Rachunek Prawdopodobieństwa i Elementy Statystyki Matematycznej

Anna Dembińska

Wydział Matematyki i Nauk Informacyjnych

Wykład 14

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

14. JEDNOCZYNNIKOWA ANALIZA WARIANCJI

Analiza wariancji (skrót ANOVA pochodzi od angielskiego **an**alysis **o**f **va**riance) to procedura służąca do porównywania średnich w wielu grupach. Jednoczynnikowa ANOVA pozwala ona na testowanie hipotezy

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

przeciwko hipotezie

$$H_1 : \text{ istnieją } i \neq j \text{ takie, że } \mu_i \neq \mu_j,$$

gdzie $k \geqslant 2$ jest liczbą badanych grup i w szczególności umożliwia sprawdzenie czy istnieje zależność pomiędzy pewnymi cechami - dokładniej czy zmienna zwana *zmienną objaśnianą* (lub *zmienną odpowiedzi*) zależy od tzw. *zmiennej objaśnianej* (nazywanej też *czynnikiem*).

Przykład 14. 1. Analiza wariancji może posłużyć np. do sprawdzenia czy papier wytwarzany przez trzech producentów różni się pod względem jaskrawości (i jeśli tak, to który z tych trzech producentów zapewnia największą jaskrawość); wtedy

- zmienna objaśniana (zmienna odpowiedzi) to jaskrawość,
- zmienna objaśniająca (czynnik) to producent, czynnik ten występuje na trzech poziomach: producent I, producent II i producent III ;

Problem czy papier wytwarzany przez trzech producentów różni się pod względem jaskrawości, możemy rozwiązywać testując hipotezę

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

przeciwko hipotezie

$$H_1 : \text{nie wszystkie } \mu_1, \mu_2, \mu_3 \text{ są równe,}$$

gdzie μ_1, μ_2, μ_3 oznaczają średnią jaskrawość papieru wytworzonego odpowiednio przez I, II i III producenta. Odrzucenie H_0 będzie świadczyło o istnieniu zależności jaskrawości papieru od tego, od którego producenta papier pochodzi.

Załóżmy, że aby sprawdzić czy papier, pochodzący od trzech producentów, różni się jaskrawością, zbadano po pięć losowo wybranych próbek papieru, pochodzących od każdego producenta, i otrzymane wyniki (wartości współczynnika odbicia) zebrano w poniższej tabelce:

producent I	producent II	producent III
60,5	60,3	60,0
60,6	60,9	60,4
60,5	60,5	59,9
61,0	60,7	60,2
60,8	60,7	59,8

.

Wprowadzamy dane do R, tworząc dwie zmienne:

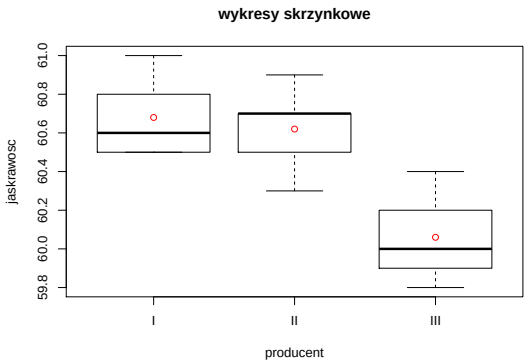
1. zmienną typu liczbowego, nazwaną np. *jaskrawosc*, zwierającą uzyskane wartości współczynnika odbicia;

2. zmienną typu *factor*, nazwaną np. *producent*, zawierającą numer producenta, od którego pochodziła dana próbka papieru:

```
> jaskrawosc=c(60.5,60.6,60.5,61,60.8,60.3,60.9,60.5,60.7,60.7,60,60.4,
                59.9,60.2,59.8)
> producent=rep(c("I","II","III"),rep(5,3))
```

Sam test analizy wariancji warto poprzedzić wstępną analizą danych pod kątem postawionego problemu - obliczyć średnie próbkowe w grupach i sporządzić wykresy skrzynkowe dla jaskrawości papieru pochodzącego od każdego z trzech producentów.

```
> srednie=tapply(jaskrawosc, producent, mean)
> plot(jaskrawosc ~ producent, main="wykresy skrzynkowe",
        xlab="producent",ylab="jaskrawosc")
> lines(1:3,srednie,type="p",col="red")
```



Rysunek 1: Wykresy skrzynkowe dla odpowiedzi w każdej grupie. Kółka oznaczają średnie grupowe.

Wykresy skrzynkowe i wartości średnich grupowych sugerują, że jaskrawość paperu pochodzącego od trzeciego producenta średnio przyjmuje wartości mniejsze niż jaskrawości papieru od pierwszego i drugiego producenta. Chcielibyśmy to przypuszczenie potwierdzić lub odrzucić, przeprowadzając formalne testy statystyczne. W tym celu konstruujemy model matematyczny:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \begin{array}{l} i = 1,2,3 \text{ - nr. producenta,} \\ j = 1,2,3,4,5 \text{ - nr. obserwacji w } i\text{-tej grupie,} \end{array}$$

gdzie

- Y_{ij} - jaskrawość j-tej próbki papieru pochodzącej od i-tego producenta,
- $\mu+\alpha_i$ - średnia jaskrawość papieru pochodzącego od i-tego producenta,
- α_i - efekt i-tego producenta,
- ε_{ij} - błąd losowy dla j-tej próbki papieru pochodzącej od i-tego producenta.

Ogólnie model jednoczynnikowej analizy wariancji jest następujący:

$$\boxed{Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \begin{array}{l} i = 1,2,\ldots,k,\ j = 1,2,\ldots,n \\ k \text{ - liczba pozimów czynnika} \\ n \text{ - liczba obserwacji na każdym pozimie czynnika,} \end{array}}$$

gdzie

- Y_{ij} - wartość zmiennej odpowiedzi dla j-tej obserwacji w i-tej grupie,
- $\mu + \alpha_i$ - wartość średnia zmiennej odpowiedzi w i-tej grupie,
- α_i - efekt i-tej grupy,
- ε_{ij} - błąd losowy dla j-tej obserwacji w i-tej grupie.

W powyższym modelu zakładamy, że dla każdego poziomu czynnika rozkład zmiennej odpowiedzi jest normalny z taką samą wariancją $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \overset{\text{ozn.}}{=} \sigma^2$. Założenie to jest równoważne założeniu, że błędy losowe ε_{ij} też mają rozkłady normalne o tej samej wariancji σ^2. Ponieważ pomiary przeprowadzamy niezależnie, tzn. Y_{ij} są niezależne, to także ε_{ij} są niezależne. Reasumując:

$$\boxed{\varepsilon_{ij} \text{ są niezależne o tym samym rozkładzie } \mathcal{N}(0, \sigma^2).}$$

Ponadto założyliśmy, że przeprowadziliśmy doświadczenie z *planem zrównoważonym* - dla każdego poziomu czynnika mamy taką samą liczbę obserwacji wynoszącą n. Wzory opisujące statystykę testową w analizie wariancji wyprowadzimy właśnie przy tym założeniu. Następnie wspominy jak je uogólnić na przypadek grup nierównolicznych.

Aby w modelu jednoczynnikowej analizy wariancji wartości μ oraz α_i, $i = 1, 2, \ldots, k$, były określone jednoznacznie, musimy coś o nich założyć. Tu stosuje się różne konwencje, np.

1). $\alpha_1 + \alpha_2 + \ldots + \alpha_k = 0$

Wtedy

- μ interpretujemy jako ogólną wartość średnią zmiennej odpowiedzi;
- α_i interpretujemy jako efekt działania i-tego pozimu czynnika względem średniej ogólnej.

2). $\alpha_1 = 0$

Wtedy poziom nr 1 przyjmujemy za poziom odniesienia a

- μ interpretujemy jako wartość średnią zmiennej odpowiedzi w grupie, która jest poziomem odniesienia;
- α_i, $i = 2, \ldots, k$ interpretujemy jako efekt działania i-tego pozimu czynnika względem poziomu czynnika nr 1.

Jest to konwencja zaimplementowana w R.

Niezależnie od wybranej konwencji, weryfikujemy hipotezę

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad \Leftrightarrow \quad H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$$

przeciwko hipotezie

$$H_1 : \text{ istnieją } i \neq j \text{ takie, że } \mu_i \neq \mu_j \quad \Leftrightarrow \quad H_1 : \text{ istnieje } i \text{ takie, że } \alpha_i \neq 0.$$

Konstrukcja testu do weryfikacji hipotez w analizie wariancji

Jeśli hipoteza H_0 jest prawdziwa i dla każdego poziomu czynnika zmienna odpowiedzi ma tą samą wariancję, to wewnątrzgrupowe rozproszenie obserwacji (zmienność wewnątrzgrupowa) oraz międzygrupowe rozproszenie obserwacji (zmienność międzygrupowa) będą w przybliżeniu równe.

Zmienność wewnątrzgrupowa mierzona jest jako średnia z wariancji w grupach:

$$\frac{S_1^2 + S_2^2 + \ldots + S_k^2}{k} = \frac{1}{k}\sum_{i=1}^{k} S_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2 \overset{\text{ozn.}}{=} \frac{1}{k(n-1)}SSE,$$

gdzie S_i^2 to wariancja obserwacji w i-tej grupie, tzn.

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

a $\bar{Y}_{i\cdot}$ to średnia obserwacji w i-tej grupie, tzn. $\bar{Y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$.

Zmienność międzygrupową mierzymy jako n razy wariancja obliczona ze średnich w grupach:

$$n \cdot \frac{1}{k-1}\sum_{i=1}^{k}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \frac{1}{k-1} \cdot n\sum_{i=1}^{k}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \overset{\text{ozn.}}{=} \frac{1}{k-1}SSA,$$

gdzie $\bar{Y}_{\cdot\cdot} = \frac{1}{k}\sum_{i=1}^{k}\bar{Y}_{i\cdot} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n} Y_{ij} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}$ to średnia ogólna.

Zachodzi równość (zwana podstawowym równaniem ANOVA):

$$SSE + SSA = SST,$$

gdzie $SST := \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$ jest tzw. całkowitą sumą kwadratów (**t**otal **s**um of **s**quares) taką, że $\frac{1}{kn-1}SST$ to wariancja obliczona na podstawie wszystkich obserwacji (czyli miara zmienności całkowitej).

Jeśli H_0 jest prawdziwa, to $\frac{1}{k(n-1)}SSE$ i $\frac{1}{k-1}SSA$ będą w przybliżeniu równe. Natomiast jeśli H_0 jest fałszywa, to $\frac{1}{k-1}SSA$ będzie miało tendencję do przyjmowania wartości większych niż $\frac{1}{k(n-1)}SSE$. Stąd pomysł na statystykę testową

$$F = \frac{\frac{1}{k-1}SSA}{\frac{1}{k(n-1)}SSE}.$$

Jeśli

- F jest "duże", to odrzucamy H_0,
- F jest "małe", to przyjmujemy H_0.

Jak ocenić czy wartość F jest duża czy mała? Można pokazać, że gdy H_0 jest prawdziwa, to statystyka testowa F ma rozkład F-Snedecora o $k-1$ i $k(n-1)$ stopniach swobody:

$$F \overset{H_0}{\sim} F_{k-1,k(n-1)}.$$

Stąd jeśli $F \geqslant f_{1-\alpha;k-1,k(n-1)}$, gdzie $f_{1-\alpha;k-1,k(n-1)}$ oznacza kwantyl rzędu $1-\alpha$ rozkładu $F_{k-1,k(n-1)}$, to H_0 odrzucamy; natomiast jeśli $F < f_{1-\alpha;k-1,k(n-1)}$ to stwierdzamy, że nie ma podstaw do odrzucenia H_0. Decyzja ta zostaje podjęta na poziomie istotności α.

Implementacja powyższego testu w R:

```
> model=lm(zmienna.odpowiedzi ~ czynnik)
> summary(model)
```

lub

```
> model=aov(zmienna.odpowiedzi ~ czynnik)
> summary(model)
```

Sprawdzanie założeń

Przed przeprowadzeniem powyższego testu trzeba sprawdzić czy są spełnione założenia analizy wariancji.

1). Czy dla każdego poziomu czynnika rozkład zmiennej odpowiedzi jest normalny?

- Jeśli n nie jest bardzo małe, to rysujemy wykresy kwantylowe dla odpowiedzi w każdej grupie:

  ```
  > par(mfrow=c(k,1))        # k oznacza liczbę grup
  > tapply(zmienna.odpowiedzi,czynnik,qqnorm)
  ```

- Jeśli n nie jest bardzo małe, to przeprowadzamy test lub testy normalności, np.:

  ```
  > tapply(zmienna.odpowiedzi,czynnik,shapiro.test)
  ```

- Jeśli n jest małe i wariancje w grupach możemy uznać za równe, to lepiej sporządzić wykres kwantylowy i przeprowadzić test Shapiro-Wilka dla wszystkich reszt $Y_{ij} - \bar{Y}_{i\cdot}$.

2). Czy wariancje w grupach są przynajmniej w przybliżeniu równe?

Przeprowadzamy test (lub testy) równości wariancji w grupach: test Levene'a lub test Bartletta (ten drugi, gdy $n \geqslant 10$ oraz, gdy rozkłady zmiennej odpowiedzi w grupach są normalne), poziom istotności dla tych testów ustalamy na poziomie $\alpha = 0,01$, bo procedury związane z analizą wariacji są w miarę odporne na odstępstwa od założenia o równych wariancjach w grupach.

```
> bartlett.test(zmienna.odpowiedzi ~ czynnik)
> library(car)
> leveneTest(zmienna.odpowiedzi ~ czynnik, center=mean)
```

Uwaga: Założenia o normalności rozkładów i równości wariancji mogą nie być spełnione z powodu istnienia obserwacji odstających. Obserwacje odstające, które można wykryć analizując wykresy skrzynkowe, należy usunąć ze zbioru danych. Jeśli po usunięciu obserwacji odstających, założenia o normalności rozkładów i równości wariancji zdecydowanie trzeba odrzucić, to można próbować tak przetransformować dane (rozpatrując np. $1/Y_{ij}$ lub $\ln(Y_{ij})$ lub $\sqrt{Y_{ij}}$ zamiast wyjściowych Y_{ij}) aby po transformacji porządane założenia były spełnione. Inną alternatywą, gdy nie są spełnione założenia analizy wariancji, jest zastosowanie nieparametrycznego testu Kruskala-Wallisa - testu tego nie będziemy tu jednak omawiać.

Porównania wielokrotne

Jeśli odrzucimy $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, to powstaje pytanie, które średnie różnią się istotnie między sobą. Aby na nie odpowiedzieć stosujemy tzw. *porównania wielokrotne* - porównujemy średnie parami, dla każdej pary stosując pewną modyfikację testu t dla próbek niezależnych o nieznanych lecz równych wariancjach. Oznaczmy przez T_{ij} statystykę testową dla testu weryfikującego

$$H_0 : \mu_i = \mu_j \text{ przeciwko } H_1 : \mu_i \neq \mu_j.$$

Wówczas

$$T_{ij} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{S\sqrt{\frac{2}{n}}},$$

gdzie $S = \sqrt{S^2}$ i S^2 jest estymatorem nieznanej wariancji σ^2, tzn. $S^2 = \frac{SSE}{N-k}$ i N to liczba wszystkich obserwacji (w szczególności w przypadku planu zrównoważonego mamy $N = kn$). Ponadto, jeśli grupy nie są równoliczne, to we wzorze na T_{ij} zastępujemy $\frac{2}{n}$ przez $\frac{1}{n_i} + \frac{1}{n_j}$, gdzie n_i to liczność i-tej grupy. Zauważmy, że powyższy wzór na T_{ij} to modyfikacja statystyki testowej testu t dla próbek niezależnych o nieznanych lecz równych wariancjach otrzymana poprzez zastąpienie estymatora wariancji opartego na dwóch porównywanych grupach estymatorem wariancji wyliczonym z wykorzystaniem wszystkich k grup. Można pokazać, że gdy H_0 jest prawdziwa, to statystyka testowa T_{ij} ma rozkład t-Studenta o $N - k$ stopniach swobody:

$$T_{ij} \overset{H_0}{\sim} t_{[N-k]}.$$

Stąd jeśli $|T_{ij}| \geqslant t_{1-\tilde{\alpha};N-k}$, gdzie $t_{1-\tilde{\alpha};N-k}$ oznacza kwantyl rzędu $1 - \tilde{\alpha}$ rozkładu $t_{[N-k]}$, to H_0 odrzucamy. Pozostaje problem jak dobrać poziom istotności pojedynczego testu $\tilde{\alpha}$ tak by poziom istotności całej procedury wynosił α. Zauważmy bowiem, że, porównując średnie parami, przeprowadzamy $\binom{k}{2} = k(k-1)/2$ testów. Gdyby za poziom istotności pojedynczego testu przyjąć po prostu α, to poziom istotności całej procedury byłby większy niż α - jest to sytuacja analogiczna np. do ciągnięcia losów: jeśli

w pojedynczym losowaniu prawdopobieństwo trafienia na los wygrywający wynosi α, to gdy losowań wykonamy więcej, prawdopodobieństwo trafienia na los wygrywający wzrośnie.

W literaturze znane są różne metody doboru poziomu istotności pojedynczego testu $\tilde{\alpha}$.

- *Procedura Bonferroniego* - za poziom istotności pojedynczego testu przyjmujemy

  $$\tilde{\alpha} = \alpha/K^*, \text{gdzie } K^* \text{ to liczba par, które porównujemy.}$$

  Niestety dla tak dobranej $\tilde{\alpha}$ poziom całej procedury jest mniejszy niż zadana α. Ponadto dla dużych wartości K^* procedura Bonferroniego staje się bezużyteczna, bo praktycznie nigdy nie odrzuca hipotezy zerowej.

  > pairwise.t.test(zmienna.odpowiedzi,czynnik,
  p.adjust.method="bonferroni")

- *Procedura Tukeya* - zbiór krytyczny dla pojedynczego testu oparty jest nie na rozkładzie t-Studenta lecz na tzw. *studentyzowanym rozkładzie rozstępu dla próby z rozkładu normalnego* czyli rozkładzie maksymalnej różnicy pomiędzy średnimi. Procedura ta jest szczególnie polecana w sytuacji porównywania średnich w grupach o tej samej liczności.

  > TukeyHSD(aov(model))

- *Procedura Scheffégo* - pozwala porównywać nie tylko średnie w parach, ale także tak zwane *kontrasty* między więcej niż dwiema średnimi. W przykładzie 14.1 pozwoli np. sprawdzić czy średnia jaskrawość papieru pochodzącego od pierwszego i drugiegi producenta jest taka sama jak średnia jaskrawość papieru pochodzącego od trzeciego producenta:

  $$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3 \text{ przeciwko } H_1 : \frac{\mu_1 + \mu_2}{2} \neq \mu_3.$$

  Ogólnie możemy testować

  $$H_0 : \sum_{i=1}^{k} c_i\mu_i = 0 \text{ przeciwko } H_1 : \sum_{i=1}^{k} c_i\mu_i \neq 0, \text{ gdzie } \sum_{i=1}^{k} c_i = 0.$$