Analiza wariancji - ANOVA

Bartosz Kozak

6 maja 2019

Wprowadzenie

Na dzisiejszych zajęciach zapoznamy się z jednoczynnikową analizą wariancji - ANOVA. Nasze rozważania rozpoczniemy od szybkiego zapoznania z teorią stanowiącą podstawę tej analizy.

Niech x_{ij} oznacza *j-otą* obserwację w *i-tej* grupie, wtedy x_{35}, oznacza piątą obserwację w trzeciej grupie; $\overline{x}_i$ oznacza średnią dla *i-tej* grupy, a $\overline{x}_.$ oznacza średnią całkowitą (średnią z wszystkich obserwacji).

Możemy rozbić obserwację na czynniki jako:

$$x_{ji} = \overline{x}_. + (\overline{x}_i - \overline{x}_.) + (x_{ij} - \overline{x}_i) \tag{1}$$

gdzie:

$\overline{x}_i - \overline{x}_.$ - odpowiada różnicy średniej grupowej od średniej całkowitej (wpływ grupy)

$x_{ij} - \overline{x}_i$ - odpowiada różnicy obserwacji od średniej grupowej (wpływ błędu obserwacji)

Koresponduje to z modelem

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij}\ N(\mu, \sigma) \tag{2}$$

gdzie:

α_i - oznacza wpływ czynnika (grupy)

ϵ_{ij} - oznacza błąd pomiaru

W modelu tym zakładamy hipotezę zerową, że wpływ czynnika (α_i) jest równy zero dla wszystkich badanych czynników. Zauważmy, że zakładamy niezależność czynnika błędu, oraz że ma on tą samą wariancję dla wszystkich czynników.

Rozważmy teraz sumy kwadratów powyższych czynników znanych jako *wariancję wewnątrzgrupową* - SSD_W

$$SSD_W = \sum_i \sum_j (x_{ij} - \overline{x}_i)^2 \tag{3}$$

oraz *wariancję międzygrupową* SSD_B

$$SSD_B = \sum_i \sum_j (x_i - \overline{x}_.)^2 = \sum_i n_i (x_i - \overline{x}_.)^2 \tag{4}$$

Możemy też wykazać, że

$$SSD_W + SSD_B = SSD_{całkowita} = \sum_i \sum_j (x_{ij} - \overline{x}_.)^2 \tag{5}$$

Należy jednak zaznaczyć, że suma kwadratów (SS) może być tylko pozytywna, więc nawet kompletnie nieinformatywne (przypadkowe) grupowanie wyjaśni jakąś część obserwowanej zmienności. Nasuwa się

więc pytanie jak małe wyjaśnienie całkowitej wariancji możemy uznać za przypadkowe? Okazuje się, że w przypadku braku systematycznych różnic między grupami należy oczekiwać, że suma kwadratów podzielona przez liczbę stopni swobody dla każdego czynniki, $k-1$ dla SSD_B oraz $N-k$ dla SSD_W, gdzie k oznacza liczbę grup, a N oznacza całkowitą liczbę obserwacji (sumę liczby obserwacji z wszystkich grup).

Możemy więc obliczyć *średnie kwadraty* jako:

$$MS_W = SSD_W/(N-k) \tag{6}$$

$$MS_B = SSD_B/(k-1) \tag{7}$$

MS_W jest połączoną wartością wariancji poszczególnych grup i jest estymatorem σ^2. W przypadku braku wpływu czynnika grupy MS_B także jest estymatorem σ^2, jednakże jeżeli wpływ czynnika grupy będzie istotny wtedy różnice między grupami, a co za tym idzie MS_B będzie "duże". Dlatego test na różnice między średnimi grup może zostać wykonany jako porównanie dwóch estymatorów wariancji. Dlatego opisywana procedura pomimo tego że służy do porównania średnich między grupami nosi nazwę *analizy wariancji*. Z poprzednich zajęć wiemy, że do porównania wariancji wykorzystujemy *testF* oraz dystrybucję F. Wartość statystyki testowej obliczamy ze wzoru

$$F = MS_B/MS_W \tag{8}$$

i porównujemy z wartością graniczną F_α z rozkładu F o liczbie stopni swobody: $(k-1)$ oraz $(N-k)$.

Testujemy hipotezę zerową postaci:

$$H_0 : \frac{MS_B}{MS_W} = 1$$

wobec hipotezy alternatywnej:

$$H_A : \frac{MS_B}{MS_W} \neq 1$$

Analiza wariancji - algorytm postępowania

Przykład 1

Przeprowadzono eksperyment w celu porównania trzech pożywek (A, B, C). Obliczono liczbę koloni bakterii (w trzech powtórzeniach) dla każdej pożywki. Otrzymane dane zapisano w pliku Cw5_1.txt. Czy pożywki wpływają na liczbę kolonii bakterii?

Do odpowiedzi na postawione pytanie wykorzystamy jednoczynnikową analizę wariancji "One-way ANOVA".

Algorytm postępowania

1. Formujemy H_0 i H_A.

H_0 : pożywka **nie wpływa** na liczbę kolonii. Jest to równoważne stwierdzeniu, że średnie liczby kolonii dla poszczególnych pożywek **są równe**.

H_A: Pożywka **wpływa** na liczbę kolonii bakterii. Jest to równoważne stwierdzeniu, że średnie liczby kolonii dla poszczególnych pożywek **nie są równe**.

2. Obliczamy statystykę F
3. Znajdujemy wartość krytyczną $F\alpha$ (dla zadanej α i odpowiednich df)
4. Porównujemy F z $F\alpha$
5. Wnioskujemy

```r
# wczytanie danych do R
dane
```

```
##   M_A M_B M_C
## 1   3   5   5
## 2   2   3   6
## 3   1   4   7
```

ad 2. Obliczamy Statystykę F.

Korzystamy z wzoru 3 i obliczamy SSD_W.

```r
# obliczamy średnie dla grup A,B oraz C
c(sr_a,sr_b,sr_c)
```

```
## [1] 2 4 6
```

```r
# obliczamy sumy kwadratów dla grup A, B oraz C
c(ss_a,ss_b,ss_c)
```

```
## [1] 2 2 2
```

```r
# obliczamy ssd_w
ssd_w
```

```
## [1] 6
```

Korzystamy z wzoru 5 i obliczamy $SSD_{całkowity}$

```r
# obliczamy średnią całkowitą
GM
```

```
## [1] 4
```

```r
# obliczamy ssd_c
ssd_c
```

```
## [1] 30
```

Następnie ponownie korzystamy z wzoru 5 i obliczamy SSD_B

```r
# obliczamy ssd_b
ssd_b
```

```
## [1] 24
```

W oparciu o wzory 6 i 7 obliczamy MS_B oraz MS_W

```r
# liczba obserwacji N
N
```

```
## [1] 9
```

```r
# liczba grup k
k
```

```
## [1] 3
```

```r
# obliczamy ssd_w
ms_b
```

```
## [1] 12
```

```r
# obliczamy ms_w
ms_w
```

```
## [1] 1
```

W oparciu o wzór 8 obliczamy statystykę testową F

```
# obliczamy F
F
```

```
## [1] 12
```

ad3 Znajdujemy wartość krytyczną F_α, dla $\alpha = 0.05$ i liczby stopni swobody $k-1$ oraz $N-k$

```
# znajdujemy Fa
Fa <- qf(0.95,(k-1),(N-k))
Fa
```

```
## [1] 5.143253
```

Po porównaniu F z F_α. Wyniki analizy anova zwyczajowo prezentuje się w tabeli analizy wariancji.

```
# wektor z wartościami sum kwadratów
ss <- c(ssd_b,ssd_w,ssd_c)
# wektor z wartościami średnich sum kwadratów
ms <- c(ms_b,ms_w,"")
# wektor z liczbami stopni swobody
df <- c((k-1),(N-k),(N-1))
# wektor z wartościami statystyki F
f <- c(F,"","")
# wektor z wartościami F krytyczne
fa <- c(round(Fa,2), "","")
# tworzymy tabelę
anova_tabela <- as.data.frame(matrix(c(df,ss,ms,f,fa),nrow = 3))
# zmieniamy nazwy kolumn i wierszy
colnames(anova_tabela) <- c("df", "SS", "MS", "F", "Fa")
row.names(anova_tabela) <- c("pożywka", "błąd", "całkowita")
# wyświetlamy tabele analizy wariancji
anova_tabela
```

```
##           df SS MS  F   Fa
## pożywka    2 24 12 12 5.14
## błąd       6  6  1
## całkowita  8 30
```

Jaki wniosek możemy wyciągnąć względem testowanej hipotezy (H_0)?

Anova w R

ANOVA może być wykonana w R wyłącznie na danych typu `stack`. Musimy więc przekształcić dane.

```
dane1 <- stack(dane)
colnames(dane1) <- c("liczba.kolonii","pozywka")
dane1
```

```
##   liczba.kolonii pozywka
## 1              3     M_A
## 2              2     M_A
## 3              1     M_A
## 4              5     M_B
## 5              3     M_B
## 6              4     M_B
```

```
## 7              5     M_C
## 8              6     M_C
## 9              7     M_C
```

Do wykonania analizy ANOVA w R użyjemy funkcji `lm` oraz `anova`

```
# budujemy model
model <- lm(liczba.kolonii~pozywka,data=dane1)
# wykonujemy ANOVA
anova(model)
```

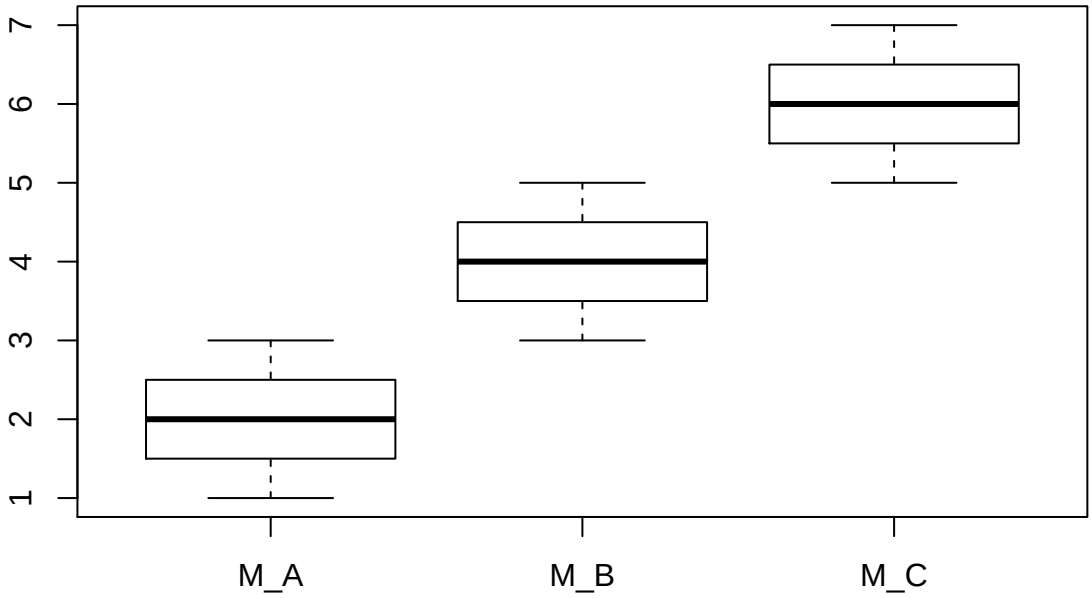
```
## Analysis of Variance Table
##
## Response: liczba.kolonii
##           Df Sum Sq Mean Sq F value Pr(>F)
## pozywka    2     24      12      12  0.008 **
## Residuals  6      6       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wartość `Pr` w tebeli wariancji odpowiada wartości pvalue znanej z testu t. Ponieważ wartość `Pr` jest mniejsza od α odrzucamy H_0 na rzecz H_A z prawdopodobieństwem 0.95%.

Graficzna prezentacja wyników - wykres pudełkowy

Często wykorzystywaną metodą prezentacji wyników w pakietach statystycznych są wykresy pudełkowe "box-plot". W programie R do stowrzenia tego typu wykresów używamy funkcji `boxplot`.

```
boxplot(liczba.kolonii~pozywka,data=dane1)
```



Zadanie do samodzielnego rozwiązania

W celu zbadania plonowania pomidora na różnego rodzaju podłożach zostało w szklarni założone doświadczenie metodą kompetnej randomizacji w 6 powtórzeniach (na powtórzenie składała się skrzynka lub mata z trzema roślinami). Badano następujące rodzaje podłoża:

- A słoma żytnia
- B słoma żytnia + torf (3:1)

- C słoma pszenna
- D słoma pszenna + kora (3:1)
- E wełna mineralna

Badano plon z rośliny (kg/roślina) w 22 tygodniu, dane zamieszczono w pliku Cw5_2.txt. Wykonaj analizę wariancji wg właściwego modelu, aby stwierdzić czy zastosowane podłoże ma pływ na uzyskany plon pomidorów (α=0.05)? Przedstaw uzyskane średnie plony dla badanych podłoży graficznie (wykres pudełkowy, lub wykres słupkowy) .