Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

06

t-분포와 평균검정

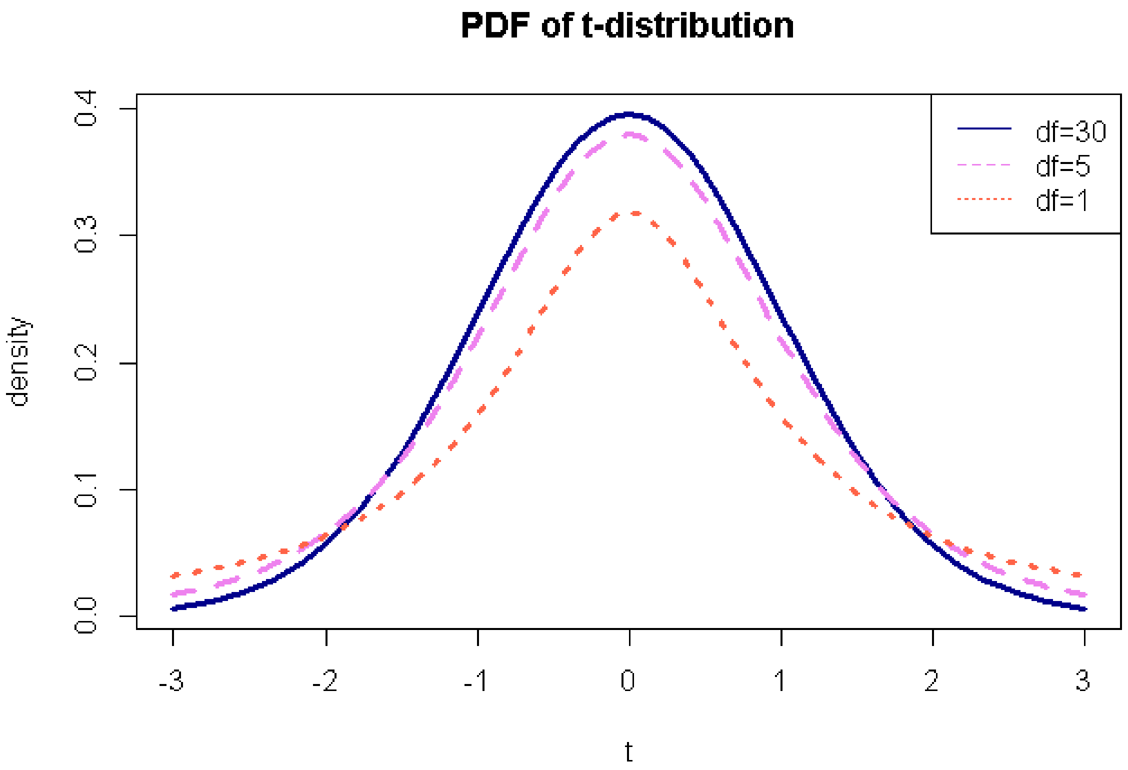
경북대학교 배준현 교수

(joonion@knu.ac.kr)

- 스튜던트의 t-분포: Student's t-distribution
  - 정규분포를 따르는 모집단으로부터 추출한 표본의 확률분포
  - 검정통계량 = t-value: $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$,
    - n: 표본크기, $\bar{X}$: 표본평균, μ: 모평균, s: 표본표준편차, $\frac{s}{\sqrt{n}}$: 표준오차
  - t-value는 t-분포를 따름
    - 종 모양의 형태를 가지면서 표본크기에 따라 종 모양이 달라짐
    - 상대적으로 정점이 낮고 양쪽 꼬리 부분이 더 두터우면서 퍼져 있는 모습
    - 표본크기가 작은 경우가 큰 경우보다 변동성이 더 큰 분포를 보임
    - 표본의 크기가 충분히 커지면 t-분포와 정규분포가 거의 유사함

- 자유도: *degree of freedom*
  - 자유도: 모집단에 대한 정보를 주는 독립적인 자료의 개수
    - 크기가 n인 표본에서 관측값의 자유도는 $n-1$
  - t-분포의 자유도: $df = n-1$
    - 표본의 크기 n에 따라 t-분포의 분산(표준편차)이 달라짐
    - 모집단의 분산: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$
    - 표본집단의 분산: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}$
    - 표본의 크기가 충분히 크면($n \geq 30$)
      - t-분포의 분산이 정규분포의 분산과 거의 유사해짐($n \approx n-1$)

```
x <- seq(-3, 3, length = 200)
curve(dt(x, df=30), min(x), max(x), lty = 1, lwd=3, col="darkblue",
      main = "PDF of t-distribution", xlab = "t", ylab = "density")
curve(dt(x, df=5), min(x), max(x), lty = 2, lwd=3, col="violet", add = T)
curve(dt(x, df=1), min(x), max(x), lty = 3, lwd=3, col="tomato", add = T)
legend("topright", legend = c("df=30", "df=5", "df=1"),
       col=c("darkblue", "violet", "tomato"), lty=c(1, 2, 3))
```



t-분포와 구간추정:

- t-분포의 특성을 이용하여 모집단 평균의 가능한 범위 예측
- 신뢰수준: 관측값이 일정 구간 내에 포함될 확률
  - 모집단으로부터 $n = 20$인 표본추출을 여러 번 반복하면
  - 표본평균의 95%는 모집단 평균과 약 두 배의 표준오차 범위 내에 있음
- 신뢰구간: 임의의 표본으로부터 산출된 표본평균과 표준오차 정보를 바탕으로
  - 95%의 신뢰도로 모집단평균이 포함되는 범위를 계산 가능
    - 표본평균의 범위: $\mu - t_{0.05,19} \times \frac{s}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{0.05,19} \times \frac{s}{\sqrt{n}}$
    - 모평균의 범위: $\bar{X} - t_{0.05,19} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{0.05,19} \times \frac{s}{\sqrt{n}}$
    - $t_{0.05,19}$: $\alpha = 0.05$, $df = 19$에 대응하는 t-value

t-분포와 t-검정:

- 모집단의 평균을 알지만 분산(표준편차)을 모를 때
  - 모집단으로부터 추출한 표본으로부터 추정된 표준오차를 통해
  - t-분포에 의존하여 가설을 검정하는 방법
- 귀무가설과 대립가설
  - 귀무가설: 표본집단은 모집단과 (평균, 비율이) 다르지 않다.
  - 대립가설: 표본집단은 모집단과 (평균, 비율이) 다르다.
- t-검정의 가정:
  - 정규성 가정: 모집단은 정규분포를 따른다.
  - 등분산성 가정: 두 집단을 비교할 때 두 집단의 분산이 동일한다.

- t-검정의 사례:
  - 대학원 박사과정 학생들은 스트레스를 많이 받아서 평균 혈압이
    - 같은 연령대의 다른 사람들에 비교하면 차이점이 있을 것 같다.
  - 귀무가설: 대학원 박사과정 학생들의 혈압은 다른 사람들과 다르지 않다.
    - 대립가설: 대학원 박사과정 학생들의 혈압은 다른 사람들과 다르다.
  - 가설검정: 귀무가설이 사실이라는 전제하에
    - 표본으로부터 얻은 t-값이 얼마나 흔하게/드물게 관찰될 수 있는가?
  - 표본추출: $n = 20, \bar{X} = 135, s = 25, \mu = 115$
    - $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{135-115}{\frac{25}{\sqrt{20}}} = 3.58$

06. t-분포와 평균검정

- 유의수준 0.05 또는 0.01에서의 t-값을 구한 후 관측된 t값과 비교

```
> t <- (135-115)/(25/sqrt(20))
> t
[1] 3.577709

> pt(3.58, df=19, lower.tail=FALSE)*2
[1] 0.001997274

> qt(0.025, df=19, lower.tail=FALSE)
[1] 2.093024

> qt(0.005, df=19, lower.tail=FALSE)
[1] 2.860935
```

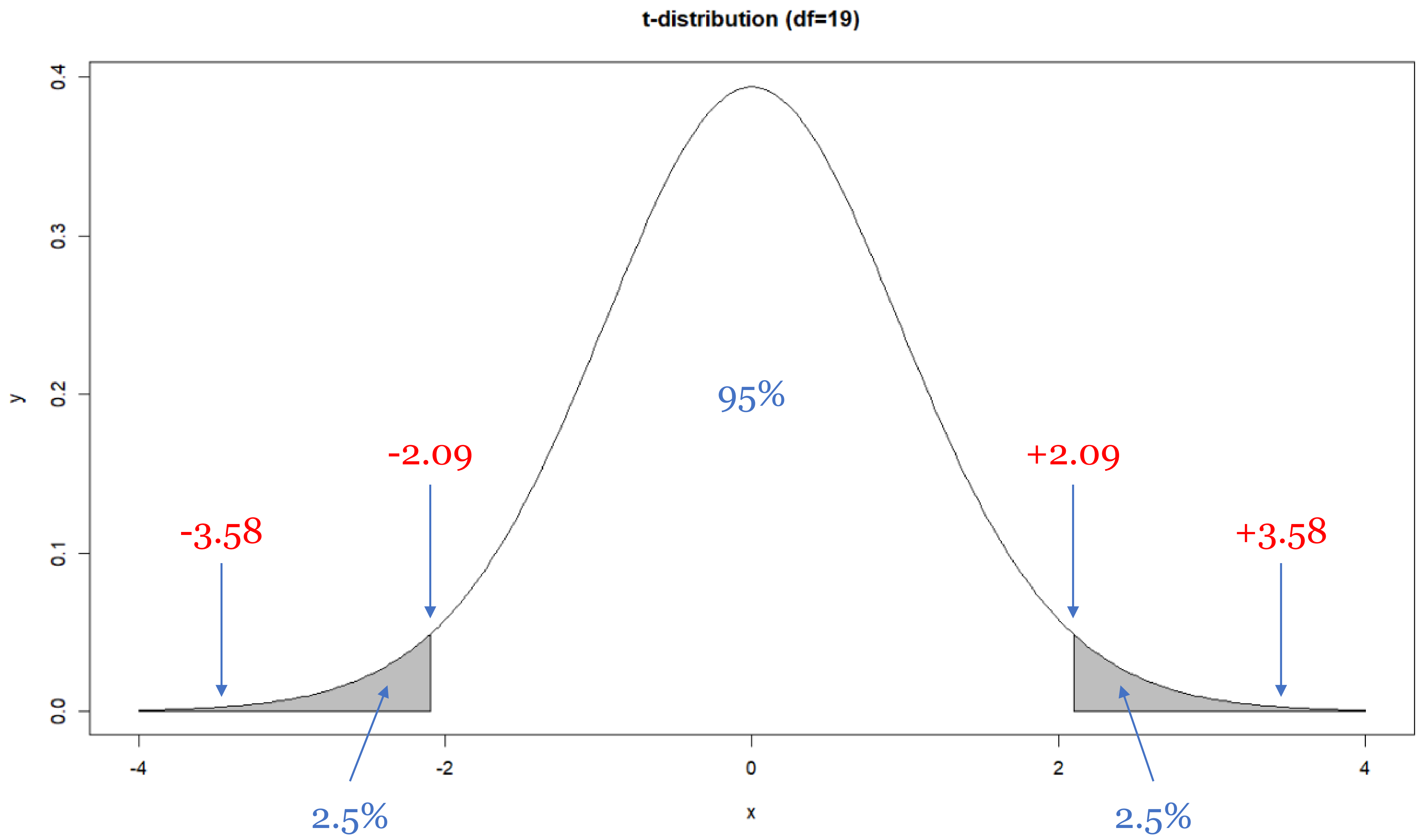
- 유의수준 0.05에서 귀무가설을 기각: 박사과정 학생들의 혈압은 다른 사람들과 같지 않다.

- 신뢰구간: *confidence interval*
  - t-분포의 특성을 이용하여 모집단평균의 가능한 범위 예측
  - 정규분포 또는 t-분포: 평균과 표준편차를 알면 신뢰도를 알 수 있음
    - 일반적으로 표본의 개수가 작을 때($n < 30$)는 t-분포를 활용
  - 신뢰도: 관측값이 일정 구간 내에 포함될 확률
    - 모집단으로부터 $n = 20$인 표본추출을 반복하면 표본평균의 95%는
    - 모집단 평균과 약 두 배의 표준오차 범위 내에 있음 (자유도: $df = 19$)

```
x <- seq(-4, 4, length=300)
y <- dt(x, df=19)
plot(x, y, type='l', main="t-distribution (df=19)")

xlim <- x[-4<=x & x<=-2.09]
ylim <- y[-4<=x & x<=-2.09]
xlim <- c(xlim[1], xlim, tail(xlim, 1))
ylim <- c(0, ylim, 0)
polygon(xlim, ylim, col="grey")

xlim <- x[2.09<=x & x<=4]
ylim <- y[2.09<=x & x<=4]
xlim <- c(xlim[1], xlim, tail(xlim, 1))
ylim <- c(0, ylim, 0)
polygon(xlim, ylim, col="grey")
```


t-distribution (df=19)
0.4
0.3
0.2
0.1
0.0
y
95%
-2.09
+2.09
-3.58
+3.58
-4
-2
0
2
4
x
2.5%
2.5%

- 구간추정:
  - 임의의 표본으로부터 산출된 표본평균과 표준오차 정보를 바탕으로
    - 95%의 신뢰도로 모집단평균이 포함되는 범위를 계산 가능
  - 표본평균의 범위: $\mu - t_{0.05,19} \times \frac{s}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{0.05,19} \times \frac{s}{\sqrt{n}}$
    - $t_{0.05,19}$: $\alpha = 0.05$(자유도=19)에 대응하는 t값(=2.09)
  - 모평균의 범위: $\bar{X} - t_{0.05,19} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{0.05,19} \times \frac{s}{\sqrt{n}}$
    - $135 - 2.09 \times \frac{25}{\sqrt{20}} \leq \mu \leq 135 + 2.09 \times \frac{25}{\sqrt{20}}$
    - $123.3 \leq \mu \leq 146.7$

- **평균검정:** *t-test*
  - **평균을 비교할 때 쓸 수 있는 가설검정 방법**
  - **단일표본 평균검정:** *one-sample* t-test
    - **표본평균을 가설로 정한 값과 비교**
  - **독립표본 평균검정:** *two-independent-samples* t-test
    - **두 집단간의 평균을 비교해서 집단의 차이에 대한 가설검정**
  - **대응표본 평균검정:** *paired-samples* t-test
    - **관측값이 서로 쌍을 이루는 경우**(예: 사전-사후)**에 대한 가설검정**

- **단일표본** 평균검정: *one-sample* t-test
  - **하나의 표본** 데이터를 이용하여 모집단의 평균이 특정값과 같은지 검정
  - 표본집단이 특정 모집단과 일치하는지 혹은 그렇지 않은지 알고 싶을 때
    - 대학원 박사과정생의 혈압은 동일 연령대의 다른 사람들의 혈압과 동일한가?
    - 가구당 소득에 대한 표본을 바탕으로 기존에 알려진 가구당 소득이 맞는가?

- MASS 패키지의 cats 데이터셋을 이용한 일표본 평균검정

```
> str(cats)
'data.frame':	144 obs. of  3 variables:
 $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
 $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

- 귀무가설: 고양이의 몸무게는 2.6kg이다.

```
> t.test(x=cats$Bwt, mu=2.6)
          One Sample t-test

data:  cats$Bwt
t = 3.0565, df = 143, p-value = 0.002673
alternative hypothesis: true mean is not equal to 2.6
95 percent confidence interval:
 2.643669 2.803553
sample estimates:
mean of x
 2.723611
```

```
> t.test(cats$Bwt, mu=2.7)

        One Sample t-test

data:  cats$Bwt
t = 0.58382, df = 143, p-value = 0.5603
alternative hypothesis: true mean is not equal to 2.7
95 percent confidence interval:
 2.643669 2.803553
sample estimates:
mean of x
 2.723611
```

- 단측검정: 고양이의 몸무게가 2.6kg보다 크다

```
> t.test(cats$Bwt, mu=2.6, alternative="greater")

	One Sample t-test

data:  cats$Bwt
t = 3.0565, df = 143, p-value = 0.001337
alternative hypothesis: true mean is greater than 2.6
95 percent confidence interval:
 2.656656      Inf
sample estimates:
mean of x 
 2.723611 
```

06. t-분포와 평균검정

```
> cats.t <- t.test(cats$Bwt, mu=2.6)
> str(cats.t)

> cats.t$p.value
> cats.t$conf.int

> t.test(cats$Bwt, mu=2.6, conf.level=0.99)
```

Any Questions?