Data Mining

Tópicos em análise de Dados

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2024

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

*A maior parte deste material foi retirado do livro: “Joaquim V. C. Assunção. **Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R**. 192 páginas. Novatec. 2021. ISBN-10 : 6586057507.”

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Mineração & análise de dados

- Atividade que pode ser realizada antes e depois da mineração.
  - Antes, com o objetivo de entender os dados para melhor formular hipóteses ou definir um tipo de algoritmo alvo.
  - Depois, com o objetivo de validar padrões, apresentar e comparar resultados.

Análise de dados

- O básico para análise de dados é também o elementar da estatística.
  - Comumente, se obtêm em conjuntos valores estatísticos como média, mediana, quartis, e desvio padrão.

//Tais valores servem para descrever conjuntos de dados; da mesma forma como peso, altura, tamanho e cor do cabelo servem para descrever a aparência de uma pessoa.

Technical demo, hands on!

- Em R, use as funções `summary()` e `sd()` para obter média, mediana, quartis, e desvio padrão.

- *Complete o gráfico adicionando os demais valores*

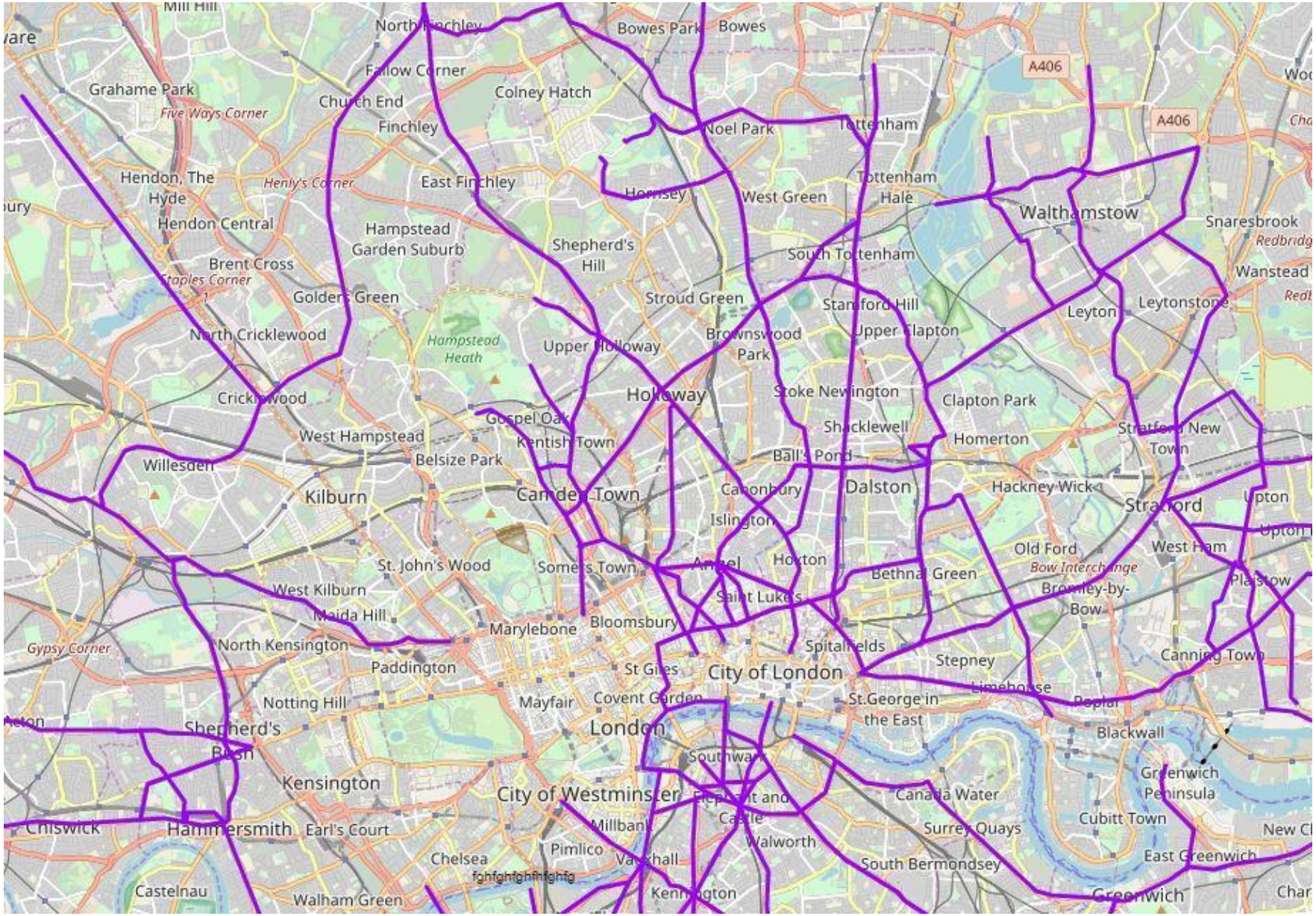
```
X <- sample(40)
plot(X, ylim = c(0,60), ylab="Valores")
text(10,56,paste('Média: ',summary(X)[4]))
```

Análise de dados

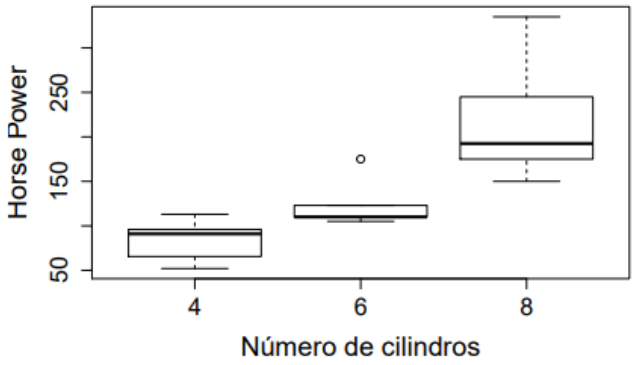
- Gráficos são importantes para entender, mostrar e comparar conjuntos de dados.
- São amplamente usados nas mais diversas tarefas, para os mais diversos fins, e com diversos níveis de complexidade.

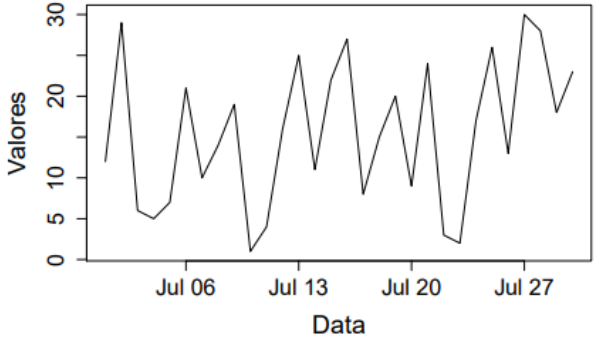
Análise de dados

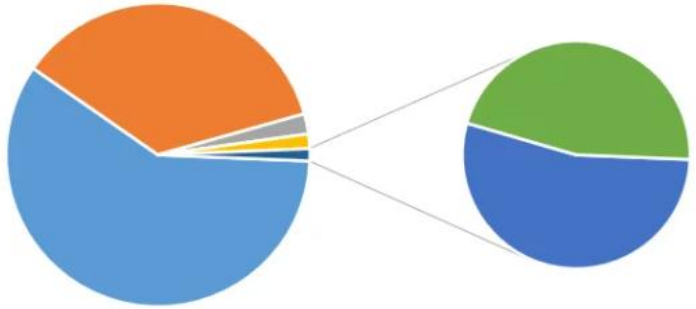
- Há diversos tipos de gráficos e incontáveis combinações que podemos fazer para adicionar mais informação em um gráfico. Porém...
- Deve-se tomar cuidado ao criar um gráfico com muitas informações. O excesso de informações não contribui para a qualidade de uma imagem.

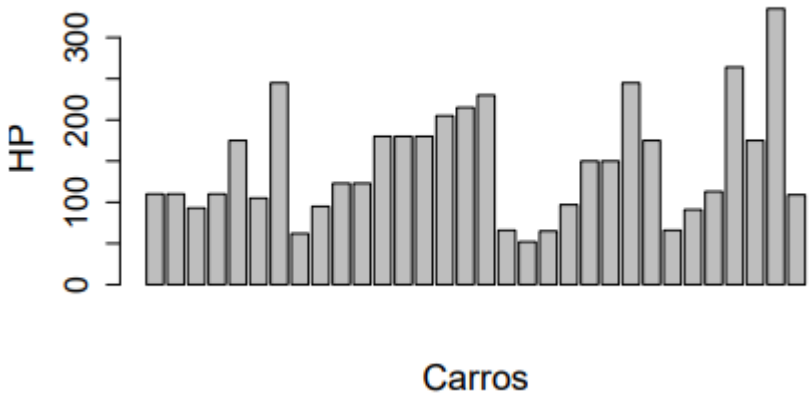

Mill Hill
North Finchley
Bowes Park
Bowes
Grahame Park
Five Ways Corner
Church End
Finchley
Colney Hatch
Noel Park
Tottenham
A406
Hendon, The
Hyde
Henly's Corner
East Finchley
Hornsey
West Green
Tottenham Hale
Walthamstow
Hendon Central
Hampstead Garden Suburb
Shepherd's Hill
South Tottenham
Snaresbrook
Wanstead
Brent Cross
Staples Corner
Golders Green
Stroud Green
Stamford Hill
Leyton
Leytonstone
North Cricklewood
Hampstead Heath
Upper Holloway
Brownswood Park
Upper Clapton
Holloway
Stoke Newington
Clapton Park
Cricklewood
Gospel Oak
Shacklewell
West Hampstead
Kentish Town
Homerton
Stratford New Town
Willesden
Belsize Park
Ball's Pond
Kilburn
Camden Town
Canonbury
Dalston
Hackney Wick
Stratford
Upton
Islington
Old Ford
West Ham
St. John's Wood
Somers Town
Arsenal
Hoxton
Bow Interchange
Bethnal Green
Plaistow
West Kilburn
Saint Luke's
Bromley-by-Bow
Maida Hill
Marylebone
Bloomsbury
Gypsy Corner
North Kensington
Spitalfields
Stepney
Canning Town
Paddington
St Giles
City of London
Limehouse
Notting Hill
Mayfair
Covent Garden
St.George in the East
Poplar
Shepherd's Bush
London
Blackwall
Southwark
Greenwich Peninsula
Kensington
City of Westminster
Canada Water
Chiswick
Hammersmith
Earl's Court
Millbank
Cubitt Town
Surrey Quays
New Cross
Pimlico
Walworth
East Greenwich
Chelsea
Vauxhall
South Bermondsey
Castelnau
Walham Green
Kennington
Greenwich

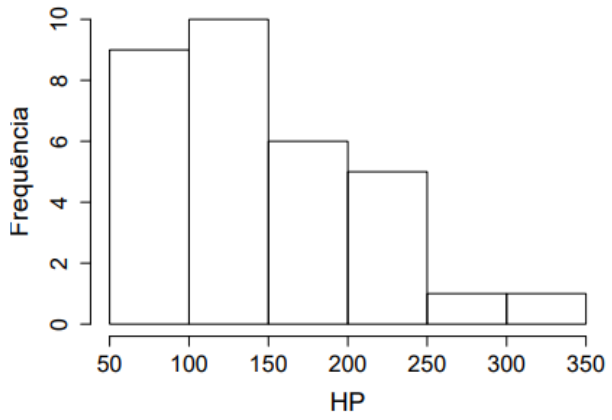
Alguns gráficos comuns para análise e visualização de dados

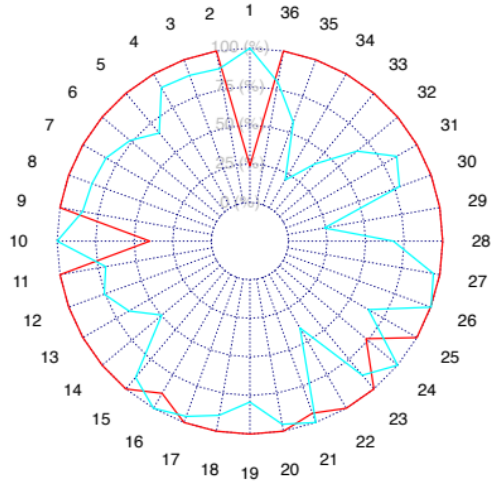












Technical demo, hands on!

- **Boxplot em R**

```
carros_HPWT <- mtcars[,c(2,4,6)]
boxplot(hp~cyl,data=carros_HPWT, xlab="Número de cilindros",
        ylab="Horse Power")
```

- **Gráfico de linha em R**

```
Y <- sample(30)
X <- seq(as.Date("2020-07-01"), as.Date("2020-07-30"), by = "days")
plot(X,Y,type='l', xlab = 'Data', ylab = 'Valores')
```

```
DF <- data.frame(X,Y)
ggplot(DF, aes(x=X, y=Y)) + geom_line()
```

Requer pacote: ***ggplot2***

Distribuições estatísticas

- Distribuições estatísticas revelam como os dados estão distribuídos de acordo com seus valores.

- Algumas das distribuições comuns são: Distribuição Normal, Uniforme, Bernoulli, Binomial, e Pareto.

Distribuições estatísticas

- Cada distribuição possui duas funções: PDF (Probability Density Function) e CDF (Cumulative Density Function).

- A densidade acumulada (CDF) é sempre um valor de 0 a 1 (100%). Já a probabilidade da densidade é um valor específico da variável em questão.

- A função PDF é comumente representada por f(x) (minúsculo), e a função CDF é representada por F(x) (maiúsculo).

Distribuição normal

- Também conhecida como bell curve (curva do sino).
- Fundamental importância para diversas áreas
- Algo em torno de 68% da população estará entre um desvio padrão da média.

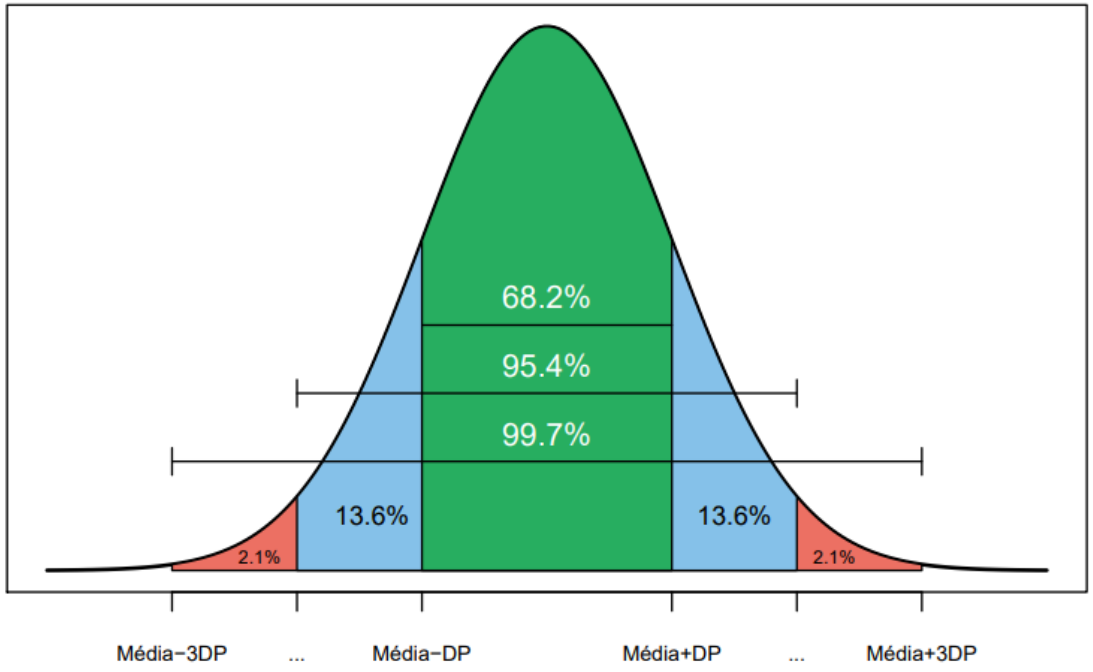


Figura 2.10: Distribuição normal.

Distribuição Uniforme

- Clássico exemplo de distribuição uniforme é a jogada de dados (objeto).

$$P(X) = \frac{1}{x_1 - x_n + 1}$$

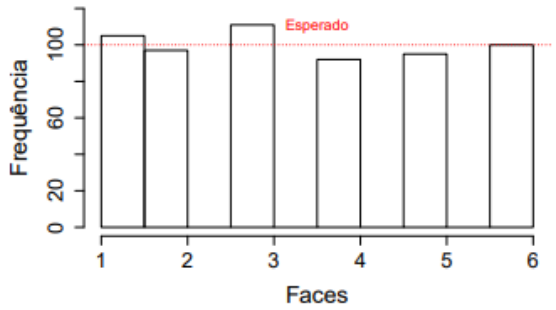


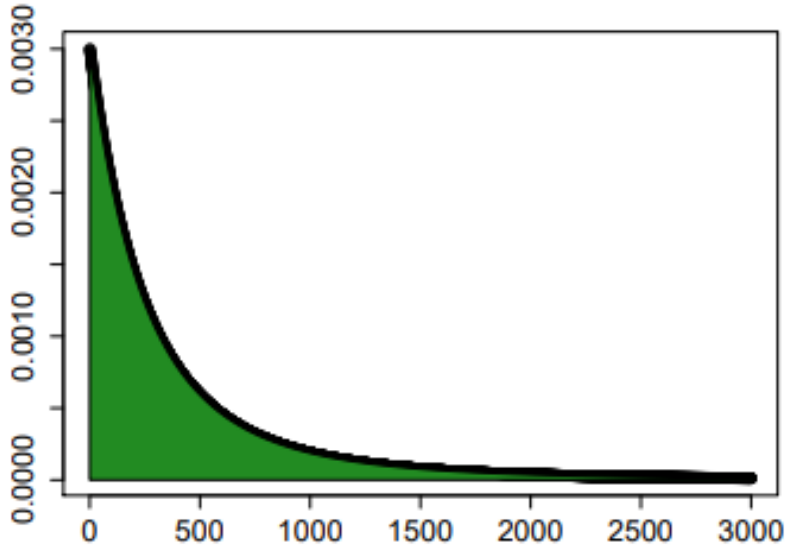
Figura 2.12: Exemplo de conjunto com distribuição uniforme.

- Por exemplo, para a jogada de dados, temos: (1/(6 − 1 + 1)) = 0.1667 para cada valor.

Distribuição de Pareto

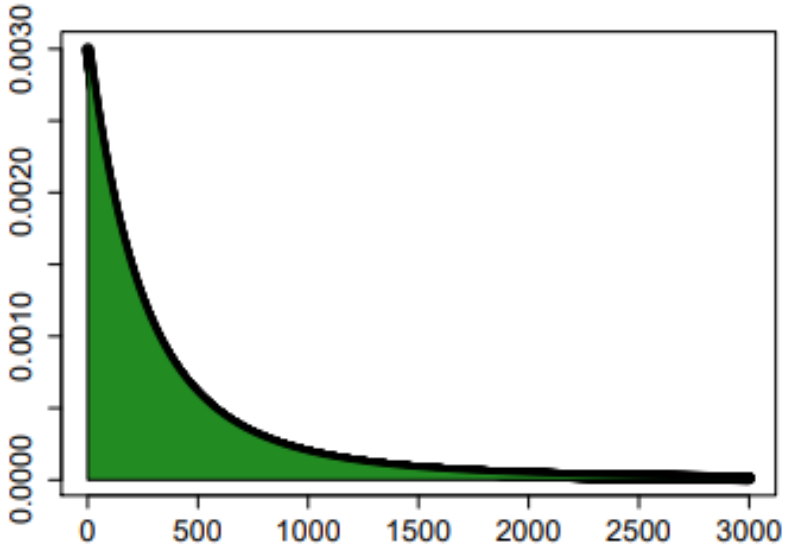
- A regra geral é 80 – 20, onde mais de 80% dos recursos são atrelados a menos de 20% dos indivíduos.

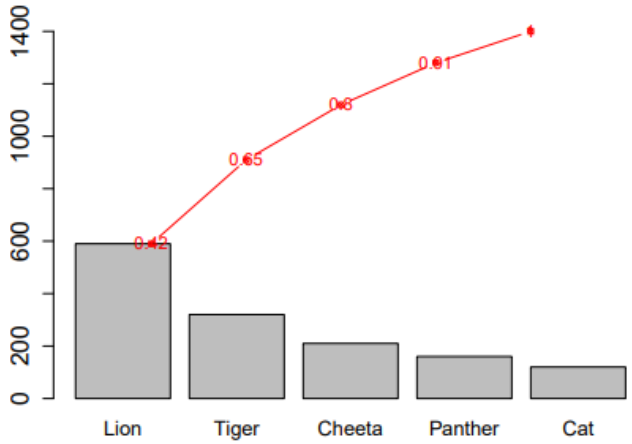
- Comumente encontramos gráficos com a densidade acumulada CDF para enfatizar a perspectiva do total.



Distribuição de Pareto

- Além disso, indivíduos da elite são mostrados a esquerda do gráfico; o que é o oposto de outras distribuições





Distribuição de Bernoulli

- A distribuição de Bernoulli é binária no sentido de que só pode haver duas saídas: “sucesso” ou “falha”.

- A densidade de Bernoulli depende diretamente da chance de sucesso atribuída a variável de probabilidade p

$$PDF(K) = \begin{cases} 1-p & SE \quad p == 0 \\ p & SE \quad p == 1 \end{cases}$$

Discretas e contínuas

- Dentre as categorias de distribuições, podemos separar em duas, discretas e continuas.

- Bernoulli e Binomial são exemplos de distribuições discretas.

- Dentre as distribuições contínuas, algumas comumente usadas são Poisson e Exponencial.