FIAP

Statistics for Machine Learning

Aula 17: Distribuições Discretas

Prof. Jones Egydio

profjones.egydio@fiap.com.br

2024

Objetivos

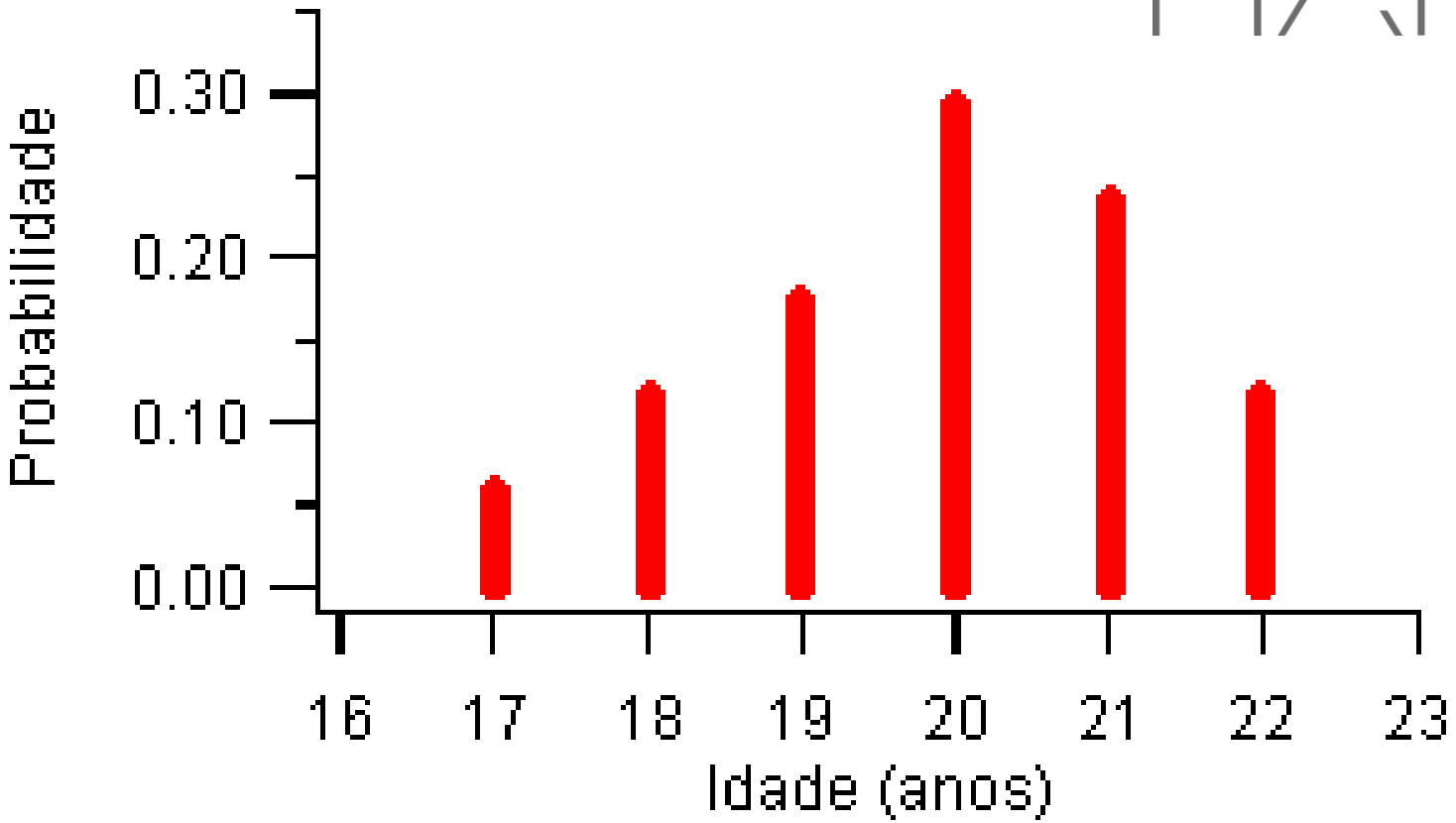
- Introduzir os conceitos de Distribuições Discretas:
  - Uniforme;
  - Bernoulli;
  - Binomial;
  - Poisson;
- Formas de representação;
- Exemplos e exercícios;
- Conclusão;
- Perguntas.

Qual a probabilidade de ocorrer algum acidente dado um histórico de acidentes?

- Anualmente ocorrem cerca de 270 milhões de acidentes e trabalho e 160 milhões de casos de doenças ocupacionais são registradas no mundo.
- + 2 milhões de pessoas perdem a vida todos os anos durante o trabalho (6,3 mil mortes por dia).
- Os custos globais chegam a 2,8 trilhões de dólares (podem comprometer até 4% do PIB mundial).

Fonte: https://www.seebcgms.org.br/noticia/acidentes-de-trabalho-matam-mais-do-que-guerras/

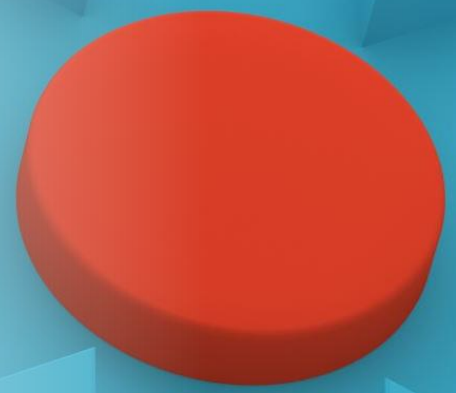
Conceitos iniciais



Distribuições de Probabilidade Discretas

Conceitos

- Variável
- Variável aleatória
- Distribuições de probabilidade
  - Distribuição Uniforme
  - Distribuição Bernoulli
  - Distribuição Binomial
  - Distribuição de Poisson



Variável

- Variável é uma característica da unidade amostral

Por exemplo, a unidade amostral seria este banco de dados.

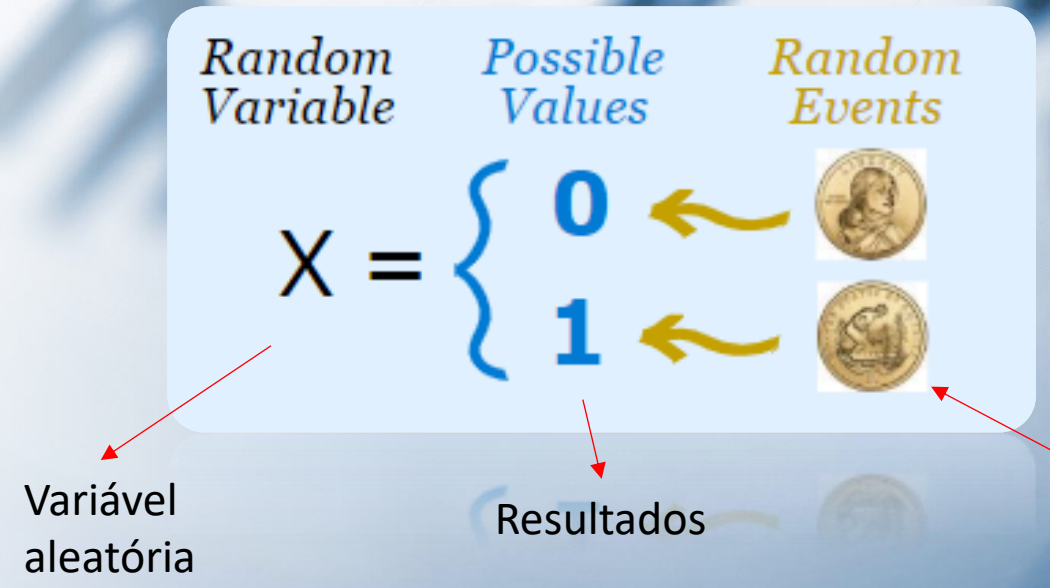
O PIB é uma característica desta unidade amostral, ou seja, uma variável (*feature*).

pais	ano	idh	corrupcao_indice	competitividade_indice	globalizacao_indice	pib	populacao
África do Sul	1997	0,6328	56	44,54	52,60298	1,49E+08	43353,632
África do Sul	1998	0,6272	57	31,11	54,51911	1,34E+08	43961,924
África do Sul	1999	0,6216	50	43,9	61,04379	1,33E+08	44526,272
África do Sul	2000	0,616	52	51,52	62,47182	1,33E+08	45064,098

Variável aleatória

Variáveis aleatórias representam resultados, em números, de processos aleatórios.

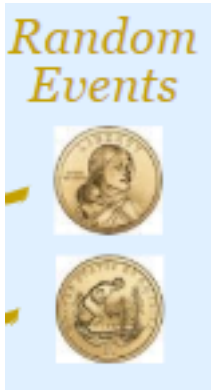
Por exemplo, jogar uma moeda, prever o resultado de um jogo, tentar cobrar alguém, alugar um carro e não ter acidente.



Tipos de variáveis aleatórias

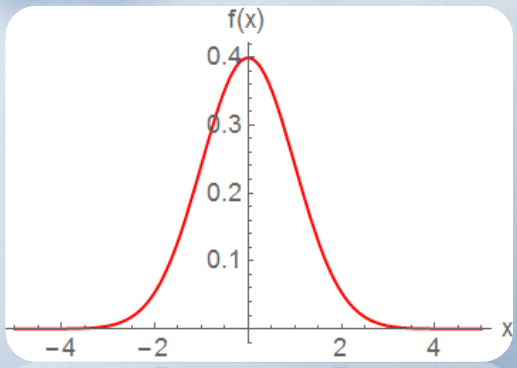
Discretas:

- 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

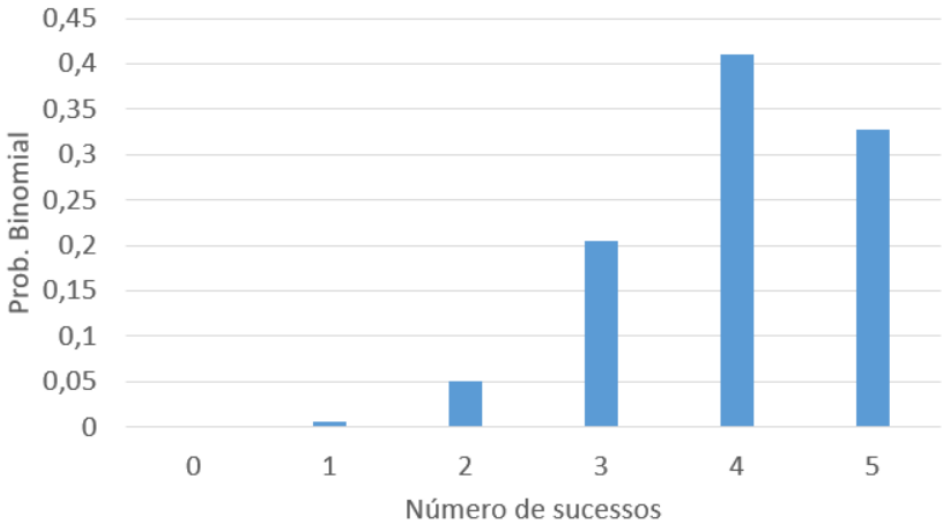
Número de sucessos	Combinações possíveis	Probabilidade de Bernoulli	Probabilidade Binomial
0	1	$0{,}2^5 = 0{,}00032$	$1 \cdot 0{,}2^5 = 0{,}00032$
1	5	$0{,}8^1 \cdot 0{,}2^4 = 0{,}00128$	$5 \cdot 0{,}8^1 \cdot 0{,}2^4 = 0{,}0064$
2	10	$0{,}8^2 \cdot 0{,}2^3 = 0{,}00512$	$10 \cdot 0{,}8^2 \cdot 0{,}2^3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

Distribuição de probabilidade

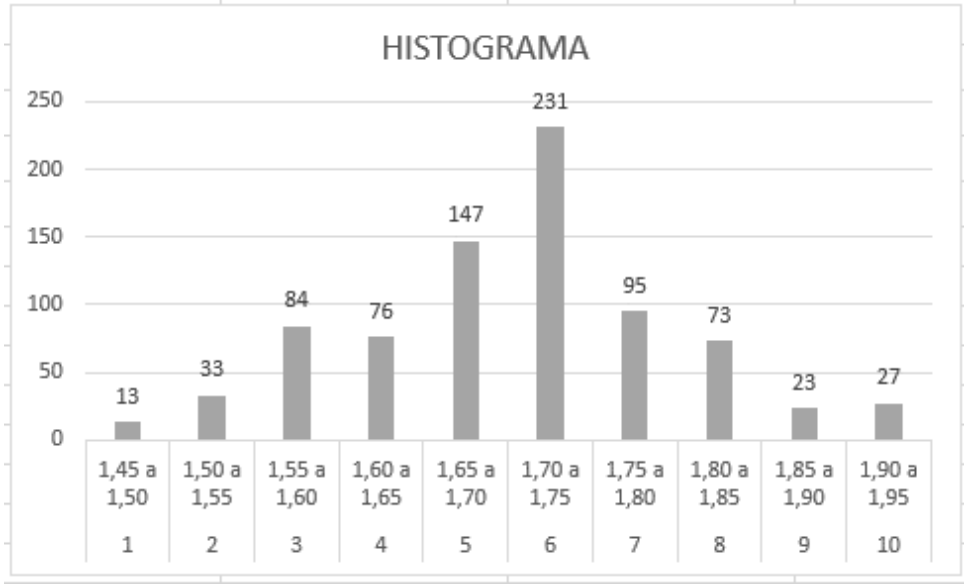
Distribuição de Probabilidades é um conjunto de resultados de eventos associados as suas respectivas probabilidades



Histograma

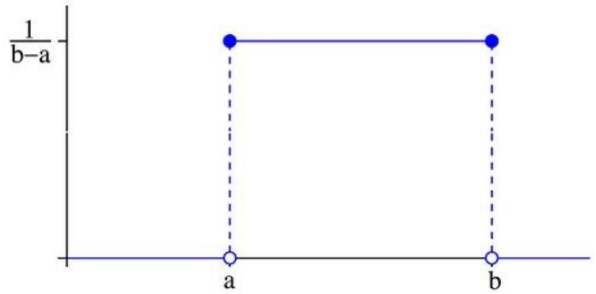
Gráficos deste tipo possuem um nome específico denominados de **Histogramas de frequência.**

Histogramas são gráficos de distribuições de probabilidades.



Distribuições de probabilidades discretas

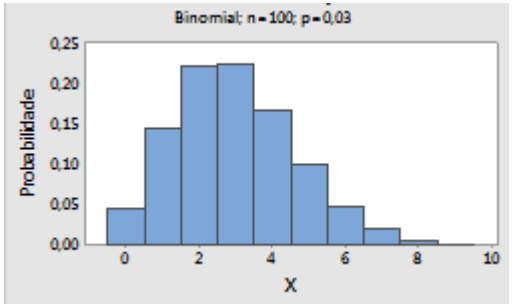
- Distribuições que serão apresentadas:



Distribuição uniforme

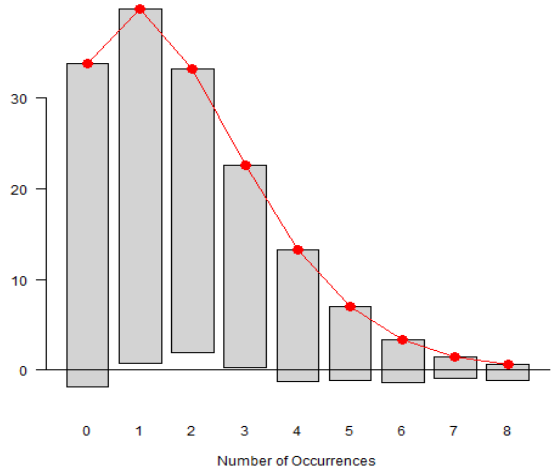
- $X \rightarrow x_1 = 1$ sucesso; $P(X=1) = p$
- $X \rightarrow x_1 = 0$ fracasso; $P(X=0) = 1 - p = q$

Distribuição bernoulli



Distribuição binomial

$$P(X = k) = \binom{n}{k} . p^k . (1-p)^{n-k}$$



Distribuição de Poisson

$$P(X = k) = \frac{e^{-\lambda} . \lambda^k}{k!}$$

Distribuição de probabilidade discreta

Uma distribuição discreta busca descrever a probabilidade associada as eventos de variáveis aleatórias discretas.

Por exemplo:

- Variável aleatória: **sucesso/fracasso.**
- Distribuição de probabilidade: **sucesso – 60%, fracasso – 40%.**

Distribuições Uniforme

- Imagine este dado de 6 lados:

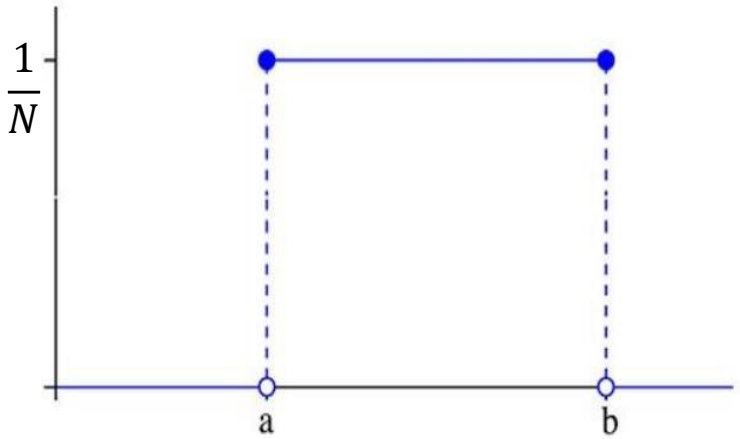
- Qual a chance de sair 1?
- Qual a chance de sair 6?

- Na distribuição uniforme todos os valores possuem a mesma probabilidade de ocorrer.

X	1	2	3	4	5	6
p(X)	1/6	1/6	1/6	1/6	1/6	1/6

Distribuições Uniforme Histograma

A distribuição tem o seguinte desenho*:



$$P(X) = \begin{cases} \frac{1}{N}, x = 1,2, \dots, N \\ 0, outras \end{cases}$$

$$média = E(X) = \frac{N+1}{2}$$

$$variância = V(X) = \frac{N^2 - 1}{12}$$

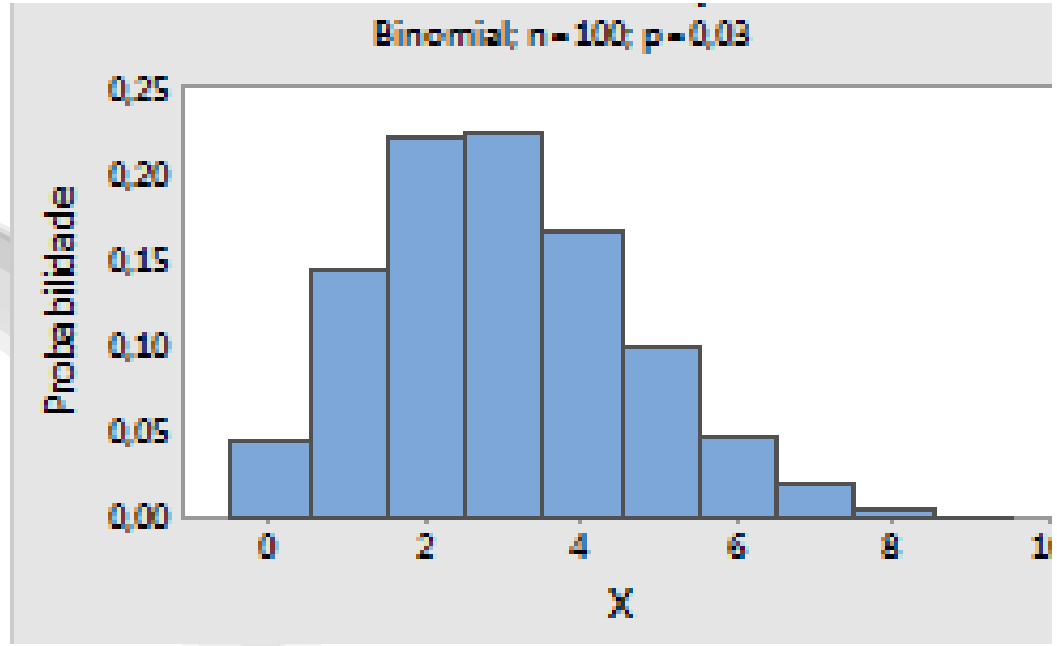
*Exemplo com números ver excel: distrib_prob.xlsx, aba dist_uniforme

Distribuição Bernoulli

- A distribuição de Bernoulli* é observada para somente a probabilidade de **um evento** em que somente 2 resultados podem ocorrer.
- Sucesso/Fracasso, Acidente/Não Acidente, Grávida/Não grávida
- **Exemplo**: Um funcionário sofrerá um acidente hoje? Teremos “0” para “não sofre acidente” com probabilidade “*p*” e “1” para “sofre acidente”, com probabilidade “1-*p*”.

*Exemplo com números ver excel: distrib_prob.xlsx, aba dist_bernoulli

Distribuição Binomial



- Imagine agora que buscamos observar um evento em que somente 2 resultados podem ocorrer (Bernoulli), porém este evento é repetido várias vezes.

- Neste caso, temos uma **Distribuição Binomial**.

Distribuição Binomial (Exemplos)

- Uma empresa envia *sms* de propaganda para 10 mil clientes, a chance de 1 cliente acessar o link é de 5%. Qual a probabilidade de 100 clientes acessarem o link dos *sms`s*?
- Uma empresa envia *email marketing* para 50 mil clientes, a chance de 1 cliente acessar o link é de 2%. Qual a probabilidade de 500 clientes acessarem o link dos *email`s*?
- A chance de acidente mensal em uma fábrica com 200 funcionários é de 3%, qual a probabilidade de 10 funcionários sofrerem acidente este mês?

Distribuição Binomial (resumo)

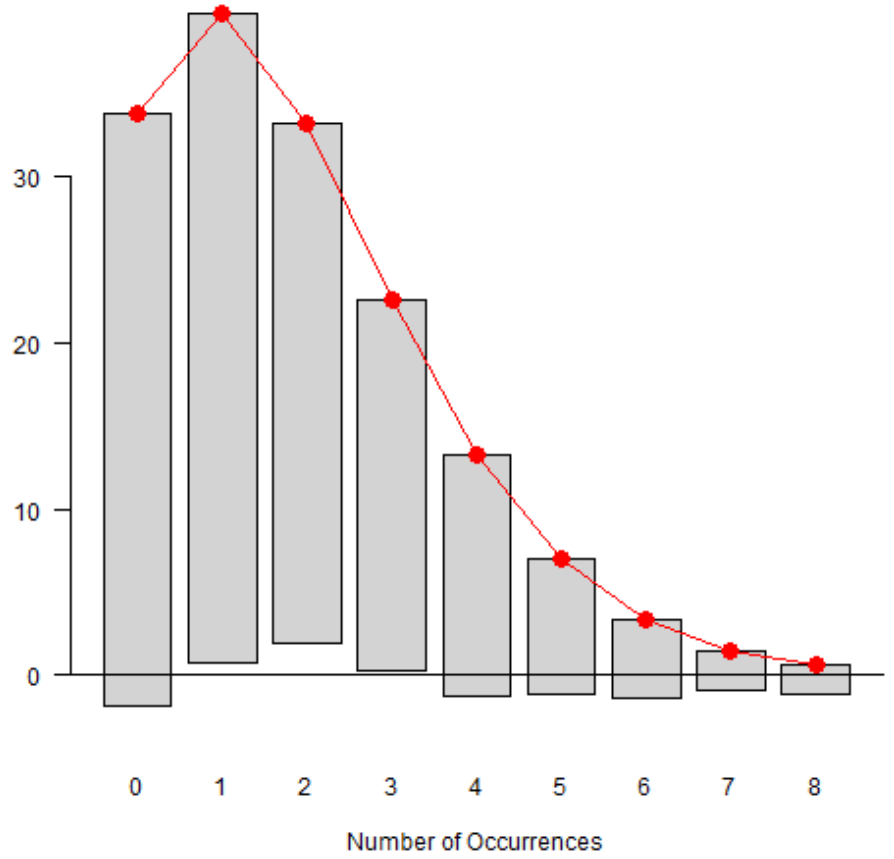
Em resumo, uma distribuição binomial possui 3 características e a seguinte fórmula:

a) consiste de *n* eventos de Bernoulli (sim/não, sucesso/fracasso)
b) cada evento é independente
c) a probabilidade de cada evento ocorrer está entre 0 e 1

$$P(X) = \frac{n!}{(n-k)!\,k!} p^k \cdot (1-p)^{n-k}$$

Distribuição Poisson

- A distribuição de Poisson é utilizada quando buscamos contar o número de eventos num intervalo pré determinado (tempo, área, volume).



Distribuição Poisson (Exemplos)

- Número de chamadas telefônicas recebidas em um intervalo de cinco minutos
- Número de falhas de um sistema bancário em um dia de operação
- Número de acidentes ocorridos em um dia na cidade de São Paulo
- Número de defeitos para cada 100 metros de tecido
- Número de carros que estacionam a cada hora.

Distribuição de Poisson (resumo)

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

a) consiste de *n* eventos que podem ser contados
b) cada evento é independente
c) deve-se obter a média de eventos dentro de um intervalo pré-determinado

$$P(X) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

λ é a média de eventos dento de um intervalo

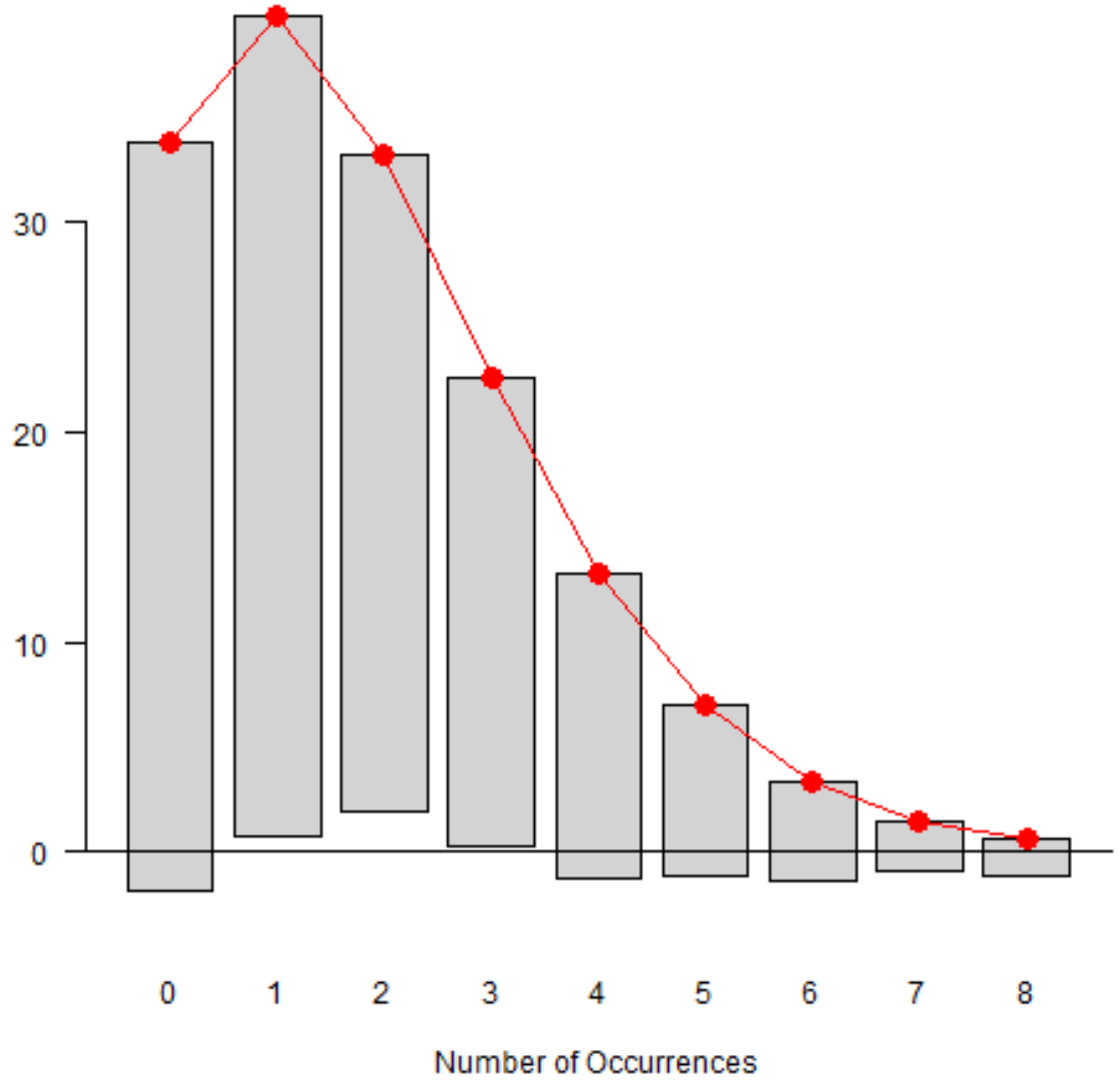
Voltando ao nosso problema

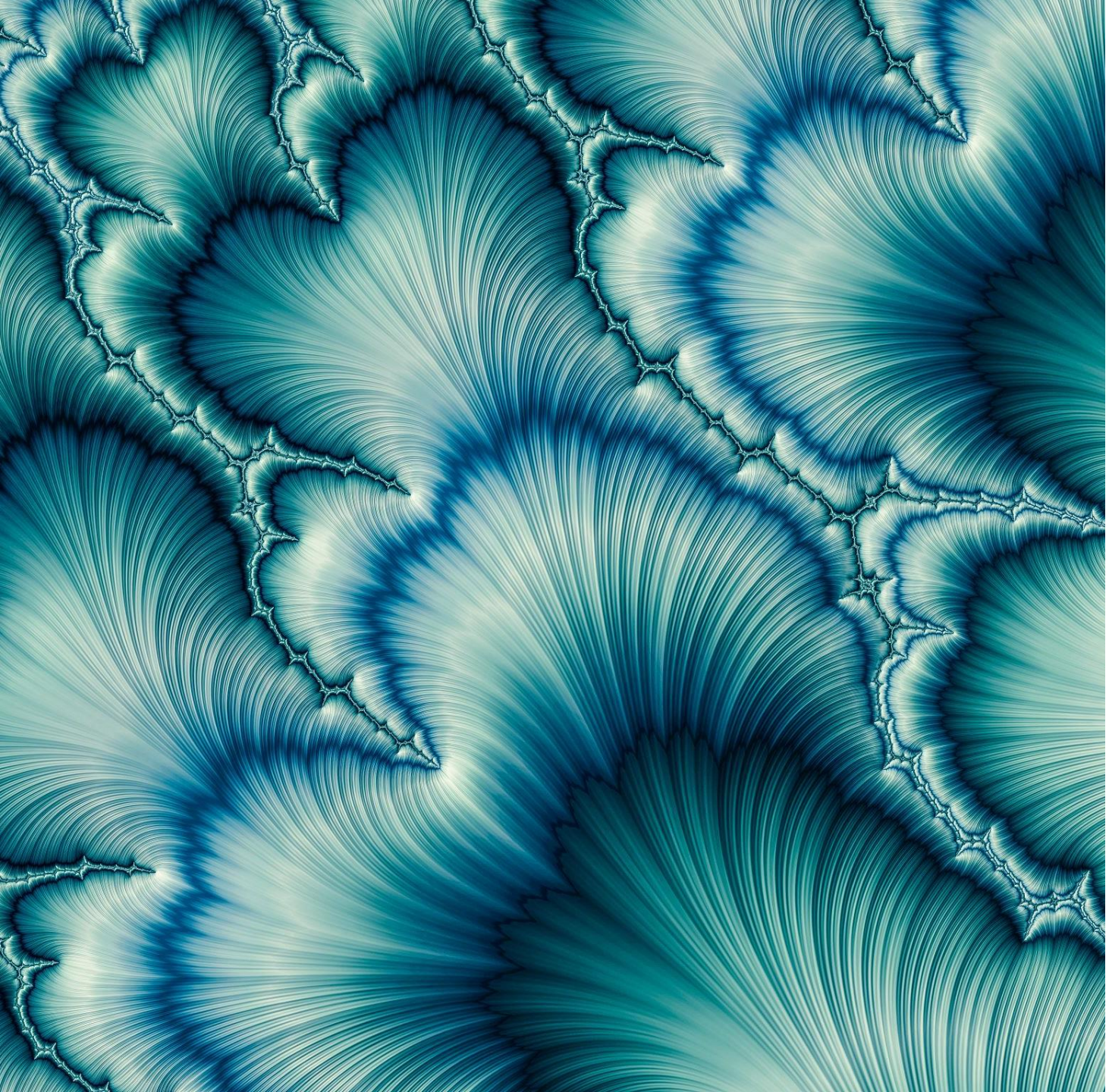
- **Qual a probabilidade de ocorrer algum acidente dado um histórico de acidentes?**
- Qual a distribuição de probabilidade você assumiria para resolver este problema?
- Se uma empresa tem média mensal de 20 acidentes. Qual a probabilidade de ocorrer 7 acidentes?
- Qual a probabilidade de ocorrer entre 15 a 17 acidentes? Mesma média.

Resolução 1

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

- R. **Poisson**



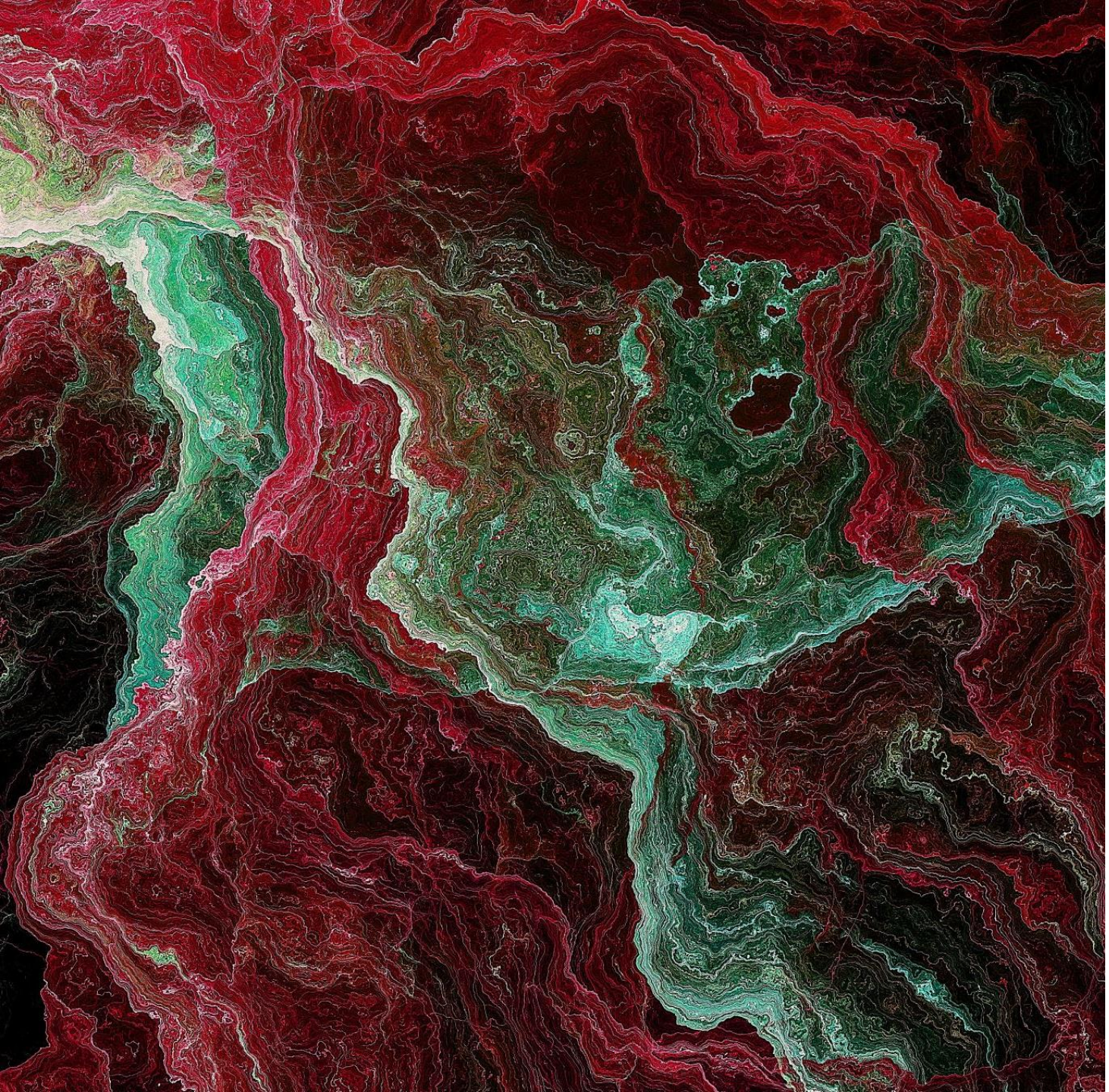


Resolução 2

Se uma empresa tem média mensal de 20 acidentes. Qual a probabilidade de ocorrer 7 acidentes?

- 20 é a média de eventos dento de um interval

$$P(X) = \frac{e^{-20}20^{7}}{7!} = 0{,}000523 = 0{,}052\%$$



FIAP

Resolução 3

- **Qual a probabilidade de ocorrer entre 15 a 17 acidentes? (mesma média, 20 acidentes)**

- 20 é a média de eventos dento de um interval

$$P(X) = \frac{e^{-15}20^{15}}{15!} + \frac{e^{-16}20^{16}}{16!} + \frac{e^{-17}20^{17}}{17!} =$$

$$= 0{,}1922 = 19{,}22\%$$

Exercícios

Você é o gerente de uma loja e sabe que, fora do horário de pico, entram, em média, 6 clientes a cada 10 minutos. Qual a probabilidade de entrarem:
a) 6 clientes na loja em um período qualquer de 10 minutos fora do horário de pico?
b) até 2 clientes num período de 10 minutos fora do horário de pico?
c) entrarem 3 clientes ou mais fora do horário de pico ao longo de 10 minutos?

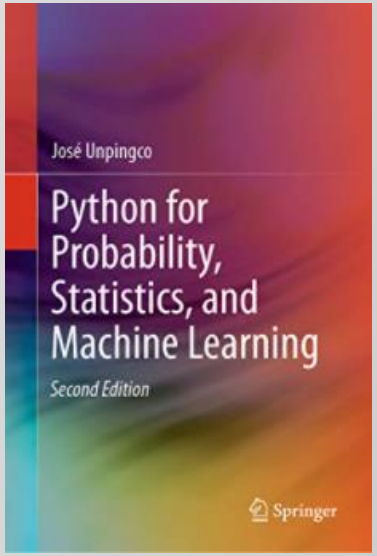
Resposta: https://www.youtube.com/watch?v=6uOLbhYeXrk

Uma prova consta de 10 testes com 5 alternativas cada um, sendo apenas uma delas correta. Um aluno que nada sabe a respeito da matéria avaliada, "chuta" uma resposta para cada teste. Qual é a probabilidade dele acertar exatamente 6 testes?

Resposta: https://www.youtube.com/watch?v=RKKF4LLpT9M

Onde estudar mais!!

Leitura



Distribuições discretas:

https://pt.khanacademy.org/math/ap-statistics/random-variables-ap/discrete-random-variables/v/discrete-probability-distribution

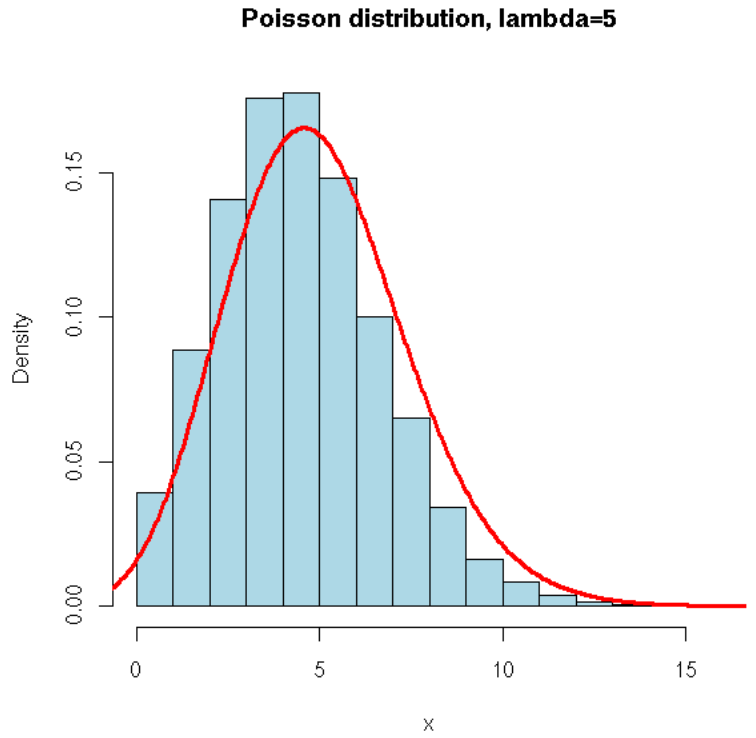
Vídeos

- Distribuição Binomial: https://pt.khanacademy.org/math/statistics-probability/random-variables-stats-library/binomial-random-variables/v/binomial-distribution
- Distribuição de Poisson: https://pt.khanacademy.org/math/statistics-probability/random-variables-stats-library/poisson-distribution/v/poisson-process-1

Parte 2: prática

Conceitos iniciais

- Distribuições de Probabilidade Discretas (Aplicações no Python)



Histograma

- Histograma com matplotlib:
- Preparação dos dados

```
dados['data_acidente'] = pd.to_datetime(dados['data_acidente'], errors='coerce')
```

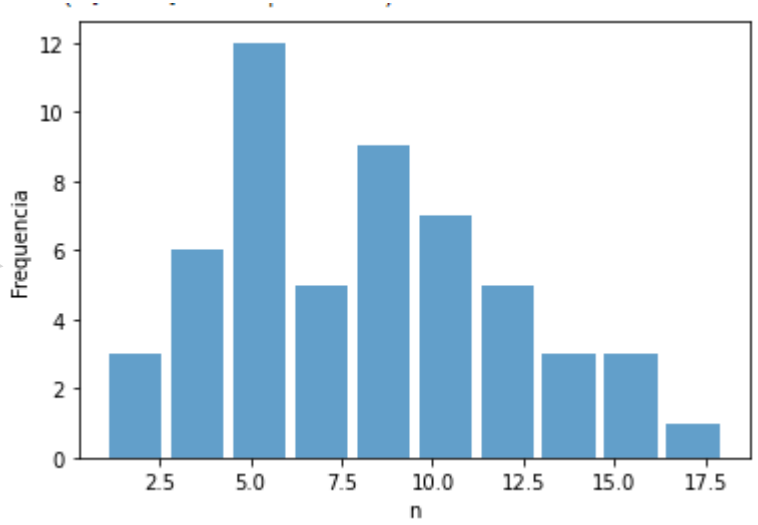
Ajustamos a coluna data_acidente para ano-mês.

```
dados['data_acidente_mensal'] = dados['data_acidente'].dt.strftime('%Y-%m')
```

```
dados_acidentes_mensal = dados.groupby('data_acidente_mensal') \
                              .size() \
                              .to_frame('n') \
                              .reset_index()
```

Esta linha de código foi adicionada pq buscamos obter o número de acidentes por mês.

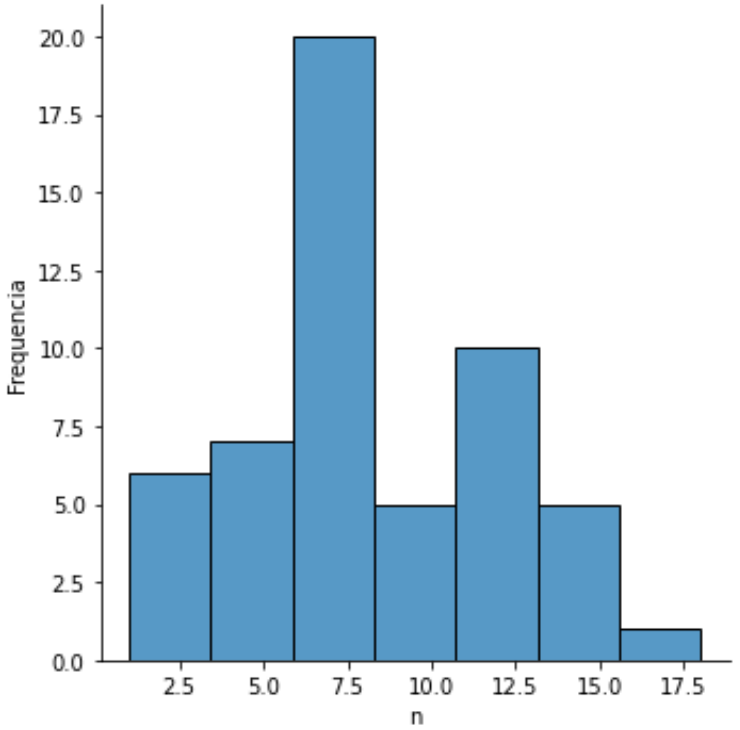
```
plt.hist(dados_acidentes_mensal['n'], alpha=0.7, rwidth=0.85)
plt.xlabel('n')
plt.ylabel('Frequencia')
```



Histograma

- Histograma com seaborn:

```
sns.displot(dados_acidentes_mensal['n'])
plt.ylabel('Frequencia')
```



Distribuições Uniforme Histograma

- Exemplo distribuição uniforme:

```
n = 6
p = 1/n
```

O dado tem 6 lados e os lados têm mesma chance de sair.

```
resultados = np.arange(1, 7)
```

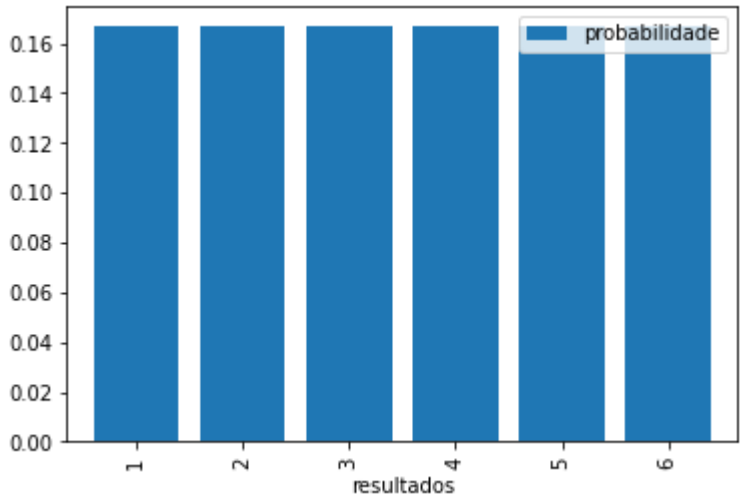
```
probabilidade = np.repeat(p, n)
```

Criamos um data frame mantendo colunas de resultados e probabilidade.

```
dados_uniforme = pd.DataFrame(resultados, columns=['resultados'])
dados_uniforme['probabilidade'] = probabilidade
```

```
dados_uniforme.plot(kind = 'bar',
                    x = 'resultados',
                    y = 'probabilidade', width = 0.8)
```

Geramos o gráfico.



Distribuição Bernoulli

- Exemplo Bernoulli

```
p = 0.2
```

```
resultados = np.arange(0, 2)
```

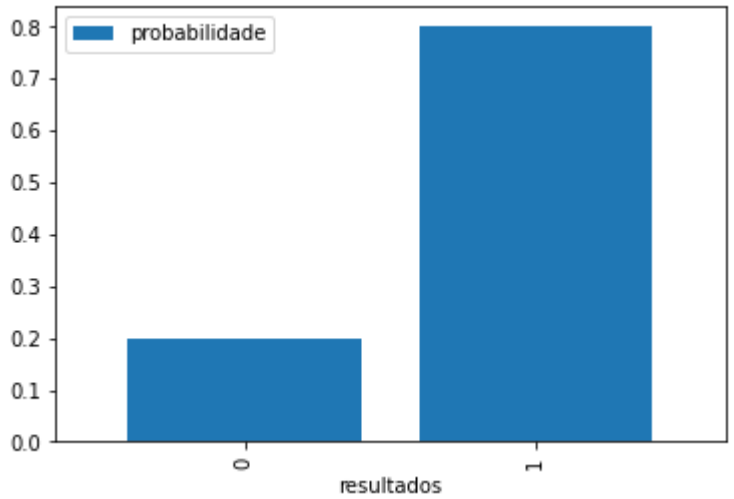
```
probabilidade = np.array([p, 1-p])
```

Criamos um data frame mantendo colunas de resultados e probabilidade.

```
dados_bernoulli = pd.DataFrame(resultados, columns=['resultados'])
dados_bernoulli['probabilidade'] = probabilidade
```

```
dados_bernoulli.plot(kind = 'bar',
                     x = 'resultados',
                     y = 'probabilidade', width = 0.8)
```

Geramos o gráfico.



Distribuição Binomial

- Exemplo Binomial
- Entender a função do Python que calcula as probabilidades da distribuição binomial:

- $P(X) = \frac{n!}{(n-k)!k!} p^k \cdot (1-p)^{n-k}$

```
binom.pmf(k, n, p)
```

Distribuição Binomial

- Exemplo Binomial:

```
n = 5
p = 0.8
```

```
n_sucessos = [i for i in range(n+1)]
```

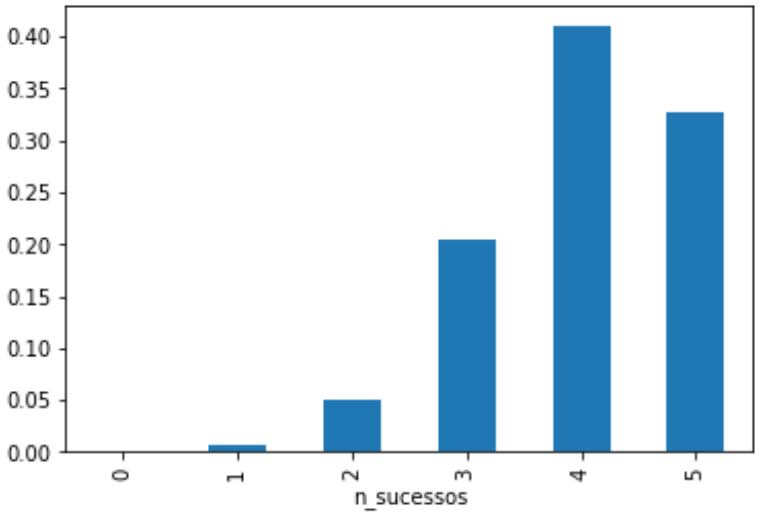
```
dados_binomial = pd.DataFrame(n_sucessos, columns=['n_sucessos'])
```

Criamos um data frame mantendo colunas de resultados e probabilidade.

```
dados_binomial['probs'] = dados_binomial['n_sucessos'].apply(lambda x: binom.pmf(k=x, n = n, p = p))
```

```
dados_binomial.plot(kind='bar',
                    x = 'n_sucessos',
                    y = 'probs',
                    legend = None)
```

Geramos o gráfico.



Distribuição de Poisson

- Exemplo Poisson
- Entender a função do Python que calcula as probabilidades da distribuição poisson:

- $P(X) = \frac{e^{-\mu}\mu^{k}}{k!}$

```
poisson.pmf(k, mu)
```

Distribuição Poisson

- Exemplo Poisson:

```
n = 15
mu = 2
```

```
resultados = [i for i in range(n+1)]
```

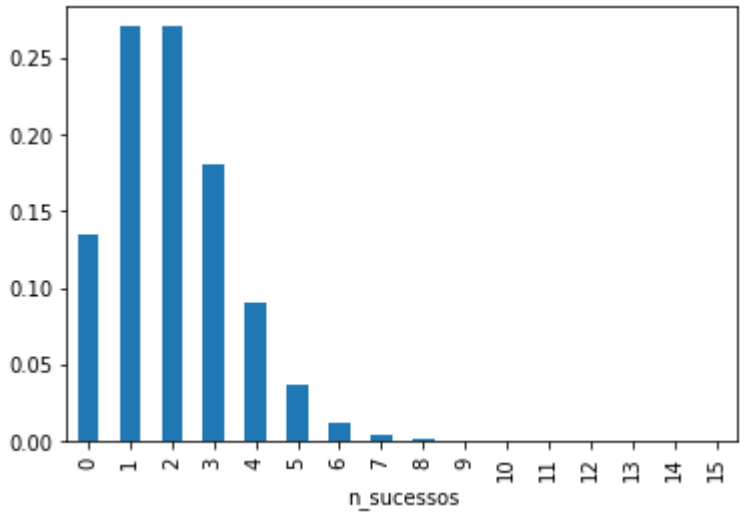
```
dados_poisson = pd.DataFrame(resultados, columns=['n_sucessos'])
```

Criamos um data frame mantendo colunas de resultados e probabilidade.

```
dados_poisson['probs'] = dados_poisson['n_sucessos'].apply(lambda x: poisson.pmf(k=x, mu=mu))
```

```
dados_poisson.plot(kind='bar',
                   x = 'n_sucessos',
                   y = 'probs',
                   legend = None)
```

Geramos o gráfico.



Problema

- Anualmente ocorrem cerca de 270 milhões de acidentes e trabalho e 160 milhões de casos de doenças ocupacionais são registradas no mundo.
- + 2 milhões de pessoas perdem a vida todos os anos durante o trabalho (6,3 mil mortes por dia).
- Os custos globais chegam a 2,8 trilhões de dólares (podem comprometer até 4% do PIB mundial).

Qual a probabilidade de ocorrer algum acidente dado um histórico de acidentes?

Fonte: https://www.seebcgms.org.br/noticia/acidentes-de-trabalho-matam-mais-do-que-guerras/

Preparação dos dados

- Carregar pacotes e dados

```
import numpy as np
import pandas as pd
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import binom, poisson
```

Pacotes.

Dados.

```
dados = pd.read_csv('acidentes.csv')
```

Uso das funções do R

- **Qual a probabilidade de ocorrer algum acidente dado um histórico de acidentes?**

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

- Se uma empresa tem média mensal de 20 acidentes. Qual a probabilidade de ocorrer 7 acidentes?

- Qual a probabilidade de ocorrer entre 15 a 17 acidentes? Mesma média.

Resolução 2

- **Se uma empresa tem média mensal de 20 acidentes. Qual a probabilidade de ocorrer 7 acidentes?**

- Vamos utilizar a distribuição de poisson
- 20 é a média de eventos dento de um intervalo
- $P(X) = \frac{e^{-20}20^7}{7!} = 0{,}000523 = 0{,}052\%$

```
media_acidentes = 20
n_acidentes = 7
```

```
poisson.pmf(k=n_acidentes,
            mu=media_acidentes)
```

```
0.0005234675866510618
```

Resolução 3

- **Qual a probabilidade de ocorrer entre 15 a 17 acidentes? (mesma média, 20 acidentes)**

- 20 é a média de eventos dento de um intervalo
- $P(X) = \frac{e^{-15}20^{15}}{15!} + \frac{e^{-16}20^{16}}{16!} + \frac{e^{-17}20^{17}}{17!} = 0{,}1922 = 19{,}22\%$

```
media_acidentes = 20

n_acidentes = np.arange(15, 18)

np.sum(poisson.pmf(k=n_acidentes, mu=media_acidentes))
```

```
0.19216411681668827
```

Análise do problema com dados

- **Qual a probabilidade de ocorrer algum acidente dado um histórico de acidentes?**
- Em nosso conjunto de dados, qual foi a média mensal de acidentes?
- Utilize a média mensal obtida anteriormente. Qual a probabilidade de ocorrer 7 acidentes?
- Qual a probabilidade de ocorrer entre 4 a 6 acidentes? (Utilize a média mensal obtida anteriormente)

Análise do problema com dados

- Em nosso conjunto de dados, qual foi a média mensal de acidentes?

```
dados['data_acidente'] = pd.to_datetime(dados['data_acidente'], errors='coerce')
```

```
dados['data_acidente_mensal'] = dados['data_acidente'].dt.strftime('%Y-%m')
```

```
dados_acidentes_mensal = dados.groupby('data_acidente_mensal') \
                              .size() \
                              .to_frame('n') \
                              .reset_index()
```

Preparamos os dados para obter a série histórica de acidentes.

```
np.mean(dados_acidentes_mensal['n'])
```

```
8.166666666666666
```

Código para obter a média mensal.

Análise do problema com dados

- Utilize a média mensal obtida anteriormente. Qual a probabilidade de ocorrer 7 acidentes?

```
media_mensal = np.mean(dados_acidentes_mensal['n'])
```

```
poisson.pmf(k = 7, mu=media_mensal)
```

```
0.13650388195181395
```

Análise do problema com dados

- Qual a probabilidade de ocorrer entre 4 a 6 acidentes? (Utilize a média mensal obtida anteriormente)

```
n_acidentes = np.arange(4, 7)
```

```
media_mensal = np.mean(dados_acidentes_mensal['n'])
```

```
np.sum(poisson.pmf(k = n_acidentes, mu=media_mensal))
```

```
0.2555945240092576
```

Obrigado!