Data Science Academy

www.datascienceacademy.com.br

Business Analytics

Distribuições de Probabilidade e Função de Densidade

Probabilidade é um valor numérico que representa uma chance, uma eventualidade ou uma possibilidade de que um determinado evento venha a acontecer.

Existem três tipos de probabilidade:

Probabilidade clássica: usada quando cada resultado no espaço amostral tem a mesma probabilidade de ocorrer. A probabilidade é baseada no conhecimento prévio do processo envolvido.

Probabilidade empírica: baseia-se em observações obtidas de experimentos aleatórios. Os resultados são baseados em dados observados e não no conhecimento prévio do processo. De acordo com a Lei dos grandes números, a medida que um experimento é repetido mais e mais vezes, a probabilidade empírica (frequência relativa) de um evento tende à sua probabilidade real.

Probabilidade subjetiva: intuição estimativa ou palpite. Normalmente baseada em experiência no passado, opinião pessoal ou análise de algum indivíduo. Pode ser útil, quando não há possibilidade de utilização da probabilidade clássica ou empírica.

Definições:

Experimento ou Fenômeno Aleatório
São aqueles que, mesmo repetidos várias vezes sob condições semelhantes, apresentam resultados imprevisíveis.

Espaço Amostral
É o conjunto de possíveis resultados de um experimento ou fenômeno aleatório, representado por S.

Evento
É qualquer subconjunto do espaço amostral S de um evento aleatório.

Probabilidade: Dado um experimento aleatório, sendo S o seu espaço amostral, admitindo que todos os elementos de S tenham a mesma chance de acontecer, ou seja, que S é um conjunto equiprovável.

Probabilidade de um evento A (A está contido em S):

$$P(A) = \frac{n(A)}{n(S)}$$

Sendo:
n(A) – número de elementos de A;
n(S) – número de elementos de S.

Eventos Complementares

Sabendo que um evento pode ocorrer ou não. Sendo **p** a probabilidade de que ele ocorra (sucesso) e **q** a probabilidade de que ele não ocorra (insucesso), para um mesmo evento existe sempre a relação:

$$p + q = 1 \Rightarrow q = 1 - p$$

Eventos Independentes

Dizemos que dois eventos são independentes quando a realização ou não realização de um dos eventos não afeta a probabilidade da realização do outro e vice-versa.

Assim, sendo p_1 a probabilidade de realização do primeiro evento e p_2 a probabilidade do segundo evento, a probabilidade de que tais eventos se realizem simultaneamente é dada por:

$$p = p_1 \cdot p_2$$

Também conhecida como regra do "e".

Eventos Mutuamente Exclusivos

Dizemos que dois ou mais eventos são mutuamente exclusivos quando a realização de um exclui a realização do(s) outro(s).

Assim, no lançamento de uma moeda, o evento “tirar cara” e o evento “tirar coroa” são mutuamente exclusivos, já que, ao se realizar um deles, o outro não se realiza.

Se dois eventos são mutuamente exclusivos, a probabilidade de que um ou outro se realize é igual à soma das probabilidades de cada um deles se realize:

$$p = p_1 + p_2$$

Também conhecida como regra do "ou".

Distribuição Binomial

Aplica-se a experimentos que satisfaçam as seguintes condições:

- O experimento deve ser repetido, nas mesmas condições, um número finito de vezes, n;
- As provas repetidas devem ser independentes, o resultado de uma não afeta o resultado da outra;
- Tem-se apenas dois resultados possíveis: sucesso ou insucesso;
- A probabilidade do sucesso em uma tentativa é p e a do insucesso é: q = 1 - p
- A probabilidade de se obter sucesso k vezes durante n tentativas é determinado por:

$$f(x) = P(x = k) = \frac{n!}{k!\,(n-k)!} \cdot p^k \cdot q^{n-k}$$

Sendo:
n = número de tentativas
K = número de sucessos
p = probabilidade de sucesso
q = probabilidade de fracasso

Distribuição de Poisson

Muitos estudos são baseados na contagem das vezes em que um evento ocorre em uma determinada área de oportunidades.

Uma área de oportunidades é uma unidade contínua ou um intervalo de tempo, volume ou área em que possa acontecer mais de uma ocorrência de um evento. Exemplos:

- Defeitos na pintura de uma geladeira nova;
- Número de falhas na rede em um determinado dia;
- Número de pulgas no pelo de um cachorro.

Pode-se utilizar a distribuição de Poisson para calcular probabilidades se:

- O interesse é encontrar o número de vezes em que um evento específico ocorre em uma determinada área de oportunidades.
- A probabilidade de que um evento específico ocorra em uma área de oportunidades é a mesma para todas as áreas de oportunidades.
- número de eventos que ocorrem em uma determinada área de oportunidades é independente do número de eventos que ocorrem em qualquer outra área de oportunidades.
- A probabilidade de que dois ou mais eventos venham a ocorrer em uma determinada área de oportunidades se aproxima de zero à medida em que a área de oportunidades de torna menor.

$$P(x) = \frac{e^{-k}k^{x}}{X!}$$

Onde:
P(x) = probabilidade de eventos
K = número esperado de eventos
e = constante matemática = 2,71828

$$P(0 \le z \le z_1)$$

X = número de eventos

Distribuição Normal

A distribuição normal pode ser considerada como a mais importante distribuição de probabilidade, pode ser aplicada em vários fenômenos. Também é conhecida como distribuição de Gauss.

A Curva Normal é um modelo teórico ou ideal que resulta muito mais de uma equação matemática do que um real delineamento de pesquisa com posterior coleta de dados.

A distribuição normal é importantíssima para a Estatística pelas seguintes razões:

- Muitas variáveis contínuas possuem distribuição que se aproximam da normal;
- Pode ser utilizada para aproximações de várias distribuições discretas;
- Proporciona a base para a inferência estatística, pois possui relação direta com o Teorema Central do Limite.

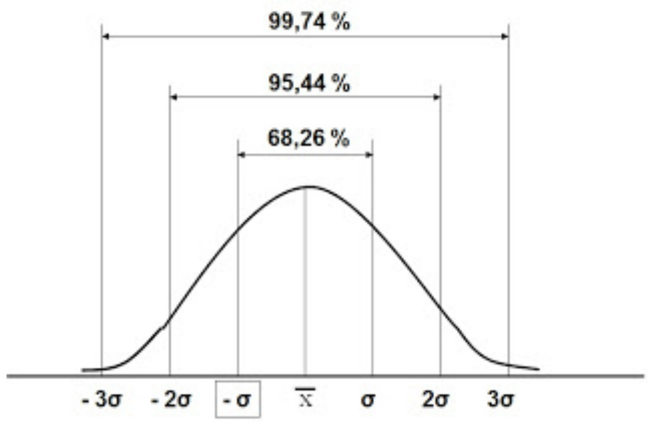
Propriedades:

- A variável "x" pode assumir qualquer valor no conjunto dos números reais.
- Seu gráfico possui formato de sino e a curva é totalmente simétrica.
- A média, a moda e a mediana possuem o mesmo valor.
- Em seu gráfico, a curva normal aproxima-se do eixo das abscissas infinitamente, mas sem alcançá-lo.
- Como a curva é simétrica, o valor de sua área é 1 e a probabilidade de ocorrer um valor menor que a média é o mesmo que ocorrer um valor maior que a média (0,5).
- A maioria dos valores agrupa-se em torno do centro (a média);
- Os valores vão se tornado cada vez menos prováveis, quanto mais distantes eles se encontram da média.

Função de Densidade da Probabilidade Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty$$

Onde:
μ = média
σ = desvio-padrão
π = constante aproximada por 3,1416
e = constante aproximada por 2,71828
x = qualquer valor da variável



Quando se tem uma variável com distribuição normal, é possível calcular a probabilidade de essa variável assumir um valor em um intervalo.

A Distribuição Normal Padronizada

A distribuição normal padronizada possui média 0 (zero) e desvio padrão 1. As probabilidades da curva normal padrão são obtidas em tabelas. A vantagem da curva padronizada é que os a média e o desvio padrão, estão definidos para qualquer escala de medida.

Para converter qualquer distribuição normal para a distribuição normal padrão deve-se utilizar a seguinte fórmula:

$$Z = \frac{x - \mu}{\sigma}$$

Tabela da Distribuição Normal Padronizada

Podemos encontrar vários tipos de tabelas que fornecem as probabilidades sob a curva normal padrão, sendo a mais utilizada a Tabela de faixa Central. Essa tabela fornece a área entre a média (0) e qualquer valor à direita (positivos). Como a curva é simétrica, podem-se obter também as probabilidades à esquerda da média (valores negativos).

A tabela fornece a área (probabilidade) no intervalo entre 0 e z_1, ou seja:

$$P(0 \leq z \leq z_1)$$

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig