Listy Biometryczne - Biometrical Letters
Vol.27 (1990), No.1,2, 47-60

ZASTOSOWANIE WYKRESÓW PROBABILISTYCZNYCH W JEDNOZMIENNEJ ANALIZIE WARIANCJI

Wiesław Wagner

Katedra Metod Matematycznych i Statystycznych
Akademii Rolniczej w Poznaniu

Streszczenie

W pracy przedstawione są rozmaite zastosowania wykresów probabilistycznych w analizie danych doświadczalnych. W szczególności omawiane są metody znajdowania odstających obserwacji w jednozmiennej analizie wariancji przy użyciu wykresów probabilistycznych typu kwantyl-kwantyl i typu wartość oczekiwana-kwantyl. Wykresy pierwszego typu stosowane są do badania kontrastów, średnich kwadratów odchyleń, średnich obiektowych oraz wariancji o jednakowej liczbie stopni swobody, natomiast wykresy drugiego typu do badania średnich kwadratów odchyleń o różnej liczbie stopni swobody.

1. WSTĘP

Analiza wariancji została zaproponowana do wnioskowania wpływu czynników doświadczalnych, wyróżnionych w układzie doświadczalnym na kształtowanie się wartości cechy na jednostkach doświadczalnych. Obejmuje ona testowanie hipotez statystycznych dotyczących efektów głównych i interakcyjnych w modelach stałych, analizę komponentów wariancyjnych w modelach losowych, albo jednocześnie jednych i drugich w modelach mieszanych. W analizie wspomnianych efektów i komponentów wykorzystujemy średnie kwadraty odchyleń, których liczbę określa odpowiedni model matematyczny analizy wariancji. Liczba stopni swobody odpowiadająca średnim kwadratom jest przeważnie różna, chociaż bywa także, (np. w doświadczeniach czynnikowych typu 2^k) jednakowa.

Przy wnioskowaniu statystycznym w jednozmiennej analizie wariancji korzysta się z testu F Fishera-Snedecora. Wymaga on spełnienia założenia o jednorodności wariancji i normalności rozkładu błędów losowych. Jedną z

Słowa kluczowe: jednozmienna analiza wariancji, kontrasty, wykres probabilistyczny typu kwantyl-kwantyl, wykres probabilistyczny typu wartość oczekiwana-kwantyl

przyczyn naruszenia tych założeń przez dane doświadczalne może być wystąpienie obserwacji odstających lub nadmiernie dużych reszt. Istnieje wiele technik wykrywania obserwacji odstających. (np. Wagner i Brzeskwiniewicz, 1986), lecz ich stosowanie w praktyce prowadzi do żmudnych obliczeń. Także metody badania reszt prowadzą do wyznaczania pewnych statystyk o określonych rozkładach (np. test nieaddytywności Tukey'a). Pomocną procedurą w takich wypadkach może być technika graficzna oparta na wykresach probabilistycznych typu: kwantyl-kwantyl (Q-Q) lub wartość oczekiwana-kwantyl (EV-Q).

Wykresy typu Q-Q stosujemy do badania średnich obiektowych wyznaczonych z ortogonalnych układów doświadczalnych, kontrastów, średnich kwadratów odchyleń dla poziomów czynników w doświadczeniach jednoczynnikowych oraz średnich kwadratów odchyleń dla kombinacji poziomów czynników w doświadczeniach wieloczynnikowych lub w doświadczeniach czynnikowych typu m^k ($m \geq 2$) o tej samej liczbie stopni swobody.

Z kolei wykresy typu EV-Q stosujemy do badania średnich kwadratów odchyleń o różnej liczbie stopni swobody.

Istotę stosowania wykresów probabilistycznych, jak i formalnego wnioskowania za ich pomocą o danych zawartych w próbie jest wykreślenie próbkowych statystyk pozycyjnych względem "reprezentatywnych" wartości otrzymanych z przyjętego rozkładu (Wilk i Gdanadesikan, 1968). Proponujemy stosowanie reprezentatywnych wartości jako odpowiednich kwantyli rozkładu (wykresy typu Q-Q) lub jako odpowiednich wartości oczekiwanych statystyk pozycyjnych rozważanego rozkładu (wykresy typu EV-Q).

2. POJĘCIE WYKRESU PROBABILISTYCZNEGO ORAZ JEGO RODZAJE

Niech X i Y oznaczają ciągłe zmienne losowe o rozkładach określonych dystrybuantami $F_X(x)$ i $F_Y(y)$, $x,y \in R$ oraz niech $q_x(\alpha)$ i $q_y(\alpha)$, $0<\alpha<1$, oznaczają kwantyle rzędu α takie, że $F_X(q_x(\alpha)) = F_Y(q_y(\alpha)) = \alpha$.

Dla rozkładów prawdopodobieństwa określonych explicite znaną dystrybuantą, kwantyle $q_x(\alpha)$ oznaczamy wprost z równania $F_X(q_x(\alpha)) = \alpha$, w przeciwnym razie stosowane są specjalne metody numeryczne.

Przykład 1. Niech X będzie zmienną losową o rozkładzie logistycznym określonym dystrybuantą

$$F_X(x) = \left\{1 + \exp(-\frac{x-\mu}{\sigma})\right\}^{-1},$$

przy $x \in R$, $\mu \in R$ i $\sigma > 0$. Kwantyl $q_x(\alpha)$ wyznaczamy z równania $F_X(q_x(\alpha)) = \alpha$. Jest on równy

$$q_x(\alpha) = \mu - \sigma \ln(\frac{1}{\alpha} - 1), \quad 0<\alpha<1.$$

Przykład 2. Niech X będzie zmienną losową o rozkładzie normalnym określonym dystrybuantą

$$F_X(x) = \frac{1}{\sigma\sqrt{2\Pi}} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\} dt \quad ,$$

przy $x \in R$, $\mu \in R$ i $\sigma > 0$. Kwantyl $q_x(\alpha)$ znajdujemy z równania całkowego

$$\frac{1}{\sigma\sqrt{2\Pi}} \int_{-\infty}^{q_x(\alpha)} \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\} dt = \alpha \quad .$$

Rozwiązanie można uzyskać wzorami przybliżonymi. Jedno z takich rozwiązań pochodzi od Odeha i Evansa (1974); podajemy je w dalszej części tego rozdziału.

Istotę wykresu probabilistycznego typu Q-Q podaje poniższy lemat, który ujmuje związek liniowy między kwantylami dwóch zmiennych losowych.

Lemat. Niech X i Y będą dwiema zmiennymi losowymi odpowiednio o ciągłych dystrybuantach $F_X(x)$ i $F_Y(y)$ oraz kwantylach $q_x(\alpha)$ i $q_y(\alpha)$ rzędu α takimi, że $F_X(q_x(\alpha)) = F_Y(q_y(\alpha)) = \alpha$. Jeżeli $Y = aX + b$, $a \neq 0$, to $q_y(\alpha) = a\, q_x(\alpha) + b$.

Dowód lematu można znaleźć w pracy Wagnera (1987).

Przykład 3. Jeżeli $X \sim N(\mu,\sigma^2)$ i $U \sim N(0,1)$ oraz $q_x(\alpha)$ i $q_u(\alpha)$ są kwantylami rzędu α tych rozkładów, to $q_x(\alpha) = \sigma q_u(\alpha) + \mu$, gdyż $X = \sigma U + \mu$.

Rozkłady zmiennych losowych X i Y przeważnie nie są znane. Dysponując próbą $y_1,\ldots,y_n$ (krótko $\{y_j\}$) n niezależnych realizacji zmiennej Y, pytamy czy można je uważać za realizacje zmiennej losowej X. Oznacza to, że poszukujemy odpowiedzi, czy próba $\{y_j\}$ pochodzi z populacji o hipotetycznym rozkładzie P określonym dystrybuantą $F_X(x)$. Odpowiedzi takich dostarcza próbkowy wykres probabilistyczny.

Definicja 1. Próbkowym wykresem probabilistycznym typu Q-Q nazywamy zbiór par $R = \{(\eta_j(\alpha_j), y_{(j)}),\ j = 1,\ldots,n\}$ przedstawionych w układzie współrzędnych prostokątnych, gdzie $\eta_j(\alpha_j)$ są kwantylami rzędu $\alpha_j (0<\alpha_j<1)$ rozkładu P, natomiast $y_{(1)} \leq \ldots \leq y_{(n)}$ są próbkowymi statystykami pozycyjnymi z próby $\{y_j\}$.

Jeżeli próba $\{y_j\}$ pochodzi z populacji o rozkładzie P, to pary punktów ze zbioru R będą układały się wokół pewnej prostej. Wnioskowanie formalne o próbie $\{y_j\}$ sprowadza się do graficznej analizy konfiguracji punktów, głównie ich odstawania od prostej.

Definicja 2. Próbkowym wykresem probabilistycznym typu EV-Q nazywamy zbiór par $R = \{(\Theta_j, y_{(j)}),\ j = 1,\ldots,n\}$ przedstawionych w układzie współrzędnych prostokątnych, gdzie Θ_j są wartościami oczekiwanymi statystyk pozycyjnych z rozkładu P, a $y_{(1)} \leq \ldots \leq y_{(n)}$ są statystykami

pozycyjnymi z próby $\{y_j\}$.

Wnioskowanie na podstawie wykresu typu EV-Q prowadzimy w sposób analogiczny jak na podstawie wykresu Q-Q.

Zależnie od określenia kwantyli wyróżniamy różne rodzaje wykresów probabilistycznych typu Q-Q: półnormalny (PNWP), normalny (NWP), chi-kwadrat (CKWP), beta (BWP), itp. Bliżej omawiamy trzy pierwsze typy wykresów, dla których podajemy wzory umożliwiające wyznaczenie odpowiednich kwantyli:

(a) rozkład normalny (Odeh i Evans, 1974)

$$q_j(\alpha_j) = t_j + \sum_{l=0}^{4} a_{1l} t_j^l \Big/ \sum_{l=0}^{4} a_{2l} t_j^l \quad , \tag{1}$$

gdzie

$$t_j = \begin{cases} (-2\ln\alpha_j)^{1/2} , & 0<\alpha_j<0.5 , \\ (-2\ln(1-\alpha_j))^{1/2} , & 0.5\le\alpha_j<1 , \end{cases}$$

$$\alpha_j = (j-a)/(n-2a+1) \quad , \quad a \in \langle 0,0.5\rangle ,$$

przy $j = 1,\ldots,n$, oraz

$a_{10} = -0.32223243$, $a_{11} = -1.0$, $a_{12} = -0.34224209$,

$a_{13} = -0.02042312$, $a_{14} = -0.45364221\cdot 10^{-4}$,

$a_{20} = 0.08954846$, $a_{21} = 0.58858157$, $a_{22} = 0.53110346$,

$a_{23} = 0.10353775$, $a_{24} = 0.00385607$.

Jeżeli $0.5\le\alpha_j<1$, to zamieniamy $q_j(\alpha_j)$ na $-q_j(\alpha_j)$;

(b) rozkład półnormalny (Gnanadesikan, 1977)

$$\hat{q}_j(\alpha_j) = q_j((1+\alpha_j)/2), \quad j = 1,\ldots,n , \tag{2}$$

gdzie α_j i $q_j(\cdot)$ są określone w (1);

(c) rozkład chi-kwadrat (Goldstein, 1973)

$$\tilde{q}_j(\alpha_j) = p\left\{\sum_{l=0}^{6} (q_j(\alpha_j)/\sqrt{p})^l \cdot w_l(p)\right\}^3 , \tag{3}$$

gdzie p jest liczbą stopni swobody, $q_j(\cdot)$ jest określone w (1), $w_l(p) = e_{0l} + e_{11}/p + e_{21}/p^2$, przy czym wartości współczynników e_{01}, $e_{11}\cdot e_{21}$ dla $l = 1,\ldots,6$, wynoszą

l	e_{01}	e_{11}	e_{21}
0	1.000088600	-0.22373680	-0.015139040
1	0.471394100	0.02607083	-0.008986007
2	0.000134803	0.01128186	0.022776790
3	-0.008553069	-0.01153761	-0.013232930
4	0.003125580	0.00516965	-0.006950356
5	-0.000842681	0.00253001	0.001060438
6	0.000097805	-0.00145012	0.002565326

W szczególności, gdy p=1, kwantyle rozkładu χ^2_1 rzędu α_j-tego wyznaczamy ze wzoru

$$\tilde{q}_j(\alpha_j) = [q_j((1+\alpha_j)/2)]^2 \quad . \tag{4}$$

Jak wspomniano wcześniej, w zagadnieniach analizy wariancji korzystamy również z wykresów probabilistycznych typu EV-Q. W tym wypadku wyznaczmy wartości oczekiwane statystyk pozycyjmych rozkładu *P*. Dla rozkładu N(0,1) obszerne tablice tych wartości zawarte są w zbiorze tablic Pearsona i Hartley'a (1972). Wyznaczanie wartości oczekiwanych statystyk pozycyjnych dla dowolnego rozkładu omawia Downton (1966). Poniżej przedstawimy jedną z metod wyznaczania wartości oczekiwanych dla statystyk pozycyjnych o rozkładach χ^2 z różnymi liczbami stopni swobody.

Niech $SM_1 \leq \ldots \leq SM_k$ będą uporządkowanymi niemalejąco średnimi kwadratami odpowiednio z $f_1,\ldots,f_k$ stopniami swobody, pochodzącymi z analizy wariancji dla pewnego modelu stałego. Te średnie kwadraty przyjmujemy z założenia jako niezależne zmienne losowe o rozkładach $\sigma^2\chi^2_{f_i}/f_i$, gdzie σ^2 jest nieznaną wariancją dla błędu. Oznaczmy przez V_i zmienną losową o rozkładzie $\chi^2_{f_i}/f_i$. Łączna gęstość rozkładu statystyk pozycyjnych $V_1 \leq \ldots \leq V_k$ jest postaci (np. Fisz, 1967)

$$\xi(v_1,\ldots,v_k) = C \prod_{i=1}^{k} \frac{f_i^{t_i} \cdot v_i^{t_i-1}}{2^{t_i}\Gamma(t_i)} \exp(-t_i \cdot v_i) \quad ,$$

gdzie C jest stałą normalizującą, $t_i = f_i/2$, określona na zbiorze $0 < v_1 \leq v_2 \leq \ldots \leq v_k < \infty$. Oznaczmy przez $\xi_i(v_i)$ gęstość brzegową i-tej statystyki pozycyjnej V_i, wówczas wartość oczekiwaną tej zmiennej wyraża całka

$$E(V_i) = \int_0^\infty v_i \xi_i(v_i) dv_i \quad .$$

Jest ona funkcją zadanych stopni swobody $f_1,\ldots,f_k$, dlatego notujemy $E(V_i|f_1,\ldots,f_k)$. Związek bezpośredni między zmiennymi SM_i oraz V_i wskazuje, że

$$E(MS_i|f_1,\ldots,f_k) = \sigma^2 E(V_i|f_1,\ldots,f_k) \equiv \Theta_i .$$

Stałe C i σ^2 wpływają jedynie na nachylenie, a nie na liniową konfigurację punktów na wykresie probabilistycznym typu EV-Q, który stanowi zbiór par $R = \{(e_i, MS_i),\ i = 1,\ldots,k\}$.

Gnanadesikan i Wilk (1970) podali wzory rekurencyjne dla obliczania $e_i(f_1,\ldots,f_k) = B_i^{(k)}(t_1,p_1;\ldots;t_kp_k)$, przy $t_i = f_i/2$, $p_i = t_i - 1$. Są one postaci

$$B_1^{(1)}(t_1,p_1) = (p_1+1)/t_1,$$

$$B_i^{(1)}(t_1,p_1;\ t_2,p_2;\ldots,t_1,p_1) = B_{i-1}^{(1-1)}(t_2,p_2;\ldots;t_1,p_1) - \left(\frac{t_2}{t_1+t_2}\right)^{p_2+1} \prod_{j=0}^{p_1} \alpha_j B_{i-1}^{(1-1)}(t_1+t_2,\ p_2+j;t_3,p_3;\ldots;t_1,p_1) \tag{5}$$

dla $i = 2,\ldots,l$; $l = 1,\ldots,k$, $\alpha_0 = 1$ oraz

$$\alpha_{j+1} = \left(\frac{t_1}{t_1+t_2} \cdot \frac{p_2+j+1}{j+1}\right)\alpha_j, \qquad j = 0,1,\ldots,p_1;$$

$$B_j^{(1)}(t_1,p_1;\ldots;t_{l-1},p_{l-1};\ t_1,p_1) = \left(\frac{t_1-1}{t_{l-1}+t_1}\right)^{p_{l-1}+1} \sum_{j=0}^{p_1} \beta_j \cdot$$

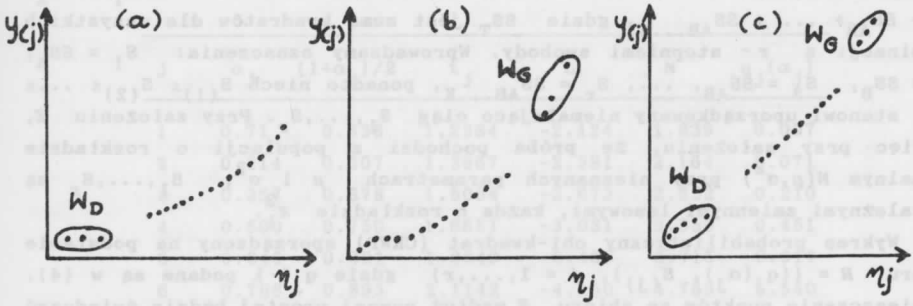
$$B_i^{(l-1)}(t_1,p_1;\ldots;t_{l-2},p_{l-2};t_{l-1}+t_1,p_{l-1}+j) \tag{6}$$

dla $i = 1,\ldots,l-1$; $l = 1,\ldots,k$, $\beta_0 = 1$ oraz

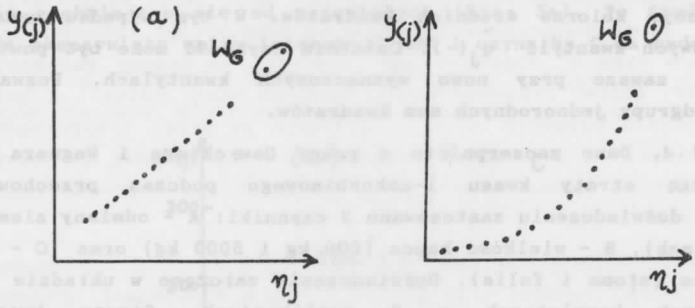
$$\beta_{j+1} = \left(\frac{t_1}{t_{l-1}+t_1} \cdot \frac{p_{l-1}+j+1}{j+1}\right)\beta_j, \qquad j = 0,1,\ldots,p_1.$$

Przy $i=1$ odpowiednim wzorem jest (6), a gdy $i=l$, to wzór (5). Przy $i = 2,3,\ldots,l-1$ można stosować dowolny wzór. Dla nieparzystych stopni swobody stosujemy interpolację. Najpierw obliczamy $B_i^{(k)}$ przy liczbie stopni swobody powiększonej o jeden, a następnie wielkość tę obliczamy przy liczbie stopni swobody pomniejszonej o jeden. Ostateczny wynik stanowi średnia z tych obliczeń.

Omówmy jeszcze ogólne własności wykresów probabilistycznych. Dla uproszczenia bierzemy tylko wykresy z pierwszej ćwiartki układu współrzędnych. Niech W_D i W_G oznaczają zbiory punktów odstających położonych odpowiednio w dolnej i górnej części wykresu (Rys. 1). Zbiory W_D lub W_G mogą w szczególności być puste. Konfiguracja punktów na wykresie może być prostoliniowa lub krzywoliniowa (Rys. 2). Pierwszy wypadek odpowiada zgodności między rozkładem próbkowych statystyk pozycyjnych, a rozkładem hipotetycznym P. Krzywoliniowa struktura odpowiada naruszeniu zgodności między wspomnianymi rozkładami, najczęściej wskazuje na odstępstwo danych od rozkładu normalnego. Obiekty, które występują w zbiorach W_G lub W_D będą wskazywały na rzeczywiste odstępstwo od przyjmowanych założeń. Z kolei obiekty odpowiadające środkowej części wykresu można uważać za nie odchylające się od przyjmowanych założeń.



Rys. 1 Wykres probabilistyczny z punktami odstającymi: (a) dolnymi, (b) górnymi, (c) dwustronnymi



Rys. 2. Wykres probabilistyczny o konfiguracji: (a) prostoliniowej, (b) krzywoliniowej

3. WYBRANE ZASTOSOWANIA WYKRESÓW TYPU Q-Q W ANALIZIE WARIANCJI

3.1. Liczba stopni swobody równa jeden.

Przedstawiamy wykorzystanie wykresów probabilistycznych do badania układu średnich kwadratów pochodzących z doświadczenia czynnikowego typu 2^k, gdzie k jest liczbą czynników doświadczalnych, każdy na dwóch poziomach. Niech A, B,...,K będą k czynnikami oraz niech $A_0, A_1, B_0, B_1, \ldots, K_0, K_1$ stanowią ich poziomy. Między poziomami istnieje 2^k kombinacji, które zapisujemy w porządku Yatesa: $1' = A_0B_0\ldots K_0$, $a = A_1B_0\ldots K_0$, $b = A_0B_1\ldots K_0$, $ab = A_1B_1\ldots K_0, \ldots$, $ab\ldots k = A_1B_1\ldots K_1$. Kombinacje traktowane jako obiekty doświadczalne mogą być rozlosowane w jednym z podstawowych układów doświadczalnych. W tych układach ocenie podlega $\binom{k}{1}$ efektów głównych, $\binom{k}{2}$ efektów interakcji dwuczynnikowej, $\binom{k}{3}$ efektów interakcji trzyczynnikowej, ..., oraz jeden efekt interakcji k-czynnikowej, łącznie $r = 2^k-1$ efektów. Sumy kwadratów dla poszczególnych

efektów obliczone schematem Yatesa mają po 1 stopniu swobody (patrz np. Gawęcki i Wagner, 1984). Przy tym spełniona jest tożsamość $SS_T = SS_A + SS_B + SS_{AB} + \ldots + SS_{AB\ldots K}$, gdzie SS_T jest sumą kwadratów dla wszystkich kombinacji z r stopniami swobody. Wprowadzamy oznaczenia: $S_1 = SS_A$, $S_2 = SS_B$, $S_3 = SS_{AB}$, ..., $S_r = SS_{AB\ldots K}$, ponadto niech $S_{(1)} \leq S_{(2)} \leq \ldots \leq S_{(r)}$ stanowi uporządkowany niemalejąco ciąg $S_1, \ldots, S_r$. Przy założeniu Z, a więc przy założeniu, że próba pochodzi z populacji o rozkładzie normalnym $N(\mu,\sigma^2)$ przy nieznanych parametrach μ i σ^2, $S_1, \ldots, S_r$ są niezależnymi zmiennymi losowymi, każda o rozkładzie χ^2_1.

Wykres probabilistyczny chi-kwadrat (CKWP) sporządzony na podstawie zbioru $R = \{(\tilde{q}_j(\alpha_j), S_{(j)}),\ j = 1,\ldots,r\}$ gdzie $\tilde{q}_j(\cdot)$ podane są w (4). Rozmieszczenie punktów ze zbioru R wzdłuż pewnej prostej będzie świadczyć o braku występowania odchyleń od rozkładu chi-kwadrat rozważanych sum kwadratów S_j. Jeśli tak nie jest, to eliminujemy najbardziej odległe punkty na wykresie probabilistycznym (wyróżniamy zbiór W_G), które odpowiadają największym sumom kwadratów i ponownie sporządzamy nowy wykres na pomniejszonym zbiorze średnich kwadratów. W tym wypadku wymaga to obliczenia nowych kwantyli $\tilde{q}_j(\cdot)$. Ostatnia czynność może być powtórzona wielokrotnie, zawsze przy nowo wyznaczonych kwantylach. Pozwala to wyodrębnić podgrupy jednorodnych sum kwadratów.

Przykład 4. Dane zaczerpnięte z pracy Gawęckiego i Wagnera (1984, s.210) dotyczą straty kwasu l-askorbinowego podczas przechowywania ziemniaków. W doświadczeniu zastosowano 3 czynniki: A - odmiany ziemniaków (Lenino i Flisak), B - wielkość kopca (500 kg i 5000 kg) oraz C - rodzaj pokrycia kopca (słoma i folia). Doświadczenie założono w układzie bloków zrandomizowanych kompletnych w 2 replikacjach. Straty kwasu l-askorbinowego w % będące średnimi z 50 kłębów były następujące:

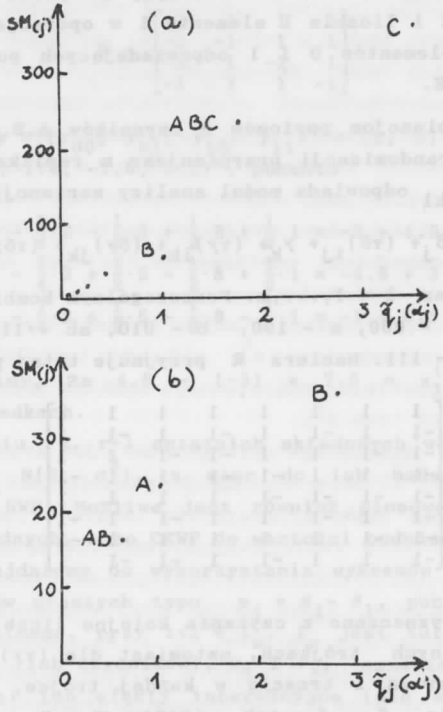
Kombinacja poziomów			Bloki	
Odmiana	Wielkość kopca	Pokrycie kopca	Kopiec przy zabudowaniu	Kopiec na otwartym polu
Lenino	500	słoma	57.2	57.4
Flisak	500	słoma	51.4	50.8
Lenino	5000	słoma	43.7	54.6
Flisak	5000	słoma	56.8	58.4
Lenino	500	folia	41.4	41.0
Flisak	500	folia	47.2	48.8
Lenino	5000	folia	44.1	44.0
Flisak	5000	folia	39.8	39.9

Średnie kwadraty odchyleń dla efektów głównych i interakcyjnych wynoszą: $SS_A = 21.855$, $SS_B = 32.776$, $SS_C = 352.500$, $SS_{AB} = 16.606$, $SS_{AC} = 4.305$, $SS_{BC} = 0.181$ i $SS_{ABC} = 227.256$, każdy z jednym stopniem swobody. Dla formalnej analizy tych średnich kwadratów stosujemy technikę graficzną wykorzystując CKWP. Niezbędne obliczenia kwantyli tego rozkładu

przedstawiamy poniżej (dla obliczenia α_j (j = 1,2,...,7) ze wzoru (1) przyjęto a = 0.5)

j	α_j	$(1+\alpha_j)/2$	t_j	L	M	$\tilde{q}_j(\alpha_j)$
1	0.71	0.536	1.2384	-2.124	1.839	0.007
2	0.214	0.607	1.3667	-2.381	2.164	0.071
3	0.357	0.678	1.5054	-2.673	2.552	0.210
4	0.500	0.750	1.6651	-3.031	3.050	0.451
5	0.643	0.821	1.8549	-3.486	3.715	0.917
6	0.786	0.893	2.1142	-4.160	4.763	1.540
7	0.929	0.964	2.5785	-5.528	7.084	3.233

Przez L i M oznaczono odpowiednio licznik i mianownik wzoru (1). Wykres CKWP zawierający n=7 punktów wskazuje na dwa średnie kwadraty SS_{ABC}, SS_C, wyraźnie odchylające się od pozostałych (Rys. 3a). Te średnie kwadraty wykazują rzeczywisty wpływ interakcji ABC i czynnika C na badaną cechę.



Rys. 3. Wykres probabilistyczny dla: (a) n = 7, (b) n = 5

Po wyeliminowaniu dwóch średnich kwadratów, analizujemy ponownie na wykresie CKWP pozostałe średnie kwadraty. Obliczenia nowych kwantyli podajemy poniżej

j	α_j	$(1+\alpha_j)/2$	t_j	L	M	$\tilde{q}_j(\alpha_j)$
1	0.1	0.55	1.2637	-2.1738	1.9002	0.014
2	0.3	0.65	1.4490	-2.5521	2.3895	0.145
3	0.5	0.75	1.6651	-3.0309	3.0497	0.451
4	0.7	0.85	1.9479	-3.7203	4.0719	1.069
5	0.9	0.95	2.4477	-5.1215	5.3689	2.701

Punkty na wykresie (Rys. 3b) układają się wokół prostej, przyjmujemy więc pięć średnich kwadratów za jednorodne. Oznacza to, że nie można się wypowiedzieć o rzeczywistym wpływie na badaną cechę źródeł zmienności, które są związane ze średnimi kwadratami SS_A, AA_B, SS_{AB}, SS_{AC} i SS_{BC}.

Badanie kontrastów oraz efektów głównych i interakcyjnych w doświadczeniach typu 2^k wygodnie jest rozważać za pomocą macierzy **R** o wymiarach rxr, której pierwszy wiersz zawiera same jedynki, a pozostałe wiersze po połowie elementy -1 i 1. Tworzymy je przez przyporządkowanie liczbie 0 elementu -1 i liczbie 1 elementu 1 w operacjach arytmetycznych modulo 2 dodawania elementów 0 i 1 odpowiadających poziomom kombinacji czynników A,B,C,...,K.

Przykład 5. Kombinacjom poziomów 3 czynników A,B,C rozlosowanych w układzie kompletnej randomizacji przypisujemy m replikacji. Obserwowanym zmiennym losowym y_{ijkl} odpowiada model analizy wariancji postaci

$$y_{ijkl} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\delta\gamma)_{jk} + (\tau\delta\gamma)_{ijk} + e_{ijkl},$$

przy $i,j,k = 0,1$ oraz $l = 1,\ldots,m$. Poszczególnym kombinacjom odpowiadają symbole Yatesa: 1' - 000, a - 100, b - 010, ab - 110, c - 001, ac - 101, bc - 011, abc - 111. Macierz **R** przyjmuje tutaj postać

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix}.$$

Elementy np. dla δ_j wyznaczono z czytania kolejno liczb 0 i 1 na pozycji drugiej w poszczególnych trójkach, natomiast dla $(\tau\gamma)_{ik}$ według zasady: sumujemy pozycje pierwszą i trzecią w każdej trójce, a otrzymane sumy przekształcamy modulo 2. I tak (0+0)mod2 = 0, (1+0)mod2 = 1, (0+0)mod2 = 0, (1+0)mod2 = 1, (0+1)mod2 = 1, (1+1)mod2 = 0,

$(0+1)\mathrm{mod}2 = 1$, $(1+1)\mathrm{mod}2 = 0$. Podobnie wyznaczono elementy w pozostałych wierszach.

Przemnażając macierz $\mathbf{R}$ przez $1/\sqrt{r}$, otrzymujemy macierz ortogonalną $\mathbf{R}_o = (1/\sqrt{r})\mathbf{R}$. Kontrasty z jednym stopniem swobody zawierają składowe $x_2,\ldots,x_r$ wektora $\mathbf{x}_o = \mathbf{R}_o\mathbf{y}$, gdzie $\mathbf{y}$ jest r-wymiarowym wektorem sum poszczególnych kombinacji. Z kolei niech $\mathbf{R} = [\mathbf{R}^{0'}, \mathbf{R}^{1'}]'$, gdzie $\mathbf{R}^0$ jest wektorem odpowiadającym pierwszemu wierszowi macierzy $\mathbf{R}$, a $\mathbf{R}^1$ jest $(r-1)$xr-wymiarową macierzą pozostałych wierszy macierzy $\mathbf{R}$. Tworzymy nową macierz

$$\mathbf{R}_1 = (1/r)[\mathbf{R}^{0'}, 2\mathbf{R}^{1'}] \quad .$$

Wektor $\mathbf{x}_1 = \mathbf{R}_1\mathbf{y}$ zawiera jako pierwszą składową średnią ogólną, a pozostałe jego elementy zawierają oceny efektów głównych i interakcyjnych. W końcu utwórzmy macierz $\mathbf{R}_2 = (1/r)\mathbf{R}\,\mathbf{D}$, gdzie $\mathbf{D} = \mathrm{diag}(1,2,2,\ldots,2)$ oraz wektor $\mathbf{x}_2 = \mathbf{R}_2\mathbf{y}$. Składowe $x_2^{(2)}, x_3^{(2)},\ldots,x_r^{(2)}$ wektora $\mathbf{x}_2$ utworzone są z dwóch zbiorów, każdy liczący po r/2 elementów, i określają specyficzne efekty obiektowe, tak że różnica ich średnich równa się $x_1^{(2)}$.

Przykład 6. Niech A i B będą dwoma czynnikami każdy o dwóch poziomach. Macierz $\mathbf{R}$ o wymiarach 4 x 4 jest tutaj postaci

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{bmatrix} .$$

Przyjmijmy, że $\mathbf{y} = (y_{00}, y_{01}, y_{10}, y_{11})' = (2, 5, 8, 1)'$. Mamy wówczas $\mathbf{x}_2 = \mathbf{R}\mathbf{y}/4 = (7.5, 1.5, -1.5, 5.5)'$, ponadto

$$x_2^{(2)} = -\frac{1}{4}\cdot 2 - \frac{1}{2}\cdot 5 + \frac{1}{2}\cdot 8 + \frac{1}{2}\cdot 1 = -3 + 4.5 = 1.5 ,$$

$$x_3^{(2)} = -\frac{1}{4}\cdot 2 + \frac{1}{2}\cdot 5 - \frac{1}{2}\cdot 8 + \frac{1}{2}\cdot 1 = -4.5 + 3 = -1.5 ,$$

$$x_4^{(2)} = -\frac{1}{4}\cdot 2 + \frac{1}{2}\cdot 5 - \frac{1}{2}\cdot 8 - \frac{1}{2}\cdot 1 = -1 + 6.5 = 5.5 .$$

Z obliczeń widzimy, że $4.5 - (-3) = 7.5 = x_1^{(2)}$. Podobnie jest w pozostałych przypadkach.

Przy założeniu Z, r-1 ostatnich składowych wektorów $\mathbf{x}$, $\mathbf{x}_1$ i $\mathbf{x}_2$ ma rozkład normalny $N(0, \sigma^2)$, a więc do ich badania stosujemy wykres probabilistyczny NWP. Możliwe jest również stosowanie wykresów PNWP do wartości bezwzględnych, albo CKWP do wartości podniesionych do kwadratu.

Obecnie przejdziemy do wykorzystania wykresów probabilistycznych do badania kontrastów prostych typu $\psi_j = \beta_l - \beta_{l'}$, pochodzących z dowolnego układu doświadczalnego, przy $l<l' \in I_w$, I_w jest tutaj zbiorem wskaźników poziomów czynnika (lub czynników), β_l i $\beta_{l'}$, oznaczają efekty główne (dla poziomów czynnika) lub efekty interakcyjne (dla kombinacji czynników), natomiast r jest liczbą badanych kontrastów. Oznaczamy przez $K_1,\ldots,K_r$ oceny kontrastów ψ_j, $j = 1,\ldots,r$ oraz niech ciąg $K_{(1)} \leq \ldots \leq K_{(r)}$ stanowi

uporządkowany niemalejąco ciąg wartości wyznaczonych kontrastów. Do kontrastów K_j stosujemy NWP na podstawie zbioru $R = \{(q_j(\alpha_j), K_{(j)}), j = 1,\ldots,r\}$. Na wykresie NWP mogą wystąpić punkty odchylające się od pewnej prostej po obu stronach (odpowiednio zbiory W_D i W_G), co pozwala wydzielić kontrasty K_j, których oceny są zbyt niskie lub zbyt wysokie. W analogiczny sposób można otrzymać wykres na zredukowanym zbiorze kontrastów, co czynione wielokrotnie pozwoli wyodrębnić podgrupy jednorodne kontrastów.

Można także zastosować wykresy PNWP na wartościach bezwzględnych, albo CKWP na sumach kwadratów K_j^2, $j = 1,\ldots,r$. Sporządzanie tych wykresów i ich interpretacja są takie jak w układzie 2^k.

3.2. Liczba stopni swobody większa niż jeden.

Niech A,B,...,K będą teraz czynnikami doświadczalnymi, odpowiednio o liczbie poziomów $m_1,m_2,\ldots,m_k$. Każda kombinacja $A_{i_1}B_{i_2}\ldots K_{i_k}$, $i_j = 1,\ldots,m_j$ niech zawiera 1 replikacji. Łącznie mamy $m = m_1m_2\ldots m_k$ kombinacji, które można rozlosować w jednym z podstawowych układów doświadczalnych. Oznaczamy przez $S^2_{i_1i_2\ldots i_k}$ wariancję dla kombinacji $(i_1,i_2,\ldots,i_k)$-tej, które niech stanowią ciąg $S_1^2, S_2^2,\ldots,S_m^2$. Przy założeniu Z zmienne S_j^2, $j = 1,\ldots,m$ mają centralne rozkłady χ^2_f z $f = 1-1$ stopniami swobody. Badanie jednorodności wariancji $\{S_j^2\}$ przeprowadzamy na CKWP w oparciu o zbiór $R = \{(q_j(\alpha_j), S^2_{(j)}), j = 1,\ldots,m\}$, gdzie $S^2_{(1)} \le \ldots \le S^2_{(m)}$ są uporządkowanymi niemalejąco wartościami $S_1^2,\ldots,S_m^2$. Punkty odchylające się od pewnej prostej (wystąpienie zbioru W_G) będą wskazywać na niejednorodność wariancji. Wydzielając podgrupę największych wariancji, powtarzamy wykres przy nowo wyznaczonych kwantylach rozkładu χ^2, by wydzielić ewentualnie dalsze podgrupy. Analiza taka pozwala wskazać możliwość wystąpienia obserwacji odstających dla tych kombinacji, które charakteryzują się dużymi wariancjami.

Wykres CKWP można także wykorzystać do badania średnich kwadratów w doświadczeniach s-poziomowych typu s^k, s>2, gdy uwzględni się średnie kwadraty o jednakowej liczbie stopni swobody. Analiza przebiega analogicznie jak w doświadczeniach typu 2^k.

Ponadto wykresy Q-Q można wykorzystać do grupowania średnich obiektowych o jednakowej liczbie replikacji pochodzących z ortogonalnych układów doświadczalnych. Jeżeli wszystkie średnie są dodatnie, to stosujemy PNWP, w przeciwnym razie NWP. Wykres sporządzony w oparciu o zbiór $R = \{(q_j(\alpha_j), \bar{x}_{(j)}), j = 1,\ldots,r\}$, gdzie r oznacza liczbę średnich. Wyodrębnienie grupy średnich, które najbardziej odchylają się od pewnej prostej prowadzi do ustalenia zbiorów W_D i W_G. Ponownie sporządzamy wykres, lecz już na zredukowanym liczebnie zbiorze średnich, wydzielając nową podgrupę. Czynność ta może być powtórzona wielokrotnie.

4. PODSUMOWANIE

Wykresy probabilistyczne można stosować jako narzędzie uzupełniające analizę wariancji. Łatwość obliczenia niezbędnych kwantyli czyni, iż wykresy typu Q-Q nie są trudne do stosowania w praktyce. Pomocna w takich przypadkach mogłaby być grafika komputerowa i odpowiednie dla tego celu sporządzone programy komputerowe. Wykresy probabilistyczne znalazły obok podanych w pracy, również zastosowanie do badania normalności jedno i wielowymiarowej (Wagner, 1987; Domański i Wagner, 1984) a także do badania reszt w modelach ortogonalnych analizy wariancji (np. Daniel, 1959; Wilk i Gnanadesikan, 1968). Przydatność wykresów typu EV-Q uzależnić należy od efektywnego programu komputerowego obliczania wartości oczekiwanych. Liczne praktyczne ilustracje tych wykresów podają Gnanadesikan i Wilk (1970).

LITERATURA

Daniel C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1**, 311-341.

Domański Cz., Wagner W. (1984). Testy wielowymiarowej normalności. *Przegląd Statystyczny*, R. XXXI, z. 3/4, 259-270.

Downton F. (1966). Linear estimates with polynomial coefficients. *Biometrika* **53**, 129-141.

Fisz M. (1967). *Rachunek prawdopodobieństwa i statystyka matematyczna.* PWN, Warszawa.

Gawęcki J., Wagner W. (1984). *Podstawy metodologii badań doświadczalnych w nauce o żywieniu i żywności.* PWN, Warszawa-Poznań.

Gnanadesikan R. (1977). *Methods for statistical data analysis for multivariate observations.* Wiley, New York.

Gnanadesikan R., Wilk M.B. (1970). A probability plotting procedure for general analysis of variance. *J. Roy. Statist. Soc.* **B 32**, 68-101.

Goldstein R.B. (1973). Chi-square quantilies. Algorithm 451, *Comm. ACM* **16**, 483-485.

Odeh R.E., Evans J.O. (1974). The percentage points of the normal distribution. *Appl. Statist.* **23**, 96-97.

Pearson E.S., Hartley H.O. (1972). *Biometrika tables for statisticians* vol. 2, University Press, Cambridge.

Wagner W. (1987). Ocena wielowymiarowej normalności na wykresach probabilistycznych typu kwanty-kwantyl. *Przegląd Statystyczny*, R. XXXIV, z.4, 343-353.

Wagner W., Brzeskwiniewicz H. (1986). Graficzne wykrywanie obserwacji odstających w modelu liniowym. *Szesnaste Colloquium Metodologiczne z Agro-Biometrii.* PAN, Warszawa, 96-107.

Wilk M.B., Gnanadesikan R. (1968). Probability plotting methods for analysis of data. *Biometrika* 55, 1-17.

Praca wpłynęła 1 września 1987;
w wersji ostatecznej 15 maja 1990

APPLICATION OF PROBABILITY PLOTS IN UNIVARIATE ANALYSIS OF VARIANCE

Summary

The paper presents different applications of probability plots in the analysis of experimental data. The methods of detecting outliers in the univariate analysis of variance using the quantile-quantile plots or the expected value-quantile plots are described. Plots of first type are used to investigate contrasts, mean squares, treatment means and variances with equal number of degrees of freedom. Plots of second type are used to investigate mean squares with different numbers of degrees of freedom.

Key words: Univariate analysis of variance, contrasts, quantile-quantile probability plots, expected value-quantile probability plots.