# R을 이용한 기초 통계 분석: 분산분석, 범주형자료 분석

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-04-12

# Agenda

- 분산분석(ANOVA, Analysis of Variance)
- 범주형 자료의 분석
  - 동일성 검정 (Test of homogeneity)
  - 독립성 검정 (Test of independence)

# 분산분석(ANOVA, Analysis of Variance)

1. 여러(3이상) 모집단의 평균비교

# 기본 용어

1. 요인 (factor): 실험에 고려한 설명변수
   - 요인의 수가 하나이면 일원분산분석 (one-way ANOVA)
2. 수준 (level): 요인이 취하는 값
3. 처리 (treatment): 요인과 수준의 조합

# 통계 모형

- 자료의 형태

| | | 요인의 수준(처리) | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | ⋯ | $p$ |
| 반복수 | 1 | $Y_{11}$ | $Y_{21}$ | ⋯ | $Y_{p1}$ |
| | 2 | $Y_{12}$ | $Y_{22}$ | ⋯ | $Y_{p2}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | $Y_{1n_1}$ | $Y_{2n_2}$ | ⋮ | $Y_{pn_p}$ |

- 모형

$$Y_{ij} = \mu + a_i + \epsilon_{ij}, i = 1, \ldots, p, j = 1, \ldots, n_i$$

  - $\mu$ : 전체 평균
  - $a_i$ : $i$번째 처리효과, $\sum_{i=1}^{p} a_i = 0$

# 통계적 추론

- 처리효과($a_i$)의 추정
- 처리효과($a_i$)의 차이 검정: $H_0 : a_1 = \ldots = a_p = 0$

# 변동의 분해 및 분산분석표

- 분산분석표

| | 자유도 | 제곱합 | 평균제곱 | F |
|---|---|---|---|---|
| 처리 | p-1 | SStr | MStr | MStr/MSE |
| 오차 | n-p | SSE | MSE | |
| 전체 | n-1 | SST | | |

- 세개이상의 모평균차에 대한 검정법: F-검정

$$F = \frac{\sum_{i=1}^{p} n_i(\overline{Y}_i - \overline{Y})^2/(p-1)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2/\sum_{i=1}^{p}(n_i-1)} \sim F_{p-1,N-p}$$

# 예제: InsectSprays

- 6종의 살충제를 뿌리고 죽은 해충의 수를 조사

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 10 | 11 | 0 | 3 | 3 | 11 |
| 7 | 17 | 1 | 5 | 5 | 9 |
| 20 | 21 | 7 | 12 | 3 | 15 |
| 14 | 11 | 2 | 6 | 5 | 22 |
| 14 | 16 | 3 | 4 | 3 | 15 |
| 12 | 14 | 1 | 3 | 6 | 16 |
| 10 | 17 | 2 | 5 | 1 | 13 |
| 23 | 17 | 1 | 5 | 1 | 10 |
| 17 | 19 | 3 | 5 | 3 | 26 |
| 20 | 21 | 0 | 5 | 2 | 26 |
| 14 | 7 | 1 | 2 | 6 | 24 |
| 13 | 13 | 4 | 4 | 4 | 13 |

# 예제: InsectSprays (conti.)

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

# 예제: InsectSprays (conti.)

*. Descriptive statistics: Mean, variance, number of elements in each cell

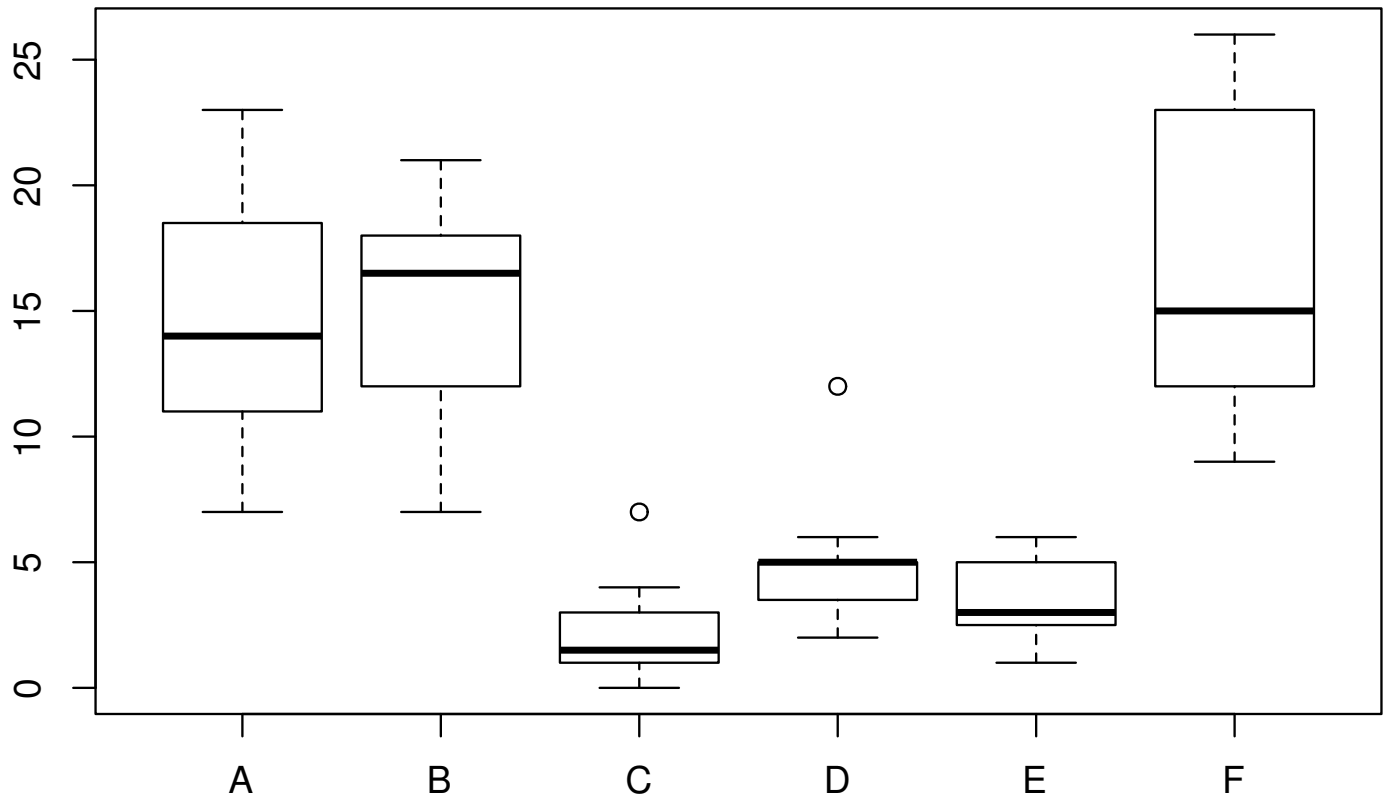
```
m <- tapply(count, spray, mean)
s <- tapply(count, spray, sd)
rbind(mean=m, sd=s)
```

```
##               A         B        C        D        E
## mean 14.500000 15.333333 2.083333 4.916667 3.500000
## sd    4.719399  4.271115 1.975225 2.503028 1.732051
##               F
## mean 16.666667
## sd    6.213378
```

# 예제: InsectSprays (conti.)

- Visualise the data – boxplot; look at distribution, look for outliers

```
boxplot(count ~ spray)
```



# 예제: InsectSprays (conti.)

- ANOVA table

```
o <- aov(count ~ spray)
summary(o)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 가설검정 :
  - p-value $< 0.001$ : Reject $H_0 : a_1 = \ldots = a_p = 0$

# 효과의 추정

```
o <- lm(count ~ spray)
summary(o)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  14.5000000   1.132156 12.8074279 1.470512e-19
## sprayB        0.8333333   1.601110  0.5204724 6.044761e-01
## sprayC      -12.4166667   1.601110 -7.7550382 7.266893e-11
## sprayD       -9.5833333   1.601110 -5.9854322 9.816910e-08
## sprayE      -11.0000000   1.601110 -6.8702352 2.753922e-09
## sprayF        2.1666667   1.601110  1.3532281 1.805998e-01
```

- Estmate for each effect: $\mu_j - \mu_A$, $j \in \{B, C, D, E, F\}$
  - $\mu_B - \mu_A = 0.83$
  - $\mu_C - \mu_A = -12.42$

# Post Hoc tests (multiple comparison, 다중비교)

- Tukey HSD(Honestly Significant Difference)

```
o<-aov(count ~ spray)
TukeyHSD(o)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = count ~ spray)
##
## $spray
##             diff        lwr       upr     p adj
## B-A    0.8333333  -3.866075  5.532742 0.9951810
## C-A -12.4166667 -17.116075 -7.717258 0.0000000
## D-A  -9.5833333 -14.282742 -4.883925 0.0000014
## E-A -11.0000000 -15.699409 -6.300591 0.0000000
## F-A   2.1666667  -2.532742  6.866075 0.7542147
## C-B -13.2500000 -17.949409 -8.550591 0.0000000
## D-B -10.4166667 -15.116075 -5.717258 0.0000002
## E-B -11.8333333 -16.532742 -7.133925 0.0000000
## F-B   1.3333333  -3.366075  6.032742 0.9603075
## D-C   2.8333333  -1.866075  7.532742 0.4920707
```

# Post Hoc tests (multiple comparison, 다중비교)

- A-B-F 와 C-D-F에 대하여 그룹화 가능함 : 그룹화 후 재 분석 필요

# 범주형자료의 분석

- 질적자료 또는 범주화된 양적자료의 분석
- 분할표 (contingency table): 범주형 자료의 분석에 사용하는 테이블 형태의 자료
  - 열 또는 행은 요인(범주형 변수)의 수준
  - 셀은 요인의 각 수준에 해당되는 관측치의 빈도

| X \ Y | $b_1$ | $b_2$ | ... | $b_c$ | 계 (total) |
|---|---|---|---|---|---|
| $a_1$ | $O_{11}$ | $O_{12}$ | ... | $O_{1c}$ | $O_{1\cdot}$ |
| $a_2$ | $O_{21}$ | $O_{22}$ | ... | $O_{2c}$ | $O_{2\cdot}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $a_r$ | $O_{r1}$ | $O_{r2}$ | ... | $O_{rc}$ | $O_{r\cdot}$ |
| 계 (total) | $O_{\cdot 1}$ | $O_{\cdot 2}$ | ... | $O_{\cdot c}$ | n |

Figure: 2원 분할표

# 범주형자료의 분석

1. 적합도 검정(goodness of fit test) : 분할표의 값들이 특정 분포를 따르고 있는지를 검정
2. 독립성 검정(test of independence) : 분할표에서 두 범주형 변수(요인)들이 서로 독립인지를 검정
   - $H_0 : p_{ij} = p_i p_j, i = 1, \ldots, n, j = 1, \ldots, m$
3. 동질성 검정(test of homogeneity) : 서로 다른 부모집단(subpopulation)에서 범주형 변수의 확률 분포가 서로 동일한지를 검정
   - $H_0 : p_{1j} = p_{2j} = \ldots p_{nj}, j = 1, \ldots, m$

- Test-statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2((r-1)(c-1))$$

# 예제: 설문조사 data (survey)

- University of Adelaide에서 237명의 학생을 대상으로 설문조사한 결과
- 변수 및 속성

| 변수 | 설명 | 변수값(범주형) |
|---|---|---|
| Sex | 성별 | “Male”, “Female” |
| Wr.Hnd | 글쓰는 손의 한뼘의 길이 (단위: Cm) | |
| NW.Hnd | 반대쪽 손의 한뼘 길이 | |
| W.Hnd | 글쓰는 손의 위치 | “Left”, “Right” |
| Fold | 팔을 접었을 때 양손의 위치 | “R on L”, “L on R”, “Neither” |
| Pulse | 혈압 (beats per minute) | |
| Clap | 박수칠 때 위로 올라가는 손 | “Right”, “Left”, “Neither” |
| Exer | 운동의 빈도 | “Freq”, “Some”, “None” |
| Smoke | 흡연의 정도 | “Heavy”, “Regul”, “Occas”, “Never” |
| Height | 키 (Cm) | |
| Age | 나이 | |

# 예제: 독립성 검정

- 운동의 빈도(Exer)와 흡연정도(Smoke)에 대한 분할표 (contingency table)

```
library(MASS)
(x <- table(survey$Exer, survey$Smoke))
```

```
##
##         Heavy Never Occas Regul
##    Freq     7    87    12     9
##    None     1    18     3     1
##    Some     3    84     4     7
```

# 예제: 독립성 검정

- Pearson's Chi-squared test: 운동의 빈도(Exer)와 흡연정도(Smoke)가 서로 독립인가?

```
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

# 예제: 동일성 검정

- 성별(Sex)과 흡연정도(Smoke)에 대한 분할표 (contingency table)

```
library(MASS)
(o <- table(survey$Sex, survey$Smoke))
```

```
##
##           Heavy Never Occas Regul
##   Female      5    99     9     5
##   Male        6    89    10    12
```

# 예제: 동일성 검정

- Pearson's Chi-squared test: 남,여 그룹에서 흡연정도(Smoke)의 비율이 서로 동일한가?

```
chisq.test(o)
```

```
##
##  Pearson's Chi-squared test
##
## data:  o
## X-squared = 3.5536, df = 3, p-value = 0.3139
```