PRAKTYCZNE ASPEKTY SZACOWANIA LICZEBNOŚCI PRÓBY W BADANIACH EMPIRYCZNYCH

Janusz Wątroba, StatSoft Polska Sp. z o.o.

Ustalenie liczby obiektów (wielkości próby) w badaniach jest jednym z tych problemów, które trzeba rozstrzygnąć w niemalże każdym projekcie badawczym. Można podać wiele powodów uzasadniających znaczenie tego zagadnienia. Niektóre z nich wynikają bezpośrednio z uwarunkowań techniczno-organizacyjnych planowanego badania (koszty, czas przeznaczony na jego realizację czy dostępność docelowej zbiorowości obiektów badawczych) i badacze zazwyczaj przy projektowaniu badania poświęcają im najwięcej uwagi. Dzieje się to kosztem odsuwania na dalszy plan innych, nie mniej ważnych kwestii, w tym uwarunkowań metodycznych czy wymogów statystycznych (chodzi głównie o wykazanie istotności statystycznej interesujących badacza efektów przy jednoczesnym zapewnieniu odpowiedniej mocy statystycznej badania). Spełnienie wymogów statystycznych jest obecnie warunkiem zapewniającym wiarygodność uzyskanych wyników w środowisku naukowym. Nie bez znaczenia jest także możliwość opublikowania wyników takiego badania w liczących się czasopismach naukowych. Tak więc ich znajomość i poprawne uwzględnienie przy planowaniu badania leży w interesie badacza.

W pierwszej części opracowania zostaną pokrótce omówione zasygnalizowane powyżej czynniki determinujące liczebność obiektów badawczych. W drugiej części zostaną zaprezentowane praktyczne przykłady szacowania liczebności próby za pomocą odpowiednich procedur dostępnych w module *Analiza mocy testu* w programie *STATISTICA* [7, 4].

Znaczenie wielkości próby w badaniach empirycznych

Prowadzenie badań empirycznych na wysokim poziomie naukowym wymaga obecnie od badaczy i zespołów badawczych sporych umiejętności menedżerskich. Zdobycie środków na badania, umiejętne zaplanowanie stosownych działań, przeprowadzenie badania oraz odpowiednie zaprezentowanie i rozpowszechnienie jego wyników wiąże się z koniecznością podejmowanie szeregu trudnych decyzji. Badacze bardzo niechętnie godzą się na ograniczenie zakresu merytorycznych celów badania. Efektem takiego podejścia jest zazwyczaj stawianie zbyt dużej liczby pytań badawczych, w dodatku bardzo szczegółowych.

Wiele z problemów, które badacz musi rozstrzygnąć, dotyczy aspektów, które można określić jako techniczno-organizacyjne. Pierwszoplanowym ograniczeniem jest tutaj

wielkość środków, które ma dyspozycji zespół badawczy, gdyż to decyduje o zakresie badań. Inny, nie mniej ważny czynnik to czas, w którym badania powinny zostać przeprowadzone. Kolejne sprawy to: dostęp do zbiorowości obiektów, której mają dotyczyć wnioski, oraz ustalenie, jaka liczba jednostek powinna być objęta badaniami. W większości projektów badawczych konieczność określenia odpowiedniej wielkości próby pojawia się zazwyczaj już na etapie planowania badania. Duże znaczenie przywiązywane do liczby badanych jednostek wynika z faktu, iż decyzja w tej sprawie zależy od wspomnianych wyżej ograniczeń techniczno-organizacyjnych, a z drugiej strony jest jednym z czynników decydujących o spełnieniu wymogów statystycznych.

Zazwyczaj badacze traktują czynniki techniczno-organizacyjne jako pierwszoplanowe i dopiero w ramach ograniczeń, jakie one nakładają, są skłonni w drugiej kolejności brać pod uwagę również uwarunkowania statystyczne. Takie podejście może prowadzić do sytuacji, w której badana zbiorowość obiektów będzie zbyt mało liczna i, co za tym idzie, badanie może prowadzić do niewystarczająco uzasadnionych wniosków. Niezależnie od tego badanie może zostać potraktowane jako nieetyczne, ponieważ wystawianie ludzi bądź zwierząt doświadczalnych na ryzyko związane z badaniami jest uzasadnione tylko pod warunkiem, że jest realna szansa na uzyskanie użytecznych informacji. Z drugiej strony badanie, które uwzględnia zbyt dużą liczbę obiektów, będzie marnowaniem ograniczonych zasobów i może wiązać się z niepotrzebnym narażaniem badanych na ryzyko.

Z przytoczonych powyżej powodów wynika jednoznacznie, że odpowiedni dobór wielkości próby jest jednym z kluczowych etapów projektowania badania. Rozważane grupy czynników determinujących wielkość próby są nawzajem powiązane, w związku z tym optymalnym podejściem wydaje się dążenie do wypracowania rozwiązania kompromisowego, które w odpowiednim stopniu będzie z jednej strony uwzględniać różnego rodzaju środki poniesione na badania, a z drugiej oczekiwane efekty badawcze.

Uwarunkowania statystyczne a wielkość próby

Jak to zostało wcześniej zasygnalizowane, oprócz uwarunkowań wynikających z kwestii techniczno-organizacyjnych przy planowaniu i realizacji badania powinno się uwzględniać również wymogi statystyczne. Konieczność brania ich pod uwagę wynika z kilku powodów. Obecnie we wszystkich dziedzinach badań naukowych, w których prowadzone są badania empiryczne, wymaga się, aby wyniki zostały opracowane za pomocą odpowiednich metod statystycznych oraz aby wnioski z badań były poparte wnioskowaniem statystycznym. W ten sposób wymogi statystyczne są obecnie powszechnie traktowane, jako warunek konieczny „podejścia naukowego". Konsekwencją tego faktu są wymagania, jakie stawiają redaktorzy czasopism naukowych i współpracujący z nimi recenzenci w przypadku starania się o opublikowanie wyników badań. Ponadto coraz częściej instytucje dysponujące środkami na badania wymagają od badaczy (starających się o różnego rodzaju granty) uzasadnienia przyjętej w badaniach liczebności próby. W wielu badaniach wymóg uzasadnienia proponowanej liczebności badanych obiektów (zwierząt doświadczalnych, osobników) jest też stawiany przez komisje bioetyczne.

Celowym zatem wydaje się krótkie omówienie kryteriów statystycznych, które należy brać pod uwagę przy ustalaniu liczby obiektów do uwzględnienia w badaniach. W grę wchodzą następujące czynniki: kryterium istotności statystycznej, moc testu statystycznego oraz wielkość efektu.

Kryterium istotności statystycznej

W większości badań empirycznych przy ocenie ich wyników pojawia się pytanie, czy zaobserwowany efekt w sposób wiarygodny odzwierciedla rzeczywiste różnice czy też należy go raczej potraktować jako przypadkowy? Aby wykazać, że stwierdzony efekt nie jest dziełem przypadku, stosuje się statystyczną procedurę weryfikacji prawdziwości odpowiedniej hipotezy statystycznej. Wynik jest uznawany za statystycznie istotny, jeśli prawdopodobieństwo jego przypadkowego wystąpienia jest niższe od przyjętego z góry (ale umownego) niskiego progu, nazywanego **poziomem istotności** (oznaczanego grecką literą α). Standardowo przyjmuje się wartość 0,05. Warto przypomnieć, że podstawy tzw. klasycznej teorii testowania hipotez statystycznych zostały sformułowane w latach 1915-1933 przez R. Fishera, J. Neymana i E. Pearsona. Jednocześnie należy dodać, że ich poglądy w kluczowych kwestiach dość mocno się różniły [6]. Współczesne podejście do testowania hipotez statystycznych stanowi pewien rodzaj konglomeratu propozycji Fishera z jednej strony i poglądów Neymana i Pearsona z drugiej. Jest to efektem pewnego zamieszania, które zostało wprowadzone przez autorów podręczników statystycznych w latach 40. ubiegłego wieku, którzy ignorowali występujące w obu proponowanych podejściach różnice pojęciowe i inne sprzeczności. Neyman i Pearson proponowali bardziej jednoznaczną terminologię, mocniejsze podstawy matematyczne oraz pewną spójną filozofię, ale sposób, w jaki problematyka ta jest przedstawiana w podręcznikach zawierających wprowadzenie do statystyki, jest bardziej zbliżony do propozycji Fishera. Znajduje to odbicie w niespójnej terminologii – przykładowo procedura testowania nie daje możliwości *przyjęcia hipotezy zerowej*, a jednocześnie powszechnie stosowany jest termin *obszar akceptacji*.

Obecnie stosowana procedura obejmuje kilka kolejnych etapów. Punktem wyjścia jest wstępna hipoteza badawcza, której prawdziwość nie jest znana. Pierwszy krok polega na postawieniu pary wzajemnie wykluczających się hipotez: zerowej i alternatywnej. Hipoteza zerowa jest ustalana w taki sposób, aby można było podjąć jedną z dwóch decyzji: odrzucić zerową i przyjąć alternatywną lub stwierdzić brak podstaw do odrzucenia hipotezy zerowej. Drugim krokiem jest rozważenie statystycznych założeń dotyczących zbiorowości jednostek, na której jest przeprowadzany test. Następnie wybiera się odpowiedni test, postać statystyki testowej oraz wyznacza jej rozkład przy założeniu prawdziwości hipotezy zerowej. Kolejny krok to ustalenie poziomu istotności α, poniżej której hipoteza zerowa jest odrzucana (standardowo przyjmowana wartość to 0,05). Jeśli procedura testowania jest przeprowadzana z wykorzystaniem odpowiedniego oprogramowania komputerowego (np. *STATISTICA*), wtedy w następnym kroku na podstawie danych obliczana jest wartość statystyki testowej oraz odpowiadająca jej wartość prawdopodobieństwa p. W ostatnim kroku podejmowana jest decyzja dotycząca hipotezy zerowej. Jeśli obliczona wartość prawdopodobieństwa p jest niższa od przyjętego wcześniej poziomu

istotności α, wówczas hipoteza zerowa jest odrzucana, na korzyść hipotezy alternatywnej. W przeciwnym przypadku stwierdza się brak podstaw do odrzucenia hipotezy zerowej.

Badacz jest zainteresowany odrzuceniem hipotezy zerowej i w ten sposób wykazaniem istotności statystycznej zaobserwowanego efektu.

Moc testu statystycznego

W ten sposób, opierając się na wynikach przeprowadzonego testu statystycznego, badacz może w stosunku do hipotezy zerowej podjąć jedną z dwóch wzajemnie wykluczających się decyzji. Pierwsza z decyzji to odrzucenie sprawdzanej hipotezy zerowej, a druga to jej nieodrzucenie. Z drugiej strony ze względu na fakt, iż testowanie hipotez statystycznych opiera się na wynikach losowej próby, w świetle przywołanej powyżej procedury testowania hipotez statystycznych decyzja odrzucenia lub stwierdzenia braku podstaw do odrzucenia sprawdzanej hipotezy może być trafna lub błędna. Sytuację tę ilustruje zamieszczony poniżej schemat.

Decyzja \ H_0	prawdziwa	fałszywa
odrzucić	*błąd I rodzaju* α	*decyzja trafna*
nie odrzucać	*decyzja trafna*	*błąd II rodzaju* β

Rys. 1. Testowanie hipotez statystycznych – błędy I i II rodzaju.

Możliwe jest popełnienie jednego z dwóch rodzajów błędu: odrzucenie hipotezy zerowej, która jest prawdziwa (jest to tzw. błąd pierwszego rodzaju) oraz nieodrzucenie hipotezy zerowej, która jest fałszywa (jest to tzw. błąd drugiego rodzaju, oznaczany jako β). Błąd pierwszego rodzaju jest bezpośrednio kontrolowany poprzez ustalenie z góry poziomu istotności testu, czyli α. Okazuje się, że obydwa błędy są ze sobą powiązane. Jednoczesne minimalizowanie ich wartości nie jest możliwe, gdyż obniżenie jednego z nich powoduje wzrost drugiego. Rozwiązaniem problemu kontroli błędów I i II rodzaju jest zaproponowana przez Neymana *koncepcja testu najmocniejszego*. Nie wchodząc w szczegóły, polega ona na ustaleniu poziomu błędu I rodzaju i wyborze takiego testu weryfikującego hipotezę zerową, który w najmniejszym stopniu naraża badacza na popełnienie błędu II rodzaju. Jest to możliwe dzięki znajomości tzw. *mocy testu*. Moc testu jest definiowana jako prawdopodobieństwo odrzucenia hipotezy zerowej, gdy w rzeczywistości jest ona fałszywa, czyli $1-\beta$. Z punktu widzenia badacza moc testu oznacza jego zdolność do wykrywania prawdziwego efektu. W większości badań przyjmuje się, że akceptowalna moc testu to 0,80 lub więcej [1, 2, 3]. Moc testu jest drugim czynnikiem, który należy uwzględniać przy ustalaniu wielkości próby. Obliczenie mocy testu jest także zalecane w przypadku, gdy mimo oczekiwań badacza badanie nie wykazało efektu.

Wielkość efektu

Przy ocenie wyników badania można wyróżnić dwa rodzaje ich istotności. Pierwszy z nich to istotność statystyczna, która pozwala stwierdzić, czy zaobserwowany efekt można przypisać wyłącznie błędowi próby. Kryterium brane pod uwagę przy ocenie istotności statystycznej zostało wcześniej omówione. Drugi rodzaj istotności to istotność praktyczna, która mówi o tym, czy zaobserwowany w badaniach efekt (np. różnica pomiędzy średnimi) może mieć znaczenie praktyczne. Niezależnie od różnego znaczenia, jakie przypisuje się obydwu rodzajom istotności wartości graniczne są w większości przypadków ustalane w sposób umowny. Często można spotkać się z sytuacją w której wykazano istotność statystyczną, ale efekt nie ma żadnego praktycznego znaczenia. W badaniach medycznych zdarza się np., że nowy sposób leczenia daje większą liczbę wyleczonych, ale jednocześnie częściej powoduje trudne do zaakceptowania skutki uboczne albo jest związany z wyższymi kosztami.

Niektórzy autorzy zwracają uwagę na to, że dla badacza pierwszoplanowe znaczenie powinna mieć istotność praktyczna [5]. Argumentują to tym, że istotność praktyczna pozwala badaczowi odpowiedzieć na dwa ważne pytania: (1) jeśli zaobserwowany efekt odzwierciedla rzeczywiste różnice, to jaka jest jego wielkość? i (2) Czy stwierdzony efekt jest wystarczająco duży, aby mógł być użyteczny praktycznie? Dla ilościowej oceny istotności praktycznej stosuje się powszechnie tzw. *wielkość efektu* [1, 2, 3].

Wielkość efektu jest statystyką wykorzystywaną do określenia wielkości efektu badawczego. Przykładowo jest ona stosowana dla określenia wielkości różnicy pomiędzy wartościami średnimi (interesującej badacza zmiennej) dla grup porównywanych w badaniu. Wielkość efektu nie bazuje jedynie na bezwzględnej różnicy pomiędzy średnimi, lecz jest wyrażana w jednostkach odchylenia standardowego. Dzięki temu wielkości efektu obliczane dla różnych zmiennych w tym samym badaniu lub nawet w różnych badaniach mogą być ze sobą porównywane. Wielkość efektu, jak widać, jest niezależna od liczebności próby.

A zatem w przypadku obliczania mocy testu po przeprowadzonych badaniach nie ma problemu z określeniem wielkości efektu. Problem pojawia się w momencie szacowania liczebności próby przed badaniami. Wtedy zaleca się skorzystanie z wyników podawanych przez innych badaczy, którzy prowadzili badania dotyczące podobnych zagadnień. Jeśli wyniki takich badań nie są dostępne, zaleca się przeprowadzenie badań pilotażowych na małej zbiorowości badanych, najlepiej pochodzącej z tej samej populacji docelowej. Jeśli to również nie jest możliwe, wówczas badacz musi po prostu przyjąć jeden lub kilka wariantów wielkości efektu, których się spodziewa. Dobrym rozwiązaniem jest skorzystanie z sugestii Cohena [2], który proponuje przyjęcie trzech wariantów wielkości efektu: 0,2 (mały efekt), 0,5 (średni efekt) i 0,8 (duży efekt).

Ogólny schemat postępowania stosowany przy szacowaniu liczebności próby

Podobnie jak przy ocenie istotności statystycznej za pomocą testów statystycznych, która przebiega w kilku następujących po sobie etapach, również szacowanie wielkości próby wymaga wykonania szeregu podobnych działań. Tak jak przy testowaniu istotności statystycznej należy zacząć od rozważenia hipotez badawczych, których będzie dotyczyło badanie. Hipotezy te trzeba w razie potrzeby przeformułować do postaci hipotez statystycznych (zerowej i alternatywnej). Następnie należy dobrać odpowiedni test statystyczny, który będzie stosowany do formalnej oceny prawdziwości sprawdzanej hipotezy zerowej. Następnie trzeba ustalić oczekiwaną wielkość efektu. W przypadku oceny zróżnicowania wskaźników proporcji wystarczy podać spodziewaną różnicę, natomiast w przypadku zmiennych ilościowych, oprócz przyjęcia oczekiwanej bezwzględnej różnicy pomiędzy średnimi, trzeba przyjąć również wielkość odchylenia standardowego. Na końcu należy ustalić poziom istności testu (α) oraz zakładaną moc testu (β). Ostatnim krokiem jest oszacowanie wielkości próby. Najlepiej zrobić to przy pomocy odpowiedniego oprogramowania komputerowego. Dobrym rozwiązaniem jest skorzystanie z programu *STATISTICA* [7].

Opisane działania można ująć w następujący schemat.

1. Postawienie hipotezy zerowej i alternatywnej.
2. Wybór testu istotności.
3. Dobór oczekiwanej wielkości zróżnicowania.
4. Dobór oczekiwanej wielkości rozproszenia.
5. Przyjęcie poziomu istotności (α).
6. Ustalenie mocy testu (β).
7. Oszacowanie wymaganej liczebności próby.

Obliczanie mocy testu i analiza liczebności próby po przeprowadzeniu badania

Dla zilustrowania praktycznych aspektów obliczania mocy testu i szacowania liczebności próby wykorzystano fragment wyników badania noworodków urodzonych przedwcześnie (pomiędzy 25 a 32 tygodniem ciąży), które po urodzeniu wymagały resuscytacji z wentylacją dodatnim ciśnieniem. W literaturze przedmiotu oraz rekomendacjach brak jednoznacznych zaleceń co do optymalnego stężenia tlenu stosowanego w resuscytacji noworodków urodzonych przedwcześnie. W związku z tym zaplanowano badanie kliniczne z randomizacją, w którym u części noworodków wentylację rozpoczynano od 100% tlenu, w drugiej grupie wentylację rozpoczynano od powietrza atmosferycznego (z zawartością 21% tlenu). Oceniano różnice pomiędzy grupami noworodków w zakresie odsetka zgonów (punkt końcowy pierwotny) oraz odsetek dysplazji oskrzelowo-płucnej (BPD), retinopatii wcześniaków (ROP) oraz krwawień do- i około komorowych (IVH), które stanowiły punkty końcowe wtórne. Zaplanowane badanie uzyskało pozytywną opinię Komisji

Bioetycznej Warszawskiego Uniwersytetu Medycznego. Wyniki badań zostały dla potrzeb niniejszego przykładu udostępnione przez kierującego projektem lek. med. Dariusza Madajczaka z Oddziału Intensywnej Terapii Noworodka Kliniki Neonatologii i Intensywnej Terapii Noworodka Warszawskiego Uniwersytetu Medycznego[1].

W prezentowanym przykładzie zostaną wykorzystane wyniki oceny zróżnicowania odsetka zgonów pomiędzy badanymi grupami noworodków. Wśród noworodków uczestniczących w badaniu stwierdzono 6 zgonów, 1 zgon w grupie noworodków, u których wentylacja rozpoczynała się od powietrza oraz 5 zgonów w grupie, w której wentylację rozpoczynano od 100% tlenu. Umieralność w pierwszej grupie wyniosła 2,3%, a w drugiej 13,9%.

Tabela 1. Liczby i odsetki zgonów i przeżyć w badanych grupach noworodków.

	Podsumowująca tabela dwudzielcza: częstości obserwowane (Noworodki.sta) Liczność oznacz. komórek > 10		
Grupa	**zgon nie**	zgon tak	Wiersz Razem
TLEN	31	5	36
%wiersza	86,11%	13,89%	
POWIETRZE	42	1	43
%wiersza	97,67%	2,33%	
Ogół	73	6	79

Dla oceny istotności zróżnicowania odsetka zgonów pomiędzy badanymi grupami noworodków zastosowano test chi-kwadrat z poprawką Yatesa. Jego wyniki przedstawia poniższa tabela.

Tabela 2. Wyniki testu chi-kwadrat.

	Statystyka: Grupa(2) x zgon(2) (Noworodki.sta)		
statystyka	**Chi-kwadr.**	df	p
Chi^2 Pearsona	3,733259	df=1	p=,05334
Chi^2 NW	3,953829	df=1	p=,04676
Chi^2 Yatesa	2,267412	df=1	p=,13212

Jak widać, w przypadku niniejszych badań ze względu na małe liczebności zgonów w porównywanych grupach noworodków powinno się wziąć pod uwagę wyniki testu chi-kwadrat z poprawką Yatesa. Wyniki przeprowadzonego testu nie dają podstaw do odrzucenia hipotezy zerowej, zakładającej brak różnic, ponieważ prawdopodobieństwo p ($p=0,13212$) przekracza przyjęty standardowy poziom istotności testu $\alpha=0,05$.

W przypadku tak małych liczb zdarzeń (w tym przypadku chodzi o zgony) wykazanie statystycznej istotności zróżnicowania jest trudne, między innymi ze względu na zwykle niską moc testu. W związku z tym, bazując na otrzymanych w niniejszym badaniu wynikach, oszacowano moc testu. Wyniki obliczeń zawiera poniższa tabela.

[1] Autor dziękuje Panu Dariuszowi Madajczakowi za wyrażenie zgody na wykorzystanie wyników przeprowadzonych badań.

Tabela 3. Wyniki analizy mocy testu.

	Moc (Noworodki.sta) Dwie frakcje, test Z H0: Pi1 = Pi2
	Wartość
Frakcja w populacji Pi1	0,1389
Frakcja w populacji Pi2	0,0233
Liczność próby N1	36,0000
Liczność próby N2	43,0000
Prawdop. bł. I rodzaju (Alfa)	0,0500
Moc (z popr. na ciągłość)	0,3304

Otrzymane wyniki pokazują, że rzeczywiście moc testu okazała się bardzo niska i wyniosła 0,3304. Zazwyczaj przyjmuje się, że wystarczająca moc testu to co najmniej 0,80. Aby ocenić, jaka liczba badanych byłaby potrzebna, aby przy stwierdzonej w niniejszych badaniach różnicy odsetka zgonów można było wykazać istotność statystyczną, przeprowadzono szacowanie liczebności próby. Podobnie jak przy analizie mocy testu wykorzystano wyniki niniejszego badania. Założono jednakowe liczebności porównywanych grup noworodków. Wyniki przedstawiono poniżej.

Tabela 4. Wyniki szacowania liczebności próby.

	Liczność próby (Noworodki.sta) Dwie frakcje, test Z H0: Pi1 = Pi2
	Wartość
Frakcja w populacji Pi1	0,1389
Frakcja w populacji Pi2	0,0233
Prawdop. bł. I rodzaju (Alfa)	0,0500
Moc docelowa	0,8000
Moc (z popr. na ciągłość)	0,8003
Liczność próby N1	103,0000
Liczność próby N2	103,0000

Otrzymane wyniki pokazują, że przy założeniu mocy testu na poziomie 0,80 należałoby przebadać po 103 noworodki, czyli łącznie 206 noworodków, aby wykazać, że zróżnicowanie odsetka zgonów pomiędzy badanymi grupami jest istotne na poziomie istotności $\alpha=0,05$.

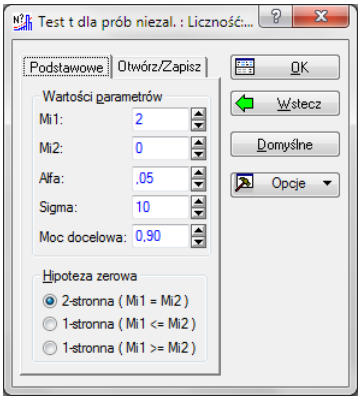
Moduł *Analiza mocy testu* w programie *STATISTICA* umożliwia przeprowadzenie w łatwy sposób szacowania liczebności próby w kilku wariantach.

Szacowanie liczebności próby przed badaniem

Dla badacza zdecydowanie bardziej użyteczna jest możliwość oszacowania liczebności próby na etapie planowania badania. W takiej sytuacji musi on jednak liczyć się z koniecznością podania lub ustalenia wielkości określonych statystyk dotyczących zmiennych, które będą podlegać badaniu.

Przyjmijmy, że badacz planuje przeprowadzenie badania wśród osób, u których zdiagnozowano nadciśnienie (o średnim poziomie skurczowego ciśnienia krwi na poziomie 180 mm Hg). Badane osoby zostaną w sposób losowy rozdzielone na dwie grupy. Jedna z grup będzie przyjmowały lek, który ma obniżyć skurczowe ciśnienie krwi, a druga grupa pozostanie nieleczona. Badacz chciałby sprawdzić, jakie liczebności próby byłyby potrzebne dla wykazania istotności statystycznej różnic dla trzech wariantów wielkości różnicy pomiędzy średnimi badanych grup: 10, 20 i 30 mm Hg, przy założeniu standardowego poziomu istotności testu i jego mocy, tzn. odpowiednio 0,05 i 0,80. Dodatkowo badacz założył, że odchylenie standardowe skurczowego ciśnienia krwi wyniosło w badanej grupie osób 30 mm Hg oraz że przy ocenie istotności statystycznej odpowiedni będzie test *t-Studenta* dla prób niezależnych.

W programie *STATISTICA* zostanie wykorzystany moduł *Analiza mocy testu*. Po wybraniu odpowiedniego testu użytkownik musi podać odpowiednie wartości dla kilku wielkości, które przedstawia zamieszczone poniżej okno programu.



Rys. 2. Okno definiowania parametrów analizy.

Po podaniu odpowiednich wartości dla wymaganych parametrów program obliczy szacowaną liczebność porównywanych grup osób. Poniżej przedstawiono wyniki dla wariantu, w którym założono obniżenie skurczowego ciśnienia krwi o 10 mm Hg.

Tabela 5. Wyniki szacowania liczebności próby (przy różnicy pomiędzy średnimi: 10 mm Hg).

	Liczność próby Dwie średnie, test t, próby niezależne H0: Mi1 = Mi2
	Wartość
Średnia populacyjna Mi1	180,0000
Średnia populacyjna Mi2	170,0000
Odch. std. w populacji (Sigma)	30,0000
Efekt standaryzowany (Es)	0,3333
Prawdop. bł. I rodzaju (Alfa)	0,0500
Wartość krytyczna t	1,9684
Moc docelowa	0,8000
Moc dla wymaganej liczności próby N	0,8021
Wymagane N (w grupie)	143,0000

Jak widać, dla wykazania, że różnica 10 mm Hg jest statystycznie istotna przy przyjęciu $\alpha=0,05$ oraz mocy testu równiej 0,80, potrzeba po 143 osoby w każdej z badanych grup.

Podobna analiza została również przeprowadzona dla dwóch pozostałych wariantów, czyli przy założeniu, że różnica średnich pomiędzy grupami wyniosła 20 i 30 mm Hg. Wyniki analizy przedstawiono poniżej.

Tabela 6. Wyniki szacowania liczebności próby (przy różnicy pomiędzy średnimi: 20 mm Hg).

	Liczność próby Dwie średnie, test t, próby niezależne H0: Mi1 = Mi2
	Wartość
Średnia populacyjna Mi1	180,0000
Średnia populacyjna Mi2	160,0000
Odch. std. w populacji (Sigma)	30,0000
Efekt standaryzowany (Es)	0,6667
Prawdop. bł. I rodzaju (Alfa)	0,0500
Wartość krytyczna t	1,9935
Moc docelowa	0,8000
Moc dla wymaganej liczności próby N	0,8076
Wymagane N (w grupie)	37,0000

Tabela 7. Wyniki szacowania liczebności próby (przy różnicy pomiędzy średnimi: 30 mm Hg).

	Liczność próby Dwie średnie, test t, próby niezależne H0: Mi1 = Mi2
	Wartość
Średnia populacyjna Mi1	180,0000
Średnia populacyjna Mi2	150,0000
Odch. std. w populacji (Sigma)	30,0000
Efekt standaryzowany (Es)	1,0000
Prawdop. bł. I rodzaju (Alfa)	0,0500
Wartość krytyczna t	2,0369
Moc docelowa	0,8000
Moc dla wymaganej liczności próby N	0,8070
Wymagane N (w grupie)	17,0000

Wyniki przeprowadzonej analizy pokazują, że przy większej różnicy pomiędzy średnimi potrzebne byłyby znacznie mniejsze liczby badanych osób. Przy różnicy 20 mm Hg dla wykazania istotności statystycznej na przyjętym poziomie istotności wystarczyłoby 37 badanych, a przy różnicy wynoszącej 30 m Hg liczba ta spada do 17 osób.

Badacz mógłby również sprawdzić inne możliwe warianty analizy, np. przy założeniu większej wartości dla odchylenia standardowego lub przy założeniu wyższej mocy.

Podsumowanie

Oszacowanie wielkości próby jest zagadnieniem, które występuje w większości projektów badawczych. Zazwyczaj badacz musi rozwiązać ten problem już na etapie planowania

badania. Jest to również jedno z tych zagadnień, z którym badacz najczęściej zwraca się do statystyka.

Praktyczne znaczenie wielkości próby wynika z faktu, iż liczebność badanej zbiorowości wpływa zarówno na aspekty techniczno-organizacyjne projektu badawczego (koszty, czas, dostęp do badanych), jak i na wymogi statystyczne (istotność statystyczna wyników oraz moc).

Statystyczne czynniki warunkujące wiarygodność wyników badania, a jednocześnie determinujące wielkość próby to: kryterium istotności statystycznej, moc testu statystycznego oraz wielkość efektu. Przy szacowaniu wielkości próby badacz musi podać lub założyć odpowiednie wartości dla tych wielkości. Dobrym rozwiązaniem jest rozważenie dla tych wielkości kilku wariantów.

Literatura

1. Bausell R. B., Li Y.-F. (2002) Power Analysis for Experimental Research. A Practical Guide for the Biological, Medical and Social Science, Cambridge University Press.
2. Cohen J., (1988) Statistical power analysis for the behavioral sciences (2nd ed.), Hilsdale, New Jersey, Lawrence Erlbaum Associates, Publishers.
3. Cohen J., (1992) A Power Primer, Psychological Bulletin, 112.
4. Harańczyk G., Gurycz J., (2006) Analiza mocy testu i jej znaczenie w badaniach empirycznych, Materiały seminaryjne „Zastosowania statystyki i data mining w badaniach naukowych", StatSoft, Kraków 2006.
5. Kirk R. E., (2001) Promoting Good Statistical Practices: Some Suggestions, Educational and Psychological Measurement, Vol. 61 No. 2, Sage Publications, Inc.
6. Lehmann E. L. (1993) The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association 88.
7. StatSoft, Inc. (2011) *STATISTICA* (data analysis software system), version 10. www.statsoft.com.

