Komputerowa Analiza Danych Doświadczalnych

Prowadząca:
dr inż. Hanna Zbroszczyk

e-mail: *gos@if.pw.edu.pl*
tel: +48 22 234 58 51
konsultacje: środa: 14-15, piątek: 13-14

www: http://www.if.pw.edu.pl/~gos/students/kadd

Politechnika Warszawska
Wydział Fizyki
Pok. 117b (wejście przez 115)

POBIERANIE PRÓBY

(z rozkładów cząstkowych)

Pobieranie próby z rozkładów cząstkowych

Populację G dzielimy na **pod-populacje**: G_1, G_2, ..., G_t.

Wielkość ***x*** opisana **gęstościami** z pod-populacji: $f_1(x)$, $f_2(x)$, ..., $f_t(x)$.

Dystrybuanty:

$$F_i(x) = \int_{-\infty}^{x} f_i(x)\,dx = P(\boldsymbol{x} < x | \boldsymbol{x} \in G_i)$$

charakter **prawdopodobieństw warunkowych**

(x musi należeć do odpowiedniej pod-populacji).

Prawdopodobieństwo całkowite:

$$F(x) = P(\boldsymbol{x} < x | \boldsymbol{x} \in G) = \sum_{1}^{t} P(\boldsymbol{x} < x | \boldsymbol{x} \in G_i) P(\boldsymbol{x} \in G_i)$$

$$F(x) = \sum_{1}^{t} P(\boldsymbol{x} \in G_i) F_i(x)$$

Gęstość prawdopodobieństwa:

$$f(x) = \sum_{1}^{t} P(\boldsymbol{x} \in G_i) f_i(x)$$

Pobieranie próby z rozkładów cząstkowych

Niech: $P(\boldsymbol{x} \in G_i) = p_i$

$$\hat{x} = E(\boldsymbol{x}) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \sum_{i=1}^{t} p_i \int_{-\infty}^{\infty} x\, f_i(x)\, dx$$

$$\hat{x} = \sum_{i=1}^{t} p_i \hat{x}_i$$

Wartość oczekiwana z populacji jest średnią z wartości oczekiwanych z poszczególnym pod-populacji mnożonych przez ich prawdopodobieństwa.

Wariancja:

$$\sigma^2(x) = \int_{-\infty}^{\infty} (x - \hat{x})^2 f(x)\, dx = \int_{-\infty}^{\infty} (x - \hat{x})^2 \sum_{i=1}^{t} p_i f_i(x)\, dx$$

$$\sigma^2(x) = \sum_{i=1}^{t} p_i \int_{-\infty}^{\infty} \{(x - \hat{x}_i) + (\hat{x}_i - \hat{x})\}^2 f_i(x)\, dx$$

co prowadzi do:

$$\sigma^2(x) = \sum_{i=1}^{t} p_i \int_{-\infty}^{\infty} \{(x - \hat{x}_i) + (\hat{x}_i - \hat{x})\}^2 f_i(x)\, dx$$

(zmienne niezależne, więc znikają kowariancje)

$$\sigma^2(x) = \sum_{i=1}^{t} p_i \{\sigma_i^2 + (\hat{x}_i - \hat{x})^2\}$$

Wariancja (średnia ważona z wariancji dla pod-populacji i wariancji wartości średniej z podpopulacji względem wartości średniej z populacji

Pobieranie próby z rozkładów cząstkowych

Każdej spod-populacji wyodrębniamy **próbkę** o liczebności n_i.

$$\sum_{i=1}^{t} n_i = n$$

Średnia arytmetyczna (z całej próby):

$$\bar{x}_p = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{t} n_i \bar{x}_i$$

Wartość oczekiwana i wariancja:

$$E(\bar{x}_p) = \boxed{\frac{1}{n}\sum_{i=1}^{t} n_i \hat{x}_i = E(\bar{x}_p)}$$

$$\sigma^2(\bar{x}_p) = E\{(\bar{x}_p - E(\bar{x}_p))^2\}$$

$$\sigma^2(\bar{x}_p) = E\{(\sum_{i=1}^{t} \frac{n_i}{n}(\bar{x}_i - \hat{x}_i))^2\}$$

$$\sigma^2(\bar{x}_p) = \frac{1}{n^2}\sum_{i=1}^{t} n_i^2 E\{(\bar{x}_i - \hat{x}_i)^2\}$$

$$\sigma^2(\bar{x}_p) = \frac{1}{n^2}\sum_{i=1}^{t} n_i^2 \sigma^2(\bar{x}_i) = \boxed{\frac{1}{n}\sum_{i=1}^{t} \frac{n_i}{n}\sigma_i^2 = \sigma^2(\bar{x}_p)}$$

Pobieranie próby z rozkładów cząstkowych

Średnia arytmetyczna nie może być estymatorem wartości średniej z populacji

(zależy od dowolnego wyboru liczebności poszczególnych pod-populacji).

Jednak, kiedy spełniony jest warunek: $p_i = \frac{n_i}{n}$

Wartość średnia może być estymowana poprzez wartości średnie poszczególnych prób wewnątrz pod-populacji:

$$\tilde{x} = \sum_{i=1}^{t} p_i \hat{x}_i$$

Zgodnie z prawem propagacji niepewności:

$$\sigma^2(\tilde{x}) = \sum_{i=1}^{t} p_i^2 \sigma^2(\bar{x}_i) = \sum_{i=1}^{t} \frac{p_i^2}{n_i} \sigma_i^2$$

Optymalny wybór **wielkości próby wewnątrz pod-populacji**:

$$n_i = \frac{n\, p_i \sigma_i}{\sum_{i=1}^{t} (p_i \sigma_i)}$$

POBIERANIE PRÓBY ZE SKOŃCZONEJ POPULACJI BEZ UZUPEŁNIEŃ

Pobieranie próby ze skończonej populacji bez uzupełnień.

- Populacja **N-elementowa**: $y_1, y_2, ..., y_N$.

- Pobierana **próba n o elementach**: $x_1, x_2, ..., x_n$.

- **Jednakowe prawdopodobieństwo** wybrania jakiegokolwiek elementu populacji:

$$E(y) = \hat{y} = \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$$

- **Średnia arytmetyczna z populacji ze skończoną liczbą elementów nie jest zmienną losową!**

- Wariancja: $\sigma^2(y) = \frac{1}{1-N}\sum_{j=1}^{N}(y_j - \bar{y})^2$

$$\sigma^2(y) = \frac{1}{1-N}\left\{\sum_{j=1}^{N} y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{N} y_j\right)^2\right\}$$

Odchylenie średnie kwadratowe, stopnie swobody

Suma kwadratów: $\sum_{j=1}^{N}(y_j-\bar{y})^2$

- Populacja **nie** jest **ograniczona**, jej elementy mogą przyjmować dowolne wartości.
- Pierwszy składnik powyższej sumy może przyjąć dowolną wartość, drugi, trzeci, … (N-1) podobnie.
- Ostatni (N – ty) składnik jest jednak ograniczony:

$$\sum_{j=1}^{N}(y_j-\bar{y})^2=0$$

Liczna stopni swobody *(ang. **NDF – Number of Degrees of Freedom**)*

dla sumy kwadratów wynosi N-1

$$\sigma^2(y)=\frac{1}{N-1}\sum_{j=1}^{N}(y_j-\bar{y})^2$$

Suma kwadratów dzielona przez liczbę stopni swobody jest **średnią kwadratową** lub **odchyleniem średnim kwadratowym**.

Pierwiastek kwadratowy z tego wyrażenia jest **pierwiastkiem ze średniego odchylenia kwadratowego** *(ang. RMS – Root Mean Square)*.

$$RMS=\sqrt{\sigma^2(y)}=\sigma(y)=\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(y_j-\bar{y})^2}$$

Pobieranie próby ze skończonej populacji bez uzupełnień – c.d

- **Element próby x_i** należy do pod-populacji y_j:

- **Prawdopodobieństwo**, że element y_j stanie się elementem próby x_i: $P = \frac{1}{N}$

- **Wartość oczekiwana**: $E = \frac{1}{N}$

- Jeśli jeden element z próby został już pobrany, **prawdopodobieństwo wybrania kolejnego** $P = \frac{1}{N-1}$

- **Prawdopodobieństwo wybrania 2 elementów**: $P = \frac{1}{N(N-1)}$

- **Wartość oczekiwana** pierwszego elementu z próby:

$$E(x_1) = \frac{1}{N}\sum_{j=1}^{N} y_j = \bar{y}$$

Pierwszy element próby **nie jest wyróżniony**,

tyle samo wynosi wartość oczekiwana każdego innego elementu próby.

Pobieranie próby ze skończonej populacji bez uzupełnień – c.d

- **Wartość oczekiwana średniej arytmetycznej** wszystkich elementów próby:

$$E(\bar{x}) = \frac{1}{n}\sum_{j=1}^{N} E(x_i) = \bar{y}$$

(Można udowodnić, że) **średnia arytmetyczna jest nieobciążonym estymatorem dla wartości średniej z populacji.**

- **Wariancja z próby**: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Można udowodnić: $E(s_x^2) = \sigma(y)$

- **Wariancja z próby jest nieobciążonym estymatorem wariancji z populacji.**

POBIERANIE PRÓBY Z ROZKŁADU NORMALNEGO

ROZKŁAD X^2

Pobieranie próby z rozkładu normalnego. Rozkład X^2.

Pobieramy próbę o elementach: $x_1, x_2, \ldots, x_n$.

Tworzymy sumę kwadratów: $x^2 = x_1^2 + x_2^2 + \ldots + x_n^2$.

Można dowieść, że wielkość ta ma dystrybuantę:

$$F(X^2) = \frac{1}{\Gamma(\lambda)2^{\lambda}} \int_0^{X^2} u^{\lambda-1} e^{-1/2u}\, du \qquad \lambda = 1/2\, n$$

n – ilość stopni swobody

$$k = \frac{1}{\Gamma(\lambda)2^{\lambda}}$$

Gęstość prawdopodobieństwa:

$$f(X^2) = k(X^2)^{\lambda-1} e^{-1/2\, X^2}$$

Pobieranie próby z rozkładu normalnego. Rozkład X^2.

Można dowieść, że:

Suma dwóch różnych rozkładów X^2 o odpowiednio n_1 oraz n_2 stopniach swobody jest rozkładem X^2 o licznie stopni swobody $n = n_1 + n_2$.

A także:

$$E(x^2) = 2\lambda = n$$

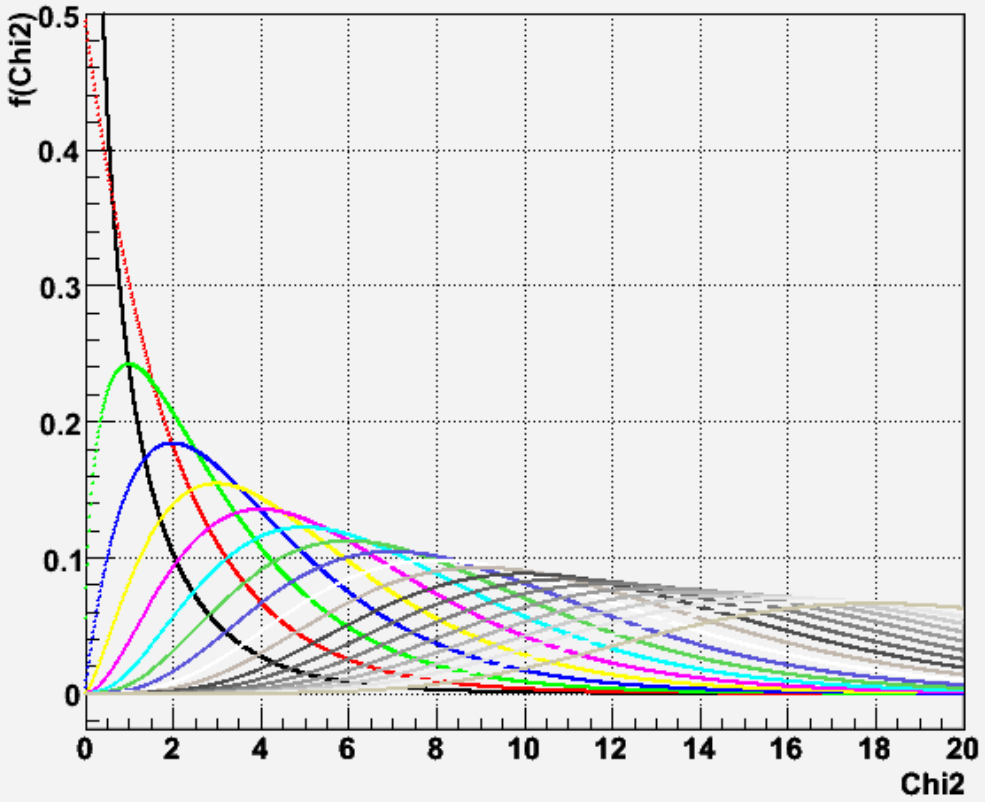
$$E\{(x^2)^2\} = 4\lambda^2 + 4\lambda = 4\left(\left(\frac{n}{2}\right)^2 + \frac{n}{2}\right)$$

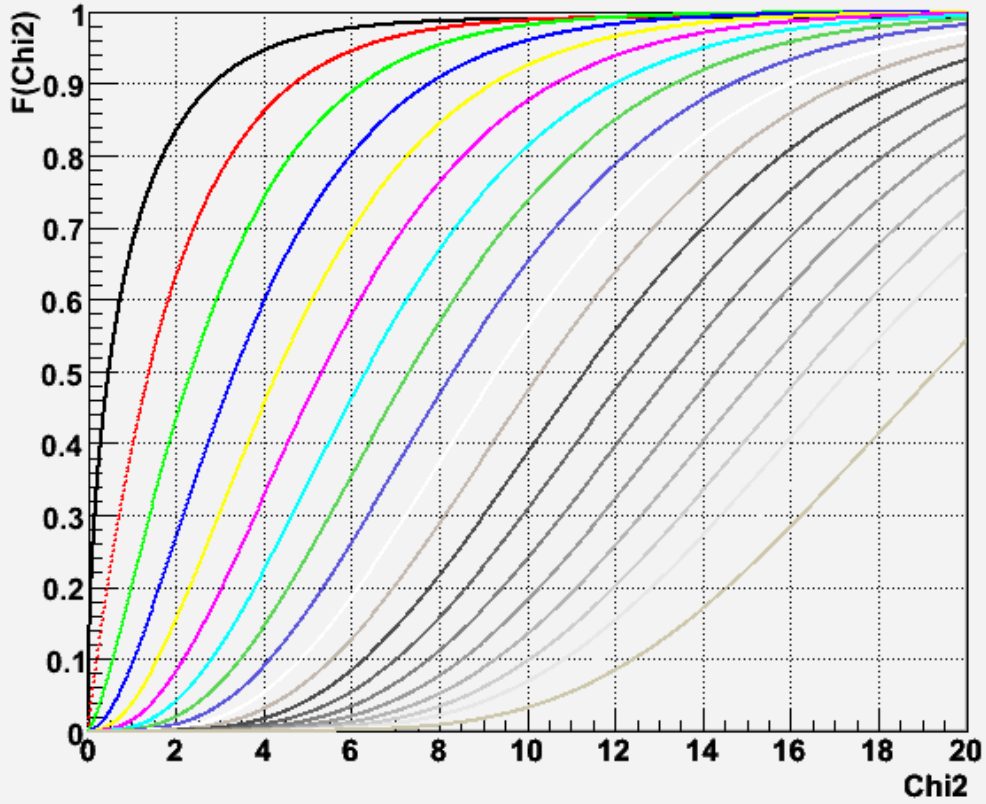
$$\sigma^2(x^2) = 4\lambda = 2n$$

Rozkład X^2

$$f(X^2) = k(X^2)^{\lambda-1} e^{-1/2\,X^2} \qquad k = \frac{1}{\Gamma(\lambda)\,2^{\lambda}} \qquad \lambda = 1/2\,n$$

n=1, 2 ,..., 20 (n - ilość stopni swobody)





Rozkład X^2

Rozkład X^2 ma szerokie zastosowanie, a w szczególności jako miara ufności uzyskanego wyniku. Im mniejsza jego wartość, tym pozornie słuszniejszy jest wynik (wartość zdefiniowana jako suma kwadratów odchyleń poszczególnych elementów próby od wartości średniej populacji).

Dystrybuanta określa prawdopodobieństwo, że zmienna losowa x^2 nie przekroczy określonej Wartości X^2.

$$F(X^2) = P(x^2 < X^2)$$

W zastosowaniach praktycznych używa się jej jako miary zaufania do wyniku wielkości:

$$W(X^2) = 1 - F(X^2)$$

$W(X^2)$ to poziom ufności.

Funkcja odwrotna do dystrybuanty określa kwantyle rozkładu X^2 (używa do testowania hipotez, dla określenia granic X^2, wewnątrz których dana hipoteza może być przyjęta jako prawdziwa):

$$X^2_F = X^2(F) = X^2(1 - W)$$

Rozkład X^2

Dotychczasowe rozważania dotyczyły sytuacji, kiedy populacja jest opisana standardowym rozkładem normalnym. Kiedy mamy do czynienia z rozkładem w postaci ogólnej:

$$x^2 = \frac{(x_1 - a)^2 + (x_2 - a)^2 + \ldots + (x_n - a)^2}{\sigma^2}$$

A jeśli poszczególne zmienne mają swoje wartości oczekiwane i odchylenia standardowe:

$$x^2 = \frac{(x_1 - a_1)^2}{\sigma_1^2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} + \ldots + \frac{(x_n - a_n)^2}{\sigma_n^2}$$

Można udowodnić, że zmienna losowa $\frac{n-1}{\sigma^2} s^2$ ma rozkład X^2 z f = n-1 stopniami swobody (s jest estymatorem wariancji).

$$s^2 = \frac{1}{n-1}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\}$$

POBIERANIE PRÓBY Z ODLICZANIEM

Pobieranie próby z odliczaniem. Próbki

Losowane jest n-elementów z populacji, sprawdzamy, czy czy są obdarzone cechą charakterystyczną. Zatrzymujemy tylko k-elementów, które wykazują tę cechę. n-k elementów odrzucamy.

Jest to pobieranie próby, której elementy podlegają rozkładowi dwumianowemu.

Parametry p, q odpowiadają odpowiednio prawdopodobieństwu wystąpienia tej cechy lub nie.

Estymator parametru p: $S(p)=\frac{k}{n}$

Wariancja: $\sigma^2(S(p))=\frac{p(1-p)}{n}$

Estymator wariancji: $s^2(S(p))=\frac{1}{n}\frac{k}{n}(1-\frac{k}{n})$

Niepewność: $\Delta k=\sqrt{s^2(S(np))}=\sqrt{k(1-\frac{k}{n})}$

Niepewność ten zależy jedynie od liczebności próbki, tzn, że jest niepewnością statystyczną.

Pobieranie próby z odliczaniem. Próbki

Rozważmy przypadek kiedy k<<n.

Zdefiniujmy parametr λ= np.

Estymator: $S(\lambda)=S(np)=k$

$$\Delta\lambda=\sqrt{k}$$

Czasami podaje się: $\Delta k=\sqrt{k}$

Trzeba jednak pamiętać, że stwierdzenie, że niepewność statystyczna liczby odliczeń jest równa pierwiastkowi kwadratowemu z liczby zliczeń ma zastosowanie tylko wtedy, kiedy k<<n

Przyjrzyjmy się bliżej interpretacji: $\Delta\lambda=\sqrt{k}$

1) liczba k nie jest mała (np. k>20). Rozkład Poissona przechodzi w rozkład Gaussa o wartości średniej oraz wariancji równej λ. Zmienna skokowa jest zastąpiona zmienną typu ciągłego. Gęstość prawdopodobieństwa jest więc opisana wzorem:

$$f(x,\lambda)=\frac{1}{\sqrt{\lambda 2\pi}}\exp\left\{-\frac{(x-\lambda)^2}{2\lambda}\right\}$$

Pobieranie próby z odliczaniem. Próbki

$$f(x,\lambda)=\frac{1}{\sqrt{\lambda 2\pi}}\exp\left\{-\frac{(x-\lambda)^2}{2\lambda}\right\}$$

Przy pomocy gęstości prawdopodobieństwa możliwe jest zdefiniowanie granic przedziału ufności przy zadanym poziomie ufności: $\beta=1-\alpha$

$$P(\lambda_m\le\lambda\le\lambda_p)=1-\alpha$$

Prawdopodobieństwo wystąpienia prawdziwej wartości λ wewnątrz przedziału ograniczonego wartościami λ_m λ_p jest równe poziomowi ufności 1 – α.

Można dowieść, że:

$$P(x<k|\lambda=\lambda_p)=1-\frac{\alpha}{2} \qquad \frac{\alpha}{2}=\psi_0\left(\frac{k-\lambda_p}{\sigma}\right)$$

$$P(x>k|\lambda=\lambda_m)=1-\frac{\alpha}{2} \qquad 1-\frac{\alpha}{2}=\psi_0\left(\frac{k-\lambda_m}{\sigma}\right)$$

ψ_0 dystrybuanta r. normalnego

Podczas, gdy: $1-\alpha=\frac{68.3}{100}$ $\lambda=k\pm\sqrt{k}$

Wzór ten określa granice przedziału ufności na poziomie ufności 68.3%.

Pobieranie próby z odliczaniem. Próbki

2) liczba k jest mała, rozkładu Poissona nie można przybliżyć rozkładem Gaussa.

$$P(x<k|\lambda=\lambda_p)=1-\frac{\alpha}{2}$$

$$P(x>k|\lambda=\lambda_m)=1-\frac{\alpha}{2}$$

Gęstością prawdopodobieństwa jest:

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

Można dowieść, że:

$$\frac{\alpha}{2}=F(k+1;\lambda_p)$$

$$1-\frac{\alpha}{2}=F(k;\lambda_m)$$

$$F(k;\lambda)=\sum_{n=0}^{k-1} f(n;\lambda)=P(\boldsymbol{k}<k)$$

F jest dystrybuanta rozkładu Poissona.

Aby wyznaczyć λ_m λ_p należy rozwiązać powyższe równania (obliczyć funkcje odwrotna do rozkładu Poissona przy ustalonym k i zadanym prawdopodobieństwie P (tu: α/2 i 1 – α/2))

Próbki z tłem

Jest wiele doświadczeń, kiedy rejestrowane są dwa (lub więcej) sygnałów (np. przypadki sygnału oraz przypadki tła).

Wartości oczekiwane liczb zdarzeń w danym doświadczeniu maja rozkład Poissona

z parametrem $\lambda = \lambda_s + \lambda_t$.

Celem doświadczenia jest wyznaczenie λ_s (więc λ_t trzeba znać wcześniej).

Rozumowanie, że należy skorzystać z metody przedstawione na poprzednich transparencjach prowadzi do błędnych wyników.

Prawdopodobieństwo zaobserwowania n zdarzeń ($n = n_s + n_t$):

$$f(n;\lambda_s+\lambda_b)=\frac{1}{n!}e^{-(\lambda_s+\lambda_t)}(\lambda_s+\lambda_t)^n$$

Prawdopodobieństwa zaobserwowania sygnału oraz tła:

$$f(n_s;\lambda_s)=\frac{1}{n_s!}e^{-\lambda_s}(\lambda_s)^{n_s}$$

$$f(n_t;\lambda_t)=\frac{1}{n_t!}e^{-\lambda_t}(\lambda_t)^{n_t}$$

Próbki z tłem

Wiemy, że przy liczbie k zanotowanych przypadków, tło nie może tej liczby przewyższać.

$$f'(n_t;\lambda_t)=\frac{f(n_t;\lambda_t)}{\sum_{n_t=0}^{k} f(n_t;\lambda_t)} \qquad (*)$$

$$f'(n;\lambda_t+\lambda_s)=\frac{f(n;\lambda_t+\lambda_s)}{\sum_{n_t=0}^{k} f(n_t;\lambda_t)}$$

Granice przedziału $\lambda_{sm}\le\lambda_s\le\lambda_{sp}$ na poziomie ufności $1-\alpha$

$$\frac{\alpha}{2}=F'(k+1;\lambda_{sp}+\lambda_t)$$

$$1-\frac{\alpha}{2}=F'(k;\lambda_{sm}+\lambda_t)$$

$$F'(k;\lambda_s+\lambda_t)=\sum_{n=0}^{k-1} f'(n;\lambda_s+\lambda_t)=P(\boldsymbol{k}<k)$$

F' to dystrybuanta odpowiadająca znormalizowanej gęstości prawdopodobieństwa (*).

KONIEC WYKŁADU 7 (8)