1. 용어정리

1) 이 강좌를 시작하며...

다양한 통계패키지를 사용하려는 분들이나 통계분석을 하려는 분들이 어려움을 겪는 부분은 각 통계패키지의 사용방법 측면과 결과해석 측면으로 나눌 수 있다.
본 과정은 이 두 가지 측면 중 결과해석과 관련된 내용을 다른 과정들보다 짧은 5차시로 진행함으로써 그 동안 학습해 본적이 있지만 업무 등에서 사용하지 않아 잊혀 진 통계지식을 리마인드 하는데 의의가 있다.
통계분석결과를 해석하기 위해 우리가 돌아보아야 할 내용들은 자료의 종류, 기초통계량, 정규분포, 표본추출분포, 가설검정 등이 있다.

2) 용어정리

우선 우리가 통계관련 강의를 수강할 때 사용하는 몇 가지 용어를 간략하게 정의해 보자.

. 모집단 : 관심을 가지고 있는 전체 집단
. 확률현상 : 다양성이 나타나는 관심을 가지고 있는 현상
. 표본 : 모집단(확률현상)의 일부분으로서 모집단에 대한 정보를 얻기 위해서 모집단으로부터 추출한 집단
. 모수 : 모집단의 특성을 나타내는 수치(모평균, 모표준편차, 모비율 등)
. 통계량 : 추출된 표본에서 관찰될 값으로부터 구해질 특성치(표본평균, 표본표준편차, 표본비율 등)
. 통계치 : 추출된 표본에서 관찰된 값으로부터 계산되어진 결과치(표본평균값, 표본표준편차값, 표본비율값 등)
. 추정량 : 모수를 추정하기 위하여 표본에서 얻게 될 값
. 추정치 : 모수를 추정하기 위하여 표본에서 얻은 값

3) 자료의 종류

우리가 통계분석을 하기 위해서는 통계 수치의 계산과정을 잘 이해하고 결과의 해석도 잘해야 하지만 이러한 과정을 진행하기 위해 수집되는(우리가 얻게 되는) 자료의 특성을 잘 이해할 필요가 있다. 우리가 얻을 수 있는 자료는 우선 수적 의미 유무에 따라 구분할 수 있다.

• 질적 자료 (qualitative data) : 관찰값이 수적 의미가 없이 범주만을 나타낸다.
• 양적 자료 (quantitative data) : 관찰값이 수적 의미를 갖고 있다.

(1) 질적자료

예를 들어 정책지지여부를 찬성과 반대로 응답한 자료에서 지지여부는 찬성을 0, 반대는 1이라고 표시해도 0, 1은 수로서의 의미가 없기 때문에 질적 자료이다. 그 외에도 질적 자료에는 거주지, 종교, 지지하는 정당에 관한 자료 등이 있다.

또한 연령을 10대, 20대, 30대 등으로 표현한다면 이런 자료도 그 사람이 속한 범주를 나타낼 뿐 직접적인 수적 의미를 나타내지 않으므로 질적 자료가 된다.

위에서 언급한 질적 자료 중 거주지, 종교, 지지하는 정당은 순수하게 범주로서의 의미만 갖는데 이렇게 순수하게 범주로서의 의미만 갖는 질적 자료를 명목형 자료(nominal data)라고 한다. 반면 10대, 20대, 30대 등으로 수집된 연령은 절대적으로 비교할 수 있는 수로서의 의미는 없지만 범주 간에 크고 작음(또는 높고 낮음)이 존재하는 질적 자료를 순위형 자료(ordinal data)라고 한다.

(2) 양적자료

직무만족도의 100점 만점의 점수로 관찰하였다고 할 때 조사된 수치는 만족도에 대한 평가가 되며 수적의미를 갖고 있다. 또한 함께 살고 있는 가족수(나를 제외)를 0명, 1명, 2명...과 같이 조사되면 이 역시 수적의미를 갖고 있고 이를 양적자료라고 한다.

그런데 직무만족도가 87.5점일 수 있을까? 가능한 값이다. 하나 더 예를 들면 신장(키)을 조사한 자료가 있다고 하자. 어떤 사람의 신장이 168.47cm로 관찰되었다면 말이 될까? 실제로 이렇게 소수 둘째자리까지 표현하지 않지만 불가능한 표현은 아니며 그 수치는 의미가 있다. 이렇게 어떤 범위 안에서 무수히 많은 값을 가질 수 있는 자료를 연속형 자료(continuous data)라고 하고 이러한 자료로 표현된 속성을 연속형 변수(continuous variable)라고 한다.
그러나 가족수와 같은 자료는 정수의 값만 가능한 자료이며, 이러한 자료를 이산형 자료(discrete data)라고 하고 이러한 자료로 표현된 속성을 이산형 변수(discrete variable)라고 부른다.

양적 자료 중 연속형 자료도 관찰되는 형태는 정수인 경우가 많아 이산형 자료와 구분이 되지 않는 경우가 많은데 이것은 측정하는 도구의 단위 때문이다. 따라서 연속형 자료와 이산형 자료는 그 관찰된 값의 형태로 구분하는 것이 아니고 관찰 가능한 값이 어떤 영역 안에 있는 모든 실수이면 연속형 자료로, 정수의 값만 가능하면 이산형 자료로 구분하면 된다.

4) 기초통계량

여러 가지 기초통계량들이 있는데 여기서 우리는 정규분포와 관련된 기초통계량을 주로 살펴보도록 하겠다.

<생각해보기!?
세 집단에서 얻은 4개의 관찰값을 비교하면서 세 집단을 표현해 줄 대표값에 대해서 생각해 보자. 얼마라고 답하겠는가?

- A집단의 표본 관찰값 : 1 2 8 9
- B집단의 표본 관찰값 : 1 4 6 9
- C집단의 표본 관찰값 : 3 4 6 7

우리가 양적자료의 요약으로 주로 사용하고 있는 평균인 5를 생각하고 있을 것이다. 평균 5를 대푯값으로 사용하는 것에 대해 어떻게 생각하는가?

대표값으로는 평균, 중앙값, 최빈수 등이 있는데, 최빈수는 관찰값이나 관찰값의 구간중 빈도가 가장 큰 값을 의미한다. 따라서 주로 질적 자료를 정리할 때 사용하면 좋다. 여기서는 먼저 평균을 다루도록 하자.

$$\text{평균} = \frac{\sum_{i=1}^{n} x_i}{n}$$

세 집단 모두 관찰값들의 합이 20이므로 평균은 5가 된다. 따라서 4개의 관찰자료를 근거로 해서 볼 때, A, B, C 세 집단의 평균은 유사하리라고 생각된다. 그러나 A, B집단은 C집단에 비해서 자료가 넓게 퍼져 있어 보이고, A집단과 B집단도 그 흩어짐의 형태가 달라 보인다. 따라서 자료를 요약할 때, 자료를 대표하는 값 이외에 자료의 퍼짐의 정도에 대한 정보가 필요함을 알 수 있다. 이제 자료의 퍼짐, 즉 산포도를 나타내는 잣대(측도)들에 대해 생각해 보자. 먼저 최대값과 최소값의 차이인 범위를 생각해 보자.

범위 = 최대값 - 최소값

A와 B집단의 범위는 '9-1=8'로서 C집단의 범위 '7-3=4'보다 두 배나 크다. 범위를 이용하면 A집단과 B집단은 퍼짐의 정도가 같은 것으로 나오고, C집단은 다른 집단에 비해 퍼짐이 작은 것으로 나온다. 그러나 A집단과 B집단도 자료의 퍼짐의 정도가 다른 것이 보인다. A집단은 양끝에 자료가 몰려 있는 반면, B집단은 양 끝은 넓게 퍼졌으나 다른 두 값은 가운데 몰려 있다. 즉 범위만으로는 A와 B자료의 퍼짐의 정도를 구분할 수가 없다.

이제 자료의 퍼짐을 재는 다른 척도를 생각해 보자.

일반적으로 퍼짐이라는 말은 '중심에서 얼마나 떨어졌는가?'를 말하는 것이니까 각 자료값이 평균으로부터 떨어진 정도, 즉 편차(자료값-평균)를 이용해서 퍼짐정도를 나타낼 수 있다.

- A집단의 편차(관찰값-평균) : -4 -3 3 4
- B집단의 편차(관찰값-평균) : -4 -1 1 4
- C집단의 편차(관찰값-평균) : -2 -1 1 2

이제 이 값들의 크기를 비교하기 위해서 이 수치들의 평균을 계산해 보자. 그런데 평균이 자료의 무게중심이므로 편차들의 합은 항상 0이다. 따라서 편차들의 평균은 집단의 산포를 재는 척도로 적합하지 않다. 편차들의 합이 0이 되는 것은 편차의 부호때문이므로, 이것을 해결하기 위해 편차의 평균을 구하는 대신 "각 관찰값과 평균값(5)과의 차이의 제곱, 즉 편차의 제곱"의 평균을 계산하는데, 이것을 분산이라고 부른다.

$$\text{분산} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

여기서 분모가 $n-1$ 이라고 생각하는 독자는 조금만 더 인내를 갖고 읽어나기기 바란다. 일단 분모를 n으로 하고, 세 집단의 분산을 구해 보자.

- A집단 : 분산 $\Rightarrow \frac{(1-5)^2 + (2-5)^2 + (8-5)^2 + (9-5)^2}{4} = \frac{50}{4} = 12.5$
- B집단 : 분산 $\Rightarrow \frac{(1-5)^2 + (4-5)^2 + (6-5)^2 + (9-5)^2}{4} = \frac{34}{4} = 8.5$
- C집단 : 분산 $\Rightarrow \frac{(3-5)^2 + (4-5)^2 + (6-5)^2 + (7-5)^2}{4} = \frac{10}{4} = 2.5$

분산은 그 집단의 편차제곱의 평균이므로, A집단의 편차제곱이 평균적으로 12.5라는 의미가 되므로, 편차는 평균적으로 $\sqrt{12.5}$=3.54라고 말 할 수 있고, 이렇게 구한 분산의 제곱근을 표준편차라고 부른다. 표준편차는 집단을 이루는 각 관찰값들이 평균으로부터 흩어져 있는 정도의 평균이라고 이해하면 될 것이다.

우리가 실제로 자료를 표현할 때, "평균±표준편차"로 나타내기도 하는데, 이때 평균은 자료의 무게 중심을, 표준편차는 자료의 퍼진 정도를 표현하고, 실제 자료가 평균보다 크거나 작기 때문에 "±"를 이용해서 표현하는 것이다.

- A집단 : 표준편차 $\Rightarrow \sqrt{12.5} = 3.54$
- B집단 : 표준편차 $\Rightarrow \sqrt{8.5} = 2.92$

- C집단 : 표준편차 $\Rightarrow \sqrt{2.5} = 1.58$

그런데 여기서는 우리의 관심대상인 전체 집단(모집단)의 일부분(표본)으로부터 자료를 얻은 관찰값이므로, 분산을 계산할 때는 편차제곱의 합을 개체의 수(n)가 아닌 '$n-1$'로 나누어서 구한다. 이때 '$n-1$'을 자유도라고 부른다.

$n-1$를 이용해 분산과 표준편차를 계산하는 식을 정리해 보면, 다음과 같다.

$$\text{분산} = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

$$\text{표준편차} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$$

이제 A, B, C집단의 분산과 표준편차를 위 공식에 따라 다시 계산해 보자.

◆ 다양한 요약값들

① 대푯값

▶ 평균(Mean)

-모든 관찰치를 사용하여 계산되고, 그 값이 유일하게 결정

-몇 개의 이상치에 민감한 영향을 받음

▶ 중위수(Median)

-모든 관찰치를 크기순으로 정렬했을 경우 50%에 위치한 값

-이상치에 의하여 영향을 받지 않고, 유일하게 결정

▶ 최빈수(Mode)

-자료에서 관측도수가 제일 많은 값으로 유일하지 않다.

-몇 개의 이상치에 영향을 받지 않는다.

② 산포도

▶ 범위(Range)

-관측된 자료에서 가장 큰 값과 가장 작은 값의 차이 ($MAX-\min$)

▶ 분산(Variance)

-각 자료 값과 평균과의 거리를 제곱하여 합을 구한 후 이를 자료의 총수로 나눈 값이며, 표본 분산은 $\frac{\sum(x_i-\overline{x})^2}{n-1}$ 이다.

▶ 표준편차(Standard Deviation)

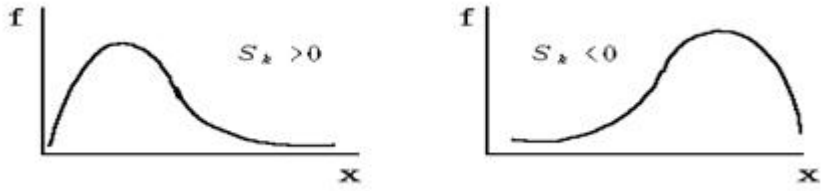
-분산의 양의 제곱근으로 나타내며, 분산보다 측정척도의 차원을 고려하며, 표본 표준편차는 $\sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}$ 이다.

▶ 변동계수(Coefficient of Variance)

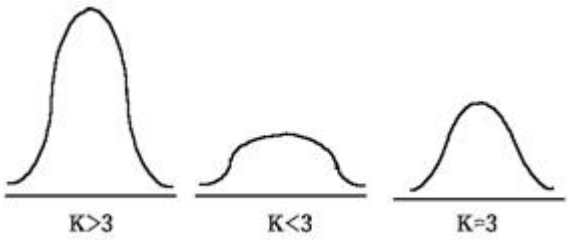
-두 개 이상의 집단의 산포 정도를 비교하는 것으로, 변동계수 CV는 $\frac{s}{\overline{x}}$ 이다.

③ 왜도와 첨도
대칭적인 상태를 중심으로 편중된 정도를 나타냄

▶ 왜도(Skewness) : 분포의 비대칭성과 편중방향을 나타내는 척도(0일 때 자료가 대칭적)



▶ 첨도(Kurtosis) : 분포의 뾰족한 정도를 나타내는 척도(0(또는 3)일 때 정규분포)



2. 정규분포의 이해

1) 정규분포

우리가 관심을 가지고 있는 전체집단의 분포에 대해 학습해본 적이 있을 것이다. 그리고 정규분포에 대해 많이 들어본 적이 있을 것이다. 정규분포가 어떤 형태였는가?
다음과 같이 가운데가 높고 이를 중심으로 좌우 대칭적인 형태를 하고 있다.

정규분포는 독일의 수학자 가우스(Gauss, 1777~1855)가 각종 물리학 실험을 수행할 때 수반되는 계측오차에 대한 확률분포로서 가우스 분포(Gauss distribution)라는 이름으로 정규분포를 제시한 이래 많은 학문 분야에서 가우스 분포를 기본 확률모형 또는 근사적인 확률모형으로 채택하였다. 특히 통계학의 초기 발전단계에는 모든 자료의 히스토그램이 가우스 분포의 곡선형태와 비슷하지 않으면 자료수집 과정이 잘못된 것이라고 믿었던 적도 있었다. 이렇게 많은 분야에서 연속형 변수로서 나타나는 현상을 표현하는 확률모형으로 이 분포가 자리를 잡음에 따라 정규분포(normal distribution)라 불리게 되었다.

정규분포는 중심을 나타내는 평균(μ)과 산포를 의미하는 표준편차(σ)에 의해 형태가 결정되는 분포로 어떤 확률현상이 정규분포를 따른다고 가정하는 것은 그 현상에서의 관찰치들로부터 그려진 y축이 상대도수밀도인 히스토그램의 모습을 다음과 같은 종모양의 함수(확률밀도함수)로 보겠다는 입장을 말하는 것이다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

어떤 확률현상이 평균이 μ이고 표준편차가 σ인 정규분포를 따른다고 할 때 우리 주위에서 자주 사용되는 정규분포의 특성은 μ와 σ에 따라서 다음과 같다.

- 이 현상의 한 관찰값이 $(\mu-1\sigma\ ,\ \mu+1\sigma)$에서 나타날 확률이 0.6826
- 이 현상의 한 관찰값이 $(\mu-2\sigma\ ,\ \mu+2\sigma)$에서 나타날 확률이 0.9544
- 이 현상의 한 관찰값이 $(\mu-3\sigma\ ,\ \mu+3\sigma)$에서 나타날 확률이 0.9974

어느 지역의 1인 가구의 생활비를 평균 170만 원, 표준편차 20만 원인 정규분포를 가정했다는 말은 이 지역에 있는 1인 가구의 약 68%의 생활비가 150만 원$(=170-1\times20)$에서 190만 원$(=170+1\times20)$으로 190만 원 이상 쓰는 가구와 150만 원 이하로 쓰는 가구가 각각 약 16%쯤이라고 보겠다는 것이다. 또한 210만 원$(=170+2\times20)$이상 생활비를 쓰는 1인 가구가 전체의 약 2.28%$(\frac{1-0.9544}{2}=0.0228)$ 라고 생각하겠다는 것이다.

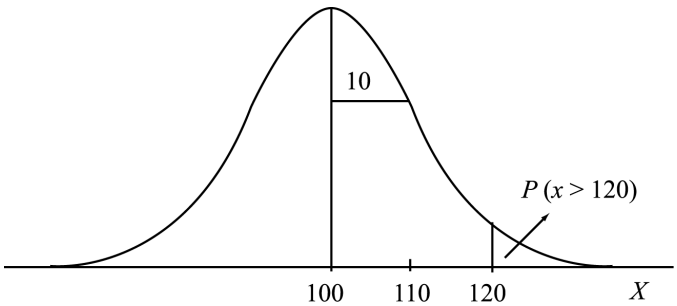
이것을 일반 통계학 교재에서는 확률변수라는 어려운 기호를 써서 다음과 같이 나타낸다.

$$P(\mu-1\sigma < X < \mu+1\sigma) = 0.6826$$
$$P(\mu-2\sigma < X < \mu+2\sigma) = 0.9544$$
$$P(\mu-3\sigma < X < \mu+3\sigma) = 0.9974$$

보다 일반적인 예로 평균이 100이고 표준편차가 10인 정규분포를 따른다고 가정하는 집단에서 120이상의 값을 갖는 개체의 비율을 알고 싶다고 하자. 이 비율은 120이상이 될 확률 $P(X\geq120)$을 의미한다.

참고로 이 확률을 구하는 방법은 히스토그램에서 y축이 상대도수밀도이면 각 기둥의 면적이 상대도수이므로 120보다 큰 쪽에 있는 기둥의 면적을 구하면 되는 것과 같이 정규분포 확률밀도함수에서 120보다 큰 쪽의 면적을 구하면 된다.

그러나 수학문제를 풀듯이 면적을 구할 필요는 없다. 지난날 우리는 표준정규분포표라는 것을 학습한바가 있으며 그 표를 이용하면 면적을 쉽게 구할 수 있게 된다.



<그림 3-3> 평균이 100이고, 표준편차가 10인 정규분포

2) 표준정규분포

그런데 정규분포 확률밀도함수에서 120보다 큰 쪽의 면적을 구하기 위한 수학적인 적분이 공식으로는 불가능하기 때문에 컴퓨터를 통해서 근사적으로 계산할 수밖에 없다. 그런데 정규분포는 μ와 σ에 따라서 분포의 모양이 달라지므로 평균이 100이고 표준편차가 10인 정규분포에서 120이상의 확률을 계산하려면 컴퓨터를 이용해서 확률을 다시 계산해야 한다.

통계학자들은 정규분포가 평균에 따라 분포가 자리하는 위치의 중심이 결정되고 표준편차에 비례해서 분포의 폭이 커지는 특성이 있다는 것을 이용해서 한 가지 정규분포의 확률만 알고 있으면 어떤 정규분포의 확률도 계산할 수 있는 표준화(standardization) 기법을 개발했다.

표준화 기법이란 개체의 관찰값과 평균의 차이를 표준편차로 나눈 값을 이 개체의 표준화 값으로 정의하는 개념이다. 표준화 값 Z를 구하는 식은 다음과 같다.

$$Z = \frac{X-\mu}{\sigma}$$

이렇게 구해진 표준화 값은 본래의 관찰값이 신장에 대한 자료이든 체중에 관한 자료이든 무엇이 되었든지 평균이 0이고 표준편차가 1인 값들이 된다.

앞의 예에서 평균이 100이고 표준편차가 10인 정규분포를 따른다고 가정한 현상에서 개체들의 표준화 값은 자동적으로 평균이 0이고 표준편차가 1인 형태의 정규분포, 즉 표준정규분포(standard normal distribution)를 따르게 된다. 대부분의 통계학 책의 부록으로 붙어있는 표준정규분포표가 이렇게 구한 것이다. 따라서 평균이 100이고 표준편차가 10인 정규분포에서 120이상일 확률을 구하라는 것은 표준정규분포에서 120의 표준화 값인 $2\left(=\frac{120-100}{10}\right)$이상인 확률을 구하는 것이 되어서 표준정규분포표를 이용하면 쉽게 구할 수 있다.

X가 정규분포를 따를 때, 그 분포의 형태가 다음과 같은 특징이 있었다.

$$P(\mu-1\sigma < X < \mu+1\sigma) = 0.6826$$
$$P(\mu-2\sigma < X < \mu+2\sigma) = 0.9544$$
$$P(\mu-3\sigma < X < \mu+3\sigma) = 0.9974$$

따라서 Z는 다음과 같은 특징을 갖게 된다.

$$P\left(-1 < Z=\frac{X-\mu}{\sigma} < 1\right) = 0.6826$$

$$P\ (-2 < Z=\frac{X-\mu}{\sigma} < 2\ ) = 0.9544$$

$$P\ \ (-3\ <\ Z=\frac{X-\mu}{\sigma}\ <3\ )\ =\ 0.9974$$

표준화 값 Z는 자료값과 평균과의 차이가 그 집단의 표준편차의 몇 배에 해당하는 지를 알려주는 값으로 어떤 값에 해당하는 표준화 값을 통해 그 값이 속한 집단에서 갖는 상대적인 위치를 알 수 있다.

다음의 예를 통해서 표준화값과 확률값을 구하는 연습을 해 보자.

예제1) 다음 자료를 보고 표준화값을 구해보자

평균 100, 표준편차 10인 정규분포		평균 80, 표준편차 4인 정규분포		평균 10, 표준편차 2인 정규분포		평균 2, 표준편차 0.1인 정규분포	
원값	표준화값	원값	표준화값	원값	표준화값	원값	표준화값
110		84					
120		44					
90		76					
85		74					
80		72					

평균과 표준편차가 다르지만 표준화값은 동일하게 평균으로부터 표준편차만큼 떨어진 위치는 절대값 1로, 평균으로부터 2배 표준편차 만큼 떨어진 위치는 절대값 2등으로 나타나는 것을 확인할 수 있다.
즉, 이를 통해 표준화 값을 통해 그 값이 속한 집단에서 갖는 상대적인 위치를 알 수 있다는 것을 다시 확인해 볼 수 있다.

예제2) 그러면 다음의 확률을 구해보자.

어떤 공장에서 생산되는 전구의 무게가 평균 100g이고 표준편차가 10g인 정규분포를 따른다고 하자. 전구 하나를 랜덤하게 뽑았을 때 이 전구의 무게가 110시간 이상일 확률을 구한다고 하면, 위의 문제와 동일한다.

평균 100, 표준편차 10인 정규분포
110이상일 확률

이를 풀면

표준화 값 Z가 $1\left(=\frac{110-100}{10}\right)$ 이상일 확률을 구하면 다음과 같다.

$$P(110 \le X) = P\left(\frac{110-100}{10} \le \frac{X-\mu}{\sigma}\right)$$
$$= P(1 \le Z) = 0.1587$$

어떤 공장에서 생산되는 전구의 무게가 평균 100g이고 표준편차가 10g인 정규분포를 따른다고 하자. 전구 하나를 랜덤하게 뽑았을 때 이 전구의 무게가 85시간 이하일 확률을 구한다고 하면, 위의 문제와 동일한다.

평균 100, 표준편차 10인 정규분포
85이하일 확률

이를 풀면

표준화 값 Z가 $-1.5\left(=\frac{85-100}{10}\right)$ 이상일 확률을 구하면 다음과 같다.

$$P(85 \ge X) = P\left(\frac{85-100}{10} \ge \frac{X-\mu}{\sigma}\right)$$
$$= P(-1.5 \ge Z) = 0.0668$$

3. 표본추출분포

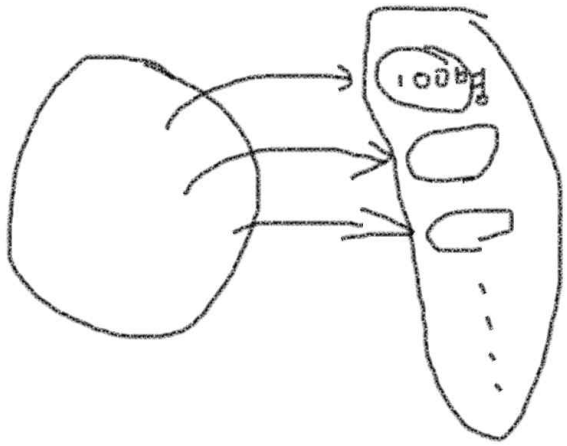
1) 표본추출분포

모집단을 전국의 만 19세 남자로 하고 그들의 키에 관심이 있다고 해보자. 만 19세 남자들의 키는 평균이 μ이고 표준편차가 σ인 정규분포를 따른다고 한다.

임의의 100명을 추출하면 100명의 표본평균은 얼마로 기대되어질까?
=>모평균 μ로 기대된다.

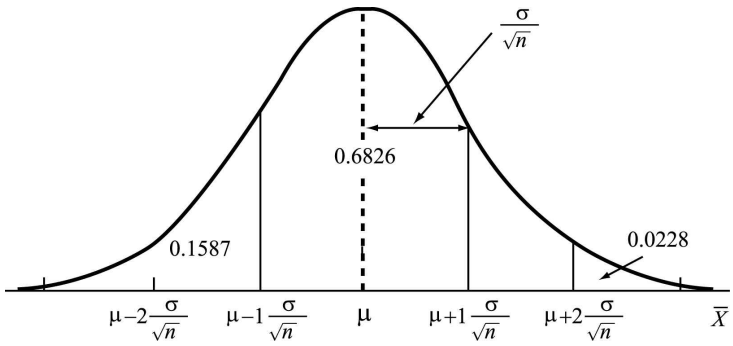
임의의 100명의 표본평균은 기대되는 값과 동일한 값이 나올까?
=>그럴 수도 아닐 수도 있다. 임의의 100명을 추출할 수 있는 경우가 무수히 많다. 무수히 많은 경우 중에 한 가지가 추출된 것으로 100명의 평균은 다양한 값들 중에 하나이다.

그런데 표본으로 추출된 100명은 평균이 μ이고 표준편차가 σ인 정규분포를 따르는 모집단에서 추출되었다. 따라서 100명의 평균들은 모평균인 μ근처 값이 주로 있을 것이다. 또한 모평균인 μ보다 큰 값도 있을 것이고 작은 값도 있을 것이다.



다시 말해, 100명의 표본평균을 가능한 한 모든 경우를 조사했다고 생각해보면 100명의 표본평균들은 다양할 것인데 모평균 μ를 중심으로 분포한다고 생각해볼 수 있다.
그래서 100명의 표본평균들의 분포는 모평균이 μ인 정규분포를 따른다고 알려져 있다.
이때 표준편차는 가능한 한 모든 경우의 100명의 표본평균이 얼마나 다른지를 나타내는 값으로 다음과 같이 나타낼 수 있으며($\frac{\sigma}{\sqrt{n}}=\frac{\sigma}{\sqrt{100}}$), 표본평균의 표준오차라고 부른다.

즉, 100명의 표본평균은 평균이 μ이고 표준오차가 $\frac{\sigma}{\sqrt{n}}=\frac{\sigma}{\sqrt{100}}$인 정규분포를 따른다고 할 수 있다.



처음에 학습했던 정규분포 때와 마찬가지로 표본평균의 분포는 정규분포이므로 $\overline{X}$가 정규분포를 따를 때, 그 분포의 형태가 다음과 같은 특징이 있다.

$$P\left(\mu - 1\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1\frac{\sigma}{\sqrt{n}}\right) = 0.6826$$

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$$

$$P\left(\mu - 3\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 3\frac{\sigma}{\sqrt{n}}\right) = 0.99744$$

따라서 Z는 다음과 같은 특징을 갖게 된다.

$$P\left(-1 < Z = \frac{X-\mu}{\sigma/\sqrt{n}} < 1\right) = 0.6826$$

$$P\left(-2 < Z = \frac{X-\mu}{\sigma/\sqrt{n}} < 2\right) = 0.9544$$

$$P\left(-3 < Z = \frac{X-\mu}{\sigma/\sqrt{n}} < 3\right) = 0.9974$$

다음의 예를 통해서 표준화값과 확률값을 구하는 연습을 해 보자.

예제1) 다음 자료를 보고 표준화값을 구해보자
1) 평균 100, 표준편차 10인 정규분포에서 임의의 100명을 추출하였을 때, 다음 100명의 평균에 대한 표준화 값을 구해보자.

평균 100, 표준편차 10인 정규분포 => 평균 100, 표준오차 1	
100명의 평균	표준화값
101	
102	
99	
98.5	
98	

2) 평균 100, 표준편차 10인 정규분포에서 임의의 25명을 추출하였을 때, 다음 25명의 평균에 대한 표준화 값을 구해보자.

평균 100, 표준편차 10인 정규분포 => 평균 100, 표준오차 2	
100명의 평균	표준화값
102	
104	
98	
97	
96	

앞 강좌에서와 마찬가지로 평균과 표준오차 및 표본크기가 다르지만 표준화값은 동일하게 평균으로부터 표준오차만큼 떨어진 위치는 절대값 1로, 평균으로부터 2배 표준오차 만큼 떨어진 위치는 절대값 2등으로 나타나는 것을 확인할 수 있다.
즉, 이를 통해 표준화 값을 통해 그 값이 속한 집단에서 갖는 상대적인 위치를 알 수 있다는 것을 다시 확인해 볼 수 있다.

예제2) 그러면 다음의 확률을 구해보자.

어떤 공장에서 생산되는 전구의 무게가 평균 100g이고 표준편차가 10g인 정규분포를 따른다고 하자. 전구 100개를 랜덤하게 뽑았을 때 이 전구의 평균 무게가 101시간 이상일 확률을 구한다고 하면, 위의 문제와 동일한다.

평균 100, 표준편차 10인 정규분포
100개 평균이 101이상일 확률

이를 풀면

표준화 값 Z가 $1\left(=\frac{101-100}{10/\sqrt{100}}\right)$ 이상일 확률을 구하면 다음과 같다.

$$P(101 \le \overline{X}) = P\left(\frac{101-100}{10/\sqrt{100}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)$$
$$= P(1 \le Z) = 0.1587$$

어떤 공장에서 생산되는 전구의 무게가 평균 100g이고 표준편차가 10g인 정규분포를 따른다고 하자. 전구를 임의로 25개 뽑았을 때 이 전구의 무게가 85시간 이하일 확률을 구한다고 하면, 위의 문제와 동일한다.

평균 100, 표준편차 10인 정규분포
97이하일 확률

이를 풀면

표준화 값 Z가 $-1.5\left(=\frac{85-100}{10/\sqrt{25}}\right)$ 이상일 확률을 구하면 다음과 같다.

$$P(97 \ge X) = P\left(\frac{97-100}{10/\sqrt{100}} \ge \frac{\overline{X}-\mu}{\sigma/\sqrt{100}}\right)$$
$$= P(-1.5 \ge Z) = 0.0668$$

그렇다면 모집단이 정규분포를 따르지 않는다면 어떻게 될까? 그래도 정규분포를 적용할 수 있을까? 이를 위해서 통계학에는 다음과 같은 중심극한정리가 있다.

모집단이 정규분포를 따르지 않아도 n이 충분히 크면 (보통 30이상이면) 표본평균의 분포는 근사적으로 정규분포를 따른다. 이때 표본평균의 기댓값은 모평균(μ)이고 표본평균의 표준오차는 $\frac{\sigma}{\sqrt{n}}$이다.

4. 가설검정

1) 가설설정

① 귀무가설이란 별다른 문제가 없는 한 나타나리라고 예상되는 현상에 대한 기존의 입장이 하나 있었다. 이와 같은 입장을 귀무가설이라고 하고 H_0이라고 쓴다.

- 공정한 주사위의 1의 눈이 나올 확률은 1/6이다.
- A제품의 불량률이 0.2(20%)라고 알려져 있다.
- A사의 직무만족도는 평균 80이다.

② 대립가설 : 귀무가설(기존의 생각)에 상반된 입장이 있다. 이것을 대립가설 또는 대안가설이라고 하고 H_1 또는 H_A라고 쓴다.

- 공정한 주사위의 1의 눈이 나올 확률은 1/6이 아니다.
- A제품의 불량률이 0.2(20%)를 넘는다.
- A사의 직무만족도는 평균 80이 아니다.

위의 내용을 정리하면 다음과 같다.

예	귀무가설(H_0)	대립가설(H_1)
주사위, 직무만족도: 양측검정	공정하다($P=1/6$).	공정하지 않다($P\neq 1/6$).
불량품 : 단측검정	불량률이 0.2이다($P=0.2$).	불량률이 0.2보다 크다($P>0.2$).

주사위의 예에서 대립가설 $P\neq 1/6$는 1/6보다 클 수도 있고 작을 수도 있다는 것을 다 포함한다. 또한 직무만족도도 $\mu\neq 80$는 80보다 클 수도 있고 작을 수도 있다는 것을 다 포함하므로 양방향 모두를 주장하기 때문에 양측검정이라고 한다. 반면, 불량품의 예는 대립가설이 $P>0.2$라는 한 방향만을 주장하기 때문에 단측검정이라고 한다.

2) 가설검정 과정

1단계 : 귀무가설(H_0)과 대립가설(H_1)을 수립한다.

2단계 : 검정을 위한 표본추출 또는 확률실험을 설계한다.

3단계 : 의사결정의 기준을 정한다.

3) 1종오류와 2종오류

10개의 제품을 조사한 후 귀무가설과 대립가설 중 어느 하나를 받아들일지 결정하는 과정에서 범할 수 있는 오류에는 어떤 것이 있을까? 표로 정리해 보자.

		의사결정	
		H_0 채택	H_0 기각
사실	H_0 true	옳은 결정	1종오류
	H_0 false	2종오류	옳은 결정

불량품의 예에서 우리가 범할 수 있는 오류는 두 가지이다. 불량률이 0.2가 맞는데(H_0이 참인 데) 불량률이 0.2가 넘는다고 결정(H_0을 기각, H_1을 채택)할 오류(1종오류)와 불량률이 기존의 생각과 달리 0.2를 넘는데(H_0이 거짓, H_1이 참인데) 불량률이 여전히 0.2라고 결정(H_0을 채택)할 오류(2종오류)가 있다. 여기서 다시 한 번 두 가지 용어를 정의해 보자.

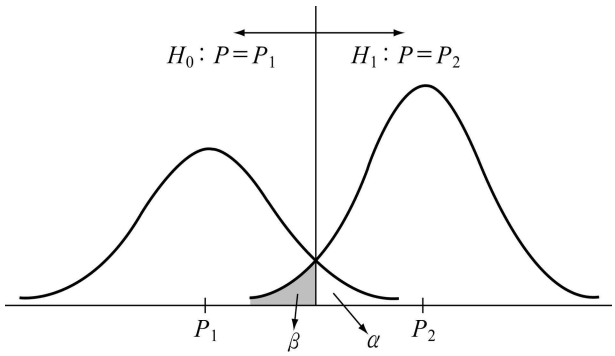
○ 1종오류(Type I error) : 귀무가설이 참일 때 귀무가설을 기각하게 되는 오류
○ 2종오류(Type II error) : 대립가설이 참일 때 대립가설을 기각하게 되는 오류

우리는 이 두 오류 중 기존의 생각(귀무가설)이 옳음에도 불구하고 기존의 생각을 틀렸다고 하는 실수(1종오류)를 가급적 범하지 않으려는 입장을 취하려고 한다.

4) 유의수준

대립가설을 채택할 기준을 정하면 그에 따른 1종오류의 최대 확률을 계산할 수 있는데 이 확률을 유의수준(α)이라고 부른다. 실제로 검정을 할 때는 유의수준 α를 미리 정해 놓고, 1종오류를 범할 확률이 α이하가 되도록 검정규칙을 정한다. 이렇게 구한 검정규칙을 유의수준 α에서의 검정규칙 이라고 부른다.

유의수준 : 연구자가 허용하는 1종오류를 범할 최대확률



α: 1종 오류를 범할 최대 확률, β: 2종 오류를 범할 최대 확률

5. 일표본 z검정

1인가구의 생활비가 $N(\mu=170,\sigma=10)$를 따른다고 알려졌다. 그런데 최근에 이 입장에 대하여 1인 가구의 생활비가 170보다 커졌다고 하는 반론이 강하게 제기되어 이를 검정하려고 한다.

입문과정에서 다룬 전형적인 일표본 평균검정문제이다. 검정은 유의수준 5%에서 수행하기로 한다.

이를 위하여 임의의 1인 가구 25명을 임의 표본으로 추출하여 얻은 표본평균 $\overline{x}$(소문자로 쓰임)는 174cm, 표본표준편차 s(소문자로 쓰임)는 9cm이다.

《 일 반 적 풀 이 》

【단계 1】가설 수립

$H_0 : \mu \le 170$

$H_1 : \mu > 170$

가설은 모집단에 관한 입장이므로 $\overline{x}$를 쓰는 것이 아니고 μ를 쓰는 것을 주목하라. 또한 등호"="가 귀무가설에 첨부되는 것도 주목하여야 한다.

【단계 2】 기각역 결정

$\overline{X}$를 크기 25인 임의 표본의 평균이라고 할 때 유의수준 5%에서 기각역을 구하는 것은 $P(\overline{X} > C \mid H_0$이 참, 즉 $\mu=170) = 0.05$ 를 만족하는 C를 구하는 것이다. 독자들은 기초통계입문과정에서 $\overline{X}$의 표본추출분포가 평균이 $\mu_{\overline{X}}=170$이고 분산이 $\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} = \frac{10^2}{25} = 4$인 정규분포임을 잘 알고 있으리라 생각한다.(이 부분은 굉장히 중요하다.) 따라서 표준정규분포의 95백분위수 $Z_{0.05}$가 1.645이므로 $\frac{c-170}{\sqrt{4}} = 1.645$ 가 되어 위로부터 $C = 170 + 2\times 1.645 = 173.29$가 되어 기각역은 $\{\overline{x} \ : \ \overline{X} > 173.29\}$이다.

【단계 3】 검정

그런데 추출된 표본에서 계산된 평균이 174로서 173.29보다 크므로 우리가 얻은 표본의 결과에 근거하여 유의수준 5%에서 1인 가구의 평균 생활비가 μ가 170만원 이하일 것이라는 귀무가설을 기각한다.

가설검정의 가설수립부터 검정단계까지 일반적인 풀이과정을 따라서 해보았다.
이 부분을 재해석해 보겠다.

【단계 1】가설 수립단계
통계적 가설검정에서는 검정 대상이 되는 귀무가설과 귀무가설의 대안으로서 제기되는 대립가설(또는 대안가설)이 있다. 통상 기존의 입장을 귀무가설로 하고 주장하고자 하는 입장을 대립가설로 세우지만 검정의 이론이 갖고 있는 한계 때문에 등호 "="는 반드시 귀무가설에 포함되어야 한다. 또한 가설을 설정하기가 모호한 경우는 실증적(자료를 근거로)으로 보일 수 없거나, 보이기가 힘든 것을 귀무가설로 한다.

【단계 2】기각역 결정과 해석
기각역을 결정할 때 항상 "귀무가설이 참일 때"가 전제되는 것과 유의수준이 핵심적인 역할을 하는 것을 알고 있을 것이다.

우리가 표본을 추출하였을 때, 추출된 표본으로부터 구해진 통계량이 기대되어지는 값은 귀무가설에서 설정한 값일 것이다. 그러나 귀무가설에서 설정한 값과 일치할까? 조금은 차이가 생길 수 있을까?
귀무가설이 참이라고 전제하더라도 통계량이 귀무가설 값과 완벽히 일치하지는 않을 것이다 (정규분포를 생각해보라)

따라서 우리는 귀무가설이 참이라면 이것에 강력하게 반하는 자료가 추출되었을 때, 더 이상 귀무가설을 참이라고 할 수 없다. 다시 말해 귀무가설이 참이라는 의심을 하게 하는 정도가 어느 정도이상 되지 않으면 일반적으로는 그냥 기존입장을 유지하는 쪽으로 가겠다는 것이다.

우리의 문제로 돌아가 보면, 귀무가설이 $\mu \leq 170$이므로 $\mu > 170$이라고 하는 대안가설에 대하여 귀무가설을 인정하려는 의지를 갖고 있는 사람으로서는 당연히 μ를 170만원으로 놓고 생각을 진행할 것이다.
1인 가구의 생활비 μ가 170만원이라고 생각하는 사람의 입장에서 이 집단에서 임으로 뽑은 25명의 평균 신장을 어느 정도로 예상하겠는가?
아마도 170만원근처라고 생각할 것이다. 그러나 임으로 추출하다보면 우연히 추출된 사람들 중에 큰 사람들이 포함되어 평균 170만원보다 클 수도 있으리라. 당연히 반대로 작은 사람들이 우연히 좀 많이 포함될 수도 있으니까 170만원보다 작을 수도 있겠으나 이 상황에서는 그것을 고려할 필요는 없을 것이다(왜? : 1인 가구의 생활비(성인전체의 평균키 μ)가 170만원가 아닐 것이라는 대립가설에서는 170만원보다 클 수도 혹은 작을 수도 있는 상황을 다 생각해 볼 것이다. 그러나 1인 가구의 생활비가 110보다 커졌다고 하는 대립가설에 대하여 이 대립가설이 사실인 경우를 생각하는 입장에서는 170만원보다 작을 수도 있는 상황을 고려할 필요가 없다.)
그렇다면 당신은 전체 1인 가구의 생활비가 170만원이지만, 25명의 평균이 어느 정도 클 때까지 우연히 큰 사람들이 표본으로 좀 많이 추출되었다고 생각하겠는가? 표본으로 추출된 25명의 평균이 171cm 혹은 172cm라면 당신의 마음은 어떠하겠는가? 25명의 평균이 180cm이어도 당신은 성인의 평균 키는 170만원인데 우연히 추출된 25명이 우연히 큰 사람들이 표본

으로 뽑혔다고 생각할 것인가?

여기서 유의수준 5%가 그 답을 준다. 즉, 유의수준 5%라는 말은 평균 신장 μ가 170만원일 경우 임의의 25명의 평균값이 나타날 수 있는 평균값 중에서 상위 5% 이상이 되면 "우연히" 큰 사람들이 많이 뽑혀서 그렇게 큰 값이 되었을 것이라고 생각하지 않고 "특별한 우연" 사건으로 간주하며 이유를 따져 왜 그렇게 되었는지 설명을 해보겠다는 것이다.

【단계 3】검정

그런데 이 예에서는 임의 추출된 25명의 신장 평균이 174cm이므로 검정관계자들은 우연히 나타난 것으로 받아들이지 않고 왜 그랬는지 따져보는 입장을 택한다. 이때 이들은 어떤 것들을 따져 볼 수 있을까?

여기서는 성인의 키가 정규분포를 따른다는 것과, 성인의 키의 표준편차가 10이라는 것이다. 성인의 키가 정규분포를 따른다는 것은 일반적으로 받아들여지는 가정이다. 하지만 가정에 다소 의심이 간다고 하더라도 표본의 크기가 25이므로 중심극한정리에 의하여 표본평균 $\overline{X}$의 표본추출분포가 정규분포에 가깝다는 사실을 사용할 수 있다. 그렇기 때문에 그렇게 심각한 문제를 일으키지 않는다고 본다.

여러분들이 알고 있는 z-표준화를 이용한다. 그러면 다음과 같은 기각역을 세울 수 있을 것이다.

$$\frac{C^{*}-170}{9/\sqrt{25}} = Z_{(0.05)} = 1.645$$

위 기각역으로부터 $C^{*} = 170 + 1.645 \times \frac{10}{\sqrt{25}} = 173.29$가 기각된다. 기각역은 $\overline{X} > C^{*}$가 된다.

이제 우리는 유의수준 5%에서 귀무가설을 기각하고 μ가 170보다 크다는 대립가설을 지지하게 된다.

【단계 3】검정의 다른 측면 : 유의수준과 유의확률

“유의수준 5%에서”의 의미를 생각해 보자. 귀무가설이 참이라면 25명의 1인 가구 생활비 평균은 170만원 근처 값이 추출될 것이다. 그런데 170만원 보다 상당히 큰 값이 추출될 수도 있다. 그러나 귀무가설이 참이라면 170만원 보다 상당히 큰 값이 추출될 수는 없을 것으로 본다. 따라서 우리가 추출한 25명의 평균이 상당히 큰 값이 아니라면 귀무가설이 참이라는 것이다. 여기서 상당히 큰 값을 어떻게 정해야 할까? 173만원? 174만원? 쉽지 않다. 우리가 관심을 갖고 있는 모집단에 따라 상당히 큰 값을 매번 고민해야 한다. 그렇다면 어떤 모집단이던 상당히 큰 값을 상위 5%이내라고 본다면 어떨까? 이를 유의수준이라고 한다.
그리고 실제로 추출된 25명의 1인 가구 생활비 평균은 174만원이었다. 이 값이 상위 몇% 인지를 알아낸다면 상당히 큰 값인지 아닌지 판단할 수 있다.
이를 위해 우리가 앞에서 학습했던 표본추출분포 내용이 다시 필요해진다.
25명의 평균의 표준화값을 구하면,

$$\frac{174-170}{10/\sqrt{25}}=2$$

로 표준화값이 2보다 더 큰값이 나올 확률을 구하면 2.28%로 이를 유의확률이라고 한다.
이 둘을 비교하면 유의수준 5%에서 귀무가설을 기각하고 μ가 170보다 크다는 대립가설을 지지한다고 말하는 것이다.