모집단

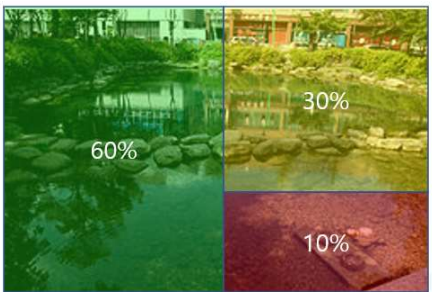
2023년 3월 16일 목요일 오후 12:15

모집단은 찌개 요리?

- 우리가 찌개음식을 만들 때, 누구나 간을 본다는 행위를 한적이 있을 것이다. 첨가한 양념이 골고루 섞였다면 한 숟가락으로 간을 본다고 해도 그 맛이 음식의 전체의 맛을 반영할 것이다.
- 통계적 추정도 이와 같은 것이다. 부분으로 전체를 판단하는 것을 우리는 통계적 추정이라고 한다.
- 하지만 아주 가끔은 간이 맛있다고 해서 음식 전체가 맛있는 경우가 100% 존재하는 것은 아니다. 조금은 다를 가능성이 있다는 점을 고려해야 한다.

랜덤 샘플링과 모평균

- 아래와 같은 연못이 존재한다. 연못의 총 합은 1이며, 각각의 구역별로 잡히는 물고기의 종류가 다른다.



서로 다른 색으로 구분되어 있는 연못이 있다.
각각의 구역별로 다른 물고기들이 무한한 숫자 크기로 살고 있다.
가장 큰 곳에는 광어가 살며, 중간 크기는 우럭이 살고, 작은 곳에는 붕어가 존재
이 말은 구체적으로 확률로 설명하면 광어는 맨 왼쪽 색의 크기 비율만큼
60%의 확률로 잡히고, 우럭은 30%, 붕어는 10% 확률로 물고기들이 잡힐 것이다.
곧, 연못의 면적은 물고기가 어느정도 쉽게 나오느냐에 대한 차이이다.
우럭이 잡힐 확률은 붕어의 6배, 우럭은 붕어의 3배 정도 잡기가 쉬울 것이다.
우리가 연못의 구성이 어떻게 되어 있는지는 몰라도 충분할 정도로 많이
반복해서 고기를 잡는다면 이 연못의 구성을 예측할 수 있을 것이다.
즉 현실에서 관측되는 데이터의 상대도수는 연못의 넓이에 그대로 반영
이러한 과정을 랜덤 샘플링(무작위 추출)의 가정이라고 한다.

- 관측을 충분히 많이 하면 모집단의 모습을 상당히 선명하게 파악할 수 있다.
- 하지만 우리의 목표는 많은 관측을 하지 않고도 모집단의 모습을 추측하는데 있다.
- 그래서 필요한 것은 많이 관측되지 않은 데이터로부터 모평균을 추측하는 방법이다.
  - 모집단의 평균을 모평균이라고 하고 μ 이라고 정의한다.
  - 모집단의 분산을 모분산이라고 하고 σ^2 이라고 정의한다.
- 평균만 가지고 모집단을 파악하기 힘들다.
  - 질문 : 데이터가 평균 주변에 어느 정도의 넓이로 퍼져 있는가?
  - 대답 : 앞서서 우리는 표준편차가 데이터가 평균에서 흩어져 있는 상태를 파악할수 있는 통계량이라고 공부하였다.

표본평균

- 우리가 알고 싶은 것은 불확실한 현상의 원천인 모집단
- 모집단이 어떤 수치로 채워져 있는가 알 수 있다면, 관측될 수치에 대해서 효과적으로 대비할 수 있기 때문
- 하지만 모집단 수치 전체의 분포 모습을 모두 정확하게 아는 것은 원칙적으로 불가능
- 몇 번 측정한 체온의 평균을 내거나, 하루하루의 매출액을 일주일 단위로 합하여 평균을 계산하기도 하는데 이렇게 관측된 데이터의 평균값은 모평균과 구별하기 위해서 표본평균($\bar{x}$)이라고 한다.

- 표본평균 = (관측된 데이터 합계) / (관측 데이터 총 개수)
- 표본평균을 구하는 이유?
  - 우연히 생긴 흩어진 데이터를 없애고 실제의 값에 가까운 값을 만들어 내고 싶기 때문에...

• 대수의 법칙

- 하나의 모집단에서 n개의 데이터를 관측하고 그 표본평균 $\bar{x}$를 만든다. 이 때 n이 크면 클수록 표본평균은 모평균 μ에 가까운 수치를 구할 가능성이 커진다.
- 모집단이 정규분포를 이루고 있는 것을 정규모집단이라고 부른다.
- 정규모집단은 표본평균을 만들어도 그 분포는 정규분포 그대로 유지한다는 성질을 가지고 있다.
  - 수학적으로 증명됨.
- 표본평균은 모집단의 평균값(모평균)을 추정하는데 이용된다. 그런 의미에서 표본평균은 모평균의 추정량이라고 한다.
- 표본평균을 모평균을 검정하는데 이용되므로 표본평균을 모평균의 검정통계량이라고 한다.

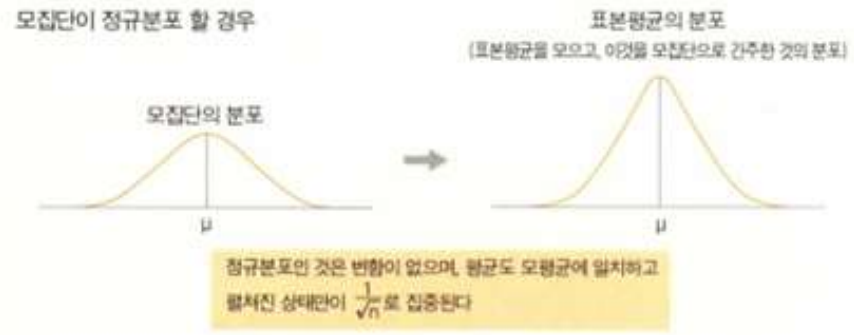
- 정규분포는 통계 분석에서 중요한 위치를 차지하고 있다. 모집단에서 반복적으로 표본을 추출해 각각 평균을 계산할 때, 이 평균들은 정규본포를 이룬다. 이것을 중심 극한 정리라 한다.
- 카지노의 크랩스(craps)라는 게임이 있다. 이 게임은 직육면체의 주사위 2 개를 사용하여 주사위 2개를 던져서 그 합을 계산하면 된다. 가장 작은 값은 (1, 1)의 값은 2이고 가장 큰 값은 (6, 6)의 값인 12이다. 아래는 그 예시이다.

2 – (1+1) – 1/36
3 – (1+2, 2+1) – 2/36
4 – (1+3, 3+1, 2+2) – 3/36
5 – (1+4, 4+1, 2+3, 3+2) – 4/36
6 – (1+6, 6+1, 2+4, 4+2, 3+3) – 5/36
7 – (1+6, 6+1, 2+5, 5+2, 3+4, 4+3) – 6/36
8 – (2+6, 6+2, 3+5, 5+3, 4+4) – 5/36
9 – (3+6, 6+3, 4+5, 5+4) – 4/36
10 – (4+6, 6+4, 5+5) – 3/36
11 – (5+6, 6+5) – 2/36
12 – (6+6) – 1/36

- 시물레이션한 결과를 보면 모집단이 정규분포 하고 있는 것을 정규모집단이라고 부르고, 이 모집단은 표본평균을 만들어도 그 분포는 정규분포 그대로 유지한다는 훌륭한 성질을 갖고 있다.

• 정규모집단에서 표본평균의 성질

- 정규모집단의 모평균을 μ, 모집단의 표준편차를 σ라고 할 때, 여기에서 관측된 데이터 x의 n개에 대한 표본평균 $\bar{x}$의 분포 역시 정규분포를 가진다. $\bar{x}$의 분포 평균값은 μ 그대로지만, 표준편차는 $\frac{\sigma}{\sqrt{n}}$가 되어서 모집단의 비해 $\frac{1}{\sqrt{n}}$로 줄어든다.

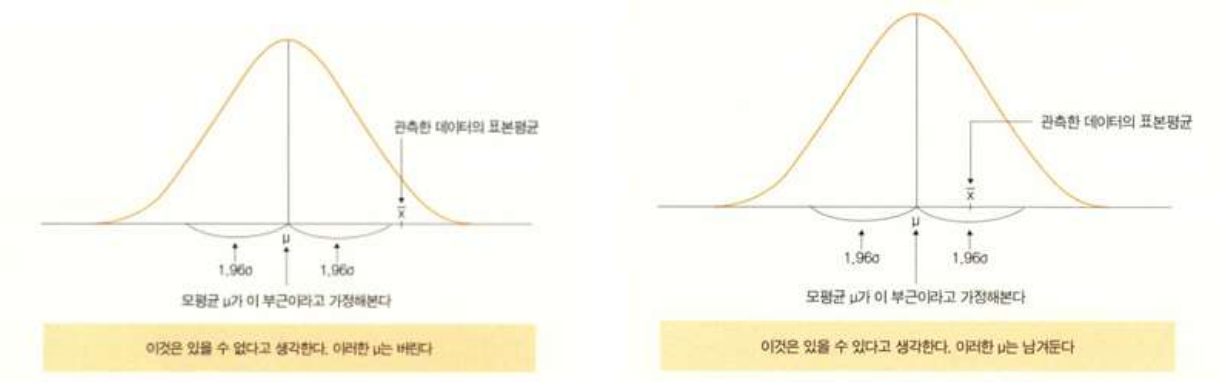


- ## 정규모집단에서 표본평균의 95% 예언적중구간
  - 모평균이 μ이고, 모집단의 표준편차가 σ인 정규분포에서 데이터 n개의 표본평균 $\bar{x}$에 대한 95% 예언적중구간은
  - $(\mu - 1.96\frac{\sigma}{\sqrt{n}})$이상 $(\mu + 1.96\frac{\sigma}{\sqrt{n}})$ 이하
  - $-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$
  - 모평균이 200, 표준편차가 10일때 관측된 데이터는 180.4이상 219.6이하의 범위로 들어간다라고 하면 95%의 확률로 맞는다.
  - 아래 예제는 모집단에서 n개씩 관측하고 표본평균값이 들어가는 범위 형태(95%)를 예측한 범위이다.

```
for n in range(4,100,8):
    upper = 200 + 1.96 * (10 / np.sqrt(n))
    lower = 200 - 1.96 * (10 / np.sqrt(n))
    print ("{} --> {} <= x <= {}".format(n, lower,upper))
```

```
4 --> 190.2 <= x <= 209.8
12 --> 194.34196736194167 <= x <= 205.65803263805833
20 --> 195.61730676410042 <= x <= 204.38269323589958
28 --> 196.29594816450958 <= x <= 203.70405183549042
36 --> 196.73333333333332 <= x <= 203.26666666666668
44 --> 197.04518882313792 <= x <= 202.95481117686208
52 --> 197.28196903849638 <= x <= 202.71803096150362
60 --> 197.46965088047781 <= x <= 202.53034911952219
68 --> 197.62315087464393 <= x <= 202.37684912535607
76 --> 197.7517258080685 <= x <= 202.2482741919315
84 --> 197.86146467568727 <= x <= 202.13853532431273
92 --> 197.95655874224067 <= x <= 202.04344125775933
```

  - 표본평균을 만드는 개수가 늘어날수록 예언하는 구간이 좁아진다.



$-1.96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$를 응용하여 수식을 정리하면 $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

- 문제 1

  편의점에서 판매하는 삼각김밥을 자동으로 만드는 기계가 있다. 이 기계는 삼각김밥의 무게를 다양하게 조절할 수 있지만, 무게에는 오차가 발생한다. 완성된 삼각김밥 무게의 모든 데이터를 모집단이라고 할 때, 그것은 정규모집단이고, 모집단의 편차가 10그램이라는 것을 알고 있다. 여기에서 25개의 삼각김밥을 만들면 그 표본평균은 80그램이었다. 제조된 삼각김밥의 무게 모평균을 95% 신뢰구간을 구간추정해 보자