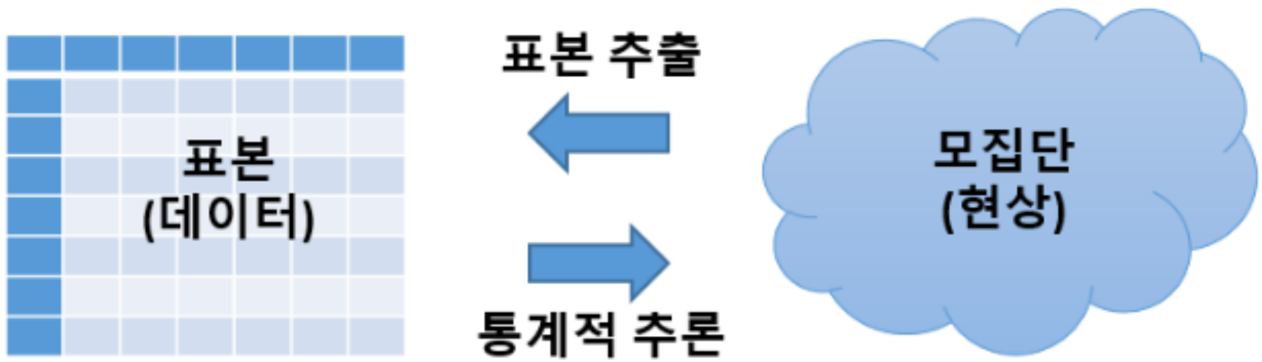
평균에 대한 추론

구간추정

통계적 추론



- 추정(estimation): 표본을 통해 모집단 특성이 어떠한가 추측
  - 미국 가계의 평균 신용카드 빚이 얼마인가?
  - 대선후보 A의 지지율이 얼마인가?

- 가설검정(testing hypothesis): 모집단 실제 값이 얼마나 되는가 하는 주장과 관련해서 표본이 가지고 있는 정보를 이용해 가설이 올바른지 판정
  - 2015년 미국 가계의 평균 신용카드 빚이 2014년에 비해 증가했다는 주장이 옳은가?
  - 국회의원 후보 A의 지지율이 0.15 이상이라는 주장이 옳은가?

평균에 대한 추론

- 관심이 되는 변수가 양적변수인 경우
  - 한 레스토랑에 오는 손님들의 평균 결제금액은 얼마인가?
  - 팁은 평균적으로 얼마나 주나?
  - 평균적으로 몇 명이 한 팀으로 식사를 하는가?

```
> summary(tips)
   total_bill         tip               sex      smoker      day         time          size     
 Min.   : 3.07   Min.   : 1.000   Female: 87   No :151   Fri :19   Dinner:176   Min.   :1.00  
 1st Qu.:13.35   1st Qu.: 2.000   Male  :157   Yes: 93   Sat :87   Lunch : 68   1st Qu.:2.00  
 Median :17.80   Median : 2.900                          Sun :76                Median :2.00  
 Mean   :19.79   Mean   : 2.998                          Thur:62                Mean   :2.57  
 3rd Qu.:24.13   3rd Qu.: 3.562                                                 3rd Qu.:3.00  
 Max.   :50.81   Max.   :10.000                                                 Max.   :6.00  
```

모집단 평균의 구간추정

- 244명의 레스토랑 손님의 평균 결제금액은 19.79불 (점추정치)
  - 전체 손님의 평균도 19.79불?
  - 다른 표본이 추출되면?
- 점추정치를 기준으로 일정 구간을 만들어 그 구간 안에 모수가 포함될 가능성을 높이는 것

- 구간추정값

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$1-\alpha$ = 신뢰계수
$t_{\alpha/2}$ = 자유도 n-1 을 가지는 t 분포의 오른쪽 꼬리 $\alpha/2$ 에 해당하는 값
s = 표본 표준편차

```
> t.test(tips$total_bill,conf.level=0.95)

        One Sample t-test

data:  tips$total_bill
t = 34.7171, df = 243, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 18.66333 20.90855
sample estimates:
mean of x
 19.78594
```

- 평균 결제금액에 대한 95% 신뢰구간=(18.66, 20.91)
- 95% 신뢰수준에서 평균 결제금액이 18.66과 20.91 사이에 들어갈 것이라고 추정
  - 신뢰수준이 올라간다면?
  - 데이터의 개수가 많아진다면?

평균에 대한 추론

단일집단 모평균에 대한 검정
(One sample t-test)

One-sample T-Test

- 모집단의 평균이 어떤 특정한 값과 같은지를 검증

예) 2006년 조사에 의하면 한국인의 1인1일 평균 알코올 섭취량은 8.1g이다. 2008년 대통령 선거로 알코올 섭취량이 **달라졌는지** 조사하기 위해 10명을 무작위로 뽑아서 조사한 결과 다음과 같은 데이터를 얻었다.

15.5,11.21,12.67,8.87,12.15,9.88,2.06,14.5,0,4.97

```
> rm(list=ls()) #workspace 지우기
> x=c(15.5,11.21,12.67,8.87,12.15,9.88,2.06,14.5,0,4.97)
```

1. 귀무가설 대립가설 설정

H_0: 2008년의 평균 알콜 섭취량이 8.1g이다.

⇔ H_0: $\mu = 8.1$

H_a: 2008년의 평균 알콜 섭취량이 8.1g이 아니다.

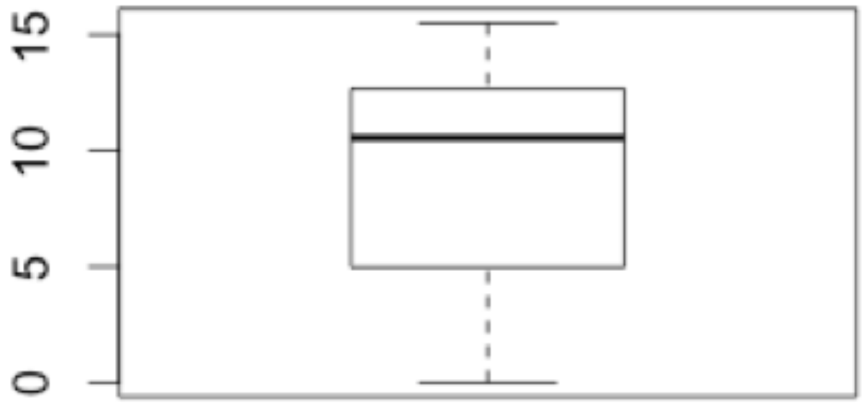
⇔ H_a: $\mu \neq 8.1$

2. 가정체크

- **가정**
  - **자료가 정규분포를 따른다.**
  - **심하게 편중되거나 극단치를 포함한 경우 표본수가 50개 (혹은 30개) 이상이다.**

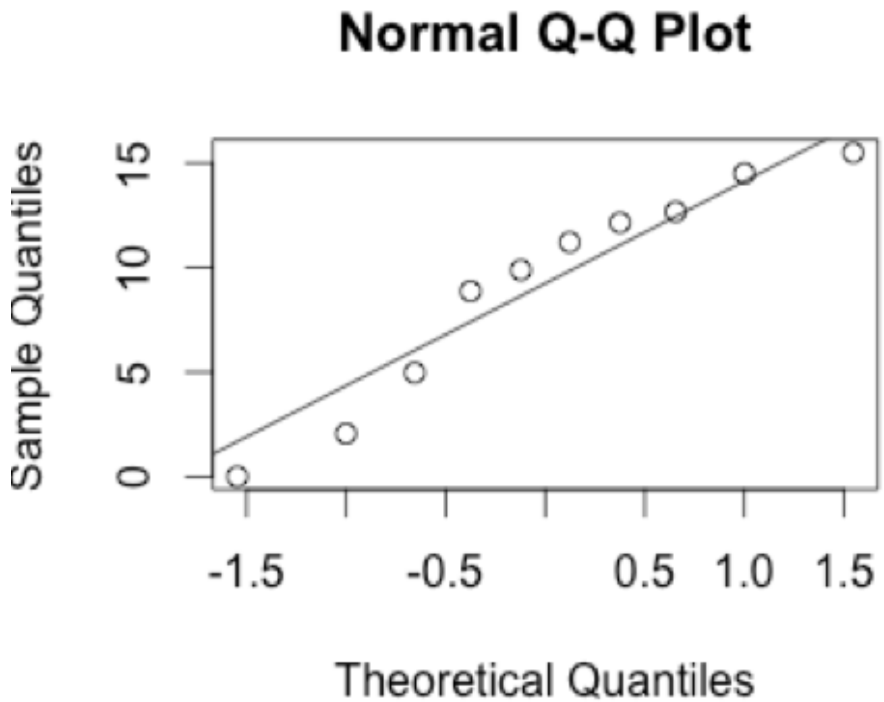
- Boxplot 또는 Histogram

```
> boxplot(x)
> hist(x)
```



- 정규확률도 (Q-Q plot)

```
> qqnorm(x)
> qqline(x)
```



- Shapiro-Wilk normality test

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9234, p-value = 0.3863
```

H_0: 데이터가 정규분포를 따른다.
H_a: 데이터가 정규분포를 따르지 않는다.

- P-value>0.05 ➔ 데이터가 정규분포를 따른다고 결정 => t-test 진행
- P-value<0.05 ➔ 데이터가 정규분포를 따르지 않는다 => 관측수가 충분히 크지 않으면 다른 Test 고려 (Shapiro-Wilks test)

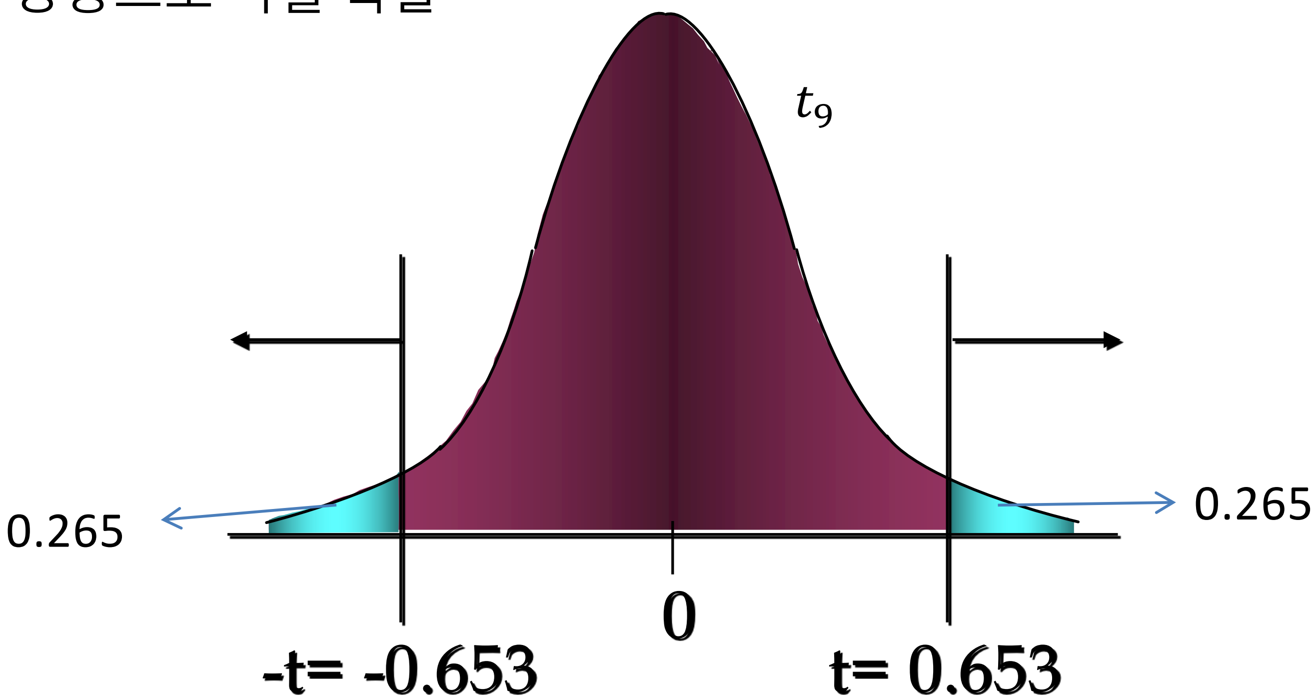
3. 검정통계량 계산: T-statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

```
> t.test(x,mu=8.1)

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.5301
alternative hypothesis: true mean is not equal to 8.1
95 percent confidence interval:
  5.436132 12.925868
sample estimates:
mean of x
    9.181
```

3. P-value 계산

모집단에서 표본추출을 반복했을 때 검정통계량들이 0.653보다 더 대립가설을 지지하는 방향으로 나올 확률



$$\Pr(|t_9| > 0.653) = 2 \times \{1 - \Pr(t_9 < 0.653)\}$$

```
> 2*(1-pt(0.653, df=9)) #p-value
[1] 0.5300818
```

3. 검정통계량과 P-value 계산 (t.test)

```
> t.test(x,mu=8.1)

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.5301
alternative hypothesis: true mean is not equal to 8.1
95 percent confidence interval:
  5.436132 12.925868
sample estimates:
mean of x
    9.181
```

T-statistics

P-value
: α보다 작으면
귀무가설 기각

대립가설

95% 신뢰구간
: 실제 알콜섭취 평균량이 이
구간 안에 포함될 것을 95%
만큼 확신한다.

점추정량
:실제 알콜섭취 평균량을
9.181로 추정한다.

4. 결론

- P-value=0.5301 > 0.05

➔귀무가설을 기각하지 못한다.

대선 후 알콜섭취 평균량이 평상시 평균량과 다르다고 할 수 없다.

양측검정 vs 단측검정

- 대선 이후 알콜 섭취량 **<u>많아졌는지</u>** 검정
- $H_0: \mu = 8.1 \ vs \ H_a: \mu > 8.1$

```
> t.test(x,mu=8.1,alter="greater")

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.265
alternative hypothesis: true mean is greater than 8.1
95 percent confidence interval:
 6.146389      Inf
sample estimates:
mean of x
    9.181
```

대립가설

- $H_0: \mu = 8.1 \ vs \ H_a: \mu < 8.1$

```
> t.test(x,mu=8.1,alter=“less")
```

p-값

- 귀무가설이 옳다면 이런 데이터 패턴이 우연히 관찰될 가능성은 얼마나 되는가?
  - 위약을 투약한 환자의 49%가 호전됐는데 신약을 투약한 환자의 91%가 호전
    - 신약이 심장병에 효과가 없는데(귀무가설) 위의 결과를 얻을 확률?
  - 자폐증을 가진 아이 59명의 뇌가 그렇지 않은 아이들 38명에 비해 최대 10% 이상 큼
    - 두 집단 아이들의 뇌 크기에 실제로 차이가 없는데 두 표본 집단에서 뇌 크기의 차이가 관찰될 확률?
- P-값이 작으면($<\alpha$) 귀무가설이 틀렸을 거라고 합리적으로 의심
  - 얼마나 작으면 기각??

유의수준 (α)

- 유의수준 : 귀무가설이 맞는데 귀무가설을 기각할 오류(제 1종 오류)를 범할 확률
  - 일반적으로 0.05, 0.01, 0.1 등의 값으로 사전에 결정
  - 귀무가설을 기각하기 위한 p-값의 임계치로 사용
    - p-값< α: 귀무가설 기각
      - 신약의 효과가 위약에 비해 통계적으로 유의하게 다름
    - p-값> α: 귀무가설을 기각할 수 없음
      - 신약과 위약의 효과가 다르다는 통계적으로 충분한 증거가 없음

제 1종 오류와 제 2종 오류

- 제 1종 오류와 제 2종 오류의 상충
  - α가 너무 크면 효과 없는 약을 있다고 판단 (제 1종 오류)
  - α가 너무 작으면 수많은 효과 있는 약을 승인 안함 (제 2종 오류)

- 어느 오류를 줄이는 것이 더 중요?
  - 스팸필터
  - 암 검사

유의수준 (α) 조정

- default: $\alpha = 0.05$

```
> t.test(x,mu=8.1,conf.level=0.99)

        One Sample t-test

data:  x
t = 0.653, df = 9, p-value = 0.5301
alternative hypothesis: true mean is not equal to 8.1
99 percent confidence interval:
  3.801088 14.560912
sample estimates:
mean of x
    9.181
```

99% 신뢰구간
: 실제 알콜섭취
평균량이 이 구간 안에
포함될 것을 99% 만큼
확신한다.

평균에 대한 추론

두 모집단 평균에 대한 검정:
독립표본 T-검정

Two-sample T-test

예: 뼈세포수

- 줄기세포를 활용하여 배양치아를 제작하려 한다. 다른 두 조건 (control, test)에 표본을 무작위로 할당하고 배양된 뼈세포의 수(resp)를 측정하였다. 두 조건에서의 평균 뼈세포의 수가 다른지 확인하고 싶다.

```
> rm(list=ls())
> dental=read.csv("dental.csv")
> dental
   treatment resp
1       test  148
2       test  190
3       test   68
4       test   79
5       test   70
6    control   40
7    control   80
8    control   64
9    control   52
10   control   45
```

관심사인 양적변수(resp)와 두 그룹을 정의하는 범주형변수(treatment)가 입력되어야 함

가설 수립

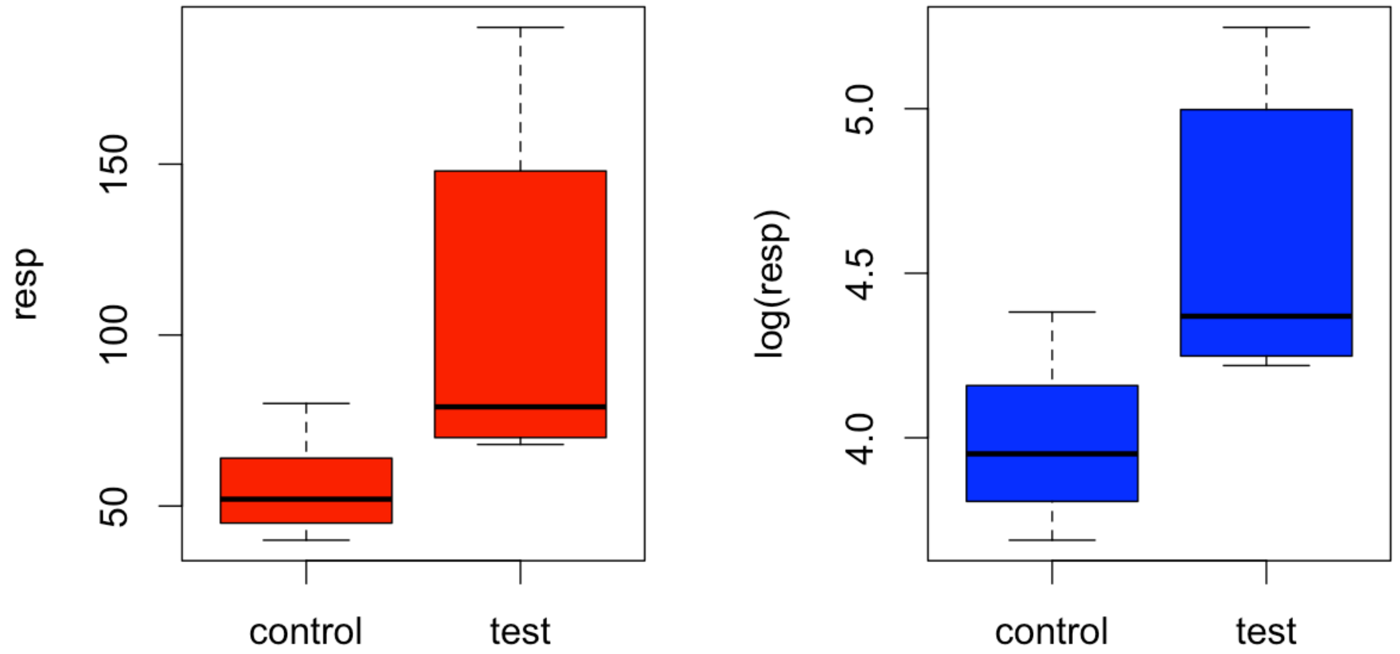
$$H_0 : \mu_1 \qquad \mu_2$$

$$H_a : \mu_1 \qquad \mu_2$$

독립표본 T-검정의 가정

- 두 집단 모두 정규분포를 따른다
  - Boxplot, histogram, Q-Q plot 등 을 통해 체크
- 정규분포를 따르지 않더라도 관측치 수가 충분히 많다면 ($n_1 + n_2$>30) 일반적으로 독립표본 T검정을 사용할 수 있다.

뼈세포수: 가정체크 (Boxplot)



- Test 군의 분산이 control 군의 분산보다 크다.
- 두 그룹 모두 편향된 분포를 가지고 있다.
- Log 변환 후 분산 차이가 좁혀졌고 분포의 편향도도 작아졌다.

Log 변환된 변수로 분석진행!

```
par(mfcol=c(1,2))
boxplot(resp~treatment, data=dental,col="red",ylab="resp")
boxplot(log(resp)~treatment, data=dental,col="blue",ylab="log(resp)")
```

검정통계량: T-statistics

- Recall: One Sample T-test

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

자료의 평균 $\bar{x}$가 모집단의 평균 μ 로부터 떨어진 상대적인 거리

표본추출을 무수히 반복했을 때 $\bar{x}$ 가 μ로부터 얼마나 흩어져있는지의 정도

- Two Sample T-test

$$T = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{Var(\overline{x_1} - \overline{x_2})}}$$

검정 통계량: T-statistics

$$\boldsymbol{Var(\overline{x_1} - \overline{x_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- $\boldsymbol{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$을 어떻게 추정할 것인가?
  ➔두 집단의 분산이 같다는 정보가 있으면 $\sigma_1 = \sigma_2$로 놓고 추정
  ➔두 집단의 분산이 다르다는 정보가 있으면 각각 추정

var.test() 이용해 두 집단의 분산이 같은지 검정

➔다르면 t.test ()

➔같으면 t.test(, var.equal=TRUE)

등분산 검정

$$H_0: \sigma_1^2 = \sigma_2^2 \quad vs. \quad H_a: \sigma_1^2 \neq \sigma_2^2$$

```
> var.test(log(resp)~treatment,data=dental)

        F test to compare two variances

data:  log(resp) by treatment
F = 0.3432, num df = 4, denom df = 4, p-value = 0.325
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.03573413 3.29636586
sample estimates:
ratio of variances
         0.3432095
```

T-statistics, p-value: 분산이 같은 경우

$$H_0: \mu_1 = \mu_2 \quad vs. \quad H_a: \mu_1 \neq \mu_2$$

```
> t.test(log(resp)~treatment,var.equal=TRUE,data=dental)

          Two Sample t-test

data:  log(resp) by treatment
t = -2.5217, df = 8, p-value = 0.03571
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.18465764 -0.05293907
sample estimates:
mean in group control    mean in group test
             3.997539              4.616337
```

0포함 안함

결론

- p-value $=0.0357<0.05$
- 귀무가설을 기각할 수 있다.
- Control과 treatment 두 그룹의 뼈세포수의 평균이 유의수준 5%하에서 차이가 있다.

Log변환 전 자료로 가설검정을 한다면?

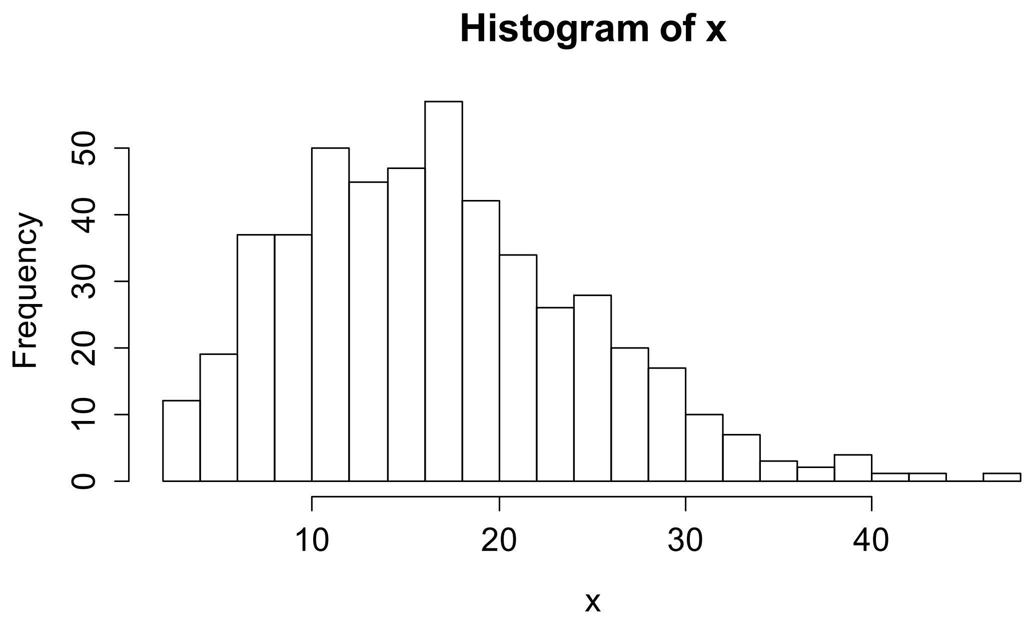
```
> var.test(resp~treatment,data=dental)

        F test to compare two variances

data:  resp by treatment
F = 0.0849, num df = 4, denom df = 4, p-value = 0.03483
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.008840233 0.815484912
sample estimates:
ratio of variances
        0.08490628

> t.test(resp~treatment,data=dental)

        Welch Two Sample t-test

data:  resp by treatment
t = -2.1333, df = 4.674, p-value = 0.08988
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -122.23919   12.63919
sample estimates:
mean in group control    mean in group test
                 56.2                 111.0
```

두 집단의 분산이 같다는 귀무가설을 기각

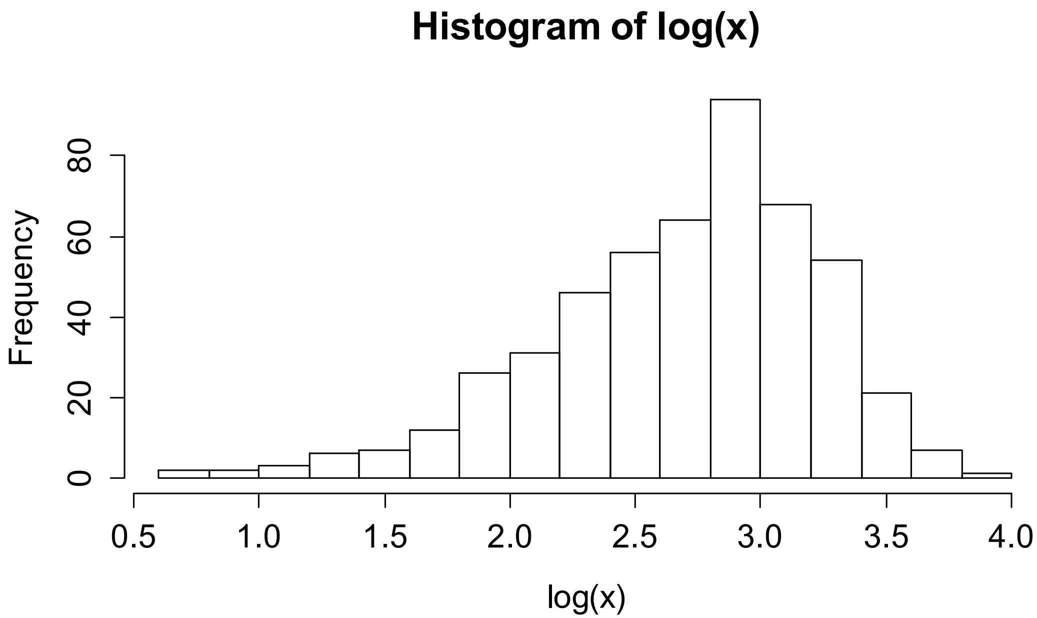
var.equal=TRUE 옵션 없이 이분산의 t-test

유의수준 5% 하에서 두 집단 평균의 차이가 없다.

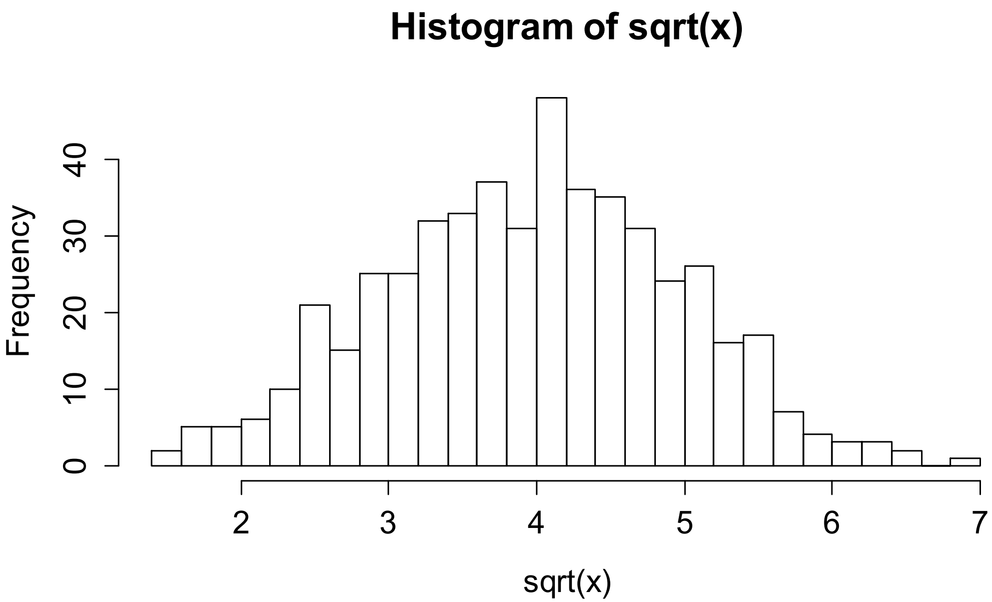
변수변환



- 많은 경우의 실제 데이터에서 오른쪽으로 꼬리가 긴 분포형태를 가진다. (소득, 키, 매출액 등)
- 여러 통계 기법들은 자료의 정규분포를 가정하므로 분석 전 변수변환이 필요한지 체크한다.




```
x=(rnorm(500)+4)^2
hist(x,20)
hist(log(x),20)
hist(sqrt(x),20)
```



평균에 대한 추론

두 모집단 평균에 대한 검정: 쌍체표본 T-검정

Paired T-test

쌍체표본 T-검정

- 쌍을 이룬 두 변수 (matched sample)의 차이를 보는 검정
- 한 집단을 대상으로 약의 복용, 치료, 교육방법 도입등
- 두 집단이더라도 쌍둥이 또는 부부처럼 변수들 간의 상관관계가 존재할 때

예)

- 10명의 비행기 조종사의 술먹기 전과 후의 특정작업에 대한 반응시간 비교

- 30쌍의 쌍둥이의 키 비교

독립표본 T-검정 vs.쌍체표본 T-검정

- 남성과 여성의 임금 차이를 조사한 코넬대학교 연구에 의하면, 남성의 임금이 여성의 임금에 비하여 높은 이유는 여성에 비해 높은 경력을 가졌기 때문인 것으로 보고되었다. 남성과 여성의 경력 차이를 비교하였다.

- 최근에 소비자들의 여가시간에 대한 매체 간의 경쟁관계가 격해지고 있다. 조사자들은 15명의 개인 자료를 활용하여 주간 케이블 TV시청시간과 주간 라디오 청취시간에 대한 자료를 수집하였다.

- 대학본부에서는 부모의 최종학력에 따른 응시자들의 SAT점수 차이를 비교하였다. 첫 번째 표본에서는 학부모들이 대학에서 학사학위를 취득한 집단의 SAT점수를 취하였다. 두 번째 표본에서는 부모들이 고등학교만 졸업한 집단의 SAT 점수를 취하였다.

예: 거식증 치료제

- 거식증치료제 FT복용 전후의 체중변화를 측정하여 FT가 체중증가에 영향이 있는지 조사
- Prewt: 복용 전 체중
- Postwt: 복용 후 체중

```
> FT=read.csv("FT.csv")
> FT
   Treat Prewt Postwt
1     FT  83.8   95.2
2     FT  83.3   94.3
3     FT  86.0   91.5
4     FT  82.5   91.9
5     FT  86.7  100.3
6     FT  79.6   76.7
7     FT  76.9   76.8
8     FT  94.2  101.6
9     FT  73.4   94.9
10    FT  80.5   75.2
11    FT  81.6   77.8
12    FT  82.1   95.5
13    FT  77.6   90.7
14    FT  83.5   92.5
15    FT  89.9   93.8
16    FT  86.0   91.7
17    FT  87.3   98.0
```

1. 귀무가설 대립가설 설정

H_0:

H_a:

2. 가정체크

- One-sample T-test 의 가정과 동일
- **<u>Postwt-Prewt</u>** 이 정규분포를 따르는지 확인
- Shpiro-Wilk test 사용

```
> with(FT, shapiro.test(Postwt-Prewt))

        Shapiro-Wilk normality test

data:  Postwt - Prewt
W = 0.9536, p-value = 0.5156
```

이 가설에 대한 p-value

H0: FT 복용전후의 차이가 정규분포를 따른다.
Ha: FT 복용전후의 차이가 정규분포를 따르지 않는다

3. 검정통계량과 p-value계산

```
> with(FT, t.test(Postwt-Prewt),alternative="greater")

        One Sample t-test

data:  Postwt - Prewt
t = 4.1849, df = 16, p-value = 0.0007003
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  3.58470 10.94471
sample estimates:
mean of x
 7.264706
```

4. 결론

- P-value=0.0007<0.05
- 귀무가설을 기각한다.
- FT 복용 후 통계적으로 유의한 체중증가가 있다.

A/B 테스트 (임의화 비교실험)

- 과거 3개월 간 DM발송 유무에 따른 평균 구매액의 차이
  - 23000원 vs. 18000원
- Optimizely (댄 시로커: 오바마의 선거참모)
  - 오바마닷컴을 방문한 유권자를 대상으로 어떤 그림이나 메세지를 노출하느냐에 따라 선호도가 어떻게 달라지는지 측정

A/B 테스트 (임의화 비교실험)

	상품구매	상품 비구매	합계
기존 디자인	9500명 (9.5%)	90500명 (90.5%)	10만명
새 디자인	9600명 (9.6%)	90400명 (90.4%)	10만명

- 새 디자인 상품구매 0.1% 상승 (1.01배)
- 의미있는 차이인가? 오차인가?

- **P-value: 실제는 아닌데도 오차나 우연에 의해 데이터와 같은 차이 (정확히는 그 이상의 극단적인 차이를 포함)가 생길 확률**