



ANALIZA MOCY TESTU I JEJ ZNACZENIE W BADANIACH EMPIRYCZNYCH

Grzegorz Harańczyk i Jerzy Gurycz, StatSoft Polska Sp. z o.o.

Planując badania empiryczne, badacz coraz częściej nie ogranicza się jedynie do postawienia hipotez, których prawdziwość będzie weryfikował, oraz ustalenia poziomu istotności wyniku, ale także planuje, jak duży efekt uzna za zadowalający i jaką chce uzyskać moc testu. W praktyce moc testu oznacza prawdopodobieństwo podjęcia na podstawie testu decyzji o istnieniu efektu, gdy on rzeczywiście istnieje.

Analiza mocy testu daje wymierne korzyści, nie tylko w wynikach badań, ale także w nakładach na nie. Dowody potwierdzające tę tezę zostaną zaprezentowane na przykładzie zastosowania analizy mocy testu w poszukiwaniu nowych leków.

Zagadnienie analizy mocy testu

Badacz podczas swojej pracy stawia pytania i przypuszczenia, następnie formułuje hipotezy, a przeprowadzając eksperymenty i badania, stara się je zweryfikować. Jednym z głównych zadań statystyki jest właśnie dostarczenie narzędzi do weryfikacji różnego rodzaju hipotez statystycznych, czyli przypuszczeń i sądów dotyczących badanego zjawiska.

Planując badanie, należy odpowiedzieć sobie na wiele pytań: „Jaką dokładność oceny interesującego nas parametru otrzymamy, przy próbie o danej wielkości?”, „Jak duże jest prawdopodobieństwo, że decyzja podjęta na podstawie testu jest prawidłowa?”, czyli „Jak duże zaufanie możemy mieć do wyników naszych badań?” albo „Jak duża musi być próbka, by osiągnąć zamierzony poziom dokładności?”. Zadaniem analizy mocy testu i oceny liczności prób jest właśnie udzielanie odpowiedzi na tak postawione pytania.

W dalszej części pokrótce przedstawimy istotę testów statystycznych oraz analizy mocy testu. Następnie zagadnienie to zostanie zilustrowane przykładem rzeczywistej analizy wykonanej dla jednej z firm zajmujących się poszukiwaniem nowych leków.

Rozpatrywany problem jest problemem dość uniwersalnym i często spotykanym w badaniach empirycznych. Przykład prezentuje kilka aspektów pracy badawczej, ze szczególnym zwróceniem uwagi na błędy II rodzaju popełniane podczas testowania hipotez statystycznych. W opisanym przykładzie te błędy mają kluczową rolę. Problem badawczy polega na tym, że znając skład chemiczny związków, nie sposób dokładnie przewidzieć ich działania,



często brak jest wiedzy o mechanizmach powodujących takie, a nie inne ich działanie na organizmach żywych. Rodzi to potrzebę wsparcia badań teoretycznych badaniami eksperymentalnymi. W tym wypadku badania takie polegają na przeszukiwaniu ogromniej ilości związków i porównywaniu ich działania na wybrane interesujące nas parametry biochemiczne.

Podczas kolejnych etapów przeprowadzana jest selekcja i w kręgu zainteresowań pozostaje coraz mniej substancji i tak aż do fazy badań klinicznych. Przy takim badaniu pojawiają się dwa skrajnie różne cele: z jednej strony mamy wyeliminować jak najwięcej związków nieaktywnych, a z drugiej nie przegapić żadnego potencjalnego leku. Wyniki przedstawione w dalszej części są fragmentem wyników pewnego wstępnego etapu takich badań. Zgodnie ze światowymi standardami poszukiwania nowych leków ich celem było wyselekcjonowanie spośród kilkudziesięciu branych pod uwagę związków kilkunastu, które weszły do dalszych badań in-vivo.

Przy analizowaniu tak ogromniej ilości związków nie można sobie pozwolić na dowolne zwiększenie liczności próby, bo wiązałoby się to z ogromnymi kosztami. Z drugiej strony być może warto zwiększyć liczbę osobników o kilka w celu podniesienia mocy testu. Istotna jest także wiedza, czy nasza procedura badawcza zawodzi, a jeśli tak, to jak często. Szczególnie ważna jest dla nas ocena ryzyka eliminowania z dalszych badań substancji aktywnych, które z jakichś powodów nie zostały uznane za dające efekt. Na te pytania postaramy się znaleźć odpowiedź w dalszej części, a teraz przejdźmy do przypomnienia zasad testowania hipotez statystycznych.

Podstawy teoretyczne

W kontekście testowania istotności statystycznej spotykamy się z dwiema sytuacjami, odrzucająco-potwierdzającą (OP) oraz przyjmująco-potwierdzającą (PP). My skupimy się w naszych rozważaniach na testowaniu odrzucająco-potwierdzającym, czyli sytuacji, gdy hipoteza zerowa jest przeciwieństwem tego, co badacz chciałby wykazać.

W przypadku klasycznej metodyki testowania hipotez statystycznych mamy do czynienia z dwoma błędami, których prawdopodobieństwa tradycyjnie oznacza się jako α i β . Przypomnijmy, że α jest prawdopodobieństwem popełnienia błędu I rodzaju, czyli błędnego odrzucenia hipotezy zerowej, gdy jest ona w rzeczywistości prawdziwa. W kontekście badań, których celem jest wykrycie zróżnicowania (istnienie zróżnicowania nazywać będziemy istnieniem efektu), popełnienie błędu pierwszego rodzaju oznacza stwierdzenie (błędne) na podstawie testu istnienia efektu dla substancji, która tak naprawdę tego efektu nie daje. Wielkość β oznacza prawdopodobieństwo popełnienia błędu drugiego rodzaju, czyli błędu polegającego na nieodrzuconiu fałszywej hipotezy zerowej. W kontekście badania istnienia efektu popełnienie tego błędu oznacza nieodrzuconie hipotezy zerowej, podczas gdy należy ją odrzucić, czyli stwierdzenie (na podstawie testu) braku efektu, podczas gdy efekt istnieje. W badaniach opisanych w przykładzie przekłada się to na „niewykrycie” działającej substancji. Beta można obliczyć jedynie dla konkretnej hipotezy alternatywnej, czyli dla konkretnej wielkości efektu.



| | | Rzeczywistość | |
|---------|-------|-------------------------|-------------------------|
| | | H_0 | H_1 |
| Decyzja | H_0 | Poprawne przyjęcie | Błąd II rodzaju β |
| | H_1 | Błąd I rodzaju α | Poprawne odrzucenie |

Moc testu statystycznego to $1-\beta$, czyli prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona fałszywa. W praktyce moc testu oznacza prawdopodobieństwo podjęcia na podstawie testu decyzji o istnieniu efektu, gdy on rzeczywiście istnieje. Moc testu zależy od konkretnych ustawień testu (w większości testów od obszaru krytycznego statystyki testowej, który to z kolei zależy od liczebności próby, rozproszenia wyników w próbie, przyjętego poziomu istotności α i przyjętych w teście założeń) oraz od wielkości efektu, którego wykrycie interesuje badacza (ta wielkość jest często określana jako efekt standaryzowany). Podobnie jak β , moc można policzyć tylko dla konkretnej alternatywy, czyli dla konkretnej wielkości efektu.

Oznacza to, że w praktyce celowe jest rozpatrywanie złożonych hipotez alternatywnych postaci $H_1 : \mu_1 > \mu_2$ (jednostronne), czyli wielu alternatyw „naraz”. Dla takiej złożonej hipotezy alternatywnej często hipotezę zerową zapisujemy jako $H_0 : \mu_1 \leq \mu_2$, taki zapis hipotezy zerowej nie ma co prawda uzasadnienia matematycznego (hipoteza zerowa musi być w postaci równości, by dało się obliczyć rozkład statystyki testowej i oceniać prawdopodobieństwa), jednak taki zapis jest powszechnie stosowany. Jeszcze raz należy podkreślić, że moc można obliczyć tylko dla konkretnych wartości testowanego efektu.

Procedura testowania

Należy jeszcze raz podkreślić, że my nie badamy prawdziwości hipotez, a jedynie to, która jest bardziej prawdopodobna. Podejmowanie decyzji na podstawie prawdopodobieństwa niesie ze sobą powszechne ryzyko popełnienia błędów, dobrze jest więc rozumieć konsekwencje tych błędów, aby mieć nad nimi kontrolę.

I tak błąd I rodzaju, oznaczający błędne potwierdzenie tezy badacza, może spowodować podjęcie inwestycji i wysiłków, a co najmniej dalszych badań, które nie mają szans powodzenia. Błąd II rodzaju w tym sposobie stawiania hipotez, jest bardzo niekorzystny dla badacza, gdyż jego słuszne przewidywania nie zostały potwierdzone jedynie na skutek losowego błędu. Z punktu widzenia badacza zmniejszenie ryzyka popełnienia błędu drugiego rodzaju powinno być może nawet istotniejsze.

Zwykle przyjmuje się, że częstość popełniania błędu I rodzaju, oznaczana przez α , powinna wynosić co najwyżej 0,05, przy czym częstość β popełniania błędu II rodzaju również nie powinna być duża. Moc testu statystycznego, definiowana jako $1-\beta$, powinna być wysoka. Powinna wynosić co najmniej 0,8, aby zapewnić wykrycie znaczących odstępstw od



hipotezy zerowej. Podane wartości są tylko umownie przyjęte, niezwykle ważną rzeczą jest rozumienie ich znaczenia i konsekwencji, jakie niesie ich zmiana.

W większości typowych sytuacji analitycznych postępowanie przy ocenie mocy testu i wymaganej liczności próby jest takie samo. Po sformułowaniu problemu, wybraniu typu hipotezy ustalamy wartość parametru odpowiadającą hipotezie zerowej. Następnie sprawdzamy moc i licznosc próby dla rozsądnego zakresu spodziewanej wielkości efektu. Kolejnym krokiem może być obliczenie wielkości próby wymaganej do wykrycia efektu o sensownej wielkości (tzn. odchyleniu od hipotezy zerowej) i przy rozsądnej mocy testu.

W przypadku niezadowolających wyników można spróbować tak zmienić parametry analizy i sposób jej przeprowadzania, aby zwiększyć moc testu. Jest kilka czynników znacznie wpływających na moc testów. Jednym z podstawowych jest sam test, który jest wykonywany. Niektóre testy statystyczne mają ze swojej natury większą moc niż inne. Kolejnym ważnym czynnikiem jest licznosc próby oraz wielkość spodziewanego efektu. W ogólnosci im większa próba, tym większa moc testu. Jednak czasem pomiary są drogie i czasochłonne, konieczne jest więc znalezienie licznosci, która będzie „wystarczająco duża” i nie będzie prowadziła do marnowania środków. Jeśli chodzi zaś o wielkość efektu, to jeżeli hipoteza zerowa jest zdecydowanie błędna, to moc testu będzie większa niż przy niewielkich rozbieżnościach. Czynnikiem, nad którym nie mamy bezpośrednio kontroli, są błędy pomiarowe i populacyjne odchylenie standardowe. Zawsze istnieje pewien wkład „szumu” pomiarowego zaciemniającego „sygnał” pochodzący od realnego, poszukiwanego efektu. Każda poprawa dokładności pomiarów i staranności prowadzenia badań poprawia moc testu. Kontrolować moc testu możemy natomiast poprzez poziom istotności α . Im większe jest α , tym większa jest moc testu, oczywiście w niektórych sytuacjach nie możemy pozwolić sobie na zwiększenie ryzyka popełnienia błędów I rodzaju.

Przykład zastosowania

Zasadniczym celem badań opisanych we wstępie była ocena statystycznej istotności różnicowania przeciętnego poziomu stężeń badanych parametrów biochemicznych przy stosowaniu placebo, leku referencyjnego i badanych związków. Działanie badanych związków oceniano na podstawie porównań ze związkiem referencyjnym, którym był lek o znanym działaniu.

Badano następujące parametry biochemiczne:

- ◆ glukoza;
- ◆ trójglicerydy;
- ◆ cholesterol;
- ◆ frakcję HDL („dobry cholesterol”);
- ◆ frakcję LDL („zły cholesterol”).

Pożądanym działaniem badanych związków jest modyfikowanie poziomu tych parametrów (ocena efektu hipoglikemizującego i wpływu na profil lipidowy). My dla potrzeb naszego



przykładu zawężymy się do poziomu glukozy i frakcji LDL, a będziemy porównywać jedynie działanie związku A ze związkiem referencyjnym.

Badania starannie przygotowano i przeprowadzono na specjalnie wyselekcjonowanych myszach. W analizowanej serii badania przeprowadzono na 6 grupach doświadczalnych, liczebność grup wynosiła po 10 osobników. W każdej grupie podawano myszom inną substancję według następującego schematu:

| numer myszy | liczba zwierząt | badany związek |
|-------------|-----------------|----------------|
| 1 – 10 | 10 | Placebo |
| 11 – 20 | 10 | Związek REF |
| 21 – 30 | 10 | Związek A |
| 31 – 40 | 10 | Związek B |
| 41 - 50 | 10 | Związek C |
| 51 – 60 | 10 | Związek D |

Substancje podawano codziennie przez 13 dni. Po upływie tego czasu oznaczono stężenie glukozy, trójglicerydów, cholesterolu całkowitego oraz jego frakcji HDL i LDL. Otrzymane wyniki zestawiono w arkuszu danych.

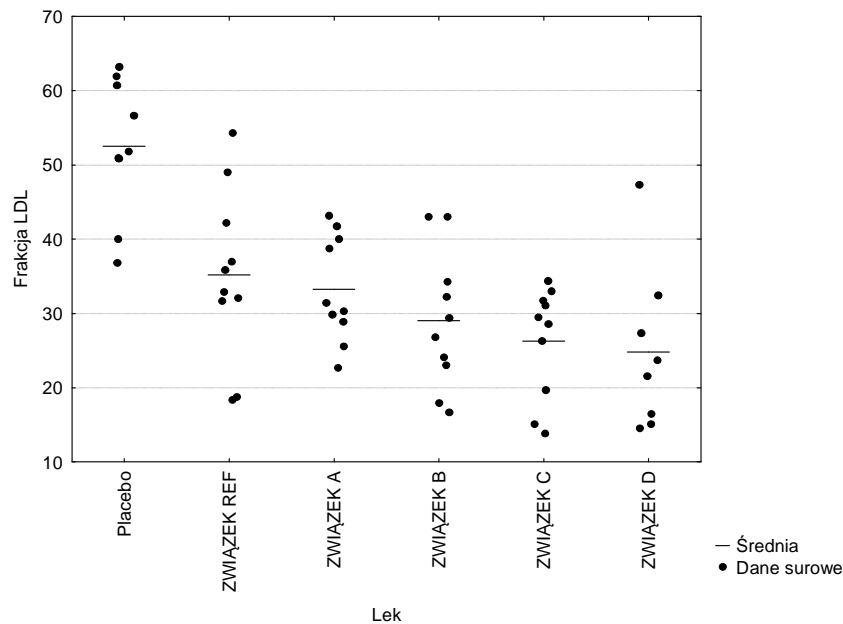
Opis danych

Zbiór danych zawierał 60 przypadków oraz 6 zmiennych.

| | 1 Lek | 2 Stężenie glukozy | 3 Stężenie trójglicerydów | 4 Stężenie cholesterolu | 5 Stężenie frakcji HDL | 6 Stężenie frakcji LDL |
|----|-------------|--------------------------|---------------------------------|-------------------------------|------------------------------|------------------------------|
| 1 | Placebo | 122,14 | 132,67 | 285,56 | 116,67 | 96,67 |
| 2 | Placebo | 547,62 | 73,33 | 107,22 | 105,00 | 50,97 |
| 3 | Placebo | 367,62 | 48,00 | 113,89 | 98,33 | 56,67 |
| 4 | Placebo | 272,38 | 59,33 | 89,44 | 103,33 | 60,69 |
| 5 | Placebo | 346,90 | 66,67 | 110,00 | 71,67 | 50,83 |
| 6 | Placebo | 193,10 | 46,00 | 112,78 | 90,00 | 36,81 |
| 7 | Placebo | 112,14 | 81,33 | 156,11 | 80,00 | 63,19 |
| 8 | Placebo | 410,24 | 100,00 | 215,00 | 85,00 | 61,94 |
| 9 | Placebo | 121,90 | 93,33 | 122,22 | 63,33 | 40,04 |
| 10 | Placebo | 159,29 | 58,00 | 158,89 | 86,67 | 51,81 |
| 11 | ZWIĄZEK REF | 192,86 | 81,33 | 74,44 | 68,33 | 32,92 |
| 12 | ZWIĄZEK REF | 229,02 | 49,33 | 76,11 | 50,00 | 32,08 |
| 13 | ZWIĄZEK REF | 207,38 | 53,33 | 68,89 | 70,00 | 36,94 |
| 14 | ZWIĄZEK REF | 194,76 | 65,33 | 117,22 | 70,00 | 54,31 |
| 15 | ZWIĄZEK REF | 154,52 | 101,33 | 140,56 | 100,00 | 49,03 |
| 16 | ZWIĄZEK REF | 258,10 | 60,67 | 82,22 | 50,00 | 31,67 |
| 17 | ZWIĄZEK REF | 311,90 | 48,00 | 92,22 | 43,33 | 18,75 |
| 18 | ZWIĄZEK REF | 297,38 | 50,67 | 67,22 | 35,00 | 18,33 |
| 19 | ZWIĄZEK REF | 201,43 | 100,67 | 117,22 | 91,67 | 35,83 |
| 20 | ZWIĄZEK REF | 235,67 | 88,67 | 104,44 | 93,33 | 42,22 |
| 21 | ZWIĄZEK A | 165,95 | 86,11 | 90,00 | 56,67 | 25,56 |
| 22 | ZWIĄZEK A | 88,10 | 72,78 | 127,14 | 74,44 | 30,32 |
| 23 | ZWIĄZEK A | 136,43 | 46,11 | 89,05 | 47,78 | 28,89 |
| 24 | ZWIĄZEK A | 206,90 | 43,33 | 114,29 | 72,22 | 43,17 |
| 25 | ZWIĄZEK A | 93,14 | 45,00 | 116,10 | 58,89 | 38,73 |

Najpierw oceniano szczegółowo wyniki oznaczeń badanych parametrów krwi pod kątem odstających obserwacji. Na podstawie położenia punktów na wykresie rozrzutu podejmowano decyzję o ewentualnym odrzuceniu odstających punktów. Za odstające uznano pomiary, które w zdecydowany sposób odbiegały od pozostałych (ich nieodrzućenie mogłoby powodować przeszacowanie lub niedoszacowanie przeciętnego poziomu danej substancji w badanej grupie zwierząt). Odrzucanie obserwacji w przypadku tak małych prób musi być jednak dokonywane bardzo ostrożnie, ponieważ może przynieść odwrotny od zamierzonego efekt.

Dla tak przygotowanych danych utworzono wykresy danych surowych mające na celu ułatwienie porównań międzygrupowych. Poniżej znajduje się wykres danych surowych z zaznaczonymi średnimi dla poziomu frakcji LDL dla badanych związków i placebo (na wykresie nie ma już obserwacji odstających). Wykres ten pozwala dokładnie prześledzić różnice, zarówno pod kątem zróżnicowania przeciętnego poziomu stężenia badanego parametru w poszczególnych grupach porównawczych, jak również rozrzutu surowych wyników w obrębie grup.



Cel analizy

Chcemy sprawdzić, czy średni poziom stężenia badanych wskaźników (cholesterol, glukoza, LDL) jest niższy w grupie, w której myszy traktowano badanym związkiem, w porównaniu do średniego poziomu w grupie, w której podawano lek referencyjny $\mu_{BADANY} < \mu_{REF}$. Tezą naszych badań jest po prostu sprawdzenie, czy któryś z badanych związków obniża interesujące nas parametry biochemiczne bardziej niż lek referencyjny. Hipoteza zerowa będzie więc zaprzeczeniem naszej tezy, toteż $H_0 : \mu_{BADANY} \geq \mu_{REF}$. Hipoteza ta będzie testowana dla każdej z badanych substancji.



Do analizy różnic pomiędzy wartościami średnich uzyskanych w wyniku przeprowadzenia oznaczeń w surowicach myszy traktowanych zawiesiną kontrolną (lek referencyjny) i badanymi związkami wykorzystano test t dla zmiennych niezależnych. Test t jest powszechnie stosowaną metodą oceny różnic między średnimi w dwóch grupach. Mogą to być próby niezależne (np. sprawdzenie różnicy ciśnienia krwi w grupie pacjentów poddanych działaniu jakiegoś leku w stosunku do grupy otrzymujących placebo) lub zależne (np. sprawdzenie różnicy ciśnienia krwi u pacjentów „przed” i „po” podaniu leku). Teoretycznie test t może być stosowany także w przypadku bardzo małych prób, jedynym warunkiem jest normalność rozkładu zmiennych oraz brak istotnych różnic między wariancjami.

Badane przez nas dane spełniają te założenia, jednak wstępne wyniki analiz nie były jednoznaczne.

W wielu przypadkach nie było podstaw do odrzucenia hipotezy zerowej, czyli uznania któregoś ze związków za lepszy niż związek referencyjny. Nasz test nie wykrywa efektów działania badanych substancji, a jego celem było wyselekcjonowanie kilkunastu związków do dalszych badań. Istnieje obawa, że niektóre ze związków nie działają z jakichś przyczyn w badanych warunkach lub ich działanie na tak małej próbie nie dało zauważalnego efektu lub zostało stłumione przez błędy losowe. Odrzucając dany związek, chcemy mieć „czyste sumienie”, być w miarę możliwości pewni, że nie straciliśmy szans na wykrycie dobrze działającego leku.

Celem dalszej analizy stało się więc obliczenie mocy testu oraz poprawienie procedury tak, aby w dalszych etapach wyniki były bardziej jednoznaczne.

Aby obliczyć moc testu w naszym przypadku należy podać populacyjne średnie dla obu populacji, licznosc pierwszej i drugiej grupy, poziom α (prawdopodobieństwo popełnienia przy testowaniu błędu I rodzaju, u nas ustalone na poziomie 0,05) oraz wartość populacyjnego odchylenia standardowego, wspólnego dla obu próbkowanych populacji. Jak już wcześniej zostało to wspomniane, analiza będzie prowadzona dla hipotezy zerowej jednostronnej $H_0 : \mu_1 \leq \mu_2$.

W dalszej części pracy zamieszczono wyniki analizy mocy testu na przykładzie stężenia glukozy oraz frakcji LDL cholesterolu (są to dwie skrajności, gdzie moc jest najmniejsza i największa). Analiza w przypadku pozostałych parametrów biochemicznych przebiega bardzo podobnie.

Wyniki analizy mocy testu w przypadku oceny zróżnicowania wartości przeciętnych frakcji LDL cholesterolu

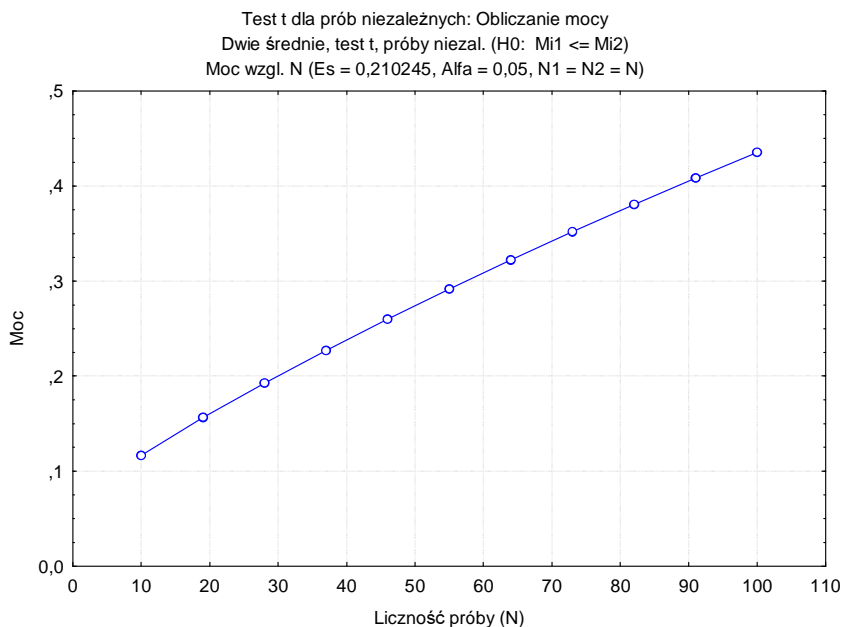
W tej części zostaną przedstawione wyniki analizy mocy testu w przypadku oceny istotności zróżnicowania przeciętnego poziomu frakcji LDL pomiędzy badanymi grupami myszy. Tak jak to zostało wcześniej powiedziane przy ocenie mocy testu za wartość referencyjną przyjęto poziom frakcji LDL dla związku referencyjnego (ZWIĄZEK REF).

Analiza mocy testu wykonywana była przy ustalonym poziomie $\alpha=0,05$. Zamieszczona poniżej tabela zawiera wyniki analizy.

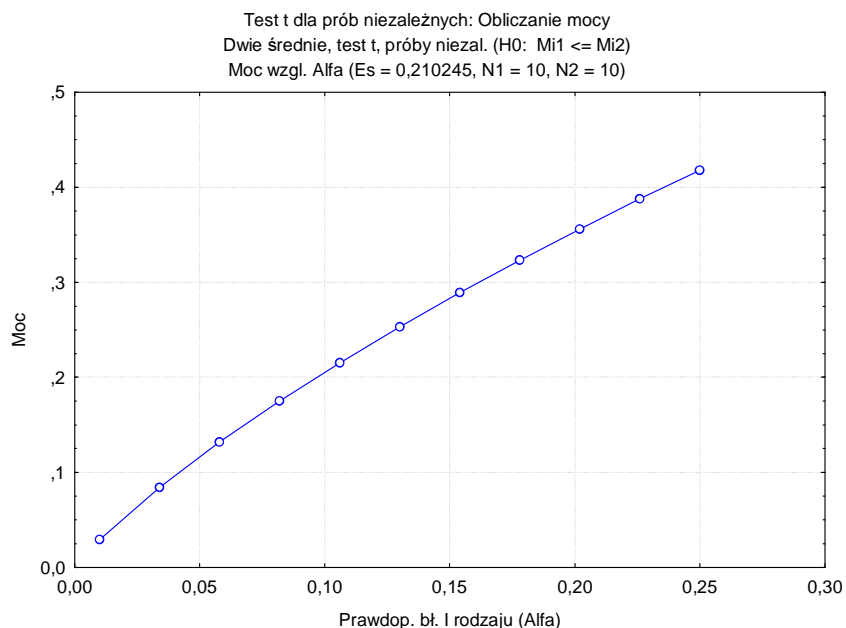
| | Dwie średnie, test t, próby niezal. H0: Mi1 <= Mi2 |
|--------------------------------|---|
| | Wartość |
| Średnia populacyjna Mi1 | 35,2100 |
| Średnia populacyjna Mi2 | 33,2400 |
| Odch. std. w populacji (Sigma) | 9,3700 |
| Efekt standaryzowany (Es) | 0,2102 |
| Liczność próby N1 | 10,0000 |
| Liczność próby N2 | 10,0000 |
| Prawdop. bł. I rodzaju (Alfa) | 0,0500 |
| Wartość krytyczna t | 1,7341 |
| Moc | 0,1166 |

Jak widać, obliczona moc testu wyniosła zaledwie blisko $0,1$. Możemy zatem stwierdzić, że przy przyjętych ustawieniach (średnie wartości frakcji LDL dla związku A i związku referencyjnego, wielkość odchylenia standardowego oraz liczebności) prawdopodobieństwa podjęcia decyzji o istnieniu efektu (na podstawie testu), gdy on rzeczywiście istnieje, jest zbyt niskie (przypomnijmy że, „dobra” moc testu to przynajmniej $0,8$).

Na podstawie wcześniejszych rozważań wiemy, że praktyczne możliwości zwiększenia mocy testu wiążą się z koniecznością zwiększenia liczebności próby, podwyższenia poziomu istotności testu α lub zmniejszeniem rozrzutu wyników. Dwie pierwsze możliwości można sprawdzić za pomocą odpowiednich wykresów. Wykresy te zamieszczono poniżej.



Wykres pozwala zauważyć, że przy przyjętej wielkości efektów dla leku referencyjnego uzyskanie wysokiej wartości mocy testu (np. rzędu $0,8$) wymagałoby bardzo dużej liczby obserwacji. Drugi z utworzonych wykresów pozwala ocenić moc testu w zależności od przyjmowanego poziomu istotności testu α .



Wykres pokazuje, że przyjęcie wyższego poziomu istotności testu (np. na poziomie 0,25) nadal nie gwarantuje zbyt wysokiej mocy testu. Oznacza to, że prawdopodobieństwo wykrycia istotnego efektu jest bardzo niskie. Powodem takiej sytuacji może być zróżnicowanie międzyosobnicze pomiędzy badanymi myszami, które ma wpływ na ich reakcje, lub niewystarczająca dokładność pomiaru badanych wskaźników. Nie musi to natomiast oznaczać, że wielkość frakcji LDL nie jest dobrym markerem do oceny efektu działania badanego związku.

Wyniki analizy mocy testu w przypadku oceny zróżnicowania wartości przeciętnego stężenia glukozy

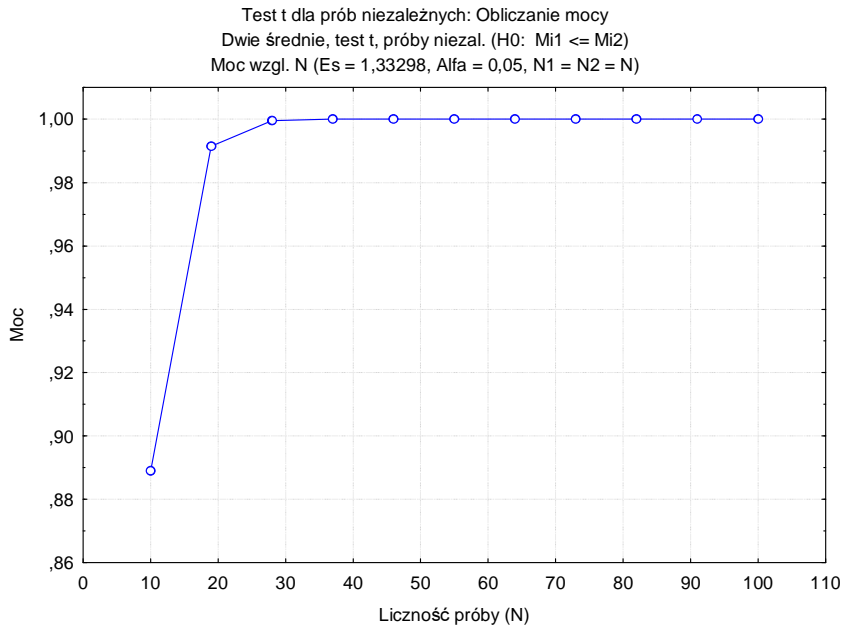
W tej części pokażemy wyniki analizy mocy testu w przypadku oceny istotności zróżnicowania przeciętnego poziomu glukozy pomiędzy badanymi grupami myszy. Tak jak powyżej będziemy porównywać działanie związku A ze związkiem referencyjnym, przy prawdopodobieństwie błędów I rodzaju ustalonym na poziomie 0,05. Zamieszczona poniżej tabela zawiera wyniki analizy.

| | Dwie średnie, test t, próby niezal. H0: $\mu_1 \leq \mu_2$ |
|--------------------------------|---|
| | Wartość |
| Średnia populacyjna μ_1 | 228,3000 |
| Średnia populacyjna μ_2 | 140,2700 |
| Odch. std. w populacji (Sigma) | 66,0400 |
| Efekt standaryzowany (Es) | 1,3330 |
| Liczność próby N1 | 10,0000 |
| Liczność próby N2 | 10,0000 |
| Prawdop. bł. I rodzaju (Alfa) | 0,0500 |
| Wartość krytyczna t | 1,7341 |
| Moc | 0,8889 |

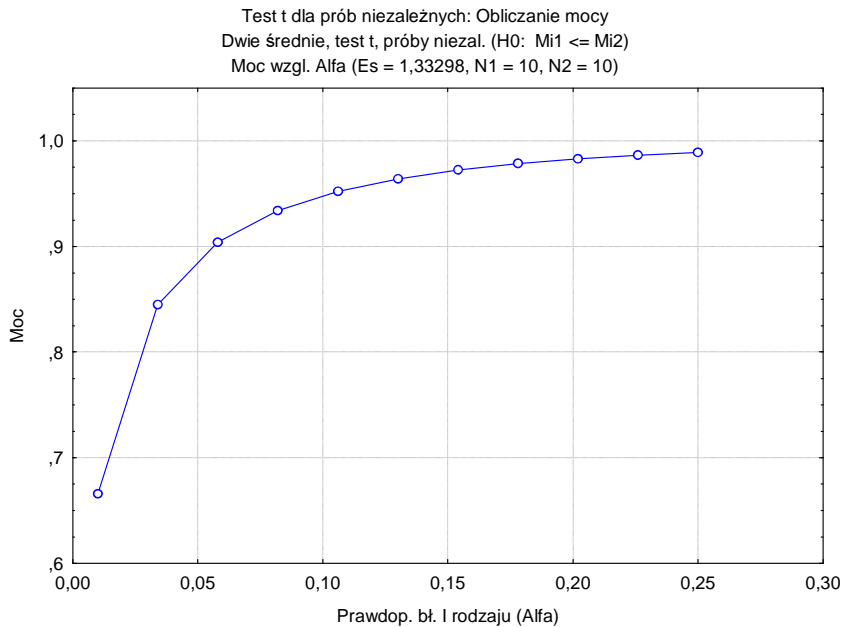


Jak widać, w przypadku stężenia glukozy obliczona moc testu wyniosła ponad 0,88. Możemy zatem stwierdzić, że przy przyjętych ustawieniach (średnie wartości poziomu glukozy dla ZWIĄZKU REF i ZWIĄZKU A, wielkość odchylenia standardowego oraz liczebności) prawdopodobieństwo podjęcia decyzji o istnieniu efektu (na podstawie testu), gdy on rzeczywiście istnieje, jest na bardzo dobrym poziomie.

Podobnie jak poprzednio, poniżej zamieszczono wykresy prezentujące zależność mocy testu od liczebności próby i przyjętego poziomu istotności testu.



Wykres pozwala zauważyć, że przy przyjętej wielkości efektów dla leku referencyjnego liczebność rzędu 10-15 obserwacji jest już wystarczająca. Drugi z utworzonych wykresów pozwala ocenić moc testu w zależności od przyjmowanego poziomu istotności testu α .





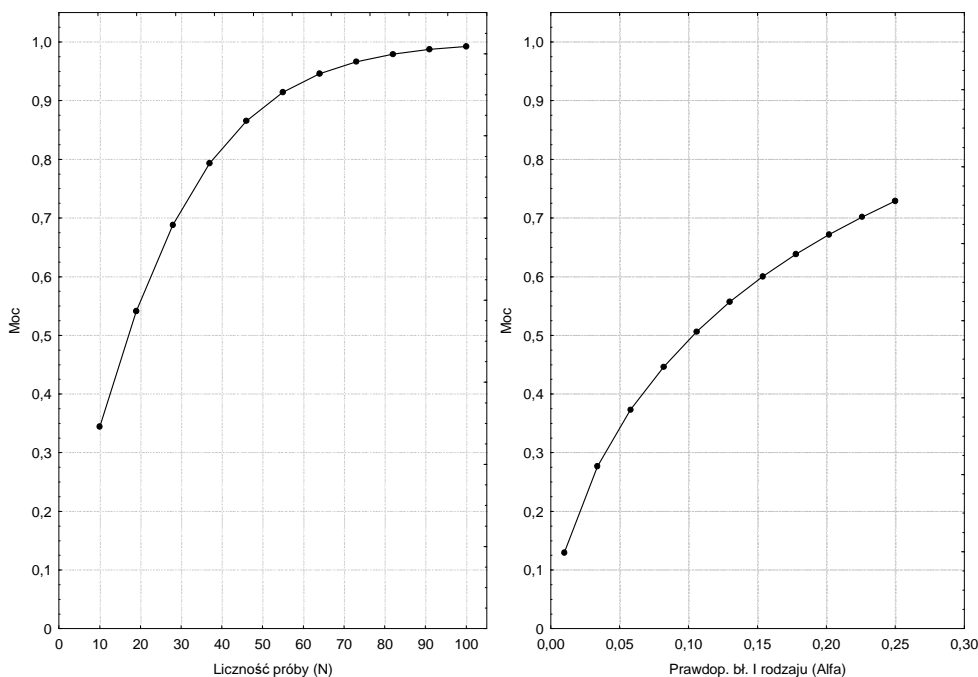
Wykres pokazuje, że przyjęcie poziomu istotności $\alpha=0,05$ daje „dobrą” moc testu (powyżej 0,88). Gdybyśmy przyjęli poziom istotności na wyższym poziomie, wówczas moc testu wzrasta nawet powyżej 0,95.

Ocena wyników analiz

W wynikach analizy dla frakcji LDL warto zwrócić uwagę na jedną rzecz, mianowicie na stosunek β do α , który można dość łatwo zinterpretować. W naszym przykładzie α zostało przyjęte na poziomie 0,05, natomiast β , równe *1-moc testu*, wyniosło 0,884. Tak więc ryzyko błędnego odrzucenia hipotezy zerowej jest oceniane przez planującego badanie jako 17 razy bardziej poważne niż błędnego zaakceptowania.

Przyjrzyjmy się jeszcze wynikom oceny istotności zróżnicowania przeciętnego poziomu frakcji LDL pomiędzy badanymi grupami myszy dla wszystkich związków badanych w tej serii. Porównanie średnich dla leku referencyjnego i związku A zostało już omówione, w przypadku pozostałych związków. Dla związków B, C oraz D sytuacja jest podobna, moc testu jest na poziomie 0,3-0,5, a zależność mocy testu od licznosci próby i poziomu α przedstawia się następująco:

Zależność mocy testu od licznosci próby i prawdopodobieństwa błędu I rodzaju



Badacz ma w tym miejscu do rozwiązania typowe zadanie optymalizacyjne. Wydaje się, że w dalszych badaniach zwiększenie licznosci próby jest najkorzystniejszym rozwiązaniem, jednak należy zwrócić uwagę także na koszty badań. Na tym etapie badań zamiast zwiększać licznosc próby dwukrotnie, co oczywiście powoduje ogromne zwiększenie kosztów badań, można po prostu post factum zwiększyć poziom α . W praktyce często uważa się, że im mniejsze α , tym lepiej, jednak powoduje to obniżenie mocy testu. Należy



w tym miejscu przypomnieć sobie cel naszych badań, zgodnie z nim podniesienie poziomu α nie jest tak szkodliwe.

W świetle uzyskanych wyników i powyższych rozważań procedurę przeprowadzania badań należałoby wzbogacić o następujący algorytm postępowania:

1. Należy określić, stężenia których substancji są krytyczne, i jaka wielkość efektu będzie oznaczać, że dana substancja jest aktywna.
2. Następnie należy odpowiednio dobierać licznosc próby oraz poziom istotności α . W ten sposób praktycznie oceniamy, jak często dopuszczalne są pomyłki. Sugerowana jest analiza sekwencyjna. Na przykład przyjęcie w pierwszej fazie badań wysokiego poziomu alfa (powiedzmy $\alpha = 0,2$), tak aby test miał jak największą moc, i dopuszczenie tym samym błędnego sygnału o występowaniu efektu, gdy w rzeczywistości go nie ma (w przykładzie to oznacza dwa razy na dziesięć) i dalsza ocena wyselekcjonowanych substancji już przy niższym poziomie α .
3. Bardzo duży wpływ na moc ma wielkość rozrzutu wyników. Jej kontrola jest niezwykle ważna. Sugeruje się, aby w następnej fazie badań spróbować zmniejszyć rozrzut wyników.

Taka poprawa procedury nie tylko obniża koszty, ale przede wszystkim zwiększa szanse na sukces w tym wieloetapowym i bardzo długim procesie badawczym.

Podsumowanie

Wykorzystywanie możliwości, jakie daje analiza mocy testu oraz wyznaczanie licznosci prób, często daje duże oszczędności przy eksperymentach, a także pomaga poprawić skuteczność testów. Aby oszacować minimalną licznosc próby, jaka daje satysfakcjonujące rezultaty lub obliczyć moc testu, należy jedynie znać lub mieć dobre estymatory wielkości efektu lub średnich w grupach i odchylenia standardowego w badanej populacji. Jeżeli jesteśmy w stanie uzyskać takie informacje poprzez przeprowadzenie badań pilotażowych, skorzystanie z wyników wcześniejszych badań lub prowadzimy badania powtarzalne, wieloetapowe, to na pewno warto przeprowadzić analizę mocy testu.

Analiza mocy testu staje się bardzo ważnym narzędziem w pracy badacza, szczególnie na etapie planowania doświadczeń, ale także już po zakończeniu badań, jako element oceny skuteczności podjętych działań.

Literatura

1. StatSoft (2006). Elektroniczny Podręcznik Statystyki PL, Kraków, WEB: <http://www.statsoft.pl/textbook/stathome.html>.
2. J. Cohen, Statistical power analysis for the behavioral sciences, LEA Publishers 1988.
3. S. McKillup, Statistics Explained, An Introductory Guide for Life Scientists, Cambridge 2006.