Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

09

F-분포와 분산분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)

09. *F*-분포와 분산분석

- **평균검정과 분산분석:**
  - **평균검정**: t-test
    - **두 개의 표본을 이용하여 각각 대응되는 두 개의 모집단 평균이 동일한지 검정**
  - **분산분석**: *ANOVA* (*analysis of variance*)
    - **세 개 이상의 표본집단에서 여러 모집단 간의 평균과 동일성을 검정**

- **분산분석 사례:**
  - ADHD(주의력결핍-과잉행동장애)에 대한 두 가지 심리치료 방법의 효과 비교
    - A: 첫 번째 심리치료, B: 두 번째 심리치료
    - 10명의 실험 참여자를 모집한 후 무작위로 A/B로 나눔
    - 심리치료 후 ADHD-RS(ADHD 평가점수) 측정

09. *F*-분포와 분산분석

- **일원 분산분석**: one-way ANOVA
  - **집단 간 일원 분산분석**: *between-groups*
    - **독립변수: 심리치료 방법**(A/B)
  - **집단 내 일원분산분석**: *within-groups*
    - **독립변수: 심리치료 기간**(4주/16주)

심리치료 A		심리치료 B	
환자	점수	환자	점수
s1		s6	
s2		s7	
s3		s8	
s4		s9	
s5		s10	

집단 간 일원분산분석

환자	기간	
	4주	16주
s1		
s2		
s3		
s4		
s5		
s6		
s7		
s8		
s9		
s10		

집단 내 일원분산분석

- **이원 분산분석**: two-way ANOVA
  - **독립변수: 심리치료 방법과 기간**
  - **주 효과**: *main effect*
    - **심리치료 방법과 기간의 영향**
  - **상호작용 효과**: *interaction effect*
    - **방법과 기간 간의 상호작용의 영향**

09. *F*-분포와 분산분석

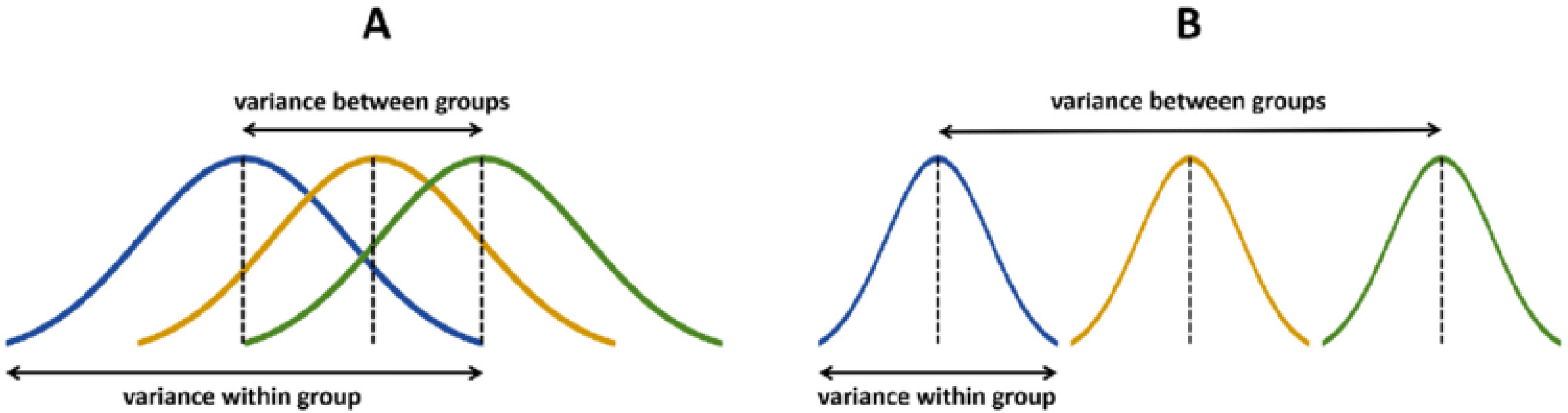
	환자	기간	
		4주	16주
심리치료 A	s1		
	s2		
	s3		
	s4		
	s5		
심리치료 B	s6		
	s7		
	s8		
	s9		
	s10		

분산분석을 위한 F-value:

- F-value: 집단 간 분산과 집단 내 분산의 비율로 계산
  - $F = \frac{\text{집단 간 분산}}{\text{집단 내 분산}}$
  - 집단 간 분산 $= \frac{\text{집단 간 제곱합}}{\text{자유도}} = \frac{\sum_g [(\overline{X_g} - \bar{X})^2 \times n_g]}{g-1}$
    - g: 집단의 개수, $\overline{X_g}$: g집단의 표본평균, $\bar{X}$: 전체 표본평균, n_g: g집단의 표본크기
  - 집단 내 분산 $= \frac{\text{집단 내 제곱합}}{\text{자유도}} = \frac{\sum_g \sum_i (X_{ig} - \overline{X_g})^2}{\sum_g (n_g - 1)}$
    - X_{ig}: g집단의 i번째 관측값

- F-검정: F-test
  - 집단 간의 평균의 차이를 검정할 때
    - 가설검정을 위한 검정통계량으로 F-value를 사용
  - 집단의 평균이 서로 다르다:
    - 집단평균의 퍼져 있는 정도를 나타내는 집단평균의 분산이 크다는 의미
    - 집단평균의 분산이 클수록 집단 간의 평균은 서로 다를 가능성이 높다.
    - 집단평균의 분산이 크더라도 각 집단 내의 분산 또한 크다면
    - 집단 간의 분포가 서로 겹쳐지는 영역이 커진다.

09. F-분포와 분산분석



https://www.researchgate.net/figure/Graphical-representation-of-the-rationale-behind-the-analysis-of-variance-ANOVA-A_fig2_329788831

09. *F*-분포와 분산분석

심리치료 A		심리치료 B	
환자	점수	환자	점수
s1	95	s6	110
s2	105	s7	125
s3	98	s8	105
s4	103	s9	113
s5	107	s10	120
표본평균	101.6	**표본평균**	114.6
표준편차	4.98	**표준편차**	7.96
전체 표본평균: 108.1 **전체 표준편차**: 9.28			

09. F-분포와 분산분석

- ADHD 데이터: 두 가지 심리치료 방법에 따른 치료효과 차이 검정

```
> adhd <- data.frame(score=c(95,105,98,103,107,110,125,105,113,120),
                     therapy=c(rep("A", 5), rep("B", 5)))
> adhd
 score therapy
1     95       A
2    105       A
3     98       A
4    103       A
5    107       A
6    110       B
7    125       B
8    105       B
9    113       B
10   120       B
```

- **집단 간 분산** $= \frac{\sum_g [(\overline{X_g} - \bar{X})^2 \times n_g]}{g-1}$

$$= \frac{(101.6-108.1)^2 \times 5 + (114.6-108.1)^2 \times 5}{2-1} = 422.5$$

```
> g <- 2
> ng <- c(5, 5)
> mg <- c(mean(adhd$score[1:5]), mean(adhd$score[6:10]))
> m <- mean(adhd$score)
> mstr <- sum(((mg-m)^2*ng) / (g-1))
> mstr
[1] 422.5
```

- 집단 내 분산 $= \frac{\sum_g \sum_i (X_{ig} - \overline{X_g})^2}{\sum_g (n_g - 1)}$

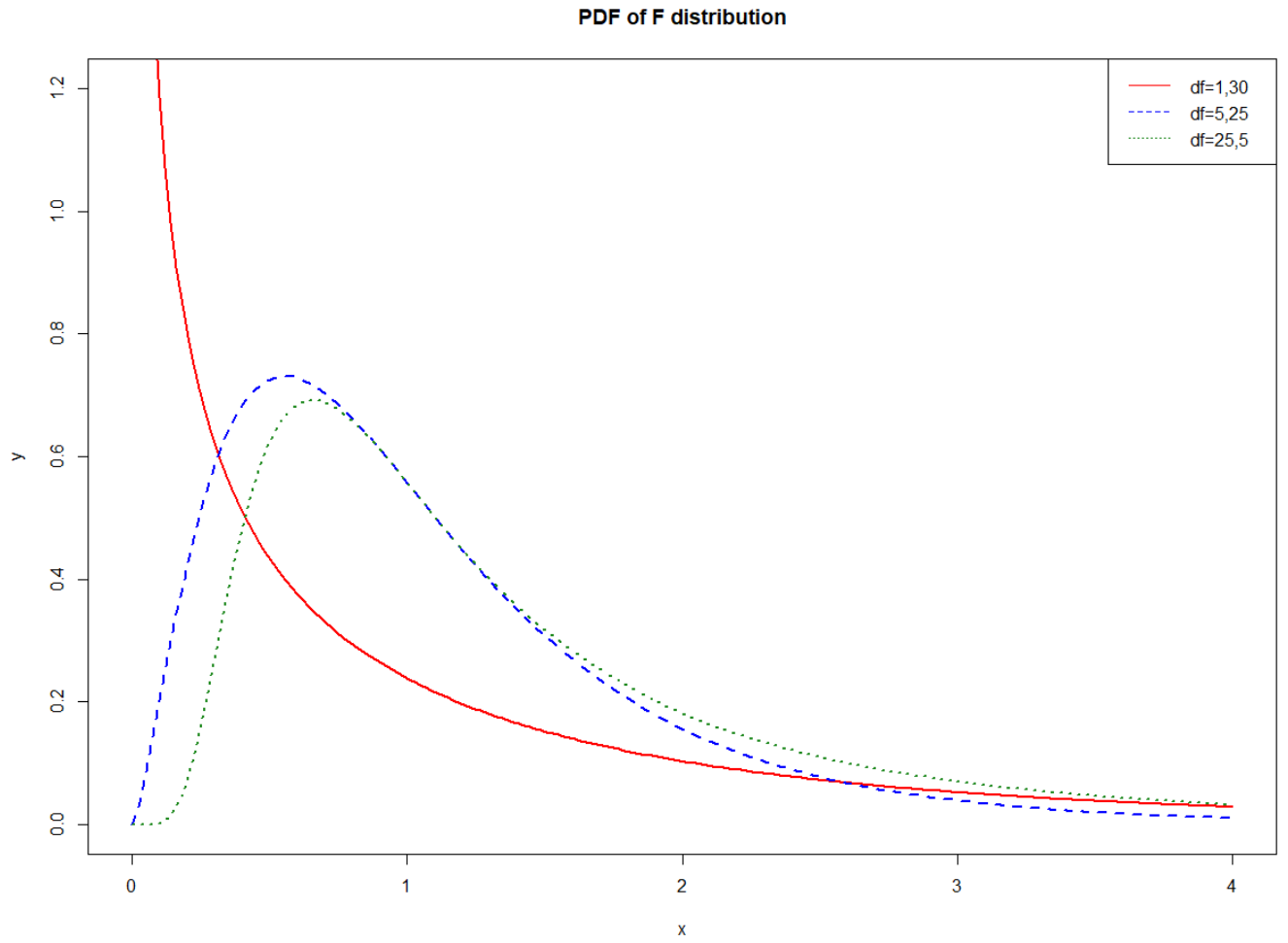
```
> Xg1 <- adhd$score[1:5]
> Xg2 <- adhd$score[6:10]
> mse <- (sum((Xg1-mg[1])^2) + sum((Xg2-mg[2])^2)) / sum(ng - 1)
> mse
[1] 44.05

> F.value <- mstr/mse
> F.value
[1] 9.591373
```

- F-분포: *F-distribution*
  - F-value는 두 개의 자유도에 의해 분포의 모양이 결정되는 F-분포를 따른다.

```
> x <- seq(0, 4, length=100)
> F.1 <- df(x, df1=1, df2=30)
> F.5 <- df(x, df1=5, df2=25)
> F.25 <- df(x, df1=25, df2=5)
> plot(x, F.1, lty=1, lwd=3, col="black", type="l", ylim=c(0, 1))
> lines(x, F.5, lty=2,lwd=3, col="blue")
> lines(x, F.25, lty=3, lwd=3, col="red")
> legend('topright', lty=c(1, 2, 3), col=c("black", "blue", "red"),
        legend=c("df = 1, 30", "df = 5, 25", "df = 25, 5"))
```

09. F-분포와 분산분석


PDF of F distribution
df=1,30
df=5,25
df=25,5
y
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4

09. F-분포와 분산분석

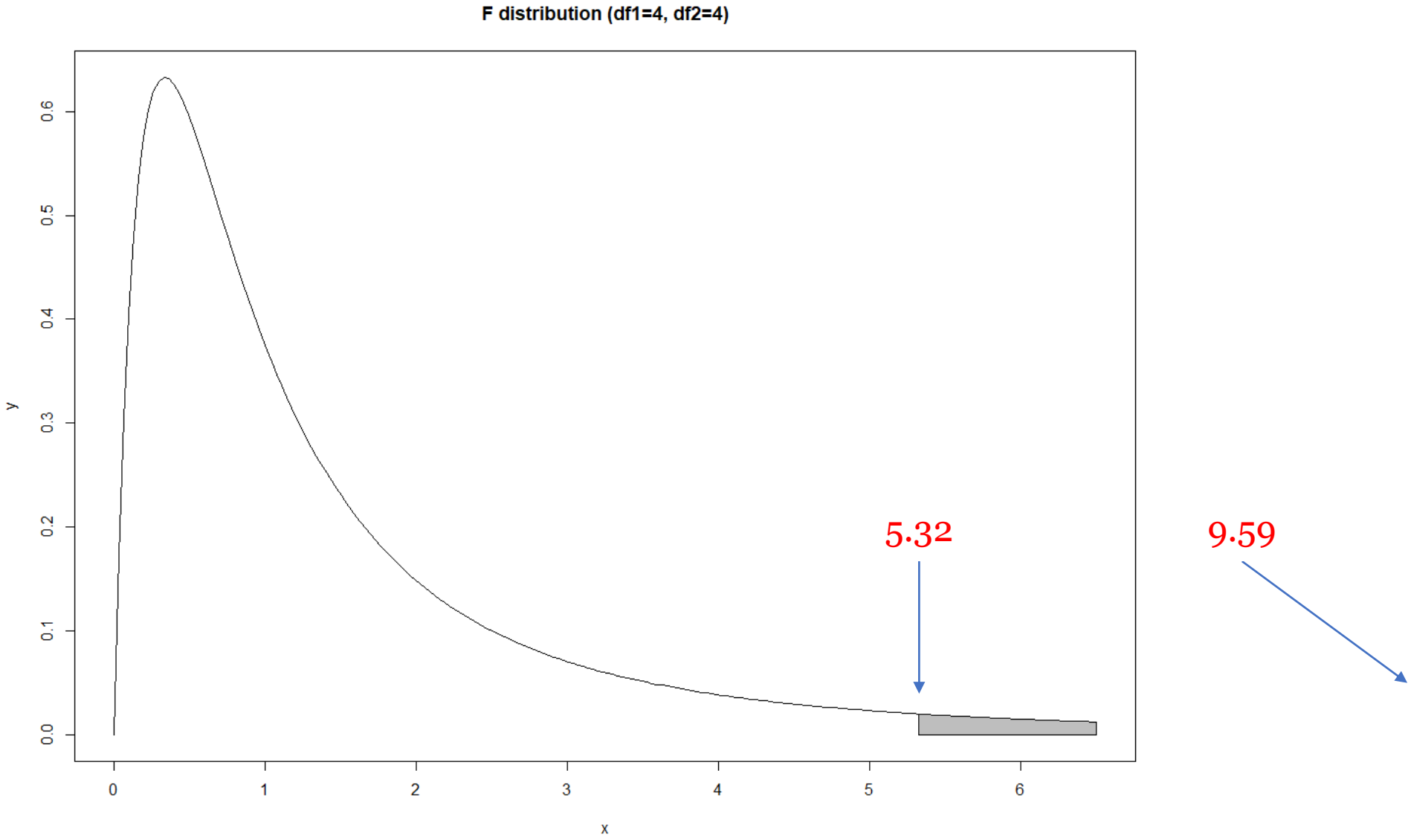
- **특정 F값에 대응되는 유의확률 구하기**

```
> pf(9.59, df1=1, df=8, lower.tail=FALSE)
[1] 0.0147376
```

- **특정 확률에 대응되는 F값 구하기**

```
> qf(0.05, df1=1, df=8, lower.tail=FALSE)
```

09. F-분포와 분산분석


F distribution (df1=4, df2=4)
y
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
5.32
9.59

09. F-분포와 분산분석

```
> str(adhd)
'data.frame':  10 obs. of  2 variables:
 $ score  : num  95 105 98 103 107 110 125 105 113 120
 $ therapy: chr  "A" "A" "A" "A" ...

> adhd.aov <- aov(score ~ therapy, data=adhd)
> summary(adhd.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
therapy      1  422.5   422.5   9.591 0.0147 *
Residuals    8  352.4    44.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> tapply(adhd$score, adhd$therapy, mean)
A     B
101.6 114.6

> tapply(adhd$score, adhd$therapy, sd)
A       B
4.97996 7.95613

> mean(adhd$score)
[1] 108.1

> sd(adhd$score)
[1] 9.279009
```

Any Questions?