Paired T-test

Paired T-test

- 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

예 : Express Deliveries

두 택배사의 배송시간 검정을
위해, 이 회사는 전국 지점들 중
무작위로 선정해서
두 개의 서류 중 하나는 UPX 로
다른 하나는 INTEX로 보냈다.

다음 슬라이더에 있는 데이터로
두 택배회사의 평균배송시간에 차이가 있다고
할 수 있는가?

지점	배송시간 (시간) UPX	INTEX	배송시간차이
Seattle	32	25	7
Los Angeles	30	24	6
Boston	19	15	4
Cleveland	16	15	1
New York	15	13	2
Houston	18	15	3
Atlanta	14	15	-1
St. Louis	10	8	2
Milwaukee	7	9	-2
Denver	16	11	5

배송시간 차이=0 ⇔ UPX와 INTEX 사이의 차이 없다
배송시간 차이≠ 0⇔ UPX와 INTEX 사이의 차이 있다.

Paired T-test = 두 변수 차이를 사용한 One sample T-test

예) 거식증 치료제

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

```
> rm(list=ls()) #workspace 지우기
> setwd("c:/data")
> FT=read.csv("FT.csv")
>
```

1. 귀무가설 대립가설 설정

H_0:

H_a:

2. 가정체크

- One-sample T-test 의 가정과 동일
- **Postwt-Prewt** 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.

Two-sample T-test

예: Par사

Par사는 골프 용품 제조업체로써 훨씬 멀리 나가는 새로운 골프공을 개발하였다.

기계장치를 이용한 드라이빙거리 시험에서, Par사 골프공 표본을 경쟁사인 Rap 사의 골프공 표본과 비교하였다. 표본통계량이 다음 슬라이더에 나와 있다.

예: Par사

	표본 #1 Par사	표본 #2 Rap사
표본규모	120 개	80 개
표본평균	275 yards	258 yards

두 모집단 평균의 차이에 대한 추정

모집단 1
Par사 골프공

μ_1 = Par사 골프공의 평균 거리

모집단2
Rap사 골프공

μ_2 = Rap사 골프공의 평균거리

$\mu_1 - \mu_2$ = 두 평균거리의 차이

Par사 골프공의 단순무작위 표본수: n_1

x_1 = Par사 공의 표본평균거리

Rap사 골프공의 단순무작위 표본수: n_2

x_2 = Rap사 공의 표본평균거리

$\overline{x_1} - \overline{x_2}$ = $\mu_1 - \mu_2$ 의 점추정치

$\mu_1 - \mu_2$ 의 점추정치

$\mu_1 - \mu_2$ 의 점추정치=

= 275 - 258

= 17 yards

여기서:

μ_1 = Par사 골프공의 모집단의 평균거리

μ_2 = Rap사 골프공의 모집단의 평균거리

One-sample t-test vs. paired t-test

- 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: Par사

유의수준 a = .01에서,
이 회사의 골프공의
평균 드라이빙거리가
Rap사 골프공의
평균 드라이빙거리보다
더 멀다고 결론지을 수 있는가 ?

가설 수립

H_0:

H_a:

여기서:

μ_1 = Par사 골프공의모집단의 평균거리

μ_2 = Rap사 골프공의 모집단의 평균거리

Two-sample T-test 의 가정

- 두 집단 모두 정규분포를 따른다
  - Boxplot, histogram 등 을 통해 체크

- 정규분포를 따르지 않더라도 관측치 수가 충분히 많다면 ($n_1 + n_2$>30) 일반적으로 two-sample t-test를 사용할 수 있다.

검정통계량: T-statistics

- Recall: One Sample T-test

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

자료의 평균 $\bar{x}$가 모집단의 평균 μ로부터 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

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

- $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$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)

결론

- P-value가 유의수준 보다 작으면 귀무가설 기각
- P-value가 유의수준 보다 크면 귀무가설을 기각할 수 없음

t.test()의 결과물에서 p-value 확인

예: 뼈세포수

- 다른 두 조건 (control, test)에서 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
```

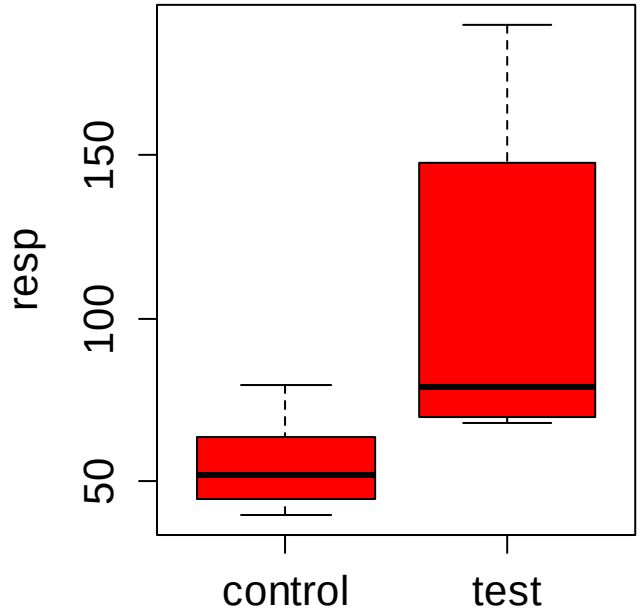
관심사인 변수(resp)와 두 그룹을 정의하는 변수 (treatment)가 입력되어야 함

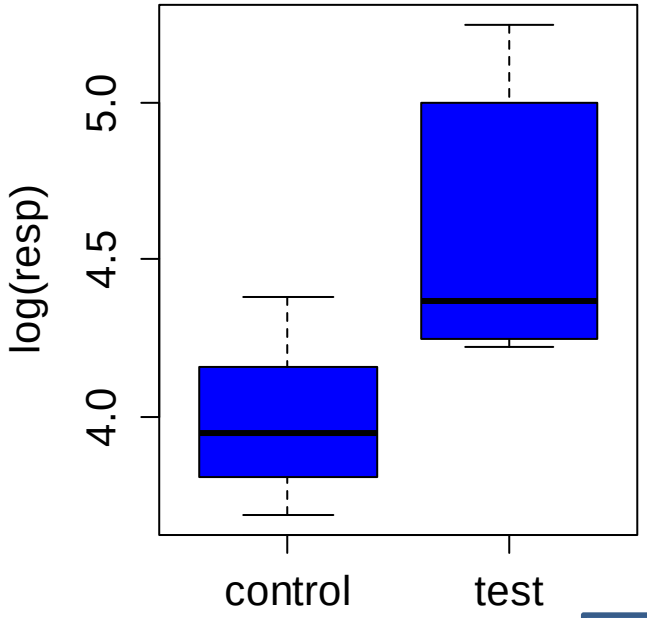
뼈세포수: 가설 수립

$H_0: \mu_1 \qquad \mu_2$

$H_a: \mu_1 \qquad \mu_2$

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)")
```

등분산 검정

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

함수	내용
var.test(종속변수~그룹변수)	그룹변수가 지정하는 그룹들 간에 종속변수의 분산의 차이가 있는지 검정 **H_0: 분산이 같다. vs H_a: 분산이 다르다**

```
> 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

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

함수	내용
t.test(종속변수~그룹변수, var.equal=TRUE)	그룹변수가 지정하는 그룹 간에 종속변수의 평균 차이가 있는지 검정 두 집단 분산이 같은 경우옵션 var.equal=TRUE를 사용. 디폴트는 var.equal=FALSE

```
> 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?

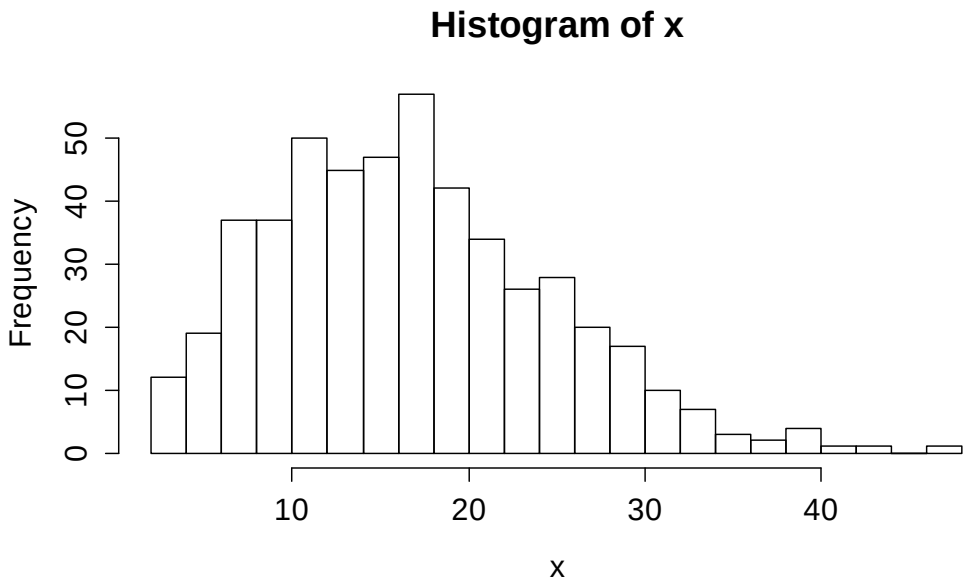
```
> var.test(resp~treatment,data=dental)

        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

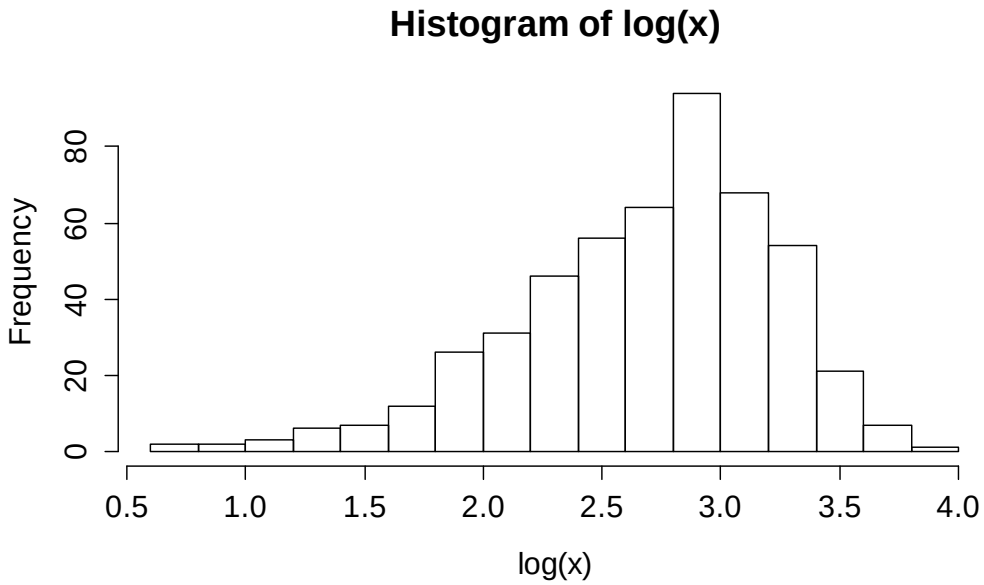
var.equal=TRUE 옵션 없이 이분산의 t-test

유의수준 5% 하에서 두 집단 평균의 차이가 없다.

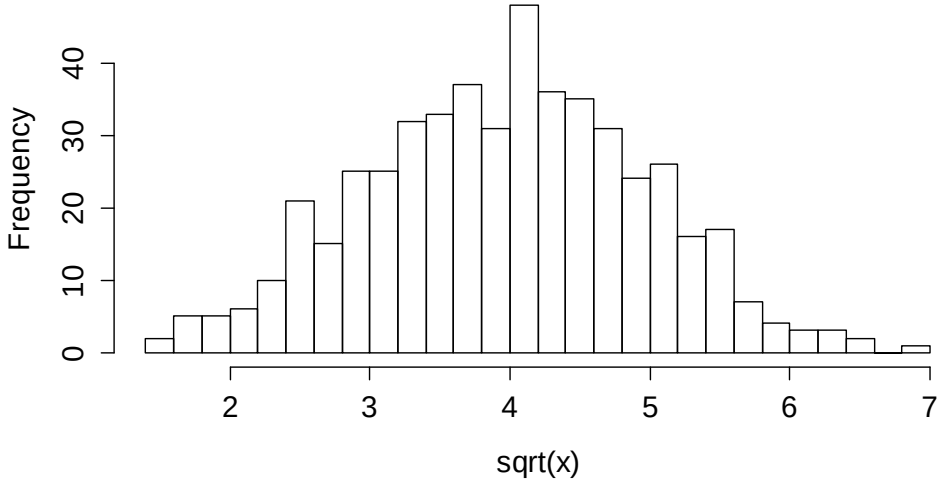
변수변환



- 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)
```



변수변환 (Box-Cox transformation)

- 데이터를 정규분포에 가깝게 만들어주는 지수변환의 값을 찾아준다. 즉, $y = x^{\lambda}$꼴이 가장 정규분포에 가깝도록 λ값 결정.
- Car 패키지 사용

```
> library(car)
> powerTransform(x)
Estimated transformation parameters
        x
0.4927693
```