통계 분석

2019년 2학기

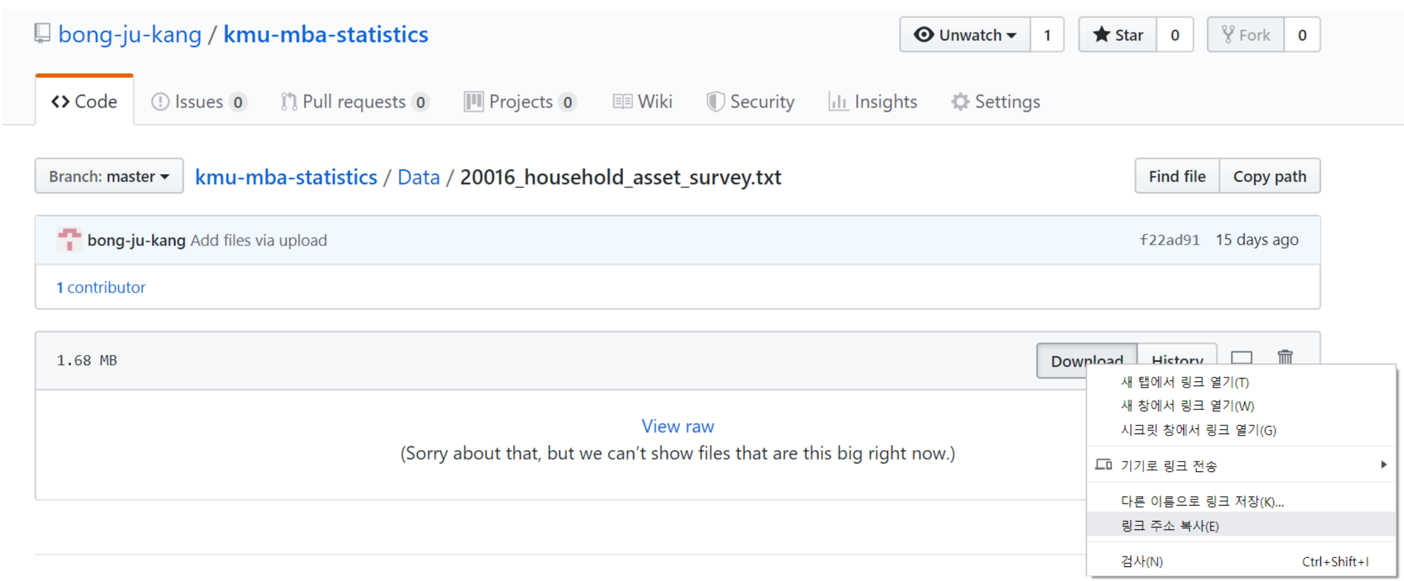
강봉주

표본 분포

표본 분포

[임의 표본(random sample)]

- [HAS: 통계청 마이크로데이터 서비스에서 제공하는 2006년 가계자산조사 데이터]



표본 분포

[**임의 표본(random sample)**]

- # 가계자산조사 > 연간자료(제공)[2006년 가계자산조사]
- url <- 'https://github.com/bong-ju-kang/kmu-mba-statistics/raw/master/Data/20016_household_asset_survey.txt'

- df <- read.csv(url, header=FALSE)
- str(df)

표본 분포

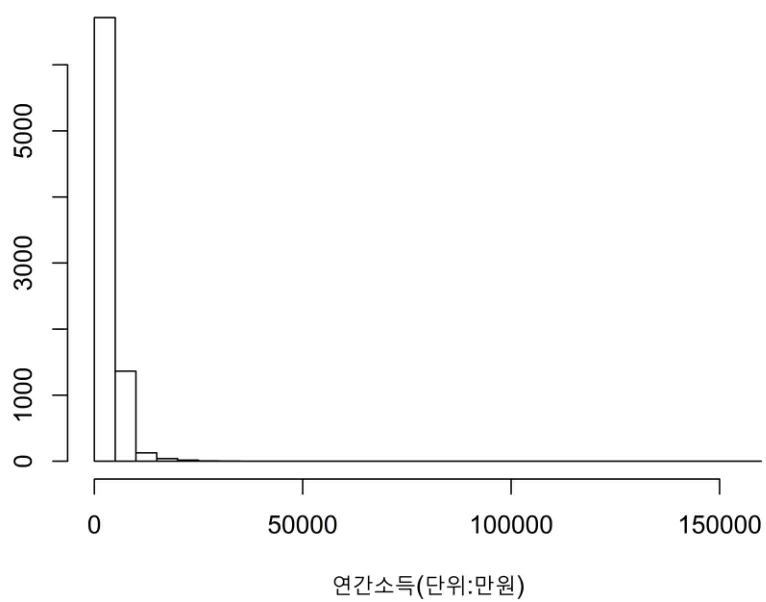
[**임의 표본(random sample)**]

예제) HAS_layout.txt를 참조하여 연간소득에 대한 히스토그램을 생성하세요.

표본 분포

[임의 표본(random sample)]

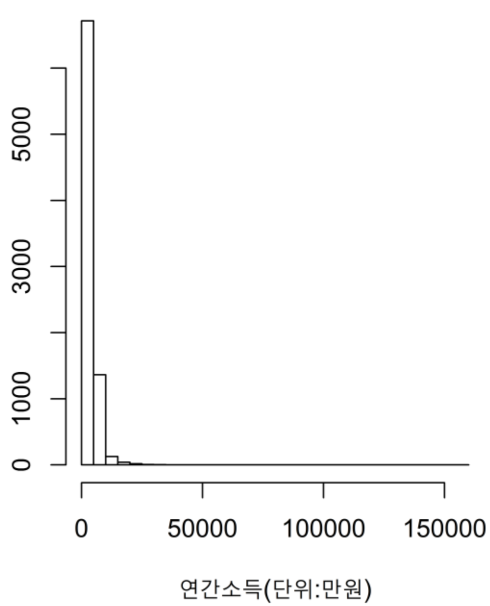
예제) HAS_layout.txt를 참조하여 연간소득에 대한 히스토그램을 생성하세요.

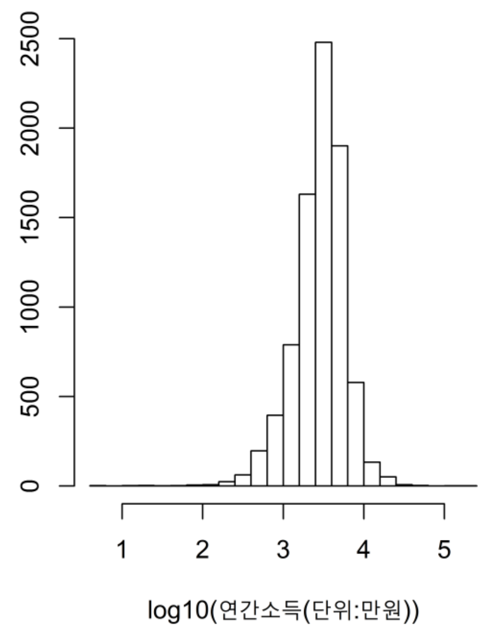


표본 분포

[임의 표본(random sample)]

- 적절한 변환을 통하여 큰 값과 작은 값이 더 보일 수 있도록





표본 분포

[**임의 표본(random sample)**]

- 모집단: 연간소득 $X \sim (\mu, \sigma) = (3613.25,\ 3594.248)$
- 크기가 100인 임의표본(random sample) $X_1, \dots, X_{100}$을 추출 후 표본 평균
- 표본의 비율은 약 1.2%

$$\bar{x} = 3481.42$$

표본 분포

[임의 표본(random sample)]

- 모집단: 연간소득 $X \sim (\mu, \sigma) = (3613.25,\ 3594.248)$
- 100인 임의표본(random sample) $X_1, \dots, X_{100}$을 추출 후 표본 평균
- 또 다른 표본

$$\bar{x} = 3518.87$$

표본 분포

[**임의 표본(random sample)**]

- 표본을 추출할 때 마다 서로 다른 추정값
- 하나의 추정값이 갖고 있는 유의미성 즉, 오차의 범위를 알고자 한다면?
- 표본 평균의 분포가 필요

표본 분포

[임의 표본(random sample)]

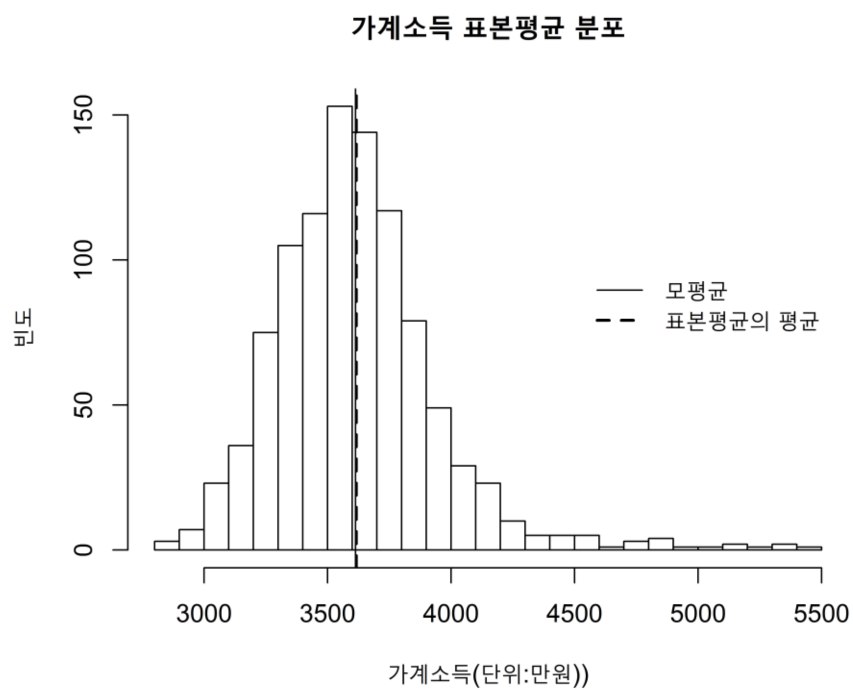
- 크기가 100인 임의 표본을 1000개 생성 후 표본 평균의 분포

```
# 크기가 100인 임의표본 1000개
result <- c()
for (i in 1:1000){
  result[i] <- mean(sample(income, 100))
}
```

표본 분포

[임의 표본(random sample)]

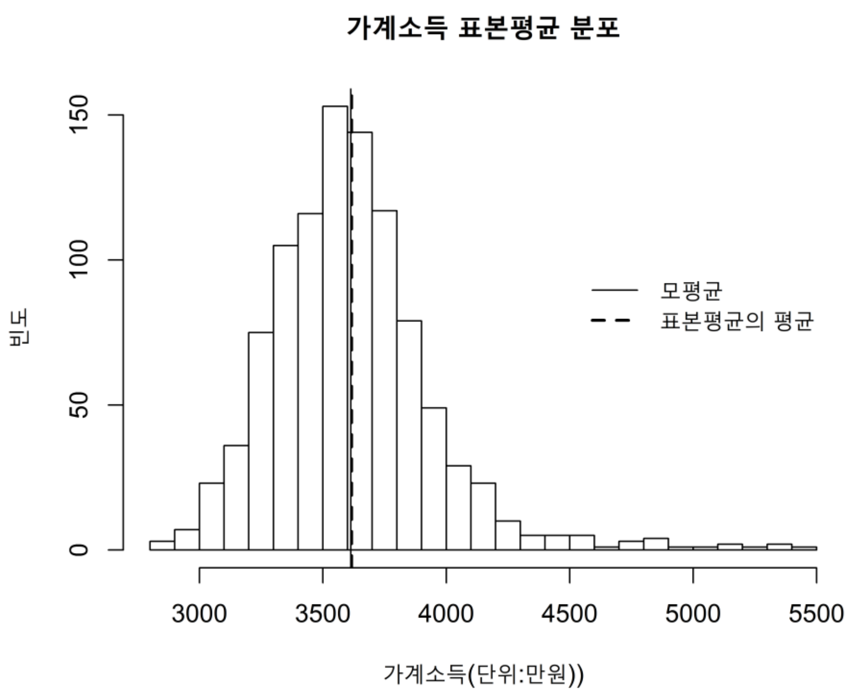
- 크기가 100인 임의 표본을 1000개 생성 후 표본 평균의 분포



표본 분포

[임의 표본(random sample)]

예제) 아래 그림을 생성하세요.



표본 분포

[**임의 표본(random sample)**]

- 표본 평균의 분포는 대략적으로 정규분포와 유사
- 이 분포를 통하여 표본 평균의 정확도 계산 가능?

$$\Pr(|\bar{X} - \mu| \leq 100)=?$$

상대 도수로 접근하면 대략 28.3%가 나옴

표본 분포

[**임의 표본(random sample)**]

$X_1, \ldots, X_n$이 **서로 독립**이고 각각은 **동일한** 확률밀도함수 $f(x)$를 갖는다고 한다면 $X_1, \ldots, X_n$의 결합확률밀도함수는 $f(x_1) \cdots f(x_n)$ 이 되며 이 때 $X_1, \ldots, X_n$을 크기(size)가 n인 임의표본(random sample)이라고 한다.

iid: independent and identically distributed

모집단의 유한모집단(finite population)인 경우에는
단순임의비복원추출(simple random sampling without replacement)

표본 분포

[표본평균의 분포]

- $X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$ 인 임의 표본
- $\bar{X} = \frac{1}{n}\sum_i X_i$

$$E(\bar{X}) = \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n}\sum_i E(X_i) = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2}Var(X_1 + \cdots + X_n) = \frac{1}{n^2}\sum_i Var(X_i) = \frac{\sigma^2}{n}$$

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

표본 분포

[표본평균의 분포]

- $X_1, \ldots, X_n \sim (\mu, \sigma^2)$ 인 임의 표본
- $\bar{X} = \frac{1}{n}\sum_i X_i$

X_i 들이 정규분포 이면 $\bar{X}$의 분포는?

표본 분포

[표본평균의 분포]

- $X_1, \ldots, X_n \sim (\mu, \sigma^2)$ 인 임의 표본
- $\bar{X} = \frac{1}{n}\sum_i X_i$

X_i 들이 정규분포 이면 $\bar{X}$의 분포는?

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

표본 분포

[표본평균의 분포]

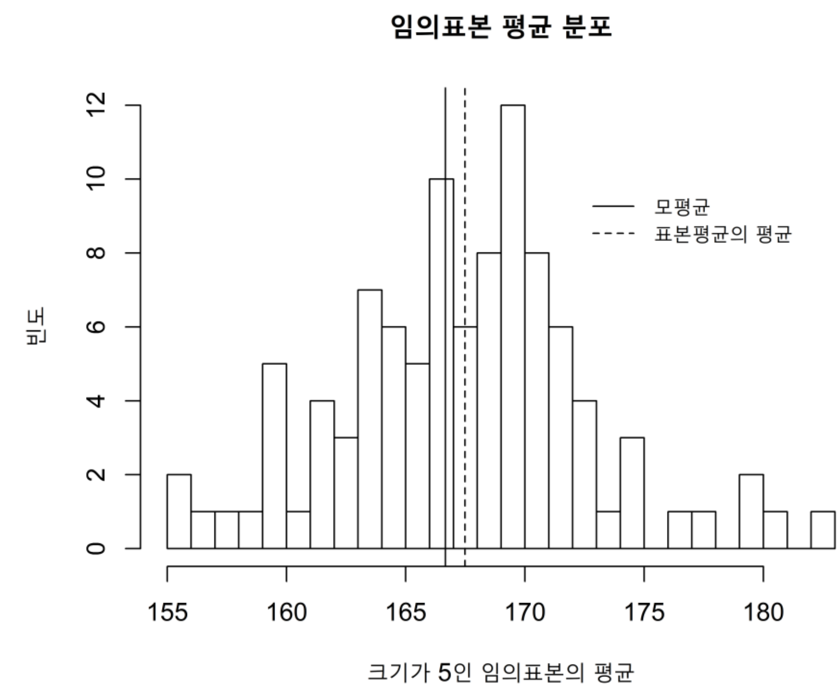
- $X_1, \dots, X_n \sim (\mu, \sigma^2)$ 인 임의 표본
- $\bar{X} = \frac{1}{n}\sum_i X_i$

X_i 들이 정규분포가 아니면 $\bar{X}$의 분포는?

표본 분포

[표본평균의 분포]

- $B\left(1000, \frac{1}{6}\right)$에서 크기가 5인 임의 표본 100개
- 표본 평균의 분포



표본 분포

[표본평균의 분포]

- $X_1, \ldots, X_n \sim (\mu, \sigma^2)$ 인 임의 표본
- $\bar{X} = \frac{1}{n}\sum_i X_i$

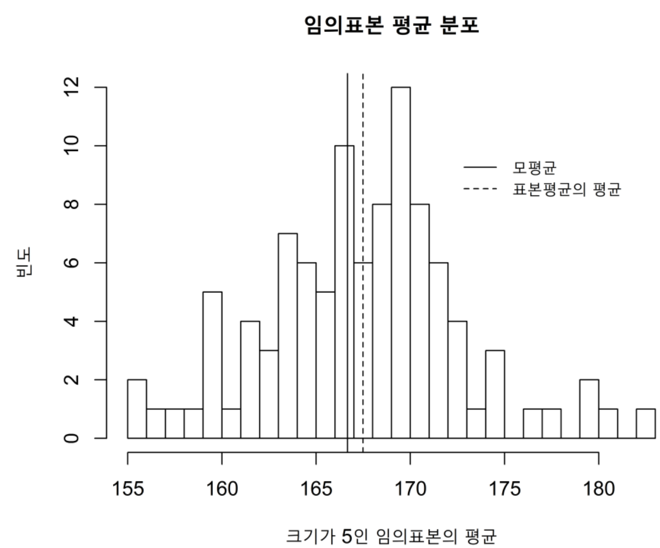
X_i 들이 정규분포가 아니면 $\bar{X}$의 분포는?

n이 크다면 정규분포로 수렴(중심 극한 정리)

표본 분포

[표본평균의 분포]

과제3) $B\left(1000, \frac{1}{6}\right)$에서 크기가 5인 임의 표본 100개에 대하여 표본평균에 대한 그래프를 생성하세요.



표본 분포

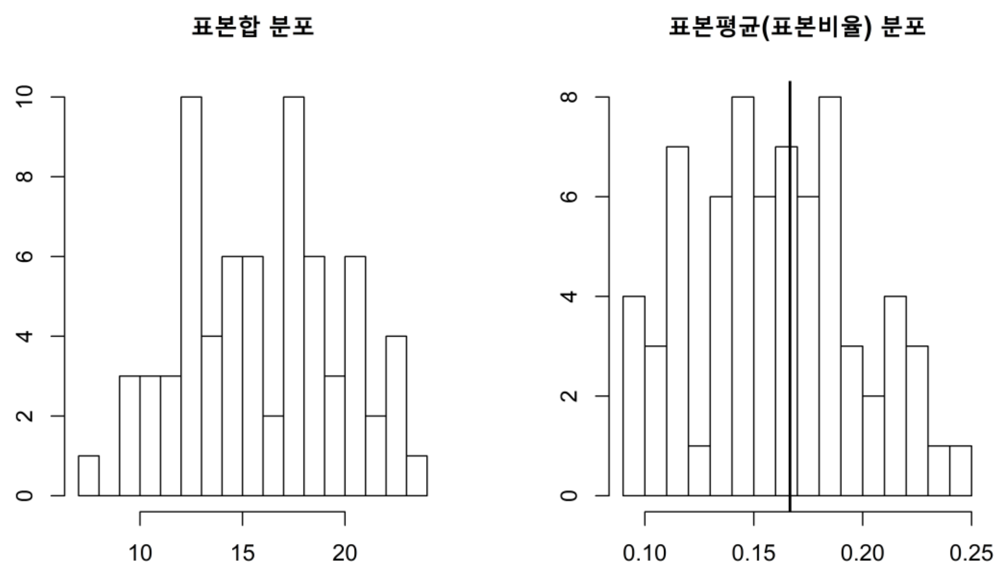
[표본평균의 분포]

예제) $X_i \sim \mathrm{B}(1, \frac{1}{6})$ 일 때 크기가 100인 임의 표본 70개에 대하여 합과 표본평균의 분포를 그래프로 표현하세요.

표본 분포

[표본평균의 분포]

예제) $X_i \sim \mathrm{B}(1, \frac{1}{6})$ 일 때 크기가 100인 임의 표본 70개에 대하여 합과 표본평균의 분포를 그래프로 표현하세요.



표본 분포

[표본평균의 분포]

- $X_i \sim \mathrm{B}(1, \frac{1}{5})$ 일 때 크기가 100인 임의 표본에 대하여 다음의 확률을 계산해보자.

$$\Pr\left(\sum X_i > 30\right) = ?$$

표본 분포

[표본평균의 분포]

- $X_i \sim \mathrm{B}(1, \frac{1}{5})$ 일 때 크기가 100인 임의 표본에 대하여 다음의 확률을 계산해보자.

$$\Pr\left(\sum X_i > 30\right) = ?$$

$$\sum X_i \sim B\left(100, \frac{1}{5}\right) \approx N\left(100 \times \frac{1}{5}, 100 \times \frac{1}{5} \times \left(1 - \frac{1}{5}\right)\right)$$
$$= N(20, 4^2)$$

표본 분포

[표본평균의 분포]

- $X_i \sim \mathrm{B}(1, \frac{1}{5})$ 일 때 크기가 100인 임의 표본에 대하여 다음의 확률을 계산해보자.

$$\Pr\left(\sum X_i > 30\right) = \Pr\left(Z > \frac{30 - 20}{4}\right)$$

$$= \Pr(Z > 2.5)$$

$$= 1 - \Phi(2.5)$$

$$= 0.0062$$

표본 분포

[카이제곱 분포(chi-squared distribution]

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$인 임의 표본
- $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$

S^2의 분포는?

표본 분포

[카이제곱 분포(chi-squared distribution)]

- $Z_1, \ldots, Z_k \sim N(0,1)$ 인 임의 표본
- $V \equiv Z_1^2 + \cdots + Z_k^2$

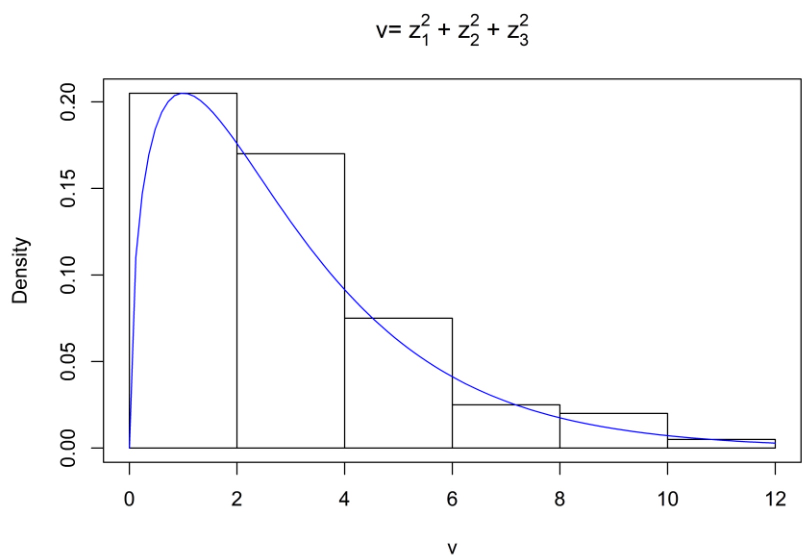
$V \sim \chi^2(k)$, k: 자유도

"자유도가 k인 카이제곱 분포를 따른다. "

표본 분포

[카이제곱 분포(chi-squared distribution)]

- 분포의 모습
- 크기가 100인 표본에 대한 분포의 모습
- $v = z_1^2 + z_2^2 + z_3^2, z_i \sim N(0,1)$



표본 분포

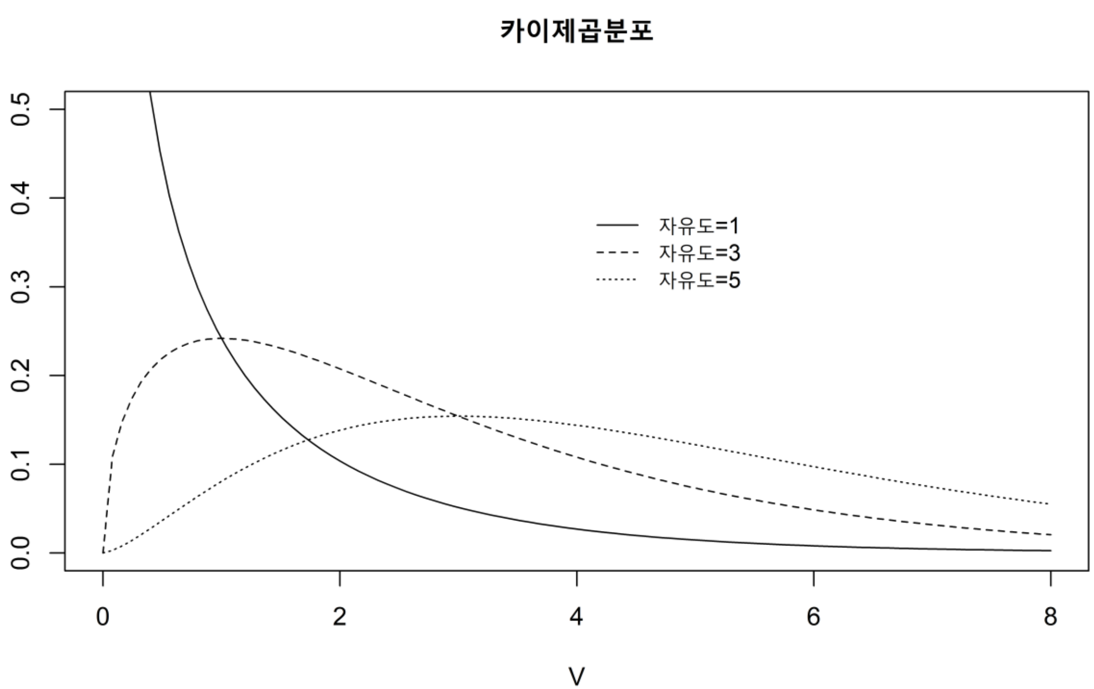
[**카이제곱 분포(chi-squared distribution]**

예제) V_1, V_2, V_3의 범위가 [0, 8] 일 때 $V_1 \sim \chi^2(1), V_2 \sim \chi^2(3), V_3 \sim \chi^2(5)$ 의 그래프를 생성하세요.

표본 분포

[카이제곱 분포(chi-squared distribution)]

예제) V_1, V_2, V_3의 범위가 [0, 8] 일 때 $V_1 \sim \chi^2(1), V_2 \sim \chi^2(3), V_3 \sim \chi^2(5)$ 의 그래프를 생성하세요.



표본 분포

[카이제곱 분포(chi-squared distribution)]

- 주요 성질
- $V \sim \chi^2(k)$
- 평균: k 분산: $2k$
- 가법성: $V_1 \sim \chi^2(k_1), V_2 \sim \chi^2(k_2)$ 이고 서로 독립이면 $V_1 + V_2 \sim \chi^2(k_1 + k_2)$

표본 분포

[카이제곱 분포(chi-squared distribution)]

- 표본 분산의 분포
- $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{S^2}{\sigma^2} = \frac{1}{n-1}\sum_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{1}{n-1}\left\{\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 - n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2\right\}$$

$$\Rightarrow \sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

$$\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n), \ \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)$$

표본 분포

[카이제곱 분포(chi-squared distribution)]

- 표본 분산의 분포
- $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{(n-1)S^2}{\sigma^2} = \sum_i \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

표본 분포

[**카이제곱 분포(chi-squared distribution]**

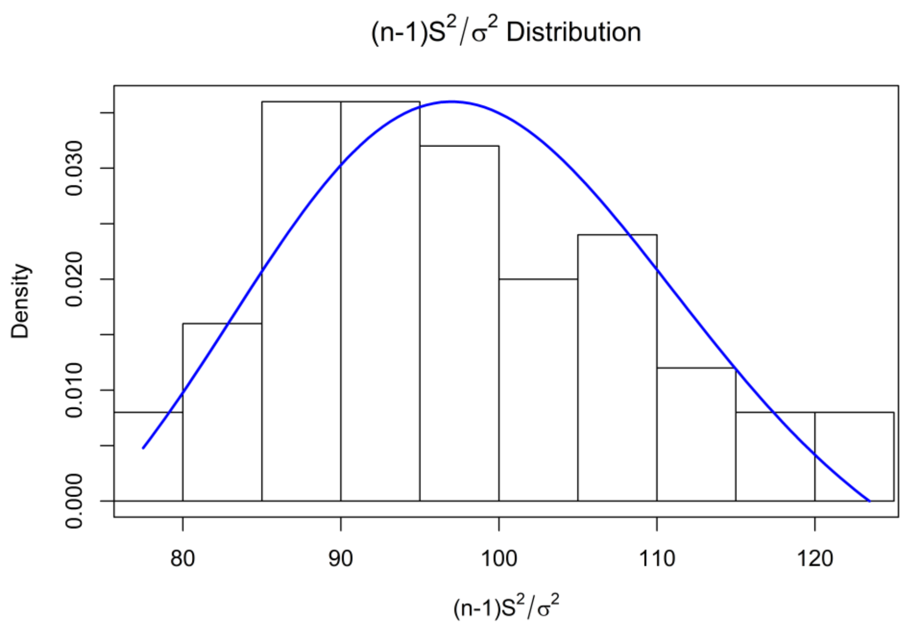
예제) $X_1, \ldots, X_{100} \overset{iid}{\sim} N(0, \sigma^2 = 3^2)$ 인 크기가 100인 표본 50개에 대하여 $\frac{(n-1)S^2}{\sigma^2}$의 분포를 그려보세요.

표본 분포

[카이제곱 분포(chi-squared distribution)]

예제) $X_1, \ldots, X_{100} \overset{iid}{\sim} N(0, \sigma^2 = 3^2)$ 인 크기가 100인 표본 50개에 대하여

$\frac{(n-1)S^2}{\sigma^2}$의 분포를 그려보세요.



표본 분포

[t 분포]

- $Z \sim N(0,1)$, $V \sim \chi^2(r)$, 서로 독립

$$T = \frac{Z}{\sqrt{V/r}}$$ 의 분포는?

표본 분포

[t 분포]

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 이지만 표준 편차를 모르는 경우?
- 모 표준편차를 표본 표준편차로 대체

$\frac{\bar{X}-\mu}{S/\sqrt{n}}$의 분포는?

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sqrt{S^2/n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{Z}{\sqrt{V/(n-1)}} = T$$

표본 분포

[t 분포]

- T 확률변수는 자유도가 $n-1$ 인 스튜던트 t 분포(Student t distribution)를 따른다

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

t분포는 고셋(William Sealy Gosset)이 Student 라는 필명으로 논문을 게재하였기 때문에 스튜던트 분포라고 하며 처음으로 피셔(Ronald Fisher)가 스튜던트분포와 t라는 문자를 사용하였다.

표본 분포

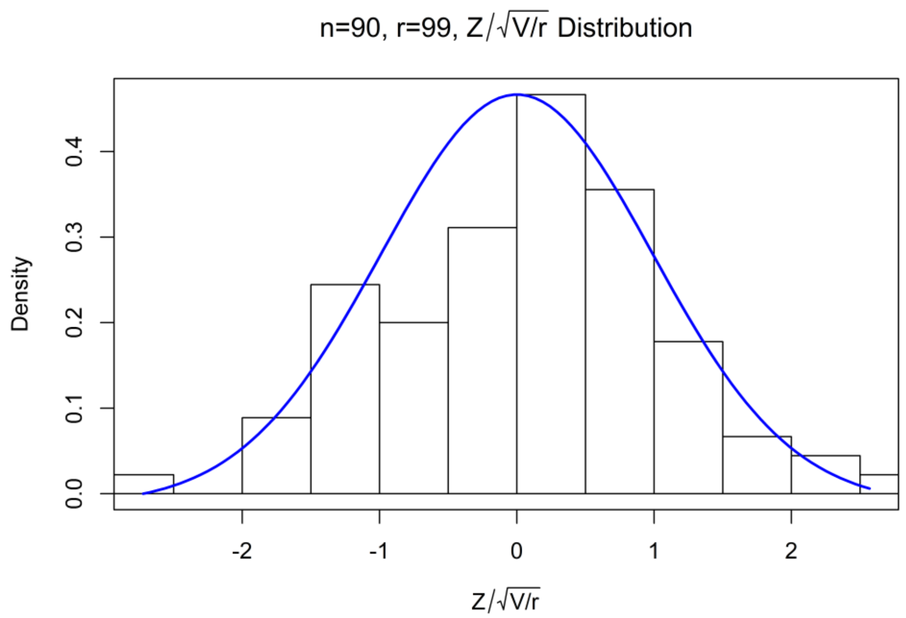
[t 분포]

예제) 자유도가 99일 때 표본의 크기가 90인 $T = \frac{Z}{\sqrt{V/r}}$의 분포를 그려보세요.

표본 분포

[t 분포]

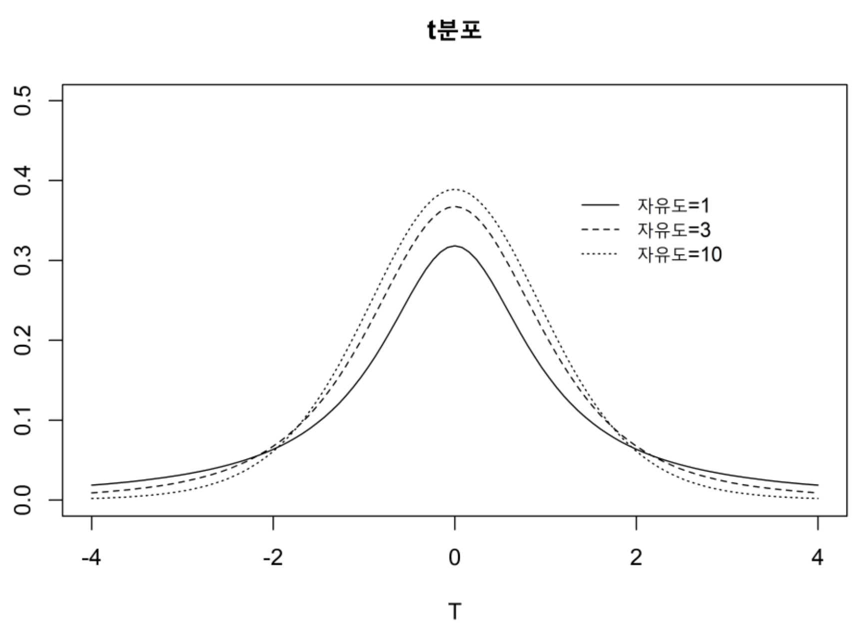
예제) 자유도가 99일 때 표본의 크기가 90인 $T = \frac{Z}{\sqrt{V/r}}$의 분포를 그려보세요.



표본 분포

[t 분포]

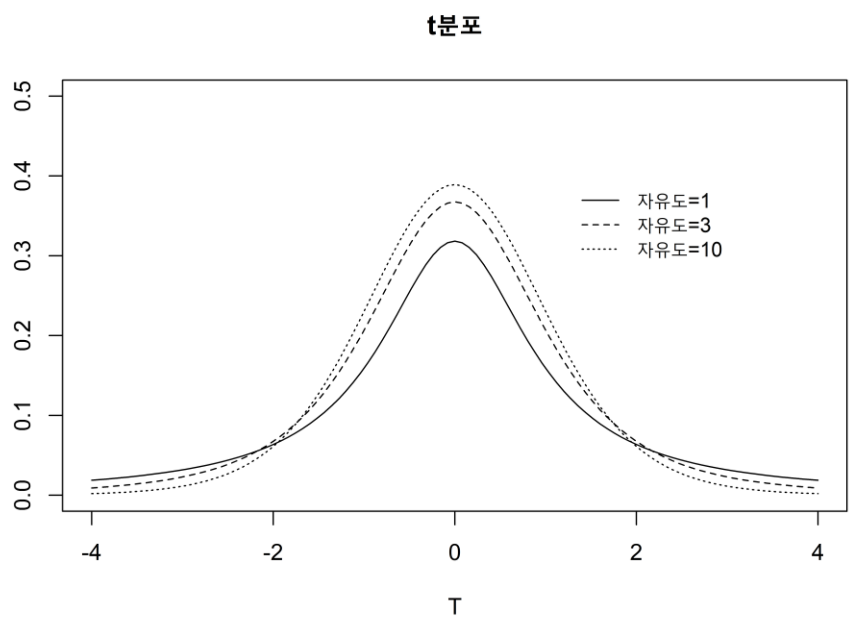
과제4) 자유도가 1, 3, 10인 t 분포에 대한 다음 그림을 프로그램으로 생성하세요.



표본 분포

[t 분포]

- 자유도가 작을 수록 꼬리 부분이 두터움
- 0에 대하여 좌우 대칭

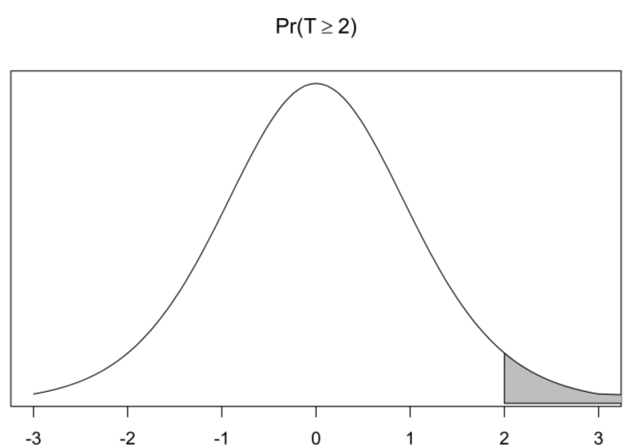


표본 분포

[t 분포]

- 0에 대하여 좌우 대칭

$$\Pr(T \geq 2) = ?$$

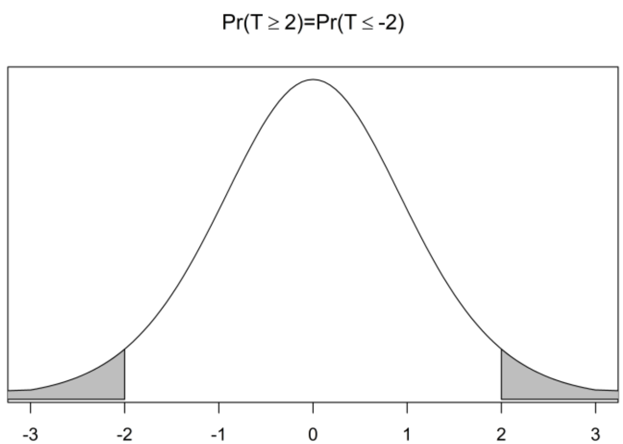


표본 분포

[t 분포]

- 0에 대하여 좌우 대칭

$$\Pr(T \geq 2) = \Pr(T \leq -2)$$




```
> pt(-2, 10)
[1] 0.03669402
```

표본 분포

[F 분포]

- $V_1 \sim \chi^2(r_1), V_2 \sim \chi^2(r_2)$ 이고 서로 독립일 때

$$F = \frac{V_1/r_1}{V_2/r_2} \text{ 의 분포는?}$$

표본 분포

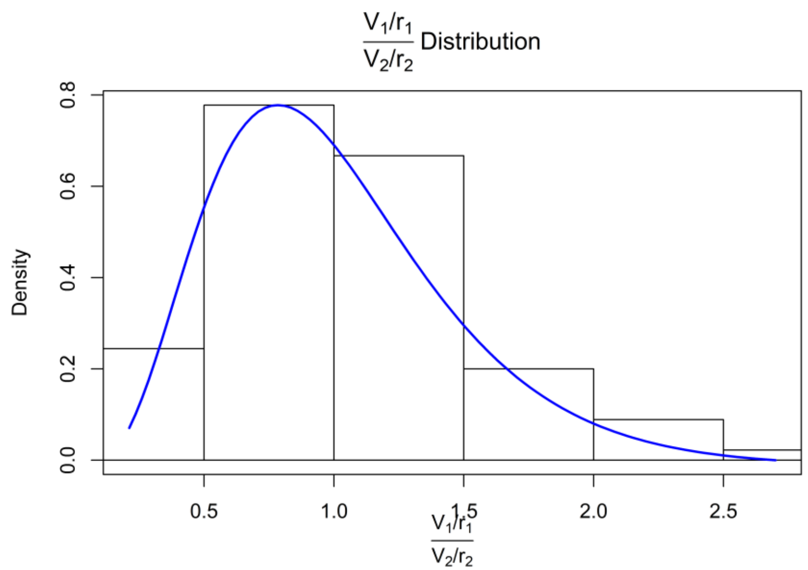
[F 분포]

예제) $F = \frac{V_1/10}{V_2/100}$ 에 대하여 표본 크기가 90인 경우에 분포의 모습에 대한 그림을 생성하세요.

표본 분포

[F 분포]

예제) $F = \frac{V_1/10}{V_2/100}$ 에 대하여 표본 크기가 90인 경우에 분포의 모습에 대한 그림을 생성하세요.



표본 분포

[F 분포]

- $V_1 \sim \chi^2(r_1), V_2 \sim \chi^2(r_2)$ 이고 서로 독립일 때

- $F = \frac{V_1/r_1}{V_2/r_2} \sim F(r_1, r_2)$
- F 확률변수는 자유도가 r_1, r_2 인 F분포를 따른다

F분포는 피셔-스네데코 분포(Fisher–Snedecor distribution) 이라고도 하며 분산분석(analysis of variance) 등에 널리 사용되는 분포이다.

표본 분포

[F 분포]

- $X_1, \dots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ 이고 $Y_1, \dots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ 이며 서로 독립일 때
- $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$

$$\frac{\dfrac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2}}{n_1-1}}{\dfrac{\dfrac{(n_2-1)S_2^2}{\sigma_2^2}}{n_2-1}} = \frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1)$$

모분산의 비에 대한 추론

표본 분포

[F 분포]

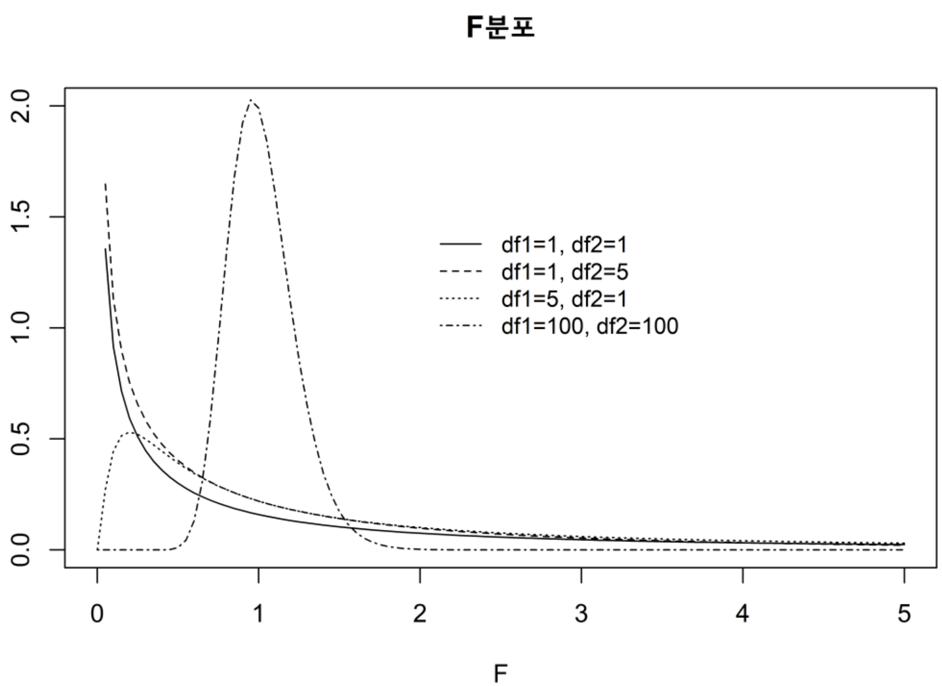
- 주요 성질
- $F \sim F(r_1, r_2)$

- $\frac{1}{F} \sim F(r_2, r_1)$
- $T = \frac{Z}{\sqrt{V/r}}$ 이면 $T^2 = \frac{Z^2/1}{V/r} \sim F(1, n-1) \because Z^2 \sim \chi^2(1)$

표본 분포

[F 분포]

과제5) 다음의 그림을 프로그램으로 생성하세요.

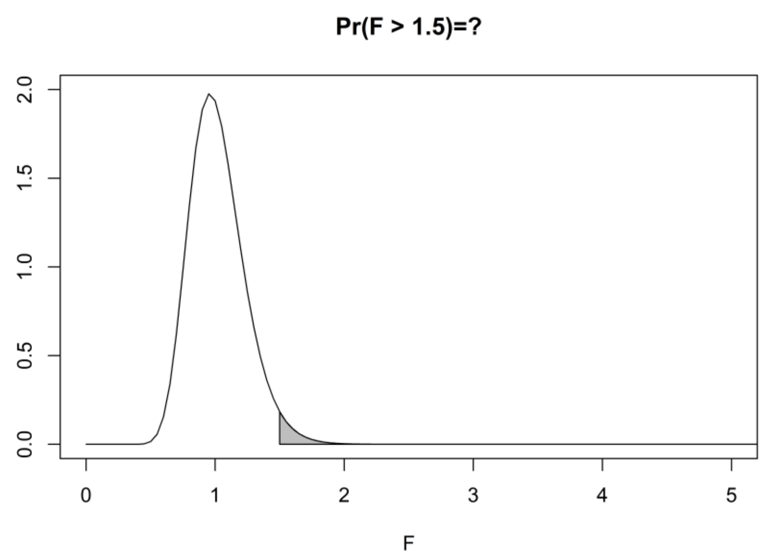


표본 분포

[F 분포]

- $F \sim F(90, 100\ )$

$$\Pr(F > 1.5) =?$$

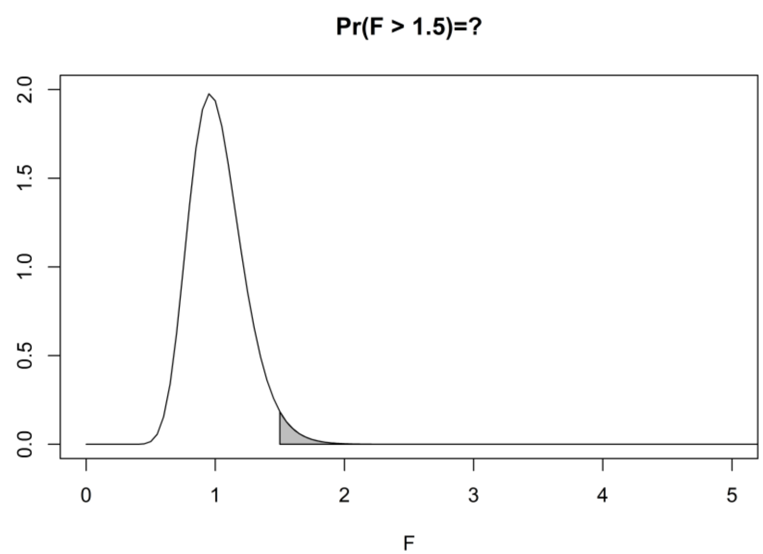


표본 분포

[F 분포]

- $F \sim F(90, 100\ )$

$$\Pr(F > 1.5) = 1 - \Pr(F \leq 1.5)?$$

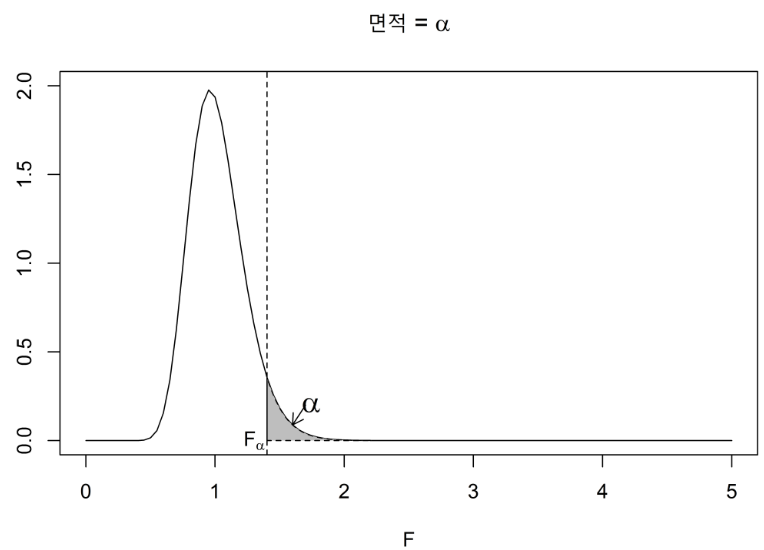


표본 분포

[F 분포]

- $F \sim F(90, 100\ )$

- $\Pr(F \geq F_{\alpha}) = \alpha$
- F_{α} 분위수(quantile)는?



표본 분포

[F 분포]

- $F \sim F(90, 100\,)$

  - $\Pr(F \geq F_{\alpha}) = \alpha \Rightarrow 1 - \alpha = \Pr(F \leq F_{\alpha})$
  - $\alpha = 0.05 \Rightarrow \Pr(F \leq F_{\alpha}) = 0.95$

```
> # 상위 alpha% 에 해당하는 분위수 찾기
> alpha <- 0.05
> f_alpha <- qf(1-alpha, df1=90, df2=100)
> print(f_alpha)
[1] 1.402047
```

표본 분포

[F 분포]

과제6) $F \sim F(90, 100)$ 일 때 다음의 그림을 생성하세요.