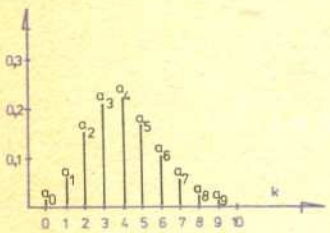
Dżem z dyni o smaku pomarańczowym, czyli o testowaniu hipotez statystycznych

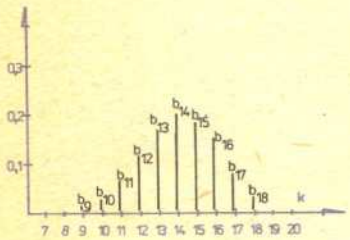
Prof. dr Jan ODERFELD, doc. dr Elżbieta PLESZCZYŃSKA

Bez względu na swoje aspiracje dżem z dyni jest dżemem z dyni, a nalepka „dżem z dyni o smaku pomarańczowym" służy jedynie celom reklamowym. Testowanie hipotez statystycznych, stosowane do wnioskowania o prawach rządzących przebiegiem eksperymentu na podstawie jego wyników, jest często reklamowane niewspółmiernie do jego rzeczywistych zalet. Sugestie „smaku pomarańczowego" wynikają z rozważań teoretycznych dotyczących testowania abstrakcyjnych, formalnie wyrażonych hipotez oraz ze starannie dobranych przykładów podawanych w podręcznikach. Ostrzeżenia przed niewłaściwym stosowaniem teorii, ukazujące różnice między „dynią" a „pomarańczą", pojawiają się zbyt rzadko, a bezkrytyczne postępowanie wielu badaczy prowadzi do poważnych nieporozumień. Postaramy się przedstawić, na czym one polegają.

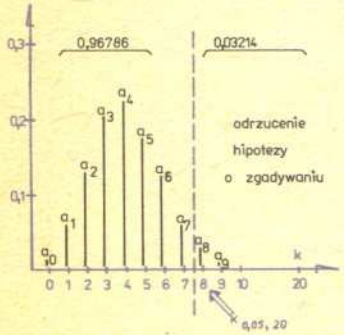
Badanie percepcji pozazmysłowej



Prawdopodobieństwa a_k wg wzoru (1) przy $n = 20$.



Prawdopodobieństwa b_k wg wzoru (2) przy $n = 20$, $p = 0{,}7$.



	$k_{0{,}05;n}$	$k_{0{,}01;n}$
6	4	5
10	5	6
20	8	9
50	16	18

Jak pisze D. S. Moore w artykule o analizie staty stycznej danych doświadczalnych (*Matematyka współczesna*, PWN, 1983), standardowy eksperyment do badania zdolności percepcji pozazmysłowej (*PP*) jest prowadzony za pomocą pięciu rodzajów kart. Eksperymentator wybiera n-krotnie po jednej karcie z przetasowanej talii zawierającej jednakową liczbę kart każdego rodzaju, a badany osobnik siedzący za zasłoną stara się tę kartę rozpoznać. Jeśli eksperyment przeprowadzany jest w sposób właściwy, to

a) eksperymentator wybiera kolejne karty bez jakiegokolwiek związku z poprzednimi i następnymi kartami,

b) w każdej kolejnej próbie każdy z rodzajów występuje z jednakowym prawdopodobieństwem $\frac{1}{5}$,

c) osobnik bez zdolności *PP* odgaduje poprawnie rodzaj karty z prawdopodobieństwem $\frac{1}{5}$,

a więc w n próbach uzyskuje k trafień z prawdopodobieństwem

$$a_k = \binom{n}{k}\left(\frac{1}{5}\right)^k\left(\frac{4}{5}\right)^{n-k} \qquad \text{dla } k = 0, 1, \ldots, n. \tag{1}$$

A jakie prawa rządziłyby przebiegiem eksperymentu, gdyby badany miał w jakimś stopniu zdolność *PP*? Można chyba założyć, że prawa te są równoważne losowaniu liczby trafień z pewnym prawdopodobieństwem $b_k (k = 0, \ldots, n)$, przy czym im liczba trafień jest większa, tym silniej przemawia za istnieniem zdolności *PP*. Formalnie wyrażamy to żądając, żeby ilorazy $\frac{b_k}{a_k}$ tworzyły ciąg niemalejący i nie były wszystkie jednakowe (gdy ilorazy są jednakowe, to $b_k = a_k$ dla $k = 0, \ldots, n$). Jeśli osobnik obdarzony zdolnością *PP* daje trafną odpowiedź z prawdopodobieństwem p większym od $\frac{1}{5}$, przy czym p jest jednakowe w kolejnych próbach, a odpowiedzi w poszczególnych próbach nie są wzajemnie uzależnione, to wtedy

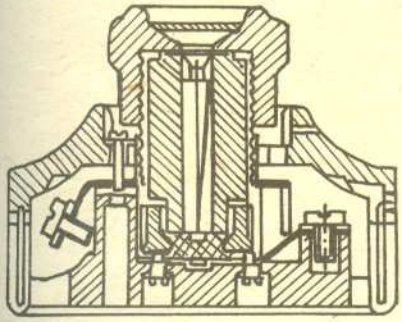
$$b_k = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

i ciąg $\frac{b_0}{a_0}, \ldots, \frac{b_n}{a_n}$ jest rosnący, przy $\frac{1}{5} < p < 1$; gdy $p = 1$, to $b_0 = \ldots = b_{n-1} = 0$ i $b_n = 1$, a więc ciąg jest niemalejący.

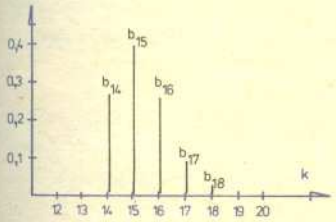
Jeśli badany uzyska liczbę trafień k na tyle wysoką, że prawdopodobieństwo osiągnięcia przez zgadywanie liczby trafień nie mniejszej niż k jest bardzo małe (mniejsze od z góry ustalonej dostatecznie małej liczby α, na przykład $\alpha = 0{,}05$), to możemy uznać ten fakt za argument potwierdzający istnienie zdolności *PP* (bądź niewłaściwego zorganizowania eksperymentu).

Liczbę α nazywa się zwykle poziomem istotności. Przy ustalonych α i n możemy na podstawie (1) wyznaczyć najmniejszą liczbę trafień oznaczoną przez $k_{\alpha,n}$, która powoduje już odrzucenie hipotezy o zgadywaniu na rzecz hipotezy o istnieniu zdolności *PP*. Liczbę trafień nie mniejszą niż $k_{\alpha,n}$ nazywa się statystycznie istotną na poziomie α.

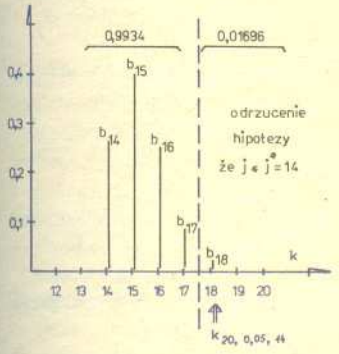
α	n	prawdopodobieństwo odrzucenia hipotezy o zgadywaniu przy założeniu (3): $b_{k_{\alpha,n}} + \ldots + b_n$			
		p			
		0,3	0,5	0,7	0,9
0,05	6	0,070	0,344	0,744	0,984
	10	0,150	0,623	0,953	1,000
	20	0,228	0,868	0,999	1,000
	50	0,431	0,997	1,000	1,000
0,01	6	0,011	0,109	0,420	0,886
	10	0,047	0,377	0,850	0,998
	20	0,113	0,748	0,995	1,000
	50	0,218	0,984	1,000	1,000



Jakie urządzenie przedstawiono na rysunku
— wyłącznik stycznikowy,
— wyłącznik topikowy,
— lampę elektronową,
— piec donicowy,
— pompę wirnikową?



Prawdopodobieństwa b_k wg wzoru (3) przy $n = 20, j = 14$.



j	l. punktów	j	l. punktów
⩽ 5	1	12	5
6	2	13	6
7	2	14	7
8	3	15	8
9	3	16	8
10	4	17	9
11	4	18	9
		⩾ 19	10

Gdyby osobnik był obdarzony zdolnością *PP*, to odrzucenie hipotezy o zgadywaniu nastąpiłoby z prawdopodobieństwem $\sum_{i \geqslant k_{\alpha,n}} b_i$. Przy założeniu (2) i przy ustalonych p i α prawdopodobieństwo to rośnie do 1 przy $n \to \infty$, a przy ustalonych n i α jest rosnącą funkcją argumentu p. Sądzimy, że takie badanie zdolności *PP* za pomocą testowania hipotezy o zgadywaniu nie powinno budzić zastrzeżeń eksperymentatorów.

Sprawdzian ogólnej orientacji technicznej dla finalistów Olimpiady Wiedzy Technicznej

W trzecim etapie Olimpiady Wiedzy Technicznej sprawdza się m.in. ogólną orientację techniczną, polegającą na dość powierzchownej, ale za to bardzo szerokiej znajomości różnych aspektów techniki. Finalista otrzymuje na przykład n schematycznych rysunków wyobrażających różne obiekty techniczne. Pod każdym rysunkiem jest 5 nazw, z których dokładnie jedna jest właściwa. Finalista oznacza krzyżykiem tę jedyną nazwę, którą uznaje za poprawną. Ocena ogólnej orientacji technicznej finalisty opiera się na łącznej liczbie trafień k.

Formalny opis sprawdzianu jest taki sam jak formalny opis badania zdolności *PP*. Pięciu rodzajom kart odpowiada pięć podpisów pod rysunkiem. Gdyby finalista nie rozpoznał żadnego rysunku i zgadywał swoje odpowiedzi dając jednakowe szanse każdej z pięciu możliwości, uzyskałby k trafień z prawdopodobieństwem a_k (patrz (1)). W przeciwnym razie prawdopodobieństwo jest równe b_k, przy czym ciąg $\frac{b_0}{a_0}, \ldots, \frac{b_n}{a_n}$ jest niemalejący. Na przykład można sobie wyobrazić, że finalista rozpoznaje j spośród n rysunków (przy czym j jest pewną liczbą naturalną nie większą niż n, charakteryzującą jego ogólną orientację techniczną) i oznacza krzyżykami odpowiadające im nazwy, a pozostałych $n-j$ krzyżyków stawia na chybił trafił z jednakową szansą $\frac{1}{5}$ dla każdej odpowiedzi. Wtedy dla ustalonej liczby j prawdopodobieństwo b_k jest równe

$$b_k = \begin{cases} 0 & \text{dla } k = 0, \ldots, j-1, \\ \binom{n-j}{k-j} \left(\frac{1}{5}\right)^{k-j} \left(\frac{4}{5}\right)^{n-k} & \text{dla } k = j, \ldots, n \end{cases} \tag{3}$$

i ciąg ilorazów $\frac{b_k}{a_k}$ jest niemalejący. Taki model ma sens wtedy, gdy odpowiedzi do rysunków są bardzo starannie dobrane, sprawdzian jest trudny, a czas jego trwania krótki. Gdyby ograniczyć się do prawdopodobieństw b_k postaci (3), to hipotezę o kompletnym braku ogólnej orientacji można by zapisać za pomocą $j = 0$, gdyż wtedy $b_k = a_k$.

Pozornie wydaje się, że przy ocenie ogólnej orientacji technicznej można by postępować analogicznie jak przy badaniu zdolności *PP* i testować hipotezę o zgadywaniu, odrzucając ją przy $k \geqslant k_{\alpha,n}$ dla obranego α. Czy jednak miałoby to sens? Trudno przypuszczać, że finalista nie ma żadnej orientacji technicznej. Podstawową różnicę między obu doświadczeniami stanowi to, że w pierwszym z nich *każde* nawet bardzo małe odchylenie od zgadywania jest dla eksperymentatora godne uwagi, a więc statystycznie istotny wynik doświadczenia, odrzucający hipotezę o zgadywaniu, budzi zrozumiałe zainteresowanie. Natomiast jurora Olimpiady interesuje raczej to, czy finalista ma *dostatecznie dużą* ogólną orientację techniczną. Formalnie można to wyrazić jako hipotezę, że liczba j rysunków rzeczywiście rozpoznanych przez finalistę (której nie należy mylić z liczbą trafień k) przekracza ustaloną przez jury liczbę j^*. Zamiast hipotezy o zgadywaniu stawiamy więc hipotezę, że prawdopodobieństwa b_k są postaci (3) przy $j \leqslant j^*$; odrzucenie takiej hipotezy prowadzi do uznania, że finalista ma dostatecznie dużą ogólną orientację techniczną.

Analogicznie jak poprzednio ustalamy poziom istotności α i wyznaczamy taką najmniejszą liczbę naturalną k_{α,n,j^*}, że dla dowolnego $j \leqslant j^*$ prawdopodobieństwo uzyskania co najmniej tylu trafień nie przekracza α. W tym celu wystarczy brać pod uwagę prawdopodobieństwa $b_1, \ldots, b_n$ przy $j = j^*$, gdyż dla dowolnego $k = 1, \ldots, n$ prawdopodobieństwo uzyskania co najmniej k trafień rośnie ze wzrostem j.

Wydaje się jednak, że organizator Olimpiady jest przede wszystkim zainteresowany oceną liczby j, to jest liczby rysunków rozpoznanych przez finalistę; każdej liczbie j gotów byłby przypisać pewną liczbę punktów (na przykład zgodnie z systemem punktacji podanym obok).

k	$\hat{j}(k)$	k	$\hat{j}(k)$
$\leqslant 4$	0		
5	2	13	12
6	4	14	13
7	4	15	14
8	6	16	16
9	7	17	17
10	8	18	18
11	9	19	19
12	11	20	20

Estymator NW przy założeniu (3) i przy $n = 20$.

Zadanie sprowadza się więc do oszacowania liczby j na podstawie liczby trafień k. Można zaproponować różne estymatory $\hat{j}(k)$ o różnych własnościach. Jednym z nich jest estymator największej wiarygodności (NW), przypisujący liczbie k taką ocenę liczby j, przy której b_k osiąga największą wartość. Na przykład przy $n = 20$ i przy $k = 9$ otrzymujemy $\hat{j}(k) = 7$. Może się zdarzyć, że dla dwóch sąsiednich liczb $j-1$ i j wartość b_k jest jednakowa; wtedy arbitralnie przyjmujemy $\hat{j}(k) = j$ zgodnie ze starą zasadą belferską: w razie wątpliwości podciągaj ocenę w górę. Na przykład przy $k = 16$ wartość prawdopodobieństwa b_{16} jest jednakowa przy $j = 15$ i przy $j = 16$.

Przy obliczaniu $\hat{j}(k)$ dla ustalonego j prawdopodobieństwo b_k przedstawia się jako iloczyn wyrażenia $\binom{n-j}{k-j} 5^j$ i wyrażenia, w którym nie występuje j, po czym poszukuje się maksymalnej wartości pierwszego czynnika.

W tabelce obok podajemy, jakie jest prawdopodobieństwo uzyskania i punktów ($i = 1, \ldots, 10$) przez finalistę rozpoznającego j rysunków (dla kilku różnych wartości j).

liczba punktów i	prawdopodobieństwo uzyskania i punktów przez finalistę rozpoznającego j rysunków			
	$j = 0$	$j = 9$	$j = 14$	$j = 17$
1	0,969	0	0	0
2	0,029	0,086	0	0
3	0,002	0,531	0	0
4	0	0,221	0	0
5	0	0,111	0	0
6	0	0,039	0,262	0
7	0	0,010	0,393	0
8	0	0,002	0,246	0
9	0	0	0,097	0,896
10	0	0	0,002	0,104

Podsumowanie

Badanie zdolności PP jest przykładem badań, w których wyróżnia się pewną klarownie sformułowaną hipotezę (brak zdolności PP). Badacz chce wiedzieć, czy na podstawie wyników doświadczenia hipotezę tę można odrzucić na rzecz ogólnikowo sformułowanej hipotezy alternatywnej (istnienie zdolności PP). Pierwszą hipotezę nazywa się „zerową", gdyż zwykle polega ona na stwierdzeniu, że przedmiot badań jest w pewnym sensie na poziomie zero. Losowy mechanizm przebiegu doświadczenia jest przy hipotezie zerowej znany i na tyle prosty, że można wyznaczyć zbiór wyników doświadczenia (zwany zbiorem krytycznym), którego prawdopodobieństwo w przypadku hipotezy zerowej nie przekracza obranego poziomu istotności α, a przy każdej sytuacji wchodzącej w skład hipotezy alternatywnej jest większe od α. Wynik doświadczenia należący do zbioru krytycznego, nazywany statystycznie istotnym, powoduje odrzucenie hipotezy zerowej na rzecz alternatywnej. Przy wyniku spoza zbioru krytycznego stwierdzamy, że nie ma podstaw do odrzucenia hipotezy zerowej; nie ma też oczywiście podstaw, żeby tę hipotezę uznać za prawdziwą, czyli taki wynik nie wnosi żadnej informacji interesującej badacza.

Oczywiście ten sposób postępowania ma sens tylko wtedy, gdy odrzucenie hipotezy zerowej rzeczywiście interesuje badacza. Tymczasem w praktyce często tak nie jest! Na przykład organizator Olimpiady nie jest zainteresowany odrzuceniem hipotezy o kompletnym braku wiedzy technicznej finalisty Olimpiady.

W wielu niedorzecznych badaniach hipoteza zerowa opisuje pewną wyidealizowaną sytuację, którą można wykluczyć a priori. Badacz chciałby natomiast wiedzieć, czy przedmiot badań jest na poziomie dostatecznie dużo różniącym się od zerowego. Zamiast hipotezy zerowej chciałby zatem testować hipotezę „prawie zerową" lub „rozmytą hipotezę zerową" (hipoteza $j \leqslant j^*$ zamiast hipotezy $j = 0$). Kłopot w tym, że hipotezę „prawie zerową" trudno nieraz formalnie opisać (w sprawdzianie olimpijskim trzeba było przyjąć odpowiednie założenia o sposobie odpowiadania finalisty w zależności od liczby rozpoznawanych rysunków), a potem zwykle trudno skonstruować dla niej zbiór krytyczny.

Jeśli na przykład interesujemy się wpływem pewnego czynnika na kształtowanie się jakiejś cechy obiektów w badanej populacji, to hipotezę zerową o braku wpływu wyraża się formalnie jako równość rozkładów cechy, gdy czynnik działa i gdy nie działa. Taką hipotezę zerową umiemy testować. Jednakże w praktyce idealna równość rozkładów jest a priori wykluczona. Zatem przy teście czułym na każde odchylenie od identyczności rozkładów uzyska się z dużym prawdopodobieństwem wynik statystycznie istotny, który będzie bez znaczenia dla eksperymentatora. Nie jest jednak łatwo wprowadzić formalnie przybliżoną równość rozkładów nie jest łatwo wyznaczyć następnie obszar krytyczny.

Co więcej, eksperymentator dochodzi często do wniosku, że zamiast *testować* hipotezę prawie zerową wolałby *estymować* pewien parametr, który charakteryzuje odstępstwo od hipotezy zerowej (na przykład liczbę rozpoznanych przez finalistę rysunków, stopień zróżnicowania rozkładów badanej cechy itp.). Zatem przed zastosowaniem jakiejś metody statystycznej warto przemyśleć najpierw, o co naprawdę chodzi — bez względu na tradycję i pozorne analogie między różnymi sytuacjami badawczymi.

— Którędy mogę się stąd wydostać — spytała Alicja.
— To zależy w dużej mierze od tego, dokąd chcesz iść — rzekł Kot.
— Właściwie jest mi wszystko jedno ...
— W takim razie możesz obrać dowolną drogę!