Walenty Ostasiewicz

# Rozkłady ucięte

## 1. Wstęp

Pojęcie „rozkład ucięty” ma dwa różne znaczenia. Rozpocznijmy od pierwszego z nich. Rozpatrzmy przypadek nieco uproszczony, ale mający duże zastosowanie w teorii ubezpieczeń.

Załóżmy, że zmienna losowa $X$ określana jest za pomocą dystrybuanty $F_X$. Zdefiniujmy nową „uciętą” zmienną losową $X^R$.

$$X^R = \begin{cases} \text{X, jeśli X} > \text{a} \\ \text{0, w przeciwnym razie} \end{cases}$$

Rozkład zmiennej $X^R$ nazywa się rozkładem uciętym. Dystrybuantę tej zmiennej losowej określa się według wzoru:

$$F_{X^R}(y) = \frac{F_X(y) - F_X(a)}{1 - F_X(a)}. \tag{1}$$

W przypadku tym wszystkie realizacje zmiennej losowej $X$ są obserwowalne. Z punktu widzenia teoretycznego nie ma żadnego problemu z estymacją parametrów obu rozkładów.

Drugie znaczenie pojęcia „rozkład ucięty” dotyczy przypadku, gdy niektóre realizacje nie są obserwowalne i właśnie one nazywane są realizacjami uciętymi.

Jeżeli symbolem $p(x,0)$ oznaczymy funkcję gęstości zmiennej losowej ciągłej lub funkcję rozkładu prawdopodobieństwa zmiennej losowej skokowej i przyjmiemy, że obserwacje zmiennej losowej $X$ należą do pewnego obszaru $T$ będącego podzbiorem przestrzeni prób, to funkcję gęstości zmiennej losowej uciętej $X^T$ określa się ze wzoru (por. [1]):

$$p^T(x,\theta) = \frac{w(x,T) \cdot p(x,\theta)}{u(T,\theta)} \tag{2}$$

gdzie

$$w(x,T) = \begin{cases} 1, \text{ jeśli x} \in T \\ 0, \text{ jeśli x} \in \text{T} \end{cases}$$

$$u(T,\theta) = E(w(X,T))$$

W przypadku tego typu rozkładów jednym z głównych problemów jest estymacja parametrów na podstawie brakujących danych.

Weźmy na przykład taką wielkość losową jak liczba przejazdów bez biletu środkami komunikacji miejskiej przypadająca na jedną osobę. „Obserwacji" w tym przypadku dokonują kontrolerzy. Można ewidencjonować, ile osób zostało raz ukaranych, ile osób zostało dwa razy ukaranych itd., ale nie wiemy, ile osób nie zostało ukaranych, mimo że jechały bez opłaty. Liczba osób, które jeżdżą bez opłaty, nie jest przecież rejestrowana.

Niżej wymienione są inne przykłady tego typu:

— Liczba osób na jedno gospodarstwo domowe, które zachorowały na określoną chorobę zakaźną, nie wiemy zaś, ile osób nie zachorowało.
— Liczba wypadków przy pracy przypadająca na jednego zatrudnionego, w określonej jednostce czasu. Gdyby liczba pracujących była stała w tym czasie, to można by łatwo obliczyć zdarzenie „zero wypadków".
— Częstość występowania słów w tekście określonego autora. Można ustalić, które słowo zostało tylko raz użyte, które dwa razy itd., ale nie można ustalić, które słowo nie zostało ani razu użyte mimo, iż jest ono znane autorowi. Gdyby były znane wszystkie słowa dostępne danemu autorowi, to łatwo by można określić częstość zdarzenia „słowo nieużywane".
— Liczba jaj znoszonych przez pewien gatunek owadów.
— Liczba jednoczesnych odkryć naukowych.

## 2. Rozkład dwumianowy

Zmienną losową $X$ o rozkładzie dwumianowym określa następujący rozkład:

$$p_i = P(X = i) = C_n^i p^i (1-p)^{n-i}, \quad \text{i} = 0, 1, ..., \text{n}$$

Załóżmy, że zdarzenie „zero" nie jest obserwowalne, mimo iż ono się realizuje. Ucięty rozkład dwumianowy, bez zdarzenia „$X = 0$", określa się następująco:

$$p_i^T = P(X^T = i) = \frac{C_n^i p^i (1-p)^{n-i}}{1-(1-p)^n}, \quad \text{i} = 1, 2, ..., \text{n}$$

Zauważmy, że

$$E(X^T) = \frac{n \cdot p}{1-(1-p)^n}$$

oraz

$$E\left(\frac{X^T}{n}\right) = \frac{p}{1-(1-p)^n}.$$

Czyli

$$E(X) < E(X^T)$$

oraz

$$E(X/n) < E(X^T/n)$$

Ponieważ ucięty rozkład dwumianowy zależy od jednego parametru *p*, to problem polega na jego estymacji.

Trudność problemu polega na tym, że nie mając możliwości obserwacji częstości zdarzenia „$X = 0$”, nie znamy też liczby wszystkich obserwacji. Nieznaną liczbę obserwacji oznaczmy symbolem $N$ zaś obserwowane częstości jako $f_1, f_2, ..., f_n$, spełniające następującą równość:

$$f_0 + f_1 + ... + f_n = N.$$

Wielkość N jest nieznana, ponieważ częstość $f_0$ występowania zdarzenia „$X=0$” nie jest znana.

Rozpatrzmy przypadek ogólniejszy, gdy nieobserwowalnych jest *k* kolejnych zdarzeń

„$X = 0$”, „$X = 1$”,..., „$X = k$ -1”.

Gdy ucięte jest jedno zdarzenie, to mamy k = 1, natomiast k = 2 oznacza nieobserwowalność zdarzeń „$X=0$” oraz „$X=1$”.

Wiadomo, że ogólną metodą estymacji parametrów jest metoda największej wiarygodności. W przypadku „normalnego” rozkładu dwumianowego zastosowanie tej metody jest wyjątkowe łatwe i służy ona niemal wzorcowym przykładem w większości podręczników ze statystyki.

W przypadku uciętych rozkładów metoda ta jest o wiele bardziej skomplikowana.

W przypadku rozkładu dwumianowego, gdy uciętych jest k najmniejszych wartości, w celu estymacji parametru *p* metodą największej wiarygodności należy rozwiązać następujące równanie względem niewiadomej *p* (por. [2, 3]):

$$A_k = \frac{np - \sum_{i=0}^{k-1} \frac{n!i}{i!(n-i)!} p^i (1-p)^{n-i}}{1 - \sum_{i=0}^{k-1} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}}$$

gdzie

$A_k$ jest to empiryczny moment rzędu pierwszego obliczany dla danych uciętych.

W przypadku gdy ucięta jest tylko wartość zerowa, równanie to ma następującą postać:

$$\frac{\sum_{i=1}^{n} i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{np}{1-(1-p)^n}.$$

Metoda prostsza rachunkowa jest przedstawiona w pracy [2, 3]. A mianowicie estymator $\hat{p}$ parametru $p$, gdy nie są obserwowalne $k$ pierwsze wartości, dany jest za pomocą wzoru:

$$\hat{p} = \frac{T_2 - k \cdot T_1}{(n-1)T_1 - (k-1)n \cdot T_0}$$

gdzie

$$T_0 = \sum_{i=k}^{n} f_i \quad , \quad T_1 = \sum_{i=k}^{n} i \cdot f_i \quad , \quad T_2 = \sum_{i=k}^{n} i^2 \cdot f_i$$

Nieznaną wielkość $N$ oblicza się wówczas, rozwiązując następujące równanie (por. [2]):

$$N - T_0 = N \sum_{i=0}^{k-1} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

Różnice między estymatorami uzyskanymi za pomocą obu metod są bardzo małe.

Podobną metodę stosuje się do estymacji parametru $\lambda$ w uciętym rozkładzie Poissona.

Gdy uciętych jest $k$ pierwszych wartości, to wzór określający estymator parametru $\lambda$ jest następujący (por. [2]):

$$\hat{\lambda} = \frac{T_2 - k \cdot T_1}{T_1 - (k-1)T_0}$$

gdzie

$$T_0 = \sum_{i=k}^{\infty} f_i\,, \quad T_1 = \sum_{i=k}^{\infty} i \cdot f_i\,, \quad T_2 = \sum_{i=k}^{\infty} i^2 f_i$$

Nieznaną wielkość $N$ określa się, rozwiązując następujące równanie:

$$T_1 - \lambda T_0 = \frac{Ne^{-\lambda}\lambda^k}{(k-1)!}.$$

Badania wykazały, że efektywność tego estymatora jest dość wysoka, przy czym wraz ze wzrostem wartości parametru λ wzrasta do 100% (por. [3]).

## 3. Gdzie umieszczać uciętą masę prawdopodobieństwa?

W celu sprecyzowania pytania postawionego w tytule, rozpatrzmy prosty przykład.

Weźmy zmienną losową $X \sim B(4, \frac{1}{2})$.

Łatwo obliczamy jej rozkład

| i | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $P(X=i)$ | $\frac{1}{16}$ | $\frac{4}{16}$ | $\frac{6}{16}$ | $\frac{4}{16}$ | $\frac{1}{16}$ |

Załóżmy, że dwie pierwsze wartości są ucięte. Obserwujemy więc tylko częstości $f_2 = P(X=2)$, $f_3 = P(X=3)$, $f_4 = P(X=4)$.

Aby określić ucięty rozkład dwumianowy, trzeba określić prawdopodobieństwa:

$$p_2^T P(X=2)\,,\ p_3^T P(X=3)\,,\ p_4^T P(X=4).$$

Zauważmy, że ucięta masa prawdopodobieństwa wynosi

$$u = P(X=0) + P(X=1).$$

W danym przypadku $u = \frac{5}{16}$.

Masą tą trzeba uzupełnić pozostałe wartości $p_2, p_3$ i $p_4$ tak, aby ucięty rozkład był rozkładem, tzn. aby $p_2^T + p_3^T + p_4^T = 1$.

Powstaje więc pytanie: które wartości prawdopodobieństw $p_2, p_3$ i $p_4$, i w jakim stopniu, powinny być zwiększone?

Tradycyjnie, zgodnie ze wzorem (1), uciętą wartość $u$ rozdziela się proporcjonalnie do wielkości $p_2, p_3$ i $p_4$.

W danym przypadku mamy więc następujący rozkład:

$$p_2^T = p_2 \frac{1}{1-u} = p_2 \cdot \frac{16}{11} = \frac{6}{16} \cdot \frac{16}{11} = \frac{6}{11},$$

$$p_3^T = p_3 \frac{1}{1-u} = p_3 \cdot \frac{16}{11} = \frac{4}{16} \cdot \frac{16}{11} = \frac{4}{11},$$

$$p_4^T = p_4 \frac{1}{1-u} = p_4 \cdot \frac{16}{11} = \frac{1}{16} \cdot \frac{16}{11} = \frac{1}{11},$$

Zamiast proporcjonalnie, moglibyśmy na przykład uciętą masę rozmieścić równomiernie:

$$p_2^T = p_2 + \frac{1}{3}u = \frac{6}{16} + \frac{5}{48} = \frac{23}{48}$$

$$p_3^T = p_3 + \frac{1}{3}u = \frac{4}{16} + \frac{5}{48} = \frac{17}{48}$$

$$p_4^T = p_4 + \frac{1}{3}u = \frac{1}{16} + \frac{5}{48} = \frac{8}{48}$$

W. Miszczak proponuje zaś „załamywać” rozkład i po „zawinięciu” części odciętej na część pozostałą dodać odpowiednie masy prawdopodobieństwa, tak jak to niżej pokazano.

$$p_2^T = p_2 + p_1 = \frac{6}{16} + \frac{4}{16} = \frac{10}{16}$$

$$p_3^T = p_3 + p_0 = \frac{4}{16} + \frac{1}{16} = \frac{5}{16}$$

$$p_4^T = p_4 + 0 = \frac{1}{16} + 0 = \frac{1}{16}$$

Problem wyboru jednego z możliwych wariantów, i jego uzasadnienia, nie był prawdopodobnie dotychczas rozpatrywany.

## Literatura

[1] Rao C.R., *Statystyka i prawda*, PWN, Warszawa 1994.
[2] Rider P.R., *Truncated Poisson distributions*, JASA, 48, 1953, 826–830.
[3] Rider P.R., *Truncated Binomial and Negative Binomial Distributions*, JASA, 50, 1955, 877–883.
[4] Ostasiewicz W., *Propedeutyka probabilistyki*, AE, Wrocław 2000.

## Summary

**Truncated Distributions**

The paper highlights two basic definitions of truncated distributions. There are discussed various type of truncation of binomial distribution, as well as are defined some open problems.