중심극한정리

중심극한정리

- 표본평균 $\bar{x}$ 분포는 어떤 분포를 따를까요?
  - 앞서 표본평균 $\bar{x}$ 분포의 중요한 특성인 기댓값과 분산(표준편차)에 대해 알아봤습니다.
  - 그렇다면 표본평균 $\bar{x}$ 분포는 어떤 모양이 될지 알아봅시다.
  - 이 과정은 상급과정에서 수리적으로 복잡한 계산을 통해 증명하지만, 우리는 R을 통해 그래프를 그려가면서 어떤 분포와 닮아가는지 확인해 봅시다.
    - 먼저 아주 특수한 경우로 모집단이 정규분포를 따를 때 이로부터 추출한 표본들의 표본평균 $\bar{x}$ 분포가 어떤 분포를 따를지 살펴봅니다.
    - 그 다음으로 좀 더 일반적인 상황으로 모집단의 분포가 임의의 분포일 때 어떤 분포를 따를지 살펴보겠습니다.

중심극한정리

- 모집단이 정규분포일 때
  - 모집단이 정규분포일 때 이로부터 추출된 표본들의 표본평균의 분포는 어떤 모양을 따를지 살펴보겠습니다.
  - 서로 다른 두 정규분포에서 4개의 표본으로부터 평균을 구하는 것을 1,000번 실시하여 표본평균의 분포가 어떤 형태를 따르는지 확인해봅니다
- Step#1) 준비과정

```
1. set.seed(9)
2. n <- 1000
3. r.1.mean <- rep(NA, n)
4. r.2.mean <- rep(NA, n)
```

중심극한정리

- 1줄 : 난수생성의 초깃값을 9로 고정합니다.
- 2줄 : 표본추출 횟수 1,000을 변수 n에 저장합니다.
- 3, 4줄 : 모집단별로 표본평균이 저장될 두 변수 r.1.mean과 r.2.mean을 결측값(NA)으로 초기화합니다.

- Step #2) 두 정규분포 $N(3, 1^2)$과 $N(170, 6^2)$ 으로부터 표본 크기가 4인 표본을 1,000번 추출하고, 각 추출마다 평균을 저장합니다.

```
5. for (i in 1:n ) {
6.   r.1.mean[i] <- mean( rnorm(4, mean=3, sd=1) )
7.   r.2.mean[i] <- mean( rnorm(4, mean=170, sd=6) )
8. }
```

중심극한정리

- 5, 8줄 : 1:1000으로 생성되는 벡터의 원소 수만큼 반복문을 만듭니다. 이 반복문으로 인해 6, 7번째 줄을 1,000번 반복합니다.
- 6줄 : $N(3, 1^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.1.mean의 i번째 원소에 저장합니다.
- 7줄 : $N(170, 6^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.2.mean의 i번째 원소에 저장합니다.

- Step #3) 표본평균들의 분포에서 평균과 표준편차를 구합니다.

```
10.options(digits=4)
11.c(mean(r.1.mean), sd(r.1.mean))
12.c(mean(r.2.mean), sd(r.2.mean))
```

- 10줄 : 출력물의 자릿수를 4로 합니다.
- 11, 12줄 : 각 정규분포로부터 추출된 표본 크기가 4인 표본평균 분포의 평균과 표준편차를 출력합니다.

중심극한정리

- 표준정규분포로부터 추출한 표본평균의 분포는 그 평균이 모집단 평균에 가깝고, 표준편차는 모집단 정규분포의 표준편차를 표본 크기의 제곱근으로 나눈 값($\frac{\sigma}{\sqrt{4}}$, 모집단 표준편차의 반) 과 비슷합니다.

- Step #4) 표본평균의 분포에 대한 히스토그램을 그리고, 그 위에 각 표본평균의 분포가 따를 것으로 생각되는 분포의 확률도표를 그려봅니다.
  - 모집단이 정규분포일 때 이로부터 추출한 표본평균의 분포는 또 다른 정규분포를 따르는 것으로 알려져 있습니다.
    - 정규분포로 부터 추출된 경우 알려진 표본평균의 분포 : $\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$

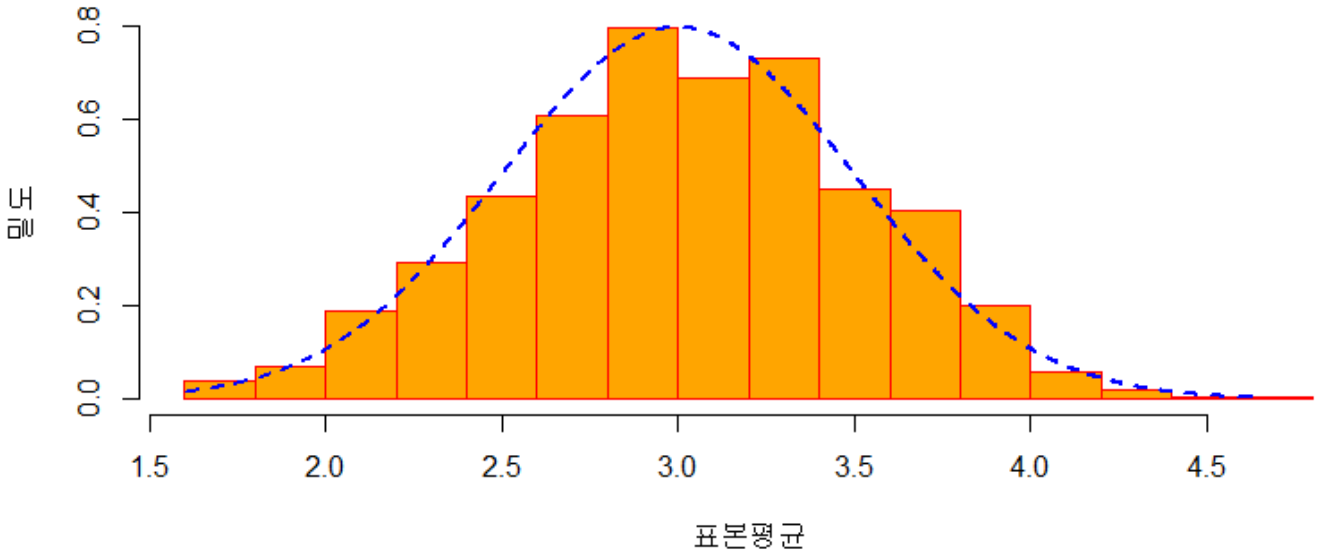
```
> c(mean(r.1.mean), sd(r.1.mean))
[1] 3.0214 0.5096
> c(mean(r.2.mean), sd(r.2.mean))
[1] 170.032   2.835
```

중심극한정리

```
14.hist(r.1.mean, prob=TRUE, xlab="표본평균", ylab="밀도", main="",
        col="orange", border="red")
15.x1 <- seq(min(r.1.mean), max(r.1.mean), length=1000)
16.y1 <- dnorm(x=x1, mean=3, sd=(1/sqrt(4)))
17.lines(x1, y1, lty=2, lwd=2, col="blue")


18.hist(r.2.mean, prob=TRUE, xlab="표본평균", ylab="밀도", main="",
        col="orange", border="red")
19.x2 <- seq(min(r.2.mean), max(r.2.mean), length=1000)
20.y2 <- dnorm( x=x2, mean=170, sd=(6/sqrt(4)) )
21.lines(x2, y2, lty=2, lwd=2, col="blue")
```

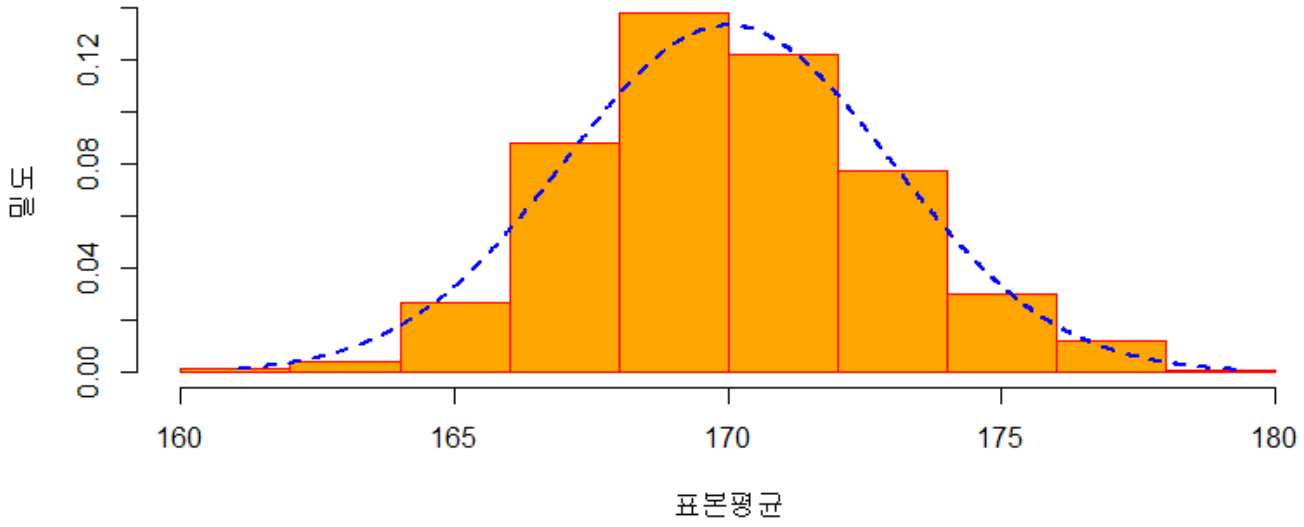
중심극한정리

- 14~17줄
  - $N(3, 1^2)$ 으로부터 표본 크기를 4로 하는 표본평균의 분포에서 평균은 모집단의 평균인 3이고, 표준편차는 $\frac{1}{\sqrt{4}}$입니다.
  - 표본평균의 히스토그램과 평균이 3이고 표준편차가 $\frac{1}{\sqrt{4}}$ 인 정규분포와 비교해 봅시다.
  - hist() 함수에서 prob으로 TRUE을 전달하면, 빈도에 대한 히스토그램이 아닌 상대빈도에 대한 히스토그램을 작성합니다.



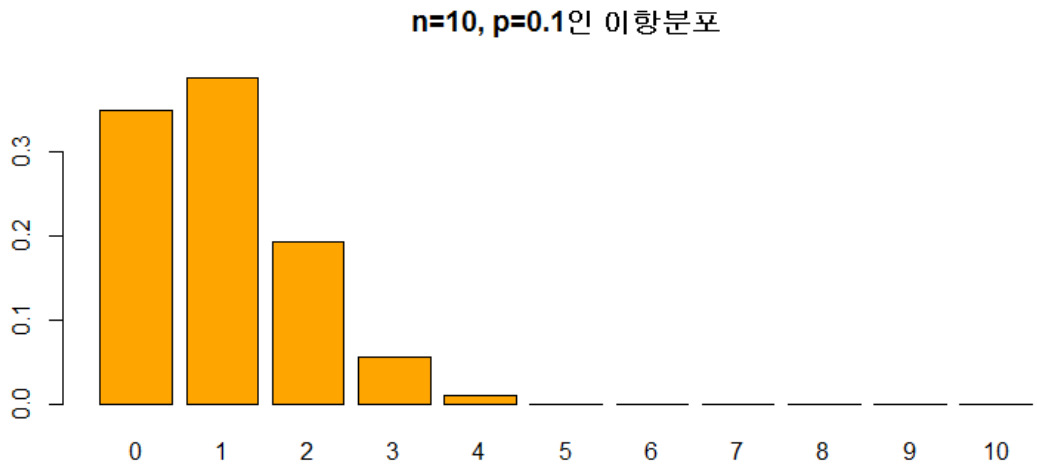
중심극한정리

- 19~22줄
  - $N(170, 6^2)$ 으로부터 표본 크기를 4로 하는 표본평균의 분포에서 평균은 모집단의 평균인 170이고, 표준편차는 $\frac{6}{\sqrt{4}}$입니다.
  - 표본평균의 히스토그램과 평균이 170이고 표준편차가 $\frac{6}{\sqrt{4}}$ 인 정규분포와 비교해 봅시다.



중심극한정리

- 모집단이 정규분포가 아닌 임의의 분포일 때
  - 조건 : 모집단의 평균과 표준편차가 (그 값을 알지 못하나) 존재합니다.
  - 예) 모집단이 시행의 횟수가 10이고 성공의 확률이 0.1인 이항분포($B(10, 0.1)$)
    - $B(10, 0.1)$ 은 꼬리가 오른쪽으로 늘어진 모양을 갖습니다.
    - 기댓값 : $np = 10 \times 0.1 = 1$
    - 표준편차 : $\sqrt{np(1-p)} = \sqrt{10 \times 0.1 \times 0.9} \approx 0.9487$
    - 표본의 크기가 2, 4, 32로 늘려가면서 표본평균의 분포를 관찰합니다.



중심극한정리

- Step #1) 자료 준비

```
7. set.seed(9)
8. t <- 10
9. p <- 0.1
10.x <- 0:10
11.n <- 1000
12.b.2.mean <- rep(NA, n)
13.b.4.mean <- rep(NA, n)
14.b.32.mean <- rep(NA, n)
```

  - 7~10줄 : 난수생성의 초기값을 9로, 시행 횟수 10을 변수 t에, 성공 확률 0.1을 변수 p에 저장하고 시행횟수가 10인 이항분포로부터 관찰 가능한 값을 변수 x에 저장합니다.
  - 11줄 : 표본을 추출할 횟수 1,000을 변수 n에 저장합니다.
  - 12~14줄 : 표본 크기에 따라 1,000번의 표본추출에서 관찰된 표본평균이 저장될 변수 b.2.mean, b.4.mean과 b.32.mean에 대해 각각 1,000개의 NA 값을 갖는 벡터로 준비합니다.

중심극한정리

- Step #2) 표본 크기별로 1000번의 표본추출로 표본평균을 구합니다.

```
16.for(i in 1:n) {
17.   b.2.mean[i] <- mean( rbinom(2, size=t, prob=p) )
18.   b.4.mean[i] <- mean( rbinom(4, size=t, prob=p) )
19.   b.32.mean[i] <- mean( rbinom(32, size=t, prob=p) )
20.}
```

- 16, 20줄 : 17~19줄을 1000번 반복(1000번의 표본추출)하는 반복문입니다.
- 17줄 : $B(10, 0.1)$로부터 2개의 표본을 추출하고,
  그 평균을 b.2.mean의 i번째 원소에 저장합니다.
- 18줄 : $B(10, 0.1)$로부터 4개의 표본을 추출하고,
  그 평균을 b.4.mean의 i번째 원소에 저장합니다.
- 19줄 : $B(10, 0.1)$로부터 32개의 표본을 추출하고,
  그 평균을 b.32.mean의 i번째 원소에 저장합니다.

중심극한정리

- Step #3) 표본평균들의 분포에서 평균과 표준편차를 구합니다

```
22.options(digits=4)
23.c(mean(b.2.mean), sd(b.2.mean))
24.c(mean(b.4.mean), sd(b.4.mean))
25.c(mean(b.32.mean), sd(b.32.mean))
```

  - 22줄 : 출력물의 자릿수를 4로 합니다.
  - 23~25줄 : $B(10, 0.1)$ 로부터 1,000번 추출된 표본 크기가 2, 4, 32인 표본평균 분포의 평균과 표준편차를 출력합니다.
    - 출력물에서 이항분포로부터 추출한 표본평균의 분포는 그 평균이 이항분포의 평균과 비교해 봅시다.
    - 표준편차는 모집단 이항분포의 표준편차를 표본 크기의 제곱근으로 나눈 값들과 비교해 봅시다.
      - $\frac{0.9487}{\sqrt{2}} \approx 0.6708, \frac{0.9487}{\sqrt{4}} \approx 0.4743, \frac{0.9487}{\sqrt{32}} \approx 0.1677$

- Step #4) 각 표본평균 분포의 히스토그램을 그리고, 그 위에 각 표본평균의 분포가 따를 것으로 알려진 정규분포의 확률도표를 작성합니다.
  - 앞서 모집단이 정규분포일 경우와 마찬가지로 정규분포를 따를 것으로 생각해봅시다. $\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$
    - n=2 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{2}} \approx 0.6708\right)^2)$, n=4 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{4}} \approx 0.4743\right)^2)$, n=32 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{32}} \approx 0.1677\right)^2)$

```
> c(mean(b.2.mean), sd(b.2.mean))
[1] 1.0090 0.6763
> c(mean(b.4.mean), sd(b.4.mean))
[1] 1.006 0.481
> c(mean(b.32.mean), sd(b.32.mean))
[1] 0.9989 0.1624
```

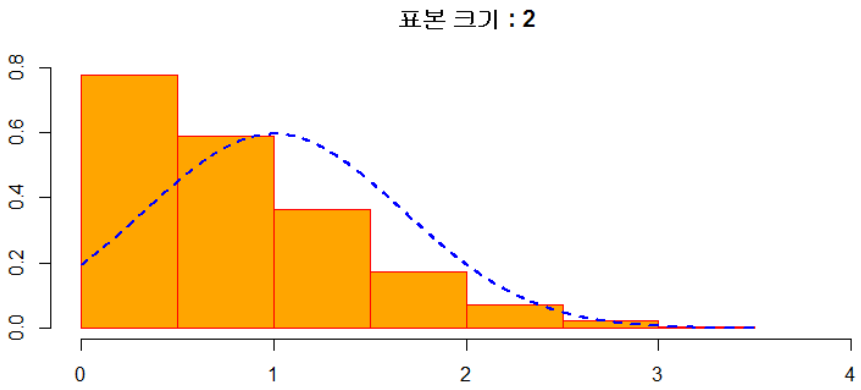
중심극한정리

```
27.hist(b.2.mean, prob=T, xlim=c(0, 4), main="표본 크기 : 2",
                ylab="", xlab="", col="orange", border="red")
28.x1 <- seq(min(b.2.mean), max(b.2.mean), length=1000)
29.y1 <- dnorm( x=x1, mean=1, sd=sqrt(0.9)/sqrt(2) )
30.lines(x1, y1, lty=2, lwd=2, col="blue")


31.hist(b.4.mean, prob=T, xlim=c(0, 4), ylim=c(0, 1.2),
   main="표본 크기 : 4", ylab="", xlab="", col="orange", border="red")
32.x2 <- seq(min(b.4.mean), max(b.4.mean), length=1000)
33.y2 <- dnorm( x=x2, mean=1, sd=sqrt(0.9)/sqrt(8) )
34.lines(x2, y2, lty=2, lwd=2, col="blue")


35.hist(b.32.mean, prob=T, xlim=c(0, 4), main="표본 크기 : 32",
   ylab="", xlab="", col="orange", border="red")
36.x3 <- seq(min(b.32.mean), max(b.32.mean), length=1000)
37.y3 <- dnorm( x=x3, mean=1, sd=sqrt(0.9)/sqrt(32) )
38.lines(x3, y3, lty=2, lwd=2, col="blue")
```

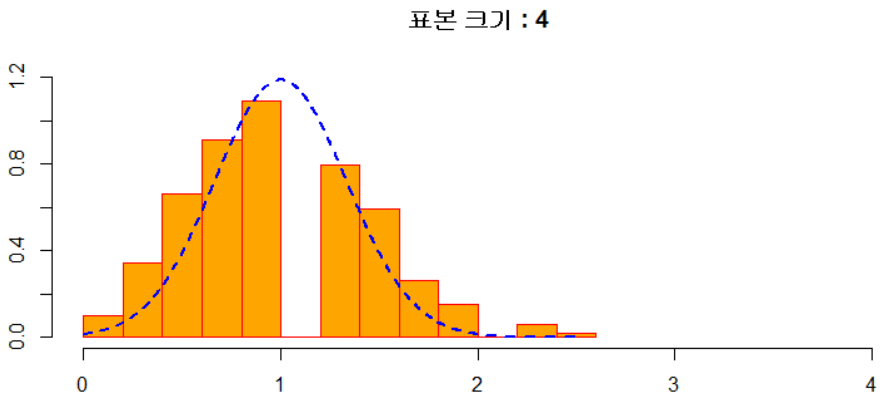
중심극한정리

- 27~30줄 : $B(10, 0.1)$ 로부터 표본 크기를 2로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{2}} \approx 0.6708$을 가집니다.
  - 평균이 1이고 표준편차가 약 0.6708인 정규분포와 비교해봅시다
- 32~35줄 : $B(10, 0.1)$ 로부터 표본 크기를 4로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{4}} \approx 0.4743$ 을 가집니다.
  - 평균이 1이고 표준편차가 약 0.4743인 정규분포와 비교해봅시다
- 37~40줄 : $B(10, 0.1)$ 로부터 표본 크기를 32로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{32}} \approx 0.1677$을 가집니다.
  - 평균이 1이고 표준편차가 약 0.1677인 정규분포와 비교해봅시다.

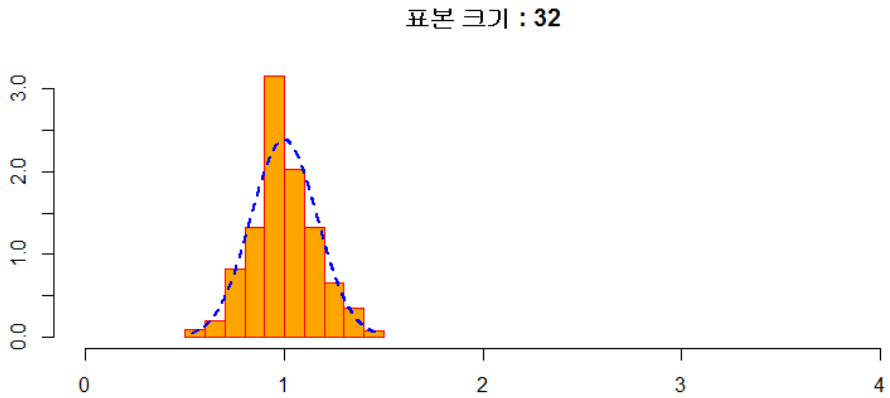
중심극한정리



여전히 표본평균의 분포는
오른쪽으로 늘어져 있습니다.



n=2 일 때 보다는 비교적 좌우대칭으로
보이지만, 중간중간 빈 구간이 보입니다.



n 이 증가할 수록 좌우대칭을 보이고,
점점 정규분포와 닮아갑니다.

중심극한정리

- 중심극한정리
  - 표본의 개수가 증가할수록 표본평균의 분포가 정규분포와 닮아감을 확인해 보았습니다.
  - 이와 같은 성질을 수리적으로 밝혀낸 것이 중심극한정리입니다.
    - 모집단의 분포와 상관없이 평균과 표준편차가 μ와 σ로 존재하는 모집단에서 추출할 때 표본의 크기 n이 충분히 크면, 표본평균의 분포가 근사적으로 정규분포를 따릅니다.
  - 중심극한정리는 모집단의 분포에 대한 사전 지식 없이도 표본평균의 분포를 알 수 있게 하여 통계학에서 유용하게 사용합니다.

중심극한정리

참고 **중심극한정리(CLT, Central Limit Theorem)**

모집단의 분포와 상관없이 모집단의 평균 μ와 표준편차 σ가 존재할 때 표본 크기 n이 충분히 크다면, 표본평균의 분포는 다음과 같이 근사적으로 정규분포를 따릅니다.

$$\overline{X} \simeq N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) \tag{4.4}$$

또한, 표본평균의 분포가 정규분포를 따르므로 다음과 같이 표준화하여 사용할 수 있습니다.

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1^2) \tag{4.5}$$