# 8. 모평균 (모 비율) 추정 및 검정

표본 자료를 정리하는 목적은 표본 자료에 대한 정보를 얻는 것이 1 차 목적이지만 그보다 더 중요한 것은 그 정보를 이용하여 알기 원하는 모집단의 값을 추정하거나 모수에 대한 가설의 진위 여부를 검정하는 것이다.

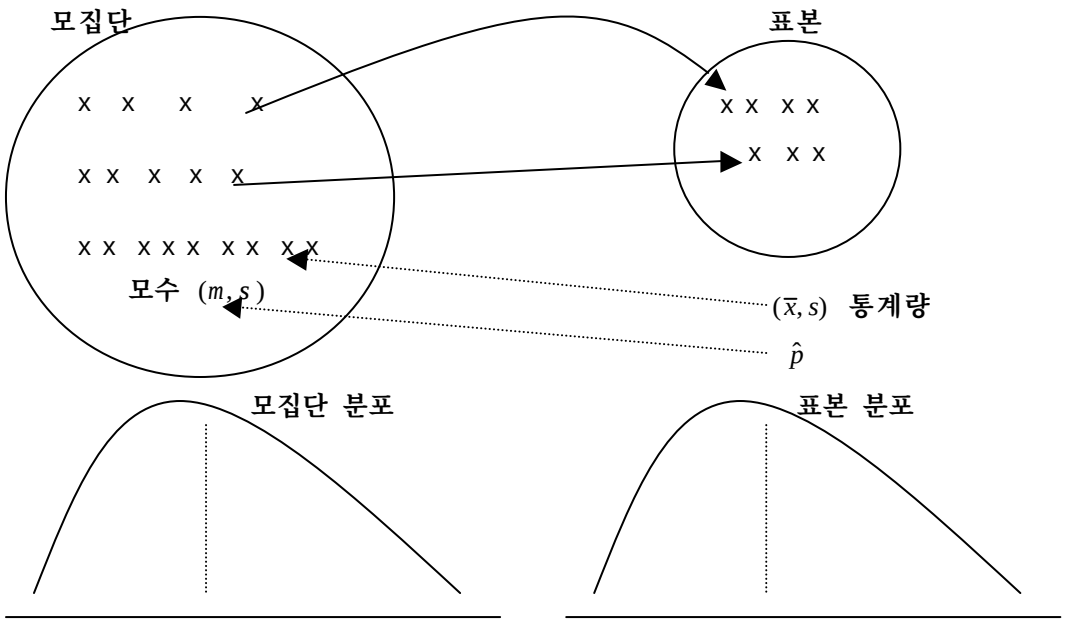
관심의 대상이 되는 모집단의 숫자 요약을 모수(parameter)라 한다. 모집단의 자료가 측정형인 경우 모수는 모집단의 평균, 중앙값, 혹은 표준 편차(분산) 등이 있고 분류형 변수의 경우 모집단의 비율이 있다. 분류형 변수의 경우 표준 편차에 (분산) 관심을 가지기도 하지만 관리 공정(품질 공학)을 제외하고는 거의 관심의 대상이 되지 않는다. 그러므로 여기서는 모집단 평균과 비율에 대한 추정 검정 방법을 살펴보기로 한다.

모수는 모집단 자료의 특성을 숫자로 나타낸 것으로 모집단 평균, 표준 편차, 중앙값, 비율 등이 있다. 이 모수에 대한 정보를 얻기 위하여 표본으로부터 계산된 적절한 숫자 요약 값이 통계량이다. 모수에 대한 정보를 얻는 과정을 통계적 추론(statistical inference)이라 하고 추론에는 추정(estimation)과 검정(hypothesis testing)이 있다.

## 8.1. 추정과 검정

### 8.1.1. 모수와 통계량

모집단 자료의 요약치를 모수(parameter)라 한다. 그 모수에 적합한 표본의 요약 값을 통계량(statistic)이라 한다.

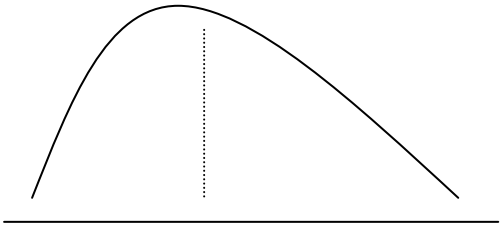


‖ **Example** ‖ 한남대학교 재학생의 IQ에 대해 관심이 있다고 하자. 이를 알아보기 위하여 한남대 재학생 중 50명을 추출하여 IQ를 측정하여 그 표본의 평균과 표준 편차를 계산하였더니 115와 10이였다면 (115, 10)이 통계량이 된다.

‖ **Example** ‖ 한남대학교 학생들의 하숙생 비율이 얼마인지에 관심을 갖는다면 모수는 한남대 재학생의 하숙 비율이 된다. 이를 알아보기 위하여 100명 학생 추출하여 하숙여부를 조사하였더니 35명이 하숙을 하고 있었다면 0.35가 통계량이 된다.

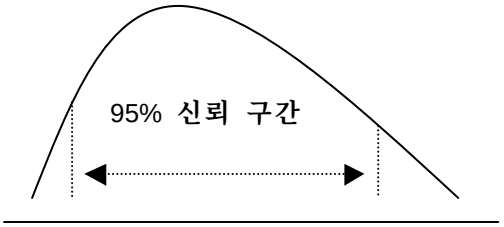
### 8.1.2. 추정

점 추정은 (point estimation) 모집단의 모수를 하나의 값으로 추정하는 방법이다. 어떤 값으로 추정하는 것이 좋은가? 통계학에서는 좋은 추정치를 Best Linear Unbiased Estimator 라 하는데 모집단의 평균은 표본의 평균으로 모집단의 비율은 표본 비율로 추정한다.



점 추정치

구간 추정은 (interval estimation) 모수 값을 구간으로 추정하는 것으로 아래 그림과 같다. 이처럼 구간을 설정하기 위해서는 분포에 대한 정보가 필요하다.




### 8.1.3. 통계적 가설 (statistical hypothesis)

모수에 대해 알고자 하는 내용을 가설로 설정하여 표본으로부터 구한 통계량을 이용하여 설정한 가설의 진위 여부를 검정하는 것을 가설 검정이라 (hypothesis testing) 한다. 통계학에서 설정되는 통계적 가설은 모수에 대한 하나의 값으로 설정된 귀무가설과 (null hypothesis)과 그에 대립하는 대립가설이 (alternative hypothesis) 있다. 우리의 관심은 대부분 대립 가설에 있으므로 이를 연구 가설이라 (research hypothesis) 한다.

‖ **Example** ‖ 한남대학교 대학생의 IQ가 120 이상인가를 알고 알아보고자 한다면

귀무가설: $m = 120$ $m =$ 한남대 학생의 평균 IQ

대립가설: $m > 120$ 단측 검정 (one-sided hypothesis)

‖ **Example** ‖ 한남대학교 대학생의 IQ가 120인가를 알고 알아보고자 한다면

귀무가설: $m = 120$ $m =$ 한남대 학생의 평균 IQ

대립가설: $m \neq 120$ 양측 검정 (two-sided hypothesis)

통계적 가설 검정은 표본으로부터 구한 통계량을 이용하여 귀무가설을 채택하거나 기각한다. 귀무가설이 기각되면 대립가설이 채택한다. 통계적 가설 검정에서는 다음과 같이 2 가지 오류(error)가 발생한다.

| 실제 / 검정 결과 | 귀무가설 참 | 귀무가설 거짓 |
|---|---|---|
| 귀무가설 기각 | 1 종 오류($a$)<br>type I error | 옳은 판단 |
| 귀무가설 채택 | 옳은 판단 | 2 종 오류($b$)<br>type II error |

1 종 오류와 2 종 오류를 동시에 최소화 하는 검정 방법은 존재하지 않으므로 1 종 오류 값을 임의로 설정하고 2 종 오류를 최소화하는 검정 방법을 사용한다. 분석자가 임의로 설정한 1 종 오류를 유의 수준이라 (significant level) 하고 일반적으로 0.1(10%), 0.05(5%), 0.01(1%)을 주로 사용한다.

신뢰구간과 가설 검정은 일대일 관계가 있다. 즉 95% 신뢰구간과 5% 유의수준 가설 검정 결과는 동일하다. 예를 들어 한남대학생 IQ 의 95% 신뢰구간을 구하였더니 $(100,130)$ 이었다면 귀무가설에서 $H_0 : m = m_0$, $m_0$ 의 값을 100 에서 130 의 값으로 설정하면 귀무가설은 채택된다.

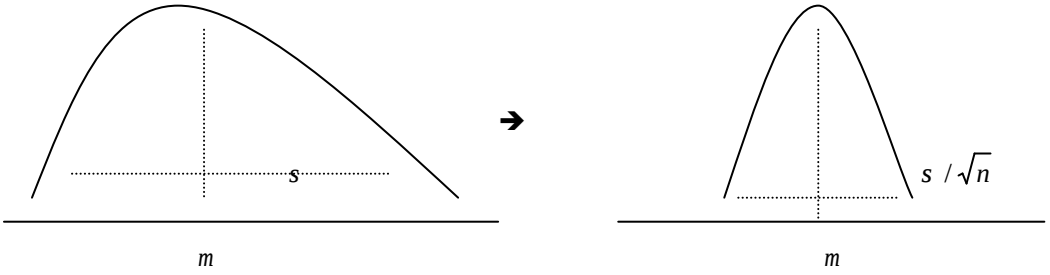
## 8.2. 표본 평균 분포

통계량의 확률 분포를 표본 분포라 (sampling distribution) 한다.

평균과 표준편차 평균이 $m$, 표준편차가 $s$ 인 모집단에서 추출된 크기 n 의 표본으로부터 계산된 표본 평균($\bar{x}$)의 평균은 $m$ 이고 표준 편차는 $s / \sqrt{n}$ 이다.

**중심 극한 정리 (표본 평균의 표본 분포) : Central Limit Theorem**

표본의 크기 n 이 큰 경우(대표본: n>20) 모집단 분포에 관계없이 표본 평균의 분포는 정규 분포를 (normal distribution) 따른다.

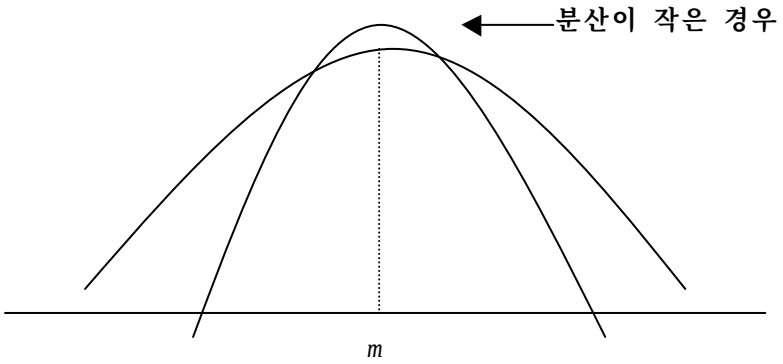



### 8.2.1. 정규 분포 (normal distribution)

Gauss (1777-1855)가 발표한 분포로 물리학 실험의 오차로부터 발견된 분포로 평균을 중심으로 좌우 대칭인 (symmetric, bell shaped) 형태를 갖는 분포 함수이다. Gauss 분포 발견 초기에는 측정된 자료의 히스토그램이 Gauss 분포와 유사하지 않으면 비정상적인 자료로 생각하여 이에 대응하는 개념으로 정규 분포라 부르게 되었다.

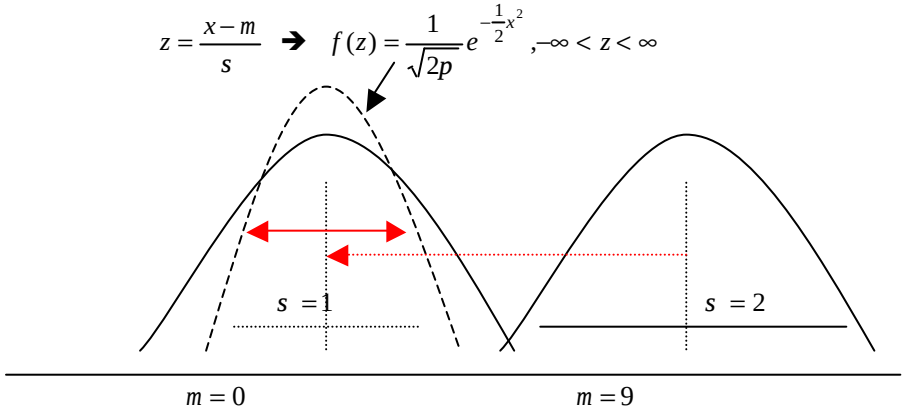
평균 $m$, 표준편차가 $s$ 인 정규분포의 확률 밀도 함수는 다음과 같다.

$$f(x)=\frac{1}{\sqrt{2p}\,s}e^{-\frac{1}{2s^2}(x-m)^2},\quad -\infty<x<\infty$$



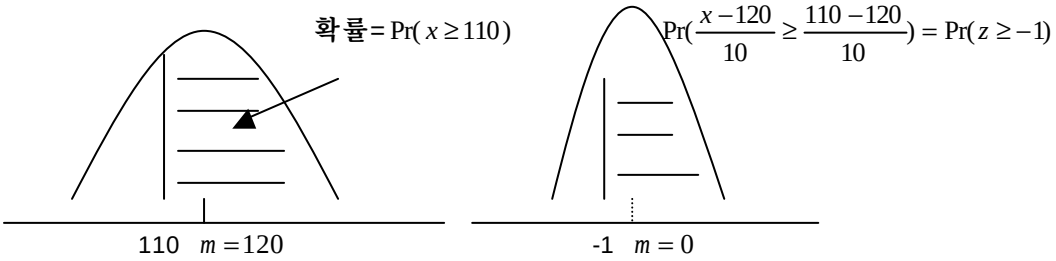

**정규 분포의 표준화**

평균 $m$, 표준편차가 $s$ 인 정규분포를 따르는 $x$ 을 다음과 같이 변형하는 것을 표준화(standardization)한다고 하며 $z$ 의 분포를 표준 정규 분포라 한다.




**표준 정규 분포 표와 확률 구하기**

표준 정규 분포를 고려하는 이유는 정규분포를 따르는 $x$ 의 구간 확률을 하나의 표로 구하기 위해서다. 예를 들어 $x$ (=학생들의 IQ) $(m = 120, s = 10)$ 분포를 따른다고 하자. IQ 가 110 이상 학생들은 몇 %(확률) 정도인가? 평균이 표준 편차가 상이하면 매번 새로운 표를 작성해야 한다. 이런 번거러움을 피하기 위하여 표준 정규 분포표를 사용한다.




표준 정규 분포표에 의해 $\Pr(z \geq -1) = 0.3413 + 0.5 = 0.8413$ 이므로 약 84%이다.

‖ **Exercise** ‖ 한남대학교 대학생의 IQ 는 평균이 120, 표준 편차가 5 인 정규 분포를 따른다고 가정하자.

1) 학생들의 IQ가 125 이상일 확률?

$$\Pr(X \geq 125) = \Pr(\frac{X-120}{5} \geq \frac{125-120}{5}) = \Pr(z \geq 1) = 1 - 0.3413 = 0.1587$$

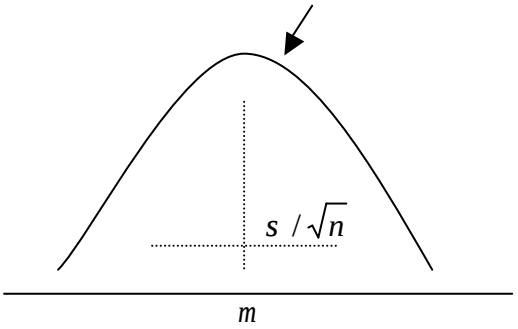
2) IQ가 110 이상, 120 이하일 확률은? $\Pr(110 < X > 120) = \Pr(-2 < z < 0) = 0.4772$

3) IQ가 110 이상인 학생들의 확률은? $\Pr(X > 110) = \Pr(z > -2) = 0.9772$

### 8.2.2. 대표본일 경우 표본 평균 분포

표본 관측치로부터 $(x_1, x_1, \ldots, x_n)$ 계산된 표본 평균$(\bar{x})$의 분포는 표본이 크기 n 이 큰 경우(대표본) 중심극한정리에 의해 정규분포(평균=$m$, 표준 편차=$s/\sqrt{n}$)를 따른다

$$\bar{x} = \sum_{i=1}^{n} x_i / n \sim Normal(m, s/\sqrt{n})$$



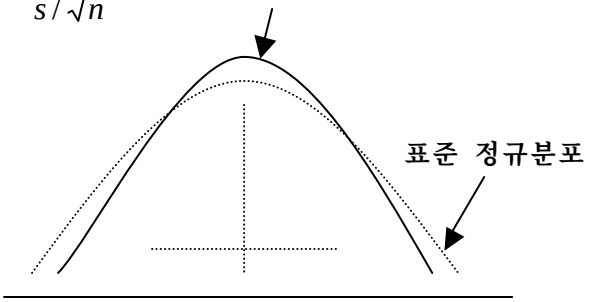

모집단의 표준편차($s$)를 모를 경우 표본의 표준편차(s)를 사용하면 된다.

$$\bar{x} = \sum_{i=1}^{n} x_i / n \sim Normal(m, s/\sqrt{n})$$

### 8.2.2. 소표본일 경우 표본 평균 분포

중심극한 정리를 사용할 수 없다. 그러나 모집단이 정규분포라는 가정이 성립하면 표본 평균의 분포는 t-분포를 따른다. (W.S. Gosset, pen-name 이 Student)

$$\frac{\bar{x} - m}{s/\sqrt{n}} \sim t(n-1) \quad (\text{평균}=0,\ \text{표준편차}=n/(n-2))$$




그러므로 표본의 크기가 커지면 t-분포는 표준 정규분포에 근사(approximate) 한다. t-분포의 경우 정규분포와는 달리 표준화를 이용하여 하나의 표를 만들 수 없다. t 분포 표는 자유도와 자주 사용되는 유의 수준(0.1, 0.05, 0.01)에 따라 만들어져 있다.

## 8.3. 대표본 모평균 추론

‖ **EXAMPLE**‖ 다음은 한남 대학생들의 IQ를 알아보기 위하여 20명을 무작위(SRS)로 추출하여 그들의 IQ를 측정한 자료이다.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 110 | 125 | 100 | 146 | 152 | 114 | 128 | 103 | 111 | 132 |
| 126 | 127 | 105 | 109 | 115 | 138 | 127 | 120 | 116 | 136 |

다음은 엑셀에서 데이터 분석 기능에 의해 IQ 자료의 기초 통계량을 구한 것이다.

| | |
|---|---|
| 평균 | 122 |
| 표준 오차 | 3.187145233 |
| 중앙값 | 122.5 |
| 최빈값 | 127 |
| 표준 편차 | 14.25334679 |
| 분산 | 203.1578947 |
| 첨도 | -0.426424555 |
| 왜도 | 0.413319889 |
| 범위 | 52 |
| 최소값 | 100 |
| 최대값 | 152 |
| 합 | 2440 |
| 관측수 | 20 |
| 신뢰 수준(95.0%) | 6.67077371 |

**점 추정**

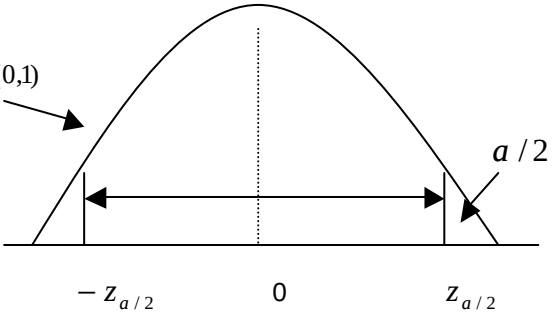
모수( $m$ ) 대한 점 추정치는 표본 평균( $\bar{x}$ ) 이다. 그러므로 한남대학교 학생들 IQ 평균은 122라고 말할 수 있다.

**(100-$a$ )% 구간 추정**

모수( $m$ ) 대한 구간 추정은 표본 평균의 분포를 이용하여야 한다. 표본의 크기가 큰 경우 표본 평균의 분포는 중심극한 정리에 의하여 정규분포를 따른다.

$\bar{x} = 122$ ~ (평균= $m$ , 표준편차= $s/\sqrt{20}$ )

$$\frac{\bar{x}-m}{s/\sqrt{n}} \sim Normal(0,1)$$




**[참고]** 신뢰 구간은 좌우 대칭으로 만드는가? 동일 신뢰 수준이면 구간의 길이가 짧을수록 좋은 추정 방법이다. 그러므로 정규분포나 t-분포는 평균을 중심으로 Bell-shaped 모양이므로 중앙을 중심으로 같은 길이의 신뢰구간을 구하게 된다.

모평균에 대한 (100-$a$ )% 신뢰구간은 $(\bar{x} - z_{a/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{a/2}\frac{s}{\sqrt{n}})$

**‖ EXAMPLE ‖** 한남대학교 학생 평균 IQ에 대한 95% 신뢰구간은

$$(122 - 1.96\frac{14.25}{\sqrt{20}}, 122 + 1.96\frac{14.25}{\sqrt{20}}) \quad \rightarrow \textbf{(115.33, 128.67)}$$

**유의 수준 $a$ 가설 검정**

모집단의 평균에 대해 분석자가 설정한 가설이 맞는지를 표본 평균을 이용하여 알아본다.

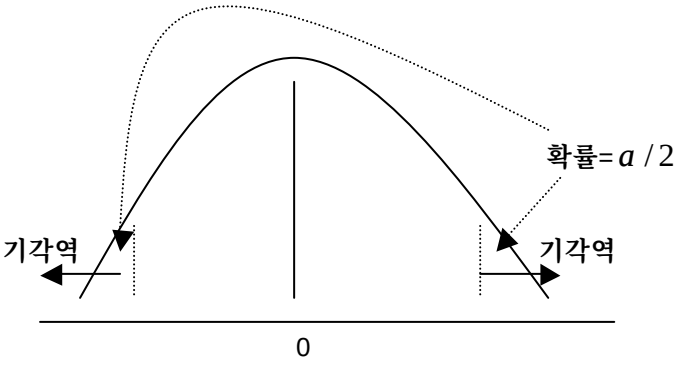
1) 귀무가설: $m = m_0$

2) 대립가설: $m \neq m_0$ (양측 검정), $m > m_0$ / $m < m_0$ (단측 검정)

3) 검정통계량: 가설 검정을 위하여 사용되는 통계량으로 표본으로부터 계산되고 그 통계량의 분포를 알아야 가설 검정이 가능하다.

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}} \sim Normal(0,1)$$

4) 결론: 검정통계량은 귀무가설이 맞다는 가정 하에서 분포함수를 구한다. 그러므로 표본으로부터 구한 검정통계량 값이 분포의 양극단에 가까운 값이면 귀무가설을 기각하게 된다.




위의 경우는 양측 검정을 하는 경우이다. 단측 검정의 경우는 한 쪽에 $a$ 를 모두 배정하고 기각역을 (critical value) 설정한다.

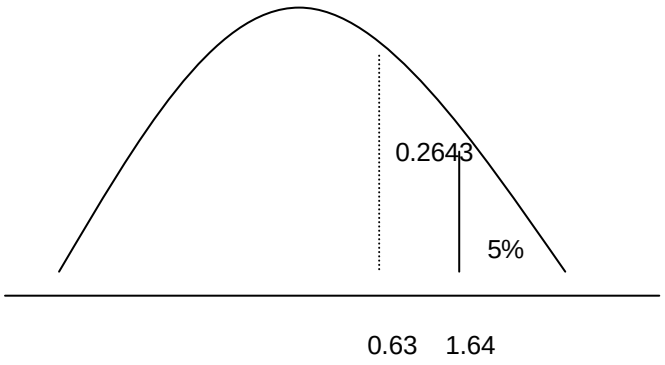
**‖ EXAMPLE ‖** 우리나라 대학생들의 IQ 평균이 120이라 한다. 한남대학교 학생들은 IQ가 120이상 이상인지 알아보고자 한다. (유의수준 5%, $a = 0.05$) 단측 검정

1) 귀무가설: $m = 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이다.

2) 대립가설: $m > 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이상이다 .

3) 검정통계량:

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}} = \frac{122 - 120}{14.25/\sqrt{20}} = 0.63 \sim Normal(0,1)$$



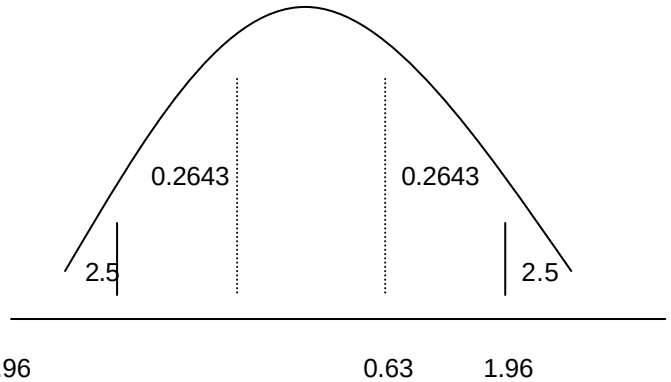

4) 결론: 검정통계량의 값(0.63)이 기각역($\{T > 1.64\}$)에 속하지 않으므로 귀무가설을 채택한다. 혹은 p-value 0.2643가 유의수준 0.05보다 크므로 귀무가설을 채택한다. 즉 한남대학생의 IQ는 우리나라 전체 평균보다 크다고 할 수 없다.

**[p-value]** 검정통계량의 값이 귀무가설을 기각할 최소의 유의 확률이다. 즉 신뢰 수준을 p 값으로 설정한다면 귀무가설을 기각할 수 있다는 것이다.

**‖ EXAMPLE ‖** 한남대학교 학생들은 IQ가 우리나라 대학생들의 평균 IQ와 같다고 할 수 있나? (유의수준 5%, $a = 0.05$) 양측 검정

1) 귀무가설: $m = 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이다.
2) 대립가설: $m \neq 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이 아니다.
3) 검정통계량:

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}} = \frac{122 - 120}{14.25/\sqrt{20}} = 0.63 \sim Normal(0,1)$$




4) 결론: 검정통계량의 값(0.63)이 기각역($\{|T| > 1.96\}$)에 속하지 않으므로 귀무가설을 채택한다. 혹은 p-value 0.5286(=0.2643*2)가 유의수준 0.05보다 크므로 귀무가설을 채택한다. 즉 한남대학생의 IQ는 우리나라 전체 평균과 같다고 할 수 있다.

**[양측검정과 신뢰구간]** 두 방법은 일대일 관계가 있다. (100-α)% 신뢰구간에 있는 모수 값을 귀무가설에 설정하면 유의수준 α하에서는 귀무가설은 기각되지 않는다.

### 8.3.1. 소표본 모평균 추론

다음 2가지 사항을 제외하고는 대표본 검정과 방법이나 절차는 동일하다.

① 모집단이 정규분포라는 가정이 필요하다. (이 가정이 무너지면 비모수 검정 방법을 사용해야 한다)

② 표본 평균의 분포는 t-분포를 따른다.

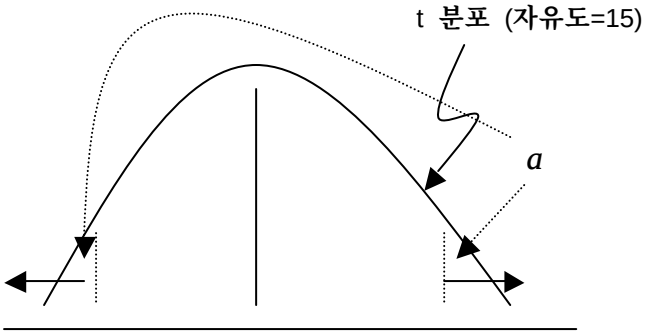
**‖ EXAMPLE ‖** 한남대학교 학생들은 IQ가 우리나라 대학생들의 평균 IQ와 (120) 같은지 알아보기 위하여 16명의 학생들을 임의 추출하여 기초 통계량을 구하였더니 다음과 같았다. 이를 이용하여 유의수준($a = 0.05$) 5% 하에서 가설 검정을 실시하고 95% 신뢰구간을 구하시오.

표본 평균 $\bar{x} = 125$, 표준편차 $s = 8$이었다.

1) 귀무가설: $m = 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이다.
2) 대립가설: $m \neq 120$, 한남대학교 학생들의 IQ 평균($m$)은 120이 아니다.
3) 검정통계량:

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}} = \frac{125 - 120}{8/\sqrt{16}} = 2.5 \sim t(n-1) = t(15)$$

4) 결론: t-분포도 정규분포와 같이 bell-shaped이므로 검정 방법은 정규 분포와 동일하다. 유의수준 0.05, 양측검정의 경우 기각역이 $\{|T| > 2.131)\}$이므로 검정통계량이 T=2.5 이므로 귀무가설이 기각된다. 한남대학교 학생들의 IQ 평균은 전체 평균보다 크다고 할 수 있다.



‖ EXAMPLE ‖ 한남대학교 학생 평균 IQ에 대한 95% 신뢰구간은

$$(\bar{x} - t_{a/2}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{a/2}(n-1)\frac{s}{\sqrt{n}})$$

$$(125 - 2.131\frac{8}{\sqrt{16}}, 125 + 2.131\frac{8}{\sqrt{16}}) \rightarrow (120.74, 129.26)$$

## 8.4 모집단 비율 추론

비율 측정을 위해 표본 자료를 수집하면 그 관측치( $x_i$ )는 "Success" ( $x_i = 1$ )나 "Fail" ( $x_i = 0$ )로 이분류 된다.(binary) 예를 들어 한남대학교 학생들의 하숙 비율을 알아보는 경우 하숙은 성공, 하숙을 하지 않으면 실패로 분류하여 하숙 학생은 $x_i = 1$, 비하숙 학생은 $x_i = 0$으로 입력한다. 다음 사실에 의해 모비율에 대한 추론을 할 수 있다.
비율 추정을 위하여 측정된 $(x_1, x_1, \ldots, x_n)$ 로부터 계산된 표본 평균 $(\bar{x} = \sum x_i / n)$ 은 우리가 구하려는 모비율( $p$ ) 점 추정치, 표본 비율(sample proportion)과 동일하다

$$\hat{p} = \frac{\#of\ success}{n} = \sum_{i=1}^{n} x_i / n$$

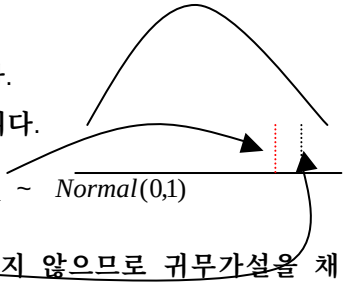
그러므로 대표본일 경우 중심극한 정리에 의해 표본 비율의 분포는 정규 분포를 따른다. 대표본이란 $\min(np, n(1-p)) > 9$을 만족하는 경우이다.

표본 비율의 평균은 $p$ 이고 표준편차는 $\sqrt{p(1-p)/n}$ 이다.

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim Normal(0,1)$$

‖ EXAMPLE ‖ 우리나라 대학생들의 하숙 비율이 27%라고 한다. 한남대학교 학생들의 하숙 비율이 전국 대학생들의 하숙 비율과 같은지 알아보기 위하여 100명을 단순 임의로 추출하여 조사하였더니 하숙생이 100명 중 35명이었다. 양측 검정, 유의수준=0.05

1) 귀무가설: $p = 0.27$, 한남 대학생들의 하숙비율은 0.27이다.
2) 대립가설: $p \neq 0.27$, 한남 대학생들의 하숙비율은 0.27아니다.
3) 검정통계량: $T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.35 - 0.27}{\sqrt{0.27 * 0.73 / 100}} = 1.81 \sim Normal(0,1)$



4) 결론: 검정통계량의 값(1.81)이 기각역($\{|T| > 1.96\}$)에 속하지 않으므로 귀무가설을 채택한다. 혹은 p-value 0.07가 유의수준 0.05보다 크므로 귀무가설을 채택한다. 즉 한남대학생의 하숙 비율은 우리나라 전체 하숙 비율과 같다고 할 수 있다.

모비율에 대한 (100-$\alpha$)% 신뢰구간은 $(\hat{p}-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p}+z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$

$(0.35-1.96\sqrt{0.35(1-0.35)/100}, 0.35+1.96\sqrt{0.35(1-0.35)/100})$ ➔ **(0.257, 0.443)**

**[소표본일 경우 모비율 검정]** 이항 분포와 p-value 를 이용하여 가설을 검정해야 한다. 다소 어려운 개념이므로 본 강의에서는 생략한다.

**HOMEWORK #6 Due the next class**

(1) 평균이 1000, 표준편차가 100 인 정규분포에서 다음확률을 구하시오.

A. Pr (1000 ≤ X ≤ 1200)

B. Pr (X > 1257)

C. Pr (X <1035)

D. Pr (857 ≤ X ≤1183)

E. Pr (X ≤ 700)

F. Pr (812 ≤ X ≤ 913)

G. Pr (X > 891)

(2) A 회사 제품에 대해 고객 만족도를 조사하여 다음 결과를 얻었다.

$$n = 65, \bar{x} = 42.95, s = 2.6424$$

A. 모집단 평균에 대한 99% 신뢰구간을 구하시오.

B. 모집단 평균에 대한 95% 신뢰구간을 구하시오.

C. A, B 중 어는 신뢰구간이 더 넓은가? 이유는?

D. 작년에는 고객 만족도의 평균이 60 이었다. A 회사 제품 고객의 만족도가 높아졌다고 할 수 있나? (유의수준=0.05)

(3) 다음의 A 회사 제품 두께를 측정한 것이다. ( $n=15$) 두께는 정규분포를 따른다고 가정하자.

| | | | | |
|---|---|---|---|---|
| 1.31 | 1.05 | 1.45 | 1.21 | 1.19 |
| 1.78 | 1.37 | 1.41 | 1.22 | 1.11 |
| 1.46 | 1.33 | 1.29 | 1.32 | 1.65 |

A. A 회사 제품 두께 평균에 대한 99% 신뢰구간을 구하시오.

B. A 회사 제품 두께 평균에 대한 95% 신뢰구간을 구하시오.

C. 규격 상 제품의 두께는 1.2 이상이어야 한다. A 회사 제품은 두께가 1.2 이상이라 할 수 있는가? (유의수준 =0.05)

(4) 한국인의 5 세 미만 아이의 남녀 비율은 0.51:0.49 이다. 대구 지역 주민 중 500 명을 조사하였더니 280 이 남자 아이였다. 그럼에도 불구하고 대구 지역에는 남자 아이들이 우리나라 전체 비율과 같다고 할 수 있나? (유의수준=0.05)