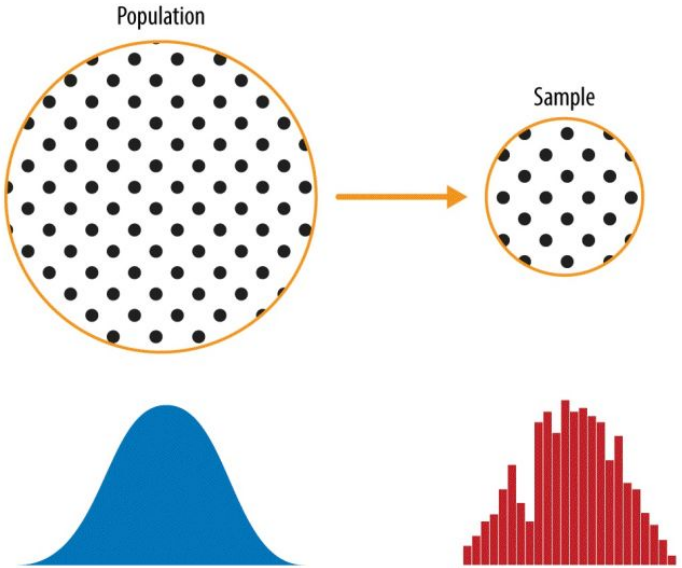
Métodos Quantitativos
Aula 04

Amostragem e Distribuições Estatísticas

Roberto Massi de Oliveira
Alex Borges Vieira

Revendo Conceito de Amostras

- Ex.: População = baralho completo / Amostra = 10 cartas do baralho



Amostragens Aleatórias ou Enviesadas

- Termos-chave para amostragem:
  - Amostra: subconjunto de uma população (média: $\bar{x}$, variância: s^2)
  - População: conjunto de dados (média: μ, variância: σ^2)
  - N(n): tamanho da população(tamanho da amostra)
  - Amostragem aleatória: elementos retirados aleatoriamente da população
  - Amostra estratificada: amostra aleatória extraída de população organizada em grupos
  - Amostra aleatória simples: amostra aleatória removida de população não estratificada
  - Amostra enviesada: amostra que representa erroneamente uma população

Viés de Seleção

Viés de Seleção

“The reviews of restaurants, hotels, cafes, and so on that you read on social media sites like Yelp are prone to bias because the people submitting them are not randomly selected; rather, they themselves have taken the initiative to write. This leads to self-selection bias — the people motivated to write reviews may be those who had poor experiences, may have an association with the establishment, or may simply be a different type of person from those who do not write reviews. Note that while self-selection samples can be unreliable indicators of the true state of affairs, they may be more reliable in simply comparing one establishment to a similar one; the same self-selection bias might apply to each.”

BRUCE, P.; BRUCE. A. Practical Statistics for Data Scientists. 1. ed. O’Reilly Media, 2017.

Viés de Seleção

https://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-busines

The Yelp Factor: Are Consumer Reviews Good for Business?

by Michael Blanding

Michael Luca shows just how much restaurant reviews on Yelp affect companies' bottom lines. The more difficult question: Are these ratings reliable as a measure of product quality?

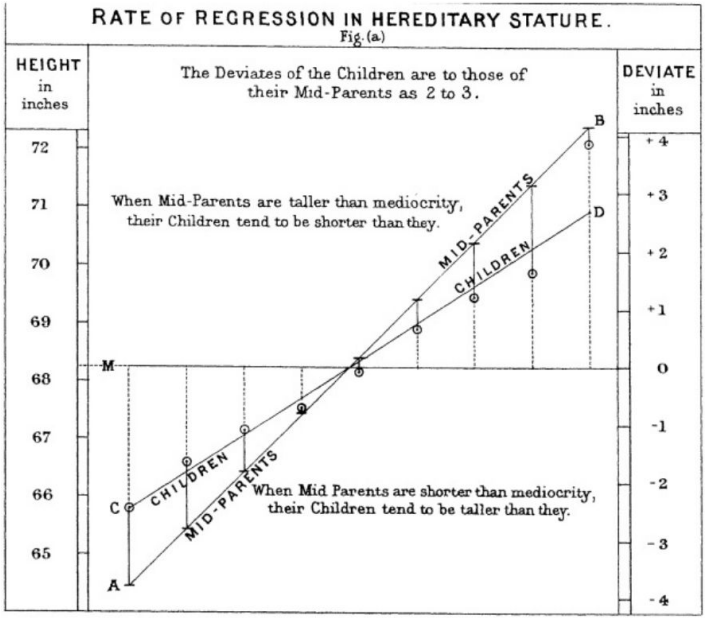
In recent years, consumer review sites including Yelp, Citysearch, and TripAdvisor have become the first stop for recommendations on everything from dinner to dentists. Along the way, they've earned a loyal following from fans, but also the ire of businesses that find themselves hurt by dyspeptic reviews.

"RESTAURANTS ARE A CLASSIC EXAMPLE IN ECONOMICS WHERE THE CONSUMER HAS TO MAKE A DECISION BASED ON VERY LITTLE INFORMATION"

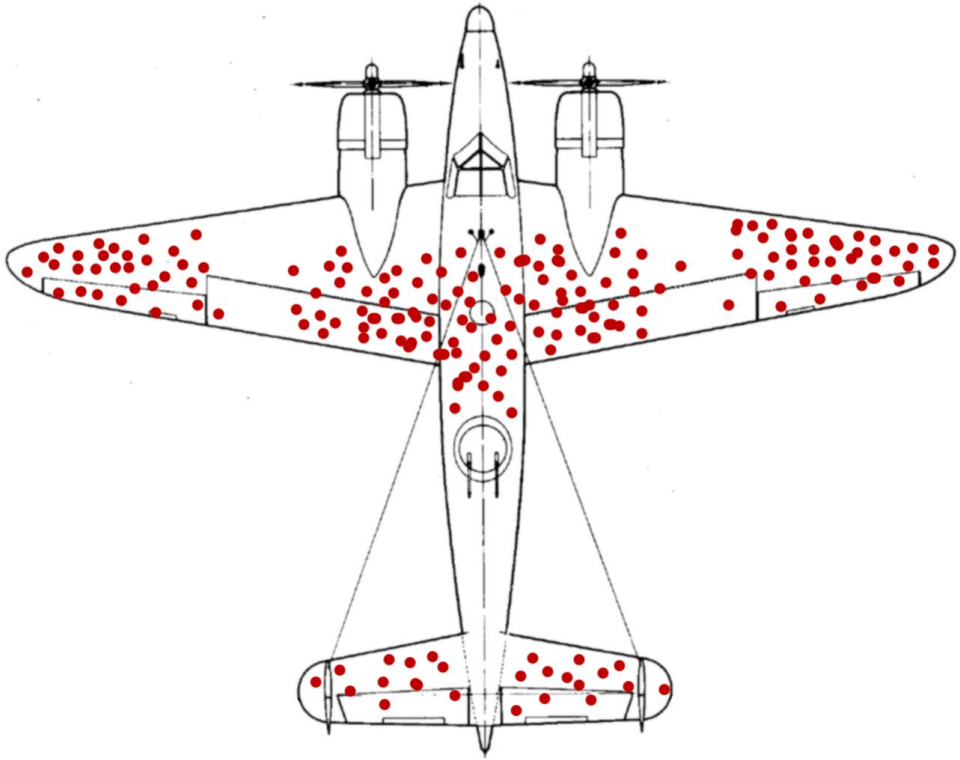
Most would agree these sites do influence consumers' decisions. In the paper Reviews, Reputation, and Revenue: The Case of Yelp.com, Harvard Business School Assistant Professor Michael Luca set to find out exactly by how much, and identify winners and losers in the process. "I have always been interested in how companies form their reputations, not only restaurants and hotels but also schools and doctors," says Luca.

Viés de Seleção

- Fenômeno da regressão para a média: quando outliers aparecem na primeira medição, valores mais próximos à média tendem a aparecer nas seguintes

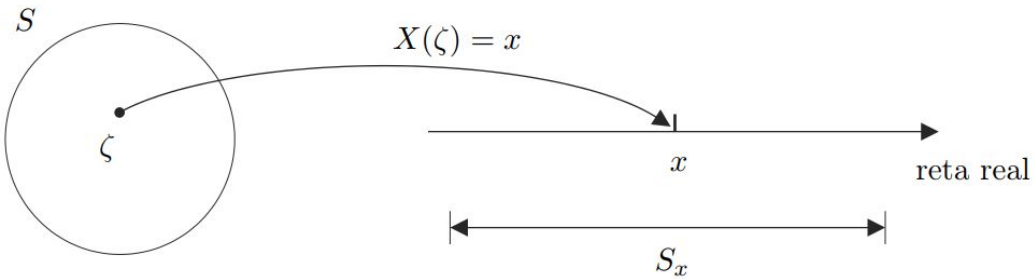


Viés de Seleção



Revisão: Variáveis Aleatórias

- Uma v.a. **X**, é uma função que associa um número real $X(\zeta)$ a cada resultado ζ no espaço amostral S de um experimento aleatório
- Podemos ver $X(\cdot)$ como uma função que mapeia os pontos amostrais $\zeta_1, \zeta_2, \ldots, \zeta_m$ em números reais $x_1, x_2, \ldots, x_n$



Revisão: Variáveis Aleatórias

- Exemplo 2.1 da apostila Ynoguti:

Especifique o espaço amostral de um experimento que consiste em jogar uma moeda 3 vezes.

Solução. O espaço amostral para este experimento é

$$S = \{CCC, CCK, CKC, CKK, KCC, KCK, KKC, KKK\},$$

onde C corresponde a “cara"e K corresponde a “coroa".

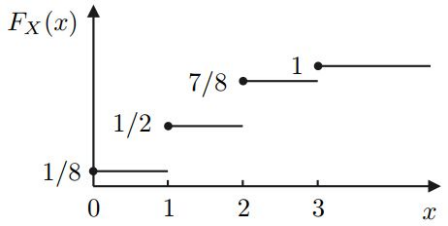
Seja X o número de caras em três jogadas da moeda. X associa a cada resultado ζ em S um número do conjunto $S_X = 0, 1, 2, 3$. A tabela abaixo lista os oito resultados de S e os valores de X correspondentes.

ζ	CCC	CCK	CKC	KCC	CKK	KCK	KKC	KKK
$X(\zeta)$	3	2	2	2	1	1	1	0

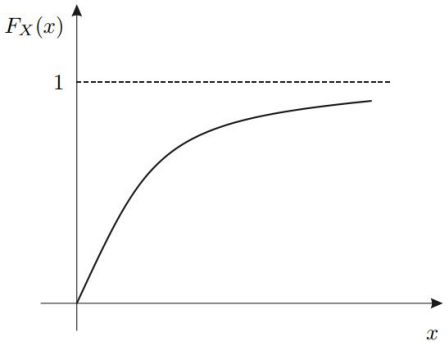
X é então uma v.a. que toma valores no conjunto $S_X = 0, 1, 2, 3$.

Revisão: Tipos de Variáveis Aleatórias

- Discretas:
  - Tomam valores de um conjunto finito
  - Aparecem geralmente em aplicações que envolvem contagens
  - Gráficos que as representam geralmente são degraus
  - Cálculos geralmente envolvem somatórios

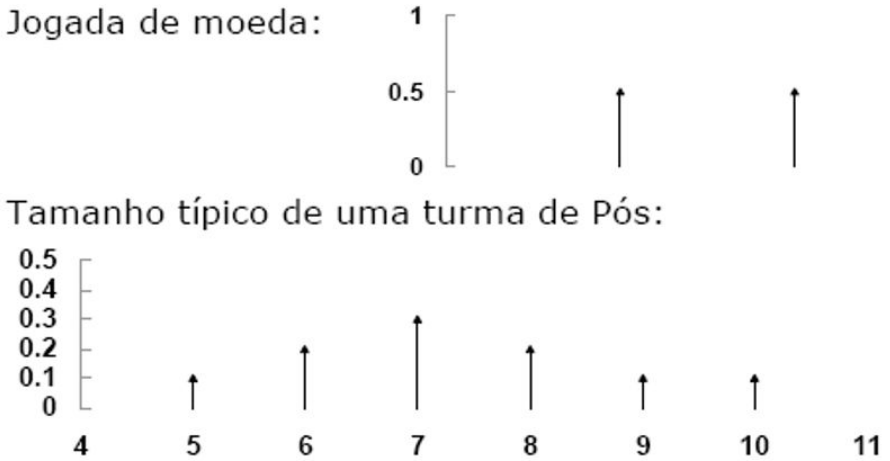


- Contínuas:
  - Tomam valores de um conjunto muito grande ou infinito
  - Gráficos que as representam geralmente são curvas contínuas
  - Cálculos geralmente envolvem integrais



Revisão: Função Massa de Probabilidade

- Conhecida como PMF
- Válida para v.a. discreta
- Associa cada valor de uma v.a. (eixo x) a uma probabilidade (eixo y)
- Geralmente representada por um gráfico de barras ou similares



Revisão: Função de Distribuição Cumulativa

- Conhecida como CDF
- Válida para v.a. discreta ou contínua
- Mapeia um valor para uma probabilidade cujo resultado é menor ou igual a x_a

$$F_{x_a}(x_a) = P(X \leq x_a)$$

- Curvas crescentes
- Se $P(X \leq x_a) = a$, esse resultado é chamado de a - percentil ou quantil

Revisão: Função Densidade de Probabilidade

- Conhecida como PDF
- Válida para v.a. contínua
- Derivada da CDF contínua

$$f(x) = \frac{dF(x)}{dx}, \; f(x) \geq 0$$

- Útil para determinar intervalos de probabilidades

$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$$

$$= \int_{x_1}^{x_2} f(x)dx$$

Distribuições Estatísticas

- Distribuições de frequências que ocorrem comumente na prática
- São descritas matematicamente por equações com poucos parâmetros
  - Distribuições de variáveis aleatórias discretas:
    - Bernoulli
    - Binomial
    - Geométrica
    - Poisson
  - Distribuições de variáveis aleatórias contínuas:
    - Uniforme
    - Exponencial
    - Gaussiana (Normal)

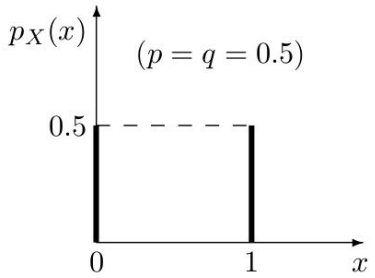
Distribuições Estatísticas (Discretas)

- Bernoulli
  - Dado um evento **A**, **x = 1** se A ocorre, **x = 0** se **A** não ocorre.
  - Ex.: Lançamento de 1 moeda. x = 1 para cara e x = 0 para coroa (booleano?)

PMF

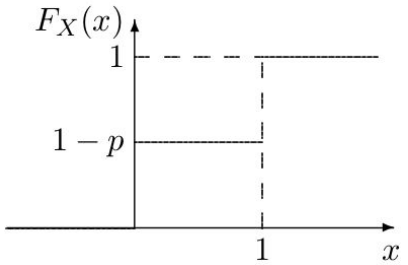
$$p_X(x) = \begin{cases} 1 - p = q, & X = 0 \\ p, & X = 1 \end{cases}$$

$$0 \leq p \leq 1$$



CDF

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$



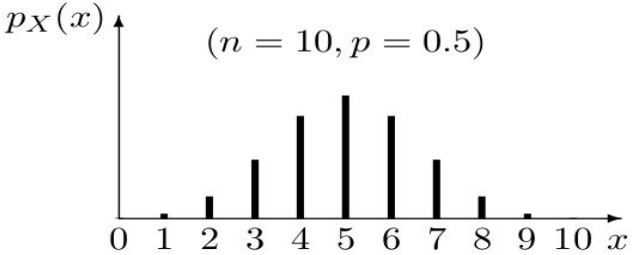
Distribuições Estatísticas (Discretas)

- Binomial
  - **X** é o número de sucessos, de probabilidade **p**, em **n** experimentos de Bernoulli
  - Ex.: Lançamento de 2 moedas, nº de pacotes que chegam no destino sem erro
  - Característica principal: numa dada amostra, tem-se nº de sucessos e de falhas

PMF

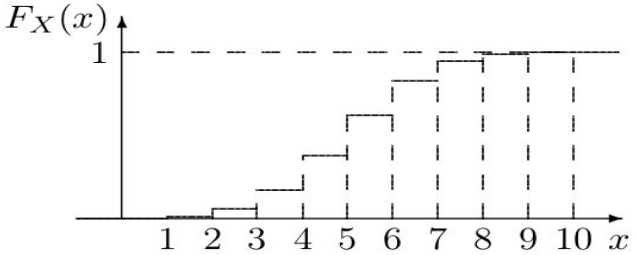
$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$x = 0, 1, \ldots, n$$



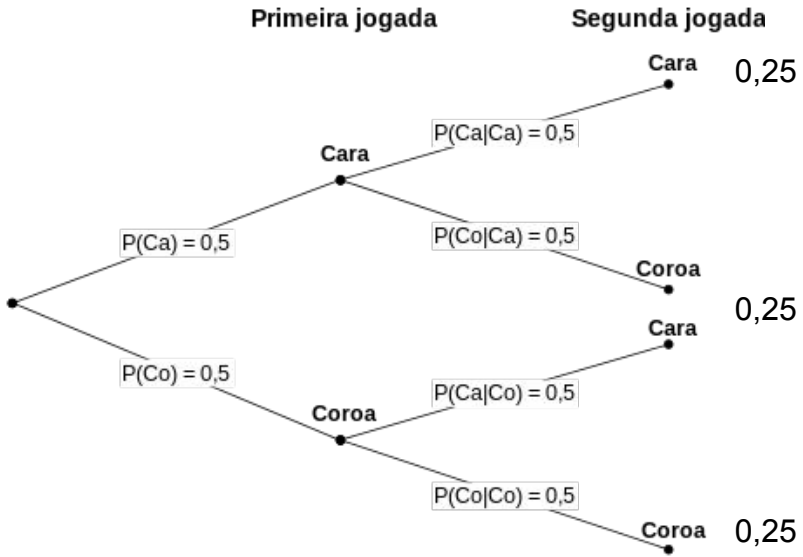
CDF

$$F_X(x) = \sum_k \binom{n}{x} p^x (1-p)^{n-x} u(x - x_k)$$



Distribuições Estatísticas (Discretas)

- Binomial
  - Ex.: No lançamento de duas moedas, podemos fazer um diagrama de árvore para facilitar o cálculo de probabilidade de cada resultado



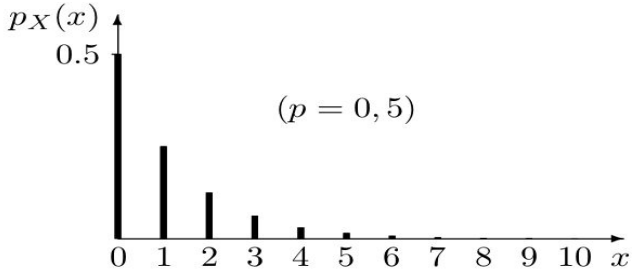
Distribuições Estatísticas (Discretas)

- Geométrica
  - Número de falhas antes do primeiro sucesso em testes de Bernoulli independentes
  - Memoryless: resultados futuros independentes de eventos passados

PMF

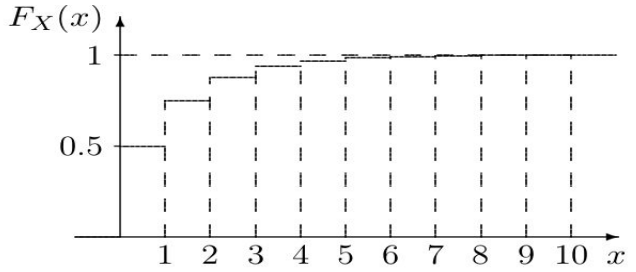
$$p_X(x) = p\,(1-p)^x$$

$$x = 0, 1, 2, \ldots$$



CDF

$$F_X(x) = \sum_{k=0}^{\infty} p\,(1-p)^k u(x-k)$$



Distribuições Estatísticas (Discretas)

- Geométrica
  - Ex.: Um dado honesto é lançado sucessivas vezes até que apareça pela primeira vez a face **1**. Seja **X** a variável aleatória que conta o número de ensaios até que corra o primeiro **1**. Qual a probabilidade de obtermos **1** no terceiro lançamento.
    - Dado é honesto: 1/6 de obtenção de cada face. Face 1 = sucesso e ocorre com **p = 1/6**. Qualquer outra face = fracasso e ocorre com a probabilidade **1-p = 5/6**. Podemos definir a variável aleatória

$$Y = \begin{cases} 1, & \text{se obtemos 1 no lançamento do dado} \\ 0, & \text{caso contrário} \end{cases}$$

    - Neste caso, **Y ~ Bernoulli(1/6)** e, se definirmos **X** como sendo a variável que representa o número de lançamentos até a obtenção do primeiro sucesso (face = 1), temos que **X ~ Geo(1/6)**. Portanto, se estamos interessados no cálculo da probabilidade de obter **1** no terceiro lançamento, precisamos calcular **P(X = 3)**, ou seja,

$$\mathbb{P}(X=3) = (1-p)^2 p = \left(1-\frac{1}{6}\right)^2 \frac{1}{6} = \left(\frac{5}{6}\right)^2 \frac{1}{6} = \frac{5^2}{6^3} \approx 0,115741.$$

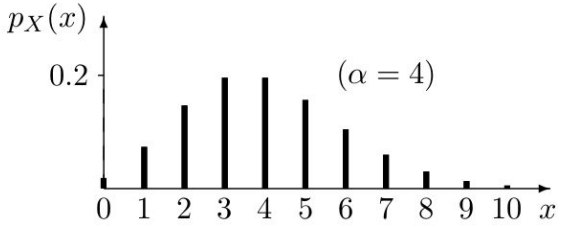
Distribuições Estatísticas (Discretas)

- Poisson
  - Número de ocorrências de um evento num dado intervalo de tempo ou região do espaço
  - Quando intervalos entre eventos são exponencialmente distribuídos com média $1/\alpha$
  - Ex.: nº de chegadas em um servidor em 1 hora

PMF

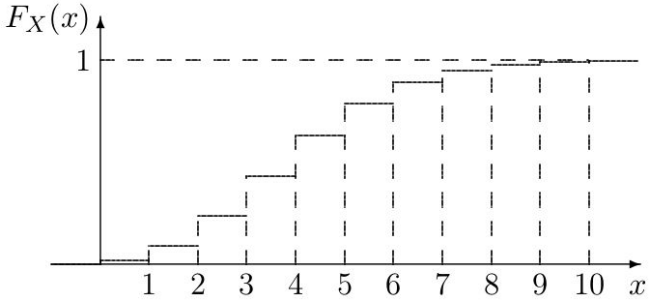
$$p_X(x) = \frac{\alpha^x}{x!} e^{-\alpha}$$

$$x = 0, 1, \ldots \quad \text{e} \quad \alpha > 0$$



CDF

$$F_X(x) = \sum_{k=0}^{\infty} \frac{\alpha^k e^{-\alpha}}{k!} u(x-k)$$



Distribuições Estatísticas (Discretas)

- Poisson
  - Ex.: Considere um processo que têm uma taxa de **0,2** defeitos por unidade. Qual a probabilidade de uma unidade qualquer apresentar:
    a) dois defeitos?
    b) um defeito?
    c) zero defeito?

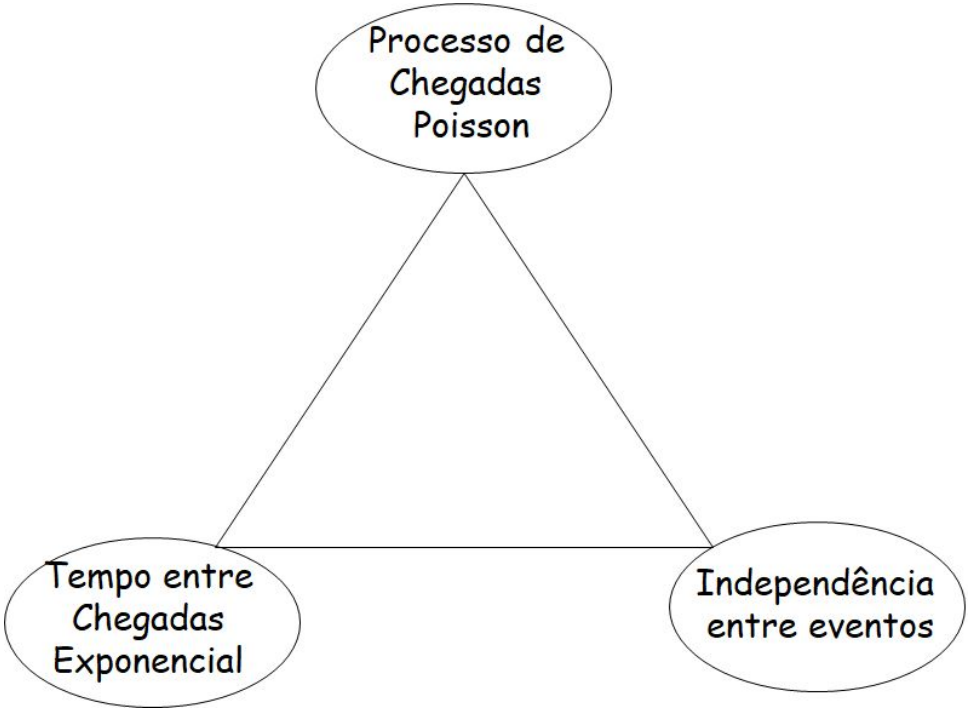
Neste caso, temos que $X \sim \text{Poisson}(\lambda)$ com $\lambda = 0,2$. Então

a) $\mathbb{P}(X = 2) = \dfrac{e^{-0,2}(0,2)^2}{2!} = 0,0164;$

b. $\mathbb{P}(X = 1) = \dfrac{e^{-0,2}(0,2)^1}{1!} = 0,1637;$

c. $\mathbb{P}(X = 0) = \dfrac{e^{-0,2}(0,2)^0}{0!} = 0,8187.$

Distribuições Estatísticas: Poisson x Exponencial

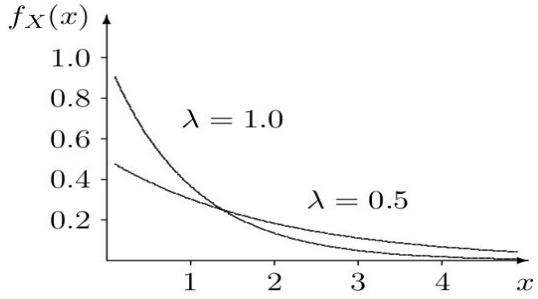


Distribuições Estatísticas (Contínuas)

- Exponencial
  - Modela tempo de duração de eventos que ocorrem segundo a distribuição de Poisson
  - Lambda é conhecido como parâmetro taxa ou parâmetro escala

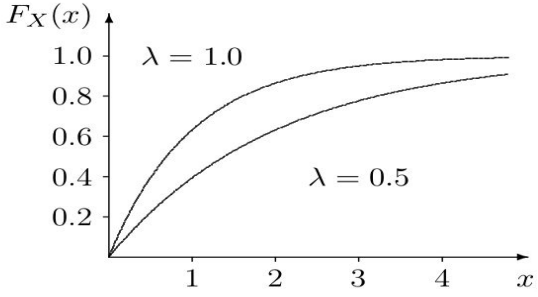
PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \text{ e } \lambda > 0 \\ 0 & \text{caso contrário} \end{cases}$$



CDF

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \lambda > 0 \\ 0 & \text{caso contrário} \end{cases}$$



Distribuições Estatísticas (Contínuas)

- Exponencial
  - Ex.: O tempo até a falha do ventilador de motores a diesel tem uma distribuição Exponencial com parâmetro $\lambda = \frac{1}{28700}$ horas. Qual a probabilidade de um destes ventiladores falhar nas primeiras **24000** horas de funcionamento?

$$\mathbb{P}[0 \leq X \leq 24000] = \int_0^{24000} f(x)dx = \int_0^{24000} \frac{1}{28700} \exp\left(-\frac{x}{28.700}\right) = 0,567.$$

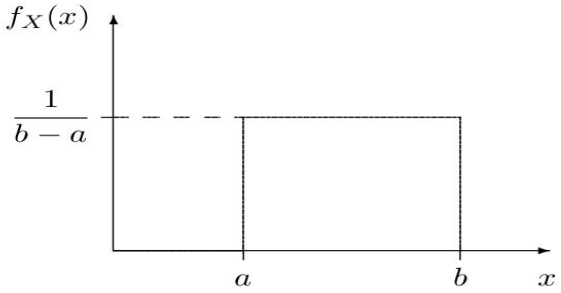
Ou seja, a probabilidade de um destes ventiladores falhar nas primeiras 24000 horas de funcionamento é de, aproximadamente, 56,7%.

Distribuições Estatísticas (Contínuas)

- Uniforme
  - Todos os valores no intervalo da reta real são equiprováveis

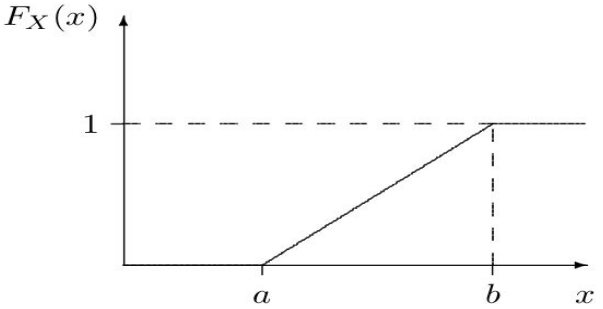
PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{caso contrário} \end{cases}$$



CDF

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$



Distribuições Estatísticas (Contínuas)

- Uniforme
  - Ex.: A ocorrência de panes em qualquer ponto de uma rede telefônica de **7 km** foi modelada por uma distribuição Uniforme no intervalo **[0, 7]**. Qual é a probabilidade de que uma pane venha a ocorrer nos primeiros **800 metros**? E qual a probabilidade de que ocorra nos **3 km** centrais da rede?

A função densidade da distribuição Uniforme é dada por $f(x) = \frac{1}{7}$ se $0 \leq x \leq 7$ e zero, caso contrário. Assim, a probabilidade de ocorrer pane nos primeiros 800 metros é

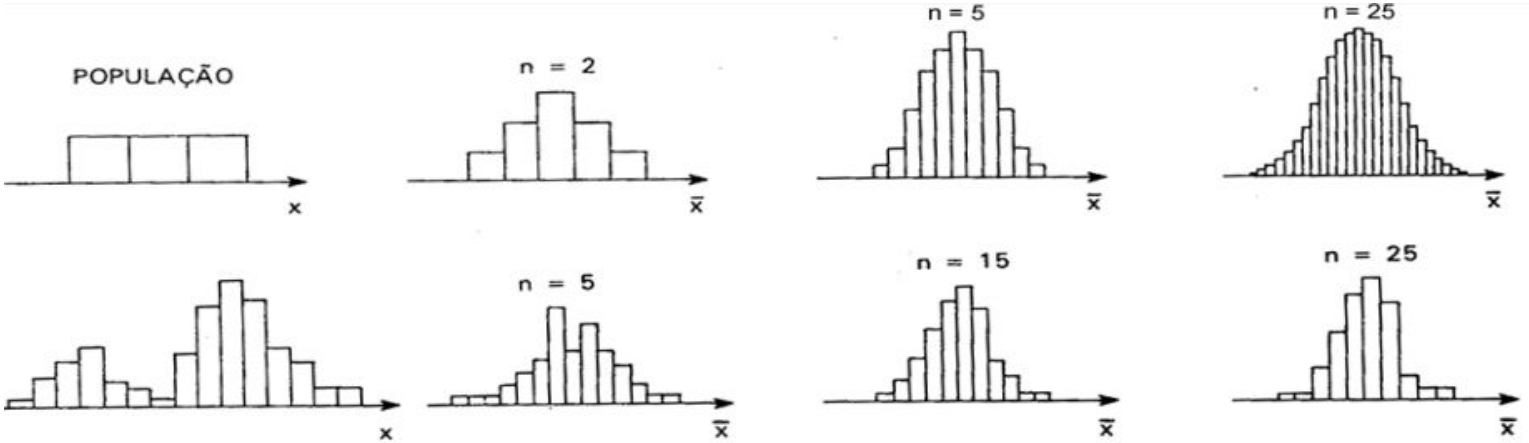
$$\mathbb{P}(X \leq 0,8) = \int_0^{0,8} f(x)dx = \frac{0,8-0}{7} = 0,1142.$$

e a probabilidade de ocorrer pane nos 3 km centrais da rede é

$$\mathbb{P}(2 \leq X \leq 5) = \int_2^5 f(x)dx = \mathbb{P}(X \leq 5) - \mathbb{P}(X \leq 2) = 5/7 - 2/7 \approx 0,4285.$$

Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)
  - Distribuição da maioria das amostras, denotada por $N(\mu,\sigma)$ (Normal unitária: $N(0,1)$)
  - Pelo fato de ocorrerem tão frequentemente, na prática, é conhecida como Normal
  - Teorema do Limite Central:
    - Conforme amostras aumentam, a distribuição da média de suas v.as tende à Normal

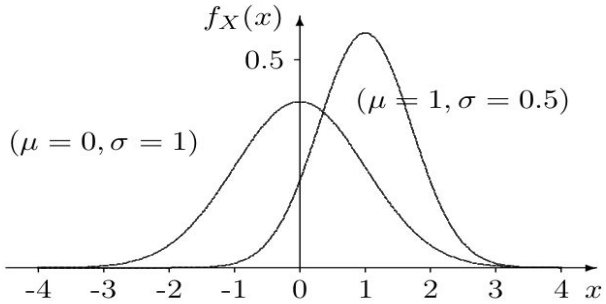


Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)

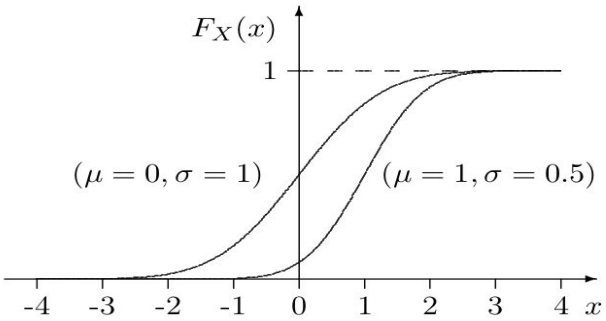
PDF

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



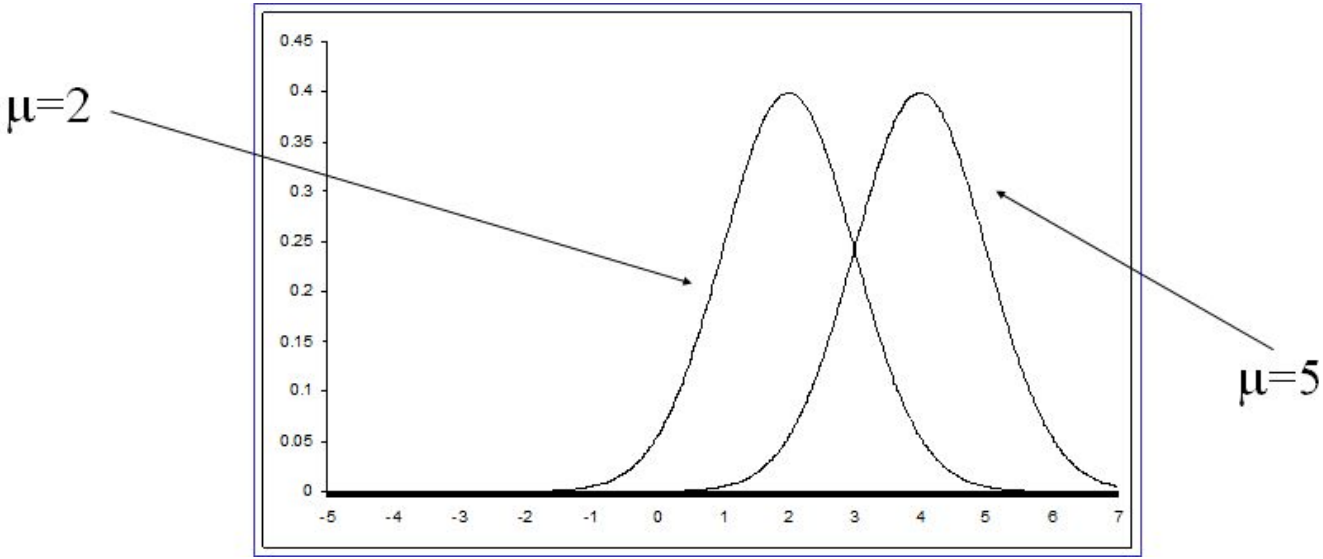
CDF

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$



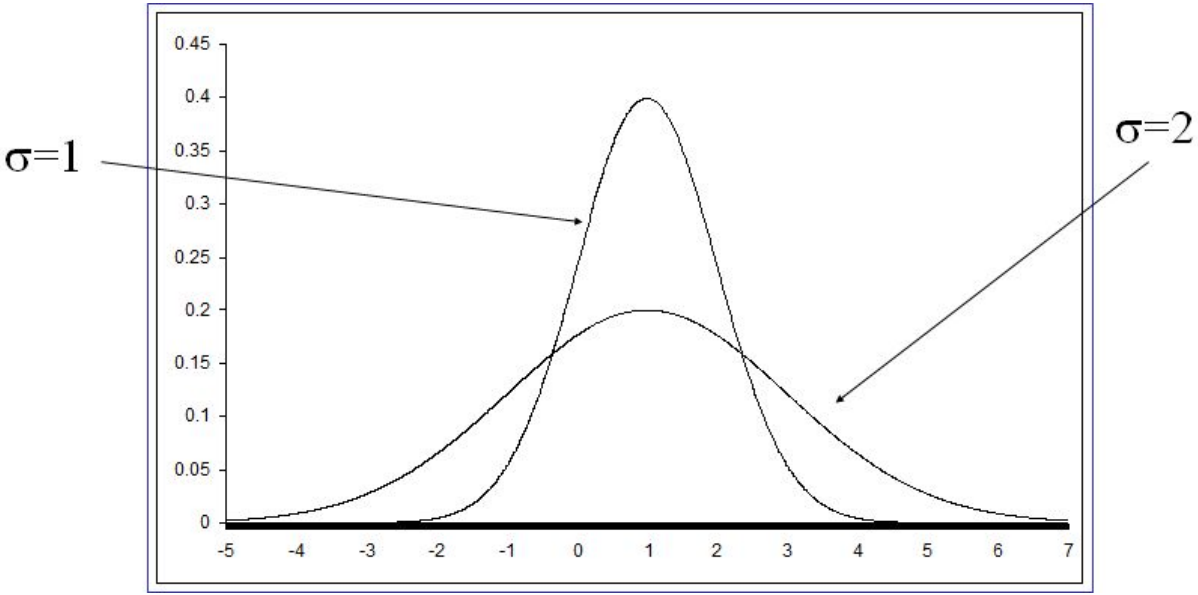
Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)



Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)



Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)
  - Teorema:

> *Se X é uma variável aleatória Gaussiana com parâmetros μ e σ, a fdc de X é*
>
> $$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
>
> *E a probabilidade de X estar no intervalo $(a, b]$ é*
>
> $$P[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

  - Usado, também, para transformar normais comuns em normais unitárias ($N(0,1)$):

$$X = Z = \frac{x - \mu}{\sigma}$$

Distribuições Estatísticas (Contínuas)

- Gaussiana (Normal)
  - Ex.: O peso médio de 800 porcos de uma fazenda é de 64 kg, e o desvio padrão é de 15 kg. Dado que esse peso é distribuído de forma normal, quantos porcos pesarão entre 42 e 73 kg.

Para resolvermos este problema primeiramente devemos padroniza-lo, ou seja,

$$Z = \frac{x-\mu}{\sigma} \sim N(0,1) \qquad Z = \frac{x-64}{15}$$

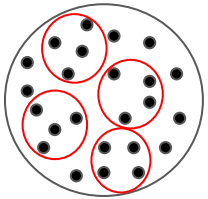
Então o valor padronizado de 42kg é de $\frac{42-64}{15} \approx -1,47$ e de 73kg é de $0,6$.

Assim a probabilidade é de

$$\mathbb{P}(-1,47 \leq Z \leq 0,6) = \mathbb{P}(Z \leq 0,6) - \mathbb{P}(Z \leq -1,47) = 0,7257 - 0,0708 = 0,6549.$$

Portanto, o número aproximado que se espera de porcos entre 42kg e 73kg é $800 \cdot 0,6549 \approx 524$.

Distribuições Estatísticas



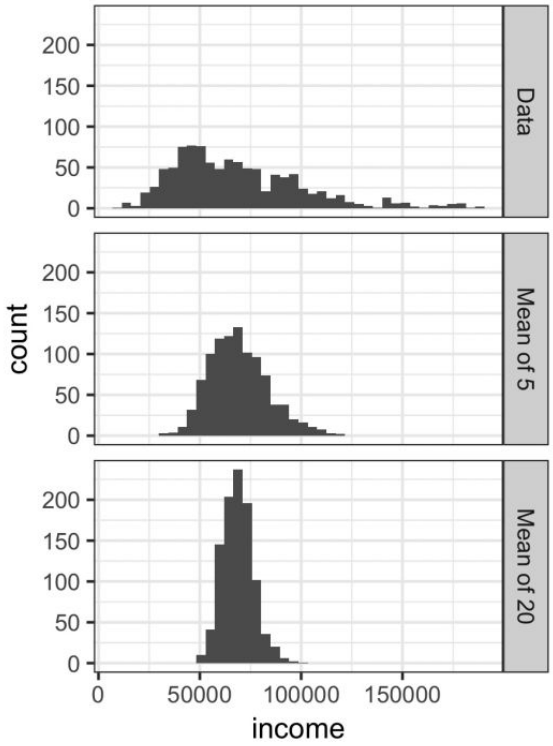
- Distribuição amostral vs Distribuição de dados
  - “Distribuição Amostral” é a distribuição de alguma estatística de várias amostras
    - Ex.: Distribuição amostral da média, da variância, etc., de cada amostra
    - Mostra como as amostras de uma dada população diferem entre si
  - “Distribuição de dados” é a distribuição das v.as em uma dada amostra

“The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of the data itself. The larger the sample that the statistic is based on, the more this is true. Also, the larger the sample, the narrower the distribution of the sample statistic.”

BRUCE, P.; BRUCE. A. Practical Statistics for Data Scientists. 1. ed. O’Reilly Media, 2017.

Distribuições Estatísticas

Ex.:



Distribuições Estatísticas

- Erro padrão: métrica que mostra a variabilidade de uma distribuição amostral
  - Calculado dividindo-se o desvio padrão **S** pela raiz do tamanho **n** da amostra

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

```
from scipy import stats

a = [2, 3, 5, 0, 1]

se = stats.sem(a)
```

- Erro Padrão vs Desvio Padrão
  - Erro Padrão mostra como as médias podem variar de uma amostra para outra. Serve para avaliar a confiabilidade da média calculada
  - Desvio Padrão é a medida de dispersão da amostra em relação à média da mesma

Distribuições Estatísticas

- Ex. 01: Numa população obteve-se desvio padrão de 1,32 com uma amostra aleatória de 121 elementos. Sabendo que para essa mesma amostra obteve-se uma média de 6,25, determine o valor mais provável para a média dos dados.
  - Para determinarmos o valor mais provável da média dos dados devemos calcular o erro padrão da estimativa. Assim, teremos:

$$S_x = \frac{1{,}32}{\sqrt{121}} = 0{,}12$$

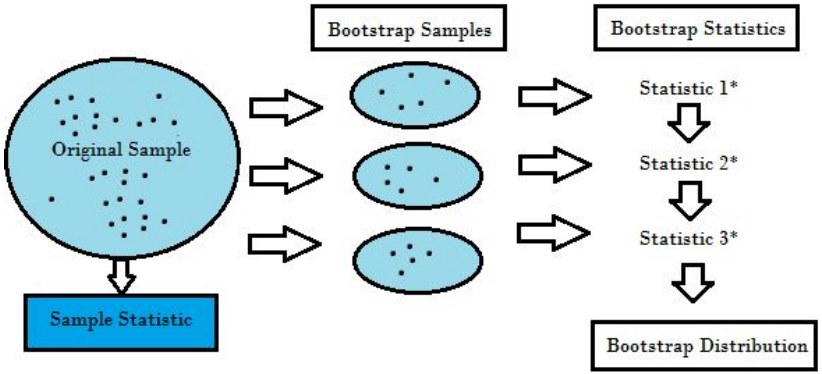
  - Finalizando, o valor mais provável para a média dos dados obtidos pode ser representado por:

$$\bar{X} = 6{,}25 \pm 0{,}12$$

Bootstrap

- Para obter distribuições amostrais muitas vezes é difícil ter acesso a várias amostras de uma dada população
- Solução: Bootstrap

1. Resample a data set *x* times,
2. Find a summary statistic (called a **bootstrap statistic**) for each of the x samples,
3. Estimate the standard error for the bootstrap statistic using the standard deviation of the bootstrap distribution.



Bootstrap

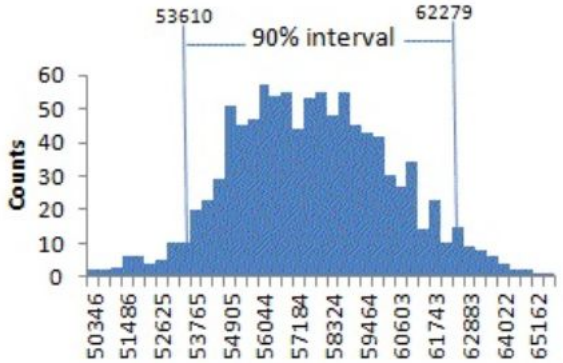
- Dica: pip install bootstrapped

```
1 import numpy as np
2 import bootstrapped.bootstrap as bs
3 import bootstrapped.stats_functions as bs_stats
4
5 mean = 100
6 stdev = 10
7
8 population = np.random.normal(loc=mean, scale=stdev, size=50000)
9
10 # take 1k 'sample ' from the larger population
11 sample  = population[:1000]
12
13 print(bs.bootstrap(sample , stat_func=bs_stats.mean))
14 >> 100.08  (99.46, 100.69)
15
16 print(bs.bootstrap(sample , stat_func=bs_stats.std))
17 >> 9.49  (9.92, 10.36)
```

- Alternativa: verificar **numpy.random.choice**

Intervalo de Confiança da Distribuição Amostral

- Dada uma amostra de tamanho **n** e uma estatística amostral de interesse:
  1. Faça o processo de bootstrap para extrair tais estatísticas de interesse
  2. Plote a distribuição amostral sobre tais estatísticas de interesse
  3. Para um intervalo de confiança de x%, desconsidere os trim[(1-[x/100]/2)]% dos extremos da distribuição plotada



AAG03 Tarefa em Dupla

- Desenvolver no Jupyter Notebook exemplos (enunciados e resoluções) que expliquem cada uma das distribuições apresentadas na aula de hoje:
  - Bernoulli
  - Binomial
  - Poisson
  - Geométrica
  - Uniforme
  - Exponencial
  - Gaussiana (Normal)
- Regras:
  1. Código e resultados devem ser explicados em Markdown com comandos LaTex
  2. Os formatos de entrega devem ser .pdf e .ipynb (código fonte+markdowns)