

# Rozkłady empiryczne i testowanie hipotez

Michał Urbański

## 1. Cel ćwiczenia

Celem ćwiczenia jest wyznaczenie parametrów rozkładów empirycznych, wyznaczenie rozkładu empirycznego i wykonanie testu hipotezy o zgodności rozkładu empirycznego z hipotetycznym. Założeniem jest to, że wyniki pomiarów są próbą losową pewnego rozkładu, którego właściwości badamy. Rozkładu tego nie znamy, ale możemy sprawdzić z jakim rozkładem dane empiryczne są w najwyższym stopniu zgodne. W modelu probabilistycznym miarą tej zgodności jest prawdopodobieństwo tego, że przy założonej hipotezie dane empiryczne są próbą z badanego rozkładu.

## 2. Wykonanie ćwiczenia

Ćwiczenie polega na wykonaniu następujących eksperymentów, w wyniku których otrzymamy trzy serie danych:

- 1) Pomiar średnicy (lub wymiaru innego przedmiotu) pręta w możliwie wielu miejscach. Średnicę mierzymy suwmiarką i mikrometrem (dwie serie danych jedna suwmiarką druga mikrometrem). Pręt nie jest idealny i w różnych miejscach uzyskuje się różne wskazania.
- 2) Pomiar czasu lotu małego (np. nakrętki od soku) ciała z wysokości 1,5-2m. Aby wyniki były powtarzalne należy zrzucić nakrętki np. z szafy. Czas należy mierzyć stoperem wykorzystując telefon komórkowy.

Każdy rodzaj pomiaru należy powtórzyć przynajmniej 100 razy przy czym każdy powinien wykonać jednakową liczbę pomiarów. Jeśli w pomiarze suwmiarką nie widać zmian (nie widać rozrzutu) pomiar wystarczy wykonać 30 razy.

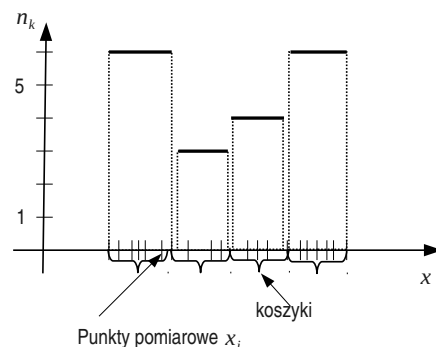
## 3. Opracowanie wyników pomiarów

Dla każdej serii danych należy wykonać następujące działania:

- 1) Wykres histogramu.
- 2) Wykres dystrybuanty empirycznej.
- 3) Wyznaczenie parametrów rozkładów.
- 4) Obliczyć wartość średnią z pomiarów, odchylenie standardowe oraz niepewność złożoną pomiaru czasu i długości (średnicy pręta).
- 5) Weryfikacja hipotezy o tym, że rozkład empiryczny jest jednym z trzech rozkładów prawdopodobieństwa: rozkład równomierny, normalny, Weibulla.
- 6) Porównanie zmierzonych czasu spadania z teorią.

### 3.1. Histogram

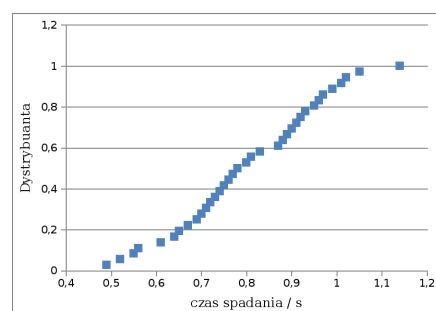
Histogram jest wykresem słupkowym, w którym każdemu przedziałowi wartości zmiennej losowej przyporządkujemy liczbę wystąpień wartości z tego przedziału. Jest to więc wykres częstości występowania zjawiska.



Rysunek 1: Konstrukcja histogramu. Nawiasami klamrowymi zaznaczone są koszyki,  $n_k$  - liczba wyników pomiaru w  $k$ -tym koszyku (binie).

W tym celu porządkujemy dane pomiarowe od najmniejszego do największego:  $x_1, x_2, \dots, x_N$ , gdzie  $N$  jest liczbą danych pomiarowych ( $x_1$  jest wartością najmniejszą a  $x_N$  - największą). Dzielimy przedział  $[x_1, x_N]$  na  $K$  odcinków, każdy o szerokości  $\Delta x = \frac{x_N - x_1}{K}$ . Dla każdego  $k$ -ego przedziału  $\bar{a}_k$  ( $k = 1, \dots, K$ ) o postaci  $\bar{a}_k = [x_1 + (k-1)\Delta x, x_1 + k\Delta x]$  wyliczamy liczbę  $n_k$  elementów  $x_i$  ( $i = 1, \dots, N$ ) ze zbioru danych  $x_1, \dots, x_N$ , które znajdują się w przedziale  $\bar{a}_k$ . Przedział  $\bar{a}_k$  nazywamy koszykiem lub binem. Zależność  $n_k$  od  $k$  jest dyskretna dlatego wykres powinien być słupkowy a nie typu XY dla funkcji ciągłych.

### 3.2. Dystrybuanta



Rysunek 2: Przykład dystrybuanty empirycznej danych pomiaru czasu spadania ciężarka wykonanych stoperem

Estymator dystrybuanty  $\tilde{F}(x)$  ma postać:

$$\tilde{F}(x) = \frac{\#(x_i \leq x)}{N} \quad (1)$$

gdzie  $\#(x_i \leq x)$  jest liczbą elementów  $x_i$  spełniających warunek  $x_i \leq x$ .

Dystrybuanta jest wykresem schodkowym, w którym w każdym punkcie pomiarowym  $x_k$  schodek wynosi  $\frac{1}{N}n_k$ ,

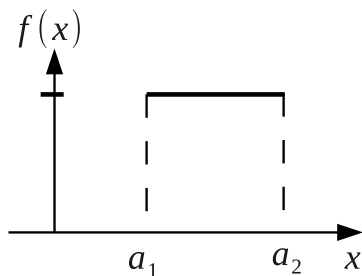
gdzie  $n_k$  jest liczbą powtórzeń wartości  $x_k$  w serii pomiarowej  $\{x_i\}_{i=1}^K$  ( $K$  - liczba danych pomiarowych o wartościach różnych).

Należy wykreślić dystrybuantę empiryczną dla pomiarów średnicy i czasu lotu, oraz na jednym rysunku dystrybuantę rozkładu równomiernego, normalnego i Weibulla o parametrach wyznaczonych metodą opisaną w punkcie następnym.

### 3.3. Wyznaczanie parametrów rozkładów

Sposób wyznaczanie parametrów rozkładu prawdopodobieństwa zależy od rozkładu.

**Jednostajny.** Rozkład jednostajny w przedziale  $[a_1, a_2]$  przedstawiony jest na rys 3.3.



Rysunek 3: Rozkład jednostajny

Rozkład jednostajny wyznaczony jest jednoznacznie przez medianę  $Med = \frac{1}{2}(a_2 + a_1)$  (czyli środek) oraz rozstęp (promień przedziału)  $L = \frac{1}{2}(a_2 - a_1)$ .

Dystrybuanta rozkładu jednostajnego opisana jest linią prostą o równaniu:

$$F(x) = \begin{cases} \frac{x-a_1}{a_2-a_1} & \text{gdy } x \in [a_1, a_2] \\ 0 & \text{gdy } x < a_1 \text{ i } x > a_2 \end{cases} \quad (2)$$

Parametry rozkładu jednostajnego można wyznaczyć na kilka sposobów:

- 1) Na podstawie danych empirycznych należy określić medianę (jako punkt środkowy uporządkowanych danych) i rozstęp równy  $L_{exp} = \frac{1}{2}(x_N - x_1)$ .
- 2) Z danych empirycznych należy wyznaczyć wartość minimalną i maksymalną  $x_N$ , po uporządkowaniu danych są to wartości pierwsza  $x_1$  i ostatnia  $x_N$ .
- 3) Metodą najmniejszych kwadratów należy wyznaczyć parametry prostej empirycznej dystrybuanty opisanej równaniem (??) Przykład podany jest na rys 4). Jeśli posta opisująca dystrybuantę ma postać  $F(x) = ax + b$  to mając postać równania ?? można wyznaczyć parametry  $a$  i  $b$ .

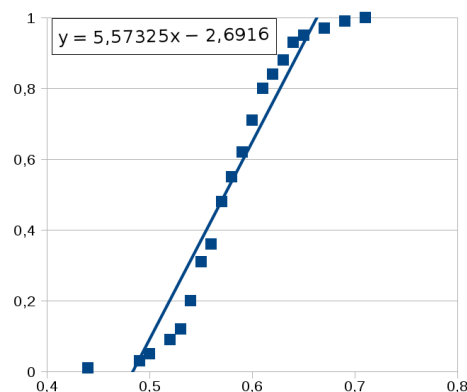
**Normalny.** Rozkład normalny określony jest przez dwa parametry: wartość oczekiwaną  $\mu$  i odchylenie standardowe  $\sigma$ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (3)$$

Parametry te wyznaczyć można z danych pomiarowych jako wartość średnią  $\bar{x}$  (estymator wartości oczekiwanej  $\mu$ ) i estymator odchylenia standardowego  $\sigma$  dany wzorem  $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$ , gdzie  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ .

**Weibulla.** Rozkład Weibulla opisany jest dystrybuantą:

$$F(x) = 1 - e^{-(\frac{x}{\lambda})^k} \quad (4)$$



Rysunek 4: Dystrybuanta empiryczna czasów spadania, podane jest równanie prostej wyznaczonej metodą najmniejszych kwadratów

gdzie  $\lambda$  i  $k$  są parametrami rozkładu. W celu wyznaczenia parametrów rozkładu Weibulla należy wykreślić empiryczną dystrybuantę w następującym układzie współrzędnych: na osi pionowej  $y = \ln(-\ln(1 - F(x)))$ , a na osi poziomej  $z = \ln x$ . Zależność (4) w nowych współrzędnych „ $y$  w funkcji  $z$ ” ma postać funkcji liniowej:  $y = kz - k \ln \lambda$ . Jeśli tą funkcję zapiszemy jako  $y = az + b$ , to współczynniki są równe  $a = k$  i  $b = -k \ln \lambda$ . Wykonać należy taki wykres dla empirycznej dystrybuanty  $\hat{F}$  wg. wzoru (1) i metodą najmniejszych kwadratów wyznaczyć nachylenie  $a$  i współczynnik stały  $b$ . Parametry rozkładu wyliczamy jako:  $k = a$  i  $\lambda = e^{-\frac{b}{a}}$ .

### 3.4. Testowanie hipotezy

W celu sprawdzenia hipotezy i tym, że dane pomiarowe są próbą z rozkładu hipotetycznego należy zastosować testu  $\chi^2$ . W tym celu należy:

(a) wyznaczyć statystykę  $\chi^2$  (zmienną  $\chi_{exp}^2$  dla danych doświadczalnych) dla każdej serii pomiarowej (średnicy i czasu):

$$\chi_{exp}^2 = \sum_{l=1}^L \frac{(n_l - Np_l)^2}{Np_l} \quad (5)$$

gdzie:  $l$  numeruje biny (koszyki),  $L$  - liczba binów (koszyków),  $n_l$  - liczba elementów w koszyku  $l$  (mogą być różne).

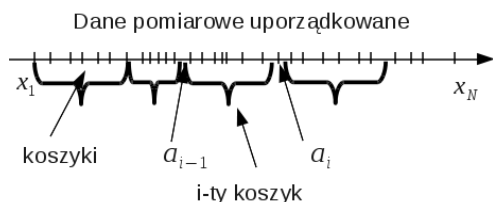
$p_l$  jest prawdopodobieństwem tego, że zmierzona wartość  $x$  jest wewnątrz  $l$ -tego koszyka:

$$\begin{aligned} p_l &= P(a_{l-1} < x < a_l) = \\ &= \int_{a_{l-1}}^{a_l} f(x) dx = F(a_l) - F(a_{l-1}) \end{aligned} \quad (6)$$

gdzie  $a_l$  jest końcem koszyków ( $l = 0, 1, \dots, L$ ),  $f(x)$  jest gęstością rozkładu prawdopodobieństwa (hipoteza), a  $F(x)$  jego dystrybuantą.

**Koszykiem** nazywamy przedział  $[a_l, a_{l+1}]$  o końcach  $a_l$  i  $a_{l+1}$ , w każdym koszyku o numerze  $l$  znajduje się  $n_l$  danych pomiarowych.

Dystrybuantę można wyznaczyć posługując się dowolnym programem typu arkusz kalkulacyjny (mogą być inne



Rysunek 5: Wyznaczanie granic koszyków z serii uporządkowanych danych pomiarowych

programy do analizy danych). Obliczenia należy zrobić dla trzech rozkładów wykorzystując parametry policzone zgodnie z punktem 3.3.

Liczba elementów w każdy koszyku powinna być większa od 10, czyli  $n_l > 10$  dla każdego  $l = 1, \dots, L$ . Trzeba więc dobrać koszyki tak aby każdym było przynajmniej 10 wartości  $x_i$ .

Aby wyznaczyć  $\chi_{exp}^2$  zgodnie ze wzorem (5) należy uporządkować dane pomiarowe  $x_i$  (w arkuszu kalkulacyjnym można porządkować jedynie liczby). Uporządkowany ciąg  $x_i$  danych dzieli się na koszyki po przynajmniej 10 danych w każdym (czyli 10 koszyków po przynajmniej 10 danych), granice koszyków oznaczamy  $a_l$ , gdzie  $l = 0, \dots, L$

(b) Wyliczyć wartość statystyki  $\chi^2$  ze wzoru (5) dla obu serii pomiarowych. Na podstawie tablic wyznaczyć prawdopodobieństwo  $p$ , że zmienna losowa  $\chi^2$  jest większa od wyznaczonego empirycznego  $\chi_{exp}^2$  (liczba stopni swobody jest równa  $L - 2$ ). Porównać wyznaczone prawdopodobieństwa  $p$  dla trzech rozkładów i wyciągnąć wnioski dotyczące podstaw odrzucenia hipotez.

### 3.5. Porównanie zmierzonego czasu spadania z obliczonym z teorii

Jeśli założymy, że pomijany jest opór powietrza i prędkość początkowa jest zerowa to czas spadania opisany jest wzorem  $t_0 = \sqrt{\frac{2H}{g}}$ , gdzie  $H$  - wysokość,  $g = 9,81m.s^2$  - przyspieszenia grawitacyjne.

Należy porównać wynik obliczenia czasu  $t_0$  spadania wyznaczonego ze wzoru powyższego ze średnią  $t_{sr}$  z serii pomiarów czasu spadania. Aby stwierdzić czy różnica pomiędzy  $t_0$  i  $t_{sr}$  mieści się w granicach błędu należy oszacować niepewności  $t_0$  i  $t_{sr}$ .

### 3.6. Analiza niepewności

Na niepewność danych pomiarowych składa się odchylenie standardowe z danych pomiarowych oraz niepewność przyrządu pomiarowego. W przypadku pomiaru czasu spadania o dokładności przyrządu decyduje refleks. W celu wyznaczenia wartości niepewności pochodzącej od refleksu należy zbadać rozkład błędu refleksu. W tym celu należy wykonać eksperyment polegający na pomiarze czasu zdarzenia o określonym czasie trwania. Należy zaprojektować taki eksperyment np. z wykorzystaniem komputera.