PRZEGLĄD STATYSTYCZNY
R. LX – ZESZYT 3 – 2013

ANDRZEJ SOKOŁOWSKI, SABINA DENKOWSKA,
KAMIL FIJOREK, MARCIN SALAMAGA

ANALIZA MOCY WYBRANYCH TESTÓW JEDNORODNOŚCI CZASÓW TRWANIA DLA POPULACJI O ROZKŁADZIE WEIBULLA[1]

1. WPROWADZENIE

Badanie niektórych zjawisk medycznych, demograficznych, społecznych, gospodarczych czy politycznych wymaga przeprowadzenia tzw. *analizy czasu trwania*. Termin ten w zależności od obszaru zastosowania ma również inne określenia: analiza przeżycia, dożycia (np. w medycynie, demografii, biologii), analiza przejścia (w ekonomii, naukach społecznych), analiza niezawodności (w naukach technicznych) (por. np. Balicki, 2006). Początkowo metody analizy trwania zjawisk znajdowały zastosowanie w statystyce medycznej i demografii, ale z czasem weszły do kanonu narzędzi badawczych innych dyscyplin naukowych. W szczególności w ostatnim czasie obserwuje się wzrost zainteresowania tymi metodami w badaniu zjawisk społeczno-gospodarczych, np. w zagadnieniach dotyczących wchodzenia przedsiębiorstw na rynek, czasu istnienia firm, rynku pracy, rynku nieruchomości, itd.

Dane wykorzystywane w analizie trwania mogą mieć charakter kompletnych bądź niekompletnych danych. Jedną z przyczyn niekompletności danych jest zjawisko cenzurowania jednostek wiążące się z ich eliminacją z pola obserwacji. Przykładowo podczas śledzenia zbioru obiektów mogą pojawić się jednostki, którym będzie towarzyszyć zdarzenie kończące ich obserwację lub też jednostki, w przypadku których takie zdarzenie nie wystąpi do zakończenia procesu obserwacji. W tej drugiej sytuacji jednostki nazywamy cenzurowanymi. Cenzurowanie może się odbywać ze względu na czas (ang. *time censoring*) lub ze względu na zakończenie badania w określonym terminie. Takie cenzurowanie nazywamy cenzurowaniem typu I (por. Balicki, 2006).

Wyróżnia się trzy podstawowe typy cenzurowania danych ze względu na czas (cenzurowanie I typu): cenzurowanie prawostronne, lewostronne oraz obustronne (przedziałowe). W praktyce najczęściej spotykane jest cenzurowanie prawostronne. Może mieć ono miejsce w jednej z dwóch sytuacji (por. Deszczyńska, 2011):

[1] Artykuł zawiera wybrane wyniki uzyskane w ramach Badań Statutowych prowadzonych w Katedrze Statystyki Uniwersytetu Ekonomicznego w Krakowie w 2012 r. (umowa nr 16/KS/6/2012/S/016).

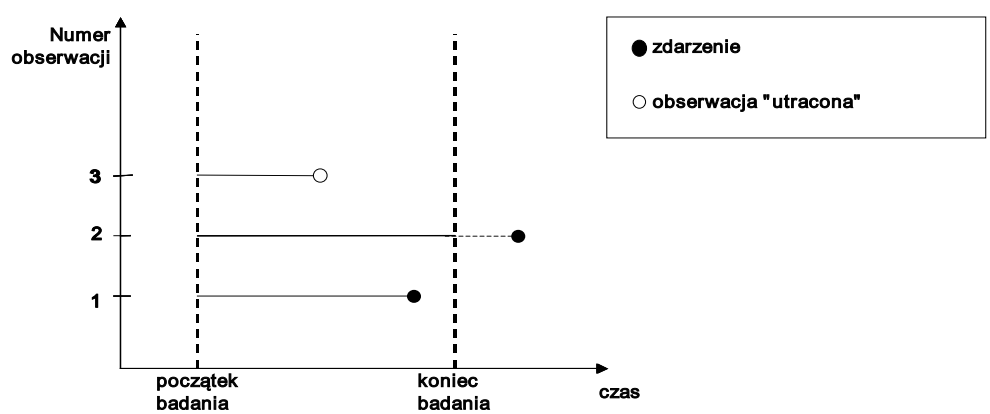
– utraty obserwowanej jednostki na skutek zdarzenia losowego (innego niż oczekiwane zdarzenie), którą przedwcześnie wyeliminowano z badania (np. zgon bezrobotnego szukającego pracy w analizie czasu pozostawania bez pracy, zniszczenie mieszkania przeznaczonego do sprzedaży (w analizie czasu sprzedaży mieszkań na rynku wtórnym) na skutek pożaru,
– zaprzestania procesu obserwacji badanej jednostki z powodu zakończenia badania, przy czym ewentualne zdarzenie związane z tą jednostką w późniejszym czasie nie zostanie już odnotowane (np. zaprzestanie obserwacji bezrobotnego, który do momentu zakończenia badania nie znalazł pracy; zakończenie obserwacji mieszkania, którego nie sprzedano do momentu zakończenia badania).

Zwróćmy uwagę, że w pierwszym przypadku cenzurowanie prawostronne ma charakter *losowy*, a w drugim ma charakter *ustalony*.

Przyjmijmy, że każda losowa jednostka ma swój własny czas trwania T, natomiast czas mierzony do momentu cenzurowania oznaczmy przez C (zakładamy też, że zmienne losowe T oraz C są niezależne dla każdej obserwowanej jednostki). Wówczas w przypadku losowego cenzurowania każdą jednostkę można scharakteryzować wektorem (Y, δ), gdzie:

$$Y = min(T, C),$$
$$\delta = I(T \leq C) = \begin{cases} 1, & gdy \quad T \leq C \\ 0, & gdy \quad T > C. \end{cases} \tag{1}$$

Zatem wskaźnik δ przyjmuje wartość 1, jeśli jednostka nie jest cenzurowana lub wartość 0, gdy jednostka jest cenzurowana.



Rysunek 1. Cenzurowanie prawostronne

Źródło: opracowanie własne na podstawie Balicki (2006).

Przykład cenzurowania prawostronnego przedstawiono na rysunku 1. Obserwację nr 3 należy traktować jako cenzurowaną prawostronnie, gdyż została utracona przed końcem badania, natomiast obserwacja nr 2 jest również cenzurowana prawostronnie, gdyż udało jej się „przetrwać" okres badania bez wystąpienia zdarzenia, które było przedmiotem badania. Jedynie w przypadku jednostki nr 1 zaobserwowano zdarzenie, które było przedmiotem badania w analizie historii zdarzeń i ta obserwacja cenzurowaną nie jest.

Przyjmijmy, że analizowane zdarzenie nastąpiło przed początkowym momentem procesu obserwacji, ale nie wiadomo dokładnie kiedy. Mówimy wówczas o cenzurowaniu lewostronnym. Z takim cenzurowaniem spotykamy się np. przy badaniu zapadalności na pewną chorobę w grupie pacjentów, spośród których niektórzy nie potrafią precyzyjnie określić kiedy w przeszłości wystąpiły u nich pierwsze objawy choroby. W sytuacji, gdy cenzurowanie jednostek następuje równocześnie przed momentem rozpoczęcia procesu obserwacji i po jego zakończeniu, to mówimy o cenzurowaniu obustronnym.

Jak już wspomniano przedmiotem badania w analizie trwania jest czas, jaki upływa od początku procesu obserwacji, aż do momentu zaistnienia zdarzenia kończącego dalszą obserwację jednostki (T), przy czym czas ten jest traktowany jako zmienna losowa o nieujemnych wartościach. Funkcja gęstości prawdopodobieństwa $f(t)$ tej zmiennej losowej pozwala określić rozkład liczby zdarzeń w czasie. Funkcję przeżycia (trwania) w analizie historii zdarzeń definiuje się następująco:

$$S(t) = 1 - F(t), \tag{2}$$

gdzie $F(t)$ oznacza dystrybuantę zmiennej losowej T ($F(t) = P(T < t)$). Funkcja przeżycia wyraża więc prawdopodobieństwo, że czas konieczny do zaistnienia zdarzenia przekroczy wartość t.

Istotnym zagadnieniem w analizie trwania jest testowanie, czy dwie populacje mają te same funkcje przeżycia. W literaturze można spotkać wiele testów do porównywania funkcji przeżycia. Część z nich doczekało się również oprogramowania w komputerowych pakietach statystycznych. Celem niniejszego opracowania jest ocena efektywności testów najczęściej stosowanych w analizie przeżycia. Cel ten osiągnięto poprzez badania symulacyjne. Godzi się zauważyć, że samo zagadnienie badania mocy testu krzywych przeżycia nie jest nowe i było już podejmowane w literaturze przedmiotu (por. Latta 1981; Magel 1991; Suciu i inni 2004; Jurkiewicz i Wycinka, 2011) w odniesieniu do wybranych testów statystycznych. W niniejszym artykule skupiono się natomiast na testach oprogramowanych w pakiecie komputerowym *STATISTICA*, w związku z tym wnioski z badań symulacyjnych mogą się przekładać na rekomendacje dla użytkowników tego pakietu statystycznego.

Badania symulacyjne przeprowadzono w środowisku R oraz programie *STATISTICA* za pomocą skryptu języka Visual Basic. W ramach symulacji generowano próby losowe z rozkładu Weibulla o wyróżnionych wartościach parametrów kształtu

i skali. Dla wybranej konfiguracji parametrów zdefiniowano rozkład referencyjny, z którym porównywane były pozostałe rozkłady. Wybór rozkładu Weibulla wynikał z tego, że jest to rozkład często stosowany w praktyce i jednocześnie dopuszczający szerokie spektrum postaci funkcji przeżycia czy funkcji hazardu. W badaniach symulacyjnych modyfikowano również rozmiary próby oraz odsetek obserwacji cenzurowanych. Za pomocą symulacji badano poziom błędu pierwszego rodzaju oraz moc porównywanych testów statystycznych służących testowaniu hipotezy zerowej głoszącej równość krzywych przeżycia w dwóch grupach.

2. TESTY SŁUŻĄCE DO PORÓWNANIA ROZKŁADÓW CZASU TRWANIA ZJAWISK

Do porównywania czasu trwania zjawisk służą testy istotności różnic pomiędzy dwoma krzywymi przeżycia. Weryfikujemy wówczas hipotezę zerową postaci $H_0 : S_1(t) = S_2(t)$. W zależności od problemu badawczego możemy rozpatrywać jedną z następujących hipotez alternatywnych: $H_1 : S_1(t) \neq S_2(t)$, $H_1 : S_1(t) > S_2(t)$ lub $H_1 : S_1(t) < S_2(t)$.

Testów do porównywania funkcji przeżycia w literaturze można spotkać wiele. Niestety, nie ma jednolitej konwencji w nazywaniu różnych statystyk testowych, a tym samym procedur na nich opartych, dlatego też Blossfeld, Golsch, Rohwer (2007, s. 79) omawiając procedury oprogramowane w pakiecie Stata, podają alternatywne nazwy pod jakimi można je spotkać w literaturze, jak i oprogramowaniu statystycznym. I tak na przykład procedurę log-rank zaimplementowaną w STATA można spotkać[2] w literaturze tematu pod nazwą Generalized Savage Test, a w pakiecie statystycznym BMDP pod nazwą Mantel-Cox log-rank. Z kolei procedura Wilcoxona-Breslowa-Gehana w programie STATA to[3] Generalized Wilcoxon (Breslow) w BMDP, a tylko Wilcoxon w SASie. Zdarza się niestety również tak, że pod tą samą nazwą w różnych pakietach statystycznych zaimplementowane są procedury istotnie różniące się algorytmicznie, a to może skutkować istotnie różnymi wynikami prowadzonych badań. Na przykład najpopularniejszą procedurę log-rank możemy spotkać w różnych wersjach, zarówno w dostępnym oprogramowaniu statystycznym, jak i w literaturze naukowej. Pyke i Thompson (1986) wyróżniają np. trzy testy log-rank: log-rank test wg Peto i Peto (por. Peto, Peto, 1972; Peto, Pike, 1973), test Coxa-Mantela (por. Cox, 1959, 1972; Mantel, 1966), test Mantela–Haenszela (por. Mantel, Heanszel, 1959).

W pakiecie *STATISTICA* dla użytkowników dostępnych jest pięć testów do porównywania dwóch funkcji przeżycia: test Wilcoxona wg Gehana, test Wilcoxona wg Peto i Peto, test log-rank, test F Coxa, test Coxa-Mantela. Biorąc pod uwagę fakt, iż w praktyce badawczej najczęściej wykorzystuje się dostępne, gotowe oprogramowanie, autorzy niniejszego opracowania zdecydowali się przeprowadzić badania symulacyjne, których celem jest porównanie efektywności testów służących do

[2] Ibidem.
[3] Ibidem.

porównywania czasu trwania zjawisk, w wersjach zaimplementowanych w pakiecie *STATISTICA*. Z powodu wspomnianego braku „jednolitej konwencji" w kwestii nazewnictwa testów badania dotyczące efektywności poprzedzone zostały szczegółową prezentacją algorytmów porównywanych testów w wersjach zaimplementowanych w pakiecie *STATISTICA*. Do opisu testów autorzy wykorzystywali zarówno oryginalne artykuły źródłowe w których opisane zostały wykorzystywane procedury, jak i wskazówki Stanisza (2007) dotyczące wersji zaimplementowanych w *STATISTICA* procedur.

Prezentację tych procedur rozpoczniemy od testów będących modyfikacjami nieparametrycznego testu Wilcoxona rangowanych znaków[4], służącego do porównywania rozkładów cech w dwóch populacjach, dla których obserwacje są zestawione w pary. Test Wilcoxona jest alternatywą dla testu par, gdy nie jest spełnione założenie o normalności rozkładów cech w obu populacjach. Gehan (1965) i bracia Peto (1972) zaproponowali modyfikacje testu Wilcoxona pod kątem stosowania go do badań czasu trwania zjawisk, gdy zazwyczaj w badaniu pojawiają się obserwacje cenzurowane (ucięte).

Test Wilcoxona wg Gehana

Na wstępie przyjmijmy za Gehanem (1965) następujące oznaczenia. Niech n_1 oznacza liczbę obserwacji w próbie pierwszej, a n_2 w próbie drugiej, przy czym:

- dla próby pierwszej: $x'_1, \ldots, x'_{s_1}$ to s_1 obserwacji cenzurowanych, a pozostałe $x_{s_1+1}, \ldots, x_{n_1}$ będą $(n_1 - s_1)$ – obserwacjami niecenzurowanymi;
- dla próby drugiej: $y'_1, \ldots, y'_{s_2}$ będą obserwacjami cenzurowanymi, natomiast $y_{s_2+1}, \ldots, y_{n_2}$ niecenzurowanymi w tej grupie.

Gehan (1965) opracował test wykorzystujący statystykę W zdefiniowaną następująco:

$$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij}, \tag{3}$$

gdzie wartość U_{ij} wyznaczana jest ze wzoru:

$$U_{ij} = \begin{cases} -1 & gdy\ x_i < y_j\ lub\ x_i \leq y'_j, \\ 0 & gdy\ x_i = y_j\ lub\ \left(x'_i, y'_j\right)^*\ lub\ x'_i < y_j\ lub\ y'_j < x_i, \\ 1 & gdy\ x_i > y_j\ lub\ x'_i \geq y_j. \end{cases} \tag{4}$$

---

[4] Niestety w polskiej literaturze statystycznej funkcjonuje niewłaściwa nazwa tego testu. *Wilcoxon signed-rank test* powinien być tłumaczony jako *Wilcoxona test znakowanych rang*. Popularna nazwa polska jest niepoprawna, bo znaki są efektem pomiaru w skali nominalnej i w żaden sposób nie można ich porangować (bo to pomiar w skali porządkowej).

* Obie obserwacje są cenzurowane.

Konstrukcja statystyki W nawiązuje (por. Gehan, 1965) do statystyk testowych: Wilcoxona, Manna-Whitney'a oraz do statystyki W Kendalla, w przypadku których nie występują obserwacje cenzurowane ani powiązane.

Gehan wykazał, że przy założeniu prawdziwości hipotezy zerowej, statystka W ma asymptotyczny rozkład normalny z wartością oczekiwaną $E(W) = 0$ i wariancją daną wzorem:

$$D^2(W) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\left(\sum_{i=1}^{s} m_i M_{i-1}\,(M_{i-1}+1) + \sum_{i=1}^{s} l_i M_i\,(M_i+1) + \sum_{i=1}^{s} m_i\,(n_1+n_2-M_i-L_{i-1})(n_1+n_2-3M_{i-1}-m_i-L_{i-1}-1)\right), \quad (5)$$

gdzie $M_j = \sum_{i=1}^{j} m_i,\ M_0 = 0,\ L_j = \sum_{i=1}^{j} l_i,\ L_0 = 0,$

m_i – liczba obserwacji niecenzurowanych w i-tej randze w uporządkowaniu rangowym obserwacji nieuciętych z różnymi wartościami,

l_i – liczba obserwacji prawostronnie cenzurowanych o wartościach większych niż obserwacja o randze i ale mniejszych niż obserwacja o randze $(i + 1)$,

s – łączna liczba obserwacji niecenzurowanych o różnych czasach przeżycia.

Mantel (1967) zaproponował znaczne uproszczenie skomplikowanej obliczeniowo procedury Gehana. W oryginalnej procedurze Gehana każda obserwacja z próby pierwszej jest porównywana ze wszystkimi obserwacjami z drugiej próby. Mantel (1967) zaproponował, by obydwie próby połączyć w jedną o $(n_1 + n_2)$ obserwacjach i każdą obserwację porównywać z pozostałymi $(n_1 + n_2 – 1)$ obserwacjami. Wartości U_i $(i = 1, \ldots, n_1 + n_2)$ wyznaczane są jako różnice pomiędzy liczbą pozostałych spośród $(n_1 + n_2 – 1)$ obserwacji definitywnie mniejszych od i-tej obserwacji, a liczbą tych spośród nich, które są definitywnie większe od tej obserwacji (por. np. Mantel, 1967; Lininger i in., 1979).

Mantel wykazał, że statystykę W Gehana można zapisać jako sumę wartości U_i odpowiadających pierwszej próbie. Wykazał również, że statystyka W ma asymptotyczny rozkład normalny o wartości oczekiwanej $E(W) = 0$ i wariancji:

$$D^2(W) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1+n_2)(n_1+n_2-1)}. \quad (6)$$

Mantel (1967) podał również prosty algorytm wyznaczania wartości U_i.

Algorytm ten polega na uporządkowaniu obu połączonych prób w porządku niemalejącym.

Wyznaczamy wartości R_{1i} przeprowadzając następujące kroki:

1. Przyporządkowujemy rangi obserwacjom od najmniejszej do największej, pomijając obserwacje prawostronnie cenzurowane.
2. Każdej obserwacji prawostronnie cenzurowanej przyporządkowujemy następną wyższą rangę.
3. Niecenzurowanym obserwacjom powiązanym (o tej samej wartości) przyporządkowujemy najmniejszą wartość rangi z przypisanych do tych obserwacji w kroku 1.
4. Redukujemy wartości rang dla obserwacji lewostronnie uciętych do 1.

Następnie wyznaczamy wartości R_{2i}:

1. Przyporządkowujemy rangi obserwacjom od wartości największej do najmniejszej, pomijając obserwacje lewostronnie cenzurowane.
2. Każdej obserwacji lewostronnie cenzurowanej przyporządkowujemy następną wyższą rangę.
3. Niecenzurowanym obserwacjom powiązanym (o tej samej wartości) przyporządkowujemy najmniejszą wartość rangi z przypisanych do tych obserwacji w kroku 1.
4. Redukujemy wartości rang dla obserwacji prawostronnie uciętych do 1.

Następnie wyznaczamy wartości statystyki $U_i = R_{1i} - R_{2i}$ dla $i = 1, \ldots, n_1 + n_2$.

Algorytm ten został zaimplementowany w pakiecie *STATISTICA* w wersji uproszczonej dla cenzurowania prawostronnego (por. Stanisz, 2007).

W pakiecie *STATISTICA* podana jest wartość empiryczna statystyki testowej[5]: $\frac{W-0,5}{D(W)}$ mającej asymptotyczny standaryzowany rozkład normalny (Stanisz, 2007).

Test Wilcoxona wg Peto i Peto

W 1972 roku bracia Peto (1972) zaproponowali modyfikację testu Wilcoxona opartą na ocenach funkcji przeżycia wyznaczonych metodą Kaplana-Meiera dla połączonych prób.

Algorytm (por. Peto, Peto, 1972; Lee, Desu, Gehan, 1975) polega na przyporządkowaniu obserwacjom niecenzurowanym t_i wartości $S(t_i) + S(t_{i-1}) - 1$, natomiast wartościom cenzurowanym prawostronnie t_i wartości $S(t_k) - 1$ wyznaczonej dla największej nieuciętej wartości $t_k \leq t_i$, gdzie S jest estymatorem funkcji przeżycia

[5] We wzorze statystyki testowej uwzględniono dyskretny charakter statystyki W (wprowadzono korektę na ciągłość).

Kaplana-Meiera dla obu połączonych prób. Suma przyporządkowanych w ten sposób wartościdla obu prób wynosi 0.

Algorytm procedury zaimplementowanej w programie *STATISTICA* jest następujący:

1. Wyznaczamy oceny funkcji przeżycia metodą Kaplana-Meiera dla połączonych prób:
   $S(t_i)$ – ocena funkcji przeżycia dla *i*-tej obserwacji niecenzurowanej.
2. Porządkujemy niemalejąco czasy przeżycia obserwacji niecenzurowanych oraz czasy obserwacji jednostek cenzurowanych dla obu prób łącznie. Najmniejszej obserwacji niecenzurowanej przyporządkowujemy odpowiadającą jej ocenę funkcji przeżycia, a następnie poszczególnym czasom przeżycia t_i przyporządkowujemy wartości W_i wyznaczone ze wzoru:

$$W_i = \begin{cases} S(t_i) + S(t_{i-1}) - 1 & \text{dla } obserwacji\ niecenzurowanej\ t_i, \\ S(t_k) - 1 & \text{dla } obserwacji\ cenzurowanej\ t_i,\ dla\ której\ t_k\ jest \\ & największą\ obserwacją\ niecenzurowaną\ taką, że\ t_k \leq t_i. \end{cases} \quad (7)$$

W programie *STATISTICA* powtarzającym się wartościom obserwacji niecenzurowanych ostatecznie przyporządkowywana jest średnia arytmetyczna odpowiadających im wartości W_i.

Przy założeniu prawdziwości hipotezy zerowej, statystyka *WW* będąca sumą wartości W_i dla próby pierwszej, ma rozkład asymptotycznie normalny z wartością oczekiwaną $E(WW) = 0$ i wariancją daną wzorem:

$$D^2(WW) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} W_i^2}{(n_1 + n_2)(n_1 + n_2 - 1)}. \quad (8)$$

W pakiecie *STATISTICA* podawana jest wartość empiryczna statystyki testowej: $\frac{WW}{D(WW)}$ mającej asymptotyczny standaryzowany rozkład normalny.

Test log-rank

Test log-rank został zaproponowany w 1966 roku przez Mantela (1966), ale nazwę tego testu zaproponowali bracia Peto (1972). Konstrukcja tego testu opiera się na logarytmie funkcji przeżycia. W literaturze tematu można spotkać różne wersje i modyfikacje oryginalnego testu log-rank Mantela np. test Mantela-Haenszela log-rank, test Coxa-Mantela logrank czy też test log-rank Peto i Peto. Poniżej przedstawimy wersję testu zaproponowaną przez Mantela (1966) z wykorzystaniem estymatora logarytmu funkcji przeżycia zaproponowanego przez Altshulera (1970). Wersja ta została zaim-

plementowana w programie *STATISTICA* pod nazwą test log-rank, a jej algorytm jest następujący (Lee, Desu Gehan, 1975, Stanisz 2007):

1. Porządkujemy rosnąco czasy przeżycia obserwacji niecenzurowanych oraz czasy obserwacji jednostek cenzurowanych dla obu prób łącznie:

$$t_{(1)} < t_{(2)} < \ldots < t_{(l)},$$

   a następnie przyporządkowujemy czasom przeżycia $t_{(i)}$ odpowiadającą im liczbę zdarzeń końcowych (niepożądanych) $m_{(i)}$.
2. W kolejnym kroku czasom przeżycia $t_{(i)}$ przyporządkowujemy $r_{(i)}$ liczbę obserwacji w zbiorze ryzyka[6] $R(t_{(i)})$, czyli liczbę obserwacji kompletnych oraz cenzurowanych z obu prób o czasie przeżycia nie krótszym niż $t_{(i)}$.
3. Logarytm funkcji przeżycia estymujemy metodą, do której rozpowszechnienia przyczynili się bracia Peto (1972), a zaproponowaną[7] przez Altshulera (1970):

$$e\left(t_{(i)}\right) = \sum_{j \le t_{(i)}} \frac{m_{(j)}}{r_{(j)}}. \tag{9}$$

4. W kolejnym kroku wyznaczane są wartości W_i zgodnie ze wzorem:

$$W_i = \begin{cases} 1 - e(t_{(i)}) & \text{dla obserwacji niecenzurowanej } t_{(i)}, \\ -e(t_{(k)}) & \text{dla obserwacji cenzurowanej } t_{(i)}, \text{ dla której } t_{(k)} \text{ jest} \\ & \text{największą obserwacją niecenzurowaną taką, że } t_{(k)} \le t_{(i)}. \end{cases} \tag{10}$$

   Zauważmy, że większym wartościom obserwacji niecenzurowanych odpowiadają mniejsze wartości W_i, natomiast obserwacje cenzurowane mają przyporządkowane ujemne wartości W_i. Suma W_i dla wszystkich obserwacji wynosi 0.
5. Test log-rank jest oparty na sumie wartości W_i dla jednej z dwóch prób (por. Lee, Desu, Gehan, 1975). W pakiecie *STATISTICA* w podsumowaniu wyników testu log-rank podawana jest suma *WW* wartości W_i dla obserwacji próby pierwszej. Przy założeniu prawdziwości hipotezy zerowej, statystyka *WW* ma rozkład taki jak w teście Wilcoxona wg Peto i Peto, czyli asymptotycznie normalny z wartością oczekiwaną zero i wariancją daną wzorem (8).
6. W pakiecie *STATISTICA* podawana jest wartość empiryczna statystyki: $\frac{WW}{D(WW)}$ mającej asymptotyczny standaryzowany rozkład normalny.

---

[6] Zbiór wszystkich kompletnych oraz cenzurowanych obserwacji o czasach przeżycia nie krótszych od t nazywany jest w literaturze zbiorem ryzyka w chwili t i oznaczany $R(t)$ (patrz np. Lee, Desu, Gehan, 1975).

[7] Metoda zaproponowana przez Altshulera (1970) dla danych prawostronnie cenzurowanych.

Test Coxa-Mantela

W 1972 roku Cox (1972) zaprezentował procedurę porównywania dwóch krzywych przeżycia. W literaturze tematu procedura ta spotykana jest zarówno pod nazwą testu Coxa (por. Desu, Lee, Gehan, 1975), jak i testu Coxa-Mantela. W pakiecie *STATISTICA* procedura Coxa (1972) występuje pod nazwą testu Coxa-Mantela.

Algorytm procedury Coxa-Mantela jest następujący (Cox, 1972, Desu, Lee, Gehan, 1975, Stanisz, 2007):

1. Porządkujemy rosnąco czasy przeżycia (obserwacji kompletnych) dla obu prób łącznie:

$$t_{(1)} < t_{(2)} < \ldots < t_{(k)},$$

a następnie przyporządkowujemy wyróżnionym czasom przeżycia $t_{(i)}$ odpowiadającą im liczbę zdarzeń końcowych (niepożądanych) $m_{(i)}$. Zauważmy, że:

$$\sum_{i=1}^{k} m_{(i)} = n_1+n_2 - (s_1 + s_2), \tag{11}$$

s_i – jest to liczba prawostronnie cenzurowanych obserwacji w i-tej próbie ($i = 1, 2$).

2. Następnie czasom przeżycia $t_{(i)}$ przyporządkowujemy $r_{(i)}$ liczbę obserwacji w zbiorze ryzyka $R(t_{(i)})$, czyli liczbę obserwacji kompletnych oraz cenzurowanych z obu prób o czasie przeżycia nie krótszym niż $t_{(i)}$.
3. Następnie zaobserwowanym czasom przeżycia $t_{(i)}$ przyporządkowujemy $A_{(i)}$ stosunek $r_{2,(i)}$ – liczby obserwacji z drugiej próby o czasie przeżycia nie krótszym niż $t_{(i)}$ oraz $r_{(i)}$:

$$A_{(i)} = \frac{r_{2,(i)}}{r_{(i)}}. \tag{12}$$

4. Wyznaczamy wartości statystyk:

$$U = n_2 - s_2 - \sum_{i=1}^{k} m_{(i)} A_{(i)}, \tag{13}$$

$$I = \sum_{i=1}^{k} \frac{m_{(i)}\left(r_{(i)} - m_{(i)}\right)}{r_{(i)} - 1} A_{(i)}\left(1 - A_{(i)}\right). \tag{14}$$

5. Następnie wyznaczamy wartość statystyki testowej C:

$$C = \frac{U}{\sqrt{I}}\,. \tag{15}$$

Cox (1972) wykazał, że przy założeniu prawdziwości hipotezy zerowej statystyka testowa C ma asymptotyczny rozkład normalny standaryzowany $N(0,1)$.

Test F Coxa

Test F Coxa został opisany przez Coxa (1964). Konstrukcja testu oparta jest na rozkładzie wykładniczym z parametrem $\lambda = 1$. Załóżmy, że obie próby pochodzą z rozkładu wykładniczego ($\lambda = 1$) o liczebnościach odpowiednio n_1, n_2, w tym odpowiednio s_1, s_2 jest liczbą obserwacji cenzurowanych. Testowanie przebiega w następujących etapach:

1. Porządkujemy niemalejąco czasy przeżycia obserwacji niecenzurowanych dla obu prób łącznie.
2. Kolejnym obserwacjom niecenzurowanym przyporządkowujemy wartości:

$$t_{rn} = \frac{1}{n} + \cdots + \frac{1}{n-r+1} \qquad \text{dla } r = 1, \ldots, (n_1 - s_1) + (n_2 - s_2), \tag{16}$$

gdzie $n = n_1 + n_2$.

W przypadku powtarzających się obserwacji niecenzurowanych, odpowiadające im wartości t_{rn} zastępujemy średnią arytmetyczną tych wartości.

Wartościom cenzurowanym przyporządkowujemy wartość:

$$t_{((n_1-s_1)+(n_2-s_2)+1),\, n} = \frac{1}{n} + \cdots + \frac{1}{s_1 + s_2}\,. \tag{17}$$

3. Następnie dla obu prób wyznaczamy średnie wartości $\bar{t}_1, \bar{t}_2$ sumując przyporządkowane w kroku 2 wartości t dla odpowiedniej próby, a następnie dzieląc przez liczbę niecenzurowanych obserwacji w tej próbie, czyli odpowiednio przez $(n_1 - s_1), (n_2 - s_2)$.

Cox (1964) pokazał, że iloraz tak wyznaczonych średnich $\bar{t}_1, \bar{t}_2$, przy założeniu prawdziwości hipotezy zerowej, ma rozkład $\boldsymbol{F}$ o $(2(n_1 - s_1), 2(n_2 - s_2))$ stopniach swobody.

3. WYNIKI EKSPERYMENTÓW SYMULACYJNYCH BADANIA MOCY TESTÓW PORÓWNYWANIA PRZEŻYĆ

Za pomocą symulacji badano poziom błędu pierwszego rodzaju oraz moc testów statystycznych opisanych w rozdziale 2. Symulacje przeprowadzono w środowisku statystycznym R oraz programie *STATISTICA* 10 za pomocą skryptu języka Visual Basic. W ramach symulacji generowano próby losowe z rozkładu Weibulla o parametrze kształtu równym 2 natomiast parametr skali przyjmował wartości 25, 30, 35, 40, 45 oraz 50. Wartość 50 definiowała rozkład referencyjny, czyli rozkład z którym porównywane były pozostałe rozkłady, włącznie z rozkładem referencyjnym. Przyjęte wartości parametru skali skutkowały następującymi wartościami median czasu przeżycia: 20,8; 25; 29,1; 33,3; 37,5; 41,6 miesięcy. Obserwacje cenzurowane generowano zgodnie z metodologią przedstawioną w Halabi i Singh (2004). Kolejnymi dwoma modyfikowanymi parametrami symulacji były rozmiar próby (równy dla obu porównywanych grup) oraz odsetek obserwacji cenzurowanych. Zestawienie scenariuszy symulacyjnych przedstawia tabela 1, natomiast rysunek 2 demonstruje porównywane krzywe przeżycia (krzywa najwyżej położona jest krzywą referencyjną).

Tabela 1.

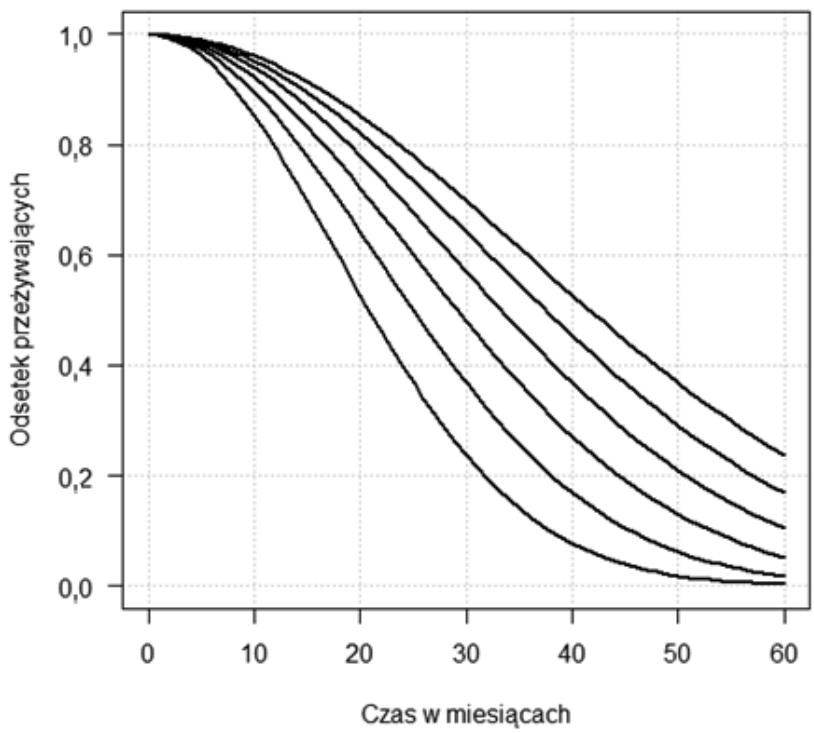
Schematy eksperymentów symulacyjnych

Rozmiary prób	Odsetek obserwacji cenzurowanych w próbie referencyjnej	Odsetek obserwacji cenzurowanych w próbie porównywanej	Parametr skali rozkładu Weibulla	Mediana w rozkładzie Weibulla o zadanej wartości parametru skali
50 100 300	0,3 0,5 0,7 0,3 0,3	0,3 0,5 0,7 0,5 0,7	25	20,8
			30	25,0
			35	29,1
			40	33,3
			45	37,5
			50	41,6

Źródło: ustalenia własne.

W rezultacie wstępnej analizy wyników symulacji okazało się, że wyniki dla testu F Coxa zdecydowanie odbiegają od wyników pozostałych testów. Nawet dla stosunkowo odległych krzywych przeżyć (mediany różniące się o około 21 miesięcy, czyli około dwukrotnie) moc tego testu wyniosła około 0,40, podczas gdy pozostałe testy osiągnęły moc równą 1 (czyli zawsze odrzucały nieprawdziwą hipotezę zerową). Podobne wyniki testu F Coxa, odbiegające od pozostałych, uzyskano również dla innych rozważanych liczebności prób, oraz procentów obserwacji cenzurowanych.

W literaturze tematu test F Coxa jest zalecany (por. np. Lee, Desu, Gehan, 1975[8]; Stanisz, 2007), w sytuacji, gdy próby są małe (mniejsze od 50), a obserwacji cenzurowanych nie ma lub jest ich bardzo niewiele. W przeprowadzonych badaniach symulacyjnych taka sytuacja nie była rozważana. Słabe wyniki testu F Coxa wynikają zapewne również z faktu, iż konstrukcja tego testu oparta jest na rozkładzie wykładniczym (z parametrem $\lambda = 1$), a w badaniu symulacyjnym próby generowane były z rozkładu Weibulla z parametrem kształtu wynoszącym 2 (rozkład Weibulla przechodzi w rozkład wykładniczy dla parametru kształtu równego jeden). Zatem test F Coxa został usunięty z dalszych rozważań.



Rysunek 2. Teoretyczne krzywe przeżyć wykorzystywane w eksperymentach symulacyjnych
Źródło: opracowanie własne.

Po kilku eksperymentach z różnymi funkcjami stwierdzono, że satysfakcjonującą aproksymację funkcji mocy testu daje funkcja logistyczna określona następującym wzorem:

[8] Test F Coxa (1964) występuje w artykule u Lee, Desu, Gehana (1975) pod nazwą testu F.

$$P(d) = \frac{1}{1 + b \cdot \exp(-c \cdot d)}, \tag{18}$$

gdzie d jest odległością pomiędzy medianami porównywanych krzywych przeżyć (miara odstępstwa od prawdziwości hipotezy zerowej). Ponadto zakłada się, że moc testu w warunkach prawdziwości hipotezy zerowej wynosi 0,05, stąd wynika wartość b równa 19. Naturalny w rozważanym zagadnieniu poziom nasycenia funkcji logistycznej wynosi 1 (licznik wyrażenia 18). Teoretyczna funkcja mocy testu była szacowana nieliniowym algorytmem Levenberga-Marquardta z wykorzystaniem kryterium najmniejszych kwadratów.

W związku z przyjętymi założeniami dotyczącymi funkcji aproksymującej (α=0,05 przy prawdziwości hipotezy zerowej oraz poziom nasycenia równy 1), w funkcji (18) pozostaje tylko jeden parametr (c). Im moduł tego parametru jest większy tym wykres funkcji mocy testu położony jest „wyżej". W związku z tym podzielono wartość c dla poszczególnych testów przez wartość c dla testu najlepszego (z największą wartością c). Otrzymujemy w ten sposób pewną miarę względnej mocy testu. Zestawienia relatywnej mocy porównywanych testów przy różnych liczbach obserwacji przedstawiono w tabelach 2–4.

Tabela 2.

Relatywna moc testów porównywania przeżyć dla n=50

Test	Proporcje obserwacji cenzurowanych (w procentach)				
	30/30	50/50	70/70	30/50	30/70
Gehana-Wilcoxona	0,939	0,875	0,903	0,875	0,841
Coxa-Mantela	1,000	1,000	1,000	1,000	1,000
Peto-Peto- Wilcoxona	0,968	0,925	0,903	0,907	0,808
Log-rank	0,977	0,953	0,916	0,891	0,618

Źródło: obliczenia własne.

Tabela 3.

Relatywna moc testów porównywania przeżyć dla n=100

Test	Proporcje obserwacji cenzurowanych (w procentach)				
	30/30	50/50	70/70	30/50	30/70
Gehana-Wilcoxona	0,883	0,889	0,897	0,902	0,880
Coxa-Mantela	1,000	1,000	1,000	1,000	1,000
Peto-Peto- Wilcoxona	0,929	0,936	0,949	0,941	0,880
Log-rank	0,983	0,985	0,932	0,918	0,732

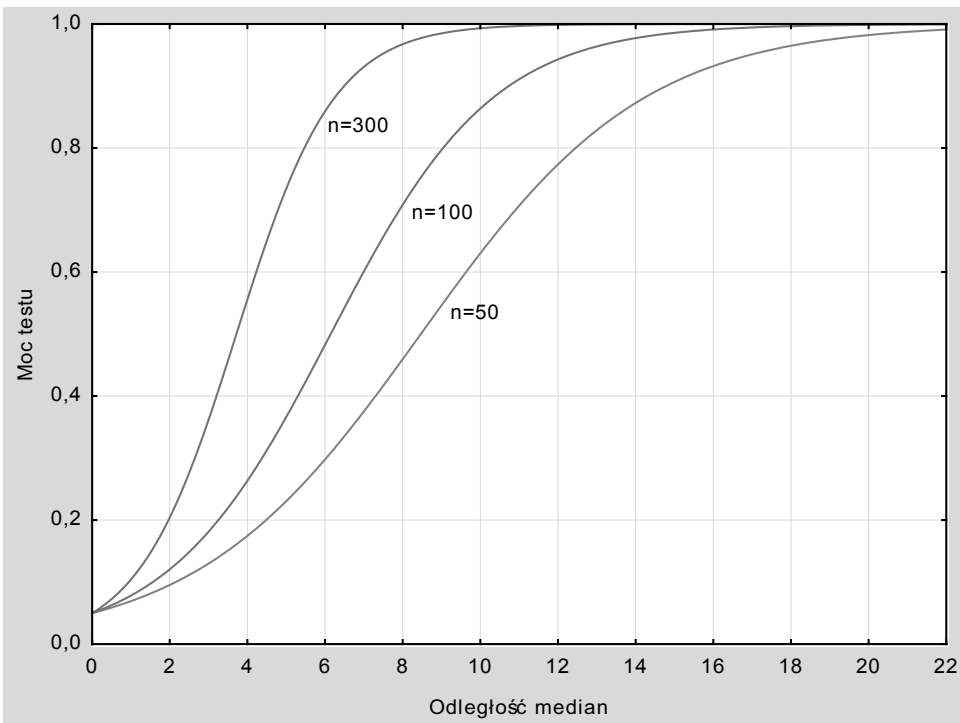
Źródło: obliczenia własne.

Tabela 4.

Relatywna moc testów porównywania przeżyć dla n=300

Test	Proporcje obserwacji cenzurowanych (w procentach)				
	30/30	50/50	70/70	30/50	30/70
Gehana-Wilcoxona	0,889	0,906	0,890	0,868	0,904
Coxa-Mantela	1,000	1,000	1,000	1,000	1,000
Peto-Peto- Wilcoxona	0,927	0,966	0,965	0,912	0,910
Log-rank	0,990	0,994	0,985	0,932	0,808

Źródło: obliczenia własne.



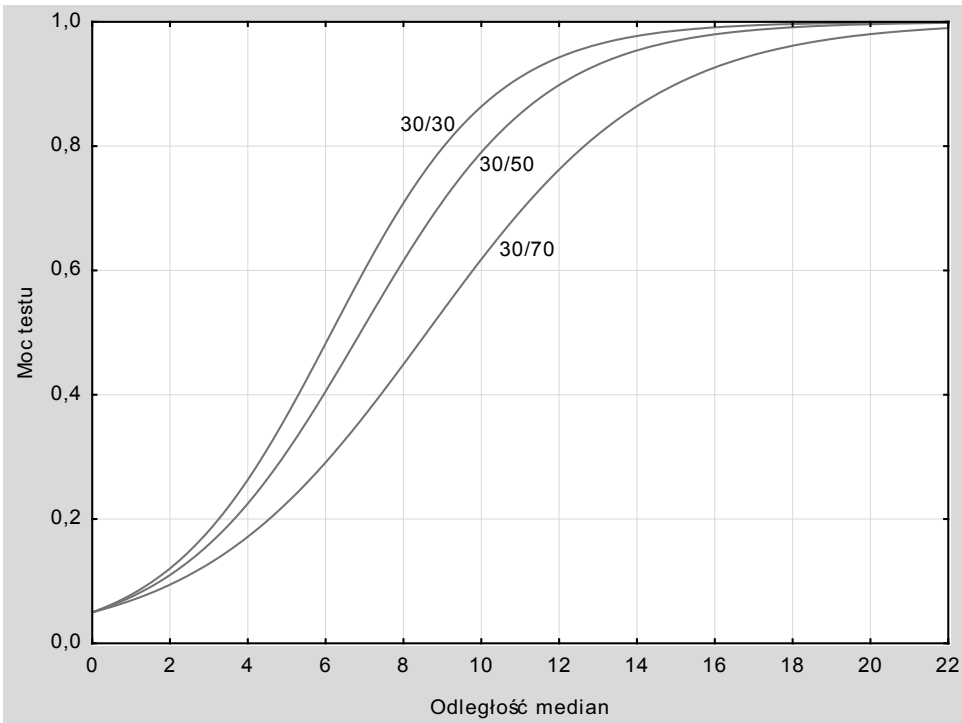
Rysunek 3. Porównanie mocy testu Cox-Mantela w zależności od liczebności próby (cenzurowanie w obu próbach 30%)

Źródło: opracowanie własne.

Wyniki symulacji przynoszą jednoznaczne rozstrzygnięcie. Dla wszystkich schematów eksperymentów symulacyjnych najlepszy okazał się test Coxa-Mantela. Na drugim miejscu plasował się na ogół test long-rank. Różnica mocy między tymi

testami zależała od proporcji obserwacji cenzurowanych. Przy małym ich procencie obydwa testy mają praktycznie porównywalną moc. Test log-rank okazał się bardzo wrażliwy na nierównomierną proporcję cenzurowania, szczególnie przy małej liczbie obserwacji kompletnych w stosunku do rozmiaru próby. W konsekwencji dodatkowe analizy zostaną zaprezentowane tylko dla testu Coxa-Mantela (rysunki 3–5).

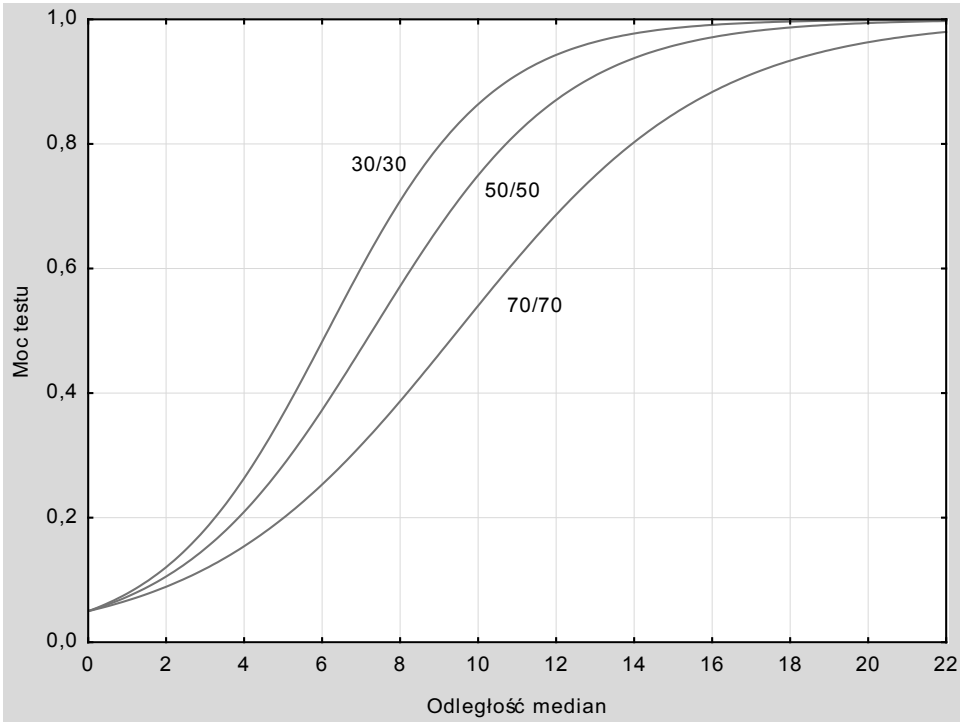
Na rysunku 3 widać oczywistą reakcję testu na zwiększanie liczebności próby – moc testu rośnie w miarę wzrostu liczebności próby. Do dalszej, bardziej szczegółowej prezentacji wyników wybrano liczebność próby równą 100.



Rysunek 4. Moc testu Coxa-Mantela przy rosnącej proporcji obserwacji cenzurowanych w drugiej próbie (C1/C2), n = 100

Źródło: opracowanie własne.

Moc testu Coxa-Mantela jest najwyższa przy wyrównanej proporcji obserwacji cenzurowanych w obu próbach. Jeżeli udział cenzurowanych staje się coraz bardziej różny, to moc testu słabnie. Zwiększanie się udziału obserwacji cenzurowanych pogarsza moc testu również w przypadku równej proporcji takich obserwacji w obydwu próbach (rysunek 5).



Rysunek 5. Moc testu Coxa-Mantela w zależności od proporcji obserwacji cenzurowanych (C1/C2), n=100

Źródło: opracowanie własne.

4. Podsumowanie

Zwiększanie liczebności próby polepsza moc zbadanych testów, co jest elementarnym wymaganiem stawianym każdemu testowi statystycznemu. Zwiększanie udziału obserwacji cenzurowanych w generowanych próbach pogarsza moc testów. W podsumowaniu można stwierdzić, że test Coxa-Mantela okazał się w przeprowadzonych badaniach nieco lepszy od zdecydowanie najpopularniejszego w praktycznych zastosowaniach testu log-rank. Różnice mocy między tymi testami stają się znaczące dopiero przy dużym procencie obserwacji cenzurowanych, szczególnie gdy cenzurowanie nie jest równomierne w porównywanych próbach. Inny ważny wniosek z badań to stwierdzenie, że test F Coxa zaimplementowany w programie *STATISTICA* powinien być stosowany z ostrożnością, gdyż nie jest to test uniwersalny.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Altshuler B., (1970), Theory for the Measurement of Competing Risks in Animal Experiments, *Math. Biosciences*, 6, 1-11.
[2] Balicki A., (2006), *Analiza przeżycia i tablice wymieralności*, PWE, Warszawa.
[3] Blossfeld H-P., Golsch K., Rohwer G., (2007), *Event History Analysis with Stata*, New Jersey: Lawrence Erlbaum Associates.
[4] Cox D. R., (1959), The Analysis of Exponentially Distributed Life-times with Two Types of Failure, *Journal of the Royal Statistical Society*, Series B (Methodological), 21, 411-421.
[5] Cox D. R., (1964), Some Applications of Exponential Ordered Scores, *Journal of the Royal Statistical Society*, Series B (Methodological), 26 (1), 103-110.
[6] Cox D. R., (1972), Regression Models and Life-Tables, *Journal of the Royal Statistical Society*, Series B (Methodological), 34 (2), 187-220.
[7] Deszczyńska A., (2011), *Model hazardów proporcjonalnych Coxa*, *Matematyka stosowana*, tom, 13/54, Instytut Matematyczny PAN, Warszawa.
[8] Gehan, E. A., (1965), A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples, *Biometrika*, 52, 203-223.
[9] Halabi S., Singh B., (2004), Sample Size Determination for Comparing Several Survival Curves with Unequal Allocations, *Statistics in Medicine*, 23, 1793-1815.
[10] Jurkiewicz T., Wycinka E., (2011), Significance Tests of Differences Between Two Crossing Survival Curves for Small Samples, w: Domański Cz., Zielińska-Sitkiewicz K. (red.), *Methodological Aspects of Multivariate Statistical Analysis, Statistical Models and Applications*, Łódź.
[11] Latta R. B., (1981), Monte Carlo Study of Some Two-Sample Rank Tests With Censored Data, *Journal of Statistical Association*, 76 (375), 713-719.
[12] Lee E.T., Desu M. M., Gehan E. A., (1975), A Monte Carlo Study of the Power of Some Two-sample Tests, *Biometrika*, 62 (2), 425-432.
[13] Lininger L., Gail M. H., Green S. B., Byar D. P., (1979), Comparison of Four Tests for Equality of Survival Curves in the Presence of Stratification and Censoring, *Biometrika*, 66 (3), 419-428.
[14] Magel R. C., (1991), Estimating the Power of the Gehan Test, *Biometrical Journal*, 88 (8), 985-997.
[15] Mantel N., (1966), Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration, *Cancer Chemotherapy Reports*, 50, 163-170.
[16] Mantel N., (1967), Ranking Procedure for Arbitrarily Restricted Observation, *Biometrics*, 23 (1), 65-78.
[17] Mantel N., Haenszel W., (1959), Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *Journal of the National Cancer Institute*, 22, 719-748.
[18] Peto R., Peto J., (1972), Asymptotically Efficient Rank Invariant Procedures, *Journal of the Royal Statistical Society*, Series A (General), 135 (2), 185-207.
[19] Peto R., Pike M. C., (1973), Conservatism of the Aproximation in the Logrank Test for Survival Data or Tumor Incidence Data, *Biometrics*, 29 (3), 579-84.
[20] Pyke D. A., Thompson J. N., (1986), Statistical Analysis of Survival and Remotal Rate Experiments, *Ecology*, 67 (1), Ecological Society of America, 240-245.
[21] http://bio.research.ucsc.edu/people/thompson/PublPDFs/025Statistical.pdf
[22] Stanisz A., (2007), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, t. 3, StatSoft.
[23] Suciu G. P., Lemeshow S., Moeschberger M., (2004), *Statistical Tests of the Equality of Survival Curves: Reconsidering the options*, w: Balakrishnan N., Rao C. R. (red.), *Handbook of Statistics 23*, *Advances in Survival Analysis,* Elsevier B. V.

ANALIZA MOCY WYBRANYCH TESTÓW JEDNORODNOŚCI CZASÓW TRWANIA DLA POPULACJI O ROZKŁADZIE WEIBULLA

Streszczenie

W ostatnim latach testy porównywania czasu trwania zjawisk znajdują coraz więcej zastosowań w analizie zagadnień ekonomicznych. Przykładami może być analiza czasu pozostawania na bezrobociu, czasu poszukiwania pracy, czasu istnienia przedsiębiorstwa, itp.

W literaturze można spotkać wiele testów do porównywania funkcji przeżycia. Autorzy niniejszego opracowania zdecydowali się przeprowadzić badania symulacyjne, których celem jest porównanie efektywności najczęściej stosowanych testów służących do porównywania czasu trwania zjawisk, w wersji zaimplementowanej w pakiecie *STATISTICA*. Za pomocą symulacji badano poziom błędu pierwszego rodzaju oraz moc następujących testów statystycznych służących testowaniu hipotezy zerowej głoszącej równość krzywych przeżycia w dwóch populacjach. Analizie poddano następujące testy: Wilcoxona wg Gehana, F Coxa, Coxa-Mantela, Wilcoxona wg Peto i Peto i log-rank. W ramach symulacji generowano próby losowe z rozkładu Weibulla.

Słowa kluczowe: funkcja przeżycia, porównywanie funkcji przeżycia, badania symulacyjne

THE POWER ANALYSIS OF TESTS FOR COMPARING SURVIVAL OF WEIBULL DISTRIBUTED POPULATIONS

Abstract

Recently, tests for comparing survival distributions become more and more popular and used in applied economics. Unemployment duration, time needed to find a new job, enterprise survival or waiting for a commodity to be sold are good examples. There are a number of tests to compare survival distributions proposed in statistical literature. The aim of this research was to analyze, by the means of computer simulations, the effectiveness of survival tests as implemented in *STATISTICA* software. The following tests have been outlined and compared: Wilcoxon, Geehan, Cox-Mantel, Peto & Peto and log rank. Random samples were generated from Weibull distribution.

Keywords: survival function, comparing survival, Monte Carlo research