Technologia informacyjna i biostatystyka w medycynie

Redakcja naukowa
Krystyna Laszki-Szcząchor
Krzysztof Kujawa

Uniwersytet Medyczny
im. Piastów Śląskich we Wrocławiu

Technologia informacyjna i biostatystyka w medycynie

Technologia informacyjna i biostatystyka w medycynie

Redakcja naukowa
Krystyna Laszki-Szcząchor
Krzysztof Kujawa

Uniwersytet Medyczny
im. Piastów Śląskich we Wrocławiu

Autorzy

Tomasz Hikawczuk, Krystyna Laszki-Szcząchor, Agnieszka Rusiecka, Andrii Zinchuk
Uniwersytet Medyczny we Wrocławiu, Centrum Analiz Statystycznych, Zespół ds. Dydaktyki Biostatystyki

Krzysztof Kujawa
Uniwersytet Medyczny we Wrocławiu, Centrum Analiz Statystycznych

Opracowanie redakcyjne
Marcin Żuk

Korekta
Aleksandra Raczkowska

Projekt typograficzny i skład
Leonard Szłapka

Projekt okładki
Monika Kolęda

Zdjęcie na okładce wygenerowano za pomocą usługi Adobe Firefly

e-ISBN 978-83-7055-698-3

© Copyright by Uniwersytet Medyczny
im. Piastów Śląskich we Wrocławiu
Wrocław 2025

Spis treści

Wstęp 6

1. Redagowanie i edycja tekstów medycznych w programie Word 7
Krystyna Laszki-Szcząchor

2. Wstawianie odsyłaczy w tekście przy użyciu programu Word 14
Krystyna Laszki-Szcząchor

3. Przygotowanie wykresów na podstawie danych 19
Andrii Zinchuk

4. Podstawowe formuły porządkujące dane w programie Excel 38
Andrii Zinchuk

5. Sortowanie i filtrowanie bazy danych w Excelu. 54
Krystyna Laszki-Szcząchor

6. Przygotowanie prezentacji w programie PowerPoint 70
Andrii Zinchuk

7. Tworzenie baz danych pacjentów w programie Access 103
Krystyna Laszki-Szcząchor

8. Wprowadzenie do programu Statistica 139
Tomasz Hikawczuk

9. Statystyka opisowa i sprawdzanie normalności rozkładu danych 146
Agnieszka Rusiecka, Tomasz Hikawczuk

10. Testowanie hipotez statystycznych. Podstawowe pojęcia. 159
Agnieszka Rusiecka

11. Tabele wielodzielcze i analiza ryzyka w medycynie 170
Agnieszka Rusiecka

12. Testy *t* i ich nieparametryczne odpowiedniki w medycynie 181
Agnieszka Rusiecka, Krzysztof Kujawa

13. Jednoczynnikowa analiza wariancji w medycynie 205
Tomasz Hikawczuk

14. Dwuczynnikowa analiza wariancji 216
Tomasz Hikawczuk

15. Korelacja Pearsona i regresja prostoliniowa 225
Tomasz Hikawczuk

16. Zastosowanie regresji logistycznej w medycynie 234
Krzysztof Kujawa, Tomasz Hikawczuk

17. Krzywa ROC w badaniach medycznych 241
Agnieszka Rusiecka, Andrii Zinchuk, Krzysztof Kujawa

Wstęp

Niniejszy skrypt obejmuje zagadnienia objęte programem nauczania przedmiotu Podstawy technologii informacyjnych i biostatystyki. Realizacja tematów z tego przedmiotu rozbita jest na dwie części: pierwsza obejmuje technologie informacyjne, natomiast druga wprowadza w tajniki statystyki medycznej. W taki też sposób podzielone są treści zawarte w tym opracowaniu. Każdy z rozdziałów skryptu zawiera merytoryczne omówienie tematu, a następnie część praktyczną z zestawem danych do wykonania samodzielnych obliczeń.

Technologie informacyjne bardzo dynamicznie się rozwijają. Związane są z wytwarzaniem danych, ich gromadzeniem, analizowaniem, a także przetwarzaniem w użyteczne informacje. Wciąż zwiększające się tempo przyrostu ilości danych, szczególnie w dziedzinach związanych z medycyną i ochroną zdrowia, wymaga od nas opanowania przynajmniej podstawowych narzędzi umożliwiających ich analizę. Z tego względu pierwsza część zajęć (rozdziały 1–6 skryptu) skupia się na tematyce wykorzystania edytorów tekstu oraz innych narzędzi do pracy z tekstem. Następnie poruszane są kwestie wykorzystania arkuszy kalkulacyjnych do gromadzenia danych i wykonywania niezbędnych obliczeń oraz analiz. Omawiana jest również tematyka tworzenia wizualizacji danych, a także ich przedstawiania przy użyciu prezentacji multimedialnych. Ta część zajęć realizowana jest przede wszystkim z wykorzystaniem programów pakietu MS Office (MS Word, MS Excel, MS Access, MS PowerPoint).

Druga część skryptu (rozdziały 7–16) poświęcona jest zagadnieniom związanym z biostatystyką. Według ogólnej definicji biostatystyka jest nauką łączącą biologię i statystykę, ale zastosowanie znajduje zarówno na płaszczyźnie klinicznej, jak i stricte naukowej. Przedmiotem jej zainteresowania są dane oraz proces ich analizy, dający w efekcie maksymalną ilość informacji pozwalających na zrozumienie istoty badanego zjawiska. Omówiono tutaj kwestie związane z zasadami planowania obserwacji i doświadczeń pozwalającymi na zbieranie rzetelnych danych. Wskazano również reguły poprawnego wnioskowania na podstawie uzyskanych wyników. Opisano podstawowe zagadnienia związane z typami zmiennych oraz badaniem rozkładów danych. Wiele miejsca poświęcono tematyce testowania hipotez statystycznych i doboru właściwych metod statystycznych dla zbioru, a także interpretacji otrzymanych wyników. Cześć praktyczna obejmuje wykonywanie niezbędnych obliczeń w programie Statistica.

Jak ważne jest opanowanie podstawowych metod technologii informacyjnych i biostatystyki, jak nie zgubić się w coraz większej ilości danych, co więcej, jak uzyskać z nich użyteczne informacje oraz niezbędną wiedzę, tego wszystkiego dowiemy się w trakcie naszych spotkań. Niech to opracowanie będzie merytorycznym wsparciem w procesie nauki.

Do zobaczenia na zajęciach.

Autorzy

Krystyna Laszki-Szcząchor

1 Redagowanie i edycja tekstów medycznych w programie Word

Pisanie i edycja tekstów medycznych jest jedną z podstawowych form przygotowania raportów, tekstów naukowych, raportów z przeprowadzonych projektów czy też sprawozdań. Poniższe ćwiczenia mają za zadanie zaznajomienie studentów z podstawami edycji tekstu medycznego oraz w późniejszej części zastosowaniem odsyłaczy w tekście. Tekst jest kompilacją przypadkowych fragmentów z wykorzystaniem różnych form formatowania.

1.1. Formatowanie tekstu w Wordzie

1.1.1. Formatowanie tekstu i listy numerowane

1.1.1.1. Krój, wielkość czcionki, interlinia, justowanie, obramowania

Fragment tekstu:

W zakresie diagnozowania schorzeń serca podstawowe znaczenie ma w dalszym ciągu interpretacja sygnału EKG, będącego odzwierciedleniem zjawisk elektrycznych w sercu.

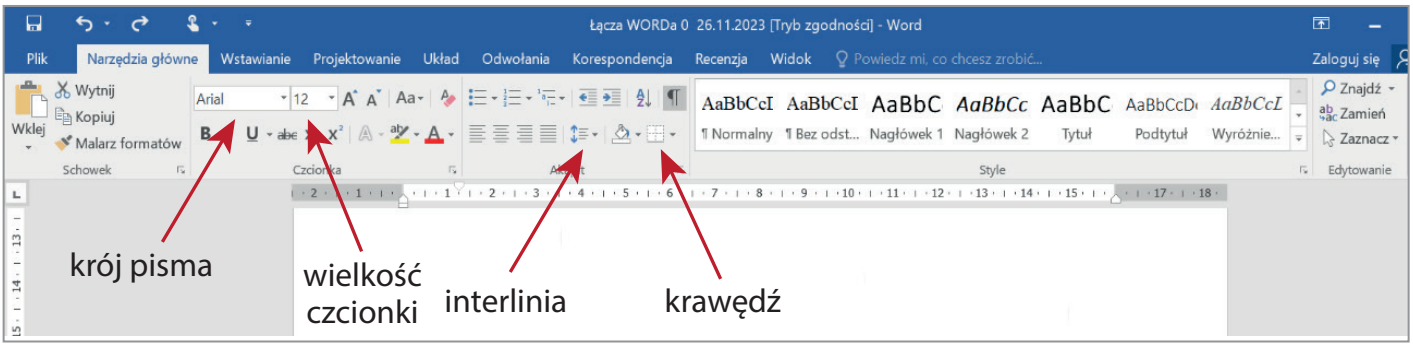
należy sformatować do postaci przedstawionej poniżej.

Tekst po edycji:

> W zakresie diagnozowania schorzeń serca podstawowe znaczenie ma w dalszym ciągu interpretacja sygnału EKG, będącego odzwierciedleniem zjawisk elektrycznych w sercu.

Dobieramy odpowiedni krój pisma, wielkość czcionki oraz interlinię ze wstążki. Tekst justujemy. Boczne linie wstawiane są z obramowania lub krawędzi (➥ **ryc. 1.1**).

Ryc. 1.1. Wstążka narzędziowa – karta Narzędzia główne



1.1.1.2. Indeksy górny i dolny, lista numerowana, symbole

Proszę dobrać odpowiednie parametry tekstu.

… postawienie diagnozy dm na podstawie wyliczonej macierzy parametrów Wk, gdzie k identyfikuje zapis danego pacjenta:

deterministyczne rozpoznanie patologii; wynik implikacji typu: Wk Dd → dm

gdzie Dd zawiera (opisuje) macierze parametrów Wk charakterystyczne dla patologii dm (Dd ⊂ D, gdzie D jest przestrzenią diagnoz…

Tekst po edycji:

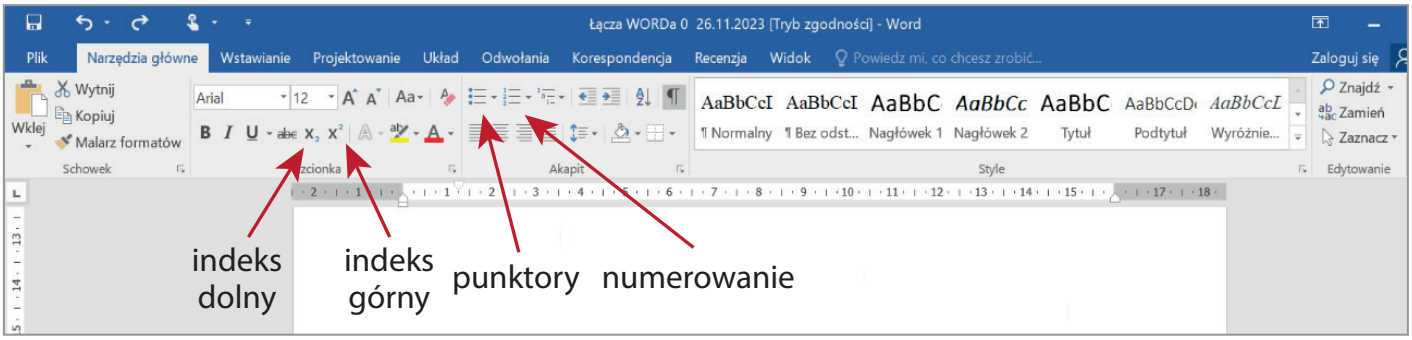
1. postawienie diagnozy d^m na podstawie wyliczonej macierzy parametrów W_k, gdzie k identyfikuje zapis danego pacjenta,
2. deterministyczne rozpoznanie patologii; wynik implikacji typu:

$$W_k \in D_d \rightarrow d^m$$

   a. gdzie D_d zawiera (opisuje) macierze parametrów W_k charakterystyczne dla patologii d^m ($D_d \subset D$, gdzie D jest przestrzenią diagnoz)

Dobieramy odpowiedni krój pisma, wielkość czcionki, tworzymy listę numerowaną, wstawiamy indeksy górne i dolne oraz symbole wg wskazówek z ➥ **ryc. 1.2**.

Ryc. 1.2. Karta Narzędzia główne



1.1.1.3. Zmiana początku listy numerowanej

Tekst do edycji:

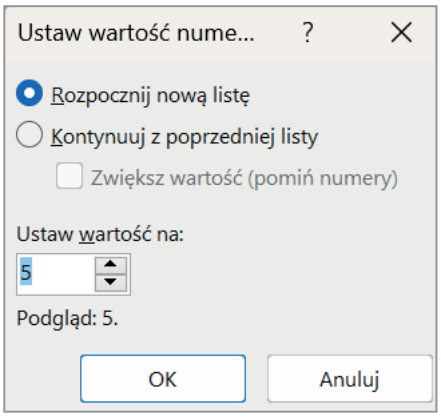
```
…1. Rozpoznanie probabilistyczne, przez wyliczenie
prawdopodobieństwa P dla każdego możliwego schorzenia
dn, tj. P(dn), gdzie n jest kolejnym numerem diagnozy
w słowniku diagnostycznym (n=1, 2, 3,.....N)…
```

Proszę tak dobrać parametr, żeby po edycji powyższy tekst wyglądał następująco:

5. Rozpoznanie probabilistyczne, przez wyliczenie prawdopodobieństwa P dla każdego możliwego schorzenia d^n, tj. P(d^n), gdzie n jest kolejnym numerem diagnozy w słowniku diagnostycznym (n=1, 2, 3,.....N)

Dobieramy odpowiedni krój, wielkość czcionki, wyrównanie tekstu, wstawiamy indeksy (na podstawie ➥ **ryc. 1.2**). Następnie przystępujemy do zmiany początku numeracji tekstu z 1 na 5. Po utworzeniu listy numerowanej przed edytowanym tekstem pojawia się liczba 1. Aby zmienić numerację, z rozwijanego menu polecenia Numeracja wybieramy pozycję Ustaw wartość numerowania (➥ **ryc. 1.3**).

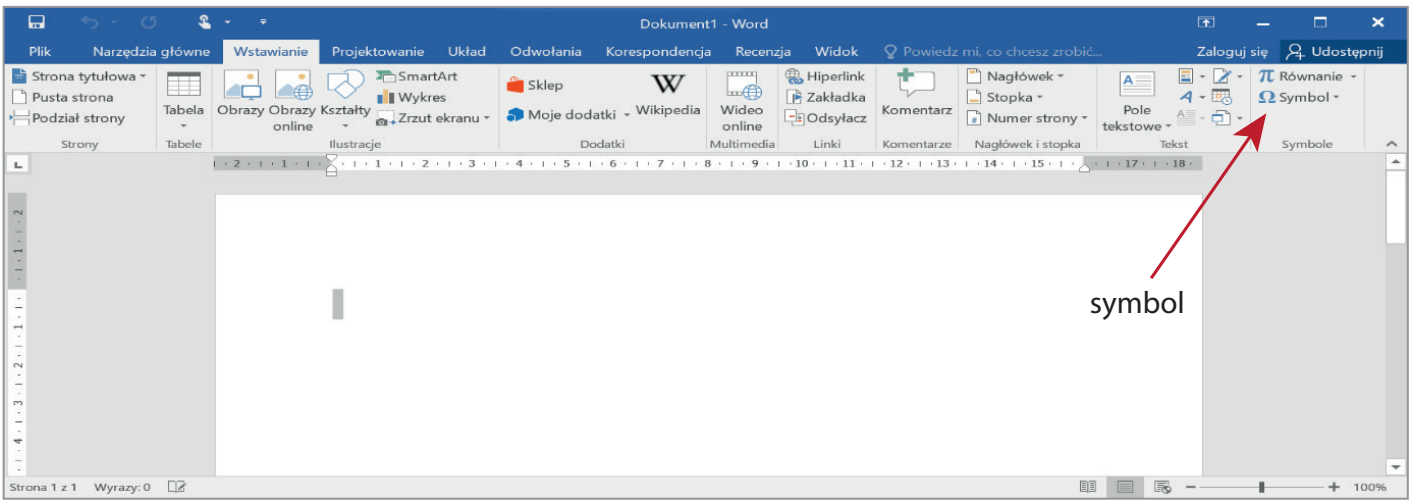
Ryc. 1.3. Okienko zmiany początku listy numerowanej



1.1.2. Edycja wzorów za pomocą polecenia Symbol

Ryc. 1.4. Umiejscowienie polecenia Symbol na karcie Wstawianie

Utwórz samodzielnie poniższe wzory, wykorzystując polecenie Symbol z karty Wstawianie (➥ **ryc. 1.4**).



Każdy wzór w opracowaniu jest oznaczony numerem w nawiasach. Do tych numerów odwołuje się autor w tekście. Tu mamy 4 wzory z różnymi numerami. Są różnie umieszczone w tekście.

$$P(dm) = maxn\{P(dn)\} \rightarrow dm \tag{4.28}$$

$$\beta_N^{|+|}\{P_{QRS}\} = 98{,}6\ \mu Vs \tag{4.30}$$

$$X(\omega) \approx \Delta t\ \Sigma x(n\ \Delta t)[\cos(\omega n\Delta t) - j\sin(\omega n\Delta t)] = \Delta t\ \Sigma x(n\Delta t)\ e\ -j\omega n\Delta t \tag{4.36}$$

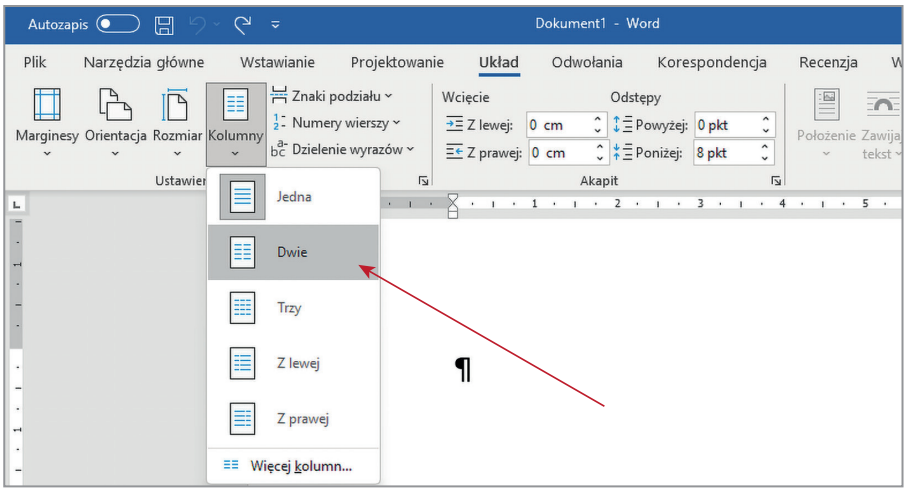
$$X(k\Delta\omega) \approx \Delta t^{*}\Sigma x(n\Delta t)\ e\ -j2k\Pi n/N = \Delta t^{*}DFTN[x(n\Delta t)] \tag{4.38}$$

1.1.3. Formatowanie tekstu w zapisie dwukolumnowym

Proszę sformatować poniższy tekst, dobierając odpowiedni krój i wielkość czcionki oraz interlinię, a następnie zastosować justowanie tekstu. Po zaznaczeniu całego tekstu należy go podzielić na kolumny (szpaltowanie, ➥ **ryc. 1.5**). Proszę dobrać odpowiednią szerokość odstępu między kolumnami.

Prawidłowa interpretacja przebiegów EKG wymaga wysokich kwalifikacji i dużego zasobu doświadczeń ze strony kardiologa. Te i szereg innych wymagań, jako istotnych elementów diagnostyki, powodują coraz szersze wprowadzanie elementów techniki komputerowej do procesu automatycznego analizowania i rozpoznawania sygnałów elektrokardiograficznych.

Ryc. 1.5. Zakładka Układ polecenie tekst dwukolumnowy



Tekst po opracowaniu powinien prezentować się następująco.

Prawidłowa interpretacja przebiegów EKG wymaga wysokich kwalifikacji i dużego zasobu doświadczeń ze strony kardiologa. Te i szereg innych wymagań, jako istotnych elementów diagnostyki, powodują coraz szersze wprowadzanie elementów techniki komputerowej do procesu automatycznego analizowania i rozpoznawania sygnałów elektrokardiograficznych.

1.1.4. Podział tekstu na akapity punktowane

Należy wydzielić poszczególne akapity w tekście i nadać im punktory użyte poniżej.

Przykładowy tekst:

Największą trudność w programach komputerowych do diagnostyki przebiegów EKG stanowi utworzenie procedur służących do rozpoznawania zaburzeń rytmu. We wdrożonym systemie podzielono je na grupy procedur służących do rozpoznawania zaburzeń rytmu. We wdrożonym systemie podzielono je na grupy obejmujące zaburzenia związane z kierunkami działania bodźców w sercu.
Zwolnienie lub przerwanie przewodzenia na jakimkolwiek poziomie określane jest mianem bloku i tak powstają kolejne grupy zaburzeń rytmu określane mianem bloków anatomicznych i czynnościowych. Bloki anatomiczne są następstwem uszkodzenia, które przerywa drogę fali pobudzenia. Na przykład wrodzony blok przedsionkowo-komorowy jest wynikiem zastąpienia tkanki przewodzącej tkanką łączną w okresie rozwoju płodowego.
Bloki czynnościowe występują, gdy potencjał czynnościowy komórek układu przewodzącego cechuje się zmniejszoną amplitudą i powolnym narastaniem fazy szybkiej depolaryzacji.

należy sformatować do postaci zaprezentowanej poniżej, używając ustawień akapitów punktowanych.

Tekst po edycji powinien wyglądać następująco:

Największą trudność w programach komputerowych do diagnostyki przebiegów EKG stanowi utworzenie procedur służących do rozpoznawania zaburzeń rytmu. We wdrożonym systemie podzielono je na grupy obejmujące zaburzenia związane z kierunkami działania bodźców w sercu.

- Zwolnienie lub przerwanie przewodzenia na jakimkolwiek poziomie określane jest mianem bloku i tak powstają kolejne grupy zaburzeń rytmu określane mianem bloków anatomicznych i czynnościowych:
  - bloki anatomiczne są następstwem uszkodzenia, które przerywa drogę fali pobudzenia. Na przykład wrodzony blok przedsionkowo-komorowy jest wynikiem zastąpienia tkanki przewodzącej tkanką łączną w okresie rozwoju płodowego;
  - bloki czynnościowe występują, gdy potencjał czynnościowy komórek układu przewodzącego cechuje się zmniejszoną amplitudą i powolnym narastaniem fazy szybkiej depolaryzacji.

1.2. Formatowanie tabel i schematów

1.2.1. Tworzenie tabeli z wynikami badań

Proszę zaprojektować poniższą ➥ **tabelę 1.1**, uzupełnić dane w komórkach i ustawić odpowiednio w tabeli, zachowując ich układ. Oznaczenie ± proszę wstawić za pomocą polecenia Symbol (patrz ➥ ryc. 1.4).

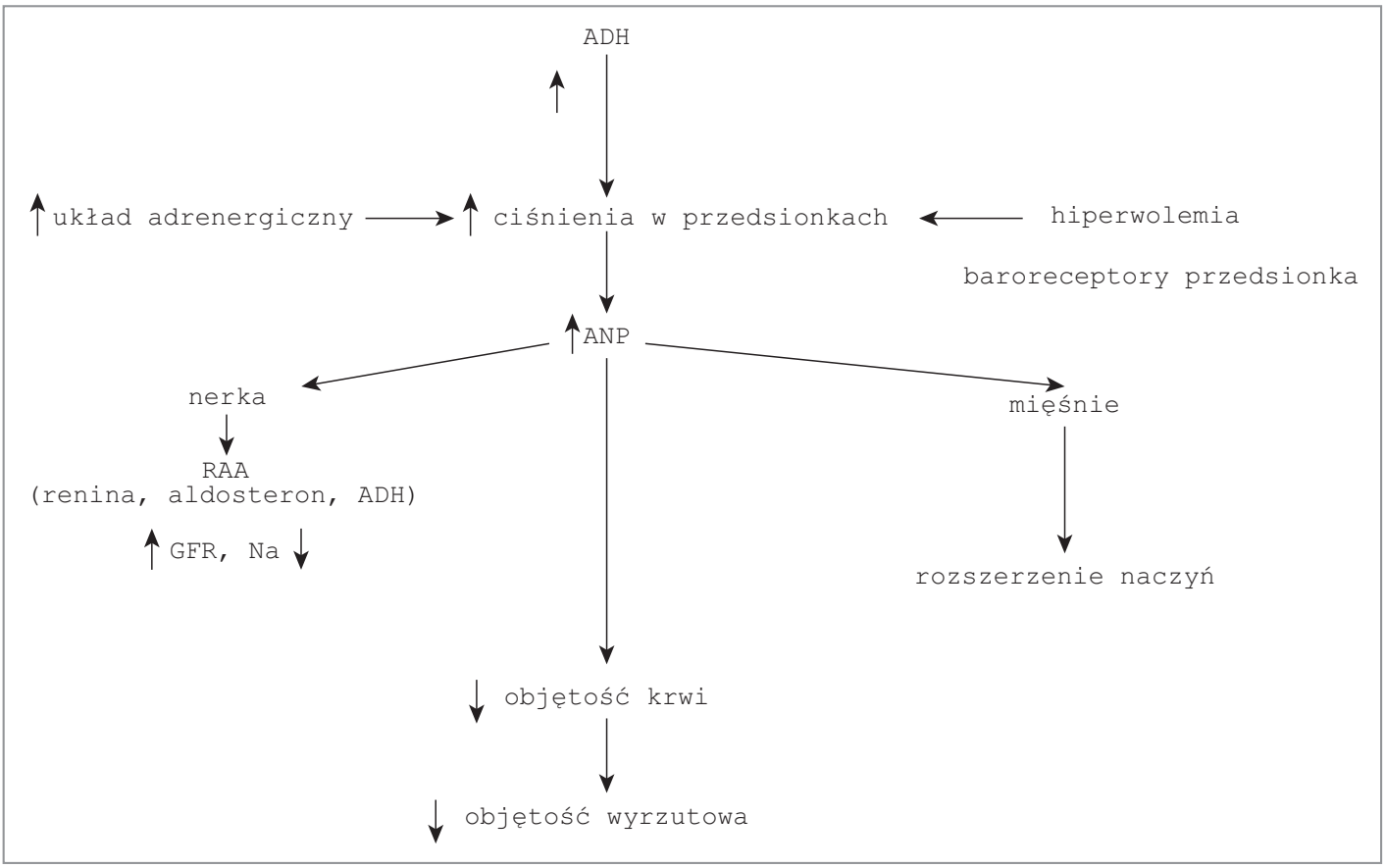
Parameters \ Group	I A	I B	II	III
t-Ca [mg/dl]	9,10 ± 0,78[cz]	9,55 ± 0,41[a]	9,22 ± 0,90	9,70 ± 0,61
P_i [mg/dl]	6,00 ± 0,69[cw]	4,93 ± 0,76	4,78 ± 0,55	4,94 ± 0,35
iPTH [pg/ml]	350,00 ± 120,84[bw]	267,25 ± 41,80[ks]	95,60 ± 69,65[x]	36,43 ± 9,14
Ca_i^{2+} [nmol/L]	71,50 ± 16,21[bfw]	47,01 ± 9,95[s]	40,96 ± 9,01[x]	26,85 ± 3,53
PMCA [umol/P_i/mg/min]	1,30 ± 0,14[bfw]	5,63 ± 1,22[ks]	4,30 ± 0,81[x]	3,13 ± 0,19
Ca × P [mg^2/dl^2]	54,43 ± 7,13[aw]	50.63 ± 3,86	46,80 ± 7,55	48,09 ± 5,12

Tabela 1.1. Tabela zawierająca wartości parametrów w poszczególnych grupach

1.2.2. Tworzenie schematu

Ryc. 1.6. Rola hormonu natriuretycznego w regulacji objętości osocza

Proszę utworzyć poniższy schemat, używając metody wstawiania pól tekstowych i autokształtów (➥ **ryc. 1.6**).



1.2.3. Tworzenie elementów blokowych

Proszę narysować, korzystając z polecenia Kształty z karty Wstawianie (➥ ryc. 1.4), strzałkę z wpisanym w środek prostokątem zawierającym podany poniżej tekst.

1.2.4. Tabela z różnym rozłożeniem komórek

Proszę wstawić ➥ **tabelę 1.2a** 5 × 8 i wpisać poniższą instrukcję resuscytacji, zmieniając komórki i tekst, żeby ostateczny jej układ był taki jak w ➥ **tabeli 1.2b**.

		Dorośli i dzieci > 8. r.ż.	Dzieci 1.–8. r.ż.	Niemowlęta
C	7. Tętno	Zbadaj tętno na tętnicy szyjnej, drugą ręką podtrzymaj żuchwę	Sprawdź tętno na tętnicy ramieniowej, drugą ręką podtrzymaj żuchwę	
Krążenie	8. Uciskanie klatki piersiowej	Uciskaj dolną część mostka 2 nadgarstkami na głębokość 4–5 cm z częstością 80–100/min	Uciskaj dolną część mostka nadgarstkiem jednej ręki na głębokość 2,5–3,5 cm z częstością 80–100/min	Uciskaj mostek 2 palcami, jeden palec poniżej linii międzysutkowej na głębokość 1,5–2,5 cm z częstością co najmniej 100/min
	9. Proporcje	1 oddech na 5 uciśnięć		
10. Ocena tętna	Co minutę przez 5 s oceniaj tętno; jeśli wystąpi brak tętna – podejmij czynności reanimacyjne, zaczynając od:			
	2 oddechy na 15 uciśnięć mostka 4 cykle	1 oddech na 5 uciśnięć mostka 10 cykli	1 oddech na 5 uciśnięć mostka 10 cykli	
	Jeśli wróci tętno, a utrzymuje się bezdech, prowadź sztuczną wentylację z częstością:			
1 oddech na 5 s	1 oddech na 4 s	1 oddech na 3 s U noworodka – 1 oddech na 2 s		

Tabela 1.2a. Instrukcja resuscytacji

Po edycji tabela powinna wyglądać następująco.

		Dorośli i dzieci > 8. r.ż.	**Dzieci 1.–8. r.ż.**	**Niemowlęta**
C	7. Tętno	Zbadaj tętno na tętnicy szyjnej, drugą ręką podtrzymaj żuchwę		Sprawdź tętno na tętnicy ramieniowej, drugą ręką podtrzymaj żuchwę
Krążenie	8. Uciskanie klatki piersiowej	Uciskaj dolną część mostka 2 nadgarstkami na głębokość 4–5 cm z częstością 80–100/min	Uciskaj dolną część mostka nadgarstkiem jednej ręki na głębokość 2,5–3,5 cm z częstością 80–100/min	Uciskaj mostek 2 palcami, jeden palec poniżej linii międzysutkowej na głębokość 1,5–2,5 cm z częstością co najmniej 100/min
	9. Proporcje	1 oddech na 5 uciśnięć		
	10. Ocena tętna	Co minutę przez 5 s oceniaj tętno; jeśli wystąpi brak tętna – podejmij czynności reanimacyjne, zaczynając od:		
		2 oddechy na 15 uciśnięć mostka **4 cykle**	1 oddech na 5 uciśnięć mostka **10 cykli**	1 oddech na 5 uciśnięć mostka **10 cykli**
		Jeśli wróci tętno, a utrzymuje się bezdech, prowadź sztuczną wentylację z częstością:		
		1 oddech na 5 s	1 oddech na 4 s	1 oddech na 3 s **U noworodka** – 1 oddech na 2 s

Tabela 1.2b. Instrukcja – rozwiązanie zadania

Krystyna Laszki-Szcząchor

2 Wstawianie odsyłaczy w tekście przy użyciu programu Word

Poniżej zaprezentowany jest fragment tekstu medycznego, do którego wstawimy odsyłacze do pozycji literatury zamieszczonej w części Piśmiennictwo na następnej stronie tego zadania. W pierwotnej wersji cyfry zostały po prostu wpisane w nawiasy umieszczone w tekście.

Założenia teoretyczne i samą koncepcję mappingu sformułował w XIX w. Waller [1]. Używając elektrometru kapilarnego, przeprowadził pomiary doświadczalne z użyciem 20 elektrod rozmieszczonych na powierzchni klatki piersiowej. Dotychczasowe prace w tej dziedzinie, wykorzystujące technikę komputerową, dotyczyły głównie analizy map w układzie klasycznym 12-elektrodowym, w układach 32-elektrodowych, 120- i 150-elektrodowych oraz w układzie sferycznym 30-elektrodowym [2, 3, 4].
Najmniej rozpoznane są mapy izochronowe. Tworzenie takiej mapy polega na łączeniu izoliniami punktów pomiarowych z powierzchni klatki piersiowej, w których wystąpił ten sam czas pobudzenia komór w sercu (ang. *ventricular activation time* – VAT). W badaniach wykorzystano mapy izochronowe do pokazania toru rozchodzenia się pobudzenia w sercu u osób zdrowych.
Badania w cylindrycznym 87-elektrodowym układzie prowadzono pod koniec lat 80. w Japonii i w latach 90. w Polsce z wykorzystaniem japońskiego aparatu firmy Fukuda Denshi umożliwiającego multielektrodowe rejestracje zapisów elektrokardiograficznych [5, 6].

Podczas pisania tego tekstu wykorzystano następujące artykuły naukowe przedstawione poniżej w ramach piśmiennictwa.

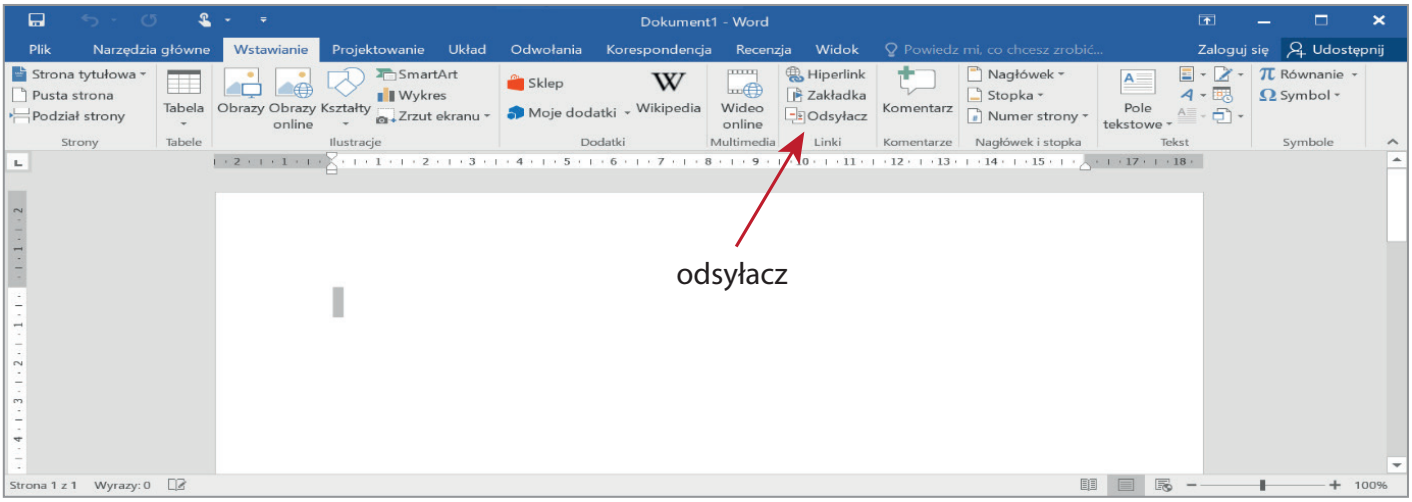
PIŚMIENNICTWO

1. Waller AD. A demonstration on men of electromotive changes accompanying the heart's beat. *J Physiol*. 1887;8:229–234.
2. Smith RE, Hyde CM. *Computer analysis of the electrocardiogram in clinical practice*. W: Manning GW, Ahuja SP, red. Electrical Activity of the Heart. Springfield, USA: Charles C Thomas Publisher; 1969:305–315.
3. Green LS, Lux L, Haws CW, Burgess MJ, Abildskov JA. Features of body surface potential maps from a large normal population. W: van Damm RT, van Oosterom A, red. *Electrocardiographic Body Surface Mapping*. Amsterdam, Holandia: Martinus Nijhoff; 1986:9–18.
4. Horacek BM, Warren JW, Field DQ, Feldman CL. Statistical and deterministic approaches to designing transformations of electrocardiographic leads. *J Electocardiol*. 2002;35(supl):41–52.
5. Hayashi H, Kojima H, Yabe S, et al. Clinical application of body surface mapping and its limitations. *J Cardiogr*. 1984;14(supl 3):41–49.
6. Jagielski J, Kałka D. Heart potential mapping. *Pol Arch Med Wewn*. 1992,88(2–3):103–115.

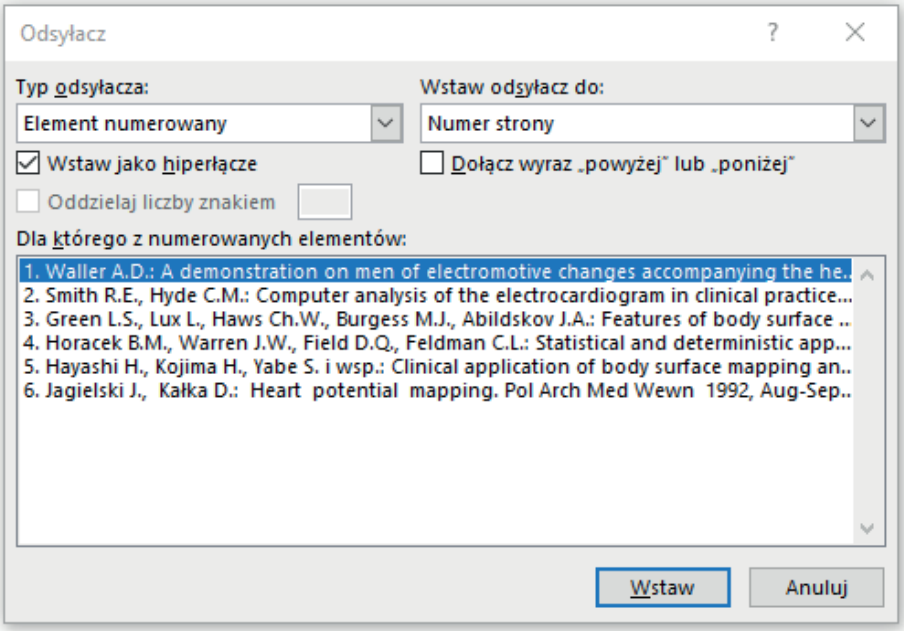
W tekście usuwamy nawias z cyfrą 1 przy słowie „Waller". Ustawiamy kursor myszy na jej miejscu, wybieramy kartę Wstawianie, a następnie polecenie Odsyłacz (➥ **ryc. 2.1**).

Pojawia się wyskakujące okienko Odsyłacz (➥ **ryc. 2.2**). Nie zmieniamy żadnych ustawień, jedynie w okienku z piśmiennictwem zaznaczamy pozycję „1. Waller…" i wciskamy przycisk Wstaw umieszczony na dole okienka.

Ryc. 2.1. Umiejscowienie polecenia Odsyłacz na karcie Wstawianie



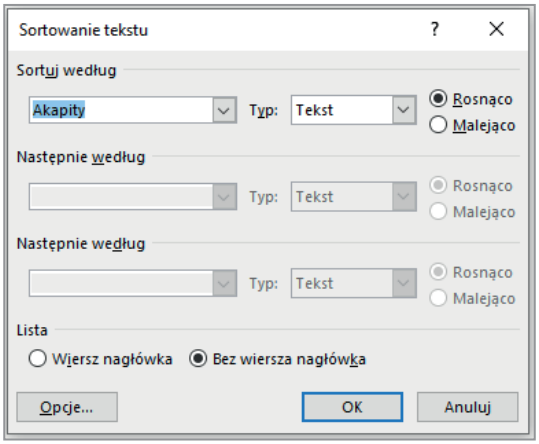
Ryc. 2.2. Okno odsyłacza z pozycjami piśmiennictwa



Z kolejnymi numerami pozycji piśmiennictwa w tekście powtarzamy tę samą sekwencję działań.

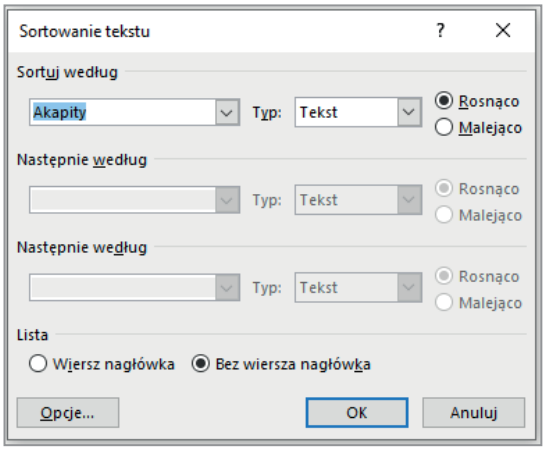
W przypadku gdy np. organizatorzy konferencji, na którą przygotowujemy pracę, wymagają alfabetycznego ustawienia kolejnych pozycji piśmiennictwa (już wcześniej przygotowanych), zaznaczamy wszystkie jego pozycje i korzystamy z polecenia Sortowanie na karcie Narzędzia główne (➥ **ryc. 2.3**).

Ryc. 2.3. Polecenie Sortowanie zaznaczone na wstążce



Po kliknięciu ikony Sortowanie pojawia się okno (➥ **ryc. 2.4**).

Ryc. 2.4. Okno sortowania



Nic w nim nie zmieniamy. Klikamy jedynie OK. Otrzymujemy listę piśmiennictwa uporządkowaną alfabetycznie.

PIŚMIENNICTWO

1. Green LS, Lux L, Haws CW, Burgess MJ, Abildskov JA. Features of body surface potential maps from a large normal population. W: van Damm RT, van Oosterom A, red. *Electrocardiographic Body Surface Mapping*. Amsterdam, Holandia: Martinus Nijhoff; 1986:9–18.
2. Hayashi H, Kojima H, Yabe S, et al. Clinical application of body surface mapping and its limitations. *J Cardiogr* 1984;14(supl 3):41–49.
3. Horacek BM, Warren JW, Field DQ, Feldman CL. Statistical and deterministic approaches to designing transformations of electrocardiographic leads. *J Electocardiol*. 2002;35(supl):41–52.
4. Jagielski J, Kałka D. Heart potential mapping. *Pol Arch Med Wewn*. 1992;88(2–3):103–115.
5. Smith RE, Hyde CM. *Computer analysis of the electrocardiogram in clinical practice. Electrical Activity of the Heart*. CH.C. Thomas, Springfield, USA. 1969:305–315.
6. Waller AD. A demonstration on men of electromotive changes accompanying the heart's beat. *J Physiol*. 1887;8:229–234.

Po kliknięciu odnośnika 1 z wciśniętym klawiszem Ctrl program odwołuje się w nowym układzie Piśmiennictwa do cytowanej publikacji Wallera, ale numery w tekście i w spisie są różne.

Aby uaktualnić numerację, zaznaczamy cały tekst artykułu. Klikamy zaznaczenie prawym przyciskiem myszy i z menu podręcznego wybieramy Aktualizuj pole. Numeracja w nawiasach kwadratowych zaktualizuje się do obecnego spisu piśmiennictwa. Artykuł przyjmuje poniższą postać. Numeracje są już zgodne.

Założenia teoretyczne i samą koncepcję mappingu sformułował w XIX w. Waller [6]. Używając elektrometru kapilarnego, przeprowadził pomiary doświadczalne z użyciem 20 elektrod rozmieszczonych na powierzchni klatki piersiowej. Dotychczasowe prace w tej dziedzinie, wykorzystujące technikę komputerową, dotyczyły głównie analizy map w układzie klasycznym 12 elektrodowym, w układach 32-elektrodowych, 120- i 150-elektrodowych oraz w układzie sferycznym 30-elektrodowym [5, 1, 3].
Najmniej rozpoznane są mapy izochronowe. Tworzenie takiej mapy polega na łączeniu izoliniami punktów pomiarowych z powierzchni klatki piersiowej, w których wystąpił ten sam czas pobudzenia komór w sercu (ang. *ventricular activation time* – VAT). W badaniach wykorzystano mapy izochronowe do pokazania toru rozchodzenia się pobudzenia w sercu u osób zdrowych
Badania w cylindrycznym 87-elektrodowym układzie prowadzono pod koniec lat 80. w Japonii i w latach 90. w Polsce z wykorzystaniem japońskiego aparatu firmy Fukuda Denshi umożliwiającego multielektrodowe rejestracje zapisów elektrokardiograficznych [2, 4].

Kiedy piśmiennictwo musimy uzupełnić o jakąkolwiek liczbę pozycji – zawsze wpisujemy je na koniec listy (tutaj pod nr. 7).

PIŚMIENNICTWO

1. Green LS, Lux L, Haws CW, Burgess MJ, Abildskov JA. Features of body surface potential maps from a large normal population. W: van Damm RT, van Oosterom A, red. *Electrocardiographic Body Surface Mapping*. Amsterdam, Holandia: Martinus Nijhoff; 1986:9–18.
2. Hayashi H, Kojima H, Yabe S, et al. Clinical application of body surface mapping and its limitations. *J Cardiogr*. 1984;14(supl 3):41–49.
3. Horacek BM, Warren JW, Field DQ, Feldman CL. Statistical and deterministic approaches to designing transformations of electrocardiographic leads. *J Electocardiol*. 2002;35(supl):41–52.
4. Jagielski J, Kałka D. Heart potential mapping. *Pol Arch Med Wewn*. 1992;88(2–3):103–115.
5. Smith RE, Hyde CM. *Computer analysis of the electrocardiogram in clinical practice. Electrical Activity of the Heart*. CH.C. Thomas, Springfield, USA. 1969:305–315.
6. Waller AD. A demonstration on men of electromotive changes accompanying the heart's beat. *J Physiol*. 1887;8:229–234.
7. Kurpesa M, Szafran B. *Interpretacja EKG. Kurs zaawansowany*. Warszawa, Polska: Wydawnictwo Lekarskie PZWL;2020.

Wstawiamy nowy odsyłacz **[7]** do tekstu.

Założenia teoretyczne i samą koncepcję mappingu sformułował w XIX w. Waller [6]. Używając elektrometru kapilarnego, przeprowadził pomiary doświadczalne z użyciem 20 elektrod rozmieszczonych na powierzchni klatki piersiowej. Dotychczasowe prace w tej dziedzinie, wykorzystujące technikę komputerową, dotyczyły głównie analizy map w układzie klasycznym 12-elektrodowym, w układach 32-elektrodowych, 120- i 150-elektrodowych oraz w układzie sferycznym 30-elektrodowym [5, 1, 3].
Najmniej rozpoznane są mapy izochronowe. Tworzenie takiej mapy polega na łączeniu izoliniami punktów pomiarowych z powierzchni klatki piersiowej, w których wystąpił ten sam czas pobudzenia komór w sercu (ang. *ventricular activation time* – VAT). W badaniach wykorzystano mapy izochronowe do pokazania toru rozchodzenia się pobudzenia w sercu u osób zdrowych [7].
Badania w cylindrycznym 87-elektrodowym układzie prowadzono pod koniec lat 80. w Japonii i w latach 90. w Polsce z wykorzystaniem japońskiego aparatu firmy Fukuda Denshi umożliwiającego multielektrodowe rejestracje zapisów elektrokardiograficznych [2, 4].

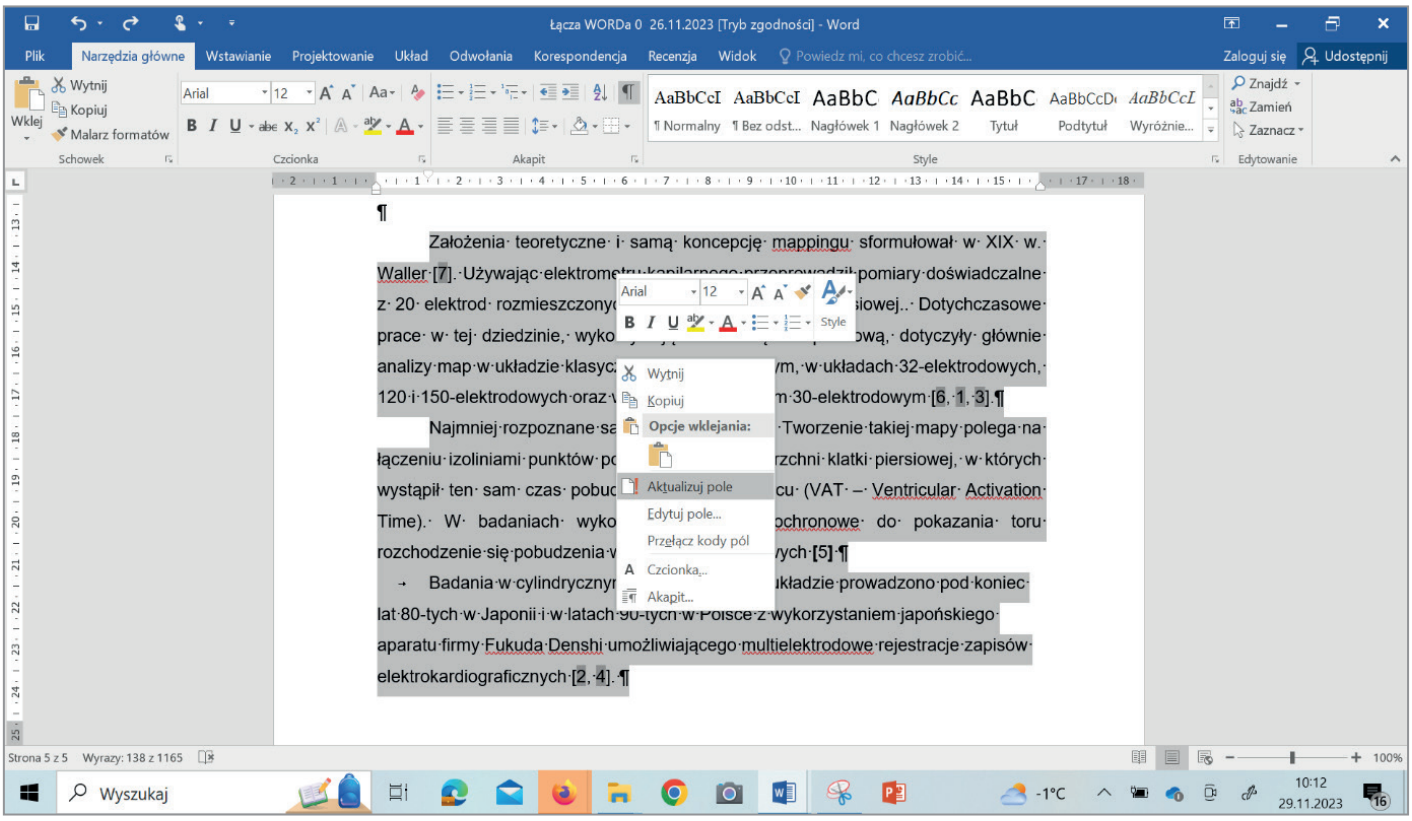
Ponownie zaznaczamy wszystkie pozycje piśmiennictwa. Ustawiamy je alfabetycznie (zgodnie z procedurą opisaną powyżej).

PIŚMIENNICTWO

1. Green LS, Lux L, Haws CW, Burgess MJ, Abildskov JA. Features of body surface potential maps from a large normal population. W: van Damm RT, van Oosterom A, red. *Electrocardiographic Body Surface Mapping*. Amsterdam, Holandia: Martinus Nijhoff; 1986:9–18.
2. Hayashi H, Kojima H, Yabe S, et al: Clinical application of body surface mapping and its limitations. *J Cardiogr*. 1984;14(supl 3):41–49.
3. Horacek BM, Warren JW, Field DQ, Feldman CL. Statistical and deterministic approaches to designing transformations of electrocardiographic leads. *J Electocardiol*. 2002;35(supl):41–52.
4. Jagielski J, Kałka D: Heart potential mapping. *Pol Arch Med Wewn*. 1992;88(2–3):103–115.
5. Kurpesa M, Szafran B. *Interpretacja EKG. Kurs zaawansowany*. PZWL. 2020.
6. Smith RE, Hyde CM. *Computer analysis of the electrocardiogram in clinical practice. Electrical Activity of the Heart*. CH.C. Thomas, Springfield, USA. 1969:305–315.
7. Waller AD. A demonstration on men of electromotive changes accompanying the heart's beat. *J Physiol*. 1887;8:229–234.

Ryc. 2.5. Zaznaczony tekst artykułu z wyświetlonym menu do aktualizacji pola

Jeszcze raz zaznaczamy tekst artykułu, aktualizujemy pole (➡ **ryc. 2.5**) i otrzymujemy efekt widoczny poniżej.



Poniżej zaprezentowano tekst uzupełniony o dopisane pozycje z ustawieniem alfabetycznym.

Założenia teoretyczne i samą koncepcję mappingu sformułował w XIX w. Waller [7]. Używając elektrometru kapilarnego, przeprowadził pomiary doświadczalne z użyciem 20 elektrod rozmieszczonych na powierzchni klatki piersiowej. Dotychczasowe prace w tej dziedzinie, wykorzystujące technikę komputerową, dotyczyły głównie analizy map w układzie klasycznym 12-elektrodowym, w układach 32-elektrodowych, 120- i 150-elektrodowych oraz w układzie sferycznym 30-elektrodowym [6, 1, 3].

Najmniej rozpoznane są mapy izochronowe. Tworzenie takiej mapy polega na łączeniu izoliniami punktów pomiarowych z powierzchni klatki piersiowej, w których wystąpił ten sam czas pobudzenia komór w sercu (ang. *ventricular activation time* – VAT). W badaniach wykorzystano mapy izochronowe do pokazania toru rozchodzenia się pobudzenia w sercu u osób zdrowych [5].

Badania w cylindrycznym 87-elektrodowym układzie prowadzono pod koniec lat 80. w Japonii i w latach 90. w Polsce z wykorzystaniem japońskiego aparatu firmy Fukuda Denshi umożliwiającego multielektrodowe rejestracje zapisów elektrokardiograficznych [2, 4].

Andrii Zinchuk

3 Przygotowanie wykresów na podstawie danych

3.1. Znaczenie wizualizacji danych w biostatystyce

Analiza i interpretacja danych są w biostatystyce niezbędne do badania wzorców (w tym trendów zmian) oraz zależności, które mogą pomóc naukowcom w zrozumieniu i wyjaśnieniu badanych procesów.

Wizualizacja danych to proces przedstawiania informacji w formie graficznej, która pozwala łatwiej interpretować złożone dane i odkrywać wzorce, trendy oraz relacje między zmiennymi. Jest ona szczególnie użyteczna w biostatystyce, ponieważ badania w tej dziedzinie często generują duże ilości złożonych danych, które mogą być trudne do zrozumienia bez wizualnej pomocy.

Rozmaite typy wykresów i diagramów, takie jak histogramy, wykresy punktowe, pudełkowe czy skrzypcowe, można stosować do przedstawiania różnych aspektów danych. Na przykład histogramy mogą być użyte do ilustracji rozkładu częstości wartości danej zmiennej, podczas gdy wykresy punktowe mogą przedstawiać zależności między dwiema zmiennymi.

Jedną z głównych cech wizualizacji danych jest jej zdolność do ukazania ukrytych wzorców i zależności, które mogą nie być oczywiste na pierwszy rzut oka. W biostatystyce wynajdywanie takich zależności może prowadzić do ważnych odkryć naukowych.

Wizualizacja to jednak przede wszystkim narzędzie komunikowania wyników. Forma graficzna pozwala zebrać rozmaite dane liczbowe w postaci czytelnego wykresu z etykietami, które w znacznym stopniu ułatwiają zrozumienie danego zjawiska.

Zwłaszcza w erze analizy dużych zbiorów liczbowych naukowcy są często zmuszeni do opracowania ogromnej ilości danych. Przejrzyste i efektywne prezentowanie wyników może być wyzwaniem. W tym świetle umiejętność tworzenia jasnych i zrozumiałych wykresów jest nieoceniona i skraca czas podjęcia trafnej decyzji.

Wizualizacja danych jest nie tylko ważnym narzędziem dla naukowców, ale również dla uczących się biostatystyki. W edukacji wizualizacje mogą pomóc studentom zrozumieć i przyswoić trudne koncepcje. W szkoleniach z zakresu statystyki często wykorzystywane są z narzędzia do wizualizacji danych, takie jak pakiety ggplot2, plotly lub RGL, pracujące w środowisku R, do nauczania interpretacji i analizy danych.

Wizualizacja danych w biostatystyce jest więc niezbędna do zrozumienia, interpretacji i komunikowania wyników analizy danych medycznych. Stosowanie odpowiednich jej technik może przyczynić się do odkrycia nowych wzorców i zależności, a także ułatwić interpretację wyników badania.

3.2. Cele przygotowania wykresów

Pierwszym celem wykorzystania wykresów w biostatystyce jest ułatwienie zrozumienia danych. Nawet najbardziej skomplikowane ich zestawienia można przedstawić w sposób przystępny za pomocą odpowiednio dobranych wykresów. Pozwalają one na łatwe identyfikowanie trendów, wyjątków czy pomiarów odstających, które mogą być kluczowe dla interpretacji wyników.

Drugim celem jest komunikacja wyników badania. Wykresy ułatwiają przedstawienie wyników innym naukowcom, studentom, a nawet szerszej publiczności. Są niezbędne w artykułach naukowych, prezentacjach czy publikacjach edukacyjnych. Wyraźne i przemyślane wykresy pomagają przedstawić wyniki badania w sposób, który jest zrozumiały dla szerokiej grupy odbiorców.

Trzecim celem jest ułatwienie procesu decyzyjnego. Wykresy pomagają w identyfikacji wzorców i trendów istotnych dla procesu decyzyjnego w badaniu naukowym czy klinicznym. Mogą np. pomóc lekarzom w identyfikacji czynników ryzyka chorób lub w monitorowaniu efektów leczenia.

Czwartym celem jest wykorzystanie ich jako narzędzi edukacyjnych. Mogą niezwykle skutecznie pomagać studentom w zrozumieniu złożonych koncepcji statystycznych i biostatystycznych. Dzięki wykorzystaniu programów do analizy danych, takich jak R, studenci mogą nauczyć się, jak tworzyć, interpretować i krytycznie oceniać wykresy.

Wykresy mogą być także narzędziem ułatwiającym analizę danych. Pozwalają na wizualizację wielowymiarowych danych, co może ujawnić zależności, które nie były wcześniej oczywiste, a to może prowadzić do nowych hipotez i kierunków badawczych.

Wykresy odgrywają kluczową rolę w biostatystyce. Ułatwiają zrozumienie i interpretację danych, pomagają w komunikacji wyników, wspierają procesy decyzyjne, służą jako narzędzia edukacyjne i pozwalają na analizę danych. Przy odpowiednim wykorzystaniu mogą przyczyniać się do postępu naukowego w dziedzinie biostatystyki.

3.3. Podstawy statystyki

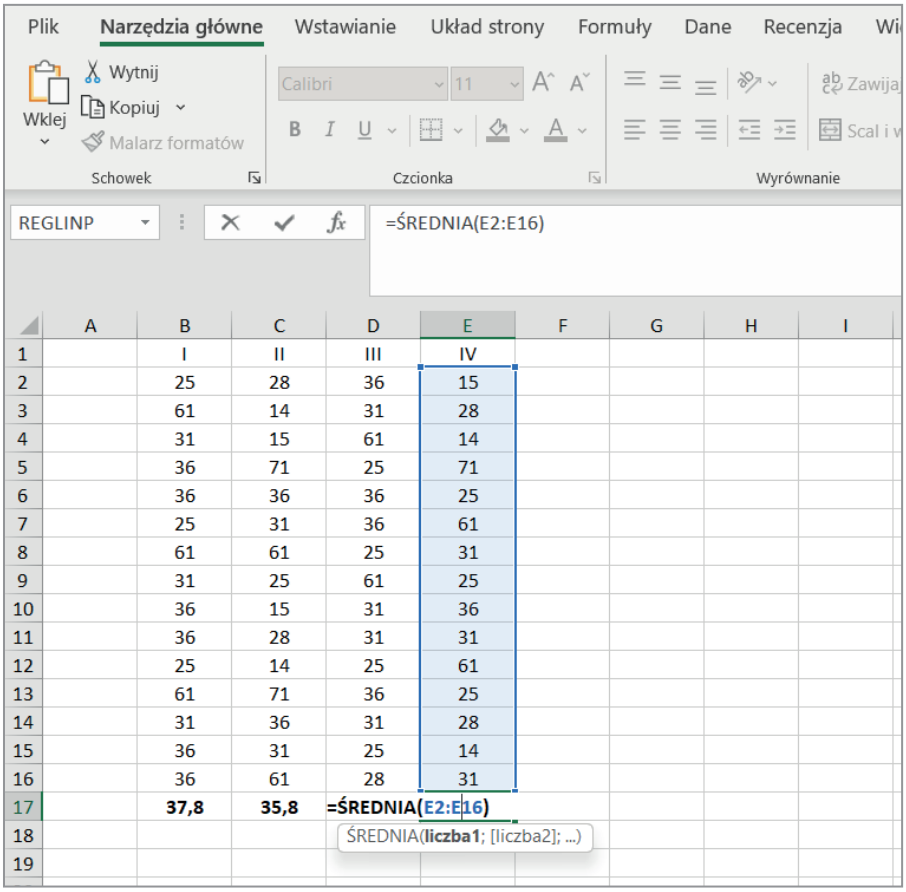
3.3.1. Omówienie podstawowych terminów statystycznych

Statystyka opisowa oferuje szereg narzędzi umożliwiających zrozumienie i podsumowanie danych. Podstawowe terminy statystyczne to: średnia, mediana, moda, wariancja i odchylenie standardowe. Rozważmy każdy z tych terminów osobno. Zastosowanie estymatorów* w postaci wartości średniej, odchylenia standardowego i wariancji ma sens tylko wtedy, gdy rozkład zmiennej jest zbliżony do normalnego. Sprawdzenie normalności rozkładu danych jest standardową procedurą, która upewnia nas co do możliwości użycia wymienionych statystyk.

Średnia, znana również jako średnia arytmetyczna, jest podstawowym miernikiem tendencji centralnej (określa taką wartość zmiennej, przy której skupiają się wartości danej zmiennej). Wartości te są jedną z podstawowych grup statystyki opisowej – pozostałymi są miary rozproszenia oraz symetrii rozkładu. Do miar tendencji centralnej zaliczamy średnią, dominantę oraz medianę. Średnią oblicza się, dodając wszystkie wartości w zbiorze, a następnie dzieląc przez liczbę tych wartości. Formuła w Excelu do obliczenia średniej dla zbioru n punktów danych (x_1, x_2, …, x_n) jest pokazana na ➥ **ryc. 3.1**.

* Estymator w statystyce to funkcja lub reguła, która służy do oszacowania wartości parametru populacji na podstawie próby losowej. Parametr populacji to wartość opisująca cechę całej populacji, np. średnia, wariancja, mediana.

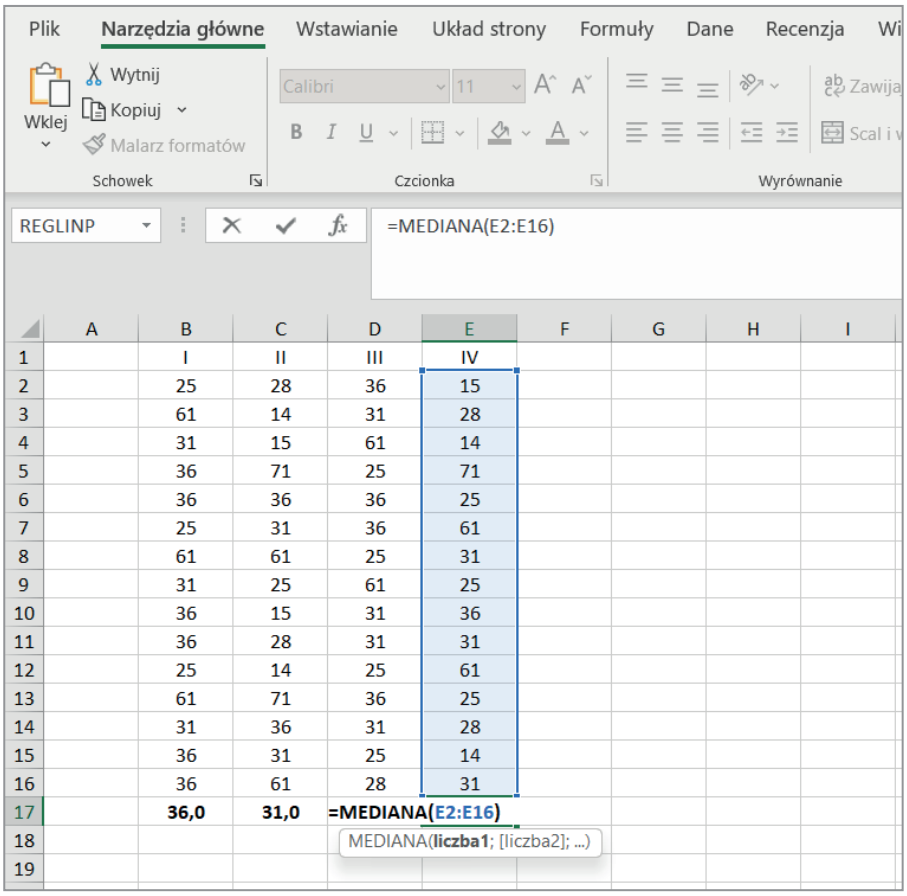
Ryc. 3.1. Określenie wartości średniej w arkuszu kalkulacyjnym



	A	B	C	D	E	F	G	H	I
1		I	II	III	IV				
2		25	28	36	15				
3		61	14	31	28				
4		31	15	61	14				
5		36	71	25	71				
6		36	36	36	25				
7		25	31	36	61				
8		61	61	25	31				
9		31	25	61	25				
10		36	15	31	36				
11		36	28	31	31				
12		25	14	25	61				
13		61	71	36	25				
14		31	36	31	28				
15		36	31	25	14				
16		36	61	28	31				
17		**37,8**	**35,8**						
18									
19									

Mediana jest punktem centralnym zestawu danych. To wartość, która dzieli zbiór danych na dwie równe części: połowa punktów danych jest mniejsza od mediany, a połowa jest większa. Jeżeli liczba punktów danych (n) jest nieparzysta, mediana to wartość środkowa. Jeśli n jest parzyste, mediana jest średnią dwóch wartości środkowych. Sposób zastosowania formuły z medianą przedstawia ➥ **ryc. 3.2**.

Ryc. 3.2. Określenie wartości mediany w arkuszu kalkulacyjnym



	A	B	C	D	E	F	G	H	I
1		I	II	III	IV				
2		25	28	36	15				
3		61	14	31	28				
4		31	15	61	14				
5		36	71	25	71				
6		36	36	36	25				
7		25	31	36	61				
8		61	61	25	31				
9		31	25	61	25				
10		36	15	31	36				
11		36	28	31	31				
12		25	14	25	61				
13		61	71	36	25				
14		31	36	31	28				
15		36	31	25	14				
16		36	61	28	31				
17		**36,0**	**31,0**						
18									
19									

Moda to wartość, która występuje najczęściej w zestawie danych. Zbiór danych może mieć jedną modę (unimodalny), dwie mody (bimodalny) lub więcej mod (multimodalny). O zestawach danych, w których żadna wartość nie powtarza się częściej niż inne, mówimy, że nie ma mody.

Wariancja jest miarą rozproszenia danych. Pokazuje, jak bardzo dane różnią się od średniej. Oblicza się ją, sumując kwadraty odchyleń każdej wartości od średniej (czyli różnic), a następnie dzieląc przez liczbę punktów danych.

Odchylenie standardowe jest również miarą rozproszenia danych, ale jest ono wyrażone w tych samych jednostkach co dane. Jest to pierwiastek kwadratowy z wariancji. Wzór do obliczania odchylenia standardowego (σ) to:

odchylenie standardowe (populacja)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

n – ilość wartości
μ – średnia arytmetyczna

odchylenie standardowe (próba)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

n – ilość wartości
$\bar{x}$ – średnia arytmetyczna

Źródło: https://techporady.pl/

3.3.2. Przykłady zastosowania podstawowych estymatorów w biostatystyce

Biostatystyka, będąca jednym z najistotniejszych narzędzi w dziedzinie badań medycznych, wykorzystuje podstawowe estymatory statystyczne w różnych kontekstach.

Średnia, jako miara tendencji centralnej, jest często stosowana do określania średniego wyniku w badaniach klinicznych lub epidemiologicznych. Może być np. użyta do obliczenia średniego ciśnienia krwi w grupie pacjentów lub do obliczenia średniego indeksu masy ciała (ang. *body mass index* – BMI) w populacji i zidentyfikowania potencjalnego problemu z otyłością.

Mediana również jest użyteczna, szczególnie gdy dane są skośne i średnia nie oddaje dobrze reprezentacji tendencji centralnej. W populacji badanej pod kątem choroby można określić medianę wieku. Mediana może być również wykorzystana do oceny średniego czasu przeżycia w grupie pacjentów chorych na raka, gdzie nieliczne przypadki bardzo długiego przeżycia mogą zniekształcać średnią.

Moda, choć rzadziej używana niż średnia lub mediana, jest przydatna w określonych przypadkach. Można ją wykorzystywać do identyfikacji najczęściej występującej wartości, np. najczęściej stosowanego leku w grupie pacjentów, oraz najczęściej występujących objawów w grupie pacjentów z konkretnym schorzeniem (np. chorych na grypę).

Wariancja i odchylenie standardowe należą do najważniejszych miar rozproszenia danych. Powszechnie używa się ich do oceny zmienności wśród wyników badań klinicznych, np. zmienności stężenia glukozy we krwi pacjentów z cukrzycą. Odchylenie standardowe może pomóc w ocenie skuteczności nowego leku na obniżenie ciśnienia krwi poprzez analizę zmienności wyników między grupą kontrolną a grupą osób leczonych.

W praktyce wykorzystanie wyżej opisanych estymatorów może znacznie pomóc w interpretacji wyników badań i ułatwić zrozumienie charakterystyki badanej populacji. Wybór odpowiedniego narzędzia zależy od rodzaju danych, celu analizy oraz kontekstu badania.

3.4. Rodzaje wykresów

3.4.1. Wykresy kolumnowe i słupkowe

Należą do najpopularniejszych metod graficznego prezentowania danych i są często stosowane w biostatystyce.

Wykres kolumnowy prezentuje dane za pomocą pionowych słupków, gdzie wysokość każdego słupka jest proporcjonalna do reprezentowanej wartości. W biostatystyce wykresy kolumnowe mogą być stosowane do porównania np. średniego ciśnienia krwi w różnych grupach pacjentów (grupa kontrolna vs. grupa leczona).

Z kolei w wykresie słupkowym słupki są położone poziomo. Wykresy słupkowe są szczególnie użyteczne, gdy mamy do czynienia z długimi nazwami kategorii, które mogą być trudne do umieszczenia na osi pionowej wykresu kolumnowego. W biostatystyce wykresy słupkowe mogą być używane do porównań np. liczby pacjentów z różnymi schorzeniami w szpitalu.

Zarówno wykresy kolumnowe, jak i słupkowe są skutecznymi narzędziami wizualizacji danych, które umożliwiają łatwe porównywanie wartości pomiędzy różnymi grupami lub kategoriami. Są niezwykle przydatne w biostatystyce, gdzie takie porównania są często niezbędne do interpretacji wyników badań.

3.4.2. Wykresy kołowe

Wykresy kołowe, znane również jako diagramy kołowe, są rodzajem grafiki, który przedstawia procentowe udziały całkowitości. Są często stosowane w biostatystyce do przedstawiania proporcji dla zmiennych jakościowych (skategoryzowanych).

W wykresie kołowym koło reprezentuje całą populację lub zbiór danych. Każda kategoria reprezentowana jest przez segment (lub „kawałek") koła, którego wielkość jest proporcjonalna do wielkości reprezentowanej kategorii.

Wykresy kołowe są szczególnie użyteczne, gdy chcemy zobrazować, jak dużą część całości stanowi każda z kategorii. W biostatystyce można ich używać np. do prezentacji udziału poszczególnych przyczyn zgonów w określonym czasie lub miejscu.

Należy jednak pamiętać, że wykresy kołowe mogą być nieczytelne, gdy liczba kategorii jest duża. Zbyt duża liczba segmentów może sprawić, że trudno będzie odróżnić poszczególne kategorie. Ponadto percepcja kątów i obszarów w wykresach kołowych może być mniej precyzyjna niż percepcja długości i wysokości, co sprawia, że wykresy kołowe mogą być mniej dokładne od wykresów słupkowych i kolumnowych.

Są one przydatnym narzędziem wizualizacji danych w biostatystyce, choć zaleca się ostrożne ich używanie, zwłaszcza gdy liczba kategorii jest duża.

3.4.3. Wykresy punktowe (rozrzutu) i liniowe

Są to dwie zasadnicze formy przedstawiania danych w kontekście zależności między dwiema zmiennymi. Są one szczególnie przydatne w biostatystyce, gdzie często bada się związki między różnymi czynnikami.

Wykres punktowy (ang. *scatter plot*) to dwuwymiarowy wykres, na którym punkty reprezentują obserwacje, a ich położenie w układzie współrzędnych wynika z wartości dwóch zmiennych. Jest to doskonałe narzędzie do przedstawiania zależności między dwiema zmiennymi ilościowymi.

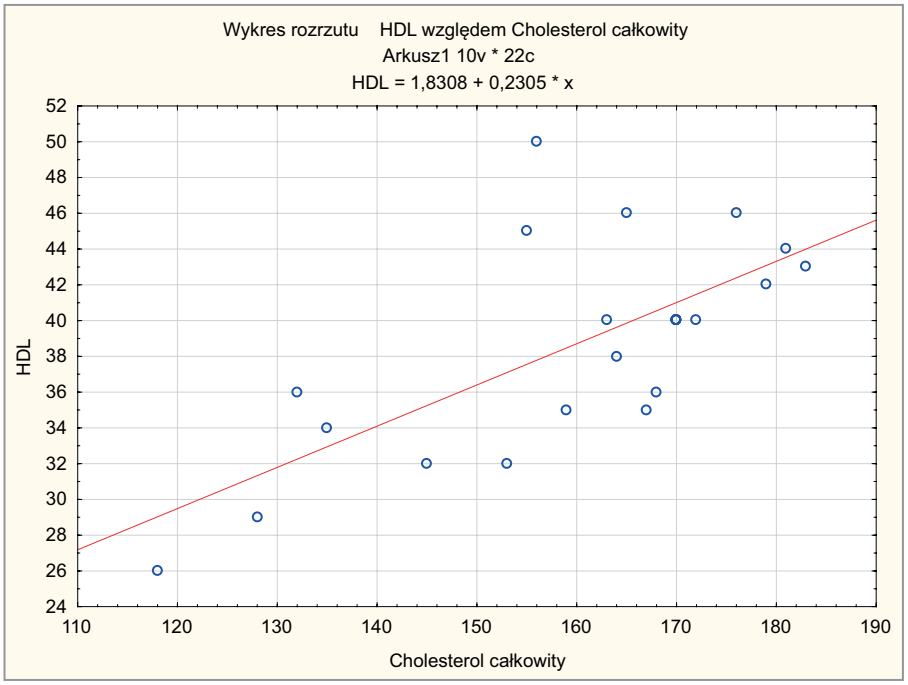
Wykres liniowy ma sens wtedy, gdy na osi X znajduje się jakaś zmienna porządkowa. Służy do przedstawiania trendów w czasie, takich jak zmiany liczby przypadków określonej choroby w różnych okresach.

Pomimo swej prostoty zarówno wykresy punktowe, jak i liniowe mogą dostarczyć cennych informacji. Na przykład kształt „chmury" punktów na wykresie punktowym

może wskazywać na siłę i kierunek związku między zmiennymi, podczas gdy kształt linii na wykresie liniowym może ujawnić istotne trendy w danych.

Są niezwykle użytecznym narzędziem do prezentacji zależności pomiędzy dwiema zmiennymi, dostarczając jednocześnie bogatych informacji o strukturze danych. W medycynie i biostatystyce mogą być wykorzystywane do badania związku między różnymi cechami pacjentów, takimi jak wiek, masa ciała, ciśnienie krwi i wyniki testów laboratoryjnych.

Wykres rozrzutu przedstawia dane jako zbiór punktów, z których każdy reprezentuje obserwację. Położenie punktu na wykresie jest determinowane przez wartości dwóch zmiennych dla danej obserwacji (➡ **ryc. 3.3**). Jedna ze zmiennych jest zazwyczaj przedstawiana na osi poziomej (oś X), a druga na osi pionowej (oś Y).



Ryc. 3.3. Wykres rozrzutu opisujący relację pomiędzy cholesterolem całkowitym i HDL

Rozkład punktów na wykresie może wskazywać na różnego rodzaju związki między zmiennymi. Jeżeli np. punkty układają się wzdłuż linii odchylonej w górę, sugeruje to dodatnią korelację między zmiennymi. Jeżeli punkty układają się wzdłuż linii odchylonej w dół, sugeruje to korelację ujemną. Natomiast jeżeli punkty są rozrzucone losowo, prawdopodobnie nie ma związku między zmiennymi.

3.4.4. Histogramy

Są to wykresy słupkowe reprezentujące rozkład częstości wartości w danym zbiorze danych (w skrócie – rozkład wartości). Histogramy są w biostatystyce szczególnie użyteczne do zilustrowania rozkładu zmiennych ciągłych, takich jak wiek, ciśnienie krwi czy stężenie glukozy we krwi.

Tworzone są przez podział danych na określoną liczbę równych przedziałów, zwanych „koszykami” lub „binami”, a następnie zliczanie obserwacji w każdym przedziale, których liczbie odpowiada wysokość słupka.

Ważne jest zrozumienie, że wybór liczby przedziałów może wpłynąć na wygląd histogramu i interpretację danych. Z jednej strony zbyt duża liczba przedziałów może prowadzić do przeuczenia wykresu, gdy nadmiar szczegółów może przysłonić ogólne trendy. Z drugiej strony zbyt mała liczba przedziałów może zbytnio upraszczać dane i pomijać ważne detale.

W przypadku danych o dużym rozrzucie często używane są histogramy logarytmiczne, w których przedziały są nierównomiernie rozłożone, zgodnie ze skalą logarytmiczną. Taki histogram pozwala na lepsze zrozumienie rozkładu danych, gdy różnice między obserwacjami są bardzo duże.

3.4.5. Wykresy pudełkowe

Wykresy pudełkowe (ang. *box plots*), określane również jako wykresy „pudełko i wąsy", są popularnym narzędziem umożliwiającym przedstawienie rozkładu zmiennych oraz identyfikację potencjalnych wartości odstających. Są rekomendowane przy porównywaniu tendencji centralnych oraz rozkładów danych między różnymi grupami, co jest często wykorzystywane w badaniach medycznych i biostatystyce.

Klasyczny wykres pudełkowy składa się z „pudełka", które reprezentuje pierwszy (Q_1) i trzeci kwartyl (Q_3) danych, linii wewnątrz pudełka, która oznacza medianę (*Me*) – *box plot* można stosować także dla średniej, *SE* i 95% *CI*, co jest zalecane przy porównywaniu średnich – i „wąsów", które rozciągają się od Q_1 do minimalnej wartości danych nie niższej niż 1,5 × *IQR* (ang. *interquartile range*, czyli Q_3–Q_1) i od Q_3 do maksymalnej wartości danych nie wyższej niż 1,5 × *IQR*.

Wykresy pudełkowe są niezwykle użyteczne do identyfikacji wartości odstających, które są zaznaczane jako indywidualne punkty występujące poza „wąsami".

3.5. Wybór odpowiedniego wykresu

3.5.1. Dobre i złe praktyki wizualizacji danych

Odpowiednia wizualizacja danych stanowi istotny element analizy biostatystycznej, zapewniając zarówno zrozumienie, jak i przekazywanie skomplikowanych zestawów danych. Jednak nieodpowiednie techniki wizualizacji mogą prowadzić do błędnej interpretacji, co z kolei może wpłynąć na podejmowane decyzje kliniczne.

Dobre praktyki wizualizacji danych zalecają utrzymanie prostoty i klarowności. Wykresy powinny być łatwe do odczytania, bez nadmiaru ozdobników lub skomplikowanych wzorców, które mogą odciągnąć uwagę od prezentowanych danych. Kolorów należy używać z rozwagą, aby uniknąć przekazywania mylących informacji.

Nieodpowiednie praktyki obejmują nadużycie kolorów, nadmierne wykorzystanie technik 3D czy skomplikowanych wzorców, które mogą wprowadzić zamieszanie lub błędny odbiór. Nieprawidłowo dobrana skala na osiach wykresu może prowadzić do nieinformacyjnej prezentacji różnic między grupami.

Ważne jest również, aby zawsze uwzględniać kontekst danych, tak aby wizualizacja była adekwatna do zawartych danych i do odbiorców, którym są one prezentowane. Dla przykładu używanie skomplikowanych wykresów może być niewłaściwe, jeśli są one przeznaczone dla osób bez podstawowej wiedzy statystycznej.

3.5.2. Wykorzystanie wykresów w biostatystyce

Wykresy słupkowe i kolumnowe są doskonałym narzędziem do ilustrowania porównań między różnymi grupami. Można za ich pomocą np. przedstawić porównanie wskaźników śmiertelności w różnych chorobach, odsetka pacjentów, którzy doświadczyli określonego efektu ubocznego leczenia, lub liczby pacjentów reagujących na określone leczenie.

Wykresów kołowych można używać do zilustrowania proporcji, np. w celu pokazania procentowego podziału pacjentów wg rodzaju diagnozy, procentowego udziału różnych typów leków przepisywanych w szpitalu czy rozkładu płci w badanej populacji. Wykresy te są również przydatne do porównania udziału różnych typów badań laboratoryjnych przeprowadzanych w szpitalu.

Wykresy punktowe i liniowe wykorzystuje się do ilustracji trendów. Mogą pokazać np. zmiany średniego stężenia cholesterolu w przedziale czasowym u pacjentów. Inne zastosowania to obrazowanie zmian ciśnienia krwi w ciągu dnia, trendów w zmianach masy ciała pacjentów podczas leczenia lub ilustracja wzrostu liczby przypadków chorób zakaźnych w określonym czasie.

Histogramy są użyteczne do prezentacji rozkładu danych, np. rozkładu wieku w populacji pacjentów, rozkładu czasu trwania leczenia, rozkładu wartości wyników testów laboratoryjnych czy liczby przyjęć do szpitala w określonym przedziale czasowym.

Wykresy pudełkowe są szczególnie użyteczne do przedstawiania rozkładu danych oraz do identyfikacji wartości odstających. Można np. przy ich użyciu zilustrować rozkład wartości ciśnienia krwi w grupie pacjentów, rozkład stężenia glukozy we krwi w różnych grupach pacjentów czy różnic w czasie trwania leczenia pomiędzy różnymi grupami.

Wykresy rozrzutu są idealne do przedstawiania związków między dwiema zmiennymi. Mogą np. pokazać związek między wiekiem a ciśnieniem krwi pacjentów, między masą ciała a stężeniem cholesterolu, między dawką leku a efektem terapeutycznym czy między czasem trwania leczenia a jego wynikiem.

Wszystkie wyżej wymienione wykresy stanowią kluczowe narzędzia wizualizacji danych w biostatystyce, a ich właściwe zrozumienie i interpretacja są niezbędne do poprawnego wykorzystania statystyki w badaniach medycznych.

3.6. Tworzenie wykresów za pomocą Excela

Dostępne są różne narzędzia, służące do tworzenia wykresów na podstawie danych statystycznych.

Excel (aplikacja pakietu Microsoft Office) oferuje szeroką gamę funkcji do analizy danych, w tym możliwość tworzenia różnych typów wykresów. Jest łatwy w użyciu i często wykorzystywany do prostych analiz i wizualizacji danych. Jest jednym z najpopularniejszych i najpowszechniej stosowanych narzędzi do analizy i prezentacji danych statystycznych oraz do zarządzania nimi. Za pomocą Excela można szybko i efektywnie tworzyć różnorodne wykresy, które umożliwiają wizualizację zależności pomiędzy danymi.

Podstawą utworzenia każdego wykresu jest zestaw danych. Do wygenerowania wykresu w Excelu potrzebujemy co najmniej jednej kolumny danych. Na podstawie takiego zbioru wartości można utworzyć np. histogram. Tworzenie skutecznego wykresu, który będzie właściwie ilustrował wyniki analizy, wymaga wyboru odpowiednich danych. Excel umożliwia również personalizację wykresów, co pozwala na dostosowanie ich do konkretnych potrzeb badacza.

3.6.1. Uruchamianie Excela i otwarcie nowego arkusza

Uruchamianie Excela może różnić się w zależności od wersji programu oraz systemu operacyjnego, dlatego poniżej przedstawiono dokładny opis, jak uruchomić program i utworzyć nowy arkusz.

3.6.1.1. Uruchamianie Excela

W systemie Windows kliknij przycisk Start (domyślnie w lewym dolnym rogu ekranu). W polu wyszukiwania menu Start (lub wciśnij na klawiaturze klawisz Windows) wpisz „Excel”. W wynikach wyszukiwania kliknij ikonę programu Excel. W zależności od konfiguracji komputera ikona skrótu do programu może się także znajdować na pulpicie lub pasku zadań.

3.6.1.2. Utworzenie nowego arkusza

Po uruchomieniu programu zwykle pojawia się ekran startowy. Tutaj można wybrać spośród kilku opcji, takich jak Pusty skoroszyt, Szablony itp. Kliknięcie w Pusty arkusz otworzy nowy, czysty arkusz kalkulacyjny.

3.6.1.3. Zapoznanie z interfejsem

Po otwarciu nowego arkusza zobaczysz interfejs Excela. Ważne elementy, na które warto zwrócić uwagę, to:

- pasek narzędziowy (wstążka) na górze ekranu, który zawiera różne zakładki (karty) takie jak Plik, Narzędzia główne, Wstaw itp.;
- kolumny i wiersze – arkusz jest podzielony na kolumny oznaczone literami (A, B, C itp.) i wiersze oznaczone numerami (1, 2, 3 itp.);
- komórki: miejsce – w którym przecinają się kolumny i wiersze, nazywane jest komórką; każda komórka ma swój unikalny adres, na przykład A1, B2 itp.

3.6.1.4. Wprowadzanie danych

To kluczowy etap przed tworzeniem wykresów w Excelu. Aby to zrobić poprawnie, śledź poniższe kroki.

Utworzenie nagłówków

Zdecyduj, jakie informacje chcesz przedstawić na wykresie. W pierwszym rzędzie wprowadź odpowiednie nagłówki dla twoich danych.

Wprowadzenie danych

W kolejnych komórkach pod nagłówkami wprowadź odpowiednie dane.

Formatowanie danych

Jeśli wprowadzasz liczby, możesz je sformatować, aby wyglądały lepiej na arkuszu. Zaznacz odpowiednie komórki, kliknij prawym przyciskiem myszy, wybierz Formatuj komórki, a następnie w otwartym oknie wybierz kategorię i odpowiednie formatowanie.

Sprawdzanie błędów

Excel ma wbudowaną funkcję sprawdzania błędów, która podkreśla potencjalne problemy we wprowadzonych danych; upewnij się, że dokładnie sprawdziłeś wprowadzone informacje i poprawiłeś ewentualne błędy. Błędy mogą być zaznaczone małymi trójkątnymi ikonami w górnym rogu komórki. Umieść kursor myszki, aby zobaczyć informacje o potencjalnym problemie.

Zapisywanie danych

Regularnie zapisuj swoją pracę, aby uniknąć przypadkowej utraty danych. Kliknij Plik w lewym górnym rogu ekranu, a następnie Zapisz. Wybierz lokalizację na dysku, nadaj nazwę plikowi i kliknij Zapisz.

Po wprowadzeniu i sprawdzeniu wszystkich danych jesteś gotowy, by przejść do kolejnego kroku i stworzenia wykresu na podstawie zebranych informacji.

3.6.1.5. Zaznaczanie danych do wykresu

Aby stworzyć wykres w Excelu, trzeba zaznaczyć odpowiednie dane, które mają być na nim przedstawione.

Zaznaczenie pojedynczej komórki

Kliknij lewym przyciskiem myszy na komórkę, którą chcesz zaznaczyć. Po tym kroku komórka staje się aktywna.

Zaznaczenie zakresu komórek

Kliknij na pierwszą komórkę zakresu. Przytrzymując lewy przycisk myszy, przeciągnij kursor do ostatniej komórki zakresu. Po zwolnieniu przycisku myszy cały zakres zostanie zaznaczony. Możesz również zaznaczyć pierwszą komórkę zakresu, a następnie, przytrzymując klawisz Shift, kliknij na ostatnią komórkę zakresu.

Zaznaczanie nieciągłego zakresu komórek

Jeśli chcesz zaznaczyć komórki, które nie sąsiadują ze sobą, kliknij na pierwszą komórkę, a następnie kolejne, przytrzymując klawisz Ctrl.

Zaznaczanie całych kolumn lub wierszy

Aby zaznaczyć całą kolumnę, kliknij na literę kolumny (np. A dla kolumny A) na górze arkusza. Aby zaznaczyć cały wiersz, kliknij na numer wiersza (np. 1 dla wiersza 1) po lewej stronie arkusza.

Zaznaczanie wielu kolumn lub wierszy

Kliknij na pierwszą kolumnę/wiersz, a następnie, przytrzymując lewy przycisk myszy, przeciągnij kursor do ostatniej kolumny/wiersza, które chcesz zaznaczyć.

Zaznaczanie wszystkich danych na arkuszu

Kliknij na mały kwadratowy przycisk w lewym górnym rogu arkusza, między literami kolumn i numerami wierszy, lub użyj skrótu klawiszowego Ctrl + A.

Wskazówka

Aby twoje dane były poprawnie reprezentowane na wykresie, należy zaznaczyć zarówno kolumny z etykietami, jak i wartościami. Na przykład jeśli masz miesiące w kolumnie A i wartości sprzedaży w kolumnie B, zaznacz obie kolumny, aby na wykresie były wyświetlane zarówno etykiety miesięcy, jak i wartości sprzedaży.

Po prawidłowym zaznaczeniu danych możesz przejść do kolejnego kroku, którym jest wybór i wstawienie odpowiedniego typu wykresu.

3.6.1.6. Wybór typu wykresu

Po zaznaczeniu odpowiednich danych w arkuszu Excela, musisz wybrać odpowiedni typ wykresu, który najlepiej przedstawi twoje informacje. Poniżej szczegółowy opis tego procesu.

Otwarcie palety wykresów

Przejdź do karty Wstawianie w menu Excela. W sekcji Wykresy zobaczysz różne ikony reprezentujące dostępne typy wykresów.

Przeglądanie dostępnych typów wykresów

Kliknij strzałkę w prawym dolnym rogu sekcji Wykresy, aby rozwinąć pełną listę dostępnych typów wykresów.

Zastanów się, jakimi danymi dysponujesz i jaki wykres najlepiej je przedstawi. Kliknij na wybrany typ, a następnie wybierz konkretny styl wykresu spośród dostępnych opcji.

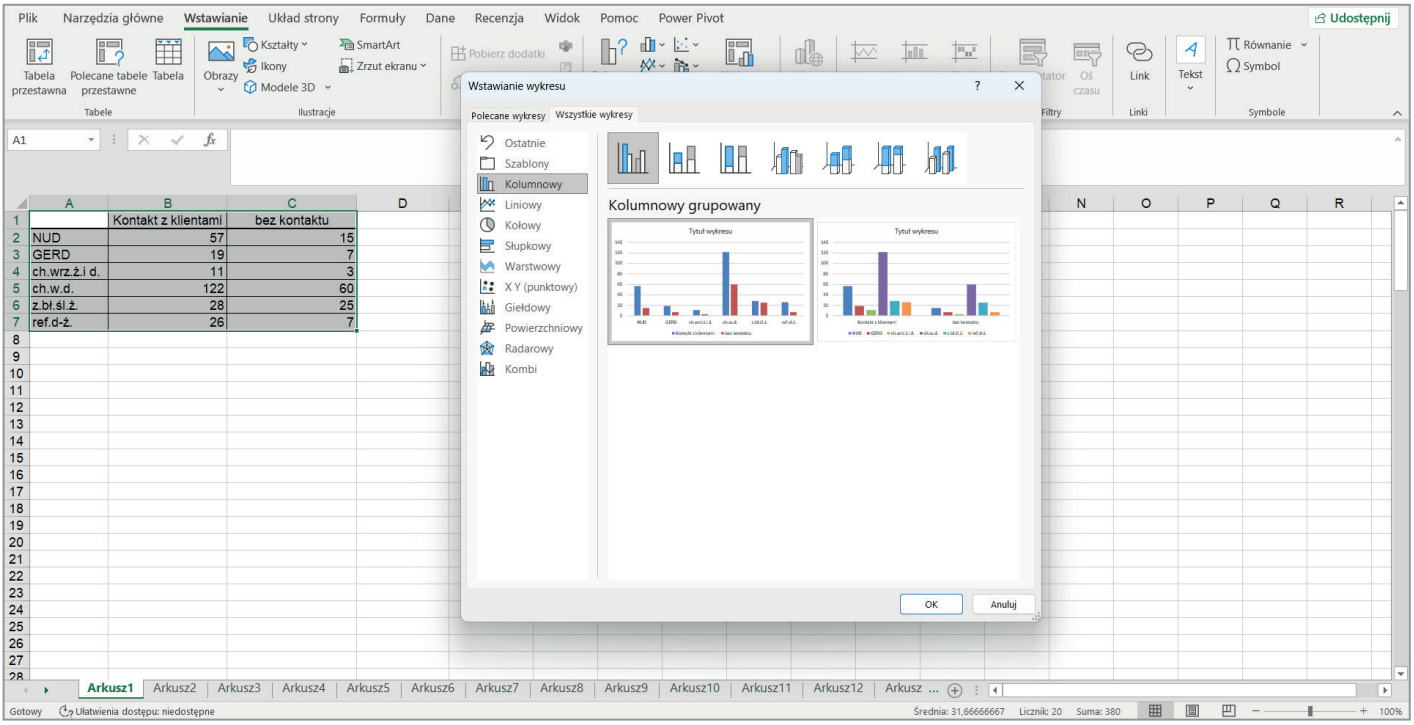
Podgląd wykresu

W wielu wersjach Excela najechanie kursorem na dany typ wykresu pokaże jego podgląd przed wstawieniem. Pozwala to na szybkie porównanie różnych opcji i wybór najlepszej dla prezentowanych danych.

Wstawianie wykresu

Kliknij na odpowiedni typ i styl wykresu (➥ **ryc. 3.4**). Wykres zostanie automatycznie wstawiony do arkusza z zaznaczonymi danymi. Początkowo może być umieszczony w nieodpowiednim miejscu lub mieć nieodpowiedni rozmiar, ale łatwo można go przeciągnąć i dostosować.

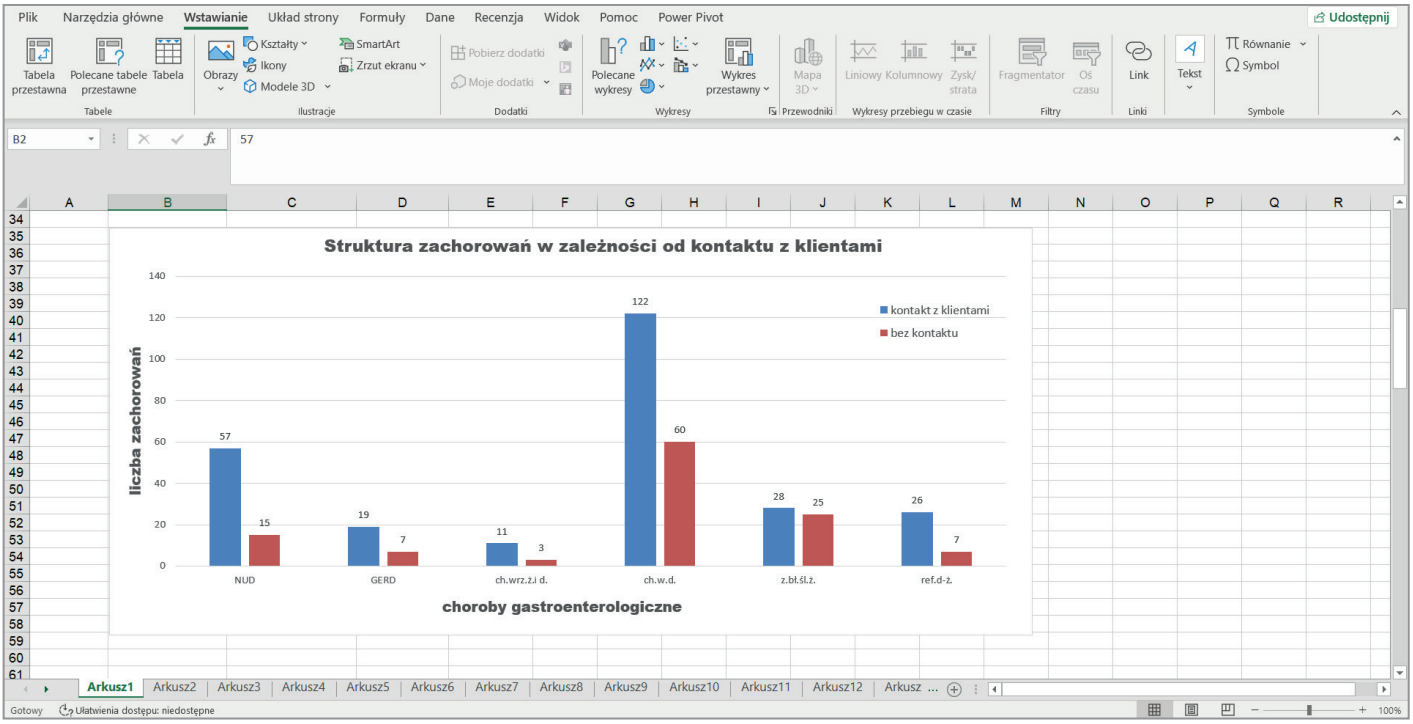
Ryc. 3.4. Wybór typu wykresu kolumnowego.



Wskazówka

Jeśli nie jesteś pewien, jaki typ wykresu wybrać, skorzystaj z opcji Polecane wykresy, jeśli jest dostępna. Excel analizuje dane i sugeruje kilka typów, które mogą być odpowiednie dla zestawu danych w arkuszu (➥ **ryc. 3.5**).

Ryc. 3.5. Gotowy wykres wstawiony do arkusza

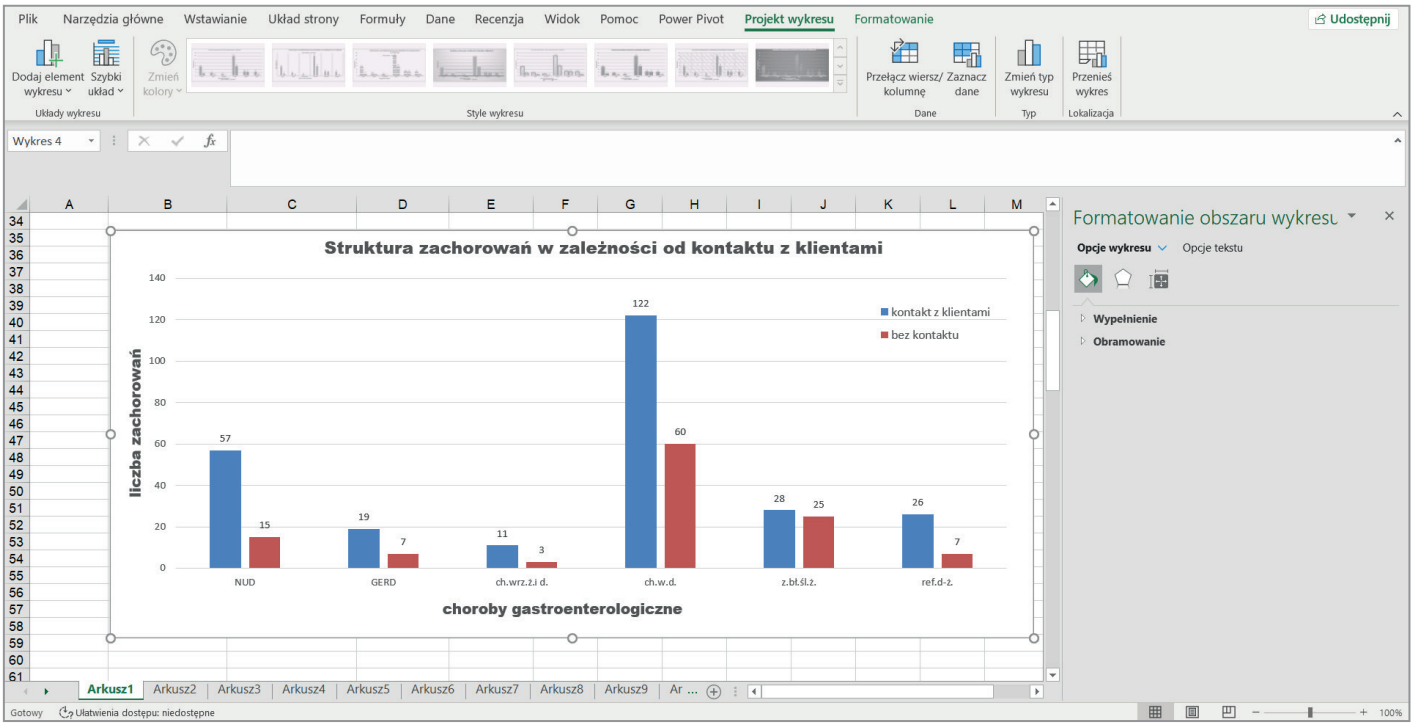


3.6.1.7. Dostosowywanie wykresu

Po wstawieniu wykresu do arkusza Excela można dostosować jego wygląd i funkcje, tak aby lepiej prezentował dane oraz był bardziej czytelny i estetyczny. Poniżej dokładny opis procesu dostosowania wykresu w Excelu.

Zaznaczanie wykresu

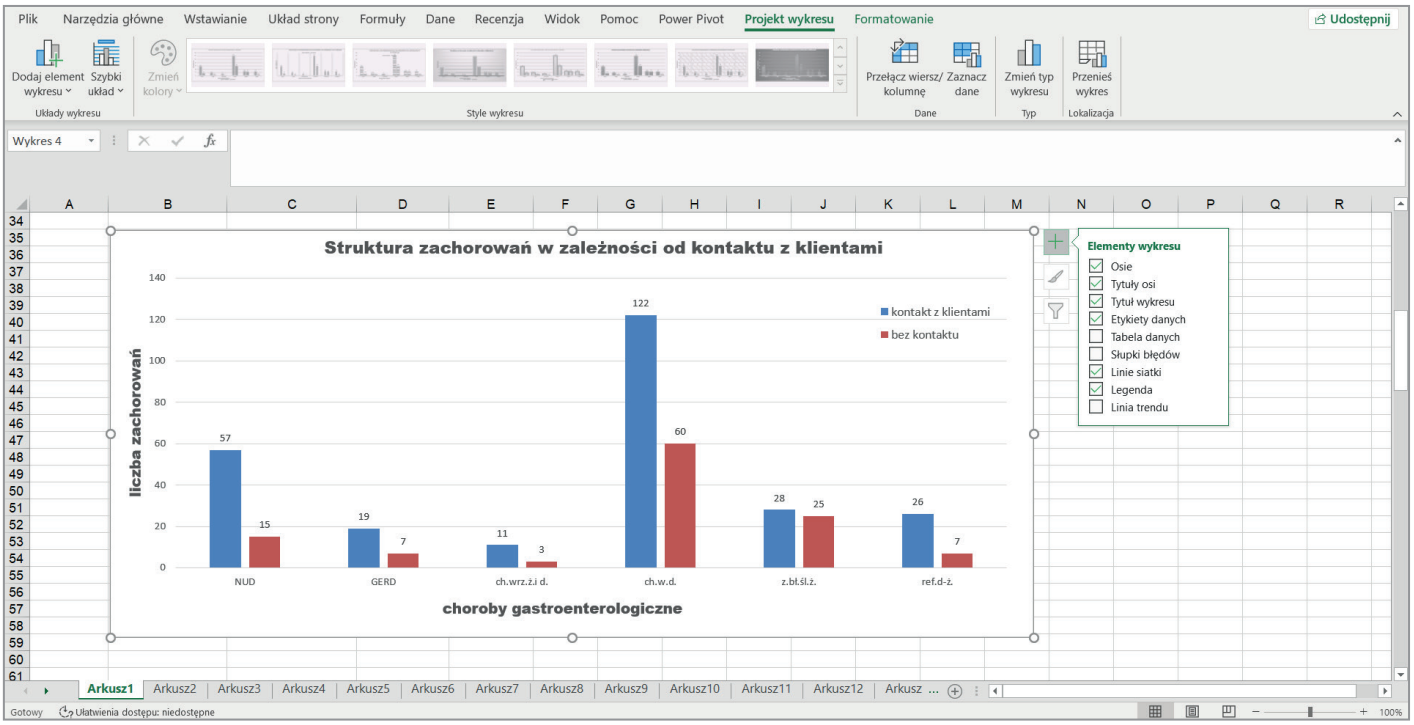
Aby dokonać jakichkolwiek zmian w wykresie, trzeba go najpierw zaznaczyć. W tym celu kliknij na wykres. Po zaznaczeniu na wstążce pojawią się dwie karty narzędzi: Projekt wykresu i Formatowanie (➥ **ryc. 3.6**).



Ryc. 3.6. Wynik kliknięcia obszaru wykresu prawym przyciskiem myszy i wyboru polecenia Formatuj obszar wykresu z menu podręcznego

Dodawanie tytułu wykresu

W zakładce Projekt wykresu kliknij polecenie Dodaj element wykresu. Wybierz Tytuł wykresu i preferowane położenie (np. „Góra wykresu"). Kliknij dwukrotnie na dodany tytuł, aby wpisać własny tekst (➥ **ryc. 3.7**).



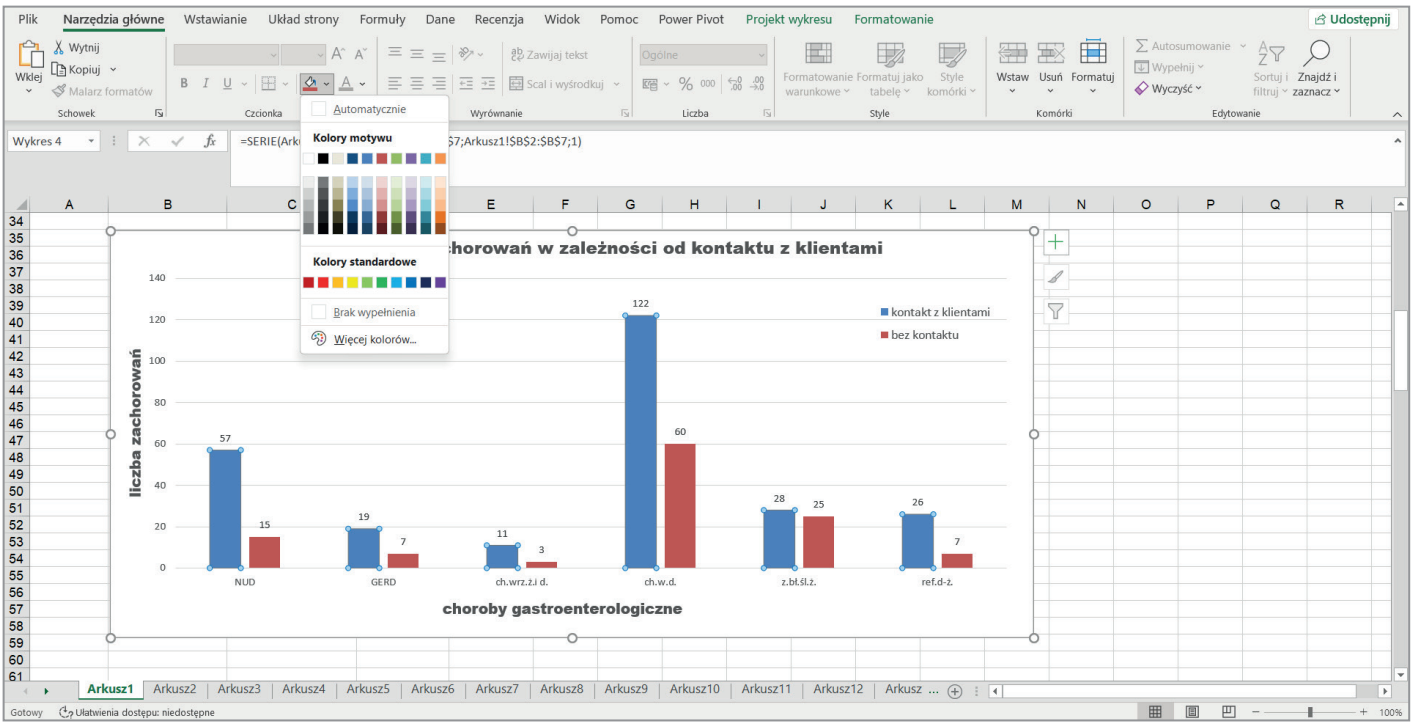
Ryc. 3.7. Po kliknięciu ikony + również można dodawać, edytować i usuwać elementy wykresu

Dodawanie tytułów osi

Podobnie jak w przypadku tytułu wykresu użyj opcji Dodaj element wykresu i wybierz Tytuły osi. Kliknij dwukrotnie dodane tytuły, aby edytować ich treść.

Zmiana kolorów i stylu serii danych

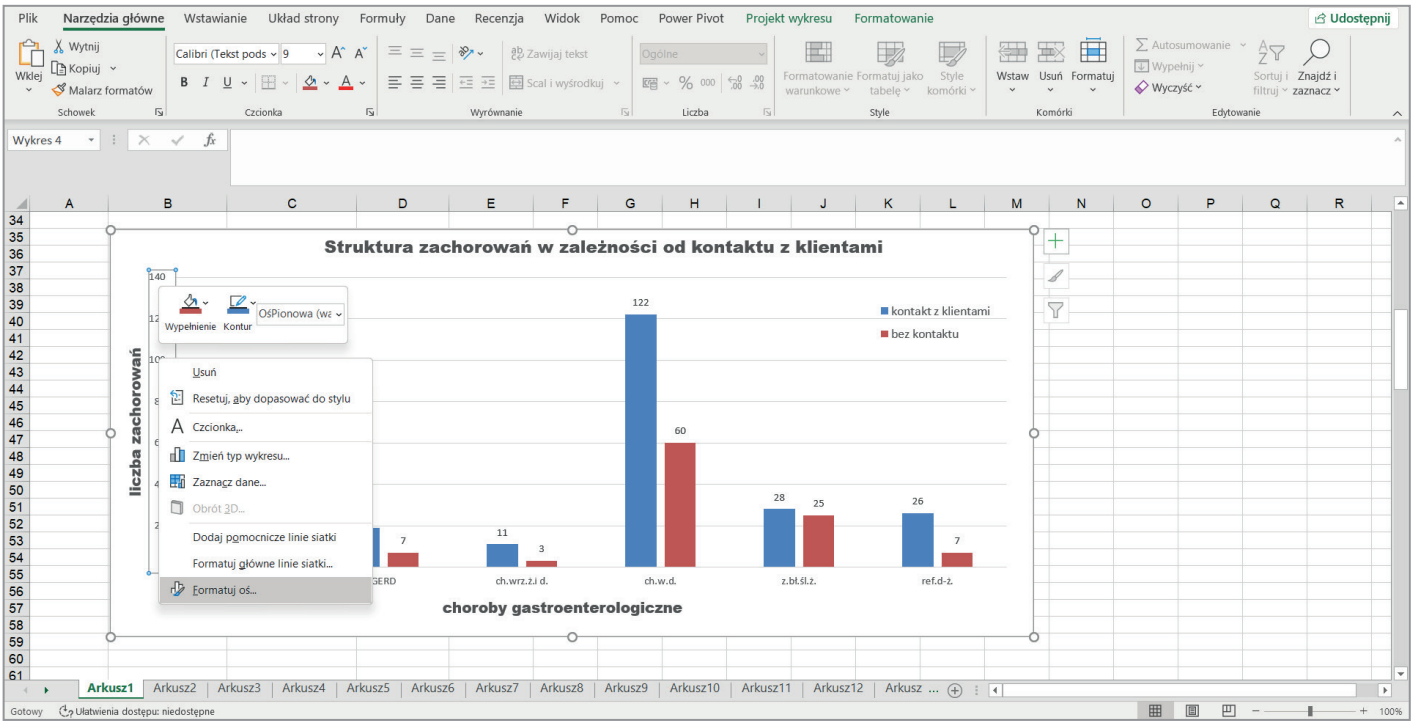
Wybierz kliknięciem konkretną serię danych na wykresie (np. jedną z kolumn wykresu kolumnowego), aby ją zaznaczyć. Następnie kliknij prawym przyciskiem myszy i wybierz Formatuj serię danych. W otwartym panelu możesz zmienić kolor, styl linii, wypełnienie i inne atrybuty wybranej serii (➥ **ryc. 3.8**).



Ryc. 3.8. Style wykresu lub serii danych można również zmienić poleceniami z karty Narzedzia główne

Dostosowywanie osi wykresu

Zaznacz którąś z osi wykresu i kliknij prawym przyciskiem myszy. Wybierz Formatuj oś. W panelu formatowania możesz dostosować zakres osi, interwał pomiędzy wartościami, kierunek wartości i inne opcje (➥ **ryc. 3.9**).

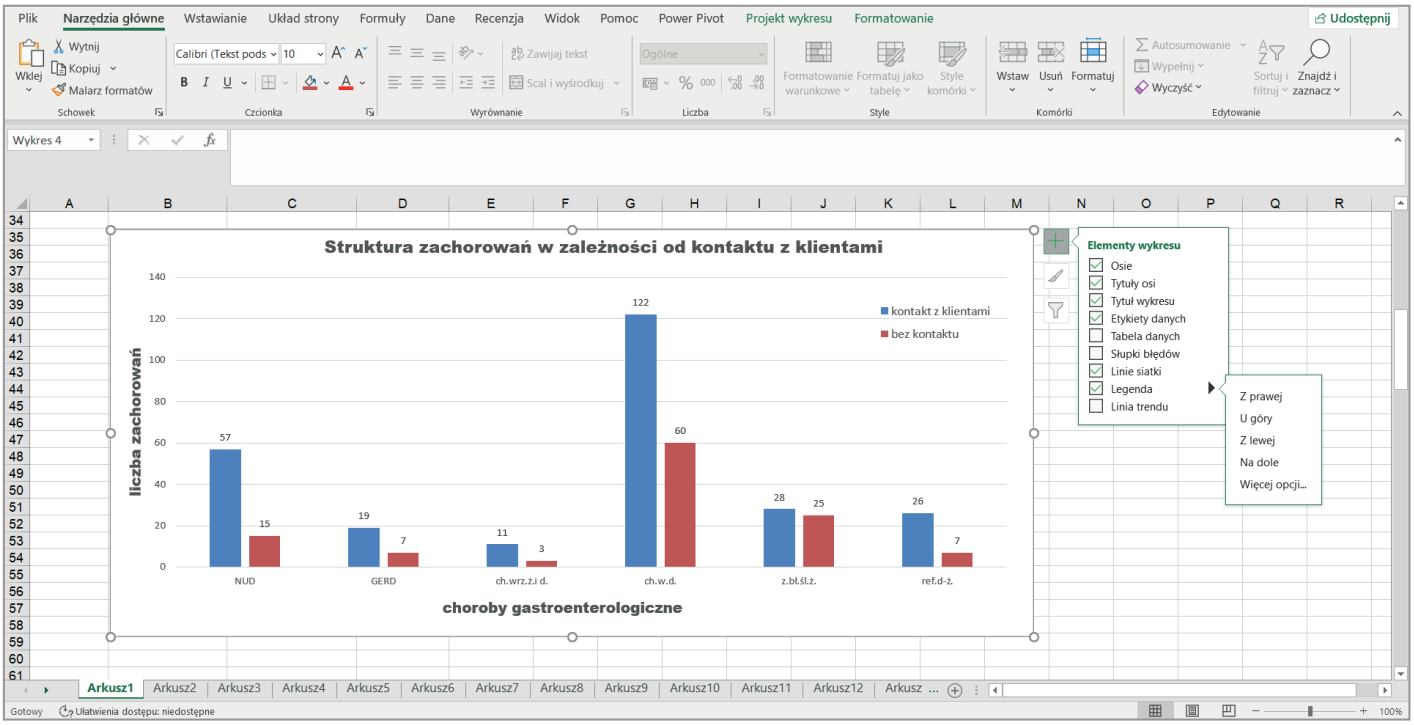


Ryc. 3.9. Po kliknięciu prawym klawiszem na któryś z elementów wykresu pojawi się menu podręczne, za pomocą którego można edytować wybrany element

Ryc. 3.10. Określanie miejsca dodania legendy za pomocą ikony + (Elementy wykresu)

Dodawanie legendy

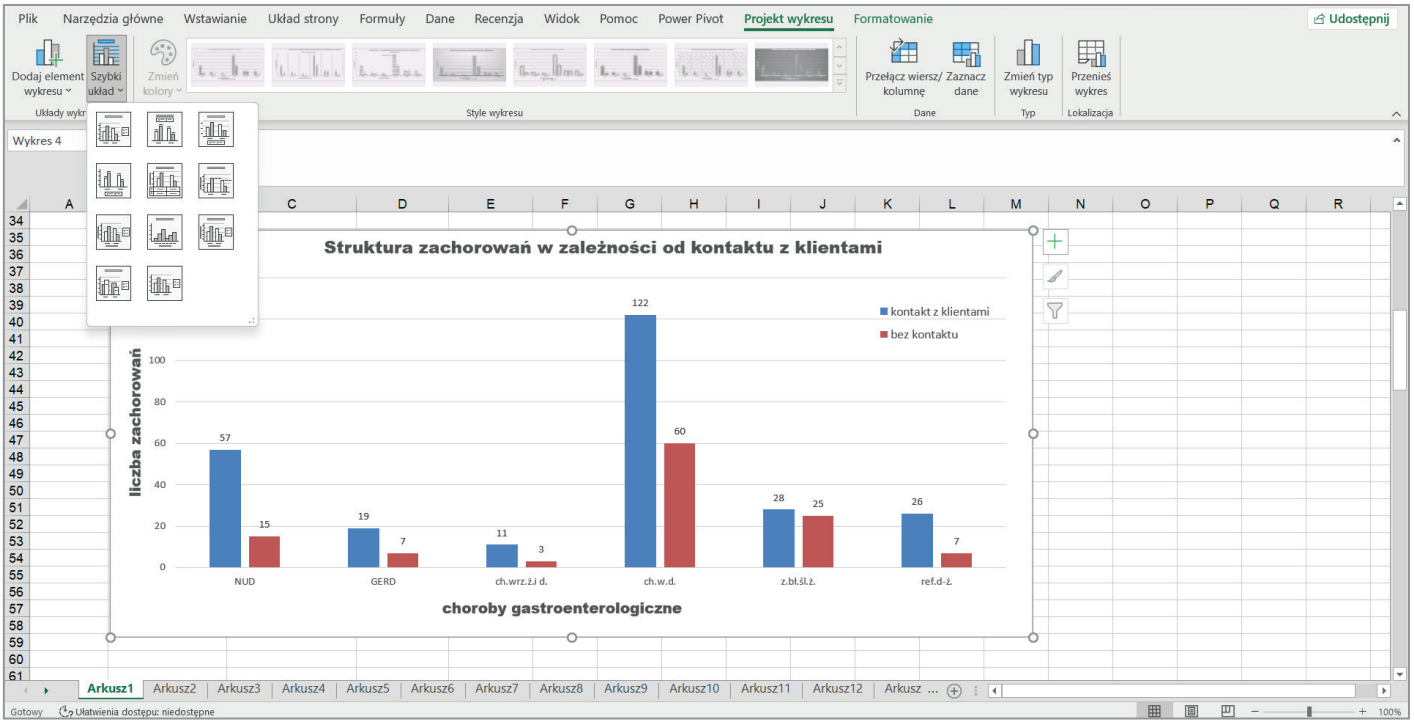
Użyj polecenia Dodaj element wykresu i wybierz Legenda. Możesz określić, gdzie legenda ma być umieszczona (np. po prawej stronie wykresu, ➥ **ryc. 3.10**).



Zmiana stylu i układu wykresu

W karcie Projekt wykresu znajdziesz grupy Style wykresu i Układy wykresu. Przeglądaj dostępne opcje i wybierz tę, która jest najbardziej odpowiednia. Za pomocą dostępnych w tych grupach narzędzi można szybko zmieniać wygląd wykresu lub dokładniej dostosowywać jego elementy (➥ **ryc. 3.11**).

Ryc. 3.11. Polecenie Szybki układ z grupy Układy wykresu



Przesuwanie i zmiana rozmiaru wykresu

Aby przesunąć wykres, umieść kursor na jego krawędzi (powinien się zmienić w krzyżujące się strzałki) i przeciągnij wykres w żądane miejsce. Aby zmienić jego rozmiar, kliknij jeden z uchwytów (małych okręgów) na krawędzi wykresu i przeciągnij, aby dostosować rozmiar (przesunięcie uchwytów na rogach skaluje wykres, zachowując proporcje, a przesunięcie uchwytów bocznych zmienia tylko wysokość lub szerokość).

Dostosowując wykres, pamiętaj, że celem jest uczynienie go jak najbardziej czytelnym i reprezentatywnym dla twoich danych. Często prostota i jasność są kluczem do skutecznego przedstawienia informacji.

3.6.1.8. Zapisywanie i udostępnianie

Zapisywanie i udostępnianie wykresu to ważne etapy, które pozwalają zachować twoją pracę oraz dzielić się nią z innymi. Poniżej znajdziesz dokładne instrukcje dotyczące tych czynności.

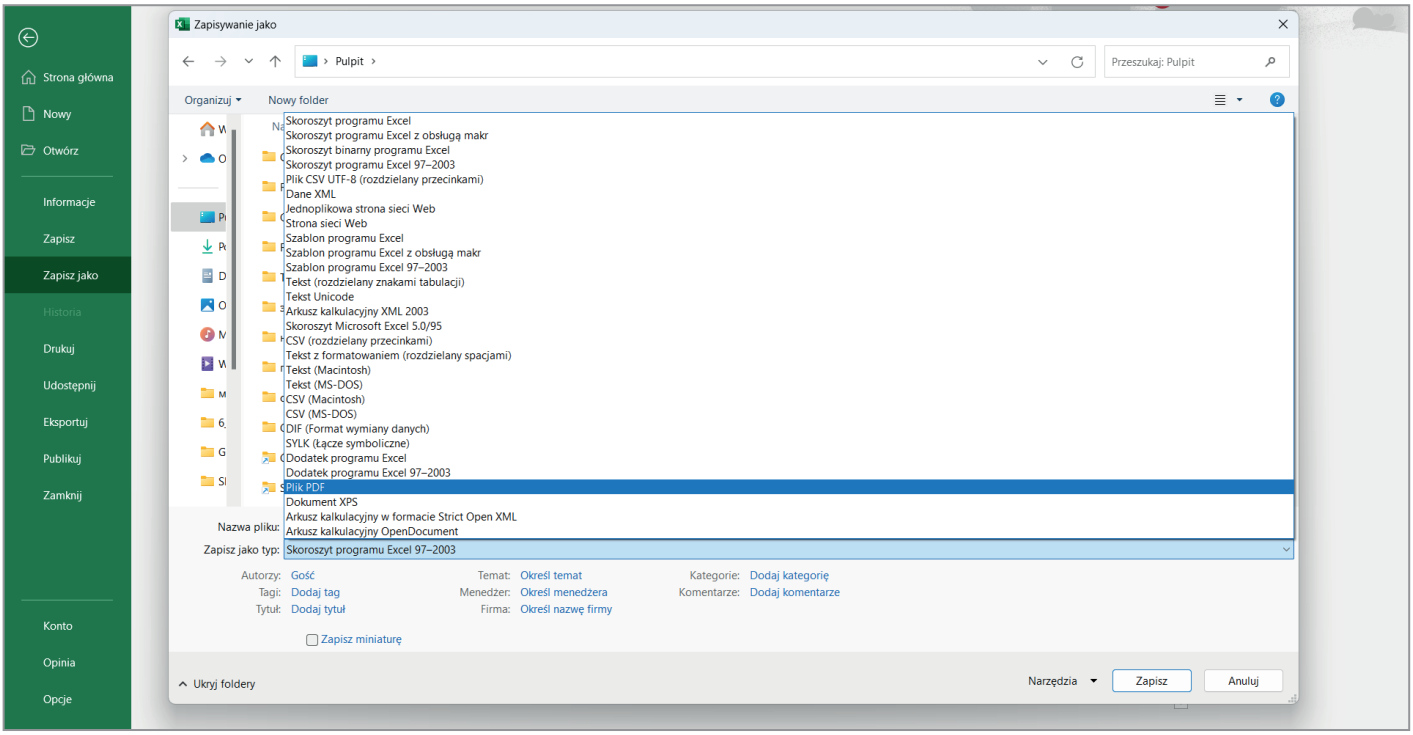
Zapisywanie arkusza z wykresem

Kliknij kartę Plik w lewym górnym rogu programu Excel. Wybierz Zapisz lub Zapisz jako, jeśli chcesz wybrać konkretną lokalizację lub zmienić format pliku. Jeśli zapisujesz plik po raz pierwszy, Excel poprosi o nadanie nazwy pliku oraz wybór lokalizacji zapisu.

Formaty plików

Excel oferuje różne formaty zapisu. Najczęściej używany to „xlsx", ale dostępne są też inne, takie jak „xls" (dla starszych wersji Excela), „csv" (dane oddzielone przecinkami) czy „pdf" (dokument w formacie PDF). Wybierz odpowiedni format w zależności od potrzeb (➡ **ryc. 3.12**).

Ryc. 3.12. Wybór formatu zapisywanego pliku



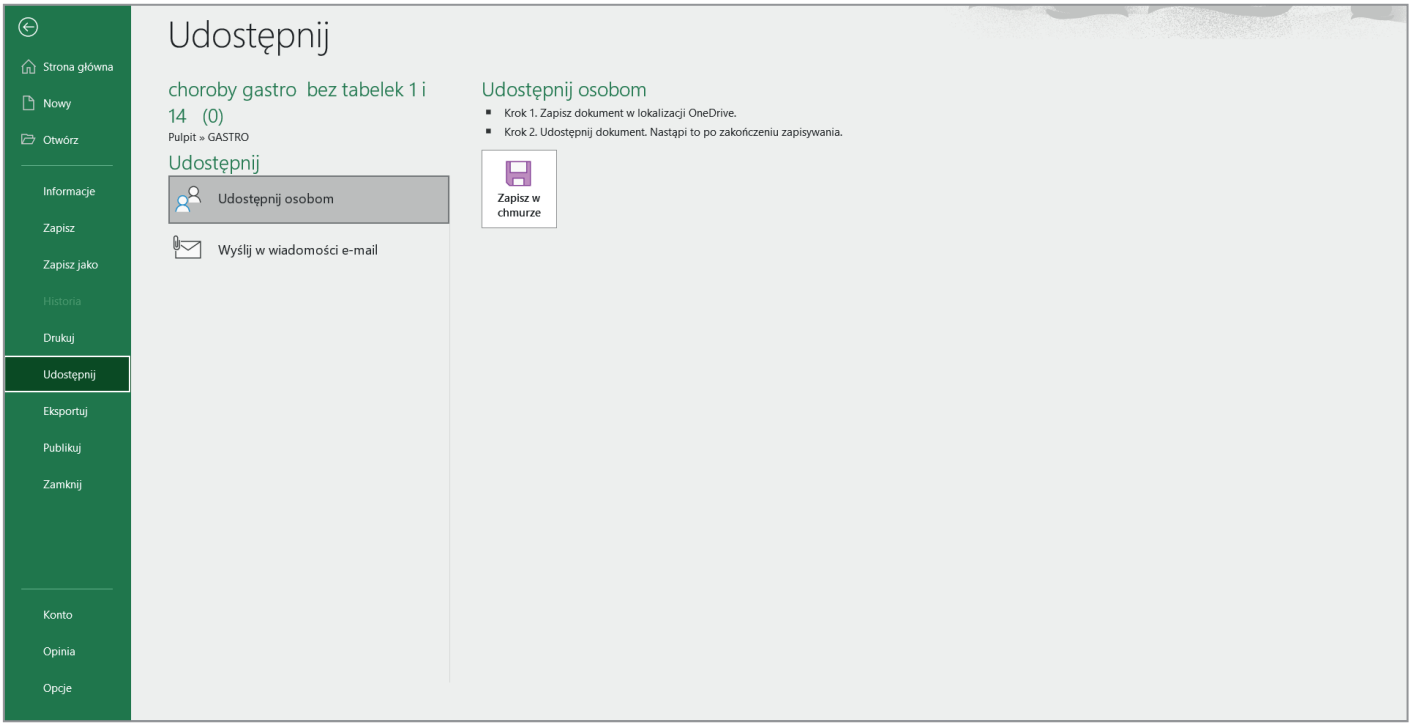
Automatyczne zapisywanie

Jeśli korzystasz z Excela wchodzącego w skład pakietu Microsoft 365, to praca może być automatycznie zapisywana w chmurze w czasie rzeczywistym. Upewnij się, że jesteś zalogowany na swoje konto Microsoft i masz dostęp do Internetu.

Udostępnianie wykresu w Excelu

Z karty Plik wybierz polecenie Udostępnij. Możesz zaprosić innych do edycji arkusza e-mailem lub uzyskać link do udostępnienia. Możesz też zdecydować, czy osoby, którym udostępniasz arkusz, mogą tylko przeglądać dane, czy też edytować je (➥ **ryc. 3.13**).

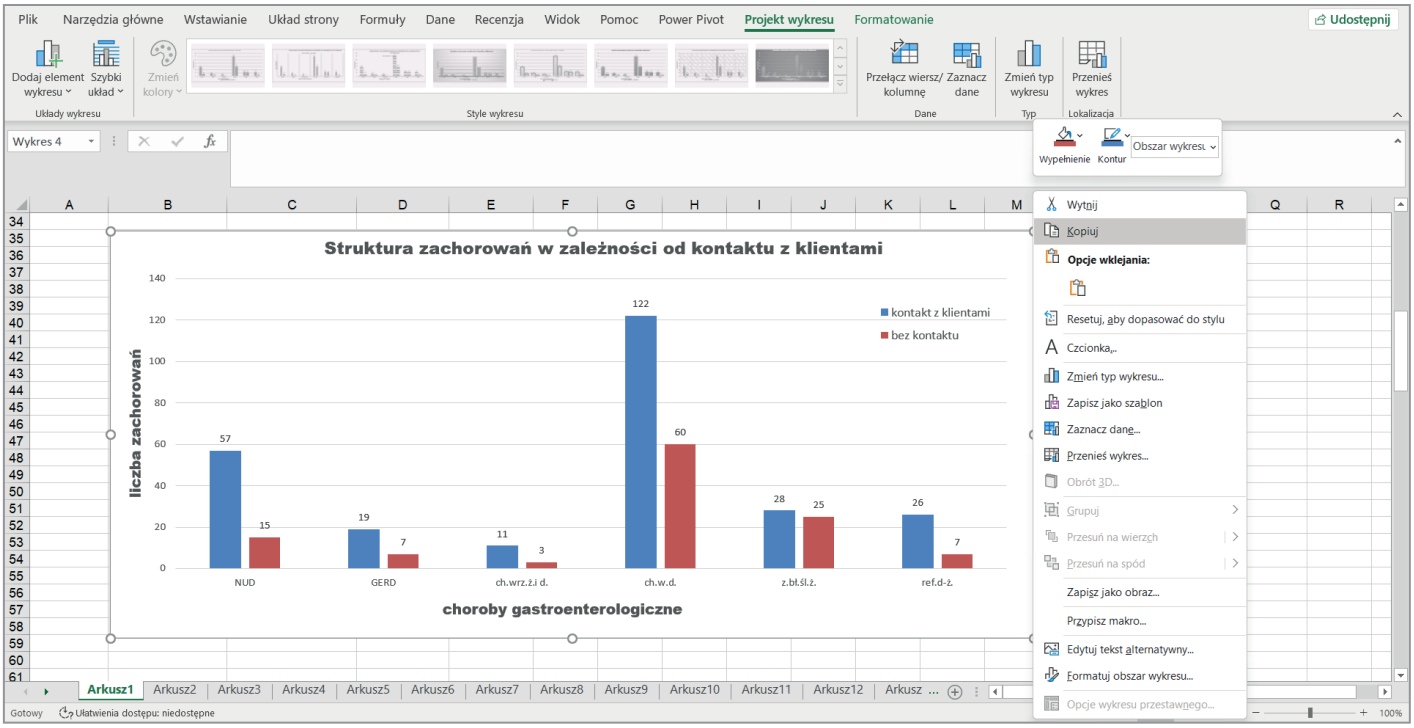
Ryc. 3.13. Okno udostępniania



Kopiowanie wykresu do innych programów

Aby skopiować wykres do innych programów, takich jak Word czy PowerPoint, kliknij wykres prawym przyciskiem myszy i wybierz Kopiuj. Następnie przejdź do programu, do którego chcesz wkleić wykres, i użyj polecenia Wklej (➥ **ryc. 3.14**).

Ryc. 3.14. Menu podręczne pojawiające się po kliknięciu wykresu prawym klawiszem myszy



Utworzony wykres ostatecznie powinien być odpowiednio zinterpretowany. Wykresy pozwalają na łatwe zobrazowanie trendów, wzorców i anomalii w danych, które mogą nie być oczywiste podczas analizy tabel. Należy jednak pamiętać, że wykresy, nawet te najbardziej zaawansowane, są tylko narzędziami do prezentacji danych, których prawdziwe zrozumienie wymaga dogłębnego zrozumienia statystyki.

Praca z programem Excel jest intuicyjna i daje dużo możliwości. Utworzone wykresy mogą posłużyć jako pomocniczy materiał do dalszej analizy statystycznej, jak również jako źródło do publikacji i prezentacji naukowych. Należy jednak pamiętać, że każde narzędzie, nawet tak proste jak Excel, wymaga odpowiedniego szkolenia i praktyki. Po poznaniu podstawowych funkcji i możliwości Excel może stać się bardzo użytecznym narzędziem dla każdego badacza.

3.7. Interpretacja wykresów

3.7.1. Jak czytać i interpretować różne rodzaje wykresów

Zrozumienie sposobów interpretacji różnych typów wykresów jest niezwykle istotne dla skutecznej analizy danych. W niniejszym podrozdziale zostanie zaprezentowana metodyka interpretacji głównych rodzajów wykresów.

Czytanie i interpretacja wykresów są z kolei bardzo istotne dla zrozumienia ukrytych w danych wzorców i trendów. Odpowiednie interpretowanie danych może np. prowadzić do opracowania nowych terapii i metod leczenia.

Te rozmaite metody wizualizacji danych umożliwiają lepsze zrozumienie, interpretację i prezentację wyników naukowcom. Należy jednak pamiętać, że każdy wykres jest tylko tak dobry, jak jego dane, i trzeba krytycznie podchodzić do źródła danych i metod ich prezentacji, aby uniknąć formułowania błędnych wniosków.

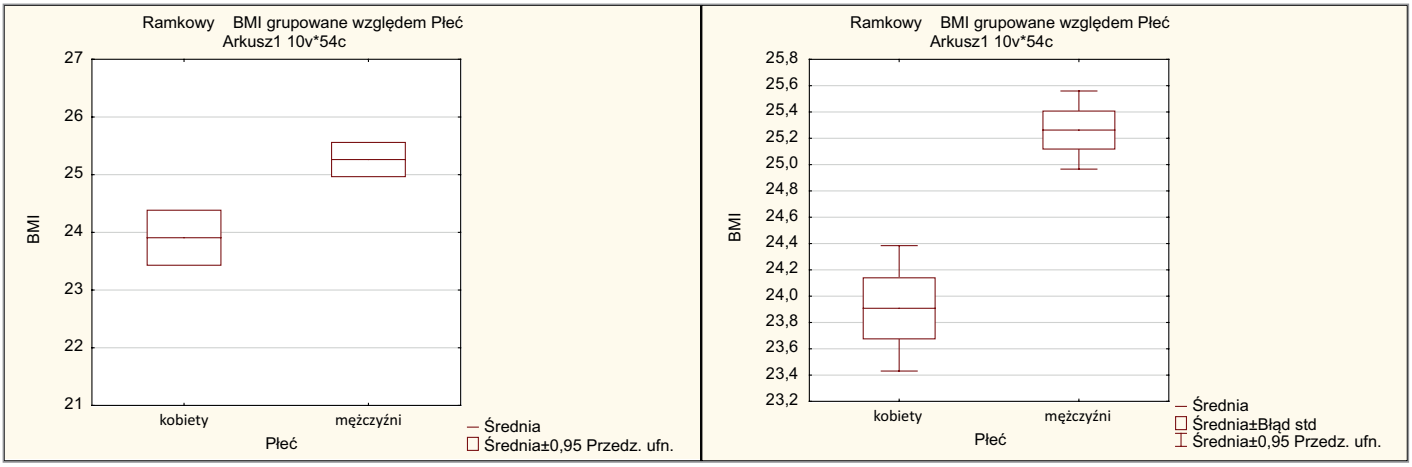
3.7.2. Powszechne pułapki i błędy przy interpretacji wykresów

Mimo że wykresy są bardzo dobrym narzędziem wizualizacji danych, niewłaściwe użycie i interpretacja mogą prowadzić do niepożądanych wniosków. Poniżej przedstawiono najczęstsze błędy i pułapki, które można napotkać przy interpretacji wykresów.

3.7.2.1. Nieodpowiednie skalowanie osi

Najbardziej powszechnym błędem jest nieodpowiednie skalowanie osi, które może zniekształcić reprezentację danych i doprowadzić do błędnej interpretacji. Oś Y powinna zawsze zaczynać się od zera, gdy przedstawiane są liczności, w przeciwnym razie różnice między wartościami mogą wydawać się większe lub mniejsze, niż są w rzeczywistości. Wyjątkowo, np. porównując wartości średniej z ich 95% przedziałem ufności, można zrezygnować z pokazywania całej skali na osi Y, ponieważ w tym przypadku ważniejsze jest stwierdzenie faktu występowania danych w określonym przedziale ufności (➥ **ryc. 3.15**).

Ryc. 3.15. Porównanie dwóch wykresów uwzględniających średnie wartości BMI u kobiet i mężczyzn (z lewej strony nieprawidłowo, z prawej prawidłowo)



3.7.2.2. Wybór nieodpowiedniego typu wykresu

Wybór nieodpowiedniego typu wykresu również może prowadzić do błędnej interpretacji. Powinien on być dopasowany do rodzaju i charakteru danych. Przykładem mogą być wykresy kołowe – często niewłaściwe do prezentowania danych, ponieważ ludzki wzrok ma trudności z porównywaniem kątów i powierzchni.

3.7.2.3. Pominięcie istotnych danych statystycznych

Pominięcie istotnych danych statystycznych, takich jak odchylenie standardowe, błąd standardowy czy przedział ufności, może prowadzić do błędnej interpretacji wykresu. Są one niezbędne do oceny zmienności danych i wiarygodności różnic między grupami.

3.7.2.4. Błędna interpretacja korelacji jako przyczynowości

Częstym błędem jest interpretowanie korelacji jako przyczynowości. To, że dwie zmienne są skorelowane, nie oznacza automatycznie, że jedna powoduje zmianę drugiej. Modulująca może być często trzecia zmienna, nieuwzględniona w analizie, a związki przyczynowe mogą być o wiele bardziej skomplikowane.

Aby uniknąć wprowadzenia w błąd siebie i innych, badacze i naukowcy powinni przygotować wykresy w sposób ułatwiający ich zrozumienie i interpretację. Świadomość tych pułapek i błędów umożliwia tworzenie bardziej precyzyjnych i wiarygodnych wykresów, co pozwala na wyciągnięcie dokładnych i rzetelnych wniosków z danych.

3.7.3. Przykłady zastosowania w biostatystyce

Ponadto interpretacja wykresów w biostatystyce jest nieodzowna dla dogłębnego zrozumienia tematu. Poniżej przedstawiono kilka przykładów, które mogą służyć jako praktyczne wytyczne.

Analiza przeżycia (ang. *survival analysis*): wykresy Kaplana–Meiera są jednym z najczęściej stosowanych narzędzi w analizie przeżycia. Prezentują szacunki przeżycia w zależności od czasu. Ważne jest zrozumienie, jak interpretować takie wykresy, szczególnie w kontekście cenzurowania (dane są „cenzurowane", gdy pacjent jest jeszcze żywy lub został utracony z obserwacji).

Analiza korelacji: wykresy punktowe są powszechnie używane do przedstawiania związku między dwiema zmiennymi. Ich interpretacja wymaga zrozumienia kierunku, siły i kształtu związku.

Analiza wielowymiarowa: metody redukcji wymiarowości, takie jak analiza głównych składowych (ang. *principal component analysis* – PCA), są często ilustrowane za pomocą dwuwymiarowych wykresów, które mogą być trudne do zrozumienia bez odpowiedniej wiedzy.

Analiza regresji: wykresy rozrzutu reszt (rezyduów) są najważniejszym narzędziem do oceny założeń modeli regresji.

Analiza szeregów czasowych: wykresy wykorzystuje się do przedstawienia danych gromadzonych w określonych odstępach czasu. Interpretacja wykresów szeregów czasowych obejmuje identyfikację trendów, sezonowości i innych wzorców.

Znając i rozumiejąc te i inne przykłady, badacze mogą efektywnie korzystać z wizualizacji danych w swoich analizach statystycznych. Ważne jest, aby pamiętać, że interpretacja każdego wykresu musi być dokonana w kontekście badania, danych i metod analizy.

3.8. Materiały dodatkowe i źródła

Internet jest bogatym źródłem zasobów i narzędzi, które mogą okazać się niezmiernie przydatne dla osób zajmujących się biostatystyką. Poniżej znajduje się zestawienie kilku wybranych stron i serwisów, które mogą znacząco przyczynić się do pogłębienia wiedzy w tej dziedzinie oraz usprawnić proces przeprowadzania analiz statystycznych.

R Project for Statistical Computing

https://www.r-project.org

Środowisko programistyczne R jest *de facto* standardem w biostatystyce. Umożliwia przeprowadzanie szerokiego zakresu analiz statystycznych i jest rozwijane przez liczną społeczność akademicką i podmioty komercyjne.

Bioconductor

https://www.bioconductor.org

Platforma do analizy danych biologicznych wykorzystująca język R. Bioconductor dostarcza narzędzia do analizy genomicznej, ale jest również przydatny w wielu innych zastosowaniach biostatystyki.

Python for Data Analysis

https://pandas.pydata.org

Biblioteka Pandas to potężne narzędzie do analizy danych w Pythonie. Jest bardzo popularna wśród naukowców i analityków danych.

GitHub

https://github.com

Umożliwia dzielenie się skryptami, wynikami i danymi, co ułatwia współpracę i powtarzalność badań. Wiele akademickich projektów związanych z biostatystyką jest dostępnych publicznie na GitHubie.

Coursera

https://www.coursera.org

Ta platforma e-learningowa oferuje wiele kursów z zakresu biostatystyki. Na przykład Uniwersytet Johnsa Hopkinsa oferuje specjalizację z *data science* obejmującą wiele aspektów analizy danych, w tym biostatystykę.

Khan Academy Statistics and Probability

https://www.khanacademy.org/math/statistics-probability

Seria darmowych lekcji wideo dotyczących statystyki i prawdopodobieństwa, dostępna na platformie Khan Academy, może być bardzo pomocna dla osób zaczynających swoją przygodę z biostatystyką.

Stack Overflow

https://stackoverflow.com

To platforma społeczności programistów, na której użytkownicy mogą zadawać pytania dotyczące kodowania i analizy danych. Wielu specjalistów z dziedziny biostatystyki regularnie z niej korzysta.

Bioinformatics Stack Exchange

https://bioinformatics.stackexchange.com

Platforma internetowa dedykowana specjalnie dla bioinformatyki, gdzie specjaliści mogą zadawać pytania i otrzymywać odpowiedzi na tematy związane z biostatystyką.

Wszystkie te zasoby są dostępne bezpłatnie, choć niektóre mogą oferować płatne funkcje premium. Często aktualizowane i utrzymywane przez aktywne społeczności są niezwykle wartościowym źródłem informacji i narzędzi dla każdego zajmującego się biostatystyką.

Piśmiennictwo

1. Pandey K, Panchal R. A Study of Real World Data Visualization of COVID-19 dataset using Python. *International Journal of Management and Humanities*. 2020;4(8):104–107. doi:10.35940/ijmh.H0834.044820

Andrii Zinchuk

4 Podstawowe formuły porządkujące dane w programie Excel

4.1. Krótki opis Excela i jego znaczenia w biostatystyce

Microsoft Excel to jeden z najpopularniejszych programów do opracowywania arkuszy kalkulacyjnych na świecie. Jego funkcjonalność sprawia, że jest powszechnie stosowany w różnych sektorach, w tym w medycynie i biostatystyce. Jest intuicyjny w użyciu i oferuje szereg narzędzi do opracowania danych, przeprowadzania obliczeń i analizy danych.

Excel jest szczególnie przydatny w biostatystyce, ponieważ umożliwia efektywne zarządzanie oraz analizę danych biologicznych i medycznych. Jest to bardzo ważne, ponieważ często są one złożone i wymagają starannej obróbki, aby można było na ich podstawie wyciągnąć właściwe wnioski. Funkcje Excela umożliwiają statystykom biologicznym przeprowadzanie podstawowych i zaawansowanych analiz statystycznych, takich jak testy *t*, analiza wariancji (ANOVA) czy regresja liniowa, co jest niezbędne do zrozumienia i interpretacji wyników badań.

Wartość Excela objawia się nie tylko w możliwościach przeprowadzania różnorodnych analiz statystycznych, ale także w elastycznym zarządzaniu danymi. Program umożliwia sortowanie, filtrowanie i przetwarzanie danych w sposób, który odpowiada potrzebom badaczy. Jest to szczególnie istotne w biostatystyce, gdzie analizowane zbiory danych mogą być bardzo duże i zróżnicowane.

Excel ma ponadto wbudowane funkcje, które pozwalają na automatyzację procesów dzięki zastosowaniu języka programowania Visual Basic for Applications (VBA) i makr. Dzięki temu badacze mogą zaoszczędzić czas poświęcany monotonnym zadaniom, skupiając się na bardziej skomplikowanych analizach statystycznych.

Program jest więc nieocenioną pomocą dla biostatystyków. Dzięki swojej wszechstronności i możliwościom personalizacji jest nie tylko niezwykle ważnym narzędziem w badaniach biologicznych i medycznych, ale także stanowi podstawę edukacji biostatystycznej.

4.2. Cel i znaczenie umiejętnego porządkowania danych w Excelu

Porządkowanie danych w Excelu ma duże znaczenie dla przetwarzania i analizy danych biologicznych i medycznych. Poprawne ich organizowanie jest niezbędne, by zapewnić wiarygodne i prawidłowe wyniki analizy statystycznej. Nieodpowiednie zarządzanie danymi może prowadzić do błędów, które z kolei mogą owocować niewłaściwymi wnioskami.

Celem porządkowania danych jest zapewnienie logicznego uporządkowania i łatwej ich dostępności. Excel oferuje szereg narzędzi do sortowania, filtrowania i grupowania, które pomagają w efektywnym zarządzaniu danymi. Odpowiednio uporządkowane dane łatwiej analizować, a wyniki takiej analizy są prostsze do interpretacji.

Badacze mogą porządkować dane w Excelu na wiele sposobów, w zależności od potrzeb. Można na przykład sortować dane wg różnych kryteriów, takich jak wartość, data lub w kolejności alfabetycznej, grupować je na podstawie określonych cech czy tworzyć tabele przestawne do skomplikowanej analizy danych. Każda z tych technik pomaga w organizacji danych, co z kolei ułatwia ich analizę.

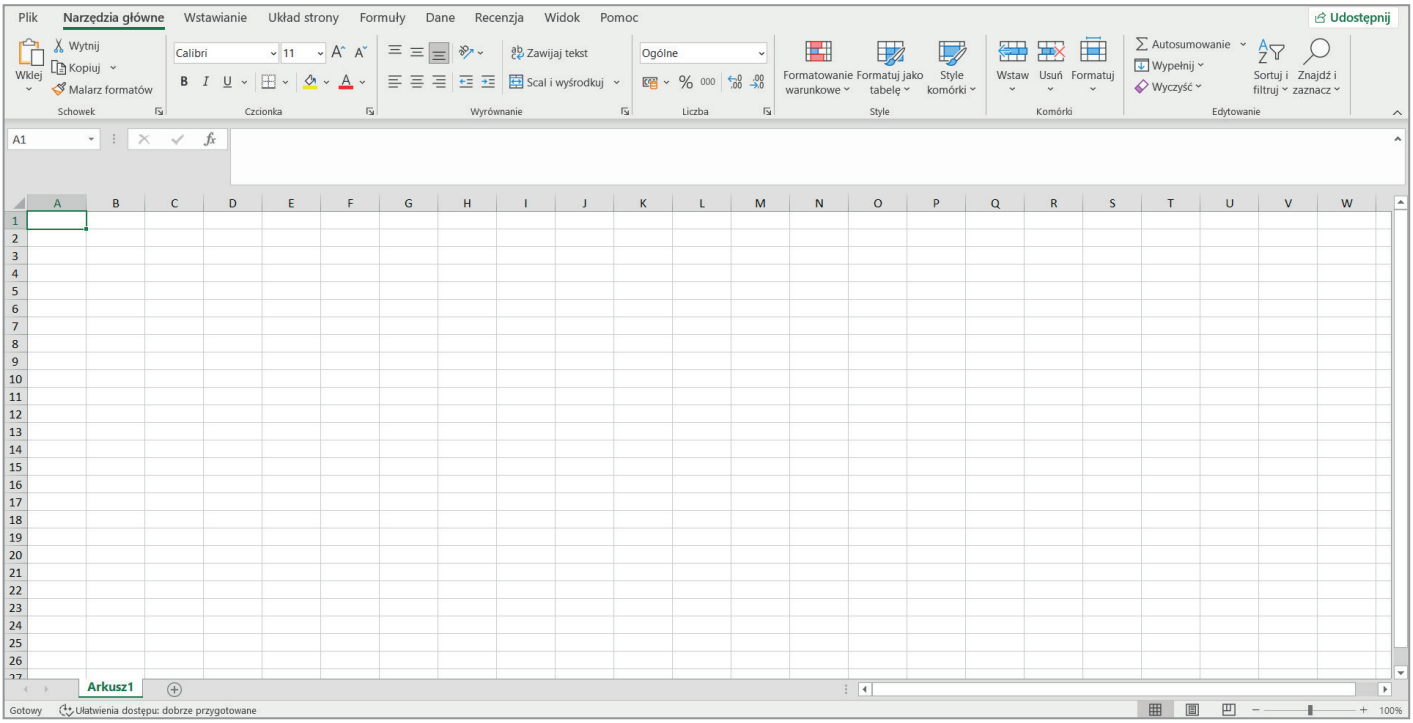
Porządkowanie danych jest ważne także dla utrzymania ich jakości. Wiarygodność wyników analizy statystycznej zależy w dużej mierze od jakości danych wejściowych. Dlatego też ich porządkowanie, w tym usuwanie duplikatów, uzupełnianie brakujących danych czy kontrolowanie spójności danych jest istotne dla zapewnienia ich jakości.

Umiejętność porządkowania danych w Excelu jest więc niezwykle cenna dla biostatystyków. Dzięki odpowiedniemu zarządzaniu nimi badacze mogą skupić się na ich analizie, a nie na radzeniu sobie z problemami związanymi z nieodpowiednią ich organizacją. Odpowiednie zarządzanie danymi to klucz do efektywnego i wiarygodnego badania biostatystycznego.

4.3. Podstawowe koncepcje Excela

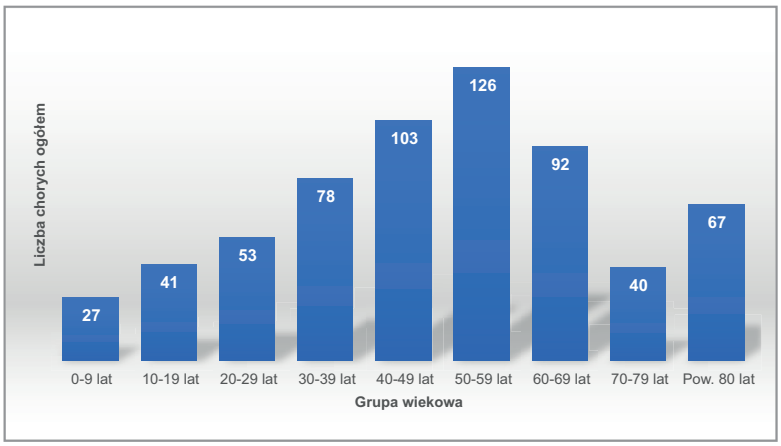
4.3.1. Wprowadzenie do arkuszy kalkulacyjnych

Arkusz kalkulacyjny to program, który umożliwia użytkownikom tworzenie tabel, wprowadzanie, organizowanie i analizowanie danych na wiele różnych sposobów. Każda tabela składa się z wierszy i kolumn, z których każda może zawierać różnego rodzaju dane, takie jak liczby, teksty, daty czy formuły (➥ **ryc. 4.1**).



Ryc. 4.1. Okno nowego arkusza kalkulacyjnego programu Excel

Excel umożliwia także tworzenie wykresów i diagramów, co ułatwia wizualizację danych i wyników analizy (➥ **ryc. 4.2**).



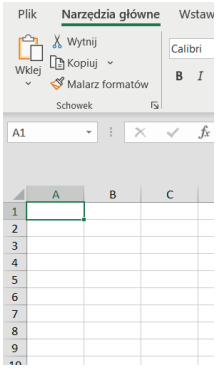
Ryc. 4.2. Przykład wykresu utworzonego w MS Excelu. Liczba chorych na SARS-Cov-2 w poszczególnych grupach wiekowych 11.03.2020–29.06.2020 r. w województwie podkarpackim.

Na podstawie: https://www.oil.rzeszow.pl/files/biuletyn/25/biuletyn-oil-merido-062020-a603aa7a.pdf

Ważne, aby zrozumieć, że o ile arkusze kalkulacyjne są potężnym narzędziem, to jednak wymagają od użytkownika umiejętności i wiedzy, aby korzystać z nich efektywnie. Błędy we wprowadzanych danych lub formułach mogą prowadzić do niepoprawnych wyników, dlatego nauka prawidłowego korzystania z tych programów jest najistotniejsza.

Niezależnie od poziomu zaawansowania arkusze kalkulacyjne są niezbędnymi narzędziami dla każdego, kto pracuje z danymi. Dlatego też zrozumienie podstawowych koncepcji stojących za nimi jest pierwszym krokiem na drodze do efektywnego wykorzystywania arkuszy.

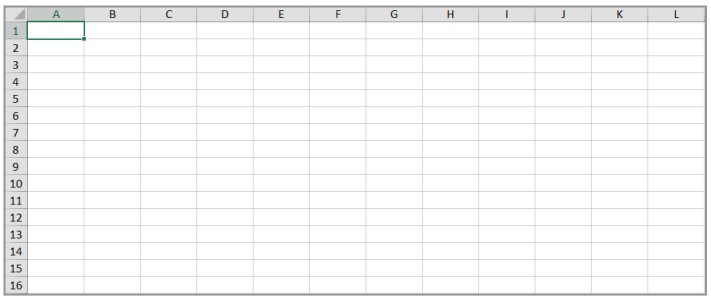
4.3.2. Komórki, kolumny, wiersze i zakresy



Ryc. 4.3. Pusty arkusz z zaznaczoną komórką A1

Podstawową strukturą arkusza kalkulacyjnego jest komórka. Jest to jego jednostka danych, która może zawierać tekst, liczby, daty lub formuły. Każda komórka w arkuszu kalkulacyjnym ma unikalny adres składający się z oznaczeń kolumny i wiersza, w których się znajduje. Na przykład komórka w pierwszej kolumnie i pierwszym wierszu będzie miała adres A1 (➥ **ryc. 4.3**). Uświadomienie sobie tego faktu jest konieczne, aby prawidłowo wprowadzać i analizować dane w arkuszu kalkulacyjnym.

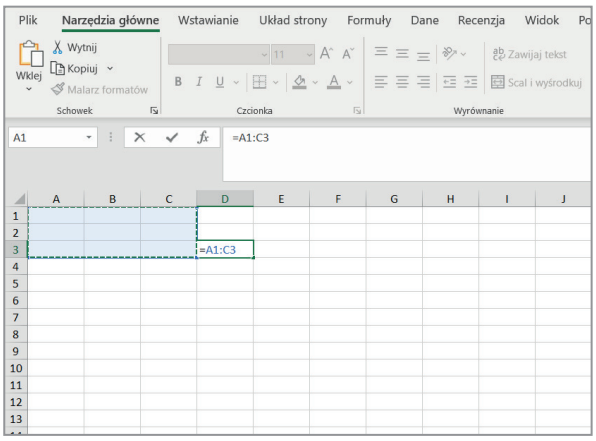
Kolumny i wiersze to dwa główne wymiary arkusza kalkulacyjnego. Kolumny są oznaczone literami (A, B, C itd.), podczas gdy wiersze są oznaczone liczbami (1, 2, 3 itd.). Każdy wiersz lub kolumna może zawierać dowolną liczbę komórek, w zależności od wielkości arkusza kalkulacyjnego (➥ **ryc. 4.4**).



Ryc. 4.4. Kolumny arkusza oznaczane są literami, a wiersze liczbami

W skład zakresu wchodzi jedna lub więcej komórek. Zakresy są bardzo istotne dla wielu operacji w arkuszu kalkulacyjnym, takich jak wprowadzanie formuł, tworzenie wykresów czy analizowanie danych. Są zazwyczaj oznaczane przez adresy ich pierw-

szej i ostatniej komórki oddzielone dwukropkiem. Na przykład zakres A1:C3 obejmuje komórki od A1 do C3, włączając wszystkie komórki między nimi (➥ **ryc. 4.5**).



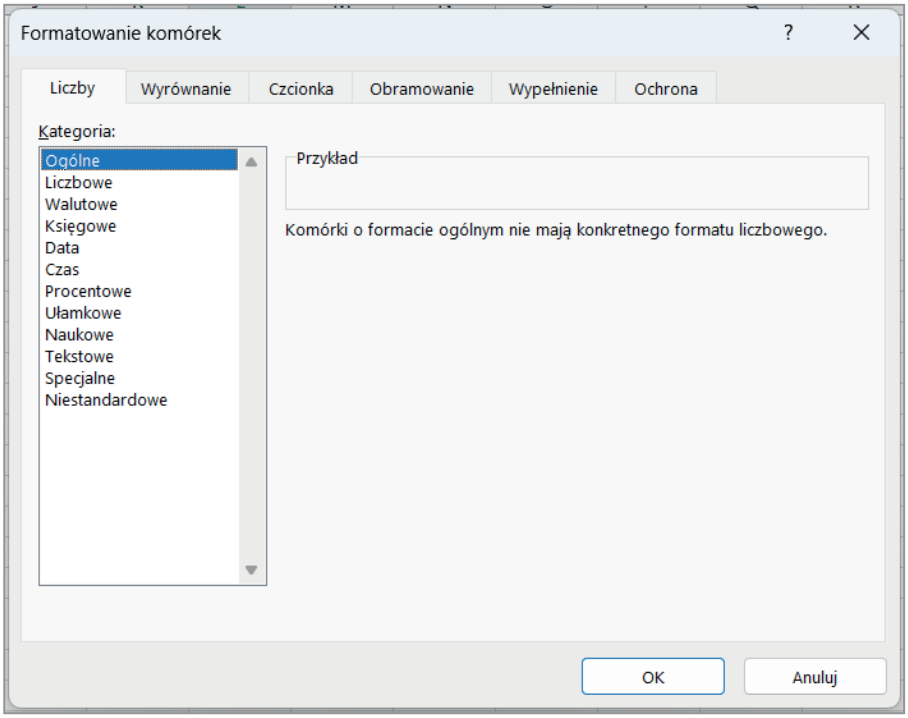
Ryc. 4.5. Zaznaczony zakres komórek A1:C3

Wszystkie te elementy współgrają ze sobą, tworząc strukturę, która umożliwia gromadzenie, analizowanie i prezentowanie danych na wiele różnych sposobów. Zrozumienie pojęć komórek, kolumn, wierszy i zakresów to podstawa efektywnego korzystania z arkuszy kalkulacyjnych i pierwszy krok do skutecznego posługiwania się Excelem niezależnie od dziedziny nauki.

4.3.3. Zasady wprowadzania danych

Wprowadzanie danych to fundamentalny aspekt pracy z arkuszami kalkulacyjnymi. Do prawidłowej analizy i interpretacji danych konieczne jest umiejętne i precyzyjne wprowadzanie informacji, czyli proces wpisywania wartości, formuł, dat, czasu lub tekstu do komórek arkusza.

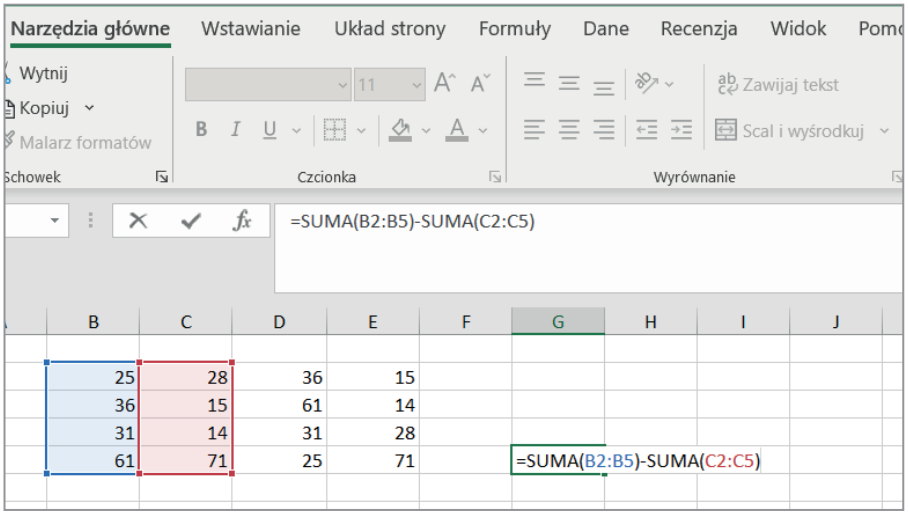
Jedną z najważniejszych cech danych jest ich typ. Excel rozróżnia różne typy danych, takie jak tekst, liczby, daty, boolean i inne. Określenie odpowiedniego ich typu ma decydujące znaczenie dla prawidłowego działania formuł i funkcji Excela (➥ **ryc. 4.6**).



Ryc. 4.6. Kategorie formatowania komórek

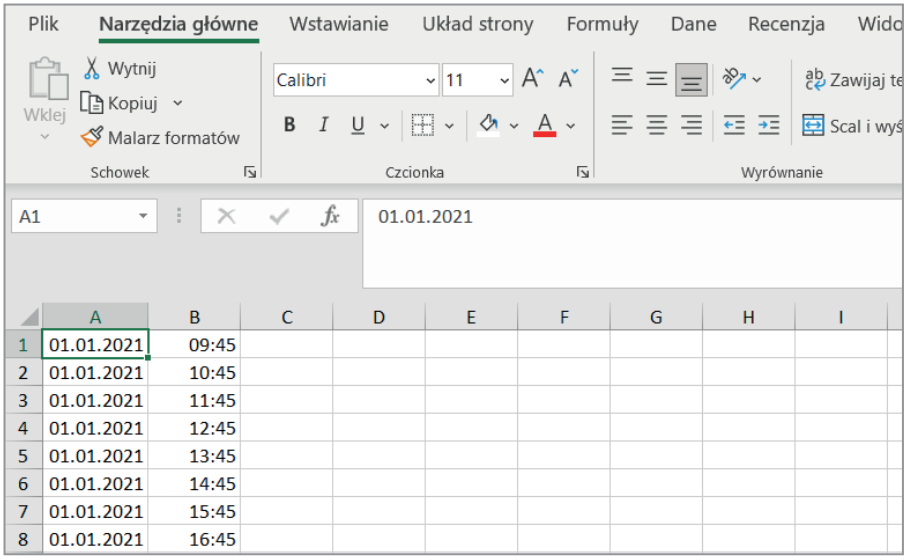
Dane można wprowadzać bezpośrednio do komórki, wpisując wartość i naciskając Enter, lub używając paska formuł, który umożliwia wprowadzanie i edycję wartości w wybranej komórce (➥ **ryc. 4.7**).

Ryc. 4.7. Formuła opisana w komórce pojawia się również na pasku formuł



Przy wprowadzaniu danych, szczególnie dla celów biostatystyki, ważne jest zachowanie spójności ich formatowania. Na przykład daty i czasy powinny być wpisywane w jednolity sposób w całym arkuszu kalkulacyjnym, aby zapewnić poprawne przetwarzanie i analizę (➥ **ryc. 4.8**).

Ryc. 4.8. Dane wprowadzone w jednolity sposób do arkusza

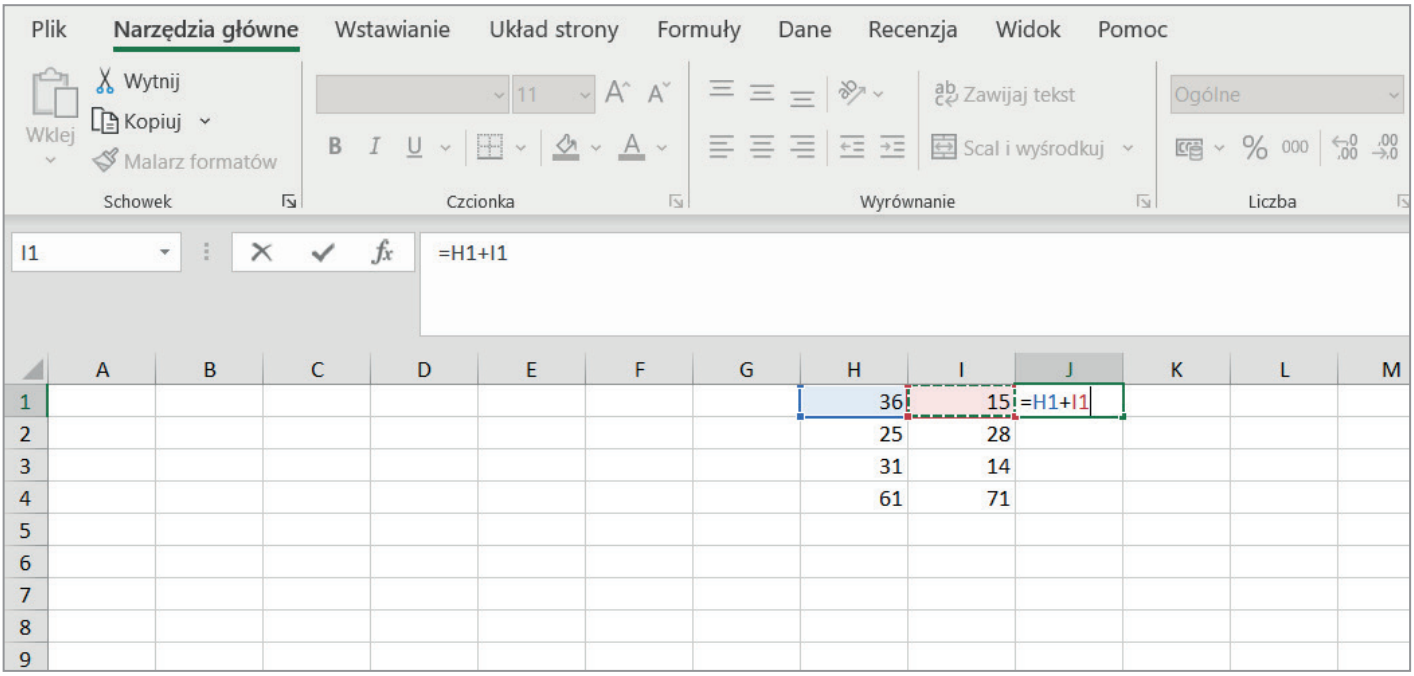


Wprowadzanie precyzyjnych i odpowiednio sformatowanych danych jest podstawą pracy z arkuszami kalkulacyjnymi. Umożliwia to skuteczną analizę i interpretację, szczególnie w medycynie i biostatystyce. Zrozumienie tego jest niezbędne do efektywnego poruszania się po arkuszu kalkulacyjnym i organizowania danych.

4.4. Formuły i funkcje w Excelu

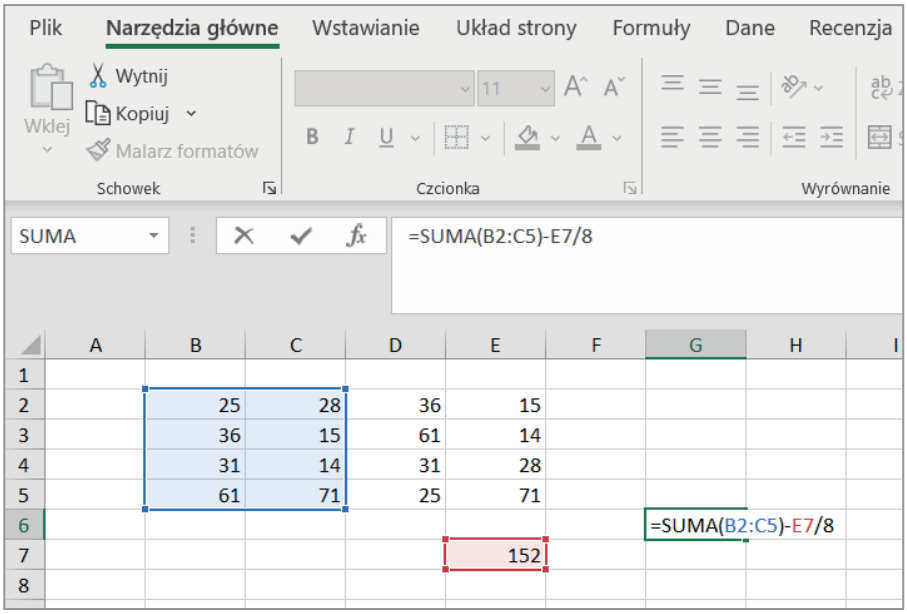
4.4.1. Wprowadzenie do formuł i funkcji

Formuły i funkcje w Excelu stanowią główny filar zdolności obliczeniowych tej aplikacji. Umożliwiają przetwarzanie, analizę i opracowanie wprowadzonych danych (➥ **ryc. 4.9**).



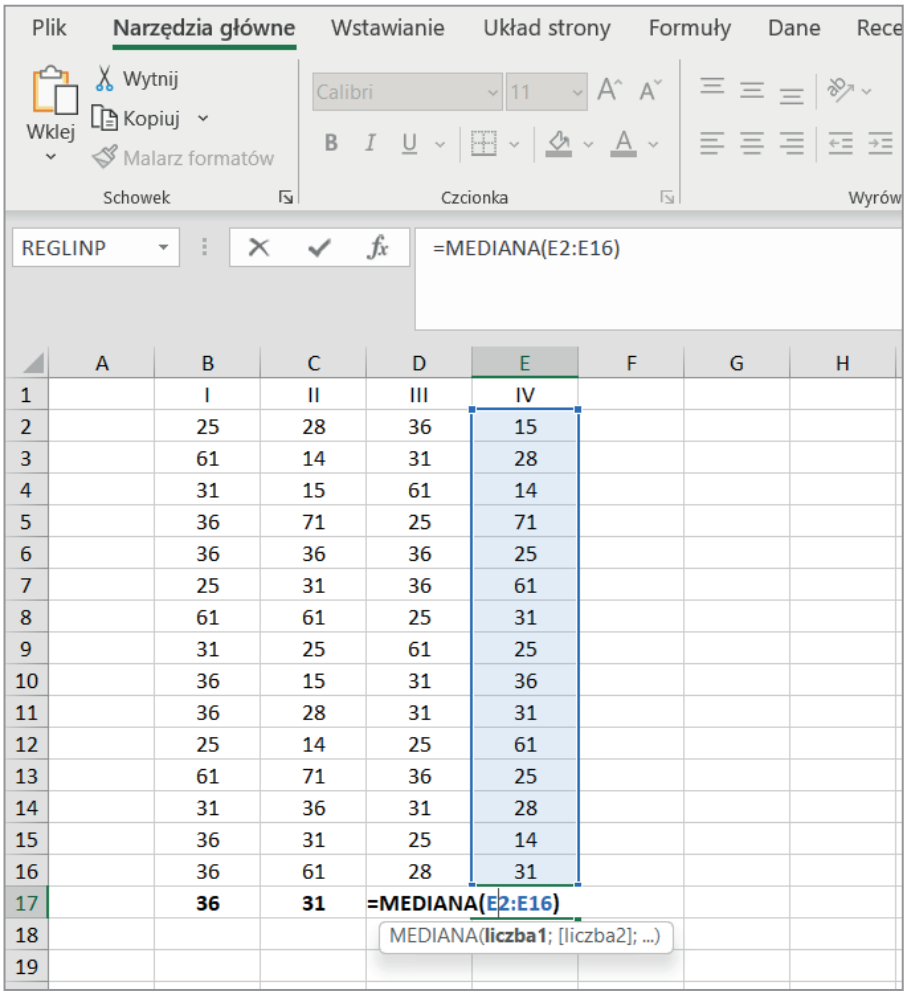
Ryc. 4.9. W komórce J1 wyświetlona zostanie suma wartości wprowadzonych do komórek H1 i I1

Formuły to wyrażenia, które wykonują obliczenia na wartościach wprowadzonych do komórek. Składają się z wartości (liczb, odwołań do komórek), operatorów matematycznych i funkcji (➥ **ryc. 4.10**).



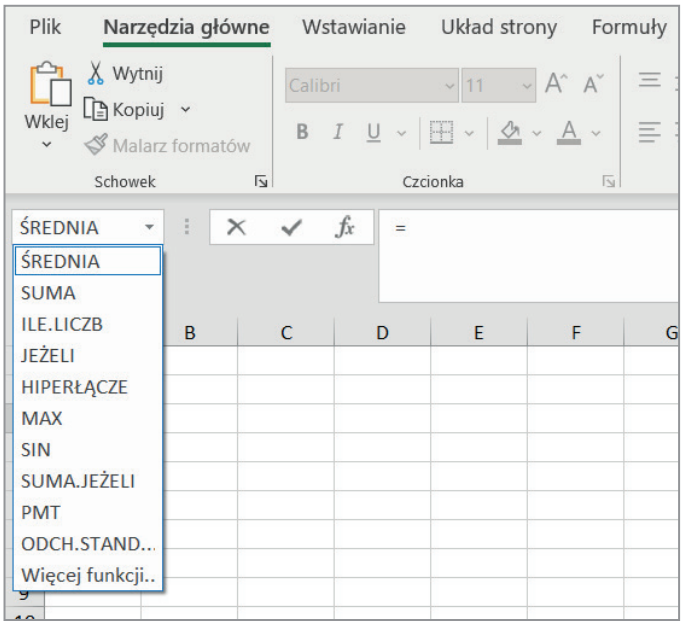
Ryc. 4.10. Formuła w Excelu

Z kolei funkcje to predefiniowane formuły, które wykonują specyficzne obliczenia na podstawie podanych argumentów. Excel zawiera szeroką gamę funkcji, od prostych matematycznych, takich jak SUMA(), do zaawansowanych statystycznych, które są często używane w biostatystyce, na przykład MEDIANA(), ŚREDNIA() czy ODCH.STANDARDOWE() (➥ **ryc. 4.11**).



Ryc. 4.11. Wyliczenie mediany z zakresu komórek E2:E16 za pomocą funkcji MEDIANA()

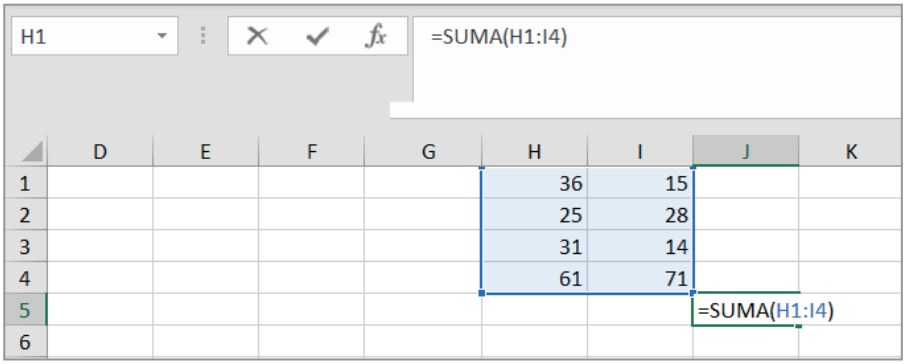
Funkcja ODCH.STANDARD.POPUL() oblicza odchylenie standardowe populacji. Używamy jej, gdy zakres wartości reprezentuje całą populację. Natomiast ODCH.STANDARD.PRÓBKI() jest wykorzystywana, gdy zakres wartości reprezentuje próbkę wartości, a nie całą populację (obliczamy odchylenie standardowe próbki, ➥ **ryc. 4.12**).



Ryc. 4.12. Rozwijane menu z najczęściej używanymi funkcjami, a wśród nich odchyleniem standardowym

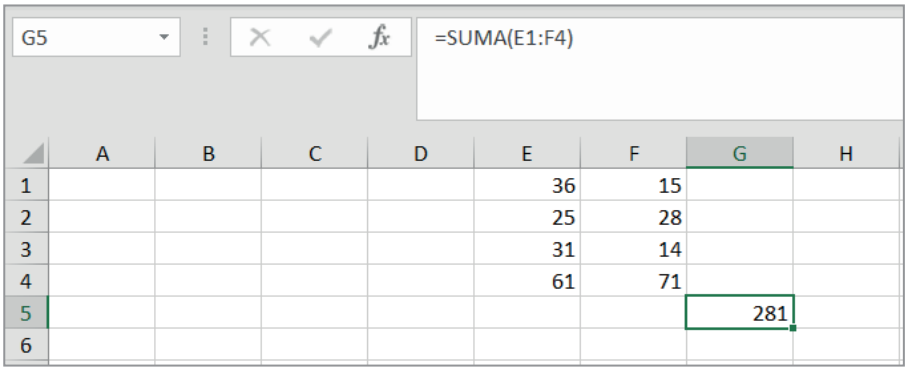
Zarówno formuły, jak i funkcje zaczynają się od znaku równości („=”), który informuje program, że zawartość komórki powinna być traktowana jak formuła lub funkcja, a nie jako statyczna wartość (➥ **ryc. 4.13**).

Ryc. 4.13. Formuły i funkcje zaczynają się od znaku „=”



Kiedy formuła lub funkcja zostanie wprowadzona do komórki, Excel automatycznie przelicza wartość komórki i wyświetla wynik obliczeń. Jednak formuła lub funkcja pozostaje widoczna w pasku formuł, umożliwiając jej edycję (➥ **ryc. 4.14**).

Ryc. 4.14. Po zatwierdzeniu formuły lub funkcji w komórce klawiszem Enter program automatycznie wylicza, a następnie wyświetla wartość wyliczenia



Zrozumienie formuł i funkcji jest niezbędne, aby efektywnie wykorzystać program, szczególnie w zakresie biostatystyki, gdzie zaawansowane obliczenia i analiza danych są na porządku dziennym.

4.4.2. Jak wprowadzać formuły

Jak wspomniano wcześniej, każda formuła w Excelu rozpoczyna się od znaku równości („=”). To sygnał dla programu, że wprowadzona zawartość powinna być traktowana właśnie jako formuła. Aby ją wprowadzić, trzeba najpierw wybrać komórkę, w której chcemy wyświetlić jej wynik, a następnie wpisać znak równości (=). Następnie można zdefiniować formułę. Może się ona składać z pojedynczych wartości (liczb), odwołań do innych komórek, operatorów matematycznych (takich jak „+”, „–”, „×”, „/”) i funkcji. Istotne jest zrozumienie, że Excel oblicza formuły w określonej kolejności, zgodnie z konwencją matematyczną dotyczącą kolejności wykonywania działań, znaną jako: nawiasy, wykładniki, mnożenie i dzielenie, dodawanie i odejmowanie (ang. BIDMAS – *brackets* (nawiasy), *indices* (indeksy), *division and multiplication* (dzielenie i mnożenie), *addition and subtraction* (dodawanie i odejmowanie).

Podczas wpisywania formuły Excel oferuje funkcję automatycznego uzupełniania, która może pomóc w szybkim znalezieniu i wprowadzeniu odpowiednich funkcji. Możliwe jest również ręczne wprowadzanie odwołań do komórek lub wykorzystanie myszy do wyboru tych, które mają być uwzględnione w formule.

Po wprowadzeniu formuły trzeba nacisnąć klawisz Enter, aby program obliczył jej wartość. Jeśli jest poprawna, wynik obliczeń pojawi się w wybranej komórce. Jeśli wystąpi błąd (np. błędna składnia formuły), zostanie wyświetlony komunikat o błędzie, który pomoże użytkownikowi zidentyfikować i skorygować problem.

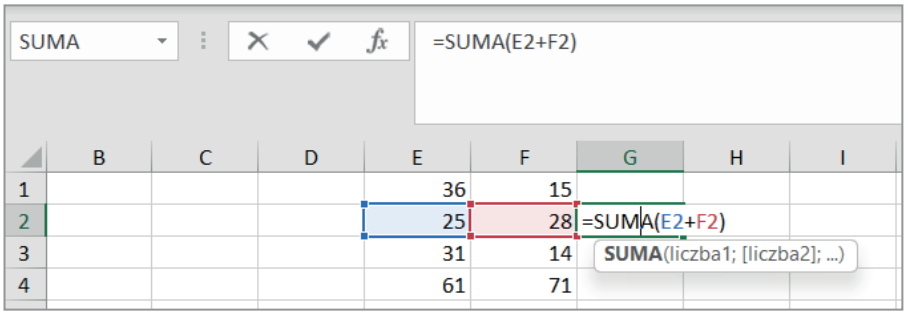
Warto zauważyć, że formuły w Excelu są dynamiczne. Oznacza to, że jeśli zmienią się wartości w komórkach, do których odwołuje się formuła, jej wynik również się

automatycznie zaktualizuje. Dynamiczne przetwarzanie danych jest jednym z głównych elementów, które czynią Excel nieocenionym narzędziem dla biostatystyków.

4.4.3. Najczęściej używane formuły i ich opis

4.4.3.1. Suma

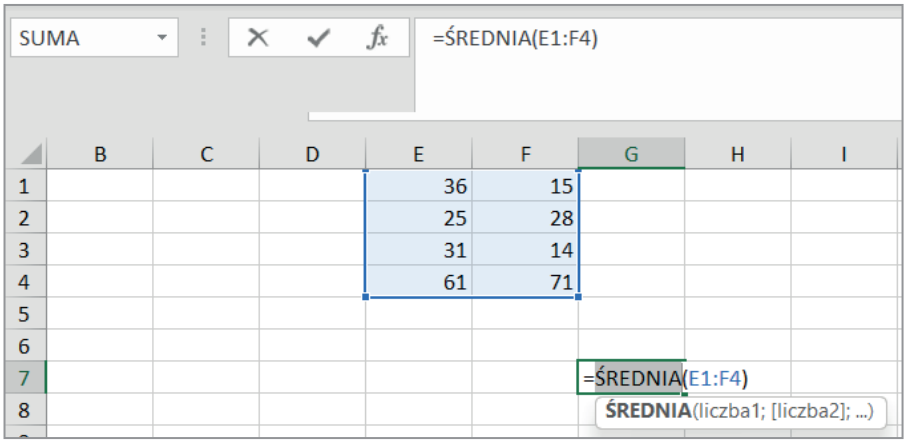
To podstawowa i często używana funkcja w Excelu. Zgodnie z nazwą oblicza ona sumę zestawu wartości. Składnia to: SUMA(liczba1; [liczba2];…), gdzie „liczba1; [liczba2];…" to zestaw wartości do dodania. Mogą być one wprowadzane jako liczby, mogą być także odwołaniami do komórek lub zakresów komórek (➥ **ryc. 4.15**).



Ryc. 4.15. Suma wartości komórek z zakresu E2:F2

4.4.3.2. Średnia

Ta funkcja oblicza średnią (arytmetyczną) zestawu wartości. Składnia funkcji to: ŚREDNIA(liczba1; [liczba2];...), gdzie „liczba1; [liczba2];…" to zestaw wartości, dla których ma być obliczona średnia. Podobnie jak w przypadku funkcji SUMA, te wartości mogą być wprowadzane bezpośrednio, a także odwołaniami do komórek lub zakresów komórek (➥ **ryc. 4.16**).

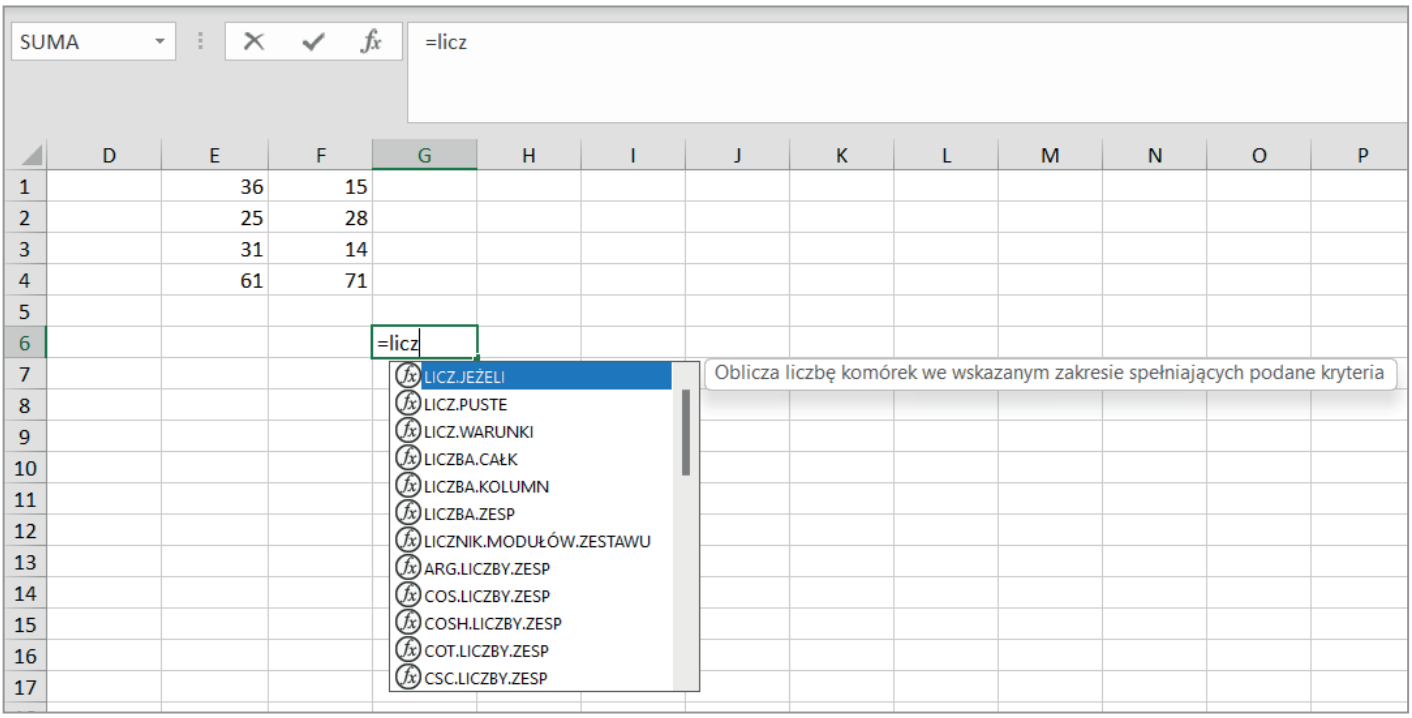


Ryc. 4.16. Średnia wartości w komórkach z zakresu E1:F4

4.4.3.3. Liczenie

Funkcja LICZ.… (LICZ.JEŻELI, LICZ.PUSTE, LICZ.WARUNKI) zwraca liczbę komórek w zakresie, które zawierają liczby. Składnia to: LICZ.…(wartość1; [wartość 2];…), gdzie „wartość1; [wartość2];…" to zestaw wartości, które mają być policzone. Ta funkcja jest szczególnie przydatna do zrozumienia wielkości próbki danych liczbowych (➥ **ryc. 4.17**).

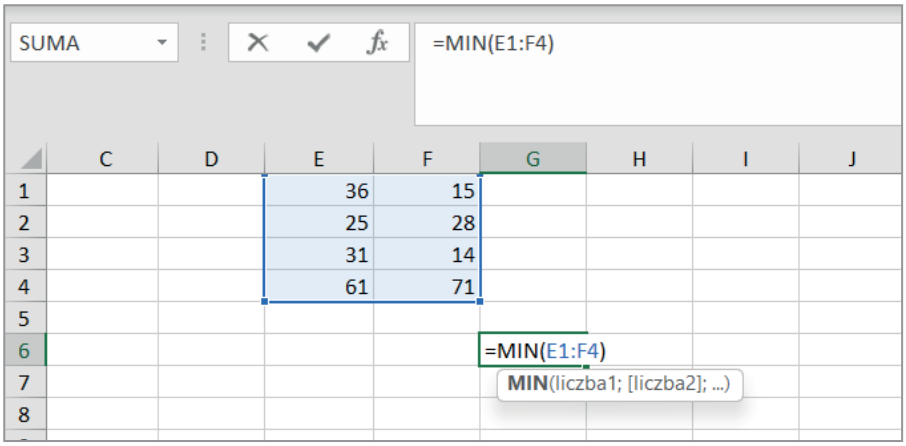
Ryc. 4.17. Podczas wpisywania nazwy funkcji pojawia się rozwijane menu z propozycjami funkcji do wstawienia



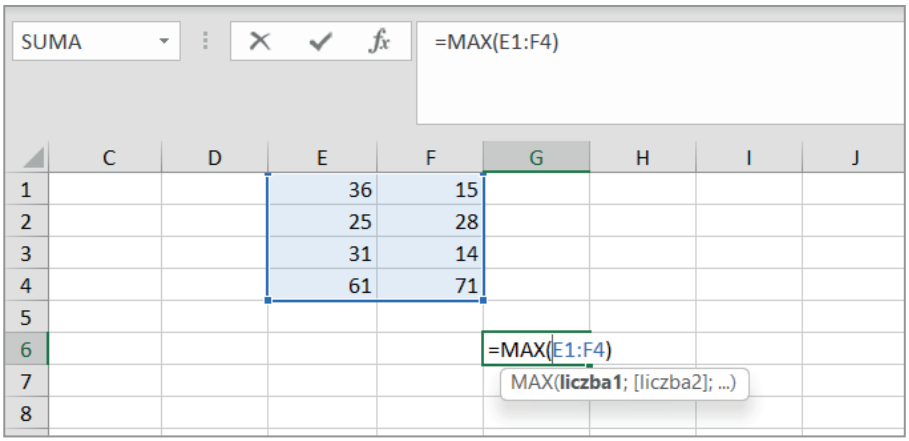
4.4.3.4. Wartości minimalne i maksymalne

Funkcje MIN i MAX służą do identyfikacji najmniejszej i największej wartości w zestawie danych. Składnia dla obu to: MIN(liczba1; [liczba2];…) lub MAX(liczba1; [liczba2];…), gdzie „liczba1; [liczba2];…" to zestaw, dla którego ma być znaleziona wartość minimalna lub maksymalna (➥ **ryc. 4.18**, ➥ **ryc. 4.19**).

Ryc. 4.18. Funkcja MIN wartości komórek z zakresu E1:F4



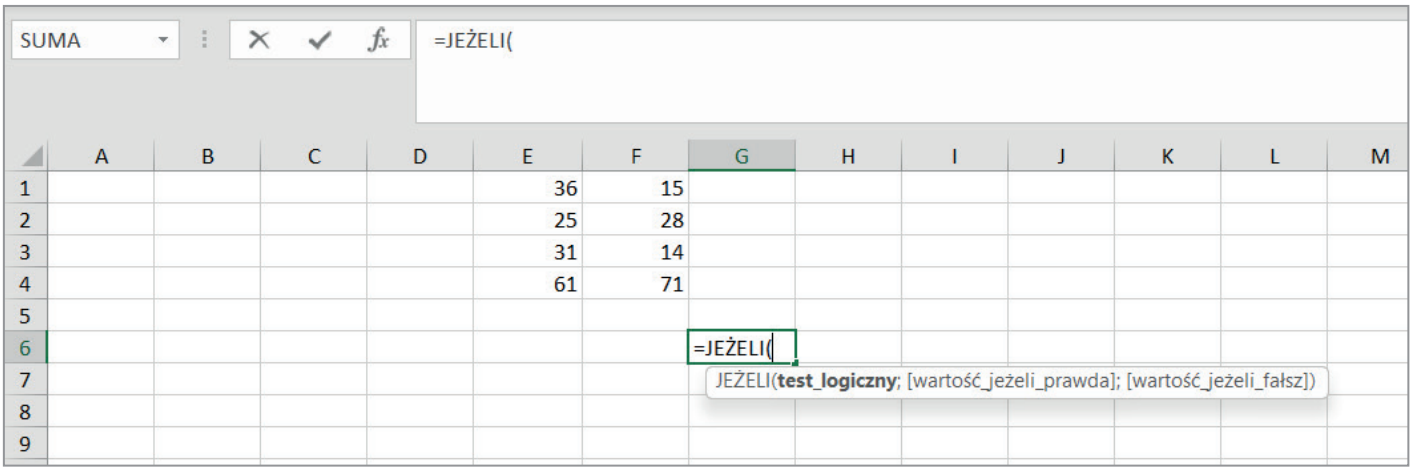
Ryc. 4.19. Funkcja MAX wartości komórek z zakresu E1:F4



4.4.3.5. Jeżeli – instrukcje warunkowe

Funkcja JEŻELI to potężne narzędzie, które umożliwia wprowadzenie logiki warunkowej do arkuszy Excela. Składnia to: JEŻELI(test_logiczny; [wartość_jeśli_prawda]; [wartość_jeśli_fałsz]), gdzie „test_logiczny" to warunek do sprawdzenia, „wartość_jeśli_prawda" to wartość, która ma być zwrócona, jeśli warunek jest spełniony, a „wartość_jeśli_fałsz" to wartość, która ma być zwrócona, jeśli warunek nie jest spełniony (➥ **ryc. 4.20**).

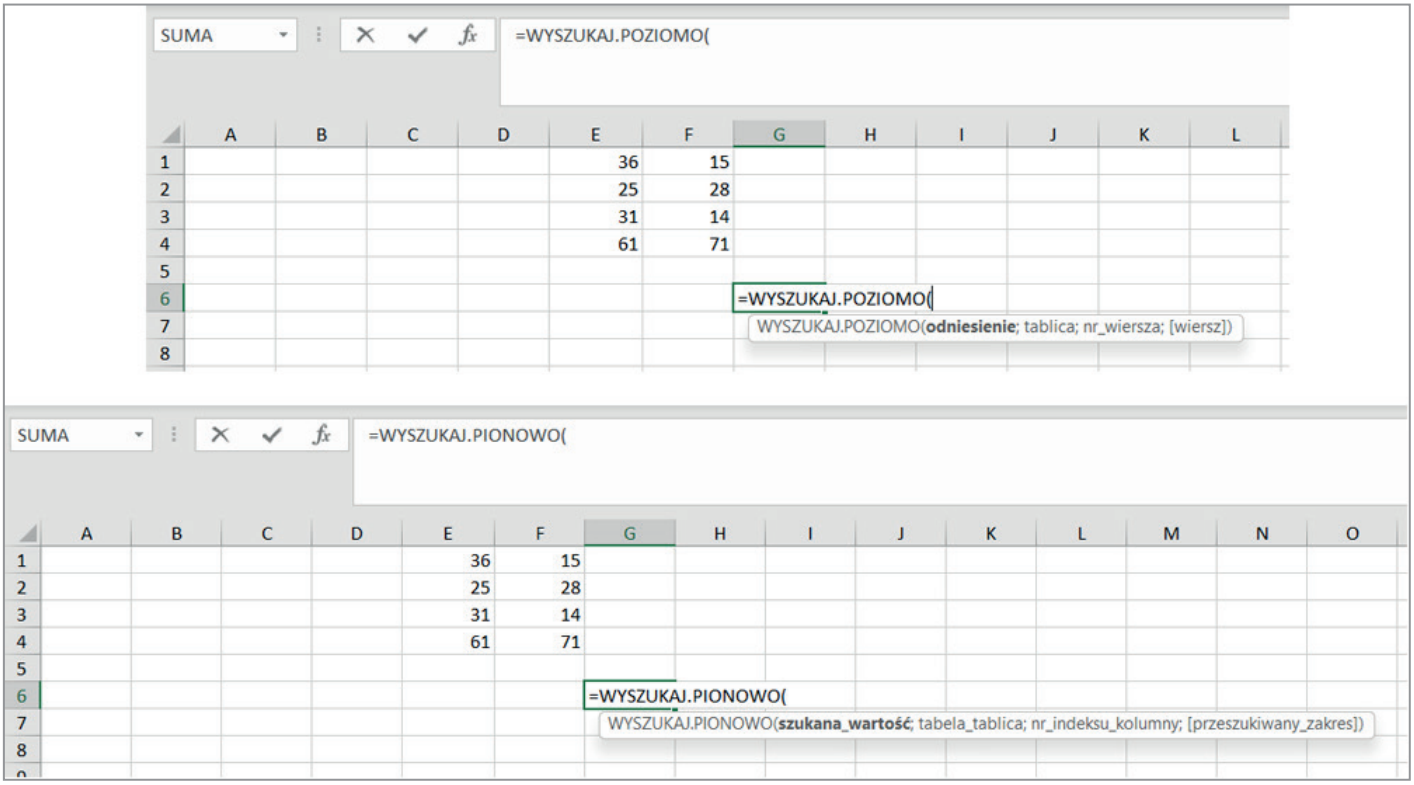
Ryc. 4.20. Zastosowanie funkcji JEŻELI



4.4.3.6. Wyszukiwanie wartości w tabeli

Funkcje WYSZUKAJ.PIONOWO (przeglądanie w pionie) oraz WYSZUKAJ POZIOMO (przeglądanie w poziome) są wykorzystywane do wyszukiwania konkretnych danych w tabeli. WYSZUKAJ.PIONOWO odnosi się do kolumn, a WYSZUKAJ.POZIOMO do wierszy. Składnia dla obu funkcji to: WYSZUKAJ.PIONOWO (szukana_wartość; tabela_tablica; nr_indeksu_kolumny; [przeszukiwany_zakres]) lub WYSZUKAJ.POZIOMO (odniesienie; tablica; nr_wiersza; [wiersz]), gdzie „szukana_wartość" to wartość, której szukamy, „_tablica" to zakres, w którym szukamy, „nr_indeksu_kolumny" lub „nr_wiersza" to numer kolumny lub wiersza, z którego chcemy zwrócić wartość, a „przeszukiwany_zakres" to opcjonalny argument określający, czy chcemy dokładne dopasowanie (FAŁSZ) lub przybliżone dopasowanie (PRAWDA) (➥ **ryc. 4.21**).

Ryc. 4.21. Arkusz z opisanymi funkcjami WYSZUKAJ.PIONOWO i WYSZUKAJ. POZIOMO



4.5. Przykłady zastosowania formuł do porządkowania danych

4.5.1. Tworzenie i interpretacja zestawień danych

W biostatystyce proces tworzenia i interpretacji zestawień danych jest niezbędny do zrozumienia zależności pomiędzy zmiennymi i ich dystrybucji oraz do identyfikacji trendów i wzorców. Excel, z jego rozbudowanym zestawem narzędzi i formuł, jest wyjątkowo przydatny w tworzeniu takich zestawień.

Zestawienie danych zaczyna się od ich gromadzenia i czyszczenia. Excel zawiera wiele formuł do sortowania i filtrowania danych, które ułatwiają ten proces. Na przykład formuły takie jak „JEŻELI", „LICZ.JEŻELI" i „SUMA.JEŻELI" mogą być użyte do sortowania danych wg określonych kryteriów, a „WYSZUKAJ.PIONOWO" i „WYSZUKAJ.POZIOMO" mogą pomóc w odnajdywaniu konkretnych informacji w dużych zbiorach danych.

Następnie można użyć formuł Excela do tworzenia zestawień krzyżowych, histogramów, wykresów punktowych, wykresów pudełkowych itp., które pomagają wizualizować dane i ułatwiają ich interpretację. Funkcje „CZĘSTOŚĆ" mogą być np. użyte do tworzenia histogramów, które pokazują dystrybucję danych.

Interpretacja zestawień danych wymaga zrozumienia metod statystycznych i zasad ich stosowania w Excelu. Na przykład wyniki testów statystycznych, takich jak testy *t*, testy χ^2, czy analizy wariancji (ANOVA), mogą być zinterpretowane przy użyciu wbudowanych funkcji Excela, takich jak „T.TEST", „CHI.TEST" i „WARIANCJA".

Ponadto Excel oferuje możliwość tworzenia zaawansowanych modeli statystycznych, takich jak regresje liniowe i nieliniowe, dzięki czemu badacze mogą analizować związki między zmiennymi i przewidywać wyniki na podstawie tych zależności.

Zrozumienie, jak skutecznie tworzyć i interpretować zestawienia danych w Excelu, umożliwia efektywne wykorzystanie tej platformy w biostatystyce. Dzięki swoim rozbudowanym formułom i narzędziom Excel jest nieocenioną pomocą dla badaczy zajmujących się zagadnieniami z zakresu medycyny w analizie i interpretacji danych.

4.5.2. Analiza danych za pomocą formuł Excela

Formuły Excela umożliwiają przetwarzanie i analizę skomplikowanych zestawów danych w celu uzyskania wartościowych informacji i ujawnienia trendów, co stanowi jego olbrzymią zaletę.

Formuły statystyczne, takie jak „ŚREDNIA", „MEDIANA", „WYST.NAJCZĘŚCIEJ", „ODCH.STANDARDOWE" i „WARIANCJA" mogą być zastosowane do obliczenia najważniejszych parametrów statystycznych, które ułatwiają podstawowe zrozumienie rozkładu danych. Formuły „LICZ.…" („LICZ.JEŻELI", „LICZ.PUSTE", „LICZ.WARUNKI") umożliwiają zliczanie elementów spełniających określone kryteria, co jest szczególnie przydatne w analizie częstości zdarzeń.

Testy hipotez można przeprowadzać przy użyciu formuł takich jak „T.TEST", „F.TEST" i „CHI.TEST". Pozwalają one przeprowadzić odpowiednie testy statystyczne, porównanie średnich wartości dwóch grup, analizę wariancji czy test χ^2, co pomaga w ocenie statystycznej istotności różnic pomiędzy grupami.

Excel umożliwia również przeprowadzanie analiz regresji liniowej i nieliniowej, co pozwala naukowcom na badanie zależności między zmiennymi i przewidywanie przyszłych wartości na ich podstawie.

Formuły „WYSZUKAJ.PIONOWO" i „WYSZUKAJ.POZIOMO" są szczególnie użyteczne w biostatystyce – dają możliwość szybkiego wyszukiwania i porównywania danych w ich dużych zestawach. Podobnie formuły „INDEX" i „DOPASUJ" mogą być używane do odnajdywania konkretnych wartości w tabelach.

Formuły logiczne, takie jak „JEŻELI", „ORAZ", „LUB" i „NIE", służą do tworzenia skomplikowanych instrukcji warunkowych, które mogą znacznie zwiększyć efektywność przetwarzania danych.

Ponadto formuły Excela można łączyć w celu tworzenia formuł bardziej skomplikowanych, które mogą być wykorzystane do przeprowadzania specjalistycznych analiz danych. Wreszcie można ich używać wraz z narzędziami Power Query do zaawansowanej transformacji i analizy danych.

Sposobność efektywnego wykorzystania formuł Excela do analizy danych jest bardzo ważna dla każdego badacza w dziedzinie biostatystyki. Dzięki temu można przekształcić surowe dane w użyteczne informacje, pomocne w zrozumieniu zjawisk biologicznych i medycznych.

4.5.3. Scenariusze, w których techniki analizy danych z wykorzystaniem formuł Excela mogą być użyteczne

Techniki analizy danych z wykorzystaniem formuł Excela mają wiele praktycznych zastosowań w dziedzinie biostatystyki. Umiejętność korzystania z tych narzędzi ma dla badaczy i naukowców nieocenione znaczenie w efektywnej analizie i interpretacji danych.

W badaniach klinicznych formuły Excela mogą być wykorzystane do analizy danych pacjentów, takich jak pomiary parametrów biochemicznych, objawy chorobowe czy wyniki testów. Formuły „ŚREDNIA", „MEDIANA" oraz „WYST.NAJCZĘŚCIEJ" mogą pomóc w ocenie centralnej tendencji tych danych, podczas gdy „ODCH.STANDARDOWE" i „WARIANCJA" – w ocenie ich rozproszenia. Może to pomóc lekarzom i naukowcom zrozumieć, jak różne czynniki wpływają na wyniki pacjentów, i pomóc w opracowywaniu nowych strategii leczenia.

W epidemiologii formuły takie jak „LICZ.…" („LICZ.JEŻELI", „LICZ.PUSTE", „LICZ.WARUNKI") można wykorzystać do zliczania liczby przypadków chorób na podstawie różnych czynników, takich jak demografia, geografia czy czas, niezbędnych do monitorowania i prognozowania rozprzestrzeniania się chorób.

W genetyce za pomocą Excela można analizować duże zestawy danych genetycznych. Narzędziem Solver naukowcy mogą przeprowadzać analizy regresji, aby zrozumieć, jak różne geny i ich mutacje wpływają na cechy fenotypowe. Formuły „WYSZUKAJ.PIONOWO" i „WYSZUKAJ.POZIOMO" ułatwiają przeszukiwanie i porównywanie danych genetycznych, co umożliwia identyfikację nowych genów związanych z chorobami.

W przypadku bioinformatyki Excel znajduje zastosowanie do przetwarzania i analizy dużych zestawów danych sekwencyjnych. Formuły takie jak „INDEX" i „DOPASUJ" pomagają szybko zlokalizować specyficzne sekwencje w dużych zestawach danych.

Formuły logiczne – „JEŻELI", „ORAZ", „LUB" i „NIE" – są szczególnie przydatne w przypadku przetwarzania i analizy danych kategorycznych, które są powszechne w badaniach medycznych i biologicznych.

Wreszcie za pomocą narzędzi Power Query można przeprowadzać zaawansowane przekształcenia i analizy danych, takie jak czyszczenie danych, łączenie różnych źródeł danych, a nawet przeprowadzanie analizy eksploracyjnej danych.

W każdym z tych scenariuszy formuły Excela umożliwiają badaczom przekształcenie surowych danych w wartościowe informacje, które będą pomocne w zrozumieniu zjawisk biologicznych i medycznych.

4.6. Podsumowanie

4.6.1. Krótkie przypomnienie głównych problemów poruszonych w rozdziale

W rozdziale ukazano, jak formuły Excela można wykorzystać do zarządzania danymi i ich analizy w medycynie i biostatystyce. W omawianym kontekście zastosowanie programu nie ogranicza się tylko do prostych obliczeń, ale stanowi potężne narzędzie do przetwarzania danych i przeprowadzania zaawansowanej analizy statystycznej.

Najpierw, w podrozdziale 4.4, omówiliśmy podstawowe formuły dostępne w Excelu, takie jak operacje arytmetyczne, formuły statystyczne, logiczne oraz wyszukiwania i odniesienia. Wyjaśniliśmy, jak można je wykorzystać do organizowania i przetwarzania danych, co jest kluczowym krokiem w każdym procesie analizy danych.

W podrozdziale 4.5 podkreśliliśmy znaczenie formuł i narzędzi w praktycznych zastosowaniach biostatystycznych, takich jak badania kliniczne, epidemiologia, genetyka i bioinformatyka. Przedstawiliśmy rozmaite scenariusze, w których techniki te mogą okazać się niezbędne dla efektywnej analizy i interpretacji danych.

Wreszcie warto podkreślić, że prawidłowe stosowanie tych narzędzi wymaga nie tylko technicznej wiedzy na temat Excela, ale także dogłębnego zrozumienia statystyki i metodologii badawczej. Tylko wtedy można skutecznie wykorzystać potencjał programu do przekształcania surowych danych w wartościowe informacje naukowe.

4.6.2. Znaczenie umiejętności porządkowania danych w biostatystyce

Umiejętność właściwego porządkowania danych jest fundamentalna w biostatystyce, a zatem w szeroko pojętej dziedzinie medycyny. Dane są nieodłącznym elementem badań naukowych, a ich właściwe gromadzenie, porządkowanie, analiza i interpretacja są podstawowym warunkiem uzyskania wiarygodnych wyników.

Excel, dzięki swoim rozbudowanym funkcjom, umożliwia przeprowadzanie analizy danych na szeroką skalę, począwszy od prostych obliczeń, a na zaawansowanej analizie statystycznej kończąc. Porządkowanie danych w Excelu, czyli ich organizacja w sposób umożliwiający łatwe przetwarzanie i analizę, jest podstawą każdego procesu badawczego.

Umiejętność ich porządkowania umożliwia optymalizację procesów ich analizy, prowadząc do szybszych i bardziej precyzyjnych wyników. Jest to szczególnie istotne w kontekście biostatystyki, gdzie analiza często obejmuje skomplikowane zestawy danych z wielu źródeł.

Porządkowanie danych umożliwia również lepsze zrozumienie i interpretację wyników. Dane, które zostały odpowiednio uporządkowane i przetworzone, są łatwiejsze do analizy i interpretacji, co prowadzi do solidniejszych i bardziej wiarygodnych wniosków. Dlatego umiejętność ich porządkowania w Excelu jest niezwykle cenna dla każdego, kto pracuje z danymi w dziedzinie medycyny i biostatystyki.

Zwiększa to efektywność pracy, ale również prowadzi do głębszego zrozumienia danych, co w rezultacie może przynieść korzyści dla pacjentów i społeczeństwa. Dlatego ważne jest, aby badacze i praktycy w dziedzinie medycyny i biostatystyki ciągle doskonalili swoje umiejętności korzystania z Excela i innych narzędzi do analizy danych.

4.6.3. Zachęta do dalszej nauki i eksploracji Excela

Jak zauważono w poprzednich częściach tego podsumowania, Excel jest bardzo ważnym narzędziem. Jednakże, aby w pełni wykorzystać potencjał programu, konieczna jest ciągła nauka i eksploracja jego możliwości.

Nauka Excela nie jest jednorazowym zadaniem, ale procesem ciągłego rozwijania umiejętności i poznawania nowych funkcji. Dzięki takim narzędziom jak VBA i Power Query aplikacja umożliwia przetwarzanie i analizę danych na poziomie zaawansowanym, który przekracza podstawowe funkcje programu. Zdolność do wykorzystania tych narzędzi może znacznie zwiększyć efektywność i precyzję analizy.

Zachęcamy do korzystania z literatury specjalistycznej w celu poszerzenia wiedzy na temat Excela. Wiele publikacji dostarcza praktyczne porady i wskazówki, które mogą pomóc w opanowaniu różnych funkcji programu. Niektóre spośród nich skupiają się na konkretnych aspektach Excela, takich jak analiza statystyczna, przetwarzanie danych czy programowanie, co pozwala na specjalistyczne pogłębienie wiedzy.

Równocześnie najistotniejsze jest zrozumienie zastosowania Excela w kontekście biostatystyki i medycyny. Także w tym przypadku warto korzystać z literatury specjalistycznej, która pokazuje, jak wykorzystać Excela w tych dziedzinach. Wyko-

rzystanie programu w praktyce może przynieść wiele korzyści, ale wymaga ciągłej nauki i dostosowywania narzędzi do własnych potrzeb. Warto wreszcie pamiętać, że nauka Excela nie jest tylko kwestią techniczną. Jest to także proces, który pozwala na lepsze zrozumienie danych, co jest bardzo ważne w pracy naukowej. Wykorzystanie tego oprogramowania do analizy danych pozwala na rozwijanie umiejętności krytycznego myślenia, niezbędnej w procesie naukowym.

Zachęcamy zatem do dalszej nauki Excela. Choć może się to wydawać skomplikowane i czasochłonne, w dłuższej perspektywie z pewnością przyniesie korzyści, zwiększając efektywność pracy, jakość badań i poziom zrozumienia danych.

4.7. Dodatki

4.7.1. Lista najczęściej używanych skrótów klawiaturowych w Excelu

Efektywne korzystanie z programu Excel wiąże się nierozłącznie z umiejętnym stosowaniem skrótów klawiaturowych, co znacznie przyspiesza i ułatwia pracę z danymi. Poniżej lista najczęściej używanych skrótów klawiaturowych w Excelu.

Nawigacja

Ctrl + strzałka (w górę/dół/lewo/prawo) – przemieszcza kursor do krawędzi bieżącego zakresu danych.

Home – przenosi do początku wiersza.

Ctrl + Home – przenosi do pierwszej komórki arkusza (domyślnie A1).

Ctrl + End – przenosi do ostatniej komórki z danymi w arkuszu.

Selekcja

Shift + strzałka (w górę/dół/lewo/prawo) – wybiera komórki w kierunku strzałki.

Ctrl + Shift + strzałka (w górę/dół/lewo/prawo) – wybiera zakres komórek od bieżącej komórki do krawędzi bieżącego zakresu danych.

Edycja

F2 – umożliwia edycję bieżącej komórki.

Ctrl + X – wycina zaznaczoną komórkę lub zakres komórek.

Ctrl + C – kopiuje zaznaczoną komórkę lub zakres komórek.

Ctrl + V – wkleja zawartość schowka do bieżącej komórki.

Ctrl + Z – cofa ostatnią akcję.

Ctrl + Y – powtarza ostatnią akcję.

Formatowanie

Ctrl + 1 – otwiera okno dialogowe Format Cells.

Ctrl + Shift + ~ – stosuje format General do zaznaczonych komórek.

Ctrl + Shift + $ – stosuje format Currency do zaznaczonych komórek.

Ctrl + Shift + % – stosuje format Percentage do zaznaczonych komórek.

Operacje na arkuszach

Ctrl + PgUp – przechodzi do poprzedniego arkusza w skoroszycie.

Ctrl + PgDn – przechodzi do następnego arkusza w skoroszycie.

Ctrl + Shift + F1 – dodaje nowy arkusz.

Ctrl + F6 – przełącza pomiędzy otwartymi skoroszytami.

Znajomość powyższych skrótów klawiaturowych oraz ich stosowanie bardzo się przydaje i pozwala na efektywne wykorzystanie potencjału Excela w przeprowadzaniu analiz biostatystycznych. Skróty te pomagają nie tylko przyspieszyć pracę, ale także uczynić ją bardziej intuicyjną i ergonomiczną, co przekłada się na zwiększenie wydajności pracy badawczej.

4.7.2. Linki do dodatkowych źródeł informacji na temat Excela

Kontynuowanie nauki i rozwijanie umiejętności jest kluczem do skutecznego wykorzystania narzędzi takich jak Excel w dziedzinie biostatystyki. Poniżej podajemy linki do niektórych pomocnych źródeł, które mogą posłużyć do dalszego eksplorowania i pogłębiania wiedzy w tym zakresie.

Microsoft Excel Help Center
https://www.free-online-training-courses.com
Wspierana przez Microsoft strona, która oferuje szereg poradników i przewodników na temat różnych aspektów korzystania z Excela. Jest aktualizowana na bieżąco, zgodnie z najnowszymi wersjami oprogramowania.

Excel Jet
https://exceljet.net
Oferuje lekcje wideo oraz tekstowe na temat różnych funkcji i technik używanych w Excelu. Serwis jest szczególnie pomocny dla osób, które preferują wzrokowy styl uczenia się.

Chandoo.org
https://chandoo.org
Na tej stronie znajdują się szczegółowe poradniki na temat zaawansowanych technik i strategii używanych w Excelu, które przede wszystkim mogą być przydatne dla osób zajmujących się biostatystyką.

Reddit r/excel
https://www.reddit.com
To forum dyskusyjne jest miejscem, w którym użytkownicy mogą zadawać pytania i udzielać odpowiedzi na temat różnych aspektów korzystania z Excela. Można tu także znaleźć rozwiązania specyficznych problemów, z jakimi mogą się spotkać użytkownicy programu.

Coursera – Excel Skills for Business
https://www.coursera.org/specializations/excel
Ten kurs online oferowany przez Macquarie University koncentruje się na możliwościach Excela szczególnie przydatnych w kontekście biznesowym, ale wiele z nich można wykorzystać w biostatystyce.

Khan Academy – Intro to Statistics
https://www.khanacademy.org/math/statistics-probability
Chociaż nie jest konkretnie związany z Excelem, ten kurs online stanowi solidne omówienie podstaw statystyki, niezbędnych, aby skutecznie wykorzystywać Excel w biostatystyce.

LinkedIn Learning – Excel for Biostatistics
https://www.linkedin.com/learning
Ten kurs online koncentruje się na wykorzystaniu Excela do celów biostatystycznych, w tym analizie danych, tworzeniu tabel przestawnych i wizualizacji danych.

Każde z powyższych źródeł stanowi wartościowy punkt wyjścia dla osób pragnących rozwijać swoje umiejętności i wiedzę na temat korzystania z Excela w biostatystyce. Jednak najważniejsze jest praktyczne zastosowanie zdobytej wiedzy, co umożliwi pogłębianie umiejętności i doskonalenie technik.

Krystyna Laszki-Szcząchor

5 Sortowanie i filtrowanie bazy danych w Excelu

Przeprowadzenie tych czynności pokazane zostanie na bazie 57 noworodków utworzonej w Excelu, którym skontrolowano dodatkowo, oprócz standardowych parametrów, parametry słuchu. Poniżej legenda bazy oraz przykładowa baza danych (➥ **ryc. 5.1**).

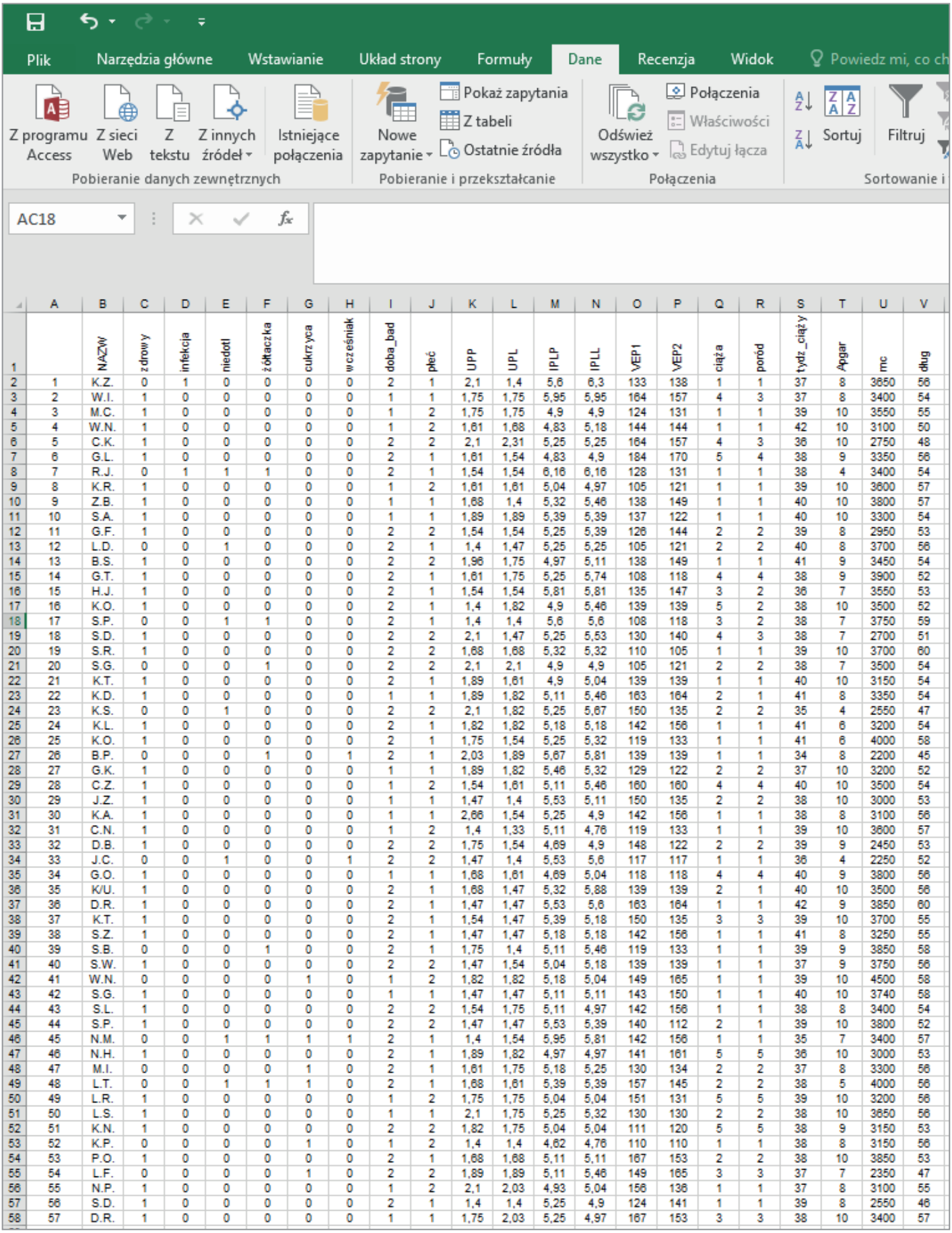
Legenda do bazy

NAZW – nazwisko pacjentki
zdrowy – 1: noworodek urodził się zdrowy, 0: noworodek urodził się chory
infekcja – 1: wystąpiła infekcja, 0: bez infekcji
niedotl – 1: wystąpiło niedotlenienie, 0: brak objawów niedotlenienia
żółtaczka – 1: urodził się z żółtaczką, 0: urodził się bez żółtaczki
cukrzyca – 1: urodził się z cukrzycą, 0: urodził się bez cukrzycy
wcześniak – 1: urodził się jako wcześniak, 0: urodzony o czasie
doba_bad – doba badania
płeć – 1: dziewczynka, 2: chłopiec
UPP – czas wystąpienia i załamka w odpowiedzi słuchowej z ucha prawego [ms]
UPL – czas wystąpienia i załamka w odpowiedzi słuchowej z ucha lewego [ms]
IPLP – czas pomiędzy załamkami I–V w odpowiedzi słuchowej z ucha prawego [ms]
IPL – czas pomiędzy załamkami I–V w odpowiedzi słuchowej z ucha lewego [ms]
VEP1 – czas odpowiedzi na bodziec wzrokowy z oka prawego [ms]
VEP2 – czas odpowiedzi na bodziec wzrokowy z oka lewego [ms]
ciąża – która ciąża
poród – który poród
tydz. ciąży – w którym tygodniu ciąży nastąpił poród
Apgar – skala Apgar: służy do oceny ogólnego stanu noworodka w pierwszych minutach życia
mc – masa ciała noworodka
dług – długość noworodka

Przedstawiona w poniższym arkuszu (➥ **ryc. 5.1**) grupa 57 noworodków została poddana badaniom słuchu i wzroku na podstawie próby potencjałów wywołanych. Przy użyciu specjalnej celowanej aparatury komputerowej u noworodków zmierzono czas ich odpowiedzi na podawane bodźce słuchowe i wzrokowe, odczytując pomiary czasu, po którym wystąpiła reakcja na bodziec. Jest to metoda określająca, czy noworodek ma prawidłowo wykształcone drogi słuchowe i wzrokowe, co jest istotne w jego dalszym rozwoju. Odpowiedzi te zostały oznaczone w tabeli symbolami UPL, UPP, IPLL, IPLP, VEP1, VEP2 (objaśnionymi powyżej w legendzie do bazy).

Żeby pokazać efekty działania funkcji sortowania i filtrowania bazy, prześledzimy, jak pracuje z nią np. lekarz neonatolog. Najważniejszą dla niego informacją jest odpowiedź na pytanie, czy noworodek urodził się zdrowy. W naszej przykładowej bazie wartości kolumny „zdrowy" są bardzo wymieszane. Łatwiej będzie lekarzowi ją przeglądać, kiedy zostaną one posortowane rosnąco wg parametru „zdrowy".

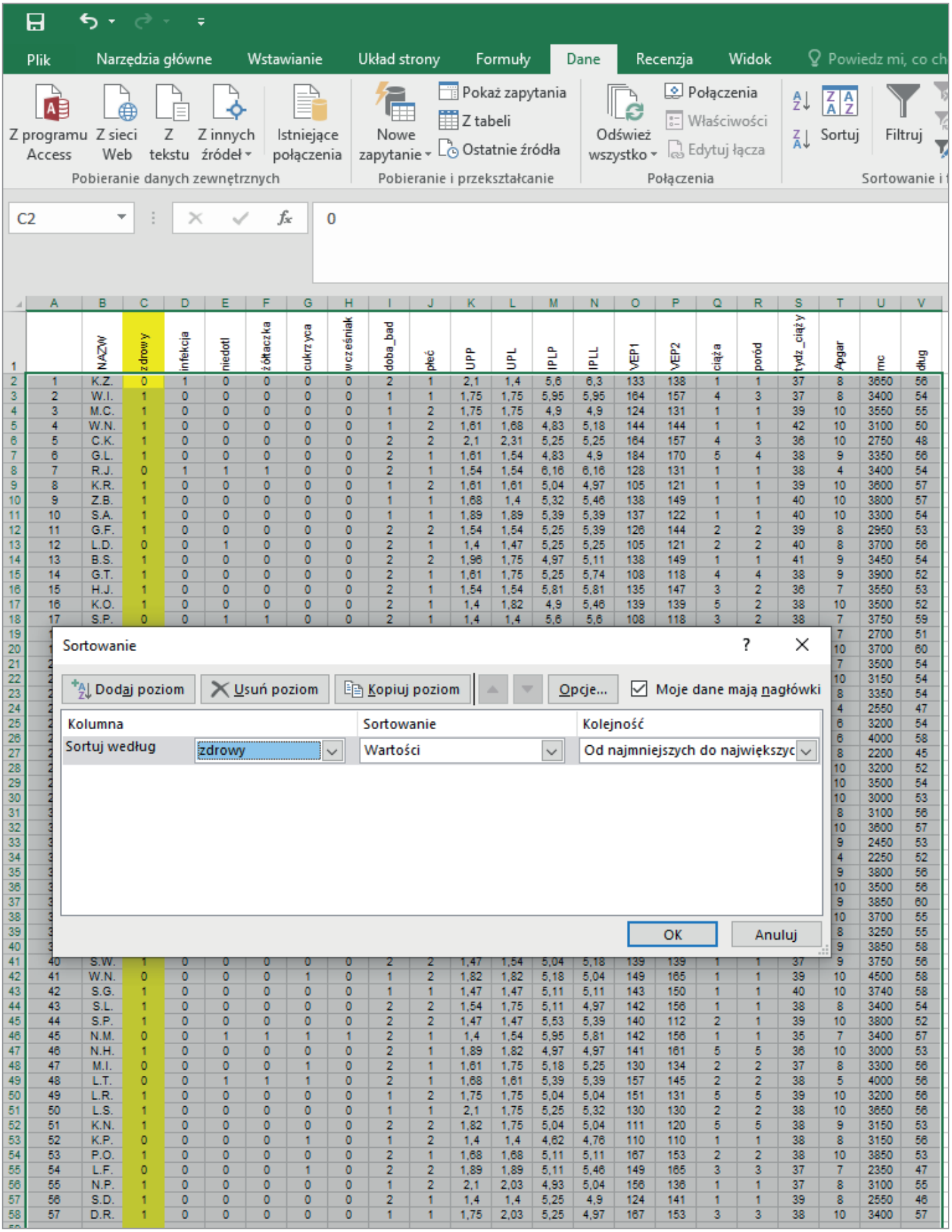
Ryc. 5.1. Przykładowa baza noworodków



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		NAZW	zdrowy	infekcja	niedotl	żółtaczka	cukrzyca	wcześniak	doba_bad	płeć	UPP	UPL	IPLP	IPLL	VEP1	VEP2	ciąża	poród	tydz_ciąży	Apgar	mc	dług
2	1	K.Z.	0	1	0	0	0	0	2	1	2,1	1,4	5,6	6,3	133	138	1	1	37	8	3650	56
3	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
4	3	M.C.	1	0	0	0	0	0	1	2	1,75	1,75	4,9	4,9	124	131	1	1	39	10	3550	55
5	4	W.N.	1	0	0	0	0	0	1	2	1,61	1,68	4,83	5,18	144	144	1	1	42	10	3100	50
6	5	C.K.	1	0	0	0	0	0	2	2	2,1	2,31	5,25	5,25	164	157	4	3	36	10	2750	48
7	6	G.L.	1	0	0	0	0	0	2	1	1,61	1,54	4,83	4,9	184	170	5	4	38	9	3350	56
8	7	R.J.	0	1	1	1	0	0	2	1	1,54	1,54	6,16	6,16	128	131	1	1	38	4	3400	54
9	8	K.R.	1	0	0	0	0	0	1	2	1,61	1,61	5,04	4,97	105	121	1	1	39	10	3600	57
10	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
11	10	S.A.	1	0	0	0	0	0	1	1	1,89	1,89	5,39	5,39	137	122	1	1	40	10	3300	54
12	11	G.F.	1	0	0	0	0	0	2	2	1,54	1,54	5,25	5,39	126	144	2	2	39	8	2950	53
13	12	L.D.	0	0	1	0	0	0	2	1	1,4	1,47	5,25	5,25	105	121	2	2	40	8	3700	56
14	13	B.S.	1	0	0	0	0	0	2	2	1,96	1,75	4,97	5,11	138	149	1	1	41	9	3450	54
15	14	G.T.	1	0	0	0	0	0	2	1	1,61	1,75	5,25	5,74	108	118	4	4	38	9	3900	52
16	15	H.J.	1	0	0	0	0	0	2	1	1,54	1,54	5,81	5,81	135	147	3	2	36	7	3550	53
17	16	K.O.	1	0	0	0	0	0	2	1	1,4	1,82	4,9	5,46	139	139	5	2	38	10	3500	52
18	17	S.P.	0	0	1	1	0	0	2	1	1,4	1,4	5,6	5,6	108	118	3	2	38	7	3750	59
19	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
20	19	S.R.	1	0	0	0	0	0	2	2	1,68	1,68	5,32	5,32	110	105	1	1	39	10	3700	60
21	20	S.G.	0	0	0	1	0	0	2	2	2,1	2,1	4,9	4,9	105	121	2	2	38	7	3500	54
22	21	K.T.	1	0	0	0	0	0	2	1	1,89	1,61	4,9	5,04	139	139	1	1	40	10	3150	54
23	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
24	23	K.S.	0	0	1	0	0	0	2	2	2,1	1,82	5,25	5,67	150	135	2	2	35	4	2550	47
25	24	K.L.	1	0	0	0	0	0	2	1	1,82	1,82	5,18	5,18	142	156	1	1	41	6	3200	54
26	25	K.O.	1	0	0	0	0	0	2	1	1,75	1,54	5,25	5,32	119	133	1	1	41	6	4000	58
27	26	B.P.	0	0	0	1	0	1	2	1	2,03	1,89	5,67	5,81	139	139	1	1	34	8	2200	45
28	27	G.K.	1	0	0	0	0	0	1	1	1,89	1,82	5,46	5,32	129	122	2	2	37	10	3200	52
29	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
30	29	J.Z.	1	0	0	0	0	0	1	1	1,47	1,4	5,53	5,11	150	135	2	2	38	10	3000	53
31	30	K.A.	1	0	0	0	0	0	1	1	2,66	1,54	5,25	4,9	142	156	1	1	38	8	3100	56
32	31	C.N.	1	0	0	0	0	0	1	2	1,4	1,33	5,11	4,76	119	133	1	1	39	10	3600	57
33	32	D.B.	1	0	0	0	0	0	2	2	1,75	1,54	4,69	4,9	148	122	2	2	39	9	2450	53
34	33	J.C.	0	0	1	0	0	1	2	2	1,47	1,4	5,53	5,6	117	117	1	1	36	4	2250	52
35	34	G.O.	1	0	0	0	0	0	1	1	1,68	1,61	4,69	5,04	118	118	4	4	40	9	3800	56
36	35	K/U.	1	0	0	0	0	0	2	1	1,68	1,47	5,32	5,88	139	139	2	1	40	10	3500	56
37	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
38	37	K.T.	1	0	0	0	0	0	2	1	1,54	1,47	5,39	5,18	150	135	3	3	39	10	3700	55
39	38	S.Z.	1	0	0	0	0	0	2	1	1,47	1,47	5,18	5,18	142	156	1	1	41	8	3250	55
40	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
41	40	S.W.	1	0	0	0	0	0	2	2	1,47	1,54	5,04	5,18	139	139	1	1	37	9	3750	56
42	41	W.N.	0	0	0	0	1	0	1	2	1,82	1,82	5,18	5,04	149	165	1	1	39	10	4500	58
43	42	S.G.	1	0	0	0	0	0	1	1	1,47	1,47	5,11	5,11	143	150	1	1	40	10	3740	58
44	43	S.L.	1	0	0	0	0	0	2	2	1,54	1,75	5,11	4,97	142	156	1	1	38	8	3400	54
45	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
46	45	N.M.	0	0	1	1	1	1	2	1	1,4	1,54	5,95	5,81	142	156	1	1	35	7	3400	57
47	46	N.H.	1	0	0	0	0	0	2	1	1,89	1,82	4,97	4,97	141	161	5	5	36	10	3000	53
48	47	M.I.	0	0	0	0	1	0	2	1	1,61	1,75	5,18	5,25	130	134	2	2	37	8	3300	56
49	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
50	49	L.R.	1	0	0	0	0	0	1	2	1,75	1,75	5,04	5,04	151	131	5	5	39	10	3200	56
51	50	L.S.	1	0	0	0	0	0	1	1	2,1	1,75	5,25	5,32	130	130	2	2	38	10	3650	56
52	51	K.N.	1	0	0	0	0	0	2	2	1,82	1,75	5,04	5,04	111	120	5	5	38	9	3150	53
53	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
54	53	P.O.	1	0	0	0	0	0	2	1	1,68	1,68	5,11	5,11	167	153	2	2	38	10	3850	53
55	54	L.F.	0	0	0	0	1	0	2	2	1,89	1,89	5,11	5,46	149	165	3	3	37	7	2350	47
56	55	N.P.	1	0	0	0	0	0	1	2	2,1	2,03	4,93	5,04	156	136	1	1	37	8	3100	55
57	56	S.D.	1	0	0	0	0	0	2	1	1,4	1,4	5,25	4,9	124	141	1	1	39	8	2550	46
58	57	D.R.	1	0	0	0	0	0	1	1	1,75	2,03	5,25	4,97	167	153	3	3	38	10	3400	57

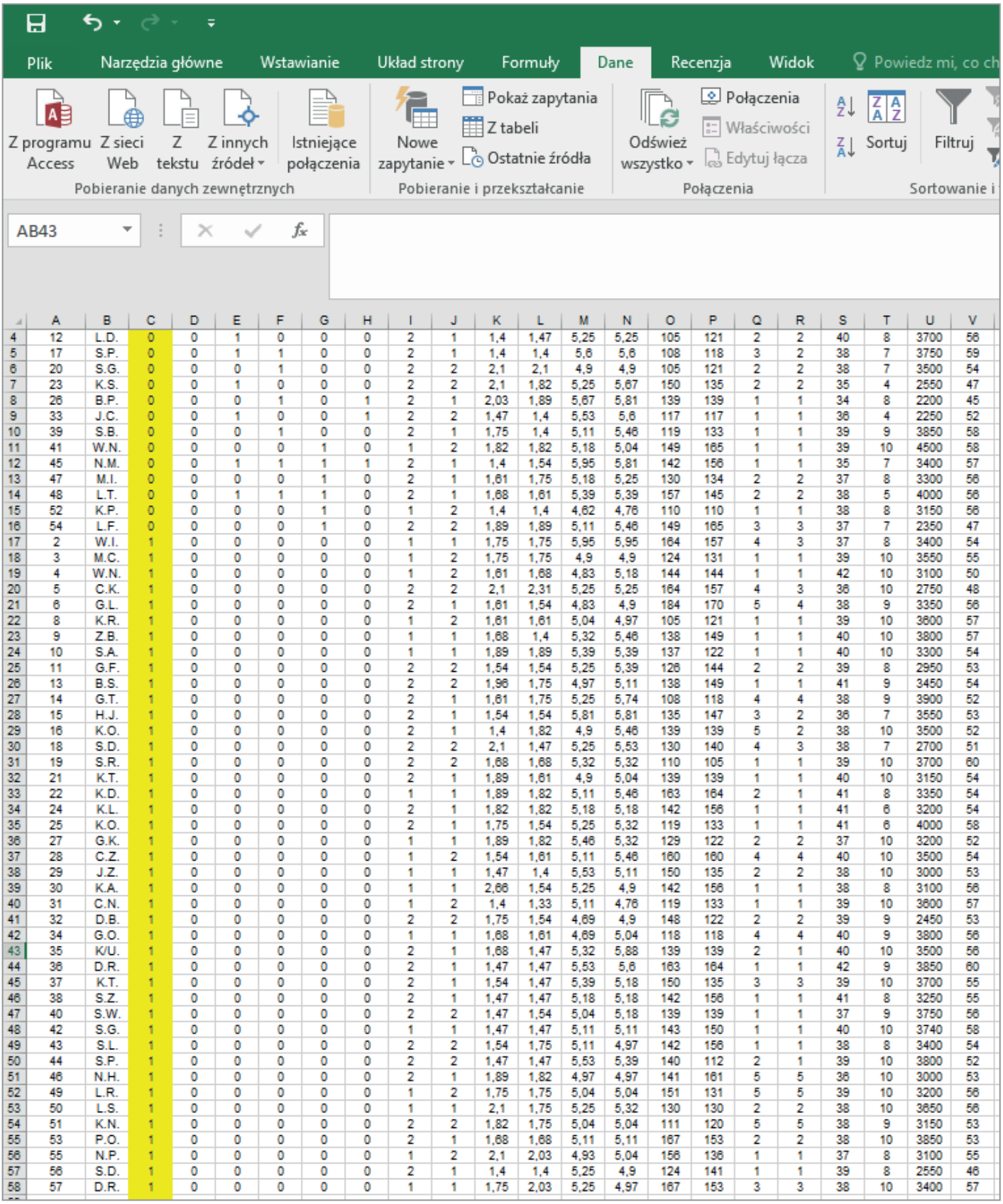
Najpierw zaznaczamy całą bazę z opisami, następnie klikamy kartę Dane na wstążce, a następnie polecenie Sortuj. Pojawia się okno dialogowe Sortowanie (➥ **ryc. 5.2**), gdzie w pozycji Sortuj według wybieramy kolumnę „zdrowy", a w pozycji Kolejność – od najmniejszych do największych.

Ryc. 5.2. Po wciśnięciu przycisku OK nastąpi sortowanie



W wybranej kolumnie ustawią się nam najpierw wartości 0 (noworodki chore), a następnie wartości 1 (noworodki zdrowe). Wyraźnie widoczny jest efekt sortowania (➡ **ryc. 5.3**). Dla neonatologa jest to układ bardzo wygodny, ponieważ noworodki chore wymagają intensywniejszej obserwacji i łatwiej je będzie wybrać z całej bazy.

Ryc. 5.3. Baza noworodków po sortowaniu wg parametru „zdrowy"

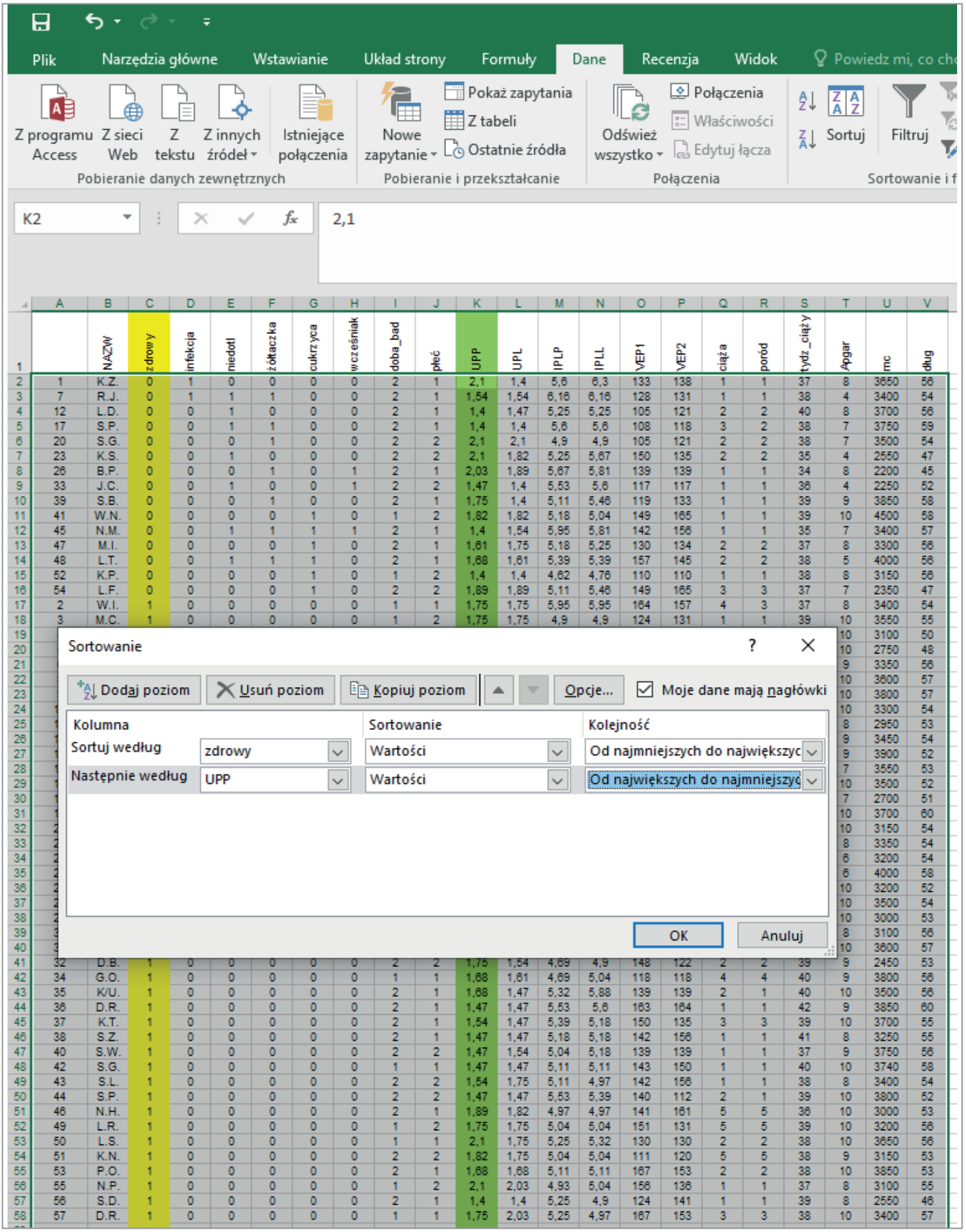


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
4	12	L.D.	0	0	1	0	0	0	2	1	1,4	1,47	5,25	5,25	105	121	2	2	40	8	3700	56
5	17	S.P.	0	0	1	1	0	0	2	1	1,4	1,4	5,6	5,6	108	118	3	2	38	7	3750	59
6	20	S.G.	0	0	0	1	0	0	2	2	2,1	2,1	4,9	4,9	105	121	2	2	38	7	3500	54
7	23	K.S.	0	0	1	0	0	0	2	2	2,1	1,82	5,25	5,67	150	135	2	2	35	4	2550	47
8	26	B.P.	0	0	0	1	0	1	2	1	2,03	1,89	5,67	5,81	139	139	1	1	34	8	2200	45
9	33	J.C.	0	0	1	0	0	1	2	2	1,47	1,4	5,53	5,6	117	117	1	1	36	4	2250	52
10	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
11	41	W.N.	0	0	0	0	1	0	1	2	1,82	1,82	5,18	5,04	149	165	1	1	39	10	4500	58
12	45	N.M.	0	0	1	1	1	1	2	1	1,4	1,54	5,95	5,81	142	156	1	1	35	7	3400	57
13	47	M.I.	0	0	0	0	1	0	2	1	1,61	1,75	5,18	5,25	130	134	2	2	37	8	3300	56
14	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
15	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
16	54	L.F.	0	0	0	0	1	0	2	2	1,89	1,89	5,11	5,46	149	165	3	3	37	7	2350	47
17	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
18	3	M.C.	1	0	0	0	0	0	1	2	1,75	1,75	4,9	4,9	124	131	1	1	39	10	3550	55
19	4	W.N.	1	0	0	0	0	0	1	2	1,61	1,68	4,83	5,18	144	144	1	1	42	10	3100	50
20	5	C.K.	1	0	0	0	0	0	2	2	2,1	2,31	5,25	5,25	164	157	4	3	36	10	2750	48
21	6	G.L.	1	0	0	0	0	0	2	1	1,61	1,54	4,83	4,9	184	170	5	4	38	9	3350	56
22	8	K.R.	1	0	0	0	0	0	1	2	1,61	1,61	5,04	4,97	105	121	1	1	39	10	3600	57
23	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
24	10	S.A.	1	0	0	0	0	0	1	1	1,89	1,89	5,39	5,39	137	122	1	1	40	10	3300	54
25	11	G.F.	1	0	0	0	0	0	2	2	1,54	1,54	5,25	5,39	126	144	2	2	39	8	2950	53
26	13	B.S.	1	0	0	0	0	0	2	2	1,96	1,75	4,97	5,11	138	149	1	1	41	9	3450	54
27	14	G.T.	1	0	0	0	0	0	2	1	1,61	1,75	5,25	5,74	108	118	4	4	38	9	3900	52
28	15	H.J.	1	0	0	0	0	0	2	1	1,54	1,54	5,81	5,81	135	147	3	2	36	7	3550	53
29	16	K.O.	1	0	0	0	0	0	2	1	1,4	1,82	4,9	5,46	139	139	5	2	38	10	3500	52
30	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
31	19	S.R.	1	0	0	0	0	0	2	2	1,68	1,68	5,32	5,32	110	105	1	1	39	10	3700	60
32	21	K.T.	1	0	0	0	0	0	2	1	1,89	1,61	4,9	5,04	139	139	1	1	40	10	3150	54
33	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
34	24	K.L.	1	0	0	0	0	0	2	1	1,82	1,82	5,18	5,18	142	156	1	1	41	6	3200	54
35	25	K.O.	1	0	0	0	0	0	2	1	1,75	1,54	5,25	5,32	119	133	1	1	41	6	4000	58
36	27	G.K.	1	0	0	0	0	0	1	1	1,89	1,82	5,46	5,32	129	122	2	2	37	10	3200	52
37	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
38	29	J.Z.	1	0	0	0	0	0	1	1	1,47	1,4	5,53	5,11	150	135	2	2	38	10	3000	53
39	30	K.A.	1	0	0	0	0	0	1	1	2,66	1,54	5,25	4,9	142	156	1	1	38	8	3100	56
40	31	C.N.	1	0	0	0	0	0	1	2	1,4	1,33	5,11	4,76	119	133	1	1	39	10	3600	57
41	32	D.B.	1	0	0	0	0	0	2	2	1,75	1,54	4,69	4,9	148	122	2	2	39	9	2450	53
42	34	G.O.	1	0	0	0	0	0	1	1	1,68	1,61	4,69	5,04	118	118	4	4	40	9	3800	56
43	35	K/U.	1	0	0	0	0	0	2	1	1,68	1,47	5,32	5,88	139	139	2	1	40	10	3500	56
44	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
45	37	K.T.	1	0	0	0	0	0	2	1	1,54	1,47	5,39	5,18	150	135	3	3	39	10	3700	55
46	38	S.Z.	1	0	0	0	0	0	2	1	1,47	1,47	5,18	5,18	142	156	1	1	41	8	3250	55
47	40	S.W.	1	0	0	0	0	0	2	2	1,47	1,54	5,04	5,18	139	139	1	1	37	9	3750	56
48	42	S.G.	1	0	0	0	0	0	1	1	1,47	1,47	5,11	5,11	143	150	1	1	40	10	3740	58
49	43	S.L.	1	0	0	0	0	0	2	2	1,54	1,75	5,11	4,97	142	156	1	1	38	8	3400	54
50	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
51	46	N.H.	1	0	0	0	0	0	2	1	1,89	1,82	4,97	4,97	141	161	5	5	36	10	3000	53
52	49	L.R.	1	0	0	0	0	0	1	2	1,75	1,75	5,04	5,04	151	131	5	5	39	10	3200	56
53	50	L.S.	1	0	0	0	0	0	1	1	2,1	1,75	5,25	5,32	130	130	2	2	38	10	3650	56
54	51	K.N.	1	0	0	0	0	0	2	2	1,82	1,75	5,04	5,04	111	120	5	5	38	9	3150	53
55	53	P.O.	1	0	0	0	0	0	2	1	1,68	1,68	5,11	5,11	167	153	2	2	38	10	3850	53
56	55	N.P.	1	0	0	0	0	0	1	2	2,1	2,03	4,93	5,04	156	136	1	1	37	8	3100	55
57	56	S.D.	1	0	0	0	0	0	2	1	1,4	1,4	5,25	4,9	124	141	1	1	39	8	2550	46
58	57	D.R.	1	0	0	0	0	0	1	1	1,75	2,03	5,25	4,97	167	153	3	3	38	10	3400	57

Ryc. 5.4. Przygotowanie bazy do sortowania wg parametru słuchowego „UPP" w układzie od wartości największych do najmniejszych (mniejsze wartości oznaczają pewność poprawności pracy drogi słuchowej)

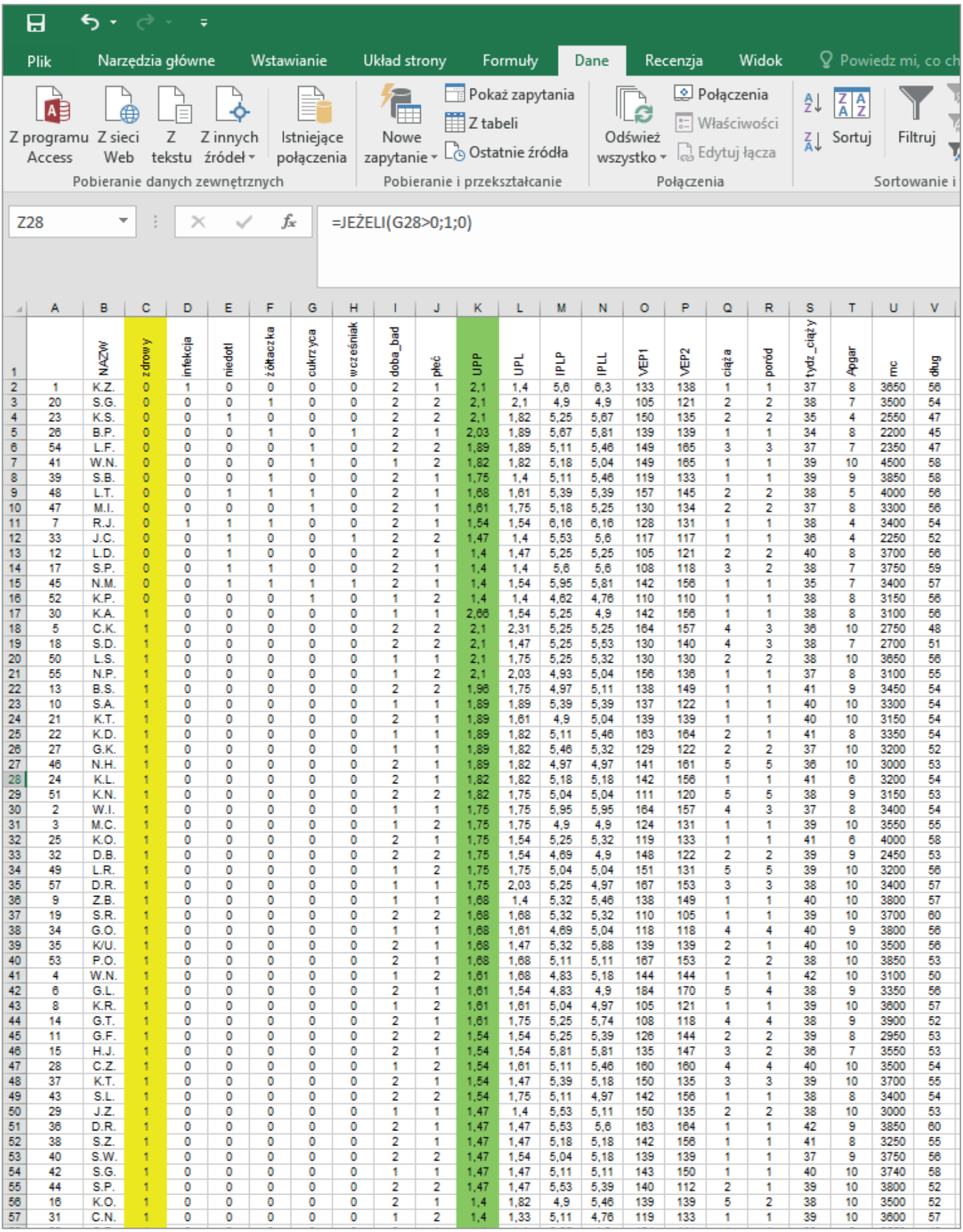
➡ **Rycina 5.4** przedstawia przygotowanie do sortowania bazy wg parametru słuchowego „UPP" od wartości największych do najmniejszych (im wartość jest mniejsza, tym praca drogi słuchowej jest bardziej poprawna).

Wartości parametru „zdrowy" nadal zostają na poziomie pierwszego sortowania. Oznacza to, że otrzymamy układ, gdzie parametry będą rozmieszczone od wartości najwięk-



Ryc. 5.5. Efekt sortowania wg parametru słuchowego „UPP" w układzie od wartości największych do najmniejszych na drugim poziomie. Pierwszy przesortowany poziom zdrowy pozostaje niezmieniony

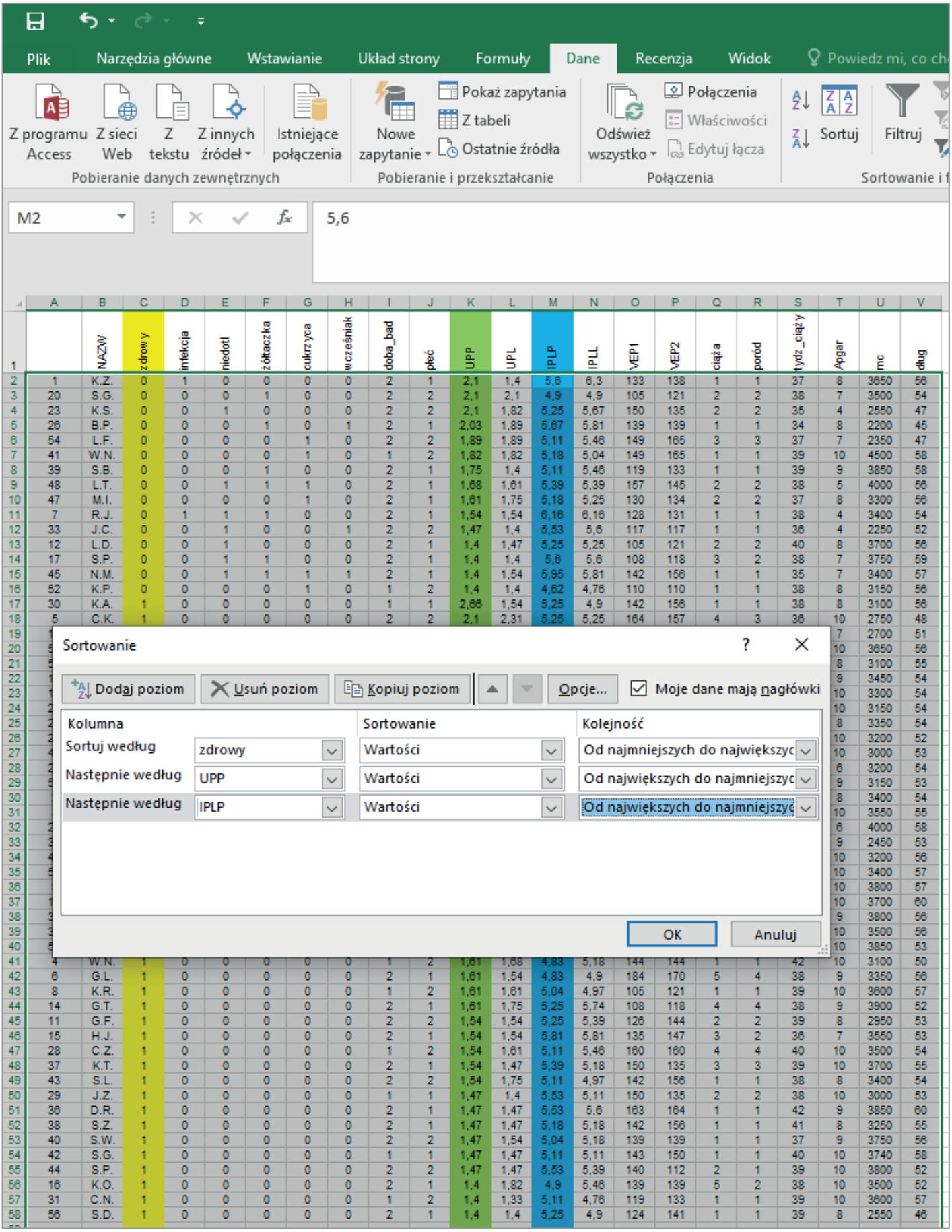
szych do najmniejszych najpierw dla grupy noworodków chorych. Niżej widoczny będzie rozkład od najwyższych do najniższych pomiarów wartości UPP, ale dla noworodków zdrowych. Na ➥ **ryc. 5.5** i ➥ **ryc. 5.6** w sposób bardzo wyraźny pokazany jest efekt sortowania danych bazy wg dwóch opisanych powyżej parametrów w nadanej im kolejności. Sukcesywne zmniejszanie się wartości w kolumnie UPP jest bardzo czytelne.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		NAZW	zdrowy	infekcja	niedotl	żółtaczka	cukrzyca	wcześniak	doba_bad	płeć	UPP	UPL	IPLP	IPLL	VEP1	VEP2	ciąża	poród	tydz_ciąży	Apgar	mc	dług
2	1	K.Z.	0	1	0	0	0	0	2	1	2,1	1,4	5,6	6,3	133	138	1	1	37	8	3650	56
3	20	S.G.	0	0	0	1	0	0	2	2	2,1	2,1	4,9	4,9	105	121	2	2	38	7	3500	54
4	23	K.S.	0	0	1	0	0	0	2	2	2,1	1,82	5,25	5,67	150	135	2	2	35	4	2550	47
5	26	B.P.	0	0	0	1	0	1	2	1	2,03	1,89	5,67	5,81	139	139	1	1	34	8	2200	45
6	54	L.F.	0	0	0	0	1	0	2	2	1,89	1,89	5,11	5,46	149	165	3	3	37	7	2350	47
7	41	W.N.	0	0	0	0	1	0	1	2	1,82	1,82	5,18	5,04	149	165	1	1	39	10	4500	58
8	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
9	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
10	47	M.I.	0	0	0	0	1	0	2	1	1,61	1,75	5,18	5,25	130	134	2	2	37	8	3300	56
11	7	R.J.	0	1	1	1	0	0	2	1	1,54	1,54	6,16	6,16	128	131	1	1	38	4	3400	54
12	33	J.C.	0	0	1	0	0	1	2	2	1,47	1,4	5,53	5,6	117	117	1	1	36	4	2250	52
13	12	L.D.	0	0	1	0	0	0	2	1	1,4	1,47	5,25	5,25	105	121	2	2	40	8	3700	56
14	17	S.P.	0	0	1	1	0	0	2	1	1,4	1,4	5,6	5,6	108	118	3	2	38	7	3750	59
15	45	N.M.	0	0	1	1	1	1	2	1	1,4	1,54	5,95	5,81	142	156	1	1	35	7	3400	57
16	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
17	30	K.A.	1	0	0	0	0	0	1	1	2,66	1,54	5,25	4,9	142	156	1	1	38	8	3100	56
18	5	C.K.	1	0	0	0	0	0	2	2	2,1	2,31	5,25	5,25	164	157	4	3	36	10	2750	48
19	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
20	50	L.S.	1	0	0	0	0	0	1	1	2,1	1,75	5,25	5,32	130	130	2	2	38	10	3650	56
21	55	N.P.	1	0	0	0	0	0	1	2	2,1	2,03	4,93	5,04	156	136	1	1	37	8	3100	55
22	13	B.S.	1	0	0	0	0	0	2	2	1,96	1,75	4,97	5,11	138	149	1	1	41	9	3450	54
23	10	S.A.	1	0	0	0	0	0	1	1	1,89	1,89	5,39	5,39	137	122	1	1	40	10	3300	54
24	21	K.T.	1	0	0	0	0	0	2	1	1,89	1,61	4,9	5,04	139	139	1	1	40	10	3150	54
25	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
26	27	G.K.	1	0	0	0	0	0	1	1	1,89	1,82	5,46	5,32	129	122	2	2	37	10	3200	52
27	46	N.H.	1	0	0	0	0	0	2	1	1,89	1,82	4,97	4,97	141	161	5	5	36	10	3000	53
28	24	K.L.	1	0	0	0	0	0	2	1	1,82	1,82	5,18	5,18	142	156	1	1	41	6	3200	54
29	51	K.N.	1	0	0	0	0	0	2	2	1,82	1,75	5,04	5,04	111	120	5	5	38	9	3150	53
30	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
31	3	M.C.	1	0	0	0	0	0	1	2	1,75	1,75	4,9	4,9	124	131	1	1	39	10	3550	55
32	25	K.O.	1	0	0	0	0	0	2	1	1,75	1,54	5,25	5,32	119	133	1	1	41	6	4000	58
33	32	D.B.	1	0	0	0	0	0	2	2	1,75	1,54	4,69	4,9	148	122	2	2	39	9	2450	53
34	49	L.R.	1	0	0	0	0	0	1	2	1,75	1,75	5,04	5,04	151	131	5	5	39	10	3200	56
35	57	D.R.	1	0	0	0	0	0	1	1	1,75	2,03	5,25	4,97	167	153	3	3	38	10	3400	57
36	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
37	19	S.R.	1	0	0	0	0	0	2	2	1,68	1,68	5,32	5,32	110	105	1	1	39	10	3700	60
38	34	G.O.	1	0	0	0	0	0	1	1	1,68	1,61	4,69	5,04	118	118	4	4	40	9	3800	56
39	35	K/U.	1	0	0	0	0	0	2	1	1,68	1,47	5,32	5,88	139	139	2	1	40	10	3500	56
40	53	P.O.	1	0	0	0	0	0	2	1	1,68	1,68	5,11	5,11	167	153	2	2	38	10	3850	53
41	4	W.N.	1	0	0	0	0	0	1	2	1,61	1,68	4,83	5,18	144	144	1	1	42	10	3100	50
42	6	G.L.	1	0	0	0	0	0	2	1	1,61	1,54	4,83	4,9	184	170	5	4	38	9	3350	56
43	8	K.R.	1	0	0	0	0	0	1	2	1,61	1,61	5,04	4,97	105	121	1	1	39	10	3600	57
44	14	G.T.	1	0	0	0	0	0	2	1	1,61	1,75	5,25	5,74	108	118	4	4	38	9	3900	52
45	11	G.F.	1	0	0	0	0	0	2	2	1,54	1,54	5,25	5,39	126	144	2	2	39	8	2950	53
46	15	H.J.	1	0	0	0	0	0	2	1	1,54	1,54	5,81	5,81	135	147	3	2	36	7	3550	53
47	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
48	37	K.T.	1	0	0	0	0	0	2	1	1,54	1,47	5,39	5,18	150	135	3	3	39	10	3700	55
49	43	S.L.	1	0	0	0	0	0	2	2	1,54	1,75	5,11	4,97	142	156	1	1	38	8	3400	54
50	29	J.Z.	1	0	0	0	0	0	1	1	1,47	1,4	5,53	5,11	150	135	2	2	38	10	3000	53
51	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
52	38	S.Z.	1	0	0	0	0	0	2	1	1,47	1,47	5,18	5,18	142	156	1	1	41	8	3250	55
53	40	S.W.	1	0	0	0	0	0	2	2	1,47	1,54	5,04	5,18	139	139	1	1	37	9	3750	56
54	42	S.G.	1	0	0	0	0	0	1	1	1,47	1,47	5,11	5,11	143	150	1	1	40	10	3740	58
55	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
56	16	K.O.	1	0	0	0	0	0	2	1	1,4	1,82	4,9	5,46	139	139	5	2	38	10	3500	52
57	31	C.N.	1	0	0	0	0	0	1	2	1,4	1,33	5,11	4,76	119	133	1	1	39	10	3600	57

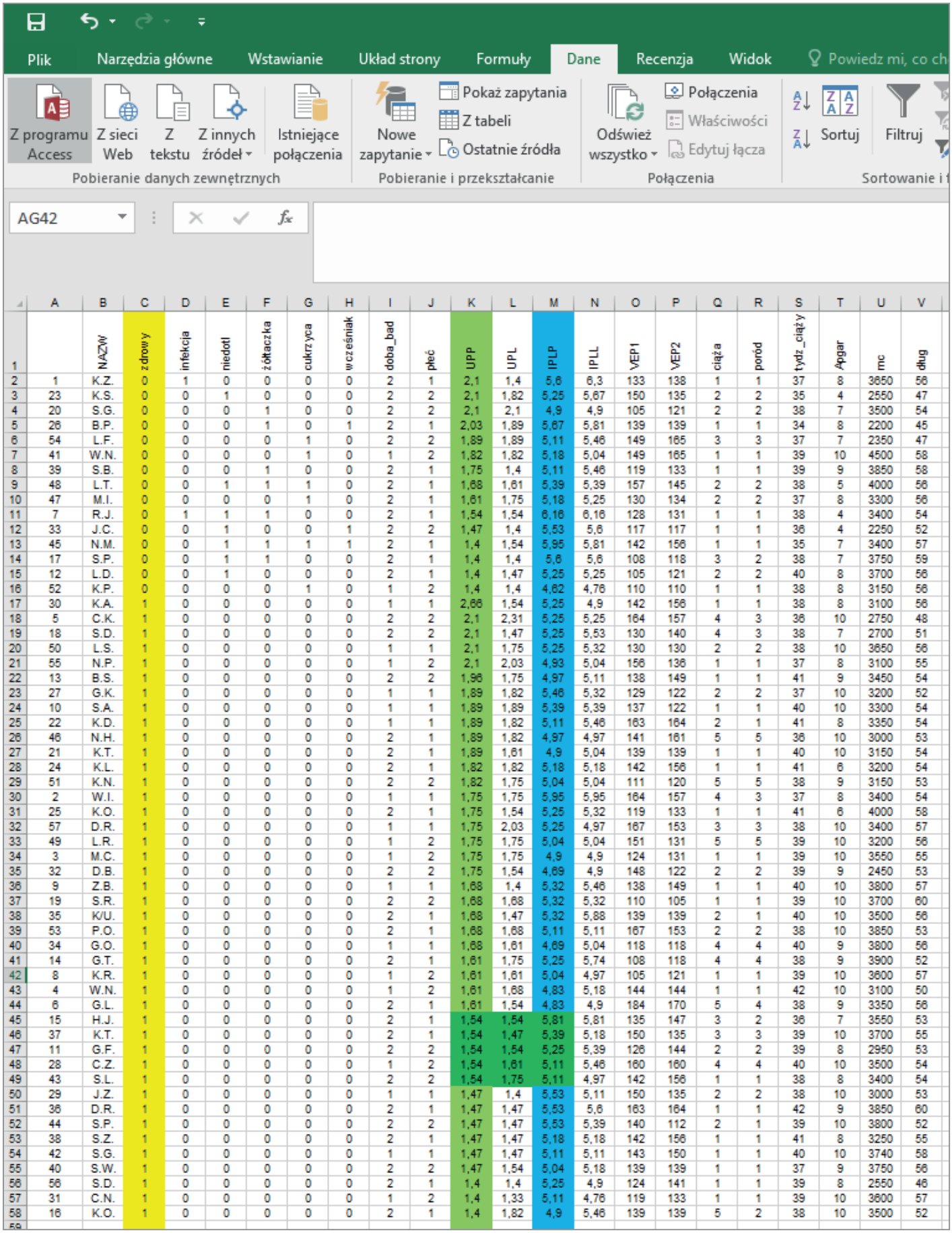
Ryc. 5.6. Przygotowanie bazy do sortowania wg drugiego parametru słuchowego „IPLP" (również dla ucha prawego) w układzie od wartości największych do najmniejszych dodanego na trzecim poziomie sortowania (mniejsze wartości oznaczają pewność poprawności pracy drogi słuchowej)

W zamieszczonej na ➥ **ryc. 5.7** bazie niestety nie widać już wyraźnie malejących wartości parametrów w kolumnie. Wynika to z dużej zmienności danych w kolumnie UPP. Efekt działania programu widać tyko na odcinkach,



Ryc. 5.7. Efekt sortowania wg parametru słuchowego „IPLP" w układzie od wartości największych do najmniejszych na 2. poziomie. Poziom zdrowy i UPP pozostają niezmienione

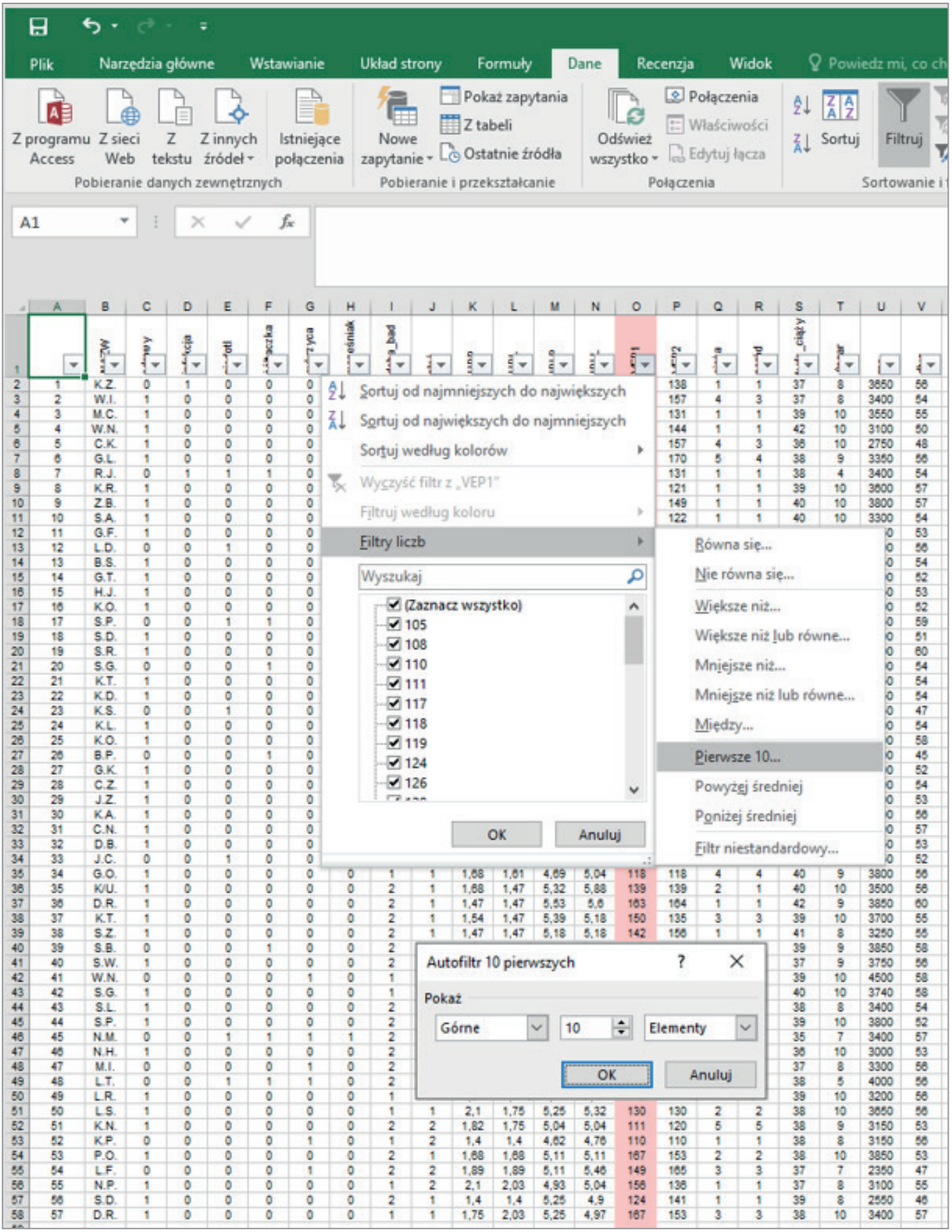
gdzie UPP ma taką samą wartość, jak np. 1,54, (ten obszar został zaznaczony na rysunku na ciemnozielono). Widać, że przy takich parametrach bazy wstawianie 3. poziomu sortowania nie daje już spektakularnych efektów. Zatem nie należy iść w tym kierunku.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		NAZW	zdrowy	infekcja	niedotl	żółtaczka	cukrzyca	wcześniak	doba_bad	płeć	UPP	UPL	IPLP	IPLL	VEP1	VEP2	ciąża	poród	tydz_ciąży	Apgar	mc	dług
2	1	K.Z.	0	1	0	0	0	0	2	1	2,1	1,4	5,6	6,3	133	138	1	1	37	8	3650	56
3	23	K.S.	0	0	1	0	0	0	2	2	2,1	1,82	5,25	5,67	150	135	2	2	35	4	2550	47
4	20	S.G.	0	0	0	1	0	0	2	2	2,1	2,1	4,9	4,9	105	121	2	2	38	7	3500	54
5	26	B.P.	0	0	0	1	0	1	2	1	2,03	1,89	5,67	5,81	139	139	1	1	34	8	2200	45
6	54	L.F.	0	0	0	0	1	0	2	2	1,89	1,89	5,11	5,46	149	165	3	3	37	7	2350	47
7	41	W.N.	0	0	0	0	1	0	1	2	1,82	1,82	5,18	5,04	149	165	1	1	39	10	4500	58
8	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
9	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
10	47	M.I.	0	0	0	0	1	0	2	1	1,61	1,75	5,18	5,25	130	134	2	2	37	8	3300	56
11	7	R.J.	0	1	1	1	0	0	2	1	1,54	1,54	6,16	6,16	128	131	1	1	38	4	3400	54
12	33	J.C.	0	0	1	0	0	1	2	2	1,47	1,4	5,53	5,6	117	117	1	1	36	4	2250	52
13	45	N.M.	0	0	1	1	1	1	2	1	1,4	1,54	5,95	5,81	142	156	1	1	35	7	3400	57
14	17	S.P.	0	0	1	1	0	0	2	1	1,4	1,4	5,6	5,6	108	118	3	2	38	7	3750	59
15	12	L.D.	0	0	1	0	0	0	2	1	1,4	1,47	5,25	5,25	105	121	2	2	40	8	3700	56
16	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
17	30	K.A.	1	0	0	0	0	0	1	1	2,66	1,54	5,25	4,9	142	156	1	1	38	8	3100	56
18	5	C.K.	1	0	0	0	0	0	2	2	2,1	2,31	5,25	5,25	164	157	4	3	36	10	2750	48
19	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
20	50	L.S.	1	0	0	0	0	0	1	1	2,1	1,75	5,25	5,32	130	130	2	2	38	10	3650	56
21	55	N.P.	1	0	0	0	0	0	1	2	2,1	2,03	4,93	5,04	156	136	1	1	37	8	3100	55
22	13	B.S.	1	0	0	0	0	0	2	2	1,96	1,75	4,97	5,11	138	149	1	1	41	9	3450	54
23	27	G.K.	1	0	0	0	0	0	1	1	1,89	1,82	5,46	5,32	129	122	2	2	37	10	3200	52
24	10	S.A.	1	0	0	0	0	0	1	1	1,89	1,89	5,39	5,39	137	122	1	1	40	10	3300	54
25	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
26	46	N.H.	1	0	0	0	0	0	2	1	1,89	1,82	4,97	4,97	141	161	5	5	36	10	3000	53
27	21	K.T.	1	0	0	0	0	0	2	1	1,89	1,61	4,9	5,04	139	139	1	1	40	10	3150	54
28	24	K.L.	1	0	0	0	0	0	2	1	1,82	1,82	5,18	5,18	142	156	1	1	41	6	3200	54
29	51	K.N.	1	0	0	0	0	0	2	2	1,82	1,75	5,04	5,04	111	120	5	5	38	9	3150	53
30	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
31	25	K.O.	1	0	0	0	0	0	2	1	1,75	1,54	5,25	5,32	119	133	1	1	41	6	4000	58
32	57	D.R.	1	0	0	0	0	0	1	1	1,75	2,03	5,25	4,97	167	153	3	3	38	10	3400	57
33	49	L.R.	1	0	0	0	0	0	1	2	1,75	1,75	5,04	5,04	151	131	5	5	39	10	3200	56
34	3	M.C.	1	0	0	0	0	0	1	2	1,75	1,75	4,9	4,9	124	131	1	1	39	10	3550	55
35	32	D.B.	1	0	0	0	0	0	2	2	1,75	1,54	4,69	4,9	148	122	2	2	39	9	2450	53
36	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
37	19	S.R.	1	0	0	0	0	0	2	2	1,68	1,68	5,32	5,32	110	105	1	1	39	10	3700	60
38	35	K/U.	1	0	0	0	0	0	2	1	1,68	1,47	5,32	5,88	139	139	2	1	40	10	3500	56
39	53	P.O.	1	0	0	0	0	0	2	1	1,68	1,68	5,11	5,11	167	153	2	2	38	10	3850	53
40	34	G.O.	1	0	0	0	0	0	1	1	1,68	1,61	4,69	5,04	118	118	4	4	40	9	3800	56
41	14	G.T.	1	0	0	0	0	0	2	1	1,61	1,75	5,25	5,74	108	118	4	4	38	9	3900	52
42	8	K.R.	1	0	0	0	0	0	1	2	1,61	1,61	5,04	4,97	105	121	1	1	39	10	3600	57
43	4	W.N.	1	0	0	0	0	0	1	2	1,61	1,68	4,83	5,18	144	144	1	1	42	10	3100	50
44	6	G.L.	1	0	0	0	0	0	2	1	1,61	1,54	4,83	4,9	184	170	5	4	38	9	3350	56
45	15	H.J.	1	0	0	0	0	0	2	1	1,54	1,54	5,81	5,81	135	147	3	2	36	7	3550	53
46	37	K.T.	1	0	0	0	0	0	2	1	1,54	1,47	5,39	5,18	150	135	3	3	39	10	3700	55
47	11	G.F.	1	0	0	0	0	0	2	2	1,54	1,54	5,25	5,39	126	144	2	2	39	8	2950	53
48	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
49	43	S.L.	1	0	0	0	0	0	2	2	1,54	1,75	5,11	4,97	142	156	1	1	38	8	3400	54
50	29	J.Z.	1	0	0	0	0	0	1	1	1,47	1,4	5,53	5,11	150	135	2	2	38	10	3000	53
51	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
52	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
53	38	S.Z.	1	0	0	0	0	0	2	1	1,47	1,47	5,18	5,18	142	156	1	1	41	8	3250	55
54	42	S.G.	1	0	0	0	0	0	1	1	1,47	1,47	5,11	5,11	143	150	1	1	40	10	3740	58
55	40	S.W.	1	0	0	0	0	0	2	2	1,47	1,54	5,04	5,18	139	139	1	1	37	9	3750	56
56	56	S.D.	1	0	0	0	0	0	2	1	1,4	1,4	5,25	4,9	124	141	1	1	39	8	2550	46
57	31	C.N.	1	0	0	0	0	0	1	2	1,4	1,33	5,11	4,76	119	133	1	1	39	10	3600	57
58	16	K.O.	1	0	0	0	0	0	2	1	1,4	1,82	4,9	5,46	139	139	5	2	38	10	3500	52

Ryc. 5.8. Wybieranie z bazy 10 najdłuższych czasów parametru wzrokowego „VEP1" dla lewego oka (możliwość niedowidzenia) w procesie filtrowania bazy

Bazę przygotowujemy do procesu filtrowania, zaznaczając na niej pojedynczą komórkę i wybierając na wstążce polecenie Filtruj (➡ **ryc. 5.8**). Na każdej kolumnie pojawiają się przyciski menu filtrowania (kwadrat z trójkątem).



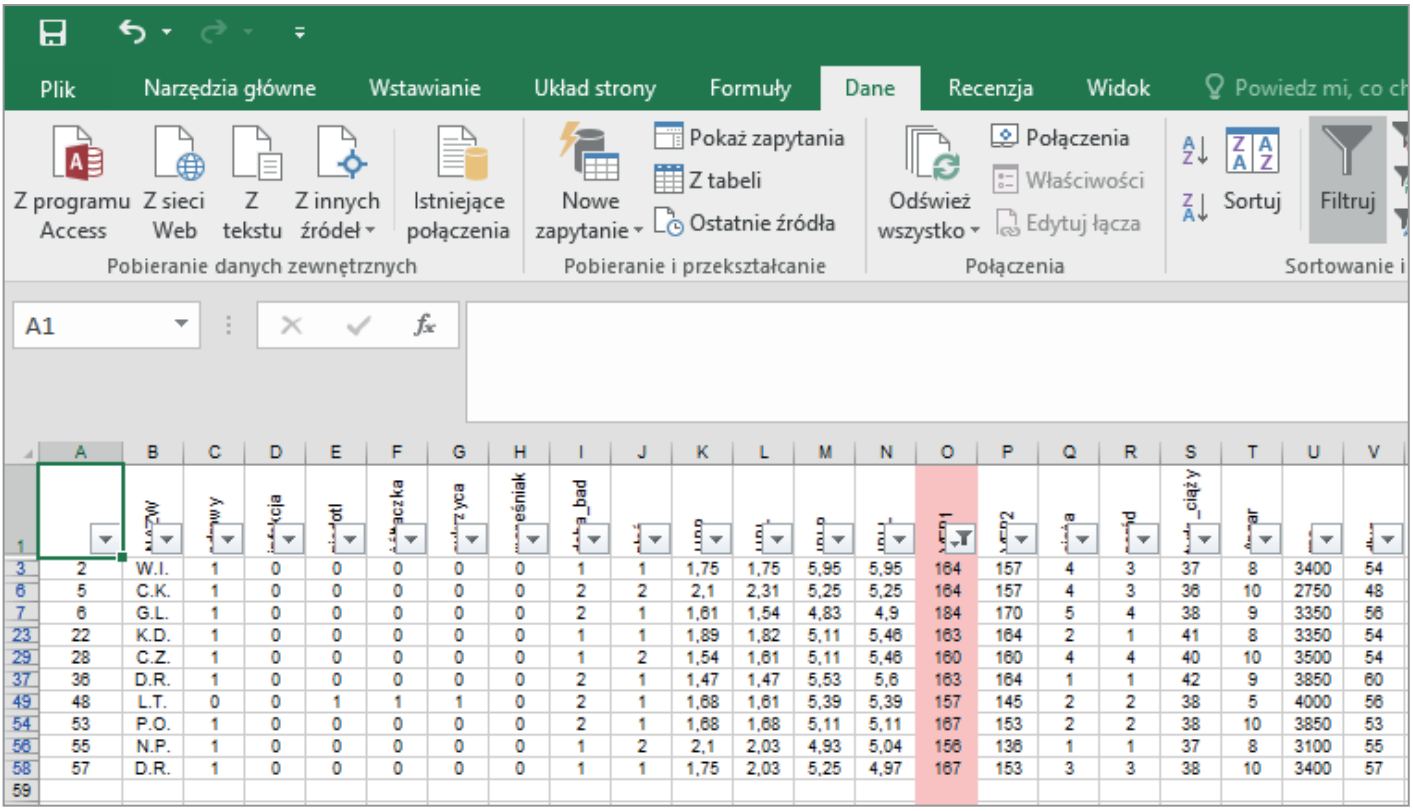
Na wybranej kolumnie VEP1 (parametr z odpowiedzi prawego oka na bodziec wzrokowy) rozwijamy menu i wybieramy z niego opcję Filtry liczb, w kolejnym rozwijanym menu zaznaczamy Pierwsze 10. Po czym w nowo otwartym okienku (u dołu ➥ **ryc. 5.8**) wybieramy kolejno Górne, 10, Elementy.

Na ➥ **ryc. 5.9** widać efekt filtrowania. W kolumnie VEP1 pozostało 10 parametrów wzrokowych z największymi wartościami czasu odpowiedzi na bodziec wzrokowy.

Na miejscu kwadratu z trójkątem w górnej części kolumny pojawił się charakterystyczny dla procesu filtrowania graficzny znak lejka. To oznacza, że odbyło się ono właśnie w tej kolumnie.

Z pozyskania parametrów filtrowania można zrezygnować, cofając ostatnio wykonywaną operację.

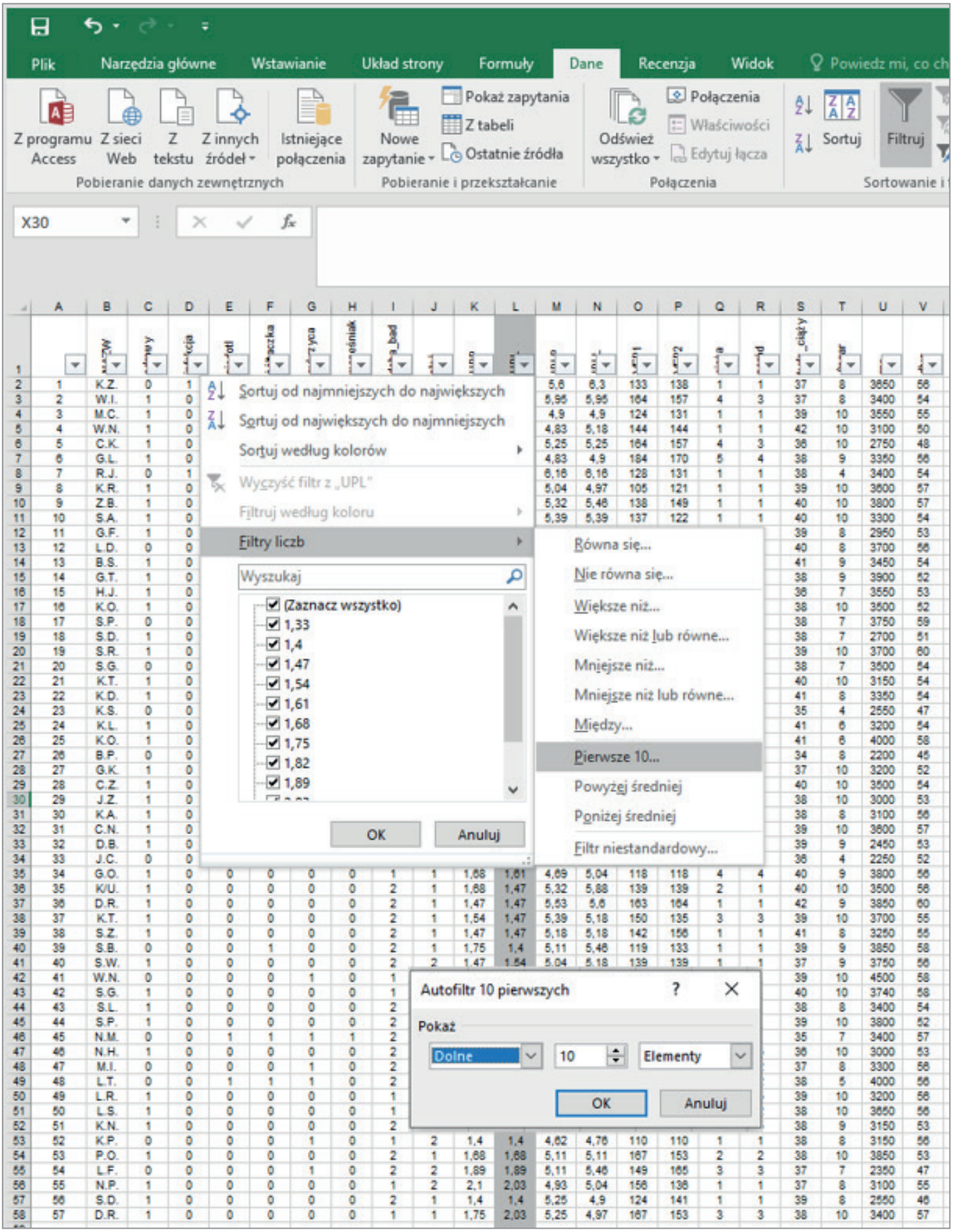
Ryc. 5.9. W kolumnie VEP1 pozostało 10 parametrów wzrokowych z największymi wartościami czasu odpowiedzi na bodziec wzrokowy



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
6	5	C.K.	1	0	0	0	0	0	2	2	2,1	2,31	5,25	5,25	164	157	4	3	36	10	2750	48
7	6	G.L.	1	0	0	0	0	0	2	1	1,61	1,54	4,83	4,9	184	170	5	4	38	9	3350	56
23	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
29	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
37	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
49	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
54	53	P.O.	1	0	0	0	0	0	2	1	1,68	1,68	5,11	5,11	167	153	2	2	38	10	3850	53
56	55	N.P.	1	0	0	0	0	0	1	2	2,1	2,03	4,93	5,04	156	136	1	1	37	8	3100	55
58	57	D.R.	1	0	0	0	0	0	1	1	1,75	2,03	5,25	4,97	167	153	3	3	38	10	3400	57
59																						

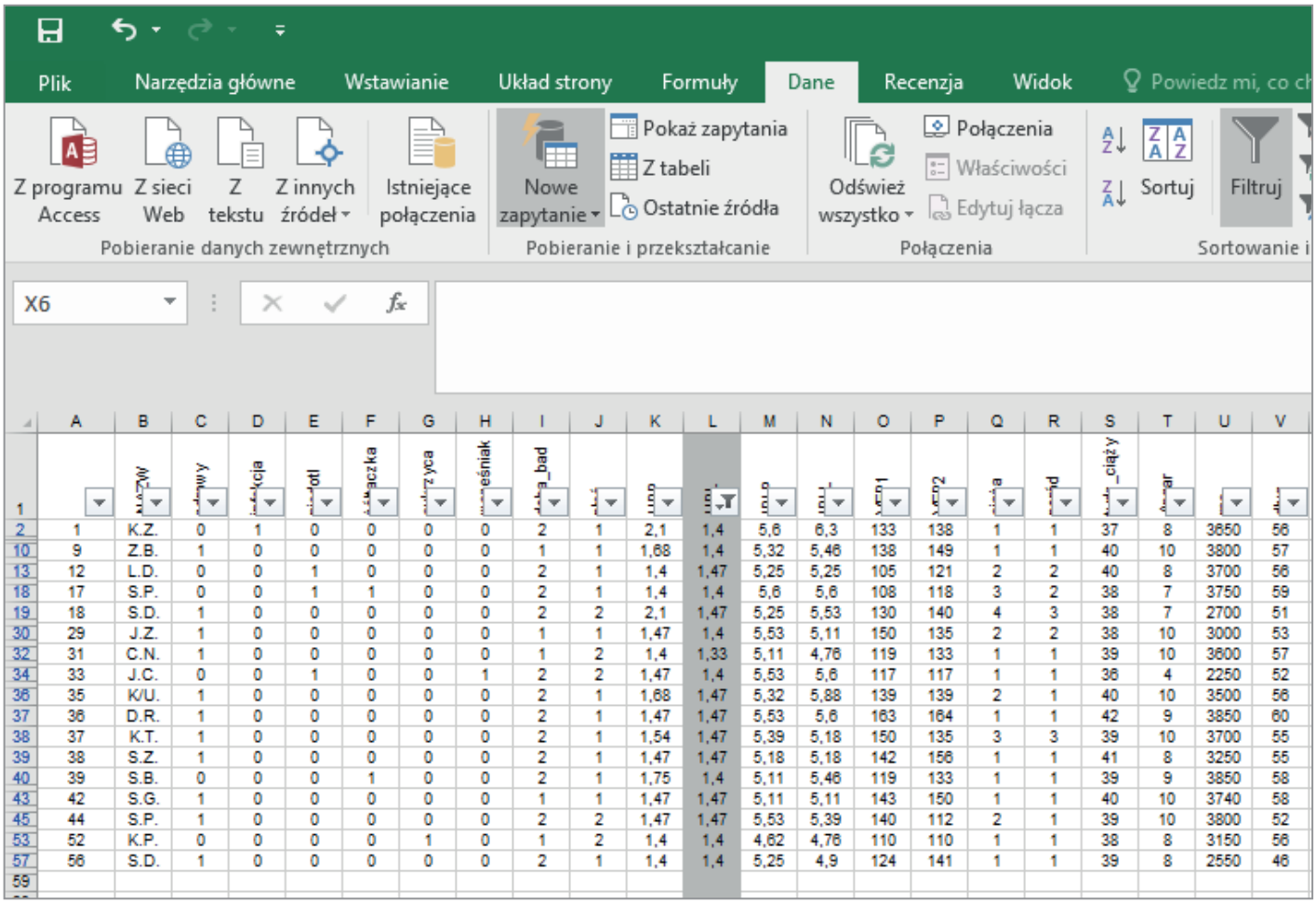
Ryc. 5.10. Filtrowanie z bazy 10 najkrótszych czasów parametru słuchowego „UPL" dla lewego ucha (najlepszy słuch)

Drugi przykład filtrowania przedstawia ➥ **ryc. 5.10**. Na wybranej kolumnie UPL (parametr z odpowiedzi z lewego ucha na bodziec słuchowy) rozwijamy menu. Wybieramy z niego opcję Filtry liczb, a z kolejnego rozwijanego menu – Pierwsze 10. Po czym w nowo otwartym okienku wybieramy kolejno: Dolne, 10, Elementy.



Ryc. 5.11. W kolumnie UPL pozostało 17 parametrów słuchowych z ucha lewego z najmniejszymi wartościami czasów odpowiedzi na bodziec słuchowy

Na ➥ **ryc. 5.11**. widać efekt filtrowania. W kolumnie UPL pozostało 17 parametrów słuchowych z ucha lewego z najmniejszymi wartościami czasu odpowiedzi na bodziec słuchowy.



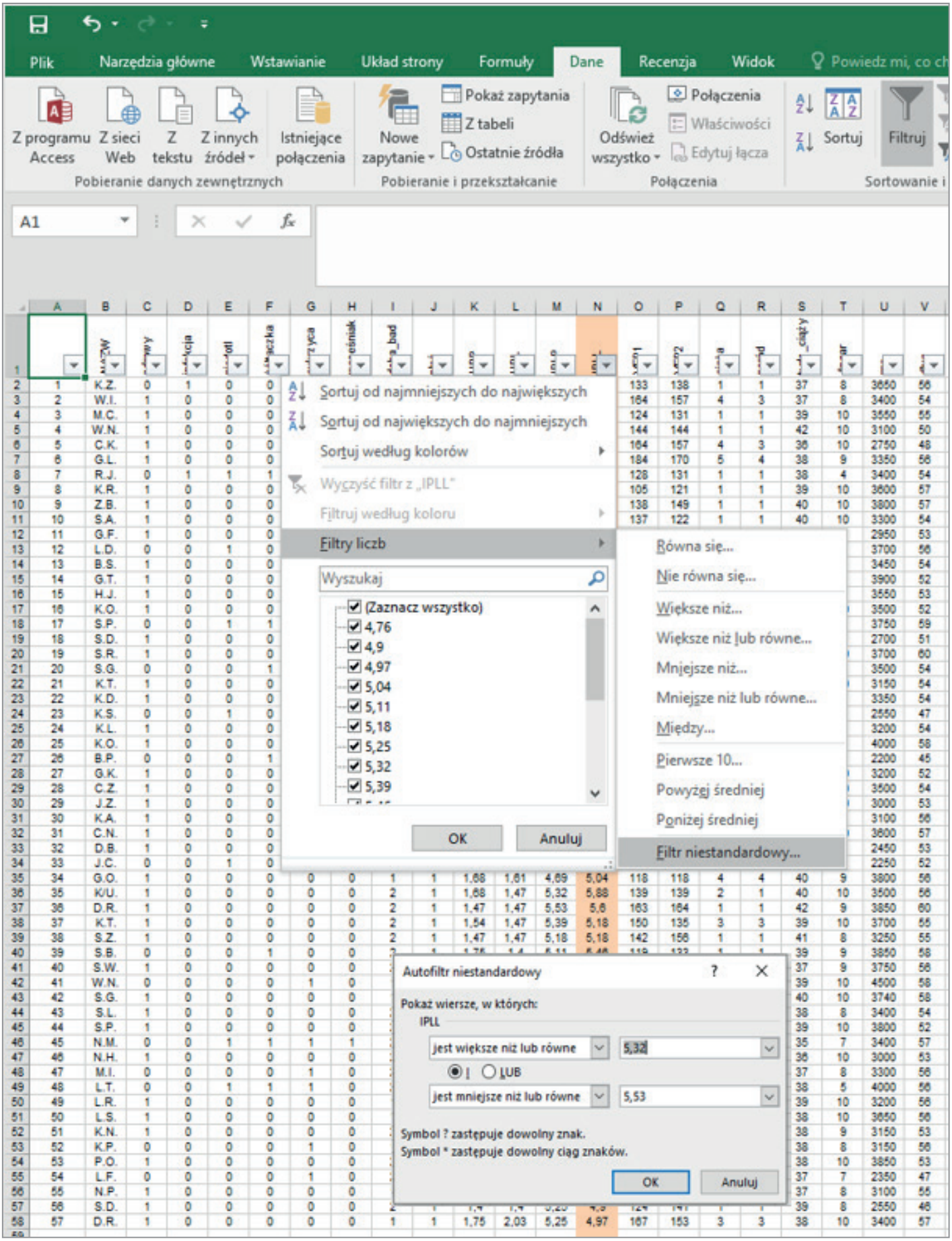
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		K.Z.	0	1	0	0	0	0	2	1	2,1	1,4	5,6	6,3	133	138	1	1	37	8	3650	56
10	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
13	12	L.D.	0	0	1	0	0	0	2	1	1,4	1,47	5,25	5,25	105	121	2	2	40	8	3700	56
18	17	S.P.	0	0	1	1	0	0	2	1	1,4	1,4	5,6	5,6	108	118	3	2	38	7	3750	59
19	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
30	29	J.Z.	1	0	0	0	0	0	1	1	1,47	1,4	5,53	5,11	150	135	2	2	38	10	3000	53
32	31	C.N.	1	0	0	0	0	0	1	2	1,4	1,33	5,11	4,76	119	133	1	1	39	10	3600	57
34	33	J.C.	0	0	1	0	0	1	2	2	1,47	1,4	5,53	5,6	117	117	1	1	36	4	2250	52
36	35	K/U.	1	0	0	0	0	0	2	1	1,68	1,47	5,32	5,88	139	139	2	1	40	10	3500	56
37	36	D.R.	1	0	0	0	0	0	2	1	1,47	1,47	5,53	5,6	163	164	1	1	42	9	3850	60
38	37	K.T.	1	0	0	0	0	0	2	1	1,54	1,47	5,39	5,18	150	135	3	3	39	10	3700	55
39	38	S.Z.	1	0	0	0	0	0	2	1	1,47	1,47	5,18	5,18	142	156	1	1	41	8	3250	55
40	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
43	42	S.G.	1	0	0	0	0	0	1	1	1,47	1,47	5,11	5,11	143	150	1	1	40	10	3740	58
45	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
53	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
57	56	S.D.	1	0	0	0	0	0	2	1	1,4	1,4	5,25	4,9	124	141	1	1	39	8	2550	46
59																						

Na miejscu kwadratu z trójkątem w górnej części kolumny ponownie pojawia się charakterystyczny dla procesu filtrowania symbol lejka, który oznacza, że obecny proces filtrowania odbył się właśnie w tej kolumnie.

Teraz odpowiemy na pytanie, dlaczego wartości zostało 17, a nie 10 – zgodnie z poleceniem. Wyjaśnienie jest proste. W omawianej kolumnie występuje 1 liczba 1,33, 8 liczb 1,4 i 8 liczb 1,47. A zatem 1 + 8 = 9. Trzeba jeszcze dodać liczbę, żeby było 10. Następna jest 1,47, ale też występuje 8 razy. Filtr nie widzi różnicy pomiędzy nimi i dlatego w kolumnie pozostawia je wszystkie, co daje sumę 17 wartości.

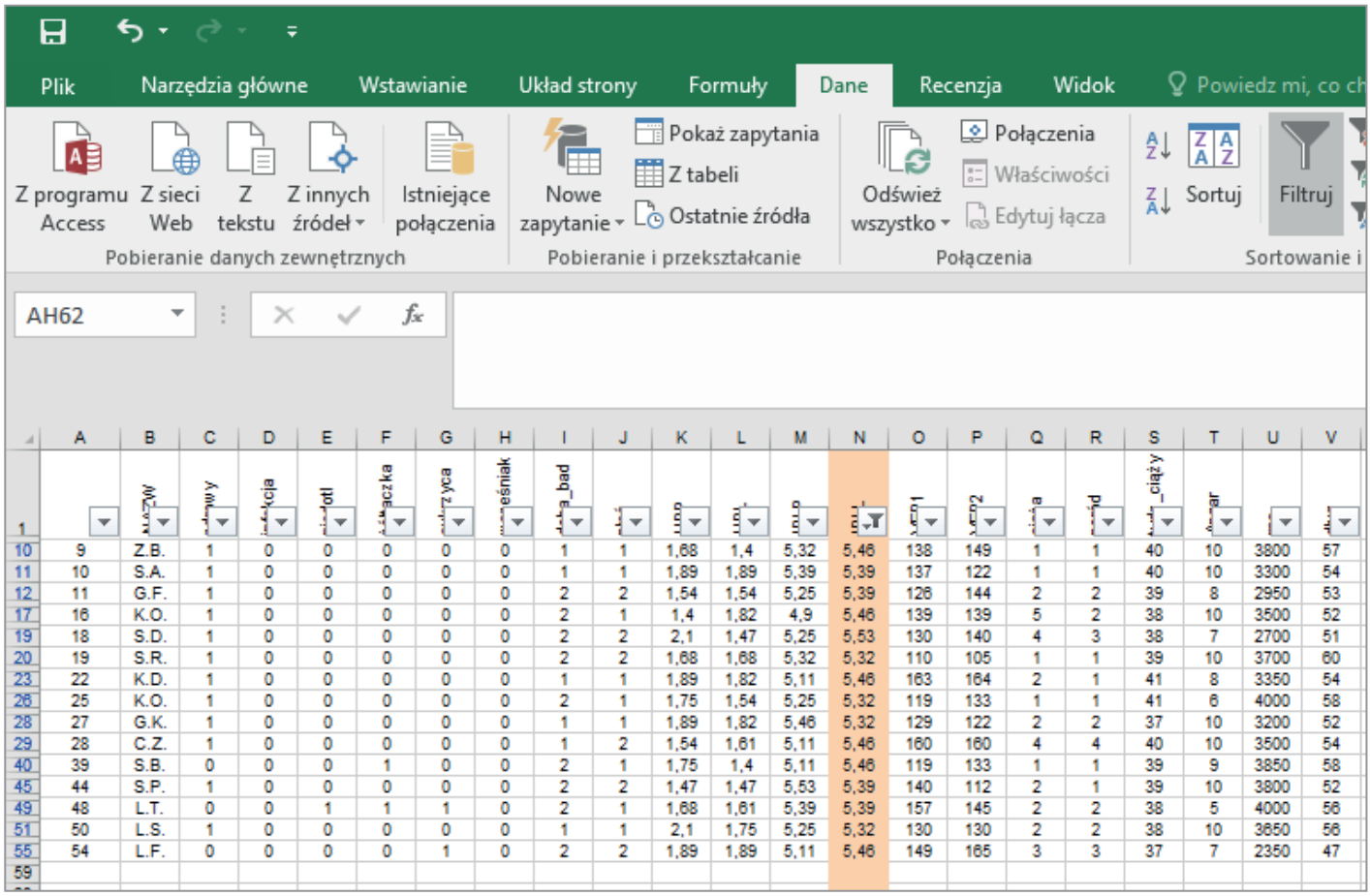
Istnieje również możliwość filtrowania wartości należących do przedziału. W tym celu ze wspomnianego wcześniej menu należy wybrać Filtr niestandardowy, a następnie z górnej

listy po lewej jest większe lub równe, a po prawej „5,32”; analogicznie z dolnych list należy wybrać jest mniejsze lub równe i „5,53”. Pozostawiamy domyślnie zaznaczony spójnik i. (➥ **ryc. 5.12**) Wynik filtrowania jest przedstawiony na ➥ **ryc. 5.13**.



Ryc. 5.12. W kolumnie IPLL filtr ma wybrać parametry słuchowe z ucha lewego, których wartości zamykają się w przedziale IPLL ≥ 5,32; IPLL ≤ 5,53

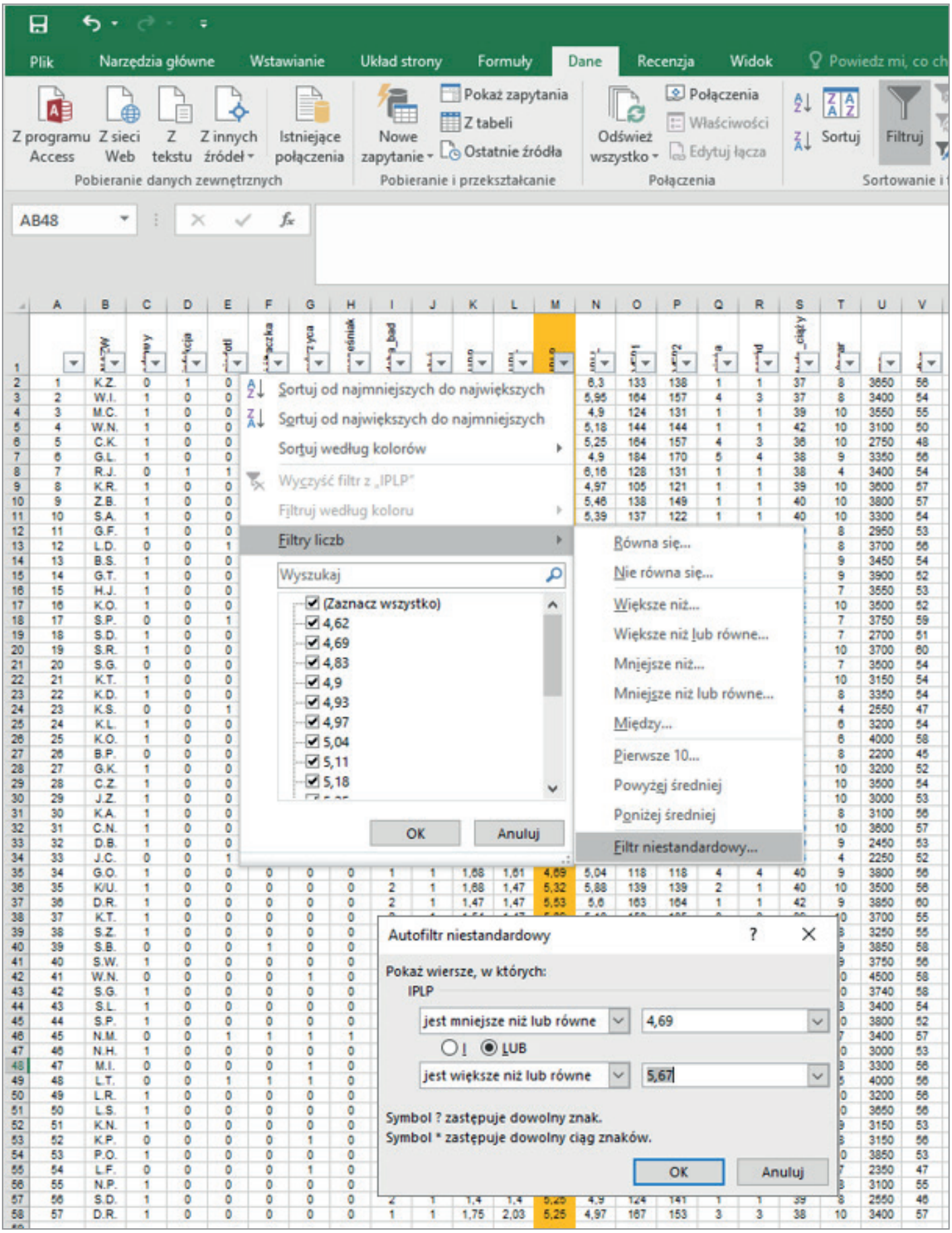
Ryc. 5.13. W kolumnie IPLL pozostały parametry słuchowe z ucha lewego, których wartości czasu zamykają się w ograniczonym zamkniętym przedziale IPLL ≥ 5,32; IPLL ≤ 5,53



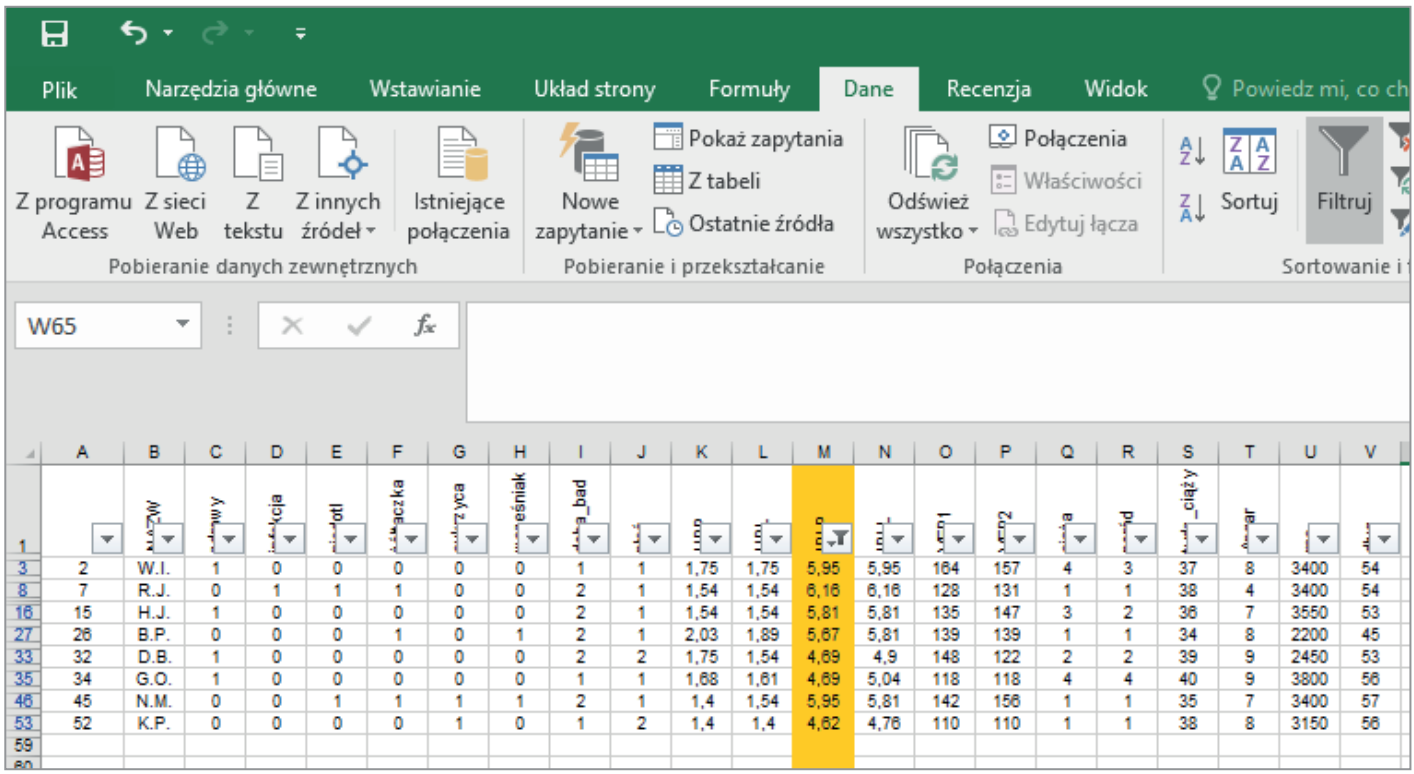
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
10	9	Z.B.	1	0	0	0	0	0	1	1	1,68	1,4	5,32	5,46	138	149	1	1	40	10	3800	57
11	10	S.A.	1	0	0	0	0	0	1	1	1,89	1,89	5,39	5,39	137	122	1	1	40	10	3300	54
12	11	G.F.	1	0	0	0	0	0	2	2	1,54	1,54	5,25	5,39	126	144	2	2	39	8	2950	53
17	16	K.O.	1	0	0	0	0	0	2	1	1,4	1,82	4,9	5,46	139	139	5	2	38	10	3500	52
19	18	S.D.	1	0	0	0	0	0	2	2	2,1	1,47	5,25	5,53	130	140	4	3	38	7	2700	51
20	19	S.R.	1	0	0	0	0	0	2	2	1,68	1,68	5,32	5,32	110	105	1	1	39	10	3700	60
23	22	K.D.	1	0	0	0	0	0	1	1	1,89	1,82	5,11	5,46	163	164	2	1	41	8	3350	54
26	25	K.O.	1	0	0	0	0	0	2	1	1,75	1,54	5,25	5,32	119	133	1	1	41	6	4000	58
28	27	G.K.	1	0	0	0	0	0	1	1	1,89	1,82	5,46	5,32	129	122	2	2	37	10	3200	52
29	28	C.Z.	1	0	0	0	0	0	1	2	1,54	1,61	5,11	5,46	160	160	4	4	40	10	3500	54
40	39	S.B.	0	0	0	1	0	0	2	1	1,75	1,4	5,11	5,46	119	133	1	1	39	9	3850	58
45	44	S.P.	1	0	0	0	0	0	2	2	1,47	1,47	5,53	5,39	140	112	2	1	39	10	3800	52
49	48	L.T.	0	0	1	1	1	0	2	1	1,68	1,61	5,39	5,39	157	145	2	2	38	5	4000	56
51	50	L.S.	1	0	0	0	0	0	1	1	2,1	1,75	5,25	5,32	130	130	2	2	38	10	3650	56
55	54	L.F.	0	0	0	0	1	0	2	2	1,89	1,89	5,11	5,46	149	165	3	3	37	7	2350	47
59																						

Ryc. 5.14. W kolumnie IPLP filtr ma wybrać parametry słuchowe z ucha prawego, których wartości czasu znajdują się poza nierównościami IPLP ≤ 4,69; IPLP ≥ 5,67

Kolejny przykład filtrowania pokaże, jak można wyłączać z wyników określony przedział, a wyniki będą stanowiły obszary wartości na lewo i na prawo na osi odniesienia od tego przedziału (➡ **ryc. 5.14**, ➡ **ryc. 5.15**).



Ryc. 5.15. W kolumnie IPLP pozostały parametry słuchowe z ucha prawego, których wartości czasu znajdują się w obszarze poza nierównościami IPLP ≤ 4,69; IPLP ≥ 5,67



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
3	2	W.I.	1	0	0	0	0	0	1	1	1,75	1,75	5,95	5,95	164	157	4	3	37	8	3400	54
8	7	R.J.	0	1	1	1	0	0	2	1	1,54	1,54	6,16	6,16	128	131	1	1	38	4	3400	54
16	15	H.J.	1	0	0	0	0	0	2	1	1,54	1,54	5,81	5,81	135	147	3	2	36	7	3550	53
27	26	B.P.	0	0	0	1	0	1	2	1	2,03	1,89	5,67	5,81	139	139	1	1	34	8	2200	45
33	32	D.B.	1	0	0	0	0	0	2	2	1,75	1,54	4,69	4,9	148	122	2	2	39	9	2450	53
35	34	G.O.	1	0	0	0	0	0	1	1	1,68	1,61	4,69	5,04	118	118	4	4	40	9	3800	56
46	45	N.M.	0	0	1	1	1	1	2	1	1,4	1,54	5,95	5,81	142	156	1	1	35	7	3400	57
53	52	K.P.	0	0	0	0	1	0	1	2	1,4	1,4	4,62	4,76	110	110	1	1	38	8	3150	56
59																						

Na kolumnie z parametrem słuchowym dla ucha prawego IPLP ustawiony zostanie filtr, który będzie dopuszczał wszystkie wartości na lewo ≤ 4,69 oraz liczby na prawo ≥ 5,67, wyłączając przedział pomiędzy nimi.

Taki przykład czasami zostaje mylnie zinterpretowany, ponieważ osoba wykonująca tego typu obliczenia sugeruje się działaniami, które przeprowadzała wcześniej na innych zbiorach parametrów (bazach). Zatem wpisuje w okienko zależności wartość IPLP ≤ 4,69, a następnie, wybrawszy nierówność IPLP ≥ 5,67, pozostawia zaznaczony łącznik i. Wynik wynosi 0, ponieważ żadna z liczb w bazie nie może spełnić tego warunku.

Żeby proces filtrowania przyniósł skutek, należy między dwoma wspomnianymi nierównościami zaznaczyć łącznik lub. Wtedy otrzymamy oczekiwany wynik.

Andrii Zinchuk

6 Przygotowanie prezentacji w programie PowerPoint

6.1. Cel i znaczenie prezentacji PowerPoint w biostatystyce

Microsoft PowerPoint od dawna stanowi kluczowe narzędzie w nauczaniu i prezentowaniu wyników badań naukowych, także w biostatystyce. Wynika to z jego zdolności do efektywnej prezentacji skomplikowanych danych i analiz w zrozumiały i przystępny sposób.

Głównym celem prezentacji PowerPoint w biostatystyce jest przekazanie wyników analiz statystycznych na tyle klarownie, aby mogły być zrozumiane przez różne grupy odbiorców, w tym studentów, innych naukowców, a także choćby kadrę zarządzającą, której przedstawiciele nie są specjalistami w dziedzinie statystyki. Dzięki wykorzystaniu wizualizacji, takich jak wykresy, diagramy i tabele, prezentacje pomagają uporządkować i ułatwić interpretację wyników analiz statystycznych.

PowerPoint ma szczególne znaczenie w edukacji biostatystycznej, gdzie jest powszechnie wykorzystywany do nauczania najważniejszych koncepcji i technik analizy danych. Nauczyciele mogą np. używać prezentacji do ilustrowania takich zagadnień jak rozkład normalny, testy statystyczne czy regresja liniowa, ułatwiając w ten sposób zrozumienie ich przez studentów.

Jest to ponadto nieocenione narzędzie dla naukowców do przedstawiania wyników ich badań. Odpowiednio przygotowana prezentacja może nie tylko ułatwić publiczne przedstawienie efektów pracy badawczej, ale także promocję wyników wśród innych naukowców, a nawet pomóc w edukacji społecznej na temat znaczenia badań.

Efektywne wykorzystanie PowerPoint w biostatystyce wymaga jednak zarówno zrozumienia koncepcji statystycznych, jak i umiejętności tworzenia klarownych, skutecznych prezentacji. Jest to wyzwanie, któremu można sprostać, rozwijając umiejętności tworzenia prezentacji i zrozumienie technik wizualizacji danych.

Wreszcie PowerPoint odgrywa kluczową rolę w biostatystyce, ponieważ pomaga naukowcom i studentom w komunikacji naukowej, która jest fundamentalnym elementem procesu naukowego. Dzięki skutecznemu prezentowaniu danych statystycznych naukowcy mogą lepiej przedstawiać wyniki swoich badań, co przyczynia się do postępu w dziedzinie biostatystyki.

6.2. Podstawowe informacje o PowerPoint

6.2.1. Krótki opis programu PowerPoint

Microsoft PowerPoint jest wszechstronnym narzędziem do tworzenia prezentacji, używanym zarówno w biznesie, jak i w nauce. Program ten pozwala na efektywne tworzenie, edytowanie i prezentowanie informacji wizualnych, co czyni go nieocenioną pomocą dla studentów, naukowców i nauczycieli, szczególnie w dziedzinie biostatystyki.

Oferuje szeroką gamę narzędzi, które umożliwiają tworzenie atrakcyjnych slajdów prezentacyjnych. Są to m.in. szablony, efekty specjalne, opcje formatowania tekstu,

wsparcie dla różnych formatów multimediów oraz rozmaite rodzaje wykresów i diagramów do przedstawiania danych. Ich szeroki wybór pozwala na dostosowanie prezentacji do potrzeb użytkownika i osiągnięcie pożądanych efektów wizualnych.

Korzystanie z PowerPointa wymaga jednak pewnych umiejętności i wiedzy technicznej. Wielkość i złożoność tego programu może początkowo wydawać się przytłaczająca, szczególnie osobom, które nie miały wcześniej do czynienia z podobnymi aplikacjami. Dlatego wiele publikacji dostarcza szczegółowych instrukcji i wskazówek na temat korzystania z różnych funkcji PowerPointa, pomagając użytkownikom lepiej zrozumieć i wykorzystać możliwości tego programu.

PowerPoint jest także kompatybilny z innymi produktami Microsoftu, takimi jak Word czy Excel, co umożliwia łatwe importowanie i eksportowanie treści między tymi programami. To sprawia, że jest on szczególnie przydatny dla naukowców i studentów, którzy muszą przedstawić skomplikowane dane statystyczne w przystępnej formie.

Program pozwala ponadto na interaktywne przedstawienie danych za pomocą narzędzi takich jak Power BI, co umożliwia tworzenie interaktywnych dashboardów i raportów, które można łatwo zintegrować z prezentacjami. To pozwala na bardziej dynamiczne i interesujące przedstawienie danych, co jest szczególnie ważne w kontekście prezentacji naukowych.

Podsumowując, PowerPoint jest niezwykle elastycznym i wszechstronnym narzędziem do tworzenia prezentacji, które oferuje wiele możliwości dla osób pracujących w dziedzinie biostatystyki. Jednak efektywne wykorzystanie możliwości programu wymaga zrozumienia jego funkcji i właściwości, które są szczegółowo omawiane w wielu publikacjach.

6.2.2. Podstawowe funkcje i narzędzia PowerPoint

Główne kategorie narzędzi aplikacji obejmują obszary takie jak zarządzanie slajdami, edycja tekstu, obróbka obrazów i multimediów, tworzenie wykresów oraz integracja z innymi produktami Microsoft.

6.2.2.1. Zarządzanie slajdami

Jedną z podstawowych funkcji PowerPoint jest zarządzanie slajdami, które obejmuje dodawanie, usuwanie, duplikowanie i reorganizowanie slajdów w prezentacji. Użytkownik może wybierać spośród różnych szablonów dostępnych w programie, które są zaprojektowane do prezentowania różnego rodzaju informacji w efektywny sposób.

6.2.2.2. Edycja tekstu

PowerPoint oferuje zaawansowane narzędzia do formatowania tekstu, takie jak dobór czcionek, rozmiarów, kolorów, efektów specjalnych i stylizacji. Program umożliwia również tworzenie list, wykorzystanie różnych stylów tytułów i podtytułów, a także formatowanie akapitów.

6.2.2.3. Obróbka obrazów i multimediów

Użytkownicy mogą wstawić i edytować różne rodzaje multimediów, w tym rysunki, fotografie, klipy wideo i dźwiękowe. PowerPoint oferuje również narzędzia do przycinania, skalowania, obracania, przesuwania i stosowania efektów artystycznych do obrazów.

6.2.2.4. Tworzenie wykresów

Jednym z najważniejszych narzędzi PowerPoint, szczególnie dla osób pracujących w dziedzinie biostatystyki, jest funkcja tworzenia wykresów. Użytkownicy mogą tworzyć różne rodzaje wykresów, w tym liniowe, kolumnowe, kołowe, punktowe

i wiele innych. PowerPoint pozwala również na ich dostosowanie, w tym zmianę kolorów, stylów, etykiet i skali.

6.2.2.5. Integracja z innymi produktami Microsoftu

PowerPoint pozwala na łatwą integrację z innymi produktami należącymi do pakietu MS Office, takimi jak Word i Excel. Użytkownicy mogą importować tabele i wykresy z Excela, a następnie edytować je bezpośrednio w PowerPoincie. Podobnie można importować tekst z Worda, zachowując przy tym formatowanie.

6.2.2.6. Prezentowanie

Przede wszystkim jednak PowerPoint pozwala prezentować slajdy na różnorodne sposoby, takie jak pokaz slajdów, widok prezentera, automatyczne przejścia między slajdami, a także narzędzia do zaznaczania i podkreślania ważnych punktów podczas prezentacji.

PowerPoint zawiera wiele narzędzi, które mogą być dostosowane do konkretnych potrzeb użytkownika, umożliwiając tworzenie profesjonalnych i efektownych prezentacji. Dlatego ważne jest, aby użytkownicy zrozumieli działanie i umiejętnie wykorzystali te narzędzia do projektowania jak najbardziej efektywnych prezentacji.

6.3. Rozpoczynanie pracy

6.3.1. Wybór tematu i planowanie prezentacji

Wybór tematu jest pierwszym i decydującym krokiem w procesie tworzenia prezentacji. Temat powinien być nie tylko interesujący dla odbiorców, ale również dobrze znany autorowi. W przypadku biostatystyki mogą obejmować różne zagadnienia, takie jak nowe metody analizy danych, wyniki badań epidemiologicznych czy analiz statystycznych.

Planowanie prezentacji jest kolejnym istotnym krokiem po wyborze tematu. Dobrze zaplanowana prezentacja powinna mieć jasny cel, logiczną strukturę i wykorzystywać odpowiednie narzędzia do przekazania głównych założeń. Celem może być informowanie, przekonywanie, edukowanie lub inspirowanie publiczności. Struktura, zwykle obejmująca wstęp, rozwinięcie i zakończenie, powinna prowadzić publiczność przez różne punkty prezentacji w sposób logiczny i spójny.

Podczas planowania prezentacji warto zastanowić się nad rodzajem informacji, która ma być przekazana, i najlepszym sposobem jej przedstawienia. Na przykład skomplikowane dane statystyczne mogą wymagać użycia wykresów, tabel lub innych form graficznych, podczas gdy koncepcje teoretyczne być może najlepiej przedstawić za pomocą tekstu lub diagramów.

Planowanie powinno również obejmować rozważenie logistyki prezentacji, takiej jak czas trwania, dostępny sprzęt i oczekiwania odbiorców. Jeśli np. istnieje ograniczenie czasowe, prezenter musi zaplanować, jak skutecznie przekazać najistotniejsze informacje w dostępnym czasie. Podobnie, jeśli prezentacja ma być wygłoszona online, trzeba zaplanować, jak najlepiej wykorzystać dostępne narzędzia do komunikacji.

Należy pamiętać, że nawet najlepiej zaplanowana prezentacja wymaga prób i modyfikacji. Trzeba przeprowadzić kilka prób, aby się upewnić, że wszystko działa prawidłowo, a przekaz jest jasny i zrozumiały. Ostateczny plan powinien być elastyczny i otwarty na zmiany, w zależności od reakcji odbiorców.

Wybór tematu i planowanie to istotne etapy tworzenia efektywnej prezentacji. Prezenter musi wybrać temat, który zna bardzo dobrze i który jest interesujący dla publiczności, a następnie tak zorganizować prezentację, aby jasno i skutecznie przekazać najważniejsze informacje.

6.3.2. Wybór odpowiedniego szablonu

Wybór odpowiedniego szablonu w PowerPoincie ma istotne znaczenie w procesie tworzenia prezentacji. Może on znacząco wpłynąć na odbiór przekazu. Szablony programu zapewniają gotową ramę do tworzenia prezentacji, obejmującą zarówno ogólną kompozycję stron, jak i projektowanie szczegółów, takich jak kolory, czcionki czy układ elementów.

Pierwszym krokiem jest zrozumienie celu prezentacji i odbiorcy. Dla prezentacji biostatystycznych, gdzie dane i ich interpretacja są najistotniejsze, preferowane mogą być szablony, które ułatwiają przedstawienie złożonych danych w czytelny sposób.

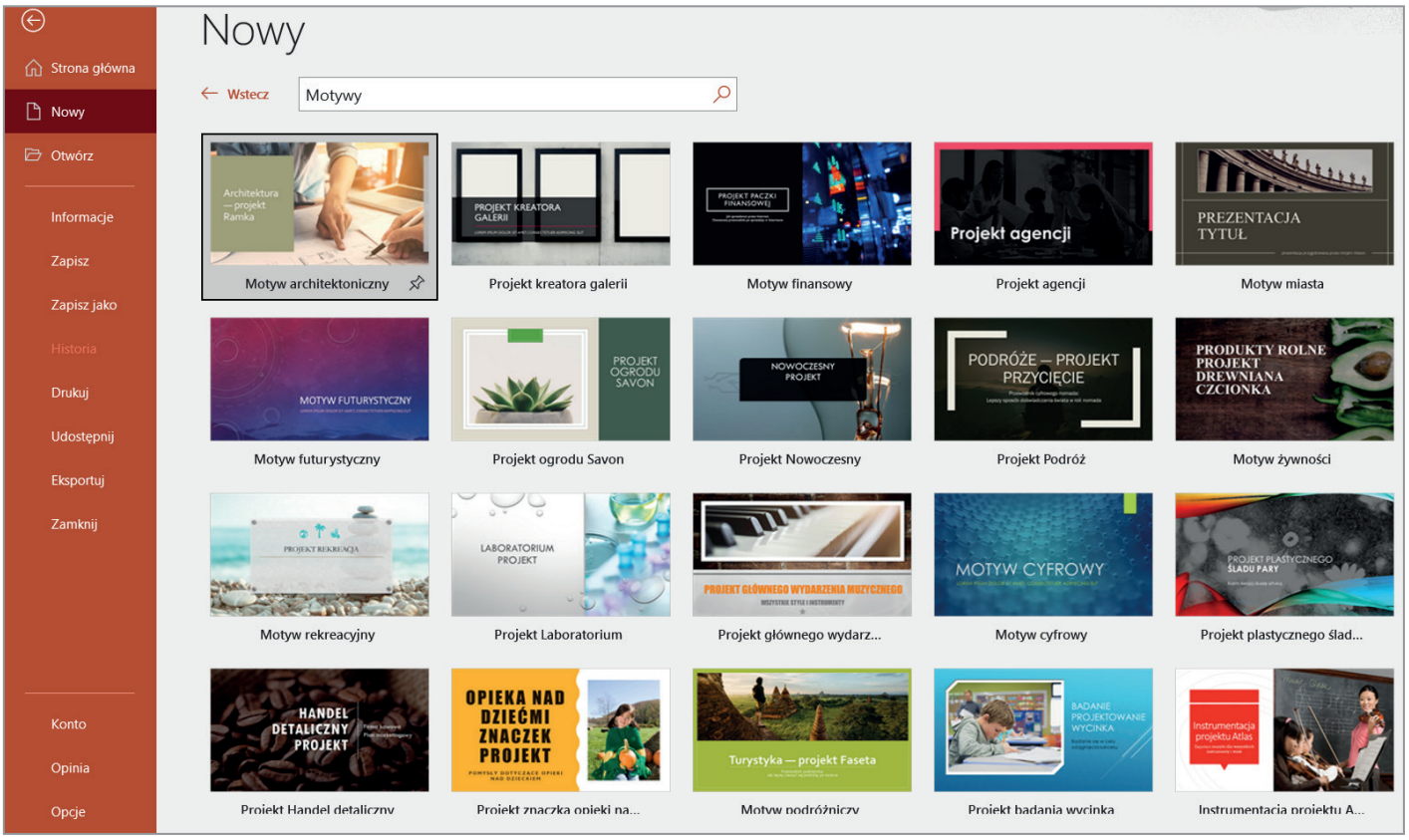
Powinien być dostosowany do kontekstu, w którym będzie prezentowany. Obejmuje to kulturę organizacyjną, oczekiwania i preferencje publiczności. Przykładowo na poważnej medycznej konferencji naukowej formalny i minimalistyczny szablon może być bardziej odpowiedni niż kolorowy i żartobliwy (➥ **ryc. 6.1**).

Ryc. 6.1. Wybór odpowiedniego szablonu jest istotny dla skuteczności komunikacji

Źródło: https://www.templatemonster.com/pl/szablony-powerpoint-86955.html#thumbnail-1 (na dole); https://naurok.com.ua/post/8-bezprograshnih-sposobiv-zipsuvati-prezentaciyu

Wybór szablonu powinien też uwzględniać indywidualne preferencje twórcy prezentacji. Niektóre mogą być bardziej intuicyjne i łatwe do użycia, a inne mogą lepiej pasować do stylu i preferencji prezentacyjnych twórcy. Dlatego warto poświęcić czas na zapoznanie się z różnymi szablonami i wybór tego, który najbardziej odpowiada naszym potrzebom (➥ **ryc. 6.2**).

Wybór odpowiedniego szablonu jest ważnym etapem procesu tworzenia prezentacji w PowerPoincie. Szablon powinien być dobrany zgodnie z celem prezentacji, oczekiwaniami publiczności, a także osobistymi preferencjami twórcy prezentacji.



Ryc. 6.2. Na ekranie uruchamiania nowej prezentacji można wybrać dla niej szablon

6.3.3. Tworzenie struktury slajdów

Kolejnym, po wyborze tematu, opracowaniu planu i wyborze szablonu, etapem w procesie tworzenia prezentacji jest ustalenie struktury slajdów. Jest to zorganizowana sekwencja slajdów, która prowadzi publiczność przez prezentację.

Podstawowa struktura powinna zawierać 3 główne elementy: wprowadzenie, część główną i zakończenie. Wprowadzenie musi zawierać krótkie przedstawienie tematu, cele prezentacji i potencjalnie tezę, którą prezentacja będzie argumentować. Część główna powinna obejmować dowody lub argumenty na potwierdzenie tezy, a zakończenie – podsumować prezentację i zapewnić miejsce na pytania lub dyskusję.

W przypadku prezentacji naukowych, np. medycznych, istotne jest również uwzględnienie dodatkowych elementów struktury, takich jak metody, wyniki i dyskusja. Metody powinny przedstawić, w jaki sposób dane zostały zgromadzone i analizowane. Wyniki prezentują główne osiągnięcia, a dyskusja interpretuje te wyniki w kontekście literatury naukowej. Każdy z tych elementów może obejmować wiele slajdów, w zależności od złożoności tematu i danych (➥ **ryc. 6.3**).

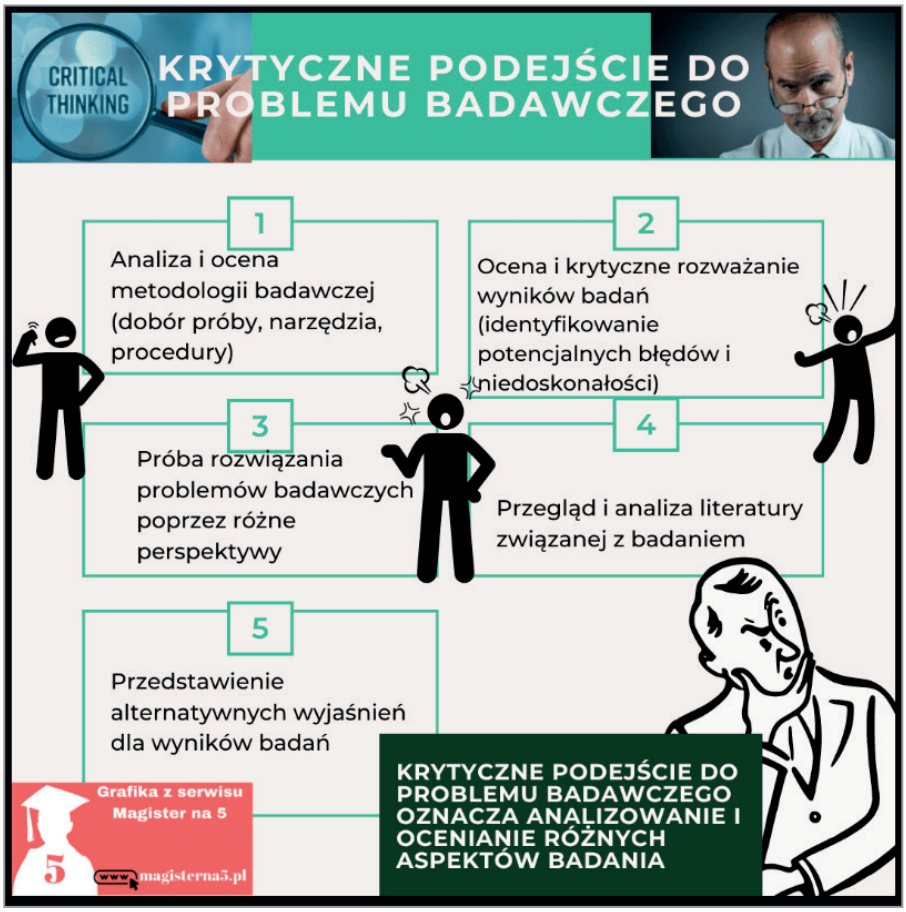
Tworzenie struktury slajdów wymaga bacznej uwagi i strategicznego myślenia. Najważniejsze jest zachowanie równowagi między przekazywaniem informacji a utrzymaniem zrozumiałej formy dla odbiorców. Zbyt wiele slajdów może zniechęcić i zdezorientować publiczność, podczas gdy zbyt mała ich liczba może nie dostarczyć wystarczających informacji. Struktura powinna być więc dostosowana do celu, złożoności tematu i potrzeb publiczności.

Ryc. 6.3. Proces tworzenia struktury slajdów w programie PowerPoint

W prezentacjach naukowych warto również uwzględnić użycie slajdów przejściowych, które pomagają widzom śledzić postęp i poruszanie się między różnymi sekcjami. Mogą one obejmować krótkie streszczenia, punkty orientacyjne lub pytania, które zapewniają ciągłość i spoistość prezentacji (➥ **ryc. 6.4**).

Ryc. 6.4. Przykład krytycznego podejścia przy tworzeniu prezentacji naukowej

Źródło: https://magisterna5.pl/wp-content/uploads/2021/12/krytyczne-podejscie-do-problemu-badawczego-w-dyskusji.png



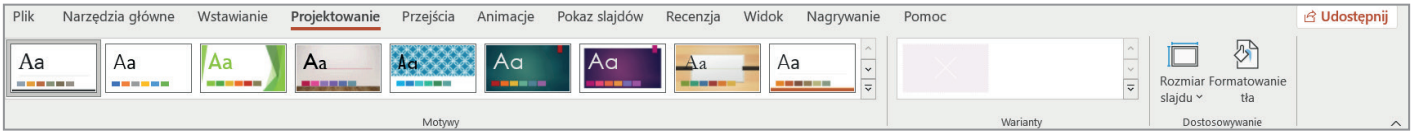
Struktura slajdów powinna służyć za drogowskaz dla publiczności, pomagając jej zrozumieć, na jakim etapie prezentacji jest i dokąd zmierza. Dobrze zaprojektowana struktura slajdów może znacznie poprawić klarowność i skuteczność prezentacji, szczególnie naukowych.

6.3.4. Kluczowe funkcje, które umożliwią wzbogacenie prezentacji

6.3.4.1. Jak zmienić tło prezentacji w PowerPoincie

W programie PowerPoint można ustawić tło, które może być jednolitym kolorem, gradientem lub obrazem. Jeśli standardowe tło nie pasuje do zadań prezentacji, zawsze możesz je zmienić w podany poniżej sposób.

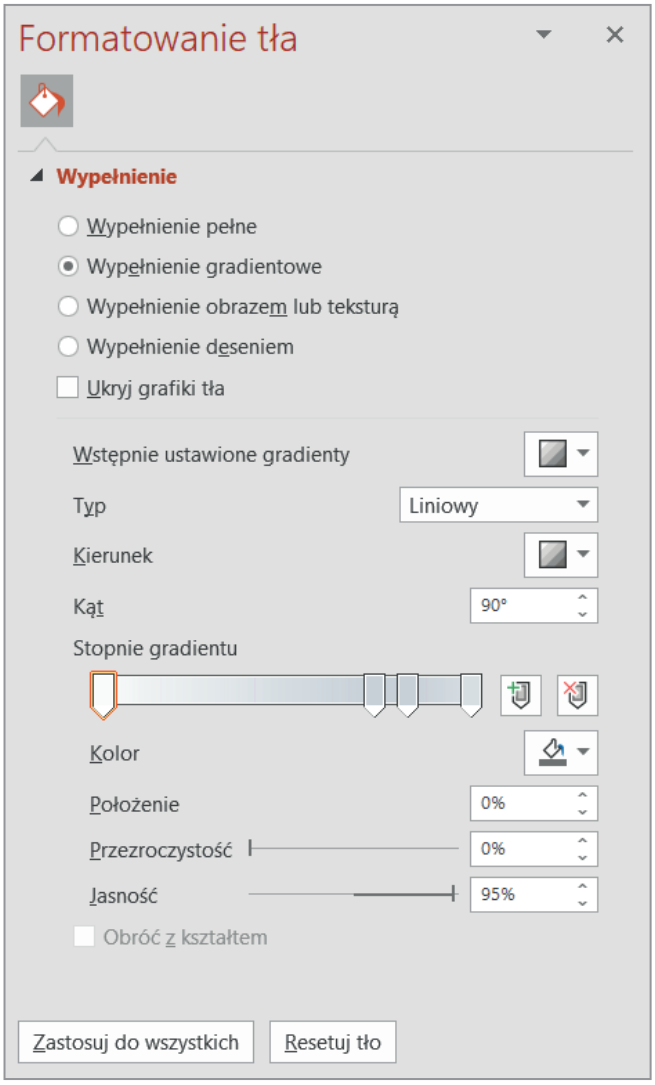
Na karcie Projektowanie w grupie Dostosowywanie kliknij Formatowanie tła (➥ **ryc. 6.5**).



Ryc. 6.5. Polecenie Formatowanie tła znajduje się po prawej stronie karty Projektowanie

Korzystając z dostępnych parametrów, możesz dostosować tło zgodnie z zamysłem (➥ **ryc. 6.6**).

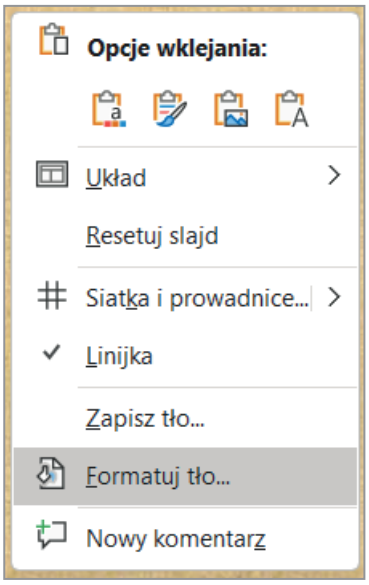
Po ustawieniu pożądanego tła zaakceptuj zmiany, klikając Zastosuj do wszystkich.



Ryc. 6.6. Zawartość okna Formatowanie tła

6.3.4.2. Jak dodać własny obraz z komputera jako tło

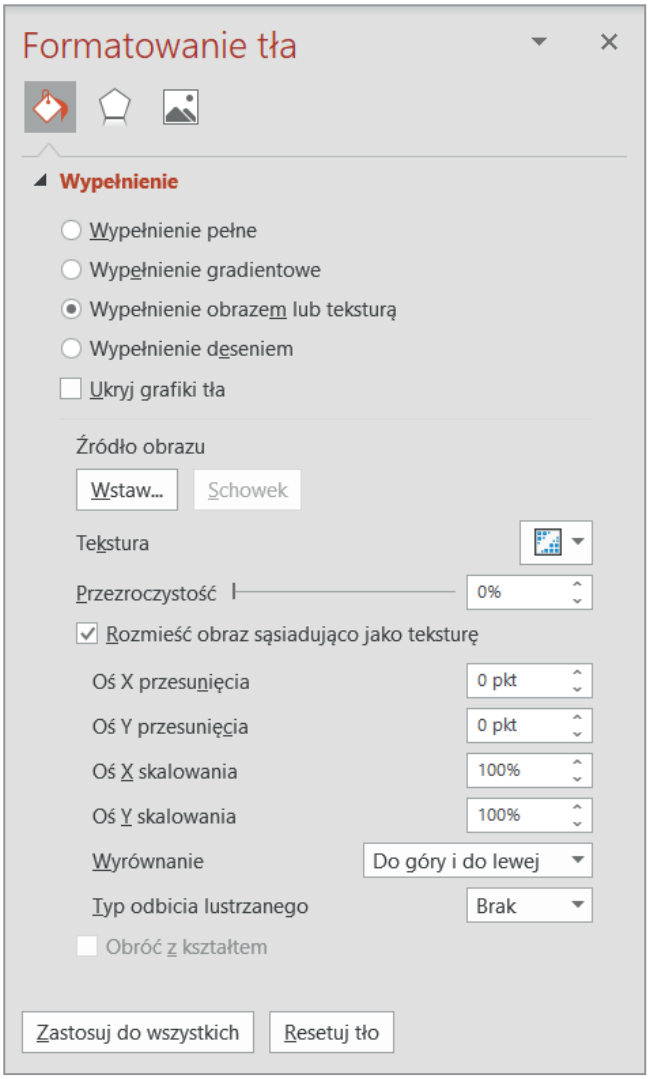
Aby dodać własny obraz z komputera jako tło, trzeba kliknąć prawym przyciskiem myszy sam slajd i wybrać opcję Formatuj tło (➡ **ryc. 6.7**).



Ryc. 6.7. Kliknięcie polecenia Formatuj tło z menu kontekstowego slajdu to pierwszy krok do wstawienia własnego obrazu jako tła

W panelu Formatowanie tła trzeba kliknąć Wypełnienie obrazem lub teksturą i dopiero wtedy w sekcji Źródło obrazu należy wybrać Wstaw (➡ **ryc. 6.8**).

Teraz można wybrać obraz z określonej lokalizacji na twardym dysku komputera.

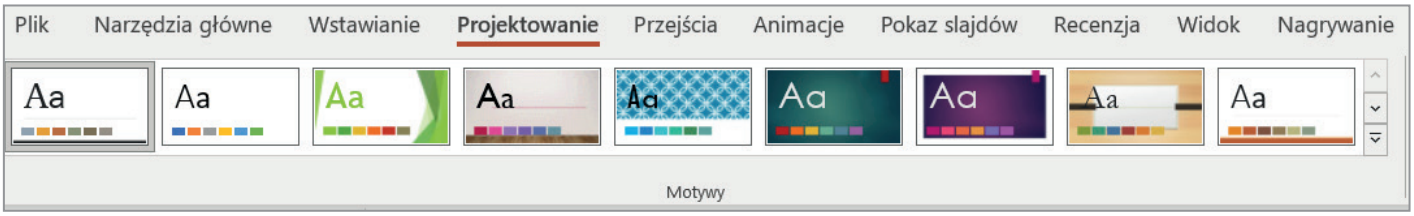


Ryc. 6.8. Po kliknięciu polecenia Formatuj tło z menu kontekstowego pojawia się znane już okno Formatowanie tła

6.3.4.3. Jak zmienić szablon w programie PowerPoint

Zwykle szablon prezentacji jest wybierany na samym początku jej tworzenia. Ale jeśli zaszła konieczność wymiany tła już podczas tworzenia prezentacji, nie ma potrzeby zaczynać wszystkiego od nowa, zawsze można dokonać tego w trakcie pracy.

Aby to zrobić w karcie Projektowanie, w grupie Motywy trzeba wybrać propozycję tła, która będzie pasować do tworzonej prezentacji (➥ **ryc. 6.9**).



Ryc. 6.9. Zawartość grupy Motywy z karty Projektowanie pozwala na zmianę kolorystyki prezentacji

Aby zobaczyć wszystkie możliwe opcje tła, należy je przełączać za pomocą strzałek znajdujących się po prawej stronie.

Po wyborze motywu można wybrać odpowiadającą nam jego wersję w grupie Warianty (➥ **ryc. 6.10**).

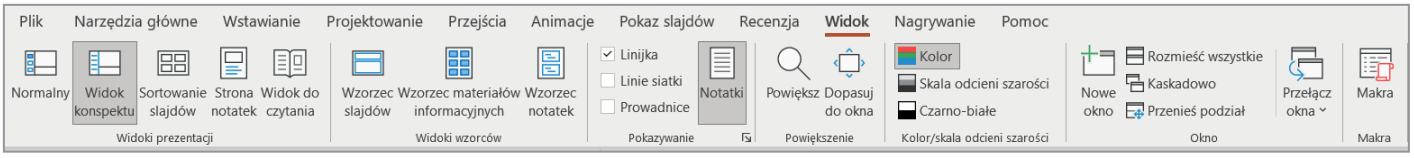
Możliwe jest również samodzielne dostosowanie szablonu, jeśli gotowe projekty nie są odpowiednie.

Ryc. 6.10. Zawartość grupy Warianty daje możliwość wyboru różnych kolorów szablonu

6.3.4.4. Jak zrobić spis treści w prezentacji

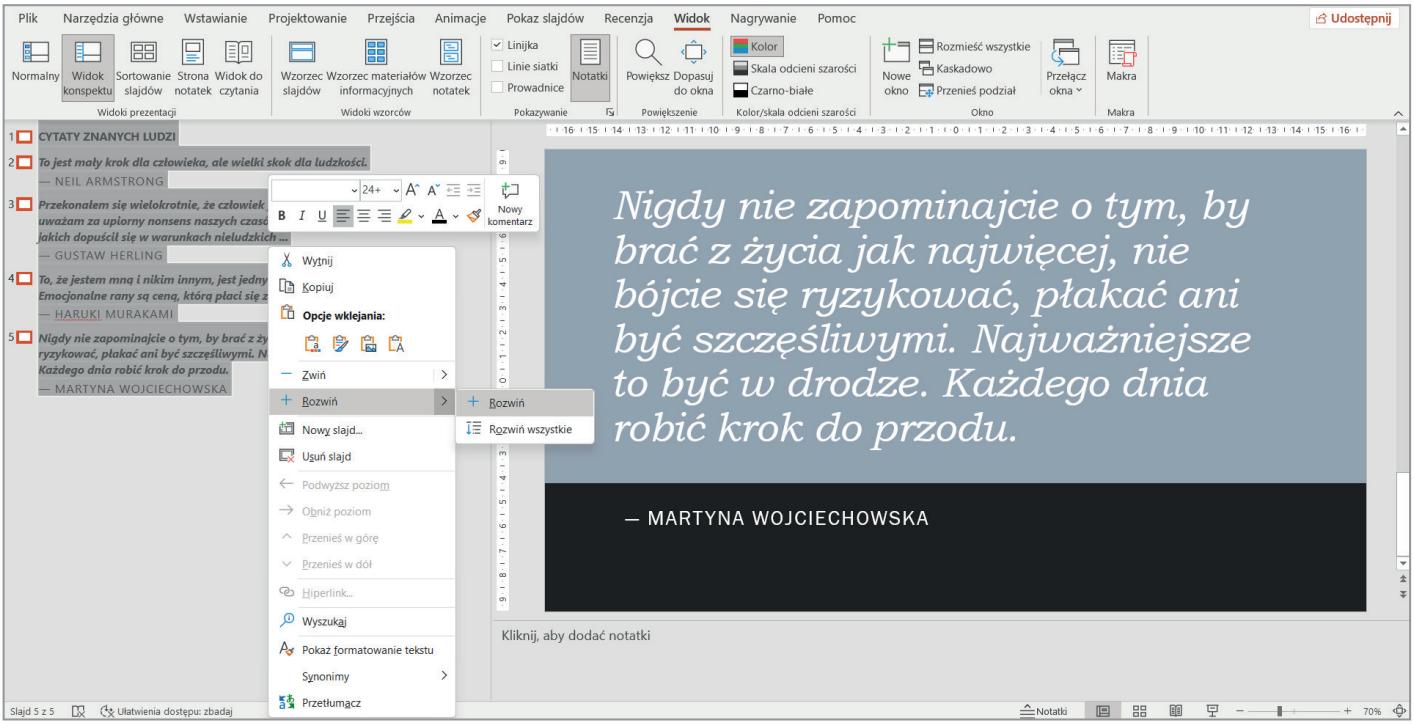
W trakcie tworzenia prezentacji możliwe jest utworzenie osobnego slajdu ze spisem treści i samodzielne wprowadzenie do niego tytułów. Możliwe jest również szybkie skopiowanie wszystkich tytułów slajdów naraz i wklejenie ich w wymaganym miejscu.

W tym celu przejdź do karty Widok i wybierz polecenie Widok konspektu (➥ **ryc. 6.11**).



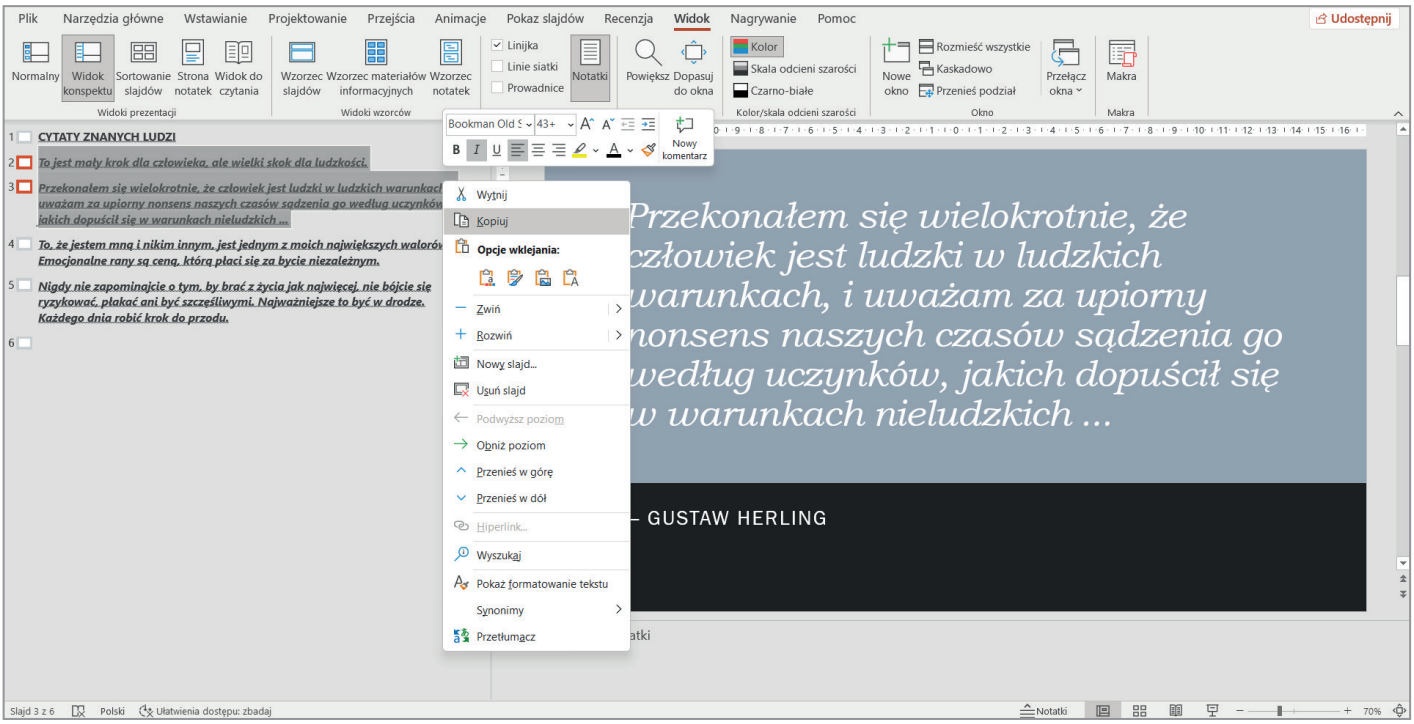
Ryc. 6.11. Polecenie Widok konspektu znajduje się w karcie Widok

Aby zobaczyć zawartość slajdów w trybie Widok konspektu, a nie tylko ich tytuły, należy kliknąć prawym przyciskiem myszy w oknie miniatur slajdów i z menu kontekstowego wybrać Rozwiń lub Rozwiń wszystko. Jeśli nie ma już potrzeby wyświetlania całej zawartości slajdu, wybierz Zwiń lub Zwiń wszystko (➥ **ryc. 6.12**).



Ryc. 6.12. Polecenie Rozwiń z menu kontekstowego widoku konspektu

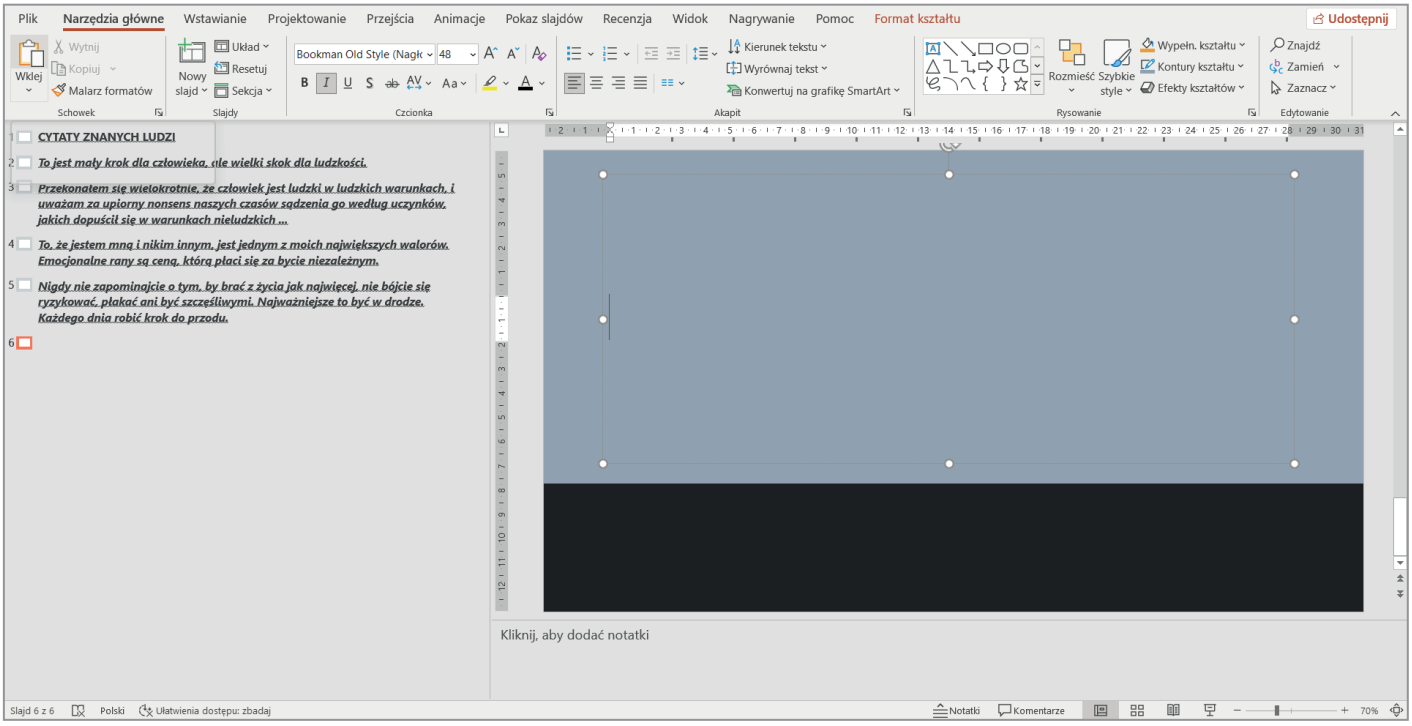
Aby skopiować tytuły slajdów, trzeba je najpierw zaznaczyć (tak jak zwykle przy zaznaczaniu tekstu), a następnie kliknąć prawym przyciskiem myszy i wybrać polecenie Kopiuj (➥ **ryc. 6.13**).



Ryc. 6.13. Tytuły slajdów można skopiować w trybie widoku konspektu

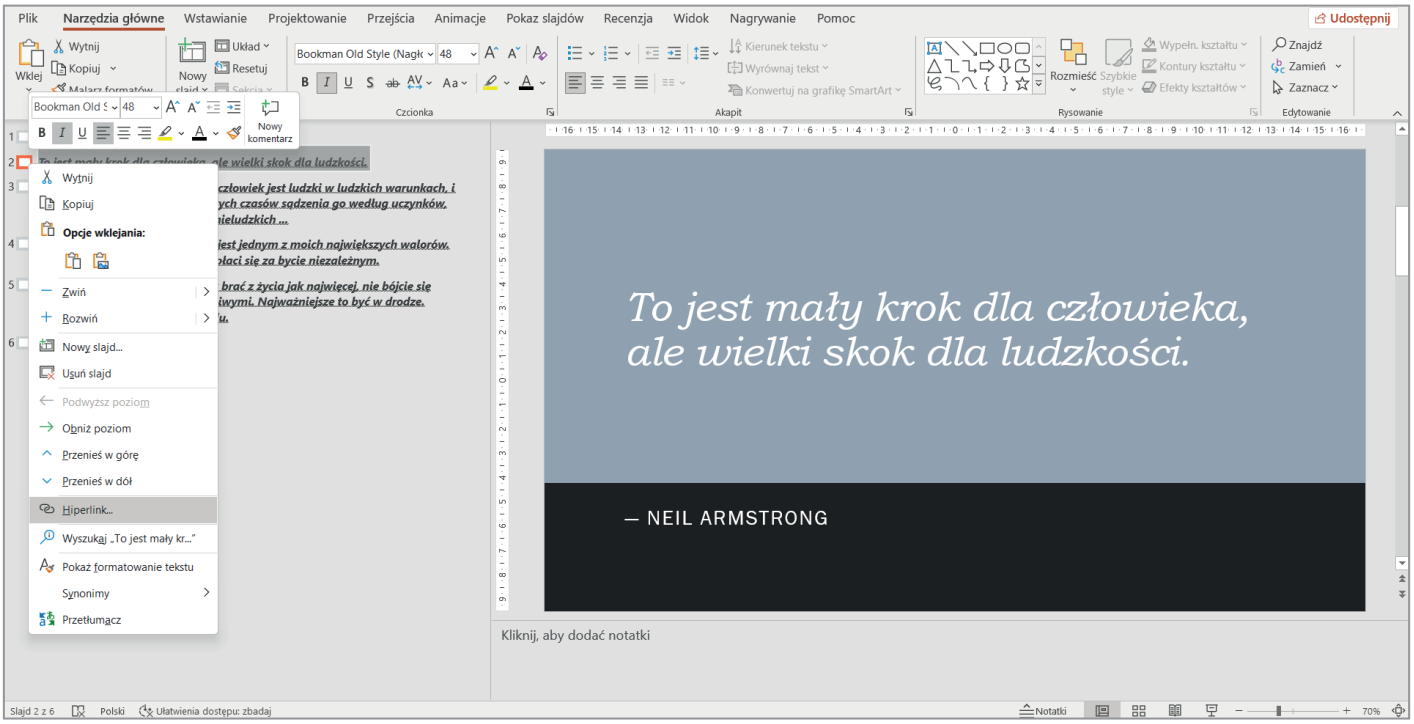
Ryc. 6.14. Na slajdzie spisu treści należy wkleić skopiowane tytuły

Po zakończeniu kopiowania trzeba przejść bezpośrednio do slajdu ze spisem treści, a następnie wkleić w nim skopiowane tytuły. Można to zrobić za pomocą kombinacji klawiszy Ctrl + V lub w dowolny inny wygodny sposób (➥ **ryc. 6.14**).



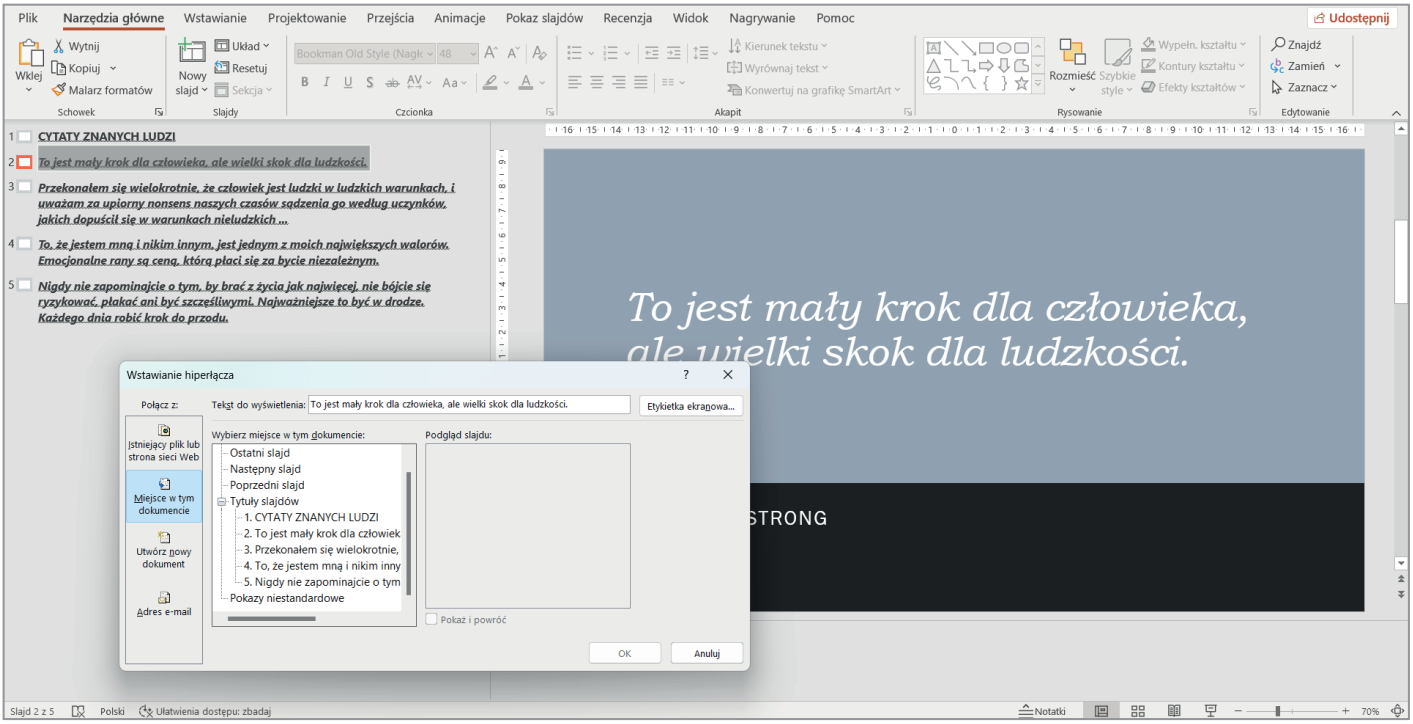
Ryc. 6.15. Tworzenie hiperłącza z tytułów w menu

Jeśli konieczne będzie dodanie hiperłącza do slajdu ze spisem treści, należy zaznaczyć niezbędne nagłówki i kliknąć prawym przyciskiem myszy, a następnie w menu podręcznym wybrać polecenie Hiperlink (➥ **ryc. 6.15**).



Następnie w wyskakującym oknie (w kolumnie po lewej stronie) należy wybrać Miejsce w tym dokumencie, a w sekcji Wybierz miejsce w tym dokumencie odpowiedni slajd i kliknąć OK, aby wstawić hiperłącze (➥ **ryc. 6.16**).

Ryc. 6.16. Okno wstawiania hiperłącza w prezentacji

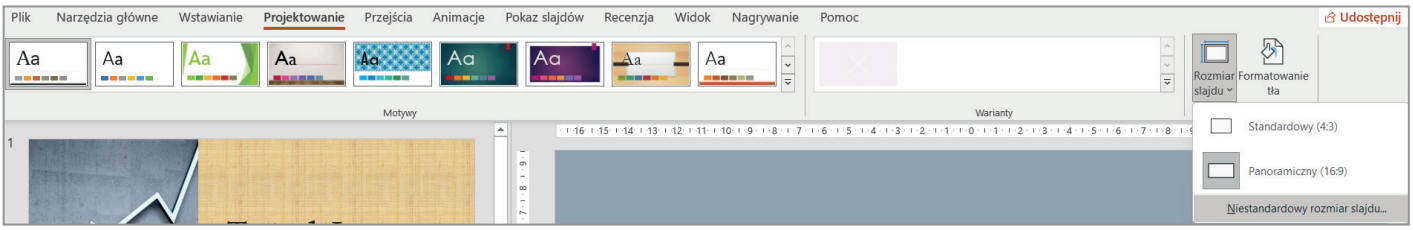


6.3.4.5. Obracanie slajdu

Jeśli domyślne ustawienia slajdów są nieodpowiednie, zawsze istnieje możliwość zmiany ich orientacji (poziomej lub pionowej) oraz rozmiarów (domyślne proporcje 4 : 3, 16 : 9 lub dowolny rozmiar slajdu). Zmiany te można wprowadzić na samym początku tworzenia prezentacji lub w trakcie prac nad nią.

W tym celu należy otworzyć kartę Projektowanie i po prawej stronie w sekcji Dostosowanie wybrać żądany rozmiar slajdu (➥ **ryc. 6.17**).

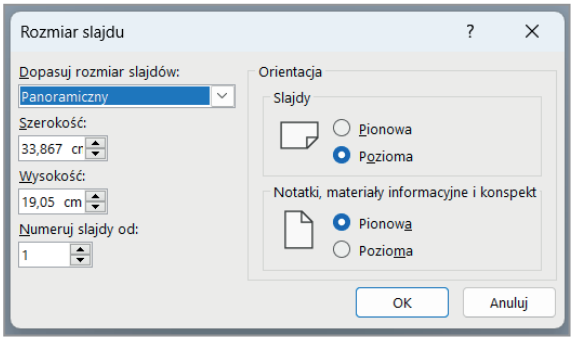
Ryc. 6.17. Zmiana rozmiaru slajdu poleceniem Rozmiar slajdu



Żeby zmienić orientację slajdu, w tym samym menu na karcie Projektowanie trzeba wybrać Niestandardowy rozmiar slajdów. W wyskakującym oknie można zmienić nie tylko orientację slajdów, ale także ustawić pożądane przez użytkownika ich niestandardowe wymiary (➥ **ryc. 6.18**).

Po wprowadzeniu niezbędnych zmian zatwierdzamy (klikamy OK).

Ryc. 6.18. Okno niestandardowych wymiarów slajdów

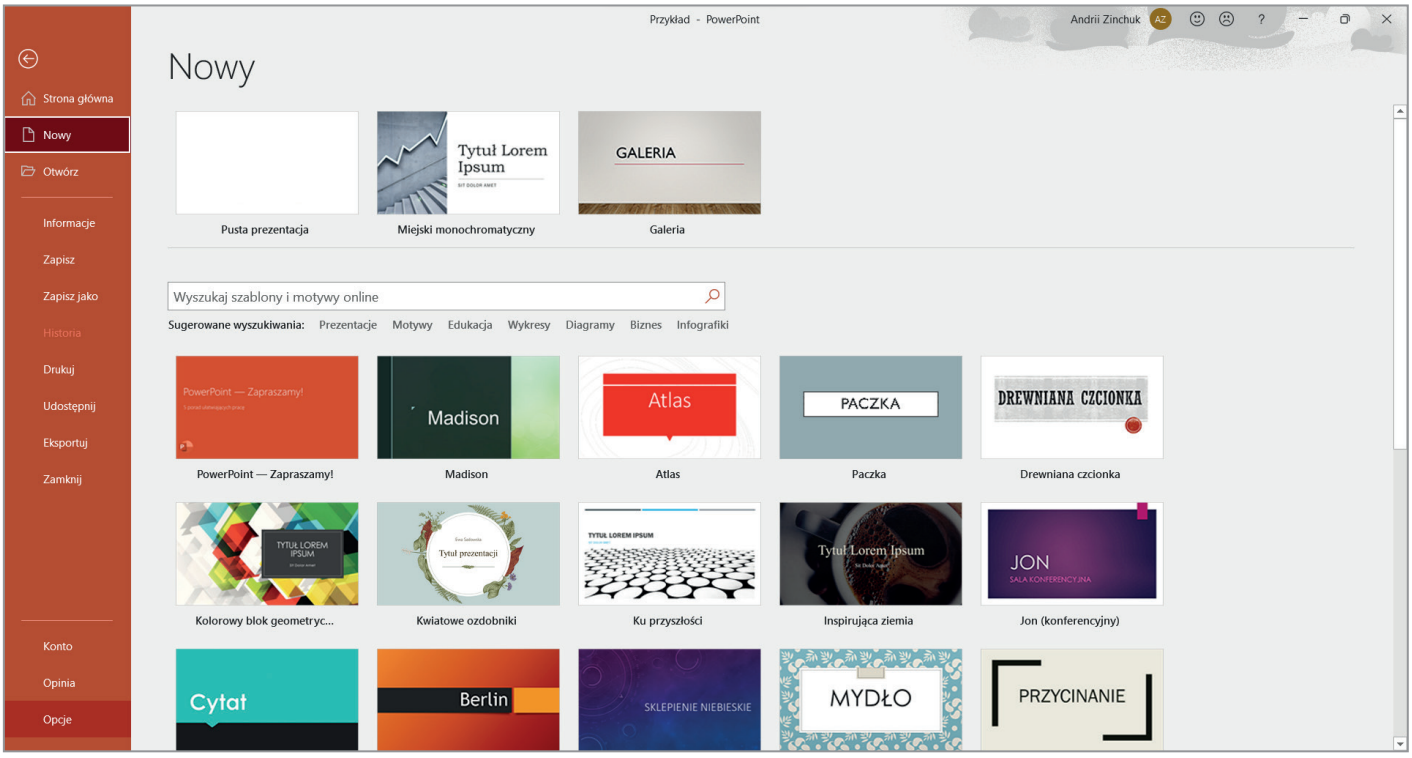


6.3.4.6. Automatyczne sprawdzanie pisowni

Jeśli w trakcie pracy zajdzie konieczność dodatkowego sprawdzenia pisowni i automatycznego poprawiania tekstu, możesz bardzo łatwo włączyć tę opcję, a jeśli nie ma takiej potrzeby, możesz ją odrzucić.

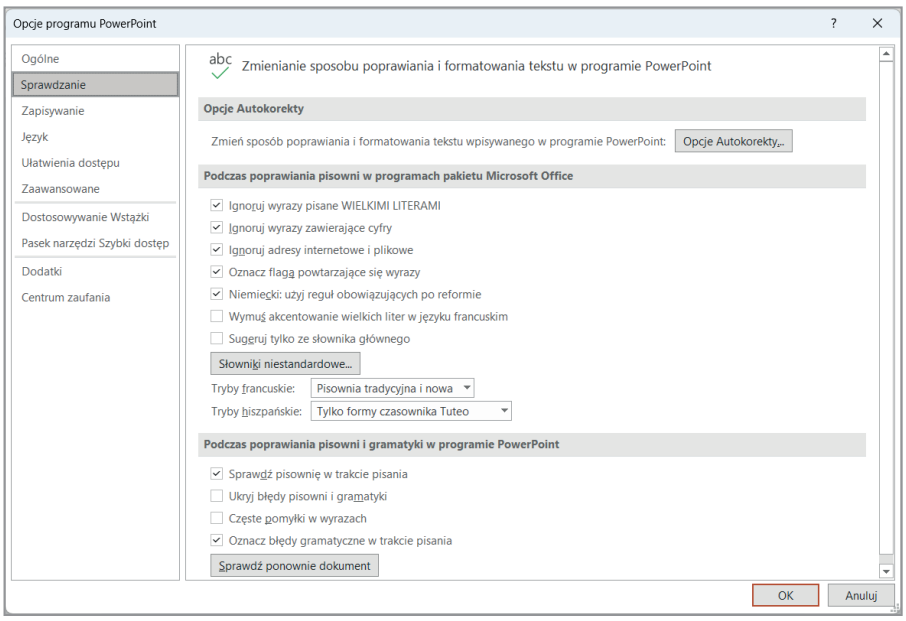
W tym celu wybierz kartę Plik, a następnie Opcje (➥ **ryc. 6.19**).

Ryc. 6.19. Okno po kliknięciu przycisku Plik (dostęp do przycisku Opcje)



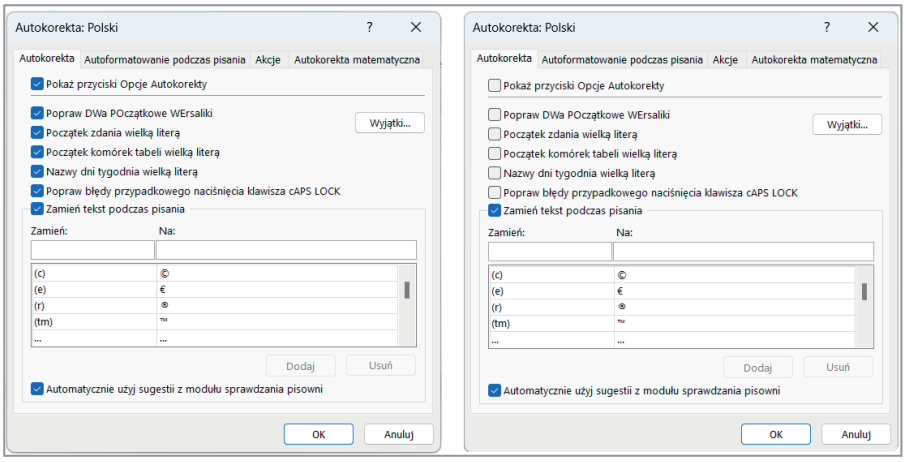
Po lewej stronie wyskakującego okna wybierz polecenie Sprawdzanie, a następnie kliknij przycisk Opcje Autokorekty, który pojawi się po prawej stronie u góry okna (➥ **ryc. 6.20**).

Ryc. 6.20. Dostęp do Opcji Autokorekty



W kolejnym wyskakującym oknie, na dole wszystkich wymienionych opcji (w pozostałych także można usunąć lub zostawić oznaczenia) według własnego uznania można odznaczyć lub zaznaczyć pole Zamień tekst podczas pisania (➥ **ryc. 6.21**).

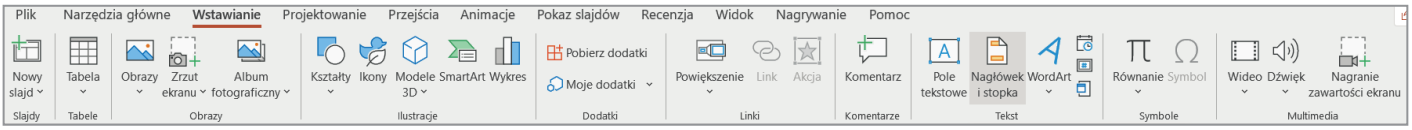
Warto pamiętać, że ta funkcja nie jest włączona w tym samym czasie dla całego pakietu Microsoft Office.



Ryc. 6.21. Możliwość odznaczenia i zaznaczenia dodatkowych opcji

6.3.4.7. Dodawanie przypisów

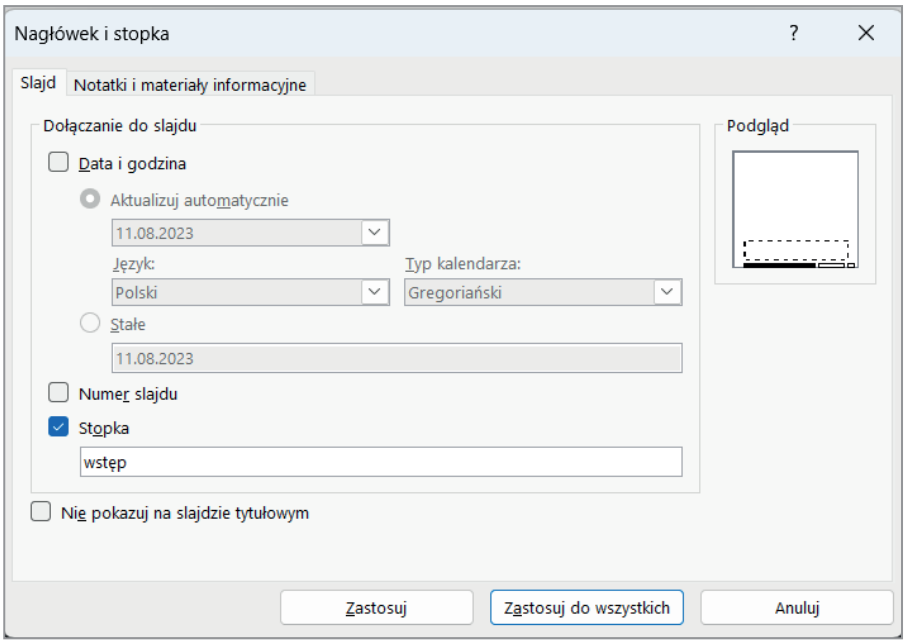
Aby dodać przypis do slajdu prezentacji, należy wybrać slajd, na którym zostanie on wstawiony, następnie wybrać kartę Wstawianie i polecenie Nagłówek i stopka w sekcji Tekst (➥ **ryc. 6.22**).



Ryc. 6.22. Widok paska narzędzi w karcie Wstawianie i wybór przycisku Nagłówek i stopka

Kolejnym krokiem jest wybór opcji Stopka i dodanie potrzebnego nam tekstu. W tym samym oknie istnieje możliwość wstawienia numeracji slajdów (jeśli jest taka potrzeba). Aby przypisy nie pojawiły się na slajdzie tytułowym, należy zaznaczyć opcję Nie pokazuj na slajdzie tytułowym.

W zależności od założeń wprowadzone powyżej zmiany można zastosować zarówno dla jednego slajdu, jak i dla całej prezentacji. W tym celu należy kliknąć przycisk Zastosuj lub Zastosuj do wszystkich (➥ **ryc. 6.23**).

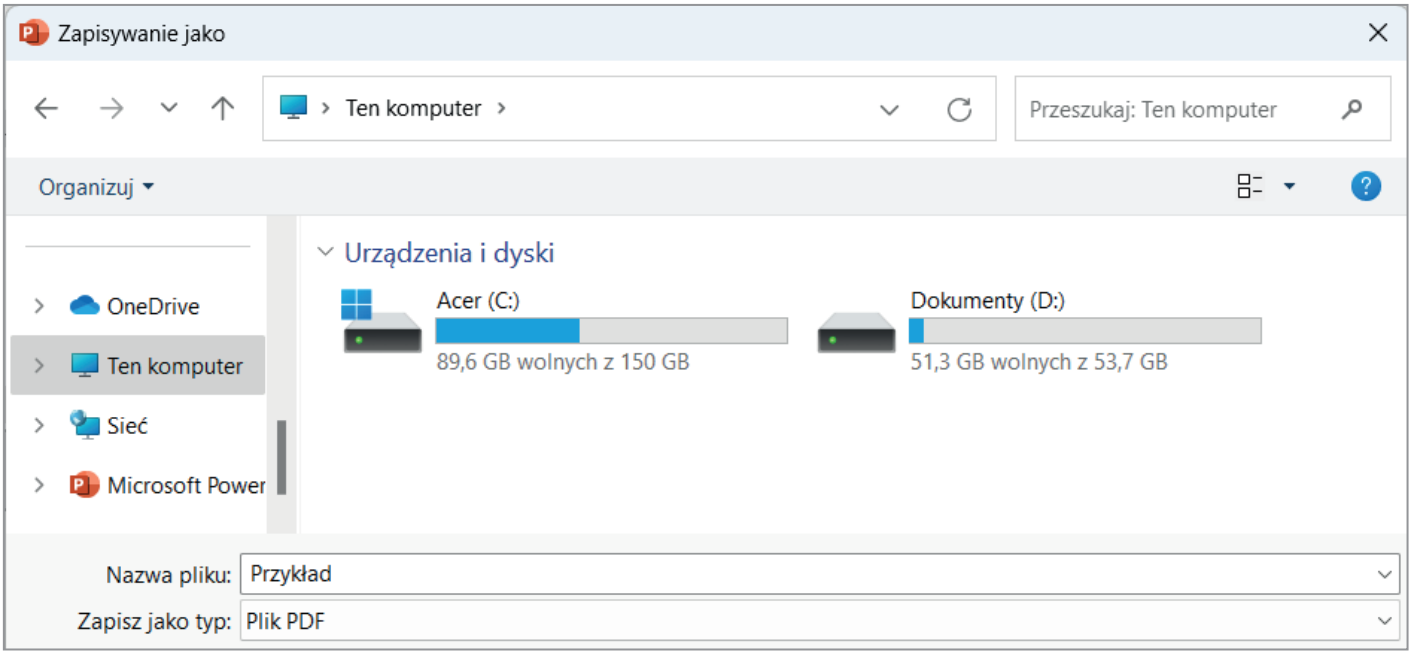


Ryc. 6.23. Okno do pracy z nagłówkiem i stopką

6.3.4.8. Zachowanie formatowania w pliku PDF

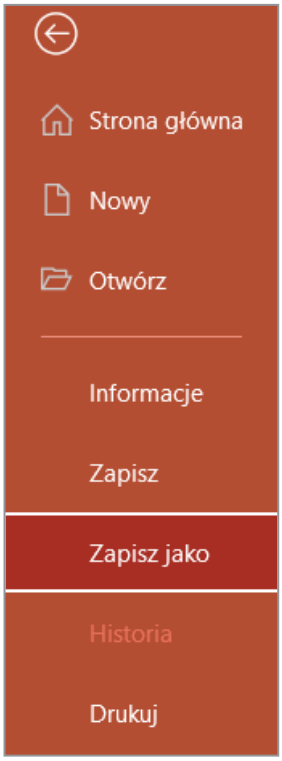
Zdarzają się przypadki, gdy podczas otwierania tej samej prezentacji na różnych komputerach (w różnych wersjach programu PowerPoint) zostaje utracone formatowanie na slajdach prezentacji lub w ogóle nie można otworzyć prezentacji. Aby uniknąć tego problemu, można po prostu zapisać prezentację w formacie PDF (➥ **ryc. 6.24**).

Ryc. 6.24. Wybór formatu PDF do zapisu prezentacji



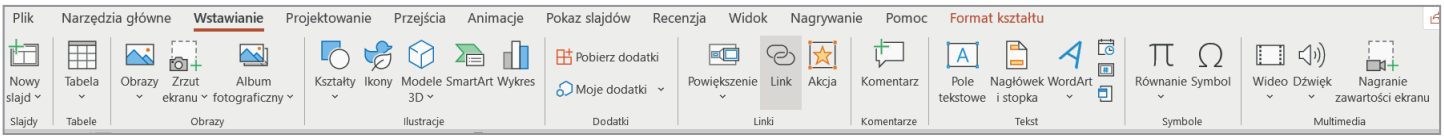
Aby to zrobić, wybierz kartę Plik, w niej otwórz sekcję Zapisz jako, a następnie po prawej stronie paska bocznego wybierz Przeglądaj. W wyskakującym nowym oknie, które się pojawi, należy wybrać lokalizację, w której chcemy zapisać nowy plik. W razie potrzeby można zmienić nazwę pliku w polu do tego przeznaczonym, a poniżej w polu Zapisz jako typ należy wybrać Plik PDF (lub inny wymagany format zapisu pliku). Na koniec klikamy przycisk Zapisz (➥ **ryc. 6.25**).

Ryc. 6.25. Wybór lokalizacji zapisu pliku



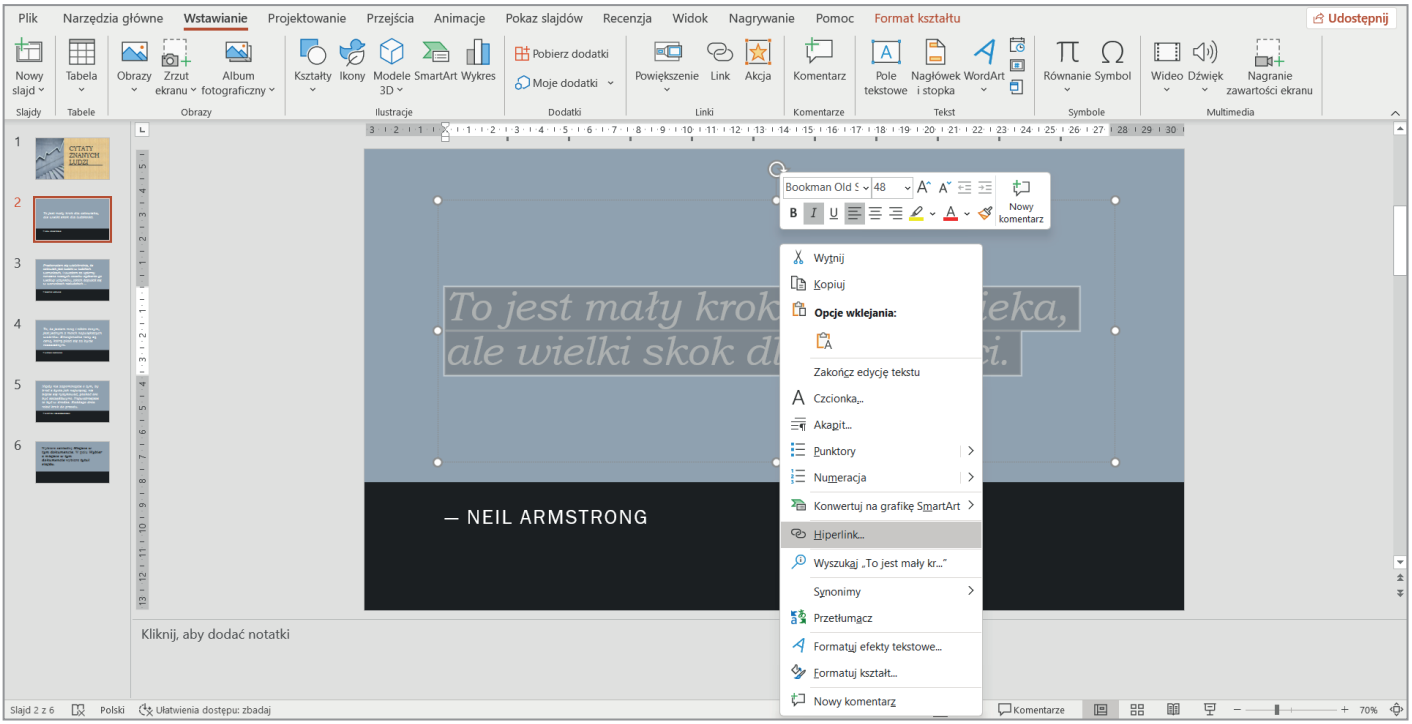
6.3.4.9. Wstawianie hiperłącza

Aby dodać hiperłącze do prezentacji, należy zaznaczyć na slajdzie prezentacji odpowiedni fragment tekstu, następnie na wstążce wybrać kartę Wstawianie i w sekcji Linki kliknąć polecenie Link (➥ **ryc. 6.26**).



Ryc. 6.26. Widok paska narzędzi w zakładce Wstawianie i wybór przycisku Link

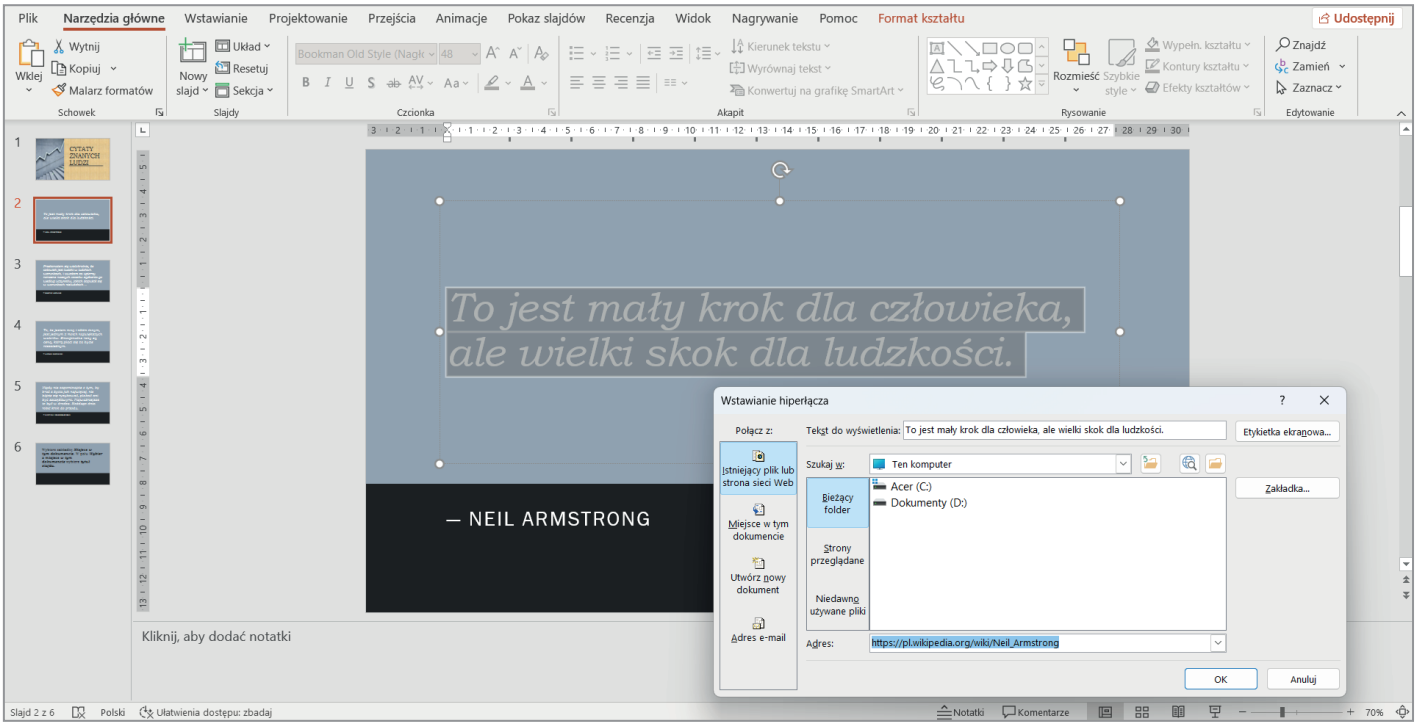
Można także kliknąć prawym przyciskiem myszy zaznaczony tekst i w nowym oknie wybrać Hiperlink (➥ **ryc. 6.27**).



Ryc. 6.27. Przykład nowego okna wyboru funkcji Hiperlink po kliknięciu prawym przyciskiem myszy na zaznaczony tekst

Następnie w wyskakującym oknie w sekcji Połącz z: należy wybrać pierwszy od góry przycisk Istniejący plik lub strona sieci Web, w sekcji Szukaj w: opcję miejsca docelowego wedle potrzeby, a potem w polu Adres w odpowiednim miejscu wkleić dane miejsca docelowego i kliknąć OK (➥ **ryc. 6.28**).

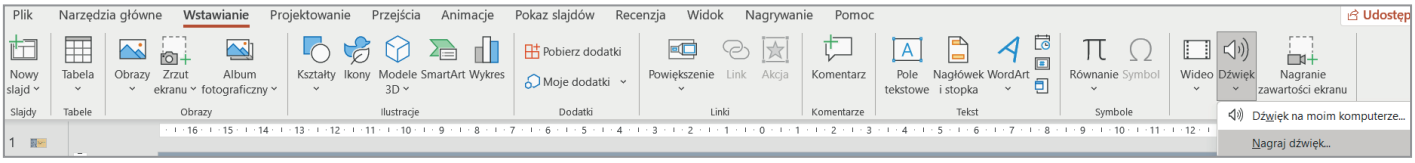
Ryc. 6.28. Kolejne okno po kliknięciu przycisku Hiperlink



6.3.4.10. Wstawianie plików audio i wideo

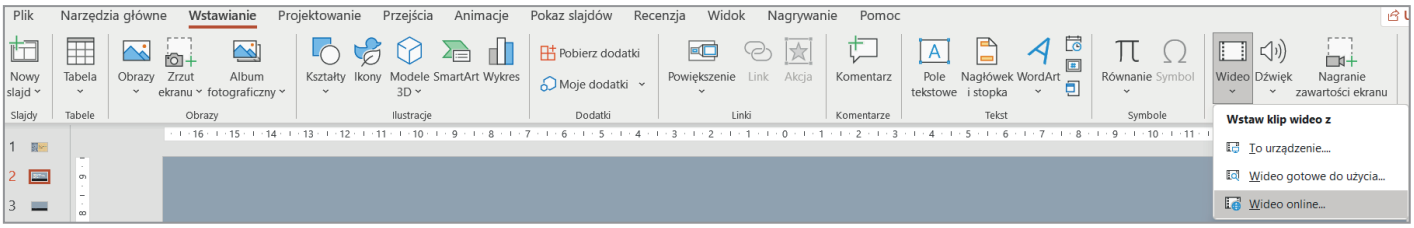
Aby wstawić do slajdu muzykę lub wideo z komputera, należy wybrać odpowiedni slajd prezentacji, który ma zawierać element dźwiękowy, a następnie z karty Wstawianie, z sekcji Multimedia wybrać polecenie Dźwięk i z rozwijanego menu opcję Dźwięk na moim komputerze. W wyskakującym oknie, które się pojawi, należy wybrać lokalizację żądanego pliku audio (➥ **ryc. 6.29**).

Ryc. 6.29. Widok paska narzędzi w zakładce Wstawianie i wybór przycisku Dźwięk



Podobne czynności należy wykonać, aby wstawić do prezentacji plik wideo, z tą różnicą, że w sekcji Multimedia należy kliknąć przycisk Wideo (➥ **ryc. 6.30**). Następnie wykonuje się te same czynności co przy wstawianiu pliku audio.

Ryc. 6.30. Widok paska narzędzi w zakładce Wstawianie i wybór przycisku Wideo

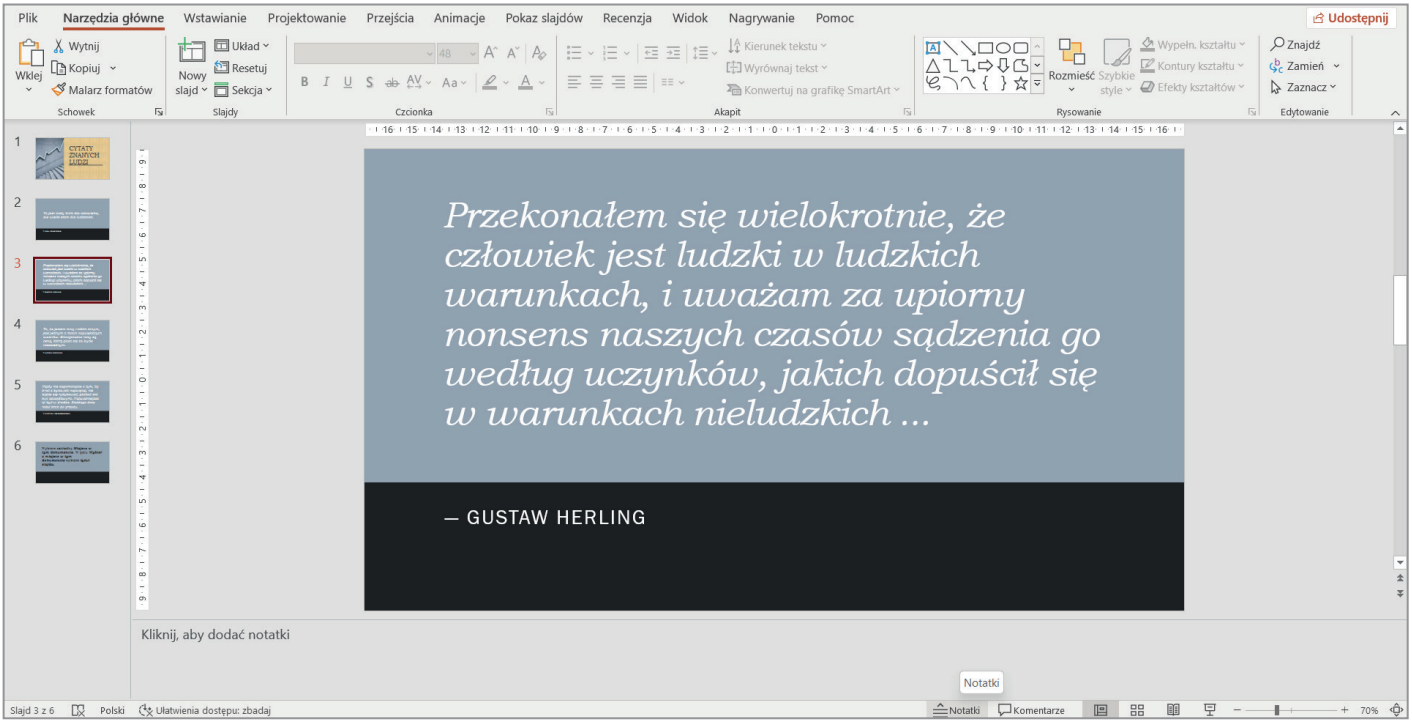


6.3.4.11. Dodawanie notatek

Dla wygody można dodać notatki, które w trakcie prezentacji nie są widoczne dla słuchaczy, a mogą bardzo pomóc prowadzącemu prezentację. W tym celu na pasku u dołu okna trzeba kliknąć lewym przyciskiem myszy Notatki (➥ **ryc. 6.31**). Bezpośrednio pod samym slajdem pojawi się pole do dodania notatek dla prelegenta (dość często pole do dodawania notatek już początkowo jest otwarte, co oznacza, że nie zawsze trzeba wykonywać podobne czynności, aby je otworzyć).

Należy pamiętać, że notatki wpisuje się dla każdego slajdu osobno.

Ryc. 6.31. Miejsce na wpisywanie notatek

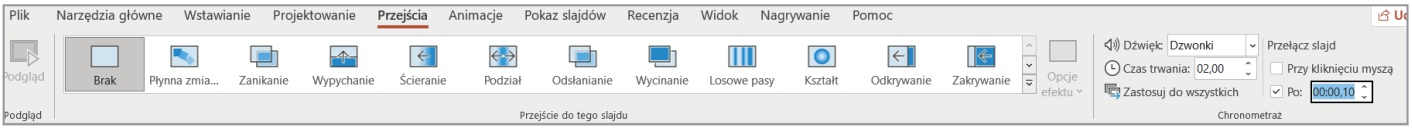


6.3.4.12. Automatyczne przełączanie slajdów

Dla wygody prelegenta istnieje także możliwość włączenia automatycznego przełączania slajdów (w określonych odstępach czasu) w trakcie prezentacji.

Aby to zrobić, wybierz kartę Przejścia na wstążce i po prawej stronie w sekcji Chronometraż określ wszystkie parametry automatycznego przełączania slajdów (czas trwania, przełączanie slajdu i inne niezbędne ustawienia, ➥ **ryc. 6.32**).

Ryc. 6.32. Widok paska narzędzi w zakładce Przejścia i wybór Chronometrażu



6.4. Tworzenie efektywnych slajdów

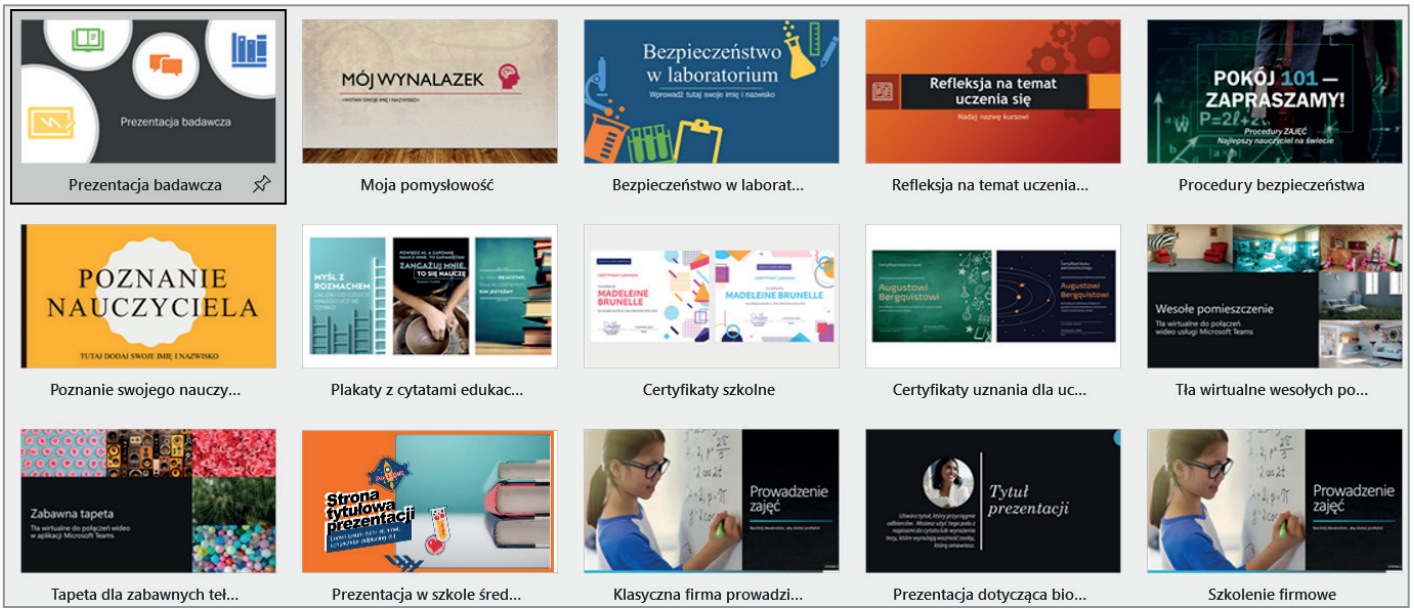
6.4.1. Zasady dotyczące tekstu i czcionek

Znaczącą rolę w projektowaniu efektywnych slajdów odgrywa tekst, jego prezentacja i czytelność, które są nieodłącznym elementem procesu komunikacji. Przestrzeganie zasad dotyczących tekstu i czcionek ułatwia odczytanie i zrozumienie prezentowanych treści.

Czcionka jest jednym z podstawowych elementów wykorzystywanych w procesie tworzenia prezentacji. Wybór właściwego kroju pisma może znacząco wpływać na odbiór prezentacji. Powszechnie zaleca się stosowanie czcionek bezszeryfowych, takich jak Arial lub Calibri, które są łatwe do odczytu na ekranie. Wielkość powinna być dostateczna, aby umożliwić odczytanie tekstu przez wszystkich uczestników – minimalny rozmiar czcionki nie powinien być mniejszy niż 24 punkty.

Kolor tekstu powinien kontrastować z tłem slajdu, aby ułatwić odczytanie. Najczęściej stosowane jest połączenie czarnego tekstu na białym tle. Inne kombinacje mogą być równie efektywne, pod warunkiem że zapewniają one odpowiednią widoczność (➥ **ryc. 6.33**).

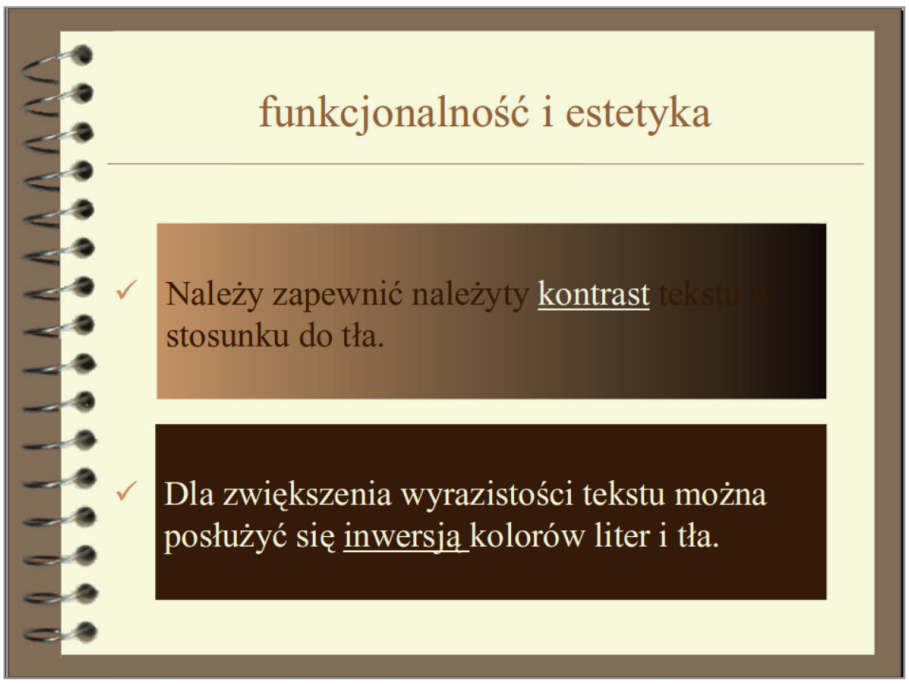
Ryc. 6.33. Przykład różnych kombinacji kolorów czcionki i tła



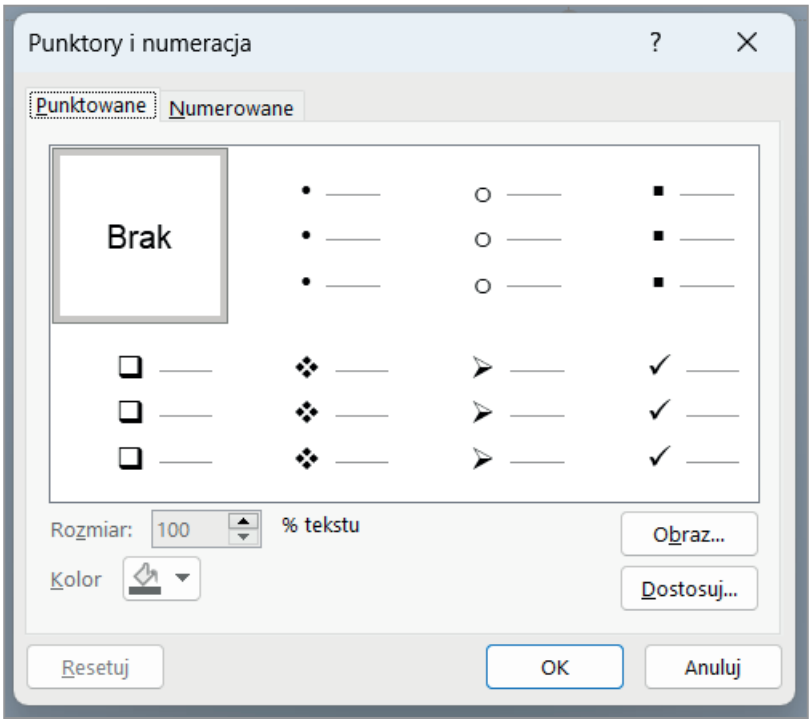
W odniesieniu do samej treści niezwykle istotna jest zasada „mniej znaczy więcej". Tekst na slajdach powinien być zwięzły, klarowny i skupiać się na najważniejszych punktach. Odbiorcy powinni być w stanie szybko przeczytać i zrozumieć treść każdego slajdu (➥ **ryc. 6.34**).

Ryc. 6.34. Przykład odpowiednio zaprojektowanego tekstu

Źródło: https://metodyka.wdfiles.com/local-files/book%3Azasady-tworzenia-prezentacji/o%20prezentacjach_6



Ciągłe bloki tekstu są zazwyczaj trudne do odczytania. Zdecydowanie zaleca się stosowanie punktorów do organizacji treści. Każdy punktor powinien reprezentować jeden konkretny pomysł lub informację. Podobnie w organizacji skomplikowanych treści i idei może pomóc stosowanie podpunktów (➡ **ryc. 6.35**).



Ryc. 6.35. Widok okna Punktory i numeracja

Zastosowanie tych zasad umożliwi stworzenie prezentacji estetycznych, czytelnych i skutecznych w komunikacji naukowej. Pamiętaj jednak, że są to wytyczne, a nie twarde zasady. Właściwy wybór czcionek, kolorów i stylów zależy od kontekstu, treści i odbiorców. Dlatego ważne, aby zawsze dostosowywać te wskazówki do określonych potrzeb i celów.

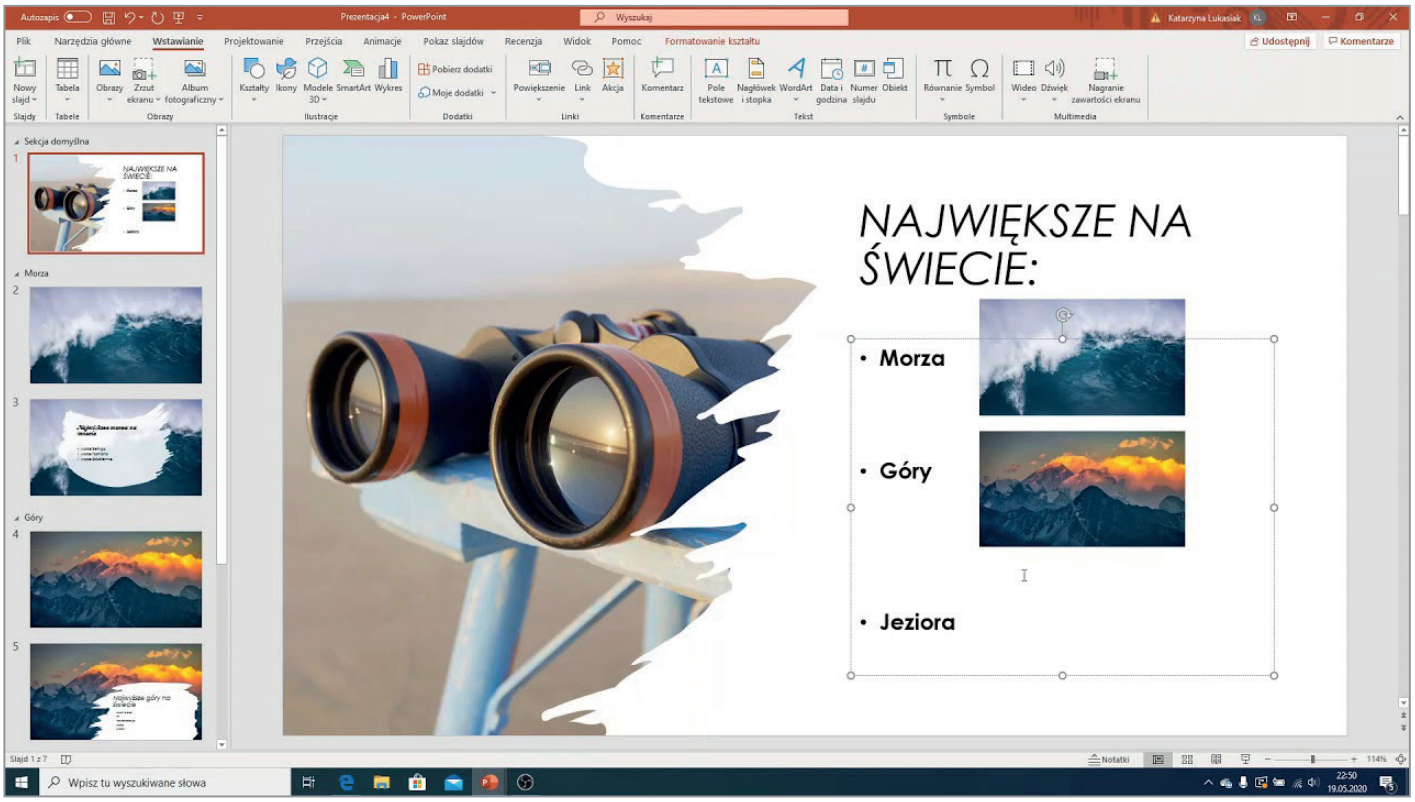
6.4.2. Wykorzystanie grafiki, tabel i wykresów

Przekazywanie informacji za pomocą obrazów, takich jak grafika, tabele i wykresy, może być bardziej skuteczne niż używanie tekstu, szczególnie w kontekście prezentowania danych naukowych i statystycznych. Wykorzystanie odpowiednio dobranych obrazów ułatwia zrozumienie złożonych koncepcji i wzmacnia przekaz zawarty w tekście.

Ryc. 6.36. Przykład efektywnego wykorzystania grafiki w prezentacji

Źródło: https://geex.x-kom.pl/lifestyle/powerpoint-jak-zrobic-efektowna-prezentacje-krok-po-kroku/

Grafika może obejmować różne rodzaje obrazów, takie jak fotografie, rysunki, diagramy, mapy i infografiki. Warto jednak pamiętać, że powinny być one odpowiednie do tematu prezentacji i przyczyniać się do zrozumienia prezentowanych treści (➡ **ryc. 6.36**).



Ryc. 6.37. Przykład wyboru tabeli dla prezentacji

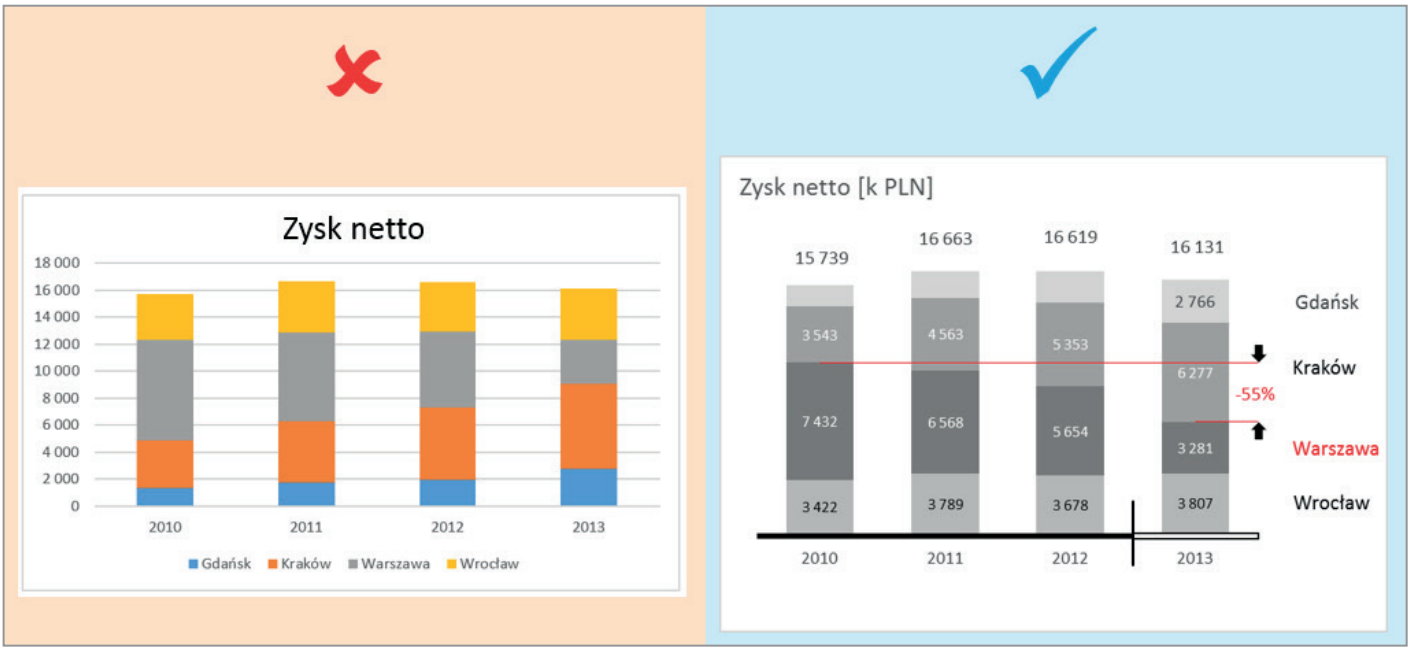
Źródło: https://www.templatemonster.com/pl/szablony-powerpoint-306050.html

Tabele są przydatne do prezentowania danych, które są zbyt skomplikowane lub nieodpowiednie do przedstawienia za pomocą wykresów. Powinny być proste i zrozumiałe, z wyraźnie oznaczonymi nagłówkami kolumn i wierszy. Mogą być szczególnie przydatne w medycynie i biostatystyce, gdzie często prezentuje się duże ilości danych (➡ **ryc. 6.37**).

Ryc. 6.38. Przykładowe wykresy stosowane w prezentacji naukowej

Źródło: https://skuteczneraporty.pl/blog/zaawansowany-wykres-do-prezentacji-powerpoint-w-20-krokach/)

Wykresy są szczególnie przydatne w prezentowaniu złożonych danych statystycznych i naukowych. Mogą skutecznie ilustrować trendy, wzorce i relacje między zmiennymi. Wybór odpowiedniego typu wykresu zależy od charakteru i złożoności danych. Wykresy słupkowe, liniowe, kołowe, punktowe są często używane do przedstawiania różnych typów danych (➥ **ryc. 6.38**).



Grafika, tabele i wykresy bardzo skutecznie przedstawiają informacje, zwłaszcza w kontekście nauki i statystyki. Ich odpowiednie użycie może znacznie ułatwić zrozumienie przedstawianych treści i sprawić, że prezentacja będzie bardziej interesująca. Warto jednak pamiętać, że wszystkie elementy graficzne powinny być stosowane z umiarem i zawsze służyć poprawie zrozumienia, a nie tylko dekoracji.

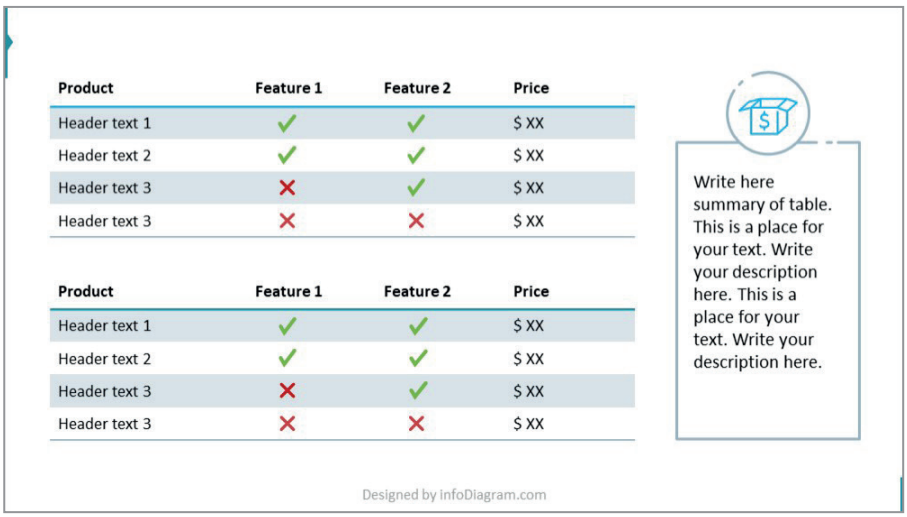
6.4.3. Jak efektywnie prezentować dane statystyczne

Przedstawianie danych statystycznych jest jednym z kluczowych elementów prezentacji naukowej, zwłaszcza w dziedzinach takich jak medycyna czy biostatystyka. Skuteczne przedstawienie ich wymaga przemyślanych decyzji dotyczących formatu, stylu i metody przekazywania informacji.

Dane można przedstawić w formie tabeli, gdy trzeba przekazać precyzyjne wartości, szczegółowe informacje lub porównać wiele zmiennych jednocześnie. Należy pamiętać, że tabele powinny być proste i łatwe do zrozumienia, a także wyraźnie podkreślać najważniejsze informacje (➥ **ryc. 6.39**).

Ryc. 6.39. Przykład tabeli z dobrze zaprezentowanymi danymi statystycznymi

Źródło: https://prezentio.com/pl/7-sposobow-atrakcyjne-tabele-powerpoint/



Product	Feature 1	Feature 2	Price
Header text 1	✓	✓	$ XX
Header text 2	✓	✓	$ XX
Header text 3	✗	✓	$ XX
Header text 3	✗	✗	$ XX

Product	Feature 1	Feature 2	Price
Header text 1	✓	✓	$ XX
Header text 2	✓	✓	$ XX
Header text 3	✗	✓	$ XX
Header text 3	✗	✗	$ XX

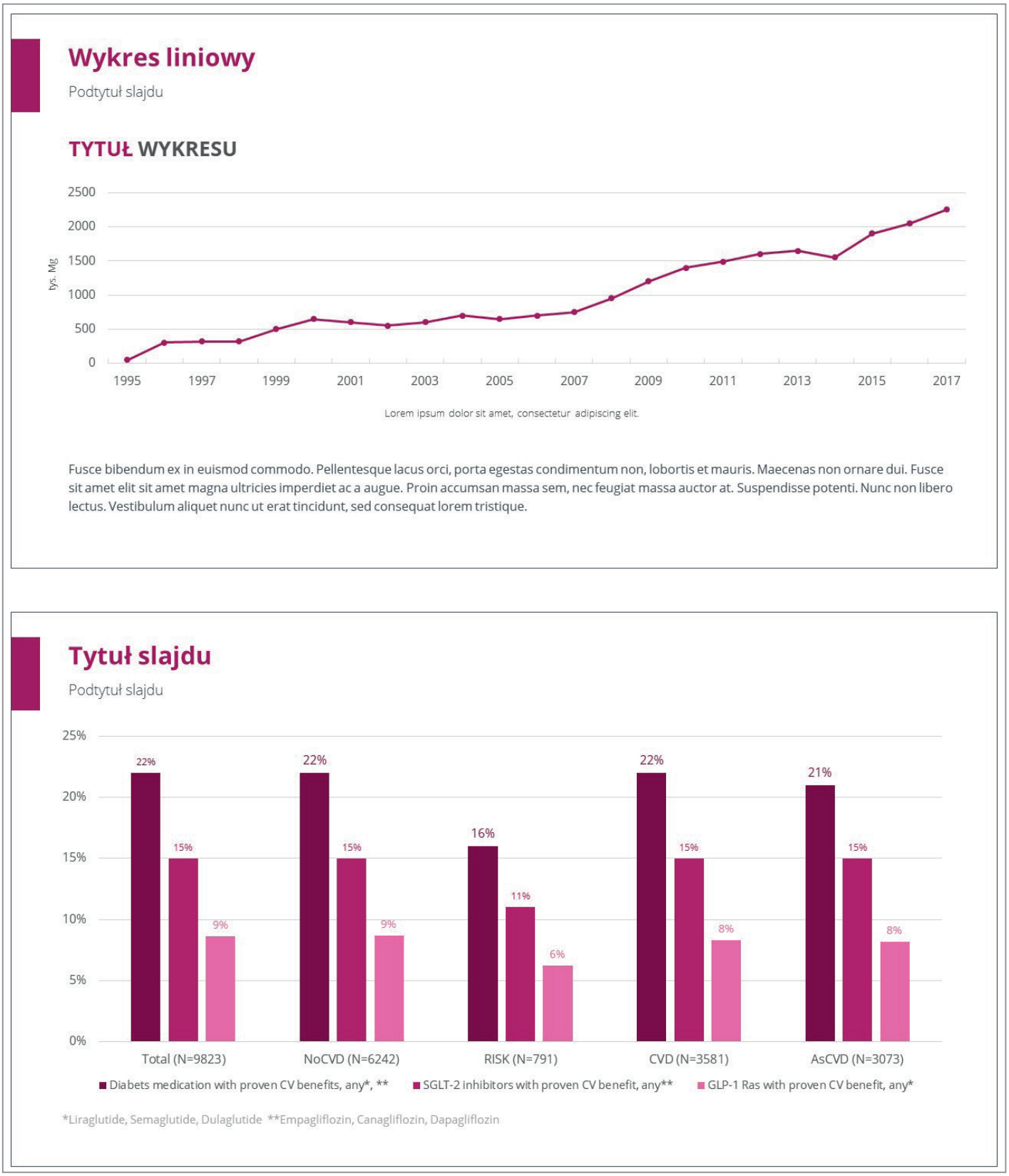
Write here summary of table. This is a place for your text. Write your description here. This is a place for your text. Write your description here.

Designed by infoDiagram.com

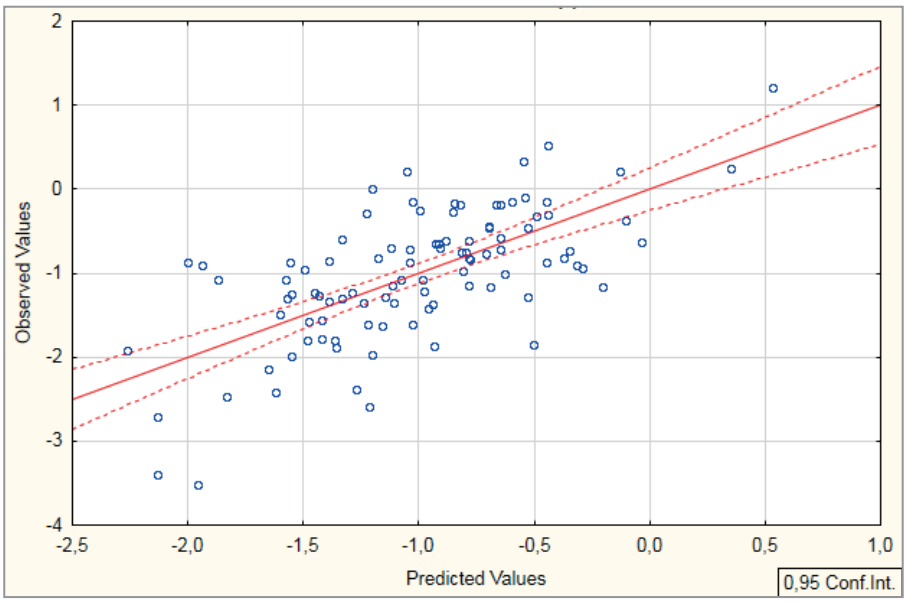
Ryc. 6.40. Wybór właściwego wykresu

Źródło: https://tworzenieprezentacji.pl/wykresy-w-powerpoint/

Wykresy mogą być skuteczniejsze w przedstawianiu zależności między zmiennymi, trendów czy rozkładów. Najlepiej dostosować typ wykresu do typu danych i celu prezentacji, np. wykresy liniowe dobrze przedstawiają trendy, podczas gdy wykresy słupkowe lepiej pokażą porównania między grupami (➥ **ryc. 6.40**).



Statystyki opisowe, takie jak średnie, mediana, zakres, odchylenie standardowe itp., mogą pomóc w przekazaniu podstawowych informacji o zbiorze danych. Jednak niewykluczone, że złożone metody statystyczne, takie jak regresja, analiza wariancji czy analiza skupień, będą wymagać specjalnego uwzględnienia w prezentacji, aby były łatwo zrozumiałe dla odbiorcy. W tym celu pomocne może być użycie wizualizacji danych lub uproszczonych schematów (➥ **ryc. 6.41**).



Ryc. 6.41. Przykład prezentowania wyników analizy regresji

Źródło: https://docplayer.pl/34183730-Analiza-regresji-urszula-smyczynska.html

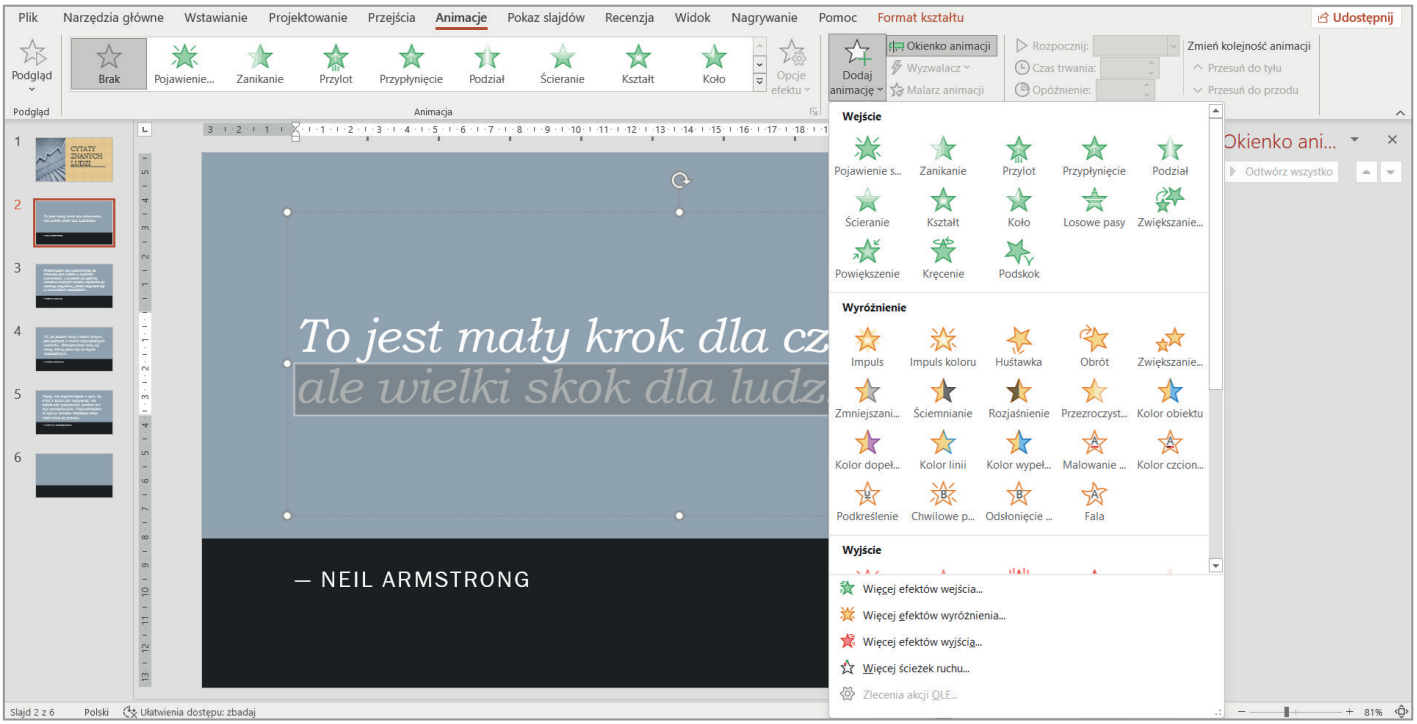
Pamiętaj, że cel prezentowania danych statystycznych to nie tylko przekazanie informacji, ale także umożliwienie odbiorcom zrozumienia i interpretacji tych danych. Dlatego ważne jest, aby podczas prezentacji zawsze podkreślać ich kontekst i znaczenie, jak również możliwe wnioski i ograniczenia.

Skuteczne przedstawienie danych statystycznych wymaga starannego planowania i dobrego zrozumienia zarówno danych, jak i odbiorców prezentacji. Wykorzystanie odpowiednich metod prezentacji danych, takich jak tabele i wykresy, w połączeniu z klarownym i precyzyjnym przekazem może znacznie zwiększyć efektywność prezentacji naukowej.

6.4.4. Użycie animacji i przejść pomiędzy slajdami

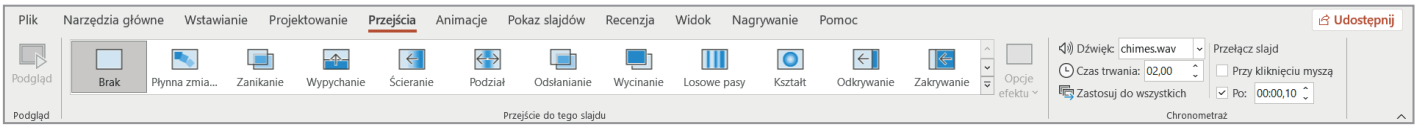
Zastosowanie animacji i przejść między slajdami może dodatkowo wzbogacić prezentację i zwiększyć jej atrakcyjność. Nie należy jednak zapominać, że ich nadużycie może przynieść efekt odwrotny do zamierzonego – zamiast ułatwić zrozumienie prezentacji, może je utrudnić.

Animacje, jeśli są stosowane umiejętnie, mogą pomóc w lepszym zrozumieniu skomplikowanych koncepcji, na przykład pokazując sekwencję zdarzeń lub zmiany w czasie. Mogą również pomóc skupić uwagę odbiorców na najważniejszych punktach prezentacji. Jednak zbyt efektowna lub niewłaściwa animacja może rozpraszać i irytować, a nawet przysłonić główne przesłanie prezentacji (➥ **ryc. 6.42**).



Ryc. 6.42. Przykład dobrego użycia animacji na slajdzie

Przejścia między slajdami, jeśli są dobrze zaplanowane i spójne, mogą nadać płynność prezentacji i pomóc utrzymać uwagę odbiorców. W przypadku długich i skomplikowanych prezentacji są również sposobem na pomoc w nawigacji i utrzymaniu struktury. Należy jednak unikać zbyt efektownych przejść, które mogą odciągać uwagę od treści prezentacji (➥ **ryc. 6.43**).



Ryc. 6.43. Widok paska narzędzi w zakładce Przejścia i wybór odpowiedniego typu przejścia między slajdami

Najważniejsze, aby pamiętać, że animacje i przejścia między slajdami są narzędziami mającymi na celu wsparcie głównego przekazu prezentacji, a nie zastąpienie go. Ważne jest, aby przy ich użyciu zawsze brać pod uwagę cel prezentacji, charakterystykę odbiorców, a także kontekst, w którym prezentacja jest wygłaszana.

Umiejętne użycie animacji i przejść pomiędzy slajdami może znacznie podnieść jakość prezentacji, zwiększając jej atrakcyjność dla odbiorców i poprawiając efektywność przekazania skomplikowanych koncepcji. Zawsze jednak należy pamiętać o zachowaniu równowagi i dostosowaniu tych narzędzi do konkretnej prezentacji i odbiorców.

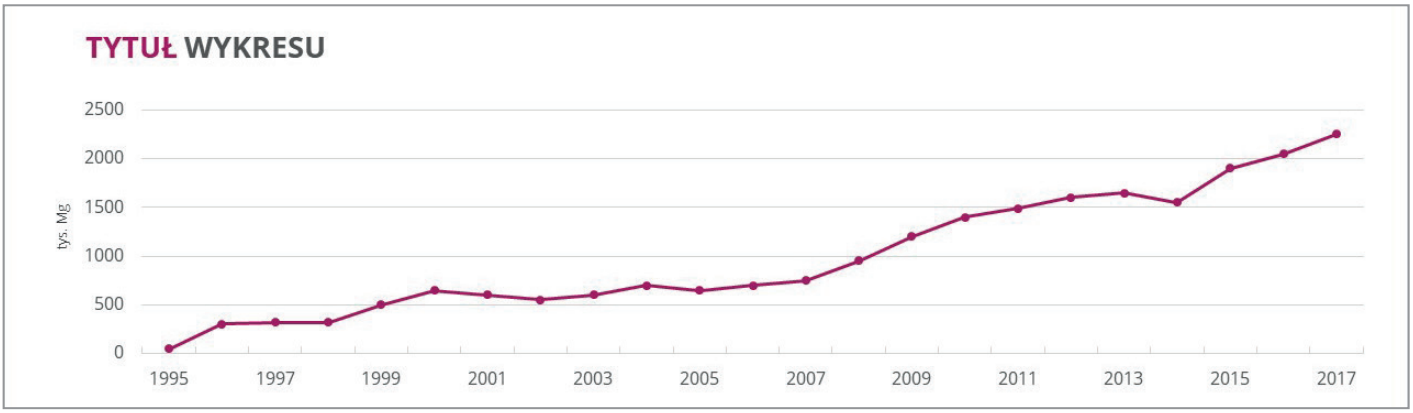
6.5. Prezentacja danych statystycznych

6.5.1. Wybór odpowiednich wykresów i diagramów do prezentacji danych

Z jednej strony wybór odpowiednich wykresów i diagramów jest istotnym aspektem prezentowania danych statystycznych, ponieważ za pomocą dobrze zaprojektowanego wykresu można przekazać skomplikowane informacje w sposób zrozumiały

i przystępny. Z drugiej natomiast niewłaściwie dobrane formy wizualizacji mogą prowadzić do nieporozumień lub nawet do błędnej interpretacji danych.

Wykresy słupkowe i kołowe, najbardziej powszechne, są doskonałe do porównywania wartości lub pokazywania proporcji. Mogą jednak być mniej efektywne, gdy prezentuje się duże ilości danych lub kiedy dane są skomplikowane lub wielowymiarowe. Możliwe, że w takich przypadkach wykresy liniowe, punktowe lub pudełkowe będą bardziej odpowiednie (➥ **ryc. 6.44**).



Ryc. 6.44. Przykład jednego z typów wykresów stosowanych do prezentacji wyników badań

Źródło: https://tworzenieprezentacji.pl/wykresy-w-powerpoint/

Wybór odpowiedniego wykresu zależy od typu danych, które chcemy przedstawić. Dane ciągłe najlepiej prezentuje się na wykresach liniowych lub punktowych, a dane kategoryczne za pomocą wykresów słupkowych lub kołowych. W niektórych przypadkach, takich jak prezentowanie rozkładów czy zależności, najbardziej odpowiednie będą wykresy pudełkowe, histogramy lub wykresy punktowe (➥ **ryc. 6.45**).

Ryc. 6.45. Przykłady wykresów dla różnych typów danych

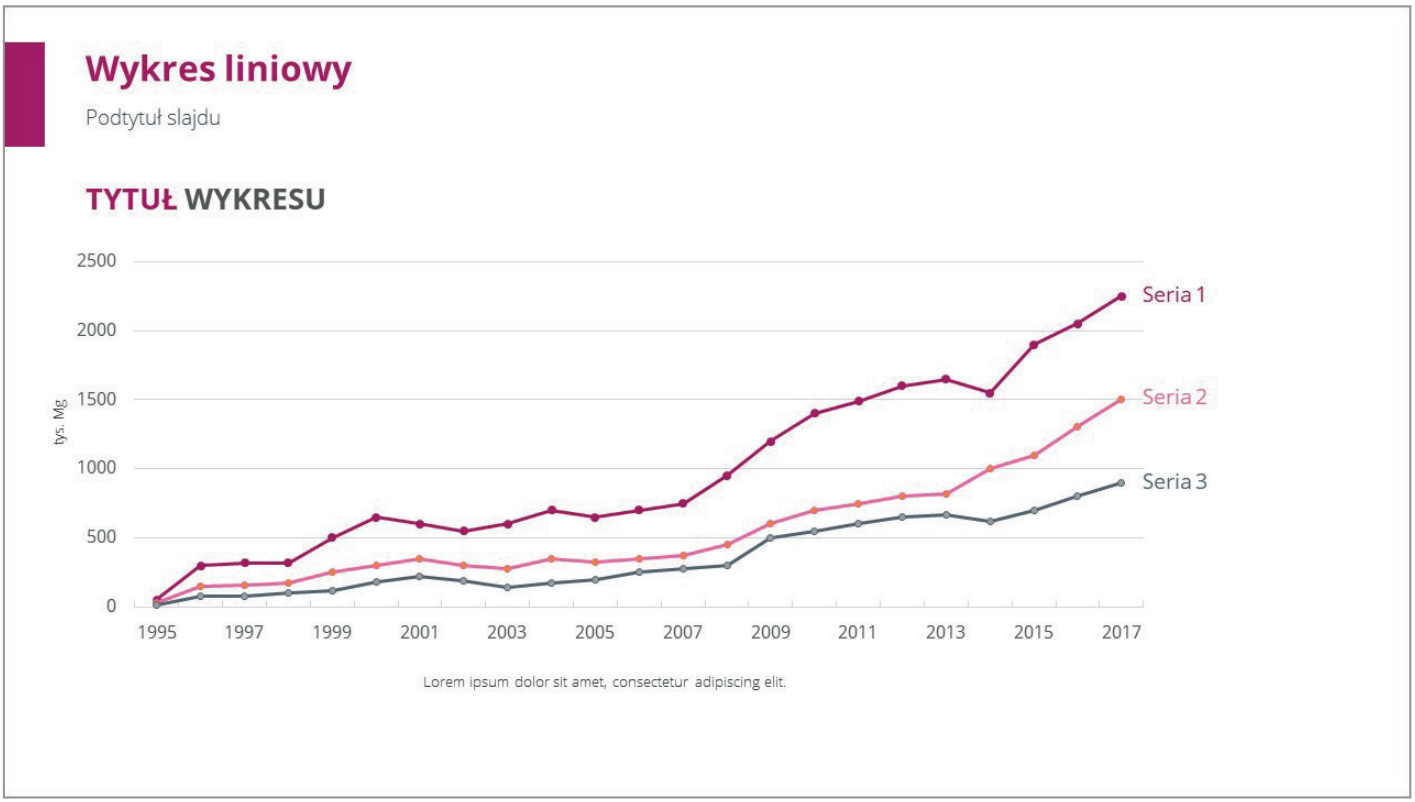
Źródło: https://prezentio.com/pl/bledy-wizualizacji-danych-powerpoint-wykresy/

Ważne, aby pamiętać, że każdy typ wykresu ma swoje mocne strony i ograniczenia i że wybór odpowiedniego wykresu zależy od charakterystyki prezentowanych danych oraz od tego, co chcemy za ich pomocą przekazać. Ponadto wykresy i diagramy powinny zawsze być jasne, zrozumiałe i łatwe do interpretacji, z odpowiednimi etykietami osi, tytułami i legendami (➥ **ryc. 6.46**).

Warto pamiętać, że właściwa prezentacja danych statystycznych to sztuka równie ważna jak umiejętność ich analizy. Stosując powyższe zasady, można pomóc odbiorcom lepiej zrozumieć i interpretować prezentowane dane.

Ryc. 6.46. Dobrze zaprojektowany wykres

Źródło: https://tworzenieprezentacji.pl/wykresy-w-powerpoint/



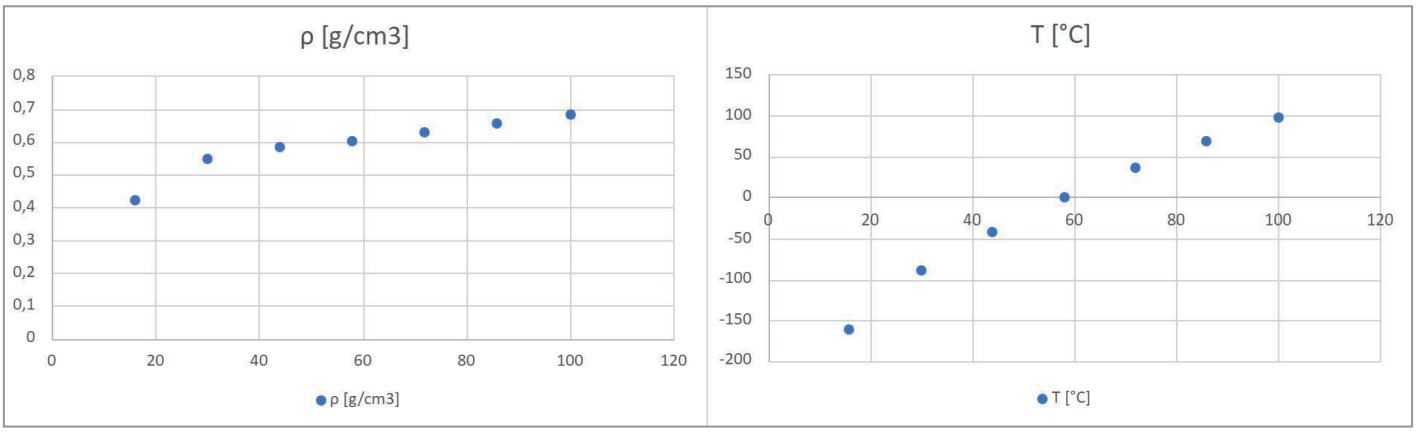
6.5.2. Prezentacja wyników analiz statystycznych

Prezentowanie wyników analiz statystycznych jest jednym z najistotniejszych etapów w procesie komunikacji naukowej. Zrozumienie tego, jak najlepiej prezentować owe wyniki, jest kluczem do efektywnego przekazywania informacji i tworzenia skutecznych argumentów naukowych.

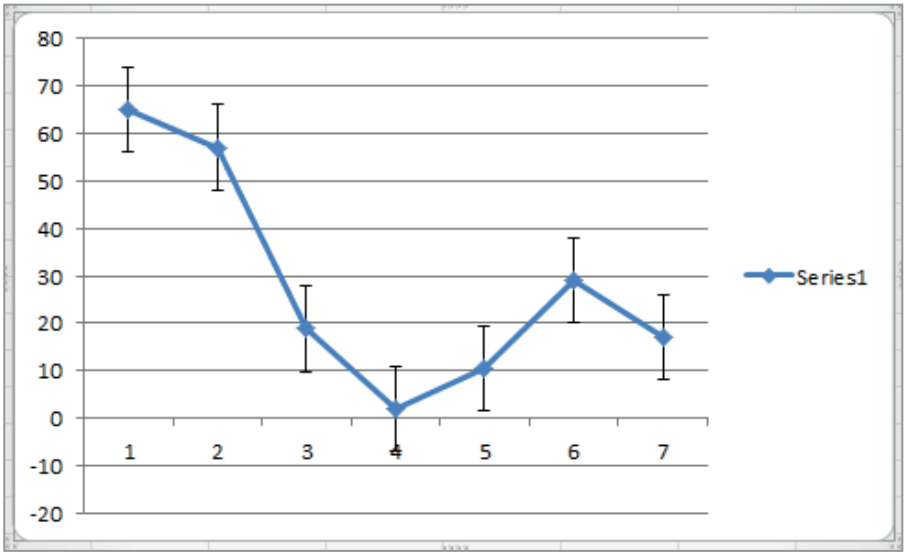
Wyniki analiz statystycznych można przedstawić na różne sposoby, w zależności od charakteru danych, celów prezentacji oraz odbiorcy. Przykładowo średnie, mediany i zakresy wartości można zaprezentować za pomocą tabel, wykresów słupkowych, punktowych lub liniowych. Do przedstawienia rozkładów wartości oraz zależności między zmiennymi często wykorzystuje się wykresy pudełkowe, histogramy czy wykresy punktowe (➥ **ryc. 6.47**).

Ryc. 6.47. Przykład przedstawienia wyników badań za pomocą prawidłowego wykresu

Źródło: https://brainly.pl/zadanie/21548645



Prezentując wyniki analiz statystycznych, zawsze należy podać stosowne miary statystyczne, takie jak wartość *p*, odchylenie standardowe czy współczynnik korelacji. Jest to kluczowe dla prawidłowego zrozumienia i interpretacji wyników (➥ **ryc. 6.48**).



Ryc. 6.48. Przykład prezentowania miar statystycznych

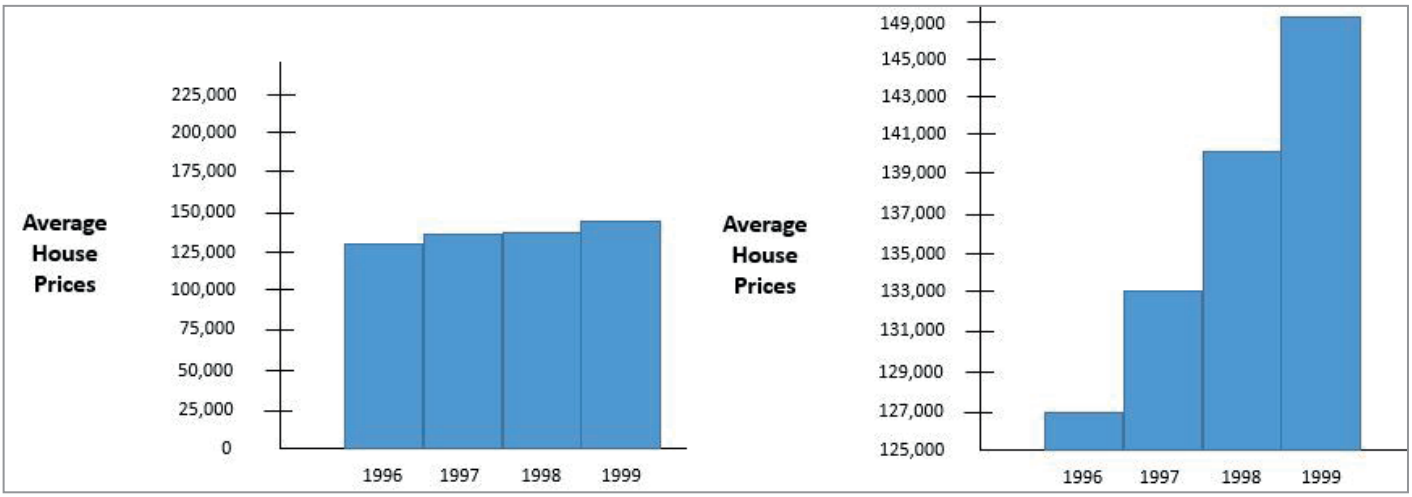
Źródło: https://www.extendoffice.com/documents/excel/863-excel-add-remove--error-bars.html

Prezentacja wyników analiz statystycznych to zadanie, które wymaga precyzyjnego przemyślenia i umiejętności korzystania z odpowiednich narzędzi wizualizacji. Powinniśmy zawsze pamiętać, że naszym głównym celem jest jak najefektywniejsze przekazanie wyników analizy w sposób zrozumiały dla odbiorcy.

6.5.3. Jak unikać błędów podczas prezentacji danych statystycznych

Niezwykle istotne jest unikanie typowych błędów, które podczas prezentacji danych mogą zmylić odbiorcę lub zniekształcić interpretację wyników. Zrozumienie tych niepożądanych rozwiązań i ich unikanie to ważny element efektywnej komunikacji naukowej.

Jednym z najczęstszych błędów jest wykorzystanie niewłaściwej skali na wykresach. Taka manipulacja może prowadzić do nieodpowiednich wniosków, dlatego warto zawsze wybierać skalę adekwatną do przedstawianych danych (➥ **ryc. 6.49**).



Ryc. 6.49. Porównanie wykresów z różnymi skalami

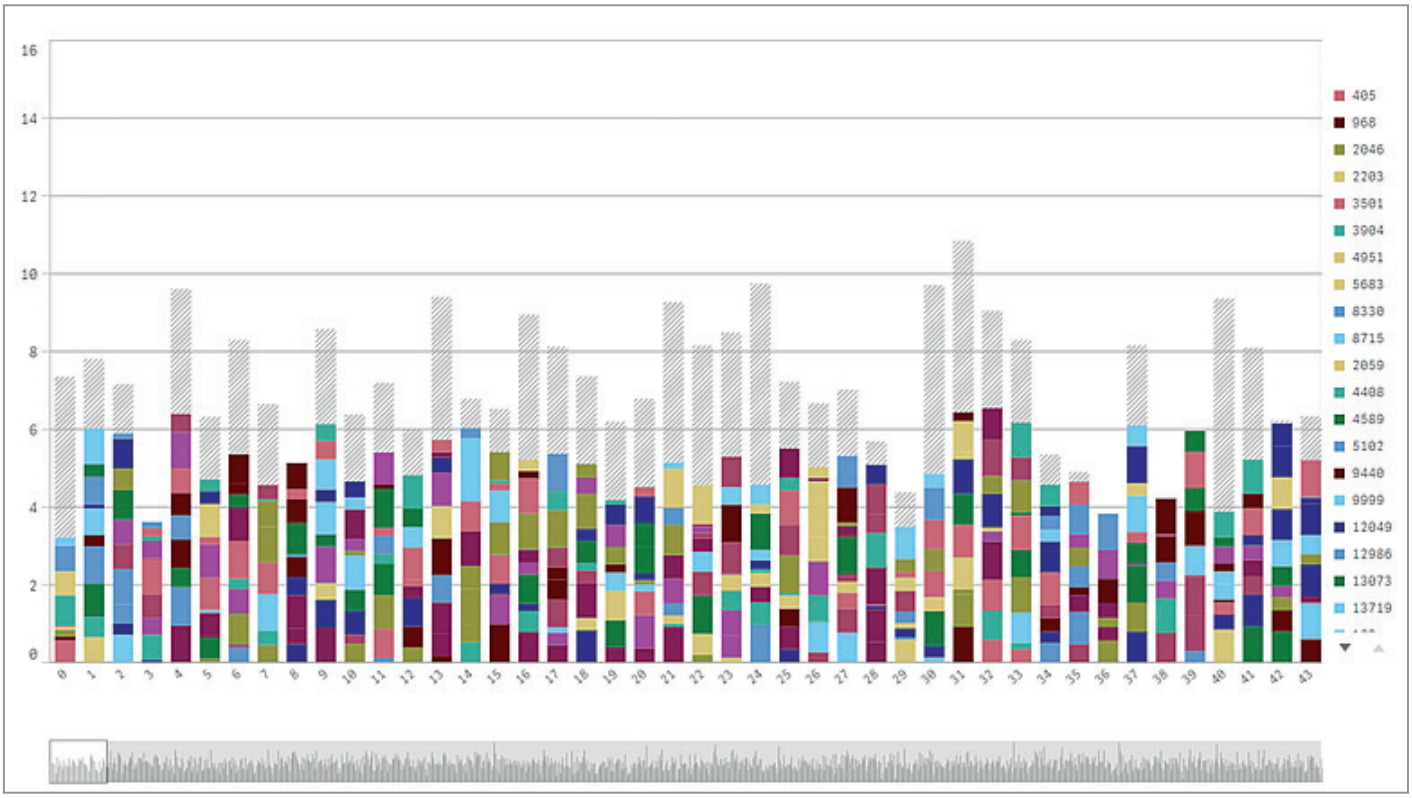
Źródło: https://study.com/learn/lesson/misleading-deceptive-bad-statistics-graphs.html

Ryc. 6.50. Przykład przeciążonego wykresu (prezentacja danych jest niewygodna do odczytania)

Źródło: https://help.qlik.com/pl-PL/sense/May2023/Subsystems/Hub/Content/Sense_Hub/Visualizations/Bar-Chart/bar-chart.htm

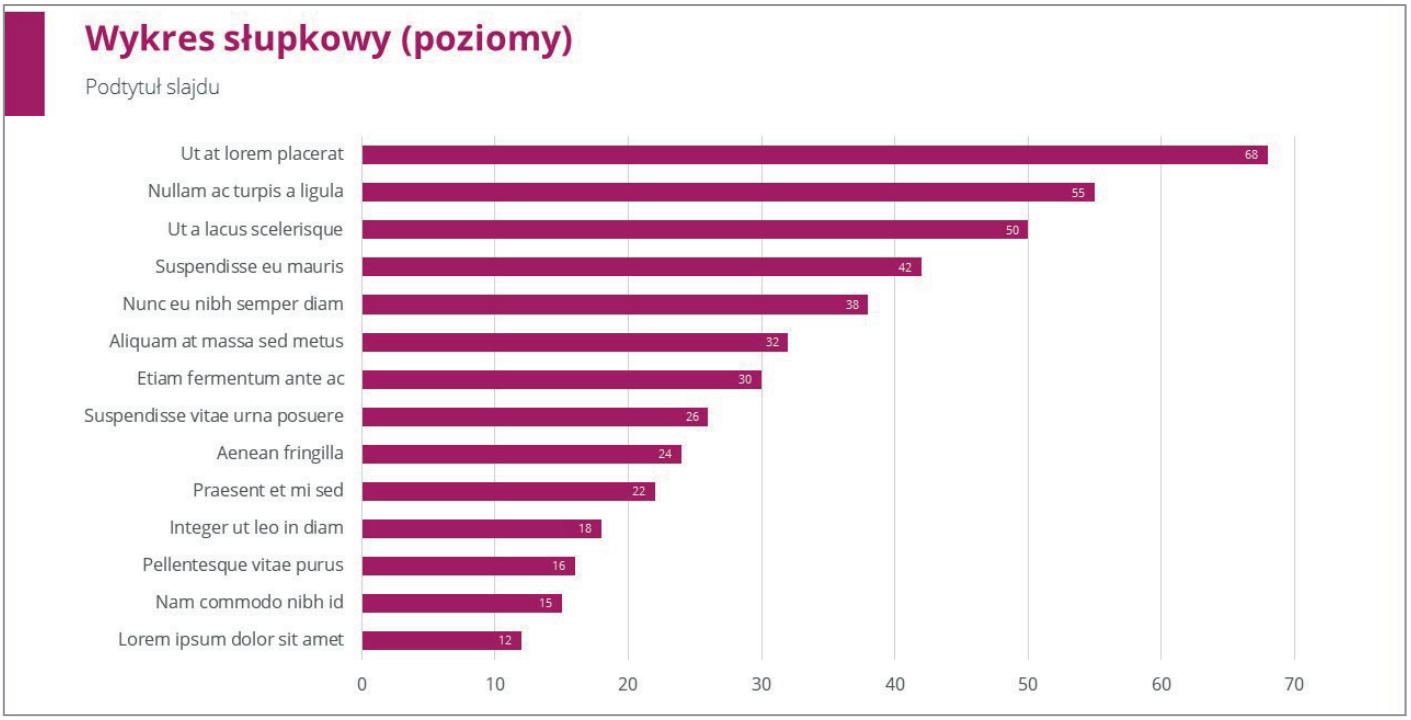
Innym częstym błędem jest prezentowanie zbyt dużej ilości danych na jednym wykresie. Może to prowadzić do przeładowania informacjami i utrudniać zrozumienie przedstawianych wyników (➥ **ryc. 6.50**).

Powszechnymi błędami są również: niedostateczne wykorzystanie opisów osi, tytułów czy legendy. Brak tych elementów może sprawić, że wyniki stają się niezrozumiałe dla odbiorcy (➥ **ryc. 6.51**).



Ryc. 6.51. Przykład wykresu bez odpowiednich oznaczeń

Źródło: https://tworzenieprezentacji.pl/wykresy-w-powerpoint/



Podczas prezentacji danych statystycznych najważniejsze jest przemyślenie, jak wyniki zostaną przedstawione odbiorcy. Należy unikać typowych błędów, takich jak niewłaściwe skalowanie wykresów, przeładowanie informacjami czy brak odpowiednich oznaczeń. Dobre praktyki prezentacji danych to podstawa skutecznej komunikacji naukowej.

6.6. Doskonalenie i edycja prezentacji

6.6.1. Przegląd i edycja prezentacji

Przegląd i edycja prezentacji są nieodłącznymi elementami procesu twórczego, mającymi na celu maksymalizację czytelności i skuteczności przekazu. Bez przeglądu i edycji łatwo przegapić błędy, które mogą mieć negatywny wpływ na odbiór prezentacji przez odbiorców.

Podczas przeglądu warto zwrócić uwagę na aspekty graficzne prezentacji, takie jak kolorystyka, czcionka, układ slajdów czy użyte ilustracje. Każdy z tych elementów powinien współgrać ze sobą, tworząc spójną całość (➥ **ryc. 6.52**).

Ryc. 6.52. Porównanie dobrze i źle zaprojektowanych slajdów

Oprócz tego kluczowe jest sprawdzenie poprawności przedstawionych danych i ich interpretacji. Błędy statystyczne mogą mieć poważne konsekwencje – z ich powodu prezentacja może być niewiarygodna

Na koniec warto również przeanalizować strukturę prezentacji i upewnić się, że jest logiczna i przejrzysta. Każda część powinna płynnie przechodzić w następną, a całość tworzyć spójną narrację.

Proces przeglądu i edycji prezentacji jest niezwykle ważny dla skutecznej komunikacji naukowej. Pozwala na wyeliminowanie ewentualnych błędów, zwiększenie czytelności i efektywności przekazu, a co za tym idzie – ułatwienie zrozumienia przedstawianych treści przez odbiorców.

6.6.2. Wprowadzanie zmian i poprawek

Wprowadzanie zmian i poprawek jest nieodzownym elementem procesu doskonalenia prezentacji. Warto zastosować zasadę iteracyjności, która polega na wielokrotnym wprowadzaniu drobnych poprawek i obserwacji ich efektów.

Pierwszym krokiem powinno być dokonanie zmian wynikających z przeglądu prezentacji. Może to obejmować modyfikacje kolorystyki, czcionek, układu slajdów, a także korektę danych i ich interpretacji.

Następnie warto skupić się na poprawie struktury prezentacji, upewniając się, że wszystkie elementy są logicznie powiązane i tworzą spójną narrację (➥ **ryc. 6.53**).

Dodatkowo powinno się przeprowadzić próbną prezentację, która może pomóc w identyfikacji obszarów wymagających dalszej poprawy. Może się np. okazać, że niektóre slajdy są przeładowane informacjami lub pewne fragmenty są niezrozumiałe dla odbiorców.

Ryc. 6.53. Struktura prezentacji z zaznaczonymi wprowadzonymi zmianami

Źródło: https://www.dr-mamczur.pl/blog/2020/diagnostyka-roznicowa-sposobem-na-pani-jest-medyczna-zagadka-2020/struktura-prezentacji-problemu-5-krokow/

Wprowadzanie zmian i poprawek jest niezbędnym elementem procesu doskonalenia prezentacji, pozwalającym na optymalizację jej formy i treści. Zastosowanie podejścia iteracyjnego umożliwia stopniowe ulepszanie i dopracowanie prezentacji.

6.6.3. Sprawdzanie zgodności i spójności danych

Jednym z najistotniejszych etapów w procesie doskonalenia prezentacji jest sprawdzanie zgodności i spójności danych. Właściwe przedstawienie danych biostatystycznych może stanowić istotne wyzwanie, a zarazem jest bardzo ważne dla zachowania wiarygodności i zrozumiałości prezentacji.

Na samym początku zawsze należy sprawdzić zgodność przedstawianych danych ze źródłem. Błędy, takie jak niedokładności lub pomyłki, mogą zniekształcić przekaz i doprowadzić do nieprawidłowych wniosków.

Kolejnym krokiem jest zapewnienie spójności danych w całej prezentacji. Niespójności, np. różne jednostki miar czy formaty danych, mogą wprowadzić w błąd i utrudnić zrozumienie prezentacji przez odbiorców.

Dodatkowo niezwykle istotne jest, aby wszelkie wykresy, tabele i grafy były zgodne z danymi, które mają reprezentować. Niewłaściwe lub mylące wizualizacje mogą wprowadzić w błąd odbiorców i przekazać nieprawdziwe informacje.

Sprawdzanie zgodności i spójności danych to bardzo istotny etap w procesie doskonalenia prezentacji. Pozwala zapewnić wiarygodność i zrozumiałość przekazu. Dokładna analiza i weryfikacja danych umożliwia eliminację błędów i niespójności, co przekłada się na wyższą jakość finalnej prezentacji.

6.6.4. Proces próbnej prezentacji i zbierania opinii

Niezbędnym etapem doskonalenia i edycji prezentacji, zwłaszcza przedstawiającej zagadnienia z zakresu medycyny i biostatystyki, jest jej próbne przedstawienie i zbieranie opinii. Prezentację ocenia wówczas grupa osób, które mogą dostarczyć cennych uwag i sugestii dotyczących jej poprawy.

Pierwszym krokiem w tym procesie jest wybór odpowiedniego zespołu do przeprowadzenia próbnej prezentacji. Powinien składać się z osób zróżnicowanych pod względem wykształcenia i zawodu, tak aby zapewnić jak najszerszą perspektywę.

Podczas próbnej prezentacji prelegent powinien zachować otwartość na konstruktywną krytykę i być gotowy do wprowadzenia poprawek do swojej prezentacji.

Bardzo ważne są zebranie i analiza opinii otrzymanych od uczestników. Mogą one dotyczyć zarówno treści prezentacji, jak i jej formy, w tym jasności przekazu, logicznej struktury czy estetyki.

Wreszcie na podstawie zebranych opinii należy rozważyć wprowadzenie zmian i usprawnień w prezentacji. Warto zwrócić uwagę na to, że etap próbnej prezentacji i zbierania opinii można powtórzyć wielokrotnie aż do uzyskania ostatecznej, pożądanej wersji. Jest niezbędny do udoskonalenia prezentacji oraz umożliwia optymalizację i maksymalizację jej skuteczności.

6.7. Prezentacja w praktyce

6.7.1. Przygotowanie do prezentacji na żywo

Kolejnym niezwykle istotnym etapem realizacji, szczególnie w obszarach naukowych, takich jak medycyna, jest przygotowanie do prezentacji na żywo. Obejmuje ono zarówno przygotowanie techniczne, jak i psychiczne prelegenta.

Przygotowanie techniczne obejmuje sprawdzenie wykorzystywanego wyposażenia i oprogramowania. Należy upewnić się, że wszystko działa prawidłowo, dzięki czemu można uniknąć niespodziewanych problemów technicznych podczas prezentacji.

Przygotowanie psychiczne jest równie ważne i obejmuje naukę wystąpień publicznych, trening redukcji stresu czy ćwiczenie prezentacji. Ważne, aby prelegent czuł się pewnie i komfortowo przed audytorium. Przekłada się to na skuteczność jego prezentacji.

W przypadku prezentacji naukowych szczególnie istotne jest także dogłębne zrozumienie prezentowanego materiału. Prelegent powinien być w stanie odpowiedzieć na pytania i wyjaśnić prezentowane koncepcje niezależnie od tego, czy są one techniczne, czy konceptualne.

Przygotowanie do prezentacji na żywo jest bardzo ważnym etapem w procesie realizacji prezentacji i obejmuje zarówno przygotowanie techniczne, jak i odpowiedni trening psychiczny. Jeśli zostały odpowiednio przeprowadzone, prelegent jest w stanie efektywnie przekazać informacje, co jest szczególnie ważne w obszarach naukowych, takich jak medycyna i biostatystyka.

6.7.2. Używanie notatek mówcy i innych pomocy

Notatki mówcy oraz dodatkowe pomoce prezentacyjne mogą znacząco ułatwić wystąpienie zwłaszcza w tak skomplikowanych dziedzinach jak medycyna i biostatystyka.

Notatki pełnią istotną rolę w organizacji prezentacji. Składają się z najważniejszych punktów, które mają zostać poruszone. To niezwykle pomocne narzędzie, szczególnie kiedy prezentacja obejmuje skomplikowane koncepcje lub analizy statystyczne, typowe dla medycyny i biostatystyki. Notatki powinny być precyzyjne i zwięzłe, co pozwoli mówcy na płynną nawigację przez materiał prezentacji bez ryzyka zagubienia się w zbyt rozbudowanych zapiskach.

Dodatkowo w notatkach można zawrzeć ważne dane techniczne lub statystyczne, które mogą okazać się pomocne w przypadku pytań od publiczności. Wymagają one jednak starannego doboru i redakcji, aby zawierały istotne informacje, a jednocześnie były zrozumiałe i łatwe do przeczytania dla mówcy.

Inne pomoce, takie jak wizualizacje danych, są niezastąpione podczas prezentacji naukowych. Ułatwiają zrozumienie skomplikowanych koncepcji i zjawisk, przekształcając złożone dane w prostsze, bardziej przystępne formy. Wykresy, diagramy, mapy ciepła czy inne formy graficzne są nieocenione w prezentacjach z zakresu biostatystyki, gdzie interpretacja danych jest najistotniejsza.

Należy jednak zwrócić uwagę, że wszelkie pomoce powinny służyć jako wsparcie dla mówcy, a nie dominować w prezentacji. Głównym źródłem informacji powinien pozostać prelegent, a materiały pomocnicze mają jedynie ułatwić zrozumienie przekazywanego materiału.

Efektywne korzystanie z notatek mówcy i innych materiałów pomocniczych jest niezbędne dla skutecznej prezentacji. Te narzędzia, gdy są właściwie stosowane, mogą znacząco podnieść jakość prezentacji i uczynić ją bardziej przystępną dla publiczności.

6.7.3. Radzenie sobie z pytaniami i interakcja z publicznością

Jednym z podstawowych aspektów udanej prezentacji jest skuteczna interakcja z publicznością. Obejmuje ona radzenie sobie z zagadnieniami, o które mogą zapytać słuchacze. Zrozumienie i przyswojenie tej umiejętności jest kluczowe, szczególnie w obszarach naukowych, takich jak medycyna i biostatystyka, gdzie treści mogą być skomplikowane i często rodzić wiele pytań.

Pierwszym krokiem jest zachowanie otwartości i gotowości na pytania. Prelegenci powinni zapewnić odpowiednie ramy czasowe dla sesji pytań i odpowiedzi. Można to dobrze zorganizować za pomocą odpowiednich narzędzi prezentacyjnych. Tworzenie atmosfery, która zachęca do zadawania pytań, jest niezwykle ważne, ponieważ zwiększa zaangażowanie publiczności i pozwala na głębsze zrozumienie prezentowanych treści.

Udzielając odpowiedzi, prelegenci powinni być spokojni i zorganizowani, mając na uwadze to, że pytania od publiczności nie są wyzwaniem, a raczej umożliwiają pogłębienie wiedzy i zrozumienie tematu. Ważne, aby odpowiedzi były jasne i precyzyjne, co może stanowić szczególne wyzwanie w przypadku skomplikowanych zagadnień medycznych albo biostatystycznych. W takich sytuacjach pomóc może odwołanie się do prezentowanych wcześniej wizualizacji danych.

Jeżeli prelegent nie jest w stanie od razu odpowiedzieć na pytanie, nie powinien tego ukrywać. Lepiej przyznać, że nie zna się odpowiedzi, sprawdzić i udzielić jej później. Ukaże to szacunek dla publiczności i podkreśli autentyczność mówcy.

Wreszcie warto zauważyć, że pytania i interakcja z publicznością mogą dostarczyć cennych informacji zwrotnych. Mogą one wskazywać na obszary, które wymagają dodatkowego wyjaśnienia, lub sugerować tematy na przyszłe prezentacje.

Zarządzanie pytaniami i interakcja z publicznością są niezbędnymi elementami dobrej prezentacji naukowej. Wymagają praktyki i przygotowania, ale korzyści, które przynoszą w formie lepszego zrozumienia, większego zaangażowania publiczności i cennej informacji zwrotnej są bezcenne.

Źródło: https://www.ewaluacja.gov.pl/media/42317/Ewaluacja_strategiczna_dzialan_info_promo_PO_KL_raport_koncowy.pdf

Krystyna Laszki-Szcząchor

7 Tworzenie baz danych pacjentów w programie Access

7.1. Zadanie: kartoteka pacjentów

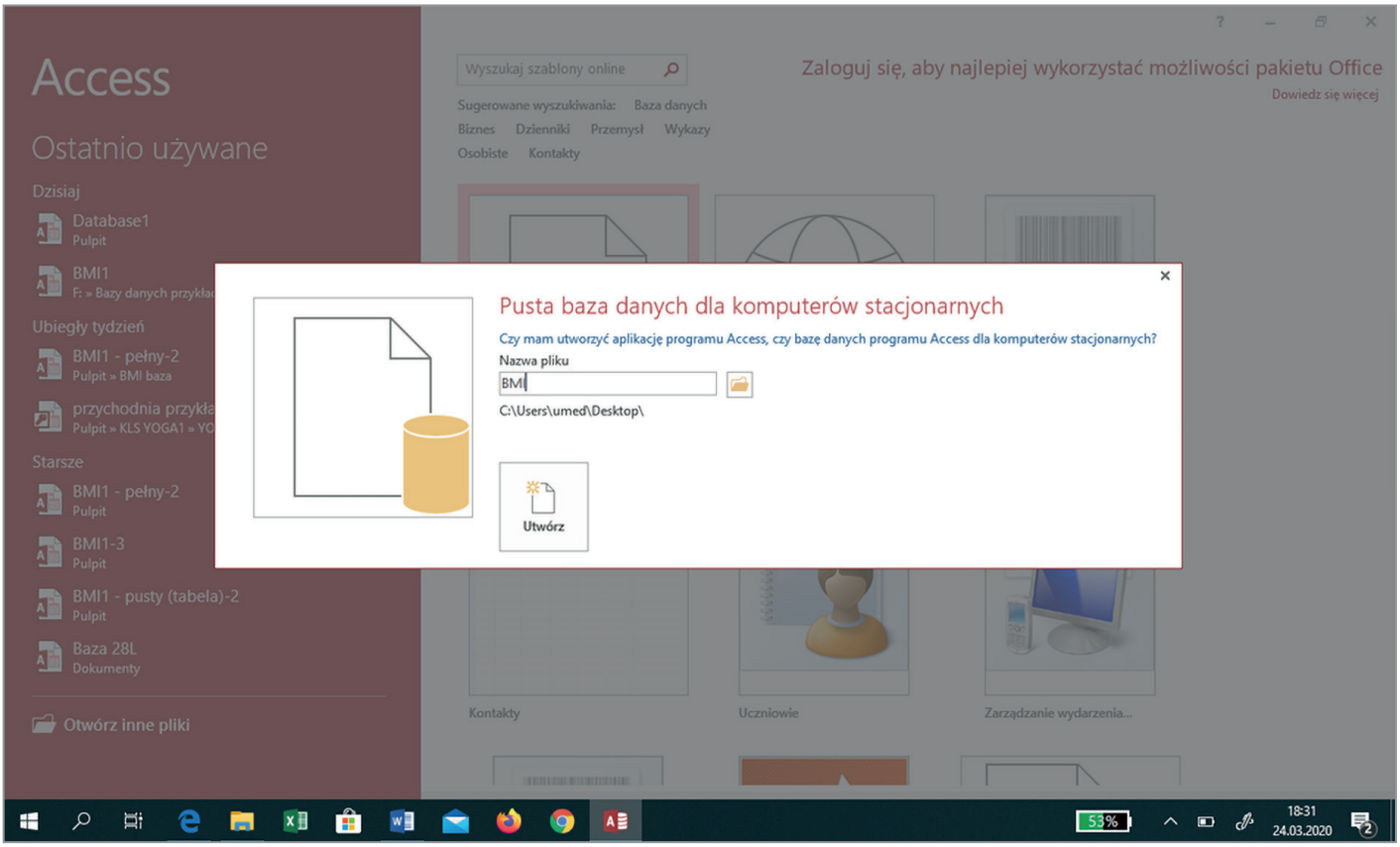
Proszę utworzyć kartotekę zawierającą dane o pacjentach poradni/gabinetu: imię, nazwisko, masę ciała i wzrost. W Internecie proszę poszukać informacji nt. współczynnika BMI i wykorzystać je do obliczenia tego wskaźnika dla każdej z osób. Następnie proszę odfiltrować dane wg zadanych kryteriów i przeprowadzić operację sortowania. Na koniec należy wykorzystać kreatory dla utworzenia formularza i raportu.

7.1.1. Rozwiązanie

Uruchamiamy aplikację do obsługi baz danych Microsoft Access (➥ **ryc. 7.1**).

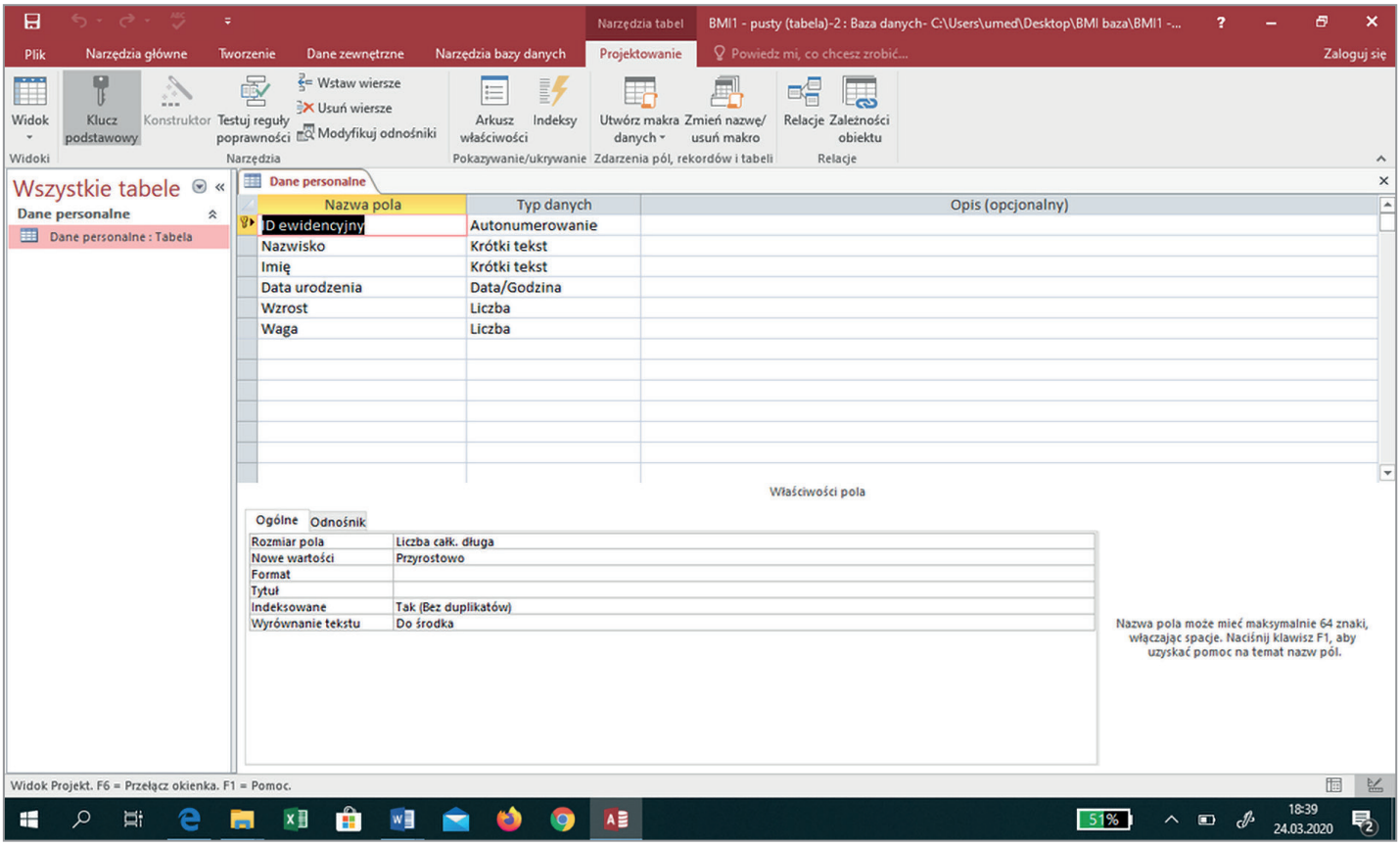
Tworzymy nową bazę danych, używając pustej bazy danych. Zapisujemy ją w pliku „BMI" na pulpicie, klikając przycisk Utwórz.

Ryc. 7.1. Tworzenie pustej bazy



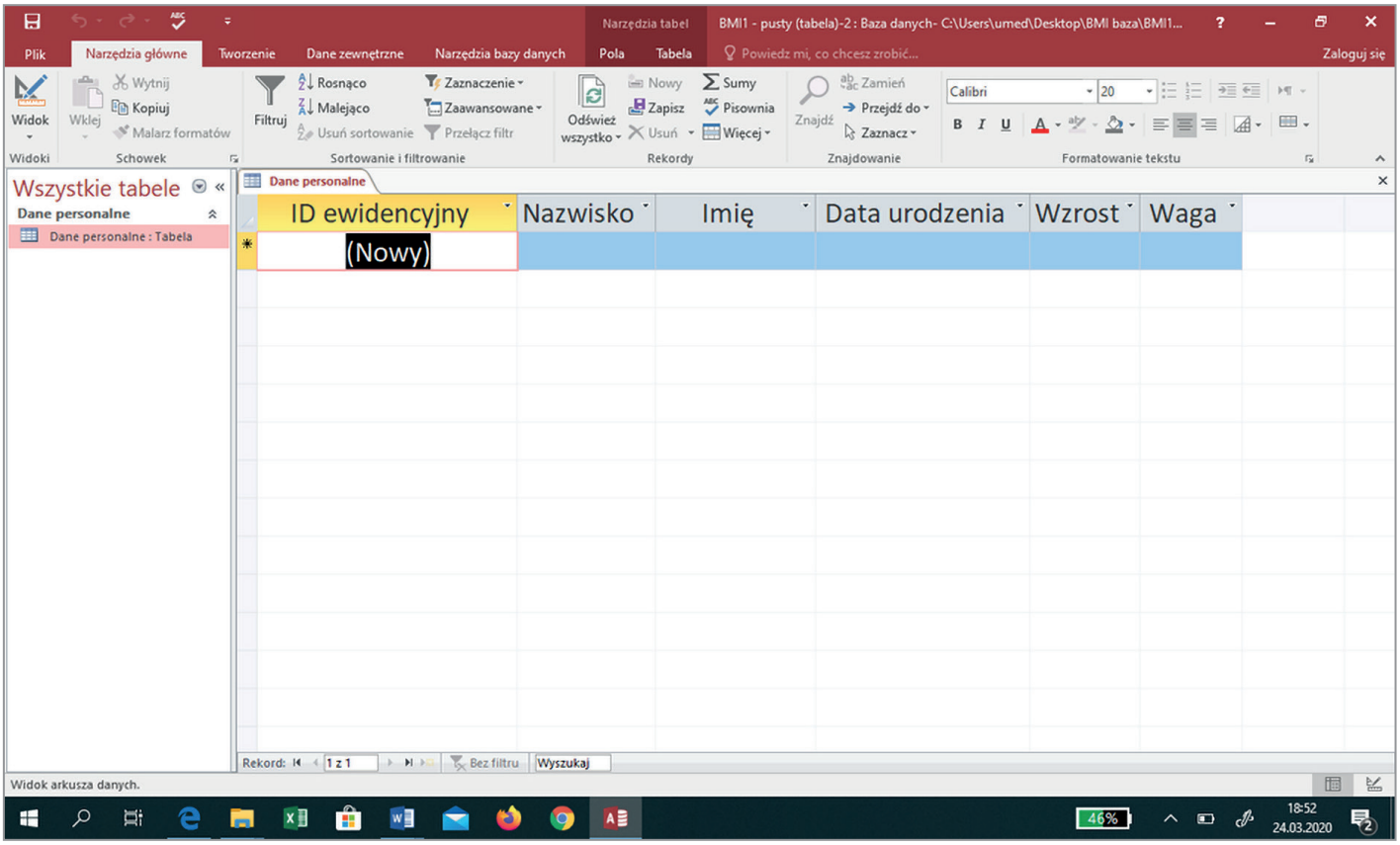
Wybieramy w otwartym oknie bazy kartę Tworzenie, następnie polecenie Projekt tabeli. Wpisujemy do niej dane jak na ➥ **ryc. 7.2**. Tabelę zamykamy. W okienku dialogowym Czy zapisać zmiany? klikamy TAK. Nadajemy nazwę „Dane personalne" i klikamy TAK, jeżeli pojawi się pytanie o ustawienie klucza podstawowego.

Ryc. 7.2. Projekt tabeli „Dane personalne"



Klikamy 2 razy tabelę „Dane personalne" w lewym oknie (różowy pasek) i otrzymujemy widok jak na ➥ **ryc. 7.3**.

Ryc. 7.3. Tabela „Dane personalne"

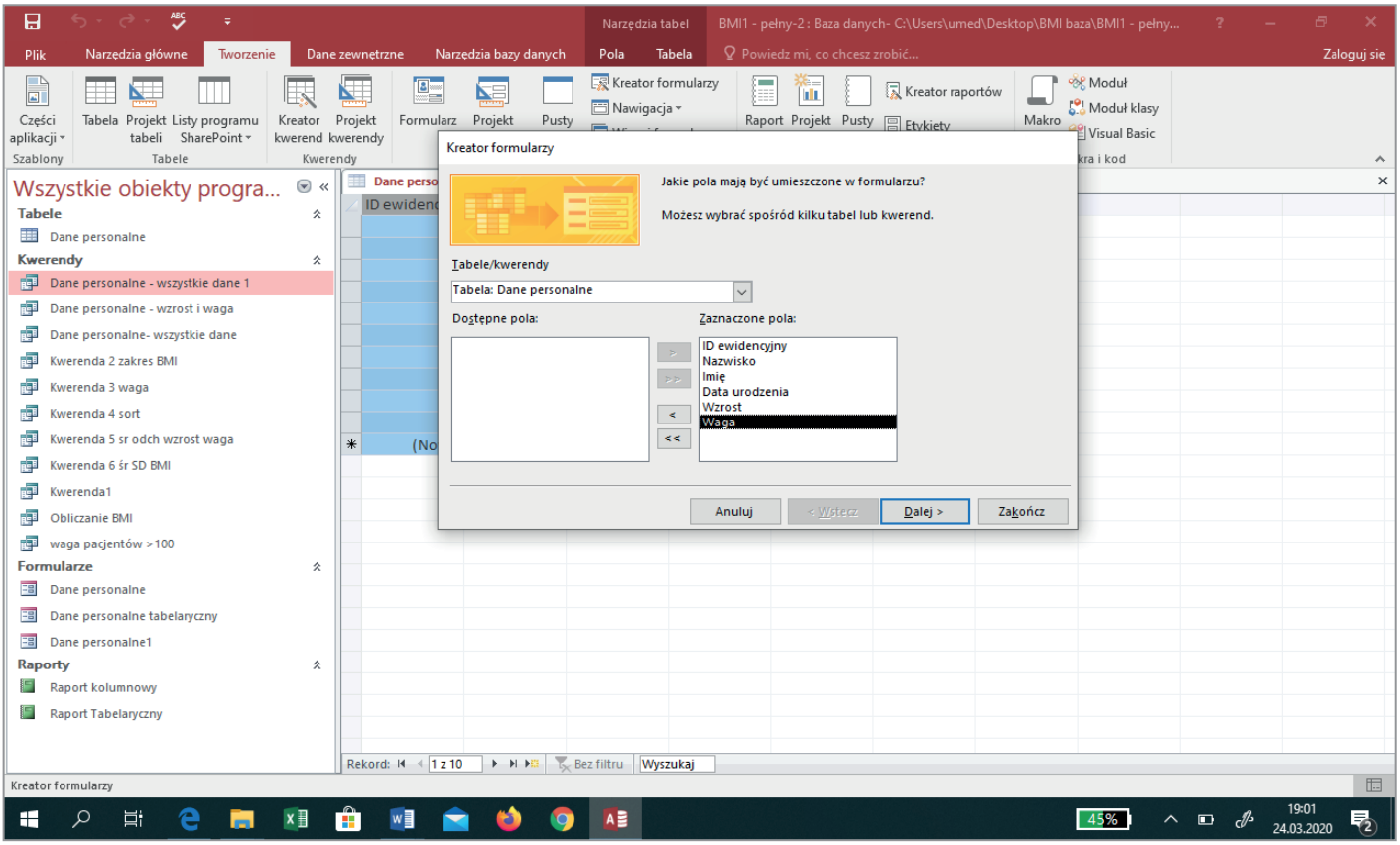


Numer ewidencyjny program wpisuje sam, natomiast resztę danych uzupełniamy wg wzoru z ➥ **tabeli 7.1**.

Dane personalne					
ID ewidencyjny	**Nazwisko**	**Imię**	**Data urodzenia**	**Wzrost**	**Waga**
1	Kowalski	Adam	15.02.1990	1,66	56
2	Nowak	Katarzyna	02.07.1991	1,72	59
3	Falek	Wojciech	09.04.1989	1,85	65
4	Wołoch	Stanisław	17.11.1990	1,71	69
5	Lorenc	Agata	01.12.1989	1,68	50
6	Derkacz	Monika	19.03.1991	1,77	81
7	Fonia	Martyna	23.10.1990	1,87	120
8	Kater	Marcin	16.11.1989	1,74	92
9	Baden	Lech	24.06.1991	1,86	111
10	Jaworska	Dominika	09.01.1990	1,85	80

Tabela 7.1. Dane pacjentów

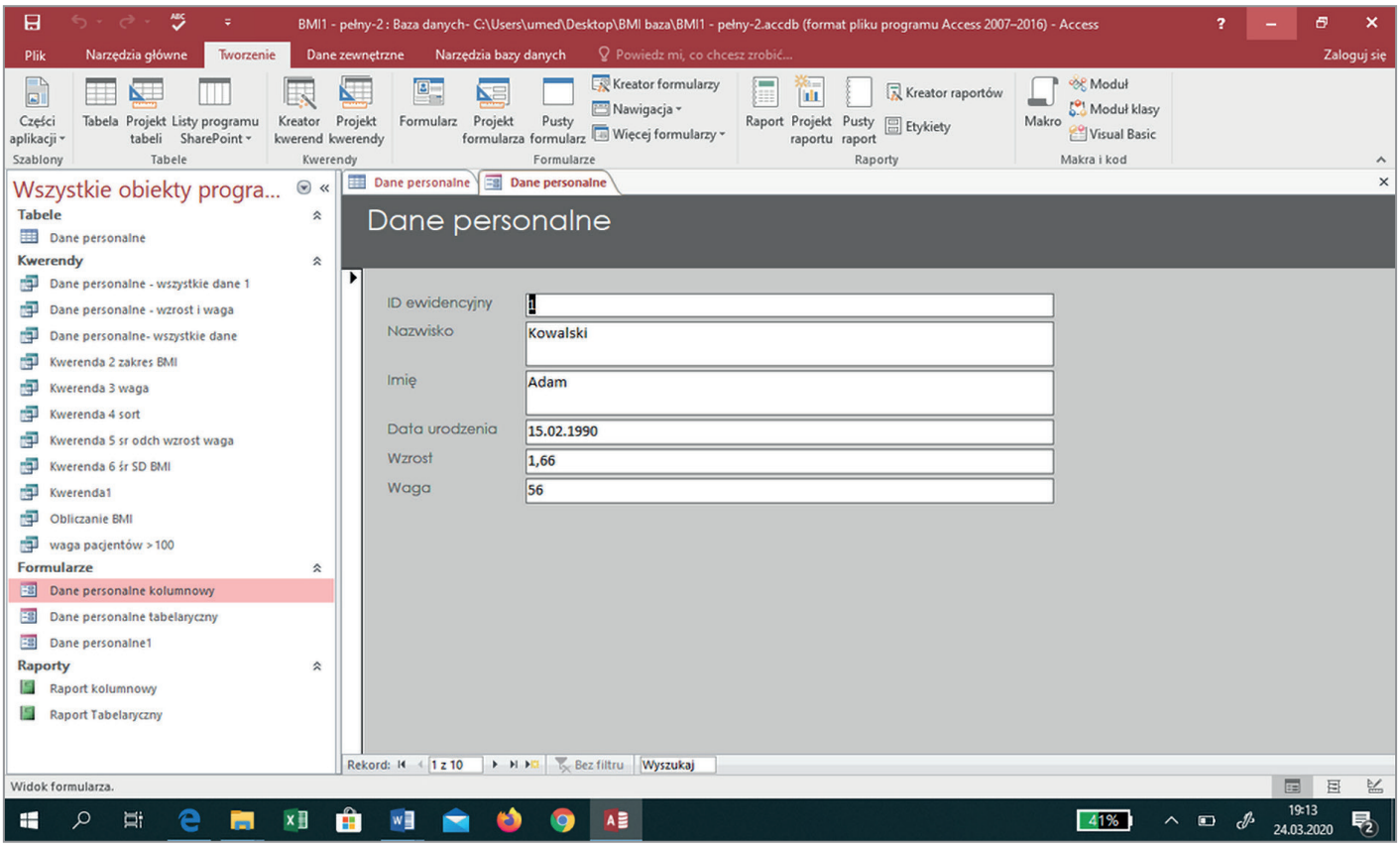
Następnie wybieramy kartę Tworzenie i po kliknięciu polecenia Kreator formularzy pojawia się okno jak na ➥ **ryc. 7.4**. Klikamy dwie strzałki >>, przenosząc w ten sposób dane do drugiego okna. Następnie klikamy przycisk Dalej (➥ **ryc. 7.4**).



Ryc. 7.4. Tworzenie raportu kolumnowego

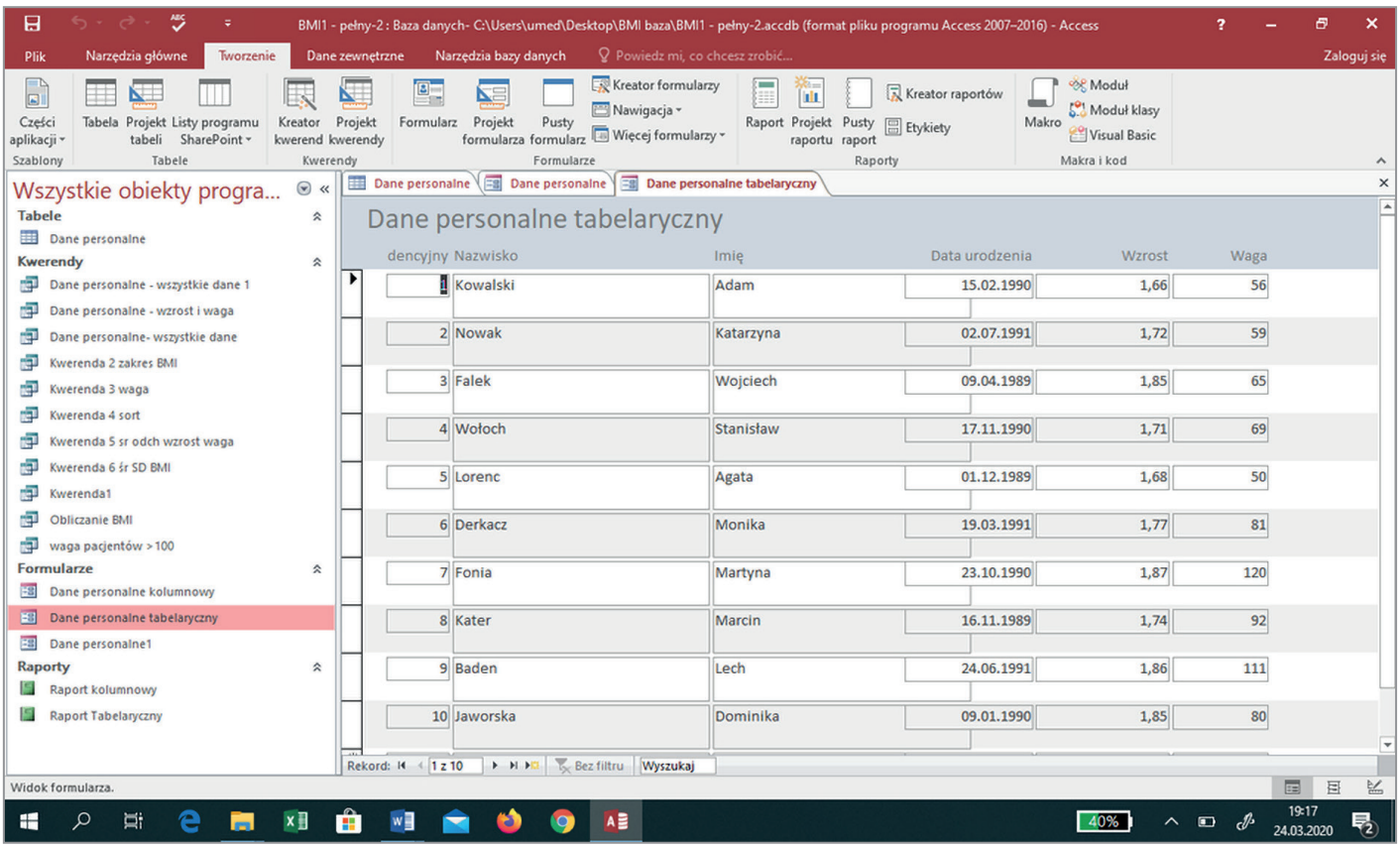
W kolejnym kroku zaznaczamy Kolumnowy. Zapisujemy pod nazwą Dane personalne kolumnowy. Strzałki na dole okna pozwalają przeglądać dane kolejnych pacjentów (➥ **ryc. 7.5**).

Ryc. 7.5. Widok raportu kolumnowego



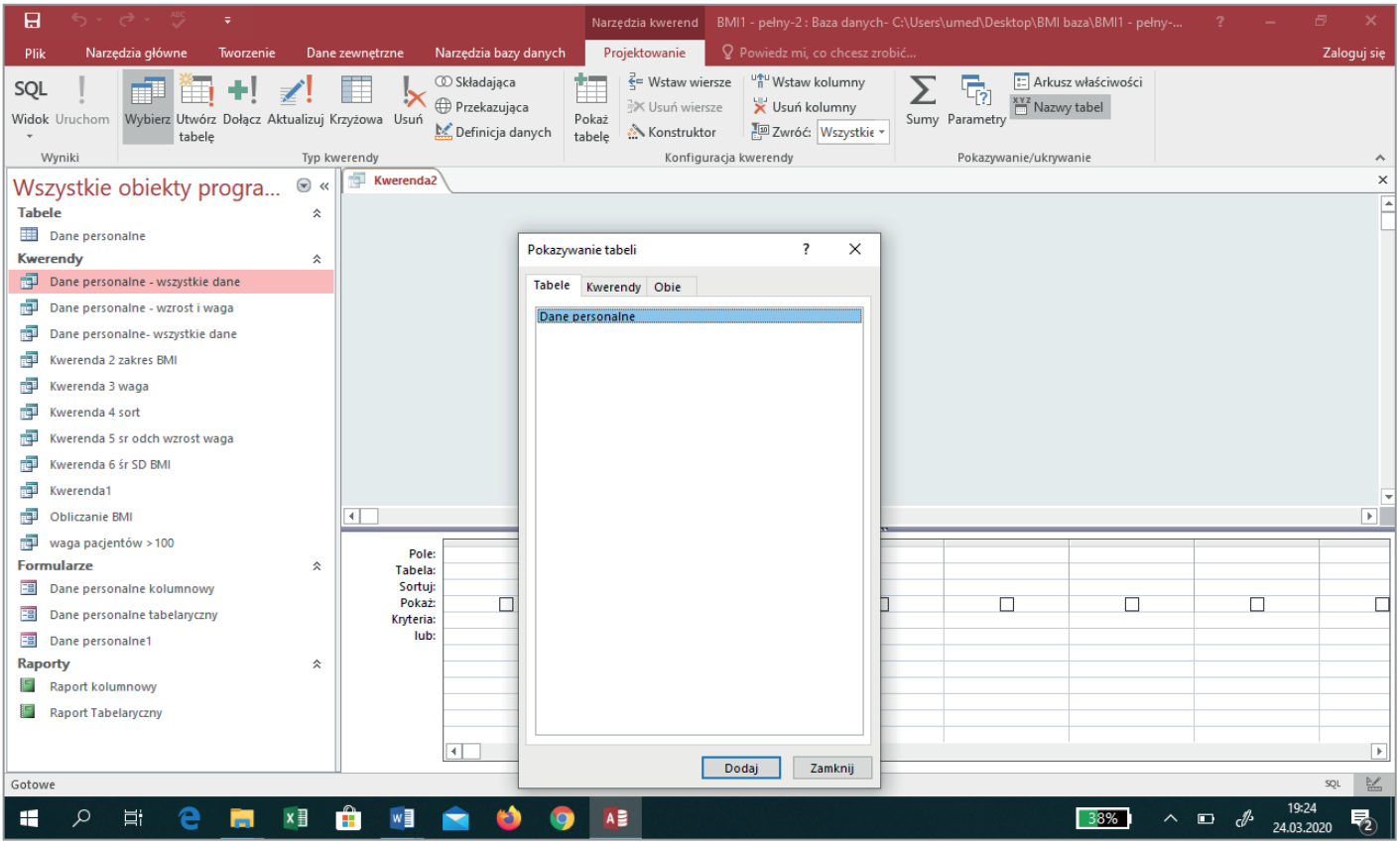
Następnie tworzymy jeszcze jeden formularz, zaznaczając tym razem Tabelaryczny i zapisując pod nazwą „Dane personalne tabelaryczny (➥ **ryc. 7.6**).

Ryc. 7.6. Widok raportu tabelarycznego



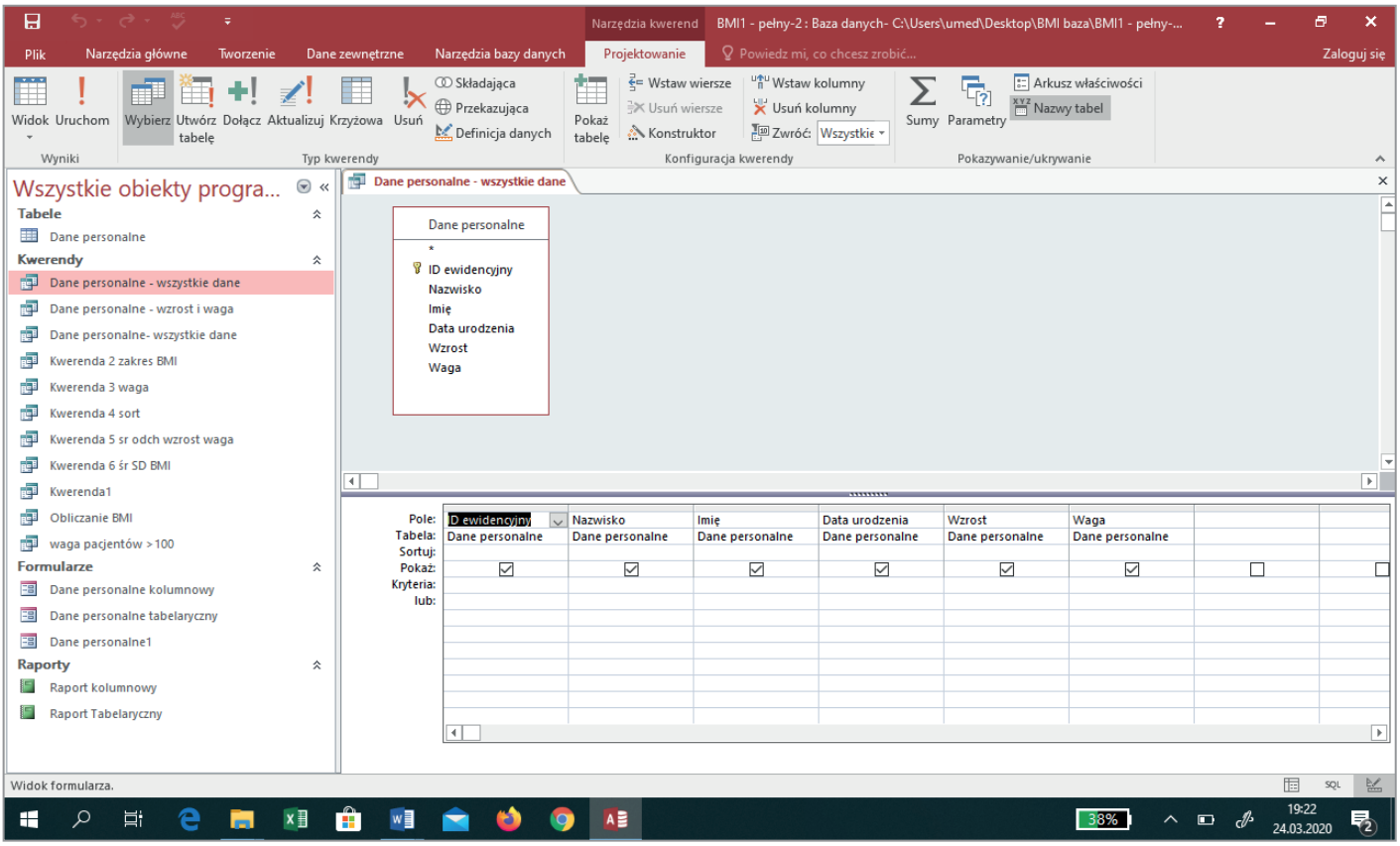
W dalszej części ćwiczenia na karcie Tworzenie klikamy polcenie Projekt kwerendy (➥ **ryc. 7.7**).

Zaznaczamy Dane personalne i na dole klikamy Dodaj.



Ryc. 7.7. Początek tworzenia kwerendy

Klikamy dwukrotnie kolejno na polach: ID ewidencyjny, nazwisko, imię, data urodzenia, wzrost, waga (pojawią się one w dolnych rubrykach, ➥ **ryc. 7.8**).

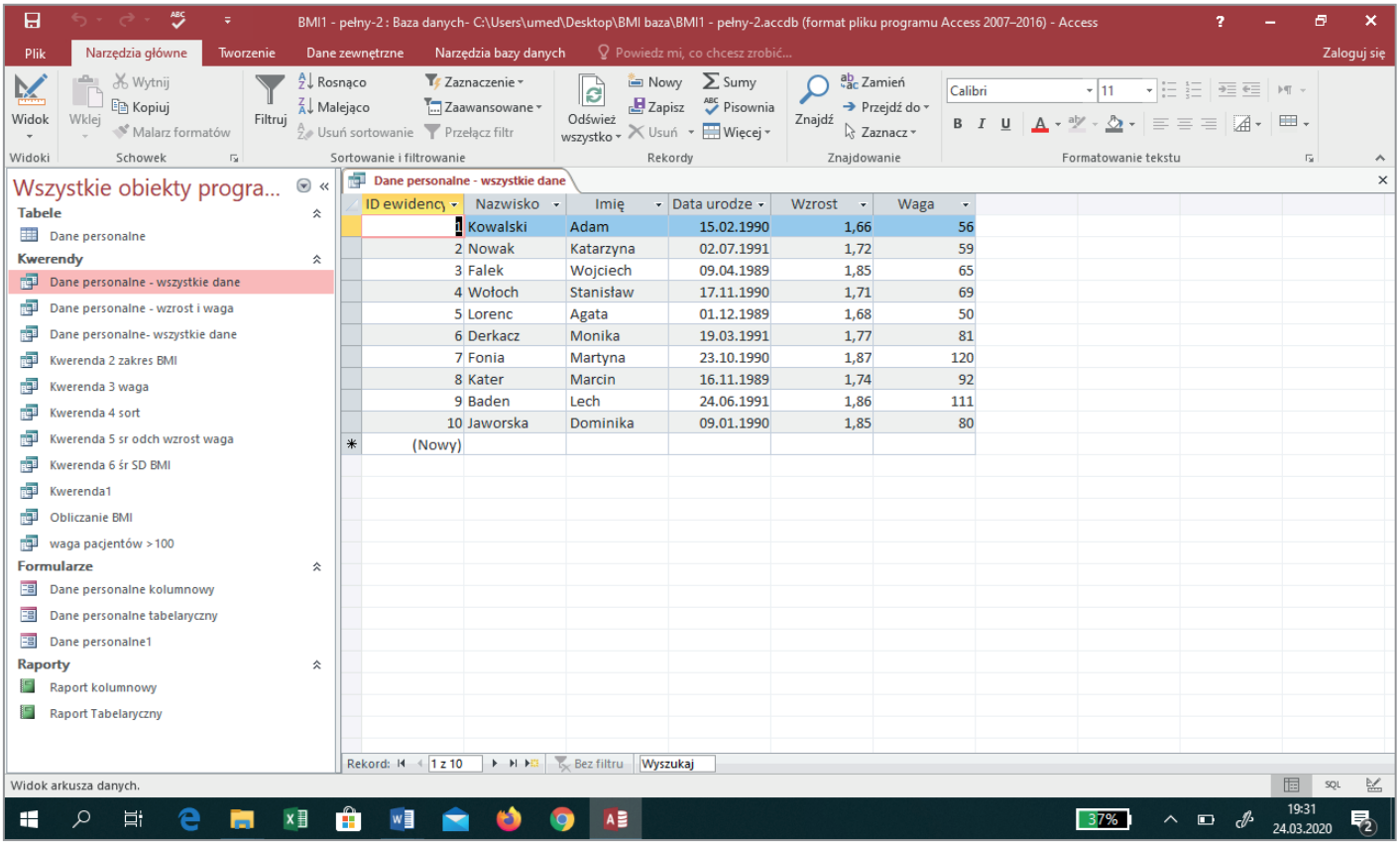


Ryc. 7.8. Projekt kwerendy do uruchomienia

Na wstążce klikamy polecenie Uruchom z czerwonym wykrzyknikiem, w wyniku czego otrzymujemy kwerendę (➥ **ryc. 7.9**).

Zapisujemy kwerendę pod nazwą „Dane personalne – wszystkie dane".

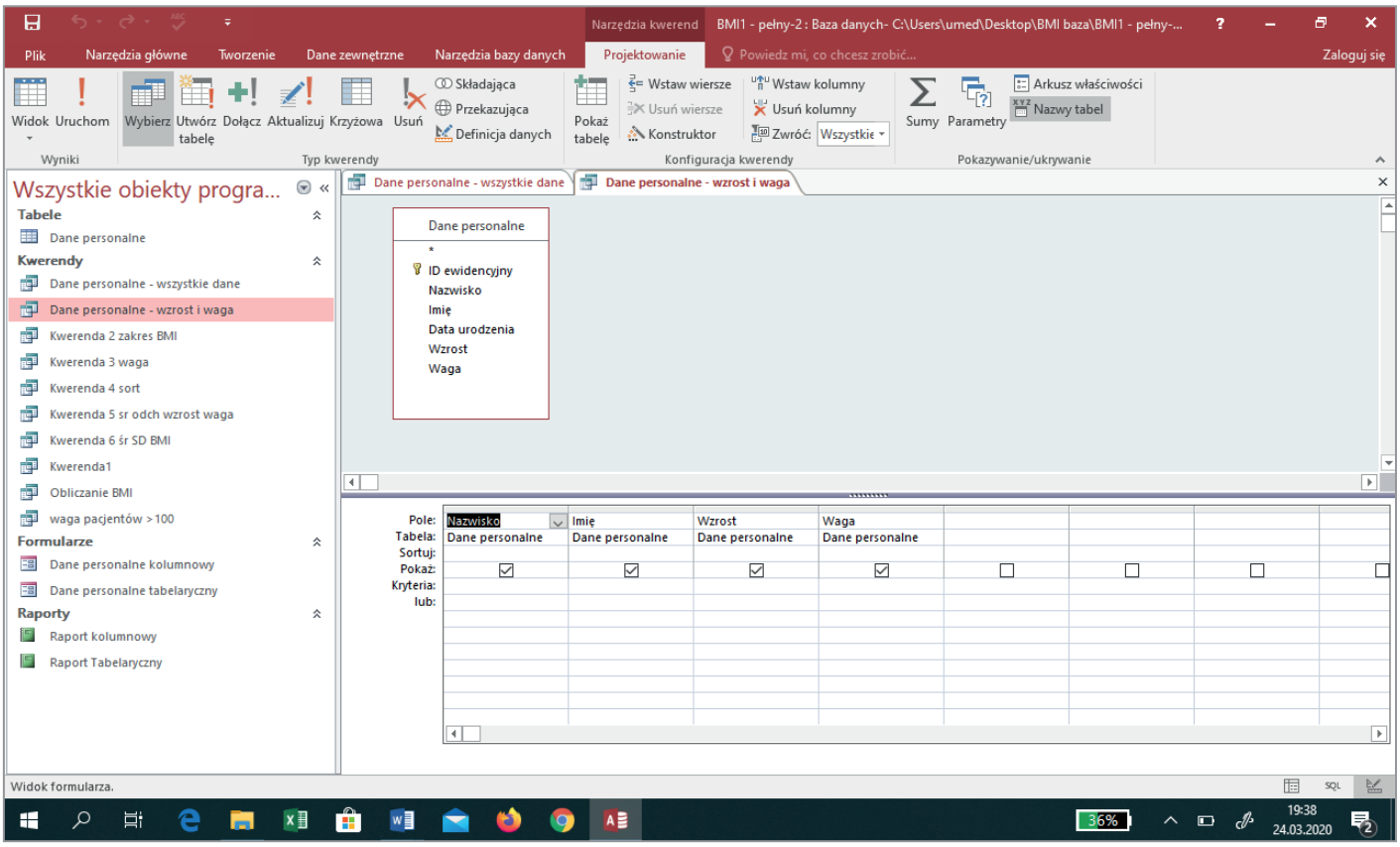
Ryc. 7.9. Wynik kwerendy z danymi personalnymi



Według powyższego schematu projektujemy kolejną kwerendę (➥ **ryc. 7.10**).

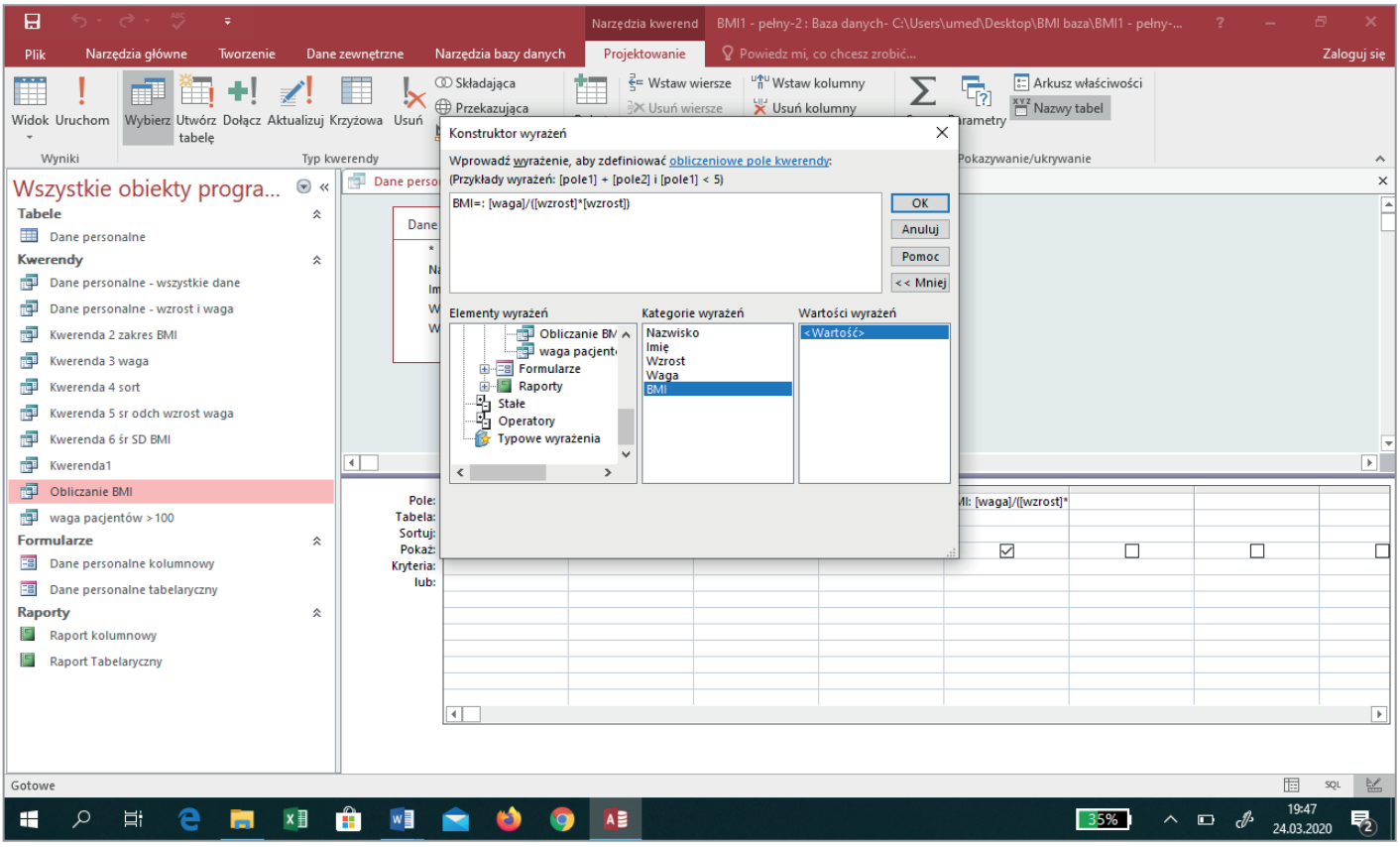
Uruchamiamy i zapisujemy jako „Dane personalne – wzrost i waga".

Ryc. 7.10. Projekt kwerendy do obliczania BMI pacjentów



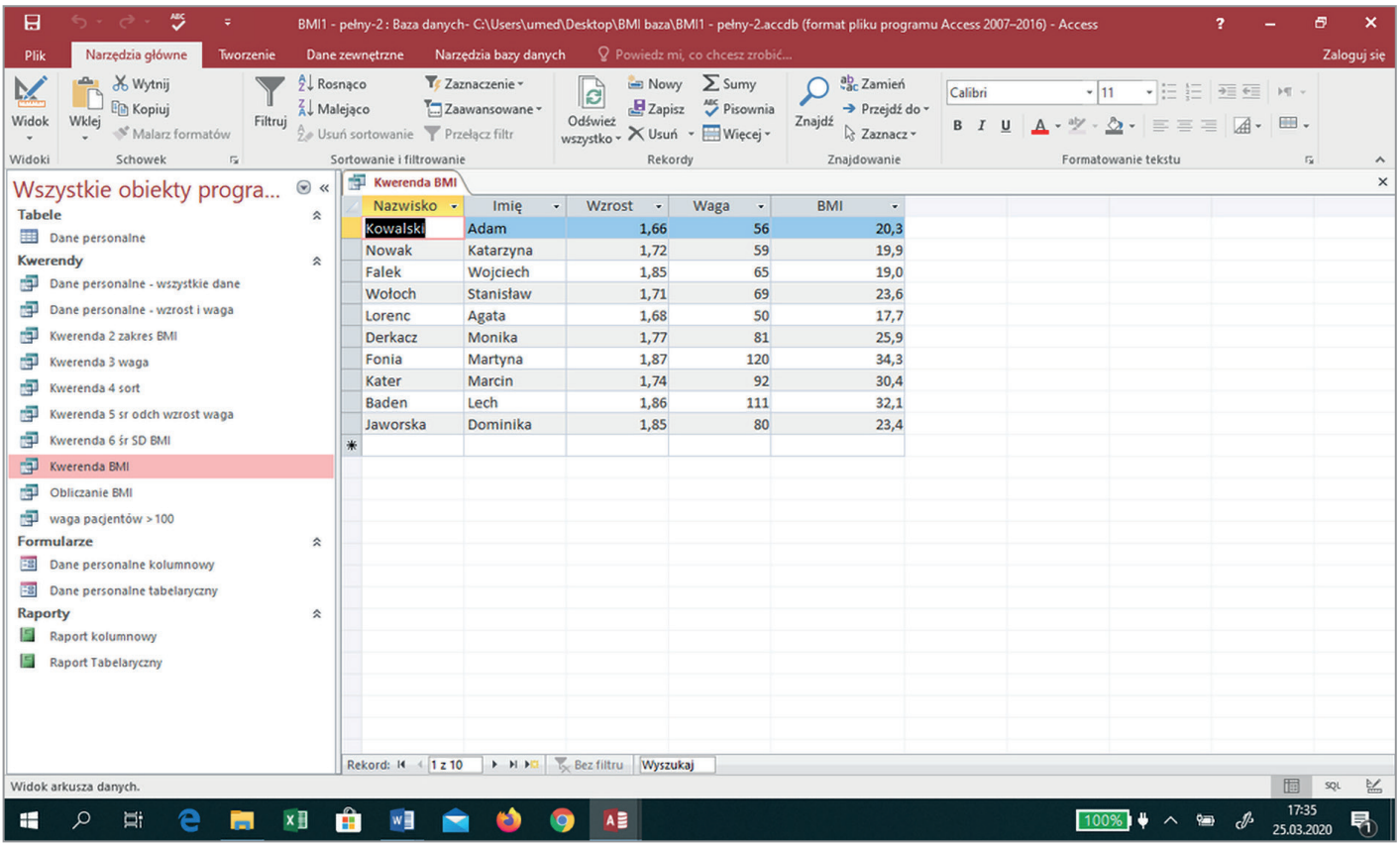
Jeszcze raz tworzymy projekt kwerendy jak wyżej, ale nie uruchamiamy. Umieszczamy kursor w komórce 5. pustej kolumny i ze wstążki wybieramy polecenie Konstruktor wyrażeń. Pojawia się okno jak na ➥ **ryc. 7.11**.

Ryc. 7.11. Obliczanie BMI za pomocą Konstruktora wyrażeń



W oknie Konstruktora wyrażeń wpisujemy wzór na obliczanie współczynnika BMI. Uruchamiamy kwerendę. Wynik przedstawiono na ➥ **ryc. 7.12**.

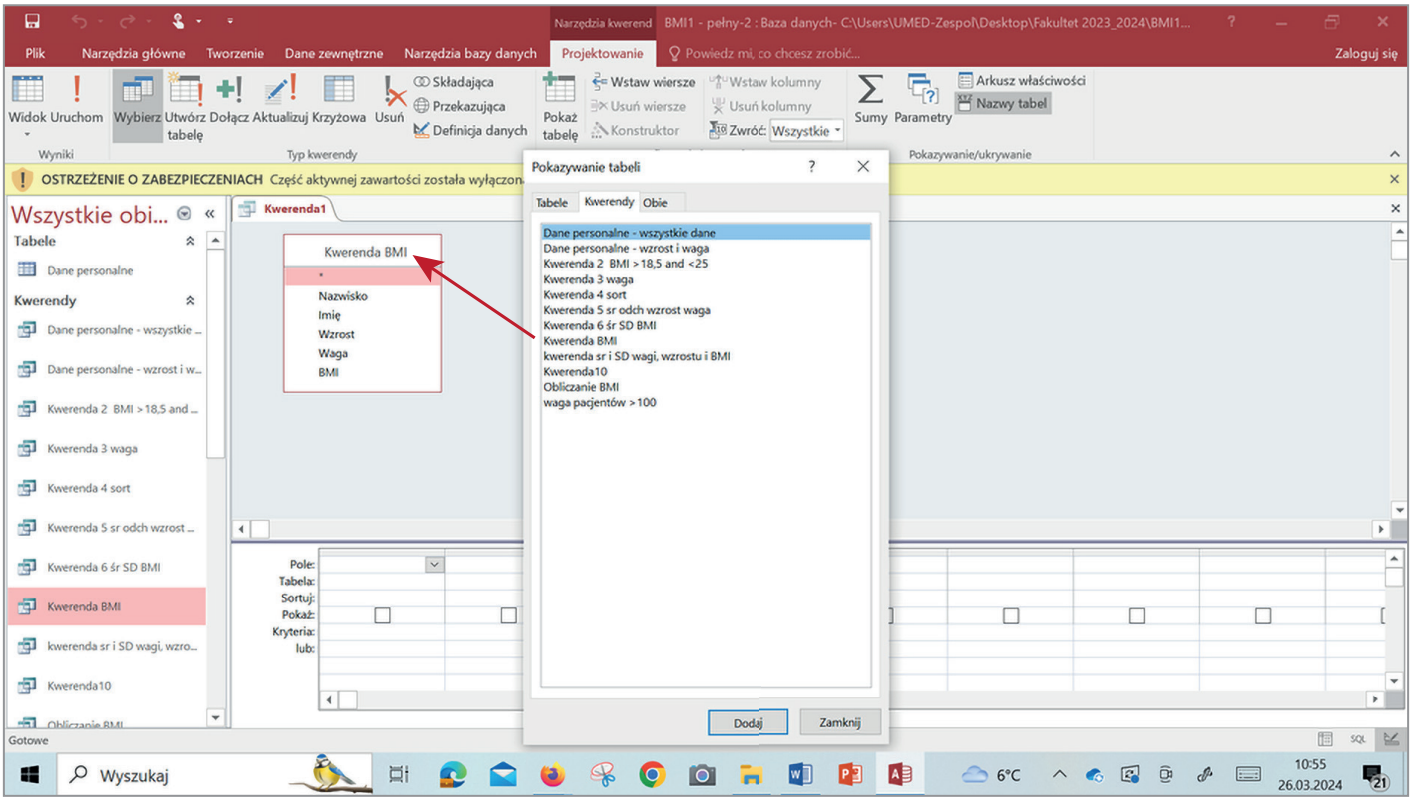
Ryc. 7.12. Wyniki obliczania BMI



Utworzymy teraz kwerendę, gdzie pokażą się tylko dane pacjentów z BMI > 18,5 i BMI < 25.

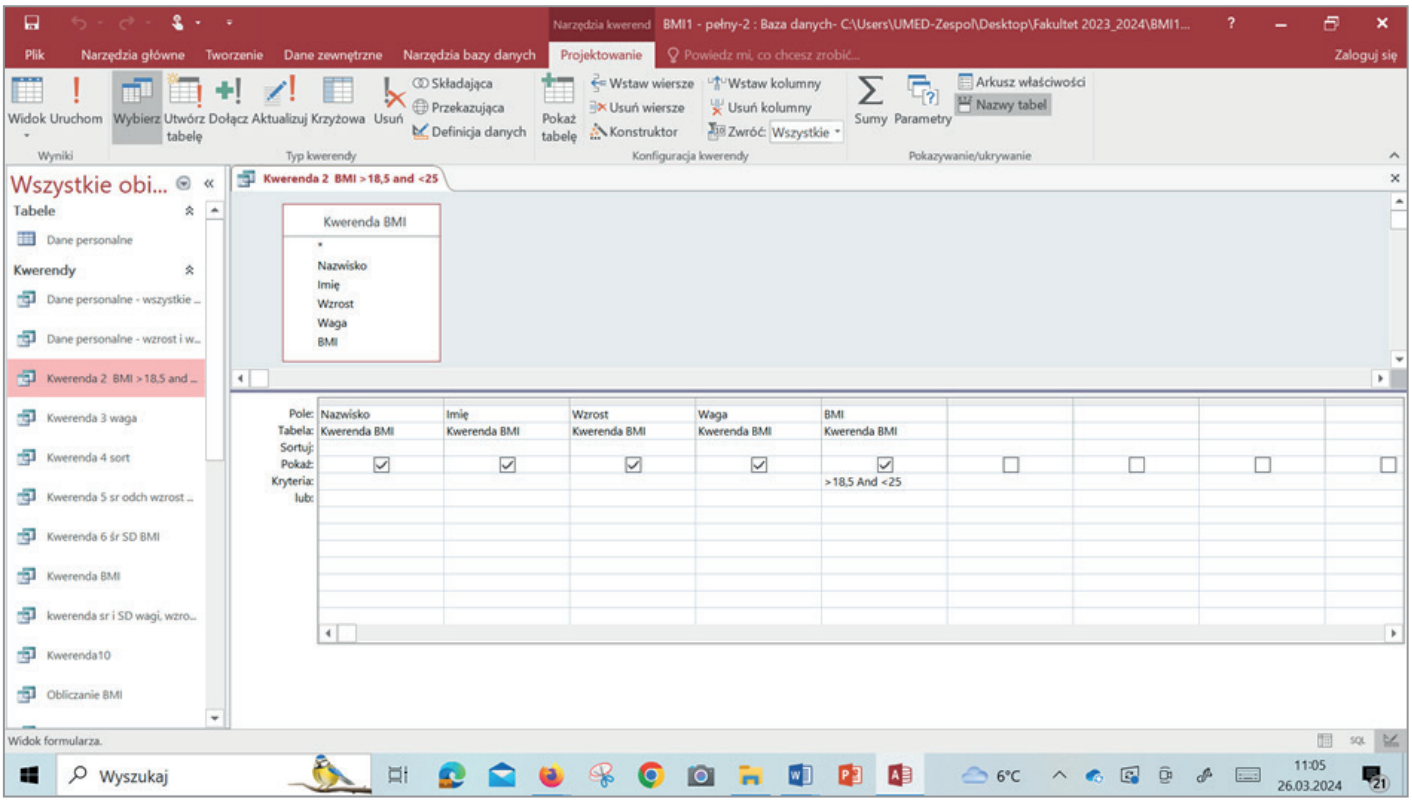
Wybieramy Tworzenie, potem projekt kwerendy i dodajemy na ekran „Kwerendę BMI", jak zaprezentowano na ➥ **ryc. 7.13**. W tym momencie zakładka Tworzenie zmienia się na Projektowanie.

Ryc. 7.13. Przygotowanie kwerendy do ustawienia warunku



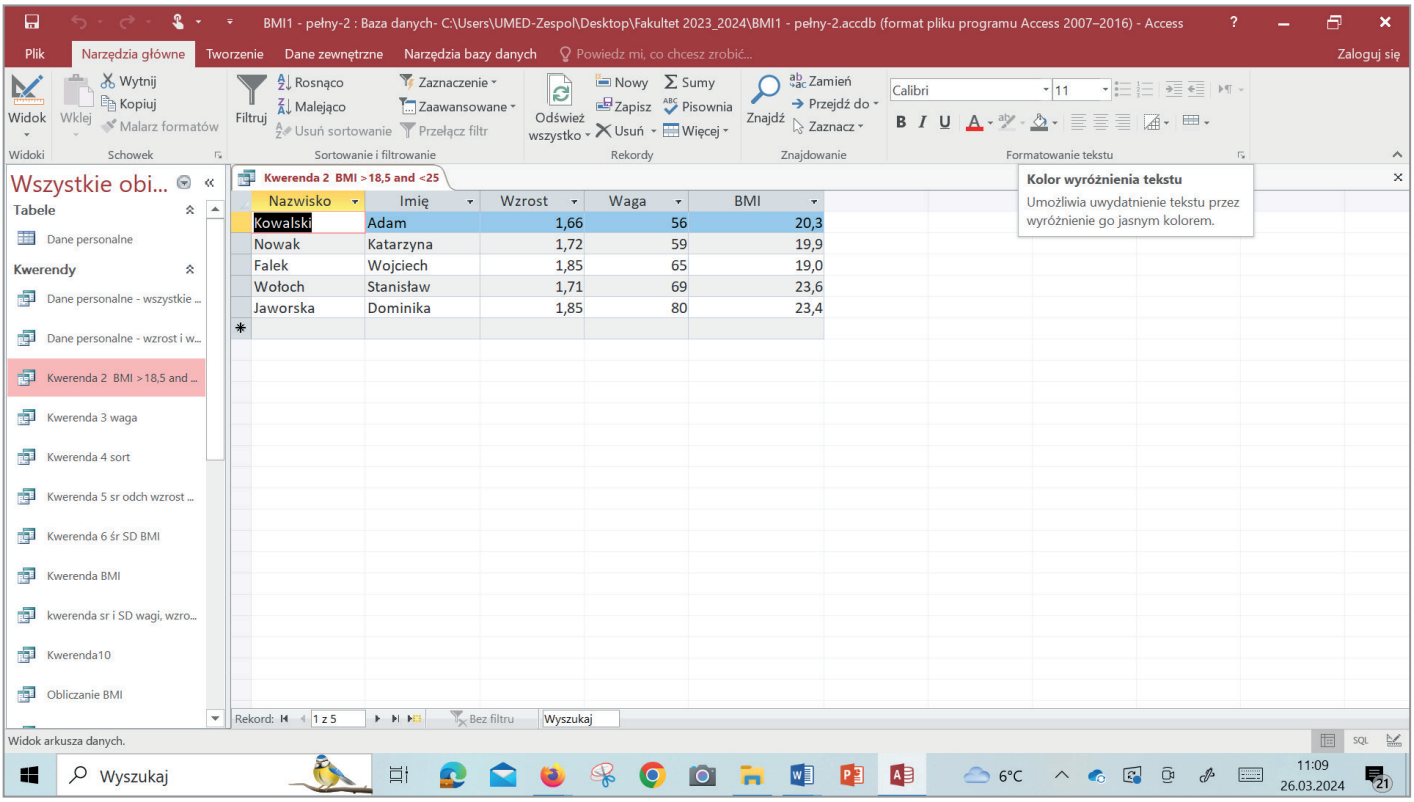
W kolumnie BMI ustawiamy kryterium > 18,5 i < 25 – BMI wynik prawidłowy (➥ **ryc. 7.14**).

Ryc. 7.14. Projekt kwerendy z kryteriami BMI: „>18,5 i < 25"



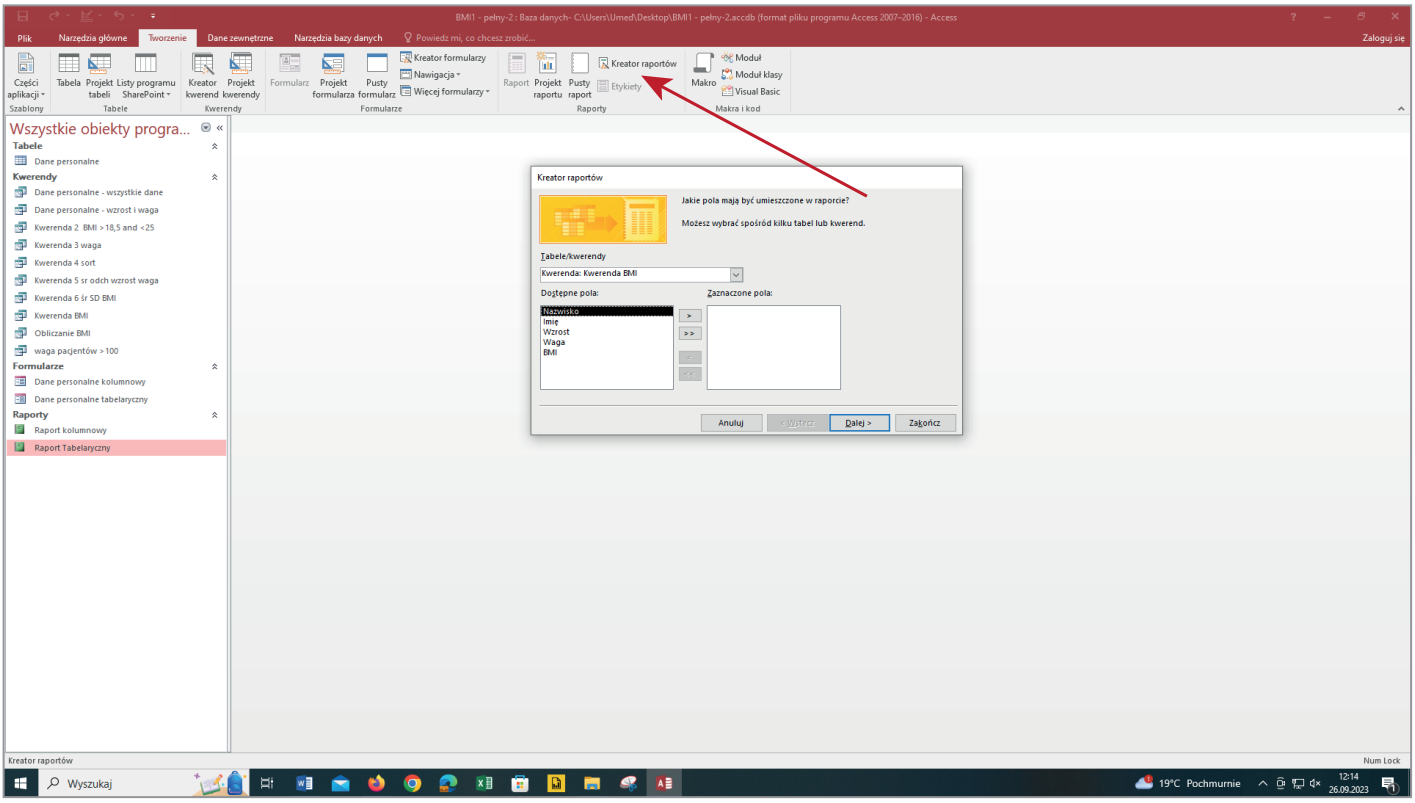
Ryc. 7.15. Wynik uruchomienia kwerendy

Wyniki oglądamy po wybraniu polecenia Uruchom (➥ **ryc. 7.15**).



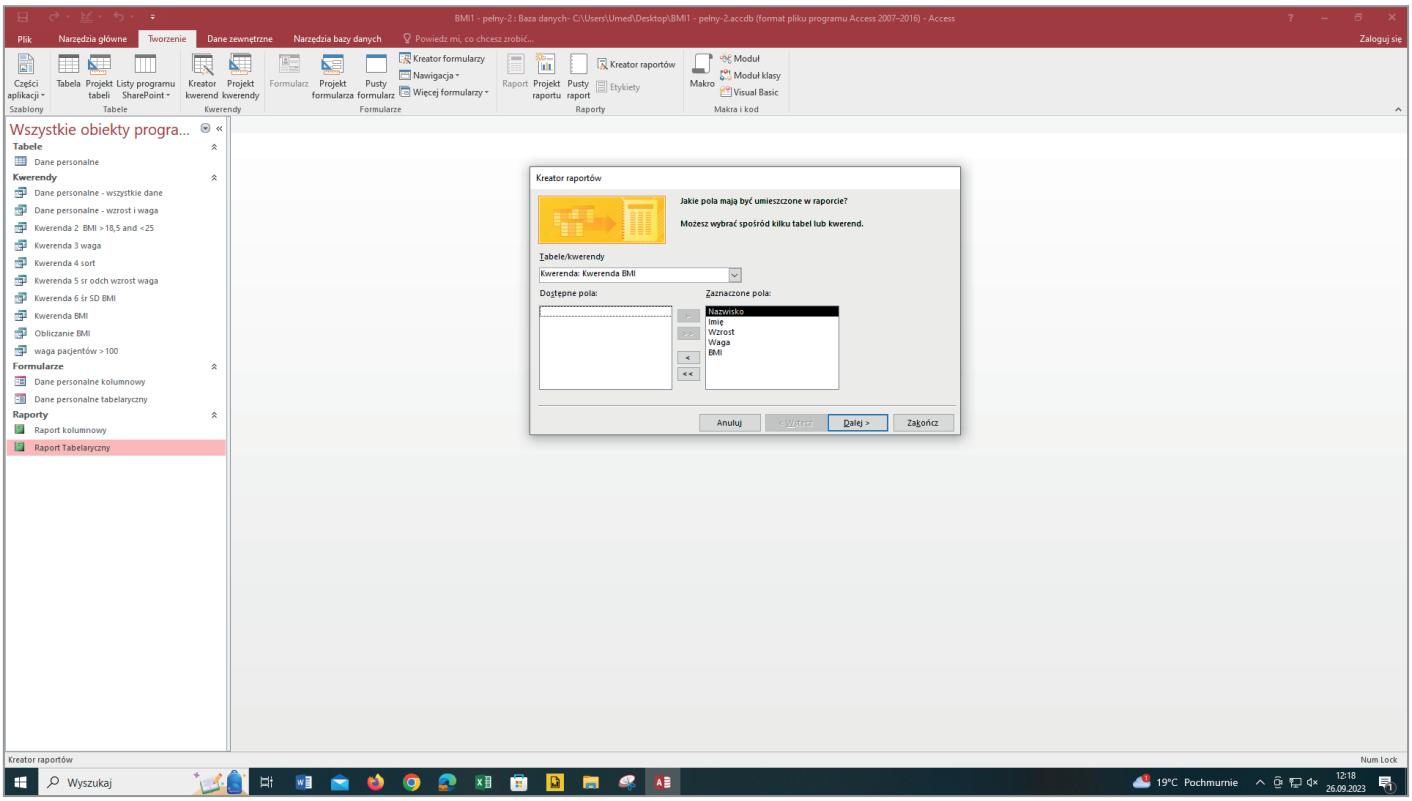
Ryc. 7.16. Tworzenie raportu z wykorzystaniem kreatora raportów

W celu utworzenia raportu korzystamy z Kreatora raportów w karcie Tworzenie. W tym przykładzie sporządzamy raport na podstawie „Kwerendy BMI" (➥ **ryc. 7.16**).



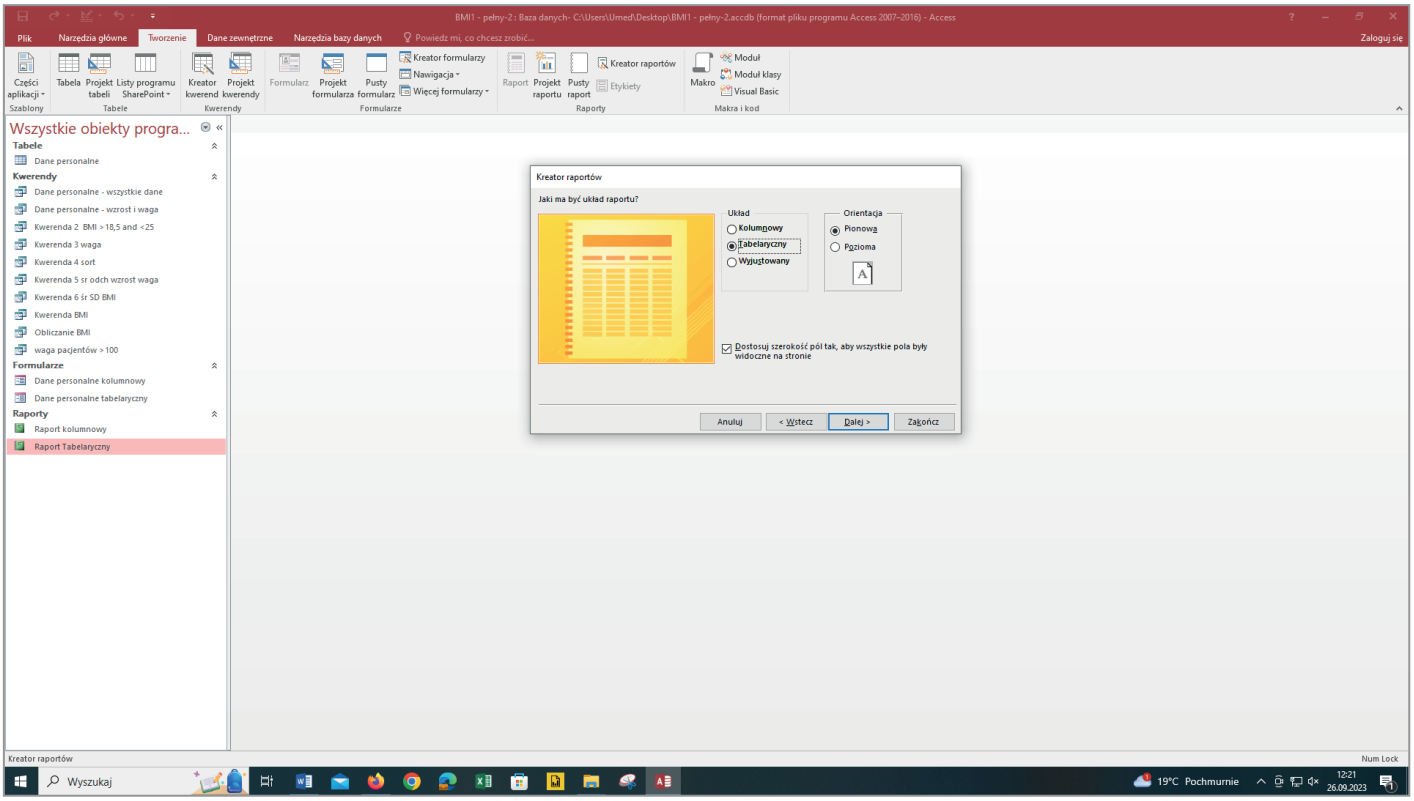
Ryc. 7.17. Przygotowanie danych do raportu

Przenosimy wszystkie dane do prawego okna. Następnie klikamy przycisk Dalej (➥ **ryc. 7.17**).



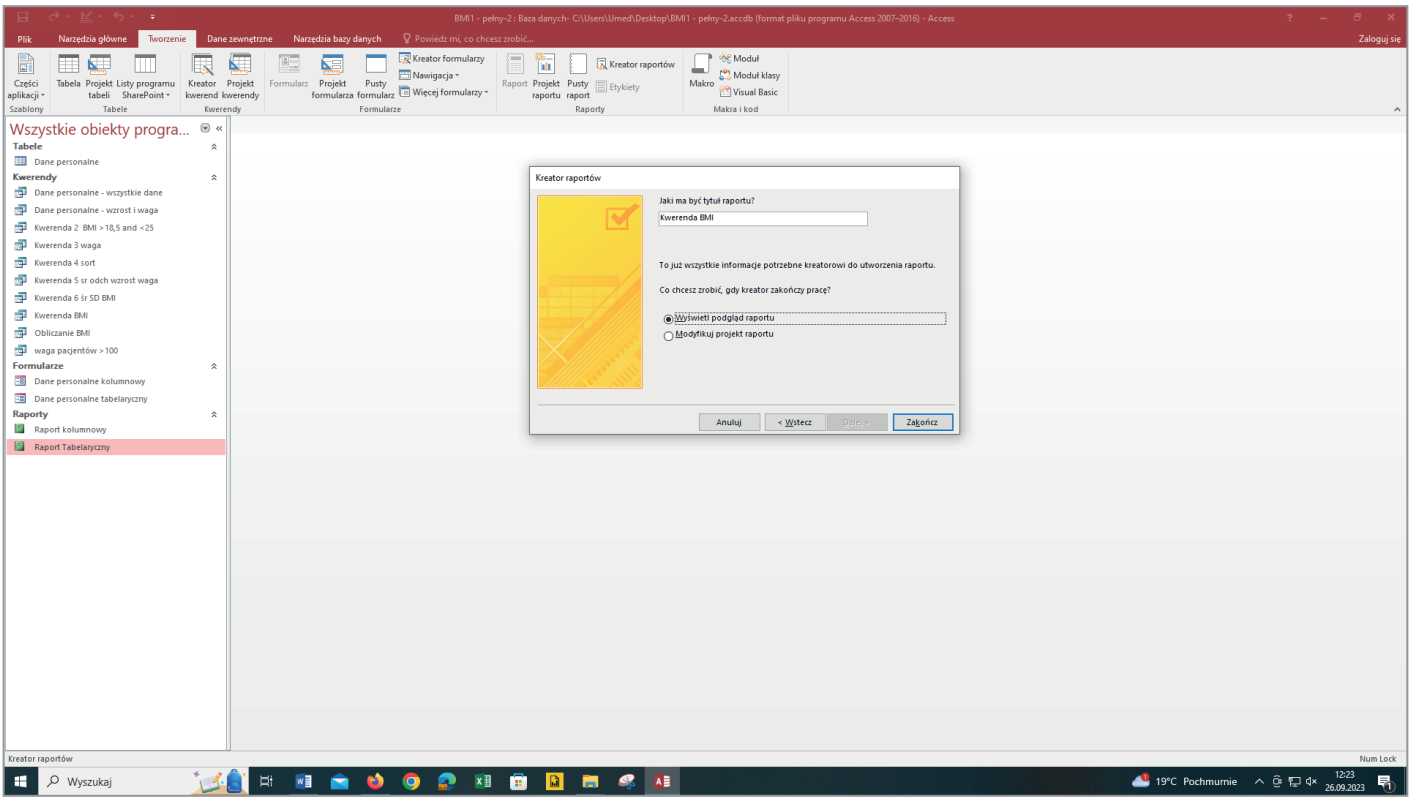
Ryc. 7.18. Wybór układu raportu

W kolejnych 2 krokach nic nie zmieniamy, tylko klikamy kolejno przycisk Dalej. Wybieramy układ Tabelaryczny i ponownie klikamy Dalej (➥ **ryc. 7.18**).



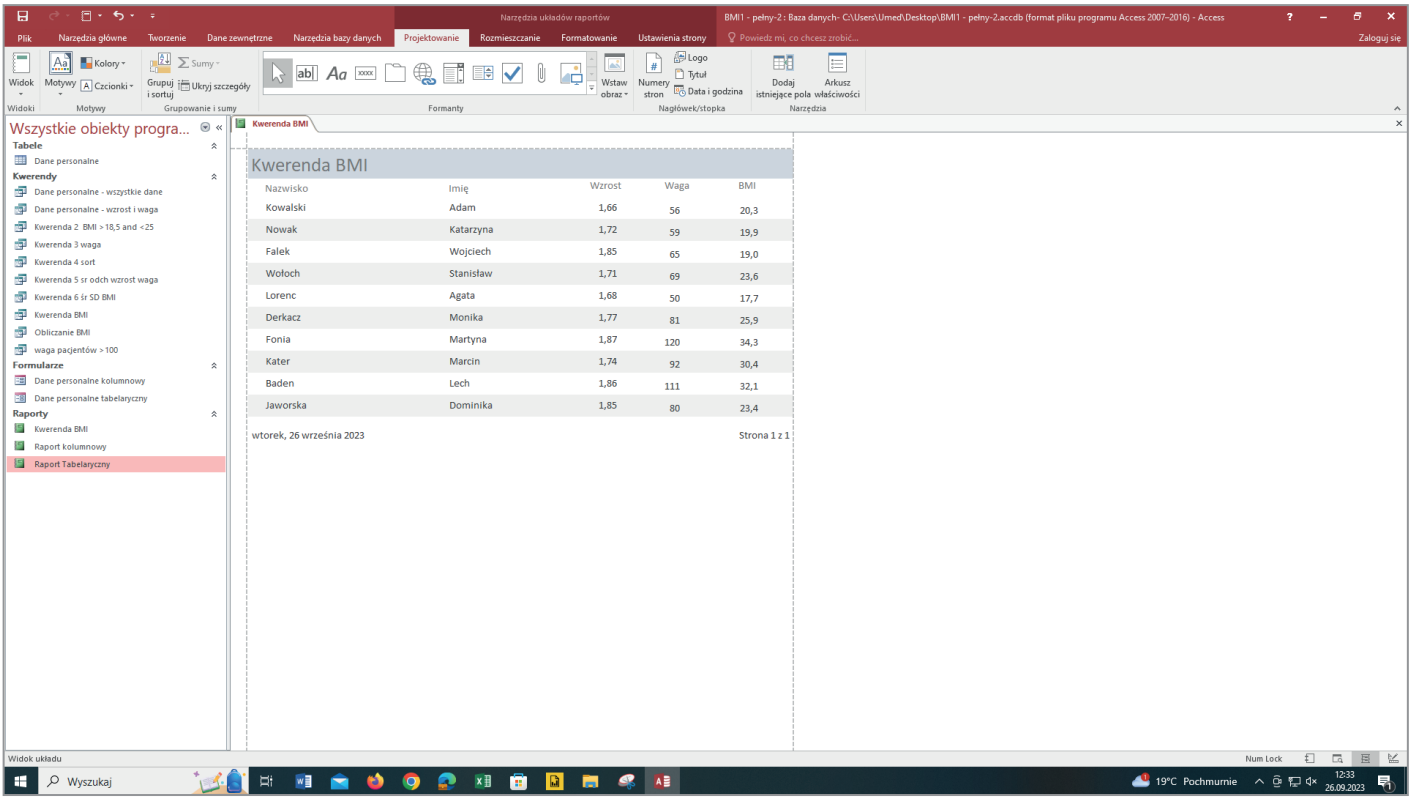
Ryc. 7.19. Zakończenie tworzenia raportu

W ostatnim kroku wpisujemy tytuł raportu („Kwerenda BMI”) i klikamy Zakończ (➥ **ryc. 7.19**).



Ryc. 7.20. Widok utworzonego raportu

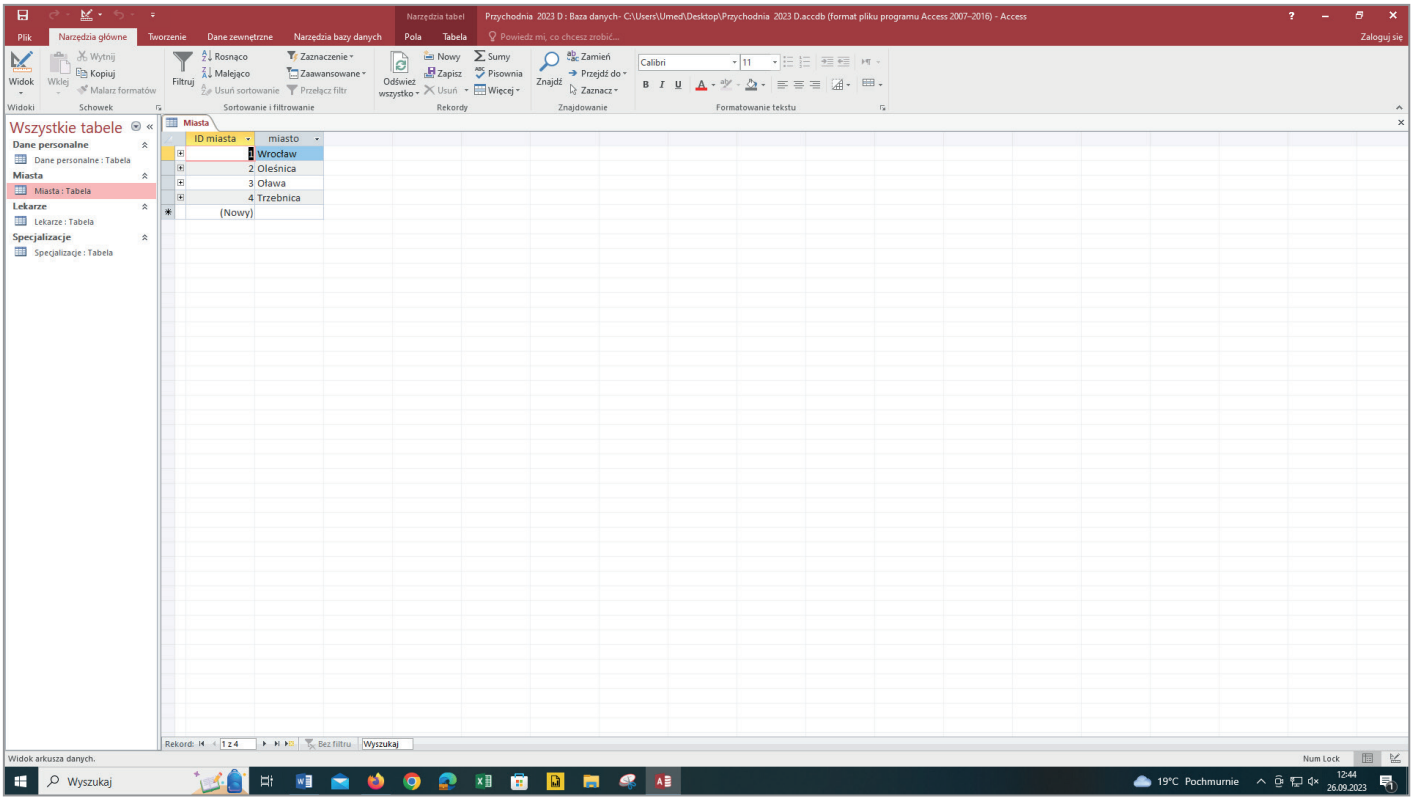
W efekcie otrzymujemy zestawienie zaprezentowane poniżej (➥ **ryc. 7.20**).



7.2. Zadanie: baza „Przychodnia”

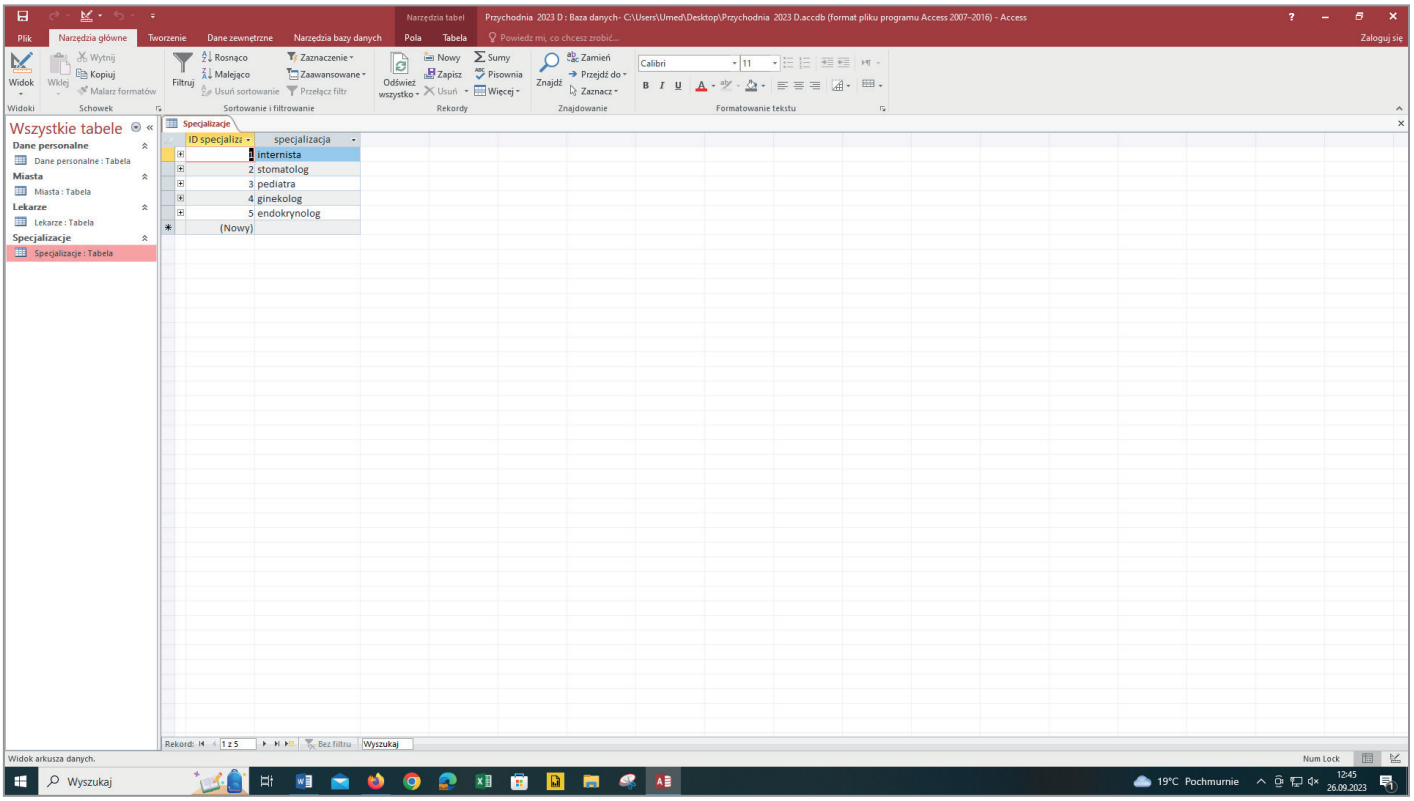
Najpierw proszę (w znany już z poprzedniego ćwiczenia sposób) utworzyć tabelę „Miasta” (➥ **ryc. 7.21**).

Ryc. 7.21. Tabela „Miasta”



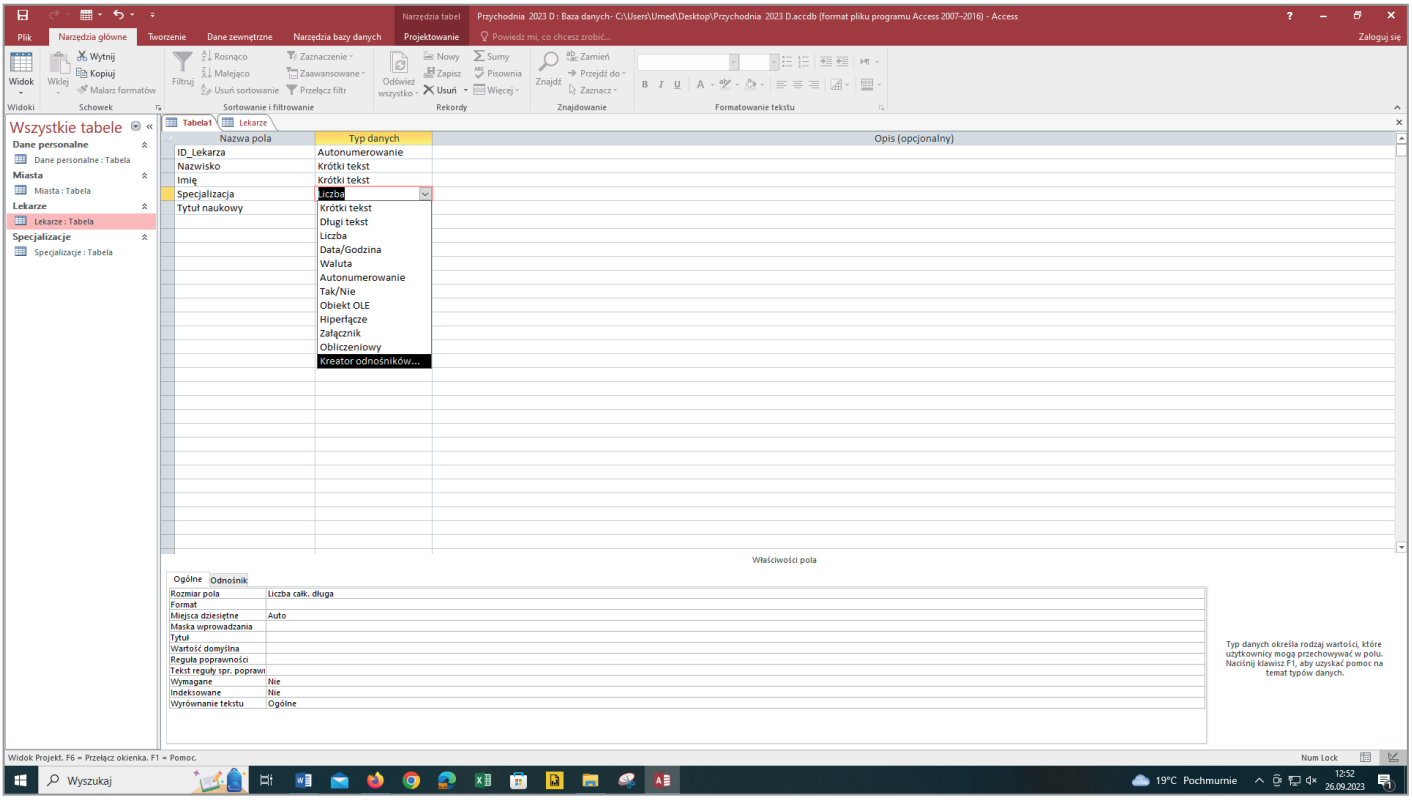
Następnie proszę utworzyć tabelę „Specjalizacje” (➥ **ryc. 7.22**). Istotna jest kolejność zakładania tabel!

Ryc. 7.22. Tabela „Specjalizacje”



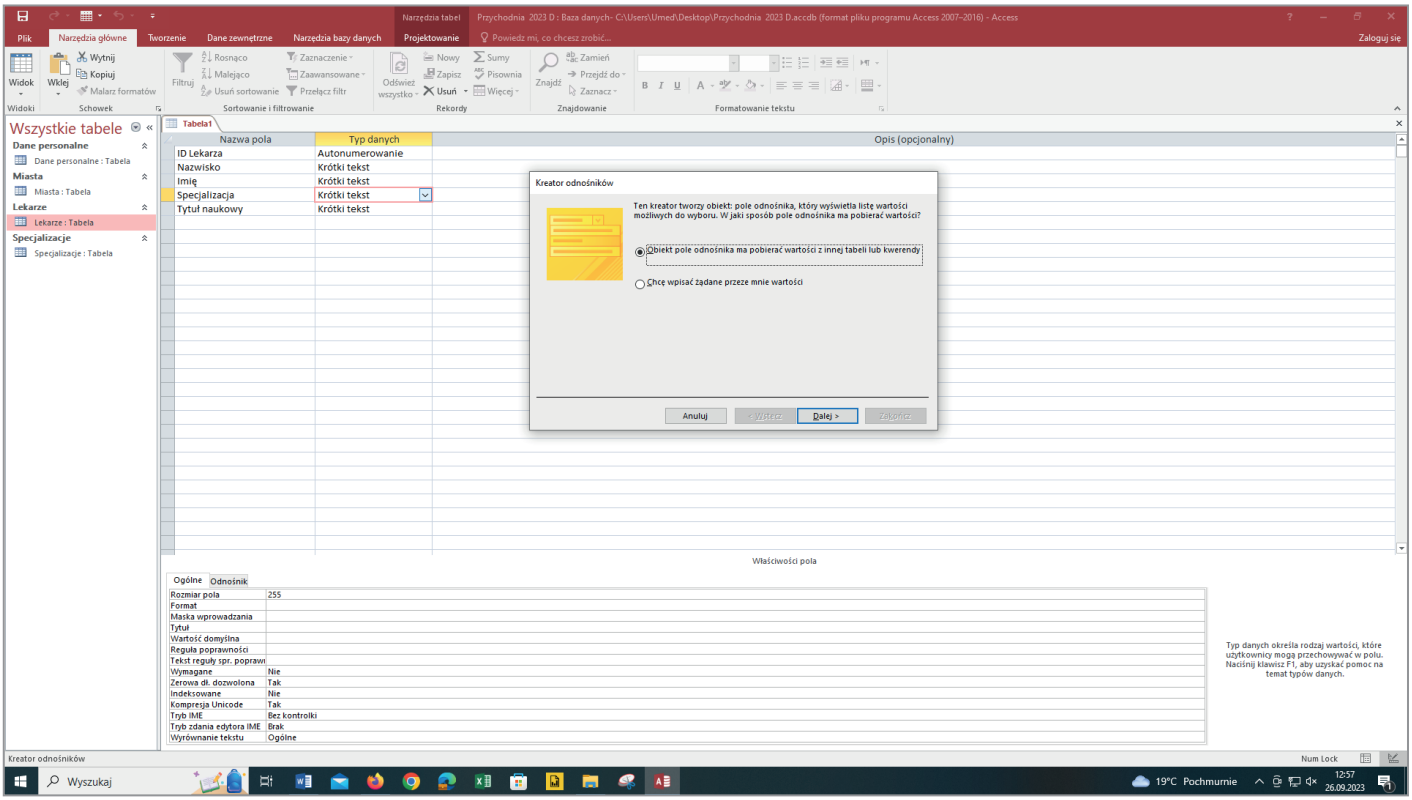
Ryc. 7.23. Tworzenie relacji do tabeli „Specjalizacja”

Teraz można już utworzyć najpierw tabelę w widoku projektu, a potem z danymi, wykorzystując odwołanie do tabeli „Specjalizacja”. W rekordzie „Specjalizacja” wybieramy typ danych Kreator odnośników (➥ **ryc. 7.23**).



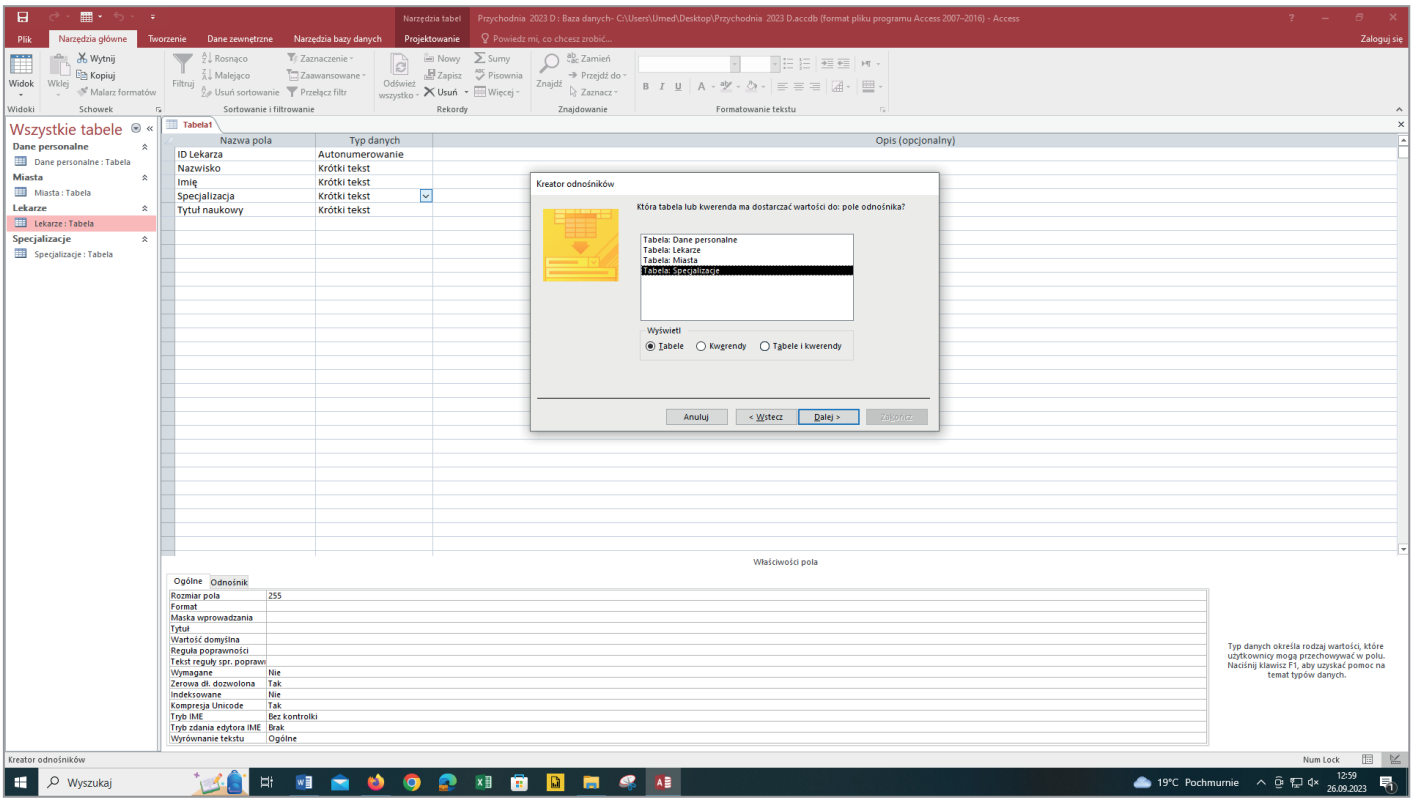
Ryc. 7.24. Przyporządkowanie odpowiedniego sposobu pobierania wartości

W wyskakującym oknie wybieramy sposób pobierania wartości z innej tabeli lub kwerendy przez kreator odnośników (➥ **ryc. 7.24**).



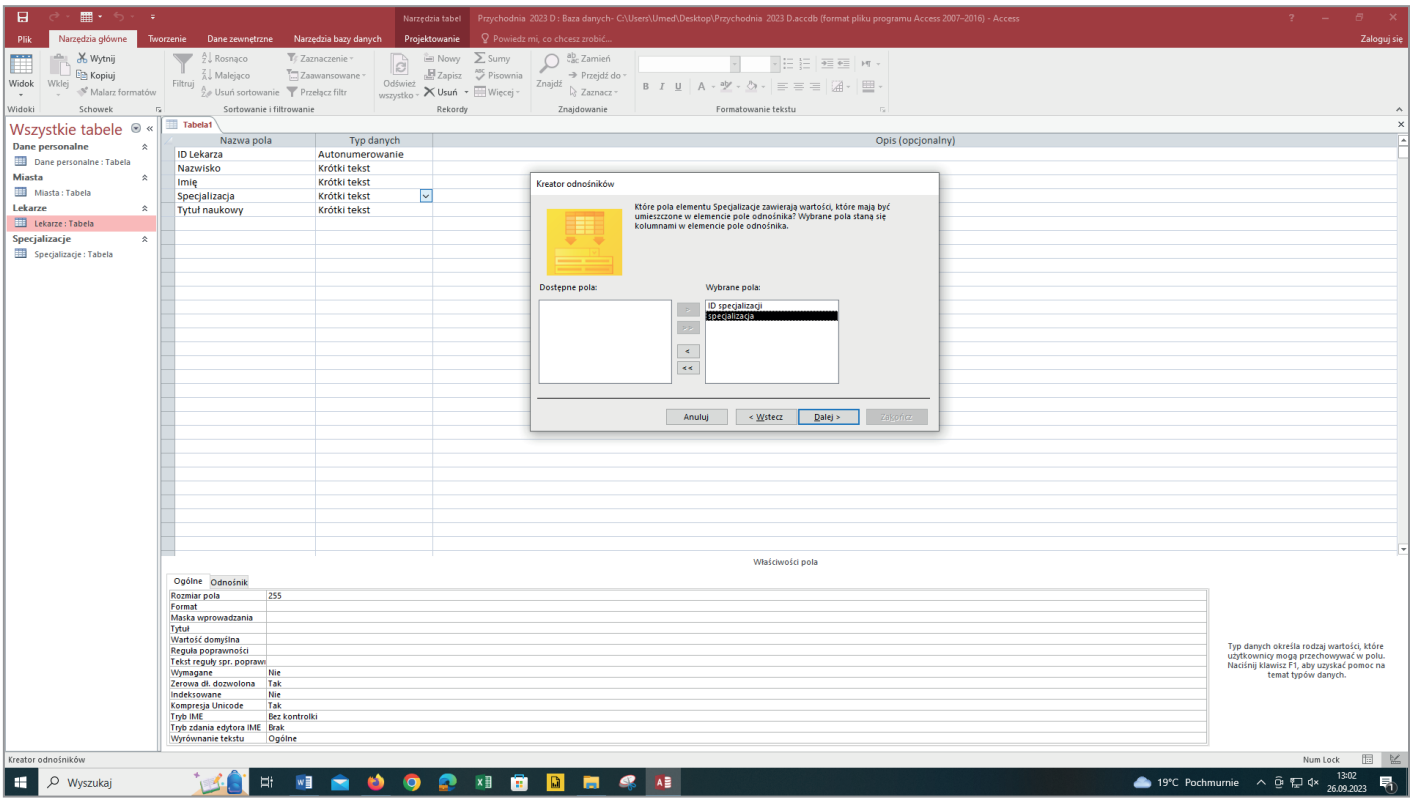
Po kliknięciu Dalej mamy możliwość wyboru tabeli, z której pobrane będą wartości (➥ **ryc. 7.25**).

Ryc. 7.25. Wybór tabeli



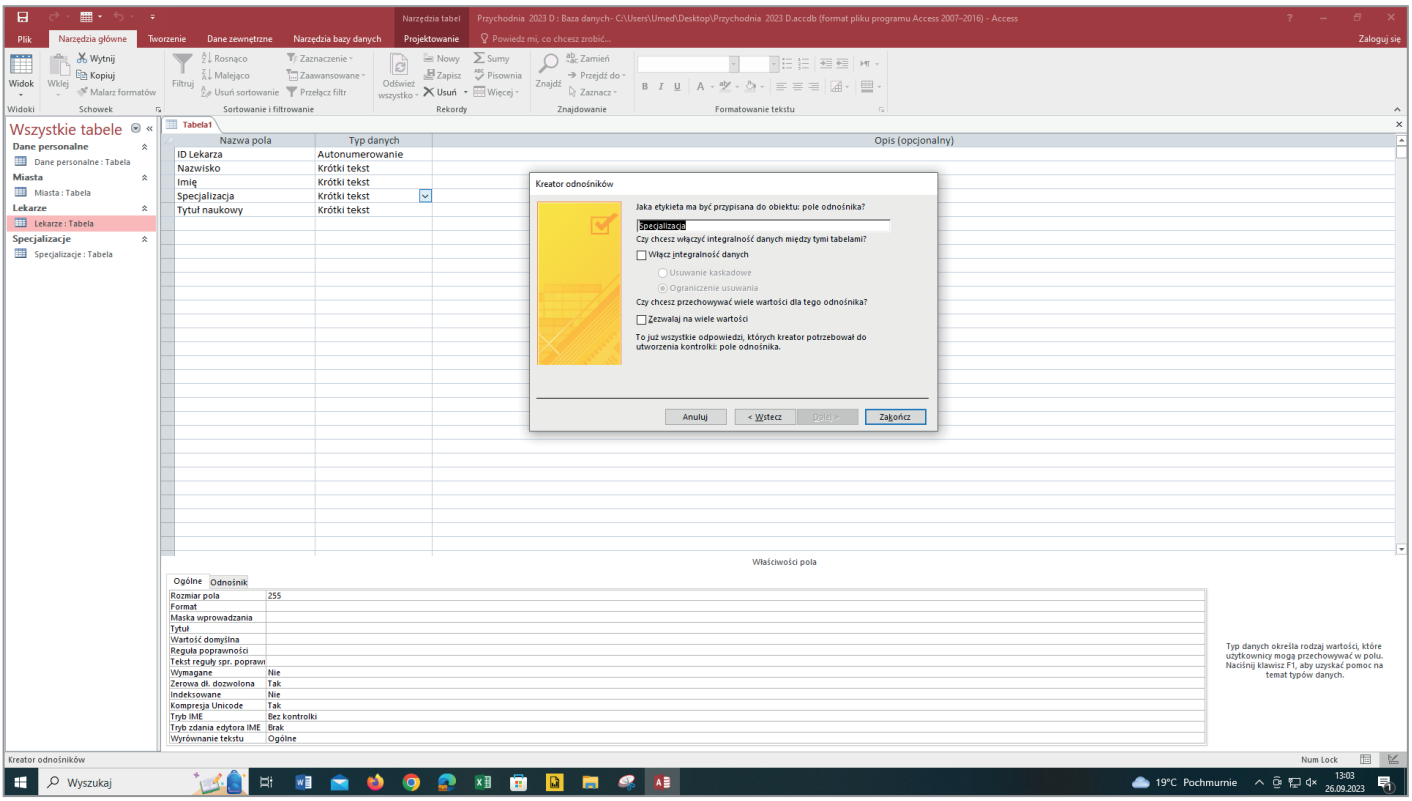
W kolejnym oknie mamy możliwość wyboru pól do umieszczenia w elemencie pola odnośnika. Pola przenosimy za pomocą przycisków ze strzałkami pomiędzy okienkami wyboru (➥ **ryc. 7.26**).

Ryc. 7.26. Przenoszenie danych



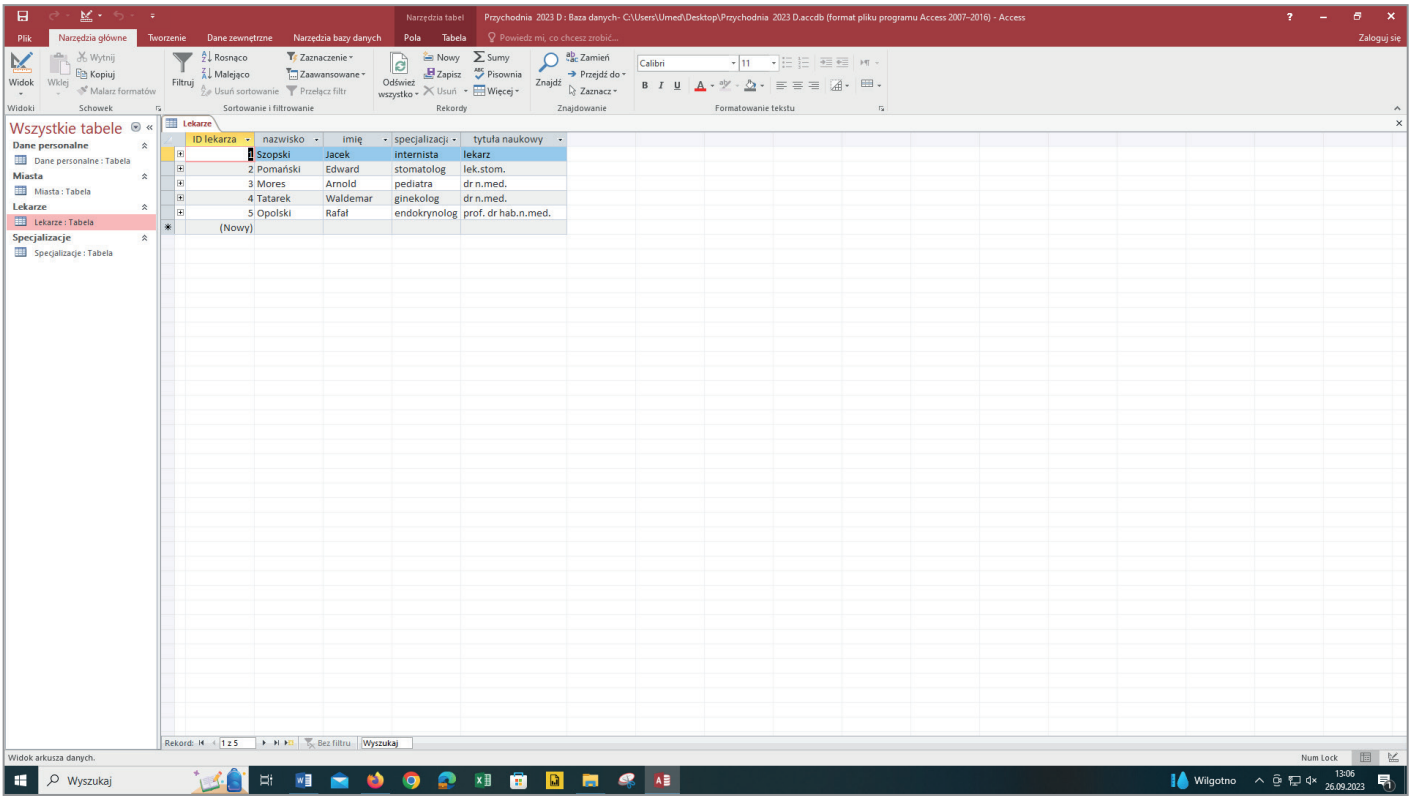
Następnie w kolejnych 3 okienkach dialogowych klikamy Dalej i w 4. Zakończ (➥ **ryc. 7.27**).

Ryc. 7.27. Zamykanie relacji



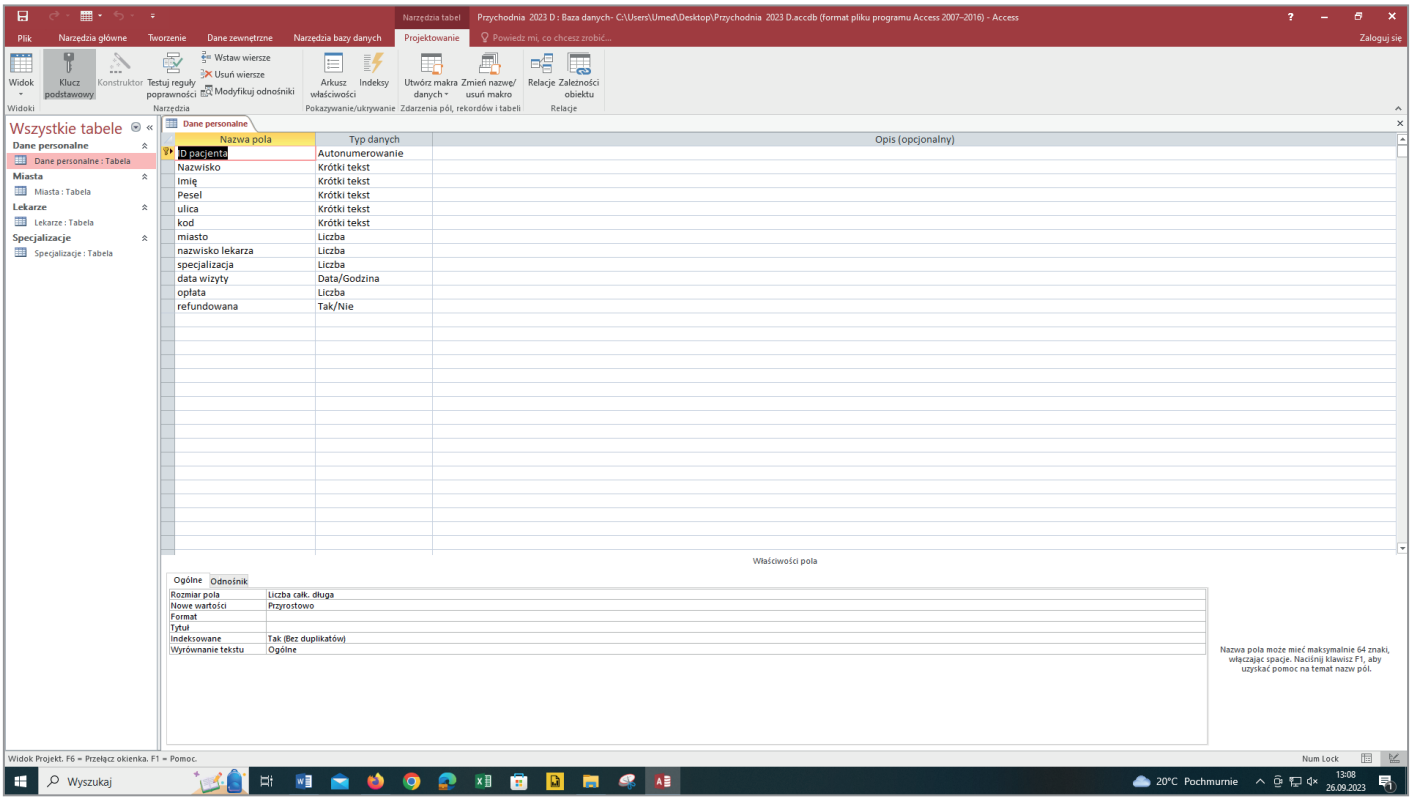
Nadajemy tabeli nazwę „Lekarze” i wypełniamy w sposób pokazany na ➥ **ryc. 7.28**, używając pola wyboru w pozycji „Specjalizacja” (➥ **ryc. 7.27**).

Ryc. 7.28. Widok tabeli „Lekarze”



Ryc. 7.29. Projekt tabeli „Dane personalne”

Teraz projektujemy tabelę „Dane personalne” (➥ **ryc. 7.29**).

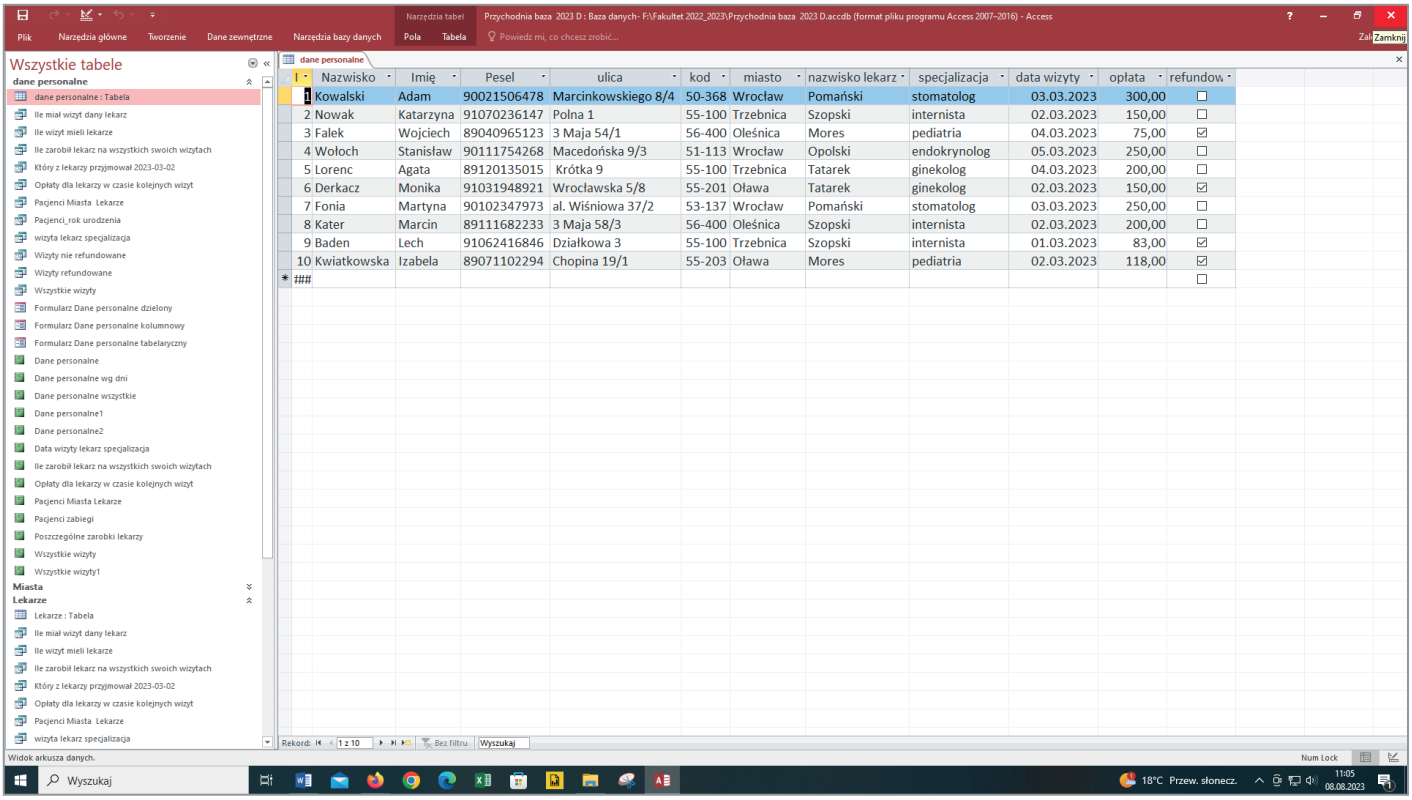


Uzupełniamy danymi z ➥ **tabeli 7.2** (➥ **ryc. 7.30**).

Dane personalne											
ID pacjenta	**Nazwisko**	**Imię**	**Pesel**	**ulica**	**kod**	**miasto**	**nazwisko lekarza**	**specjalizacja**	**data wizyty**	**opłata**	**refundowana**
1	Kowalski	Adam	90021506478	Marcinkowskiego 8/4	50-368	Wrocław	Pomański	stomatolog	03.03.2023	300,00	Nie
2	Nowak	Katarzyna	91070236147	Polna 1	55-100	Trzebnica	Szopski	internista	02.03.2023	150,00	Nie
3	Falek	Wojciech	89040965123	3 Maja 54/1	56-400	Oleśnica	Mores	pediatria	04.03.2023	75,00	Tak
4	Wołoch	Stanisław	90111754268	Macedońska 9/3	51-113	Wrocław	Opolski	endokrynolog	05.03.2023	250,00	Nie
5	Lorenc	Agata	89120135015	Krótka 9	55-100	Trzebnica	Tatarek	ginekolog	04.03.2023	200,00	Nie
6	Derkacz	Monika	91031948921	Wrocławska 5/8	55-201	Oława	Tatarek	ginekolog	02.03.2023	150,00	Tak
7	Fonia	Martyna	90102347973	al. Wiśniowa 37/2	53-137	Wrocław	Pomański	stomatolog	03.03.2023	250,00	Nie
8	Kater	Marcin	89111682233	3 Maja 58/3	56-400	Oleśnica	Szopski	internista	02.03.2023	200,00	Nie
9	Baden	Lech	91062416846	Działkowa 3	55-100	Trzebnica	Szopski	internista	01.03.2023	83,00	Tak
10	Kwiatkowska	Izabela	89071102294	Chopina 19/1	55-203	Oława	Mores	pediatria	02.03.2023	118,00	Tak

Tabela 7.2. Przykładowe dane pacjentów

Ryc. 7.30. Tabela w bazie z danymi pacjentów

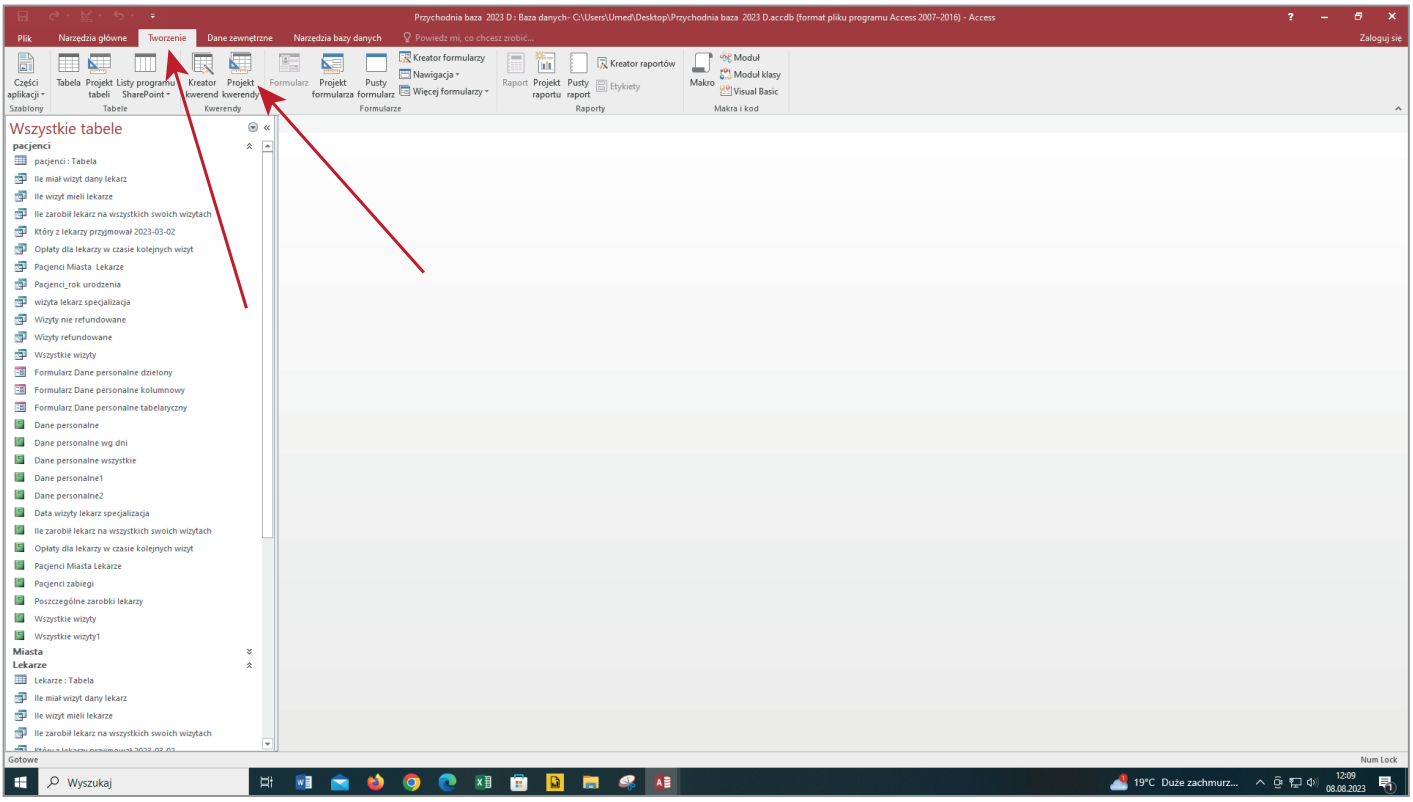


I	Nazwisko	Imię	Pesel	ulica	kod	miasto	nazwisko lekarz	specjalizacja	data wizyty	opłata	refundow
1	Kowalski	Adam	90021506478	Marcinkowskiego 8/4	50-368	Wrocław	Pomański	stomatolog	03.03.2023	300,00	☐
2	Nowak	Katarzyna	91070236147	Polna 1	55-100	Trzebnica	Szopski	internista	02.03.2023	150,00	☐
3	Falek	Wojciech	89040965123	3 Maja 54/1	56-400	Oleśnica	Mores	pediatria	04.03.2023	75,00	☑
4	Wołoch	Stanisław	90111754268	Macedońska 9/3	51-113	Wrocław	Opolski	endokrynolog	05.03.2023	250,00	☐
5	Lorenc	Agata	89120135015	Krótka 9	55-100	Trzebnica	Tatarek	ginekolog	04.03.2023	200,00	☐
6	Derkacz	Monika	91031948921	Wrocławska 5/8	55-201	Oława	Tatarek	ginekolog	02.03.2023	150,00	☑
7	Fonia	Martyna	90102347973	al. Wiśniowa 37/2	53-137	Wrocław	Pomański	stomatolog	03.03.2023	250,00	☐
8	Kater	Marcin	89111682233	3 Maja 58/3	56-400	Oleśnica	Szopski	internista	02.03.2023	200,00	☐
9	Baden	Lech	91062416846	Działkowa 3	55-100	Trzebnica	Szopski	internista	01.03.2023	83,00	☑
10	Kwiatkowska	Izabela	89071102294	Chopina 19/1	55-203	Oława	Mores	pediatria	02.03.2023	118,00	☑
*	###										☐

7.2.1. Przykład tworzenia kwerendy

Ryc. 7.31. Przygotowanie do tworzenia projektu kwerendy

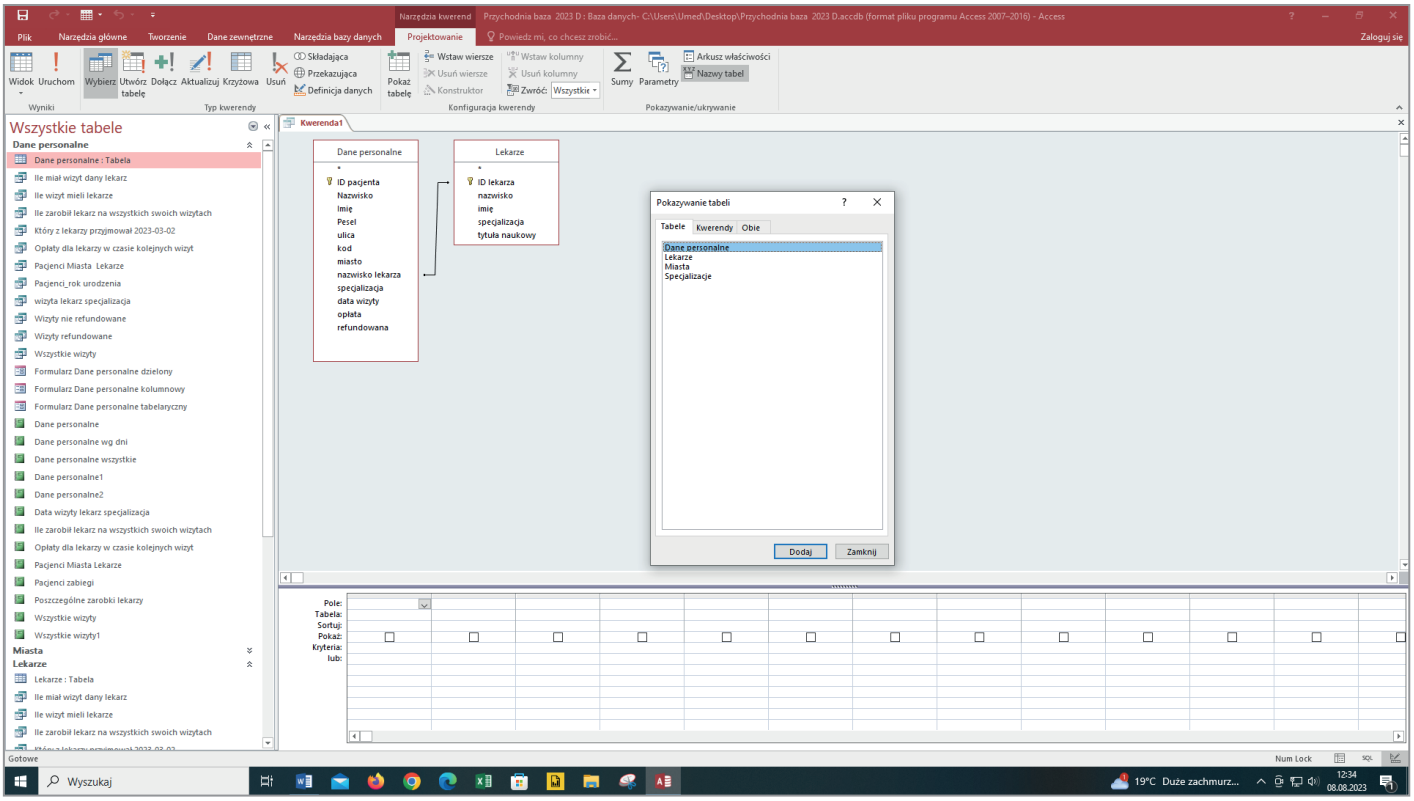
Wybieramy kartę Tworzenie, a następnie ze wstążki polecenie Projekt kwerendy (➡ **ryc. 7.31**).



Otrzymujemy ekran widoczny na ➥ **ryc. 7.32**.

Zamykamy okno Pokazywanie tabeli.

Ryc. 7.32. Wybór odpowiednich tabel



Rozpoczynamy tworzenie kwerendy, która pokaże lekarzy przyjmujących w dniu 2 marca 2023 roku wraz z przyjętymi przez nich pacjentami.

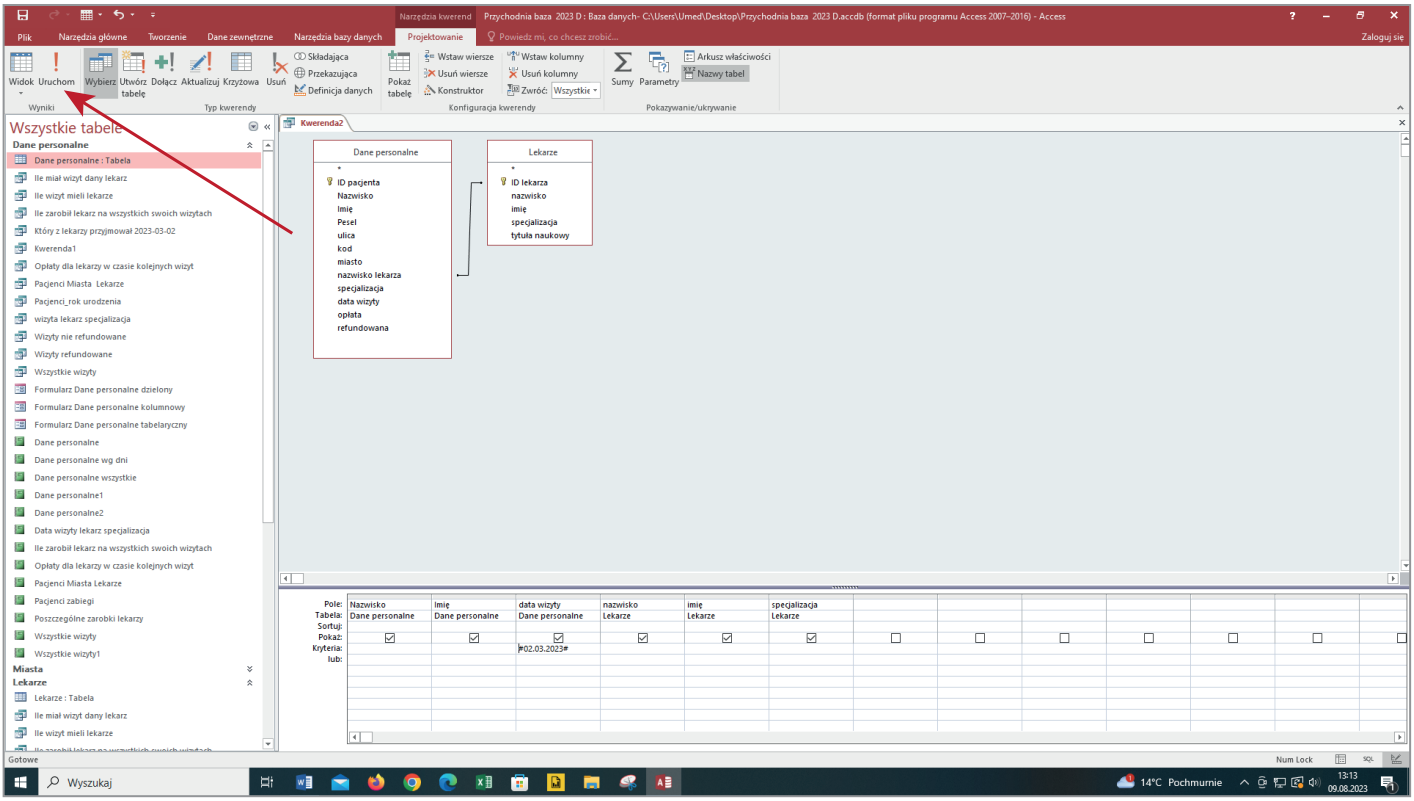
Najpierw klikamy dwukrotnie myszką nazwisko, imię i datę wizyty w tabeli „Dane personalne”, następnie klikamy dwukrotnie nazwisko, imię i specjalizację w tabeli „Lekarze”.

Każda z wybranych pozycji przeniesiona zostaje do tabeli w dolnej części okna i ułożona w kolejnych kolumnach. Pod komórkami pól, w wierszu Tabela pojawiają się informacje, z której tabeli pochodzi wybrany parametr.

W kolumnie „Data wizyty”, w wierszu „Kryteria” wpisujemy datę, np. 2.03.2023.

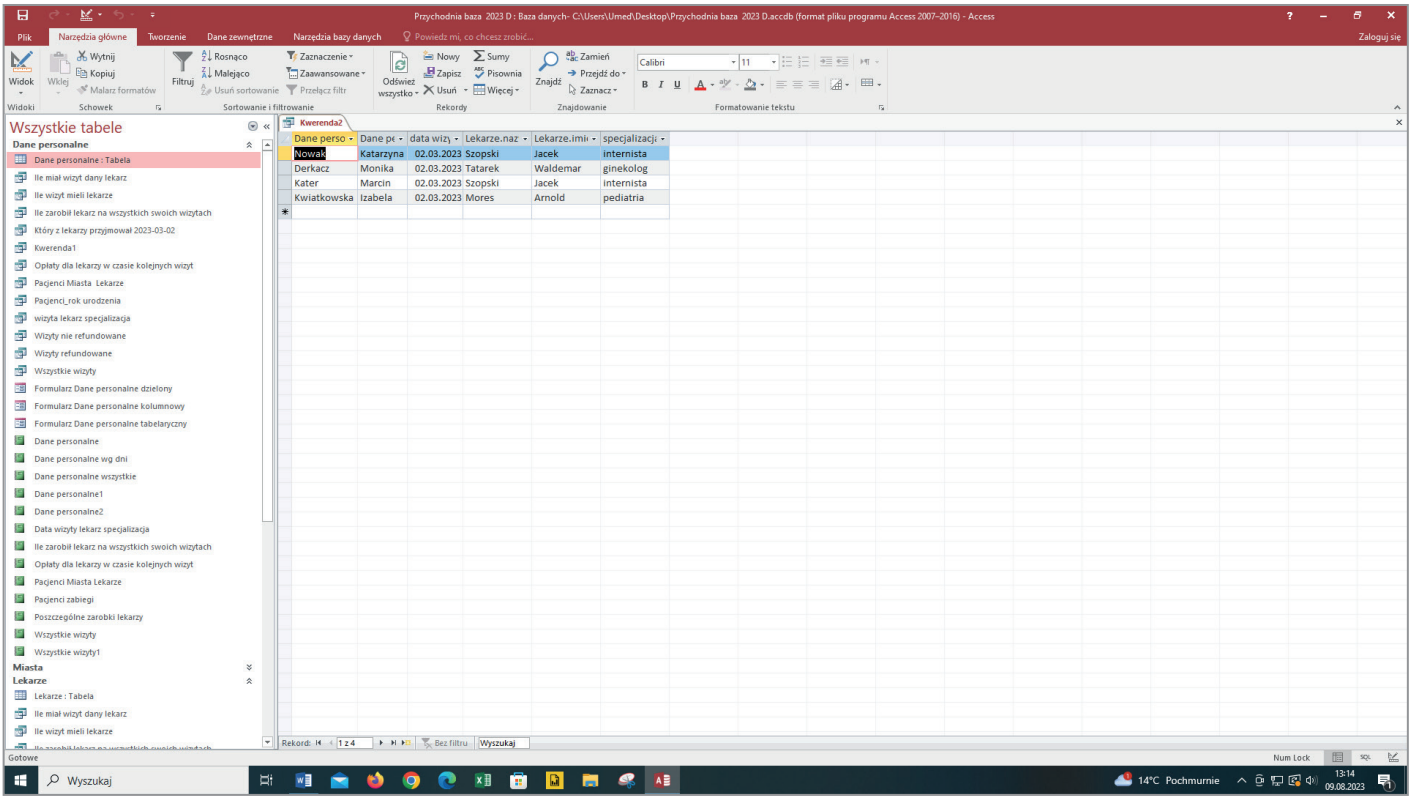
Następnie klikamy na wstążce polecenie Uruchom ! (➥ **ryc. 7.33**).

Ryc. 7.33. Uruchamianie utworzonej kwerendy



Otrzymujemy na ekranie wynik działania zaprojektowanej kwerendy (➥ **ryc. 7.34**). W tabeli możemy wyczytać pożądane przez nas informacje. Inne kwerendy tworzymy na tych samych zasadach.

Ryc. 7.34. Wynik kwerendy

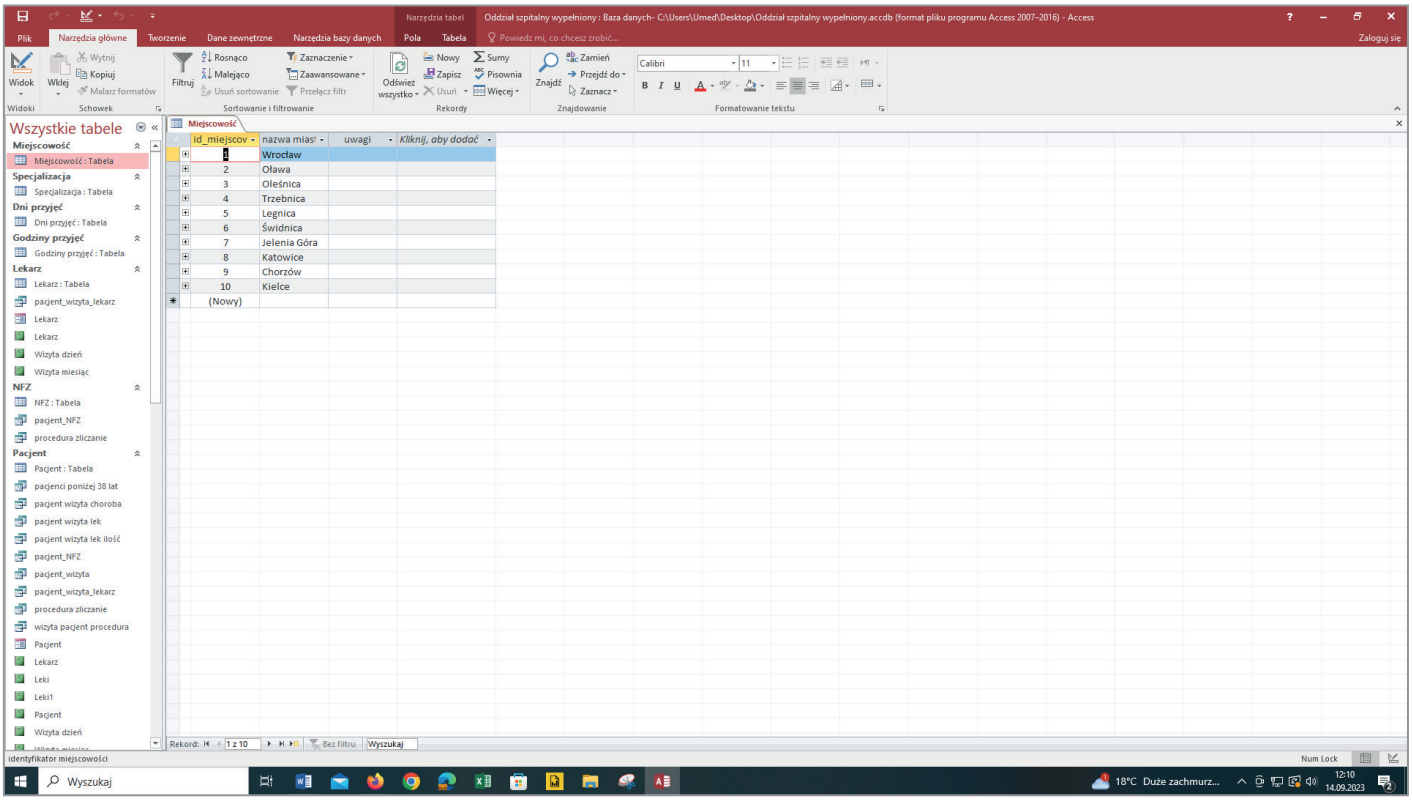


7.3. Baza „Oddział szpitalny”

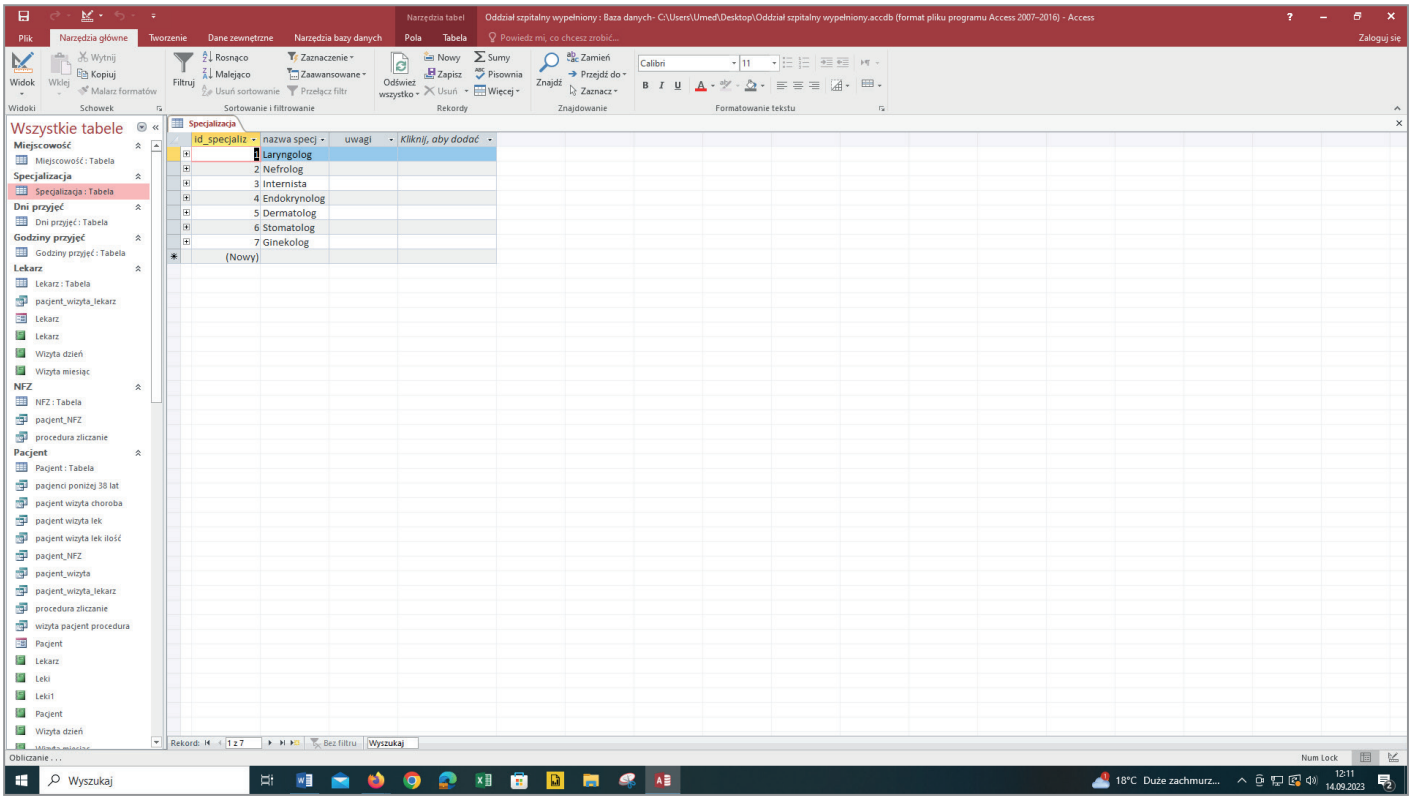
Najpierw proszę (w znany już z poprzednich ćwiczeń sposób) kolejno utworzyć tabele: „Miejscowość”, „Specjalizacja”, „Dni przyjęć”, „Godziny przyjęć” i „NFZ” (ta ostatnia będzie zawierać informacje o oddziałach NFZ, ➥ **ryc. 7.35–7.39**).

Proszę pamiętać, że istotna jest kolejność tworzenia tabel.

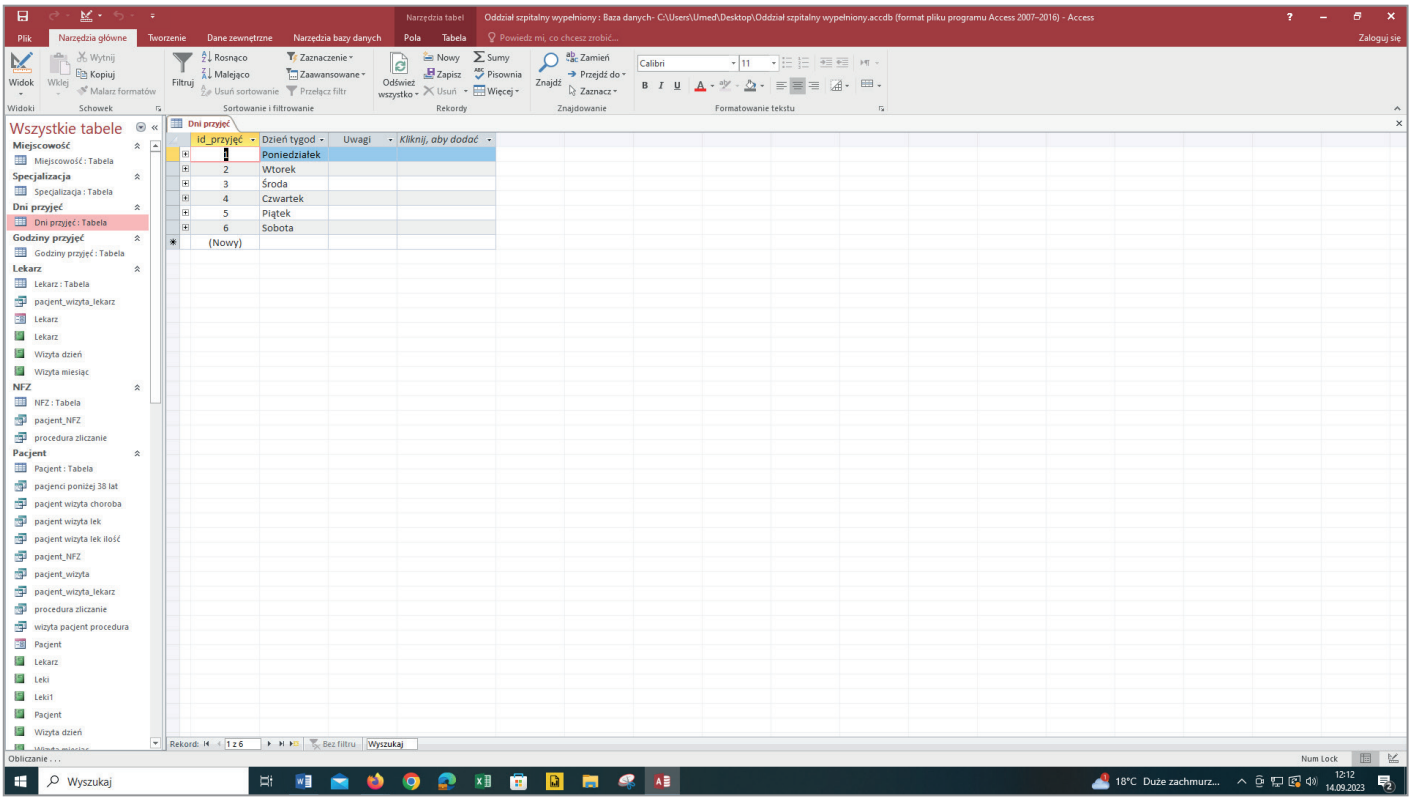
Ryc. 7.35. Tabela „Miejscowość”



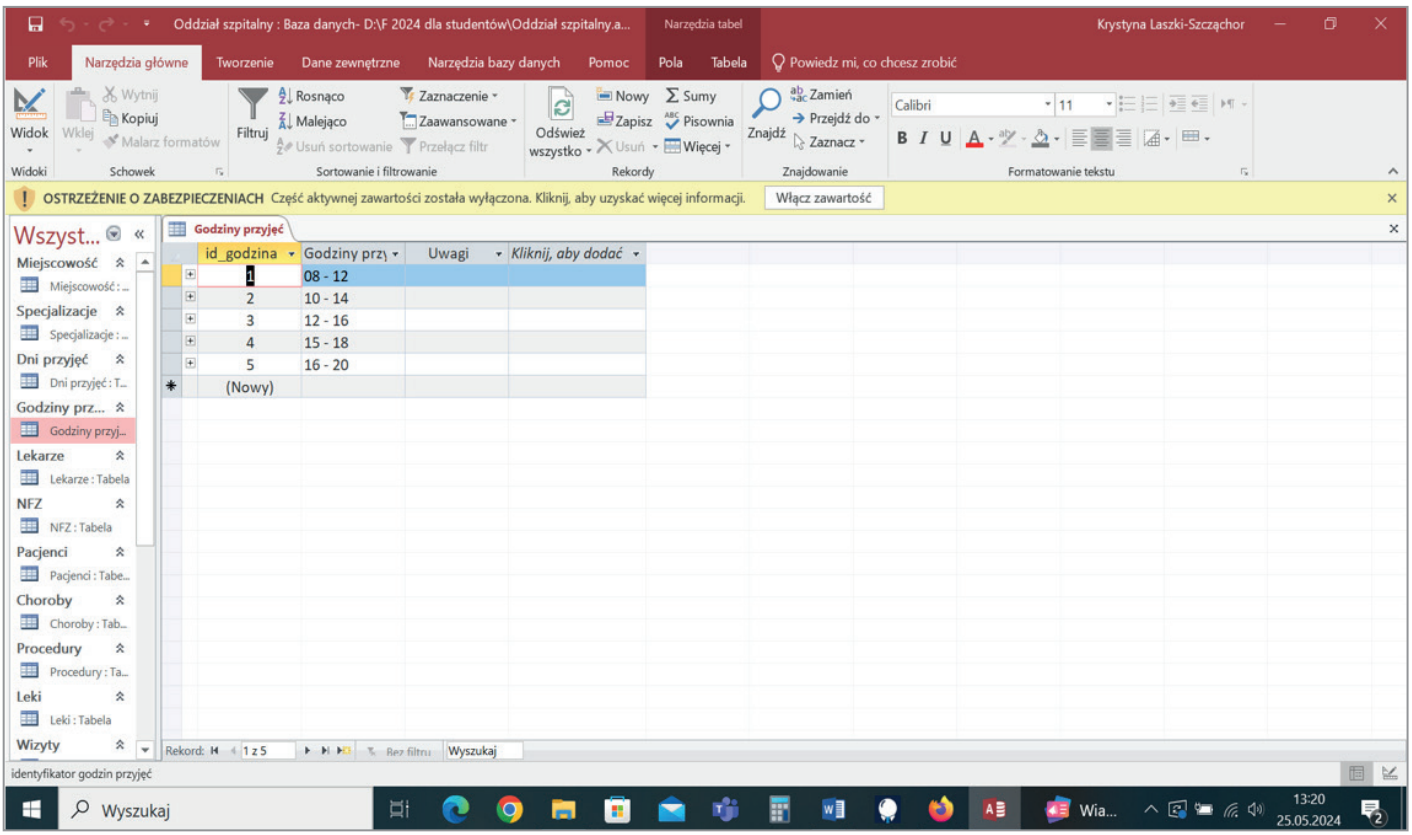
Ryc. 7.36. Tabela „Specjalizacja”



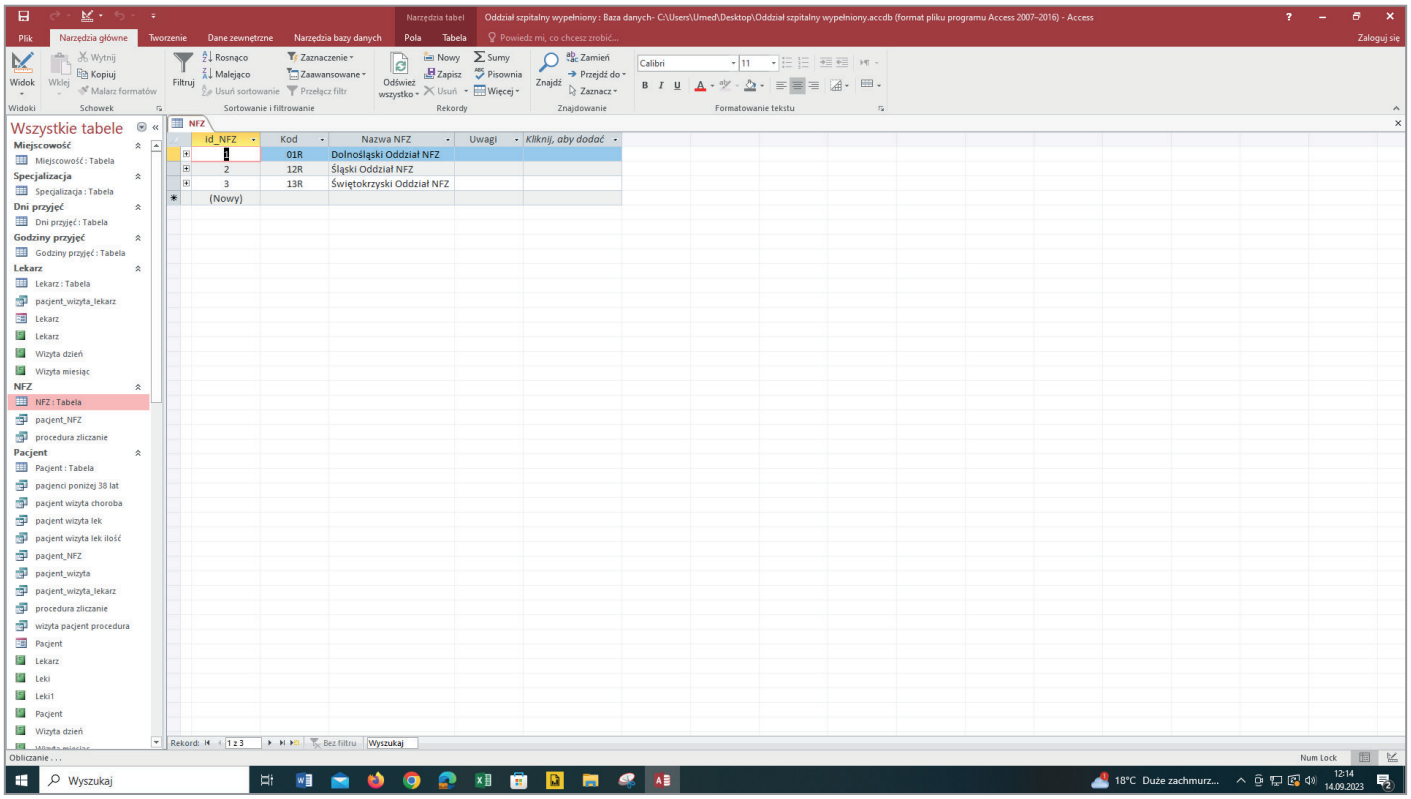
Ryc. 7.37. Tabela „Dni przyjęć”



Ryc. 7.38. Tabela „Godziny przyjęć”

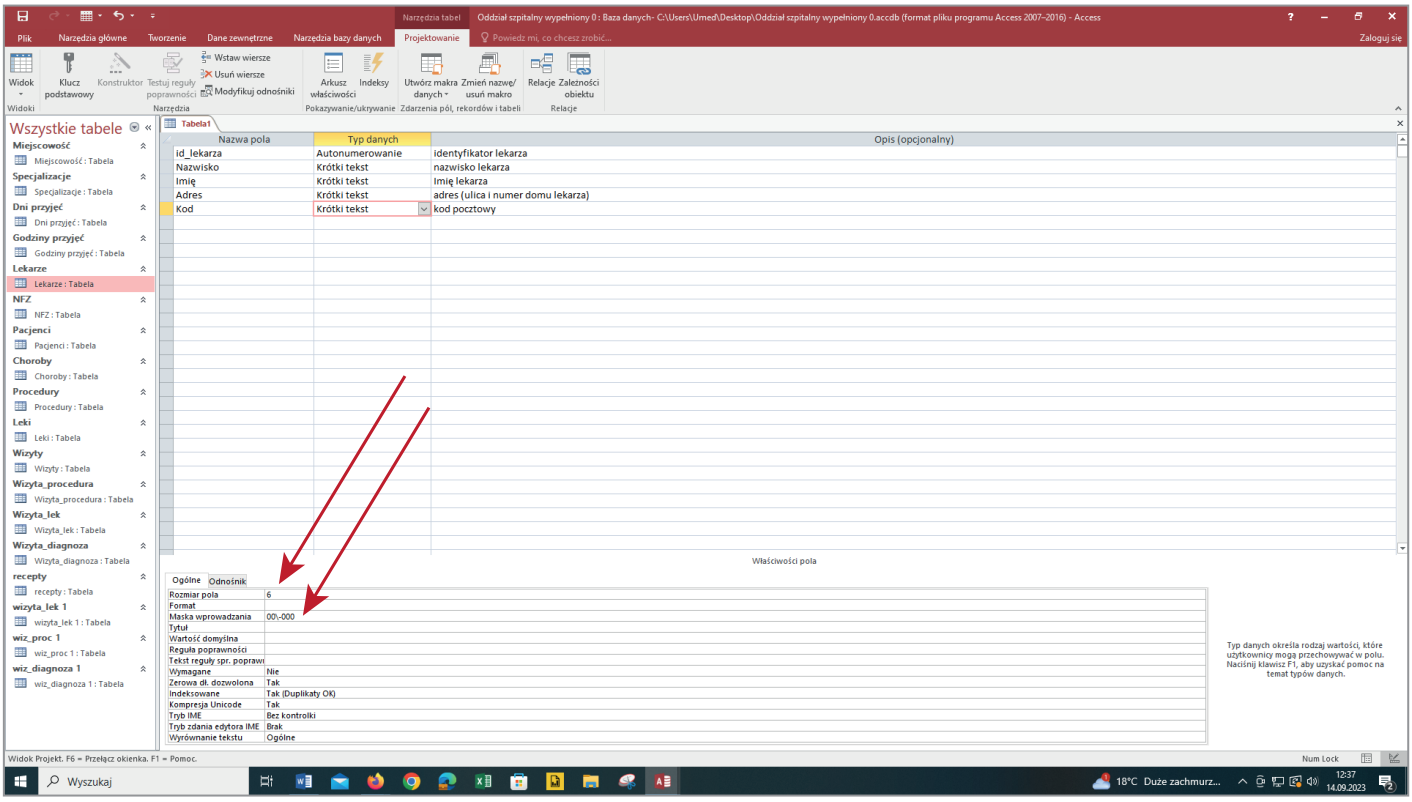


Ryc. 7.39. Tabela zawierająca oddziały NFZ



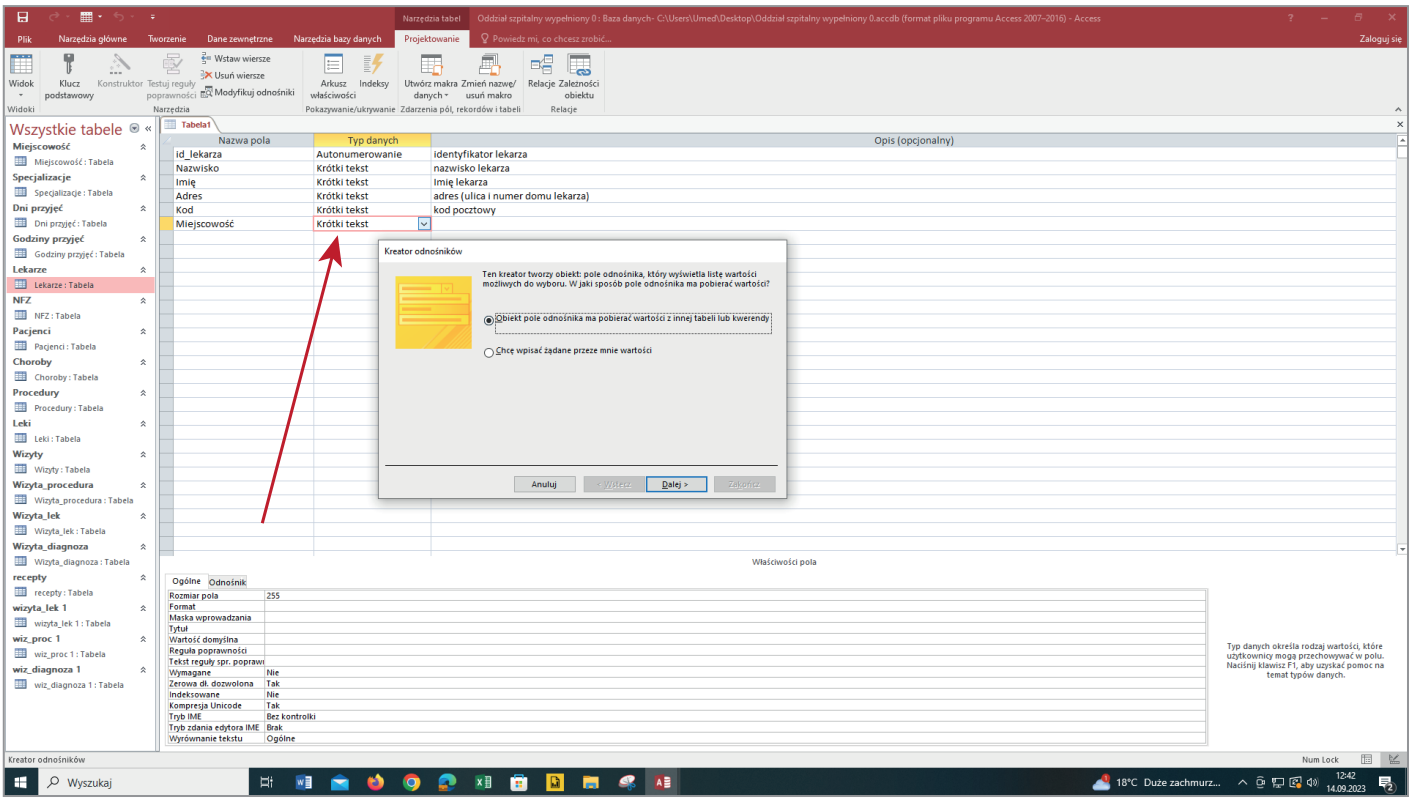
Ryc. 7.40. Widok projektu tabeli „Lekarze"

Jako typ danych pozycji Kod wybieramy Krótki tekst, a w tabeli na dole okna ustawiamy rozmiar pola na 6 i maskę wprowadzania na 00/-000 (➡ **ryc. 7.40**).



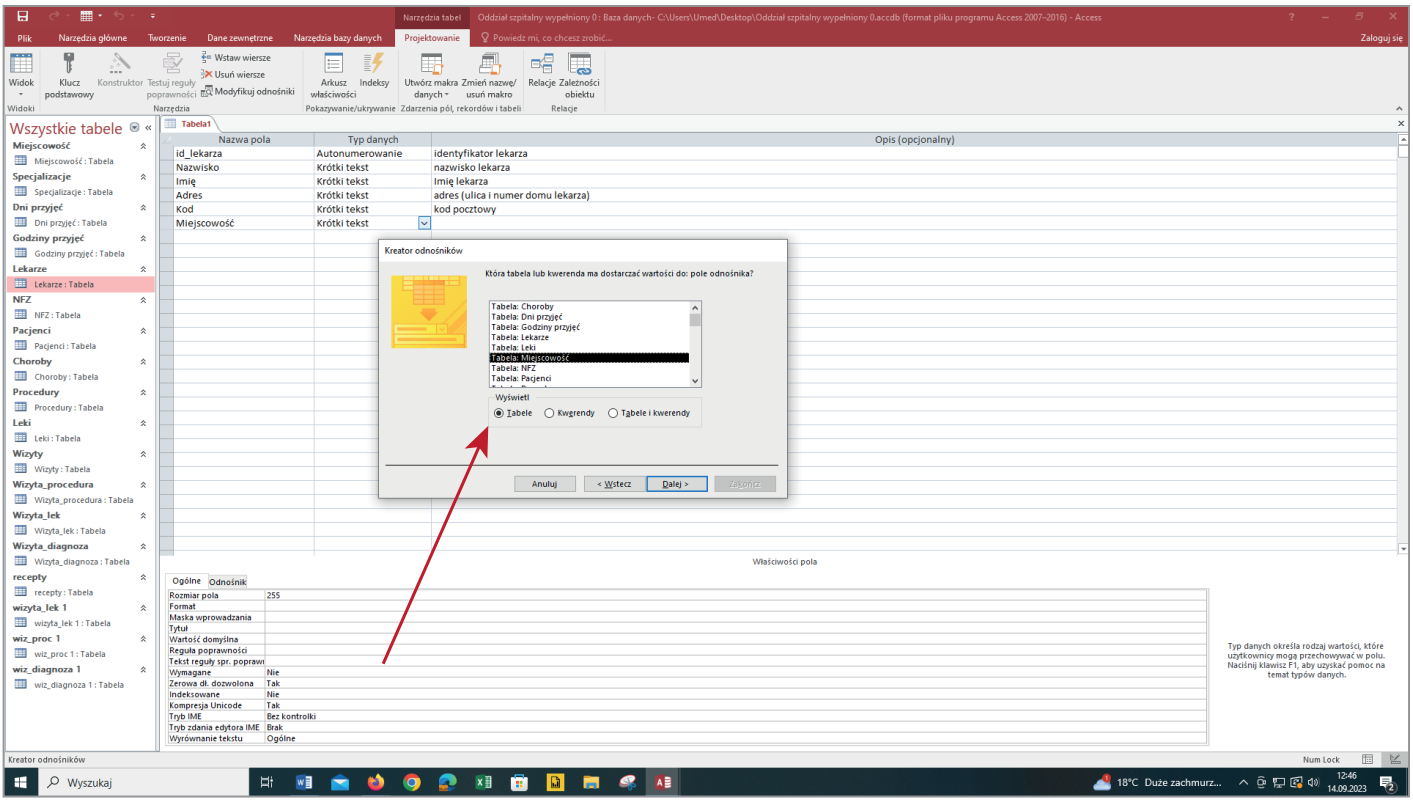
Ryc. 7.41. Wybieranie Kreatora odnośników jako typu danych

W polu miejscowość wybieramy jako typ danych Kreator odnośników (➥ **ryc. 7.41**).



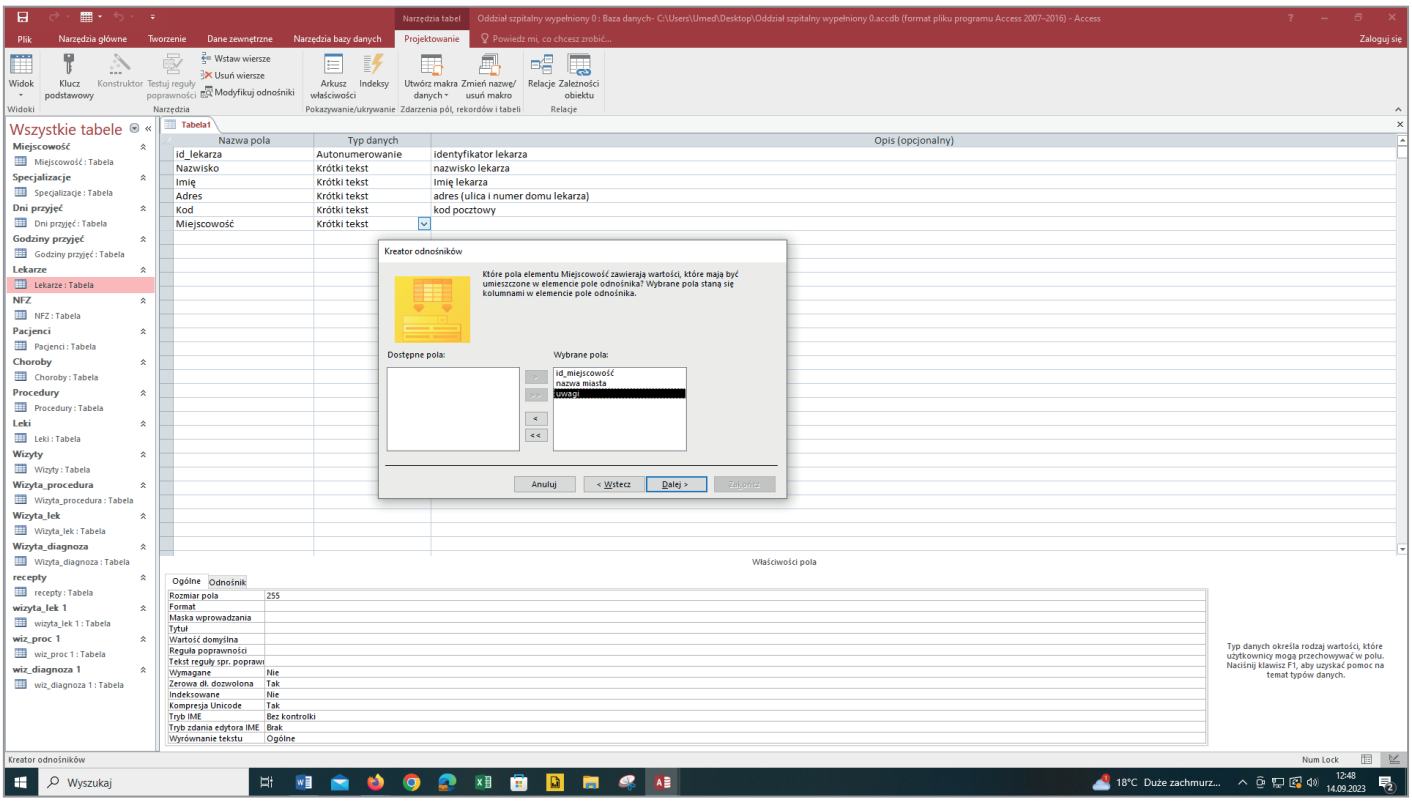
Ryc. 7.42. Tworzenie relacji z tabelą „Miejscowość"

Potem kliknięciu Dalej w następnym oknie tworzymy relację z tabelą „Miejscowość" i klikamy Dalej (➥ **ryc. 7.42**).



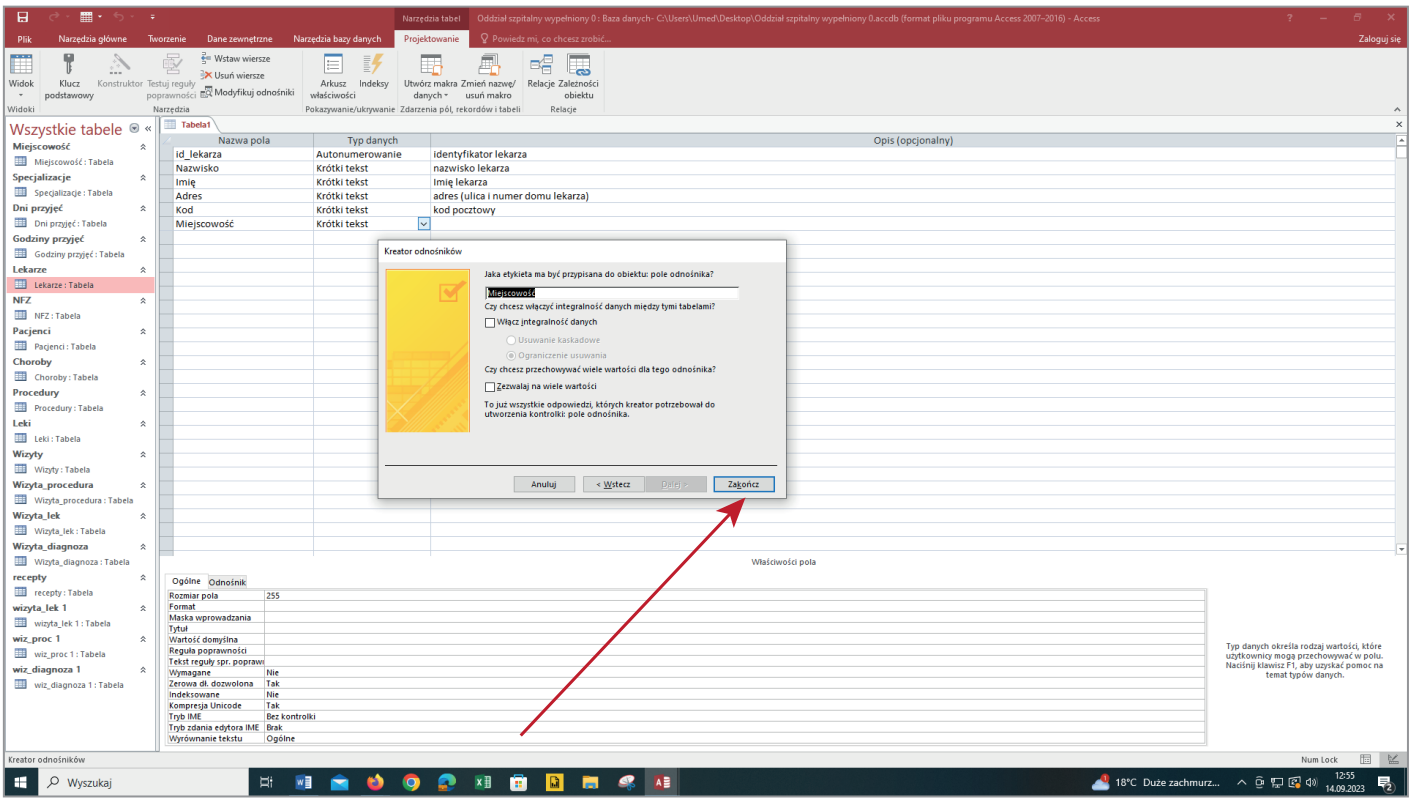
Ryc. 7.43. Przenoszenie danych z tabeli

W następnym oknie wybieramy pola, które zawierają interesujące nas dane. Przenosimy je strzałkami pomiędzy okienkami, aby umieścić je w tabeli (➥ **ryc. 7.43**).



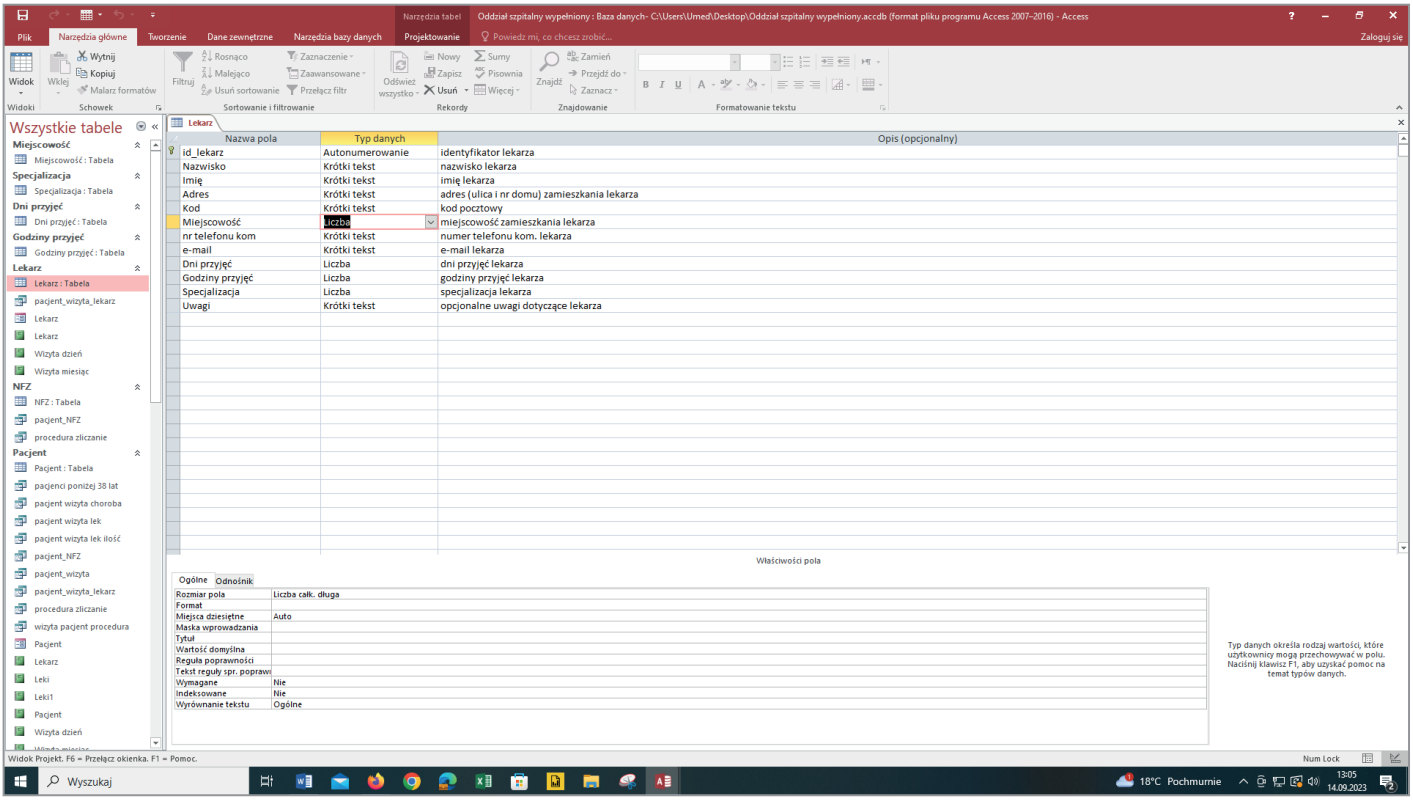
Ryc. 7.44. Zakończenie tworzenia tabeli

Ponownie klikamy przycisk Dalej w 3 kolejnych oknach dialogowych, a następnie Zakończ (➥ **ryc. 7.44**).



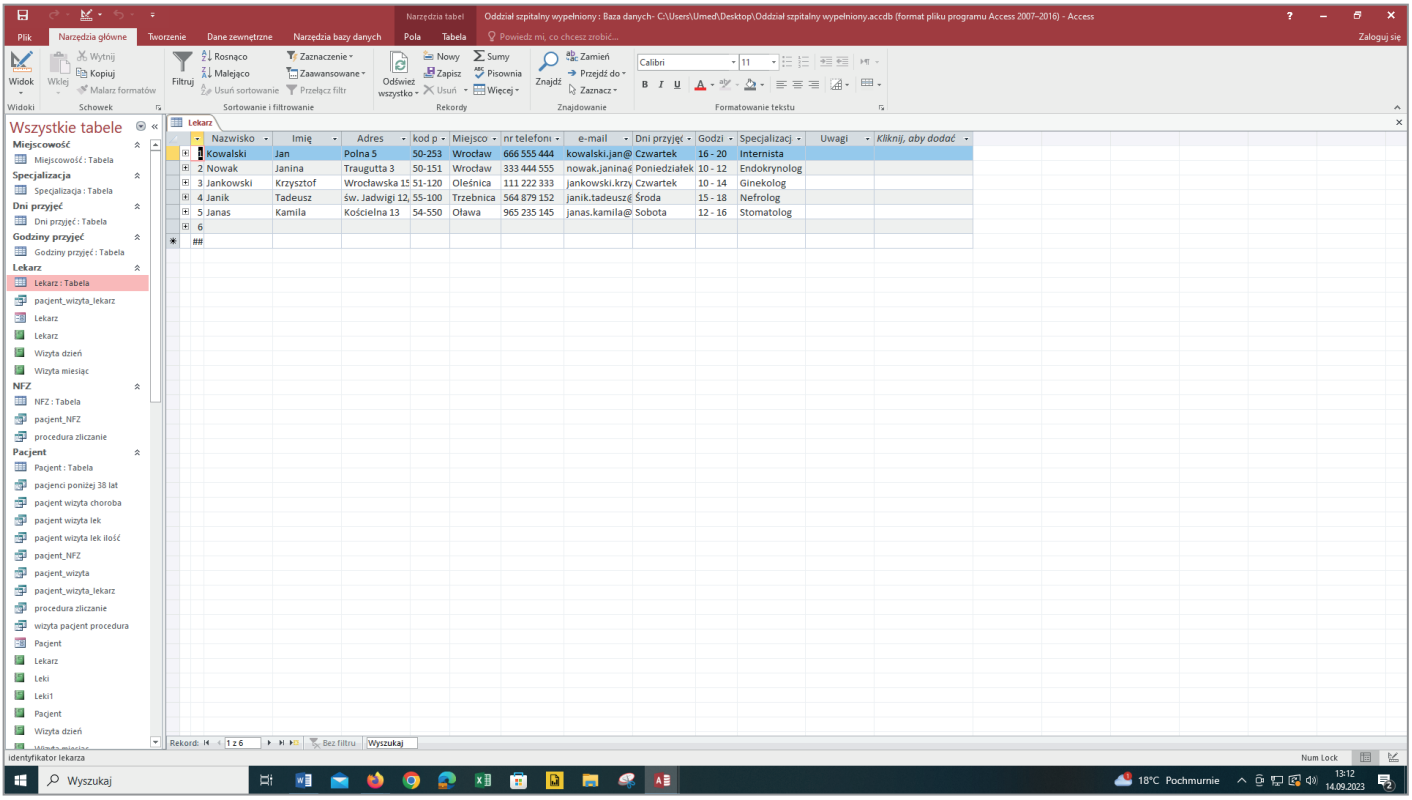
Ryc. 7.45. Pełny widok projektu tabeli „Lekarze”

W dalszej części projektu tabeli „Lekarze” w rekordach „Dni przyjęć”, „Godziny przyjęć” i „Specjalizacja” jako typ danych analogicznie wybieramy Kreator odnośników i poruszamy się po kolejnych oknach, jak opisano powyżej (➥ **ryc. 7.45**).



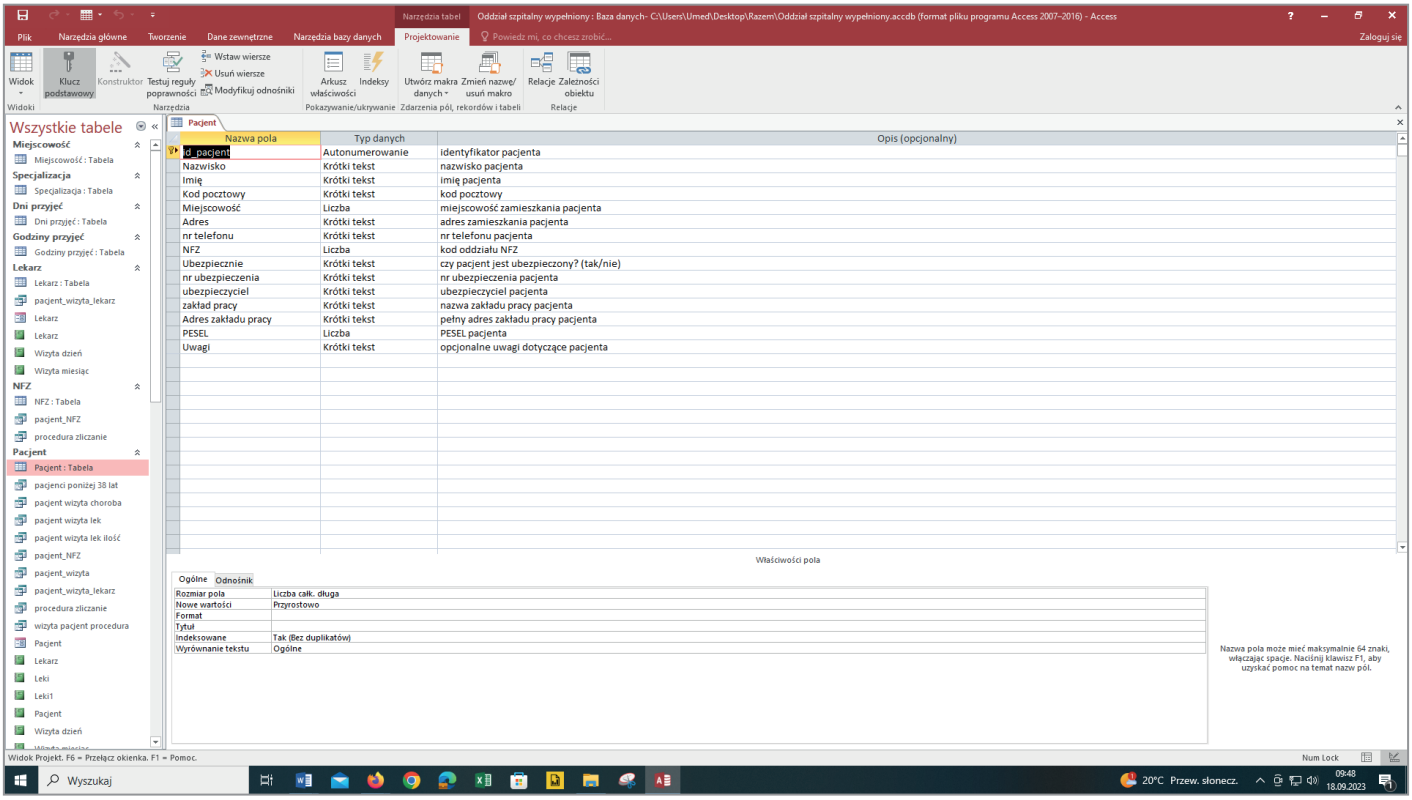
Ryc. 7.46. Uzupełnianie danych w tabeli

Następnie otwieramy projekt tabeli i wpisujemy lub wybieramy dane dotyczące lekarzy (➥ **ryc. 7.46**).



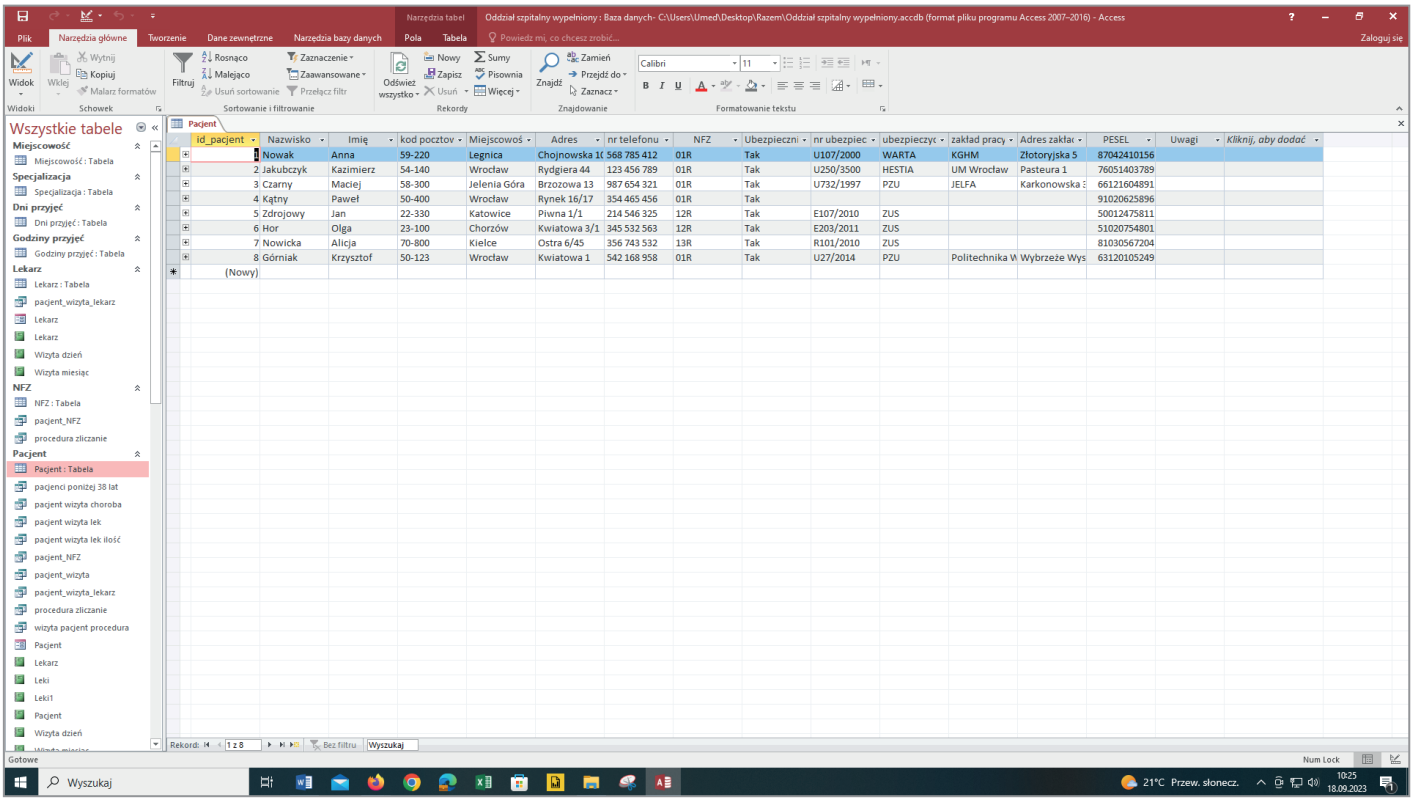
W dalszej kolejności analogicznie do projektu tabeli „Lekarze” tworzymy projekt tabeli „Pacjenci”. Kreatora odnośników używamy w pozycjach „Miejscowość” i „NFZ” (➥ **ryc. 7.47**).

Ryc. 7.47. Projekt tabeli „Pacjenci”

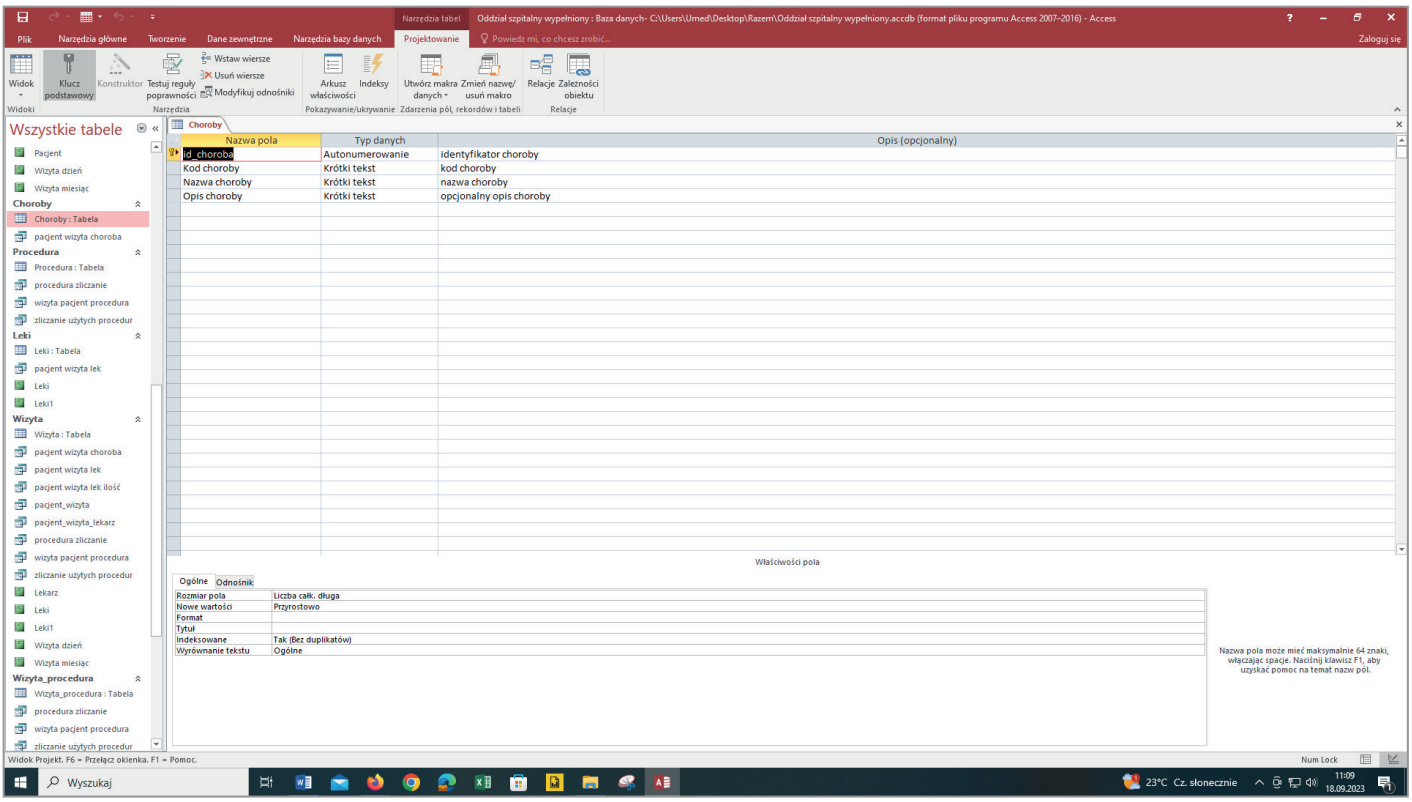


Ryc. 7.48. Tabela z danymi pacjentów

Otwieramy tabelę i wpisujemy dane pacjentów tak, jak widać na ➥ **ryc. 7.48**.

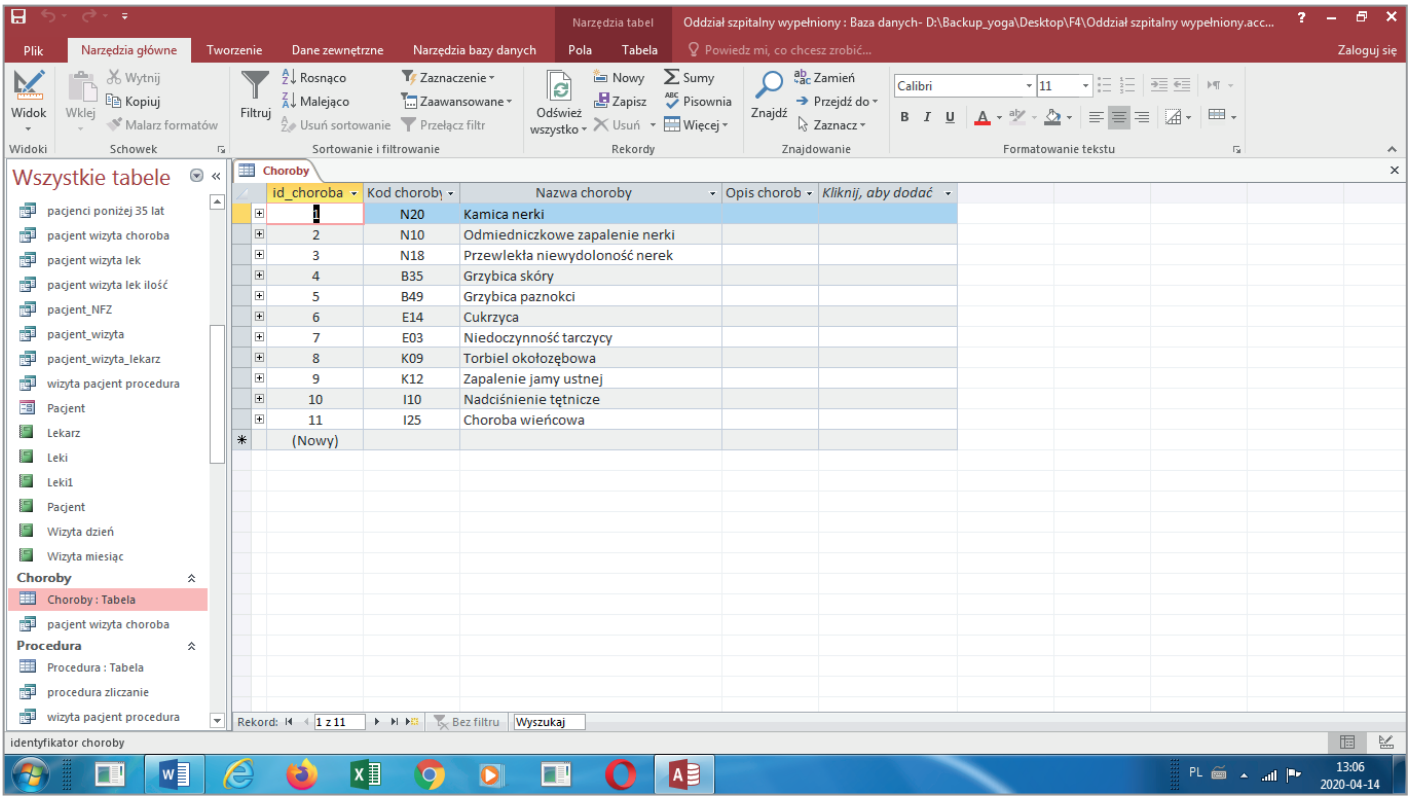


Kolejnym krokiem jest utworzenie tabeli „Choroby” z oznaczeniem i nazwami chorób (➥ **ryc. 7.49**).



Ryc. 7.49. Projekt tabeli „Choroby”

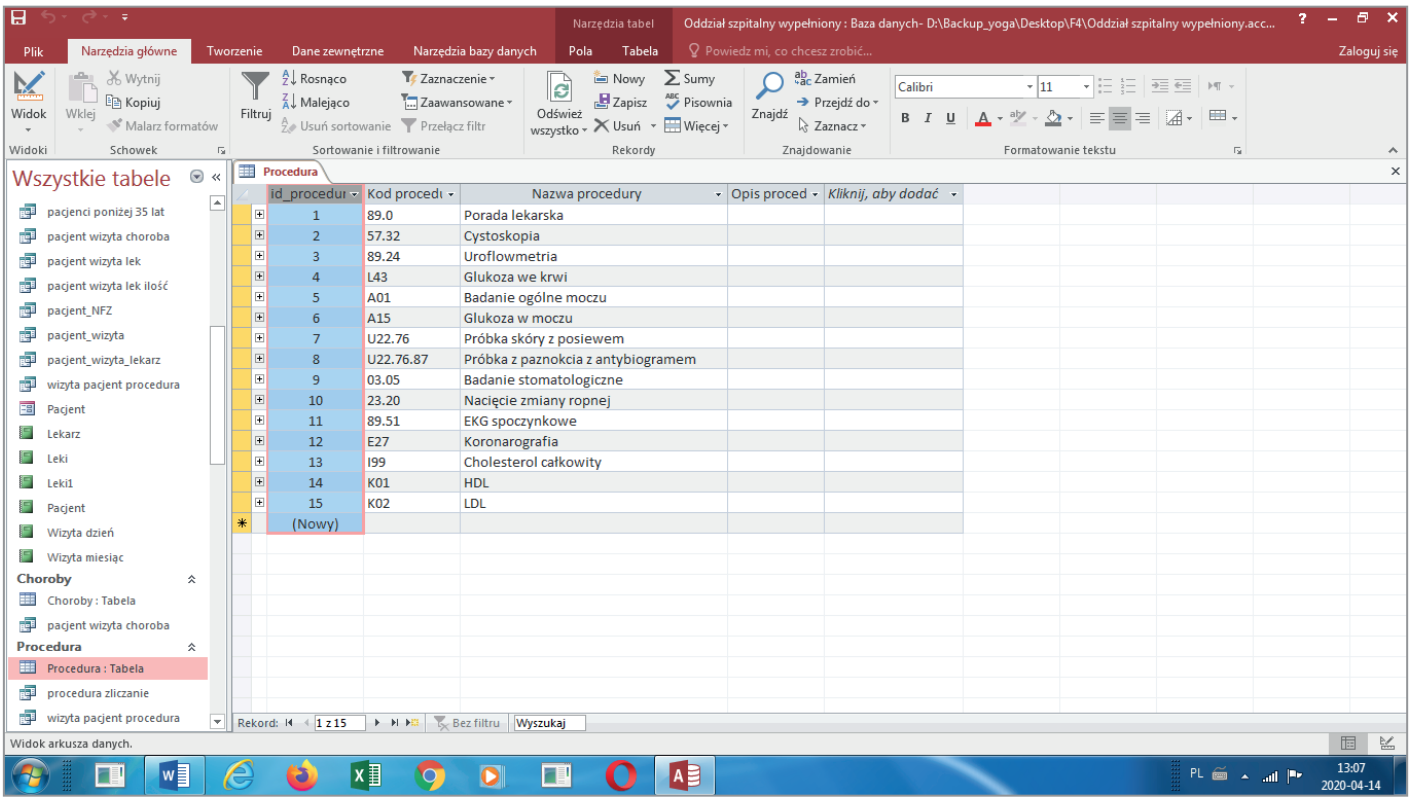
Tabelę otwieramy i uzupełniamy, wpisując kody i nazwy chorób (➥ **ryc. 7.50**).



Ryc. 7.50. Tabela z danymi chorób

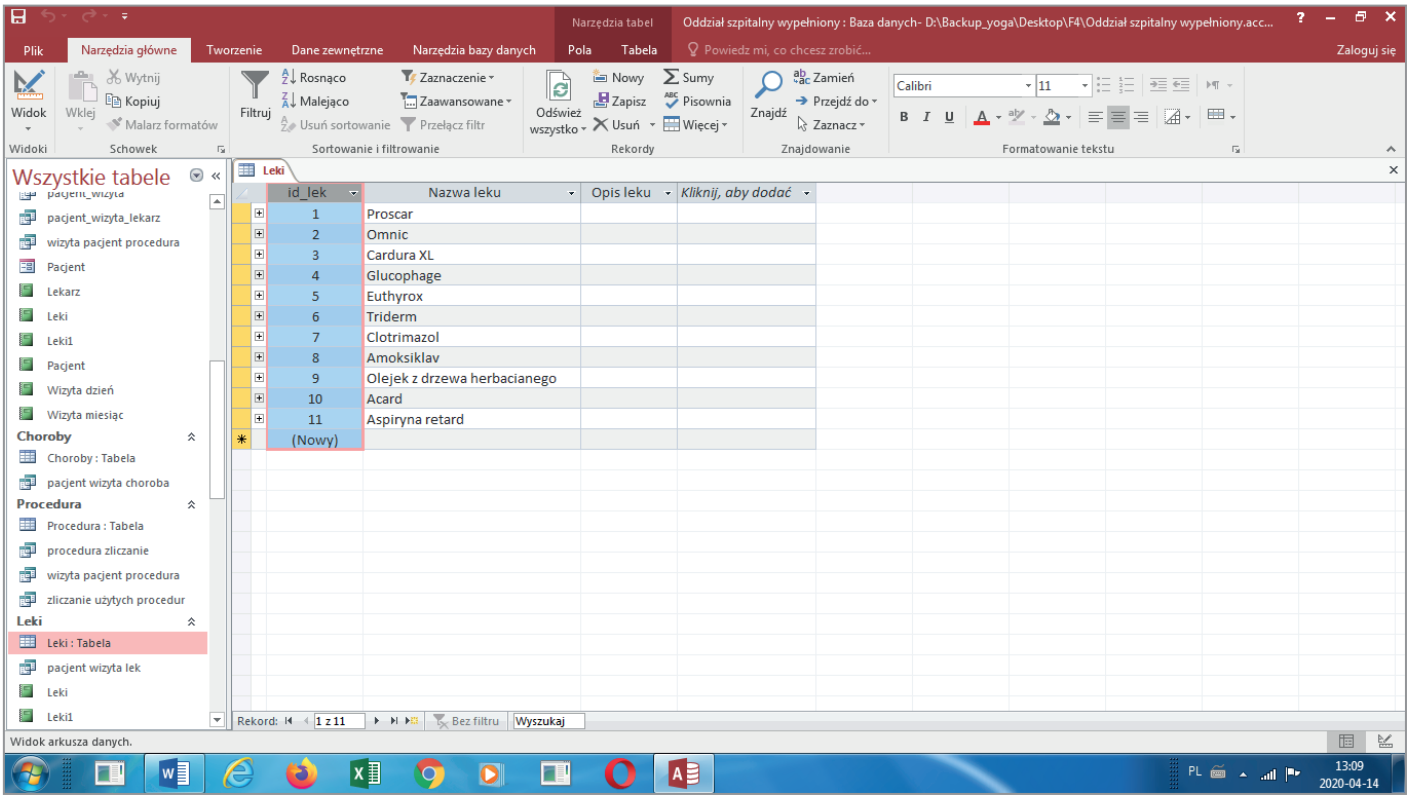
Ryc. 7.51. Uzupełniona tabela „Procedury”

Teraz proszę o zaprojektowanie tabeli „Procedury” na podstawie poprzednich kroków. Uzupełniamy dane, jak widać na poniższej ➥ **ryc. 7.51**.



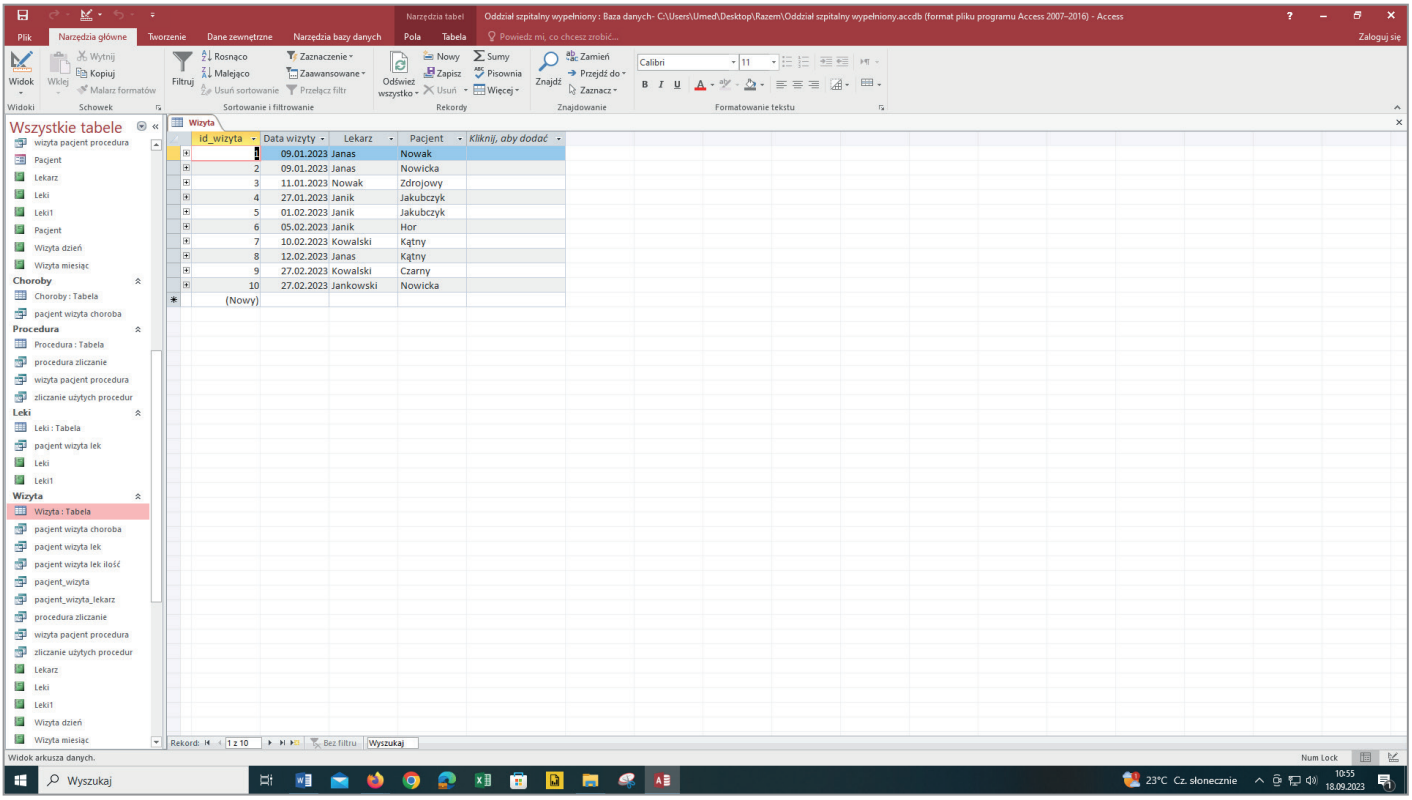
Ryc. 7.52. Uzupełniona tabela „Leki”

Dodatkowo należy utworzyć tabelę „Leki” (➥ **ryc. 7.52**).



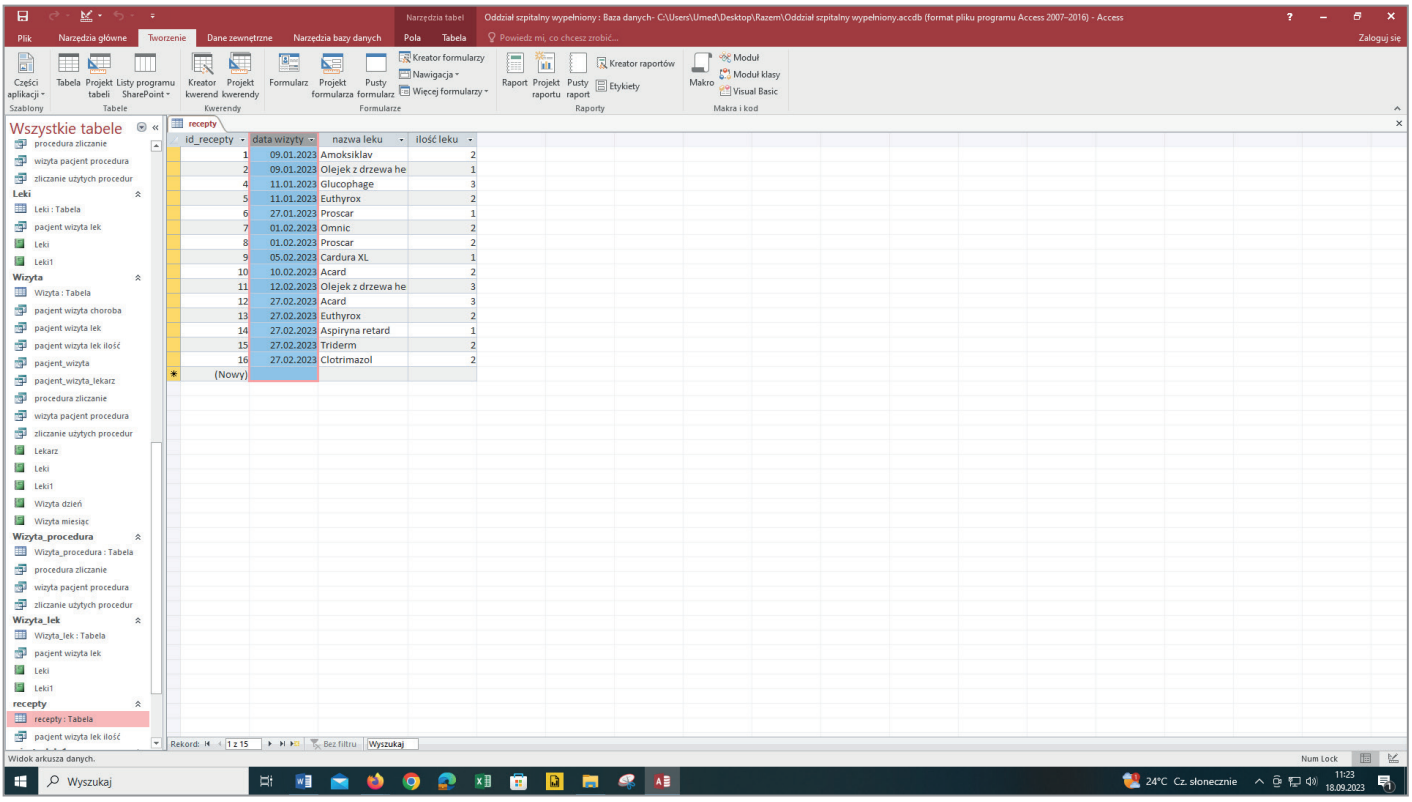
W kolejnej tabeli „Wizyta” w projekcie, analogicznie do wcześniejszych przykładów, używamy kreatora odnośników do tabel „Lekarze” i „Pacjenci”. Uzupełnioną o dane tabelę pokazuje ➥ **ryc. 7.53**.

Ryc. 7.53. Tabela „Wizyta”



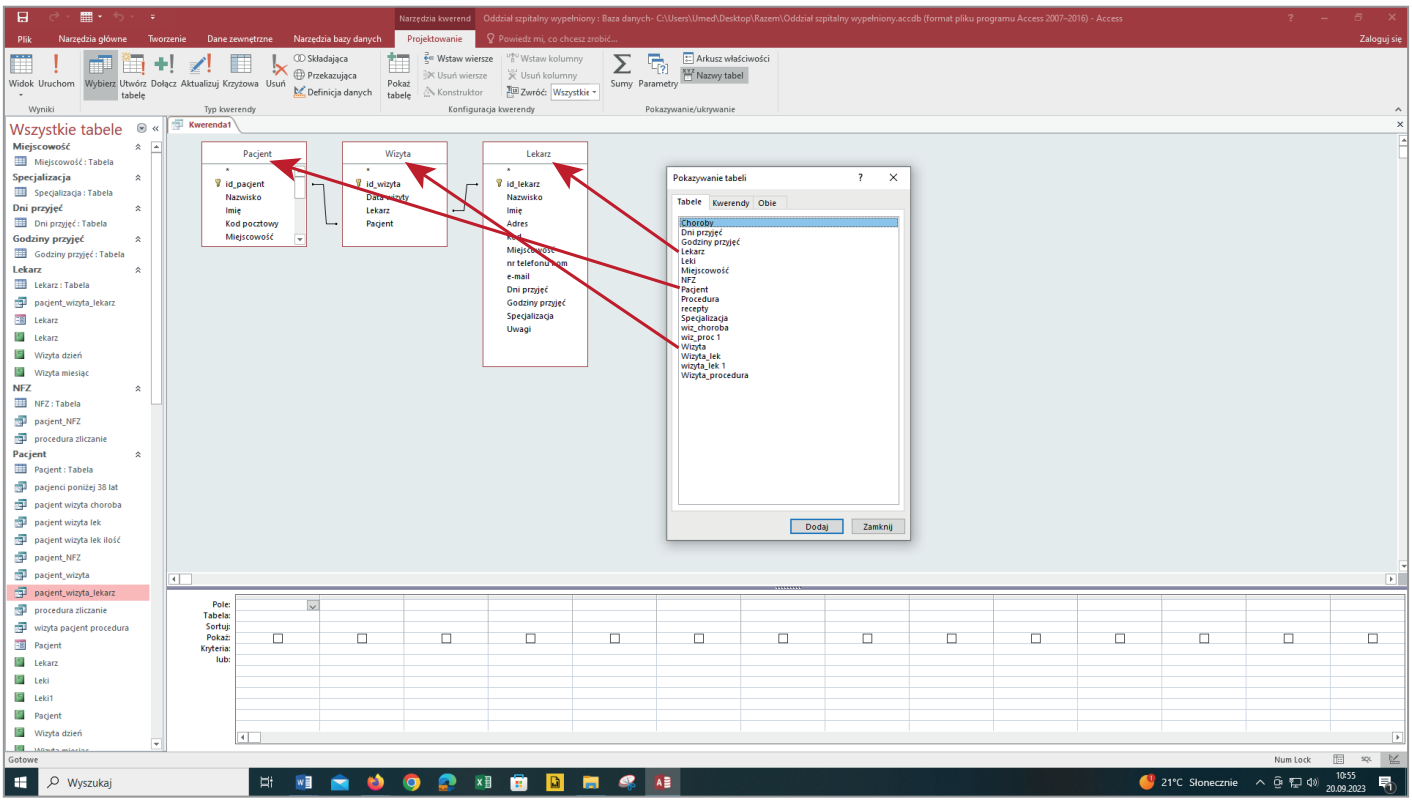
Ryc. 7.54. Tabela „Recepty”

Kolejnym krokiem jest utworzenie tabeli „Recepty” (➥ **ryc. 7.54**).



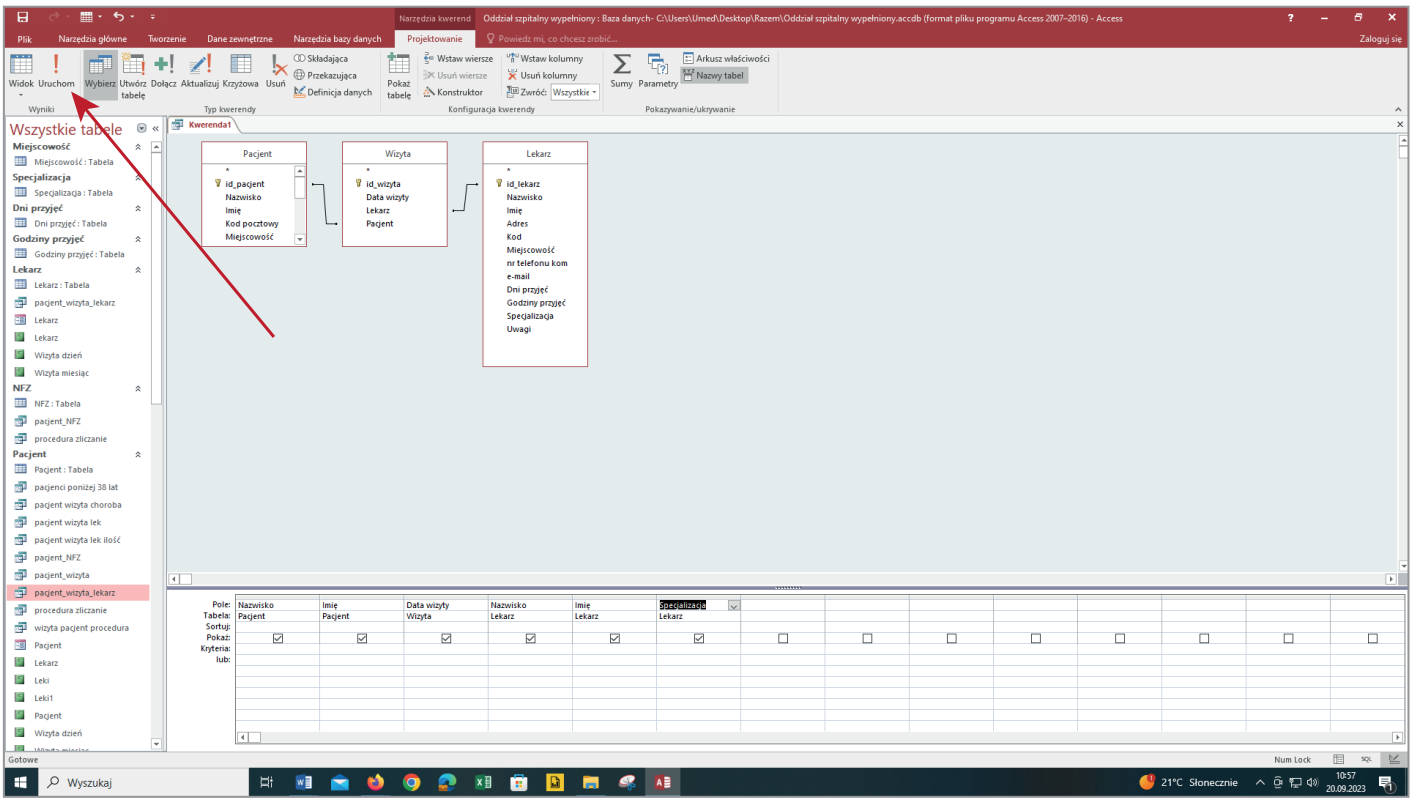
Teraz przystąpimy do tworzenia kwerend, tzn. tabel przedstawiających dane dotyczące interesującego nas zagadnienia. W omówionym przykładzie chcemy się dowiedzieć, który pacjent, w jakim dniu, u którego lekarza miał wizytę. Na wstążce wybieramy kartę Tworzenie, a następnie polecenie Projekt kwerendy. W oknie pojawia się spis tabel, które mamy w naszej bazie. Wybieramy tabele: „Pacjent", „Wizyta", „Lekarz", zaznaczając je kolejno i klikając przycisk Dodaj (➥ **ryc. 7.55**).

Ryc. 7.55. Tworzenie projektu kwerendy „pacjent_wizyta_lekarz"



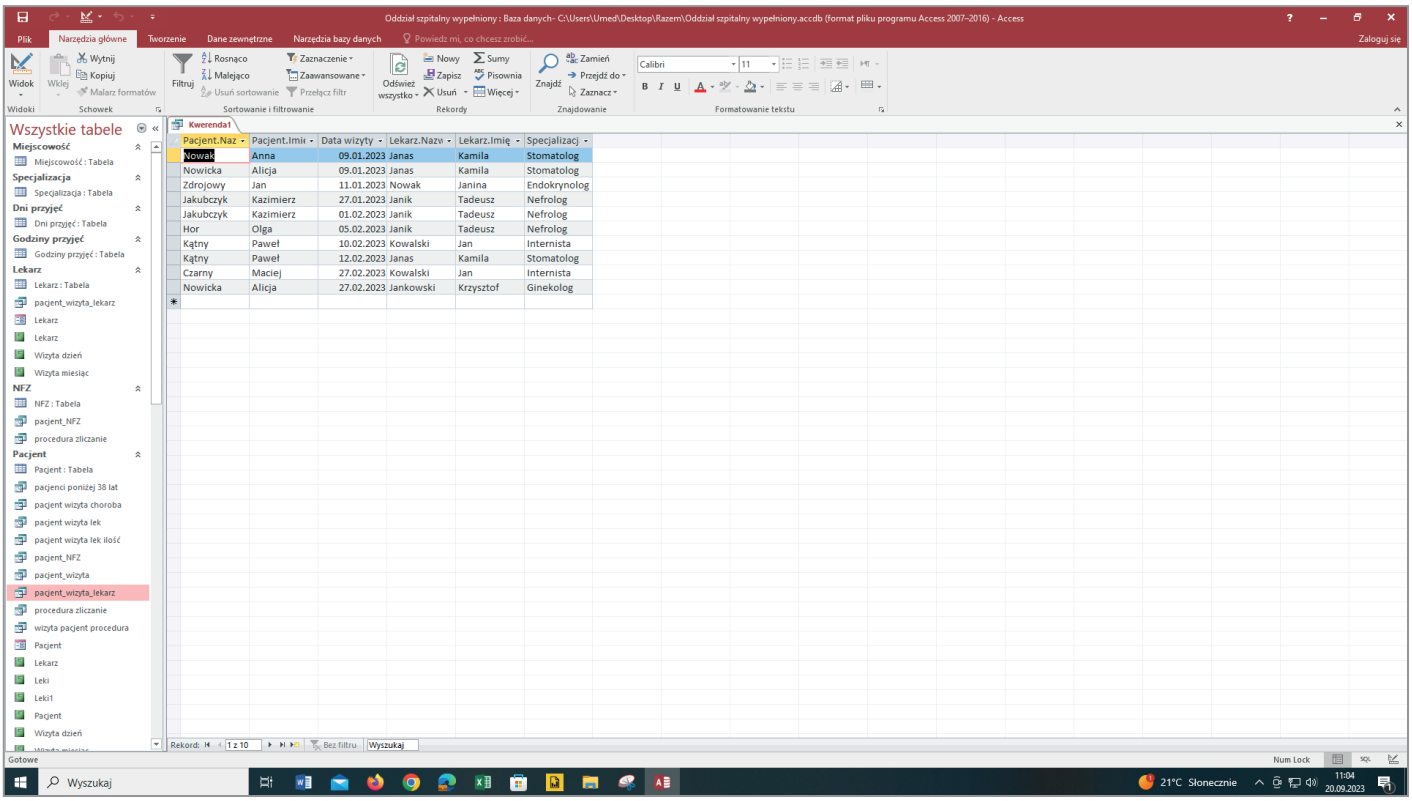
Podwójnym kliknięciem wybieramy „Nazwisko" i „Imię", „Datę wizyty" oraz „Nazwisko", „Imię" i „Specjalizację lekarza", które pojawiają się w tabeli u dołu okna. Klikamy polecenie Uruchom na wstążce (➥ **ryc. 7.56**).

Ryc. 7.56. Uruchamianie kwerendy



Ryc. 7.57. Tabela wyników kwerendy „pacjent_wizyta_lekarz”

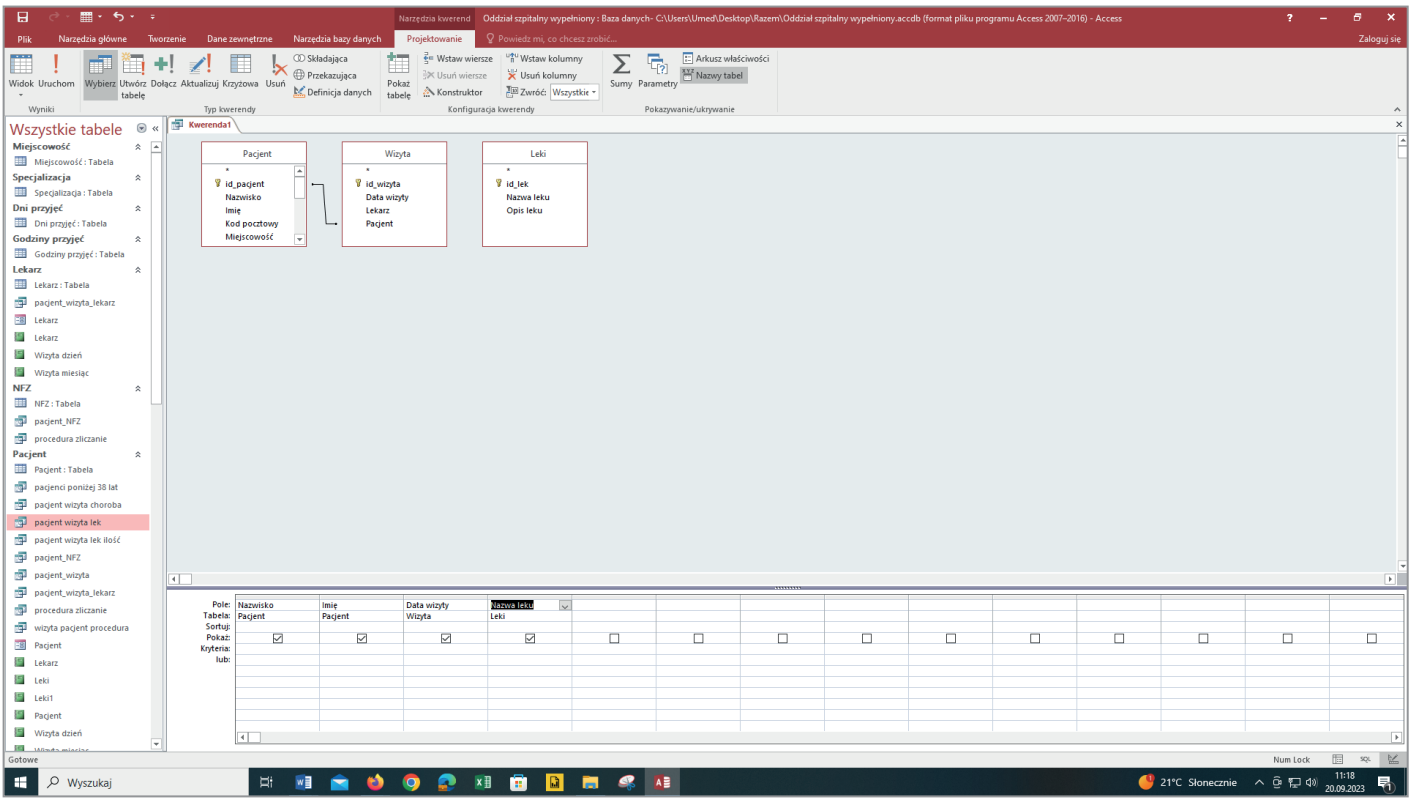
Otrzymujemy tabelę, która przedstawia interesujące nas zagadnienie (➥ **ryc. 7.57**).



Ryc. 7.58. Kwerenda do problemu „jakie leki przepisano pacjentowi na kolejnych wizytach?”

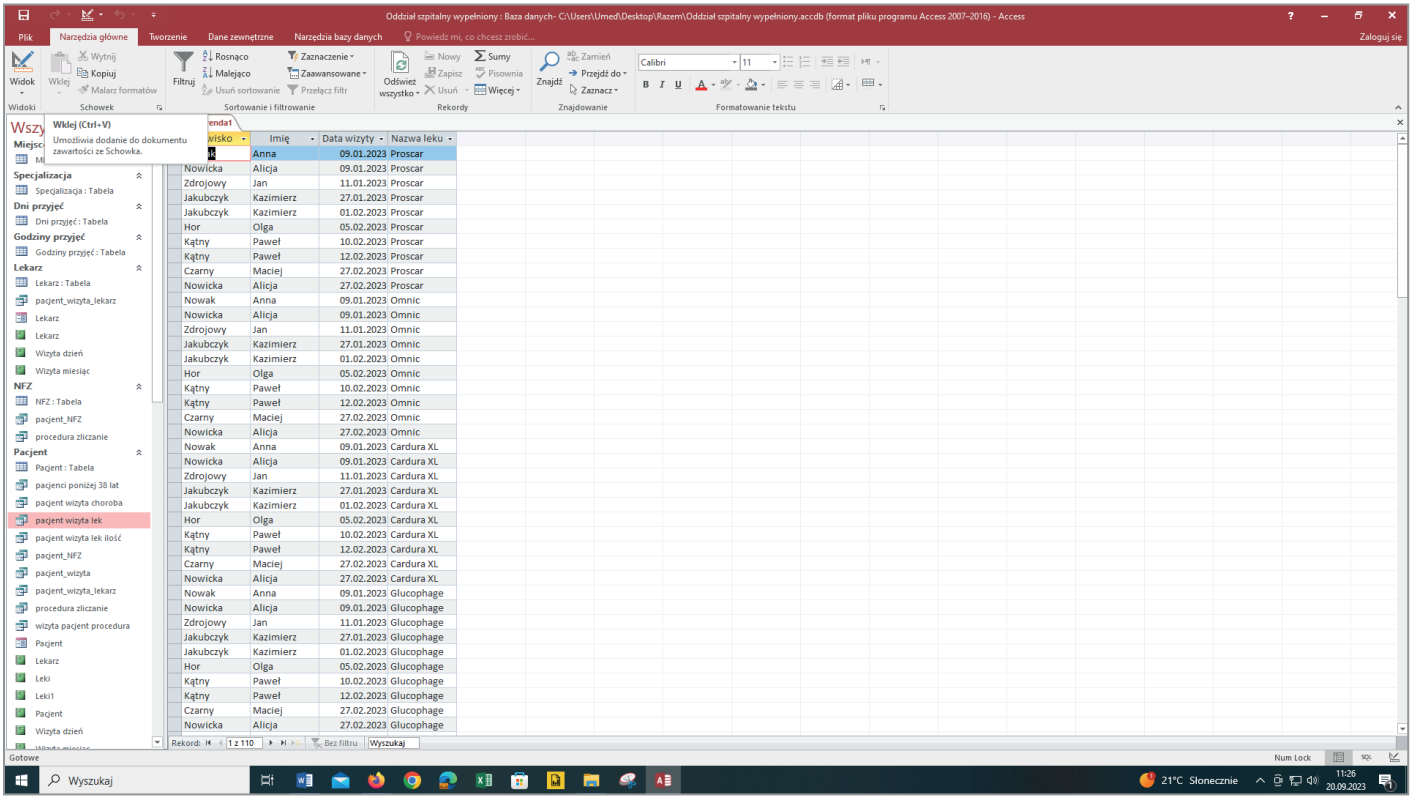
Teraz sformułujmy problem: jakie leki przepisano pacjentowi na kolejnych wizytach?

W przedstawiony powyżej sposób tworzymy nowy projekt kwerendy i wybieramy tabele: „Pacjent”, „Wizyta”, „Leki”. Dwukrotnym kliknięciem wybieramy z poszczególnych tabel dane: „Nazwisko” i „Imię” pacjenta, „Data wizyty” i „Nazwa leku”, które pojawiają się w dole okna (➥ **ryc. 7.58**).



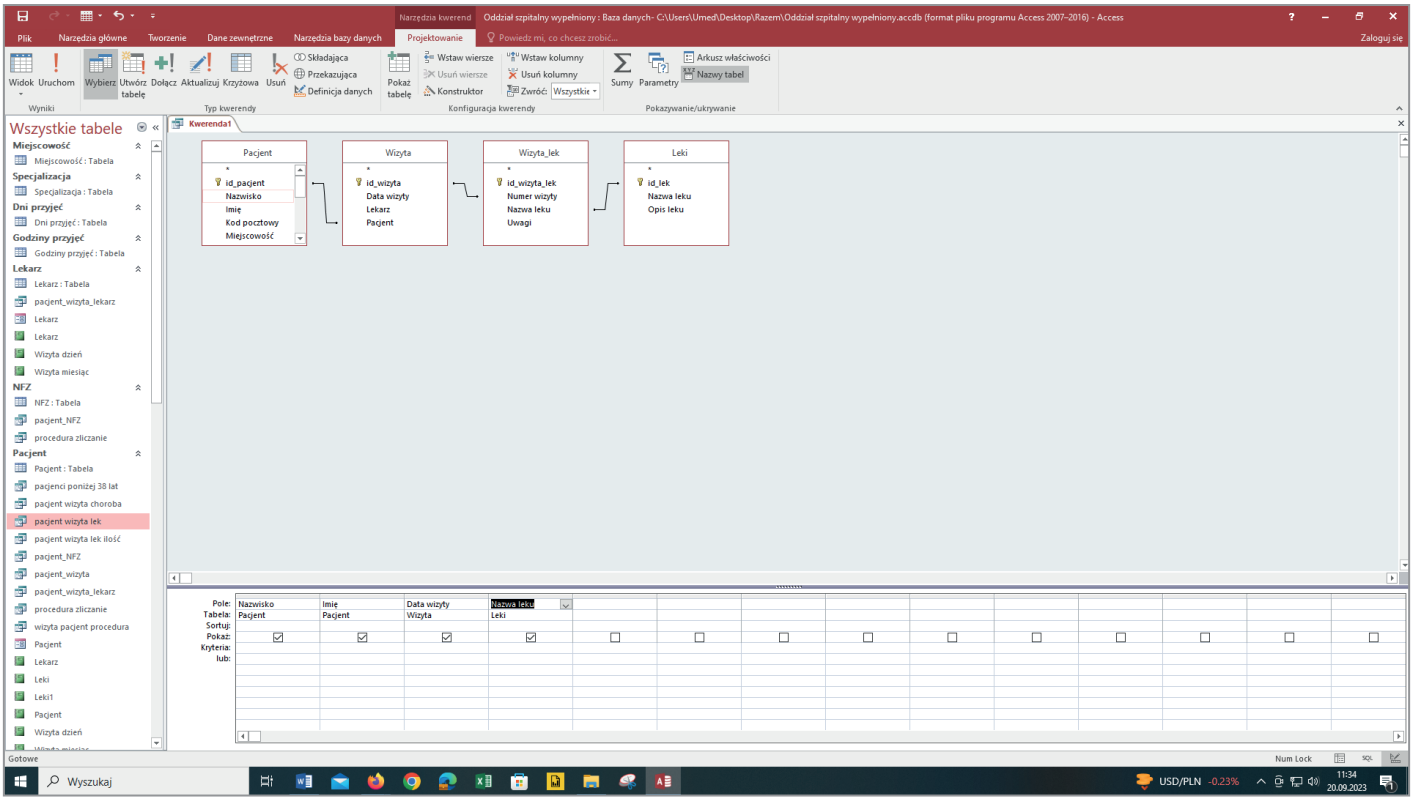
Ryc. 7.59. Efekt działania kwerendy. Wynik jest nieprawidłowy, ponieważ zawiera powtórzenia

Uruchamiamy kwerendę i otrzymujemy wynik przedstawiony na ➥ **ryc. 7.59**.



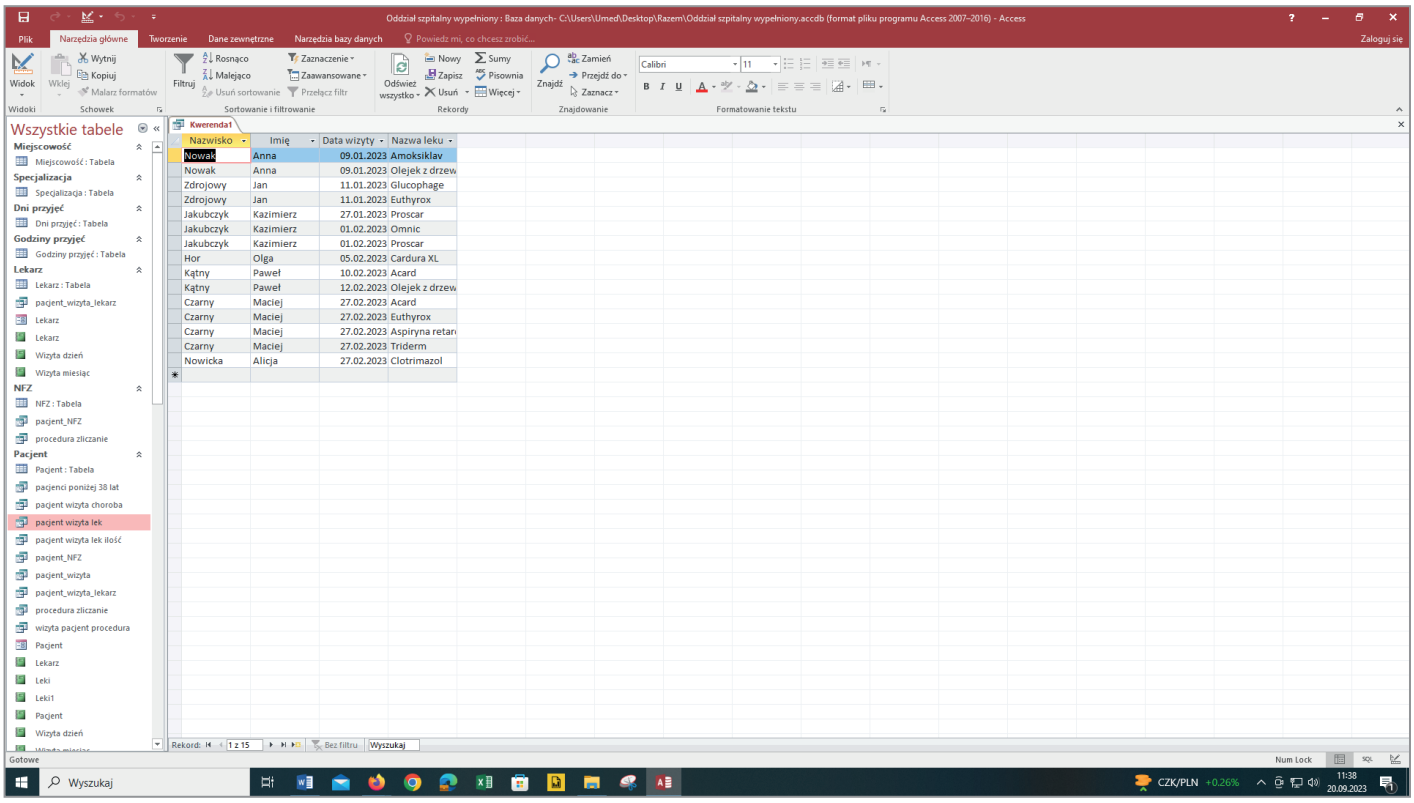
Od razu widzimy, że kwerenda nie działa poprawnie, ponieważ w tabeli występują liczne powtórzenia. Są one wynikiem tego, że tabela „Leki" nie jest powiązana z żadną inną. Zatem musimy jeszcze utworzyć i wprowadzić tzw. tabele łącznikowe (w tym przypadku „wizyta_lek"). Przykład jest zaprezentowany na ➥ **ryc. 7.60**. Wybieramy podwójnymi kliknięciami dane: „Nazwisko" i „Imię" pacjenta, „Data wizyty" i „Nazwa leku" oraz uruchamiamy kwerendę.

Ryc. 7.60. Kwerenda z ➥ **ryc. 7.58**. uzupełniona o tabelę „wizyta_lek"



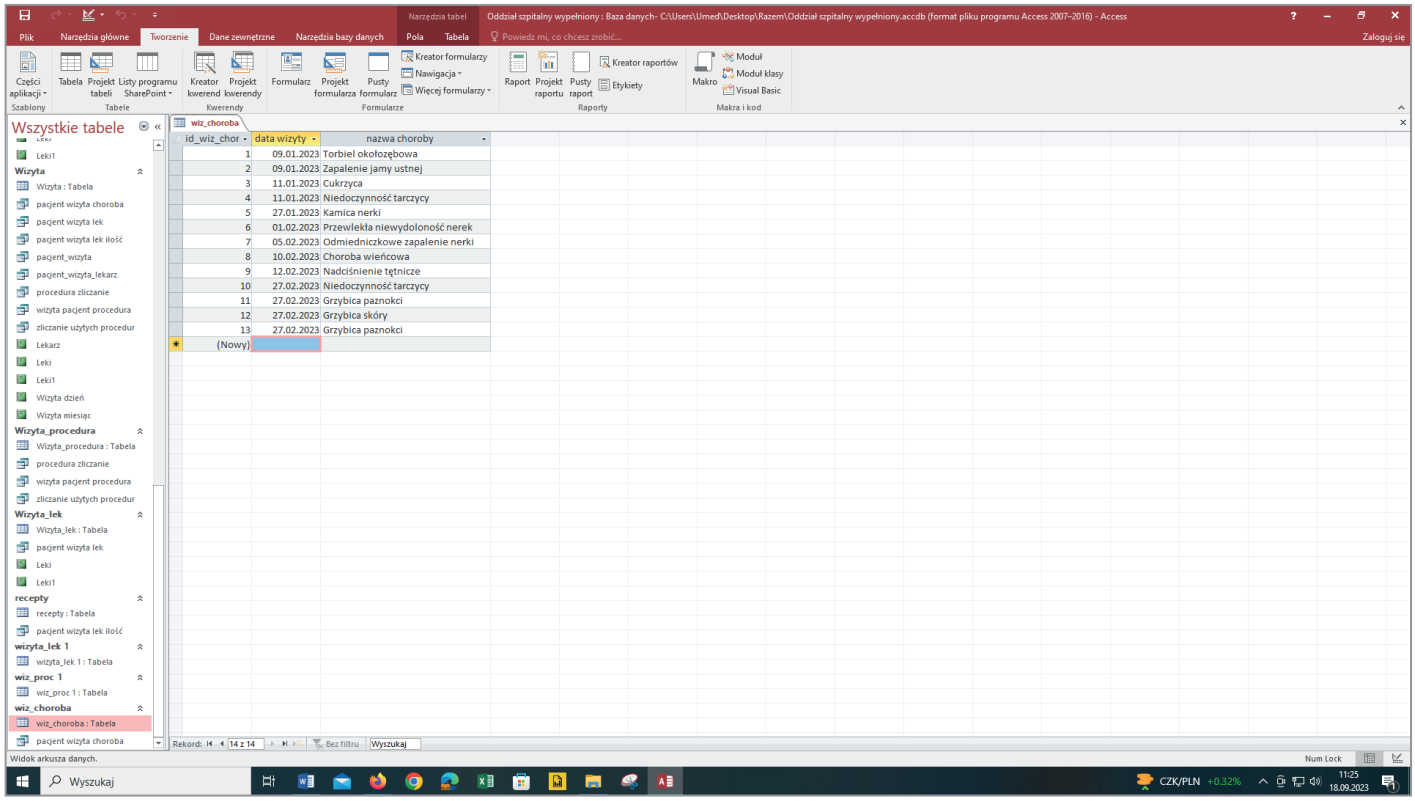
Ryc. 7.61. Prawidłowy wynik działania uzupełnionej kwerendy

Otrzymujemy prawidłowy wynik bez powtórzeń (➥ **ryc. 7.61**).

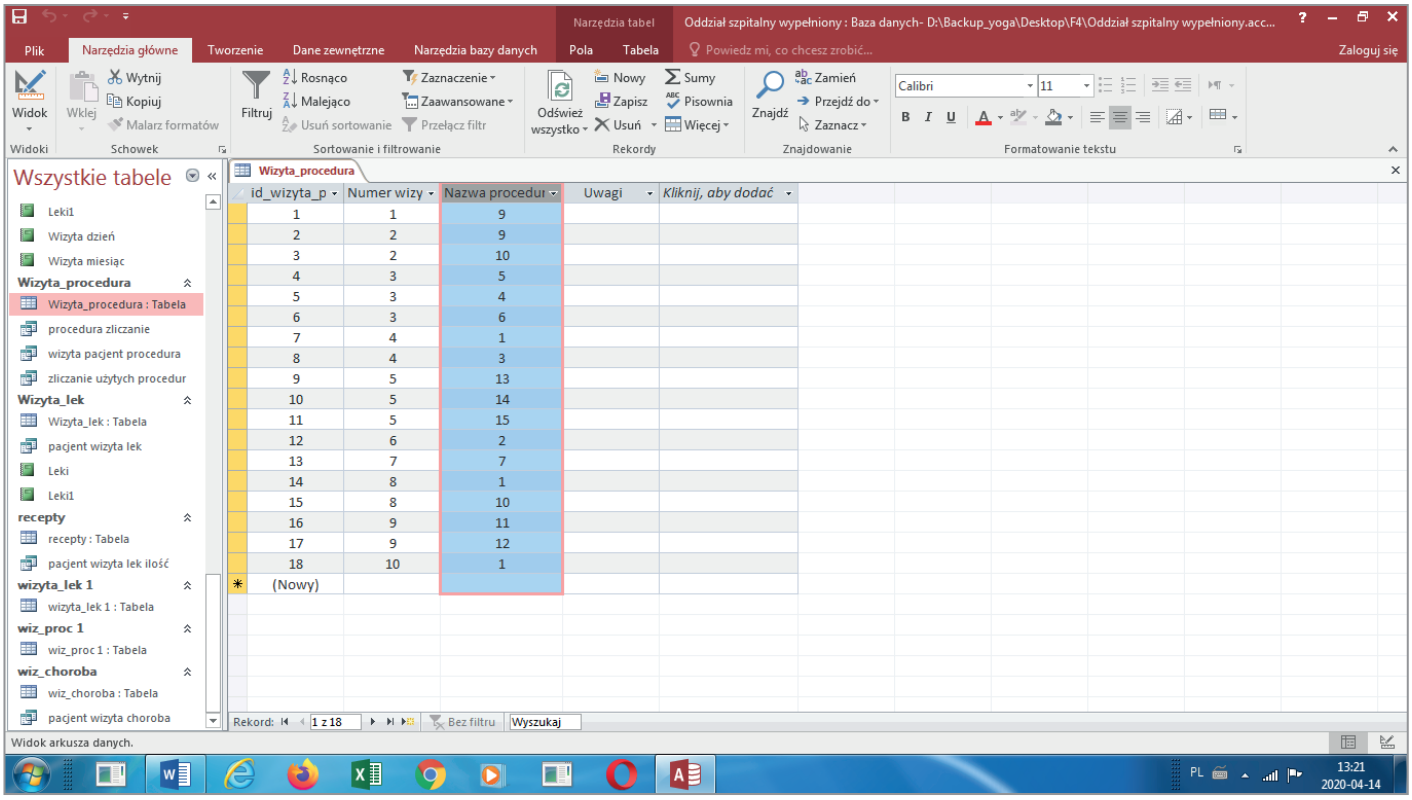


Ryc. 7.62. Tabela łącznikowa „wiz_choroba"

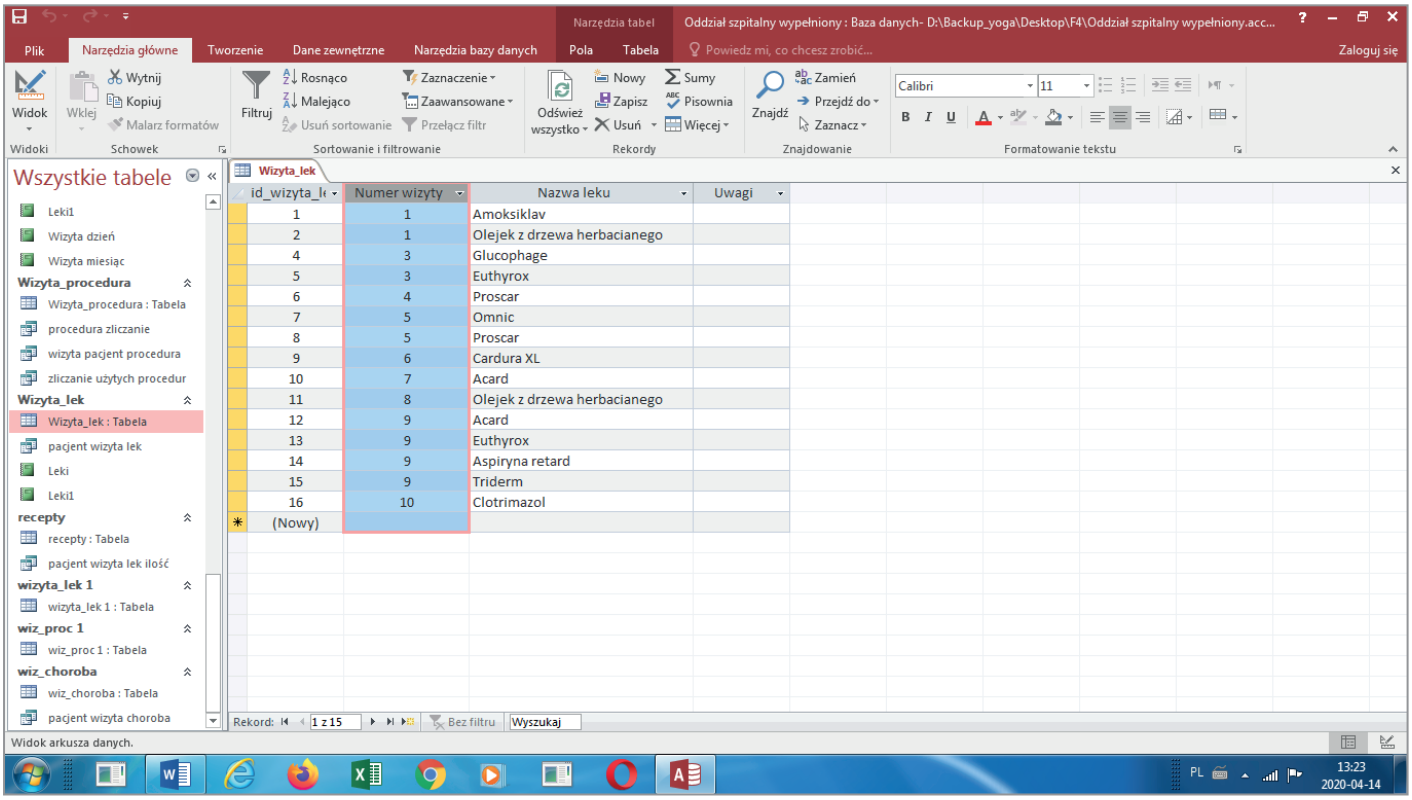
W celu zapewnienia poprawnego funkcjonowania naszej bazy potrzebne będą jeszcze następujące tabele łącznikowe: „wiz_choroba", „Wizyta_procedura", „Wizyta_lek" (➥ **ryc. 7.62 –7.64**)



Ryc. 7.63. Tabela łącznikowa „Wizyta_procedura”

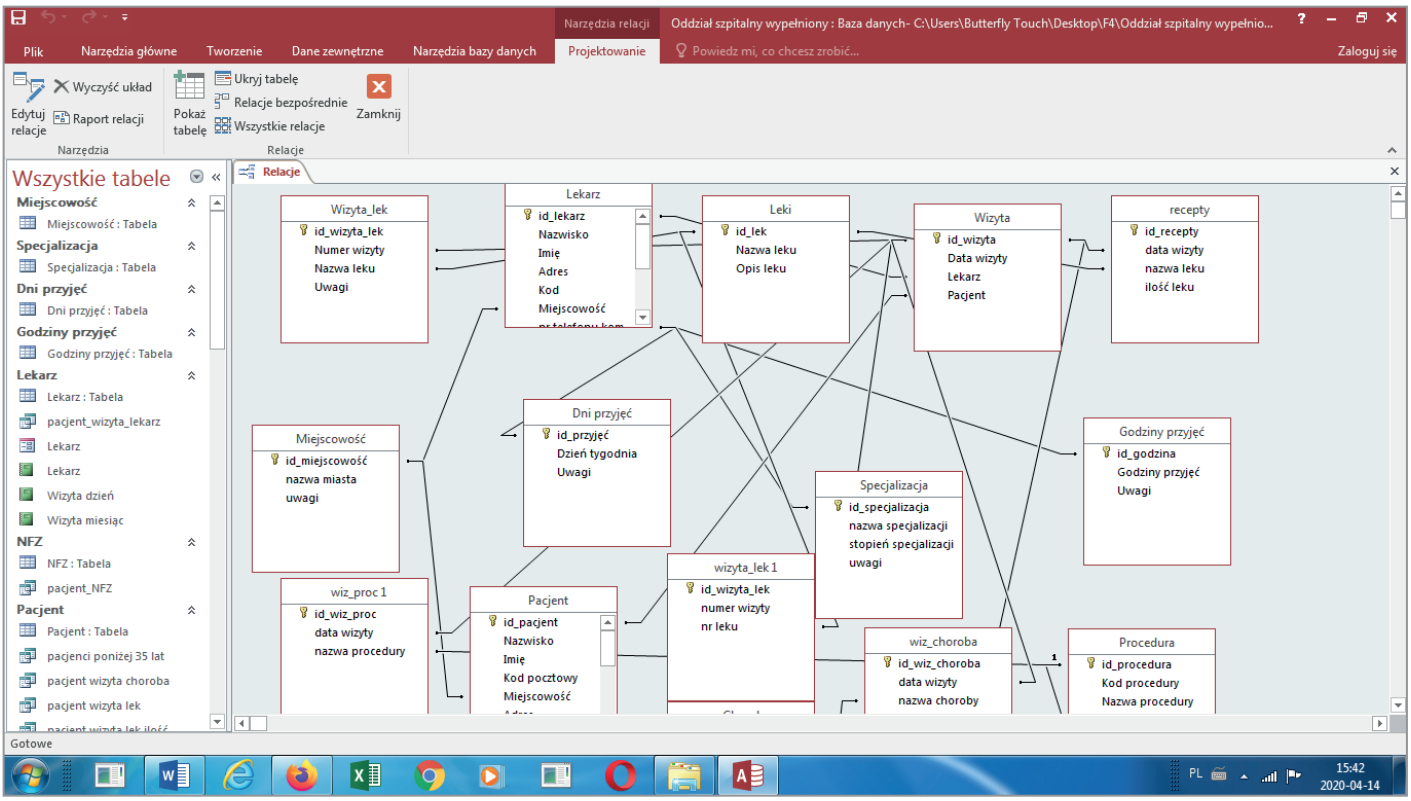


Ryc. 7.64. Tabela łącznikowa „Wizyta_lek”



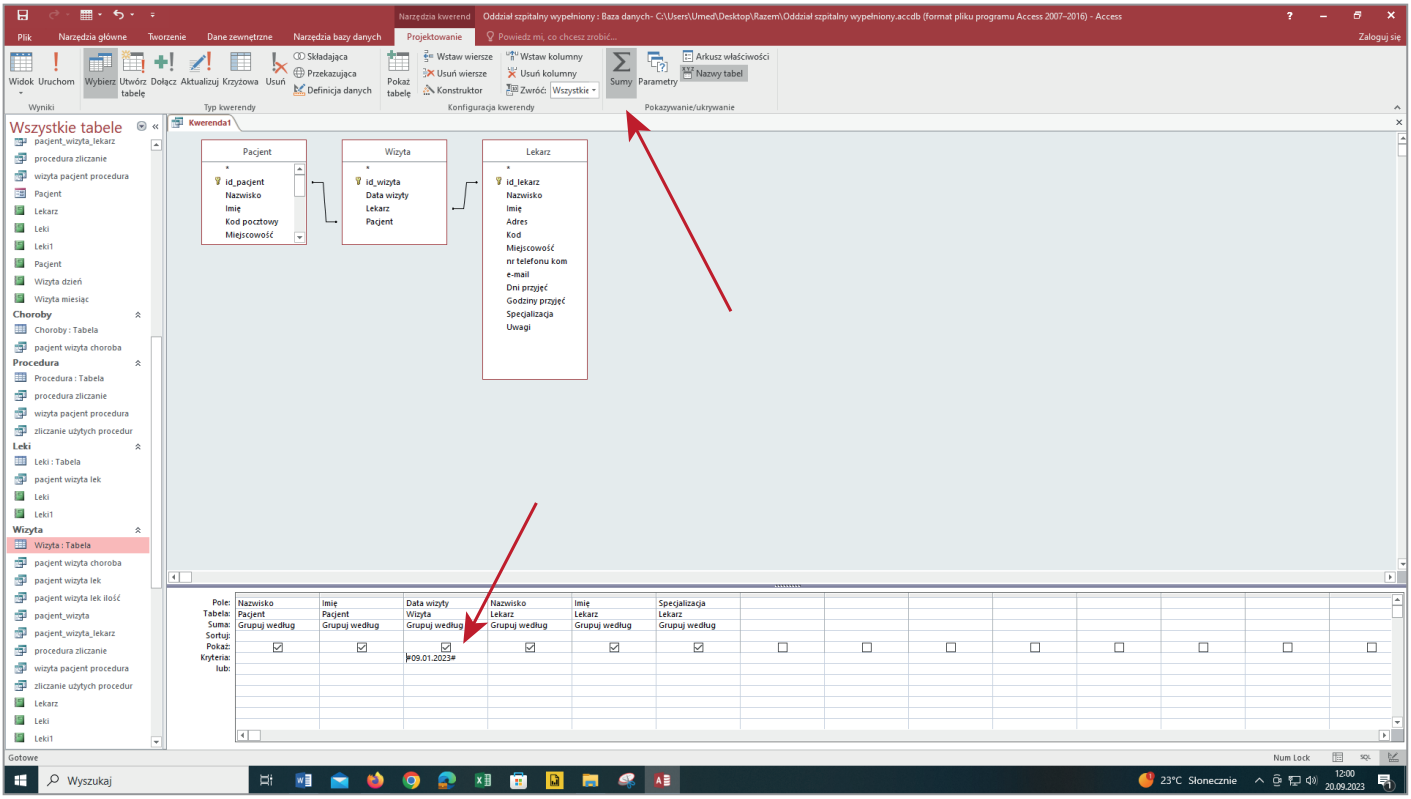
Ryc. 7.65. Schemat relacji dla całej bazy

Po utworzeniu wyżej wymienionych tabel otrzymujemy schemat relacji między wszystkimi tabelami w bazie przedstawiony w poniższym oknie (➥ **ryc. 7.65**).



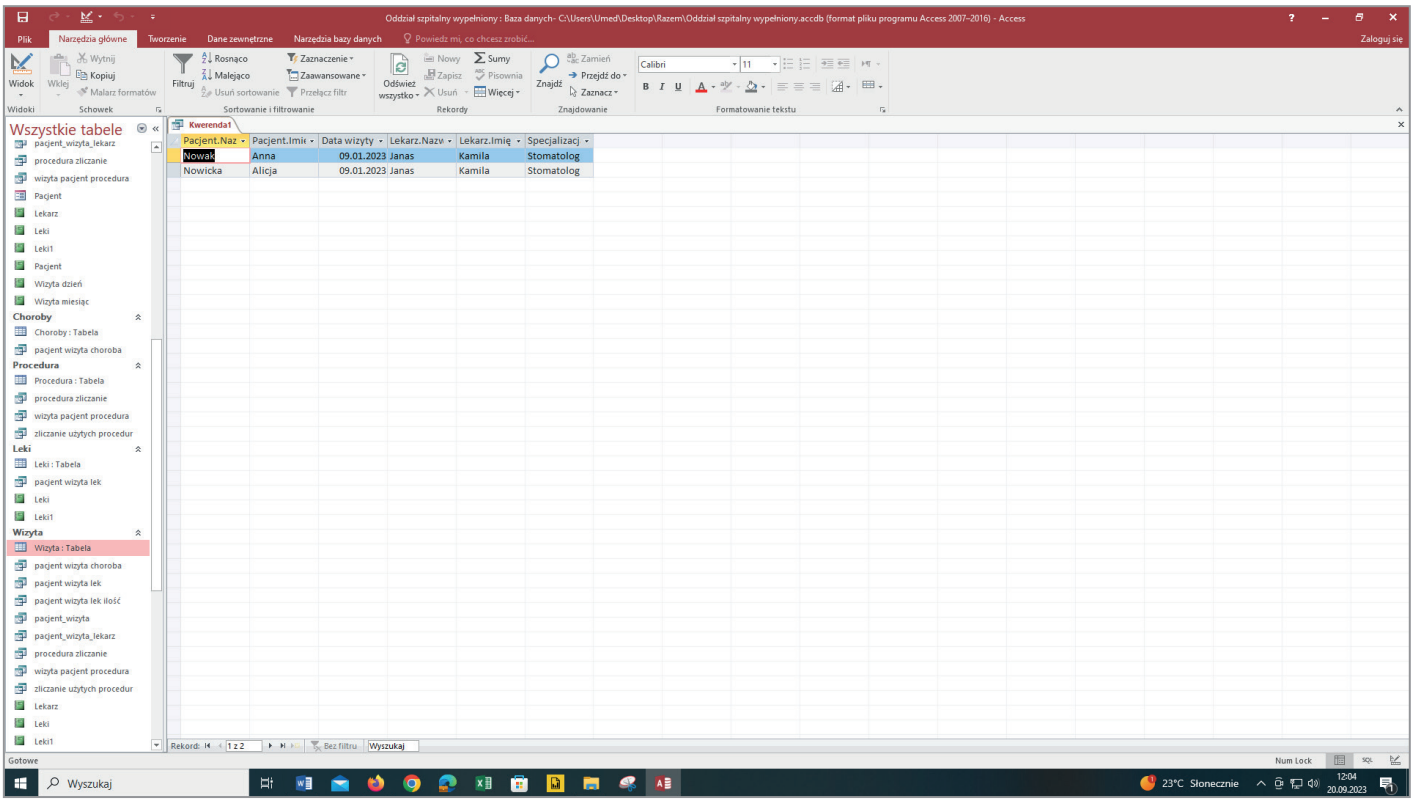
Ryc. 7.66. Projekt kwerendy do pytania „którzy lekarze przyjmowali w dniu 9.01.2023?"

Utworzymy jeszcze przykładową kwerendę odpowiadającą na pytanie, którzy lekarze przyjmowali w dniu 9.01.2023 roku. Należy w tym celu wybrać w projekcie tabele, jak pokazuje ➥ **ryc. 7.66**.



W pozycji „Kryteria” (w dolnej tabeli) klikamy na wstążce polecenie Sumy z karty Projektowanie i wpisujemy do komórki datę 09.01.2023. Wynik kwerendy przedstawiono na ➥ **ryc. 7.67**.

Ryc. 7.67. Wynik powyższej kwerendy



Na podstawie przedstawionych przykładów możemy tworzyć dowolne zapytania i odpowiadać na nie, projektując odpowiednie kwerendy. Nie powinno to już sprawiać żadnych trudności.

Raporty tworzymy do wszystkich baz jednakowo na podstawie schematów pokazanych na ➥ ryc. 7.16–ryc. 7.20.

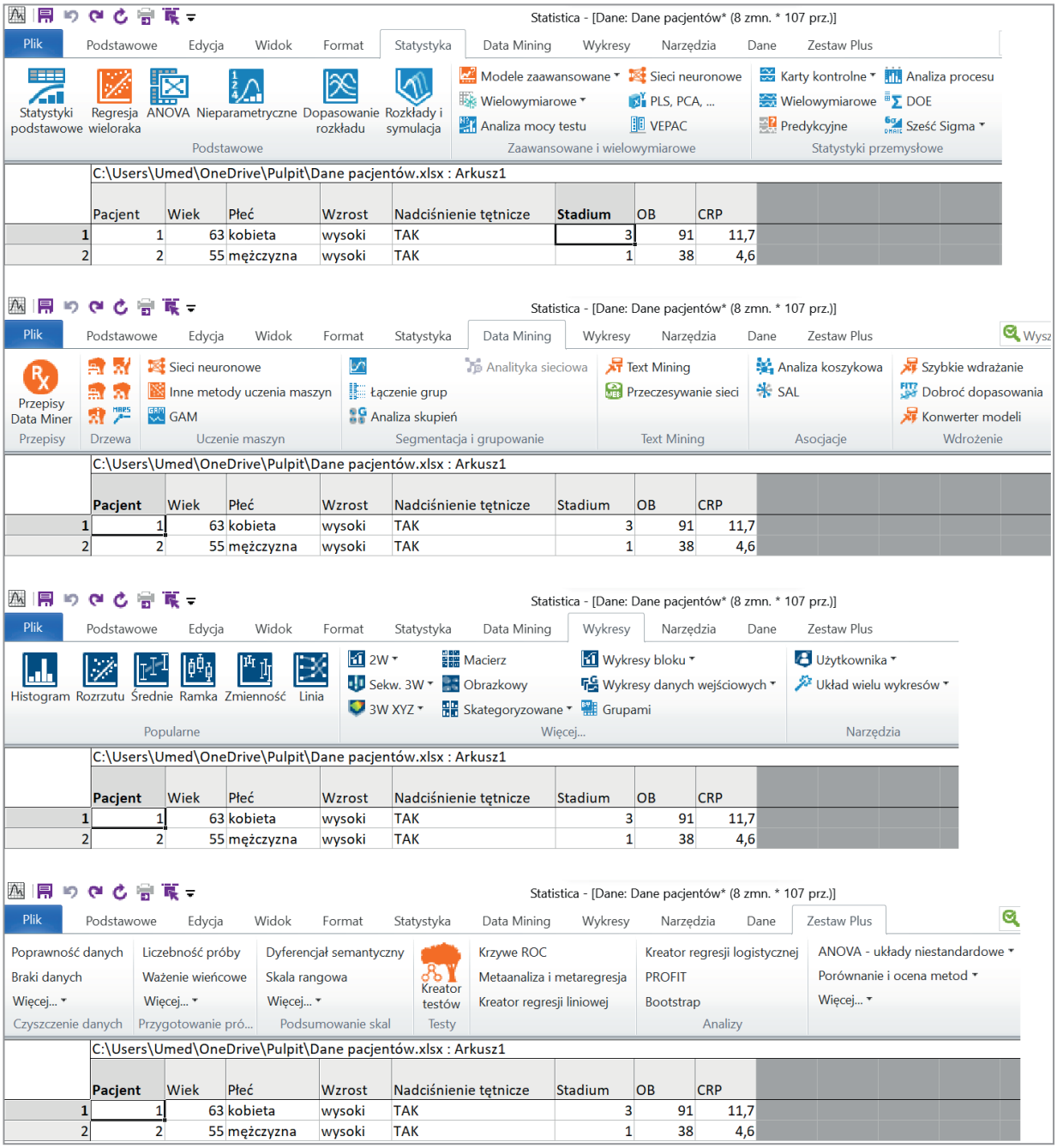
Tomasz Hikawczuk

8 Wprowadzenie do programu Statistica

Program Statistica umożliwia przeprowadzanie analizy statystycznej za pomocą testów parametrycznych lub nieparametrycznych w zależności od przygotowanego wcześniej układu doświadczalnego. Pozwala również na analizę na danych retrospektywnych zebranych w ramach quasi-eksperymentu. Są to dane obserwacyjne, które można porównywać za pomocą statystyki opisowej lub analizować, np. przy użyciu korelacji bez wcześniejszego planowania eksperymentu i randomizacji układu doświadczalnego. Wadą tego typu porównania jest niższa dokładność wyników w porównaniu z zaplanowanymi uprzednio układami doświadczalnymi. Jednak w tego typu analizie można uzyskać pewne wnioski również przy zastosowaniu technik *data mining*, które następnie mogą być weryfikowane za pomocą zaplanowanych układów randomizowanych.

Przy rozpatrywaniu tematyki testów statystycznych główną zakładką, z której wybierane będą odpowiednie testy, jest Statystyka (➥ **ryc. 8.1**). Jednak w trakcie pracy z danymi niezbędne jest ich wczytywanie i tworzenie nowych arkuszy (Plik), przygotowywanie wykresów (Wykresy), wyszukiwanie prawidłowości w bazach danych (Data Mining) czy też zastosowanie testów statystycznych charakterystycznych dla badań medycznych (Zestaw Plus).

Zbiory danych można poddać analizie pod kątem różnorakich kryteriów, uzależnionych oczywiście od wcześniej przyjętego układu doświadczalnego lub na zasadzie quasi-eksperymentu bez wcześniejszej randomizacji danych, opierając się na statystykach opisowych i/lub metodach korelacji. Quasi-eksperyment nie spełnia istotnych kryteriów formalnych z punktu widzenia planu doświadczenia, jednak istnieje możliwość znalezienia struktury w przypadku danych odpowiadającej eksperymentowi kontrolowanemu i uwzględnieniu wyników za pomocą statystyk opisowych lub korelacji lub też porównania liczebności obserwowanych i oczekiwanych w przypadku zastosowania testu χ^2. W zakładce Statystyka zgromadzone są różnego rodzaju testy parametryczne i nieparametryczne, które można zastosować do porównywania zmiennych i analizy frekwencji obserwacji np. w tabelach 2×2. Szczegółowo zostaną one opisane w dalszej części skryptu. Zakładka Data Mining pozwala na przeprowadzenie analizy danych na zasadzie drzew klasyfikacyjnych, analizy skupień, łączenie w grupy w zależności od określonych kryteriów, przeprowadzenie analizy koszykowej, zastosowanie mechanizmów uczenia maszynowego, tworzenie sieci neuronowych oraz szybkie tworzenie modeli zjawisk w przypadku danych bez wcześniejszego planowania doświadczeń na podstawie określania tendencji występowania określonych procesów. Zakładka Wykresy pozwala m.in. na porównanie zmiennych za pomocą wykresów, m.in. histogramów (opisujących frekwencję danych w przedziałach), tworzenie wykresów rozrzutu i dopasowanie równań regresji do danych z próby, porównaniu średnich/median (wykres ramka-wąsy). Natomiast w części Zestaw Plus znajdują się narzędzia i testy znajdujące zastosowanie w badaniach medycznych z zakresu metaanalizy, liczebności próby, regresji liniowej, krzywych ROC (krzywa charakterystyki operacyjnej odbiornika, ang. *receiver operating characteristic curve*) itp., pozwalających na zebranie, uporządkowanie i analizę danych na podstawie testów i wykresów specyficznych dla badań medycznych.



Ryc. 8.1. Główne cele zastosowania programu Statistica w analizie danych statystycznych

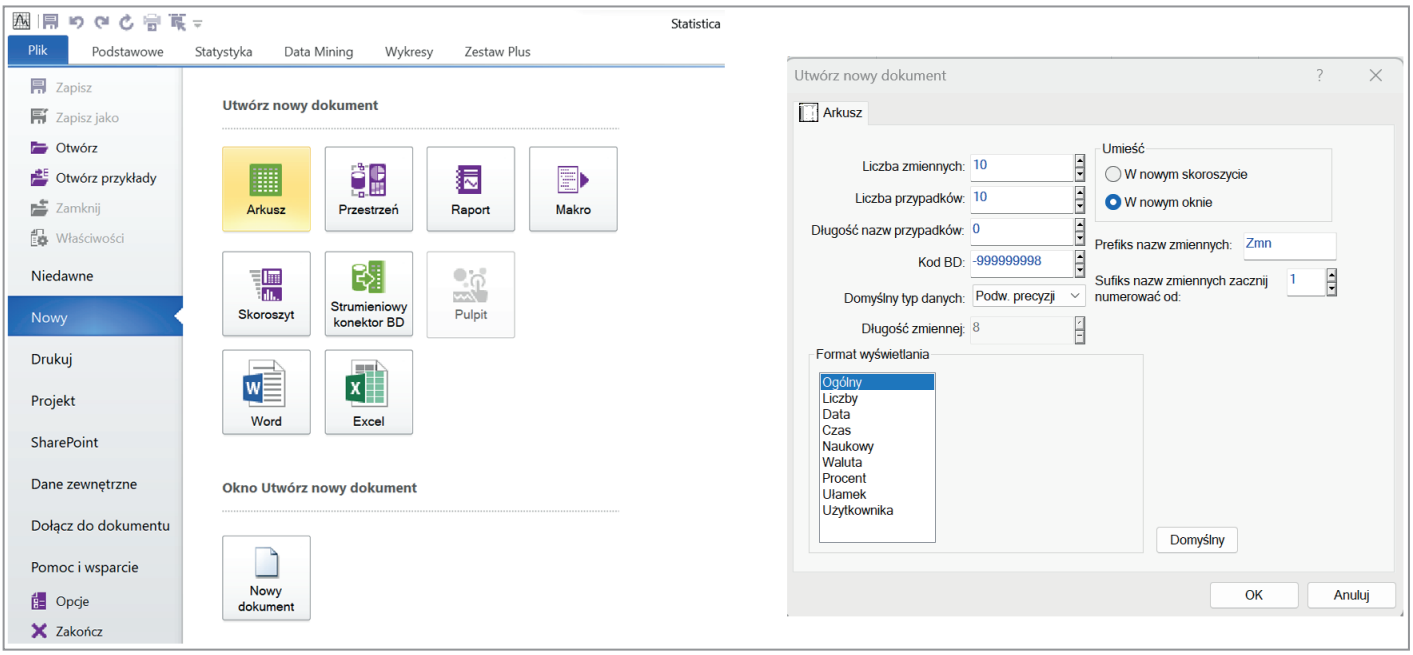
8.1. Tworzenie nowego arkusza w programie Statistica

Nowy arkusz można przygotować na 2 sposoby: wkleić wcześniej przygotowane do analizy dane (kopiuj-wklej) lub otworzyć gotowy arkusz zapisany w Excelu.

Kiedy dysponujemy danymi zapisanymi w pliku Excela („xls", „xlsx"), tabeli w pliku tekstowym Worda („doc", „docx") bądź jakimkolwiek innym formacie umożliwiającym bezproblemowe przeniesienie danych poprzez skopiowanie, a następnie wklejenie do programu statystycznego, można na samym początku przed przystąpieniem do analizy utworzyć nowy arkusz o określonych rozmiarach w zależności od potrzeb. W tym celu wybieramy zakładkę Plik, a następnie klikamy w ikonkę Nowy,

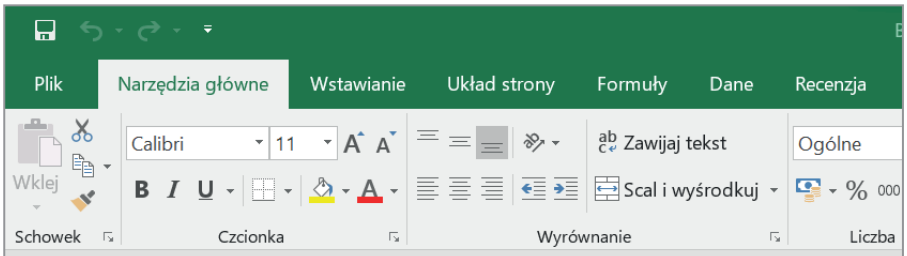
Ryc. 8.2. Tworzenie nowego arkusza w programie Statistica

i kolejno – Arkusz. Po kliknięciu ikony Arkusz pojawia się okienko pozwalające na określenie cech arkusza, w tym np. jego wielkości (➥ **ryc. 8.2**).



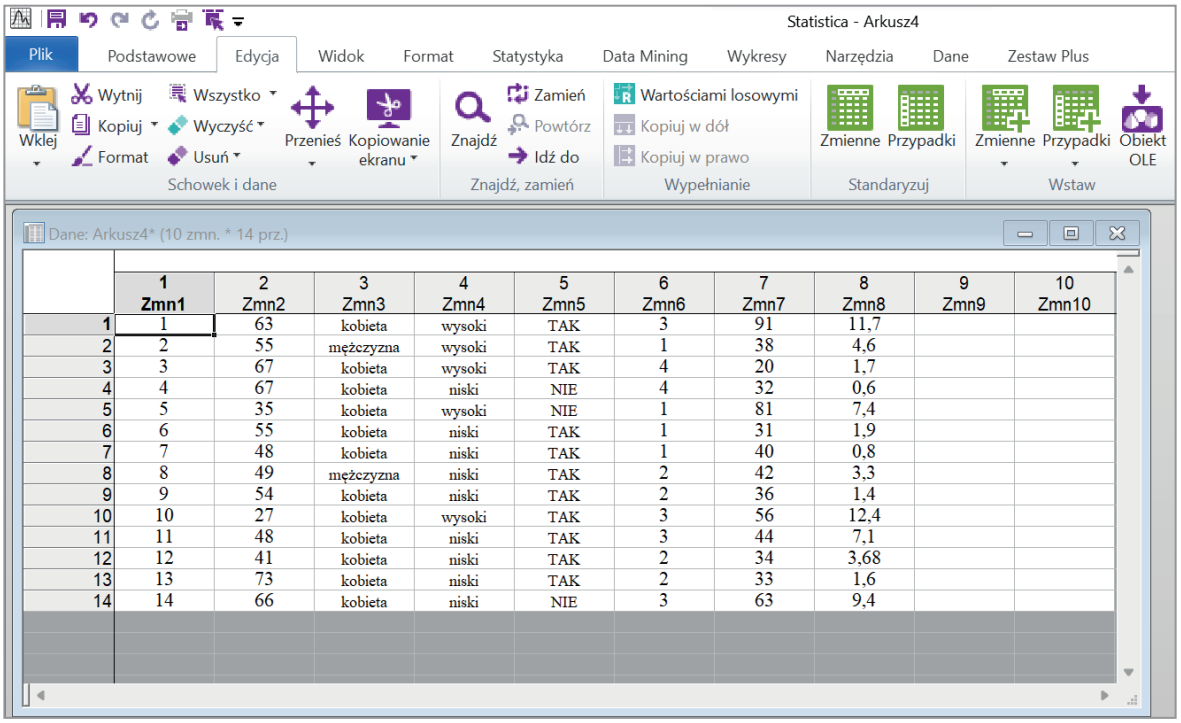
Dane można w prosty sposób skopiować z arkusza kalkulacyjnego i wkleić do programu Statistica (➥ **ryc. 8.3**).

Ryc. 8.3. Dane 14 pacjentów z danej kliniki/przychodni w arkuszu Excela



	A	B	C	D	E	F	G	H
1	**Pacjent**	**Wiek**	**Płeć**	**Wzrost**	**Nadciśnienie tętnicze**	**Stadium**	**OB**	**CRP**
2	1	63	kobieta	wysoki	TAK	3	91	11,7
3	2	55	mężczyzna	wysoki	TAK	1	38	4,6
4	3	67	kobieta	wysoki	TAK	4	20	1,7
5	4	67	kobieta	niski	NIE	4	32	0,6
6	5	35	kobieta	wysoki	NIE	1	81	7,4
7	6	55	kobieta	niski	TAK	1	31	1,9
8	7	48	kobieta	niski	TAK	1	40	0,8
9	8	49	mężczyzna	niski	TAK	2	42	3,3
10	9	54	kobieta	niski	TAK	2	36	1,4
11	10	27	kobieta	wysoki	TAK	3	56	12,4
12	11	48	kobieta	niski	TAK	3	44	7,1
13	12	41	kobieta	niski	TAK	2	34	3,68
14	13	73	kobieta	niski	TAK	2	33	1,6
15	14	66	kobieta	niski	NIE	3	63	9,4

Dodatkowo, jeśli wklejamy dane, można wstępnie ustawić rozmiar arkusza na 10 × 10 komórek, natomiast w trakcie przeklejania wartości zmiennych program sam dostosuje jego wielkość. W efekcie w programie uzyskujemy dane, na których możliwe jest już przeprowadzanie obliczeń (➥ **ryc. 8.4**).

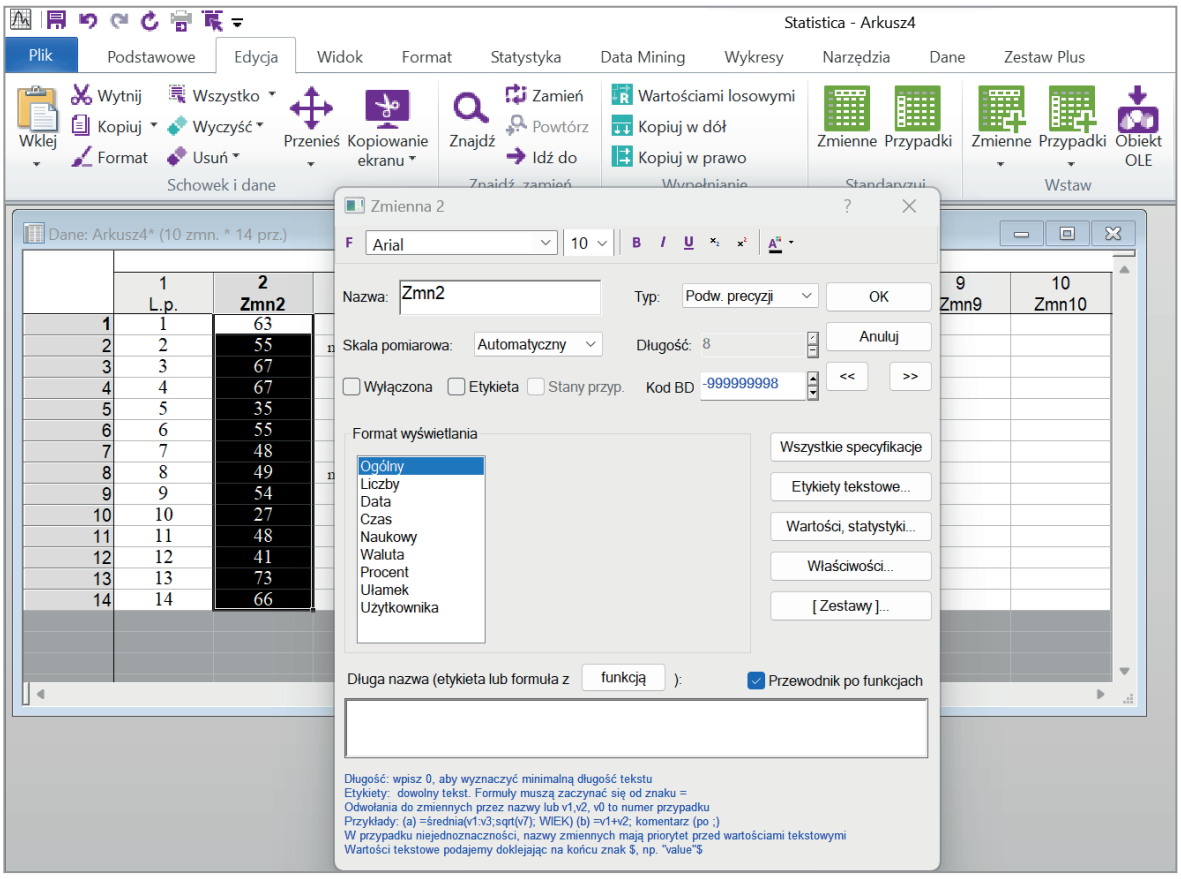


	1 Zmn1	2 Zmn2	3 Zmn3	4 Zmn4	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	1	63	kobieta	wysoki	TAK	3	91	11,7		
2	2	55	mężczyzna	wysoki	TAK	1	38	4,6		
3	3	67	kobieta	wysoki	TAK	4	20	1,7		
4	4	67	kobieta	niski	NIE	4	32	0,6		
5	5	35	kobieta	wysoki	NIE	1	81	7,4		
6	6	55	kobieta	niski	TAK	1	31	1,9		
7	7	48	kobieta	niski	TAK	1	40	0,8		
8	8	49	mężczyzna	niski	TAK	2	42	3,3		
9	9	54	kobieta	niski	TAK	2	36	1,4		
10	10	27	kobieta	wysoki	TAK	3	56	12,4		
11	11	48	kobieta	niski	TAK	3	44	7,1		
12	12	41	kobieta	niski	TAK	2	34	3,68		
13	13	73	kobieta	niski	TAK	2	33	1,6		
14	14	66	kobieta	niski	NIE	3	63	9,4		

Ryc. 8.4. Dane po wklejeniu do programu Statistica (wielkość arkusza dostosowała się automatycznie)

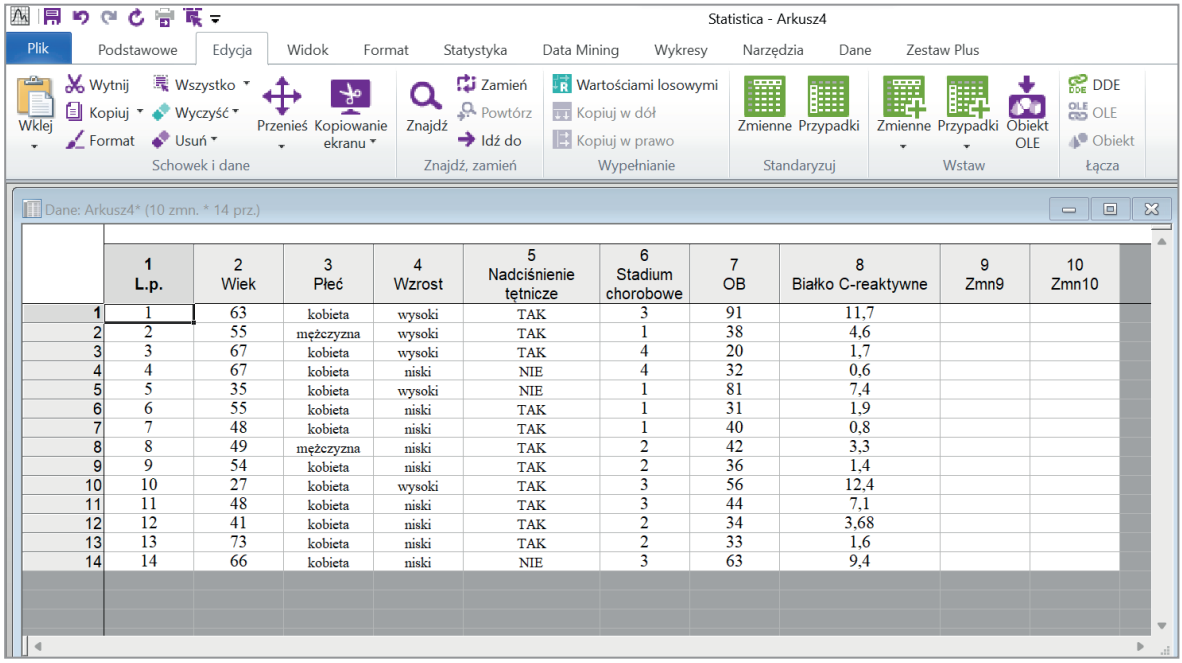
Czasem w czasie kopiowania wyników z arkusza kalkulacyjnego zmienne tekstowe wymagają utworzenia tzw. etykiet w celu łatwiejszego dla czytelnika prezentowania zmiennych jakościowych.

Dodatkowo możliwa jest również zmiana nagłówków zmiennych. W przypadku dużej ich liczby znacznie ułatwia to wybór z listy. Po dwukrotnym kliknięciu na Zmn2, otworzy się okno, w którym zamiast jej nazwy można wpisać Wiek (➥ **ryc. 8.5**).



Ryc. 8.5. Zmiana nazwy zmiennej w arkuszu

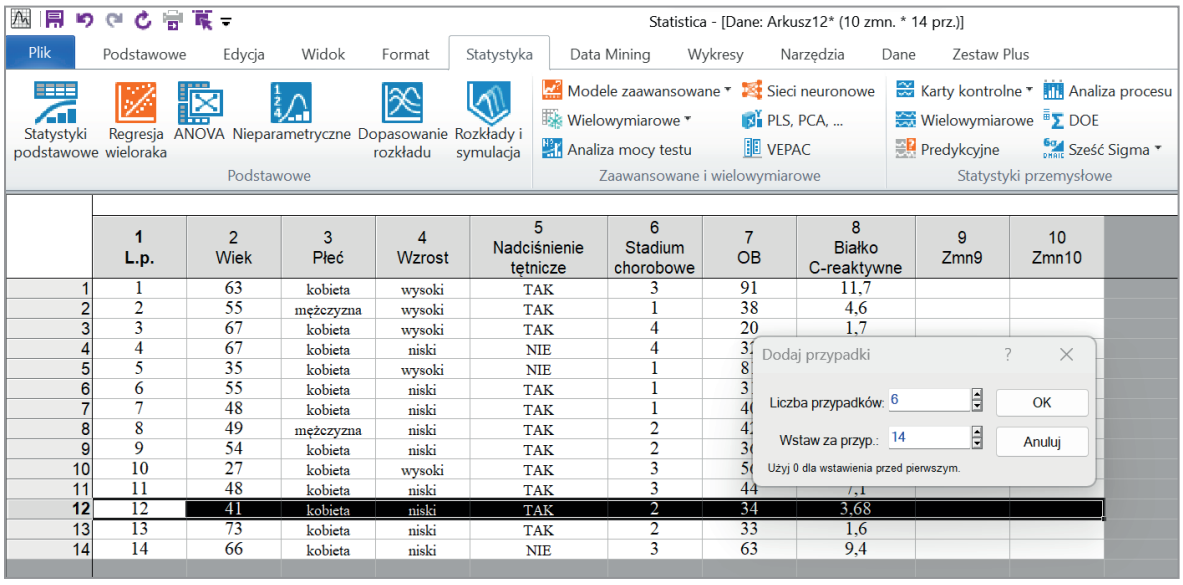
Na tym etapie można też dodać lub zmienić charakter zmiennej (ogólna, liczba, data, czas, procent, […] inna zdefiniowana przez użytkownika). Efekt jest podobny do zawartości arkusza kalkulacyjnego, z którego przeniesiono dane (➥ **ryc. 8.6**).



	1 L.p.	2 Wiek	3 Płeć	4 Wzrost	5 Nadciśnienie tętnicze	6 Stadium chorobowe	7 OB	8 Białko C-reaktywne	9 Zmn9	10 Zmn10
1	1	63	kobieta	wysoki	TAK	3	91	11,7		
2	2	55	mężczyzna	wysoki	TAK	1	38	4,6		
3	3	67	kobieta	wysoki	TAK	4	20	1,7		
4	4	67	kobieta	niski	NIE	4	32	0,6		
5	5	35	kobieta	wysoki	NIE	1	81	7,4		
6	6	55	kobieta	niski	TAK	1	31	1,9		
7	7	48	kobieta	niski	TAK	1	40	0,8		
8	8	49	mężczyzna	niski	TAK	2	42	3,3		
9	9	54	kobieta	niski	TAK	2	36	1,4		
10	10	27	kobieta	wysoki	TAK	3	56	12,4		
11	11	48	kobieta	niski	TAK	3	44	7,1		
12	12	41	kobieta	niski	TAK	2	34	3,68		
13	13	73	kobieta	niski	TAK	2	33	1,6		
14	14	66	kobieta	niski	NIE	3	63	9,4		

Ryc. 8.6. Dane z arkusza kalkulacyjnego wraz ze zmienionymi nagłówkami dla zmiennych

Gdy zaistnieje konieczność dodania kolejnych przypadków, można, klikając prawym klawiszem myszy na szarym tle z lewej strony, rozwinąć menu i wybrać polecenie Dodaj przypadki. Pozwoli to na dodanie brakującej liczby wierszy (➥ **ryc. 8.7**).



	1 L.p.	2 Wiek	3 Płeć	4 Wzrost	5 Nadciśnienie tętnicze	6 Stadium chorobowe	7 OB	8 Białko C-reaktywne	9 Zmn9	10 Zmn10
1	1	63	kobieta	wysoki	TAK	3	91	11,7		
2	2	55	mężczyzna	wysoki	TAK	1	38	4,6		
3	3	67	kobieta	wysoki	TAK	4	20	1,7		
4	4	67	kobieta	niski	NIE	4	3			
5	5	35	kobieta	wysoki	NIE	1	8			
6	6	55	kobieta	niski	TAK	1	3			
7	7	48	kobieta	niski	TAK	1	4			
8	8	49	mężczyzna	niski	TAK	2	4			
9	9	54	kobieta	niski	TAK	2	3			
10	10	27	kobieta	wysoki	TAK	3	5			
11	11	48	kobieta	niski	TAK	3	44	7,1		
12	12	41	kobieta	niski	TAK	2	34	3,68		
13	13	73	kobieta	niski	TAK	2	33	1,6		
14	14	66	kobieta	niski	NIE	3	63	9,4		

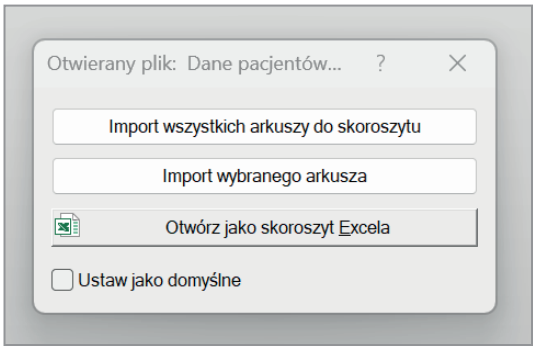
Ryc. 8.7. Dodawanie przypadków do bieżącego arkusza w programie Statistica

Potrzebne jest dodanie 6 pacjentów do uzyskania 20 osób w sumie z 14 początkowych, tak aby nie przesunąć pozostałych danych, co ma szczególne znaczenie podczas stosowania np. zmiennych grupujących, gdzie dane osoby są już przypisane do odpowiednich grup, a wklejenie 6 wierszy wyników w obrębie istniejących danych może zaburzyć całą analizę wariancji, przesuwając wyniki pomiędzy grupami.

Podsumowując, warto przygotować sobie wstępnie wszystkie dane w arkuszu kalkulacyjnym Excel i przenosić do programu statystycznego, ale w razie potrzeby istnieje możliwość dodawania i usuwania poszczególnych zmiennych i przypadków. Oczywiście należy zwrócić uwagę na to, czy miejsce wklejania danych lub ich usuwania nie wpływa na zmianę całkowitego układu doświadczenia, zwłaszcza gdy dane w analizie połączone są ze zmiennymi grupującymi.

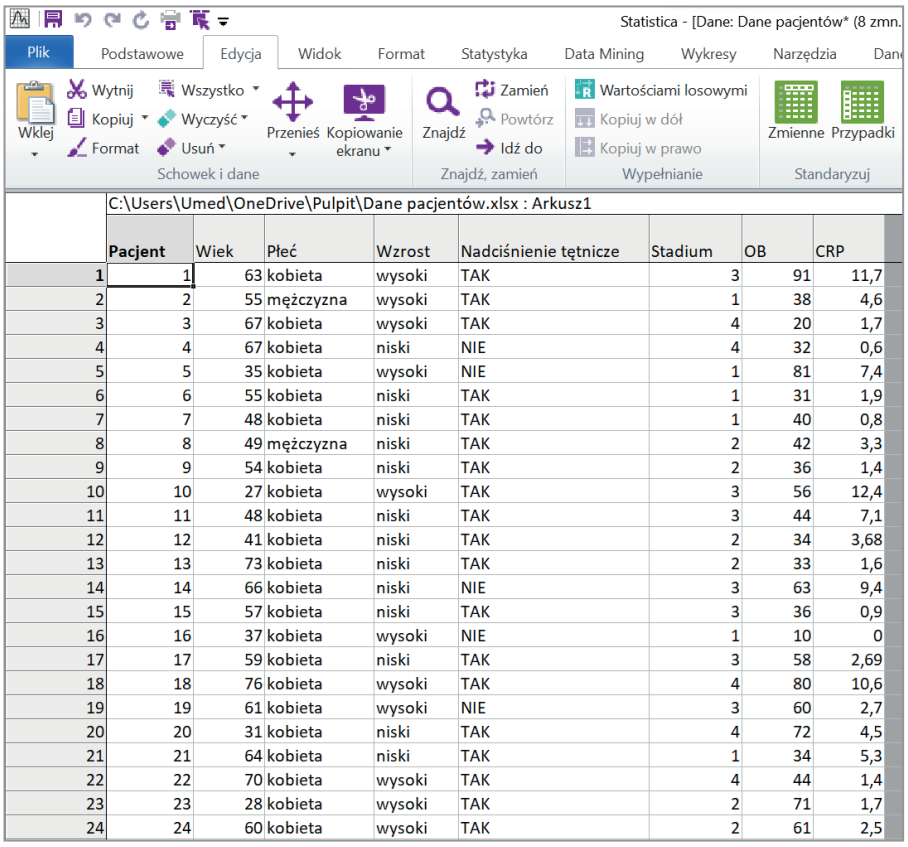
8.2. Otwieranie w programie Statistica arkusza z danymi przygotowanymi w Excelu

Gdy baza danych lub arkusz wynikowy zawiera znaczną liczbę rekordów, kopiowanie danych do programu może być uciążliwe, stąd też łatwiej otworzyć określony arkusz w programie bezpośrednio z pliku „xlsx". Taki arkusz kalkulacyjny z danymi pacjentów można bezpośrednio wczytać z zakładki Plik. Wybieramy ikonę Otwórz, a następnie w eksploratorze plików arkusz kalkulacyjny z danymi pacjentów (➥ **ryc. 8.8**). Możemy importować wszystkie arkusze lub pojedynczy arkusz z danymi. Istnieje również możliwość otwarcia pliku w postaci skoroszytu w programie Excel.



Ryc. 8.8. Otwieranie danych z arkusza kalkulacyjnego

Po wyborze tego sposobu przeniesienia danych istnieje możliwość ich zgromadzenia w jednym arkuszu. Efekt całej operacji wygląda tak jak na ➥ **ryc. 8.9**. Po przeniesieniu danych można już na nich bez problemów pracować w programie statystycznym. Czasem w momencie gdy przenoszone są dane tekstowe, program może wyświetlić komunikat odnośnie nadania etykiet poszczególnym wartościom zmiennych tekstowych. Akceptacja tej operacji pozwala na wygodniejszą pracę z danymi, gdy w analizie uwzględniane są zmienne jakościowe.



	Pacjent	Wiek	Płeć	Wzrost	Nadciśnienie tętnicze	Stadium	OB	CRP
1	1	63	kobieta	wysoki	TAK	3	91	11,7
2	2	55	mężczyzna	wysoki	TAK	1	38	4,6
3	3	67	kobieta	wysoki	TAK	4	20	1,7
4	4	67	kobieta	niski	NIE	4	32	0,6
5	5	35	kobieta	wysoki	NIE	1	81	7,4
6	6	55	kobieta	niski	TAK	1	31	1,9
7	7	48	kobieta	niski	TAK	1	40	0,8
8	8	49	mężczyzna	niski	TAK	2	42	3,3
9	9	54	kobieta	niski	TAK	2	36	1,4
10	10	27	kobieta	wysoki	TAK	3	56	12,4
11	11	48	kobieta	niski	TAK	3	44	7,1
12	12	41	kobieta	niski	TAK	2	34	3,68
13	13	73	kobieta	niski	TAK	2	33	1,6
14	14	66	kobieta	niski	NIE	3	63	9,4
15	15	57	kobieta	niski	TAK	3	36	0,9
16	16	37	kobieta	wysoki	NIE	1	10	0
17	17	59	kobieta	niski	TAK	3	58	2,69
18	18	76	kobieta	wysoki	TAK	4	80	10,6
19	19	61	kobieta	wysoki	NIE	3	60	2,7
20	20	31	kobieta	niski	TAK	4	72	4,5
21	21	64	kobieta	niski	TAK	1	34	5,3
22	22	70	kobieta	wysoki	TAK	4	44	1,4
23	23	28	kobieta	wysoki	TAK	2	71	1,7
24	24	60	kobieta	wysoki	TAK	2	61	2,5

Ryc. 8.9. Dane pacjentów wczytane bezpośrednio z pliku arkusza kalkulacyjnego

Praca z danymi może wydawać się trudnym zadaniem, jednak wypracowanie schematów postępowania i kryteriów doboru testu statystycznego pozwala na łatwiejsze rozeznanie się w całym zbiorze uzyskanych wyników. Natomiast wcześniejsze zaplanowanie doświadczeń pozwala zawęzić zakres stosowanych testów – najczęściej do dylematu doboru testów parametrycznych i nieparametrycznych pod kątem kryterium normalności rozkładu danych oraz uwzględnienia homogeniczności ich wariancji. Oczywiście nic nie stoi na przeszkodzie, by analizować wstępnie dane w ramach quasi-eksperymentu. Dzięki zastosowaniu technik *data mining* można w jego trakcie wyszukiwać tendencje, prawidłowości i zależności w ramach badań retrospektywnych. Na tej podstawie możliwe jest budowanie modeli zjawisk dla sprawdzenia ich efektywności przy zastosowaniu badań prospektywnych, uwzględniając przy tym współczynnik dopasowania danych do utworzonego modelu zjawiska oraz odpowiednią liczbę obserwacji w ramach grup doświadczalnych lub zmiennych w przypadku analizy regresji prostoliniowej.

Agnieszka Rusiecka,
Tomasz Hikawczuk

9 Statystyka opisowa i sprawdzanie normalności rozkładu danych

Cel ćwiczenia. Określenie podstawowych estymatorów według danych zebranych w przychodni z zastosowaniem statystyki opisowej, sprawdzenie normalności rozkładów danych oraz tego, czy dane pochodzą z tej samej populacji.

9.1. Statystyka opisowa w praktyce

Dane statystyczne pozwalają nam wnioskować o całej populacji. Brak możliwości zebrania z niej wszystkich wyników jest najczęstszym problemem w analizie statystycznej. Najczęstszą przyczyną jest wielkość samej populacji, a tym samym wynikające z tego ograniczenia czasowe, nakład pracy czy też koszty przeprowadzenia analiz. Stąd najczęściej wnioskowanie o populacji jest przeprowadzane na podstawie próby. Próba pobrana z populacji musi wykazywać kilka podstawowych cech. Pierwszą z nich jest losowość, definiowana jako identyczna szansa trafienia każdego elementu ze zbioru danych z populacji do próby. Drugą reprezentatywność, która określa stopień odzwierciedlenia struktury danych w populacji na podstawie pobranej próby, a kolejną liczebność próby, która jest szczególnie przydatna w określeniu mocy testu pod względem statystycznej weryfikacji danego zjawiska. Dodatkowo warto na wykresie normalności lub histogramie sprawdzić, czy dane mają postać krzywej Gaussa i dominuje pojedyncza wartość modalna. Wielokrotna wartość modalna w analizie wskazuje na dane pochodzące z więcej niż jednej populacji i najczęściej przejawia się to rozkładem bimodalnym uzyskanym w trakcie dopasowania rozkładu danych.

Odpowiednio pobrana z populacji próba pozwala na weryfikację hipotezy zerowej (H_0) i alternatywnej w trakcie przeprowadzania testu statystycznego. Aby jednak móc przeprowadzić takie wnioskowanie, należy w pierwszej kolejności określić poziom istotności testu, zwykle oznaczany jako „alfa", oraz poziom ufności, czyli wynik wyrażenia $1 - \alpha$. Przedział ufności wyraża zakres wyników, dla którego nie ma podstaw do odrzucenia H_0 i wynosi on najczęściej 95% lub 99%. W zapisie określa się go jako odległość 1,96 odchylenia standardowego w obydwu kierunkach od średniej z populacji. Stąd też wartość średnia powinna znajdować się w podanym zakresie, aby móc zaakceptować H_0:

$$X - 1{,}96 \leq \mu \leq X + 1{,}96$$

Przedział ufności jest uzależniony od dokładności i charakteru badań. Gdy porównujemy daną zmienną w 2 grupach na podstawie zebranych danych, wówczas przedział ufności na poziomie 95% będzie jak najbardziej wystarczający. Natomiast gdy przeprowadzane są dokładne badania, biorące również pod uwagę wcześniejsze planowanie badań i metody doboru osób do badań, aby jak najbardziej ujednorodnić zmienność, wnioskowanie może uwzględniać większy przedział ufności dla H_0 (0,99). Poziom istotności (poziom α) określa zaś punkt dla przedziału krytycznego, w przypadku którego możliwe jest odrzucenie hipotezy zerowej.

Statystyka opisowa jest podstawowym narzędziem służącym do opisu danych w pojedynczym lub kilku zbiorach danych pacjentów w próbie, które niekoniecznie mają zaplanowany układ doświadczalny. Bardzo się przydaje do opisu w badaniach obserwacyjnych, gdzie nie jest zaplanowane działanie określonego czynnika doświadczalnego. Pozwala ona na określenie podstawowych estymatorów opisujących dany

zbiór wartości liczbowych uzyskanych w przeprowadzonym doświadczeniu lub zgromadzonych w bazach danych na podstawie obserwacji lub przeprowadzonego doświadczenia. Estymator jest funkcją służącą do określenia nieznanego parametru populacji na podstawie pobranej z niej próby.

Pobrana z populacji próba jest opisana za pomocą litery n oznaczającej liczbę elementów próby, na podstawie której odbywa się wnioskowanie o danej populacji. Dodatkowo doświadczenia z przygotowanym planem eksperymentu uwzględniają 2 wartości: N i n, gdzie N określa liczbę obserwacji w całym układzie doświadczalnym, natomiast n w poszczególnych grupach doświadczalnych. Statystyka opisowa charakteryzuje miary tendencji centralnej (średnia, mediana, wartość modalna) i miary zmienności w postaci odchylenia standardowego, wariancji, kwantyli, wartości maksymalnej i minimalnej, w pewnym stopniu odzwierciedlając opis graficzny histogramu lub krzywej rozkładu normalnego. Statystyki opisowe obejmują także „skośność", czyli ocenę symetryczności rozkładu normalnego, oraz kurtozę opisującą gęstość obserwacji w danym przedziale i definiującą rozkład normalny w kierunku leptokurtycznego (z cięższymi „ogonami" i większym zagęszczeniem danych w okolicy wartości średniej) lub platykurtycznego (lżejsze „ogony" i mniejsze zagęszczenie danych w okolicy średniej).

9.2. Miary tendencji centralnej

Wśród estymatorów stosowanych do opisu tendencji centralnej wyróżnia się wartość średnią, medianę oraz wartość modalną. Wybór właściwej miary tendencji centralnej zależy od typu analizowanej zmiennej i typu rozkładu częstości jej wartości.

Średnia arytmetyczna jest właściwym estymatorem w sytuacji, gdy ma się do czynienia z rozkładami normalnymi i homogenicznością wariancji w przypadku zaplanowanego doświadczenia w grupach, przy czym charakter porównań w doświadczeniu determinuje zastosowanie (zob. poniżej) odpowiedniego rodzaju średniej (arytmetyczna, geometryczna, harmoniczna). Chociaż w przypadku wartości modalnej problem w analizie pojawia się w momencie, gdy mamy do czynienia z modą wielokrotną, co wskazuje na pobranie danych do próby z różnych populacji.

9.2.1. Średnia arytmetyczna i inne rodzaje średniej

Średnia arytmetyczna jest jedną z podstawowych miar tendencji centralnej stosowaną do opisu danych pojedynczej próby lub populacji, jak również do porównywania różnic między grupami doświadczalnymi z zastosowaniem testów parametrycznych, takich jak: test t Studenta dla zmiennych zależnych i niezależnych oraz analiz wariancji i kowariancji. Stosuje się ją wtedy, gdy rozkład wartości danej zmiennej jest zbliżony do rozkładu normalnego (o tym, czy tak jest, decydują odpowiednie testy, takie jak Shapiro–Wilka lub Kołmogorowa–Smirnowa). Definicja średniej arytmetycznej opisuje ją jako sumę wszystkich pomiarów w danym zbiorze dzieloną przez ich liczebność:

$$\bar{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

W przypadku zbioru danych wartość średnią określa się symbolem $\bar{x}$, natomiast litera n oznacza liczbę obserwacji, na podstawie których jest określana wartość średniej arytmetycznej. Czasem, aby ograniczyć wpływ pomiarów odstających na wartość tego estymatora, stosuje się tzw. średnią przyciętą, gdzie usuwa się wartości górne i dolne (np. po 2,5% wartości najmniejszych i największych).

Oprócz średniej arytmetycznej zastosowanie znajduje średnia geometryczna wyrażająca się wzorem:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}$$

Stosowana jest w przypadku badania średniego tempa zmian zjawisk.

Natomiast najrzadziej stosowana jest w statystyce średnia harmoniczna. Stosuje się ją wtedy gdy chcemy uśrednić wartości podawane w przeliczeniu na daną jednostkę,

jak np. prędkość (km/h) lub gęstość zaludnienia (n/km^2). Średnią harmoniczną opisuje się wzorem:

$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

9.2.2. Mediana

Medianę definiuje się jako wartość wyznaczającą środek pomiarów w danym zbiowyników po uporządkowaniu wszystkich pomiarów w grupie od wartości najniższej do najwyższej (w przypadku nieparzystej liczby danych medianą jest wartość środkowa, a w przypadku liczby parzystej - średnia arytmetyczna z dwóch wartości środkowych). Jest statystyką porządkową i jej obliczenie opiera się na porządkowych właściwościach danych. Mediana jest odpowiednią miarą tendencji centralnej, gdy zbiór danych wykazuje dużą skośność lub bimodalność. Obliczenia z zastosowaniem mediany mogą być stosowane w przypadku braku założenia normalności rozkładu.

9.2.3. Wartość modalna

Wartość modalną (modę) stosuje się w celu określenia tendencji centralnej w zbiorze danych, jednak ma ona ograniczone zastosowanie. Określa wartość najczęściej występującą w danym zbiorze. Wartość modalna w danym zbiorze wyników może być również modą wielokrotną, w przypadku gdy największa liczba obserwacji jest identyczna dla dwóch lub więcej wartości.

Średnia arytmetyczna, mediana i moda wykazują różne położenie w stosunku do wierzchołka krzywej normalności w zależności od charakteru krzywej rozkładu normalnego. Jeśli rozkład danych jest idealnie symetryczny, średnia arytmetyczna, mediana i moda mają identyczną wartość. W przypadku rozkładów unimodalnych, ale niesymetrycznych, o umiejscowieniu poszczególnych miar tendencji centralnej decyduje skośność zbioru danych.

9.2.4. Centyle

Centyle są estymatorami stosowanymi w uporządkowanych zbiorach danych do oceny rozkładu częstości wartości danej zmiennej. Dzieli on populację lub próbę pobraną z populacji na określone części, w zależności od charakteru podziału całości danych na odcinki. Podział na 100 części definiuje percentyle, 10 – decyle, 5 – kwintyle, 4 – kwartyle, przy podziale na 2 części mediana wyznacza punkt środkowy.

9.3. Miary zmienności

Zmienność charakteryzuje różnorodność wyników uzyskiwanych w danym zbiorze, na co wpływ mają cechy różnicujące populację, z której została pobrana próba. W zależności od charakteru pobranej próby zmienność może być większa lub mniejsza.

Miary zmienności pozwalają na dokładniejszy opis całości danych z uwzględnieniem ich rozproszenia wobec miary centralnej.

9.3.1. Rozstęp

Rozstęp stanowi różnicę pomiędzy najwyższą i najniższą wartością w analizowanym zbiorze danych. Pozwala określić zakres uzyskanych danych, jednak ma pewne wady związane z wnioskowaniem opartym na tym rodzaju estymatora. Najistotniejszą jest to, że jest niestabilny i w przypadku małych prób jest obarczony dużym ryzykiem błędu, jednak jego stabilność rośnie bardzo szybko wraz ze zwiększaniem się liczby obserwacji w badanej próbie. Rozstęp jest zatem estymatorem zależnym od wielkości próby. Charakteryzuje się asymptotycznym wzrostem wraz z wielkością próby.

9.3.2. Wariancja i odchylenie standardowe

Wariancja jest jedną z najczęściej stosowanych miar zmienności w analizie statystycznej. Można ją oznaczać zarówno w populacji, jak i próbie pobranej z populacji generalnej. Wariancja jest kwadratem odchylenia standardowego będącego sumą kwadratów różnic między poszczególnymi wartościami a ich średnią, podzieloną przez liczbę wartości. Wariancję można określić na dwóch poziomach: populacji oraz próby. Wariancja z populacji jest obliczana na podstawie poniższego wzoru:

$$s^2 = \frac{\Sigma(x-\mu)^2}{N}$$

gdzie:
x – wartość pomiaru,
μ – średnia w danej populacji,
N – liczebność populacji.

Natomiast w badanej próbie przyjmuje ona następującą postać:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

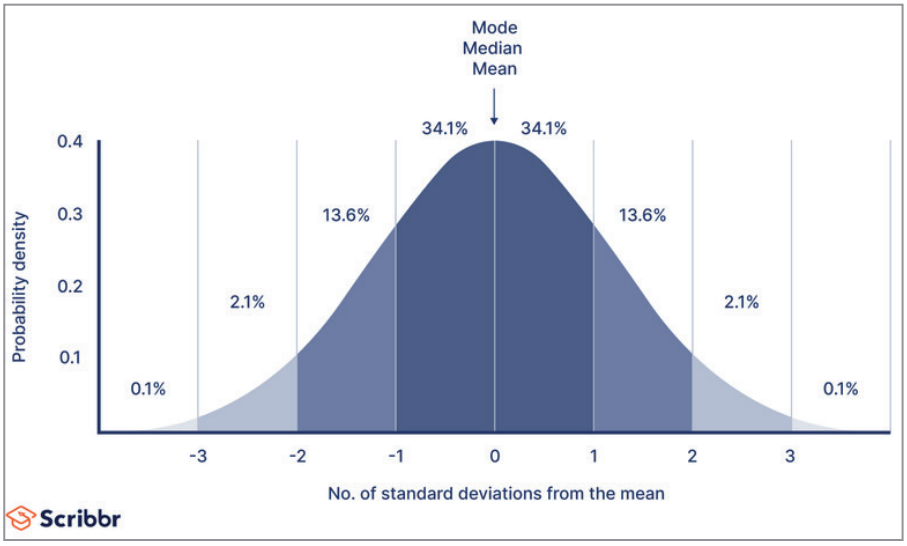
gdzie:
s^2 – wariancja z próby,
n – liczba pomiarów.

Z kolei w charakterystyce analizowanego zbioru danych średniej arytmetycznej najczęściej towarzyszy odchylenie standardowe (s), które obliczane jest jako pierwiastek z wariancji:

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

9.4. Miary asymetrii

Miary asymetrii odnoszą się do możliwych wariacji rozkładu odbiegających od krzywej charakteryzującej rozkład normalny, tj. mającej charakterystyczny, symetryczny, dzwonowaty kształt (➡ **ryc. 9.1**).



Ryc. 9.1. Standardowa krzywa rozkładu normalnego z gęstością prawdopodobieństwa i zakresem 6 sigma

Źródło: https://www.scribbr.com/statistics/normal-distribution/

W pewnych przypadkach wierzchołek krzywej może być przesunięty względem środka symetrii w prawą lub lewą stronę, czyli wykazywać skośność, i wówczas potwierdzenie normalności rozkładu danych będzie możliwe przy zastosowaniu testów normalności, np. Shapiro–Wilka, Kołmogorowa–Smirnowa, Andersona–Darlinga itd. Stąd też, aby rozróżnić tego typu charakter rozmieszczenia gęstości danych, warto wprowadzić dwa współczynniki charakteryzujące przesunięcie wierzchołka

w prawą lub lewą stronę, określające prawo- lub lewoskośność oraz gęstość danych w okolicach punktu centralnego, pozwalające zakwalifikować dany rozkład jako lepto- lub platykurtyczny (odpowiednio: spiczasty i spłaszczony) i odróżnić tego typu rozkłady od rozkładu normalnego.

Wartości obserwacji pomiarowych mają często niesymetryczne rozkłady i zróżnicowany stopień koncentracji wokół danej miary tendencji centralnej (mediany lub średniej). Asymetria takich próbek wzrasta przy zmniejszaniu się liczby ich elementów. Jednakże w wielu przypadkach występujących w praktyce dysponuje się tylko małą liczbą obserwacji pomiarowych i z różnych przyczyn nie ma możliwości jej zwiększenia. Ograniczenie to spowodowane jest:

- brakiem większej liczby obiektów badanych (np. przy walidacji metody stosowanej tylko w kilku akredytowanych laboratoriach);
- dużym kosztem badań lub ograniczonym czasem ich wykonania;
- niemożnością ponownego przeprowadzenia pomiarów, np. w badaniach odległych terenowo i w medycynie, w badaniach procesów jednorazowych i badaniach niszczących obiekt lub zmieniających nieodwracalnie jego właściwości.

9.5. Rozkład normalny

Dane statystyczne najczęściej przyrównuje się w pierwszej kolejności do rozkładu normalnego, który obrazuje krzywa Gaussa (➥ **ryc. 9.1**). Jest to krzywa dzwonowa, odzwierciedlająca rozkład wartości zmiennych charakteryzujących wiele cech i zjawisk przyrodniczych. Jednak w rzeczywistości błędy świadomie i nieświadomie popełniane w trakcie pobierania prób, polegające na zbyt małej liczebności obserwacji, wpływają na pozyskiwanie danych o rozkładach różniących się od rozkładu normalnego pod względem współczynnika skośności i kurtozy. Współczynnik skośności określa asymetryczne przesunięcie wierzchołka krzywej na prawo lub lewo, charakteryzując odpowiednio rozkład prawo- i lewoskośny. Natomiast współczynnik kurtozy opisuje zagęszczenie danych w okolicach wartości średniej, uzyskując w ten sposób rozkład lepto- lub platykurtyczny. W statystyce istnieje wiele narzędzi do sprawdzania normalności rozkładu, co jest istotne, gdyż determinuje możliwość użycia testów parametrycznych, których moc jest większa niż testów nieparametrycznych, dostosowanych do zmiennych o rozkładach różniących się od normalnego. Czasami zwiększenie liczebności próby zwiększa prawdopodobieństwo uzyskania rozkładu normalnego. Jeśli dane nie mają rozkładu normalnego, wówczas niezbędne będzie zastosowanie testów nieparametrycznych.

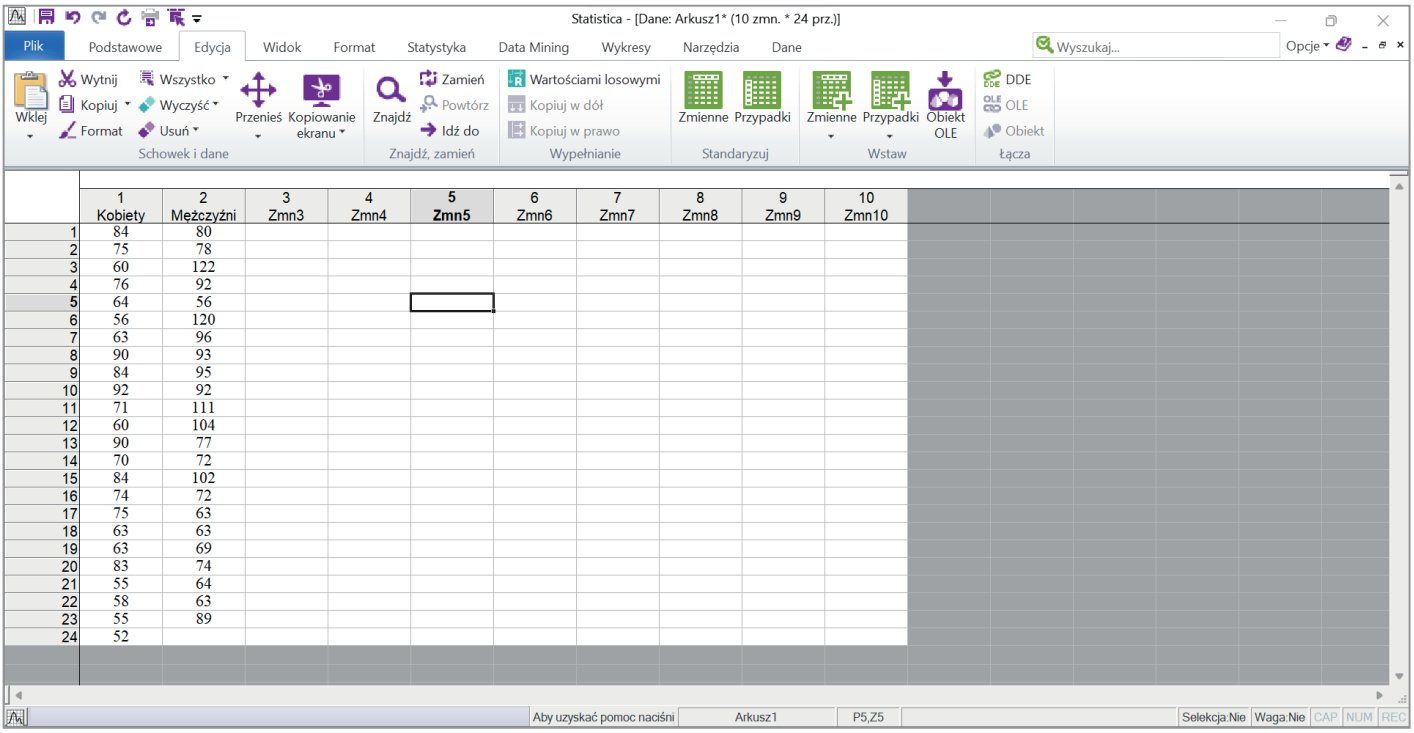
9.6. Przykład obliczeń

W przychodni uzyskano dane dotyczące masy ciała pacjentów mieszkających w pewnym mieście. Proszę o przeanalizowanie danych pod kątem podstawowych estymatorów charakteryzujących zarówno kobiety, jak i mężczyzn z tej przychodni i określić: miary tendencji centralnej (średnią arytmetyczną, medianę i wartość modalną); miary rozrzutu (wartość minimalną i maksymalną) oraz miary zmienności (odchylenie standardowe i wariancja). Na koniec proszę sprawdzić, czy w grupie kobiet i mężczyzn dane mają rozkład normalny, przygotowując histogram naniesiony na krzywą normalności. Testowanie normalności przeprowadza się za pomocą testu Shapiro–Wilka.

Masa ciała (kg)												
Kobiety	84	75	60	76	64	56	63	90	84	92	71	60
	90	70	84	74	75	63	63	83	55	58	55	52
Mężczyźni	80	78	122	92	56	120	96	93	95	92	111	104
	77	72	102	72	63	63	69	74	64	63	89	

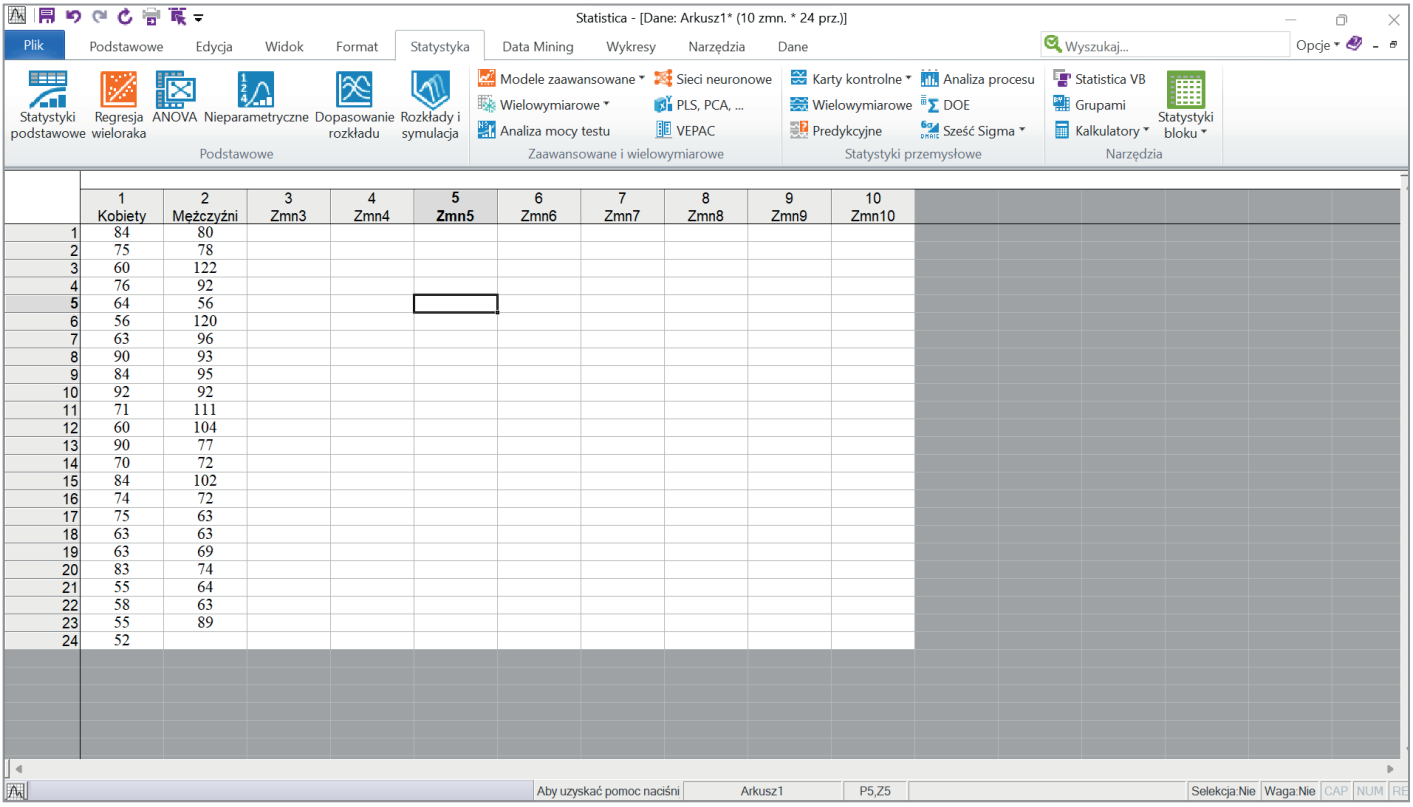
Zacząć należy od wpisania lub przeniesienia danych z pliku do programu statystycznego w 2 kolumnach oddzielnie dla kobiet i mężczyzn. Dodatkowo po dwukrotnym kliknięciu w pierwszą i drugą zmienną można zmienić nazwy poszczególnych kolumn w celu łatwiejszej orientacji w programie. Efekt działań jest przedstawiony na ➥ **ryc. 9.2**.

Ryc. 9.2. Dane początkowe do obliczeń statystycznych związanych ze statystyką podstawową



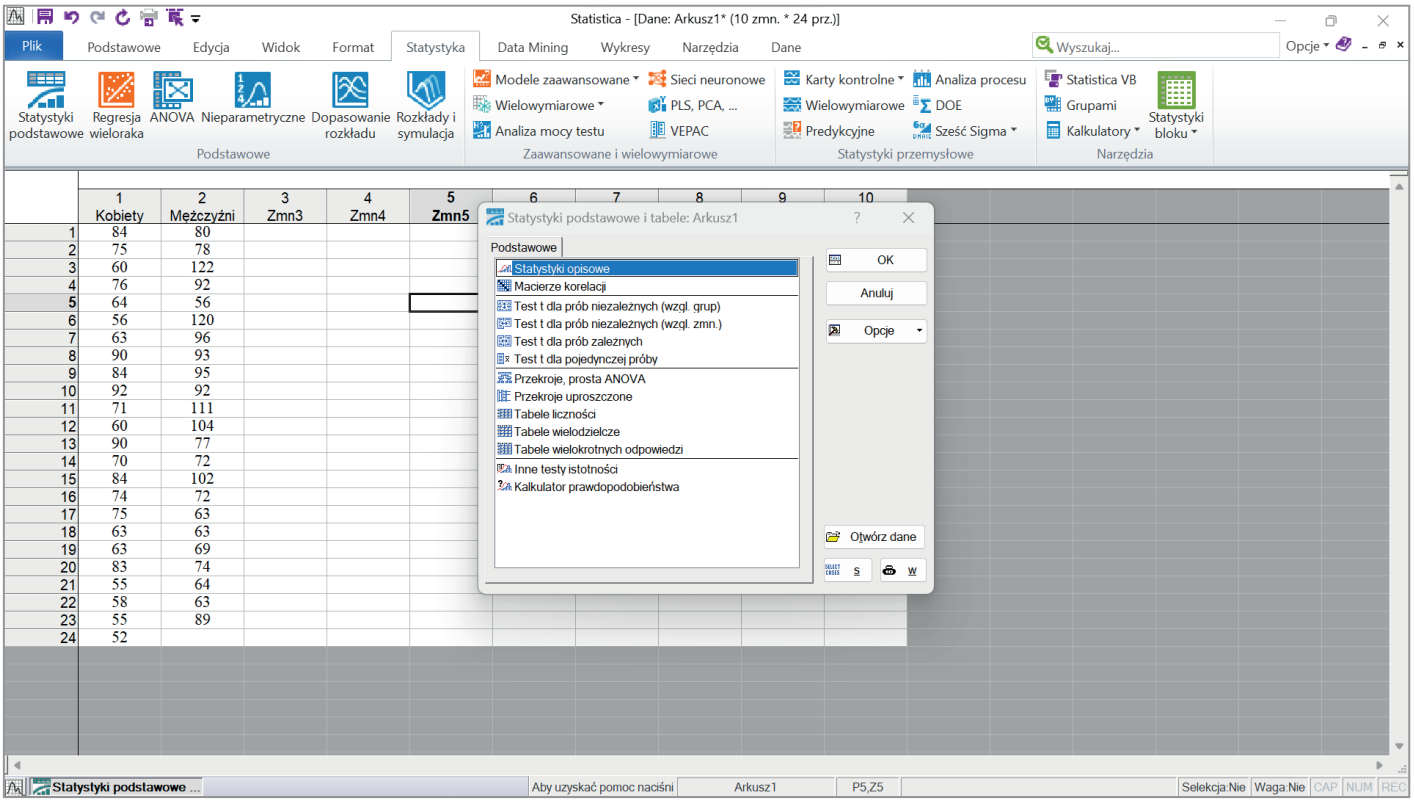
Po kliknięciu zakładki Statystyka wyświetla się menu, z którego wybieramy Statystyki podstawowe (➥ **ryc. 9.3**)

Ryc. 9.3. Zakładka Statystyka i opcje wyboru obliczeń statystycznych



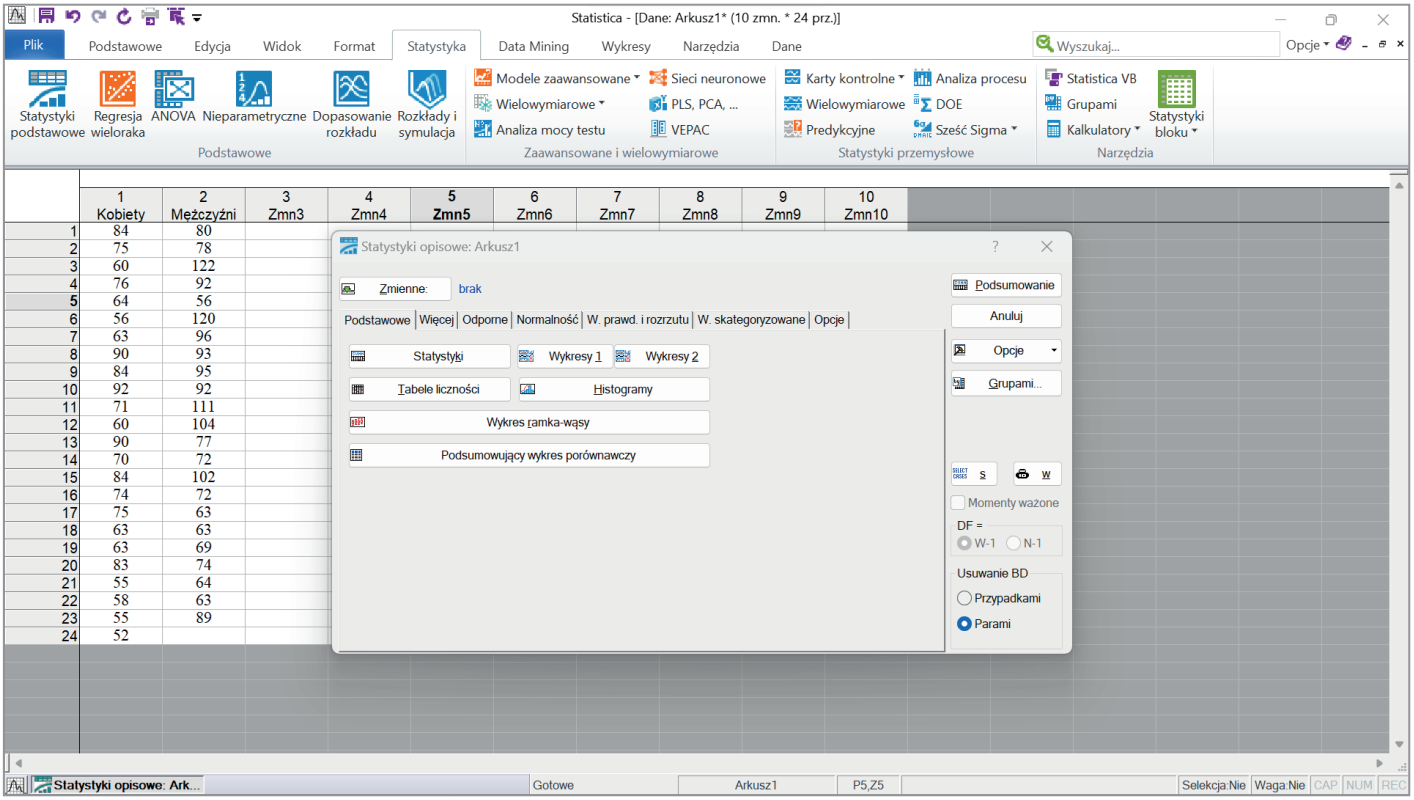
Po kliknięciu w Statystyki podstawowe pojawia się okno z możliwością obliczenia statystyk opisowych, testów *t*, analizy wariancji jednoczynnikowej, tabel liczności i wielodzielczych oraz innych testów istotności. Zaczynamy od wyboru Statystyk opisowych (➡ **ryc. 9.4**).

Ryc. 9.4. Okno Statystyki podstawowe



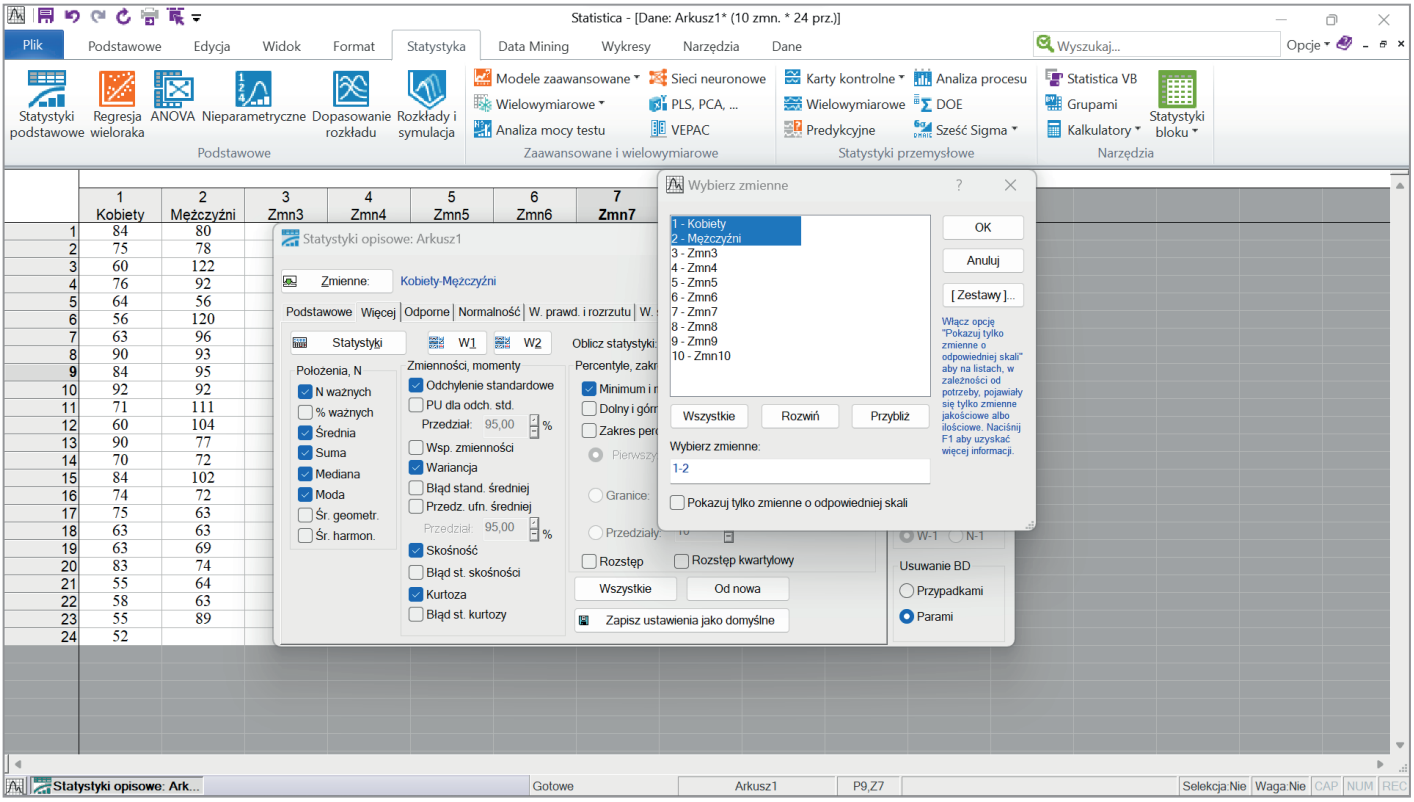
Po kliknięciu Statystyk opisowych pojawia się kolejne okno z możliwością wyboru zmiennej oraz m.in. zakładkami Podstawowe z możliwością utworzenia wykresów, Więcej z możliwością obliczenia estymatorów i Normalność pozwalającą na skonstruowanie histogramu i zweryfikowanie normalności odpowiednim testem. W przypadku liczby obserwacji $n \leq 50$ rekomendowane jest użycie testu Shapiro–Wilka, a przy $n > 50$ – testu Kołmogorowa–Smirnowa (➡ **ryc. 9.5**).

Ryc. 9.5. Okno z zakładkami i możliwością wyboru zmiennych



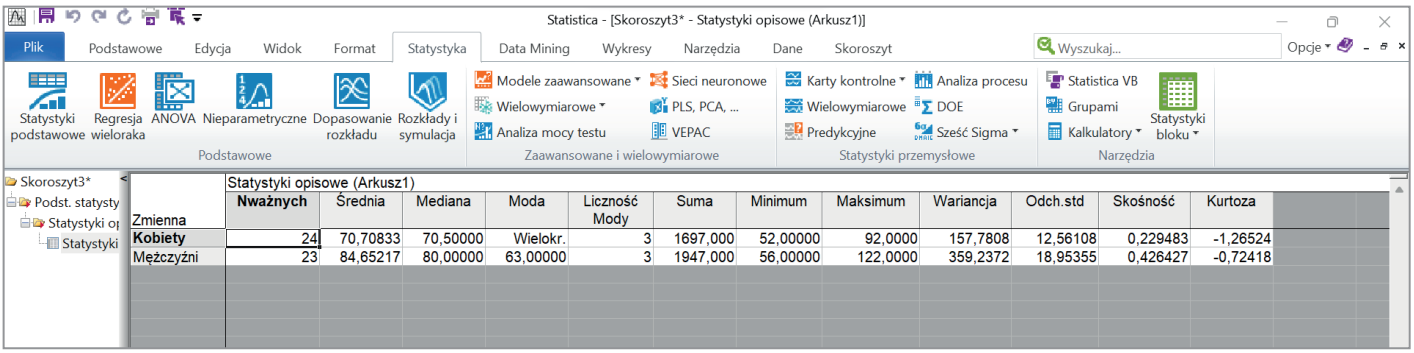
Ryc. 9.6. Oznaczenie estymatorów i zaznaczenie zmiennych

Przechodzimy do zakładki Więcej i zaznaczamy poszczególne estymatory dla zbioru danych (➥ **ryc. 9.6**), następnie dokonujemy wyboru zmiennych i klikamy na Podsumowanie.



Ryc. 9.7. Wartości estymatorów dla grup kobiet i mężczyzn

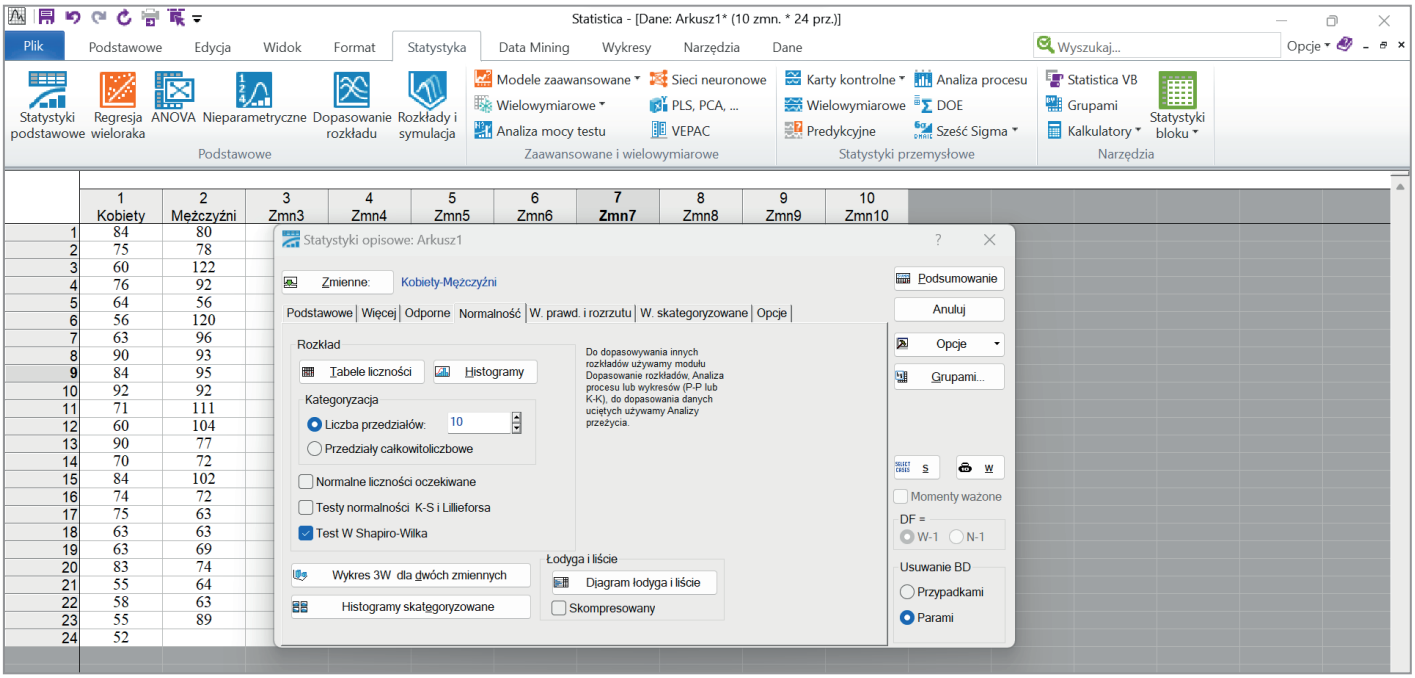
W efekcie uzyskujemy wartości dla kobiet i mężczyzn (przedstawione na ➥ **ryc. 9.7**).



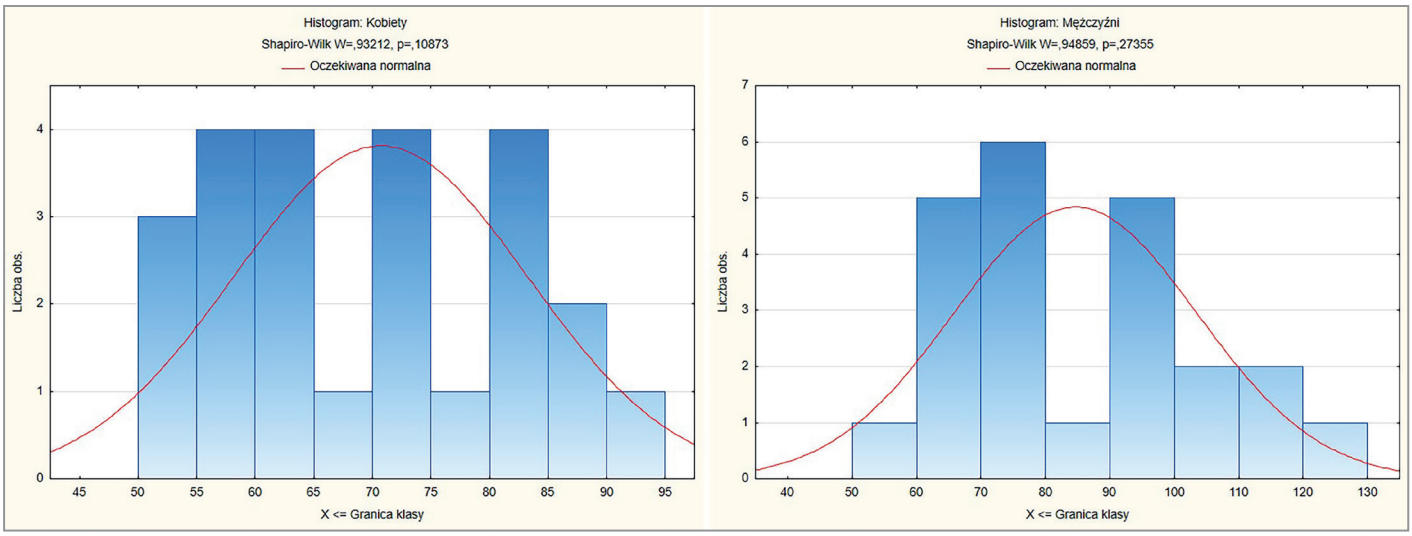
Zmienna	Nważnych	Średnia	Mediana	Moda	Liczność Mody	Suma	Minimum	Maksimum	Wariancja	Odch.std	Skośność	Kurtoza
Kobiety	24	70,70833	70,50000	Wielokr.	3	1697,000	52,00000	92,0000	157,7808	12,56108	0,229483	-1,26524
Mężczyźni	23	84,65217	80,00000	63,00000	3	1947,000	56,00000	122,0000	359,2372	18,95355	0,426427	-0,72418

Samą prezentację wizualną można wzbogacić przez dodanie wykresu oraz zastosowanie testu normalności. W tym celu w zakładce Normalność należy zaznaczyć test Shapiro–Wilka i można kliknąć przycisk Histogramy, aby dopasować dane do rozkładu normalnego (z ewentualną zmianą liczby przedziałów, ➥ **ryc. 9.8**) i dokładnej weryfikacji testem (➥ **ryc. 9.9**).

Ryc. 9.8. Zakładka dotycząca badania normalności rozkładu w poszczególnych grupach doświadczalnych

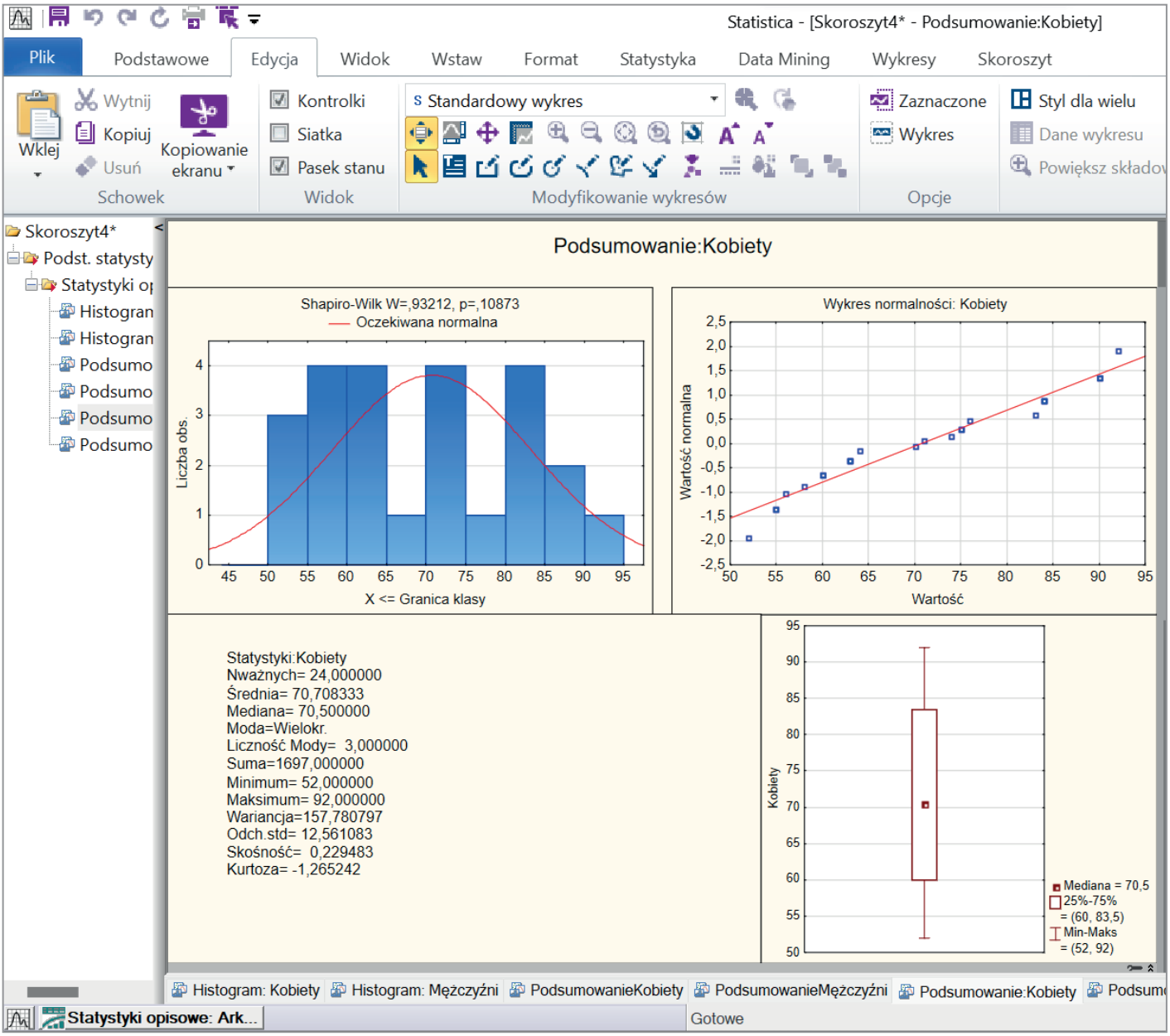


Ryc. 9.9. Oddzielne histogramy dla poszczególnych grup danych (kobiety, mężczyźni) z naniesioną wartością testu Shapiro–Wilka i wartością *p*

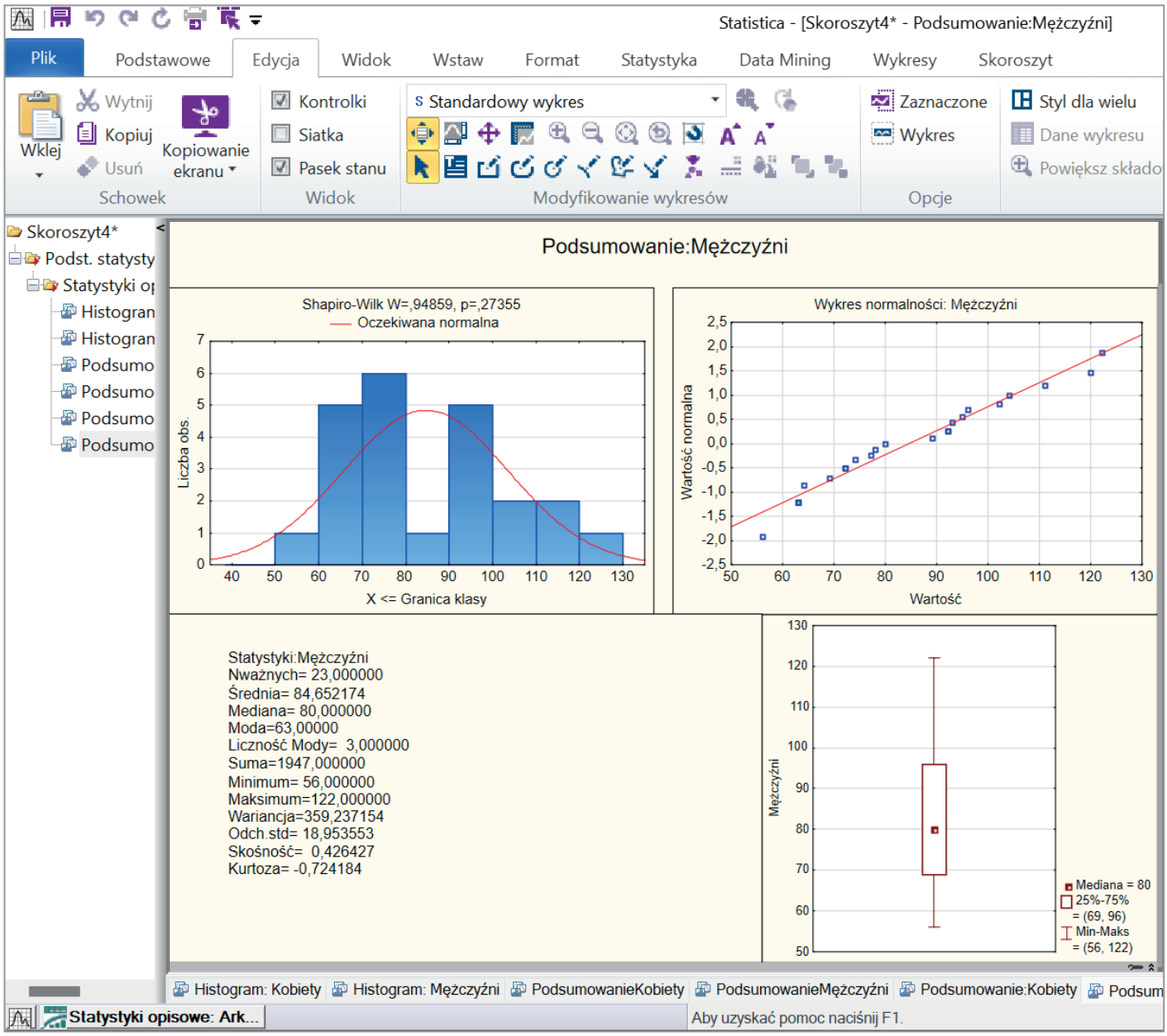


Ryc. 9.10. Podsumowanie dotyczące danych kobiet zawierające wykresy normalności i estymatory oraz wykres pudełkowy z wąsami

Na podstawie obliczeń w programie możemy również uzyskać podsumowanie danych dla kobiet i mężczyzn (➥ **ryc. 9.10** i ➥ **ryc. 9.11**).



Ryc. 9.11. Podsumowanie dotyczące danych mężczyzn uwzględniające wykresy normalności i estymatory oraz wykres pudełkowy z wąsami



Rezultatem użycia statystyk opisowych jest ➥ **tabela 9.1**.

Estymator	Kobiety	Mężczyźni
Liczba obserwacji (n)	24	13
Średnia (M)	70,7	84,6
Odchylenie standardowe (SD)	12,6	18,9
Mediana (Me)	70,5	80,0
Moda (Mo)	55, 60, 63, 75, 84, 90	63
Wariancja (s^2)	157,8	359,2
Wartość $p_{Shapiro-Wilka}$	0,109	0,274
Rozkład normalny	TAK	TAK

Tabela 9.1. Wartości estymatorów masy ciała oraz wyniki użycia testu Shapiro–Wilka dla dwóch grup doświadczalnych wyodrębnionych spośród pacjentów przychodni

9.6.1. Zadanie 1

Dane zebrane poniżej przedstawiają urodzeniową masę ciała noworodków (w kg). Proszę sprawdzić na podstawie normalności rozkładu, czy dane pochodzą z jednej populacji, oraz określić podstawowe estymatory ją opisujące.

3,99	3,31	3,61	3,60	3,51
3,89	3,70	3,83	4,13	2,71
3,18	2,90	3,41	3,36	3,53
2,74	3,27	3,42	3,54	2,38
3,60	3,65	3,23	3,65	2,34
3,73	4,08	2,86	3,69	3,34

Estymator	**Urodzeniowa masa ciała noworodków**
Liczba obserwacji (n)	
Średnia (M)	
Odchylenie standardowe (SD)	
Mediana (Me)	
Moda (Mo)	
Wariancja (s^2)	
Wartość $p_{Shapiro-Wilka}$	
Rozkład normalny	

Tabela 9.2. Urodzeniowa masa ciała noworodków (w kg). Dane do zadania 1

9.6.2. Zadanie 2

Podczas wizyty w przychodni zebrano dane do określenia wskaźników BMI (stosunek masy ciała do wzrostu, ang. *body mass index*) i WHR (stosunek obwodu talii do obwodu bioder, ang. *waist-to-hip ratio*). Proszę sprawdzić, czy dane przedstawione w ➥ **tabeli 9.3** w przypadku poszczególnych zmiennych mają rozkład normalny (histogram z naniesionym testem Shapiro–Wilka), oraz określić, o ile to możliwe, podstawowe wartości estymatorów dla każdej zmiennej.

Wiek	Płeć	Masa ciała [kg]	Wzrost [cm]	Obwód w talii [cm]	Obwód w biodrach [cm]
30	M	70	180	75	96
25	M	71	191	84	96
20	M	81	177	87	97
25	M	72	193	82	95
45	K	72	170	79	96
18	M	97	183	103	98
42	K	58	162	81	93
29	K	47	162	62	89
20	K	56	166	67	83
34	K	67	157	90	104
22	M	86	188	87	106
25	M	125	190	127	130
33	M	109	182	111	109
34	M	102	174	112	106
28	K	49	165	64	90
22	M	63	160	85	94
27	K	54	155	77	99
26	M	86	178	95	98
32	K	78	175	102	128
23	M	77	188	86	98
24	M	76	177	89	91
30	K	76	170	104	110
18	M	80	185	84	100
35	K	50	160	68	92
28	K	67	157	92	100

Tabela 9.3. Parametry ciała pacjentów. Dane do wykorzystania w zadaniu 2

Wyszczególnienie	Wiek	Masa ciała [kg]	Wzrost [cm]	Obwód w talii [cm]	Obwód w biodrach [cm]
Liczba obserwacji (n)					
Średnia (M)					
Odchylenie standardowe (SD)					
Mediana (Me)					
Moda (Mo)					
Wariancja (s^2)					
Wartość $p_{Shapiro–Wilka}$					
Rozkład normalny					

Tabela 9.4. Rozwiązanie zadania 2

Agnieszka Rusiecka

10 Testowanie hipotez statystycznych. Podstawowe pojęcia

Przeprowadzenie badań naukowych wymaga często wieloetapowych działań, których celem jest uzyskanie rzetelnego i obiektywnego obrazu rzeczywistości. Najczęściej jest to jednak tylko obraz pewnego wycinka interesującego nas zagadnienia. Wiąże się to z ograniczeniami, które napotykamy w pracy badawczej. Najbardziej miarodajne wyniki można uzyskać po analizie danych zebranych w całej populacji. Często jednak ze względu na złożoność, ograniczoną dostępność, dużą liczność zbiorowości lub czasochłonność i kosztowność procedur nie ma możliwości przeprowadzenia badań całej interesującej nas populacji. W takiej sytuacji najczęściej decydujemy się na wybór próby.

Ustalmy definicje, których będziemy odtąd używać:

- **populacja** – inaczej zbiorowość statystyczna – zbiór obiektów (osób, przedmiotów, zdarzeń) logicznie ze sobą powiązanych (ale nie identycznych), poddany badaniu statystycznemu;
- **próba** – wybrana część populacji, która jest badana w celu uogólnienia wyników na całą populację.

Skoro zatem najczęściej zmuszeni jesteśmy prowadzić badania naukowe nad fragmentem populacji, którym jest próba, a następnie na ich podstawie wysnuwać wnioski w odniesieniu do ogólnej zborowości, musimy liczyć się z ryzykiem, że nasze wyniki nie będą w pełni reprezentatywne dla populacji. Ryzyko popełnienia błędu możemy jednak minimalizować przez wybór reprezentatywnej próby, a także odpowiednie procedury próbkowania i dobór odpowiedniej liczebności próby. **Próba reprezentatywna** powinna odpowiadać cechom całej populacji (z wyjątkiem liczebności), a jej struktura powinna być zbliżona do struktury populacji statystycznej ze względu na badane cechy. Taki dobór próby pozwala na oszacowanie parametrów statystycznych dla całej populacji. Najczęściej wyliczanymi statystykami są: średnia arytmetyczna i odchylenie standardowe, które określone na podstawie danych z próby powinny być zbliżone do wartości parametrów w populacji, podobnie jak wszystkie inne miary rozkładu. Załóżmy np., że chcemy dowiedzieć się, jaki jest średni wzrost studentów we Wrocławiu. Wiemy, że populacja studentów w tym mieście wynosi 110 tys. (N). Ze względu na wielkość populacji nie jesteśmy w stanie zmierzyć wszystkich osób. Decydujemy się zatem na wybranie losowej próby 500 studentów (n). Na podstawie wykonanych pomiarów obliczamy wartość średniej arytmetycznej $\bar{x}$. Nie będzie ona oczywiście identyczna z wartością średniej populacji ogólnej (μ). Gdybyśmy zdecydowali się na powtórzenie badania i losowanie kolejnej próby 500 studentów z tej samej populacji, średnia arytmetyczna zapewne również nie będzie identyczna z wynikami pierwszej próby, a oba wyniki będą jedynie oszacowaniem wartości parametru populacyjnego μ. Różnicę między średnią z próbki $\bar{x}$ a średnią szacowaną dla populacji μ nazywamy błędem próbkowania. Problem polega na tym, że nigdy do końca nie wiemy, jak duży jest ten błąd. Żeby go oszacować, należałoby zmierzyć wzrost u wszystkich studentów Wrocławia, ale wtedy żadne szacunki nie byłyby potrzebne, ponieważ dysponowalibyśmy kompletem danych. Pewne natomiast jest to, że im większą część populacji przebadamy, tym pewniejsze będą nasze szacunki. Oznacza to, że im bardziej wzrasta liczebność próby, tym bardziej wartość średniej arytmetycznej $\bar{x}$ przybliża się do średniej całej populacji μ. Jeśli więc chcemy zapewnić dużą wiarygodność w naszych doświadczeniach, powinniśmy wybierać duże próby do badań. Jaką wielkość w takim razie powinna mieć próba, aby była reprezentatywna i pozwalała na wnioskowanie o całej populacji, a jednocześnie w praktyce wciąż była moż-

liwa do zebrania? W rozwiązaniu tego dylematu pomagają badaczom metody estymacji minimalnej liczebności próby losowej. Najczęściej wykorzystywaną procedurą statystyczną jest **testowanie hipotez**, zwane także badaniem istotności statystycznej. Jest to sprawdzanie sądów o populacji sformułowanych bez zbadania jej całości (na podstawie danych pochodzących z reprezentatywnej próby). **Hipotezą statystyczną** nazywamy dowolne przypuszczenie dotyczące rozkładu populacji lub wartości parametru rozkładu. Tradycyjnie dzielimy hipotezy statystyczne na dwie grupy, gdzie hipotezy parametryczne dotyczą wartości parametrów statystycznych populacji (np. średniej lub wariancji), a hipotezy nieparametryczne dotyczą innych niż średnia i wariancja cech rozkładu, takich jak mediana, lub ogólnego wzorca rozkładu wartości.

Procedura weryfikacyjna zazwyczaj składa się z pięciu podstawowych kroków.

10.1. Sformułowanie hipotez

W praktyce formułujemy dwie przeciwstawne hipotezy w taki sposób, że uznanie jednej z nich za nieprawdziwą narzuca konieczność zaakceptowania drugiej. Przy czym sposób ich formułowania jest z góry narzucony, a możliwe jest odrzucenie, z określonym prawdopodobieństwem, wyłącznie hipotezy zerowej (H_0). Nie ma natomiast możliwości udowodnienia jej prawdziwości. Również sama treść hipotezy jest ściśle określona. Umownie przyjęto, że H_0 w uproszczeniu zawsze zakłada, że nie ma różnic (lub zależności) między zmiennymi lub, innymi słowy, różnica równa jest 0 (lub zależności brak). Hipoteza alternatywna (H_1) przeciwstawia się hipotezie zerowej, zakłada zatem występowanie różnicy lub różnic. Z punktu widzenia badacza najczęściej dąży się do obalenia H_0 i przyjęcia hipotezy alternatywnej, a więc zależy nam na pokazaniu odmienności, wskazaniu na czynnik różnicujący.

1. $\boldsymbol{H_0}$ poddawana jest procedurze weryfikacyjnej, w której zakładamy, że różnica między analizowanymi parametrami lub rozkładami wynosi 0 (lub miara zależności, np. współczynnik korelacji, wynosi 0).

   Przykładowo stężenie hemoglobiny przed leczeniem i po leczeniu **jest takie samo** (nie różni się) lub: $\mu_1 = \mu_2$.
2. $\boldsymbol{H_1}$ jest przeciwstawna do weryfikowanej. Możemy ją zapisać na trzy sposoby w zależności od sformułowania badanego problemu:
   H_1: stężenie hemoglobiny przed leczeniem i po leczeniu **różni się**, $\mu_1 \neq \mu_2$,
   H_1: stężenie hemoglobiny przed leczeniem **jest niższe** niż po leczeniu, $\mu_1 < \mu_2$,
   H_1: stężenie hemoglobiny przed leczeniem **jest wyższe** niż po leczeniu, $\mu_1 > \mu_2$.

10.2. Przyjęcie odpowiedniego poziomu istotności

W wyniku testowania hipotez musimy podjąć decyzję odnośnie do prawdziwości hipotezy zerowej. Może ona zostać odrzucona w sytuacji, gdy zebrane dane z dużym prawdopodobieństwem wskazują, że jest fałszywa. W przeciwnym razie musimy uznać, że nie ma podstaw do jej odrzucenia. Nie oznacza to jednak, że jest ona w pełni prawdziwa. Taką pewność zyskalibyśmy jedynie, gdyby przebadana została cała populacja. Jak zatem podjąć decyzję o przyjęciu lub odrzuceniu hipotezy zerowej? Niestety musimy się pogodzić z możliwością pomyłki: błędnego odrzucenia „prawdziwej" H_0 lub błędnego przyjęcia „fałszywej" H_0. Ryzyko pomyłki określa się prawdopodobieństwem jej popełnienia i opisuje przy pomocy błędu I rodzaju i błędu II rodzaju (➡ **tabela 10.1**).

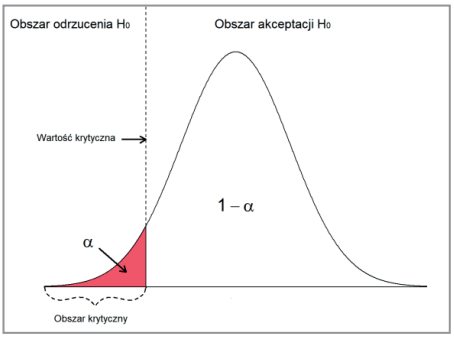
		Rzeczywistość	
		H_0 prawdziwa	**H_0 fałszywa**
Wynik testu	**odrzucić H_0**	**błąd I rodzaju** (α)	decyzja prawidłowa (moc testu: $1 - \beta$)
	nie odrzucać H_0	decyzja prawidłowa ($1 - \alpha$)	**błąd II rodzaju** (β)

Tabela 10.1. Testowanie hipotez statystycznych, błędy I i II rodzaju

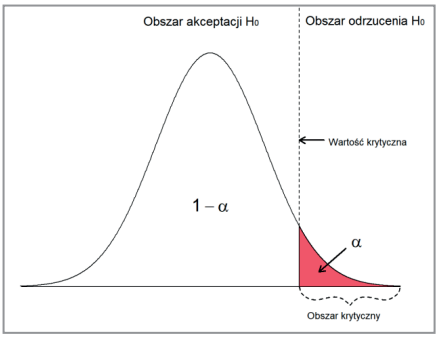
Przez **błąd I rodzaju** rozumiemy prawdopodobieństwo błędnego odrzucenia prawdziwej H_0. Prawdopodobieństwo popełnienia błędu I rodzaju nazywamy **poziomem istotności** i oznaczamy jako α. **Błąd II rodzaju** to prawdopodobieństwo przyjęcia H_0, gdy jest ona w rzeczywistości fałszywa. Prawdopodobieństwo popełnienia błędu II rodzaju oznaczać będziemy symbolem β. Wielkość błędu II rodzaju oraz moc testu są zależne od liczebności próby, wielkości efektu na tle losowej zmienności obserwowanej w populacji, ale również przyjętego poziomu istotności α. Tak więc wartości błędu I i II rodzaju również są ze sobą powiązane. Zmniejszenie prawdopodobieństwa α powoduje wzrost prawdopodobieństwa β, tzn. niższy poziom α oznacza mniejsze ryzyko odrzucenia prawdziwej H_0, a jednocześnie większe ryzyko przyjęcia fałszywej H_0. Dlatego pewnym kompromisem w tej sytuacji jest narzucenie z góry pewnej akceptowalnej dla badacza wartości α. W naukach medycznych i biologicznych najczęściej przyjmuje się wartość $\alpha = 0{,}05$ lub mniejszą. Poziom istotności α powinien być ustalony jeszcze przed rozpoczęciem testowania i jego wielkość jest arbitralna. Jak zatem stwierdzić, czy wynik jest istotny?

10.3. Wyznaczenie obszaru krytycznego testu przy ustalonym poziomie istotności

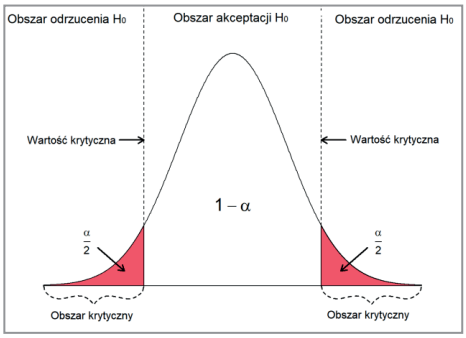
Każda statystyka testowa, jako zmienna losowa, posiada określony rozkład charakterystyczny dla testowania H_0. Znając ten rozkład, możemy określić, które wartości tej statystyki są mało prawdopodobne, a więc przeczą H_0, a których prawdopodobieństwo jest wyższe, czyli przemawiają na jej korzyść. Obszary, które jej przeczą, nazywamy **obszarami krytycznymi testu**, natomiast te wspierające H_0 – **obszarami akceptacji hipotezy zerowej**. Obszar krytyczny może składać się z 1 części i wtedy jest to obszar krytyczny jednostronny (➥ **ryc. 10.1–10.2**) lub z 2 części – w tym przypadku mamy do czynienia z obszarem dwustronnym (➥ **ryc. 10.3**). Postać zbioru uzależniona jest od sposobu sformułowania hipotezy alternatywnej wobec H_0. Wielkość obszaru krytycznego wyznaczają punkty krytyczne będące granicą między obszarem krytycznym a obszarem akceptacji H_0. Wartości krytyczne wynikają z rozkładu statystyki testowej przy ustalonym poziomie istotności α.



Ryc. 10.1. Lewostronny obszar krytyczny



Ryc. 10.2. Prawostronny obszar krytyczny



Ryc. 10.3. Dwustronny obszar krytyczny

Źródło: https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Hypothesis-Tests/Introduction-to-Hypothesis-Testing/Critical-Value-and-the-p-Value-Approach/index.html

Przeanalizujmy wyznaczanie obszaru krytycznego dla statystyki t. W pierwszym kroku formułujemy hipotezy dla testu. Hipotezę zerową formułujemy jako brak efektu lub różnicy.

Np. H_0: Średni poziom hemoglobiny u kobiet i mężczyzn nie różni się.

Hipoteza alternatywna (H_1): formułujemy ją jako istnienie efektu lub różnicy. Taką hipotezę można sformułować na trzy sposoby:

1. H_1 (dwustronna): Średni poziom hemoglobiny u kobiet i mężczyzn różni się (bez wskazania czy średni poziom hemoglobiny u kobiet jest wyższy czy niższy niż u mężczyzn).
2. H_1 (lewostronna): Średni poziom hemoglobiny jest niższy u kobiet.
3. H_1 (prawostronna): Średni poziom hemoglobiny jest wyższy u kobiet.

W kolejnym kroku, w zależności od testu i dostępnych danych, należy wybrać odpowiednią statystykę testową np. statystykę t. Statystyka t przy założeniu prawdziwości H_0 ma rozkład t Studenta o $n - 2$ stopniach swobody.

W zależności od tego jaką postawiliśmy hipotezę alternatywną, obszar krytyczny dla ustalonego α (z reguły $\alpha = 0{,}05$) będzie wyglądał tak jak pokazano na ➥ **ryc. 10.1–10.2**. Następnie za pomocą programów, obliczających p z wykorzystaniem odpowiedniej funkcji, odczytujemy prawdopodobieństwo tego, że wartość statystyki t, przy założeniu, że H_0 jest prawdziwa, będzie leżała w obszarze krytycznym. Dla hipotezy dwustronnej obszar ten można opisać jako sumę obszarów $(-\infty,-t\alpha/2]$ $[t\alpha/2,\infty)$, a jego wartość wynosi α. Jeśli pobierzemy próbę i okaże się, że obliczona z niej wartość t należy do obszaru krytycznego, to można stwierdzić, że zaszło zdarzenie, któremu towarzyszyło małe prawdopodobieństwo (rzędu wielkości α) prawdziwości H_0. W takiej sytuacji H_0 odrzucamy i przyjmujemy hipotezę H_1 (➥ **ryc. 10.3**).

Obszar krytyczny w przypadku hipotezy lewostronnej będzie mieścił się w przedziale prawdopodobieństwa $(-\infty,t\alpha]$ widocznym na ➥ **ryc. 10.2**. W przypadku hipotezy prawostronnej mamy prawostronny obszar krytyczny $[t\alpha,\infty)$ przedstawiony na ➥ **ryc. 10.1**.

10.4. Wybór odpowiedniego testu statystycznego i obliczenie jego wartości na podstawie danych pochodzących z próby

Testy, w których kontrolowany jest poziom istotności α, nazywamy **testami istotności**. Decyzja o wyborze odpowiedniego testu statystycznego i sprawdzeniu jego założeń należy do najważniejszych w całym procesie testowania. Uzależniona może być od takich czynników jak np. typ rozkładu badanej populacji czy wielkość próby. Wybór niewłaściwego testu istotnie podważa wartość całego późniejszego wnioskowania. Założenia do każdego testu statystycznego opisane zostały w dalszych częściach tego opracowania. W praktyce staramy się zawsze wybierać testy o największej mocy, tak by prawdopodobieństwo przyjęcia fałszywej H_0 było jak najmniejsze, tzn. takie, gdzie przy z góry ustalonej wartości α wartość β jest najmniejsza. Zwykle przyjmuje się, że β nie powinno przekraczać 0,2, co oznacza moc testu nie mniejszą niż 80%, aby wykryć różnicę w stosunku do twierdzenia zawartego w H_0.

Wybrawszy test, należy przystąpić do obliczenia tzw. wartości testowej na podstawie danych pochodzących z próby. Przykładowo w teście t Studenta jest to wartość t, w teście Manna–Whitneya – wartość U itd. Rezultatem obliczenia wartości testowej jest ustalenie wartości **istotności statystycznej** (ang. *p-value* – p). Wartość p jest prawdopodobieństwem prawdziwości H_0. Wartość p dla danego testu statystycznego wyliczana jest najczęściej na podstawie wartości statystyki testowej dla określonej liczby stopni swobody. Przez liczbę **stopni swobody** (ang. *degrees of freedom* – *df*) rozumiemy liczbę niezależnych wyników obserwacji pomniejszoną o liczbę związków, które łączą te wyniki ze sobą. Dlaczego jest to tak istotna miara statystyczna wykorzystywana w testach statystycznych? Związane jest to z występowaniem błędu oszacowania poszukiwanego parametru w populacji na podstawie wylosowanej próby. Jeżeli w analizach uwzględnialibyśmy liczbę zebranych obserwacji, a nie liczbę stopni swobody dla liczby obserwacji, oszacowanie parametru byłoby obciążone błędem systematycznym.

10.5. Podjęcie decyzji o hipotezach statystycznych

Decyzję o odrzuceniu lub nieodrzuceniu H_0 podejmujemy na podstawie obliczonej wartości statystyki testowej p. Po wyznaczeniu przedziału krytycznego sprawdzamy, czy wartość p należy do obszaru krytycznego czy też do obszaru akceptacji H_0.

Jeżeli obliczona wartość statystyki testowej należy do zbioru krytycznego, należy odrzucić H_0, a następnie przyjąć H_1. Jeżeli obliczona wartość statystyki testowej nie należy do zbioru krytycznego, oznacza to, że nie ma podstaw do odrzucenia H_0. Należy pamiętać, że nieodrzucenie hipotezy nie dowodzi jej prawdziwości.

W praktyce porównujemy wartość statystyki testowej p do przyjętego wcześniej poziomu istotności (α). Jeśli wartość p jest mniejsza lub równa wartości α, możemy odrzucić H_0 i przyjąć H_1, natomiast gdy $p > \alpha$, to nie ma podstaw do odrzucenia H_0.

10.5.1. Moc testu

Planując doświadczenie, chcemy, by cechowało się ono wysoką istotnością statystyczną, a więc żeby błąd I rodzaju był jak najmniejszy (α). Jaka jest zatem różnica między α i wartością p? Zakłada się, że α to prawdopodobieństwo przyjęte na poziomie planowania doświadczenia, np. zdecydowaliśmy, że będziemy testowali hipotezy na poziomie istotności $\alpha < 0{,}01$. Taka wartość oznacza, że mamy 1 szansę na 100, że odrzucimy H_0, która jest prawdziwa (błąd I rodzaju). Wartość p natomiast stosuje się, aby wskazać istotność statystyczną wyliczoną dla przyjętego wcześniej poziomu istotności na podstawie danych, którymi dysponujemy. Wartość p wylicza się więc już po wykonaniu doświadczenia. Jeśli wartość p jest większa niż teoretyczna wartość α, należy odrzucić H_0 i przyjąć H_1.

Trzeba również zadbać o moc testu, która w praktyce zależy od tego, jak bardzo różni się nasze wnioskowanie o próbie od rzeczywistości obserwowanej w populacji. Na statystyczną moc testu wpływają takie czynniki jak: wielkość badanej próby, wielkość efektu badanego czynnika (ang. *effect size*) oraz zakładanego poziomu α. W testach o dużej mocy statystycznej ryzyko popełnienia błędu II rodzaju (β) jest minimalne, zatem fałszywa H_0 zostanie odrzucona. W słabych testach statystycznych istnieje duże prawdopodobieństwo przyjęcia H_0, która nie jest zgodna z prawdą.

W próbach o dużej liczebności i przy danym założonym poziomie α, łatwiej (czyli z większą mocą) można wykryć nawet mały wpływ badanego czynnika, czyli np. małą różnicę w średnich przy porównaniu pewnych grup. Jednak wracamy wtedy do dylematu z początku rozdziału: jak duża powinna być próba, aby nasze wnioskowanie o populacji było miarodajne, a jednocześnie jej wielkość miała uzasadnienie w kosztach i nakładzie pracy badacza? Innymi słowy, jak duża powinna być próba, aby osiągnąć zamierzoną dokładność? Na te pytania należy sobie odpowiedzieć już na etapie planowania eksperymentu. Drugim czynnikiem wpływającym na moc testu jest wielkość efektu wywołanego przez czynnik badany. W praktyce badacz powinien zdecydować, jaka minimalna wielkość efektu jest znacząca klinicznie. Duży wpływ badanego czynnika (odzwierciedlony np. dużą różnicą średnich porównywanych grup) przybliża do odrzucenia H_0. Powinniśmy pamiętać również o wpływie zmienności parametrów na wiarygodność naszego wnioskowania: wysoka zmienność badanych zmiennych zmniejsza precyzję wnioskowania. Zmienność ta może wynikać z istniejącej zmienności biologicznej, stosowanych metod pomiarowych lub błędów metodologicznych, ale zawsze ma wpływ na ocenę precyzji naszego wnioskowania. Dlatego proponowane często miary wielkości wpływu (czyli wielkości efektu) uwzględniają zmienność danych. Na przykład powszechnie stosowaną miarą wielkości wpływu danego czynnika na średnią wartość pewnej zmiennej jest wskaźnik d Cohena,

$$d = \frac{m_1 - m_2}{SD}$$

gdzie m_1, m_2 to średnie w grupach 1 i 2, a SD – odchylenie standardowe (przy założeniu, że jest ono jednakowe w obu grupach). Zatem im większe SD (czyli większa zmienność danych), tym mniejszy wpływ d. Moc testu uzależniona jest również od typu zastosowanej procedury statystycznej, ponieważ testy różnią się pod tym względem – parametryczne są zazwyczaj mocniejsze niż nieparametryczne.

10.5.2. Wielkość efektu

W tradycyjnej analizie testowaniu poddawane są dwie przeciwstawne hipotezy, a jej celem jest wysnucie wniosków o populacji na podstawie danych pochodzących z próby oraz znalezienie odpowiedzi na postawione wcześniej pytanie badawcze. Takie podejście oparte jest na ocenie prawdopodobieństwa wystąpienia zdarzenia zawartego w H_0. Przypomnijmy: w trakcie analiz obliczany jest wynik istotności (wartość p), który porównany do wartości α (najczęściej równej 0,05) prowadzi do odrzucenia lub przyjęcia H_0 (a raczej uznania, że nie ma podstaw do jej odrzucenia). Należy przy tym pamiętać, że przyjęta wartość p na poziomie 0,05 jest umowna.

Należy również mieć na uwadze, że na wartość istotności duży wpływ mają takie czynniki jak liczebność próby, więc stawianie wniosków końcowych na podstawie jedynie tej wartości łatwo może prowadzić do błędów w podejmowaniu decyzji.

Liczba obserwacji ma znaczenie przy wyznaczaniu stopni swobody, które razem z obliczoną wartością testową stanowią podstawę do wyznaczenia istotności statystycznej p.

Zwiększenie liczby stopni swobody przez zwiększenie liczebności badanej próby może więc sprawić, że nawet najmniejsza różnica między grupami (bez praktycznej wartości) będzie uznana za istotną. Należy jednak pamiętać, że samo określenie tego, czy wynik jest istotny statystycznie, nie jest wystarczające do dokonania pełnej interpretacji uzyskanych wyników i wyciągnięcia trafnych wniosków. Aby więc przeprowadzane wnioskowanie było miarodajne, wykorzystuje się wielkość efektu (ang. *effect size*). Wielkość efektu to miara siły związku między zmiennymi w badaniu. Nie zależy od wielkości próby. Jest to miara statystyczna, która służy do oszacowania siły związku między zmiennymi, wyrażona jako liczba, niezależnie od zastosowanej jednostki pomiaru.

Wartość istotności statystycznej i wskaźniki siły efektu uzupełniają się, a wnioski płynące z jednego i drugiego pozwalają na rzetelną ocenę uzyskanych wyników. Przeprowadzając badania, można uzyskać 1 z 4 sytuacji:

- istotność statystyczna: ($p < 0{,}05$) + silny, akceptowalny efekt,
- brak istotności statystycznej: ($p \geq 0{,}05$) + słaby, nieakceptowalny efekt,
- istotność statystyczna: ($p < 0{,}05$) + słaby, nieakceptowalny efekt,
- brak istotności statystycznej: ($p \geq 0{,}05$) + silny, akceptowalny efekt.

Dwie pierwsze sytuacje są łatwe do interpretacji. Tutaj poziom istotności i wielkość efektu są spójne. Natomiast sytuacja 3. i 4. mogą stanowić problem i prowadzić do błędów wnioskowania, jeśli będą się opierać jedynie na wartości p. W sytuacji 3. uzyskano istotność statystyczną, ale uzyskane wyniki mają niewielkie praktyczne znaczenie. Natomiast przykład 4. pokazuje, że wynik nie może zostać wykorzystany do wnioskowania o całej populacji.

Istnieje wiele wskaźników, które mogą być wykorzystane do oceny wielkości efektu, a testy statystyczne charakteryzować się będą różnymi wyliczonymi wielkościami siły efektu. W ➥ **tabeli 10.2** wymienione są najczęściej wykorzystywane miary efektu w odniesieniu do testów statystycznych omawianych w skrypcie.

Test statystyczny	**Miara siły efektu**	**Interpretacja**
Test t Studenta dla zmiennych niezależnych	d Cohena g Hedgesa δ Glassa	0,2 – efekt słaby 0,5 – efekt umiarkowany 0,8 – efekt silny
Test t Studenta dla zmiennych zależnych	d Cohena	0,2 – efekt słaby 0,5 – efekt umiarkowany 0,8 – efekt silny
Jednoczynnikowa analiza wariancji	η^2 – eta kwadrat ω^2 – omega kwadrat	0,01 – efekt słaby 0,06 – efekt umiarkowany 0,14 – efekt silny
Test χ^2	V Cramera wskaźnik fi (φ)	0,1 – efekt mały 0,3 – efekt przeciętny 0,5 – efekt duży
Test Wilcoxona	r_c – współczynnik korelacji dwuseryjnej dla par dopasowanych	0,1 – efekt słaby 0,3 – efekt umiarkowany 0,5 – efekt silny
Test U Manna–Whitneya	r_g – współczynnik korelacji dwuseryjnej Glassa	0,1 – efekt słaby 0,3 – efekt umiarkowany 0,5 – efekt silny
Test Kruskala–Wallisa	ε^2 – epsilon kwadrat	przyjmuje wartości od 0 (brak związku) do 1 (maksymalny związek)

Tabela 10.2. Miary siły efektu dla wybranych testów statystycznych

10.6. Zadania praktyczne

Jak w praktyce określić minimalną liczebność próby? Do tego celu wykorzystamy wcześniej zebrane wyniki badań (np. pochodzące z badań wstępnych lub z publikacji naukowych), ponieważ musimy znać średnią i odchylenie standardowe z próby, a także ustalić minimalną wielkość efektu akceptowaną przez badacza jako ważną – czy to z poznawczego, czy np. z klinicznego punktu widzenia. Dane do zadania znajdują się w ➥ **tabeli 10.3**.

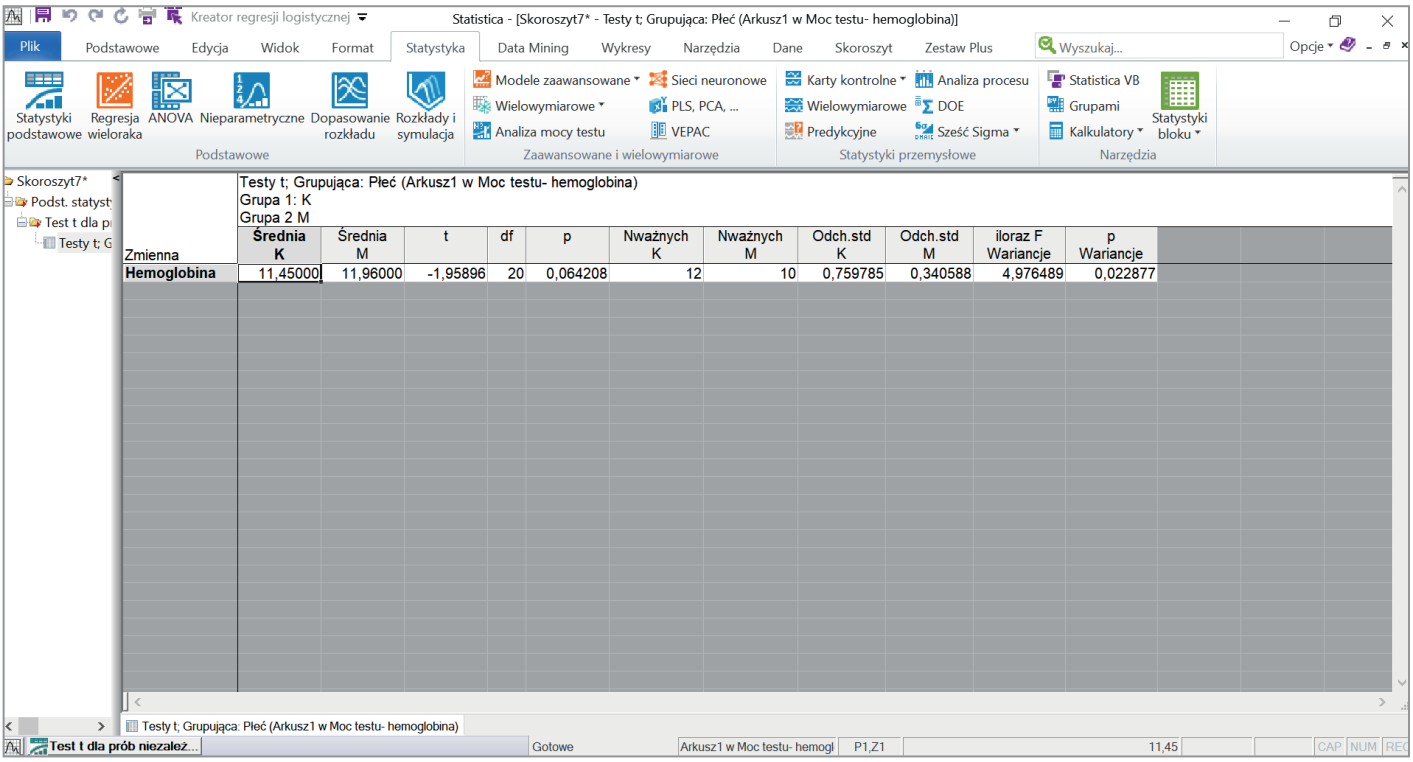
Chcemy ocenić różnicę stężenia hemoglobiny między 2 grupami – kobiet i mężczyzn. W poniższej tabeli znajdują się dane dla 12 kobiet i 10 mężczyzn. Obserwujemy zatem 2 niezależne grupy pacjentów, gdzie zmienną grupującą jest „PŁEĆ".

Kobiety	12,0	12,3	11,8	11,3	12,7	11,3	11,2	11,9	11,6	10,3	10,2	10,8
Mężczyźni	11,7	11,8	12,3	11,5	12,1	12,2	11,6	11,7	12,5	12,2		

Tabela 10.3. Wartości stężenia hemoglobiny – dane do zadania

Do porównania średnich poziomów hemoglobiny u kobiet i mężczyzn wykorzystamy test *t* Studenta dla prób niezależnych (dokładna procedura opisana została w rozdziale 12. *Testy* t *i ich nieparametryczne odpowiedniki*). W wyniku procedury otrzymujemy arkusz wynikowy (➥ **ryc. 10.4**).

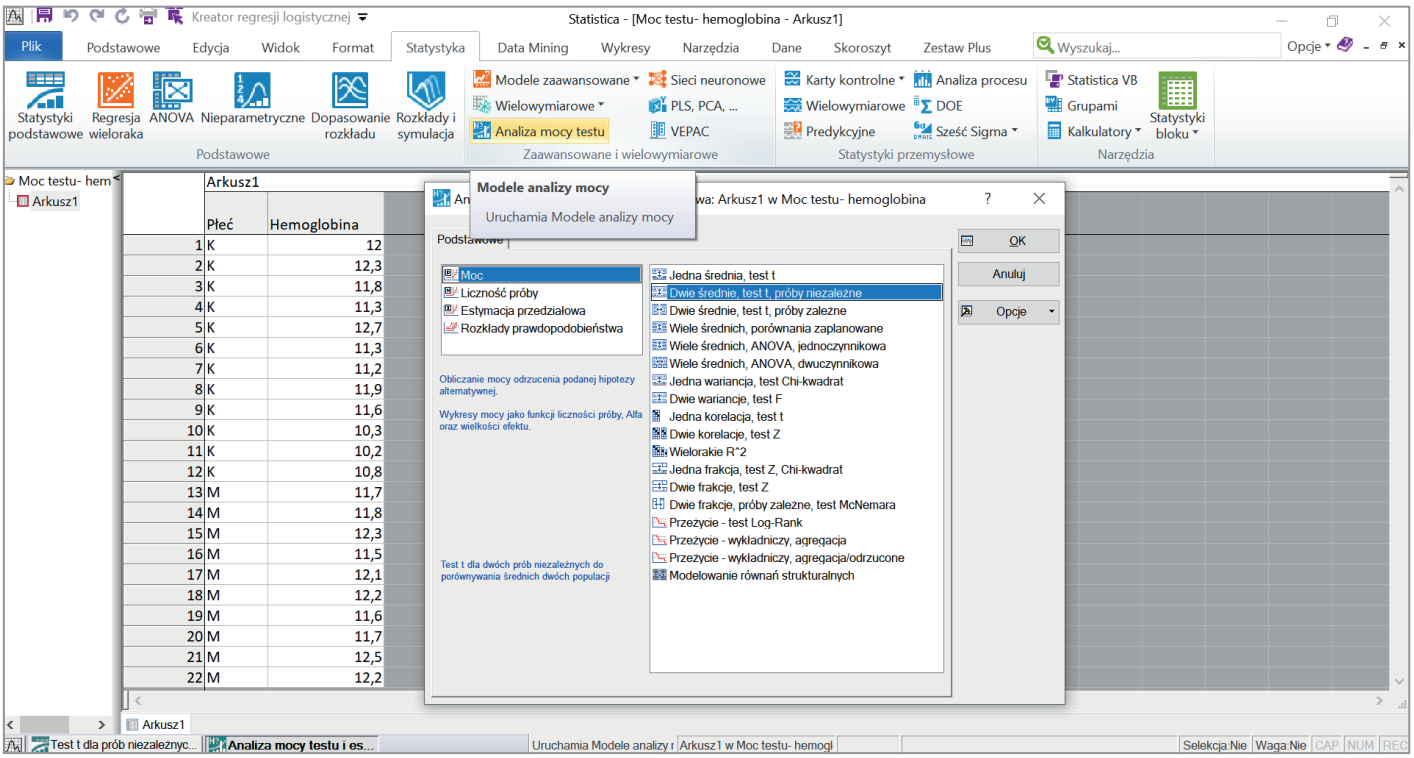
Ryc. 10.4. Wyniki testu *t* Studenta dla grup niezależnych



Po analizie wyników można dostrzec, że stężenie hemoglobiny u mężczyzn jest wyższe niż u kobiet (u kobiet średnia wynosi 11,45 z odchyleniem standardowym równym 0,76, a u mężczyzn średnia to 11,96 z odchyleniem standardowym 0,34). Różnica pomiędzy średnią dla kobiet i mężczyzn (średnia 1 – średnia 2) jest ujemna i wynosi –0,51 z przedziałem ufności 95% [–1,05–0,03]. Różnica ta nie jest jednak wystarczająca, aby mówić o istotności statystycznej ($p < 0,05$), pomimo że wynik znajdował się na jej granicy. Moc tego testu była niewystarczająca.

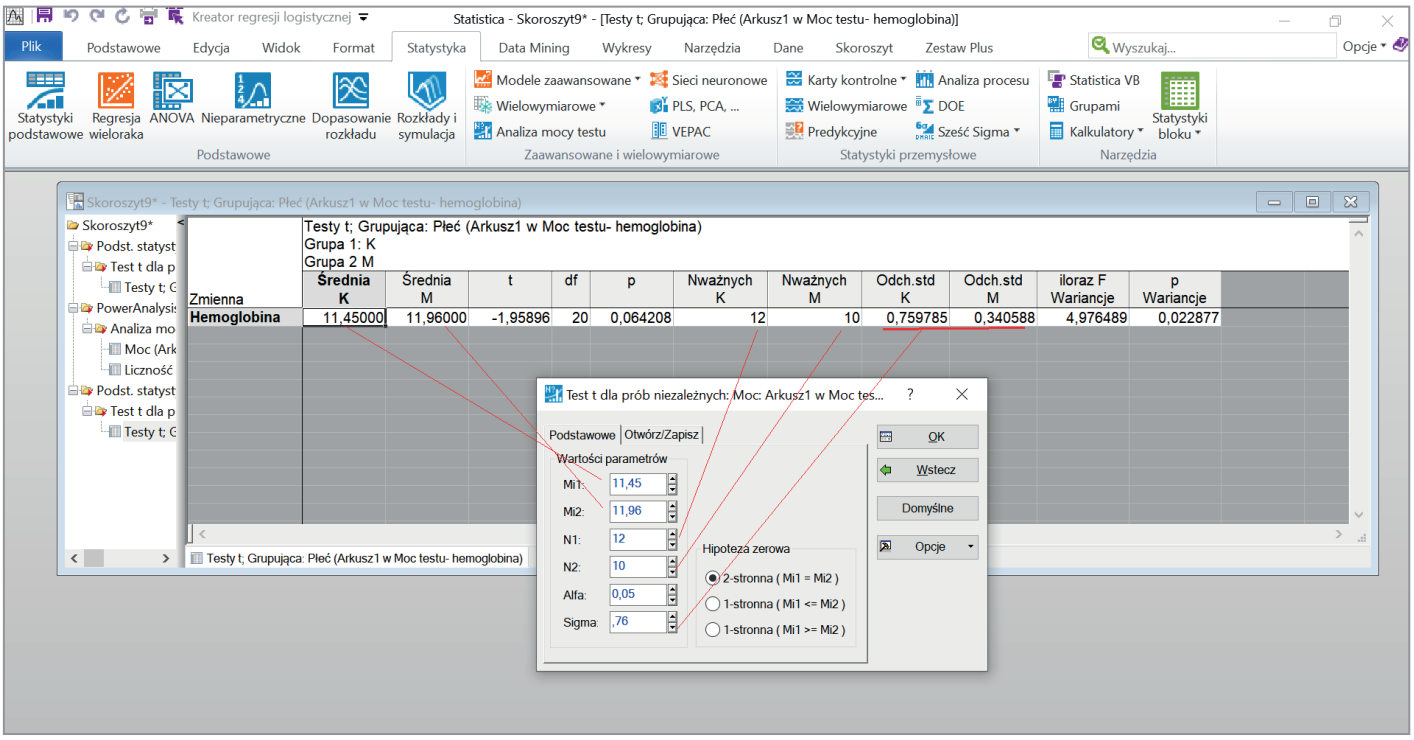
Ryc. 10.5. Analiza mocy testu dla 2 prób niezależnych

Sprawdźmy, jak duża jest moc testu dla naszego przykładu. W pakiecie Statistica możemy wykorzystać moduł Analiza mocy testu, gdzie wybieramy następnie typ analizy, tj. moc testu lub minimalną liczebność próby (➥ **ryc. 10.5**).



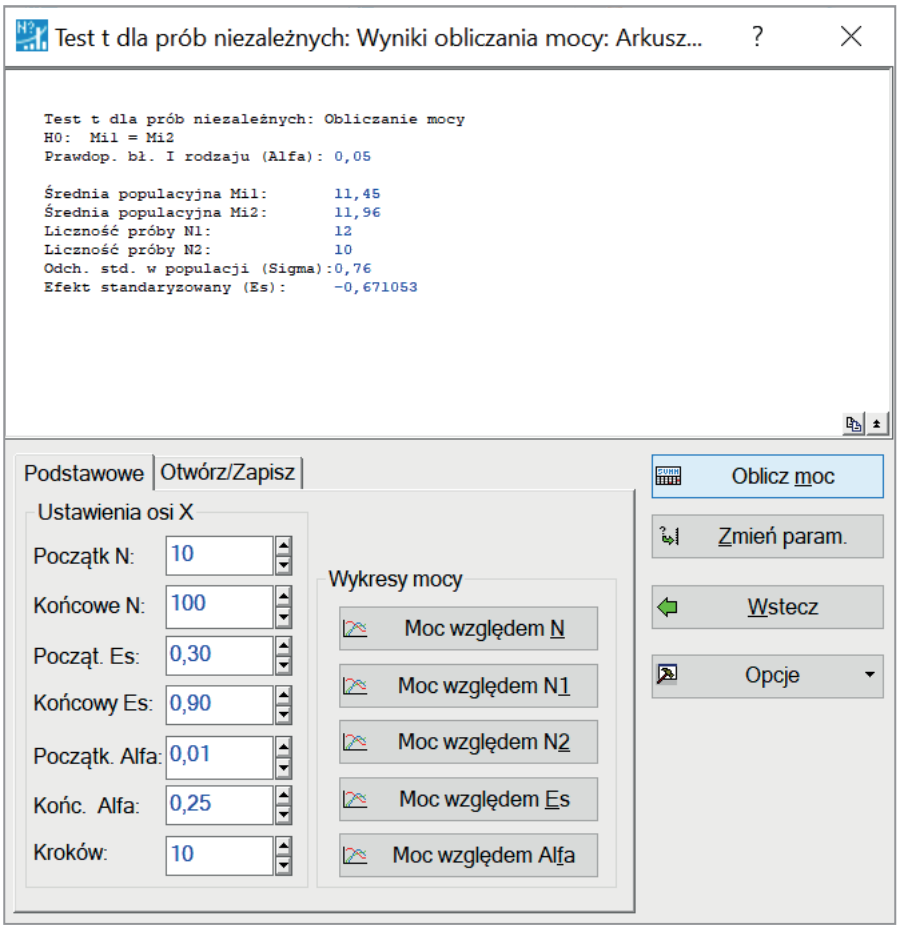
Ryc. 10.6. Obliczanie mocy testu dla prób niezależnych

W nowym oknie należy uzupełnić dane dotyczące średnich i wyższe z obliczonych odchyleń standardowych wg. rezultatów testu *t* Studenta. Wybieramy wyższą wartość odchylenia, ponieważ będzie to podejście bezpieczne (im mniejsze *SD* tym, łatwiej odrzucić H_0). Wartość α zostawiamy na poziomie 0,05 (➥ **ryc. 10.6**).



Po wypełnieniu wskazanych pól klikamy OK, a następnie w nowym oknie klikamy Oblicz moc (➡ **ryc. 10.7**).

Ryc. 10.7. Obliczanie mocy testu dla prób niezależnych



Efektem analizy jest skoroszyt wynikowy z obliczoną mocą testu (➡ **ryc. 10.8**).

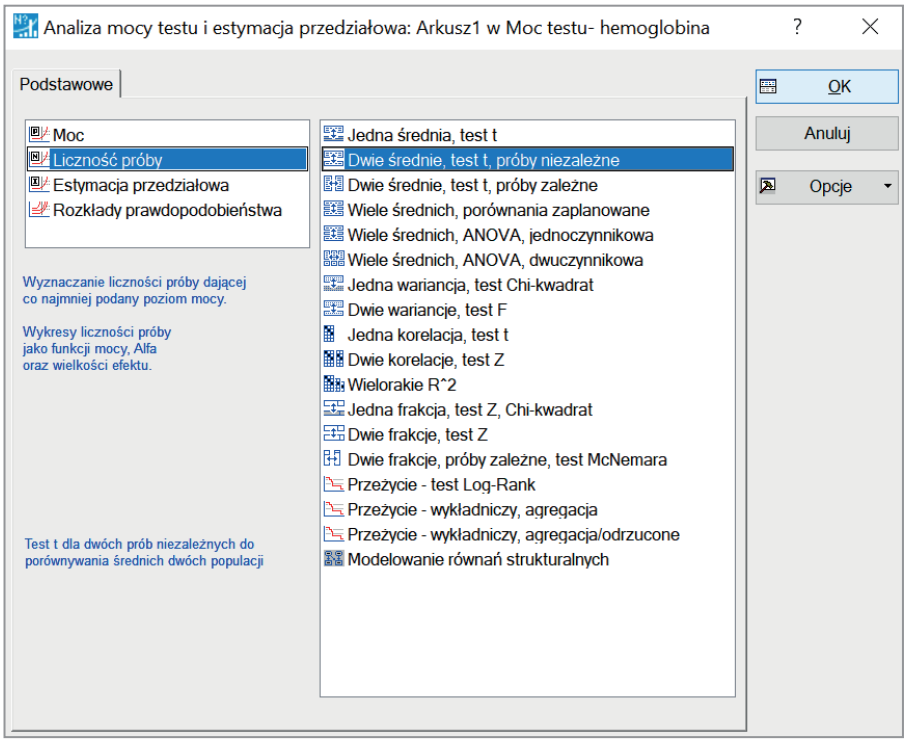
Ryc. 10.8. Moc testu dla prób niezależnych

	Moc (Arkusz1 w Moc testu- hemoglobina) Dwie średnie, test t, próby niezależne H0: Mi1 = Mi2
	Wartość
Średnia populacyjna Mi1	11,4500
Średnia populacyjna Mi2	11,9600
Odch. std. w populacji (Sigma)	0,7600
Efekt standaryzowany (Es)	-0,6711
Liczność próby N1	12,0000
Liczność próby N2	10,0000
Prawdop. bł. I rodzaju (Alfa)	0,0500
Wartość krytyczna t	2,0860
Moc	0,3203

Dla naszego przykładu, porównującego stężenie hemoglobiny u kobiet i mężczyzn, moc testu wynosi 0,32, czyli 32%. Jest ona wyraźnie niższa od przyjmowanego zazwyczaj poziomu 0,80 (80%). W takiej sytuacji należy sprawdzić, jak dużą liczebnością grup powinniśmy dysponować, aby przy otrzymanych powyżej wynikach można było wykazać istotność statystyczną na poziomie $\alpha = 0{,}05$.

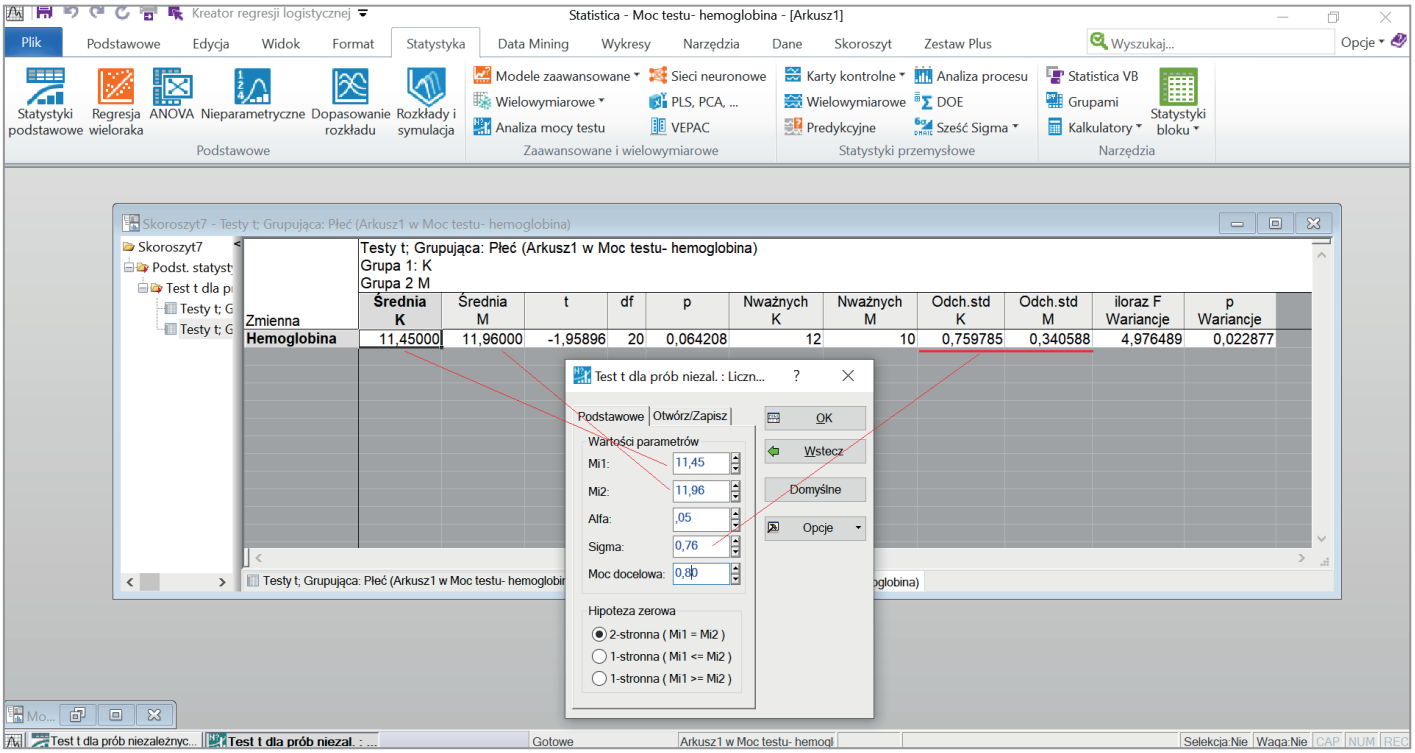
Obliczenia minimalnej liczebności prób możemy dokonać w tym samym module, w którym obliczaliśmy moc testu. Otwieramy zatem zakładkę Statystyka, wybieramy Analiza mocy testu, a następnie Liczebność próby. Z listy dostępnych testów wybieramy Dwie średnie, test t, próby niezależne (➥ **ryc. 10.9**).

Ryc. 10.9. Obliczanie liczebności próby



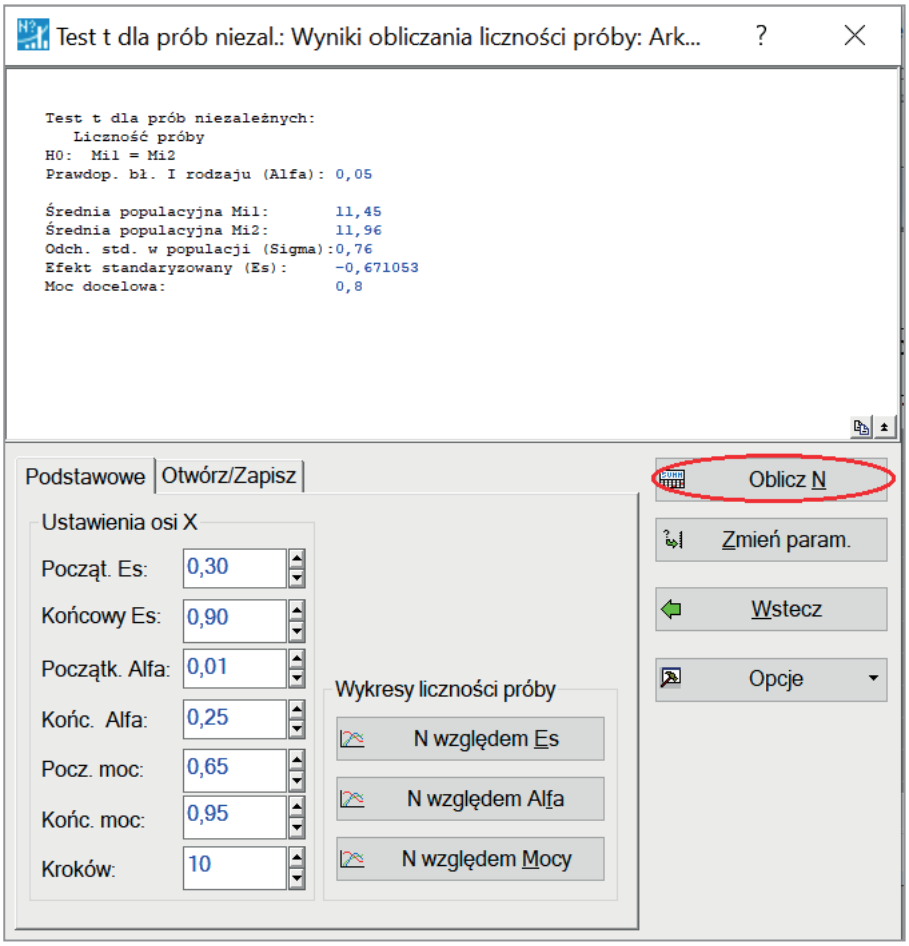
W nowym oknie należy uzupełnić dane dotyczące średnich i wyższe z obliczonych odchyleń standardowych wg rezultatów testu *t* Studenta (➥ **ryc. 10.10**). Korzystamy z odchylenia wyższego z dwóch dostępnych, ponieważ jest to podejście bezpieczne (im mniejsze *SD*, tym łatwiej odrzucić H_0). Wartość α zostawiamy na 0,05, a moc testu przyjmujemy na poziomie 0,80. Po wprowadzeniu danych klikamy OK.

Ryc. 10.10. Minimalna liczebność próby



W nowym oknie klikamy Oblicz N (➡ **ryc. 10.11**).

Ryc. 10.11. Obliczanie minimalnej liczebności próby



W wyniku analizy otrzymujemy arkusz wynikowy (➡ **ryc. 10.12**).

Ryc. 10.12. Minimalna liczebność próby

	Liczność próby (Arkusz1 w Moc testu- hemoglobina) Dwie średnie, test t, próby niezależne H0: Mi1 = Mi2
	Wartość
Średnia populacyjna Mi1	11,4500
Średnia populacyjna Mi2	11,9600
Odch. std. w populacji (Sigma)	0,7600
Efekt standaryzowany (Es)	-0,6711
Prawdop. bł. I rodzaju (Alfa)	0,0500
Wartość krytyczna t	1,9944
Moc docelowa	0,8000
Moc dla wymaganej liczności próby N	0,8017
Wymagane N (w grupie)	36,0000

Liczność próby (Arkusz1 w Moc testu- hemoglobina)

Z przedstawionej analizy wynika, że aby moc testu była na akceptowalnym dla nas poziomie, liczebność w obu próbach powinna wynosić po 36 obserwacji.

Jeśli do wykonania analizy możemy użyć kilku różnych testów, do wnioskowania w wynikach doświadczenia zawsze należy użyć tego o wyższej mocy.

Agnieszka Rusiecka

11 Tabele wielodzielcze i analiza ryzyka w medycynie

11.1. Zmienne jakościowe

W badaniach medycznych mamy do czynienia nie tylko z wartościami mierzalnymi (np. masa ciała pacjentów, stężenie glukozy itp.), ale analizujemy także dane jakościowe. Dla przypomnienia, zmienne jakościowe bazują na słownym opisie. Dane te są pozyskiwane w badaniach empirycznych przy wykorzystaniu ankiet, wywiadów czy obserwacji. Zmienne jakościowe dzielimy na 2 kategorie:

- nominalne, kiedy wartości reprezentują kategorie bez wewnętrznego rangowania, a ich poszczególne kategorie są rozłączne; przykładami zmiennych mogą być grupa krwi (0, A, B, AB) czy nazwa oddziału, na którym przebywają pacjenci (oddział pediatryczny, reumatologiczny, nefrologiczny); zmienne te mogą mieć również charakter binarny (dwumianowe – stan pacjenta: zdrowy, chory; płeć: kobieta, mężczyzna);
- porządkowe, kiedy wartości zmiennych reprezentują kategorię z wewnętrznym rangowaniem, czyli dają się w sposób sensowny uporządkować; przykładem takiej zmiennej jest skala mierząca poziom bólu (brak, słaby, średni, silny) albo zadowolenia z usługi (niezadowolony, średnio zadowolony, bardzo zadowolony).

11.2. Test χ^2 Pearsona

Dysponując danymi typu jakościowego, za cel najczęściej obierzemy ocenę zależności między zmiennymi, np. zależność pomiędzy występowaniem cukrzycy a miejscem zamieszkania pacjentów. W jaki sposób dokonać analizy tego typu danych?

Pierwszym krokiem powinno być zebranie wszystkich danych i zliczenie ich w poszczególnych kategoriach przy wykorzystaniu tabeli wielodzielczej. Liczebności w ostatnim wierszu i w ostatniej kolumnie tabeli nazywamy sumami lub rozkładami brzegowymi (➥ **tabela 11.1**).

Tabela wielodzielcza stanowi kombinację dwóch lub więcej tabel liczności w taki sposób, by przedstawiała rozkład liczności wszystkich powstałych podgrup. Analizując te liczności, można określić relacje, jakie zachodzą między wskazanymi zmiennymi. Żeby zrozumieć kolejne etapy stosowania testu χ^2, w pierwszej kolejności rozwiążemy zadania bez użycia pakietu Statistica.

Przeanalizujmy następujące zadanie.

W celu zbadania skuteczności 2 leków A i B przeprowadzono eksperyment na 2 jednorodnych pod względem stanu zaawansowania choroby grupach o liczebności 50 i 75 osób. Po leczeniu zbadano stan chorych. Wyleczono 12 spośród 50 osób leczonych lekiem A, a spośród 75 osób leczonych lekiem B wyleczono 30.

Najpierw skupimy się na utworzeniu tabeli wielodzielczej podsumowującej liczności obu zmiennych, wraz z sumami brzegowymi (➥ **tabela 11.1**). Możemy ją nazwać tabelą danych obserwowanych. Ze względu na to, że każda z naszych

zmiennych reprezentuje dwie kategorie (zmienna „Lek": A lub B; zmienna „Stan chorych": wyleczony, niewyleczony), otrzymamy najmniejszą z możliwych tabel wielodzielczych, tj. tabelę dwudzielczą albo inaczej – często spotykane określenie – tabelę 2 × 2.

Wartości obserwowane

	Wyleczony	Niewyleczony	
Lek A	12	38	50
Lek B	30	45	75
	42	83	125

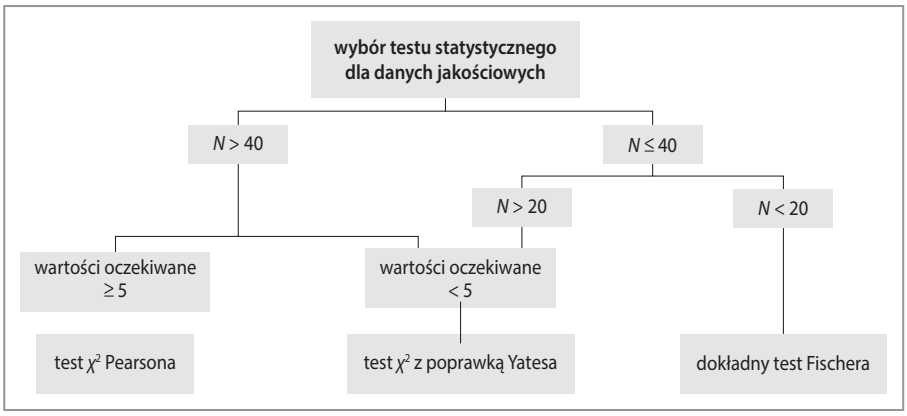
Tabela 11.1. Tabela dwudzielcza

Wynika z niej, że lek B jest skuteczniejszy niż lek A.

Należy jednak rozważyć, czy różnica ta jest istotna statystycznie, tj. czy na tej podstawie możemy sądzić, że lek B jest skuteczniejszy dla całej populacji.

Następny etap analizy zakłada więc testowanie hipotezy statystycznej, a co za tym idzie, wybór odpowiedniego testu statystycznego. Ponieważ nasze próby są niezależne (pacjenci leczeni lekiem A to inna grupa niż ta leczona lekiem B), możemy posłużyć się schematem dla prób niezależnych (➥ **ryc. 11.1**). Wybór jest dość prosty – do zmiennych jakościowych należy użyć testu χ^2.

Ryc. 11.1. Schemat wyboru testu statystycznego dla tabel wielodzielczych



Jego twórcą był Karl Pearson, który zaproponował metodę pozwalającą sprawdzić, czy istnieje związek między zmiennymi zawartymi w tabeli wielodzielczej. Statystyka χ^2 jest jednak często modyfikowana w celu lepszego dostosowania funkcji matematycznych do analizowanych danych (poprawki do testu χ^2 zostaną omówione w podrozdziale 11.3 *Test χ^2 z poprawką Yatesa*).

Hipoteza zerowa dla testu χ^2 mówi, że nie ma zależności między badanymi zmiennymi. W naszym zadaniu przykładowa H_0 może zostać sformułowana w taki sposób:

H_0: Leki A i B są tak samo skuteczne w leczeniu choroby.

Hipoteza alternatywna, która stanowi przeciwieństwo dla H_0, może brzmieć:

H_1: Skuteczność leków A i B w leczeniu choroby różni się.

Test χ^2 opiera się na porównaniu częstości obserwowanych (➥ **tabela 11.2**) z częstościami oczekiwanymi wyliczonymi przy założeniu, że H_0 jest prawdziwa (a więc nie ma zależności między badanymi zmiennymi).

Częstości oczekiwane obliczamy, wykorzystując sumy brzegowe z tabeli wielodzielczej wg wzoru:

$$\text{wartość oczekiwana} = \frac{(\text{suma kolumny}) \times (\text{suma wiersza})}{(\text{suma całkowita})}$$

Dla naszego zadania tabela oczekiwana będzie się prezentowała następująco:

Wartości oczekiwane

	Wyleczony	Niewyleczony	
Lek A	$\frac{(42 \times 50)}{125} = 16{,}8$	$\frac{(83 \times 50)}{125} = 33{,}2$	50
Lek B	$\frac{(42 \times 75)}{125} = 25{,}2$	$\frac{(83 \times 75)}{125} = 49{,}8$	75
	42	83	125

Tabela 11.2. Tabela wartości oczekiwanych

Porównania tabeli wartości obserwowanych oraz wartości oczekiwanych dokonamy przy użyciu wzoru:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

gdzie O oznacza wartości z tabeli wartości obserwowanych, a E – wartości z tabeli wartości oczekiwanych.

Łatwo zauważyć, że gdyby między badanymi zmiennymi nie było żadnej zależności, to między tabelą wartości obserwowanych a tabelą wartości oczekiwanych nie powinny występować duże rozbieżności. Kiedy zaś stan rzeczywisty zaczyna się różnić od rozkładu teoretycznego (wartości oczekiwanych), wzrasta wartość χ^2. W naszym zadaniu funkcja ta będzie wyglądała następująco:

$$\chi^2 = \frac{(12 - 16{,}8)^2}{16{,}8} + \frac{(38 - 33{,}2)^2}{33{,}2} + \frac{(30 - 25{,}2)^2}{25{,}2} + \frac{(45 - 49{,}8)^2}{49{,}8} = 3{,}44$$

Uzyskany wynik to wartość naszej statystyki χ^2. Czy 3,44 to dużo czy mało? Żeby to sprawdzić, musimy sięgnąć do tablic statystycznych dla rozkładu χ^2. By z nich skorzystać, należy jeszcze poznać liczbę stopni swobody *df*. W przypadku testu niezależności χ^2 obliczamy ją ze wzoru

$$df = (k - 1)(w - 1)$$

gdzie k i w to ilość poziomów zmiennych, dla których obliczamy χ^2, czyli innymi słowy liczba kolumn i liczba wierszy z kategoriami w naszej tabeli wielodzielczej. W tym przypadku liczba stopni swobody wynosi $df = (2 - 1)(2 - 1) = 1$. Przyjęliśmy poziom istotności $\alpha = 0{,}05$. Wartość $\chi^2 = 3{,}44$ odpowiada wartości $p = 0{,}06$. Tym samym przyjmujemy H_0, która mówi, że nie ma zależności między stanem pacjenta a stosowanym lekiem.

11.3. Test χ^2 z poprawką Yatesa

Jak wspomniano wcześniej, pierwotna formuła testu χ^2 Pearsona jest często modyfikowana i dostosowywana do danych, którymi dysponujemy. Na poprzedniej stronie przedstawiono poglądowy schemat zastosowania poprawek dla testu χ^2 Pearsona (➥ **ryc. 11.1**). Warto jednak zaznaczyć, że podlegał on krytyce. Dziś uważa się, że poprawka Yatesa jest zbyt konserwatywna, a dla małych prób (tj. gdy w co najmniej 20% komórek liczność oczekiwana jest mniejsza niż 5) lepiej używać testu Fishera.

Wzór funkcji χ^2 z poprawką Yatesa przedstawia się następująco:

$$\chi^2 = \sum \frac{(|O_i - E_i| - 0{,}5)}{E_i}$$

Poprawkę Yatesa stosujemy, gdy liczebność badanej próby to $20 < N < 40$ oraz gdy którakolwiek z liczebności oczekiwanych jest mniejsza od 5. Poprawka Yatesa jest testem bardziej konserwatywnym niż test χ^2 Pearsona, co oznacza, że rzadziej niż w tym teście odrzuca H_0.

11.4. Test χ^2 w pakiecie Statistica

11.4.1. Rozwiązanie na podstawie danych wprowadzonych do arkusza

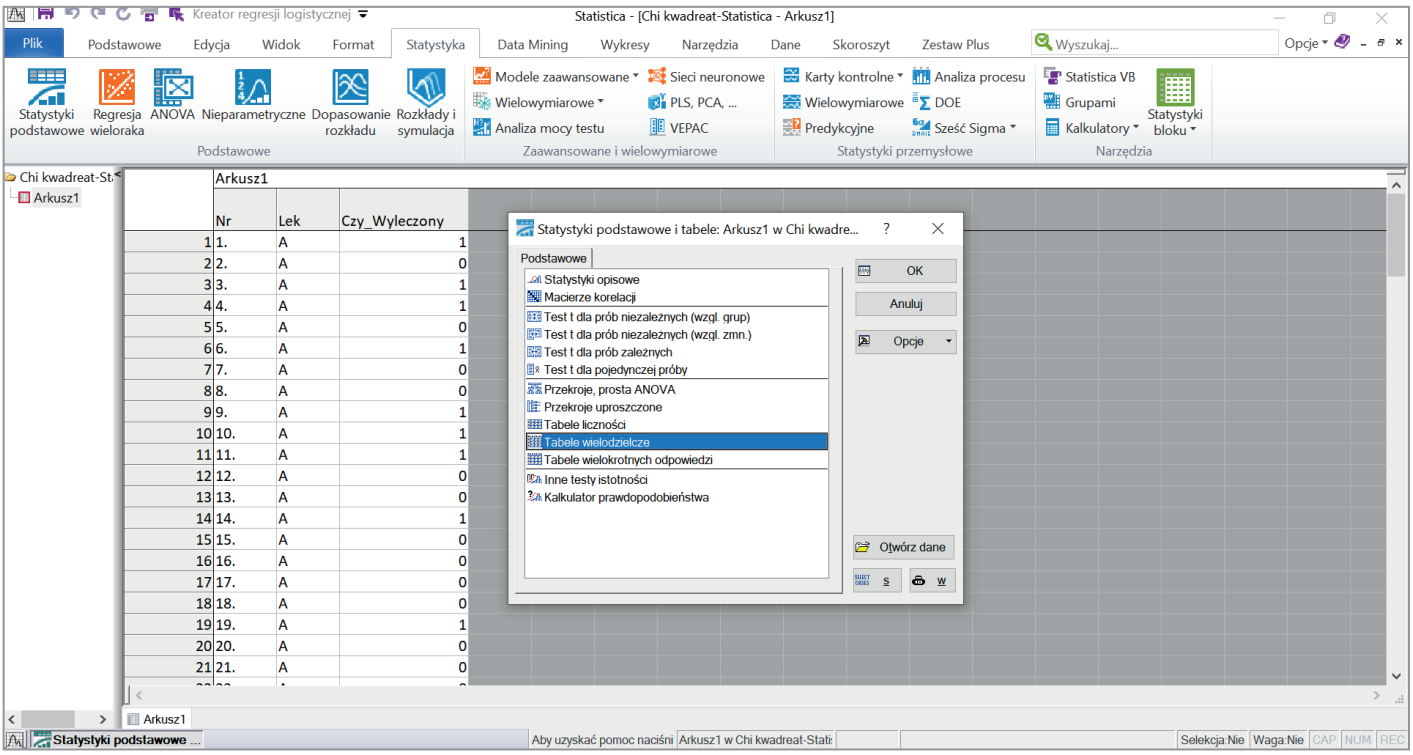
Rozwiążmy zadanie jeszcze raz, tym razem wykorzystując odpowiednie funkcje programu Statistica. Możemy to zrobić na 2 sposoby, korzystając z różnych kart. Dane dotyczące pacjentów najczęściej zebrane są w arkuszach kalkulacyjnych i w takiej formie mogą też zostać poddane analizie. Wprowadzamy je więc do arkusza przypadek po przypadku zgodnie z treścią zadania, tak by 12 spośród 50 osób leczonych lekiem A oznaczyć jako wyleczone, a spośród 75 osób leczonych lekiem B oznaczyć w ten sposób 30 osób (wg schematu z ➥ **tabeli 11.3**).

ID pacjenta	Lek	Czy_Wyleczony
1	A	1
2	A	0
3	B	1
4	B	1
5	A	0
6	B	1
7	A	0
…	…	…

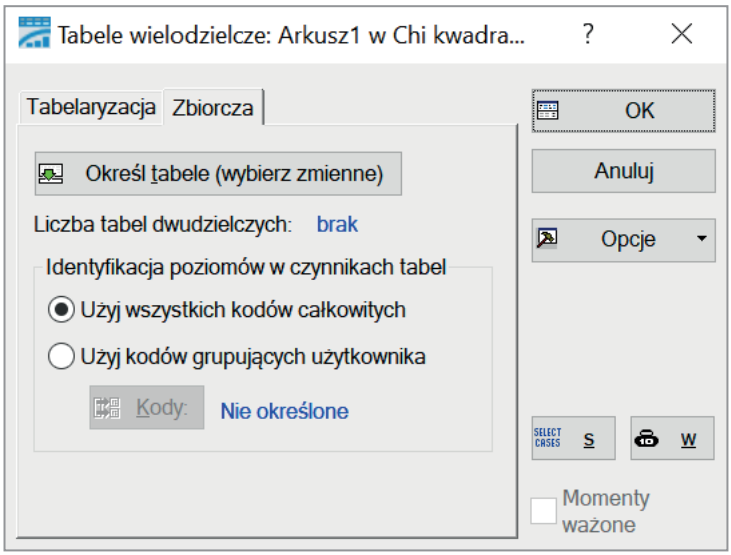
Tabela 11.3. Fragment arkusza obejmującego informacje o stosowanym leku i stanie pacjenta

W programie Statistica do analizy tabel wielodzielczych służy opcja Tabele wielodzielcze. Wybieramy polecenie Statystyki podstawowe, a następnie Statystyki opisowe i Tabele wielodzielcze (➥ **ryc. 11.2**).

Ryc. 11.2. Wybór metody Tabele wielodzielcze

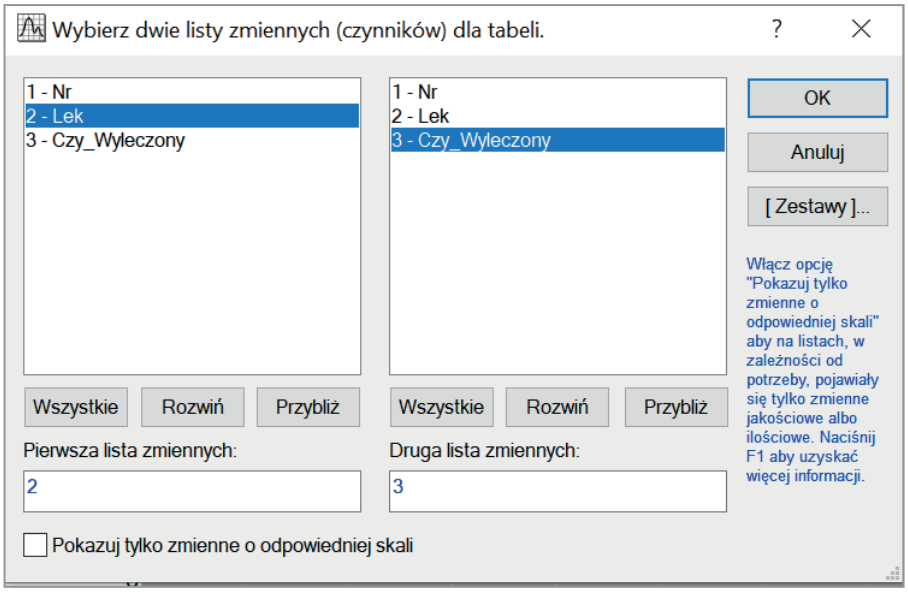


Przechodzimy do okna Tabele wielodzielcze. W module tym możemy wybrać 2 grupy statystycznych analiz dotyczących tablic zbiorczych oraz tablic wielokrotnych odpowiedzi. Interesujące nas opcje zawiera zakładka Zbiorcza (➥ **ryc. 11.3**).



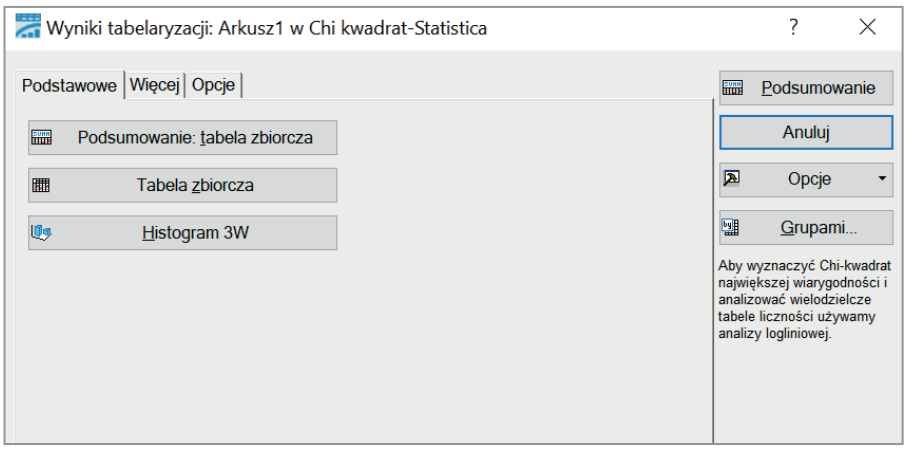
Ryc. 11.3. Wybór metody oraz zmiennych w karcie Tabele wielodzielcze

W dalszej kolejności wybieramy zmienne. Po kliknięciu przycisku Zmienne otworzy się okno z 2 listami zmiennych. Jedna z nich jest tabelaryzowana w kolumnach, a druga – w wierszach (➥ **ryc. 11.3**). Klikamy OK i po powrocie do wcześniejszego okna (➥ **ryc. 11.4**) jeszcze raz OK.



Ryc. 11.4. Wybór zmiennych dla tabel wielodzielczych

Przechodzimy do następnego okna z możliwością wyboru odpowiednich opcji i metod podsumowania analizy (➥ **ryc. 11.5**).



Ryc. 11.5. Wybór opcji i metod podsumowania tabel wielodzielczych

W tym oknie znajdują się 3 zakładki, w których możemy dokonać wyboru opcji analizy. W zakładce Podstawowe przy użycia przycisku Podsumowanie: tabela zbiorcza możemy utworzyć tabelę dwudzielczą (➥ **ryc. 11.6**).

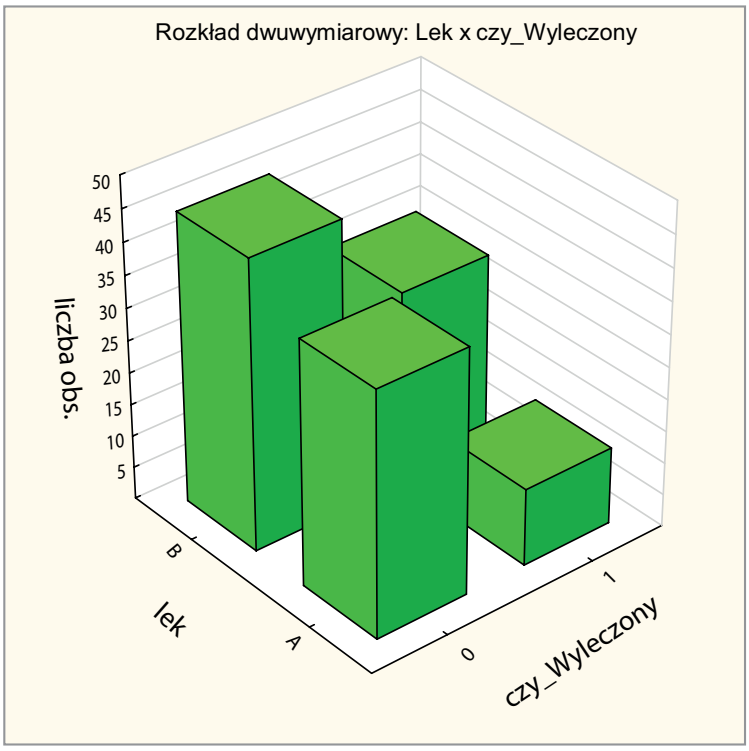
Ryc. 11.6. Tabela dwudzielcza

Tabela liczności (Arkusz1 w Chi kwadrat-Statistica)
Liczność oznacz. komórek > 10
(Nie oznaczono sum brzegowych)

Lek	Czy_Wyleczony 0	Czy_Wyleczony 1	Wiersz Razem
A	38	12	50
B	45	30	75
Ogół	83	42	125

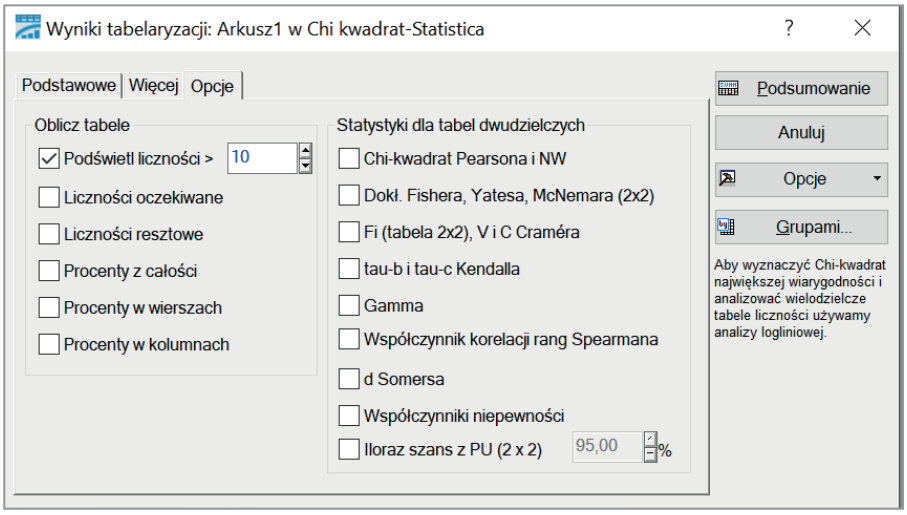
Jeśli do analizy wybraliśmy więcej niż 2 zmienne, możemy również utworzyć tabelę zbiorczą (po kliknięciu przycisku Tabela zbiorcza). Wygląda ona jak połączenie wielu tabel dwudzielczych. To funkcja przydatna, kiedy chcemy przejrzeć wszystkie podsumowania w jednym miejscu. Dodatkowo dla wskazanych tabel możemy utworzyć trójwymiarowy histogram (przycisk Histogram 3W, ➥ **ryc. 11.7**).

Ryc. 11.7. Histogram 3W



Żeby obliczyć statystykę testową dla naszego rozkładu danych, należy dokonać wyboru odpowiednich opcji, dlatego po wstępnej analizie wykonanej w zakładce Podstawowe przechodzimy do zakładki Opcje (➥ **ryc. 11.8**).

Ryc. 11.8. Opcje wyboru analizy dla tabel wielodzielczych



Karta Opcje podzielona jest na 2 części: Oblicz tabele oraz Statystyki dla tabel dwudzielczych.

W pierwszej części można dokonać wyboru wyświetlania wyników. Do wyboru mamy następujące opcje:

- podświetlenie liczności – umożliwia podświetlenie w tabeli wszystkich wartości innych niż zadana; domyślnie ustawiona jest wartość 10, a więc na czerwono podświetlane będą wszystkie wyniki > 10;
- liczności oczekiwane – umożliwia wyświetlenie w każdej komórce, oprócz wartości obserwowanych, również wartości oczekiwanych;
- liczności resztowe – w każdej komórce zostaną wyświetlone wartości będące różnicą między wartościami obserwowanymi a oczekiwanymi;
- procent z całości – w każdej komórce wyświetlane są wartości procentowe obliczane względem całkowitej liczności próby;
- procenty w wierszach – w każdej komórce wyświetlane są wartości procentowe obliczone z całkowitej liczebności w danym wierszu;
- procenty w kolumnach – w każdej komórce wyświetlane są wartości procentowe obliczone z całkowitej liczebności w danej kolumnie.

Wybór zależy od użytkownika. Najczęściej interesują nas wartości procentowe, dlatego dobrze jest zaznaczyć wszystkie dotyczące ich opcje (procenty z całości, procenty z wiersza, procenty z kolumny, ➥ **ryc. 11.9**). Taki układ umożliwia dokładną analizę danych.

Ryc. 11.9. Tabela dwudzielcza z przykładowymi opcjami wyświetlania

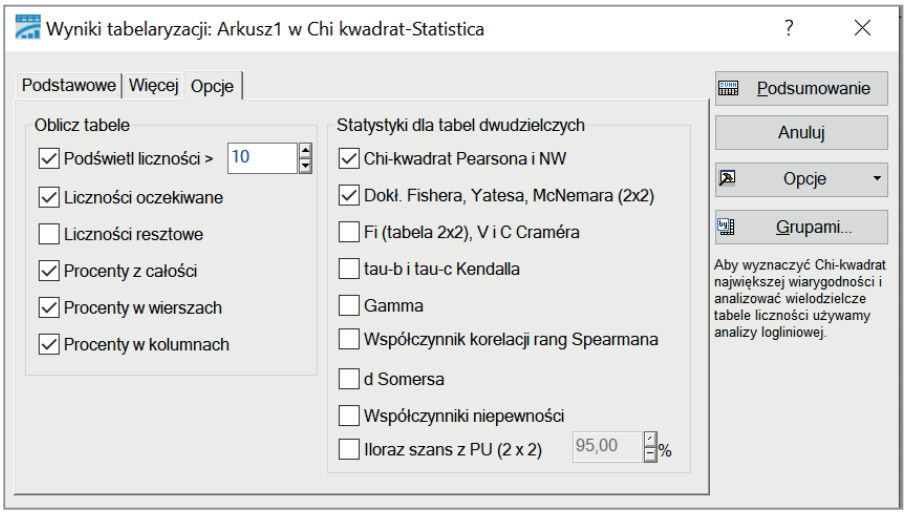
Lek	Podsumowująca tabela dwudzielcza: częstości observ Liczność oznacz. komórek > 10		
	Czy_Wyleczony 0	Czy_Wyleczony 1	Wiersz Razem
A	38	12	50
%kolumny	45,78%	28,57%	
%wiersza	76,00%	24,00%	
% ogółu	30,40%	9,60%	40,00%
B	45	30	75
%kolumny	54,22%	71,43%	
%wiersza	60,00%	40,00%	
% ogółu	36,00%	24,00%	60,00%
Ogół	83	42	125
% ogółu	66,40%	33,60%	100,00%

Druga część karty Opcje, Statystyki dla tabel dwudzielczych (➥ **ryc. 11.8**), umożliwia wybór statystyk, które pozwolą opisać zależności między zmiennymi. W tej części można wybrać wiele testów statystycznych, jednak omówione zostaną tylko wybrane, najczęściej stosowane możliwości:

- test niezależności χ^2 Pearsona umożliwia obliczenie statystyki, która szczegółowo została opisana w rozdziale 11.2;
- test χ^2 największej wiarygodności (NW) sprawdza tę samą hipotezę co χ^2 Pearsona, jednak różni się drogą dochodzenia do wyniku, w tym wypadku obliczenia opierają się, jak wskazuje nazwa, na teorii największej wiarygodności; najczęściej wyniki są bardzo podobne do χ^2 Pearsona;
- test χ^2 z poprawką Yatesa – poprawkę tę stosuje się w przypadku małych tabel 2×2, jeżeli liczebności oczekiwane są mniejsze niż 5;
- dokładny test Fishera w programie Statistica jest obliczany tylko dla tabel 2×2; stosowany jest dla bardzo małych liczebności całkowitych i małej liczby wartości oczekiwanych; wyniki są podane dla testu jedno- i dwustronnego.

Procedura wyboru odpowiedniego testu przedstawiona jest na ➥ **ryc. 11.1**.

Wróćmy do naszego zadania. Ze względu na to, że dysponujemy w nim tabelą dwudzielczą, w której całkowita liczebność badanych jest większa niż 40, a wartości oczekiwane w żadnej kategorii nie są mniejsze niż 5, możemy zastosować test χ^2 Pearsona. Dla porównania zaznaczmy również opcję z poprawką Yatesa (➥ **ryc. 11.10**).



Ryc. 11.10. Okno z zaznaczonymi opcjami analizy

Żeby po zaznaczeniu wybranych opcji wygenerować wyniki analizy, należy skorzystać z zakładki Więcej. Na niej, oprócz przycisków Podsumowanie: tabela zbiorcza oraz Tabela zbiorcza (zostały omówione wcześniej), znajduje się również przycisk Dokładne tabele dwudzielcze, po kliknięciu którego zostanie wygenerowana dokładna tabela dwudzielcza (➥ **ryc. 11.8**) oraz tabela z wynikami analizy statystycznej (➥ **ryc. 11.11**).

statystyka	Statystyka: Lek(2) x Czy_Wyleczony		
	Chi-kwadr.	df	p
Chi^2 Pearsona	3,442341	df=1	p=,06355
Chi^2 NW	3,526903	df=1	p=,06038
Chi^2 Yatesa	2,762538	df=1	p=,09650
dokł. Fishera, 1-stronny			p=,04715
2-stronny			p=,08219
Chi^2 McNemara (A/D)	,7205882	df=1	p=,39595
(B/C)	17,96491	df=1	p=,00002

Ryc. 11.11. Wyniki analizy statystycznej dla tabeli dwudzielczej

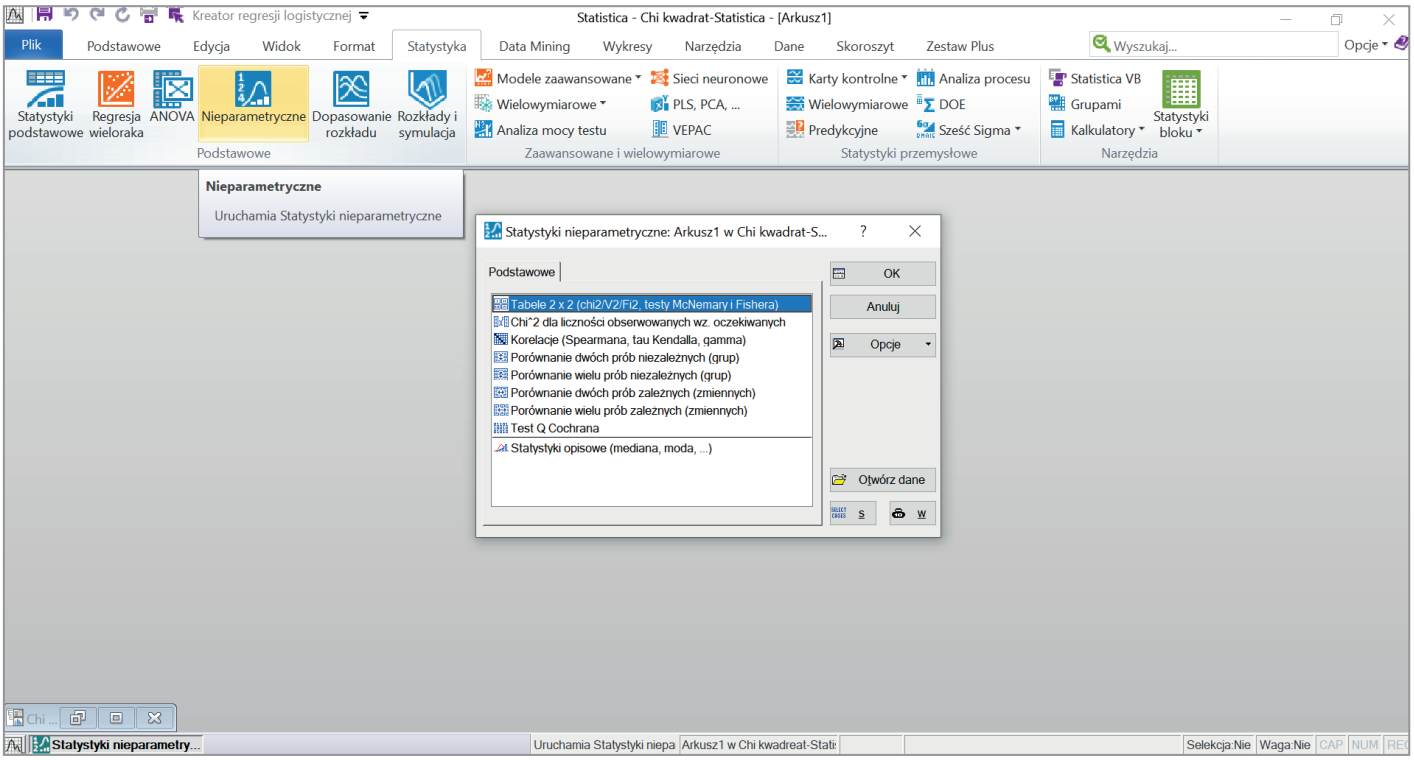
Z powyższej analizy wynika, że zarówno wartość testu χ^2 Pearsona, jak i χ^2 z poprawką Yatesa są większe niż przyjęty próg istotności statystycznej 0,05. W takiej sytuacji nie mamy podstaw do odrzucenia H_0, a więc nie ma zależności między stanem pacjenta a stosowanym lekiem.

11.4.2. Obliczenia na podstawie znanych liczebności

W **podrozdziale 11.4.1** przedstawiono sposób obliczenia testu χ^2 na podstawie danych zgromadzonych w arkuszu kalkulacyjnym. Najczęściej takie arkusze zawierają wiele zmiennych, które opisują kolejnych pacjentów. Czasami jednak dysponujemy samymi liczebnościami grup. W takiej sytuacji również można wykonać analizę testem χ^2.

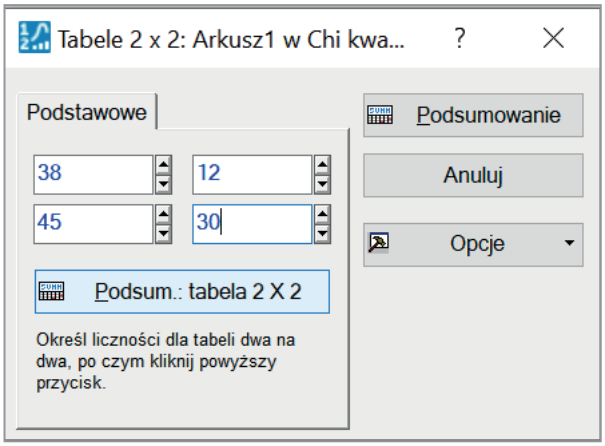
Ryc. 11.12. Wybór metody analizy tabel 2 × 2

Wystarczy wybrać kartę Statystyka, wybrać polecenie Nieparametryczne, a następnie z listy Tabela 2 × 2 (chi2/V2/Fi2, testy McNemary i Fishera) (➡ **ryc. 11.12**).



Po kliknięciu OK pojawi się okno dialogowe umożliwiające wprowadzenie liczebności w poszczególnych kategoriach. Skorzystajmy z danych z zadania w poprzednim podrozdziale (➡ **ryc. 11.13**).

Ryc. 11.13. Wprowadzone liczebności w tabeli 2 × 2



Klikamy w przycisk Podsum.: tabela 2 X 2 i otrzymujemy okno z wynikami analizy (➡ **ryc. 11.14**).

Ryc. 11.14. Wyniki analizy tabeli 2 × 2

	Tabela 2x2 (Arkusz1 w Chi kwadrat-Stat		
	Kolumna1	Kolumna2	Wiersz Razem
Liczności, wiersz	38	12	50
Procent całości	30,400%	9,600%	40,000%
Liczności, wiersz	45	30	75
Procent całości	36,000%	24,000%	60,000%
Razem w kol.	83	42	125
Procent całości	66,400%	33,600%	
Chi-kwadrat (df=1)	3,44	p= ,0636	
V-kwadrat (df=1)	3,41	p= ,0646	
Chi-kwadrat skoryg. Yatesa	2,76	p= ,0965	
Fi-kwadrat	,02754		
dokł. p Fishera, jednostr.		p= ,0472	
dwustr.		p= ,0822	
Chi-kwadrat McNemary A/D	,72	p= ,3960	
Chi-kwadrat McNemary B/C	17,96	p= ,0000	

Uzyskaliśmy wyniki takie jak w poprzednim podrozdziale, zatem nie ma potrzeby ich tutaj omawiać. Warto pamiętać, że analizę taką możemy wykonać bez konieczności wprowadzania danych do arkusza.

11.5. Miary oceny ryzyka

W poprzednim podrozdziale omówiono test χ^2, jego zastosowanie i interpretację wyników. Warto ponownie zwrócić uwagę, że sama wartość χ^2 oraz odpowiadająca jej wartość p mówią jedynie, czy zależność między badanymi zmiennymi istnieje lub nie istnieje. Nie możemy jednak stwierdzić, jaka jest siła danego powiązania czy wielkości efektu, ponieważ są one zależne od liczebności próby (N). Z poprzedniego rozdziału wiemy, że kiedy rośnie liczebność badanej próby, rośnie także wartość χ^2 oraz maleje wartość p. Ale, zgodnie z intuicją, wielkość efektu nie powinna się zmieniać wraz ze zmianą liczebności próby. Sugeruje się zatem, aby wraz z wartościami dla testu χ^2 były raportowane miary oceny ryzyka. Jednymi z najczęściej stosowanych miar są ryzyko względne oraz iloraz szans. Obie te miary są wyliczane na podstawie danych umieszczonych w tabeli dwudzielczej.

11.5.1. Ryzyko względne

Przez ryzyko względne (ang. *relative risk* – *RR*) rozumiemy stosunek prawdopodobieństwa wystąpienia zdarzenia (punktu końcowego) w grupie narażonej do prawdopodobieństwa wystąpienia zdarzenia w grupie nienarażonej. Grupę nienarażoną często nazywamy grupą kontrolną.

Zebrane dane w bardzo łatwy sposób można umieścić w tabeli dwudzielczej (➡ **tabela 11.4**).

Narażenie	**Przypadki**	**Kontrole**	**Przypadki + kontrole**
Narażeni	a	b	a + b
Nienarażeni	c	d	c + d
Narażeni + nienarażeni	a + c	b + d	N = a + b + c + d

Tabela 11.4. Tabela dwudzielcza do oceny ryzyka względnego

Wartość ryzyka względnego wyliczamy ze wzoru:

$$RR = \frac{c(a+b)}{a(c+d)}$$

Jak interpretować współczynnik ryzyka względnego? Współczynnik $RR = 1$ określa sytuację, w której prawdopodobieństwa wystąpienia zdarzenia w grupach nienarażonej i narażonej na działanie badanego czynnika są podobne (zbliżone). Z kolei $RR > 1$ oznacza dodatnią zależność, czyli sytuację, w której prawdopodobieństwo wystąpienia zdarzenia jest większe w grupie narażonej na badany czynnik (np. wśród palaczy rak płuc występuje częściej niż wśród osób niepalących). Jeżeli $RR = 2$, oznacza to, że wystąpienie zdarzenia w grupie narażonej jest 2 razy bardziej prawdopodobne niż w grupie nienarażonej; analogicznie $RR = 3$ oznacza trzykrotnie większe prawdopodobieństwo, a $RR = 1{,}5$ oznacza prawdopodobieństwo o 50% większe. Sytuację odwrotną, czyli zależność ujemną, oznacza $RR < 1$. Przy takich wartościach wystąpienie zdarzenia w grupie narażonej jest mniejsze niż w grupie kontrolnej. Przy $RR = 0{,}5$ prawdopodobieństwo wystąpienia zdarzenia w grupie narażonej jest dwukrotnie mniejsze niż w grupie kontrolnej.

Wykonajmy obliczenia na następującym przykładzie: sprawdzono, jak często u osób zaszczepionych i niezaszczepionych przeciwko COVID-19 zaobserwowano objawy infekcji. Dane te ujęto w tabeli dwudzielczej (➥ **tabela 11.5**).

Lek	**COVID-19**	
	wystąpiło (1)	**nie wystąpiło (0)**
Brak szczepienia (0)	55 (a)	95 (b)
Szczepienie (1)	25 (c)	125 (d)

Tabela 11.5. Dane do wyliczenia ryzyka względnego

Wartość p dla testu χ^2 w naszym przykładzie wynosi 0,0001 i pozwala na odrzucenie H_0 oraz przyjęcie wniosku o istnieniu zależności między częstością zachorowań na COVID-19 a wykonaniem szczepienia ochronnego. Interesuje nas kształt tej zależności. Korzystając z wzoru, obliczamy wartość RR. Dla naszych danych wynosi ono 0,455. Możemy zatem powiedzieć, że u osób szczepionych ryzyko wystąpienia infekcji COVID-19 jest prawie o połowę mniejsze niż u tych, którzy nie poddali się szczepieniu.

11.5.2. Iloraz szans

Bardzo podobną miarą do ryzyka względnego jest wskaźnik ilorazu szans (ang. *odds ratio* – *OR*). Jednakże miary te nie są tożsame. Iloraz szans dwóch porównywanych grup definiuje się jako stosunek możliwości wystąpienia zdarzenia dla grupy kontrolnej (bez szczepienia) do możliwości wystąpienia zdarzenia w grupie narażonej (szczepionej).

Aby dobrze zrozumieć pojęcie ilorazu szans, należy najpierw zrozumieć znaczenie szansy. Szansą nazywamy stosunek liczby wystąpień danego zdarzenia do liczby braku wystąpień danego zdarzenia. Na przykład przebadano losowo 1000 kobiet w wieku rozrodczym pod względem zachorowalności na raka macicy. Odnotowano 56 przypadków zachorowania, a w 944 przypadkach (1000 – 56) nie stwierdzono choroby. Na podstawie tych danych określono szansę na wystąpienie raka macicy u kobiet w tej grupie wiekowej na 5,6%. Zatem szansa na zachorowanie na raka macicy w wieku rozrodczym wyniosła 2,5%.

W odróżnieniu od ryzyka względnego, gdzie dzielimy prawdopodobieństwa (proporcję liczby wystąpień zdarzenia do wszystkich obserwacji (wystąpienia + niewystąpienia), przy wyliczaniu ilorazu szans dzielimy między sobą szanse (proporcję liczby wystąpień zdarzenia do sytuacji, kiedy to zdarzenie nie występuje).

Wartość ilorazu szans możemy wyliczyć z wzoru:

$$OR = \frac{ad}{bc}$$

Dla naszego przykładu $OR = 1{,}211$. Zatem szansa wystąpienia infekcji COVID-19 u osób niezaszczepionych jest 1,21 razy większa niż wśród zaszczepionych. Wartości ilorazu szans interpretuje się następująco:

- $OR = 1$ – szansa taka sama w obu grupach;
- $OR > 1$ – szansa wystąpienia zdarzenia w grupie nienarażonej (nieszczepionej) jest większa niż w narażonej (szczepionej);
- $OR < 1$ – szansa wystąpienia zdarzenia w grupie nienarażonej (nieszczepionej) jest mniejsza niż w grupie narażonej (szczepionej).

Podobnie jak przy obliczaniu ryzyka względnego musimy pamiętać, że *OR* podlega zmienności losowej, dlatego w pierwszej kolejności należy zweryfikować istotność tej statystyki za pomocą testu χ^2.

Agnieszka Rusiecka,
Krzysztof Kujawa

12 Testy *t* i ich nieparametryczne odpowiedniki w medycynie

12.1. Testy *t* Studenta

Cel ćwiczenia. Porównanie średnich arytmetycznych obliczonych na podstawie danych z dwóch prób.

12.1.1. Rodzaje testów *t* Studenta

Porównanie populacji, określenie różnic między nimi oraz poznanie przyczyny ich powstawania to najczęstsze zadania, które stawia przed sobą badacz. To właśnie znajdowanie różnic między populacjami stanowi rzeczywistą wartość prowadzonej analizy statystycznej. Porównania tego najczęściej będziemy próbowali dokonać przez zestawienie wartości średnich arytmetycznych obliczonych na podstawie pomiarów z prób pochodzących z badanej populacji.

Testy *t* Studenta wykorzystywane są do porównania zmiennych ilościowych (np. stężenia hemoglobiny, masy ciała itp.) w dwóch grupach, czyli zmienna grupująca jest dychotomiczna (np. płeć: kobieta–mężczyzna; stan pacjenta: chory–zdrowy).

Istnieją jednak trzy rodzaje testów *t* Studenta, a dobór odpowiedniego uzależniony jest od tego, w jaki sposób zaprojektowano doświadczenie.

Testy *t* Studenta można podzielić na:
- testy *t* Studenta dla dwóch prób niezależnych,
- testy *t* Studenta dla dwóch prób zależnych,
- testy *t* Studenta dla jednej próby.

12.2. Próby niezależne i próby zależne

12.2.1. Próby niezależne

Jak sama nazwa testu wskazuje, próby muszą być wobec siebie niezależne, a więc wyniki pomiaru jednej nie są w żaden sposób powiązane z pomiarami drugiej. Celem testu *t* Studenta dla prób niezależnych jest porównanie ze sobą dwóch różnych grup obserwacji pod względem jednej cechy ilościowej. Wykorzystuje się go również do porównania próby eksperymentalnej (np. poddanej leczeniu określonym preparatem) z próbą kontrolną (grupa pacjentów zażywająca placebo).

Przypuśćmy, że chcemy ocenić różnicę stężenia hemoglobiny w grupie kobiet i mężczyzn. W ➥ **tabeli 12.1** podane są dane dla 12 kobiet i 10 mężczyzn. Obserwujemy więc dwie niezależne grupy pacjentów, utworzone ze względu na płeć.

Kobiety	12,5	12,3	10,1	12,2	13,5	10,2	10,8	11,9	10,1	9,8	10,2	10,8
Mężczyźni	10,4	11,8	12,3	13,6	14,5	13,2	12,6	11,9	12,5	12,2		

Tabela 12.1. Próby niezależne. Stężenie hemoglobiny zmierzone u 12 kobiet i 10 mężczyzn

12.2.2. Próby zależne

Próby zależne (inaczej powiązane) to zbiory pomiarów dotyczących tej samej grupy obiektów.

Poniższy przykład (➥ **tabela 12.2**) wskazuje dwie próby zależne dotyczące 12 pacjentów, którym zmierzono stężenie hemoglobiny dwukrotnie: przed leczeniem i po jego zastosowaniu.

ID pacjenta	Hemoglobina	
	przed leczeniem	po leczeniu
1	12,17	13,97
2	11,42	14,76
3	11,64	11,98
4	11,23	13,05
5	12,52	12,47
6	13,34	12,23
7	11,68	15,12
8	12,60	15,20
9	11,89	13,77
10	12,50	14,20
11	11,86	12,29
12	11,62	13,96

Tabela 12.2. Próby zależne. Stężenie hemoglobiny u pacjentów przed leczeniem i po jego podjęciu

12.3. Założenia dla testów *t* Studenta

Aby zastosować określony test statystyczny, w pierwszej kolejności należy sprawdzić, czy spełnione są jego założenia. Tylko w sytuacji, gdy test dostosowany jest do typu danych oraz ich rozkładu, będzie można w pełni zaufać jego wynikom. Złe dopasowanie do danych spowoduje nietrafne interpretacje i prowadzić może do błędnych wniosków.

Test *t* Studenta wykorzystujemy do porównania ze sobą dwóch populacji. Służy do znalezienia odpowiedzi na pytanie, czy średnie wartości interesującej nas zmiennej zależnej w dwóch populacjach różnią się od siebie w sposób istotny statystycznie, czy też mają one charakter przypadkowy.

12.3.1. Rozkład normalny zmiennej

Rozkład wartości zmiennej zależnej w obu grupach badanych (np. stężenie hemoglobiny w grupie kobiet i w grupie mężczyzn) musi mieć rozkład normalny lub zbliżony do normalnego. Normalność rozkładu możemy sprawdzić przy użyciu testu statystycznego, np. testu Shapiro–Wilka lub Kołmogorowa–Smirnowa (zostały one opisane w rozdziale 9. *Statystyka opisowa i sprawdzanie normalności rozkładu danych*). Warto jednak zaznaczyć, że moc tych testów (tj. zdolność do wykrywania różnic istotnych statystycznie, czyli w tym przypadku istotnych odchyleń od roz-

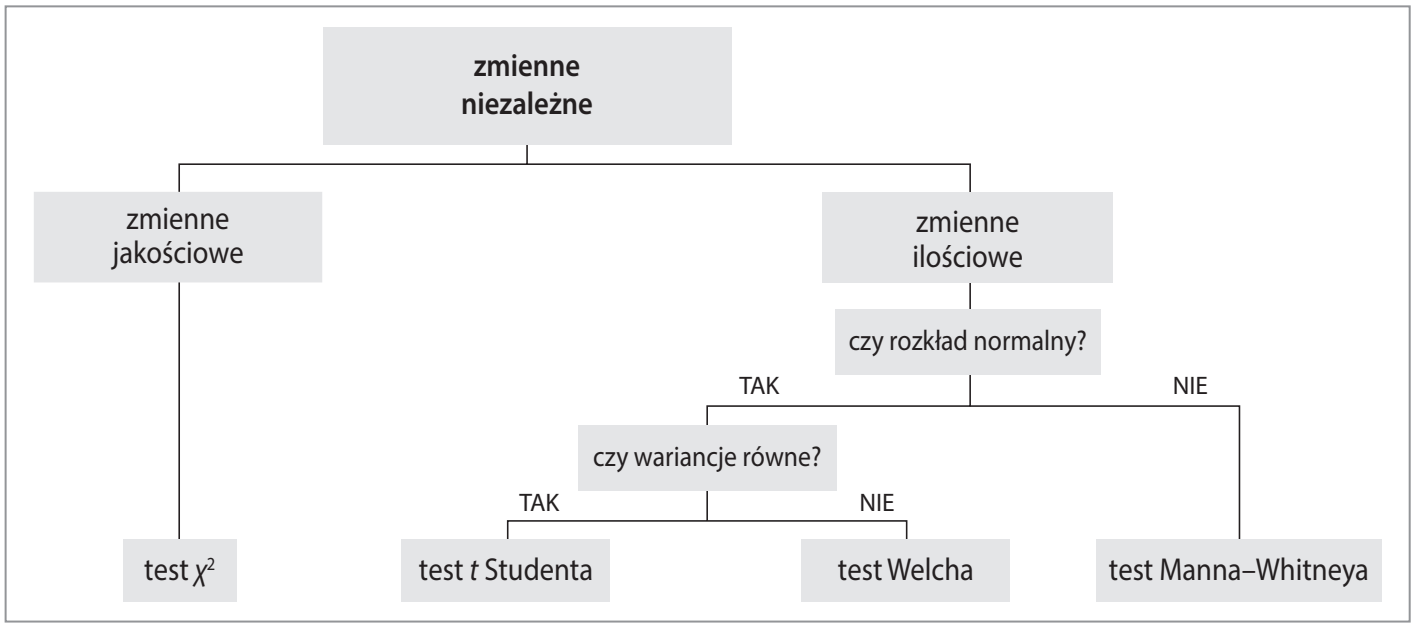
kładu normalnego) zależy od liczebności próby. Gdy wynosi ona kilka lub kilkanaście, moc nawet najsilniejszego z tych testów, czyli testu Shapiro–Wilka, wynosi kilka–kilkanaście procent i sprawdzanie normalności rozkładu jest nieefektywne. W takich przypadkach należy użyć testu nieparametrycznego.

12.3.2 Jednorodność wariancji

W przypadku stosowania testu *t* Studenta dla prób niezależnych wymagane jest, poza spełnieniem warunku o rozkładzie normalnym danych dla każdej z prób, również spełnienie warunku o jednorodności wariancji. Do sprawdzenia tego założenia służy test Levene'a lub Browna–Forsythe'a. Hipoteza zerowa, którą chcemy zweryfikować, zakłada jednorodność wariancji. Gdy to założenie nie jest spełnione, należy wykonać test *t* Studenta z poprawką Welcha, która uwzględnia niejednorodność wariancji (➥ **ryc. 12.1**).

12.4. Test *t* Studenta dla dwóch prób niezależnych

Wybór odpowiedniego testu statystycznego uzależniony jest od sposobu, w jaki zadamy pytanie badawcze, czy rozkład w obu grupach jest normalny oraz czy spełnione jest założenie o równości wariancji. Na poniższym schemacie blokowym przedstawiony jest algorytm doboru testu dla dwóch zmiennych niepowiązanych (➥ **ryc. 12.1**).



Ryc. 12.1. Schemat blokowy wyboru testu statystycznego dla grup niezależnych

Z powyższego schematu wynika, że test *t* Studenta należy stosować jedynie w sytuacji, kiedy dla obu porównywanych grup spełniony jest warunek o rozkładzie normalnym oraz równych wariancjach. W sytuacji gdy rozkładu normalnego danych nie ma (w jednej lub obu próbach), należy zastosować nieparametryczny test Manna–Whitneya oparty na rangowaniu wyników zmiennej zależnej.

Kiedy zaś warunek o rozkładzie normalnym jest spełniony, ale wariancje w próbkach nie są równe, do porównania grup należy użyć testu *t* Studenta z poprawką Welcha (zwanego w skrócie testem Welcha). Jest on uogólnieniem testu *t* Studenta na populacje o różnych wariancjach.

12.5. Próby niezależne – przykład obliczeń

12.5.1. Testy parametryczne

Wróćmy do naszego przykładu dla prób niezależnych (dla przypomnienia: próby niezależne reprezentowane przez dwie rozdzielne grupy pacjentów).

Chcemy ocenić różnicę stężenia hemoglobiny w grupie kobiet i mężczyzn. W ➥ **tabeli 12.3** podane są dane dla 12 kobiet i 10 mężczyzn. Obserwujemy dwie niezależne grupy pacjentów, gdzie zmienną grupującą jest PŁEĆ.

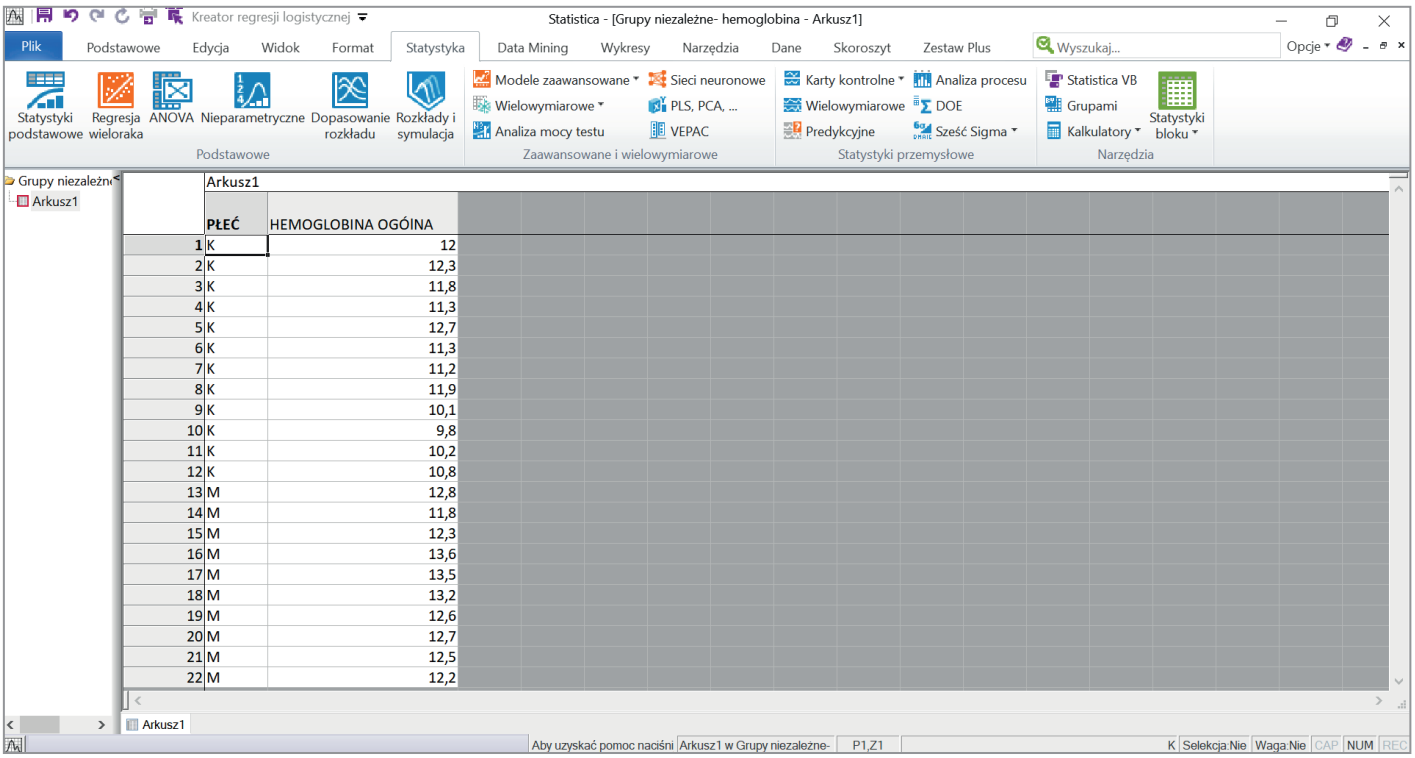
Kobiety	12,0	12,3	11,8	11,3	12,7	11,3	11,2	11,9	10,1	9,8	10,2	10,8
Mężczyźni	12,8	11,8	12,3	13,6	13,5	13,2	12,6	12,7	12,5	12,2		

Tabela 12.3. Próby niezależne reprezentowane przez 2 grupy pacjentów

Aby móc wykonać obliczenia w programie Statistica, należy wprowadzić w nim dane do arkusza.

Istnieje kilka możliwości przedstawienia danych w arkuszu, ale najczęściej wykorzystywany sposób polega na ich wpisywaniu przypadek po przypadku. W takiej sytuacji każdy wiersz przedstawia informacje o jednym pacjencie, a każda kolumna reprezentuje jedną badaną zmienną (np. płeć pacjenta, wiek, stężenie hemoglobiny itp.). To sposób najbardziej intuicyjny. Umożliwia łatwe i przejrzyste przedstawienie wszystkich zebranych o pacjencie danych. Należy jednak wyróżnić tzw. zmienną grupującą, która umożliwia zaliczenie pacjenta do jednej z porównywanych grup. W naszej sytuacji będzie nią zmienna PŁEĆ. Będzie dzięki temu można podzielić pacjentów na kobiety (zmienna PŁEĆ = K) oraz mężczyzn (zmienna PŁEĆ = M). Rezultat został przedstawiony na ➥ **ryc. 12.2**.

Ryc. 12.2. Dane początkowe do obliczeń statystycznych dla grup niezależnych

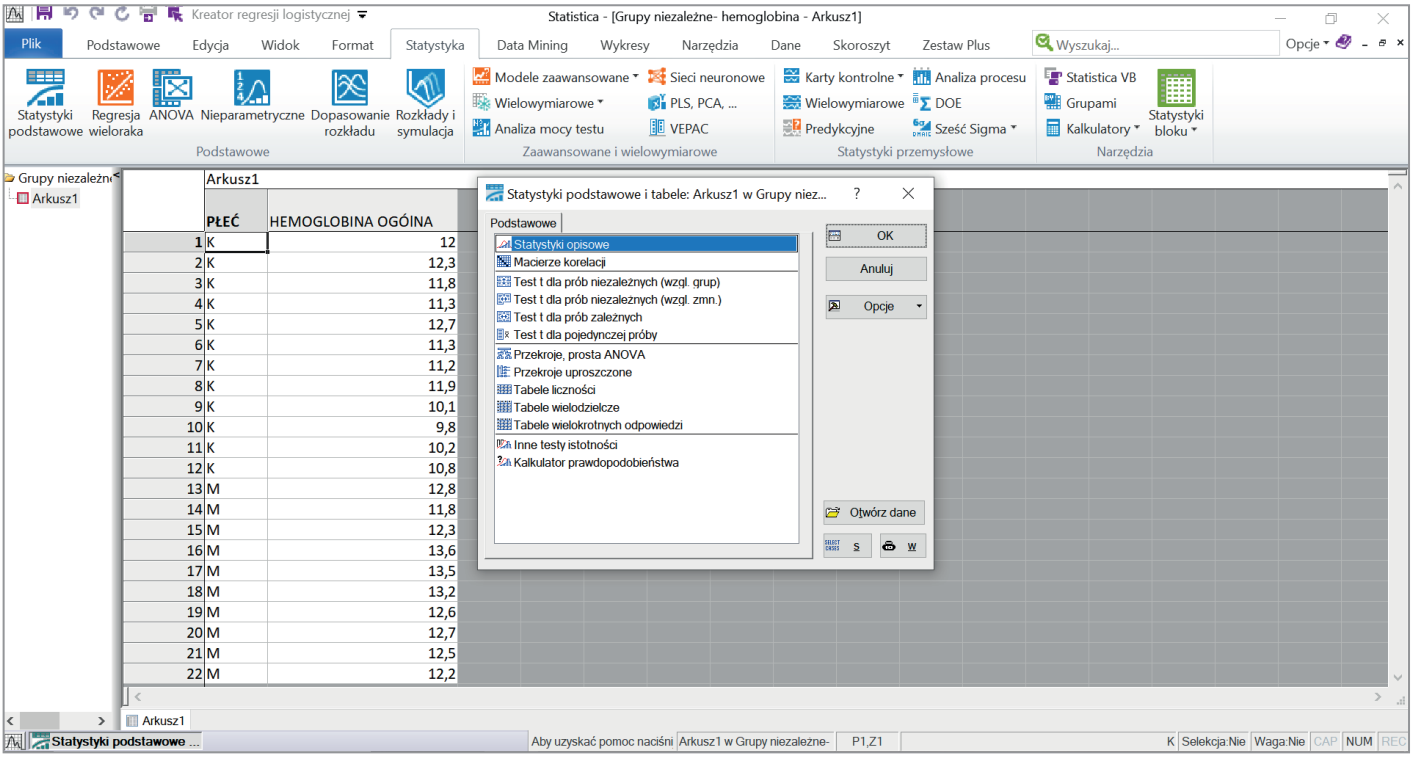


Druga opcja, tj. test *t* dla prób niezależnych (wzgl. zmn.), również umożliwia obliczenie statystyki testowej *t* Studenta, ale w tym przypadku dane wejściowe do obliczeń powinny być przedstawione w 2 kolumnach (w pierwszej znajdowałyby się stężenia hemoglobiny dla kobiet, a w drugiej dla mężczyzn). Nie jest to jednak wygodny sposób, gdy arkusz zawiera więcej informacji o pacjencie niż tylko np. stężenie hemoglobiny.

Ze względu na to, że nasze dane są wprowadzane pacjent po pacjencie ze zmienną grupującą „PŁEĆ", z dostępnej listy wybieramy Test t dla prób niezależnych (wzgl. grup).

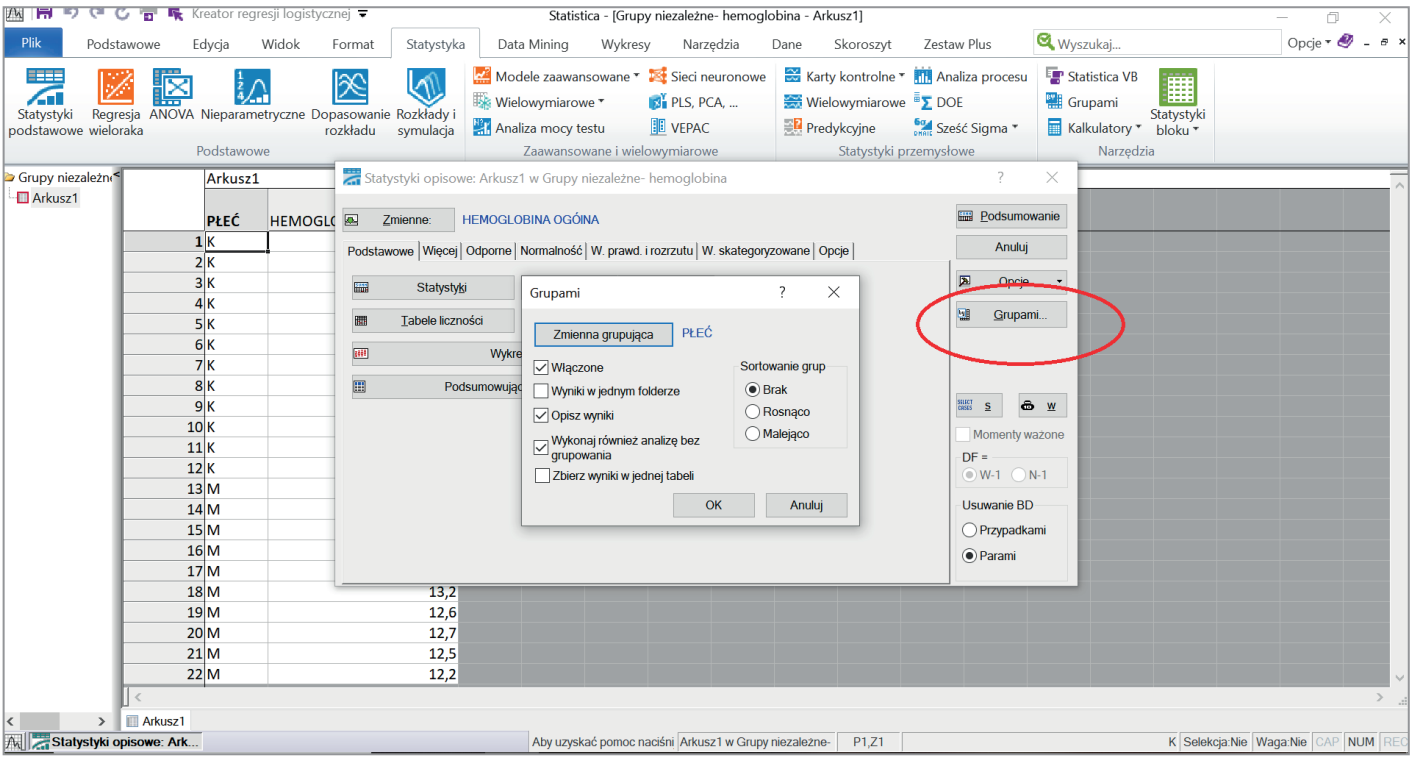
Analizę danych warto zacząć od obliczenia statystyk opisowych. W tym celu przechodzimy do karty Statystyka i wybieramy Statystyki podstawowe, a następnie Statystyki opisowe (➥ **ryc. 12.3**).

Ryc. 12.3. Wybór metody statystycznej



Dokonujemy wyboru zmiennych. W tym celu klikamy przycisk Zmienne, gdzie wybieramy zmienną zależną („HEMOGLOBINA OGÓLNA"), a następnie przycisk Grupami znajdujący się po prawej stronie okna, gdzie wybieramy zmienną grupującą („PŁEĆ"). Wybór zatwierdzamy przyciskiem OK (➥ **ryc. 12.4**).

Ryc. 12.4. Statystyki opisowe



Jak opisano w rozdziale 9, ta część analizy umożliwia obliczenie statystyk podstawowych, ale również sprawdzenie normalności rozkładów obu zmiennych.

W wyniku analizy otrzymujemy podsumowanie – osobno dla obu płci (➥ **ryc. 12.5**, ➥ **ryc. 12.6**).

Ryc. 12.5. Statystyki opisowe dla grupy kobiet

	PŁEĆ=K Statystyki opisowe (Arkusz in Grupy niezależne- hemoglobina)				
Zmienna	**Nważnych**	Średnia	Minimum	Maksimum	Odch.std
HEMOGLOBINA OGÓINA	12	11,28333	9,800000	12,70000	0,915357

Na podstawie wykonanych obliczeń widzimy, że średnie stężenie hemoglobiny (średnia ± odchylenie standardowe) w grupie kobiet wynosi 11,3 ± 0,92, a w grupie mężczyzn 12,7 ± 0,58. Interesuje nas, czy różnica w poziomie hemoglobiny u kobiet i mężczyzn jest istotna statystycznie, tzn. czy obserwowana różnica wynika z przypadku, czy rzeczywiście stężenie hemoglobiny w populacji kobiet jest niższe niż obserwowane u mężczyzn.

Ryc. 12.6. Statystyki opisowe dla grupy mężczyzn

	PŁEĆ=M Statystyki opisowe (Arkusz in Grupy niezależne- hemoglobina)				
Zmienna	**Nważnych**	Średnia	Minimum	Maksimum	Odch.std
HEMOGLOBINA OGÓINA	10	12,72000	11,80000	13,60000	0,575036

Innymi słowy, jeśli założymy, że stężenie hemoglobiny u kobiet i mężczyzn jest takie samo, to jakie jest prawdopodobieństwo, że różnica w jego wartościach w dwóch losowo wybranych próbach kobiet i mężczyzn jest tak duża lub większa niż w tym konkretnym badaniu, które już przeprowadziliśmy.

Odpowiedź na nasze pytanie da nam odpowiednio dobrany test statystyczny. Test *t* Studenta dla prób niezależnych służy weryfikacji hipotezy **o równości** dwóch średnich w dwóch populacjach – między dwiema teoretycznymi średnimi.

Teoretyczne średnie można symbolicznie zapisać w następujący sposób:

- μ_1 to średnia cechy w jednej populacji (średnie stężenie hemoglobiny dla populacji kobiet),
- μ_2 to średnia cechy w drugiej populacji (średnie stężenie hemoglobiny dla populacji mężczyzn).

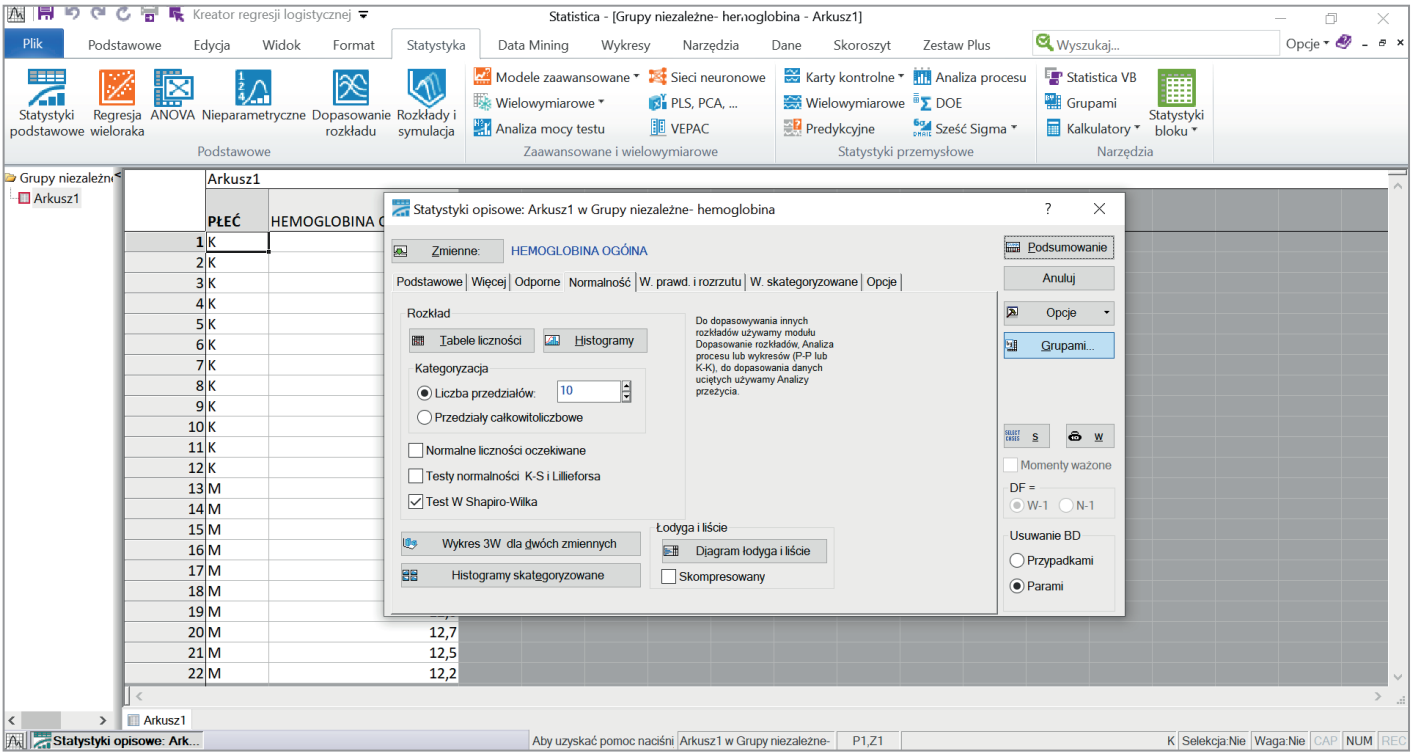
W ten sposób H_0 w teście *t* Studenta przedstawia się następującym wzorem:

$$H_0 : \mu_1 = \mu_2$$

Zanim jednak przejdziemy do testu *t* Studenta, sprawdzimy założenia dla niego. Tylko w takiej sytuacji będziemy mogli mieć pewność, że jest poprawnie dobrany do analizy danych, którymi dysponujemy, a wyniki tych analiz są wiarygodne.

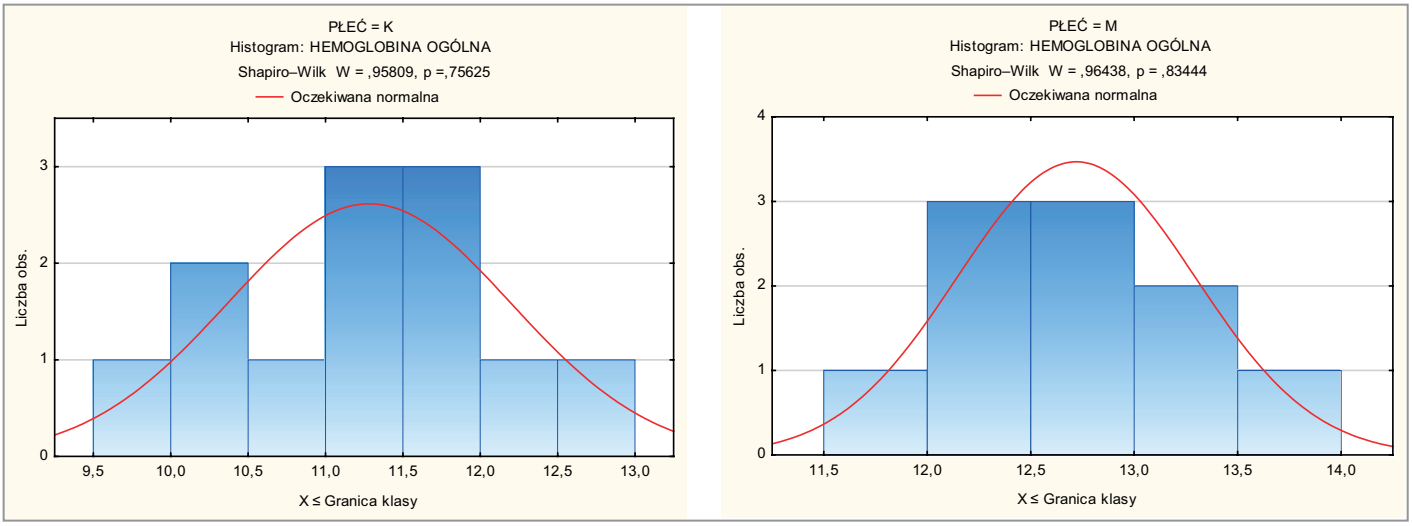
Kolejnym krokiem jest sprawdzenie normalności rozkładu danych w obu grupach. Jak opisano w poprzednich rozdziałach, normalność rozkładu można sprawdzić w części Statystyki opisowe, w zakładce Normalność (➥ **ryc. 12.7**). W tym przypadku do sprawdzenia rozkładu zostanie użyty test Shapiro–Wilka. A następnie tworzymy histogramy.

Ryc. 12.7. Sprawdzenie normalności rozkładów dla hemoglobiny w grupach



Ryc. 12.8. Histogramy dla kobiet i mężczyzn z wynikami testu normalności

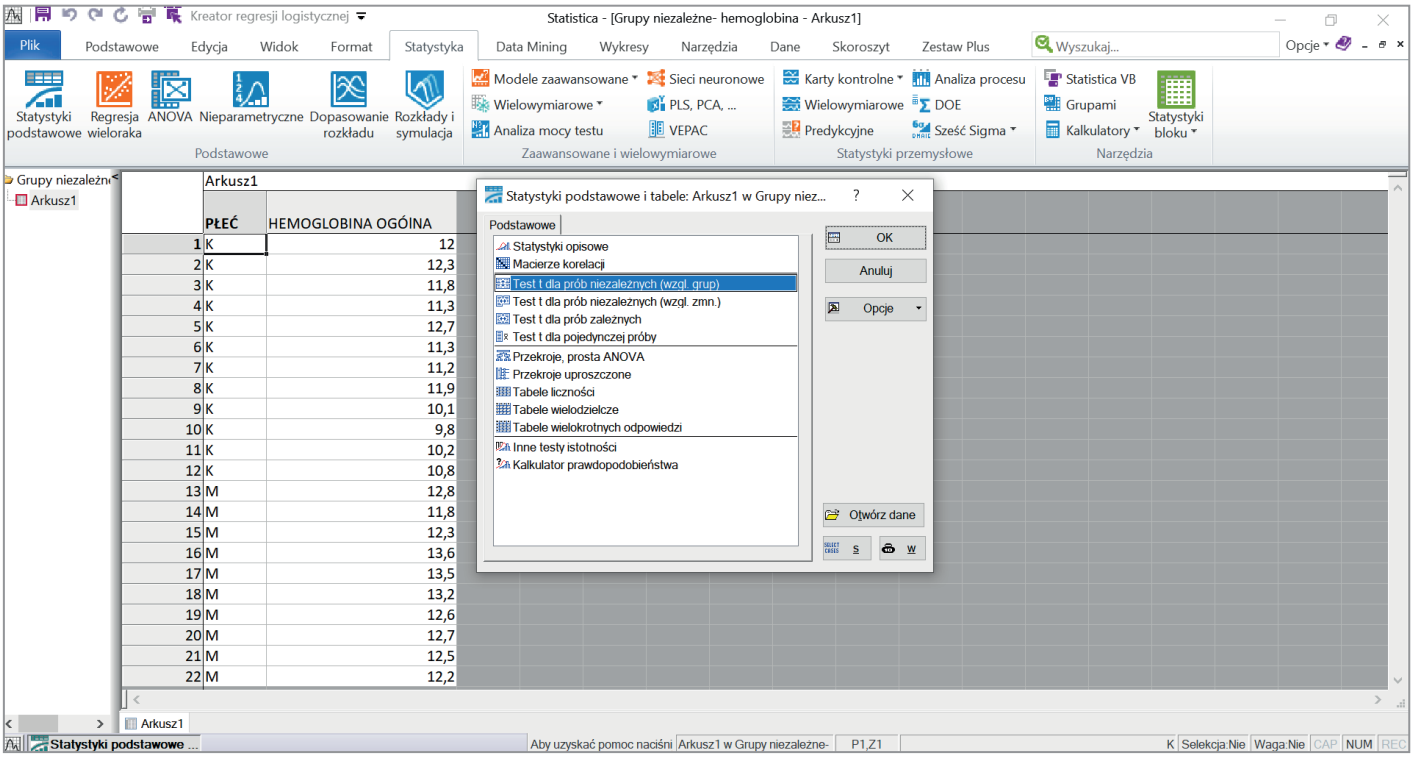
W rezultacie otrzymujemy histogramy dla kobiet i mężczyzn. Wartość testu Shapiro–Wilka można odczytać z wykresów (➥ **ryc. 12.8**).



Wartości p w obu grupach dla testu Shapiro–Wilka są większe od 0,05, a więc nie mamy podstaw do odrzucenia H_0 o normalności rozkładu. Oznacza to, że pierwsze założenie dla testu t Studenta, o normalności rozkładów w porównywanych grupach, jest spełnione. Należy jednak pamiętać o małej mocy tego testu (choć najsilniejszego z testów normalności rozkładów) w przypadku tak małych prób. To oznacza konieczność ostrożnego przyjmowania wyników analizy zgodności rozkładu zmiennej z rozkładem normalnym.

Ryc. 12.9. Opcje wyboru obliczeń statystycznych dostępnych z polecenia Statystyki podstawowe karty Statystyka

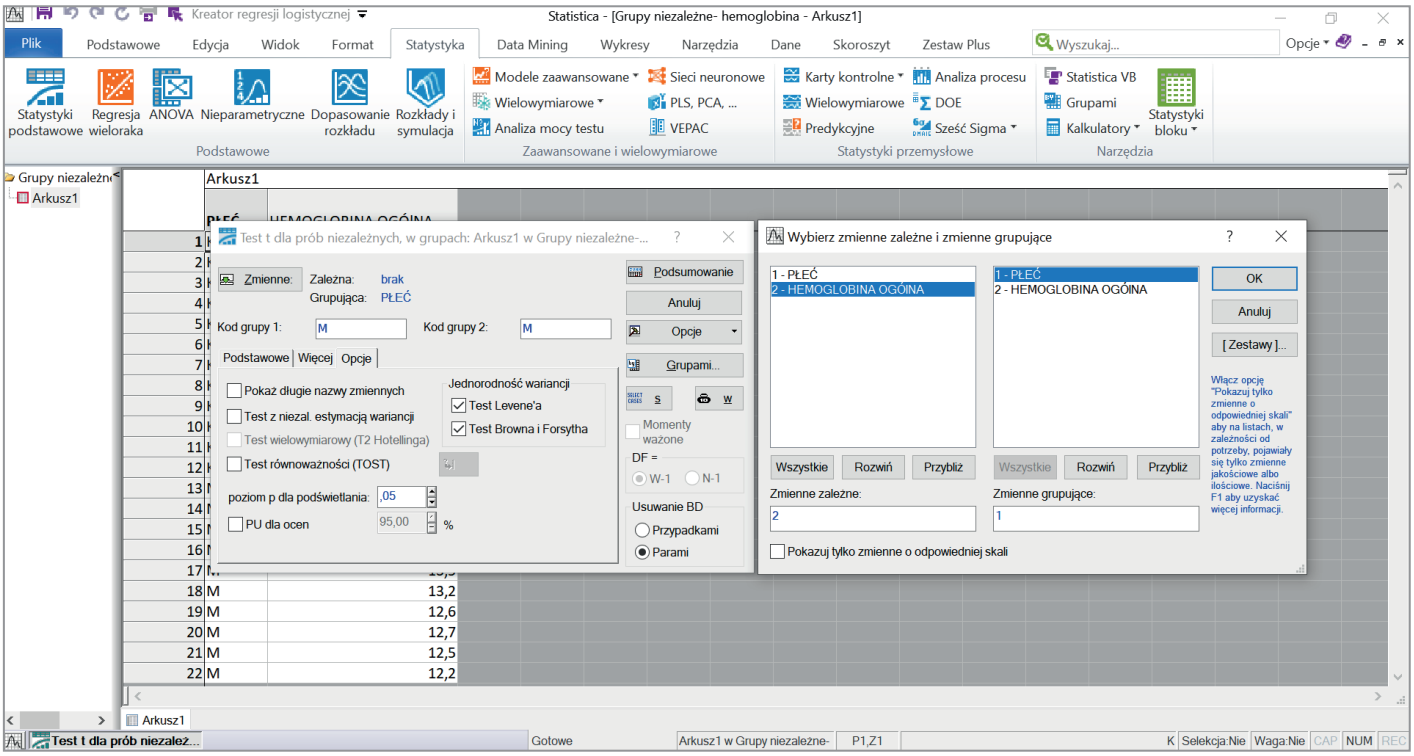
Sprawdzenie wariancji wykonamy w tym samym module, który służy do wykonania testu *t* Studenta. W tym celu przechodzimy do karty Statystyka, z której wybieramy Statystyki podstawowe. Z dostępnej listy wybieramy Test t dla prób niezależnych (wzgl. grup) i klikamy OK (➥ **ryc. 12.9**).



Przechodzimy do okna wyboru zmiennych oraz zakładki Podstawowe z możliwością obliczenia statystyki *t* Studenta i utworzenia wykresów. Dodatkowo w zakładce Więcej również istnieje możliwość skonstruowania histogramu i sprawdzenia normalności rozkładu danych, podobnie jak w sekcji Statystyki opisowe. Zakładka Opcje posiada funkcję sprawdzenia równości wariancji przy użyciu testu Levene'a oraz testu Browna–Forsythe'a.

Ryc. 12.10. Test *t* Studenta dla prób niezależnych. Wybór zmiennych i sprawdzenie jednorodności wariancji

W pierwszej kolejności wybieramy Zmienne (zmienna zależna to „HEMOGLOBINA OGÓLNA", zmienna grupująca to „PŁEĆ") oraz zaznaczamy pola wyboru testu Levene'a oraz Browna–Forsythe'a (➥ **ryc. 12.10**).



Po zaznaczeniu wskazanych pól klikamy przycisk Podsumowanie. W wyniku analizy zostanie utworzony arkusz wynikowy. Do analizy jednorodności wariancji potrzebujemy jego końcowego fragmentu (➥ **ryc. 12.11**).

Ryc. 12.11. Wyniki dla testów Levene'a oraz Browna i Forsythe'a

Zmienna	Testy t; Grupująca: PŁEĆ (Arkusz1 w Grupy niezależne- hemoglobina) Grupa 1: K Grupa 2 M							
	iloraz F Wariancje	p Wariancje	Levene'a F(1,df)	df Levene'a	p Levene'a	Brn-Fors F(1,df)	df Brn-Fors	p Brn-Fors
HEMOGLOBINA OGÓINA	2,533908	0,173445	2,063959	20	0,166278	2,014224	20	0,171229

Wartości *p* dla obu testów są wyższe niż 0,05, a więc nie ma podstaw do odrzucenia H_0 o równości wariancji. Oznacza to, że założenie o jednorodności wariancji jest spełnione. Jednak i tym razem należy zachować ostrożność w interpretacji wyników ze względu na małą liczebność próby, gdyż moc testów statystycznych uzależniona jest m.in. od wielkości próby.

Zgodnie ze schematem z ➥ ryc. 12.1 po sprawdzeniu i potwierdzeniu założeń o normalności rozkładu danych i jednorodności wariancji możemy przystąpić do porównania stężenia hemoglobiny u kobiet i mężczyzn przy użyciu testu *t* Studenta.

12.5.1.1. Wyniki porównania przy użyciu testów *t* Studenta

Wyniki porównania stężenia hemoglobiny w grupie kobiet i mężczyzn przy użyciu testów *t* Studenta znajdziemy w początkowej części arkusza wygenerowanego do sprawdzenia jednorodności wariancji (➥ **ryc. 12.12**).

Ryc. 12.12. Wyniki testów *t* Studenta

Zmienna	Testy t; Grupująca: PŁEĆ (Arkusz1 w Grupy niezależne- hemoglobina) Grupa 1: K Grupa 2 M								
	Średnia K	Średnia M	t	df	p	Nważnych K	Nważnych M	Odch.std K	Odch.std M
HEMOGLOBINA OGÓINA	11,28333	12,72000	-4,29735	20	0,000351	12	10	0,915357	0,575036

W arkuszu tym wyświetlone są kolejno:

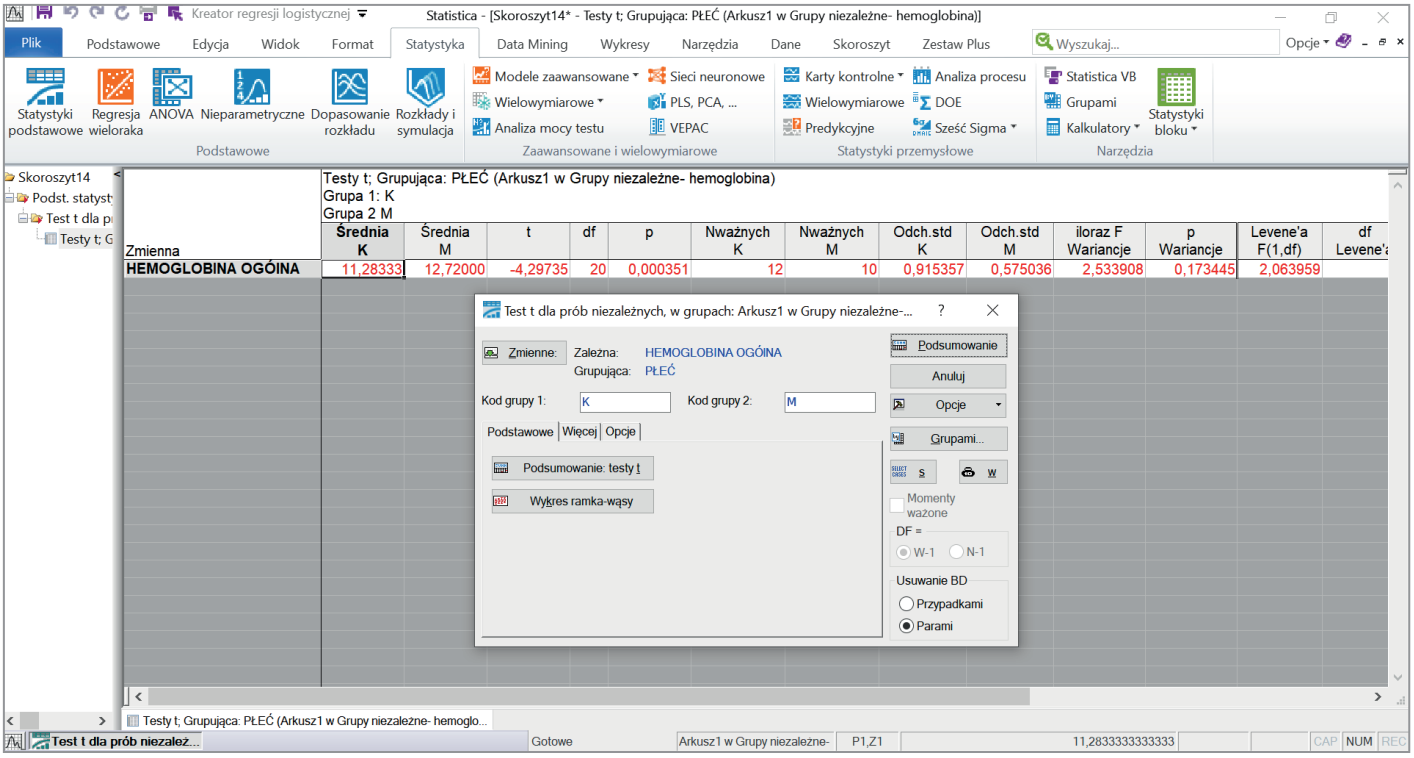
- średnie stężenie hemoglobiny dla kobiet,
- średnie stężenie hemoglobiny dla mężczyzn,
- wartość testu *t*,
- liczbę stopni swobody dla testu *t*,
- poziom prawdopodobieństwa *p* (ang. *p*-value),
- liczebność grupy kobiet,
- liczebność grupy mężczyzn,
- odchylenie standardowe dla grupy kobiet,
- odchylenie standardowe dla grupy mężczyzn,

Najistotniejsza jest dla nas wartość *p*. Wynosi ona 0,000351, a więc jest mniejsza od wartości $\alpha = 0{,}05$. Przy $p < \alpha$ decydujemy się na odrzucenie H_0 o równości średnich i przyjęcie H_1. Oznacza to, że średnie stężenie hemoglobiny w populacji kobiet jest istotnie niższe niż w populacji mężczyzn. Wyniki istotne statystycznie są w programie wyświetlone na czerwono.

12.5.1.2. Wykres ramka-wąsy

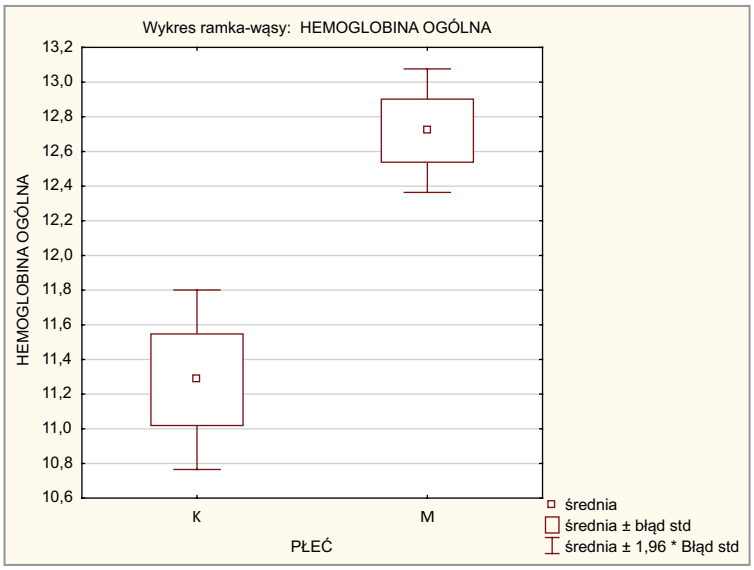
Aby wykonać dla danych wykres ramka-wąsy umożliwiający porównanie średnich stężeń hemoglobiny u kobiet i mężczyzn, przechodzimy w ramach tej samej analizy do zakładki Podstawowe, a następnie klikamy przycisk Wykres ramka-wąsy (➥ **ryc. 12.13**).

Ryc. 12.13. Wykonanie wykresu ze statystykami opisowymi dla porównywanych grup



Otrzymujemy wykres ramka-wąsy przedstawiający i porównujący wybrane statystyki opisowe dla poziomów hemoglobiny w grupie kobiet i grupie mężczyzn. Potwierdza on wyższe średnie stężenie hemoglobiny w populacji mężczyzn. Statistica daje możliwość tworzenia różnych wykresów ramka-wąsy (tj. z różnie definiowanymi ramkami i wąsami), ale zaleca się wybranie takiego wariantu, w którym wąsy reprezentują przedział ufności (obliczany wyrażeniem 1,96 × błąd standardowy). Umożliwia on łatwą ocenę, czy różnice są istotne statystycznie. Tak jest w przypadku, gdy przedziały ufności nie zachodzą na siebie, a im dalej sa one położone od siebie (wg osi X), tym większa jest istotność statystyczna danej różnicy między średnimi (➥ **ryc. 12.14**).

Ryc. 12.14. Wykres ramka–wąsy przedstawiający stężenie hemoglobiny u kobiet i mężczyzn

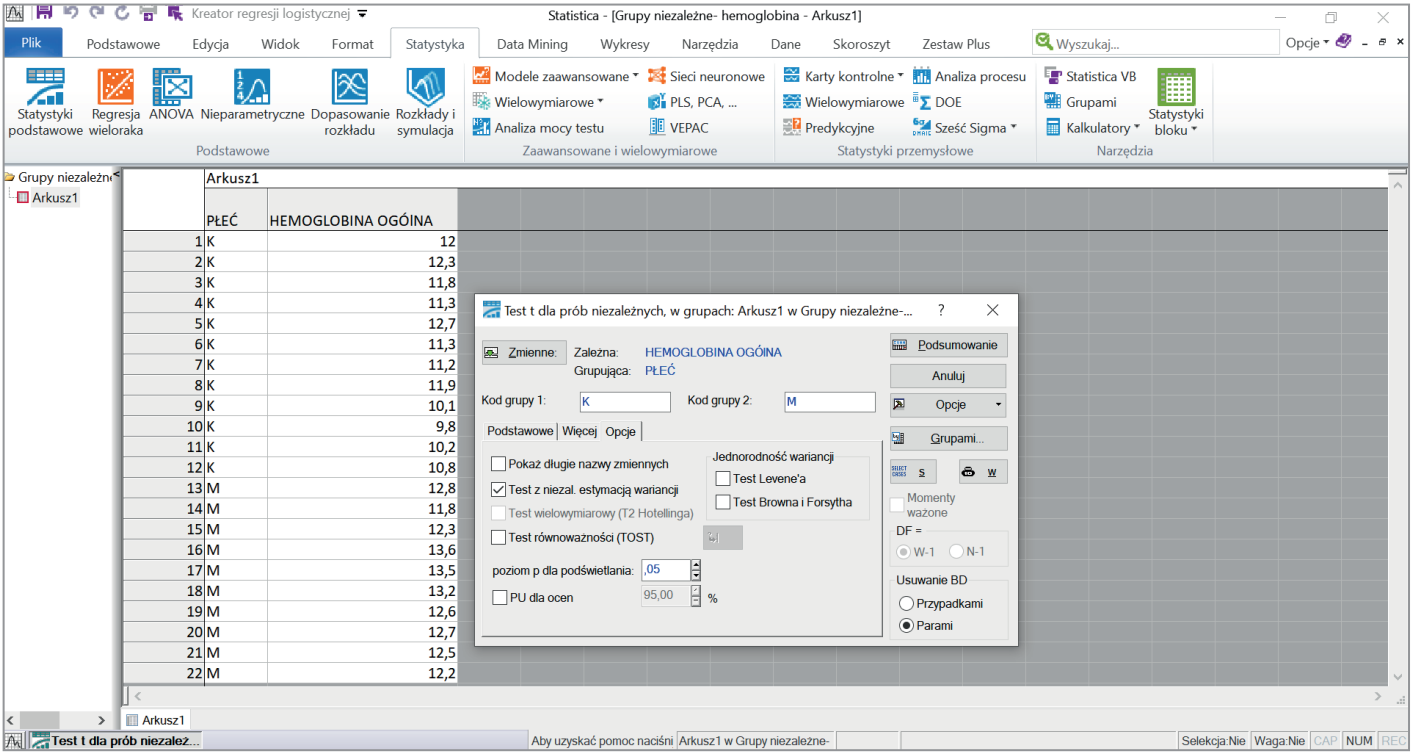


12.5.1.3. Test Welcha dla niejednorodnych wariancji

Nierzadko dochodzi do sytuacji, że w trakcie sprawdzania założeń dla testów *t* orientujemy się, iż pomimo normalnego rozkładu danych wariancje w porównywanych próbach nie są równe, a hipotezę zerową dla testów Levene'a oraz Browna–Forsythe'a o jednorodności wariancji należy odrzucić ($p < 0{,}05$). Oznacza to, że nie wszystkie założenia dla testu *t* Studenta są spełnione i nie należy go stosować w przypadku naszego układu danych. Jak wtedy postępować? Zgodnie ze schematem z ➥ ryc. 12.1 do porównania średnich stężeń hemoglobiny w grupie kobiet i mężczyzn należałoby zastosować test Welcha.

W tym celu wybieramy Testy t dla prób niezależnych, w grupach, dodajemy zmienne, tak jak opisano wcześniej, a następnie w oknie Opcje zaznaczamy pole wyboru Test z niezależną estymacją wariancji (➥ **ryc. 12.15**).

Ryc. 12.15. Pole wyboru testu Welcha



Klikamy przycisk Podsumowanie. W wyniku tego działania otrzymujemy arkusz wynikowy z dodatkowymi polami, gdzie pole „t oddz. est. war." przedstawia wartość testu Welcha, a pole „p dwustron." odpowiadającą mu wartość *p*, na podstawie której należałoby podjąć decyzję dotyczącą równości średnich poziomów hemoglobiny w populacjach kobiet i mężczyzn. W naszej sytuacji wartość *p* jest niższa niż poziom $\alpha = 0{,}05$, a więc odrzucamy H_0 i przyjmujemy H_1, co prowadzi do wniosku, że stężenie hemoglobiny istotnie różni się w populacji kobiet i mężczyzn (➥ **ryc. 12.16**).

Ryc. 12.16. Test Welcha, wyniki analizy

Testy t; Grupująca: PŁEC (Arkusz1 w Grupy niezależne- hemoglobina)
Grupa 1: K
Grupa 2 M

Zmienna	Średnia K	Średnia M	t	df	p	t oddz. est.war.	df	p dwustron	Nważnych K	Nważnych M	Odch.std K	Odch.std M	iloraz F Wariancje	p Wariancje
HEMOGLOBINA OGÓINA	11,28333	12,72000	-4,29735	20	0,000351	-4,47888	18,74692	0,000265	12	10	0,915357	0,575036	2,533908	0,173445

12.5.2. Testy nieparametryczne

12.5.2.1. Test Manna–Whitneya dla rozkładów innych niż normalne

Jeśli założenie o normalności rozkładu danych nie jest spełnione (lub nie ma możliwości jego rzetelnego sprawdzenia – w przypadku małych prób), do porównania dwóch prób wykorzystujemy testy *U* Manna–Whitneya, Walda–Wolfowitza lub Kołmogorowa–Smirnowa, należące do grupy testów nieparametrycznych. Najczęściej używany jest ten pierwszy, który stanowi najmocniejszą alternatywę dla testu *t* Studenta. Musimy jednak pamiętać, że o ile test *t* Studenta porównuje ze sobą średnie arytmetyczne, o tyle test *U* Manna–Whitneya sprawdza, czy dane z dwóch populacji różnią się co do wartości (wielkości). Czyli H_0 dla tego testu weryfikuje H_0, że dwie losowo wybrane próby pochodzą z dwóch populacji o takich samych danych. Test ten posługuje się rangami. Aby nadać odpowiednie rangi wynikom (w programach do analizy statystycznej ten krok wykonywany jest automatycznie), należy je:

- uporządkować,
- przyporządkować liczby naturalne każdej wartości od najmniejszej do największej (lub odwrotnie – od największej do najmniejszej),
- jeśli jakaś wartość powtarza się, wyliczyć średnią arytmetyczną z rang.

12.5.2.2. Przykład obliczeń

Chcemy porównać stężenie glukozy w dwóch grupach pacjentów, przyjmujących dwa różne leki: A i B. Obie grupy są 17-osobowe, w pierwszej grupie pacjenci przyjmują lek A, w drugiej – lek B.

Dane znajdują się w ➥ **tabeli 12.4**.

Lek A	84	89	110	93	91	117	99	93	89	96	109	104	112	87	89	89	129
Lek B	109	114	135	118	116	142	124	118	114	121	134	129	137	112	114	114	154

Tabela 12.4. Stężenie glukozy pacjentów z grup A i B

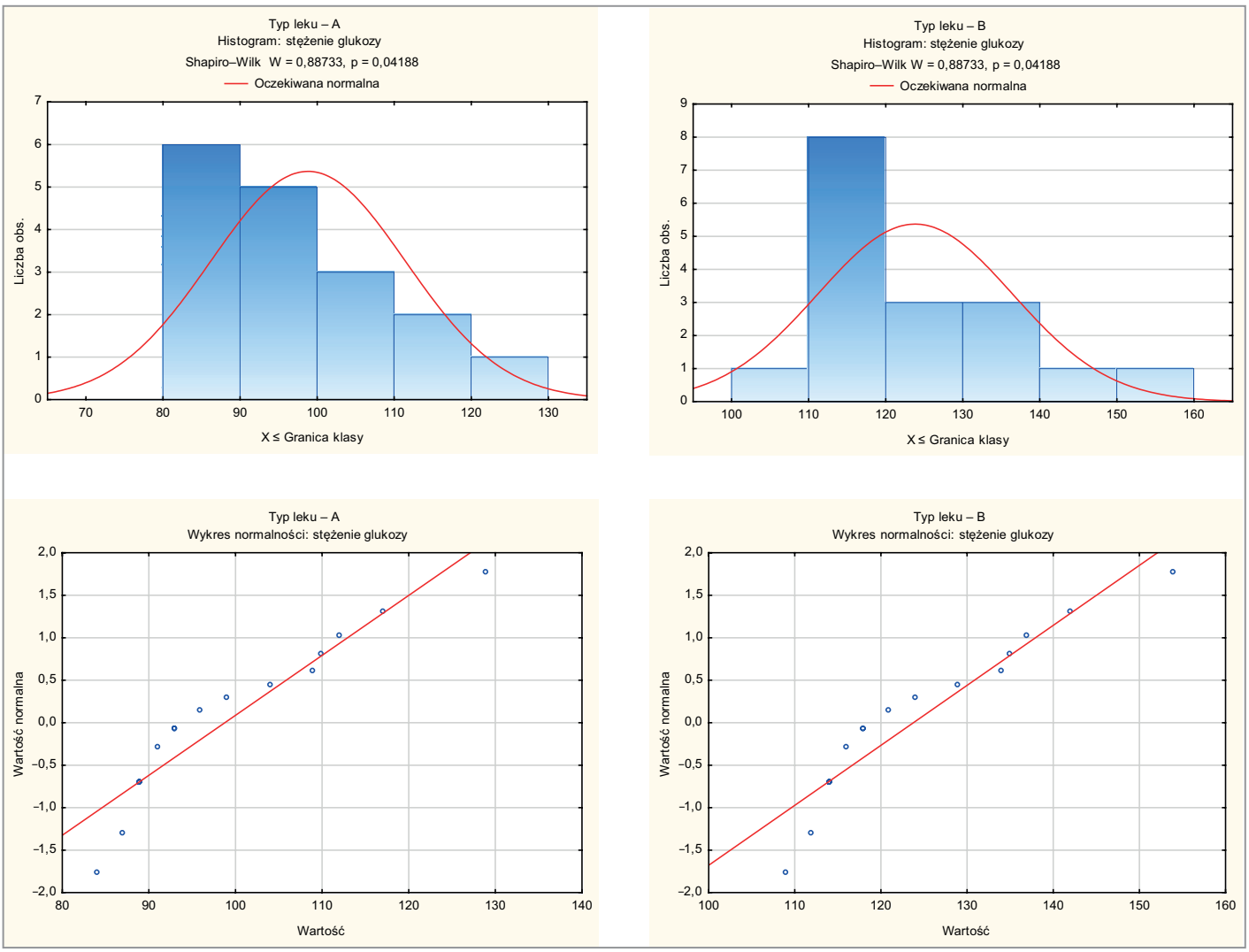
Dane wprowadzamy do arkusza grupami, pacjent po pacjencie. Stężenie glukozy u wszystkich pacjentów będzie zawierała pierwsza kolumna, natomiast w drugiej zostaną umieszczone informacje o typie przyjmowanego leku (zmienna grupująca), co pozwoli na zaliczenie pacjenta do jednej z porównywanych grup.

Ponieważ mamy do czynienia z dwiema grupami niezależnymi, do wyboru odpowiedniego testu statystycznego możemy użyć schematu z ➥ ryc. 12.1. Zaczynamy więc od sprawdzenia rozkładu badanych zmiennych. W tym celu należy wykonać

test Shapiro–Wilka (polecenie Statystyki podstawowe, następnie Statystyki opisowe, zakładka Normalność).

W wyniku tej analizy otrzymujemy 2 histogramy z wartościami *p* dla testu Shapiro–Wilka: osobno dla grupy pacjentów przyjmujących lek A i dla przyjmujących lek B. W zakładce Wykres prawd. I rozrzutu można również wykonać wykresy normalności (➡ **ryc. 12.17**).

Ryc. 12.17. Rozkład stężenia glukozy w grupie pacjentów przyjmujących lek A i lek B

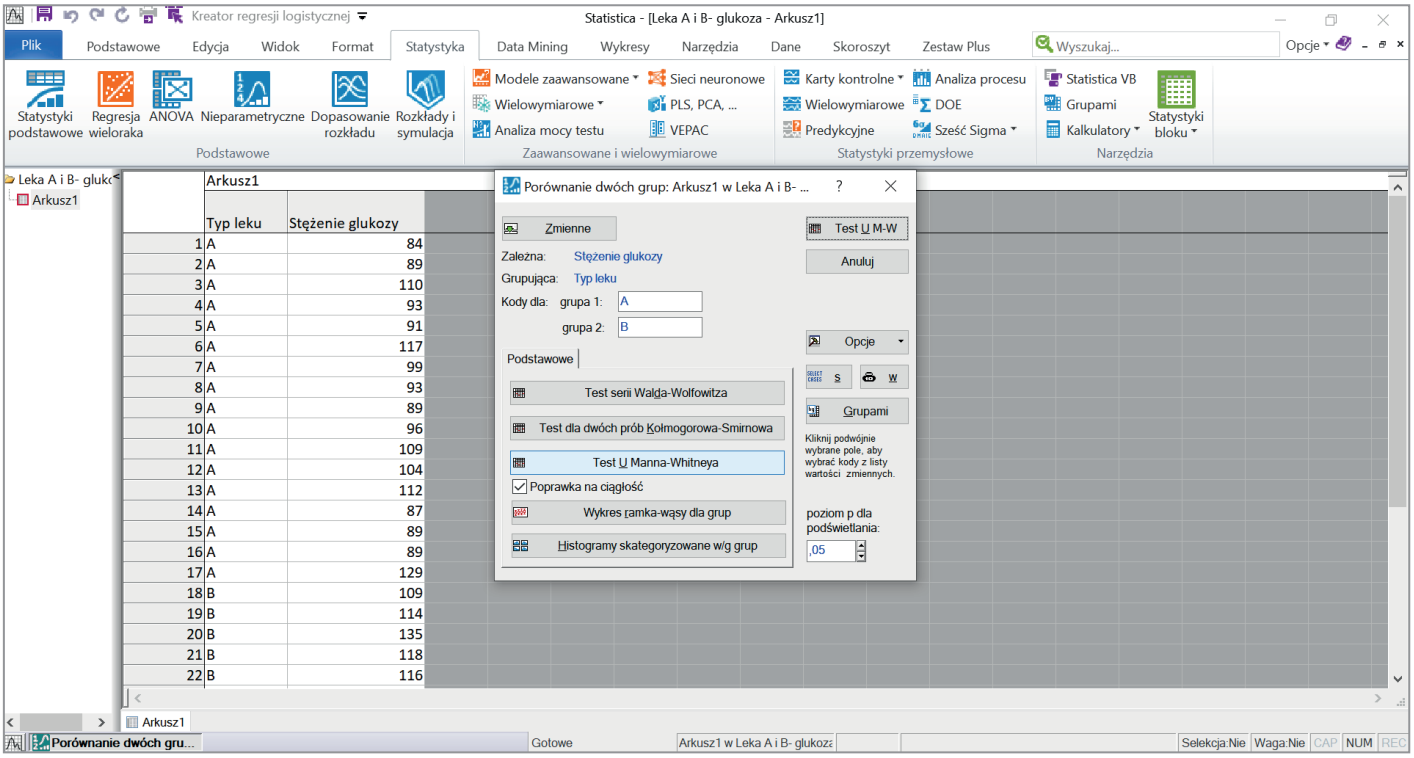


W obu przypadkach wartości *p* są mniejsze niż 0,05, co pozwala nam na odrzucenie H_0 o normalności rozkładów zarówno w grupie A, jak i B. Widzimy zatem, że do porównania nie możemy użyć testu *t* Studenta dla grup niezależnych, ponieważ nie jest spełnione założenie o normalności rozkładu. Nasze wyniki nie byłyby w tej

sytuacji miarodajne. Zachodzi zatem konieczność użycia nieparametrycznego testu, który jest odpowiednikiem testu *t* Studenta. Taki jest test Manna–Whitneya.

Interesujący nas test dla dwóch prób niezależnych wybierzemy poleceniem Nieparametryczne. Z listy dostępnych testów wybieramy Porównanie dwóch prób niezależnych (grup), co przedstawiono na ➥ **ryc. 12.18**.

Ryc. 12.18. Wybór testu *U* Manna–Whitneya



Dokonujemy wyboru zmiennych, gdzie zmienną zależną jest „Stężenie glukozy", a zmienną grupującą – „Typ leku", a następnie klikamy w przycisk Test *U* Manna–Whitneya. Otrzymujemy arkusz przedstawiony na ➥ **ryc. 12.19**.

Ryc. 12.19. Wyniki testu *U* Manna–Whitneya

Test U Manna-Whitneya (z poprawką na ciągłość) (Arkusz1 w Leka A i B- glukoza)
Względem zmiennej: Typ leku
Zaznaczone wyniki są istotne z p <,05000

Zmienna	Sum.rang A	Sum.rang B	U	Z	p	Z popraw.	p	N ważn. A	N ważn. B	2*1str. dokł. p
Stężenie glukozy	174,5000	420,5000	21,50000	-4,21933	0,000025	-4,22741	0,000024	17	17	0,000003

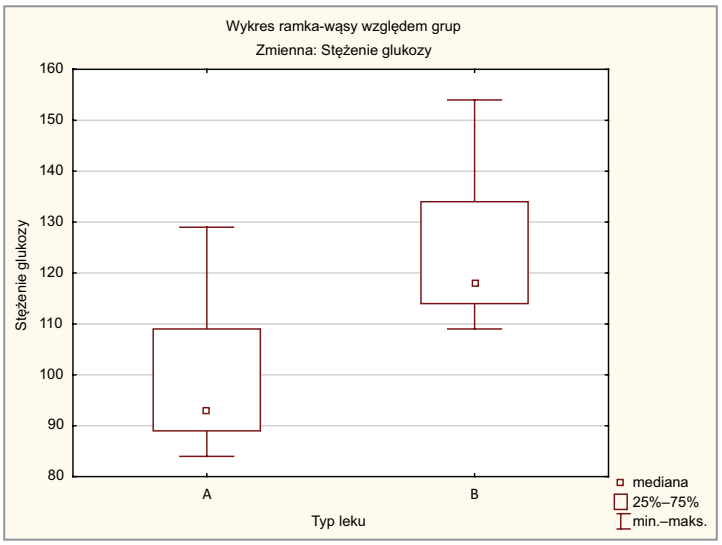
Przeanalizujmy kolejne wartości od lewej:

- suma rang w grupie pacjentów leczonych lekiem A,
- suma rang w grupie pacjentów leczonych lekiem B,
- wartość testu Manna–Whitneya dla liczebności < 20,
- wartość testu Manna–Whitneya dla liczebności > 20,
- poziom istotności dla testu Manna–Whitneya,
- wartość testu skorygowanego (związane jest to z występowaniem rang wiązanych) – dla liczebności > 20,
- poziom istotności wyliczany dla sytuacji, gdy występują rangi wiązane – dla liczebności > 20,
- liczebność grupy pacjentów leczonych lekiem A,
- liczebność grupy pacjentów leczonych lekiem B,
- poziom istotności dla małych liczebności, gdy nie występują rangi wiązane.

Decyzję o przyjęciu lub odrzuceniu H_0 podejmiemy na podstawie wartości zaznaczonej ramką (➥ **ryc. 12.19**). Ponieważ wartość ta jest niższa niż 0,05, decydujemy

się na odrzucenie H_0 i przyjęcie H_1. Oznacza to, że leki w różny sposób wpływają na stężenie glukozy u pacjentów.

Żeby zobaczyć, jak wyglądają różnice między badanymi grupami, warto wykonać wykres. W tym celu klikamy na przycisk Wykres ramka-wąsy (➥ **ryc. 12.20**).



Ryc. 12.20. Porównanie stężenia glukozy dla grupy osób przyjmujących lek A z grupą przyjmującą lek B

Z wykresu wynika, że stężenie glukozy u pacjentów przyjmujących lek B jest wyższy niż w grupie przyjmującej lek A. Warto zwrócić uwagę, że na wykresie przedstawiono mediany jako miary centralne, a pierwszy i trzeci kwartyl (w legendzie jako 25%–75%) oraz rozstęp jako miary rozrzutu wokół mediany.

12.6. Próba zależna – przykład obliczeń

12.6.1. Testy parametryczne

Na początek kilka słów przypomnienia. Pojęciem „pary zależne" określamy pary wyników, które są w jakiś sposób ze sobą powiązane, np. wyniki stężenia hemoglobiny w grupie pacjentów, która została poddana badaniu dwukrotnie w różnych warunkach, np. przed leczeniem i po nim. Oznacza to, że dysponujemy dwoma pomiarami od każdego pacjenta, które zostały zebrane w pewnym odstępie czasowym.

Innym przykładem par zależnych są pomiary np. danej tkanki prawidłowej oraz nowotworu w jej sąsiedztwie.

Zatem żeby sprawdzić, czy zastosowane leczenie wpłynęło w jakiś sposób na stężenie hemoglobiny (podwyższyło je lub obniżyło), należałoby obliczyć różnicę *d* między tymi wartościami (zmienna „Różnice", ➥ **tabela 12.5**).

ID pacjenta	Hemoglobina		Różnice *d* = po leczeniu – przed leczeniem
	przed leczeniem	po leczeniu	
1	12,17	13,97	1,8
2	11,42	14,76	3,34
3	12,25	13,34	1,09
4	10,76	13,05	2,29
5	12,52	12,47	–0,05
6	13,34	12,89	–0,45
7	11,68	15,12	3,44
8	12,6	14,46	1,86
9	11,89	13,77	1,88
10	12,5	14,2	1,7
11	11,86	12,29	0,43
12	11,62	13,96	2,34

Tabela 12.5. Różnice stężenia hemoglobiny przed leczeniem i po nim

Jeśli popatrzymy na różnicę, możemy zauważyć, że u większości pacjentów stężenie hemoglobiny po leczeniu wzrosło. Jedynie u 2 pacjentów (5. i 6.) po leczeniu zanotowano jego zmniejszenie po leczeniu.

Chcielibyśmy jednak wiedzieć, czy zaobserwowane różnice są istotne i wynikają z zastosowania tego konkretnego schematu leczenia, a podobne zmiany obserwowalibyśmy w całej populacji, czy też może są wynikiem przypadku i podobnego stanu rzeczy nie należy się spodziewać w populacji, z której pochodzi próba badana.

Żeby odpowiedzieć na te pytania, należy użyć testu statystycznego, który pomoże nam zweryfikować H_0. W tym przypadku zakłada ona, że zastosowanie leczenia nie wpłynie na stężenie hemoglobiny, nie spowoduje zatem ani jego spadku, ani wzrostu. Możemy ją zapisać w formie równania matematycznego:

$$H_0: D = 0$$

gdzie D oznacza różnicę między wynikami przed leczeniem i po leczeniu w całej populacji.

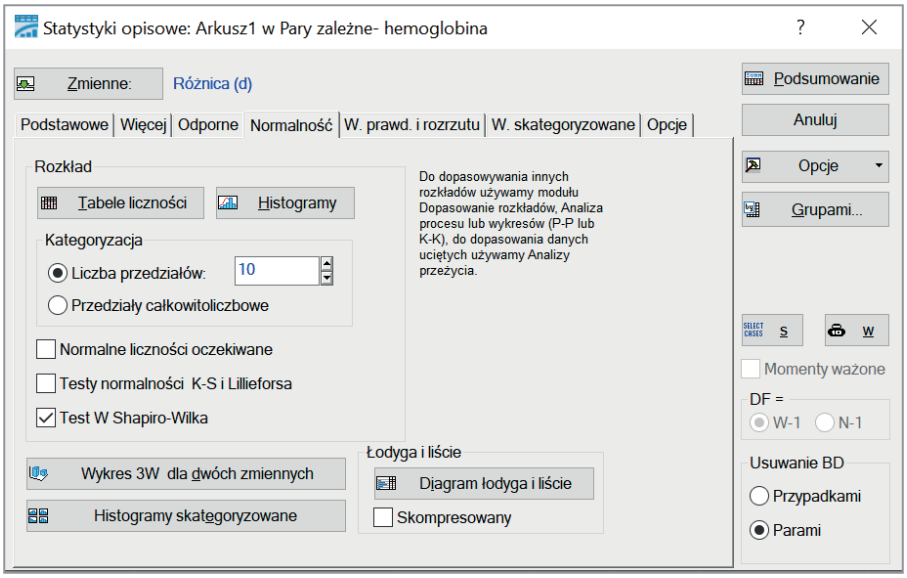
Jak zatem zapisać hipotezę alternatywną?

$$H_1: D \neq 0$$

Tak zapisana H_1 wskazuje, że po zastosowaniu leczenia stężenie hemoglobiny się zmieni, nie precyzuje natomiast, czy wzrośnie czy spadnie. W pierwszej kolejności należy więc obliczyć statystyki opisowe dla obu grup wyników (przed leczeniem i po

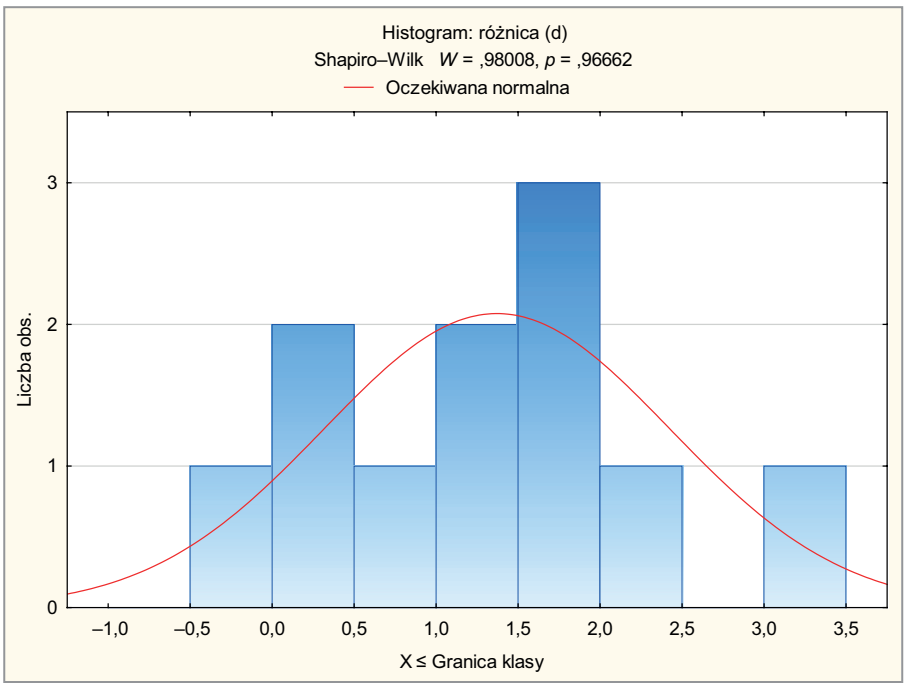
nim): średnie i odchylenia standardowe lub mediany i kwartyle pierwszy i trzeci, gdy rozkłady tych zmiennych różnią się od rozkładu normalnego. Żeby zaś wybrać odpowiedni test statystyczny, musimy również sprawdzić normalność rozkładu różnic (*d*).

W zakładce Statystyka wybieramy Statystyki podstawowe, a następnie Statystyki opisowe. Klikamy przycisk Zmienne i zaznaczamy zmienną „Różnica (*d*)". Wybór zatwierdzamy przyciskiem OK (➥ **ryc. 12.21**).



Ryc. 12.21. Wybór zmiennych w Statystykach opisowych

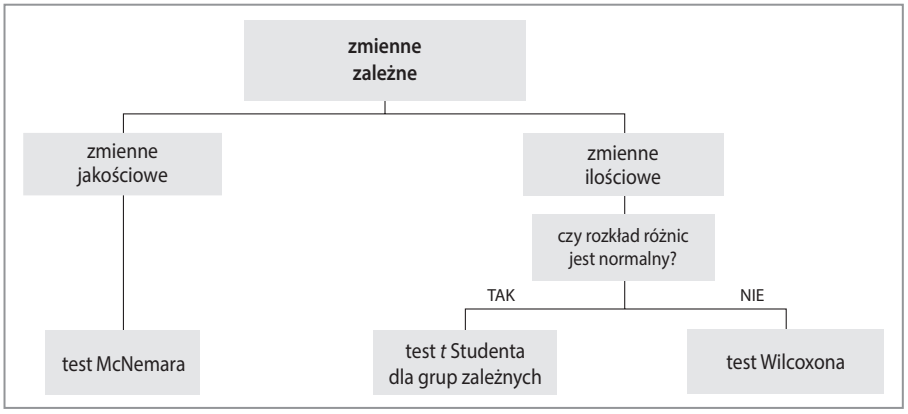
Przechodzimy do zakładki Normalność i zaznaczamy pole wyboru Test W Shapiro–Wilka, a następnie klikamy przycisk Histogramy. Otrzymujemy wykres przedstawiający rozkład różnic wraz z wartością *p* dla testu normalności Shapiro–Wilka (➥ **ryc. 12.22**).



Ryc. 12.22. Rozkład zmiennej „Różnica (*d*)" z wartością testu *W* Shapiro–Wilka

Widzimy już, że zmienna „Różnica (*d*)" charakteryzuje się rozkładem normalnym (wartość $p > 0{,}05$), więc korzystając ze schematu blokowego, możemy dokonać wyboru odpowiedniego testu statystycznego (➥ **ryc. 12.23**).

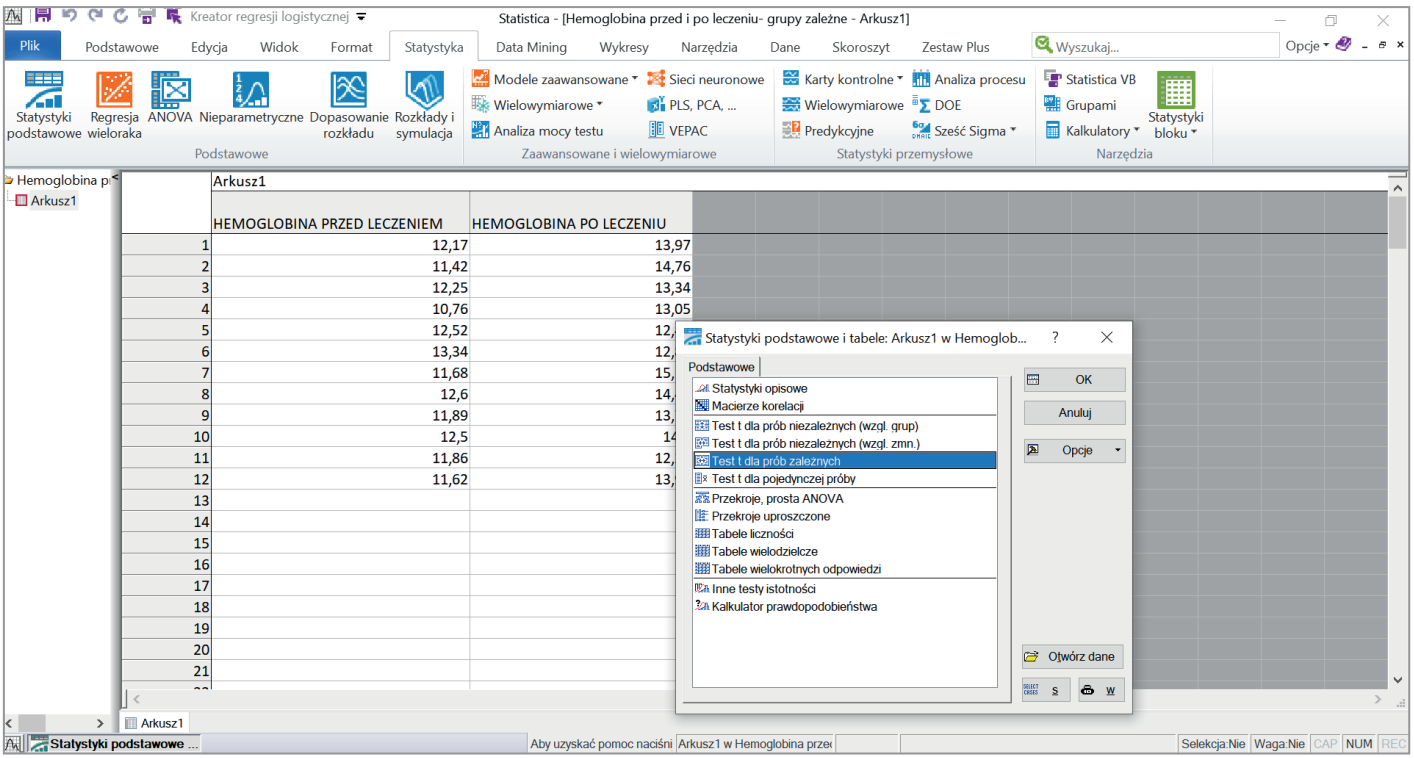
Ryc. 12.23. Schemat blokowy wyboru testu statystycznego dla grup zależnych



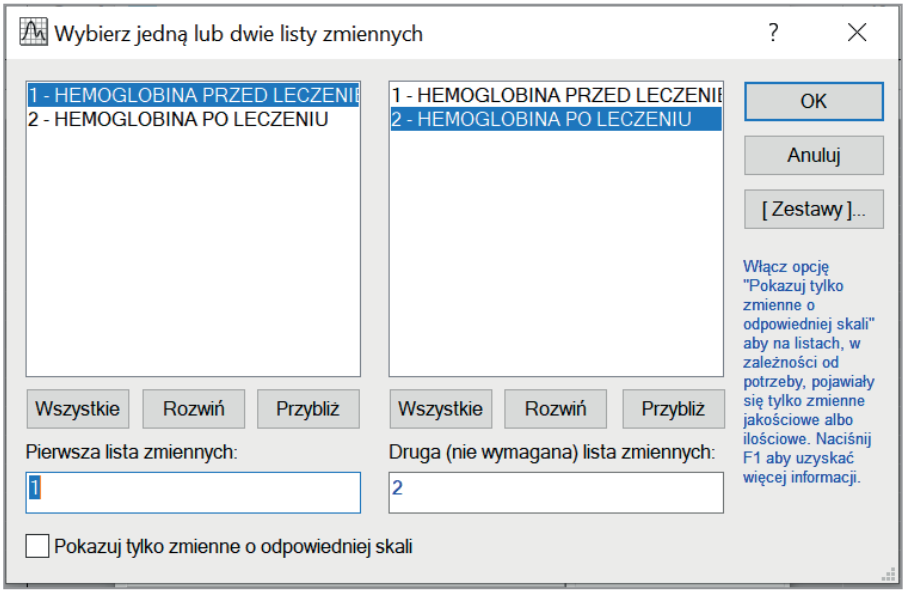
Ponieważ nasze dane (stężenie hemoglobiny przed leczeniem i po nim) są danymi ilościowymi o rozkładzie normalnym, do dalszego porównania należy wykorzystać test *t* Studenta dla grup zależnych.

Z karty Statystyka wybieramy zatem Statystyki podstawowe, a następnie z listy wybieramy Test t dla prób zależnych (➥ **ryc. 12.24**).

Ryc. 12.24. Wybór testu *t* Studenta dla prób zależnych



Przechodzimy do wyboru zmiennych. W tym przypadku najlepiej będzie zaznaczyć na pierwszej liście zmiennych stężenie hemoglobiny przed leczeniem, a na drugiej stężenie hemoglobiny po leczeniu (➥ **ryc. 12.25**). Następnie klikamy OK.



Ryc. 12.25. Wybór zmiennych dla testu *t* Studenta dla prób zależnych

Otrzymujemy arkusz przedstawiający statystyki opisowe oraz wyniki testu *t* dla prób zależnych (➥ **ryc. 12.26**).

Ryc. 12.26. Wyniki testu *t* Studenta dla prób zależnych

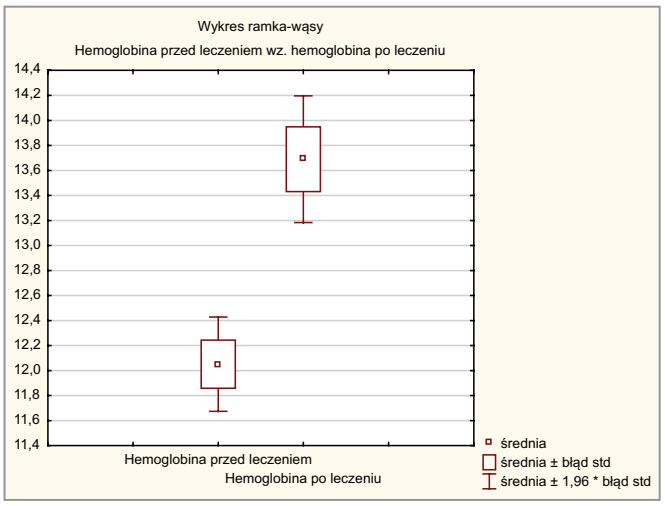
Zmienna	Test T dla prób zależnych (Arkusz1 w Hemoglobina przed i po leczeniu- grupy zależne) Zaznaczone różnice są istotne z p < ,05000									
	Średnia	Odch.st.	Ważnych	Różnica	Odch.st. Różnica	t	df	p	Ufność -95,000%	Ufność +95,000%
HEMOGLOBINA PRZED LECZENIEM	12,05083	0,666667								
HEMOGLOBINA PO LECZENIU	13,69000	0,895494	12	-1,63917	1,212379	-4,68355	11	0,000668	-2,40948	-0,868858

Przeanalizujmy kolejne wartości:

- Średnia – wartości średnich arytmetycznych dla stężenia hemoglobiny przed leczeniem i po nim,
- Odch.st. – wartości odchyleń standardowych dla stężenia hemoglobiny przed leczeniem i po nim,
- Ważnych – liczba przypadków branych do analizy,
- Różnica – różnica między średnimi stężeniami hemoglobiny przed leczeniem i po nim,
- Odch.st. Różnica – odchylenie standardowe dla różnic,
- *t* – wartość testu *t* Studenta dla zmiennych zależnych,
- *df* – liczba stopni swobody dla testu *t* Studenta,
- *p* – poziom prawdopodobieństwa prawdziwości H_0 dla testu *t* Studenta dla par zależnych,
- Ufność –95,000%, +95,000% – przedział 95% ufności dla średniej z wartości różnic między obydwoma pomiarami dla danego pacjenta.

Wartość $p < 0{,}05$ pozwala nam na odrzucenie H_0 oraz przyjęcie H_1. Widzimy, że średnie stężenie hemoglobiny przed leczeniem wynosi średnio 12,05 ± 0,67, natomiast po leczeniu jest to 13,69 ± 0,90. Możemy zatem wysnuć wniosek, że w wyniku leczenia stężenie hemoglobiny wzrasta, a różnica ta jest istotna statystycznie ($p < 0{,}05$).

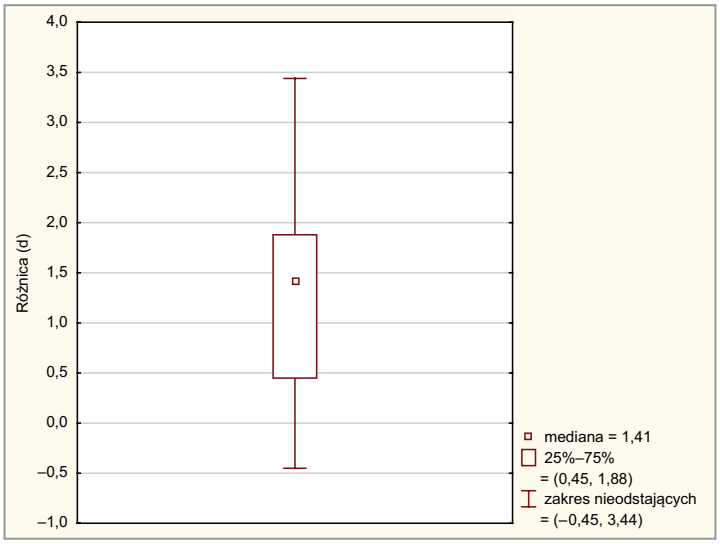
Przedstawmy ten wynik w formie graficznej. W oknie analizy klikamy przycisk Wykresy Ramka-wąsy i otrzymujemy wykres (➥ **ryc. 12.27**).



Ryc. 12.27. Porównanie stężenia hemoglobiny przed leczeniem i po nim

Należy jednak zauważyć, że w teście dla pomiarów powiązanych statystyki opisowe (np. przedstawione w formie wykresów pudełkowych) mają drugorzędne znaczenie. Najważniejsza jest charakterystyka różnic między pomiarami wykonanymi przed interwencją i po niej. Zdarza się czasami, że różnice między pomiarami są bardzo małe, ale konsekwentnie zmieniają się w jednym kierunku, np. stężenie hemoglobiny zmienia się nieznacznie, ale mimo wszystko konsekwentnie zwiększa się w porównaniu z pomiarami wykonanymi przed interwencją. Tego typu zmiany są trudne do dostrzeżenia na wykresie pudełkowym porównującym pomiary przed i po.

Z tego względu dobrze jest zastosować wykres pudełkowy dla różnic (➥ **ryc. 12.28**).

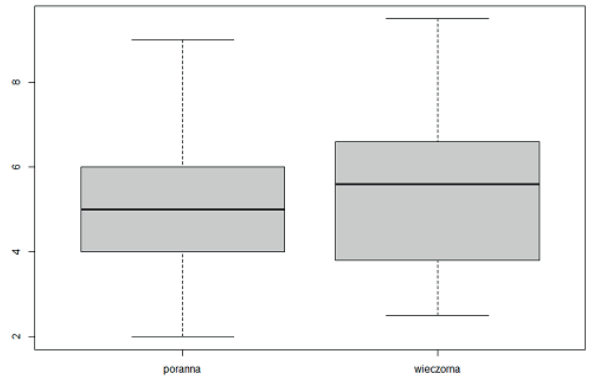


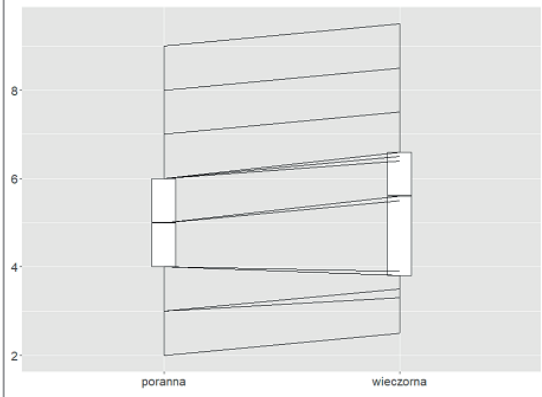
Ryc. 12.28. Wykres pudełkowy dla zmiennej „Różnica (*d*)"

Na wykresie dostrzec można, że mediana dla różnic jest dodatnia, a więc stężenie hemoglobiny po leczeniu u większości osób rośnie w porównaniu z pomiarami wykonanymi przed nim (➥ **ryc. 12.29**). Wykres jest rezultatem zastosowania pakietu „PairedData" w środowisku R.

Ryc. 12.29. *Box plot* z różnic

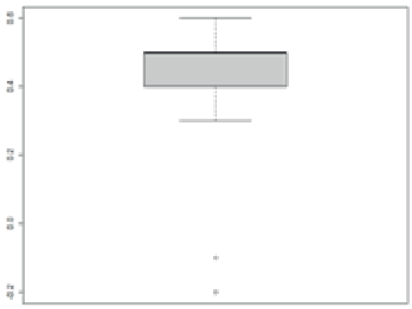
```
poranna wieczorna
1 2 2.5
2 3 3.5
3 3 3.3
4 4 3.9
5 4 3.8
6 5 5.5
7 5 5.6
8 6 6.5
9 6 6.6
10 6 6.4
11 7 7.5
12 8 8.5
13 9 9.5
```






```
    Wilcoxon signed rank test with continuity correction

data: poranna and wieczorna
V = 3, p-value = 0.002811
alternative hypothesis: true location shift is not equal to 0
```



12.6.2. Testy nieparametryczne

Podobnie do prób niezależnych również testy dla prób zależnych mają swoje odpowiedniki nieparametryczne. Stosujemy je w sytuacji, gdy różnice między powiązanymi parami nie mają rozkładu normalnego. Odpowiednikiem nieparametrycznym testu *t* Studenta dla prób zależnych jest test Wilcoxona dla par powiązanych, (➥ **ryc. 12.23**).

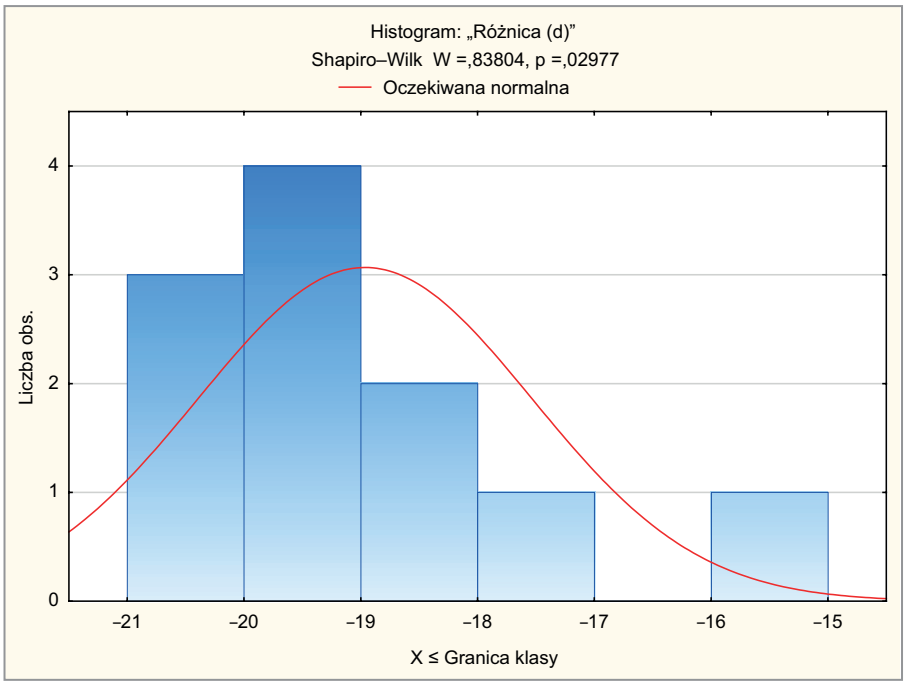
Wyobraźmy sobie, że chcemy porównać wpływ stosowania określonej diety na masę ciała pacjentów. W tym celu ważymy ich na początku badania oraz po 6 miesiącach stosowania diety. Zapis naszych danych wygląda jak w ➥ **tabeli 12.6**.

ID pacjenta	Masa ciała przed dietą	Masa ciała po 6 miesiącach stosowania diety	Różnica (*d*) = masa ciała po 6 miesiącach – masa ciała przed dietą
1	99,5	81,4	–18,1
2	87,7	67,7	–20,0
3	102,3	82,2	–20,1
4	101,1	81,3	–19,8
5	98,3	78,7	–19,6
6	91,3	75,7	–15,6
7	76,5	57,5	–19,0
8	105,3	87,9	–17,4
9	103,9	83,6	–20,3
10	103,1	84,2	–18,9
11	97,3	77,5	–19,8

Tabela 12.6. Dane dotyczące różnicy masy pacjentów w rozważanym przykładzie

W pierwszej kolejności należy zbadać rozkład różnic między masą po diecie oraz przed dietą (zmienna „Różnica (*d*)”).

W karcie Statystyka wybieramy kolejno: Statystyki podstawowe, Statystyki opisowe i zakładkę Normalność. Z dostępnych testów wybieramy test *W* Shapiro–Wilka i klikamy przycisk Histogramy (➥ **ryc. 12.30**).



Ryc. 12.30. Rozkład zmiennej „Różnica (*d*)”

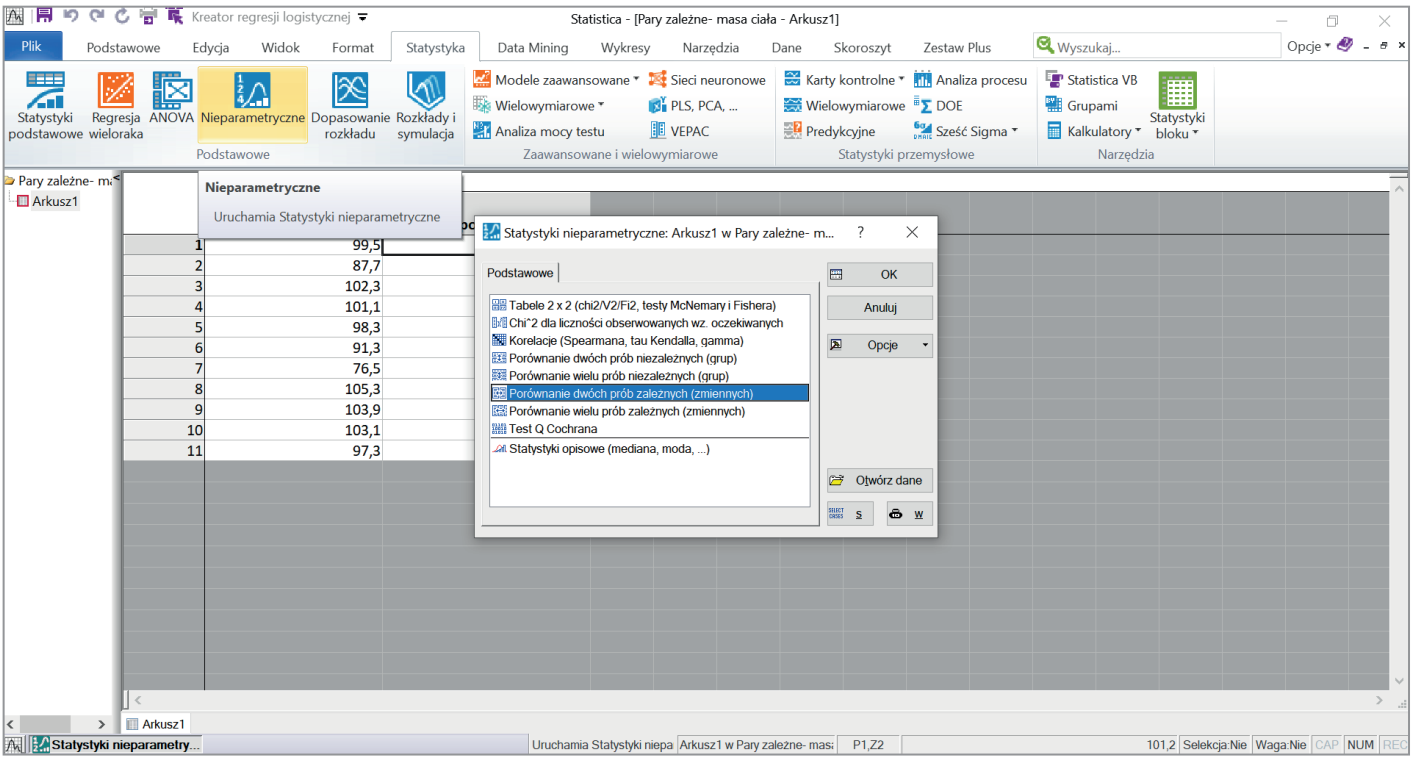
Dla obu kategorii można stwierdzić, że dla testu Shapiro–Wilka $p < 0{,}05$. Na tej podstawie odrzucamy H_0 o rozkładzie normalnym. Jest to cenna informacja, która decyduje o wyborze odpowiedniego testu statystycznego do porównania masy ciała oraz skuteczności stosowanej diety. Posługując się ➥ ryc. 12.23, stwierdzamy, że dla par zależnych o rozkładzie innym niż normalny należy użyć testu Wilcoxona dla par zależnych. Wykorzystujemy go również w przypadku, gdy zmienne mierzone są na skali porządkowej (w teście *t* Studenta zmienne muszą być mierzone na skali ilościowej). Jest oparty na rangowaniu różnic pomiędzy 1. i 2. pomiarem. Hipoteza zerowa w naszym przypadku, podobnie jak dla testu *t* Studenta, brzmi:

$$H_0\colon D = 0$$

a więc masa ciała nie zmieniła się w efekcie stosowania diety.

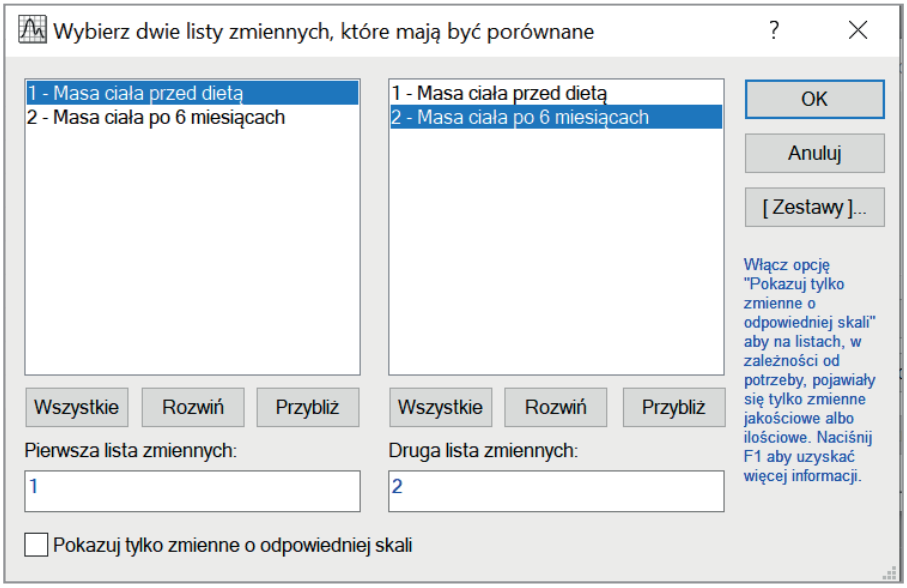
Żeby w programie Statistica wykonać test Wilcoxona dla par zależnych, należy skorzystać z polecenia Nieparametryczne i wybrać Porównanie dwóch prób zależnych (zmiennych), ➥ **ryc. 12.31**.

Ryc. 12.31. Wybór testu Wilcoxona dla par zależnych



Dokonujemy wyboru zmiennych, tak by na 1. liście znalazły się wartości przed dietą, a na 2. – zmienna „Masa ciała po 6 miesiącach stosowania diety" (➥ **ryc. 12.32**).

Ryc. 12.32. Wybór zmiennych dla testu Wilcoxona



Ryc. 12.33. Wyniki analizy testu Wilcoxona dla par zależnych

Następnie klikamy przycisk Test kolejności par Wilcoxona i otrzymujemy arkusz (➥ **ryc. 12.33**).

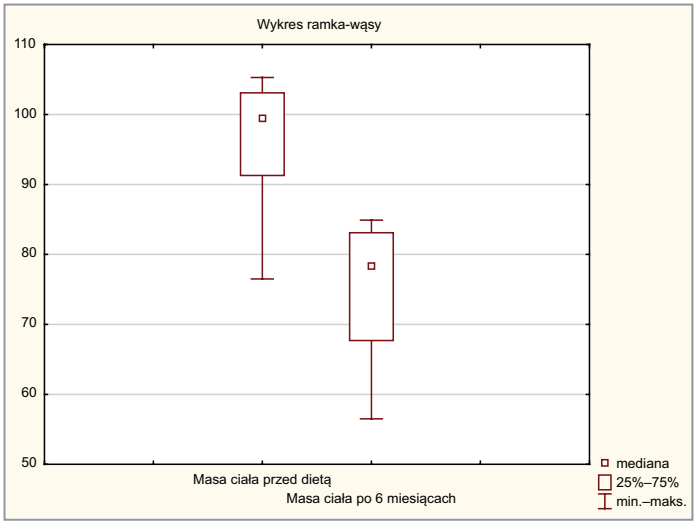
Para zmiennych	Test kolejności par Wilcoxona (Arkusz1 w Pary zależne- masa ciała) Zaznaczone wyniki są istotne z p <,05000			
	N Ważnych	T	Z	p
Masa ciała przed dietą & Masa ciała po 6 miesiącach	11	0,00	2,934058	0,003346

Przeanalizujmy kolejne pola:

- *N* ważnych – liczebność grup (czyli liczba porównywanych par);
- *T* – wartość testu Wilcoxona dla grup o liczebności ≤ 25;
- Z – wartość testu Wilcoxona dla grup o liczebności > 25;
- *p* – wartość prawdopodobieństwa testowego dla testu Wilcoxona.

Ponieważ w naszym przypadku wartość *p* jest mniejsza niż przyjęty wcześniej poziom istotności, tj. 0,05, możemy odrzucić H_0 i przyjąć H_1, która mówi, że stosowanie diety wpływa na masę ciała. Ale czy masa ciała spada, czy rośnie w wyniku stosowania diety? Żeby się tego dowiedzieć, należy obliczyć wybrane statystyki opisowe lub wykonać wykres. Trzeba pamiętać, że przy rozkładzie danych w co najmniej jednej grupie pomiarów różniących się od rozkładu normalnego zamiast średniej arytmetycznej należy zastosować medianę, a jako miarę rozproszenia wybrać kwartyle i rozstęp (➥ **ryc. 12.34**).

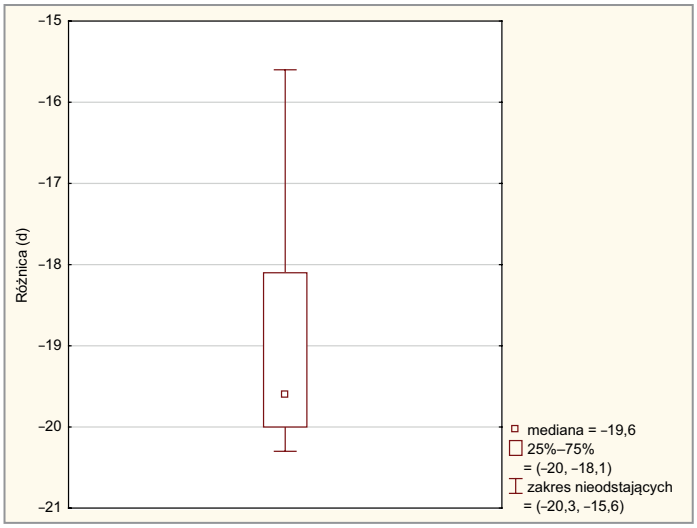
Ryc. 12.34. Wpływ stosowania diety na masę ciała



Z wykresu wynika, że masa ciała wyraźnie spadła w ciągu 6 miesięcy stosowania diety. Ponieważ jednak, jak wspomniano powyżej, wykres pudełkowy jest zawodny, kiedy zmiany między pomiarami przed i po są niewielkie, mimo iż konsekwentnie rosną lub maleją, należy wykonać wykres pudełkowy dla różnic (➥ **ryc. 12.35**).

Na wykresie dostrzec można, że u wszystkich pacjentów zanotowano utratę masy ciała w wyniku stosowania diety, przy czym 75% straciło ponad 18 kg.

Ryc. 12.35. Wykres pudełkowy dla zmiennej „Różnica (*d*)"



Tomasz Hikawczuk

13 Jednoczynnikowa analiza wariancji w medycynie

Cel ćwiczenia. Określenie różnic statystycznie istotnych pomiędzy grupami doświadczalnymi przy zastosowaniu jednoczynnikowej analizy wariancji lub jej nieparametrycznego odpowiednika.

Badania retrospektywne w postaci badań kliniczno-kontrolnych, gdzie wyniki są zgromadzone w bazach danych, umożliwiają m.in. przeprowadzenie prostej jednoczynnikowej analizy wariancji. W analizie wariancji nie jest wymagana równa liczba obserwacji w badanych grupach doświadczalnych. Ułatwia to w znacznym stopniu zastosowanie tego typu testów statystycznych. Ważnym kryterium doboru testu (parametrycznego lub nieparametrycznego) w tym przypadku jest występowanie rozkładu normalnego we wszystkich porównywanych grupach. Jeśli liczba obserwacji nie przekracza 50 w grupie doświadczalnej, wówczas najbardziej odpowiednim testem statystycznym do określenia normalności rozkładu danych będzie test Shapiro–Wilka. Gdy mamy do czynienia z większą liczbą obserwacji (powyżej 50), znacznie bardziej przydatny będzie test Kołmogorowa–Smirnowa.

W większości układów doświadczalnych z dwoma lub większą liczbą czynników doświadczalnych niezbędne jest przygotowanie odpowiedniego planu doświadczenia umożliwiającego zastosowanie zrandomizowanego układu doświadczalnego, a następnie przeprowadzenie analizy wariancji w oparciu o moc testu w zależności od liczby replikacji w poszczególnych grupach i standaryzowanej miary efektu (RMSSE).

W zastosowaniu analizy wariancji przydatny jest rozkład Fishera–Snedecora i wartość F testu, która wyraża stosunek pomiędzy zmiennością międzygrupową i wewnątrzgrupową w przeprowadzanym doświadczeniu. Klasyczna analiza wariancji oprócz normalności rozkładu danych w poszczególnych grupach doświadczalnych wymaga również homogenności wariancji. Brak jednego z tych uwarunkowań nie neguje zastosowania testu analizy wariancji, jednak w tym przypadku będzie to analiza oparta na wartościach rangowanych, tj. test Kruskala–Wallisa. Są jednak poprawki umożliwiające wykonanie testów post-hoc nawet jeśli wariancje nie są jednakowe, np. test Gamesa–Howella lub Dunnetta T3.

13.1. Zmienności: międzygrupowa, wewnątrzgrupowa i ogólna

Zastosowanie testu analizy wariancji pozwala na podzielenie zmienności ogólnej w danej próbie (zbiorze danych) na zmienności międzygrupową i wewnątrzgrupową, a następnie na analizę ich sumy kwadratów w zależności od liczby stopni swobody przy zastosowaniu testu F (➥ **tabela 13.1**). Zmienność międzygrupowa jest efektem działania czynnika doświadczalnego i wpływa na różnicowanie się wartości średnich w poszczególnych grupach doświadczalnych. Drugim rodzajem zmienności jest zmienność wewnątrzgrupowa powstała w wyniku działania czynników losowych nieuwzględnionych w doświadczeniu, a wynikająca ze zmienności osobniczej. Cały sens obliczeń sprowadza się w pierwszej kolejności do określenia sumy kwadratów zmienności całkowitej, która następnie może być podzielona na sumę kwadratów zmienności wewnątrzgrupowej (*SS* błędu) i sumę kwadratów

zmienności międzygrupowej (*SS* efektu). Następnie po uwzględnieniu stopni swobody w przypadku każdego z rodzajów zmienności wartość sumy kwadratów dzieli się przez liczbę stopni swobody, w wyniku czego uzyskuje się wartość średniego kwadratu, która następnie umożliwia obliczenie wartości *F* testu Fishera–Snedecora (➥ **tabela 13.1**).

Zmienna	**Suma kwadratów (*SS*)**	**Stopnie swobody**	**Średni kwadrat (*MS*)**	**Test *F***
Całkowita	całkowita	$n - 1$	–	$F = \frac{MS \text{ efektu}}{MS \text{ błędu}}$
Międzygrupowa	efektu	$k - 1$	$MS = \frac{SS \text{ efektu}}{k - 1}$	
Wewnątrzgrupowa	błędu	$n - k$	$MS = \frac{SS \text{ błędu}}{n - k}$	
n – liczba obserwacji, k – liczba grup				

Tabela 13.1. Sposób porównania zmienności międzygrupowej i wewnątrzgrupowej

Oczywiście dziś nie przelicza się sum kwadratów w sposób klasyczny. Dane są przenoszone do programu komputerowego, który oblicza wszystkie potrzebne wartości.

13.2. Jednoczynnikowa analiza wariancji

Najprostszym rodzajem analizy wariancji jest analiza jednoczynnikowa. Umożliwia ona przygotowanie układu doświadczalnego składającego się z kilku grup doświadczalnych, przy czym liczba replikacji nie musi być w nich równa. Założeniami poprawnie przeprowadzonej analizy wariancji jest losowy dobór elementów próby, normalność rozkładu w każdej z nich oraz jednakowość wariancji we wszystkich grupach. Dobór elementów do doświadczenia ma charakter losowy (pacjentów, zwierząt), ważnym kryterium w tym przypadku jest normalność rozkładu i homogenność wariancji w badanych grupach. Efektem końcowym danego doświadczenia jest wyznaczenie modelu doświadczalnego w postaci równania:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

gdzie y_{ij} to wartość obserwowanej zmiennej zależnej y (i oznacza liczbę elementów danej grupy, natomiast j liczbę grup); μ to ogólna średnia, wpływ czynników wspólnych; α_i – wpływ grupy; ε_{ij} – wpływ czynników losowych.

Wystąpienie jednorodności wariancji jest zwykle rzadsze w przypadku retrospektywnych danych obserwacyjnych gromadzonych w danym szpitalu lub przychodni, gdyż mogą pochodzić one z różnych populacji. W przypadku wystąpienia rozkładu normalnego i niejednorodności wariancji, analogicznie do testu *t* Studenta stosujemy poprawkę Welcha. Przy braku normalności rozkładu powinno się zastosować analizę wariancji za pomocą testu Kruskala–Wallisa, który bazuje na rangach wartości i nie wymaga rozkładu normalnego. Opracowano także testy, które umożliwiają analizę istotności różnic między poszczególnymi grupami. Często używa się wówczas testu Dunna z poprawką Bonferroniego kontrolującą błąd I rodzaju, określanego w Statistice ogólnikowo „porównaniami wielokrotnymi".

13.2.1. Przykład

W celu porównania stężenia hemoglobiny w 3 grupach mężczyzn chorych na białaczkę (grupy utworzono ze względu na stopień zaawansowania choroby: grupa I – 1. stopień, grupa II – 2. stopień, grupa III – 3. stopień) wylosowano próby o liczebności 10 osób. Uzyskano wartości podane w ➥ **tabeli 13.2**.

Grupa I	Grupa II	Grupa III
11,8	11,5	10,5
11,5	10,9	10,7
10,8	10,9	10,8
12,1	11,1	10,2
10,9	10,4	10,3
11,8	10,7	10,1
11,6	11,2	10,1
11,9	11,7	10,8
10,9	11,4	10,4
11,7	10,2	10,3

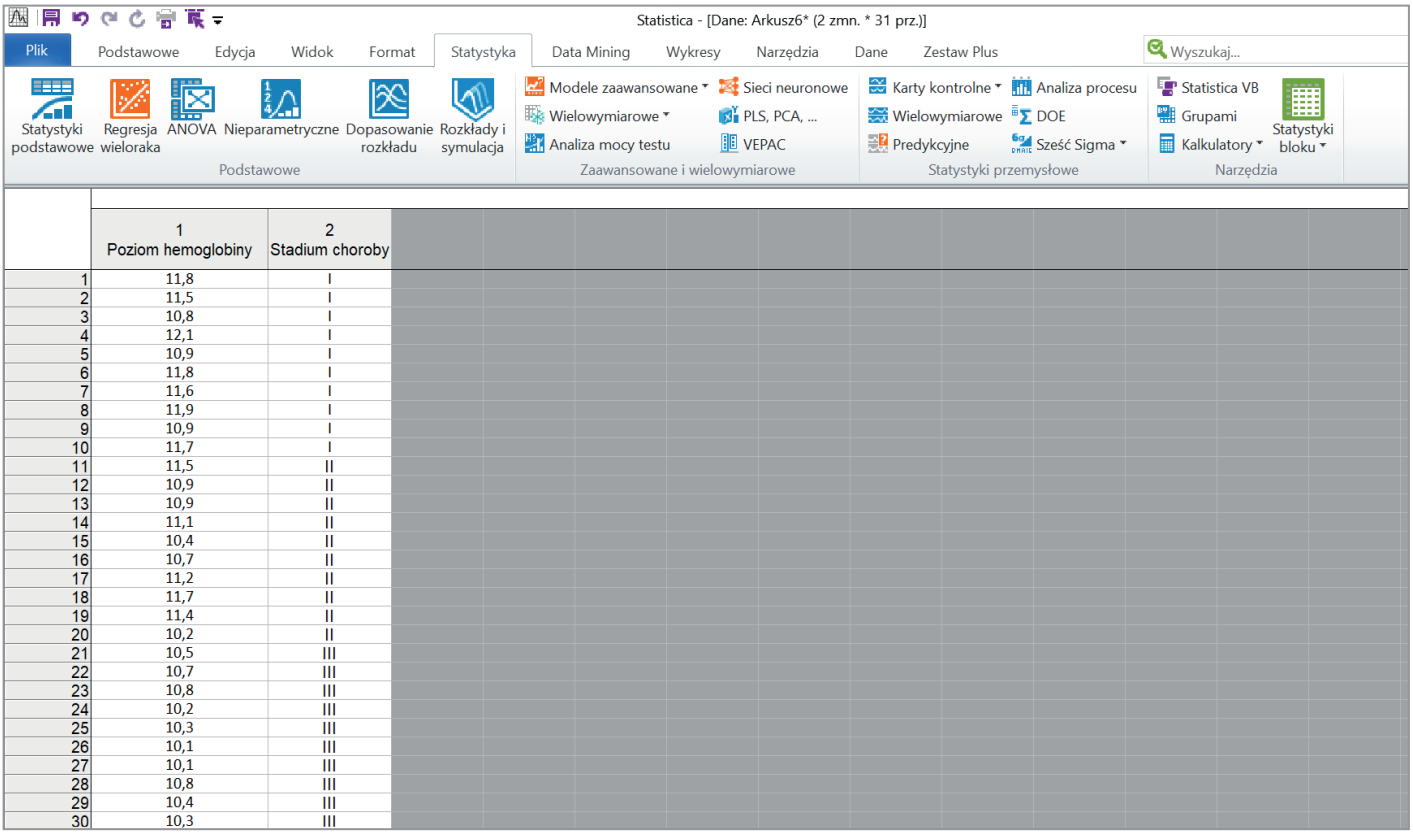
Tabela 13.2. Wartości stężenia hemoglobiny w 3 grupach doświadczalnych

Proszę sprawdzić, czy stężenie hemoglobiny różni się istotnie w poszczególnych stadiach choroby.

Na samym początku formułujemy H_0: brak różnic w wartościach stężenia hemoglobiny we krwi pacjentów pomiędzy 3 grupami osób chorych na białaczkę zróżnicowanych w zależności od stadium zaawansowania choroby.

Analizę rozpoczynamy od przeniesienia danych z pliku przygotowanego w Excelu. Nagłówki można nazwać bezpośrednio w programie Statistica lub otworzyć w programie plik zawierający już gotowe zweryfikowane dane. W tym przypadku w pierwszej kolumnie znajduje się zmienna zależna – stężenie hemoglobiny, natomiast w drugiej zmienna grupująca – stadium choroby (➥ **ryc. 13.1**).

Ryc. 13.1. Stężenie hemoglobiny u poszczególnych pacjentów w grupach doświadczalnych w zależności od stopnia zaawansowania białaczki

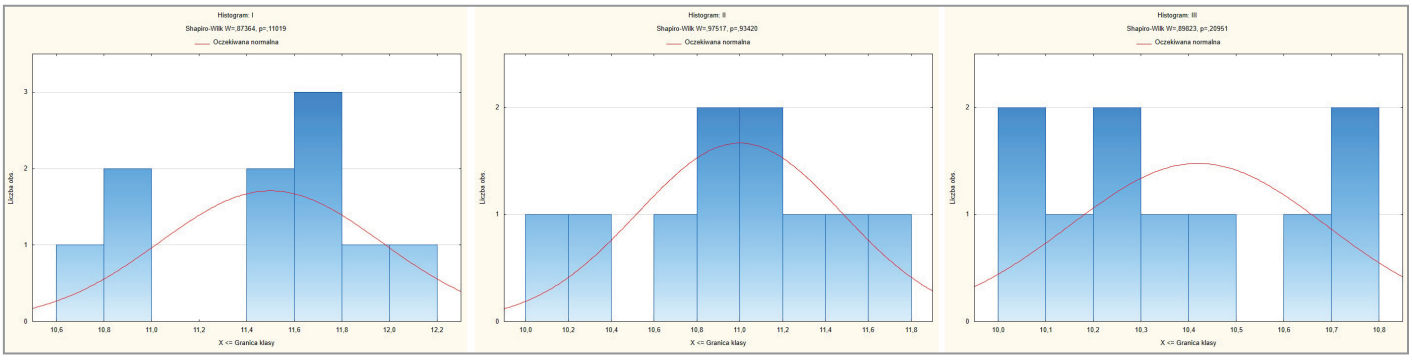


	1 Poziom hemoglobiny	2 Stadium choroby
1	11,8	I
2	11,5	I
3	10,8	I
4	12,1	I
5	10,9	I
6	11,8	I
7	11,6	I
8	11,9	I
9	10,9	I
10	11,7	I
11	11,5	II
12	10,9	II
13	10,9	II
14	11,1	II
15	10,4	II
16	10,7	II
17	11,2	II
18	11,7	II
19	11,4	II
20	10,2	II
21	10,5	III
22	10,7	III
23	10,8	III
24	10,2	III
25	10,3	III
26	10,1	III
27	10,1	III
28	10,8	III
29	10,4	III
30	10,3	III

Przed rozpoczęciem analizy wariancji należy sprawdzić rozkłady danych w grupach doświadczalnych. Ważne jest, aby na podstawie obserwacji dla każdej grupy w kolumnie „Stadium choroby” rozkłady były podane oddzielnie.

W przypadku tego typu danych i przy liczbie obserwacji do 50 najbardziej odpowiednim testem jest test Shapiro–Wilka, który weryfikuje H_0 o zgodności z rozkładem normalnym na podstawie wartości p tego testu. Jeśli znajduje się ona w obszarze krytycznym ($p < 0{,}05$), wówczas H_0 odrzuca się, co jest równoznaczne ze stwierdzeniem, że rozkład danych badanej zmiennej nie jest zgodny z normalnym. W pozostałych przypadkach ($p > 0{,}05$) H_0 jest przyjmowana. W odniesieniu do analizy wariancji wszystkie 3 grupy doświadczalne wykazują wartości wyższe od wartości granicznej obszaru krytycznego (0,05), stąd też w każdym z przypadków mamy do czynienia z rozkładem normalnym (➥ **ryc. 13.2**).

Ryc. 13.2. Histogramy dopasowane do krzywej normalności oraz wartość p testu Shapiro–Wilka



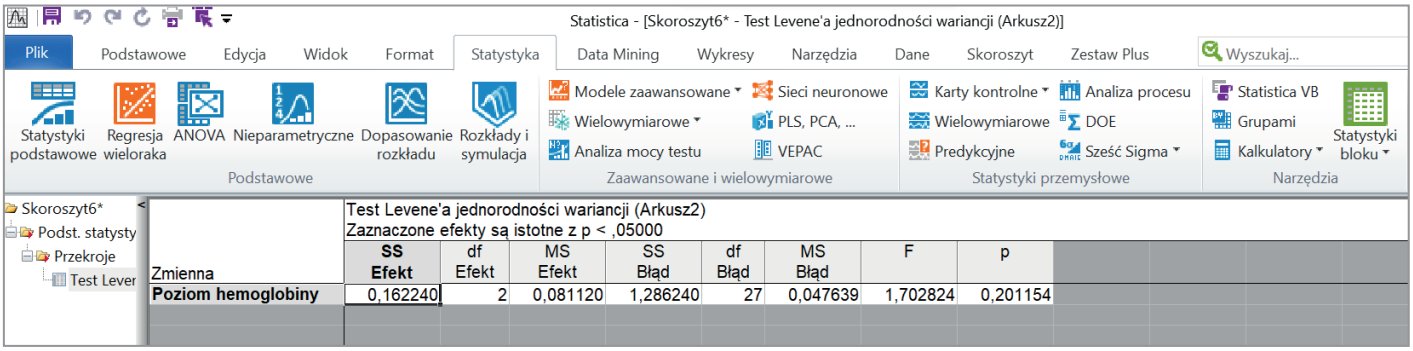
W przykładzie grupa doświadczalna zawiera tylko 10 elementów ($n = 10$), zatem trzeba pamiętać, że możliwości tego testu co do wykrywania różnic względem rozkładu normalnego są silnie ograniczone – jego moc jest wówczas mała (w przypadku tego zbioru damych wynosi 0,14). Dlatego przy tak małych próbach należy stosować nieparametryczną analizę wariancji.

Analiza wariancji wymaga dodatkowo sprawdzenia jednorodności wariancji w poszczególnych grupach. Standardowo w tym celu stosowany jest test Levene'a. Brak homogenności wariancji w grupach doświadczalnych wymaga zastosowania w porównaniu istotności różnic w grupach testu Welcha, a także Gamesa–Howella i Dunnetta T3. Wspomniano wcześniej, że niskie liczby replikacji w grupach doświadczalnych sugerują nam zastosowanie nieparametrycznej analizy wariancji, natomiast dla celów ćwiczeniowych spróbujemy sprawdzić jednorodność wariancji testem Levene'a.

Aby przejść do tego testu, należy na wstążce rozpocząć od wyboru karty Statystyka, następnie wybrać Statystyki podstawowe i w pojawiającym się oknie kliknąć 7. możliwość od dołu lub od góry: Przekroje, prosta ANOVA. Po zatwierdzeniu wybieramy zmienne: „zależna: poziom hemoglobiny" i „grupująca: stadium choroby". Następnie po kolejnym zatwierdzeniu i przejściu do zakładki Testy Anova wybieramy test Levene'a.

W przypadku danych dotyczących stężenia hemoglobiny (➥ **ryc. 13.3**), weryfikując H_0 pod kątem homogenności wariancji, stwierdzamy, że nie ma podstaw do jej odrzucenia ($p = 0{,}20$). Stąd też w dalszym kroku można przejść do testu jednoczynnikowej analizy wariancji, opierając się na dystrybuancie testu F i wartości p odczytanej na podstawie liczby stopni swobody.

Ryc. 13.3. Wyniki testu Levene'a weryfikujące jednorodność wariancji

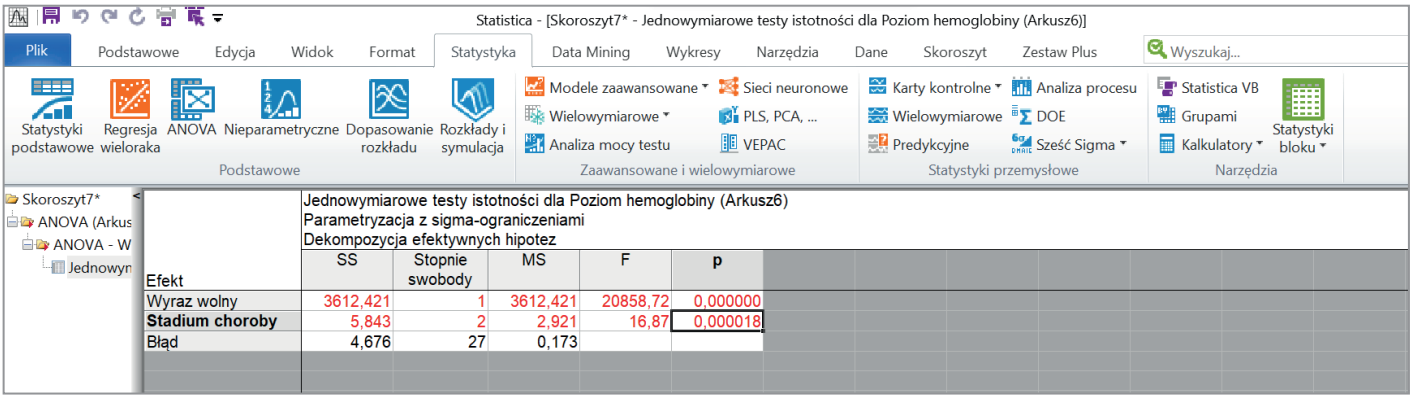


Test Levene'a jednorodności wariancji (Arkusz2)
Zaznaczone efekty są istotne z p < ,05000

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
Poziom hemoglobiny	0,162240	2	0,081120	1,286240	27	0,047639	1,702824	0,201154

Aby przejść do jednoczynnikowej analizy wariancji, należy na wstążce rozpocząć od wyboru karty Statystyka, a następnie użyć polecenia ANOVA. W wyskakującym okienku wybrać opcję Jednoczynnikowa analiza wariancji. Po zatwierdzeniu wybieramy zmienne: „zależna: poziom hemoglobiny" i „grupująca: stadium choroby", i klikamy Wszystkie Efekty. Otrzymujemy wartości z ➥ **ryc. 13.4**.

Ryc. 13.4. Wyniki testu jednoczynnikowej analizy wariancji



Jednowymiarowe testy istotności dla Poziom hemoglobiny (Arkusz6)
Parametryzacja z sigma-ograniczeniami
Dekompozycja efektywnych hipotez

Efekt	SS	Stopnie swobody	MS	F	p
Wyraz wolny	3612,421	1	3612,421	20858,72	0,000000
Stadium choroby	5,843	2	2,921	16,87	0,000018
Błąd	4,676	27	0,173		

W tym przypadku w pierwszej kolumnie otrzymujemy wartości sumy kwadratów (*SS*) dla stadium chorobowego i dla błędu, odpowiednio: 5,843 i 4,676. Liczba stopni swobody pozwala na określenie średniego kwadratu z sumy: dla stadium chorobowego z liczby zabiegów (k) wynosi $k - 1$, czyli $3 - 1 = 2$.

Natomiast dla błędu z liczby replikacji liczba swobody jest równa $n - k = 30 - 3 = 27$ (➥ **tabela 13.1**).

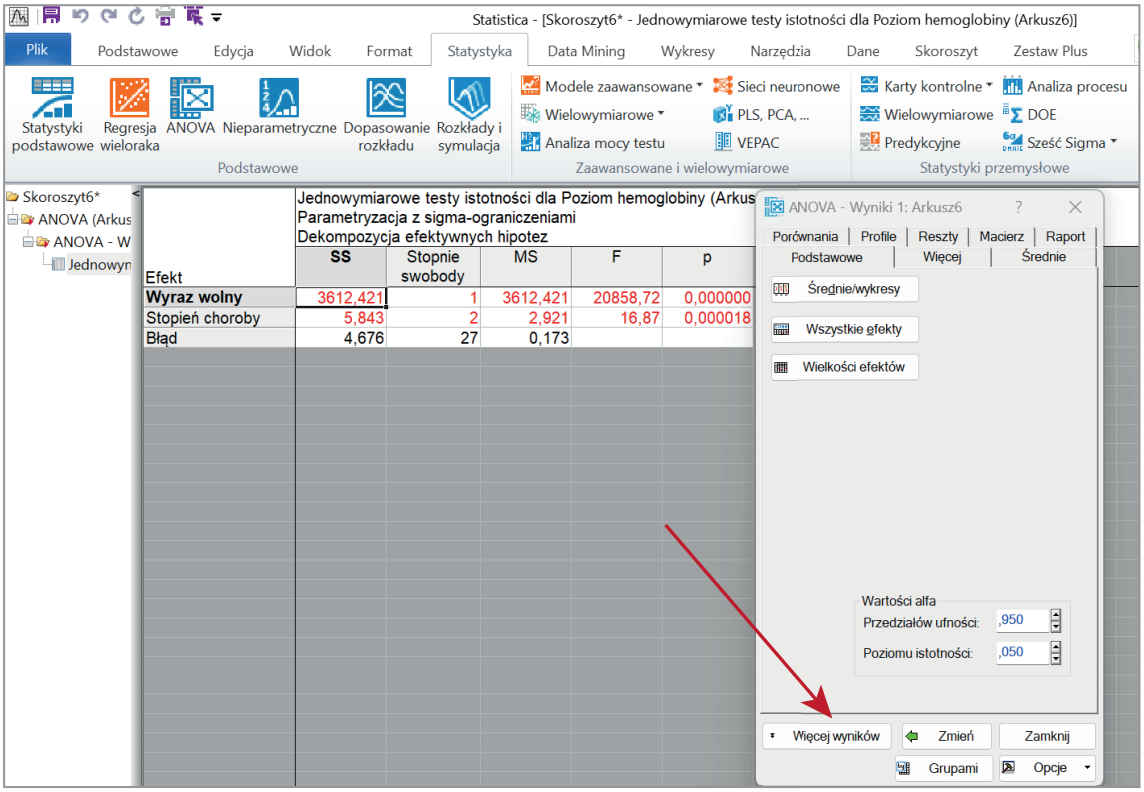
W konsekwencji uzyskujemy wartość średniego kwadratu (*ŚK*: 2,92 dla czynnika doświadczalnego i 0,173 dla błędu), co po podstawieniu do wzoru:

$$F = \frac{MS\ grup}{MS\ błędu}$$

pozwala na określenie $F = 16{,}87$ i odczyt $p = 0{,}000$. Jest to istotność statystyczna wpływu stopnia rozwoju choroby wyznaczana za pomocą tzw. testu ogólnego ANOVA.

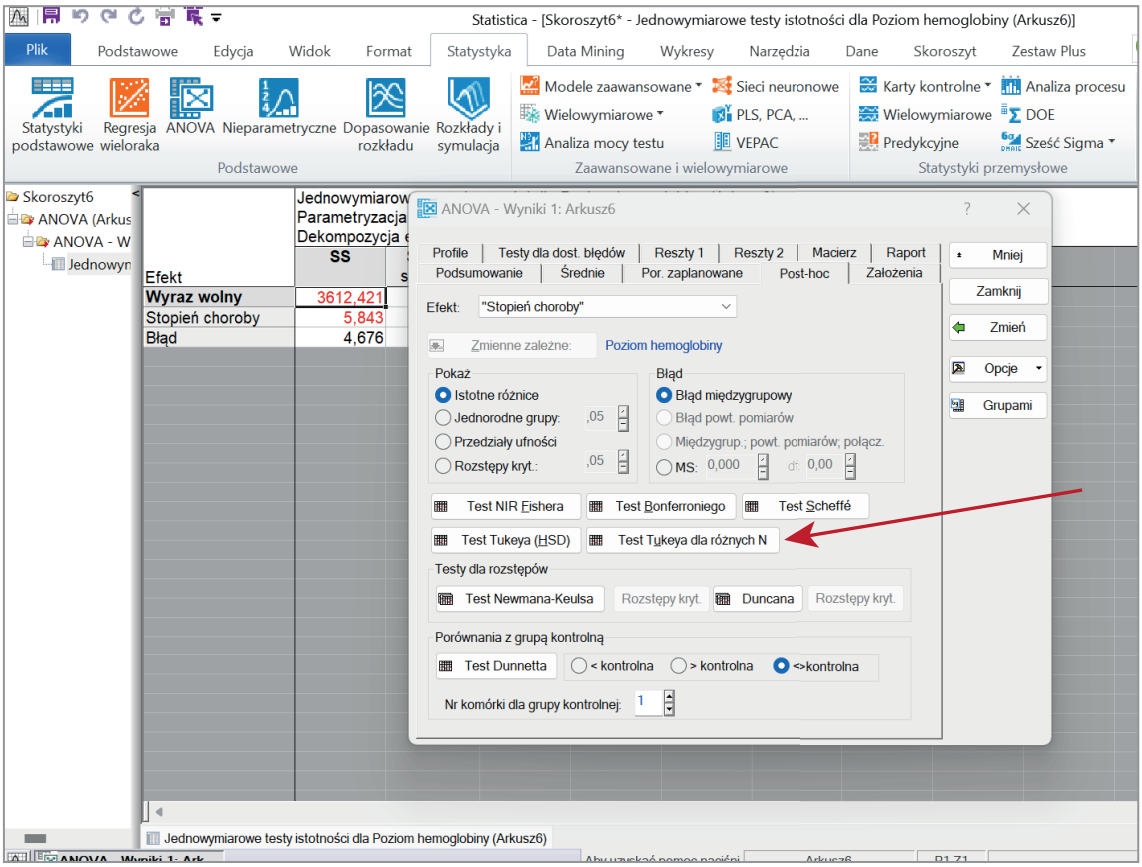
W następnym kroku niezbędne może być zastosowanie tzw. testu *post hoc* do dokładnego określenia różnic pomiędzy poszczególnymi grupami. Najlepszym testem do porównania poszczególnych grup będzie test HSD Tukeya, który zestawia pary pomiarów, a poziom błędu doświadczenia jest ustalany dla każdego z porównań na poziomie zbioru. Jest on o wiele bardziej konserwatywny niż testy LSD Fishera i Duncana, których dokładność jest kwestionowana. Spośród testów post hoc duże znaczenie ma również test Bonferroniego, zwłaszcza gdy liczba replikacji w grupach jest niewielka, np. wynosi trzy powtórzenia. W przypadku tego testu poziom istotności p = 0,05 jest dzielony przez trzy i za istotny statystycznie uważany jest wynik poniżej 0,0167.

W tym celu w lewym dolnym rogu wybieramy ANOVA – Wyniki, tak aby powrócić do okna wyboru, i klikamy Więcej wyników (➥ **ryc. 13.5**). Następnie można wybrać zakładkę Post-hoc i wybrać odpowiedni test w zależności od liczby grup doświadczalnych i zakładanej konserwatywności (Duncana, HSD Tukeya lub inny).



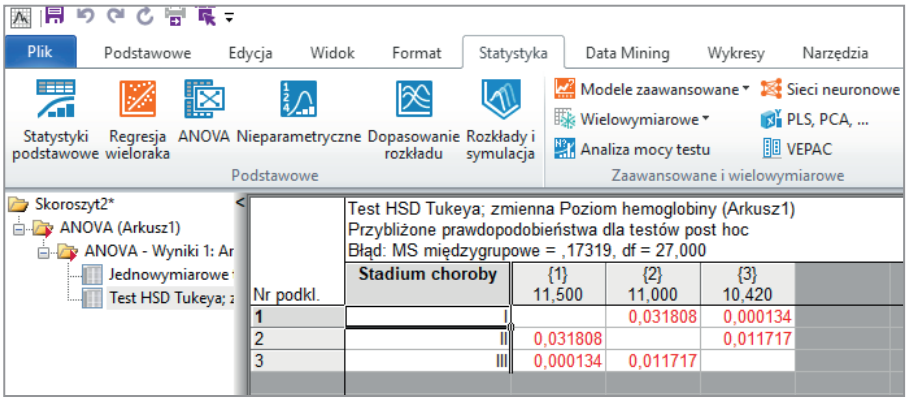
Ryc. 13.5. Wybór testu *post hoc*

W tym przypadku mamy 3 grupy doświadczalne. Można wybrać test Duncana, który jest mniej konserwatywny niż test HSD Tukeya (➥ **ryc. 13.6**).



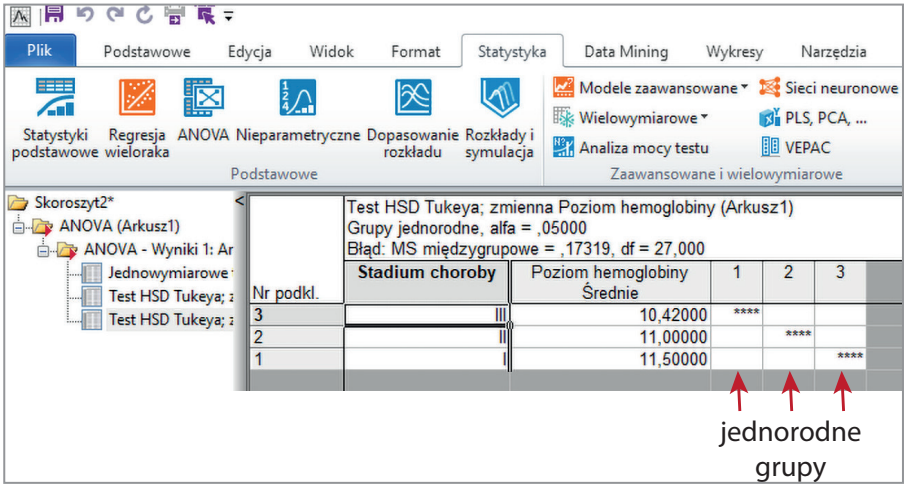
Ryc. 13.6. Zakładka testów *post hoc* i wybór określonego rodzaju testu

W efekcie otrzymujemy zestawienie w odniesieniu do wszystkich trzech grup doświadczalnych z zastosowaniem testu rozstępów HSD (ang. Honestly Significant Difference) Tukeya pomiędzy poszczególnymi grupami doświadczalnymi określające dokładnie, pomiędzy którymi różnica istotna statystycznie występuje na poziomie istotności $p < 0{,}05$ lub $p < 0{,}01$ (➥ **ryc. 13.7**).



Ryc. 13.7. Test HSD Tukeya w postaci testów rozstępu

Dodatkowo w celu łatwiejszej interpretacji różnic statystycznych pomiędzy grupami na poziomie $p < 0{,}05$ możliwe jest wyświetlenie jednorodnych grup, dzięki czemu można stwierdzić statystycznie istotne różnice pomiędzy wszystkimi grupami doświadczalnymi (➥ **ryc. 13.8**). Gwiazdki oznaczają znaczące różnice na poziomie istotności $p < 0{,}05$, co można równoznacznie przedstawić w formie tabeli z zapisem literowym w postaci indeksów górnych: a, b, c. Różne litery przy wartości średniej oznaczają istotną statystycznie różnicę pomiędzy grupami (jednakowe litery określałyby brak różnic pomiędzy grupami), czyli w tym przypadku każde stadium różni się pomiędzy sobą statystycznie (➥ **tabela 13.3**).



Ryc. 13.8. Wyświetlenie jednorodnych grup (1, 2, 3) w teście HSD Tukeya i określenie różnic statystycznych w poszczególnych stadiach białaczki. Każda z cyfr oznacza oddzielną grupę na poziomie istotności $p < 0{,}05$

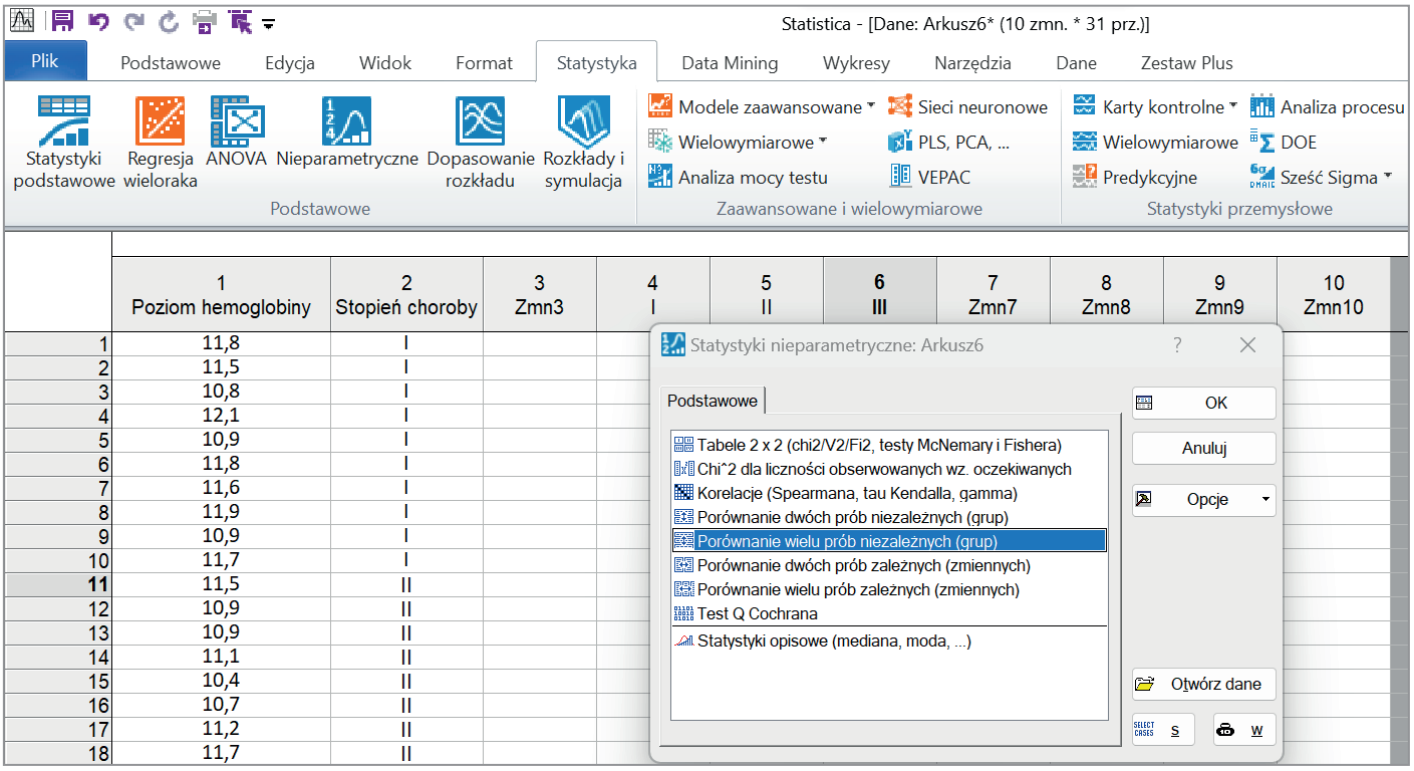
Stadium I	Stadium II	Stadium III
$11{,}5^{a}$	$11{,}0^{b}$	$10{,}4^{c}$

Tabela 13.3. Określenie istotności różnic pomiędzy poszczególnymi stadiami białaczki

Widzimy zatem, że wraz z kolejnym stadium białaczki następuje obniżenie stężenia hemoglobiny u osób z danej grupy doświadczalnej i różnice są istotne pomiędzy każdą z grup.

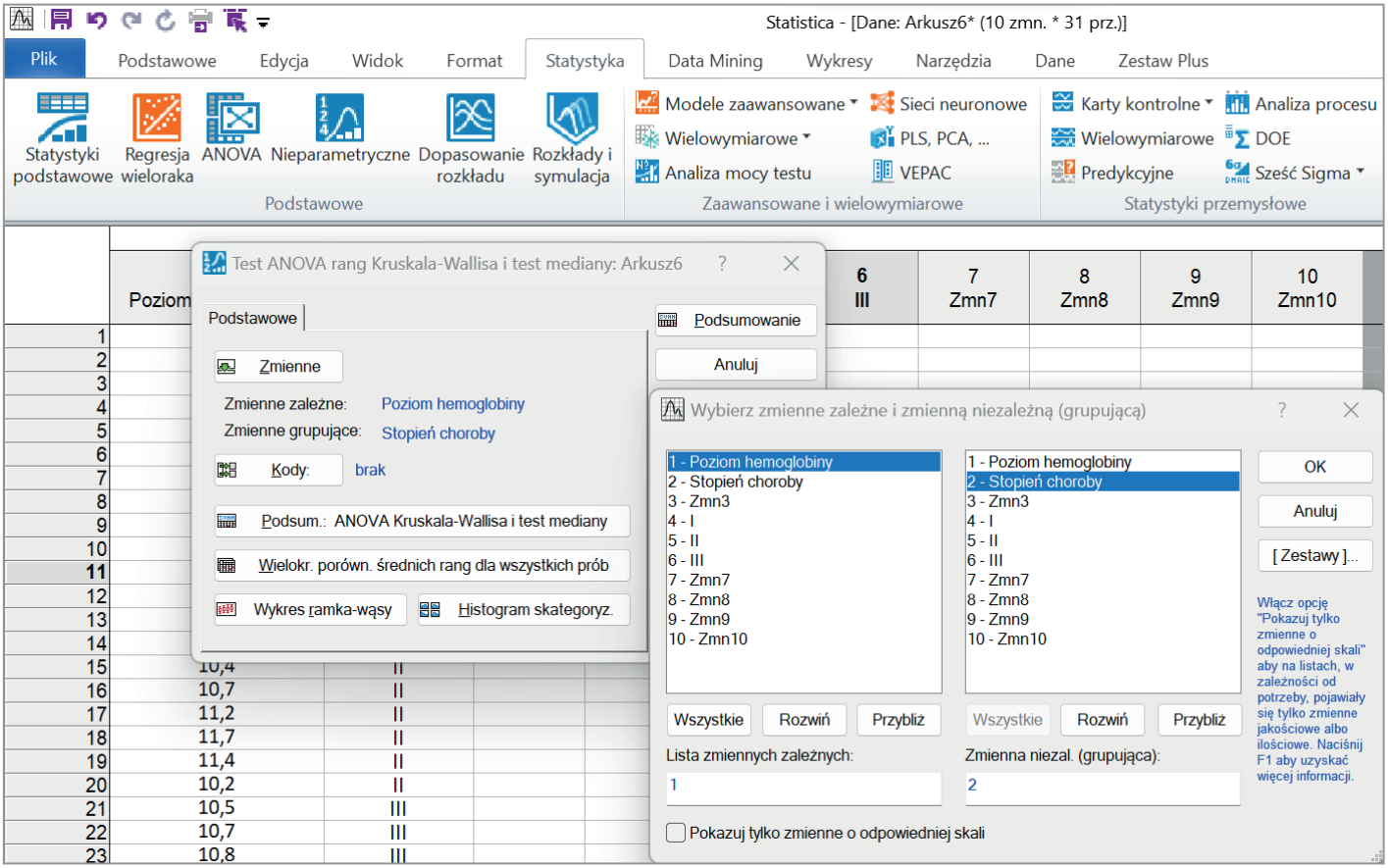
Ryc. 13.9. Wybór testów porównania wielu prób niezależnych z okna statystyk nieparametrycznych

Gdyby dane nie miały rozkładu normalnego, można by je przeanalizować odpowiednikiem nieparametrycznego testu analizy wariancji ANOVA Kruskala–Wallisa. Wówczas w karcie Statystyka należy wybrać Nieparametryczne, a następnie kliknąć Porównanie wielu prób niezależnych (➥ **ryc. 13.9**).



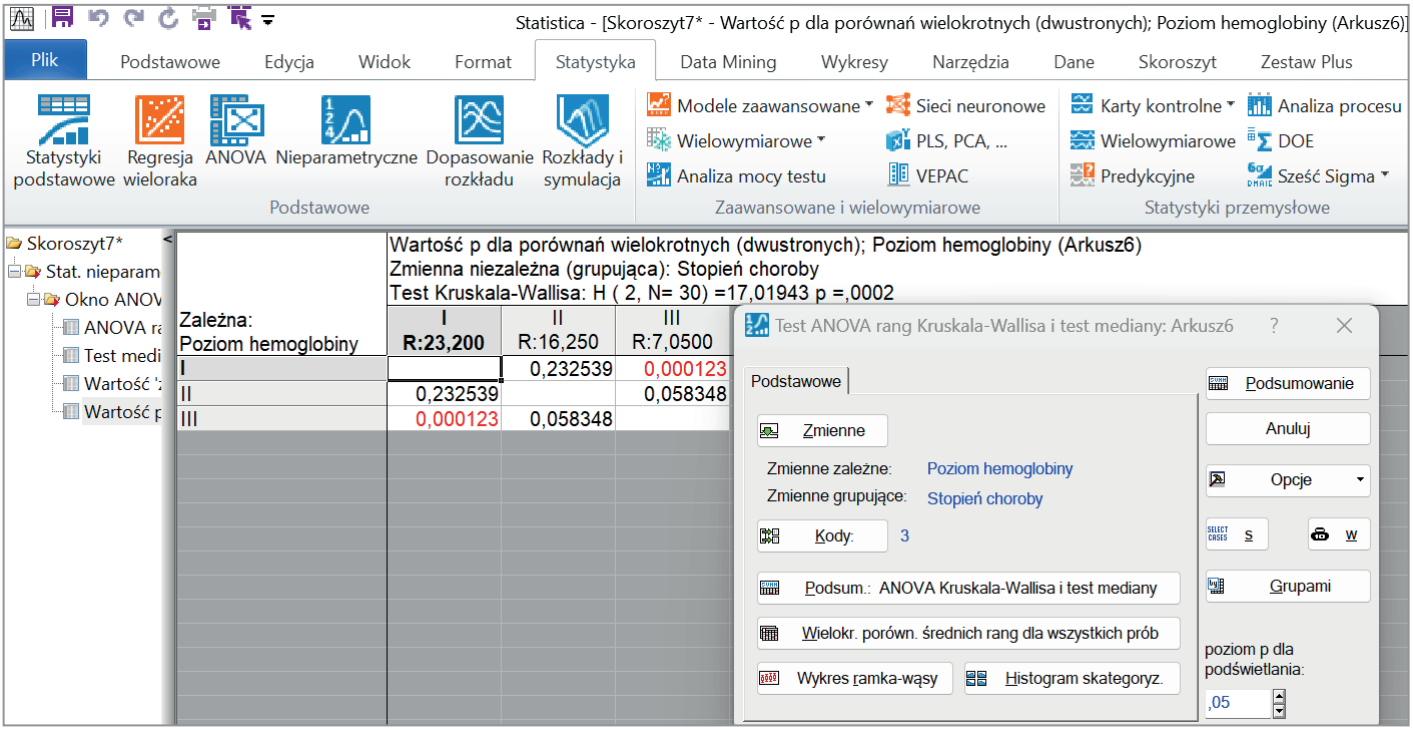
Ryc. 13.10. Wybór ANOVA Kruskala–Wallisa i testu mediany z menu

Dobieramy zmienne w podobny sposób jak w przypadku analizy jednoczynnikowej. W 1. kolumnie badana zmienna, a w 2. zmienna grupująca. Następnie wybieramy analizę ANOVA Kruskala–Wallisa i test mediany (➥ **ryc. 13.10**).



ANOVA Kruskala–Wallisa jest rozwinięciem testu *U* Manna–Whitneya. Przy jego użyciu można porównywać również grupy o różnej liczebności. Test nie wymaga rozkładu normalnego danych. Po jego zastosowaniu uzyskuje się przedstawione na ➥ **ryc. 13.11** wyniki dla testu *H* opartego na sumie rang i średnich rangach przy $p = 0{,}0002$; stąd można wyciągnąć wniosek o istnieniu różnic pomiędzy poszczególnymi grupami doświadczalnymi. Dodatkowo do porównania grup można użyć testu mediany, który opiera się na oczekiwaniu, że w próbach pobranych z populacji powyżej mediany będzie się znajdowało tyle samo pomiarów, ile poniżej. W efekcie konstruowana jest tabela wielodzielcza 2 × 3, a następnie na podstawie testu χ^2 jest określana różnica w poszczególnych stadiach białaczki. Test ten ma na celu stwierdzenie, czy przy braku normalności rozkładu mediany pochodzą z tej samej populacji. Jest on odporny na wartości odstające w badaniach i wtedy znajduje główne zastosowanie, jednak charakteryzuje się niską mocą w porównaniu z typową ANOVA Kruskala–Wallisa.

Ryc. 13.11. Wyniki testu nieparametrycznego związanego z ANOVA Kruskala–Wallisa



Pomiary wielokrotne (a precyzując – test Dunna z poprawką Bonferroniego) pozwalają na stwierdzenie istotnej różnicy pomiędzy grupą I i III (➥ **ryc. 13.11**). W stadium II spadek stężenia hemoglobiny nie jest tak znaczący w porównaniu ze stadium I, co może wskazywać istotny wzrost niedokrwistości pacjentów z tą chorobą w III stadium chorobowym ($p < 0{,}05$).

13.2.2. Ćwiczenie 1

Proszę sprawdzić, czy istnieje związek pomiędzy poziomem ekspozycji słonecznej i ilością wypitych płynów (l) przez osoby uczestniczące w doświadczeniu.

Wystawienie na ekspozycję słoneczną			
Brak	**Niskie**	**Średnie**	**Wysokie**
3,8	4,0	5,4	5,3
3,4	4,2	5,2	5,4
2,2	4,6	3,7	4,6
2,9	3,3	4,5	3,8
3,4	3,8	4,8	4,6
3,4	3,9	4,8	5,0
3,2	4,3	5,2	3,9
2,8	4,7	5,3	3,6
2,7	4,4	5,5	4,6
2,9	3,8	5,5	4,5

Tabela 13.4. Ilość wypitych płynów w zależności od ekspozycji na słońce

13.2.3. Ćwiczenie 2

Proszę sprawdzić, czy stadium białaczki ma wpływ na wartość OB (mm/h).

Stadium białaczki			
I	**II**	**III**	**IV**
23,5	26,2	56,4	20,0
26,7	22,4	34,8	49,7
19,2	21,1	27,3	44,8
24,8	20,5	39,2	27,3
18,7	44,0	22,4	37,3
21,1	37,8	35,9	13,0
16,6	45,4	37,3	33,5
16,2	14,9	39,7	44,8
21,9	31,1	12,4	46,7
	43,5	37,3	42,9
	18,6	35,4	
		18,6	
		39,2	

Tabela 13.5. Wskaźniki OB pacjentów w poszczególnych stadiach choroby

13.2.4. Ćwiczenie 3

Laboratorium farmakologiczne bada wpływ 4 leków na koncentrację NO w plazmie szczurów ($n = 12$). Tabela przedstawia dane na temat koncentracji NO (µmol/l) u osobników w poszczególnych grupach doświadczalnych. Proszę sprawdzić, czy różnice są istotne statystycznie i na podstawie średnich lub rang ocenić, który lek jest najbardziej efektywny (➡ **tabela 13.6**)

Placebo	Lek 1	Lek 2	Lek 3	Lek 4
51	64	78	82	20
73	50	54	14	36
99	84	90	69	46
112	70	34	53	49
79	38	40	87	30
84	46	55	75	67
58	91	90	42	40
139	67	103	83	84
127	76	85	8	101
110	56	44	103	56
89	61	25	68	70
82	20	51	49	14

Tabela 13.6. Dane dotyczące badania na temat wpływu leków na koncentrację NO w plazmie szczurów

Tomasz Hikawczuk

14 Dwuczynnikowa analiza wariancji

Cel ćwiczenia. Określenie różnic statystycznie istotnych pomiędzy grupami doświadczalnymi przy zastosowaniu testu dwuczynnikowej analizy wariancji

Analiza wariancji dla układów czynnikowych pozwala na porównanie ze sobą 2 lub więcej czynników doświadczalnych. Podobnie jak w przypadku analizy jednoczynnikowej istnieje możliwość przeprowadzenia obliczeń na nierównej liczebności podgrup każdego z czynników doświadczalnych, aczkolwiek wpływa to na efektywność całego układu blokowego. W przypadku dwuczynnikowej analizy wariancji oprócz działania poszczególnych czynników doświadczalnych należy uwzględnić również interakcję, czyli obopólny wpływ jednego czynnika doświadczalnego na wartość zmiennej objaśnianej w zależności od wartości drugiego czynnika doświadczalnego. W doświadczalnictwie zapis analizy wariancji przedstawia się w formie liczby zastosowanych wariantów w przypadku danego czynnika: 2 × 3 oznacza 2 warianty 1. czynnika, np. płeć, i 3 warianty 2. czynnika, czyli np. dawki leku. Zapis 4 × 3 oznacza 4 warianty 1. czynnika i 3 warianty 2. czynnika. W przypadku wystąpienia interakcji jej efekt można przedstawić za pomocą wykresu. Można również porównywać średnie pomiędzy podgrupami, stosując testy *post hoc*, np. test HSD Tukeya. W analizie wariancji duże znaczenie ma wcześniejsze zaplanowanie układu doświadczalnego uwzględniające liczbę wariantów w obrębie czynnika doświadczalnego i uwzględnienie mocy testu ze względu na standaryzowaną miarę efektu (RMSSE) oraz liczbę replikacji w poszczególnych podgrupach. O wiele łatwiej jest wstępnie zaplanować dany układ doświadczalny w postaci badania prospektywnego na podstawie badań retrospektywnych lub danych zebranych w przychodni, klinice czy szpitalu, niż wyszukiwać układ dwuczynnikowy w danych retrospektywnych. Stąd też test dwuczynnikowej analizy wariancji znajduje zastosowanie w typowym doświadczalnictwie i porównywaniu efektów terapii, działania leków, czy też w doświadczeniach ze zwierzętami.

14.1. Przykład

W doświadczeniu chciano sprawdzić, czy istnieje różnica w czasie ustąpienia bólu u mężczyzn i kobiet. W tym celu losowo przydzielono 30 mężczyzn i 30 kobiet do 3 wariantów terapii (A, B, C, po 10 mężczyzn i 10 kobiet w każdym). Poszczególne osoby nie wiedziały, do jakiego rodzaju terapii zostały przydzielone, otrzymały jednak instrukcje, aby w razie bólu stawów przyjąć lek i zapisywać czas po jego ustąpieniu. Proszę sprawdzić, czy występuje istotna różnica w czasie ustąpienia bólu w przypadku płci i rodzaju terapii.

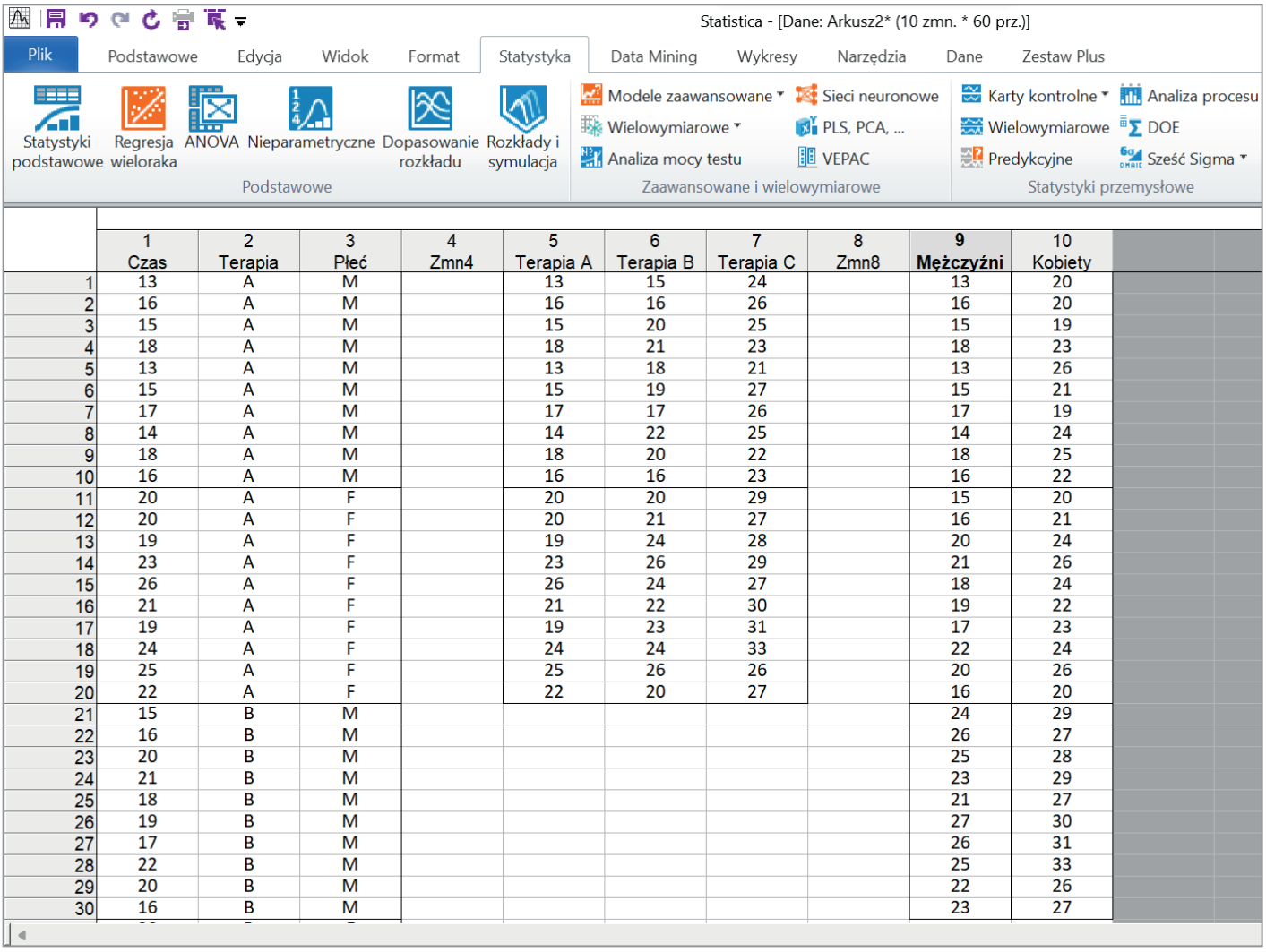
Terapia A		Terapia B		Terapia C	
mężczyźni	**kobiety**	**mężczyźni**	**kobiety**	**mężczyźni**	**kobiety**
13	20	15	20	24	29
16	20	16	21	26	27
15	19	20	24	25	28
18	23	21	26	23	29
13	26	18	24	21	27
15	21	19	22	27	30
17	19	17	23	26	31
14	24	22	24	25	33
18	25	20	26	22	26
16	22	16	20	23	27

Tabela 14.1. Czas ustąpienia bólu (w min) w zależności od terapii. Dane do ćwiczenia

Hipotezy zerowe w teście ANOVA zakładają brak różnic w wartościach średnich pomiędzy grupami doświadczalnymi oraz brak interakcji pomiędzy czynnikami doświadczalnymi.

Pierwszym krokiem jest przygotowanie danych do obliczeń. Dla ułatwienia warto tego typu schemat blokowy przygotować w Excelu, a następnie przenieść do programu statystycznego. W 3 kolumnach wprowadzone są: czas jako zmienna zależna oraz rodzaj terapii i płeć jako zmienne niezależne (➥ **ryc. 14.1**). Dodatkowo w celu sprawdzenia normalności rozkładu podzielono dane na poszczególne czynniki doświadczalne: terapie A, B, C oraz płeć. Stąd też trzeba uzyskać jedenaście rozkładów normalnych, aby móc przejść do testu dwuczynnikowej analizy wariancji.

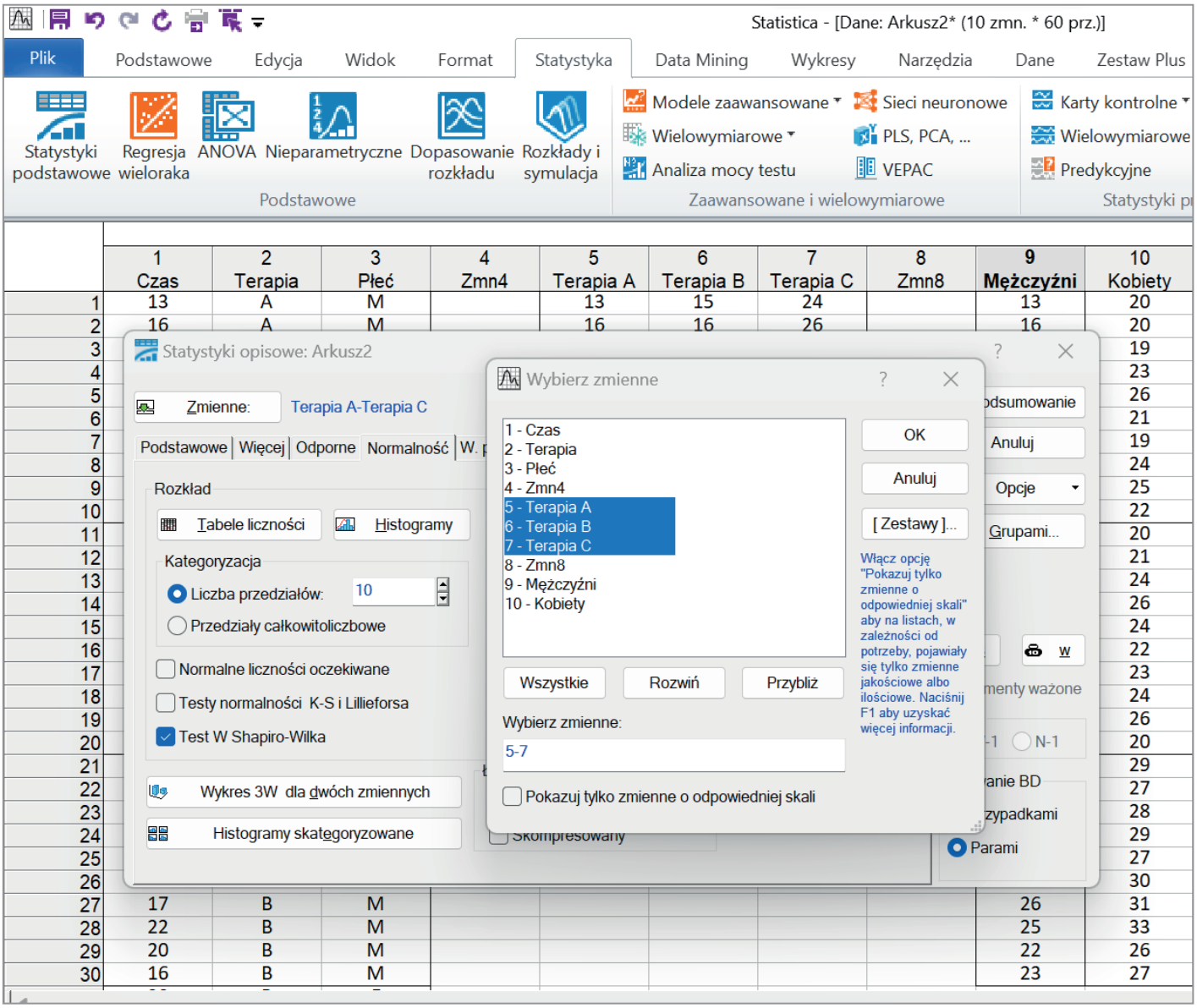
Ryc. 14.1. Dane do obliczeń statystycznych



	1 Czas	2 Terapia	3 Płeć	4 Zmn4	5 Terapia A	6 Terapia B	7 Terapia C	8 Zmn8	9 Mężczyźni	10 Kobiety
1	13	A	M		13	15	24		13	20
2	16	A	M		16	16	26		16	20
3	15	A	M		15	20	25		15	19
4	18	A	M		18	21	23		18	23
5	13	A	M		13	18	21		13	26
6	15	A	M		15	19	27		15	21
7	17	A	M		17	17	26		17	19
8	14	A	M		14	22	25		14	24
9	18	A	M		18	20	22		18	25
10	16	A	M		16	16	23		16	22
11	20	A	F		20	20	29		15	20
12	20	A	F		20	21	27		16	21
13	19	A	F		19	24	28		20	24
14	23	A	F		23	26	29		21	26
15	26	A	F		26	24	27		18	24
16	21	A	F		21	22	30		19	22
17	19	A	F		19	23	31		17	23
18	24	A	F		24	24	33		22	24
19	25	A	F		25	26	26		20	26
20	22	A	F		22	20	27		16	20
21	15	B	M						24	29
22	16	B	M						26	27
23	20	B	M						25	28
24	21	B	M						23	29
25	18	B	M						21	27
26	19	B	M						27	30
27	17	B	M						26	31
28	22	B	M						25	33
29	20	B	M						22	26
30	16	B	M						23	27

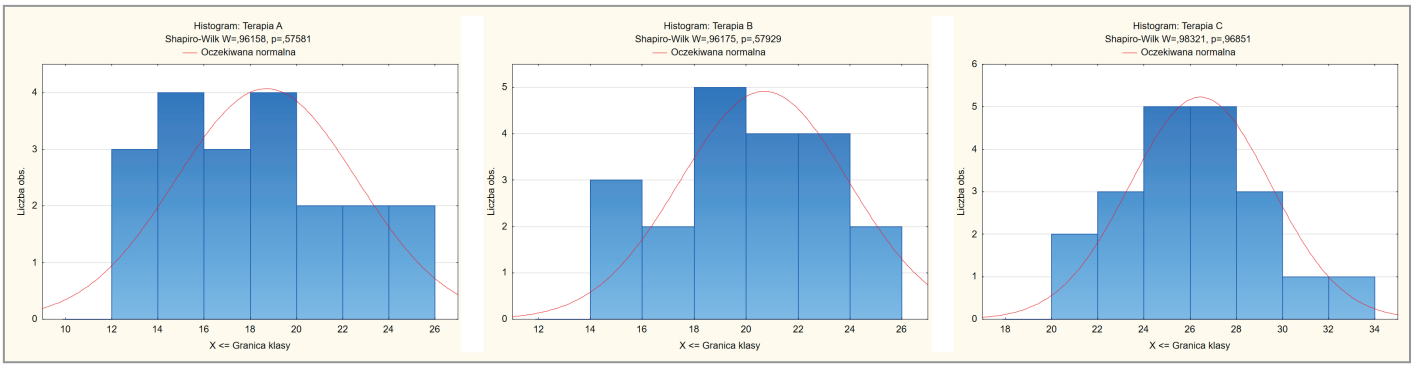
Następnie można przejść do sprawdzenia normalności rozkładów poszczególnych grup doświadczalnych w znany już sposób. Z karty Statystyka wybieramy Statystyki podstawowe, w pojawiającym się oknie klikamy Statystyki opisowe i zatwierdzamy kliknięciem OK. Otrzymujemy wówczas okno z ➥ **ryc. 14.2** umożliwiające wybór zmiennych, a następnie po kliknięciu zakładki Normalność można wybrać test Shapiro–Wilka odpowiedni dla zbiorów danych do 50 replikacji w grupie doświadczalnej. Następnie klikamy Histogram w celu weryfikacji hipotezy zerowej o braku różnic pomiędzy danym rozkładem a rozkładem normalnym w przypadku poszczególnych rodzajów terapii.

Ryc. 14.2. Sprawdzenie normalności rozkładów w przypadku terapii



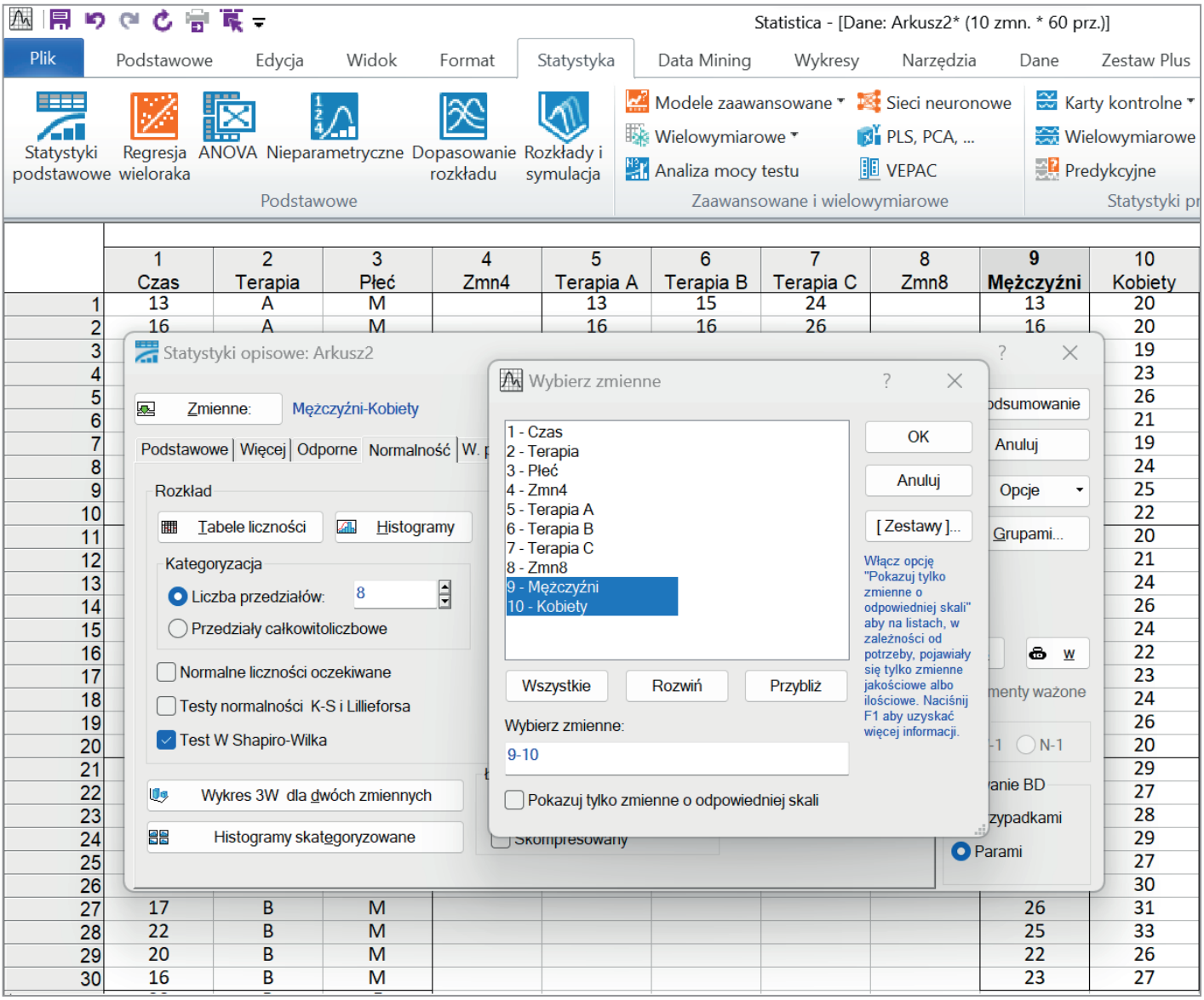
Ryc. 14.3. Sprawdzenie rozkładu normalnego w danych z poszczególnych rodzajów terapii

W wyniku przeprowadzonego testu Shapiro–Wilka w przypadku każdej z terapii uzyskano wartość p znajdującą się w przedziale ufności dla hipotezy zerowej (➥ **ryc. 14.3**).



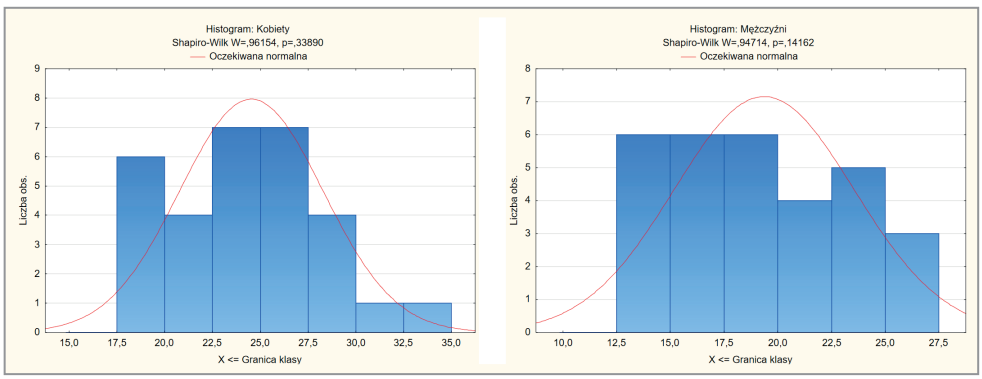
W każdym przypadku możliwość popełnienia błędu przy odrzuceniu H_0 jest większa niż 50%, więc nie możemy jej odrzucić. Tym samym potwierdzamy normalność rozkładu danych w przypadku każdej z terapii. W podobny sposób możemy przeprowadzić test Shapiro–Wilka dla poszczególnych płci (➥ **ryc. 14.4**).

Ryc. 14.4. Wybór danych do określenia normalności rozkładu danych w przypadku płci



Podobnie jak w przypadku terapii również tutaj uzyskano rozkłady normalne (➥ **ryc. 14.5**).

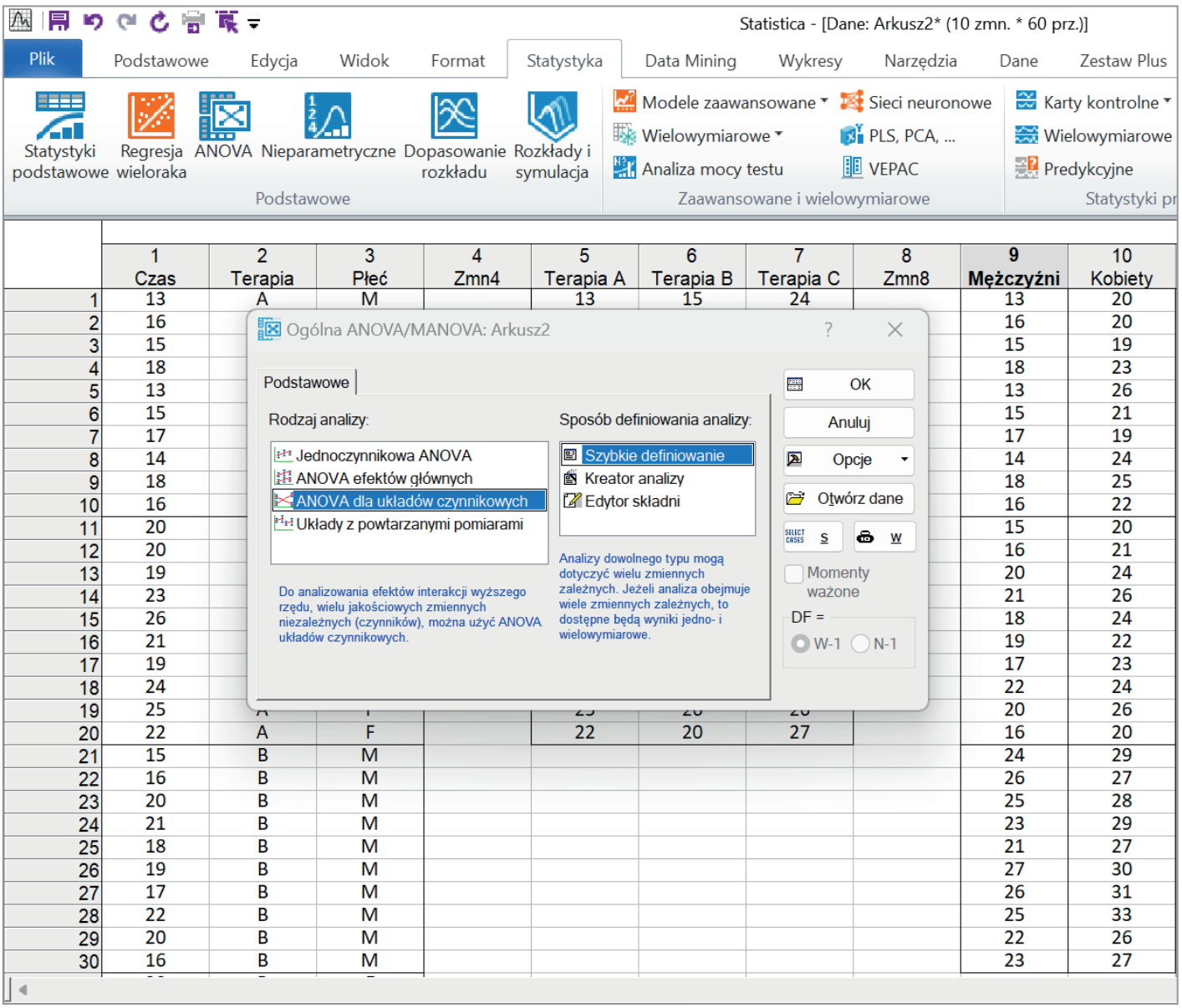
Ryc. 14.5. Wyniki testu Shapiro–Wilka dla płci



Kolejnym ważnym krokiem po potwierdzeniu normalności rozkładów jest sprawdzenie homogenności wariancji, jednak można tego dokonać dopiero na późniejszym etapie (ewentualnie w statystykach podstawowych, ale wymaga to oddzielnego testowania dla poszczególnych czynników doświadczalnych).

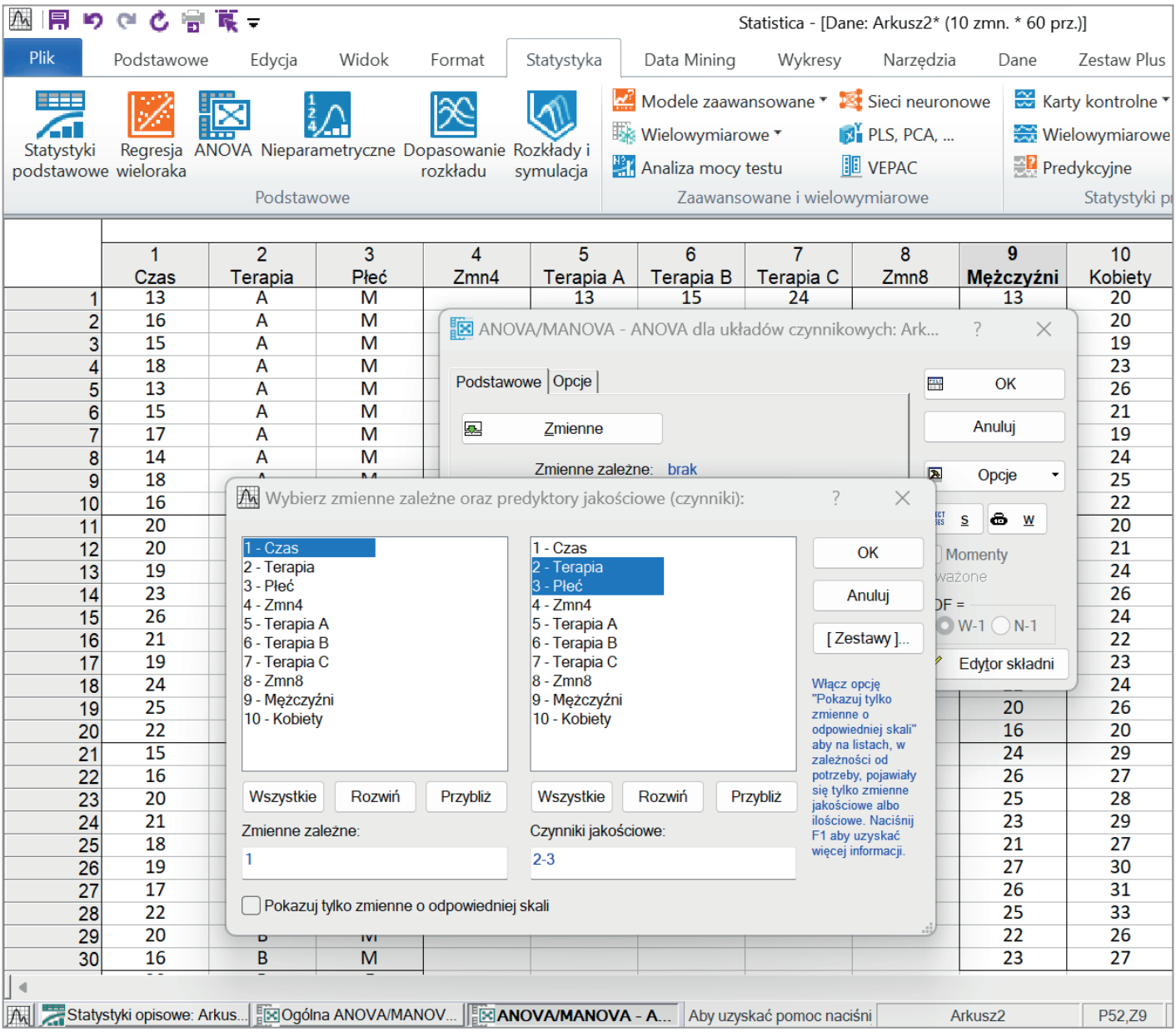
Stąd też przechodzimy do wyboru testu. W karcie Statystyka wybieramy ANOVA, co umożliwi wybranie poszczególnych rodzajów testów z ➥ **ryc. 14.6**. Dla danego doświadczenia układ jest czynnikowy, z uwzględnieniem efektu interakcji, stąd też taki test wybieramy. W tym oknie w przypadku innych danych mamy możliwość wyboru ANOVY dla efektów głównych, nieuwzględniającej efektu interakcji. Układ z powtarzanymi pomiarami jest rozwinięciem testu *t* dla zmiennych zależnych i umożliwia pomiary na tych samych osobach w kilku turach, np. z zastosowaniem kilku leków w określonej kolejności jeden po drugim. Wybieramy ANOVA dla układów czynnikowych, zatwierdzamy kliknięciem OK (➥ **ryc. 14.6**).

Ryc. 14.6. Wybór analizy wariancji

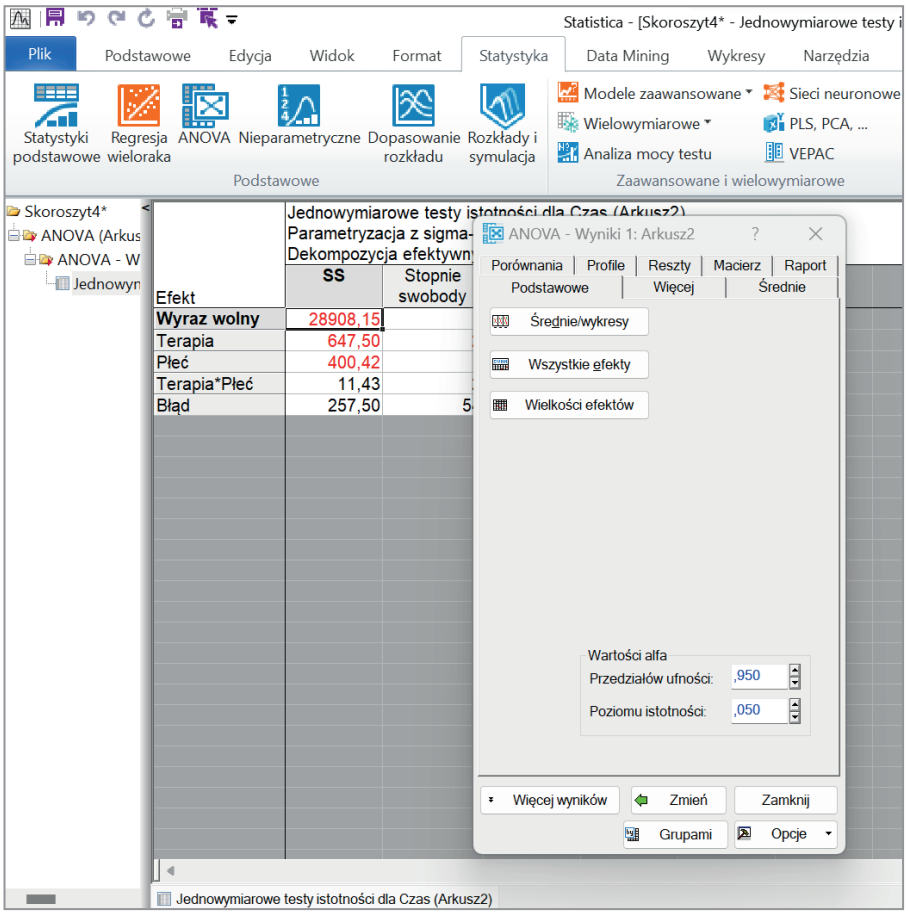


Ryc. 14.7. Wybór zmiennych do analizy dla układów czynnikowych

Pojawia się następnie okno wyboru zmiennych (➥ **ryc. 14.7**). Zmienną zależną jest czas, natomiast czynnikami jakościowymi – terapia i płeć.

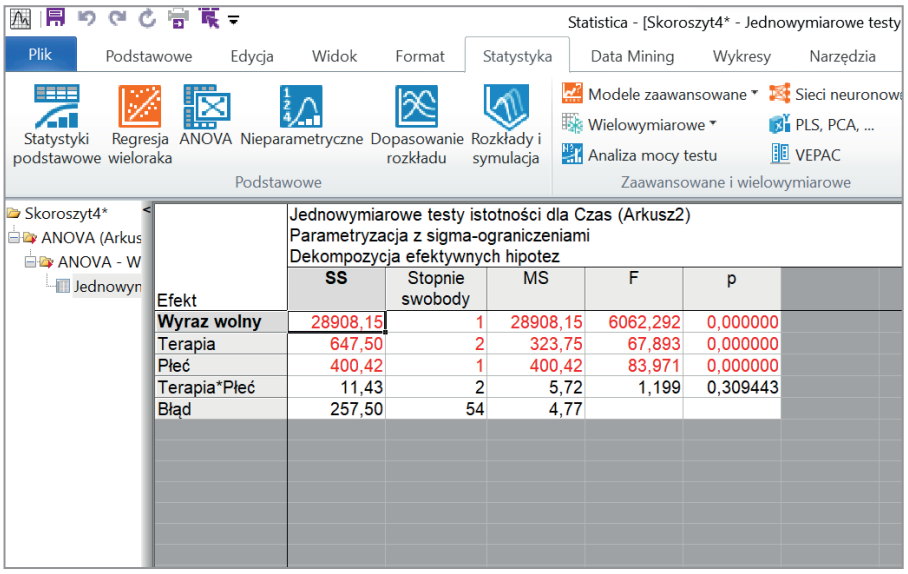


Zatwierdzamy wybór zmiennych kliknięciem OK. Następnie potwierdzamy OK po raz drugi i otrzymujemy okno, przy użyciu którego po wybraniu wszystkich efektów możemy weryfikować hipotezy zerowe (➥ **ryc. 14.8**).



Ryc. 14.8. Przejście do sekcji Więcej wyników

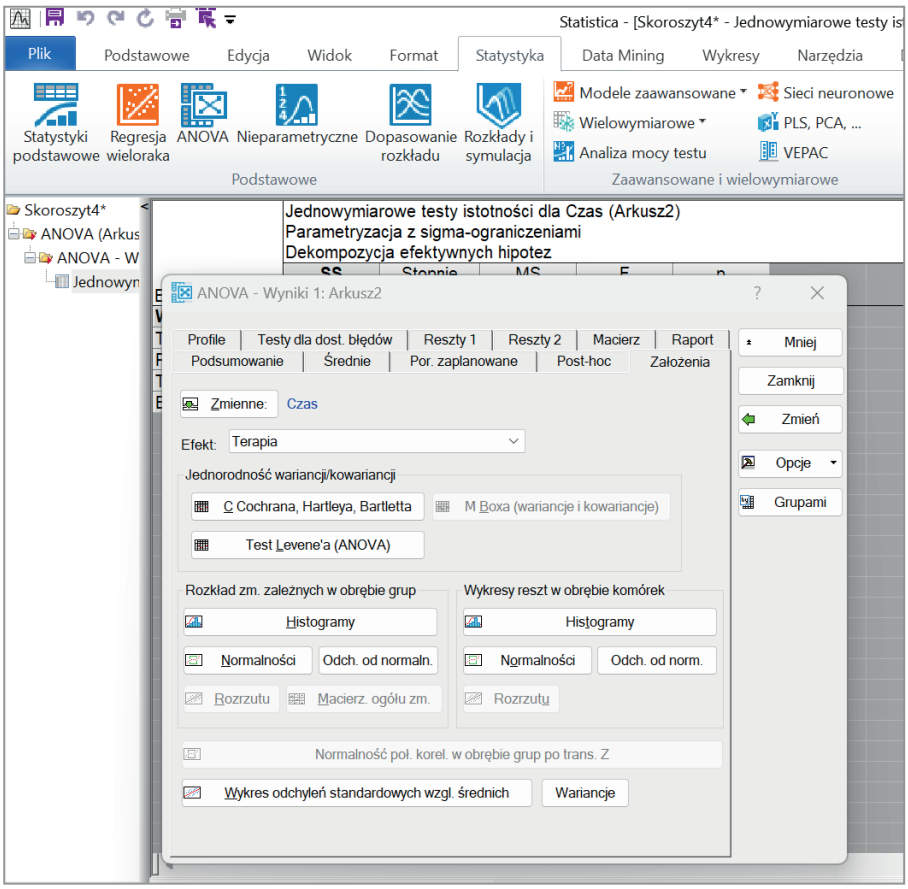
Jednowymiarowe testy istotności potwierdziły znaczący wpływ na wyniki czasu ustąpienia bólu w przypadku rodzaju terapii i płci (➥ **ryc. 14.9**).



Ryc. 14.9. Wyniki testu analizy wariancji dla czynników doświadczalnych w ramach czasu ustąpienia bólu

Nie stwierdzono interakcji pomiędzy badanymi czynnikami doświadczalnymi, natomiast istotne różnice ($p < 0{,}05$) pomiędzy wartościami średnimi w grupach stwierdzono w przypadku rodzaju terapii i płci. Dodatkowo powrót do okna analizy (na dolnym pasku) i kliknięcie przycisku Więcej wyników pozwala na przejście do zakładki Założenia, w której można sprawdzić homogenność wariancji testem Levene'a (➥ **ryc. 14.10**).

Ryc. 14.10. Wynik testu Levene'a dla określenia homogenności wariancji



Test ten przeprowadza się dla każdego efektu: terapii, płci i interakcji terapia ∗ płeć. Analizując wyniki z ➥ **ryc. 14.11**, możemy zauważyć, że H_0 zakładająca jednorodność wariancji w grupach jest potwierdzona we wszystkich 3 przypadkach: 3 wariantami terapii, 2 wariantami płci i 3 × 2 wariantami interakcji ($p > 0{,}05$).

Ryc. 14.11. Test homogenności dla grup w obrębie czynnika doświadczalnego

	Test Levene'a jednorodności wariancji (Arkusz2) Efekt: Terapia Stopnie swobody dla każdego F : 2, 57			
	MS Efekt	MS Błąd	F	p
Czas	3,816667	3,840351	0,993833	0,376477
	Test Levene'a jednorodności wariancji (Arkusz2) Efekt: Płeć Stopnie swobody dla każdego F : 1, 58			
	MS Efekt	MS Błąd	F	p
Czas	3,585185	4,152370	0,863407	0,356638
	Test Levene'a jednorodności wariancji (Arkusz2) Efekt: Terapia*Płeć Stopnie swobody dla każdego F : 5, 54			
	MS Efekt	MS Błąd	F	p
Czas	0,536667	1,185185	0,452812	0,809400

Na koniec pozostaje już tylko określenie różnic pomiędzy grupami. W tym celu, wracając do okna z ➥ **ryc. 14.10**, klikając na zakładce, można wybrać test *post hoc* HSD Tukeya, który wskaże na poziomie istotności $p < 0{,}05$, pomiędzy którymi grupami występują różnice i która terapia jest najkorzystniejsza na podstawie różnic istotnych statystycznie w wartościach średnich poszczególnych grup doświadczalnych.

Pod względem czasu ustąpienia bólu u kobiet i mężczyzn każda terapia różni się istotnie na poziomie $p < 0{,}05$. Jednak ból ustępuje najszybciej po zastosowaniu terapii A, poniżej 19 min, natomiast najmniej efektywna jest terapia C, gdzie na wygaszenie bólu potrzeba ok. 26,5 min (➥ **ryc. 14.12**). Oczywiście jeśli brakuje leku z terapii A, to można zastosować lek B lub C, aczkolwiek dodatkowy czas dla osób silniej odczuwających ból nie będzie przyjemnym odczuciem.

Ryc. 14.12. Test *post hoc* dla rodzajów terapii

Nr podkl.	Terapia	{1} 18,700	{2} 20,700	{3} 26,450
1	A		0,014900	0,000121
2	B	0,014900		0,000121
3	C	0,000121	0,000121	

Test HSD Tukeya; zmienna Czas (Arkusz2)
Przybliżone prawdopodobieństwa dla testów post hoc
Błąd: MS międzygrupowe = 4,7685, df = 54,000

Z kolei w przypadku płci wszystkie leki działają szybciej u mężczyzn niż u kobiet ($p = 0{,}000$), co przedstawiono na ➥ **ryc. 14.13**.

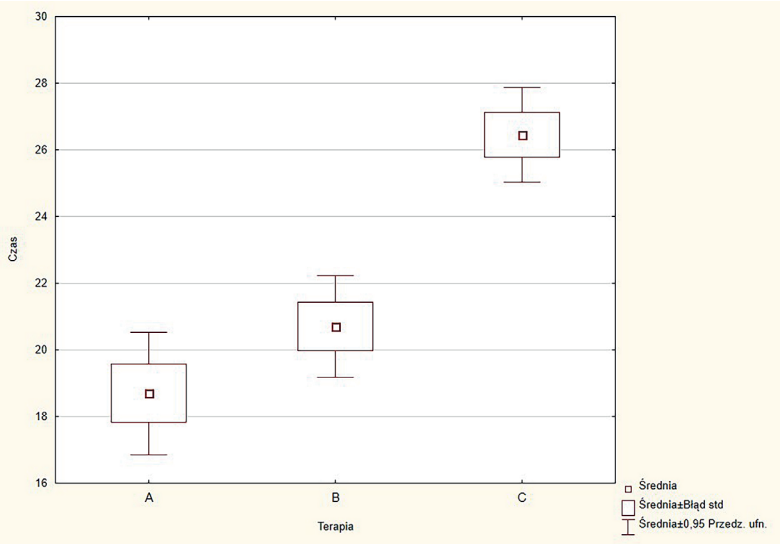
Ryc. 14.13. Test *post hoc* dla płci

Nr podkl.	Płeć	{1} 19,367	{2} 24,533
1	M		0,000113
2	F	0,000113	

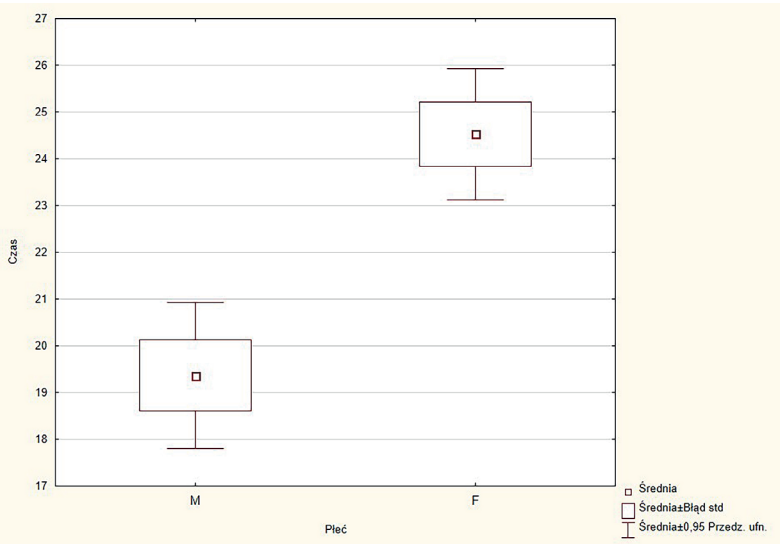
Test HSD Tukeya; zmienna Czas (Arkusz2)
Przybliżone prawdopodobieństwa dla testów post hoc
Błąd: MS międzygrupowe = 4,7685, df = 54,000

Natomiast ➥ **ryc. 14.14** i ➥ **ryc. 14.15** przedstawiają w postaci wykresów pudełkowych średni czas ustąpienia bólu w zależności odpowiednio od rodzaju terapii i płci.

Ryc. 14.14. Wykres przedstawiający średni czas ustąpienia bólu w zależności od rodzaju terapii



Ryc. 14.15. Wykres przedstawiający średni czas ustąpienia bólu w zależnosci od płci



Tomasz Hikawczuk

15 Korelacja Pearsona i regresja prostoliniowa

W tym rozdziale przedstawimy korelację Pearsona i regresję prostoliniową jako miary opisujące relację pomiędzy dwiema zmiennymi.

Cel ćwiczenia. Określenie relacji pomiędzy dwiema zmiennymi ciągłymi w postaci korelacji i przewidywanie wartości danej zmiennej za pomocą jednoczynnikowej regresji prostoliniowej.

Relacja pomiędzy dwiema zmiennymi może być określana za pomocą współczynnika korelacji Pearsona. Wskazuje on, co dzieje się z wartościami liczbowymi drugiej zmiennej (tzn. czy rosną, czy maleją, czy pozostają bez zmian), gdy wartości pierwszej zmiennej ulegają wzrostowi. Oprócz obliczenia współczynnika korelacji testuje się również hipotezę dotyczącą jego statystycznej istotności. Hipoteza zerowa zakłada, że współczynnik korelacji wynosi 0. Z kolei współczynnik r określa siłę i kierunek korelacji Pearsona.

Ważnym kryterium wyboru współczynnika korelacji (spośród korelacji Pearsona, Spearmana lub tau Kendalla) jest określenie normalności rozkładu danych. Analizę statystyczną zaczynamy od zastosowania testu Shapiro–Wilka (gdy $n < 50$) weryfikującego normalność rozkładu obu korelowanych zbiorów danych. W przypadku potwierdzenia normalności siłę związku między zmiennymi powinno się określać za pomocą współczynnika korelacji Pearsona (r). Jest to test parametryczny. Gdy założenia normalności nie są spełnione, niezbędne będzie przypisanie rang wartościom zmiennych i przeprowadzenie nieparametrycznej analizy korelacji Spearmana lub korelacji tau Kendalla. Tę ostatnią stosujemy, jeśli wystąpią tzw. rangi powiązane, czyli jednakowe. Przy większej liczbie obserwacji ($n > 50$) do sprawdzania normalności rozkładu stosowane są inne testy: Kołmogorowa–Smirnowa lub Andersona–Darlinga.

W przypadku danych parametrycznych wygodnym i przejrzystym rozwiązaniem jest sporządzenie wykresu rozrzutu, który zobrazuje relację z naniesionym na nim równaniem regresji prostej. Pozwoli to na:

- usunięcie pomiarów odstających,
- sprawdzenia założenia zależności prostoliniowej (kluczowej podczas zastosowania korelacji Pearsona).

Określenie współczynnika korelacji Pearsona polega na porównaniu zależności i siły relacji pomiędzy 2 zmiennymi. Jest on wyznaczany na podstawie wzoru:

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{x})\cdot(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}}$$

gdzie:
n – liczba par obserwacji,
x – wartość 1. zmiennej,
y – wartość 2. zmiennej,
$\bar{x}$ – wartość średnia 1. zmiennej,
$\bar{y}$ – wartość średnia 2. zmiennej.

Współczynnik korelacji Pearsona przyjmuje wartość w zakresie od –1 do +1 i określa siłę związku (korelacji) pomiędzy dwiema zmiennymi (➡ **tabela 15.1**). Znak korelacji pokazuje tendencję rosnącą lub malejącą jednej z korelowanych zmiennych w odniesieniu do drugiej zmiennej.

Zakres współczynnika *r*	Siła korelacji
0,8–1,0	bardzo silna
0,6–0,79	silna
0,4–0,59	średnia
0,2–0,39	słaba
0–0,19	brak i bardzo słaba

Tabela 15.1. Siła korelacji w zależności od wartości współczynnika *r*[1]

Współczynnik korelacji *r* określa liniową zależność pomiędzy dwiema zmiennymi. Wartość ta podniesiona do potęgi drugiej wyznacza współczynnik determinacji (r^2). Opisuje on proporcję wspólnej zmienności dwóch zmiennych oraz określa, w jakim stopniu część zmienności *Y* jest wyjaśniona zmiennością *Y*. Współczynnik determinacji pozwala także na wyjaśnienie prostoliniowych relacji pomiędzy zmiennymi przy zastosowaniu równania regresji prostoliniowej.

Drugą możliwością określenia relacji pomiędzy zmiennymi jest współczynnik regresji, uwzględniany również w najprostszej formie równania regresji prostoliniowej na podstawie próby. Równanie to określa, o ile zmienia się wartość drugiej zmiennej, gdy wartości pierwszej zmiennej zmieniają się o jedną jednostkę.

Równanie regresji prostoliniowej przyjmuje następującą postać:

$$y = ax + b$$

gdzie:
y – zmienna objaśniana,
x – zmienna objaśniająca,
a – współczynnik kierunkowy prostej,
b – współczynnik przesunięcia prostej.

Celem regresji prostoliniowej niezbędne jest stwierdzenie występowania zależności prostoliniowej pomiędzy zmiennymi, co wymaga uwzględnienia w analizie wartości współczynnika kierunkowego (a) i wyrazu wolnego (b) oraz sprawdzenia poniższych założeń regresji prostoliniowej:

- związek pomiędzy *X* i *Y* jest związkiem liniowym,
- składniki błędów losowych mają rozkład normalny.

Hipoteza zerowa w przypadku równania regresji zakłada jej brak, co oznacza wartość równą lub bliską 0 w przypadku współczynnika kierunkowego prostej.

Wraz ze wzrostem wartości współczynnika korelacji Pearsona wzrostowi ulega również wartość współczynnika determinacji, co umożliwia dokładniejszą predykcję zmiennej *Y* na podstawie równania regresji prostoliniowej w badanej populacji ze znanymi wartościami zmiennej *X*.

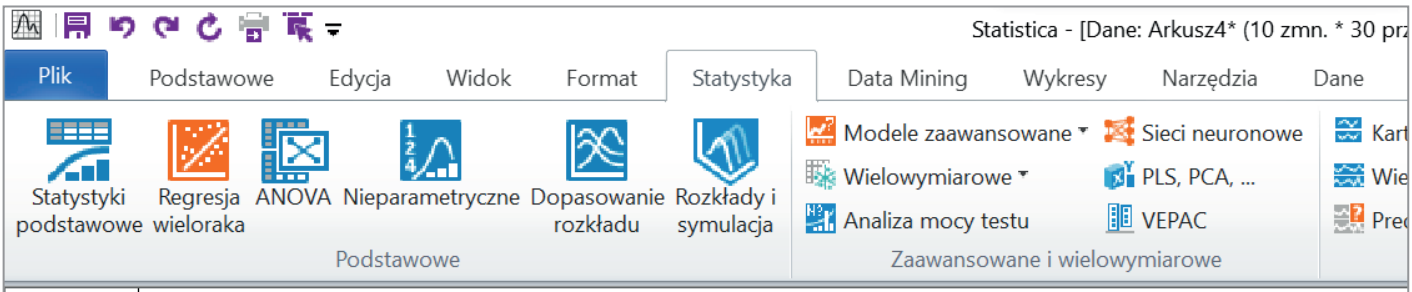
W rezultacie zastosowania regresji liniowej uzyskujemy równanie regresji utworzone metodą najmniejszych kwadratów, obrazowane za pomocą linii o takim nachyleniu, przy którym suma kwadratów odległości punktów od linii jest najmniejsza. Dokładność przewidywania wartości zmiennej zależnej (*y*) na podstawie znajomości zmiennej *x* jest tym wyższa, im wartość współczynnika determinacji r^2 jest bliższa 1.

15.1. Przykład

Proszę przeanalizować, czy istnieje istotna statystycznie korelacja pomiędzy odczynem Biernackiego (OB) w mm, a stężeniem białka C-reaktywnego (ang. *C-reactive proteine* – CRP) w mg/l w próbkach krwi 25 pacjentów. Na podstawie danych z próby utwórz prosty model predykcyjny w postaci równania regresji liniowej i spróbuj przewidzieć wartości dla białka CRP w przypadku OB wynoszącego 53, 25 i 61 mm.

Na ➥ **ryc. 15.1** widać dane uzyskane w doświadczeniu z lekami zebrane w jedną grupę. Aby wyznaczyć współczynnik korelacji i w następnym kroku równanie regresji, niezbędne jest uwzględnienie rodzaju rozkładu dystrybuanty danych. Sprawdzenie jej normalności odbywa się za pomocą testu Shapiro–Wilka. Możemy je przeprowadzić, wybierając z karty Statystyka Statystyki podstawowe, a następnie Statystyki opisowe.

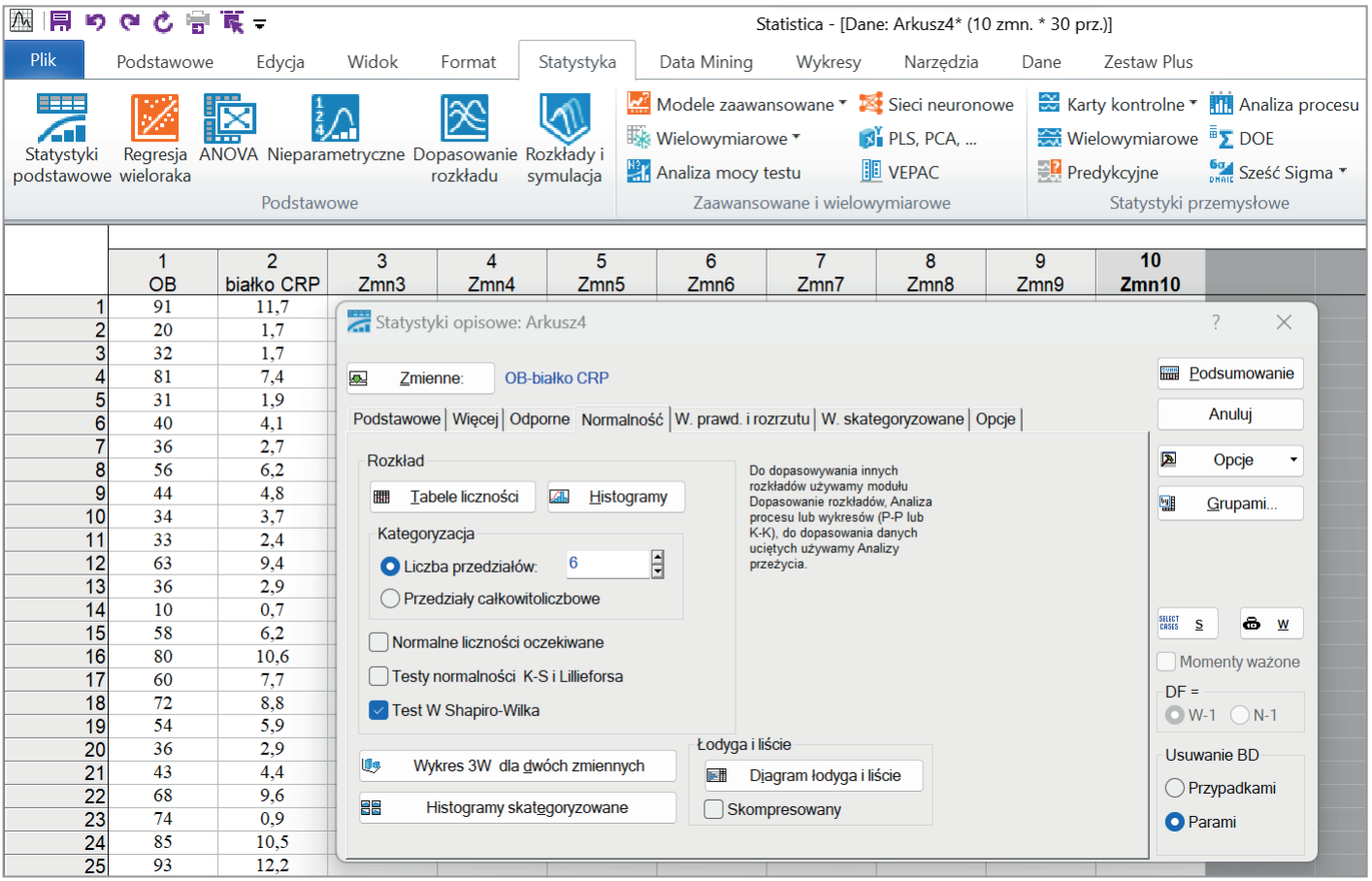
Ryc. 15.1. Dane pacjentów przychodni przedstawiające OB oraz stężenie białka CRP



	1 OB	2 białko CRP	3 Zmn3	4 Zmn4	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9
1	91	11,7							
2	20	1,7							
3	32	1,7							
4	81	7,4							
5	31	1,9							
6	40	4,1							
7	36	2,7							
8	56	6,2							
9	44	4,8							
10	34	3,7							
11	33	2,4							
12	63	9,4							
13	36	2,9							
14	10	0,7							
15	58	6,2							
16	80	10,6							
17	60	7,7							
18	72	8,8							
19	54	5,9							
20	36	2,9							
21	43	4,4							
22	68	9,6							
23	74	0,9							
24	85	10,5							
25	93	12,2							

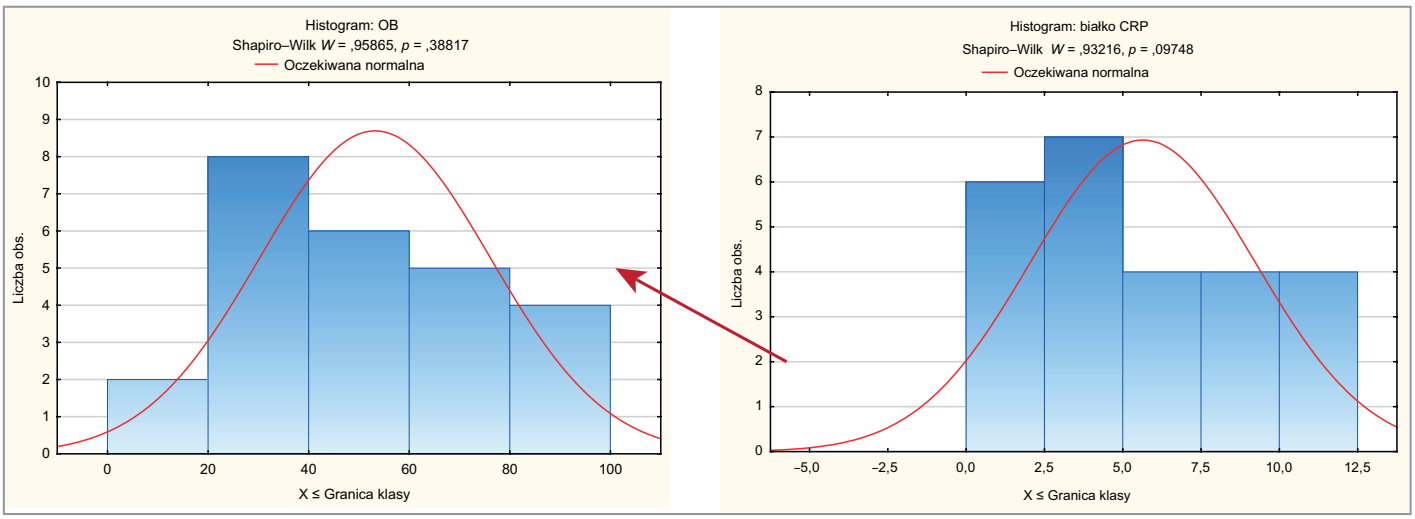
Ryc. 15.2. Wybór zmiennych i wartości testu Shapiro–Wilka naniesionych na histogramy

Po wybraniu zmiennych i testu normalności w zakładce Normalność można wybrać histogramy (➥ **ryc. 15.2**).



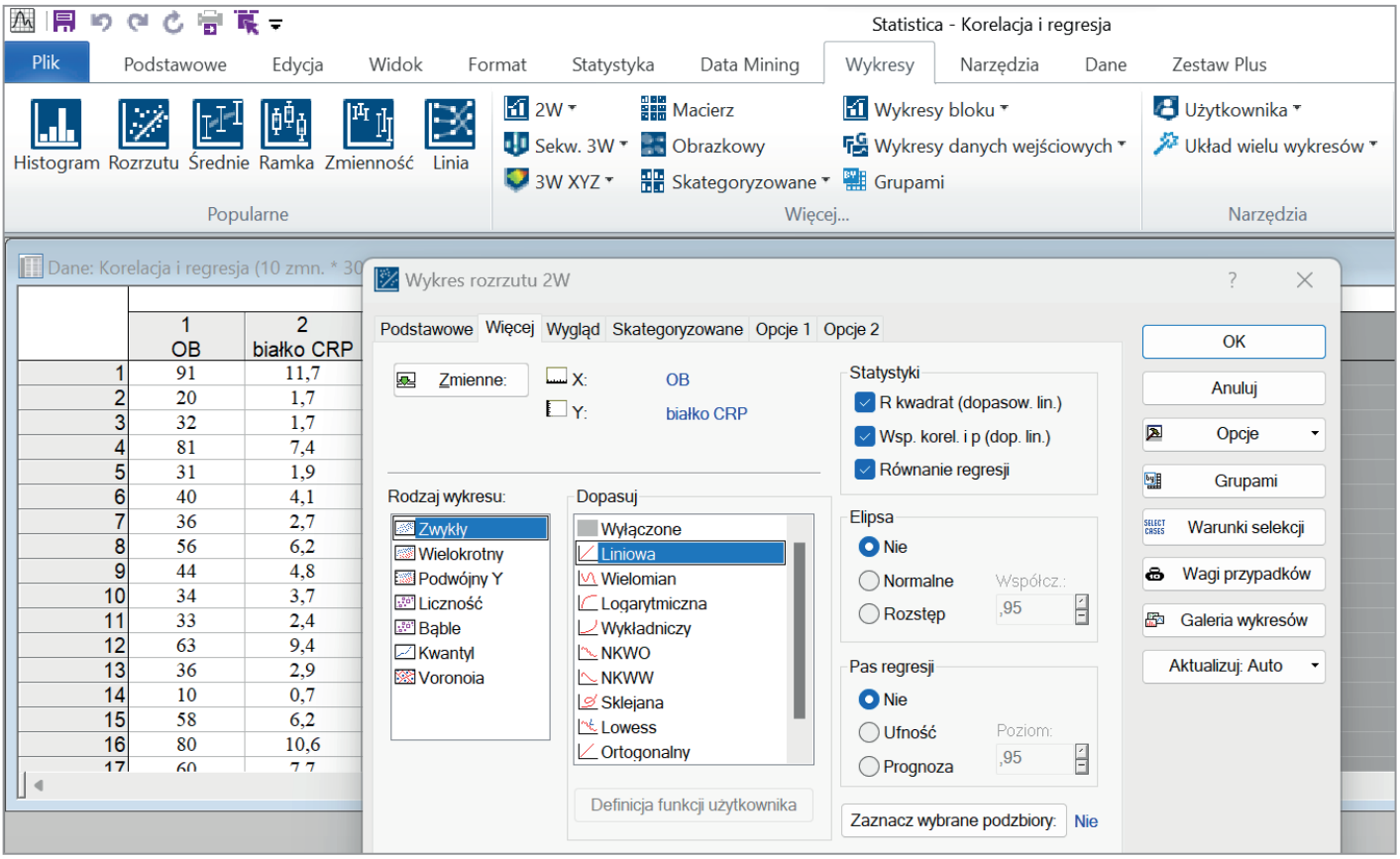
Ryc. 15.3. Porównanie dystrybuant danych z krzywą normalności

W ten sposób uzyskuje się informacje o dystrybuantach danych z ➥ **ryc. 15.3**.



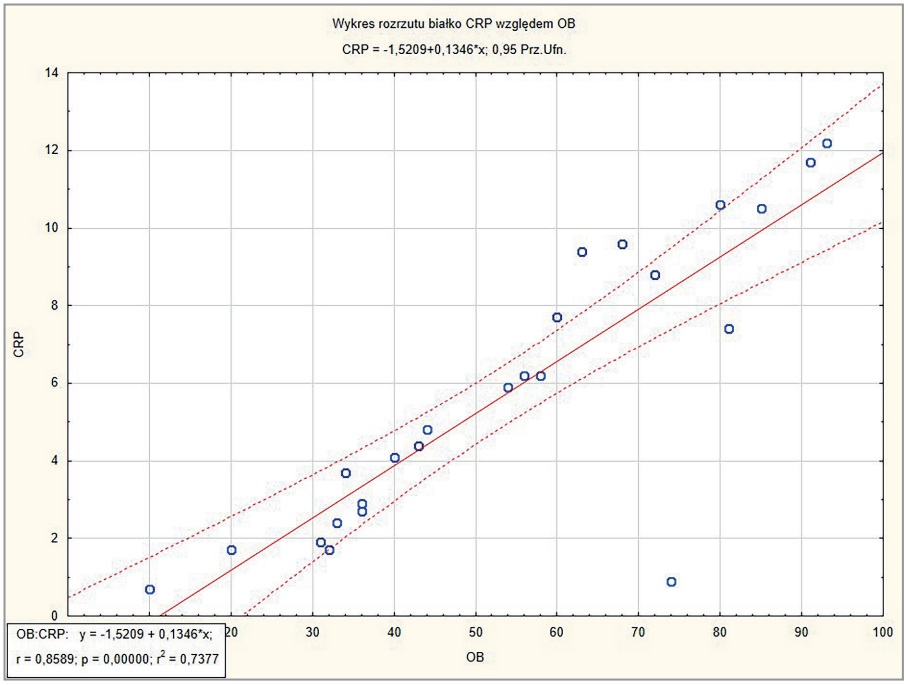
Na podstawie uzyskanych histogramów można stwierdzić, że obydwie wartości p (0,388 i 0,097) nie pozwalają nam na odrzucenie H_0 mówiącej o braku różnicy między uzyskanym rozkładem empirycznym danych a rozkładem normalnym (➥ **ryc. 15.3**). Zatem dzięki temu, że rozkład obu zmiennych nie różni się od rozkładu normalnego, związek między nimi można określić za pomocą współczynnika korelacji Pearsona (test parametryczny). Wykres rozrzutu sugeruje związek prostoliniowy, więc użycie tego współczynnika jest w pełni zasadne. Istnieje możliwość bezpośredniego utworzenia równania regresji prostoliniowej poprzez wybranie z karty Statystyka polecenia Statystyki opisowe i następnie macierzy korelacji. Jednak podobny efekt można uzyskać, wybierając z karty Wykresy polecenie Rozrzut (➥ **ryc. 15.4**). W zakładce Więcej okna Wykres rozrzutu 2W wybieramy Zmienne, w tym przypadku OB i białko CRP, a następnie w polu Statystyki (prawy górny róg) zaznaczamy do obliczenia R kwadrat, Współczynnik korelacji i równanie regresji.

Ryc. 15.4. Wybór zmiennych i statystyk dla wykresu rozrzutu danych



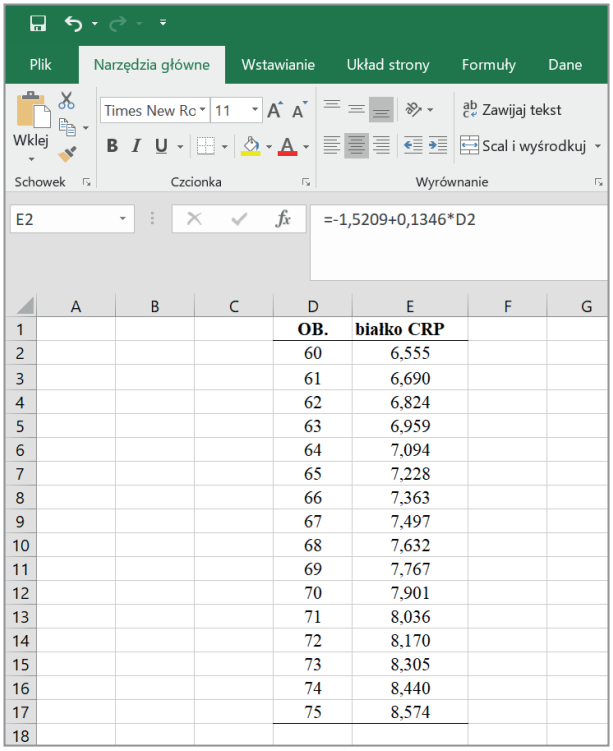
Otrzymujemy wykres rozrzutu z naniesioną prostą oraz wartościami opisującymi korelację i dopasowanie uzyskanego modelu regresji prostej do danych w próbie. Jednak o ile korelację możemy obliczać bez konieczności rozważenia, która zmienna jest objaśniająca, a która objaśniana, o tyle w przypadku regresji prostoliniowej dobór zmiennej objaśniającej i objaśnianej musi wynikać z natury badanego zjawiska (należy wcześniej zapoznać się z teorią i przygotować plan doświadczenia), a więc zmienne nie mogą być dobierane losowo.

W teście korelacji Pearsona przedstawionym na wykresie (➡ **ryc. 15.5**) uzyskano współczynnik korelacji $r = 0{,}859$, co wskazuje na silną pozytywną korelację pomiędzy 2 zmiennymi. Czyli wraz ze wzrostem OB zwiększa się również stężenie białka C-reaktywnego. Wysoki współczynnik r wpływa ponadto na występowanie statystycznie potwierdzonej istotności relacji pomiędzy 2 zmiennymi ($p = 0{,}000$) i dopasowania modelu regresji prostej do danych zebranych w próbie ($r^2 = 0{,}74$). W konsekwencji możemy w populacji, z której pobrano daną próbę, dokonać predykcji wartości białka C-reaktywnego na podstawie znajomości OB.



Ryc. 15.5. Wykres rozrzutu uwzględniający korelację i równanie regresji dla OB i białka CRP

W tym celu stosujemy utworzone równanie regresji $y = -1{,}5209 + 0{,}1346 \times x$ i pod zmienną x podstawiamy dowolną wartość rzeczywistą z zakresu OB. Można też, stosując formułę dla serii wartości w programie Excel, dokonać predykcji danych z określonego zakresu lub odczytywać bezpośrednio na wykresie (➡ **ryc. 15.6**), oczywiście uwzględniając wcześniej przedział ufności dla współczynnika kierunkowego a i procent wyjaśnienia zmienności zmiennej zależnej.



	A	B	C	D	E	F	G
1				OB.	białko CRP		
2				60	6,555		
3				61	6,690		
4				62	6,824		
5				63	6,959		
6				64	7,094		
7				65	7,228		
8				66	7,363		
9				67	7,497		
10				68	7,632		
11				69	7,767		
12				70	7,901		
13				71	8,036		
14				72	8,170		
15				73	8,305		
16				74	8,440		
17				75	8,574		
18							

Ryc. 15.6. Zastosowanie równania regresji prostej i predykcja zawartości białka C-reaktywnego w zakresie wartości dla OB 60–75

15.2. Ćwiczenie 1

Proszę określić, na podstawie danych z ➥ **tabeli 15.2**, czy występuje istotna korelacja pomiędzy poszczególnymi frakcjami cholesterolu (TC–HDL, TC–LDL, HDL–LDL w mg/dl; TC – cholesterol całkowity, ang. *total cholesterol*; HDL – lipoproteina o dużej gęstości, ang. *high-density lipoprotein*; LDL – lipoproteina o małej gęstości. ang. *low-density lipoprotein*) w próbkach krwi 23 pacjentów. Proszę przygotować równanie regresji liniowej i spróbować na podstawie utworzonego równania regresji prostej dokonać predykcji TC z wykorzystaniem znanej wartości HDL i LDL: 204 mg/dl.

TC	HDL	LDL
167	10	22
155	27	101
131	21	67
194	24	127
229	27	125
183	26	121
201	21	97
247	22	136
173	23	102
206	20	104
209	24	140
227	30	140
185	24	96
217	26	123
183	22	127
162	21	107
200	27	114
156	34	87
155	29	95
206	20	134
184	28	119
204	29	133
185	29	101

Tabela 15.2. Dane pacjentów do ćwiczenia 1

15.3. Ćwiczenie 2

Proszę sprawdzić, na podstawie danych z ➥ **tabeli 15.3**, czy występują istotne statystycznie korelacje w przypadku chorób serca między różnymi zmiennymi w postaci ciśnienia skurczowego (mmHg), LDL (mmol/l) i wskaźnika otłuszczenia ciała (ang. *body adiposity index* – BAI, w %), a następnie utworzyć równanie regresji prostej w celu predykcji wartości LDL i BAI z zakresu ciśnienia skurczowego 130–160 mmHg.

Ciśnienie skurczowe	LDL	BAI
118	3,6	23,1
122	4,4	28,6
118	3,5	22,7
130	3,9	38,0
130	3,5	27,8
126	5,5	24,8
128	3,4	16,2
136	4,6	14,6
134	3,8	19,4
140	4,6	31,0
158	3,0	32,3
108	4,4	22,4
114	1,9	10,1
140	1,9	17,2
143	2,3	17,2
148	2,4	29,0
128	4,0	22,0
130	4,3	19,4
126	3,4	24,1
145	3,2	16,9
126	1,7	12,3
122	2,9	26,5
116	4,7	34,0

Tabela 15.3. Dane pacjentów do wykorzystania w ćwiczeniu 2

15.4. Ćwiczenie 3

Proszę sprawdzić, korzystając z danych zawartych w ➥ **tabeli 15.4**, czy istnieją istotne statystycznie korelacje pomiędzy trzema zmiennymi: cholesterolem całkowitym (mg/dl), koncentracją glukozy (mmol/l) i frakcją wyrzutową serca (%) u tych samych pacjentów. Jeśli istnieją, proszę wyznaczyć współczynniki korelacji (r) i dopasowania (r^2) oraz równanie regresji, a następnie spróbować przewidzieć wynik dla glukozy przy zawartości cholesterolu całkowitego 206 mg/dl.

Cholesterol całkowity	Glukoza	Frakcja wyrzutowa serca
207	4,6	46
176	4,9	55
229	4,8	45
178	4,8	59
182	4,7	50
180	5,2	47
225	5,0	54
168	4,4	46
156	4,7	45
154	3,8	52
190	4,5	52
137	4,5	55
168	5,1	45
190	4,6	51
197	3,9	54
147	5,4	55
183	5,0	50
154	4,8	49
125	4,3	51
179	4,3	50
159	5,1	55
174	4,6	47

Tabela 15.4. Dane pacjentów do wykorzystania w ćwiczeniu 3

Piśmiennictwo

1. Liang Y, Abbott D, Howard N, Lim K, Ward R, Elgendi M. How effective is pulse arrival time for evaluating blood pressure? Challenges and recommendations from a study using the MIMIC database. *J Clin Med*. 2019;8(3):337. https://doi.org/10.3390/jcm8030337

Krzysztof Kujawa,
Tomasz Hikawczuk

16 Zastosowanie regresji logistycznej w medycynie

Cel ćwiczenia. Utworzenie modelu regresji logistycznej opisującego zapadalność na nowotwory płuc w zależności od wskaźnika BMI, stężenia leukocytów we krwi oraz faktu palenia lub niepalenia papierosów.

Jedną z popularniejszych metod predykcji w modelowaniu występowania zjawisk w medycynie jest regresja logistyczna. Prognozuje ona zmienną jakościową. Istnieją modele regresji logistycznej, które służą objaśnianiu zmiennych mogących przyjmować kilka stanów (np. wystąpienie jednej z kilku chorób), jednak tu zajmiemy się prostszą wersją modelu służącego do predykcji stanu zmiennej mogącej przyjmować dwa stany (zmienna dychotomiczna, np. zdrowy/chory). Na podstawie regresji logistycznej można zbudować model zależności danego zjawiska od jego predyktorów (zmiennych objaśniających). Można jej również użyć jako narzędzia predykcyjnego, np. służącego do kalkulacji ryzyka wystąpienia komplikacji po zabiegu, przeprowadzonej operacji lub przeszczepieniu narządu. Model ten nie wymaga założenia rozkładu normalnego predyktorów, ale w przypadku stosowania modelu wieloczynnikowego zmienne objaśniające nie mogą być ze sobą silnie skorelowane, tj. nie powinna występować współliniowość. Ponadto kluczowym założeniem jest liniowa zależność logitu ilorazu szans (ang. *odds ratio – OR*) od każdego z predyktorów.

Funkcja logistyczna pozwala modelować zjawiska, w których zmiana wartości następuje po przekroczeniu wartości progowej danego parametru.

Prawdopodobieństwo wystąpienia danego zdarzenia w tego typu modelu wyprowadzonym na podstawie funkcji logistycznej oblicza się wg następującego wzoru:

$$f_z = \frac{e^z}{1 + e^z}$$

gdzie e oznacza liczbę Eulera, a z stanowi równanie opisujące zależności pomiędzy zmiennymi składowymi rozpatrywanego przypadku. Przykładowo model logistyczny może być zastosowany do przewidywania wystąpienia covidowego zapalenia płuc. Pacjentów można przebadać pod kątem 8 zmiennych: wieku, płci, występowania chorób serca i naczyń krwionośnych, nadciśnienia tętniczego, wskaźnika BMI, cukrzycy, przewlekłych chorób płuc i obniżonej odporności. Budując następnie model na podstawie tych zmiennych, można określać iloraz szans wystąpienia danego zjawiska u osób z określonymi parametrami, uwzględniając, jak każda składowa wpływa na wystąpienie stanu chorobowego i, co ważne, czy ten wpływ jest istotny statystycznie.

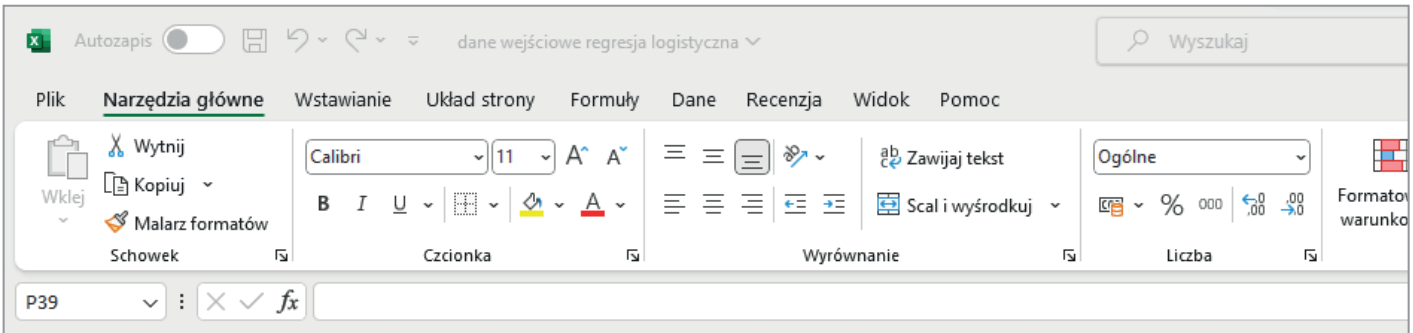
Model logistyczny uzyskany na podstawie wcześniejszych danych z uwzględnieniem najważniejszych czynników ryzyka można w praktyczny sposób wykorzystać do predykcji np. ryzyka zgonu wynikającego z późniejszych powikłań. Dodatkowo jakość predykcji można sprawdzać dla zdefiniowanej liczby biorców w danym okresie po określeniu prawdopodobieństwa ryzyka wystąpienia powikłań. Oczywiście im większa liczba dostępnych danych do wyznaczenia modelu logistycznego, tym bardziej wiarygodne wyniki można uzyskać ze względu na dokładniejsze dopasowanie równania regresji logistycznej do danych.

Weryfikacja modelu logistycznego jest niezbędna do sprawdzenia jego dopasowania metodą regresji logistycznej na podstawie wcześniej zebranych danych empirycznych w grupie badanych osób/próbie w przeniesieniu na proces leczenia nawet w postaci określenia efektywności przeprowadzonych zabiegów wyrażonych w postaci wartości %. Tego typu walidacja pozwala zwiększyć dokładność dopasowania równania do danych i zweryfikować konieczność zwiększenia liczby pacjentów od których niezbędne będzie pobranie danych celem zwiększenia liczby n w analizie, zmiany liczby zmiennych niezależnych lub też konieczności wdrożenia nowego modelu budowanego równolegle z dotychczas istniejącym na podstawie próbkowania w czasie rzeczywistym z zastosowaniem metod AI.

16.1. Przykład

Ryc. 16.1. Dane zebrane w procesie leczenia służące do określenia równania regresji logistycznej

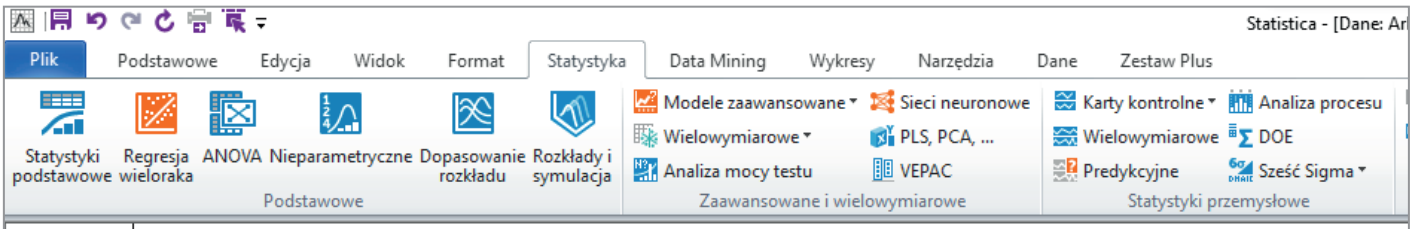
Na podstawie danych zgromadzonych w procesie leczenia pacjentów utworzymy model regresji logistycznej opisujący wystąpienie choroby w zależności od następujących czynników: wskaźnika BMI, stężenia leukocytów we krwi i liczby wypalonych papierosów przez dany zbiór pacjentów (➥ **ryc. 16.1**).



	A	B	C	D	E	F	G	H	I	J	K	L
1	BMI	Leukocyty (tys)	Palacz	Choroba	BMI	Leukocyty (tys)	Palacz	Choroba	BMI	Leukocyty (tys)	Palacz	Choroba
2	27,2	8	7	Tak	22,9	4	7	Tak	26,2	9	11	Nie
3	24,3	17	0	Nie	26,5	5	48	Tak	27,8	10	0	Nie
4	34,6	11	15	Tak	28,2	7	21	Nie	22,2	6	8	Tak
5	32,9	6	0	Nie	20,9	6	40	Tak	18,0	7	0	Nie
6	33,7	3	0	Nie	29,4	4	0	Nie	23,5	10	0	Nie
7	18,0	9	44	Tak	33,1	8	0	Nie	32,0	7	27	Tak
8	24,6	11	0	Nie	33,5	7	0	Tak	24,1	8	0	Nie
9	22,5	4	0	Tak	21,5	9	41	Tak	20,3	9	0	Nie
10	27,1	8	26	Tak	18,5	9	0	Nie	21,9	9	14	Tak
11	25,5	18	43	Nie	19,9	11	0	Nie	19,9	8	7	Tak
12	21,8	8	2	Tak	24,9	8	20	Nie	21,4	7	0	Tak
13	24,3	15	0	Tak	29,3	8	1	Tak	41,7	6	0	Nie
14	26,3	5	17	Nie	24,7	10	5	Tak	30,3	3	0	Nie
15	23,4	6	0	Nie	19,0	7	0	Tak	25,8	6	20	Tak
16	19,5	22	44	Nie	33,5	7	0	Tak				
17	27,2	12	2	Tak	30,4	7	0	Tak				
18	28,4	19	8	Tak	25,7	21	0	Nie				
19	19,7	4	35	Tak	22,0	7	33	Tak				
20	21,1	9	43	Tak	22,7	9	48	Tak				
21	21,7	10	24	Tak	23,1	26	0	Tak				
22	25,8	6	19	Tak	31,2	7	0	Nie				
23	33,2	11	0	Nie	20,8	9	45	Tak				
24	27,7	10	1	Tak	23,9	14	42	Tak				
25	21,1	12	0	Nie	30,0	6	0	Nie				
26	26,0	13	34	Tak	22,3	8	41	Nie				
27	18,6	8	14	Tak	22,3	4	26	Tak				
28	24,2	7	0	Nie	22,0	13	0	Tak				
29	27,1	16	27	Nie	31,8	8	35	Tak				
30	25,6	13	46	Tak	24,5	5	12	Tak				

Ryc. 16.2. Dane empiryczne i wybór kreatora regresji logistycznej

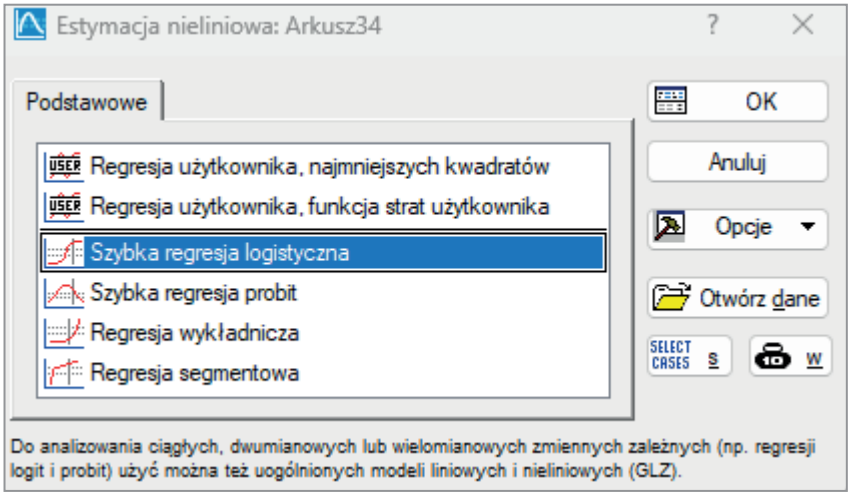
Po przeniesieniu danych z Excela regresja logistyczna uwzględniająca zmienną dychotomiczną i zmienne ilościowe możliwa jest do wyboru z karty Statystyka za pomocą polecenia Modele zaawansowane, a następnie Estymacja nieliniowa (➡ **ryc. 16.2**).



	1 BMI	2 Leukocyty (tys.)	3 Liczba papierosów	4 choroba	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	27,2	8	7	Tak						
2	24,3	17	0	Nie						
3	34,6	11	15	Tak						
4	32,9	6	0	Nie						
5	33,7	3	0	Nie						
6	18,0	9	44	Tak						
7	24,6	11	0	Nie						
8	22,5	4	0	Tak						
9	27,1	8	26	Tak						
10	25,5	18	43	Nie						
11	21,8	8	2	Tak						
12	24,3	15	0	Tak						
13	26,3	5	17	Nie						
14	23,4	6	0	Nie						
15	19,5	22	44	Nie						
16	27,2	12	2	Tak						
17	28,4	19	8	Tak						
18	19,7	4	35	Tak						
19	21,1	9	43	Tak						
20	21,7	10	24	Tak						
21	25,8	6	19	Tak						
22	33,2	11	0	Nie						
23	27,7	10	1	Tak						
24	21,1	12	0	Nie						
25	26,0	13	34	Tak						

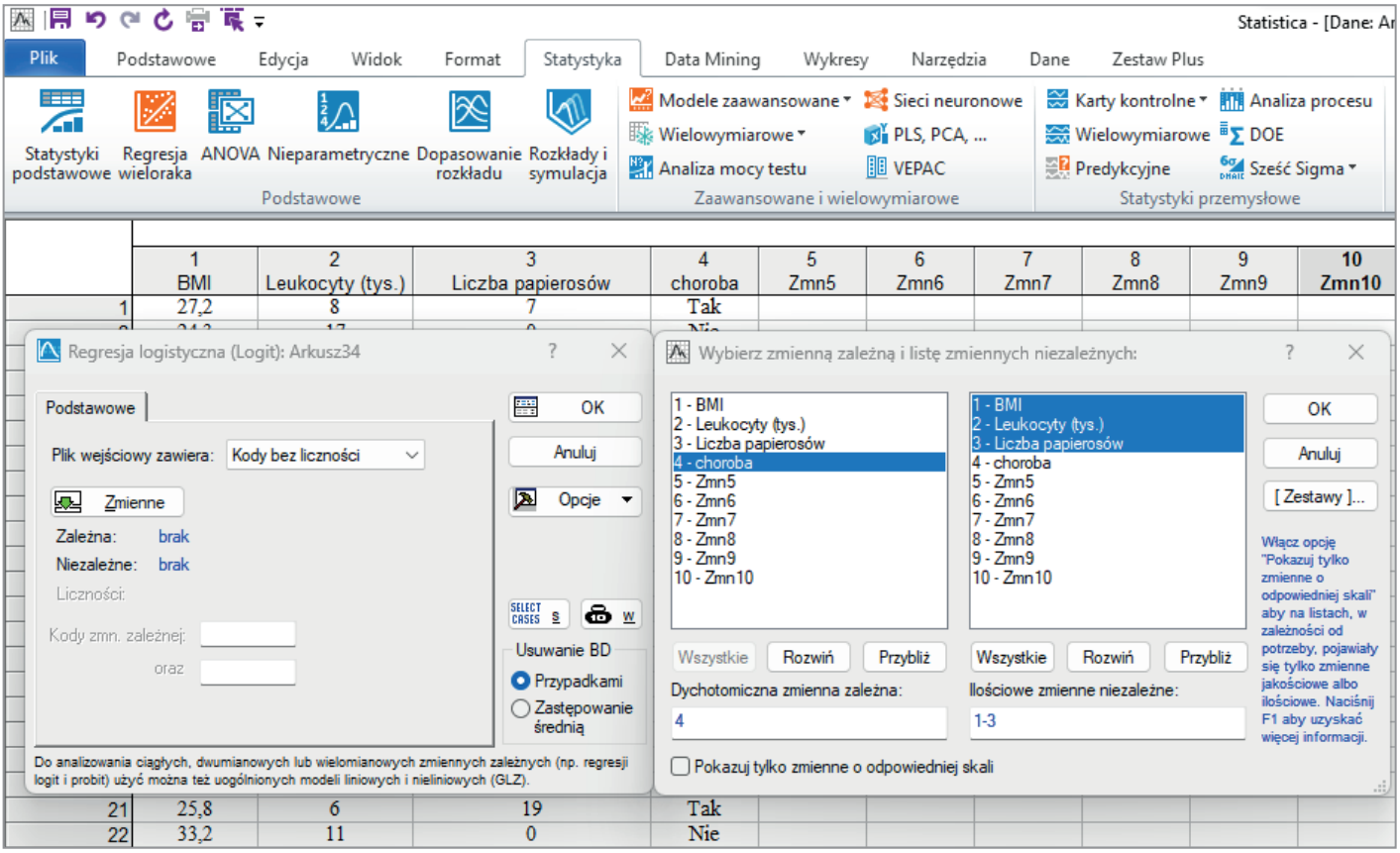
Otwiera się okno estymacji nieliniowej, gdzie możemy wybrać Szybką regresję logistyczną (➡ **ryc. 16.3**).

Ryc. 16.3. Okno estymacji nieliniowej



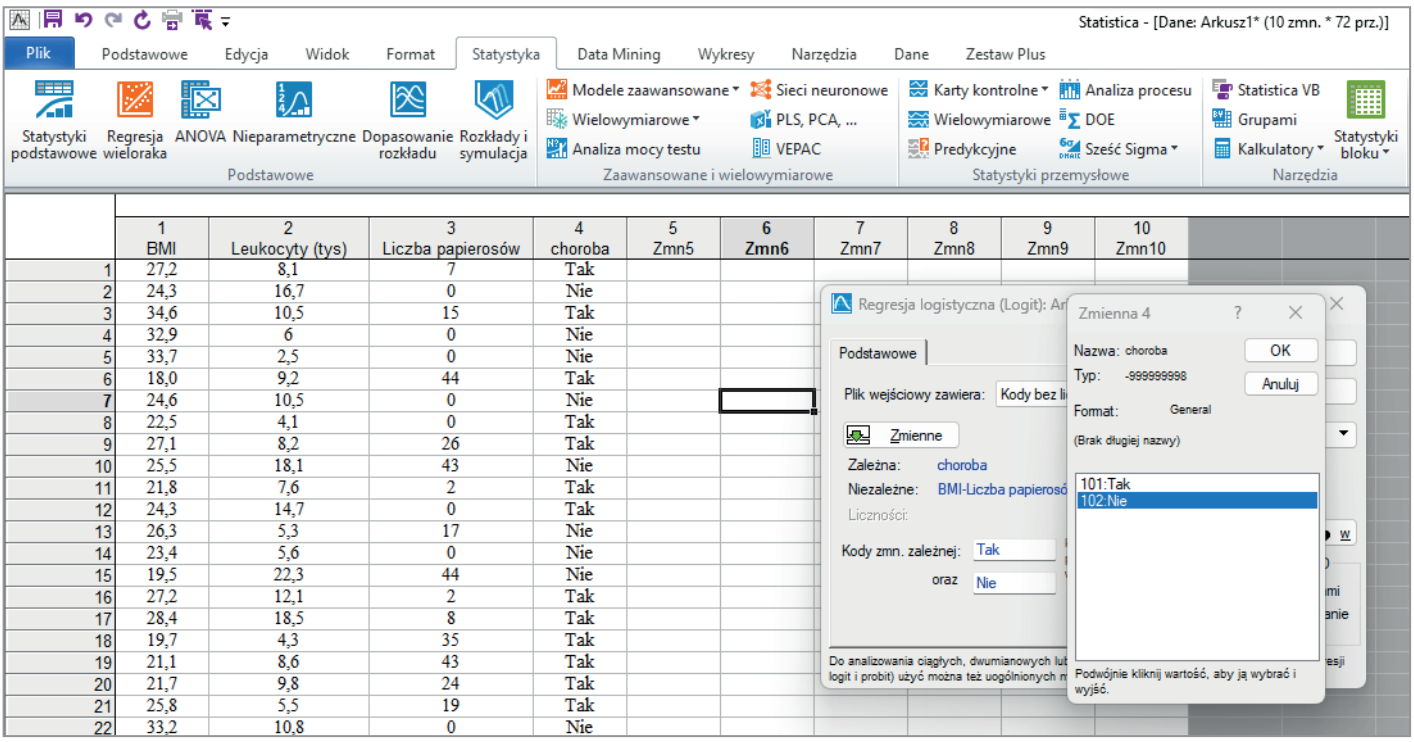
Po zatwierdzeniu wyboru Szybkiej regresji logistycznej możliwy jest wybór zmiennych uwzględnianych w danej analizie (➥ **ryc. 16.4**). Dychotomiczną zmienną zależną z dwiema możliwościami uzyskanego wyniku jest wystąpienie stanu chorobowego. W tym celu weryfikowana jest H_0 przy zastosowaniu testu χ^2 i poziomie istotności $p = 0{,}05$.

Ryc. 16.4. Wybór analizowanych zmiennych



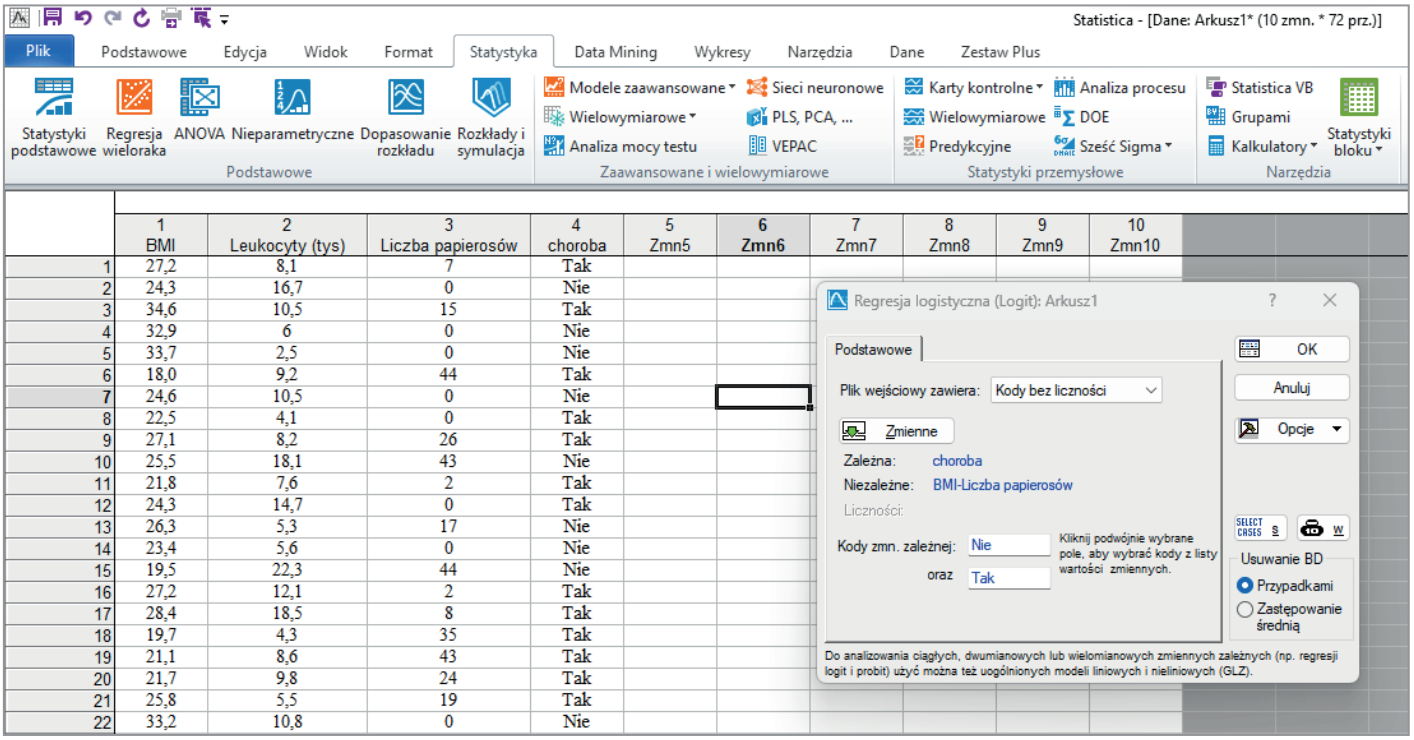
Dodatkowo zmieniamy sposób modelowania z braku występowania choroby na jej występowanie. Klikając dwa razy na kody zmiennej zależnej, mamy możliwość zmiany modelowania procesu w kierunku określenia prawdopodobieństwa wystąpienia choroby. W tym celu klikamy dwukrotnie Kody zmiennej zależnej i zamieniamy Tak na Nie, a w wierszu oraz zmieniamy Nie na Tak (➥ **ryc. 16.5**).

Ryc. 16.5. Zmiana kodowania w kierunku obliczenia prawdopodobieństwa wystąpienia choroby



Ryc. 16.6. Efekt zmiany kodowania

W efekcie otrzymujemy kodowanie jak na ➥ **ryc. 16.6**.

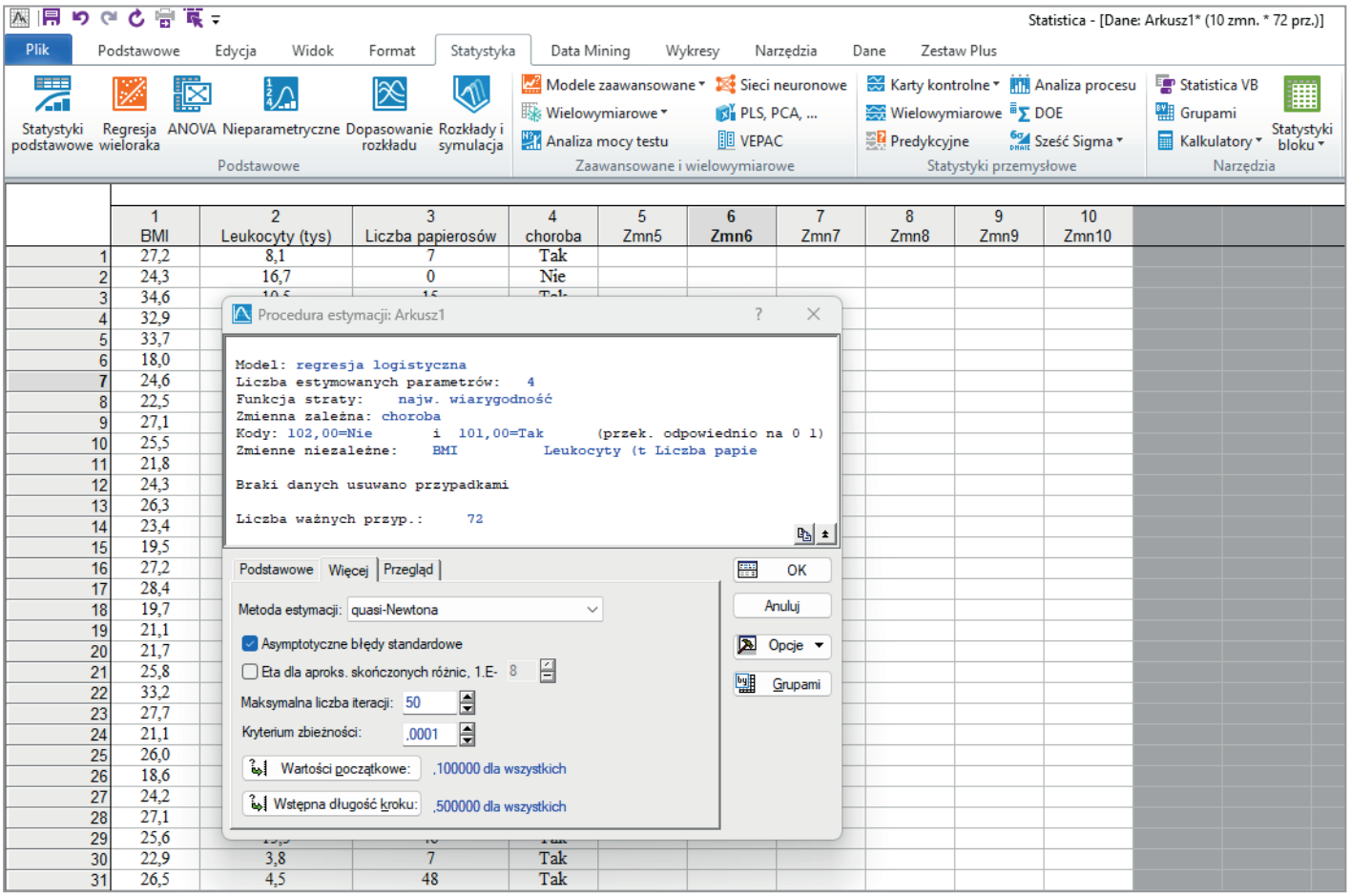


	1 BMI	2 Leukocyty (tys)	3 Liczba papierosów	4 choroba	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	27,2	8,1	7	Tak						
2	24,3	16,7	0	Nie						
3	34,6	10,5	15	Tak						
4	32,9	6	0	Nie						
5	33,7	2,5	0	Nie						
6	18,0	9,2	44	Tak						
7	24,6	10,5	0	Nie						
8	22,5	4,1	0	Tak						
9	27,1	8,2	26	Tak						
10	25,5	18,1	43	Nie						
11	21,8	7,6	2	Tak						
12	24,3	14,7	0	Tak						
13	26,3	5,3	17	Nie						
14	23,4	5,6	0	Nie						
15	19,5	22,3	44	Nie						
16	27,2	12,1	2	Tak						
17	28,4	18,5	8	Tak						
18	19,7	4,3	35	Tak						
19	21,1	8,6	43	Tak						
20	21,7	9,8	24	Tak						
21	25,8	5,5	19	Tak						
22	33,2	10,8	0	Nie						

Natomiast zmienne ilościowe rozpatrywane w tym przypadku to: BMI danej osoby, stężenie leukocytów w tys./µl oraz liczba wypalanych dziennie papierosów. Wybór zmiennych zatwierdzamy kliknięciem OK.

W kolejnym etapie tworzony jest model regresji (➥ **ryc. 16.7**). Analizowane są 4 zmienne: 1 jako zmienna dychotomiczna, pozostałe 3 w postaci zmiennych ilościowych. Jeżeli w którejkolwiek komórce wiersza danego pacjenta brakuje danych,

Ryc. 16.7. Procedura estymacji regresji logistycznej



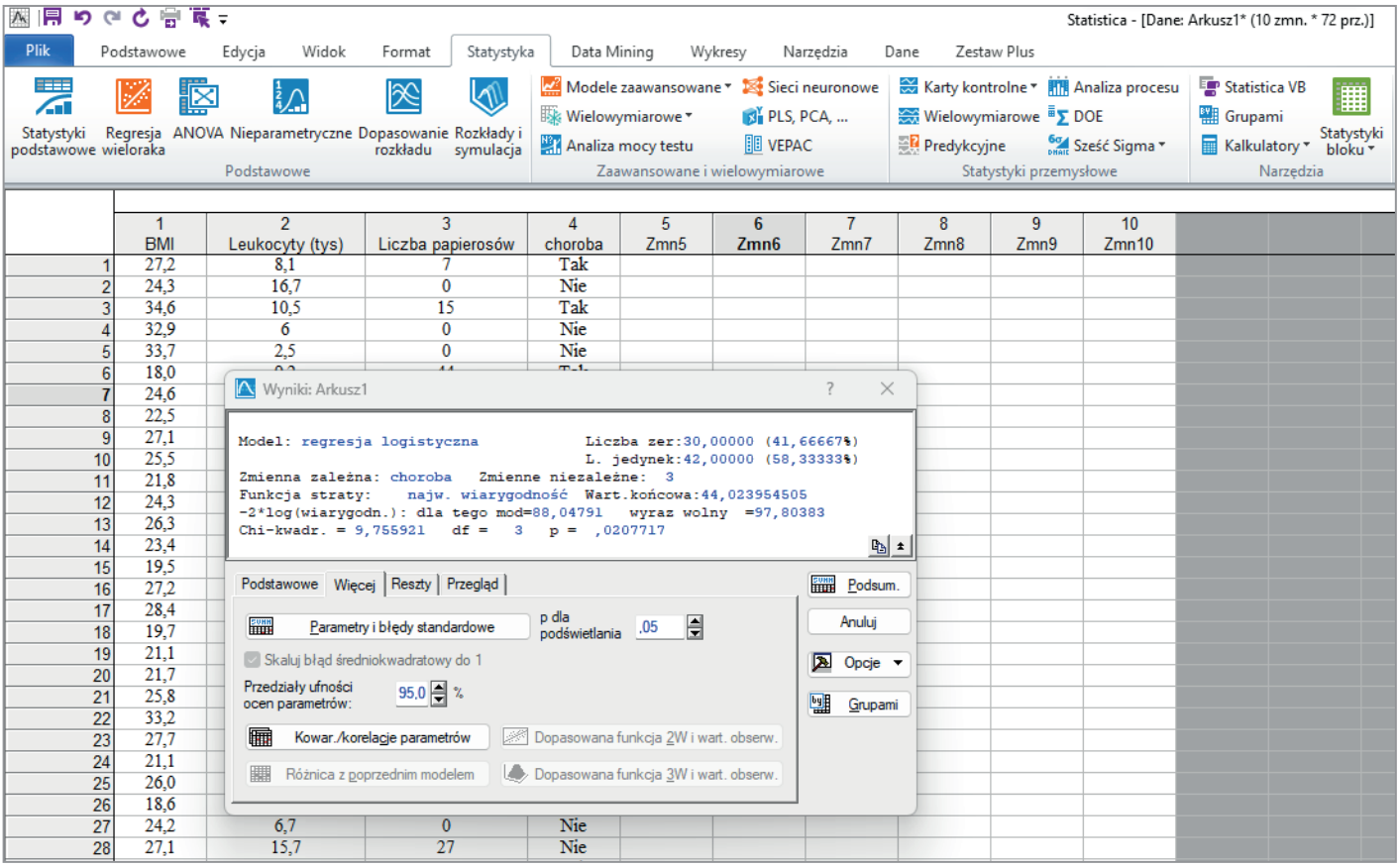
	1 BMI	2 Leukocyty (tys)	3 Liczba papierosów	4 choroba	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	27,2	8,1	7	Tak						
2	24,3	16,7	0	Nie						
3	34,6	[illegible]	[illegible]	[illegible]						
4	32,9									
5	33,7									
6	18,0									
7	24,6									
8	22,5									
9	27,1									
10	25,5									
11	21,8									
12	24,3									
13	26,3									
14	23,4									
15	19,5									
16	27,2									
17	28,4									
18	19,7									
19	21,1									
20	21,7									
21	25,8									
22	33,2									
23	27,7									
24	21,1									
25	26,0									
26	18,6									
27	24,2									
28	27,1									
29	25,6	[illegible]	[illegible]	[illegible]						
30	22,9	3,8	7	Tak						
31	26,5	4,5	48	Tak						

nie są one brane pod uwagę w modelu końcowym (oczywiście w przypadku wykrycia tego typu braku istnieje możliwość ich dodania do arkusza, jeśli są uwzględnione w kartotece lub bazie danych). Do wyboru jest kilka metod estymacji. W tym przypadku zastosowano Estymację quasi-Newtona i włączono Asymptotyczne błędy standardowe. Kolejnym krokiem jest kliknięcie OK (➥ **ryc. 16.7**).

Uzyskujemy w ten sposób informację nt. testu χ^2. Hipoteza zerowa testu zakłada brak wpływu którejś z analizowanych zmiennych na występowanie stanu chorobowego. Wartość $p = 0{,}021$ dla testu χ^2 nie pozwala na jej przyjęcie, w wyniku czego przyjmuje się H_1 wskazującą, że wystąpienie choroby jest zależne od którejś zmiennej (➥ **ryc. 16.8**).

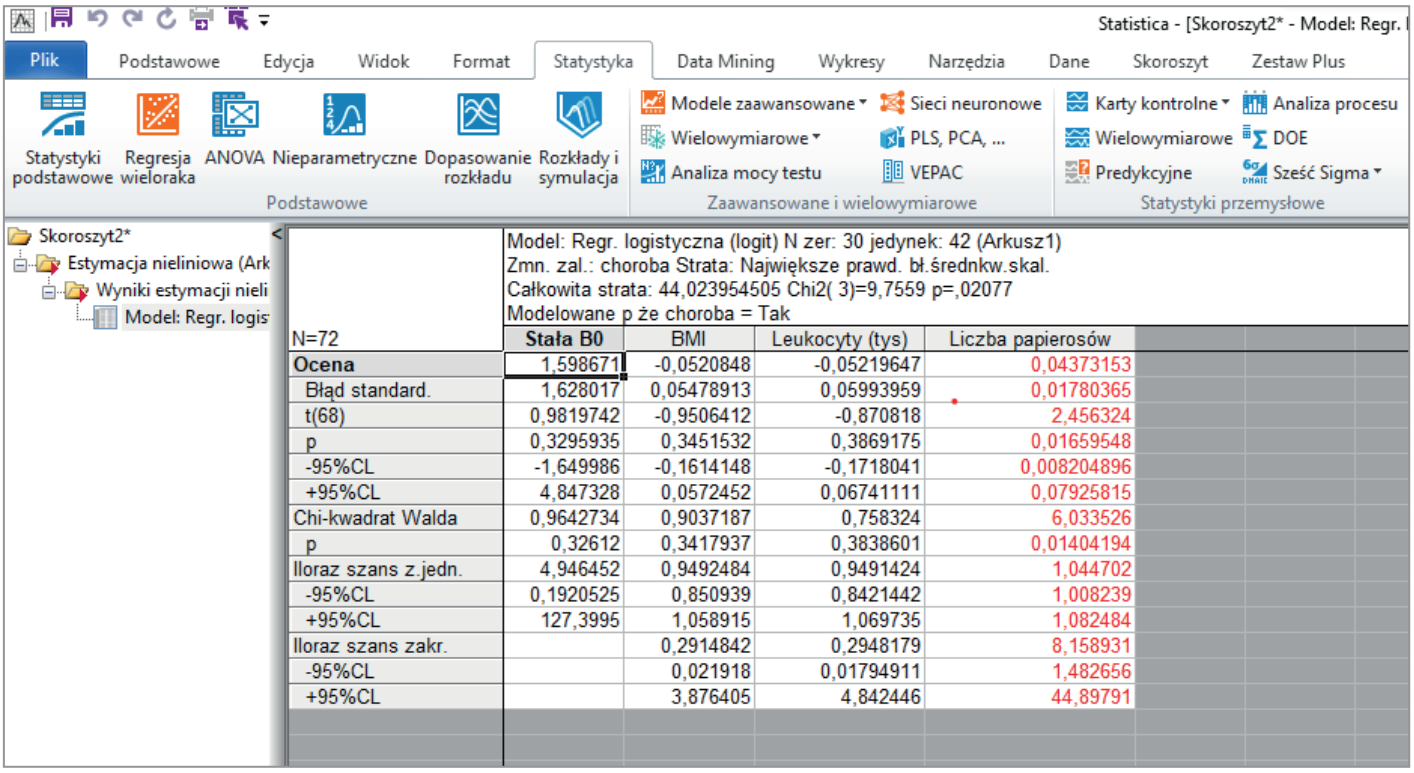
W zakładce Więcej można ustalić szczegóły dotyczące przedziału ufności i weryfikacji istotności wpływu danej zmiennej (➥ **ryc. 16.8**; standardowo ustawiamy $p < 0{,}05$; istnieje jednak możliwość zawężenia poziomu istotności do $p < 0{,}01$ lub zwiększenia w zależności od charakteru badań).

Ryc. 16.8. Wyniki modelowania regresji logistycznej



Ryc. 16.9. Wartości poszczególnych współczynników równania regresji

Następnie klikamy Podsumowanie i uzyskujemy wartość wyrazu ogólnego równania i wartości współczynników dla poszczególnych zmiennych stanowiących składowe równania modelu regresji logistycznej (➥ **ryc. 16.9**).



Model: Regr. logistyczna (logit) N zer: 30 jedynek: 42 (Arkusz1)
Zmn. zal.: choroba Strata: Największe prawd. bł.średnkw.skal.
Całkowita strata: 44,023954505 Chi2(3)=9,7559 p=,02077
Modelowane p że choroba = Tak

N=72	Stała B0	BMI	Leukocyty (tys)	Liczba papierosów
Ocena	1,598671	-0,0520848	-0,05219647	0,04373153
Błąd standard.	1,628017	0,05478913	0,05993959	0,01780365
t(68)	0,9819742	-0,9506412	-0,870818	2,456324
p	0,3295935	0,3451532	0,3869175	0,01659548
-95%CL	-1,649986	-0,1614148	-0,1718041	0,008204896
+95%CL	4,847328	0,0572452	0,06741111	0,07925815
Chi-kwadrat Walda	0,9642734	0,9037187	0,758324	6,033526
p	0,32612	0,3417937	0,3838601	0,01404194
Iloraz szans z.jedn.	4,946452	0,9492484	0,9491424	1,044702
-95%CL	0,1920525	0,850939	0,8421442	1,008239
+95%CL	127,3995	1,058915	1,069735	1,082484
Iloraz szans zakr.		0,2914842	0,2948179	8,158931
-95%CL		0,021918	0,01794911	1,482656
+95%CL		3,876405	4,842446	44,89791

Z wyników widocznych na rycinie płynie wniosek, że liczba wypalonych dziennie papierosów wiąże się z wystąpieniem stanu chorobowego. Potwierdza to też iloraz szans, w którym grupa narażona w zakresie przedziału ufności jest podatna na występowanie choroby, wynoszący 1,48–44,9 w zależności od liczby wypalonych dziennie papierosów. W równaniu uwzględniamy pierwszy wiersz wynikowy (Ocena). Współczynniki zostały przedstawione w ➥ **tabeli 16.1**.

Wyszczególnienie	Wartość współczynnika	Wartość p	-95% CI	+95% CI
Stała	1,599			
BMI	0,052	0,345	-0,161	0,057
Poziom leukocytów	-0,052	0,387	-0,172	0,067
Liczba papierosów dziennie	0,044	0,017	0,008	0,079

Tabela 16.1. Wartości współczynników składowych równania regresji logistycznej, wraz z przedziałami ufności i wartością *p*

W efekcie podstawiamy dane do równania regresji logistycznej:

$$P(x) = \frac{e^{1,599 - 0,052 \times BMI - 0,052 \times \text{stężenie leukocytów} + 0,044 \times \text{liczba papierosów dziennie}}}{1 - e^{1,599 - 0,052 \times BMI - 0,052 \times \text{stężenie leukocytów} + 0,044 \times \text{liczba papierosów dziennie}}}$$

17 Krzywa ROC w badaniach medycznych

Agnieszka Rusiecka,
Andrii Zinchuk,
Krzysztof Kujawa

17.1. Wprowadzenie do analizy ROC

Krzywa ROC (krzywa charakterystyki operacyjnej odbiornika; ang. *receiver operating characteristic curve*) to narzędzie, które znalazło szerokie zastosowanie w medycynie i biostatystyce, zwłaszcza w ocenie jakości testów diagnostycznych. Krzywą ROC wykorzystywano już podczas II wojny światowej podczas oceny zdolności operatorów radarów do wykrywania sygnałów w obecności zakłóceń. Dzisiaj to główne narzędzie służące do oceny testów diagnostycznych w medycynie oraz do oceny klasyfikatorów w biostatystyce.

Podstawowym celem analizy ROC jest graficzne przedstawienie zależności między czułością a specyficznością testu w całym zakresie możliwych punktów odcięcia (zob. niżej). Czułość (sensytywność) opisuje zdolność testu do prawidłowego identyfikowania osób chorych, natomiast specyficzność (swoistość) – do rozróżniania osób zdrowych (bardziej szczegółowy opis znajduje się w podrozdziale 17.3).

Na osi X układu współrzędnych zwierającego krzywą ROC umieszczony jest odsetek fałszywie dodatnich wyników (1-specyficzność), podczas gdy na osi Y przedstawiona jest czułość testu. Idealny test diagnostyczny miałby punkt w lewym górnym rogu wykresu (czułość = 1, 1-specyficzność = 0). W praktyce większość testów diagnostycznych daje krzywą, która przebiega między tym ideałem a przekątną, która reprezentuje test bez wartości diagnostycznej.

Obszar pod krzywą ROC (ang. *area under the curve* – AUC) przedstawia jednolitą miarę jakości testu, niezależnie od konkretnego punktu odcięcia. Punkt odcięcia to konkretny próg lub wartość używana do podziału wyników klasyfikacji na 2 klasy (np. dodatnie i ujemne). Określa, które przewidywane prawdopodobieństwa lub wyniki klasyfikatora zostaną uznane za pozytywne, a które za negatywne. Zmiana punktu odcięcia może mieć wpływ na czułość i swoistość klasyfikatora. Biorąc pod uwagę, że pole powierzchni całego układu współrzędnych wynosi 1 (ponieważ długość obu osi także wynosi 1), AUC = 1 wskazywałoby na idealny test (maksymalna czułość i specyficzność). Z kolei AUC = 0,5 wskazuje na brak wartości diagnostycznej testu. Większość testów używanych w diagnostyce reprezentuje moc diagnostyczną wyrażającą się wielkościami AUC 0,8–0,95. Co więcej, możliwe jest (i powinno się to robić) obliczenie przedziału ufności AUC, co pozwala na ocenę istotności statystycznej danego klasyfikatora.

Zastosowanie krzywej ROC w medycynie jest niezwykle ważne, ponieważ pozwala ocenić jakość testu diagnostycznego w kontekście jego czułości i specyficzności. Dzięki analizie ROC lekarze i badacze mogą dokładnie określić, jaki poziom czułości lub specyficzności jest akceptowalny dla danego testu w konkretnym zastosowaniu klinicznym.

W biostatystyce analiza ROC jest niezastąpiona przy ocenie skuteczności klasyfikatorów, szczególnie w sytuacjach, gdy klasy są niezrównoważone. Klasyfikator odnosi się do modelu, który dzieli dane na dwie klasy, takie jak np. wartości „wysokie" i „niskie". Krzywa ROC pozwala zaś na wizualizację zależności pomiędzy czułością (prawdopodobieństwem prawidłowej identyfikacji przypadków pozytywnych) a swoistością (prawdopodobieństwem prawidłowej identyfikacji przypadków negatywnych) przy różnych progach klasyfikacji. Poza medycyną analiza ROC znalazła zastosowanie również w innych dziedzinach, takich jak ekonomia czy inżynieria.

Dzięki temu lekarze i badacze mogą podejmować świadome decyzje dotyczące zastosowania danego testu w praktyce klinicznej.

17.2. Znaczenie krzywej ROC w badaniach medycznych

Współczesna medycyna opiera się na rzetelnych i precyzyjnych testach diagnostycznych, które stanowią fundament podejmowania kluczowych decyzji klinicznych. W tym kontekście krzywa ROC jest niezastąpionym narzędziem pozwalającym analizować relację między czułością a specyficznością testu. Dzięki niej możliwa jest obiektywna ocena jakości testów oraz ich porównanie, co ułatwia wybór najbardziej efektywnych rozwiązań diagnostycznych.

Krzywa ROC umożliwia również wyznaczenie optymalnego punktu odcięcia, równoważącego czułość i specyficzność, co pozwala dostosować testy do specyficznych wymagań klinicznych. Jest to szczególnie istotne w przypadku biomarkerów, które odgrywają kluczową rolę w diagnozowaniu i monitorowaniu chorób. Analiza ROC pozwala określić ich wartość i przydatność w praktyce medycznej.

Poza tradycyjnymi testami diagnostycznymi krzywa ROC znajduje zastosowanie w ocenie zaawansowanych klasyfikatorów, takich jak systemy diagnostyki obrazowej. Ponadto umożliwia analizę różnic w wydajności testów w różnych grupach pacjentów.

Dzięki wizualnemu przedstawieniu relacji między korzyściami testu, krzywa ROC pomaga lepiej zrozumieć mocne i słabe strony poszczególnych testów. Umożliwia to podejmowanie bardziej świadomych, precyzyjnych decyzji diagnostycznych.

Krzywa ROC jest kluczowym narzędziem wspierającym lekarzy i badaczy w tworzeniu skuteczniejszych metod diagnostycznych. Jej wszechstronność i precyzja przekładają się na wyższy poziom opieki nad pacjentami, czyniąc ją nieodzownym elementem współczesnej medycyny opartej na danych.

17.3. Czym jest czułość (sensytywność) i specyficzność (swoistość)

Ważnym elementem analizy medycznej jest trafność oceny stosowanych testów diagnostycznych. Istotne jest, aby podział na pacjentów chorych i zdrowych był rzeczywisty, z możliwie jak najmniejszą liczbą pomyłek. O rzetelności zastosowanego testu decyduje jego powtarzalność oraz odtwarzalność. Przez powtarzalność (ang. *repeatability*) rozumiemy możliwość osiągnięcia takich samych wyników przez tego samego badacza. Natomiast odtwarzalność (ang. *reproducibility*) pozwala na otrzymanie takich samych wyników przez innych badaczy.

Każdy test diagnostyczny może w niektórych przypadkach uzyskać wyniki różne od rzeczywistych. Miarami stosowanymi do oceny wiarygodności testu są czułość (sensytywność) i specyficzność (swoistość) testu.

Czułość, sensytywność (ang. *sensitivity*) testu opisuje zdolność wykrywania osób rzeczywiście chorych (posiadających daną cechę). Jeśli więc badamy grupę osób chorych, to czułość daje nam informacje, jaki procent z nich ma pozytywny wynik testu. Czułość na poziomie 100% oznaczałaby, że wszystkie osoby chore zostały poprawnie rozpoznane.

Specyficzność, swoistość (ang. *specificity*) opisuje zdolność wykrywania osób rzeczywiście zdrowych (bez danej cechy). Jeśli więc badamy grupę osób zdrowych, to swoistość daje nam informacje, jaka część z nich ma negatywny wynik testu.

Do omówienia czułości i swoistości testu wykorzystamy tabelę dwudzielczą (➡ **tabela 17.1**):

		Wyniki diagnozy klinicznej, określanej jako złoty standard		Suma
		chorzy	zdrowi	
Wyniki testu przesiewowego	dodatni	prawdziwie dodatni (PD)	fałszywie dodatni (FD)	PD + FD
	ujemny	fałszywie ujemni (FU)	prawdziwie dodatni (PU)	FU + PU
Suma		PD + FU	FD + PU	N

Tabela 17.1. Wyniki testu diagnostycznego

Dla tak sklasyfikowanych wyników czułość i swoistość można wyrazić wzorami:

$$\text{czułość} = \frac{PD}{PD + FU} \times 100\%$$

$$\text{swoistość} = \frac{PU}{FD + PU} \times 100\%$$

17.4. Przykłady zastosowań krzywej ROC w badaniach klinicznych

Współczesne badania kliniczne często polegają na ocenie i porównaniu różnych metod diagnostycznych, a krzywa ROC jest jednym z najistotniejszych narzędzi w tej dziedzinie. Daje ona nie tylko możliwość oceny skuteczności jednego testu, ale umożliwia również porównanie skuteczności wielu testów w jednym badaniu.

17.4.1. Rak nerek i *logistic kernel*

W niedawnym badaniu Khan i Akbar stosowali krzywą ROC do oceny wrażliwości biomarkera *logistic kernel* w diagnozie raka nerek. Wyniki wykazały, że biomarker ten jest znacznie czulszy od tradycyjnych metod diagnostycznych.[1]

17.4.2. Metoda gamma w fizyce medycznej

Szczurek badał przydatność metody gamma, stosując krzywą ROC w kontekście fizyki medycznej podczas XVI Kongresu Polskiego Towarzystwa Fizyki Medycznej. Analiza ta pomogła w identyfikacji granic dla różnych parametrów metody gamma, które były najbardziej odpowiednie dla różnych zastosowań klinicznych.[2]

17.4.3. Zastosowanie w epidemiologii

Kohn i Newman przedstawili innowacyjne podejście do interpretacji krzywej ROC i obszaru pod krzywą ROC, nazywając je „The Walking Man Approach". Chociaż ich metoda jest bardziej teoretyczna, ma potencjalne zastosowanie w praktycznej ocenie testów diagnostycznych w badaniach klinicznych.[3]

Krzywa ROC ma szerokie zastosowanie w badaniach klinicznych. Umożliwia ocenę i porównanie różnych metod diagnostycznych. Dzięki jej wszechstronności stała się standardem w wielu dziedzinach medycyny od onkologii po fizykę medyczną.

17.5. Zadania

Poniższe zadanie pomoże nam opanować obliczanie czułości i swoistości w programie Statistica.

Przeprowadzono badanie w kierunku skuteczności testu na występowanie alergii na lateks. Wzięło w nim udział 135 pacjentów. Za punkt odniesienia przyjęto dane z wywiadu (występowanie objawów po kontakcie z lateksem zadeklarowane w kwestionariuszu), na podstawie których nadwrażliwość na lateks stwierdzono u 29 badanych. Dla testu prowokacji (wywołanie reakcji, np. na alergen) rękawiczkami lateksowymi uzyskano wyniki prawdziwie dodatnie u 23 pacjentów, prawdziwie ujemne u 82, fałszywie dodatnie u 6 i fałszywie ujemne u 24.

Wprowadzamy dane do arkusza kalkulacyjnego (fragment danych, ➥ **ryc. 17.1**).

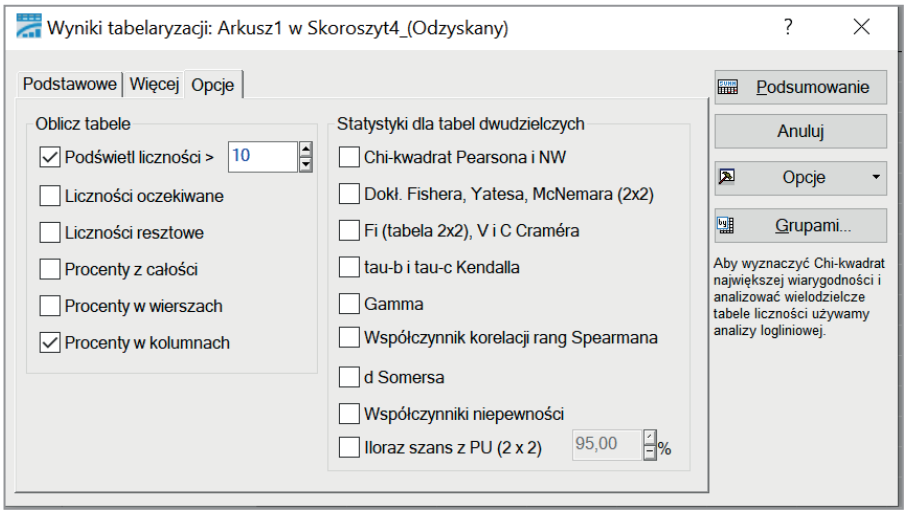
Ryc. 17.1. Fragment danych do analizy swoistości i czułości testu

	Arkusz1	
	WYWIAD uczulenie na lateks	TEST prowokacji
16	TAK	TAK
17	TAK	TAK
18	TAK	TAK
19	TAK	TAK
20	TAK	TAK
21	TAK	TAK
22	TAK	TAK
23	TAK	TAK
24	TAK	NIE
25	TAK	NIE
26	TAK	NIE
27	TAK	NIE
28	TAK	NIE
29	TAK	NIE
30	NIE	NIE
31	NIE	NIE
32	NIE	NIE
33	NIE	NIE
34	NIE	NIE
35	NIE	NIE
36	NIE	NIE

Arkusz1

Po wprowadzeniu danych klikamy Statystyki opisowe i tabele, a następnie wybieramy Tabele wielodzielcze. Przechodzimy na zakładkę Zbiorcza i wybieramy Zmienne. Na pierwszej liście zmiennych wybieramy „Wywiad uczulenie na lateks", na drugiej „Test prowokacji". Klikamy OK i ponownie OK. W oknie Wyniki tabelaryzacji przechodzimy na kartę Opcje i po lewej stronie okna zaznaczamy Procenty w kolumnach (➥ **ryc. 17.2**).

Ryc. 17.2. Wybór opcji dla tabeli wielodzielczej



Klikamy Podsumowanie. Otrzymamy tabelę wynikową (➥ **ryc. 17.3**).

Ryc. 17.3. Arkusz wynikowy

	Tabela liczności (Arkusz1 w Skoroszyt4_(Odzyskany)) Liczność oznacz. komórek > 10 (Nie oznaczono sum brzegowych)			
	WYWIAD uczulenie na lateks	TEST prowokacji TAK	TEST prowokacji NIE	Wiersz Razem
Liczba	TAK	23	6	29
% z kolumny		48,94%	6,82%	
Liczba	NIE	24	82	106
% z kolumny		51,06%	93,18%	
Liczba	Ogół	47	88	135

Z otrzymanej tablicy odczytujemy, że czułość testu wynosi 48,94%, a swoistość 93,18%.

Badacze często stają przed wyzwaniem obliczenia czułości i swoistości testu, gdy wyniki testów diagnostycznych mają charakter zmiennych ciągłych (np. poziom swoistych IgE lub średnica bąbla w teście punktowym) i mogą przybrać dowolną wartość w zakresie pomiarowym testu. Stają wtedy przed dylematem, od jakiej wartości wynik może być uznany za dodatni. W takiej sytuacji do wyboru optymalnych punktów odcięcia wyników pozytywnych od negatywnych, a zatem optymalizacji czułości i swoistości testu, służą krzywe ROC. Najczęściej stosuje się krzywą ROC przedstawiającą zależność częstości występowania wartości prawdziwie dodatnich (czułość) od wartości fałszywie dodatnich (1-swoistość) na każdym poziomie prawdopodobieństwa. Krzywa ROC pozwala więc na ustalenie najlepszego kompromisu pomiędzy czułością a swoistością testu, a tym samym osiągnięcie największej skuteczności np. w wykrywaniu uczulenia na konkretny alergen. Należy wyznaczyć optymalny punkt odcięcia w celu ustalenia wartości pozwalającej na rozróżnienie stanów: wystąpienia i braku wystąpienia badanego zjawiska przy określonym prawdopodobieństwie. Punkt ten powinien charakteryzować się zarówno wysoką czułością, jak i swoistością. W praktyce jednak czułość i swoistość mogą nie zostać uznane za równie ważne. W niektórych sytuacjach fałszywie ujemne wyniki mogą być ważniejsze niż fałszywie dodatnie, wówczas optymalny punkt odcięcia powinien charakteryzować się stosunkowo wysoką specyficznością. Jeśli jednak nie mamy założeń co do czułości i specyficzności, wówczas do wyznaczenia optymalnego punktu odcięcia możemy użyć indeksu Youdena (J). Przyjmuje on postać:

$$J = \textit{czułość} + \textit{specyficzność} - 1.$$

W takiej sytuacji optymalny punkt odcięcia odpowiada przypadkowi, dla którego wartość J osiąga maksimum (maksymalna możliwa wartość dla indeksu Youdena to 1, minimalna to 0).

Do wyznaczenia najlepszego punktu odcięcia stosuje się również metodę graficzną, gdzie punkt przecięcia wykresów czułości i specyficzności dla interesującego nas parametru wskazuje poszukiwaną wartość. Jest to wartość o maksymalnej liczbie wystąpień zjawiska, które zostały poprawnie sklasyfikowane.

Wykonajmy kolejne zadanie.

Firma opracowuje nowy test ELISA do wykrywania kleszczowego zapalenia mózgu u pacjentów. Sprawdzano stężenie chemokin CCL4 w płynie mózgowo-rdzeniowym u pacjentów, których stan zdrowia był znany oraz u pacjentów chorych (których wyniki potwierdzono metodami referencyjnymi). Wyniki testu ELISA przedstawia ➥ **tabela 17.2**.

test ELISA	**liczba zdrowych**	**liczba chorych**
20–40	42	11
40–60	21	19
60–80	12	23
> 80	5	87
N	80	140

Tabela 17.2. Rozkład częstości testu Elisa

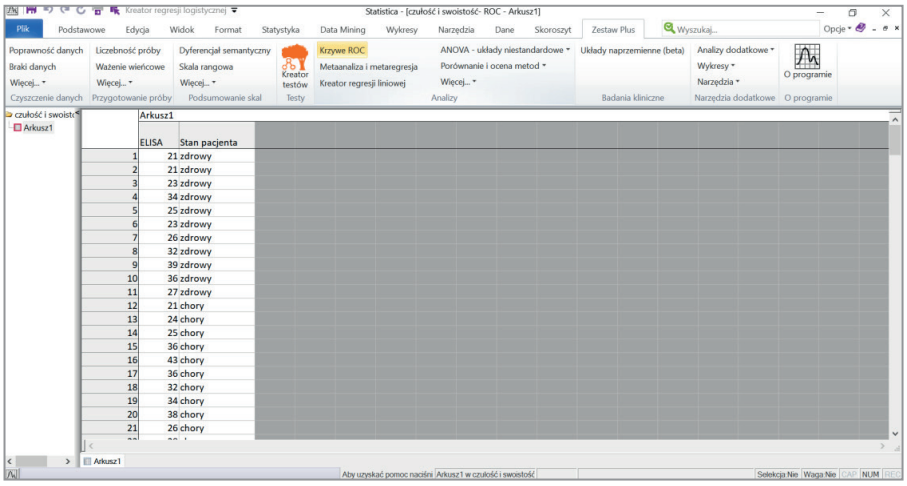
Ocenę skuteczności nowego testu należy zacząć od wyliczenia jego czułości i swoistości przy odpowiednich wartościach progowych. Trzeba pamiętać, że warunkiem wyliczenia czułości i swoistości jest wskazanie, który wynik jest prawidłowy, a który nieprawidłowy. W naszym zadaniu zaproponujemy 5 progów, dla których wyliczymy czułość i swoistość. Dane rozmieszczono w ➥ **tabeli 17.3**.

test ELISA	Czułość	Swoistość
40	53%	92%
60	78%	86%
80	98%	38%

Tabela 17.3. Czułość i swoistość testu ELISA

Łatwo można zaobserwować, że kiedy czułość testu rośnie, maleje jego swoistość i odwrotnie – swoistość testu rośnie przy malejącej czułości. Należy pamiętać, że testy charakteryzujące się dużą czułością rzadko pomijają osoby chore i warto je stosować, gdy chcemy wykryć większość przypadków chorobowych. Czułość wynosi 100%, jeżeli każda osoba chora zostanie tak oznaczona (żaden przypadek choroby nie zostanie przeoczony). Natomiast test o wysokiej swoistości rzadko kwalifikuje zdrowych jako chorych i sprawdza się raczej, gdy chcemy wykryć osoby zdrowe w badanej grupie. Swoistość wynosząca 100% oznacza, że każda osoba zdrowa jest rozpoznana jako zdrowa. Jak zatem wybrać wartość progową na skali ciągłej, pozwalającą nam zdecydować o skuteczności danego testu? Wykonamy wykres krzywych ROC dla danych z powyższego zadania. Ten typ analizy jest dostępny w programie Statistica w ramach Zestawu Medycznego, Zestawu do Analiz Marketingowych i Rynkowych oraz Zestawu Plus.

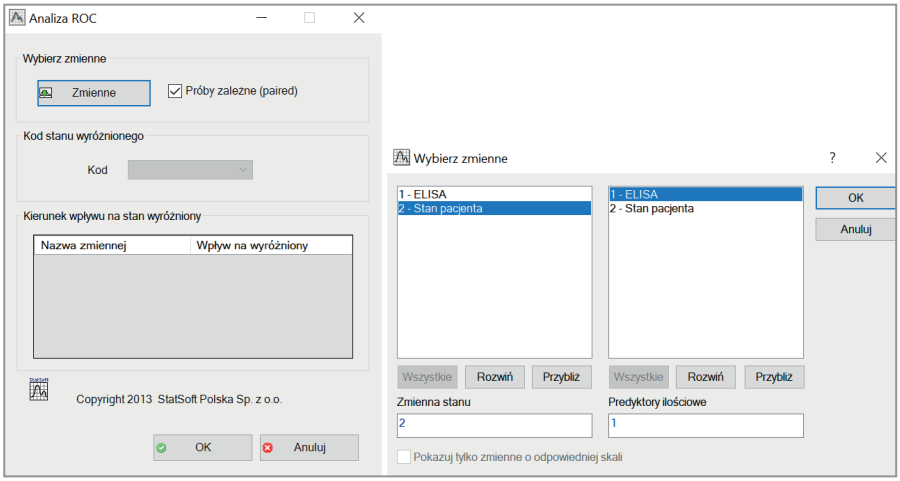
W głównym menu programu Statistica, należy wybrać kartę Zestaw Plus lub jej odpowiednik (w zależności od zainstalowanego dodatku), a z niej polecenie Krzywe ROC (➥ **ryc. 17.4**).



Ryc. 17.4. Wybór analizy ROC w menu programu Statistica

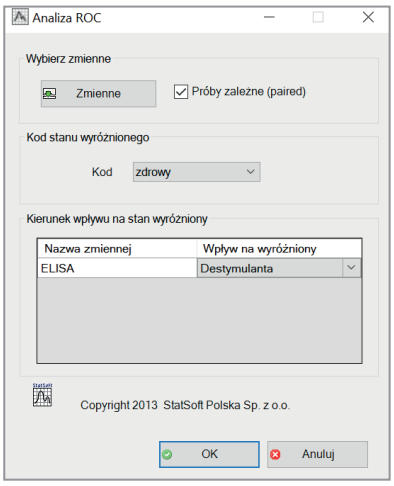
Klikamy przycisk Zmienne i wskazujemy zmienną stanu oraz predyktor ilościowy. Jako zmienną stanu wybieramy „Stan pacjenta" (zdrowy/chory), natomiast jako predyktor ilościowy – „ELISA". Klikamy OK (➥ **ryc. 17.5**).

Ryc. 17.5. Wybór odpowiednich zmiennych

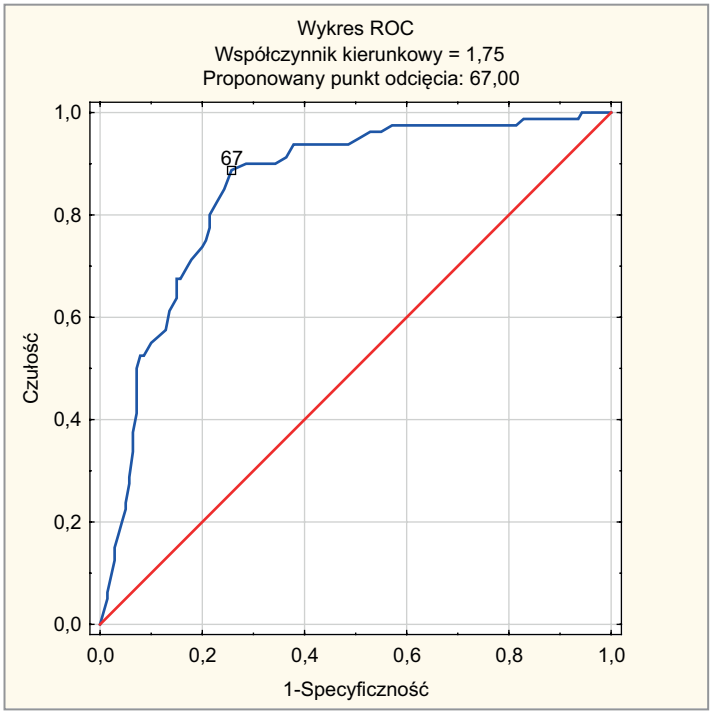


Na rozwijanej liście Kod wskazujemy kod stanu wyróżnionego. W naszym przypadku jest to kod „zdrowy". Musimy również zdecydować, czy wynik testu ELISA jest stymulantą czy destymulantą. Jeśli podejrzewamy, że większe wartości zmiennej ilościowej wpływają na częstsze występowanie stanu wyróżnionego, możemy założyć, że zmienna ilościowa będzie stymulantą. Na podstawie danych z ➥ tabeli 17.1 możemy stwierdzić, że wyższe wartości testu ELISA oznaczają więcej stwierdzonych przypadków zachorowań. Nasza zmienna jest więc destymulantą stanu „zdrowy" (➥ **ryc. 17.6**).

Ryc. 17.6. Wybór ustawień dla metody



Po ustawieniu odpowiednich parametrów klikamy OK i przechodzimy do nowego okna. Klikamy przycisk Wykres ROC. W wyniku analizy otrzymujemy wykres krzywych ROC wraz z zaznaczonym punktem odcięcia (➥ **ryc. 17.7**).



Ryc. 17.7. Wykres ROC z zaznaczonym punktem odcięcia

Pole pod krzywą ROC jest miarą zdolności testu do rozgraniczenia wyników prawidłowych i nieprawidłowych. Im większa jest powierzchnia pod krzywą, tym większa moc diagnostyczna testu. W praktyce im bardziej krzywa jest wygięta w stronę lewego górnego rogu (większa powierzchnia pod wykresem), tym wynik jest dokładniejszy. Interpretację krzywej ROC znacznie ułatwia dodanie przekątnej, która jest linią referencyjną i obliczenie pola pod powierzchnią krzywej. Im bardziej więc krzywa zbliża się do przekątnej, tym test jest mniej dokładny. Należy również pamiętać, że AUC podajemy wraz z przedziałem ufności, ponieważ wartość ta jest wyliczana na podstawie danych z próby i stanowi jedynie oszacowanie prawdziwej, nieznanej wartości AUC dla populacji. Przedział ufności wskazuje nam, że poszukiwana przez nas rzeczywista wartość mieści się w pewnym przedziale z założonym prawdopodobieństwem.

Dla danych z naszego przykładu wartość progowa wynosi 67 (➥ **ryc. 17.8**).

Ryc. 17.8. Raport wyników dla krzywej ROC wraz z podaniem przedziałów ufności

	P.uf.metodą BCA Liczba prób bootstrap:1000		
	1 **Oszacowanie**	2 Gr. dolna 95% PU	3 Gr. górna 95% PU
Punkt odcięcia	67	113	113
Indeks Youdena	0,6304	0,0375	0,0375

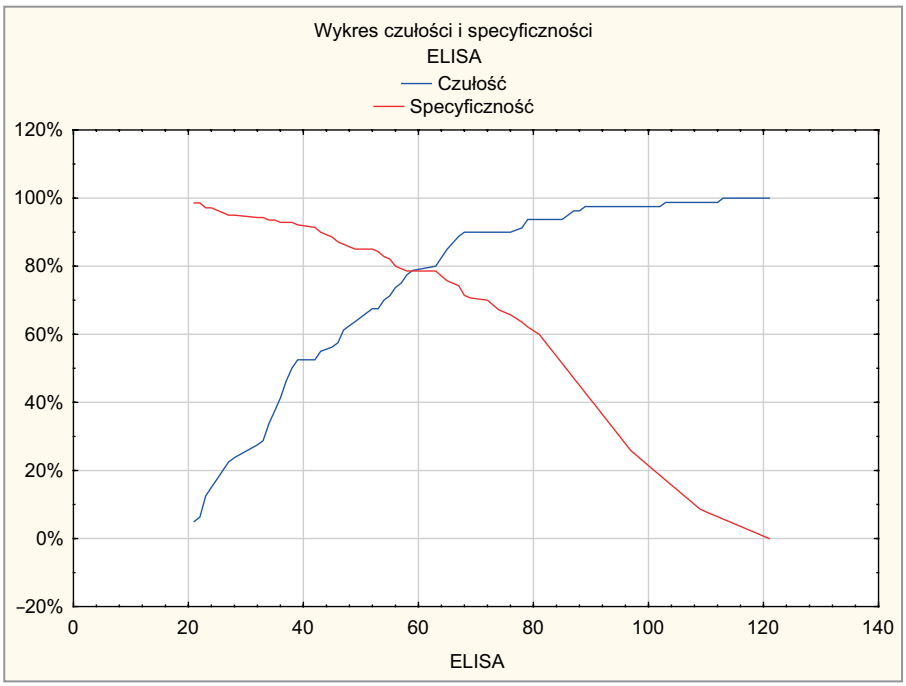
Klikając przycisk Raport, otrzymujemy m.in. wartość pola pod wykresem krzywej ROC (➥ **ryc. 17.9**).

Ryc. 17.9. Raport wyników dla krzywej ROC

	Zmienna: ELISA					
	AUC	SE	AUC Dolny 95%	AUC Górny 95%	z	p
1	0,852	0,027	0,8	0,904	13,194	0,0000

Dla naszego przykładu AUC, czyli pole pod wykresem krzywej ROC, wynosi 0,852 (95% przedział ufności 0,8–0,904; jak wspomniano wyżej, większość testów używanych w diagnostyce reprezentuje moc diagnostyczną wyrażoną AUC pomiędzy 0,8 i 0,95).

Możemy również wygenerować wykres przedstawiający zależność między czułością a specyficznością testu (➡ **ryc. 17.10**).



Ryc. 17.10. Zależność między czułością a specyficznością testu

Pozytywny wynik badania o wysokiej czułości i swoistości nie musi jednak oznaczać wysokiego prawdopodobieństwa tego, że osoba jest chora, jeżeli w badanej populacji choroba występuje dość rzadko (np. badania przesiewowe generalnie zdrowej populacji).

Dlatego do oceny, czy pacjent jest chory, gdy znane są wyniki testu, używamy następujących miar:

- wartości predykcyjnej dodatniej (ang. *positive predictive value* – PPV) – prawdopodobieństwa choroby, gdy wyniki testu są dodatnie; można ją wyliczyć, stosując wzór:

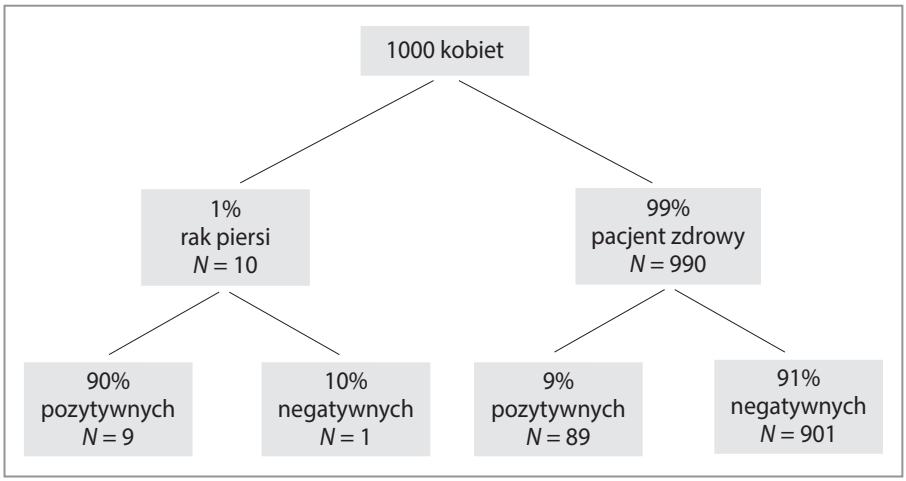
$$WP\,(+) = \frac{PD}{PD + FD} \times 100\%$$

- wartości predykcyjnej ujemnej (ang. *negative predictive value* – NPV) – prawdopodobieństwa, że pacjent jest zdrowy, gdy wyniki są ujemne; obliczamy ją ze wzoru:

$$WP\,(-) = \frac{PU}{PU + FU} \times 100\%$$

Należy pamiętać, że wartość predykcyjna zależy od częstości występowania zjawiska (np. choroby) w populacji, z której pochodzi pacjent.

Wyobraźmy sobie sytuację, że badamy populację 1000 kobiet pod kątem występowania raka piersi. Wiemy, że częstość występowania choroby w populacji wynosi 1%. Dysponujemy testem o czułości wynoszącej 90% i swoistości równej 91%. Jakie jest prawdopodobieństwo, że osoba jest chora? W celu analizy posłużymy się schematem (➥ **ryc. 17.11**).



Ryc. 17.11. Analiza prawdopodobieństwa wykrycia choroby przy zastosowaniu testów o określonej czułości i swoistości

Ocenę prawdopodobieństwa wystąpienia choroby można obliczyć następująco:

$$P\,(\text{rak piersi} \mid \text{test pozytywny}) = \frac{0{,}01 \times 0{,}9}{0{,}01 \times 0{,}9 + 0{,}99 \times 0{,}09} = \frac{9}{9 + 89} = 0{,}09$$

Jak widzimy, tylko ok. 10% wyników jest prawdziwych, a zdecydowana większość pozytywnych wyników jest fałszywa.

Piśmiennictwo

1. Khan JA, Akbar A. Logistic kernel: A sensitive biomarker for kidney cancer by ROC curve. *International Journal of Applied Sciences & Development*. 2023;2:120–132. doi:10.37394/232029.2023.2.13
2. Szczurek Ł, Juszkat R, Szczurek J, Sosnowski P. Ocena przydatności metody gamma w oparciu o krzywą ROC. *Inż Fiz Med*. 2019;8(3):218.
3. Kohn M, Newman T. The Walking Man Approach to Interpreting the ROC curve and area under the ROC Curve. *J Clin Epidemiol*. 2023;162:182–186. doi:10.1016/j.jclinepi.2023.07.020

W podręczniku opisano najważniejsze zagadnienia dotyczące podstaw technologii informacyjnej i biostatystyki zarówno w naukach medycznych, jak i w praktyce systemu ochrony zdrowia.

W pierwszej części, dotyczącej technologii informacyjnej, przedstawiono sposoby wykorzystania programów pakietu MS Office do zbierania, przetwarzania i prezentowania danych. Ukazano możliwości edycji tekstu, analizy i przedstawiania informacji w arkuszach kalkulacyjnych i bazach danych, a także tworzenia klarownych i skutecznych prezentacji multimedialnych.

Część dotycząca biostatystyki oparta jest na wykorzystaniu programu Statistica. Ukazuje zasady planowania obserwacji i doświadczeń i opisuje procesy poprawnego wnioskowania na podstawie danych. Przedstawia pojęcia związane ze statystyką opisową, typami zmiennych, sposobami badania rozkładów danych, testowaniem hipotez statystycznych i odpowiednim interpretowaniem wyników.

Uniwersytet Medyczny
im. Piastów Śląskich we Wrocławiu

ISBN 978-83-7055-698-3
9 788370 556983