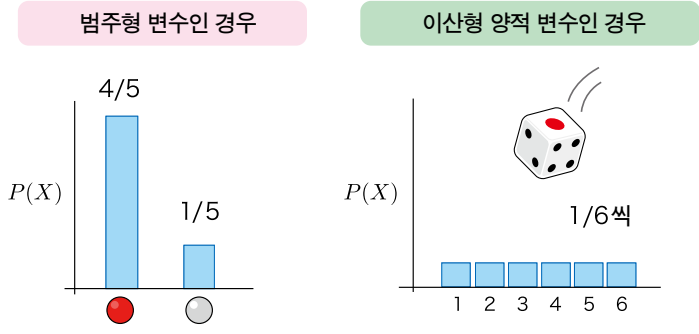
$P(X$=흰 구슬$)=1/5$의 식으로 표현할 수 있습니다. 여기서 X와 같이 확률이 달라지는 변수를, **확률변수**라 부릅니다.* 그리고 확률변수가 실제로 취하는 값(여기서는 붉은 구슬 또는 흰 구슬)을 **실현값**이라 합니다.

확률변수가 '붉은 구슬/흰 구슬' 같은 범주이거나, '주사위 눈'처럼 서로 떨어진 숫자인 경우 이산형 확률변수, 키(cm)와 같이 연속한 값인 경우 연속형 확률변수라 합니다. 둘의 차이는 확률분포에서 나타납니다.

● 확률분포

확률분포란 가로축에 확률변수를, 세로축에 그 확률변수의 발생 가능성을 표시한 분포입니다. 확률변수가 이산형인 경우 세로축이 확률 그 자체를 나타냅니다(**그림 3.4.2**). 히스토그램과 마찬가지로 확률변수가 범주일 때는 가로축 순서에 의미는 없습니다.

◆ 그림 3.4.2 **이산형 확률분포**

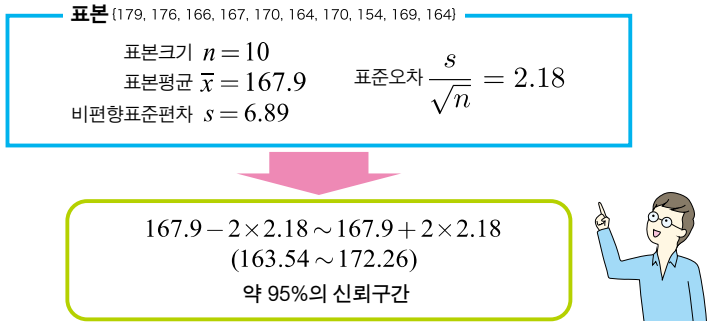


가로축은 확률변수, 세로축은 확률

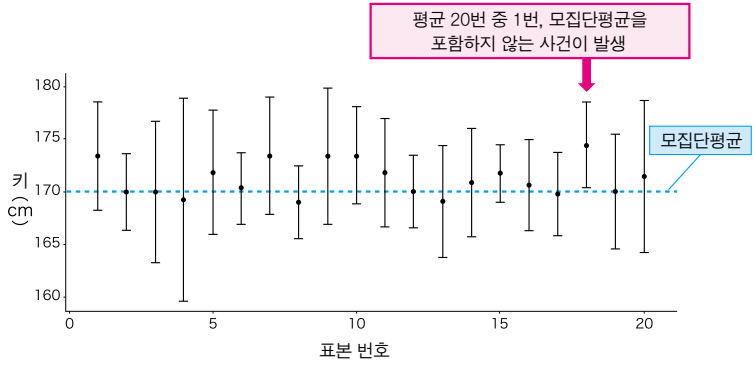
* 측도론을 이용한 확률론에서는 확률변수 X를 사건을 나타내는 집합의 요소를 실수로 변환하는 함수라고 정의합니다. 예를 들어 붉은 구슬이 나오면 100원, 흰 구슬이 나오면 500원을 받는 뽑기라면 X({붉은 구슬})=100, X({흰 구슬})=500이 됩니다. 이 책에서는 간단하게 이해하고자 확률적으로 변하는 변수를 확률변수라고 하고 이야기를 진행합니다.

그림 4.2.7 용어 오기 표본오차 => 표준오차

◆ 그림 4.2.7 **신뢰구간의 예**



◆ 그림 4.2.8 **직감적으로 95% 신뢰구간 이해하기**



μ=170인 모집단에서 표본크기 n=10인 표본을 추출하고, 95% 신뢰구간을 그리는 작업을 20번 반복했습니다. 검은색 동그라미가 표본평균, 위아래로 늘인 선이 95% 신뢰구간을 나타냅니다. 18번째 표본으로 그린 95% 신뢰구간은 모집단평균 μ(파란색 선)를 포함하지 않습니다.

지 않습니다. 즉, 95% 신뢰구간이란 평균적으로 20번 중 1번 정도 벗어난다는, 달리 말하면 20번 중 19번은 구간에 모집단평균을 포함한다는 뜻입니다.

t 분포와 95% 신뢰구간

지금까지 '약 95%'라는 표현으로 뭉뚱그려 이야기를 진행했습니다만, 이쯤

그림 6.3.2 적합도 계산 값 오기 4.33 => 13.8

어떤 주사위를 60번 던져, 1부터 6까지의 눈이 각각 5, 8, 10, 20, 7, 10번씩 나왔다고 합시다(**그림 6.3.2**). 이 데이터가 각 눈이 나올 확률이 모두 1/6인 정상 주사위에서 얻은 것인지를 해석해 보겠습니다.

카이제곱검정의 적합도검정에서는 먼저 귀무가설의 확률분포에서 얻을 수 있는 기대도수를 계산합니다. 기대도수란 전체 개수에 각 확률을 곱한 값입니다. 즉, 가장 나타나기 쉬운 실현값인 셈입니다. 예를 들어 전체 60번에서 1/6 확률이라면, 기대도수는 각각 10이 됩니다.

다음으로, **그림 6.3.2**에서 보듯이 각 눈의 (실제 출현도수−기대도수)^2/(기대도수)를 계산하고, 이를 더한 값을 구합니다. 이 검정통계량을 χ^2값(카이제곱값)이라 부르는데, 귀무가설이 옳다면 이는 χ^2분포(카이제곱분포)라는 확률분포를 따릅니다. 이 분포 안에서 실제로 얻은 χ^2값의 위치를 구하여 p값을 도출합니다.

예시 데이터에서는 p=0.017이 되어 대립가설을 채택하므로, 통계적으로 유의미하게 이 주사위는 올바른 주사위가 아니라고 판단할 수 있습니다.

여기서는 각 눈이 똑같이 1/6씩 나온다는 균등한 확률분포를 가설로 세웠습니다만, 임의의 확률분포도 사용할 수 있습니다. 예를 들어 한국인의 혈액형

◆ 그림 6.3.2 **적합도검정**

적합도검정

얻은 출현도수(개수)가 이론적인 비율(이산확률분포)에 따라 얻어진 것인지를 조사하는 검정

예 : 올바른 주사위인가?

	1	2	3	4	5	6	합계
출현도수	5	8	10	20	7	10	60
이론적 비율 (확률)	1/6	1/6	1/6	1/6	1/6	1/6	1
기대도수	10	10	10	10	10	10	60

$$\chi^2 = \sum \frac{(\text{출현도수} - \text{기대도수})}{(\text{기대도수})}$$

$$= (5\text{-}10)^2/10 + \cdots + (10\text{-}10)^2/10$$

$$= 13.8$$

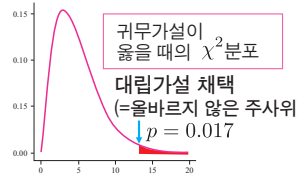


그림 8.2.1 범례 누락 가로축: 저온/고온 세로축: 줄기 길이

두 번째 요인(설명변수 x_2)을 '저온/고온'으로 하여, 두 요인이 줄기 길이에 어떤 영향을 미치는지를 분석한다고 합시다. 상호작용을 고려하지 않는 모형이라면 [줄기 길이]=절편+b_1x_1(비료 없음=0 또는 비료 있음=1)+b_2x_2(저온=0 또는 고온=1)+ε이 됩니다.

이 모형에서 비료 효과는 저온이든 고온이든 상관없으며, 온도 변화의 영향 역시 비료 유무와는 상관없습니다. 그러나 실제 현상이라면 비료 효과가 온도에 따라 달라질 가능성이 있습니다. 이에 상호작용항 $c_1x_1x_2$를 추가한 이원배치 분산분석을 시행할 수 있습니다.

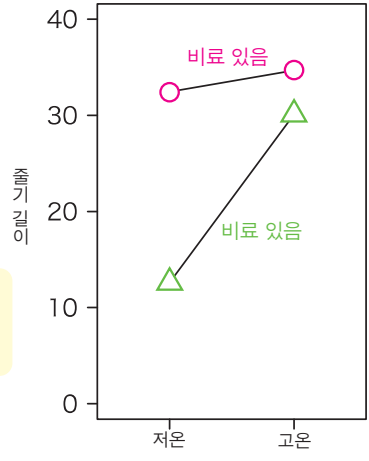
추가된 상호작용항을 x_2에 관해 정리하면 $(b_2+c_1x_1)x_2$이므로, 지금까지 일정했던 b_2를 대신하여 $(b_2+c_1x_1)$을 x_2에 곱하게 됩니다. 즉, x_2가 0(저온)에서 1(고온)이 될 때의 줄기 길이는 비료 유무 x_1에 따라 달라진다는 것을 뜻합니다. **그림 8.2.1**은 이러한 상호작용이 있는 예를 보여줍니다. 온도가 변했을 때의 효과는 비료 유무에 따라 달라진다는 것을 알아볼 수 있습니다.

◆ 그림 8.2.1 **이원배치 분산분석**

줄기 길이(cm)	비료	온도
35	있음	높음
31	있음	낮음
27	없음	높음
13	없음	낮음
⋮	⋮	⋮

주효과 …1개 요인으로 좁힌 효과

상호작용 …2개 요인의 상승효과



2개 요인인 비료와 온도의 차이에 의해 줄기 길이가 달라지는가를 조사한 실험 데이터입니다. 오른쪽 그림의 ○와 △는 각 조건에서 줄기 길이의 평균값을 나타냅니다.