3. Wyznaczyć wartości:

$$|F(x_i)-F_n(x_i)|, \quad i=1, \ldots, n,$$

gdzie $F(x_i)$ jest wartością dystrybuanty rozkładu normalnego $N(\bar{x}; s)$.

4. Spośród obliczonych modułów różnic wybrać największą:

$$D'_n = \max_{1 \leqslant i \leqslant n} |F(x_i)-F_n(x_i)|. \tag{6.47}$$

5. Porównać wartości D'_n z wartością krytyczną $D'_{n\alpha}$ odczytaną z tablicy 7, zamieszczonej na końcu książki, dla przyjętego poziomu istotności α. Jeżeli $D'_n > D'_{n\alpha}$, to hipotezę H_0 odrzucamy.

6.3.5. Test normalności Shapiro–Wilka

Przyjmijmy założenia o badanej populacji i postaci hipotez H_0 i H_1 takie jak w teście Kołmogorowa–Lillieforsa. W teście Shapiro–Wilka [1965] funkcja testowa określona jest wzorem:

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n,i}(x_{(n-i-1)}-x_{(i)})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{6.48}$$

gdzie $a_{n,i}$ są stałymi zależnymi od (n, i). Zapis $\left[\frac{n}{2}\right]$ oznacza część całkowitą liczby $\frac{n}{2}$.

Wartość współczynników $a_{n,i}$ oraz wartości krytyczne zostały stablicowane przez Shapiro–Wilka [1965] dla $n \leqslant 50$. Dla $n > 50$ współczynniki i wartości krytyczne zostały przedstawione przez Domańskiego, Gadeckiego, Wagnera [1989] (por. tablica 9 zamieszczona na końcu książki).

Hipotezę H_0 o normalności odrzuca się na poziomie istotności α, jeżeli zachodzi $W \leqslant W_\alpha$. Wartości stałych podane zostały w tablicy 8 zamieszczonej na końcu książki.

6.3.6. Test Davida–Hellwiga

Załóżmy, że z populacji badanej ze względu na zmienną X o ciągłej dystrybuancie F została wylosowana próba o elementach $(X_1, X_2, \ldots, X_n)$ i chcemy zweryfikować hipotezę H_0, że próba ta pochodzi z populacji generalnej

o hipotetycznej dystrybuancie F_0, tzn. hipotezę H_0: $F=F_0$, wobec hipotezy alternatywnej H_1: $F \in \mathcal{F}$, gdzie $F_0 \notin \mathcal{F}$. Możemy wykorzystać do tego celu test Davida–Hellwiga.

Sprawdzianem testu jest statystyka postaci:

$$K_n = \text{card}\{j: m_j = 0\}, \tag{6.49}$$

gdzie m_j $(j=1, 2, \ldots, m)$ oznacza liczbę elementów należących do celi, czyli zbioru $M_j=(z_{j-1}; z_j)$, przy czym $P(z_{j-1} < X < z_j) = \frac{1}{m}$ przy założeniu prawdziwości hipotezy H_0, a m jest liczbą cel, czyli zbiorów, na jakie został podzielony zbiór liczb rzeczywistych $\mathcal{R}$. Symbol $\text{card}(C)$ oznacza moc (liczebność) zbioru C.

Statystyka (6.49) testu Davida–Hellwiga ma rozkład nazywany rozkładem pustych cel (por. prace Davida [1950] i Hellwiga [1965]).

Tablice niektórych kwantyli rozkładu pustych cel dla $n=2, 3, \ldots, 100$ zostały podane w tablicy 12 zamieszczonej na końcu książki.

Przedstawimy też modyfikację testu Davida–Hellwiga stosowaną do weryfikacji H_0: $F \in \mathcal{F}$, gdzie $\mathcal{F}$ jest dystrybuantą rozkładu normalnego o nieznanych parametrach.

1. Wyznaczyć przedziały M_j $(j=1, 2, \ldots, n)$ zwane celami, dzieląc przedział $(-\infty, \infty)$ na m części.
2. Uporządkować w kolejności niemalejącej wyniki próby. Niech tworzą one ciąg $x_{(1)}, \ldots, x_{(n)}$.
3. Wyznaczyć wartości $u_{(i)} = (x_{(i)}-\bar{x})/\hat{s}$ dla $i=1, \ldots, n$, gdzie $\bar{x}$ i $\hat{s}$ oznaczają średnią arytmetyczną i odchylenie standardowe dla wyników próby.
4. Odczytać wartość dystrybuanty rozkładu normalnego $N(0; 1)$ dla $u_{(i)}$, czyli $\Phi(u_{(i)})$ dla $i=1, \ldots, n$.
5. Obliczyć $l_i = \text{entier}[m\Phi(u_{(i)})]+1$ dla $i=1, \ldots, n$ oraz $m_j = \text{card}\{i: l_i=j\}$ dla $i=1, \ldots, m$.
6. Wyznaczyć wartość statystyki testu, która przyjmuje postać:

$$K_n^* = \text{card}\{j: m_j = 0\}. \tag{6.50}$$

7. Wyznaczyć z tablicy 12 rozkładu K_n^*, zamieszczonej na końcu książki, wartość krytyczną K_α dla danego poziomu istotności α.
8. Hipotezę H_0 odrzucić, gdy zachodzi nierówność $K_n^* \leqslant K_\alpha$, w przeciwnym przypadku podjąć decyzję: nie ma podstaw do odrzucenia hipotezy H_0.

Jeśli chcemy zweryfikować hipotezę H_0, że rozkład populacji generalnej jest normalny oraz znamy parametry rozkładu, to posłużymy się wartościami krytycznymi z tablicy 12.

Zauważmy, że niektórzy badacze wykorzystując materiał statystyczny, nie zwracają uwagi na formalne założenia dotyczące każdej metody statystycz-

nej. Prześledźmy te niebezpieczeństwa na przykładzie statystyki K_n Davida [1950], która została zdefiniowana przy założeniu, że obserwacje w próbie są generowane niezależnie ze znanego rozkładu. Przedmiotem naszych rozważań jest test Davida–Hellwiga dla złożonej hipotezy o normalności rozkładu. Oznacza to, że nie znamy parametrów rozkładu, więc szacujemy je z próby, a w konsekwencji, zgodnie z procedurą przedstawioną wcześniej, otrzymamy statystkę K_n^*, która oznacza również liczbę pustych cel, ale mającą inny rozkład niż statystyka K_n dana wzorem (6.49). Wniosek ten udowodniony został metodą Monte Carlo w pracy Domańskiego i Tomaszewicza [1989].

Niech $n=m$, tzn. liczebność próby równa jest liczbie cel.

Dla każdego $n=m=5, 6, ..., 100$ zostało wygenerowanych 10 000 n-elementowych prób z rozkładu normalnego. Dla każdej próby została wyznaczona liczba pustych cel. W ten sposób dla każdego n $(n=m)$ otrzymano empiryczny rozkład prawdopodobieństwa:

$$\hat{p}(n, k) \approx P(K_n^* = k), \quad \text{gdzie } k \doteq 0, 1, ..., n-1. \tag{6.51}$$

Na tej podstawie ustosunkowano się do problemu, czy kwantyle statystyki K_n można stosować do tworzenia obszaru krytycznego przy zastosowaniu testu, którego sprawdzianem jest statystyka K_n^*. Analizę przeprowadzono, opierając się na testach zrandomizowanych oraz na kwantylach interpolowanych. Interpolowane (prawostronne) kwantyle statystyki K_n definiuje się następująco:

$$k_i^d(n, \alpha) = k^d(n, \alpha) - p_r^d(n, \alpha), \tag{6.52}$$

gdzie:

$$k^d(n, \alpha) = \min \{h: P(K_n \geqslant h) \leqslant \alpha\} \tag{6.53}$$

jest kwantylem całkowitym, natomiast:

$$p_r^d(n, \alpha) = \frac{\alpha - P(K_n \geqslant k^d(n, \alpha))}{P(K_n = k^d(n, \alpha) - 1)} \tag{6.54}$$

prawdopodobieństwem randomizacyjnym.

Rozmiar zrandomizowanego testu Davida–Hellwiga znajduje się na podstawie kwantyli (6.52):

$$\alpha^d(n, \alpha) = P(K_n^* \geqslant k^d(n, \alpha)) + p_r(n, \alpha) P(K_n^* = k^d(n, \alpha) - 1). \tag{6.55}$$

Dobrą oceną rozmiaru tego testu jest:

$$\hat{\alpha}(n, \alpha) = \sum_{h=k^d(n, \alpha)}^{n-1} \hat{p}(n, h) + \hat{p}_r(n, \alpha) \hat{p}(n, k^d(n, \alpha) - 1), \tag{6.56}$$

gdzie p_r i $\hat{p}_r$ są prawdopodobieństwami randomizacyjnymi odpowiednio dla dokładnego i empirycznego rozkładu.

Na podstawie przeprowadzonych badań okazuje się, że zastosowanie kwantyli rozkładu zmiennej K_n do budowy obszaru krytycznego dla K_n^* jest niezasadne. Błędy są dość duże i dochodzą do 20% dla n bliskiego 100.

Podjęta została zatem próba wyznaczenia obszaru krytycznego testu ze sprawdzianu K_n^* w inny sposób. Empiryczne kwantyle interpolowane z rozkładu (6.51) są następujące:

$$k_i^e(n, \alpha) = k^e(n, \alpha) - p_r^e(n, \alpha), \tag{6.57}$$

gdzie:

$$k^e(n, \alpha) = \min \{h: \sum_{j=h}^{n-1} \hat{p}(n, j) \leqslant \alpha\},$$

$$p_r^e(n, \alpha) = \frac{\alpha - \sum_{j=k^e(n, \alpha)}^{n-1} \hat{p}(n, j)}{\hat{p}(n, k^e(n, \alpha) - 1)}.$$

Ciąg kwantyli rozkładu zmiennej K_n^* jest dość „gładki", wyrównane kwantyle wydają się zatem być bardziej użyteczne niż empiryczne (6.57). Aproksymowane kwantyle interpolowane określone są za pomocą wzoru:

$$k_i^a(n, \alpha) = \gamma_{-2}(\alpha) n^{-2} + \gamma_{-1}(\alpha) n^{-1} + \gamma_0(\alpha) + \gamma_1(\alpha) n + \gamma_2(\alpha) n^2, \tag{6.58}$$

gdzie współczynniki $\gamma_j(\alpha)$ zostały wyznaczone za pomocą metody najmniejszych kwadratów na podstawie (6.57) dla $n=5, 6, ..., 100$.

Współczynniki $\gamma_j(\alpha)$ są podane w tablicy 6.5.

Tablica 6.5
Współczynniki $\gamma_j(\alpha)$

$\gamma_j(\alpha)$	$j=-2$	$j=-1$	$j=0$	$j=1$	$j=2$
$\gamma_j(0{,}01)$	0,11131	−0,15444	1,80138	4,51059	−0,02985
$\gamma_j(0{,}05)$	−0,03547	−0,22875	1,41346	4,19997	−0,01436
$\gamma_j(0{,}10)$	−0,07374	−0,50914	1,34094	4,02583	−0,00727

Wykorzystując wielkości współczynników $\gamma_j(\alpha)$, zbudowane zostały tablice wartości krytycznych dla testu Davida–Hellwiga w przypadku złożonych hipotez normalności (por. tablica 11 zamieszczona na końcu książki). W świetle przeprowadzonych badań dla małych i średnich prób $(n \leqslant 100)$ można sformułować następujące wnioski.

1. W przypadku testowania złożonych hipotez o normalności rozkładu zmiennej losowej nie można stosować kwantyli rozkładu statystyki K_n