1 PROJETANDO ESTUDOS

PROJETO EXPERIMENTAL

DESIGN EXPERIMENTAL

O design experimental é o plano que o pesquisador cria para conduzir um experimento. Ele inclui as etapas específicas a serem seguidas para garantir que os resultados sejam confiáveis e válidos.

O projeto experimental busca minimizar viés, controlar variáveis indesejadas e fornecer uma base sólida para a interpretação dos resultados, contribuindo para o avanço do conhecimento em uma determinada área.

Conceitos fundamentais:

* Controle:
  + Definição: Envolve a comparação do grupo experimental com um grupo de controle que não recebe o tratamento.
  + Importância: Isola e identifica os efeitos do tratamento, permitindo a determinação se as mudanças observadas são devidas ao tratamento ou a outros fatores.
* Randomização:
  + Definição: Atribuição aleatória dos participantes aos grupos de tratamento e controle.
  + Importância: Garante que os grupos sejam comparáveis antes do tratamento, reduzindo o viés na atribuição e permitindo a atribuição de mudanças às intervenções.
* Replicação:
  + Definição: Coleta de uma amostra suficientemente grande.
  + Importância: Aumenta a confiabilidade dos resultados e contribui para a generalização dos achados para a população maior.
* Bloqueio:
  + Definição: Utilizado quando há variáveis conhecidas que podem afetar a variável resposta. Agrupa os participantes em blocos com base nessas variáveis.
  + Importância: Ajuda a controlar variáveis de confusão, garantindo que o efeito do tratamento seja mais claramente observado.

ATRIBUIÇÃO DE AMOSTRAS ALEATÓRIAS

* Amostras aleatórias: todos os indivíduos da população têm mesma probabilidade de serem selecionados paras o estudo
* Atribuição aleatória: ocorre apenas em ambientes experimentais

2 EXPLORAÇÃO DE DADOS NUMÉRICOS

VISUALIZAÇÃO DE DADOS NUMÉRICOS

GRÁFICO DE DISPERSÃO

* Utilizado para observar a relação entre duas variáveis numéricas
* Variável explicativa (eixo x): interfere/ causa efeito na variável resposta
* Variável resposta (eixo y): representa a variável que está sendo observada ou medida em resposta à variável explicativa.

HISTOGRAMA

* Visualizar a distribuição de uma variável numérica
* Visão de densidade dos dados e distribuição
* Agrupa os dados em intervalos (barras) e mostra a frequência de observações em cada intervalo.

DOTPLOT (DIAGRAMA DE PONTOS)

* Útil quando valores individuais são interessantes
* Cada ponto no gráfico representa uma observação individual, proporcionando uma visão direta dos dados.

BOXPLOT (GRÁFICO DE CAIXA)

O boxplot é um gráfico utilizado para representar a distribuição de uma variável numérica, já que ele fornece uma representação visual dessa distribuição, incluindo mediana, quartis e possíveis *outliers.* Seus elementos (Figura 4) são:

* *Outliers:* valores discrepantes se comparados com o conjunto de dados; São valores representados fora dos limites superior e inferior (PERES, 2022).
* Limite superior e inferior (PERES, 2022).
* Mediana: é a medida que separa os dados ordenados em 50% superiores e 50% inferiores, sendo, dessa forma, o ponto médio dos dados. No gráfico *boxplot* a mediana coincide com o 2º quartil PERES, 2022).
* 1º e 3º quartis: É semelhante a lógica da mediana, no entanto, ao invés de dividir os dados na metade, é divido por quatro (cada parte/ quartil contendo 25% dos dados) O primeiro quartil (Q1) é o valor abaixo do qual 25% dos dados estão localizados. O terceiro quartil (Q3) é o valor abaixo do qual 75% dos dados estão localizados. O segundo quartil é a mediana, o valor que divide a distribuição ao meio. (PERES, 2022)
* A amplitude interquartil (IQR) é a diferença entre Q3 e Q1 e é representada pelo próprio box no boxplot (a IQR é proporcional ao seu comprimento). (PERES, 2022)
* A "whisker" (linha que se estende para fora do box) geralmente se estende até 1,5 vezes o IQR a partir de Q1 e Q3, e valores fora dessa faixa são considerados outliers. (PERES, 2022).

Figura 2 - Elementos do gráfico *boxplot.*

Uma imagem contendo Gráfico

Descrição gerada automaticamente

Fonte: Peres (2022).

MAPA DE INTENSIDADE

* Envolve distribuição espacial
* Pode ser usado para visualizar como a intensidade ou o valor de uma variável numérica varia em diferentes regiões geográficas

CARACTERÍSTICAS DE DISTRIBUIÇÃO - Termos de Assimetria

1. Inclinado para a Esquerda (Assimetria Negativa):

- A cauda da distribuição estende-se mais para a esquerda.

- A maioria dos valores está concentrada na parte direita da distribuição.

- A média é menor que a mediana.

2. Simétrico:

- A distribuição é equilibrada, sem inclinação evidente para a esquerda ou para a direita.

- A média é aproximadamente igual à mediana.

- A curva de densidade de probabilidade é simétrica em torno do ponto central.

3. Inclinado para a Direita (Assimetria Positiva):

- A cauda da distribuição estende-se mais para a direita.

- A maioria dos valores está concentrada na parte esquerda da distribuição.

- A média é maior que a mediana.

CARACTERÍSTICAS DE MODALIDADE DA DISTRIBUIÇÃO

1. Unimodal:

- Possui um único pico ou modo na distribuição.

- A maioria dos valores concentra-se em torno de um valor central.

2. Bimodal:

- Possui dois picos ou modos distintos na distribuição.

- Indica a presença de duas diferentes concentrações ou grupos de valores.

3. Uniforme:

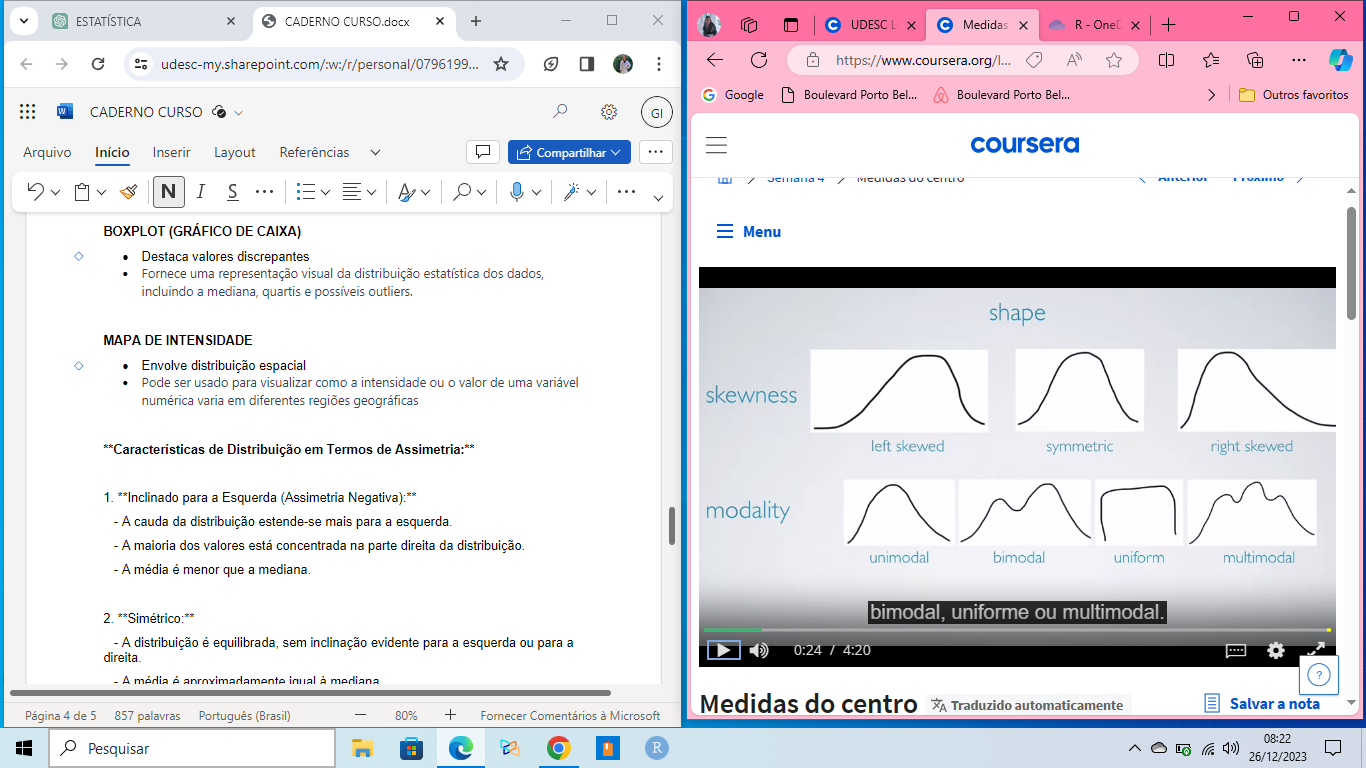
- Todos os valores têm uma frequência aproximadamente igual.

- A distribuição é plana, sem picos ou vales evidentes.

4. Multimodal:

- Possui mais de dois picos ou modos.

- Indica a presença de múltiplos grupos ou concentrações de valore



MEDIDAS DO CENTRO

DEFINIÇÃO: Uma medida de centro é um valor no centro, ou meio, do conjunto de dados.

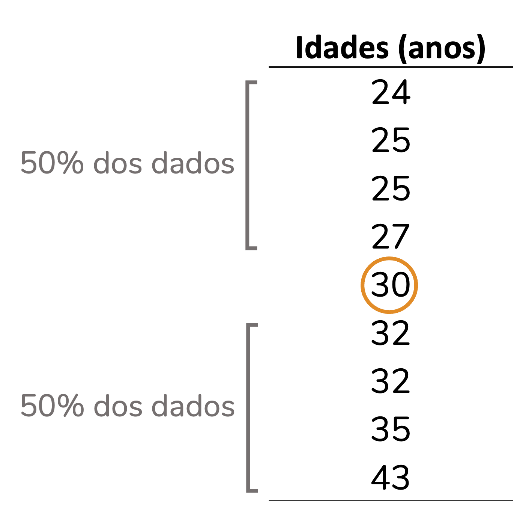
CENTRO DA DISTRIBUIÇÃO

Média: média aritmética dos valores

* + Sensível a valores extremos (outliers) que podem distorcê-la.

Mediana: ponto médio dos dados

* + Representa o ponto médio dos dados quando ordenados.
  + É a medida que separa os 50% inferiores dos 50% superiores dos dados.



Modo: Refere-se ao valor observado com maior frequência na distribuição.

* Nem sempre é um valor muito útil, especialmente em distribuições contínuas.
* Pode ser mais informativo em distribuições discretas ou multimodais.
* Quando dois valores ocorrem com maior frequência, cada um é uma moda, e o conjunto de dados é bimodal.
* Quando mais de dois valores ocorrem com maior frequência, cada um é uma moda, e o conjunto de dados é multimodal.
* Quando nenhum valor se repete, dizemos que não há moda.

RELAÇÃO ENTRE MÉDIA, MEDIANA E ASSIMETRIA

* Distribuição inclinada para a esquerda: média < mediana
* Distribuição simétrica: são aproximadamente iguais
* Distribuição inclinada para a direita: média > mediana
* Em uma distribuição inclinada para a esquerda, a mediana tende a ser maior do que a média, portanto, esperamos que a média/mediana seja menor do que 1.

MEDIDAS DE PROPAGAÇÃO/DISPERSÃO/VARIAÇÃO

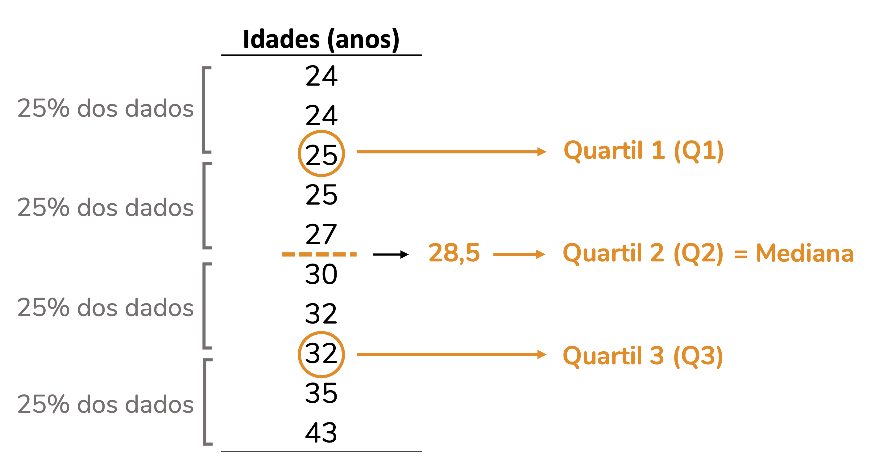
DISPERSÃO

* Indica o grau de afastamento de um conjunto de números em relação à sua média.
* Variância:
  + A média do desvio em relação a média ao quadrado
  + Indica o quanto os dados estão espalhados ao redor da média.
  + Quanto maior a variância, maior a dispersão dos dados.
* Desvio padrão
  + O desvio padrão é a raiz quadrada da variância.
  + Fornece uma medida da dispersão dos dados em termos da unidade original da variável.
  + Desvios padrão maiores indicam maior variabilidade.

DISTRIBUIÇÃO

**Intervalo interquartis**

A lógica por trás dos quartis é muito semelhante à lógica da mediana. Mas, ao invés de dividir o conjunto de dados em duas partes iguais, como a mediana, os quartis dividem esse conjunto em quatro (daí vem o nome quartis), cada um contendo 25% (um quarto) dos dados. Veja no exemplo abaixo. Note que o quartil 2 corresponde à mediana.



**Quartis:**

* O primeiro quartil (Q1) é o valor abaixo do qual 25% dos dados estão localizados.
* O terceiro quartil (Q3) é o valor abaixo do qual 75% dos dados estão localizados.
* O segundo quartil é a mediana, o valor que divide a distribuição ao meio.
* O amplitude interquartil (IQR- interquartile range) é a faixa da metade dos dados e é calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1).
* Representa a dispersão dos dados na região central da distribuição.
* Menos sensível a valores extremos do que a variância e o desvio padrão.
* É uma medida prontamente disponível em boxplots, onde a "caixa" representa o IQR.

ESTATÍSTICA ROBUSCA

Refere-se a medidas que são menos sensíveis a valores extremos (outliers) em um conjunto de dados. Ela é especialmente útil quando lidamos com distribuições que possuem observações atípicas que podem distorcer as estimativas tradicionais.

Medidas Robustas:

**Mediana:** é uma medida de tendência central que é menos afetada por valores extremos do que a média.Ela representa o ponto médio dos dados ordenados e não é influenciada por valores extremos.

**Intervalo Interquartil (IQR):** O IQR é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) e é uma medida de dispersão robusta. Ele é menos suscetível a valores extremos do que o desvio padrão.

Importância e Aplicações:

Redução de Impacto de Outliers:

Em conjuntos de dados com valores extremos, a estatística robusta reduz o impacto desses valores, proporcionando uma visão mais equilibrada da distribuição.

Para que Serve:

A estatística robusta serve para melhorar a confiabilidade e a validade das análises estatísticas em situações onde valores extremos podem distorcer os resultados. Ela proporciona uma visão mais resistente e equilibrada da distribuição dos dados, tornando as conclusões mais confiáveis em cenários com heterogeneidade ou presença de observações atípicas.

TRANSFORMANDO DADOS

A transformação de dados é um procedimento estatístico que busca modificar a escala ou a distribuição dos dados originais. O principal objetivo é alcançar uma relação mais linear entre as variáveis, o que pode ser benéfico em análises estatísticas e modelagem.

Exemplos de Transformações: Logarítmica (Log): A transformação logarítmica é frequentemente usada para reduzir a variação em dados que mostram uma tendência exponencial.

3 EXPLORAÇÃO DE DADOS CATEGÓRICOS

RELAÇÃO ENTRE VARIÁVEIS CATEGÓRICAS

Gráficos de barras: Os gráficos de barras são uma ferramenta visual fundamental para exibir a distribuição de variáveis categóricas. As categorias podem ser apresentadas em qualquer ordem, facilitando a interpretação.

Gráficos de barras segmentado: Cada barra é dividida em segmentos correspondentes às diferentes categorias, permitindo uma análise mais detalhada.

Gráfico circula (pizza): É útil para destacar proporções relativas entre categorias, mas pode ser menos eficaz na comparação precisa de tamanhos absolutos.

Gráfico mosaico: Uma alternativa interessante que exibe proporções proporcionais em um formato de mosaico.

Histogramas são utilizados para variáveis numéricas: Eixo é uma reta numérica -> a ordem das barras não pode ser alterada

4 INTRODUÇÃO A PROBABILIDADE

A estatística é dividida em:

* Probabilidade: estudo da aleatoriedade e incerteza;
* Estatística Descritiva: Organizar e Representar os dados (gráficos, planilhas etc)
* Estatística Inferencial: Realizar inferências nos dados coletados, utilizando a probabilidade

Entender probabilidade é como lidar com a incerteza e a chance nos eventos da vida. Vamos explorar alguns conceitos fundamentais de uma maneira mais acessível.

DEFINIÇÃO

Probabilidade é uma maneira de medir a chance de algo acontecer. Quando usamos a letra P, estamos falando sobre a probabilidade de um evento específico. Por exemplo, P(chuva) seria a probabilidade de chover.

ESPAÇO AMOSTRAL

O espaço amostral é um termo complicado para dizer "todos os possíveis resultados de um experimento". Se estamos falando sobre lançar um dado, o espaço amostral seria todos os números que poderíamos tirar (1, 2, 3, 4, 5, 6).

EVENTOS SIMPLES E COMPLEXOS

Um evento é basicamente algo que acontece. Se tirarmos um 4 ao lançar o dado, isso é um evento simples. Se falamos sobre obter um número par (2, 4, 6), isso é um evento mais complexo.

A REGRA DO EVENTO RARO

A regra do evento raro é como uma dica: se a chance de algo acontecer é super pequena, talvez nossa ideia inicial esteja errada. Por exemplo, se a chance de ganhar na loteria é quase zero, talvez não seja uma boa aposta.

LEI DOS GRANDES NÚMEROS

Essa lei nos diz que, à medida que fazemos um experimento mais e mais vezes, a probabilidade de um evento acontecer se aproxima da probabilidade "real". Isso significa que, com muitas tentativas, nossas previsões ficam mais próximas da realidade.

EVENTOS DISJUNTOS E A REGRA GERAL DE ADIÇÃO

Eventos Disjuntos:

Eventos disjuntos são aqueles que não podem ocorrer ao mesmo tempo. Se um evento A está acontecendo, o evento B não está. A probabilidade de ambos acontecerem simultaneamente é zero.

Probabilidade de A e B Acontecerem (Eventos Disjuntos):

Se temos eventos disjuntos A e B, calcular a probabilidade de ambos ocorrerem é simples. Basta somar a probabilidade de A com a probabilidade de B.

P(A∩B)=P(A)+P(B)

Probabilidade de A ou B Acontecer (Regra Geral de Adição):

Agora, se queremos calcular a probabilidade de A ou B acontecerem (não ambos ao mesmo tempo), usamos a regra geral de adição. É a probabilidade de A mais a probabilidade de B, mas precisamos subtrair a probabilidade de ambos acontecerem, para não contar duas vezes os eventos simultâneos.

P(A∪B)=P(A)+P(B)−P(A∩B)

INDEPENDÊNCIA

Eventos Independentes:

Dois ou mais eventos são considerados independentes quando a ocorrência de um evento não influencia a probabilidade da ocorrência dos outros. Em outras palavras, a informação sobre a ocorrência de um evento não fornece nenhuma informação adicional sobre a ocorrência dos outros.

Regra Geral para Verificar a Independência:

Para determinar a independência entre processos aleatórios, examinamos a dependência nos dados de amostra. Se observarmos diferenças nas probabilidades condicionais com base na amostra, podemos sugerir dependência. O próximo passo seria realizar um teste de hipótese para confirmar a independência ou dependência.

Regra do Produto:

Se A e B são eventos independentes, a probabilidade de ambos ocorrerem é simplesmente o produto de suas probabilidades individuais.

P(A∩B)=P(A)⋅P(B)

Essa regra é útil quando queremos calcular a probabilidade de dois eventos independentes acontecerem em conjunto. Portanto, se a igualdade não for respeitada, significa que os eventos não são independentes

PROBABILIDADE CONDICIONAL

Quando queremos calcular a probabilidade de ocorrência de A, considerando que B já ocorreu, representamos isso como P(A∣B) (probabilidade de A dado B). A fórmula é dada por:

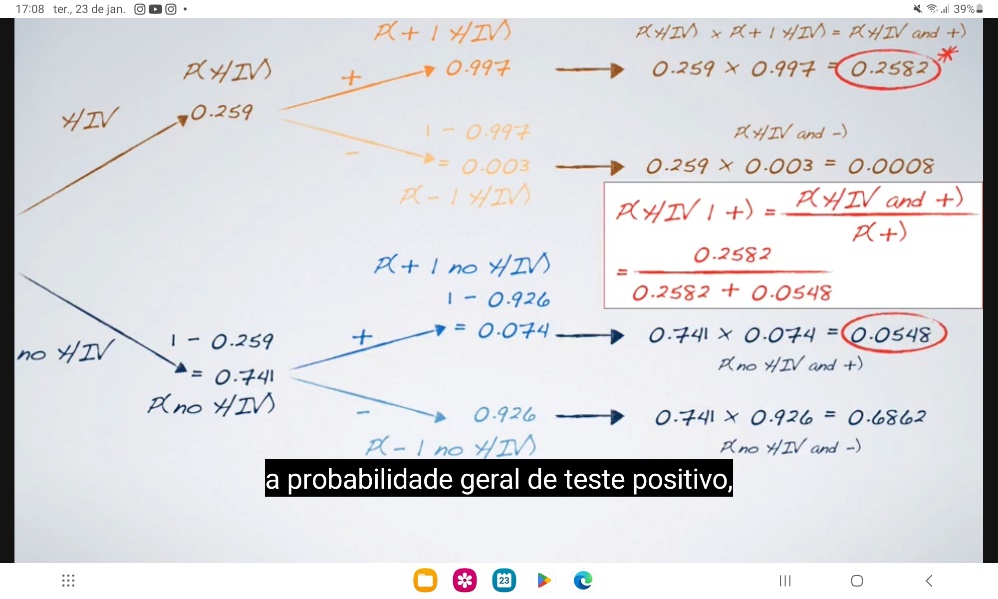
P(A∣B)= P(A∩B) / P(B)

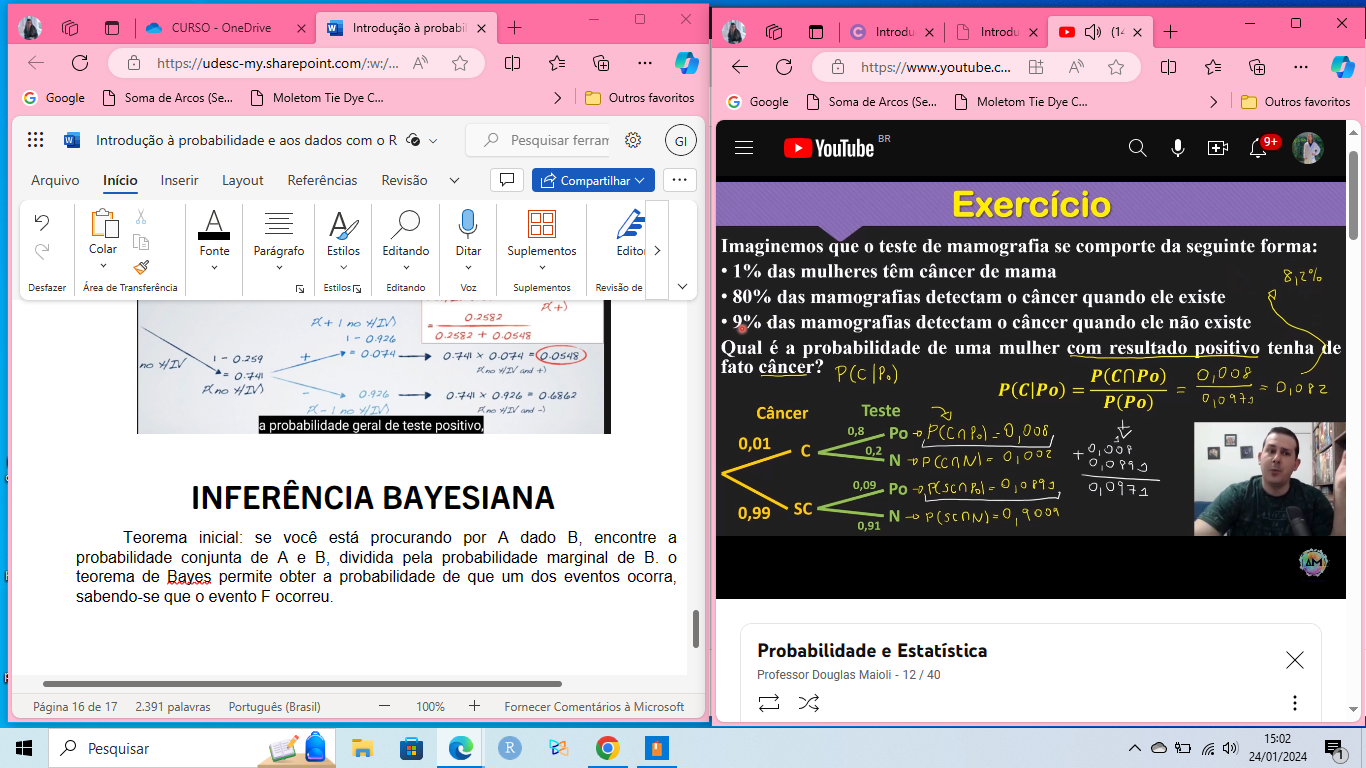
Em palavras, a probabilidade de A ocorrer, dado que B ocorreu, é igual à probabilidade de ocorrência simultânea de A e B dividida pela probabilidade de ocorrência de B.

Calculamos probabilidades condicionais usando o Teorema de Bayes, que afirma que a probabilidade de A dado B é a probabilidade de A e B dividida pela probabilidade de B (evento ao qual estamos condicionados).

ARVORE DE PROBABILIDADE

A árvore de probabilidade é uma representação gráfica que ajuda a visualizar e calcular probabilidades em situações que envolvem múltiplos eventos inter-relacionados.





INFERÊNCIA BAYESIANA

O Teorema de Bayes é uma ferramenta na teoria da probabilidade que nos permite calcular a probabilidade de um evento com base na probabilidade de outro evento relacionado. Em termos simples, se estamos procurando a probabilidade de um evento A dado que o evento B ocorreu, usamos o teorema de Bayes.

Fórmula do Teorema de Bayes: (A∣B)=P(B∣A)⋅P(A) / P(B)

Este teorema nos orienta sobre como atualizar nossas crenças (probabilidades) quando novas evidências (eventos) são apresentadas.

Probabilidade Posterior: A probabilidade posterior é a probabilidade da hipótese que estabelecemos, levando em consideração os dados recentemente observados. Em outras palavras, após coletar dados, atualizamos nossas crenças iniciais (probabilidades anteriores) para obter as chamadas probabilidades posteriores.

Passos para Aplicação:

* Definir uma Probabilidade Anterior: Antes de observar os dados, temos uma probabilidade inicial com base em conhecimento prévio ou crenças.
* Coletar Dados: Observar dados ou evidências relacionadas ao evento em questão.
* Calcular a Probabilidade Posterior: Utilizando o Teorema de Bayes, calculamos a probabilidade posterior com base nos dados recém-observados.
* Atualizar a Probabilidade Anterior: A probabilidade posterior torna-se a nova probabilidade anterior para futuras atualizações, criando um processo iterativo.

Essa abordagem Bayesian proporciona uma maneira robusta de incorporar novas informações em nossas análises e tomadas de decisão, permitindo uma adaptação contínua à medida que mais dados são disponibilizados.

3 DISTRIBUIÇÕES DE PROBABILIDADE

As distribuições de probabilidade são ferramentas fundamentais na teoria da probabilidade e na estatística. É um modelo matemático que relaciona um certo valor da variável aleatória com a sua probabilidade de ocorrência. Elas modelam a variabilidade de um conjunto de dados e fornecem uma estrutura matemática para entender e analisar eventos aleatórios. Duas distribuições de probabilidade comumente utilizadas são a distribuição normal e a distribuição binomial.

**Distribuição Normal:**

*Características Principais:*

* Unimodal e simétrica, com uma forma de sino.
* Totalmente definida por dois parâmetros: média (μ) e desvio padrão (σ).
* A curva normal descreve a variabilidade de muitos fenômenos naturais e humanos.

*Utilização:*

* Modela dados contínuos, como medidas físicas, pontuações em testes, etc.
* Permite calcular probabilidades associadas a diferentes valores em uma distribuição contínua.
* Facilita análises estatísticas e inferências.

**Distribuição Binomial:**

*Características Principais:*

* Modela experimentos binários, onde existem apenas dois resultados possíveis: sucesso ou fracasso.
* Cada ensaio é independente e tem uma probabilidade constante de sucesso (denotada por p).
* Os ensaios são realizados um número fixo de vezes (n).

*Utilização:*

* Modela situações de contagem, como o número de sucessos em uma sequência de ensaios independentes.
* Aplica-se em áreas como controle de qualidade, estudos de confiabilidade, entre outros.

****DISTRIBUIÇÃO NORMAL****

A distribuição normal é caracterizada por ser unimodal e simétrica, apresentando uma curva em forma de sino. Essa distribuição é totalmente definida por dois parâmetros: a média e o desvio padrão.

Regras para variabilidade dos dados normalmente distribuídos em torno da média:

**68% dos dados:** Localizam-se dentro de um desvio padrão da média.

**95% dos dados:** Encontram-se dentro de dois desvios padrão da média.

**99,7% dos dados:** Situam-se dentro de três desvios padrão da média.

É possível avaliar se uma distribuição é normal ou não através de m gráfico de probabilidade normal. Para isso, os dados são plotados no eixo y e os quartis teóricos são plotados no eixo x (linha reta). É possível fazer, também, através de um histograma.

**Pontuações padronizadas / Pontuação Z:**

A pontuação Z, ou escore padronizado, representa o número de desvios padrão pelos quais uma observação está acima ou abaixo da média. É calculada pela fórmula:

**Percentis:**

Os percentis indicam a porcentagem de observações que fica abaixo de um determinado ponto de dados. Em uma distribuição normal, as pontuações Z também são usadas para calcular os percentis, geralmente consultando tabelas específicas.

Em R**:** Na linguagem R, as funções úteis são 'pnorm' para probabilidades e 'qnorm' para quantis (valores de corte) que utilizam o percentil como a primeira entrada, a média como a segunda e o desvio padrão como a terceira. Por exemplo:

qnorm(0.9, mean = 1500, sd = 300)

Essa função retornará o valor correspondente ao percentil 90% em uma distribuição normal com média 1500 e desvio padrão 300.

DISTRIBUIÇÃO BINOMINAL

**Definição:** A distribuição binomial é uma distribuição discreta de probabilidade aplicável quando um experimento é repetido n vezes, cada tentativa sendo independente e com uma probabilidade de sucesso p. Ela se aplica a situações em que várias tentativas independentes são realizadas, e cada uma tem apenas dois resultados possíveis.

**Condições para uma Variável Aleatória Binomial:** Para que uma variável aleatória siga uma distribuição binomial, é necessário atender a algumas condições:

Os testes devem ser independentes.

O número de tentativas (n) deve ser fixo.

Cada tentativa deve ter apenas dois resultados possíveis: sucesso ou fracasso.

A probabilidade de sucesso (p) deve ser constante em cada tentativa.

**Configurações e Cálculo:** A distribuição binomial é caracterizada por k sucessos em n ensaios independentes de Bernoulli, com probabilidade de sucesso p. O cálculo da probabilidade de k sucessos em n ensaios é dado por

Onde o número binomial é dado por:

**Notações Importantes:**

* p representa a probabilidade de sucesso.
* 1−p representa a probabilidade de falha.
* n é o número de tentativas/ experimentos.
* k é o número de sucessos.

**Função "choose" em R:**

A função "choose" (n choose k) calcula o número de maneiras de escolher k sucessos em n tentativas.

Por exemplo, em R, para calcular o número de combinações de 2 sucessos em 9 tentativas, você usaria ***choose(9, 2).***

**Função "dbinom" em R:**

A função "dbinom" em R calcula a probabilidade de obter um número específico de sucessos (k) em um determinado número de ensaios (size), com uma probabilidade de sucesso constante (p). Por exemplo***, dbinom(8, size = 10, p = 0.13)*** calcularia a probabilidade de obter exatamente 8 sucessos em 10 ensaios, com uma probabilidade de sucesso de 0.13.

Calculadora de distribuição: [Distribution Calculator (shinyapps.io)](https://gallery.shinyapps.io/dist_calc/)