Cientista de Dados

Aula 5 - Amostragem e Distribuições - Parte 2

Profa. Dra. Deborah Fernandes (INF/UFG)
Prof. Msc. Márcio Giovane C. Fernandes (UEG)

Programa de Residência em TI

Cientista de Dados

Contextualização

Amostragem

Distribuições estatísticas

Distribuições em Ciência de Dados

- Gaussiana
- Binomial
- Poisson
- Exponencial
- Qui-quadrado, T-Student, F

1 - Contextualização

Desafios do mundo real

- Por que não analisamos sempre a **população inteira**? (custos, tempo, inviabilidade técnica).
  - Não temos os dados de toda a população
  - Não possuímos os logs de acessos de uma rede social
  - Alto custo
  - Tempo para coleta e organização
  - Inviabilidade técnica

1 - Contextualização

Desafios do mundo real

- Papel da amostragem em **Big Data** e em situações com **dados limitados**.
  - Custo e tempo para processar 100% dos dados

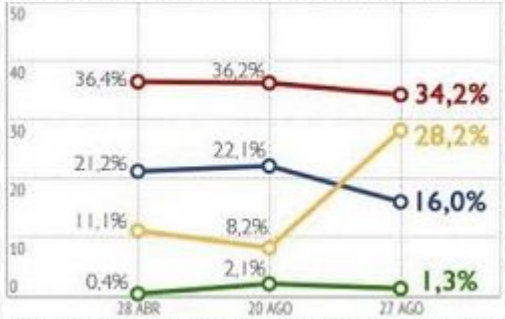
- “Se você tivesse 100 milhões de registros, analisaria todos ou faria uma amostra?”

1 - Contextualização

O papel da amostragem

- A amostragem é a base da **inferência estatística**
  - permite tirar conclusões a partir de um subconjunto
  - extrapolação da amostra para a população.

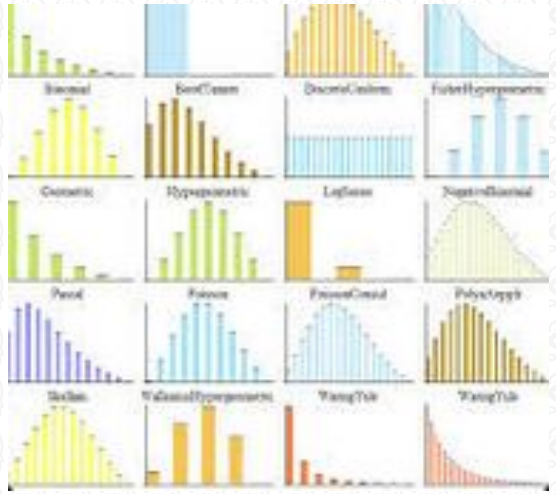
📌 **Exemplos**
- pesquisas eleitorais
- controle de qualidade
- treino de modelos de *machine learning*



1 - Contextualização

O porquê das distribuições

- Os dados não se comportam de forma aleatória, sem padrão, eles seguem distribuições
- Entender distribuições nos permite:
  - Prever a probabilidade de eventos
  - Modelar incertezas
  - Escolher técnicas estatísticas adequadas



📌 **Exemplos**

- Tempo de resposta de um sistema web - geralmente segue uma distribuição exponencial
- Altura das pessoas segue uma distribuição normal

1 - Contextualização

O trabalho do cientista de dados

- Sem compreender amostragem e distribuição:
  - Não há como fazer inferência estatística confiável;
  - Não conseguimos avaliar a significância dos resultados
  - Corremos o risco de generalizar conclusões erradas.

"Na ciência de dados, não trabalhamos apenas com dados; trabalhamos com a incerteza dos dados."

2 - Amostragem

Definição: seleção de subconjuntos representativos da população.

Tipos de Amostragem probabilística:

1. **Aleatória simples** – cada elemento tem igual chance de ser escolhido.
   a. Exemplo: Sortear clientes de uma base com `random.sample()`.

2 - Amostragem

Definição: seleção de subconjuntos representativos da população.

Tipos de Amostragem probabilística:

2. Sistemática – seleção em intervalos fixos.

a. Exemplo: A cada 10 registros
b. Boa para bases muito grandes
c. Risco: se os dados tiverem padrão periódico, pode enviesar.

2 - Amostragem

Tipos de Amostragem probabilística:

3. **Estratificada** – divisão em subgrupos (estratos) e amostragem proporcional.
   a. Exemplo: Em uma pesquisa de renda, estratificar por região ou faixa etária
   b. Muito usada quando há desbalanceamento de classes.

4. **Por conglomerados** – sorteio de grupos inteiros ao invés de indivíduos
   a. Exemplo: Escolher cidades inteiras para pesquisar, em vez de indivíduos aleatórios.

2 - Amostragem

Tipos de Amostragem não probabilísticas:

Quando não temos probabilidade conhecida; é mais simples e pode gerar viés.

1. **Conveniente**
   a. Escolhe apenas quem está acessível
   b. Exemplo: pedir respostas a alunos presentes em sala
2. **Bola de Neve**
   a. Um participante indica outro
   b. Útil em populações difíceis de alcançar
3. **Intencional / Julgamento**
   a. Escolha feita com base em conhecimento prévio do pesquisador;

2 - Amostragem

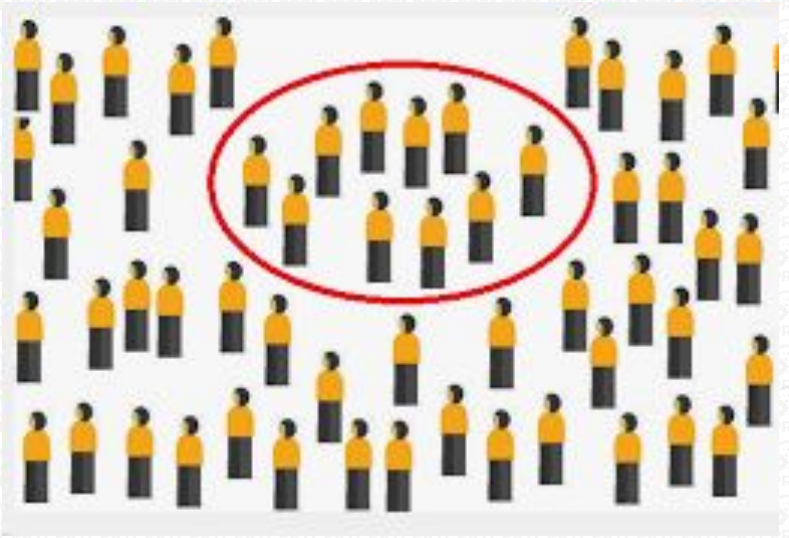
Tamanho da amostra

- Quanto maior a amostra, maior a precisão, maior o custo;
- Lei dos grandes números: quanto maior a amostra, mais próxima a média amostral da média populacional

Aplicações em Ciência de Dados

- Amostragem para treino/teste e validação em ML
- Técnicas de *undersampling/oversampling* em desbalanceamento de classes.

2 - Amostragem



Como calcular o tamanho de uma amostra?

1. Parâmetros que influenciam o cálculo

- **N** → Tamanho da população (se conhecida).
- **Z** → Valor crítico da distribuição normal (depende do nível de confiança).
  - 90% → Z ≈ 1,64
  - 95% → Z ≈ 1,96
  - 99% → Z ≈ 2,58

- **p** → Proporção esperada da característica de interesse (se não se sabe, usa-se p=0,5 → cenário mais conservador).
- **q** = 1 − p.
- **E** → Margem de erro tolerada (ex.: 5% = 0,05).

2 - Amostragem

Como calcular o tamanho de uma amostra?

2. Fórmula geral para população infinita

Quando a população é muito grande (Big Data, redes sociais etc.):

$$n = \frac{Z^2 \cdot p \cdot q}{E^2}$$

3. Correção para população finita

Se a população é conhecida (N):

$$n = \frac{N \cdot Z^2 \cdot p \cdot q}{E^2 \cdot (N-1) + Z^2 \cdot p \cdot q}$$

2 - Amostragem

Como calcular o tamanho de uma amostra?

📌 **Exemplo prático:**

Imagine que você queira estimar a proporção de pessoas que usam IA generativa em uma população de **10.000 alunos universitários**, com:

- Nível de confiança = 95% → Z = 1,96
- Margem de erro = 5% → E = 0,05
- Proporção esperada = 50% (p = 0,5; q = 0,5)
  Substituindo:

$$n = \frac{10000 \cdot (1,96^2 \cdot 0,5 \cdot 0,5)}{0,05^2 \cdot (10000 - 1) + (1,96^2 \cdot 0,5 \cdot 0,5)}$$

$$n \approx 370$$

Ou seja, **370 pessoas** já seriam suficientes para inferir sobre os 10.000 com 95% de confiança

e 5% de margem de erro.

2 - Amostragem

Definição: seleção de subconjuntos representativos da população.

Prática com Python

- **Extrair amostras com** (`pandas.sample()` e `sklearn.model_selection.train_test_split`).

3 - Distribuições Estatísticas

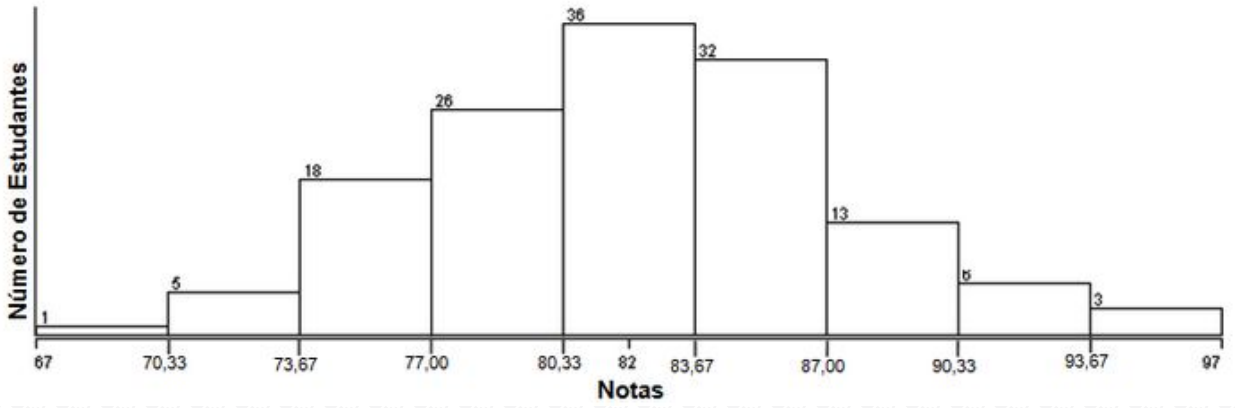
O que é uma distribuição?

Representa como os valores de uma variável estão organizados ao longo de um intervalo. Pode ser apresentada como:

a. **Distribuição empírica** → obtida a partir dos dados (histogramas, gráficos de densidade).

b. **Distribuição teórica** → modelo matemático que descreve o comportamento esperado (ex.: Normal, Binomial, Poisson).

📌 **Exemplo: Distribuição de notas de estudantes**



TI

3 - Distribuições Estatísticas

O que é uma distribuição amostral?

- Quando extraímos várias amostras da mesma população, os **valores das médias** dessas amostras também formam uma distribuição.

📌 **Exemplo:**

Tirar várias amostras de tamanho 30 de um dataset → calcular as médias → plotar histograma das médias.

3 - Distribuições Estatísticas

Lei dos Grandes Números

- Quanto maior o tamanho da amostra, mais a média amostral se aproxima da média populacional.

- Fundamenta a ideia de que aumentar a amostra reduz a margem de erro.

📌 **Exemplo:**

Comparar médias calculadas com amostras de tamanho 10, 100 e 1.000.

3 - Distribuições Estatísticas

3.4 Teorema Central do Limite (TCL)

- Pilar da estatística: **independentemente da distribuição da população**, a distribuição das médias amostrais tende a ser **Normal** quando n é suficientemente grande (tipicamente n > 30).
  - médias de amostras tendem a seguir uma **Normal**, mesmo que a população não seja normal.

- Justifica o uso da Normal em muitos métodos estatísticos e algoritmos de ML.

📌 **Exemplo prático com simulação:**

Gerar 1.000 amostras de uma **distribuição Exponencial** → calcular a média de cada amostra → mostrar que a curva das médias tende para a Normal.

3.4 Teorema Central do Limite (TCL)- Exemplo

Teoria do Limite Central (TLC)

Solução do Problema com o TLC

1.Descrição do Cenário:

1. O valor das diárias de um servidor pode ter uma distribuição desconhecida ou assimétrica.
2. Queremos prever o comportamento da média amostral de valores de diárias ao selecionar amostras aleatórias de diferentes tamanhos.

2.Aplicação do TLC:

1. Se retirarmos várias amostras aleatórias de tamanhos n (ex.: 10, 30, 50) e calcularmos a média de cada amostra, as médias amostrais formarão uma distribuição que será aproximadamente normal, independentemente da forma original dos dados.

3.Experimento Prático:

1. Extrairemos amostras aleatórias de diferentes tamanhos (n = 50,70,90). Para cada tamanho, extrairemos 1000 amostras (1000 amostras de tamanho 50, 1000 amostras de tamanho 70, 1000 amostras de tamanho 90).
2. Calcularemos as médias de cada amostra.
3. Verificaremos se as médias amostrais seguem uma distribuição normal conforme o tamanho da amostra aumenta.

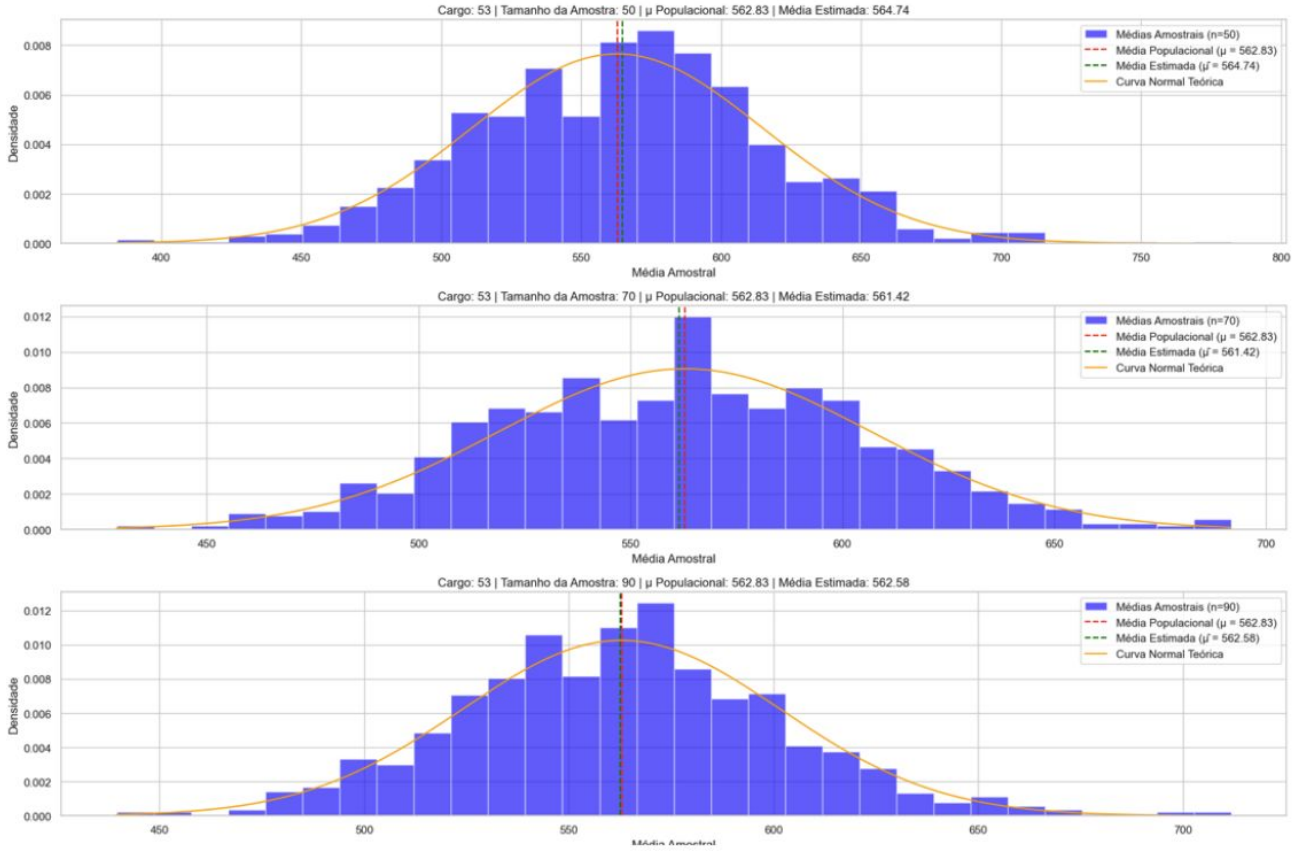
3.4 Teorema Central do Limite (TCL)- Exemplo

```
Comparação entre Média Populacional e Estimada, e Desvios Padrão Populacional e Estimado:
Média Populacional (μ): 562.83
Desvio Padrão Populacional (σ): 368.89

Tamanho da Amostra: 50
- Média Populacional (μ): 562.83
- Média Estimada (μ̂): 564.74
- Diferença Absoluta da Média: 1.91
- Diferença Relativa da Média: 0.34%
- Desvio Padrão Populacional (σ): 368.89
- Desvio Padrão Estimado (s): 368.42
- Diferença Absoluta do Desvio Padrão: 0.47
- Diferença Relativa do Desvio Padrão: 0.13%

Tamanho da Amostra: 70
- Média Populacional (μ): 562.83
- Média Estimada (μ̂): 561.42
- Diferença Absoluta da Média: 1.41
- Diferença Relativa da Média: 0.25%
- Desvio Padrão Populacional (σ): 368.89
- Desvio Padrão Estimado (s): 368.02
- Diferença Absoluta do Desvio Padrão: 0.87
- Diferença Relativa do Desvio Padrão: 0.24%

Tamanho da Amostra: 90
- Média Populacional (μ): 562.83
- Média Estimada (μ̂): 562.58
- Diferença Absoluta da Média: 0.25
- Diferença Relativa da Média: 0.05%
- Desvio Padrão Populacional (σ): 368.89
- Desvio Padrão Estimado (s): 368.78
- Diferença Absoluta do Desvio Padrão: 0.11
- Diferença Relativa do Desvio Padrão: 0.03%
```



3 - Distribuições Estatísticas

Resumindo ...

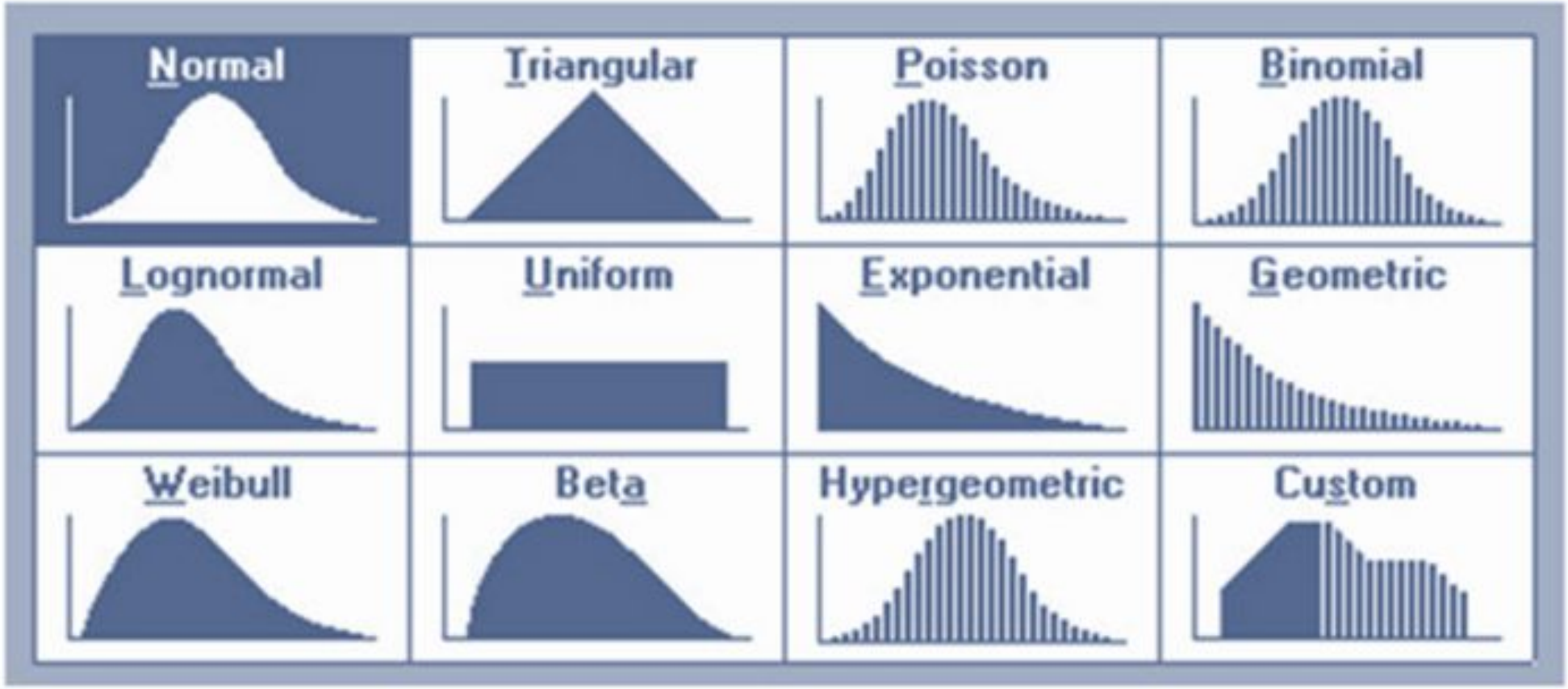
São o modelo matemático que usamos para entender o comportamento dos dados.

- **Distribuição de frequência** (histogramas, gráficos de densidade).
- **Distribuição amostral**: a média ou proporção de uma amostra segue uma distribuição.
- **Lei dos Grandes Números** (quanto maior a amostra, mais próxima da média populacional).
- **Teorema Central do Limite (TCL)**: médias de amostras tendem a seguir uma **Normal**, mesmo que a população não seja normal.

4 - Distribuições em Ciência de Dados

1. **Normal (Gaussiana)** – base de muitos algoritmos.
2. **Binomial** – experimentos de sucesso/fracasso.
3. **Poisson** – eventos raros (ex.: falhas em sistemas, acessos por minuto).
4. **Exponencial** – tempo entre eventos.
5. **Qui-quadrado, t-*Student*, F** – fundamentais para testes de hipóteses.
6. Outras

4 - Distribuições mais usadas em Ciência de Dados



Fonte: https://ucreanop.com/wp-content/uploads/2021/08/06-Medir.pdf

4 - Distribuições mais usadas em Ciência de Dados

4.1 Distribuição Normal (Gaussiana)

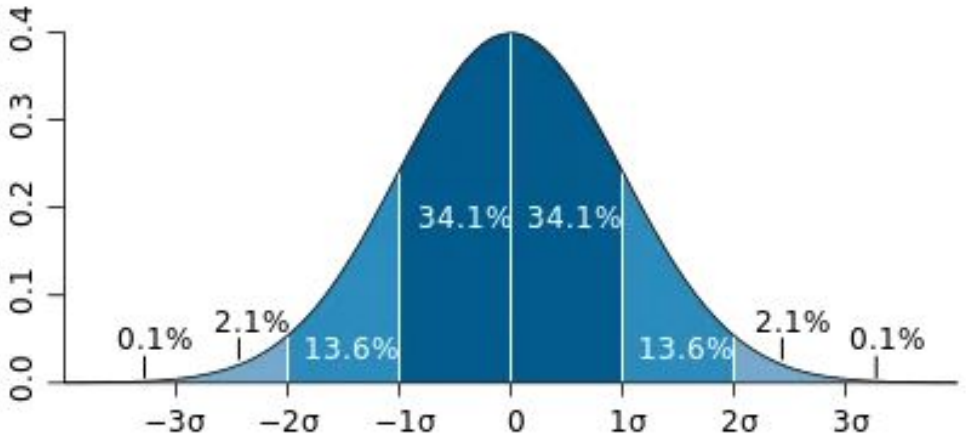
- É uma distribuição de probabilidade contínua e simétrica que representa o comportamento de um fenômeno natural de forma aleatória.
- A curva de distribuição normal representa o comportamento de diversos processos nas empresas e muitos fenômenos comuns, como:
  - altura ou peso de uma população,
  - a pressão sanguínea de um grupo de pessoas,
  - o tempo que estudantes gastam em uma prova.
- Nela, a **média, mediana e moda** dos dados possuem o mesmo valor.

fonte: https://voitto.com.br/blog/artigo/distribuicao-normal

4 - Distribuições em Ciência de Dados

4.1 Distribuição Normal (Gaussiana)

- **Características:** simétrica, forma de sino, definida por média (μ) e desvio padrão (σ).
- **Importância:** Representa a maior parte dos fenômenos naturais
- **Aplicações em *Data Science*:**
  - Detecção de *outliers* (são os valores que ficam **muito distantes da média**, fora da região onde esperamos que a maior parte dos dados esteja, valores que caem fora de **±3σ da média**)
  - Processos de controle de qualidade.



https://www.inf.ufsc.br/~andre.zibetti/probabilidade/normal.html
https://www.blog.psicometriaonline.com.br/distribuicao-normal/
https://geokrigagem.com.br/distribuicao-normal-o-que-e-e-sua-grande-importancia-na-estatistica/
https://www.lampada.uerj.br/arquivosdb/_book/estimadores.html

4.1 Distribuição Normal - Exemplo: Gaussiana no controle de qualidade

(1) Hipótese central:

Em muitos processos industriais e de serviços, as **variáveis de interesse** (ex.: diâmetro de uma peça, tempo de resposta de um sistema, concentração de um reagente) tendem a seguir uma **distribuição Normal**, por causa da soma de pequenas variações aleatórias (ruído do processo, tolerâncias de máquina, etc).

(2) Definir média e desvio padrão do processo

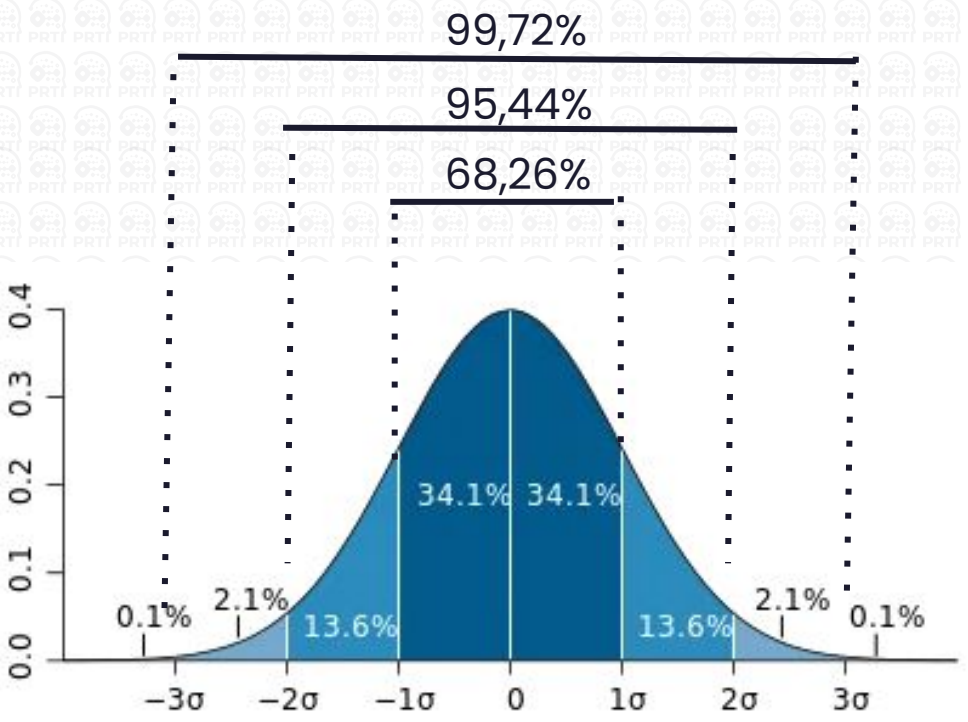
- Mede-se a **média (μ)** do processo → valor central esperado.
- Mede-se o **desvio padrão (σ)** → indica a variação natural.

📌 Exemplo:

O diâmetro esperado de um parafuso é **10 mm** com σ = 0,1 mm.

(3) Definir limites de controle

- Normalmente, limites de **±3σ** em torno da média são usados (baseado na **regra 68-95-99,7**).
- Se o processo estiver **sob controle**, espera-se que 99,7% das peças estejam dentro desse intervalo.



📌 Exemplo:

- ❏ Limite inferior: 10 – 0,3 = **9,7 mm**
- ❏ Limite superior: 10 + 0,3 = **10,3 mm**

4.1 Distribuição Normal - Exemplo: Gaussiana no controle de qualidade

(4) Monitorar amostras (gráficos de controle)

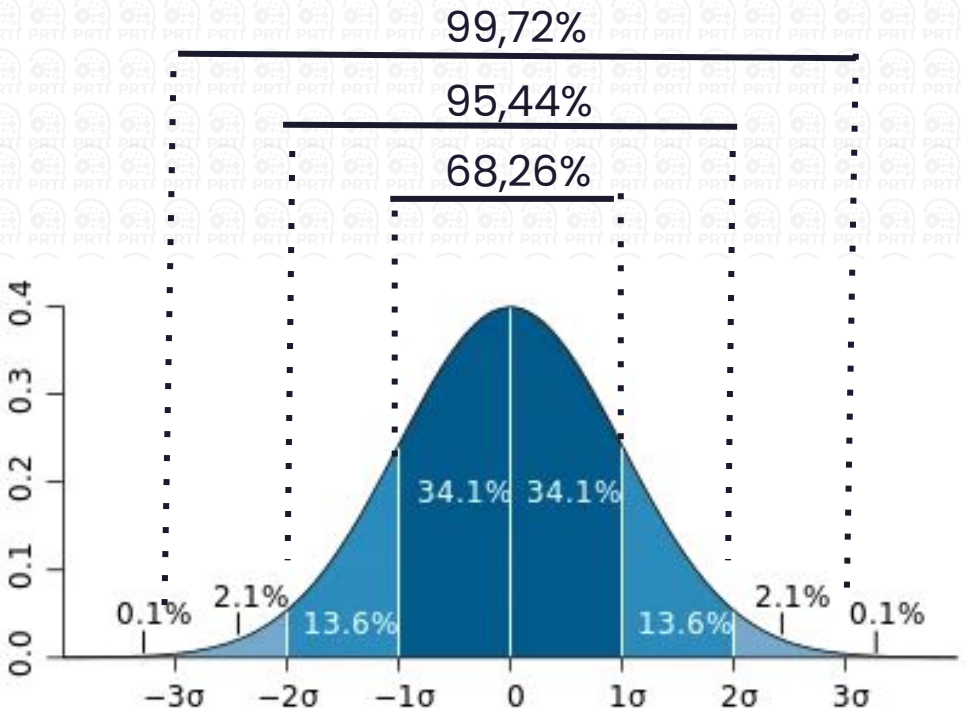
- Seleciona-se amostras do processo ao longo do tempo.
- Calcula-se a média e plota-se em um **gráfico de controle**.
- Se os pontos caem fora dos limites (±3σ) → indicam **anomalias ou falhas no processo**.

  📌 Exemplo:

  Se algumas peças aparecem com 9,4 mm → fora de 3σ → o processo precisa ser revisado (máquina desregulada, matéria-prima com defeito).

(5) Interpretação prática

- Valores dentro dos limites → variação natural do processo (ruído aceitável).
- Valores fora dos limites → variação anormal (sinal de falha, erro, desajuste).

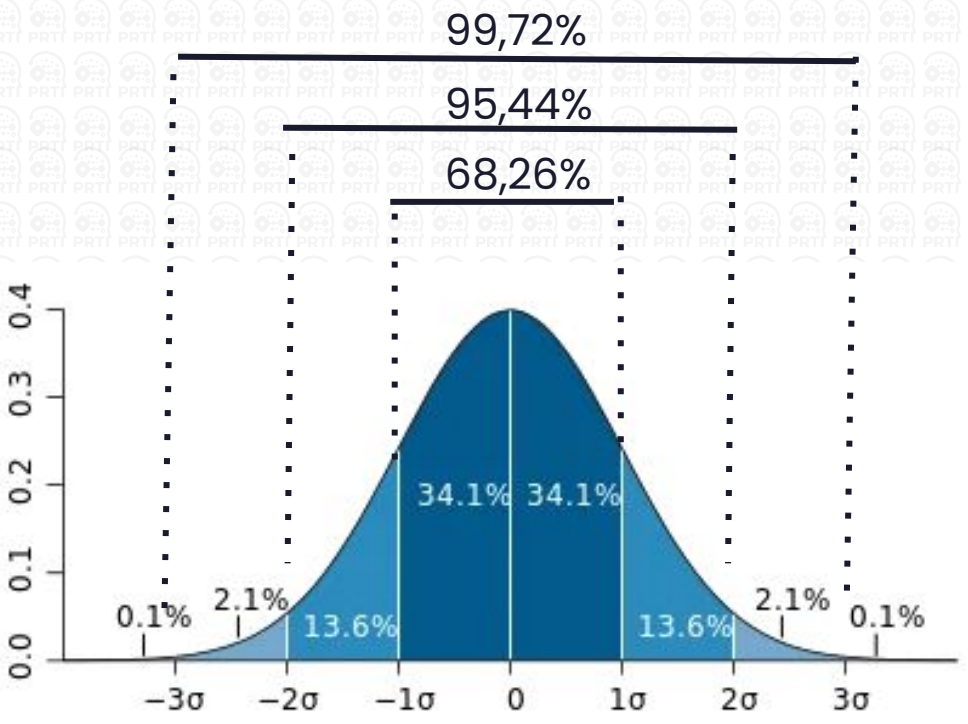


4.1 Distribuição Normal - Exemplo: Gaussiana no controle de qualidade

Aplicações práticas

- **Indústria**: controle de medidas, pesos, resistência de materiais.
- **Saúde**: monitoramento de exames (ex.: níveis de glicose em pacientes).
- **TI**: tempos de resposta de servidores (detecção de picos anômalos).
- **Finanças**: detecção de operações fora do padrão (ex.: riscos de fraude).

Resumindo: A Gaussiana permite definir **limites de controle (±3σ)**. Se o processo for estável, quase todos os valores ficam dentro desses limites. Quando pontos caem fora, temos **outliers que indicam problemas reais no processo** → é assim que a estatística ajuda a manter a qualidade.



4 - Distribuições mais usadas em Ciência de Dados

4.1 Distribuição Normal (Gaussiana)

- **Cálculo em Python**
  - biblioteca `scipy.stats` e a função `norm`.
  - Geração de variáveis aleatórias (`norm.rvs`);
  - Cálculo da probabilidade de um valor (norm.pdf) e a probabilidade acumulada (`norm.cdf`)
  - Definição de intervalos de confiança (`norm.interval`)

Resumão

Distribuição	Tipo	Parâmetros	Forma / Característica	Aplicações em Ciência de Dados
Normal (Gaussiana)	Contínua	Média (μ), Desvio padrão (σ)	Curva simétrica em sino, centrada em μ	Modelagem de erros, regressão (permite **prever valores futuros)**, z-score(é uma régua universal: mede quão “longe” um valor está do normal esperado), detecção de outliers, testes de hipótese
Binomial	Discreta	n (número de tentativas), p (probabilidade de sucesso)	Conta sucessos em n tentativas; discreta e finita	Modelar cliques em campanhas (CTR), testes de software, classificação binária
Poisson	Discreta	λ (taxa média de eventos)	Eventos raros em intervalo fixo; valores ≥ 0, sem limite superior	Modelar acessos por minuto em sites, falhas em sistemas, tráfego de rede
Qui-quadrado (χ^2)	Contínua (não simétrica)	Graus de liberdade (k)	Assimétrica à direita; soma de quadrados de variáveis normais	Testes de independência (tabelas de contingência), ajuste de modelos, variabilidade
t-Student	Contínua	Graus de liberdade (n-1)	Parecida com a Normal, mas caudas mais pesadas (mais c	Muito usada para verificar se um modelo explica significativamente mais variabilidade que outro.

4 - Distribuições mais usadas em Ciência de Dados

4.2 Distribuição Binomial

- Modelo de probabilidade discreta.
- **Parâmetros:** n (número de tentativas), p (probabilidade de sucesso).
- **Situação:** É utilizado para o cálculo da probabilidade de ocorrer um número específico de "sucessos" em um conjunto fixo ***n*** de "tentativas" independentes, onde cada tentativa tem apenas dois resultados possíveis (sucesso ou fracasso) e a probabilidade ***p*** de sucesso é a mesma em todas elas.
- **Aplicações:**
  - Testar taxas de clique (CTR) em campanhas digitais.
  - Avaliar número de aprovações em testes de software.

4 - Distribuições mais usadas em Ciência de Dados

4.2 Distribuição Binomial

- **Principais propriedades:**
  - É discreto, e pode tomar valores de 0 a n, onde n é o tamanho da amostra
- **Cálculo em Python**
  - Biblioteca `scipy.stats` e função `binom`.
  - Comando `binom.rvs(n, p, size)`, gera variáveis aleatórias que seguem uma distribuição binomial, onde `n` é o número de tentativas, `p` é a probabilidade de sucesso em cada tentativa, e `size` define o número de amostras a serem geradas.

4 - Distribuições mais usadas em Ciência de Dados

4.2 Distribuição Binomial

Cálculo da probabilidade binomial

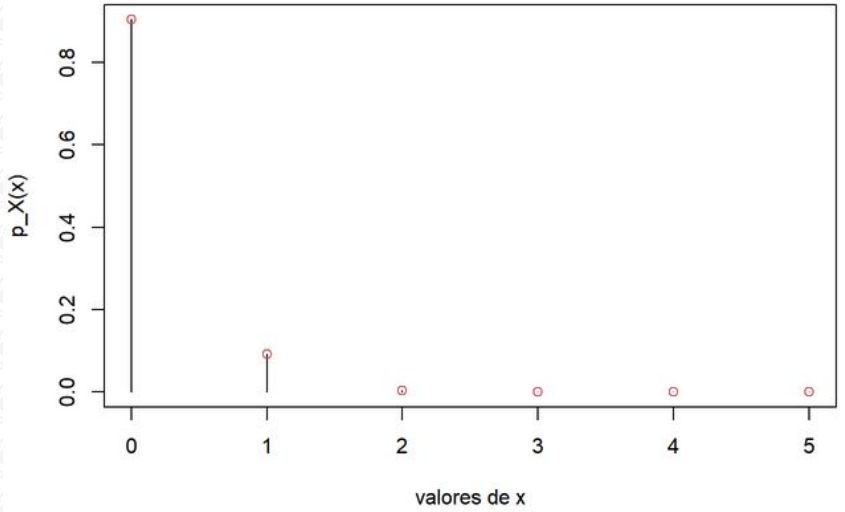
Exemplo Binomial 1:

Baseado em estudos anteriores, a probabilidade de um certo componente elétrico estar em condições operacionais satisfatórias é de 0.98. Os componentes são amostrados item por item, a partir de uma produção (contínua). Em uma amostra de cinco componentes, quais são as probabilidades de se encontrarem,

- a. zero
- b. exatamente um
- c. exatamente dois
- d. dois ou mais
- e. ao menos quatro, itens defeituosos?

Resposta Exemplo B1:

Os requisitos para a aplicação do modelo binomial foram satisfeitos. $n = 5, P(defeituoso) = 0.02$ Assumiremos como $p = 0.02$ a probabilidade de encontrarmos um item defeituoso. Aplicando o modelo probabilístico binomial para responder as questões temos:




```
           0           1           2           3           4           5
0.9039207968 0.0922368160 0.0037647680 0.0000768320 0.0000007840 0.0000000032
```

Fonte: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/binomial.html

4 - Distribuições mais usadas em Ciência de Dados

4.3 Distribuição Poisson

- Modelo de probabilidade discreta que calcula **a probabilidade de um número específico de ocorrências (k) de um evento em um intervalo fixo de tempo ou espaço,** quando os eventos ocorrem aleatoriamente, de forma independente e com uma taxa média (λ) constante.
- Quando usar?
  - Ela é usada para eventos raros,
  - Os eventos são independentes um do outro.
  - A taxa média de ocorrência (λ) não muda durante o intervalo de tempo ou espaço em questão.
  - Você está contando o número de vezes que um evento ocorre em um intervalo fixo (contagens inteiras não negativas).

4 - Distribuições mais usadas em Ciência de Dados

4.3 Distribuição Poisson

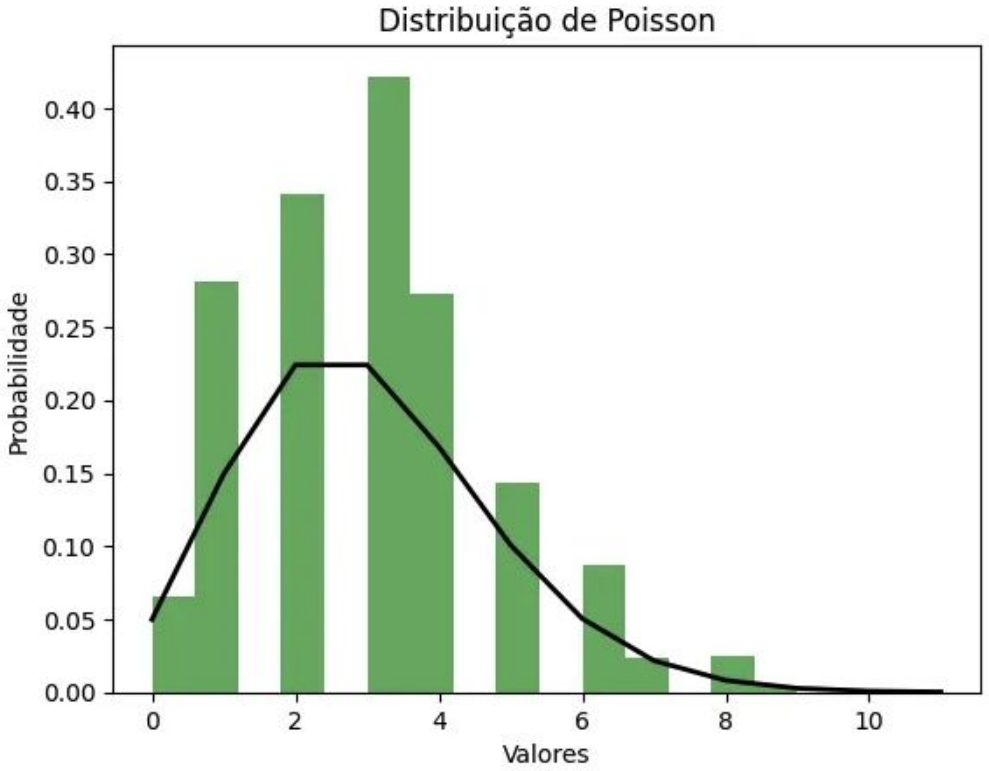
Exemplos de Aplicação

- O número de emails que recebe em uma hora.
- O número de erros de digitação em uma página de livro.
- O número de acidentes de trânsito em um determinado cruzamento ao longo de um mês
- número de defeitos em um produto,
- número chamadas para uma central de atendimento ou
- o número de clientes numa loja por hora

4 - Distribuições mais usadas em Ciência de Dados

4.3 Distribuição Poisson

- `numpy.random` para gerar amostras aleatórias;
- `scipy.stats` para cálculos de probabilidade
- No NumPy `np.random.poisson(lam, size)`,
- No SciPy `scipy.stats.poisson.rvs(mu, size)`
- `lam` ou `mu` representam o parâmetro lambda , a taxa de eventos.



4 - Distribuições mais usadas em Ciência de Dados

4.4 Distribuição Exponencial

- Distribuição de probabilidade contínua usada para modelar o tempo até que um evento ocorra
- Distribuição contínua que se aplica a variáveis com valores não negativos (≥ 0)
- É frequentemente associada ao processo de Poisson, que modela a ocorrência de eventos em intervalos

Parâmetro de taxa (λ):

- Define a taxa de ocorrência dos eventos, ou seja, quantos eventos se esperam por unidade de tempo ou espaço. Quanto maior λ, mais rápidos são os eventos.

4 - Distribuições mais usadas em Ciência de Dados

4.4 Distribuição Exponencial

Propriedade: Falta de memória**,** significa que o tempo que ainda falta para o próximo evento ocorrer não depende de quanto tempo já passou.

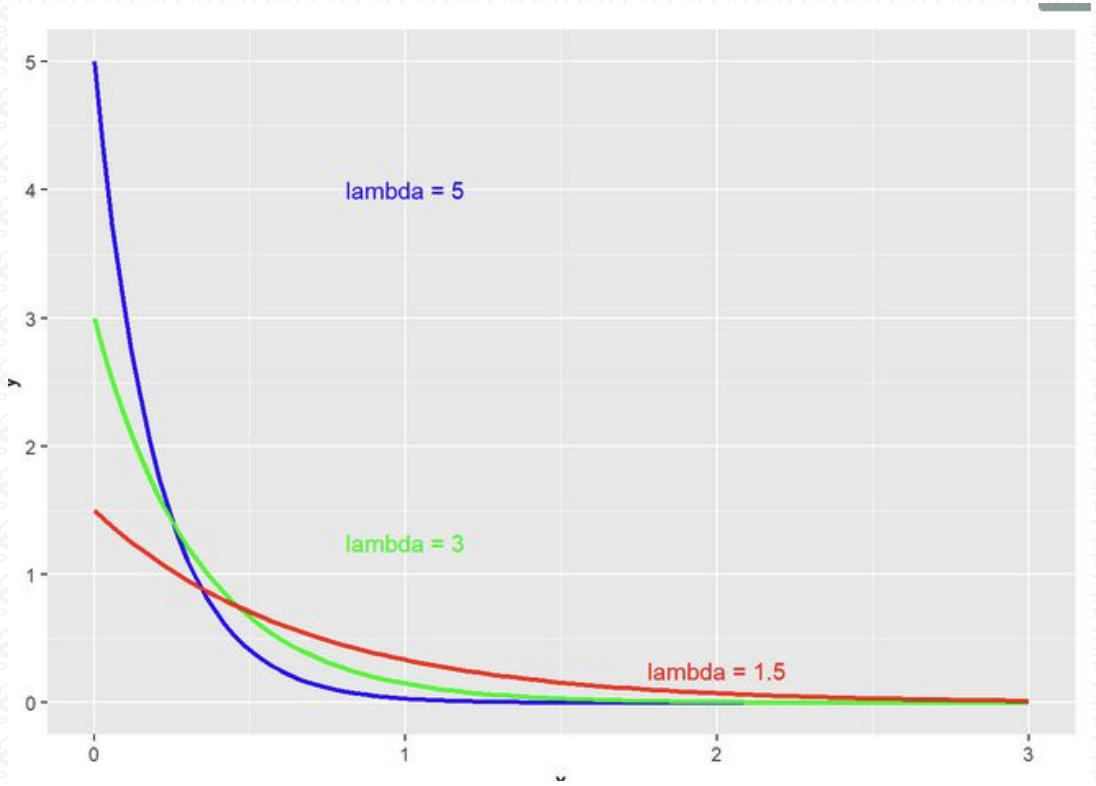
Aplicações

- **Confiabilidade de sistemas:** Prever a vida útil de equipamentos eletrônicos ou mecânicos, como a probabilidade de um celular continuar funcionando após um certo tempo.
- **Processos de atendimento:** Estimar o tempo entre a chegada de clientes a um serviço de atendimento.
- **Engenharia:** Avaliar o tempo até o recebimento de uma peça defeituosa numa linha de produção.

4 - Distribuições mais usadas em Ciência de Dados

4.4 Distribuição Exponencial

- Seu principal uso é modelar o tempo entre ocorrências de eventos independentes que acontecem a uma taxa média constante.
- Exemplo: Considere um serviço de atendimento, a distribuição exponencial pode descrever o tempo que um cliente espera antes de ser atendido, assumindo que a probabilidade de um cliente chegar é constante ao longo do tempo.
- Parâmetros:
    - Lambda (taxa): Representa a frequência média de eventos por unidade de tempo.
        - Lambda maior indica que os eventos ocorrem com maior frequência revelando um tempo médio entre eventos menor.
        - Lambda menor significa eventos mais esporádicos com um tempo médio maior.
- Sem memória: Indica que a probabilidade de um evento ocorrer a seguir não depende de quanto tempo já passou.



4 - Distribuições mais usadas em Ciência de Dados

4.5 Qui-quadrado, t-*Student*, F

- São distribuições e testes de hipóteses estatísticos usados em inferências estatísticas com aplicações distintas:
    - **Qui-quadrado:** Compara frequências observadas e esperadas. Para dados categóricos, analisa frequências de associações.
        - Testa independência ou aderência a uma distribuição teórica
    - **T-student:** Compara médias de dois grupos. Para dados numéricos.
        - usado quando a variância populacional é desconhecida e a amostra é pequena
    - **F:** Compara variâncias entre dois ou mais grupos ou múltiplos grupos de médias.

4 - Distribuições mais usadas em Ciência de Dados

4.5. 1 Qui-quadrado

- É uma distribuição de probabilidade e um teste estatístico para analisar a relação entre variáveis categóricas (ou qualitativas), comparando dados observados com dados esperados para verificar se há uma associação ou diferença significativa.
- Se o valor calculado do Qui-Quadrado foi maior que um valor crítico, rejeita-se a hipótese nula, indicando uma associação entre as variáveis.

4 - Distribuições mais usadas em Ciência de Dados

4.5. 1 Qui-quadrado

- É utilizada:
  - Em **teste de independência**: Para verificar se duas variáveis categóricas estão relacionadas ou se são independentes.
  - Em **teste de aderência**: Avaliar se as frequências (dados) observadas em uma amostra seguem uma distribuição esperada.
- Aplicações: Testar se uma determinada característica (ex. preferência de cor) diferente entre diferentes grupos (ex. crianças/adolescentes).

4 - Distribuições mais usadas em Ciência de Dados

4.5. 1 Qui-quadrado - Funcionamento

1. **Formulação da hipótese nula:** Assume-se que não há relação entre as variáveis
2. **Cálculo das frequências Esperadas:** São calculadas as frequências que seriam esperadas sob as condições de não associação
3. **Cálculo da estatística Qui-Quadrado:**

$$\chi^2 = \sum_{i=1}^{m \times n} \frac{(O_i - E_i)^2}{E_i}$$

- **Oi** = frequências reais observadas
- **Ei** = frequências esperadas em cada categoria
- **m** e **n** = quantidade de linhas e colunas da tabela de contingências
- **Tabela de contingência:** Essa é uma tabela cruzada ou tabela bidirecional. Você usa para mostrar uma variável em uma linha e outra em uma coluna com sua contagem de frequência. É um tipo de tabela de distribuição de frequência das variáveis categóricas.

4 - Distribuições mais usadas em Ciência de Dados

4.5. 1 Qui-quadrado - Funcionamento

4. Comparação com o valor crítico

- A estatística calculada é comparada com um valor tabelado da distribuição qui-quadrado, considerando graus de liberdade e o nível de significância.

5. Interpretação

- Se o X2 calculado for maior que o valor crítico, a hipótese nula é rejeitada, o que sugere uma associação estatisticamente significativa.

https://www.datacamp.com/pt/tutorial/chi-square-test-in-spreadsheets
https://www.blog.psicometriaonline.com.br/qui-quadrado-teste-de-independencia/

4 - Distribuições mais usadas em Ciência de Dados

4.5. 1 Qui-quadrado - Exemplo

Um exemplo comum é testar se existe uma relação entre o gênero e a preferência por uma marca de refrigerante, onde a hipótese nula é que não há associação. Após calcular o valor estatístico qui-quadrado e compará-lo com um valor crítico, a hipótese nula é rejeitada se o valor calculado for maior, indicando uma associação significativa.

https://www.datacamp.com/pt/tutorial/chi-square-test-in-spreadsheets
https://www.blog.psicometriaonline.com.br/qui-quadrado-teste-de-independencia/

4 - Distribuições mais usadas em Ciência de Dados

4.5.1 T-Student

- É uma distribuição de probabilidade e um teste de hipótese paramétrico.
- É utilizada: Para comparar as médias de dois grupos quando os dados seguem uma distribuição normal e a variância da população é desconhecida ou estimada a partir de uma amostra pequena.
- Aplicações: Comparações de médias de resultados entre dois grupos (A e B) para ver se há uma diferença significativa.

4 - Distribuições mais usadas em Ciência de Dados

4.5. 3 - Distribuição F

- É uma distribuição de probabilidade usada para comparar as variâncias de dois ou mais grupos de dados.
- É utilizada:
  - Em Análise de Variância (ANOVA): Para determinar se as médias de três ou mais grupos são estatisticamente diferentes.
  - Em comparação de variâncias: Para testar a hipótese de que as variâncias de duas populações são iguais.
- Aplicação: Verifica se a variabilidade dos resultados de uma experiência é diferente entre diferentes tratamentos em um estudo.

5 - Aplicações em Ciência de Dados

- **Treino e teste em Machine Learning** → amostragem é a base para evitar overfitting.
- **Bootstrapping** e **Cross-validation** → uso de amostras para validar modelos.
- **Estatística inferencial**: intervalos de confiança e testes de hipótese dependem das distribuições.
- **Amostragem em Big Data** → reduzir volume mantendo representatividade.

Perguntas

Obrigado(a)!