R을 이용한 통계 기초(5일차)

인하대학교 대학원 통계학과 국성희

Contents

1. 지난 시간 강의 내용 복습
2. 일원 배치 분산분석
3. 연습문제

1. 지난 시간 복습

- 반복문과 조건문, 그리고 결측값 핸들링
- 그래프 그리기
- 범주형 데이터 와 이변량 데이터 분석
- 통계적 가설검정과 T-test, 카이제곱 검정 (연습문제)

2. 분산분석

- 분산분석 : 두 모집단 뿐 아니라 셋 이상의 모집단 간의 평균을 비교하는 방법
  (analysis of variance, ANOVA)

- 여러 모집단에서의 관측자료를 효과적으로 분석하고 해석하게 해 준다.

(ex: 소비자 단체에서 여러 종류의 건전지 중에서 어느 건전지의 수명이 오래가는지
또는 어느 농업 연구가는 여러 품종의 볍씨 중 어느 종의 수확량이 가장 높은 지 등에 관심을 갖는 경우)

2. 분산분석

- 2.1 일원 배치 분산분석

요인(source)	자유도(df)	제곱합(SS)	평균제곱합(MS)	F-통계량	기각값
treatment(처리)	k-1	SSR	MSR	F-통계량	p-value
error(오차)	N-k	SSE	MSE		
Total	N-1	SST			

- 처리효과에 대한 가설 검정

H0 : 처리효과가 없다.
H1 : 적어도 하나는 처리효과가 존재한다.

2. 분산분석

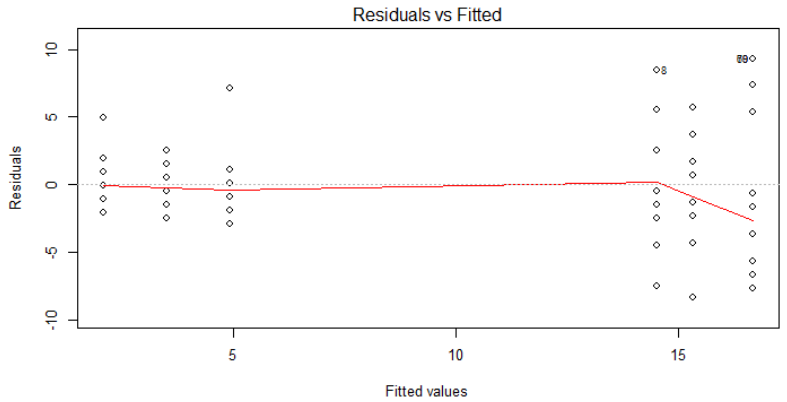
- 2.1 일원 배치 분산분석

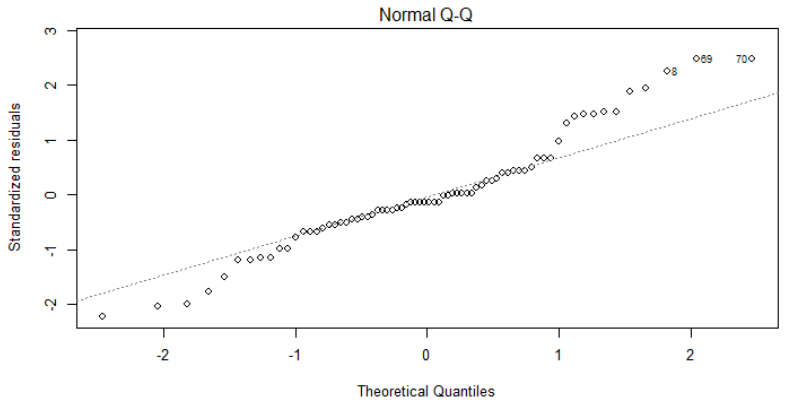
  InsectSprays 데이터를 이용하여 spray종류에 따라 count 평균이 같다고 할 수 있는지 분산분석을 해라.

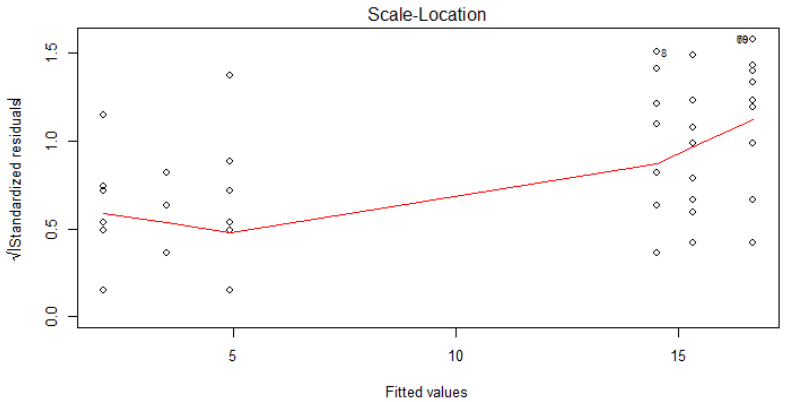
```
data(InsectSprays)

data=InsectSprays
aov.spray1=aov(count~spray, data=data)

summary(aov.spray1)

par(mfrow=c(2,2))
plot(aov.spray1)
```

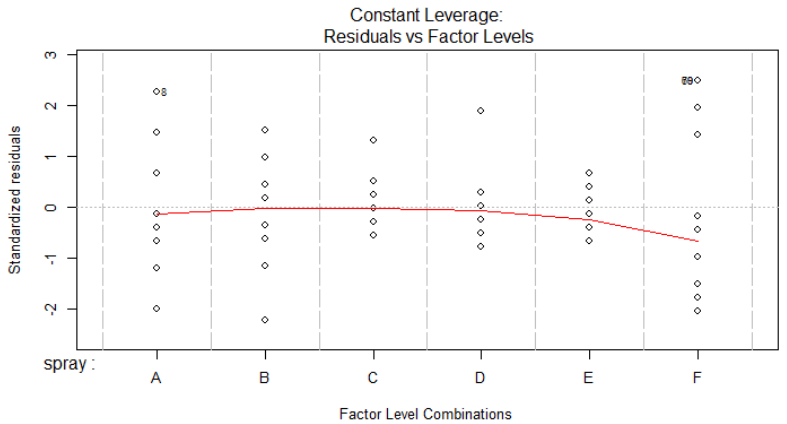
2. 분산분석

- 2.1 일원 배치 분산분석









- 적합값과 잔차 그래프
- Q-Q plot
- 적합값과 sqrt(표준화 잔차)
- 요인값에 대한 표준화 잔차

2. 분산분석

- 2.1 일원 배치 분산분석

  - 세 종류의 건전지 수명에 차이가 있는지 알아보고자 한다. 각 회사에서 5개씩 건전지를 선택하여 수명실험으로 다음의 데이터를 얻었고 이들은 정규분포를 따른다고 할 수 있다고 한다.

  - A : 100, 96, 98, 96, 92
  - B : 76, 80, 84, 84, 78
  - C : 108, 100, 101, 98, 96

2. 분산분석

- 2.1 일원 배치 분산분석

```
a=c(100,96,98,96,92)
b=c(76,80,84,84,78)
c=c(108,100,101,98,96)
life=data.frame(a,b,c)

#data 형태 변형
b.life=stack(life)

#plot 확인
par(mfrow=c(1,2))
boxplot(values~ind, data=b.life)
stripchart(life)
result=aov(values~ind,data=b.life)
summary(result)
```

2. 분산분석

- 2.1 일원 배치 분산분석

```
#data 재 생성
type=c(rep("a",5),rep("b",5),rep("c",5))
y=c(a,b,c)
ty=as.factor(type)
life.aov=aov(y~ty)
summary(life.aov)

#사후분석
life.tukey=TukeyHSD(life.aov, "ty", ordered=T)
life.LSD=pairwise.t.test(y,ty,p.adjust="none",pool.sd=T)
plot(life.tukey)
```

3. 연습문제

- 1. 6쌍의 쌍둥이 형제의 공격성 정도를 측정하는 심리검사를 실시하여 다음의 데이터를 얻었다.

	1	2	3	4	5	6
형	86	71	77	68	91	72
동생	88	77	76	64	96	72

(a) txt파일로 만들어 읽어들 brother.txt로 저장하시오.

(b) (a)에서 만든 brother.txt파일을 읽어 형과 동생 점수를 각각 older, younger에 저장하고 각 변수의 평균과 표준편차를 구하시오

3. 연습문제

- 2. 다음은 학생들의 국어 성적과 영어 성적 데이터이다.

	Korean
kim	93
lee	76
park	87
oh	92
yang	98
min	75
jung	82

	English
kim	90
lee	94
park	88
oh	75
yang	79
min	87
jung	88
choi	90

(a) 두 데이터셋을 합하시오(각 이름이 모두 나오게.. 합해주세요)

(b) (a)에서 만들어진 데이터셋에서 이름을 알파벳순으로 정렬하시오.

3. 연습문제

- 3. UsingR 패키지를 인스톨한 후 내장되어있는 데이터셋 primes를 이용하여 히스토그램을 그리시오

- 4. B회사 직원들의 월별 핸드폰 요금을 조사하여 얻은 데이터이다.

  20870 39400 65000 45000 35890 29000 56770

  23000 38550 59800 39880 56780 35220 48990

(a) 평균 핸드폰 요금을 추정하시오.

(b) 평균 핸드폰 요금의 95% 신뢰구간을 추정하시오.

(c) 핸드폰 요금의 분산과 표준편차와 범위를 추정하시오

(d) 상자그림을 그리시오

(e) 히스토그램을 그리고 density 선을 그리시오

3. 연습문제

- 5. 다음은 기존 자외선 차단 로션과 새로 개발한 자외선 차단 로션 E,F에 대하여 효과 차이가 있는지 알아보고자 한다. 8명의 여성에게 왼쪽 팔에는 E로션을, 오른쪽 팔에는 F 로션을 바르고 일정 시간 지난 후 피부 그을린 정도를 측정하여 데이터를 얻었다.
  두 로션간의 차이가 있다고 할 수 있는지 유의수준 5%에서 검정하시오.

	1	2	3	4	5	6	7	8
E(왼쪽 팔)	90	88	78	65	78	60	89	73
F(오른쪽 팔)	80	78	75	69	73	62	79	70

3. 연습문제

- 6. 여성 화장품 신제품 출시 후 일정량을 바르게 한 후 구매여부에 대한 의견을 조사하여 정리한 데이터이다. 유의수준 5%에서 여성 나이대와 구매여부에 대한 독립성 검정을 하시오.

구매여부	여성나이대		
	20대	30대	40대
안하겠다	24	20	50
고려해보겠다	32	30	20
하겠다	54	60	15

3. 연습문제

- 7. 한 정부기관에서 산업폐수를 정화하는 필터 4종류 A,B,C,D를 비교하여 여과능력에 차이가 있는지 분석하시오

필터종류			
A	B	C	D
26	19	28	29
23	20	28	29
27	20	29	27
	17		34

Q&A