Analiza danych

Mariusz Przybycień

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza

Wykład 1

Literatura

- Statystyka dla fizyków, R.J. Nowak, PWN 2002.
- Statystyka dla fizyków. Ćwiczenia, R.J. Nowak, PWN 2002.
- Data Analysis Techniques for Physical Scientists, C. Pruneau, Cambridge 2017.
- Statistical Methods for Data Analysis in Particle Physics, L. Lista, Springer 2017.
- Statistical Data Analysis, G. Cowan, Oxford 1998.
- Data Analysis in High Energy Physics, O. Behnke, K. Kröninger, G. Schott, T. Schrörner-Sadenius, Wiley-VCH 2013.
- Statistical Analysis Techniques in Particle Physics, I. Narsky, F.C. Porter, Wiley-VCH 2014.
- Probability and Statistics for Particle Physics, C. Mana, Springer 2017.

$\Rightarrow$ http://home.agh.edu.pl/mariuszp

$\Rightarrow$ Wymagania wstępne - znajomość rachunku prawdopodobieństwa i statystyki matematycznej na poziomie kursu dla II roku I stopnia FT na WFiIS AGH.

Analiza danych w fizyce cząstek elementarnych

- Rejestrujemy przypadki zderzeń w akceleratorze (np. $e^{\pm}e^{-}$, $e^{\pm}p$, pp, p+A, A+A, $e^{\pm}$+A, ...) i dla każdego mierzymy pewien zestaw charakterystyk:

  pędy cząstek, ich rodzaje, liczbę muonów, energie dżetów, ...

- Rozkłady mierzonych charakterystyk porównujemy z przewidywaniami teoretycznymi (modele teoretyczne + symulacja odpowiedzi detektora w środowisku Geant4).
- Na podstawie tych porównań staramy się oszacować swobodne parametry modeli, ich niepewności (statystyczne i systematyczne):

  przekroje czynne, masy nowych cząstek, temp. przejścia do stanu QGP, ...

- ... a także ocenić jak dobrze poszczególne modele teoretyczne opisują dane eksperymentalne.
- Staramy się także stwierdzić czy w danych nie występują znaczące odstępstwa od modeli, wskazujące na zaobserwowanie nowych procesów/cząstek, itp.
- Aby to wszystko zrobić w sposób ścisły i ilościowy potrzebne jest podejście wykorzystujące rachunek prawdopodobieństwa i metody statystyczne.
- W dalszej części tego wykładu zakładam znajomość rachunku p-twa i statystyki matematycznej na poziomie kursu dla II roku FT na WFiIS AGH.

Charakterystyki rozkładów prawdopodobieństwa

Rozważamy zmienną losową ciągłą x o funkcji gęstości $f(x)$ oraz funkcję $y = g(x)$.

- Wartość oczekiwana:

$$\mu_{\mathrm{x}} \equiv \mathcal{E}[\mathrm{x}] = \int_{-\infty}^{+\infty} x\, f(x)\, dx, \qquad \mu_{\mathrm{y}} = \mathcal{E}[\mathrm{y}] = \int_{-\infty}^{+\infty} g(x)\, f(x)\, dx$$

- Wariancja:

$$\sigma_{\mathrm{x}}^2 \equiv \mathcal{V}[\mathrm{x}] = \mathcal{E}[(\mathrm{x} - \mu_x)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\, dx = \mathcal{E}[\mathrm{x}^2] - \mathcal{E}^2[\mathrm{x}]$$

- Momenty rzędu n - zwykły m_n oraz centralny μ_n:

$$m_n \equiv \mathcal{E}[\mathrm{x}^n] = \int_{-\infty}^{+\infty} x^n\, f(x)\, dx, \qquad \mu_n = \mathcal{E}[(\mathrm{x} - \mu_{\mathrm{x}})^n] = \int_{-\infty}^{+\infty} (x-\mu)^n\, f(x)\, dx$$

- Moment centralny mieszany μ_{rs} rzędu $r+s$ (μ_{11} nazywamy kowariancją):

$$\mu_{rs} \equiv \mathcal{E}[(\mathrm{x} - \mu_{\mathrm{x}})^r (\mathrm{y} - \mu_{\mathrm{y}})^s] = \iint_{-\infty}^{+\infty} (x-\mu_{\mathrm{x}})^r (y-\mu_{\mathrm{y}})^s\, f(x,y)\, dxdy$$

- Współczynnik korelacji: $\rho(\mathrm{x},\mathrm{y}) = \dfrac{\mathrm{cov}(\mathrm{x},\mathrm{y})}{\sigma_{\mathrm{x}}\, \sigma_{\mathrm{y}}}$

Wektor losowy, macierz kowariancji

Niech $\vec{\mathrm{x}}$ bedzie wektorem losowym (tzn. takim którego składowe są zmiennymi losowymi). Wówczas wektorem wartości oczekiwanych jest $\vec{\mu} = \mathcal{E}[\vec{\mathrm{x}}]$ natomiast macierzą kowariancji nazywamy:

$$\mathrm{V}[\vec{\mathrm{x}}] = \mathcal{E}\left[(\vec{\mathrm{x}} - \vec{\mu})(\vec{\mathrm{x}} - \vec{\mu})^{\mathrm{T}}\right]$$

gdzie $v_{ij} = \mathcal{E}\left[(\mathrm{x}_i - \mu_i)(\mathrm{x}_j - \mu_j)\right], \quad i, j = 1, 2, ..., n.$

W szczególności: $v_{ii} = \mathcal{V}\left[\mathrm{x}_i\right]$ oraz $v_{ij} = \mathrm{cov}[\mathrm{x}_i, \mathrm{x}_j]$.

Twierdzenie: Niech będzie dany wektor losowy $\vec{\mathrm{x}}$ wraz z wektorem wartości oczekiwanych $\vec{\mu}$ oraz macierzą kowariancji Λ. Wówczas wektor wartości oczekiwanych i macierz kowariancji wektora losowego $\vec{\mathrm{y}} = \mathrm{B}\vec{\mathrm{x}} + \vec{b}$, gdzie B jest macierzą $m \times n$, a $\vec{b}$ stałym wektorem, dane są przez:

$$\mathcal{E}\left[\vec{\mathrm{y}}\right] = \mathrm{B}\vec{\mu} + \vec{b} \qquad \text{oraz} \qquad \mathrm{V}[\vec{\mathrm{y}}] = \mathrm{B}\mathrm{V}[\vec{\mathrm{x}}]\mathrm{B}^{\mathrm{T}}$$

Dowód:

$$\begin{aligned}
\mathcal{E}[\vec{\mathrm{y}}] &= \mathrm{B}\,\mathcal{E}\left[\vec{\mathrm{x}}\right] + \vec{b} = \mathrm{B}\vec{\mu} + \vec{b} \\
\mathrm{V}[\vec{\mathrm{y}}] &= \mathcal{E}\left[(\vec{\mathrm{y}} - \mathcal{E}[\vec{\mathrm{y}}])(\vec{\mathrm{y}} - \mathcal{E}[\vec{\mathrm{y}}])^{\mathrm{T}}\right] = \mathcal{E}\left[\mathrm{B}(\vec{\mathrm{x}} - \vec{\mu})(\vec{\mathrm{x}} - \vec{\mu})^{\mathrm{T}}\mathrm{B}^{\mathrm{T}}\right] = \\
&= \mathrm{B}\,\mathcal{E}\left[(\vec{\mathrm{x}} - \vec{\mu})(\vec{\mathrm{x}} - \vec{\mu})^{\mathrm{T}}\right]\mathrm{B}^{\mathrm{T}} = \mathrm{B}\mathrm{V}[\vec{\mathrm{x}}]\mathrm{B}^{\mathrm{T}}
\end{aligned}$$

Rozkład jednorodny (płaski, jednostajny)

- Dyskretny: $P_k = \frac{1}{n}, \quad k = 1, ..., n, \qquad \mathcal{E}[\mathrm{k}] = \frac{1}{2}(n+1), \qquad \mathcal{V}[\mathrm{k}] = \frac{n^2-1}{12}$
- Ciągły: $f(x) = \frac{1}{b-a}, \quad a < x < b, \qquad \mathcal{E}[\mathrm{x}] = \frac{1}{2}(b-a), \qquad \mathcal{V}[\mathrm{x}] = \frac{(b-a)^2}{12}$
- Dystrybuanta dowolnej zmiennej losowej ma rozkład płaski na przedziale $[0, 1]$:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \quad \left\{ y = F(x) \right\} \quad g(y) = f(x(y)) \left| \frac{dx}{dy} \right| = f(x) \frac{1}{f(x)} = 1$$

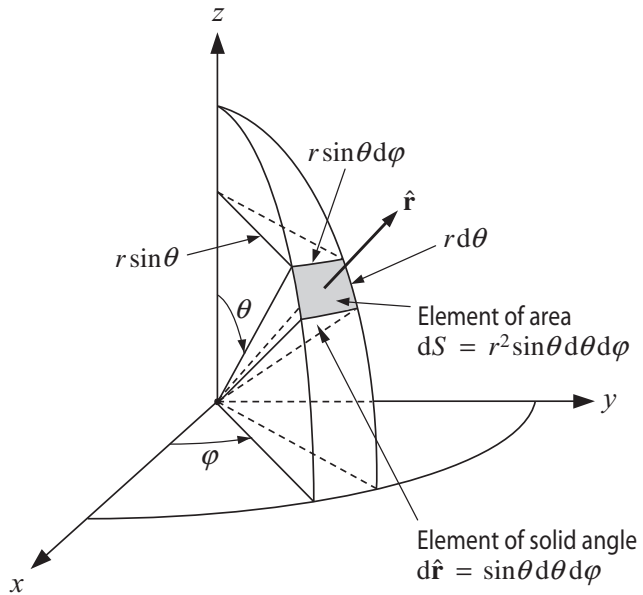
- Rozkład jednorodny na sferze o promieniu $r = 1$:

$$P(\Omega)\,d\Omega = \frac{1}{4\pi} \sin\theta\,d\theta d\varphi = P(\theta, \varphi)\,d\theta d\varphi$$

$$P(\theta, \varphi) = \frac{\sin\varphi}{4\pi} \quad \Rightarrow \quad \begin{cases} P(\varphi) = \int_0^{\pi} P(\theta, \varphi)\,d\theta = \frac{1}{2\pi} \\ P(\theta) = \int_0^{2\pi} P(\theta, \varphi)\,d\varphi = \frac{\sin\theta}{2} \end{cases}$$

$$u = F(\varphi) = \frac{\varphi}{2\pi} \quad \Rightarrow \quad \varphi = 2\pi u$$

$$v = F(\theta) = \frac{1 - \cos\theta}{2} \quad \Rightarrow \quad \theta = \cos^{-1}(2v - 1)$$

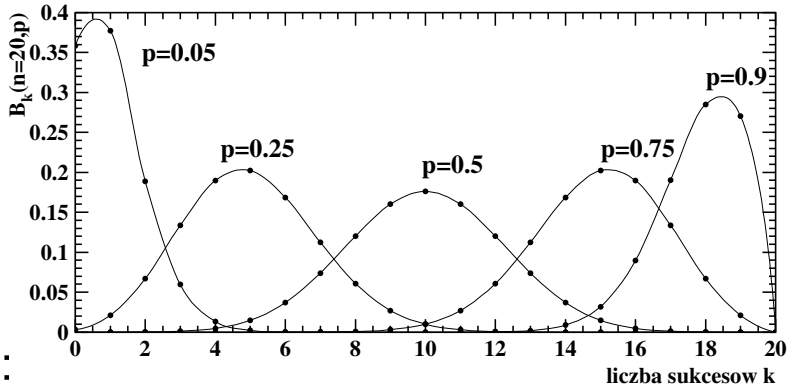


Rozkłady dwu- i wielomianowy oraz geometryczny

- Rozkład dwumianowy zmiennej losowej k:

$$\mathcal{B}_k(n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n$$

$$\mathcal{E}[\mathrm{k}] = np, \quad \mathcal{V}[\mathrm{k}] = np(1-p)$$



- Rozkład wielomianowy ($k_1 + k_2 + ... + k_j = n$):

$$\mathcal{W}_{k_1 k_2 ... k_j}(n, p_1, p_2, ..., p_j) = \frac{n!}{k_1! k_2! \cdots k_j!} p_1^{k_1} p_2^{k_2} \cdots p_j^{k_j}$$

$$\mathcal{E}[\mathrm{k}_i] = np_i, \quad \mathcal{V}[\mathrm{k}_i] = np_i(1-p_i)$$

- Rozkład geometryczny:

$$\mathcal{G}_k(p) = p(1-p)^{k-1}, \quad k = 1, 2, ... \qquad \mathcal{E}[\mathrm{k}] = \frac{1}{p}, \quad \mathcal{V}[\mathrm{k}] = \frac{1-p}{p^2}$$

- Rozkład ujemny dwumianowy:

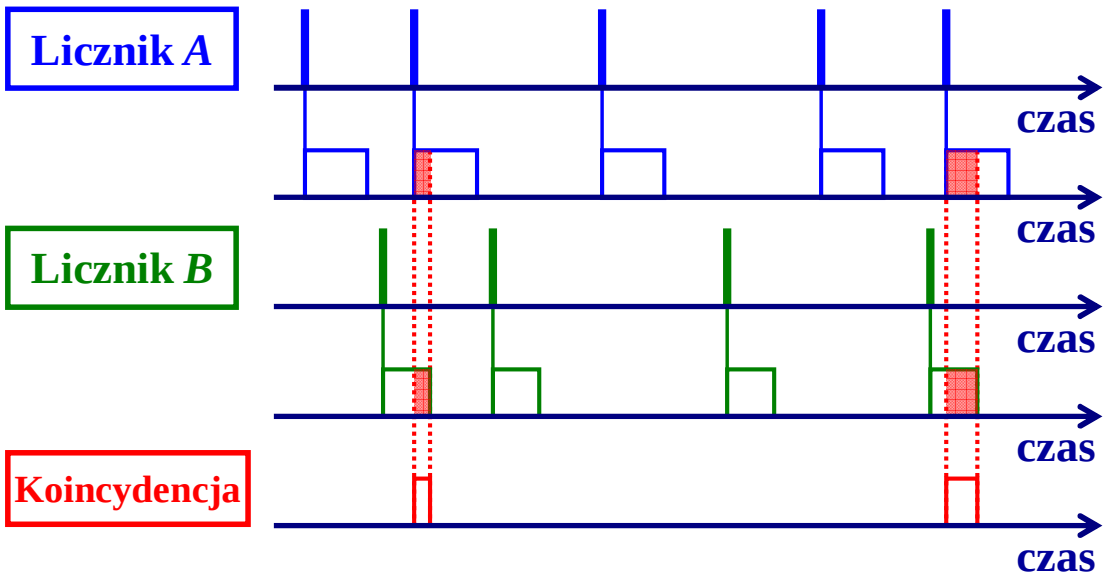
$$\mathcal{U}_k(n,p) = \binom{k-1}{n-1} p^k (1-p)^{n-k}, \quad n = 1, 2, ... \quad k = n, n+1, ...$$

$$\mathcal{E}[\mathrm{k}] = \frac{n}{p}, \quad \mathcal{V}[\mathrm{k}] = \frac{n(1-p)}{p^2}$$

Koincydencje przypadkowe

Rozważmy dwa liczniki Geigera-Müllera (A i B) umieszczone w poziomie w pewnej odległości od siebie i rejestrujące promieniowanie kosmiczne.

- $f_{\mathrm{A}}, f_{\mathrm{B}}$ - typowe częstości rejestracji promieni przez każdy z liczników.
- $\tau_{\mathrm{A}} = 1/f_{\mathrm{A}},\ \tau_{\mathrm{B}} = 1/f_{\mathrm{B}}$ - typowe odstępy czasu pomiędzy kolejnymi cząstkami.
- Impulsy wyjściowe mają standardowy kształt prostokątny o czasach trwania T_{A} i T_{B} które są znacznie krótsze niż τ_{A} i τ_{B}.
- Impulsy wyjściowe wysyłamy na układ koincydencyjny, który daje sygnał gdy impulsy wejściowe przekrywają się choć minimalnie.



Niech czas obserwacji $\tau \gg \tau_{\mathrm{A}}, \tau_{\mathrm{B}}$, wtedy p-two sygnału na wyjściu detektorów wynosi odpowiednio:

$$\frac{\tau}{\tau_{\mathrm{A}}} T_{\mathrm{A}} = \tau f_{\mathrm{A}} T_{\mathrm{A}} \quad \Rightarrow \quad P_{\mathrm{A}} = f_{\mathrm{A}} T_{\mathrm{A}}$$

$$\frac{\tau}{\tau_{\mathrm{B}}} T_{\mathrm{B}} = \tau f_{\mathrm{B}} T_{\mathrm{B}} \quad \Rightarrow \quad P_{\mathrm{B}} = f_{\mathrm{B}} T_{\mathrm{B}}$$

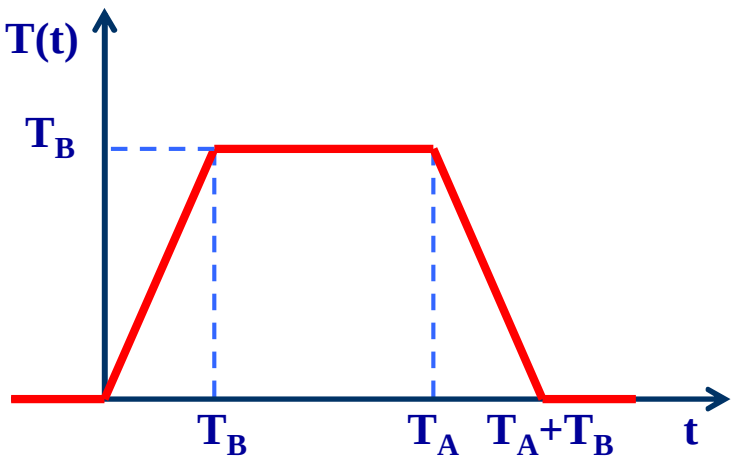
Zdarzenia są niezależne, więc:

$$P = P_{\mathrm{A}} P_{\mathrm{B}} = f_{\mathrm{A}} T_{\mathrm{A}} f_{\mathrm{B}} T_{\mathrm{B}}$$

Koincydencje przypadkowe

- Oszacowanie średniego czasu trwania pojedynczej koincydencji ($T_A > T_B$):

$$T(t) = \begin{cases} t & 0 \leqslant t \leqslant T_B \\ T_B & T_B \leqslant t \leqslant T_A \\ T_A + T_B - t & T_A \leqslant t \leqslant T_A + T_B \end{cases}$$



$$\overline{T} = \frac{1}{T_A + T_B} \left(\int_0^{T_B} t\,dt + \int_{T_B}^{T_A} T_B\,dt + \int_{T_A}^{T_A+T_B} (T_A + T_B - t)\,dt \right) = \frac{T_A T_B}{T_A + T_B}$$

- Częstość koincydencji przypadkowych dana jest przez:

$$f_{AB} = \frac{P\tau/\overline{T}}{\tau} = f_A T_A f_B T_B \left(\frac{1}{T_A} + \frac{1}{T_B} \right) = f_A f_B (T_A + T_B)$$

- Podobnie, dla potrójnych koincydencji otrzymujemy:

$$\frac{1}{\overline{T}_{ABC}} = \frac{1}{\overline{T}_{AB}} + \frac{1}{T_C} = \frac{1}{T_A} + \frac{1}{T_B} + \frac{1}{T_C}$$

$$f_{ABC} = f_A T_A f_B T_B f_C T_C \left(\frac{1}{T_A} + \frac{1}{T_B} + \frac{1}{T_C} \right)$$

Rozkład wykładniczy i Weibulla

- Rozkład wykładniczy opisuje czas oczekiwania na zdarzenie:

$$Exp(t;\lambda) = \lambda e^{-\lambda t} \quad \text{lub} \quad Exp(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \quad \text{gdzie} \quad \tau = \frac{1}{\lambda}$$

$$\mathcal{E}[\mathrm{t}] = \tau = \frac{1}{\lambda} \qquad \mathcal{V}[\mathrm{t}] = \tau^2 = \frac{1}{\lambda^2}$$

- Częstość koincydencji przypadkowych:

λ_A, λ_B - intensywności rejestracji promieniowania przez liczniki A i B,
T_A, T_B - czasy trwania impulsów generowanych przez te liczniki.

Niech koincydencja będzie inicjowana przez licznik A, wtedy:

$$f(t_A, t_B; \lambda_A, \lambda_B) = \lambda_B(\lambda_A + \lambda_B) \exp\left(-\lambda_A t_A - \lambda_B t_B\right) \qquad 0 \leqslant t_A \leqslant t_B < \infty$$

P-two koincydencji:

$$P(t_A - t_B \leqslant T_A) = \lambda_B(\lambda_A + \lambda_B) \int_0^{\infty} dt_A \, e^{-\lambda_A t_a} \int_{t_A}^{t_A + T_A} dt_B \, e^{-\lambda_B t_B} = 1 - e^{-\lambda_B T_A}$$

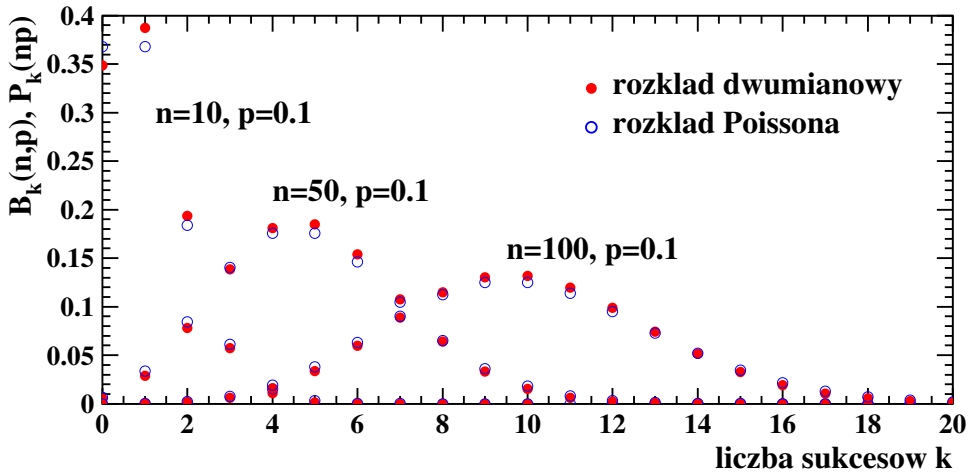
Łączna liczba koincydencji przypadkowych inicjowanych przez detektory A i B:

$$f_{AB} = \lambda_A \left(1 - e^{-\lambda_B T_A}\right) + \lambda_B \left(1 - e^{-\lambda_A T_B}\right) \approx \lambda_A \lambda_B (T_A + T_B)$$

- Rozkład Weibulla: $f(x;\alpha,\sigma) = \frac{\alpha}{\sigma}\left(\frac{x}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\sigma}\right)^{\alpha}\right)$

Rozkład Poissona

- Rozkład Poissona opisuje p-two wystąpienia k zdarzeń w zadanym przedziale czasu Δt jeśli czas oczekiwania na każde kolejne zdarzenie podlega rozkładowi wykładniczemu $Exp(\lambda)$.



- Rozkład Poissona:

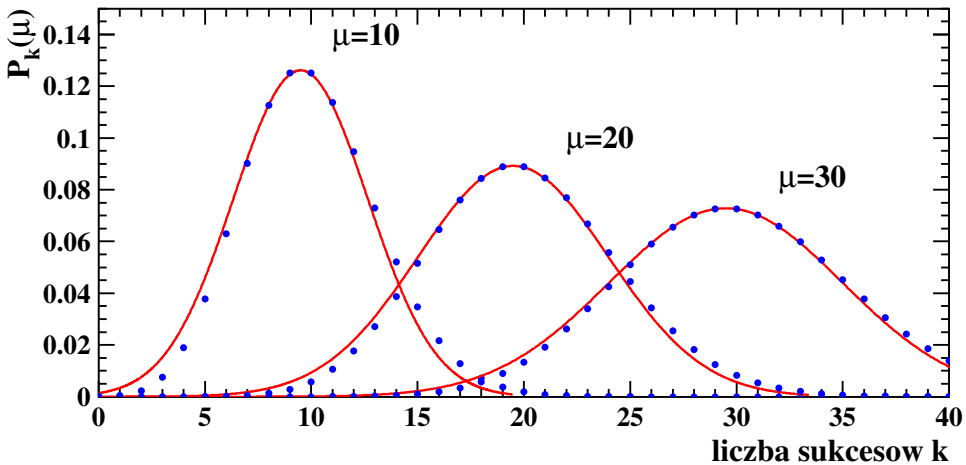
$$\mathcal{P}_k(\mu) = \frac{\mu^k}{k!} e^{-\mu}, \quad \mu > 0,\ k = 0, 1, ...$$

$$\mathcal{E}[\mathrm{k}] = \mu, \quad \mathcal{V}[\mathrm{k}] = \mu$$

Dla $n \to \infty,\ p \to 0,\ np \to \mu$:

$$\mathcal{B}_k(n, p) \longrightarrow \mathcal{P}_k(\mu)$$

- Parametr $\mu = \lambda \Delta t$ określa średnią liczbę zdarzeń w przedziale czasu Δt.
- Rozkład Poissona może także odnosić się do zdarzeń w przestrzeni (λ wtedy oznacza gęstość $[1/\mathrm{m}^n]$).



- Dla dużych wartości μ mamy:

$$\mathcal{P}_k(\mu) \to \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-(\mu-0.5))^2}{2\sigma^2}}$$

gdzie $\sigma = \sqrt{\mu}$.

Rozkład normalny (Gaussa)

- Rozkład normalny (Gaussa):

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

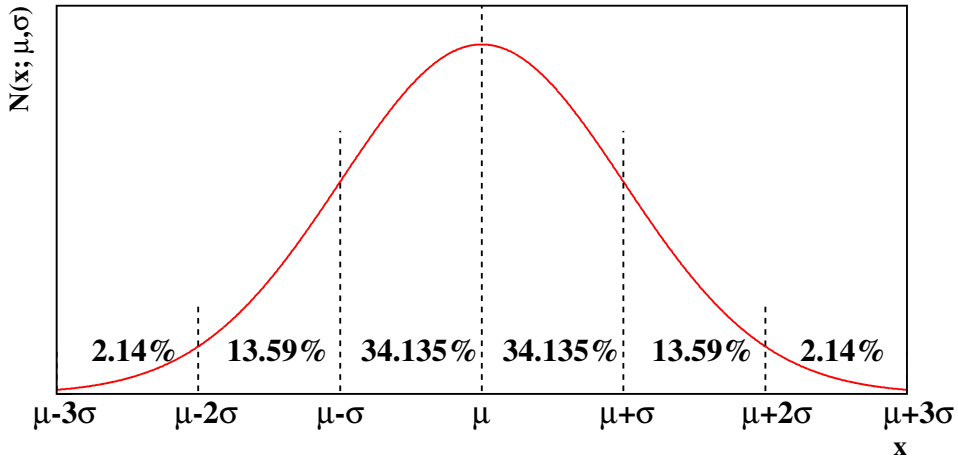
$$\mathcal{E}[\mathrm{x}] = \mu, \qquad \mathcal{V}[\mathrm{x}] = \sigma^2$$

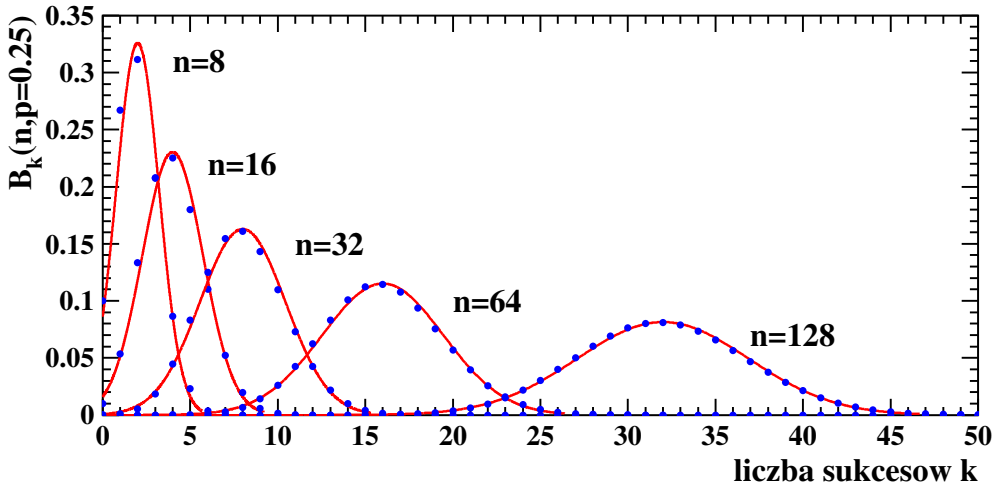
- Dystrybuanta standaryzowanego rozkładu Gaussa:

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\right)\right]$$

gdzie $\mathrm{erf}(x)$ to funkcja błędu:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$





- Dwuwymiarowy rozkład normalny:

$$\mathcal{N}(x,y;\mu_\mathrm{x},\mu_\mathrm{y},\sigma_\mathrm{x},\sigma_\mathrm{y}) =$$

$$= \frac{1}{2\pi\,\sigma_\mathrm{x}\sigma_\mathrm{y}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_\mathrm{x}}{\sigma_\mathrm{x}}\right)^2 + \left(\frac{y-\mu_\mathrm{y}}{\sigma_\mathrm{y}}\right)^2 - 2\rho\left(\frac{x-\mu_\mathrm{x}}{\sigma_\mathrm{x}}\right)\left(\frac{y-\mu_\mathrm{y}}{\sigma_\mathrm{y}}\right)\right]\right\}$$

Rozkład logarytmiczno-normalny

- Wielowymiarowy (m) rozkład normalny:

$$\mathcal{N}(\vec{\mathrm{x}};\vec{\mu},\mathrm{V}) = \frac{1}{(2\pi)^{m/2}|\mathrm{V}|^{1/2}}\,\exp\left[-\frac{1}{2}(\vec{\mathrm{x}}-\vec{\mu})^{\mathrm{T}}\mathrm{V}^{-1}(\vec{\mathrm{x}}-\vec{\mu})\right]$$

- Rozkład logarytmiczno-normalny:

$$p_{\ln}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}\,\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), \quad x>0$$

$$\mathcal{E}[\mathrm{x}] = \exp\left(\mu+\frac{1}{2}\sigma^2\right), \quad \mathcal{V}[\mathrm{x}] = \left[\exp\left(\sigma^2\right)-1\right]\exp\left(2\mu+\sigma^2\right)$$

$$F(x) = \Phi\left(-\frac{\ln x-\mu}{\sigma}\right)$$

- Zastosowanie rozkładu log-normalnego do iloczynów wielu zm. losowych:

  Rozważmy n zmiennych losowych $\mathrm{z}_1,\mathrm{z}_2,...,\mathrm{z}_n$ oraz ich iloczyn $\mathrm{x}=\prod_{i=1}^{n}\mathrm{z}_i$.
  Definiujemy zm. losowe w_i poprzez $\mathrm{z}_i = e^{\mathrm{w}_i}$ oraz $\mathrm{y}=\sum_{i=1}^{n}\mathrm{w}_i$

  Ponieważ $\mathrm{x}=e^{\mathrm{y}} \quad \Rightarrow \quad p(x)=p(y(x))\left|\frac{dy}{dx}\right| = p(y(x))\frac{1}{x}$

  Dla dużego n, na podstawie CTG, zmienna y ma rozkład nornalny, a więc zmienna x ma rozkład log-normalny.

Centralne Twierdzenie Graniczne

- Niech będzie dany ciąg niezależnych zmiennych losowych $\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_n$ pochodzących z dowolnego rozkładu o skończonej wartości oczekiwanej μ oraz wariancji σ^2. Rozkład zmiennej losowej:

$$\mathrm{z}_n = \frac{\overline{\mathrm{x}} - \mu}{\sigma/\sqrt{n}} \qquad \text{gdzie} \qquad \overline{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

dąży do standaryzowanego rozkładu normalnego.