통계학 입문

2021-04-25

데이터과학융합스쿨

손 대 순 교수

여러가지 확률분포에 대한 이야기를 하고 있습니다. 이산형 확률변수에서는 이항분포와 포아송 분포를 다루었고, 연속형에서는 정규분포 이야기를 했습니다. 다른 분포도 많이 있지만, 분포의 종류를 많이 아는 것보다, 표현방법, 확률밀도함수, 확률계산방법 등의 개념이 더욱 중요해서 이 부분을 강조하고자 했어요. 복잡한 분포들이 많이 나오면 훨씬 혼동스러울 것 같아서, 이 정도로만 합니다. ^^

정규분포표에서 찾는 확률은 R에서 pnorm함수를 가지고 찾을 수도 있답니다. ^^

```
> pnorm(0.4)
> pnorm(2.4)
> pnorm(-0.4)
```

이런 명령어도 입력해 보면서, 표와 비교해 보고 이들의 관계를 잘 정리해 보세요.

이번 시간에는 통계학에서 가장 중요하다고 해도 될 정리에 대해 소개하고자 합니다. '중심극한정리(Central Limit Theorem)'라고 하는 것인데요. 이름은 처음 들을지 몰라도 이미 고등학교에서 경험한 내용이기는 합니다.

> 중심극한정리:
> 평균이 μ이고 분산이 σ^2인 모집단으로부터 추출한 크기가 n인 확률표본의 표본평균 $\overline{X}$는 n이 증가할수록 모집단의 분포유형에 상관없이 근사적으로 정규분포 $N(\mu, \sigma^2/n)$을 따른다.

중심극한정리는 위의 내용입니다. 별 것 없죠. 이에 대해 명확히 이해하면 되니 큰 부담은 없으리라 생각하지만 워낙 중요한 정리이니, 차근히 설명하겠습니다. 다음의 사항을 중요하게 보아야

합니다.

1) 평균이 μ이고 분산이 σ^2인 모집단으로부터

2) n이 증가할수록

3) 표본평균은 근사적으로 정규분포를 따른다.

4) 평균은 그대로이고, 분산은 n으로 나누어준 값을 가진다.

이상의 4가지에 초점을 두어 설명합니다.

1) 평균이 μ이고 분산이 σ^2인 모집단으로부터

모집단에도 분명 평균과 분산이 있겠지요. 이 부분에서 말하고자 하는 것은 분포를 특정하지 않았다는 것이 중요합니다. 위의 설명에서 '분포유형에 상관없이'에 해당하는 것이 강조되는 포인트 입니다. 모집단이 어떤 분포이든, 아주 이상한 분포를 가졌다 하더라도 이 중심극한정리가 성립한다라는 것이죠. 때문에 아주 강력한 이론입니다.

2) n이 증가할수록

중심극한정리는 극한의 개념이지요. 중심극한정리를 수식으로 한 번 볼까요? 슬쩍 아래 링크에 들어갔다가 와 보세요.

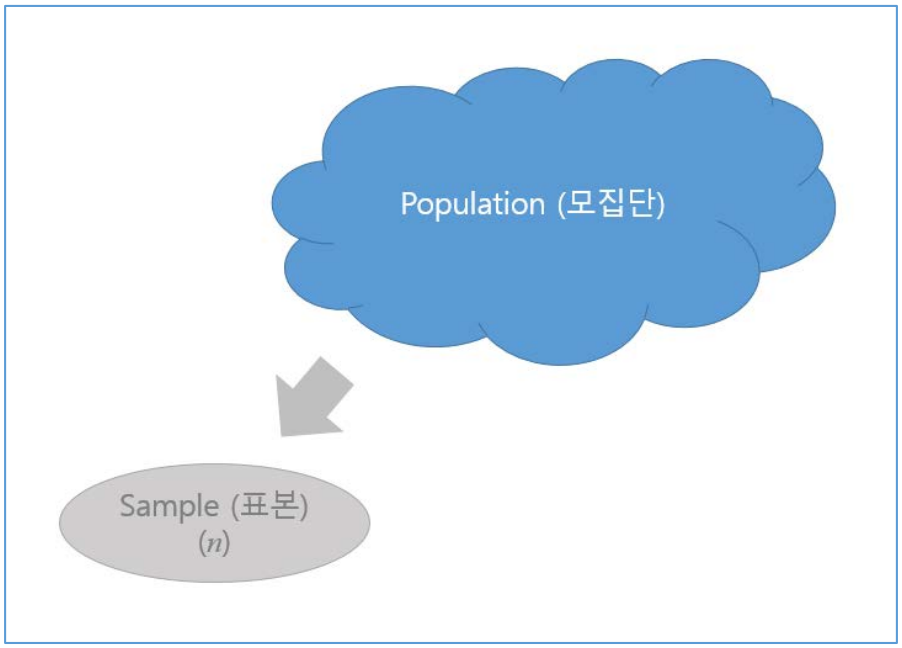
https://en.wikipedia.org/wiki/Central_limit_theorem

놀랐죠? 간단하게 여러분들에게 설명하려 하지만, 이를 수리적으로 증명하는 과정은 입문수준을 벗어나므로, 넘어가도록 하겠습니다. 다만, 극한의 개념을 사용하고 있다는 것은 볼 수 있습니다. 그렇죠? n이 무한대로 갈 때 성립하는 정리입니다.

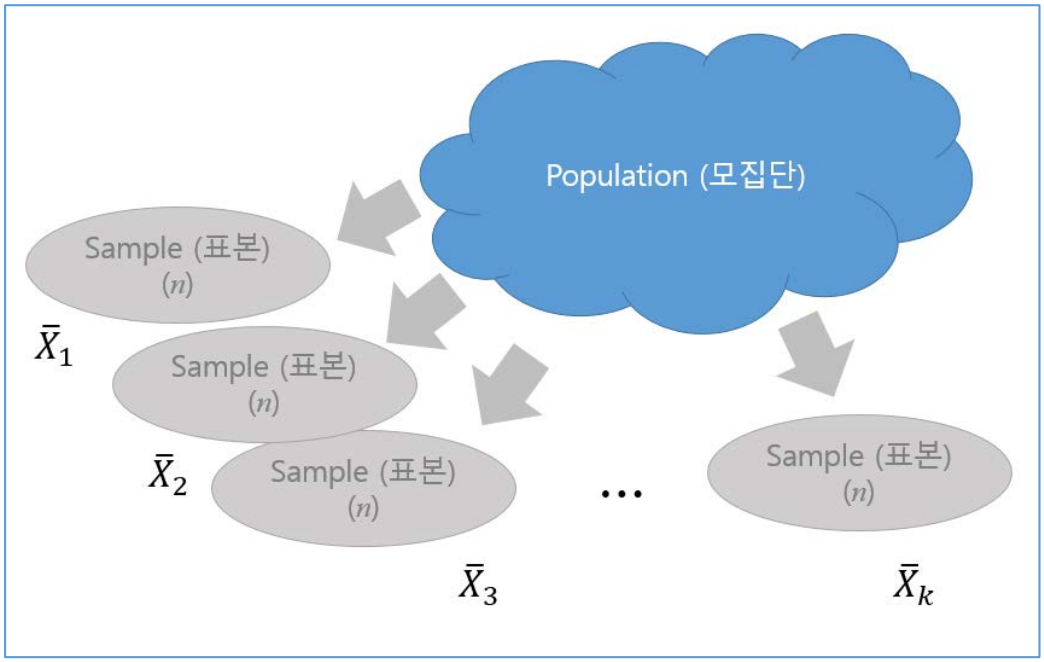
3) 표본평균은 근사적으로 정규분포를 따른다.

결국 모집단의 분포에 상관없이 "표본평균"은 정규분포를 따른다. 이 말이 중심극한정리입니다. 무엇이? 표본평균이!! 정규분포를 따른다.

표본 평균이 정규분포를 따른다는 말이 무슨 뜻일까요?



모집단에서 표본을 뽑습니다. 표본의 크기가 n개라고 하죠. 이 표본의 평균을 우리는 표본평균으로 정의하고 $\bar{X}$라고 표시하기로 했습니다. 이러한 표본을 여러 개 뽑아 보죠.



크가가 n인 표본을 여러 개 만들 수 있을 것이고, 이 표본에는 각각 표본 평균들이 있겠죠. 이 표본평균들을 모아서 분포를 그려보면 정규분포한다는 말입니다. 그런데 조건이 있죠? 표본 수, n이 커야 한다는 것이죠. n이 2, 3 같이 작은 수면 안된다는 것입니다. 충분히 n이 커야 근사적으로 정규분포한다는 뜻이죠. 그렇다면 얼마나 커야 충분히 큰 것인가? 하는 질문이 생기죠? 사실 이것은 절대적인 기준이 없습니다. 크면 클수록 정규분포한다는 성질이니, 어느 한 점에서 끊어낼

수는 없는 노릇이죠. 통상적으로 30 이상이면 정규분포에서 확률 계산 하는데 무리가 없다고 합니다만, 절대적인 기준은 아닙니다.

잘 생각해 봅시다. 표본의 크기가 충분히 크면, 표본평균이 모집단의 평균에 잘 접근하지 않겠어요? 오차가 있어도, 그 주변으로 조금 나겠죠. 표본의 크기가 아주 작다면, 모집단의 평균에서 벗어나는 경우가 많을 것 같지 않아요? 어떻게 보면 상식적인 것 같기도 해요.

4) 평균은 그대로이고, 분산은 n으로 나누어준 값을 가진다.

그래서, 표본평균의 평균은 모집단의 평균 그대로 μ이고 분산이 σ^2/n이 됩니다. 표본이 크기가 커지면, 모집단의 평균과 표본평균의 평균이 모집단의 평균과 같아진다는 것은 직관적으로 이해할 수 있지요? 그럼 분산을 봅시다. 분산은 퍼진 정도를 측정하는 지표입니다. n이 크면 클수록 표본평균은 모집단의 평균에 더 정밀하게 접근할 것입니다. 흔들림이 없겠죠. 심지어 표본의 크기가 무한대라고 해 봅시다. 이는 결국 모집단 전체를 표본으로 뽑는다는 의미라고 보면 됩니다. 모집단 전체가 표본이니, 반복해서 표본을 뽑는다 하여도, 표본의 평균이 정확하게 모집단의 평균과 일치할 것입니다. 즉 분산이 0이 된다는 뜻이예요. 그러니 n이 무한대로 간다면 표본평균의 분산, σ^2/n은 0으로 수렴하겠지요.

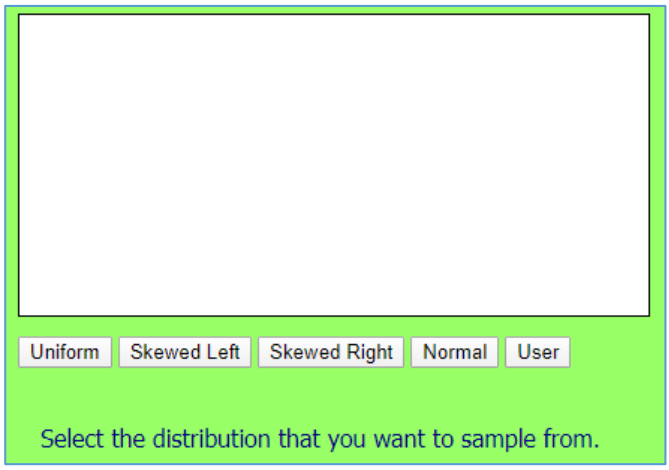
그렇다면, 이 어렵지 않은 정리가 왜 그렇게 중요할까요?

통계학에서 많은 경우 정규분포 가정을 합니다. 그런데, 모집단이 어떤 분포를 가지는지는 쉽게 알 수가 없지요. 그래서, 평균을 주로 많이 사용합니다. 표본의 평균은 정규분포하는 성질을 가지고 있으므로, 정규분포가정을 할 수가 있게 되고, 이를 이용하여 확률을 계산할 수가 있게 되는 것이지요. 많은 이론들이 이를 토대로 하고 있으므로 많은 통계적 방법들의 이론적 토대가 되는 정리라고 할 수 있어요.

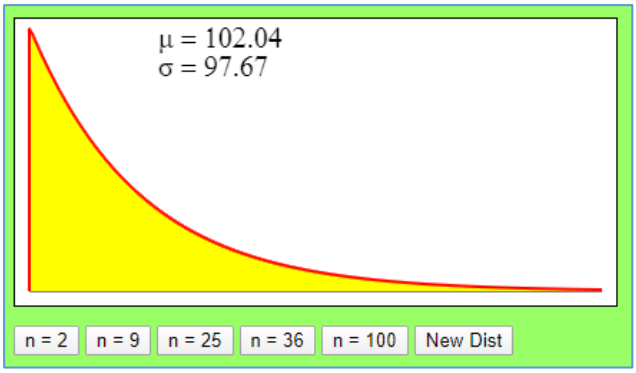
중심극한정리를 증명하는 방법은 수식을 이용하는 방법과 시뮬레이션을 이용하는 방법이 있습니다. 수식을 이용하는 방법은 적률생성함수라는 것을 이용하는데, 본 강의의 수준을 넘어서는 것이니, 여기에서는 하지 않을게요. 다만, 시뮬레이션은 해 볼 수 있을 것 같아요.

http://www.ltcconline.net/greenl/java/Statistics/clt/cltsimulation.html

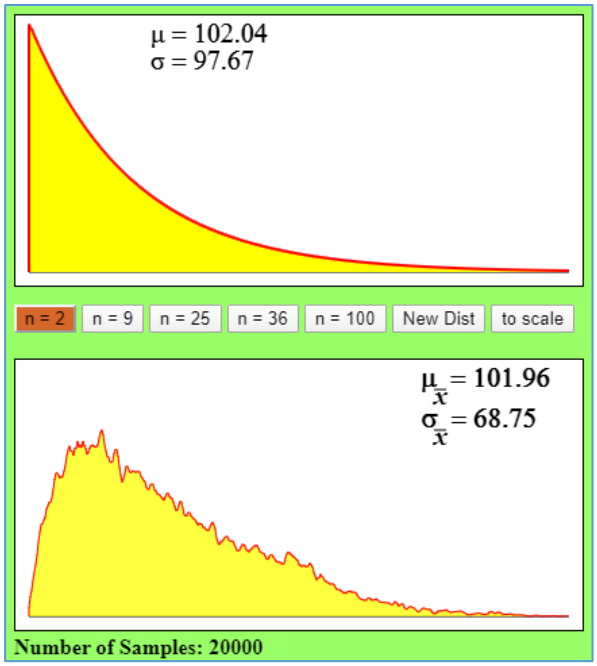
위의 링크에서 시뮬레이션을 제공하고 있습니다. 여러분들이 한 번 해 보면 이해가 쉬울 것 같습니다. 한 번 따라서 해 볼까요?



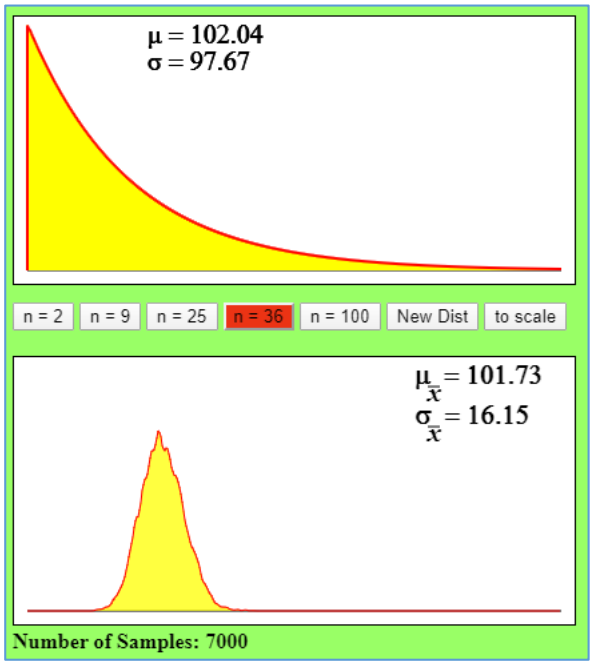
이 창은 모집단의 분포를 정하는 창입니다. Right skewed distribution을 한 번 선택해 보죠.



선택해 보면, 평균이 102.4이고 표준편차가 97.67인, 오른쪽으로 꼬리가 긴 분포가 나옵니다. 그리고, 아래에 버튼이 생기죠. 표본의 크기가 2인 것부터 100인 경우까지 선택할 수 있도록 나오네요. 우선 표본의 크기를 2로 선택해 봅시다.

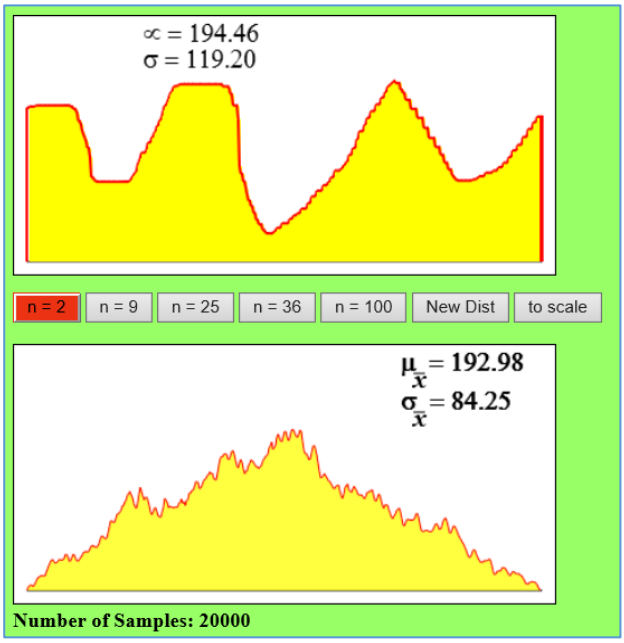


아래 적혀진 것처럼, 표본의 크기가 2인 샘플을 20000번 표본을 뽑은 것이예요. 표본 평균이 20000개 있는 것이고, 이를 그림으로 표현한 것입니다. 어때요? 원래의 그림처럼 right skewed distribution을 가지고 있죠? 그렇다면 표본의 크기를 36으로 키워 봅시다.

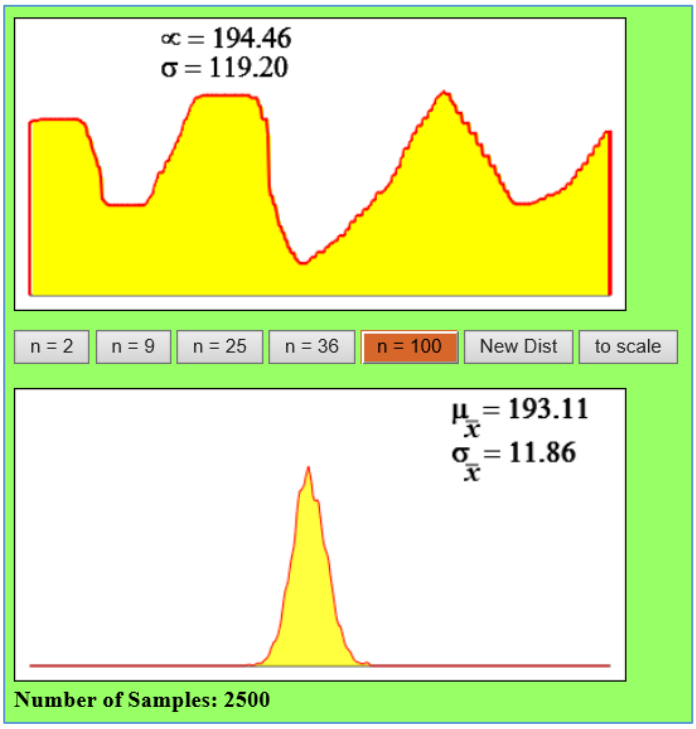


어떤가요? 많이 정규분포에 가까워졌죠? 이렇게 모집단이 다른 분포를 가지고 있어도 표본평균의 분포가 정규분포한다는 것이 중심극한정리입니다. 여러분이 분포를 만들어서 넣을 수도 있게 해 놓았네요. 다른 분포도 직접 해 보세요.

분포 선택에서 'User'를 클릭해서 여러분이 분포를 만들어서 넣어 보세요. 아래 그림처럼 아주 이상한 모집단의 분포를 그려 넣었습니다.



n=2일때는 표본평균의 분포가 위와 같군요. 표본의 크기를 늘려보면, 점점 정규분포에 가까워지고 있음을 확인할 수 있습니다.



이해 되죠?

여러분이 고등학교 때, '표준화'라고하는 식을 사용한 적이 있습니다.

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

이 식인데요. 이는 데이터를 의미하는 X를 표준화 하는 것이었죠. 그래서 X의 표준편차인 σ로 나누어 준 것입니다. 지난 시간에 확률계산을 위해 정규분포표를 이용할 때 사용했던 방법이죠. 만일 $\bar{X}$라면 어떻게 해야 할까요?

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

을 사용해야 합니다. 중심극한정리에 의해서 $\bar{X}$의 표준편차가 $\sigma/\sqrt{n}$이기 때문이죠. 이는 고등학교 통계문제에 단골로 등장합니다. 한 번 볼까요?

평균이 71.5이고 분산이 25인 정규분포를 따르는 모집단으로부터 표본의 크기가 100인 확률표본 $X_1, X_2, \ldots, X_{100}$ 을 추출하였다.

<표>는 Z가 표준정규분포를 따르는 확률변수일 때 $P(Z \le z)$를 나타낸 것이다. 표본평균 $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$ 가 72.5보다 클 확률은?

z	$P(Z \le z)$	z	$P(Z \le z)$	z	$P(Z \le z)$
0.2	0.5793	1.2	0.8849	2.2	0.9861
0.4	0.6554	1.4	0.9192	2.4	0.9918
0.6	0.7257	1.6	0.9452	2.6	0.9953
0.8	0.7881	1.8	0.9641	2.8	0.9974
1.0	0.8413	2.0	0.9772	3.0	0.9987

① 0.4207 ② 0.2743 ③ 0.1151

④ 0.0228 ⑤ 0.0139

위의 문제는 다음의 과정으로 확률을 계산할 수 있습니다.

$$X \sim N(71.5, 5^2)$$

중심극한정리에 의해

$$\bar{X} \sim N(71.5\ , 5^2/100)$$

그러므로,

$$P(\bar{X} > 72.5) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{72.5-71.5}{5/10}\right) = P(Z > 2) = 1 - P(Z \le 2) = 0.0228$$

이해할 수 있죠? 위의 문제에서는 모집단의 분포가 정규분포라고 주어졌지만, 그렇지않다 하더라도 문제에는 지장이 없습니다.

중심극한정리 (Central Limit Theorem, CLT)가 어떻게 사용되는지 이해할 수 있겠죠?

다음시간에는 이산형 분포의 대표 주자인 이항분포를 이용하여 연속형 근사 방법을 다루어 보겠습니다.

오늘도 수고 많았어요.