추정 (Point Estimation)
추정하고자 하는 하나의 모수에 대하여 n개의 확률변수로 하나의 통계량을 만들고, 주어진 표본으로부터 그 값을 계산하여 하나의 수치를 제시하려고 하는 것이다

→ 추정량(estimator) : 모수를 추정하기 위해 만들어진 통계량
→ 추정치(estimate) : 표본에서 구한 추정량의 값

Example) 해당 도시의 중학교 1학년 남학생의 평균키 추정량 : $\overline{X}=\frac{1}{n}(X_1+\cdots+X_n)$

주어진 표본이 30명이라 하고 $\overline{x}=160.20cm$이라 하면 평균키의 추정치는 $160.53cm$이다.

표준오차(Standard error, $S.E.$) : 추정량의 정확도를 측정하는 도구. (추정량의 표준편차=표준오차**)

Example) 추정량이 표본평균인 경우 표본평균의 표준편차인 $\frac{\sigma}{\sqrt{n}}$ 이 표준오차이다. ($S.E.(\overline{X})=\frac{\sigma}{\sqrt{n}}$)

→$\overline{X}$를 가지고 μ를 추정할 경우 n이 클수록 표준오차가 작아져서 좀 더 정확한 추정이 가능해진다.

→ 추정된 표준오차 : $\frac{s}{\sqrt{n}}$, ($s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$)

(σ를 모르는 경우 σ의 추정량으로 표본표준편차 s를 사용 : $\hat{\sigma}=s$)

모평균에 대한 점추정 요약

모수 : 모집단의 평균 μ

자료 : 평균이 μ이고 표준편차가 σ인 모집단에서 임의추출한 표본 $X_1,\cdots,X_n$

추정량 : 표본평균 $\overline{X}$

표준오차 : $S.E.(\overline{X})=\dfrac{\sigma}{\sqrt{n}}$, 추정된 표준오차 : $\dfrac{s}{\sqrt{n}}$

→임의표본(random sample) : 모집단으로부터 임의추출된 크기가 n인 표본 $X_1,\cdots,X_n$이 서로 독립, 모두 모집단의 분포와 같은 분포를 갖는 것

Example) 어느 전기부품공장에서 자사의 제품 중 전기퓨즈의 평균수명(μ)을 알아보기 위해 40개의 표본을 추출하고 그 수명(x_i)를 조사하여 다음과 같은 결과를 얻었다.

$$\Sigma x_i = 88 \qquad \Sigma x_i^2 = 212$$

평균수명을 추정하고 그 추정량의 표준오차를 구하라.

sol) 평균수명(μ)의 추정량은 표본평균($\overline{X}$)이므로 μ의 추정치 : $\overline{x}=\dfrac{88}{40}=2.2$

모집단의 표준편차가 알려져 있지 않으므로 표본표준편차를 이용하여 표준오차를 추정한다. 우선,

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}=\sqrt{\frac{1}{39}(212-40(2.2)^2)}=0.687$$

따라서 표본평균의 추정된 표준오차는 $\dfrac{s}{\sqrt{n}}=\dfrac{0.687}{\sqrt{40}}=0.109$ 이다.

**점추정량의 선택 기준 (추정량의 성질)

(1) **불편성**(Unbiasedness)

- 어떤 추정량 $\hat{\theta}$ 가 $E(\hat{\theta})=\theta$ 를 만족하면 $\hat{\theta}$ 를 θ의 **불편추정량**(undiased estimator) 이라 한다.
- 추정량의 기댓값이 모수와 같이 않을 때 **편의**(bias) 가 있다고 표현한다.

(2) **효율성**(Efficiency) : 추정량의 표준오차(표준편차)가 작은 것

- 추정량 $\hat{\theta}$의 표준오차를 $S.E.(\hat{\theta})$라 하면, 모수 θ의 불편추정량 $\widehat{\theta_1},\widehat{\theta_2}$ 에 대하여

$S.E.(\widehat{\theta_1})<S.E.(\widehat{\theta_2}) \Leftrightarrow \widehat{\theta_1}$ 은 $\widehat{\theta_2}$ 보다 더 **효율성 있는 추정량** 이라 한다.

(3) **일치성**(Consistency)

- 표본 크기 n이 증가함에 따라 통계량이 점차 모수에 접근하게 될 때 그 통계량을 **일치추정량** 이라 한다.
- 크기가 n인 표본에서 구한 통계량을 $\widehat{\theta_n}$ 이라 할 때, $\lim\limits_{n\to\infty}E(\widehat{\theta_n}-\theta)=0$

→ 편의 : 추정량이 모수에 근접하지 않고 한쪽으로 기울어져 치우친 상태

→ '일치성'이란 표본 크기가 클 수록 추정량이 보다 신뢰성이 있음을 의미

모수(parameter)	추정량(estimator)	성질
μ	$\hat{\mu} = \bar{x}$	불편성, 효율성, 일치성 모두 만족
p	$\hat{p}$	불편성, 효율성, 일치성 모두 만족
σ^2	$\widehat{\sigma^2} = s^2$	불편성(n-1) ※표본분산은 다른 여러 종류의 산포도를 측정하는 단위보다 상대적으로 효율성과 일치성이 있기 때문에 사용!!

예제) {1,2,3,4,5}에서 표본크기가 3인 표본을 비복원으로 추출한다.

$$\mu = \frac{1+2+3+4+5}{5} = 3, \qquad \sigma^2 = \frac{1}{5}\{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2\} = 2$$

(1) 세 통계량 $\overline{X}$, $\widetilde{X}$, $X_{(1)}$ 값 중 불편추정량은 어떤 것인가 그리고 불편추정량 중 효율성이 좋은 것은?

표본	$\overline{X}$	$\widetilde{X}$	$X_{(1)}$
{1,2,3}	2	2	1
{1,2,4}	2.33	2	1
{1,2,5}	2.67	2	1
{1,3,4}	2.67	3	1
{1,3,5}	3	3	1
{1,4,5}	3.33	4	1
{2,3,4}	3	3	2
{2,3,5}	3.33	3	2
{2,4,5}	3.67	4	2
{3,4,5}	4	4	3

$$E(\overline{X}) = \frac{2+2.33+\cdots+4}{10} = 3$$

$S_{\overline{X}}^2 = 0.37038, S_{\overline{X}} = 0.60859$

$$E(\widetilde{X}) = \frac{2+2+\cdots+4}{10} = 3$$

$S_{\widetilde{X}}^2 = 0.6667, \ S_{\widetilde{X}} = 0.81652$

$$E(X_{(1)}) = \frac{1+1+\cdots+3}{10} = 1.5$$

불편추정량 : 표본평균($\overline{X}$)과 중앙값($\widetilde{X}$)

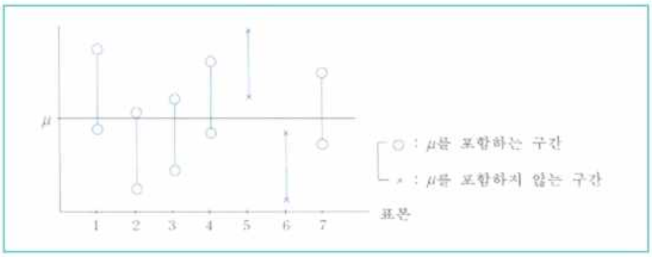
효율성 좋은 것 : 표본평균($\overline{X}$)

=> 최종 추정량 선택: 표본평균($\overline{X}$)

구간추정 (Interval Estimation)

하나의 수치를 구하는 것이 아니라 추정량의 분포를 이용하여 표본으로부터 모수 값을 포함하리라고 예상되는 구간을 구하여 제시하는 것. 이 때 제시되는 구간을 **신뢰구간**(confidence interval) 이라 한다. 그리고 여러 번 구한 신뢰구간 중 추정하고자 하는 모수를 포함하고 있는 신뢰구간의 비율을 **신뢰수준**(confidence level) 이라 한다.

Example) '모평균 μ에 대한 95% 신뢰구간' 은 예를 들면 100차례 표본을 추출하여 신뢰구간을 구하였을 때, 그 100개의 신뢰구간 중 95개가 모수 μ를 포함하리라고 예상한다는 의미이다.



[신뢰구간의 의미]

모평균 μ에 대한 신뢰구간

(1) 표본의 크기(n)가 큰 경우

표본의 크기가 n이고 평균과 표준편차가 $\overline{X}$와 s로 주어질 때 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은

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

단, 모집단의 표준편차(σ)가 알려져있으면 $s\rightarrow\sigma$로 대체한다.

(2) 표본의 크기($n<30$)가 작은 경우 (단, 모집단이 정규분포인 경우)

모집단의 분포가 정규분포이고 표본크기가 작을 때 $\dfrac{\overline{X}-\mu}{s/\sqrt{n}}$은 표준 정규분포가 아닌 t-분포를 따르게 된다.

따라서 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\overline{X}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{X}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \text{ 혹은 } \overline{X}\pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

이 때 표준편차(σ)는 알려져 있지 않은 것으로 간주한다.

→ 신뢰수준은 보통 90%, 95%, 99% 등을 사용한다.

→ $\overline{X}\pm z_{\alpha/2}\frac{s}{\sqrt{n}}$, $\overline{X}\pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ 를 $100(1-\alpha)\%$**오차범위**(error margin) 라고 한다.

→ $z_{\alpha/2}$, $t_{\alpha/2}(n-1)$ 는 각 분포에서 상위 $\alpha/2$의 확률을 주는 값.

신뢰수준	90%	95%	99%
$z_{\alpha/2}$	1.645	1.96	2.575

Note) (1) 표본의 크기가 큰 경우 μ에 대한 $100(1-\alpha)\%$ 신뢰구간의 유도

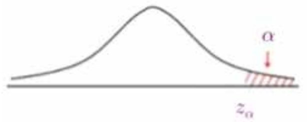
모집단에서 크기가 n인 충분히 큰 표본을 추출한다면, 표본평균 $\overline{X}$의 분포는 중심극한정리(CLT)에 의해

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$\alpha = 1 - P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right)$ 로 두면, $1-\alpha = P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right)$ 이고 이 때 $z_{\alpha/2}$ 는 상위 $\alpha/2$의 확률을 주는 값으로 정한다. 그러면

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

$$= P\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$



이므로, 신뢰구간은 $\left(\overline{X}-z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}+z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$ 이 된다.

(2) 표본 크기가 작고, 모표준편차 σ를 모르는 경우 μ에 대한 $100(1-\alpha)\%$ 신뢰구간의 유도

$t = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$ 를 이용하면

$$1-\alpha == P\left(-t_{\alpha/2}(n-1) \le \frac{\overline{X}-\mu}{s/\sqrt{n}} \le t_{\alpha/2}(n-1)\right)$$

$$= P\left(\overline{X}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \overline{X}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$$

이므로, 신뢰구간은 $\left(\overline{X}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{X}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$ 이 된다.

- **t 분포 (t-distribution)**

정규모집단 $N(\mu,\sigma^2)$으로부터 추출된 표본을 $X_1,\cdots,X_n$이라고 할 때, 표본평균과 표본분산을

$$\overline{X} = \frac{\Sigma X_i}{n}, \quad s^2 = \frac{\Sigma(X_i-\overline{X})^2}{n-1}$$

이라고 정의하면, 표준화된 확률변수

$$t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$$

는 자유도가 $(n-1)$인 t 분포를 따른다고 하고, 이를 기호로써 $t(n-1)$로 표현한다.

→ 자유도가 커질수록 표준정규분포에 가까워진다 ⇔ 표본의 크기가 클 수록 근사적으로 정규분포를 따른다

→ $t_\alpha(r)$: 상위 α의 확률을 주는 값. 이 때 r은 자유도($d.f.$)

Example) 우주선 제작에 쓰기 위해서 새롭게 개발된 합금의 평균 장력 μ를 추정하고자 한다. 15개의 새 합금조각으로부터 장력을 측정한 결과 평균이 39.3, 표준편차가 2.6으로 나타났다.
새롭게 개발된 합금의 평균장력 μ에 대한 90% 신뢰구간을 구하라.

sol) $\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = 39.3 \pm 1.761\frac{2.6}{\sqrt{15}} = 39.3 \pm 1.18$ 혹은 $(38.12, 40.48)$

2. 모비율에 대한 추정

모집단의 크기가 매우 커서 그에 비해 표본의 크기가 작은 경우, 표본의 추출과정은 거의 서로 독립이므로 X의 분포는 모수가 (n,p)인 이항분포를 따른다. → $X \sim B(n,p)$: 평균 np, 표준편차 $\sqrt{np(1-p)}$

이로부터 표본비율인 $\hat{p} = \frac{\overline{X}}{n}$ 의 기댓값과 표준편차를 구하면 $E(\hat{p}) = p$, $S.E.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.

모비율의 점추정

모수 : 모집단에서 A 라는 특성을 갖는 집단의 비율 : 모비율 p

자료 : 크기가 n인 표본에서 A 라는 특성을 갖는 개체의 수 : X

추정량 : 표본비율 $\hat{p} = \frac{\overline{X}}{n}$

표준오차 : $S.E.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$, 추정된 표준오차 : $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

* 표본의 크기가 큰 경우에 $\hat{p}$는 근사적으로 평균이 p이고 분산이 $\frac{p(1-p)}{n}$ 인 정규분포를 따른다.

이를 표준화 하면 $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$ 이다.

Example) 어느 통신판매회사에서는 하나의 새로운 상품을 그 달의 특별상품으로 판매하고자 한다. 그 회사에서 관리하는 만 명의 고객명단에서 임의로 250명을 추출하여 안내책자를 발송한 결과 그 중 70명이 구입을 희망하였다고 한다. 이 표본자료에 근거해서 전체 고객 중 그 상품에 대한 구입을 원하는 모비율(p)을 추정하고, 그 추정량의 표본오차를 구하여라.

sol) 모비율 p에 대한 추정량은 표본비율이므로 p의 추정치 : $\hat{p} = \frac{70}{250} = 0.28$ 이고,

$\hat{p}$의 표준오차를 추정하면, $\hat{p}$의 추정된 표준오차 : $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.28 \times 0.72}{250}} = 0.028$ 이다

모비율의 구간추정

****모비율 p에 대한 $100(1-\alpha)\%$ 신뢰구간**

- 모비율 p는 알려져 있지 않으나 표본 크기 n이 충분히 크면 p 대신에 표본비율 $\hat{p}$를 사용할 수 있다.

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \text{ 혹은 } \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example) 한 도시의 노동인구 500명 중 41명이 실업자로 나타났다고 하였다. 그 도시의 실업률 p에 대한 95% 신뢰구간을 구하여라.

sol) p에 대한 95% 신뢰구간은

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{41}{500} \pm 1.96\sqrt{\frac{0.082 \cdot 0.918}{500}} = 0.082 \pm 0.024 \text{ 혹은 } (0.058, 0.106) \text{ 이다.}$$

3. 표본 크기의 결정

-오차가 d 이하로 될 확률이 최소한 $100(1-\alpha)\%$가 되도록 하기 위해서 필요한 최소의 표본크기

모수	표본크기
모평균	$n \geq \left(z_{\alpha/2}\frac{\sigma}{d}\right)^2$
모비율	$n \geq \left(z_{\alpha/2}\frac{\sigma}{d}\right)^2 p(1-p)$ *사전정보 없을 시 $p=0.5$ 이용

유도 → 모집단이 정규분포이고, 표준편차 σ가 주어져 있을 때

-모평균의 경우 오차가 d 이하로 될 확률이 최소한 $100(1-\alpha)\%$: $P(|\overline{X}-\mu| \leq d) \geq 1-\alpha$

-표준화된 확률변수의 분포가 표준정규분포를 따르므로 $P\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$

$\Rightarrow z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq d$, 이 식을 n에 대하여 풀면 $n \geq \left(z_{\alpha/2}\frac{\sigma}{d}\right)^2$ 이 된다.

※ 모집단이 정규분포라는 가정이 없다면, 표본의 크기가 30이상으로 구해졌을 때에만 사용가능하다.

Example) 호수의 성분을 연구하는 한 학자가 호수성분의 하나인 인산염의 무게 분포에 관심이 있다고 한다. 그 학자가 단위부피당 평균무게(μ)를 추정하고자 하는데, 이제끄지의 자료에 의하면 단위부피당 인산염의 무게의 표준편차 $\sigma=4$를 갖는다고 한다. 평균무게 μ에 대한 오차가 0.75 이하가 될 확률이 최소한 90% 가 되기 위해서는 어느 정도 크기의 표본을 조사하여야 하겠는가?

sol) $z_{0.05}=1.645,\ d=0.75.\ \sigma=4 \Rightarrow n \geq \left(1.645\left(\frac{4}{0.75}\right)\right)^2 = 76.97$, 따라서 n은 최소한 77이상이어야 한다.

Example) 공중보건에 관한 조사에서 시력장애자의 비율을 추정하고자 한다. 만약 이 조사에서 얻어지는 p에 대한 추정량의 오차가 0.05 이하가 될 확률이 최소한 98%가 되기를 원한다면 몇명을 대상으로 시력장애 여부를 조사하여야겠는가?

sol) $n \geq 0.5(1-0.5)\left(\frac{2.33}{0.05}\right)^2 = 542.9$, 따라서 최소한 543이어야 한다.

4. 모표준편차에 대한 추정

- 모집단이 정규분포를 따른다고 가정한다.

  점추정의 경우 : σ^2의 추정량으로 s^2 사용하고, 모표준편차의 추정량으로 표본표준편차 s를 사용

  구간추정의 경우 : s^2의 분포가 필요! → χ^2(카이제곱) 분포와 연관

모표준편차의 구간추정

- **χ^2 분포 (χ^2 distribution)**

정규모집단 $N(\mu,\sigma^2)$으로부터 추출된 표본을 $X_1,\cdots,X_n$이라고 할 때

$$\chi^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

은 자유도가 $(n-1)$인 χ^2 분포를 따른다고 하고, 이를 기호로써 $\chi^2(n-1)$로 표현한다.

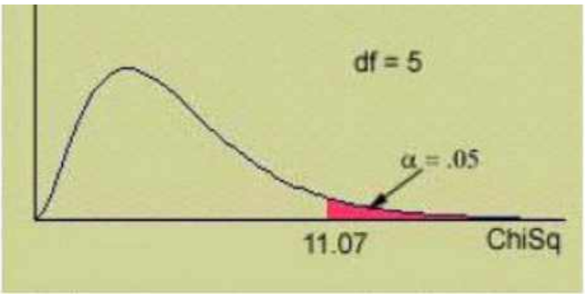
→이 분포는 정규분포나 t분포와 달리 확률밀도함수가 양수쪽에만 퍼져있고 오른쪽에 긴 꼬리를 갖는 비대칭형이다.

→자유도가 클수록 0으로부터 멀리 떨어져서 넓게 분포한다.

→$\chi^2_\alpha(r)$: 자유도가 r인 확률변수 χ^2에 대하여 상,하위 α의 확률을 주는 값

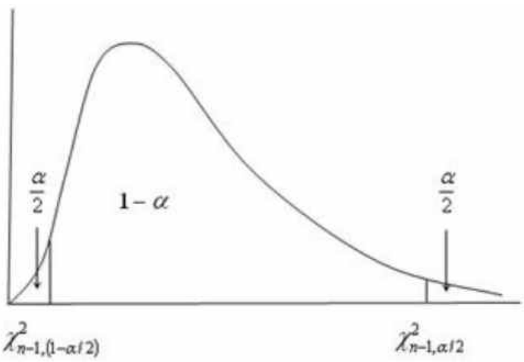
※ χ^2 분포의 확률 계산하기

자유도$(n-1)$와 α값에 따라 χ^2 분포표를 이용하여 상위 α의 확률을 구한다.



$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$	df
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	14.4494	16.8119	18.5476	6
12.0170	14.0671	16.0128	18.4753	20.2777	7

Example) χ^2 분포표에서 자유도가 17인 χ^2 분포의 상, 하위 5%의 확률을 주는 값을 찾아라.



sol) 상위 5%의 확률을 주는 값은 $\chi^2_{0.05}(17) = 27.59$이고, 하위 5%의 확률을 주는 값은 $\chi^2_{0.05}(17) = 8.67$이다.

****모표준편차 σ의 $100(1-\alpha)\%$ 신뢰구간**

$$\left(s\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}\,,\ s\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}} \right)$$

Note) 표본분산 s^2을 이용한 σ^2의 신뢰구간을 우선 구하자.

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ 임을 이용하면 $P(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)) = 1-\alpha$

$\Rightarrow P(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}) = 1-\alpha$

$\Rightarrow \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \right)$: σ^2의 신뢰구간

$\Rightarrow \left(s\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}\,,\ s\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}} \right)$: σ의 신뢰구간

Example) 볼트와 너트를 생산하는 한 공장에서는 제품의 품질이 얼마나 균일하게 유지되는지를 검사하려고 10개의 볼트를 추출하여 지름을 측정하고 그 표준편차를 구하였더니 0.4였다. 그 공장에서 생산되는 볼트의 지름이 정규분포를 따른다는 가정 하에 σ의 90% 신뢰구간을 구하여라.

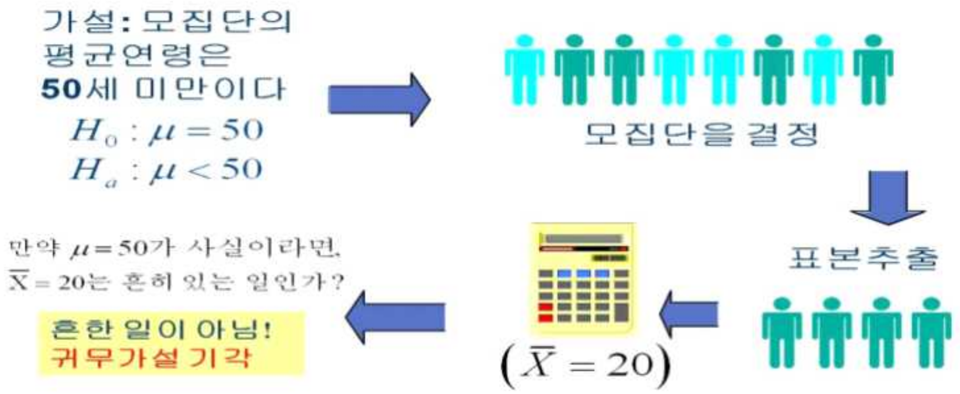
sol) $n = 10$, $\chi^2_{0.05}(9) = 16.92$, $\chi^2_{0.95}(9) = 3.33$

σ^2의 90% 신뢰구간 : $\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \right) = \left(\frac{9\times(0.4)^2}{16.92}, \frac{9\times(0.4)^2}{3.33} \right) = (0.085, 0.432)$

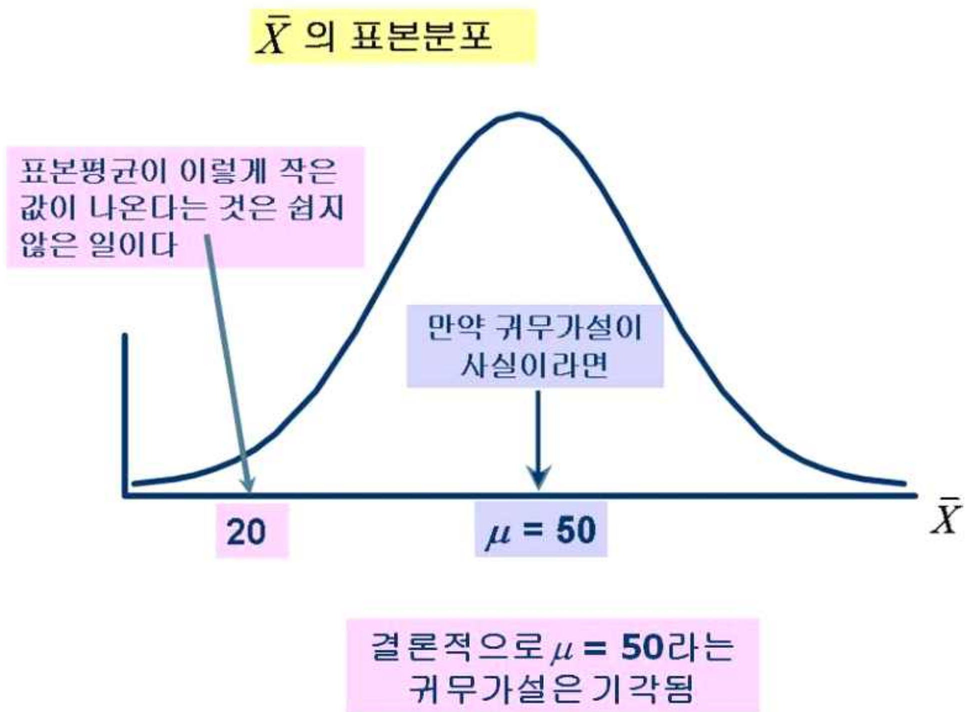
$\therefore \sigma$의 신뢰구간은 $(\sqrt{0.085}\,, \sqrt{0.432}\,) = (0.29, 0.66)$ 이다.

5. 가설 검정

가설검정 (Testing of Hypothesis)
모집단의 모수나 분포에 대하여 가설을 설정하고, 이 가설의 옳고 그름을 표본의 정보를 이용하여 검정하는 방법



[가설검정 과정 및 원리]



통계적 가설 (Statistical Hypothesis)
모집단의 특성 또는 모수에 대한 대립되는 두 가지 주장에 대하여 통계적으로 다루기에 편리하도록 정리한 것

- 가설의 종류

(1) **대립가설** (H_1 : alternative hypothesis)
  - 표본에서 나타난 정보를 통해서 밝히고자 하는 가설 → "입증하려는 가설"

(2) **귀무가설** (H_0 : null hypothesis)
  - 대립가설에 반대되는 가설
    → "대립가설을 입증할 수 없을 때 대립가설을 무효화 시키면서 받아들이는 가설"

- 가설검정 규칙

귀무가설		대립가설	
$H_0 : \theta = \theta_0$	$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$	양측검정(two-sided test)
$H_0 : \theta \le \theta_0$ $H_0 : \theta \ge \theta_0$		$H_1 : \theta > \theta_0$ $H_1 : \theta < \theta_0$	단측검정(one-sided test)

→ 귀무가설은 모수를 특정한 값으로 표현한다. $H_0 : \theta = \theta_0$ (θ_0 : 모수의 특정한 값)
→ 대립가설은 귀무가설에서 지적한 모수의 값이 아닌 어떤 영역으로 나타낸다.

Example) 어느 전구의 제조공정에서는 평균수명이 1500시간이고 표준편차가 100이 되도록 품질관리를 하고 있다. 품질을 개선하기 위하여 어느 팀에서 새로운 공법을 개발하였으며 새 공법에 의하면 평균수명이 증가한다고 주장한다. (다만, 새 공법의 특성은 기존의 공법과 같아서 표준편차가 100으로 가정할 수 있으며, 생산비용도 같다고 가정한다.)

- **검정통계량** (test statistic)
- 검정에 사용하기 위하여 표본자료에서 구한 통계량
- 일반적으로 검정하려는 모수의 점추정량 또는 점추정량을 표준화한 것을 사용

검정하려는 모수		표준화된 검정통계량
μ (정규분포 모집단의 평균)	σ를 알 때	$Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$
	σ를 모를 때	$T = \dfrac{\overline{X} - \mu_0}{s / \sqrt{n}} \sim t(n-1)$
p (모비율)	$np \ge 5, n(1-p) \ge 5$	$Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$

- 검정의 오류 (test error)

\- 가설을 채택하거나 기각할 때 확률적으로 틀릴 가능성

<오류의 종류>

실제상황 / 검정결과	H_0가 사실	H_1이 사실
H_0를 기각못함	옳은결정 : $1-\alpha$	**제 2종 오류** (type II error) : β
H_0를 기각 (H_1채택)	**제 1종 오류** (type I error) : α	옳은 결정 : $1-\beta$

→ 제 1종 오류 : 귀무가설이 맞을 때 귀무가설을 기각하는 오류

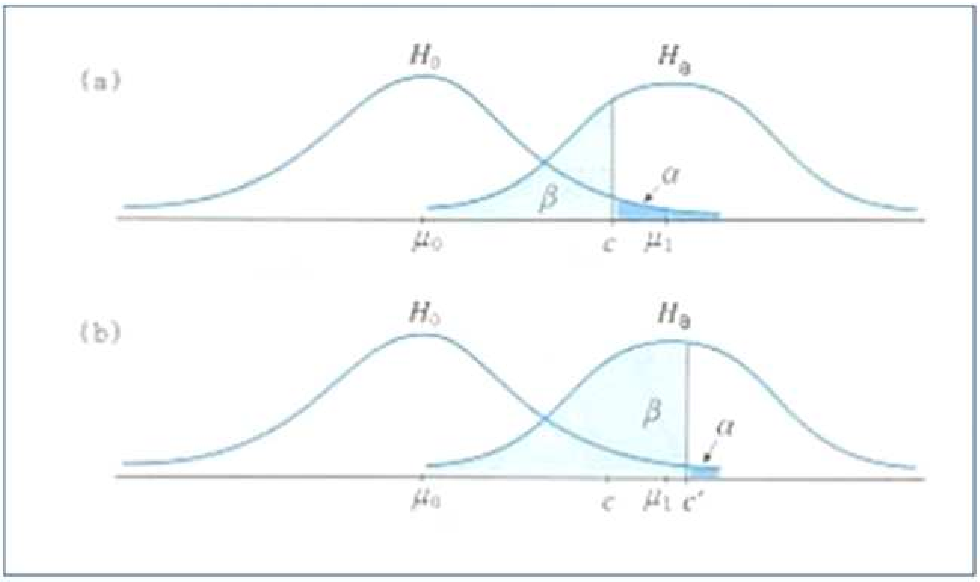
→ 제 2종 오류 : 귀무가설이 틀리는데 귀무가설을 기각하지 않는 오류

→ 두 개의 가설 H_1과 H_0을 쓸 때 H_0 대 H_1 으로 쓴다.

→ $\alpha = P(\text{제1종 오류}) = P(H_0\text{기각} \mid H_0\text{사실})$
$\beta = P(\text{제2종 오류}) = P(H_0\text{채택} \mid H_1\text{사실})$

※ 검정력(power, $1-\beta$) : $1-\beta = P(H_0\text{기각} \mid H_1\text{사실})$

Example) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ 의 가설에서 H_1의 대표값이 μ_1일 때, 임계값 c에 의해서 α, β를 결정



→ *이상적인 목표는 α, β를 동시에 작게하는 것이지만, α가 작아지면 β가 커지는 모순이 발생*

→ *α를 정하고 그 중 β를 최소로 하는 가설검정 시행*

- **유의수준 (significance level : α) 결정** : <u>제 1종 오류의 최대값</u>

- $\alpha = \max P(\text{제1종 오류}) = \max P(H_0\text{기각} \mid H_0\text{사실})$
- 귀무가설이 참인데도 불구하고 이를 기각하는 확률로 위험률이라고도 한다.
- 자료와 관련된 학문 분야에서 이미 사용되고 있는 유의수준 또는 논문 등을 참고하여 유의수준을 결정한다.

- **기각역 (rejection region / critical region)**

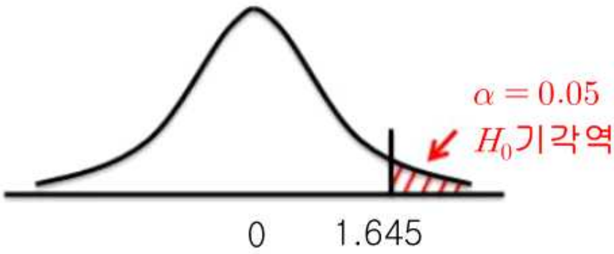
- 귀무가설을 기각할 수 있는 검정통계량의 영역으로 대립가설의 형태에 따라 기각역이 달라진다.

→ <u>기각역에 의한 판정 : 유의수준에 따라 기각역을 정하고, 검정통계량이 기각역에 속하면 귀무가설을 기각</u>

Example) $H_1 : \mu > \mu_0$ $(\alpha = 0.05)$

검정통계량 : $\overline{X} \Rightarrow R : \overline{X} \ge c,\ c = \mu_0 + 1.645\frac{s}{\sqrt{n}}$

검정통계량 : $\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \Rightarrow R : \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \ge 1.645$



- **유의확률 (significance probability / p-value)**

- *<u>귀무가설이 사실일 때</u>* 표본이 대립가설 방향으로 검정통계량의 값보다 더 어긋나게 될 확률
- 주어진 검정통계량의 값으로부터 H_0를 기각하게 하는 최소의 유의수준

⇒ p값이 작은 경우 : 귀무가설이 참일 때 나오기 힘든 경우가 나온 것.
i.e. p값이 작을수록, 귀무가설을 기각하는 쪽으로 신뢰

$p-value \le \alpha$: 유의수준 α에서 H_0 기각 (통계적으로 유의하다)

$p-value > \alpha$: 유의수준 α에서 H_0 기각 못함 (통계적으로 유의하지 않다)

※ *가설검정 절차*

① 귀무가설과 대립가설을 설정한다.

② 주어진 문제의 특성에 따라 유의수준 α를 결정한다.

③ 표본 자료에서 검정통계량을 계산한다.

④ 가설검정결론

(i) 기각역 사용
- 유의수준에 따라 기각역을 구한다.
- 검정통계량이 기각역에 속하면 H_0을 기각한다.

(ii) p값(유의확률) 사용
- 검정통계량을 이용하여 p값을 구한다.
- $p \le \alpha$ 이면 H_0를 기각한다.

5-1. 모평균의 가설검정

① 가설

귀무가설	대립가설	
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	양측검정(two-sided test)
	$H_1 : \mu > \mu_0$	단측검정(one-sided test)
	$H_1 : \mu < \mu_0$	단측검정(one-sided test)

② 유의수준 α 정하기 : 0.01, 0.05

③ 검정통계량 : 귀무가설이 $H_0 : \mu = \mu_0$ 인 경우 검정통계량은

σ를 알 때	σ를 모를 때, n이 클 때 $(n \geq 30)$	σ를 모를 때, n이 작을 때 ※ 모집단 정규분포 가정
$Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$	$Z = \dfrac{\overline{X} - \mu_0}{s / \sqrt{n}} \sim N(0,1)$	$T = \dfrac{\overline{X} - \mu_0}{s / \sqrt{n}} \sim t(n-1)$

④ 기각역 & 유의확률

			대립가설		
			$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$
					
검정통계량	$Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$ 또는 $Z = \dfrac{\overline{X} - \mu_0}{s / \sqrt{n}}$	기각역	$R = \{Z \geq z_\alpha\}$	$R = \{Z \leq -z_\alpha\}$	$R = \begin{Bmatrix} Z \leq -z_{\alpha/2}, \\ Z \geq z_{\alpha/2} \end{Bmatrix}$
		유의확률 (p-값)	$P(Z \geq z)$	$P(Z \leq z)$	$z > 0 : 2 \times P(Z \geq z)$ $z < 0 : 2 \times P(Z \leq z)$
	$T = \dfrac{\overline{X} - \mu_0}{s / \sqrt{n}}$	기각역	$R = \{T \geq t_\alpha(n-1)\}$	$R = \{T \leq -t_\alpha(n-1)\}$	$R = \begin{Bmatrix} T \leq -t_{\alpha/2}(n-1), \\ T \geq t_{\alpha/2}(n-1) \end{Bmatrix}$
		유의확률 (p-값)	$P(T \geq t)$	$P(T \leq t)$	$t > 0 : 2 \times P(T \geq t)$ $t < 0 : 2 \times P(T \leq t)$

Example) 어느 다이어트 방법을 소개하는 책자에서 주장하기를 그 다이어트 방법을 이용하면 5주 동안 10kg 넘게 체중을 줄일 수 있다고 한다. 그 다이어트 방법을 이용하여 56명을 대상으로 5주 동안의 체중 감소량을 조사하였더니 평균이 10.5kg, 표준편차가 4.5kg이었다고 한다. 이 자료에 근거해서 그 책자의 주장이 옳다고 할 수 있는가? 유의수준 5%에서 검정하시오.

sol) $n=56,\ \overline{x}=10.5,\ s=4.5$

① 가설 : $H_0 : \mu=10,\ H_1 : \mu>10$

② 검정통계량 : $z=\dfrac{10.5-10}{4.5/\sqrt{56}} \fallingdotseq 0.83$

③ 기각역 : $R : \{Z \geq z_{0.05}=1.645\}$

유의확률(p-값) : $P(Z \geq 0.83)=1-P(Z<0.83)=1-0.7967=0.2033$

④ 결론 : (i) 기각역 기준 : 검정통계량이 $z=0.83$으로 임계치 1.645보다 작으므로 H_0 기각역에 속하지 않는다. 그러므로 H_0를 기각할 수 없다. 즉, 유의수준 5%에서 책자의 주장이 옳다고 할 수 없다.

(ii) 유의확률 기준 : 유의확률이 0.2033으로 유의수준 0.05보다 크므로 H_0를 기각할 수 없다. 즉, 유의수준 5%에서 책자의 주장이 옳다고 할 수 없다.

Example) 한국 청소년들의 하루 평균 TV시청 시간에 대하여 가설검정을 한다. 한편은 시청 시간이 하루 3시간이라고 주장하고 다른 한편은 “그렇지 않다”라고 주장한다. 그래서 100명의 청소년을 대상으로 시청 시간을 조사해 본 결과 아래와 같은 결과가 나왔다. 유의수준 5%로 검정하시오.

sol) $\overline{x}=2.75$시간이고, $\sigma=1$

① 가설 : $H_0 : \mu=3,\ H_1 : \mu \neq 3$

② 검정통계량 : $z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{2.75-3}{1/\sqrt{100}}=-2.5$

③ 기각역 $R : \{|z| \geq z_{0.025}=1.96\} \Rightarrow R : \{z<-1.96,\ z>1.96\}$

유의확률(p-값) : $P\{|Z|>2.5\}=P(z<-2.5)+P(z>2.5)=0.0062\times 2=0.0124$

④ 결론

- $z=-2.5$는 기각역 -1.96 보다 작으므로 기각역에 속한다. 따라서 유의수준 5%에서 귀무가설을 기각할 수 있다. 즉, 이 자료에 의하면 유의수준 5%에서 청소년들의 하루 평균 TV시청이 3시간이 아니라고 할 수 있다.
- p값이 0.0124로 유의수준 $\alpha=0.05$보다 작으므로 귀무가설을 기각할 수 있다. 즉, 이 자료에 의하면 유의수준 5%에서 청소년들의 하루 평균 TV시청이 3시간이 아니라고 할 수 있다.

Example) 어느 도시의 보건복지과에서는 그 도시의 상수원인 어느 호수의 수질에 관심이 있다고 한다. 수질을 나타내는 하나의 수치로 단위부피당 평균 세균 수가 있는데, 그 수가 200이상이면 상수원으로 적합하지 않다고 한다. 호수의 열 군데에서 물을 떠서 조사한 결과 단위부피당 세균수가 다음과 같이 나타났다

175 190 215 198 184 207 210 193 196 180

이 자료료부터 호수의 단위부피당 평균 세균 수(μ)가 200보다 적다고 주장할 수 있겠는가?
(단, 호수의 단위부피당 세균 수는 정규분포를 따른다고 가정하자)

sol) $n=10,\ \overline{x}=194.8$

$$s^2=\frac{\Sigma x^2-\frac{(\Sigma x)^2}{n}}{n-1}=\frac{381024-\frac{1948^2}{10}}{9}=172.62,\ s=\sqrt{172.62}\fallingdotseq 13.14$$

① 가설 : $H_0:\mu=200.\ H_1:\mu<200$

② 검정통계량 : $t=\dfrac{\overline{X}-\mu_0}{s/\sqrt{n}}=\dfrac{194.8-200}{13.14/\sqrt{10}}=-1.25$

③ 기각역 : $R:\{T\le\ -t_{0.05}(9)=-1.833\}$

④ 결론 : $t=-1.25>-t_{0.05}(9)=-1.833$ 이므로 H_0을 기각할 수 없다. 즉, 호수의 단위 부피당 평균 세균 수(μ)가 200보다 적다고 할 만한 충분한 근거가 없다.

5-2. 모비율의 가설검정 (대표본인 경우)

① 가설

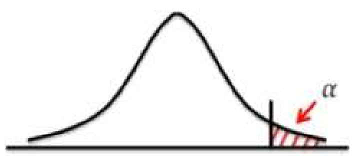
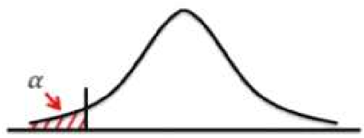
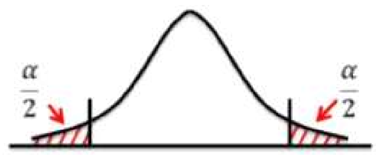
귀무가설		대립가설
	$H_1 : p \neq p_0$	양측검정(two-sided test)
$H_0 : p = p_0$	$H_1 : p > p_0$	단측검정(one-sided test)
	$H_1 : p < p_0$	단측검정(one-sided test)

② 유의수준 α 정하기 : 0.01, 0.05

③ 검정통계량 : 귀무가설이 $H_0 : p = p_0$ 인 경우 검정통계량

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

④ 기각역 & 유의확률

	대립가설		
	$H_1 : p > p_0$	$H_1 : p < p_0$	$H_1 : p \neq p_0$
			
기각역	$R = \{Z \geq z_\alpha\}$	$R = \{Z \leq -z_\alpha\}$	$R = \begin{Bmatrix} Z \leq -z_{\alpha/2}, \\ Z \geq z_{\alpha/2} \end{Bmatrix}$
유의확률 (p-값)	$P(Z \geq z)$	$P(Z \leq z)$	$z > 0 :\ 2 \times P(Z \geq z)$ $z < 0 :\ 2 \times P(Z \leq z)$

Example) 어떤 특정 암의 경우에 수술을 시행한 후 완치되는 비율(5년 이상 생존비율)이 30%라고 한다. 이 암에 걸린 60명의 환자를 대상으로 수술 뿐 아니라 수술 전후에 일정기간 방사선치료를 병행하였더니 60명 중 27명이 완치되었다고 한다. 이 자료로부터 수술만 하는 것보다 방사선 치료를 병행하는 것이 암의 완치율(p)을 높이는데 효과가 있다고 할 수 있는지 검정하라.

sol) ① 가설 : $H_0 : p = 0.3,\ H_1 : p > 0.3$

② 검정통계량 : $z = \dfrac{\hat{p} - p_0}{aqrt\dfrac{p_0(1-p_0)}{n}} = \dfrac{0.45 - 0.3}{\sqrt{\dfrac{0.3 \times 0.7}{60}}} = 2.546 \ \left(\hat{p} = \dfrac{27}{60} = 0.45\right)$

③ 기각역 : $R : \{Z > z_{0.05} = 1.645\}$

$p-$값 : $P(Z \geq 2.54) = 0.0055$

④ 결론 : 귀무가설을 기각할 수 있다. 즉, 자료로부터 방사선치료의 병행이 암치료에 있어서 완치율을 높이는데 효과가 있다고 할 수 있다.

5-3. 모표준편차에 대한 가설검정 (모집단이 정규분포인 경우)

① 가설

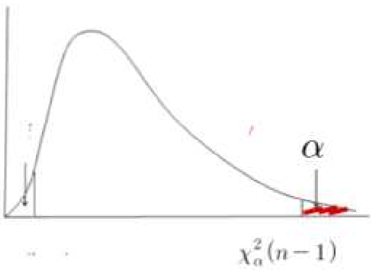
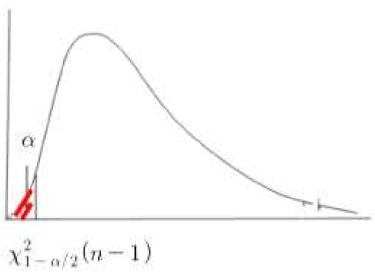
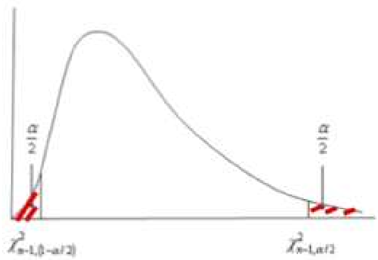
귀무가설		대립가설
	$H_1 : \sigma \neq \sigma_0$	양측검정(two-sided test)
$H_0 : \sigma = \sigma_0$	$H_1 : \sigma > \sigma_0$	단측검정(one-sided test)
	$H_1 : \sigma < \sigma_0$	단측검정(one-sided test)

② 유의수준 α 정하기 : 0.01, 0.05

③ 검정통계량 : 귀무가설이 $H_0 : \sigma = \sigma_0$ 인 경우 검정통계량

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(0,1)$$

④ 기각역 & 유의확률

	대립가설		
	$H_1 : \sigma > \sigma_0$	$H_1 : \sigma < \sigma_0$	$H_1 : \sigma \neq \sigma_0$
			
기각역	$R : \{\chi^2 \geq \chi^2_\alpha(n-1)\}$	$R : \{\chi^2 \leq \chi^2_{1-\alpha}(n-1)$	$R : \left\{\begin{matrix} \chi^2 \geq \chi^2_{\alpha/2}(n-1), \\ \chi^2 \leq \chi^2_{1-\alpha/2}(n-1) \end{matrix}\right\}$

Example) 볼트와 너트를 생산하는 한 공장에서는 제품의 품질이 얼마나 균일하게 유지되는지를 검사하려고 10개의 볼트를 추출하여 지름을 측정하고 그 표준편차를 구하였더니 0.4였다. σ가 0.2 보다 크다고 할 수 있는지 유의수준 0.05로 검정하라.

sol) ① 가설 : $H_0 : \sigma = 0.2$ vs $H_1 : \sigma > 0.2$

② 검정통계량 : $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \times 0.4^2}{0.2^2} = 36$

③ 기각역 : $R : \chi^2 \geq \chi^2_{0.05}(9) = 16.92$

④ 결론 : 검정통계량의 값이 기각역에 포함되므로 주어진 자료로부터 σ가 0.2보다 크다고 할 수 있다.

5-4. 신뢰구간과 양측검정의 관계

모수 θ에 대한 $100(1-\alpha)\%$의 신뢰구간이 (L, U)로 구해졌을 때,
가설 $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ 에 대하여 유의수준 α로 검정을 시행할 때의 결론은 다음과 같다.

$$\theta_0 \in (L, U) \Leftrightarrow H_0\text{를 기각할 수 없다.}$$
$$\theta_0 \notin (L, U) \Leftrightarrow H_0\text{를 기각할 수 있다.}$$

※ 신뢰구간은 한번 구하는 것으로 동시에 여러 귀무가설에 대한 <u>양측 검정</u> 결론을 얻을 수 있다.
$\Rightarrow$ 신뢰구간이 좀 더 포괄적 추론의 과정

Example) 정규모집단으로부터 크기가 9인 표본을 추출하여 평균과 표준편차를 계산하였더니 각각 $\bar{x} = 8.3$과 $s = 1.2$였다고 한다. 모평균 μ에 대한 95% 신뢰구간을 구하고, 유의수준 5%로 가설 $H_0 : \mu = 8.5$ vs $H_1 : \mu \neq 8.5$를 검정하라.

sol) • μ에 대한 95% 신뢰구간

$$\overline{X} \pm t_{0.025,8}\frac{s}{\sqrt{n}} = 8.3 \pm 2.306\frac{1.2}{\sqrt{9}} = 8.3 \pm 0.9224,\ \text{또는 } (7.38, 9.22)$$

• 가설검정

가설 : $H_0 : \mu = 8.5$ vs $H_1 : \mu \neq 8.5$

검정통계량 : $t = \dfrac{8.3 - 8.5}{1.2/\sqrt{9}} = -0.5$, 기각역 : $|t| > 2.306$

제 12 장. 두 모집단의 비교

Example) -사무직과 생산직의 임금의 비교
-수도권과 지방 고등학생들 간의 학력 비교
-우리나라와 다른 나라의 물가 비교
-두 종류의 강의방법에 의한 학습효과 비교
-두 치료제의 효능 비교
-두 종자의 생산량 비교
-두 생산 공장의 불량률 비교
-새로운 치료제의 효능

Example) -두 치료제의 효능 비교 :
10명의 환자에게 서로 다른 두 치료제를 투여함으로써 A와 B 두 치료제의 효능을 비교하려고 한다. 먼저 5명의 환자를 무작위로 추출하여 A 치료제를 투여하고, 나머지 5명의 환자에게는 B 치료제를 투여하여 치료효과를 측정한다.

-새로운 고혈압 치료제의 효능 :
고혈압을 가지고 있는 10명의 환자에게 새로운 치료제를 투여함으로써 약의 효능을 비교하려고 한다. 약 복용 전의 혈압, 일주일간 복용 후의 혈압을 재어 차이가 있는지 비교한다.

※ 두 개의 독립표본

서로 독립인 두 개의 모집단에서 각각 랜덤표본을 추출한 후 자료를 얻어 비교하는 방법

확률표본 $X_1, \cdots, X_n$: $N(\mu_1, \sigma_1^2)$ 인 모집단으로부터 추출
확률표본 $Y_1, \cdots, Y_m$: $N(\mu_2, \sigma_2^2)$ 인 모집단으로부터 추출

$\Rightarrow$ $X_1, \cdots, X_n$ 와 $Y_1, \cdots, Y_m$는 서로 독립인 확률표본이다.

※ 두 모집단의 평균차이에 대한 추정

(1) $\mu_1 - \mu_2$ 의 점추정량 : $\widehat{\mu_1 - \mu_2} = \overline{X} - \overline{Y}$

(2) $\mu_1 - \mu_2$ 의 구간추정 :

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \Rightarrow \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

(1) $\mu_1 - \mu_2$ 의 점추정량 : $\widehat{\mu_1 - \mu_2} = \overline{X} - \overline{Y}$

(2) $\mu_1 - \mu_2$ 의 구간추정 :

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \Rightarrow \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$