Rozkłady empiryczne i wyznaczanie estymatorów wartości oczekiwanej i odchylenia standardowego

Michał Urbański

1. Cel ćwiczenia

Celem ćwiczenia jest wyznaczenie parametrów rozkładów empirycznych, wyznaczenie rozkładu empirycznego (histogramu i dystrybuanty empirycznej), wyznaczanie niepewności gdy mamy serię danych pomiarowych (próbę losową) i porównanie uzyskanych rozkładów z kilkoma rozkładami znanymi w matematyce (rozkład jednostajny, rozkład normalny). W eksperymencie mierzyć będziemy czas spadania niewielkiego przedmiotu oraz średnicę pręta (lub grubość blaszki). Zakładać będziemy, 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. Zgodność rozkładu empirycznego z rozkładem hipotetycznym weryfikuje się metodami statystycznymi (test chi kwadrat) ale w tym ćwiczeniu porównamy jedynie wykresy w sposób jakościowy nanosząc na jednym wykresie rozkład empiryczny i rozkład hipotetyczny.

2. Wykonanie ćwiczenia

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

1) Pomiar średnicy pręta (lub grubości innego przedmiotu) 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 1-2m. Aby wyniki były powtarzalne należy zrzucać nakrętki np. z szafy. Czas należy mierzyć stoperem wykorzystując telefon komórkowy.

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

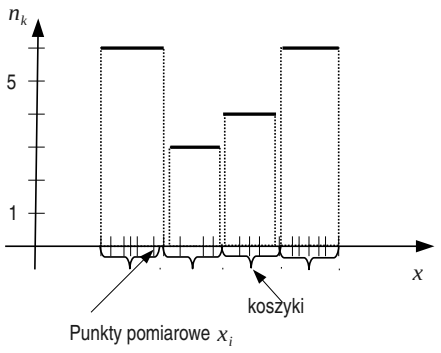
3. Opracowanie wyników pomiarów

Dla każdej serii danych należy wykonać następujące obliczenie i wykresy:

1) Wykres histogramu.
2) Wykres dystrybuanty empirycznej.
3) Wyznaczenie parametrów rozkładów.
4) Obliczenia wartości średniej z pomiarów, odchylenia standardowego oraz niepewności złożonej (uwzględniającej błędy aparaturowe i refleksu) pomiaru czasu i długości (średnicy pręta lub grubości przedmiotu).
5) Porównanie wykresów rozkładu empirycznego z rozkładem równomiernym, normalnym i ewentualnie dowolnym innym ustalonym przez piszącego sprawozdanie.
6) Porównanie zmierzonego czasu spadania z teorią opisującą czas spadania swobodnego w polu grawitacyjnym ziemskim.

3.1. Histogram

Histogram jest wykresem słupkowym, w którym każdemu przedziałowi wartości zmiennej losowej przyporządkujemy liczbę wystąpień wartości zmiennej losowej z tego przedziału. Przedziały nazywamy koszykami lub binami. 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).

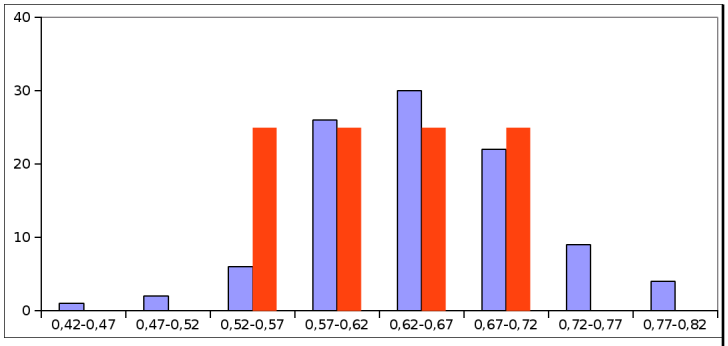
W celu wykonani histogramu porządkujemy dane pomiarowe od najmniejszego do największego: $x_1, x_2, \ldots, x_N$, gdzie N jest liczbą danych pomiarowych (x_1 jest wartością najmniejszą a x_N - największą). Dzielimy przedział $[x_1, x_N]$ na K odcinków, każdy o szerokości $\Delta x = \frac{x_N - x_1}{K}$. Dla każdego k-ego przedziału $\bar{a}_k$ ($k = 1, \ldots, K$) o postaci $\bar{a}_k = [x_1 + (k-1)\Delta x, x_1 + k\Delta x]$ wyliczamy liczbę n_k elementów x_i ($i = 1, \ldots, N$) ze zbioru danych $x_1, \ldots, x_N$, 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 powinie być słupkowy a nie typu XY dla funkcji ciągłych.

Na wykresie 2 pokazano histogram dla przykładowych danych i pokazano porównanie z histogramem dla rozkładu jednostajnego wyznaczonego na rys. 4.

3.2. Dystrybuanta

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

$$\tilde{F}(x) = \frac{\sharp(x_i \leq x)}{N} \tag{1}$$



Rysunek 2: Histogram czasów spadania dla danych jak na rys 4. Niebieskie słupki- histogram empiryczny, czerwone - histogram wyznaczony z równania na wykresie 4. Na osi poziomej podano przedziały czasów w sekundach.

gdzie $\sharp(x_i \leq x)$ 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).

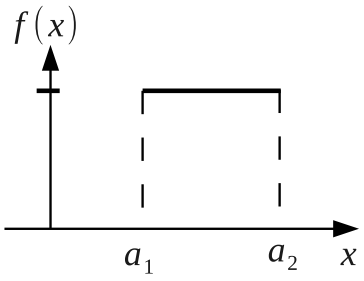
W sprawozdaniu należy wykreślić dystrybuantę empiryczną dla pomiarów średnicy i czasu lotu, oraz na jednym rysunku dystrybuantę rozkładu równomiernego, normalnego i ewentualnie innego rozkładu, o parametrach wyznaczonych metodą opisaną w punkcie następnym. Na rys. 4 pokazano przykład empirycznej dystrybuanty dla danych takich ja na wykresie 2 i pokazano wykres prostej będącej dystrybuantą rozkładu jednostajnego najlepiej pasującego do danych pomiarowych.

3.3. Wyznaczanie parametrów rozkładów

Sposób wyznaczanie parametrów rozkładu prawdopodobieństwa zależy od rozkładu.

3.3.1. Rozkład jednostajny.

Rozkład jednostajny w przedziale $[a_1, a_2]$ przedstawiony jest na rys 3.3.1.



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) $L = \frac{1}{2}(a_2 - a_1)$.

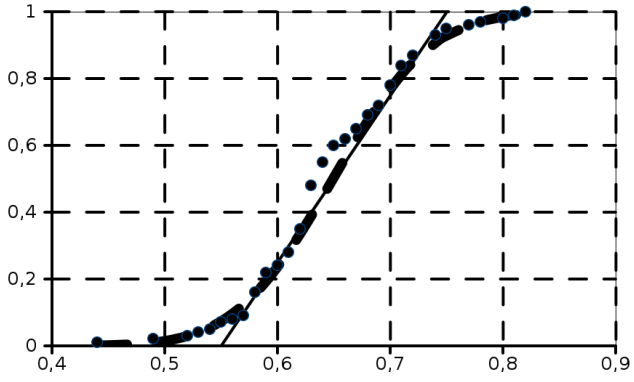
Dystrybuanta rozkładu jednostajnego opisana jest linią prostą o równaniu:

$$F(x) = \begin{cases} 0 & gdy \quad x < x_1 \\ \frac{x-a_1}{a_2-a_1} & gdy \quad x \in [a_1, a_2] \\ 0 & gdy \quad xx > a_2 \end{cases} \tag{2}$$

Parametry rozkładu jednostajnego można wyznaczyć na kilka sposobów, jednak w tym ćwiczeniu wykorzystamy metodę uproszczona polegająca na dobraniu wartości a_1 i a_2 (dwóch parametrów równania (3.3.1) tak aby wykres prostej pokrywał się najlepiej (w ocenie na „oko") z wykresem dystrybuanty. Najprościej jest to wykonać ręcznie na wykresie dystrybuanty empirycznej. Punkty przecięcia prostej z prostymi poziomymi zera i jedynki wyznaczają a_1 i a_2: czyli $F(a_1) = 0$ i $F(a_2) = 1$. Korzystając z komputera należy wykreślić na jednym rysunku funkcję daną wzorem (3.3.1) i wykres empiryczny i tak dobrać parametry a_1 i a_2 aby prosta najlepiej pasowała do danych dystrybuanty empirycznej. Należy parametry te umieścić w dwóch komórkach arkusza a wzór na dystrybuantę należy wpisać korzystając z tych komórek.

Prostej nie należy prowadzić przez wartości maksymalną i minimalna bowiem zazwyczaj dystrybuanta ma płaskie wykresy na krańcach przedziału zmienności wartości pomiarów. W przykładzie na rysunku 4 zakres zmienności czasów to przedział [0.44, 0.82] natomiast parametry, które dają dobre przyleganie prostej do głównej części wykresu wynoszą $a_1 = 0,55$ i $a_2 = 0,75$. Łatwo zauważyć, że prosta poprowadzona przez wartości brzegowe 0,44 i 0.82 nie będzie dobrze pasować do danych empirycznych.

Również współczynniki prostej wyznaczone metodą najmniejszych kwadratów dla całości danych nie są zadowalające z punktu widzenia obserwacji wzrokowej. Dopasowanie prostej metodą najmniejszych kwadratów dla prostej nie jest dobre ponieważ faktycznie dystrybuanta jest krzywą łamaną składająca się z dwóch odcinków płaskich i jednej ukośnej.



Rysunek 4: Przykładowa dystrybuanta empiryczna czasów spadania (kropki), dystrybuanty rozkładu jednostajnego (linia ciągła) i rozkładu normalnego (linia przerywana). Prosta będąca dystrybuantą rozkładu jednostajnego została poprowadzona przez obszar danych układających się na możliwie prostej części wykresu z pominięciem „zagięć". Prosta na rysunku jest wyznaczona z równania (3.3.1) dla parametrów $a_1 = 0,55s$ i $a_2 = 0,75s$. Widać, że rozkład normalny lepiej pasuje do rozkładu empirycznego.

3.3.2. Rozkład normalny.

Rozkład normalny określony jest przez dwa parametry: wartość oczekiwaną μ i odchylenie standardowe σ. Wzór na funkcję gęstości prawdopodobieństwa ma postać:

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

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

Dystrybuanta rozkładu normalnego może być wyliczona przy pomocy odpowiedniej funkcji arkusza kalkulacyjnego (np. w Gnumericu funkcja tan nazywa się „normdist". Funkcja ta ma cztery argumenty; (zmienna, wartość średnia, odchylenie standardowe,liczba). Jesli chcemy obliczeń dystrybuanty liczba musi być 1.

3.4. Porównanie wykresów empirycznych z rozkładami hipotetycznymi

W celu porównania rozkładu empirycznego z rozkładem hipotetycznym należy wykonać następujące wykresy:

1) na jednym wykresie narysować dystrybuanty: empiryczną, rozkładu jednostajnego i rozkładu normalnego i ewentualnie innego.
2) na jednym wykresie narysować histogram: dla danych empirycznych, oraz rozkładu jednostajnego i rozkładu normalnego i ewentualnie innego.

Przykład wykresu dla dystrybuanty rozkładu empirycznego oraz jednostajnego i normalnego pokazany jest na rys. 4. Parametry rozkładu normalnego wyznaczone są metodą opisaną w punkcie 3.3.2.

W celu porównania histogramu empirycznego z rozkładem jednostajnym i normalnym należy wyznaczyć prawdopodobieństwa p_l dla każdego l-tego przedziału (czyli koszyka) na podstawie następującego wzoru:

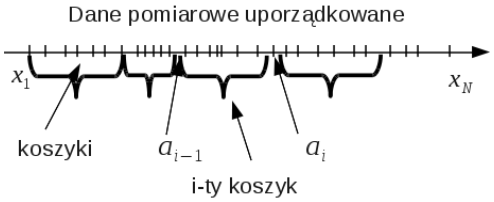
$$p_l = P\left(a_{l-1} < x < a_l\right) = \tag{4}$$

$$= \int_{a_{l-1}}^{a_l} f(x)dx = F(a_l) - F(a_{l-1}) \tag{5}$$

gdzie a_l jest końcem koszyków ($l = 0, 1, \ldots, L$), $f(x)$ jest gęstością rozkładu prawdopodobieństwa (hipoteza), a $F(x)$ jego dystrybuantą i p_l jest prawdopodobieństwem tego, że zmierzona wartość x jest wewnątrz l-tego koszyka.

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. W celu budowy histogramu koszyki powinny mieć jednakową szerokość (ok sześciu koszyków). Należy więc przedział w którym obserwujemy mierzoną zmienną (przedział zmienności) podzielić na jednakowe przedziały.

Dystrybuantę można wyznaczyć posługując się dowolnym programem typu arkusz kalkulacyjny (mogą być inne programy do analizy danych). Obliczenia należy zrobić dla wszystkich rozważanych rozkładów (jednostajnego, normalnego i ewentualnie innego) wykorzystując parametry policzone zgodnie z punktem 3.3.



Rysunek 5: Wyznaczanie granic koszyków z serii uporządkowanych danych pomiarowych

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

$$t_0 = \sqrt{\frac{2H}{g}} \tag{6}$$

gdzie H - wysokość, $g = 9,81m.s^2$ - przyśpieszenia grawitacyjne.

Należy porównać wynik obliczenia czasu t_0 spadania wyznaczonego na podstawie wzoru (6) 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.