

Jednoczynnikowa analiza ANOVA w R Studio

Jacek Matulewski, wersja z 31 maja 2025

Kurs R: <https://jacekmatulewski.fizyka.umk.pl/dydaktyka/r/>

Cel analizy

Jednoczynnikowa analiza wariancji (ANOVA) służy do sprawdzenia, czy średnie w więcej niż dwóch grupach różnią się istotnie statystycznie. Wymaga spełnienia kilku założeń, m.in. normalności i homogeniczności wariancji.

0. Aktualizacja R i R Studio

R.Version()

Wydruk:

```
$platform
[1] "x86_64-w64-mingw32"

$arch
[1] "x86_64"

$os
[1] "mingw32"

$crt
[1] "ucrt"

$system
[1] "x86_64, mingw32"

$status
[1] ""

$major
[1] "4"

$minor
[1] "5.0"

$year
[1] "2025"

$month
[1] "04"

$day
[1] "11"

$`svn rev`
[1] "88135"

$language
[1] "R"

$version.string
[1] "R version 4.5.0 (2025-04-11 ucrt)"

$nickname
[1] "How About a Twenty-Six"
```

1. Dane (iris)

Przegląd danych do analizy:

```
summary(iris)
levels(iris$Species)
table(iris$Species)
```

Wydruk:

```
setosa versicolor virginica
    50         50         50
```

2. Sprawdzenie założeń

a) Normalność rozkładu

Dla każdej grupy (gatunku irysa) sprawdzamy, czy rozkład zmiennej zależnej jest normalny. Zaczynamy od działek kielicha:

```
iris$Sepal.Length
iris$Sepal.Length[iris$Species=="setosa"]
shapiro.test(iris$Sepal.Length[iris$Species=="setosa"])      #p=0.4595
shapiro.test(iris$Sepal.Width[iris$Species=="setosa"])      #p=0.2715
shapiro.test(iris$Sepal.Length[iris$Species=="versicolor"]) #p=0.4647
shapiro.test(iris$Sepal.Width[iris$Species=="versicolor"]) #p=0.338
shapiro.test(iris$Sepal.Length[iris$Species=="virginica"])  #p=0.2583
shapiro.test(iris$Sepal.Width[iris$Species=="virginica"])   #p=0.1809
```

Wszystkie rozkłady są normalne.

Można mniejszą liczbą instrukcji korzystając z *dplyr*:

```
#wersja skrócona
library(dplyr)
iris %>%
  select(Sepal.Length, Sepal.Width, Species) %>%
  group_by(Species) %>%
  summarise(across(.fns = ~shapiro.test(.)$p.value)) -> result
result #tabela wartości p
```

b) Homogeniczność wariancji (test Levene'a)

Skupmy się na szerokościach działek kwiatowych.

```
lm.iris.Sepal.Width <- lm(Sepal.Width~Species, data=iris) #model
summary(lm.iris.Sepal.Width)

# https://cran.r-project.org/web/packages/car/index.html
install.packages("car", dependencies = TRUE)
library(car)
leveneTest(lm.iris.Sepal.Width) #p=0.5555, wariancje są homogeniczne
```

Wydruk:

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.5902 0.5555
147
```

Podobnie, jak w przypadku testu normalności rozkładu Shapiro-Wilka, test Levene'a potwierdza, że rozkłady w podgrupach są homogeniczne, gdy $p > 0.05$.

Wobec tego dla $p = 0.5555$ rozkłady szerokości działek kwiatowych są homogeniczne (mają podobne wariancje).

3. Jednoczynnikowa ANOVA

Jeśli dane spełniają założenia:

```
anova(lm.iris.Sepal.Width)
```

Wydruk:

Analysis of Variance Table

Response: Sepal.Width

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	11.345	5.6725	49.16	< 2.2e-16 ***
Residuals	147	16.962	0.1154		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Istotność statystyczna $p < 2.2 \cdot 10^{-16}$, więc gatunek istotnie różnicuje szerokość działek kielicha irysów w analizowanym zbiorze danych.

Alternatywnie funkcja, która jest zgodna z testami post-hoc:

```
aov.iris.Sepal.Width <- aov(Sepal.Width~Species, data=iris)
summary(aov.iris.Sepal.Width)
```

Wydruk (wynik jest identyczny):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	11.35	5.672	49.16	<2e-16 ***
Residuals	147	16.96	0.115		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4. Analizy post-hoc

Jeśli wynik analizy ANOVA jest istotny statystycznie, to przy normalnych rozkładach w podgrupach i homogenicznych wariancjach, możemy do porównania grup użyć testu Tukeya:

```
TukeyHSD(aov.iris.Sepal.Width)
```

Wydruk:

Tukey multiple comparisons of means
95% family-wise confidence level

```
Fit: aov(formula = Sepal.Width ~ Species, data = iris)
```

\$Species

	diff	lwr	upr	p adj
versicolor-setosa	-0.658	-0.81885528	-0.4971447	0.0000000
virginica-setosa	-0.454	-0.61485528	-0.2931447	0.0000000
virginica-versicolor	0.204	0.04314472	0.3648553	0.0087802

Różnice między wszystkimi grupami są istotne statystycznie.

5. Wariant ANOVA dla rozkładu niehomogenicznego

Zajmijmy się długościami działek kwiatowych. Wiemy już, że rozkład jest normalny. Sprawdzamy homogeniczność.

```
lm.iris.Sepal.Length <- lm(Sepal.Length~Species, data=iris) #model
summary(lm.iris.Sepal.Length)
leveneTest(lm.iris.Sepal.Length) #p=0.002259, wariancje nie są homogeniczne
```

Levene's Test for Homogeneity of Variance (center = median)

```
      Df F value    Pr(>F)
group   2  6.3527 0.002259 **
      147
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Istotność jest mniejsza niż 0.05, więc rozkład długości płatków dla różnych gatunków nie jest homogeniczny (wariancja jest różna). Należy użyć testu Welcha.

```
install.packages("rstatix")
library(rstatix)
aov.Sepal.Length <- welch_anova_test(Sepal.Length~Species, data = iris)
aov.Sepal.Length$p
```

Wynik:

A tibble: 1 × 7

	.y. <chr>	n <int>	statistic <dbl>	DFn <dbl>	DFd <dbl>	p <dbl>	method <chr>
1	Sepal.Length	150	138.91	2	92.21115	1.51e-28	Welch ANOVA

1 row

Różnice są istotne statystycznie ($p = 1.51 \cdot 10^{-28}$), więc można przeprowadzić testy post-hoc:

```
games_howell_test(Sepal.Length~Species, data = iris)
```

Wynik:

A tibble: 3 × 8

	.y. <chr>	group1 <chr>	group2 <chr>	estimate <dbl>	conf.low <dbl>	conf.high <dbl>	p.adj <dbl>
1	Sepal.Length	setosa	versicolor	0.930	0.7192050	1.1407950	2.86e-10
2	Sepal.Length	setosa	virginica	1.582	1.3362452	1.8277548	0.00e+00
3	Sepal.Length	versicolor	virginica	0.652	0.3761745	0.9278255	5.58e-07

3 rows | 1-8 of 8 columns

Długości płatków kielicha różnią się istotnie pomiędzy wszystkimi parami gatunków.

6. Nieparametryczna alternatywa – test Kruskala-Wallisa

Jeśli dane nie spełniają warunku normalności rozkładu we wszystkich grupach, konieczne jest użycie nieparametrycznej wersji ANOVA, czyli testu Kruskala-Wallisa.

Weryfikacja normalności dla szerokości płatków:

```
shapiro.test(iris$Petal.Width[iris$Species=="setosa"])$p
shapiro.test(iris$Petal.Width[iris$Species=="versicolor"])$p
shapiro.test(iris$Petal.Width[iris$Species=="virginica"])$p
```

Wydruk:

```
> shapiro.test(iris$Petal.Width[iris$Species=="setosa"])$p
[1] 8.658573e-07
> shapiro.test(iris$Petal.Width[iris$Species=="versicolor"])$p
[1] 0.0272778
> shapiro.test(iris$Petal.Width[iris$Species=="virginica"])$p
[1] 0.08695419
```

Tylko rozkład dla gatunku *virginica* jest normalny.

Wobec tego konieczny jest test nieparametryczny ANOVA:

```
kruskal.test(Petal.Width~Species, data=iris)
```

Wydruk:

```
kruskal-wallis rank sum test

data: Petal.Width by species
kruskal-wallis chi-squared = 131.19, df = 2, p-value < 2.2e-16
```

Wniosek: gatunek różnicuje długości płatków ($p < 2.2 \cdot 10^{-16}$)

Po więcej informacji zob.

<https://statsandr.com/blog/kruskal-wallis-test-nonparametric-version-anova/>

Testy post-hoc dla testu Kruskala-Wallisa

```
install.packages("FSA")
library(FSA)
dunnTest(Petal.Width~Species, data=iris)
```

Wynik:

Comparison <chr>	Z <dbl>	P.unadj <dbl>	P.adj <dbl>
setosa - versicolor	-5.895978	3.724694e-09	7.449388e-09
setosa - virginica	-11.451936	2.299512e-30	6.898535e-30
versicolor - virginica	-5.555959	2.760924e-08	2.760924e-08

3 rows

7. Wizualizacja wyników

```
require(ggplot2)
require(ggsignif)

ggplot(iris, aes(x = Species, y = Petal.Length)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal()
```

Wynik:

