R을 이용한 기초통계학

5강: 추정과 가설검정

수원대학교 데이터과학부 김 진 흠		서울대학교 보건환경연구소 이 보 라

The 19th ASIAN INSTITUTE IN STATISTICAL GENETICS AND GENOMICS

INTRODUCTORY STATISTICS USING R

추정

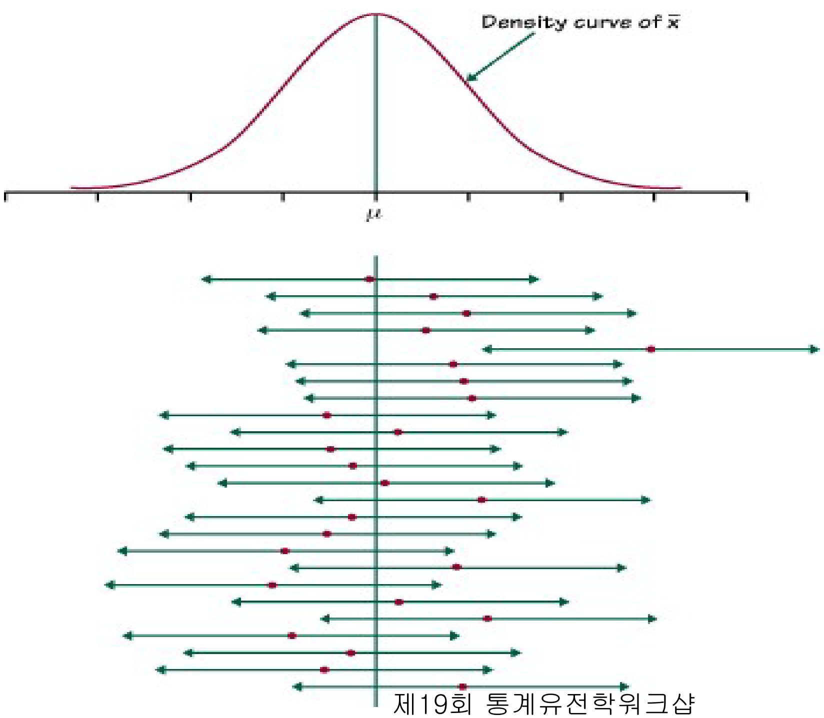
- 관심 있는 미지의 수값(=모수, parameter)을 주어진 자료로부터 추정하는 것
- 점추정: 모수를 **한 값**으로 추정하는 것
  - 모평균 ← 표본평균
  - 모비율 ← 표본비율
- 구간추정: 모수가 포함될 **구간**을 추정
  - 95% 신뢰구간

구간추정

- 모수가 포함될 구간을 추정
- $100 \times (1-\alpha)\%$ 신뢰구간(confidence interval: CI)
  - $P[L < \theta < U] = 1 - \alpha$
  - L과 U는 주어진 자료로부터 추정
- 방법
  - **추정량** ± (추정량 분포의 **상위 누적확률** $\boldsymbol{\frac{\alpha}{2}}$에 해당하는 값) × (추정량의 **표준오차**)

신뢰구간의 뜻

- 95% 신뢰구간의 해석: “100개의 data set으로부터 100개의 서로 다른 신뢰구간을 구했을 때 이 중에서 95개 정도는 미지의 모수를 포함한다”



모평균에 대한 점추정

- 모수: μ
- Data: $X_1, \dots, X_n \sim iid\ (\mu, \sigma^2), \sigma^2$: unknown
- **점추정량**: $\hat{\mu} = \bar{X}$
- **표준오차**(standard error; SE): $\mathbf{SE}(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})} = \sqrt{\mathrm{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$
- $\sigma \leftarrow S$:표본표준편차

모평균에 대한 구간추정

- When population is normal,

  $\bar{X} \pm \boldsymbol{t_{\frac{\alpha}{2}}(n-1)} \times \mathrm{SE}(\bar{X})$

  - $t_\alpha(n)$: 자유도가 n인 t−분포에서 상위 누적확률 α에 해당하는 값

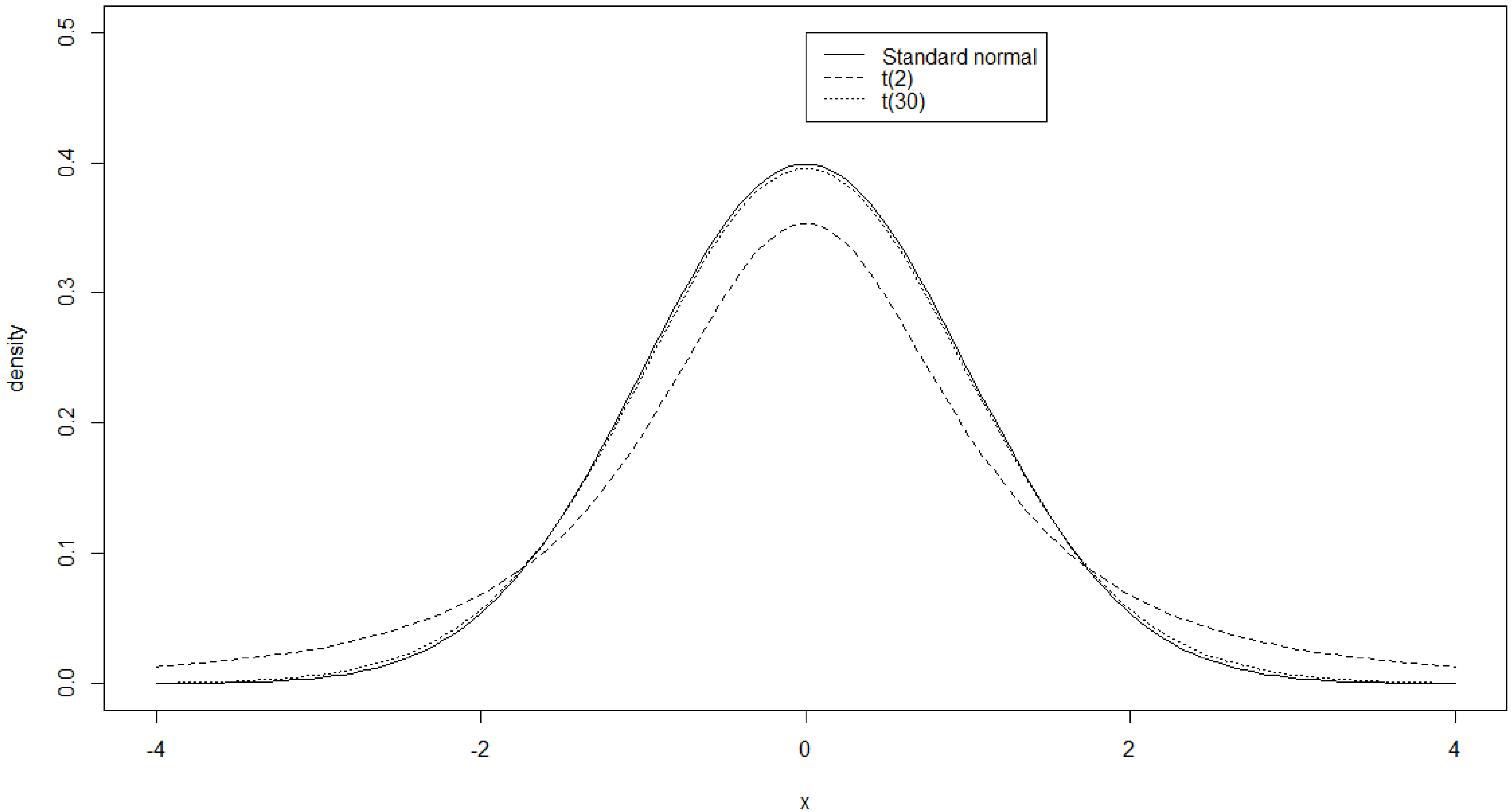
- When n: large, by CLT

  $\bar{X} \pm \boldsymbol{z_{\frac{\alpha}{2}}} \times \mathrm{SE}(\bar{X})$

t -분포 ➜ R code로 이동

```
> x <- seq(-4,4,by = 0.05)
> Std.Normal <- dnorm(x)
> t2 <- dt(x,df = 2)
> t30 <- dt(x,df = 30)
> plot(x,Std.Normal,type = "l",ylim = c(0,0.5),ylab = "density")
> lines(x,t2,lty = 2)
> lines(x,t30,lty = 3)
> legend(0,0.5,lty = 1:3,c("Standard normal","t(2)","t(30)"))
```

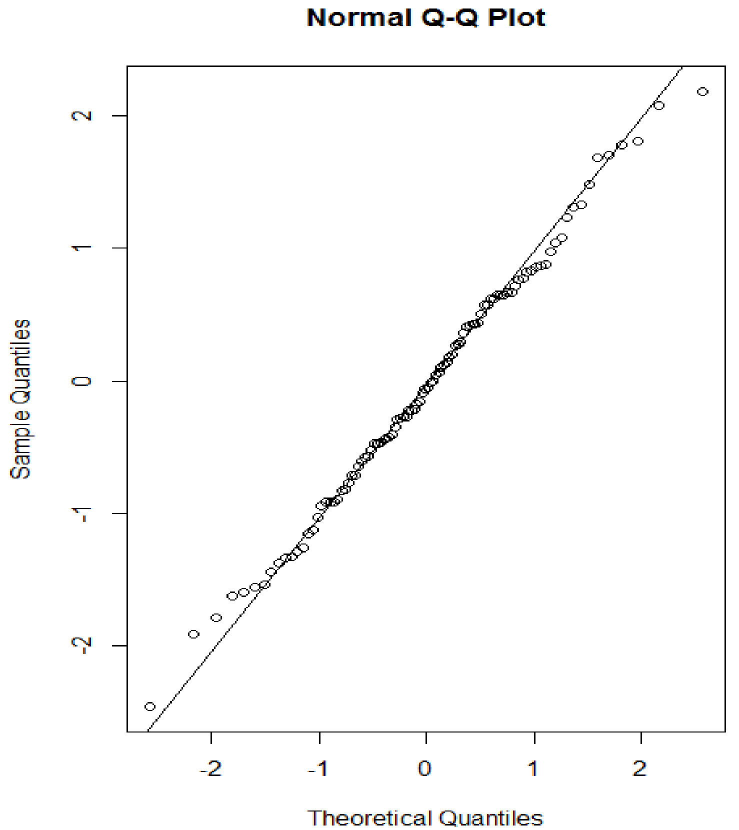
t –분포

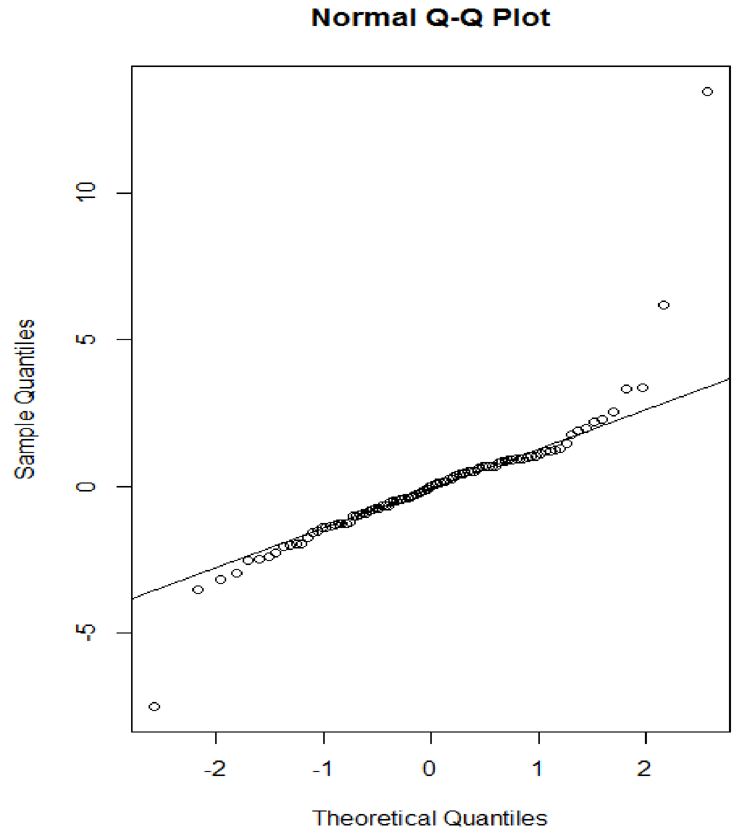


QQ plot of $N(0,1)$ and $t(2)$

```
> par(mfrow = c(1,2))
> r.normal <- rnorm(100)
> r.t <- rt(100,df = 2)
> qqnorm(r.normal); qqline(r.normal)
> qqnorm(r.t); qqline(r.t)
```

QQ plot of $N(0,1)$ and $t(2)$

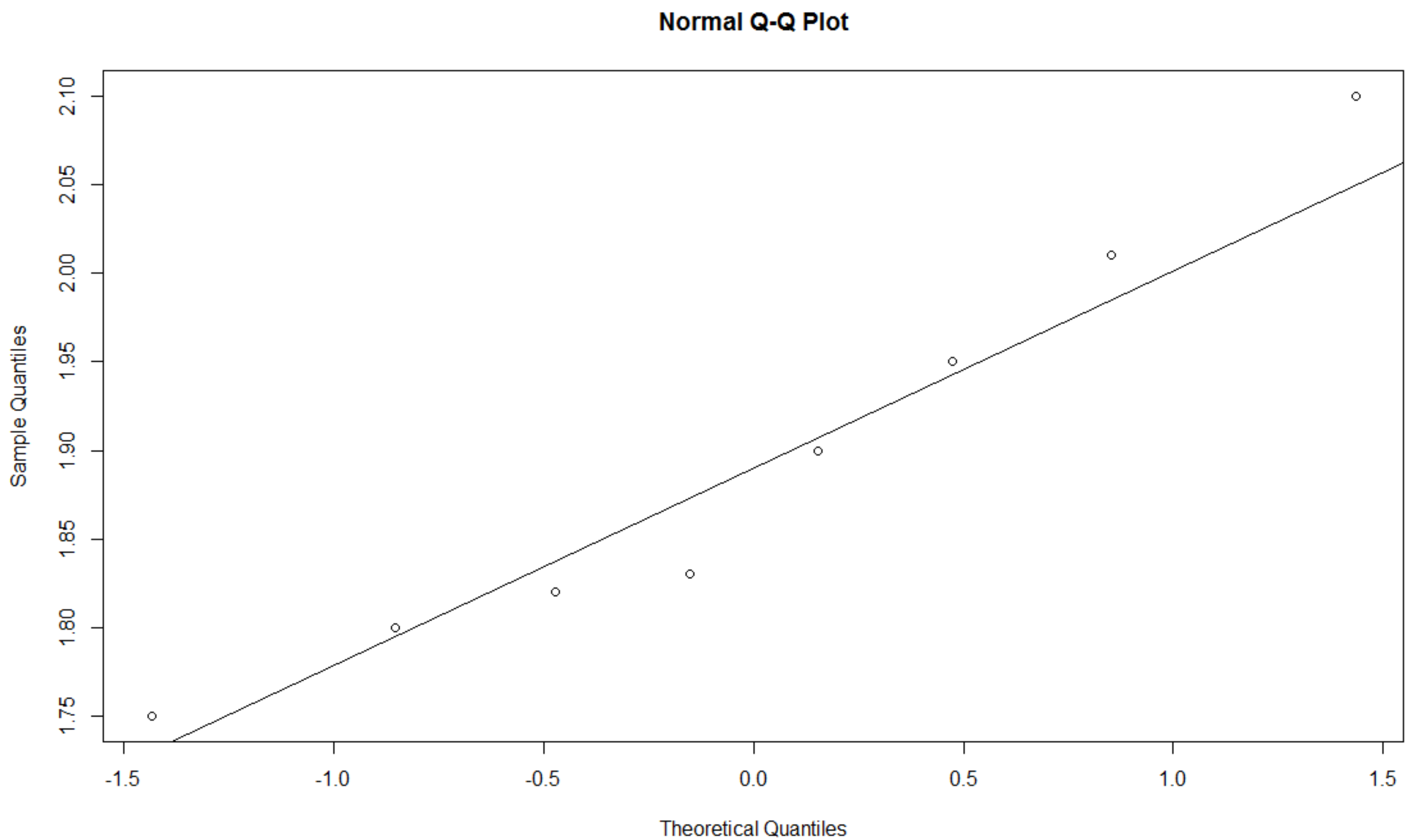




t.test() 함수 다루기

```
> ozs <- c(1.95,1.80,2.10,1.82,1.75,2.01,1.83,1.90)
> t.test(ozs,conf.level = 0.8)

        One Sample t-test

data:  ozs
t = 45.253, df = 7, p-value = 6.724e-10
alternative hypothesis: true mean is not equal to 0
80 percent confidence interval:
 1.835749 1.954251
sample estimates:
mean of x
    1.895

> qqnorm(ozs); qqline(ozs)  # approximately linear
```

t.test() 함수 다루기


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
1.75
1.80
1.85
1.90
1.95
2.00
2.05
2.10
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

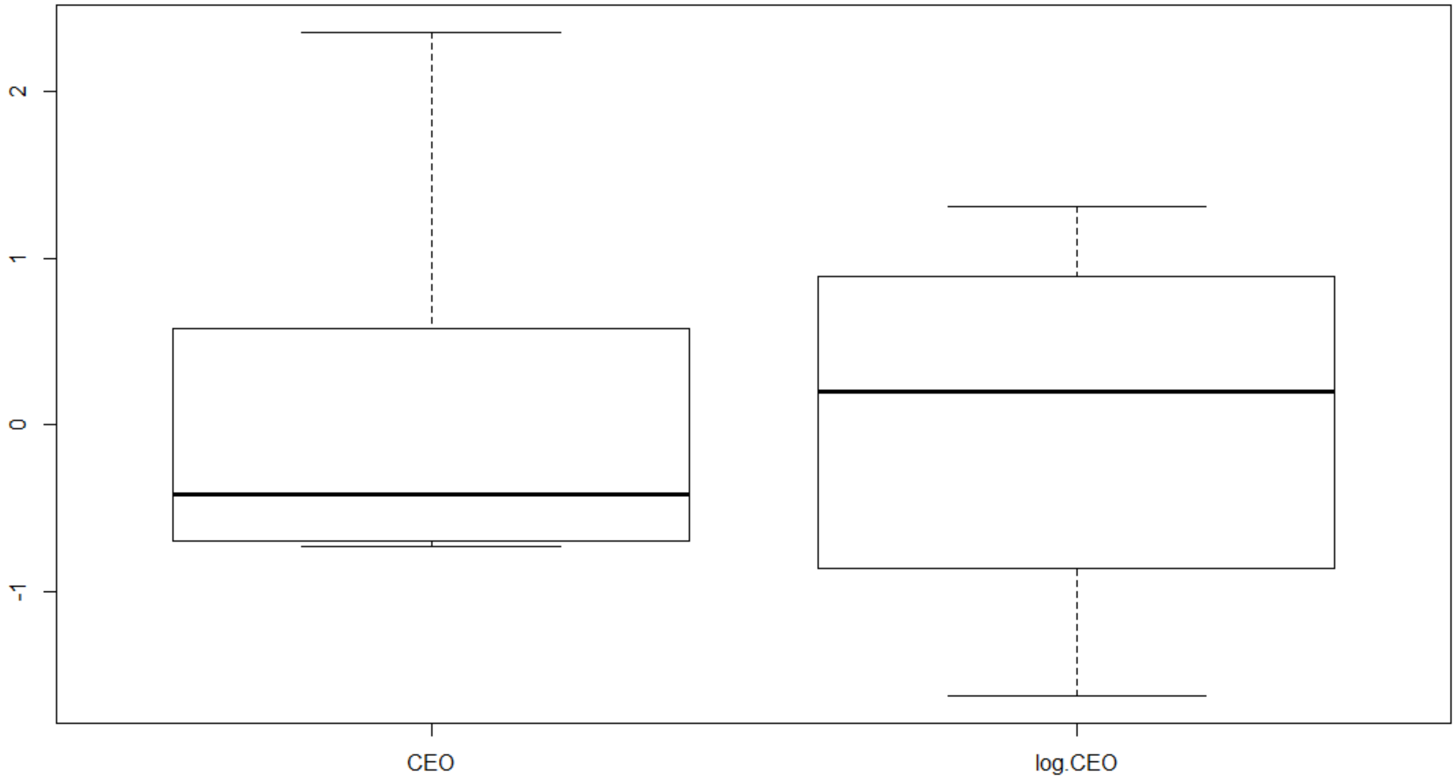
비모수 구간추정

- 언제 필요한가? 자료의 크기가 **작고**, 정규성에서 많이 **벗어날 때**
- wilcox.test() 함수 이용: 모집단의 분포가 **대칭성**을 만족할 때 타당!
- 대칭성을 **만족하지 못할** 때는 **sign test**에 기초한 구간추정 방법 이용

For the top 200 CEOs' pay in 2000 ➜ R code로 이동

```
> pay.00 <- c(110,12,2.5,98,1017,540,54,4.3,150,432)
> wilcox.test(log(pay.00),conf.int = TRUE,conf.level=0.9)

        Wilcoxon signed rank test

data:  log(pay.00)
V = 55, p-value = 0.001953
alternative hypothesis: true location is not equal to 0
90 percent confidence interval:
 2.963463 5.539530
sample estimates:
(pseudo)median
      4.344732

> exp(c(2.963,5,540))
[1]  1.935595e+01
[2]  1.484132e+02
[3] 3.303849e+234
> boxplot(list(scale(pay.00),scale(log(pay.00))),names=c("CEO","log.CEO"))
```

For the top 200 CEOs' pay in 2000



가설검정

- Idea: 어떤 가설이 '옳다'는 것을 증명하기는 어려우나 **'틀리다' 는 것을 증명**하기는 쉬움.
  그 이유는 틀린 사례를 찾으면 되기 때문에!

대립가설(alternative hypothesis)	귀무가설(null hypothesis)
H_1	H_0
연구의 주된 목적이 되는 가설	목적과 반대되는 가설
'옳다'고 주장하고 싶은 가설	'틀리다'고 주장하고 싶은 가설

가설검정의 원리

- '가짜 가설(H_0)이 틀리다'는 것을 보임으로써 '주장하고 싶은 가설(H_1)이 옳다'는 것을 증명
- 'H_0가 옳다'고 가정한 후에 **모순**(contradiction)을 유도하는 방법

두 종류의 오류

- 검정 결과는 H_0을 기각(reject) 또는 기각 안함(accept)
- 두 종류의 오류가 있음

실제현상 / 검정결과	H_0이 사실	H_1이 사실
H_0을 기각 안함	옳은 결정	Type II Error
H_0을 기각	Type I Error	옳은 결정

유의수준과 검정력

- $P_I = P(H_0$을 기각 | $H_0)$ "false positive rate(위양률)"
- $P_{II} = P(H_0$을 기각 안 함| $H_1)$ "false negative rate(위음률)"

  $= 1 - P(H_0$을 기각| $H_1)$

  $= 1 -$ Power
- 좋은 검정법은 P_I을 작게 하고 P_{II}를 작게 하는 검정법
- P_I의 상한값을 정해 놓고 그 값을 만족하는 검정법 중에서 P_{II}를 작게 하는 검정법을 선택
- 유의수준(significance level, α): P_I의 상한값. $\alpha =$ 0.01, 0.05, 0.1

유의확률(P–값)

- **P–값**은 "H_0이 참일 때, 현재 관측된 값 이상으로 H_1을 지지할 결과가 나올 확률"이며, **작을수록** H_0 을 기각할 근거가 강해짐

유의성 검정 절차

- H_0과 H_1을 선택
  - 주장하고 싶은 가설은 H_1
- 유의수준 결정
- 자료 수집 및 검정통계량 계산
- P −값 계산
- 결론
  - P−값 $< \alpha \rightarrow$ **H_0를 기각**
  - P−값 $\geq \alpha \rightarrow H_0$를 기각하지 못함

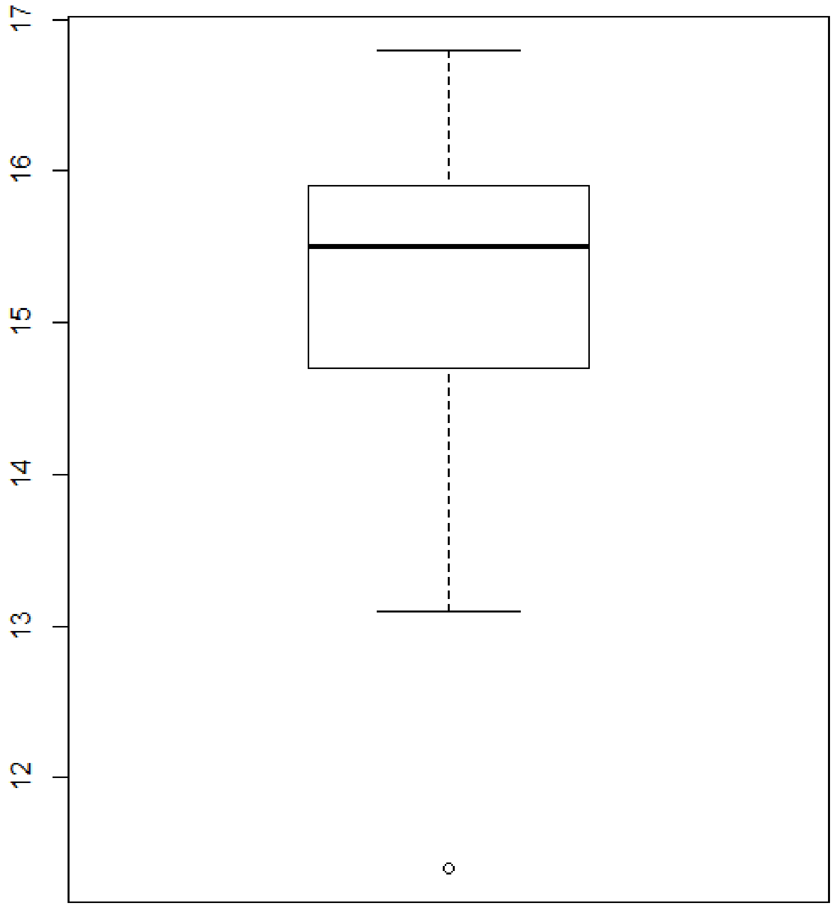
모평균에 대한 유의성 검정

- Hypothesis: $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$
- Test statistic: $T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
- Null distribution:
  - When normal population, $T \sim t(n-1)$
  - When large sample, $T \sim N(0,1)$
- $P-$값: $P(|T| \geq |t_0| | H_0)$

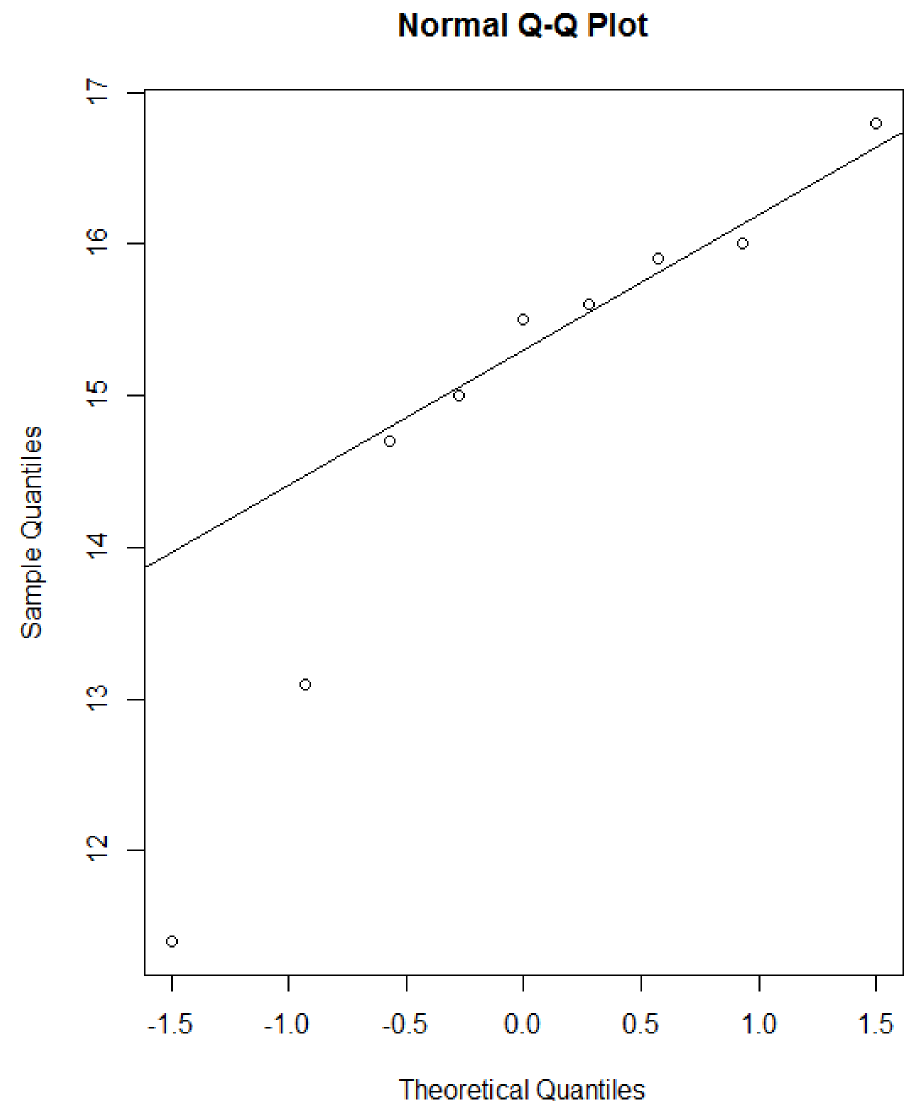
Does the actual mpg of a new SUV match the advertised 17mpg?

➜R code로 이동

```
> mpg <- c(11.4,13.1,14.7,15.0,15.5,15.6,15.9,16.0,16.8)
> t.test(mpg,mu = 17,alt = "less")

        One Sample t-test

data:  mpg
t = -3.8011, df = 8,
p-value = 0.002614
alternative hypothesis: true mean is less than 17
95 percent confidence interval:
     -Inf 15.92166
sample estimates:
mean of x
 14.88889

> par(mfrow = c(1,2))
> boxplot(mpg); qqnorm(mpg); qqline(mpg)
```



17
16
15
14
13
12



Normal Q-Q Plot
Sample Quantiles
17
16
15
14
13
12
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

비모수 방법: 부호순위검정

- Hypothesis: H_0: mediam $= m_0$ 대 H_1: median $\neq m_0$
- Test statistic
  - $W^+ = \sum_{i:X_i>m_0} \text{rank}\,(|X_i - m_0|)$
- Null distribution: Follows a discrete distribution
- $P-$값: $2 \times P(W^+ \geq \max(w_0^+, 2M - w_0^+)|H_0)$
  - $M = \frac{n(n+1)}{4}$

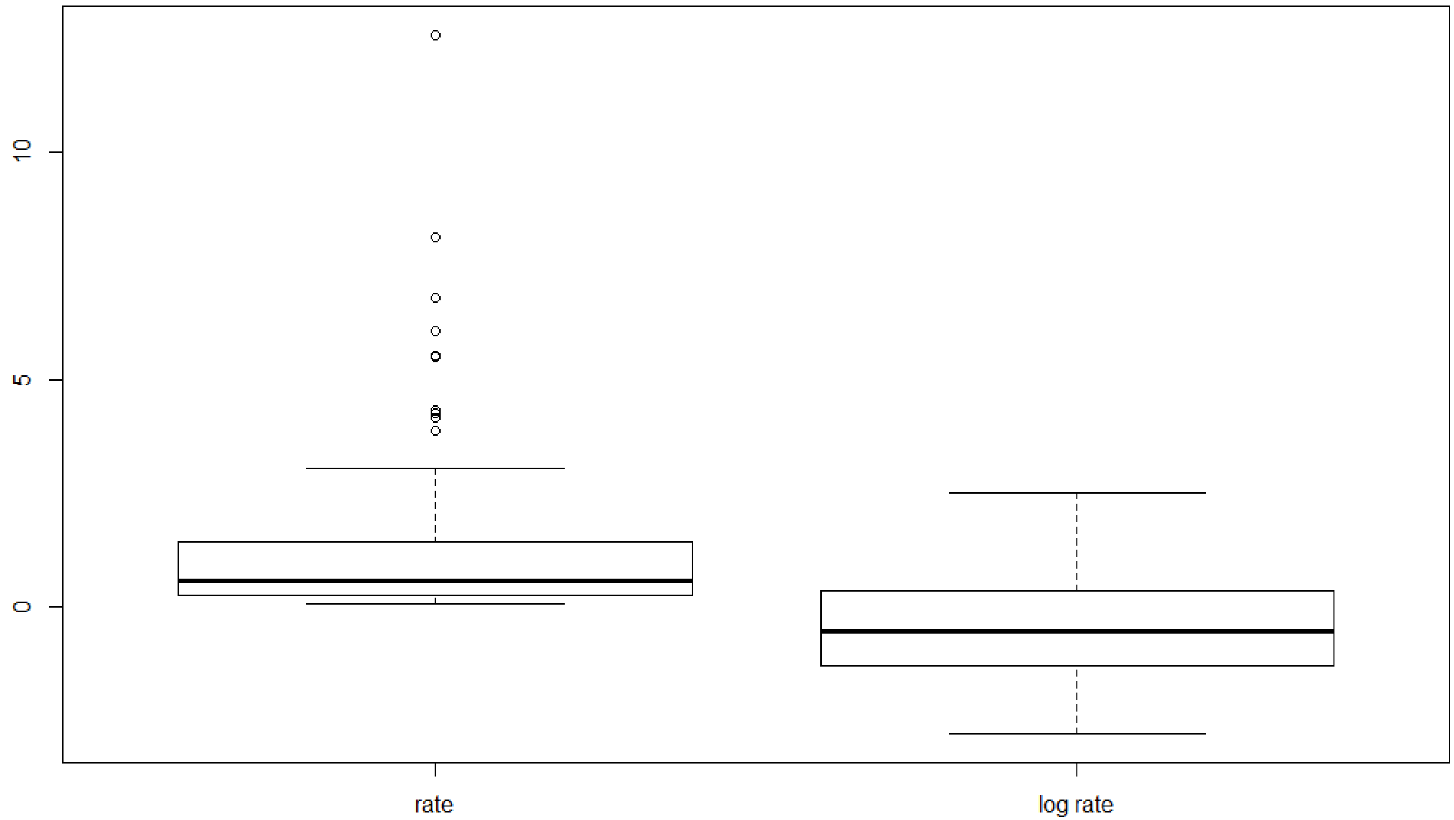
Signed rank test for the number of recruitments ➜ R code로 이동

```
> library(UsingR)
> salmon.rate
 [1] 0.0032957090 0.0045423985 0.0013499171 0.0053712171 0.0007302881 0.0079389928
 [7] 0.0006189738 0.0678289591 0.0118366780 0.0007570559 0.0416221059 0.0549750430
[13] 0.0012448352 0.0119910491 0.0060911195 0.0020834481 0.0016403706 0.0078468743
[19] 0.0009929378 0.0032190597 0.0119298698 0.1257444561 0.0012397337 0.0052303656
[25] 0.0027723679 0.0027207213 0.0058788433 0.0032394675 0.0607477124 0.0277212451
[31] 0.0051846602 0.0019769926 0.0129629634 0.0071289647 0.0069694656 0.0305683227
[37] 0.0010909469 0.0219024030 0.0019444685 0.0025799307 0.0067309012 0.0181346222
[43] 0.0021727548 0.0115980875 0.0030833961 0.0040328250 0.0015227455 0.0389437874
[49] 0.0033940044 0.0182083964 0.0015692376 0.0017217865 0.0019702113 0.0079029624
[55] 0.0049259929 0.0089105039 0.0115364575 0.0240510046 0.0009342104 0.0206367805
[61] 0.0085526033 0.0068403764 0.0262958709 0.0031215829 0.0040817614 0.0104582401
[67] 0.0811382070 0.0039190238 0.0032196817 0.0551593492 0.0160465372 0.0019534335
[73] 0.0426852284 0.0029152562 0.0074490351 0.0054607565 0.0213064378 0.0079174646
[79] 0.0432410779 0.0084867735 0.0216046367 0.0052675222 0.0048499662
```

Signed rank test for the number of recruits

```
> wilcox.test(log(salmon.rate),mu = log(0.005),alt = "greater")

        Wilcoxon signed rank test
        with continuity
        correction

data:  log(salmon.rate)
V = 2077, p-value = 0.065
alternative hypothesis: true location is greater than -5.298317

> boxplot(list(salmon.rate *100,log(salmon.rate * 100)),
+         names=c("rate","log rate"))
```

Signed rank test for the number of recruits



정규성 검토

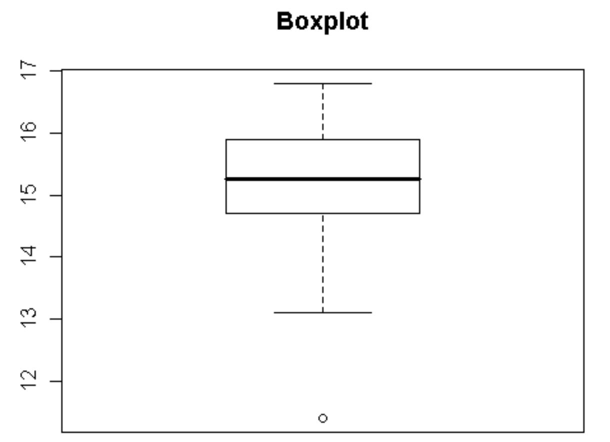
- Graph
  - 줄기-잎-그림: 대칭적?
  - 상자그림: 대칭적?
  - Normal probability plot: 직선?
- Goodness-of-fit 검정
  - H_0: 자료가 정규분포를 따른다
  - P-값 > 0.05 → H_0 기각 안함

실습: 정규성 검토

→ R code로 이동

- 줄기-잎-그림

```
> stem(ex4)

  The decimal point is at the |

  10 | 4
  12 | 1
  14 | 770569
  16 | 08
```

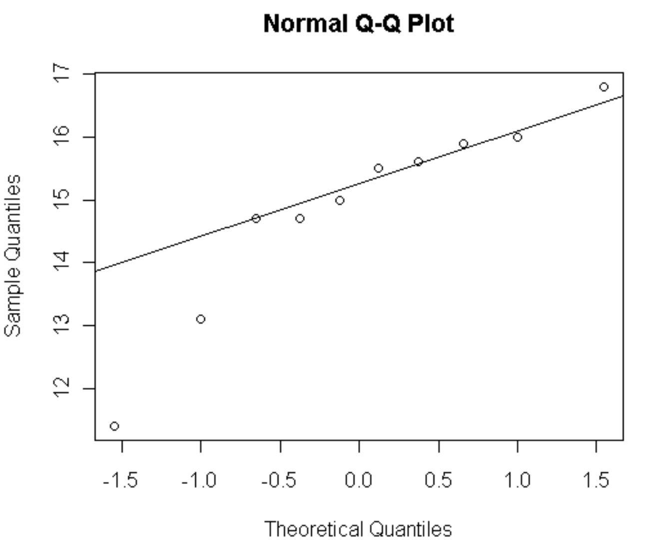


- 상자그림

```
> boxplot(ex4, main="Boxplot")
```

- 정규확률도

```
> qqnorm(ex4); qqline(ex4)
```



실습: 정규성 검토

- 샤피로-윌크(Shapiro-Wilk) 정규성 검정

```
> shapiro.test(ex4)

        Shapiro-Wilk normality test

data:  ex4
W = 0.89071, p-value = 0.1727
```

비정규성 자료

➜ R code로 이동

- Box–Cox 변환: $g(y) = I(\lambda \neq 0)\frac{y^{\lambda}-1}{\lambda} + I(\lambda = 0)\log y$

```
> library(MASS)
> bc <- boxcox(ex4 ~ 1, lambda = seq(-6, 6))
> lambda <- bc$x[which.max(bc$y)]
> lambda
[1] 3.69697
> bcex4 <- (ex4 ^ lambda - 1) / lambda
> shapiro.test(bcex4)

        Shapiro-Wilk normality test

data:  bcex4
W = 0.94025, p-value = 0.5558
```

- 비모수 검정

모비율에 대한 점추정

- 모수: $p \in (0,1)$
- Data: n번의 베르누이 시행에서 성공한 횟수 (X)
- 점추정량: $\hat{p} = \frac{X}{n}$
- 표준오차: $\mathrm{SE}(\hat{p}) = \sqrt{\mathrm{Var}(\hat{p})} = \frac{p(1-p)}{n}$
  - Why? $X \sim B(n,p)$
  - $p \leftarrow \hat{p}$

모비율에 대한 구간추정

- $100 \times (1-\alpha)\%$ 신뢰구간: $\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

prop.test() 함수 다루기

➜ R code로 이동

```
> prop.test(466,1013,conf.level = 0.95)

        1-sample proportions test with continuity correction

data:  466 out of 1013, null probability 0.5
X-squared = 6.3179, df = 1, p-value = 0.01195
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4290475 0.4912989
sample estimates:
        p
0.4600197
```

Exact CI for p

```
> binom.test(466,1013,conf.level = 0.95)

	Exact binomial test

data:  466 and 1013
number of successes = 466, number of trials = 1013, p-value =
0.01192
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4289889 0.4912836
sample estimates:
probability of success
             0.4600197
```

모비율에 대한 유의성 검정

- Hypothesis: $H_0: p = p_0$ vs. $H_1: p \neq p_0$
- Test statistic: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- Null distribution: When n: large, $Z \sim N(0,1)$
- $P-$값: $P(|Z| \geq |z_0| \,|\, H_o)$

Does the figure of the year-2001 show an increase from 11.3%?

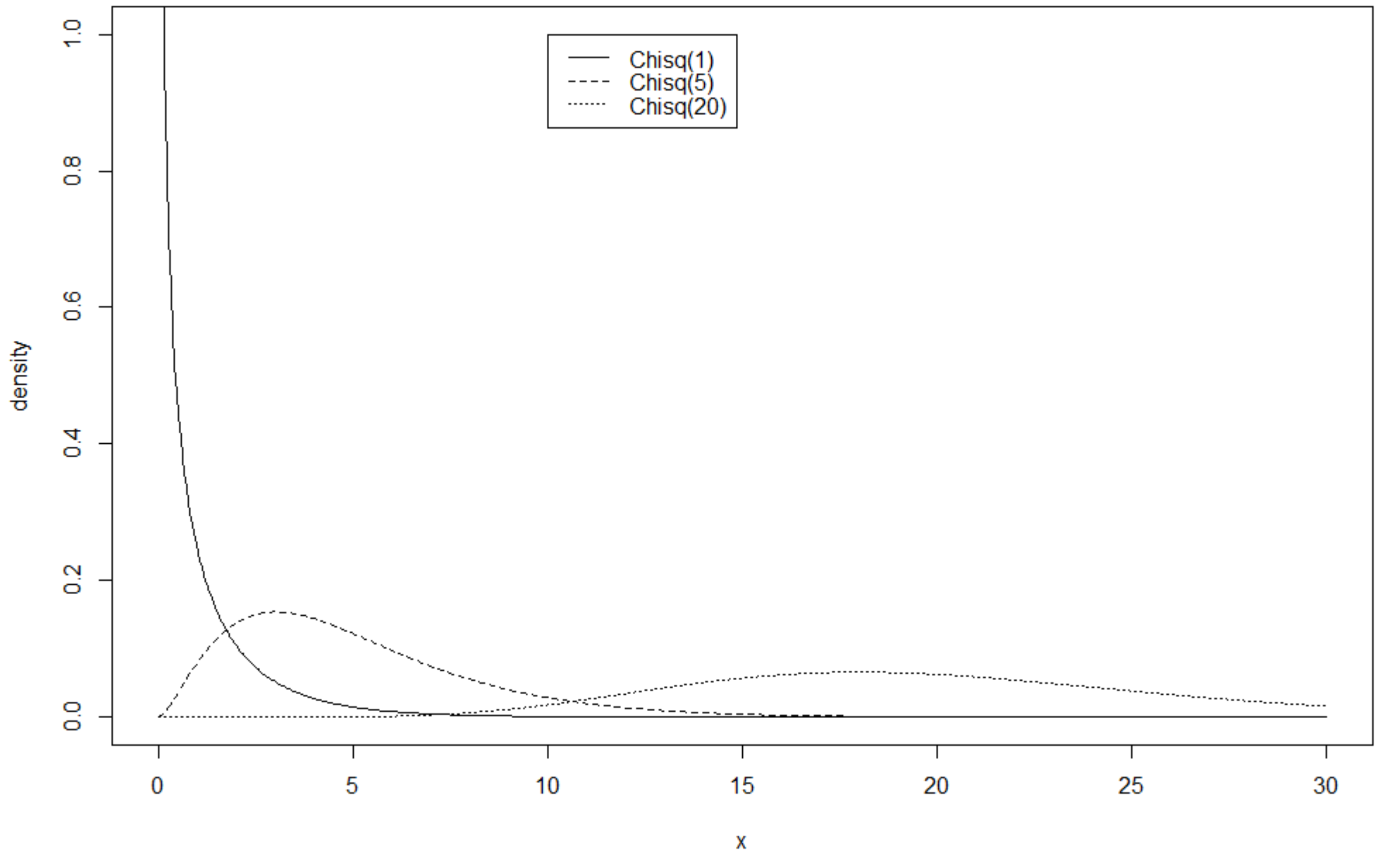
→ R code로 이동

```
> prop.test(x=5800,n=50000,p=0.113,alt="greater")

        1-sample proportions test with continuity correction

data:  5800 out of 50000, null probability 0.113
X-squared = 4.4597, df = 1, p-value = 0.01735
alternative hypothesis: true p is greater than 0.113
95 percent confidence interval:
 0.1136553 1.0000000
sample estimates:
    p
0.116
```

χ^2 –분포

```
> x <- seq(0,30,by = 0.05)
> chis1 <- dchisq(x,df = 1)
> chis5 <- dchisq(x,df = 5)
> chis20 <- dchisq(x,df = 20)
> plot(x,chis1,type = "l",ylim = c(0,1),ylab = "density")
> lines(x,chis5,lty = 2)
> lines(x,chis20,lty = 3)
> legend(10,1,lty = 1:3,c("Chisq(1)","Chisq(5)","Chisq(20)"))
```

χ^2 –분포


Chisq(1)
Chisq(5)
Chisq(20)
density
x
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

요약

- 추정
  - 점추정
  - 구간추정
- 가설검정의 원리
- 모평균에 대한 추론
  - 모수적 방법
  - 비모수적 방법
- 정규성 검토
- 모비율에 대한 추론