



## JAK PLANOWAĆ DOŚWIADCZENIA NAUKOWE Z WYKORZYSTANIEM METOD STATYSTYCZNYCH? TESTOWANIE HIPOTEZ STATYSTYCZNYCH: MIĘDZY MOCĄ STATYSTYCZNĄ A NIEMOCĄ DECYZYJNĄ

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

Nasze badawcze ambicje z reguły przerastają dostępne możliwości poznania. Najczęściej zdarza się, że nie mamy możliwości lub woli badania całej populacji interesujących nas obiektów. Nie zmienia to jednak faktu, iż zwykle pragniemy wypowiedzieć się o dużej zbiorowości interesujących nas obiektów, a nie jedynie o fragmencie tej zbiorowości, którą naprawdę mogliśmy/zdołaliśmy przebadać. Tradycje myślenia indukcyjnego (wnioski o charakterze uniwersalnym na mocy dokonanych nielicznych szczegółowych badań) w nauce bywają jednak często także zwodnicze. Ryzykujemy, iż zgromadzone przez nas dane charakteryzują się niewielką/niewystarczającą reprezentatywnością w odniesieniu do ogólnej zbiorowości, a zatem nasz uogólniający, nadto uniwersalny wniosek może łatwo stać się nadużyciem. Pokusa takiego działania jest jednak na tyle wielka, iż skłonni jesteśmy najczęściej podjąć takie ryzyko. Przyczyny naszej decyzji bywają na ogół bardzo trywialne: cała zbiorowość jest zbyt liczna, aby w całości poddać ją analizie, lub też – jest zbyt kosztowne i czasochłonne, aby badaniem objąć obszerny fragment populacji generalnej. Dlatego też podejmujemy ryzyko wydawania bardziej ogólnych i uniwersalnych opinii na podstawie danych zebranych dla niewielkiego wycinka ogólnej zbiorowości. W sytuacjach takich powinniśmy jednak zatroszczyć się o to, by badane przez nas fragmenty dużej zbiorowości były reprezentatywne dla całej zbiorowości, o jakiej chcemy wygłaszać opinie. Właściwe procedury próbkowania oraz zasadne dobieranie liczebności próby, którą chcemy przebadać, to dwie podstawowe troski uczciwego badacza [1].

### **Dobieranie elementów badanej próby**

Na podstawie małego fragmentu populacji ogólnej staramy się oszacować tzw. parametry statystyczne badanej próby, w celu scharakteryzowania badanego parametru/zmiennej. Średnia i odchylenia standardowe to najbardziej typowe parametry charakteryzujące badaną populację. To jak bliskie są wartości estymowane tych parametrów prawdziwym wartościom charakteryzującym całą (ogólną) populację, wskazuje nam, jak dobrze nasza badana próba reprezentuje całą badaną populację. Załóżmy na przykład, że zamiarem



naszym jest ocena średniej masy ciała chłopców w wieku 15-17 lat, zamieszkujących pewne miasto liczące, powiedzmy, 100000 mieszkańców. Teoretyczna średnia  $\mu$  jest parametrem odzwierciedlającym uśrednioną masę ciała chłopców w wieku 15-17 lat zamieszkujących badane miasto. Zakładając, że całkowita populacja chłopców w tym wieku w badanym mieście liczy 17500 osób, zdecydujemy się na zbadanie losowej próby 100 chłopców, zamiast badać całą grupę 17500 chłopców. Jest oczywiste, że liczba chłopców, których zamierzamy zbadać ( $n$ ), będzie dość niewielka w porównaniu z całą dostępną populacją ( $N = 17500$ ). Gdy tylko wybierzemy naszą badaną próbę 100 chłopców,  $\bar{x}$  będzie średnią charakteryzującą badaną populację. Wartość  $\bar{x}$  oczywiście nigdy nie będzie identyczna z wartością średniej populacji ogólnej  $\mu$ . Będzie ona zawsze nieco odbiegać od liczby  $\mu$  o wartość, którą moglibyśmy nazwać błędem próbkowania (doboru badanej próby) lub nieprecyzją pomiaru, ponieważ wielkość  $\bar{x}$  obejmuje szum informacyjny związany z tym, co moglibyśmy nazwać „przypadkowością losowania próby”. Oznacza to, że każde losowanie może „dostarczyć” nieco odmiennych wartości  $\bar{x}$ , bliskich, lecz jednak nie identycznych z rzeczywistą wartością  $\mu$  w populacji ogólnej. Każdej wylosowanej badanej próbie można przypisać jakiś błąd próbkowania, problem nasz polega jednak na tym, że nigdy nie wiemy, jak duży jest to błąd. *Nota bene*, jeśli wiedzielibyśmy, jak duży jest to błąd, moglibyśmy bezbłędnie oszacować rzeczywistą wartość mierzonego parametru i nie byłoby potrzeby jakiegokolwiek aproksymacji.

W oparciu o teoretyczne rozważania nie jesteśmy oczywiście w stanie przewidzieć, co konkretnie wydarzy się w toku wykonywanego doświadczenia, lecz raczej, jaka będzie prawidłowość/tendencja zdarzeń w ogólnej interesującej dla nas zbiorowości o liczebności  $N$ . Charakterystykę mierzonej zmiennej określają parametry rozkładu badanej populacji, z której losujemy naszą badaną próbę, takie jak np. miara tendencji centralnej, czyli średnia próby  $\bar{x}$ ; dla elementów tej próby wykonujemy określoną liczbę powtórzeń pomiaru. Im bardziej wrasta liczebność naszej losowej próby ( $N$ ), tym bardziej nasze mierzone  $\bar{x}$  przybliży się do średniej populacji ogólnej  $\mu$ . Z uwagi na naturalnie występującą zmienność w każdej populacji elementów, a także na błędy pomiarowe towarzyszące każdym pomiarom, w naszej próbie zawsze znajdzie się niewielka frakcja wyników o wartościach mniejszych lub większych od  $\mu$ . Rozkład mierzonych wartości wokół rzeczywistego  $\mu$  jest dla nas odzwierciedleniem faktu, iż nasze  $\bar{x}$  jest jedynie niedoskonałą miarą ogólnopopulacyjnego  $\mu$ , oraz ukazuje rozmiar „szumu informacyjnego” towarzyszącego naszym pomiarom. Analizując równanie opisujące zależność między odchyleniem standardowym (miara zmienności wewnątrzpopulacyjnej) i błędem standardowym (miara nieprecyzji pomiarowej)

$$SEM = \frac{SD}{\sqrt{N}}$$

dostrzegamy, że nieprecyzja naszych pomiarów, czyli nasz błąd doświadczalny, maleje (a zatem precyzja rośnie) wraz z rosnącym  $N$ . Przy wystarczająco wysokich liczebnościach możemy mieć pewność, że nasza estymowana wartość średnia badanej populacji staje się



coraz bliższa rzeczywistej wartości  $\mu$ ; a zatem moglibyśmy powiedzieć, że duża liczebność próby poprawia naszą precyzję oceny.

Ogólna prawidłowość między wielkością próby a błędem próbkowania (losowego doboru elementów próby), pokazująca, że im większa liczebność ( $n$ ), tym mniejszy błąd doświadczalny, wskazuje nam, że zwiększając liczebność próby losowej redukujemy tym samym błąd naszej oceny. Stąd, aby zapewnić odpowiednio wysoką precyzję w naszych doświadczeniach, powinniśmy świadomie dobierać duże liczebnie próby do przebadania. Oznacza to jednak także wzrost czaso- i kosztocłonności naszych badań. Cały czas dążymy zatem do kompromisu między tym, jak racjonalnie obniżyć niezbędną liczebność badanej próby, aby precyzyjność naszych pomiarów nie spadła poniżej pewnej krytycznej granicy wyznaczającej w ogóle jakąkolwiek użyteczność zbieranych przez nas danych. Kompromis ten – z jednej strony – wyznacza zatem wielkość badanej próby na tyle wysoką, aby próba nasza nadal pozostawała reprezentatywna dla charakterystyki zbiorowości ogólnej (o której chcemy się autorytatywnie wypowiadać), z drugiej – na tyle niską, abyśmy w ogóle korzystali z faktu doboru próby losowej (zamiast badać całą zbiorowość). Procedury służące estymacji minimalnej liczebności losowej badanej próby służą właśnie do tego, aby ułatwić nam szybkie podejmowanie decyzji w kwestii właściwego kompromisu między wielkością próby, istotnością oraz mocą statystyczną wniosku.

## Logika testowania hipotez statystycznych

Testowanie hipotez statystycznych jest niewątpliwie najpowszechniej wykorzystywaną procedurą statystyczną. Ogólny algorytm działania procedur statystycznych zmierzających do zweryfikowania wiarygodności naszych sądów dotyczących zgodności lub niezgodności danych doświadczalnych z teorią badacza nazywa się testowaniem istotności oraz polega na stawianiu hipotez statystycznych oraz orzekaniu o ich prawdziwości lub fałszywości.

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.

*Przykład:* Kwas acetylosalicylowy (ASA) przenika przez dwuwarstwą lipidową błon biologicznych.

Pojedyncza hipoteza statystyczna powinna dotyczyć fragmentu hipotezy badawczej; stąd każdą koncepcję powinniśmy sprowadzić do kilku/kilkunastu hipotez statystycznych – każda z nich będzie rewidowała słuszność pojedynczych porównań. Hipotezy statystyczne zestawia się parami: hipotezie podstawowej (tzw. zerowej) przeciwstawia się hipotezę przeciwną (alternatywną) – w taki sposób, że jedna jest zaprzeczeniem drugiej. Precyzowanie hipotez polega na zestawieniu par przeciwieństw, np. stwierdzeń:

Hipoteza zerowa ( $H_0$ )	–	fakt A jest prawdziwy
Hipoteza alternatywna ( $H_A$ )	–	fakt A jest fałszywy



Taki parytet ma niezwykle doniosłe implikacje w praktyce doświadczalnej. Przede wszystkim hipotezy te są całkowicie dopełniające się: odrzucenie jednej z hipotez (jako nieprawdziwej) narzuca konieczność zaakceptowania drugiej. Stwarza to badaczowi wspaniałe warunki korzystania z narzędzia popperowskiej falsyfikacji. Pamiętając, iż weryfikacja pozytywna jest nie do spełnienia na poziomie logicznym (David Hume, a później Karl R. Popper), staramy się w naszym wywodzie myślowym sfalsyfikować jedną z hipotez (i odrzucić ją), aby automatycznie móc zaakceptować drugą (alternatywną), bez potrzeby jej udowadniania. Nasze doświadczenie podpowiada nam, jak sformułować brzmienie weryfikowanych hipotez, aby odpowiedź/odpowiedzi na stawiane sobie pytania były jak najpełniejsze.

*Przykład:* Badając hipotezę o ASA, pragniemy wykazać, czy związek ten może przenikać przez dwuwarstwą lipidową błony na drodze biernego lub ułatwionego transportu. Zgodnie z takim założeniem oczekujemy, iż stosownym modelem błony biologicznej, który posłużyć nam może do zweryfikowania naszej koncepcji badawczej, będą np. błony liposomalne lub tzw. „czarne błony lipidowe” (PLM, BLM), złożone z lipidów błonowych, lecz pozbawione białek. Jeżeli potwierdzimy występowanie transportu ASA przez takie modelowe błony, to wnioskujemy – *per analogiam* – iż prawdopodobne jest także przenikanie ASA przez błony komórkowe. Należy sobie jednak zdawać sprawę z faktu, że wykazanie takiego podobieństwa/analogii nie jest oczywiście wystarczającym dowodem na to, że mechanizm działania jest taki sam – jest jednak prostym sposobem weryfikacji, czy podążać dalej tym torem rozumowania. Jeżeli stwierdzilibyśmy, że transport następuje, to w dalszym rozumowaniu zaplanujemy doświadczenie wykazujące występowanie transportu ASA przez błony komórek wybranych do badań. Przeciwnie, jeżeli nie stwierdzimy występowania transportu ASA przez błony modelowe, nie będzie to dla nas dowodem na to, iż transport nie zachodzi w ustroju, a jedynie wskaże, iż w układzie modelowym przenikanie takie nie zachodzi. Możliwe do sprecyzowania hipotezy statystyczne mogłyby mieć brzmienie:

- ◆ hipoteza 1: stężenie ASA w kompartmentcie docelowym układu pomiarowego nie wzrasta w czasie;
- ◆ hipoteza 2: stężenie ASA w kompartmentcie docelowym układu pomiarowego wzrasta w czasie.

lub

- ◆ hipoteza 1: stężenie ASA wewnątrz komórki nie wzrasta w czasie;
- ◆ hipoteza 2: stężenie ASA wewnątrz komórki wzrasta w czasie.

Zwróćmy uwagę, iż w przykładzie tym zakładamy stosowanie testu jednostronnego: oczekujemy bowiem określonego kierunku zmian, a nie dowolnych (jakichkolwiek) zmian; to drugie (oczekiwanie, że transport ASA odbywałby się wbrew gradientowi stężeń) jawi się nam jako niespójność logiczna w odniesieniu do idei tego doświadczenia.

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 prawdopo-



dobieństwem), ale nigdy udowodnienie jej prawdziwości. Od właściwego sprecyzowania hipotezy statystycznej zależy to, czy będziemy mogli dowieść (z określonym prawdopodobieństwem) jej słuszności lub fałszywości.

Jakie właściwie brzmienie powinna mieć dobrze sformułowana hipoteza zerowa? Umownie przyjęto, aby hipoteza zerowa zakładała niewystępowanie różnic, natomiast hipoteza alternatywna wskazuje na występowanie jednej lub wielu różnic. Jako badacze, oczekujemy najczęściej wykazania lub wykrycia pewnej odmienności, charakterystycznej cechy odróżniającej, występowania jakiegoś efektu działania badanego czynnika, zależności itp. Toteż nasza teoria badacza na ogół pokrywa się z brzmieniem hipotezy alternatywnej. Skoro tak, to możemy powiedzieć, że hipoteza zerowa będzie najczęściej zaprzeczeniem naszej teorii badacza.

Hipoteza może być odrzucona, jeżeli materiał dowodowy pozwala nam orzec z dużym prawdopodobieństwem, że hipoteza jest fałszywa. Z drugiej strony, hipoteza nie może być odrzucona, jeżeli nie mamy podstaw do jej zaprzeczenia. Zaprzeczeniem hipotezy zerowej jest hipoteza alternatywna. Tylko jedna z nich może być prawdziwa, a wtedy druga musi być fałszywa, ponieważ obie hipotezy obejmują wszystkie możliwe warianty/możliwości. Konsekwencją niespełnienia równości  $\mu_1 = \mu_2$  musi być zaakceptowanie braku równości,  $\mu_1 \neq \mu_2$ . Hipoteza zerowa postuluje, że  $\mu_1 = \mu_2$ . W rzeczywistości testujemy równość  $\bar{X}_1 = \bar{X}_2$  i zakładamy, że wartości średnie dla badanych prób są reprezentatywne dla populacji generalnej.

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.

Skoro hipoteza zerowa jest nieprawdziwa, to znaczy że prawdziwa jest hipoteza alternatywna. Ponieważ prawie nigdy nie znamy wartości rzeczywistych charakteryzujących miary położenia i rozproszenia dla danej zbiorowości, a jedynie dostrzegamy „poblask” rzeczywistości na podstawie analizy próby losowej, przeto o prawdziwości czy fałszywości hipotez statystycznych możemy orzekać z określonym prawdopodobieństwem mniej lub bardziej różnym od 1. Wartość tego prawdopodobieństwa precyzują dwa błędy statystyczne testowania hipotez.

Jak już wspomniano wyżej, 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. 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  $\beta$ ).

- ◆ Prawdopodobieństwo błędu I rodzaju alfa ( $\alpha$ ) to prawdopodobieństwo błędnego odrzucenia hipotezy  $H_0$  w przypadku, gdy jest ona prawdziwa. Błąd I rodzaju popełniamy, jeżeli mylnie odrzucamy prawdziwą hipotezę zerową. Popełniając go, postulujemy różnicę, której *de facto* nie ma.
- ◆ Prawdopodobieństwo błędu II rodzaju beta ( $\beta$ ) to prawdopodobieństwo błędnego odrzucenia hipotezy  $H_1$  w przypadku, gdy jest ona poprawna. Błąd statystyczny II rodzaju popełniamy, jeżeli nie odrzucamy hipotezy zerowej, wtedy gdy jest ona fałszywa. Popełniamy go, ukrywając istniejącą różnicę. Prawdopodobieństwo błędu II rodzaju oraz moc testu bardzo istotnie zależą od liczebności próby oraz wielkości minimalnej różnicy, jaką badacz chce wykryć. Moc testu to tak naprawdę zdolność testu do wykrycia istotnej różnicy, jeżeli takowa naprawdę istnieje.

Na użytek praktyczny zapamiętamy, że istotność wyniku testu statystycznego to prawdopodobieństwo popełnienia błędu  $\alpha$ , zaś prawdopodobieństwo odrzucenia fałszywej hipotezy zerowej to moc testu (rys. 1).

		świat realny - oparty na faktach	
		$H_0$ 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 odrzucać $H_0$	<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. Jak dobrać wartości  $\alpha$  i  $\beta$  i czym się przy takim doborze kierować? Obowiązująca konwencja jest tu o wiele bardziej sztywna dla  $\alpha$  niż dla  $\beta$ : podczas gdy ogólnie akceptuje się utrzymywanie istotności ( $\alpha$ ) na poziomie 0,05 lub niższym, co do  $\beta$  przyjmujemy, że nie powinno ono przekraczać 0,2, co oczywiście oznacza, iż zadowolamy się mocą testu nie mniejszą niż 80%, aby wykrywać zasadną różnicę w stosunku do tego, co „mówi” hipoteza zerowa.



Jak zatem stwierdzić, czy wynik jest rzeczywiście istotny? Nie można niestety zupełnie uniknąć pewnej umowności co do tego, jaki poziom istotności skłonni jesteśmy uznać za rzeczywiście istotny. Oznacza to, że wybór poziomu istotności, powyżej którego wynik będzie odrzucany jako nieistotny, jest wyborem arbitralnym. Trafna ocena tego, jakie powinno być  $\alpha$ , determinuje w olbrzymim stopniu specyfika określonego problemu badawczego, a zatem trafny dobór jest pochodną doświadczenia naukowego badacza. W praktyce oznacza to, że ostateczna decyzja w tym względzie zależy od wielu czynników, od tego czy wynik był przewidziany *a priori* czy też jedynie był odkryty *post hoc* (po fakcie) w wyniku analiz i porównań przeprowadzonych na określonej zbiorowości danych, od zebranego materiału doświadczalnego, jak i od tradycji panującej w danej dziedzinie badań. W wielu dziedzinach badań jako typową wartość graniczną poziomu istotności przyjmuje się  $p = 0,05$ . Poniżej tej wartości wynik oceniany jest jako statystycznie istotny. Zauważmy, że jest to wartość, która niesie w sobie dość duże ryzyko popełnienia błędu (5%). W badaniach biomedycznych wyniki istotne na poziomie  $p < 0,01$  uważa się powszechnie jako statystycznie istotne, zaś wyniki istotne na poziomie  $p < 0,005$  lub  $p < 0,001$  postrzega się jako wysoce istotne. Należy jednak mieć świadomość, że tego typu klasyfikacje nie są niczym innym niż tylko konwencjami o dużej dozie dowolności, opartymi na doświadczeniu badawczym.

Zauważmy, że w przypadku gdybyśmy znali wyniki dla całej populacji generalnej, hipoteza zerowa musiałaby być albo prawdziwa albo fałszywa z prawdopodobieństwem 100% – tym samym ryzyko popełnienia błędu I lub II rodzaju byłoby zerowe.

Przedstawiony powyżej schemat poprawnego lub błędnego wnioskowania (popełniania błędów statystycznych) nawiązuje do zasad legislacyjnych przy orzekaniu winy lub niewinności (rys. 2).

		świat realny - oparty na faktach	
		jest niewinny	jest winny
werdykt	nie jest winny		<b>błąd II rodzaju</b>
	jest winny	<b>błąd I rodzaju</b>	

Rys. 2. Zasady opisujące trafność stawiania werdyktu w sądownictwie podczas orzekania o winie lub niewinności podsądnego oraz możliwe błędy decyzyjne.

Jeżeli podsądny jest niewinny, a sąd orzeka jego winę, to popełniany jest błąd I rodzaju (podsądny wysłany zostaje do więzienia „za niewinność”). Z drugiej strony, jeżeli podsądny jest winny, a sąd orzeka jego niewinność, to popełniany jest błąd II rodzaju (sąd uwalnia winowajcę). Jeżeli sąd orzeka „nie jest winny”, nie oznacza to, że podsądny nie popełnił winy. Podobnie – w testowaniu hipotez statystycznych – decyzja nie brzmi „przyjąć hipotezę zerową”, lecz „nie odrzucać hipotezy zerowej” – to nie to samo. Dlatego w testowaniu statystycznym nigdy nie możemy udowodnić prawdziwości hipotezy zerowej



– możemy ją jedynie odrzucić. Kiedy orzekamy, że wynik jest nieistotny statystycznie, nie znaczy to, że przyjmujemy hipotezę zerową, po prostu jej nie odrzucamy. Pamiętając, że to nie to samo, należy właściwie wyrażać i budować hipotezy statystyczne, tak aby dawały nam najbardziej wiarygodne podstawy do udowadniania hipotez naukowych.

## Wnioskowanie typu R-S oraz typu A-S

Załóżmy, że realizując nasze zadanie badawcze, jesteśmy zainteresowani, aby wykazać, że dziewczęta w określonym przedziale wieku są średnio niższe niż chłopcy w takiej samej kategorii wiekowej. Możemy to wyrazić za pomocą zapisu matematycznego:  $\bar{x}_{chłopcy} > \bar{x}_{dziewczeta}$ . Przeczenie podpowiada nam, że zapis ten jest prawdziwy, ale pragniemy uzyskać dowód statystyczny. W celu zweryfikowania słuszności tej nierówności, musimy zestawić dwie przeciwstawne brzmiące hipotezy, a następnie zebrać dane i policzyć statystykę testu. Pamiętajmy, że hipoteza zerowa ( $H_0$ ), będąca logicznym dopełnieniem hipotezy alternatywnej ( $H_A$ ), może być jedynie odrzucona (a nie przyjęta). Wierzmy, że w naszym konkretnym przypadku prawdziwe jest to, co stwierdza hipoteza alternatywna. Obliczając statystykę testu przy weryfikowaniu tej pary hipotez, pragniemy wykazać, że z wysokim prawdopodobieństwem  $H_0$  jest fałszywa i stąd powinna być odrzucona. W ten sposób, odrzucając (fałszywą) hipotezę zerową, akceptujemy to, w co wierzymy (przyjmując hipotezę alternatywną). Zatem odrzucenie hipotezy zerowej wspiera w tym przypadku teorię badacza, a taki typ rozumowania nazywa się wnioskowaniem typu R-S (*reject-support*), ponieważ odrzucając  $H_0$  dostarczamy argumentów na poparcie naszej teorii („że dziewczynki są średnio niższe od chłopców”). W praktyce doświadczenia, dla których stosujemy wnioskowanie typu R-S, dotyczą z reguły porównywania dwóch średnich: grupy kontrolnej i badanej, a teoria badacza zakłada, że czynnik badany (np. leczenie, zabieg itp.) wywołuje zmianę badanego parametru/zmiennej. Badacz wykorzystuje określony test inferencyjny i stara się odrzucić hipotezę zerową zakładającą brak wpływu testowanego czynnika. W testowaniu typu R-S popełnienie błędu I rodzaju ( $\alpha$ ) oznacza asymilację wyniku fałszywie dodatniego. Z punktu widzenia badacza błąd  $\alpha$  jest skrajnie niepożądany, gdyż oznacza stratę czasu i pieniędzy, zwłaszcza w przypadku, gdy taki fałszywie dodatni wynik jest interesujący z teoretycznego punktu widzenia. Napędza to dalsze badania w tym samym kierunku, badania najczęściej bezcelowe, gdyż nie przywiodą do tych samych wyników i wniosków, które były fałszywe. Niemożność powielenia wyników (fałszywie) dodatnich może rodzić zdezorientowanie w środowisku naukowym (niespójne dane obserwacyjne) i frustrację (np. zwątpienie we własne kompetencje). Z drugiej strony, nie mniejszą „tragedią” jest popełnienie błędu II rodzaju ( $\beta$ ) w testowaniu R-S, gdyż „prawdziwa” teoria zostaje niesłusznie odrzucona. W ten sposób, możemy stracić szansę na zaimplementowanie do praktyki nowego sposobu leczenia, diagnostyki itp., co do których mylnie nie wykazaliśmy przewagi w stosunku do grupy referencyjnej (kontroli). Tracimy (na jakiś czas) wartościową naukowo procedurę, sposób myślenia, interesującą teorię, gdyż fałszywie negatywny wynik osłabia nasz entuzjazm w kierunku kontynuowania tego kierunku badań. Ostatecznie, korzystnie jest



ograniczać wielkość obu błędów, choć w praktyce doświadczalnej, zwłaszcza dla małych prób, zwykle szukamy kompromisu między niskimi  $\alpha$  i  $\beta$ .

Przeciwstawna logika rozumowania towarzyszy wnioskowaniu typu A-S (*accept-support*). W testowaniu A-S zależy nam właśnie na nieodrzućeniu  $H_0$ , która – jak wierzymy – wspiera teorię badacza. Z przypadkami takimi mamy do czynienia np. w naukach farmaceutycznych, gdy pragniemy dowieść identyczności działania preparatu, leku, jakości procedury izolowania, oczyszczania itp. Teoria badacza jest wspierana przez nieodrzućenie hipotezy zerowej. Stąd w sytuacjach takich błąd I rodzaju ( $\alpha$ ) jest wynikiem fałszywie ujemnym dla naszej teorii (zaprzeczamy niefałszywej  $H_0$ ), podczas gdy błąd II rodzaju ( $\beta$ ) stanowi wynik fałszywie dodatni (akceptujemy fałszywą  $H_0$ ). W konsekwencji, minimalizowanie ryzyka popełnienia błędu I rodzaju w testowaniu typu R-S jest tożsame z maksymalizowaniem poziomu ufności w prawdziwość teorii badacza w testowaniu A-S.

Dylemat, przed którym stoi badacz, polega na poszukiwaniu możliwości testowania z jak największą mocą statystyczną w próbach losowych o umiarkowanej liczności. Niewielkie próby implikują jednak niewielką moc, czyli małe zaufanie badacza w wiarygodność wyniku testu statystycznego. Z drugiej strony, nadmiernie wysoka moc statystyczna także może stanowić problem w praktyce badawczej. W testowaniu typu R-S nawet drobne, trywialne różnice między porównywanymi średnimi w bardzo licznych grupach mogą nas przywozić do odrzućania hipotezy zerowej, niezależnie od występowania rzeczywistej różnicy. Jest to jeszcze bardziej krytyczną przeszkodą w testowaniu typu A-S, ponieważ zbyt liczne próby zwiększają prawdopodobieństwo mylnego odrzućenia  $H_0$ , a zatem działają przeciwko teorii badacza, nawet wtedy gdy teoria przystaje do zebranych danych niemal idealnie. W takich przypadkach oczywiście nadmiernie wysoka precyzja działa przeciwko badaczowi (tabela 1).

Tabela 1. Porównanie błędów statystycznych popełnianych przy testowaniu hipotez.

<b>Błąd I rodzaju</b>	<b>Błąd II rodzaju</b>
oznaczenie: $\alpha$	oznaczenie: $\beta$
definicja: mylne odrzućenie prawdziwej $H_0$	definicja: mylne przyjęcie fałszywej $H_0$
ustalany a priori	zależny od innych parametrów
nie zależy od wielkości próby, jeśli ustalony a priori	silnie zależy od wielkości próby oraz istotności
wzrasta wraz z liczbą testów wykonanych na badanej próbie (wymaga poprawki na testowanie wielokrotne)	może być oceniony jedynie jako funkcja rzeczywistego testowanego efektu w badanej populacji
	maleje ze wzrostem liczebności
	maleje wraz ze wzrostem liczby przeprowadzanych testów i ocenianych punktów końcowych

W podsumowaniu, testując hipotezy według schematu R-S, jesteśmy intuicyjnie zainteresowani odrzućeniem  $H_0$ , zatem popłaca minimalizowanie ryzyka błędu I rodzaju



( $\alpha$ , istotności). Odpowiednio wysoka liczebność próby jest dla badacza korzystna, dlatego zawsze opłaca się ocenić minimalną liczebność badanej próby przed rozpoczęciem doświadczenia. Powinniśmy zadbać o to, aby moc wnioskowania statystycznego nie była za niska, a więc zadbajmy także o małe ryzyko popełnienia błędu II rodzaju ( $\beta$ ). Należy pamiętać wszelako, że zbyt wysoka moc statystyczna działa przeciwko nam, ponieważ sprawia, że niewielkie trywialne różnice mogą osiągać wysoką istotność statystyczną i doprowadzać do niesłusznego odrzucenia hipotezy zerowej. Z drugiej strony, wybierając testowanie zgodnie ze schematem A-S, zależy nam, aby nie odrzucać  $H_0$ . Powinniśmy kontrolować wielkość błędu II rodzaju ( $\beta$ ) i unikać nieodrzucaenia fałszywej  $H_0$ , ale zależy nam także na minimalizowaniu ryzyka błędu I rodzaju ( $\alpha$ ), aby nie pozbyć się pochoinnie prawdziwej hipotezy zerowej. Paradoksalnie, próby bardzo liczne są niepożądane z punktu widzenia badacza, gdyż nadmiernie wysoka moc sprzyja odrzuceniu prawdziwej hipotezy zerowej i naszej słusznej teorii, w oparciu o wykrycie trywialnych różnic zarejestrowanych z powodu naturalnie występującej zmienności danych pomiarowych.

## Istotność statystyczna

Konwencjonalnie terminu „istotność” używamy, aby wyrazić naszą ufność w to, że mylnie nie odrzucamy prawdziwej hipotezy zerowej. Poziom istotności równy jest prawdopodobieństwu popełnienia błędu statystycznego I rodzaju ( $\alpha$ ) w takim sensie, że wyższa istotność oznacza numerycznie mniejszą wartość prawdopodobieństwa (tabela 1). W literaturze panuje mały zamęt terminologiczny: stosowane są oba oznaczenia, zarówno „ $\alpha$ ” jak i „p”, do wyrażania istotności statystycznej. Jak się w tym połąpać? Przyjęło się, że z oznaczenia grecką literą „ $\alpha$ ” korzystamy przy definiowaniu prawdopodobieństwa *a priori*, na etapie planowania doświadczenia, podczas gdy termin „wartość p” (lub „P”) stosujemy raczej, aby wskazać na istotność *a posteriori*, wyestymowaną na podstawie policzonej statystyki testu dla zebranych danych doświadczalnych. Co istotne, w zależności od tego, czy dokonujemy weryfikacji hipotez w kategoriach binarnych czy analogowych, orzekamy o fałszywości lub niefałszywości hipotezy zerowej lub też – dokładnie obliczamy prawdopodobieństwo pomyłki przy mylnym odrzuceniu hipotezy zerowej na podstawie konkretnej wartości obliczonej statystyki testu. W pierwszym przypadku podejmujemy decyzję odrzucenia (lub nie)  $H_0$ , po prostu orzekając „prawda – fałsz” na podstawie dobranego *a priori* poziomu istotności ( $\alpha$ ). Powiedzmy, że zdecydowaliśmy, iż będziemy testowali parę hipotez statystycznych na poziomie istotności  $\alpha < 0,001$ . Oznacza to, że jeśli odrzucimy hipotezę zerową, będzie istniała mniej niż jedna szansa na 1000, że popełnimy błąd decyzyjny. Po zebraniu wyników doświadczenia obliczamy oczywiście statystykę testu i porównujemy jej wartość z wartością teoretyczną odczytaną dla przyjętych warunków (przyjętej *a priori* istotności  $\alpha=0,001$  oraz liczby stopni swobody zależnej m.in. od liczebności próby). Jeżeli policzona, „doświadczalna” wartość statystyki testu jest większa od (lub równa) teoretycznej, odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną. Podejście takie jest typowo jakościowe: mamy dwie możliwości decyzyjne – odrzucić lub nie odrzucać  $H_0$ . Jeżeli nasza obliczona statystyka testu jest o wiele większa od teoretycznej, istnieje szansa,



iż nasz zapis „ $\alpha < 0,001$ ” oznacza w rzeczywistości „mniej niż 1 na 5000” lub nawet „mniej niż 1 na 100000”. Nie jest to dla nas ważne, gdyż odrzuciliśmy  $H_0$  na minimalnym, satysfakcjonującym nas poziomie istotności. W sytuacji, gdy wyrażamy istotność jako dokładną wartość prawdopodobieństwa *a posteriori* (policzonego na podstawie statystyki testu wyliczonej z danych doświadczalnych), stosujemy tzw. podejście ilościowe (analogowe). W takich sytuacjach nie musimy się kierować żadną konwencją przy ocenie, że np. prawdopodobieństwo 1% można uważać za wystarczająco wysoką istotność, podczas gdy prawdopodobieństwo 5% nie spełnia tego kryterium. Podejście analogowe jest szczególnie powszechnie stosowane w analizach wielowymiarowych, gdy dokładne wartości prawdopodobieństwa dla poszczególnych zmiennych w modelu są dla nas często cenną wskazówką przy ocenie i porównywaniu różnych modeli wieloparametrowych.

## Moc statystyczna i wielkość próby

Mówi się, że umiejętne „zonglowanie” tymi dwoma parametrami leży u podstaw dobrego planowania doświadczeń. Techniki oceny mocy statystycznej oraz estymacji wielkości próby pozwalają nam zdecydować: (a) jak duża powinna być nasza próba, abyśmy mogli wiarygodnie i precyzyjnie wnioskować o słuszności/fałszywości stawianych hipotez, oraz (b) jakie jest prawdopodobieństwo, że test, którym się posługujemy, będzie w stanie wykryć w określonej, szczególnej sytuacji pożądaną przez nas zmianę pod wpływem testowanego czynnika. Zarówno estymacja niezbędnej liczebności próby, jak i oszacowanie mocy statystycznej są tak niezmiernie ważne, gdyż bez tej oceny ryzykujemy, że nasza próba może być albo zbyt mało liczna, albo zbyt liczna. W pierwszym przypadku, jeśli nasza próba jest za mało liczna, badacz nie może z wystarczającą precyzją udzielić wiarygodnych odpowiedzi na stawiane pytania. W drugim przypadku ryzykujemy stratę energii, czasu i pieniędzy, nie odnosząc dodatkowych korzyści z przebadania nadmiernie licznej próby losowej.

## Dobieranie mocy statystycznej wnioskowania

Planując doświadczenie, powinniśmy zadbać o to, aby moc naszego wnioskowania była wystarczająco wysoka do wiarygodnej falsyfikacji naszej hipotezy zerowej. Pamiętamy, że w testach o dużej mocy statystycznej ryzyko popełnienia błędu II rodzaju ( $\beta$ ) jest minimalne. Oznacza to, że fałszywa  $H_0$  jest zawsze odrzucana, badacz potwierdza swoją teorię, i dlatego przeprowadzanie eksperymentu ma w ogóle sens. Liczne czynniki wpływają na wielkość mocy statystycznej. Należą do nich: (a) wielkość badanej próby, (b) wielkość oczekiwanego efektu działania testowanego czynnika, (c) zmienność mierzonego parametru/zmiennej oraz (d) typ stosowanej procedury statystycznej.

Wiemy już, że liczniejsze próby warunkują większą moc wnioskowania. Jesteśmy jednak także świadomi, że zwiększanie liczebności naszej próby to potrzeba większego wysiłku z naszej strony, tak w zakresie energii czy czasu, jak i większe koszty eksperymentu.



Dlatego tak zasadna jest ocena *a priori*, jaka powinna być niezbędna, wystarczająca i zadowalająca liczebność naszej próby. Oczekując wyraźnego efektu badanego czynnika, możemy założyć, że z dużym prawdopodobieństwem będziemy mogli sfalsyfikować hipotezę zerową. Im większa oczekiwana różnica, tym większa moc naszego wnioskowania. Pamiętajmy jednak, że oceniany przez nas efekt winien być istotny z praktycznego punktu widzenia (np. w praktyce klinicznej, diagnostycznej itd.), a nie jedynie w kategoriach statystycznych. Wysoka zmienność badanych/mierzonych zmiennych, czy to w wyniku małej precyzyjności stosowanych przez nas metod pomiarowych czy w wyniku naturalnie istniejącej zmienności biologicznej, zmniejsza oczywiście moc naszego wnioskowania, i odwrotnie – wysoka precyzja i duża spójność danych pomiarowych poprawia moc wnioskowania statystycznego. Nie wszystkie testy charakteryzują się jednakowo wysoką mocą statystyczną: parametryczne są na ogół mocniejsze niż nieparametryczne. Toteż w dużej mierze od nas zależy mądry wybór procedury, jaką zechcemy stosować, aby zminimalizować ryzyko popełnienia błędu II rodzaju.

## Szacowanie wielkości próby

Zanim przeprowadzimy eksperyment, warto postawić sobie następujące pytania:

- ◆ Ile powtórzeń powinniśmy zebrać, aby z wystarczającą mocą wnioskować o istotności badanego efektu?
- ◆ Czy możemy mieć pewność, iż przebadawszy próbę losową o wystarczająco dużej liczebności, będziemy w stanie wnioskować o występowaniu lub braku istotnych różnic?
- ◆ Na czym opiera się nasza estymacja liczebności próby? Innymi słowy, jakie parametry powinniśmy znać, aby ocenić minimalną wielkość badanej próby?
- ◆ A może szkoda czasu na niepewne estymacje liczebności? Może powinniśmy kontynuować badania do czasu, aż wystarczy nam środków finansowych oraz czasu na zbieranie danych?
- ◆ Czy estymacji takiej powinniśmy dokonać *a priori* (zanim przeprowadzimy doświadczenie) czy *a posteriori* (po zebraniu danych)? Innymi słowy, czy potrzebujemy wiedzieć przed eksperymentem, ile danych zebrać? Czy raczej ocenimy po wykonaniu badania, z jaką mocą możemy się wypowiedzieć na temat zauważonych różnic?

Odpowiedź na powyższe pytania jest jednoznaczna. Estymacja liczebności badanej próby **musi** być przeprowadzona przed rozpoczęciem doświadczenia, aby mieć pewność, że wnioskowanie statystyczne będzie przeprowadzone z odpowiednią mocą statystyczną. Praktykujemy ją np. w sytuacji, gdy:

- ◆ dostrzegamy już na „pierwszy rzut oka”, że porównywane grupy różnią się między sobą,
- ◆ nie występują rzeczywiste różnice i nie wykażemy ich niezależnie od liczebności próby, zbierając bardzo dużą liczbę powtórzeń mnożymy tylko niepotrzebnie koszty eksperymentu.



Logika stojąca za estymacją liczebności próby w obu takich skrajnych przypadkach jest podobna. Tracimy czas, energię i pieniądze na gromadzenie niepotrzebnie licznych danych, co nie przyczynia się ani do zasadnego zwiększenia istotności poszukiwanych różnic, ani mocy, z jaką o nich orzekamy. Zbyt wczesne zakończenie eksperymentu ma także swoje wady. Za małą liczebność próby grozi nam niewystarczająca moc i błędnym przyjęciem fałszywej hipotezy zerowej. Nawet jeśli badany przez nas doświadczalny efekt jest oczywisty, a nie mamy cierpliwości, aby ukończyć doświadczenie, możemy nie wykazać statystycznie takiego efektu, co sprawi, że wykonywanie doświadczenia traci jakikolwiek sens.

Planując doświadczenie, powinniśmy zastanowić się, co stanowić ma dla nas minimalny efekt, który pragniemy wykryć, jaka jest minimalna moc, z którą pragniemy wykryć ten efekt, i przy jakiej liczebności próby wykryjemy tenże efekt z pożądaną mocą. Szacując minimalną liczebność badanej próby powinniśmy znać:

- ◆ zmienność wewnątrzgrupową (tj. SD, SEM; określające stopień zróżnicowania w naszej próbie, a także to, jaka jest nieprecyzja naszych pomiarów),
- ◆ wymiar efektu doświadczalnego (tzn. wielkość oczekiwanej różnicy między porównywanymi grupami, stopień dyskryminacji między nimi); jak bardzo ma się różnić wartość parametru mierzonego przed i po zadziałaniu testowanego czynnika,
- ◆ istotność z jaką pragniemy wykryć testowany efekt (czyli ryzyko, że popełnimy błąd I rodzaju, odrzucając hipotezę zerową, która nie byłaby w rzeczywistości fałszywa), oraz
- ◆ moc naszego wnioskowania (czyli ryzyko popełnienia błędu II rodzaju).

Warto podkreślić, że wszystkie z wymienionych wyżej parametrów (być może z wyjątkiem pierwszego w pewnych rzadkich sytuacjach) są dobierane przez badacza. To badacz decyduje, jakie są wartości tych parametrów, które posłużą mu do późniejszej oceny liczebności próby. Czyni to, zanim wykona jakiegokolwiek pomiary, opierając się wyłącznie na właściwym zaplanowaniu badania oraz czerpiąc ze swojego doświadczenia i wiedzy w zakresie problematyki, której poświęcone mają być planowane eksperymenty. Jako badacze, to my sami kreujemy myślowy obraz naszego doświadczenia, zanim przeprowadzimy je w praktyce. Może się to wydawać nie do wiary, w jaki sposób jesteśmy np. w stanie ocenić stopień zmienności jakiejś cechy, zanim zbierzemy dane doświadczalne. Niewiara taka wynika często z faktu, że możemy być na początku słabo przygotowani koncepcyjnie do stworzenia sobie wizerunku myślowego naszego badania. Przedyskutujmy pokrótce każdy z parametrów wykorzystywanych w celu oceny liczebności próby.

1. Zmienność wewnątrzgrupowa badanej cechy może być pochodną błędu pomiarowego (np. na skutek nieprecyzji techniki, którą się posługujemy), naturalnego zróżnicowania biologicznego lub obu. Jedynie w wyjątkowo rzadkich przypadkach jesteśmy zupełnie pozbawieni informacji na temat tego, jaka jest ta zmienność. Mogłoby się tak zdarzyć np. wtedy, gdy badamy jakąś cechę lub używamy jakiejś metodologii, które nigdy przedtem przez nikogo nie były badane czy stosowane, tzn. gdy nasze działania są absolutnie nowatorskie pod każdym względem. Najczęściej jednak tak nie jest. Możemy wyekstrapolować informacje na temat zmienności z badań innych uczonych, którzy wprawdzie nie zajmowali się identycznym problemem naukowym, z jakim my sami się zmagamy, lecz stosowali uniwersalnie przyjęte techniki pomiarowe. Jest jedynie



sprawą uczciwego i sprawnego przeszukania literatury naukowej, aby dowiedzieć się o stopniu takiej zmienności i zaimplementować tę informację do własnych celów. Jeżeli nie mamy dostępu do takiej informacji, to pozostaje nam oczywiście wykonanie kilku pomiarów w badaniu pilotowym w celu oszacowania takiej zmienności.

2. Podobnie sami decydujemy o wielkości efektu działania czynnika, który badamy. To od nas zależy, jaki efekt będziemy uważać za „satisfakcjonująco” wysoki, aby uznać, że testowany przez nas czynnik faktycznie działa. Nie zadowolamy się dowolnym, nawet istotnym statystycznie efektem, lecz jedynie takim, który oceniamy jako zasadny w odniesieniu do konkretnego problemu naukowego. Podejście takie jawi się jako szczególnie racjonalne np. w odniesieniu do badań klinicznych. Testując nowy lek, oczekujemy zdefiniowanego przez nas stopnia zmian parametrów biochemicznych pod wpływem tego leku, nie zaś jakichkolwiek zmian, także niewielkich i nieważnych w praktyce klinicznej, choćby istotnych ze statystycznego punktu widzenia. Jaki duży powinien być taki efekt, zależy od tego, co badamy. Na przykład, efekty przekraczające 20% są postrzegane jako interesujące w praktyce klinicznej.
3. Pamiętajmy, że poziom istotności oznacza ryzyko popełnienia błędu I rodzaju (odrzućcenia niefałszywej hipotezy zerowej). W częściej przeprowadzanym testowaniu typu R-S niezbyt wysoka istotność (umiarkowanie niskie  $\alpha$ ) oznacza po prostu nie nazbyt pochopne akceptowanie teorii badacza. Przy wyjątkowo wysokich wartościach poziomu istotności przyjmujemy, że nasza teoria prawie zawsze będzie słuszna. W praktyce klinicznej oznacza to wiarygodną diagnozę oraz pewne prognozowanie. Decydując o dobraniu właściwej istotności na etapie planowania doświadczenia, zapominamy na moment o eksploracyjnej mocy tego parametru. Spoglądamy na poziom istotności w kategoriach jakościowych (prawda czy fałsz), a nie ilościowych (jakie będzie dokładnie prawdopodobieństwo, że pomyliłem się zbyt pochopnie odrzucając  $H_0$ ?). Decyzja w kategoriach jakościowych wymaga oczywiście przyjęcia pewnej konwencji: co ma być zdefiniowane jako „prawda”, a co jako „fałsz”. Konwencja taka narzuca, rzecz jasna, przyjęcie arbitralnej wartości granicznej „zadowolającej” istotności, co – siłą rzeczy – jest podejściem mniej informatywnym niż stosowanie miar analogowych. Choć decydujemy w oparciu o kryteria binarne, musimy zdecydować, gdzie leży granica i jaka istotność uzasadnia przyjęcie naszej teorii badacza. W naukach społecznych i niektórych przyrodniczych przyjęło się, że graniczna istotność winna wynosić nie mniej niż 5% (czyli  $\alpha=0.05$ ), rzadziej 1% lub 0.1% (tzw. „wysoka istotność statystyczna”). Co to znaczy, że  $\alpha=0.05$ ? Oznacza to oczywiście, że na każde 100 podjętych decyzji na temat prawdziwości lub nieprawdziwości badanych hipotez mylimy się w 5 przypadkach. Możemy powiedzieć z ufnością w 95 na 100 przypadków, że wypowiadamy prawdę. Czy to dużo czy mało? Zależy to oczywiście od rozważanego problemu naukowego. Jeżeli postawimy diagnozę u 100 pacjentów i u 5 nasza diagnoza będzie fałszywa, czy jest to do przyjęcia czy też dyskwalifikuje nasz osąd? Wskazuje to, że w niektórych badaniach, np. klinicznych, nie ma niczego takiego jak graniczny standard istotności. Poziom ryzyka popełnienia błędu I rodzaju zależy od problemu badawczego, który rozważamy. Zwykle w sytuacjach takich ustawiamy ten poziom o wiele niżej niż 0.05, ale to my decydujemy, jaka powinna być to wartość,



w oparciu o naszą wiedzę, doświadczenie i w nawiązaniu do konkretnego pytania. Próba deklarowania jakichkolwiek norm w tym zakresie ma nas po prostu fałszywym poczuciem decyzyjnego bezpieczeństwa.

4. Moc wykrywania poszukiwanego efektu to ostatni z wybieranych przez nas parametrów. Pamiętajmy, że musimy ją dobrać mądrze: nie za nisko, aby wiarygodnie wykryć interesujący efekt, i nie za wysoko, aby nie wykazywać istnienia efektów nierzeczywistych. Zwykle dobiera się moc w zakresie 80-90% (czyli nasze ryzyko  $\beta=0.1-0.2$ ).

Istnieje wiele pakietów statystycznych, zarówno profesjonalnych [2, 3], jak i publicznie dostępnych w sieci [4-7], które umożliwiają obliczanie liczebności próby lub ocenę mocy statystycznej na podstawie poziomu istotności oraz efektu doświadczalnego. Bardziej zaawansowane programy oferują także szczegółową analizę graficzną zależności między mocą a liczebnością próby dla różnych wielkości dyskryminacji oraz istotności, dobieranych przez użytkownika. Daje to użytkownikowi możliwość porównania kilku ocen w zależności od warunków wstępnych dla estymacji. Niektóre programy zapewniają szybką kalkulację liczebności na podstawie danych wsadowych, jednak pakiety takie mogą być mało przyjazne dla niedoświadczonych badaczy. Prezentacja graficzna daje pełniejszy wgląd we wzajemne relacje między mocą, istotnością, zmiennością oraz dyskryminacją, i jeżeli to tylko możliwe, warto z niej zawsze korzystać [2]. Zmieniając poszczególne parametry wsadowe na etapie planowania mamy możliwość oceny, do jakiego stopnia poszukiwana przez nas różnica będzie się objawiać w zależności od szczególnych warunków doświadczenia, jakie planujemy [2, 3]:

## Uwagi końcowe

Podsumowując, dobra znajomość charakterystyki (m.in. mocy) testu, który planujemy wykorzystać do analizy zebranych przez nas danych, jest dla nas niezwykle korzystna na etapie planowania doświadczenia, zanim jeszcze zbierzemy nasze dane. Przede wszystkim powinniśmy zdawać sobie sprawę z faktu, że nawet niewielki wzrost zmienności ocenianego efektu może silnie wpłynąć na niezbędną liczebność naszej badanej próby, a zatem spowodować, że nie będziemy w stanie wykryć oczekiwanego efektu z pożądaną istotnością i mocą. W sytuacji takiej nie zawsze będziemy przygotowani na zareagowanie na takie zmiany warunków naszego doświadczenia, np. w sensie finansowym lub czasowym. Aby zaskoczenie było mniejsze i mniej traumatyczne, dobrą praktyką jest zaplanowanie doświadczenia w kilku alternatywnych wariantach (np. dotyczących oceny liczebności). Dobrze jest „pobawić” się nieco dobozem parametrów służących ocenie liczebności naszej próby, zamiast wybierać najbardziej optymistyczny dla nas wariant takiej oceny. Oczywiście jest, że powinniśmy to zrobić, zanim rozpoczniemy zbieranie danych, chociażby w celu ukształtowania sobie szerszego własnego spojrzenia na różne możliwe scenariusze w zaplanowaniu eksperymentu.



## Literatura

1. Watala C. (2006) How to plan an experiment? I. Randomization: current fad or (ever)lasting fashion? Arch Med Sci 2: 58-65.
2. *STATISTICA* for Windows [Computer program manual]. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104; email: [info@statsoftinc.com](mailto:info@statsoftinc.com), 2008.
3. StatsDirect statistical software, version 2.3.7. StatsDirect Ltd., 2004.
4. Simple Interactive Statistical Analysis (SISA) [Computer software]. Retrieved 13-11-2006 from <http://home.clara.net/sisa/power.htm>.
5. UCLA's Dept. of Statistics Power Calculator [Computer software]. Retrieved 13-11-2006 from <http://stat.ubc.ca/~rollin/stats/ssize/>.
6. Lenth R.V. (2006) Java Applets for Power and Sample Size [Computer software]. Retrieved 13-11-2006, from <http://www.stat.uiowa.edu/~rlenth/Power>.
7. Schoenfeld DA. Statistical considerations for clinical trials and scientific experiments [Computer software]. Retrieved 13-11-2006 from [http://hedwig.mgh.harvard.edu/sample\\_size/size.html](http://hedwig.mgh.harvard.edu/sample_size/size.html).