통계학실험(005)
과제 4 (5장)
2022-18608 문현호

1

예제 1. 주어진 자료는 전체 부동산에 대한 자료이므로 모집단으로 생각할 수 있다. 거래가 이루어진 전체 부동산의 집의 크기의 평균값은 얼마인가? 모분산은 얼마인가?

```
ames = pd.read_csv("../data/ames.csv")

living_area = ames["Gr.Liv.Area"]

print("Mean value of total : ", living_area.mean())
print("Variance of total : ", living_area.var())

'''
Mean value of total :  1499.6904436860068
Variance of total :  255539.23531322062

'''
```

- 계산에 따라, 집의 크기의 모평균은 1499.69, 모분산은 255539가 나왔다.

2
예제 2. 모집단에서 크기가 60인 랜덤 표본을 선택하자. 모집단 평균에 대한 점추정값은 얼마인가?

```
sp60 = living_area.sample(n=60)
print("Mean value of the sample : ", sp60.mean())

'''
Mean value of the sample :  1499.75

'''
```

- 계산에 따라, 모집단 평균에 대한 추정값은 1499.75이다.

3
예제 3. 예제 2에서 선택된 표본을 이용하여 모평균에 대한 95% 신뢰구간을 구해보자. 이때, 모분산은 예제 1에서 구한 값을 사용하도록 한다. 이 신뢰구간은 모평균을 포함하는가?

```
mean = living_area.mean()
std = living_area.std()/math.sqrt(60)
print(norm.ppf(0.025, mean, std), " ~ ", norm.ppf(0.975, mean, std))

'''
1371.7813972553886  ~  1627.599490116625

'''
```

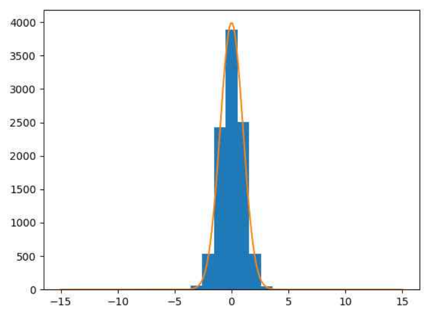
- 위의 표본에서 얻은 모분산, 모표준편차를 이용해서 모평균에 대한 95% 신뢰구간은 1371.78 ~ 1627.60 정도였다. 모평균인 1499.69는 이 신뢰구간 안에 포함된다.

add 1
예제1. 정규분포 N(0, 5^2)에서 크기가 25인 sample을 추출하여 표본평균을 구하는 시행을 10000번 반복하고, 각 시행에서의 표본평균을 벡터에 저장하도록 한다. 표본평균 벡터로 histogram을 그리고 아래의 코드를 시행하여 얻게 되는 plot과 비교하여라.

```
means = []
for _ in range(10000):
    a = norm.rvs(loc=0, scale=5, size=25)
    means.append(a.mean())
plt.hist(means, bins=np.linspace(-15, 15, num=30))

# 아래 코드
x = np.linspace(-15, 15, num=10000)
plt.plot(x, 10000*norm.pdf(x))
plt.show()
```

'''



'''

- 표본평균의 표준편차는 모표준편차/sqrt(표본의 크기) = 1이기 때문에, 표준편차가 1인 정규분포와 그 양상이 똑같아진다. 위 그림에서는 정규분포의 값에 10000을 곱하여, 히스토그램과 넓이가 같도록 하였기 때문에 값이 같다.

add 2
예제2. 다음은 대한민국 고등학생들의 키에 대한 표본 자료이다. 모분산은 50이다.
[177.4, 161.5, 158.3, 179.0, 174.2, 182.2, 162.6, 170.3, 162.9, 155.4, 175.9, 174.7, 177.3, 164.5, 168.1, 162.9, 177.3, 165.4, 159.8, 159.4]
(1) 이 자료로부터 대한민국 고등학생 키의 평균, 표준편차를 추정하고 대한민국 고등학생 평균 키의 추정량에 대한 표준오차를 추정하여라.
(2) 95% 신뢰구간을 구하여라.

```
samples = [177.4, 161.5, 158.3, 179.0, 174.2, 182.2, 162.6, 170.3, 162.9, 155.4,
           175.9, 174.7, 177.3, 164.5, 168.1, 162.9, 177.3, 165.4, 159.8, 159.4]
df = pd.DataFrame(samples)
print(df.mean())
print(df.std())
print(df.std()/math.sqrt(20))

mean = df.mean()
std = math.sqrt(50)/math.sqrt(20)
print(norm.ppf(0.025, mean, std), " ~ ", norm.ppf(0.975, mean, std))

'''
0      168.455
dtype: float64
0      8.167231
dtype: float64
0      1.826248
dtype: float64
[165.35602484]  ~  [171.55397516]

'''
```

- 대한민국 고등학생의 키는 평균 168.455, 표준편차는 8.167231, 그리고 이를 이용해 구한 평균 키 추정량에 대한 표준오차는 1.826248로 추정되었다.
- 주어진 모표준편차를 이용한 모평균의 95% 신뢰구간은 165.356~171.554였다.

add 3
예제3. 어느 배터리의 평균수명이 평균 100(시간) 이고 표준편차 30(시간)인 정규분포
를 따른다고 하자. 이때, 신제품 배터리 n=16개를 시험 생산한 결과 표본평균=111(시간)으로 나타났다. 이 결과를 통해 신제품 배터리의 평균수명이 기존 배터리와 다르다고 말할 수 있는가? 유의수준 5%에서 이를 확인하여라.
(1) 문제 상황에 맞는 귀무가설과 대립가설을 서술하여라.
(2) 위의 가설검정 상황에 맞는 검정통계량과 유의확률을 구하여라.
(3) 신제품 배터리의 평균수명의 95% 신뢰구간을 구하여라.
(4) 유의수준 5%에서 가설검정한 결과를 서술하여라

```
mean = 100
std = 30/4
print(norm.ppf(0.025, mean, std), " ~ ", norm.ppf(0.975, mean, std))

'''
85.30027011594959   ~   114.69972988405041

'''
```

- 귀무가설, 신제품 배터리의 평균수명은 기존 배터리와 같다 / 대립가설, 신제품의 평균수명은 기존과 다르다.
- 검정통계량 : 표본평균 111, 개수 16개. 유의확률은 5%로 생각하면 된다.
- 위 계산에 따라, 95% 신뢰구간은 85.3 ~ 114.7로 나타났다.
- 유의수준 5%에서의 가설검정 결과, 귀무가설을 기각할 수 없다. 따라서 신제품 배터리의 평균수명은 기존 배터리와 같다고 생각할 수 있다.