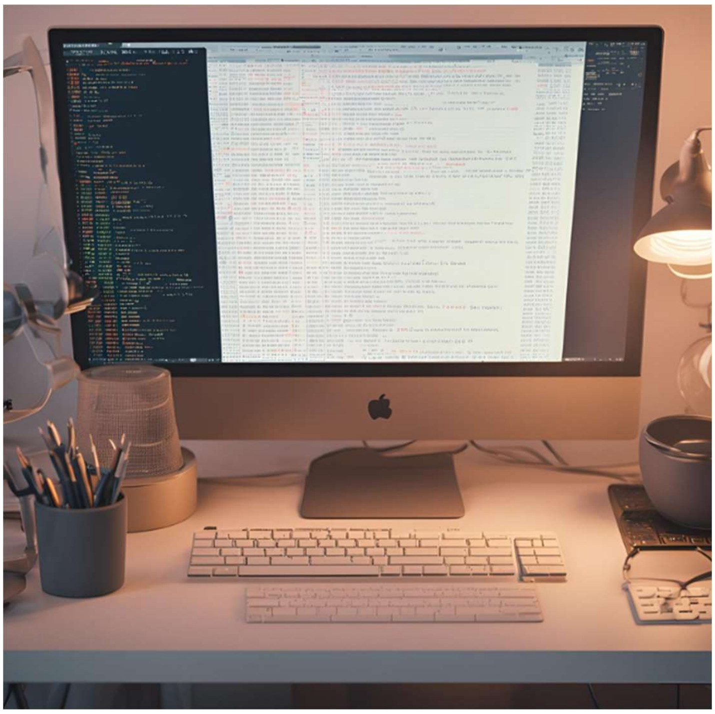
Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Ćwiczenie nr 8
Testy statystyczne w Excelu oraz języku R – część 3



POLITECHNIKA BYDGOSKA im. Jana i Jędrzeja Śniadeckich

POLITECHNIKA BYDGOSKA Wydział Technologii i Inżynierii Chemicznej

POLITECHNIKA BYDGOSKA Wydział Medyczny

PRACOWNIA KOMPUTEROWA

Wstęp

Statystyka odgrywa kluczową rolę w analizie danych medycznych, umożliwiając wnioskowanie na podstawie wyników eksperymentalnych. W zależności od charakterystyki danych i postawionych hipotez, stosuje się różne rodzaje testów statystycznych. Kluczowym aspektem wyboru odpowiedniego testu jest zrozumienie struktury danych, rozkładu zmiennych oraz celu analizy.

W tych ćwiczeniach omówimy:

1. Testy parametryczne – stosowane, gdy dane spełniają określone założenia, takie jak normalność rozkładu i jednorodność wariancji (na przykładzie analiza wariancji ANOVA).
2. Testy nieparametryczne – używane, gdy założenia testów parametrycznych nie są spełnione (np. test U Manna-Whitneya, test Kruskala-Wallisa).
3. Sprawdzanie założeń statystycznych – przed wyborem testu parametrycznego należy zweryfikować:
   - Normalność rozkładu za pomocą testów takich jak test Shapiro-Wilka lub test Kołmogorowa-Smirnowa. Kiedy założenie normalności nie jest spełnione i wyniki odbiegają znacząco od rozkładu normalnego, można posłużyć się różnego rodzaju transformacjami danych, np. logarytmowaniem, potęgowaniem.
   - Jednorodność wariancji (homogeniczność) przy użyciu testów np. Levene'a
   - Równoliczność obserwacji w grupach. Kategorie zmiennej niezależnej powinny być statystycznie równoliczne (nie chodzi o idealną równoliczność poszczególnych grup, niewielkie różnice między grupami mogą być). Aby sprawdzić, czy analizowane grupy różnią się istotnie statycznie pod względem liczebności, można zastosować test zgodności Chi-kwadrat.
4. Testy post-hoc – stosowane po wykryciu istotnych różnic w testach wieloczynnikowych (np. ANOVA) w celu identyfikacji, między którymi grupami występują różnice (np. test Tukeya).
5. Obliczenie siły efektu. W przypadku jednoczynnikowej analizy wariancji miarą siły efektu jest współczynnik eta-kwadrat (η^2).

   $$\eta^2 = \frac{\text{Sum Sq}_{\text{between}}}{\text{Sum Sq}_{\text{total}}}$$

   Pozwala on określić siłę związku między zmienną niezależną a zmienną zależną lub ujmując to w sposób bardziej statystyczny, jaka proporcja całkowitej wariancji zmiennej zależnej wyjaśniana jest przez badany efekt.

Analiza wariancji (ANOVA) to statystyczna metoda służąca do porównywania średnich w więcej niż dwóch grupach w celu ustalenia, czy różnice między nimi są istotne statystycznie. Umożliwia analizę wpływu jednej lub więcej zmiennych niezależnych (czynników) na zmienną zależną, a jej zastosowanie zależy od struktury badania.

Podstawowe rodzaje testów ANOVA

1. Jednoczynnikowa ANOVA (One-Way ANOVA) - używamy, gdy porównujemy średnie w więcej niż dwóch grupach na podstawie jednego czynnika.

   Przykład: Analiza wpływu diety (np. dieta A, B, C) na masę ciała.

Założenia: Dane w każdej grupie pochodzą z populacji o rozkładzie normalnym. Grupy mają jednorodne wariancje. Obserwacje są niezależne.
Wynik testu informuje, czy istnieje statystycznie istotna różnica między średnimi w grupach.

2. Dwuczynnikowa ANOVA (Two-Way ANOVA) - stosujemy, gdy chcemy zbadać wpływ dwóch czynników na zmienną zależną, a także interakcję między tymi czynnikami.

   Przykład: Analiza wpływu diety (dieta A, B, C) i aktywności fizycznej (niska, średnia, wysoka) na masę ciała.
   Założenia są takie same jak w jednoczynnikowej ANOVA, z dodatkowym aspektem interakcji między czynnikami.
   Wynik testu pozwala określić czy każdy z czynników wpływa istotnie na zmienną zależną oraz czy istnieje interakcja między czynnikami, czyli czy wpływ jednego czynnika zależy od poziomu drugiego.

Przedstawiając wyniki analizy wariancji powinno się dołączyć krótki opis testowanych zmiennych, wartość testu F, stopnie swobody oraz wartości poziomu istotności dla każdej zmiennej niezależnej.

Cel

Celem bieżącego ćwiczenia jest opanowanie zastosowania testu ANOVA w języku R oraz Excelu, oraz zrozumienie, w jakich sytuacjach i dlaczego stosujemy konkretne testy. Dzięki temu możliwe będzie poprawne planowanie analiz oraz interpretacja wyników badań.

Studenci nauczą się, jak:

- Wykonywać testy ANOVA jedno- i dwu-czynnikowe wraz z testami post-hoc aby sprawdzić zależności i różnice w zestawach danych.
- Interpretować wyniki testów statystycznych.

Przebieg ćwiczenia

1. Testowanie statystyczne danych za pomocą testu jednoczynnikowego ANOVA

Treść zadania ze strony Statystyka od A do Z (https://www.statystyka.eu):
Wiadomo, że związki chemiczne stosowane w leczeniu nowotworów mogą powodować obniżenie poziomu hemoglobiny we krwi (niedokrwistość). W przypadku pewnego związku chemicznego stosowanego w leczeniu nowotworów (Lek A) podejrzewano, że przy długotrwałym stosowaniu powoduje niedokrwistość (stężenie hemoglobiny we krwi poniżej 11g/dl) w większym stopniu niż inne leki tego typu.
Do badania włączono grupę 24 osób z rozpoznaniem nowotworu. 10 z nich podawano wspomniany lek A. Pozostałym pacjentom podawano inne leki o podobnym działaniu. 7 pacjentów zażywało lek B, a 7 lek C.
W momencie przystąpienie do badania u wszystkich pacjentów poziom hemoglobiny we krwi był prawidłowy. Po zakończonej obserwacji u pacjentów ponownie wykonano morfologię krwi. Wyniki badania poziomu hemoglobiny u badanych były następujące:

Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Lek A	Lek B	Lek C
10,2	14,3	10,4
8,7	14,1	12
12,5	17	13,6
13,8	13,2	13,5
7,6	11,6	14,7
8,2	10,9	15,3
9,8	9,3	14,9
10,9		
11,6		
14,2		

a) Przeprowadzanie analizy w języku R

- Wejdź na stronę z kompilatorem R online, np. https://rdrr.io/snippets/
- Sprawdźmy testem chi-kwadrat czy grupy są statystycznie równoliczne bowiem z tabelki widać, że mamy 10 osób którym podawano Lek A i tylko 7 gdzie podawano Lek B i Lek C:

```
# Liczebności obserwowane
obserwacje <- c(10, 7, 7)
# Oczekiwane liczebności (np. równy podział czyli powinno być pacjentów 8,8,8)
oczekiwane <- rep(sum(obserwacje) / 3, 3)
# Test Chi-kwadrat
test <- chisq.test(x = obserwacje, p = oczekiwane / sum(oczekiwane))
print(test)
```

Odpowiedz na pytanie: czy są statystycznie równoliczne?

- Sprawdźmy testem Shapiro-Wilka czy grupy mają rozkład normalny

```
LekA = c(10.2, 8.7, 12.5, 13.8, 7.6, 8.2, 9.8, 10.9, 11.6, 14.2)
LekB = c(14.3, 14.1, 17, 13.2, 11.6, 10.9, 9.3)
LekC = c(10.4, 12, 13.6, 13.5, 14.7, 15.3, 14.9)
shapiro.test(LekA)
shapiro.test(LekB)
shapiro.test(LekC)
```

Odpowiedz na pytanie: czy wszystkie grupy są normalne?

- Sprawdźmy testem Levene'a czy grupy mają jednorodną wariancję

```
library(car)
grupy <- factor(rep(c("LekA", "LekB", "LekC"),
                times = c(length(LekA), length(LekB), length(LekC))))
wyniki <- c(LekA, LekB, LekC)
dane <- data.frame(grupy, wyniki)
# Test Levene'a
test_levene <- leveneTest(wyniki ~ grupy, data = dane)
print(test_levene)
```

Odpowiedz na pytanie: czy wariancje są jednorodne?

- Jeśli na wszystkie poprzednie pytania odpowiedzi były twierdzące to wykonujemy test ANOVA jednoczynnikowy:

```
# Jednoczynnikowy ANOVA
anova_model <- aov(wyniki ~ grupy, data = dane)
# Wyniki ANOVA
```

```
anova_summary <- summary(anova_model)
print(anova_summary)
```

Odpowiedz na pytanie: czy są różnice między grupami?

- Jeśli są różnice to sprawdź między którymi grupami testem post-hoc:

```
# Test post-hoc Tukeya
tukey_results <- TukeyHSD(anova_model)
print(tukey_results)
```

Między którymi grupami jest statystycznie istotna różnica?

- Aby sprawdzić siłę efektu korzystamy z wyników ANOVA i wyciągamy wartości SS:

```
# Wyciąganie sum kwadratów
ss_between <- anova_summary[[1]]$`Sum Sq`[1]  # Suma kwadratów między grupami
ss_total <- sum(anova_summary[[1]]$`Sum Sq`)  # Całkowita suma kwadratów
# Obliczenie eta-kwadrat
eta_squared <- ss_between / ss_total
print(eta_squared)
```

Wartość η^2 wskazuje, jaka część całkowitej zmienności wyników jest wyjaśniana przez czynnik grupujący (w tym przypadku grupy leków). Czy siła efektu jest duża?

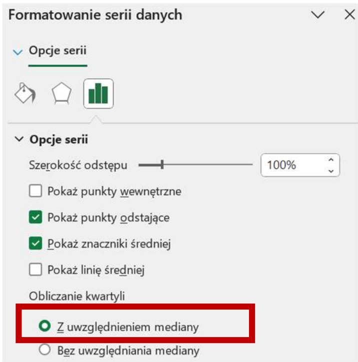
Ogólne wytyczne dla η^2:

- $\eta^2 = 0.01$: **Mała siła efektu** (1% wariancji wyjaśnionej).
- $\eta^2 = 0.06$: **Średnia siła efektu** (6% wariancji wyjaśnionej).
- $\eta^2 = 0.14$: **Duża siła efektu** (14% wariancji wyjaśnionej).

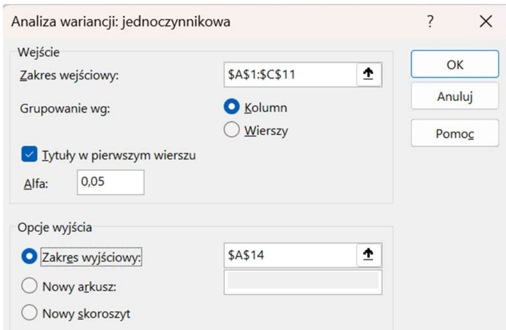
b) Przeprowadzanie analizy w Excelu

Przepisz tabelkę z treści zadania do Excela i wykonaj następujące kroki:

- Zrób wykres Skrzynka i wąsy dla trzech różnych leków. Zaznacz 3 kolumny wraz z nagłówkami i wybierz Wstaw->Polecane wykresy->Wszystkie wykresy->Skrzynka i wąsy. Dodaj etykiety do wykresu. Zaznacz po kolei każde pudełko i sformatuj aby kwartyle liczone były z uwzględnieniem mediany. Czy na podstawie wykresów możesz wywnioskować czy rozkład jest podobny do normalnego?
  Wykonaj dodatkowo statystykę opisową dla tych 3 grup danych aby upewnić się co do skośności i kurtozy (czy nie są za duże – jak w poprzednich zadaniach).



- Wykonaj analizę wariancji ANOVA (Dane->Analiza danych-> Analiza wariancji: jednoczynnikowa)



Pamiętaj o zaznaczeniu „Tytuły w pierwszym wierszu” jeśli „Zakres wejściowy” zaznaczony był wraz z nagłówkami kolumn.

- Policz siłę efektu na podstawie wyliczeń ANOVA:

ANALIZA WARIANCJI						
ódło warian	*SS*	*df*	*MS*	*F*	*Wartość-p*	*Test F*
Pomiędzy	36,15744	2	18,07872	3,639936	0,043933	3,4668
W obrębie	104,3021	21	4,966769			
Razem	140,4596	23				

Siła efektu η^2 to SS pomiędzy podzielone przez SS Razem (odpowiednie pola pokazano powyżej).

Uwaga: w Excelu można by zrobić test post-hoc ale licząc ręcznie np. używając metody NIR (Najmniej Istotnej Różnicy, ang. Least Significant Difference, LSD).

2. Testowanie statystyczne danych za pomocą testu dwuczynnikowego ANOVA

Załóżmy, że mamy dane o skuteczności leczenia dla różnych grup pacjentów, którzy zostali podzieleni według wieku i rodzaju terapii. Dane są zgromadzone w następującej tabeli:

Wiek	Terapia A	Terapia B
Młodzi	85	82
Młodzi	87	81
Młodzi	90	83
Młodzi	88	79
Młodzi	86	80
Młodzi	89	84
Starsi	78	74
Starsi	79	73
Starsi	77	75
Starsi	76	72
Starsi	75	74
Starsi	78	76

Sprawdź czy wiek i rodzaj terapii wpływają na skuteczność leczenia.

a) Przeprowadzanie analizy w języku R

Utwórz ramkę danych na podstawie powyższej tabelki. Ramka w formacie długim potrzebnym do testów statystycznych wygląda następująco (pamiętaj nie używać polskich literek w nazwach zmiennych):

```
# Dane wejściowe
dane <- data.frame(
  Wiek = factor(c(rep("Mlodzi", 12), rep("Starsi", 12))),
```

```
  Terapia = factor(c(rep("A", 6), rep("B", 6), rep("A", 6), rep("B", 6))),
  Skutecznosc = c(85, 87, 90, 88, 86, 89, 82, 81, 83, 79, 80, 84, 78, 79, 77, 76, 75, 78,
  74, 73, 75, 72, 74, 76)
)
print(dane)
```

- Sprawdź czy spełnione są założenia potrzebne do analizy wariancji ANOVA
  - sprawdzenie normalności rozkładu

Aby przeprowadzić test Shapiro-Wilka potrzebna jest ramka danych w formacie szerokim. Zamieńmy więc tą ramkę na szeroką. Indeks potrzebny jest aby funkcja pivot_wider nie pisała ostrzeżeń, że dane nie są unikalne:

```
library(tidyr)
# Dodanie indeksu
dane$Indeks <- ave(dane$Skutecznosc, dane$Wiek, dane$Terapia, FUN = seq_along)
# Przekształcenie na format szeroki
dane_wide <- dane %>%
  pivot_wider(names_from = c(Wiek, Terapia), values_from = Skutecznosc)
# Wyświetlenie wyniku
print(dane_wide)
# Sprawdzamy normalność rozkładu
shapiro.test(dane_wide$Mlodzi_A)
shapiro.test(dane_wide$Mlodzi_B)
shapiro.test(dane_wide$Starsi_A)
shapiro.test(dane_wide$Starsi_B)
```

  - sprawdzenie jednorodności wariancji

```
library(car)
# Test Levene'a dla zmiennej Skutecznosc, z grupowaniem według Wiek i Terapia
levene_result <- leveneTest(Skutecznosc ~ Wiek * Terapia, data = dane)
print(levene_result)
```

Jeśli oba powyższe testy potwierdziły wszystkie hipotezy 0 (czyli p-value > 0.05 dla poziomu ufności 0.05) to można wykonać analizę wariancji ANOVA dwuczynnikową. Uwaga: nie trzeba sprawdzać czy grupy są równo liczne bo widać to z tabelki, że tak.

- Wykonujemy analizę dwuczynnikową ANOVA

```
# Dwuczynnikowa analiza ANOVA
anova_wynik <- aov(Skutecznosc ~ Wiek * Terapia, data = dane)
# Wyświetlenie wyników
summary(anova_wynik)
```

Interpretacja

Tam gdzie p-value < 0.05 to istnieje **istotna różnica** w skuteczności leczenia między grupami wiekowymi. Testowane były 3 rzeczy:

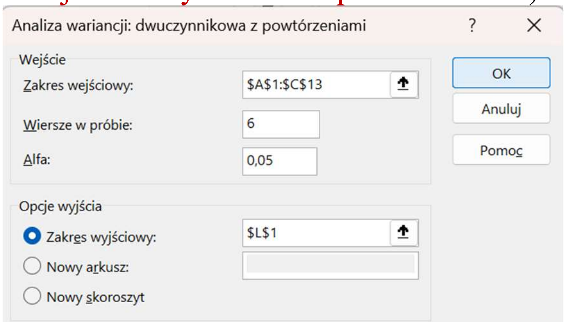
- różnica w skuteczności między grupami wiekowymi (Mlodzi vs. Starsi),
- różnica w skuteczności między terapią A a terapią B,
- interakcja między wiekiem a terapią, czy wpływa na skuteczność.

- Jeśli któryś z tych trzech testów wskazał różnice to można sprawdzić między czym a czym są te różnice wykonując test post-hoc:

```
# Test post-hoc Tukeya
tukey_wynik <- TukeyHSD(anova_wynik)
print(tukey_wynik)
```

b) Przeprowadzanie analizy w Excelu

Przepisz do Excela tabelkę z zadania i wykonaj analizę dwuczynnikową ANOVA z powtórzeniami (ponieważ grupa Mlodzi i Starsi powtarza się w danych). (Dane->Analiza danych-> Analiza wariancji: dwuczynnikowa z powtórzeniami)



Porównaj wyniki otrzymane z języka R i Excela.

3. Testowanie statystyczne danych za pomocą testu Kruskala-Wallisa

Wykonajmy porównanie skuteczności trzech różnych metod leczenia bólu pooperacyjnego w grupach pacjentów. Poziom bólu został oceniony w skali od 0 do 10, gdzie 0 oznacza brak bólu, a 10 oznacza najsilniejszy możliwy ból. W leczeniu bólu zastosowano 3 metody:

- **Leczenie A** (metoda farmakologiczna)
- **Leczenie B** (metoda fizjoterapeutyczna)
- **Leczenie C** (metoda chirurgiczna)

Dane pacjentów znajdują się w pliku pacjenci_bol_cw8.xlsx (link poda prowadzący).

Za pomocą konwertera online (np. https://tableconvert.com/excel-to-rdataframe) wczytaj dane z pliku Excela do języka R (kopiując tabelkę Excela w odpowiednie miejsce konwertera):

```
dane <- data.frame(
  Leczenie = c("C", "A", "C", "C", "A", "A", "C", "B", "C", "C", "C", "C", "A", "C", "B", "A", "B", "B", "B", "B", "A", "A", "B", "B", "A", "A", "A", "C", "C", "C", "B", "C", "B", "B", "C", "B", "C", "C", "A", "C", "A", "C", "C", "A", "A", "C", "B", "A", "B", "B"),
  Poziom_bolu = c("6,8", "7", "5,9", "5,8", "8,4", "9,4", "6,4", "7,6", "5,2", "6,4", "6,3", "9,3", "7,5", "4,9", "3,7", "5,4", "7,3", "6,3", "6,7", "4,2", "6,6", "6", "6,5", "6,7", "8,3", "7,1", "6,1", "6,4", "5", "5", "4,1", "4,6", "6,6", "5,1", "4,6", "5,6", "4,3", "4,7", "2,9", "3,9", "5,8", "2,7", "5,1", "6,1", "4,5", "2", "4,4", "3,9", "4,6", "4,9")
)
```

Zamieńmy przecinki na kropki i dane tekstowe na liczbowe:

```
dane$Poziom_bolu <- as.numeric(gsub(",", ".", dane$Poziom_bolu))
print(dane)
# Konwertujemy kolumnę 'Leczenie' na zmienną czynnikową
```

```
dane$Leczenie <- factor(dane$Leczenie, levels = c("A", "B", "C"))
# Przeprowadzamy test Kruskala-Wallisa
kruskal_test <- kruskal.test(Poziom_bolu ~ Leczenie, data = dane)
print(kruskal_test)
```

Odpowiedz na pytanie: czy są istotne statystyczne różnice między metodami leczenia?
Jeśli test wykaże istotne różnice to sprawdzimy między czym a czym są te różnice testem post-hoc Dunn'a:

```
# Przeprowadzenie testu Dunn's
library(FSA)
dunn_test <- dunnTest(Poziom_bolu ~ Leczenie, data = dane, method = "bh")
print(dunn_test)
```

Czy w tym przykładzie zastosowaliśmy dobry test? Sprawdź to licząc normalność rozkładu oraz jednorodność wariancji:

```
shapiro.test(dane$Poziom_bolu[dane$Leczenie == "A"])  # Test dla grupy A
shapiro.test(dane$Poziom_bolu[dane$Leczenie == "B"])  # Test dla grupy B
shapiro.test(dane$Poziom_bolu[dane$Leczenie == "C"])  # Test dla grupy C
library(car)
leveneTest(Poziom_bolu ~ Leczenie, data = dane)
```

Jeśli wykorzystaliśmy zły test to wykonaj ponownie zadanie z poprawnie dobranym testem (kieruj się poprzednimi przykładami – starczy zmienić 2 linijki).