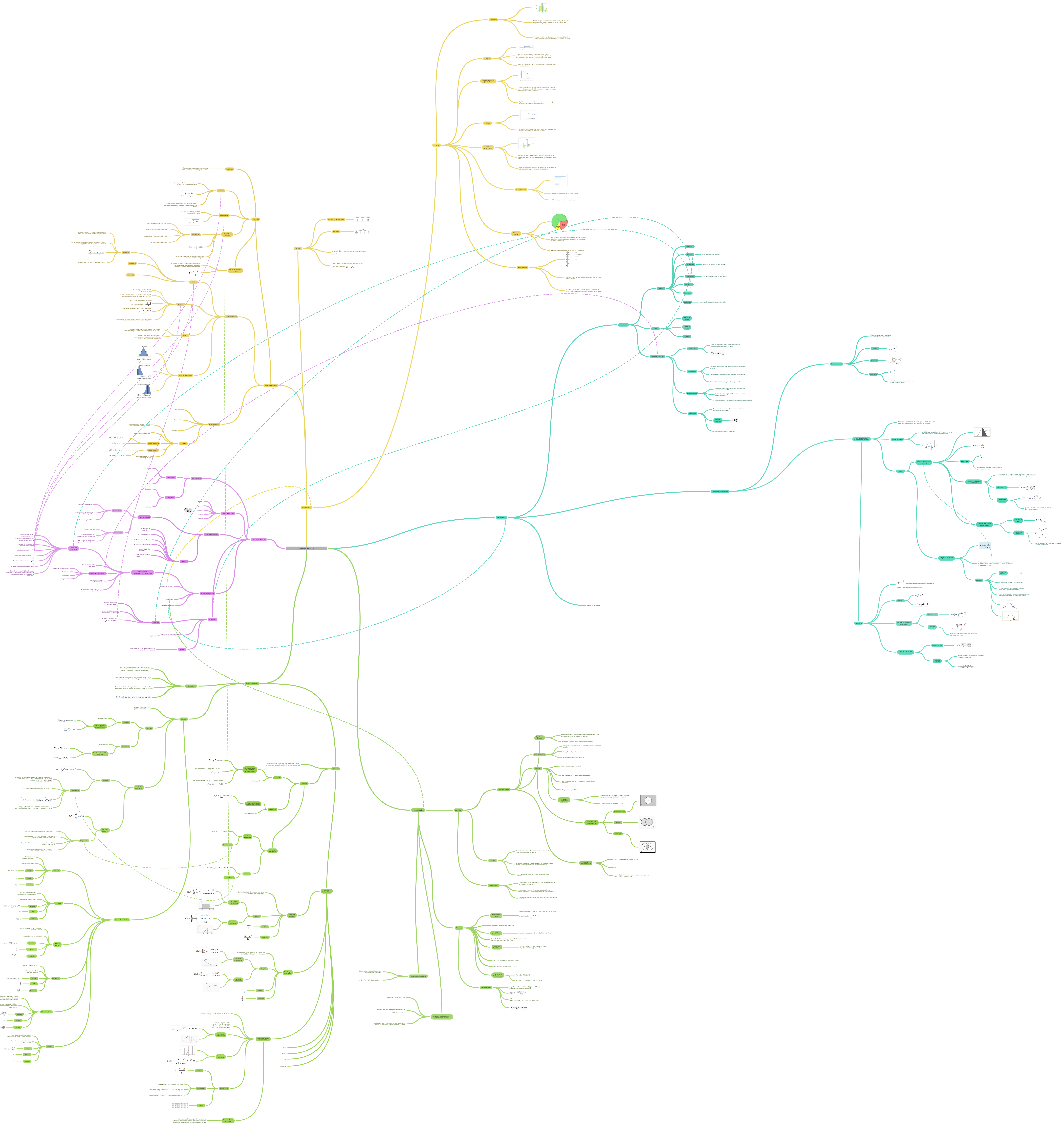

Estatística Básica

- Descritiva
  - Gráficos
    - Histograma
      - Representação gráfica em colunas de um conjunto de dados previamente tabulado e dividido em classes de valores uniformes ou não uniformes.
      - Pode ser Simétrico (ou Gaussiano ou Unimodal), Distorcido à Direita, Distorcido à Esquerda, Bimodal, Multimodal ou Platô.
    - Boxplot
      - É uma caixa que apresenta cinco medidas para se fazer análises estatísticas: a mediana, o primeiro quartil, o terceiro quartil, o menor valor e o maior valor no conjunto de dados.
      - Serve para visualizar o centro, a dispersão e a assimetria de um conjunto de dados
    - Gráfico de Dispersão (Scatter Plot)
      - Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo y e o valor da outra variável no eixo x.
      - O gráfico de dispersão é útil para verificar como dois conjuntos de dados comparáveis concordam entre si.
    - Linhas
      - Um gráfico de linhas é muitas vezes usado para visualizar uma tendência nos dados em intervalos de tempo
    - Colunas ou Barras Vertical
      - Os gráficos de colunas são úteis para mostrar alterações de dados durante um período ou para ilustrar a comparação entre itens.
      - Os gráficos de colunas podem ser agrupados, empilhados ou 100% empilhados (para visualizar distribuições).
    - Barras Horizontal
      - O propósito é o mesmo do de barras vertical
      - Utilizasse quando se tem muitas categorias.
    - Setores ou Pizza
      - Ativos: 62%; Inativos: 25,9%; Pensões: 12,1%
      - Um diagrama circular em que os valores de cada categoria estatística representada são proporcionais às respectivas medidas dos ângulos
      - Evitar este gráfico quando houver mais de 4 categorias
    - Ramo e Folhas
      - 1 | 6 7 7 8 8 9 9 9
      - 2 | 0 0 1 1 2 3 3 4 4 4 8 9 9
      - 3 | 0 1 2 2 2 1 7 8 8
      - 4 | 2 3 6 6 6 6 8 9
      - 5 | 2 3 3 4 4 8
      - 6 | 0 2 5 6
      - 7 | 1 4
      - Dispositivo para apresentação de dados quantitativos em um formato gráfico
      - São úteis para mostrar a densidade relativa e a forma dos dados, dando ao leitor uma rápida visão geral da distribuição.
  - Tabelas
    - Distribuição de Frequência
    - Cruzadas
    - O símbolo ├ é utilizado para representar o intervalo aberto/fechado.
    - Uma maneira de determinar o número de classes é através do cálculo $K = \sqrt{n}$
  - Medidas Descritivas
    - Dispersão
      - Amplitude
        - É definida como sendo a diferença entre o maior e o menor valor do conjunto de dados
      - Variância
        - Utilizada para identificar quanto próximo ou afastado o valor está da média
        - $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$
        - A variância tem a desvantagem de apresentar unidade de medida igual ao quadrado da unidade de medida dos dados.
      - Desvio Padrão
        - Utilizado para voltar a Variância para a medida original
        - $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$
      - Coeficiente de Variação
        - Classificação
          - < 10%: homogeneidade muito alta
          - de 10% a 20%: homogeneidade alta
          - de 20% a 30%: homogeneidade média
          - 30%: homogeneidade baixa
        - $CV_{(\%)} = \frac{s}{\bar{X}} \cdot 100\%$
        - É utilizado, quando se compara amostras com médias ou medidas desiguais.
      - Escore Padronizado (Escore Z)
        - O objetivo de se calcular o escore z é expressar em unidades de desvio padrão quanto um determinado número está distante da média
        - $z = \frac{x - \bar{x}}{s}$
    - Tendência Central
      - Média
        - Aritmética
          - A média aritmética é a medida sintetizadora mais adequada quando não existem muitos Outliers.
          - O seu valor é obtido através da soma de todos os valores e dividido pelo tamanho da amostra ou população.
          - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
          - Também conhecida como Esperança Matemática
        - Ponderada
        - Harmônica
      - Mediana
        - É o valor que divide o conjunto de dados ao meio
        - Para calcular a mediana é necessário que o conjunto de dados esteja organizado em ordem crescente
        - Se n é ímpar, a mediana é dada pelo valor que ocupa a posição $\frac{n+1}{2}$
        - Se n é par, a mediana será a média dos valores que ocupam as posições $\frac{n}{2}$ e $\frac{n+2}{2}$
        - A mediana deve ser usada sempre que possível como medida representativa de distribuições fortemente assimétricas.
      - Moda
        - Dado um conjunto de valores, a Moda (Mo) desses valores será aquele que se repetir o maior número de vezes
        - Uma amostra que não tem moda que é chamada de amodal, duas modal é bimodal e três ou mais é chamada multimodal
      - Forma de Distribuição
        - Simétrica: Moda = Média = Mediana
        - Assimétrica à direita: Média > Mediana > Moda
        - Assimétrica à esquerda: Média < Mediana < Moda
    - Posição Relativa
      - Quartis
      - Decis
      - Percentis
      - Outliers
        - São valores muito grandes ou muito pequenos em relação aos demais
        - Alteram enormemente a média e a variabilidade dos dados
        - Outlier Moderado
          - $OS = Q_3 + 1,5 \times A$
          - $OI = Q_1 - 1,5 \times A$
        - Outlier Extremo
          - $OS = Q_3 + 3 \times A$
          - $OI = Q_1 - 3 \times A$
        - Amplitude no cálculo de Outlier é a diferença de Q3 - Q1
- Conceitos Básicos
  - Tipo de Dados
    - Qualitativos
      - Nominal
      - Ordinal
    - Quantitativos
      - Discretos
      - Contínuos
  - Objetivo Estatística
    - Coletar
    - Organizar
    - Descrever
    - Analisar
    - Intepretar
  - Estudos Estatísticos
    - Tipos de Estudos
      - Observacional
        - Amostras Independentes
        - Características da População não são manipuladas
        - Ex. Censo, Pesquisa Eleitoral
      - Experimental
        - Amostras Pareadas
        - O grupo de indivíduos é manipulado para avaliar efeito
        - Ex. Eficácia de Tratamentos, Reações de Processos Químicos
    - Fases
      1. Identificação do Problema
      2. Coleta de Dados
      3. Tratamento dos Dados
      4. Análise e interpretação
      5. Apresentação dos resultados
      6. Obtenção de relatório técnico
  - Áreas da Estatística
    - Estatística Descritiva
    - Probabilidade
    - Estatística Inferencial
    - Amostragem e Planejamento de Experimentos
      - Processo de escolha da amostra
      - Métodos de Amostragem
        - Aleatória (Casual) Simples
        - Sistemática
        - Estratificada
        - Conglomerado
      - Parte inicial de qualquer estudo estatístico
      - Amostra é um subconjunto de elementos de uma população
      - Estimador ou Estatística
        - O tamanho da amostra é denotada como n
        - Descreve numericamente uma característica da amostra
        - O resultado de um parâmetro estimador é a estimativa
        - A média é denotada como $\bar{x}$
        - A mediana é denotada como $\tilde{x}$
        - A variância é denotada como s^2
        - O desvio padrão é denotado como S
        - Graus de Liberdade (GL) é o número de elementos da amostra(n), menos o número de parâmetros (medidas da população) já estimados.
  - População
    - Parâmetro
      - O tamanho da população é denotado como N
      - Descreve numericamente uma característica da população
      - A média é denotada como μ para população
    - É o conjunto de todos os resultados, respostas, medidas ou contagens a serem estudados.
    - Censo
      - É o conjunto de dados relativos a todos os elementos de uma população.
- Inferencial
  - Amostragem
    - Conceitos
      - População
      - Amostra: Subconjunto de uma população
      - Amostragem: Processo de Seleção de uma Amostra
      - Erro Amostral: Erro que ocorre pelo uso de uma amostra
      - Parâmetro
      - Estimador
      - Estimativa: Valor numérico determinado pelo estimador
    - Viés
      - Seleção de Dados
      - Coleta de Dados
      - Estimação
    - Tipos de Amostragem
      - Casual Simples
        - Todas os elementos da população tem a mesma probabilidade de serem selecionados
        - $P(X_j = x_i) = \frac{1}{N}$
      - Estratificação
        - Utilizasse uma variável "critério" para dividir a população em estratos
        - Dentro de cada estrato existe uma grande homogeneidade
        - Entre estratos existe uma grande heterogeneidade
      - Conglomerados
        - Utilizasse uma categoria e faz-se o entendimento do comportamento dela.
        - Dentro de cada conglomerado existe uma grande heterogeneidade
        - Entre cada conglomerado existe uma grande homogeneidade
      - Sistemática
        - Considerando uma população de tamanho N, tiramos uma amostra de tamanho n.
        - Fator de estimação: $s = [\frac{N}{n}]$
        - População deve estar ordenada.
  - Estimação de Parâmetros
    - Estimativa Pontual
      - É uma estimativa de um único valor para um parâmetro populacional
      - Média: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
      - Variância: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
      - Proporção
        - $\hat{p} = \frac{x}{n}$
        - x = nº de itens na amostra que apresentam a característica de interesse
    - Estimativa Intervalar ou Intervalo de Confiança
      - Um intervalo de valores possíveis no qual se admite, sob certa probabilidade, esteja contido o parâmetro populacional
      - Grau de Confiança
        - Probabilidade 1 - α de o intervalo de confiança conter o verdadeiro valor do parâmetro populacional
      - Média
        - Variância Populacional Conhecida
          - $\bar{x} \mp z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$
          - Valor Crítico
            - $z_{\frac{\alpha}{2}}$
            - Fronteira que separa as médias amostrais passíveis de ocorrerem
          - Tamanho Populacional Conhecido
            - Se a população é finita é necessário modificar a margem de erro com a inclusão de um fator de correção para população finita
            - Margem de Erro: $E = z \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
            - Tamanho da Amostra
              - $n = \frac{N \cdot \sigma^2 \cdot (z)^2}{(N-1) \cdot E^2 + \sigma^2 \cdot (z)^2}$
              - Quando resultado ser fracionado, considerar o próximo valor inteiro
          - Tamanho Populacional Desconhecido
            - Margem de Erro: $E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
            - Tamanho da Amostra
              - $n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E}\right)^2$
              - Quando resultado ser fracionado, considerar o próximo valor inteiro
        - Variância Populacional Desconhecida
          - $E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$
          - Considera-se as mesmas regras da variância populacional conhecida, porém utiliza-se a tabela T-Student ao invés da de distribuição normal.
          - t-Student
            - Graus de Liberdade: n-1
            - A distribuição t-Student tem média t = 0
            - O desvio padrão da distribuição t-Student varia com o tamanho da amostra
            - Se o tamanho da amostra aumenta, a distribuição t-Student se aproxima da distribuição normal
      - Proporção
        - $\hat{p} = \frac{x}{n}$, sendo que a proporção será considerada 50% caso não se tenha o número de sucessos
        - Validação
          - $n.\hat{p} \geq 5$
          - $n.(1 - \hat{p}) \geq 5$
        - Tamanho Populacional Desconhecido
          - Margem de Erro: $E = (z) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
          - Tamanho Amostra
            - $n = \frac{z_{\frac{\alpha}{2}}^2 \cdot \hat{p}.(1-\hat{p})}{E^2}$
            - Quando resultado ser fracionado, considerar o próximo valor inteiro
        - Tamanho Populacional Conhecido
          - Margem de Erro: $E = (z) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
          - Tamanho Amostra
            - Quando resultado ser fracionado, considerar o próximo valor inteiro
            - $n = \frac{N \cdot \hat{p}(1-\hat{p}) \cdot (z)^2}{\hat{p}(1-\hat{p}) \cdot (z)^2 + (N-1) \cdot E^2}$
  - Testes de Hipóteses
- Probabilidade
  - Conceitos
    - Definições Básicas
      - Experimento Aleatório
        - São experimentos cujos resultados podem ser diferentes, ainda que sejam repetidos sob condições idênticas
        - É possível descrever todos os possíveis resultados.
      - Espaço Amostral
        - O conjunto de todos os possíveis resultados de um experimento aleatório
        - Ex. {Ouro, Paus, Copas, Espadas}
        - É representado pela letra ômega Ω
      - Eventos
        - Subconjunto do espaço amostral
        - São enumeráveis no casos de dados discretos
        - São colocados na forma de intervalo no caso de dados contínuos
        - É representado pela letra A
        - Eventos Equiprováveis
          - São eventos on P(e1) = P(e2)=...=P(en), logo que possuem a mesma probabilidade de ocorrer
          - A probabilidade de cada evento é 1/n.
      - Operações com Eventos Aleatórios
        - Complementação
        - União
        - Intersecção
      - Axiomas (de Kolmogorov)
        - P(A) ≥ 0, para qualquer evento A em Ω
        - P(Ω) = 1
        - Se A e B são dois eventos de Ω e mutuamente exclusivos então: P(A U B) = P(A) + P(B)
    - Clássico
      - Probabilidade que calcula analiticamente as chances de determinado fenômeno acontecer
      - O conceito clássico só pode ser utilizado em situações onde o espaço amostral é enumerável, finito e equiprovável.
      - P(A) = Número de casos favoráveis / Número de casos possíveis
    - Frequentista
      - Probabilidade de um evento como a proporção do número de vezes que o evento ocorre
      - A medida que o número de repetições do experimento cresce, a frequência relativa se aproxima da probabilidade P(A).
      - P(A) = Número de vez que A ocorreu / Número de Experimentos realizados
  - Teoremas
    - Probabilidade Total: Se os eventos A1, A2, A3...An formam uma partição do espaço amostral, então: $\bigcup_{i=1}^{n} A_i = \Omega$
    - Se ø é um conjunto vazio, então P(ø) = 0
    - Evento Complementar: Se Ac é o complemento de A, então P(Ac) = 1 - P(A)
    - Se A e B são dois eventos quaisquer e Bc é o complemento de B, então: P(A ∩ Bc) = P(A) - P(A ∩ B)
    - Soma das Probilidades: Se A e B são dois eventos quaisquer, então: P(A U B) = P(A) + P(B) - P(A ∩ B)
    - Se A ⊂ (é subconjunto) B, então P(A) ≤ P(B)
    - Para um evento A qualquer: 0 ≤ P(A) ≤ 1
    - Produto das Probabilidades
      - P(A ∩ B) = P(A|B).P(B)
      - P(A ∩ B ∩ C) = P[A|(B ∩ C)].P(B|C).P(C)
    - Teorema Bayes
      - As probabilidades condicionais P(A|B) e P(B|A) expressam diferentes valores de probabilidades
      - $P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$
      - onde: P(A|B).P(B) = P(A ∩ B) = P(B ∩ A) = P(B|A).P(A)
      - $P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$
  - Probabilidade Condicional
    - Definimos como a probabilidade de A ocorrer dado que B ocorre.
    - P(A|B) = P(A ∩ B)/P(B), onde P(B) ≠ 0
  - Independência Estocástica ou Eventos Independentes
    - P(A|B) = P(A) ou P(B|A) = P(B)
    - Dois eventos A e B são ditos independentes se: P(A ∩ B) = P(A).P(B)
    - Probabilidade de um evento ocorrer sem que dependa do fato de os outros eventos terem ou não ocorrido
- Variáveis Aleatórias
  - Definições
    - Em matemática, é definida como uma função que associa a todo evento pertencente a uma partição do espaço amostral Ω a um único número real (R)
    - É comum a representação das variáveis aleatórias por letras maiúsculas e dos valores assumidos por letras minúsculas
    - O uso da variável aleatória permite descrever resultados de um experimento aleatório por meio de número ao invés de palavras
    - $X : \Omega \to \mathbb{R} \Leftrightarrow \{\omega : X(\omega) \leq x\} \in \mathcal{F}, \; \forall x \in \mathbb{R}$
  - Discretas
    - Assume valores que podem ser contados
    - Funções
      - Univariada
        - Unidimensional
        - Função de massa de probabilidade
          - $P(x_i) \geq 0$ para todo x_i
          - $\sum_i P(x_i) = 1$
      - Multivariada
        - Vetor Aleatório
        - Função de distribuição acumulada
          - F(x) = P(X ≤ x)
          - F(x) = $\sum_{x_i \leq x} p(x_i)$
    - Fórmulas Importantes
      - Variância
        - Var(x) = $\sum_{i=1}^{n} x_i^2 \cdot p(x_i) - E(X)^2$
        - Propriedades
          - A variância também não possui a propriedade de linearidade do valor médio, isto é, Var(aX+b)≠aVar(X)+b. Em vez disso, teremos Var(aX+b)=a²Var(X)
          - Se c é uma constante, então Var(X+c) = Var(X)
          - Suponha-se que c seja uma constante e X seja uma variável aleatória, então Var(cX) = c²Var(X)
          - Se (X , Y) for uma variável aleatória bidimensional, e se X e Y forem independentes. Então, Var(X+Y) = Var(X) + Var(Y)
      - Média ou Esperança
        - $E(X) = \sum_{i=1}^{n} x_i \cdot p(x_i)$
        - Propriedades
          - Se X = c, onde c é uma constante, então E(X) = c
          - Suponha-se que c seja uma constante e X seja uma variável aleatória. Então E(cX) = cE(X)
          - Sejam X e Y duas variáveis aleatórias quaisquer. Então, E(X+Y) = E(X) + E(Y)
          - Para funções lineares, se Y = aX + b, onde a e b são constantes, então E(Y) = aE(X) + b
    - Modelos Probabilísticos
      - Bernoulli
        - Probabilidade de Sucesso ou Fracasso
        - Ex. Acertar chute ao gol
        - Função: P(sucesso) = p
        - Média: p
        - Variância: p(1-p)
      - Binomial
        - Nro de sucessos (p) dentre n repetições de um experimento
        - Acertar 2 de 5 chutes ao gol
        - Função: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
        - Média: n.p
        - Variância: n.p(1-p)
      - Binomial Negativa
        - Nro de tentativas (k) para acontecer o r-ésimo sucesso
        - Acertar 2 chutes na tentativa 4
        - Função: $P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$
        - Média: $\frac{r}{p}$
        - Variância: $\frac{r(1-p)}{p^2}$
      - Geométrica
        - Nro de tentativas (k) até acontecer o primeiro sucesso
        - Acertar o primeiro chute na terceira tentativa
        - Função: $P(X = k) = p(1-p)^{k-1}$
        - Média: $\frac{1}{p}$
        - Variância: $\frac{1-p}{p^2}$
      - HiperGeométrica
        - Nro. de elementos (k) de uma amostra com n elementos, obtida sem reposição de um universo de N elementos com r elementos com determinada característica
        - Ter 3 celulares com defeito de uma amostra de 10 celulares retirados sem reposição de um lote de 50 celulares que tinham 10 com defeito
        - Função: $P(X = k) = \frac{\binom{r}{k} \cdot \binom{N-r}{n-k}}{\binom{N}{n}}$
        - Média: np
        - Variância: $np(1-p)\frac{N-n}{N-1}$
      - Poisson
        - Nro. de sucessos (k) dentro de um intervalo de tempo, volume, espaço
        - Um callcenter receber 3 chamas em 1 minuto
        - Função: $P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$
        - Média: λ
        - Variância: λ
  - Contínuas
    - Assume qualquer valor dentro de um intervalo, ou seja, se houver um número incontável de resultados possíveis
    - Funções
      - Univariada
        - Função de massa de densidade de probabilidade
          - f(x) ≥ 0, para todo x
          - A área definida por f(x) é igual a 1, ou seja, $\int_{-\infty}^{+\infty} f(x)dx = 1$
          - Para quaisquer a e b, com -∞ < a < b < +∞, teremos $P(a \leq X \leq b) = \int_a^b f(x)dx$
        - Unidimensional
      - Multivariada
        - Função densidade de probabilidade conjunta: $F(x) = \int_{-\infty}^{x} f(t)\, dt$
        - Unidimensional
    - Fórmulas Importantes
      - Média ou Esperança
        - $E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$
        - Propriedades
      - Variância
        - Var(X) = $\int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx - [E(X)]^2$
        - Propriedades
    - Modelos Probabilísticos
      - Uniforme e Contínuo
        - Ex: A probabilidade de um ponto estar entre 1 e 1,5 em um intervalo de 0 a 2.
        - Funções
          - Função de Probabilidade: $f(x) = \begin{cases} \frac{1}{b-a}, & se\ a \leq x \leq b \\ 0, & caso\ contrário \end{cases}$
          - Função de Distribuição: $F(x) = \begin{cases} 0, & se\ x \leq a \\ \frac{x-a}{b-a}, & se\ a \leq x \leq b \\ 1, & se\ x \geq b \end{cases}$
        - Média: $\frac{a+b}{2}$
        - Variância: $\frac{(b-a)^2}{12}$
      - Distribuição Exponencial
        - Probabilidade entre o intervalo de ligações em um Call Center ficar entre 2 e 3 minutos
        - Funções
          - Função de Probabilidade: $f(x) = \begin{cases} 0, & se\ x \leq 0 \\ \lambda e^{-\lambda x}, & se\ x \geq 0 \end{cases}$
          - Função de Distribuição: $F(x) = \begin{cases} 0, & se\ x \leq 0 \\ 1 - e^{-\lambda x}, & se\ x \geq 0 \end{cases}$
        - Média: $\frac{1}{\lambda}$
        - Variância: $\frac{1}{\lambda^2}$
      - Distribuição Normal ou Gaussiana
        - É uma distribuição simétrica em torno da média
        - -1 σ a 1 σ (Sigma) = 68%
        - -2 σ a 2 σ (Sigma) = 95,44%
        - -3 σ a 3 σ (Sigma) = 99,73%
        - -4 σ a 4 σ (Sigma) = 99,994%
        - Função de Probabilidade: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$
        - Função de Distribuição: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$
        - Padronização
          - Fórmula: $z = \frac{x - \mu}{\sigma}$
          - Probabilidade
            - Probabilidade P(Z ≤ z): use esse valor direto
            - Probabilidade P(0 ≤ Z ≤ z): some 0,5 para obter P(Z ≤ z)
            - Probabilidade P(Z > z): faça 1 - P(Z > z) para obter P(Z ≤ z)
          - Tabela
        - Teorema Central do Limite
          - Esse teorema afirma que quando o tamanho da amostra aumenta, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal.
      - Gama
      - Weibull
      - Beta
      - Lognormal