R에서 제공하는
검정함수

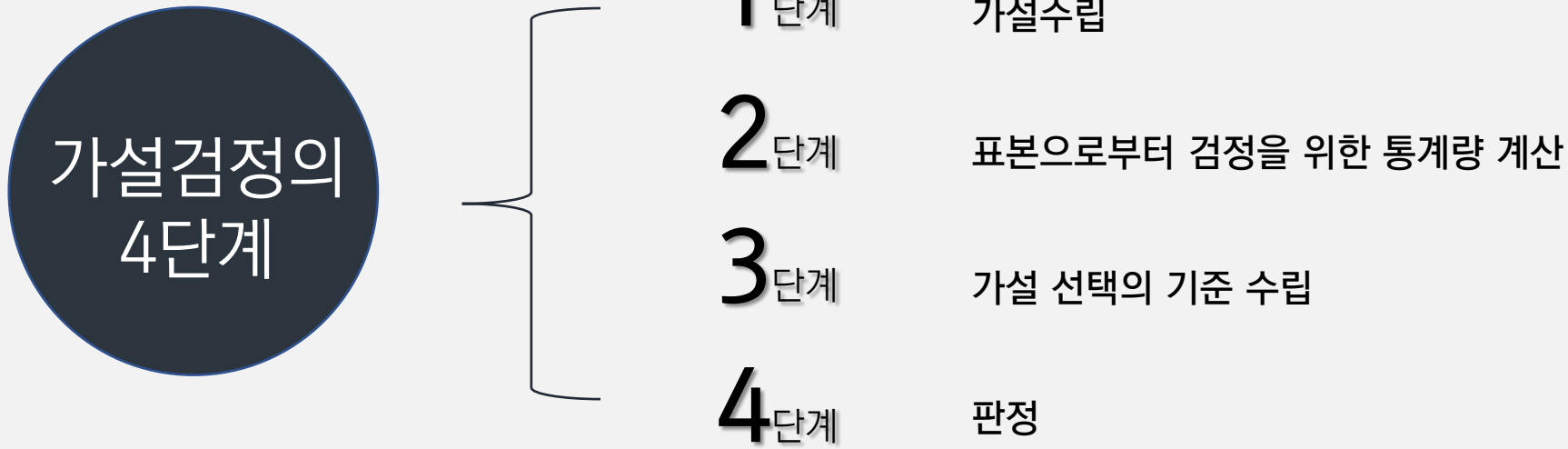
201701190 안중찬


1. 가설검정이란?

1 가설 검정이란? _가설검정의 4단계

● 가설 검정의 과정

모집단 특성의 상태에 대한 주장인 가설에 대해 표본으로부터 얻은 정보를 바탕으로 이를 채택할 지, 기각할 지를 판단함으로써, 모집단의 상태에 대해 결정하는 과정으로 아래의 4단계로 이루어짐



1 가설 검정이란? _검정통계량과 판정의 오류

검정통계량이란?

H_0 의 채택 및 기각 여부를 확인하기 위해 표본을 통해
관찰된 값을 사용하는 통계량

가설 검정 시 판정의 오류

제 1종 오류 : 귀무가설이 참인데 대립가설을 선택하는 오류

제 2종 오류 : 귀무가설이 거짓인데 귀무가설을 선택하는 오류

유의수준(α) : 제 1종오류를 범할 확률의 최대 허용 한계

유의수준(β) : 제 2종 오류를 범할 확률의 최대 허용 한계

유의수준의 역할 : 기각역 수립

임계값 : 기각역과 채택역

가설 검정 시 오류 표

		실제 상태	
판정		H_0	H_1
	H_0	O	X
	H_1	X	O


2.t-검정

2 t-검정_t-검정이란?

t-검정이란?

비용이나 시간이 많이 소요되는 실험의 경우에는 소 표본(일반적으로, n≤ 30)으로부터 통계적 추론을 해야 함
n이 크지 않을 때(소 표본일 때), 표본평균 $\bar{X}$ 의 표본분포는 어떤 것인지 모르기 때문에, 모집단의 분포가 정규분포라고 가정할 수 있을 경우 (σ를 모르는 경우), t-분포를 사용하여 검정을 함

t-검정 할 때 알아야 하는 값

표본표준편차 S : σ을 모르기 때문에, σ의 직관적인 추정량인 S를 사용
자유도 : $n-1$
상위 $\alpha \times 100\%$ **점에서 자유도(d.f.)가 n 인 점** : $t_\alpha(n)$
소 표본일 경우, 정규 모집단의 모평균에 대한 가설 검정을 위하여 사용되는 **검정통계량** : $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$

2 t-검정_귀무가설, 대립가설, 기각역, 유의확률

- 유의수준 α에서 각 귀무가설, 대립가설, 기각역과 유의확률

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
양쪽검정	$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$P(T > t_\alpha) = \alpha/2$ $P(T < t_\alpha) = \alpha/2$
왼쪽 한쪽검정	$H_0 : \mu \geq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$P(T < t_\alpha) = \alpha$
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(T > t_\alpha) = \alpha$

2 t-검정_소 표본일 때, 유의확률 구하기

R로 유의확률 구하기

오른쪽 한쪽 검정을 실행하였을 때, 검정통계량 : $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ =0.73이 나왔다고 하자. 이 때, 자유도가 14인 t-분포에서의 분포확률을 구해야 유의확률을 구할 수 있을 것이다.

유의확률은 $P(T > t_\alpha)$ = 1- $P(T > t_\alpha)$ = 1- $P(T > 0.73, \text{자유도 } 14)$

R 코드

```
> # 자유도가 14인 t-분포에서 구한 검정통계량 0.73에 대한 유의
확률 P(T>0.753)
> 1-pt(0.73,df=14)
[1] 0.2387141
```

2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

〈예제〉

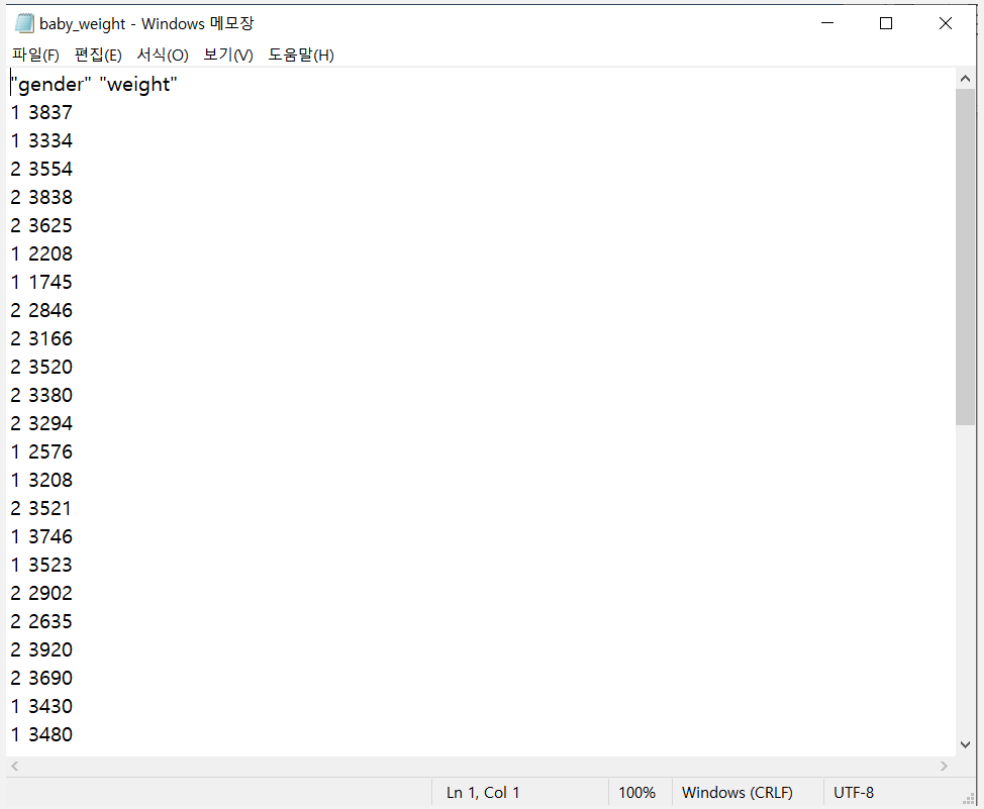
아래 주어진 자료는 여아 신생아 18명의 몸무게이다.

3837	3334	2208	1745	2576	2308
3746	3523	3430	3480	3116	3428
2184	2383	3500	3866	3542	3278

이 자료 들에서 '여아 신생아 몸무게가 2800g이 넘는다' 라는 가설을 뒷받침하는 강력한 증거를 조사원을 찾고자 한다. 이 가설을 검증하라.

표본의 개수는 18개 → 소 표본 검정 필요 → t-분포 활용

● 여아 신생아 18명의 몸무게를 txt파일에 데이터로 저장



2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

데이터 불러오기

```
> #여아 신생아 18명의 몸무게에 대한 t-검정
> ###데이터 불러오기
> data<-read.table("C:/Users/sec/Desktop/baby_weight.t
xt",header=T)
> str(data)
'data.frame':   44 obs. of  2 variables:
 $ gender: int  1 1 2 2 2 1 1 2 2 2 ...
 $ weight: int  3837 3334 3554 3838 3625 2208 1745 284
6 3166 3520 ...
> names(data)<-c("gender","weight")
> tmp<-subset(data,gender==1)
> ###데이터 확인하기
> weight<-tmp[[2]]
> weight
 [1] 3837 3334 2208 1745 2576 3208 3746 3523 3430
[10] 3480 3116 3428 2184 2383 3500 3866 3542 3278
```

평균과 표준편차 구하기

```
> ###평균과 표준편차, 표본의 개수 구하기
> xbar<-mean(weight)
> xbar
[1] 3132.444
> s<-sd(weight)
> s
[1] 631.5825
> n<-length(weight)
> n
[1] 18
>
```

2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

가설 세우기

'여아 신생아 몸무게가 2800g이 넘는다' 를 대립 가설로 내세운다면,
$H_1 : \mu > 2800$이 될 것이다.
이 때, 오른쪽 한쪽 검정이므로, 귀무 가설은

$H_0 : \mu \leq 2800$

$H_0 : \mu = 2800$

둘 중 하나가 될 수 있다.

이 때, $H_0 : \mu = 2800$을 채택한다.

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(T > t_\alpha) = \alpha$

귀무가설일 때, 검정통계량 구하기

```
> ###귀무가설일 때, 검정통계량 구하기
> mu0<-2800
> (t.st<-(xbar-mu0)/(s/sqrt(n)))
[1] 2.233188
>
```

$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = 2.23318$이라는 사실을 알 수 있음

2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

분위 수 구하기

이 때, 대립가설이 $H_1 : \mu > 2800$이므로, 기각역은 $R : T > t_\alpha$일 것이다.
이 때, 유의수준을 α=0.05로 잡고(95%일 때, 검정), 자유도가 17이므로,
이 때, t-분포에서 α=0.05일 때, t_α를 구하기 위해서 t-분포에서 분위수를 계산해주는 qt함수를 사용하여 분위수를 계산해줌

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(T > t_\alpha) = \alpha$

```
> ###유의수준이 0.05일 때, 자유도 17에서 t_a값(분위수) 구하기
> alpha<-0.05 # 유의수준 0.05로 놓기
> (c.u<-qt(1-alpha, df=n-1))# 자유도가 n-1인 t-분포에서 P
(T>cu)=0.05가 되는 분위수 구하기
[1] 1.739607
>
```

가설 검정하기

```
> ###귀무가설일 때, 검정통계량 구하기
> mu0<-2800
> (t.st<-(xbar-mu0)/(s/sqrt(n)))
[1] 2.233188
>
```

```
> ###유의수준이 0.05일 때, 자유도 17에서 t_a값(분위수) 구하기
> alpha<-0.05 # 유의수준 0.05로 놓기
> (c.u<-qt(1-alpha, df=n-1))# 자유도가 n-1인 t-분포에서 P
(T>cu)=0.05가 되는 분위수 구하기
[1] 1.739607
>
```

t_α=2.233188이므로, 기각역은 $R : T > 2.233188$ 이 된다. 이 때, 유의수준이 0.05일 때, 자유도 17에서 cu값(분위수)는 1.739607이므로, 기각역 안에 들지 못한다.
그러므로, 대립가설은 기각되지 못하고, '여야 신생아의 몸무게의 평균이 2800g보다 크다는 가설은 통계적으로 유의한 결론을 얻을 수 있다.

2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

유의 확률 구하기

```
> ###검정통계량에 대한 유의확률 구하기 (P(T>2.233))
> 1-pt(2.23318, df=n-1)
[1] 0.01963452
> 
```

대립가설이 $H_1 : \mu > 2800$이므로 유의확률은 $P(T > t_\alpha) = \alpha$
가 될 것이다. 이 때 t_α값은 2.23318이므로, 이를 넣어서 계산해주면, α가 나온다.
이 때, $P(T > t_\alpha)$=1- $P(T \leq t_\alpha)$이므로, 1-pt함수(누적 분포를 계산해주는 함수)
로 계산해주면 된다.

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(T > t_\alpha) = \alpha$

t.test() 함수

```
> ###t.test()
> t.test(weight,mu=2800,alternative="greater")

        One Sample t-test

data:  weight
t = 2.2332, df = 17, p-value = 0.01963
alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
 2873.477      Inf
sample estimates:
mean of x
 3132.444
```

지금까지 했던 모든 과정을 한번에 끝내주는 함수

2 t-검정_ 소 표본일 때, 단일 모집단의 평균에 대한 가설 검정

유의 확률 구하기

```
> ###검정통계량에 대한 유의확률 구하기 (P(T>2.233))
> 1-pt(2.23318, df=n-1)
[1] 0.01963452
> |
```

```
> ###귀무가설일 때, 검정통계량 구하기
> mu0<-2800
> (t.st<-(xbar-mu0)/(s/sqrt(n)))
[1] 2.233188
> |
```

대립가설이 H_1
가 될 것이다
이 때, $P(T >$
로 계산해주면 된다.

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
오른쪽 한쪽검정	$H_0 : \mu = \mu_0$		

```
> xbar<-mean(weight)
> xbar
[1] 3132.444
```

t.test() 함수

검정의 종류	alternative
양쪽검정	"two.sided" (기본값)
왼쪽 한쪽검정	"less"
오른쪽 한쪽검정	"greater"

```
> ###t.test()   mu0를 설정
> t.test(weight,mu=2800,alternative="greater")

        One Sample t-test

data:  weight
t = 2.2332, df = 17, p-value = 0.01963
alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
 2873.477      Inf
sample estimates:
mean of x
 3132.444
```

자유도

95% 신뢰구간

```
> ###검정통계량에 대한 유의확률 구하기 (P(T>2.233))
> 1-pt(2.23318, df=n-1)
[1] 0.01963452
> |
```


3.모비율검정

3 모비율검정_모비율 검정에서 알아야 하는 것

모비율 검정 할 때 알아야 하는 값

모비율 p에 대한 추정량으로 **표본비율** : $\hat{p} = \frac{X}{n}$

표준오차 : $S.E.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

H_0가설하에서 단일 모집단의 모비율 검정에서 사용하는 **검정통계량** : $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1^2)$

→ p_0 : H_0가설하에서의 모비율, n : 표본의 개수, $\sqrt{\frac{\widehat{p_0}(1-\widehat{p_0})}{n}}$: H_0가설하에서의 표준오차

유의수준 α에서 각 귀무가설, 대립가설, 기각역과 유의확률

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
양쪽검정	$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$P(Z > z_\alpha) = \alpha/2$ $P(Z < z_\alpha) = \alpha/2$
왼쪽 한쪽검정	$H_0 : \mu \geq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$P(Z < z_\alpha) = \alpha$
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(Z > z_\alpha) = \alpha$

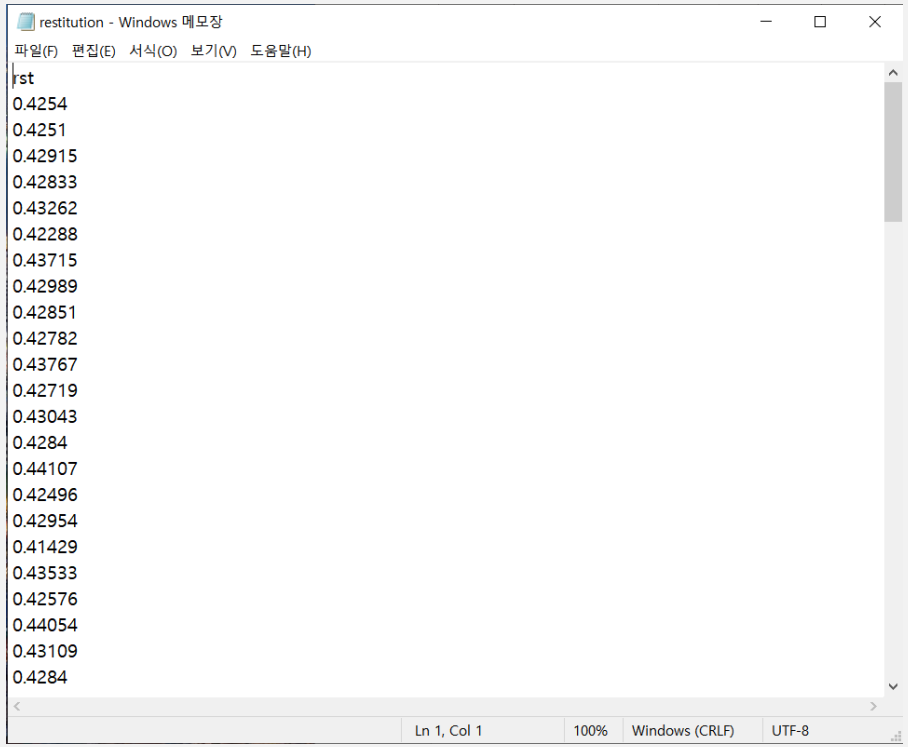
3 모비율검정_모비율 검정을 예제로 알아보기

〈예제〉

납품 받을 야구공을 임의로 샘플 100개를 추출하여 반발계수가 0.4134에서 0.4374 범위를 정상으로 인정하고, 불량률이 10%를 넘기면 납품을 받지 않는다고 가정하고 유의수준 α=0.05에 대해 모비율에 대한 가설검정을 실시하라.

모비율에 대한 가설 검정 실시 → Z-분포 활용

- 납품 받을 야구공 100개의 반발계수에 대한 샘플을 txt파일로 저장하기



3 모비율검정_모비율 검정을 예제로 알아보기

데이터 불러오기

```
> ###데이터 불러오기
> tmp<-read.table("C:/Users/sec/Desktop/restitution.tx
t",header=T)
> head(tmp)
      rst
1 0.42540
2 0.42510
3 0.42915
4 0.42833
5 0.43262
6 0.42288
>
```

Ifelse() 함수로 각 샘플이 불량인지 아닌지 파악하기

```
> ###ifelse 함수를 통하여 각 샘플이 불량인지 아닌지 파악하기
> rel<-ifelse(tmp$rst<0.4134|tmp$rst>0.4374,1,0)
> rel
  [1] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0
 [23] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [45] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
 [67] 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0
 [89] 0 0 0 0 0 0 0 0 0 0 0 0
>
```

Ifelse(test,1,0)
Test에 대해서 참이면 1로 표시하고, 거짓이면 0으로 표시하라
이 때, test를 tmp$rst〈0.4134|tmp$rst〉0.4374으로 놓았다
|는 or를 뜻하며, 두 조건 중 하나만 만족해도 1로 표시될 것이다.
즉, 1이 나오면 불량이라는 뜻이다.

3 모비율검정_모비율 검정을 예제로 알아보기

Phat 정의하기

```
> ###phat 정의
> n <- length(rel)
> n
[1] 100
> sumrel <- sum(rel)
> sumrel
[1] 11
> phat<- sumrel / n
> phat
[1] 0.11
> 
```

Length()함수로 표본의 개수를 n으로 정의하고,
불량인 1들을 모두 더해서 불량의 개수를 sum()함수로
Sumrel로 정의한 후, phat을 sumrel/n으로 정의하기

가설 설정하기

불량률이 10%를 넘기면 납품을 받지 않는다고 가정했으므로,

$$H_1 : p_{불량} > 0.1$$

이 대립가설이 될 것이다.
그렇다면, 귀무가설은

$$H_0 : p_{불량} \leq 0.1 \text{ 또는 } H_0 : p_{불량} = 0.1$$

이 될 것이다.

이 2개 중에서, $H_0 : p_{불량} \leq 0.1$ 를 귀무가설로 설정하도록 하겠다. 왜냐하면, 불량일 확률이 0.1일 때만 물건을 납품받지 않는 것이 아니라, 0.1이하일 때, 전부를 받기 때문이다.

3 모비율검정_모비율 검정을 예제로 알아보기

p_0설정

$H_0 : p_{불량} \leq 0.1$ 이기 때문에,

H_0가설하에 불량률이 0.1이라고 할 수 있다. 이를 정의해주면 된다.

```
> ###p_0정의
> p0 <- 0.1
>
```

검정통계량 구하기

```
> ###검정통계량 구하기
> (z <- (phat - p0) / sqrt( ( p0*(1-p0) )/n ) )
[1] 0.3333333
>
```

$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1^2)$을 구하면 된다.

3 모비율검정_모비율 검정을 예제로 알아보기

분위 수 구하기

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
오른쪽 한쪽검정	$H_0 : \mu \leq \mu_0$ $H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P(Z > z_\alpha) = \alpha$

유의수준 α=0.05에 대해 모비율에 대한 가설검정을 실시
한다고 하였으므로, α=0.05로 설정하고, $H_1 : p_{불량} > 0.1$이 대립 가설이므로, 기각역은 R : $(Z > z_\alpha)$ 이 될 것이다.
그러므로, Z-분포에서 α=0.05일 때, Z_α를 구하기 위해서 Z-분포에서 분위수를 계산해주는 qnorm함수를 사용하여 분위수를 계산해줌

```
> ###유의수준이 0.05일 때, z_a값(분위수) 구하기
> alpha<-0.05
> (c.u<-qnorm(1-alpha))
[1] 1.644854
>
```

3 모비율검정_모비율 검정을 예제로 알아보기

가설 검정하기

```
> ###검정통계량 구하기
> (z <- (phat - p0) / sqrt( ( p0*(1-p0) )/n ) )
[1] 0.3333333
> 
```

```
> ###유의수준이 0.05일 때, z_a값(분위수) 구하기
> alpha<-0.05
> (c.u<-qnorm(1-alpha))
[1] 1.644854
> 
```

z_α =0.333333이므로, 기각역은 $R : Z > 0.3333333$ 이 된다. 이 때, 유의수준이 0.05일 때, 자유도 17에서 t_α값(분위수)는 1.644854이므로, 기각역 안에 들어간다. 그러므로, 귀무가설은 기각되지 못하고, 대립가설이 기각된다. 그러므로, 귀무가설인 “공장에서 생산된 야구공의 불량은 10% 미만이다 ‘라는 가설은 통계적으로 유의한 결론을 얻을 수 있다.

P값 구하기

```
> ###p 값 구하기
> (p.value<-1-pnorm(z))
[1] 0.3694413
> 
```

대립가설이$H_1 : p_{불량} > 0.1$이므로 유의확률은 $P(Z > Z_\alpha) = \alpha$
가 될 것이다. 이 때 z_α값은 0.333333이므로, 이를 넣어서 계산해주면, α가 나온다. 이 때, $P(Z > z_\alpha)$=1- $P(Z \leq z_\alpha)$이므로, 1-pnorm함수(누적분포를 계산해주는 함수)로 계산해주면 된다.

3 모비율검정_모비율 검정을 예제로 알아보기

Prop.test() 함수

```
> ###prop.test() 함수
> prop.test( sumrel, n, p=0.1, alternative="greater",
  correct=FALSE)

        1-sample proportions test without
        continuity correction

data:  sumrel out of n, null probability 0.1
X-squared = 0.11111, df = 1, p-value =
0.3694
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.0684615 1.0000000
sample estimates:
   p
0.11

>
```

지금까지 했던 모든 과정을 한번에 끝내주는 함수

```
> ###p_0정의
> p0 <- 0.1
>
```

검정의 종류	alternative
양쪽검정	"two.sided" (기본값)
왼쪽 한쪽검정	"less"
오른쪽 한쪽검정	"greater"

```
> ###p 값 구하기
> (p.value<-1-pnorm(z))
[1] 0.3694413
>
```

이 때, 유의확률이 유의수준 0.05보다 더 크므로,
귀무가설은 기각되지 못하고 대립가설은 기각된다.

95% 신뢰구간


4.등 분산 검정

4 등 분산 검정_등 분산 검정이란?

등분산 검정이란?

등분산 검정은 서로 독립인 두 모집단의 평균 차이를 검정할 때, 두 집단의 분산은 서로 동일하다는 가정을 만족하는지 검정하는 방법이다.
이렇게 분산의 동일성을 검정하는 R 함수는 var.test() 함수이다.

〈예제〉

남아 신생아 26명과 여아 신생아 18명의 몸무게가 기록된 자료에서, 유의수준 0.05
로 하여 여아와 남아의 분산이 서로 동일한지 검정하라.

여아와 남아의 분산이 서로 동일한지 검정 → 등 분산 검정 활용

4 등 분산 검정_등 분산 검정을 예제로 알아보기

가설 수립

$$H_0 : \sigma_{여아몸무게}^2 / \sigma_{남아몸무게}^2 = 1$$
vs.
$$H_1 : \sigma_{여아몸무게}^2 / \sigma_{남아몸무게}^2 \neq 1$$

분산이 서로 같은지를 알아보는 검정이므로 , 귀무가설은 분산이 서로 동일하다 일 것이고, 대립가설은 분산이 동일하지 않다 일 것이다.

분위수 구하기

```
> # F-분포(양쪽검정, 자유도 17, 25)의 임계값
> qf(0.975, df1=17,df2=25)
[1] 2.359863
>
```

유의수준 0.05에서 검정한다고 했으므로, 자유도가 17과 25인 F-분포에서 $P(F < c_u)$ 가 되게 하는 분위수를 구해준다.
이 때, 양측 검증을 하지만, 분산분석이나 카이 제곱분석에서는 한쪽만 따져준다.

4 등 분산 검정_등 분산 검정을 예제로 알아보기

검정 통계량과 p-값 구하기

```
> # 등분산 검정
> data <- read.table("C:/Users/sec/Desktop/baby_weigh
t.txt", header=T)
> var.test(data$weight ~ data$gender)

        F test to compare two variances

data:  data$weight by data$gender
F = 2.1771, num df = 17, denom df = 25,
p-value = 0.07526
alternative hypothesis: true ratio of variances is not
 equal to 1
95 percent confidence interval:
 0.9225552 5.5481739
sample estimates:
ratio of variances
          2.177104

> |
```

유의확률

유의수준 0.05보다 큰 0.07526을
띄고 있으므로, 귀무가설을 채택할 수 있다.

검정통계량은 2.1771이므로, 유의수준이 0.05 일 때, F-분포에서
$P(F < 2.357863)$
안에 들어가므로, 채택역 안에 들어가게 되고
귀무가설을 채택할 수 있게 된다.

검정통계량

5. 서로 독립인 두 모집단의 모평균 차이 검정

5 서로 독립인 두 모집단의 차이 검정

서로 독립인 두 모집단의 모평균 차이 검정

〈예제〉

남아 신생아 26명과 여아 신생아 18명의 몸무게가 기록된 자료에서, 유의수준 0.05로 하여
남아의 몸무게가 여아의 몸무게보다 더 많이 나간다는 것을 검정하라.

가설 수립

검정의 종류에 따라 두 모집단의 모평균 차이를 검정할 때 수립되는 가설은 검정에 따라 아래와 같다.

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
양쪽검정	$H_0 : \mu_1 - \mu_2 = 0$	$H_1 : \mu_1 - \mu_2 \neq 0$	$P(t \geq t_\alpha) = \alpha/2$ $P(t \leq t_\alpha) = \alpha/2$
왼쪽 한쪽검정		$H_1 : \mu_1 - \mu_2 < 0$	$P(t \leq -t_\alpha) = \alpha$
오른쪽 한쪽검정		$H_1 : \mu_1 - \mu_2 > 0$	$P(t \geq t_\alpha) = \alpha$

이 때, 남아의 몸무게가 여아의 몸무게보다 더 많이 나간다는 것을 검정하라고 하였으므로,
대립가설은 $H_1 : \mu_{\text{여아의 몸무게}} - \mu_{\text{남아의 몸무게}} < 0$이 될 것이고, 이 때 귀무가설은 $H_0 : \mu_{\text{여야의 몸무게}} - \mu_{\text{남아의 몸무게}} = 0$ 가 될 것이다.

5 서로 독립인 두 모집단의 차이 검정

- t.test() 함수

```
> #서로 독립인 두 모집단: 모평균의 차이 검정
> t.test(data$weight~data$gender,mu=0,alternative="les
s",var.equal=T)

        Two Sample t-test

data:  data$weight by data$gender
t = -1.5229, df = 42, p-value = 0.06764
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
     -Inf 25.37242
sample estimates:
mean in group 1 mean in group 2
       3132.444        3375.308

>
```

귀무가설에서 $H_0 : \mu_{\text{여아의 몸무게}} - \mu_{\text{남아의 몸무게}} = 0$
이므로, mu=0

검정의 종류	alternative
양쪽검정	"two.sided" (기본값)
왼쪽 한쪽검정	"less"
오른쪽 한쪽검정	"greater"

대립가설은 $H_1 : \mu_{\text{여아의 몸무게}} - \mu_{\text{남아의 몸무게}} < 0$이므로,
왼쪽 한쪽검정이다. 그러므로 "less"를 넣어주면 된다.

앞에서의 분산검정에서 분산이 같다는 것이 검정되었으므로,
var.equal=T로 분산의 동일성 여부를 전달하면 된다.

5 서로 독립인 두 모집단의 차이 검정

- t.test() 함수

```
> #서로 독립인 두 모집단: 모평균의 차이 검정
> t.test(data$weight~data$gender,mu=0,alternative="les
s",var.equal=T)

        Two Sample t-test

data:  data$weight by data$gender
t = -1.5229, df = 42, p-value = 0.06764
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
     -Inf 25.37242
sample estimates:
mean in group 1 mean in group 2
       3132.444        3375.308

> |
```

유의 확률은 0.06764가 나왔다. 유의수준이 0.05일 때, 유의 확률이 이것보다 높으므로, 이는 신뢰할 수 있다.

유의확률

검정통계량 -1.5229

5 서로 독립인 두 모집단의 차이 검정

t.test() 함수

```
> #서로 독립인 두 모집단: 모평균의 차이 검정
> t.test(data$weight~data$gender,mu=0,alternative="les
s",var.equal=T)

        Two Sample t-test

data:  data$weight by data$gender
t = -1.5229, df = 42, p-value = 0.06764
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
     -Inf 25.37242
sample estimates:
mean in group 1 mean in group 2
       3132.444        3375.308

>
```

검정통계량 -1.5229

분위수 구하기

유의수준 0.05에서 검정한다고 했으므로, 자유도가 42인 t-분포에서 $P(t \leq -t_\alpha)$=0.05가 되게 하는 분위수 $-t_\alpha$ 를 구해준다.

```
> #t-분포(단측검정, 자유도 42)의 임계값
> qt(0.05, df=42)
[1] -1.681952
>
```

이 때, 그 결과는 위와 같이 나오고, 검정 통계량이 -1.5229이므로, 이는 $t \leq -1.681952$ 안에 들어간다고 할 수 있다. 즉, 대립가설은 기각되고, 귀무가설이 맞다는 의미가 된다. 다시 말하면, 남아 몸무게 평균이 여아 몸무게의 평균보다 크지 않은 것으로 판단된다.


6.서로 대응인 두 모집단에
대한 모평균의 차이 검정

6 서로 대응인 두 모집단의 모평균 차이 검정

서로 대응인 두 모집단의 모평균 차이 검정

대응인 두 집단의 평균 비교는 동일한 관찰대상으로부터 처리 이전의 관찰과 처리 이후 관찰을 통해
처리가 어떠한 영향을 미쳤는지 밝히는데 많이 사용된다.

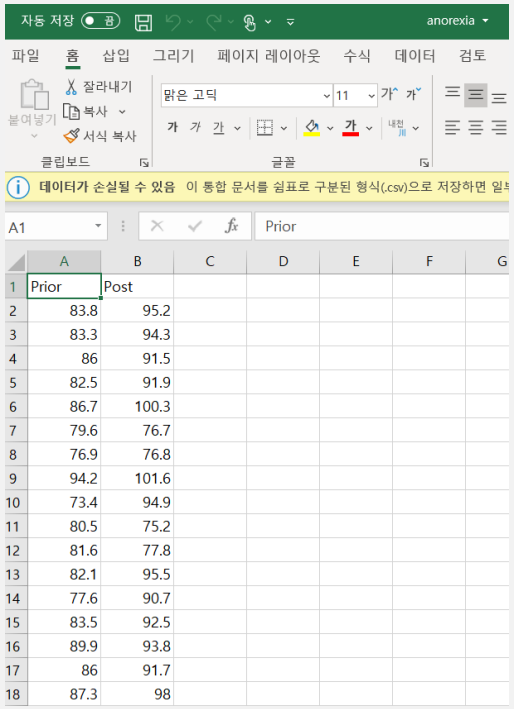
〈예제〉

식욕 부진증 치료법의 효과 검정

17명의 여학생들로부터 신경성 식욕부진 등의 치료요법 시행 전(prior)과 시행 후(post)에 각각 몸무게를 측정한 자료에서 시행 전과 시행 후의 체중은 정규분포를 따른다고할 때, 유의수준 0.05에서 신경성 식욕부진의 치료요법이 효과가 있음을 검정하라.

6 서로 대응인 두 모집단의 모평균 차이 검정

데이터 csv 파일로 저장하기



	A	B	C	D	E	F	G
1	Prior	Post					
2	83.8	95.2					
3	83.3	94.3					
4	86	91.5					
5	82.5	91.9					
6	86.7	100.3					
7	79.6	76.7					
8	76.9	76.8					
9	94.2	101.6					
10	73.4	94.9					
11	80.5	75.2					
12	81.6	77.8					
13	82.1	95.5					
14	77.6	90.7					
15	83.5	92.5					
16	89.9	93.8					
17	86	91.7					
18	87.3	98					

데이터 불러오기

```
> data <- read.csv("C:/Users/sec/Desktop/anorexia.cs
v",header=T)
> str(data)
'data.frame':   17 obs. of  2 variables:
 $ Prior: num  83.8 83.3 86 82.5 86.7 79.6 76.9 94.2 7
3.4 80.5 ...
 $ Post : num  95.2 94.3 91.5 91.9 100.3 ...
> data
   Prior  Post
1   83.8  95.2
2   83.3  94.3
3   86.0  91.5
4   82.5  91.9
5   86.7 100.3
6   79.6  76.7
7   76.9  76.8
8   94.2 101.6
9   73.4  94.9
10  80.5  75.2
11  81.6  77.8
12  82.1  95.5
13  77.6  90.7
14  83.5  92.5
15  89.9  93.8
16  86.0  91.7
17  87.3  98.0
>
```

6 서로 대응인 두 모집단의 모평균 차이 검정

가설 수립

검정의 종류에 따라 두 모집단의 모평균 차이를 검정할 때 수립되는 가설은 검정에 따라 아래와 같다.

검정의 종류	귀무가설	대립가설	
양쪽검정	$H_0 : \mu_D = 0$	$H_1 : \mu_D \neq 0$	차이가 있음 여부
왼쪽 한쪽검정	$H_0 : \mu_D \geq 0$ $H_0 : \mu_D = 0$	$H_1 : \mu_D < 0$	처리로 인해 사후 관찰 값이 줄어듦의 여부
오른쪽 한쪽검정	$H_0 : \mu_D \leq 0$ $H_0 : \mu_D = 0$	$H_1 : \mu_D > 0$	처리로 인해 사후 관찰 값이 증가함의 여부

이 때, 신경성 식욕부진의 치료요법이 효과가 있음을 검정하라고 하였으므로, 시행 전에 비해서 시행 후에 체중이 증가했을 것이다.
이 때, μ_D=시행 전- 시행후이므로, 대립가설은 $H_1 : \mu_D < 0$이 되고, 귀무가설은 $H_0 : \mu_D \geq 0$ 또는 $H_0 : \mu_D = 0$이 될 것이다.

6 서로 대응인 두 모집단의 모평균 차이 검정

검정통계량 구하기

검정통계량은 n개의 관찰대상으로부터 각 대응별 차이의 (표본)평균을 $\bar{D}$, (표본)표준편차를 S_D 로 나타낼 때 귀무가설이 $H_0 : \mu_D = 0$이므로,
검정통계량은 $T = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t(n-1)$이 될 것이다.

```
> (n <- length(data$Prior - data$Post))
[1] 17
> (m <- mean( data$Prior - data$Post ))
[1] -7.264706
> (s <- sd (data$Prior - data$Post))
[1] 7.157421
> (round( t.t <- m/(s / sqrt(n)),3) )
[1] -4.185
>
```

이를 R코드로 만들어보면 다음과 같이 만들어진다.
이에 따라서, 검정 통계량 -4.185를 얻을 수 있었다.

t.test()를 통해 더 쉽게 검정통계량 구하기

```
> t.test(data$Prior, data$Post, paired=T, alternative
="less")

        Paired t-test

data:  data$Prior and data$Post
t = -4.1849, df = 16, p-value = 0.0003501
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
      -Inf -4.233975
sample estimates:
mean of the differences
              -7.264706

>
```

위처럼 하나하나 식을 r코드로 짜서 구하는 것보다,
t.test()를 이용하여 구하는 것이 더 쉽다는 것을 알 수 있다.
이 때, paired=T는 두 집단의 표본이 대응이 된다는 이야기를 해주고 있고,
대립가설이$H_1 : \mu_D < 0$이 되고, 왼측 한쪽검정 이므로 “less”를 사용해준다.
이 때, 검정통계량은 -4.1849라는 사실을 알게 된다.

6 서로 대응인 두 모집단의 모평균 차이 검정

- 분위수 구하고 검정하기

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
왼쪽 한쪽검정	$H_0 : \mu_1 - \mu_2 = 0$	$H_1 : \mu_1 - \mu_2 < 0$	$P(t \leq -t_\alpha) = \alpha$

```
> alpha <- 0.05
> qt(alpha, df=16)
[1] -1.745884
>
```

유의수준 0.05에서 검정한다고 했으므로, 자유도가 16인 t-분포에서 $P(t \leq -t_\alpha)$=0.05가 되게 하는 분위수 $-t_\alpha$ 를 구해준다.

이 때, 기각역은 $t \leq -1.745884$ 이므로, 검정통계량 -4.1849가 기각역 안에 들어간다는 것을 알 수 있다.

그러므로, 귀무가설이 기각되고 대립가설이 채택된다. 즉, 식욕부진증 치료 요법이 효과가 있는 것으로 판단된다.

```
> t.test(data$Prior, data$Post, paired=T, alternative
="less")

        Paired t-test

data:  data$Prior and data$Post
t = -4.1849, df = 16, p-value = 0.0003501
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
      -Inf -4.233975
sample estimates:
mean of the differences
              -7.264706

>
```

6 서로 대응인 두 모집단의 모평균 차이 검정

- 분위수 구하고 검정하기

검정의 종류	귀무가설	대립가설	기각 역과 유의수준
왼쪽 한쪽검정	$H_0 : \mu_1 - \mu_2 = 0$	$H_1 : \mu_1 - \mu_2 < 0$	$P(t \leq -t_\alpha) = \alpha$

```
> pt(t.t, df=16)
[1] 0.0003501266
> pt(-4.1849,df=16)
[1] 0.0003501325
>
```

대립가설이 $H_1 : \mu_D < 0$ 이므로 유의확률은$P(t \leq -t_\alpha) = \alpha$가 될 것이다.

이 때 t_α값은 -4.1849이므로, 이를 넣어서 계산해주면,
α가 나온다. 이 때, $P(T > t_\alpha)$=1- $P(T \leq t_\alpha)$이므로,
1-pt함수(누적 분포를 계산해주는 함수)로 계산해주면 된다.

이 때, 앞에서 식을 써서 계산한 검정 통계량을 넣어도 무방하다. 이 결과 유의 확률이
유의수준 0.05보다 작으므로 귀무가설은 기각된다.

```
> t.test(data$Prior, data$Post, paired=T, alternative
="less")

        Paired t-test

data:  data$Prior and data$Post
t = -4.1849, df = 16, p-value = 0.0003501
alternative hypothesis: true difference in means is le
ss than 0
95 percent confidence interval:
      -Inf -4.233975
sample estimates:
mean of the differences
              -7.264706

>
```


7.모집단이 세 개 이상일
경우의 평균 비교검정

7 모집단이 세 개 이상일 경우의 평균 비교 검정

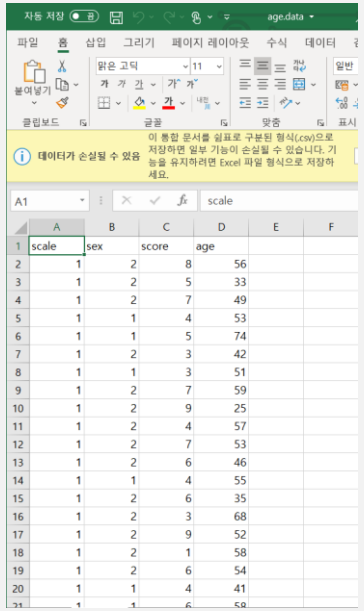
- 모집단이 세 개 이상일 경우의 평균 비교 검정

모집단이 세 개 이상일 경우의 평균 비교 검정을 하기 위해서는 우선 오차제곱합(SSE)와 처리제곱합(SST)을 구해줘야 한다.
그래야 평균의 동일성에 대한 F-검정을 실행할 수 있기 때문이다.

평균의 동일성에 대한 F-검정이란,
귀무가설이 $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$일 때, 대립가설은 H_1 : 모든 μ_i가 같은 것은 아니다. 로 표현되고,
이 때, 기각역은 $F = \frac{SS_T/(k-1)}{SSE/(n-k)} \geq F_a(k-1, n-k)$를 가지는 것을 통해 검정하는 것을 말한다.

7 모집단이 세 개 이상일 경우의 평균 비교 검정

데이터 csv 파일로 저장하기



	A	B	C	D
1	scale	sex	score	age
2	1	2	8	56
3	1	2	5	33
4	1	2	7	49
5	1	1	4	53
6	1	1	5	74
7	1	2	3	42
8	1	1	3	51
9	1	2	7	59
10	1	2	9	25
11	1	2	4	57
12	1	2	7	53
13	1	2	6	46
14	1	1	4	55
15	1	2	6	35
16	1	2	3	68
17	1	2	9	52
18	1	2	1	58
19	1	2	6	54
20	1	1	4	41

데이터 불러오기

```
> ad <- read.csv("C:/Users/sec/Desktop/age.data.csv",h
eader=T)
> str(ad)
'data.frame':   150 obs. of  4 variables:
 $ scale: int  1 1 1 1 1 1 1 1 1 1 ...
 $ sex  : int  2 2 2 1 1 2 1 2 2 2 ...
 $ score: int  8 5 7 4 5 3 3 7 9 4 ...
 $ age  : int  56 33 49 53 74 42 51 59 25 57 ...
> View(ad)
>
```

	scale	sex	score	age
1	1	2	8	56
2	1	2	5	33
3	1	2	7	49
4	1	1	4	53
5	1	1	5	74
6	1	2	3	42
7	1	1	3	51
8	1	2	7	59
9	1	2	9	25
10	1	2	4	57
11	1	2	7	53
12	1	2	6	46
13	1	1	4	55
14	1	2	6	35
15	1	2	3	68
16	1	2	9	52
17	1	2	1	58
18	1	2	6	54

7 모집단이 세 개 이상일 경우의 평균 비교 검정

오차 제곱합 구하기

우선 SSE인 오차제곱합부터 구해야 된다.
이 때, 오차제곱합은 $SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$ 이기 때문에 이를 가지고 문제를 해결한다.

각 집단별로 나누어 정리하기

```
> y1 <- ad$age[ad$scale=="1"]
> y2 <- ad$age[ad$scale=="2"]
> y3 <- ad$age[ad$scale=="3"]
> 
```

각 집단 별 평균 구하기

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$$

```
> y1.mean <- mean(y1)
> y2.mean <- mean(y2)
> y3.mean <- mean(y3)
> y1.mean
[1] 45.94
> y2.mean
[1] 45.68
> y3.mean
[1] 47.92
> 
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

오차 제곱합 구하기

우선 SSE인 오차제곱합부터 구해야 된다.
이 때, 오차제곱합은 $SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$ 이기 때문에 이를 가지고 문제를 해결한다.

각 집단 별 편차 제곱 합 구하기

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$$

```
> sse.1 <- sum( (y1 - y1.mean)^2 )
> sse.2 <- sum( (y2 - y2.mean)^2 )
> sse.3 <- sum( (y3 - y3.mean)^2 )
> sse.1
[1] 10244.82
> sse.2
[1] 9048.88
> sse.3
[1] 10845.68
>
```

각 집단 별 편차 제곱 합을 모두 더해 오차 제곱 합 구하기

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$$

```
> (sse <- sse.1 + sse.2 + sse.3)
[1] 30139.38
>
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

오차 제곱합 구하기

이 때, 오차제곱합 의 자유도는 $\sum_{i=1}^{k} n_i - \mathrm{k}$이므로,
이는 R 코드로 아래와 같이 출력할 수 있다.

각 집단 별 편차 제곱 합 구하기

각 집단 별 편차 제곱 합을 모두 더해 오차 제곱 합 구하기

```
> (dfe <- (length(y1)) + (length(y2)) + (length(y3))-
3)
[1] 147
>
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

처리 제곱합 구하기

그다음 SST인 처리제곱합부터 구해야 된다.
이 때, 처리제곱합은 $SST = \sum_{i=1}^{3} n_i(\bar{y_i} - \bar{y})^2$ 이기 때문에 이를 가지고 문제를 해결한다.

- 전체 평균을 구해 변수 y에 저장

$$SST = \sum_{i=1}^{3} n_i(\bar{y_i} - \bar{y})^2$$

```
> y <- mean(ad$age)
> y
[1] 46.51333
> |
```

- 각 처리별로 처리의 평균과 전체 평균과의 편차제곱합을 구하고, 각 처리의 표본 개수와 곱함

$$SST = \sum_{i=1}^{3} n_i(\bar{y_i} - \bar{y})^2$$

```
> sst.1 <- length(y1) * sum((y1.mean - y)^2)
> sst.2 <- length(y2) * sum((y2.mean - y)^2)
> sst.3 <- length(y3) * sum((y3.mean - y)^2)
> sst.1
[1] 16.43556
> sst.2
[1] 34.72222
> sst.3
[1] 98.93556
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

오차 제곱합 구하기

우선 SSE인 오차제곱합부터 구해야 된다.
이 때, 오차제곱합은 $SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2$ 이기 때문에 이를 가지고 문제를 해결한다.

- 각 처리별로 구한 값을 모두 더해 처리제곱합을 구한 후 변수 sst에 저장하고 출력

$$SST = \sum_{i=1}^{3} n_i(\bar{y_i} - \bar{y})^2$$

```
> (sst <- sst.1 + sst.2 + sst.3)
[1] 150.0933
> 
```

- 처리제곱합의 자유도는 $(k-1)$ 이므로, 아래와 같이 유도될 수 있음

```
> as.factor(ad$scale) #as.factor() 요인(factor)으로 변환
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [23] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [45] 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [67] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [89] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
[111] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[133] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
Levels: 1 2 3
> nlevels(as.factor(ad$scale)) #ad$scale의 factor 개수#
nlevels():level의 개수(출력 결과의 개수가 처리집단의 수)
[1] 3
> (dft <- nlevels(as.factor(ad$scale)) - 1)
[1] 2
>
> dft<-3-1
> dft
[1] 2
> #식을 이용하여 계산 할 수도 있다.
> 
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

검산

- 총 제곱 합 구하기

총 제곱합 = 처리제곱합 + 잔차제곱합 이므로, 이를 통해 검산이 가능하다.
이 때, 총 제곱합은 $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$이므로 이를 R로 아래와 같이 구현 가능하다.

```
> ( tsq <- sum( (ad$age - y)^2 ) )
[1] 30289.47
>
```

- 검산하기

총 제곱합 = 처리제곱합 + 잔차제곱합이므로, 처리제곱합과 잔차제곱합을 더해준 것이 총 제곱합과 같아야 한다.
이는 아래와 같이 보일 수 있다.

```
> ( ss <- sst + sse )
[1] 30289.47
>
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

F-검정 실시

이 때, 앞에서 평균의 동일성에 대한 F-검정이란, 귀무가설이 $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$일 때, 대립가설은 H_1 : 모든 μ_i가 같은 것은 아니다. 로 표현되고, 이 때, 기각역은 $F = \frac{SS_T/(k-1)}{SSE/(n-k)} \geq F_a(k-1, n-k)$를 가지는 것을 통해 검정하는 것을 말한다고 하였고, SS_T와 SSE, 또한 SS_T의 자유도인 $(k-1)$ 과 SSE의 자유도인 $(n-k)$를 밝혀냈으므로, 검정통계량을 구할 수 있다.이를 R로 구현하면 아래와 같다.

- 처리평균제곱합을 구해 변수 mst에 저장

$$F = \frac{SS_T/(k-1)}{SSE/(n-k)}$$

```
> mst <- sst / dft
> mst
[1] 75.04667
> |
```

- 오차 평균 제곱합을 구해 변수 mse에 저장

$$F = \frac{SS_T/(k-1)}{SSE/(n-k)}$$

```
> mse <- sse / dfe
> mse
[1] 205.0298
> |
```

- 처리평균제곱합(mst)를 오차평균제곱합(mse)로 나눈값을 f.t에 저장하고 출력

$$F = \frac{SS_T/(k-1)}{SSE/(n-k)}$$

```
> (f.t <- mst / mse)
[1] 0.3660281
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

판정

기각역은 $F \geq F_a(k-1, n-k)$이므로, 오른쪽 단측 검정이라는 사실을 알 수 있다. 유의수준이 0.05라고 한다면, 우리는 유의수준이 0.05일 때, 유의수준 0.05에서 검정한다고 했으므로, 자유도가 2,147인 F-분포에서 $P(F \leq F_\alpha(2,147))$=0.05가 되게 하는 분위수 $F_\alpha(2,147)$를 구해준다. 이를 R로 나타내면, 옆에 보이는 것과 같다.

```
> ##유의수준이 0.05라고 놓았으므로, alpha를 0.05라고 놓고
 시작
> alpha<-0.05
> ##유의수준이 0.05일 때, qf()함수를 통하여 자유도가 2,147
인 F-분포에서 유의수준 0.05일 때 분위수를 구함
> (tol<-qf(1-alpha,2,147))
[1] 3.057621
>
```

이 때, 검정통계량이 0.3660281이므로, 기각역 $F \geq 3.057621$안에 들어가지 않으므로 귀무가설이 채택된다.

```
> (f.t <- mst / mse)
[1] 0.3660281
>
```

또한, 검정 통계량이 저장된 f.t.값보다 클 확률을 구해 p값을 계산해 보아도 유의수준 0.05보다 큰 값이 나오므로, 귀무가설이 채택됨

```
> #검정통계량이 저장된 f.t값보다 클 확률을 구해 p.value에
 저장하고 출력
> (p.value <- 1 - pf(f.t, 2, 147))
[1] 0.6941136
>
```

결론적으로, 3개의 모평균이 차이가 없다는 것을 유도해낼 수 있다

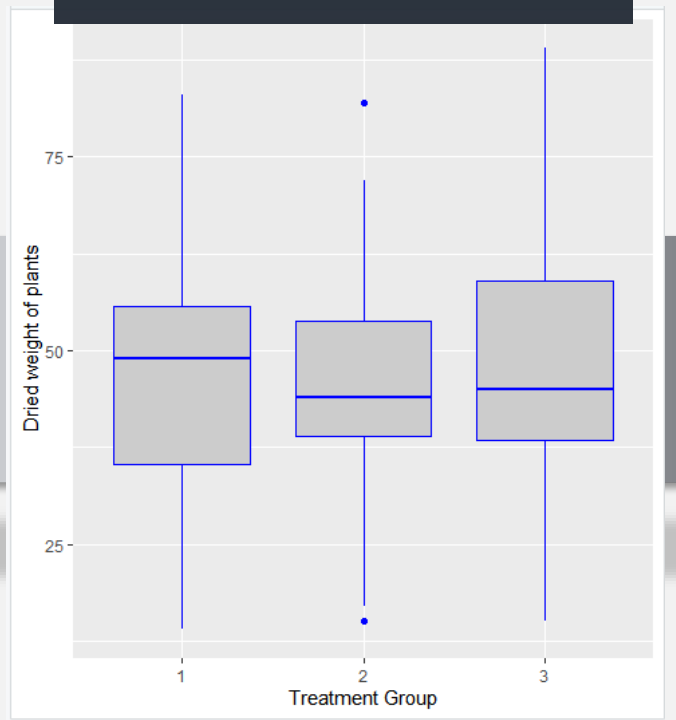
7 모집단이 세 개 이상일 경우의 평균 비교 검정

- 평균의 동일성 검증을 boxplot을 통해 시각화하기

R코드

```
> #변수 scale의 label을 재정의
> ad$scale = factor(ad$scale, labels = c("1", "2",
 "3"))
> require(ggplot2)
> ggplot(ad, aes(x = scale, y = age)) +
+     geom_boxplot(fill = "grey80", colour = "blue") +
+     scale_x_discrete() + xlab("Treatment Group") +
+     ylab("Dried weight of plants")
> |
```

boxplot



결과

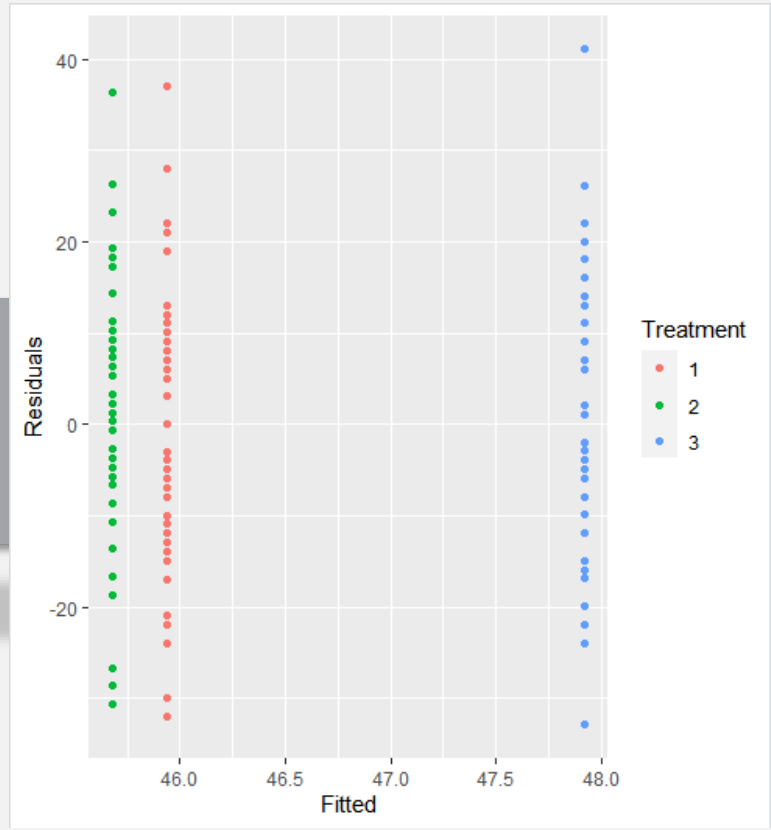
ggplot 패키지를 통해 평균의 동일성 검증을 시각화해 보았다.
이 때, 파란색 선이 평균인데, 서로 많은 차이가 없다는 것을 알 수 있다. 즉, 귀무가설 H_0 : $\mu_1 = \mu_2 = \mu_3$가 기각되지 않은 이유를 시각적으로 알 수 있다.

7 모집단이 세 개 이상일 경우의 평균 비교 검정

- Data.mod() 를 통해 데이터가 어떻게 분포되어 있는지 시각화하기

R코드

```
> data.mod=data.frame(Fitted = fitted(data.mod1),
+ Residuals=resid(data.mod1),Treatment=ad$scale)
> ggplot(data.mod, aes(Fitted, Residuals, color = Trea
tment))+geom_point()
>
```



Data.mod() 함수를 통해, 표본이 어떻게 분포되어 있는지 시각적으로 확인가능하다

7 모집단이 세 개 이상일 경우의 평균 비교 검정

lm() 함수로 회귀분석과 F-검정 단순화 시키기

```
> data.mod1 = lm(age ~ scale, data = ad)
> summary(data.mod1)

Call:
lm(formula = age ~ scale, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max
-32.920  -8.680  -0.310   9.875  41.080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.940      2.025  22.687   <2e-16
scale2        -0.260      2.864  -0.091    0.928
scale3         1.980      2.864   0.691    0.490

(Intercept) ***
scale2
scale3
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.32 on 147 degrees of freed
om
Multiple R-squared:  0.004955,  Adjusted R-squared:  -
0.008583
F-statistic: 0.366 on 2 and 147 DF,  p-value: 0.6941

>
```

```
> #검정통계량이 저장된 f.t값보다 클 확률을 구해 p.value에
 저장하고 출력
> (p.value <- 1 - pf(f.t, 2, 147))
[1] 0.6941136
>
```

유의 확률은 0.6941136이 나왔다. 유의수준이 0.05일 때, 유의 확률이 이것보다 높으므로, 이는 신뢰할 수 있다.

유의확률

```
> (f.t <- mst / mse)
[1] 0.3660281
>
```

검정통계량 0.3660281

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Anova() 함수로 분산분석표 제시

```
> anova(data.mod1)
Analysis of Variance Table

Response: age
            Df  Sum Sq Mean Sq F value Pr(>F)
scale        2   150.1  75.047   0.366 0.6941
Residuals  147 30139.4 205.030
> |
```

자유도

```
> as.factor(ad$scale) #as.factor() 요인(factor)으로 변환
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [23] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [45] 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [67] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [89] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
[111] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[133] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
Levels: 1 2 3
> nlevels(as.factor(ad$scale)) #ad$scale의 factor 개수#
nlevels():level의 개수(출력 결과의 개수가 처리집단의 수)
[1] 3
> (dft <- nlevels(as.factor(ad$scale)) - 1)
[1] 2
>
> dft<-3-1
> dft
[1] 2
> #식을 이용하여 계산 할 수도 있다.
> |
> (dfe <- (length(y1)) + (length(y2)) + (length(y3))-
3)
[1] 147
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Anova() 함수로 분산분석표 제시

```
> anova(data.mod1)
Analysis of Variance Table

Response: age
           Df  Sum Sq Mean Sq F value Pr(>F)
scale       2   150.1  75.047   0.366 0.6941
Residuals 147 30139.4 205.030
> |
```

sst

```
> (sst <- sst.1 + sst.2 + sst.3)
[1] 150.0933
> |
```

sse

```
> (sse <- sse.1 + sse.2 + sse.3)
[1] 30139.38
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Anova() 함수로 분산분석표 제시

```
> anova(data.mod1)
Analysis of Variance Table

Response: age
           Df  Sum Sq Mean Sq F value Pr(>F)
scale       2   150.1  75.047   0.366 0.6941
Residuals 147 30139.4 205.030
> |
```

검정통계량

```
> (f.t <- mst / mse)
[1] 0.3660281
> |
```

유의확률

```
> #검정통계량이 저장된 f.t값보다 클 확률을 구해 p.value에
 저장하고 출력
> (p.value <- 1 - pf(f.t, 2, 147))
[1] 0.6941136
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

confint() 함수로 회귀분석시 95% 신뢰구간 구하기

```
> confint(data.mod1)
                2.5 %    97.5 %
(Intercept) 41.938142 49.941858
scale2      -5.919482  5.399482
scale3      -3.679482  7.639482
```

Confint() 함수를 사용하면, 회귀분석을 할 때 신뢰구간을 구할 수 있다. 이 때, 다른 신뢰구간을 구하기 위해서는 뒤를 바꾸어 주면 된다.
현재는 기본 값인 confint(res, level=0.95)로 되어있지만, level을 바꾸면 다른 신뢰구간을 구하는 것 역시 가능하다.

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Aov () 함수로 분산분석표 제시하기

```
> r.aov<-aov(age~scale, data=ad)
> summary(r.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
scale         2    150   75.05   0.366  0.694
Residuals   147  30139  205.03
> |
```

자유도

```
> as.factor(ad$scale) #as.factor() 요인(factor)으로 변환
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [23] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [45] 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [67] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [89] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
[111] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[133] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
Levels: 1 2 3
> nlevels(as.factor(ad$scale)) #ad$scale의 factor 개수#
nlevels():level의 개수(출력 결과의 개수가 처리집단의 수)
[1] 3
> (dft <- nlevels(as.factor(ad$scale)) - 1)
[1] 2
>
> dft<-3-1
> dft
[1] 2
> #식을 이용하여 계산 할 수도 있다.
> |
> (dfe <- (length(y1)) + (length(y2)) + (length(y3))-
3)
[1] 147
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Aov () 함수로 분산분석표 제시하기

```
> r.aov<-aov(age~scale, data=ad)
> summary(r.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
scale         2    150   75.05   0.366  0.694
Residuals   147  30139  205.03
> |
```

sst

```
> (sst <- sst.1 + sst.2 + sst.3)
[1] 150.0933
> |
```

sse

```
> (sse <- sse.1 + sse.2 + sse.3)
[1] 30139.38
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Aov () 함수로 분산분석표 제시하기

```
> r.aov<-aov(age~scale, data=ad)
> summary(r.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
scale         2    150   75.05   0.366  0.694
Residuals   147  30139  205.03
> |
```

검정통계량

```
> (f.t <- mst / mse)
[1] 0.3660281
> |
```

유의확률

```
> #검정통계량이 저장된 f.t값보다 클 확률을 구해 p.value에
 저장하고 출력
> (p.value <- 1 - pf(f.t, 2, 147))
[1] 0.6941136
> |
```

7 모집단이 세 개 이상일 경우의 평균 비교 검정

Bartlett.test() 함수로 바틀렛 검정을 통하여 검정하기

```
> bartlett.test(age~scale,data=ad)

        Bartlett test of homogeneity of variances

data:  age by scale
Bartlett's K-squared = 0.41183, df = 2,
p-value = 0.8139

> |
```

이 때, 유의확률이 0.8139이므로, 유의수준 0.05에서 귀무가설을 기각하지 못함
그러므로, 등분산 가정을 만족한다고 볼 수 있다.
이렇게 등분산 가정의 만족 유무를 따지기 위해서는 bartlett함수로 판별하면 된다.

8. 일원배치 분산 분석

8 일원배치 분산분석

일원배치 분산분석

적어도 한 개의 집단엔 차이를 보인다라는 대립가설과 모든 집단이 같다라는 귀무가설을 세우고 어떤 가설이 타당한지 따지는 것이 일원배치 분산분석이다.
단, 일원배치 분산 분석이 t-검정과 다른 이유는, 독립변수가 3개 이상이라는 것이다. 그래서 임의로 독립변수가 3개인 데이터를 만들고,
일원배치 분산분석을 통해 집단간의 차이 여부를 일원배치 분산분석으로 파악하겠다.

가설 세우기

매일 측정하였을 때, 온도 상태에 따라서 나오는 결과의 차이가 존재하는가라는 의문에 대해서,
귀무가설은 차이가 없다, 대립가설은 차이가 있다고 가설을 세우고 시작한다.

8 일원배치 분산분석

데이터 만들기

```
> #일원배치 분산분석
> ##데이터 만들기
> ###매일 측정하였을 때, 온도 상태에 따라서 나오는 결과의 차이가 존재하는가?
> ###라는 가설을 세우고, 귀무가설은 차이가 없다. 대립가설은 차이가 있다로 검정은 시작된다.
> ###이 때, 온도 상태의 그룹은 3개가 있는데, group 1,2,3이 바로 그 것이다.
> # group 1 : temperature condition 1
> # group 2 : temperature condition 2
> # group 3 : temperature condition 3
> ###이 때, 우리는 그 온도 상태에 대한 각각의 결과값들을 r에 넣으려고 한다.
> ###group 1,2,3에 해당하는 각각의 결과값들을 y1,2,3에 저장한다.
> y1 <- c(50.5, 52.1, 51.9, 52.4, 50.6, 51.4, 51.2, 52.2, 51.5, 50.8)
> y2 <- c(47.5, 47.7, 46.6, 47.1, 47.2, 47.8, 45.2, 47.4, 45.0, 47.9)
> y3 <- c(46.0, 47.1, 45.6, 47.1, 47.2, 46.4, 45.9, 47.1, 44.9, 46.2)
> y1
 [1] 50.5 52.1 51.9 52.4 50.6 51.4 51.2 52.2 51.5
[10] 50.8
> y2
 [1] 47.5 47.7 46.6 47.1 47.2 47.8 45.2 47.4 45.0
[10] 47.9
> y3
 [1] 46.0 47.1 45.6 47.1 47.2 46.4 45.9 47.1 44.9
[10] 46.2
>
```

group 1,2,3에 해당하는 각각의 결과값들을 y1,2,3 저장

```
> ###그 후, 모든 결과값들을 y에 저장한다.
> y <- c(y1, y2, y3)
> y
 [1] 50.5 52.1 51.9 52.4 50.6 51.4 51.2 52.2 51.5
[10] 50.8 47.5 47.7 46.6 47.1 47.2 47.8 45.2 47.4
[19] 45.0 47.9 46.0 47.1 45.6 47.1 47.2 46.4 45.9
[28] 47.1 44.9 46.2
>
```

모든 결과값들을 y에 저장

```
> ###그 다음, 이 값들을 데이터 프레임으로 옮기기 위한 사전 절차를 실시할 것이다.
> ###데이터 프레임에 그냥 자료를 넣게되면, 어지러워질 수도 있기 때문에, 각각의 그룹에 해당하는
> ###자료들 앞에 1,2,3,을 붙여주어야 한다. 그 사전작업을 실시한다.
> n <- rep(10, 3)
> n
[1] 10 10 10
>
> group <- rep(1:3, n)
> group
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3
[24] 3 3 3 3 3 3 3
>
```

각각의 그룹에 해당하는 자료들 앞에 1,2,3,을 붙여주는 사전작업을 실시

현재, rep()함수를 사용하여, 3개의 10을 만들어 준 후, 각각의 1,2,3을 10번씩 순차대로 반복

8 일원배치 분산분석

데이터 만들기

```
> # data.frame으로  표 만들어 주기. 이 때, 1열에는 y의 값
을 순차적으로 넣어주고, 2열에는
> # group의 값을 넣어주라고 명령한다.
> group_df <- data.frame(y, group)
> group_df
      y group
1  50.5     1
2  52.1     1
3  51.9     1
4  52.4     1
5  50.6     1
6  51.4     1
7  51.2     1
8  52.2     1
9  51.5     1
10 50.8     1
11 47.5     2
12 47.7     2
13 46.6     2
14 47.1     2
15 47.2     2
16 47.8     2
17 45.2     2
18 47.4     2
19 45.0     2
20 47.9     2
21 46.0     3
22 47.1     3
23 45.6     3
24 47.1     3
25 47.2     3
26 46.4     3
27 45.9     3
28 47.1     3
29 44.9     3
30 46.2     3
>
```

데이터 프레임으로 옮기기
data.frame으로 표 만들어 주기
1열에는 y의 값
2열에는 group의 값

```
> ###그 다음, 이 1,2,3들이 숫자가 아니라 범주형 자료이므로,
 이를 변환하기 위한 작업을 한다.
> sapply(group_df, class)
        y     group
"numeric" "integer"
> group_df <- transform(group_df, group = factor(gro
up))
> sapply(group_df, class)
        y     group
"numeric"  "factor"
>
```

그 다음, 이 1,2,3들이 숫자가 아니라
범주형 자료이므로,
이를 변환하기 위한 작업 실시

8 일원배치 분산분석

Aov () 함수로 분산분석표 제시하기

```
> #일원배치분산분석 시행
> aov(y ~ group, data = group_df)
Call:
   aov(formula = y ~ group, data = group_df)

Terms:
                  group Residuals
Sum of Squares  156.302    19.393
Deg. of Freedom       2        27

Residual standard error: 0.8475018
Estimated effects may be unbalanced
> |
```

sst.와 sse를 구해주고, 각각의 자유도 구하기

sst,sse, F-검정통계량, 유의확률 구해주기

```
> summary(aov(y ~ group, data = group_df))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 156.30   78.15   108.8 1.2e-13 ***
Residuals   27  19.39    0.72
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ##sst,sse, F-검정통계량, 유의확률 구해주기
> ##유의 확률이 유의수준 0.05보다 더 매우 작게 나와서 귀무
가설을 기각하고 대립가설을
> ##채택하게 되어 온도 조건 1/2/3에 따라서 결과의 차이가 있
다라고 말할 수 있다.
> |
```

유의 확률이 유의수준 0.05보다 더 매우 작게 나와서 귀무가설을 기각하고
대립가설을 채택하게 되어 온도 조건 1/2/3에 따라서
결과의 차이가 있다 라고 말할 수 있다.

8 일원배치 분산분석

Bartlett.test() 함수로 바틀렛 검정을 통하여 오차의 등분산성 검정하기

```
> #bartlett.test() 함수로 오차의 등분산성 검정하기
> bartlett.test(y ~ group, data = group_df)

        Bartlett test of homogeneity of variances

data:  y by group
Bartlett's K-squared = 1.6565, df = 2,
p-value = 0.4368

> #이 때는 유의확률의 유의수준인 0.05보다 더 크게 나왔으므
로, 오차의 등분산성 가정을 만족한다고
> #말할 수 있을 것이다.
>
```

이 때는 유의확률의 유의수준인 0.05보다 더 크게 나왔으므로,
오차의 등분산성 가정을 만족한다고 말할 수 있을 것이다.


9.정리 – PlantGrowth

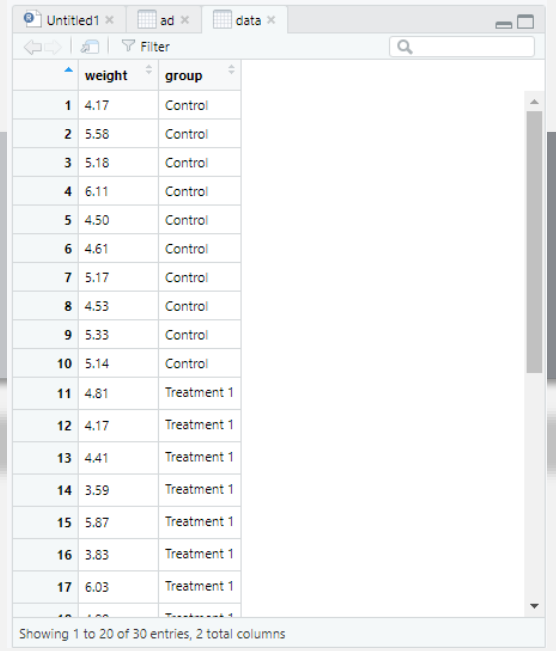
9 정리_PlantGrowth

- PlantGrowth를 이용하여 지금까지 했던 내용 복습하기

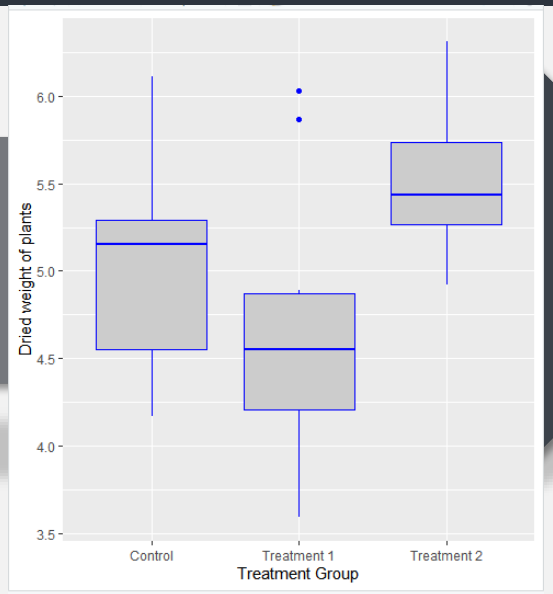
R코드

```
> ###R에서 제공하는 데이터인 PlantGrowth를 이용하여 boxpl
ot을 그려보고
> ###회귀분석 데이터와 F검정값과 유의확률 이끌어내기
> data <- PlantGrowth
> View(data)
> data$group = factor(data$group, labels = c("Contro
l", "Treatment 1", "Treatment 2"))
> require(ggplot2)
> ggplot(data, aes(x = group, y = weight)) +
+     geom_boxplot(fill = "grey80", colour = "blue")
 +
+     scale_x_discrete() + xlab("Treatment Group") +
+     ylab("Dried weight of plants")
>
```

View



Boxplot



9 정리_PlantGrowth

- PlantGrowth를 이용하여 지금까지 했던 내용 복습하기

```
> data.mod1 = lm(weight ~ group, data = data)
> summary(data.mod1)

Call:
lm(formula = weight ~ group, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
                  Estimate Std. Error t value
(Intercept)         5.0320     0.1971  25.527
groupTreatment 1   -0.3710     0.2788  -1.331
groupTreatment 2    0.4940     0.2788   1.772
                  Pr(>|t|)
(Intercept)         <2e-16 ***
groupTreatment 1    0.1944
groupTreatment 2    0.0877 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6234 on 27 degrees of fre
edom
Multiple R-squared:  0.2641,    Adjusted R-squared:
 0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591

> |
```

Summary()

회귀분석 데이터와 F검정값과 유의확률을 구할 수 있다.


부족한 피피티 봐주셔서 감사합니다!