

TERESA LEDWINA (Wrocław)

## O pewnych testach statystycznych i ich porównywaniu

Probabilistyka konstruuje i bada modele zjawisk losowych. Statystyka stara się odpowiedzieć na pytanie, jaki model probabilistyczny i na ile dobrze opisuje obserwowane zjawisko.

Jednym z kilku głównych działów statystyki jest teoria testów. Zanim wprowadzimy formalnie rozważane przez nas zagadnienie testowania i pojęcie testu, przedstawimy nieco informacji na temat najwcześniejszych prac próbujących rozstrzygnąć czy dany model „pasuje” do danych. Rozwiązaniem starszym i mniej popularnym w środowisku matematyków oraz ich autorom poświęcimy nieco więcej uwagi, niż tym dobrze znanym choćby z prac probabilistycznych.

### 1. Konstrukcje testów

**1.1. *Od J. Arbuthnotta do K. Pearsona.*** Za pierwszą pracę z omawianej dziedziny uważa się notkę J. Arbuthnotta [1]. Reprint tej pracy znaleźć można w książce Kendalla i Placketta [34] na str. 30–34. Ciekawe komentarze o tej publikacji zawarto w powyższej książce na str. 35–37 oraz w książce S. Stiglera [64] na str. 225–226. Nadmienimy również, że ta krótka notka J. Arbuthnotta jest uznawana za jeden z milowych kamieni w rozwoju statystyki (por. Kotz i Johnson [36] str. viii–ix). Arbuthnott opublikował w niej liczby noworodków płci męskiej i żeńskiej ochrzczonych w Londynie w latach 1629–1710. Dane te pokazują, że we wszystkich osiemdziesięciu dwóch latach ujętych w jego wykazie, liczby noworodków chłopców są większe od liczb dziewczynek. Przyjmując założenie, że zdarzenie losowe ma szansę wynoszącą  $1/2$ , autor wywnioskował stąd, że szansa na to, że w kolejnych 82 latach urodzi się więcej chłopców niż dziewczynek wynosi  $(1/2)^{82}$ . Nie negując stochastycznego charakteru samego procesu pojawiania się typu płci, autor konkluduje, że tak małe prawdopodobieństwo zaobserwowania uzyskanego wyniku wyklucza losowość (w powyższym rozumieniu). Jednocześnie autor stwierdza, że chłopcy wiodą bardziej ryzykowne życie i nieco większa liczba męskich noworodków gwarantuje równą proporcję płci w wieku

dojrzałym. To z kolei jest warunkiem monogamii, zgodnej z prawem naturalnym. Ostateczny wniosek jest taki, że wykazany brak losowości wskazuje na istnienie Boskiej Opatrzności. Niejako na usprawiedliwienie autora dodajmy, że poglądy na to, że rozmaite „regularności” zdarzeń losowych są przejawem boskiego porządku, nie były w osiemnastym wieku rzadkością. Nadmienmy również, że Arbuthnott był lekarzem i znanym satyrykiem. Dla zarobku przetłumaczył na język angielski i nawet nieco uzupełnił probabilistyczny traktat Huyghensa [15]. To tłumaczenie było pierwszą pracą o prawdopodobieństwie opublikowaną w języku angielskim.

W następnych dwóch stuleciach pojawiło się zaledwie kilka dalszych prób orzekania o zgodności modelu probabilistycznego z obserwacjami. E. Lehmann [43] str. 126, wymienia w tym kontekście prace D. Bernoulliego [4], P. Laplace’a [39], J. Gavarreta [14], W. Lexisa [44], [45] i F. Edgewortha [12].

Następną przełomową pracą, która zapoczątkowała powszechne stosowanie testów była praca K. Pearsona z 1900 r. (patrz [58] i [37]). K. Pearson miał wszechstronne wykształcenie. Oprócz matematyki studiował prawo, fizykę, biologię, filozofię i średniowieczny język niemiecki. Od 1884 r. K. Pearson pracował jako profesor matematyki stosowanej i mechaniki w University College w Londynie. Początkowo jego zainteresowania i publikacje koncentrowały się na filozofii, religii i sztuce. Spotkanie w 1890 r. zoologa W. Weldona i dyskusje z nim zaowocowały założeniem w 1901 r. znakomitego czasopisma *Biometrika* i pracami K. Pearsona drukowanymi w latach 1893–1905 między innymi w rozprawach o matematycznej teorii ewolucji. W szczególności, około roku 1900 K. Pearson zajął się problemem weryfikowania zgodności rozkładu obserwacji z zadaniem rozkładem. Wspomniana wyżej praca zawiera właśnie rozwiązanie tego problemu. Przedstawimy je w zastosowaniu do jednego konkretnego modelu, który będzie nam również służył jako ilustracja w dalszej części tego opracowania. Opis poprzedzamy wprowadzeniem niezbędnych oznaczeń.

*1.1.1. Podstawowe oznaczenia.* Przypuśćmy, że w wyniku pomiarów pewnej cechy w populacji otrzymaliśmy  $n$  liczb  $y_1, \dots, y_n$ . Będziemy się ograniczać do sytuacji, gdy pomiary są dokonywane w sposób gwarantujący, że  $y_1, \dots, y_n$  mogą być interpretowane jako wartości niezależnych zmiennych losowych  $Y_1, \dots, Y_n$  o jednakowym rozkładzie. Oznaczmy przez  $G$  dystrybuantę zmiennej losowej  $Y_i$ ,  $i = 1, \dots, n$ . Jeśli  $Y_i$  jest określona na przestrzeni probabilistycznej  $(\Omega, \mathcal{B}, P)$ , to  $G(y) = P(Y_i \leq y)$ ,  $y \in R$ . Skrótowo będziemy pisać  $Y_i \sim G$  i oznaczać przez  $P_G$  łączny rozkład  $Y_1, \dots, Y_n$ .

Skupimy uwagę wyłącznie na przypadku, gdy  $G$  jest funkcją ciągłą. Wypisujemy przypuszczenie, zwane hipotezą testowaną lub zerową, mówiące, że  $G = G_0$ , gdzie  $G_0$  jest daną ciągłą dystrybuantą, np. dystrybuantą rozkładu gaussowskiego  $N(0, 1)$ . Prostą i naturalną ilustracją problemu jest weryfikacja działania generatora liczb losowych.

Dla dowolnej ciągłej dystrybuanty  $G_0$  zachodzi  $Y \sim G_0 \Leftrightarrow X = G_0(Y) \sim F_0$ , gdzie  $F_0$  jest dystrybuantą rozkładu jednostajnego na odcinku  $(0,1)$ , to znaczy

$$F_0(x) = \begin{cases} 0 & \text{dla } x \leq 0, \\ x & \text{dla } 0 < x \leq 1, \\ 1 & \text{dla } x > 1. \end{cases}$$

Tak więc problem weryfikacji hipotezy  $G = G_0$  w oparciu o obserwacje zmiennych  $Y_1, \dots, Y_n$  możemy zastąpić problemem weryfikacji hipotezy  $F = F_0$  w oparciu o wartości zmiennych  $X_1 = G_0(Y_1), \dots, X_n = G_0(Y_n)$ . Reasumując, całe poniższe opracowanie dotyczyć będzie problemu weryfikacji hipotezy zerowej

$$H_0 : F = F_0$$

w oparciu o niezależne zmienne losowe  $X_1, \dots, X_n$  o dystrybuancie  $F$  z rozkładu skoncentrowanego na  $(0,1)$ . Przez  $P_F$  będziemy oznaczać łączny rozkład tych zmiennych.

Na zakończenie wprowadzimy jeszcze dodatkową dystrybuantę

$$F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n},$$

gdzie  $\#B$  oznacza licznosc zbioru  $B$ .  $F_n(x)$  jest nazywana dystrybuantą empiryczną zmiennych  $X_1, \dots, X_n$ . Z dystrybuantą empiryczną w sposób naturalny wiąże się proces empiryczny  $\{U_n(x), x \in [0, 1]\}$  gdzie

$$U_n(x) = \sqrt{n}[F_n(x) - x],$$

będący standaryzowaną różnicą między  $F_n(x)$  i  $F_0(x)$ .

*1.1.2. Test chi-kwadrat K. Pearsona.* W zastosowaniu do weryfikacji  $H_0 : F = F_0$ , w oparciu o  $X_1, \dots, X_n$ , rozumowanie i rozwiązanie K. Pearsona wyglądało następująco. Wybieramy liczbę naturalną  $k$ ,  $k \geq 2$ , oraz liczby  $0 < a_1 < \dots < a_{k-1} < 1$ . Dzielimy  $(0,1)$  na klasy  $(a_0, a_1], (a_1, a_2], \dots, (a_{k-1}, a_k)$ , gdzie  $a_0 = 0$ ,  $a_k = 1$ . Zliczamy licznosci  $N_j$  wartości zmiennych  $X_1, \dots, X_n$ , które wpadły do  $j$ -tej klasy,  $j = 1, \dots, k$ . Zauważamy, że przy prawdziwości  $H_0$  wektor  $(N_1, \dots, N_k)$  ma rozkład wielomianowy z parametrami  $n$  i  $p_0 = (p_{01}, \dots, p_{0k})$ ,  $p_{0j} = a_j - a_{j-1}$ ,  $j = 1, \dots, k$ . Dla zmierzenia stopnia zgodności rozkładu  $X_i$  z  $F_0$  K. Pearson zaproponował porównanie obserwowanych licznosci  $N_j$  z ich oczekiwanymi wartościami  $np_{0j}$ , liczonymi przy założeniu, że  $F = F_0$ . W tym celu wprowadził statystykę (kryterium)

$$\chi^2 = \sum_{j=1}^k \frac{(N_j - np_{0j})^2}{np_{0j}}.$$

Duże wartości tego kryterium świadczą o słabym dopasowaniu rozkładu  $F_0$ . K. Pearson udowodnił, że przy prawdziwości  $H_0$  i  $n \rightarrow \infty$  zachodzi  $\chi^2 \xrightarrow{D} \chi_{(k-1)}^2$ , gdzie  $\chi_{(s)}^2$  oznacza zmienną losową o rozkładzie chi-kwadrat

z  $s$  stopniami swobody. Wobec powyższego, przynajmniej dla w miarę licznych prób, mając dane i obliczoną dla nich wartość  $\chi_{obl}^2$  statystyki  $\chi^2$ , można z tablic rozkładu  $\chi_{(k-1)}^2$  odczytać wartość  $P_{F_0}(\chi^2 \geq \chi_{obl}^2)$ . Ta liczba daje pogląd na ile „typowa” jest zaobserwowana wartość  $\chi_{obl}^2$ , przy prawdziwości  $H_0$ . Jeśli to prawdopodobieństwo jest bardzo małe, to prawdziwość  $H_0$  jest mocno wątpliwa.

Jak widać z tego opisu, rozwiązanie K. Pearsona wymagało wielu subiektywnych decyzji. Nie jest jasne jak, wybierać  $k$  i liczby  $a_j$ ,  $j = 1, \dots, k - 1$ , a sprawa decyzji o akceptacji  $H_0$  pozostała kwestią odczuć statystyka. O pewnych rozwiązaniach powyższych kwestii wspominamy w rozdziałach 1.3.1 i 1.4. Tym niemniej, w porównaniu do wcześniejszych dywagacji, praca K. Pearsona stanowiła ogromny postęp.

J. Haldane napisał o tym rozwiązaniu: „*This has turned out to be an immensely powerful tool, and is used on a huge scale*” (por. [57], str. 432). G. Barnard (patrz [37], str. 1) przypomina, że czasopismo *Science* zaliczyło w 1984 r. tę pracę K. Pearsona do dwudziestu najważniejszych osiągnięć wieku dwudziestego, w szerokim spektrum nauk ścisłych.

W następnym rozdziale omawiamy pokrótce dwa podobne rozwiązania, zaproponowane znacznie później przez znakomitych statystyków i probabilistów. Dla porównania statystyki K. Pearsona z ich propozycjami wyrazimy  $\chi^2$  jako funkcję procesu empirycznego  $\{U_n(x), x \in [0, 1]\}$ . Bezpośrednio z definicji mamy

$$\chi^2 = \sum_{j=1}^k \frac{[U_n(a_j) - U_n(a_{j-1})]^2}{a_j - a_{j-1}}.$$

Tak więc, dla danego  $k$  i ustalonych  $a_0, \dots, a_k$ ,  $(\chi^2)^{\frac{1}{2}}$  jest seminormą na przestrzeni  $D[0, 1]$ , liczoną w punkcie  $U_n$ .

**1.2. Statystyki Craméra-von Misesa i Kolmogorowa-Smirnowa.** H. Cramér [7] (przedruk w [8]) i R. von Mises [46] rozważali między innymi statystyki typu

$$\int_0^1 [U_n(x)]^2 K(x) dx,$$

gdzie  $K$  jest pewną nieujemną funkcją wagową. W rzeczywistości obaj formułowali problem w terminach wyjściowych zmiennych losowych  $Y_1, \dots, Y_n$  i dystrybuanty  $G_0$  (por. rozdział 1.1.1.). W takim ujęciu funkcje wagowe warto uzależnić od  $G_0$ , co przytomnie zauważył N. Smirnow. N. Smirnow [62] wyznaczył również asymptotyczny rozkład

$$W = \left\{ \int_0^1 [U_n(x)]^2 dx \right\}^{\frac{1}{2}}.$$

Seminorma  $W$  jest we współczesnej literaturze nazywana statystyką Craméra-von Misesa. Warto zaznaczyć, że rozważania H. Craméra [7] były ciekawsze i głębsze, niż to co ostatecznie weszło do powszechnego obiegu. H. Cramér bazował na centralnym twierdzeniu granicznym i rozwinięciu Charliera granicznej dystrybuanty. W jego pracy jest analizowanych pięć dużych zbiorów danych oraz stopień dopasowania ich rozkładów empirycznych przez skorygowany (*via* uwzględnienie kolejnych członów rozwinięcia) rozkład normalny. E. Phragmén (por. [8], str. v) tak scharakteryzował H. Craméra: „*He belonged to a generation of mathematicians for which it was self evident that mathematics constitutes one of the highest forms of human thoughts, perhaps even the highest. ... This does not however mean that they underestimated the importance of ‘using theoretical knowledge to obtain know-how’.*” Należy żałować, że N. Smirnow spopularyzował uproszczony wariant rozwiązania H. Craméra.

W 1933 r. A. Kołmogorow napisał pracę [35], w której rozważył inną seminormę procesu empirycznego. Mianowicie, zdefiniował

$$D = \sup_{x \in [0,1]} |U_n(x)|$$

i znalazł rozkład graniczny zmiennej  $D$ . Angielskie tłumaczenie tej przełomowej, w powszechnym odczuciu, pracy można znaleźć w [37]. Należy podkreślić, że w przeciwieństwie do H. Craméra, A. Kołmogorow przynajmniej do czasu drugiej wojny światowej nie interesował się statystyką. We wspomnianej pracy A. Kołmogorow jasno wyklada swoje motywacje. Skoro zachowanie normy  $L_2$  procesu empirycznego (statystyka  $W$ ) badał R. von Mises [46], to on zajmie się zachowaniem normy supremum (statystyka  $D$ ). W pracy w ogóle nie pojawiają się słowa: test, statystyka itp. To N. Smirnow opracował rozmaite warianty  $D$ , wyznaczył ich rozkłady graniczne i wprowadził niejako tę seminormę do praktyki statystycznej. W uznaniu jego wkładu,  $D$  jest nazywana statystyką Kołmogorowa-Smirnowa. Oczywiście, tak jak w przypadku statystyki  $\chi^2$ , duże wartości  $W$  i  $D$  wskazują na istotne odstępstwa od rozkładu hipotetycznego, zadanego dystrybuantą  $F_0$ . Również, podobnie jak rozwiązanie K. Pearsona,  $W$  i  $D$  są dziś standardowymi narzędziami, omawianymi we wszystkich podręcznikach i zaimplementowanymi w pakietach statystycznych. Więcej informacji na temat tych statystyk można znaleźć w ciekawym artykule M. Stephensa (patrz [37], str. 93–113) oraz w monografiach J. Durbina [11] oraz R. D’Agostino i M. Stephensa [9].

### 1.3. Test Neymana.

1.3.1. *Uwagi i pojęcia wstępne.* Wkład J. Neymana w zbudowanie ogólnej teorii testowania hipotez jest fundamentalny. Wiele szczegółów na ten temat można znaleźć m.in. w omówieniach L. LeCama i E. Lehmana [40]

oraz T. Ledwiny [42]. W niniejszym opracowaniu ograniczymy się do przypomnienia paru mało znanych faktów i naświetlenia pewnych aspektów, które będą istotne dla zrozumienia rewolucyjności pracy J. Neymana [49], dotyczącej interesującego nas problemu testowania. Pracę tę krótko omówimy w rozdziale 1.3.2.

Jak widać z przedstawionego wyżej materiału, problem testowania został postawiony przez przyrodników i pierwsze prace nie wykraczały daleko poza dostarczenie jakiegoś narzędzia pozwalającego porównywać „obserwowane” z „oczekiwanym”. Szczęśliwym trafem J. Neyman był doskonale wykształconym matematykiem. Przypomnijmy, że jego nauczycielem i mistrzem na studiach w Charkowie był S. Bernstein. Praca dyplomowa J. Neymana, licząca 530 str., dotyczyła całki Lebesgue’a i została wyróżniona złotym medalem. Te zainteresowania J. Neyman pogłębiał później na rocznym stypendium w Paryżu u E. Borela i H. Lebesgue’a. Po przyjeździe do Polski w 1921 r., z życiowej konieczności, J. Neyman zajął się statystyką i w 1924 r. uzyskał z niej doktorat na Uniwersytecie Warszawskim. Już wtedy ujawnił się niezwykły talent J. Neymana. W skład doktoratu weszły jego prace z doświadczeń rolniczego, opublikowane po polsku między innymi w *Rocznikach Nauk Rolniczych*. Część tych wyników została w 1990 r. przetłumaczona na język angielski ([63]). W komentarzu do tego tłumaczenia D. Rubin ([60], str. 472) napisał: ... „*It is a honour to be asked to discuss this document, which reinforces Neyman’s place as one of our greatest statistical thinkers...*”. Po doktoracie, J. Neyman spędził rok akademicki 1924/25 u K. Pearsona. Nawiązanie współpracy z E. Pearsonem, synem Karola, zaowocowało stworzeniem podstaw nowoczesnej teorii testowania hipotez. W szczególności, ich przełomowa praca z 1933 r. (patrz [53] lub przedruk w [66]) wprowadziła formalizację zagadnienia testowania w języku pewnego problemu optymalizacyjnego. Spróbujemy krótko przedstawić ich punkt widzenia na przykładzie zagadnienia, którym się zajmujemy w niniejszym opracowaniu. Takie rozwiązanie ograniczy znacznie ogólność prezentacji, równocześnie pozwoli na uniknięcie wprowadzania szeregu dodatkowych oznaczeń i założeń.

Przypomnijmy, że statystyk dysponuje wartościami  $n$  niezależnych zmiennych losowych  $X_1, \dots, X_n$  o ciągłej dystrybuancie  $F$ , każda. Hipoteza testowana  $H_0$  orzeka, że  $F = F_0$ , gdzie  $F_0$  jest dystrybuantą rozkładu jednostajnego na  $(0, 1)$ . Zauważmy również, że już w pracach [51] i [52] (przedruki w [66]) J. Neyman z E. Pearsonem wprowadzili istotną innowację w stawianiu samego problemu testowania. Mianowicie, oprócz hipotezy testowanej wyróżnili również hipotezę alternatywną. Tak więc, jeśli przez  $\mathcal{F}$  oznaczymy klasę wszystkich ciągłych dystrybuant na  $(0, 1)$ , to problem weryfikacji zgodności z  $F_0$  wyglądałby w ich ujęciu na przykład tak: testujemy hipotezę zerową

$$H_0 : F = F_0$$

przeciwko alternatywie

$$A_0 : F \in \mathcal{F}_0,$$

gdzie  $\mathcal{F}_0$  jest pewną interesującą z praktycznego lub poznawczego punktu widzenia podklasą  $\mathcal{F}$ . Oczywiście, można wziąć w szczególności  $\mathcal{F}_0 = \mathcal{F}$ . Zajmijmy się tak postawionym problemem testowania. Dla formalności wywodu zanotujmy, że  $\mathbf{X} = (X_1, \dots, X_n)$  przyjmuje wartości w  $\mathcal{X} = (0, 1)^n$  i wprowadźmy przestrzeń probabilistyczną  $(\mathcal{X}, \mathcal{B}, P_F)$ , gdzie  $\mathcal{B}$  jest  $\sigma$ -ciałem zbiorów borelowskich na  $\mathcal{X}$ , a  $P_F$  oznacza miarę produktową na  $\mathcal{B}$  o brzegowych dystrybuantach  $F$ . Podkreślmy, że wszystkie rozważania w tym rozdziale są robione dla ustalonego  $n$ .

Testem statystycznym nazywać będziemy dowolną funkcję mierzalną  $\phi(\mathbf{x})$ ,  $\mathbf{x} \in \mathcal{X}$ , przyjmującą wartości w zbiorze  $\{0, 1\}$ . Dla wartości  $\mathbf{x}$  takich, że  $\phi(\mathbf{x}) = 0$  będziemy akceptować  $H_0$ , a dla  $\mathbf{x}$ , dla których  $\phi(\mathbf{x}) = 1$ , będziemy akceptować  $A_0$  (odrzucając tym samym  $H_0$ ). Kryteria wprowadzone przez K. Pearsona oraz H. Craméra i R. von Misesa prowadziły do akceptacji  $H_0$ , gdy wartości  $\chi_{obl}^2$  i  $W_{obl}$  były subiektywnie mało prawdopodobne. Aby wprowadzić pewne standardy, J. Neyman z E. Pearsonem zaproponowali, aby wybrać pewną małą liczbę  $\alpha$ , zwaną poziomem istotności, wyznaczyć liczbę  $c_\alpha$ , zwaną wartością krytyczną, akceptować  $H_0$ , gdy obliczona wartość rozważanego kryterium nie przekracza  $c_\alpha$  i odrzucać  $H_0$  (akceptować  $A_0$ ) w przypadku przeciwnym. Wartość  $c_\alpha$  ustala się tak, aby prawdopodobieństwo odrzucenia hipotezy testowanej, gdy jest prawdziwa, wynosiło  $\alpha$ . Ten postulat J. Neymana i E. Pearsona w naszym problemie można krótko zapisać w postaci warunku

$$(1) \quad \int_{\mathcal{X}} \phi(\mathbf{x}) dP_{F_0}(\mathbf{x}) = E_{P_{F_0}} \phi(\mathbf{X}) = \alpha.$$

Dla ilustracji, zamiast liczyć  $P_{F_0}(\chi^2 \geq \chi_{obl}^2)$ , jak proponował K. Pearson, i decydować czy uznać je za małe czy też nie, porównujemy  $\chi_{obl}^2$  z wartością  $c_\alpha$  spełniającą  $P(\chi_{(k-1)}^2 \geq c_\alpha) = \alpha$ . Jeśli  $\chi_{obl}^2 \geq c_\alpha$ , to  $H_0$  odrzucamy na poziomie istotności  $\alpha$ . Typowe poziomy istotności to 0.05 i 0.01. Takie ujęcie ułatwia rozumienie konkluzji wyciąganych przez rozmaitych statystyków dla tych samych zbiorów obserwacji.

Oznaczmy teraz przez  $\mathcal{T}_\alpha$  klasę wszystkich mierzalnych  $\phi : \mathcal{X} \rightarrow \{0, 1\}$ , które spełniają (1).

Nawiązując do ilustracji z rozdziału 1.1.1., widzimy, że test, służący do sprawdzania czy nowy generator rzeczywiście dostarcza wyników z rozkładu jednostajnego, jest na poziomie istotności  $\alpha$  lub, równoważnie, należy do  $\mathcal{T}_\alpha$ , jeśli szansa zdyskwalifikowania przez ten test dobrego generatora wynosi  $\alpha$ .

Drugi postulat J. Neymana i E. Pearsona zawierał istotę ich innowacji. W języku rozważanego tu problemu testowania można go sformułować następująco. W klasie  $\mathcal{T}_\alpha$  szukamy funkcji  $\phi_0$  maksymalizującej  $E_{P_F} \phi(\mathbf{X})$  dla

$F \in \mathcal{F}_0$ . Tak więc  $\phi_0$  ma spełniać

$$(2) \quad \arg \max_{\phi \in \mathcal{T}_\alpha} E_{P_F} \phi(\mathbf{X}) = E_{P_F} \phi_0(\mathbf{X}), \quad \forall F \in \mathcal{F}_0.$$

Ponieważ  $\phi(\mathbf{x}) = 1$  prowadzi do akceptacji  $A_0$ , warunek (2) oznacza, że w klasie  $\mathcal{T}_\alpha$  szukamy funkcji maksymalizującej, jednostajnie po  $\mathcal{F}_0$ , prawdopodobieństwo stwierdzenia, że  $F \in \mathcal{F}_0$ , w sytuacji, gdy  $F$  rzeczywiście należy do  $\mathcal{F}_0$ .

W rozważanym przez nas przykładzie, drugi postulat J. Neymana i E. Pearsona oznacza, że szukamy testu, który rzadko (z prawdopodobieństwem  $\alpha$ ) będzie dyskwalifikował dobry generator i jednocześnie maksymalnie często odrzucał wadliwe generatory.

$E_{P_F} \phi(\mathbf{x})$ , traktowana jako funkcja  $F \in \mathcal{F}_0$ , nazywana jest funkcją mocy testu  $\phi$ . Krótko mówiąc, J. Neyman z E. Pearsonem postawili formalny problem wyznaczenia testu jednostajnie najmocniejszego w klasie testów na poziomie istotności  $\alpha$ .

Jest chyba oczywiste, że możliwości rozwiązania tak postawionego problemu optymalizacyjnego są ograniczone. We wspomnianej pracy [53] podano np. rozwiązanie, gdy  $\mathcal{F}_0 = \{F_1\}$ , gdzie  $F_1$  jest ustaloną dystrybucją. Rozważono też pewną parametryczną rodzinę rozkładów  $\{F_\theta, \theta \in \Theta \subset R^q\}$ , pewne problemy testowania o jej parametrach i wyznaczono testy jednostajnie najmocniejsze. Dalsze prace J. Neymana i E. Pearsona ([54], [55]; przedruk w [66]) oraz innych statystyków doprowadziły do kompletnego i eleganckiego rozwiązania tak postawionego problemu i problemów pochodnych, wynikających z pewnych dalszych ale naturalnych ograniczeń na klasę  $\mathcal{T}_\alpha$ . Jeszcze raz podkreślmy, że wszystkie te wyniki uzyskano dla dowolnej, ale ustalonej wielkości próby  $n$ . Współczesne ujęcie tych wyników można znaleźć w monografii E. Lehmana [43]. Na zakończenie tej krótkiej prezentacji zacytujmy opinię L. Le Cama i E. Lehmana [40] na temat znaczenia omawianych wyżej prac J. Neymana i E. Pearsona. Chyba trudno trafniej i krócej podsumować ich zasługi. *„The impact of this work has been enormous. It is, for example, hard to imagine hypothesis testing today without the concept of power, which provides the basis both for the determination of sample size and for any comparisons among competing tests. And the optimum properties of the classical normal theory tests are not only aesthetically pleasing but serve as benchmarks against which the performance of simpler or more robust tests can be gauged. However, the influence of the work goes far beyond its implications for hypothesis testing. By deriving tests as the solutions of clearly defined optimum problems, Neyman and Pearson established a pattern for Wald’s general decision theory and for the whole field of mathematical statistics as it has developed since then.”*

**1.3.2. Gładki test Neymana.** Jak widać z powyższego, już około 1936 roku J. Neyman zdawał sobie sprawę z tego, że lista problemów, w których,



przy ustalonej wielkości próby  $n$ , istnieją testy jednostajnie najmocniejsze w klasie  $\mathcal{T}_\alpha$  lub w pewnych interesujących podklasach tej klasy, została wyczerpana. Dalszy postęp w dziedzinie definicji i konstrukcji testów optymalnych wymagał kolejnej istotnej innowacji. I J. Neyman takiej innowacji dokonał już w 1937 r. (por. [49] lub przedruk w [65]). Ta jego praca została określona przez L. Le Cama i E. Lehmana [40] jako genialna. Jeśli się wie, jak ogromne wymagania stawiał sobie i innym Le Cam, waga tych słów zasługuje na szczególną uwagę. We wspomnianej pracy [49], dedykowanej pamięci zmarłego w 1936 r. Karola Pearsona, J. Neyman rozważył problem testowania jednostajności rozkładu. Używając oznaczeń z poprzedniego rozdziału, problem wyglądał następująco.

Testujemy

$$H_0 : F = F_0$$

przeciwko

$$A : F \neq F_0, \quad F \in \mathcal{F},$$

na bazie obserwacji niezależnych zmiennych  $X_1, \dots, X_n$ . W przeciwieństwie do swych słynnych poprzedników, J. Neyman wyprowadził postać statystyki testowej na podstawie rozważań optymalizacyjnych.

Pierwszy krok, który zrobił J. Neyman, polegał na wyróżnieniu podklasy alternatyw. J. Neymana interesowała konstrukcja testu, który byłby czuły na ‘gładkie’ odstępstwa od hipotetycznego modelu jednostajnego. ‘Gładkie’ odstępstwa obejmowały: zmianę średniej, wariancji, skośności, stopnia spłaszczenia rozkładu itp. W celu sformalizowania tego postulatu, J. Neyman ograniczył po pierwsze rozważania do dystrybuant  $F$  posiadających gęstość  $f$  względem miary Lebesgue’a na  $(0,1)$ . Wówczas równoważną postacią  $H_0$  jest

$$H_0 : f = f_0 \equiv 1.$$

Do modelowania wspomnianych ‘gładkich’ odstępstw od  $f_0$  J. Neyman zaproponował rodzinę gęstości  $\mathcal{F}_k(\theta)$  złożoną z elementów postaci

$$f_k(x; \theta) = c_k(\theta) \exp \left\{ \sum_{j=1}^k \theta_j L_j(x) \right\},$$

gdzie  $k$  jest ustaloną liczbą naturalną,  $\theta = (\theta_1, \dots, \theta_k) \in R^k$  jest wektorem nieznanych parametrów,  $\{L_j\}_{j \geq 1}$  jest układem ortonormalnych wielomianów Legendre’a na  $(0, 1)$  a  $c_k(\theta)$  stałą normującą. W ten sposób wyjściowy problem testowania został zawężony do weryfikacji

$$H_0^* : \theta = 0$$

przeciwko

$$A_0^* : \theta \neq 0$$

w rodzinie gęstości  $\mathcal{F}_k(\theta)$ .

Drugi krok konstrukcji polegał na wyznaczeniu testu optymalnego dla tego zawężonego problemu testowania. Niebagatelną sprawą było samo określenie optymalności testu. Rozważając tę kwestię, J. Neyman podał rozwiązanie, które po trzech dekadach od momentu wprowadzenia weszło do powszechnego użycia i na długie lata było podstawowym narzędziem statystyki asymptotycznej. Mianowicie, J. Neyman zaproponował badanie i porównywanie funkcji mocy testów przy  $n \rightarrow \infty$  i jednoczesnym „ściągnięciu” parametrów alternatyw do 0 w tempie  $1/\sqrt{n}$ . Przy takim podejściu J. Neyman wyznaczył test lokalnie asymptotycznie optymalny dla  $H_0^*$  przeciwko  $A_0^*$ , w klasie testów na poziomie istotności  $\alpha$ , spełniających pewne dodatkowe naturalne warunki symetrii. To asymptotycznie optymalne rozwiązanie okazało się naturalne i łatwe do interpretacji. J. Neyman oznaczył statystykę testową, zapewne przez analogię do rozwiązania K. Pearsona, symbolem  $\Psi_k^2$ . Otrzymany test odrzuca  $H_0$  dla dużych wartości  $\Psi_k^2$ , gdzie

$$\begin{aligned}\Psi_k^2 &= n \sum_{j=1}^k \left\{ \frac{1}{n} \sum_{i=1}^n L_j(X_i) \right\}^2 = n \sum_{j=1}^k \left\{ \int_0^1 L_j(x) dF_n(x) \right\}^2 \\ &= \sum_{j=1}^k \left\{ \int_0^1 L_j(x) dU_n(x) \right\}^2,\end{aligned}$$

podczas gdy  $F_n$  jest dystrybuantą empiryczną zmiennych  $X_1, \dots, X_n$  a  $\{U_n(x), x \in [0, 1]\}$  procesem empirycznym. Tak więc  $\Psi_k^2$  jest unormowaną sumą kwadratów  $k$  pierwszych współczynników Fouriera  $F_n$  w bazie  $\{L_j\}_{j \geq 1}$ . Z drugiej strony, postać wielomianów Legendre’a implikuje, że małe wartości  $\Psi_k^2$  oznaczają nieistotne zmiany kolejnych empirycznych momentów: średniej, wariancji itd. J. Neyman przedyskutował również zachowanie funkcji mocy testu  $\Psi_k^2$  w zależności od wyboru  $k$  i do praktycznych celów zarekomendował  $\Psi_3^2$  lub  $\Psi_4^2$ . Zanotujmy, że J. Neyman doskonale zdawał sobie sprawę z tego, że wybór liczby składników w  $\Psi_k^2$  ma decydujące znaczenie dla zachowania funkcji mocy tego testu. Jego wybór gwarantował stabilną moc tylko przy bardzo ‘gładkich’ odstępstwach od postulowanego modelu.

W pracy [50] dedykowanej H. Cramérowi, J. Neyman dokonał następnego przełomowego odkrycia, podając konstrukcję klasy lokalnie asymptotycznie optymalnych testów dla problemów testowania z parametrami zakłócającymi. Tego typu problemy są kluczowe w praktyce statystycznej. Należy podkreślić, że istota pomysłu J. Neymana, pozwalającego wyeliminować wpływ parametrów zakłócających, została natychmiast podchwycona i jest eksploatowana do dziś w analizie złożonych modeli parametrycznych i semiparametrycznych. Niestety, wydaje się, że zasługi J. Neymana w tej materii nie są właściwie eksponowane. Warto też chyba nadmienić, że test

$\Psi_k^2$  był praktycznie zapomniany przez niemal 50 lat, a wprowadzona w [50] klasa testów optymalnych nie była właściwie doceniona przez niemal 40 lat.

*1.3.3. Klasa gładkich testów Neymana.* Jest rzeczą oczywistą, że rozwiązanie J. Neymana można uogólnić używając innych układów w miejsce wielomianów Legendre'a. W niniejszym opracowaniu będziemy rozważać głównie klasę gładkich testów odrzucających  $H_0$  dla dużych wartości

$$(3) \quad N_k = N_k(\Upsilon) = n \sum_{j=1}^k \left\{ \frac{1}{n} \sum_{i=1}^n \Upsilon_j(X_i) \right\}^2,$$

gdzie  $\{\Upsilon_j\}_{j \geq 1}$  jest ortonormalnym układem w  $L_2([0, 1], \lambda)$ ,  $\lambda$  miarą Lebesgue'a. Naturalnie, można pójść dalej i wprowadzić układy podwójnie indeksowane  $\{\Lambda_{j,k}\}$ ,  $1 \leq j \leq k$ ,  $k = 1, 2, \dots$  dopuszczając w ten sposób do zastosowania np. funkcji sklejanych czy falek. W szczególności, biorąc układ  $\{\Lambda_{j,k}^0\}$  zortogonalizowanych funkcji

$$b_{j,k}(x) = \{I_{(a_{j-1}, a_j]}(x) - p_{0j}\} p_{0j}^{-1/2},$$

gdzie  $I_A(\cdot)$  jest indykátorem zbioru  $A$ , a pozostałe oznaczenia wzięto z rozdziału 1.1.2, można pokazać, że statystyka  $\chi^2$  K. Pearsona ma postać

$$(4) \quad \chi^2 = n \sum_{j=1}^{k-1} \left\{ \frac{1}{n} \sum_{i=1}^n \Lambda_{j,k-1}^0(X_i) \right\}^2.$$

Statystyki o strukturze (3) lub, ogólniej,

$$(5) \quad n \sum_{j=1}^k \left\{ \frac{1}{n} \sum_{i=1}^n \Lambda_{j,k}(X_i) \right\}^2$$

nazywa się wspólnie gładkimi statystykami rzędu  $k$ .

*1.3.4. Inne rozwiązania i inne problemy testowania.* Oprócz przedstawionych tu statystyk istnieje wiele innych, będących funkcjami dystrybuanty empirycznej lub procesu empirycznego. Rozważa się również testy oparte na procesach kwantylowych, spacjach, funkcjach charakterystycznych i estymatorach gęstości. Do konstrukcji testów wykorzystuje się ponadto rozmaite charakteryzacje rozkładów, własności momentów itp. Przegląd tego typu rozwiązań można znaleźć w książce R. D'Agostino i M. Stephensa [9].

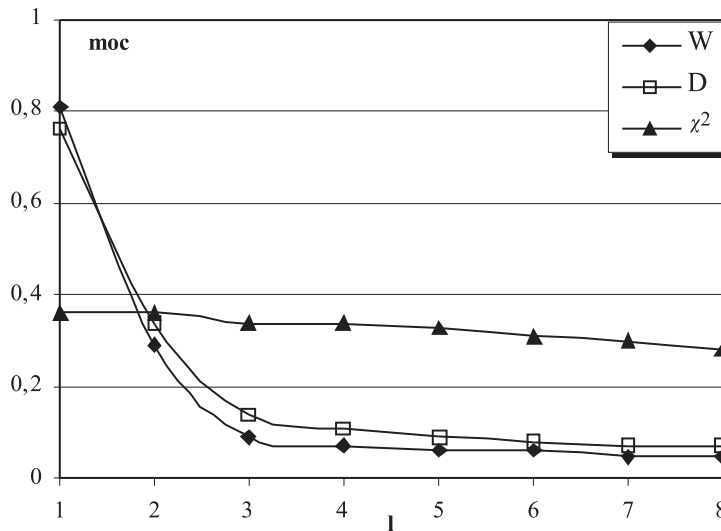
Analogiczne konstrukcje opracowano i nadal się opracowuje dla problemów testowania z parametrami zakłócającymi, rozmaitych schematów zbierania danych (np. obserwacje zależne lub cenzurowane), procesów stochastycznych, danych funkcyjnych itp. Istniejąca liczba rozmaitych wariantów i uogólnień przedstawionych idei jest przeogromna.

Pod koniec lat osiemdziesiątych zeszłego stulecia zaczęto robić pierwsze badania symulacyjne. Wykazały one między innymi, że omówione w niniejszym opracowaniu rozwiązania są proste, ale dość słabe. Dokładniej,

ich moc jest duża tylko dla specjalnych odstępstw od  $H_0$ . Dla ilustracji istoty tego problemu przedstawimy symulowane moce testu  $\chi^2$ , Craméra-von Misesa ( $W$ ) i Kolmogorowa-Smirnowa ( $D$ ) dla problemu testowania  $H_0 : f = f_0 \equiv 1$ , w sytuacji, gdy obserwacje pochodzą z rozkładu o gęstości

$$f(x; d, l) = 1 + d \cos(l\pi x) \quad d \in (0, 1], \quad l = 1, 2, \dots$$

W symulacjach wzięto poziom istotności  $\alpha = 0.05$ ,  $n = 100$  i wykonano 10000 powtórzeń. Statystyka  $\chi^2$  jest liczona dla podziału  $(0, 1)$  na 15 równych podprzedziałów.



**Rys. 1.** Symulowane moce testów Craméra-von Misesa ( $W$ ), Kolmogorowa-Smirnowa ( $D$ ) i K. Pearsona ( $\chi^2$ ) przy  $n = 100$ ,  $\alpha = 0,05$  i alternatywach  $1 + (0,4) \cos(l\pi x)$ ,  $l = 1, \dots, 8$ .

Wyniki pokazują, że moc testów  $W$  i  $D$  jest bardzo wysoka przy zaburzeniu  $f_0$  pierwszym cosinusem, a potem gwałtownie spada wraz ze wzrostem liczby oscylacji. To pośrednio implikuje, że dla alternatyw  $f$  o dużym pierwszym współczynnikiem Fouriera, w bazie cosinusów, moc tych testów też będzie wysoka. Natomiast dla alternatyw  $f$ , w których dominują wyższe składowe, moc wspomnianych testów będzie stosunkowo niska. Z kolei test  $\chi^2$ , przy powyższym dość typowym doborze klas, ma stabilną, ale niezbyt wysoką moc.

Zauważona słabość klasycznych i niezwykle popularnych testów spowodowała pod koniec lat dziewięćdziesiątych spore zainteresowanie w znalezieniu nowych, bardziej efektywnych rozwiązań. Pewne istotne ulepszenia przedstawili G. Neuhaus [48], J. Hart i R. Eubank [13] oraz P. Bickel i Y. Ritov [5]. T. Ledwina [41] zaproponowała powiązanie konstrukcji J. Neymana

z 1937 r. z nowoczesnymi metodami doboru modelu. Powstała w ten sposób procedura nazwała testem adaptacyjnym. Szczegóły tej konstrukcji podamy poniżej.

**1.4. Adaptacyjne testy Neymana.** Przypomnijmy, że rozważamy problem testowania jednostajności

$$H_0 : F = F_0$$

przeciwko

$$A : F \neq F_0, \quad f \in \mathcal{F},$$

gdzie  $\mathcal{F}$  jest klasą wszystkich ciągłych dystrybuant rozkładów skoncentrowanych na  $(0, 1)$ .

Myśląc realistycznie, jest oczywiste, że mając, powiedzmy,  $n = 100$  obserwacji niezależnych zmiennych losowych o dystrybuancie  $F$  nie jesteśmy w stanie zidentyfikować wszystkich możliwych alternatyw. Innymi słowy, przy konstrukcji testu warto uwzględnić jakieś realistyczne postulaty co do spektrum odstępstw od modelu, jakie przede wszystkim chcielibyśmy „wyłapać” z dużą częstością, przy umiarkowanych wartościach  $n$ . Postulat J. Neymana [49], aby w pierwszej kolejności skupić uwagę na ‘gładkich’ odstępstwach od modelu, wydaje się rozsądny. Z drugiej strony jest także naturalne, aby wraz ze wzrostem  $n$  stopień złożoności alternatyw, które możemy wykryć z dużym prawdopodobieństwem, również rósł.

Dla spełnienia tych dwóch powyższych postulatów T. Ledwina [41] zaproponowała, aby wyróżnić w  $\mathcal{F}$  rosnącą wraz z  $n$  rodzinę modeli  $\mathcal{F}_1^*(\theta) \subset \mathcal{F}_2^*(\theta) \subset \dots \subset \mathcal{F}_{d(n)}^*(\theta)$ , gdzie  $d(n) \rightarrow \infty$  przy  $n \rightarrow \infty$ , a elementy  $\mathcal{F}_k^*(\theta)$  mają postać

$$f_k(x; \theta) = c_k(\theta) \exp \left\{ \sum_{j=1}^k \theta_j \Upsilon_j(x) \right\},$$

podczas gdy  $\{\Upsilon_j\}_{j \geq 1}$  jest pewnym układem ortonormalnym. Przypomnijmy, że dla każdego ustalonego  $k$ , używając konstrukcji J. Neymana [49] z 1937 r., jesteśmy w stanie skonstruować optymalny test dla testowania  $f_0$  przeciwko  $f_k$ . Test ten odrzuca  $H_0$  dla dużych wartości  $N_k = N_k(\Upsilon) = n \sum_{j=1}^k \left\{ \frac{1}{n} \sum_{i=1}^n \Upsilon_j(X_i) \right\}^2$ . Cały problem z użyciem konstrukcji J. Neymana do naszych celów polega na tym, że nie wiemy jakiemu  $f_k$ , choćby w przybliżeniu, podlegają obserwacje. Jak nadmieniliśmy uprzednio, bezmyślny lub nietrafiony wybór  $k$  w  $N_k$  ma dramatyczne skutki dla mocy testu. Wobec tego w omawianej pracy zaproponowano aby, mając dane, sprawdzić niejako najpierw jak wygląda najlepiej dopasowany do nich model z listy  $\mathcal{F}_1^*(\theta) \subset \mathcal{F}_2^*(\theta) \subset \dots \subset \mathcal{F}_{d(n)}^*(\theta)$ , gdzie  $\mathcal{F}_k^*(\theta) = \{f_k(x; \theta) : \theta \in R^k\}$ . Dopiero w następnym kroku, znając ten model, a w szczególności jego wymiar  $k_0$  wybrany na podstawie posiadanych obserwacji, sprawdzić testem  $N_{k_0}$ , czy jednostajność należy odrzucić.

Powyższe rozumowanie doprowadziło do pytania o odpowiednią metodę doboru modelu  $\mathcal{F}_k^*(\theta)$ . Istnieje wiele rozwiązań w tym zakresie (por. Lanterman [38]). T. Ledwina [41] zaproponowała użycie reguły Schwarza [61], która została wyprowadzona na gruncie bayesowskiej teorii decyzji. Są znane również uzasadnienia tej metody na bazie teorii kodowania (por. Barron i Cover [3] oraz Lanterman [38]).

Prezentację reguły Schwarza ograniczymy do rozważanej powyżej sytuacji. Przypuśćmy, że niezależne zmienne losowe  $X_1, \dots, X_n$  mają rozkład o gęstości  $f_k(x; \theta)$ , każda. Przy ustalonych wartościach  $x_1, \dots, x_n$  tych zmiennych wprowadźmy następujące oznaczenia

$$L_k(\theta) = \log \prod_{i=1}^n f_k(x_i; \theta) \quad \text{oraz} \quad \hat{L}_k = \sup_{\theta \in R^k} L_k(\theta) = L_k(\hat{\theta}_k).$$

Zauważmy, że  $\hat{L}_k$  dobiera parametr  $\hat{\theta}_k = \hat{\theta}_k(x_1, \dots, x_n)$ , przy którym pojawienie się wyników  $x_1, \dots, x_n$  jest najbardziej prawdopodobne, mówiąc skrótowo. Zanotujmy również, że próba użycia  $\arg \max_{1 \leq k \leq d(n)} \hat{L}_k$  jako kryterium doboru wymiaru modelu  $\mathcal{F}_k^*(\theta)$  prowadziłyby do wyboru maksymalnego dopuszczalnego wymiaru  $d(n)$ . Zdefiniujmy dodatkowo funkcję

$$\mathcal{L}_k = \hat{L}_k - \frac{1}{2}k \log n$$

pomniejszającą  $\hat{L}_k$  o karę na wymiar modelu i wielkość próby. Optymalizacyjne rozważania Schwarza doprowadziły do reguły  $S$ , podającej optymalny wymiar modelu i danej wzorem

$$S = \min\{k, 1 \leq k \leq d(n) : \mathcal{L}_k = \max_{1 \leq j \leq d(n)} \mathcal{L}_j\}.$$

Przedstawiona argumentacja wskazuje, że  $S$  dobiera możliwie „oszczędny”, ale w miarę dobrze dopasowany model.

Reasumując, powyższe rozumowanie dało w wyniku adaptacyjną wersję gładkiej statystyki Neymana postaci  $N_S$ .

Własności  $S$  i  $N_S$  zostały wnikliwie przebadane w pracach W. Kallenberg i T. Ledwiny [31] oraz T. Ingłota i T. Ledwiny [21].

Praktyczne wyznaczenie reguły Schwarza  $S$  wymaga użycia metod numerycznych i obliczeń komputerowych. Ponadto, kwestia przeniesienia konstrukcji tej reguły na modele nieparametryczne i semiparametryczne nie wydaje się oczywista. Z tych względów w pracy W. Kallenberg i T. Ledwiny [32] zaproponowano tak zwaną uproszczoną regułę Schwarza, która w omawianym tu przypadku testowania jednostajności przyjmuje szczególnie prostą postać. Oznaczmy tę nową regułę przez  $S1$ . Mamy

$$S1 = \min\{k, 1 \leq k \leq d(n) : N_k(\Upsilon) - k \log n \geq N_j(\Upsilon) - j \log n, \\ j = 1, \dots, d(n)\}.$$

$S1$  jest prosta, bardzo dobrze naśladuje  $S$  i daje się w naturalny sposób uogólniać na bardziej złożone problemy wnioskowania. Dla krótkości prezentacji, w dalszej części tego opracowania, ograniczymy uwagę do układu  $\Upsilon_j(x) = C_j(x) = \sqrt{2} \cos(\pi j x)$ ,  $j = 1, 2, \dots$ , i związanej z nim uproszczonej reguły Schwarza  $S1$ . Dla zwięzłości będziemy pisać  $N_k$  w miejsce  $N_k(C)$ .

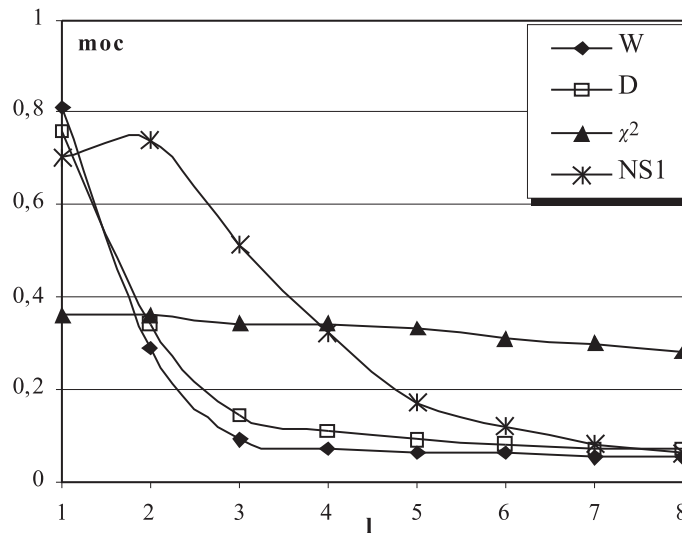
Asymptotyczne zachowanie  $S1$ , związanej z układem  $\{C_j\}_{j \geq 1}$ , przy założeniu prawdziwości  $H_0$ , opisuje następujące twierdzenie.

**TWIERDZENIE 1.** *Jeśli ciąg  $\{d(n)\}$  spełnia  $d(n)\sqrt{\log n/n} \rightarrow 0$ , to przy prawdziwości  $H_0$ , mamy*

$$\lim_{n \rightarrow \infty} P_{F_0}(S1 = 1) = 1.$$

Symulacje pokazują, że zbieżność jest szybka. Np. przy  $n = 50$  i  $n = 100$  frakcje  $\{S1 = 1\}$  wynoszą  $94/100$  i  $96/100$ , odpowiednio. Dowód twierdzenia 1 eksploatuje elegancką wykładniczą nierówność Prohorowa [59] dla rozkładu norm wektorów losowych. Twierdzenie 1 natychmiast implikuje, że granicznym rozkładem  $N_{S1}$  przy  $H_0$  jest  $\chi^2_{(1)}$ .

Rozdział ten zakończymy prezentacją i dyskusją zachowania symulowanych mocy testu opartego na  $N_{S1}$ . Rozważamy tę samą sytuację co na rys. 1 i uzupełniamy go o wyniki dla  $N_{S1}$  (patrz rys. 2).



**Rys. 2.** Symulowane moce testów  $W$ ,  $D$ ,  $\chi^2$  i  $N_{S1}$  przy  $n = 100$ ,  $\alpha = 0,05$  i alternatywach  $1 + (0, 4) \cos(l\pi x)$ ,  $l = 1, \dots, 8$ .

Widać, że używając  $N_{S1}$ , dla alternatyw generowanych przez pierwsze trzy lub cztery cosinusy uzyskano znacznie większe moce niż mają testy klasyczne. Jak stwierdzono uprzednio, podobny efekt uzyskuje się dla innych

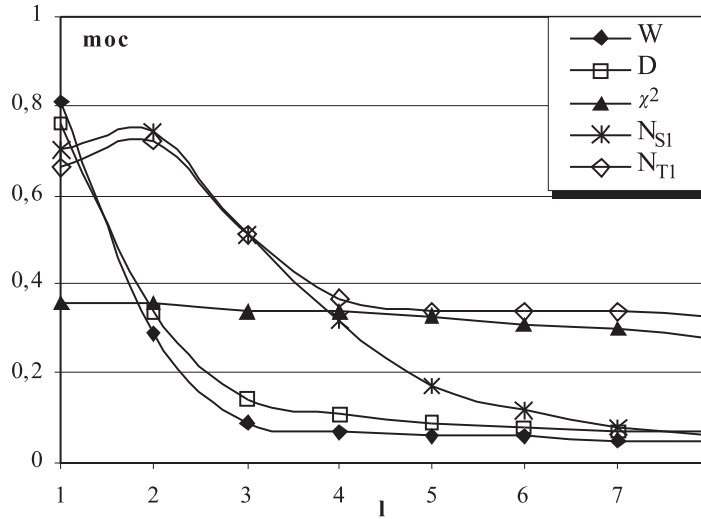
alternatyw o dominujących pierwszych czterech współczynnikach Fouriera. Powstaje też pytanie, czy taki wynik można było uznać za zadowalający. Jeśli chodzi o klasyczne problemy testowania zgodności z zadanyim rozkładem, to, naszym zdaniem, odpowiedź jest pozytywna. Bardzo obszerne badania symulacyjne pokazały, że przedstawione rozwiązanie i jego uogólnienia na problemy testowania z parametrami zakłócającymi (patrz np. Kallenberg i Ledwina [32]) oraz zagadnienia nieparametryczne (np. testowanie niezależności i problem dwóch prób; por. W. Kallenberg i T. Ledwina [33] oraz A. Janic-Wróblewska i T. Ledwina [28]) dały moce porównywalne lub przewyższające moce najlepszych istniejących rozwiązań poszczególnych problemów. Wyniki porównań z rozwiązaniami G. Neuhausa [48], R. Eubanka i J. Harta [13] oraz P. Bickela i Y. Ritova [5] też były satysfakcjonujące. Powodem takiej sytuacji jest to, że rozkłady alternatywne, jakie się rozważa w modelowaniu wielu zjawisk, są, w sensie analitycznym, bardzo gładkie i kilka pierwszych wyrazów w ich rozwinięciu w szereg Fouriera zawiera zasadniczą informację o ich kształcie. Informacja ta jest wychwytywana przez adaptacyjne testy przy stosunkowo niewielkich  $n$ . Oczywiście, wraz ze wzrostem  $n$  czułość wszystkich testów rośnie i tym samym rozszerza się zakres ich czułości. Jednakże prace teoretyczne nad porównaniem adaptacyjnych testów z rozwiązaniami klasycznymi i współczesnymi pokazały, że w pewnym sensie moce testów adaptacyjnych rosną szybciej wraz z  $n$ . Zagadnienia te dyskutujemy pokrótce w poniższym rozdziale.

Na zakończenie tej dyskusji nadmienimy, że rozwiązania, które zaproponowaliśmy, dają się rozszerzyć na dużo bardziej złożone problemy, niż te dotychczas dyskutowane. W takich zagadnieniach wysoko oscylujące alternatywy występują w sposób naturalny. Praca nad takimi zagadnieniami była motywacją do pewnych modyfikacji konstrukcji  $N_{S1}$ , powiedzmy, w celu uzyskania testu, którego zakres czułości, przy danym  $n$ , zostałby zwiększony dla wysokich oscylacji przy jednoczesnym ewentualnym minimalnym spadku mocy dla niskich oscylacji. Takie rozwiązanie przedstawiono w pracy T. Ingłota i T. Ledwiny [27]. Rys. 3 zawiera symulowane moce tego nowego rozwiązania ( $N_{T1}$ ) oraz tych uprzednio rozważanych. Prezentację samego rozwiązania pomijamy.

Zanotujmy również, że, uwzględniając reprezentację statystyki  $\chi^2$  daną (4), było w miarę oczywiste, że wypracowana metodologia konstrukcji adaptacyjnych testów może być również użyta do wyboru, na podstawie obserwacji, liczby przedziałów oraz ich końców. Takie rozwiązania podano w pracach M. Bogdan [6] oraz T. Ingłota i A. Janic-Wróblewskiej [17].

**2. Porównywanie testów.** Jak wspomnieliśmy uprzednio, J. Neyman [49] zaproponował badanie i porównywanie asymptotycznej mocy testów przy ciągach alternatyw zbieżnych do hipotezy zerowej w tempie  $1/\sqrt{n}$ .





**Rys. 3.** Symulowane moce testów  $W$ ,  $D$ ,  $\chi^2$ ,  $N_{S1}$  i  $N_{T1}$  przy  $n = 100$ ,  $\alpha = 0,05$  i alternatywach  $1 + (0, 4) \cos(l\pi x)$ ,  $l = 1, \dots, 8$ .

Może warto w tym miejscu nadmienić, jak się formalnie rozumie stwierdzenie, że ciąg gęstości  $\{f_n\}$  zbiega do  $f_0$  w tempie  $1/\sqrt{n}$ . Mianowicie, oznacza ono, że

$$\limsup_{n \rightarrow \infty} nH^2(f_n, f_0) < \infty,$$

gdzie

$$H(f_n, f_0) = \left\{ \int_0^1 (f_n^{1/2} - f_0^{1/2})^2 d\lambda \right\}^{1/2}$$

jest odległością Hellingera między  $f_n$  i  $f_0$ . Ten pomysł Neymana zdominował asymptotyczną statystykę w dwudziestym wieku. W pewnym momencie okazało się jednak, że takie podejście przestaje wystarczać. Już praca Neuhaus'a [47] pokazała, że informacja o asymptotycznej mocy tak prostej statystyki jak  $W$  Craméra-von Misesa, jaką można w ten sposób uzyskać, jest dość ograniczona i raczej opisowa niż ilościowa. Dodatkowe przejścia graniczne pozwoliły Neuhausowi na stwierdzenie, że test odrzucający  $H_0$  dla dużych wartości  $W$  jest optymalny jedynie przy alternatywie postaci  $1 + d \cos(\pi x)$ ,  $d \approx 0$ , a dla alternatyw postaci  $1 + d \cos(j\pi x)$ ,  $d \approx 0$ ,  $j \geq 1$ , jego asymptotyczna moc maleje wraz ze wzrostem  $j$ . Te opisowe wyniki są zgodne z wynikami symulacji (por. rys. 1; dodajmy, że maksymalna moc, jaką można uzyskać dla tej alternatywy przy  $j = 1$ ,  $n = 100$  i  $\alpha = 0.05$ , wynosi 0.9). W odniesieniu do  $N_{S1}$  rozważanie lokalnych alternatyw zbieżnych w tempie  $1/\sqrt{n}$  daje wręcz absurdalne wyniki (por. T. Inglot i T. Ledwina [22]).

Statystycy wypracowali również wiele alternatywnych metod porównywania testów. Część z nich nie ma, niestety, sensownej praktycznej interpretacji. Większość standardowych procedur daje się głównie stosować do statystyk, które mają w przybliżeniu strukturę liniową i ich zastosowanie do  $N_{S1}$  okazało się niemożliwe (por. dyskusja w [22]).

Powstała więc potrzeba znalezienia jakiegoś sensownego podejścia, pozwalającego zbadać zachowanie mocy  $N_{S1}$  przy dużych  $n$  i porównać je z zachowaniem mocy innych testów. W serii prac napisanych w ciągu ostatnich dziesięciu lat przez T. Inglota i T. Ledwinę oraz współpracowników (por. [10], [18], [19], [21]–[26]) rozwinięto tak zwane podejście pośrednie proponowane w pracach Kallenberg [29], [30]. Okazało się, że to rozwiązanie dało się zastosować do porównywania testów klasycznych, np. Craméra-von Misesa i Kołmogorowa-Smirnowa, jak i nowych konstrukcji Neuhausa [48], Bickela i Ritova [5], Eubanka i Harta [13], Ledwiny [41], Barauda i innych [2], itp. Prace na temat asymptotycznego porównywania testów są na ogół bardzo techniczne. Nie inaczej jest w przypadku stosowania podejścia pośredniego. Właściwie, należy to wyraźnie powiedzieć, samo podejście jest dużo bardziej złożone niż klasyczne rozwiązania (por. Inglot i Ledwina [21] oraz Inglot [16]). Jego precyzyjny opis znacznie wykracza poza ramy niniejszego opracowania. Ograniczymy się więc do zasygnalizowania, z czym się wiąże liczenie tak zwanej efektywności pośredniej Kallenberg i jak wygląda przykładowy rezultat jej zastosowania do statystyki  $W$ . Podamy również krótki komentarz na temat analogicznych wyników dla  $N_{S1}$ . W naszych pracach rozwinięliśmy dwa pomysły Kallenberg: liczenie efektywności i liczenie niedoboru mocy. Liczenie niedoboru mocy jest subtelniejsze i wymaga nieco mocniejszych narzędzi niż liczenie efektywności. Liczenia niedoboru mocy nie będziemy tu w ogóle dyskutować.

Poniżej przedstawiamy pewne fragmenty argumentacji leżącej u podstaw formalnej definicji efektywności Kallenberg. Prezentację ograniczamy do problemu liczenia efektywności testu Craméra-von Misesa względem innego testu odrzucającego  $H_0$  dla dużych wartości pewnej statystyki  $V$ .

Generalnie, statystyk chciałby mieć pewien miernik mówiący mu o ile i jakich sytuacjach jeden test jest lepszy od drugiego. Typową próbą odpowiedzi na tak postawione pytanie jest podanie jakiegoś wariantu efektywności względnej. Efektywność względną wypiszemy dla problemu testowania jednostajności, przy dodatkowym założeniu, że dystrybuanty  $F$  posiadają gęstość  $f$  względem miary Lebesgue'a na  $(0,1)$ . To ograniczenie nie jest potrzebne przy wprowadzaniu samej definicji efektywności, ale pozwoli nieco skrócić późniejszą dyskusję. Rozważamy więc

$$H_0 : f = f_0 \equiv 1$$

przeciwko

$$A : f \neq f_0.$$

Niech  $\alpha$  będzie danym poziomem istotności, a  $\beta$  zadaną liczbą,  $0 < \alpha < \beta < 1$ . Niech  $N_V(\alpha, \beta, f)$  oznacza minimalną liczbę obserwacji potrzebną na to, aby test na poziomie istotności  $\alpha$ , odrzucający  $H_0$  dla dużych wartości  $V$ , miał moc co najmniej  $\beta$  przy alternatywie  $f$ . Analogicznie definiuje się  $N_W(\alpha, \beta, f)$ . Relatywną efektywnością  $W$  względem  $V$  nazywa się iloraz

$$(6) \quad e_{W,V}(\alpha, \beta, f) = \frac{N_V(\alpha, \beta, f)}{N_W(\alpha, \beta, f)}.$$

Ta wielkość ma prostą i intuicyjną interpretację, ale jest praktycznie nie do policzenia. Dla wybrnięcia z tego kłopotu próbuje się rozważać rozmaite wielkości pochodne. Np. zastosowanie pomysłu Pitmana z 1949 r., eksplloatującego podejście Neymana [49], polegałoby na próbie wyliczenia  $\lim_{n \rightarrow \infty} e_{W,V}(\alpha, \beta, f_n)$ , gdzie  $f_n \rightarrow f_0$  w tempie  $1/\sqrt{n}$ . Wariant Bahadura z 1967 r. to  $\lim_{n \rightarrow \infty} e_{W,V}(\alpha_n, \beta, f)$ , gdzie  $\alpha_n \rightarrow 0$  w tempie wykładniczym. Niestety, w rozważanym przez nas przypadku oba te pomysły nie prowadzą do kryteriów, które można by bezpośrednio zastosować do porównywania testów. Neuhaus [47] i Nikitin [56] zastosowali dodatkowe przejścia graniczne, aby dostać dla  $W$  coś w miarę czytelne.

W ogólnej sytuacji Kallenberg [30] podał, rozwiązanie pośrednie między tymi wprowadzonymi przez Pitmana i Bahadura. Mianowicie zaproponował, aby rozważać  $\alpha_n \rightarrow 0$  wolniej niż wykładniczo i  $f_n \rightarrow f_0$  wolniej niż w tempie  $1/\sqrt{n}$ . Formalnie rzecz ujmując, mają być spełnione następujące warunki:

$$\lim_{n \rightarrow \infty} \alpha_n = \lim_{n \rightarrow \infty} n^{-1} \log \alpha_n = 0, \quad \lim_{n \rightarrow \infty} H(f_n, f_0) = 0$$

oraz

$$\lim_{n \rightarrow \infty} nH^2(f_n, f_0) = \infty.$$

Dla wyegzekwowania warunku  $0 < \beta < 1$ , występującego w definicji relatywnej efektywności i gwarantującego nietrywialność uzyskiwanych wyników, tempa zbieżności  $f_n$  i  $\alpha_n$  muszą być powiązane. Pomysł Kallenberga z grubsza wyglądał tak, że próbuje się wyliczyć wielkość

$$(7) \quad \mathcal{E}_{W,V} = \lim_{n \rightarrow \infty} e_{W,V}(\alpha_n, \beta, f_n).$$

I rzeczywiście, okazało się, że przy drobnej, ale sensownej modyfikacji definicji  $N_V(\alpha, \beta, f)$ , podanej wyżej, pewnych założeniach na zachowanie się przy  $H_0$  ogonów rozkładów statystyk  $W$  i  $V$  oraz przy zachodzeniu prawa wielkich liczb dla  $W$  i  $V$  przy  $\{f_n\}$ , granica (7) istnieje dla dużej klasy ciągów  $\{f_n\}$  i odpowiednich ciągów  $\{\alpha_n\}$ . Szczegóły, dyskusję warunków i przykłady zastosowań można znaleźć np. w pracach W. Kallenberga [30], T. Inglota i T. Ledwiny [21], [22] i [25] oraz T. Inglota [16]. Tu podamy jedynie garść informacji na temat zachowania ogonów rozkładu  $W$ , które są istotne dla istnienia efektywności Kallenberga oraz postać  $\mathcal{E}_{W,V}$  dla wybranej prostej klasy  $\{f_n\}$  i wybranej konkurencyjnej statystyki  $V$ .

Przypomnijmy, że Bahadur żądał, aby  $\alpha_n \rightarrow 0$  w tempie wykładniczym. To założenie wymuszało istnienie wielkich odchyłeń dla rozważanych statystyk testowych. W szczególności, w jego podejściu kluczową rolę odgrywała funkcja

$$\mathcal{I}(x) = - \lim_{n \rightarrow \infty} \frac{1}{nx^2} \log P_{F_0}(W \geq x\sqrt{n}).$$

Podejście Kallenberg'a wiąże się z wolniejszą asymptotyką  $\{\alpha_n\}$ . To z kolei implikuje, że w omawianym tu przykładzie centralną rolę pełnią oszacowania dla  $P_{F_0}(W \geq x_n\sqrt{n})$  oraz granice

$$(8) \quad c_W = - \lim_{n \rightarrow \infty} \frac{1}{nx_n^2} \log P_{F_0}(W \geq x_n\sqrt{n}),$$

gdzie  $x_n \rightarrow 0$  oraz  $nx_n^2 \rightarrow \infty$ . Dla ilustracji przytaczamy oszacowanie, które można uzyskać z rezultatów pracy T. Inglota i T. Ledwiny [20].

**TWIERDZENIE 2.** *Jeśli  $x_n \rightarrow 0$  i  $nx_n^2 \rightarrow \infty$ , to dla każdego  $\gamma \in (2, 3)$  zachodzi*

$$(9) \quad P_{F_0}(W \geq x_n\sqrt{n}) = \exp \left\{ - \frac{1}{2} \pi^2 nx_n^2 + O(nx_n^\gamma) + O(\log[nx_n^2]) \right\}.$$

Oczywiście, (9) natychmiast implikuje (8) z  $c_W = \pi^2/2$ .

Dla sprawdzenia jak sprawnie statystyka  $W$  wydobywa informację o gęstości  $f$ , zawartą w próbie  $X_1, \dots, X_n$ , będziemy liczyć jej efektywność względem najlepszego testu dla weryfikowania  $H_0 : f = f_0$  przeciwko  $A_1 : f = f_n$ , gdzie  $f_n$  jest  $n$ -tym elementem ciągu gęstości  $\{f_n\}$ . Taki najlepszy (najmocniejszy) test został skonstruowany w 1933 r. przez J. Neymana i E. Pearsona. Odrzuca on  $H_0$  dla dużych wartości statystyki  $\sum_{i=1}^n \log f_n(X_i)$  lub równoważnie, jej standaryzowanej wersji

$$V = \frac{1}{\sqrt{n}\sigma_{0n}} \sum_{i=1}^n \{\log f_n(X_i) - e_{0n}\},$$

gdzie

$$e_{0n} = \int_0^1 \log f_n d\lambda \quad \text{oraz} \quad \sigma_{0n}^2 = \int_0^1 (\log f_n)^2 d\lambda - e_{0n}^2.$$

Jest chyba jasne, że w rozważanym zagadnieniu używanie  $V$  nie jest możliwe, bo  $f_n$  nie jest znane. Test ten jest jednakże rodzajem pewnego wzorca pokazującego jaka maksymalna moc jest teoretycznie możliwa do uzyskania w punkcie  $f_n$ , przy danym  $n$  i  $\alpha$ . Przy takim wyborze  $V$  zawsze mamy  $\mathcal{E}_{W,V} \leq 1$ . Jeśli dla jakiegoś ciągu  $\{f_n\}$  zachodzi równość, to mówimy, że  $W$  jest optymalny przy  $\{f_n\}$ . Porównując także inne testy do wzorca  $V$  dostajemy jednocześnie informację o ich wzajemnej relacji. Odpowiednik (8) dla  $V$  można wydedukować z raportu Booka (1996) (por. twierdzenie 5.8 w pracy [21]).

Rozwinięta przez nas teoria pozwala liczyć  $\mathcal{E}_{W,V}$  dla dowolnych ciągów alternatyw spełniających  $H(f_n, f_0) \rightarrow 0$  i  $nH^2(f_n, f_0) \rightarrow \infty$ . Jednakże dla prostoty prezentacji przedstawimy postać efektywności  $\mathcal{E}_{W,V}$  jedynie dla wybranej, standardowej w literaturze, klasy ciągów  $\{f_n\}$ . Mianowicie, rozważymy wszystkie ciągi o elementach postaci

$$(10) \quad f_n(x) = 1 + \theta_n a(x), \quad \theta_n \rightarrow 0, \quad n\theta_n^2 \rightarrow \infty,$$

gdzie

$$(11) \quad a \in \mathcal{A} = \left\{ a : \sup_{x \in [0,1]} |a(x)| < \infty, \int_0^1 a d\lambda = 0, \int_0^1 a^2 d\lambda = 1 \right\}.$$

Poniższe twierdzenie pochodzi z pracy T. Inglota i T. Ledwiny [22].

**TWIERDZENIE 3.** *Dla każdego ciągu alternatyw  $\{f_n\}$  danego przez (10) i (11) oraz  $\{\alpha_n\}$  zdefiniowanego wzorem (12) zachodzi*

$$\mathcal{E}_{W,V} = (\pi \|A\|)^2,$$

gdzie

$$A(x) = \int_0^x a d\lambda, \quad \|A\| = \left\{ \int_0^1 A^2 d\lambda \right\}^{1/2} = \left\{ \sum_{j=1}^{\infty} \left( \frac{1}{\pi j} \right)^2 c_j^2 \right\}^{1/2},$$

$$c_j = \int_0^1 a C_j d\lambda,$$

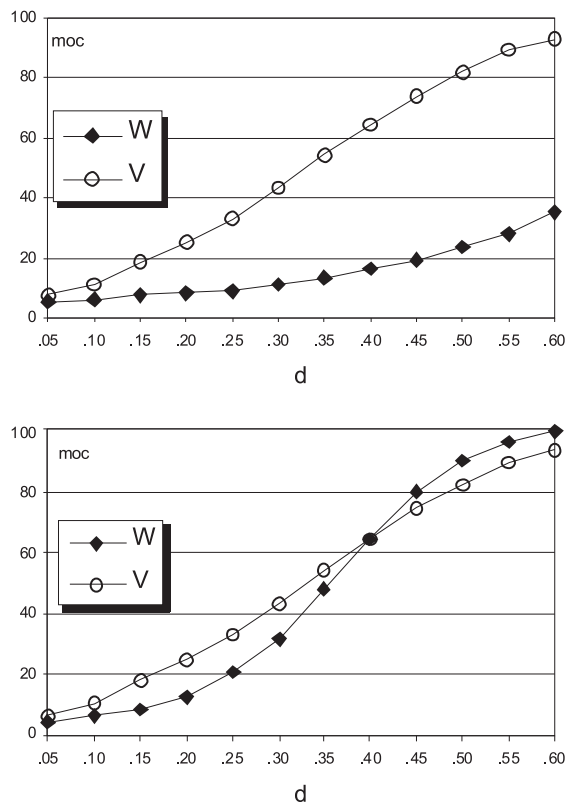
a  $C_j(x) = \sqrt{2} \cos(j\pi x)$ . Związana z  $\{f_n\}$ , postać  $\alpha_n$ , gwarantująca niezdegenerowaną asymptotyczną moc  $\beta \in (0, 1)$ , jest dana przez

$$(12) \quad \alpha_n = \exp \left\{ -\frac{1}{2} \pi^2 (n\theta_n^2) \|A\|^2 + O(\sqrt{n}\theta_n) \right\}.$$

Ponieważ  $\int a^2 = 1$ , powyższy wynik implikuje, że  $\mathcal{E}_{W,V} = 1$  wtedy i tylko wtedy, gdy  $a(x) = C_1(x)$ . Dla każdego innego  $a \in \mathcal{A}$  mamy  $\mathcal{E}_{W,V} < 1$ . W szczególności, dla  $a(x) = C_j(x)$ ,  $j = 2, 3, \dots$  otrzymujemy

$$(13) \quad \mathcal{E}_{W,V} = 1/j^2.$$

Wynik ten jest zgodny z opisowymi rezultatami Neuhaus'a [47]. Interpretacja (13) jest następująca. Dla uzyskania przez  $W$  takiej samej mocy jaką ma  $V$ , wyliczone na bazie  $n$  obserwacji, należy wziąć dla powyższych lokalnych alternatyw  $nj^2$  obserwacji. Oczywiście, w praktyce mamy ustalone  $\alpha$  i  $n$  oraz ustalone alternatywy. Badania symulacyjne pokazują jednakże, że lokalna teoria ma sensowne odzwierciedlenie w rzeczywistej sytuacji. Dla ilustracji



**Rys. 4.** Empiryczne moce testów Neymana-Pearsona ( $V$ ) i Craméra-von Misesa ( $W$ ) przy  $\alpha = 0,05$  i alternatywach  $1 + d \cos(2\pi x)$ ,  $d = 0,05, \dots, 0,60$ . Góra: empiryczne moce  $V$  i  $W$  przy  $n = 50$ . Dół: empiryczna moc  $V$  przy  $n = 50$  i empiryczna moc  $W$  przy  $n = 200$ .

przytaczamy jeden z rysunków z pracy [24] (rys. 4). Analogiczne rozważania i ilustracje dla testu Kołmogorowa-Smirnowa podano w [25].

Zastosowanie efektywności Kallenberg'a do porównywania  $N_{S1}$  i  $V$  jest dużo bardziej złożone. W szczególności, dla  $N_{S1}$  nie istnieją umiarkowane odchylenia w pełnym zakresie  $\{x_n\}$  (odpowiednik (8)). Ponadto, oprócz zgrania temp zbieżności  $\{f_n\}$  i  $\{\alpha_n\}$ , trzeba jeszcze uwzględnić tempo w jakim przyrasta długość listy modeli  $d(n)$ . Tym niemniej daje się pokazać (por. T. Inglot i T. Ledwina [21], [22] i T. Inglot [16]), że przy szerokiej klasie alternatyw  $\{f_n\}$  zbieżnych do  $f_0$  w tempie wolniejszym niż  $1/\sqrt{n}$  zachodzi  $\mathcal{E}_{N_{S1},V} = 1$ . Była to pierwsza znana nam konstrukcja o takiej własności. Można też pokazać, że podobny wynik zachodzi dla  $N_{T1}$ .

W 2003 r. Baraud, Huet i Laurent podali alternatywną konstrukcję, która też jest optymalna w sensie Kallenberg'a (por. T. Inglot i T. Ledwina [26]).

Na zakończenie chcielibyśmy przedstawić kilka uwag na temat interpretacji wyniku  $\mathcal{E}_{N_{S1},V} = 1$ . Nie można zapomnieć, że jest to wynik graniczny,

osiągnięty dla pewnej pomocniczej abstrakcyjnej sytuacji i jego przeniesienie na sytuację rzeczywistą, gdzie mamy ustalone  $n$ ,  $\alpha$  i alternatywę (dla nas nieznaną), nie może być literalne. Zbieżność do wartości 1 nie jest jednostajna po klasie alternatyw. Tak więc, jeśli rozpatrujemy np. empiryczne moce  $N_{S1}$  przy alternatywach typu  $1 + dC_j(x)$  i danym  $n$  oraz  $\alpha$  (por. rys. 2. i podane tam warunki), to nie należy oczekiwać, że dla wszystkich  $j = 1, 2, \dots$  moc testu  $N_{S1}$  będzie jednakowo bliska mocy testu najlepszego; to jest .90. Konstrukcja  $N_{T1}$  pokazuje, że dla skończonego  $n$  inny asymptotycznie optymalny test może mieć nieco inną funkcję mocy (por. rys. 3.). Wynik  $\mathcal{E}_{N_{S1},V} = 1$  oznacza, że konstrukcja T. Ledwiny (1994), wraz ze wzrostem  $n$ , wyciąga, w pewnym sensie, całą dostępną informację o alternatywie, która jest zawarta w próbie  $X_1, \dots, X_n$ . Wynik ten wskazuje również, że nazwa adaptacyjny test Neymana jest adekwatna i dobrze uzasadniona. Konstrukcja  $N_{T1}$  pokazała z kolei, że dla skończonego  $n$  asymptotycznie optymalne rozwiązanie można jeszcze poprawić.

**Podziękowania.** Niniejszy tekst pokrywa się w dużej mierze z treścią wykładu wygłoszonego na XVI Zjeździe Matematyków Polskich we Wrocławiu w 2005 r. Autorka dziękuje Organizatorom za zaproszenie, a prof. dr hab. R. Dudzie za propozycję przesłania tego tekstu do *Wiadomości Matematycznych*.

#### Literatura

- [1] J. Arbuthnott, *An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes*, Philos. Trans. Roy. Soc., London., **27** (1710), 186–190.
- [2] Y. Baraud, S. Huet, B. Laurent, *Adaptive tests of linear hypotheses by model selection*, Ann. Statist. **31** (2003), 225–251.
- [3] A. R. Barron, T. M. Cover, *Minimum complexity density estimation*, IEEE Trans. Inf. Th. **37** (1991), 1034–1054.
- [4] D. Bernoulli, *Quelle est la cause physique de l'inclinaison de planètes...*, Recueil des Pièces qui ont Remporté le Prix de l'Académie Royale des Sciences **3** (1734), 95–122.
- [5] P. J. Bickel, Y. Ritov, *Testing for goodness of fit: a new approach*, w: A. K. Md. E. Saleh (redaktor), *Nonparametric Statistics and Related Topics*, North-Holland, Amsterdam 1992, 51–57.
- [6] M. Bogdan, *Data driven versions of Pearson's chi-square test for uniformity*, J. Statist. Comput. Simulation **52** (1995), 217–237.
- [7] H. Cramér, *On the composition of elementary errors. Second paper: statistical applications*, Skand. Aktuarietidskrift. **11** (1928), 141–180.
- [8] H. Cramér, *Collected Works, I*, A. Martin-Löf (edytor), Springer, Berlin 1993.
- [9] R. B. D'Agostino, M. A. Stephens, *Goodness-of-Fit Techniques*, Marcel Dekker, New York 1986.
- [10] G. R. Ducharme, T. Ledwina, *Efficient and adaptive nonparametric test for the two-sample problem*, Ann. Statist. **31** (2003), 2036–2058.

- [11] J. Durbin, *Distribution Theory for Tests Based on the Sample Distribution Function*, SIAM, Philadelphia 1973.
- [12] F. Y. Edgeworth, *Methods of Statistics, Jubilee volume of the Statist. Soc.*, E. Stanford, London 1885.
- [13] R. L. Eubank, J. D. Hart, *Testing goodness of fit in regression via order selection criteria*, Ann. Statist. **20** (1992), 1412–1425.
- [14] J. Gavarret, *Principes généraux de statistique médicale*, Paris 1840.
- [15] C. Huyghens, *De ratiociniis in ludo aleae, Exercitationum Mathematicarum*, 517–534, Leiden 1657.
- [16] T. Inglot, *Generalized intermediate efficiency of goodness of fit tests*, Math. Methods Statist. **8** (1999), 487–509.
- [17] T. Inglot, A. Janic-Wróblewska, *Data driven chi-square test for uniformity with unequal cells*, J. Stat. Comput. Simulation **73** (2003), 545–561.
- [18] T. Inglot, W. C. M. Kallenberg, T. Ledwina, *Vanishing shortcoming of data driven Neyman's tests*, *Asymptotic Methods in Probability and Statistics*, w: B. Szyszkowicz (redaktor), *A volume to honour Miklós Csörgő*, Elsevier, Amsterdam 1978, 811–829.
- [19] T. Inglot, W. C. M. Kallenberg, T. Ledwina, *Vanishing shortcoming and asymptotic relative efficiency*, Ann. Statist. **28** (2000), 215–238
- [20] T. Inglot, T. Ledwina, *On probabilities of excessive deviations for Kolmogorov-Smirnov, Cramér-von Mises and chi-square statistics*, Ann. Statist. **18** (1990), 1491–1495.
- [21] T. Inglot, T. Ledwina, *Asymptotic optimality of data driven Neyman's tests for uniformity*, Ann. Statist. **24** (1996), 1982–2019.
- [22] T. Inglot, T. Ledwina, *Intermediate approach to comparison of some goodness-of-fit tests*, Ann. Inst. Statist. Math. **53** (2001), 810–834.
- [23] T. Inglot, T. Ledwina, *Asymptotic optimality of data driven smooth tests for location-scale family*, Sankhy-a, Ser. A. **63** (2001), 41–71.
- [24] T. Inglot, T. Ledwina, *On consistent minimax distinguishability and intermediate efficiency of Cramér-von Mises test*, J. Statist. Plan. Inference. **124** (2004), 453–474.
- [25] T. Inglot, T. Ledwina, *Intermediate efficiency of some max-type statistics*, J. Statist. Plan. Inference (w druku).
- [26] T. Inglot, T. Ledwina, *Asymptotic optimality of new adaptive test in regression model*, Annales de L'Institut Henri Poincaré, Probab. & Stat. (w druku).
- [27] T. Inglot, T. Ledwina, *Towards data driven selection of a penalty function for data driven Neyman tests*, Linear Algebra Appl., A volume to honour Ingram Olkin (w druku).
- [28] A. Janic-Wróblewska, T. Ledwina, *Data driven rank test for two-sample problem*, Scand. J. Statist. **27** (2000), 281–297.
- [29] W. C. M. Kallenberg, *Asymptotic Optimality of Likelihood Ratio Tests in Exponential Families*, Mathematical Centre Tracts No. 77, Amsterdam 1978.
- [30] W. C. M. Kallenberg, *Intermediate efficiency, theory and examples*, Ann. Statist. **11** (1983), 170–182.
- [31] W. C. M. Kallenberg, T. Ledwina, *Consistency and Monte Carlo simulation of data driven version of smooth goodness of fit tests*, Ann. Statist. **23** (1995), 1594–1608.
- [32] W. C. M. Kallenberg, T. Ledwina, *Data driven smooth tests when the hypothesis is composite*, J. Amer. Statist. Assoc. **92** (1997), 1094–1104.



- [33] W. C. M. Kallenber g, T. Ledwina, *Data driven rank tests for independence*, J. Amer. Statist. Assoc. **94** (1999), 285–301.
- [34] M. G. Kendall, R. L. Plackett (redaktorzy), *Studies in the History of Statistics and Probability*, II. Charles Griffin & Co Ltd, London 1977.
- [35] A. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*, Ist. Ital. Attuari, G., **4** (1933), 1–11.
- [36] S. Kotz, N. L. Johnson (redaktorzy), *Breakthroughs in Statistics, I (Foundations and Basic Theory)*, Springer-Verlag, New York 1992.
- [37] S. Kotz, N. L. Johnson, (edytorzy), *Breakthroughs in Statistics, II (Methodology and Distributions)*, Springer-Verlag, New York 1992.
- [38] A. D. Lanterman, *Schwarz, Wallace, and Rissanen: intertwining themes in theories of model selection*, Int. Statist. Rev. **69** (2001), 185–212.
- [39] P. S. Laplace (1773), *Mémoire sur l'inclinaison moyenne des orbites des comètes*, Mem. Acad. Roy. Sci. Paris, **VII** (1776), 503–524.
- [40] L. Le Cam, E. L. Lehmann, *J. Neyman On the occasion of his 80th birthday*, Ann. Statist. **2** (1974), vii–xiii.
- [41] T. Ledwina, *Data driven version of the Neyman smooth test of fit*, J. Amer. Statist. Assoc. **89** (1994), 1000–1005.
- [42] T. Ledwina, *Idee Neymana w teorii testowania hipotez*, w: S. Domoradzki, Z. Pawlikowska-Brozek, D. Węglowska (redaktorzy), *XII Szkoła Historii Matematyki*, Wydawnictwo AGH (1999), 37–44.
- [43] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York 1986.
- [44] W. Lexis, *Einleitung in die Theorie der Bevölkerungsstatistik*, Strassburg 1875.
- [45] W. Lexis, *Zur Theorie der Massenerscheinungen in der Menschlichen Gesellschaft*, Wagner, Freiburg 1877.
- [46] R. von Mises, *Vorlesungen aus dem Gebiete der Angewandten Mathematik, 1, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik*, Deuticke, Leipzig 1931.
- [47] G. Neuhäus, *Asymptotic power properties of the Cramér-von Mises test under contiguous alternatives*, J. Multivariate Anal. **6** (1976), 95–110.
- [48] G. Neuhäus, *Addendum to: „Local asymptotics for linear rank statistics with estimated score functions”*, Ann. Statist. **16** (1988), 1342–1343.
- [49] J. Neyman, *‘Smooth’ test for goodness of fit*, Scand. Aktuarietidskr. **20** (1937), 149–199.
- [50] J. Neyman, *Optimal asymptotic tests of composite statistical hypotheses*, Probability and Statistics (The Harald Cramér Volume), 213–234, Almqvist and Wiksells, Uppsala 1959.
- [51] J. Neyman, E. S. Pearson, *On the use and interpretation of certain certain test criteria for purposes of statistical inference. Part I*, Biometrika **20** A (1928), 175–240.
- [52] J. Neyman, E. S. Pearson, *On the use and interpretation of certain test criteria for purposes of statistical inference. Part II*, Biometrika **20** A (1928), 263–294.
- [53] J. Neyman, E. S. Pearson, *On the problem of most efficient tests of statistical hypotheses*, Phil. Trans. Roy. Soc. A **231** (1933), 289–337.
- [54] J. Neyman, E. S. Pearson, *Contributions to the theory of testing statistical hypotheses. Part I. Unbiased critical regions of type A and type  $A_1$* , Statist. Res. Mem. **1** (1936), 1–37.

- [55] J. Neyman, E. S. Pearson, *Contributions to the theory of testing statistical hypotheses. Part II. Certain theorems on unbiased critical regions of type A*, Statist. Res. Mem. **2** (1938), 25–57.
- [56] Ya. Yu. Nikitin, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge Univ. Press, Cambridge 1995.
- [57] E. S. Pearson, M. G. Kendall (redaktorzy), *Studies in the History of Statistics and Probability*, Carles Griffin & Co Ltd, London 1970.
- [58] K. Pearson, *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling* London, Edinburgh and Dublin Philos. Mag. and J. of Sci. **50** (1900), 157–172.
- [59] A. V. Prohorov, *On sums of random vectors*, Theory Probab. Appl. **18** (1973), 186–188.
- [60] D. B. Rubin, *Comment: Neyman (1923) and causal inference in experiments and observational studies*, Statist. Sci. **5** (1990), 472–480.
- [61] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), 461–464.
- [62] N. V. Smirnov, *Sur la distribution de  $\omega^2$  (criterium de M. R. von Mises)*, C. R. Acad. Sci. (Paris) **202** (1936), 449–452.
- [63] J. Sława-Neyman, *On the application of probability theory to agricultural experiments. Essay on principles. Section 9*, Statist. Sci. **5** (1990), 465–472.
- [64] S. M. Stigler, *The History of Statistics. The Measurement of Uncertainty before 1900*, The Belknap Press, Cambridge, 1986.
- [65] *A Selection of Early Statistical Papers of J. Neyman*, University of California Press, Berkeley 1967.
- [66] *Joint Statistical Papers of J. Neyman & E. S. Pearson*, University of California Press, Berkeley 1967.