Statystyka w badaniach medycznych

Statystyka w badaniach medycznych

Jerzy A. Moczko,
Grzegorz H. Bręborowicz,
Ryszard Tadeusiewicz

Springer PWN
Warszawa 1998

Redakcja
Elżbieta Wieteska

Indeks
Małgorzata Teperek

Projekt okładki
Andrzej Przygodzki

Springer PWN
Wydanie I
Arkuszy drukarskich 8
Skład i łamanie: Auto Graf, Warszawa
Druk ukończono w maju 1998 r.
Druk i oprawa: Toruńskie Zakłady Graficzne „Zapolex"
Toruń, ul. Gen. Sowińskiego 2/4

ISBN 83-86637-96-X

Spis treści

Wstęp

Truizmem jest stwierdzenie, że wszyscy często posługują się statystyką, jednak mało kto ją tak naprawdę zna i rozumie. Wśród ludzi mających na codzień niewielki kontakt z naukami ścisłymi utarło się wręcz przekonanie, że statystyka jest pewną techniką matematyczną pozwalającą na sprytne manipulowanie danymi, tak by w majestacie nauki móc przedstawiać uzyskane wyniki w sposób wygodny dla eksperymentatora. Powstało nawet powiedzenie (często przypisywane Oscarowi Wilde'owi): *„małe kłamstwo, łgarstwo, statystyka"*, mające obrazować, jak perfidnie metody statystyczne ukrywają istotę rzeczy. Tymczasem statystyka jest takim samym działem matematyki, jak trygonometria, algebra, rachunek różniczkowy i całkowy itp. jest bowiem częścią rachunku prawdopodobieństwa. Wykorzystują ją niemal na każdym kroku inne nauki ścisłe takie jak fizyka czy chemia; jest również niezbędna w naukach technicznych, a prawidłowo używana stanowi podstawowe narzędzie ekonomisty (jako tak zwana ekonometria). Nikt z nas nie zaneguje stwierdzenia, iż gwałtowny rozwój techniki zmienił całkowicie oblicze tego świata, a stało się to z pewnością poniekąd także za sprawą statystyki. Również fizyka ciała stałego oraz mechanika kwantowa, dzięki którym nastąpił tak kolosalny przewrót w technologii, bazują w dużej mierze na rachunku prawdopodobieństwa. Wreszcie to, co napędza świat – to znaczy pieniądz i ekonomia – także podlegają bardzo surowemu osądowi statystyki i są za jej pomocą sterowane i kontrolowane. Skoro więc statystyka „sprawdza się" w wielu ważnych dziedzinach, to może nie powinno się jej traktować z tak dużą nieufnością? Może fakt, że statystyki się nadużywa (niestety) częściej niż na przykład rachunku różniczkowego nie powinien obciążać samej dziedziny, lecz tych, którzy nadużyć się dopuszczają?

Próbując przychylniej i życzliwiej spojrzeć na statystykę w pierwszej kolejności dostrzegamy, że jest ona naprawdę potrzebna. Zjawiska, które obserwujemy w technice, naukach ścisłych, ekonomii lub biologii i medycynie są zawsze obarczone pewnymi przypadkowymi zakłóceniami. Tylko statystyka jest narzędziem pozwalającym na eliminację skutków tej przypadkowości i na w miarę pewne i skuteczne działanie w niestałym, pełnym nieoczekiwanych zdarzeń świecie. Otacza nas niewyobrażalnie wielka liczba generatorów chaosu, które powodują, że nic nie zdarza się dwa razy tak

samo i każda nowa sytuacja jest w jakimś sensie unikatowa i jedyna w swoim rodzaju. Urządzenia techniczne nigdy nie są wykonywane dokładnie w ten sam sposób, gdyż nigdy nie mamy do czynienia z identycznymi pod każdym względem surowcami ani z dokładnie takim samym procesem technologicznym. Dlatego poszczególne wyprodukowane egzemplarze różnią się między sobą. Równocześnie jednak – w interesie nas wszystkich – inżynier musi zagwarantować, że mimo tych różnic każdy samochód będzie bezpieczny, a każdy telewizor będzie odbierał program TV i nie będzie wymagał ciągłych napraw. Jedynym sposobem osiągnięcia tej pewności działania niepewnej techniki jest statystyczna kontrola jakości. Podobnie dokonując jakiegokolwiek pomiaru czy obserwacji, musimy zdawać sobie sprawę, że mierzona czy obserwowana wielkość nie jest tą prawdziwą, obiektywną, idealną wartością poszukiwanego parametru czy ujawnionej relacji, dlatego, że narzędzie pomiarowe zawsze wprowadza jakiś błąd, ponieważ akurat ten konkretny obiekt badania miał swoje własne, indywidualne, niepowtarzalne cechy, które omyłkowo przyjęliśmy za wyraz pewnej ogólnej tendencji, wreszcie dlatego, że sami popełniamy różne pomyłki i błędy. A jednak chcemy na podstawie naszych pomiarów i obserwacji budować teorie na temat obiektywnej rzeczywistości, chcemy wierzyć, że poznajemy świat, chcemy konstruować naukę. Jest to niemożliwe bez odpowiedniej statystycznej analizy wyników pomiarowych – jak inaczej odróżnimy to, co stałe i niezmienne od tego, co ulotne i przypadkowe?

Statystyka służy więc do wydobywania prawdy z chaosu, ochrony przed skutkami niepewności wynikającej z przypadkowości wielu ważnych czynników i pomaga uzyskać pewność i skuteczność w warunkach niepewności. Do tego naprawdę nie wystarczy sam tylko zdrowy rozsądek ani intuicja – tu konieczne jest wsparcie matematyki, a narzędziem matematycznym zapewniającym to wsparcie jest właśnie statystyka.

Okazuje się jednakże, że nie tylko technika i nauki ścisłe posługują się statystyką. Niezwykle intensywnie z narzędzi statystycznych zaczęły korzystać również psychologia, socjologia oraz nauki rolnicze. Niemal wszystkie prezentowane w czasopismach medycznych czy biologicznych publikacje naukowe, przedstawiające wyniki badań eksperymentalnych, stosują mniej lub bardziej skomplikowane procedury statystyczne. Dla przeciętnego lekarza, mającego stosunkowo nikły kontakt z matematyką, okazują się one jednak z reguły mało zrozumiałe, w związku z czym fragmenty tekstu zawierające dyskusję statystyczną, po prostu przy czytaniu pomija.

Podobny problem powstaje przy pisaniu publikacji. Szanujące się czasopismo medyczne, biologiczne czy przyrodnicze nie przyjmie obecnie do druku artykułu zawierającego wyniki badań doświadczalnych bez odpo-

wiedniej ich weryfikacji matematycznej. W tej sytuacji autor publikacji musi skorzystać z usług zawodowego statystyka. Wprawdzie dzięki coraz szerszej dostępności mikrokomputerów i świetnego oprogramowania statystycznego mógłby sam dokonać obliczeń – ale do tego potrzebna jest pewna wiedza. Podręczniki statystyki mogą tej wiedzy dostarczyć, są one jednak pisane z reguły przez statystyków dla statystyków i w związku z tym najeżone wzorami matematycznymi, które wprawiają laików w stan osłupienia połączonego z przerażeniem. Fakt ten jest powodem, dla którego dla większości lekarzy i biologów arkana statystyki pozostają obszarem „wiedzy tajemnej".

Tak jednak być nie musi. Aby sensownie używać statystyki, nie trzeba wcale znać jej zasad od strony używanego aparatu matematycznego czy szczegółów algorytmów obliczeniowych. Współczesne komputerowe pakiety statystyczne są tak dalece przyjazne dla użytkownika (*user-friendly* lub jak kto woli *idiot-proof*), że z reguły nie ma on żadnych trudności w przeprowadzenia obliczeń. Wykonuje je komputer, sam stosując wszystkie potrzebne wzory, formuły i tabele, których użytkownik takiego programu wcale nie musi znać ani się nimi interesować. Nie znaczy to jednak, że można samemu prowadzić obliczenia statystyczne bez żadnych ograniczeń. Problemy wymagające określonego poziomu fachowej wiedzy nieuchronnie się pojawią, jednakże nie przy ustalaniu jak liczyć, ale co i w jakim celu – a więc na przykład przy doborze odpowiedniego testu statystycznego i przy interpretacji wyników. Tego żaden komputer nigdy sam nie zrobi i taką właśnie – operacyjną – wiedzę statystyczną trzeba koniecznie posiadać, by bezpiecznie pokonywać drogi i bezdroża statystyki.

Na rynku wydawniczym znajduje się wiele świetnych pozycji literaturowych, zarówno z zakresu samej statystyki, jak i opisujących posługiwanie się konkretnymi programami (pakietami statystycznymi). Zamiarem autorów niniejszej książki nie było więc napisanie kolejnego dzieła tego rodzaju, ponieważ nie wydawało się to potrzebne. Chcieliśmy natomiast zaproponować Czytelnikowi informację, pozwalającą na świadome i skuteczne dokonywanie właściwego wyboru tej najlepszej i najbardziej poprawnej metody obliczeniowej spośród setek stosowanych i dostępnych w odpowiednich programach technik statystycznych. Będziemy przy tym wyraźnie kierować się przydatnością określonych technik w badaniach medycznych, co jest o tyle warte podkreślenia, że chcąc korzystać z tej książki w innych zastosowaniach (na przykład w ekonomii czy socjologii) Czytelnik musi zweryfikować rekomendowane tu techniki i metody, uwzględniając metodologię uprawianej dyscypliny naukowej.

Jednak nawet ograniczenie do sfery statystyki medycznej i biologicznej (zwanej często także biometrią) nie do końca definiuje zawartość tej

książki. Dodajmy więc dla uściślenia, że chcemy w niej odpowiedzieć na dwa najczęściej zadawane pytania:

– jakiego typu testy należy stosować w najczęściej spotykanych sytuacjach badawczych;

– jak należy interpretować wyniki uzyskane z obliczeń statystycznych.

Tylko tyle i aż tyle wiadomości chcemy tu zawrzeć. Nie ma w tej książce niemal żadnych wzorów, ich wyprowadzeń ani dowodów. Zakładamy, że wzory powinni znać twórcy oprogramowania realizującego obliczenia statystyczne, natomiast użytkownik programu, w szczególności lekarz lub biolog – może się bez nich z powodzeniem obyć. Nieco prostej symboliki matematycznej zdecydowaliśmy się wprowadzić jedynie tam, gdzie wzór w sposób istotny ułatwia zrozumienie tekstu. W wielu miejscach zrezygnowaliśmy też ze ścisłości matematycznej na korzyść intuicji. Chcielibyśmy bowiem, aby niniejsza książeczka była dla Czytelnika czymś w rodzaju przewodnika, ułatwiającego zrozumienie i poprawne stosowanie podstawowych testów statystycznych. Ponieważ jednak zdecydowana większość dostępnych pakietów statystycznych (programów komputerowych) jest angielskojęzyczna (w tym sensie, że po angielsku są podawane przez program wszystkie informacje i proponowane opcje) zdecydowaliśmy się w wybranych miejscach podawać oprócz polskiej terminologii statystycznej odpowiedniki w języku angielskim. Powinno to ułatwić posługiwanie się pakietami statystycznymi i świadome osiąganie założonych celów.

Całość tekstu jest ilustrowana konkretnymi przykładami z zakresu medycyny, co powinno ułatwić przyswojenie podawanych reguł i ogólnych zaleceń. Gorąco zachęcamy Czytelnika do przeliczania podanych w tekście przykładów za pomocą posiadanego przez Niego oprogramowania statystycznego. Pozwoli to na ściślejsze powiązanie przedstawianej tutaj teorii z codzienną praktyką badań naukowych.

Pojęcia podstawowe

Skala pomiarowa

Punktem wyjścia do obliczeń statystycznych jest zawsze zgromadzony zbiór danych. Mają one różny charakter – mogą to być spostrzeżenia, wyniki badań, obserwacje itp. W celu ujednolicenia terminologii jednak wszystkie te informacje nazywać będziemy **wynikami pomiarów** (*measurements*). Przyjmując taki termin zgadzamy się z poglądem, że w swojej pracy zawodowej (a także w codziennym życiu) stale dokonujemy rozmaitych pomiarów, angażując do tego czasami mniej lub bardziej skomplikowaną aparaturę pomiarową, a częściej po prostu nasze narządy zmysłów. Gdy na targu przyglądamy się owocom, starając się wybrać te najdorodniejsze, a jednocześnie najtańsze – to dokonujemy (nie zdając sobie z tego sprawy) pewnego pomiaru i pewnej procedury klasyfikacyjnej. Jeśli w praktyce klinicznej mierzymy tętno, ciśnienie skurczowe i rozkurczowe, wykonujemy testy laboratoryjne, osłuchujemy pacjenta, zbieramy od niego dane do wywiadu chorobowego itd. – to także dokonujemy serii pomiarów. Nawet gdy obserwując preparat mikroskopowy szkicujemy podstawowe elementy obserwowanego obrazu i próbujemy rozpoznać badaną tkankę – również jest to pewien pomiar. Wszystkie wymienione tu czynności (oraz setki innych, które składają się na naszą aktywność zawodową i pozazawodową) nie są niczym innym niż pomiarami, mimo iż każdy z nich dokonywany jest w inny sposób i za pomocą innych narzędzi.

Rozmaite są również dane uzyskiwane w ich wyniku – rozmaite pod względem przenoszonych treści i ich znaczenia, a także – co dla nas tutaj będzie znacznie ważniejsze – pod względem charakteru. Niektóre z nich są *stricte* danymi liczbowymi, takimi jak wzrost, waga, tętno, ciśnienie skurczowe i rozkurczowe, niektóre zaś opisowymi (np. kolor skóry, stopień zadyszki, płeć). Dane liczbowe zwane są również danymi **ilościowymi** (*quantitative data*), dane opisowe natomiast – **jakościowymi** (*qualitative data*). Rozróżnienie to jest dość istotne, ponieważ w przypadku danych ilościowych jest możliwe łatwe i naturalne prowadzenie obliczeń matematycznych (można dane np. uśredniać, ustalać ile razy jedna jest większa od

drugiej, badać ich różnice albo swobodnie wstawiać je do wzorów matematycznych). W odniesieniu do danych opisowych takie operacje są absolutnie niedopuszczalne – o czym często się zapomina, stosując statystykę w sposób amatorski – i dostarczają absolutnie bezsensownych wyników.

Dane ilościowe są w statystyce szczególnie wygodne i użyteczne, ponieważ niosą ze sobą informację zarówno o kierunku wzrostu natężenia cechy, jak i o odległości między wynikami pomiarów. Ich odległość w każdym punkcie skali jest równoważna. Różnica masy ciała pacjenta równa 5 kg jest taka sama, gdy weźmiemy parę pomiarów 65 i 60 kg czy też 82,5 i 77,5 kg. Skala o tego typu własnościach nazywana jest **skalą interwałową** (*interval scale*). Umożliwia ona całkowicie swobodne stosowanie dowolnych technik statystycznych.

Dane jakościowe natomiast trzeba traktować w statystyce ze szczególną ostrożnością, gdyż nie wszystkie techniki i metody obliczeń statystycznych mają dla nich dobrze zdefiniowany sens. W dodatku dane jakościowe mogą się znacznie między sobą różnić. Występują wśród nich dane czysto kategorialne, a więc takie, które pozwalają zaliczyć pomiar do jakiejś klasy obiektów bez jakiegokolwiek ich uporządkowania (np. płeć pacjenta, jednostka chorobowa, na którą cierpi, rasa – biała, czarna, żółta – wyznanie religijne). O danych takich mówimy, że zostały wyrażone w **skali nominalnej** (*nominal scale*). Możemy tylko stwierdzić, że są takie same lub że się różnią między sobą, bez sensu natomiast są w tym przypadku takie pojęcia, jak „większy" czy „mniejszy".

Drugi typ danych jakościowych pozwoli zarówno na dokonanie podziału kategorialnego, jak i na określenie kierunku wzrostu natężenia cechy charakteryzującej podział. Jeżeli pytamy pacjenta o natężenie bólu, możemy uzyskać odpowiedzi typu: ból słaby, umiarkowany, silny, nie do wytrzymania. W przeciwieństwie do poprzedniego typu danych jakościowych, gdzie nie było sensu ani możliwości określenia kierunku wzrostu natężenia mierzonej cechy (nie możemy np. powiedzieć, że osoba chora na gruźlicę jest mniej lub bardziej chora od osoby chorej na boreliozę) – tu możemy wyraźnie uporządkować wyniki w kierunku np. wzrostu natężenia bólu. Oczywiście pomiar ten może być (i najczęściej jest) pomiarem subiektywnym, niemniej jednak możemy go poddać dalszej analizie. Odpowiadającą takiemu przypadkowi skalę pomiarową nazywamy **skalą porządkową** (*ordinal scale*).

Często wyniki pomiaru w skali porządkowej są wyrażane symbolicznie liczbami, co staje się źródłem błędów interpretacyjnych. Weźmy dla przykładu serię wyników opisujących stan noworodka w pierwszej minucie po porodzie wyrażony w jedenastopunktowej skali wg V. Apgar. Użyte tu

liczby są bardziej symbolami niż prawdziwymi liczbami w jakich wyrażamy pomiary w skali interwałowej, takie jak masa ciała, wzrost, wiek itp. Nie mamy bowiem zdefiniowanej odległości między poszczególnymi punktami na skali. Mimo że liczbowo różnica między stanem 0 i 2 jest taka sama, jak między 8 a 10, to z klinicznego punktu widzenia oznaczają one zupełnie co innego. Jak wiadomo, pierwsza para liczb opisuje zróżnicowanie między noworodkiem nieżywym a noworodkiem w stanie bardzo ciężkim, druga natomiast – między noworodkiem z niewielkim odchyleniem od stanu prawidłowego a w pełni zdrowym. Intuicyjnie czujemy, że „odległość" między elementami pierwszej pary jest zdecydowanie większa od „odległości" pary drugiej. Każdy jednak pakiet statystyczny, do którego wprowadzimy serię wyników wyrażonych w skali Apgar, będzie traktował obie „odległości" jako identyczne.

Z podobnych przyczyn błędem byłoby obliczenie średniej arytmetycznej z serii pomiarów, których wyniki są wyrażone w skali porządkowej. Gdy przyjrzymy się jednak literaturze położniczej, często niestety znajdziemy wśród wyników obliczoną średnią arytmetyczną z serii pomiarów w skali Apgar. Nie ma to oczywiście najmniejszego sensu i jest ewidentnym błędem. Jakiego typu miernika należałoby użyć dla tej skali – pokażemy w rozdziale poświęconym **statystyce opisowej** (*descriptive statiscs*). Zapamiętajmy, że odległość między kategoriami w skali porządkowej nie musi być identyczna (i najczęściej nie jest) z odległością między odpowiadającymi im liczbami, które służą wyłącznie jako kody pomiarowe. Pamiętać o tym musi prowadzący badania, opracowujący wyniki program komputerowy nie będzie w tym pomocny. Dla komputera liczba to liczba, jeśli więc dostanie on serię danych, będących zakodowanymi liczbowo danymi rejestrowanymi w skali porządkowej, to oczywiście na żądanie obliczy nam średnią arytmetyczną, geometryczną, wariancję, odchylenie standardowe – traktując wyniki pomiarów jako standardowe liczby, a nie symbole. Dlatego w statystycznym opracowywaniu wyników badań biologicznych i medycznych istnieje ścisły podział kompetencji: komputer wykonuje obliczenia i prezentuje wyniki (na przykład zestawia je w postaci atrakcyjnych graficznie wykresów) badacz natomiast (i tylko on!) określa, jakie obliczenia należy wykonać, na jakich danych i co zrobić z wynikami. Musi on więc rozumieć opisany wyżej podział na różne typy skal pomiarowych – komputer będzie wprawdzie za niego sprawnie i wydajnie liczył, nie będzie jednak w stanie dokonywać istotnych wyborów – czyli po prostu myślał!

Zapamiętajmy: Każdy typ danych pomiarowych wymaga stosowania innych technik statystycznych. Możemy przykładowo chcieć oszacować najbardziej reprezentatywny wynik pomiaru dla danej grupy. W przypadku

danych jakościowych możemy to zrobić jedynie przez wyznaczenie najczęściej występującego w niej wyniku; sięgnąć po wygodne (ale często nadużywane!) oszacowanie wartości średniej możemy dopiero wtedy, gdy jesteśmy pewni, że mamy do czynienia z danymi ilościowymi. Podobne rozróżnienie jest potrzebne, gdy chcemy stwierdzić, jak rozproszone są uzyskane wyniki pomiarów wokół wyniku najbardziej reprezentatywnego (na przykład średniego) lub jaki jest rozkład mierzonej cechy. Odpowiedzi na te pytania są domeną statystyki opisowej.

Często parametry uzyskane w wyniku zastosowania statystyki opisowej porównuje się, aby stwierdzić, czy zachodzi pomiędzy nimi określona relacja (np. czy w grupie pacjentów z nadciśnieniem stosujących dany lek hipotensyjny ciśnienie tętnicze jest istotnie niższe niż w grupie pacjentów z nadciśnieniem otrzymujących placebo). Możemy również chcieć zbadać dynamikę zmian określonego parametru w czasie. Dobrym przykładem może być próba odpowiedzi na pytanie, jak zmienia się w czasie wartość parametru FEV w grupie osób z astmą oskrzelową leczonych preparatem Intal. Tego typu zagadnieniami zajmuje się dział statystyki zwany testowaniem hipotez (*hypothesis testing*). Tutaj także obowiązuje podział na metody, które można stosować wyłącznie do danych ilościowych i metody (z reguły nieco bardziej złożone), które można stosować także do danych jakościowych*.

Problemem, z którym wyjątkowo często spotykamy się w biologii i medycynie, jest stwierdzenie czy między pewną liczbą zmierzonych cech istnieje zależność i czy jest to zależność przypadkowa, czy istotna statystycznie. Spotykamy się bowiem często z pytaniami w rodzaju: Czy istnieje zależność masy urodzeniowej noworodka od poziomu estriolu u ciężarnej zmierzonym w okresie okołoporodowym? Jeśli tak, to czy jest istotnie wyrazem jakiejś prawidłowości, czy też jest dziełem przypadku? Czy jest dodatnia (im wyższy poziom estriolu, tym wyższa masa urodzeniowa), czy ujemna (im wyższy poziom estriolu, tym niższa masa urodzeniowa)? Czy jest liniowa (punkty odpowiadające poszczególnym pomiarom układają się mniej lub bardziej dokładnie wzdłuż linii prostej), czy nieliniowa (punkty leżą wzdłuż jakiejś innej linii, na przykład parabolicznej (efekt nasila się dla większych poziomów estriolu) lub logarytmicznej (przy dużych poziomach

---

* Jednakże proszę zauważyć, że z danych ilościowych zawsze można łatwo zrobić dane jakościowe dowolnego typu (na przykład wprowadzając jakąś kategoryzację do wyrażonych ilościowo pomiarów; w najprostszym przypadku można je podzielić na „duże", „średnie" i „małe"). Dlatego wszystkie metody i testy statystyczne, które dają się zastosować do danych jakościowych, można także bez trudu zastosować do danych ilościowych, natomiast odwrotna procedura jest absolutnie niemożliwa.

efekt słabnie)? Odpowiedzi na tego typu pytania udziela **analiza korelacji i regresji** (*correlation and regression analysis*).

Aby usystematyzować omówione do tej pory zagadnienia, a przez to ułatwić wybór odpowiedniego testu statystycznego, podamy serię prostych przykładów. Na początku tego rozdziału stwierdziliśmy, że pojęcie pomiaru jest nadzwyczaj szerokie, a dane uzyskane w wyniku akcji pomiarowej mogą mieć różne właściwości. Właściwości te wpływają – z oczywistych względów – na dalsze możliwości wnioskowania. Zacznijmy od stwierdzenia podstawowego i w oczywisty sposób prawdziwego, że pomiar ilościowy niesie więcej informacji niż pomiar jakościowy. Na czym jednak dokładnie to polega?

Dla przykładu rozważmy pomiar temperatury ciała pacjentów. Mając odpowiedni przyrząd dokonujemy pomiaru ilościowego, wyrażając jego wynik w skali interwałowej, jaką jest skala Celsjusza (czy też używana w krajach anglosaskich skala Fahrenheita). Na podstawie tak przeprowadzonego pomiaru możemy podzielić pacjentów np. na trzy grupy: z temperaturą prawidłową, grupę ze stanem podgorączkowym i grupę z temperaturą wysoką. Co więcej, możemy dokładnie określić, o ile stopni miał temperaturę wyższą pacjent pierwszy od drugiego. Jeśli zastosujemy jakąś formę terapii, której efektem będzie obniżenie temperatury, będziemy mogli pokusić się o określenie szybkości spadku temperatury po zażyciu leku (na przykład w stopniach na godzinę) lub powiązać wielkość obniżenia się temperatury z dawką leku. Ilościowy pomiar temperatury umożliwi nam zatem ustawienie pomiarów w kierunku wzrostu natężenia cechy (temperatura normalna, stan podgorączkowy, temperatura wysoka) oraz określenie odległości między punktami pomiarowymi (np. pacjent pierwszy ma temperaturę wyższą o 1,3 stopnia Celsjusza od pacjenta drugiego). Tymczasem gdybyśmy w miejsce pomiaru ilościowego zastosowali pomiar jakościowy (pacjent „ma gorączkę" lub „nie ma gorączki"), możliwości dalszego wnioskowania i subtelniejszych ocen byłyby dużo skromniejsze.

Przytoczony wcześniej przykład odczuwania bólu przez pacjenta jest pomiarem, którego wyniki wyrażono w skali porządkowej. Jeżeli pierwszy pacjent określa ból związany z zabiegiem jako słaby, a drugi jako bardzo silny, to możemy stwierdzić, że zapewne drugi pacjent cierpi bardziej od pacjenta pierwszego. To, że stopień odczuwania bólu może być własnością osobniczą (zależną np. od predyspozycji psychicznych pacjenta, motywacji, nastroju) nie zmienia tego faktu. Nie potrafimy jednakże określić, o ile pacjent pierwszy jest bardziej cierpiący – nie mamy zatem określonej odległości między wynikami pomiarowymi. Pomiar w skali porządkowej dostarcza nam więc wyłącznie informacji o kierunku wzrostu natężenia mierzonej

cechy, nic natomiast nie mówi o odległościach między punktami pomiarowymi.

Najsłabsza ze skal pomiarowych, skala nominalna, mówi jeszcze mniej. Pomiary wykonane z użyciem tej skali dostarczają wyłącznie informacji, do jakiej grupy kategorii należy badany obiekt, nie pozwalają natomiast na uporządkowanie tych kategorii, a tym bardziej na określenie odległości między nimi. Dla przykładu pomiar polegający na określeniu płci osób pewnej grupy pozwala wyłącznie na zliczenie przypadków przynależnych do danej kategorii (płci), bezsensowna jest natomiast każda próba uporządkowania tych kategorii według stopnia natężenia cechy pomiarowej lub odległości między cechami.

Należy pamiętać, że zawsze możemy dokonać redukcji skali pomiarowej, lecz nigdy nie przejdziemy od skali pomiarowej słabszej do silniejszej bez powtórzenia pomiaru. Przykładowo, jeżeli w grupie pacjentów dokonaliśmy pomiaru wzrostu w skali interwałowej (tzn. wyraziliśmy wzrost w metrach lub centymetrach), to zawsze możemy przejść do skali porządkowej zakładając jakiś podział kategorialny w skali porządkowej (np. wzrost niski – do 160 cm włącznie, średni – od 160 cm do 175 cm włącznie, wysoki – powyżej 175 cm). Za pomocą takiego odwzorowania każdemu pomiarowi w skali interwałowej jednoznacznie przyporządkowujemy jego odpowiednik w skali porządkowej. Warto jednak zwrócić uwagę, że po dokonaniu odwzorowania i skasowaniu danych źródłowych odwzorowanie odwrotne nie będzie możliwe. Mając serię pomiarów wykonanych w skali porządkowej, nie potrafimy przejść do skali interwałowej – chociaż czasem możemy „udawać", że mamy do czynienia ze skalą interwałową (na przykład przypisując każdemu osobnikowi z danej kategorii wartość wzrostu typową dla danej kategorii). Takie „udawanie" może czasem pozwolić na przeprowadzenie pewnego użytecznego wnioskowania, trzeba jednak pamiętać, że podobnie jak nie jesteśmy w stanie sprawić, by wszyscy ludzie średniego wzrostu mieli dokładnie 167,5 cm, tak samo nie możemy stosować metod statystycznych właściwych dla skali interwałowej po dokonaniu kategoryzacji pomiarów. Aby mieć prawo do stosowania wygodnych i potężnych narzędzi statystycznych właściwych dla danych ilościowych, musimy w rozważanym przykładzie powtórzyć badanie w skali interwałowej.

Podobne rozumowanie można przeprowadzić dla przejść między pomiarami w skali interwałowej i nominalnej oraz porządkowej i nominalnej. Podsumowując, możemy zawsze przejść ze skali silniejszej do słabszej, co jednak nieuchronnie łączy się z utratą informacji, nigdy zaś nie możemy skutecznie przejść ze skali słabszej do silniejszej. Planując jakieś badania, pamiętajmy zawsze by wykonywać pomiary w najsilniejszej z możliwych

skalach pomiarowych gdyż może to nas uchronić przed koniecznością powtarzania badań, w razie potrzeby stosowania procedur statystycznych wymagających większej ilości informacji.

W dalszych rozdziałach zobaczymy, że określenie skali pomiarowej jest nieodzowne do prawidłowego doboru testu statystycznego. Raz jeszcze podkreślamy, że wyboru tego nie dokona za nas komputer i pełna odpowiedzialność spoczywa zawsze na badaczu.

Wspomniane w tym rozdziale działy statystyki (statystyka opisowa, testowanie hipotez i badania siły związku statystycznego) nie wyczerpują wszystkich jej możliwości, lecz można śmiało powiedzieć, że w badaniach medycznych znajdują najczęstsze zastosowanie. W związku z tym, tymi właśnie działami zajmiemy się w dalszej części podręcznika.

Populacja i próba statystyczna

Statystyka jest narzędziem, które umożliwia obiektywne wnioskowanie na podstawie wyników serii eksperymentów. Eksperymenty przeprowadzane są na pewnym zbiorze elementów. Zbiory takie będziemy nazywali populacjami. Jeżeli liczba elementów podlegającego badaniu zbioru jest skończona, to **populację** nazywamy **skończoną** (*finite population*). W przeciwnym razie mówimy o **populacji nieskończonej** (*infinite population*). W eksperymentach medycznych niemal zawsze będziemy mieli do czynienia z populacjami skończonymi. Jeżeli prowadzone przez nas badanie będzie obejmowało wszystkie elementy populacji, to określimy je jako **wyczerpujące**. Badania wyczerpujące dostarczają niewątpliwie więcej informacji niż tzw. badania cząstkowe, omówione w dalszej części rozdziału, nie zawsze jednak można je przeprowadzić. Niemożliwe jest np. przeprowadzenie badania wyczerpującego, gdy pomiar łączy się ze zniszczeniem lub zużyciem obiektu badania. Gdybyśmy chcieli metodą wyczerpującą zbadać stopień zanieczyszczenia otrzymanej partii insuliny, nie zostałoby nam leku do prowadzenia leczenia. Czasami badanie wyczerpujące może być zbyt trudne do wykonania lub zbyt kosztowne. Na przykład nie można dokonać badań wyczerpujących na populacji ludzi dorosłych na całym świecie, chociaż dostarczyłyby one z pewnością ciekawych wyników. Podobny problem występuje także wtedy, gdy badana przez nas populacja ewoluuje w czasie wykonywania badania.

W wymienionych przypadkach nasuwa się więc następujące rozwiązanie. Spróbujmy z badanej populacji wybrać pewien podzbiór elementów i wykonajmy badanie na tym podzbiorze. Takie badanie nazywamy **cząstkowym**, a wybrany podzbiór określamy jako **próbę statystyczną** (*statistical sample*). Wnioski, które wyciągniemy na podstawie wyników badania próby, będziemy chcieli potem odnieść do całej populacji. Tą techniką wykonuje się mnóstwo różnych badań statystycznych – na przykład wszelkie rankingi przedwyborcze opierają się właśnie na takich badaniach cząstkowych.

Czy jednak zawsze uzyskane wyniki możemy odnieść do całej populacji? Okazuje się, że wnioskowanie o populacji na podstawie badania próby można przeprowadzić tylko w jednym przypadku – gdy próba ta będzie **reprezentatywna** (*representative sample*). W przeciwnym razie wyniki będą dotyczyły wyłącznie próby, czyli z naukowego i praktycznego punktu widzenia mogą być mniej przydatne.

Teoria doboru próby reprezentatywnej jest złożona i trudna do popularnego ujęcia. Ograniczymy się więc jedynie do skrótowego podania wa-

runków koniecznych, których spełnienie pozwala założyć jej reprezentatywność. Po pierwsze dobór elementów musi być losowy (*random sampling*). Innymi słowy, każdy element populacji musi mieć jednakowe szanse wejścia do próby. Jeżeli nasze badanie dotyczyłoby zmian w układzie nerwowym wywoływanym przez cukrzycę, a do próby wybralibyśmy wyłącznie mężczyzn, to tak wybrana próba nie byłaby reprezentatywna (kobiety również chorują na cukrzycę, a ich metabolizm różni się subtelnie w wielu aspektach od metabolizmu mężczyzn)*.

Podobnie nie dozwolone przy konstruowaniu próby reprezentatywnej są wszelkie inne preferencje nie uzasadnione przyczynami natury merytorycznej (na przykład preferencje dla wieku – wybór wyłącznie starszych pacjentów może istotnie zafałszować obraz zjawiska i należy go unikać, chociaż może się okazać, że prowadzenie badań na osobach w podeszłym wieku jest łatwiejsze, ponieważ mają więcej czasu i na ogół chętniej odwiedzają lekarzy). Ogólnie należy dążyć do tego, by w wybranej próbie rozkład wszystkich cech był możliwie najbardziej zbliżony do rozkładu tych cech w całej badanej populacji. Oczywiście dotyczy to w praktyce tylko tych cech, które możemy łatwo ustalić (wiek, płeć, środowisko, stan rodzinny, nałogi itp.), więc wybrana próba zawsze jest w jakimś sensie mało reprezentatywna (bo rozkład tych cech, których nie kontrolujemy, może być w badanej próbie zdecydowanie inny niż w populacji). Tego błędu do końca uniknąć się nie da, gdyż w celu wybrania w pełni reprezentatywnej próby trzeba by było najpierw wykonać badania wyczerpujące...

W praktyce stosuje się losowanie przypadków, które włączymy do próby, z pełnej listy elementów branej pod uwagę populacji**. Można to sobie ułatwić wykorzystując odpowiednie programy komputerowe. Czasem proces losowania zastępujemy procesem pseudolosowego wyboru (na przykład możemy włączać do badań co dziesiątego zgłaszającego się pacjenta),

---

* Wybór samych kobiet natomiast będzie mógł być uznany za reprezentatywny np. w przypadku analizowania nowotworów szyjki macicy.

** W wielu przypadkach prawidłowe zdefiniowanie populacji nie jest łatwe. Wyobraźmy sobie, że przeprowadziliśmy badania kliniczne na małej grupie pacjentów (próbie) chorych na stosunkowo rzadko występującą jednostkę chorobową. Uzyskane wyniki chcielibyśmy uogólnić i przenieść na populację. Co jednakże stanowi taką populację: czy są to wszyscy pacjenci z tą jednostką chorobową, którzy zgłaszają się do ośrodka, w którym prowadzimy badanie, czy też wszyscy pacjenci z tą jednostką chorobową z terenu miasta, regionu, kraju, kontynentu, świata? Zwróćmy przy tym także uwagę na fakt, że populacja osób leczonych w określonych wiodących szpitalach klinicznych (przykładowo w uniwersyteckich) będzie zasadniczo różniła się od populacji osób leczonych w typowych szpitalach miejskich. W tych pierwszych bowiem z oczywistych względów częściej napotkamy przypadki bardziej skomplikowane, nietypowe, powikłane itp.

mając nadzieję, że losowy proces pojawiania się pacjentów zgłaszających się po poradę w jakimś stopniu imituje prawidłowe losowanie. Należy jednak zdawać sobie sprawę, że losowość próby jest wtedy bardzo problematyczna i w badaniach, których wyniki są szczególnie istotne należy ją zweryfikować. Do losowego wyboru składu próby, a także do zbadania, czy wybrana próba wykazuje (z punktu widzenia cech, które uznamy za znaczące) znamiona próby losowej – można użyć ogólnie dostępnych programów statystycznych, które z reguły potrafią wspomóc użytkownika przy konstruowaniu próby losowej i przy weryfikacji jej losowości.

Drugi warunek pozwalający na traktowanie wyników uzyskanych dla próby jako wiarygodnych dla całej populacji jest związany z odpowiednią liczebnością próby*. Oszacowanie minimalnej wielkości próby opiera się na wynikach dość skomplikowanych obliczeń i zależy od wielu czynników (m.in. od tego, jak mocno zaznacza się badana przez nas cecha na tle losowych czynników maskujących jej działanie). Jeśli zaznacza się silnie z dużą powtarzalnością, a czynniki przypadkowe nie występują albo są mało istotne – można poprzestać na próbie o niezbyt dużej liczebności. Gdy jednak manifestuje się w sposób bardzo subtelny i ulotny, a równocześnie na wynikach obserwacji prowadzonych dla różnych pacjentów bardzo silnie ważą ich cechy osobnicze oraz inne czynniki losowe, wówczas potrzebujemy naprawdę licznej próby – w razie jej braku nie będziemy mogli przeprowadzić rzetelnego wnioskowania. Tworząc próbę statystyczną, musimy bezwarunkowo podchodzić bardzo poważnie do zagadnienia prawidłowego określenia jej minimalnej liczebności**.

---

* Intuicyjnie wyczuwamy, że im większa będzie liczebność badanej próby, tym bardziej uzyskane w wyniku jej badania parametry będą zbliżone do tych, które uzyskalibyśmy badając całą populację. Zwiększenie liczebności próby podwyższa jednak koszt i wydłuża czas realizacji badania.

**Zastosowanie zbyt licznej próby niczym w tym zakresie nie grozi – zawsze lepiej jest poczynić więcej obserwacji, niż za mało. Pojawia się jednak wtedy problem kosztów badań (wszak decydując się na badanie próby, a nie populacji, czynimy to właśnie po to, by uzyskać potrzebne odpowiedzi mniejszym kosztem). Jeśli z braku odpowiedniej wiedzy i doświadczenia zaplanujemy badania na zbyt licznej próbie – wyniki będą niewątpliwie dobre i prawdziwe, tyle tylko, że zapłacimy za nie o wiele więcej niż było trzeba!

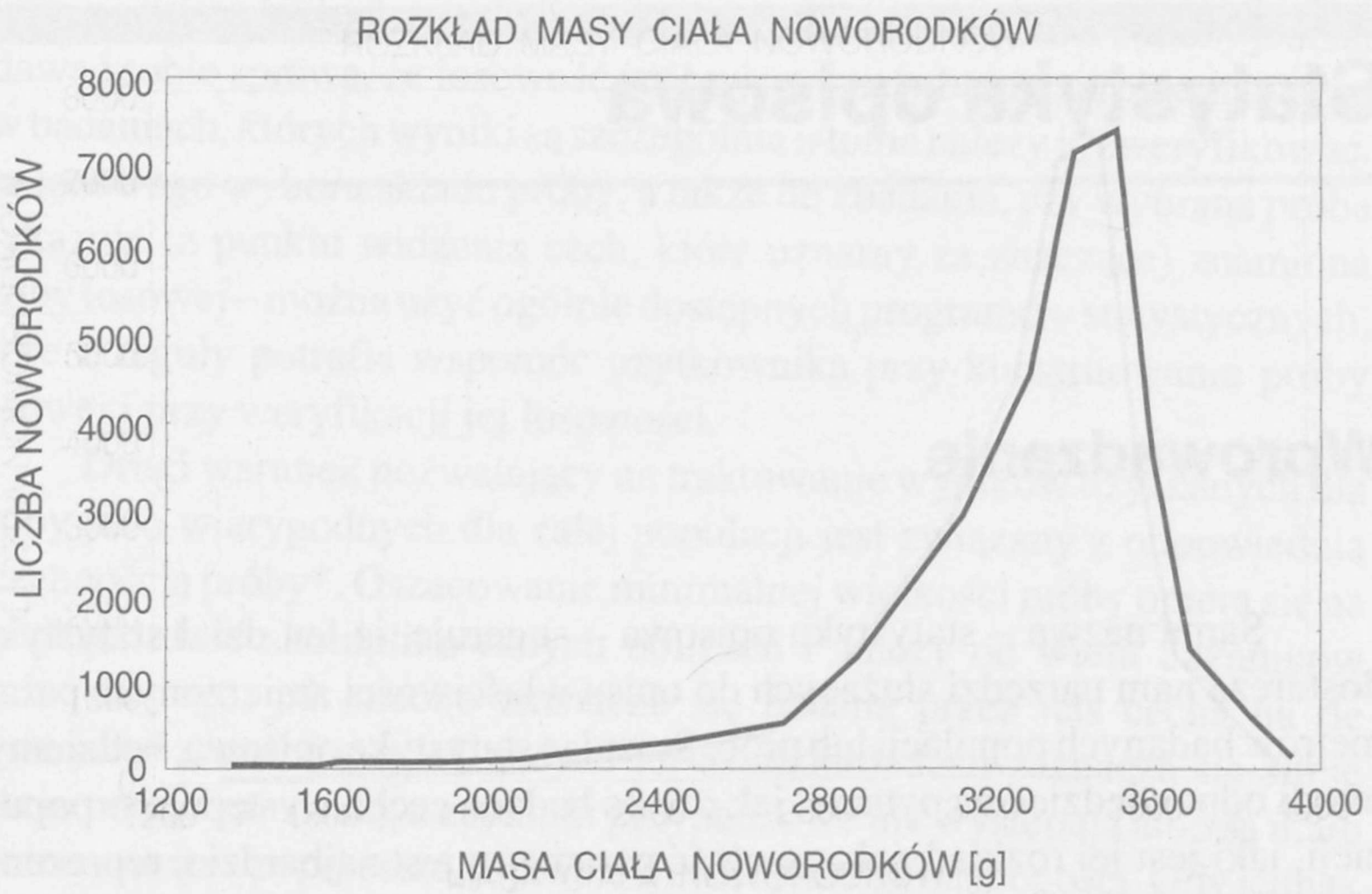
Statystyka opisowa

Wprowadzenie

Sama nazwa – statystyka opisowa – sugeruje, iż ten dział statystyki dostarcza nam narzędzi służących do opisu właściwości zmierzonych parametrów badanych populacji lub prób. Stosując statystykę opisową, będziemy mogli odpowiedzieć na pytanie, jak często badana cecha występuje w populacji, jaki jest jej rozkład, jaka wartość parametru jest najbardziej reprezentatywna w analizowanej próbie itp. Wartości uzyskane w wyniku zastosowania statystyki opisowej są również podstawą dla dalszych, bardziej zaawansowanych metod wnioskowania statystycznego, takich jak testowanie różnic między grupami czy też badanie związków między wybranymi cechami.

Kluczową rolę w statystyce opisowej odgrywają dwie grupy miar: **miary tendencji centralnej** (*central tendency measures*) oraz **miary rozproszenia** (*dispersion measures*). Wszystkie miary tendencji centralnej oszacowują „najbardziej typową" wartość parametru charakteryzującą badaną grupę, lecz każda z nich skupia się na innym aspekcie jego rozkładu. Do najczęściej stosowanych miar tendencji centralnej należą: **średnia arytmetyczna, mediana i wartość modalna**. Zajmiemy się nimi w dalszej części rozdziału, nie omówimy natomiast innych, również niekiedy spotykanych miar tendencji centralnej, takich jak średnia geometryczna i (o wiele rzadziej stosowana) średnia harmoniczna.

Należy pamiętać, że nie jest rzeczą obojętną, którą z miar wybierzemy do scharakteryzowania badanej przez nas populacji. W zrozumieniu istotnych różnic między miarami tendencji centralnej pomoże nam pojęcie **rozkładu danych** (*data distribution*). Wyobraźmy sobie, że dokonujemy pomiaru masy ciała noworodka w bardzo dużej grupie przypadków (dane wyrażone w skali interwałowej). Jeżeli wykreślimy liczbę zmierzonych wartości masy ciała noworodka (oś zmiennej zależnej y) w funkcji wartości masy ciała (oś zmiennej niezależnej x), to otrzymamy krzywą przedstawioną na rycinie 1. Tak skonstruowana krzywa przedstawia rozkład badanej cechy w próbie, a więc w naszym przypadku rozkład masy ciała noworodka



Ryc. 1. Rozkład masy ciała noworodków

w zbadanej przez nas grupie przypadków. W rozpatrywanym tu przykładzie im więcej dokonamy pomiarów, tym bardziej uzyskana krzywa będzie przypominała symetryczny dzwon. Jeżeli (co jest oczywiście rzeczą niemożliwą) przeanalizowalibyśmy grupę o nieskończonej liczbie przypadków, otrzymalibyśmy **rozkład normalny** (*normal distribution*) zwany również rozkładem Gaussa. Jego właściwości omówimy w dalszej części książki. W praktyce badana przez nas próba zawsze ma ograniczoną liczbę elementów, lecz warto zapamiętać, że im więcej elementów włączymy do naszego badania, tym lepiej zostanie odwzorowany rzeczywisty rozkład badanej zmiennej w analizowanej populacji.

Nie wszystkie mierzone wielkości mają rozkład normalny. Dlatego podamy teraz, jakie mogą być inne typy rozkładów i ich właściwości. Rozkład danych jest **ciągły** (*continuous*), jeśli badany parametr może przyjmować nieskończenie wiele leżących nieskończenie blisko siebie wartości lub **dyskretny** (*discrete*), jeśli liczba dostępnych wartości jest ściśle określona. Masa ciała, wzrost, wiek, ciśnienie skurczowe i rozkurczowe, temperatura to przykładowe zmienne opisywane rozkładem ciągłym (np. w zakresie temperatur 36–42° Celsjusza istnieje nieskończenie wiele wartości pomiarowych, z których my, jedynie ze względu na dokładność naszego narzędzia pomiarowego, jakim jest termometr, wybieramy do rozważań

pewną skończoną ich liczbę). Przykładami rozkładów ciągłych używanych w statystyce są – obok opisanego rozkładu normalnego – rozkłady **t-Studenta, F-Fishera-Snedecora,** rozkład **chi-kwadrat.**

W rozkładach dyskretnych z natury rzeczy (a nie z powodu ograniczeń wynikających z metody pomiaru) występuje skończona liczba wartości pomiarowych. Kwalifikacja płci, ocena przeżycie – zgon, liczba dzieci w rodzinie, liczba poronień – to typowe rozkłady dyskretne. Dla przykładu, liczba dzieci w rodzinie może przyjmować wartości całkowite: 1, 2, 3,..... nigdy ułamkowe: 1,3; 2,7 itp. Jakkolwiek dokładnie przeprowadziliśmy pomiar, to liczba dzieci będzie zawsze przyjmowała wartości całkowite. Przykładami rozkładów dyskretnych mogą być rozkład **dwumianowy** (*binomial distribution*) i jego graniczna postać – **rozkład Poissona** (*Poisson distribution*).

Dlaczego wobec tego akurat rozkład normalny odgrywa kluczową rolę w statystyce, skoro jest tak wiele innych? Znaczenie rozkładu normalnego wynika nie tylko z tego, że reprezentuje on szczególnie dużo obserwowanych rozkładów, lecz także z powodu jego szczególnej roli w teorii doboru próby. Na ogół w miarę wzrostu liczby analizowanych przypadków ich rozkład staje się coraz bliższy rozkładowi normalnemu (dokładny rozkład normalny jest osiągalny dopiero dla nieskończonej liczby przypadków). Co więcej, jeśli na końcowy, obserwowany przez nas wynik badania, ma wpływ wiele niezależnych czynników (taka sytuacja zdarza się często w badaniach biologicznych i medycznych), to wypadkowe zjawisko ma rozkład normalny nawet wtedy, gdy poszczególne wpływające na jego przebieg czynniki mają (obiektywnie) inne rozkłady. Własność ta jest konsekwencją niezwykle ważnego w statystyce **centralnego twierdzenia granicznego** (*central limit theorem*). Rozkładowi normalnemu poświęcimy w naszych dalszych rozważaniach szczególnie dużo miejsca.

Po wprowadzeniu pojęcia rozkładu i skal pomiarowych możemy przyjrzeć się bliżej właściwościom poszczególnych miar tendencji centralnej oraz miar rozproszeń.

Miary tendencji centralnej

Średnia arytmetyczna

Najbardziej znaną i najczęściej stosowaną (choć nie zawsze słusznie) miarą tendencji centralnej jest **średnia arytmetyczna** (*arithmetic mean*). Jej popularność wynika między innymi z nieskomplikowanej metody oszacowania, polegającej na zsumowaniu wszystkich pomiarów i podzieleniu sumy przez ich liczbę. Już sam sposób obliczania wskazuje, że można ją stosować wyłącznie wtedy, gdy wyniki pomiarów są wyrażone w skali interwałowej, konieczna jest bowiem informacja o odległości między punktami pomiarowymi – zarówno w przypadku skali porządkowej, jak i nominalnej tą informacją nie dysponujemy.

Przykład 1

Wyobraźmy sobie, że zmierzyliśmy wzrost pacjentów, a wyniki przedstawiliśmy w skali porządkowej. Wyglądają one następująco: 5 osób o wzroście niskim, 10 o średnim, 7 wysokich, 3 bardzo wysokie.

W wyniku pomiaru nie dowiedzieliśmy się, jaki konkretnie jest wzrost poszczególnych osób, a jedynie, czy mieści się on w pewnym określonym przedziale. Obliczenie średniej arytmetycznej jako (5+10+7+3)/4 jest oczywiście bezsensowne, gdyż daje nam informację nie o średnim wzroście, lecz o średniej liczbie badanych w poszczególnych grupach wzrostu. Widzimy zatem, że obliczenie średniej arytmetycznej dla wyników przedstawionych w skali porządkowej nie jest możliwe.

Ciekawe są również właściwości średniej arytmetycznej. Pierwszą z nich jest jej wysoka czułość na skrajne wartości wyników pomiarów.

Przykład 2

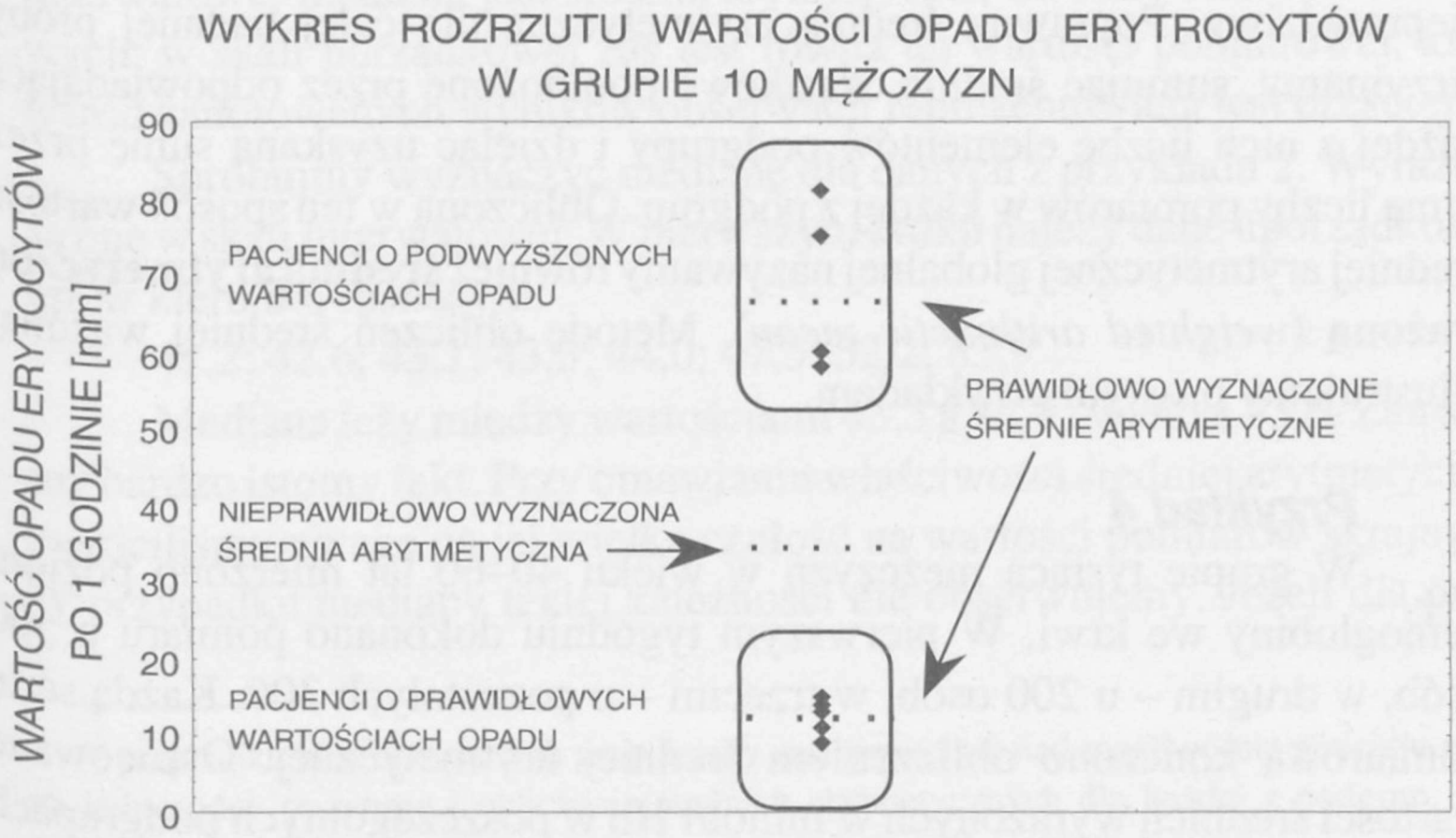
Pomiar masy ciała w 8-osobowej grupie dzieci dał następujące wyniki wyrażone w kilogramach: 41,2; 47,5; 52,2; 43,3; 44,0; 83,9; 42,6; 43,1. Średnia arytmetyczna masy ciała w tej grupie wynosi zatem 49,7 kg (pomiaru dokonano z dokładnością do 0,1 kg, z taką też dokładnością możemy więc oszacować wartość średnią). Błędem byłoby podanie wartości średniej równej 49,72 kg lub 49,725 kg). Przypatrzmy się jednak szóstemu pomiarowi. Jego wynik wyraźnie odbiega od wyników pozostałych pomiarów. Gdybyśmy go wyeliminowali, średnia masy ciała wynosiłaby (w 7-osobowej grupie dzieci) 44,8 kg. Zwróćmy uwagę, że wynik pojedynczego pomiaru zmienił wartość średnią o mniej więcej 11%. Wpływ pojedynczego „odstającego" wyniku pomiaru na wartość średniej arytmetycznej zależy zarówno od

stopnia jego oddalenia od pozostałych wyników (im większe oddalenie, tym większa deformacja średniej), jak i od liczby pomiarów w grupie (im większa liczba pomiarów, tym mniejszy wpływ pojedynczego odstającego wyniku). Musimy oczywiście pamiętać, że wynik może być zarówno znacznie większy (jak w naszym przykładzie), jak i zdecydowanie mniejszy od pozostałych wartości. W tym ostatnim przypadku otrzymalibyśmy oczywiście zaniżenie wartości średniej arytmetycznej.

Bardzo często odstające od reszty wyniki pomiarów określa się jako „błędy grube" i eliminuje z obliczeń. Nie zawsze jest to jednak słuszne. Pojawienie się takiego wyniku (lub całej grupy tego typu wyników) może świadczyć o **niejednorodności badanej grupy** (*nonhomogenous group*). W naszym przykładzie wartość 83,9 kg jest rzeczywiście podejrzana, jednakże jej źródłem może być nie tylko błąd pomiarowy, lecz np. fakt, że grupa dzieci poddanych badaniu była niejednorodna pod względem wieku lub sposobu odżywiania się rodzin, z których pochodziły.

Przykład 3

W grupie 10 mężczyzn w wieku powyżej 50 lat zaobserwowano następujące wartości opadu erytrocytów po 1 godzinie (wyrażone w mm): 11, 9, 58, 14, 75, 13, 14, 60, 81, 15. Obliczenie średniej arytmetycznej opadu w tej grupie daje wynik 35. Jeżeli przedstawimy te dane graficznie (ryc. 2), od razu zauważymy niejednorodność ich rozkładu. Część z nich tworzy



Ryc. 2. Przykład prawidłowego i nieprawidłowego obliczania średniej arytmetycznej dla niejednorodnej grupy pomiarów

6-elementową podgrupę prawidłowych wartości opadu po pierwszej godzinie (11, 9, 14, 13, 14, 15), pozostałe – 4-elementową grupę o wartościach podwyższonych (58, 75, 60, 81). Podwyższonych wartości opadu nie traktujemy jako błędów grubych, lecz stwierdzamy w rozważanej grupie pacjentów niejednorodność ze względu na wartość opadu. Obliczona przez nas uprzednio wartość średnia, mimo iż prawidłowa z matematycznego punktu widzenia, w tej sytuacji nie daje się zinterpretować medycznie. Prawidłowe postępowanie powinno polegać na policzeniu dwóch średnich arytmetycznych: jednej dla grupy o wartościach prawidłowych (13), drugiej – dla grupy o wartościach podwyższonych (68).

Przy ocenie jednorodności analizowanych danych nieocenione usługi oddaje prosty **wykres rozrzutu punktów pomiarowych** (*scatterplot*). Z reguły pozwala on na wykrycie skupień wyników pomiarów i daje możliwość uniknięcia pomyłek w oszacowaniu parametrów statystyki opisowej.

Zdarza się, że czasami otrzymujemy wyniki porcjami. Dla każdej z nich obliczamy średnią arytmetyczną. Czy na podstawie tak otrzymanych **średnich arytmetycznych cząstkowych** (*partial arithmetic means*) jesteśmy w stanie obliczyć średnią arytmetyczną dla całej badanej próby bez konieczności prowadzenia obliczeń od początku? Okazuje się, że jest to możliwe. Najprostsza sytuacja występuje, gdy liczba pomiarów w każdej porcji była identyczna. Wystarczy wtedy dodać do siebie średnie arytmetyczne cząstkowe, a sumę podzielić przez liczbę podgrup. W praktyce najczęściej zdarza się jednak, że każda z podgrup zawiera inną liczbę wyników. W takiej sytuacji postępowanie opisane poprzednio dałoby wynik nieprawdziwy. Poprawną średnią arytmetyczną dla całej badanej próby otrzymamy, sumując średnie cząstkowe pomnożone przez odpowiadającą każdej z nich liczbę elementów podgrupy i dzieląc uzyskaną sumę przez sumę liczby pomiarów w każdej z podgrup. Obliczoną w ten sposób wartość średniej arytmetycznej globalnej nazywamy również **średnią arytmetyczną ważoną** (*weighted arithmetic mean*). Metodę obliczeń średniej ważonej zilustrujemy prostym przykładem.

Przykład 4

W grupie tysiąca mężczyzn w wieku 40–60 lat mierzono poziom hemoglobiny we krwi. W pierwszym tygodniu dokonano pomiaru u 500 osób, w drugim – u 200 osób, w trzecim – u pozostałych 300. Każdą serię pomiarową kończono obliczeniem średniej arytmetycznej. Oszacowane wartości średnich wyrażonych w mmol/l Hb w poszczególnych podgrupach wynosiły odpowiednio: 9,17; 10,12; 9,85. Aby obliczyć średnią arytmetyczną ważoną w całej 1000-osobowej grupie obliczono wartość wyrażenia

$(500 \times 9{,}17 + 200 \times 10{,}12 + 300 \times 9{,}85)/1000$, co dało w wyniku 9,56. Zwróćmy uwagę, że obliczenie wartości wyrażenia $(9{,}17 + 10{,}12 + 9{,}85)/3$ daje nieprawidłową, zawyżoną wartość średniej arytmetycznej (równą 9,71).

Podsumujmy teraz najważniejsze własności średniej arytmetycznej:

– jest ona miernikiem tendencji centralnej dla pomiarów wykonanych w skali interwałowej,

– jest bardzo czuła na wyniki pomiarów znacznie odbiegające od przeciętnych,

– nie nadaje się do opisu niejednorodnych grup pomiarowych,*

– dobrze opisuje tylko dane o symetrycznym rozkładzie**.

Mediana

Mediana (*median*) jest miarą tendencji centralnej przeznaczoną do opisu danych wyrażonych w skali porządkowej. Oczywiście miara ta może być również wykorzystywana do pomiarów w skali interwałowej, gdzie często bywa bardziej reprezentatywna od nagminnie używanej (i nadużywanej) średniej arytmetycznej. Wyznaczenie mediany opiera się na wykorzystaniu informacji o kierunku wzrostu natężenia cechy, niesionej przez obie wymienione skale pomiarowe. Chcąc wyznaczyć medianę, należy zatem uporządkować posiadane dane w porządku rosnącym (lub malejącym) i wybrać pomiar środkowy. Jest to łatwe i naturalne w przypadku nieparzystej liczby pomiarów. Gdy liczba pomiarów jest parzysta, to w przypadku skali interwałowej mediana jest średnią arytmetyczną dwóch pomiarów środkowych, w skali porządkowej zaś jest równa tej wartości pomiarowej, która wśród zanotowanych wyników obserwacji reprezentowana jest częściej.

Spróbujmy wyznaczyć medianę dla danych z przykładu 2. Wyrażone są one w skali interwałowej. W pierwszym kroku należy dane uporządkować np. w kierunku rosnącym:

41,2; 42,6; 43,1; 43,3; 44,0; 47,5; 52,2; 83,9

Mediana leży między wartościami 43,3 a 44,0 i wynosi 43,6. Zauważmy bardzo istotny fakt. Przy omawianiu właściwości średniej arytmetycznej zwróciliśmy uwagę na jej wielką czułość na wartości pomiarów skrajnych. W przypadku mediany takiej zależności nie obserwujemy. Jeżeli dla przy-

---

* W przypadku niejednorodnych grup należy zastanowić się nad możliwością podziału grupy na jednorodne podgrupy i obliczenia średnich arytmetycznych dla każdej z podgrup, gdyż średnia wyznaczona dla całej grupy może nie mieć prawidłowej interpretacji.

** Gdy rozkład danych pomiarowych nie jest symetryczny, należy rozważyć konieczność użycia innego miernika tendecji centralnej, takiego jak mediana lub średnia geometryczna.

kładu zwiększylibyśmy ostatni pomiar nawet o 1000 jednostek (celowo wprowadzony błąd gruby), nasz szereg danych przyjąłby postać 41,2; 42,6; 43,1; 43,3; 44,0; 47,5; 52,2; 1083,9; ale mediana nie zmieniłaby swojej wartości.

Przykład 5

W grupie 10 pacjentów przeprowadzono ocenę jakościowej próby Benedicta na zawartość glukozy w moczu. Uzyskano następujące wyniki:

+, ++, –, –, –,++++, +, –, +, –.

Wyniki wyrażone są w skali porządkowej, co uniemożliwia użycie średniej jako miary tendencji centralnej. W tej sytuacji wyznaczymy medianę. Uporządkowanie wyników pomiarów w kierunku wzrostu natężenia cechy daje szereg:

–, –, –, –, –,+, +, +, ++, ++++

którego mediana wynosi „–" (pomiar środkowy mieści się między pomiarem piątym „–" a szóstym „+", lecz pomiar „–" pojawia się częściej niż pomiar „+").

Podsumujmy najważniejsze właściwości mediany:

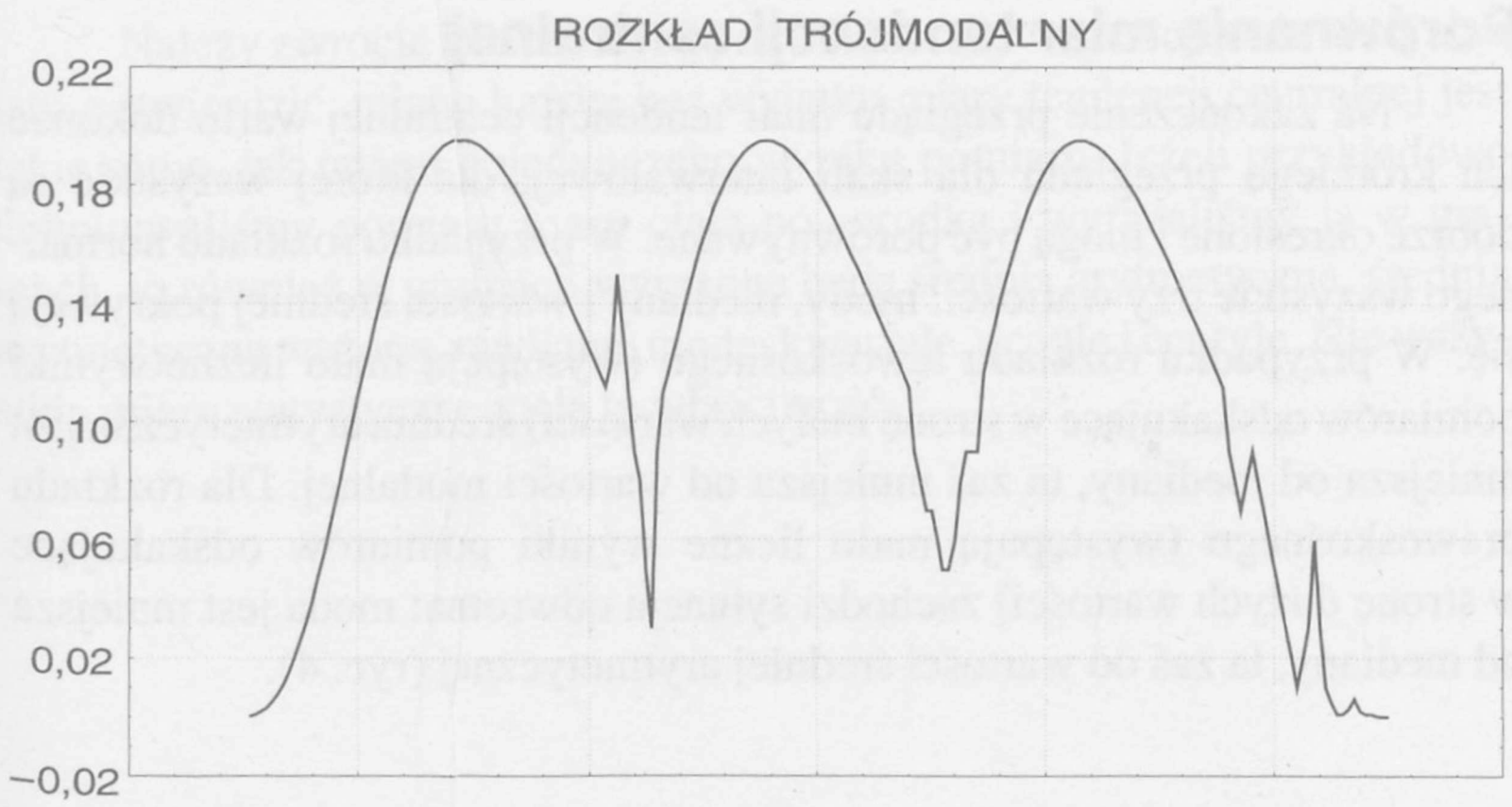
– medianę można wyznaczyć z szeregu danych wyrażonych zarówno w skali interwałowej, jak i porządkowej,

– obliczając ją wykorzystuje się informację o kierunku wzrostu natężenia cechy, a nie o odległości między poszczególnymi wynikami pomiarów, co powoduje, że nie jest ona czuła na wartości pomiarów skrajnych,

– dla skali interwałowej powinno się stosować medianę zamiast średniej arytmetycznej dla tych pomiarów, których rozkład jest wyraźnie skośny.

Mediana dzieli uporządkowany szereg danych na połowy. W analogiczny sposób można wprowadzić wielkości dzielące uporządkowany szereg na cztery równe części (**kwartyle** – *quartiles*), na dziesięć równych części (**decyle** – *deciles*), wreszcie na 100 równych części (**centyle** – *centiles*). Drugi kwartyl, piąty decyl i pięćdziesiąty centyl są równe medianie. Wprowadza się również pojęcie skali percentylowej (*percentile score*), która nie jest jednakże miarą tendencji centralnej, lecz informuje o procencie obserwacji, których wynik jest mniejszy od określonej wartości lub jej równy.

Warto zauważyć, że nie istnieje pojęcie ważonej mediany jako odpowiednika ważonej średniej arytmetycznej. Nie ma zatem możliwości obliczenia mediany dla całej grupy na podstawie wartości median obliczonych dla podgrup.



Ryc. 3. Przykłady rozkładu wielomodalnego

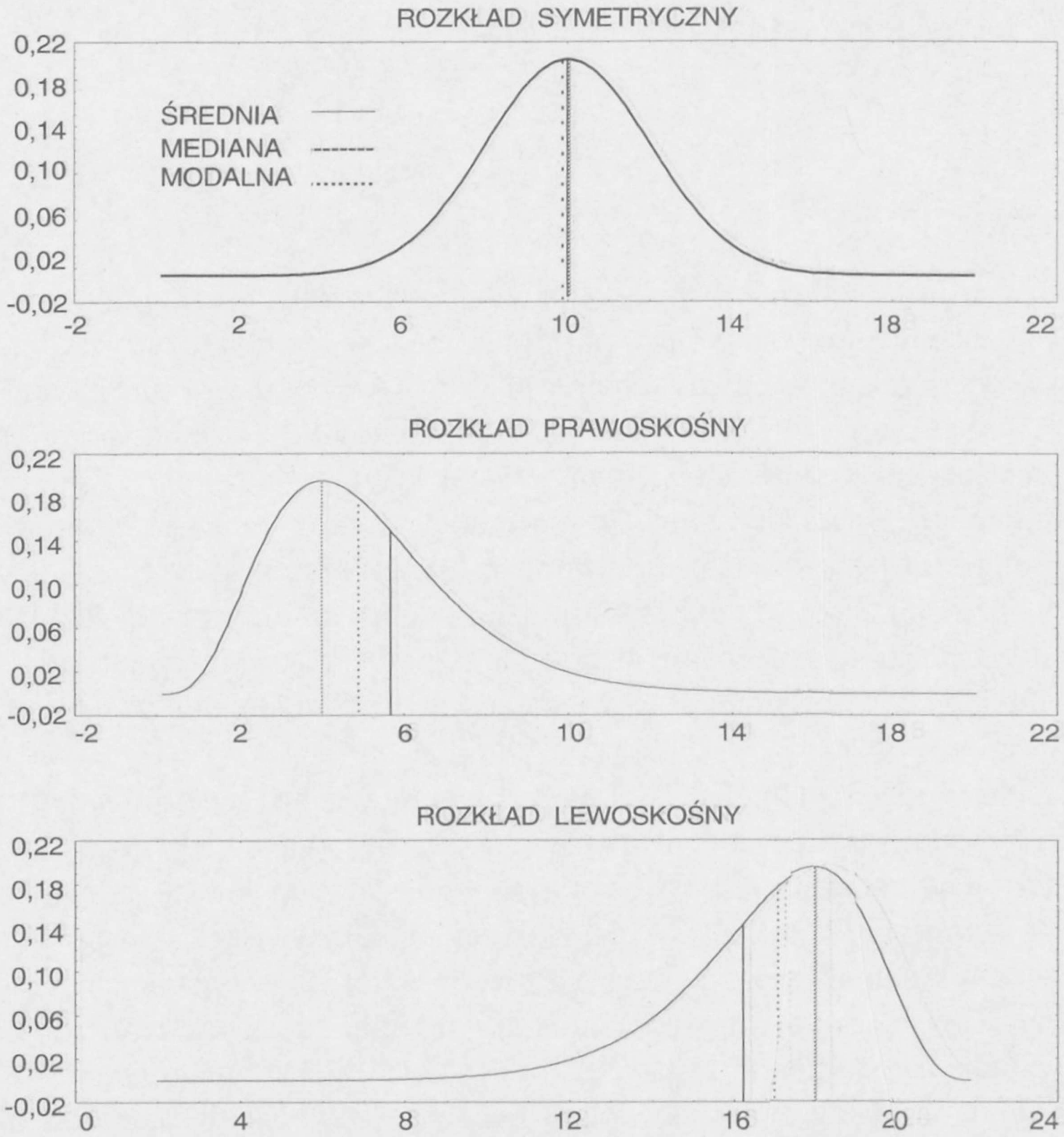
Modalna

Jak już wspomnieliśmy w poprzednich rozdziałach, skala nominalna daje wyłącznie informację o przynależności badanego obiektu do określonej klasy. Nie istnieje tu pojęcie odległości wyników pomiarów ani kierunku wzrostu natężenia cechy. Nie ma zatem możliwości obliczenia w tej skali średniej arytmetycznej ani mediany. W tej sytuacji jako miarę tendencji centralnej wprowadzono **wartość modalną**, zwaną również krótko **modą** lub **dominantą** (*mode*). Jest nią ta wartość, która w szeregu danych powtarza się najczęściej. Jeżeli przedstawimy dane w postaci ich krzywej rozkładu (identycznie jak to zrobiliśmy na ryc. 1), to moda będzie się znajdowała na osi odciętych w punkcie odpowiadającym maksymalnemu wzniesieniu krzywej.

Jak widać, definicja ta jest bardzo nieprecyzyjna, gdyż mogą występować rozkłady bimodalne (dwa wyniki pomiarów pojawiają się jednakowo często), wielomodalne (wiele wyników pomiarów pojawia się jednakowo często), wreszcie rozkłady bez określonej modalnej (wszystkie pomiary występują z jednakową częstością) – ryc. 3. W takich przypadkach jednoznaczne wyznaczenie wartości modalnej jest niemożliwe. Jest to cena, jaką przychodzi często płacić za użycie zbyt „słabej" skali pomiarowej. Warto jednak zwrócić uwagę, iż modalna może być również wyznaczona dla wyników pomiarów wyrażonych w skali porządkowej i interwałowej, jest więc najbardziej uniwersalną miarą tendencji centralnej.

Porównanie miar tendencji centralnej

Na zakończenie przeglądu miar tendencji centralnej warto dokonać ich krótkiego przeglądu dla skali interwałowej, dla której wszystkie są dobrze określone i mogą być porównywane. W przypadku rozkładu normalnego wszystkie trzy wartości: mody, mediany i wartości średniej pokrywają się. W przypadku rozkładu lewoskośnego (występują mało liczne wyniki pomiarów odskakujące w stronę małych wartości) średnia arytmetyczna jest mniejsza od mediany, ta zaś mniejsza od wartości modalnej. Dla rozkładu prawoskośnego (występują mało liczne wyniki pomiarów odskakujące w stronę dużych wartości) zachodzi sytuacja odwrotna: moda jest mniejsza od mediany, ta zaś od wartości średniej arytmetycznej (ryc. 4).



Ryc. 4. Wartość średnia arytmetyczna, mediana i modalnadla rozkładu normalnego i rozkładów skośnych

Należy zwrócić jeszcze uwagę na miano miar tendencji centralnej. Jak łatwo stwierdzić, miano każdej bez wyjątku miary tendencji centralnej jest takie samo, jak miano pojedynczego wyniku pomiaru. Jeżeli przykładowo dokonywaliśmy pomiaru masy ciała noworodka i wyrażaliśmy ją w gramach, to również w gramach wyrażone będą średnia arytmetyczna, średnia arytmetyczna ważona, mediana, moda, kwartyle, decyle i centyle. Nie wszystkie miary statystyczne mają tę właściwość.

Miary rozproszeń

Potrzeba stosowania miar rozproszenia

Omówione w poprzednim rozdziale miary tendencji centralnej wyznaczały pewną wielkość najbardziej reprezentatywną dla badanej próby. W przytoczonym dalej przykładzie zobaczymy, że scharakteryzowanie grupy pomiarów wyłącznie takim miernikiem jest zdecydowanie niewystarczające.

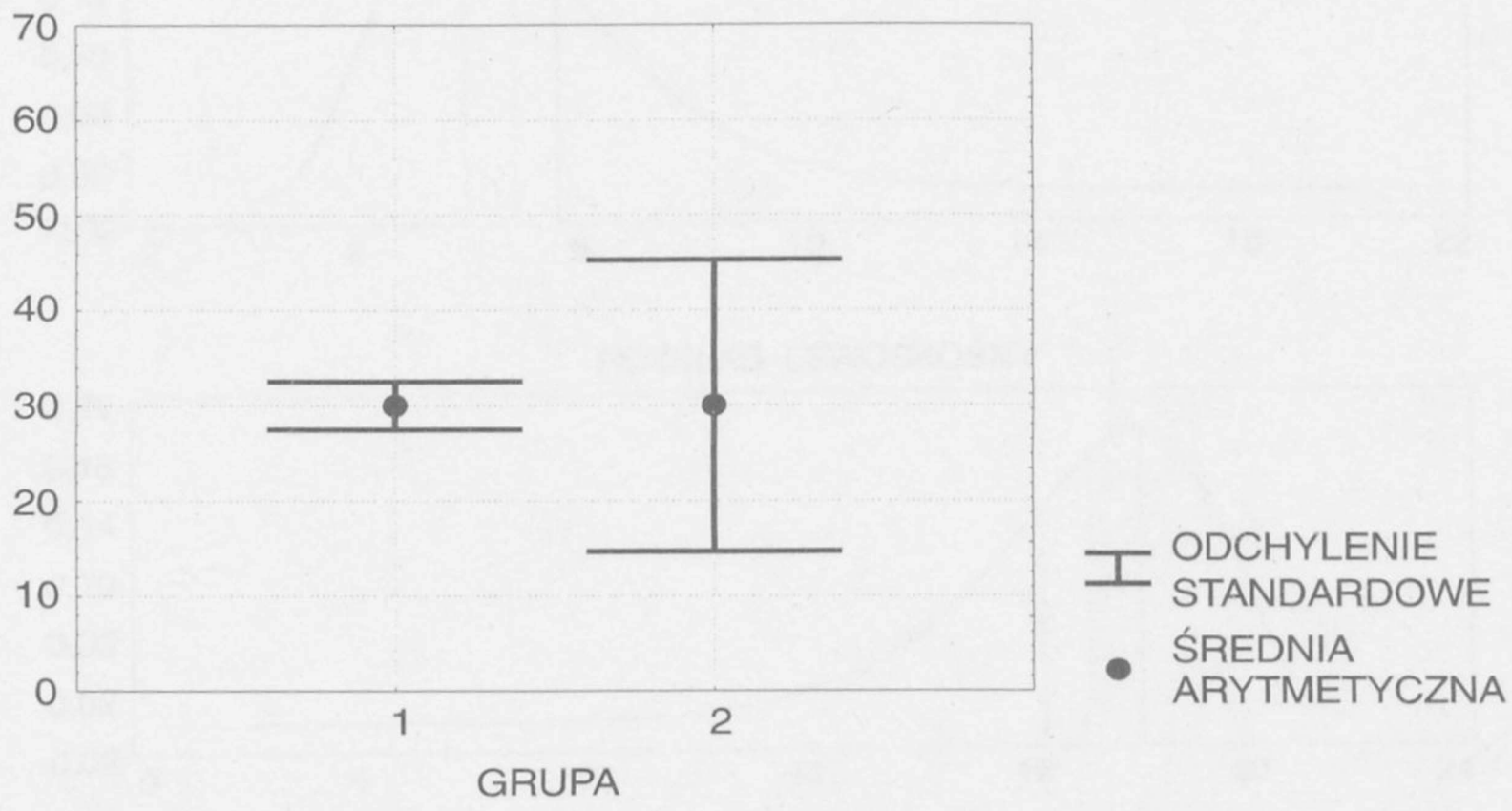
Przykład 6

Weźmy pod uwagę dwie pięcioosobowe grupy pacjentów. W każdej z grup obliczono wartości średniej arytmetycznej:

Grupa 1. Wiek: 30, 28, 27, 33, 32 – średnia arytmetyczna wieku: 30 lat

Grupa 2. Wiek: 12, 48, 28, 19, 43 – średnia arytmetyczna wieku: 30 lat

Z tego prostego przykładu wynika, że zdecydowanie różne zestawy danych mogą mieć identyczne średnie arytmetyczne. Analogiczne przykłady możemy zbudować dla mediany i wartości modalnej. Widać zatem, że samo podanie miary tendencji centralnej może być nie wystarczające do pełnego scharakteryzowania analizowanych danych. Jeżeli narysujemy wykres rozkładu danych w obu grupach, to stwierdzimy, że wyniki pomiarów w grupie pierwszej są bardziej skupione wokół wartości średniej arytmetycznej niż wyniki w grupie drugiej (ryc. 5). Zatem mimo iż średnia arytmetyczna wieku



Ryc. 5. Zestawy danych o identycznej średniej arytmetycznej i o różnym rozproszeniu wokół wartości średniej

w obu grupach jest identyczna, to rozproszenie wyników jest inne. W związku z tym wprowadza się pojęcie **miar rozproszenia** (*dispersion*), które charakteryzują stopień rozrzutu pomiarów wokół miary tendencji centralnej. Im większa jest wartość miary rozproszenia, tym bardziej wyniki pomiarów są rozrzucone wokół miary tendencji centralnej. Na ogół oznacza to również, że miara tendencji centralnej słabiej te wyniki pomiarów reprezentuje, ponieważ są one bardzo zróżnicowane. Niewielkie wartości miary rozproszenia pozwalają z kolei uznać miarę tendencji centralnej za dobrego reprezentanta całej próby.

Odchylenie standardowe

Najczęściej stosowaną miarą rozproszenia jest **odchylenie standardowe** (*standard deviation*). Odchylenie standardowe odzwierciedla stopień rozproszenia pomiarów wokół średniej arytmetycznej. Oczywiście może ono być wyznaczane wyłącznie dla tych danych, dla których możliwe jest poprawne wyznaczenie średniej arytmetycznej. Wyznacza się je, obliczając najpierw różnice między wynikami poszczególnych pomiarów i średnią arytmetyczną, następnie wyznaczone różnice podnosi się do kwadratu, a uzyskane kwadraty sumuje. Powyższą sumę dzieli się przez liczbę pomiarów i z uzyskanego wyrażenia oblicza pierwiastek kwadratowy. Tak zdefiniowana wielkość stanowi oszacowanie **odchylenia standardowego w populacji** (*population standard deviation*). Niekiedy zamiast dzielenia przez liczbę pomiarów dokonuje się dzielenia przez liczbę pomiarów pomniejszoną o jeden*. W ten sposób definiujemy **odchylenie standardowe w próbie** (*sample standard deviation*). Różnica między wartościami tych dwóch wielkości jest tym większa, z im mniejszą próbą mamy do czynienia. Dla przykładu odchylenie standardowe w próbie złożonej z 5 pomiarów będzie około 12% wyższe od odchylenia standardowego w populacji, w próbie 50-elementowej różnica ta spadnie do 1%, w 100-elementowej do 0,5%. Niestety, często się zdarza, że dokumentacja towarzysząca pakietom statystycznym nie precyzuje, czy program oblicza odchylenie standardowe w próbie czy w populacji, co w przypadku małych prób może doprowadzić do niejednoznaczności uzyskiwanych wyników. Jeśli jednak badacz ma wybór i jeśli badania prowadzone były metodą wyboru próby i pomiaru badanego parametru wyłącznie w próbie – jest rzeczą właściwą posługiwanie się wyłącznie techniką obliczania odchylenia standardowego w próbie,

* Dokładne uzasadnienie tego zabiegu opiera się na matematycznej teorii estymacji (pojęcia estymatorów obciążonych i nieobciążonych) i nie jest możliwe do ścisłego przedstawienia w tej książce w związku z założeniem o unikaniu stosowania matematyki wyższej.

a nie odchylenia standardowego w populacji – chociaż na ogół prowadzi to do mniej korzystnych rozstrzygnięć statystycznych (na przykład pewnych interesujących badacza różnic nie można uznać za znamienne statystycznie). Taka jest jednak niestety cena, jaką płaci się za badanie próby zamiast całej populacji. Cena ta – jak łatwo się domyślić – jest tym niższa, im większą próbę uda się przebadać.

Przyjrzyjmy się teraz ogólnym własnościom odchylenia standardowego. Po pierwsze zauważmy, że ta miara rozproszenia ma zastosowanie wyłącznie do interwałowej skali pomiarowej. Oszacowanie rozproszenia w postaci odchylenia standardowego jest możliwe jedynie wówczas, gdy znamy odległości między punktami, a tej informacji nie dostarcza nam ani skala porządkowa, ani nominalna.

Po drugie, gdy w grupie danych wszystkie wyniki pomiarów są identyczne, to oczywiście średnia arytmetyczna jest równa tej stałej wartości i rozproszenie wyników pomiarów wokół średniej jest równe zero. Zatem wartość zerowa odchylenia standardowego wystąpi wyłącznie w przypadku identycznych wyników pomiarów w badanej grupie. Każdy inny rozkład wyników pomiarów ma odchylenie standardowe większe od zera. Z samej zasady obliczeń wynika natomiast, że odchylenie standardowe nie może być nigdy ujemne.

Po trzecie odchylenie standardowe jest wyrażone w tych samych jednostkach, co pomiar – podobnie jak miary tendencji centralnej.

Po czwarte wartość liczbowa odchylenia standardowego może być większa niż wartość średniej arytmetycznej. Bardzo prostym przykładem takiej sytuacji jest szereg danych:

–3,0; –2,0; –1,0; 0,0; 1,0; 2,0; 3,0

Jego wartość średnia wynosi 0,0, podczas gdy odchylenie standardowe jest równe 2,2.

Wysoka wartość odchylenia standardowego może wskazywać na występowanie w badanej grupie wyników pomiarów znacznie odbiegających od reszty (np. błędy grube), na małą dokładność prowadzonego pomiaru (mało precyzyjne urządzenie pomiarowe) lub na istnienie niejednorodności pod względem mierzonego parametru w obrębie tej grupy (podobnie jak w przykładzie 3).

Błąd standardowy średniej arytmetycznej

Z odchyleniem standardowym mylony jest czasami **błąd standardowy średniej arytmetycznej** (SEM – *standard error of the mean*). Nie jest on miarą rozproszenia wyników pomiarowych, lecz określa dokładność, z jaką możemy obliczyć wartość średniej arytmetycznej w populacji, dys-

ponując średnią arytmetyczną w analizowanej próbie. Dokładność ta jest ściśle związana z rozproszeniem pomiarów w próbie i oblicza się ją, dzieląc odchylenie standardowe przez pierwiastek kwadratowy z liczby pomiarów. Widać stąd, że jednostka, w której wyrażamy błąd standardowy średniej, jest taka sama jak jednostka poszczególnego pomiaru czy też średniej arytmetycznej, ale jego wartość jest zawsze mniejsza niż wartość odchylenia standardowego. Błąd standardowy średniej określa rozproszenie całej serii średnich arytmetycznych obliczonych w wielu próbach reprezentatywnych, wybranych z tej samej populacji, wokół prawdziwej średniej arytmetycznej w populacji. Jak widać jest to coś całkiem innego niż rozproszenie samych wyników wokół wartości średniej.

Stwierdziliśmy, że odchylenie standardowe mówi o stopniu rozproszenia pomiarów wokół średniej arytmetycznej. Wyobraźmy sobie następujący problem.

Przykład 7

W dwudziestoosobowej grupie dzieci zmierzono wzrost i masę ciała, uzyskując następujące wyniki:

Wzrost: 152,0 ± 13,2 cm

Masa: 48,3 ± 9,7 kg

(Proszę zwrócić uwagę na powszechnie stosowaną notację: wartość średnia arytmetyczna ± odchylenie standardowe).

Interesuje nas, czy badana grupa jest bardziej zróżnicowana pod względem wzrostu, czy też pod względem masy? Stwierdziliśmy, że im wyższa jest wartość odchylenia standardowego, tym wyższe jest rozproszenie wyników, a więc i ich zróżnicowanie. W związku z tym chciałoby się powiedzieć, że dzieci są bardziej zróżnicowane pod względem wzrostu. I tu popełnilibyśmy błąd. Nie wolno nam bowiem bezpośrednio porównywać wielkości wyrażonych w różnych jednostkach. Stwierdzenie, że odchylenie standardowe 9,7 kg jest mniejsze od 13,2 cm jest oczywiście bezsensowne.

Aby odpowiedzieć na interesujące nas pytanie, wprowadzimy nową miarę rozproszenia, którą jest **współczynnik zmienności** (*variability index*). Definiuje się go jako stosunek odchylenia standardowego do wartości średniej arytmetycznej. Jak widać, współczynnik zmienności nie posiada miana (czasami wyraża się go w procentach, mnożąc obliczony iloraz przez 100%).

Niemianowane wielkości możemy już bez kłopotu porównywać. Wracając do naszego przykładu, współczynnik zmienności dla wzrostu wynosi 13,2/152,0 × 100% = 8,7%, natomiast dla masy ciała 9,7/48,3 × 100% = 20,1%. Badana grupa jest więc przeszło dwukrotnie bardziej zróżnicowana pod względem masy ciała niż pod względem wzrostu.

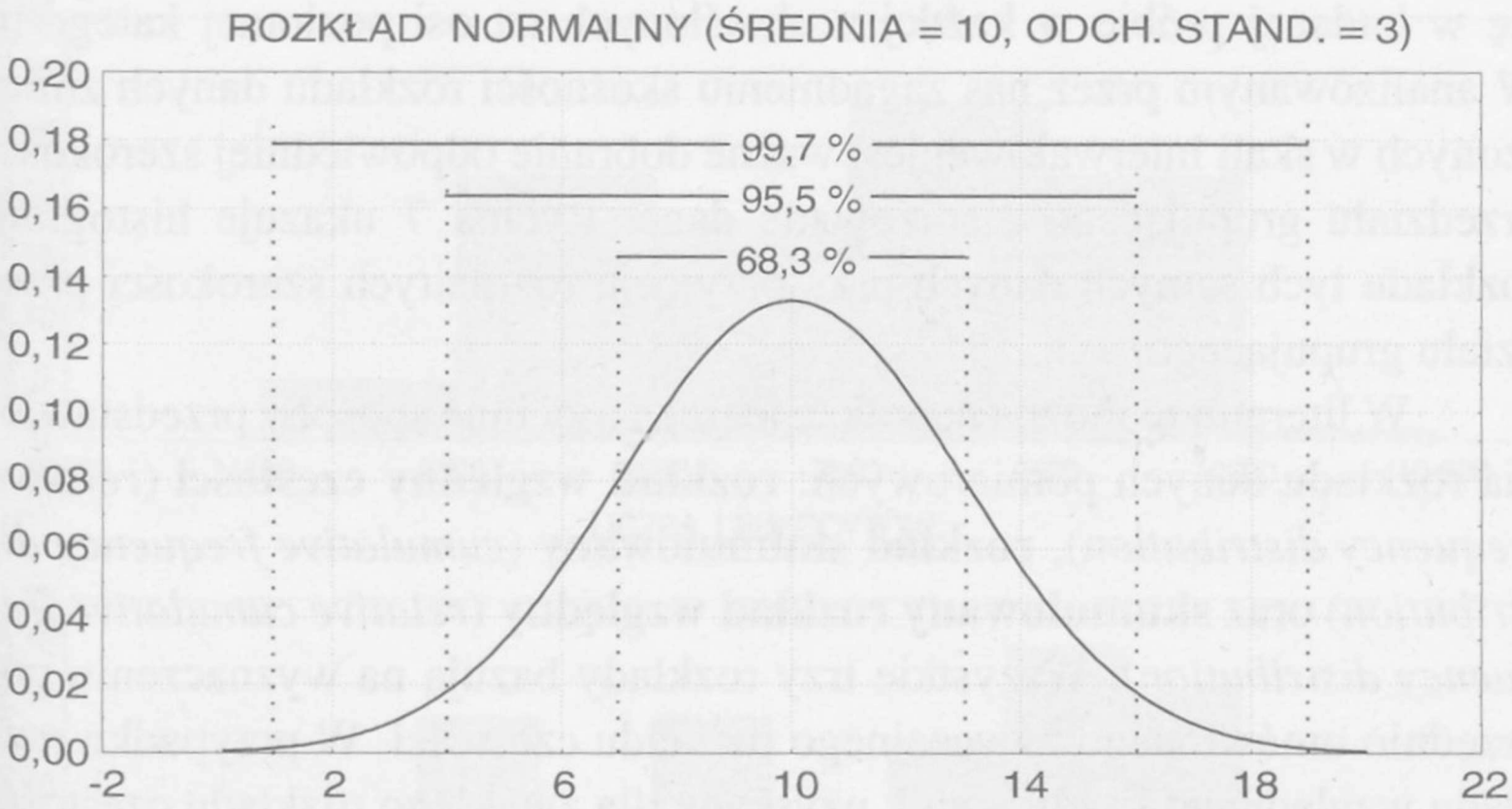
I tutaj nasuwają się dwie uwagi. Po pierwsze współczynnik zmienności nie zawsze jest dobrze zdefiniowany (jest nie określony, gdy średnia arytmetyczna równa się zero). Po drugie, odchylenia standardowe możemy porównywać wyłącznie wtedy, gdy są one wyrażone w takich samych jednostkach. Dla przykładu, jeżeli zmierzono wzrost w dwóch grupach lub większej ich liczbie, to wolno nam porównywać zarówno średnie arytmetyczne, jak i odchylenia standardowe, nie uciekając się do obliczenia współczynników zmienności. We wszystkich innych przypadkach trzeba odwoływać się jednak do współczynnika zmienności.

Wariancja

Kolejną miarą rozproszenia jest **wariancja** (*variance*). Oblicza się ją jako kwadrat odchylenia standardowego. Mówiąc ściślej, zawsze najpierw oszacowuje się wariancję, a z niej dopiero (poprzez pierwiastkowanie) oblicza się odchylenie standardowe – dla nas jednakże kolejność obliczeń nie będzie miała zasadniczego znaczenia, gdyż i tak obliczenia dokonuje komputer wyposażony w pakiet oprogramowania statystycznego. Łatwo się domyślić, że jednostką, w której wyrażona jest wariancja będzie jednostka, w której wyrażono wynik pomiaru podniesiona do kwadratu. Dla statystyków wariancja odgrywa kluczową rolę ze względu na jej „eleganckie" własności matematyczne, przeciętny praktyk jednak, korzystający ze statystyki pomocniczo do opisu swoich obserwacji i pomiarów, woli określać rozproszenia przez odchylenie standardowe – między innymi ze względu na prostotę interpretacji wyników (łatwo sobie wyobrazić cm^2, ale już nie tak prosto kg^2).

Eliminacja błędów grubych na podstawie odchylenia standardowego

Odchylenie standardowe ma jeszcze jedną ciekawą właściwość. Wspomnieliśmy już, że wiele różnych danych pochodzi z populacji o **rozkładzie normalnym** (*normal distribution*). Rozkład ten jest odwzorowany symetryczną krzywą w kształcie dzwonu, a jego średnia arytmetyczna odpowiada punktowi maksymalnego wypiętrzenia krzywej. Jeżeli skonstruujemy przedział (średnia – odchylenie standardowe, średnia + odchylenie standardowe), to znajdzie się w nim 68,3% wszystkich wyników pomiarów. Przedział dwukrotnie szerszy (średnia – dwa odchylenia standardowe, średnia + dwa odchylenia standardowe) zawiera 95,5% wszystkich pomiarów, zaś trzykrotnie – 99,7%. Innymi słowy, jeżeli dane pochodzą z populacji o rozkładzie normalnym, to mamy jedynie 0,3% szansy na napotkanie pomiaru, którego wynik różni się od wartości średniej więcej niż o trzy



Ryc. 6. Procentowy rozkład danych pochodzących z populacji o rozkładzie normalnym w przedziałach o szerokościach będących wielokrotnościami odchylenia standardowego

odchylenia standardowe (ryc. 6). Czasami stosuje się to jako kryterium do określania błędów grubych – pomiary których wyniki różnią się od wartości średniej więcej niż o trzy odchylenia standardowe traktuje się jako pomiary obarczone błędem grubym i eliminuje się z dalszej analizy.

Problem oceny skośności rozkładu

Stosowanie opisanego kryterium eliminacji błędów grubych jest uzasadnione jedynie w wypadku danych pochodzących z populacji o rozkładzie normalnym. W wypadku rozkładów skośnych możemy w ten sposób wyeliminować w pełni wartościowe wyniki pomiarów. Jak zatem ocenić, czy rozkład danych jest symetryczny, czy skośny?

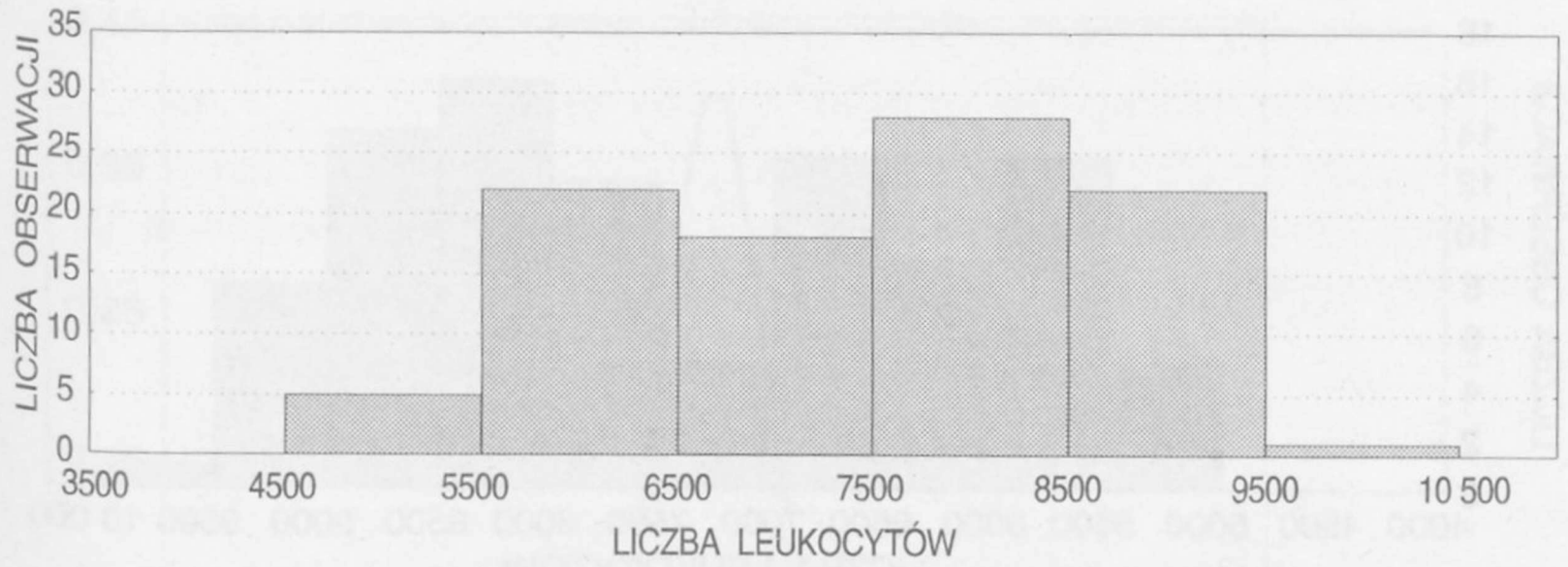
Istnieją precyzyjne mierniki statystyczne skośności zbudowane na tzw. trzecim momencie rozkładu danych. My jednak ograniczymy się tu do sposobu prostszego i w większości przypadków bardzo skutecznego – konstrukcji **tablicy rozkładu częstości** (*frequency distribution*). Tablica ta wskazuje, jak często w badanej próbie pojawił się wynik pomiaru określonej wartości. Można jej używać w dowolnej skali pomiarowej. Wyniki zawarte w tabeli rozkładu częstości prezentuje się graficznie najczęściej w postaci histogramu. Na osi poziomej zaznaczamy klasy kategorii (w przypadku skali nominalnej), uporządkowane klasy kategorii (w skali porządkowej) lub wartości liczbowe (mogą być również pogrupowane) w przypadku skali interwałowej. Oś pionowa przedstawia liczbę przypadków pojawiających

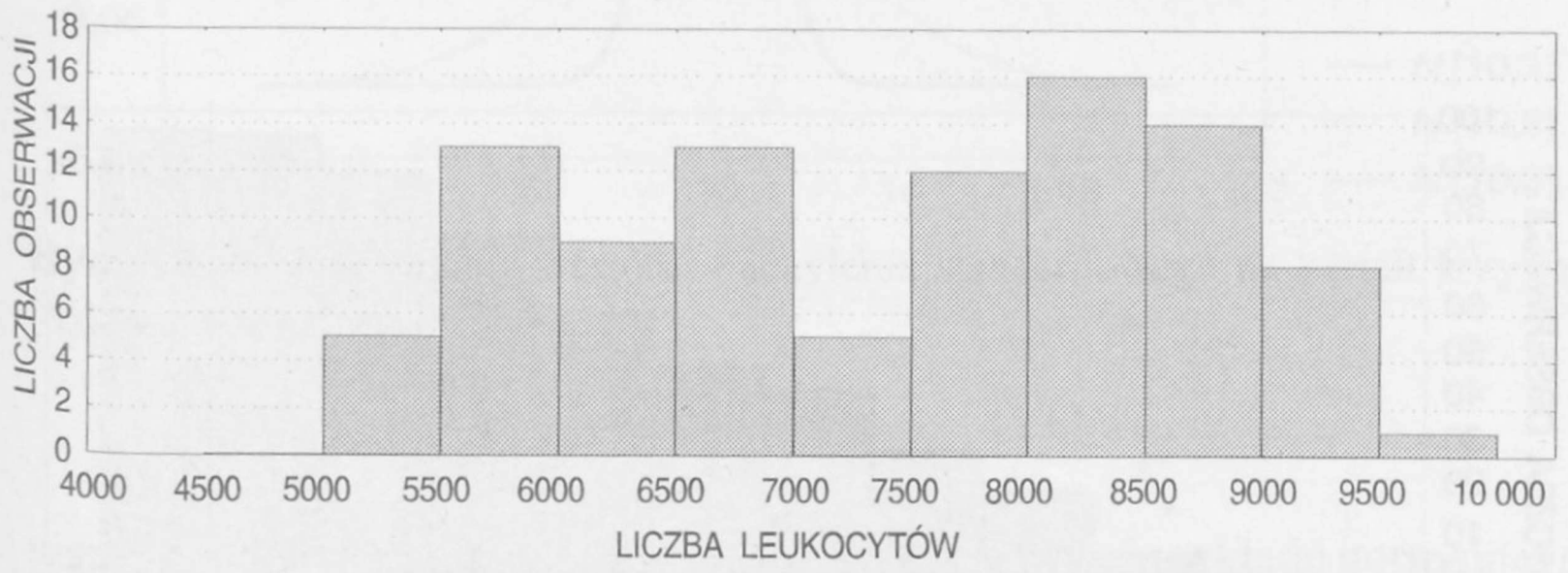
się w badanej próbie w każdej z określonych na osi poziomej kategorii. W analizowanym przez nas zagadnieniu skośności rozkładu danych zmierzonych w skali interwałowej jest ważne dobranie odpowiedniej szerokości przedziału grupującego analizowane dane. Rycina 7 ukazuje histogram rozkładu tych samych danych przy przyjęciu rozmaitych szerokości przedziału grupującego.

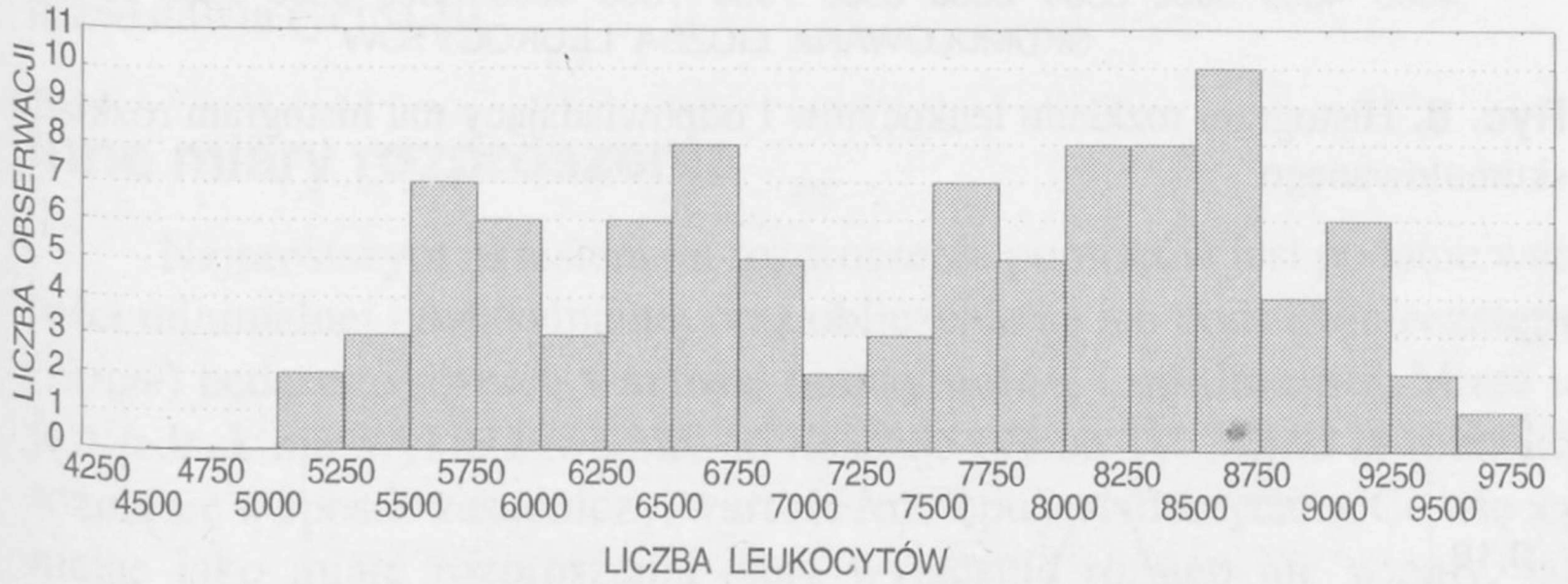
W literaturze można napotkać jeszcze trzy inne sposoby przedstawienia rozkładu danych pomiarowych: **rozkład względny częstości** (*relative frequency distribution*), **rozkład skumulowany** (*cumulative frequency distribution*) oraz **skumulowany rozkład względny** (*relative cumulative frequency distribution*). Wszystkie trzy rozkłady bazują na wyznaczeniu poprzednio omówionego zwyczajnego rozkładu częstości. W przypadku rozkładu względnego każdy wynik uzyskany dla zwykłego rozkładu częstości dzielimy przez łączną liczbę pomiarów w badanej próbie i czasami (jeżeli wynik chcemy wyrazić w procentach) mnożymy przez 100%. Względny rozkład częstości określa zatem, jaki odsetek całkowitej liczby pomiarów należy do danej kategorii.

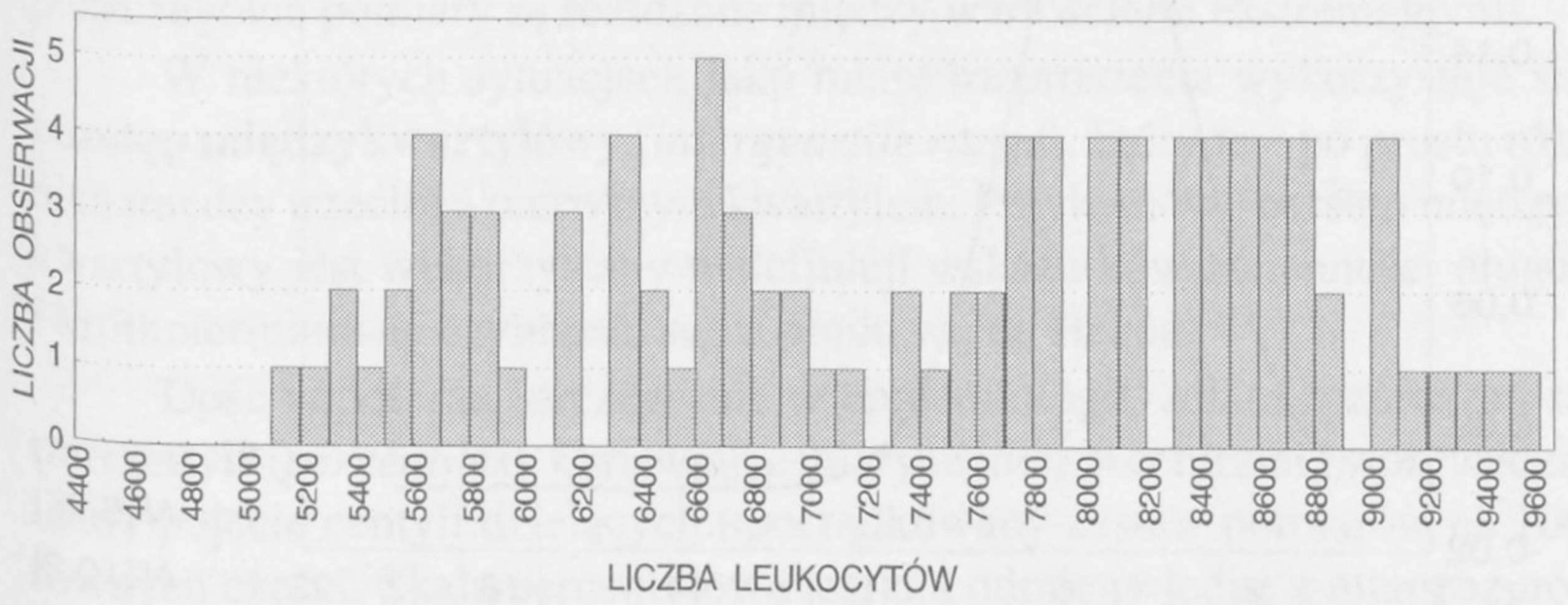
Skumulowany rozkład częstotliwości wskazuje łączną liczbę przypadków, które wpadają do określonego przedziału lub do wszystkich przedziałów leżących poniżej przedziału analizowanego. Przykładowo wartość skumulowanego rozkładu częstości dla wartości 3 wg skali Apgar oznacza liczbę przypadków z oceną stanu noworodka równą 0, 1, 2 lub 3. Z definicji wynika również, że rozkład skumulowany nie może być skonstruowany dla pomiarów w skali nominalnej (brak możliwości uporządkowania kategorii). Skumulowany rozkład względny uzyskujemy, dzieląc wartości rozkładu skumulowanego przez łączną liczbę przypadków w analizowanej próbie i, jeśli chcemy wyrazić rozkład w procentach, mnożąc uzyskaną wartość przez 100%. Porównanie histogramów rozkładu danych oraz rozkładu skumulowanego przedstawia rycina 8.

Średnia arytmetyczna i odchylenie standardowe określają w sposób jednoznaczny kształt krzywej rozkładu normalnego. Z tego powodu symboliczna notacja normalności rozkładu danych ma postać $N[\bar{x}, SD]$, gdzie $\bar{x}$ oznacza średnią arytmetyczną zaś SD – odchylenie standardowe. Wzrost wartości średniej powoduje przesunięcie dzwonu w prawo (w stronę większych wartości zmiennej niezależnej), jej spadek przesuwa krzywą w lewo (ryc. 9). Zwiększenie wartości odchylenia standardowego oznacza większe rozproszenie wyników i zgodnie z intuicją spłaszczenie krzywej rozkładu (fachowo określa się ją jako *platykurtyczną*). Gdy odchylenie standardowe maleje, krzywa rozkładu wysmukla się (krzywa *leptokurtyczna*). Wpływ

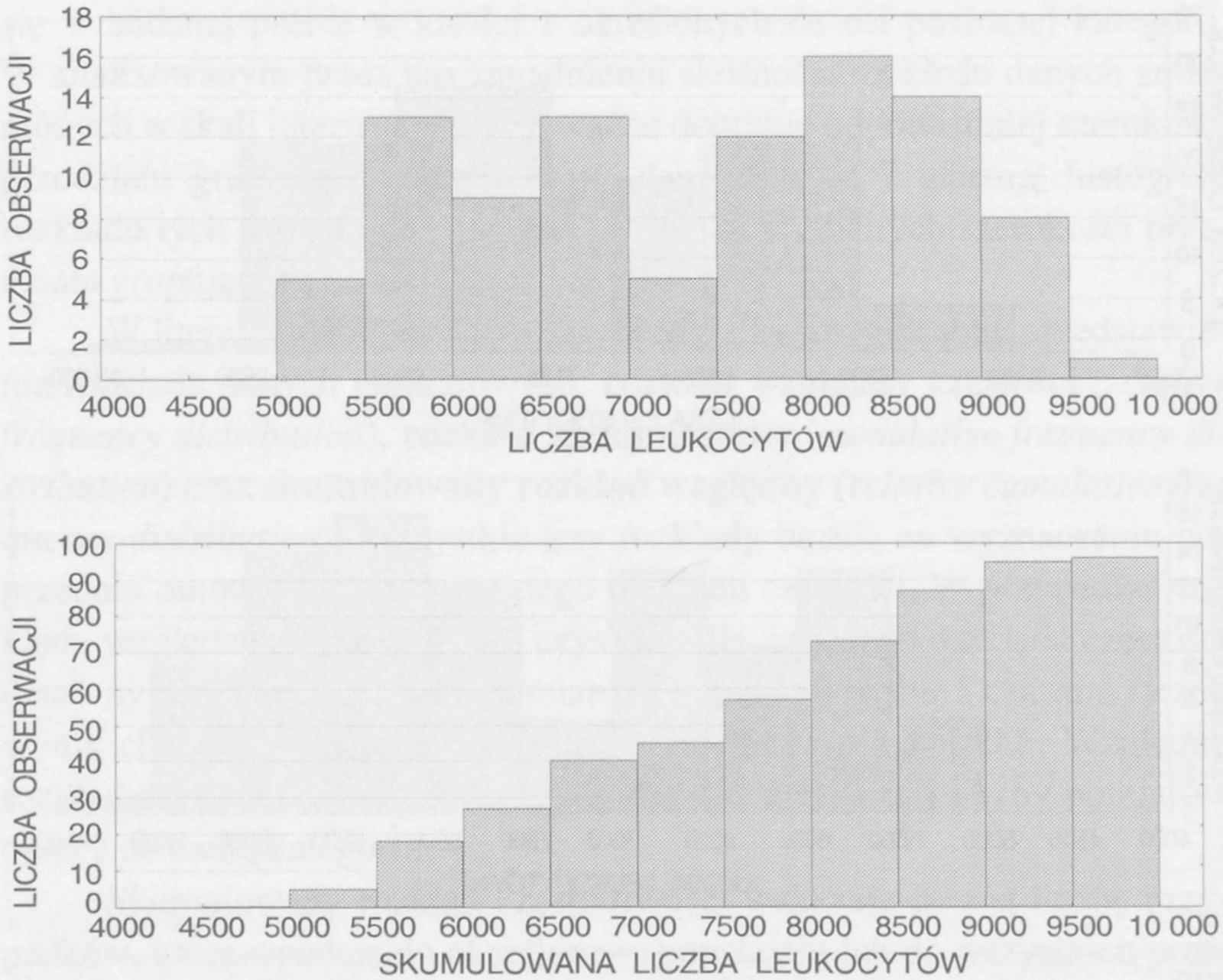




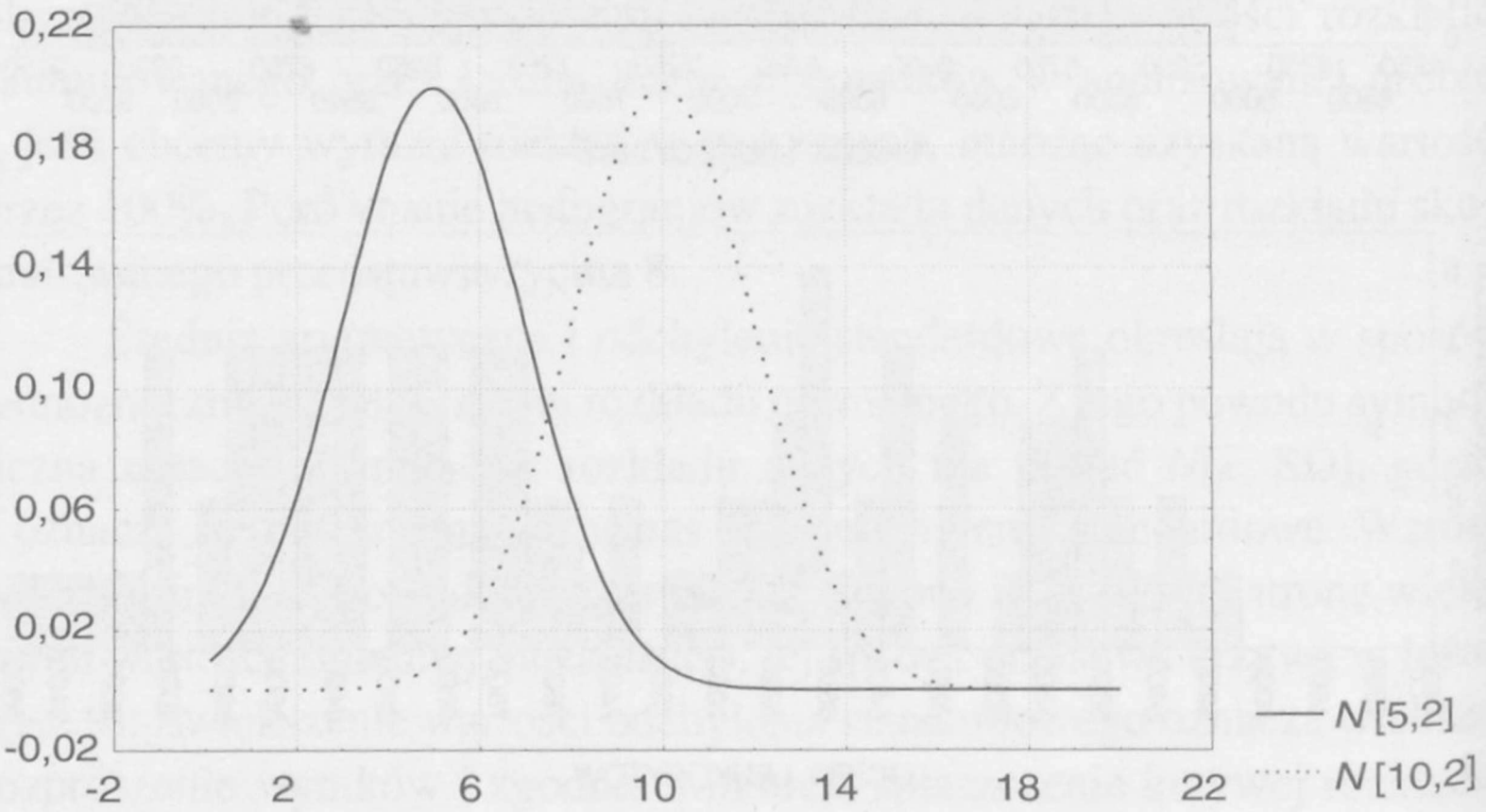




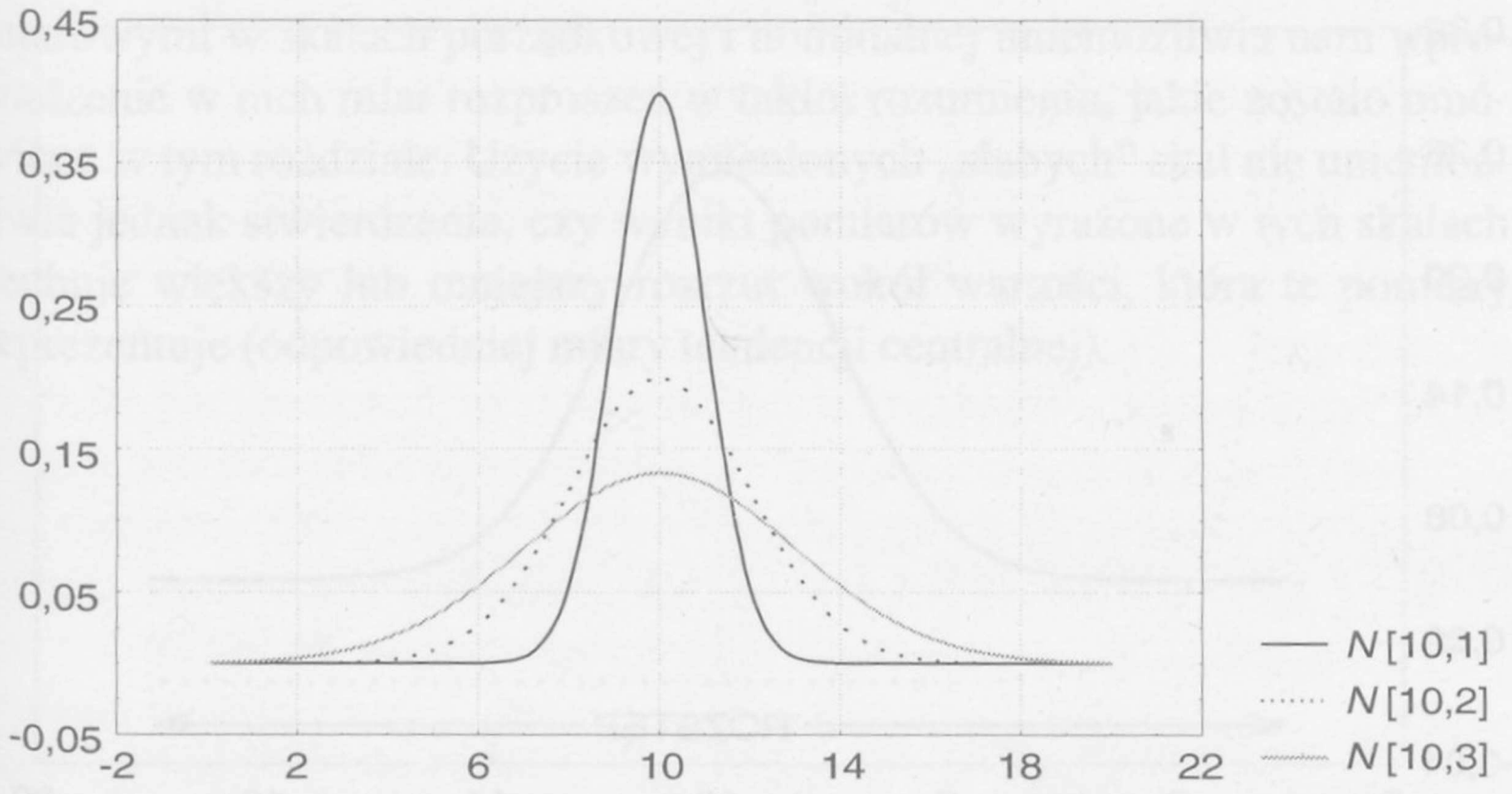
Ryc. 7. Zależność kształtu histogramu tego samego zestawu danych od przyjętej wartości szerokości przedziału grupującego wyniki



Ryc. 8. Histogram rozkładu leukocytów i odpowiadający mu histogram rozkładu skumulowanego



Ryc. 9. Wpływ zmiany wartości średniej arytmetycznej na położenie „dzwonu" krzywej rozkładu normalnego



Ryc. 10. Wpływ zmiany wartości odchylenia standardowego na kształt krzywej rozkładu normalnego

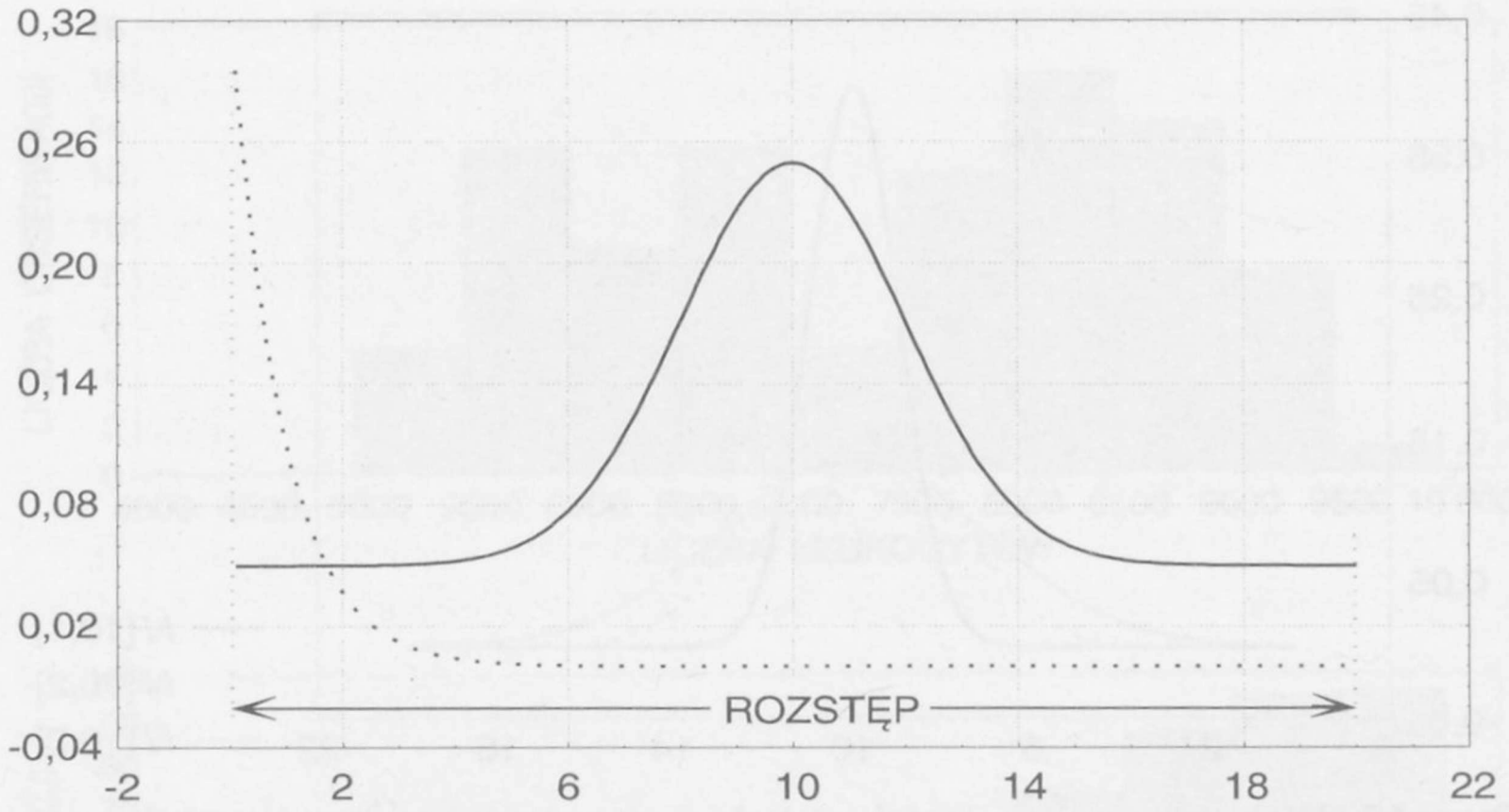
wartości odchylenia standardowego na kształt krzywej rozkładu normalnego przedstawia rycina 10.

Inne miary rozproszenia

Najprostszym określeniem rozproszenia pomiarów jest podanie wartości minimalnej i maksymalnej oraz obliczenie na ich podstawie **rozstępu** (*range*) będącego różnicą **wartości maksymalnej** i **minimalnej**. Miara ta jest jednak niezwykle zawodna, co ilustruje rycina 11. Mimo iż rozkłady różnią się w sposób zasadniczy, wartość rozstępu jest identyczna. Co więcej, mając jako miarę rozproszenia dany wyłącznie rozstęp nie wiemy, jak poszczególne pomiary są rozłożone między wartościami ekstremalnymi.

W niektórych sytuacjach jako miarę rozproszenia wykorzystuje się **rozstęp międzykwartylowy** (*interquartile range*), który jest po prostu różnicą między trzecim a pierwszym kwartylem. Przykładowo rozstęp międzykwartylowy jest wykorzystany w definicji wskaźników zmienności długo- i krótkoterminowej czynności serca płodu wg de Haana.

Dość popularną (szczególnie w epidemiologii) miarą rozproszeń są **percentyle** (*percentiles*). Omawiając miary tendencji centralnej wprowadziliśmy pojęcie centyli dzielących uporządkowany zestaw pomiarów na 100 równych części. Skala percentylowa jest dla odmiany jedną z miar rozproszenia i określa procent wszystkich obserwacji, których wartość jest mniejsza lub równa wartości określonego centyla.



Ryc. 11. Miara rozstępu w przypadku krańcowo różnych rozkładów danych może być identyczna

Przykład 8

W grupie 20 mężczyzn oznaczono poziom kreatyniny w moczu (wyrażony w mg). Uporządkowane rosnąco wyniki przedstawiają się następująco: 12, 17, 25, 28, 29, 29, 30, 30, 30, 31, 32, 32, 32, 32, 33, 33, 34, 34, 35, 36.

Wyznacz wartość piątego, trzydziestego i osiemdziesiątego percentyla.

Ponieważ mamy dwadzieścia uporządkowanych pomiarów, skok o jeden pomiar jest równy 5 procentom na skali percentylowej. Zatem centyl piąty odpowiada pomiarowi pierwszemu (12 mg), trzydziesty – pomiarowi szóstemu (29 mg), a osiemdziesiąty – pomiarowi szesnastemu (33 mg). Pięć procent pomiarów jest więc mniejszych lub równych 12 mg, trzydzieści procent – 29 mg, a osiemdziesiąt procent – 33 mg.

Pamiętajmy jednak, aby zawsze przed wyznaczeniem wartości centyli uporządkować dane pomiarowe w poprządku niemalejącym.

Opisane przez nas wielkości nie wyczerpują wszystkich możliwych miar rozproszeń (np. nie omówiliśmy tu własności odchylenia przeciętnego), stanowią jednakże reprezentatywny przegląd miar najczęściej wykorzystywanych w naukach medycznych.

Na zakończenie zwróćmy uwagę na następujący, nadzwyczaj istotny fakt. Wszystkie bez wyjątku miary rozproszeń są zdefiniowane wyłącznie dla skali interwałowej. Brak informacji o odległości między punktami po-

miarowymi w skalach porządkowej i nominalnej uniemożliwia nam wprowadzenie w nich miar rozproszeń w takim rozumieniu, jakie zostało omówione w tym rozdziale. Użycie wymienionych „słabych" skal nie uniemożliwia jednak stwierdzenia, czy wyniki pomiarów wyrażone w tych skalach cechuje większy lub mniejszy rozrzut wokół wartości, która te pomiary reprezentuje (odpowiedniej miary tendencji centralnej).

Testowanie hipotez

Błędy pomiarowe i ich pochodzenie

W poprzednim rozdziale, omawiając najczęściej stosowane miary statystyki opisowej, wprowadziliśmy punktowe wielkości charakteryzujące tendencję centralną i rozproszenie pomiarów. Założyliśmy przy tym w sposób niejawny, że każdy pomiar dokonywany jest nieskończenie dokładnie. W praktyce jest to oczywiście nierealne, w związku z czym wprowadzone przez nas miary punktowe są pewnego rodzaju idealizacją rzeczywistości. Obecnie spróbujemy do tematu błędów pomiarowych powrócić.

Każdy bez wyjątku pomiar jest obarczony błędem. Znana dobrze fizykom (i nie tylko) zasada nieoznaczoności Heisenberga stwierdza, że każda czynność pomiarowa zakłóca stan obiektu mierzonego, przez co sam pomiar staje się niedokładny.

Innymi słowy wynik pomiaru x jest jedynie lepszym lub gorszym przybliżeniem wartości rzeczywistej x_0. Różnica między tymi dwiema wielkościami zwana jest **błędem bezwzględnym** (*absolute error*), natomiast jej stosunek do wartości pomiaru jest określany jako **błąd względny** (*relative error*).

Niepewność pomiarową można zminimalizować przez przyjęcie odpowiedniej metody pomiaru, ale nigdy nie można jej wyeliminować całkowicie. Błędy pomiarowe dzieli się na trzy podstawowe klasy:

– **błędy systematyczne** (*systematic errors*) – ich źródłem jest zwykle przyrząd pomiarowy, obserwator lub przyjęta metoda pomiaru. Przykładowo błąd systematyczny może być związany z dokładnością przyrządu (którą określa najmniejsza działka na skali przyrządu). W takim przypadku można go zmniejszyć stosując bardziej dokładny przyrząd (np. pomiar długości linijką ze skalą milimetrową daje dokładność pomiaru 1 mm, noniuszem – 0,1 mm, śrubą mikrometryczną – 0,01 mm). Błąd systematyczny może wynikać również z tego, że obserwator stale źle wykonuje pomiar (na przykład odczytuje wskazanie przyrządu pomiarowego patrząc pod złym kątem, w wyniku czego pojawia się błąd paralaksy). W takim przypadku błąd

można zmniejszyć szkoląc operatora. I wreszcie źródłem błędu systematycznego może być metoda pomiarowa. Wtedy zmniejszenie błędu osiąga się przez ponowne przeanalizowanie stosowanej metody pomiaru i skorygowanie nieprawidłowo wykonywanych czynności.

– **błędy grube** (*biased errors*) – ich źródłem może być pomyłka przy odczycie wyniku pomiaru, nieprawidłowo wykalibrowany przyrząd, zła metoda pomiarowa. Charakterystyczną cechą błędów grubych jest to, że wynik pomiaru obarczony błędem grubym tak bardzo odbiega od wartości prawidłowej (lub chociażby tylko prawdopodobnej), że stosunkowo łatwo jest się domyślić, że jest błędny.

– **błędy losowe** (*random errors*) – związane z odstępstwem od założonego modelu (wielokrotny pomiar średnicy tej samej kulki może dawać rozmaite wyniki, gdyż kulka rzeczywista nie jest idealną kulą), z wpływem zmysłów eksperymentatora, z dokonywaniem pomiarów nie na jednym obiekcie, lecz na pewnym zbiorze obiektów, dla którego jest zdefiniowana wielkość mierzona. W badaniach biologicznych będziemy mieli do czynienia szczególnie często z tą ostatnią przyczyną, określaną zwykle jako zmienność osobnicza. Błędów losowych nie potrafimy wyeliminować, możemy jednak oszacować ich wielkość i wpływ na wiarogodność naszych pomiarów.

Omawiając własności średniej arytmetycznej, stwierdziliśmy, że wartość ta jest szczególnie czuła na skrajne wartości wyników pomiarów. Nie oznacza to jednak, że średnia nie jest czuła na inne wartości wyników. Zmiana wyniku jakiegokolwiek pomiaru powoduje większą lub mniejszą zmianę średniej arytmetycznej. Jeżeli zatem każdy z nich jest nam znany jedynie w lepszym lub gorszym przybliżeniu, to również średnia arytmetyczna oszacowana na ich podstawie nie może być oszacowana jako punkt, lecz jako pewien przedział, w którym ten punkt może się znaleźć z większym lub mniejszym prawdopodobieństwem. Naszym zadaniem jest znalezienie tego przedziału.

Widzimy zatem, że w przeciwieństwie do statystyki opisowej, w której średnią arytmetyczną traktowaliśmy deterministycznie jako punkt, w teorii testowania hipotez mamy do czynienia z określonym probabilistycznie obszarem, w którym ta średnia może się znaleźć z określonym prawdopodobieństwem.

Może się nam to podobać lub nie, ale tak naprawdę nie możemy odpowiedzieć w sposób jednoznaczny na pytanie, czy na przykład dwie średnie różnią się od siebie. Jesteśmy natomiast w stanie stwierdzić, czy omawiane średnie różnią się od siebie z określonym prawdopodobieństwem. Badanie istotności różnic może oczywiście dotyczyć nie tylko średnich

arytmetycznych, ale i wariancji, rozkładów pomiarów, współczynników korelacji itp. Za każdym razem gdy stwierdzamy, że jakieś wyznaczone na podstawie próby statystycznej parametry różnią się w dwóch analizowanych populacjach (lub większej ich liczbie), pojawia się pytanie, czy jest to wyraz jakiejś stałej tendencji, czy wynik przypadkowego zbiegu okoliczności. Odpowiedź ma z reguły znaczenie fundamentalne.

W tym momencie Czytelnik mógłby odnieść wrażenie, że niepewność pomiarowa jest związana tylko z silną skalą interwałową, nie dotyczy natomiast skali porządkowej i nominalnej. Niestety, rzeczywistość jest bardziej złożona – błędy pojawiają się przy stosowaniu każdej skali, a niektóre skale są na nie po prostu bardziej lub mniej podatne. Wynika to z faktu, że w każdej ze znanych skal pomiarowych jest możliwe subiektywne (związane ze zmiennością osobniczą) omyłkowe ustalenie (lub błędne zapisanie – to także się zdarza!) wyniku dokonanego pomiaru.

Jakby nie dość było wymienionych kłopotów, musimy sobie ponadto uświadomić, że prowadząc dowolne pomiary i obserwacje popełniamy niestety jeszcze jeden rodzaj błędu – **błąd próbkowania** (*sampling error*). W rozdziale omawiającym pojęcie próby i populacji stwierdziliśmy, że rzadko się zdarza, by eksperymentator miał możność zebrania danych w obrębie całej populacji (badanie wyczerpujące). W związku z powyższym pobiera on z populacji pewne podzbiory zwane próbami. Rozkład częstości danych w każdej pobranej próbie różni się mniej lub bardziej od rzeczywistego rozkładu w populacji. Zjawisko to często jest określane jako **zmienność próbkowania** (*sampling variation*). Im mniejsza jest pobierana próba, tym większe jest prawdopodobieństwo, że rozkłady w próbie i populacji będą się różniły.

Możemy to sprawdzić za pomocą bardzo prostego eksperymentu. Weźmy nieuszkodzoną monetę i dokonajmy wielu serii rzutów po 100 rzutów w każdej serii. Spodziewamy się, że liczba otrzymanych w każdej serii wyrzuconych orłów i reszek będzie za każdym razem identyczna i równa 50. Wynika to z faktu, że prawdopodobieństwo wyrzucenia orła lub reszki jest takie samo i wynosi 0,5. W tablicy 1 zestawiono przykładowe wyniki z 10 serii rzutów.

Jeślibyśmy taki eksperyment przeprowadzili wielokrotnie (powiedzmy sto tysięcy razy) i spróbowali wykreślić histogram rozkładu częstotliwości otrzymania określonej liczby orłów (lub, co jest całkowicie równoważne, określonej liczby reszek), to otrzymalibyśmy wykres przypominający rozkład normalny. Pik histogramu przypadałby dla liczby 50, im bardziej natomiast oddalalibyśmy się od tej wartości, przypadków byłoby coraz mniej. Histogram przypominałby krzywą rozkładu normalnego z wartością

Tablica 1. Wyniki przykładowego eksperymentu wielokrotnej serii rzutów monetą w seriach po sto rzutów

Nr serii rzutu	Liczba orłów	Liczba reszek
1	55	45
2	48	52
3	42	58
4	49	51
5	57	43
6	54	46
7	55	45
8	48	52
9	48	52
10	50	50

średnią wynoszącą 50 i z pewną niezerową wariancją.* To, że wbrew naszym przewidywaniom nie zawsze w serii rzutów otrzymalibyśmy 50 orłów jest właśnie wynikiem błędu próbkowania. Pamiętajmy, że nawet najmniej prawdopodobny wariant (otrzymanie stu orłów lub nieotrzymanie orłów) ma szansę pojawić się w naszym eksperymencie. Niemożliwą natomiast rzeczą jest uzyskanie w serii 100 rzutów 101 lub większej liczby orłów. Widzimy zatem, że także pobranie próby z badanej populacji jest obarczone błędem.

Zadaniem działu statystyki zwanego **testowaniem hipotez** (*hypothesis testing*) będzie próba odpowiedzi na następujące pytanie: czy dwie próby (lub większa ich liczba) pochodzą z tej samej czy też z różnych populacji?

Nasze rozważania rozpoczniemy od porównywania parametrów statystyki opisowej wyznaczonych dla danych zmierzonych w skali interwałowej.

---

* Autorzy chcą podkreślić użycie zwrotu „histogram przypominałby krzywą rozkładu normalnego ...", gdyż w rzeczywistości z uwagi na na dyskretyzację pomiaru (orzeł lub reszka) rozkład prawdopodobieństwa jest opisany rozkładem dwumianowym. Rozkład normalny jest natomiast jednym z możliwych rozkładów opisujących zmienną ciągłą.

Przedziały ufności

Omawiając pewne właściwości rozkładu normalnego i odchylenia standardowego jako miary rozproszenia zauważyliśmy, że w przedziałach $\bar{x} \pm SD$, $\bar{x} \pm 2 \times SD$ oraz $\bar{x} \pm 3 \times SD$ mieści się odpowiednio 68,3; 95,5 oraz 99,7% wszystkich pomiarów, jeśli pochodziły one z populacji o rozkładzie normalnym. Możemy teraz spróbować sformułować pytanie odwrotne: Jak szeroki powinien być przedział, aby zawierał na przykład 95,0; 99,0 lub 99,9% pomiarów? Inaczej – przez jaką liczbę należy przemnożyć wartość odchylenia standardowego, by poza tak skonstruowanym przedziałem znalazło się odpowiednio nie więcej niż 5,0; 1,0 lub 0,1% pomiarów?

Liczby te można znaleźć w tablicach statystycznych, lecz „zna" je również zdecydowana większość współczesnych pakietów statystycznych. Dla zaspokojenia ciekawości Czytelnika podamy jedynie, że przedział o krańcach: $\bar{x} - 1{,}96 \times SD, \bar{x} + 1{,}96 \times SD$ pokrywa 95% pomiarów pobranych z populacji o rozkładzie normalnym. Jeśli mnożnik 1,96 zastąpimy przez 2,58, to rozszerzony przedział pokryje aż 99% pomiarów, wreszcie przy zastosowaniu mnożnika 3,29 uzyskamy przedział pokrywający 99,9% wszystkich pomiarów.

Do tej pory interesowało nas prawdopodobieństwo znalezienia określonego odsetka danych pomiarowych. W poprzednim rozdziale stwierdziliśmy, że z uwagi na błąd każdego z pomiarów wyznaczona na ich podstawie średnia arytmetyczna nie jest wartością punktową. Spróbujmy teraz oszacować prawdopodobieństwo, że leży ona w pewnym określonym przedziale. Skorzystamy tu z prostego twierdzenia, które mówi, że jeżeli uzyskane dane mają rozkład normalny o wartości średniej $\bar{x}$ i odchyleniu standardowym SD (symboliczny zapis $N[\bar{x}, SD]$), to skonstruowana na ich podstawie średnia arytmetyczna ma również rozkład normalny o parametrach $N[\bar{x}, SD/\sqrt{M}]$, gdzie M jest liczbą danych pomiarowych. Przypomnijmy, że wartość $SD/\sqrt{M}$ to nic innego, jak błąd standardowy średniej arytmetycznej SEM. Zbudujmy teraz przedziały o krańcach: średnia – mnożnik × SEM, średnia + mnożnik × SEM, gdzie mnożnik przyjmuje jedną z wartości: 1,96; 2,58 lub 3,29. W tych przedziałach odpowiednio z prawdopodobieństwem 95%, 99% lub 99,9% znajduje się prawdziwa wartość średniej arytmetycznej dla populacji (użyta do konstrukcji przedziałów średnia jest średnią arytmetyczną pomiarów z badanej próby a nie z całej populacji). Szansa, że prawdziwa wartość średniej populacyjnej leży poza tymi przedziałami wynosi odpowiednio 5,1 lub 0,1%.

Tak skonstruowane przedziały nazywamy **przedziałami ufności** (*confidence intervals*), mnożniki 1,96; 2,58 i 3,29 natomiast wartościami

krytycznymi rozkładu normalnego (*critical values of the normal distribution*) na **poziomie ufności** (lub **poziomie istotności** – *confidence level, p-value*) odpowiednio: 0,05 (5%), 0,01 (1%) oraz 0,001 (0,1%). W dalszej części książki będziemy wymiennie używać terminów poziom istotności, poziom istotności, oraz *p-value*.

Zakładaliśmy do tej pory, że wyniki naszych pomiarów pochodzą z populacji o rozkładzie normalnym. Wiemy jednak, że prawidłowe przybliżenie rozkładu normalnego osiąga się jedynie przy bardzo dużej liczbie pomiarów (prawo wielkich liczb). W badaniach medycznych natomiast spotykamy często małe liczebności grup, ponieważ przeprowadzenie dużej liczby obserwacji jest zwykle bardzo trudne i kosztowne, a niekiedy wręcz niemożliwe (rzadko występujące choroby).

Czy skonstruowane poprzednio przedziały będą rzeczywiście pokrywały żądany procent pomiarów? Odpowiedź niestety jest negatywna. Im mniejsza jest liczba pomiarów, tym szerszy musi być przedział pokrywający zadany ich procent. Większe muszą zatem też być mnożniki, przez które należy przemnożyć odchylenie standardowe, by w uzyskanym przedziale „zamknąć" potrzebny odsetek pomiarów. Ich wartości będą zależne od dwóch parametrów: tak zwanej **liczby stopni swobody** (*degrees of freedom*) którą w najprostszych przypadkach otrzymuje się, odejmując jedynkę od liczby analizowanych pomiarów, oraz od poziomu ufności. Poszukiwane przez nas wartości mnożników wyznaczających przedziały zmienności dla średnich arytmetycznych można znaleźć w tablicach statystycznych wartości krytycznych **rozkładu t-Studenta** (*t-Student distribution*). Wartości te – podobnie jak wartości krytyczne rozkładu Gaussa (czyli rozkładu normalnego) – są generowane automatycznie w komputerowych pakietach statystycznych. Nie wchodząc w szczegóły matematyczne, rozkład t-Studenta jest uogólnionym rozkładem normalnym. Modyfikacja polega na wzięciu pod uwagę innej niż nieskończona liczby pomiarów. Rozkład normalny jest zatem szczególnym przypadkiem rozkładu t-Studenta – gdy liczba stopni swobody jest nieskończona. Rozkład t-Studenta tym bardziej odbiega od rozkładu normalnego, im mniejsza liczba stopni swobody jest brana pod uwagę przy ustalaniu rozkładu. Pozostałe własności rozkładu t-Studenta, a więc symetria, zależność od średniej arytmetycznej i odchylenia standardowego są identyczne, jak dla rozkładu Gaussa.

Przedział ufności konstruujemy, dodając (odejmując) do (od) średniej arytmetycznej następujące wyrażenie: odchylenie standardowe pomnożone przez odpowiednią wartość krytyczną rozkładu t-Studenta i podzielone przez pierwiastek kwadratowy z liczby pomiarów.

Teraz, gdy wiemy już, co to są przedziały ufności dla średniej arytmetycznej, jak je konstruować i interpretować, zastanówmy się nad ich przydatnością. Przedziały ufności nadają się świetnie do intuicyjnego zrozumienia idei testowania hipotez. Spróbujemy to wyjaśnić na prostym przykładzie.

Przykład 9

W dwóch grupach szczurów badano przyrost masy ciała między 30 a 90 dniem życia. Dieta pierwszej grupy, w której było 10 szczurów charakteryzowała się wysoką zawartością białka, podczas gdy dieta grupy drugiej (8 szczurów) była niskobiałkowa. Czy rodzaj diety ma wpływ na przyrost masy ciała badanych szczurów?

Przyrost masy w grupie z dietą wysokobiałkową (wyrażony w gramach): 106,0; 111,0; 146,0; 102,0; 97,0; 125,0; 130,0; 109,0; 115,0; 99,0.

Przyrost masy w grupie z dietą niskobiałkową (wyrażony w gramach): 108,0; 78,0; 85,0; 101,0; 102,0; 82,0; 90,0; 99,0.

Dla grupy pierwszej średnia arytmetyczna wynosi 114,0 g, a odchylenie standardowe 15,5 g, dla grupy drugiej odpowiednio: 93,1 g i 10,9 g.

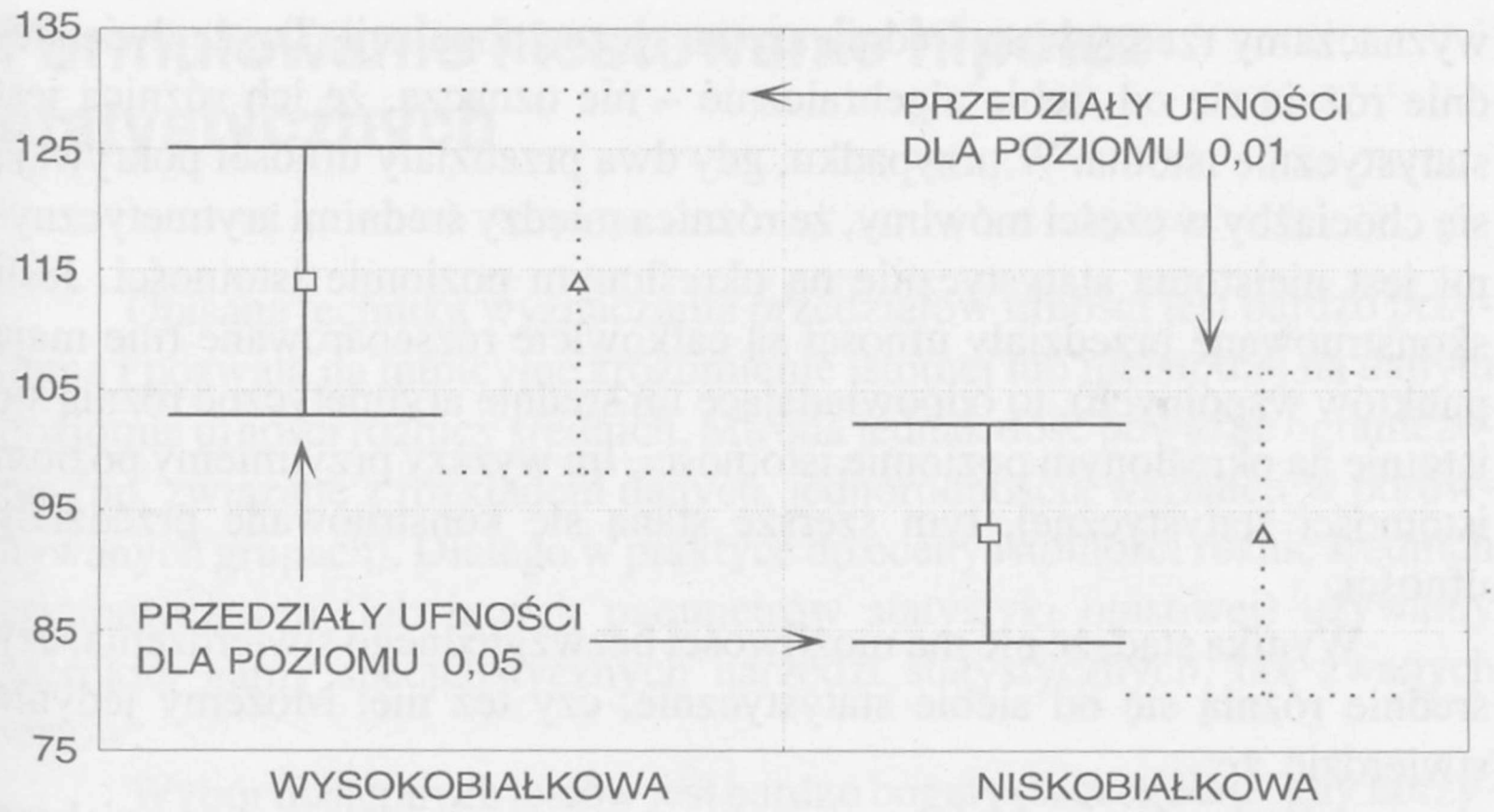
Skonstruujmy przedziały ufności dla średnich w obu grupach. Dla poziomu ufności 0,05 wartość krytyczna rozkładu t-Studenta dla grupy pierwszej (9 stopni swobody) wynosi 2,3; dla drugiej (7 stopni swobody) – 2,4.

Dla diety wysokobiałkowej na poziomie 0,05 dolny kraniec przedziału ufności wynosi: $114{,}0 - 2{,}3 \times 15{,}5/\sqrt{10} = 102{,}7$; górny zaś $114+2{,}3 \times 15{,}5/\sqrt{10} = 125{,}3$.

Dla diety ubogiej w białko na poziomie 0,05 dolny kraniec przedziału ufności wynosi: $93{,}1 - 2{,}4 \times 10{,}9/\sqrt{8} = 83{,}9$, górny zaś $93{,}1 + 2{,}4 \times 10{,}9/\sqrt{8} = 102{,}3$

Jak łatwo zauważyć, oba przedziały nie mają elementów wspólnych – są dobrze rozseparowane. Możemy zatem z prawdopodobieństwem 95% stwierdzić, że dieta wysokobiałkowa powodowała istotny przyrost masy szczura w porównaniu z dietą ubogą w białko. Proponujemy teraz Czytelnikowi dokonanie analogicznych obliczeń dla poziomu ufności 0,01 (mnożniki odpowiednio przyjmą wartości 3,2 oraz 3,5). Tym razem przedziały ufności będą częściowo na siebie zachodziły, w związku z czym na tym poziomie ufności nie możemy stwierdzić, czy algebraiczna różnica średnich mas w obu grupach była spowodowana rodzajem diety, a nie błędem próbkowania. Ilustracja graficzna tego zjawiska przedstawiona jest na rycinie 12.

Proszę również zwrócić uwagę na pewien drobiazg. Dokonanie analogicznych obliczeń za pomocą komputera doprowadzi nas do nieco innych wartości liczbowych (np. dla grupy wysokobiałkowej dolny kraniec prze-



Ryc. 12. Przedziały ufności na poziomie 0,05 i 0,01 dla zestawu danych z przykładu 9

działu ufności wyniesie 102,9229 zamiast obliczonego przez nas 102,7, górny zaś odpowiednio 125,0771 zamiast 125,3). Różnice te wynikają z przyjętego przez nas w obliczeniach „ręcznych" zaokrąglenia danych i wyników do jednego miejsca po przecinku dziesiętnym (np. przyjęliśmy wartość krytyczną rozkładu t-Studenta jako równą 2,3 zamiast wartości dokładniejszej 2,262156).

Podsumowując nasze rozważania stwierdzamy, że szerokość przedziału ufności zależy od trzech parametrów:

– poziomu ufności (zwanego również poziomem istotności statystycznej) – im większą chcemy mieć pewność, że w przedziale ufności mieści się rzeczywista średnia arytmetyczna, tym przedział będzie szerszy przy tych samych wartościach pozostałych dwóch parametrów;

– liczby analizowanych danych w grupie – grupy bardziej liczne powodują zawężenie przedziału ufności przy tych samych wartościach pozostałych dwóch parametrów;

– stopnia rozproszenia danych wyrażonego przez odchylenie standardowe – im większy jest rozrzut wyników pomiarów, tym szerszy jest przedział ufności przy tych samych wartościach pozostałych dwóch parametrów.

Pamiętajmy, że przedział ufności nie kwantyfikuje rozproszenia danych (czynią to miary rozproszenia), lecz oszacowuje niepewność z jaką

wyznaczamy rzeczywistą średnią arytmetyczną populacji. To, że dwie średnie różnią się od siebie algebraicznie – nie oznacza, że ich różnica jest statystycznie istotna. W przypadku, gdy dwa przedziały ufności pokrywają się chociażby w części mówimy, że różnica między średnimi arytmetycznymi jest nieistotna statystycznie na określonym poziomie istotności. Jeśli skonstruowane przedziały ufności są całkowicie rozseparowane (nie mają punktów wspólnych), to odpowiadające im średnie arytmetyczne różnią się istotnie na określonym poziomie istotności. Im wyższy przyjmiemy poziom istotności statystycznej, tym szersze staną się konstruowane przedziały ufności.

Wynika stąd, że nie ma możliwości bezwzględnego stwierdzenia, czy średnie różnią się od siebie statystycznie, czy też nie. Możemy jedynie stwierdzić, że:

– im bardziej oddalone są w rzeczywistości średnie, tym większą mamy pewność, że różnią się one między sobą istotnie;

– dwie średnie różniące się istotnie statystycznie na przykład na poziomie 0,05 mogą się nie różnić statystycznie na poziomie wyższym (np. 0,01 lub 0,001) – co zilustrowano na rycinie 12.

Analogiczne przedziały ufności jak dla średnich arytmetycznych można konstruować dla różnicy średnich, proporcji, różnicy lub stosunku dwóch proporcji, wariancji itp., aczkolwiek z oczywistych względów korzystamy wtedy z całkiem innych wzorów matematycznych.

Formułowanie i testowanie hipotez statystycznych

Opisana technika wyznaczania przedziałów ufności jest bardzo przydatna i pozwala na intuicyjne zrozumienie istotnej lub nieistotnej na danym poziomie ufności różnicy średnich. Ma ona jednak dość poważne ograniczenia (np. związane z rozkładem danych, jednorodnością wariancji w porównywanych grupach). Dlatego w praktyce do oceny istotności różnic średnich arytmetycznych (lub innych parametrów statystyki opisowej) używamy szerokiej gamy specjalistycznych narzędzi statystycznych, tak zwanych testów.

Wybór dostępnych testów jest bardzo bogaty (szczególnie, gdy korzysta się z gotowych programów komputerowych służących do statystycznego opracowywania danych, których twórcy dosłownie prześcigają się w liczbie i różnorodności testów oferowanych użytkownikowi oraz rozbudowanej grafice prezentacyjnej). Jednak to pozornie korzystne bogactwo technik i metod analizy danych w istocie sprawia kłopot (szczególnie początkującym badaczom), ponieważ muszą dokonać wyboru najwłaściwszej techniki w stosunku do posiadanych danych i rozważanego problemu naukowego – a to może się okazać bardzo trudne. To czy test jest w danej sytuacji optymalny, zależy od wielu czynników, takich jak wykorzystana skala pomiarowa, schemat prowadzenia badania, liczba porównywanych grup, normalność rozkładu danych (lub jej brak). Prawidłowy wybór testu jest jednak warunkiem koniecznym uzyskania wiarygodnych wyników – żaden program komputerowy nie jest w stanie podjąć decyzji za badacza.

Poruszony tu problem ma szerszy wymiar, musi więc być przedyskutowany nieco dokładniej. Generalnie nauczenie się posługiwania dowolnym pakietem oprogramowania statystycznego jest stosunkowo proste. Szczególnie „przyjazne" dla potencjalnego użytkownika są aplikacje korzystające z platformy WINDOWS – opanowanie sposobu przygotowania danych, wyboru odpowiednich opcji i generacji grafiki może zająć przeciętnie uzdolnionej osobie od kilku do kilkunastu godzin. O wiele bardziej złożony jest jednak problem doboru odpowiedniej do rozwiązywanego zadania metodyki oraz interpretacja uzyskanych wyników. Dla zrozumienia tego problemu musimy wprowadzić cała serię pojęć wstępnych.

W badaniu statystycznym spotkamy się często z pojęciem **zmiennej** (*variable*). Przez zmienne rozumiemy to wszystko, co mierzymy, kontrolujemy i czym manipulujemy w badaniu. Zmienne podlegają dwom typom badań: **badaniu korelacyjnemu** (*correlational research*) i **badaniu ekspe-**

rymentalnemu (*experimental research*). W badaniu pierwszego typu eksperymentator nie wpływa (lub stara się nie wpływać) na żadną z mierzonych zmiennych i mierzy wyłącznie zależności między nimi. W badaniu eksperymentalnym manipulujemy pewnymi zmiennymi (tzw. zmienne niezależne) i mierzymy wpływ tej manipulacji na pozostałe zmienne (zależne).

Jeżeli mierzymy wyłącznie dwie zmienne – na przykład poziom cholesterolu i wartość ciśnienia tętniczego krwi – mamy do czynienia z typowym badaniem korelacyjnym. Natomiast gdy manipulujemy jedną ze zmiennych (np. poziomem cholesterolu przez ustalenie określonego sposobu odżywiania) i określamy wpływ tej manipulacji na drugą zmienną (ciśnienie tętnicze), przeprowadzamy właśnie badanie eksperymentalne. Warto w tym miejscu przypomnieć, że realizacja pomiaru każdej zmiennej może się odbywać w jednej z trzech skal: interwałowej, porządkowej i nominalnej, więc techniki opracowania danych muszą to uwzględniać. Niezależnie jednak od użytej skali pomiarowej możemy stwierdzić, że zmienne są od siebie uzależnione, gdy ich wartości są rozłożone w jakiś spójny sposób (np. jeżeli jedna ze zmiennych rośnie, to druga też rośnie lub na odwrót, gdy pierwsza zmienna rośnie, to druga maleje). W przeciwnym razie możemy twierdzić, że zmienne nie są od siebie zależne, a więc nie są skorelowane.

Każdy związek między zmiennymi jest określony dwiema cechami: **siłą związku** (*relationship strength*) oraz jego **pewnością** (prawdopodobieństwem) (*relationship reliability*). Im większa jest siła związku, tym lepiej na podstawie wartości jednej zmiennej możemy określić („przewidzieć") wartość zmiennej z nią związanej. Pewność związku określa natomiast stopień prawdopodobieństwa tego, że pobierając inny zestaw próbek z tej samej populacji uzyskamy taką samą siłę związku. Siła związku i jego pewność są ze sobą w pewnym stopniu powiązane. W próbce o określonej wielkości im większa jest siła związku między zmiennymi, tym bardziej ten związek jest pewny.

Musimy przy tym zdawać sobie sprawę z faktu, że wykryty statystycznie związek między cechami nie implikuje w żadnym razie istnienia mechanizmu przyczynowo – skutkowego wiążącego rozważane zmienne. Wnioskowanie bowiem można przeprowadzić wyłącznie w jedną stronę: jeśli nie wykryliśmy związku statystycznego, to zapewne nie istnieje także związek przyczynowy. Jeśli jednak istnieje związek statystyczny, to zależność przyczynowa może występować lub nie. Okazuje się bowiem, że można wykryć związek statystyczny nawet w dwu całkowicie przypadkowo dobranych zestawach danych.

Aby pokazać, że jest to możliwe, posłużymy się dość karkołomnym przykładem. Z jednego z podręczników doświadczalnictwa rolniczego za-

czerpnięto dane dotyczące wysokości czaszki ryjówki w różnych miesiącach roku 1947. Drugi zestaw danych to liczba dzieci urodzonych w jednym z poznańskich szpitali w poszczególnych miesiącach 1983 roku. Zastosowanie testu korelacji liniowej Pearsona (będziemy mówić o nim dokładniej dalej) dało wynik wysoce znamienny statystycznie (korelacja istotna na poziomie $p < 0{,}01$). Gdyby podejść bezkrytycznie do uzyskanego wyniku można by wyciągnąć wniosek, że im wyższa była wysokość czaszki ryjówki w danym miesiącu roku 1947, tym więcej rodziło się dzieci trzydzieści sześć lat później. Ten przykład powinien nas przestrzec przed nieostrożnym stosowaniem pewnych narzędzi statystycznych i działaniem typu „skorelujmy wszystko ze wszystkim i może coś z tego wyjdzie", które niestety jest dość powszechne w środowisku lekarskim.

Prawdopodobieństwo pojawienia się istotnego związku między przypadkowymi zestawami danych jest tym większe, im mniej liczebna jest analizowana próbka. Jest to intuicyjnie oczywiste, ponieważ im mniej jest analizowanych pomiarów, tym mniejsza jest liczba możliwych kombinacji ustawienia danych, a co z tym się wiąże – większe prawdopodobieństwo takiego losowego (czytaj: całkiem przypadkowego) ustawienia danych, które wskazywać będzie na istnienie związku między zmiennymi.

Dodatkowym czynnikiem (nie występującym w przytoczonym wyżej przykładzie, ale często spotykanym w praktyce) może być pojawianie się silnego związku statystycznego wtedy, gdy obydwa badane czynniki są uzależnione od jakiegoś innego, nie obserwowanego i nie analizowanego wspólnego czynnika (na przykład klimatycznego). Ujawniające się w takich warunkach bardzo znamienne statystycznie korelacje dowodzą jedynie współwystępowania określonych zjawisk, ale nie ich wzajemnego związku przyczynowego. Dla przykładu, w pewnym szpitalu po rozpoczęciu leczenia jednostki chorobowej *X* wprowadzono nowy, dający świetne efekt lek *Y*. Obniżył on w istotny sposób śmiertelność pacjentów z powodu jednostki chorobowej *X*. Jednocześnie zaobserwowano jednak, że od czasu wprowadzenia tego leku wzrosła liczba zakażeń wewnątrzszpitalnych. Mogłoby to sugerować, że stosowany lek *Y* wywołuje (dodatkowe w stosunku do normalnie występujących) zakażenia, a więc ma pewne szkodliwe działanie uboczne. Tymczasem rzeczywistość jest całkiem inna. Lek *Y*, ratując życie pacjentów, spowodował ich dłuższy pobyt w szpitalu, co oczywiście zwiększyło prawdopodobieństwo zakażenia wewnątrzszpitalnego. A więc próba wycofania leku *Y* z użycia (na podstawie błędnej przesłanki, że zwiększa on liczbę zakażeń wewnątrzszpitalnych) byłabyw działaniem szkodliwym.

Wprowadzimy teraz definicję **hipotezy statystycznej** (*statistical hypothesis*). Załóżmy, że w dwóch grupach pacjentów zmierzyliśmy średni

poziom leukocytów i chcemy stwierdzić, czy średnie te różnią się między sobą istotnie statystycznie, czy też różnica wynika z błędów losowych wynikających z niedokładności pomiaru lub samej procedury pobierania próby z populacji. Oznaczmy średnią w pierwszej grupie przez $\bar{x}$, w drugiej przez $\bar{y}$. Sformułujemy teraz tzw. **hipotezę zerową** (*null hypothesis*): średnie w obu grupach nie różnią się od siebie istotnie statystycznie (w terminologii przedziałów ufności oznacza to, że istnieje pewien obszar, w którym oba przedziały zachodzą na siebie). Zapisujemy to symbolicznie w postaci: H_0: $\bar{x} = \bar{y}$.

Jeszcze raz podkreślamy, że przedmiotem hipotezy jest brak różnicy statystycznej, gdyż różnica algebraiczna może występować. Innymi słowy hipoteza zerowa stwierdza, że ewentualnie występująca różnica między średnimi jest niepowtarzalna i wynika wyłącznie z błędu związanego z pobieraniem próbek.

W stosunku do hipotezy zerowej możemy zbudować **hipotezę alternatywną** (*alternative hypothesis*), która mówi, że średnie w obu grupach różnią się od siebie istotnie statystycznie, czyli pochodzą z różnych populacji (w terminologii przedziałów ufności oba przedziały nie mają obszaru wspólnego). Zapiszemy to symbolicznie jako: H_1: $\bar{x} \neq \bar{y}$.

Błędy pierwszego i drugiego rodzaju

Oba przytoczone stwierdzenia możemy rozpatrywać z punktu widzenia absolutu – czyli rzeczywistej sytuacji w badanej populacji – lub wyniku jakiegoś testu statystycznego, który może (ale nie musi!) potwierdzać rzeczywistą sytuację. Możemy mieć do czynienia z czterema przypadkami przedstawionymi w tablicy 2.

Tablica 2. Definicja błędu pierwszego i drugiego rodzaju

Rzeczywistość / Wynik testu	$H_0 : \bar{x} = \bar{y}$	$H_1 : \bar{x} \neq \bar{y}$
$H_0 : \bar{x} = \bar{y}$	O.K.	błąd drugiego rodzaju
$H_1 : \bar{x} \neq \bar{y}$	błąd pierwszego rodzaju	O.K.

Jeżeli wynik przeprowadzonego przez nas testu statystycznego wykazuje brak istotnej różnicy między średnimi (stwierdza słuszność hipotezy H_0) i tak jest w rzeczywistości, to mamy do czynienia z sytuacją prawidłową (wynik testu potwierdza stan rzeczywisty). Podobną sytuację mamy w przypadku, gdy test wykazuje istnienie istotnej statystycznie różnicy między średnimi i różnica ta istnieje naprawdę.

Mogą się jednak zdarzyć dwa odmienne przypadki. Pierwszy z nich nazywa się **błędem pierwszego rodzaju** (*type I error, alpha error*) i polega na tym, że wynik testu wykazuje istnienie istotnej statystycznie różnicy, podczas gdy w rzeczywistości ta różnica nie jest istotna. Badacz odrzuca zatem hipotezę zerową, stwierdzając istotność różnicy średnich, choć tak naprawdę wynika ona (na przykład) wyłącznie z błędu próbkowania. W literaturze statystycznej błąd ten oznacza się najczęściej symbolem α.

Prawdopodobieństwo popełnienia błędu pierwszego rodzaju (a więc odrzucenia hipotezy zerowej, gdy w rzeczywistości była ona prawdziwa) jest ściśle związane ze znanym już nam z dyskusji nad przedziałami ufności poziomem istotności (*p-value*). Im bardziej obniżymy wartość błędu pierwszego rodzaju, tym mniejsze grozi nam niebezpieczeństwo stwierdzenia istotności różnicy średnich, podczas gdy różnica ta jest w rzeczywistości nieistotna. A zatem im bardziej obniżymy wartość błędu pierwszego rodzaju, tym większa jest istotność statystyczna ewentualnie stwierdzonych różnic. Trzy najczęściej stosowane poziomy istotności to 0,05, 0,01 oraz 0,001. Pierwszy z nich oznacza, że w pięciu przypadkach na sto analizowanych mamy szansę popełnienia błędu pierwszego rodzaju, drugi zmniejsza tę

szansę do jednego na sto, trzeci zaś – do jeden na tysiąc. Wybór jednej z tych wartości zależy oczywiście od badacza, jednak zdecydowanie trudniej jest dowieść istotności na poziomie 0,001 niż na poziomie 0,05.

Wartość *p-value* odpowiada na następujące pytanie: jeżeli hipoteza zerowa jest w rzeczywistości słuszna (a więc średnie naprawdę nie różnią się istotnie statystycznie), to jakie jest prawdopodobieństwo zdarzenia, że dwie losowo wybrane próbki będą miały taką jak zmierzona przez nas (lub wyższą) różnicę średnich?

Im niższą przyjmiemy wartość *p-value* (czyli im wyższa ma być istotność statystyczna), tym mniej jest prawdopodobne, że rejestrowana różnica jest spowodowana błędem losowym, a nie badanym przez nas czynnikiem. Parametr *p-value* kwantyfikuje zatem prawdopodobieństwo tego, że uzyskany wynik jest wyłącznie dziełem przypadku.

Dwa czynniki istotnie zwiększają szansę pojawienia się błędu pierwszego rodzaju: małe liczebności badanych prób oraz zbyt duża liczba przeprowadzanych analiz. O ile pierwszy czynnik jest w miarę oczywisty, o tyle drugi wymaga bliższego komentarza. Załóżmy więc, że badamy istotność różnic średnich między dziesięcioma próbami. Oznacza to konieczność przeprowadzenia 45 porównań. Można się zatem spodziewać, że przy przyjęciu poziomu istotności $p < 0{,}05$ w około dwóch przypadkach (w jednym na każde 20 porównań) stwierdzimy istotną statystycznie różnicę nawet wtedy, gdy wszystkie próby pochodzą z tej samej populacji (a więc mają takie same średnie). Wskazuje to zatem, że do porównań między wieloma grupami musimy stosować specjalnie skonstruowane testy porównań wielokrotnych (*multiple comparison tests*) zamiast analizy porównawczej typu „każda grupa z każdą pozostałą".

Mogłoby się wydawać, że przy testowaniu hipotez statystycznych należałoby dążyć do stosowania jak najwyższych poziomów istotności (minimalizować jak tylko się da możliwość wystąpienia błędu pierwszego rodzaju). Niestety, nieodłączną konsekwencją tego postępowania stanie się zwiększenie prawdopodobieństwa popełnienia **błędu drugiego rodzaju** (*type II error, beta error*). Spojrzawszy ponownie na tabl. 2, zobaczymy, że błąd drugiego rodzaju polega na stwierdzeniu na podstawie przeprowadzonego testu statystycznego, iż średnie nie różnią się istotnie statystycznie, podczas gdy w rzeczywistości taka różnica istnieje. Innymi słowy, błąd drugiego rodzaju określa stopień akceptowanego przez eksperymentatora ryzyka zagubienia rzeczywiście istniejącej różnicy między średnimi. Błąd ten onacza się często literą beta. Na pojawienie się błędu drugiego rodzaju ma istotny wpływ zbyt mała liczebność próby oraz duże rozproszenie wyników pomiarów w próbie.

Niezwykle istotny jest fakt, że zmniejszaniu prawdopodobieństwa wystąpienia błędu jednego typu towarzyszy zawsze wzrost prawdopodobieństwa pojawienia się błędu typu drugiego. Jedyną metodą równoczesnego zmniejszenia błędu obu typów jest powiększenie liczebności próby, na której dokonujemy analizy.

Z błędem drugiego rodzaju jest stowarzyszona **moc testu** (*test power*) używanego do rozstrzygnięcia, która z hipotez jest słuszna. Jest ona zdefiniowana jako (*1-beta*) i odpowiada na pytanie: Jeżeli różnica (lub związek) między zmiennymi w całej populacji jest wyrażona jakąś wartością, to jakie jest prawdopodobieństwo znalezienia istotnej różnicy (istotnego związku) równej tej wartości w losowo wybranych próbach o określonej wielkości? Warto zwrócić uwagę, że moc testu jest ściśle zależna od liczebności analizowanej próby. Zależność tę często wykorzystuje się do znalezienia minimalnej wielkości próby, która pozwoli na znalezienie różnicy między zmiennymi o zadanej wielkości.

Dociekliwego Czytelnika będzie na pewno nurtowało pytanie, którego typu błąd jest korzystniej minimalizować. Niestety, nie ma na nie jednoznacznej odpowiedzi. Spróbujmy posłużyć się następującym prostym przykładem, za pomocą którego zilustrujemy jednocześnie fakt, iż hipoteza zerowa i alternatywna nie muszą być konstruowane wyłącznie w aspekcie istotnej różnicy między średnimi arytmetycznymi lub jej braku.

Najbardziej ogólna definicja hipotezy zerowej mówi, że dwie pobrane zmienne (lub większa ich liczba) pochodzą z tej samej populacji. Hipoteza alternatywna stwierdza w tym przypadku, że zmienne te pochodzą z różnych populacji. Różnica średnich jest jedynie szczególnym przypadkiem tak sformułowanych hipotez. Hipotezy badawcze i testy statystyczne możemy więc również stosować do porównywania istotności różnic innych parametrów charakteryzujących rozkład próby – na przykład wariancji. Można też testować zgodność między dwiema grupami lub większą ich liczbą. Testy te mogą służyć do porównywania zgodności rozkładów danych w badanych próbach, istotności współczynników korelacji, proporcji itp.

Przykład 10

W badaniach prowadzonych nad nowym lekiem staramy się rozstrzygnąć kwestię, czy wykazuje on niebezpieczne dla pacjenta działania uboczne (np. działa uszkadzająco na mięsień sercowy). Możemy skonstruować hipotezę zerową H_0 która mówi, że lek nie daje niebezpiecznych skutków ubocznych. Hipoteza alternatywna H_1 ma natomiast postać: lek powoduje skutki uboczne – uszkadza mięsień sercowy. Popełnienie błędu pierwszego rodzaju będzie zatem polegało na stwierdzeniu, że lek uszkadza serce,

podczas gdy w rzeczywistości lek nie daje takiego skutku. Błąd drugiego rodzaju to wnioskowanie, że lek nie daje komplikacji, podczas gdy w rzeczywistości powoduje nieodwracalne uszkodzenie mięśnia sercowego. Jest rzeczą oczywistą, że z punktu widzenia ochrony pacjenta należy zminimalizować błąd drugiego rodzaju. Interes firmy farmaceutycznej wprowadzającej lek wymaga natomiast minimalizacji błędu pierwszego rodzaju.

Podsumowując, błąd pierwszego rodzaju to akceptowalne przez eksperymentatora prawdopodobieństwo znalezienia w próbie, którą bada istotnej różnicy, podczas gdy ta różnica w populacji nie występuje. Błąd drugiego rodzaju to akceptowalne przez eksperymentatora prawdopodobieństwo zagubienia istotnej różnicy w badanej próbie, podczas gdy ta różnica jest istotna statystycznie w populacji, z której próba pochodzi.

Testy jedno- i dwustronne

Konstruowana przez nas do tej pory hipoteza alternatywna H_1: $\bar{x} = \bar{y}$ nie zakładała konkretnego znaku różnicy między średnimi – po prostu stwierdzała, że średnie różnią się od siebie istotnie statystycznie. Taki typ hipotezy nazywamy **hipotezą dwustronną** (*two-tailed hypothesis*). W praktyce, gdy podejrzewamy, jaki powinien być kierunek zmian, można stosować również drugi typ hipotez alternatywnych – **hipotezy jednostronne** (*one-tailed hypothesis*). Ich postać jest następująca:

H_2: $\bar{x} > \bar{y}$
lub
H_3: $\bar{x} < \bar{y}$

Dla przykładu załóżmy, że badamy ciśnienie tętnicze w grupie chorych leczonych nowym preparatem, który je obniża. Symbolem $\bar{x}$ oznaczmy średnią arytmetyczną ciśnienia przed terapią, symbolem $\bar{y}$ – po przeprowadzeniu leczenia. Przyjęcie hipotezy alternatywnej H_1: $\bar{x} \neq \bar{y}$ pozwoli nam wyłącznie na sprawdzenie, czy nastąpiła istotna zmiana ciśnienia krwi (jego wzrost lub spadek), czy też nie. Hipoteza natomiast H_2: $\bar{x} > \bar{y}$ jednoznacznie wskazuje na kierunek zmian (obniżenie ciśnienia pod wpływem działania leku hipotensyjnego), a o to w istocie nam chodzi. W zasadzie hipotezy jednostronne powinny być stosowane wtedy, gdy jest możliwa zmiana tylko w jednym kierunku.

Testy porównań wielokrotnych

Do tej pory posługiwaliśmy się przykładem badania istotności różnicy parametrów między dwiema grupami. Nasze rozumowanie możemy rozszerzyć na porównywanie parametrów między większą liczbą grup. W przypadku porównywania średnich hipoteza zerowa przyjmie postać:

H_0: $\bar{x} = \bar{y} = \bar{z} = ...$
a hipoteza alternatywna
H_1: $\bar{x} \neq \bar{y} \neq \bar{z} \neq ...$

W tym miejscu wprowadzimy pojęcie modelu badania. Gdy porównujemy ze sobą dwie grupy lub większą ich liczbę, skonstruowane w ten sposób, że każdy obiekt badania może należeć do wyłącznie jednej z nich, mówimy o **modelu zmiennych nie powiązanych** (*unpaired model*). Liczebność każdej z grup może być inna. Dla przykładu jeżeli badamy zależność wyniku terapii od dawki badanego leku, to grupę leczonych pacjentów dzielimy na szereg podgrup, w których stosujemy różne dawki. Żaden z pacjentów nie może należeć do więcej niż jednej podgrupy – każdy otrzymuje ściśle określoną dawkę leku. Jest to klasyczny model zmiennych nie powiązanych.

Rozważmy teraz inną sytuację. Mamy pewną grupę pacjentów, którym podajemy ten sam lek w identycznej dawce. Interesuje nas dynamika zmian badanego parametru w trakcie kuracji. Zwróćmy uwagę, że mamy tu do czynienia z badaniem ciągle tych samych obiektów, lecz w różnych momentach czasowych (np. przed rozpoczęciem leczenia, w pierwszym, siódmym, czternastym i dwudziestym ósmym dniu kuracji). Taki model nazywamy **modelem zmiennych powiązanych** (*paired model, matched model*). Liczebność każdej porównywanej grupy musi być taka sama. Jeżeli się zdarzy, że z jakichkolwiek przyczyn zestaw naszych danych jest niekompletny (np. zagubiono wynik badania lub pacjent nie zgłosił się w określonym czasie), należy odpowiadający mu przypadek całkowicie wyłączyć z analizy. Niektóre pakiety statystyczne umożliwiają zastąpienie brakujących danych wartością średnią pozostałych wyników albo liniową lub nieliniową procedurą interpolacyjną, lecz działania tego typu należy prowadzić nadzwyczaj ostrożnie.

Podstawy metodyki testowania hipotez statystycznych

Po wprowadzeniu pojęcia hipotezy zerowej i alternatywnej oraz błędów pierwszego i drugiego rodzaju możemy przejść do testowania hipotez, a więc próby rozstrzygnięcia, która z nich (zerowa czy alternatywna) jest bardziej prawdopodobna. Tym samym chcemy odpowiedzieć na pytanie, czy badane próby pochodzą z tej samej populacji, czy też naprawdę różnią się od siebie i pochodzą z różnych populacji. Wprowadzana przez nas metodyka jest niezwykle silnym i uniwersalnym narzędziem. Będzie miała zastosowanie w każdej skali pomiarowej, przy dowolnej liczbie porównywanych grup i dowolnie przyjętym modelu badania.

Schemat naszego postępowania będzie w zasadzie wszędzie taki sam. Jedynymi różnicami będą kształt przyjętej hipotezy zerowej i alternatywnej oraz postać funkcji testującej. Na tę ostatnią wpływają następujące czynniki: skala pomiarowa, w której przeprowadzono eksperyment, schemat prowadzonego badania (zmienne powiązane lub nie powiązane), typ rozkładu danych w populacji, z której pobrano dane do analizy oraz liczba porównywanych grup.

Każdy proces testowania hipotez składa się z pięciu etapów:

- skonstruowania hipotezy zerowej i alternatywnej,
- ustalenia wielkości próby,
- zebrania danych,
- przeprowadzenia analizy statystycznej za pomocą odpowiedniego testu w celu ustalenia prawdopodobieństwa, że hipoteza zerowa jest słuszna,
- pozostawienia lub odrzucenia hipotezy zerowej.

W pierwszym etapie musimy podjąć decyzję sformułowania kształtu obu hipotez. Będzie on zależał od tego, w jakiej skali przeprowadziliśmy pomiar, jak wiele grup chcemy między sobą porównywać, czy zmiany których się spodziewamy mogą zachodzić tylko w jednym czy w obu kierunkach itp.

Warto zwrócić uwagę, że w praktyce eksperymentatorzy bardzo często odrzucają punkt drugi, co niestety stanowi błąd w sztuce badawczej! W istocie niezwykle ważnym problemem jest ustalenie, jak wielka musi być próba, aby zapewniła zarówno kliniczną, jak i statystyczną istotność uzyskiwanych wyników. Niemal każdy pakiet oprogramowania statystycznego

zawiera procedury wyznaczania tego parametru, nie ma więc konieczności przytaczania tu jakichkolwiek wzorów. Istotny jest natomiast sam sposób prowadzenia analizy. Po pierwsze eksperymentator musi *a priori* ustalić wielkość różnicy, którą będzie z punktu widzenia klinicznego uważał za istotną. Dla przykładu trzeba rozstrzygać, czy spadek liczby białych ciałek krwi o 1000 jest już dla lekarza zmianą istotną z punktu widzenia klinicznego, czy też nie. Czy różnica masy ciała noworodków rzędu 100 gramów ma konsekwencje kliniczne, czy może jest już ważna różnica wynosząca 50 gramów?

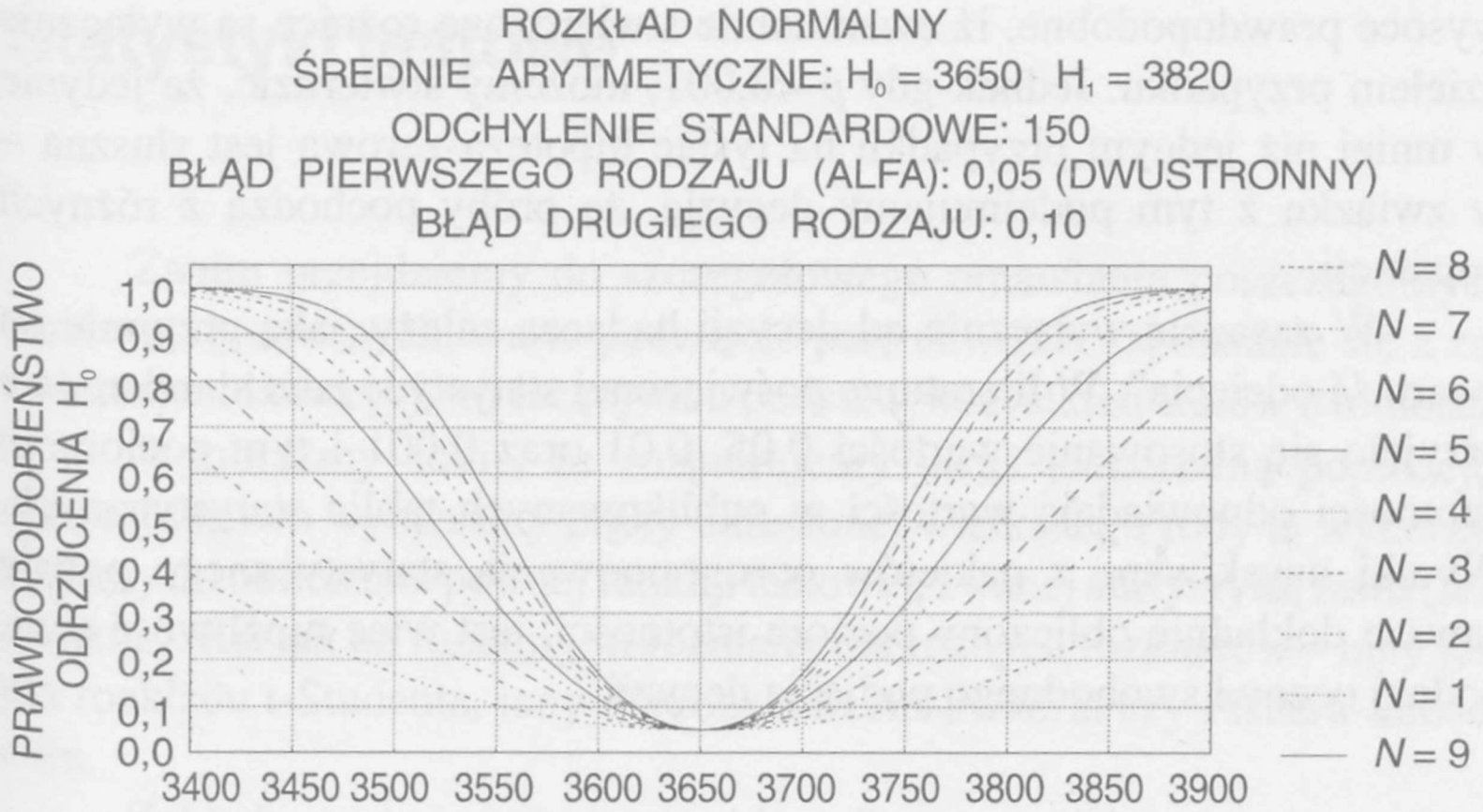
Im mniejszą różnicę chcemy być w stanie wykryć, określając ją jednocześnie jako istotną statystycznie, tym analizowana próba musi być większa. Widać stąd, że nie oszacowując wielkości próby przed zebraniem danych, możemy przeprowadzać obliczenia na zbiorach zbyt małych, nie mogących z natury rzeczy ujawnić zmian istotnych dla badacza.

Dla ustalenia niezbędnej wielkości próby musimy znać również szacunkową wielkość odchylenia standardowego w badanej populacji. Z reguły estymacji tej wartości dokonujemy, prowadząc badanie pilotowe na losowo dobranej próbie.

Musimy również założyć dla naszego badania prawdopodobieństwa popełnienia błędu pierwszego rodzaju. Im mniejsze ma być ryzyko odrzucenia hipotezy zerowej spowodowane błędem losowym (czyli ryzyko znalezienia fikcyjnych różnic), tym większa musi być próba.

Ostatni parametr niezbędny do wyznaczenia wielkości próby to moc stosowanego testu. Określa ona prawdopodobieństwo znalezienia istotnych różnic o ściśle zadanych wielkościach. Im większa ma być moc testu, tym większa musi być próba. Niektóre pakiety statystyczne (np. CSS: STATISTICA) pozwalają nie tylko na oszacowanie minimalnej wielkości próby, ale podają również w formie wykresu (tzw. *operating characteristic curve*), zależności między wielkością próby a mocą testu. Na rycinie 13 przedstawiliśmy przykładowe krzywe zależności mocy testu od liczebności próby dla następujących danych początkowych: chcemy ocenić, jaka powinna być wielkość próby, aby na poziomie istotności 0,05 móc zidentyfikować różnicę między średnimi masy ciała noworodka w dwóch próbach, wynoszącymi odpowiednio 3650 i 3820 g, (różnica mas 170 g,) przy odchyleniu standardowym 150 g.

Z obliczeń wynika, że minimalna wielkość próby przy przyjętym prawdopodobieństwie popełnienia błędu drugiego rodzaju równym 0,1 wynosi 9. Proszę zwrócić uwagę, że jeżeli zmniejszymy liczebność próby na przykład do 5 przypadków, to – nie zmieniając pozostałych założonych parametrów – będziemy mogli wykryć różnicę mas nie mniejszą niż 220



Ryc. 13. Wpływ liczebności próby na możliwość detekcji określonej różnicy masy przy założonym poziomie prawdopodobieństwa popełnienia błędu pierwszego i drugiego rodzaju

gramów. Gdy przy założonym odchyleniu standardowym 150 g i liczebności próby równej 9 różnica mas przekroczy 220 g, to prawdopodobieństwo popełnienia przez nas błędu drugiego rodzaju (stwierdzenia braku różnic, podczas gdy są one istotne i nie wynikają z błędów losowych) zmniejsza się prawie do zera.

Trzeci etap opisywanej metodyki testowania hipotez to zbieranie danych. Pamiętajmy, że nawet najtrafniej dobrane procedury statystyczne nie poprawią uzyskiwanych wyników, jeśli dane eksperymentalne były zebrane nieprawidłowo. Zagadnienie to nie wchodzi jednak w zakres tematyki niniejszej książki i dlatego odsyłamy Czytelnika do specjalistycznej literatury.

W kolejnym etapie badacz musi podjąć decyzję, który test statystyczny powinien w danej sytuacji zastosować. Jak już wspomnieliśmy, zależy od skali i pomiarowej modelu badania, liczebności próby oraz typu rozkładu danych. Szczegóły doboru testu omówimy w kolejnym rozdziale.

Ostatnia faza testowania hipotez to podjęcie decyzji o przyjęciu lub odrzuceniu hipotezy zerowej oraz interpretacja uzyskanych wyników.

Gdy stwierdzamy, że próby pochodzą z różnych populacji mamy na myśli, że oszacowana wartość prawdopodobieństwa *p-value* jest na tyle mała, że możemy odrzucić hipotezę zerową. Jeżeli dla przykładu otrzymamy wartość $p < 0{,}17$, to nie będziemy odrzucali hipotezy zerowej, gdyż jest

wysoce prawdopodobne, iż ewentualnie stwierdzone rożnice są wyłącznie dziełem przypadku. Jednak gdy p <0,001, możemy stwierdzić, że jedynie w mniej niż jednym przypadku na tysiąc hipoteza zerowa jest słuszna – w związku z tym podejmujemy decyzję, że próby pochodzą z różnych populacji.

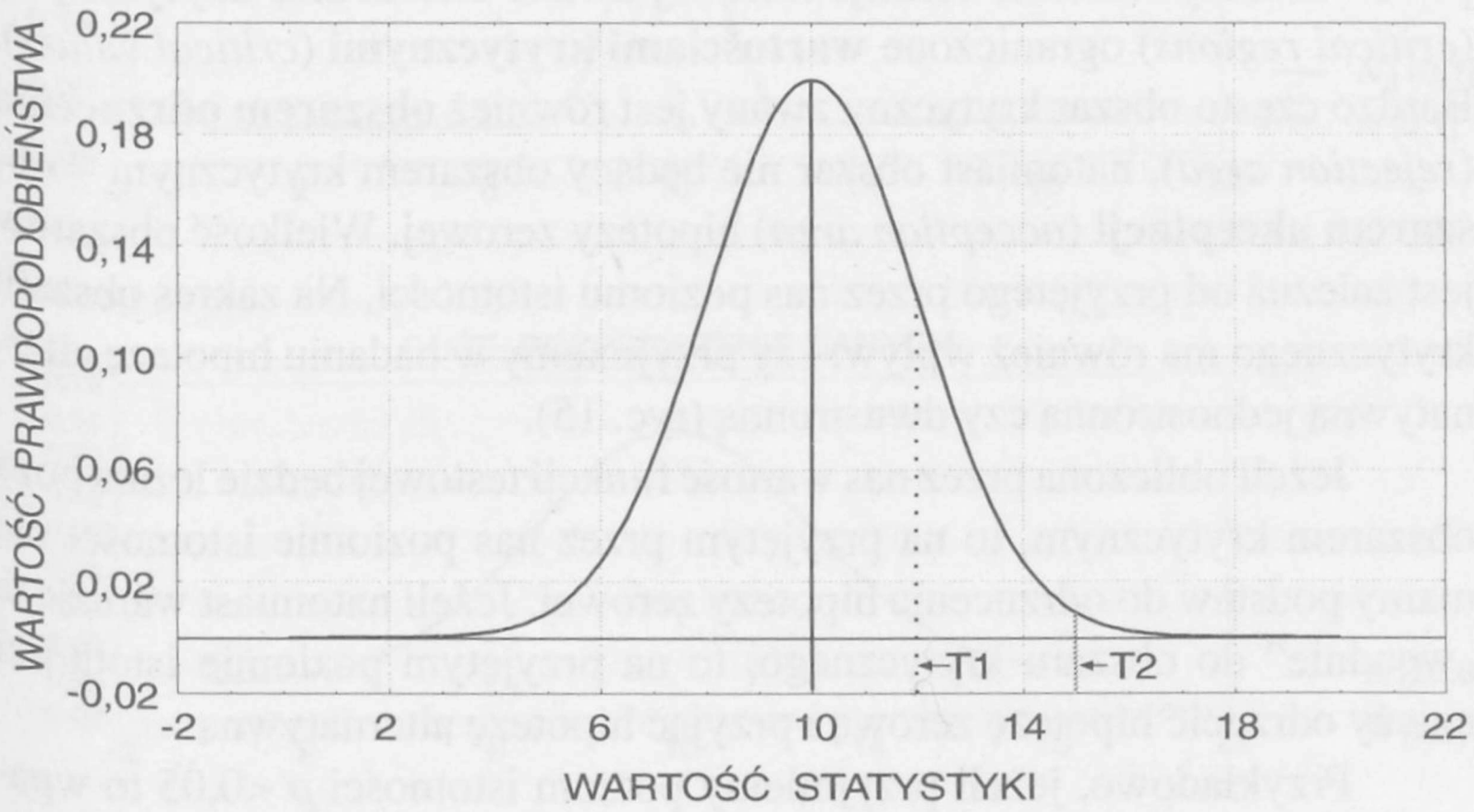
W zasadzie wyłącznie od decyzji badacza zależy, jaką przyjmie on „wartość odcięcia". W literaturze poświęconej statystyce jako standardowe przyjęło się stosowanie wartości 0,05, 0,01 oraz 0,001 i tym poziomom istotności odpowiadają wartości w publikowanych tablic statystycznych. Wyniki uzyskiwane z pakietów oprogramowania statystycznego podają zawsze dokładnie obliczony poziom istotności, jest więc możliwość swobodnej oceny i swobodnego podjęcia decyzji.

Statystyki testowe

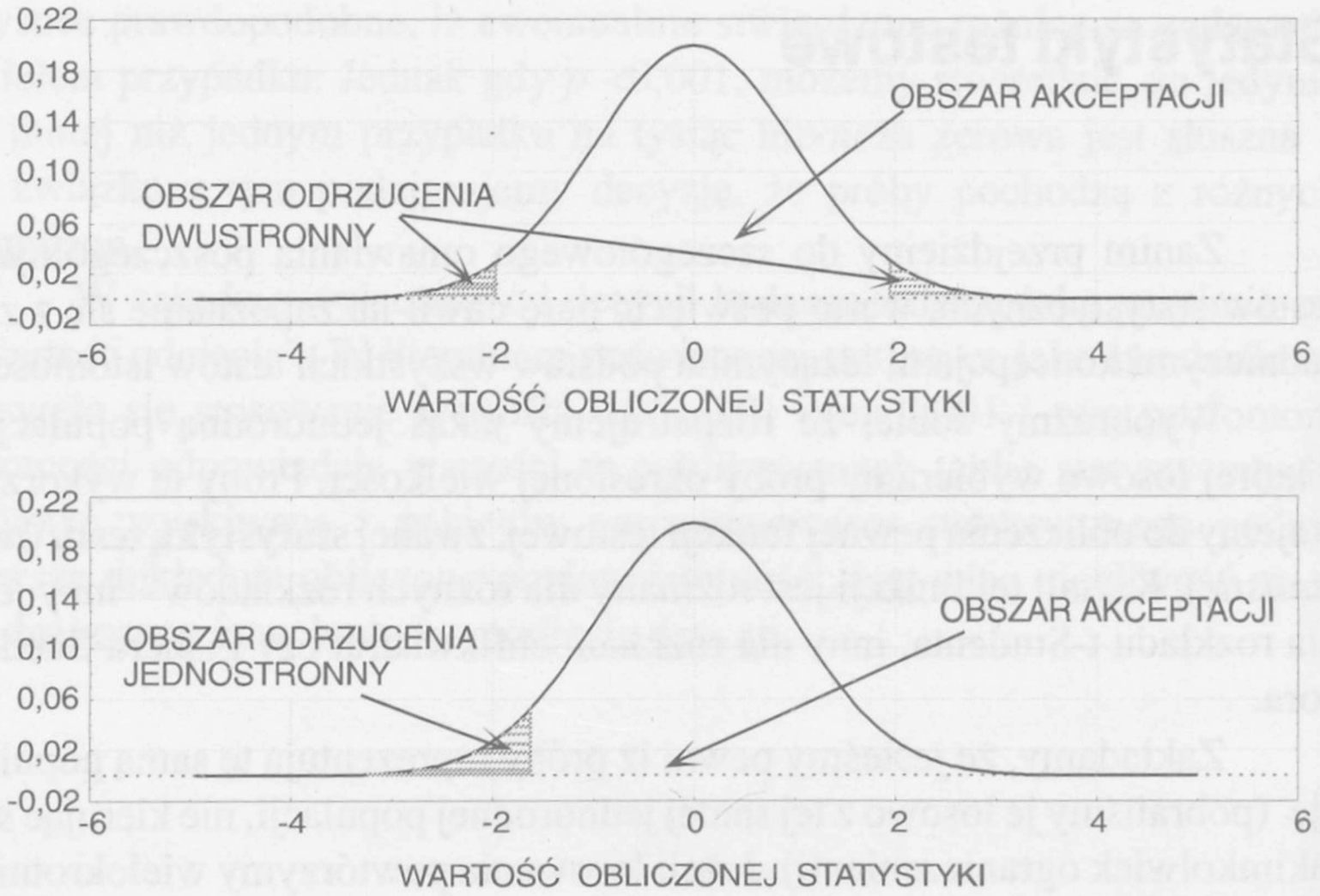
Zanim przejdziemy do szczegółowego omawiania poszczególnych testów statystycznych, warto poświęcić parę chwil na zapoznanie się z zasadniczymi koncepcjami leżącymi u podstaw wszystkich testów istotności.

Wyobraźmy sobie, że rozpatrujemy jakąś jednorodną populację, z której losowo wybieramy próby określonej wielkości. Próby te wykorzystujemy do obliczenia pewnej funkcji testowej, zwanej **statystyką testu** (*test statistic*). Kształt tej funkcji jest rozmaity dla różnych rozkładów – inny jest dla rozkładu t-Studenta, inny dla rozkładu chi-kwadrat czy Fishera-Snedecora.

Zakładamy, że jesteśmy pewni iż próby reprezentują tę samą populację (pobraliśmy je losowo z tej samej jednorodnej populacji, nie kierując się jakimkolwiek ograniczeniami). Jeżeli losowania powtórzymy wielokrotnie, to uzyskamy serię rozmaitych wartości funkcji testowej. Możemy je przedstawić w postaci krzywej rozkładu częstości. Krzywa ta reprezentuje wszystkie mogące się pojawić wartości funkcji testowej, jeżeli stosowane do jej obliczenia losowe próby pochodziły z tej samej jednorodnej populacji. Rozkład ten ma swoją wartość średnią i odchylenie standardowe. Im bardziej wartość statystyki testowej będzie się różniła od wartości średniej jej rozkładu, tym mniej prawdopodobne jest jej znalezienie (ryc. 14). Statystycy



Ryc. 14. Zależność prawdopodobieństwa znalezienia określonej wartości statystyki od odległości od wartości średniej arytmetycznej jej rozkładu



Ryc. 15. Krzywa rozkładu wartości statystyki dla funkcji testowej – jednostronny i dwustronny obszar krytyczny

wyznaczyli za nas tego typu krzywe dla rozmaitych rozkładów (t-Studenta, chi-kwadrat, Fishera-Snedecora, Weibulla itp.) i możemy z nich korzystać w naszych badaniach.

Idea postępowania jest następująca. Pod krzywą rozkładu ustala się pewne zakresy wartości funkcji testowej zwane **obszarami krytycznymi** (*critical regions*) ograniczone **wartościami krytycznymi** (*critical values*). Bardzo często obszar krytyczny zwany jest również **obszarem odrzucenia** (*rejection area*), natomiast obszar nie będący obszarem krytycznym – **obszarem akceptacji** (*acception area*) hipotezy zerowej. Wielkość obszarów jest zależna od przyjętego przez nas poziomu istotności. Na zakres obszaru krytycznego ma również wpływ, czy przyjmiemy w badaniu hipotezę alternatywną jednostronną czy dwustronną (ryc. 15).

Jeżeli obliczona przez nas wartość funkcji testowej będzie leżała poza obszarem krytycznym, to na przyjętym przez nas poziomie istotności nie mamy podstaw do odrzucenia hipotezy zerowej. Jeżeli natomiast wartość ta „wpadnie" do obszaru krytycznego, to na przyjętym poziomie istotności należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną.

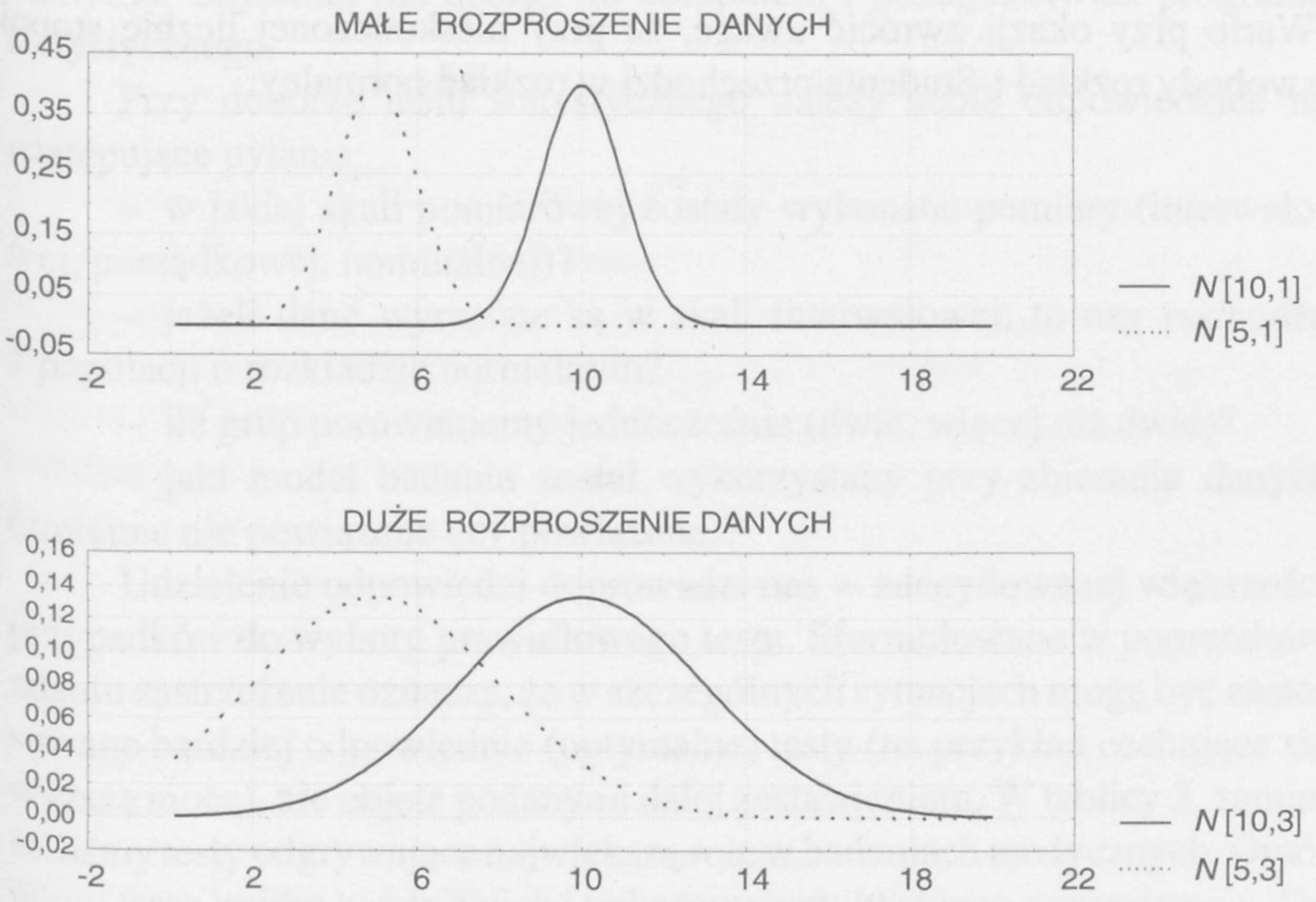
Przykładowo, jeżeli przyjmiemy poziom istotności $p < 0{,}05$ to wpadnięcie obliczonej funkcji testowej do obszaru krytycznego ograniczonego wartością krytyczną dla tego poziomu oznacza, że jedynie w co najwyżej

pięciu przypadkach na sto stwierdzona przez nas różnica mogła być wywołana błędem próbkowania, w dziewięćdziesięciu pięciu natomiast jest wywołana faktem, że próby pochodzą z różnych populacji.

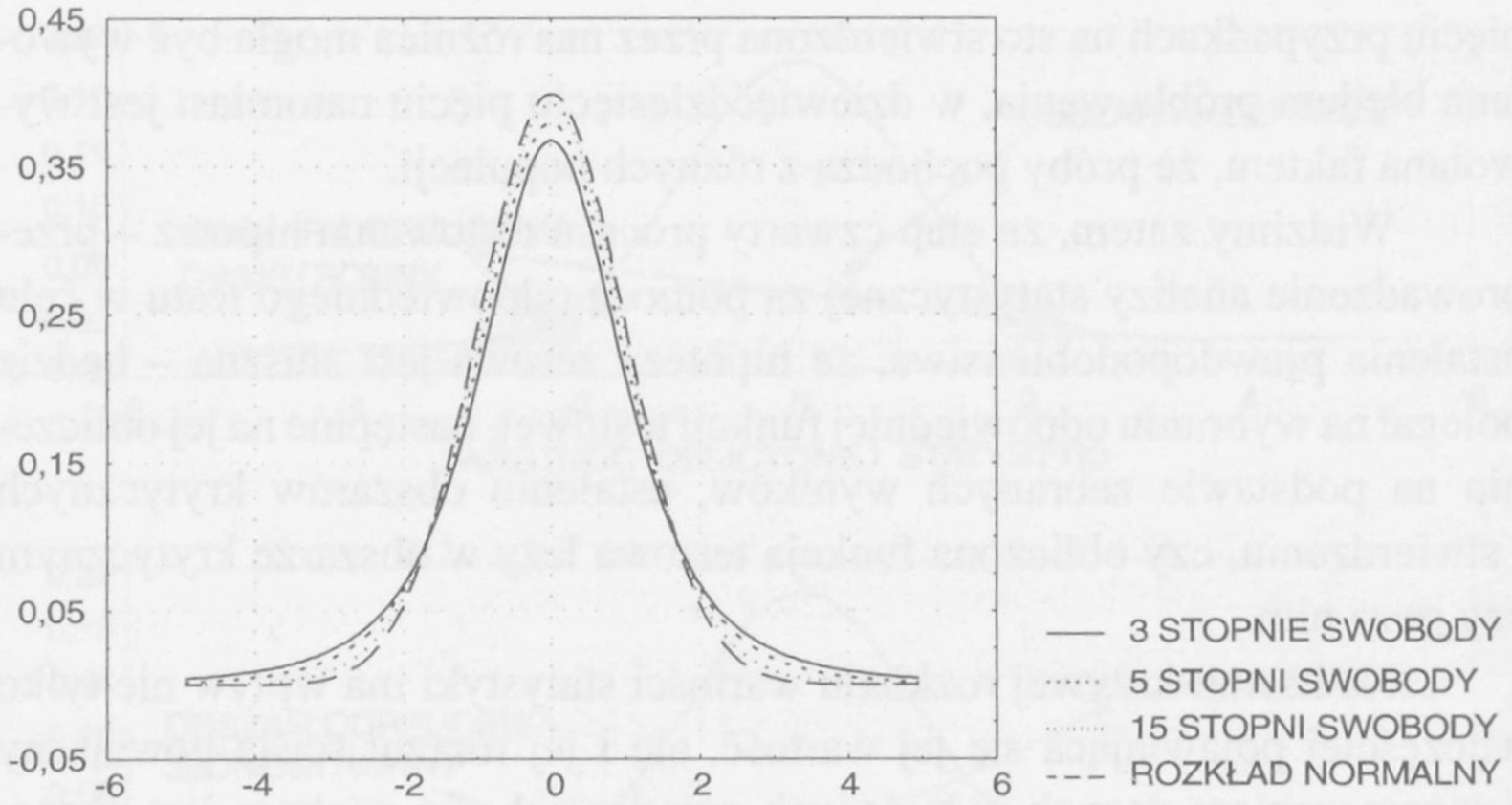
Widzimy zatem, że etap czwarty procesu testowania hipotez – przeprowadzenie analizy statystycznej za pomocą odpowiedniego testu w celu ustalenia prawdopodobieństwa, że hipoteza zerowa jest słuszna – będzie polegał na wybraniu odpowiedniej funkcji testowej, następnie na jej obliczeniu na podstawie zebranych wyników, ustaleniu obszarów krytycznych i stwierdzeniu, czy obliczona funkcja testowa leży w obszarze krytycznym czy poza nim.

Na kształt krzywej rozkładu wartości statystyki ma wpływ nie tylko najczęściej pojawiająca się jej wartość, ale i jej rozrzut ściśle powiązany z rozproszeniem danych w badanych populacjach. Im większe jest to rozproszenie, tym mniejsze jest prawdopodobieństwo, że porównywane podgrupy pochodzą z różnych populacji (ryc. 16).

Kształt funkcji rozkładu jest również uwarunkowany tak zwaną **liczbą stopni swobody** (*number of degrees of freedom*), o której wspomnieliśmy przy okazji omawiania przedziałów ufności. Szczegółowe omawianie tej



Ryc. 16. Wpływ rozproszenia porównywanych danych na prawdopodobieństwo stwierdzenia, że dane pochodzą z różnych populacji



Ryc. 17. Wpływ liczby stopni swobody na kształt rozkładu t-Studenta

wielkości wykracza poza przyjęty przez nas zakres materiału. Warto jednak wiedzieć, że wielkość ta jest zależna od wielkości analizowanych prób i dla każdej funkcji wyrażona innym wzorem. Przykładowy wpływ zmiany liczby stopni swobody na kształt rozkładu t-Studenta jest pokazany na rycinie 17. Warto przy okazji zwrócić uwagę, że przy nieskończonej liczbie stopni swobody rozkład t-Studenta przechodzi w rozkład normalny.

Przegląd ważniejszych testów statystycznych

Uwagi ogólne

Po wprowadzeniu w poprzednim rozdziale podstawowych pojęć z teorii testowania hipotez możemy przejść do zagadnień praktycznych. Podamy teraz Czytelnikowi wskazówki, w jakich sytuacjach badawczych należy stosować wybraną grupę testów statystycznych. Z góry zastrzegamy, że nie omawiamy tu wszystkich istniejących testów, lecz wyłącznie te, które mają największe praktyczne znaczenie w badaniach medycznych. Nie podajemy również konkretnych wzorów na obliczenie statystyki testowej, gdyż zakładamy, że Czytelnik ma dostęp do komputera i jakiegokolwiek programu statystycznego.

Przy doborze testu statystycznego należy sobie odpowiedzieć na następujące pytania:

– w jakiej skali pomiarowej zostały wykonane pomiary (interwałowej, porządkowej, nominalnej)?

– jeżeli dane wyrażone są w skali interwałowej, to czy pochodzą z populacji o rozkładzie normalnym?

– ile grup porównujemy jednocześnie (dwie, więcej niż dwie)?

– jaki model badania został wykorzystany przy zbieraniu danych (zmienne nie powiązane czy powiązane)?

Udzielenie odpowiedzi doprowadzi nas w zdecydowanej większości przypadków do wyboru prawidłowego testu. Sformułowane w poprzednim zdaniu zastrzeżenie oznacza, że w szczególnych sytuacjach mogą być zastosowane bardziej odpowiednie (optymalne) testy (na przykład cechujące się wyższą mocą), nie objęte podanymi dalej zestawieniem. W tablicy 3. zamieściliśmy testy odgrywające największą rolę w badaniach medycznych. Omówimy teraz krótko każdy z nich i pokażemy przykład jego zastosowania. Do obliczeń autorzy użyli pakietów statystycznych: InStat v.2.05a firmy GraphPad Software, CSS:STATISTICA v.5.0 firmy StatSoft oraz StatXact-3

Tablica 3. Wybór najczęściej stosowanych testów w badaniach medycznych i warunki ich stosowania

Liczba grup / Skala	Warunki dodatkowe	Dwie grupy	Dwie grupy	Więcej niż dwie grupy	Więcej niż dwie grupy
		zmienne nie powiązane	zmienne powiązane	zmienne nie powiązane	zmienne powiązane
Interwałowa	normalność rozkładu	test t-Studenta (nie powiązane)	test t-Studenta (powiązane)	analiza wariancji	analiza wariancji
Interwałowa	brak normalności rozkładu	test Manna-Whitneya	test Wilcoxona	test Kruskala-Wallisa	test Friedmana
Porządkowa	–	test Manna-Whitneya	test Wilcoxona	test Kruskala-Wallisa	test Friedmana
Nominalna	–	test chi-kwadrat lub dokładny test Fishera-Snedecora	test znaków lub test McNemary	test chi-kwadrat lub dokładny test Fishera-Snedecora	test Q-Cochrana

v.3.0.2 firmy Cytel Software. Tam, gdzie nie zaznaczono tego w sposób jawny przyjęto poziom istotności statystycznej α=0,05.

Test t-Studenta dla zmiennych nie powiązanych (*t-Student test for unpaired data*)

Do użycia tego testu wyniki pomiarów muszą być przedstawione w skali interwałowej i zebrane w dwóch grupach według modelu zmiennych nie powiązanych. Dane w obu porównywanych grupach powinny pochodzić z populacji o rozkładzie normalnym. Nie istnieje teoretycznie dolna granica liczby danych, dla której test ten można by stosować jednak pamiętajmy, że próby o większej liczebności lepiej odwzorowują rozkład Gaussa*. Na pewno wiarogodność otrzymanych wyników będzie więc wyższa przy analizie grup zawierających ponad 30 pomiarów niż grup o mniejszej liczebności.

Przy stosowaniu testu t-Studenta pamiętajmy, że duża skośność rozkładu powoduje, iż średnia arytmetyczna nie jest dobrą miarą tendencji centralnej. W takiej sytuacji istnieją dwa wyjścia. Można zastosować:

– odpowiednią transformację danych przykładowo przez logarytmowanie lub pierwiastkowanie i użycie testu t-Studenta dla zmiennych przetransformowanych (tego rodzaju transformacje powodują zmniejszenie skośności rozkładu danych),

– słabszy test nie zakładający jakiegokolwiek kształtu rozkładu danych (tzw. *distribution-free test*) – odpowiedni będzie w tym przypadku test Manna-Whitneya.

Stosowana w teście t-Studenta hipoteza zerowa dotyczy braku istotnej różnicy między wartościami średnimi pomiarów w obu grupach. Hipoteza alternatywna zakłada istnienie istotnej statystycznie (nie spowodowanej przypadkiem) różnicy między tymi średnimi. Pamiętajmy, że hipotezę alternatywną możemy sformułować jako dwustronną lub jednostronną. Weryfikacji, która z hipotez jest słuszna dokonujemy obliczywszy wartość statystyki t-Studenta dla zmiennych nie powiązanych i sprawdzając, czy leży ona w obszarze krytycznym rozkładu na danym poziomie istotności (przyjmujemy wtedy za słuszną hipotezę alternatywną), czy też poza tym obszarem

---

* Symulacyjne badania komputerowe, w których wykorzystano tzw. metody Monte-Carlo wykazały, że konsekwencje związane z naruszeniem normalności rozkładu danych dla większości opartych na tym założeniu testów statystycznych nie są aż tak duże, jak to początkowo podejrzewano. Nie wszystkie jednak testy mają taką samą czułość na odstępstwa od tego założenia. Dla przykładu test Fishera-Snedecora stosowany w weryfikacji istotności różnic wariancji jest znacznie bardziej czuły na odstępstwa od normalności rozkładu niż test t-Studenta dla zmiennych powiązanych lub nie powiązanych.

(pozostajemy przy hipotezie zerowej). Przykładowo, jeżeli obliczona wartość statystyki t-Studenta będzie co do modułu (wartości bezwzględnej) większa od wartości krytycznej rozkładu t-Studenta na przyjętym poziomie istotności alfa, to należy hipotezę zerową odrzucić. W przeciwnym razie przyjmujemy słuszność hipotezy zerowej.

Zdecydowana większość pakietów statystycznych podaje obok wartości obliczonej statystyki omawianą już kilkukrotnie wartość *p-value*. Jeżeli jest ona większa od założonego poziomu istotności statystycznej (najczęściej przyjmuje się wartości alfa = 0,05, 0,01 lub 0,001), to przyjmuje się za słuszną hipotezę zerową. W przeciwnym razie odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej.

Liczba stopni swobody statystyki testu t-Studenta wynosi $(N_1 + N_2 - 2)$, gdzie N_1, N_2 – to odpowiednio liczebność pierwszej i drugiej grupy.

Użycie testu t-Studenta dla zmiennych nie powiązanych wymaga jeszcze jednego, silnego założenia o jednorodności wariancji w porównywanych grupach. Warunek ten testuje się za pomocą testu Fishera-Snedecora budując parę hipotez:

$H_0 : SD^2_x = SD^2_y$; różnica wariancji nie jest istotna statystycznie (innymi słowy wariancje są jednorodne, homogeniczne),

$H_1 : SD^2_x \neq SD^2_y$; różnica wariancji jest istotna statystycznie (wariancje są niejednorodne, niehomogeniczne).

Weryfikacji, która z hipotez jest bardziej prawdopodobna dokonujemy podobnie, jak dla hipotez dotyczących różnicy średnich. Gdy wariancje w obu grupach różnią się istotnie statystycznie, do porównania istotności różnic średnich musimy użyć zmodyfikowanego testu t-Studenta znanego jako test Welcha.

Przykład 11

Wyobraźmy sobie, że chcemy stwierdzić czy typ stosowanej diety (z dużym albo małym spożyciem soli kuchennej ma wpływ na średnie ciśnienie skurczowe w grupie badanych osób. Badanie przeprowadzono w grupie 2300 mężczyzn w wieku od 20 do 25 lat po jednorocznym stosowaniu określonego typu diety. Uzyskane wyniki ujęto w tabl. 4.

Użyto skali pomiarowej interwałowej. Porównano dwie grupy wyników zebranych w schemacie badań nie powiązanych (żaden pacjent należący do grupy spożywającej pokarmy o dużej zawartości soli nie może jednocześnie należeć do grupy ograniczającej spożycie soli i vice versa). Na tym etapie nie możemy jeszcze stosować testu t-Studenta dla zmiennych nie powiązanych, ponieważ nie jesteśmy pewni, czy wariancje w obu grupach

Tablica 4. Wartość ciśnienia skurczowego w zależności od stosowanego typu diety

	Wysoka zawartość soli kuchennej w pożywieniu	Niska zawartość soli kuchennej w pożywieniu
Średnia wartość ciśnienia skurczowego	135,1	126,8
Odchylenie standardowe	11,2	10,5
Liczba pacjentów	1250	1050

są homogeniczne. W celu rozstrzygnięcia tego problemu użyjemy testu Fishera-Snedecora, budując przed tym parę hipotez:

$H_0 : SD^2_x = SD^2_y$;

$H_1 : SD^2_x \neq SD^2_y$;

Otrzymana z obliczeń wartość statystyki F jest równa 1,138, a *p-value* wynosi 0,0149, co wskazuje, że przy przyjętym przez nas standardowo poziomie istotności $\alpha = 0,05$ wariancje są niehomogeniczne i nie wolno nam stosować testu t-Studenta dla zmiennych nie powiązanych, tylko test Welcha.

Konstruujemy teraz kolejną parę hipotez (tym razem do testowania różnicy średnich, a nie wariancji):

$H_0 : \bar{x} = \bar{y}$;

$H_1 : \bar{x} \neq \bar{y}$;

W wyniku użycia testu Welcha otrzymujemy wartość statystyki $t = 18,316$ przy 2270 stopniach swobody i *p-value* <0,0001. Wskazuje to na konieczność odrzucenia hipotezy zerowej oraz wyciągnięcie wniosku, że zaobserwowana różnica ciśnienia skurczowego 8,3 mm Hg jest istotna statystycznie i że dieta o dużej zawartości soli kuchennej istotnie podnosi wartość ciśnienia skurczowego krwi.

Spróbujmy teraz zrobić dodatkowy eksperyment. Załóżmy, że analogiczne wyniki średnich i odchyleń standardowych uzyskaliśmy, badając stukrotnie mniejsze grupy pacjentów, tj. odpowiednio 12 i 10 mężczyzn.

Tym razem przeprowadzony test Fishera–Snedecora nie wykaże istotnych różnic między wariancjami (F = 1,138, *p-value* = 0,4301), możemy zatem do porównywania średnich wartości ciśnienia skurczowego użyć zwykłego testu t-Studenta dla zmiennych nie powiązanych. Proszę zwrócić uwagę, że w obu układach wartość statystyki Fishera-Snedecora jest taka sama (F = 1,138), a różnica *p-value* wynika ze znacznie zmniejszonej liczebności grup.

Użycie testu t-Studenta daje wartość statystyki t = 1,780 przy 20 stopniach swobody i *p-value* = 0,0903.Oznacza to, że przy tak silnie zmniej-

szonej liczebności grup ta sama różnica ciśnień skurczowych (8,3 mm Hg) nie jest już istotna statystycznie, a jej źródłem nie jest rodzaj stosowanej diety, lecz błąd próbkowania.

Ten prosty przykład pokazuje nam, jak silny wpływ na wnioskowanie statystyczne ma liczebność analizowanej grupy, o czym zresztą wspominaliśmy, omawiając pojęcie próby reprezentatywnej.

Test t-Studenta dla zmiennych powiązanych (*t-Student test for paired data*)

Warunki stosowania testu t-Studenta dla zmiennych powiązanych są identyczne jak testu t-Studenta dla zmiennych nie powiązanych z jedną tylko różnicą, która dotyczy modelu prowadzenia badania i jest uwidoczniona w samej nazwie testu. Liczba stopni swobody wynosi w tej wersji testu $N - 1$, gdzie N to liczba par porównywanych danych (pamiętajmy, że model zmiennych powiązanych wymaga sparowanych danych, co oznacza, że liczba pomiarów w obu porównywanych grupach jest identyczna).

Przykład 12

W klinicznym badaniu pediatrycznym dokonano próby oceny wpływu podania aspiryny na obniżenie temperatury ciała. W grupie dwunastu pięcioletnich dziewczynek chorych na grypę zmierzono temperaturę bezpośrednio przed podaniem określonej dawki aspiryny oraz godzinę po jej podaniu. Otrzymane wyniki zestawiono w tabl. 5.

Użyto interwałowej skali pomiarowej. Porównano dwie grupy wyników zebranych w modelu zmiennych powiązanych. Do analizy możemy zatem zastosować test t-Studenta dla zmiennych powiązanych.

Tablica 5. Dane dotyczące wpływu podania aspiryny na obniżenie temperatury ciała

Nr dziecka	Temperatura ciała przed podaniem aspiryny	Temperatura ciała po podaniu aspiryny
1	39,1	37,6
2	39,6	37,8
3	38,8	37,9
4	39,4	38,4
5	38,4	37,7
6	38,2	37,9
7	39,2	38,3
8	39,5	37,8
9	39,3	38,2
10	39,0	38,4
11	38,8	38,5
12	38,6	37,5

W wyniku obliczeń otrzymujemy następujące parametry statystyki opisowej:

temperatura przed podaniem aspiryny: $39{,}0 \pm 0{,}4^{\circ}C$

temperatura po podaniu aspiryny: $38{,}0 \pm 0{,}3^{\circ}C$

Musimy stwierdzić, czy zaobserwowana różnica temperatur jest wywołana rzeczywiście podaniem leku, czy też wynika z błędu próbkowania. W tym celu konstruujemy hipotezę zerową ($H_0 : \bar{x} = \bar{y}$; różnica temperatur nie jest istotna statystycznie) oraz hipotezę alternatywną ($H_1 : \bar{x} \neq \bar{y}$; różnica temperatur jest istotna statystycznie). Przyjmujemy poziom istotności statystycznej $\alpha = 0{,}05$. Otrzymana wartość statystyki t-Studenta wynosi 6,689 (11 stopni swobody), co daje wartość prawdopodobieństwa popełnienia błędu pierwszego rodzaju *p-value* <0,0001. Możemy zatem odrzucić hipotezę zerową (czyniąc to ryzykujemy, że popełnimy pomyłkę co najwyżej w jednym przypadku na dziesięć tysięcy) i stwierdzić, że zaobserwowane obniżenie temperatury o $1{,}0^{\circ}$ C nie wynika z wariancji próbkowania, lecz z rzeczywistego działania leku.

Test Manna-Whitneya (*Mann-Whitney test*)

Test Manna-Whitneya znany również pod nazwą testu Wilcoxona dla sumy rang (*Wilcoxon rank sum test*) jest stosowany dla porównania dwóch grup danych zebranych według modelu zmiennych nie powiązanych, gdy pomiarów dokonano:

– w skali interwałowej, lecz istnieje silne odstępstwo rozkładu tych danych od rozkładu normalnego (nie dające się na przykład usunąć przez zastosowanie odpowiedniej transformacji danych);

– w skali porządkowej.

W pierwszym przypadku można nadal hipotezę zerową formułować jako brak istotnej różnicy średnich arytmetycznych (bo takowe w skali interwałowej można obliczyć), w drugim zaś hipoteza zerowa będzie zakładać, że badane próby pochodzą z tych samych populacji (rozkłady danych w obu grupach nie różnią się istotnie statystycznie) – bez jawnego odwoływania się do pojęcia średniej (bo w przypadku skali porządkowej nie można jej oczywiście obliczać).

Często jest zadawane pytanie: czy błędem będzie zastosowanie testu Manna-Whitneya w skali interwałowej, gdy dane pochodzą z populacji o rozkładzie normalym? Oczywiście błędem nie jest, jednak test Manna-Whitneya jest testem słabszym od testu t-Studenta dla zmiennych nie powiązanych. Moc testu Manna-Whitneya wynosi około 95% mocy testu t-Studenta dla zmiennych nie powiązanych. Widać zatem, że test Manna-Whitneya będzie nieco bardziej konserwatywny (będzie preferował utrzymanie hipotezy zerowej), czyli w większej liczbie przypadków badanie statystyczne nie przyniesie oczekiwanego rozstrzygnięcia. Podobna uwaga dotyczy zastosowania testu Wilcoxona w miejsce testu t-Studenta dla zmiennych powiązanych w przypadku, gdy dane pochodzą z populacji o rozkładzie normalnym.

Przykład 13

Badania mające na celu ocenę efektywności leczenia litem w grupie pacjentów maniakalno-depresyjnych, przeprowadzano w dwóch różnych ośrodkach. Przed stosowaniem terapii litowej każdy z pacjentów otrzymał do wypełnienia kwestionariusz; miał w nim dokonać oceny swojego stanu psychicznego w czterostopniowej skali: 1 – czuję się dobrze, 2 – zazwyczaj czuję się dobrze, rzadko bywam nerwowy, 3 – okres dobrego samopoczucia i okres wzmożonej nerwowości jest taki sam, 4 – zawsze jestem silnie

pobudzony. Poniżej zestawiono dane uzyskane od pacjentów w każdym z ośrodków psychiatrycznych.

Ośrodek 1.	3, 4, 1, 1, 3, 2, 3, 4, 4, 3, 2, 4, 4, 4
Ośrodek 2.	1, 2, 1, 3, 2, 4, 1, 2, 1, 3, 1, 2, 2 , 2, 1, 3

Chcemy ocenić, czy stan początkowy pacjentów w obu ośrodkach jest jednakowy. Wyniki zostały zebrane w typowej skali porządkowej. Łatwo jest określić kierunek wzrostu natężenia badanej cechy, nie uda nam się natomiast ocenić odległości. Do porównania mamy dwie grupy wyników zebranych w modelu zmiennych nie powiązanych. Informacje te wskazują na konieczność zastosowania testu nieparametrycznego Manna–Whitneya.

Każdy pakiet statystyczny, do którego wprowadzimy nasze dane, potraktuje je jako dane liczbowe (skalę interwałową) i oczywiście obliczy m.in. takie parametry, jak średnie arytmetyczne i odchylenia standardowe. Dla naszego zestawu danych otrzymalibyśmy dla ośrodka pierwszego wielkości 3,0000 ± 1,1090, dla drugiego zaś 1,9374 ± 0,9287. Pamiętamy, że wartości te nie mają jednak żadnego sensu. W skali porządkowej nie mamy określonych odległości między punktami, więc uzyskanych wyników nie powinniśmy w żadnym razie przytaczać. Prawidłową miarą tendecji centralnej jest w naszym przypadku mediana. Wynosi ona dla obu ośrodków odpowiednio 3 oraz 2. Musimy teraz podjąć decyzję: czy różnica obu median wynika z błędu próbkowania, czy też jest wywołana istotnie różnym stanem wyjściowym pacjentów w obu ośrodkach. Budujemy zatem parę hipotez o następującej postaci:

H_0 : mediana 1 = mediana 2; (mediany nie różnią się istotnie statystycznie, wyjściowy stan pacjentów w obu ośrodkach jest jednakowy)

H_1 : mediana 1 ≠ mediana 2; (mediany różnią się istotnie statystycznie, wyjściowy stan pacjentów w obu ośrodkach jest różny)

Użycie testu Manna-Whitneya da nam wartość statystyki równą 53 oraz *p-valu*e równe 0,01455. Widzimy zatem, że należy odrzucić hipotezę zerową i stwierdzić, że stan wyjściowy pacjentów przed rozpoczęciem terapii litowej był w obu ośrodkach istotnie różny („średni" stan pacjenta w ośrodku drugim jest lepszy niż w ośrodku pierwszym).

Test Wilcoxona (*Wilcoxon's test*)

W pakietach statystycznych test ten jest często nazywany testem Wilcoxona dla znakowanych rang (*Wilcoxon's signed rank test*). Uwagi dotyczące zakresu jego stosowania są identyczne, jak w przypadku testu Manna–Whitneya, ale dane powinny zostać zebrane według modelu zmiennych powiązanych.

Przykład 14

W przykładzie 12. przyjęliśmy ad hoc, że dane w obu analizowanych grupach nie mają rozkładu istotnie różniącego się od rozkładu normalnego. Spróbujmy uzyskane tam wyniki zweryfikować testem, który nie zakłada istnienia normalności rozkładu. Pozostałe warunki zadania pozostają identyczne – takie, jak w przykładzie 12.

W wyniku zastosowania testu Wilcoxona otrzymamy wartość statystyki $W = 78{,}0$, co daje prawdopodobieństwo popełnienia błędu pierwszego rodzaju *p-value* = 0,0005. Przy przyjętym *a priori* poziomie $\alpha = 0{,}05$ możemy odrzucić hipotezę zerową o nieistotnej różnicy średnich i stwierdzić, że zaobserwowany spadek temperatury o jeden stopień Celsjusza został wywołany zaaplikowaniem dzieciom aspiryny. Przyjmując hipotezę alternatywną, popełniamy błąd w co najwyżej pięciu przypadkach na dziesięć tysięcy. Proszę zwrócić uwagę, że test Wilcoxona jest nieco bardziej konserwatywny niż test t-Studenta (wskazuje na to *p-value* równa odpowiednio dla tych testów 0,0005 i 0,0001).

Test chi-kwadrat (*Chi-square test*) i dokładny test Fishera (*Fisher's exact test*)

Test chi-kwadrat przeznaczony jest dla danych zebranych w skali nominalnej według schematu zmiennych nie powiązanych. Należy zaznaczyć, że może on służyć zarówno do porównywania dwóch, jak i większej liczby grup. Dane są zestawiane w postaci **tablic kontyngencji** (*contingency tables*), które w zależności od liczby porównywanych grup mogą mieć rozmaite wymiary. Tablica kontyngencji podaje, ile razy zachodzi określona koincydencja wartości wyników pomiarów obu porównywanych zmiennych. Na przykład element leżący w pierwszym wierszu i pierwszej kolumnie tablicy kontyngencji pokazuje, w ilu obserwacjach miała miejsce taka sytuacja, że pierwsza z mierzonych zmiennych miała wartość wymienioną na liście jej wartości na pierwszym miejscu i równocześnie druga zmienna miała także wartość wymienioną z kolei na liście jej wartości na pierwszym miejscu. Liczba stopni swobody dla testu chi-kwadrat jest równa $(w - 1) \times (k - 1)$, gdzie w oznacza liczbę wierszy tablicy kontyngencji (czyli liczbę możliwych wartości pierwszej zmiennej), a k – liczbę kolumn (czyli liczbę możliwych wartości drugiej zmiennej).

Przykład 15

W grupie 79 pacjentów chorych na tę samą chorobę zastosowano dwa różne schematy leczenia. Wyniki terapii zostały przedstawione w poniższej tablicy kontyngencji o wymiarze 2 x 2 (pierwsza zmienna przyjmuje wartości: *przeżył, zmarł*; druga zmienna przyjmuje wartości: *leczenie metodą 1, leczenie metodą 2*. Tablica jest często nazywana **tablicą wartości obserwowanych** (*observed values*). Chcemy odpowiedzieć na pytanie, czy obie metody leczenia dają taki sam efekt końcowy.

Sposób leczenia / Wynik	Przeżyło	Zmarło	Razem
Metoda 1	20	12	32
Metoda 2	32	15	47
Razem	52	27	79

Procent przeżycia po zastosowaniu pierwszej metody wynosi 20/32 × 100% = 62,5%, w drugiej zaś 32/47 × 100% = 68,09%. Na pozór wydaje

się więc, że druga metoda jest lepsza. Czy jednak około pięcioprocentowa różnica rzeczywiście przemawia na korzyść stosowania drugiej metody leczenia, czy też należy ją traktować jako błąd wynikający ze zmienności próbkowania? Nasza hipoteza zerowa przyjmie więc postać: „uzyskane wyniki pochodzą z tej samej populacji". Innymi słowy stawiamy hipotezę, że nie ma istotnej statystycznie różnicy w przeżyciu pacjenta, niezależnie czy leczymy go pierwszą, czy też drugą metodą.

Podstawą wnioskowania w teście chi-kwadrat jest porównanie rzeczywistych (obserwowanych) wartości kontyngencji z **wartościami oczekiwanymi** (*expected values*). Wartości oczekiwane w każdej **komórce** (*cell*) tablicy oblicza się na podstawie tablicy kontyngencji z danymi eksperymentalnymi w bardzo prosty sposób – mnożymy odpowiadające komórce sumy brzegowe przez siebie i iloczyn dzielimy przez liczbę przypadków.

Sposób leczenia / Wynik	Przeżyło	Zmarło
Metoda 1	$52 \times 32 / 79 = 21{,}06$	$27 \times 32 / 79 = 10{,}94$
Metoda 2	$52 \times 47 / 79 = 30{,}94$	$27 \times 47 / 79 = 16{,}06$

Warunkiem poprawności stosowania testu chi-kwadrat jest, aby wartość oczekiwana pomiarów w każdej komórce tablicy kontyngencji była nie mniejsza niż pięć. Jak widać, w rozważanym przykładzie żadna z wartości oczekiwanych w tablicy kontyngencji nie jest mniejsza od pięciu, możemy zatem użyć testu chi-kwadrat. Gdyby choć jedna komórka w tablicy wartości oczekiwanych miała wartość mniejszą lub równą pięć, do testowania hipotez należałoby użyć dokładnego testu Fishera. Niektórzy statystycy wprowadzają uproszczone kryterium – liczebność w żadnej z komórek tablicy wartości obserwowanych nie może być mniejsza od sześciu. Uzyskiwana metodą testu chi-kwadrat *p-value* jest wartością przybliżoną, zbliżającą się asymptotycznie do wartości prawdziwej w przypadku bardzo dużych prób. Obliczona tą metodą *p-value* jest systematycznie zaniżana. W celu uniknięcia tego błędu stosuje się **poprawkę Yatesa** (*Yates' correction*) zwiększającą zachowawczość testu chi-kwadrat. Pamiętajmy, że we wszystkich trzech sytuacjach (test chi-kwadrat bez poprawki Yatesa, z poprawką, dokładny test Fishera) sformułowanie hipotezy zerowej jest takie samo: nie obserwuje się zależności między sposobem leczenia a jego efektem, tzn. procent pacjentów, których leczenie zostało zakończone sukcesem nie różni się istotnie

statystycznie w obu metodach leczenia. Hipoteza alternatywna stwierdza statystycznie istotną zależność między metodą leczenia a jej efektem.

Dla danych z naszego przykładu otrzymujemy następujące wyniki:

Wartość statystyki chi-kwadrat bez poprawki Yates'a **chi** $\chi^2 = 0{,}2640$ (1 stopień swobody) – odpowiadająca *p-value* = 0,6074

Wartość statystyki chi-kwadrat poprawką Yatesa **chi** $\chi^2 = 0{,}0741$ (1 stopień swobody) – odpowiadająca *p-value* = 0,7855

Dokładna *p-value* otrzymana testem Fishera wynosi 0,6360

Jak widzimy, wszystkie trzy metody doprowadziły nas do tego samego stwierdzenia, że nie mamy podstaw do odrzucenia hipotezy zerowej. Obie metody leczenia są więc jednakowo skuteczne, a zaobserwowana różnica procentowa jest spowodowana błędem próbkowania. Test Fishera wyznacza nam zawsze dokładną *p-value*, zaś test chi-kwadrat lepszą lub gorszą jej aproksymację.

Jak już wspomnieliśmy test chi-kwadrat może służyć do porównywania większej liczby grup (gdy zastosowaliśmy więcej niż dwie różne metody leczenia albo wynik leczenia sklasyfikowaliśmy na więcej niż dwa poziomy, na przykład *poprawa, pogorszenie, zgon*). W takiej sytuacji postępujemy identycznie, jak w przypadku porównywania dwóch grup. Jedyny problem może stanowić sytuacja, gdy wartość oczekiwana w jednej z komórek tablicy kontyngencji będzie równa pięć lub mniejsza. Tutaj także właściwe postępowanie polega na skorzystaniu z dokładnego testu Fishera. O ile dokładny test Fishera dla tablic 2×2 znajduje się już niemal w każdym współczesnym pakiecie oprogramowania statystycznego, to z uwagi na wielką złożoność algorytmu obliczeniowego (np. algorytm Mehty i Patela) dla tablic większych niż 2×2 powoduje jego słabe rozpowszechnienie (można go np. znaleźć w pakiecie StatXact Turbo firmy CYTEL Co). Gdy nie mamy możliwości skorzystania z tego testu, często skuteczną metodą uzyskania potrzebnej oceny z użyciem klasycznego testu chi-kwadrat jest zwiększanie liczebności krytycznych komórek przez łączenie niektórych wierszy lub kolumn tablicy kontyngencji (oczywiście, jeżeli ma to sens merytoryczny) i zmodyfikowanie w ten sposób wartości oczekiwanych.

Test znaków (*sign test*) i test McNemary (*McNemar's test*)

Oba testy wymienione w tytule podrozdziału są używane do porównań dwóch grup pomiarów wykonanych w skali nominalnej według modelu zmiennych powiązanych. Wybór, który z tych dwóch testów należy zastosować w danej sytuacji badawczej, zależy od liczebności grup. Gdy liczba par pomiarów jest mniejsza od 20, korzystniej jest wykorzystywać test znaków. W przeciwnym razie stosujemy test McNemary, który jest pewną modyfikacją testu chi-kwadrat uwzględniającą powiązanie prób.

Przykład 16

W badaniu chcemy porównać zgodność dwóch metod w wykrywaniu guza sutka: metody BAC (biopsja aspiracyjna cienkoigłowa) oraz radiologicznej metody mammograficznej. U każdej z 20 pacjentek wykonano oba badania, a ich wyniki zestawiono w tabl. 6. Znakiem „+" oznaczyliśmy wykrycie nowotworu.

Tablica 6. Wyniki biopsji aspiracyjnej cienkoigłowej i mammografii w grupie 20 pacjentek

Nr pacjentki	BAC	Mammografia
1	–	–
2	–	–
3	+	–
4	+	+
5	–	–
6	+	–
7	–	–
8	+	+
9	+	+
10	–	–
11	+	–
12	+	–
13	–	–
14	+	–
15	–	+
16	+	–
17	+	–
18	–	–
19	–	–
20	–	–

Jak wynika z treści zadania mamy do czynienia z porównaniem dwóch grup wyników zebranych w modelu zmiennych powiązanych (każda pacjentka miała wykonane oba badania), a wyniki pomiarów przedstawiono w skali nominalnej. W celu uzyskania odpowiedzi na pytanie, czy obie metody dają zgodne wyniki, musimy użyć testu McNemary. Przed przystąpieniem do obliczeń trzeba jednak przekonstruować tablicę wyników do postaci:

Metoda	Mammografia		
BAC	Wynik	+	–
	+	3	7
	–	1	9

Hipoteza zerowa zakłada, że uzyskane obiema metodami wyniki nie różnią się istotnie statystycznie. Hipoteza alternatywna stwierdza istnienie różnic diagnostycznych*. Użycie testu McNemary z poprawką Yatesa daje nam wartość statystyki W = 3,125 i *p-value* = 0,0771. Wynik nie daje nam podstaw do odrzucenia hipotezy zerowej, stwierdzamy zatem, że obie techniki dają zgodne rezultaty, a zaobserwowane rozbieżności wynikają wyłącznie z fluktuacji statystycznych.

* Proszę zwrócić uwagę, że nie badamy tutaj, która z metod jest lepsza lub gorsza, lecz jedynie zgodność uzyskanych wyników. Jeżeli natomiast uznamy metodę BAC jako wzorzec prawidłowej diagnozy, to możemy spróbować odpowiedzieć na pytanie, czy mammografia jest równie skuteczną metodą wykrywania guza sutka jak punkcja cienkoigłowa.

Analiza wariancji (*analysis of variance*)

Test analizy wariancji służy do dokonywania porównań wielu grup pomiarów. Mogłoby się wydawać, że taki test jest zbyteczny, ponieważ jeżeli chce się porównać ze sobą więcej niż dwie grupy pomiarów, należy po prostu wielokrotnie zastosować testy, których używaliśmy do tej pory do porównań dwóch grup między sobą. Takie postawienie sprawy niesie jednak ze sobą bardzo poważne niebezpieczeństwo. Duża liczba porównań może doprowadzić do pojawienia się przypadkowych (losowych) różnic nie wynikających ze stanu rzeczywistego. Zjawisko to omówiliśmy dokładnie w rozdziale „Pojęcia podstawowe". Testy zawarte w tabl. 3 przeznaczone do porównań większej liczby grup zawierają specjalne czynniki korekcyjne, zmniejszające ryzyko wystąpienia fałszywych wyników.

Metody analizy wariancji służą właśnie do porównywania wielu grup pomiarów, przy czym (wbrew swojej nazwie) są wykorzystywane do testowania różnic średnich arytmetycznych w wielu grupach. Istnieje wiele modeli analizy wariancji dedykowanych bardzo specyficznym układom porównywanych danych. Modele te zależą od liczby czynników i liczby poziomów poszczególnych czynników, a także od tego, czy mają strukturę prostą, albo też użyte są pomiary powtarzane, czy użyty model analizy wariancji opiera się na technice efektów stałych, losowych lub mieszanych itp. Pełna dyskusja metod analizy wariancji jest zbyt obszerna i skomplikowana, by próbować ją tutaj streszczać. Poprzestaniemy więc tylko na kilku podstawowych informacjach. Znakomity, pełny przegląd metod analizy wariancji znajdzie Czytelnik w pracy J. Brzezińskiego i R. Stachowskiego [3].

Scharakteryzujemy teraz krótko istotę „działania" tej grupy testów statystycznych, a w tabl. 7 przedstawimy kilka przykładowych układów danych, które można analizować za pomocą analizy wariancji.

Każdy model analizy wariancji zakłada, że pomiary wykonane zostały w skali interwałowej. Bardzo ważnym założeniem jest normalność rozkładu danych w badanych populacjach, gdyż używany w analizie wariancji test Fishera-Snedecora jest znacznie bardziej czuły na odstępstwa od normalności niż test t-Studenta. W przypadku rozkładu innego niż normalny należy albo dokonać próby przetransformowania danych, albo użyć odpowiednich testów z klasy technik statystycznych niezależnych od rozkładu (*distribution-free tests*), takich jak test Kruskala-Wallisa lub test Friedmana. Oprócz normalności rozkładu niektóre modele wymagają spełnienia dodatkowych założeń, takich jak jednorodność wariancji, symetria macierzy wariancji

Tablica 7. Przykładowe modele analizy wariancji

Cytostatyk I	**Cytostatyk II**	Cytostatyk III
---- ---- ----	---- ---- ----	---- ---- ----

a. Układ jednoczynnikowy bez powtarzanych pomiarów (*one-way ANOVA without repeated measures*) – porównanie poziomu białych ciałek krwi w trzech grupach pacjentów leczonych różnymi cytostatykami (I, II, III).

Cytostatyk / **Płeć**	**I**	**II**	**III**
Mężczyźni	----	----	----
Kobiety	----	----	----

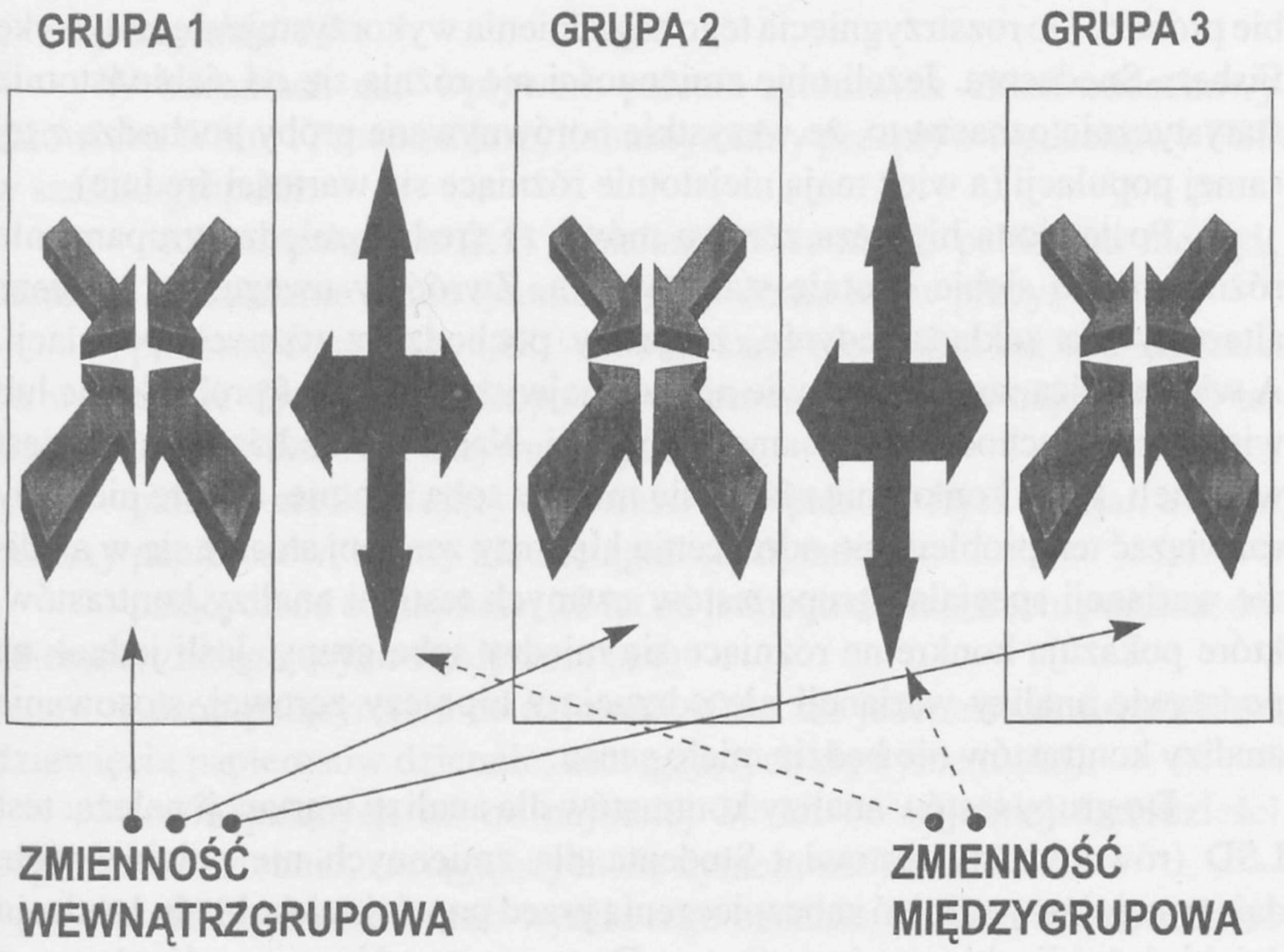
b. Układ dwuczynnikowy bez powtarzanych pomiarów (*two-way ANOVA without repeated measures*) – porównanie poziomu białych ciałek krwi w trzech grupach pacjentów leczonych różnymi cytostatykami ze zróżnicowaniem na płeć.

Cytostatyk I		**Cytostatyk II**		**Cytostatyk III**	
przed	po	przed	po	przed	po
---- ----	---- ----	---- ----	---- ----	---- ----	---- ----

c. Układ jednoczynnikowy z powtarzanymi pomiarami (*one-way ANOVA with repeated measures*) – porównanie poziomu białych ciałek krwi w trzech grupach pacjentów leczonych różnymi cytostatykami z uwzględnieniem poziomu leukocytów przed podaniem cytostatyku i po jego podaniu.

Cytostatyk / **Płeć**	**I**		**II**		**III**	
	przed	po	przed	po	przed	po
Mężczyźni	---- ----	---- ----	---- ----	---- ----	---- ----	---- ----
Kobiety	---- ----	---- ----	---- ----	---- ----	---- ----	---- ----

d. Układ dwuczynnikowy z powtarzanymi pomiarami (*two-way ANOVA with repeated measures*) – porównanie poziomu białych ciałek krwi w trzech grupach pacjentów leczonych różnymi cytostatykami z uwzględnieniem płci oraz poziomu leukocytów przed podaniem cytostatyku i po jego podaniu.



Ryc. 18. Idea zmienności międzygrupowej i wewnątrzgrupowej w analizie wariancji

i kowariancji, równość macierzy wariancji-kowariancji itp. Jeśli są spełnione odpowiednie warunki, analiza wariancji może służyć zarówno do badania danych zebranych w schemacie zmiennych powiązanych, jak i nie powiązanych.

Wyobraźmy sobie następującą sytuację. Mamy porównać ze sobą kilka grup wyników pomiarów. Na razie nie precyzujemy, według jakiego schematu dane te zostały zebrane, lecz oznaczamy każdy z pomiarów ogólnie symbolem x_{ij} gdzie indeks i oznacza numer grupy, zaś j numer pomiaru w grupie. Analiza wariancji oszacowuje średnie $\bar{x}_i$ dla każdej z grup oraz średnią centralną $\bar{x}$ (wartość średnia po połączeniu pomiarów z wszystkich grup w jeden duży zbiór).

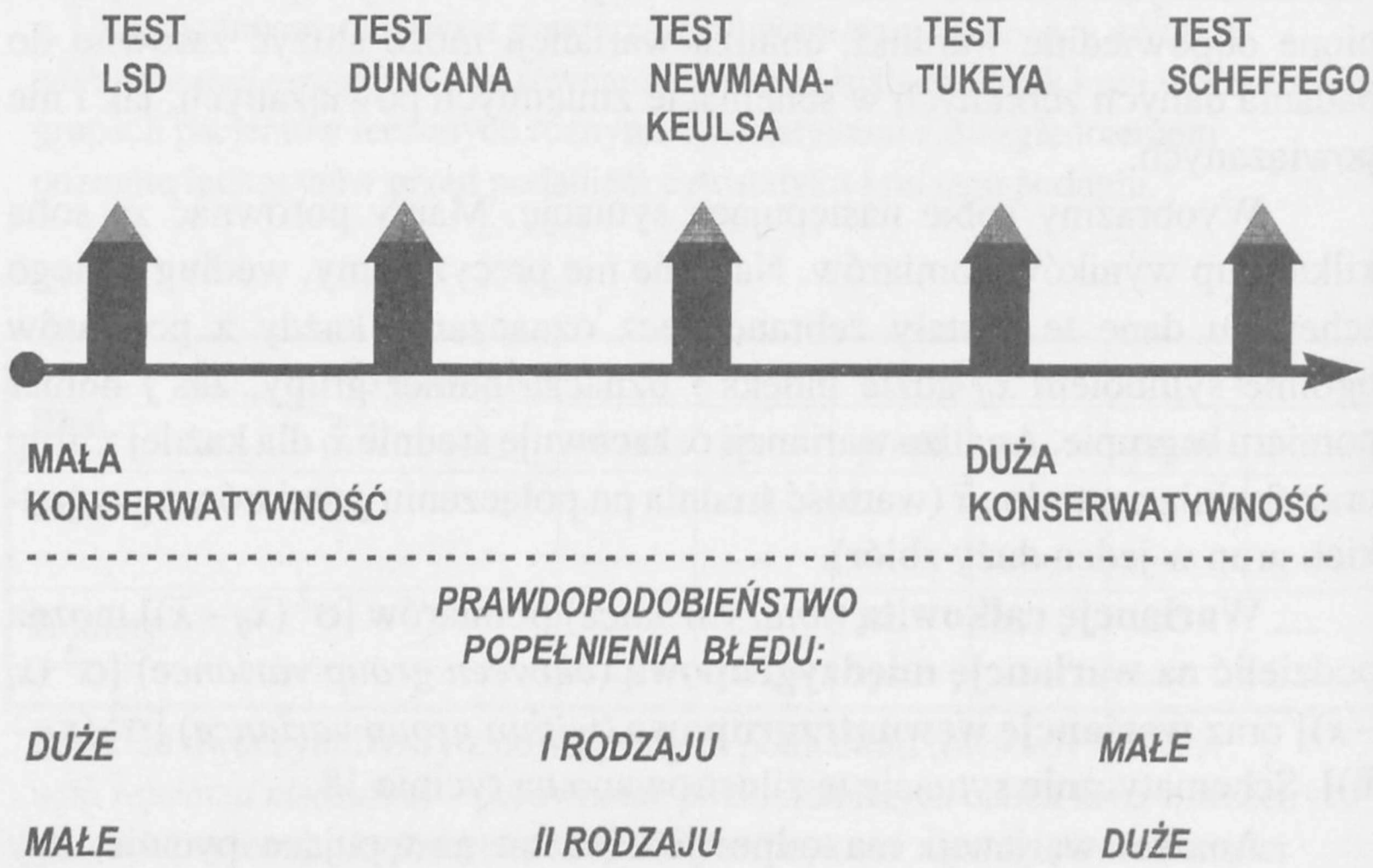
Wariancję całkowitą (*total variance*) pomiarów [$\sigma^2 (\bar{x}_{ij} - \bar{x})$] można podzielić na **wariancję międzygrupową** (*between group variance*) [$\sigma^2 (x_i - x)$] oraz **wariancję wewnątrzgrupową** (*within group variance*) [$\sigma^2 (x_{ij} - \bar{x}_i)$]. Schematycznie sytuację tę zilustrowano na rycinie 18.

Analiza wariancji ma odpowiedzieć na następujące pytanie: czy zmienność pomiarów między grupami jest taka sama, jak zmienność w obrę-

bie próbek? Do rozstrzygnięcia tego zagadnienia wykorzystuje się statystykę Fishera-Snedecora. Jeżeli obie zmienności nie różnią się od siebie istotnie statystycznie oznacza to, że wszystkie porównywane próby pochodzą z tej samej populacji (a więc mają nieistotnie różniące się wartości średnie).

Postawiona hipoteza zerowa mówi, że średnie między grupami nie różnią się od siebie istotnie statystycznie. Zwróćmy uwagę, że hipoteza alternatywna zakłada jedynie, że próby pochodzą z różnych populacji. A więc analiza wariancji powie nam co najwyżej, że jakieś próby (dwie lub więcej) nie pochodzą z tej samej populacji. Nadal nie będziemy natomiast wiedzieli, które konkretnie różnią się między sobą istotnie, a które nie. Aby rozwiązać ten problem, po odrzuceniu hipotezy zerowej stosuje się w analizie wariancji specjalną grupę testów zwanych testami analizy kontrastów, które pokazują konkretne różniące się między sobą grupy. Jeśli jednak na podstawie analizy wariancji nie odrzucimy hipotezy zerowej, stosowanie analizy kontrastów nie będzie miało sensu.

Do grupy testów analizy kontrastów dla analizy wariancji należą: test LSD (równoważny testowi t-Studenta dla zmiennych nie powiązanych, dający najniższy stopień zabezpieczenia przed popełnieniem błędu I rodzaju dzięki dużej liczbie porównań), test Duncana, test Newmana-Keulsa, test Tukeya oraz test Scheffego. Wymienione testy ustawiono w kierunku wzrostu ich konserwatywności. Porównanie prawdopodobieństwa popełnienia błędów I i II rodzaju dla każdego z testów analizy kontrastów przedstawiono na rycinie 19.



Ryc. 19. Stopień konserwatywności wybranych testów analizy kontrastów

Przykład 17

W badaniach nad wpływem palenia tytoniu na układ oddechowy człowieka White i Froeb zmierzyli maksymalny przepływ wydechowy FEF w sześciu grupach:

– osób niepalących (*A1*) – badani ani sami nie palili, ani nie byli narażeni na działanie dymu tytoniowego w domu lub w pracy;

– pasywnych palaczy (*A2*) – badani sami nie palili i nie byli narażeni na wdychanie dymu papierosowego w domu, lecz od co najmniej 20 lat pracowali w środowisku aktywnych palaczy;

– palaczy nie inhalujących dymu (*A3*) – palaczy fajek i cygar, a także palaczy papierosów, którzy nie zaciągali się dymem;

– palących od co najmniej 20 lat od jednego do dziesięciu papierosów dziennie, zaciągających się dymem (*A4*)

– osób palących od co najmniej 20 lat od jedenastu do trzydziestu dziewięciu papierosów dziennie, zaciągających się dymem (*A5*);

– osób palących od co najmniej 20 lat co najmniej czterdzieści papierosów dziennie, zaciągających się dymem (*A6*).

W tablicy 8 zestawiono wyniki tego badania. Chcemy odpowiedzieć na pytanie, czy wartości maksymalnego przepływu wydechowego różnią się istotnie statystycznie między badanymi grupami pacjentów.

Tablica 8. Wartości średnie i odchylenia standardowe maksymalnego przepływu wydechowego w sześciu grupach pacjentów

Grupa	Średnia FEF [l/s]	Odch. stand. FEF [l/s]	Liczebność grupy
A1	3,78	0,79	200
A2	3,30	0,77	200
A3	3,32	0,86	50
A4	3,23	0,78	200
A5	2,73	0,81	200
A6	2,59	0,82	200

Użyta w badaniu skala interwałowa oraz sześć grup badawczych w układzie zmiennych niepowiązanych wskazują na konieczność użycia prostego modelu analizy wariancji. Obok założenia, że dane zostały pobrane z populacji o rozkładzie nieistotnie różniącym się od rozkładu normalnego, konieczne jest sprawdzenie, czy wariancje wyników są homogeniczne (nie różnią się od siebie istotnie statystycznie).

Użyjemy do tego celu testu Bartletta, stawiając następującą hipotezę zerową:

$H_0 : SD^2_{A1} = SD^2_{A2} = SD^2_{A3} = SD^2_{A4} = SD^2_{A5} = SD^2_{A6}$

oraz hipotezę alternatywną

$H_1 : SD^2_{A1} \neq SD^2_{A2} \neq SD^2_{A3} \neq SD^2_{A4}$

Otrzymujemy wartość statystyki Bartletta B=1,7 i odpowiadającą jej *p-value* = 0,8889, co wskazuje na brak podstaw do odrzucenia hipotezy zerowej. Możemy zatem przystąpić do drugiej fazy badania – sprawdzenia za pomocą testu analizy wariancji, czy średnie arytmetyczne różnią się od siebie istotnie statystycznie, czy też obserwowane różnice są fluktuacjami losowymi. Konstruujemy jak zwykle parę hipotez:

$H_0 : \bar{x}_{A1} = \bar{x}_{A2} = \bar{x}_{A3} = \bar{x}_{A4} = \bar{x}_{A5} = \bar{x}_{A6}$

$H_1 : \bar{x}_{A1} \neq \bar{x}_{A2} \neq \bar{x}_{A3} \neq \bar{x}_{A4} \neq \bar{x}_{A5} \neq \bar{x}_{A6}$

Uzyskana wartość statystyki Fishera w teście analizy wariancji wynosi 57,990, natomiast *p-value* <0,0001 wskazuje na istnienie istotnych różnic między średnimi. Proszę zwrócić uwagę, że nie wiemy jednak, czy wszystkie średnie różnią się od siebie istotnie statystycznie, czy też tylko niektóre pary. Aby to zbadać, użyjemy któregoś spośród testów wielokrotnych porównań, np. testu Newmana-Keulsa. Test ten wskazuje, że istotne różnice występują między wszystkimi kombinacjami par z wyjątkiem par: *A5–A6, A3–A4, A2–A3* oraz *A2–A4* (wszystkich możliwych porównań jest 6×5 / 2 = 15).

Test Kruskala-Wallisa (*Kruskal-Wallis test*)

Użycie testu Kruskala-Wallisa wymaga danych zebranych w kilku grupach według schematu zmiennych nie powiązanych. W zasadzie test jest przeznaczony dla pomiarów w skali porządkowej, lecz stosuje się go również wtedy, gdy dane w skali interwałowej nie spełniają jednego z podstawowych warunków stosowania analizy wariancji – normalności rozkładu. Hipoteza zerowa dla pomiarów w skali porządkowej brzmi: pomiary we wszystkich porównywanych grupach pochodzą z tej samej populacji. W skali interwałowej możemy dodać – średnie w porównywanych grupach nie różnią się od siebie istotnie statystycznie.

Podobnie jak w przypadku analizy wariancji, test Kruskala-Wallisa przy ewentualnym odrzuceniu hipotezy zerowej nie wskazuje konkretnie, które z grup różnią się między sobą istotnie. W tym przypadku również stosuje się testy analizy kontrastów, np. test Millera, Page'a lub Dunna. Testy te zawierają specjalne czynniki korekcyjne minimalizujące prawdopodobieństwo uzyskania przypadkowych różnic przy dużej liczbie porównań.

Przykład 18

W eksperymencie przeprowadzonym na królikach próbowano ocenić działanie przeciwzapalne czterech leków: indometacyny, aspiryny, piroxicamu i BW755C. W pierwszej fazie badania zwierzętom zakroplono do oka jednakową dawkę kwasu arachidonowego powodującego wystąpienie stanu zapalnego. Po dziesięciu minutach w każdej z czterech grup zaaplikowano odpowiedni lek i po dalszych 15 minutach zbadano efekt jego działania. Posłużono się przy tym następującą skalą: „0" – zupełne zniesienie efektu stanu zapalnego, oko całkowicie otwarte, „+++" – zupełny brak ustąpienia stanu zapalnego, oko całkowicie zamknięte. Symbole „+" oraz „++" oznaczały stany pośrednie między wymienionymi. Wyniki działania poszczególnych leków zebrano w tabl. 9.

Pomiaru dokonano w typowej skali porządkowej. Porównujemy wyniki działania czterech różnych leków, a więc mamy do czynienia z modelem zmiennych nie powiązanych. Te założenia wskazują na konieczność wyboru testu Kruskala-Wallisa. Po przyjęciu poziomu istotności statystycznej konstruujemy parę hipotez badawczych:

$H_0 : Me_1 = Me_2 = Me_3 = Me_4$

$H_1 : Me_1 \neq Me_2 \neq Me_3 \neq Me_4$

gdzie symbolem Me_i oznaczyliśmy medianę skuteczności działania i-tego leku.

Tablica 9. Wyniki uzyskane w eksperymencie badania skuteczności czterech wybranych leków w usuwaniu stanu zapalnego.

Indometacyna	Aspiryna	Piroxicam	BW755C
++	+	+++	+
+++	+++	+	0
+++	+	++	0
+++	++	+	++
+++	++	+++	0
0		+++	0
++			+

Zdecydowana większość pakietów statystycznych nie dopuszcza wprowadzania jako danych symboli nienumerycznych. W związku z tym dokonujemy prostej konwersji danych według następującego schematu: 0→0, + → 1, ++ → 2, +++ → 3. Proszę zwrócić uwagę, że zmiana oznaczeń nie powoduje ani dodania, ani też utraty informacji, użyte bowiem przez nas liczby 0, 1, 2, 3 przestają być liczbami – stają się zwykłymi symbolami. Nie mamy zatem odległości między punktami, ale wyłącznie kierunek wzrostu natężenia cechy mierzonej (stopnia skuteczności zniesienia stanu zapalnego). Dlatego pamiętajmy, że choć pakiet statystyczny poda nam średnie arytmetyczne, odchylenia standardowe i przedziały ufności w każdej z badanych grup, to nie mają one żadnego sensu i nie należy ich wykorzystywać. Z pewną ostrożnością należy również podchodzić do otrzymanej wartości mediany. Jakkolwiek ta miara tendencji centralnej jest przeznaczona dla skali porządkowej, to czasami jest źle wyznaczana przez pakiety statystyczne. Zdarza się to przy obliczaniu wariancji dla parzystej liczby pomiarów (błąd może się pojawić, ale nie musi*).

Przypomnijmy sobie sposób obliczania mediany: jest to wynik środkowego pomiaru w uporządkowanym nie malejąco szeregu danych. Jeśli liczba pomiarów jest nieparzysta, znalezienie pomiaru środkowego jest jednoznaczne. Problem pojawia się przy liczbie parzystej. Spójrzmy na dane dotyczące piroxicamu. Uporządkowany szereg tych pomiarów ma postać: 1, 1, 2, 3, 3, 3. Zatem pomiar środkowy leży pomiędzy pomiarem 2 i 3. Jeżeli dane byłyby wyrażone w skali interwałowej, to wartość mediany wynosiłaby

---

* Błędnie policzoną medianę możemy rozpoznać po tym, że przyjmuje ona wartość spoza zestawu zdefiniowanych na skali pomiarowej symboli. W naszym przykładzie dobrze zdefiniowanymi symbolami są 0, 1, 2, 3, a wyznaczona przez komputer w trzeciej grupie wartość mediany wynosi 2,5.

2,5 (obliczone jako (2+3)/2). I taki wynik daje nam program statystyczny, gdyż traktuje on dane jako szereg liczb, a nie symboli. My jednak wiemy, że odległość na naszej skali pomiarowej nie jest zdefiniowana. Nie istnieje żaden symbol postaci 2,5, wspomniana operacja nie ma więc sensu. Prawidłowa wartość mediany dla naszego szeregu danych wynosi 3. Pomiar środkowy jest bowiem otoczony przez symbole 2 i 3, a symbol 3 występuje częściej. W pozostałych grupach zawierających nieparzystą liczbę pomiarów nie mamy tego rodzaju problemów. Zatem wartości mediany dla analizowanych grup wynoszą odpowiednio: 3, 2, 3, 0.

Po tej wstępnej dyskusji możemy przejść do wykonania testu Kruskala-Wallisa. Obliczona wartość statystyki wynosi KW = 9,531, a odpowiadająca jej *p-value* 0,0230. Wskazuje to, że mamy podstawę do odrzucenia hipotezy zerowej (oczywiście jeśli przyjęliśmy przed rozpoczęciem obliczeń dopuszczalny poziom popełnienia błędu pierwszego rodzaju alfa = 0,05). Musimy zatem stwierdzić, które z median różnią się od siebie istotnie statystycznie. W tym celu wykonujemy test Dunna (jest to test wielokrotnych porównań dla testu Kruskala-Wallisa). W wyniku jego użycia stwierdzamy, że istotna statystycznie jest różnica między medianami 3 oraz 0, pozostałe kombinacje dają różnice nieistotne statystycznie (jest wysoce prawdopodobne, że są one wywołane błędem próbkowania, a nie rzeczywistą różnicą w sile działania leku). Ostatecznie możemy dać odpowiedź, że zdecydowanie najlepszy efekt znoszenia stanu zapalnego daje lek BW775C, pozostałe leki niestotnie różnią się od siebie w działaniu i dają słaby efekt przeciwzapalny.

Test Friedmana (*Friedman's test*)

Testu Friedmana używamy w przypadku danych zebranych w wielu grupach według schematu zmiennych powiązanych. Wszystkie pozostałe uwagi są identyczne, jak opisane w teście Kruskala-Wallisa. Oba bowiem testy należą do tak zwanej klasy **testów nieparametrycznych** (*nonparametric distribution-free tests*), czyli testów, których można używać niezależnie od typu rozkładu danych pomiarowych. W związku z tym ich zastosowanie wtedy, gdy dane pochodzą z populacji o rozkładzie normalnym nie jest błędem. Zmniejsza się jedynie nieznacznie moc testu (o mniej więcej 5%). W przypadku danych o rozkładzie normalnym testy nieparametryczne są nieco bardziej konserwatywne niż testy analizy wariancji. Oznacza to, że za ich pomocą nieco trudniej jest wykryć istniejącą w rzeczywistości różnicę średnich. Jako testy analizy kontrastów stosuje się najczęściej test Millera, Dunna i Page'a. Niestety tylko nieliczne pakiety statystyczne zawierają testy analizy kontrastów dla testu Kruskala-Wallisa i testu Friedmana. Wyczerpujący przegląd metod nieparametrycznych jest zawarty w pracy M. Hollandera i D.A. Wolfe'a [5].

Korelacja – badanie zależności

Wprowadzenie

Omówione w poprzednim rozdziale metody miały za zadanie udzielić odpowiedzi na pytanie: czy pomiary zawarte w dwóch grupach lub większej ich liczbie pochodzą z tej samej czy też z różnych populacji. **Metody korelacyjne** (*correlation methods*), które obecnie przedstawimy, służą do wykrycia istnienia ewentualnego związku między dwiema zmiennymi lub większą ich liczbą oraz oszacowania siły i istotności statystycznej tego związku.

Pamiętajmy, że istnienie związku między zmiennymi nie jest jednoznaczne z przyczynowością – mogą istnieć przypadkowe związki między zmiennymi i w żadnym razie nie wolno nam na ich podstawie wysnuwać wniosków, co do związków przyczynowych (patrz przykład w rozdziale „Pojęcia podstawowe"). Podobnie jak przy testowaniu hipotez związek między zmiennymi należy rozpatrywać w zależności od skali pomiarowej, w której zebrane zostały wyniki pomiarów. Jedynie w najsilniejszej skali – interwałowej – możemy mówić o kształcie związku, jego sile i istotności statystycznej. W skalach porządkowej i nominalnej możemy określić wyłącznie jego siłę i istotność statystyczną.

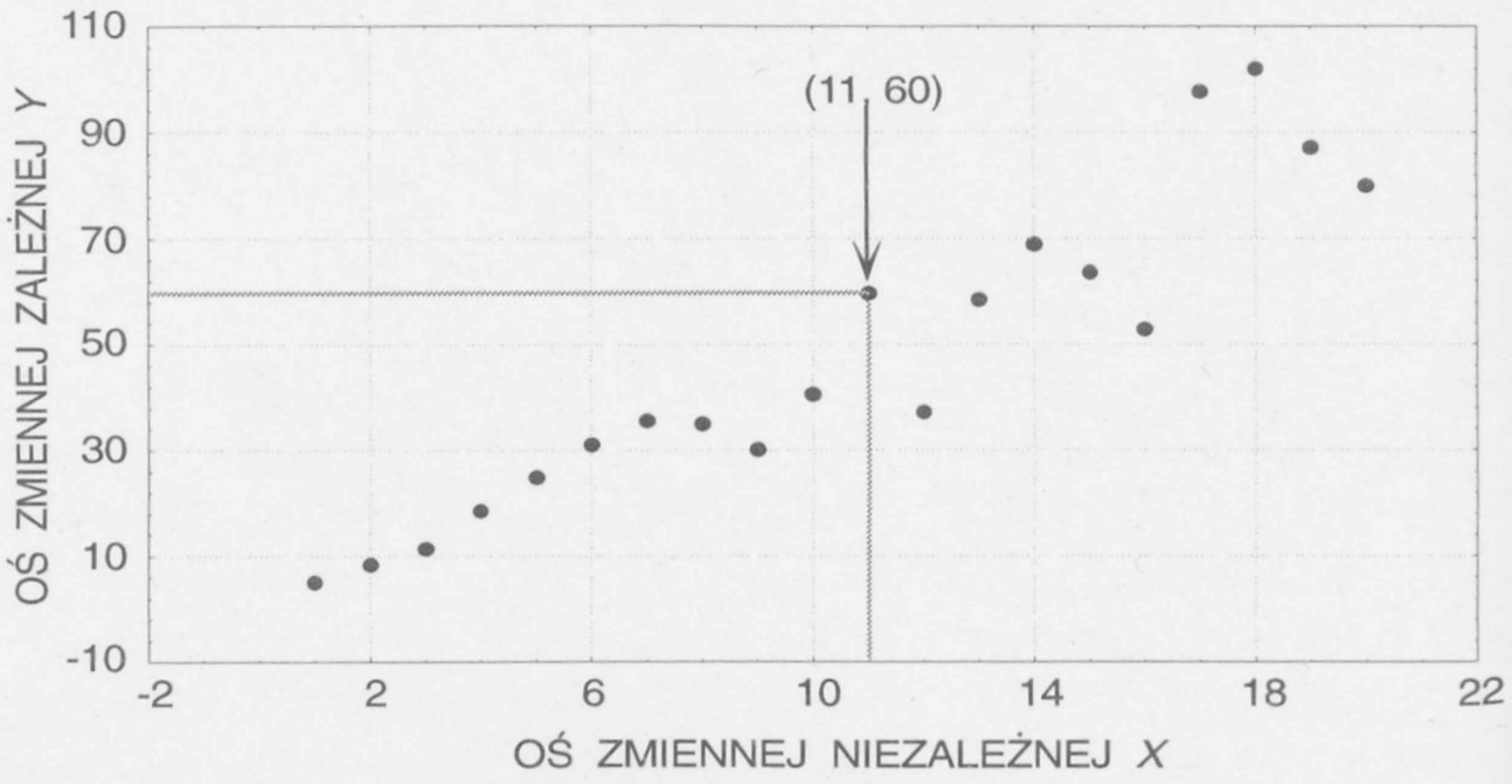
Współczynnik korelacji Pearsona

Zacznijmy omawianie metod korelacyjnych od skali interwałowej. Dla uproszczenia zajmiemy się wyłącznie związkami między dwiema zmiennymi. Każdy wynik pomiaru możemy zatem przedstawić na płaszczyźnie w postaci punktu leżącego na przecięciu dwóch prostych przechodzących przez osie wykresu w punktach odpowiadających wartościom korelowanych zmiennych (ryc. 20). Na osi odciętych (oś pozioma) umieszcza się wartości zmiennej niezależnej, na osi rzędnych (oś pionowa) – wartości zmiennej zależnej. Wykres taki, przedstawiający rozrzut par punktów doświadczalnych (tzw. scattergram), pozwala w pierwszym przybliżeniu określić kształt badanego związku.

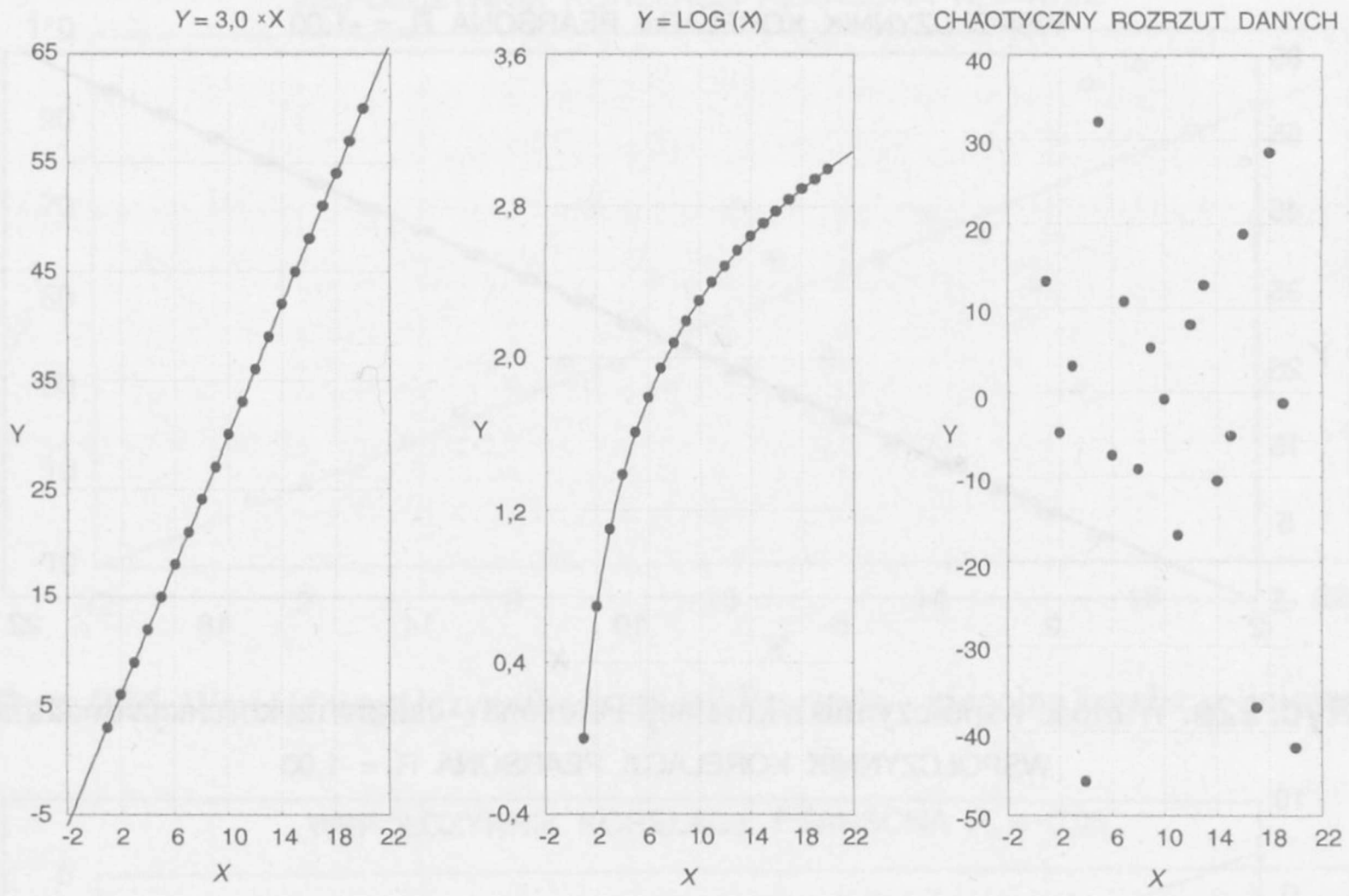
Jeżeli punkty będą miały tendencję do układania się wzdłuż linii prostej, możemy podejrzewać, że zachodzi między badanymi zmiennymi zależność liniowa, gdy układają się wzdłuż jakiejś krzywej – prawdopodobnie mamy do czynienia z zależnością nieliniową, wreszcie gdy punkty są rozrzucone chaotycznie na całej płaszczyźnie, najprawdopodobniej między zmiennymi nie ma żadnej zależności (ryc. 21).

Najczęściej badanym związkiem między zmiennymi jest związek liniowy. Jego siłę oszacowuje się, wyznaczając **współczynnik korelacji liniowej Pearsona** (*Pearson's linear correlation coefficient*). Może on przyjmować wartości z zakresu od –1 do +1.

Wartość +1 oznacza całkowitą dodatnią zależność liniową, co możemy interpretować w sposób następujący: wszystkie punkty pomiarowe leżą



Ryc. 20. Zasada tworzenia wykresu korelacyjnego (*scattergramu*)



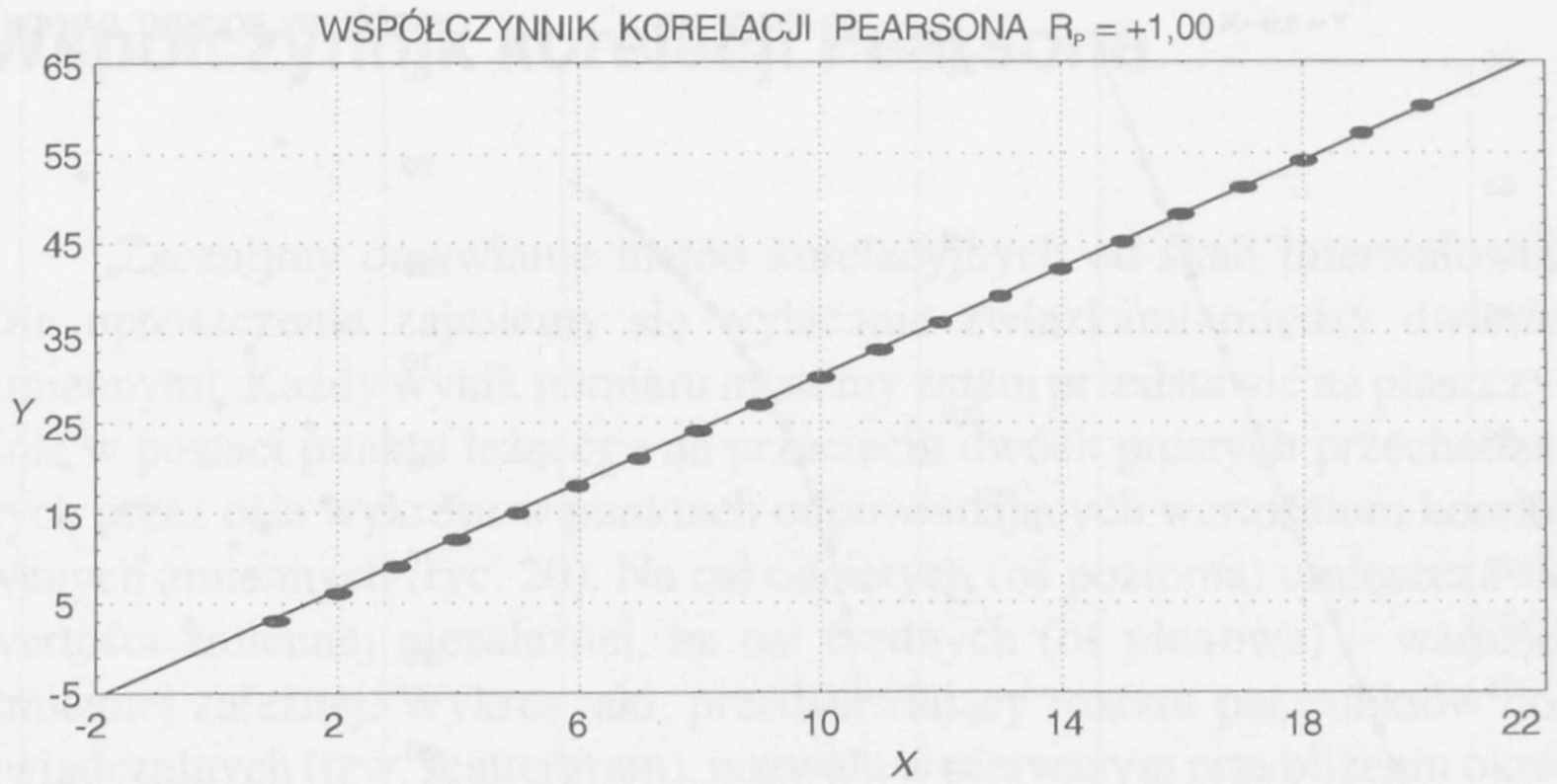
Ryc. 21. Przykładowe wykresy rozrzutu punktów dla zależności liniowej, nieliniowej oraz braku zależności

idealnie na linii prostej, a wzrostowi zmiennej niezależnej odpowiada wzrost zmiennej zależnej (ryc. 22a).

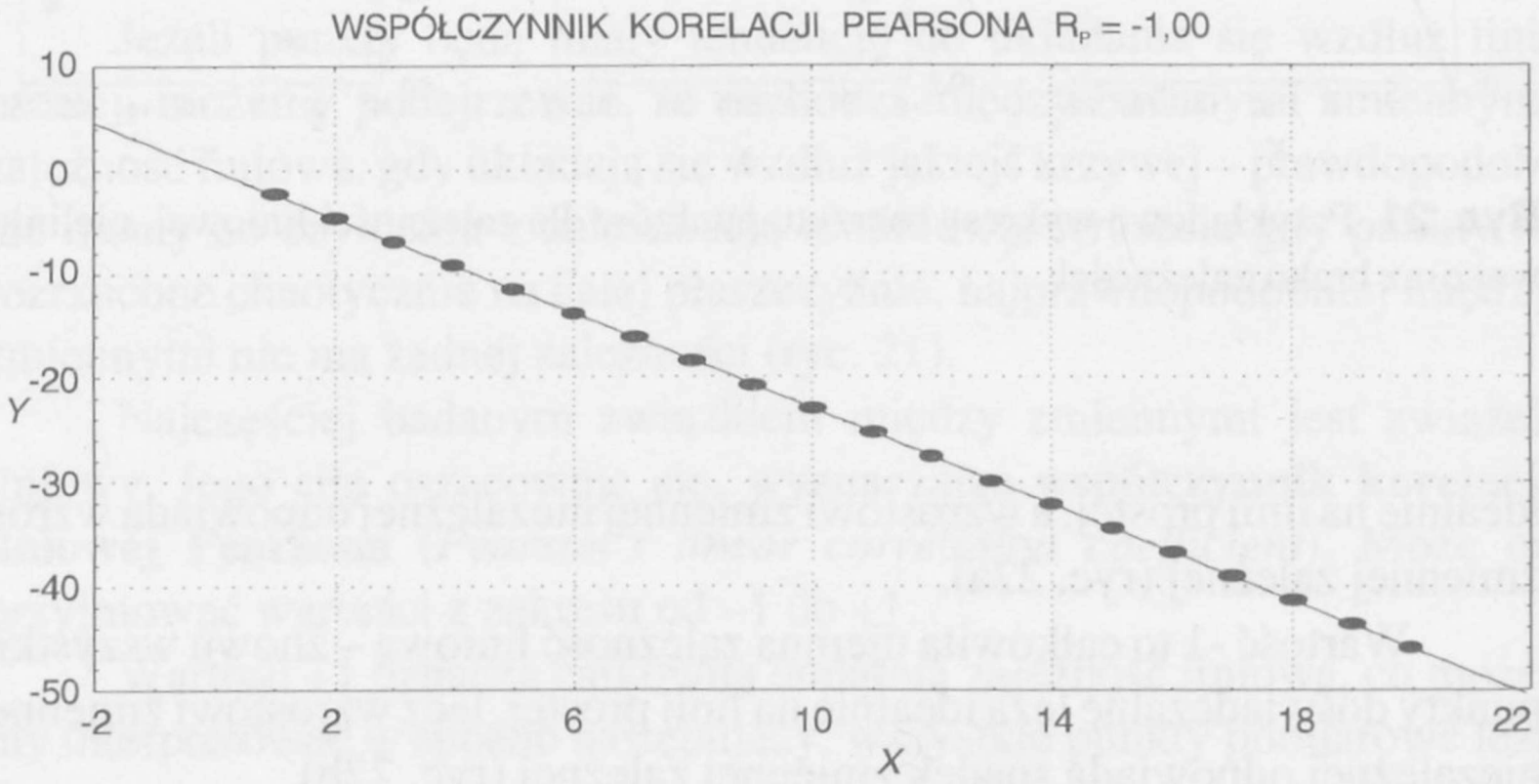
Wartość -1 to całkowita ujemna zależność liniowa – znowu wszystkie punkty doświadczalne leżą idealnie na linii prostej, lecz wzrostowi zmiennej niezależnej odpowiada spadek zmiennej zależnej (ryc. 22b).

Wartość 0 odpowiada całkowitemu brakowi zależności liniowej – przykładowo punkty mogą być rozrzucone chaotycznie, tak jak na ryc. 22c.

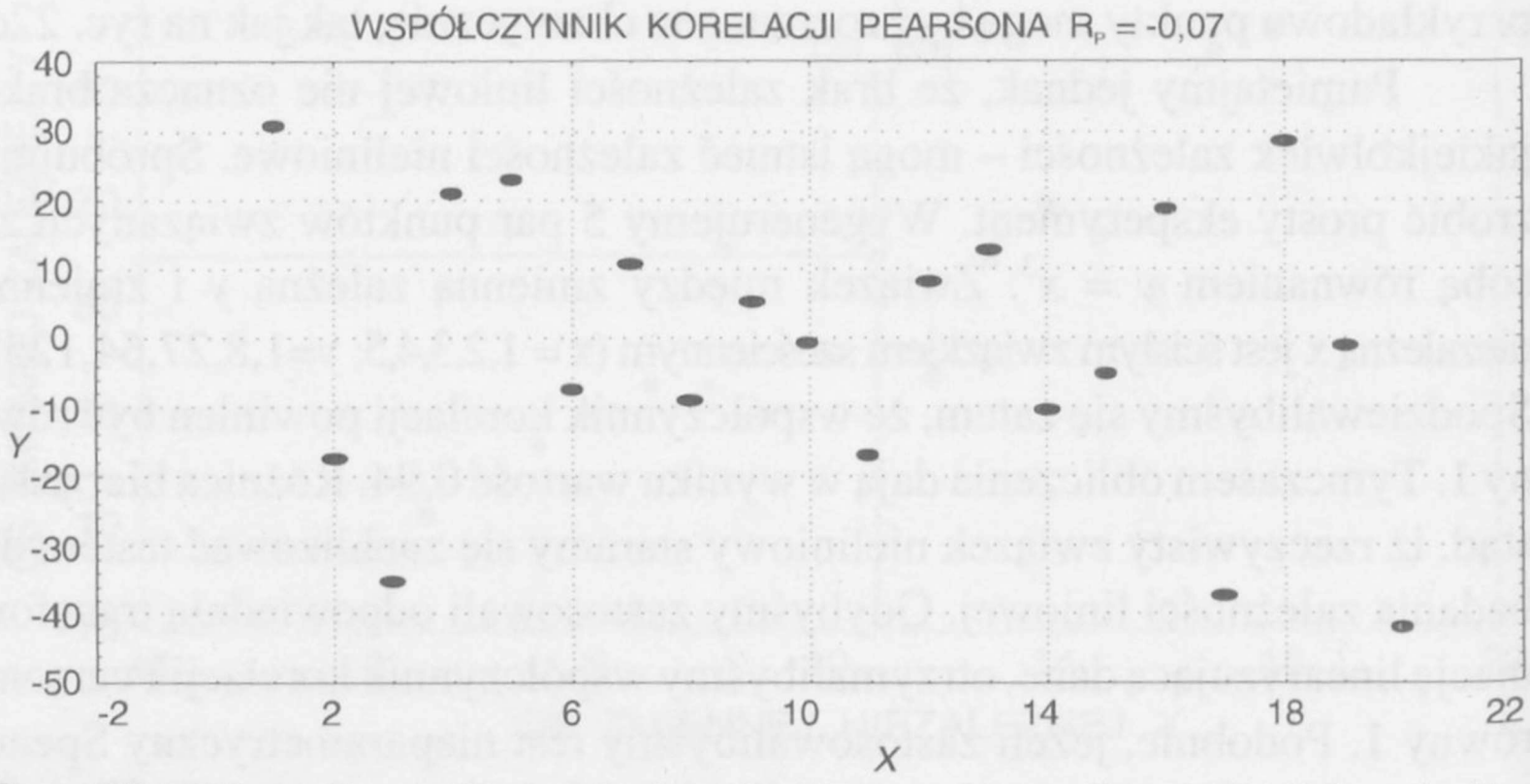
Pamiętajmy jednak, że brak zależności liniowej nie oznacza braku jakiejkolwiek zależności – mogą istnieć zależności nieliniowe. Spróbujmy zrobić prosty eksperyment. Wygenerujemy 5 par punktów związanych ze sobą równaniem $y = x^3$. Związek między zmienną zależną y i zmienną niezależną x jest ścisłym związkiem sześciennym (x = 1,2,3,4,5; y=1,8,27,64,125). Spodziewalibyśmy się zatem, że współczynnik korelacji powinien być równy 1. Tymczasem obliczenia dają w wyniku wartość 0,94. Różnica bierze się stąd, iż rzeczywisty związek nieliniowy staramy się zanalizować testem do badania zależności liniowej. Gdybyśmy zastosowali odpowiednią transformację linearyzującą dane, otrzymalibyśmy współczynnik korelacji Pearsona równy 1. Podobnie, jeżeli zastosowalibyśmy test nieparametryczny Spearmana, badający wyłącznie stopień zależności między zmiennymi, a abstra-



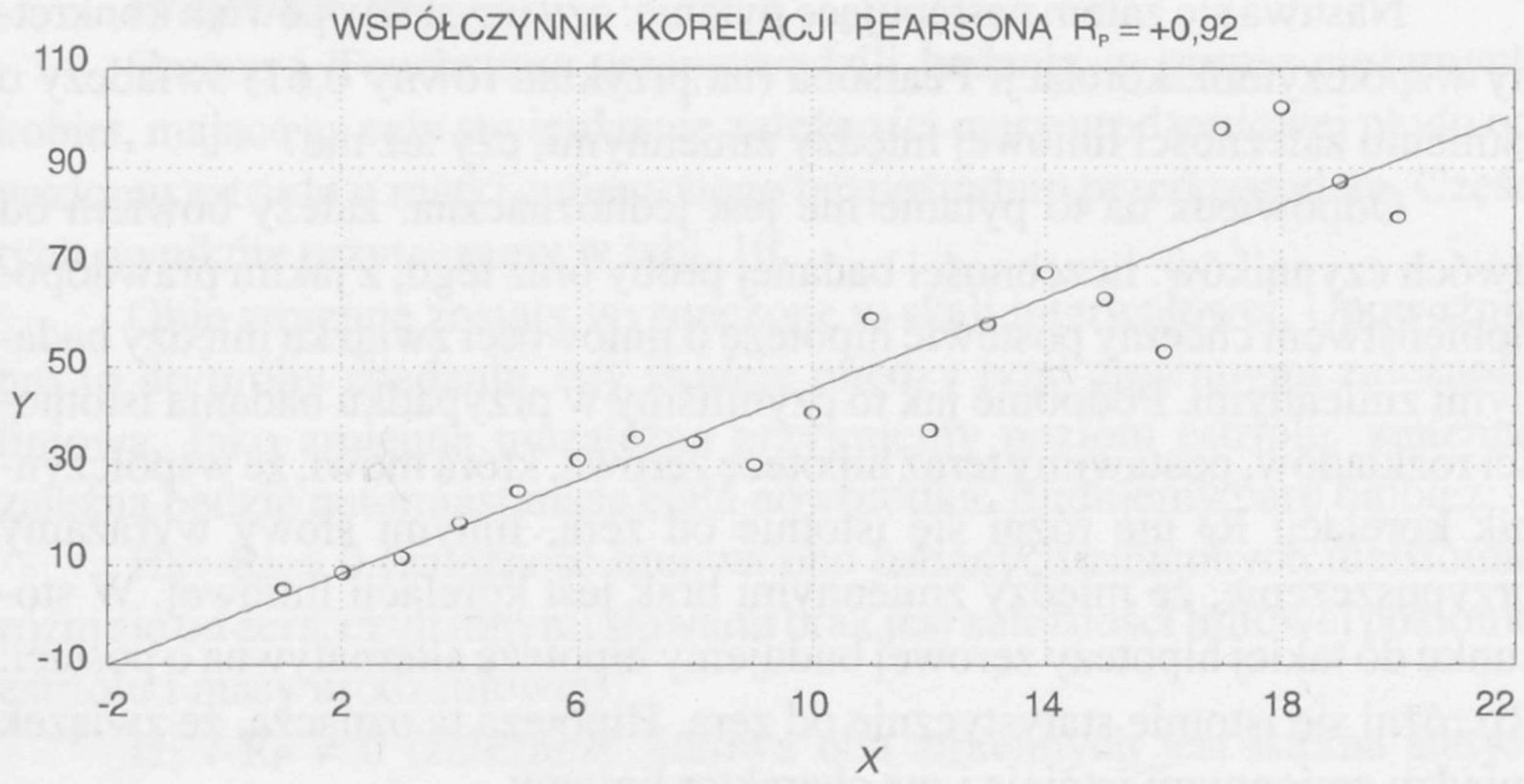
Ryc. 22a. Wartość współczynnika korelacji Pearsona – całkowita korelacja dodatnia



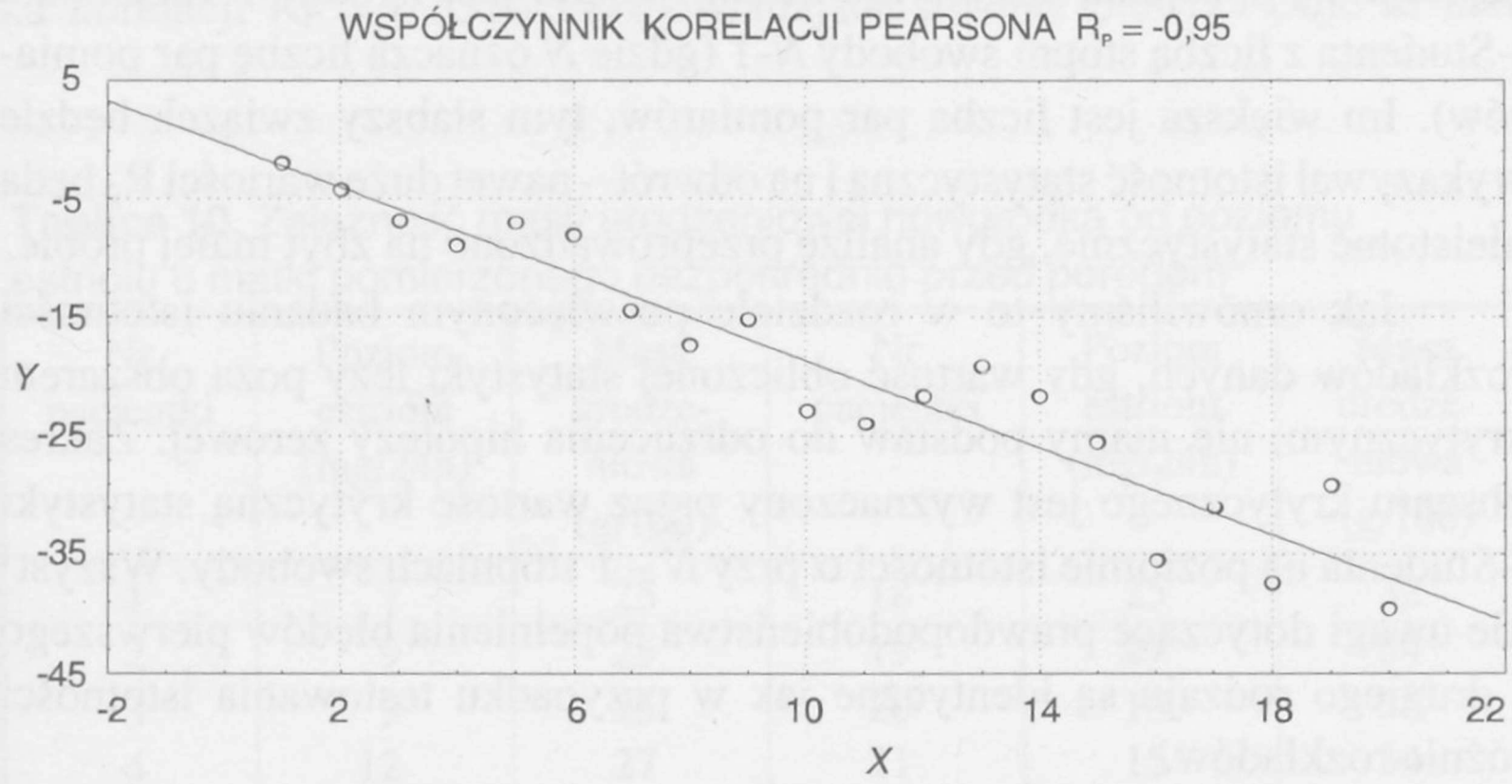
Ryc. 22b. Wartość współczynnika korelacji Pearsona – całkowita korelacja ujemna



Ryc. 22c. Wartość współczynnika korelacji Pearsona – całkowity brak korelacji



Ryc. 22d. Wartość współczynnika korelacji Pearsona – niepełna korelacja liniowa



Ryc. 22e. Wartość współczynnika korelacji Pearsona – niepełna korelacja liniowa ujemna

hujący od kształtu tej zależności, również uzyskalibyśmy w wyniku pełną korelację dodatnią.

Najczęściej będziemy jednak mieli do czynienia z sytuacją, gdzie współczynnik korelacji Pearsona będzie przyjmował wartości pośrednie z przedziału <–1, +1> (ryc. 22 d, e). Im bardziej liniowy będzie związek między zmiennymi, tym punkty będą leżały bliżej prostej regresji i tym bardziej wartość współczynnika korelacji Pearsona będzie się zbliżała do wartości +1 lub –1.

Nasuwa się zatem następujące pytanie: czy uzyskany pewien konkretny współczynnik korelacji Pearsona (na przykład równy 0,61) świadczy o istnieniu zależności liniowej między zmiennymi, czy też nie?

Odpowiedź na to pytanie nie jest jednoznaczna, zależy bowiem od dwóch czynników: liczebności badanej próby oraz tego, z jakim prawdopodobieństwem chcemy postawić hipotezę o liniowości związku między badanymi zmiennymi. Podobnie jak to czyniliśmy w przypadku badania istotności rozkładów, postawimy teraz hipotezę zerową, która mówi, że współczynnik korelacji R_P nie różni się istotnie od zera. Innymi słowy wyrażamy przypuszczenie, że między zmiennymi brak jest korelacji liniowej. W stosunku do takiej hipotezy zerowej budujemy hipotezę alternatywną o postaci: Rp różni się istotnie statystycznie od zera. Hipoteza ta oznacza, że związek między zmiennymi istnieje i ma charakter liniowy.

Do weryfikacji, która z tych hipotez badawczych jest słuszna, wykorzystuje się specjalną statystykę, której rozkład jest zgodny z rozkładem t-Studenta z liczbą stopni swobody N-1 (gdzie N oznacza liczbę par pomiarów). Im większa jest liczba par pomiarów, tym słabszy związek będzie wykazywał istotność statystyczną i na odwrót – nawet duże wartości R_p będą nieistotne statystycznie, gdy analizę przeprowadzono na zbyt małej próbie.

Jak omówiliśmy to w rozdziale poświęconym badaniu istotności rozkładów danych, gdy wartość obliczonej statystyki leży poza obszarem krytycznym, nie mamy podstaw do odrzucenia hipotezy zerowej. Zakres obszaru krytycznego jest wyznaczony przez wartość krytyczną statystyki t-Studenta na poziomie istotności α przy $N-1$ stopniach swobody. Wszystkie uwagi dotyczące prawdopodobieństwa popełnienia błędów pierwszego i drugiego rodzaju są identyczne jak w przypadku testowania istotności różnic rozkładów.

Niezwykle ważną rolę w interpretacji siły związku pełni kwadrat współczynnika korelacji pomnożony przez 100%. Wyraża on bowiem procent zmienności zmiennej zależnej, spowodowanej zmiennością zmiennej niezależnej. W naszym przykładzie z funkcją $y=x^3$ współczynnik korelacji liniowej wynosił 0,94, związek liniowy zatem tłumaczył jedynie około 89% ($0{,}94 \times 0{,}94 \times 100\%$) zmienności zmiennej zależnej y jako spowodowaną przez zmienność zmiennej niezależnej. Aż 11% zmienności zmiennej y było spowodowane innymi czynnikami niż zmienność zmiennej x (w istocie chodziło tu o efekty nieliniowe). Wspomniany współczynnik nosi nazwę **współczynnika determinacji** (*determination coefficient*).

Przykład 19

Greene i Touchstone przeprowadzili badania w grupie ciężarnych kobiet, mające na celu stwierdzenie zależności masy urodzeniowej płodu od poziomu estriolu u matki zmierzonego bezpośrednio przed porodem. Część tych wyników przytaczamy w tabl. 10.

Obie zmienne zostały wyznaczone w skali interwałowej. Upoważnia nas to do próby zbadania, czy istnieje między tymi zmiennymi zależność liniowa. Jako zmienną niezależną przyjmiemy poziom estriolu, zmienną zależną będzie natomiast masa ciała noworodka. Budujemy parę hipotez:

$H_0 : R_P = 0$ (zależność liniowa obu badanych zmiennych nieistotnie różni się od zera, czyli innymi słowami brak jest zależności liniowej poziomu estriolu i masy urodzeniowej)

$H_1 : R_P \neq 0$ (zależność liniowa obu zmiennych jest istotna statystycznie).

Stosując test korelacyjny Pearsona otrzymamy wartość współczynnika korelacji $R_P = 0{,}5325$ i odpowiadającą *p-value* 0,0012. Daje to nam

Tablica 10. Zależność masy urodzeniowej noworodka od poziomu estriolu u matki pomierzonego bezpośrednio przed porodem

Nr pacjentki	Poziom estriolu (mg/24h)	Masa urodze-niowa (g/100)	Nr pacjentki	Poziom estriolu (mg/24h)	Masa urodze-niowa (g/100)
1	7	25	18	25	32
2	9	25	19	27	34
3	9	25	20	15	34
4	12	27	21	15	34
5	14	27	22	15	34
6	16	27	23	15	35
7	16	24	24	16	35
8	14	30	25	19	34
9	16	30	26	18	35
10	16	31	27	17	36
11	17	30	28	18	37
12	19	31	29	20	38
13	21	30	30	22	40
14	24	28	31	25	39
15	15	32	32	24	43
16	16	32	33	21	28
17	17	32	34	25	31

podstawę do odrzucenia hipotezy zerowej i stwierdzenia, że między analizowanymi zmiennymi istnieje zależność liniowa. Na podstawie wyliczonych przez program współczynników nachylenia prostej oraz jej punktu przecięcia z osią zmiennej zależnej: możemy zatem zbudować liniowy model tej zależności w postaci:

masa urodzeniowa (wyrażona w g/100) = 0,5109 × poziom estriolu – 44,9580

Teoretycznie moglibyśmy używać tego modelu do interpolacji i ekstrapolacji. Przyjrzyjmy się jednak współczynnikowi determinacji. Wynosi on 0,2835. Oznacza to, że nasz model, chociaż istotny statystycznie, tłumaczy jedynie 28,35% zmienności masy urodzeniowej. Oznacza to, że aż 71,65% tej zmienności jest spowodowanych innymi czynnikami niż poziom estriolu u matki. Musimy zatem bardzo ostrożnie traktować wartości uzyskiwane na podstawie modelu i lepiej byłoby, gdyby udało nam się znaleźć inne czynniki mające wpływ na masę urodzeniową. Moglibyśmy wtedy zbudować model liniowy oparty na wielu zmiennych niezależnych. Można również próbować dokonać transformacji naszych danych i zbudować model nieliniowy. Obie jednakże opcje wychodzą poza podstawowy zakres metod przewidzianych do prezentacji w tej książce.

Test Spearmana

Omawiając „klasyczny" współczynnik korelacji Pearsona podkreślaliśmy, że wykrywany niekiedy przez ten współczynnik brak związku liniowego między zmiennymi nie oznacza w istocie braku jakiegokolwiek związku między nimi. Dlatego jeśli nie możemy się doszukać w zgromadzonych danych potrzebnego nam związku warto jest jeszcze wykonać, prócz testu Pearsona, dodatkowy test korelacyjny – **test Spearmana** (*Spearman's test*). Jest to test nieparametryczny (a więc z klasy *distribution-free tests*).

Test Spearmana jest w zasadzie przeznaczony do badania związków między zmiennymi uzyskanymi w skali porządkowej. Można go jednak z powodzeniem wykorzystywać również do badania związku w sytuacjach, gdy jedna ze zmiennych jest uzyskana w skali porządkowej, a druga w interwałowej lub gdy obie zmienne pochodzą ze skali interwałowej, lecz nie jesteśmy pewni liniowego charakteru związku między nimi.

Współczynnik korelacji Spearmana R_{Ss} podobnie jak współczynnik R_P Pearsona przyjmuje wartości z przedziału $<-1, +1>$. Jego istotność testuje się w identyczny sposób, jak istotność współczynnika R_P, a jego kwadrat pomnożony przez 100 procent również wyraża procent zmienności zmiennej zależnej spowodowanej zmiennością zmiennej niezależnej. Gdybyśmy oszacowali siłę korelacji Spearmana dla danych uzyskanych z funkcji $y=x^3$ uzyskamy wartość $R_S = +1$, co wskazuje na pełny związek między zmienną x oraz y i pełne, 100-procentowe wytłumaczenie zmienności zmiennej y zmiennością zmiennej x (wartość współczynnika korelacji liniowej $R_P = 0{,}98$). Dowodzi to, że test Spearmana pozwala wykrywać również zależności typu nieliniowego.

Związek między zmiennymi w skali nominalnej

Podobnie jak w skali interwałowej i porządkowej, związek między zmiennymi jest również możliwy do oszacowania dla pomiarów wyrażonych w najsłabszej skali – skali nominalnej. Jako miary korelacji używamy wtedy **ryzyka względnego** (*relative risk*) lub **stosunku szans** zwanego również **ilorazem szans** (*odds ratio*).

Pojęcie ryzyka względnego wprowadza się używając takich samych tablic kontyngencji, jakie stosowaliśmy do oszacowania wartości statystyki chi-kwadrat przy testowaniu hipotez. Dla lekarza najbardziej przemawiająca będzie nie sucha definicja, lecz prezentacja istoty rozważanego postępowania na konkretnym przykładzie.

Załóżmy, że jakaś grupa badanej populacji jest narażona na pewien czynnik ryzyka. Liczbę osób, u których wystąpiła (związana z istnieniem czynnika ryzyka) choroba oznaczamy jako A, natomiast B określa liczbę osób, u których mimo narażenia choroba nia pojawiła się. Okazuje się jednak, że jednostka chorobowa pojawia się również wśród ludzi, którzy nie byli narażeni na badany czynnik (C). Ostatnia wreszcie podgrupa to osoby zdrowe nie narażone na czynnik ryzyka (D). Jak wspomnieliśmy najprościej zobrazować istniejącą sytuację za pomocą tablicy kontyngencji.

Jednostka chorobowa / Czynnik ryzyka	Występuje	Nie występuje
Obecny	A	B
Nieobecny	C	D

Pytanie brzmi: czy ekspozycja na czynnik ryzyka zwiększa prawdopodobieństwo zachorowania na daną jednostkę chorobową? W tym miejscu odejdziemy od przyjętej przez nas reguły i podamy (z uwagi na jego prostotę) wzór na obliczenie ryzyka względnego:

$$\text{ryzyko względne} = [A/(A+B)] \,/\, [C/(C+D)]$$

Słownie możemy wyrazić ten wzór następująco: ryzyko względne jest stosunkiem prawdopodobieństwa wystąpienia choroby w grupie narażonej na czynnik ryzyka do prawdopodobieństwa wystąpienia tej choroby w grupie nie narażonej na czynnik ryzyka.

Jak widzimy z definicji zmienność ryzyka względnego mieści się w przedziale od 0 do plus nieskończoności. Gdy jego wartość wynosi 1,0 czynnik ryzyka nie ma żadnego wpływu na zachorowalność na badaną jednostkę chorobową (nie ma związku między ekspozycją na czynnik ryzyka a występowaniem choroby). Dla wartości ryzyka względnego >1,0 zachorowalność w grupie osób eksponowanych na czynnik ryzyka wzrasta, dla wartości <1,0 dochodzimy wręcz do dziwnego wniosku: ekspozycja na czynnik, który nazwaliśmy ryzykiem, zapobiega zachorowalności na badaną chorobę. Widzimy zatem, że ryzyko względne może stanowić miarę związku między zmiennymi wyrażonymi w skali nominalnej.

Równie prosta jest definicja stosunku szans. Odnoszący się do tej samej tablicy kontyngencji wzór na stosunek szans jest następujący:

iloraz szans = $(A/C) / (B/D)$

Zakres zmienności stosunku szans jest identyczny jak zakres zmienności ryzyka względnego i, podobnie, wyraża on związek między badanymi zmiennymi.

W literaturze medycznej bardzo często spotyka się cztery dodatkowe terminy związane z opisem związku między dwiema zmiennymi wyrażonymi w skali nominalnej. Są to **wartość predykcyjna dodatnia**, **wartość predykcyjna ujemna**, **czułość** i **specyficzność**. Ich właściwości omówimy na następującym przykładzie.

Chcemy ocenić, jaka jest przydatność stosowania kardiotokograficznego testu niestresowego do oceny stanu płodu. W rozważaniach ograniczymy się jedynie do czterech możliwości zebranych w poniższej tablicy kontyngencji. Pamiętajmy, że w naszym przykładzie wynik negatywny testu oznacza zapis prawidłowy, wynik pozytywny natomiast stanowi podejrzenie stanu nieprawidłowego. (Często zdarza się odwrotna interpretacja wyniku pozytywnego i negatywnego!)

Wynik testu / Stan płodu	Stan prawidłowy	Stan zagrożenia
Test negatywny	*A*	*B*
Test pozytywny	*C*	*D*

Wartość predykcyjna dodatnia (*positive predictive value*) jest zdefiniowana jako stosunek $A/(A + B)$ i odpowiada na pytanie: jeżeli wynik testu był prawidłowy (test negatywny) to jakie jest prawdopodobieństwo tego, że płód jest rzeczywiście w stanie prawidłowym?

Wartość predykcyjna ujemna (*negative predictive value*) definiujemy jako $D/(C + D)$, co daje nam odpowiedź na pytanie: jeżeli wynik testu wskazuje na istnienie nieprawidłowości (test pozytywny), to jakie jest prawdopodobieństwo tego, że płód jest rzeczywiście zagrożony?

Czułość testu (*sensitivity*) definiowana jako stosunek $A/(A + C)$ mówi o tym, jakie jest prawdopodobieństwo zarejestrowania testu negatywnego, jeżeli płód znajduje się rzeczywiście w stanie niezagrożonym.

Specyficzność testu (*specifity*) określana jako stosunek $D/(D + B)$ odpowiada na pytanie: jak prawdopodobne jest uzyskanie wyniku pozytywnego testu niestresowego, gdy płód jest rzeczywiście zagrożony.

W rozdziale tym omówiliśmy najczęściej stosowane metody badania związków między dwiema zmiennymi. Istnieją oczywiście techniki pozwalające na badanie takich związków jak jedna zmienna zależna – wiele zmiennych niezależnych oraz wiele zmiennych zależnych – wiele zmiennych niezależnych. Jest to domeną tak zwanej **analizy metod wielowymiarowych** (*multivariate analysis*). Z uwagi na bardziej złożoną strukturę stosowanych tam metod (niezbędna jest znajomość rachunku macierzowego) poprzestaniemy jedynie na uwadze, że techniki takie istnieją, a programy umożliwiające ich zastosowanie są zawarte w większości pakietów statystycznych.

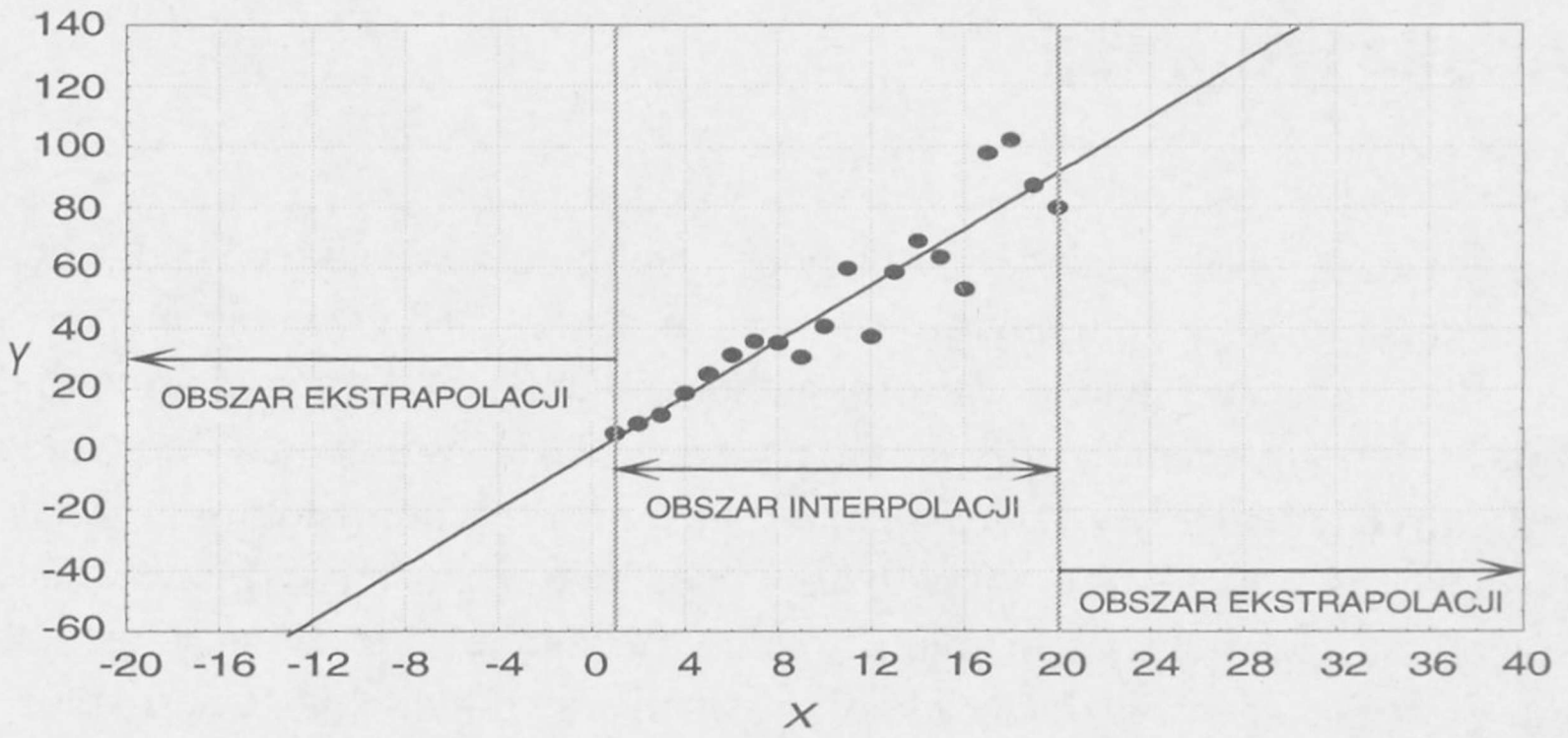
Analiza regresji

Drugą obok analizy korelacyjnej techniką używaną w badaniach zależności między zmiennymi jest **analiza regresji** (*regression analysis*). Pozwala ona na predykcję wartości jednej zmiennej – tak zwanej **zmiennej zależnej** (*dependent variable*) na podstawie wartości jednej lub wielu **zmiennych niezależnych** (*independent variables*). W poprzednim rozdziale poświęconym korelacji omówiliśmy podstawowe techniki pozwalające stwierdzić, czy między badanymi zmiennymi istnieje jakieś powiązanie, jak duża jest siła tego związku i jak bardzo jest on prawdopodobny. Często może jednak nasuwać się pytanie dodatkowe: jeżeli stwierdziliśmy, że związek między zmiennymi istnieje, to czy możemy na podstawie wartości jednej zmiennej przewidzieć, jaka powinna być wartość zmiennej powiązanej? Zagadnieniem tym zajmuje się właśnie analiza regresji.

Zanim podejmiemy jakąkolwiek akcję predykcyjną, powinniśmy przeprowadzić pełną analizę korelacyjną. Dla skali interwałowej musimy zatem wstępnie ustalić (np. na podstawie rysunku rozrzutu punktów – scattergramu), jakiego poszukujemy modelu zależności (liniowy, kwadratowy, eksponencjalny, logarytmiczny itp.). Dla przyjętego modelu przeprowadzamy pełną analizę korelacyjną, to znaczy oszacowujemy siłę związku i jego prawdopodobieństwo. Na podstawie wartości kwadratu współczynnika korelacji stwierdzamy, jaki procent zmienności zmiennej zależnej wynika z samej zmienności zmiennej niezależnej. Jeżeli z takiej wstępnej analizy uzyskamy wynik, że korelacja jest na zadanym poziomie istotna statystycznie i procent wariancji jest dla nas satysfakcjonujący, to możemy przejść do procedury predykcyjnej. Obejmuje ona dwa podstawowe schematy: **interpolację** (*interpolation*) oraz **ekstrapolację** (*extrapolation*).

Interpolacja polega na przewidywaniu wartości zmiennej zależnej na podstawie wartości zmiennej niezależnej leżących wewnątrz obszaru wcześniej obserwowanych zmian zmiennej niezależnej (a więc tego zakresu zmiennej niezależnej, na podstawie którego zbudowano wykorzystywany model – ryc. 23). Interpolacja jest więc z reguły procedurą bezpieczną – zakłada się tu jedynie ciągłość funkcji wyrażającej zależność obu zmiennych.

Ekstrapolacja jest przewidywaniem wartości zmiennej zależnej dla tych wartości zmiennej niezależnej, które leżą poza obszarem zmienności tej zmiennej użytym do budowy modelu predykcyjnego. W przeciwieństwie do interpolacji, ekstrapolacja bywa często zabiegiem ryzykownym. Budując bowiem model regresji bazujemy na pewnym zakresie zmienności zmiennej



Ryc. 23. Zasada interpolacji i ekstrapolacji

niezależnej. Kształt związku, siłę korelacji i jej wiarogodność badamy jedynie dla tego zakresu zmienności, nie możemy natomiast niczego powiedzieć o charakterze związku poza przebadanym obszarem zmienności. Być może związek pozostaje taki sam, ale również prawdopodobne jest, że zmienia on charakter (kształt, siłę, staje się nieistotny statystycznie itp.) – (ryc. 24). Przyjmuje się zwykle, że ekstrapolacja w obszarze nie przekraczającym 10% rozstępu obszaru zmienności zmiennej niezależnej, dla którego zbudowano model predykcyjny jest procedurą bezpieczną. Nie zawsze jednak musi to być prawdą.

Z reguły większość modeli predykcyjnych w skali interwałowej bazuje na tak zwanej metodzie **najmniejszych kwadratów** (*least squares method*). Istota tej metody polega na minimalizacji sumy kwadratów błędów predykcji. Błędy te są w zasadzie nieuniknione, rzadko bowiem się zdarza, aby dopasowywana prosta lub krzywa pasowała idealnie do wszystkich punktów doświadczalnych. W zasadzie zdarza się to jedynie w dwóch sytuacjach:

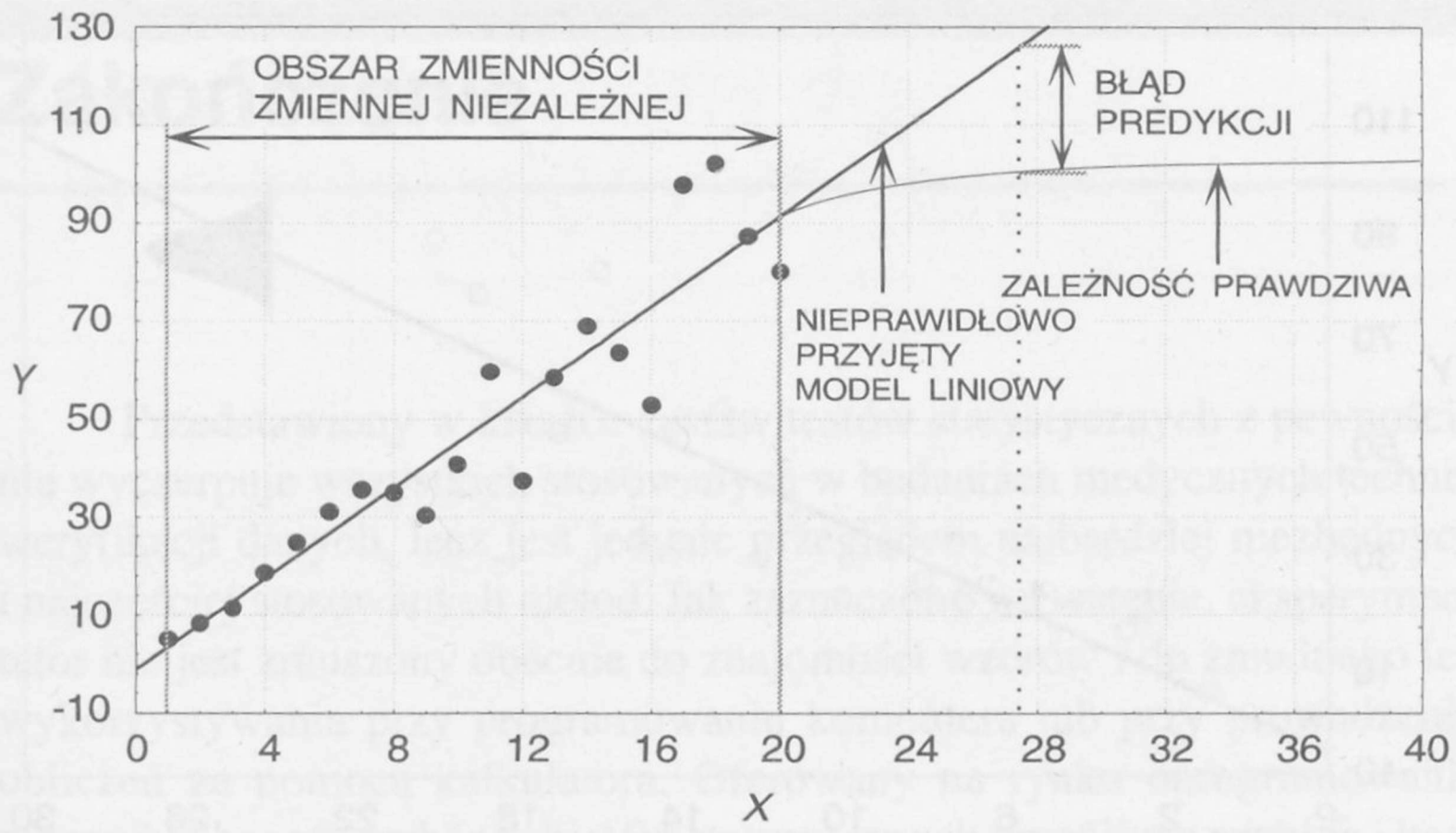
– gdy zależność między badanymi zmiennymi ma charakter funkcyjny (np. $y = \log x$)

– gdy rząd dopasowywanej funkcji jest równy liczbie punktów doświadczalnych.

Obie sytuacje są niezwykle rzadkie, gdyż:

– błędy pomiarowe i błędy próbkowania skutecznie deformują teoretycznie występującą zależność funkcyjną,

– dopasowanie funkcji wielomianowej wyższego rzędu niż trzy stwarza trudności w interpretacji związku. Z reguły w modelowaniu stosuje się



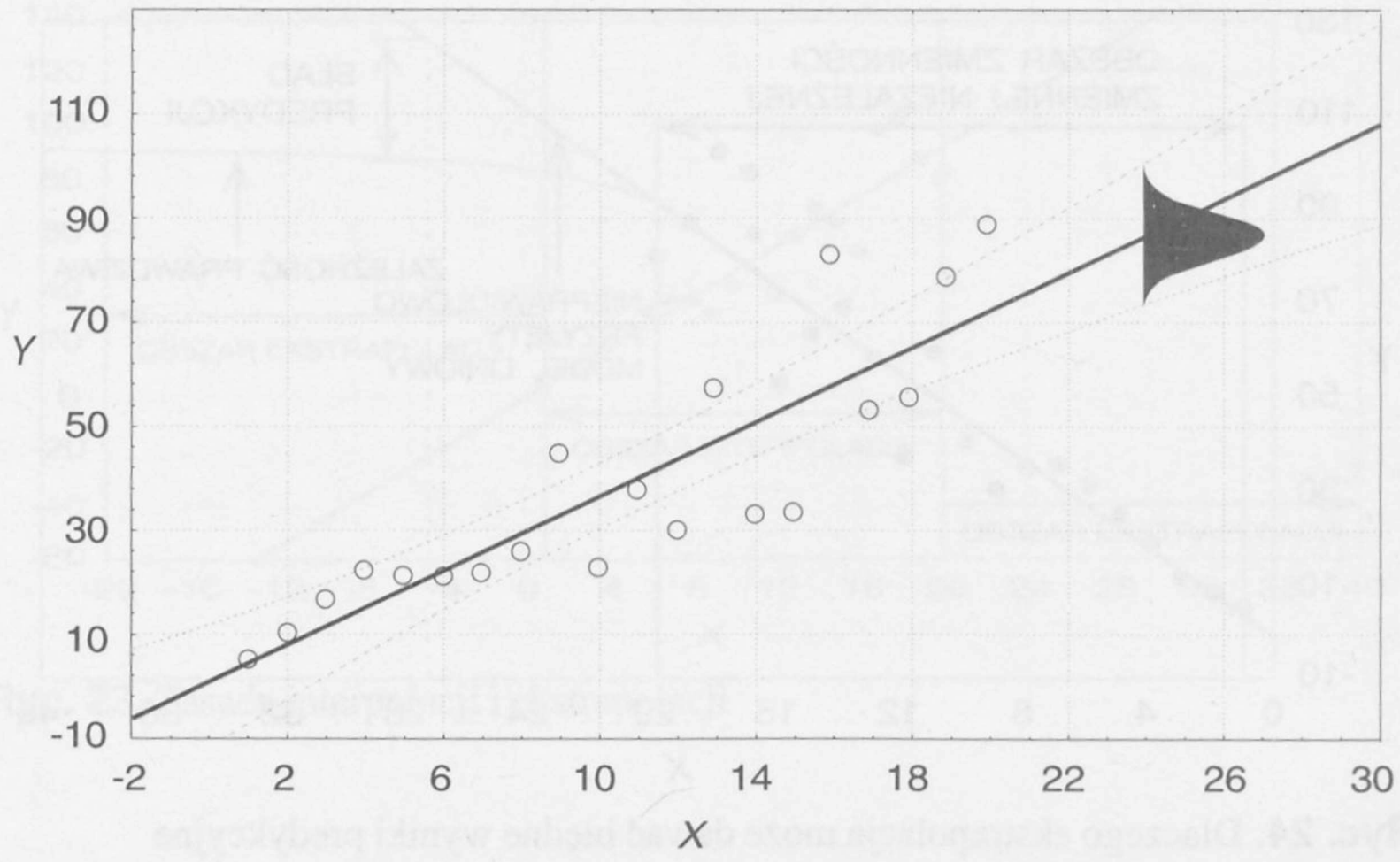
Ryc. 24. Dlaczego ekstrapolacja może dawać błędne wyniki predykcyjne

związki typu liniowego, kwadratowego, sześciennego, eksponencjalnego i logarytmicznego.

W związku z tym prostą (lub krzywą) regresji prowadzi się w ten sposób, by suma kwadratów odległości punktów doświadczalnych od tej prostej (lub krzywej) była jak najmniejsza. Stosuje się przy tym podstawowe metody rachunku różniczkowego, co pozwala na znalezienie takich parametrów dopasowywanej prostej (lub krzywej), by postawiony warunek minimalizacji błędu był spełniony.

Znalezienie przewidywanej wartości zmiennej zależnej na podstawie zadanej wartości zmiennej niezależnej polega po prostu na podstawieniu wartości zmiennej niezależnej do równania prostej (lub krzywej) regresji, której parametry zostały znalezione metodą najmniejszych kwadratów. Musimy pamiętać o tym, że znalezione metodą interpolacji lub ekstrapolacji wartości zmiennej zależnej są obarczone błędem, podobnie jak i wartości zmiennych pomiarowych. W zasadzie każdy pakiet statystyczny oprócz znalezienia i wykreślenia krzywej regresji oblicza i zaznacza na wykresie krzywe obszaru ufności na zadanym przez badacza poziomie istotności statystycznej (ryc. 25). Ich interpretacja jest taka sama, jak interpretacja przedziałów ufności zbudowanych wokół średniej arytmetycznej.

Do tej pory mówiliśmy o technice regresyjnej stosowanej w przypadku, gdy zarówno zmienna zależna, jak i niezależna były wyrażone w skali interwałowej. Szczególnym przypadkiem regresji jest **regresja logistyczna**



Ryc. 25. Rozkład estymowanych metodą predykcji wartości zmiennej zależnej i krzywe obszaru ufności

(*logistic regression*) i **techniki modelowania log-linear** (*log-linear modelling techniques*). Pierwsza z nich ma zastosowanie, gdy zmienna niezależna jest wyrażona w dowolnej z trzech skal, natomiast zmienna zależna w skali nominalnej lub porządkowej dychotomicznej (występują jedynie dwie wartości pomiarowe np. *przeżył, zmarł*). Drugą stosujemy w przypadku, gdy zmienna niezależna jest wyrażona w skali nominalnej lub porządkowej, a zmienna zależna w wielostopniowej (niedychotomicznej) skali nominalnej lub porządkowej.

Zakończenie

Przedstawiony w książce zestaw testów statystycznych z pewnością nie wyczerpuje wszystkich stosowanych w badaniach medycznych technik weryfikacji danych, lecz jest jedynie przeglądem najbardziej niezbędnych i najczęściej stosowanych metod. Jak zaznaczono we wstępie, eksperymentator nie jest zmuszony obecnie do znajomości wzorów i do żmudnego ich wykorzystywania przy programowaniu komputera lub przy prowadzeniu obliczeń za pomocą kalkulatora. Oferowany na rynku oprogramowania niezwykle bogaty wybór pakietów statystycznych umożliwia szybkie i bezbłędne dokonanie analiz, lecz w żadnym razie nie zwalnia od prawidłowego zaprojektowania eksperymentu, zebrania danych, wyboru optymalnego w danej sytuacji testu statystycznego oraz interpretacji wyników. Należy jednakże podkreślić, że coraz częściej pojawiają się w ofercie oprogramowania pakiety statystyczne, które ułatwiają wybór prawidłowego testu oraz interpretację otrzymanych wyników. Przegląd najczęściej stosowanych w badaniach medycznych programów statystycznych można znaleźć w pracy J. Moczko [6].

Zadania-problemy do samodzielnego rozwiązania

Celem tej części podręcznika jest zaproponowanie kilku zadań-problemów, które Czytelnik powinien samodzielnie rozwiązać na podstawie materiału przedstawionego w podręczniku. W obliczeniach jest wskazane wykorzystanie dostępnych pakietów statystycznych, gdyż zgodnie z przyjętym przez nas założeniem niepodawania wzorów jest to jedyna droga przeprowadzania tego typu obliczeń.

Zadanie 1

Wyniki badań populacji 10 000 ciężarnych kobiet w aspekcie oceny wystąpienia zagrożenia płodu w czasie porodu przedstawiono w tablicy 11. Jako czynnik ryzyka w analizowanej populacji określono opóźniony wewnątrzmaciczny rozwój płodu.

Tablica 11. Wyniki badań ciężarnych kobiet

Objawy zagrożenia płodu / Czynnik ryzyka	Występują	Brak	Ogółem
Występuje	80	920	1000
Brak	180	8820	9000
Ogółem	260	9740	10000

Proszę określić ryzyko względne, iloraz szans oraz wartości predykcyjne badanego objawu, którym jest opóźniony wzrost wewnątrzmaciczny płodu, w prognozowaniu wystąpienia objawów zagrożenia płodu.

Zadanie 2

Jednym z parametrów określających wydolność oddechową noworodka jest opór w drogach oddechowych. W trakcie badań przeprowadzonych u 28 noworodków uzyskano następujące wyniki:
183,9; 158,5; 137,3; 163,1; 159,3 204,2; 166,2; 190,7; 174,6
201,4; 112,8; 190,2; 182,0; 206,3; 160,8; 221,5; 154,7;

187,6; 166,0; 235,6; 152,4; 190,7; 144,2; 160,7; 165,9; 130,7; 211,0; 194,7

Proszę określić z jakim typem skali pomiarowej mamy do czynienia oraz scharakteryzować przytoczone dane za pomocą właściwych opisowych charakterystyk statystycznych.

Zadanie 3

Załóżmy, że przedstawiona w zadaniu 2. zmienna ma rozkład normalny, co jest zgodne z uzyskanymi wcześniej charakterystykami opisowymi. Opierając się na wcześniej obliczonych charakterystykach, proszę podać przedziały, w których znajduje się 68,3% oraz 95,5% wyników wszystkich obserwacji.

Zadanie 4

Korzystając z wartości zmiennej przedstawionej w zadaniu 2. i przyjmując, że ma ona rozkład normalny, proszę skonstruować przedziały ufności dla wartości średniej na poziomach istotności statystycznej 0,05, 0,01 oraz 0,001.

Zadanie 5

W styczniu na oddziale położniczym miały miejsce 763 porody. Rozkład masy urodzeniowej przedstawiono w tabl. 12. Proszę przedstawić graficznie rozkład analizowanej zmiennej (różne sposoby).

Tablica 12. Rozkład masy urodzeniowej

Masa urodzeniowa	Liczba porodów
4001 – 4500	15
3501 – 4000	73
3001 – 3500	302
2501 – 3000	211
2001 – 2500	120
1501 – 2000	33
1001 – 1500	7
500 – 1000	2

Zadanie 6

Celem przeprowadzonych badań epidemiologicznych było określenie ewentualnej zależności między warunkami pracy zawodowej a przebiegiem i zakończeniem ciąży. Jednym z ocenianych elementów była praca ciężarnej przy monitorze komputera we wczesnym okresie ciąży a wystąpieniem

poronienia, porodu przedwczesnego lub ukończenia ciąży w prawidłowym czasie. Uzyskano następujące wyniki:

Tablica 13. Zależność przebiegu i zakończenia ciąży od warunków pracy zawodowej

Zakończenie ciąży / Czas spędzany przy komputerze	Poronienie	Poród przedwczesny	Poród o czasie
Nie pracuje na komputerze	343	30	256
Okazyjnie	55	12	57
Często (do 4 godzin na dobę)	41	10	45
Bardzo często (powyżej 4 godzin na dobę)	57	6	57

Zadanie polega na doborze odpowiedniego testu statystycznego i odpowiedzi na pytanie, czy istnieje istotna statystycznie zależność między czasem spędzonym przy monitorze komputera a czasem ukończenia ciąży.

Zadanie 7

W celu zbadania nowego leku pod kątem jego skuteczności w terapii nadciśnienia tętniczego przeprowadzono badania u 15 chorych przed podaniem doustnym badanego środka oraz 2 godziny po nim. Mierzono ciśnienie rozkurczowe. Uzyskano następujące wyniki:

Przed podaniem leku		Po podaniu leku	
120	135	105	80
115	140	95	100
145	130	120	85
140	130	90	90
135	125	115	85
135	110	100	70
145	135	140	105
125		100	

Proszę zaproponować właściwy test oraz stwierdzić, czy badany lek ma istotny statystycznie wpływ na obniżenie rozkurczowego ciśnienia tętniczego.

Zadanie 8

W badaniach mających na celu wczesne rozpoznanie dysplazji oskrzelowo-płucnej jednym z ocenianych parametrów jest podatność dynamiczna. Badania przeprowadzono w 5. dobie życia u 21 noworodków, u których nie rozpoznano tego powikłania oraz u 8 noworodków, u których stwierdzono dysplazję oskrzelowo-płucną w kolejnych dobach życia. Uzyskano następujące wyniki:

Dzieci bez dysplazji		Dzieci z dysplazją
0,54	0,71	0,35
0,79	0,77	0,41
0,62	0,71	0,37
0,53	0,52	0,58
0,68	0,38	0,33
0,67	0,69	0,56
0,79	0,58	0,54
0,76	0,67	042
0,89	0,62	
0,79	0,72	
0,58		

Proszę wybrać właściwy test, który pozwoli na stwierdzenie, czy istnieją istotne statystycznie różnice wartości średnich podatności dynamicznej między badanymi grupami.

Zadanie 9

W dwóch ośrodkach leżących w różnych regionach kraju przeprowadzono badania wzrostu dzieci w wieku przedszkolnym (5–6 lat). Celem badań było stwierdzenie ewentualnych różnic w zakresie tej zmiennej. W I regionie zbadano 25 dzieci, w II regionie natomiast 30 dzieci. (*UWAGA:* aby uzyskane wyniki były reprezentatywne liczebności te powinny być wielokrotnie wyższe).

I region					II region				
110	106	114	120	112	109	126	110	109	115
116	109	109	118		112	103	109	114	121
115	119	121	126		119	116	122	117	110
107	123	114	117		108	118	117	115	114
121	125	116	108		112	120	112	119	113
117	118	110	115		119	116	110	108	118

Zadanie polega na porównaniu wzrostu dzieci w obu grupach. Przy wyborze właściwego testu statystycznego można założyć, że wzrost jest cechą o rozkładzie normalnym.

Zadanie 10

W celu potwierdzenia ewentualnej zależności między dwiema zmiennymi losowymi dokonano serii 20 pomiarów, których wyniki są przedstawione poniżej:

X		*Y*	
10	42	14	38
7	34	9	32
9	23	16	27
14	12	19	17
16	23	19	28
23	26	27	32
43	18	37	19
21	17	26	19
35	25	31	27
23	32	29	33

Należy obliczyć wartość współczynnika korelacji liniowej oraz zbadać jego istotność statystyczną na poziomie 0,05.

Rozwiązania zadań wraz z krótkimi komentarzami

Przedstawione rozwiązania opierają się na wyliczeniach dokonanych za pomocą pakietu STATGRAPHICS. Wykorzystanie innych pakietów statystycznych powinno jednak prowadzić do identycznych lub bardzo zbliżonych wyników.

Zadanie 1

Zgodnie z wcześniej przedstawionymi wzorami ryzyko względne (RW) oraz iloraz szans (OR) wynoszą odpowiednio:

RW = (80/1000) : (180/9000) = 4

OR = (80×8820) : (180×920) = 4,26

Uzyskane wartości zarówno ryzyka względnego, jak i ilorazu szans wskazują na silną zależność między analizowanym czynnikiem ryzyka (opóźniony wzrost wewnątrzmaciczny płodu) a badanym zjawiskiem (objawy zagrożenia płodu).

Obliczając parametry określające wartość prognostyczną analizowanego czynnika ryzyka uzyskujemy wyniki:

pozytywna wartość predykcyjna 80/1000 = 0,08

negatywna wartość predykcyjna 1 – 180/9000 = 0,98

czułość 80/260 = 0,31

swoistość 8820/9740 = 0,91

Na postawie analizy uzyskanych wyników można stwierdzić, że istnieje silny związek między badanym czynnikiem ryzyka a badanym powikłaniem. Jednak czynniki prognostyczne, np. niska czułość, wskazują, że decyzja o wprowadzeniu programu prewencyjnego na podstawie obserwowanego opóźnienia wewnątrzmacicznego wzrostu płodu może być obarczone dużym błędem. Pozytywna wartość predykcyjna, wynosząca 0,08 wskazuje, że w wypadku 92% pacjentek z grupy wysokiego ryzyka płód niesłusznie zostanie uznany za zagrożony.

Zadanie 2

Analizowana zmienna ma charakter ilościowy (skala interwałowa). W analizie opisowej należy pamiętać o następujących parametrach:

wartość średnia	175,25
mediana	170,40
odchylenie standardowe	28,21
błąd standardowy średniej	5,33
wartość minimalna	112,8
wartość maksymalna	235,6
rozstęp	122,8

Zadanie 3

Wspomniane w zadaniu przedziały to:
68,3% – (147,04 ÷ 203,46)
95,5% – (118,83 ÷ 231,67)

Zadanie 4

Przedział ufności dla wartości średniej zmiennej z przykładu 2. powinniśmy obliczyć, wykorzystując następujące informacje:

wartość średnia	175,25
odchylenie standardowe	28,21
liczebność	28

Tak więc odpowiednie przedziały ufności są następujące:
przedział ufności na poziomie *p-value* = 0,05 – (164,31 ÷ 186,19)
przedział ufności na poziomie *p-value* = 0,01 – (160,48 ÷ 190,02)

Zadanie 5

Prezentacja graficzna przedstawionych w zadaniu danych powinna zawierać: histogram lub histogram rozkładu skumulowanego.

Zadanie 6

Właściwym testem oceniającym zależność między przedstawionymi w zadaniu zmiennymi jest test chi-kwadrat. Obliczona wartość testu chi-kwadrat w tym przypadku wynosi 13,823 (*p-value* = 0,0317). Przy 6 stopniach swobody, gdy wartość graniczna dla *p-value* = 0,05 wynosi 12,593, świadczy to o istotności różnicy rozkładów analizowanych zmiennych w obu grupach na poziomie istotności statystycznej *p-value* <0,05.

Zadanie 7

W zadaniu tym najkorzystniej jest wykorzystać test t-Studenta dla zmiennych powiązanych – jest to możliwe, gdyż badania były dwukrotnie

wykonywane u tych samych osób. Wartość średnia różnicy wynosi 32,3, co dla obliczonej statystyki t daje istotność na poziomie *p-value* <0,001.

Zadanie 8

W zadaniu tym badamy wartości zmiennej w dwóch różnych grupach dzieci. Należy podkreślić stosunkowo małą liczebność drugiej grupy. Jest to w badaniach medycznych sytuacja dość częsta – z różnych powodów (np. rzadkie występowanie choroby, ograniczone fundusze, względy etyczne) uzyskanie większej liczby informacji jest trudne, a mimo to chcemy ustosunkować się do ewentualnych różnic między badanymi grupami. Zastosowanie testu t-Studenta nie jest tutaj szczególnie wskazane, ponieważ można byłoby nam postawić zarzut, że nie sprawdziliśmy normalności rozkładów, co w tej sytuacji byłoby mało zasadne. Dlatego korzystniejszym jest zastosowanie testu nieparametrycznego, a w tym konkretnym przypadku testu Manna-Whitneya. Wartość obliczonej znormalizowanej statystyki Manna-Whitneya wynosi 3,37, co świadczy o istotnych różnicach między badanymi grupami na poziomie *p-value* <0,01.

Zadanie 9

Zadanie to jest klasycznym przykładem możliwości zastosowania w ocenie różnic między wartościami średnimi testu t-Studenta dla zmiennych nie powiązanych. Upoważnia nas do tego między innymi fakt, iż wiadomo że rozkład wzrostu w populacji jest rozkładem normalnym. Obliczona wartość statystyki t wynosi 0,74. Wskazuje to na konieczność przyjęcia hipotezy H_0 mówiącej o braku różnic między badanymi wartościami średnimi.

Zadanie 10

W przedstawionej sytuacji wartość współczynnika korelacji wynosi 0,95 (*p-value* <0,001), a współczynnik determinacji jest równy 0,9025. Wartość ta świadczy o bardzo dużej liniowej zależności między badanymi zmiennymi. Należy jednak podkreślić, że w badaniach medycznych wartości tak duże obserwuje się bardzo rzadko (prawdopodobnie większość zależności jest bardziej skomplikowana – krzywoliniowa). Istotne jest również to, że ocena analizowanej prawidłowo zależności była badana próba powinna być odpowiednio liczna – wielu autorów przyjmuje, że większa od 30.

Literatura wybrana

1. Armitage P.: *Metody statystyczne w badaniach medycznych*. PZWL, Warszawa 1978.

2. Blalock H.: *Statystyka dla socjologów*. PWN, Warszawa 1977.

3. Brzeziński J., Stachowski R.: *Zastosowanie analizy wariancji w eksperymentalnych badaniach psychologicznych*. PWN, Warszawa, 1984.

4. Hermansen M.: *Biostatistics – some basic concepts*. Caduceus Medical Publishers, New York 1990.

5. Hollander M., Wolfe D.A.: *Nonparametric statistics*. John Wiley & Sons, New York 1973.

6. Moczko J.: *Automatyzacja obliczeń statystycznych* (rozdział w monografii pod redakcją W. Wajsa, wydawnictwa AGH, Kraków – przygotowane do druku).

7. Motulsky H.: *Intuitive Biostatistics*. Oxford University Press, New York 1995.

8. Rosner B.: *Fundamentals of Biostatistics*. PWS, Massachusetts 1986.

9. Siegel S.: *Nonparametric Statistics for Behavioral Sciences*. Prentice Hall, 1956.

Indeks

R

S

Ś

T

W

Z