7. DOPASOWYWANIE I TRANSFORMACJA ROZKŁADÓW

Dopasowywanie rozkładu jest procedurą poszukiwania rozkładu teoretycznego, który najlepiej odpowiada danym empirycznym. Dobrze zidentyfikowany rozkład teoretyczny pozwala na prawidłowe szacowanie prawdopodobieństw wystąpienia określonych zdarzeń, pozwala więc na podejmowanie właściwych decyzji.

Identyfikacja rozkładu teoretycznego, który najlepiej odpowiada zaobserwowanym danym to:

- wybór typu rozkładu,
- ustalenie wartości parametrów tego rozkładu.

Parametry rozkładu teoretycznego są na ogół szacowane przy pomocy:

- *metody największej wiarygodności*

  Idea metody sprowadza się do spostrzeżenia, że dane z próby są najbardziej prawdopodobne dla parametrów rozkładu odpowiadających parametrom rozkładu populacji generalnej. W celu znalezienia parametrów rozkładu budowana jest *funkcja wiarygodności,* która odpowiada prawdopodobieństwu uzyskania wartości otrzymanych z próby. Ostatecznie, poszukiwane są takie wartości parametrów dla których funkcja wiarygodności osiąga wartość maksymalną.
- *metody momentów*

  Idea metody sprowadza się do założenia, że momenty teoretyczne rozkładu odpowiadają momentom z próby. W celu znalezienia parametrów rozkładu należy znaleźć związki opisujące zależność momentów teoretycznych od parametrów rozkładu. Ostatecznie, parametry rozkładu są znajdowane w wyniku rozwiązania układu równań, w którym momenty teoretyczne zastępowane są momentami z próby.

Zgodność rozkładu empirycznego z wybranym rozkładem teoretycznym może być oceniana poprzez analizę wykresów: histogramów z nałożoną funkcją gęstości rozkładu teoretycznego, wykresów *Q–Q* czy wykresów *P–P* lub poprzez testowanie hipotez o zgodności rozkładów.

W wielu przypadkach znajomość rozkładu danych jest niewystarczająca. Większość analiz statystycznych dotyczy zmiennych o rozkładzie normalnym. W praktyce rozkład zmiennych często odbiega od rozkładu normalnego – w takim przypadku przed wykonaniem odpowiedniej analizy konieczne jest takie przekształcenie zmiennych aby po wykonaniu przekształcenia ich rozkład był bliski normalnemu.

Rzeczywisty rozkład zmiennej decyduje o transformacji, którą należy zastosować, np.:

a) przekształcenie logarytmiczne

wykorzystywane dla zmiennych o rozkładzie zbliżonym do rozkładu logarytmiczno-normalnego (wariancja jest zbliżona do pierwiastka ze średniej),

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \log(X)$$

lub gdy zmienna X przyjmuje wartości < 0

$$Y = \log(X + c), \quad (c \text{ – przesunięcie zmiennej}),$$

b) przekształcenie pierwiastkowe

wykorzystywane dla zmiennych o rozkładzie prawostronnie skośnym, dla których wariancja jest zbliżona do średniej,

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \sqrt{X},$$

lub gdy zmienna X przyjmuje wartości < 0

$$Y = \sqrt{X + c}, \quad (c \text{ – przesunięcie zmiennej}),$$

c) przekształcenie Blissa

wykorzystywane dla zmiennych reprezentujących dane dotyczące proporcji w sytuacji, gdy proporcje te są mniejsze od 0,2 lub większe od 0,8,

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \arcsin\left(\sqrt{X}\right),$$

lub gdy zmienna X przyjmuje wartości < 0

$$Y = \arcsin\left(\sqrt{X + c}\right), \quad (c \text{ – przesunięcie zmiennej}).$$

Przekształcenia *a*) – *c*) są szczególnymi przypadkami przekształceń należącymi do rodziny *przekształceń Boxa – Coxa*.

Przekształcenie Boxa – Coxa

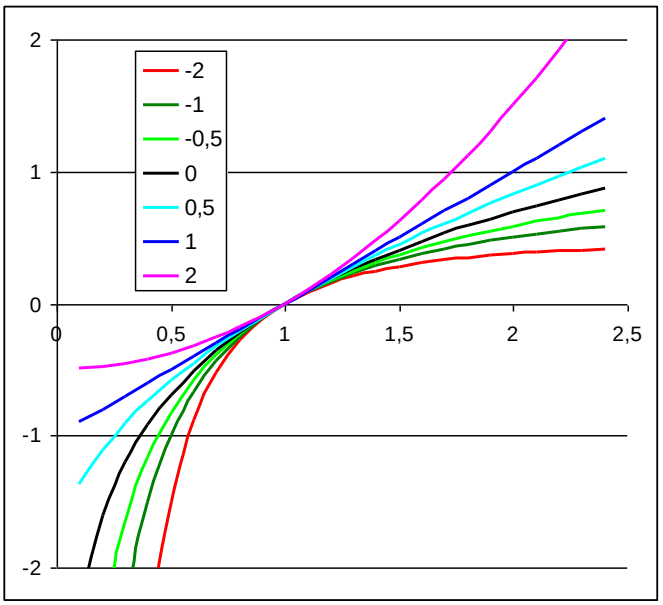
Transformacja przeprowadzana jest zgodnie ze wzorem:

$$Y = \frac{X^{\lambda} - 1}{\lambda},$$

lub gdy zmienna X przyjmuje wartości < 0:

$$Y = \frac{(X + c)^{\lambda} - 1}{\lambda},$$

(c – przesunięcie zmiennej).



Parametr λ decyduje o rodzaju przekształcenia, np.:

- dla $\lambda = 1$ *brak przekształcenia,*
- dla $\lambda = 0{,}5$ *przekształcenie pierwiastkowe,*
- dla $\lambda = 0$ *przekształcenie logarytmiczne,*
- dla $\lambda = -0{,}5$ *odwrotność pierwiastka,*
- dla $\lambda = -1$ *odwrotność.*

Przekształcenie dla $\lambda = 0$ można pokazać wykorzystując *regułę de l'Hospitala*:

$$\lim_{\lambda \to 0} \frac{X^{\lambda} - 1}{\lambda} \overset{H}{=} \lim_{\lambda \to 0} \frac{X^{\lambda} \log(X)}{1} = \log(X).$$

Nieznany parametr λ potrzebny do przekształcenia zmiennej której rozkład odbiega od normalnego może być wyznaczony na kilka sposobów:

- na podstawie wykresu normalności zmiennej *Y* (wykres ten przedstawia zależność kwantyli empirycznych od kwantyli rozkładu normalnego, którego parametry są uzależnione od wartości przekształconej zmiennej *Y*, maksymalizacja współczynnika korelacji liniowej zmiennych przedstawianych na wykresie prowadzi do wyznaczenia optymalnej wartości parametru λ),
- maksymalizując *funkcję największej wiarygodności* (stosując *metodę największej wiarygodności*) postaci:

$$L(\lambda) = -\frac{n}{2} \log(s^2) + (\lambda - 1) \sum_{i=1}^{n} \log(X_i),$$

gdzie: s^2 – odchylenie standardowe zmiennej *Y*, n – rozmiar próby.

Parametr λ nie może być wyznaczony analityczne – konieczne jest wykorzystanie numerycznej metody optymalizacyjnej (można zastosować np. *metodę złotego podziału*, która wykonuje minimalizację funkcji wyznaczając kolejne przybliżenia minimum w zadanym przedziale, na rys. [*a*, *c*]).