



## JAK SKUTECZNIE WYKORZYSTYWAĆ METODY STATYSTYCZNE W PLANOWANIU I PRZEPROWADZANIU EKSPERYMENTU NAUKOWEGO?

*Cezary Watała, Uniwersytet Medyczny w Łodzi, Zakład Zaburzeń Krzepnięcia Krwi KDL;  
Uniwersytecki Szpital Kliniczny nr 2 im. WAM*

Niewłaściwie dobrane lub zastosowane metody analizy statystycznej wyników badania naukowego mogą prowadzić do budowania nieprawdziwych wniosków płynących z takiej analizy. Zły dobór metod statystycznych może dotyczyć zarówno postępowania analitycznego *a priori*, jak i *a posteriori*. Z jednej strony fakt, że błędy są popełniane już na wstępnym etapie planowania eksperymentu naukowego (w praktyce jest to najczęściej po prostu nieplanowanie doświadczeń od strony statystycznej) urąga dobrem obyczajom w pracy naukowej oraz prowadzi do trudności lub nawet niemożności poprawnego opracowania wyników badania, a często także do zwielokrotnienia jego kosztów. Z drugiej strony, gdy zebrane już wyniki analizujemy przy zastosowaniu niewłaściwie dobranych metod statystycznych – wnioski naszych badań mogą być wypaczeniem lub przekłamaniem rzeczywistości (np. stwierdzając zależności nieistniejące w rzeczywistości).

### **Algorytm przeprowadzenia eksperymentu naukowego – spojrzenie statystyka**

Schemat postępowania w planowaniu doświadczenia nie jest dowolny. Przystępując do wykonania eksperymentu naukowego w myśl sprawdzenia pomysłu, koncepcji czy hipotezy, musimy wiedzieć, jak brzmi ta hipoteza i czego mamy bronić lub co obalać. Przypadkowe i szeroko zakrojone zbieranie wyników, a następnie porządkowanie ich w różnych konfiguracjach w celu sprawdzenia zależności statystycznych czy różnic między nimi jest nieporozumieniem i przeczy racjonalnemu podejściu w pracy naukowej. Wyniki doświadczeń naukowych zbieramy zawsze i opracowujemy z myślą o udowodnieniu postawionej wcześniej hipotezy, a nie odwrotnie. Próby „wykrojania” koncepcji badawczej na podstawie zebranych wcześniej (najczęściej w sposób chaotyczny, gdyż bez wcześniejszego zamysłu celowego działania) danych pomiarowych jest działaniem po omacku i może sprawiać wrażenie manipulacji naukowej. W takim rozumieniu idei pracy naukowej przez doświadczonych badaczy, wpojenie sobie nawyku właściwego planowania pracy doświadczalnej jest podstawą uczciwej i wiarygodnej działalności naukowej. Oczywiście



konieczność budowania schematu czy planu naukowego nie wyklucza sytuacji, gdy przypadkowo zgromadzimy dane doświadczalne pozwalające na udowodnienie jakiejś koncepcji, z której wcześniej nie zdawaliśmy sobie sprawy. Przykładem takiej sytuacji mogą być doniesienia kazuistyczne, których koncepcje powstają w oparciu o zbierane na gorąco obserwacje. Algorytm właściwego zaplanowania eksperymentu można przedstawić następująco:

- ◆ Określenie problemu badawczego można sprowadzić do próby odpowiedzi na następujące pytania:

Co nas interesuje i czego jeszcze nie wiadomo? Co chcielibyśmy udowodnić? Niemożliwe jest uzyskanie nowej wiedzy na podstawie badań bez sprecyzowania tego co chcemy zbadać - problem badawczy precyzujemy nie na podstawie tego, co możemy zmierzyć, mając do dyspozycji określony warsztat aparaturowy oraz zaplecze finansowe, lecz na podstawie tego, co nas intryguje, co jest napędzane naszą pasją badawczą, która steruje naszym rozwojem naukowym. To właśnie ta pasja pcha nas w kierunku stawiania sobie kolejnych przyczynkowych zapytań w oparciu o gromadzoną wiedzę o problemie/przedmiocie badań. Wszystko to uzupełniane jest przez gromadzone z upływem czasu doświadczenie badawcze. Nasza rosnąca kompetencja badawcza zwiększa szanse sukcesu naukowego, a tym samym jest motorem polepszania warunków realizacji badań. Pamiętajmy jednak, że nawet stworzenie najbardziej dogodnych warunków ekonomicznych nie jest samo w sobie w stanie doprowadzić do wykreowania doświadczonego pracownika naukowego.

- ◆ Sprecyzowanie, jak brzmi hipoteza badawcza: co chcemy sprawdzić i udowodnić lub na jakie pytanie/a odpowiedzieć?

Naszym podstawowym zadaniem jest znalezienie sposobu, jak z ogólnej hipotezy badawczej wykreować weryfikowalne matematycznie hipotezy statystyczne (zob. poniżej).

- ◆ Wybranie właściwego testu statystycznego

Aby z powodzeniem pokonać ten etap powinniśmy przed przystąpieniem do każdego badania naukowego postawić sobie następujące pytanie: Czy model badawczy przystaje do modelu opracowania statystycznego wyników? Zastanawianie się, w jaki sposób będziemy analizować dane jeszcze przed ich zebraniem, jest bezwzględnie konieczne, ponieważ decyzja o tym, czy mamy do czynienia ze zmiennymi ciągłymi czy dyskretnymi, zależnymi czy niezależnymi itd., decyduje o tym, jaki test można wykorzystać do analizy statystycznej.

Pomimo dużej różnorodności wariantów metod statystycznych niekiedy zdarza się, że układ doświadczalny nie spełnia idealnie warunków procedury statystycznej. Dotyczy to np. modeli analizy wariancji, porównań wielokrotnych, porównań wielu grup z powtórzeniami itp. Niekiedy niepoprawny schemat wykonania doświadczenia może wręcz wykluczać poprawne użycie jakiegokolwiek dostępnego testu oraz narzuca konieczność budowania własnego indywidualnego modelu testu dostosowanego do



analizy własnego układu eksperymentalnego. Pomijając fakt, że nie zawsze czujemy się kompetentni do wykonania takiej modyfikacji matematycznej, jest to niewygodne i niepraktyczne rozwiązanie. Ponieważ naturalnie o wiele prościej jest korzystać z już wcześniej opracowanych i sprawdzonych procedur statystycznych (stosowanych zgodnie z ich przeznaczeniem), zamiast wymyślać lub opracowywać własne (jedynie nielicznych na to stać!), najbardziej racjonalne wydaje się zebrać dane w taki sposób, aby można je było analizować właśnie za pomocą tych istniejących, gotowych i sprawdzonych metod statystycznych. To jeszcze jeden argument za tym, aby najpierw planować dobór metod statystycznych, a potem wykonywać doświadczenie.

Właściwe dobranie metody analizy statystycznej jest trudne i wymaga sporo doświadczenia i dobrej znajomości teoretycznych podstaw metod statystycznych stosowanych w badaniach naukowych. Przyjemną metodą pozyskiwania takiego doświadczenia, chociaż bardzo czasochłonną, może być przeglądanie przykładów opracowań statystycznych konkretnych wyników pracy naukowej w naukach medycznych.

- ◆ Właściwy wybór/dobór próby badanej.

Należy sobie odpowiedzieć na pytania: Jakie cechy powinna mieć właściwie dobrana grupa kontrolna? Kto ma ją stanowić, aby była to reprezentatywna próba? Co ma decydować o tym, że tę grupę nazwiemy kontrolną – jakie kryteria? Czy to, że jej przedstawiciele nie mają określonej choroby, czy to, że nie biorą określonego leku?

Ile pomiarów musimy wykonać, aby udowodnić słuszność hipotezy statystycznej? Czy jeżeli przebadamy dużą (w naszym przekonaniu) liczbę osobników możemy mieć pewność, że wiarygodnie wypowiemy się o braku lub występowaniu różnic istotnych statystycznie? Na czym opieramy swoją ocenę, jak duża powinna być grupa? Czy prowadzimy badania dopóty, dopóki wystarczy nam środków finansowych, czy wykonujemy ściśle określoną liczbę pomiarów? Czy takiej estymacji dokonujemy *a priori* czy *a posteriori*?

Do właściwej oceny pożądanej liczebności grupy służą metody estymacji liczebności próby badanej (zobacz niżej).

- ◆ Zebranie danych

W olbrzymiej większości badań w naukach przyrodniczych lub medycznych powinniśmy uwzględnić dwa podstawowe wymagania opracowania statystycznego:

Wymaganie 1: każdy pomiar jest niezależny od innych pomiarów, chyba że mamy do czynienia z próbami z powtórzeniami; wymaganie to stanowi założenie większości testów statystycznych; jeżeli nie jest spełnione, to istnieje ryzyko, że wynik wnioskowania statystycznego może być obarczony błędem.

Wymaganie 2: elementy próby badanej powinny być wybrane w sposób losowy; służą do tego metody randomizacji omówione w dużym skrócie poniżej.

W badaniach medycznych problem losowego doboru pacjentów/ochotników może stwarzać problemy w praktyce, np. możemy stanąć wobec dylematu, czy jest w ogóle możliwe, aby dobrać ochotników do grupy kontrolnej oraz pacjentów do grupy badanej



w sposób całkowicie losowy i jednocześnie uwzględnić wymagane podobieństwo grup pod względem np. wieku, skoro wiadomo, iż badana jednostka chorobowa pojawia się jedynie w określonych grupach wiekowych? Czy w porównywalnej pod względem wieku grupie kontrolnej możemy wtedy wykluczyć wystąpienie jakichkolwiek czynników, które mogłyby wpływać na badane przez nas parametry? Jak przeprowadzić randomizację, jeżeli pacjenci, którzy reprezentują interesującą nas grupę badawczą, spotykani są jedynie sporadycznie w materiale klinicznym?

- ◆ Zastosowanie właściwego testu statystycznego oraz zdecydowanie o wyniku doświadczenia

Pamiętając, że możliwe jest jedynie odrzucenie hipotezy zerowej (z określonym prawdopodobieństwem), ale nigdy udowodnienie jej prawdziwości, hipotezy statystyczne należy budować w ten sposób, aby informatywność wynikająca z ich odrzucenia była jak największa.

## **Budowanie i weryfikacja hipotez statystycznych - błędy statystyczne i co z nich wynika**

Omawiając zasady formułowania hipotez badawczych, należy rozróżnić dwie kwestie. Z jednej strony mówimy o hipotezie badawczej, która jest stwierdzeniem precyzującym istnienie jakiejś zależności, różnicy, mechanizmu funkcjonowania, prawdopodobieństwa zachodzenia procesu itp. Jest to jakby hipotetyczny scenariusz procesu biologicznego. Z drugiej strony mamy hipotezę w ujęciu statystycznym, która sprowadza się do potwierdzenia równości/nierówności matematycznej. Pojedyncza hipoteza statystyczna dotyczy fragmentu hipotezy badawczej; stąd każdą koncepcję badawczą można sprowadzić do kilku/kilkunastu hipotez statystycznych – każda z nich będzie rewidowała słuszność pojedynczych porównań.

Formuła stawiania hipotez statystycznych jest ustalona – nie ma tutaj dużej dowolności, jakie powinno być brzmienie hipotezy zerowej, a jakie hipotezy alternatywnej. Wynika to z faktu, że możliwe jest jedynie odrzucenie hipotezy zerowej (z określonym prawdopodobieństwem), ale nigdy udowodnienie jej prawdziwości. Przyjęło się, że hipotezy statystyczne zestawia się parami w taki sposób, aby hipoteza podstawowa (tzw. zerowa, zakładająca niewystępowanie różnic,  $\mu_1 = \mu_2$ ) i przeciwstawna do niej hipoteza alternatywna (zakładająca występowanie różnic,  $\mu_1 \neq \mu_2$ ) wzajemnie się wykluczały.

Zasadą udowadniania prawdziwości nierówności  $\mu_1 \neq \mu_2$  przy użyciu testu statystycznego jest obliczanie tzw. statystyki testu w oparciu o zebrane dane pomiarowe. Jeżeli statystyka porównania dwóch średnich jest równa zero, to oznacza to, że dwie średnie są identyczne. Im bardziej wartość testu odbiega od wartości 0, tym większe jest prawdopodobieństwo, że średnie różnią się istotnie od siebie w sposób nieprzypadkowy. Innymi słowy, im większa jest wartość obliczonej statystyki, tym mniejsze są szanse, że hipoteza zerowa jest prawdziwa oraz że obliczona różnica jest dziełem przypadku a nie prawidłowością. O prawdziwości czy fałszywości hipotez statystycznych możemy orzekać z określonym

prawdopodobieństwem mniej lub bardziej różnym od 1, a nigdy z zupełną pewnością. Jeżeli nie mamy podstaw do zaprzeczenia hipotezy, to nie może być ona odrzucona, ale nie oznacza to, że jest prawdziwa. Oznacza to, iż zakładamy możliwość pomyłki: błędnego odrzucenia „prawdziwej” hipotezy zerowej lub błędnego przyjęcia „fałszywej” hipotezy zerowej. Ryzyko takiej pomyłki, zdefiniowane jako prawdopodobieństwo jej popełnienia, określa wartości dwóch błędów statystycznych testowania hipotez. Jeżeli mylnie odrzucamy prawdziwą hipotezę zerową, to popełniamy błąd I rodzaju (błąd  $\alpha$ ), jeżeli zaś mylnie nie odrzucamy fałszywej hipotezy zerowej, to popełniamy błąd statystyczny II rodzaju (błąd  $\beta$ ). Pamiętajmy, że istotność wyniku testu statystycznego to prawdopodobieństwo popełnienia błędu  $\alpha$ , zaś prawdopodobieństwo odrzucenia fałszywej hipotezy zerowej to moc testu (rys. 1).

		świat realny	
		$H_0$ jest prawdziwa	$H_0$ jest fałszywa
wynik testu	odrzuć $H_0$	<b>błąd I rodzaju</b> (prawdopodobieństwo = istotność)	<b>wniosek słuszny</b> (prawdopodobieństwo = moc testu)
	nie odrzuć $H_0$	<b>wniosek słuszny</b> (prawdopodobieństwo = 1 - istotność)	<b>błąd II rodzaju</b> (prawdopodobieństwo = 1 - moc testu)

Rys. 1. Zasada testowania hipotez statystycznych oraz definicja istotności i mocy statystycznej testu.

Obrazuje to, dlaczego staramy się wybierać zawsze testy o możliwie największej mocy – właśnie po to, aby zminimalizować ryzyko przyjęcia „fałszywej” hipotezy zerowej. Silne testy prowadzą nas pewniej do wiarygodnego odrzucenia nieprawdziwej hipotezy zerowej, o ile testowana różnica naprawdę istnieje.

## Planowanie i schemat doświadczenia – użyteczność kwadratów łacińskich i grecko-łacińskich

W przypadku stosowania analizy wariancji trudności początkującym badaczom może nasręczać właściwe zdefiniowanie czynników. W najprostszym układzie, z jednym czynnikiem (jedną zmienną grupującą), mówimy o ANOVA w klasyfikacji pojedynczej lub jednoczynnikowej. Jeżeli czynników jest więcej (w praktyce niechętnie przekraczamy liczbę 2), poza mocą dyskryminatywną każdego z czynników musimy zwracać uwagę na występowanie interakcji między czynnikami. Innymi słowy, przyglądamy się, czy każdy z czynników pozostaje niezależny. Jeżeli występują istotne interakcje między badanymi czynnikami (zmiennymi grupującymi), to informatywność analizy oczywiście maleje, gdyż żadnemu nie możemy przypisać niezależnego wpływu na badany parametr. W praktyce liczba czynników może niepokojąco rosnąć w przypadkach, gdy układ doświadczalny staje się bardziej złożony, zarówno od strony koncepcyjnej, jak i technicznej strony wykonania doświadczenia. Na przykład złożone badanie wpływu kilku substancji czynnych na



określony rodzaj komórek jest niekiedy trudno przeprowadzić w oparciu o materiał pochodzący od jednego pacjenta/ochotnika i nie ma możliwości, aby w pożądanym ramach czasowych wykonał to jeden laborant. Złożony układ doświadczalny może rodzić ryzyko niepożądanych interakcji między różnymi realnie występującymi czynnikami o potencjalnym wpływie na wyniki doświadczenia. Szczęśliwie w sytuacjach takich możemy niekiedy pominąć efekty interakcji. Sytuacja taka występuje dość często w praktyce, wówczas gdy na przykład (1) nie możemy przeprowadzić pełnego układu doświadczalnego z przyczyn ekonomicznych, lub też gdy (2) wiemy, że w danej populacji efekt interakcji jest na tyle nieistotny dla dalszej części naszego wnioskowania, że możemy go pominąć. Na przykład planujemy przeprowadzenie badania, w którym chcemy sprawdzić skuteczność 4 różnych antagonistów receptora dla fibrynogenu w hamowaniu agregacji płytek krwi. Ponieważ od każdego dawcy możemy uzyskać najwyżej 10 ml krwi pełnej, ilość materiału niezbędnego do wykonania pełnego badania wyklucza jego przeprowadzenie w pełnej wersji u każdego z dawców. Do badania wykorzystamy więc krew uzyskaną od 4 niezależnych dawców. Badania należy przeprowadzić jak najszybciej po pobraniu krwi, toteż badania wszystkich antagonistów nie jest w stanie przeprowadzić w tym samym czasie jeden laborant. Nasz plan zakłada, że każdy z 4 wyznaczonych techników laboratoryjnych będzie w danym czasie badał wpływ jednego antagonisty, wykorzystując do tego krew pobraną od jednego pacjenta.

Pełne doświadczenie zestawione według takiego hierarchicznego układu z uwzględnieniem wszystkich czynników (rodzaj antagonisty, dawca krwi, laborant), tzn. takie w którym każda kombinacja laboranta, antagonisty oraz dawcy krwi pojawia się przynajmniej jeden raz wymagałaby  $4 \times 4 \times 4 = 64$  grup. Jednakże możemy nie mieć środków ani czasu, aby przeprowadzić próby we wszystkich kombinacjach, a ponadto wydaje się mało prawdopodobne, aby np. osoba laboranta występowała w interakcji z dawcą krwi lub rodzajem antagonisty w stopniu, który mógłby mieć jakieś racjonalne praktyczne znaczenie. Biorąc to pod uwagę, moglibyśmy w rzeczywistości zrealizować jedynie tzw. układ kwadratu łacińskiego obejmującego 4 rodzaje antagonistów (A, B, C i D) i 16 osobnych grup badanych (rys. 2).

	dawca krwi			
	1	2	3	4
<i>laborant 1</i>	A	B	C	D
<i>laborant 2</i>	B	C	D	A
<i>laborant 3</i>	C	D	A	B
<i>laborant 4</i>	D	A	B	C

Rys. 2. Schemat kwadratu łacińskiego. Literami A, B, C i D oznaczono poszczególne grupy czynnika zdefiniowanego jako rodzaj antagonisty. Grupy czynnika „laborant” rozmieszczono w wierszach, grupy czynnika „dawca krwi” – w kolumnach kwadratu.

Widzimy, że układ ten jest układem hierarchicznym niekompletnym w tym sensie, że nie wszystkie kombinacje grup dla poszczególnych czynników są uwzględnione w modelu. Na przykład laborant 1 będzie badał płytki krwi od dawcy 1 z dodatkiem antagonisty A,

podczas gdy laborant 3 będzie badał krew od tego samego dawcy z dodatkiem antagonisty C. Co więcej, poszczególne grupy czynnika zdefiniowanego jako rodzaj antagonisty (A, B, C i D) są rozmieszczane w przypadkowy sposób w macierzy wyznaczonej przez czynniki dawca krwi i laborant. Podobne rozwiązania są bardzo często stosowane w praktyce planowania badań w sytuacjach, gdy niektóre efekty interakcji możemy pominąć bez szkody dla wyników analizy.

Innym przykładem zastosowania tej metody w zaplanowaniu doświadczenia może być sytuacja, w której pragniemy zbadać, jaka jest skuteczność wabienia korników do przynęt zawierających 7 różnych testowanych związków chemicznych (właściwa zmienna grupująca) rozlokowanych na 7 różnych przecinkach leśnych (kolumny) przez 7 kolejnych dni (rzędy). Analizujemy 49 różnych układów zamiast  $7 \times 7 \times 7 = 343$ .

	przecinka leśna						
	1	2	3	4	5	6	7
<i>dzień 1</i>	A7	B6	C5	D4	E3	F2	G1
<i>dzień 2</i>	G6	A5	B4	C3	D2	E1	F7
<i>dzień 3</i>	F5	G4	A3	B2	C1	D7	E6
<i>dzień 4</i>	E4	F3	G2	A1	B7	C6	D5
<i>dzień 5</i>	D3	E2	F1	G7	A6	B5	C4
<i>dzień 6</i>	C2	D1	E7	F6	G5	A4	B3
<i>dzień 7</i>	B1	C7	D6	E5	F4	G3	A2

Rys. 3. Schemat kwadratu grecko-łacińskiego. Literami A, B, C, D, E i F oznaczono poszczególne rodzaje testowanych feromonów, cyframi 1-7 – poszczególne stężenia każdego z atraktantów. Kolumny 1-7 oznaczają przecinki leśne, gdzie wystawiono próbki feromonów wabiących, w rzędach umieszczono kolejne dni prowadzenia badań.

Bardziej zaawansowaną formą kwadratów łacińskich są tzw. kwadraty grecko-łacińskie, gdzie analizujemy wpływ dwóch istotnych dla nas czynników (zmiennych grupujących) oraz dwóch dodatkowych czynników, których efekty interakcji decydujemy się pominąć. Kwadrat grecko-łaciński można byłoby np. wykorzystać w sytuacji, gdy chcemy zbadać skuteczność 7 różnych feromonów wabiących korniki, każdy w 7 różnych stężeniach, rozlokowanych na 7 różnych przecinkach leśnych (kolumny) przez 7 kolejnych dni (rzędy) (rys. 3). Analizowalibyśmy 49 różnych układów zamiast 7 (feromony: A, B, C, D, E i F) x 7 (stężenia: 1-7) x 7 (przecinki, kolumny) x 7 (dni, rzędy) = 2401. Cóż za oszczędność czasu i pieniędzy!

## Jak dobrać istotność testu? Prosta weryfikacja czy eksploracja statystyczna?

Zastanawiając się nad tym, czym właściwie jest istotność testu statystycznego, zauważamy, że większość nowoczesnych obszernych pakietów statystycznych oferuje nam nie tylko prostą odpowiedź zero-jedynkową TAK/NIE przy weryfikacji prawdziwości hipotezy



zerowej, lecz pozwala na dogłębniejszą eksplorację i dokładniejszą ocenę prawdopodobieństwa, z jakim możemy popełnić błąd, odrzucając prawdziwą hipotezę zerową. Tradycyjnie zwykło się przyjmować, przynajmniej w naukach przyrodniczych czy biomedycznych, iż  $p < 0.05$  stanowi wystarczające kryterium uprawniające do odrzucenia hipotezy zerowej. Zauważmy jednak, iż przyjmowana wartość  $p$  (co ważne i zgodne z regułami „sztuki statystycznej”: przyjmowana w oparciu o nasze założenia wstępne, koniecznie przed weryfikacją statystyczną i wykonaniem obliczeń, nie zaś po wyliczeniu wartości statystyki testu i dopasowaniu tablicowego „ $p$ ”) ma określony sens statystyczny. Oznacza, że godzimy się z faktem, iż na 100 podejmowanych decyzji odrzucenia „nieprawdziwej” hipotezy zerowej w 5 przypadkach popełnimy błąd statystyczny I rodzaju, czyli odrzucimy hipotezę, która niekoniecznie jest nieprawdziwa. Powinniśmy mieć świadomość, że to co dobre w niektórych rodzajach badań w naukach przyrodniczych, nie jest często do zaakceptowania np. w badaniach medycznych. Bo czyż łatwo pogodzimy się z 5 mylnymi diagnozami na każde 100 podejmowanych decyzji klinicznych? Na pewno nie. Toteż zwłaszcza w badaniach medycznych tak duże znaczenie ma podejście eksploracyjne, nie zero-jedynkowe, lecz właśnie analogowe, wyznaczające dokładną wartość prawdopodobieństwa niesłusznego odrzucenia prawdziwej hipotezy zerowej, czyli wykrycia różnic (wywołanych np. przez toczący się proces patologiczny), tam gdzie w rzeczywistości one nie występują.

## Test jednostronny czy obustronny?

Nasze pytanie o różnicę między wynikiem a wartością hipotetyczną można postawić w dwojaki sposób. Możemy zapytać, czy istnieje w ogóle jakakolwiek istotna różnica – w górę lub w dół – nieważne, czy nasz wynik będzie mniejszy albo większy od teoretycznej przyjętej *a priori* wartości, ale powinien być od niej różny. W takim przypadku pomijamy znak obliczonej wartości statystyki testu, ponieważ oczekujemy, iż w przypadku istotnej różnicy wynik ten będzie położony albo na lewym (gdy będzie mniejszy od teoretycznej wartości  $\mu=0$ ) albo na prawym (wtedy gdy będzie większy od hipotetycznej wartości  $\mu=0$ ) krańcu rozkładu, w jego najbardziej peryferyjnych regionach nieobjętych tym obszarem pola pod krzywą, który odpowiada przyjętemu przez nas prawdopodobieństwu (np. dla wartości prawdopodobieństwa 95% to „resztkowe” pole na peryferiach rozkładu będzie wynosiło 5%, czyli po 2.5% po każdej stronie wartości średniej umieszczonej centralnie). W takim przypadku mamy do czynienia z **testem obustronnym** – niezależnie od tego, w którym obszarze istotności statystycznej symetrycznego rozkładu – prawym czy lewym – znajdzie się wynik, nazwiemy go wynikiem istotnie statystycznie różnym od wartości hipotetycznej. Test obustronny wybieramy w sytuacjach, gdy nie mamy wystarczającej wiedzy o badanym zjawisku, a w szczególności, gdy nie znamy kierunku oczekiwanych zmian. Test weryfikuje, czy jest prawdopodobne występowanie różnic w badanej zbiorowości w odniesieniu do grupy referencyjnej, np. grupy kontrolnej.

Z drugiej strony nasze pytanie może być bardziej konkretne, kiedy na przykład pragniemy wykazać, że nasz wynik jest istotnie wyższy od przyjętej wartości hipotetycznej. W takiej



sytuacji oczekujemy, że nasz istotnie różny wynik znajdzie się konkretnie w prawym obszarze istotności statystycznej. Typ testu, który weryfikuje powyższe złożenie nazywamy **testem jednostronnym**. Test jednostronny znajduje zastosowanie, gdy nasza znajomość badanego zjawiska/procesu wystarcza do określenia, jakich zmian powinniśmy się spodziewać. W rzeczywistości z takimi właśnie przypadkami spotykamy się najczęściej. Z reguły dobrze wiemy, jakich zmian, zgodnie z racjonalnym postrzeganiem mechanizmu badanego zjawiska, winniśmy oczekiwać. Rejestrowany odmienny kierunek zmian może być dla nas wręcz źródłem zaniepokojenia, a nie zadowolenia, iż jakiegokolwiek zmiany są postrzegane.

Wbrew temu, co niekiedy zakładają mało doświadczeni badacze, wybierając (świadomie) test jednostronny, nie jest naszym zamierzeniem przyporządkowanie określonym wartościom statystyki testu mniejszych wartości istotności statystycznej, czyli większego prawdopodobieństwa, iż poprawnie odrzuciliśmy nieprawdziwą hipotezę zerową, lecz zbadanie czy rejestrowany trend/kierunek zmian jest zgodny z naszymi oczekiwaniami oraz czy jest statystycznie istotny w stosunku do naszej grupy referencyjnej (np. gdy brak czynnika indukującego zmiany). Przy dobieraniu testu do naszych potrzeb (jednostronny lub obustronny) powinniśmy się zatem kierować naszą aprioryczną wiedzą o badanym zjawisku/procesie, w oparciu o racjonalne przesłanki doświadczenia, nie zaś dokonywać wyboru na podstawie zadowalającej nas wartości poziomu istotności, kierując się perspektywą wykazywania wyższych istotności różnic. Ta ostatnia możliwość jest kusząca, ale należy pamiętać, że wybór taki wiąże się także z większym ryzykiem niesłusznie odrzuconej hipotezy zerowej oraz fałszywego wnioskowania o występowaniu efektu. Na przykład, badając skuteczność leku nasennego na wydłużenie snu, zakładamy *a priori*, że ochotnicy przyjmujący ten lek będą spali dłużej niż ochotnicy otrzymujący *placebo*, a nie że osoby w obu grupach będą przesypany różną ilość czasu. Uwzględnienie w analizie takiego badania także lewostronnego obszaru istotności (tzn. wpływu leku na skrócenie czasu snu) podważa bowiem w ogóle sens interesowania się tym lekiem jako środkiem nasennym. Naszym jedynym racjonalnym wyborem będzie więc tutaj test jednostronny. Tak samo jest zresztą z olbrzymią większością zastosowań testu sparowanego – *a priori* zakładamy występowanie jakiegoś ukierunkowanego efektu. Inaczej jeżeli porównujemy określony parametr u pacjentów reprezentujących różne jednostki chorobowe: *a priori* nie zawsze możemy przewidzieć kierunek różnic.

## Randomizacja – na czym polega i jakie z niej mamy korzyści?

Jednym z podstawowych oczekiwań i naturalnych zachowań badacza jest chęć ekstrapolacji uzyskanych wyników z przebadanej grupy na ogólne wnioski dotyczące większej populacji. Najlepszym rozwiązaniem byłoby przebadanie całej (lub olbrzymiej większości elementów) populacji, ale z wielu względów praktycznych zadanie to jest niemożliwe do przeprowadzenia, a dodatkowe czynniki natury logicznej i rozumowej mogą nas do tego zniechęcić. Wszystkie pozostałe metody są metodami przybliżającymi, zaokrąglającymi, przyjmującymi pewne założenia. Randomizacja pozwala lub ułatwia nam sprostanie jednemu z podstawowych założeń naszego badania, jakim jest reprezentatywności próby wybranej do przebadania i analizy. Wnioski wyciągane z badań, w których nie zachowano



odpowiedniej randomizacji w doborze elementów tej próby, nie mają oczywiście bardziej uniwersalnego znaczenia i powinny być ograniczone wyłącznie do przebadanej grupy, nie można ich uogólniać i ekstrapolować na większe i/lub inne grupy. W przeciwnym razie wyciągane wnioski mogą być błędne. Termin „randomizacja” oznacza przypadkowy, losowy – a więc obiektywny – dobór badanych osób, obiektów, komórek (*random* - ang. przypadkowy, losowy). Przeprowadzenie (*randomness*) zakłada (z definicji) całkowity brak jakiegokolwiek schematyczności oraz możliwości przewidzenia wyniku.

Zauważmy, że chociaż przypadkowość „rządzi się” prawami teorii prawdopodobieństwa, w swojej definicji jest całkowitym zaprzeczeniem celów, jakie ta teoria sobie wyznacza. Randomizacja jest przeprowadzana albo w celu wyboru konkretnych osób/obiektów do badań, albo w celu przyporządkowania badanych obiektów do odpowiedniej procedury diagnostycznej, doświadczalnej, leczniczej itp. Randomizacja może być przeprowadzona wieloma różnymi metodami, np. przez program komputerowy, na zasadzie wyciągania z worka różnokolorowych kulek, rzutu monetą, przygotowania zaklejonych i nieopisanych kopert z informacją o podaniu leku lub placebo, lub na „chybił trafił”. W przypadku dużych grup badanych dobór losowy jest zastępowany badaniem kolejnych pacjentów (tzw. *consecutive cases analysis*). Dobrym praktycznym sprawdzianem tego, czy zebrane dane pochodzą od losowo wybranych elementów próby (pacjentów, ochotników, zwierząt laboratoryjnych, próbek produktu w procesie kontroli jakości, itp.) jest przyjmowanie przez nie rozkładu normalnego.

Przykłady:

1. Wybrano w sposób losowy 100 szczurów Wistar z większej grupy ponad 300 zwierząt oznakowanych numerami od 100 do 412. Spośród wybranych do doświadczenia zwierząt 50 szczurów przydzielono do grupy szczurów, u których wywoływano cukrzycę doświadczalną na drodze iniekcji streptozotocyny.

2. Sto pięćdziesiąt kolejnych osób z chorobą niedokrwinną serca przydzielono w sposób losowy do grupy otrzymującej kwas acetylosalicylowy doustnie w dawkach powtarzanych co 24 godziny lub jego pochodną lizynową we wlewie dożylnym. Dobór do odpowiedniej z grup badanych przeprowadzono w oparciu o losowe wyciąganie czarnej lub białej kulki z worka, w którym przed badaniem umieszczono po 75 kul białych i 75 czarnych.

Randomizacja może być prosta (*simple randomization*), jeżeli polega na wygenerowaniu ciągu liczb losowych przyporządkowanych w umowny sposób elementom próby badanej, lub ograniczona (*restricted randomization*), jeżeli zależy nam na osiągnięciu zrównoważenia grup pod względem ich rozmiaru oraz podstawowej charakterystyki wyjściowej. Dwie podstawowe metody służące zrównoważeniu wielkości grup oraz wyrównaniu ich charakterystyki znane są jako blokowanie i stratyfikacja (warstwowanie). Podstawowe korzyści płynące ze stosowania randomizacji to zminimalizowanie dwóch podstawowych zagrożeń wiarygodności badania: (a) obciążenia (*bias*, rodzaj systematycznego błędu, który prowadzi do przekłamania wyniku testu lub zależności) oraz zmiennych zakłócających (*confounders*).

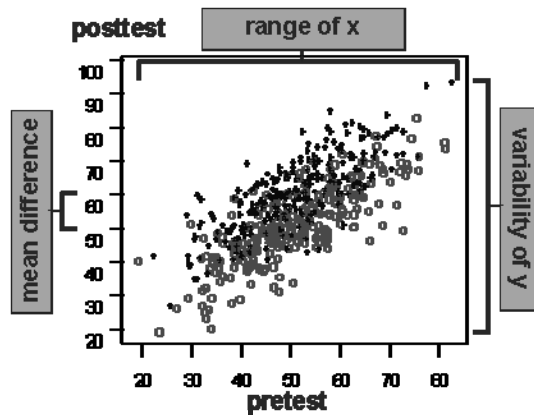


## Układ zrandomizowanych bloków

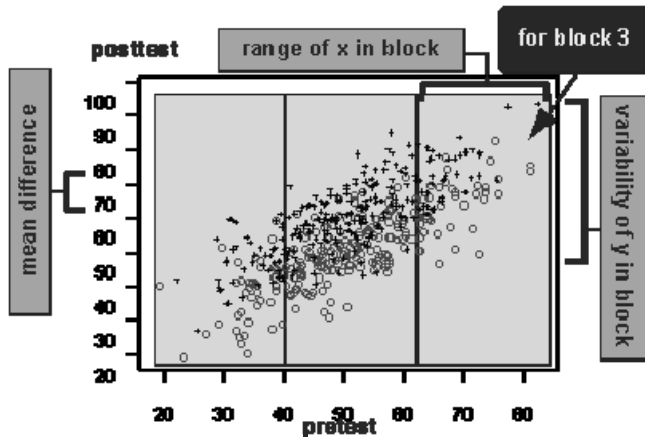
Układ zrandomizowanych bloków jest dla badacza odpowiednikiem randomizacji w warstwach. Podobnie jak w klasycznej metodzie randomizacji, celem stosowania takiego układu jest zmniejszenie szumu lub ograniczenie wewnątrzgrupowej zmienności próby. Jak posłużyć się tą techniką w praktyce? Badacz powinien podzielić badaną próbę na mniejsze podgrupy (bloki) o większej homogenności (jednorodności) niż cała grupa. Podgrupy te są odpowiednikami warstw w procedurze randomizacji. Następnie przeprowadzamy wewnątrz każdej z jednorodnych podgrup lub bloków taką samą analizę, jaka miała być przeprowadzona w całej grupie. Ideą takiego postępowania jest to, że zmienność rejestrowana w każdym bloku (podgrupie) jest mniejsza niż zmienność obserwowana w całej grupie (próbie). Skoro tak, to oczywiście wpływ badanego czynnika (czynników) jawi się o wiele bardziej wyraźnie w podgrupie (bloku) niż w całej grupie. Kiedy efekt ten zostanie „uwspólniony” dla wszystkich bloków, to oczywiście całkowity efekt będzie wyraźniejszy niż w sytuacji, gdybyśmy blokowania nie zastosowali. Widać to na prostym przykładzie przedstawionym na rys. 4. Załóżmy, że zamierzamy wykonać jedynie prostą randomizację *post hoc* wyników badania przeprowadzonego w grupie studentów. Można przypuszczać, że podgrupy studentów są względnie jednorodne w zakresie np. roku studiów. Pogrupujemy więc studentów w trzech blokach odpowiadających rocznikowi studiów. Jeżeli nasze założenie miałyby okazać się poprawne, powinniśmy oczekiwać, że zmienność w obrębie jednego roku jest mniejsza niż całkowita zmienność dla wszystkich studentów. Stąd też badany wpływ czynnika powinien być o wiele wyraźniej widoczny w każdym z bloków niż w całej grupie obejmującej wszystkich studentów. Przeprowadźmy zatem nasze doświadczenie (badanie wpływu testowanego czynnika) niezależnie w każdym z trzech wyróżnionych bloków (podgrup).

Zwróćmy uwagę na kilka charakterystycznych cech tej metody. Po pierwsze, postronny obserwator z zewnątrz może w ogóle nie zauważyć, że grupujemy studentów w blokach, ponieważ ten sam schemat postępowania doświadczalnego jest praktykowany w każdym z bloków. Nie ma zatem powodu, aby ludzi z jednego bloku oddzielać od uczestników badania zaszeregowanych do innych bloków. Podział na bloki istnieje jedynie w naszej wyobraźni, nie jest jednak usankcjonowany żadnymi innymi kryteriami stanowiącymi przedmiot samej analizy. Innymi słowy, blokowanie nie wpływa w żaden sposób na cokolwiek, co czynimy w doświadczeniu z elementami próby (studentami). Procedura blokowania (przydziału do podgrup) jest strategią grupowania studentów podczas analizy danych w celu zminimalizowania szumu, czyli wariancji wewnątrz grupy – jest typową strategią *post hoc*, gdyż dotyczy samej analizy. Z drugiej strony powinniśmy zauważyć, iż blokowanie będzie skuteczne jedynie wtedy, gdy nasz podział na podgrupy będzie zasadny, tzn. gdy zmienność w poszczególnych blokach będzie rzeczywiście mniejsza od całkowitej zmienności w grupie wszystkich badanych studentów. Jeżeli nasze założenia o większej jednorodności wewnątrz bloków nie będą słuszne, to procedura blokowania może *de facto* obniżyć moc dyskryminacji w badaniu wpływu interesującego nas czynnika.

a



b



Rys. 4. Wpływ procedury blokowania na stosunek wielkości obserwowanej zmienności do średniej zmiany w odpowiedzi na zadziałanie czynnika. Wartości przed działaniem badanego czynnika przedstawiono na osi odciętych (pretest), te zmierzone po zadziałaniu czynnika – na osi rzędnych (posttest). Symbolem „+” oznaczono przypadki dla grupy badanej, symbolem „o” – przypadki dla grupy referencyjnej (porównawczej); (a) bez procedury blokowania, (b) z wykorzystaniem blokowania.

W jaki sposób blokowanie według opisanego powyżej schematu może obniżyć szum i poprawić dyskryminację wpływu badanego czynnika? Przyjrzyjmy się wykresom na rys. 4. Przedstawiają one porównanie wyników analizy przeprowadzonej bez zastosowania techniki blokowania (analiza dla całej grupy) oraz wyników analizy wykonanej z wykorzystaniem blokowania (analiza w poszczególnych podgrupach).

Zauważmy, że dla poszczególnych wartości parametru mierzonych przed zadziałaniem czynnika (pretest) można zaobserwować wzrost po zadziałaniu czynnika (posttest), a średnia różnica wynosi około 10 jednostek. Teraz dzielimy całą grupę na trzy niezależne



jednorodne podgrupy (bloki) na podstawie wartości przed zadziałaniem czynnika (pretest). Zobaczymy, co się dzieje w trzecim bloku. Średnia różnica wciąż pozostaje taka sama (wzrost o około 10 jednostek w stosunku do wartości przed zadziałaniem czynnika, „pre-test”). Jednak cała zmienność wewnątrz bloku jest o wiele mniejsza niż dla całej niepodzielonej grupy. Pamiętajmy, że efekt działania czynnika określony jest wartością ilorazu sygnału do szumu. Nasz sygnał to średnia różnica po zadziałaniu czynnika. Nasz szum to zmienność w obrębie grupy. Rysunek pokazuje, że na drodze blokowania nie zmieniliśmy wartości sygnału, lecz obniżyliśmy znacząco zmienność wewnątrz grupy. Zatem efekt działania czynnika będzie się charakteryzował znacznie mniejszym szumem w stosunku do takiej samej wartości sygnału. Blokowanie poprawiło znacząco dyskryminację wpływu badanego czynnika, dzięki zwiększeniu jednorodności podgrup, w których ten efekt ocenialiśmy.

## **Prawdopodobieństwo a proporcja – potoczne znaczenie prawdopodobieństwa**

Te dwa pojęcia są często mylone w praktyce, dlatego warto przypomnieć ich definicje.

*Prawdopodobieństwo* to wielkość antycypowana *a priori*; nie w oparciu o naszą dokładną wiedzę o badanym zjawisku, które jeszcze nie zaistniało, które dopiero ma się zdarzyć, a my dopiero zarejestrujemy jego wynik. Tak rozumiane prawdopodobieństwo może być obliczone jedynie dla przypadkowych sekwencji zdarzeń. W myśl popperowskiego świata skłonności prawdopodobieństwo jest równoważne skłonności do stania się, do wydarzenia się czegoś.

*Proporcja* natomiast jest obliczana na podstawie obserwacji *a posteriori*; dlatego jej wartość może być obliczona zarówno dla przypadkowych, jak i nieprzypadkowych sekwencji zdarzeń.

Aby odnieść do praktyki różnicę między tymi dwoma pojęciami, spróbujmy odpowiedzieć na pytanie: jakie jest „prawdopodobieństwo”, że palenie tytoniu przyczyni się do powstania nowotworu płuc? Gdybyśmy rozważali powyższy problem w kategoriach eksperymentu naukowego, to czy moglibyśmy sobie wyobrazić przypadkową sekwencję zdarzeń, z których niektóre przyczynią się do powstania choroby a inne nie? Czy byłoby możliwe pokierowanie przez badacza takim badaniem eksperymentalnym? Z pewnością nie. To, czego dociekamy w pytaniu, to nie „prawdopodobieństwo”, lecz proporcja, znana skądinąd z wyników badań retrospektywnych publikowanych w literaturze naukowej. To fragment wiedzy opartej na doświadczeniach z przeszłości, nie zaś antycypacja w naszej wyobraźni. Zatem w sensie heurystycznym lub idealistycznym „prawdopodobieństwo” może być rozumiane jako miara zaufania lub wiary w to, że pewne stwierdzenia są prawdziwe, a inne nie.



## Szacowanie wielkości próby badanej – po co i jak je przeprowadzamy?

Zasadniczym wymaganiem w planowaniu badań naukowych jest oszacowanie wielkości próby, jaką zamierzamy przebadać. Stosujemy je m.in. po to, aby nie zbierać niepotrzebnie dużej liczby danych w sytuacji, gdy: (a) dostrzegamy już na „pierwszy rzut oka”, że porównywane grupy różnią się między sobą, (b) nie występują rzeczywiste różnice i nie wykażemy ich niezależnie od liczebności próby, zbierając bardzo dużą liczbę powtórzeń mnożymy tylko niepotrzebnie koszty eksperymentu, podczas gdy moglibyśmy wykorzystać te środki na sprawdzenie innej koncepcji badawczej. Stosowanie estymacji właściwej liczebności próby powinno być nawykiem każdego rzetelnego badacza, a niewykorzystanie tej metody może być uważane za niekompetencję w prowadzeniu badań naukowych. Niestety doświadczenie uczy, że ocena liczebności grupy badanej przed wykonaniem badań jest bardzo rzadko stosowaną praktyką, a liczebność taka oceniana jest na czysto arbitralnych zasadach. Jest to praktyka uważana przez licznych badaczy za nieetyczną. Wykonując niepotrzebnie bardzo dużą liczbę powtórzeń, nie tylko mnożymy niepotrzebnie koszty eksperymentu, podczas gdy moglibyśmy wykorzystać te środki na sprawdzenie innej koncepcji badawczej. W badaniach klinicznych wiąże się to nie tylko z podawaniem większej liczbie osób *placebo*, ale także realnym opóźnieniem wprowadzania do praktyki klinicznej korzystnej strategii farmakologicznej.

W przypadku testów porównań dla zmiennych ciągłych metody estymacji liczebności próby opierają się na kilku założeniach:

- ♦ próby mają rozkład normalny - gdy liczebność próby bardzo wzrasta, wówczas średnie prób podlegają rozkładowi normalnemu nawet w sytuacji, gdy odpowiednia zmienna w populacji nie ma rozkładu normalnego lub nie jest wystarczająco dobrze zmierzona,
- ♦ musimy zdefiniować, z jakim prawdopodobieństwem pragniemy orzec o występowaniu lub braku różnic,
- ♦ estymowana liczebność zależy od mocy stosowanego testu, czyli musimy założyć, jak duże ryzyko błędu II rodzaju (prawdopodobieństwo nieodrzućcia hipotezy zerowej, gdy jest ona fałszywa) dopuszczamy.

W przypadku badań populacyjnych/epidemiologicznych (na liczebnościach grup) wymagane jest z reguły bardzo precyzyjne określenie, jakiego wyniku spodziewamy się po przeprowadzeniu badania. Na przykład w analizie porównania śmiertelności wśród niemowląt karmionych odżywką w stosunku do tych karmionych piersią samo stwierdzenie większego ryzyka nie zadowala nas – pragniemy jeszcze wiedzieć, ile razy ryzyko takie jest większe. Wielkość próby będzie na przykład inna w przypadku 4-krotnego i dwukrotnego ryzyka. Należy także pamiętać, że z uwagi na zmienność wyników (widoczną szczególnie wyraźnie w przypadku małych prób) obserwowany wzrost ryzyka może być za mały, aby wykazać jego istotność. Dlatego powinniśmy a priori określić prawdopodobieństwo, z jakim chcielibyśmy wnioskować o istotności różnic na danym poziomie istotności, czyli powinniśmy ustalić moc wnioskowania. W ten sposób możemy na przykład określić, że badanie dostarcza wartościowych wyników, jeżeli z prawdopodobieństwem 90% możemy stwierdzić, że ryzyko względne śmierci niemowląt karmionych butelką w stosunku do tych



karmionych piersią jest na przyjętym poziomie istotności (np. 5%) przynajmniej tak wysokie jak 2.

## Transformacje „surowych” danych – po co i jak je przeprowadzamy?

W praktyce dość często zdarza się, że zebrane przez nas obserwacje nie spełniają wymagań niezbędnych dla zastosowania testów i metod, które są szczególnie użyteczne, dogodne i których lubimy używać. Tak jest na przykład ze stosowaniem testu t Studenta – jest on tak popularny i chętnie wykorzystywany, że najczęściej nie sprawdzamy nawet, czy nie są naruszone warunki usprawiedliwiające jego zastosowanie. Dwa z takich przeciwwskazań, którym przypisuje się największe znaczenie, to naruszenie normalności rozkładu oraz niejednorodność wariancji.

Test t Studenta jest względnie odporny na naruszenie tych warunków, ale już analiza wariancji nie jest. Gdy mamy do czynienia ze zmiennymi o rozkładach ciągłych, niespełniających warunku normalności rozkładu, w przypadkach lewo- lub prawoskośnych rozkładów, nierównych wariancji porównywanych prób czy nieliniowych zależności między zmiennymi, korzystamy często z narzędzia transformacji „surowych” danych. Czy transformując matematycznie uzyskane wyniki, nie ingerujemy i nie wypaczamy poszukiwanych różnic, zależności itp.? Nie, o ile tej samej procedurze transformacyjnej (temu samemu działaniu matematycznemu) poddajemy obie (wszystkie) porównywane grupy. Zależnie od sytuacji i charakteru danych stosuje się różne przekształcenia matematyczne i nie ma tutaj zbyt dużej dowolności. Inne transformacje są użyteczne w przypadku rozkładów lewoskośnych, inne dla rozkładów prawoskośnych, inne w przypadku heteroscedastyczności zmiennych, jeszcze inne w różnych wariantach nieliniowych zależności między zmiennymi. Transformacja odwrotnej proporcjonalności jest na przykład silniejsza, zaś pierwiastkowa słabsza niż logarytmiczna i dlatego dobiera się je w zależności od stopnia skośności rozkładu. W przypadku danych procentowych lub proporcji (których rozkłady są raczej bardziej dwumianowe niż normalne, odstępstwa od normalności są szczególnie rażące dla niskich i wysokich %, tzn. 0-30% i 70-100%) stosuje się często transformacje arcus sinus. Dzięki procedurze transformacji doprowadzamy do zwiększania jednorodności wyników w porównywanych grupach, a skoro maleje zmienność wewnątrzgrupowa (szum) przy niezmiennionej wartości sygnału (średnia różnica między grupami, zmienność międzygrupowa), to oczywiście rośnie moc dyskryminatywna testów wykorzystywanych do badania istotności różnic. Ponieważ transformacja pomaga w normalizacji rozkładu, jej działanie można porównać do zabiegu zwiększania liczebności porównywanych grup.

## Dane odstające – jak zdecydować, co odrzucić a co zostawić?

Odstającymi nazywamy nietypowe (z definicji), niepasujące do innych, rzadko występujące obserwacje w próbie. Wierzmy, że odstające obserwacje są manifestacją losowego błędu, który chcielibyśmy kontrolować i eliminować częstość obserwacji odstających



i niepasujących do ogółu. Niestety nie jest znana żadna metoda sprawdzająca się przy automatycznym usuwaniu odstających obserwacji. Dlatego też jesteśmy zdani na analizę rozkładów pojedynczych zmiennych oraz wykresów rozrzutu dla par lub kilku zmiennych. Usuwanie zmiennych w oparciu o intuicyjne przeświadczenie o ich „inności” może graniczyć z manipulacją danymi, dlatego staramy się dobrać jak najbardziej obiektywne metody statystyczne i stosujemy często równolegle kilka technik weryfikacji ich „nie-dopasowania” do reszty danych. Z samej definicji odstających obserwacji wynika, że są to dane o skrajnych wartościach w monotonicznym szeregu obserwacji, obdarzone na tyle dużym błędem losowym, że nie mieszczą się w zakresie zmienności wyznaczonym przez pozostałe obserwacje próby. Ponieważ wiele czynników może być odpowiedzialnych za generowanie takich nietypowych wyników, bardzo pożądane jest zweryfikowanie przyczyn, które złożyły się na ten błąd. Próba wyeliminowania tego błędu przy powtarzaniu doświadczenia/pomiaru jest dla nas najlepszą weryfikacją występowania przypadkowości lub regularności w odstawianiu niektórych wyników. W tym miejscu warto zastanowić się nad przyczyną występowania takich regularnie odstających obserwacji. Czy przypadki te nie zasługują na naszą szczególną uwagę jako elementy unikalne, być może reprezentujące inną zbiorowość niż ta analizowana przez nas (np. pacjenci z rzadko spotykaną jednostką chorobową, inny gatunek/podgatunek o zachodzącym areale zasięgu)?

## Najczęściej popełniane błędy w statystycznej analizie danych

Ocenia się, że około 60% wszystkich prac oryginalnych publikowanych w zakresie nauk biomedycznych czy farmakologicznych zawiera błędy opracowania statystycznego danych (De Muth, 1999). Najczęściej spotykane błędy w medycznej literaturze naukowej obejmują:

- ◆ niewłaściwe planowanie doświadczenia i/lub sformułowanie hipotezy badawczej,
- ◆ stosowanie błędu standardowego zamiast odchylenia standardowego lub odwrotnie jako miary rozproszenia,
- ◆ testowania hipotez statystycznych i doboru testów parametrycznych oraz nieparametrycznych,
- ◆ błędy wynikające z niespełnienia warunków normalności rozkładu i/lub jednorodności wariancji,
- ◆ niepoprawnego oszacowania lub nieoszacowania właściwej wielkość próby badanej,
- ◆ niewłaściwego stosowania testów sparowanych i niesparowanych,
- ◆ niestosowania wielokrotnych testów  $t$  jako rozwinięcia metod analizy wariancji,
- ◆ niewłaściwego stosowania testu  $\chi^2$  i testu dokładnego Fishera.



## Literatura

1. Afifi, A.A. and Clark, V. (1990) *Computer-aided multivariate analysis*, 2nd Ed., New York-London-Melbourne, Van Nostrand Reinhold, ss. 1-463.
2. Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research*, 3rd Ed., Blackwell Science, Oxford-London-Edinburgh, ss. 1-630.
3. De Muth J.E. (1999) *Basic Statistics and Pharmaceutical Statistical Applications*. Marcel Dekker, Inc., New York-Basel, ss. 596.
4. Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D., Hearst, N., Newman, T.B. *Designing Clinical Research*. 2<sup>nd</sup> ed. Lippincott Williams & Wilkins, Philadelphia, 2001, ss. vii-336.
5. Katz, M.H. *Multivariable Analysis. A Practical Guide for Clinicians*. Cambridge University Press, Cambridge, 2001, ss. vii-192.
6. Lang, T.A., Secic, M. *How to Report Statistics in Medicine. Annotated Guidelines for Authors, Editors, and Reviewers*. ACP Series, Philadelphia, 1997, ss. vii-367.
7. Siegel, S. and Castellan, N.J., Jr. (1988) *Nonparametric statistics for the behavioral sciences*, 2nd Ed., New York, McGraw-Hill Book Company, ss. i-399.
8. Sokal, R.R. and Rohlf, F.J. (1981) *Biometry - The principles and practice of statistics in biological research*, 2nd Ed., San Francisco, W.H. Freeman & Co., ss. 1-862.
9. Stanisław, A. (2001) *Przystępny kurs statystyki w oparciu o program STATISTICA PL na przykładach z medycyny*, 2nd Ed., StatSoft Polska, Kraków.
10. Stanisław, A. (2000) *Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny. Tom II*, StatSoft Polska, Kraków.
11. Zar, J. (1999) *Biostatistical analysis*, 4th Ed., Prentice-Hall International, Inc. Simon & Schuster/A Viacom Company, Upper Saddle River, N.J., ss. 1-663.
12. Zieliński, T. (1999) *Jak pokochać statystykę, czyli STATISTICA do poduszki*, StatSoft Polska, Kraków.