Analiza danych

Mariusz Przybycień

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza

Wykład 2

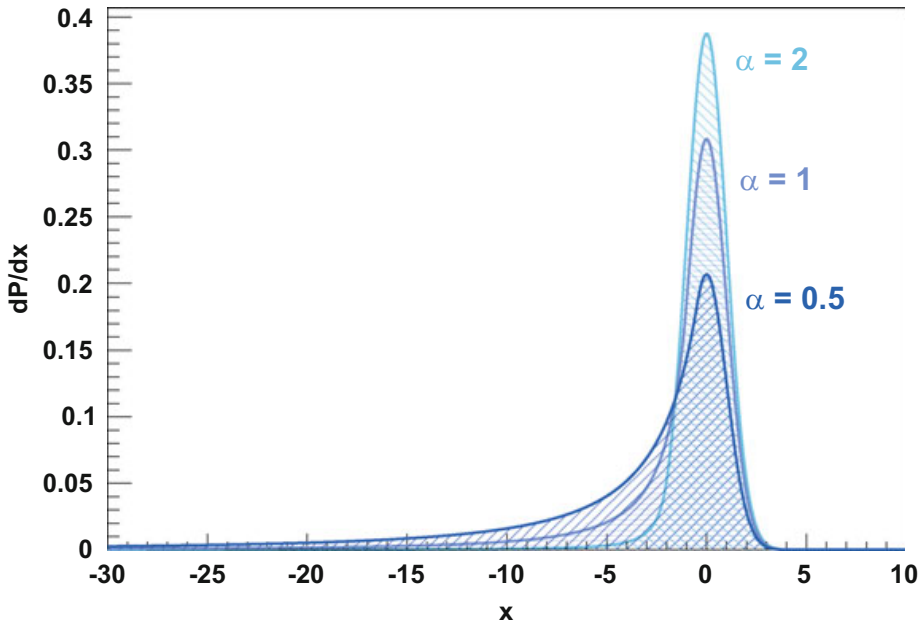
Funkcja Crystal Ball

- W wielu sytuacjach (np. odpowiedź detektora, rozkłady pewnych zmiennych) jedynie w przybliżeniu mają rozkłady Gaussa. Często zdaża się, że rozkłady te charakteryzują się brakiem symetrii i długim ogonem po jednej lub drugiej stronie.

- Współpraca Crystal Ball (eksperyment w SLAC) zaproponowała wykorzystanie do opisu takich zmiennych funkcję:

$$p(x;\alpha,n,\mu,\sigma) =$$

$$N\cdot\begin{cases} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] & \text{dla} \quad \frac{x-\mu}{\sigma} > -\alpha \\ A\cdot\left(B-\frac{x-\mu}{\sigma}\right)^{-n} & \text{dla} \quad \frac{x-\mu}{\sigma} \leqslant -\alpha \end{cases}$$



gdzie N jest stałą normalizacyjną, a parametry A oraz B zapewniają ciągłość funkcji i pierwszej pochodnej:

$$A = \left(\frac{n}{|\alpha|}\right)^n e^{-\alpha^2/2}, \qquad B = \frac{n}{|\alpha|} - |\alpha|$$

Relatywistyczny rozkład Breita-Wignera

- Relatywistyczny rozkład B-W opisuje p-two produkcji rezonansów (cząstek niestabilnych) w zależności od dostępnej energii E:

$$f(E) = \frac{k}{(E^2 - M^2)^2 + \Gamma^2 M^2}, \qquad k = \frac{2\sqrt{2} M \Gamma \gamma}{\pi \sqrt{M^2 + \gamma}}, \qquad \gamma = \sqrt{M^2 (M^2 + \Gamma^2)}$$

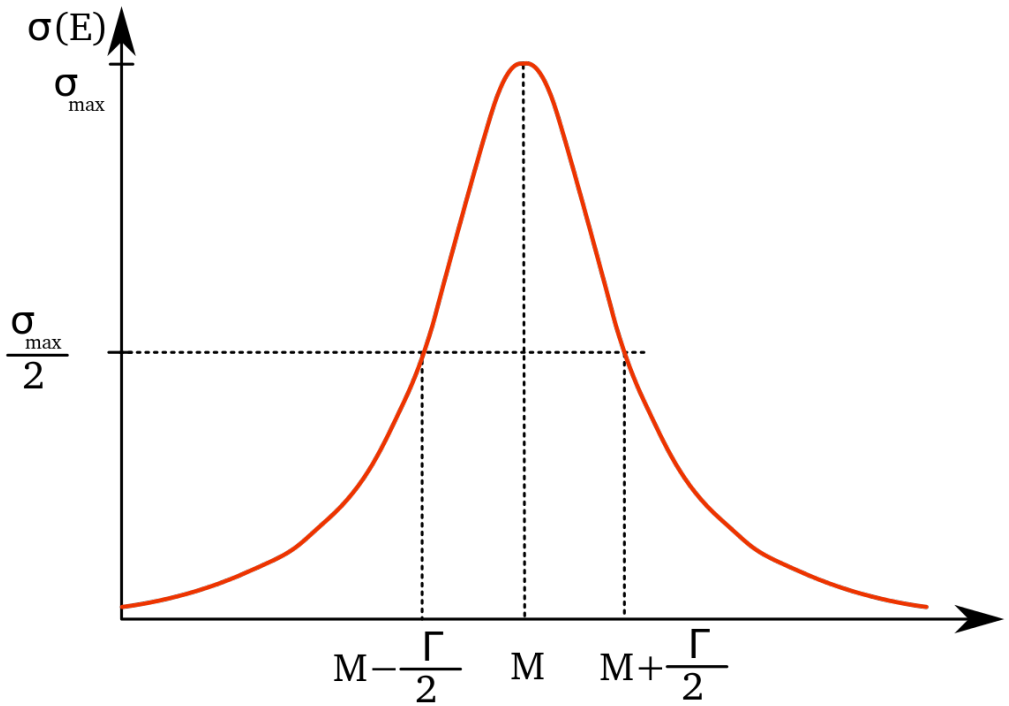
- Postać rozkładu wynika z propagatora cząstki niestabilnej, który w układzie spoczynkowym jest proporcjonalny do kwantowo-mechanicznej amplitudy opisującej rozpad rezonansu:

$$\Psi \propto \frac{\sqrt{k}}{(E^2 - M^2) + iM\Gamma}$$

- W przypadku reakcji

$$a + b \to R + X \to c + d + X$$

  - $E = M_{cd}$ - energia w CMS,
  - Γ - szerokość rozpadu związana ze średnim czasem życia $\tau = \hbar/\Gamma$.



Rozkład energii produktu rozpadu

- Przykład: Rozważmy rozpad $a \to b + c$ cząstki a o masie spoczynkowej m_a, energii E_a i pędzie $\vec{p}_a$ w LAB na dwie cząstki o masach spoczynkowych m_b i m_c.
  - wielkości oznaczone przez $\star$ to wielkości mierzone w układzie własnym cząstki a,
  - niech $\theta, \theta^\star$ oznaczają kąt emisji cząstki b względem pędu $\vec{p}_a$,
  - oznaczenia: $p_b^\star = p_c^\star \equiv p^\star, \quad \beta = p_a/E_a, \quad \gamma = E_a/m_a, \quad B = p_b^\star/E_b^\star$

$$
\begin{aligned}
p_b \cos\theta &= \gamma(p^\star \cos\theta^\star + \beta E_b^\star) = \gamma E_b^\star (B\cos\theta^\star + \beta) \\
p_b \sin\theta &= p^\star \sin\theta^\star \\
E_b &= \gamma(E_b^\star + \beta p^\star \cos\theta) = \gamma E_b^\star (1 + \beta B \cos\theta^\star)
\end{aligned}
$$

Dozwolony zakres energii cząstki b:

$$
\left.
\begin{aligned}
E_- &= \gamma E_b^\star - \gamma\beta p^\star \\
E_+ &= \gamma E_b^\star + \gamma\beta p^\star
\end{aligned}
\right\} \quad \Rightarrow \quad \Delta E \equiv E_+ - E_- = 2\gamma\beta p^\star = 2\,\frac{p_a p^\star}{m_a}
$$

Ponieważ: $\quad f(x = \cos\theta^\star, \varphi) = \dfrac{1}{4\pi}, \quad -1 \leqslant x \leqslant 1, \quad 0 \leqslant \varphi \leqslant 2\pi$

więc:

$$
g(E_b, \varphi) = f(x(E_b), \varphi) \left| \frac{dx}{dE_b} \right| = \frac{1}{2\pi\,\Delta E}, \quad E_- \leqslant E_b \leqslant E_+, \quad 0 \leqslant \varphi \leqslant 2\pi
$$

Randomizacja

- Rozkład energii cząstki b otrzymujemy jako rozkład brzegowy:

$$g_1(E_b) = \int_0^{2\pi} g(E_b,\varphi)d\varphi = \frac{1}{\Delta E} \qquad E_- \leqslant E_b \leqslant E_+$$

$$f(E_b) = g\left(E_b|E_a^{(1)}\right)P_1 + g\left(E_b|E_a^{(2)}\right)P_2 + g\left(E_b|E_a^{(3)}\right)P_3$$

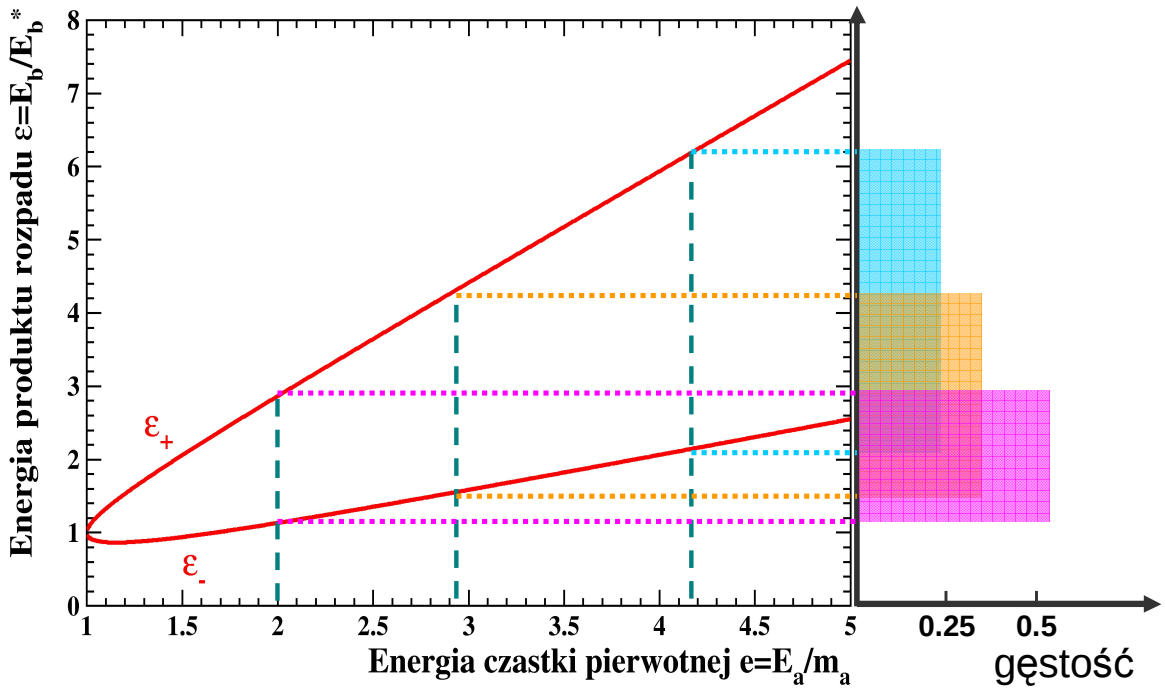
$$f(E_b) = \int_{E_a^-(E_b)}^{E_a^+(E_b)} g(E_b|E_a)h(E_a)dE_a = \frac{m_a}{2p^\star}\int_{E_a^-(E_b)}^{E_a^+(E_b)} \frac{1}{E_a^2-m_a^2}h(E_a)dE_a$$

$$\varepsilon_\pm \equiv \frac{E_\pm}{E_b^\star} =$$

$$= \frac{E_a}{m_a} \pm B\sqrt{\left(\frac{E_a}{m_a}\right)^2 - 1}$$

- Ogólnie:

$$f(x) = \int_{-\infty}^{+\infty} g(x|y)h(y)dy$$



Splatanie rozkładu wykładniczego

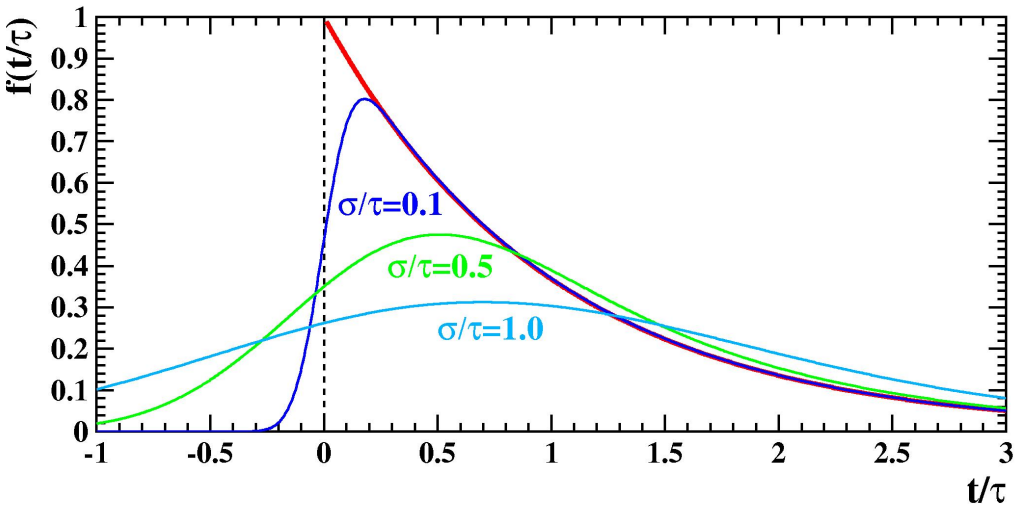
Przykład: Mierzymy czas t oczekiwania na rozpad jądra promieniotwórczego za pomocą przyrządu którego zdolność rozdzielcza ma kształt gaussowski:

$$h(t) = \lambda \exp(-\lambda t) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) \qquad r(t'|t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t'-t)^2}{2\sigma^2}\right)$$

Splatamy teoretyczną funkcję rozkładu z funkcją zdolności rozdzielczej:

$$f(t') = \int_0^\infty r(t'|t)h(t)\,dt = \frac{\lambda}{\sqrt{2\pi}\sigma} \int_0^\infty \exp\left(-\lambda t - \frac{(t'-t)^2}{2\sigma^2}\right) dt =$$

$$= \frac{\lambda}{\sqrt{2\pi}} \exp\left(-\lambda t' + \frac{\lambda^2\sigma^2}{2}\right) \int_{-\infty}^{t'/\sigma - \lambda\sigma} \exp\left(-\frac{y^2}{2}\right) dy$$



- Splatanie prowadzi do wygładzania rozkładów, tym silniejszego im mniej precyzyjnie przeprowadzamy pomiar.
- Zdarza się, że w wyniku mało precyzyjnego pomiaru otrzymujemy wielkości mierzone w niefizycznym obszarze.

Statystyka porządku

Niech $\mathrm{X}_1, \mathrm{X}_2, ...$ będą zmiennymi losowymi o rozkładzie zadanym dystrybuantą $F(x)$.

Definicja: Dla $k = 1, 2, ..., n$ niech $\mathrm{X}_{(k)}$ oznacza k-tą z kolei najmniejszą wartość spośród $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$. Wówczas niemalejący ciąg $(\mathrm{X}_{(1)}, \mathrm{X}_{(2)}, ..., \mathrm{X}_{(n)})$ nazywamy statystyką porządku, a $\mathrm{X}_{(k)}$ k-tą zmienną w porządku.

Wniosek: Statystykę porządku otrzymuje się z nieuporządkowanej próbki poprzez permutację taką w wyniku której mamy: $\mathrm{X}_{(1)} \leqslant \mathrm{X}_{(2)} \leqslant \cdots \leqslant \mathrm{X}_{(n)}$

Uwaga: Statystyka porządku zależy także od n: $\mathrm{X}_{(k)}$ jest k-tą najmniejszą wartością spośród n zmiennnych $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$.
Czasem stosuje się bardziej precyzyjną notację $\mathrm{X}_{1:n}, \mathrm{X}_{2:n}, ..., \mathrm{X}_{n:n}$.

Wartości ekstremalne i ich rozkłady:

$$\mathrm{X}_{(1)} = \min\{\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n\} \quad \text{oraz} \quad \mathrm{X}_{(n)} = \max\{\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n\}$$

$$F_{\mathrm{X}_{(1)}}(x) = 1 - P\left(\mathrm{X}_{(1)} > x\right) = 1 - P\left(\mathrm{X}_1 > x, \mathrm{X}_2 > x, ...\mathrm{X}_n > x\right) =$$

$$= 1 - \prod_{k=1}^{n} P\left(\mathrm{X}_k > x\right) = 1 - \left(1 - F(x)\right)^n$$

$$F_{\mathrm{X}_{(n)}}(x) = P\left(\mathrm{X}_1 \leqslant x, \mathrm{X}_2 \leqslant x, ...\mathrm{X}_n \leqslant x\right) = \prod_{k=1}^{n} P\left(\mathrm{X}_k \leqslant x\right) = \left(F(x)\right)^n$$

Statystyka porządku - rozkłady

Dla rozkładów ciągłych znajdujemy funkcje gęstości prawdopodobieństwa:

$$f_{\mathrm{X}_{(1)}}(x) = n\left(1 - F(x)\right)^{n-1} f(x) \quad \text{oraz} \quad f_{\mathrm{X}_{(n)}}(x) = n\left(F(x)\right)^{n-1} f(x)$$

Przykład: W biegu na 100 metrów czasy uzyskane przez zawodników mają rozkład płaski na przedziale $(9.6, 10)$ sekund. W finale uczetniczy 8 zawodników. Ile wynosi prawdopodobieństwo, że zwycięzca pobije rekord 9.69 s?

$$F(x) = 2.5x - 24 \quad \Rightarrow \quad F_{\mathrm{X}_{(1)}}(x) = 1 - (25 - 2.5x)^8 \quad \Rightarrow \quad p = F_{\mathrm{X}_{(1)}}(9.69) \approx 0.8699$$

Twierdzenie: Dla $k = 1, 2, ..., n$ mamy:

$$F_{\mathrm{X}_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \int_0^{F(x)} y^{k-1}(1-y)^{n-k} dy$$

czyli

$$F_{\mathrm{X}_{(k)}}(x) = F_{\beta(k,n+1-k)}\left(F(x)\right) \qquad \text{gdzie } \beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

Wniosek: W szczególności dla rozkładu płaskiego $u(0,1)$ mamy

$$\mathrm{X}_{(k)} \in \beta(k, n+1-k), \quad k = 1, 2, ..., n$$

Przykład: Wylosowano 100 liczb z rozkładu płaskiego $u(0,1)$. Znajdź prawdopodobieństwo, że druga z kolei najmniejsza liczba nie jest mniejsza niż 0.002.

$$F(x) = x \quad \Rightarrow \quad p = 1 - F_{\mathrm{X}_{(2)}}(0.002) \approx 0.9826$$

Gęstość prawdopodobieństwa zmiennej uporządkowanej

Twierdzenie: Dla rozkładu ciągłego o funkcji gestości $f(x)$, funkcja gęstości zmiennej $\mathrm{X}_{(k)}$ dla $k = 1, 2, ..., n$ dana jest przez pochodną $F_{\mathrm{X}_{(k)}}(x)$:

$$f_{\mathrm{X}_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} f(x) =$$
$$= f_{\beta(k,n+1-k)}\left(F(x)\right) \cdot f(x)$$

Dowód: Stosujemy regułę całkową Leibniza do $F_{\mathrm{X}_{(k)}}(x)$:

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)\,dt\right) = f(x,b(x))\frac{db}{dx} - f(x,a(x))\frac{da}{dx} + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt$$

Przykład:

Łączny rozkład ekstremów

Twierdzenie: Łączna gęstość prawdopodobieństwazmiennych $\mathrm{X}_{(1)}$ i $\mathrm{X}_{(n)}$ dana jest przez:

$$f_{\mathrm{X}_{(1)},\mathrm{X}_{(n)}}(x,y) = \begin{cases} n(n-1)\left(F(y)-F(x)\right)^{n-2} f(y)f(x), & \text{dla } x<y \\ 0 & \text{dla pozostałych } (x,y) \end{cases}$$

Dowód: Korzystamy z niezależności zmienny $\mathrm{X}_{(k)}$ oraz rozłączności zdarzeń $(\mathrm{X}_{(1)} \leqslant x)$ i $(\mathrm{X}_{(1)} > x)$:

$$P\left(\mathrm{X}_{(1)} > x, \mathrm{X}_{(n)} \leqslant y\right) = P\left(x < \mathrm{X}_k \leqslant y,\, k=1,2,...,n\right) =$$

$$= \prod_{k=1}^{n} P\left(x < \mathrm{X}_k \leqslant y\right) = \left(F(y)-F(x)\right)^n, \quad \text{dla} \;\; x<y$$

$$P\left(\mathrm{X}_{(1)} \leqslant x, \mathrm{X}_{(n)} \leqslant y\right) + P\left(\mathrm{X}_{(1)} > x, \mathrm{X}_{(n)} \leqslant y\right) = P\left(\mathrm{X}_{(n)} \leqslant y\right)$$

$$F_{\mathrm{X}_{(1)},\mathrm{X}_{(n)}}(x,y) = F_{\mathrm{X}_{(n)}}(y) - P\left(\mathrm{X}_{(1)} > x, \mathrm{X}_{(n)} \leqslant y\right) =$$

$$= \begin{cases} \left(F(y)\right)^n - \left(F(y)-F(x)\right)^n, & \text{dla} \;\; x<y \\ \left(F(y)\right)^n, & \text{dla} \;\; x \geqslant y \end{cases}$$

Aby otrzymać funkcję gęstości należy obliczyc pochodną $\dfrac{\partial^2}{\partial x \partial y} F_{\mathrm{X}_{(1)},\mathrm{X}_{(n)}}(x,y)$.

Łączny rozkład ekstremów, zakres

Gęstość prawdopodobieństwa zmiennej losowej $\mathrm{R}_n = \mathrm{X}_{(n)} - \mathrm{X}_{(1)}$ znajdujemy jako rozkład brzegowy dwuwymiarowej zmiennej losowej $(\mathrm{R}_n, \mathrm{U} = \mathrm{X}_{(1)})$:

$$f_{\mathrm{R}_n}(r) = n(n-1) \int_{-\infty}^{\infty} \left(F(u+r) - F(u)\right)^{n-2} f(u+r) f(u) du, \quad \text{dla} \;\; r > 0$$

Przykład: Dla zmiennej losowej $\mathrm{X} \in U(0,1)$ mamy:

$$f_{\mathrm{R}_n}(r) = n(n-1) \int_0^{1-r} (u+r-u)^{n-2} \cdot 1 \cdot 1 \cdot du = n(n-1) r^{n-2}(1-r), \quad 0 < r < 1$$

Wartość oczekiwana: $\mathcal{E}\left[\mathrm{R}_n\right] = \dfrac{n-1}{n+1}$.

Przykład: Niech $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$ będą niezależnymi zmiennymi losowymi z rozkładu wykładniczego $Exp(\lambda)$. Znajdź (a) $f_{\mathrm{X}_{(1)},\mathrm{X}_{(n)}}(x,y)$, (b) $f_{\mathrm{R}_n}(r)$.

$$\begin{aligned} f_{\mathrm{X}_{(1)},\mathrm{X}_{(n)}}(x,y) &= n(n-1)\left(1 - e^{-\lambda y} - (1 - e^{-\lambda x})\right)^{n-2} \cdot \lambda e^{-\lambda y} \cdot \lambda e^{-\lambda x} = \\ &= n(n-1)\left(e^{-\lambda x} - e^{-\lambda y}\right)^{n-2} \lambda^2 e^{-\lambda(x+y)}, \quad \text{dla} \;\; 0 < x < y \end{aligned}$$

$$\begin{aligned} f_{\mathrm{R}_n}(r) &= n(n-1) \int_0^{\infty} \left(e^{-\lambda u} - e^{-\lambda(u+r)}\right)^{n-2} \lambda^2 e^{-\lambda(2u+r)} du = \\ &= (n-1)\lambda \left(1 - e^{-\lambda r}\right)^{n-2} e^{-\lambda r}, \quad \text{dla} \;\; r > 0 \end{aligned}$$

Warunkowe rozkłady statystyki porządku

Przykład cd: Dystrybuanta zakresu dla zmiennych z rozkładu wykładniczego:

$$F_{R_n}(r) = (1 - e^{-\lambda r})^{n-1} = (F(r))^{n-1}$$

Rezultat zgodny z "brakiem pamięci" rozkładu wykładniczego - $(\mathrm{X}_{(2)} - \mathrm{X}_{(1)}, \mathrm{X}_{(3)} - \mathrm{X}_{(1)}, ..., \mathrm{X}_{(n)} - \mathrm{X}_{(1)})$ można interpretować jako statystykę porządku dla próbki $n-1$ elementowej, gdzie $R_n = \mathrm{X}_{(n)} - \mathrm{X}_{(1)}$ jest jej wartością maksymalną.

Przykład: Niech $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ będą niezależnymi zmiennymi losowymi z rozkładu wykładniczego $Exp(\lambda = 1)$. Znajdź $\mathcal{E}\left[\mathrm{X}_{(3)} | \mathrm{X}_{(1)} = x\right]$.

$$f_{\mathrm{X}_{(1)},\mathrm{X}_{(3)}}(x,y) = 3 \cdot 2 \left(e^{-x} - e^{-y}\right) e^{-(x+y)}, \quad \text{dla } \ 0 < x < y$$

$$f_{\mathrm{X}_{(3)}|\mathrm{X}_{(1)}=x}(y) = \frac{f_{\mathrm{X}_{(1)},\mathrm{X}_{(3)}}(x,y)}{f_{\mathrm{X}_{(1)}}(x)} = \frac{6\left(e^{-x} - e^{-y}\right) e^{-(x+y)}}{3e^{-3x}} =$$

$$= 2\left(e^{-x} - e^{-y}\right) e^{2x-y} \quad \text{dla } \ 0 < x < y$$

Warunkowa wartość oczekiwana jest więc równa:

$$\mathcal{E}\left[\mathrm{X}_{(3)} | \mathrm{X}_{(1)} = x\right] = \int_x^\infty 2y \left(e^{-x} - e^{-y}\right)^{n-2} e^{2x-y} dy = \left\{ y = u + x \right\} =$$

$$= \int_0^\infty 2(u+x) \left(e^{-x} - e^{-(u+x)}\right) e^{2x-u-x} du = x + \frac{3}{2}$$

Łączny rozkład statystyki porządku

Twierdzenie: Łączna gęstość prawdopodobieństwazmiennych $\mathrm{X}_{(1)}, \mathrm{X}_{(2)}, ..., \mathrm{X}_{(n)}$ dana jest przez:

$$f_{\mathrm{X}_{(1)},\mathrm{X}_{(2)},...,\mathrm{X}_{(n)}}(y_1, y_2, ..., y_n) = \begin{cases} n! \prod_{k=1}^{n} f(y_k) & \text{dla } y_1 < y_2 < .. < y_n \\ 0 & \text{w pozostałych przypadkach} \end{cases}$$

Przykład: Niech $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ będą niezależnymi zmiennymi losowymi z rozkładu płaskiego $U(0,1)$. Znajdź wszystkie rozkłady brzegowe statystyki porządku otrzymanej z tych zmiennych.

Łączny rozkład p-twa: $f_{\mathrm{X}_{(1)},\mathrm{X}_{(2)},\mathrm{X}_{(3)}}(y_1, y_2, y_3) = 6$ dla $0 < y_1 < y_2 < y_3 < 1$

Rozkłady brzegowe:

$$\begin{aligned}
f_{\mathrm{X}_{(1)},\mathrm{X}_{(2)}}(y_1, y_2) &= \int_{y_2}^{1} 6\, dy_3 = 6(1 - y_2) \quad 0 < y_1 < y_2 < 1 \\
f_{\mathrm{X}_{(1)},\mathrm{X}_{(3)}}(y_1, y_3) &= \int_{y_1}^{y_3} 6\, dy_2 = 6(y_3 - y_1) \quad 0 < y_1 < y_3 < 1 \\
f_{\mathrm{X}_{(2)},\mathrm{X}_{(3)}}(y_2, y_3) &= \int_{0}^{y_2} 6\, dy_1 = 6y_2 \quad 0 < y_2 < y_3 < 1 \\
f_{\mathrm{X}_{(1)}}(y_1) &= \int_{y_1}^{1} 6(1 - y_2)\, dy_2 = 3(1 - y_1)^2 \quad 0 < y_1 < 1 \\
f_{\mathrm{X}_{(2)}}(y_2) &= \int_{0}^{y_2} 6(1 - y_2)\, dy_1 = 6y_2(1 - y_2) \quad 0 < y_2 < 1 \\
f_{\mathrm{X}_{(1)}}(y_1) &= \int_{y_1}^{1} 6(y_3 - y_1)\, dy_3 = 3(1 - y_1)^2 \quad 0 < y_1 < 1 \\
f_{\mathrm{X}_{(3)}}(y_3) &= \int_{0}^{y_3} 6(y_3 - y_1)\, dy_1 = 3y_3^2 \quad 0 < y_3 < 1
\end{aligned}$$