FIAP

Statistics for Machine Learning

Aula 18: Distribuições Contínuas

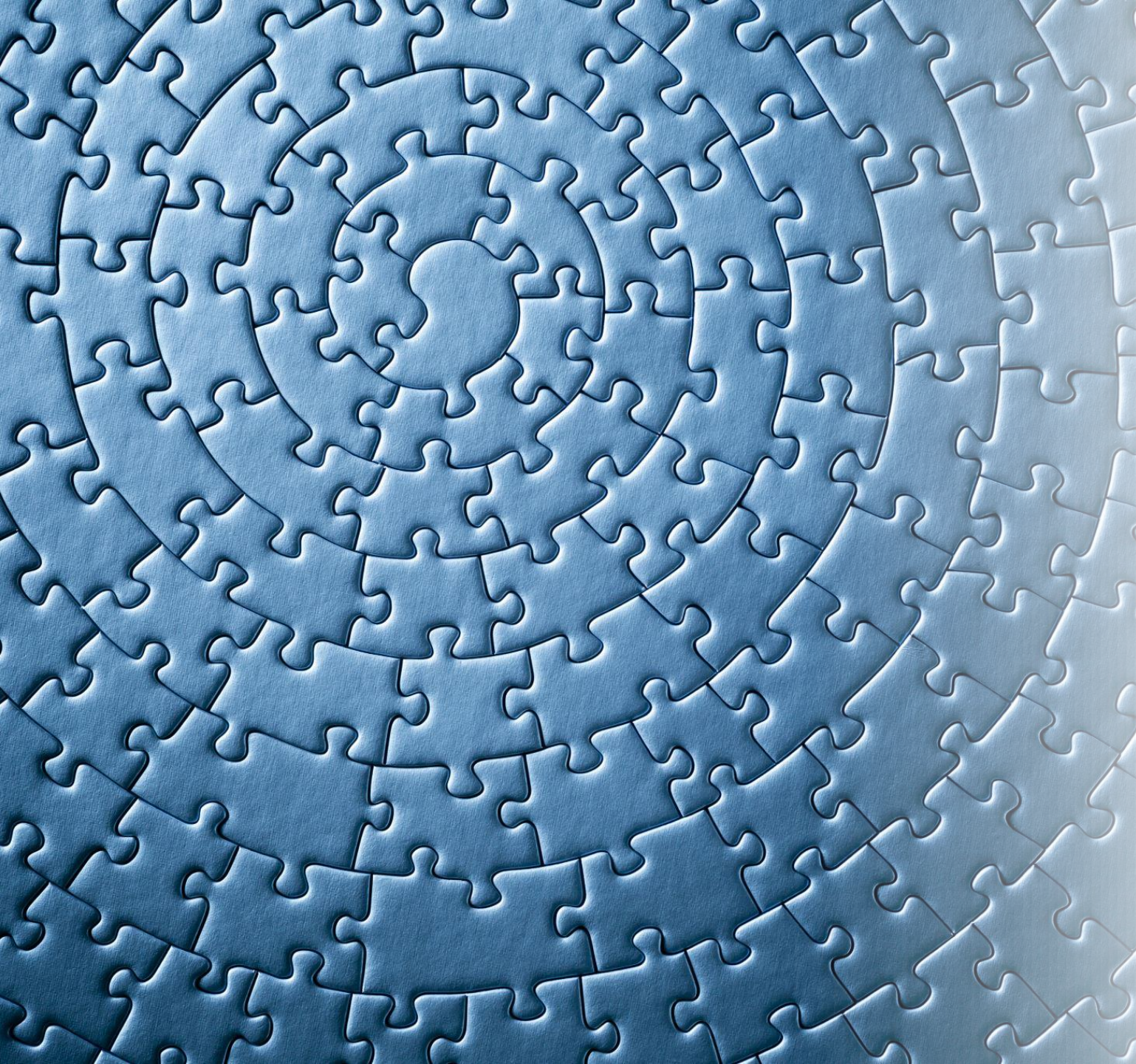
Prof. Jones Egydio

profjones.egydio@fiap.com.br

2024

Objetivos

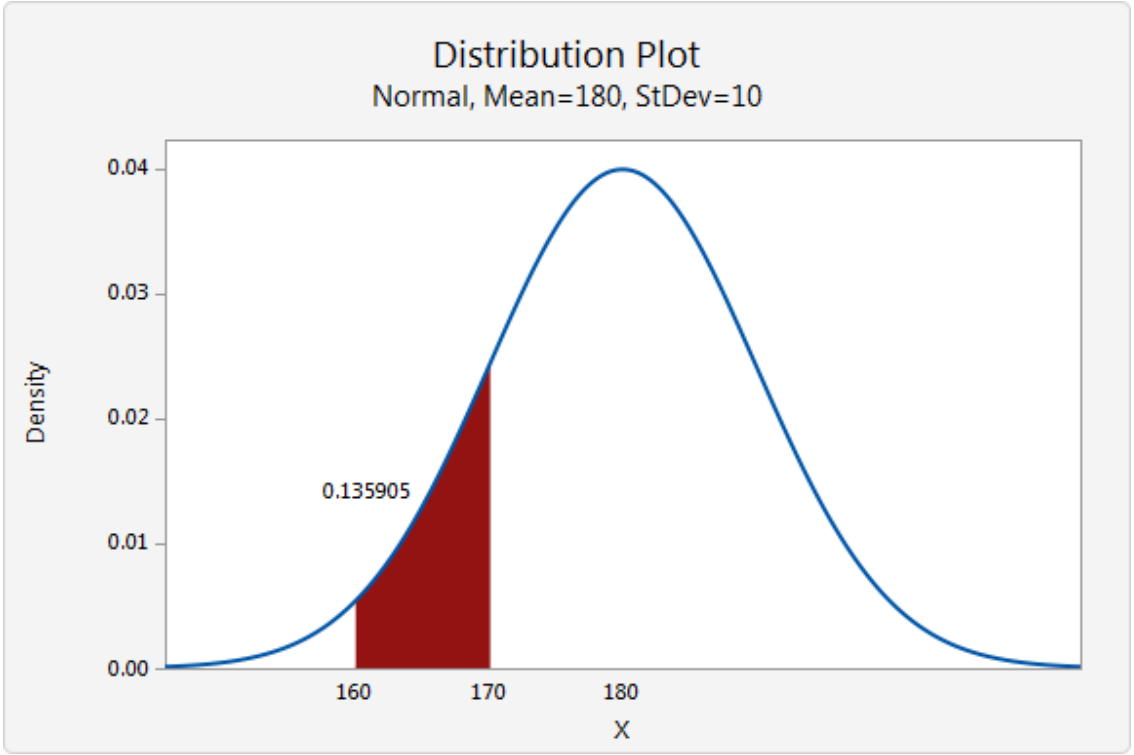
- Introduzir os conceitos de Distribuições Contínuas:
  - Exponencial;
  - Normal;
  - T-Student;
  - Qui-quadrado;
  - F;
- Formas de representação;
- Exemplos e exercícios;
- Conclusão;
- Perguntas.



Qual a probabilidade das vendas ocorrerem acima da meta?

- Projeção de vendas é uma das atividades importantes no mundo dos negócios.
- Por meio das projeções, indústria e varejistas podem definir ofertas, promoções e definição de preços.
- Como as distribuições de probabilidade podem auxiliar nesta atividade?

Conceitos iniciais



Distribuições de Probabilidade Contínuas

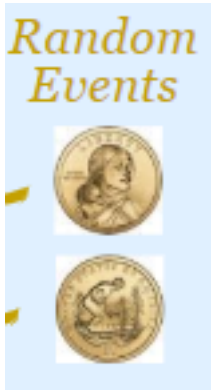
Conceitos

- Distribuições de probabilidade
  - Distribuições de probabilidade contínuas
  - Distribuição exponencial
  - Distribuição normal
  - Distribuição t
  - Distribuição Qui-Quadrado
  - Distribuição F

Tipos de variáveis aleatórias

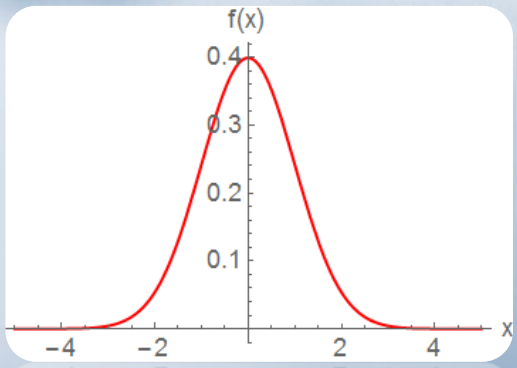
Discretas:

- Possui um número finito de resultados.
- Exemplo: Vivo/morto, grávida/não grávida, sucesso/fracasso, vitória/derrota/empate.



Contínuas:

- Possui uma grande amplitude de resultados, em que se torna praticamente impossível contar o número de resultados possíveis.
- Exemplo: Peso, altura, renda, velocidade de um carro



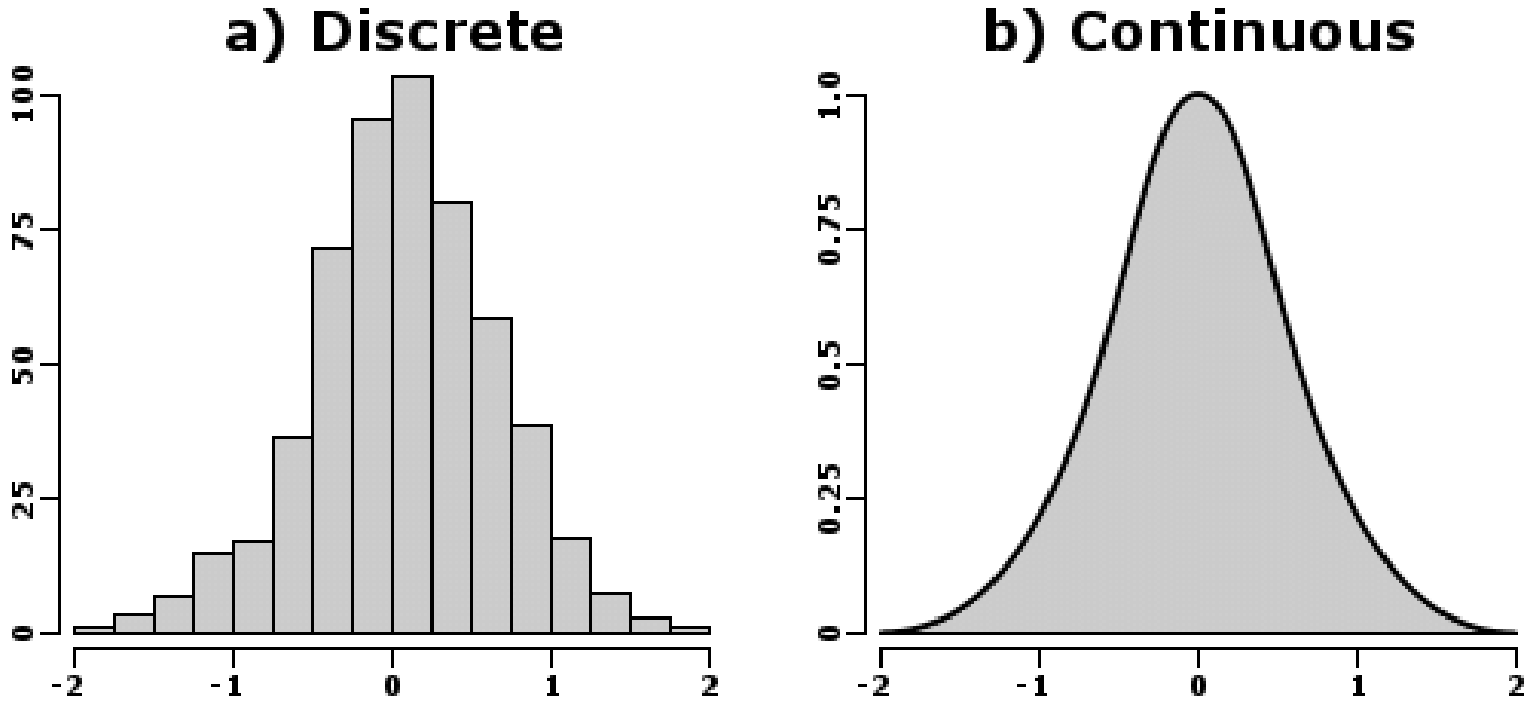
Combinação + Probabilidade = Distribuição de Probabilidade

Número de sucessos	Combinações possíveis	Probabilidade de Bernoulli	Probabilidade Binomial
0	1	$0{,}2^5 = 0{,}00032$	$1 \cdot 0{,}2^5 = 0{,}00032$
1	5	$0{,}8^1 \cdot 0{,}2^4 = 0{,}00128$	$5 \cdot 0{,}8^1 \cdot 0{,}2^4 = 0{,}0064$
2	10	$0{,}8^2 \cdot 0{,}2^3 = 0{,}00512$	$10 \cdot 0{,}8^2 \cdot 0{,}2^3 = 0{,}0512$
3	10	$0{,}8^3 \cdot 0{,}2^2 = 0{,}02048$	$10 \cdot 0{,}8^3 \cdot 0{,}2^2 = 0{,}2048$
4	5	$0{,}8^4 \cdot 0{,}2^1 = 0{,}08192$	$5 \cdot 0{,}8^4 \cdot 0{,}2^1 = 0{,}4096$
5	1	$0{,}8^5 = 0{,}3277$	$1 \cdot 0{,}8^5 = 0{,}3277$

Nesta tabela temos um conjunto de resultados de eventos associados as suas respectivas probabilidades. Ou seja, denominamos isto de **Distribuição de Probabilidades**.

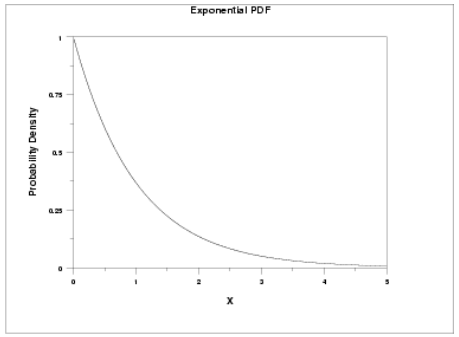
A probabilidade binomial é o resultado de uma sequência de eventos em que somente 2 resultados podem ocorrer. Neste exemplo, sucesso ou fracasso.

Diferença entre distribuições de probabilidade discretas e contínuas.

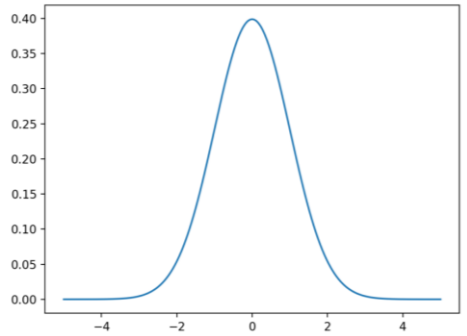


Distribuições de probabilidades contínuas

Distribuições que serão apresentadas:

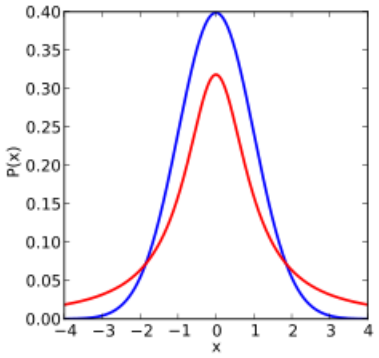


Distribuição exponencial

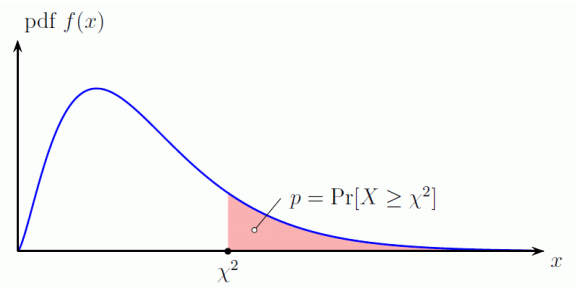


$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

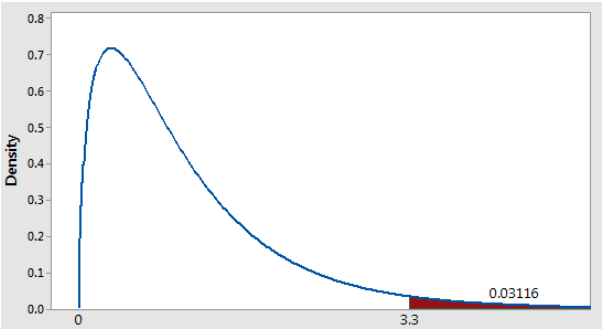
Distribuição normal



Distribuição t



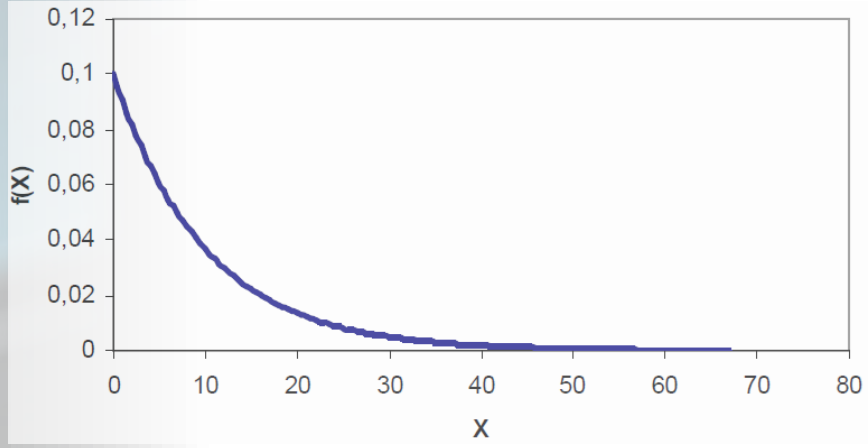
Distribuição de Qui-Quadrado



Distribuição de F

Distribuição Exponencial

Quando temos um problema de determinar a probabilidade de um evento ocorrer até um determinado momento (data, hora), temos uma **Distribuição Exponencial.**



Distribuição Exponencial

- Probabilidade de tempo de atendimento na fila do banco ser inferior a 10 minutos.

- Probabilidade de uma máquina apresentar defeito em 5 anos, sabendo que o tempo médio de vida é 15 anos.

Distribuição Exponencial

Em resumo, uma distribuição exponencial possui 3 características e a seguinte fórmula:

a) Consiste de 1 evento ocorrer até o fim do processo observado

b) Cada evento é independente

c) A probabilidade de cada evento ocorrer está entre 0 e 1

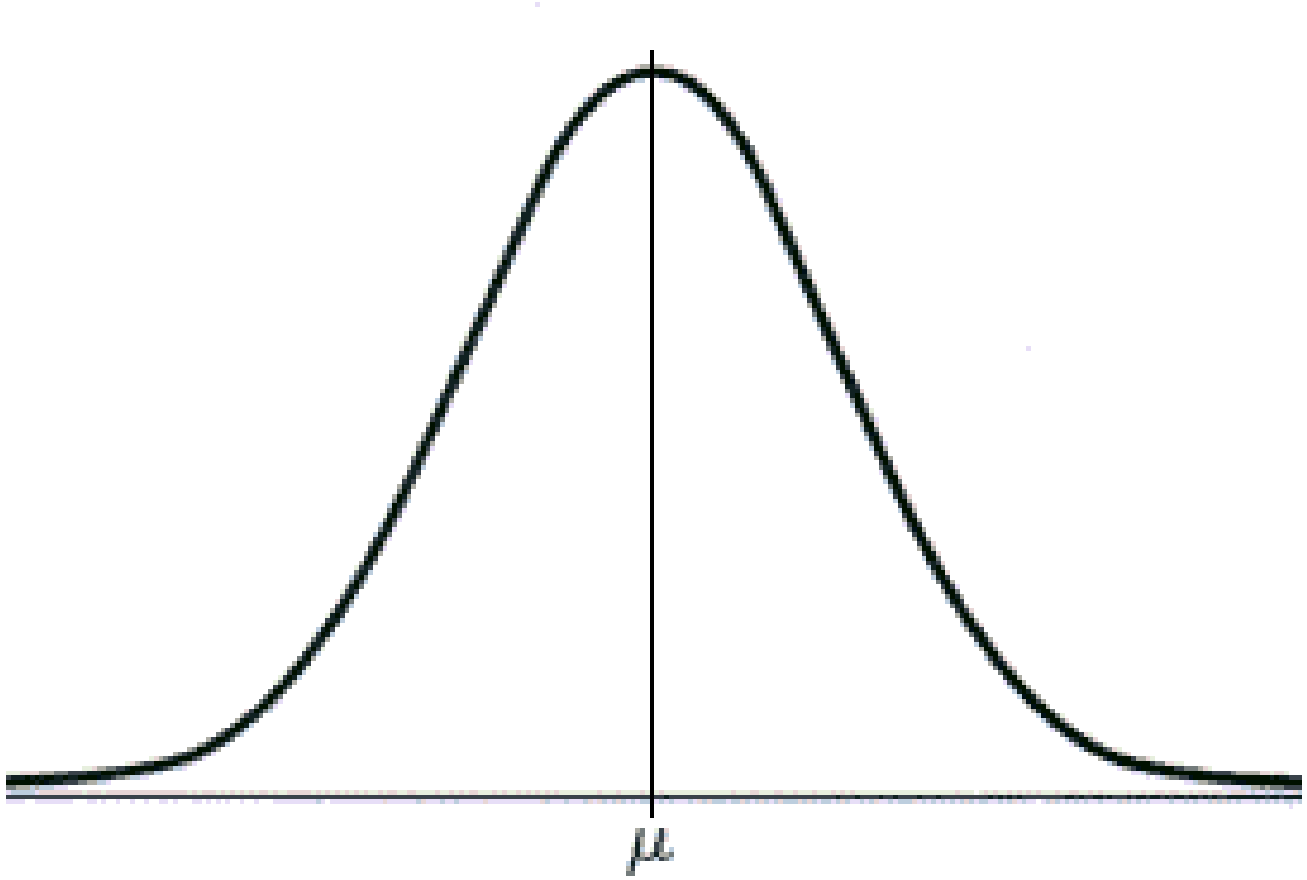
$$P(X) = 1 - e^{-x/\lambda}$$

Onde λ é a média.

Distribuição Normal

- A distribuição normal (gaussiana) é uma das distribuições mais importantes da área de análise de dados.

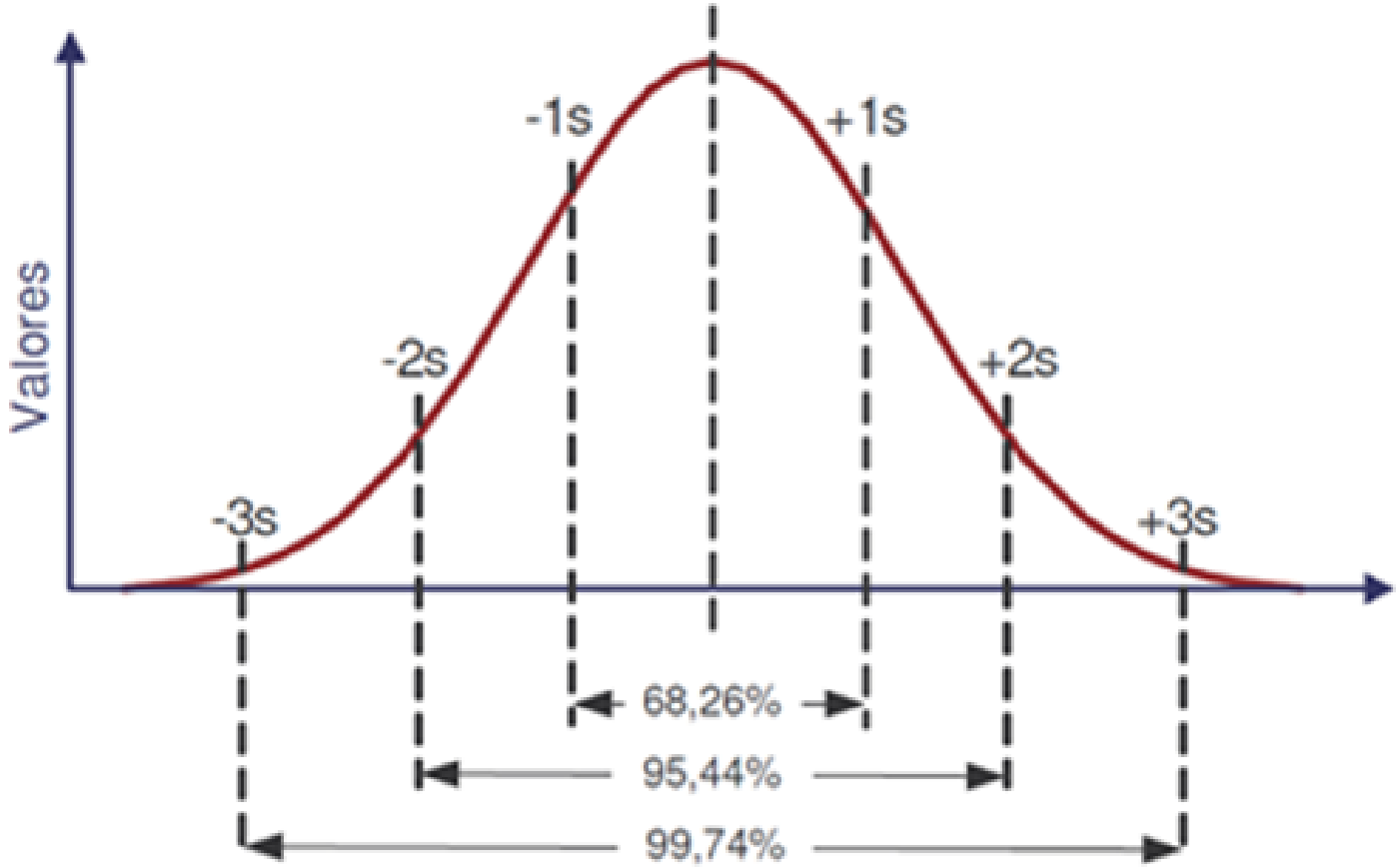
- É uma distribuição simétrica e tem formato de sino. Isso ocorre porque a maioria dos resultados se concentram em torno da média.



*Exemplo com números ver excel:
distrib_prob_continuas.xlsx, aba dist_normal

Distribuição Normal

- Possui a média no centro
- O tamanho do desvio padrão define as probabilidades.
- A probabilidade de um evento ocorrer dentro do primeiro desvio padrão é 68,26%.
- A probabilidade de um evento ocorrer dentro do segundo desvio padrão é 95,44%.
- A probabilidade de um evento ocorrer dentro do terceiro desvio padrão é 99,74%.



Distribuição Normal – exemplos

- O comportamento de uma ação de um mercado financeiro pode ser definido como os retornos sendo a média e o risco sendo o desvio padrão. Por meio destes parâmetros podemos presumir probabilidades próximas a distribuição normal.

- Quando estudamos as rendas das pessoas de uma cidade podemos encontrar uma renda média e verificar o quanto as rendas podem variar de um pessoa para outra.

- Quando estudamos as alturas das pessoas podemos encontrar uma altura média e verificar o quanto as alturas podem variar de um pessoa para outra

Distribuição Normal

Em resumo, uma distribuição exponencial possui 3 características e a seguinte fórmula:

a) Distribuição construída por meio da média e desvio-padrão dos eventos

b) Cada evento é independente

c) A distribuição é simétrica, sendo que os valores se concentram perto da média.

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}$$

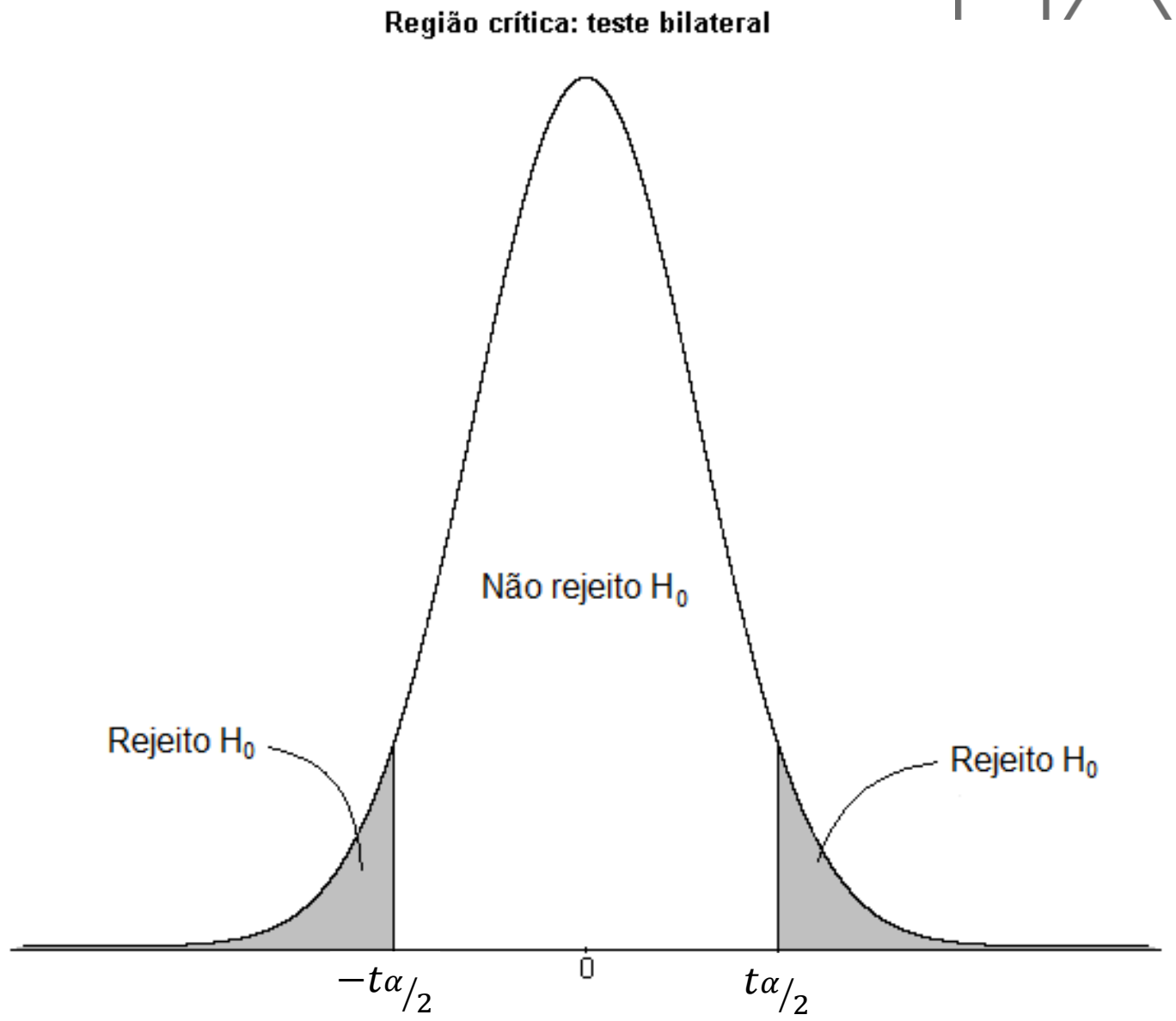
Distribuição t-Student

- A distribuição t de Student possui importância na realização de testes de hipóteses.
- A principal utilização seria na comparação de médias de 2 populações.
- Por exemplo, qual cidade possui renda maior, cidade A ou cidade B?

Distribuição t

- A distribuição t é simétrica e possui forma bastante similar com a distribuição normal.

- Porém, esta distribuição possui caudas e altura maior que a distribuição normal quando a amostra é pequena.



Distribuição t

Em resumo, uma distribuição t possui 3 características e a seguinte fórmula:

a) Possui caudas e altura maiores que a dist. normal

b) Cada evento é independente

c) Principal uso em testes de hipóteses

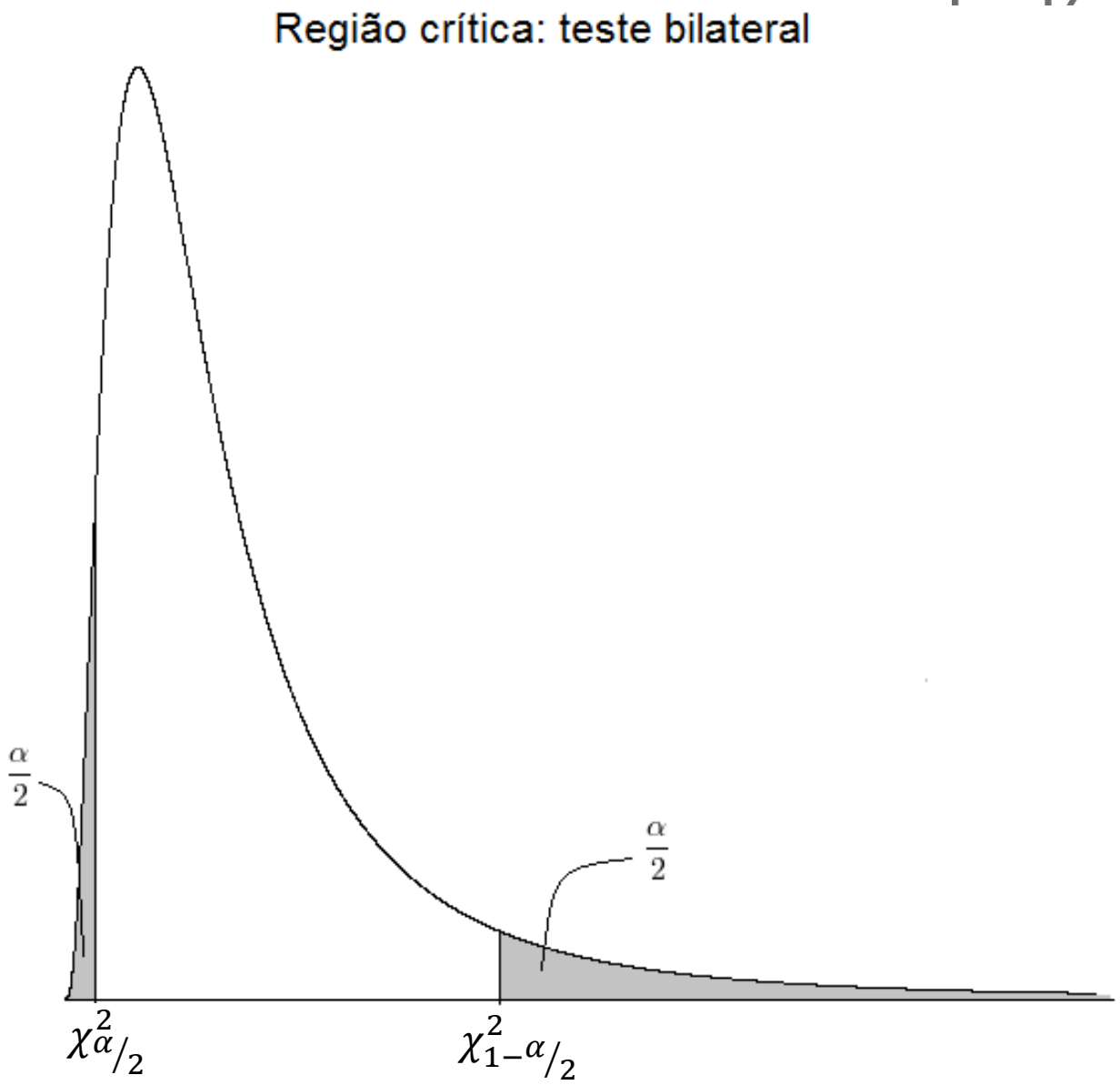
$$t = \frac{x - \bar{x}}{s/\sqrt{n}}$$

Distribuição Qui-Quadrado

- A distribuição Qui-Quadrado possui importância na realização de testes de hipóteses.
- Possui diversas aplicações como testar se uma população A possui distribuição de renda similar a alguma distribuição esperada.
- Outro exemplo de uso seria verificar a associação entre dois grupos. Por exemplo, pode-se testar a aderência de homens e mulheres a cursos de tecnologia.

Distribuição Qui-Quadrado

- A distribuição Qui-Quadrado é assimétrica.
- Somente assume valores positivos.



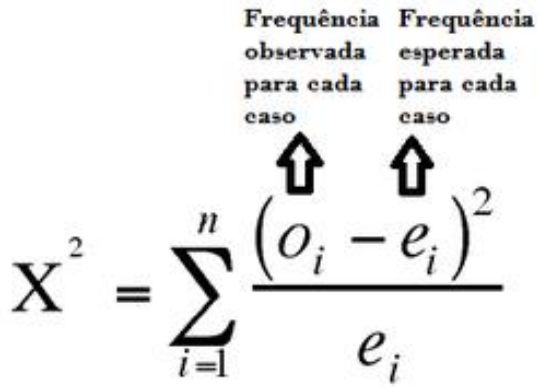
Distribuição t

Em resumo, uma distribuição Qui-Quadrado possui 3 características e a seguinte fórmula:

a) Assume somente valores positivos

b) Tende a ser assimétrica

c) Principal uso em testes de hipóteses

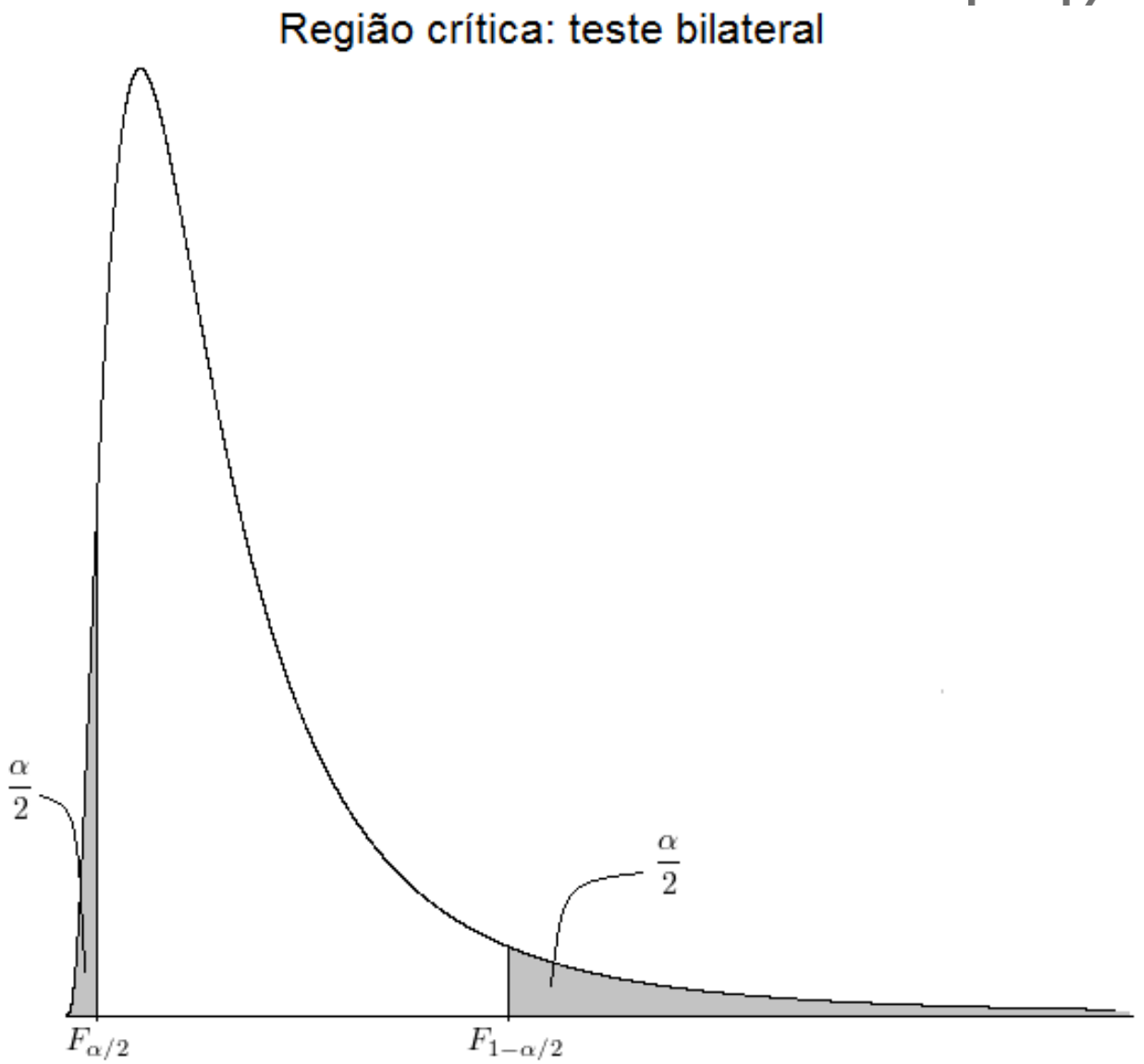


Distribuição F

- A distribuição F possui importância na realização de testes de hipóteses.
- Ela é gerada pelo quociente entre duas variáveis.
- Possui aplicações em diversos modelos estatísticos como regressão linear e ANOVA (Analysis of Variance).

Distribuição F

- A distribuição F é assimétrica.
- Somente assume valores positivos.



Distribuição F

Em resumo, uma distribuição F possui 3 características:

a) assume somente valores positivos

b) Tende a ser assimétrica

c) Principal uso em testes de hipóteses

$$F = \frac{S_1^2}{S_2^2}$$

Voltando ao nosso problema

Qual a probabilidade das vendas ocorrerem acima da meta?

- Projeção de vendas é uma das atividades importantes no mundo dos negócios.
- Por meio das projeções, indústria e varejistas podem definir ofertas, promoções e definição de preços.
- Como as distribuições de probabilidade podem auxiliar nesta atividade?

Voltando ao nosso problema

Qual a probabilidade das vendas ocorrem acima da meta? (Vamos considerar que em média a empresa vende 50 unidades por semana com desvio padrão de 10 unidades.)

- Qual a distribuição de probabilidade você assumiria para resolver este problema?
- Qual a probabilidade da empresa vender entre 30 e 45 unidades?
- O Marketing definiu uma meta de vendas para a semana de 85 unidades. Qual a probabilidade desta meta ser atingida?

Qual a probabilidade da empresa vender entre 30 e 45 unidades?

Adotamos uma distribuição normal e fazemos o seguinte:

Primeiro calculamos a probabilidade até 45 unidades:

$$P(X \leq 45) = 30{,}85\%$$

Depois calculamos a probabilidade até 30 unidades:

$$P(X \leq 30) = 2{,}28\%$$

Finalmente, fazemos a diferença para obter a probabilidade:

$$P(30 \leq X \leq 45) = 30{,}85\% - 2{,}28\% = 28{,}58\%$$

O Marketing definiu uma meta de vendas para a semana de 85 unidades. Qual a probabilidade desta meta ser atingida?

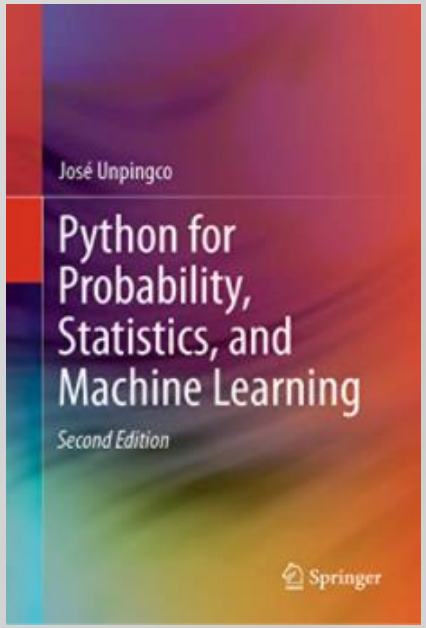
Adotamos uma distribuição normal e fazemos o seguinte:

Primeiro calculamos a probabilidade até 85 unidades, depois subtraímos 1 – probabilidade.

$$P(X \geq 85) = 1 - P(X \leq 85) = 0{,}0233\%$$

Onde estudar mais!!

- Leitura



- Vídeos

- Distribuições contínuas: https://pt.khanacademy.org/math/ap-statistics/random-variables-ap/continuous-random-variables/v/probabilities-from-density-curves

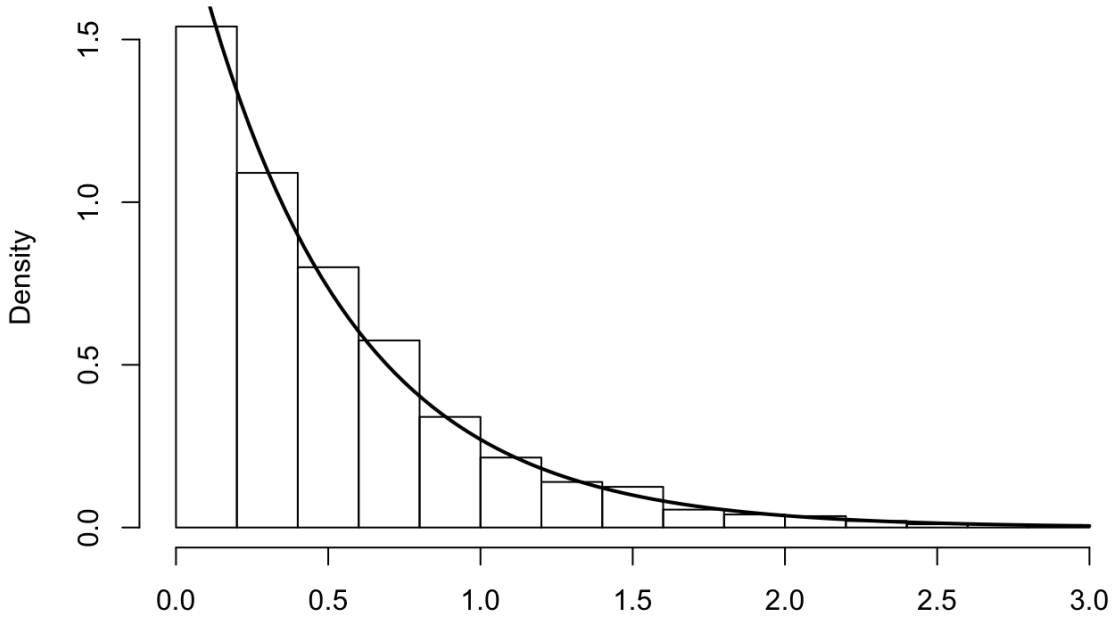
- Distribuição Normal: https://pt.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/more-on-normal-distributions/v/introduction-to-the-normal-distribution

- Distribuições contínuas: https://pt.khanacademy.org/math/ap-statistics/random-variables-ap/continuous-random-variables/v/probabilities-from-density-curves

Parte 2
Prática

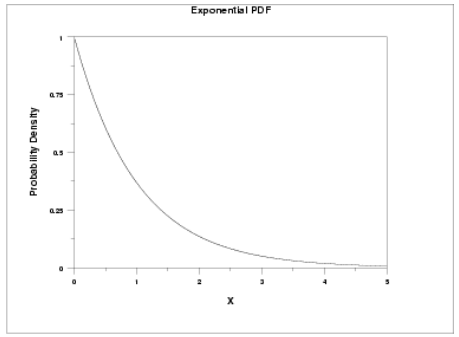
Conceitos iniciais

- Distribuições de Probabilidade Contínuas (Aplicações no Python)

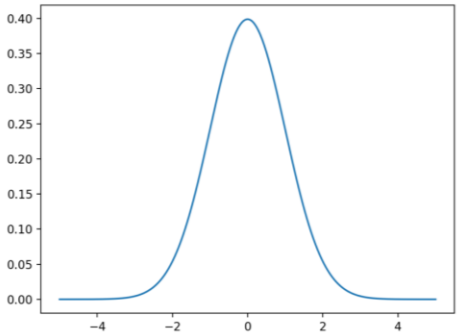


Distribuições de probabilidades contínuas

- Distribuições que serão apresentadas:

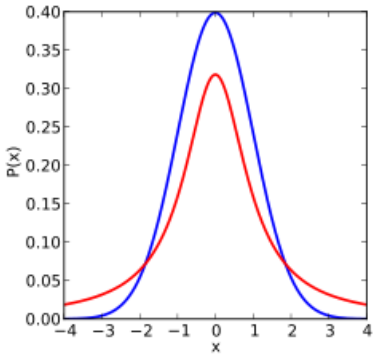


Distribuição exponencial

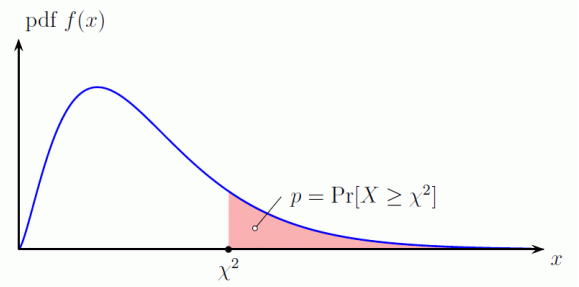


$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

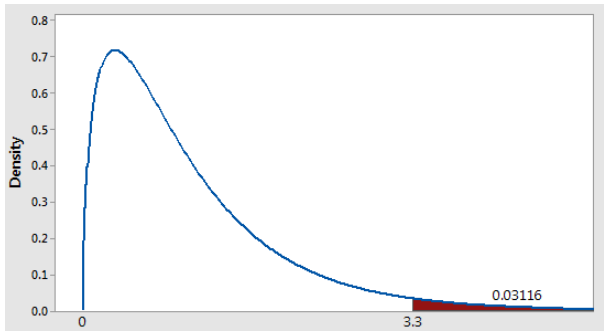
Distribuição normal



Distribuição t



Distribuição de Qui-Quadrado



Distribuição de F

Distribuição Exponencial (resumo)

- Em resumo, uma distribuição exponencial possui 3 características e a seguinte fórmula:

- a) consiste de 1 evento ocorrer até o fim do processo observado
- b) cada evento é independente
- c) a probabilidade de cada evento ocorrer está entre 0 e 1

- $P(X) = 1 - e^{-\frac{x}{\lambda}}$, λ é a média.

```
expon.pdf(x, scale)
```

Distribuição Exponencial

- Exemplo distribuição exponencial:

```
n = 60
duracao_media = 12
```

```
resultados = np.arange(61)
```

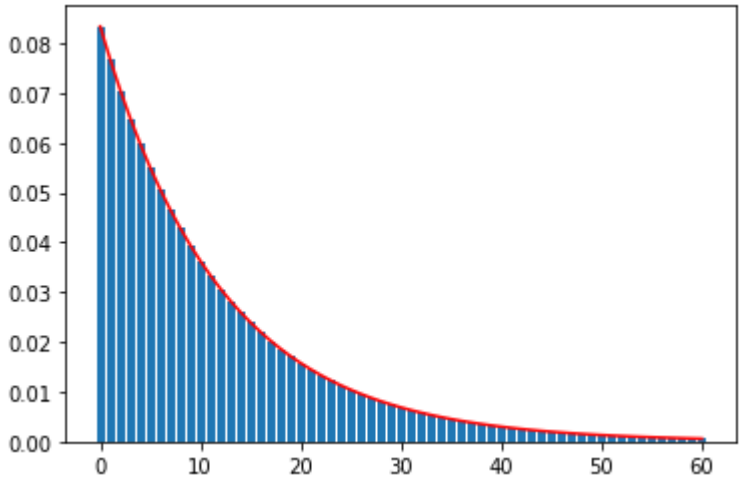
```
probabilidades = expon.pdf(x = resultados, scale = duracao_media)
```

Simular 60 eventos e mantem uma duração média de 12.

```
dados_exponencial = pd.DataFrame(np.transpose([resultados, probabilidades]),
                                 columns = ['resultados', 'probabilidades'])
```

```
plt.bar(dados_exponencial['resultados'], dados_exponencial['probabilidades'])
plt.plot(dados_exponencial['resultados'], dados_exponencial['probabilidades'], color = 'red')
```

Geramos o gráfico.



Distribuição Normal (resumo)

- Em resumo, uma distribuição normal possui 3 características e a seguinte fórmula:

- a) distribuição construída por meio da média e desvio-padrão dos eventos
- b) cada evento é independente
- c) A distribuição é simétrica, sendo que os valores se concentram perto da média.

- $P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}$.

```
norm.pdf(x, scale, loc)
```

Distribuição Normal

- Exemplo distribuição normal:

```
n = 70
media = 30
desvio_padrao = 6
```

Define media e desvio padrão

```
resultados = np.linspace(0, 70, 141)
```

A função linspace é utilizada para gerar sequencias com intervalos definidos. Neste exemplo utilizamos “0.5” (141 espaços entre 0 e 70).

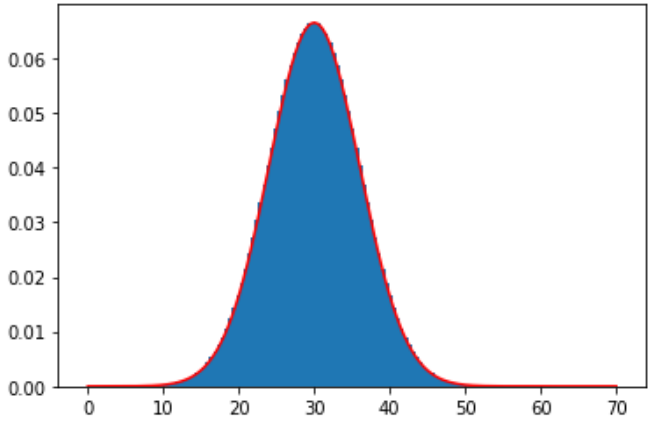
```
probabilidades = norm.pdf(x = resultados, loc = media, scale = desvio_padrao)
```

A função norm.pdf gera as probabilidades da distribuição normal.

```
dados_normal = pd.DataFrame(np.transpose([resultados, probabilidades]),
                            columns = ['resultados', 'probabilidades'])
```

```
plt.bar(dados_normal['resultados'], dados_normal['probabilidades'])
plt.plot(dados_normal['resultados'], dados_normal['probabilidades'], color = 'red')
```

Geramos o gráfico.



Distribuição t (resumo)

- Em resumo, uma distribuição t possui 3 características e a seguinte fórmula:

- a) possui caudas e altura maiores que a dist. normal
- b) cada evento é independente
- c) principal uso em testes de hipóteses

- $t = \frac{x-\bar{x}}{s/\sqrt{n}}$; esta função é uma simplificação.

```
t.pdf(x, loc, scale, df)*2
```

Distribuição t

- Exemplo Distribuição t:

Define media e desvio padrão

```
n = 70
media = 30
desvio_padrao = 6
```

```
resultados = np.linspace(0, 70, 141)
```

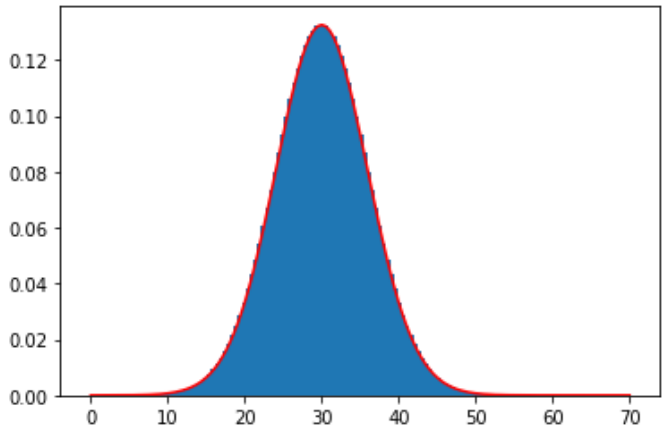
A função t.pdf gera as probabilidades da distribuição t.

```
probabilidades = t.pdf(x=resultados, loc=media, scale=desvio_padrao, df=n-1)*2
```

```
dados_t = pd.DataFrame(np.transpose([resultados, probabilidades]),
                       columns = ['resultados', 'probabilidades'])
```

```
plt.bar(dados_t['resultados'], dados_t['probabilidades'])
plt.plot(dados_t['resultados'], dados_t['probabilidades'], color = 'red')
```

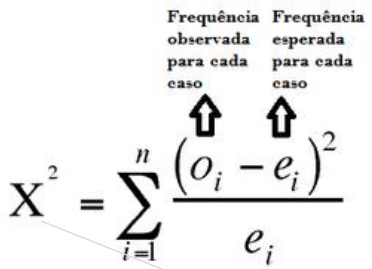
Geramos o gráfico.



Distribuição Qui-Quadrado (resumo)

- Em resumo, uma distribuição Qui-Quadrado possui 3 características e a seguinte fórmula:

- a) assume somente valores positivos
- b) Tende a ser assimétrica
- c) principal uso em testes de hipóteses



Também é uma simplificação.

```
chi2.pdf(x, df)
```

Distribuição Qui-Quadrado

- Exemplo distribuição Qui-Quadrado:

```
n = 10
```

```
resultados = np.linspace(0, 30, 61)
```

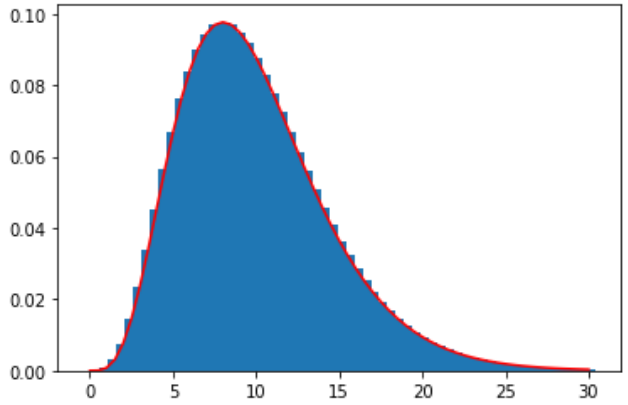
```
probabilidades = chi2.pdf(x=resultados, df=n)
```

A função chi2.pdf gera as probabilidades da distribuição qui-quadrado.

```
dados_chi2 = pd.DataFrame(np.transpose([resultados, probabilidades]),
                          columns = ['resultados', 'probabilidades'])
```

```
plt.bar(dados_chi2['resultados'], dados_chi2['probabilidades'])
plt.plot(dados_chi2['resultados'], dados_chi2['probabilidades'], color = 'red')
```

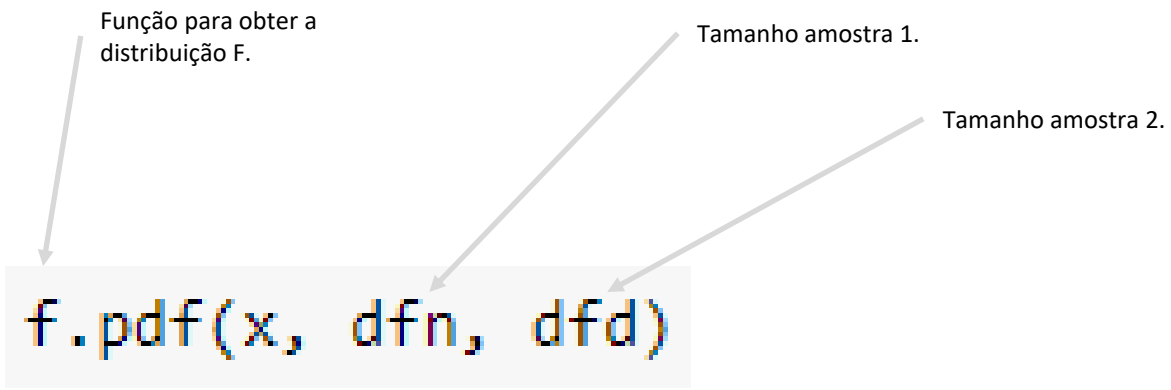
Geramos o gráfico.



Distribuição F (resumo)

- Em resumo, uma distribuição F possui 3 características e a seguinte fórmula:

- a) assume somente valores positivos
- b) Tende a ser assimétrica
- c) principal uso em testes de hipóteses



Distribuição F

- Exemplo distribuição F:

```
n1 = 5
n2 = 4
```

```
resultados = np.linspace(0, 30, 61)
```

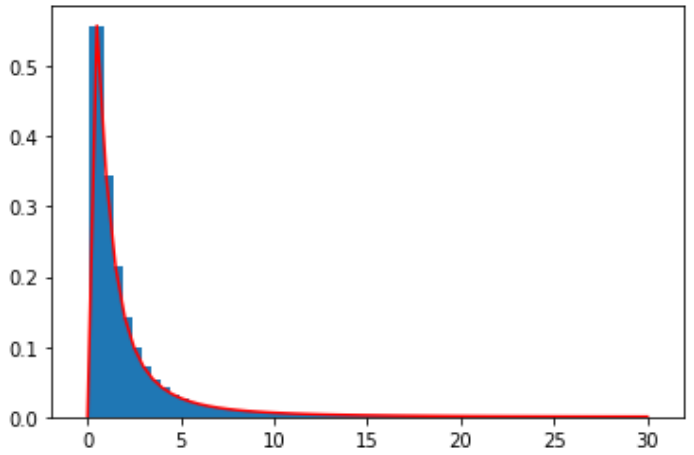
```
probabilidades = f.pdf(x=resultados, dfn=n1-1, dfd=n2-1)
```

A função “f.pdf” gera as probabilidades da distribuição F.

```
dados_f = pd.DataFrame(np.transpose([resultados, probabilidades]),
                       columns = ['resultados', 'probabilidades'])
```

```
plt.bar(dados_f['resultados'], dados_f['probabilidades'])
plt.plot(dados_f['resultados'], dados_f['probabilidades'], color = 'red')
```

Geramos o gráfico.



Problema

- Projeção de vendas é uma das atividades importantes no mundo dos negócios.
- Por meio das projeções, indústria e varejistas podem definir ofertas, promoções e definição de preços.
- Como as distribuições de probabilidade podem auxiliar nesta atividade?

Qual a probabilidade das vendas ocorrerem acima da meta?

Problemas

- **Qual a probabilidade das vendas ocorrem acima da meta? (Vamos considerar que em média a empresa vende 50 unidades por semana com desvio padrão de 10 unidades.)**

- Qual a distribuição de probabilidade você assumiria para resolver este problema?

- Qual a probabilidade da empresa vender entre 30 e 45 unidades?

- O Marketing definiu uma meta de vendas para a semana de 85 unidades. Qual a probabilidade desta meta ser atingida?

Problemas

- Qual a probabilidade da empresa vender entre 30 e 45 unidades?
- Adotamos uma distribuição normal e fazemos o seguinte:

```
media = 50
dp = 10
```

Define media e desvio padrão.

```
prob_35 = norm.cdf(x = 30, loc = media, scale = dp)
```

Obtém as probabilidades até 30 e até 40. Utilizamos uma nova função "norm.cdf", esta função nos apresenta a probabilidade acumulada.

```
prob_45 = norm.cdf(x = 45, loc = media, scale = dp)
```

```
prob_45 - prob_35
```

Subtrai as duas áreas.

```
0.2857874067778077
```

Resultado final.

Problemas

- O Marketing definiu uma meta de vendas para a semana de 85 unidades. Qual a probabilidade desta meta ser atingida?

```
media = 50
dp = 10
```

Define media e desvio padrão.

```
prob_85 = norm.cdf(x = 85, loc = media, scale = dp)
```

Obtém as probabilidades até 85.

```
1 - prob_85
```

Subtrai a área depois de 85 unidades

```
0.00023262907903554 01
```

Resultado final.

Problemas

- **Carregar o arquivo “vinho_nacional.csv” e responda as perguntas:**

- Qual a média mensal de vendas de vinhos nacional?
- Qual o desvio padrão de vendas de vinhos nacional?
- Qual a probabilidade da empresa vender entre 330 e 370 garrafas de vinho nacional?
- O Marketing definiu uma meta de vendas para o mês de 370 garrafas de vinho nacional. Qual a probabilidade desta meta ser superada?

Problemas

- Qual a média mensal de vendas de vinhos?
- Qual o desvio padrão de vendas de vinhos?

```
dados.agg(media_vendas = pd.NamedAgg('vendas_vinho_nacional', 'mean'),
          dp_vendas = pd.NamedAgg('vendas_vinho_nacional', 'std'))
```

	vendas_vinho_nacional
media_vendas	339.050000
dp_vendas	7.301586

Média e desvio padrão

Problemas

- Qual a probabilidade da empresa vender entre 330 e 370 garrafas?

```
media_vendas = dados_resumo[dados_resumo['index'] == 'media_vendas'].values[0][1]
```

```
dp_vendas = dados_resumo[dados_resumo['index'] == 'dp_vendas'].values[0][1]
```

```
p_330 = norm.cdf(x = 330, loc = media_vendas, scale = dp_vendas)
```

```
p_370 = norm.cdf(x = 370, loc = media_vendas, scale = dp_vendas)
```

```
p_370 - p_330
```

```
0.8924005653958927
```

A empresa possui 89,24% de chance de vender entre 330 e 370 garrafas.

Problemas

- O Marketing definiu uma meta de vendas para o mês de 370 garrafas. Qual a probabilidade desta meta ser superada?

```
media_vendas = dados_resumo[dados_resumo['index'] == 'media_vendas'].values[0][1]
```

```
dp_vendas = dados_resumo[dados_resumo['index'] == 'dp_vendas'].values[0][1]
```

```
p_370 = norm.cdf(x = 370, loc = media_vendas, scale = dp_vendas)
```

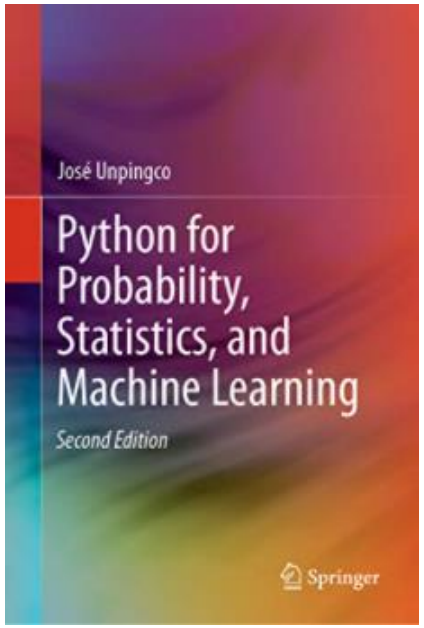
```
(1 - p_370) * 100
```

```
0.00112356312291783 67
```

A empresa possui 0,0011% de chance de superar a meta.

Onde estudar mais!!

- Leitura



- Distribuições contínuas:
  https://pt.khanacademy.org/math/ap-statistics/random-variables-ap/continuous-random-variables/v/probabilities-from-density-curves

- Vídeos

- Distribuições contínuas:
  https://pt.khanacademy.org/math/ap-statistics/random-variables-ap/continuous-random-variables/v/probabilities-from-density-curves

- Distribuição Normal:
  https://pt.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/more-on-normal-distributions/v/introduction-to-the-normal-distribution

Obrigado!