Tipos de vieses em Dados

Prof. Dr. Ronan Assumpção Silva

Disciplina de Ética e Viéses em Dados





Erros em dados

Os erros encontrados em dados ocorrem na coleta, organização, representação ou análise. É um erro tido como não intencional e aleatório.

Exemplo: após um questionário ser aplicado, a transcrição da folha de papel para o meio digital possui erro(s) de digitação.

Viés em dados?

Definição:

Viés em dados ocorre quando uma informação não é precisa e falha na representação da população como um todo. Como consequência tem-se resultados distorcidos e desigualdade. Portanto, é importante identificar e evitar de tempos em tempos.

Reflexão

Quando precisamos estimar o perfil político dos eleitores, um estado isolado é suficientemente capaz de representar os demais estados do Brasil?

Exemplo

Considere que a população do estudo são os brasileiros. Ao coletar dados de apenas um dos estados, podemos representar a população toda?

Possibilidades de erro:

- Imprecisão
- Desigualdade
- Incompleto
- Incompatível



Tipos de vieses:

Resposta

Seleção

Social

Automação

Dados implícitos

Atribuição de grupo

Viés de resposta

Ocorre quando a resposta é falsa ou incompleta. Ou ainda pode representar poucas opiniões. Tudo isso pode acarretar em conclusões generalistas, erradas e tendenciosas.

Exemplo: Um questionário é aplicado para detectar o perfil dos eleitores, porém os leitores não respondem sinceramente ou são evasivos para não reverlar a opinião.

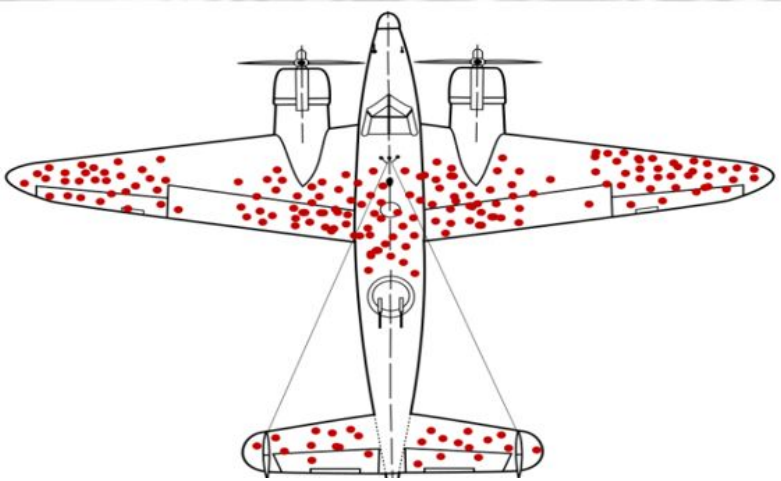
Viés de seleção/amostra

A amostra não é apropriada, seja pela quantidade ou pela distribuição destas amostras.

Exemplo: Ao analisar a opinião dos trabalhadores sobre o plano de carreira não é considerado por exemplo a igualdade de gênero.

Viés de amostragem (2º exemplo)

Seção do avião	Balas	%
Motores (2)	39	0.193
Fuselagem	78	0.386
Sistema de combustível	31	0.154
Outras seções	54	0.267



Qual é a seção do avião que requer mais atenção quanto a reforço na proteção?

Na segunda guerra Abraham Wald concluiu que era necessário reforçar a proteção do motor, mesmo sendo a parte menos atingida de um avião que retorna a base ou é resgatado.

Qual é a motivação do cientista?

Viés social

É chamado de social ou de rótulo o viés de gênero e de raça. Quando os dados possuem essas características, ferramentas que são baseadas nesses dados serão influenciadas.

Exemplo: Uma ferramenta de Inteligência Artificial explicar diferentemente conceitos quando sabe-se que o usuário é homem ou mulher.

Viés de automação de dados

Esse tipo de viés acontece quando é considerado mais importante a fonte de dados gerada por sistemas do que por pessoas. Como os dados são importantes para o funcionamento dos sistemas, esses dados devem representar corretamente o cenário.

Exemplo: Sistemas com reconhecimento de padrão alimentados principalmente por dados artificiais.

Viés de dados implícitos

Considere uma base de dados composta por suposições. É a opinião da pessoa baseada em suas experiências. Pode não estar relacionada conscientemente ao viés/preconceito.

Exemplo: A descrição de uma família tradicional.

O que acontece na sua casa acontece com todos?

Viés de atribuição de grupo

Esse viés é considerado quando a distribuição de dados entre os indivíduos e grupos é feito uniformemente, no sentido de que o comportamento e características de cada um são considerados o mesmo.

Exemplo: Na tratativa da desigualdade, colocar toda a população em programas sociais.

Viés de perspectiva/confirmação

Pode favorecer ao financiador, ao autor ou a um grupo específico ao interpretar os dados usando a subjetividade.

Exemplo 1: É como os comerciais tratam os produtos do anunciante em comparação com os concorrentes.

Exemplo 2: Imagine interpretar os dados forçando os padrões encontrados a se encaixarem com sua conclusão preestabelecida.

Viés de omissão de variável

Ao analisar os dados, uma variável pode afetar profundamente o entendimento dos padrões. Logo, ao omitir essa variável pode facilitar a interpretação, os gráficos e as conclusões.

Exemplo: Considere uma base de dados sobre abastecimento de água. Podemos concluir que 84,9% da população brasileira é atendida e apenas 15,1% da população não é atendida. Porém, ao considerar as regiões, temos a região Norte com 35,8% da população sem acesso (Brasil Escola, 2024).

Como indentificar?

Abordagens chave:

- Avaliar a fonte de dados
- Buscar valores discrepantes (*outliers*)

Avaliar a fonte de dados

- Saber como os dados foram gerados
- Descobrir se os dados foram verificados e por uma fonte/método confiável?
- Refletir sobre a competência do sistema/pessoa que gerou os dados

Buscar valores discrepantes (*outliers*)

- Procure por valores discrepantes, fantasiosos, mistura de unidades de medida
- Padrões fora de realidade
- Encontre dados em falta ou incompletos
- Investigue o motivo destas anomalias acima

Possíveis caminhos para identificar:

- Construção de figuras, gráficos, visualizações e estimativas

Como evitar?

- Encontrar bases de dados confiáveis
- Soluções computacionais são projetadas por humanos, logo, podem ter viés também. Minimizar o viés humano no desenvolvimento dos sistemas e na coleta de dados precisa ser considerado
- Reamostragem: abordagem que produz amostras a partir das encontradas em uma base de dados. Há diferentes técnicas para este fim

Obrigado

ronan.silva@ufpr.br

ronan.silva@ifpr.edu.br