FIAP

MBA

MBA EM DATA SCIENCE & AI

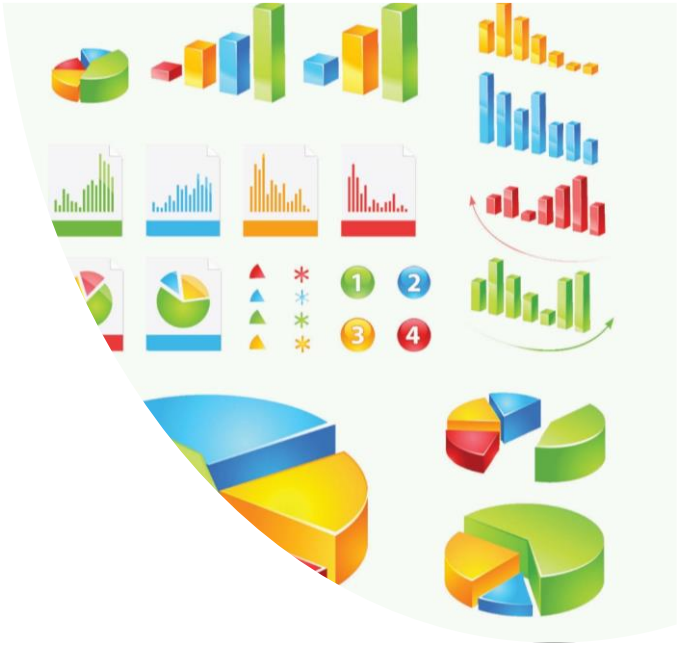
APPLIED STATISTICS

AULA 3

Análise gráfica

Teoria de probabilidade

Distribuição de probabilidades



Plots

Análises Gráficas
com o Python

Por onde começar

- Você já conhece minimamente seus dados?
- Quais perguntas quer responder?

Pacotes de plotagens mais conhecidos

- **Matplotlib**
  - **Descrição:** Um dos pacotes de visualização mais utilizados em Python, conhecido por sua flexibilidade e capacidade de criar uma ampla variedade de gráficos, desde gráficos simples até figuras complexas.
  - **Uso Comum:** Gráficos de linha, barras, histogramas, gráficos de dispersão, etc.
- **Seaborn**
  - **Descrição:** Construído sobre o Matplotlib, o Seaborn oferece uma interface mais amigável e ferramentas para criar gráficos estatísticos mais atraentes e informativos.
  - **Uso Comum:** Mapas de calor, gráficos de violino, gráficos de distribuição, etc.
- **Plotly**
  - **Descrição:** Uma biblioteca de gráficos interativos que pode ser usada em notebooks Jupyter e também em aplicações web. Oferece gráficos em 2D e 3D.
  - **Uso Comum:** Gráficos interativos, gráficos de linhas, gráficos de dispersão 3D, mapas, etc.

seaborn

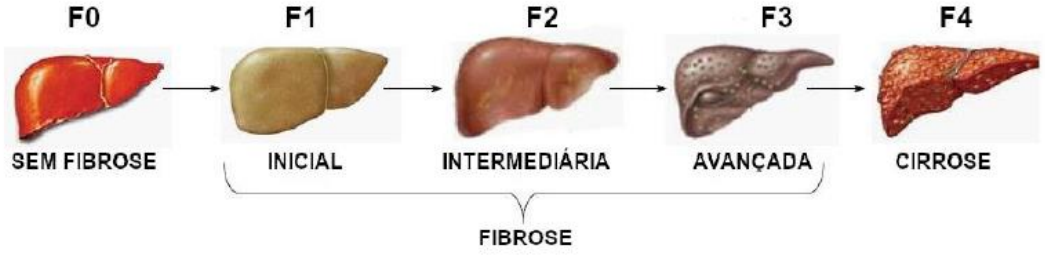
HTTPS://SEABORN.PYDATA.ORG/

Preparando o terreno

```
import seaborn as sns
import matplotlib.pyplot as plt
```

Exemplo

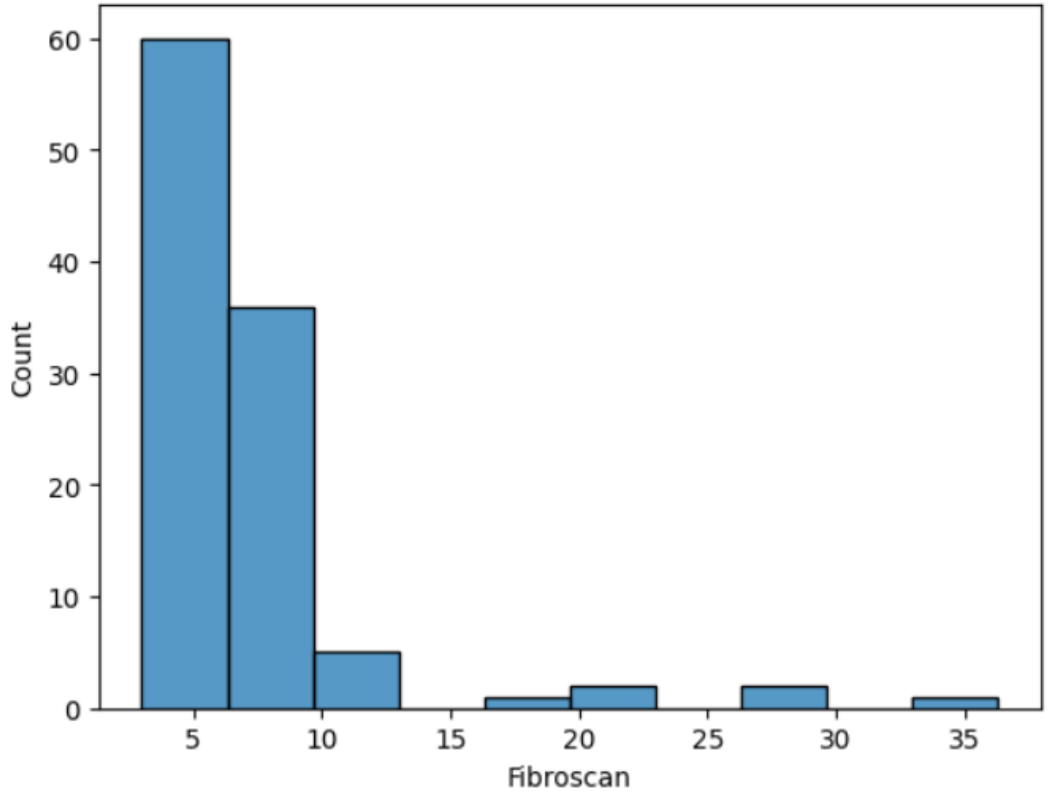
1) Leia a base basefibrose.csv
2) Calcule as medidas de resumo para as variáveis contínuas
3) Calcule a distribuição absoluta e relativa da variável Grau de Fibrose.
4) Faça os gráficos de histograma e boxplot.
5) Faça o mesmo por Grau de Fibrose.
6) Faça um gráfico de barras cruzando Grau de Fibrose com sexo



Histograma

```
[28] sns.histplot(x = 'Fibroscan', bins = 10, data = df)
```

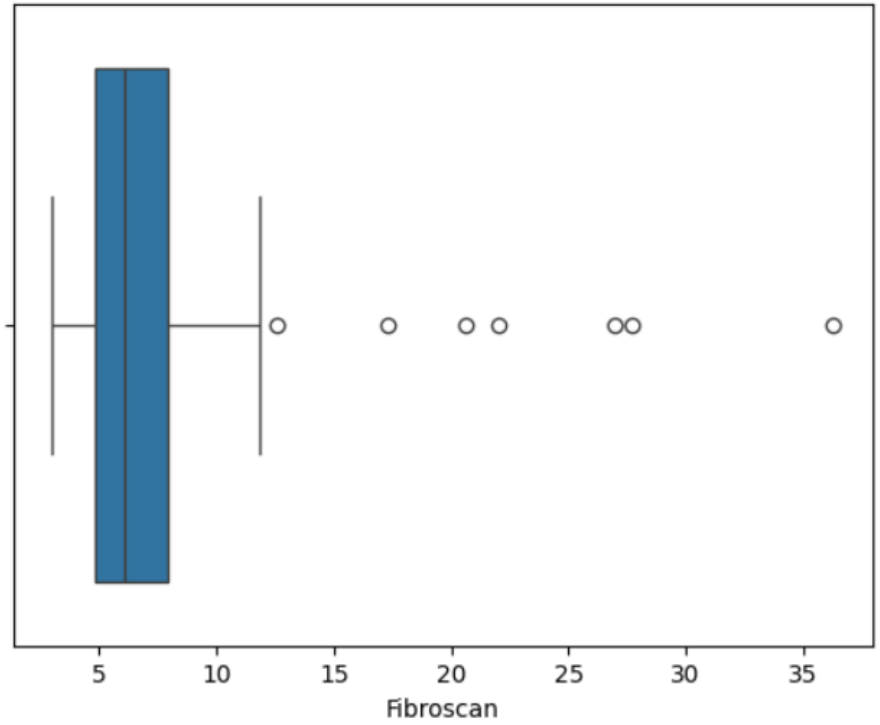
```
<Axes: xlabel='Fibroscan', ylabel='Count'>
```



Boxplots

```
sns.boxplot(x = 'Fibroscan' ,  data = df)
```

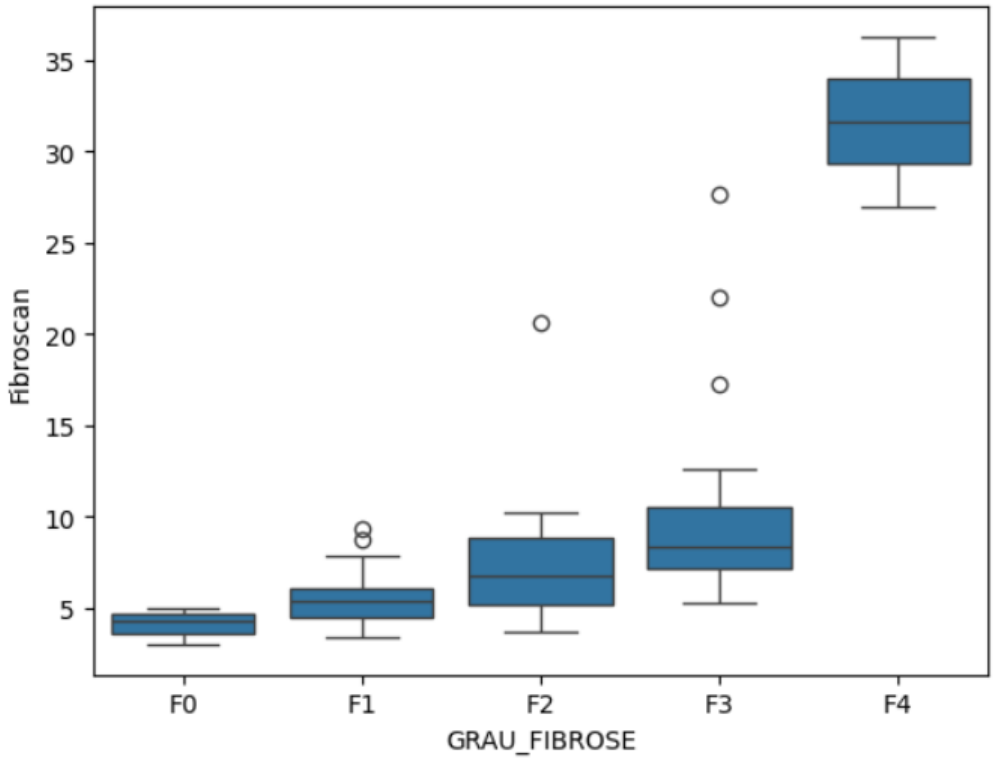
```
<Axes: xlabel='Fibroscan'>
```



Boxplots por variável

```
sns.boxplot(x = 'GRAU_FIBROSE', y = 'Fibroscan' ,  data = df, order = ['F0', 'F1', 'F2', 'F3', 'F4'])
```

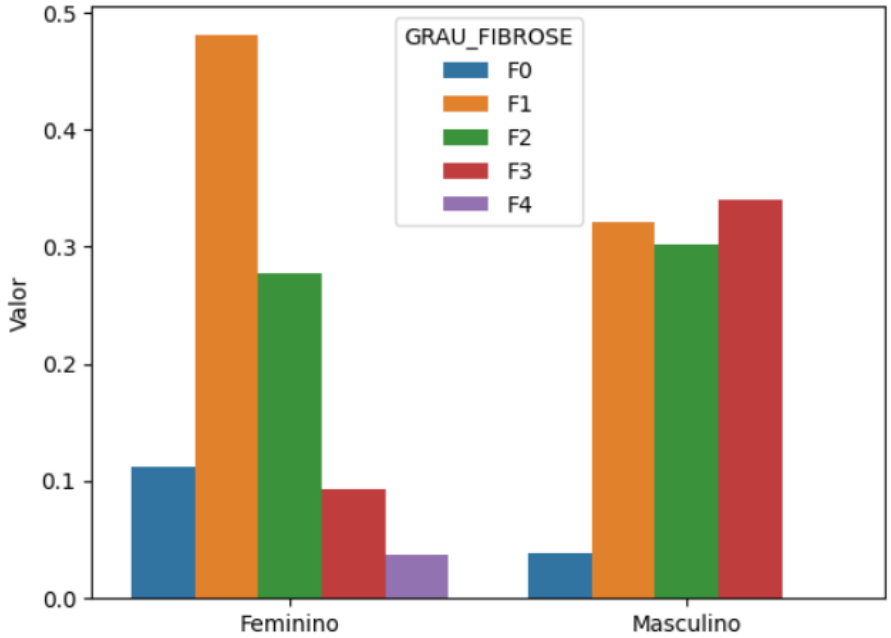
```
<Axes: xlabel='GRAU_FIBROSE', ylabel='Fibroscan'>
```



Gráficos de barras

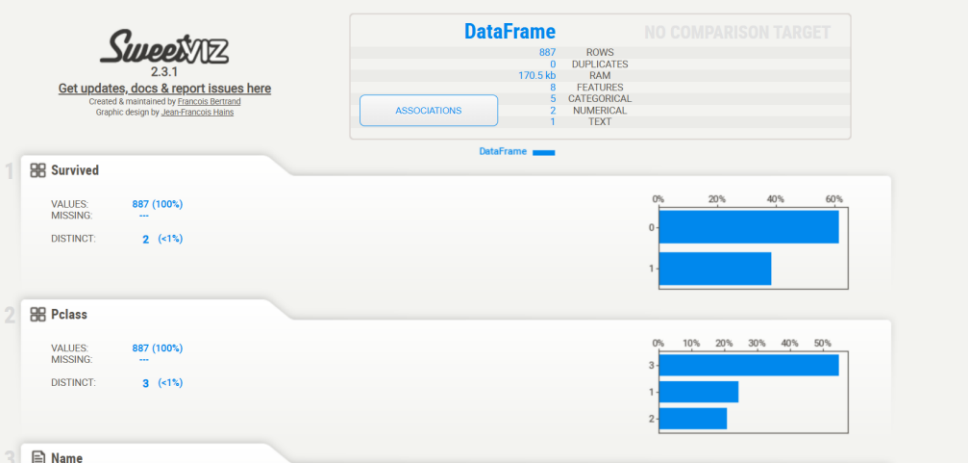
```
df_bar = df.groupby(['Sexo'])['GRAU_FIBROSE'].value_counts(normalize=True).rename('proportion').reset_index()
```

```
sns.barplot(x='Sexo', y='proportion', hue='GRAU_FIBROSE', data = df_bar, hue_order = ['F0', 'F1', 'F2', 'F3', 'F4'])
plt.ylabel('Valor')
plt.show()
```



Relatatório Sweet ViZ

```
!pip install sweetviz
import sweetviz
relatorio = sweetviz.analyze(df)
relatorio.show_html('RELATORIO.html')
```



Como medir incerteza? (Probabilidade?)

Experimento Aleatório: procedimento que, ao ser repetido sob as mesmas condições, pode fornecer resultados diferentes.

Exemplos:

E_1: Lançamento de um dado e observar a face superior.

E_2:Lançamento de uma moeda quatro vezes e observar o número de caras.

E_3: Acompanhar os 30 alunos matriculados na disciplina e observar o número de aprovados.

E_4: Ligar uma lâmpada nova e observar o seu tempo de duração (em minutos).

Espaço Amostral (Ω): Conjunto de todos os resultados possíveis de um experimento aleatório.

Aos experimentos aleatórios exemplificados anteriormente estão associados os seguintes espaços amostrais, respectivamente:

$\Omega_1 = \{ 1, 2, 3, 4, 5, 6 \}.$

$\Omega_2 = \{ 0, 1, 2, 3, 4 \}.$

$\Omega_3 = \{ 0, 1, 2, \ldots 28, 29, 30 \}.$

$\Omega_4 = \{ t \in R \mid t \geq 0 \}.$

Evento: É um subconjunto de elementos do espaço amostral.

Aos espaços amostrais exemplificados anteriormente estão associados os seguintes eventos, respectivamente

A_1 = { 2, 4, 6 }, ou seja, obter uma face par.

B_2 = { 2 }, ou seja, obter duas caras.

C_3 = { 24, 25, 26, 27, 28, 29, 30 }, ou seja, pelo menos 80% de alunos aprovados na disciplina.

D_4 = { $t \geq 10000$ }, ou seja, a lâmpada durar pelo menos 10000 minutos.

Probabilidade: é uma medida da incerteza associada aos resultados do experimento aleatório.

Para calcularmos a probabilidade de um determinado evento A acontecer num determinado espaço amostral Ω, realizamos a seguinte conta

$$P(A) = \frac{\#A}{\#\Omega}$$

Número de casos favoráveis sobre o número de casos possíveis

Probabilidade: é uma medida da incerteza associada aos resultados do experimento aleatório.

Para calcularmos a probabilidade de um determinado evento A acontecer num determinado espaço amostral Ω, realizamos a seguinte conta

$$P(A) = \frac{\#A}{\#\Omega}$$

Visão clássica

Número de casos favoráveis sobre o número de casos possíveis

Por exemplo, ao lançarmos uma moeda equilibrada sabemos que, teoricamente, cada face tem a mesma probabilidade de ocorrência, isto é, P(C) = P($\overline{C}$) = ½.

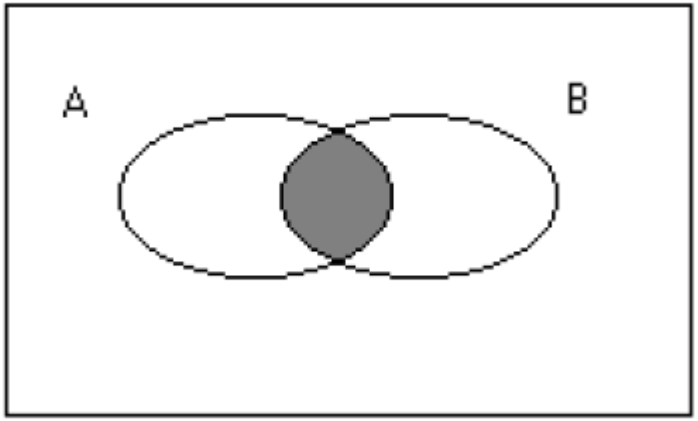
Axiomas de Probabilidade

Dado um espaço amostral, Ω, suponha que estamos estudando um evento A. A probabilidade do evento A ocorrer é denotada por P(A). A função P(A) só será uma probabilidade se ela satisfaz três condições básicas:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1)+P(A_2)+P(A_3)+...$, se os eventos A_1, A_2,... forem disjuntos (isto é, mutuamente exclusivos).

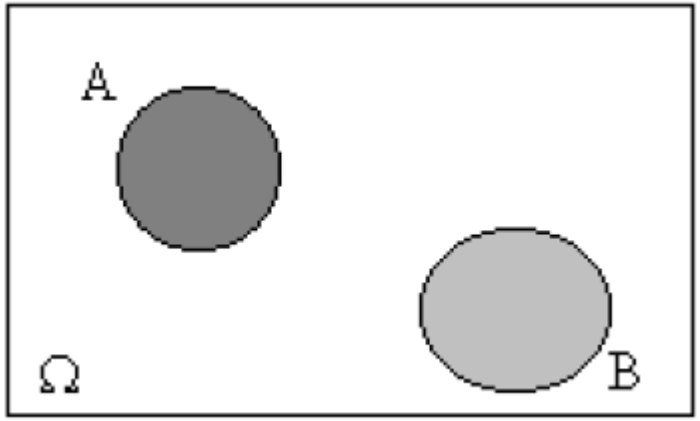
Intersecção de eventos

A interseção de dois eventos A e B corresponde à ocorrência simultânea dos eventos A e B. Contém todos os pontos do espaço amostral comuns a A e B. É denotada por $A \cap B$. A interseção é ilustrada pela área hachurada do diagrama de Venn abaixo.



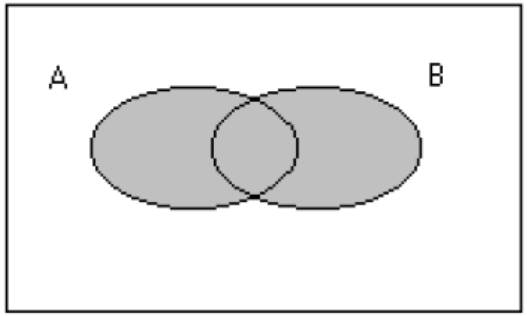
Eventos disjuntos ou mutualmente exclusivos

Dois eventos A e B são chamados disjuntos ou mutuamente exclusivos quando não puderem ocorrer juntos, ou seja, quando não têm elementos em comum, isto é, $A \cap B = \varnothing$. O diagrama de Venn a seguir ilustra esta situação.



União de eventos

A união dos eventos A e B equivale à ocorrência de A, ou de B, ou de ambos, ou seja, a ocorrência de pelo menos um dos eventos A ou B. É denotada por $A \cup B$. A área hachurada na figura abaixo ilustra esta situação.

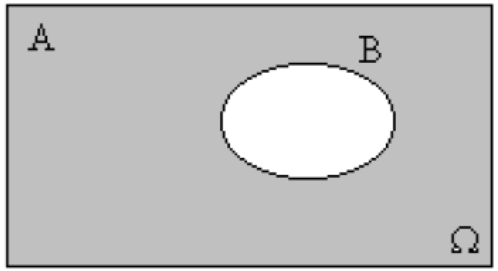


Para encontrar a união de dois eventos deve-se utilizar a seguinte fórmula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Eventos complementares

Dois eventos A e B são complementares se sua união corresponde ao espaço amostral e sua interseção é vazia. O diagrama a seguir ilustra tal situação.



- Para dois eventos A e B serem complementares: $A \cup B = \Omega$ e $A \cap B = \varnothing$. Além disso, $A^C = B$ e $B^C = A$, ou seja, o complementar do evento A ocorre quando o evento A não ocorrer!
- Pode-se observar também que: $P(A) = 1 - P(B)$ e que $P(A^C) = P(B) = 1 - P(A)$.

Exemplo 1

- Estudo da relação entre o hábito de fumar e a causa da morte, entre 1000 empresários.

Fumante	Causa da morte			Total
	Câncer (C)	Doença cardíaca (D)	Outros (O)	
Sim (F)	135	310	205	650
Não (F^C)	55	155	140	350
Total	190	465	345	1000

Um indivíduo é selecionado aleatoriamente entre os observados na amostra. Determine as seguintes probabilidades:

a.) Ser fumante.
b.) Ter morrido de câncer.
c.) Não ser fumante e ter morrido de doença cardíaca.
d.) Ser fumante ou ter morrido de outras causas.

Solução 1

a.) $P(F) = \frac{650}{1000} = 0{,}65$

b.) $P(C) = \frac{190}{1000} = 0{,}19$

c.) $P(F^c \cap D) = \frac{155}{1000} = 0{,}155$

d.) $P(F \cup O) = P(F) + P(O) - P(F \cap O) = \frac{650}{1000} + \frac{345}{1000} - \frac{205}{1000} = 0{,}790$

Probabilidade Condicional

Em diversas situações práticas, a probabilidade de ocorrência de um evento A se modifica quando dispomos de informação sobre a ocorrência de um outro evento associado.

A probabilidade condicional de A dado B é a probabilidade de ocorrência do evento A, sabido que o evento B já ocorreu. Pode ser determinada dividindo-se a probabilidade de ocorrência de ambos os eventos A e B pela probabilidade de ocorrência do evento B, como é mostrado a seguir:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \ , \ P(B) > 0$$

Da definição de probabilidade condicional, deduzimos a regra do produto de probabilidades que é uma relação bastante útil:

$$P(A \cap B) = P(A \mid B).P(B) \ , \ P(B) > 0$$

Independência de eventos

Dois eventos A e B são independentes se a ocorrência de um deles não afeta a probabilidade de ocorrência do outro, ou seja, $P(A \mid B) = P(A)$ e $P(B \mid A) = P(B)$. Se dois eventos A e B são independentes então $P(A \cap B) = P(A).P(B)$.

Exemplo 2

- Estudo da relação entre criminoso e vitima

Criminoso	Vítima			Total
	Homicídio (H)	Furto (F)	Assalto (A)	
Estranho (E)	12	379	727	1118
Conhecido (C)	39	106	642	787
Ignorado (I)	18	20	57	95
Total	69	505	1426	2000

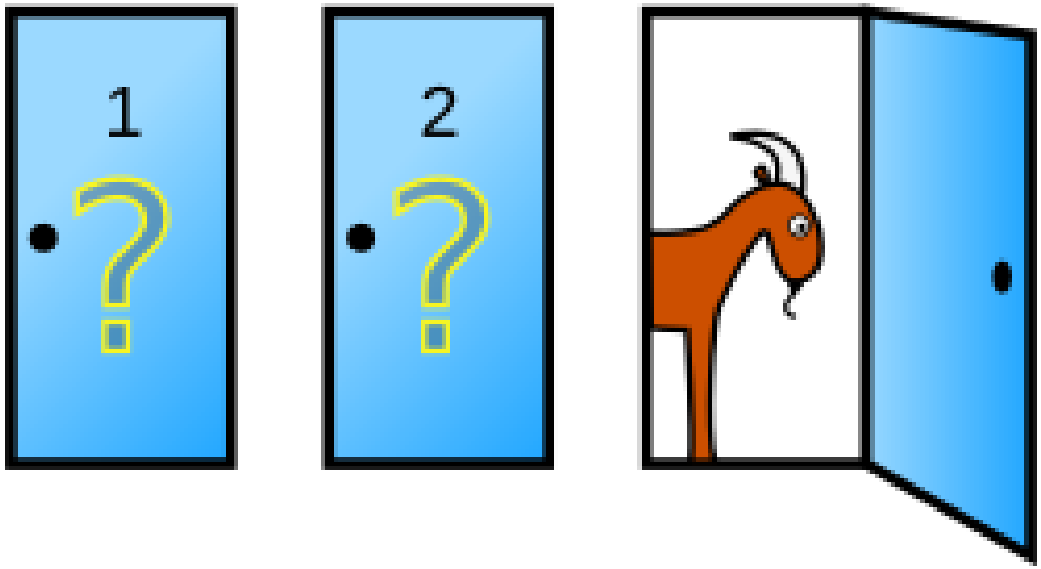
Solução 2

a.) $P(H \cup E) = P(H) + P(E) - P(H \cap E) = \frac{69}{2000} + \frac{1118}{2000} - \frac{12}{2000} = 0{,}587$

b.) $P(C \mid A) = \frac{P(C \cap A)}{P(A)} = \frac{642/2000}{1426/2000} = 0{,}450$

c.) $P(F \mid E) = \frac{P(F \cap E)}{P(E)} = \frac{379}{1118} = 0{,}338$

Problema de Monty Hall



Problema de Monty Hall

O jogo consistia no seguinte: Monty Hall, o apresentador, apresentava três portas aos concorrentes. Atrás de uma delas estava um **prêmio (um carro)** e, atrás das outras duas, **dois bodes**.

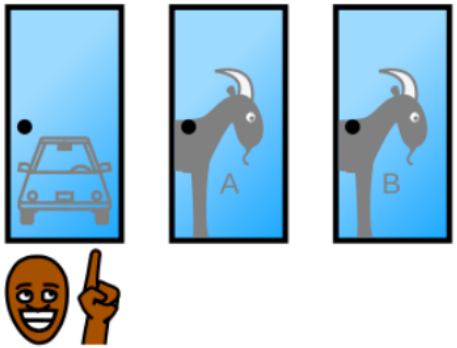
- Na 1.ª etapa o concorrente escolhe uma das três portas (que ainda não é aberta);
- Na 2.ª etapa, Monty abre uma das outras duas portas que o concorrente não escolheu, revelando que o carro não se encontra nessa porta e revelando um dos bodes;
- Na 3.ª etapa Monty pergunta ao concorrente se quer decidir permanecer com a porta que escolheu no início do jogo ou se ele pretende mudar para a outra porta que ainda está fechada para então a abrir. Agora, com duas portas apenas para escolher — pois uma delas já se viu, na 2.ª etapa, que não tinha o prêmio — e sabendo que o carro está atrás de uma das restantes duas, o concorrente tem que tomar a decisão.

Problema de Monty Hall

Qual é a estratégia mais lógica? Ficar com a porta escolhida inicialmente ou mudar de porta? Com qual das duas portas ainda fechadas o concorrente tem mais probabilidades de ganhar? Por quê?

Solução

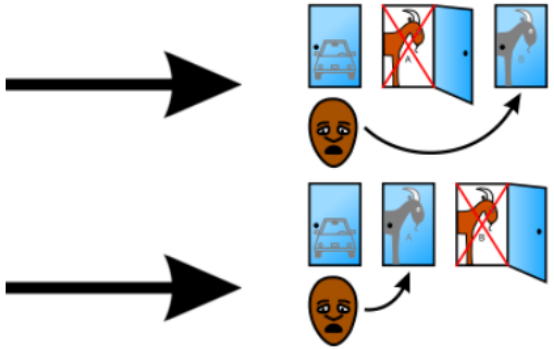
1.



Jogador escolhe carro

(probabilidade 1/3)

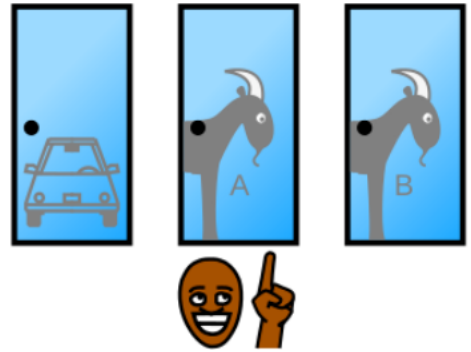
Apresentador revela um dos bodes



Trocar perde.

Solução

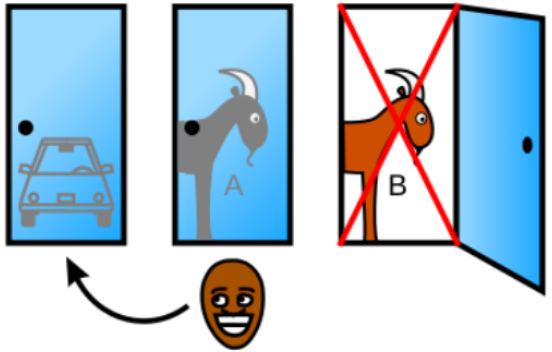
2.



Jogador escolhe Bode A

(probabilidade 1/3)

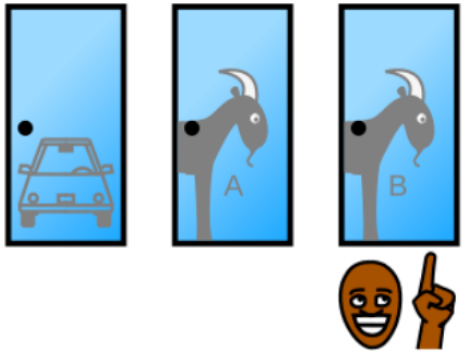
Apresentador tem que revelar Bode B



Trocar ganha.

Solução

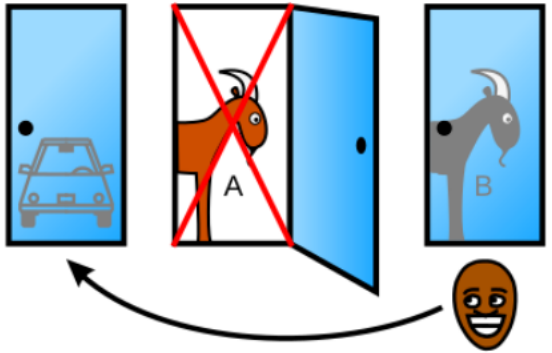
3.



Jogador escolhe Bode B

(probabilidade 1/3)

Apresentador tem que revelar Bode A



Trocar ganha.

Solução

Conclusão, vale a pena trocar de porta!

Variável Aleatória

Definição: Uma **variável aleatória (v.a.)** é uma função que descreve os resultados de um experimento através de valores numéricos.

Uma variável aleatória pode ser classificada em dois tipos:

- Variável Aleatória Discreta
- Variável Aleatória Contínua

Variável aleatória Discreta

Definição: Denomina-se X uma **variável aleatória discreta** se o número de valores possíveis de X for um conjunto de **pontos finito** ou **infinito enumerável**.

Exemplos:

- Número de ações vendidas de uma empresa.
- Número de erros de transmissão em um processo.
- Número de aparelhos defeituosos em uma produção.

Variável aleatória Contínua

Definição: Denomina-se X uma **variável aleatória contínua** se ela assume valores num **intervalo de números reais.**

Exemplos:

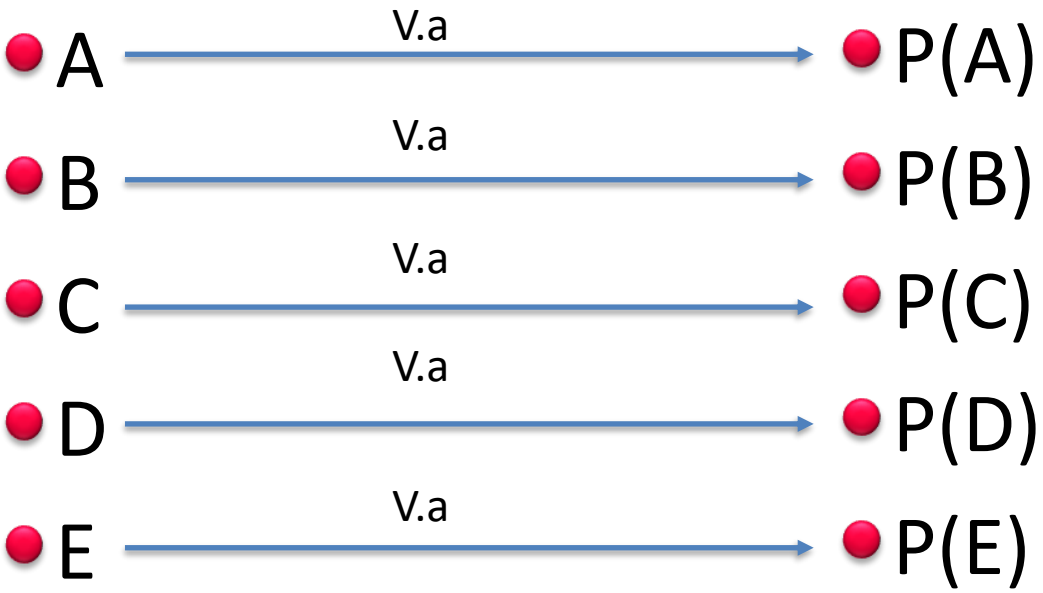
- resistência de um material;
- concentração de CO2 na água
- tempo de vida de um componente eletrônico;
- tempo de resposta de um sistema computacional;

Distribuição de probabilidade

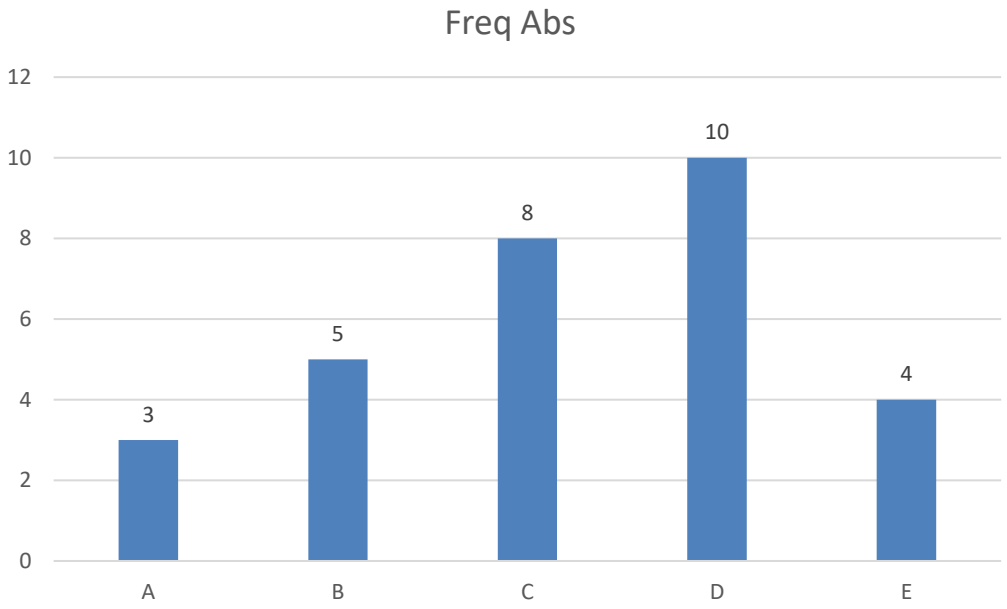
Definição: A distribuição de probabilidade descreve as probabilidades dos resultados numéricos de um experimento, ou seja, associa **uma probabilidade** de cada valor de uma variável aleatória.

Existem, na literatura, diversas distribuições de probabilidades para modelar **variáveis aleatórias discretas e contínuas.**

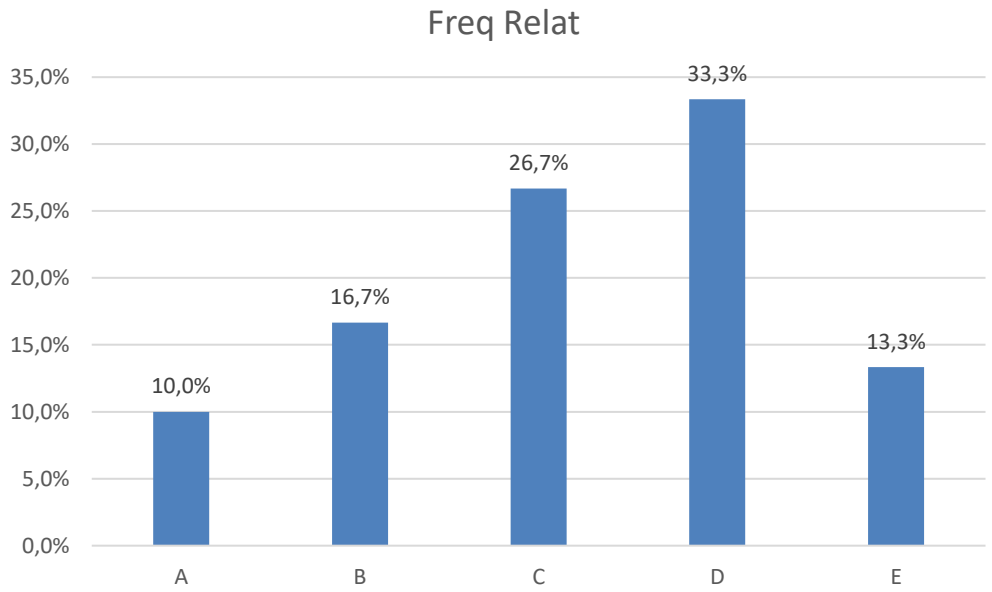
Variável aleatória



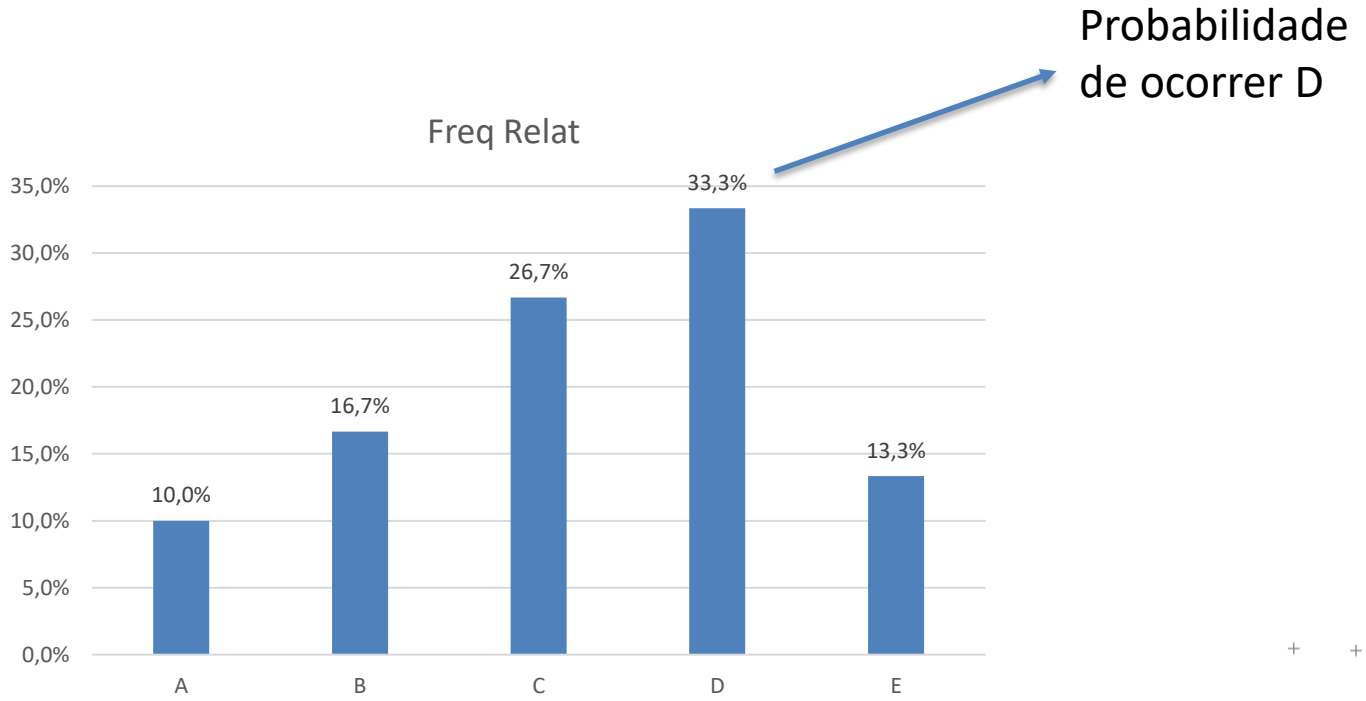
Voltando um pouco...



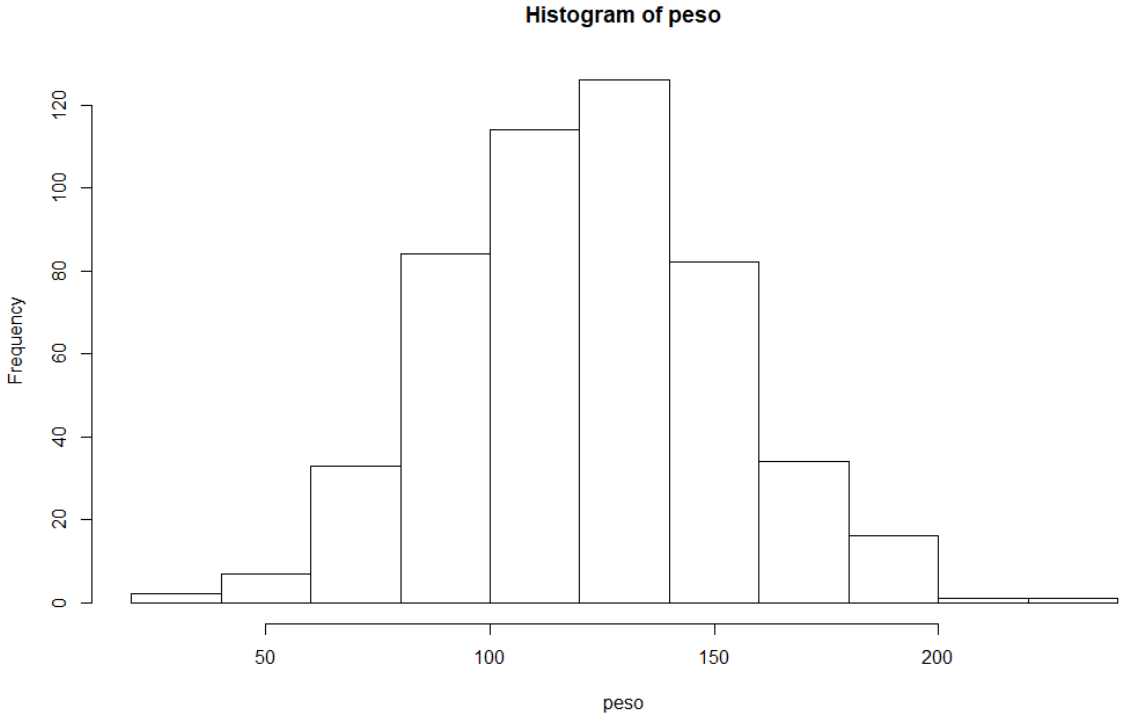
Voltando um pouco...



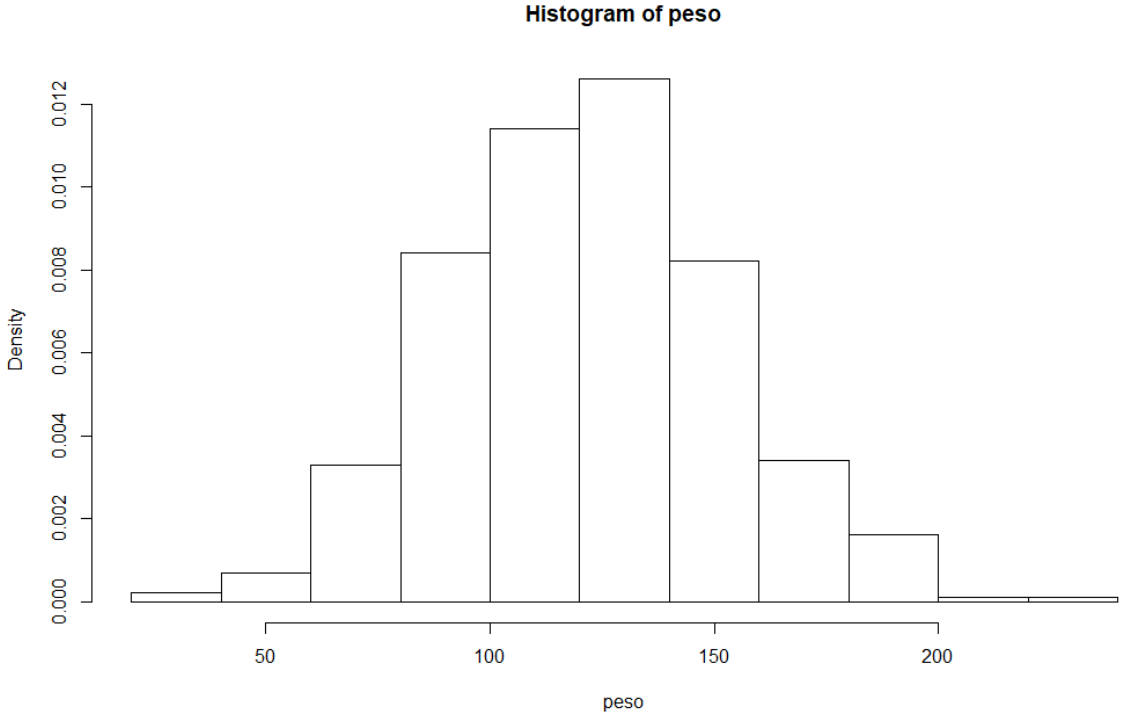
Distribuição de Probabilidade



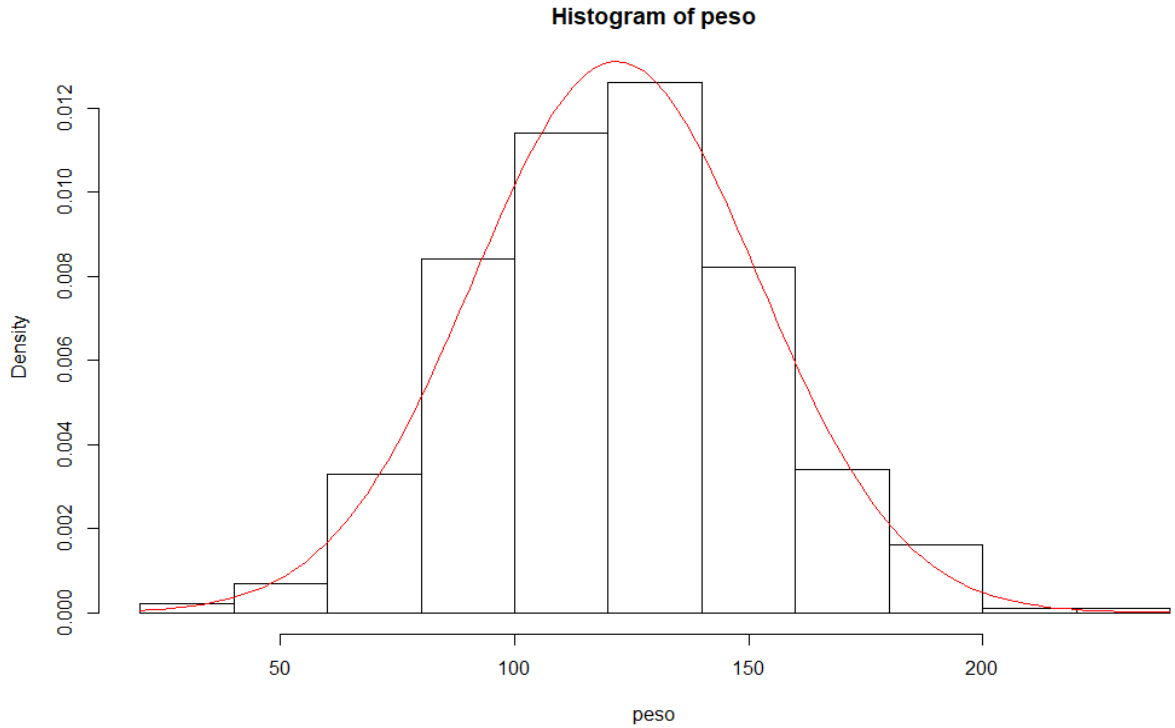
Variável peso: Frequência



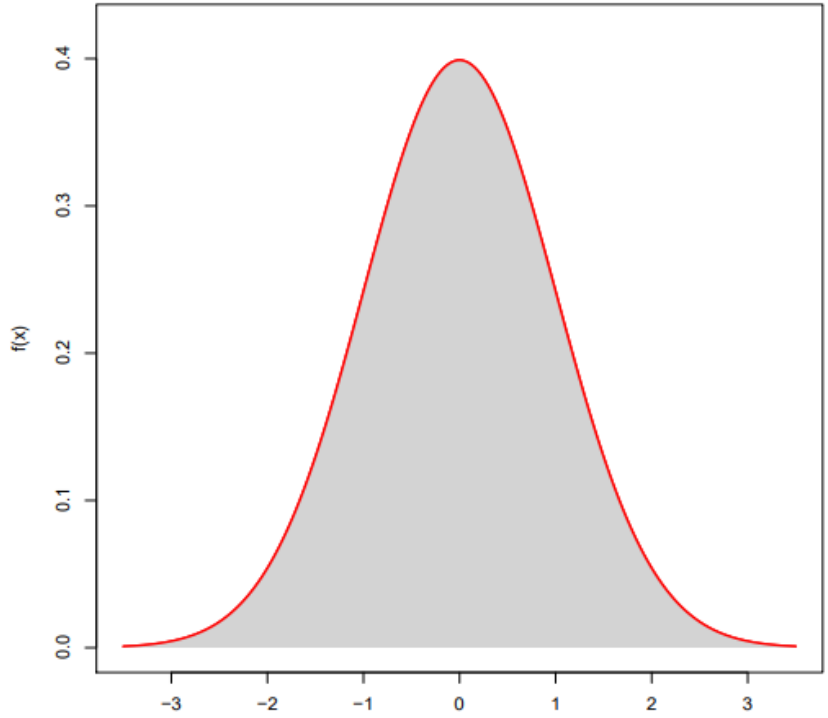
Variável peso: Probabilidade



Variável peso: Fitada por uma N(120,30)



Distribuição Normal



Distribuição Normal

Distribuição Normal

Se X é uma variável aleatória com distribuição normal de média μ e variância σ^2, a função densidade de probabilidade de X é definida por

Distribuição Normal

FIAP MBA+

Distribuição Normal

Se X é uma variável aleatória com distribuição normal de média μ e variância σ^2, a função densidade de probabilidade de X é definida por

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

para $-\infty < x, \mu < +\infty$ e $\sigma > 0$. Notação: $X \sim N(\mu, \sigma^2)$.

Distribuição Normal

FIAP MBA+

Padronização

Se $X \sim N(\mu, \sigma^2)$ e $Z \sim N(0, 1)$ (normal padrão), então

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right),$$

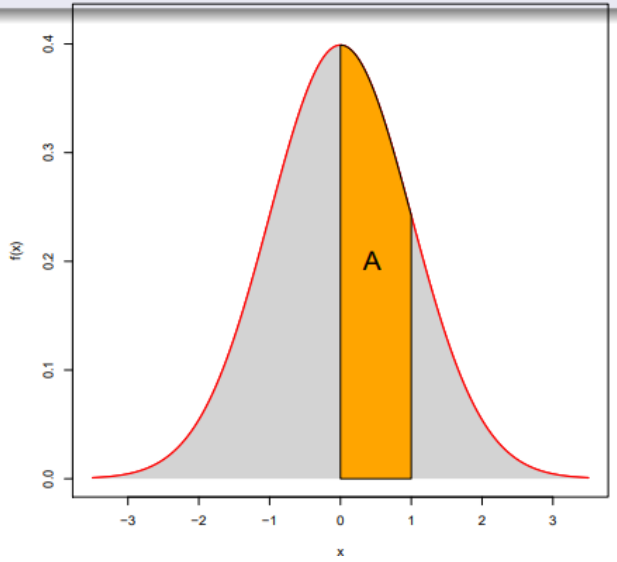
ou seja, todos os cálculos podem ser feitos pela normal padrão.

Distribuição Normal

Cálculo de probabilidades

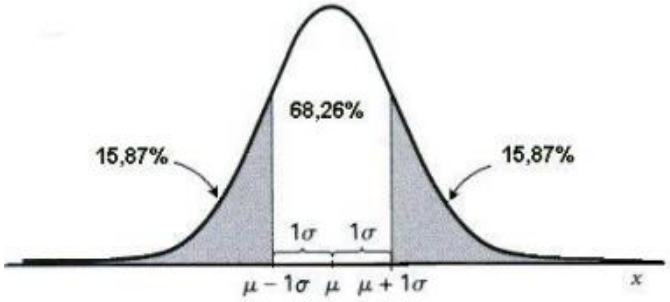
Por exemplo, a probabilidade $A = P(0 \leq X \leq 1)$ pode ser calculada pela diferença

$$P(X \leq 1) - P(X \leq 0) = 0,841 - 0,5 = 0,341.$$

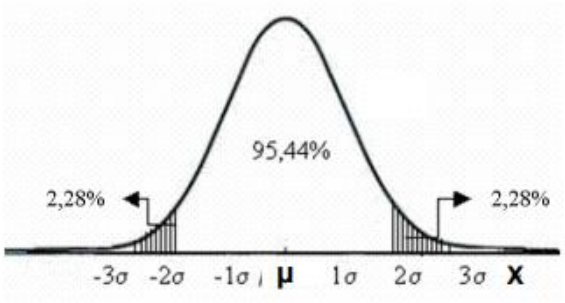


Distribuição Normal

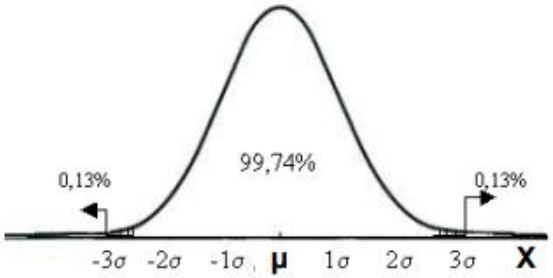
$$Z \sim N(0,1)$$



$$P[(\mu-\sigma) < X < (\mu+\sigma)] = 68.25\%$$



$$P[(\mu-2\sigma) < X < (\mu+2\sigma)] = 95.44\%$$

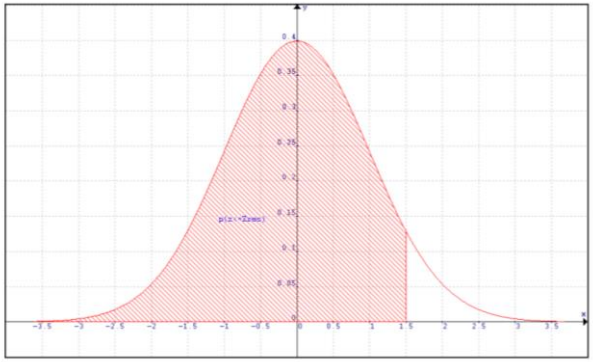


$$P[(\mu-3\sigma) < X < (\mu+3\sigma)] = 99.74\%$$

Tabela da distribuição Normal

FIAP MBA+

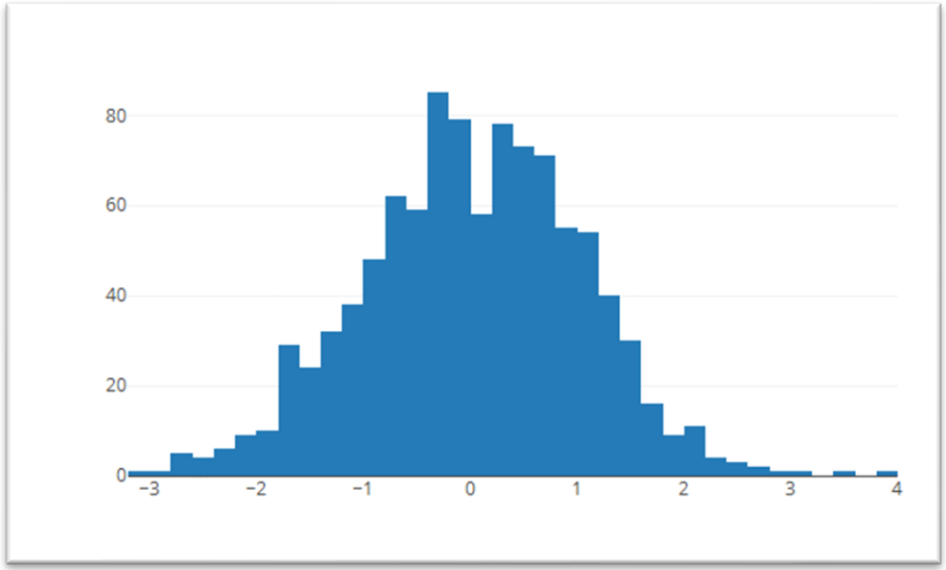
P(Z<1.38)=?



Zres até a 1ª casa decimal	Segunda casa decimal de Zres									
	0	1	2	3	4	5	6	7	8	9
0,0	**0,5000**	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
...	...	...	...	...	...	...	...	...	...	...
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	**0,9162**	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
...	...	...	...	...	...	...	...	...	...	...

Gerando números aleatórios

- Família Normal



Trabalhando no Python

Distribuição Normal para o cálculo de probabilidade ***P(X≤x)***

norm.cdf(x, m, s)

Distribuição Normal para o cálculo de probabilidade ***P(X>x)***

norm.sf(x, m, s)

Cálculo inverso: Informa o valor de x a partir de uma probabilidade ***acumulada***

norm.ppf(p, m, s)

onde:

m= média
s= desvio padrão
p = representa a probabilidade acumulada até x

Trabalhando no Python

Distribuição Normal para o cálculo de probabilidade ***P(X≤x)***

norm.cdf(x, m, s)

Distribuição Normal para o cálculo de probabilidade ***P(X>x)***

norm.sf(x, m, s)

Cálculo inverso: Informa o valor de x a partir de uma probabilidade ***acumulada***

norm.ppf(p, m, s)

Importante:

Para usar as funções de cálculo de probabilidade para a distribuição normal no Python é necessário primeiramente que você importe a função norm:

from scipy.stats import norm

Exemplo 1

Suponha que as medidas da corrente em um pedaço de fio sigam a distribuição normal, com um média de 10 miliamperes e uma variância de 5 miliamperes. Qual a probabilidade:

Como temos uma variável (X: medida da corrente em um pedaço de fio) com distribuição normal com μ=10 e σ^2=5, é necessário padroniza-la para poder consultar as probabilidades disponíveis na tabela da distribuição normal padrão. A padronização de uma variável X ~ N (μ, σ^2) em uma variável Z ~ N (0, 1) é realizada efetuando o seguinte cálculo:

$$z = \frac{x - \mu}{\sigma}$$

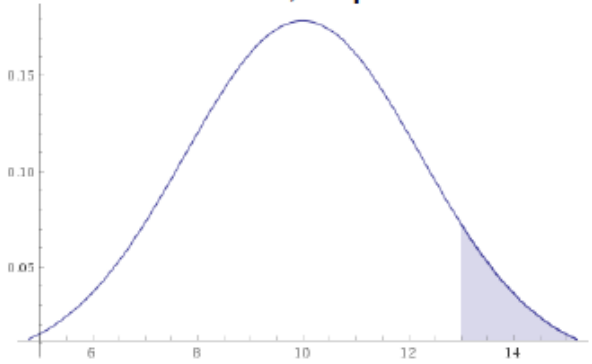
Exemplo 1

a) Da medida da corrente ser de no máximo 12 miliamperes.
Graficamente, a probabilidade desejada pode ser representada da seguinte maneira:

$$P(X \leq 12) = P\left(Z \leq \frac{12-10}{\sqrt{5}}\right) = P(Z \leq 0{,}89) = 0{,}5 + 0{,}3133 = 0{,}8133$$

Exemplo 1

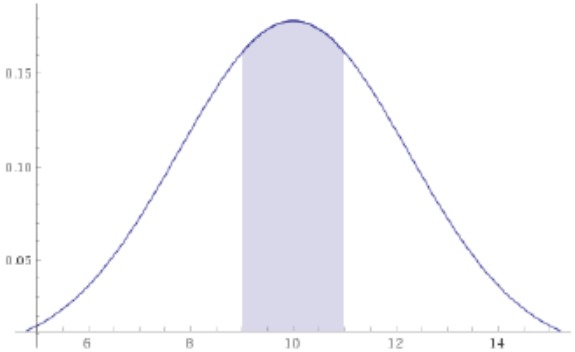
b) Da medida da corrente ser de pelo menos 13 miliamperes.
Graficamente, a probabilidade desejada pode ser representada da seguinte maneira:



$$P(X \geq 13) = P\left(Z \geq \frac{13-10}{\sqrt{5}}\right) = P(Z \geq 1{,}34) = 0{,}5 - 0{,}4099 = 0{,}0901$$

Exemplo 1

c) Um valor entre 9 e 11 miliamperes.
Graficamente, a probabilidade desejada pode ser representada da seguinte maneira:



$$P(9 < X < 11) = P\left(\frac{9-10}{\sqrt{5}} < Z < \frac{11-10}{\sqrt{5}}\right) = P(-0{,}45 < Z < +0{,}45)$$
$$= 0{,}1736 + 0{,}1736 = 0{,}3472$$

Exemplo 1

d) Maior do que 8 miliamperes.
Graficamente, a probabilidade desejada pode ser representada da seguinte maneira:

$$P(X > 8) = P\left(Z > \frac{8-10}{\sqrt{5}}\right) = P(Z > -0{,}89) = 0{,}5 + 0{,}3133 = 0{,}8133$$

Exercícios

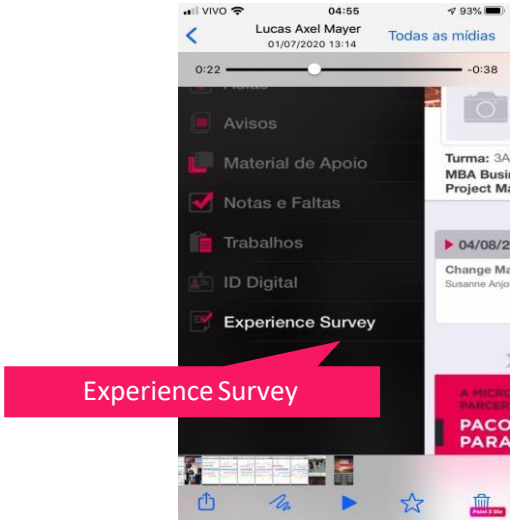
Considere que a pontuação obtida por diferentes candidatos em um concurso público segue uma distribuição aproximadamente normal, com média igual a 140 pontos e desvio padrão igual a 20 pontos. Suponha que um candidato é escolhido ao acaso. Calcule as probabilidades a seguir:

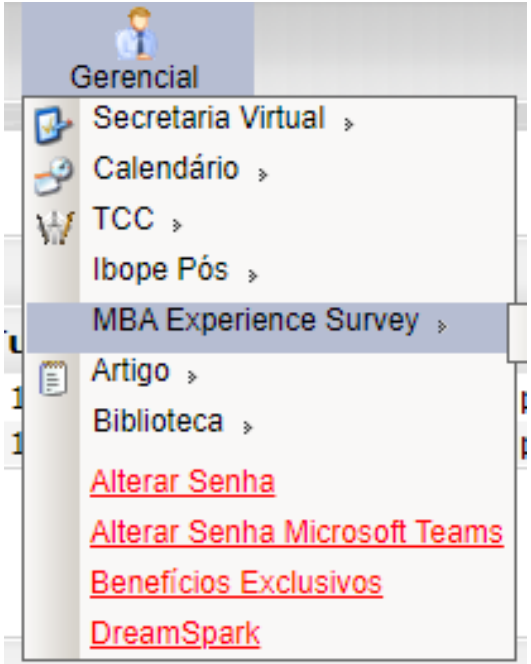
a) Apresentar uma pontuação entre 140 e 165,6.
b) Apresentar uma pontuação entre 127,4 e 140.
c) Apresentar uma pontuação entre 117,2 e 157.
d) Apresentar uma pontuação inferior a 127.
e) Apresentar uma pontuação superior a 174,2.
f) Apresentar uma pontuação inferior a 167,4.
g) Apresentar uma pontuação entre 155,4 e 168,4.

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

in/lafphd

profleandro.ferreira@fiap.com.br

FIAP MBA+


FIAP