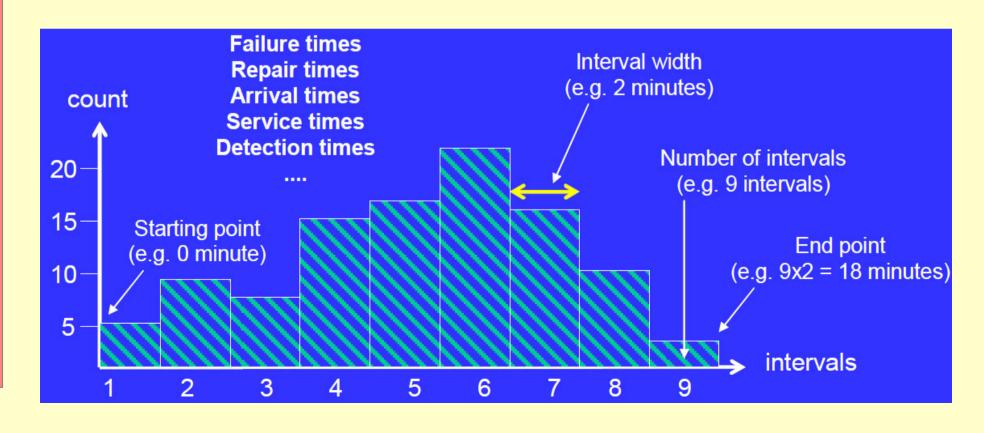
Symulacja komputerowa w informatyce

wykład 5b Rozkłady prawdopodobieństwa

Plan wykładu

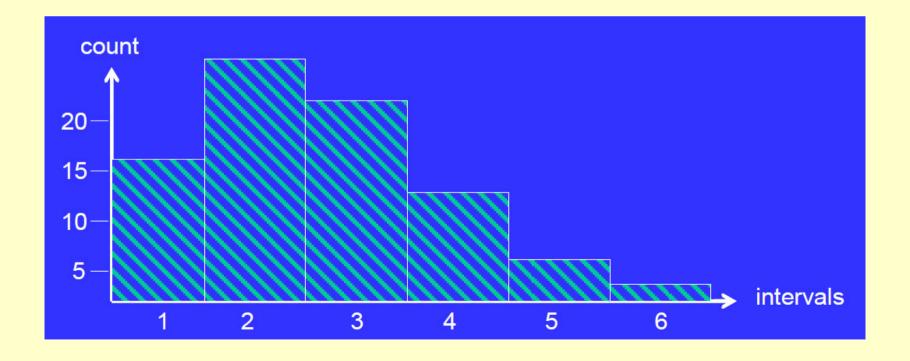
- Histogramy
- Rozkłady prawdopodobieństwa
- Wybór rozkładu prawdopodobieństwa
- Ocenianie poprawności dopasowania

- · graficzne przedstawienie ztabelaryzowanych częstości (zestawienie przedziałów danych i liczby próbek dla nich).
- próbki danych są powszechnie reprezentowane jako liczba wystąpień pewnych zdarzeń lub zakończenie procesu.

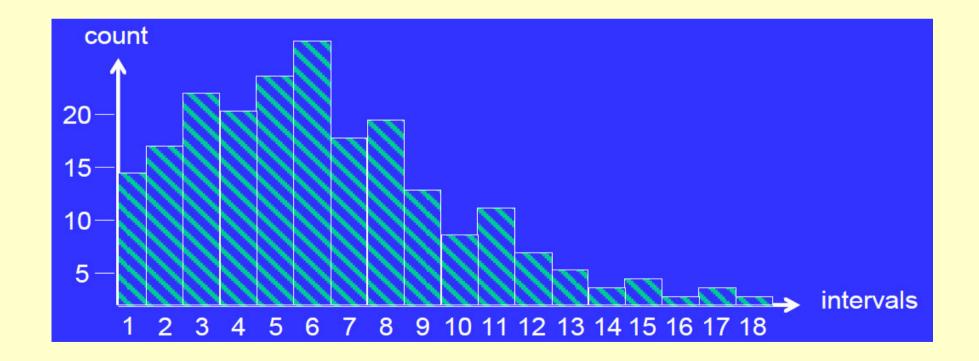


- · nie ma zdefiniowanej reguły, wyboru poprawnych parametrów histogramu.
- · do wykonania (z powtórzeniami):
 - dopasowanie punktu startowego (początkowego),
 - dopasowanie szerokości przedziału,
 - ustalenie liczby przedziałów obejmujących wszystkie dane.
- · Wybierz odpowiedni histogram dla reprezentacji próbek danych.

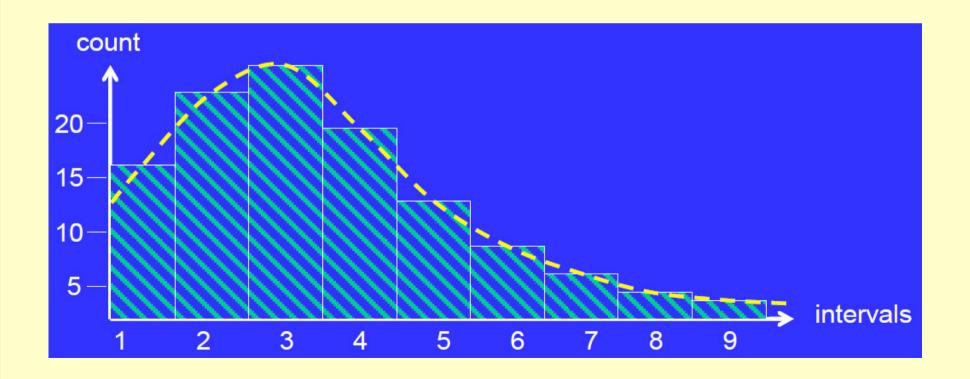
- · Jeżeli szerokości przedziałów są za duże:
 - Wykres będzie mało dokładny
 - Szczegóły dotyczące kształtu danych zostaną utracone.



- · Jeżeli szerokości przedziałów są za małe,
 - Wykres będzie zbyt zakłócony (zaszumiony)
 - Ogólny zarys kształtu danych zostanie utracony.

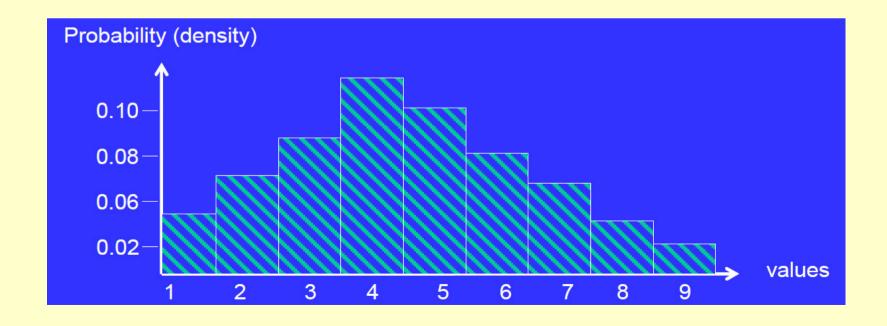


- · Nie istnieje najlepszy histogram.
- · Najlepiej spróbować umieścić co najmniej 3 do 5 próbek w każdym przedziale.



Rozkłady prawdopodobieństwa

- · Opisuje wartości i prawdopodobieństwa związane ze zdarzeniami losowymi:
- funkcja rozkładu prawdopodobieństwa,
- funkcja gęstości.

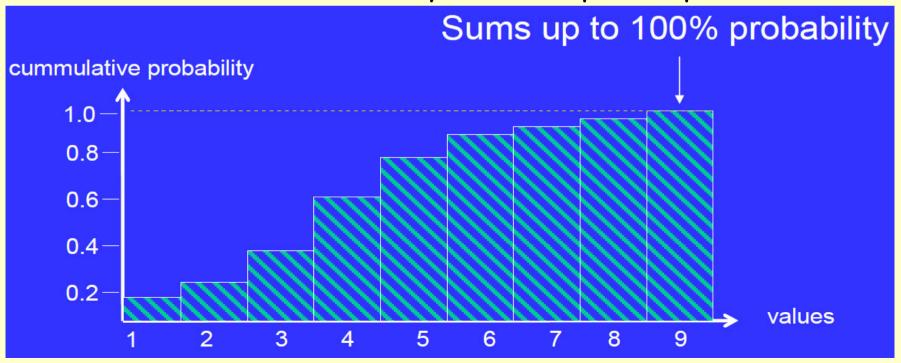


Rozkłady prawdopodobieństwa

Dystrybuanta rozkładu

· Opisuje wartości i narastające prawdopodobieństwo związane ze zdarzeniem losowym.

Sumy do 100% prawdopodobieństwa



(W odniesieniu do wartości)

- · Rodzaje według liczby wartości:
 - rozkłady dyskretne:
 - skończona lub przeliczalna liczba różnych wartości
 - rozkłady ciągłe:
 - · nieprzeliczalna ilość różnych wartości
- · Rodzaje według zakresu wartości:
 - Rozkłady nieujemne
 - Rozkłady ograniczone
 - Rozkłady nieograniczone

(W odniesieniu do sposobu reprezentacji)

- następujące rozkłady prawdopodobieństwa są powszechnie stosowane do modelowania wejścia symulacji:
- Rozkłady standardowe
- Rozkłady empiryczne

Rozkłady standardowe

Rozkłady parametrów matematycznych,
 które zwykle mają parametry lokalizacji i skali,
 oraz zero, jeden lub dwa parametry kształtu.

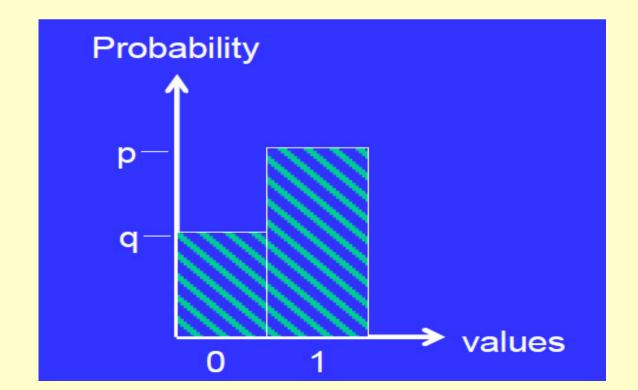
[exa rys]

Powszechnie używane rozkłady standardowe

Nonnegative Cont.	Nonnegative Dis.	Unbounded cont.	Bounded Cont.	Bounded Dis.
Chi-square	Geometric	Cauchy	Beta	Bernoulli
Erlang	Logarithmic	Error	Johnson S _B	Binomial
Exponential	Nagative binomial	Exponential power	Power function	Uniform
F	Poisson	Extreme value	Triangular	
Gamma		Johnson S _u	Uniform	
Inverse gaussian		Laplace		
Inverted weibull		Logistic		
Log-laplace		Normal		
Log-normal		Pareto		
Pearson type 5		Student's t		
Pearson type 6				
Random walk				
Rayleigh				
Wald				
Weibull				

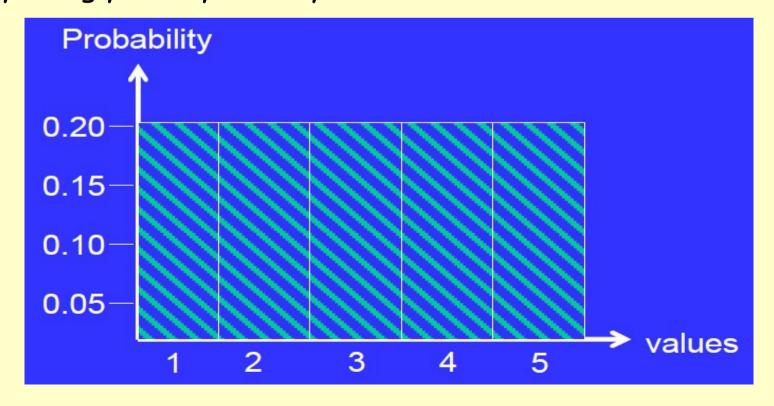
Rozkład Bernoulli'ego (dwupunktowy)

- dyskretny ograniczony rozkład prawdopodobieństwa, który przyjmuje;
- Wartość 1 z prawdopodobieństwem sukcesu p,
- Wartość O z prawdopodobieństwem porażki q = 1 p.



Rozkład jednorodny (inne nazwy: jednostajny, prostokątny, równomierny)

- ograniczony rozkład prawdopodobieństwa, który przyjmuje wartości pomiędzy a i b, gdzie a<b, i prawdopodobieństwa wszystkich wartości są równe.
- · Może być ciągły lub dyskretny.



Rozkład trójkątny

ograniczony ciągły rozkład prawdopodobieństwa
 z dolnym limitem równym a, maksimum w c, a górny limit w b.

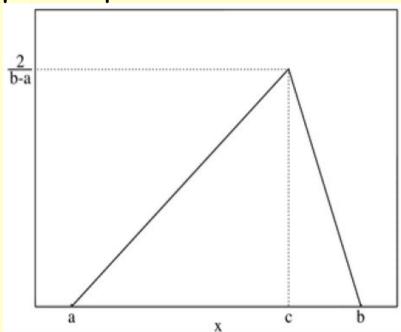
$$f(x|a,b,c) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c \le x \le b \end{cases}$$

$$0 & \text{for any other case}$$

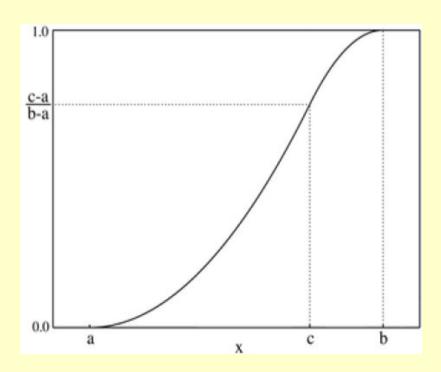
Rozkład trójkątny

ograniczony ciągły rozkład prawdopodobieństwa
 z dolnym limitem równym a, maksimum w c, a górny limit w b.

funkcja gęstości prawdopodobieństwa

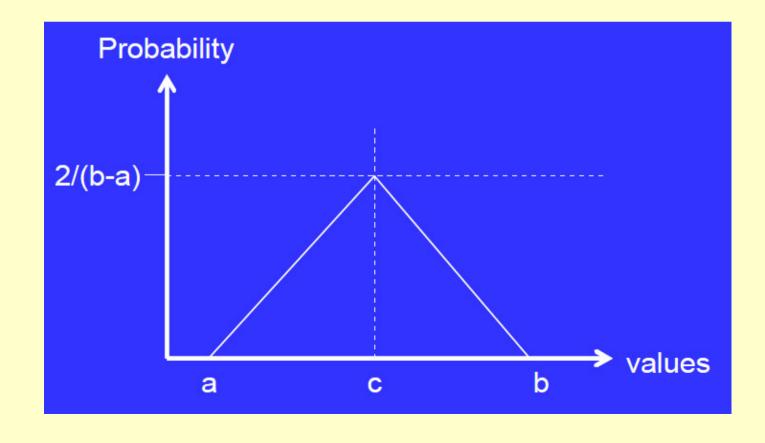


dystrybuanta



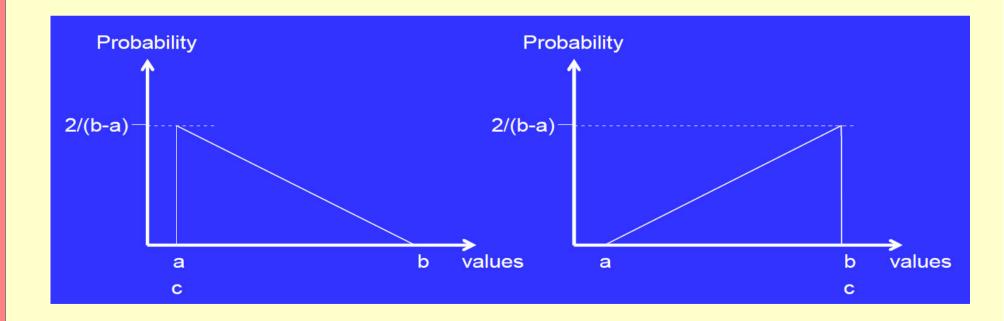
Rozkład trójkątny symetryczny

 Rozkład trójkątny z c znajdującym się w centrum pomiędzy a i b.



Rozkład trójkątny dwu-punktowy

 Rozkład trójkątny ze znanymi a i b, gdzie c jest równe a lub b.

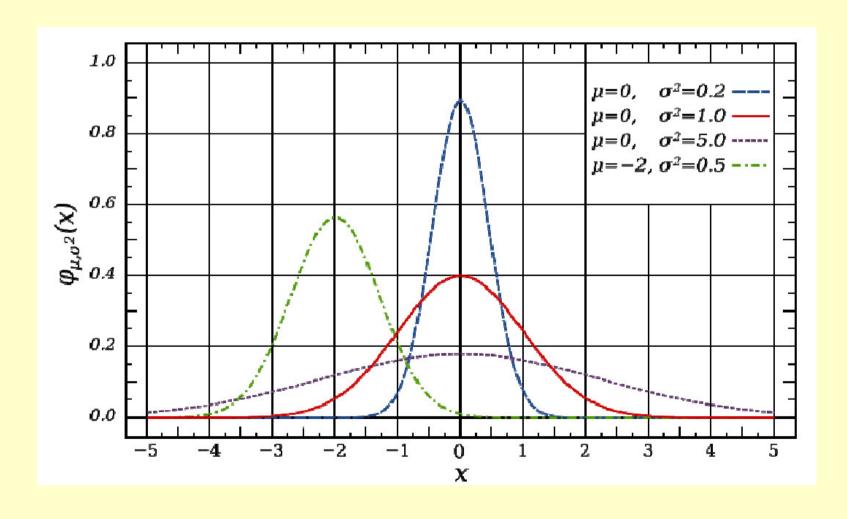


Rozkład Normalny (Gaussa)

- · ważna rodzina rozkładów prawdopodobieństwa ciągłych nieograniczonych, mająca zastosowanie na wielu polach.
- Zdefiniowana przez dwa parametry:
 - Lokalizacja: μ, średnia
 - Skala: σ², wariancja
 (kwadrat odchylenia standardowego)
- · Standardowy rozkład normalny:
 - Rozkład normalny z
 - średnią= 0 i wariancją =1.

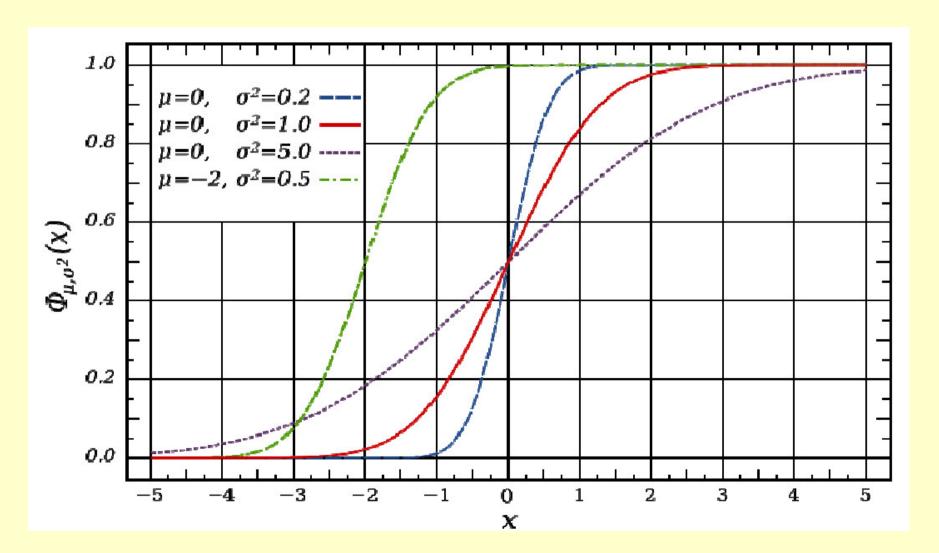
Rozkład Normalny (Gaussa)

Funkcja gęstości prawdopodobieństwa w kształcie dzwonu



Rozkład Normalny (Gaussa)

Dystrybuanta rozkładu normalnego



Rozkład Normalny (Gaussa)

- · Ważne:
 - Model zjawisk ilościowych w naukach przyrodniczych i społecznych. (*)
 - Wiele pomiarów, od psychologicznych do zjawisk fizycznych, może być w przybliżane [aproksymowane] w różnym stopniu poprzez rozkład normalny.
 - Najbardziej powszechnie używana rodzina rozkładów w statystyce.
 - Wiele testów statystycznych jest opartych na założeniu normalności.

Rozkład Normalny (Gaussa)

Funkcja gęstości prawdopodobieństwa

 σ > 0 jest odchyleniem standardowym μ jest średnią (wartość oczekiwana)

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \quad x\in\mathbb{R},$$

Rozkład Normalny (Gaussa)

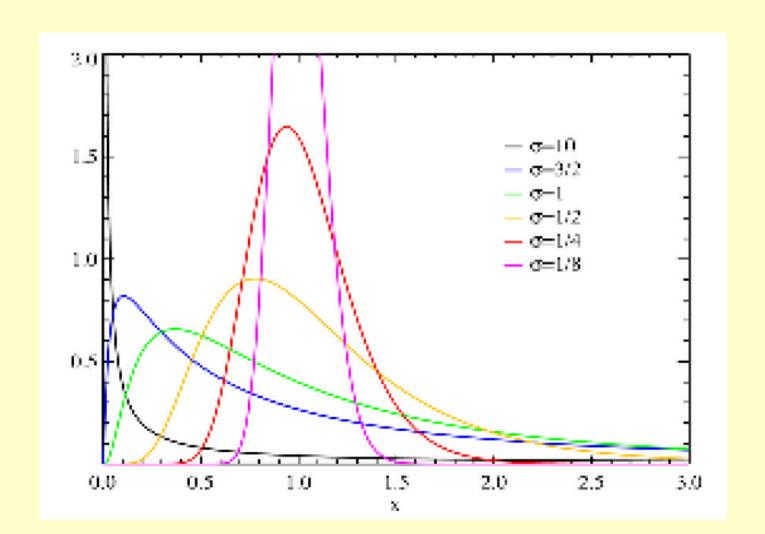
Podstawowy mechanizm tworzący rozkład normalny można wyobrazić sobie jako nieskończoną liczbę niezależnych zdarzeń losowych (dwumianowych), które generują wartości danej zmiennej. Przykładowo, istnieje prawdopodobnie prawie nieograniczona liczba czynników determinujących wzrost człowieka (olbrzymia liczba genów, sposób odżywiania, przebyte choroby itd.). Tak więc należy spodziewać się, że w populacji wzrost podlega rozkładowi normalnemu.

Rozkłady logarytmiczno-normalne

- Rozkłady prawdopodobieństwa nieujemne ciągłe posiadające pojedyncze "ramię" dla każdej zmiennej losowej, której logarytm jest rozkładem normalnym.
- · Zdefiniowane przez dwa parametry:
- Średnia: μ
- Odchylenie standardowe: σ

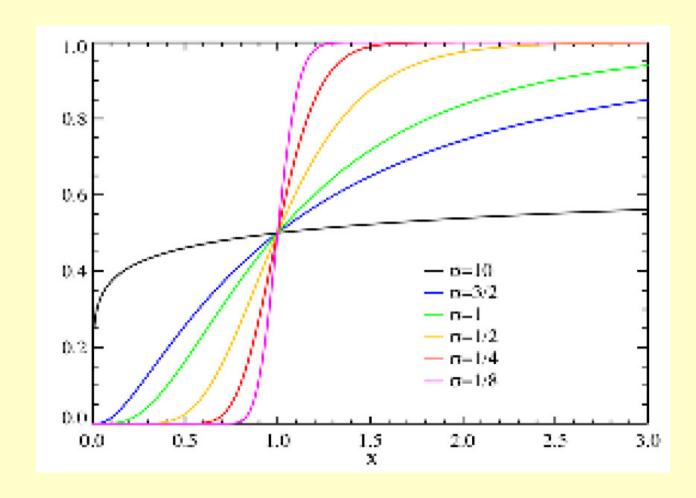
Rozkłady logarytmiczno-normalne

Funkcja gęstości prawdopodobieństwa ($\mu = 0$)



Rozkłady logarytmiczno-normalne

Dystrybuanta rozkładu ($\mu = 0$)



Rozkłady logarytmiczno-normalne

Funkcja gęstości prawdopodobieństwa dla x>0 $\sigma>0$ to odchylenie standardowe μ jest średnią

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

Rozkłady logarytmiczno-normalne

Rozkład logarytmicznie normalny jest często lepszym od rozkładu normalnego przybliżeniem rozkładów cech, w których istotne są stosunki pomiędzy wartościami, a nie różnice pomiędzy nimi. Na przykład przybliżony rozkład logarytmicznie normalny mają kursy akcji giełdowych, gdzie ważniejsze jest o ile procent zmniejszyła się lub zwiększyła wartość akcji, a nie o ile złotych.

Rozkład tego typu jest często wykorzystywany do modelowania rozkładu takich zmiennych jak dochody osobiste lub wiek w momencie zawierania pierwszego małżeństwa.

Rozkłady wykładnicze

- Rozkłady prawdopodobieństwa nieujemne ciągłe z parametrem (lambda) A, który opisuje czasy między zdarzeniami w procesie Poissona.
- Występuje naturalnie opisując długości międzyczasów przybyć (wejść) wydarzeń w jednorodnym procesie Poissona.

....

Rozkłady wykładnicze

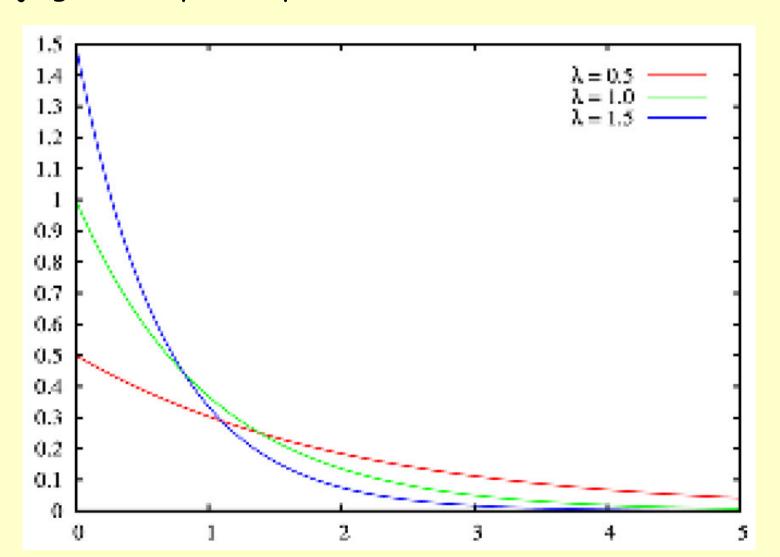
- Proces Poissona:
- Proces, w którym zdarzenia występują ciągłe, niezależnie od siebie ze stałą średnią częstością.
- · Zdefiniowana przez jeden parametr:
- A > 0: często nazywany parametrem rate (częstości, wzrostu, stopa).

Rozkłady wykładnicze

Przyjmijmy, że T oznacza czas pomiędzy kolejnymi wystąpieniami rzadkiego zdarzenia, które zachodzi średnio λ razy na jednostkę czasu. Wówczas T podlega rozkładowi wykładniczemu z parametrem λ (lambda). Rozkład ten jest często wykorzystywany do modelowania przedziałów czasu pomiędzy kolejnymi zdarzeniami losowymi. Przykłady zmiennych podlegających temu rozkładowi to: odstęp czasu pomiędzy przejazdami samochodów przez skrzyżowanie, czas bezawaryjnej pracy urządzeń elektronicznych lub czas pojawienia się klientów w sklepie spożywczym.

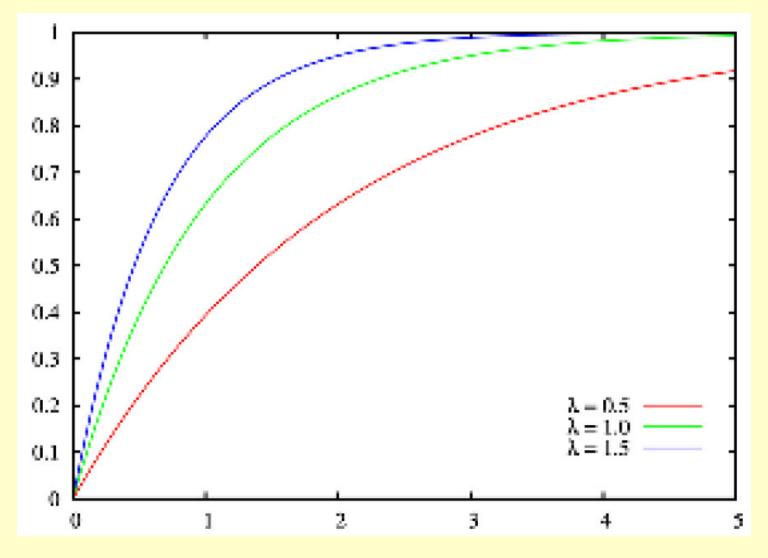
Rozkłady wykładnicze

Funkcja gęstości prawdopodobieństwa



Rozkłady wykładnicze

Dystrybuanta rozkładu



Rozkłady wykładnicze

Funkcja gęstości prawdopodobieństwa

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} &, x \ge 0, \\ 0 &, x < 0. \end{cases}$$

A > 0 jest parametr rate

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} &, x \ge 0, \\ 0 &, x < 0. \end{cases}$$

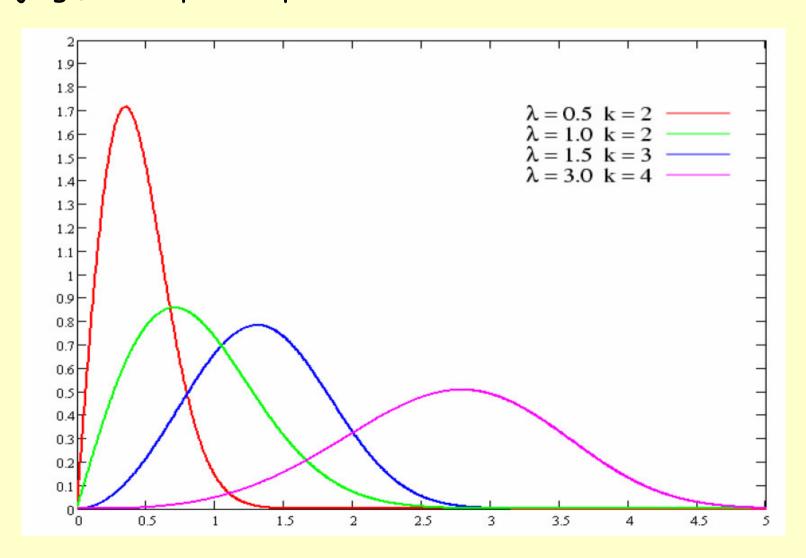
β > 0 jest parametrem skali rozkładu i jest odwrotnością parametru λ rate.

Rozkłady Weibulla

- Rozkłady prawdopodobieństwa nieujemne ciągłe używane do opisania wielkości rozkładu cząstek.
- Zdefiniowane przez dwa parametry :
 - Kształt: k
 - Skala: A

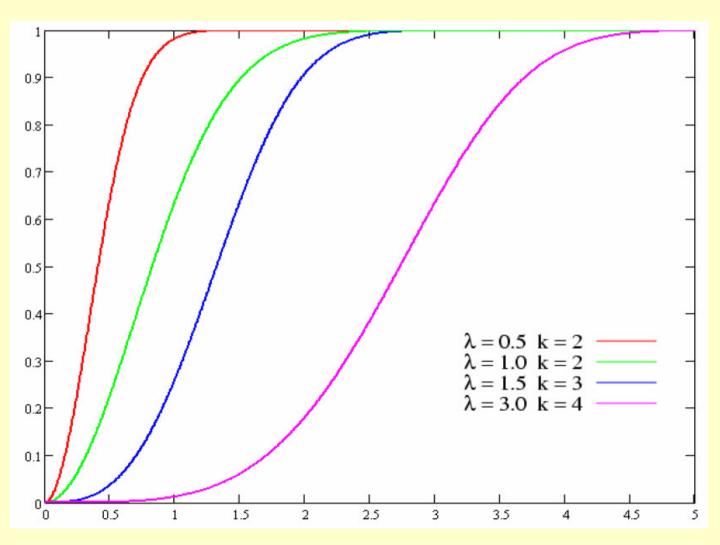
Rozkłady Weibulla

Funkcja gęstości prawdopodobieństwa



Rozkłady Weibulla

Dystrybuanta rozkładu



Rozkłady Weibulla

Funkcja gęstości prawdopodobieństwa

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

k to kształt A to skala

Gdy k = 1, rozkład Weibulla redukuje się do rozkładu wykładniczego. Gdy k = 3,4 rozkład Weibulla wydaje się podobny do rozkładu normalnego.

Rozkład Weibulla

Rozkład wykładniczy jest często wykorzystywany do modelowania pomiarów czasu bezawaryjnej pracy w sytuacji gdy wskaźnik defektów (ryzyko) jest stały w określonym przedziale czasu.

W przypadku gdy prawdopodobieństwo defektu zmienia się w czasie bardziej stosownym typem rozkładu jest rozkład Weibulla.

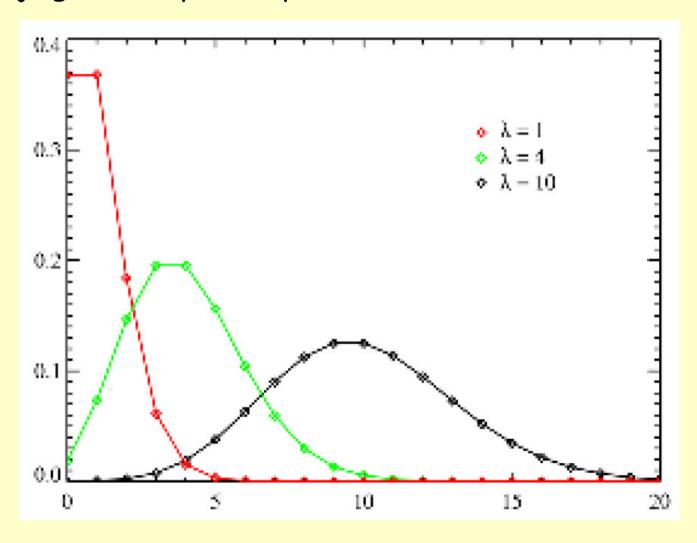
Dlatego też rozkład Weibulla jest często stosowany do oceny niezawodności (np. przekaźników elektronicznych, łożysk kulkowych, itp.

Rozkłady Poissona

- nieujemny rozkład dyskretny.
- wyraża prawdopodobieństwo liczby zdarzeń zachodzących w ustalonym okresie czasu.
- Skupia się na liczbie wystąpień zdarzeń dyskretnych (nazywanych "wejściami"), które mają miejsce w określonym odstępie czasowym (o określonej długości).
- · Zdefiniowane przez dwa parametry:
- k: liczba wystąpień zdarzenia
- (lambda) λ > 0: oczekiwana liczba wystąpień zdarzenia w określonym przedziale czasowym.

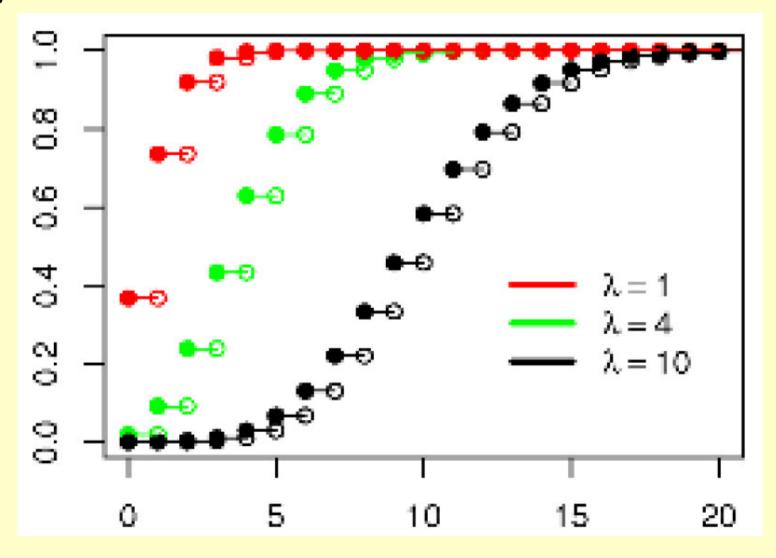
Rozkłady Poissona

Funkcja gęstości prawdopodobieństwa



Rozkłady Poissona

Dystrybuanta rozkładu



Rozkłady Poissona

Funkcja gęstości prawdopodobieństwa

Prawdopodobieństwo, że pojawi się dokładnie k wystąpień zdarzenia jest równa:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

k: liczba wystąpień zdarzenia $\lambda > 0$: przewidywana liczba zdarzeń w określonym przedziale czasowym.

Rozkłady Poissona

Rozkład Poissona jest czasami nazywany rozkładem zdarzeń rzadkich. Przykłady zmiennych o rozkładzie Poissona to: liczba wypadków na osobę, liczba wygranych w Toto-Lotku lub liczba awarii występujących w procesie produkcyjnym.

Praca Poissona skupiała się na niektórych zmiennych losowych N, wyrażających, między innymi, liczbę dyskretnych zdarzeń, które odbywają się w przedziale czasu, o określonej długości. Rozkład Poissona może być stosowany do systemów z dużą liczbą możliwych zdarzeń, z których każde jest bardzo rzadkie. Klasycznym przykładem jest rozpad jąder atomowych.

Rozkłady Poissona

Rozkład ten

wyraża prawdopodobieństwo szeregu wydarzeń mających miejsce w określonym czasie, gdy te wydarzenia występują ze znaną średnią częstotliwością i w sposób niezależny od czasu jaki upłynął od ostatniego zajścia takiego zdarzenia.

Rozkład Poissona jest czasami nazywany prawem małych liczb, ponieważ jest to rozkład prawdopodobieństwa ilości wystąpień zdarzenia, które zdarza się rzadko ale ma bardzo wiele możliwości, aby się zdarzyć.

Rozkłady empiryczne

- alternatywą dla standardowych rozkładów jest użycie rozkładów empirycznych.
- funkcja dystrybuanty rozkładu,
 która przypisuje prawdopodobieństwo 1/n dla każdego elementu próbki, które zawierają n jako liczbę próbek.

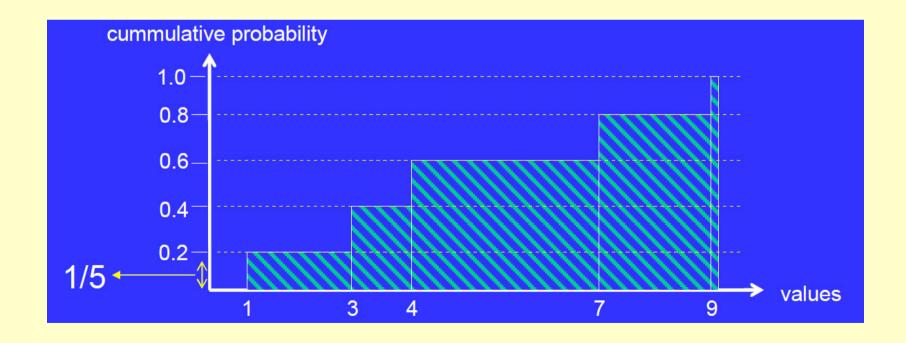
Rozkłady empiryczne

- · W postaci ogólnej:
 - Prawdopodobieństwo o wartości mniejszej lub równej x (innymi słowy dystrybuanta x)

$$F_n(x) = \frac{\text{number of elements in the sample} \le x}{n} = \frac{1}{n} \sum_{i=1}^n I(X_i \le x),$$

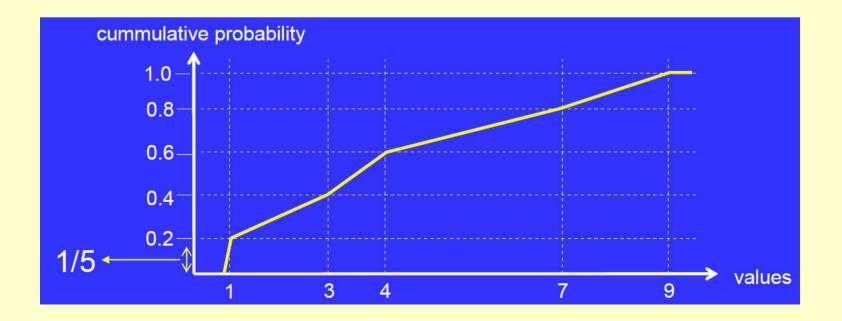
Rozkłady empiryczne

 ogólna forma to krok funkcji, który rośnie przy każdej unikalnej wartości próbki obserwowanej proporcjonalnie do całkowitej ilości takich wartości.



Rozkłady empiryczne

 alternatywa forma wykorzystywana dla rozkładów ciągłych zastępująca kroki, liniową interpolacją pomiędzy kolejnymi punktami.

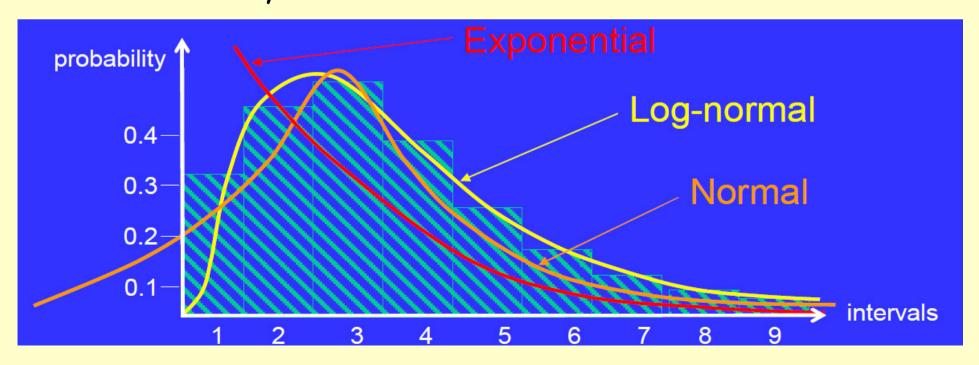


Rozkłady empiryczne

- · Wady:
- może reprezentować jedynie ograniczone rozkłady w obserwowanym zakresie danych.
 - · Możemy dodawać oszacowane ramię na początku i końcu.
- jakość reprezentacji jest całkowicie zależna od jakości dostępnej próbki.
- prawdopodobieństwo, że historia się powtórzy jest dokładnie równe zero.

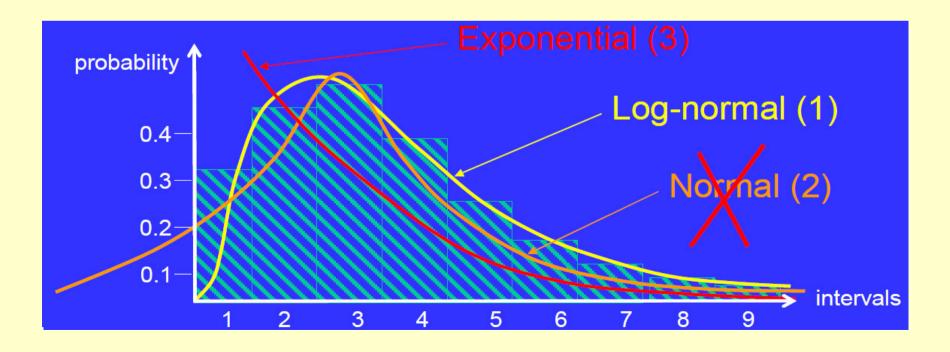
Wybór rozkładu prawdopodobieństwa

- Wykorzystanie wiedzy o losowości do określenia jakiś granic wartości, które może on wygenerować.
- Dopasowanie jak największej liczby standardowych rozkładów do danych.



Wybór rozkładu prawdopodobieństwa

- Użyj zestawu kryteriów oceniającym poprawność dopasowania wybranego rozkładu do danych.
- Jeśli którykolwiek z często używanych modeli rozkładów jest niezgodny w zakresie limitów wartości, należy ich rodzinę wykluczyć.



Wybór rozkładu prawdopodobieństwa

- Użyj rozsądnego zestawu kryteriów, w celu określenia czy najlepiej dopasowany rozkład jest rozsądną reprezentacją danych.
- Jeśli najlepszy zapewnia rozsądną reprezentację danych,
 - Użyj go w symulacji,
- · W przeciwnym razie,
 - Użyj rozkładu empirycznego do bezpośredniej reprezentacji danych.

Ocena poprawności dopasowania

- Należy rozważyć szereg miar poprawności dopasowania zamiast tylko jednej
 - Ponieważ każda będzie nierzetelna w niektórych przypadkach.
- · Nie bądź zależny od miar poprawności dopasowania
 - nie polegaj na niezbyt czystych próbkach danych (np. ignorować próbki problematyczne)
 - uwaga na parametry dostarczone przez użytkownika (np. konfiguracje histogramu).
 - miary mogą dostarczyć niezgodne wyniki.

Ocena poprawności dopasowania

- · w kontekście modelowania wejścia symulacji:
 - Klasyczne metody poprawności dopasowania w statystyce nie są całkowicie odpowiednie dla ostatecznej oceny jakości dopasowania.
 - metody statystyczne mają określone założenia, które czasami nie są prawdziwe/poprawne dla modelowania symulacji.
 - graficzne metody heurystyczne powinny być również stosowane do oceny: która jest najlepsza, i która jest wystarczająco dobra.

Ocena poprawności dopasowania

- Może być użyta do oceny:
 - Histogramów
 - Funkcji dystrybuanty empirycznej przykładowych danych.

Ocena poprawności dopasowania

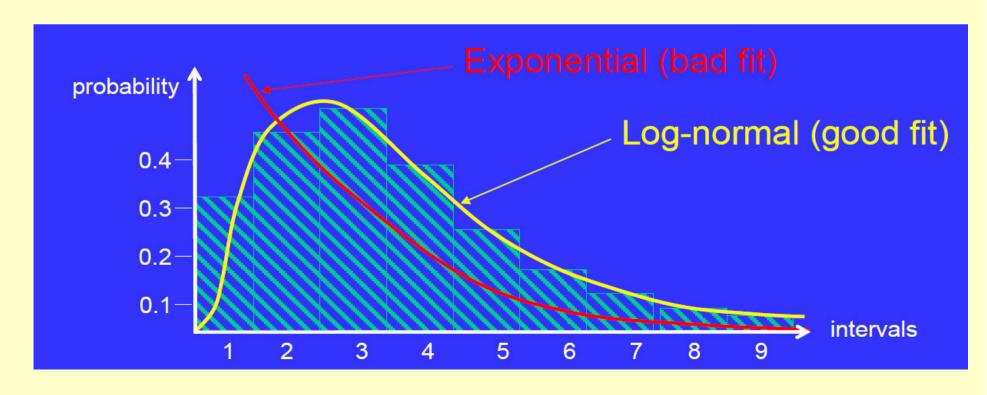
Histogramy

- Histogram jest to oszacowanie funkcji gęstości prawdopodobieństwa zdarzeń w systemie świata rzeczywistego.
- Jest więc uzasadnione, aby go wykreślić i porównać funkcje gęstości dopasowanych modeli dla danego histogramu.

Ocena poprawności dopasowania

Histogramy

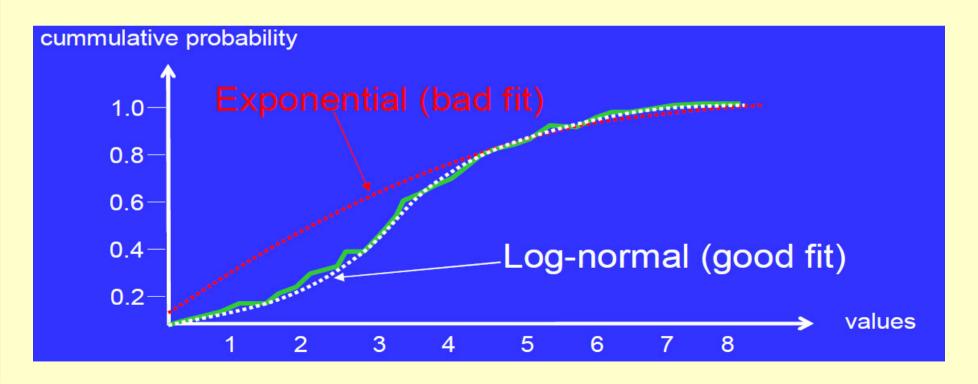
 Histogramy przedstawiają przykładowe próbki, ale można je łatwo przekształcić do jednostek gęstości przez podzielenie przez całkowitą liczbę próbek.



Ocena poprawności dopasowania

Funkcja dystrybuanty

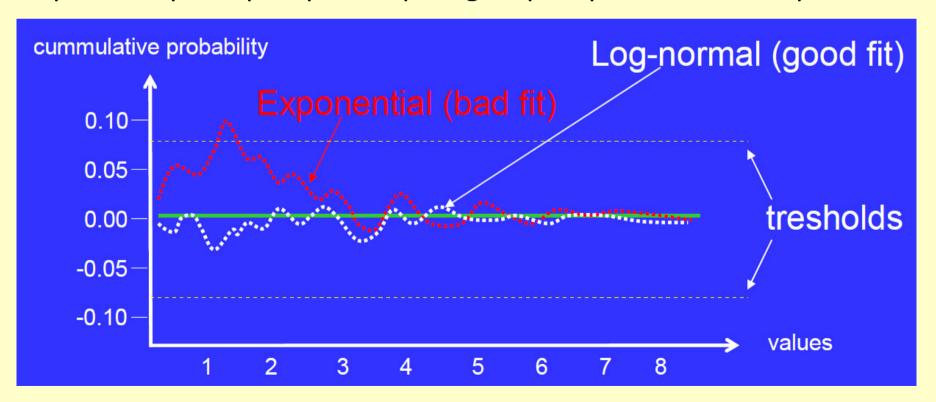
 Funkcja dystrybuanty empirycznej przykładowych danych może być użyta do porównania z dopasowanym modelem.



Ocena poprawności dopasowania

Funkcja dystrybuanty

 Różnice (błędy) między częścią empiryczną i modelem dopasowanym dystrybuanty mogą być wykreślone na wykresie.



Ocena poprawności dopasowania

Statystyka testowa

- Statystyka testowa:
 - Procedura operacyjna poprawności dopasowania
 - W celu obliczenia funkcji obserwowanych danych i dopasowanego modelu,
 - Porównanie błędów o wartości krytycznej, aby zaakceptować lub odrzucić hipotezę dopasowania.
- Odrzucenie oznacza, że nie istnieją wystarczające dowody aby stwierdzić, że oba rozkłady pasują do siebie.
- Jeśli nie ma ekstremalnych błędów, lepiej nie odrzucać, ale określić poziom dopasowania, aby znaleźć najlepsze lub wystarczająco dobre rozwiązanie.

Ocena poprawności dopasowania

Statystyka testowa: Chi-kwadrat

- Dzielimy zakres danych próbki na M przedziałów (podobne do histogramów).
- · Pierwszy i ostatni przedział może być przedłużony do
- [+ /] nieskończoności, aby pokryć cały zakres zmiennej losowej.
- Oblicz różnice między danymi przykładowymi i dopasowywanym modelem (błędy) w każdym przedziale.
- Zsumować błędy wszystkich przedziałów, aby otrzymać łączny błąd.
- · Test ten jest powszechnie stosowany.

Ocena poprawności dopasowania

Statystyka testowa: Chi-kwadrat

M = liczba przedziałów

T_s = Liczba próbek

O_k = liczba próbek, w przedziale k-ym

 E_k = Prawdopodobieństwo dla k-ego przedziału dopasowanego

modelu x T

Total error =
$$\sum_{k=1}^{M} \frac{(O_k - E_k)^2}{E_k}$$

$$E_k$$
Błąd kwadratowy

Waga, ta jest odwrotnie proporcjonalna do ilości oczekiwanych punktów (większa waga wstawiana gdy rzadkie zdarzenia).

Ocena poprawności dopasowania

Statystyka testowa: Chi-kwadrat

- Przez zmianę konfiguracji przedziałów mogą być generowane sprzeczne wyniki.
- W związku z tym, wyniki nie powinny budzić ufności pojedynczo (trzeba sprawdzić kilka konfiguracji).

Generatory liczb pseudolosowych

Literatura:

- 1) Tyszer Jerzy: Symulacja cyfrowa, WNT, Warszawa 1990
- 2) Banks J.:Introduction To Simulation, Proceedings of the 1999 Winter Simulation Conference (WSC),
- 3) Dr.Çağatay ÜNDEĞER: Modeling & Simulation, Bilkent Üniversitesi, Ankara 2008,
- 4) Kotulski Zbigniew: "Generatory liczb losowych: algorytmy, testowanie, zastosowania", MATEMATYKA STOSOWANA 2, 2001
- 5) Jarnicki J.: Komputerowe generatory liczb losowych, Polit. Wrocław.
- 6) Wikipedia (Rozkłady prawdopodobieństwa)

Generatory liczb pseudolosowych

Wykład powstał na podstawie:

Prof. Tadeusz Nowicki: Symulacja komputerowa - materiały dydaktyczne, Kielce 2011

Dr.Çağatay ÜNDEĞER: Modeling & Simulation, Bilkent Üniversitesi, Ankara 2008