Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską
NCBR Narodowe Centrum Badań i Rozwoju

Materiały do kursu: *Komputerowa analiza danych z wykorzystaniem programu Statistica*

dr inż. Anna Olszewska, prof. PB

Białystok, 2026 r.

Spis treści

1. Metodologia badań statystycznych i techniki doboru próby

1.1. Wprowadzenie do procesów statystycznych i zjawisk masowych

Pojęcie procesu statystycznego jest nierozerwalnie związane z badaniem tzw. zjawisk masowych. **Zjawiskiem masowym** nazywamy takie procesy lub zdarzenia, które charakteryzują się dużą częstotliwością występowania, powtarzalnością oraz wielością obiektów, których dotyczą. W przeciwieństwie do zjawisk jednostkowych, gdzie wynik pojedynczego eksperymentu może być dziełem czystego przypadku, zjawiska masowe ujawniają stabilne, dające się matematycznie opisać prawidłowości (nazywane prawidłowościami wielkich liczb). Do klasycznych przykładów należą akty kupna-sprzedaży na giełdzie, dobowe fluktuacje ruchu sieciowego na serwerach, procesy demograficzne zachodzące w społeczeństwie czy też powtarzalne cykle produkcyjne na zautomatyzowanych liniach montażowych.

Analiza dowolnego procesu statystycznego wymaga zrozumienia, że na każdy badany obiekt oddziałuje skomplikowany układ sił zewnętrznych i wewnętrznych, który w metodologii statystycznej dzieli się na dwie zasadnicze kategorie: czynniki główne oraz czynniki uboczne. **Czynniki główne**, nazywane również systematycznymi, to fundamentalne determinanty procesu, które działają z jednakową siłą i w tym samym kierunku na wszystkie jednostki podlegające badaniu. To one kształtują ogólny trend, bazowy poziom zjawiska oraz jego zasadniczy kierunek rozwoju. Z kolei **czynniki uboczne**, określane jako losowe, to całe spektrum drobnych, trudnych do przewidzenia i często niemożliwych do kontrolowania impulsów. Czynniki losowe nie działają jednakowo – mogą aktywować się tylko przy niektórych obiektach, w różnych momentach i z odmienną intensywnością. Istnienie czynników ubocznych jest bezpośrednią przyczyną tzw. zmienności wewnątrz zbiorowości. Sprawia ona, że dwa obiekty podlegające identycznemu działaniu czynników głównych (np. dwa smartfony wyprodukowane na tej samej taśmie, przez te same maszyny i z tych samych komponentów) będą różnić się mikro-detalami lub ostateczną żywotnością ogniwa. Zadaniem statystyki jest odfiltrowanie szumu generowanego przez czynniki losowe w celu precyzyjnego zidentyfikowania działania czynników systematycznych.

1.2. Podstawowe pojęcia statystyki matematycznej

Przed uruchomieniem jakiejkolwiek procedury analitycznej w programie komputerowym, badacz musi zdefiniować strukturę logiczną swojego eksperymentu za pomocą trzech elementarnych pojęć statystyki matematycznej: populacji generalnej, jednostki

statystycznej oraz próby. **Populacja generalna** (często nazywana zbiorowością statystyczną, masą statystyczną lub uniwersum) stanowi kompletny, nieograniczony lub ograniczony zbiór wszystkich hipotetycznych i rzeczywistych obiektów, które są przedmiotem zainteresowania badawczego. Kluczową cechą populacji jest nieidentyczność jej elementów pod względem analizowanych cech. Gdyby wszystkie elementy były identyczne, statystyka straciłaby rację bytu, gdyż zbadanie jednego obiektu dawałoby pełną wiedzę o całości.

Jednostka statystyczna to element składowy populacji, stanowiący najmniejszą, niepodzielną strukturę podlegającą bezpośredniej obserwacji, zliczeniu lub fizycznemu pomiarowi. W zależności od kontekstu badania, jednostką statystyczną może być człowiek (np. w badaniach opinii publicznej), pojedyncze przedsiębiorstwo (w analizach makroekonomicznych), pojedyncza partia materiału budowlanego (w testach wytrzymałościowych) czy też konkretny przedział czasu (np. godzina pracy reaktora w analizie wydajności).

Z przyczyn ekonomicznych, fizycznych, a często także logistycznych (np. gdy badanie ma charakter niszczący, jak testowanie odporności poduszek powietrznych na wybuch), zbadanie całej populacji generalnej jest niemożliwe. Wówczas statystycy posługują się **próbą** (zbiorowością próbną), czyli ściśle określonym podzbiorem populacji generalnej, który został fizycznie poddany procedurze pomiarowej. Aby wnioskowanie na podstawie próby miało jakąkolwiek wartość naukową, musi zostać spełniony rygorystyczny postulat reprezentatywności. **Reprezentatywność** oznacza, że struktura próby pod względem kluczowych właściwości musi stanowić wierne, proporcjonalne odzwierciedlenie struktury całej populacji generalnej – można powiedzieć, że próba musi być jej miniaturowym, nieskrzywionym obrazem.

1.3. Klasyfikacja badań statystycznych

Metodologia statystyczna klasyfikuje badania według dwóch głównych osi podziału: kryterium obszaru (czyli zakresu objęcia populacji) oraz kryterium czasu (czyli częstotliwości i ciągłości zbierania danych). Z punktu widzenia obszaru, badania dzielimy na pełne i częściowe. **Badania pełne** (całkowite) polegają na przeprowadzeniu pomiaru na każdej jednostce statystycznej tworzącej populację generalną. Najbardziej wyrazistym przykładem takiego podejścia jest spis powszechny ludności i mieszkań, pełna inwentaryzacja środków trwałych w przedsiębiorstwie lub stuprocentowa, zautomatyzowana kontrola jakości w systemach produkcyjnych. Badania pełne dają stuprocentową pewność wyniku, są jednak kosztowne i czasochłonne.

Alternatywą są **badania częściowe** (niepełne), w których obserwacji poddaje się wyłącznie wybraną część populacji. W ramach badań częściowych kluczowe miejsce zajmują badania reprezentacyjne. Są one oparte na matematycznych regułach losowego doboru próby, co przy użyciu aparatu statystyki matematycznej pozwala na precyzyjne uogólnianie uzyskanych wyników na całą populację z mierzalnym błędem i poziomem ufności. Oprócz nich w badaniach częściowych wyróżnia się metody ankietowe (często oparte na dobrowolności odpowiedzi) oraz badania monograficzne, które polegają na ekstremalnie szczegółowej, wszechstronnej analizie zaledwie kilku celowo wybranych jednostek (np. dogłębne studium przypadku jednego, konkretnego kryzysu w wybranym przedsiębiorstwie).

Biorąc pod uwagę kryterium czasu, badania statystyczne przyjmują postać ciągłą, okresową lub doraźną. **Badania ciągłe** polegają na permanentnej, nieprzerwanej rejestracji faktów i zdarzeń w momencie ich wystąpienia. Przykładem jest bieżące rejestrowanie zgonów i urodzeń w Urzędach Stanu Cywilnego, ciągły monitoring emisji gazów przez systemy kominowe fabryki czy automatyczne logowanie błędów sieciowych przez systemy operacyjne. **Badania okresowe** są realizowane systematycznie, w sztywno zdefiniowanych i powtarzających się odstępach czasu. Zaliczamy do nich przykładowo raporty finansowe spółek giełdowych, comiesięczne analizy stopy bezrobocia czy przeprowadzane co kilka lat spisy rolne. **Badania doraźne** mają z kolei charakter unikalny i jednorazowy. Są one uruchamiane ad hoc w celu natychmiastowego zdiagnozowania niespodziewanego problemu decyzyjnego lub rynkowego, np. nagły sondaż opinii publicznej po wprowadzeniu nowej ustawy lub awaryjne badanie satysfakcji klientów po wykryciu globalnej usterki w oprogramowaniu dystrybuowanym przez firmę.

1.4. Techniki doboru próby badawczej

Wybór odpowiedniej techniki doboru próby jest najważniejszym etapem planowania eksperymentu, ponieważ błędy popełnione w tym miejscu są nienaprawialne na etapie analizy matematycznej. Spektakularne porażki badawcze z przeszłości stanowią tego najlepszy dowód. W 1936 roku amerykański magazyn Literary Digest podjął próbę przewidzenia wyników wyborów prezydenckich. Rozesłano imponującą liczbę 10 milionów kart pocztowych do osób wybranych z książek telefonicznych oraz rejestrów posiadaczy samochodów. Mimo uzyskania ogromnej liczby ponad 2 milionów odpowiedzi, magazyn opublikował prognozę zapowiadającą miażdżące zwycięstwo Alfa Landona (57%) nad Franklinem Rooseveltem (43%). Rzeczywisty wynik wyborów okazał się całkowicie odwrotny: Roosevelt zdobył 62% głosów. Przyczyną kompromitacji był fundamentalny błąd metodologiczny – w czasach

Wielkiego Kryzysu telefony i samochody były dobrem luksusowym, przez co próba została drastycznie skrzywiona w stronę ludzi zamożnych o poglądach republikańskich. W tym samym czasie młody statystyk George Gallup, posługując się wielokrotnie mniejszą (zaledwie kilkutysięczną), ale poprawnie dobraną metodą kwotową próbą, bezbłędnie przepowiedział triumf Roosevelta. Lekcja ta udowodniła, że gigantyczna liczebność próby nigdy nie skompensuje braku jej reprezentatywności.

Współczesna metodologia dzieli techniki doboru na nielosowe oraz losowe. W **metodach nielosowych (nieprobabilistycznych)** o włączeniu danej jednostki do próby decyduje subiektywna decyzja badacza, przypadek lub specyficzny cel. Niemożliwe jest tutaj określenie matematycznego prawdopodobieństwa udziału danej jednostki w badaniu, co formalnie uniemożliwia stosowanie zaawansowanych teorii estymacji i testów parametrycznych. Do tych metod zalicza się **dobór przypadkowy** (oparty na czystej dostępności jednostek, np. przepytywanie przechodniów na ulicy), **dobór celowy** (gdzie ekspert intencjonalnie wybiera obiekty uznane przez niego za idealnie typowe dla danego zagadnienia) oraz wspomniany **dobór kwotowy**. Dobór kwotowy opiera się na precyzyjnej wiedzy o strukturze populacji (np. wiemy z danych urzędowych, że w danym mieście żyje 52% kobiet i 48% mężczyzn, z czego 20% to osoby w wieku emerytalnym). Badacz narzuca ankieterom sztywne limity (kwoty) osób o takich cechach, jakie muszą przebadać. Ostatnią popularną metodą nielosową jest **metoda kuli śnieżnej**, wykorzystywana przy badaniu populacji rzadkich, ukrytych lub trudnodostępnych (np. czynni programiści niszowych języków programowania, pacjenci z rzadkimi mutacjami genetycznymi). Proces polega na zlokalizowaniu i przebadaniu pierwszych kilku jednostek, które następnie proszone są o wskazanie i skontaktowanie badacza z kolejnymi osobami spełniającymi kryteria.

Metody losowe (probabilistyczne) gwarantują, że każda pojedyncza jednostka populacji generalnej posiada dokładnie znane, matematycznie definiowalne i większe od zera prawdopodobieństwo znalezienia się w próbie. Podstawą tych metod jest posiadanie kompletnego operatu losowania (czyli dokładnej, aktualnej listy wszystkich jednostek populacji). Najprostszą formą jest **losowanie proste indywidualne**, realizowane za pomocą generatorów liczb losowych (np. w programie Excel) lub tablic losowych. Może mieć ono charakter zależny (losowanie zwrotne, gdzie wylosowany obiekt po zarejestrowaniu wraca do bazy i może zostać wylosowany ponownie) lub niezależny (losowanie bezzwrotne).

W przypadku bardzo dużych populacji stosuje się **losowanie systematyczne**. Polega ono na uszeregowaniu listy, wyznaczeniu interwału losowania k jako ilorazu wielkości populacji N

i planowanej próby n ($k = N/n$), wylosowaniu pierwszej jednostki startowej z przedziału od 1 do k, a następnie automatycznym pobieraniu z listy co k-tej jednostki. Kolejną zaawansowaną metodą jest **losowanie warstwowe**. Stosuje się je, gdy populacja jest wyraźnie niejednorodna, ale można ją podzielić na wewnętrznie spójne podgrupy zwane warstwami (np. podział przedsiębiorstw na mikro, małe, średnie i duże). Losowanie proste przeprowadza się wtedy niezależnie wewnątrz każdej warstwy. Może mieć ono charakter proporcjonalny (zachowujemy identyczne udziały warstw w próbie, jak w populacji) lub optymalny (liczba losowanych jednostek z danej warstwy zależy nie tylko od jej liczebności, ale też od stopnia wewnętrznego zróżnicowania cechy w tej warstwie – im większa zmienność, tym większa próba z danej warstwy). Ostatnią grupą są **losowania zespołowe i wielostopniowe**. W losowaniu zespołowym jednostką losowania nie jest pojedynczy obiekt, ale ich naturalne, terytorialne lub organizacyjne skupisko (tzw. zespół, np. całe klasy w szkole, całe szpitale w kraju). W losowaniu wielostopniowym proces przebiega kaskadowo: najpierw losuje się jednostki stopnia pierwszego (np. województwa), z nich jednostki stopnia drugiego (np. powiaty), a dopiero na samym końcu konkretne obiekty końcowe.

1.5. Skale pomiarowe cech i ich odwzorowanie w programie Statistica

Proces analizy danych wymaga od użytkownika jednoznacznego zdefiniowania skali pomiarowej dla każdej wprowadzanej zmiennej. **Skala pomiarowa** to formalny system reguł przyporządkowywania liczb lub symboli badanym cechom obiektów. Typ skali determinuje zestaw dopuszczalnych operacji matematyczno-statystycznych oraz definiuje rygor założeń dla przyszłych testów istotności. Wyróżniamy cztery klasyczne skale pomiarowe, uporządkowane hierarchicznie od najsłabszej do najsilniejszej.

Pierwszą i najprostszą jest skala **nominalna (jakościowa)**. Służy ona wyłącznie do klasyfikacji, rozróżniania i etykietowania obiektów. Liczby przypisane wariantom cechy nominalnej pełnią jedynie rolę unikalnych symboli identyfikacyjnych i nie posiadają żadnych właściwości numerycznych – nie można ich dodawać, porządkować ani wyciągać z nich średniej. Jedyne dozwolone relacje logiczne na tej skali to stwierdzenie równości lub różności między dwoma obiektami. Typowymi przykładami są: płeć, podział na kraje pochodzenia, grupy krwi, kolory powłok lakierniczych czy numery PESEL. W programie Statistica zmienne te mogą być wprowadzane jako ciągi tekstowe, jednak dla optymalizacji obliczeń najczęściej koduje się je liczbami (np. 1 dla kobiet, 2 dla mężczyzn), nadając im jednocześnie tzw. wartości tekstowe, które automatycznie pojawiają się na wykresach i w raportach.

Drugim poziomem w hierarchii jest skala **porządkowa (rangowa)**. Posiada ona wszystkie właściwości skali nominalnej (pozwala na klasyfikację), ale dodatkowo umożliwia uporządkowanie i uszeregowanie obiektów według stopnia nasilenia badanej cechy. Na tej skali dozwolone są relacje porządku: większy, mniejszy oraz relacje równoważności. Skala porządkowa ma jednak fundamentalne ograniczenie – odległości między kolejnymi punktami skali są nieznane i niemierzalne. Nie można stwierdzić, czy różnica między oceną „bardzo dobra" a „dobra" jest dokładnie taka sama, jak między „dostateczna" a „niedostateczna". Klasycznymi przykładami są skale ocen szkolnych, stopnie wojskowe, klasyfikacja sportowa, poziomy wykształcenia oraz niezwykle popularna w badaniach ankietowych skala Likerta (od „zdecydowanie się nie zgadzam" do „zdecydowanie się zgadzam"). W programie Statistica dane porządkowe są kluczowe dla nieparametrycznych testów statystycznych.

Trzeci stopień to **skala przedziałowa (interwałowa)**. Posiada ona cechy skali porządkowej, a dodatkowo charakteryzuje się tym, że odległości (interwały) między jednostkami skali są stałe, precyzyjne i mierzalne, ponieważ wyraża się je w zbiorze liczb rzeczywistych. Na tej skali w pełni uprawnione są operacje dodawania i odejmowania wartości. Cechą skali przedziałowej jest obecność tzw. zera umownego (zera relatywnego). Zero na tej skali nie oznacza absolutnego braku badanej cechy, lecz jest jedynie punktem przyjętym przez twórcę danej skali. Przykładem jest pomiar temperatury w stopniach Celsjusza lub Fahrenheita – wartość 0 °C to umowna temperatura zamarzania wody, a nie całkowity brak temperatury (ponieważ istnieją temperatury ujemne). Konsekwencją istnienia zera umownego jest zakaz obliczania ilorazów: nie wolno stwierdzić, że dzień, w którym odnotowano 40 °C, był „dwa razy cieplejszy" od dnia z temperaturą 20 °C.

Najwyższym poziomem jest **skala ilorazowa (stosunkowa)**. Reprezentuje ona najwyższy stopień rygoru matematycznego, posiadając stałe interwały oraz, co najważniejsze, zero absolutne (bezwzględne). Zero absolutne na skali ilorazowej oznacza całkowity, fizyczny brak badanej właściwości. Poniżej zera nie istnieją żadne wartości. Dzięki temu na tej skali można wykonywać wszystkie operacje matematyczne, włącznie z mnożeniem i dzieleniem. Uprawnione jest tu stwierdzenie, że dany detal waży „trzy razy więcej" lub maszyna pracowała „o połowę krócej". Przykładami są: wiek, masa ciała, wzrost, dystans, przychody finansowe, czas trwania operacji procesora czy liczba wyprodukowanych braków. W środowisku Statistica zmienne pochodzące ze skali przedziałowej i ilorazowej są traktowane wspólnie jako zmienne ciągłe (ilościowe). W strukturze programu stanowią one rdzeń analiz ilościowych, podczas gdy

zmienne nominalne i porządkowe są wykorzystywane jako czynniki grupujące lub zmienne niezależne jakościowe.

Przykład

Dane wykorzystane w projekcie pochodzą z wewnętrznego audytu operacyjno-finansowego oraz analizy efektywności pracy, które zrealizowano w dużej firmie z sektora usług biznesowych. Zarząd przedsiębiorstwa podjął próbę zdiagnozowania, jakie czynniki najsilniej decydują o produktywności zespołów w realiach pracy stacjonarnej i zdalnej. Głównym celem analizy było sprawdzenie czy programy rozwojowe oraz realne obciążenie pracą i doświadczenie mają bezpośredni, mierzalny wpływ na wyniki biznesowe organizacji.

W badaniu wzięło udział $N = 120$ pracowników zatrudnionych na etatach w roku 2026. Analizowaną zbiorowością statystyczną (populacją) byli zatem wszyscy pracownicy operacyjnych, natomiast jednostką statystyczną był pojedynczy pracownik, u którego mierzono wybrane właściwości. Aby poprawnie przygotować te informacje do przetworzenia w programie *Statistica*, każdą mierzoną cechę statystyczną przypisano do odpowiedniej skali pomiarowej.

Pierwszą grupę parametrów tworzą zmienne ilościowe, czyli wyrażone bezpośrednio w liczbach. Naszą zmienną objaśnianą (zależną), którą chcemy modelować, jest *liczba zrealizowanych projektów (wydajność)*. Reprezentuje ona skalę stosunkową (ilorazową), ponieważ posiada naturalny punkt zero (wartość 0 oznacza całkowity brak zrealizowanych operacji). Kolejną cechą liczbową jest *roczna liczba godzin szkoleń* (X1). Została ona również osadzona na skali stosunkowej (ilorazowej), ponieważ posiada naturalny punkt zero, oznaczający sytuację, w której pracownik nie uczestniczył w żadnym kursie rozwojowym. Ostatnią zmienną w tej grupie jest *staż pracy w danej organizacji (wyrażony w pełnych latach)* (X2). To także cecha mierzona na skali stosunkowej, gdzie wartości reprezentują upływ czasu.

Drugą grupę stanowią zmienne jakościowe, których głównym celem jest podział badanych ludzi na mniejsze podgrupy. *Format pracy* to dychotomiczna cecha na skali nominalnej, dzieląca próbę na pracowników zdalnych oraz stacjonarnych. Drugą cechą nominalną jest *dział*, w którym dana osoba wykonuje obowiązki (HR, IT lub Sprzedaż). Całość schematu uzupełniają dwa pytania kwestionariuszowe, w których pracownicy oceniali poziom swojej *autonomii* oraz *stresu* na pięciostopniowej skali Likerta (od 1 – zdecydowanie się nie zgadzam, do 5 – zdecydowanie się zgadzam). Cechy te reprezentują skalę porządkową (rangową).

2. Przygotowanie, prezentacja i opisowa analiza danych strukturalnych

2.1. Grupowanie danych i tworzenie szeregów statystycznych

Pierwszym fundamentalnym krokiem po zebraniu surowego materiału empirycznego i poprawnym wprowadzeniu go do bazy danych programu Statistica jest jego uporządkowanie. Surowe dane, występujące w postaci pierwotnych, chaotycznych ciągów liczbowych, są całkowicie nieczytelne dla analityka i uniemożliwiają interpretację prawidłowości. Proces porządkowania polega na transformacji tych danych w tzw. **szereg statystyczny**, czyli logicznie uporządkowany według określonych kryteriów zbiór obserwacji. Najprostszą i najbardziej podstawową formą prezentacji jest **szereg szczegółowy**, w którym wszystkie zaobserwowane wartości badanej cechy statystycznej zostają uszeregowane w sposób rosnący, co formalnie zapisujemy jako ciąg relacji $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$, gdzie indeks dolny oznacza pozycję danej obserwacji po sortowaniu, a $n$stanowi całkowitą liczebność próby. Taki układ pozwala badaczowi na natychmiastowe zidentyfikowanie wartości skrajnych, czyli minimum oraz maksimum, a także ogólnego obszaru zmienności analizowanego zjawiska.

W sytuacji, gdy liczba zebranych obserwacji n jest duża, szereg szczegółowy staje się zbyt rozbudowany i przestaje spełniać funkcję syntetyzującą. Wówczas w metodologii statystycznej stosuje się procedurę grupowania, która prowadzi bezpośrednio do utworzenia szeregu rozdzielczego. **Szereg rozdzielczy** definiuje się jako zbiorowość statystyczną podzieloną na relatywnie jednorodne klasy lub warianty według badanej cechy jakościowej bądź ilościowej, z jednoczesnym przyporządkowaniem liczebności, oznaczanej jako n_i, lub częstości, oznaczanej jako ω_i i definiowanej jako iloraz liczebności danej klasy do liczebności całkowitej, czyli $\omega_i = n_i/n$. W zależności od natury badanej zmiennej wyróżnia się dwa zasadnicze typy szeregów rozdzielczych. Pierwszym z nich są **szeregi rozdzielcze punktowe**, które konstruuje się dla cech skokowych, zwanych dyskretnymi, przyjmujących niewielką liczbę unikalnych wariantów, takich jak liczba błędów w oprogramowaniu czy liczba dzieci w gospodarstwie domowym. Drugim typem są **szeregi rozdzielcze przedziałowe**, zwane klasowymi, które stają się niezbędne w przypadku cech o charakterze ciągłym, takich jak czas, wzrost czy przychód, bądź cech skokowych o bardzo szerokim spektrum unikalnych wartości. W procesie budowy szeregu przedziałowego tworzy się tzw. **przedziały klasowe** (interwały), definiując dla każdego z nich dolną granicę, oznaczaną jako x_{0i}, oraz górną granicę, oznaczaną jako x_{1i}. Kluczowym parametrem takiego przedziału jest jego **rozpiętość**, czyli szerokość klasy

oznaczana jako h_i, którą oblicza się poprzez odjęcie granicy dolnej od granicy górnej, co przyjmuje postać prostego równania $h_i = x_{1i} - x_{0i}$.

Dodatkowo, każdy z tych szeregów może być prezentowany w formie prostej, która podaje surową liczebność n_i przypadających na dany wariant, bądź w formie skumulowanej. Liczebność skumulowaną dla i-tego przedziału, oznaczaną jako n_{ski}, oblicza się jako sumę liczebności danej klasy oraz wszystkich klas poprzedzających, co matematycznie wyraża się wzorem $n_{ski} = \sum_{j=1}^{i} n_j$. Odpowiednio, częstość skumulowana, oznaczana symbolem ω_{ski}, określa sumaryczny udział jednostek posiadających wartość cechy mniejszą lub równą górnej granicy danego przedziału klasowego, a wyznacza się ją poprzez zsumowanie kolejnych częstości, co zapisujemy jako $\omega_{ski} = \sum_{j=1}^{i} \omega_j$, bądź poprzez bezpośredni stosunek liczebności skumulowanej do liczebności próby, czyli $\omega_{ski} = n_{ski}/n$. Zastosowanie formy skumulowanej pozwala analitykowi na błyskawiczne określenie, jaki procent badanej populacji nie przekracza z góry założonego progu wartości.

2.2. Wizualizacja danych w programie Statistica

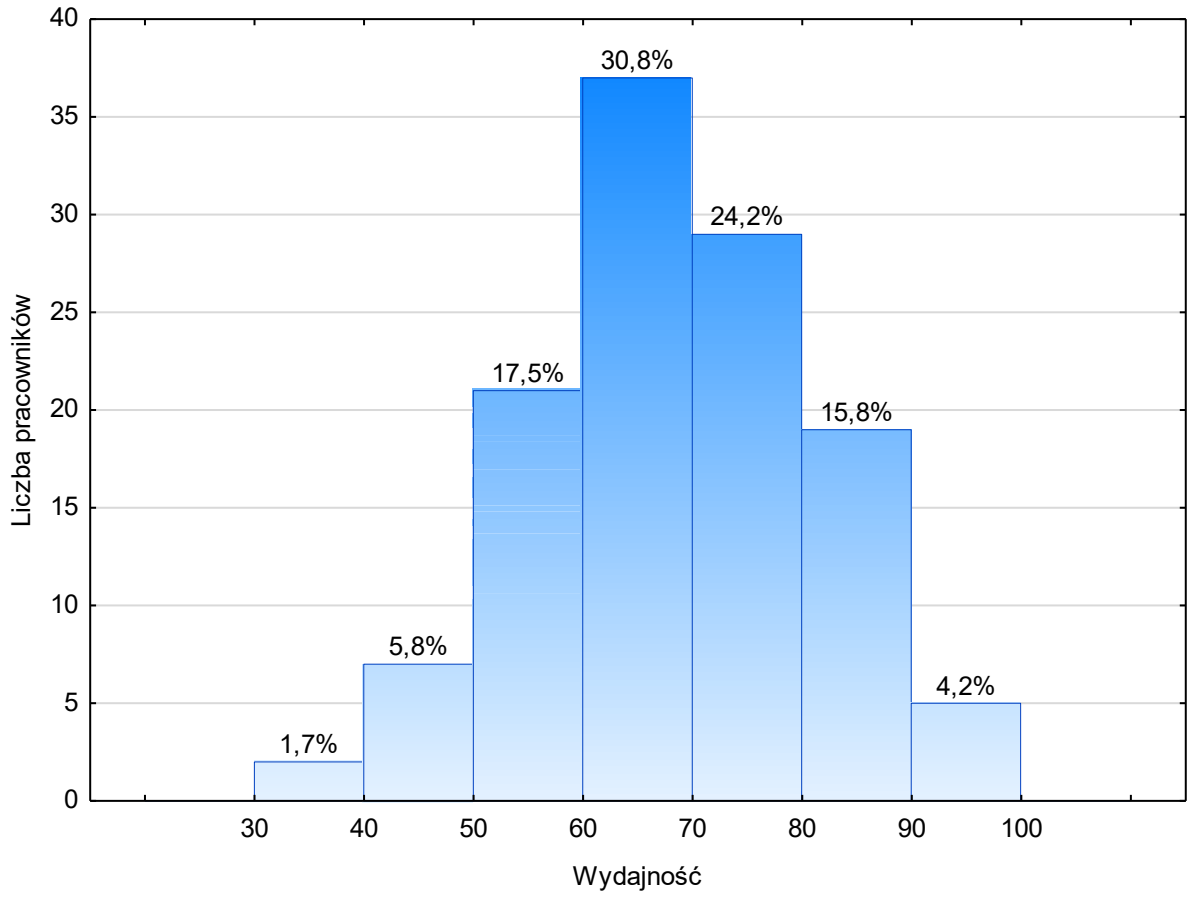
Graficzna prezentacja wyników stanowi kluczowy i nadrzędny element eksploracyjnej analizy danych, ponieważ pozwala ludzkiemu umysłowi na natychmiastowe i intuicyjne wykrycie ogólnego kształtu rozkładu empirycznego, a także na szybkie zlokalizowanie potencjalnych anomalii, błędów grubych czy obserwacji odstających. Program Statistica oferuje system generowania grafiki statystycznej, która dopasowuje się do charakteru badanych zmiennych oraz zastosowanych skali pomiarowych. Dla cech o charakterze jakościowym, a więc mierzonych na skali nominalnej bądź porządkowej, najwłaściwszym narzędziem wizualizacji są wykresy słupkowe oraz wykresy kołowe. Na wykresie kołowym wielkość powierzchni poszczególnych wycinków koła odzwierciedla strukturę próby, reprezentując udziały procentowe lub surowe liczebności konkretnych kategorii w odniesieniu do całości zbiorowości.

W przypadku zmiennych ilościowych o charakterze ciągłym, które zostały pogrupowane w strukturę przedziałową, fundamentalnym i niezastąpionym narzędziem jest histogram częstości. Wykres ten konstruuje się w prostokątnym układzie współrzędnych, gdzie na osi odciętych, oznaczanej jako X, zaznacza się precyzyjnie granice poszczególnych przedziałów klasowych, natomiast na osi rzędnych, oznaczanej jako Y, nanosi się surowe liczebności n_i lub gęstość liczebności, definiowaną jako stosunek n_i/h_i, co staje się wymogiem w sytuacji, gdy szereg rozdzielczy składa się z przedziałów o nierównej szerokości. Nad każdym przedziałem

klasowym wznosi się prostokątny słupek, którego pole powierzchni jest ściśle proporcjonalne do liczebności wpadającej do danego interwału. Środowisko programu Statistica pozwala na automatyczne generowanie histogramów z jednoczesnym nałożeniem na wykres ciągłej, czerwonej linii teoretycznej gęstości rozkładu normalnego, co daje badaczowi narzędzie do wstępnej, wzrokowej oceny stopnia zgodności danych empirycznych z modelem teoretycznym Gaussa. W zaawansowanych analizach makroekonomicznych i geograficznych wykorzystuje się również kartogramy, czyli mapy konturowe, na które za pomocą odpowiednio dobranej skali barwnej lub gęstości kreskowania nanosi się przestrzenne natężenie badanego zjawiska w poszczególnych regionach.

Przykład

W celu oceny zebranego materiału empirycznego przeprowadzono analizę kształtu rozkładu wydajności pracowników (rys. 1).



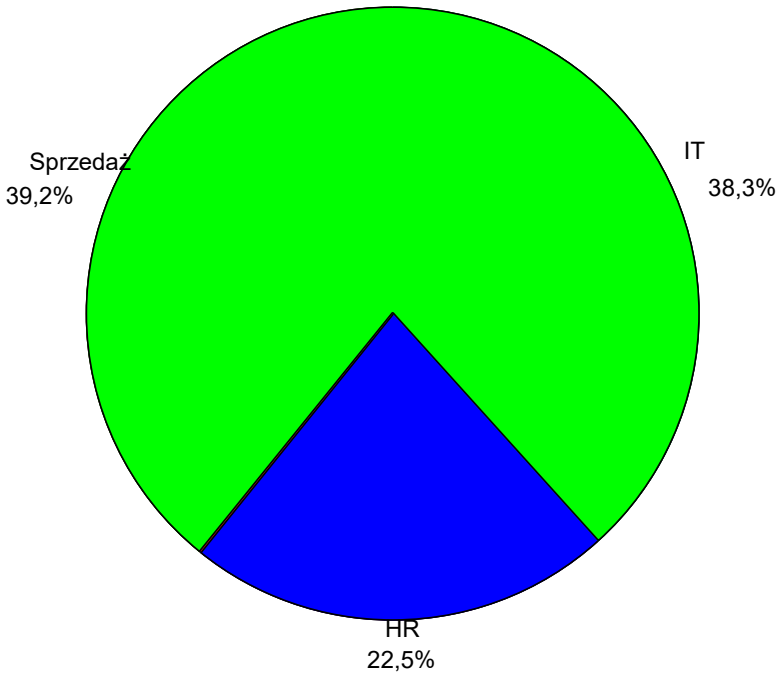
Rys. 1. Histogram rozkładu wydajności pracowników

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogramy, wybranie zmiennej, a następnie w zakładce *Etykiety punktów* zaznaczenie wyświetlania wartości jako procenty oraz włączenie podziału na równe przedziały klasowe o szerokości równej np. 10.

Na podstawie histogramu wydajności (rys. 1) można zauważyć, że badana cecha (wydajność) ma rozkład jednomodalny i kształt zbliżony do rozkładu symetrycznego. Największa część próby, bo aż 30,8% (co odpowiada dokładnie 37 pracownikom), osiąga wydajność w środkowym przedziale między 60 a 70 zrealizowanych w ciągu miesiąca projektów. W sąsiednich przedziałach również notuje się duże skupienie osób: wydajność na poziomie 50–60 projektów osiąga 21 osób (17,5%), natomiast 70-80 projektów na miesiąc realizuje 29 pracowników (24,2%). Skrajne wartości wydajności w tej zbiorowości występują sporadycznie. W przedziale najniższych wyników, czyli między 30 a 40 zrealizowanych projektów, znajduje się zaledwie 2 pracowników (1,7%). Z kolei najwyższą produktywnością, mieszczącą się w granicach 90-100 projektów na miesiąc, wykazuje się tylko 5 osób (4,2%).

W celu scharakteryzowania struktury zatrudnienia w badanej próbie na Rys. 2. przedstawiono procentowy udział pracowników w podziale na trzy główne piony organizacyjne firmy.

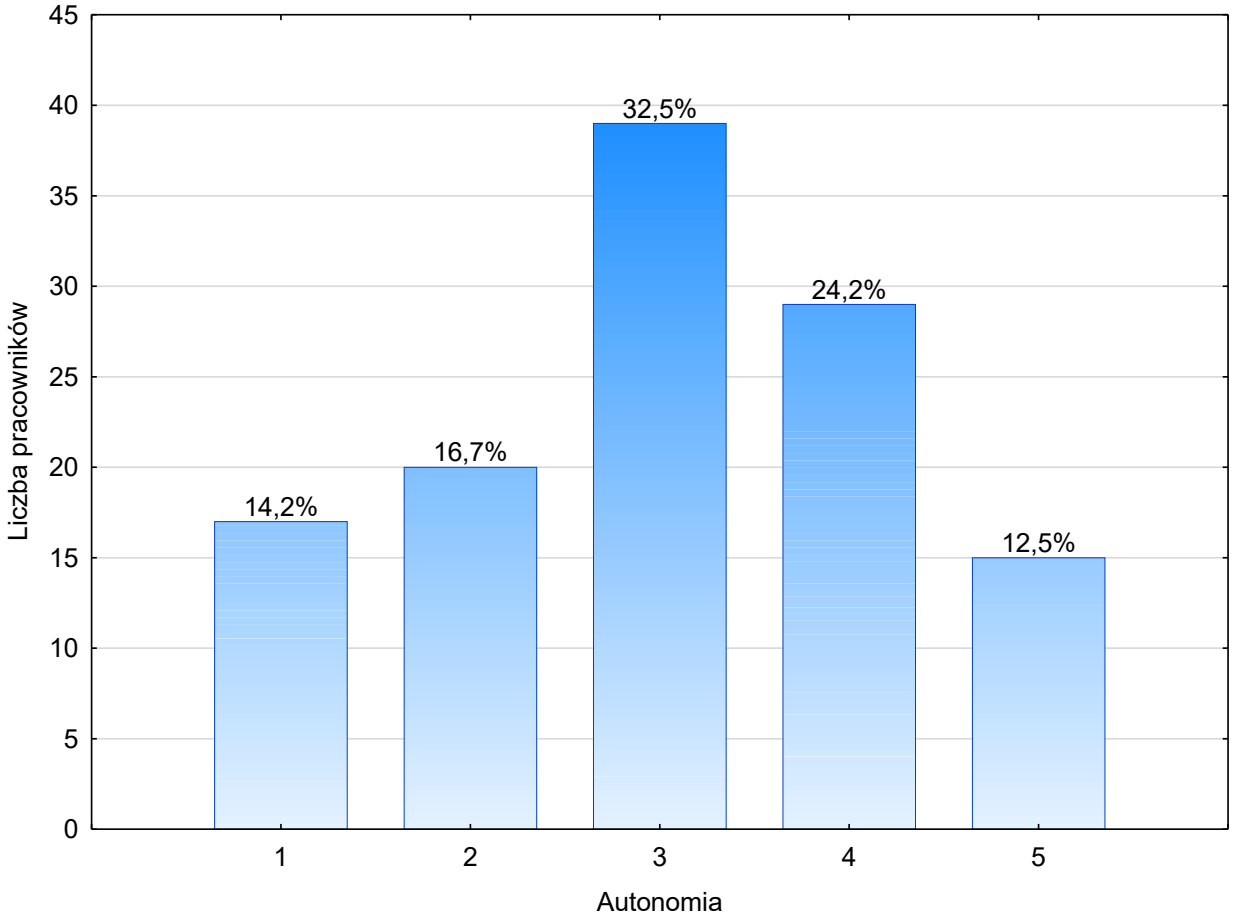


Rys. 2. Struktura zatrudnienia badanych pracowników z podziałem na działy

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy 2W → Wykresy kołowe wybranie zmiennej, a następnie w zakładce *Etykiety punktów* zaznaczenie wyświetlania nazwy kategorii wraz z odpowiadającymi im wartościami procentowymi.

W celu zbadania subiektywnych odczuć pracowników dotyczących środowiska pracy, na Rys. 3. przedstawiono strukturę odpowiedzi na pytanie dotyczące poziom poczucia autonomii w miejscu pracy analizowanej próby pracowników.



Rys. 3. Struktura odpowiedzi pracowników na pytanie o poziom autonomii w pracy

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogramy, wybranie zmiennej, a następnie w zakładce *Etykiety punktów* zaznaczenie wyświetlania wartości jako procenty oraz w zakładce *Histogram* ustawienie szerokości słupka na np. 70 %.

Analiza struktury zaprezentowanej na Rys. 3 pozwala na ocenę nastrojów pracowników w kontekście posiadanej swobody decyzyjnej. Największa część respondentów, bo aż 32,5% (co odpowiada dokładnie 39 pracownikom), wybrała wariant 3, oznaczający postawę neutralną. Pozytywną ocenę swojego poziomu niezależności (wariant 4 – „zgadzam się") zadeklarowało 29 osób (24,2%), natomiast pełną swobodę decyzyjną (wariant 5 – „zdecydowanie się zgadzam") odczuwa 15 pracowników (12,5%). Łącznie opinie o charakterze pozytywnym reprezentuje zatem 36,7% badanej zbiorowości (44 osoby).

Po przeciwnej stronie skali plasują się pracownicy odczuwający deficyt autonomii. Niski poziom swobody w działaniu (wariant 2 – „nie zgadzam się") wskazało 20 osób (16,7%),

z kolei skrajnie negatywną ocenę (wariant 1 – „zdecydowanie się nie zgadzam") zadeklarowało 17 pracowników (14,2%). Sumarycznie postawy negatywne obejmują 30,9% całej próby (37 osób). Rozkład odpowiedzi zbliżony jest do rozkładu symetrycznego z dominantą ulokowaną w punkcie neutralnym.

2.3. Miary tendencji centralnej (położenia)

Opisowa analiza statystyczna zmierza do syntetycznego scharakteryzowania całej badanej zbiorowości za pomocą pojedynczych, łatwych w interpretacji wskaźników liczbowych. Pierwszą i najbardziej intuicyjną grupą takich wskaźników są miary tendencji centralnej, zwane potocznie miarami położenia, które informują badacza o tym, wokół jakiej centralnej wartości skupia się największa część analizowanych obserwacji. W metodologii statystycznej miary te dzieli się na dwie zasadnicze rodziny, mianowicie miary klasyczne, które są obliczane na podstawie wartości wszystkich jednostek wchodzących w skład próby, oraz miary pozycyjne, których wartość nie zależy od wszystkich obserwacji bezpośrednio, lecz od pozycji, jaką konkretna jednostka zajmuje w uporządkowanym szeregu statystycznym.

Głównym i najpowszechniej stosowanym reprezentantem miar klasycznych jest **średnia arytmetyczna**, oznaczana symbolem $\overline{x}$. W zależności od sposobu przygotowania i uporządkowania danych, oblicza się ją na dwa sposoby. Dla szeregu szczegółowego stosuje się wzór na średnią prostą, który polega na zsumowaniu wszystkich wartości i podzieleniu wyniku przez ich liczbę, co zapisujemy jako:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

W sytuacji, gdy dysponujemy danymi pogrupowanymi w szereg rozdzielczy przedziałowy, konieczne jest zastosowanie wzoru na średnią ważoną, gdzie wagami stają się liczebności poszczególnych klas n_i, a jako reprezentanta każdej klasy przyjmuje się jej środek oznaczany jako $\dot{x}_i$, obliczany jako średnia z dolnej i górnej granicy interwału, czyli $\dot{x}_i = (x_{0i} + x_{1i})/2$. Pełny wzór na średnią ważoną dla k przedziałów przyjmuje wówczas postać:

$$\overline{x} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot n_i}{n}$$

Średnia arytmetyczna charakteryzuje się jednak bardzo wysoką wrażliwością na obecność wartości skrajnych i odstających w próbie, co oznacza, że jedna nietypowo wysoka wartość może zawyżyć wynik, dając fałszywy obraz rzeczywistości. Oprócz niej w analizach specyficznych wyróżnia się **średnią harmoniczną**, oznaczaną jako $\overline{x}_h$, stosowaną przy badaniu

cech będących stosunkiem innych wielkości, jak prędkość czy wydajność, którą dla szeregu szczegółowego oblicza się jako iloraz liczebności próby do sumy odwrotności poszczególnych wartości, zgodnie z formułą:

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Kolejną miarą klasyczną jest **średnia geometryczna**, oznaczana jako $\overline{x}_g$, która stanowi pierwiastek n-tego stopnia z iloczynu wszystkich obserwowanych wartości, co zapisujemy matematycznie jako:

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Średnia geometryczna jest narzędziem niezastąpionym w analizie dynamiki zjawisk, gdzie służy do wyliczania średniego tempa zmian procesów w czasie.

Do grupy miar pozycyjnych zalicza się przede wszystkim medianę oraz dominantę. **Mediana**, oznaczana symbolem Me i nazywana wartością środkową, dzieli uporządkowany szereg szczegółowy dokładnie na dwie równe części, co oznacza, że połowa jednostek przyjmuje wartości mniejsze lub równe medianie, a druga połowa wartości większe bądź równe. Dla szeregu szczegółowego jej wyznaczenie zależy od parzystości liczby obserwacji i jest opisywane funkcją warunkową:

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{dla } n \text{ nieparzystych} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{dla } n \text{ parzystych} \end{cases}$$

Gdy dane są zamknięte w szeregu przedziałowym, konkretną wartość mediany szacuje się za pomocą matematycznego wzoru interpolacyjnego, który przyjmuje następującą formę:

$$Me = x_{0m} + \frac{\frac{n}{2} - n_{sk,m-1}}{n_m} \cdot h_m$$

gdzie symbol x_{0m} oznacza dolną granicę przedziału zawierającego medianę (czyli pierwszego przedziału, w którym liczebność skumulowana przekracza lub równa się wartości $n/2$), symbol $n_{sk,m-1}$ to liczebność skumulowana przedziału bezpośrednio poprzedzającego przedział mediany, n_m to surowa liczebność przedziału mediany, natomiast h_m stanowi jego szerokość. Mediana, w przeciwieństwie do średniej arytmetycznej, jest miarą odporną na wpływ obserwacji skrajnych.

Uzupełnieniem rodziny miar pozycyjnych, obok mediany, są parametry dzielące zbiorowość na cztery równe części, czyli kwartyle: **kwartyl pierwszy**, oznaczany symbolem Q_1 i nazywany kwartylem dolnym, oraz kwartyl trzeci, oznaczany jako Q_3 i określany mianem kwartyla górnego (mediana stanowi formalnie kwartyl drugi Q_2). Kwartyl dolny odcina w uporządkowanym szeregu 25% jednostek o najniższych wartościach cechy, co oznacza, że jedna czwarta zbiorowości przyjmuje wartości mniejsze lub równe Q_1, natomiast 75% jednostek przyjmuje wartości od niego większe bądź równe. Odwrotnie interpretuje się **kwartyl górny**, który odcina 75% obserwacji, co wskazuje, że trzy czwarte badanych obiektów charakteryzuje się wartościami cechy mniejszymi lub równymi Q_3, a pozostałe 25% przyjmuje wartości wyższe.

W algorytmach programu Statistica, w przypadku surowego szeregu szczegółowego, pozycje tych parametrów wyznaczane są na podstawie rangi percentylowej (odpowiednio jako 25. i 75. percentyl). Jeżeli natomiast dane zostaną pogrupowane w szereg rozdzielczy przedziałowy, do precyzyjnego oszacowania wartości obu kwartyli stosuje się wzory interpolacyjne, które wykazują analogię strukturalną do algorytmu wyznaczania mediany. Dla **kwartyla dolnego** Q_1 wzór ten przyjmuje postać:

$$Q_1 = x_{0Q1} + \frac{\frac{n}{4} - n_{sk,Q1-1}}{n_{Q1}} \cdot h_{Q1}$$

gdzie symbol x_{0Q1} oznacza dolną granicę przedziału zawierającego kwartyl pierwszy (identyfikowanego jako pierwszy przedział, w którym liczebność skumulowana osiąga lub przekracza wartość $n/4$), symbol $n_{sk,Q1-1}$ reprezentuje liczebność skumulowaną klasy poprzedzającej przedział kwartylny, n_{Q1} to surowa liczebność przedziału kwartyla pierwszego, natomiast h_{Q1} odzwierciedla jego szerokość.

Dla **kwartyla górnego** Q_3, ze względu na odcinanie trzech czwartych zbiorowości, kluczowy człon pozycji przyjmuje wartość $3n/4$, co przekłada się na następującą formułę interpolacyjną:

$$Q_3 = x_{0Q3} + \frac{\frac{3n}{4} - n_{sk,Q3-1}}{n_{Q3}} \cdot h_{Q3}$$

w której x_{0Q3} definiuje dolną granicę przedziału zawierającego kwartyl trzeci (czyli pierwszą klasę, dla której liczebność skumulowana jest większa bądź równa $3n/4$), $n_{sk,Q3-1}$ oznacza liczebność skumulowaną przedziału poprzedzającego, n_{Q3} to liczebność samego przedziału kwartylnego, a h_{Q3} stanowi rozpiętość tego interwału.

Uzupełnieniem miar pozycyjnych jest **dominanta**, oznaczana jako D lub Mo i nazywana modą, która wskazuje wartość najczęściej występującą w zbiorowości. W szeregu przedziałowym najpierw lokalizuje się przedział o największej liczebności, zwany przedziałem modalnym, a następnie precyzyjną wartość dominanty aproksymuje się wzorem interpolacyjnym opartym na różnicach liczebności sąsiednich klas:

$$D = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot h_d$$

w którym x_{0d} definiuje dolną granicę przedziału modalnego, n_d to jego liczebność, n_{d-1} oznacza liczebność klasy poprzedzającej, n_{d+1} to liczebność klasy następującej po przedziale modalnym, a h_d odzwierciedla szerokość tego interwału.

2.4. Miary zróżnicowania (dyspersji i zmienności)

Sama znajomość miar tendencji centralnej jest dalece niewystarczająca do rzetelnego i poprawnego opisania zbiorowości statystycznej. Dwie niezależne grupy obiektów mogą charakteryzować się identyczną średnią arytmetyczną, jednak struktura wewnętrzna tych grup może być skrajnie odmienna – jedna może być wysoce jednorodna, gdzie wszystkie jednostki skupiają się blisko średniej, druga zaś skrajnie zróżnicowana. Do identyfikacji i ilościowego opisu tego zjawiska służą miary zróżnicowania, zwane miarami dyspersji lub rozproszenia, które określają, jak silnie poszczególne wartości odchylają się od poziomu centralnego.

Podstawową i wyjściową klasyczną miarą dyspersji jest **wariancja**, oznaczana jako s^2, która stanowi średnią arytmetyczną kwadratów odchyleń poszczególnych wartości cechy od ich ogólnej średniej arytmetycznej. Dla szeregu szczegółowego wariancję oblicza się za pomocą wzoru:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Natomiast dla szeregu rozdzielczego przedziałowego wariancja przyjmuje postać ważoną, w której odchylenia środków przedziałów $\dot{x}_i$ od średniej są mnożone przez liczebności klas n_i, co zapisujemy jako:

$$s^2 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 \cdot n_i}{n}$$

Uwaga:

W procesie wnioskowania statystycznego realizowanego na małych próbach, w celu uzyskania estymatora nieobciążonego, w mianowniku powyższych wzorów zamiast wartości n stosuje się poprawkę na stopnie swobody w postaci $n - 1$.

Ponieważ wariancja wyrażona jest w jednostkach podniesionych do kwadratu, co uniemożliwia bezpośrednią i intuicyjną interpretację, w praktyce badawczej znacznie częściej wyznacza się **odchylenie standardowe**, oznaczane symbolem s, które jest pierwiastkiem kwadratowym z wariancji, czyli $s = \sqrt{s^2}$. Odchylenie standardowe informuje analityka, o ile przeciętnie poszczególne jednostki w próbie różnią się (odchylają in plus lub in minus) od wyliczonej średniej arytmetycznej. Aby dokonać oceny siły rozproszenia w sposób całkowicie niezależny od miana badanej cechy i skali danych, oblicza się **klasyczny współczynnik zmienności**, oznaczany jako V_s, będący procentowym stosunkiem odchylenia standardowego do średniej arytmetycznej, co wyraża się wzorem:

$$V_s = \frac{s}{\overline{x}} \cdot 100\%$$

Końcowym etapem analizy dyspersji jest wyznaczenie tzw. typowego obszaru zmienności, czyli przedziału skupiającego wokół środka najbardziej reprezentatywne, typowe jednostki. W ujęciu klasycznym typowy obszar zmienności określa przedział domknięty będący wynikiem działania $\langle \overline{x} - s \ ; \ \overline{x} + s \rangle$.

2.5. Analiza asymetrii i skośności rozkładu

Ostatnim, trzecim fundamentalnym filarem statystyki opisowej, obok badania położenia i dyspersji, jest szczegółowa ocena kształtu rozkładu empirycznego, ze szczególnym uwzględnieniem jego symetrii bądź asymetrii. Rozkład empiryczny uznaje się za idealnie **symetryczny**, jeżeli jednostki położone po obu stronach miary centralnej są rozmieszczone w dokładnie taki sam sposób, co w rzeczywistości matematycznej objawia się równością trzech głównych miar, czyli zachodzi relacja $\overline{x} = Me = D$. W praktycznych badaniach rynkowych idealna symetria występuje niezmiernie rzadko, a rozkłady charakteryzują się skośnością, czyli asymetrią, wywołaną obecnością nielicznych, ale bardzo skrajnych obserwacji po jednej ze stron rozkładu.

Wyróżnia się dwa zasadnicze kierunki tego zjawiska. Pierwszym jest **asymetria prawostronna**, zwana dodatnią, która charakteryzuje się tym, że zdecydowana większość badanych obiektów skupia się wokół wartości niskich (poniżej średniej), natomiast ramię rozkładu wyciąga się daleko w stronę wartości wysokich. W rozkładzie o asymetrii dodatniej miary układają się zazwyczaj w relację nierówności $D < Me < \overline{x}$. Klasycznym przykładem jest rozkład wynagrodzeń w przedsiębiorstwach, gdzie większość pracowników zarabia relatywnie niewiele, a wąska grupa kadry zarządzającej otrzymuje pensje bardzo wysokie, co

silnie i sztucznie zawyża średnią arytmetyczną. Drugim kierunkiem jest **asymetria lewostronna**, zwana ujemną, gdzie większość jednostek przyjmuje wartości wysokie, a ramię rozkładu rozciąga się w stronę wartości niskich, co generuje zazwyczaj relację odwrotną, czyli $\overline{x} < Me < D$. Przykładem może być wiek kobiet rodzących dzieci we współczesnym społeczeństwie bądź oceny z egzaminu, który okazał się bardzo prosty dla większości studentów.

Do precyzyjnego, ilościowego pomiaru siły i kierunku skośności w programie Statistica stosuje się **klasyczny współczynnik skośności**, oznaczany jako A_s, oparty na trzecim momencie centralnym rozkładu i definiowany skomplikowanym ułamkiem:

$$A_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$

Interpretacja współczynnika skośności jest następująca: wartość równa dokładnie zero oznacza rozkład idealnie symetryczny, wartości większe od zera wskazują na asymetrię prawostronną, natomiast wartości mniejsze od zera informują o wystąpieniu asymetrii lewostronnej. Im większa jest bezwzględna wartość danego współczynnika, tym silniejsze i bardziej skrajne jest zniekształcenie kształtu badanego rozkładu empirycznego.

2.6. Analiza spłaszczenia i koncentracji rozkładu

Dopełnieniem kompleksowej analizy struktury rozkładu empirycznego, po scharakteryzowaniu jego położenia, dyspersji oraz asymetrii, jest ocena stopnia koncentracji oraz spłaszczenia obserwacji wokół średniej arytmetycznej. Do pomiaru tego zjawiska służy **kurtoza** (zwana miarą skupienia), która pozwala badaczowi określić, jak intensywnie wartości empiryczne koncentrują się w centrum rozkładu w porównaniu do teoretycznego, wzorcowego rozkładu normalnego (krzywej Gaussa).

Punktem odniesienia podczas wyznaczania kurtozy jest czwarty moment centralny (μ_4) oraz odchylenie standardowe podniesionym do potęgi czwartej (s^4), co dla szeregu szczegółowego wyraża się następującym wzorem ciągłym:

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4}$$

gdzie symbol n oznacza liczebność próby, x_i to poszczególne wartości zmiennej, $\overline{x}$ definiuje średnią arytmetyczną, natomiast s odzwierciedla odchylenie standardowe z próby.

Należy jednak pamiętać, że program Statistica w oknie podsumowania statystyk opisowych raportuje tę miarę pod nazwą „Kurtoza”, stosując w rzeczywistości tzw. eksces

(współczynnik przebiegu). Algorytm programu automatycznie odejmuje od klasycznej kurtozy wartość 3 (gdyż dla idealnego rozkładu normalnego kurtoza wynosi dokładnie 3).

Interpretacja wartości liczbowej kurtozy (ekscesu) generowanej przez program Statistica pozwala na sklasyfikowanie badanego rozkładu empirycznego pod kątem jego smukłości do jednej z trzech grup:

- **rozkład mezokurtyczny (normalny) –** występuje wtedy, gdy uzyskany z programu wynik jest równy lub ekstremalnie bliski zeru, co oznacza, że koncentracja jednostek wokół średniej jest zbliżona jak w teoretycznym rozkładzie normalnym.
- **rozkład leptokurtyczny (smukły) –** ma miejsce wówczas, gdy wartość raportowana przez program jest dodatnia. W takim układzie obserwuje się koncentrację wartości cechy wokół średniej arytmetycznej (wykres jest „strzelisty"), co w praktyce oznacza, że większość badanych obiektów przyjmuje wartości bardzo zbliżone do średniej, a na krańcach rozkładu mogą pojawiać się nieliczne, ale skrajnie oddalone obserwacje (tzw. grube ogony).
- **rozkład platokurtyczny (spłaszczony) –** definiuje sytuację, w której kurtoza przyjmuje wartości ujemne, co świadczy o słabej koncentracji wokół środka rozkładu, co oznacza, że wartości empiryczne są bardziej równomiernie „rozlane" po całym obszarze zmienności, a sam wykres gęstości jest relatywnie płaski.

W codziennej praktyce analitycznej w programie Statistica kurtoza i asymetria odgrywa rolę podczas wstępnej weryfikacji założeń przed uruchomieniem testów parametrycznych. Wartości kurtozy i współczynnika asymetrii znacząco odbiegające od zera stanowią dla badacza wyraźny sygnał o naruszeniu założenia o normalności rozkładu, co zmusza do zastosowania metod transformacji danych bądź przejścia do metodologii nieparametrycznej.

Przykład

W celu dokonania wszechstronnej oceny parametrów ilościowych charakteryzujących badaną zbiorowość, w Tabeli 1. zestawiono podstawowe statystyki opisowe dla trzech kluczowych zmiennych metrycznych: wydajności, liczby godzin szkoleń oraz stażu pracy.

Tabela 1. Zbiorcze zestawienie statystyk opisowych dla zmiennych ilościowych

Zmienna	N ważnych	Średnia	Mediana	Minimum	Maksimum	Dolny kwartyl	Górny kwartyl	Rozstęp	Odch.std	Wsp.zmn.	Skośność	Kurtoza
Wydajność	120	68,97	69	39	100	60,5	78,5	61	13,14	19,05	0,031	-0,100

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Statystyki podstawowe → Statystyki opisowe, wybranie zmiennej, a następnie w zakładce *Więcej* zaznaczenie podanych miar.

Analiza tabeli 1 wskazuje, że poddano analizie 120 pracowników zatrudnionych na etatach w pewnym przedsiębiorstwie z sektora usług biznesowych. Pracowników tych analizowano pod kątem wydajności mierzonej liczbą zrealizowanych miesięcznie projektów.

Najniższa wydajność wśród pracowników badanych przez przedsiębiorstwo wynosiła 39 zrealizowanych w miesiącu projektów, zaś najwyższa 100 projektów, co daje różnicę pomiędzy najbardziej i najmniej wydajnym pracownikiem wynoszącą 61 projektów. Przynajmniej 25% poddanych analizie pracowników miało wydajność nie wyższą niż 60,5 projektu w miesiącu, zaś przynajmniej 75% było nie mniej wydajnych niż 60,5 projektu w miesiącu. Co najmniej 75% pracowników miało wydajność równą lub mniejszą niż 78,5 projektu w miesiącu, zaś co najmniej 25% miało nie mniej niż 78,5 projektu zrealizowanego w miesiącu. Wydajność co najmniej połowy pracowników nie przekraczała 69 projektów w miesiącu. Przeciętnie badani pracownicy cechowali się wydajnością wynoszącą około 68,97 zrealizowanego projektu w miesiącu, przy czym wydajność tych pracowników różniła się od tej średniej wydajności przeciętnie o 13,14 projektu w miesiącu. Zatem zróżnicowanie pracowników ze względu na wydajność jest słabe, bo wynosiło 19,05%.

Ponad połowa pracowników miała wydajność od 55,83 projektu do 82,11 projektu w miesiącu. Rozkład wydajności pracowników jest rozkładem o niewielkiej asymetrii prawostronnej, co oznacza, że niewiele ponad połowa pracowników miała wydajność niższą od średniej. Rozkład wydajności pracowników jest w niewielkim stopniu spłaszczony w porównaniu do rozkładu normalnego.

.

3. Teoretyczne rozkłady zmiennych losowych w komputerowej analizie danych

3.1. Rozkłady skokowe (dyskretne)

W komputerowej analizie danych fundamentalne znaczenie ma podział zmiennych losowych na skokowe (dyskretne) oraz ciągłe. **Zmienna losowa o rozkładzie skokowym** to taka, która przyjmuje skończoną lub co najwyżej przeliczalną liczbę unikalnych wartości, a każdemu z tych wariantów przypisane jest określone prawdopodobieństwo. Najprostszą formą w tej rodzinie jest **rozkład zero-jedynkowy**, stanowiący model matematyczny dla doświadczeń, w których mogą zaistnieć tylko dwa wykluczające się zdarzenia, tradycyjnie nazywane sukcesem oraz porażką. Prawdopodobieństwo zaistnienia sukcesu określa się parametrem p, co formalnie zapisujemy jako $P(X = 1) = p$, natomiast prawdopodobieństwo porażki wynosi $q = 1 - p$, czyli $P(X = 0) = q = 1 - p$.

Rozwinięciem tego modelu dla serii n niezależnych powtórzeń jest rozkład dwumianowy, znany również jako **rozkład Bernoulliego**. Opisuje on prawdopodobieństwo uzyskania dokładnie k sukcesów w n próbach, gdzie w każdej z nich prawdopodobieństwo sukcesu wynosi p. Funkcja prawdopodobieństwa rozkładu dwumianowego opiera się na symbolu Newtona i przyjmuje postać:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

W programie Statistica rozkład ten jest wykorzystywany m.in. do modelowania procesów kontroli jakości, gdzie bada się liczbę wadliwych elementów w pobranej partii towaru.

W sytuacjach, gdy liczba prób n dąży do nieskończoności, a prawdopodobieństwo sukcesu p jest ekstremalnie małe (bliskie zeru), rozkład dwumianowy dąży asymptotycznie do rozkładu Poissona, będącego klasycznym modelem dla tzw. zdarzeń rzadkich. Parametrem wejściowym jest tutaj $\lambda = np$, reprezentujący średnią liczbę zdarzeń w określonym przedziale czasu lub przestrzeni. Funkcja prawdopodobieństwa rozkładu Poissona wyraża się wzorem:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

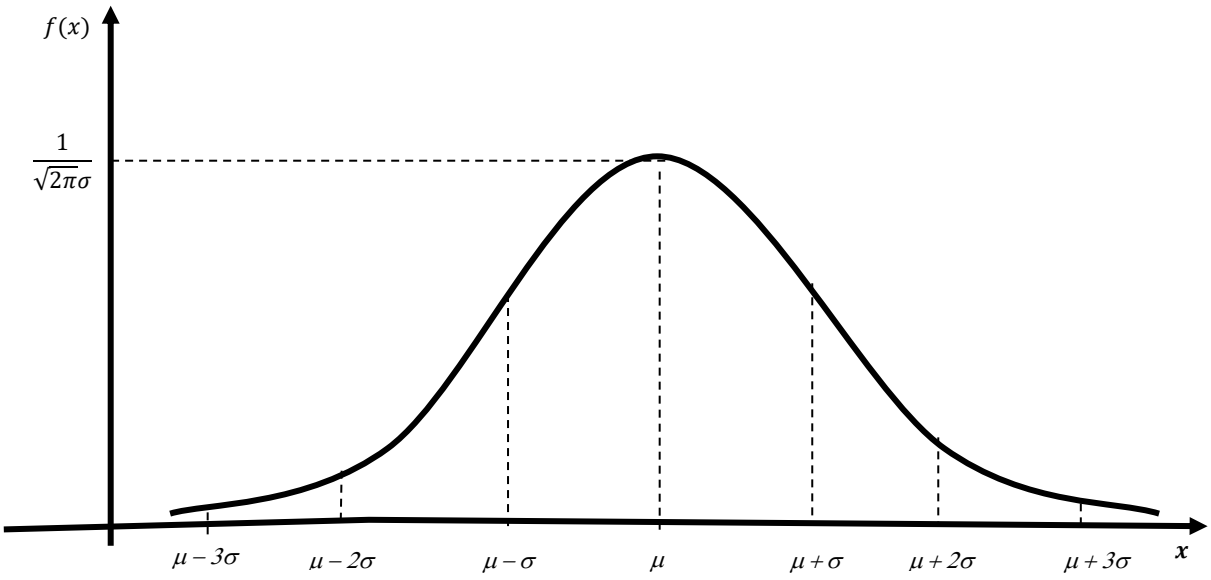
gdzie e stanowi podstawę logarytmu naturalnego. Model ten służy do analizy intensywności zgłoszeń awarii, wypadków czy błędów na liniach montażowych. Program Statistica udostępnia dedykowany kalkulator rozkładów dyskretnych, który pozwala na natychmiastowe obliczenie prawdopodobieństw punktowych oraz skumulowanych dla dowolnie zdefiniowanych parametrów n, p czy λ.

3.2. Rozkład normalny (Gaussa-Laplace'a)

Przejście do analizy cech mierzonych na skalach silnych (przedziałowej i ilorazowej) wymaga zastosowania rozkładów ciągłych, wśród których absolutnie nadrzędną rolę odgrywa rozkład normalny, zwany rozkładem Gaussa-Laplace'a. Wynika to bezpośrednio z Centralnych Twierdzeń Granicznych, według których suma dużej liczby niezależnych czynników losowych o dowolnych rozkładach dąży w granicy do rozkładu normalnego. Zmienna losowa X podlega rozkładowi normalnemu o parametrach μ (średnia w populacji) oraz σ (odchylenie standardowe w populacji), co zapisujemy jako $X \sim N(\mu, \sigma)$, jeżeli jej funkcja gęstości prawdopodobieństwa opisana jest wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wykres tej funkcji tworzy charakterystyczną, idealnie symetryczną krzywą dzwonową.



Ponieważ parametry μ oraz σ mogą przyjmować nieskończenie wiele kombinacji, w celach obliczeniowych algorytmy programu Statistica dokonują standaryzacji rozkładu normalnego. Proces ten polega na transformacji zmiennej X w zmienną standaryzowaną U poprzez odjęcie od każdej wartości średniej i podzielenie przez odchylenie standardowe, co wyraża się jawnym wzorem transformacji $U = (X - \mu)/\sigma$. W wyniku tej operacji uzyskujemy standaryzowany rozkład normalny o parametrach zerowej średniej i jednostkowego odchylenia standardowego, oznaczany jako $U \sim N(0,1)$. Program Statistica wykonuje te operacje automatycznie, eliminując konieczność korzystania z tradycyjnych tablic statystycznych.

3.3. Pozostałe rozkłady wykorzystywane we wnioskowaniu statystycznym

Oprócz rozkładu normalnego, statystyka matematyczna i procedury weryfikacji hipotez w programie Statistica opierają się między innymi na trzech teoretycznych rozkładach pochodnych, które powstają w wyniku matematycznych przekształceń niezależnych zmiennych o rozkładzie normalnym. Są to rozkłady: chi-kwadrat (χ^2), t-Studenta oraz F Fishera-Snedecora. Ich znajomość jest niezbędna do zrozumienia, skąd program czerpie tzw. wartości krytyczne w testach istotności.

Pierwszym z nich jest rozkład chi-kwadrat, oznaczany symbolem χ^2. Powstaje on jako suma kwadratów k niezależnych zmiennych losowych, z których każda podlega standaryzowanemu rozkładowi normalnemu. Parametrem jednoznacznie definiującym ten rozkład jest liczba stopni swobody. Rozkład χ^2 jest rozkładem asymetrycznym, przyjmującym wyłącznie wartości nieujemne. W miarę wzrostu liczby stopni swobody, kształt rozkładu chi-kwadrat dąży asymptotycznie do rozkładu normalnego. Program Statistica wykorzystuje ten rozkład do weryfikacji hipotez dotyczących wariancji jednej populacji oraz w nieparametrycznych testach zgodności i niezależności.

Drugim fundamentalnym modelem jest rozkład t-Studenta, opracowany przez Williama Gosseta. Definiuje się go jako stosunek zmiennej o standaryzowanym rozkładzie normalnym do pierwiastka z niezależnej zmiennej o rozkładzie chi-kwadrat, podzielonej przez jej liczbę stopni swobody. Wykres rozkładu t-Studenta jest, podobnie jak rozkład normalny, symetryczny względem zera i ma kształt dzwonowy, jednak charakteryzuje się większym rozproszeniem (ma „grubsze ogony”). W miarę jak liczba stopni swobody dąży do nieskończoności (w praktyce dla prób o liczebności $n > 30$), rozkład t-Studenta staje się identyczny ze standaryzowanym rozkładem normalnym $N(0,1)$. Rozkład ten jest rdzeniem algorytmów programu Statistica służących do testowania istotności średnich (testy t-Studenta dla jednej i dwóch prób) oraz oceny istotności współczynników korelacji i regresji.

Trzecim rozkładem pochodnym jest rozkład F Fishera-Snedecora, powstały w kontekście analizy wariancji. Stanowi on stosunek dwóch niezależnych zmiennych losowych o rozkładach chi-kwadrat, z których każda została podzielona przez odpowiadającą jej liczbę stopni swobody. Rozkład F jest zdefiniowany przez parę stopni swobody: stopnie swobody licznika oraz stopnie swobody mianownika. Podobnie jak rozkład chi-kwadrat, przyjmuje on wyłącznie wartości dodatnie i jest prawostronnie asymetryczny. W programie Statistica rozkład F stanowi podstawę działania modułu analizy wariancji (ANOVA), gdzie służy do porównywania

wariancji międzygrupowych i wewnątrzgrupowych, a także do testowania łącznej istotności strukturalnej modeli regresji wielorakiej (test Walda).

4. Teoria estymacji

4.1. Podstawowe pojęcia teorii estymacji ora zwłasności estymatorów

Proces wnioskowania statystycznego, realizowany w programie Statistica, opiera się w znacznej mierze na teorii **estymacji**, czyli szacowaniu nieznanych parametrów rozkładu populacji generalnej na podstawie danych uzyskanych z próby losowej. Estymację dzielimy zasadniczo na **parametryczną**, która dotyczy wnioskowania o konkretnych wartościach liczbowych (takich jak średnia czy wariancja) przy znanym typie rozkładu oraz **nieparametryczną**, służącą do wyznaczania ogólnej postaci funkcyjnej tego rozkładu. W ramach podejścia parametrycznego wyróżnia się **estymację punktową**, gdzie nieznany parametr populacji jest reprezentowany przez pojedynczą wartość liczbową (np. średnia z próby stanowi punktowy estymator średniej w populacji), oraz **estymację przedziałową**. Ta druga polega na konstrukcji tzw. **przedziału ufności**, czyli losowego przedziału o granicach obliczanych na podstawie próby, który z góry określonym, wysokim prawdopodobieństwem – nazywanym współczynnikiem ufności i oznaczanym symbolem $1 - \alpha$ – pokrywa rzeczywistą, nieznaną wartość szacowanego parametru populacji.

Aby funkcja statystyczna z próby mogła stać się poprawnym estymatorem, musi spełniać pożądane własności matematyczne. Pierwszą z nich jest **nieobciążoność**, oznaczająca, że wartość oczekiwana estymatora jest równa rzeczywistej wartości szacowanego parametru. Drugą cechą jest **efektywność**, która wskazuje, że dany estymator charakteryzuje się najmniejszą wariancją (najmniejszym rozrzutem ocen) wśród wszystkich nieobciążonych rozwiązań. Trzecia cecha to **zgodność**, czyli własność asymptotyczna gwarantująca, że wraz ze wzrostem liczebności próby n do nieskończoności, wartość estymatora zbiega stochastycznie do prawdziwej wartości parametru w populacji.

Kluczowym elementem planowania każdego eksperymentu badawczego oraz procedur estymacji przedziałowej jest **maksymalny dopuszczalny błąd szacunku**, oznaczany literą d. Definiuje się go jako połowę długości przedziału ufności, co oznacza, że wyznacza on maksymalną akceptowalną odległość między wartością estymatora uzyskaną z próby a rzeczywistą wartością parametru w populacji generalnej. Precyzja badania rośnie wraz ze spadkiem wartości d, co jednak w naturalny sposób wymaga zwiększenia nakładów na powiększenie zbioru danych. Przed uruchomieniem procedury zbierania danych badacz musi

podjąć decyzję, jak dużą próbę należy wylosować, aby przy zdefiniowanym współczynniku ufności $1-\alpha$ nie przekroczyć założonego progu błędu d. Program Statistica dostarcza w tym zakresie algorytmy matematyczne w ramach modułu analizy mocy i wielkości próby.

4.2. Estymatory punktowe i przedziałowe

W procedurach komputerowych programu Statistica punktowe szacowanie parametrów uzupełnione jest wyznaczaniem ich granic przedziałowych, co pozwala ocenić precyzję wnioskowania na podstawie rozkładów prawdopodobieństwa statystyk testowych. Podczas analizy wartości średniej populacji, oznaczanej symbolem μ, punktowym estymatorem jest średnia z próby $\overline{x}$. W modelu przy znanej wariancji σ^2 przedział ufności konstruuje się w oparciu o standaryzowany rozkład normalny $N(0,1)$, co wyraża się wzorem:

$$P\left(\overline{x}-u_\alpha\frac{\sigma}{\sqrt{n}}<\mu<\overline{x}+u_\alpha\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

gdzie symbol u_α jest wartością krytyczną rozkładu normalnego spełniającą warunek $P(-u_\alpha<U<u_\alpha)=1-\alpha$.

W sytuacji, gdy wariancja jest nieznana, a dysponujemy małą próbą, gdzie $n\leq 30$, program wykorzystuje nieobciążony estymator wariancji z próby $\hat{s}^2$. Granice przedziału ufności dla średniej wyznaczane są wtedy w oparciu o rozkład t-Studenta o liczbie stopni swobody $df=n-1$, co zapisujemy jako:

$$P\left(\overline{x}-t_\alpha\frac{\hat{s}}{\sqrt{n}}<\mu<\overline{x}+t_\alpha\frac{\hat{s}}{\sqrt{n}}\right)=1-\alpha$$

gdzie t_α to wartość odczytywana przez program z tablic rozkładu t-Studenta.

Przy nieznanej wariancji i dużej próbie, gdy $n>30$, zgodnie z centralnymi twierdzeniami granicznymi, rozkład t-Studenta zbiega do rozkładu normalnego, co pozwala programowi na automatyczne zastosowanie statystyki u_α w miejsce t_α przy wariancji z próby s^2.

Szacowanie parametru zmienności, czyli wariancji populacji σ^2, oparte jest na punktowym estymatorze $\hat{s}^2$. Przedział ufności dla wariancji opiera się na asymetrycznym rozkładzie chi-kwadrat i przyjmuje postać:

$$P\left(\frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2}}}<\sigma^2<\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2}}}\right)=1-\alpha$$

gdzie wartości mianowników reprezentują odpowiednie kwantyle rozkładu chi-kwadrat odczytywane przez algorytmy oprogramowania.

Ostatnim podstawowym parametrem jest wskaźnik struktury, czyli proporcja p, stosowana w populacjach o rozkładzie dwupunktowym. Punktowym estymatorem jest tutaj frakcja z próby m/n, definiowana jako stosunek liczby elementów wyróżnionych m do liczebności próby n. Dla dużych prób, gdzie $n > 100$, konstruuje się przedział ufności wykorzystujący aproksymację rozkładem normalnym, co zapisujemy wzorem:

$$P\left(\frac{m}{n} - u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right) = 1-\alpha$$

Wszystkie te struktury przedziałowe można również zobrazować graficznie, generując w programie Statistica np. wykresy ramka-wąsy.

4.3. Szacowanie minimalnej liczebności próby

Wyznaczenie minimalnej wielkości próby przed przystąpieniem do właściwych badań realizowane jest w ramach trzech głównych modeli teoretycznych. Model I znajduje zastosowanie, gdy populacja generalna charakteryzuje się rozkładem normalnym $N(\mu,\sigma)$, a wariancja populacji σ^2 lub odchylenie standardowe σ jest znane z danych historycznych lub założeń procesowych. Jeśli badacz stawia wymóg, aby przy ustalonym współczynniku ufności $1-\alpha$ maksymalny błąd szacunku średniej nie przekroczył z góry określonej wartości d, minimalną liczebność próby n wyznacza się za pomocą wzoru

$$n = \frac{u_\alpha^2\sigma^2}{d^2}$$

w którym u_α to wartość zmiennej normalnej standaryzowanej $N(0,1)$ odczytana w taki sposób, aby zachodziła równość prawdopodobieństwa $P(-u_\alpha < U < u_\alpha) = 1-\alpha$. W sytuacji, gdy uzyskana wartość n przewyższa możliwości organizacyjne eksperymentu, mniejszą liczebność można wyznaczyć wyłącznie poprzez celowe zwiększenie dopuszczalnego błędu szacunku d.

Model II rozwiązuje problem planowania, gdy wariancja populacji σ^2 jest nieznana. Procedura wymaga wtedy przeprowadzenia wstępnego badania pilotażowego na małej próbie o liczebności n_0 elementów. Z uzyskanych danych wylicza się nieobciążoną wariancję z próby $\hat{s}^2$. Niezbędną liczebność próby właściwej n program oblicza wówczas według formuły:

$$n = \frac{t_\alpha^2\hat{s}^2}{d^2}$$

gdzie t_α oznacza wartość zmiennej rozkładu t-Studenta odczytaną dla zdefiniowanego współczynnika ufności $1-\alpha$ oraz dla liczby stopni swobody równej $n_0 - 1$, tak aby spełniony był warunek $P(-t_\alpha < t < t_\alpha) = 1-\alpha$. Po wyznaczeniu wartości n stosuje się regułę

decyzyjną mówiącą, że jeśli obliczona liczebność próby spełnia nierówność $n < n_0$, liczebność próby wstępnej jest wystarczająca i badanie można zakończyć, jeśli zaś zachodzi relacja $n > n_0$, badacz musi wylosować z populacji i dodatkowo przebadać jeszcze $n - n_0$ elementów.

Model III odnosi się do sytuacji, gdy populacja generalna realizuje rozkład dwupunktowy z parametrem p, reprezentującym odsetek elementów wyróżnionych w zbiorowości. Obliczenia prowadzi się w dwóch wariantach metodologicznych, gdzie w wariancie pierwszym, gdy znany jest spodziewany rząd wielkości frakcji p, niezbędną wielkość próby wyznacza się bezpośrednio ze wzoru:

$$n = \frac{u_\alpha^2 \cdot p \cdot (1-p)}{d^2}$$

w którym u_α to wartość krytyczna rozkładu normalnego $N(0,1)$ odpowiadająca współczynnikowi ufności $1 - \alpha$. W wariancie drugim, gdy rząd wielkości parametru p jest całkowicie nieznany, w celu zapewnienia maksymalnego bezpieczeństwa estymacji przyjmuje się założenie najmniej korzystne strukturalnie. Ponieważ funkcja matematyczna iloczynu osiąga swoje maksimum dla wartości $p = 0{,}5$, gdzie przyjmuje najwyższą możliwą wartość równą $\frac{1}{4}$, otrzymujemy bezpieczny wzór na wielkość próby, niezależny od rzeczywistej struktury populacji, który przyjmuje postać:

$$n = \frac{u_\alpha^2}{4d^2}$$

Przykład

Analiza przedziałów ufności zaprezentowanych w Tabeli 2. pozwala na uogólnienie wyników z próby 120 pracowników na całą populację generalną zatrudnioną w przedsiębiorstwie, ze współczynnikiem ufności $1 - \alpha = 0{,}95$.

Tabela 2. Przedziały ufności dla średnich oraz odchyleń standardowych

Zmienna	Średnia	Ufność -95%	Ufność +95%	Odch. Std.	P. ufności odch. std -95%	P. ufności odch. std +95%
Wydajność	68,97	66,59	71,34	13,14	11,66	15,05

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Statystyki podstawowe → Statystyki opisowe, wybranie przedziałów ufności i wskazanie współczynnika ufności, np. 95%.

Wyniki analizy pozwalają stwierdzić, że z 95% prawdopodobieństwem rzeczywista, nieznana średnia wydajność dla całej populacji pracowników mieści się w granicach od 66,6 do 71,3 zrealizowanego projektu w miesiącu. Z kolei z 95% ufnością można przyjąć, że przeciętne odchylenie standardowe wydajności pracowników w całym przedsiębiorstwie zawiera się w przedziale od 11,7 do 15,1 projektu w miesiącu.

5. Procedury weryfikacji hipotez i testy istotności

5.1. Metodologia weryfikacji hipotez statystycznych i procedury decyzyjne

Procedura weryfikacji hipotez statystycznych stanowi, obok teorii estymacji, drugi główny filar wnioskowania statystycznego realizowanego w programie Statistica. **Hipoteza statystyczna** to każdy sąd lub przypuszczenie dotyczące parametru bądź postaci rozkładu populacji generalnej, wydane bez przeprowadzenia pełnego badania całkowitego. Metodologia ta dzieli hipotezy na parametryczne, które precyzują konkretną wartość numeryczną parametru (np. średniej lub wariancji) w rozkładzie populacji o znanym typie, oraz nieparametryczne, które określają ogólną postać funkcjonalną bądź zgodność całego rozkładu populacji. Formalny proces testowania opiera się na zestawieniu dwóch wykluczających się twierdzeń: **hipotezy zerowej**, oznaczanej jako H_0, która stanowi bazowe założenie o braku zmian, braku różnic lub braku wpływu i podlega bezpośredniej procedurze weryfikacyjnej oraz **hipotezy alternatywnej**, oznaczanej jako H_1, która jest twierdzeniem konkurencyjnym, przyjmowanym w momencie odrzucenia hipotezy zerowej. Jako klasyczny przykład można wskazać sytuację, w której firma logistyczna monitoruje efektywność i chce sprawdzić, czy średni czas dostawy przesyłek wynosi dokładnie 24 godziny, co prowadzi do sformułowania hipotezy zerowej w postaci $H_0: \mu = 24$ wobec hipotezy alternatywnej $H_1: \mu \neq 24$.

Sercem matematycznym każdego testu statystycznego jest **statystyka testowa** (sprawdzian testu), czyli określona funkcja zmiennych z próby, której rozkład prawdopodobieństwa przy założeniu prawdziwości hipotezy H_0 jest badaczowi dokładnie znany (może to być rozkład normalny, t-Studenta, chi-kwadrat czy F Fishera-Snedecora). Na podstawie tego rozkładu oraz przyjętego poziomu istotności α konstruuje się obszar krytyczny, czyli podzbiór przestrzeni próby o tej unikalnej właściwości, że jeśli obliczona na podstawie danych empirycznych wartość statystyki testowej w nim się znajdzie, podejmuje się decyzję o odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej. Granicę tego obszaru

wyznacza **wartość krytyczna**, stanowiąca punkt odcięcia w rozkładzie prawdopodobieństwa statystyki.

W zależności od sformułowania hipotezy alternatywnej, obszar krytyczny może przyjąć postać dwustronną (gdy H_1 używa znaku różności $\neq$) bądź jednostronną (prawostronną przy znaku większości $>$ lub lewostronną przy znaku mniejszości $<$). Choć klasyczna procedura wymaga manualnego porównywania statystyki z wartością krytyczną, program Statistica opiera proces weryfikacji na komputerowym wskaźniku **p-value** (p-wartości). P-value to najmniejszy poziom istotności, przy którym zaobserwowana statystyka testowa pozwala na odrzucenie hipotezy zerowej, odzwierciedlający prawdopodobieństwo uzyskania wyniku równie skrajnego lub jeszcze bardziej skrajnego niż ten zaobserwowany w próbie, przy założeniu, że hipoteza H_0 jest prawdziwa. Reguła decyzyjna wskazuje, że jeśli wartość $p < \alpha$, wówczas wynik jest statystycznie istotny i hipotezę H_0 należy odrzucić; w przeciwnym wypadku, gdy $p > \alpha$, stwierdza się brak podstaw do odrzucenia H_0, co metodologicznie nie oznacza udowodnienia jej prawdziwości, a jedynie brak dostatecznych dowodów empirycznych na jej negację.

5.2. Błędy decyzji statystycznych i moc testu

Wnioskowanie statystyczne na podstawie próby losowej ze swej natury obarczone jest ryzykiem podjęcia błędnej decyzji, ponieważ badacz nie dysponuje pełną wiedzą o populacji generalnej. Teoria weryfikacji hipotez precyzuje dwa rodzaje błędów. Błąd I rodzaju polega na odrzuceniu hipotezy zerowej w sytuacji, gdy w rzeczywistości jest ona prawdziwa. Maksymalne dopuszczalne prawdopodobieństwo popełnienia tego błędu jest kontrolowane przez badacza i nazywane poziomem istotności, oznaczanym symbolem α. Najczęściej przyjmowanymi w nauce i praktyce inżynierskiej poziomami α są wartości 0,1; 0,05; 0,01 lub 0,001, przy czym wybór mniejszego poziomu istotności (np. 0,01 zamiast 0,05) zmniejsza ryzyko fałszywego odkrycia, ale jednocześnie rozszerza przedział ufności i utrudnia odrzucenie hipotezy zerowej. Intuicyjną analogią błędu I rodzaju w procesie sądowym jest skazanie osoby całkowicie niewinnej.

Z kolei błąd II rodzaju polega na nieodrzuceniu (przyjęciu) hipotezy zerowej w sytuacji, gdy w rzeczywistości jest ona fałszywa. Prawdopodobieństwo popełnienia tego błędu oznacza się symbolem β. W kontekście procesowym błąd II rodzaju odpowiada uniewinnieniu rzeczywistego przestępcy. Bezpośrednio z parametrem β powiązana jest moc testu, definiowana jako dopełnienie prawdopodobieństwa błędu II rodzaju do jedności, czyli

wyrażana wzorem $1 - \beta$. Wysoka moc testu oznacza dużą zdolność procedury statystycznej do poprawnego odrzucenia hipotezy zerowej, gdy jest ona w rzeczywistości fałszywa.

Pomiędzy prawdopodobieństwami α i β istnieje relacja odwrotna: przy stałej liczebności próby próba zmniejszenia poziomu α automatycznie powoduje wzrost wartości β (spadek mocy testu). Jedynym matematycznym sposobem na jednoczesne obniżenie prawdopodobieństwa obu błędów jest zwiększenie liczebności próby losowej n, co stanowi pomost łączący teorię testowania z minimalną wielkością próby opisaną powyżej.

5.3. Parametryczne testy istotności dla jednej populacji

Parametryczne testy istotności dla jednej populacji służą do weryfikacji hipotez, w których wartość wyliczona z próby jest porównywana z teoretyczną bądź nominalną wartością odniesienia. Pierwszą grupą testów jednej populacji są testy istotności jednej średniej (μ). Ich procedura realizowana jest w ramach trzech głównych modeli teoretycznych. Model I wykorzystywany jest wówczas, gdy populacja generalna ma rozkład normalny $N(\mu, \sigma)$, a odchylenie standardowe σ jest dane. Wówczas sprawdzianem hipotezy zerowej $H_0: \mu = \mu_0$ jest statystyka u o rozkładzie standaryzowanym $N(0,1)$, wyznaczana wzorem:

$$u = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$$

gdzie $\overline{x}$ to średnia z próby, a n to jej wielkość. Przykładowo, gdy firma logistyczna o znanym historycznym odchyleniu $\sigma = 2$ godziny bada, czy obecny średni czas dostaw różni się od normy $\mu_0 = 24$ godziny, na podstawie próby $n = 36$ dostaw o średniej $\overline{x} = 23{,}5$ godziny program wylicza statystykę u i generuje p-value, które bada się na poziomie istotności np. $\alpha = 0{,}05$.

Model II wykorzystywany jest gry populacja generalna ma rozkład normalny oraz nieznaną wariancję σ^2 i małą liczebność próba ($n \leq 30$). Sprawdzianem hipotezy jest statystyka t podlegająca rozkładowi t-Studenta o $n - 1$ stopniach swobody, opisywana równaniem:

$$t = \frac{\overline{x} - \mu_0}{s}\sqrt{n-1} = \frac{\overline{x} - \mu_0}{\hat{s}}\sqrt{n}$$

w którym symbol $s$oznacza obciążone, a $\hat{s}$ nieobciążone odchylenie standardowe z próby.

Model III wykorzystywany jest gry populacja generalna ma dowolny rozkład oraz nieznaną wariancję σ^2 i dużą liczebność próba ($n > 30$). Sprawdzianem hipotezy zerowej staje się statystyka u aproksymowana rozkładem normalnym $N(0,1)$, wyznaczana wzorem:

$$u = \frac{\overline{x} - \mu_0}{s}\sqrt{n}$$

Kolejnym testem jednej populacji jest test istotności wariancji jednej populacji (σ^2). Test ten służy do weryfikacji, czy stabilność (zmienność) procesu różni się od wartości teoretycznej σ_0^2. Dla prób małych ($n \leq 30$) statystyka testowa podlegająca rozkładowi chi-kwadrat z $n-1$ stopniami swobody, wyliczana ze wzoru:

$$\chi^2_{obl} = \frac{n \cdot s^2}{\sigma_0^2} = \frac{(n-1) \cdot \hat{s}^2}{\sigma_0^2}$$

Dla prób dużych ($n > 30$) obliczoną wartość chi-kwadrat przekształca się do statystyki u rozkładu normalnego $N(0,1)$ jawnym wzorem transformacji $u = \sqrt{2\chi^2} - \sqrt{2n-3}$.

Ostatnim opisanym w niniejszych materiałach testem jednej populacji jest test istotności wskaźnika struktury (proporcji p). Wykorzystywany jest, gdy populacja ma rozkład dwupunktowy, a liczebność próby spełnia wymóg $n > 100$, test istotności dla hipotezy $H_0: p = p_0$ realizuje się za pomocą statystyki u o rozkładzie normalnym $N(0,1)$, opisanej wzorem:

$$u = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

gdzie m to liczba elementów wyróżnionych.

5.4. Parametryczne testy istotności dla dwóch populacji

W praktyce badawczej i analizach inżynierskich znacznie częściej niż ocena jednej grupy zachodzi potrzeba porównania dwóch niezależnych populacji ze sobą, na przykład w celu zweryfikowania, czy wydajność dwóch linii produkcyjnych, czas dostaw realizowanych przez dwie konkurencyjne firmy kurierskie bądź odsetek wadliwych komponentów w dwóch fabrykach różnią się od siebie w sposób nieprzypadkowy.

Podczas weryfikacji hipotezy zerowej o braku różnic między średnimi w dwóch populacjach, oznaczanej formalnie jako $H_0: \mu_1 = \mu_2$ wobec odpowiednio sformułowanej hipotezy alternatywnej, wybór właściwego sprawdzianu zależy od liczebności prób oraz stopnia znajomości parametrów rozproszenia populacji generalnych. Model I znajduje zastosowanie w sytuacji, gdy populacje generalne charakteryzują się rozkładami normalnymi $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$, bądź rozkładami do nich zbliżonymi, a odchylenia standardowe obu populacji σ_1 i σ_2 są badaczowi dokładnie znane. Sprawdzianem tak postawionej hipotezy jest statystyka u podlegająca standaryzowanemu rozkładowi normalnemu $N(0;1)$, którą oblicza się za pomocą wzoru:

$$u = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

gdzie symbole $\overline{x}_1$ oraz $\overline{x}_2$ oznaczają średnie obliczone odpowiednio z pierwszej i drugiej próby, natomiast n_1 i n_2 definiują ich liczebności.

Model II opisuje sytuację, w której populacje generalne mają rozkłady normalne, jednak ich odchylenia standardowe σ_1 i σ_2 pozostają nieznane, lecz na podstawie założeń teoretycznych bądź wcześniejszych testów można uznać je za jednakowe, co zapisujemy jako równość $\sigma_1 = \sigma_2$. Jeżeli pobrane z populacji próby są małe, to znaczy ich liczebności spełniają warunki $n_1 \leq 30$ oraz $n_2 \leq 30$, sprawdzianem hipotezy zerowej jest statystyka t o rozkładzie t-Studenta z liczbą stopni swobody zdefiniowaną jako $n_1 + n_2 - 2$. Formuła obliczeniowa implementowana w tym teście dla prób niezależnych przyjmuje następującą postać:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{n_1\hat{s}_1^2 + n_2\hat{s}_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

w której symbole s_1 oraz s_2 odzwierciedlają obciążone odchylenia standardowe z prób, podczas gdy $\hat{s}_1$ i $\hat{s}_2$ reprezentują nieobciążone estymatory tych odchyleń.

Jeżeli odchylenia standardowe populacji są nieznane, ale pobrane próby mają charakter duży, co oznacza spełnienie nierówności $n_1 > 30$ oraz $n_2 > 30$, algorytm programu przechodzi do Modelu III, w którym sprawdzianem hipotezy zerowej staje się statystyka u podlegająca aproksymacji rozkładem normalnym $N(0; 1)$, wyliczana bezpośrednio ze wzoru:

$$u = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Przed uruchomieniem porównania średnich dla małych prób w Modelu II, należy zweryfikować założenie o jednorodności, czyli równości wariancji w obu populacjach, sprawdzając hipotezę zerową $H_0: \sigma_1^2 = \sigma_2^2$ wobec hipotezy alternatywnej $\$H_1: \sigma_1^2 > \sigma_2^2\$$. Sprawdzianem w tym teście istotności dwóch wariancji jest statystyka F_{obl}, konstruowana jako stosunek większej wariancji nieobciążonej z próby do wariancji mniejszej, tak aby jej wartość była zawsze większa lub równa jedności, co opisuje wzór:

$$F_{obl} = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{\frac{n_1}{n_1 - 1} \cdot s_1^2}{\frac{n_2}{n_2 - 1} \cdot s_2^2} \quad \text{przy warunku, że} \quad \hat{s}_1^2 > \hat{s}_2^2$$

Statystyka ta przy założeniu prawdziwości hipotezy zerowej podlega rozkładowi F-Snedecora z liczbą stopni swobody licznika równą $n_1 - 1$ oraz mianownika równą $n_2 - 1$.

Ostatnią procedurą porównawczą dla dwóch populacji jest test istotności dwóch wskaźników struktury, stosowany, gdy populacje te mają rozkłady dwupunktowe z parametrami frakcji elementów wyróżnionych p_1 oraz p_2. Jeśli pobrane liczebności prób są bardzo duże i spełniają warunek komputerowy $n_1 > 100$ oraz $n_2 > 100$, sprawdzianem hipotezy zerowej $H_0: p_1 = p_2$ o braku różnic strukturalnych jest statystyka u podlegająca rozkładowi normalnemu $N(0; 1)$. Formuła obliczeniowa wiąże udziały empiryczne z prób za pomocą połączonej, średniej proporcji i przyjmuje postać:

$$u = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}$$

w której zmienne pomocnicze definiuje się jako wspólną frakcję próby łącznej $\overline{p} = \frac{m_1+m_2}{n_1+n_2}$ oraz jako wskaźnik liczebności prób wyrażany ułamkiem $n = \frac{n_1 n_2}{n_1+n_2}$, gdzie symbole m_1 i m_2 oznaczają fizyczną liczbę obiektów wyróżnionych w obu badaniach.

5.5. Nieparametryczny test dwóch prób niezależnych – Test U Manna-Whitneya

W sytuacjach badawczych, gdy celem analizy jest porównanie poziomów położenia (przeciętnych) w dwóch niezależnych populacjach, a założenia klasycznego parametrycznego testu t-Studenta zostały naruszone, wymagane jest przejście do statystyki nieparametrycznej. Powszechnie stosowanym narzędziem w tym obszarze jest test U Manna-Whitneya (znany również jako test Wilcoxona-Manna-Whitneya). Test ten nie stawia żadnych wymagań dotyczących normalności rozkładu cechy w populacjach generalnych oraz cechuje się niewrażliwością na obecność nawet bardzo skrajnych obserwacji odstających. Kluczowym warunkiem jego poprawności formalnej jest to, aby badana zmienna zależna była mierzona na skali co najmniej porządkowej, co umożliwia operację porządkowania i rangowania oraz aby obie próby o liczebnościach oznaczanych jako n_1 oraz n_2 były pobrane w sposób całkowicie niezależny.

Matematyczna procedura tego testu, realizowana automatycznie przez program Statistica, opiera się na analizie wzajemnego przenikania się rang obu grup. Na pierwszym etapie algorytm łączy obie próby w jeden wspólny szereg szczegółowy, sortuje wszystkie obserwacje rosnąco i przypisuje im rangi od 1 do $n_1 + n_2$. Następnie, po rozdzieleniu obiektów z powrotem na pierwotne grupy, obliczane są sumy rang dla każdej z nich, oznaczane odpowiednio jako R_1

oraz R_2. Na ich podstawie program wyznacza dwie statystyki pomocnicze U_1 oraz U_2, wplatając w strukturę obliczeń następujące równania:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \text{oraz} \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Ostatecznym sprawdzianem weryfikowanej hipotezy zerowej H_0, mówiącej o identyczności rozkładów (równości median) w obu populacjach, jest statystyka U, definiowana jako mniejsza z uzyskanych wartości pomocniczych, co zapisujemy matematycznie jako $U = \min(U_1, U_2)$. W przypadku małych prób program odczytuje wartości krytyczne ze specjalnych tablic statystycznych, natomiast dla prób dużych ($n_1, n_2 > 20$) statystyka U podlega asymptotycznej aproksymacji rozkładem normalnym $N(0,1)$ za pomocą standaryzowanej statystyki Z. Wygenerowanie przez program wartości $p < \alpha$ (najczęściej 0,05) stanowi podstawę do odrzucenia hipotezy zerowej i stwierdzenia, że analizowane dwie grupy różnią się od siebie w sposób istotny statystycznie.

6. Analiza zgodności rozkładów i testy normalności

6.1. Wprowadzenie do testów nieparametrycznych

W klasycznej metodologii wnioskowania statystycznego większość standardowych parametrycznych testów istotności, takich jak test t-Studenta czy analiza wariancji (ANOVA), opiera się na rygorystycznym zestawie trzech podstawowych założeń, do których należą:

- wymóg, aby dane pochodziły z populacji o rozkładzie normalnym,
- postulat losowości i reprezentatywności próbek,
- założenie o niezależności poszczególnych pomiarów, co oznacza, że wartości obserwowane nie mogą na siebie wzajemnie wpływać.

W praktyce analitycznej bardzo często dochodzi jednak do sytuacji, w której przynajmniej jedno z tych fundamentalnych założeń zostaje naruszone. Niespełnienie tych kryteriów sprawia, że klasyczne testy parametryczne przestają być miarodajne, a bezkrytyczne opieranie się na ich wynikach może prowadzić do fałszywych wniosków oraz całkowicie błędnych decyzji badawczych lub zarządczych. Dodatkowo, w wielu badaniach dane ze swej natury mierzone są na słabszych skalach, na przykład na skali jakościowej nominalnej lub porządkowej (rangowej), co z góry wyklucza poprawność stosowania parametrów takich jak średnia arytmetyczna czy odchylenie standardowe.

Rozwiązaniem tego problemu w programie Statistica jest zastosowanie nieparametrycznych testów istotności, które są dedykowane dla hipotez zerowych

precyzujących ogólną postać rozkładu lub relacje między strukturami populacji generalnych, bez wnikania w konkretne wartości parametrów liczbowych. Testy te charakteryzują się unikalnymi zaletami metodologicznymi, ponieważ nie wymagają one rygorystycznych założeń dotyczących dokładnego kształtu rozkładu danych w populacji oraz wykazują wysoką odporność na obecność skrajnych obserwacji odstających i niestandardowych rozkładów o silnej skośności lub kurtozie. Matematyczna istota większości testów nieparametrycznych polega na tym, że surowe wartości liczbowe są przez algorytmy oprogramowania transformowane w rangi (pozycje w szeregu), co pozwala na eliminację wpływu znacznych odchyleń i asymetrii.

6.2. Nieparametryczne testy zgodności rozkładów

Przed uruchomieniem jakiejkolwiek zaawansowanej procedury parametrycznej badacz jest zobligowany do przeprowadzenia tzw. testu zgodności, którego celem jest zweryfikowanie, czy empiryczny rozkład danych uzyskany z próby jest zgodny z założonym rozkładem teoretycznym, najczęściej z rozkładem normalnym.

Pierwszym klasycznym narzędziem realizującym to zadanie jest test zgodności chi-kwadrat (χ^2), który polega na bezpośrednim porównywaniu liczebności empirycznych, oznaczanych jako n_i, czyli zaobserwowanych w poszczególnych klasach szeregu rozdzielczego, z liczebnościami teoretycznymi, oznaczanymi jako np_i, czyli wynikającymi z matematycznych właściwości hipotetycznego rozkładu. Sprawdzianem hipotezy zerowej o zgodności rozkładów jest statystyka χ^2_{obl}, wyliczana przez program według następującego wzoru:

$$\chi^2_{obl} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

w którym symbol k oznacza liczbę przedziałów klasowych, a np_i to oczekiwana liczebność w i-tej klasie. Statystyka ta podlega rozkładowi chi-kwadrat o liczbie stopni swobody zdefiniowanej jako $v = k - r - 1$, gdzie r to liczba parametrów hipotetycznego rozkładu szacowanych z próby (dla rozkładu normalnego $r = 2$, gdy szacuje się średnią i wariancję). Należy pamiętać o ograniczeniu tego testu: liczebności oczekiwane w każdym przedziale muszą być dostatecznie duże, co formalnie opisuje warunek $np_i \geq 5$. Jeśli w niektórych klasach liczebność teoretyczna jest mniejsza, użytkownik musi dokonać manualnego lub automatycznego łączenia sąsiednich przedziałów.

Alternatywnym testem zgodności jest test Kołmogorowa-Smirnowa, który w przeciwieństwie do testu chi-kwadrat nie wymaga grupowania danych w przedziały klasowe, dzięki czemu nie dochodzi do utraty informacji pierwotnej. Istota tego testu opiera się na analizie maksymalnej bezwzględnej różnicy między dystrybuantą empiryczną wyznaczoną z próby, oznaczaną jako $F_n(x)$, a dystrybuantą teoretyczną założonego rozkładu, oznaczaną jako $F_0(x)$. Sprawdzianem hipotezy jest statystyka D, definiowana jako:

$$D = \sup_x |F_n(x) - F_0(x)|$$

Test Kołmogorowa-Smirnowa jest dedykowany przede wszystkim dla prób o liczebnościach małych i średnich.

Choć opisane powyżej testy zgodności mogą weryfikować dopasowanie do dowolnego rozkładu (np. Poissona czy wykładniczego), w komputerowej analizie danych kluczowe znaczenie ma specyficzna ocena normalności. Spośród dostępnych procedur, najbardziej czułym, rekomendowanym testem oceny normalności dla małych i średnich prób (do kilkudziesięciu) jest test Shapiro-Wilka.

Statystyka testowa Shapiro-Wilka, oznaczana symbolem W, opiera się na analizie regresji i korelacji ocen z próby uporządkowanej względem oczekiwanych wartości statystyk pozycyjnych z rozkładu normalnego, a jej algorytm obliczeniowy wyraża się wzorem:

$$W = \frac{\left(\sum_{i=1}^{m} a_i\,(n)(x_{n-i+1} - x_i)\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

w którym x_i to elementy próby uszeregowane w szereg szczegółowy rosnący, $\overline{x}$ stanowi średnią arytmetyczną, a symbole $a_i(n)$ oznaczają współczynniki wagowe stablicowane dla określonej liczebności próby n, generowane wewnętrznie przez oprogramowanie. Wartość statystyki W przyjmuje wartości z zamkniętego przedziału od 0 do 1, przy czym wartości skrajnie bliskie jedności świadczą o idealnym dopasowaniu danych do krzywej dzwonowej rozkładu normalnego.

Weryfikacja hipotezy zerowej H_0, mówiącej, że badana zmienna losowa ma rozkład normalny, wobec hipotezy alternatywnej H_1, stwierdzającej, że rozkład zmiennej różni się istotnie od rozkładu normalnego, odbywa się w programie Statistica na podstawie wyliczonej p-wartości. Jeśli uzyskany z programu wynik spełnia relację $p < \alpha$ (gdzie α to najczęściej 0,05), hipotezę zerową o normalności należy odrzucić, co blokuje możliwość stosowania testów parametrycznych. Jeśli natomiast $p > \alpha$, stwierdza się brak podstaw do odrzucenia normalności, co uprawnia analityka do przeprowadzenia metod klasycznych.

Przykład

W celu zweryfikowania, czy przeciętna miesięczna wydajność pracowników w badanym przedsiębiorstwie różni się od odgórnie założonej normy wydajności, przeprowadzone zostało wnioskowanie statystyczne za pomocą parametrycznego testu istotności dla jednej populacji (testu t-Studenta). Jako wartość odniesienia (normę docelową dla stanowiska) przyjęto wartość $\mu_0 = 70$ zrealizowanych projektów w miesiącu. Przed uruchomieniem procedury weryfikacyjnej sformułowano hipotezę zerową oraz hipotezę alternatywną. Jako hipotezę zerową przyjęto, że przeciętna miesięczna wydajność pracowników w populacji generalnej jest równa normie wynoszącej siedemdziesiąt zrealizowanych projektów w miesiącu. Z kolei jako hipotezę alternatywną, sformułowano sąd, że przeciętna miesięczna wydajność pracowników w populacji różni się w sposób istotny od założonej normy siedemdziesięciu zrealizowanych projektów. Wyniki przeprowadzonego testu t-Studenta dla jednej próby zestawiono w Tabeli 3.

Tabela 3. Wyniki testu t-Studenta dla jednej próby

Zmienna	Średnia	Odch.st.	Ważnych	Bł. std.	Odniesienie	t	df	p
Wydajność	68,97	13,14	120	1,199	70,000	-0,861	119	0,391

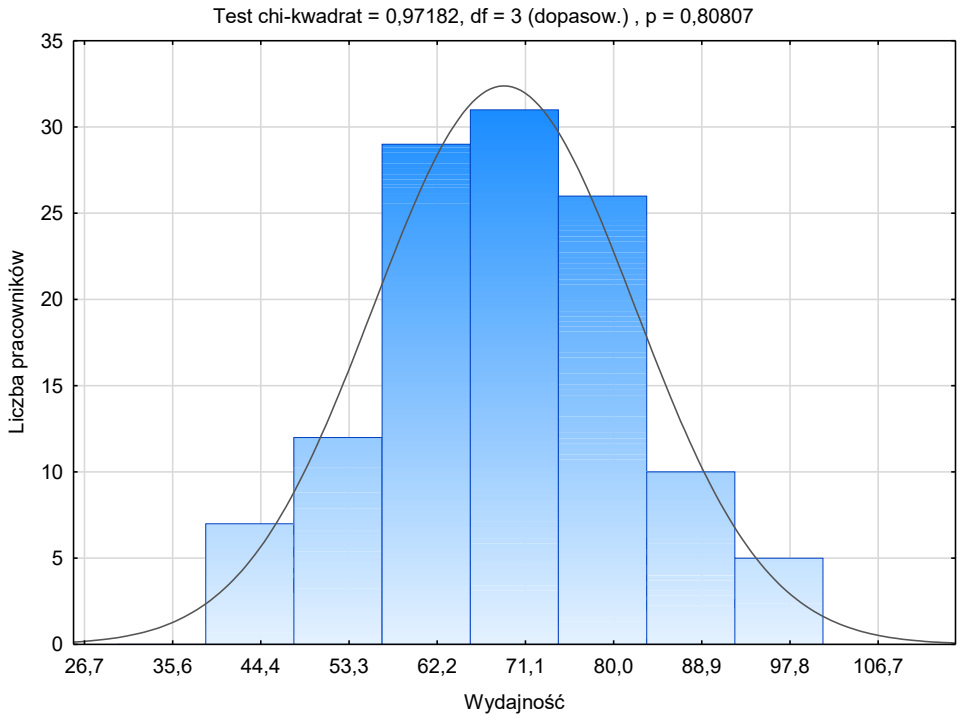
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Test t dla pojedynczej próby***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) i wpisano wartość odniesienia równą 70.

Średnia liczba zrealizowanych projektów obliczona bezpośrednio z próby badawczej wyniosła niemal 69 projektu w miesiącu. Wartość tę porównano z odgórnie założoną normą teoretyczną, która wynosiła 70 zrealizowanych projektów w miesiącu. W wyniku przeprowadzonej procedury otrzymano prawdopodobieństwo testowe $p = 0{,}391$. Ponieważ uzyskany wynik jest większy od standardowo przyjętego poziomu istotności α wynoszącego $0{,}05$, oznacza to brak podstaw do odrzucenia hipotezy zerowej. Ostatecznie stwierdza się, że średnia wydajność uzyskana z próby pracowników nie różni się w sposób istotny od założonej normy efektywności, a drobne odchylenie ma charakter czysto przypadkowy.

W celu formalnego sprawdzenia założeń niezbędnych do poprawnego przeprowadzenia parametrycznego testu t-Studenta, zweryfikowano założenie zgodności miesięcznej wydajności pracowników z rozkładem normalnym. Ze względu na dużą liczebność badanej próby, wynoszącą 120 osób, do weryfikacji tego kryterium zastosowano nieparametryczny test zgodności chi-kwadrat. Zgodnie z procedurą badawczą, sformułowano hipotezę zerową, że rozkład liczby zrealizowanych projektów w miesiącu w populacji generalnej jest zgodny

z rozkładem normalnym. Jako hipotezę alternatywną sformułowano sąd stwierdzający, że empiryczny rozkład wydajności pracowników różni się w sposób istotny od rozkładu normalnego. Wyniki tej procedury weryfikacyjnej zostały przedstawione graficznie oraz liczbowo na Rys. 4.



Rys. 4. Histogram rozkładu wydajności pracowników z nałożoną krzywą rozkładu normalnego oraz wynikami testu zgodności chi-kwadrat
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Dopasowanie rozkładu → Rozkład normalny → Wykres rozkładu obserwowanego i oczekiwanego, wybranie zmiennej, a w razie braku zgodności (poziom p<0,05) w zakładce *Parametry* dopasowanie np. liczby kategorii i powrót do zakładki *Podstawowe* i ponowny wybór wykresu.

W wyniku przeprowadzonego testu zgodności chi-kwadrat uzyskano poziom istotności $p = 0{,}808$, który jest większy od przyjętego poziomu istotności $\alpha = 0{,}05$. Wynik ten oznacza brak podstaw do odrzucenia hipotezy zerowej. Rozkład wydajności pracowników nie różni się zatem istotnie od rozkładu normalnego, co pozwala stwierdzić, że nie ma podstaw do uznania, że założenie o normalności cechy nie zostało spełnione.

W kolejnej analizie przyjęto za cel zweryfikowanie, czy przeciętna miesięczna wydajność pracowników, mierzona liczbą zrealizowanych projektów w miesiącu, różni się w sposób istotny statystycznie w zależności od przyjętej formy pracy. Wyniki przeprowadzonego testu t-Studenta dla dwóch prób niezależnych zestawiono w Tabeli 4.

Tabela 4. Wyniki testu t-Studenta dla dwóch prób

Zmienna	Średnia Zdalny	Średnia Stacjonarny	t	df	p	N ważnych Zdalny	N ważnych Stacjonarny	Odch.std Zdalny	Odch.std Stacjonarny	iloraz F	p wariancje	Levene'a	df Levene'a	p Levene'a
Wydajność	69,6	67,9	0,720	118	0,473	74	46	12,71	13,88	1,192	0,498	0,610	118	0,436

Źródło: opracowanie własne przy użyciu programu Statistica.

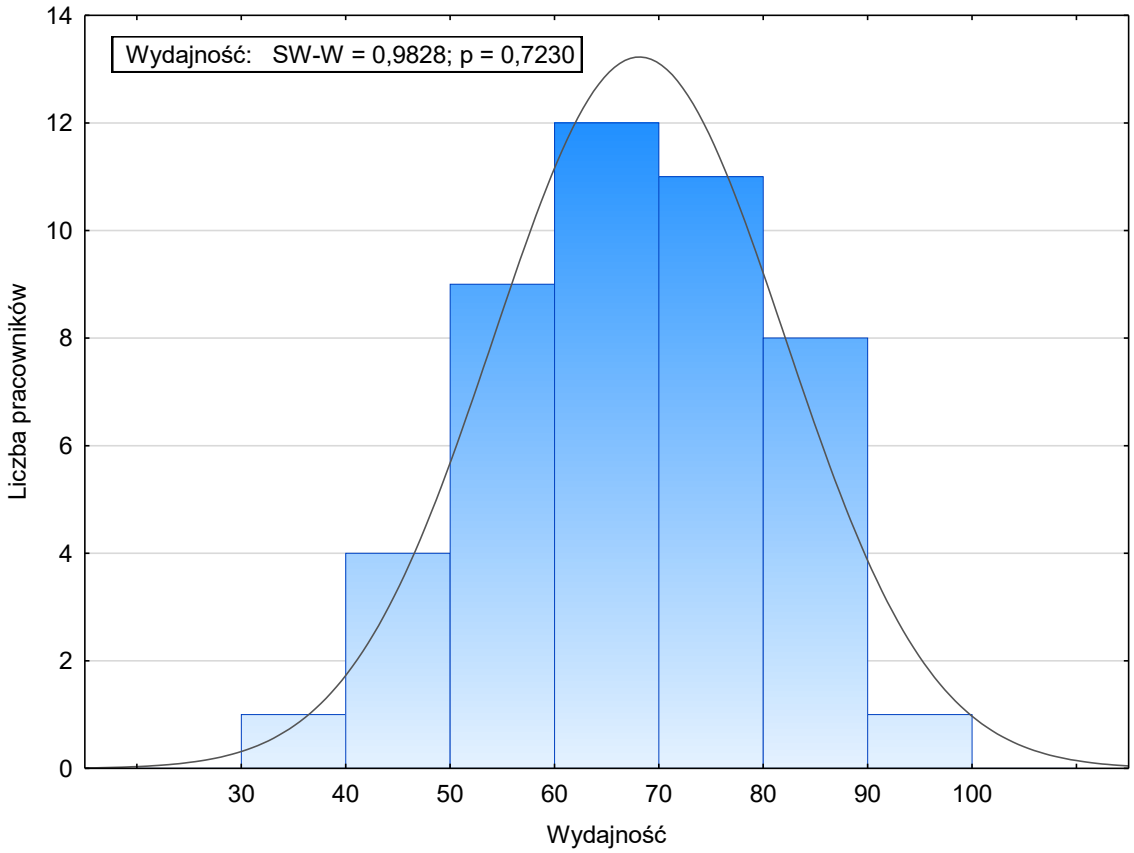
Zakładka ***Statystyki podstawowe → Test t dla prób niezależnych (wzgl. grup) próby***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) oraz grupującą (tu *Forma pracy*) oraz w zakładce *Opcje* zaznaczono test Levene'a.

Realizując powyższy cel, sformułowano hipotezę zerową mówiącą o tym, że średnia wydajność pracowników wykonujących obowiązki w formie stacjonarnej jest taka sama jak średnia wydajność pracowników wykonujących obowiązki w formie zdalnej. Sformułowano hipotezę alternatywną, mówiącą o tym, że średnie te różnią się między sobą.

Analizując otrzymane wyniki zaprezentowane w Tabeli 4. należy zauważyć, że średnia wydajność pracowników pracujących stacjonarnie wynosiła ponad 68 projektów w miesiącu, zaś pracujących zdalnie ponad 69 projektów w miesiącu. Podana różnica nie jest istotna, o czym świadczy poziom $p = 0{,}473$. Jest on większy od przyjętego poziomu istotności $\alpha = 0{,}05$, więc nie ma podstaw do odrzucenia hipotezy o równości średnich. Zatem nie ma podstaw uważać, że średnie wydajności w obu formach pracy różnią się między sobą.

Warunkiem właściwego stosowania testu t porównania dwóch średnich jest zgodność z rozkładem normalnym w obu grupach (pomiarach wydajności dla obu form pracy) oraz jednorodność wariancji w analizowanych podpopulacjach. Sprawdzając założenie o jednorodność wariancji postawiono hipotezę zerową o równości wariancji wydajności w obu formach pracy wobec hipotezy alternatywnej, że jedna z wariancji jest większa. Uzyskany w teście Levene'a poziom $p = 0{,}436$ jest większy od przyjętego poziomu istotności $\alpha = 0{,}05$, co oznacza brak podstaw do odrzucenia hipotezy zerowej i potwierdza, że nie można uznać, że założenie o jednorodności wariancji nie zostało spełnione.

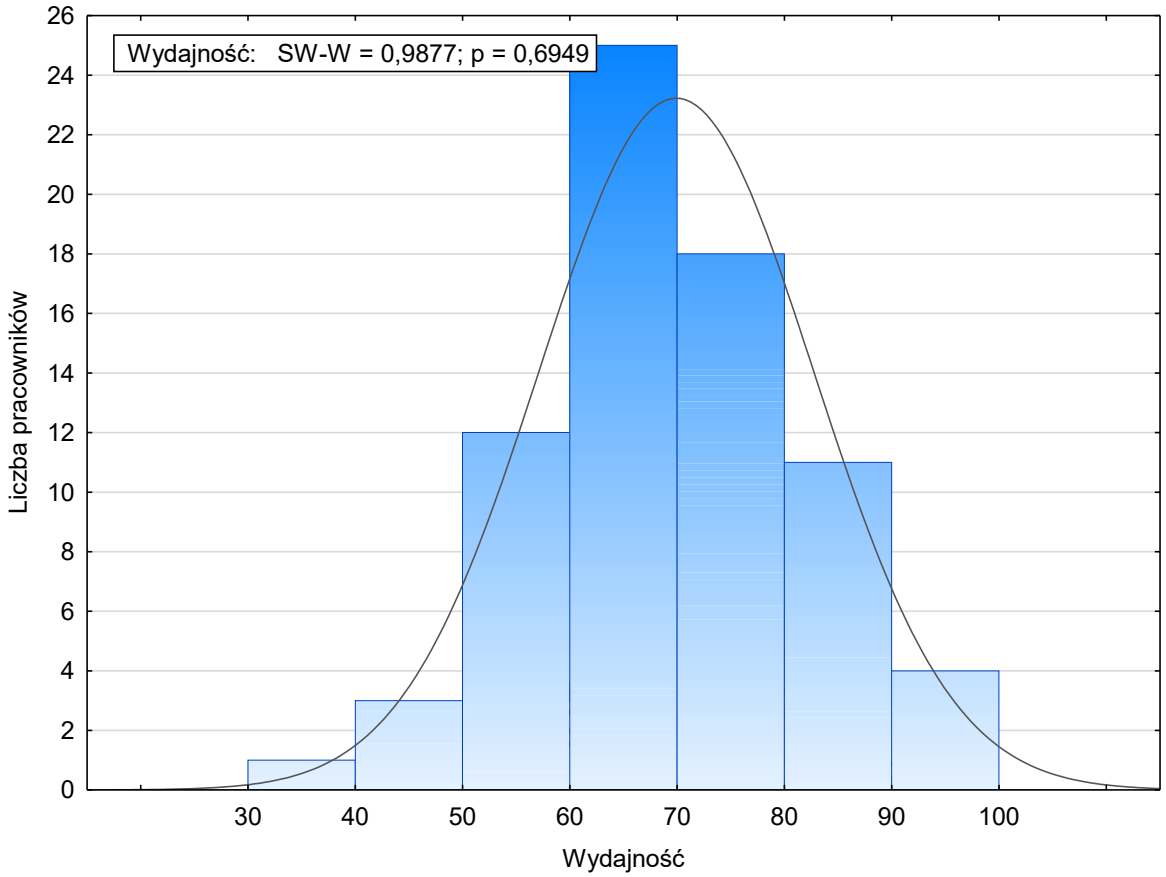
W celu oceny stopnia dopasowania danych empirycznych do modelowej krzywej dzwonowej Gaussa przeprowadzono, ze względu na próby liczące kilkadziesiąt pomiarów test Shapiro-Wilka, którego wyniki przedstawiono odpowiednio na Rys. 5 oraz Rys. 6 dla grupy wykonującej obowiązki w formie stacjonarnej oraz w formie zdalnej.



Rys. 5. Histogram rozkładu wydajności dla formy pracy stacjonarnej z nałożoną krzywą rozkładu normalnego oraz wynikami testu Shapiro-Wilka
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogram, wybranie zmiennej (tu *Wydajność*), w zakładce *Więcej* zaznaczony test Shapiro-Wilka, wybranie przycisku *Grupami* i wskazanie zmiennej grupującej (tu *Forma pracy*).

W pierwszym kroku postawiono hipotezę zerową mówiącą, że rozkład pomiarów wydajności pracowników wykonujących obowiązki stacjonarnie jest zgodny z rozkładem normalnym, wobec hipotezy alternatywnej, że nie jest zgodny z rozkładem normalnym. Otrzymano poziom $p = 0{,}723$, który jest większy od przyjętego poziomu istotności $\alpha = 0{,}05$, co oznacza, że nie ma podstaw uważać, że rozkład liczby zrealizowanych projektów w miesiącu dla formy stacjonarnej nie jest zgodny z rozkładem normalnym, czyli nie mamy podstaw by uważać, ze założenie nie zostało spełnione.



Rys. 6. Histogram rozkładu wydajności dla formy pracy zdalnej z nałożoną krzywą rozkładu normalnego oraz wynikami testu Shapiro-Wilka
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogram, wybranie zmiennej (tu *Wydajność*), w zakładce *Więcej* zaznaczony test Shapiro-Wilka, wybranie przycisku *Grupami* i wskazanie zmiennej grupującej (tu *Forma pracy*).

W kolejnym kroku postawiono hipotezę zerową mówiącą, że rozkład pomiarów wydajności pracowników wykonujących obowiązki zdalnie jest zgodny z rozkładem normalnym, wobec hipotezy alternatywnej, że nie jest zgodny z rozkładem normalnym. Otrzymano poziom $p = 0{,}695$, który jest większy od przyjętego poziomu istotności $\alpha = 0{,}05$, co oznacza, że nie ma podstaw uważać, że rozkład liczby zrealizowanych projektów w miesiącu dla formy zdalnej nie jest zgodny z rozkładem normalnym, czyli nie mamy podstaw by uważać, że założenie nie zostało spełnione.

Celem kolejnej analizy było zweryfikowanie czy poziom autonomii deklarowany przez pracowników różni się w sposób istotny statystycznie w zależności od przyjętej formy pracy (zdalnej lub stacjonarnej). Ponieważ zmienna Autonomia ma charakter porządkowy, do analizy porównawczej zastosowano nieparametryczny test U Manna-Whitneya dla dwóch prób niezależnych. Wyniki tej procedury zestawiono w Tabeli 5.

Tabela 5. Wyniki testu U Manna-Whitneya

Zmienna	Sum.rang Zdalny	Sum.rang Stacjonarny	U	Z	p
Autonomia	4891	2369	1288	2,232	0,026

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Nieparametryczne → Porównanie dwóch prób niezależnych (grup)***, w oknie konfiguracji wybrano zmienną zależną (tu *Autonomia*) oraz grupującą (tu *Forma pracy*).

Postawiono hipotezę zerową mówiącą o tym, że postać rozkładu poziomu autonomii pracowników wykonujących obowiązki w formie zdalnej jest taka sama jak postać rozkładu poziomu autonomii pracowników wykonujących obowiązki w formie stacjonarnej. Sformułowano hipotezę alternatywną, mówiącą o tym, że rozkłady te różnią się od siebie.

Analizując otrzymane wyniki zaprezentowane w Tabeli 5 należy zauważyć, że suma rang dla pracowników pracujących zdalnie wynosiła 4891, zaś dla pracowników pracujących stacjonarnie 2369. Podana różnica w strukturze rangowej jest istotna statystycznie, o czym świadczy uzyskany poziom $p = 0{,}026$. Otrzymany poziom p jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$, więc należy odrzucić hipotezę o równości rozkładów. Zatem postać rozkładu oraz ogólny poziom autonomii różnią się w zależności od formy wykonywania pracy.

W celu wizualizacji struktury ocen oraz porównania rozkładów zmiennej Autonomia pomiędzy pracownikami wykonującymi obowiązki w formie zdalnej a pracującymi stacjonarnie, na Rys. 7 przedstawiono wykres ramka-wąsy.

Rys. 7. Wykres ramka-wąsy poziomu autonomii w zależności od formy pracy
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Nieparametryczne → Porównanie dwóch prób niezależnych (grup)***, w oknie konfiguracji wybrano zmienną zależną (tu *Autonomia*) oraz grupującą (tu *Forma pracy*) oraz wybranie *Wykres ramka-wąsy dla grup*.

Analiza graficzna zaprezentowana na wykresie (Rys. 7.) potwierdza wcześniejsze wyniki testu U Manna-Whitneya o istotnych różnicach między grupami. Co prawda mediana poziomu autonomii w obu podgrupach jest identyczna i wynosi 3,0, to prostokątne ramki, reprezentujące zakres od dwudziestego piątego do siedemdziesiątego piątego centyla (środkowe 50% odpowiedzi), w obu przypadkach są odmienne – u osób pracujących zdalnie jest to zakres od 3,0 do 4,0, a u pracujących stacjonarnie – od 2,0 do 4,0. Zarówno wartości skrajne (od 1,0 do 5,0), jak i ogólny rozstęp danych są w obu grupach identyczne, jednak przesunięcie dolnej granicy kwartylnej w grupie pracowników zdalnych wizualizuje zaobserwowaną w teście statystycznym istotną różnicę w strukturze rozkładów.

7. Analiza wariancji (ANOVA) oraz jej nieparametryczna alternatywa

7.1. Jednoczynnikowa analiza wariancji (ANOVA)

Klasyczna analiza wariancji, znana powszechnie pod akronimem ANOVA (od angielskiego *Analysis of Variance*), została opracowana przez Ronalda Fishera jako zaawansowane narzędzie wnioskowania parametrycznego dedykowane do jednoczesnego

porównywania średnich wartości cechy ciągłej w wielu niezależnych populacjach. Choć celem procedury jest weryfikacja hipotezy zerowej o równości średnich w k grupach, czyli $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ wobec hipotezy alternatywnej $H_1: \exists_{(i \neq j)} \mu_i \neq \mu_j$ stwierdzającej istnienie przynajmniej jednej pary średnich różniących się w sposób istotny, to matematyczna istota testu opiera się na analizie zmian poziomu wariancji. Gdyby badacz próbował porównywać wiele prób za pomocą seryjnych testów t-Studenta, doprowadziłoby to do gwałtownego wzrostu prawdopodobieństwa popełnienia błędu I rodzaju (kumulacja poziomu α). ANOVA rozwiązuje ten problem poprzez zastosowanie jednego, globalnego sprawdzianu opartego na dekompozycji całkowitej sumy kwadratów odchyleń.

Istotą algorytmu ANOVA realizowanego w programie Statistica jest podział całkowitej zmienności badanej zmiennej zależnej, opisywanej przez całkowitą sumę kwadratów (SS_T – *Total Sum of Squares*), na dwa niezależne komponenty: zmienność międzygrupową (SS_B – *Between-group Sum of Squares*) oraz zmienność wewnątrzgrupową (SS_W – *Within-group Sum of Squares*, zwaną sumą kwadratów dla błędu). Zmienność międzygrupowa odzwierciedla odchylenia średnich grupowych od średniej ogólnej i reprezentuje wpływ czynnika klasyfikującego (systematycznego), natomiast zmienność wewnątrzgrupowa mierzy rozproszenie obserwacji wokół ich własnych średnich grupowych, co stanowi odzwierciedlenie czystego szumu losowego. Całkowite równanie dekompozycji przyjmuje postać $SS_T = SS_B + SS_W$.

Aby dokonać poprawnego porównania, program Statistica przelicza surowe sumy kwadratów na tzw. średnie kwadraty odchyleń (MS – *Mean Squares*), dzieląc je przez odpowiadające im liczby stopni swobody. Dla zmienności międzygrupowej średni kwadrat wynosi $MS_B = SS_B/(k-1)$, gdzie k to liczba grup, a dla zmienności wewnątrzgrupowej wynosi $MS_W = SS_W/(n-k)$, gdzie n to całkowita liczebność próby łącznej. Sprawdzianem hipotezy zerowej jest statystyka F_{obl}, definiowana jako stosunek średniego kwadratu międzygrupowego do wewnątrzgrupowego, co wyraża się wzorem:

$$F_{obl} = \frac{MS_B}{MS_W} = \frac{\frac{SS_B}{k-1}}{\frac{SS_W}{n-k}}$$

Statystyka ta przy założeniu prawdziwości hipotezy zerowej podlega teoretycznemu rozkładowi F Fishera-Snedecora z liczbą stopni swobody licznika $v_1 = k-1$ oraz stopni swobody mianownika $v_2 = n-k$. Jeśli zmienność generowana przez czynnik systematyczny (licznik) jest znacznie większa niż szum losowy (mianownik), statystyka F_{obl} przyjmuje

wysokie wartości, co skutkuje wygenerowaniem przez program bardzo małej p-wartości ($p < \alpha$) i prowadzi do odrzucenia hipotezy H_0 o równości średnich.

7.2. Analiza porównań wielokrotnych – testy post-hoc

Weryfikacja globalnej hipotezy zerowej w jednoczynnikowej analizie wariancji (ANOVA) informuje badacza wyłącznie o tym, czy zaobserwowane różnice między średnimi lub medianami k porównywanych grup są statystycznie istotne. Odrzucenie hipotezy o identyczności rozkładów nie wskazuje jednak, które konkretnie pary grup różnią się między sobą. Próba rozwiązania tego problemu poprzez seryjne uruchamianie klasycznych testów dla dwóch prób (np. testu t-Studenta) prowadzi do matematycznego błędu kumulacji prawdopodobieństwa błędu I rodzaju, gdzie rzeczywisty poziom istotności dla całej rodziny porównań drastycznie wzrasta ponad założone $\alpha = 0{,}05$. Narzędziem dedykowanym do bezpiecznej, szczegółowej identyfikacji różnic międzygrupowych przy zachowaniu pełnej kontroli nad globalnym poziomem błędu są testy porównań wielokrotnych, zwane testami post-hoc (po fakcie).

Historycznie pierwszym narzędziem jest test Najmniejszej Istotnej Różnicy (NIR) Fishera. Test ten charakteryzuje się najwyższą mocą (zdolnością do wykrywania minimalnych różnic), ponieważ polega na seryjnym wykonywaniu testów t-Studenta dla par średnich, wykorzystując w strukturze mianownika globalny średni kwadrat błędu wewnątrzgrupowego (MS_W) z modelu ANOVA. Krytyczną wartość graniczną, powyżej której różnica średnich empirycznych uznawana jest za istotną, algorytm wyznacza ze wzoru:

$$NIR = t_{\alpha,v}\sqrt{MS_W\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

w którym $t_{\alpha,v}$ to wartość z rozkładu t-Studenta dla stopni swobody błędu $v = n - k$, a symbole n_i, n_j oznaczają liczebności podgrup. Ponieważ test NIR nie posiada mechanizmu korekty błędu dla wielokrotnych porównań, jego stosowanie jest dopuszczalne w metodologii wyłącznie wtedy, gdy globalny test F analizy wariancji okazał się statystycznie istotny.

W celu pełnego zabezpieczenia przed błędem I rodzaju, program Statistica implementuje bardziej konserwatywne testy, takie jak test Tukeya (HSD) oraz test Scheffégo. Test Tukeya bazuje na rozkładzie studentyzowanej rozpiętości i służy do precyzyjnego porównywania wszystkich możliwych par średnich. Bardziej konserwatywnym rozwiązaniem jest test Scheffégo, który pozwala na testowanie nie tylko par, ale dowolnych kontrastów (kombinacji

liniowych średnich), wyznaczając najszersze przedziały tolerancji, co minimalizuje ryzyko popełnienia błędu I rodzaju kosztem mocy testu.

7.2. Weryfikacja założeń klasycznej ANOVA

Prawidłowe stosowanie jednoczynnikowej analizy wariancji ANOVA w programie Statistica wymaga spełnienia założeń, do których należą: mierzalność zmiennej zależnej na silnej skali (przedziałowej lub ilorazowej), niezależność obserwacji w analizowanych grupach, zgodność rozkładu badanej cechy z rozkładem normalnym wewnątrz każdej z wydzielonych podpopulacji oraz jednorodność (homoskedastyczność), czyli równość wariancji warunkowych na wszystkich poziomach analizowanego czynnika. O ile małe odstępstwa od normalności rozkładu są przez algorytm ANOVA relatywnie dobrze tolerowane (szczególnie przy dużych i równolicznych próbach), o tyle naruszenie założenia o równości wariancji drastycznie zniekształca wyniki testu F, prowadząc do fałszywych wniosków.

Do weryfikacji tego założenia program Statistica udostępnia dwa alternatywne narzędzia diagnostyczne: test Bartletta oraz test Levene'a. Klasyczny test Bartletta opiera się na porównaniu średniej arytmetycznej wariancji grupowych z ich średnią geometryczną, a statystyka testowa B jest wyliczana przez oprogramowanie za pomocą formuły:

$$B = \exp\left(\frac{M}{c}\right) \quad \text{gdzie} \quad M = (n-k)\ln\left(\frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)\,\hat{S}_{y_i}^2\right) - \sum_{i=1}^{k}(n_i-1)\ln\hat{S}_{y_i}^2$$

gdzie symbol n_i oznacza liczebność i-tej grupy, $\hat{S}_{y_i}^2$ to nieobciążony estymator wariancji w i-tej podpróbie, k to liczba grup, a c stanowi stałą poprawkową zwiększającą dokładność aproksymacji. Główną wadą testu Bartletta jest jego wysoka wrażliwość na odstępstwa od rozkładu normalnego – jeśli dane w grupach mają rozkład asymetryczny, test ten może błędnie wykazać brak jednorodności wariancji.

Problem ten eliminuje test Levene'a, który charakteryzuje się wysoką odpornością na naruszenie założenia o normalności. Matematyczna istota testu Levene'a polega na przekształceniu surowych danych i uruchomieniu standardowej procedury ANOVA na obliczonych odchyleniach bezwzględnych obserwacji y_{ij} od średnich swoich grup $\overline{y}_i$, co zapisujemy jako $z_{ij} = |y_{ij} - \overline{y}_i|$. Statystyka testowa W (będąca odpowiednikiem statystyki F) wyliczana jest ze wzoru

$$W = \frac{n-k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i \left(\overline{z}_{i\cdot} - \overline{z}_{\cdot\cdot}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(z_{ij} - \overline{z}_{i\cdot}\right)^2}$$

gdzie $\overline{z}_{i\cdot}$ oznacza średnią z odcholeń w i-tej grupie, a $\overline{z}_{\cdot\cdot}$ stanowi średnią ogólną ze wszystkich odchyleń.

Zarówno w teście Bartletta, jak i Levene'a reguła decyzyjna w programie Statistica jest tożsama: uzyskanie wysokiej p-wartości ($p > 0{,}05$) oznacza brak podstaw do odrzucenia hipotezy zerowej, co potwierdza homoskedastyczność i uprawnia do interpretacji klasycznej ANOVA. Wynik przeciwny ($p < 0{,}05$) nakazuje odrzucić założenie o równości wariancji, co zmusza badacza do zastosowania poprawki Welcha bądź przejścia do nieparametrycznego testu Kruskala-Wallisa.

7.3. Nieparametryczna alternatywa dla ANOVA – test Kruskala-Wallisa

W sytuacjach, gdy formalna weryfikacja założeń wykaże jawne naruszenie jednorodności wariancji za pomocą testu Bartletta lub Levene'a, gdy rozkłady w podgrupach charakteryzują się ekstremalną asymetrią, bądź gdy analizowana zmienna zależna została zmierzona wyłącznie na słabej skali porządkowej (rangowej), klasyczna analiza ANOVA traci swą miarodajność. Bezpieczną i powszechnie stosowaną alternatywą nieparametryczną w programie Statistica jest wówczas test Kruskala-Wallisa, który stanowi bezpośrednie uogólnienie testu U Manna-Whitneya dla struktury wielu próbek niezależnych. Założenia tego testu są minimalne i wymagają jedynie, aby zmienna zależna była mierzona na skali co najmniej porządkowej oraz aby obserwacje w analizowanych grupach były względem siebie niezależne.

Matematyczna istota testu Kruskala-Wallisa polega na całkowitym porzuceniu surowych wartości liczbowych na rzecz ich rang. Algorytm łączy wszystkie obserwacje ze wszystkich k grup w jeden wspólny szereg szczegółowy, porządkuje je rosnąco i przypisuje im kolejne rangi od 1 do n. Następnie oblicza się sumy rang dla każdej z wyodrębnionych grup, oznaczane jako R_i. Sprawdzianem hipotezy zerowej, mówiącej o tym, że wszystkie próby pochodzą z tej samej populacji (lub z populacji o identycznych medianach), jest statystyka H, wyliczana przez zgodnie z wzorem:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

gdzie n_i to liczebność i-tej grupy, a n stanowi całkowitą liczebność połączonych próbek. Statystyka H przy założeniu prawdziwości hipotezy zerowej podlega rozkładowi chi-kwadrat o $k-1$ stopniach swobody. W sytuacji, gdy w danych występuje duża liczba identycznych

wartości, algorytmy programu Statistica automatycznie wprowadzają poprawkę na tzw. rangi wiązane, modyfikując mianownik wzoru. Uzyskanie wartości $p < \alpha$ informuje badacza o istnieniu istotnych statystycznie różnic w poziomach położenia (medianach) analizowanych grup.

W odniesieniu do nieparametrycznej analizy ANOVA, realizowanej za pomocą testu Kruskala-Wallisa, pojęcie testu porównań szczegółowych (post-hoc) odnosi się najczęściej do procedury opracowanej przez O. J. Dunna. Gdy globalny sprawdzian wykaże istotne statystycznie zróżnicowanie rozkładów ($p < 0{,}05$), badacz staje przed koniecznością precyzyjnego wskazania odmiennych grup bez ryzyka sztucznego zawyżenia błędu I rodzaju. Algorytm nieparametrycznego testu post-hoc w programie Statistica opiera się na porównywaniu bezwzględnych różnic między średnimi rangami poszczególnych podprób, wyznaczanymi na etapie globalnego testowania rangowego. Dla każdej badanej pary grup wyliczana jest standaryzowana statystyka Z, której mianownik uwzględnia poprawkę na tzw. rangi wiązane oraz liczebności porównywanych podgrup. Aby skutecznie kontrolować globalny poziom istotności ($\alpha = 0{,}05$) dla całej rodziny jednoczesnych porównań, program automatycznie implementuje klasyczną poprawkę Bonferroniego, polegającą na proporcjonalnym dzieleniu nominalnego prawdopodobieństwa przez liczbę wszystkich możliwych zestawień parowych. W rezultacie w oknie raportu otrzymuje się macierz skorygowanych p-wartości, w której czerwone wyróżnienia jednoznacznie wskazują te kombinacje grup, których mediany różnią się w sposób statystycznie istotny i nieprzypadkowy.

Przykład

Celem przeprowadzonego badania było zweryfikowanie, czy przeciętna miesięczna wydajność pracowników, mierzona liczbą zrealizowanych projektów w miesiącu, różni się w sposób istotny statystycznie w zależności od działu, w którym są oni zatrudnieni. Analizą objęto trzy grupy pracowników reprezentujące dział IT, HR oraz dział Sprzedaży. Ponieważ badanie dotyczyło porównania wartości średnich w więcej niż dwóch populacjach, jako narzędzie weryfikacji hipotez badawczych zastosowano jednoczynnikową analizę wariancji.

W teście ANOVA postawiono hipotezę zerową mówiącą o tym, że średnia wydajność pracowników w działach IT, HR oraz Sprzedaż jest taka sama ($\mu_1 = \mu_2 = \mu_3$). Sformułowano hipotezę alternatywną, mówiącą o tym, że co najmniej dwa działy różnią się między sobą średnią wydajnością. Wyniki testu ANOVA przedstawiono w tabeli 6.

Tabela 6. Wyniki jednoczynnikowej analizy wariancji (ANOVA)

Zmienna	SS	df	MS	SS	df	MS	F	p
Wydajność	4986,015	2	2493,007	15559,85	117	132,990	18,7458	0,000

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Przekroje, prosta ANOVA***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) oraz grupującą (tu *Dział*) oraz wybranie w zakładce *Podstawowe* przycisku *Analiza wariancji.*

Uzyskano poziom $p < 0{,}001$, który jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$, co prowadzi do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Zatem średnie wydajności w przynajmniej dwóch działach różnią się między sobą w sposób istotny statystycznie. Gdyby poziom p był większy od 0,05, wówczas nie byłoby podstaw do odrzucenia hipotezy zerowej, co oznaczałoby, że struktura organizacyjna nie różnicuje efektywności personelu.

Samo odrzucenie hipotezy o równości średnich w modelu ANOVA nie wskazuje jednak, pomiędzy którymi dokładnie działami występują wykryte różnice. W celu ich precyzyjnego zlokalizowania przeprowadzono analizę porównań wielokrotnych za pomocą testu post-hoc, którego szczegółową macierz poziomów istotności przedstawiono w Tabeli 7.

Tabela 7. Macierz poziomów istotności p dla porównań wielokrotnych (test NIR)

Dział		{1} M=78,324	{2} M=63,000	{3} M=66,821
IT	{1}		0,000	0,000
HR	{2}	0,000		0,135
Sprzedaz	{3}	0,000	0,135	

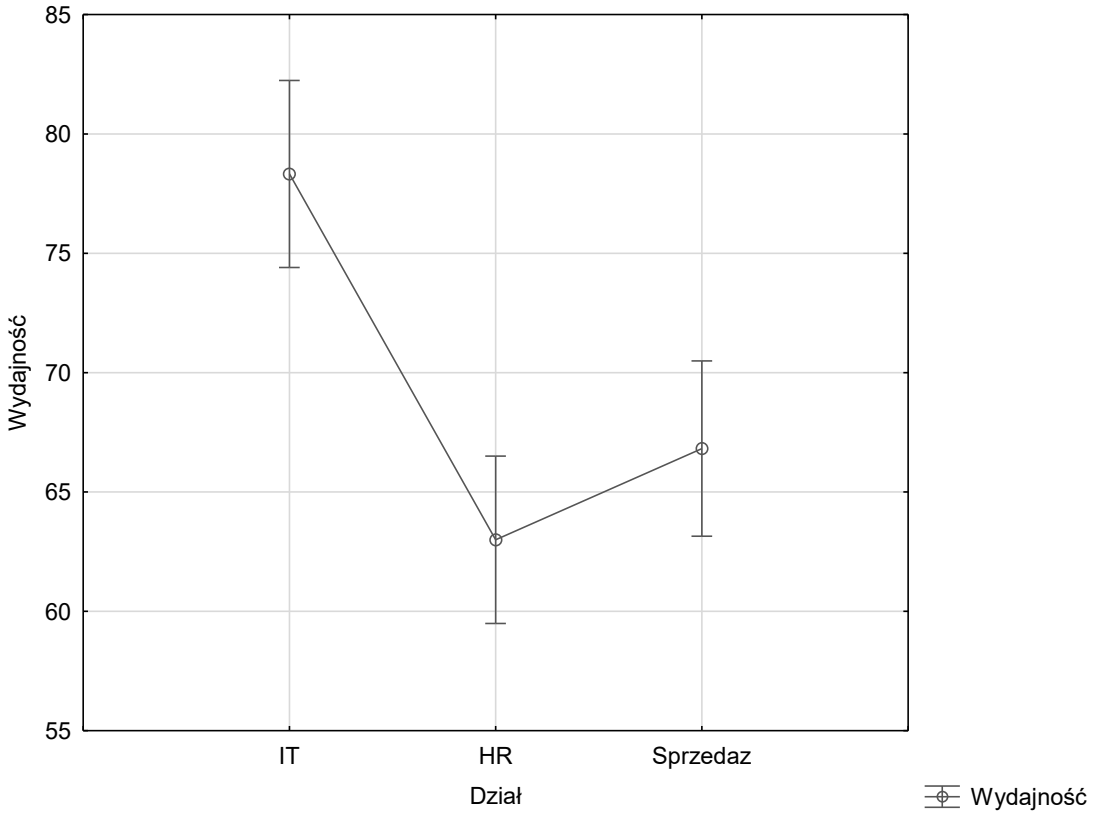
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Przekroje, prosta ANOVA***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) oraz grupującą (tu *Dział*) oraz wybranie zakładki *Post-hoc* i w tej zakładce *Test NIR i porównanie zaplanowane.*

Biorąc pod uwagę średnie grupowe, najwyższą wydajnością charakteryzuje się dział IT, gdzie średnia wyniosła ponad 78 projektów, na drugim miejscu uplasował się dział Sprzedaż ze średnią bliską 67 projektów, a najniższy wynik odnotowano w dziale HR, gdzie średnia ukształtowała się na poziomie 63 projektów w miesiącu. Analizując bezpośrednie zestawienia par grup w Tabeli 7, w przypadku porównania działu IT z działem HR otrzymano poziom $p < 0{,}001$. Ponieważ wartość ta jest mniejsza od 0,05, różnica między średnią wydajnością tych jednostek jest istotna statystycznie na korzyść działu IT. Podobną zależność zaobserwowano

przy porównaniu działu IT z działem Sprzedaż, gdzie uzyskany poziom $p < 0{,}001$ również okazał się mniejszy od poziomu istotności $\alpha = 0{,}05$, co potwierdza istotną statystycznie wyższą efektywność pracowników sektora technologicznego. Ostatnie porównanie, dotyczące relacji między działem HR a działem Sprzedaż, wykazało poziom $p = 0{,}135$. Jest on większy od przyjętego poziomu istotności $\alpha = 0{,}05$, co oznacza brak istotnych statystycznie różnic pomiędzy wydajnością pracowników tych dwóch pionów. Podsumowując, uzyskane wyniki pozwalają stwierdzić, że dział IT generuje znacząco wyższą produktywność niż pozostałe działy, podczas gdy pracownicy HR i Sprzedaży prezentują zbliżony potencjał, a występująca między nimi niewielka rozbieżność ma charakter czysto przypadkowy.

Dopełnieniem przeprowadzonej analizy wariancji jest graficzna prezentacja średnich grupowych wraz z przedziałami ufności, którą przedstawiono na Rysunku 8.



Rys. 8. Wykres średnich wydajności z 95% przedziałem ufności w zależności od działu
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Przekroje, prosta ANOVA***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) oraz grupującą (tu *Dział*) oraz wybranie w zakładce *Podstawowe* przycisku *Wykresy interakcji.*

Analiza graficzna zaprezentowana na wykresie (Rys. 8.) stanowi bezpośrednie odzwierciedlenie oraz wizualne potwierdzenie wniosków uzyskanych za pomocą testów post-

hoc. Punkt położony najwyżej na osi pionowej jednoznacznie wskazuje, że najwyższą przeciętną produktywnością charakteryzuje się dział IT, którego wąs błędu (przedział ufności) jest całkowicie odseparowany i nie pokrywa się z przedziałami wyznaczonymi dla pozostałych dwóch jednostek organizacyjnych. Taka separacja graficzna wizualizuje bardzo wysoką istotność statystyczną różnic wykazaną w teście post-hoc. Z kolei punkty odpowiadające działom HR oraz Sprzedaż znajdują się na wyraźnie niższym poziomie. Co kluczowe, pionowe wąsy błędu reprezentujące 95% przedziały ufności dla tych dwóch działów w nachodzą na siebie. Wizualny efekt nakładania się przedziałów ufności stanowi graficzny dowód na to, że różnice w poziomach wydajności pomiędzy pracownikami działu HR a działu Sprzedaż są zjawiskiem czysto przypadkowym i nie noszą znamion istotności statystycznej, co jest w pełni spójne z wcześniejszym wynikiem prawdopodobieństwa testowego wynoszącym $p = 0{,}135$.

W ramach przeprowadzonej analizy wariancji, konieczne jest również formalne zweryfikowanie założeń parametrycznych, do których należą jednorodność wariancji pomiędzy analizowanymi działami oraz zgodność rozkładu badanej zmiennej z rozkładem normalnym w każdej z podgrup.

W celu sprawdzenia pierwszego z założeń, jakim jest homogeniczność wariancji, przeprowadzono test Levene'a, którego wyniki przedstawiono w Tabeli 8. Postawiono w nim hipotezę zerową zakładającą równość wariancji wydajności w trzech badanych działach, wobec hipotezy alternatywnej mówiącej, że wariancje te różnią się od siebie.

Tabela 8. Wyniki testu jednorodności wariancji Levene'a

Zmienna	SS	df	MS	SS	df	MS	F	p
Wydajność	40,830	2	20,415	5823,306	117	49,772	0,410	0,664

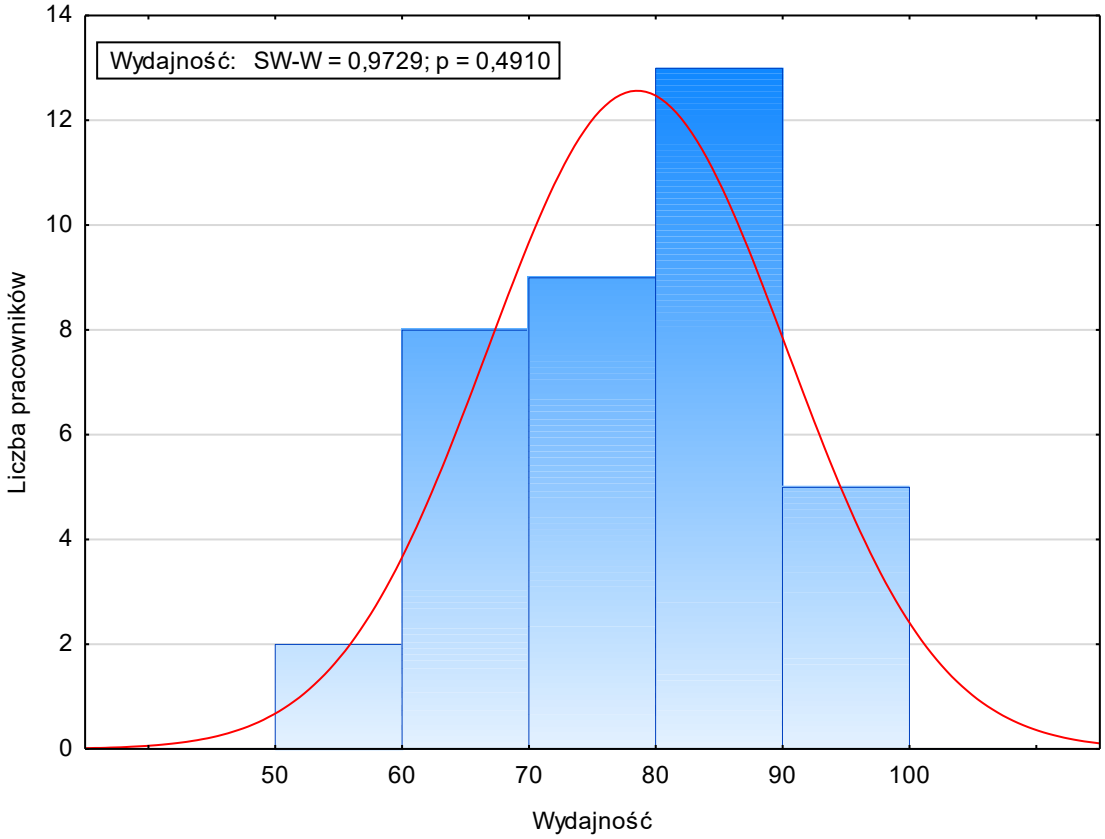
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Przekroje, prosta ANOVA***, w oknie konfiguracji wybrano zmienną zależną (tu *Wydajność*) oraz grupującą (tu *Dział*) oraz wybranie zakładki *Post-hoc* i przycisku *Test Levene'a.*

Wynik testu Levene'a dał poziom $p = 0{,}664$, który jest większy od przyjętego poziomu istotności $\alpha = 0{,}05$. W konsekwencji prowadzi to do wniosku, że brak jest podstaw do odrzucenia hipotezy zerowej, co potwierdza, że wariancje wewnątrz analizowanych działów nie różnią się istotnie.

Drugie z założeń, jakim była zgodność rozkładu badanej zmiennej z rozkładem normalnym w każdej z podgrup, ze względu na podpróby liczące po kilkadziesiąt pomiarów, w każdym z trzech działów przeprowadzono test Shapiro-Wilka. W procedurze tej dla każdej podpopulacji sformułowano hipotezę zerową mówiącą, że rozkład pomiarów wydajności jest

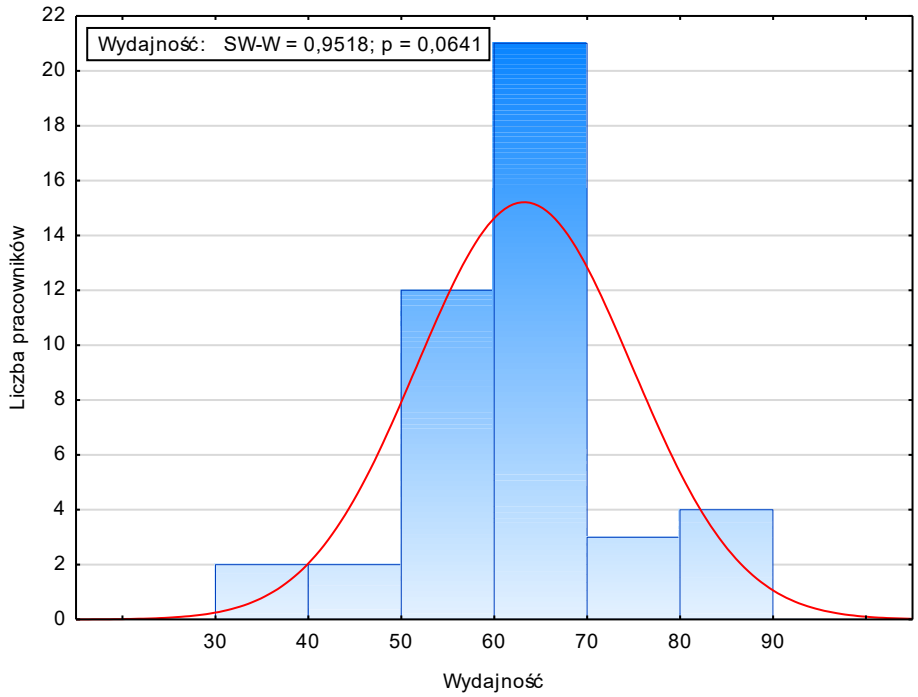
zgodny z rozkładem normalnym, wobec hipotezy alternatywnej, że nie jest on zgodny z rozkładem normalnym. Wyniki przedstawiono odpowiednio na rys. 9 dla IT, na rys. 10 dla działu HR oraz na rys. 11 dla działu Sprzedaży.



Rys. 9. Histogram rozkładu wydajności dla działu IT z nałożoną krzywą rozkładu normalnego oraz wynikami testu Shapiro-Wilka

Źródło: opracowanie własne przy użyciu programu Statistica.

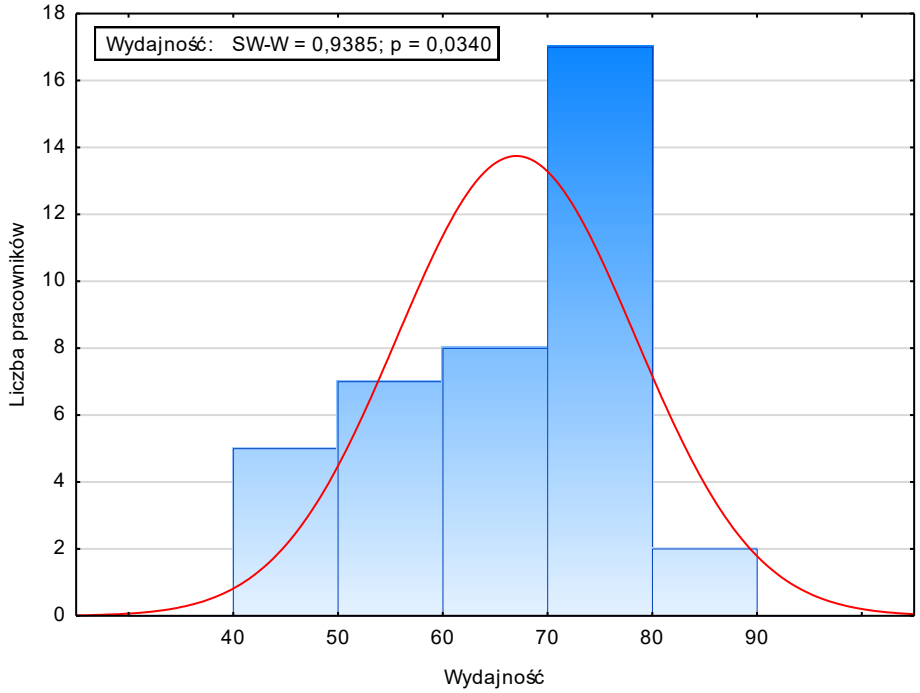
Zakładka Wykresy → Histogram, wybranie zmiennej (tu *Wydajność*), w zakładce *Więcej* zaznaczony test Shapiro-Wilka, wybranie przycisku *Grupami* i wskazanie zmiennej grupującej (tu *Dział*).



Rys. 10. Histogram rozkładu wydajności dla działu HR z nałożoną krzywą rozkładu normalnego oraz wynikami testu Shapiro-Wilka

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogram, wybranie zmiennej (tu *Wydajność*), w zakładce *Więcej* zaznaczony test Shapiro-Wilka, wybranie przycisku *Grupami* i wskazanie zmiennej grupującej (tu *Dział*).



Rys. 11. Histogram rozkładu wydajności dla działu Sprzedaż z nałożoną krzywą rozkładu normalnego oraz wynikami testu Shapiro-Wilka

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka Wykresy → Histogram, wybranie zmiennej (tu *Wydajność*), w zakładce *Więcej* zaznaczony test Shapiro-Wilka, wybranie przycisku *Grupami* i wskazanie zmiennej grupującej (tu *Dział*).

Dla pracowników działu IT uzyskano poziom $p = 0{,}491$, dla działu HR wartość wyniosła $p = 0{,}064$, W obu tych przypadkach wartości prawdopodobieństwa testowego są większe od przyjętego poziomu istotności $\alpha = 0{,}05$, co oznacza brak podstaw do odrzucenia hipotezy zerowej o zgodności z rozkładem normalnym. Odmienną sytuację odnotowano w dziale HR, dla którego wartość wyniosła $p = 0{,}034$. Ponieważ ten wynik jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$, w tym jednym przypadku odrzucamy hipotezę zerową na rzecz alternatywnej. Oznacza to, że rozkład miesięcznej wydajności wewnątrz działu HR różni się w sposób istotny od teoretycznego rozkładu normalnego, czyli założenie o normalności w tej podpopulacji nie zostało spełnione.

Niezbędna weryfikacja obu powyższych założeń wykazała, że choć wariancje pozostały homogeniczne, to brak normalności rozkładu w jednej z podgrup formalnie narusza kryteria klasycznej, parametrycznej procedury jednoczynnikowej analizy wariancji.

Celem przeprowadzonego kolejnego badania było zweryfikowanie czy rozkład poziomu autonomii deklarowanego przez pracowników różni się w sposób istotny statystycznie w zależności od działu, w którym są oni zatrudnieni. Analizą objęto trzy niezależne grupy pracowników reprezentujące dział IT, HR oraz dział Sprzedaży. Ponieważ zmienna Autonomia ma charakter jakościowy-porządkowy jako narzędzie weryfikacji hipotez badawczych zastosowano nieparametryczną alternatywę dla klasycznej analizy ANOVA — test Kruskala-Wallisa dla wielu prób niezależnych. Kompleksowe wyniki tego wnioskowania statystycznego, obejmujące test główny oraz macierz nieparametrycznych porównań wielokrotnych, zestawiono kolejno w Tabeli 9 oraz Tabeli 10.

Tabela 9. Wyniki nieparametrycznego testu Kruskala-Wallisa

Dział	Test Kruskala-Wallisa: H (2, N= 120) =10,80165 p = 0,0045			
	Kod	N ważnych	Suma Rang	Średnia Ranga
IT	1	37	2618	70,757
HR	2	44	2089	47,477
Sprzedaż	3	39	2553	65,461

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Nieparametryczne → Porównanie wielu prób niezależnych (grup)***, w oknie konfiguracji wybrano zmienną zależną (tu *Autonomia*) oraz grupującą (tu *Dział*).

W głównym teście Kruskala-Wallisa, zaprezentowanym w Tabeli 8, testowano hipotezę zerową mówiącą o tym, że postać rozkładu poziomu autonomii w działach IT, HR oraz Sprzedaż jest taka sama, wobec hipotezy alternatywnej stwierdzającej, że rozkłady te różnią się między sobą dla co najmniej dwóch działów. Uzyskano poziom $p = 0{,}005$. Ponieważ otrzymany poziom p jest mniejszy od przyjętego poziomu istotności $\alpha = 0{,}05$, obliguje to do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Zatem postać rozkładu oraz ogólny poziom autonomii różnią się między analizowanymi działami w sposób istotny statystycznie.

Samo odrzucenie hipotezy zerowej w teście Kruskala-Wallisa informuje jedynie o istnieniu ogólnych rozbieżności, nie wskazuje jednak, pomiędzy którymi konkretnie jednostkami organizacyjnymi one występują. W celu ich precyzyjnego zlokalizowania przeanalizowano średnie rangi przypisane poszczególnym grupom oraz macierz nieparametrycznych porównań wielokrotnych, zaprezentowaną w Tabeli 10.

Tabela 9. Macierz poziomów istotności p dla wielokrotnych porównań średnich rang

Dział	IT R: 70,757	HR R: 47,477	Sprzedaż R: 65,462
IT		0,008	1,000
HR	0,008		0,056
Sprzedaż	1,000	0,056	

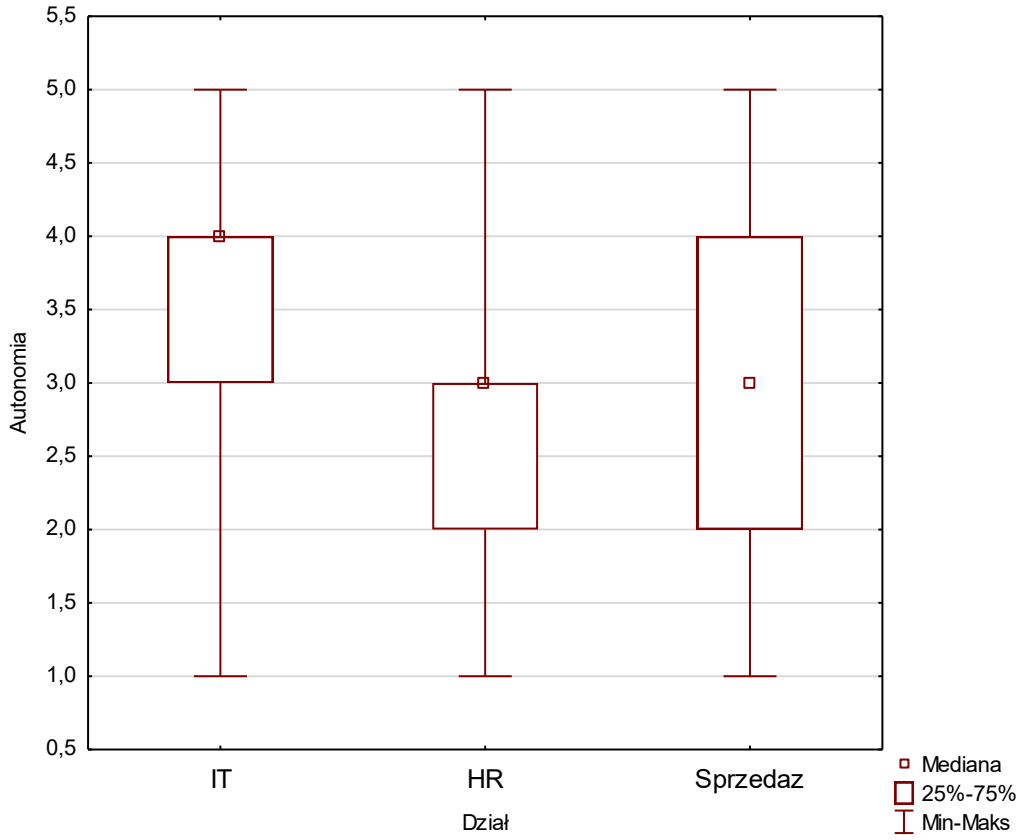
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Nieparametryczne → Porównanie wielu prób niezależnych (grup)***, w oknie konfiguracji wybrano zmienną zależną (tu *Autonomia*) oraz grupującą (tu *Dział*), przycisk *Wielokrotne porównanie średnich rang dla wszystkich prób.*

Analizując bezpośrednie zestawienia par działów zawarte w Tabeli 10, w przypadku porównania działu IT z działem HR otrzymano poziom p=0,008. Ponieważ wartość ta jest mniejsza od poziomu $\alpha = 0{,}05$, różnica w rozkładach autonomii pomiędzy działem IT a działem HR jest istotna statystycznie na korzyść sektora IT. Odmienną zależność zaobserwowano przy porównaniu działu IT z działem Sprzedaż, gdzie uzyskany poziom istotności ukształtował się na poziomie $p = 1{,}000$. Wynik ten jest znacznie większy od poziomu $\alpha = 0{,}05$, co oznacza brak istotnych różnic w poczuciu autonomii pomiędzy pracownikami tych dwóch pionów. Ostatnie porównanie, dotyczące relacji między działem HR a działem Sprzedaż, wykazało poziom $p = 0{,}056$. Wartość ta znajduje się minimalnie powyżej przyjętego kryterium istotności $\alpha = 0{,}05$, co formalnie oznacza brak podstaw do odrzucenia hipotezy zerowej i wskazuje, że rozkład autonomii w dziale HR nie różnicuje się w sposób

istotny statystycznie od rozkładu w dziale Sprzedaż, choć zaobserwowana tendencja lokuje się na samej granicy istotności. Podsumowując, przeprowadzone wnioskowanie nieparametryczne pozwala stwierdzić, że struktura organizacyjna przedsiębiorstwa różnicuje poczucie autonomii personelu, gdzie pracownicy działu IT oraz działu Sprzedaż wykazują zbliżony, wysoki poziom niezależności w pracy, podczas gdy pracownicy zatrudnieni w dziale HR cechują się istotnie niższym poczuciem autonomii w porównaniu do działu IT.

W celu wizualnej prezentacji struktury odpowiedzi oraz porównania rozkładów zmiennej Autonomia pomiędzy pracownikami poszczególnych pionów organizacyjnych, na Rysunku 9. przedstawiono wykres skrzynkowy (typu ramka-wąsy) przygotowany dla trzech analizowanych grup.



Rys. 12. Wykres ramka-wąsy poziomu autonomii w zależności od działu
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Nieparametryczne → Porównanie wielu prób niezależnych (grup)***, w oknie konfiguracji wybrano zmienną zależną (tu *Autonomia*) oraz grupującą (tu *Dział*), przycisk *Wykres ramka-wąsy.*

Analiza graficzna zaprezentowana na Rys. 12 w wizualizuje zależności wykazane w nieparametrycznym teście Kruskala-Wallisa. Obserwując położenie miar centralnych, można zauważyć, że mediana dla działu IT wynosi 4,0, co plasuje tę grupę najwyżej pod względem

poczucia samodzielności, podczas gdy w działach HR oraz Sprzedaż wartość środkowa ukształtowała się na identycznym poziomie równym 3,0. Kluczowe dla zrozumienia istotności statystycznej testu jest jednak porównanie całego układu skrzynek (środkowych 50% obserwacji). Prostokątna ramka dla działu IT (zakres od 3,0 do 4,0) jest wyraźnie przesunięta ku górnym rejestrom skali i niemal w całości pokrywa się z ramką działu Sprzedaż (zakres od 2,0 do 4,0), co graficznie wyjaśnia brak istotnych różnic w teście post-hoc pomiędzy tymi dwoma obszarami. Z kolei skrzynka odpowiadająca pracownikom działu HR jest wyraźnie przesunięta w dół i zamyka się w granicach od 2,0 do 3,0. Fakt, że górna granica ramki dla działu HR (trzeci kwartyl) kończy się dokładnie tam, gdzie zaczyna się dolna granica ramki działu IT (pierwszy kwartyl), stanowi wizualny dowód na separację tych rozkładów i potwierdza wykazaną istotność statystyczną.

Na szczególną uwagę zasługuje relacja pomiędzy działem HR a działem Sprzedaż, gdzie uzyskany poziom istotności ukształtował się na granicy progu decyzyjnego ($p = 0{,}056$). Choć jest on formalnie większy niż przyjęty poziom istotności $\alpha = 0{,}05$, co oznacza brak podstaw do odrzucenia hipotezy zerowej o braku różnic, to w praktyce badawczej interpretuje się go jako trend statystyczny. Istnienie tej tendencji potwierdza analiza graficzna (Rys. 9.), na której ramka kwartylna działu Sprzedaż jest wyraźnie przesunięta ku wyższym ocenom w stosunku do działu HR. Ostateczny brak formalnej istotności wynika najprawdopodobniej z ograniczonej liczebności analizowanych podgrup oraz niższej mocy testu nieparametrycznego, co pozwala przypuszczać, że na większej próbie różnica ta by mogła okazać się istotna statystycznie.

8. Analiza współzależności cech

8.1. Typy powiązań między zmiennymi

W klasycznej analizie danych punktem wyjścia do badania **współzależności** cech są zbiory dwuwymiarowe, w których dla każdej pojedynczej jednostki statystycznej określa się jednocześnie wartości dwóch różnych cech, tradycyjnie oznaczanych jako X oraz Y. Mamy wówczas do czynienia z zbiorem n jednostek, którym przyporządkowane są pary cech (x_i, y_i) dla indeksu $i = 1,2, \dots, n$. W metodologii statystycznej relacje zachodzące między dwiema zmiennymi losowymi klasyfikuje się pod kątem ich charakteru na związki funkcyjne oraz stochastyczne. **Zależność funkcyjna** reprezentuje model w pełni deterministyczny, w którym każdej precyzyjnie określonej wartości zmiennej X odpowiada dokładnie jedna, ściśle wyznaczona wartość zmiennej Y. Model ten, choć powszechny w fizyce czy matematyce

teoretycznej, w naukach ekonomicznych, społecznych czy inżynierii jakości występuje niezmiernie rzadko.

W realnej praktyce analitycznej badacz operuje na **związkach stochastycznych**, które charakteryzują się tym, że wraz ze zmianą wartości zmiennej X zmianie ulega rozkład prawdopodobieństwa zmiennej Y. Szczególnym, najpowszechniej analizowanym przypadkiem zależności stochastycznej jest **zależność korelacyjna**. O relacji korelacyjnej mówimy wtedy, gdy określonym wartościom jednej zmiennej odpowiadają różne wartości drugiej zmiennej, ale w średnich swoich wielkościach wykazują one mierzalną prawidłowość kierunkową. Korelacja może przyjąć charakter dodatni, gdy wzrostowi wartości zmiennej X towarzyszy przeciętny wzrost wartości zmiennej Y, bądź ujemny, kiedy wraz ze wzrostem wartości zmiennej X obserwuje się przeciętny spadek wartości zmiennej Y. Zadaniem statystyki jest ilościowe określenie siły (natężenia) oraz kierunku tego typu powiązań.

8.2. Analiza korelacji danych pogrupowanych i tablice korelacyjne

W sytuacjach, gdy liczba zebranych par obserwacji n jest bardzo duża, surowy szereg szczegółowy dla dwóch obserwowanych cech staje się nieczytelny. Wówczas należy dokonać zagregowania danych poprzez konstrukcję tzw. **tablicy korelacyjnej (tabeli wielodzielczej)**. Tablica korelacyjna stanowi dwuwymiarowy szereg rozdzielczy, w którym wiersze reprezentują warianty lub przedziały klasowe zmiennej X, kolumny odzwierciedlają warianty zmiennej Y, natomiast wewnątrz poszczególnych komórek tabeli nanosi się liczebności powiązań, oznaczane jako n_{ij}, które informują, ile jednostek statystycznych posiada jednocześnie i-ty wariant cechy X oraz j-ty wariant cechy Y.

Na podstawie struktury tablicy korelacyjnej program wyznacza tzw. **rozkłady brzegowe** (sumaryczne liczebności dla poszczególnych wierszy i kolumn) oraz **rozkłady warunkowe**, czyli rozkłady jednej cechy przy ustalonym, sztywnym wariancie drugiej zmiennej. Analiza rozkładów warunkowych pozwala na obliczenie średnich warunkowych (np. średniej wartości cechy Y pod warunkiem, że cecha X przyjmuje konkretną wartość x_i) oraz wariancji warunkowych. Wykres łączący kolejne wartości średnich warunkowych nazywamy linią regresji empirycznej. Do pomiaru ogólnej siły powiązania korelacyjnego dla danych tablicowych, bez względu na to, czy zależność ma charakter prostoliniowy, czy krzywoliniowy, stosuje się stosunek korelacyjny, oznaczany jako e_{xy} lub e_{yx}. Miernik ten bazuje na analizie wariancji warunkowych i przyjmuje wartości z przedziału domkniętego $\langle 0; 1 \rangle$, gdzie wartość 0

oznacza całkowity brak powiązania korelacyjnego między średnimi warunkowymi, a wartość 1 świadczy o istnieniu ścisłego związku funkcyjnego.

8.3. Analiza korelacji liniowej zmiennych ciągłych oraz współczynnik Pearsona

W sytuacji, gdy badacz operuje na zmiennych ciągłych mierzonych na skalach silnych (przedziałowej lub ilorazowej) i zakłada, że zależność między nimi ma charakter prostoliniowy, podstawowym narzędziem pomiaru staje się kowariancja oraz współczynnik korelacji liniowej Pearsona. Kowariancja, oznaczana symbolem cov_{xy}, to średnia arytmetyczna iloczynu odchyleń wartości obu zmiennych od ich własnych średnich arytmetycznych, wyznaczana dla szeregu szczegółowego wzorem:

$$cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

Kowariancja informuje wyłącznie o kierunku zależności (dodatnia wartość świadczy o zależności dodatniej, ujemna o ujemnej), jest jednak miarą mianowaną i zależną od skali danych, co uniemożliwia bezpośrednie porównywanie siły związków dla różnych cech.

Aby wyeliminować tę wadę, Karl Pearson dokonał standaryzacji kowariancji, dzieląc ją przez iloczyn odchyleń standardowych obu zmiennych, co doprowadziło do stworzenia współczynnika korelacji liniowej, oznaczanego symbolem r lub r_{xy}, wyliczanego według wzoru:

$$r_{xy} = \frac{cov_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Współczynnik Pearsona jest miarą niemianowaną i przyjmuje wartości z przedziału $\langle -1; 1\rangle$. Wartość $r = 0$ oznacza całkowity brak korelacji liniowej, wartości dodatnie ($r > 0$) wskazują na korelację dodatnią, natomiast ujemne ($r < 0$) na korelację ujemną, przy czym im bliższa jedności jest bezwzględna wartość współczynnika $|r|$, tym silniejsza zależność liniowa łączy obie zmienne.

Uzyskanie określonej wartości r z próby nie gwarantuje jeszcze, że zależność ta występuje w całej populacji generalnej, gdyż wynik może być dziełem przypadku. Z tego powodu pwarto przeprowadzić parametryczny test istotności współczynnika korelacji, weryfikujący hipotezę zerową $H_0: \rho = 0$ (brak korelacji w populacji) wobec hipotezy alternatywnej $H_1: \rho \neq 0$. Sprawdzianem hipotezy jest statystyka t podlegająca teoretycznemu rozkładowi t-Studenta o $v = n - 2$ stopniach swobody, opisywana formułą:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Na podstawie tej statystyki program generuje p-wartość: jeśli $p < \alpha$ (najczęściej 0,05), hipotezę zerową się odrzuca, stwierdzając, że korelacja liniowa w populacji jest statystycznie istotna.

8.4. Analiza korelacji dla danych porządkowych – współczynnik rang Spearmana

Gdy założenia dotyczące stosowania współczynnika Pearsona zostaną naruszone, na przykład, gdy dane charakteryzują się nieliniowym (ale monotonicznym) charakterem powiązania, gdy w próbie występują silne obserwacje odstające, bądź gdy analizowane cechy zostały zmierzone na słabej skali porządkowej, właściwym rozwiązaniem staje się nieparametryczny współczynnik korelacji kolejnościowej (rang) Spearmana, oznaczany jako R_{xy}.

Matematyczna istota wyznaczania współczynnika Spearmana polega na zastąpieniu surowych wartości cech X oraz Y ich rangami (pozycjami), przypisywanymi niezależnie dla każdej zmiennej po ich uprzednim uszeregowaniu w szeregi szczegółowe rosnące. Po dokonaniu procedury rangowania, dla każdej jednostki statystycznej oblicza się różnicę między przypisanymi jej rangami, oznaczaną jako $d_i = R(x_i) - R(y_i)$. Jeśli w bazie danych nie występują tzw. rangi wiązane (identyczne wartości cech), współczynnik korelacji rangowej Spearmana wylicza się za pomocą wplecionego w tekst, uproszczonego wzoru:

$$R_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Miernik ten, analogicznie do współczynnika Pearsona, przyjmuje wartości z przedziału domkniętego $\langle -1; 1 \rangle$, a jego interpretacja pod kątem siły i kierunku związku jest tożsama. Weryfikacja istotności statystycznej współczynnika Spearmana dla prób o liczebności $n > 10$ realizowana jest przez program również w oparciu o aparat rozkładu t-Studenta o $n-2$ stopniach swobody za pomocą analogicznej statystyki testowej.

8.5. Test niezależności chi-kwadrat dla cech jakościowych

W sytuacji, gdy badacz operuje na zmiennych o charakterze niemierzalnym, sklasyfikowanych na słabych skalach nominalnych bądź porządkowych (np. płeć, wykształcenie, typ usterki, region geograficzny), stosowanie klasycznych współczynników korelacji liniowej Pearsona jest całkowicie niedozwolone. Narzędziem dedykowanym do weryfikacji, czy między dwiema takimi cechami jakościowymi zachodzi jakikolwiek mierzalny

związek stochastyczny, jest nieparametryczny test niezależności chi-kwadrat (χ^2). Podobnie jak w przypadku analizy danych pogrupowanych, punktem wyjścia jest konstrukcja tablicy wielodzielczej (tablicy kontyngencji) o wymiarach $k \times l$, gdzie k to liczba wariantów cechy X, a l oznacza liczbę wariantów cechy Y. Sformułowana hipoteza zerowa H_0 zakłada niezależność badanych cech w populacji generalnej (brak związku), wobec hipotezy alternatywnej H_1 stwierdzającej, że cechy X i Y są ze sobą powiązane stochastycznie.

Matematyczna procedura testowa polega na zestawieniu zaobserwowanych w badaniu empirycznym liczebności komórkowych, oznaczanych jako n_{ij}, z tzw. liczebnościami teoretycznymi (oczekiwanymi), oznaczanymi symbolem np_{ij}, które pojawiłyby się w tablicy, gdyby hipoteza o pełnej niezależności była idealnie prawdziwa. Dla każdej komórki tablicy program wyznacza liczebność teoretyczną jako iloraz iloczynu odpowiednich sum brzegowych (sumy danego wiersza $n_{i\cdot}$ i sumy danej kolumny $n_{\cdot j}$) do całkowitej liczebności próby n, co opisuje wzór $np_{ij} = \left(n_{i\cdot} \cdot n_{\cdot j}\right)/n$. Sprawdzianem hipotezy zerowej jest statystyka χ^2_{obl}, obliczana jako sumaryczna wielkość unormowanych kwadratów odchyleń liczebności rzeczywistych od teoretycznych we wszystkich komórkach, zgodnie z formułą:

$$\chi^2_{obl} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}}$$

Statystyka ta przy założeniu prawdziwości hipotezy zerowej podlega teoretycznemu rozkładowi chi-kwadrat o liczbie stopni swobody zdefiniowanej wzorem $v = (k-1)(l-1)$. Należy pamiętać o ograniczeniu tego testu: liczebności oczekiwane tabeli muszą spełniać warunek $np_{ij} \geq 5$, a żadna komórka nie może mieć liczebności oczekiwanej równej zero.

Reguła decyzyjna w programie Statistica opiera się na wygenerowanej p-wartości: jeśli $p < \alpha$ (najczęściej 0,05), hipotezę zerową o niezależności odrzucamy, uzyskując formalny dowód na istnienie istotnego statystycznie związku między badanymi cechami jakościowymi.

W kontekście komputerowej weryfikacji współzależności cech za pomocą testu niezależności chi-kwadrat, sam wynik liczbowy statystyki oraz jej p-wartość informują badacza jedynie o tym, *czy* zależność jest istotna statystycznie. Nie dają one jednak odpowiedzi na pytanie, jak ta zależność się układa wewnątrz próby i które kategorie determinują uzyskany wynik. Kluczem do pełnej, merytorycznej interpretacji struktury powiązań jest szczegółowa analiza procentów wierszowych oraz procentów kolumnowych wewnątrz tablicy wielodzielczej (tabeli kontyngencji).

Procenty wierszowe oblicza się, przyjmując sumę brzegową danego wiersza za podstawę odniesienia, czyli za 100%. Algorytm programu dzieli liczebność danej komórki n_{ij} przez sumę brzegową wiersza $n_{i.}$ i mnoży wynik przez sto procent, co formalnie zapisujemy jako $\%_{wiersza} = (n_{ij}/n_{i.}) \cdot 100\%$. Procenty wierszowe pozwalają na analizę rozkładu warunkowego cechy kolumnowej (Y) na poszczególnych poziomach cechy wierszowej (X). Odpowiadają one na pytanie: „Jak kształtuje się struktura cechy Y wewnątrz konkretnej, wyodrębnionej grupy X?”. Przykładowo, jeśli cechą wierszową jest wykształcenie (podstawowe, średnie, wyższe), a kolumnową preferencje zakupu produktu (A, B, C), to analiza procentów wierszowych pozwoli stwierdzić, że na przykład aż 75% osób z wykształceniem wyższym wybiera produkt A, podczas gdy w grupie z wykształceniem podstawowym odsetek ten wynosi zaledwie 10%. Porównywanie profili wierszowych pozwala precyzyjnie wskazać kierunek zależności.

Procenty kolumnowe wyznacza się w sposób odwrotny, przyjmując za bazę (100%) sumę brzegową danej kolumny. Każda liczebność komórkowa n_{ij} jest dzielona przez sumę brzegową kolumny $n_{.j}$ i wyrażana w procentach, zgodnie ze wzorem $\%_{kolumny} = (n_{ij}/n_{.j}) \cdot 100\%$. Mierniki te służą do badania struktury cechy wierszowej (X) wewnątrz zdefiniowanych kategorii cechy kolumnowej (Y). Pozwalają odpowiedzieć na pytanie: „Jaki jest profil (skład) grupy osób, które wybrały dany wariant Y, pod względem cechy X?”. Przykładowo, nawiązując do poprzedniego eksperymentu, analiza procentów kolumnowych pozwoli ustalić, jaki odsetek wśród wszystkich konsumentów, którzy zdecydowali się na produkt A, stanowią osoby z wykształceniem wyższym, średnim oraz podstawowym. Jeśli 80% nabywców produktu A to osoby z wyższym wykształceniem, badacz uzyskuje informację profilową o strukturze rynku.

Przykład

Celem kolejnego etapu badania było zweryfikowanie istnienia, kierunku oraz siły współzależności pomiędzy miesięczną wydajnością pracowników, liczbą odbytych szkoleń a stażem pracy w przedsiębiorstwie. W celu określenia charakteru tych relacji przeprowadzono analizę korelacji r-Pearsona, której szczegółowe wyniki w postaci współczynników korelacji oraz odpowiadających im poziomów istotności p zestawiono w Tabeli 10.

Tabela 10. Macierz współczynników korelacji r-Pearsona

Zmienna	Wydajność	Szkolenia	Staż
Wydajność	1,000 p= ---	0,583 p=0,000	0,358 p=0,000
Szkolenia	0,583 p=0,000	1,000 p= ---	-0,203 p=0,026
Staż	0,358 p=0,000	-0,203 p=0,026	1,000 p= ---

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe*** → ***Macierz korelacji***, w oknie konfiguracji wybrano dwie listy zmiennych (tu w obu listach: *Wydajność, Szkolenia, Staż*), w zakładce opcje zaznaczono *Wyświetl r, N i p.*

Analizując uzyskane dane empiryczne, w pierwszym kroku oceniono relację pomiędzy wydajnością pracowników a liczbą odbytych szkoleń. Współczynnik korelacji dla tej pary zmiennych ukształtował się na poziomie $0{,}583$ przy poziomie istotności $p < 0{,}001$. Ponieważ uzyskana wartość prawdopodobieństwa testowego jest mniejsza od przyjętego poziomu istotności $\alpha = 0{,}05$, związek ten jest istotny statystycznie. Dodatnia wartość współczynnika wskazuje na dodatni kierunek korelacji o umiarkowanej sile. Oznacza to, że wraz ze wzrostem liczby odbytych szkoleń miesięczna wydajność pracowników również wykazuje istotną tendencję wzrostową, a współczynnik determinacji wskazuje, że aktywność szkoleniowa wyjaśnia w 34,0% zmienność tej wydajności.

W następnej kolejności poddano analizie związek pomiędzy wydajnością a stażem pracy zatrudnionych osób. W tym przypadku współczynnik korelacji osiągnął wartość $0{,}358$ przy poziomie istotności $p < 0{,}001$. Wynik ten ponaddto jest mniejszy od $\alpha = 0{,}05$, co potwierdza istotność statystyczną tej współzależności. Wartość współczynnika świadczy o słabej, lecz wyraźnej korelacji dodatniej, a współczynnik determinacji wskazuje, że staż pracy wyjaśnia w 12,8% zmienność miesięcznej wydajności pracowników. Pozwala to wnioskować, że pracownicy o dłuższym stażu pracy w firmie osiągają wyższą miesięczną wydajność, jednak ta zależność jest słabsza niż w przypadku szkoleń.

Ostatnim analizowanym elementem była relacja łącząca liczbę odbytych szkoleń oraz staż pracy pracowników. Przeprowadzone badanie wykazało dla tej pary współczynnik korelacji na poziomie $-0{,}203$ przy poziomie istotności $p = 0{,}026$. Ponieważ wartość p jest mniejsza od przyjętego kryterium $\alpha = 0{,}05$, zależność ta jest istotna statystycznie, choć charakteryzuje się słabą siłą. Ujemny znak współczynnika wskazuje na ujemny kierunek relacji. Oznacza to, że pracownicy o dłuższym stażu pracy w przedsiębiorstwie wykazują tendencję do uczestniczenia w mniejszej liczbie szkoleń (lub odwrotnie: osoby o krótszym stażu przechodzą

ich relatywnie więcej). Podsumowując, wszystkie analizowane zmienne wykazują wzajemne, istotne statystycznie powiązania.

Celem kolejnego etapu analizy było zweryfikowanie, czy istnieje istotna statystycznie zależność pomiędzy formą wykonywanej pracy (zdalną lub stacjonarną) a działem, w którym zatrudnieni są pracownicy (IT, HR, Sprzedaż). Ponieważ badanie dotyczyło współwystępowania dwóch zmiennych o charakterze jakościowym, do weryfikacji hipotezy o niezależności zastosowano test χ^2 Pearsona. Dodatkowo, w celu określenia siły ewentualnego związku, obliczony został współczynniki kontyngencji oraz V Craméra. Zbiorcze wyniki przedstawiono w Tabeli 11.

Tabela 11. Wyniki testu niezależności χ^2 oraz miar siły związku

Statystyka	Chi-kwadr.	df	p
Chi^2 Pearsona	0,688	df=2	p=0,709
Chi^2 NW	0,684	df=2	p=0,710
Fi	0,076		
Wsp. kontyngencji	0,076		
V Craméra	0,076		

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Tabele wielodzielcze***, w oknie konfiguracji wybrano przycisk *Określ tabelę* (tu lista 1: *Format pracy,* lista 2: *Dział*), w zakładce *Opcje* zaznaczono *Chi-kwadrat Pearsona i NW* oraz *Fi (tabela 2x2, v i C Cramera,* a następnie w zakładce *Więcej* przycisk *Dokładne tabele dwudzielcze.*

W teście χ^2 Pearsona postawiono hipotezę zerową mówiącą o tym, że forma pracy oraz dział zatrudnienia są zmiennymi niezależnymi w populacji generalnej, wobec hipotezy alternatywnej zakładającej istnienie współzależności pomiędzy tymi cechami. Analizując dane empiryczne zaprezentowane w Tabeli 11, dla statystyki testowej $\chi^2 = 0{,}688$ przy liczbie stopni swobody $df = 2$, program wyznaczył poziom istotności $p = 0{,}709$. Ponieważ wyznaczony poziom p jest większy od przyjętego kryterium istotności $\alpha = 0{,}05$, brak jest podstaw do odrzucenia hipotezy zerowej. Nie ma podstaw uważać, że zmienne są zależne.

Konkluzję o braku powiązań potwierdzają również wskaźniki siły związku, które ze względu na niespełnienie warunku istotności samego testu χ^2 mają charakter wyłącznie uzupełniający. Wszystkie obliczone parametry – współczynnik kontyngencji oraz V Craméra – przybrały identyczną, niską wartość wynoszącą 0,076. Tak znikomy poziom wskaźników wskazuje na brak związku. Podsumowując, uzyskane wyniki pozwalają stwierdzić, że podział pracowników na formę pracy zdalnej i stacjonarnej odbywa się w sposób niezależny od

struktury działowej przedsiębiorstwa, a zaobserwowane w próbie drobne wahania liczebności mają charakter czysto losowy.

W celu dokładnego zobrazowania struktury zatrudnienia w analizowanym przedsiębiorstwie, w Tabeli 12 zaprezentowano zestawienie liczebności oraz rozkładów procentowych (w układzie kolumnowym, wierszowym oraz ogólnym) dla zmiennych jakościowych *Format pracy* oraz *Dział*.

Tabela 12. Tabela krzyżowa liczebności i odsetek dla zmiennych *Forma pracy* pracy oraz *Dział*

Format pracy	Dział IT	Dział HR	Dział Sprzedaż	Wiersz
Zdalny	24	28	22	74
%kolumny	64,9%	63,6%	56,4%	
%wiersza	32,4%	37,8%	29,7%	
% ogółu	20,0%	23,3%	18,3%	61,7%
Stacjonarny	13	16	17	46
%kolumny	35,1%	36,4%	43,6%	
%wiersza	28,3%	34,8%	37,0%	
% ogółu	10,8%	13,3%	14,2%	38,3%
Ogół	37	44	39	120
% ogółu	30,8%	36,7%	32,5%	100,0%

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Statystyki podstawowe → Tabele wielodzielcze***, w oknie konfiguracji wybrano przycisk *Określ tabelę* (tu lista 1: *Format pracy,* lista 2: *Dział*), w zakładce *Opcje* zaznaczono *Procenty z całości*, *Procenty w wierszach*, *Procenty w kolumnach.*

Analizując zebrany materiał badawczy dla łącznej próby $N = 120$ pracowników, można zauważyć, że w przedsiębiorstwie wyraźnie dominuje model pracy zdalnej, w którym obowiązki wykonuje łącznie 74 pracowników, co stanowi 61,7% ogółu zatrudnionych. W formie stacjonarnej pracuje natomiast 46 osób (38,3% ogółu). Pod względem liczebności poszczególnych pionów organizacyjnych, największą grupę stanowią pracownicy działu HR (44 osoby; 36,7%), na drugim miejscu plasuje się pion Sprzedaży (39 osób; 32,5%), a najmniej liczną grupą są specjaliści z działu IT (37 osób; 30,8%).

Szczegółowy rozkład formatów pracy wewnątrz poszczególnych jednostek organizacyjnych (% kolumny) wykazuje bardzo wysoki stopień jednolitości, co bezpośrednio tłumaczy brak istotności statystycznej w teście χ^2. Zarówno w dziale IT, jak i w dziale HR, odsetek osób pracujących zdalnie utrzymuje się na bardzo zbliżonym, wysokim poziomie i wynosi odpowiednio 64,9% (24 osoby) oraz 63,6% (28 osób). W dziale Sprzedaży udział pracowników zdalnych jest minimalnie niższy, ukształtował się bowiem na poziomie 56,4% (22 osoby), co przekłada się na nieco większy niż w pozostałych grupach udział pracowników

stacjonarnych (43,6%; 17 osób). Analizując strukturę wewnątrz form pracy (% wiersza), okazuje się, że wśród wszystkich 74 pracowników zdalnych największy odsetek stanowią reprezentanci działu HR (37,8%), a najmniejszy – działu Sprzedaży (29,7%). Z kolei w grupie 46 pracowników stacjonarnych sytuacja się odwraca: najwięcej osób reprezentuje pion sprzedaży (37,0%), a najmniej – IT (28,3%). Podsumowując, rozkład procentowy potwierdza, że struktura formy pracy rozkłada się równomiernie w poprzek całej struktury działowej firmy.

9. Modelowanie zależności – analiza regresji liniowej

9.1. Klasyczny model regresji liniowej jednej zmiennej

W statystycznej analizie danych, po ilościowym określeniu siły i kierunku powiązania korelacyjnego między cechami, kolejnym naturalnym krokiem badawczym jest próba przyczynowo-skutkowego opisania zachodzącego zjawiska za pomocą modelu matematycznego. Narzędziem realizującym to zadanie jest analiza regresji, która pozwala na badanie wpływu jednej lub wielu zmiennych niezależnych (objaśniających), oznaczanych symbolem X, na zmienną zależną (objaśnianą), oznaczaną jako Y. Najpowszechniej stosowaną formą tego podejścia jest klasyczny model regresji liniowej jednej zmiennej, w którym zakłada się, że zależność między badanymi cechami w populacji generalnej można opisać za pomocą funkcji prostej z uwzględnieniem losowego błędu (składnika losowego), co formalnie przyjmuje postać równania $Y = \beta_0 + \beta_1 X + \varepsilon$, gdzie parametry β_0 i β_1 to nieznane współczynniki strukturalne populacji, a symbol ε oznacza składnik losowy (resztowy) odzwierciedlający wpływ wszystkich czynników ubocznych.

Aby wnioskowanie na podstawie tego modelu było poprawne matematycznie, muszą zostać spełnione założenia:

- zmienna objaśniająca X musi być nielosowa,
- składnik losowy ε musi charakteryzować się wartością oczekiwaną równą zero, stałą wariancją dla wszystkich obserwacji (homoskedastyczność), brakiem autokorelacji oraz rozkładem normalnym.

Ponieważ parametry populacyjne są nieznane, program Statistica dokonuje ich szacowania na podstawie próby empirycznej, wyznaczając model empiryczny opisany równaniem $\hat{y} = b_0 + b_1 x$, w którym b_0 oraz b_1 są ocenami parametrów strukturalnych. Do wyznaczenia tych współczynników algorytm oprogramowania stosuje klasyczną Metodę Najmniejszych Kwadratów (MNK). Istotą MNK jest zminimalizowanie sumy kwadratów różnic (reszt) pomiędzy rzeczywistymi wartościami empirycznymi zmiennej objaśnianej y_i a

wartościami teoretycznymi $\hat{y}_i$ wynikającymi z równania prostej. W wyniku tej minimalizacji otrzymujemy wzory na współczynniki regresji, gdzie współczynnik kierunkowy prostej b_1 oblicza się jako stosunek kowariancji do wariancji zmiennej X:

$$b_1 = \frac{cov_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Natomiast wyraz wolny prostej b_0 wyznacza się na podstawie wyliczonego parametru b_1 oraz średnich arytmetycznych obu cech, według równania transformacji:

$$b_0 = \overline{y} - b_1\overline{x}$$

Uzyskanie konkretnych wartości liczbowych współczynników b_1 oraz b_0 pozwala badaczowi na dokonanie ścisłej interpretacji merytorycznej badanej zależności. Wyraz wolny b_0 informuje o teoretycznym poziomie zmiennej objaśnianej Y w hipotetycznej sytuacji, gdy zmienna objaśniająca X przyjmuje wartość równą zero, przy czym jego interpretacja ma sens tylko wtedy, gdy wartość zero mieści się w obszarze zmienności danych empirycznych. Z kolei kluczowe znaczenie poznawcze ma współczynnik kierunkowy regresji b_1, nazywany również efektem krańcowym. Współczynnik b_1 informuje, o ile jednostek (swojego miana) zmieni się (wzrośnie, gdy $b_1 > 0$, lub spadnie, gdy $b_1 < 0$) przeciętna wartość zmiennej objaśnianej Y, jeżeli wartość zmiennej objaśniającej X wzrośnie o jedną własną jednostkę.

9.2. Klasyczny model regresji wielorakiej

W rzeczywistych procesach masowych zmienna objaśniana Y rzadko zależy tylko od jednego czynnika, dlatego naturalnym rozwinięciem analizy jest model regresji wielorakiej, który pozwala na jednoczesne badanie wpływu całego zestawu k zmiennych objaśniających, oznaczanych jako $X_1, X_2, \ldots, X_k$. Klasyczny model regresji wielorakiej w populacji generalnej przyjmuje postać liniową opisaną ogólnym równaniem:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

w którym parametr β_0 to wyraz wolny, symbole $\beta_1, \beta_2, \ldots, \beta_k$ oznaczają cząstkowe (regresyjne) współczynniki strukturalne, a ε stanowi wielowymiarowy składnik losowy.

Algorytmy programu Statistica dokonują estymacji tego układu przy użyciu macierzowego rozwinięcia Metody Najmniejszych Kwadratów, wyznaczając równanie empiryczne w postaci $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$, gdzie wartości b_i stanowią nieobciążone oceny nieznanych parametrów β_i. Cząstkowe współczynniki kierunkowe b_i (raportowane przez program również jako ustandaryzowane współczynniki *beta*) posiadają unikalną wartość interpretacyjną, ponieważ informują badacza, o ile jednostek zmieni się

przeciętna wartość zmiennej objaśnianej Y, jeżeli wartość danej zmiennej objaśniającej X_i wzrośnie o jedną własną jednostkę, przy z góry założonym warunku *ceteris paribus*, czyli przy zachowaniu wszystkich pozostałych zmiennych objaśniających w modelu na niezmienionym poziomie. Wyraz wolny b_0 informuje o teoretycznym poziomie zmiennej objaśnianej Y w sytuacji, gdy wszystkie zmienne objaśniające przyjmują wartość równą zero.

9.3. Wnioskowanie o parametrach strukturalnych modelu

Wartości współczynników wyliczone z próby są jednak wyłącznie ocenami punktowymi i wymagają formalnego sprawdzenia pod kątem ich istotności statystycznej w populacji generalnej. Program Statistica automatycznie wyznacza błędy szacunku współczynników, oznaczane jako $s(b_i)$, i przeprowadza testy istotności dla każdego parametru z osobna, weryfikujące hipotezę zerową $H_0: \beta_i = 0$ (zmienna X_i nie ma istotnego wpływu na Y) wobec hipotezy alternatywnej $H_1: \beta_i \neq 0$. Sprawdzianem tak postawionych hipotez jest statystyka t podlegająca teoretycznemu rozkładowi t-Studenta o liczbie stopni swobody zdefiniowanej jako $v = n - k - 1$ dla modelu wielorakiego, obliczana według wzoru:

$$t = \frac{b_i}{s(b_i)}$$

Jeśli stowarzyszona z tym testem p-wartość spełnia relację $p < \alpha$ (najczęściej 0,05), hipotezę zerową się odrzuca na rzecz hipotezy alternatywnej, uznając wpływ danej zmiennej objaśniającej za statystycznie istotny i nieprzypadkowy.

9.4. Kompleksowa weryfikacja założeń i ocena jakości modeli regresyjnych

Aby zbudowane równanie regresji (zarówno liniowej jednej zmiennej, jak i wielorakiej) mogło zostać uznane za poprawne należy przeprowadzić procedurę weryfikacji założeń opartą na analizie uzyskanych reszt empirycznych $e_i = y_i - \hat{y}_i$ oraz ocenie dobroci dopasowania. Pierwszym etapem jest wyznaczenie współczynnika determinacji R^2, który mierzy, jaka część całkowitej zmienności zmiennej Y została wyjaśniona przez wprowadzone do modelu zmienne objaśniające, wyliczanego ze wzoru:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Do przetestowania łącznej istotności strukturalnej całego modelu (czyli hipotezy zerowej mówiącej, że współczynnik determinacji w populacji jest równy zero) program wykorzystuje test Walda oparty na statystyce F_{obl} podlegającej rozkładowi F Fishera-Snedecora o stopniach swobody $v_1 = k$ oraz $v_2 = n - k - 1$, wyznaczanej ze wzoru:

$$F_{obl} = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

Sama wysoka wartość R^2 nie gwarantuje poprawności modelu, jeśli naruszone zostaną założenia dotyczące składnika losowego, co weryfikuje się w programie Statistica za pomocą następujących testów resztowych:

1. **Założenie o normalności rozkładu składnika losowego**

   Składnik losowy powinien podgać rozkładowi normalnemu o wartości oczekiwanej równej zero. Naruszenie tego założenia uniemożliwia poprawną weryfikację statystyczną parametrów. Założenie to jest realizowane poprzez wygenerowanie reszt empirycznych i poddanie ich testowi zgodności, np. testem Shapiro-Wilka lub chi-kwadrat, gdzie pożądanym wynikiem jest wartość $p > 0{,}05$, świadcząca o braku podstaw do odrzucenia normalności rozkładu błędów.

2. **Założenie o braku autokorelacji składnika losowego**

   Błędy dla poszczególnych obserwacji nie mogą być od siebie zależne, co jest szczególnie istotne przy danych o charakterze szeregów czasowych. Do wykrywania autokorelacji pierwszego rzędu program Statistica implementuje Test Durbina-Watsona, którego statystyka d opiera się na kwadratach różnic sąsiednich reszt: $d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$.

   Wynik bliski wartości 2 oznacza pożądany brak autokorelacji, podczas gdy wartości bliskie 0 lub 4 sygnalizują obecność odpowiednio dodatniej lub ujemnej autokorelacji reszt.

Przykład

W celu określenia siły oraz kierunku wpływu aktywności rozwojowej i doświadczenia na efektywność pracowników, przeprowadzono wielokrotną analizę regresji liniowej. Jako zmienną zależną wskazano miesięczną Wydajność, natomiast w charakterze zmiennych niezależnych (predyktorów) wprowadzono do modelu liczbę odbytych Szkoleń oraz Staż pracy. Zbiorcze wyniki szacowania parametrów modelu zestawiono w Tabeli 13.

Tabela 13. Wyniki wielokrotnej regresji liniowej dla zmiennej zależnej *Wydajność*

	Podsumowanie regresji zmiennej zależnej: Wydajność R= 0,759 R^2=0,577 Popraw. R2=0 ,569 F(2,117)=79,689 p<0,000 Błąd std. estymacji: 8,622					
N=120	b*	Bł. std.	b	Bł. std.	t(117)	p
W. wolny			32,350	3,362	9,621	0,000
Szkolenia	0,684	0,061	0,842	0,076	11,134	0,000
Staż	0,497	0,061	3,610	0,447	8,084	0,000

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Regresja wieloraka***, wskazano zmienne (tu Zmienne zależne: *Wydajność,* Lista zmiennych niezależnych: *Szkolenia, Staż*), po naciśnięciu przycisku *OK* wybrano *Podsumowanie: wynki regresji.*

Pierwszym etapem analizy była ocena globalnego dopasowania oraz istotności zbudowanego modelu regresji. Współczynnik korelacji wielokrotnej ukształtował się na wysokim poziomie $R = 0{,}759$, co świadczy o silnym związku liniowym między zmiennymi objaśniającymi a zmienną zależną. Współczynnik determinacji $R^2 = 0{,}577$ informuje, że model uwzględniający staż oraz liczbę szkoleń wyjaśnia 57,7% całkowitej wariancji (zmienności) miesięcznej wydajności pracowników. Pozostałe 42,3% stanowią czynniki losowe lub zmienne niewłączone do badania. Wynik testu istotności całego modelu okazał się istotny statystycznie ($F(2{,}117) = 79{,}689;\ p < 0{,}001$), co wskazuje, że wprowadzone predyktory w sposób istotny poprawiają przewidywanie wydajności w stosunku do modelu opartego wyłącznie na średniej.

W drugim kroku poddano analizie parametry cząstkowe poszczególnych zmiennych niezależnych. Wyraz wolny modelu wynosi 32,350 ($p < 0{,}001$) i reprezentuje teoretyczny poziom wydajności dla pracownika bez stażu, który nie odbył żadnego szkolenia. Oba wprowadzone predyktory okazały się istotnymi statystycznie stymulatorami wydajności, na co wskazują poziomy $p < 0{,}001$ (mniejsze od $\alpha = 0{,}05$).

Zmienna Szkolenia charakteryzuje się współczynnikiem kierunkowym $b = 0{,}842$. Oznacza to, że przy kontrolowaniu wpływu stażu pracy, każde kolejne odbyte szkolenie przekłada się na średni wzrost miesięcznej wydajności pracownika o niemal jeden (0,842) projekt w miesiącu. Z kolei dla zmiennej Staż współczynnik ten wyniósł $b = 3{,}610$, co wskazuje, że przy stałej liczbie szkoleń, każdy dodatkowy rok stażu pracy zwiększa wydajność średnio o 3,610 projekty zrealizowanego w miesiącu.

Na podstawie uzyskanych współczynników niestandaryzowanych b sformułowano równanie regresji opisujące badany proces:

$$\textit{Wydajność} = 32,350 + 0,842 \cdot \textit{Szkolenia} + 3,610 \cdot \textit{Staż}$$

Uzupełnieniem i formalnym potwierdzeniem prawidłowości dopasowania wielokrotnej regresji liniowej jest tabela analizy wariancji (ANOVA) dla modelu regresji, zaprezentowana w Tabeli 14.

Tabela 14. Wyniki analizy wariancji (ANOVA) dla modelu regresji zmiennej zależnej *Wydajność*

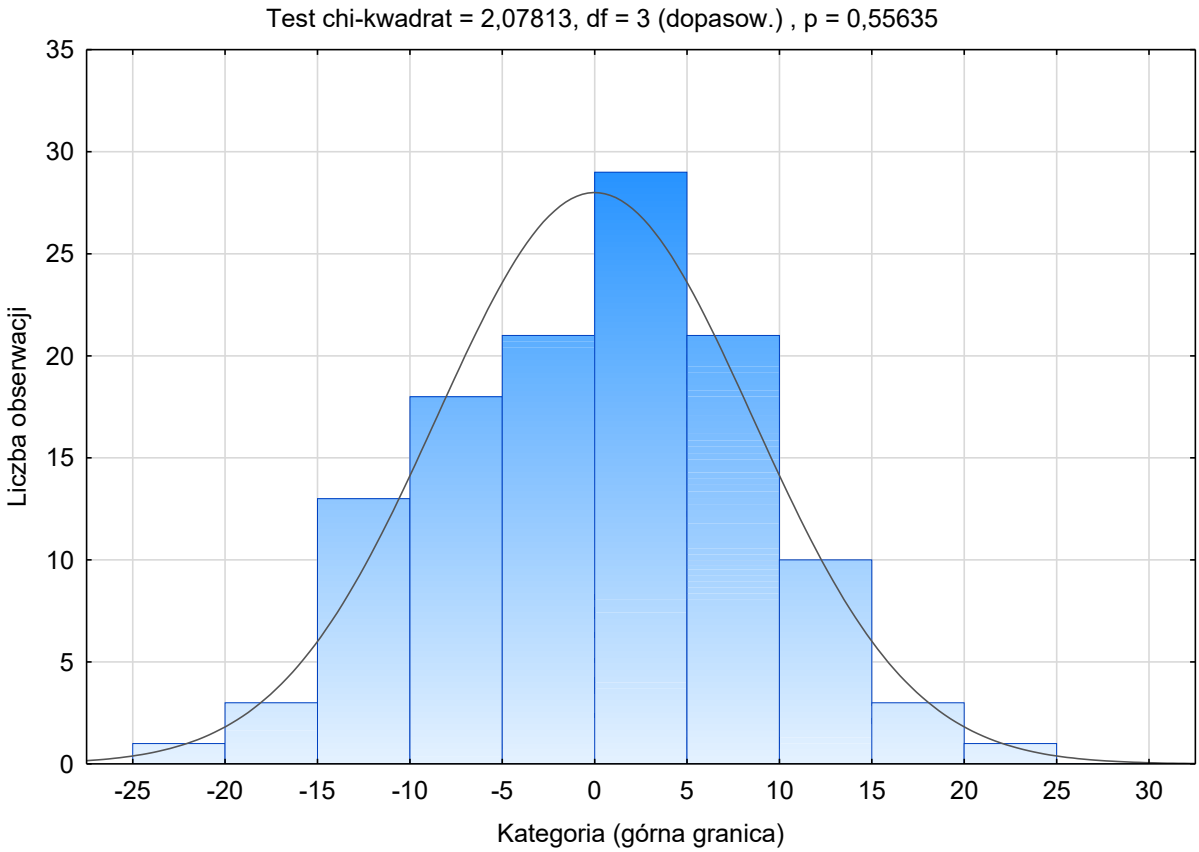
Efekt	Suma	df	Średnia	F	p
Regres.	11848,09	2	5924,046	79,689	0,000
Reszta	8697,77	117	74,340		
Razem	20545,87				

Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Regresja wieloraka***, wskazano zmienne (tu Zmienne zależne: *Wydajność,* Lista zmiennych niezależnych: *Szkolenia, Staż*), po naciśnięciu przycisku *OK* wybrano w zakładce *Więcej* przycisk *ANOVA (sum. dobroć dopasow.).*

Uzyskano prawdopodobieństwo testowe $p < 0,001$, które jest mniejsze od poziomu istotności $\alpha = 0,05$, więc hipoteza zerowa o braku wpływu predyktorów zostaje odrzucona. Wynik ten potwierdza, że kombinacja liniowa zmiennych Szkolenia oraz Staż w sposób istotny statystycznie wyjaśnia wariancję wydajności pracowników.

Ostatnim, kluczowym etapem weryfikacji poprawności i stabilności zbudowanego modelu regresji wielokrotnej była formalna ocena rozkładu składników resztowych (błędów estymacji). Jednym z fundamentalnych założeń klasycznej metody najmniejszych kwadratów (KMNK) jest warunek, aby reszty modelu charakteryzowały się rozkładem normalnym o średniej równej zero. W celu oceny tego kryterium, na Rysunku 15 zaprezentowano histogram częstości reszt z nałożoną teoretyczną krzywą Gaussa oraz wynikami testu zgodności χ^2.



Rys. 13. Histogram rozkładu reszt modelu regresji z nałożoną krzywą rozkładu normalnego
Źródło: opracowanie własne przy użyciu programu Statistica.

Zakładka ***Dopasowanie rozkładu***, w oknie konfiguracji wybrano rozkład normalny, zmienną *Reszty* (powstała ona po wybraniu w oknie *Regresji wielorakiej* zakładki *Reszty, założenia, predykcja* przycisku *Wykonaj analizę reszt* – przekopiowanie z utworzonego pliku zmiennej *Reszty* do nowego arkusza) i po zaakceptowaniu naciśnięto przycisk *Wykres rozkładu obserw. i oczekiwanego.*

Średnia wartość reszt ukształtowała się na poziomie równym 0, co wskazuje na brak systematycznego błędu w prognozach wydajności stawianych przez model. Przeprowadzony test zgodności χ^2 uzyskano poziomu istotności $p = 0{,}556$. Uzyskane prawdopodobieństwo testowe jest większe od standardowego kryterium istotności $\alpha = 0{,}05$, co oznacza brak podstaw do odrzucenia hipotezy zerowej mówiącej o zgodności struktury reszt z rozkładem normalnym. Oznacza to, że nie ma podstaw do odrzucenia hipotezy o normalności rozkładu reszt, co w połączeniu z ich zerową wartością oczekiwaną wskazuje, że wnioskowanie o istotności całego modelu jest wiarygodne, mimo dość niskiej wartości współczynnika determinacji.

10. Metodyka wyboru procedur statystycznych

10.1. Strategia wyboru testu statystycznego

Najtrudniejszym etapem samodzielnej pracy z analizą danych w programie Statistica nie jest techniczne kliknięcie odpowiedniej opcji w menu, lecz poprawny, metodologiczny wybór procedury statystycznej, która jest adekwatna do postawionego problemu badawczego oraz natury zebranych danych empirycznych. Proces ten wymaga od analityka przejścia przez sekwencyjną ścieżkę pytań kontrolnych, tworzących tzw. operacyjne drzewo decyzyjne. Pierwszym węzłem tego drzewa jest identyfikacja celu analizy: badacz musi rozstrzygnąć, czy jego intencją jest opis struktury jednej zmiennej (statystyka opisowa), weryfikacja różnic między wieloma grupami (testy istotności dla średnich lub wariancji), czy też modelowanie współzależności i poszukiwanie związków przyczynowo-skutkowych (korelacja i regresja).

Drugim, krytycznym punktem zwrotnym jest określenie skali pomiarowej zmiennej zależnej. Jeżeli cecha ma charakter jakościowy i jest mierzona na skali nominalnej bądź porządkowej, badacz zostaje automatycznie skierowany do gałęzi metod nieparametrycznych, takich jak test zgodności chi-kwadrat lub analiza korelacji rangowej Spearmana. Jeśli zmienna jest ciągła (skala przedziałowa lub ilorazowa), kluczowym krokiem staje się weryfikacja założeń o normalności rozkładu (np. testem Shapiro-Wilka) oraz jednorodności wariancji (np. testem Bartletta). Spełnienie tych kryteriów otwiera drogę do metod parametrycznych (testy t-Studenta, ANOVA, regresja liniowa). W przypadku ich naruszenia, algorytm decyzyjny nakazuje odrzucić metody klasyczne i zastosować ich odporne, nieparametryczne odpowiedniki.

10.2. Zbiorczy schemat interpretacji raportów komputerowych

Bez względu na stopień skomplikowania uruchomionej procedury matematycznej, program Statistica opiera swoje arkusze wynikowe na zunifikowanym systemie raportowania, którego właściwe odczytanie decyduje o sukcesie analizy. Kluczowym elementem, na którym analityk musi skupić uwagę w pierwszej kolejności, jest skojarzona z danym testem p-wartość (*p-value*), prezentowana w komórkach tabeli. Program Statistica stosuje bardzo intuicyjne rozwiązanie wizualne: wszystkie współczynniki i wartości statystyk, dla których prawdopodobieństwo testowe jest mniejsze od standardowego poziomu istotności ($p < 0{,}05$), są automatycznie podświetlane na kolor czerwony. Czerwony kolor w systemie Statistica jest dla badacza bezpośrednim sygnałem oznaczającym odrzucenie określonej hipotezy zerowej

i przyjęcie hipotezy alternatywnej, co pozwala na natychmiastowe stwierdzenie istotności statystycznej badanych różnic, wpływów czy korelacji.

Interpretacja ta musi być jednak prowadzona ze ścisłym uwzględnieniem kierunku weryfikacji. W testach istotności (np. porównanie średnich lub współczynników regresji) dąży się do uzyskania niskiego, czerwonego p-value ($p < 0{,}05$), gdyż dowodzi to istnienia rzeczywistych, nieprzypadkowych efektów. Odwrotna logika obowiązuje natomiast podczas testowania założeń modelu (tzw. testy diagnostyczne). W testach normalności czy jednorodności wariancji pożądanym wynikiem jest wysokie p-value, zaznaczone w programie standardową, czarną czcionką ($p > 0{,}05$). Czarny kolor czcionki oznacza w tym przypadku brak podstaw do odrzucenia hipotezy zerowej, co potwierdza, że dane spełniają kryteria normalności i homoskedastyczności, dając badaczowi zielone światło do kontynuowania zaawansowanego modelowania parametrycznego.

10.3. Przegląd najczęstszych błędów interpretacyjnych i metodologicznych

Końcowa faza komputerowej analizy danych wiąże się z ryzykiem popełnienia krytycznych błędów interpretacyjnych, wynikających najczęściej z niezrozumienia filozofii testowania statystycznego. Najpowszechniejszym błędem metodologicznym jest tzw. utożsamianie braku podstaw do odrzucenia hipotezy zerowej z jej ostatecznym udowodnieniem. Jeśli program Statistica zwraca wysokie p-value ($p > 0{,}05$), oznacza to wyłącznie, że zebrana próba losowa jest zbyt mała lub zbyt mało zróżnicowana, aby wykryć badaną zależność – nie jest to jednak dowód na absolutny brak tej zależności w populacji. W rzetelnych raportach naukowo-badawczych należy unikać kategorycznych sformułowań, że zmienne „nie różnią się" lub „nie korelują", zastępując je poprawnym zapisem mówiącym, że „nie potwierdzono istotnych różnic" lub „nie znaleziono podstaw do odrzucenia hipotezy o braku korelacji".

Drugim poważnym błędem jest ignorowanie kontekstu założeń resztowych w modelowaniu regresyjnym. Często analitycy bezkrytycznie uznają model za doskonały tylko na podstawie wysokiej wartości współczynnika determinacji (np. $R^2 \approx 0{,}95$), całkowicie pomijając fakt, że test Durbina-Watsona wykazał silną autokorelację reszt, a test Shapiro-Wilka odrzucił normalność składnika losowego. Taki model, mimo wysokiego dopasowania do danych historycznych, posiada zerową wartość prognostyczną i doprowadzi do błędów w przypadku próby prognozowania przyszłych zjawisk. Należy być świadomym, że

komputerowa analiza danych to proces holistyczny, w którym poprawność formalna całego łańcucha założeń jest ważniejsza niż pojedynczy, satysfakcjonujący wskaźnik liczbowy.

Bibliografia

[1] Aczel A.D. *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa 2010.

[2] Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K., *Statystyka opisowa. Przykłady i zadania*, CeDeWu, Warszawa 2024.

[3] Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K., *Statystyka matematyczna. Przykłady i zadania*, CeDeWu, Warszawa 2023.

[4] Dobosz M., *Wspomagana komputerowo statystyczna analiza wyników badań*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2004.

[5] Ferguson G.A., Takane Y., *Analiza statystyczna w psychologii i pedagogice*, Wydawnictwo Naukowe PWN, Warszawa 2003.

[6] King B.M., Minium E.W., *Statystyka dla psychologów i pedagogów*, Wydawnictwo Naukowe PWN, Warszawa 2020.

[7] Luszniewicz A. (red.), *Statystyka w zarządzaniu*, Wydawnictwo WSFiZ w Białymstoku, Białystok 2008.

[8] Olszewska A.M., Madras-Kobus B., *Opis statystyczny w zarządzaniu. Zbiór zadań*, Oficyna Wydawnicza Politechniki Białostockiej, Białystok 2025.

[9] Rabiej M., *Statystyka z programem Statistica*, Helion, Gliwice 2012.

[10] Sobczyk M., *Statystyka opisowa*, Wydawnictwo C.H. Beck, Warszawa 2010.

[11] Sobczyk M., *Statystyka matematyczna*, Wydawnictwo C.H. Beck, Warszawa 2010.

[12] Stanisz A. *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 1: Statystyki podstawowe*, StatSoft Polska, Kraków 2006.

[13] Stanisz A., *Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny. Tom 2: Liniowe i nieliniowe modele regresji oraz analiza wariancji*, StatSoft Polska, Kraków 2007.

[14] Stanisz A., *Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny. Tom 3: Analizy wielowymiarowe*, StatSoft Polska, Kraków 2007.

[15] Tarka D., Olszewska A.M., *Elementy statystyki. Opis statystyczny*, Oficyna Wydawnicza Politechniki Białostockiej, Białystok 2018.

[16] Walesiak M., Gatnar E., Bąk A., *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

[17] Witkowski M., *Statystyka matematyczna w zarządzaniu*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań 2010.

[18] Zeliaś A., *Metody statystyczne*, Polskie Wydawnictwo Ekonomiczne (PWE), Warszawa 2000.