PDF Gaussiana e Propriedades

Guilherme de Alencar Barreto

gbarreto@ufc.br

Grupo de Aprendizado de Máquinas – GRAMA
Programa de Pós-Graduação em Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://www.researchgate.net/profile/Guilherme_Barreto2/

Conteúdo da Apresentação

1. Objetivo Geral
2. PDF Gaussiana
3. Transformação de Box-Cox
4. Transformação Z-score e sua inversa
5. PDF Gaussiana Padronizada
6. Geração de Números Aleatórios
7. Histograma de Frequência
8. Exemplos no Matlab/Octave

Objetivo Geral

Objetivo Geral da Aula

Introduzir noções elementares sobre função densidade de probabilidade normal e algumas de suas propriedades.

Parte I

FDP Gaussiana e Propriedades

Importância da FDP Gaussiana

- A FDP gaussiana é uma das mais importantes em ETI e em Ciências de um modo geral, pois é usada como modelo das flutuações aleatórias (ruído) que distorcem valores medidos de um determinada variável.
- Outras aplicações da FDP gaussiana:

1. Ruídos em canais de comunicações.
2. Robôs manipuladores (repetibilidade).
3. Modelos de ruído em imagens digitais.
4. Ruídos de medida em instrumentação eletrônica.
5. Detecção de anomalias em Monitoramento de Processos.

Definição

- Uma VA contínua é chamada de $normal$ ou $gaussiana$ se sua FDP é dada por

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}} \quad (1)$$

ou

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \quad (2)$$

em que $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}_{\geq 0}$ são constantes.

- O domínio da variável X é a reta dos números reais, ou seja, $-\infty < X < +\infty$.

Probabilidade como a área sob a curva $f_X(x)$

- Pode-se mostrar que, para uma variável X contínua com FDP $f_X(x)$, probabilidades são dadas por

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx, \tag{3}$$

  ou seja, a $\Pr(a \leq X \leq b)$ é dada pela área sob a curva no intervalo $[a, b]$.

- Por isso, em um gráfico de PDF, não se lê probabilidades no eixo vertical, mas na área sob a curva.

Probabilidade como a área sob a curva $f_X(x)$

Dois resultados interessantes derivam da Equação 3.

1. Probabilidade de uma variável contínua assumir um valor específico é zero.
   Demonstração: Fazendo $a = b$ na Equação (3), tem-se

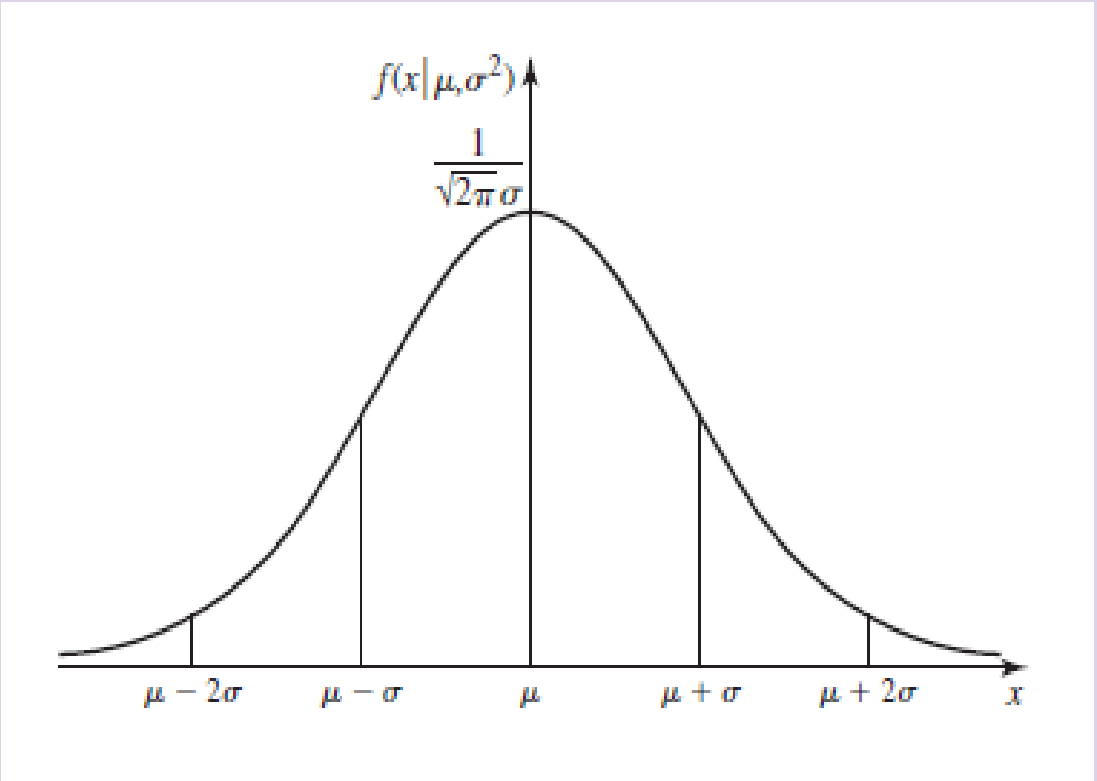
$$\Pr(a \leq X \leq a) = \Pr(X = a) = \int_a^a f_X(x)dx = 0. \qquad (4)$$

2. Área total sob a curva de uma FDP é 1.
   Demonstração: Fazendo $a = -\infty$ e $b = +\infty$ na Equação (3), tem-se

$$\Pr(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f_X(x)dx = 1. \qquad (5)$$

Juntando as informações anteriores, pode-se esboçar o gráfico da FDP gaussiana para uma V.A. de média μ e variância σ^2.

Gráfico da FDP Gaussiana



Definição (cont.-1)

- A FDP gaussiana é especificada por dois parâmetros:

$$\mu = \text{média populacional} \quad (6)$$
$$\sigma^2 = \text{variância populacional} \quad (7)$$

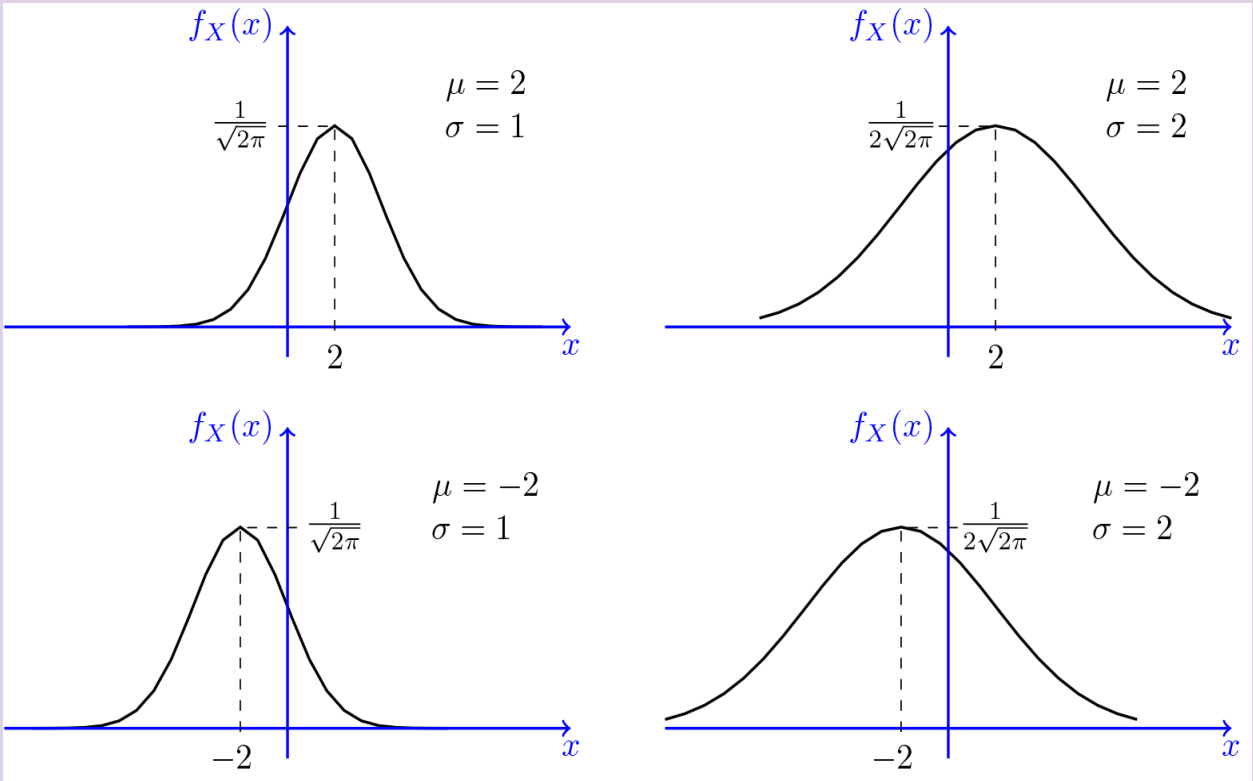
- O desvio padrão populacional é dado por $\sigma = \sqrt{\sigma^2}$.
- Que podem ser estimadas pelas seguintes fórmulas:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(média amostral)} \quad (8)$$
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \text{(variância amostral)} \quad (9)$$
$$s = \sqrt{s^2} \quad \text{(desvio padrão amostral)} \quad (10)$$

Definição (cont.-2)

- É equivalente fazer uma das seguintes afirmações:
  1. Uma variável X segue uma lei de probabilidade dada pela função gaussiana, cujos parâmetros são a média μ e a variância σ^2.
  2. Uma variável X está *distribuída* segundo uma FDP gaussiana de média μ e variância σ^2.
  3. Notação simplificada: $X \sim N(\mu, \sigma^2)$

Gráfico da FDP Gaussiana para diferentes médias e variâncias

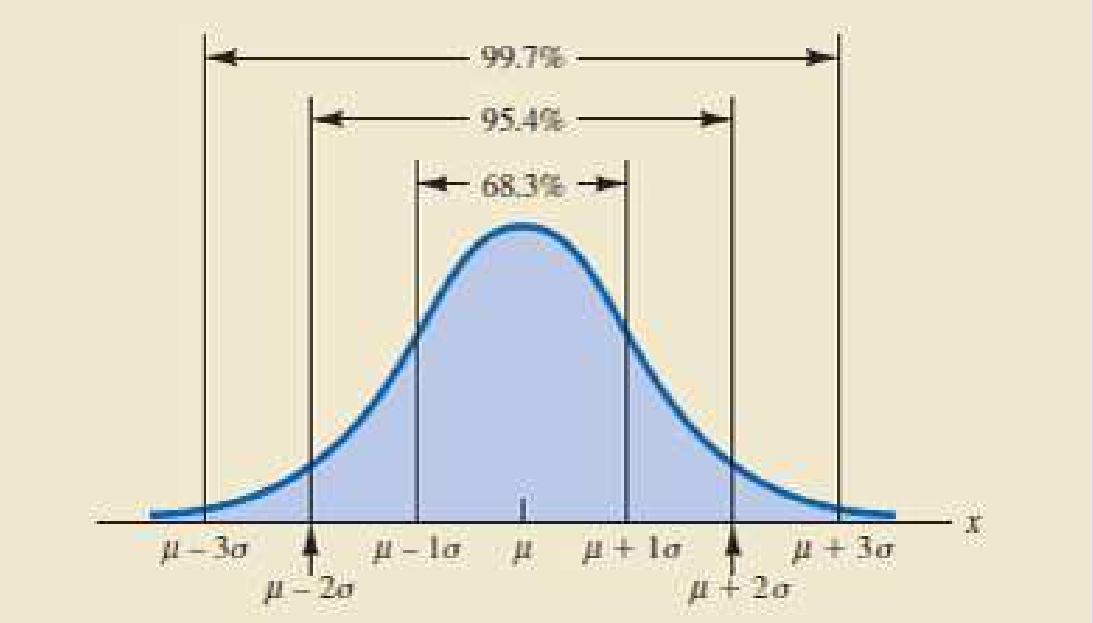


Cálculo da Variância Amostral

X_1	$d_1 = x_1 - \bar{x}_1$	$d_1^2 = (x_1 - \bar{x}_1)^2$
480	-54=480-534	2916
500	-34=500-534	1156
380	-154=380-534	23716
1100	+566=1100-534	320356
1100	+566=1100-534	320356
230	-304=230-534	92416
490	-44=490-534	1936
250	-284=250-534	80656
300	-234=300-534	54756
510	-24=510-534	576
$\bar{x}_1 = 534$	$\sum(x_1 - \bar{x}_1) \approx 0$	$\sum(x_1 - \bar{x}_1)^2 = 898840$
$s^2 = \sum_{n=1}^{10}(x_1(n) - \bar{x}_1)^2/(10-1) = 898840/9 = 99871,11$		

Table: Tabela para cálculo passo-a-passo da variância amostral.

Gráfico da FDP Gaussiana e Probabilidades Intervalares



Entendendo a FDP Gaussiana

- Com base no gráfico anterior da FDP Gaussiana temos que:
  1. 68% dos valores de X estarão no intervalo $[\mu - \sigma, \mu + \sigma]$.

$$\Pr(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0{,}68$$

  2. 95% dos valores de X estarão no intervalo $[\mu - 2\sigma, \mu + 2\sigma]$.

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}95$$

  3. 99% dos valores de X estarão no intervalo $[\mu - 3\sigma, \mu + 3\sigma]$.

$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0{,}99$$

- Esta propriedade da FDP gaussiana é usada para detecção de observações atípicas ou anômalas.

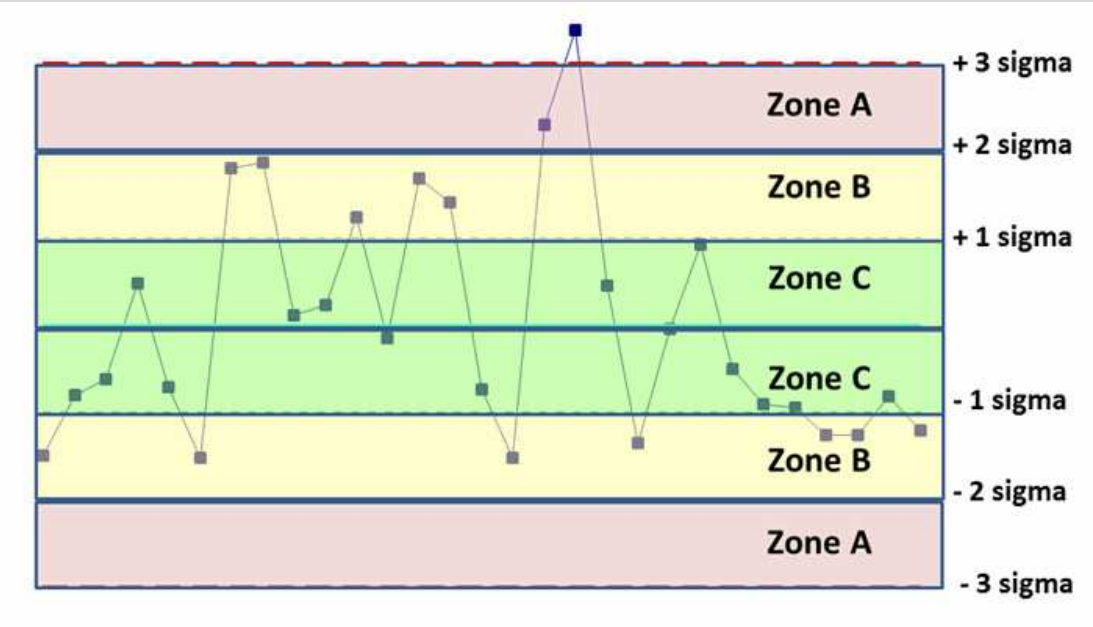
Aplicação em Detecção de Anormalidades em Processos

- Suponha que estejamos monitorando (medindo e armazenando) uma grandeza física X que representa o estado de um processo industrial.
- Se há razões para assumir que $X \sim N(\mu, \sigma^2)$, então podemos usar a seguinte regra de decisão:

$$\begin{array}{ll} \text{SE} & X \notin [\mu - 3\sigma, \mu + 3\sigma], \\ \text{ENTÃO} & X \text{ é um valor anômalo.} \end{array}$$

- No caso de detectar alguma anormalidade, o operador deve ser avisado (e.g. por meio de um alarme sonoro ou visual).

Aplicação em Controle de Processos



Zona A: Normal (zona de maior probabilidade de ocorrências).

Zona B: Normal? (ainda dentro do normal, mas exigindo atenção).

Zona C: Suspeita forte de anomalia.

Transformação de Box-Cox

- Muitas variáveis não seguem uma distribuição normal.
- Porém, normalidade é uma importante suposição para muitas técnicas e métodos estatísticos.
- Nestes casos, é possível realizar uma operação, conhecida como *transformação de Box-Cox*[a], sobre a variável x_i:

$$x_i^* = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & \text{se } 0 < \lambda < 1. \\ \ln x_i, & \text{se } \lambda = 0. \end{cases} \tag{11}$$

- Com isso, variáveis não gaussianas terão uma distribuição com uma forma mais simétrica, mais próxima da normal.

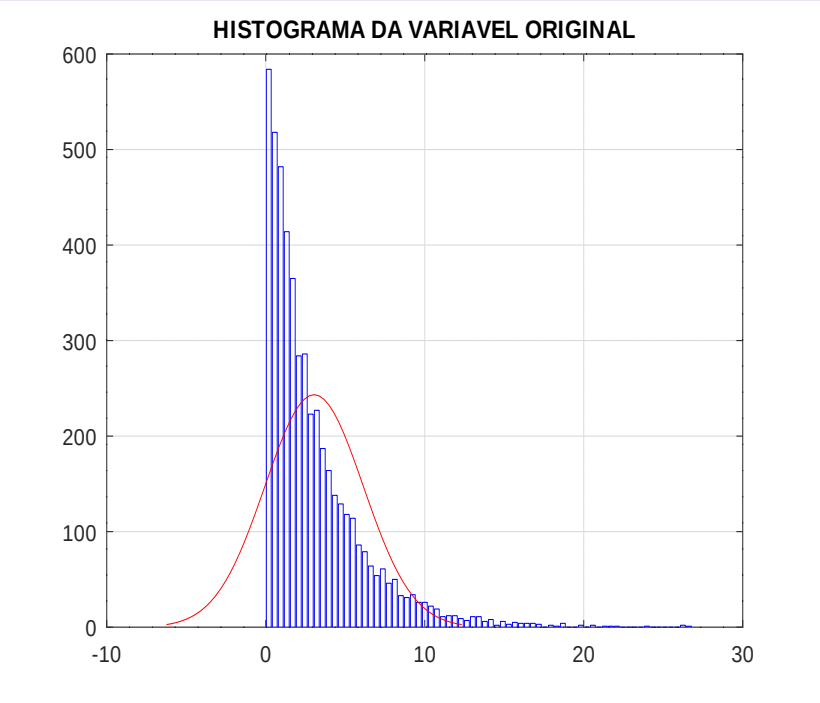
[a] G.E.P. Box and D.R. Cox, 'An Analysis of Transformations', Journal of the Royal Statistical Society B, 26:211-252 (1964).

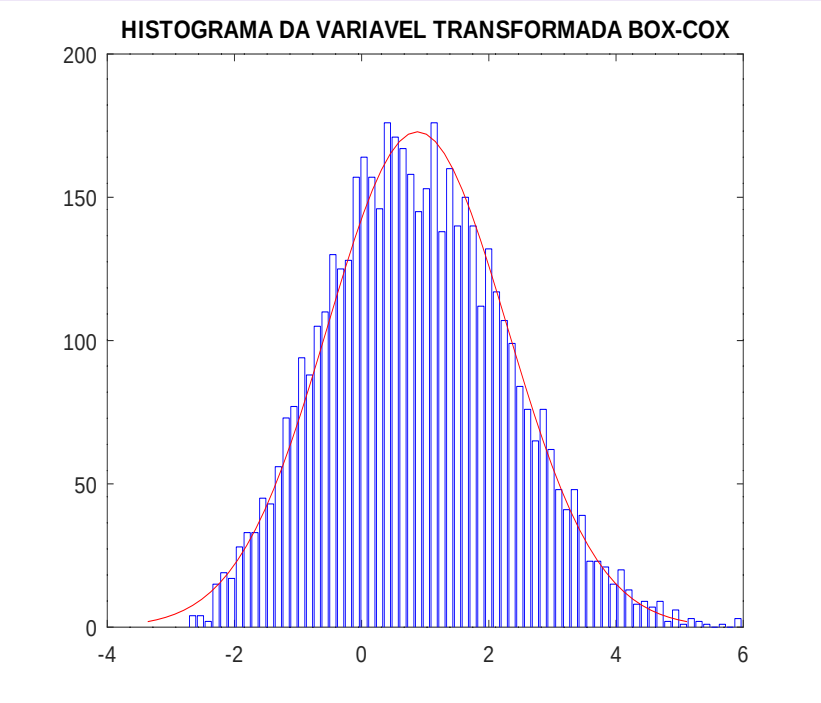
Exemplo da Transformação de Box-Cox

- Seja uma variável aleatória X que segue uma FDP exponencial de parâmetro $\gamma = 3$.
- No Octave, 5000 observações de X são geradas pelo seguinte comando: `>> x=exprnd(3,5000,1)`
- O histograma de X, com uma gaussiana ajustada às observações de X, é visualizada pelo seguinte comando:
  `>> histfit(x)`
- Aplica-se a transformação de Box-Cox às observações de X, para um expoente $\lambda = 0.2$, pela seguinte operação:

```
>> lamb=0.2;
>> xstar=(x.^lamb - 1)/lamb;
>> figure; histfit(xstar)
```

Resultado da aplicação da transformação de Box-Cox





Exemplo da Transformação de Box-Cox

- Uma questão interessante que surge com a aplicação da transformação de Box-Cox é como saber se a distribuição resultante é similar o suficiente de uma gaussiana.
- Para isso, podemos usar 2 métodos. O primeiro deles é qualitativo, e se baseia numa comparação visual das curvas da função distribuição acumulada (FDA) da distribuição transformada e da gaussiana de mesma média e variância.
- O segundo é métodos é quantitativo, e se baseia na aplicação do teste de hipótese de Kolmogorov-Smirnov (HS).
- Iremos detalhar os procedimentos a seguir.

Avaliação Qualitativa da Transformação de Box-Cox

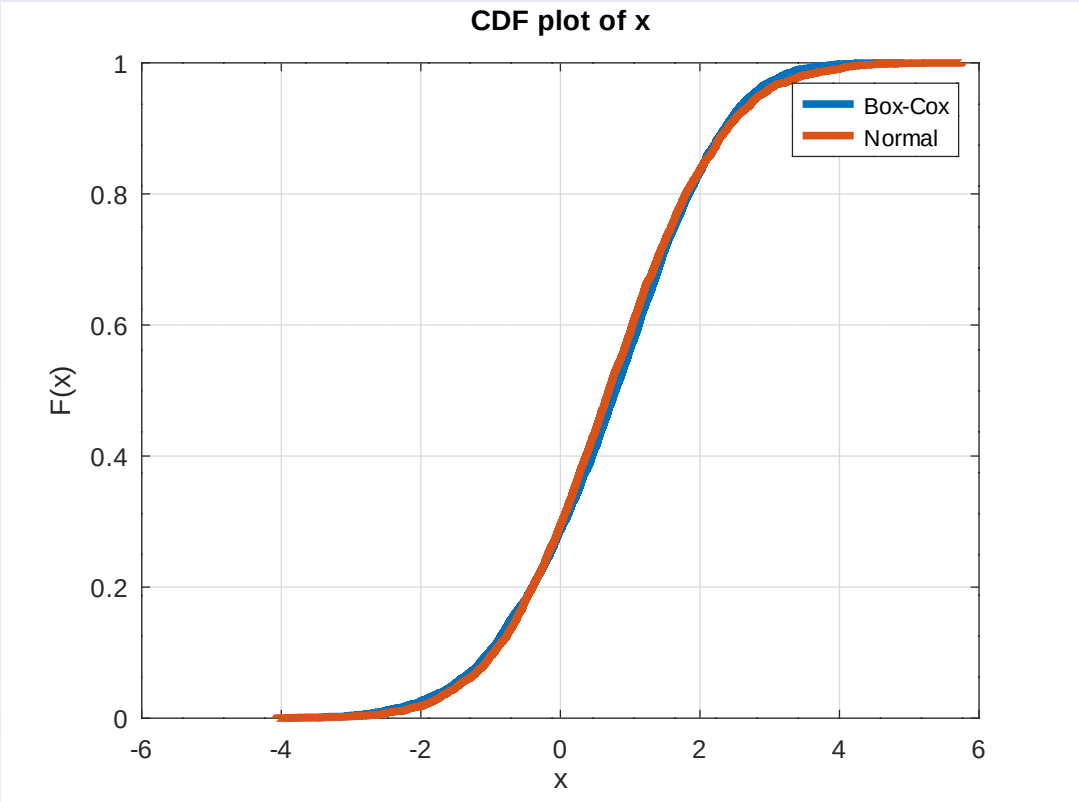
- Este procedimento compara a FDA da amostra transformada por Box-Cox e a FDA de uma amostra normalmente distribuída com mesma média e mesma variância que a amostra Box-Cox.
- A comparação das FDAs resulta mais acurada do que pela comparação dos histogramas das duas amostras.
- Em muitas aplicações, este procedimento é suficiente, uma vez que se busca uma boa aproximação e não uma confirmação da gaussianidade entre as amostras.

Avaliação Qualitativa da Transformação de Box-Cox

- Usando o vetor `xstar` com as medidas transformadas por Box-Cox, realizar as seguintes operações no Octave/Matlab:

```
>> mxstar=mean(xstar); % media das medidas em xstar
>> dpxstar=std(xstar); % desvio padrao das medidas em xstar
>> xnorm=normrnd(mxstar,dpxstar,5000,1); % N(mxstar,dpstar)
>> h1=cdfplot(xstar); % FDA empirica de xstar
>> hold on; % mantem figura para proximo plot
>> h2=cdfplot(xnorm); % FDA empirica de N(mxstar,dpstar)
```

Verificando se distribuição resultante é gaussiana comprando a sua FDA com a Normal de mesma média e variância.



Avaliação Quantitativa da Transformação de Box-Cox

- Este procedimento também compara as FDAs da amostra transformada por Box-Cox e a de uma amostra normalmente distribuída de mesma média e variância.
- Trata-se, porém, de um procedimento mais formal por ser um teste de hipóteses. Para isso, calcula-se uma estatística de teste baseada na maior distância vertical entre as duas FDAs comparando-a com o valor-p, para um certo nível de significância α.
- Este procedimento é mais rigoroso que a simples inspeção visual, uma vez que busca confirmar $(H = 0)$ ou não $(H = 1)$ a gaussianidade entre as amostras.
- Em $Python$: `scipy.stats.kstest`

Avaliação Quantitativa da Transformação de Box-Cox

- Usando o vetor `xstar` com as medidas transformadas por Box-Cox, realizar as seguintes operações no Octave/Matlab:

```
>> mxstar=mean(xstar); % media das medidas em xstar
>> dpxstar=std(xstar); % desvio padrao das medidas em xstar
>> xnorm=normrnd(mxstar,dpxstar,5000,1); % N(mxstar,dpstar)
>> H=kstest2(xstar,xnorm); % Teste Kolmogorov-Smirnov
H=1
```

Comentário: Pelo resultado acima a distribuição da amostra transformada por Box-Cox não pode ser considerada como sendo compatível com uma gaussiana.

FDP Gaussiana Padronizada

- Após a aplicação da transformação de Box-Cox, é usual fazer com que o conjunto de medidas possua $\mu = 0$ e $\sigma^2 = 1$.
- Esta propriedade pode ser imposta a qualquer conjunto de medidas, através da seguinte transformação:

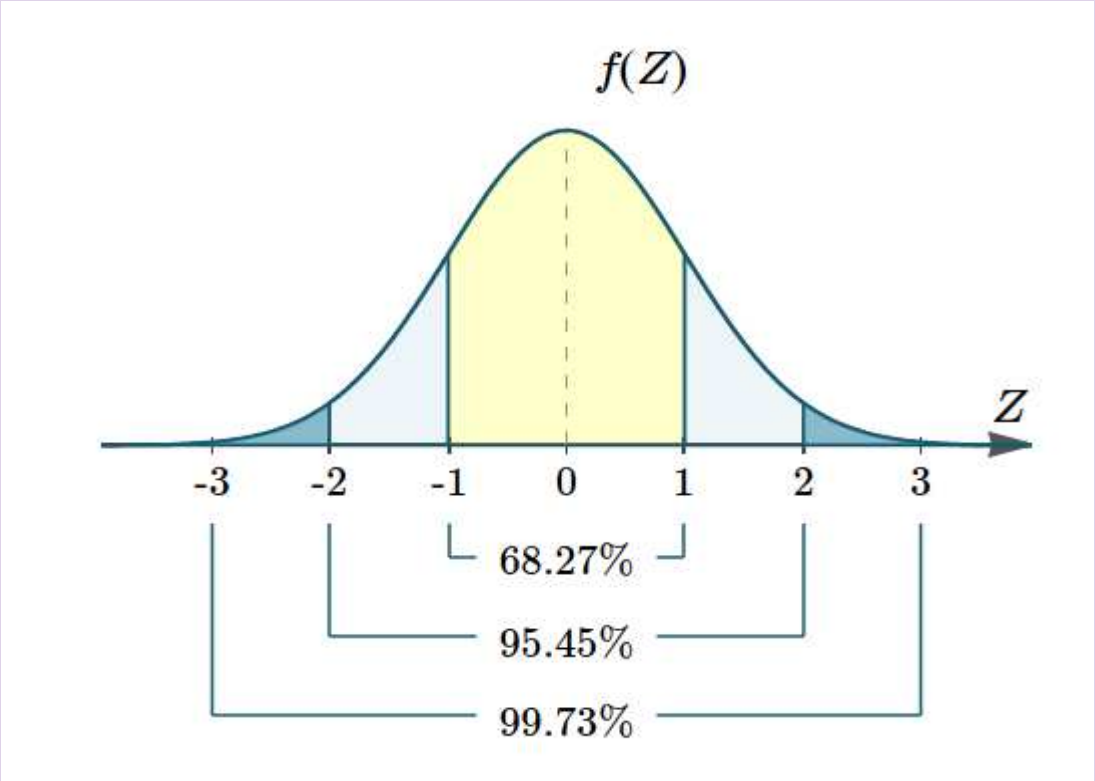
$$z_i = \frac{x_i - \mu}{\sigma}, \qquad i = 1, 2, \ldots, N. \tag{12}$$

- Para voltar para aos valores originais, aplica-se a seguinte transformação inversa:

$$x_i = \sigma z_i + \mu, \qquad i = 1, 2, \ldots, N. \tag{13}$$

- Para o caso de variáveis gaussianas, a nova FDP é chamada de *gaussiana padronizada*, sendo denotada por $Z \sim N(0, 1)$.

Gráfico da FDP Gaussiana de Média=0 e Variância=1



Entendendo a FDP Gaussiana Padronizada

- Com base no gráfico anterior da FDP gaussiana padronizada temos que:
  1. 68% dos valores de Z estarão no intervalo $[-1, +1]$.

$$\Pr(-1 \leq Z \leq +1) \approx 0{,}68$$

  2. 95% dos valores de Z estarão no intervalo $[-2, +2]$.

$$\Pr(-2 \leq Z \leq +2) \approx 0{,}95$$

  3. 99% dos valores de Z estarão no intervalo $[-3, +3]$.

$$\Pr(-3 \leq Z \leq +3) \approx 0{,}99$$

- Esta propriedade da FDP gaussiana padronizada é usada para detecção de observações atípicas ou anômalas.

Aplicação em Detecção de Anormalidades em Processos

- Suponha que estejamos monitorando (medindo e armazenando) uma grandeza física X que representa o estado de um processo industrial.
- Se há razões para assumir que $X \sim N(\mu, \sigma^2)$, então podemos usar a seguinte regra de decisão:

$$\begin{array}{ll} \text{SE} & Z = \frac{X-\mu}{\sigma} \notin [-3, +3], \\ \text{ENTÃO} & X \text{ é um estado anormal.} \end{array}$$

- No caso de detectar alguma anormalidade, o operador deve ser avisado.

Calculando Probabilidades a Partir das Amostras - Parte 1

- Vamos agora abordar o problema de calcular probabilidades $P(a \leq x \leq b)$ a partir de um conjunto de observações de uma certa variável aleatória X.
- Suponha que temos um conjunto de N medidas de uma certa variável X, ou seja,

$$\mathcal{X} = \{X(1), X(2), X(3), \ldots, X(N)\}.$$

- Vamos gerar artificialmente $N = 5000$ medidas de uma variável aleatória $X \sim N(100, 4)$.
- **Matlab/Octave**: `>> X=normrnd(100,sqrt(4),5000,1);`
- **Scilab**: `--> X=grand(5000,1,'nor',100,sqrt(4));`

Calculando Probabilidades a Partir das Amostras - Parte 2

- De posse do conjuntos de medidas $\mathcal{X}$, vamos determinar $P(X \leq 104)$ da seguinte maneira.
- O método consiste simplemente em contar quantos elementos do conjunto $\mathcal{X}$ são menores que ou iguais a 104.
- No Matlab/Octave/Scilab, usamos os seguintes comandos:

```
>> P=length(find(X<=104))/5000;
```

Gerando e Visualizando VAs Gaussianas no Octave/Matlab

- Os comandos abaixo geram e visualizam um conjunto de observações de uma variável contínua gaussiana de média $\mu = 10$ e variância $\sigma^2 = 2$.

```
>> mi=10; vari=2; % parametros da distribuicao
>> N=5000; % No.  de observacoes desejadas
>> X=normrnd(mi,sqrt(vari),N,1); % VAs gaussianas
>> mi=mean(X); % media amostral
   mi = 9.9907
>> s2=var(X,1); % variancia amostral
   s2 = 1.9932
>> histfit(X,20); % histograma de X
```

Entendendo o Histograma

Passo 1 - Calcular a média ($\bar{x}$) e a desvio padrão (s) amostrais de X.

Passo 2 - Definir o domínio de X: $[\bar{x} - k \cdot s, \bar{x} + k \cdot s]$ (e.g. $k = 5$).

Passo 3 - Discretizar o domínio de X em $M + 1$ pontos x_i, $i = 0, 1, \ldots, M$.

Passo 4 - Para cada intervalo $\Delta x_i = x_i - x_{i-1}$, $i = 1, \ldots, M$, determinar:

$$\boxed{C(\Delta x_i) = \text{No. de observações de } X \in \Delta x_i}$$

Passo 5 - Desenhar o gráfico de barras $C(\Delta x_i) \times \Delta x_i$.

Gráfico da FDP Empírica (Histograma)