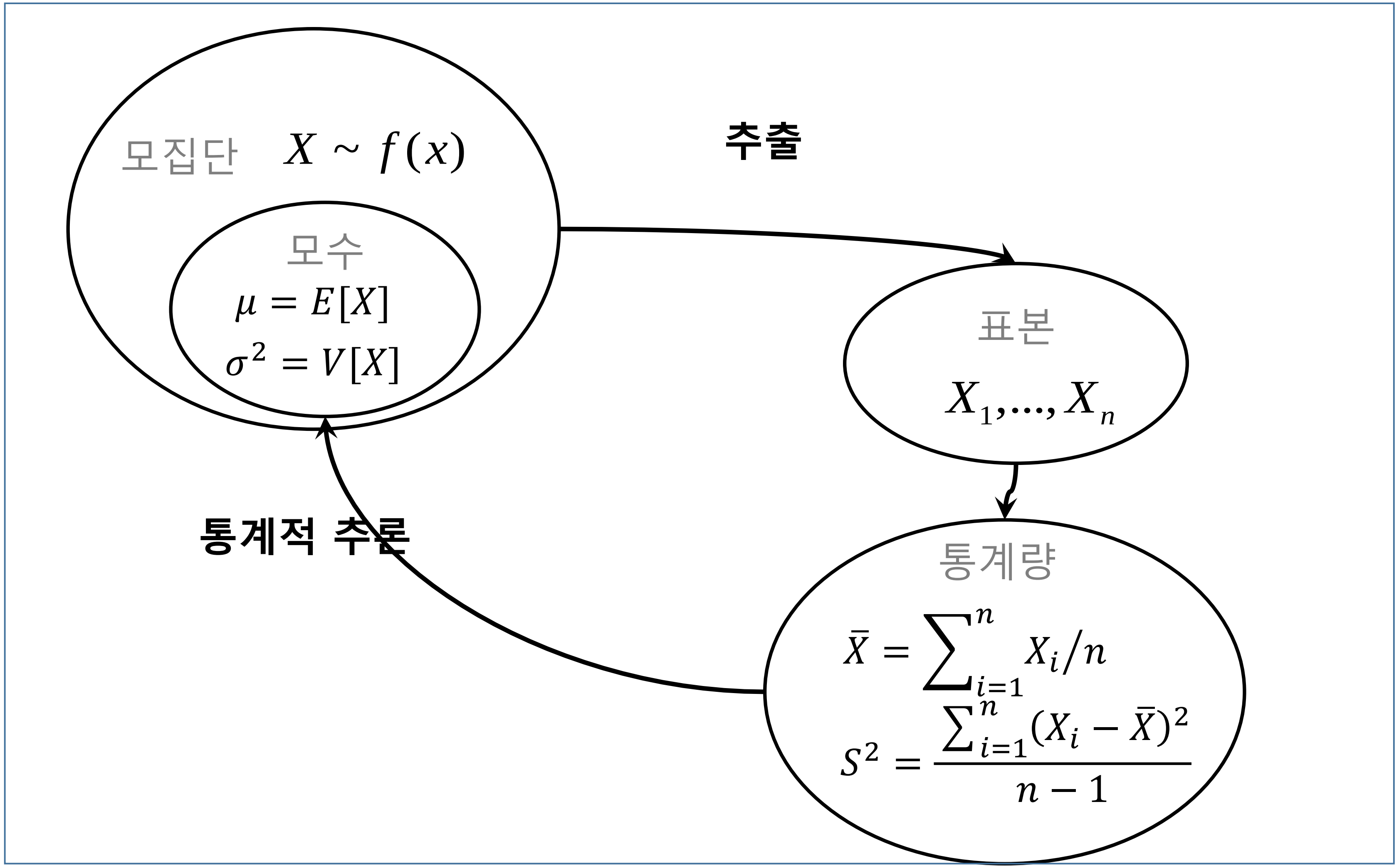
5. 확률표본과 표본분포

확률표본과 표본분포

- **통계적 추론 관련 기본 개념**



확률표본과 표본분포

- **모집단의 분포와 확률표본**

미지인 모집단의 확률변수를 X로 그 분포를 $f(x)$로 나타낸다고 할 때,

- **확률표본** (**iid** 표본, Random Sample)

  모집단 $f(x)$로부터의 확률표본 $X_1, X_2, \ldots, X_n$ 은 다음의 두가지 성질을 만족하는 모집단 $f(x)$ 로부터의 표본을 뜻함.

  - $X_1, X_2, \ldots, X_n$은 서로 **독립**임
  - $X_1, X_2, \ldots, X_n$는 모두 **동일**하게 $f(x)$의 분포를 따름

확률표본과 표본분포

- **통계량과 표본분포**
  - **통계량** (Statistics) : 확률표본 $X_1, X_2, \ldots, X_n$ 의 함수
    - 통계량의 예

$$\bar{X} = \sum_{i=1}^{n} X_i/n \ , \qquad S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, \qquad \max(X_i)\ , \qquad median(X_i)$$

  - **표본분포** (표집분포, Sampling Distribution) : 통계량의 확률분포

주요 통계량과 그 성질

- **주요 통계량**

  $X_1, X_2, \ldots, X_n$이 기대값이 μ이고 분산이 σ^2인 모집단의 분포 $f(x)$로부터의 확률 표본이라고 가정할 때,

  - **표본평균** (Sample Mean)
    - 정의

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

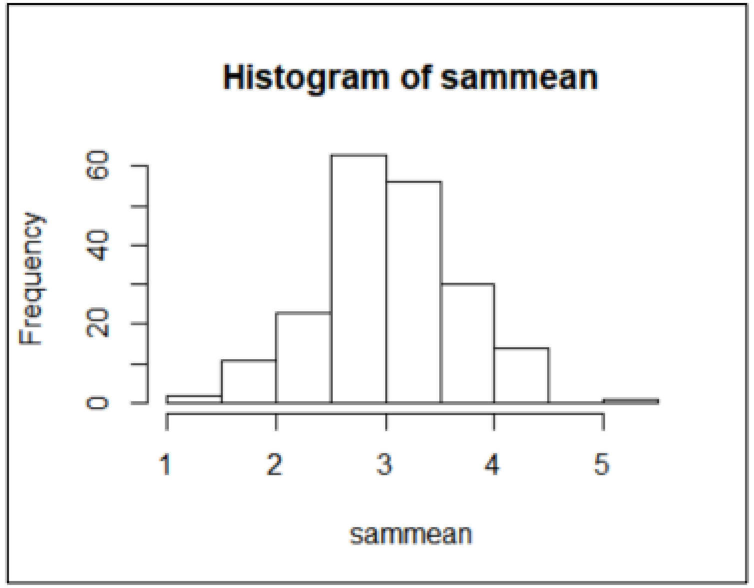
    - 성질

$$E[\bar{X}] = \mu \qquad and \qquad V[\bar{X}] = \frac{\sigma^2}{n}$$

► 표본평균 $\bar{X}$의 평균은 모집단의 평균 μ와 같으며,
표본의 크기 n이 클수록 그 분산이 0에 가까워져,
결국 표본의 크기가 클 때, $\bar{X}$는 μ 근처에 밀집되어 분포하게 됨을 알 수 있음.

주요 통계량과 그 성질

```
> sammean <- rep( 0, times=200 )
> for ( i in 1:200 ){
+ sam10 <- rnorm(10, mean=3, sd=2)
+ sammean[i] <- mean( sam10 )
+ }
> hist( sammean )
```



주요 통계량과 그 성질

```
> mean( sammean )
[1] 3.043031
> sd( sammean )
[1] 0.6408593
> sqrt( 2^2/10 )
[1] 0.6324555
```

주요 통계량과 그 성질

- **표본분산** (Sample Variance)
  - 정의

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

  - 성질

$$E[S^2] = \sigma^2$$

주요 통계량과 그 성질

- **표본비율** (Sample Proportion)

  $X_1, X_2, \ldots, X_n$이 모성공확률이 p 인 베르누이 분포 $f(x)$로부터의 확률 표본이라고 가정할 때,

  - 정의

$$\hat{p} = \sum_{i=1}^{n} X_i / n$$

  - 성질

$$E[\hat{p}] = p \qquad and \qquad V[\hat{p}] = \frac{p(1-p)}{n}$$

중심극한의 정리

- **중심극한의 정리** (Central Limit Theorem)

  평균이 μ이고 분산이 σ^2인 임의의 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \ldots, X_n$을 추출하면, 그 표본의 평균 $\bar{X}$는 n이 충분히 클 때 ($n > 30$),

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

  따라서

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

  이 성립한다.

중심극한의 정리

- **중심극한의 정리** (Central Limit Theorem)
  평균이 μ이고 분산이 σ^2인 임의의 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \ldots, X_n$을 추출하면, 그 표본의 평균 $\bar{X}$는 n이 충분히 클 때 ($n > 30$),

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

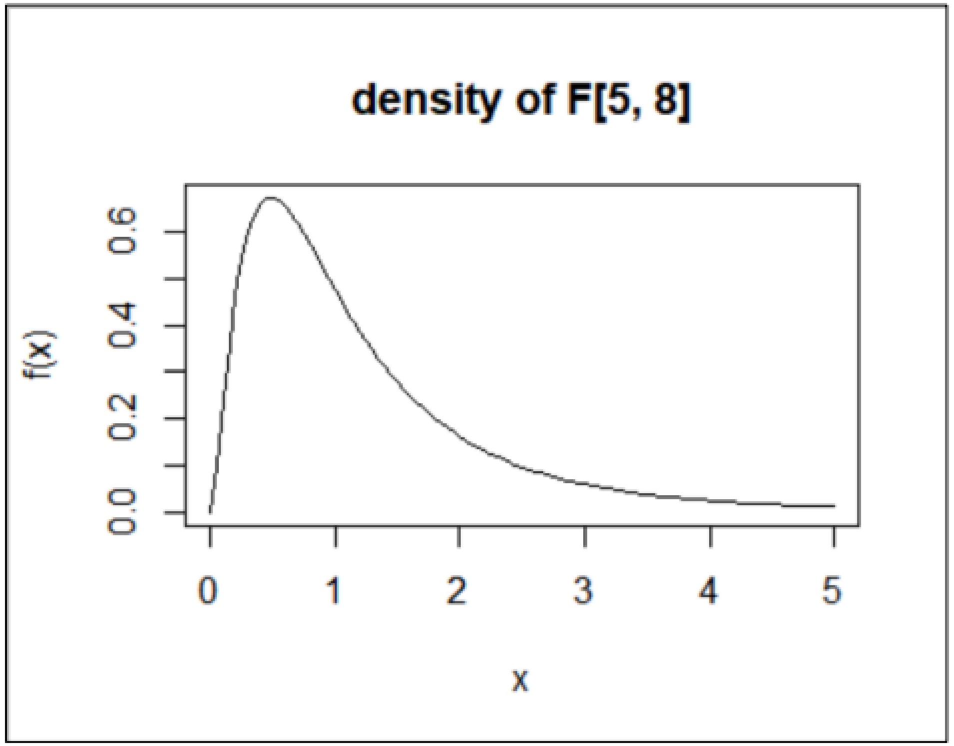
  따라서

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

  이 성립한다.

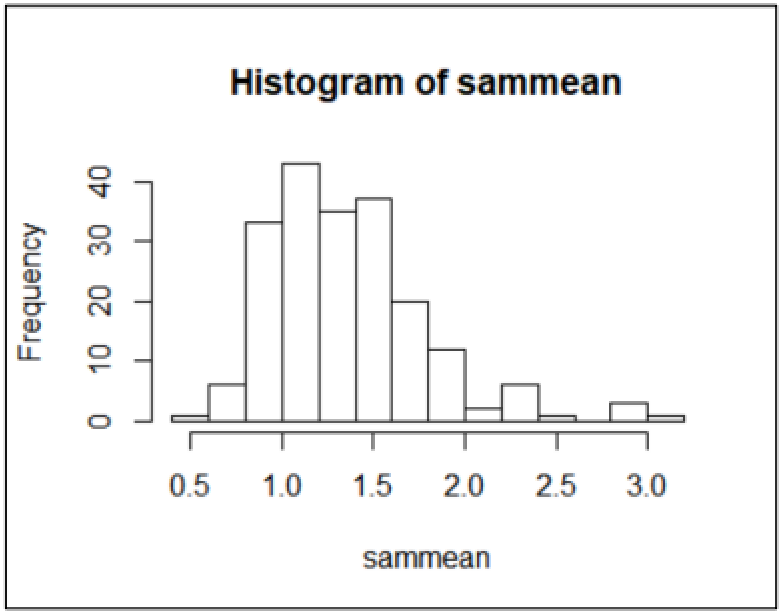
중심극한의 정리

```
> curve( df(x, df1=5, df2=8), from=0, to=5, ylab="f(x)",
+ main="density of F[5, 8]")
```



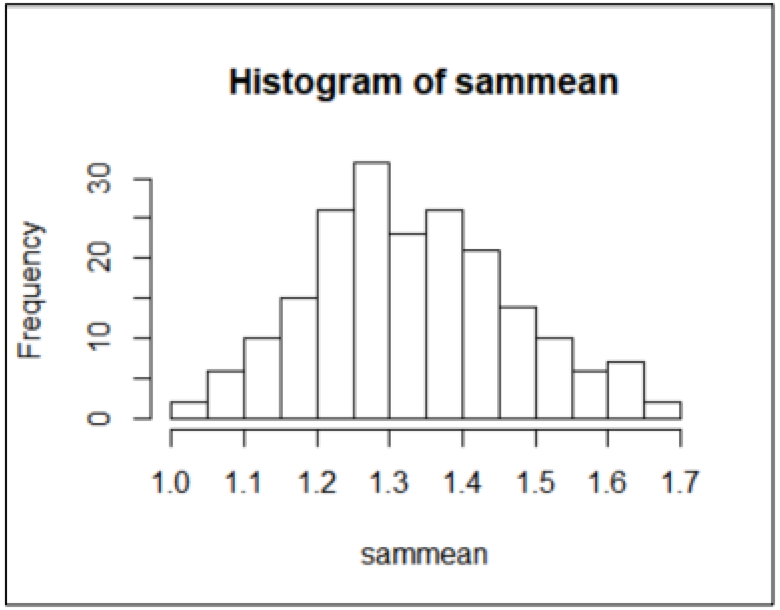
중심극한의 정리

```
> sammean <- rep( 0, times=200 )
> for ( i in 1:200 ){
+ sam10 <- rf(10, df1=5, df2=8)
+ sammean[i] <- mean( sam10 )
+ }
> hist( sammean, breaks=10 )
```



중심극한의 정리

```
> sammean <- rep( 0, times=200 )
> for ( i in 1:200 ){
+ sam100 <- rf(100, df1=5, df2=8)
+ sammean[i] <- mean( sam100 )
+ }
> hist( sammean, breaks=10 )
```



중심극한의 정리

- **예제**
  - 모평균이 10이고 모표준편차가 5인 어느 모집단 분포에서 크기 50의 무작위 표본을 추출한다고 가정하면, 표본평균의 분포는 무엇인가?

중심극한의 정리

- **예제**
  - 어느 지역 가구주의 월 평균소득은 300만원, 표준편차는 25만원이라고 한다. 이 지역에서 25명의 가구주를 임의로 뽑았을 때, 이들의 월 소득 평균이 290만원에서 310만원 사이에 있을 확률을 구하여라.

주요 통계량의 표본분포

- **표본평균 $\bar{X}$의 분포**
  - 모평균이 μ이고 모분산이 σ^2인 **정규** 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \dots, X_n$을 추출한 경우

$$\bar{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right) \ \rightarrow \ Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

  - 모평균이 μ이고 모분산이 σ^2인 **임의의** 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \dots, X_n$을 추출하였으며, **표본의 크기 $\boldsymbol{n}$이 충분히 큰 경우** ($n \geq 30$),

$$\bar{X} \dot{\sim} N\left(\mu,\ \frac{\sigma^2}{n}\right) \rightarrow \ Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \quad \text{(중심극한정리)}$$

주요 통계량의 표본분포

- 예제
  - 모분포인 $Normal[\mu, 3^2]$ 로부터 9개의 확률표본 $X_1, \dots, X_9$을 추출하였다고 하자. 이 경우 표본평균 $\bar{X}$ 가 모평균 μ 와의 차이의 절대값이 1보다 작을 확률 ($P[-1 < \bar{X} - \mu < 1]$)은 얼마인가?

주요 통계량의 표본분포

- 예제
  - 어느 해 인도에 투자하는 펀드들의 연 수익률의 평균은 7.98 %이며 표준 편차는 10.1 %로 알려져 있다고 가정하자. 이들 펀드 중 64개의 표본을 랜덤하게 추출하였을 때, 다음 물음에 답하여라.
    - 64개 표본의 연 수익률의 평균의 기대값은 얼마인가?
    - 64개 표본의 연 수익률의 평균의 분산은 얼마인가?
    - 64개 표본의 연 수익률의 평균이 10%를 넘을 확률은 얼마인가?

주요 통계량의 표본분포

- **표본비율 $\hat{\boldsymbol{p}}$의 분포**
  - 모성공비율이 p 인 베르누이 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \dots, X_n$을 추출하였으며, **표본의 크기 $\boldsymbol{n}$이 충분히 큰 경우** ($np \geq 5$ 이고 $n(1-p) \geq 5$),

$$\hat{p} \sim N\left(p,\ \frac{p(1-p)}{n}\right) \ \rightarrow \ Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1) \ \text{(중심극한정리)}$$

주요 통계량의 표본분포

- 예제
  - 역사적으로 매년 NYSE 상장 주식 중 61%가 가격이 상승한다고 가정하자. NYSE의 종목 중 100개를 임의로 추출하였을 때, 100개의 표본 종목 중 어느 한 해 동안 주가가 오른 경우의 비율이 55 %에서 65 %사이일 확률은 얼마인가?

주요 통계량의 표본분포

- 예제
  - 어느 금융회사의 고객 중 첫 해에 이탈하는 고객의 비율은 3%라고 한다. 이 주장이 사실이라면, 이 금융회사 고객 중 400명을 무작위로 선택하여 조사하였을 때, 이 중 첫 해에 이탈하는 고객이 20명이 넘을 확률은 얼마인가?

주요 통계량의 표본분포

- **표본분산 S^2의 분포**
  - 모평균이 μ이고 모분산이 σ^2인 **정규** 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \dots, X_n$을 추출한 경우

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

주요 통계량의 표본분포

- 예제
  - 모분포인 $Normal[10, 3^2]$ 로부터 20개의 확률표본 $X_1, \ldots, X_{20}$을 추출하였다고 할 때, 표본분산 S^2 이 12.799보다 클 확률을 구하여라.

주요 통계량의 표본분포

- 예제
  - 2000년~2020년 사이의 어느 회사 고수익률 펀드의 월수익률에 대한 표준편차는 4%였으며, 최근까지 이 수준은 잘 유지되고 있다고 주장하고 있다. 이 주장이 사실이라고 할 때, 최근 2년 간 월수익률에 대한 표준편차가 4.4%를 넘길 확률은 얼마인가? 단, 월 수익률은 독립적이며, 정규분포를 따르는 것으로 가정할 것.

주요 통계량의 표본분포

- **표본평균 $\bar{X}$와 표본분산 S^2의 분포**
  - 모평균이 μ이고 모분산이 σ^2인 **정규** 모집단으로부터 크기가 n인 확률표본 $X_1, X_2, \ldots, X_n$을 추출한 경우

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

주요 통계량의 표본분포

- 예제
  - 모분포인 $Normal[3, 2^2]$ 로부터 9개의 확률표본 $X_1, \dots, X_9$을 추출하여 구한 표본평균을 $\bar{X}$, 표본표준편차를 S라고 할 때, $Y\left(= \frac{\bar{X}-3}{S}\right)$가 0.5보다 클 확률을 구하여라.