Probabilidade

1. Introdução

A **Probabilidade** é uma área dentro da matemática e estatística que visa estudar e entender os eventos de comportamento **aleatório**, cujos resultados destes eventos são **probabilísticos** ou **estocásticos.** Em sistemas na natureza, existem poucos deles que de fato se comportam aleatoriamente, mas é conveniente tratar estes sistemas como estocásticos a fim de avaliar as probabilidades.

Alguns exemplos clássicos de aplicação de probabilidade são o lançamento de objetos como moedas e dados, seleção de bolas com ou sem repetições dentre outras aplicações, mas o estudo da probabilidade serve de fundamental ferramenta para a análise de dados, onde cria-se modelos que buscam generalizar relações que não podem ser determinadas diretamente, pois o acesso ao total da população destes dados é limitado, muitas dessas análises sendo feitas a partir de amostras. Dessa forma, entender os principais conceitos de probabilidade são cruciais para o desenvolvimento dentro da análise de dados.

2. Espaço Amostral

O **espaço amostral** é o conjunto de **todos os resultados possíveis** de um determinado fenômeno aleatório, ou seja, todos os valores possíveis que um fenômeno possa assumir. O espaço amostral comumente é representado pela letra grega ω. Os subconjuntos de ω são denominados **eventos** e comumente representados por letras maiúsculas (A, B entre outras).

Por exemplo, em um estudo onde o objetivo é avaliar os resultados do lançamento de um dado honesta e perfeito, ou seja o dado é simétrico e homogêneo, a definição do espaço amostral ω é dado por: $\omega = \{1, 2, 3, 4, 5, 6\}$

Levando em conta que o espaço amostral abrange todos os resultados possíveis, alguns exemplos de eventos ou subconjuntos desse espaço amostral estão indicados abaixo:

- **Evento A** = {1} : o resultado do lançamento do dado seja igual a 1;
- **Evento B** = {2, 4, 6} : o resultado do lançamento do dado seja um número par;
- **Evento C** = {5, 6}: o resultado do lançamento do dado seja maior do que 4;

4. Propriedades de Eventos de Probabilidade

Para que uma função P seja denominada como a **probabilidade de determinado evento**, esta função deve satisfazer algumas propriedades:

- A probabilidade de um evento A ocorrer será tal que $0 \leq P(A) \leq 1$;
- A probabilidade do espaço amostral será $P(\omega) = 1$, pois o espaço amostral abrange todos os eventos possíveis;
- A probabilidade conjunta de n eventos é dada por $P(\cup A_j) = \sum_{j=1}^{n} P(A_j)$, no caso destes eventos sejam mutuamente exclusivos.

5. Métodos de Avaliação da Probabilidade

No que tange definir a probabilidade de um determinado evento, pode ser feito o levantamento da probabilidade de duas maneiras:

1. A partir das propriedades teóricas do determinado evento

Então voltando ao exemplo do lançamento de um dado, foi definido que o espaço amostral seria igual a $\omega = 1, 2, 3, 4, 5, 6$. Novamente levantando as considerações sobre a característica física do dado, ou seja um dado homogêneo e perfeito, onde todas as faces são simétricas e têm igual chance de ocorrer, a probabilidade de ocorrer cada um dos lados do dado pode ser definida como:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6} = 0,166...$$

2. A partir de exaustivas experimentações

Ainda no problema onde serão feitos lançamentos de dados, uma forma de chegar nos valores das probabilidades para cada uma das faces do dado seria justamente repetir várias vezes o lançamento do dado e anotando os valores dos resultados, à medida que são feitos muitos experimentos (neste exemplo, serão simulados 1 milhão de lançamentos), a tendência é que os valores das probabilidades **convirjam para o valor teórico**. Abaixo segue um exemplo de implementação em *Python* de uma simulação de 1 milhão de lançamentos de dados e as frequências relativas das faces:

```python
# Definindo uma semente aleatória
np.random.seed(42)

# Número de experimentos
N = 1000000

# Definindo os resultados de um dado a partir de números aleatórios entre 1 e 6
amostra = np.random.randint(low = 1, # limite inferior
                            high = 7, # limite superior (não incluso)
```

```
                                   size = N) # quantidade de experimentos a ser feito

# Convertendo a amostra para uma Séries e calculando as frequências relativas
print(pd.Series(amostra).value_counts(normalize = True))
```

6. Probabilidade da União de Eventos

Dados dois eventos aleatórios quaisquer A e B, o cálculo da probabilidade da união desses eventos é determinada da seguinte forma:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Exemplificando com uma aplicação, qual seria a probabilidade que em um lançamento de dado saia um número par ou maior ou igual a 3?

Primeiramente, define o espaço amostral para o dado sendo $\omega = 1, 2, 3, 4, 5, 6$. Os eventos A (saiu um número par no dado) e B (saiu um número maior ou igual a 3) são definidos da seguinte forma:

- Evento A = $\{2, 4, 6\}$, ou seja a probabilidade do evento A será $P(A) = \frac{N_A}{N_\omega} = \frac{3}{6} = \frac{1}{2}$
- Evento B = $\{3, 4, 5, 6\}$, ou seja a probabilidade do evento B será $P(B) = \frac{N_B}{N_\omega} = \frac{4}{6} = \frac{2}{3}$

Agora, resta definir a probabilidade do evento AB, que seria ao lançar um dado sair um número par e um número maior ou igual a 3:

- Evento $AB = \{4, 6\}$, ou seja a probabilidade do evento AB será $P(AB) = \frac{N_{AB}}{N_\omega} = \frac{2}{6} = \frac{1}{3}$

Definidas todas as probabilidades, aplica-se a fórmula da união de eventos:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = \frac{5}{6}$$

Uma observação importante a ser feita é que ao lembrar das propriedades da probabilidade, no caso de os eventos A e B serem mutuamente exclusivos (quando um evento ocorre, o outro não), pode-se simplificar a formúla da união de eventos como a seguir:

$$P(A \cup B) = P(A) + P(B)$$

7. Probabilidade Condicional

Sejam dois eventos aleatórios A e B. Define-se a probabilidade condicional entre esses eventos pela seguinte relação:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Esta relação só será válida se a probabilidade do evento B for maior que zero, ou seja, $P(B) > 0$. A forma como é interpretada a notação $P(A|B)$ seria **qual a probabilidade do evento A ocorrer, dado que o evento B já ocorreu**. Em problemas de probabilidade condicional, a principal diferença é que o espaço amostral de referência para a probabilidade irá mudar.

Por exemplo, qual a probabilidade de, ao lançar um dado, sair um número ímpar, dado que o número que saiu é maior do que 2?

Deve-se primeiro notar que o espaço amostral para o dado é $\omega = \{1, 2, 3, 4, 5, 6\}$. Os eventos A (saiu um número ímpar no dado) e B (saiu um número maior do que 2) são definidos da seguinte forma:

- Evento $A = \{1,\ 3,\ 5\}$, ou seja a probabilidade do evento A será $P(A) = \frac{N_A}{N_\omega} = \frac{3}{6} = \frac{1}{2}$
- Evento $B = \{3, 4, 5, 6\}$, ou seja a probabilidade do evento B será $P(B) = \frac{N_B}{N_\omega} = \frac{4}{6} = \frac{2}{3}$

Agora define-se a probabilidade do evento $A \cap B$, que seria ao lançar um dado sair um número ímpar e que este seja um número maior do que 2:

- Evento $A \cap B = \{3, 5\}$, ou seja a probabilidade do evento $A \cap B$ será $P(A \cap B) = \frac{N_{AB}}{N_\omega} = \frac{2}{6} = \frac{1}{3}$

Por fim, calcula-se a probabilidade condicional $P(A|B)$, ou seja **a probabilidade do evento A ocorrer, dado que o evento B já ocorreu**:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

Vale notar que a probabilidade condicional é frequentemente utilizada para calcular a probabilidade da ocorrência simultânea de dois eventos, $P(AB)$:

$$P(A \cap B) = P(A|B)P(B)$$

Uma outra forma de obter esse resultado é notando que $P(A \cap B) = P(B \cap A)$. Assim, também temos:

$$P(A \cap B) = P(B|A)P(A)$$

8. Eventos Independentes

Dados dois eventos aleatórios A e B, caso esses eventos sejam **independentes** entre si, pode-se deduzir a seguinte relação sobre a probabilidade condicional:

$$P(A|B) = P(A)$$

Isto é válido pois o evento A ocorrerá **independente** do que passar pelo evento B. Aplicando agora esta relação na fórmula da probabilidade condicional, obtém-se o resultado a seguir:

$$P(A \cap B) = P(A)P(B)$$

9. Teorema de Bayes

O **Teorema de Bayes** é um resultado importante que pode ser deduzido a partir da probabilidade condicional. Sejam os eventos A e B e a definição da probabilidade condicional:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Utilizando a definição, se invertermos os eventos, isto é, calculando a probabilidade condicional do evento B dado que ocorreu o A, teremos:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Em ambas expressões acima temos $P(A \cap B)$. Igualando as relações consegue-se chegar no seguinte resultado:

$$P(A|B)P(B) = P(B|A)P(A)$$

A forma mais convencional de ser apresentada a relação do Teorema de Bayes é como observado abaixo:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Um exemplo de aplicação deste teorema seria na avaliação de um determinado teste B em identificar uma doença A. Sabe-se que 1% das pessoas possui uma certa doença A, além disso o teste B não é totalmente confiável: ele acerta 90% dos casos positivos e 95% dos casos negativos. Qual a probabilidade de uma pessoa que fez o teste estar realmente doente, dado que o teste deu positivo?

Montando o Teorema de Bayes a partir das condições apresentadas, tem-se que:

$$P(A = doente|B = positivo) = \frac{P(B = positivo|A = doente)P(A = doente)}{P(B = positivo)}$$

O próximo passo seria construir/calcular a probabilidade $P(B = positivo)$:

$$P(B = positivo) = P(B = positivo A = doente) + P(B = positivo A = saudável)$$

Importante ressaltar neste ponto que a pessoa pode apresentar um teste positivo, estando doente ou não. Agora aplicando a fórmula da probabilidade condicional a ambos os termos:

$$P(B = positivo) = P(B = positivo|A = doente)P(A = doente) + P(B = positivo|A = saudável)P(A = saudável)$$

Voltando a definição original do Teorema de Bayes, tem-se que:

$$P(A = doente|B = positivo) = \frac{P(B = positivo|A = doente)P(A = doente)}{P(B = positivo|A = doente)P(A = doente) + P(B = positivo|A = saudável)P(A = saudável)}$$

Definindo as probabilidades necessárias a partir do enunciado:

- $P(A = doente) =$ 1%; $P(A = saudável) = 1 - P(A = doente) =$ 99%;
- $P(B = positivo|A = doente) =$ 90%;
- $P(B = positivo|A = saudável) = 1 - P(B = negativo|A = saudável) =$ 5%.

Aplicando os valores no Teorema de Bayes chega-se ao seguinte resultado:

$$P(A = doente|B = positivo) = \frac{(90\%)(1\%)}{(90\%)(1\%) + (5\%)(99\%)} = 15,38\%$$

Materiais Complementares

Artigo sobre Teorema de Bayes e Probabilidade publicado pelo Lauro de Oliveira;

Referências

Pedro A. Morettin, Wilton O. Bussab, Estatística Básica, 8ª edição

Peter Bruce, Andrew Bruce & Peter Gedeck, Practical Statistics for Data Scientists, 50+ Essential Concepts Using R and Python, 2ª edition

Ron Larson & Betsy Farber, Estatística Aplicada, 6ª edição.