7. DOPASOWYWANIE ROZKŁADU DO DANYCH

Dopasowywanie rozkładu jest procedurą poszukiwania rozkładu teoretycznego, który najlepiej odpowiada danym empirycznym. Dobrze zidentyfikowany rozkład teoretyczny pozwala na prawidłowe szacowanie prawdopodobieństw wystąpienia określonych zdarzeń, pozwala więc na podejmowanie właściwych decyzji. Identyfikacja rozkładu teoretycznego który najlepiej odpowiada zaobserwowanym danym to nie tylko wybór typu rozkładu ale również ustalenie wartości parametrów tego rozkładu. Parametry te na ogół są szacowane przy pomocy metody największej wiarygodności czy metody momentów. Zgodność rozkładu empirycznego z wybranym rozkładem teoretycznym może być oceniana poprzez analizę wykresów: histogramów z nałożoną funkcją gęstości rozkładu teoretycznego, wykresów *Q–Q* czy wykresów *P–P* lub poprzez testowanie hipotez o zgodności rozkładów.

7.1. Dopasowywanie rozkładu teoretycznego w MATLAB-ie

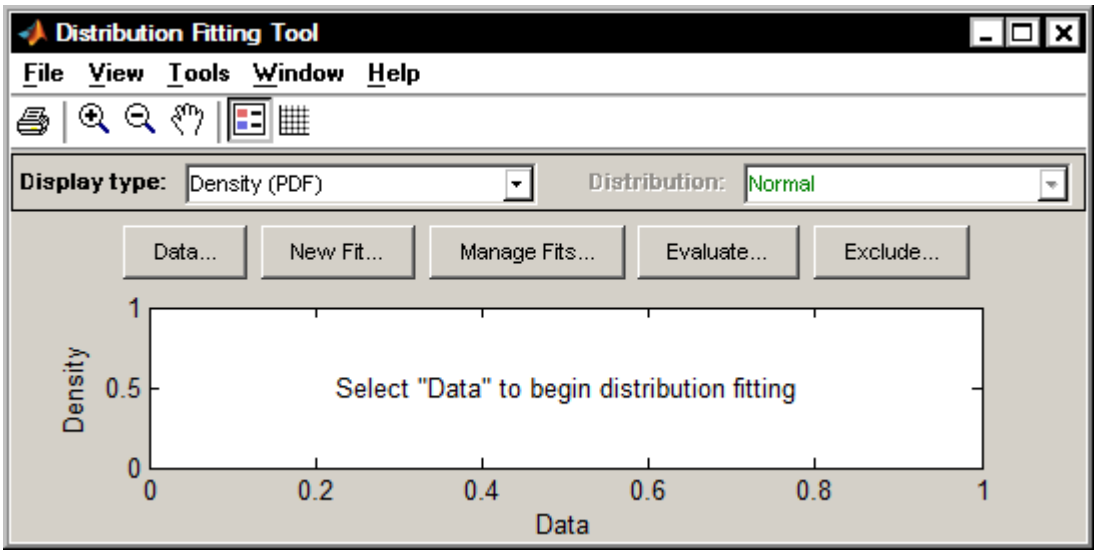
Znalezienie odpowiedniego rozkładu teoretycznego dla zbioru danych empirycznych ułatwia w MATLAB-ie narzędzie `dfittool`. Jest to właściwie odrębna aplikacja, która może korzystać z danych znajdujących się w obszarze roboczym (Workspace) MATLAB-a po ich zaimportowaniu. Podobnie wyniki w postaci zbioru punktów odpowiadających funkcji gęstości czy dystrybuancie rozkładu teoretycznego mogą być eksportowane do obszaru roboczego MATLAB-a.
Praca z aplikacją sprowadza się do wyboru przez użytkownika określonego typu rozkładu teoretycznego. W odpowiedzi, w oparciu o metodę największej wiarygodności, wyznaczane są parametry rozkładu teoretycznego dające najlepsze dopasowanie obydwu rozkładów. Dodatkowo, jakość dopasowania można ocenić na wybranym przez użytkownika wykresie. Sposób wykorzystania narzędzia `dfittool` zostanie omówiony na następującym przykładzie.

Załóżmy, że zebrane zostały dane pochodzące z rozkładu χ^2 o 3 stopniach swobody:

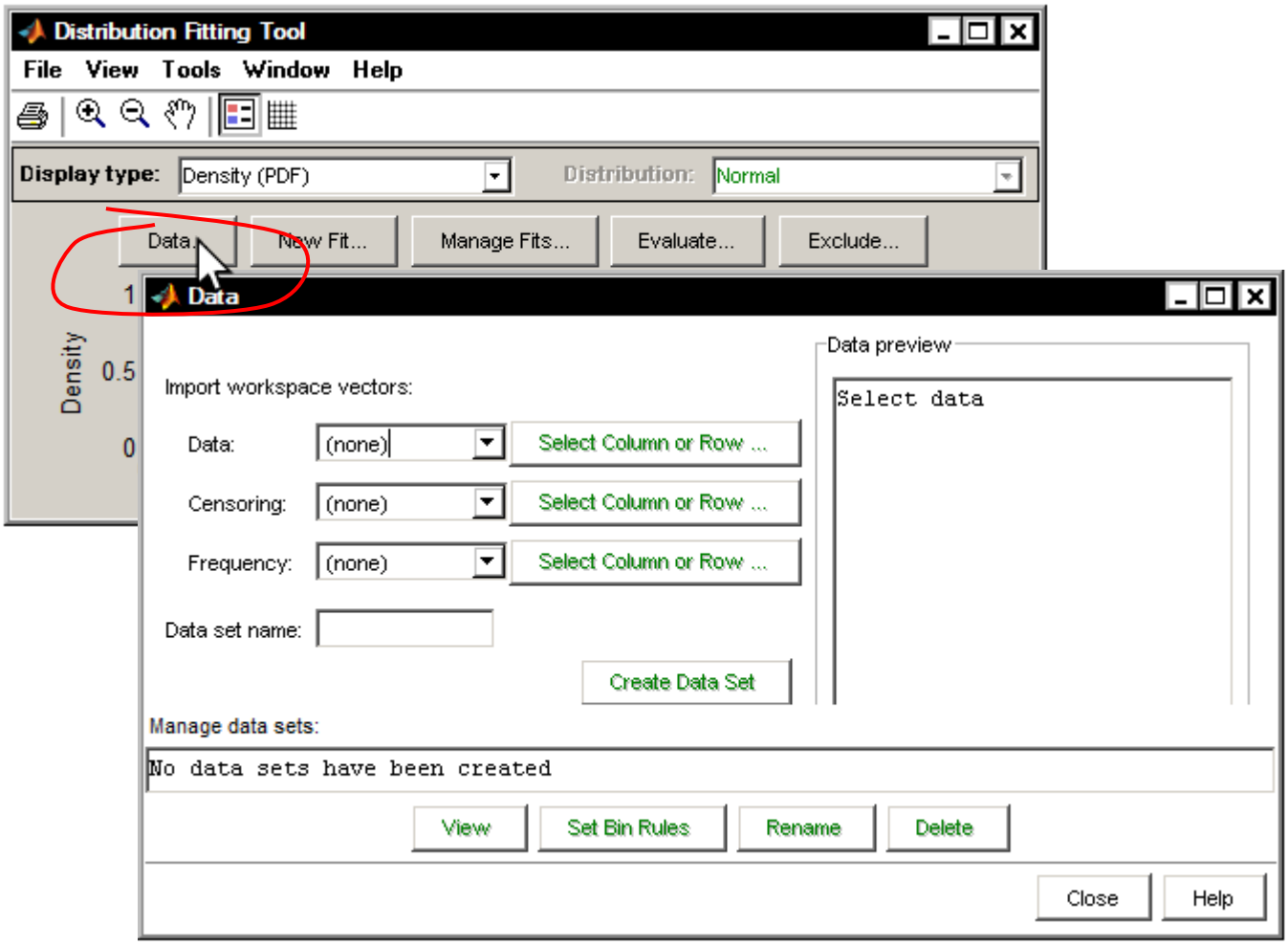
```
x=chi2rnd(3,100,1);
```

Okno główne aplikacji wyświetlane jest po wywołaniu polecenia:

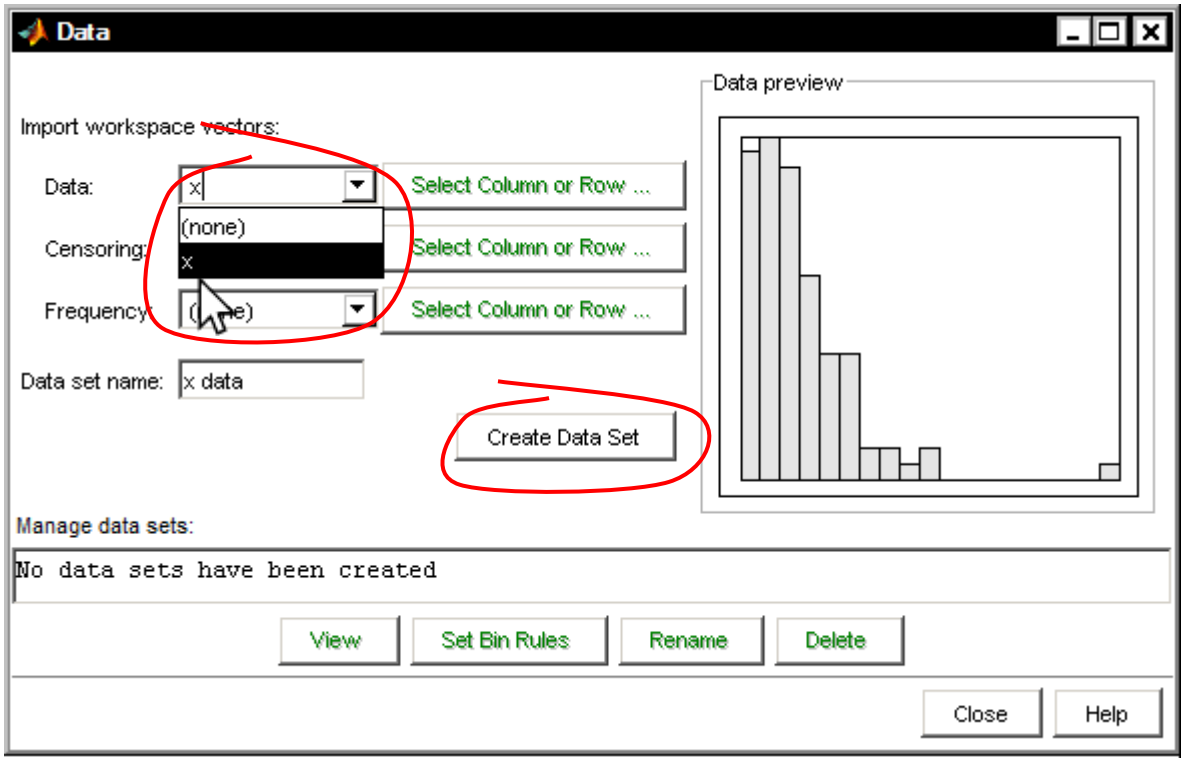
```
dfittool
```



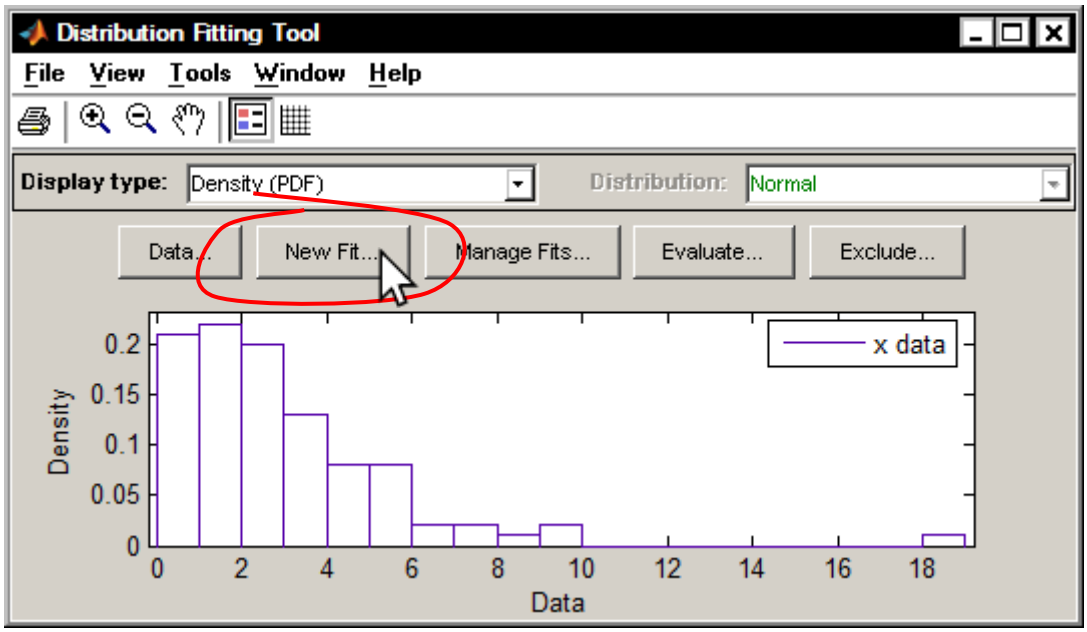
Wyniki pomiarów należy zaimportować do programu klikając przycisk „`Data`” i wskazując właściwy wektor w wyświetlonym po kliknięciu oknie:



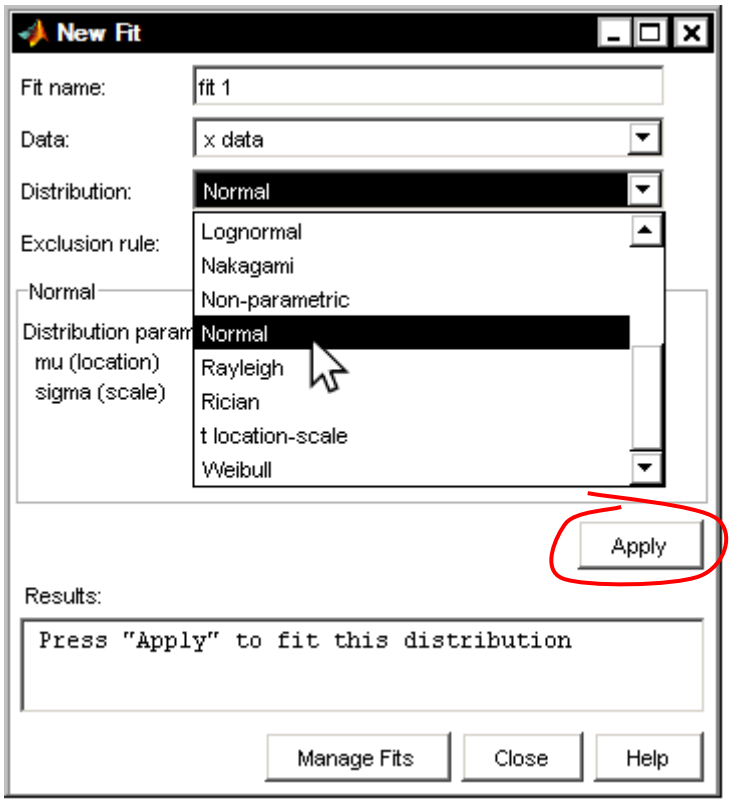
Dane obszaru roboczego są widoczne w liście: `Data`. Po wskazaniu właściwego zbioru należy je zaimportować klikając ma przycisku „`Create Data Set`”.



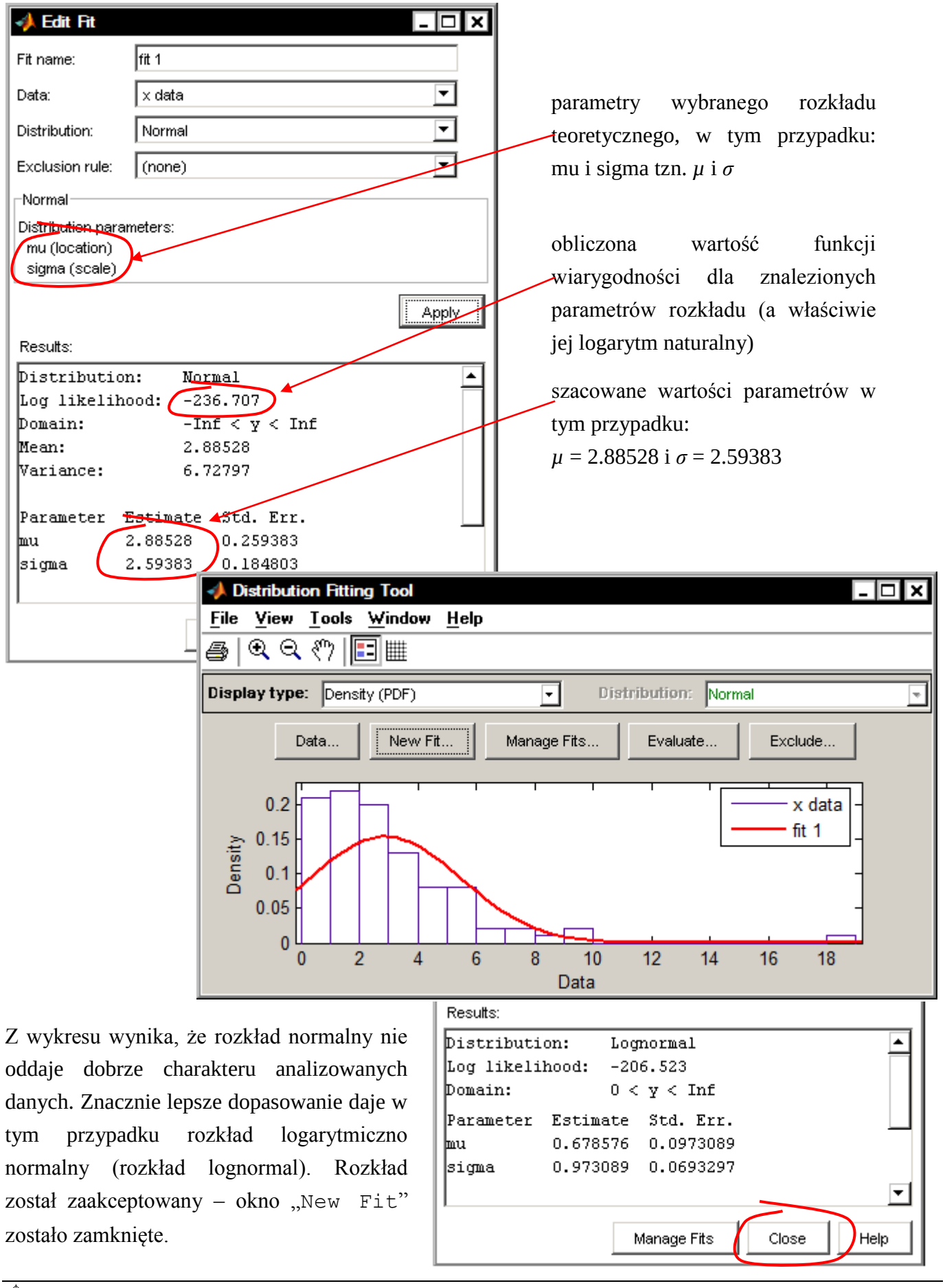
Po zaimportowaniu danych i zamknięciu okna „Data” w oknie głównym programu wyświetlany jest automatycznie histogram przedstawiający rozkład zaimportowanych danych. Najlepiej pasujący do tego rozkładu rozkład teoretyczny można ustalić z pomocą okna „New Fit”, które wyświetlane jest po naciśnięciu przycisku: „New Fit”.



W oknie „New Fit” po wskazaniu typu rozkładu i należy kliknąć przycisk „Apply”. Automatycznie zostaną wyznaczone najlepsze (maksymalizujące funkcję wiarygodności) wartości parametrów rozkładu.



Wyznaczone wartości parametrów wyświetlane są w dolnej części okna, w oknie głównym na wykres rozkładu empirycznego nakładany jest wyznaczony rozkład teoretyczny.



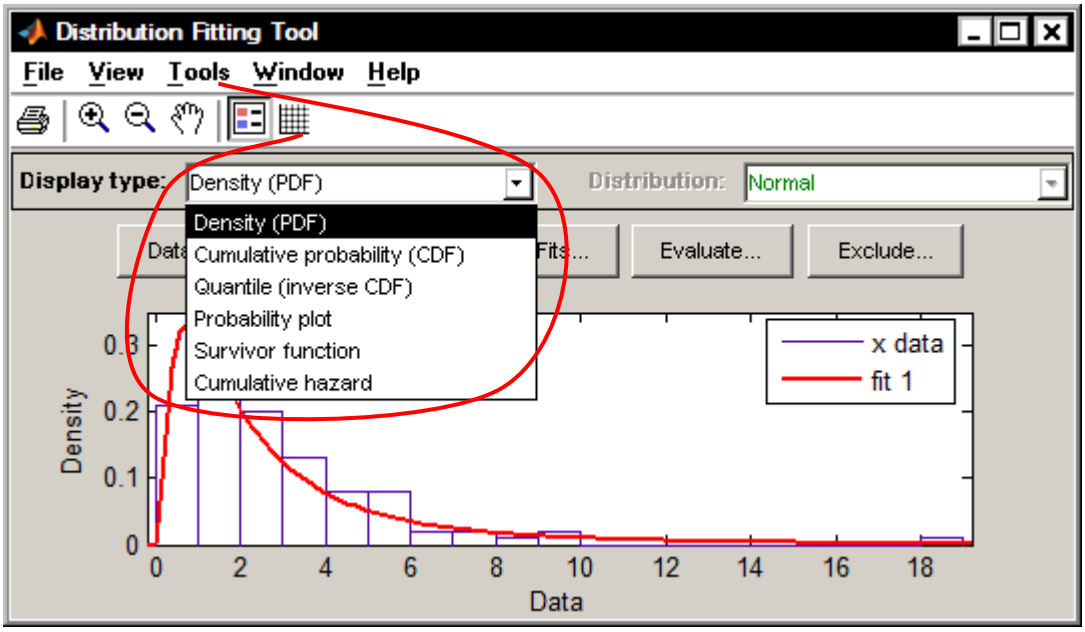
parametry wybranego rozkładu teoretycznego, w tym przypadku: mu i sigma tzn. μ i σ

obliczona wartość funkcji wiarygodności dla znalezionych parametrów rozkładu (a właściwie jej logarytm naturalny)

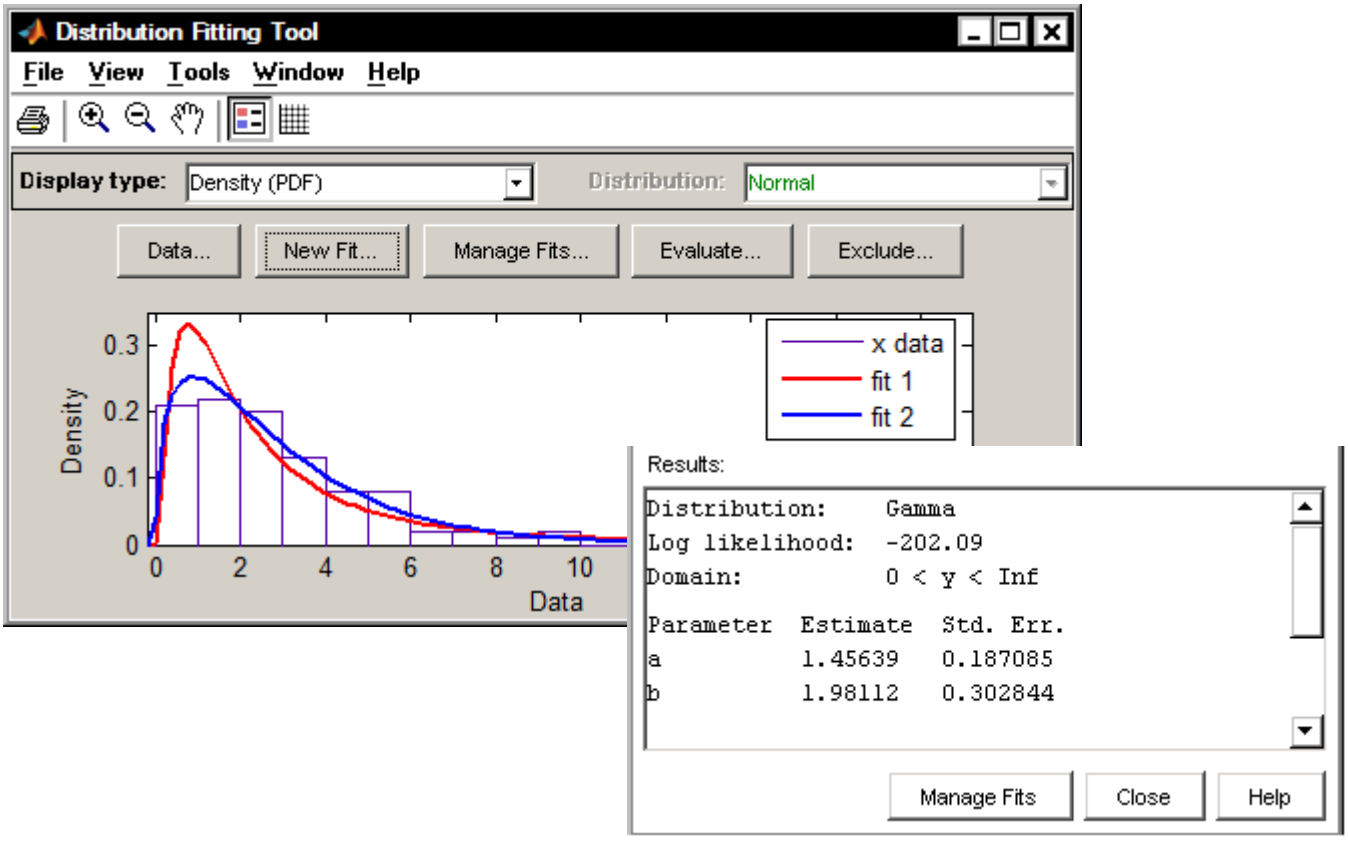
szacowane wartości parametrów w tym przypadku:
$\mu = 2.88528$ i $\sigma = 2.59383$

Z wykresu wynika, że rozkład normalny nie oddaje dobrze charakteru analizowanych danych. Znacznie lepsze dopasowanie daje w tym przypadku rozkład logarytmiczno normalny (rozkład lognormal). Rozkład został zaakceptowany – okno „`New Fit`” zostało zamknięte.

W oknie głównym aplikacji na wykres rozkładu empirycznego nałożony został wyznaczony powyżej rozkład logarytmiczno normalny. Rozkład ten opisany jest w legendzie jako „`fit 1`” – jest to domyślna nazwa nadawana przez program, którą użytkownik może zmienić podając własną na etapie dopasowywania rozkładów (tzn. w poprzednim oknie) w polu „`Fit name`”.



Okno główne pozwala na wykonanie porównania obydwu rozkładów także na innych wykresach – patrz lista rozwijalna „`Display type`”. Dodatkowo, możliwe jest również jednoczesne porównanie rozkładu danych z kilkoma rozkładami teoretycznymi. Dodanie kolejnego rozkładu teoretycznego odbywa się w taki sam sposób jak dodanie pierwszego rozkładu, tzn. poprzez wykorzystanie okna „`New Fit`”. Na poniższym rysunku analizowane dane zostały porównane również z rozkładem Gamma.



Ocenę jakości dopasowania można oprzeć na obliczanych dla każdego rozkładu teoretycznego wartościach logarytmu funkcji wiarygodności (`Log likelihood`). Najczęściej używanymi wskaźnikami używanymi do oceny dopasowania są tzw. kryteria informacyjne:

Akaike'go: $AIC = -2\ln(L) + 2k$,

Schwarza (bayesowskie): $BIC = -2\ln(L) + k\ln(n)$,

gdzie: L – wartość funkcji wiarygodności, k – liczba parametrów rozkładu teoretycznego, n – rozmiar próby.

Przyjmuje się, że im mniejsza wartość kryterium tym lepsze dopasowanie. Kryteria informacyjne bazując na wartościach funkcji wiarygodności (im większa wartość funkcji tym dopasowanie jest lepsze) wprowadzają dodatkowe kary za złożoność rozkładu – kara jest tym większa im większa jest liczba parametrów rozkładu. Taka konstrukcja premiuje więc rozkłady prostsze – o mniejszej liczbie parametrów.

Wartości logarytmu funkcji wiarygodności otrzymane w wyniku porównania analizowanych danych z wybranymi rozkładami teoretycznymi przedstawia poniższa tabela.

Rozkład	*Parametry*	*Wartość funkcji* $\ln(L)$
normalny	$\mu = 2.885280$, $\sigma = 2.593832$	-236.707
logarytmiczno normalny	$\mu = 0.678576$, $\sigma = 0.973089$	-206.523
gamma	$\alpha = 1.456390$, $\beta = 1.981120$	-202.090

Ze względu na to, ze wszystkie rozważane rozkłady teoretyczne mają dwa parametry więc ocenę jakości dopasowania można przeprowadzić w oparciu o wartość logarytmu funkcji wiarygodności. W tym przypadku funkcja wiarygodności osiąga największą wartość dla rozkładu gamma – więc można uznać, że jest on najlepiej dopasowany do analizowanego rozkładu danych. W rzeczywistości dane pochodziły z rozkładu χ^2 o 3 stopniach swobody, rozkład χ^2 jest właściwie rozkładem gamma o parametrach $\alpha = n/2$ i $\beta = 2$ (n – ilość stopni swobody rozkładu χ^2) – więc otrzymany wynik jest zgodny z oczekiwaniami.

Wyniki przeprowadzonej analizy można sprawdzić testując hipotezę o zgodności rozkładu empirycznego z wybranym rozkładem teoretycznym. W tym celu należy przyjąć określone wartości badanego rozkładu teoretycznego. Można je albo przepisać z okna w którym dopasowywano parametry rozkładu albo też wyznaczyć posługując się jedną z dostępnych funkcji (niektóre rozkłady dostępne w oknie `dfittool` nie mają swoich odpowiedników w postaci funkcji `pdf`, `cdf` czy `inv`). Najbardziej uniwersalną funkcją jest funkcja `mle` (Maximum-Likelihood Estimation – estymacja metodą największej wiarygodności):

```
[par, przedz] = mle(x, 'distribution', dist)
```

gdzie:

`x` – wektor zawierający wartości dla których przeprowadzane będą obliczenia;

`par`, `przedz` – oszacowane parametry rozkładu wraz z przedziałami ufności,

`dist` – rozkład teoretyczny, dostępne wartości:

Lp.	Wartość 'dist'	Rozkład
1	'Beta'	beta
2	'Bernoulli'	Bernoulliego
3	'binomial'	dwumianowy
4	'Discrete uniform'	jednostajny dyskretny
5	'Exponential'	wykładniczy
6	'Extreme value'	wartości ekstremalnych
7	'Gamma'	gamma
8	'Geometric'	geometryczny
9	'lognormal'	logarytmiczno normalny
10	'negative binomial'	Pascala – ujemny rozkład dwumianowy
11	'Normal'	normalny
12	'Poisson'	Poissona
13	'Rayleigh'	Rayleigha
14	'Uniform'	jednostajny
15	'Weibull'	Weibulla

Estymację metodą największej wiarygodności wykorzystują także wyspecjalizowane funkcje (numery po prawej odpowiadają numerom rozkładów z powyższej tabeli):

```
[par, przedz] = betafit(x, alfa)                          (1)
[par, przedz] = binofit(x, n, alfa)                       (3)
[par, przedz] = expfit(x, alfa)                           (5)
[par, przedz] = evfit(x, alfa)                            (6)
[par, przedz] = gamfit(x, alfa)                           (7)
[par, przedz] = lognfit(x, alfa)                          (9)
[par, przedz] = nbinfit(x, alfa)                         (10)
[xs, sigma, p_xs, p_s] = normfit(x, alfa)                (11)
[par, przedz] = poissfit(x, alfa)                        (12)
[par, przedz] = raylfit(x, alfa)                         (13)
[par, przedz] = unifit(x, alfa)                          (14)
[par, przedz] = wblfit(x, alfa)                          (15)
```

gdzie:

`x` – wektor zawierający wartości dla których przeprowadzane będą obliczenia;

`par`, `przedz` – oszacowane parametry rozkładu wraz z przedziałami ufności;

`xs`, `s`, `p_xs`, `p_s` – średnia, odchylenie standardowe, przedziały ufności dla średniej i odchylenia standardowego rozkładu normalnego

`alfa` – poziom istotności wykorzystywany przy wyznaczaniu przedziałów ufności, domyślnie 0.05;

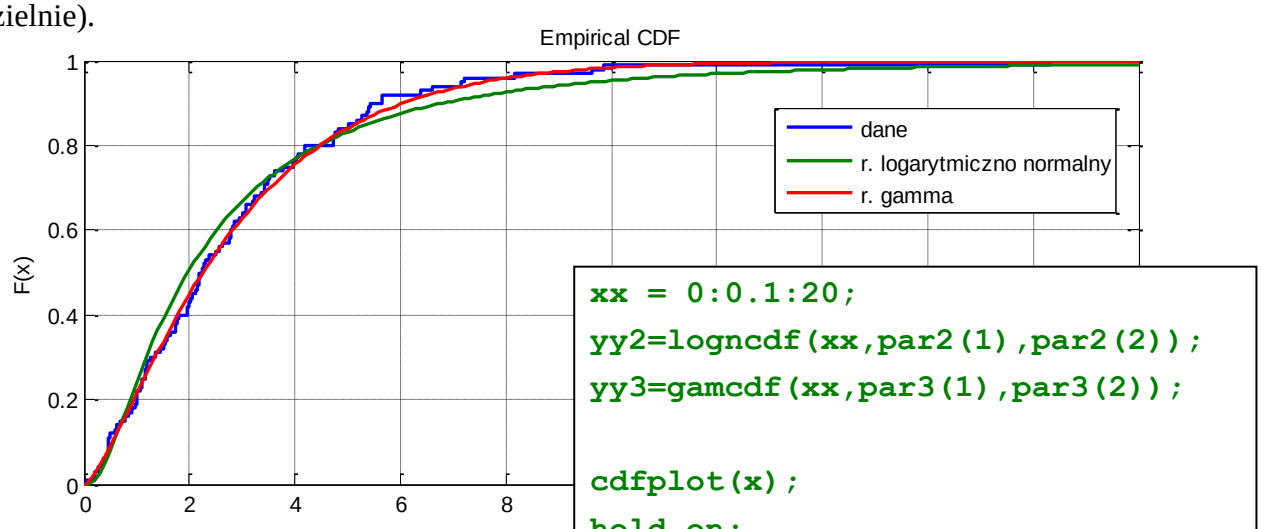
W powyższym przykładzie analizowano dopasowanie rozkładów: normalnego, logarytmiczno normalnego i rozkładu gamma. Zgodność tych rozkładów z badanym rozkładem empirycznym można również ocenić w oparciu o, dostępny w MATLAB-ie, test zgodności Kołmogorowa – Smirnowa. Test

ten wyznacza maksymalną odległość pomiędzy dystrybuantami obydwu rozkładów co dodatkowo pozwala na interpretację geometryczną wyniku. Poniżej przeprowadzone zostały testy zgodności dla badanych rozkładów teoretycznych.

```
alfa = 0.05;

%zgodność z rozkładem normalnym
par1 = mle(x,'distribution', 'normal')
par1 = 2.8853   2.5808
cdf = [x normcdf(x, par1(1), par1(2))];
[h p d]=kstest(x, cdf, alfa)
h = 1, p = 0.0381, d = 0.1389
%zgodność z rozkładem logarytmiczno normalnym
par2 = mle(x,'distribution', 'lognormal')
par2 = 0.6786   0.9682
cdf = [x logncdf(x, par2(1), par2(2))];
[h p d]=kstest(x, cdf, alfa)
h = 0, p = 0.3230, d = 0.0941
%zgodność z rozkładem gamma
par3 = mle(x,'distribution', 'gamma')
par3 = 1.4564   1.9811
cdf = [x gamcdf(x, par3(1), par3(2))];
[h p d]=kstest(x, cdf, alfa)
h = 0, p = 0.9973, d = 0.0394
```

Przyjęty poziom istotności $\alpha = 0.05$ pozwala na odrzucenie hipotezy o zgodności rozkładu z rozkładem normalnym, natomiast nie można odrzucić hipotez o zgodności z rozkładem logarytmiczno normalnym i z rozkładem gamma. Na lepsze dopasowanie rozkładu gamma wskazuje mniejsza wartość odległości pomiędzy dystrybuantami. Dopasowanie to można ocenić analizując wykres przedstawiający dystrybuanty (można skorzystać z możliwości narzędzia `dfittool` lub przygotować wykres samodzielnie).




```
xx = 0:0.1:20;
yy2=logncdf(xx,par2(1),par2(2));
yy3=gamcdf(xx,par3(1),par3(2));

cdfplot(x);
hold on;
plot(xx,yy2,'g',xx,yy3, 'r');
hold off
```