상담연구통계이해

[상담연구방법과 통계분석의 이해] 표준오차의 분모에 왜 루트 N이 들어가는가?

헤아림
2016. 6. 20. 8:55

이웃추가

[상담연구방법과 통계분석의 이해] 표준오차의 분모에 왜 루트 N이 들어가는가?

표준오차의 공식은 아래와 같다.

$$\frac{s}{\sqrt{n}}$$

표준오차는 왜 표준편차 S를 $\sqrt{n}$ 으로 나누는가? 무슨 사정이 있어 그러는가?

이를 이해하기 위해 우리는 길을 좀 돌아가야 할 필요가 있겠다.

▶ 모집단 1,2,3에 대한 모평균과 모분산을 구하기

모집단에 1,2,3 이라는 자료가 있다고 가정하고, 모평균과 모분산을 구하는 방식에 두가지가 있다.

① 모평균 $\mu=\frac{\Sigma X}{N}$, 모분산 $\sigma^2=\frac{\sum(X_i-\overline{X})^2}{N}$

② 모평균 E(X)= $\sum$확률변수×확률 , 모분산 Var(X)=$E(X-\mu)^2=E(\text{측정값}-\text{평균})^2$

①번 방식은 측정값을 나열해놓고 다 더한 후 계산하는 일반적인 방식이다. 예를 들어 자료 (1,2,3)에 대한 평균을 구하면 $\overline{X}=\frac{1+2+3}{3}$ =2

②번 방식으로 구한 평균을 E(X)으로 표현한다. 이는 확률변수와 확률이 제시된 표를 통해 구한 평균을 의미한다. 그리고 이 방식으로 구한 평균을 1번 방식으로 구한 평균과 구분하기 위해 기댓값(expected value)이라고 따로 부른다.
2번 방식으로 계산하려면 아래와 같은 표를 만들게 된다.

<표 1. 기댓값>

확률변수	1	2	3
확률	1/3	1/3	1/3

확률변수란 쉽게 말하자면 수집된 데이터 점수를 '종류'의 개념으로 바꾼 것이다. 그러니까 (1,2,3)이 수집되었을 때, 확률변수는 1이라는 종류, 2라는 종류, 3이라는 종류를 의미한다. 만약 (1,2,3,4)라고 한다면 확률변수는 4가지가 된다. 만약 (1,1,3,4)라면 확률변수는 3가지가 있는 것이다.

그리고 <표1>의 확률이란 그 확률변수가 취하는 확률을 의미한다. 이것은 확률 변수마다 몇 개의 값이 수집되었는지 숫자를 센 후에 자료의 총 개수로 나눈다. 그러면 그 확률변수가 나타날 가능성을 확률로 표현할 수 있게 된다. 1,2,3이 각각 한번씩 나타났고 총 3개의 자료가 있으니, 각각의 확률은 1/3이다.
이제 E(X)= $\sum$확률변수×확률=1×1/3+2×1/3+3×1/3=6/3=2

모평균과 기댓값이 같다는 것을 볼 수 있다. 이것은 불변의 법칙이다. 언제나 μ=E(X)는 같은 값이 나온다.
그런데 왜 굳이 기댓값이란 방식으로 평균을 구하는가? 그 이유는 모집단이 단지 수집된 자료의 나열로만 제시되는 경우도 있지만, 때로는 <표1>처럼 수집된 자료의 종류(확률변수)와 그 자료의 확률로 제시될 경우도 있기 때문이다. 이런 식으로 표를 제시하는 경우는 대체로 어떤 실험이나 조사를 무수히 반복했을 때 예상되는 평균값을 생각해보자는 의미가 있다.

무수히 반복이라는 말에서 짐작할 수 있듯이 기댓값은 현실적인 결과보다는 이론적으로 나올 수 있는 결과와 관련있다. 그래서 기대값은 실용적인 통계분석기법으로 활용되기 보다는 실용적인 통계분석기법이 나올 수 있도록 토양이 되어주는 편이다. 마치 인체의 움직임을 실용적인 통계분석기법이라고 한다면, 기댓값은 인체의 뼈처럼 살속에 있어 보이지는 않지만 그 통계분석기법이 존재하도록 이론적으로 받쳐주는 역할을 한다고 볼 수 있다.

<표1>의 자료가 모집단의 자료라고 가정했다. 우리는 여기서 기댓값의 공식과 비슷한 스타일의 공식으로 분산을 구할 수 있다. (사실 이 계산방식은 언제나 모분산을 구하는 것이기 때문에 자료가 모집단이라는고 가정을 하든 안하든 상관은 없다)
<표1>의 형태로 모분산을 구할 때, 이 모분산을 Var(X)로 표현한다. 그리고 Var(X)는 그리스어로 σ^2으로 쓸 때도 많다. 이 둘이 같다는 것을 기억하자.

Var(X)=$E(X-\mu)^2=E(\text{측정값}-\text{평균})^2$
=$\sum(\text{측정값}-\text{평균})^2\times$확률 <-E가 사라지고 시그마 기호가 나타나면서 확률이 곱해지는 것에 주의
=$(1-2)^2\times1/3+(2-2)^2\times1/3+(3-2)^2\times1/3$
=(1+1)/3=0.66666.... = σ^2

만약 1,2,3을 표본 분산을 구하는 공식을 적용한다면, 1이 된다.

꽃보다 심리학 헤아림 심리연구소

그러나 모분산을 계산하는 공식으로 하면 $\frac{\sum(X_i-\overline{X})^2}{N}=\frac{2}{3}$=약 0.67 로 나온다.

그래서 <표1>의 형태로 분산을 구할 때 모분산을 구하는 것이지, 표본분산을 구한다고 생각하면 혼동될 수 있으니 주의하자.

우리는 <표1>의 형태로 기댓값과 모분산을 구하는 공식을 알게되었다.

▶ 표본평균분포에서 모평균과 모분산, 표준오차

이제 표준오차에 관한 이야기를 진행해보자.

표준오차는 '표본평균들의 세계'에서 표본평균들 간의 표준편차를 의미한다고 했다. 그래서 우리는 지금 표본평균들의 세계를 따로 만들어봐야 한다.

모집단의 자료가 1,2,3 이라고 했다. 이 세 개의 자료에서 우리는 표본의 크기가 일정한 표본 집단을 만들어야 한다. 또한 이 때 중복 가능한 표본 집단을 만든다. (중복가능한 표본 집단이 만들어진다는 가정하에 표준오차라는 개념이 만들어졌기 때문이다)

표본의 크기를 2로 통일해서 표본집단을 만들면 아래와 같다.

아래와 같이 표본집단이 구해질 것이다.

표본집단	각 집단의 표본 자료		각 표본집단의 평균
1	1	1	1
2	1	2	1.5
3	1	3	2
4	2	1	1.5
5	2	2	2
6	2	3	2.5
7	3	1	2
8	3	2	2.5
9	3	3	3

1번 공식의 방식으로 표본평균의 평균과 표본 평균의 분산을 구해보자.

표본평균의 모평균 $\mu=\frac{\Sigma X}{N}$=2

그래서 모집단에서 표본크기를 동일하게 하여 찾아낸 모든 표본집단의 평균들에 대한 평균은 모집단 평균과 같다는 것을 알 수 있다.

표본평균의 모분산 $\sigma^2=\frac{\sum(X_i-\overline{X})^2}{N}$=약 0.333

모집단에서 표본크기를 동일하게 하여 찾아낸 모든 표본집단의 평균들에 대한 분산은 모집단의 분산과 다름을 알 수 있다. 그 대신 언제나 표본평균의 모분산은 모집단의 모분산을 표본집단의 표본 크기로 나눈 값과 동일하게 계산된다.

앞에서 우리는 모집단 1,2,3의 모분산이 0.6666... 임을 알았고, 이 값을 표본집단의 사례수 2로 나누면 표본평균들의 분산이 나온다. 그래서 일단 여기서

표본평균의 분포에서 표본평균들의 분산= $\frac{\sigma^2}{n}=\frac{0.666\cdots}{2}$ =0.333....

(σ^2은 모집단 1,2,3에서 모분산, n은 표본집단의 표본 크기 2이다.모집단의 표본 크기 3이 아님에 주의)

표본평균들의 분산에 루트를 씌우면 표준편차를 구할 수 있다.
$\sqrt{\frac{\sigma^2}{n}}=\frac{\sigma}{\sqrt{n}}=\sqrt{0.3333}$ =0.577 (σ^2이 σ가 되었기 때문에 σ는 모표준편차라고 부르게 된다)

우리는 표본평균들의 분포에서 계산한 표준편차가 표준오차의 개념이라는 것을 기억하고 있다. 즉 0.577이 모집단의 표준오차인 것이다.
지금 표준오차의 공식으로 $\frac{\sigma}{\sqrt{n}}$ 이 사용될 수 있음을 보았다.
그런데 우리가 실제적으로 쓰는 공식은 $\frac{s}{\sqrt{n}}$ 이다. 여기에는 s, 즉 표본 표준편차가 들어간다. 모표준편차대신 표본표준편차를 쓴다.

우리가 실제 통계에서는 하나의 표본집단을 갖고 분석을 한다. 표본집단에서 계산된 표준편차는 모표준편차를 모를 때, 모표준편차를 대신할 수 있다는 이론적인 가정이 있기 때문에 $\frac{s}{\sqrt{n}}$ 와 같은 공식을 사용할 수 있게 된다.

지금부터는 2번 공식으로 표준오차를 이해해보자.

우선 아래와 같은 표를 구할 수 있다.

<표 2. 표본평균들의 기댓값>

확률변수 $\overline{X}$	1	1.5	2	2.5	3
확률	1/9	2/9	3/9	2/9	1/9

표본평균들의 평균을 구해보자.

E($\overline{X}$)= $\sum$확률변수×확률=1×1/9+1.5×2/9+2×3/9+2.5×2/9+3×1/9= 2

(* E( ) 안에 $\overline{X}$이 들어간다는 것에 주의하자. 지금 표본평균들의 평균을 구하는 것이기 때문이다)

우리는 앞서 2가 모평균임을 알고 있다.

그래서 표분평균 분포에서 표본평균들의 기댓값 E($\overline{X}$)는 언제나 모평균과 같다는 원칙이 나온다.

E($\overline{X}$)=μ

이제 모분산을 구해보자. <표2>의 형태에서 모분산을 구하는 방법은

Var($\overline{X}$)=$E(\overline{X}-\mu)^2$

$=(1-2)^2\times1/9+(1.5-2)^2\times2/9+(2-2)^2\times3/9+(2.5-2)^2\times2/9+(3-2)^2\times1/9$

$=\frac{1+0.5+0+0.5+1}{9}=\frac{3}{9}$

=0.3333....

$\sqrt{Var(\overline{X})}$=표본평균들의 표준편차= 이는 곧 표준오차 =0.577

1번 계산 방식과 2번 계산 방식은 동일한 값을 계산한다는 것을 이해할 수 있다.

▶ 표준오차에 대한 이론적인 공식 유도

지금은 Var($\overline{X}$)에 숫자를 대입하지 말고 이론적으로 계산 공식을 이끌어내보자.

Var($\overline{X}$)=Var($\frac{(X_1+X_2+\cdots+X_n)}{n}$) <- $\overline{X}$ 는 $\frac{(X_1+X_2+\cdots+X_n)}{n}$이기 때문에

<- 여기서 n은 표본집단의 표본 크기이다.

$=\frac{1}{n^2}Var(X_1+X_2+\cdots+X_n)$

<- n이 n^2으로 빠져나오는 이유는 분산의 공식이 갖는 특징 때문이다. Var(bX)=b^2Var(X)

$=\frac{1}{n^2}(Var(X_1)+Var(X_2)+\cdots+Var(X_n))$

<-여기서 아래로 내려가기 전에 많이 헷갈린다. 그 이유는 $Var(X_1)$에 대한 표현 때문이다. X에 붙은 1,2,n이라는 숫자에 너무 신경쓰지 말자. 그 대신 모집단에서 꺼낸 표본집단의 분포에 대해 분산을 나타내는 것으로 생각하자. 즉 $Var(X_1)$, $Var(X_2)$, $Var(X_3)$ 은 같은 모집단에서 추출하되 서로 영향을 주지 않는 독립된 분포를 나타낸다고 생각하자. 이론적으로 그렇게 가정을 해놓는 것이다. 그러면 아래의 공식을 응용하여 그 다음 과정으로 넘어갈 수 있다.

Var(X+Y)=Var(X)+Var(Y)=$\sigma_x^2+\sigma_y^2$

이 공식은 어떤 확률변수와 어떤 확률변수를 합할 경우의 분산은 각 확률변수의 분산을 합한 것과 같다는 의미가 있다.

$=\frac{1}{n^2}(\sigma^2+\sigma^2+\cdots+\sigma^2)$ <- 동일한 σ^2으로 표현되는 이유는 표본집단의 사례수가 충분한 크기를 갖는다면, 같은 모집단에서 추출한 표본집단의 분산은 동일할 것으로 가정하기 때문이다.

$=\frac{n\sigma^2}{n^2}$ <- σ^2이 n개 만큼 있기 때문에 아래와 같이 표현된다. 여기서 n을 하나씩 지울 수 있다.

$=\frac{\sigma^2}{n}$

이 분산의 공식에 루트를 씌우면 다시 $\sqrt{\frac{\sigma^2}{n}}=\frac{\sigma}{\sqrt{n}}$ 임을 알 수 있다. 우리가 실제적으로 표본집단에서 구한 표준편차로 모평균의 분포 상태를 추측해야 하므로 모표준편차 대신에 표본표준편차를 쓴다.

그래서 아래와 같은 공식을 쓰게 된다.

표준오차= $\frac{S}{\quad}$