Jest to zapis odczytu wygłoszonego na XXXV Szkole Matematyki Poglądowej, *Porządek*, sierpień 2005.

Porządki w statystyce

Andrzej DĄBROWSKI, Wrocław

Pewnego letniego popołudnia w końcu lat 20 zebrała się w Cambridge grupka przyjaciół. Przy okrągłym stole na tarasie zasiedli profesorowie uniwersytetu wraz z żonami. Patrząc na filiżanki wypełnione herbatą z mlekiem jedna z dam stwierdziła: – A ja umiem rozpoznać czy najpierw nalano herbatę, czy mleko. Być może stwierdzenie to przeszłoby bez echa, gdyby nie uwaga, może niezbyt grzeczna, jednego z gości: – To trzeba sprawdzić! I natychmiast przed zaskoczonymi słuchaczami około 30-letni mężczyzna, z brodą w stylu van Dycka, przedstawił plan doświadczenia. Część filiżanek – jak proponował – bez wiedzy osoby testowanej zostałaby wypełniona najpierw herbatą a później mlekiem, zaś część, w odwrotnej kolejności. I dopiero wtedy, oceniając liczbę prawidłowo rozpoznanych sytuacji, przekonamy się, czy rzeczywiście można rozróżnić kiedy nalano mleko do herbaty. – Ale czy eksperyment przekona nas o tym na pewno? – pytano.

Opis tej rozmowy, wraz ze szczegółowo rozpisanymi wariantami doświadczenia, znalazł się w drugim rozdziale książki, napisanej w 1935 roku przez bohatera tej anegdoty – Ronalda Fishera[i]. Książka ta – *The Design of Experiments* – wraz z wydaną 10 lat wcześniej *Statistical Methods for Research Workers* była przełomem w nauce XX wieku. Odtąd wyniki doświadczeń zostały poddane wnikliwej analizie, a wnioski z tej analizy oparte zostały na gruncie ścisłego rozumowania matematycznego. Paradoksalnie, zasady dedukcji obowiązujące w matematyce musiały być zrewidowane, aby je dostosować do specyfiki eksperymentalnego poznawania rzeczywistości.

[i] Ronald Fisher (1890–1962)

W matematyce kryterium, czy interesujący nas obiekt ma jakąś własność, musi być bezbłędne. Na przykład, gdy chcemy rozstrzygnąć, czy liczba jest parzysta, można sprawdzić parzystość jej ostatniej cyfry. Kryterium ostatniej cyfry jest bezbłędne: gdy napotkamy liczbę, której ostatnia cyfra jest parzysta, mamy pewność, że jest to liczba parzysta. Gdy napotkamy liczbę, której ostatnia cyfra jest nieparzysta, mamy pewność, że ta liczba nie może być parzysta.

Dla eksperymentu z herbatą nawet prawidłowe rozpoznanie kolejności dodawania mleka dla wszystkich filiżanek nie daje pewności, że osoba testowana tę umiejętność posiada. Jak w takim razie znajdować testy, które najlepiej potrafią wykrywać jakąś własność? I jak ocenić ich jakość?

Testy

Odpowiedź na to ważne dla nauki pytanie dały prace z lat 1928 -1933 Jerzego Neymana[ii] i Egona Pearsona[iii]. Od początku XX wieku zaczęto proponować różne kryteria, zwane testami, mające weryfikować hipotezy pojawiające się w naukach doświadczalnych. Ojciec Egona – Karl – był twórcą, w 1900 roku, jednego z pierwszych takich testów – słynnego testu χ^2 – którego po raz pierwszy użyto do sprawdzenia rzetelności ruletki w Monte Carlo. Paradoksy pojawiające się przy stosowaniu testów wymagały wyjaśnienia teoretycznego, a zasady ich tworzenia – kodyfikacji.

[ii] Jerzy Neyman (1894–1981)

[iii] Egon Pearson (1895–1980)

Testy, tworzone na początku XX wieku, były testami, mającymi wykazać istotność weryfikowanej hipotezy. Poziom istotności hipotezy, określany był jako prawdopodobieństwo, że użyty test stwierdzi, iż jest ona fałszywa, gdy w rzeczywistości była prawdziwa.

W rozumowaniu matematycznym wszystkie kryteria (testy) mają poziom istotności 0. Na przykład, prawdopodobieństwo że ostatnia cyfra liczby parzystej jest nieparzysta wynosi 0. Ogłaszając komunikat, że według ustalonych przez nas kryteriów (według wybranego przez nas testu) testowana osoba nie potrafi odróżnić, kiedy wlano mleko do herbaty, musimy z dodatnim prawdopodobieństwem się pomylić.

Testy istotności były konstruowane tak, aby miały jak najniższy poziom istotności, czyli uwzględniały sytuację, gdy hipoteza była prawdziwa. Neyman

zauważył, że przyczyną wielu paradoksów jest brak kontroli nad zachowaniem się testu, gdy hipoteza nie jest prawdziwa. Co więcej, tak naprawdę weryfikuje się prawdziwość zdania przeciwnego niż rozważane. Badając rzetelność ruletki w Monte Carlo, test χ^2 musiał zawierać kryterium mierzące na podstawie danych jej *nierzetelność*. Dopiero gdy ta nierzetelność jest mało prawdopodobna można wnioskować, że ruletka jest rzetelna. Przypomina to rozumowanie niewprost, używane w matematyce.

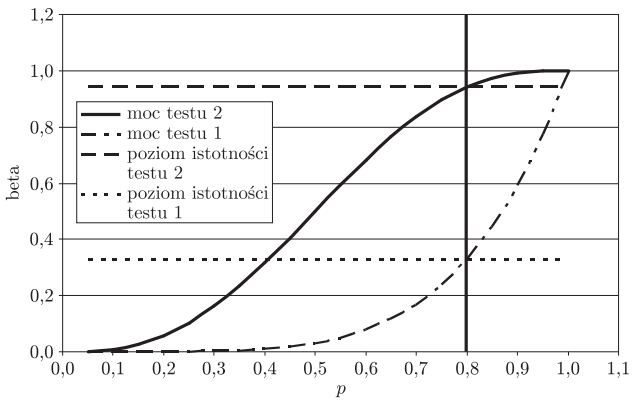
Neyman podzielił zbiór zdań o badanym obiekcie na dwie grupy: ta, do której przynależność jest weryfikowana przez test – i ta grupa nazywana jest *hipotezą* konkurencyjną, oraz grupę zdań, zwaną *hipotezą zerową*, której test nie weryfikuje. Na przykład, gdy chcemy sprawdzić, czy jesteśmy chorzy na grypę, czy nie, i posługujemy się termometrem to „testem na grypę" jest przekroczenie czerwonej kreski na termometrze. Tu „grypa" jest hipotezą alternatywną a wszystkie stany zdrowia, inne niż grypa stanowią hipotezę zerową.

Test $T(x)$, gdzie x jest wektorem, którego współrzędne są obserwowanymi wartościami, stanowi warunek, który jest funkcją zebranych danych. Spełnienie tego warunku oznacza przyjęcie hipotezy alternatywnej. Jeżeli ten warunek ma postać nierówności $T(x) \geq c$, to stała c nosi nazwę *wartości krytycznej* testu.

Narzędziem do badania jakości testu, wprowadzonym w 1933 roku przez Neymana i Pearsona jest *funkcja mocy* testu $\beta(h)$, gdzie parametr h przebiega zbiór wszystkich rozważanych hipotez, należących do hipotezy zerowej lub alternatywnej. Funkcja mocy jest prawdopodobieństwem, że test jest spełniony, gdy prawdziwa jest hipoteza h. Należy oczekiwać, że w dobrze skonstruowanym teście jego moc ma duże wartości dla hipotezy alternatywnej, a małe dla hipotezy zerowej. Maksymalna wartość funkcji mocy na zbiorze, określanym przez hipotezę zerową, zwana jest *poziomem istotności* (rozmiarem) testu.

Na przykład w teście „termometrowym" moc oznacza prawdopodobieństwo, że słupek rtęci przekroczy czerwoną kreskę. Dla hipotezy alternatywnej, czyli gdy pacjent jest chory na grypę, prawdopodobieństwo to jest wysokie, a dla hipotezy zerowej, gdy nie jest chory na grypę, prawdopodobieństwo to jest znacznie mniejsze.

Porównajmy funkcje mocy dla dwóch testów herbacianych z hipotezą alternatywną: „Osoba badana potrafi odróżnić kolejność mleka i herbaty" i hipotezą zerową: „Osoba badana nie potrafi odróżnić kolejności mleka i herbaty", gdy próbuje się 5 filiżanek herbaty. Niech $T(x)$ będzie liczbą poprawnie zidentyfikowanych filiżanek, x wektorem (losowym) decyzji, do której filiżanki nalano najpierw mleko. Test T_1 niech ma wartość krytyczną 5, test T_2 – wartość krytyczną 3. W teście T_1 uznajemy, że osoba badana potrafi odróżnić kolejność mleka i herbaty, gdy poprawnie zidentyfikuje 5 filiżanek, w teście drugim – gdy zidentyfikuje co najmniej 3 filiżanki.



Rys. 1

Wartość funkcji mocy zależy od tego, co oznacza sformułowanie „potrafi odróżnić". Realistyczna definicja powinna dopuścić popełnianie błędów przy rozpoznawaniu. Na przykład można przyjąć, że „potrafi odróżnić" oznacza, iż prawdopodobieństwo poprawnej odpowiedzi jest większe niż 0,8 dla każdej filiżanki, a „nie potrafi odróżnić" – prawdopodobieństwo jest nie większe niż 0,8. Obie hipotezy dają się więc opisać za pomocą parametru p, a moc jest funkcją tego parametru. Takie hipotezy nazywamy parametrycznymi.

Dla testu T_1 jego funkcja mocy wyraża się wzorem $\beta_1(p) = p^5$, dla testu T_2 moc jest postaci $\beta_2(p) = 10p^3(1-p)^2 + 5p^4(1-p) + p^5$.

Moc testu T_2 na obszarze hipotezy „potrafi odróżnić" ($p > 0{,}8$) jest większa niż moc testu T_1. Oznacza to, że T_2 lepiej niż T_1 wykrywa zdolności do odróżniania

kolejności wlewania mleka. Odbywa się to jednak kosztem wartości funkcji mocy na obszarze hipotezy „nie potrafi odróżnić" ($p \leq 0{,}8$), która mierzy prawdopodobieństwo błędnej decyzji podjętej na podstawie testu. Maksymalne prawdopodobieństwo błędu, a więc poziom istotności wynosi 0,94 dla testu T_2, a 0,33 dla testu T_1.

Reguła wyboru, zaproponowana jeszcze przez Fishera, głosi, że w pierwszej kolejności wybieramy testy, których poziom istotności, a więc maksymalne prawdopodobieństwo błędu, nie przekracza z góry ustalonej wartości. Spośród nich, o ile to jest możliwe, wybieramy test o największej mocy na obszarze hipotezy alternatywnej. Zazwyczaj jako dopuszczalną granicę maksymalnego prawdopodobieństwa błędu przyjmuje się wartość 0,05[iv].

[iv] W teście herbacianym dla 5 filiżanek nie istnieje taka wartość krytyczna, aby istniał test na poziomie istotności 0,05. Dopiero przy 14 filiżankach można taki test skonstruować.

Porównanie średnich

Podstawowym problemem, z którym ma do czynienia statystyk, jest porównanie danych. Pytania: czy nowy lek jest lepszy od starego, jaki sposób treningu daje najlepsze wyniki na zawodach, należy odczytać jako porównanie wartości typowych, zazwyczaj reprezentowanych przez wartości średnie. A więc pytamy, czy średnia liczba wyleczonych z użyciem nowego leku jest większa niż średnia liczba osób wyleczonych z użyciem starego leku. Pytamy, czy średnie czasy uzyskane przez zawodników są różne dla różnych sposobów treningu, a jeśli tak, to w jaki sposób można je uporządkować?

Najprostszym zagadnieniem tego typu jest pytanie dotyczące porównania średnich μ_1 i μ_2 dla pewnej cechy liczbowej, charakteryzującej grupy G_1 i G_2. Dane o wartościach tej cechy w grupie G_1 zostały zebrane w postaci k niezależnych pomiarów $X_1, X_2, \ldots, X_k$, dane o wartościach w grupie G_2 zostały zebrane w postaci l niezależnych pomiarów $Y_1, Y_2, \ldots, Y_l$. Pomiary dla obu grup też zostały zebrane niezależnie od siebie. Hipoteza alternatywna H_1 ma postać: $\mu_2 > \mu_1$, hipoteza zerowa $\mathrm{H}_0 : \mu_2 \leq \mu_1$.

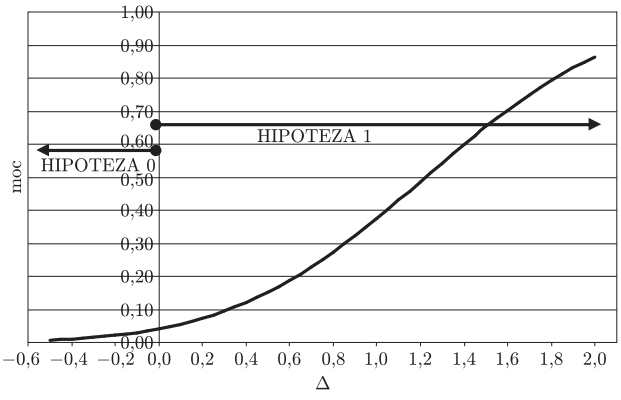
[v] William Gosset (1876–1937)

Problem ten rozwiązał w 1908 William Gosset[v], posługujący się pseudonimem Student. Założył on, że cecha X ma rozkład normalny $N(\nu_i, \sigma)$ o średniej μ_i w grupie G_i oraz o wspólnym odchyleniu standardowym w obu grupach. Test **t Studenta** o wartości krytycznej t_0 jest postaci $t \geq r_0$. Wyrażenie $t = \dfrac{\overline{Y} - \overline{X}}{s}\sqrt{\dfrac{kl}{k+l}}$ zwane jest statystyką t Studenta. Wielkości $\overline{X}$ i $\overline{Y}$ są średnimi arytmetycznymi pomiarów w grupach G_1 i G_2, a s jest oszacowaniem wspólnego odchylenia standardowego σ. Funkcja mocy tego testu zależy od parametru $\Delta = \dfrac{\mu_2 - \mu_1}{\sigma}$ oraz od wartości k i l. Hipoteza alternatywna H_1 jest określona przez warunek $\Delta > 0$, hipoteza zerowa H_0 – przez warunek $\Delta \leq 0$.

Rozpatrzmy dla przykładu rzeczywiste dane o długości bezawaryjnej pracy kompresorów w elektrowniach atomowych USA. Są one konserwowane na dwa sposoby (wyniki w grupach, odpowiednio, G_1 i G_2):

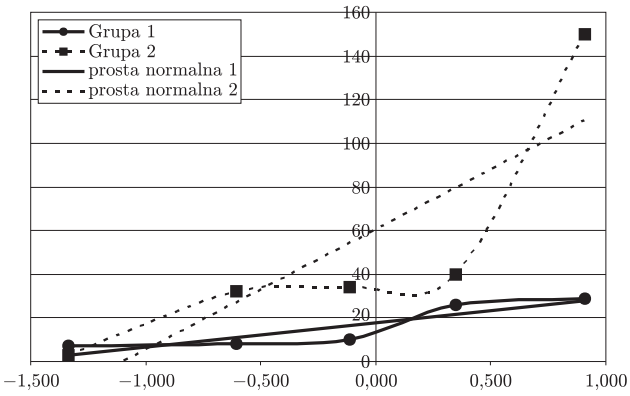
G_1:	7	8	10	26	29
G_2:	3	32	34	40	150

Interesuje nas odpowiedź, czy czas bezawaryjnej pracy kompresorów w grupie G_2 jest większy niż czas w grupie G_1. Na oko widać, że ten czas jest większy w grupie G_2.



Rys. 2

Gdy przyjmiemy założenia testu Studenta, to funkcja mocy tego testu z wartością krytyczną 2 ma wykres, z którego można odczytać, że poziom istotności jest około 0,05 (dokładnie 0,04). Gdy średnie różnią się o jedno odchylenie standardowe ($\Delta = 1$), test wskaże na hipotezę H_1 z prawdopodobieństwem 0,38, a gdy średnie różnią się o dwa odchylenia standardowe ($\Delta = 2$) – z prawdopodobieństwem 0,86.



Rys. 3

Dla naszych danych wartość statystyki Studenta wynosi $t = 1{,}387$, co jest poniżej wartości krytycznej 2. Test wskazuje na hipotezę H_0 a więc, że czas bezawaryjnej pracy kompresorów w grupie G_2 nie przekracza czasu bezawaryjnej pracy w grupie G_1. Jest to sprzeczne z intuicją.

Niezgodność z intuicją w tym przypadku daje się łatwo wytłumaczyć. Przyjęliśmy założenia o normalności rozkładu, które w tym przypadku nie są spełnione, szczególnie dla danych z grupy G_2. Można się o tym przekonać, na wykresie prawdopodobieństwa normalnego, w którym dane mające rozkład normalny leżą na prostej normalnej.

Pomysł Wilcoxona

Z takimi problemami często spotykał się na początku lat 40-tych chemik Frank Wilcoxon[vi]. W czasie swoich badań nad skutecznością środków ochrony roślin, nad którymi wtedy pracował, często korzystał zarówno z testu Studenta, jak i metod analizy wariancji Fishera. I nierzadko zdarzało się, że rozwiązania proponowane przez klasyczne metody statystyczne dawały wyniki przeczące intuicji. Zdawał sobie sprawę, że na ogół takie dane nie spełniają założeń testu Studenta. Nie znalazł jednak w literaturze żadnych wskazówek, jak postąpić.

[vi] Frank Wilcoxon (1892–1965)

Zupełnie desperacko, tak jak pewnie nie zrobiłby tego żaden matematyk, zamiast obserwowanych wartości wstawiał ich rangi. Rangi obserwacji są to numery wszystkich obserwacji uporządkowanych rosnąco. Oto rangi dla naszych danych

G_1:	7	8	10	26	29
rangi	**2**	**3**	**4**	**5**	**6**
G_2:	3	32	34	40	150
rangi	**1**	**7**	**8**	**9**	**10**

Wartość statystyki Studenta dla rang wynosi $t = 1{,}732$ – jeszcze daleko od wartości krytycznej 2, ale zdecydowanie bliżej do przyjęcia hipotezy H_1 niż w przypadku poprzednim. Również rangi bardziej przypominają rozkład normalny niż bezpośrednie obserwacje.

Okazuje się, że statystyka Studenta jest rosnącą funkcją sumy rang w grupie G_2, rozkład zaś sumy rang nie zależy od rozkładu oryginalnych obserwacji, o ile tylko mają gęstość. Co więcej, z centralnego twierdzenia granicznego wynika, że suma rang po standaryzacji ma rozkład normalny. Objaśnia to niezwykły fenomen, że w wielu sytuacjach, w których nie mógł być stosowany test Studenta, „sztuczka" Wilcoxona z zamianą obserwacji na rangi się sprawdzała: wyniki były zgodne z intuicją, a rozkład rang w grupach był zbliżony do normalnego. Test Wilcoxona oparty na sumie rang w grupie G_2 nosi nazwę *testu sumy rang*.

Wilcoxon nie był pewien, ile warte są jego pomysły. Postanowił wysłać artykuł do dobrego czasopisma *Biometrics*, zawierający wyniki teoretyczne, które uzyskał. Liczył na to, że albo recenzenci znajdą błąd w jego rozumowaniu, albo wskażą literaturę, gdzie takie wyniki już się pokazały. Ku jego zdumieniu artykuł został przyjęty do druku w 1945 roku. Test sumy rang stał się niezwykle popularny wśród praktyków dzięki temu, że w roku 1947 umieszczono je w poradniku *Some Rapid Approximate Statistical Procedures*. Osiągnięcia Wilcoxona zapoczątkowały też nowy dział badań – *statystykę nieparametryczną*.

Użycie rang, zamiast bezpośrednich obserwacji stanowiło jednak krok wstecz w rozwoju nauk eksperymentalnych.

Skale

Wyniki obserwacji zapisuje się w różny sposób, zależny od zawartości informacji, które one wnoszą. Sposób zapisu informacji nazywamy *skalą*.

Najprostsze obserwacje pozwalają odróżnić jeden obiekt od drugiego. Takie informacje zapisuje się nadając nazwy obiektom. Po to, by zapisać, jakie kwiaty rosną na łące wystarczy wymienić ich nazwy. Każde odkrycie nowego obiektu w astronomii czy geografii wiąże się z nadaniem mu indywidualnej nazwy. Taka skala nazywa się *nominalną*. Płeć, imię, kolor oczu wyrażone są w skali nominalnej. Wszystkie nauki oparte na obserwacjach przechodzą etap nazywania obiektów. Skala nominalna dzieli obserwacje na grupy obserwacji podobnych do siebie według wybranego kryterium. Obiekt zapisany w skali nominalnej uważany jest za najbardziej typowy, jeśli należy do grupy o największej liczności. Taka grupa nazywa się *modą* lub *dominantą*.

Niektóre nazwy nadaje się, aby uporządkować obserwacje. Jak w znanej grze dziecięcej, użycie nazwy *zimno*, *ciepło*, *gorąco*, *parzy* pozwala ocenić, jak daleko jesteśmy od szukanego obiektu. Nazwy medali sportowych *złoty*, *srebrny*, *brązowy* pozwalają odtworzyć porządek na mecie. Inny sposób uporządkowania to zanotowanie kolejności na mecie, czyli rangi. Ranga pozwala odtworzyć porządek, ale nic więcej. Ktoś, kto zajął czwarte miejsce na mecie, czyli ma rangę 4, wcale nie przyszedł 4 razy później niż zwycięzca, mający rangę 1. Skala pozwalająca uporządkować obiekty według jakiegoś kryterium nazywa się *porządkową*. Obiekt opisany w skali porządkowej jest typowy, jeśli tyle samo obiektów jest na skali przed nim, a tyle samo po nim. Taka wielkość nazywa się *medianą*. Czasami skalę porządkową koduje się za pomocą liczb, co prowadzi do nieporozumień. Liczenie średniej arytmetycznej jako wskaźnika typowości dla danych w skali porządkowej jest błędem[vii].

Prawdziwy postęp w nauce dokonał się, kiedy wprowadzono jednostki miar, które pozwalają opisać obiekty za pomocą liczb i odpowiedzieć na pytanie, ile razy coś jest większe, lub o ile jest większe. Taka skala nazywa się *liczbowa*. Znane jest powiedzenie: obserwacje są nominalne lub porządkowe, pomiary i teoria w skali liczbowej. Wartością typową w skali liczbowej jest średnia (arytmetyczna, harmoniczna geometryczna itp.). Historia prób znalezienia opisu jakiegoś zjawiska za pomocą liczb bywa fascynująca[viii].

Pomysł Wilcoxona pozornie cofnął analizę statystyczną z bogatej skali liczbowej, w której zanotowano obserwacje, do ubogiej skali porządkowej. Zadawano sobie pytania, o ile gorszy od testu Studenta jest test sumy rang, gdy mamy do czynienia z rozkładem normalnym, i czy nie opłaca się opracować innych testów dopasowanych do innych niż normalny rozkładów.

Efektywność Pitmana

Odpowiedź nadeszła z dalekiego kraju i również od statystycznego outsidera. Wilcoxon nie był matematykiem – był samoukiem statystycznym, który swoją wiedzę oparł na wspomnianej książce Fishera *Statistical Methods for Research Workers*[ix]. Człowiek, który postawił kropkę nad i, Edwin Pitman[x], był wprawdzie matematykiem, ale statystykę wykładał na uniwersytecie w Hobart na Tasmanii trochę z przymusu[xi]. Jak sam napisał, był matematykiem , który zabłąkał się w statystyce. Edukację statystyczną Pitman również oparł na książce Fishera.

Swoją pracę Pitman rozpoczyna od porównania funkcji mocy testów Studenta i Wilcoxona, gdy dane pochodzą z rozkładu normalnego (rys. 4).

Jak widać, funkcja mocy dla testu Wilcoxona jest niewiele słabsza niż dla testu Studenta. Ale czym opisać to podobieństwo?

Pitman wprowadza pojęcie efektywności względnej testu Wilcoxona względem testu Studenta. Jak wiadomo, funkcja mocy zależy od $\Delta = \frac{\mu_2 - \mu_1}{\sigma}$ i liczby obserwacji w grupach k i l. Jeżeli założymy, że $k = l$ i $2k = N$ jest łączną liczbą obserwacji, to można porównać oba testy przy tej samej wartości Δ i różnych N.

Dla każdej wartości mocy można wybrać parę liczb NT i NW taką, że przy NT obserwacjach test Studenta ma taką sama moc jak test Wilcoxona przy NW obserwacjach. Dla mocy 0,9 i $\Delta = 0{,}5$ przy założeniu, że dane mają rozkład

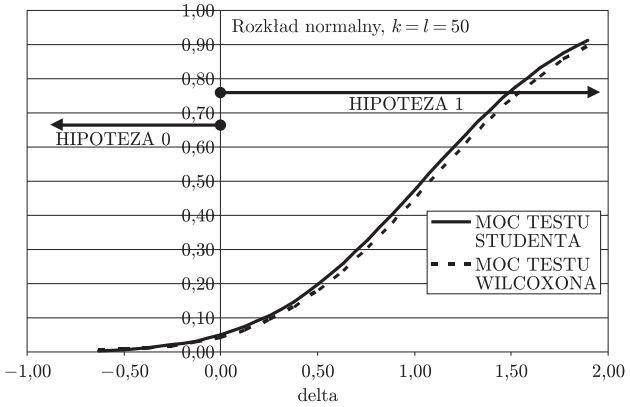
[vii] Na przykład w mineralogii skala twardości Mohsa jest porządkowa i umożliwia rozpoznanie, którym z dwóch porównywanych minerałów można zrobić rysę na drugim. Wybrano 10 wzorcowych minerałów ustalających bazowe wartości skali: diament ma wartość 10, korund 9. Najmniejszą wartość spośród wzorcowych minerałów – 1 – ma talk. Diament ma wyższą wartość niż korund i talk, bo można nim zrobić rysę na obu minerałach. Dla substancji nie wymienionych na liście przyjmuje się wartość pośrednią: na przykład paznokieć ma twardość 2,5 bo rysuje gips, który ma twardość 2 i nie rysuje na kalcycie, który ma twardość 3. Diament jednak nie jest 4 razy twardszy od paznokcia, mimo że ma twardość 10, paznokieć 2,5. Ocena szkolna jest w skali porządkowej i uczeń, który otrzymał 4, nie jest dwa razy lepszy od ucznia, który otrzymał 2. W USA, gdzie oceny są od A do F nie ma miejsca na takie pomyłki.

[viii] Na przykład historia wynalezienia termometru i związanej z nią skali temperatur.

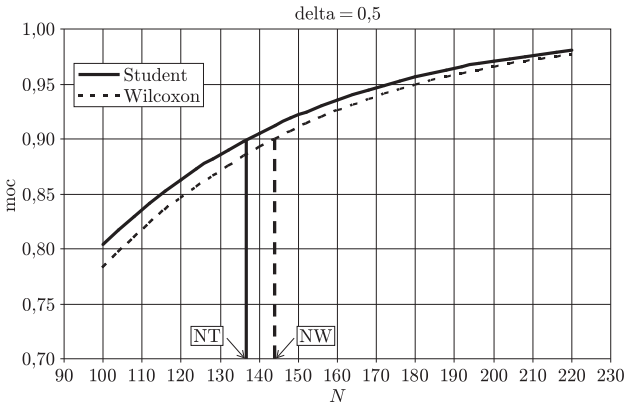
[ix] Można ją przeczytać pod adresem http://psychclassics.yorku.ca/Fisher/Methods/ Tekst jest opatrzony reprodukcjami wzorów i tabel z oryginalnego wydania z 1925 roku.

[x] Edwin Pitman (1897–1993)

[xi] Pitman, urodzony w Melbourne brał udział w pierwszej wojnie światowej w Europie. Po jej zakończeniu skończył studia matematyczne w Melbourne. Pierwszą pracę podjął jako matematyk na uniwersytecie w Nowej Zelandii. Poważna propozycja pracy przyszła z Hobart. Uniwersytet był bardzo biedny i zatrudniał niewielu pracowników. Pitman wykładał całą matematykę i fizykę. Warunkiem przyjęcia do pracy było przygotowanie wykładu z „nowej dziedziny nauki – statystyki".



Rys. 4



Rys. 5

normalny, te liczby wynoszą $NT = 137$ i $NW = 144$. Test Wilcoxona musi nadrobić spadek mocy poprzez zwiększenie liczby obserwacji. Spadek ten można scharakteryzować stosunkiem NT/NW, zwanym efektywnością względną. W tym przypadku efektywność względna wynosi 0,95 (rys. 5).

Ten sam rachunek można wykonać dla dowolnego rozkładu prawdopodobieństwa. Efektywność względna zależy od NT, poziomu istotności, wartości Δ i mocy testu. W pracy, którą Pitman wysłał w roku 1948 do jednego z najlepszych na świecie czasopism statystycznych, *Annals of Mathematical Statistics*, pokazał, że gdy $\Delta \approx \dfrac{1}{\sqrt{NT}}$ i $NT \to \infty$ i to NT/NW ma granicę równą $\frac{e}{\pi} \approx 0{,}955$, niezależnie od poziomu istotności, mocy i granicy $\sqrt{NT}\Delta$.

Jak widać, przy zamianie testu Studenta na test Wilcoxona tracimy na efektywności niewiele, bo około 5%, gdy dane mają rozkład normalny. Ale jak to jest dla innych rozkładów?

Dla dowolnej pary rozkładów o dystrybuantach F i G takich, że $G(t) = F(t - \Delta)$ można pokazać, że gdy $\Delta \approx \dfrac{1}{\sqrt{NT}}$, $NT \to \infty$ i dystrybuanta F ma gęstość f oraz wariancję σ^2 to NT/NW ma granicę

$$e(F) = 12\sigma^2\left(\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right)^2.$$

Okazuje się, że dla każdej dystrybuanty F mamy $e(F) \geq 0{,}864$, czyli że strata przy zamianie testu Studenta na test Wilcoxona nigdy nie przekracza 14%. Ten wynik był zaskoczeniem dla całego środowiska statystycznego. Ale to nie koniec sensacji.

Dla rodziny rozkładów gamma sparametryzowanych parametrem p efektywność Pitmana można jawnie wyliczyć:

$$e(p) = \frac{3p\Gamma^2(2p)}{2^{4(p-1)}(2p-1)^2\Gamma(p)}.$$

Tablica efektywności Pitmana dla niektórych wartości p:

p	0,5	1	2	3	4	5	10	∞
$e(p)$	∞	3	1,5	1,27	1,17	1,12	1,03	0,955

Jak widać, test Wilcoxona dla asymetrycznych rozkładów gamma jest wielokrotnie bardziej efektywny niż test Studenta. Dla rozkładu χ^2, gdzie $p = 0{,}5$ efektywność jest nieskończona, dla rozkładu wykładniczego ($p = 1$) jest trzykrotnie wyższa. Dla rozkładu normalnego ($p = \infty$) efektywność jest mniejsza od 1, ale test Studenta był przystosowany specjalnie do tego rozkładu.

A więc, przynajmniej w statystyce, na ogół prostota popłaca, a czasem nawet nieskończenie wiele razy.