1. Introdução

Neste tópico, serão abordados alguns conceitos primordiais como variável aleatória e seus tipos, definição de alguns tipos de distribuições e suas respectivas aplicações em casos onde deve-se levantar as probabilidades de eventos descritos por estas funções.

2. Tipos de Variáveis Aleatórias

Uma **variável aleatória** é uma variável cujo valor depende de fatores não determinísticos, ou melhor dizendo **probabilísticos**. Detalhando um pouco mais sobre as variáveis aleatórias, existem dois tipos mais comuns de variáveis aleatórias que podem ser aplicadas, sendo elas:

> **Discreta:** são as variáveis aleatórias que a distribuição de probabilidade assume apenas valores enumeráveis;
>
> **Contínua:** são as variáveis aleatórias que a distribuição de probabilidade assume valores contínuos dentro de um intervalo pré-determinado.

Para entender profundamente os conceitos de cada um dos tipos de variáveis aleatórias, deve-se definir as propriedades de distribuições de probabilidade para cada um dos casos.

3. Distribuições de Probabilidade

Uma **distribuição de probabilidade** é uma função que descreve o comportamento aleatório de um fenômeno dependente do acaso. A distribuição de probabilidade pode **modelar incertezas** e descrever fenômenos físicos, biológicos, econômicos, entre outros.

A representação matemática para a distribuição de probabilidade é feita utilizando o que chama-se de **função de probabilidade** (caso discreto) **função de densidade de probabilidade** (caso contínuo). Para cada um dos tipos de variáveis aleatórias, tem-se uma definição de densidade de probabilidade, conforme descritos a seguir:

> **Discretas:** Para o caso das variáveis aleatórias discretas a definição matemática para a função de probabilidade (também chamada de *função massa de probabilidade*), pode ser dada da seguinte forma: $F(X) = P(X = x)$

Ou seja, para o caso discreto, a variável aleatória é uma função que assume um valor real para cada elemento do espaço amostral. Partindo das definições gerais para probabilidade, os resultados a seguir são sempre válidos:

$$0 \leq P(X = x) \leq 1$$

$$P(X = x) \geq 0$$

$$\sum_{i=1}^{\infty} P(X = x_i) = 1$$

$$P(a \leq X \leq b) = \sum_{x=a}^{x=b} P(X = x)$$

> **Contínuas:** Para o caso das variáveis aleatórias contínuas a definição matemática para a probabilidade, P, utilizando a função de densidade de probabilidade p, pode ser dada da seguinte forma: $P(X) = \int p(x)\, dx$

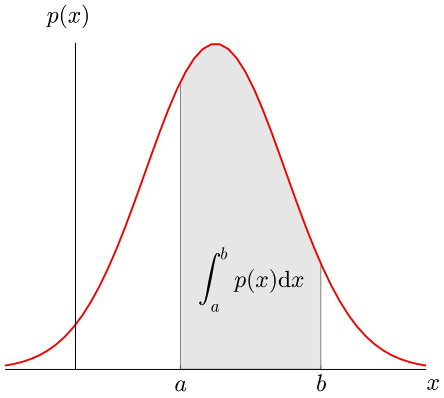
Da mesma forma que para o caso discreto, define-se algumas relações conhecidas para o caso das variáveis aleatórias contínuas:

$$p(x) \geq 0, \forall x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)\,dx$$

Uma outra forma de interpretarmos a probabilidade de uma dada distribuição entre o intervalo a e b, seria como a área abaixo ao gráfico da função de densidade de probabilidade:



Fonte : Thomas Haslwanter

3.1. Valor Esperado (Esperança) e Variância

Assim como no caso da Estatística Descritiva, no casos de modelos probabilísticos também existem parâmetros de posição e variabilidade utilizados para caracterizar uma distribuição de probabilidade:

> **Valor Esperado (Esperança):** O valor esperado seria o produto da variável aleatória x e sua respectiva probabilidade, funcionando como se fosse uma média ponderada para as probabilidades. O cálculo do valor esperado é definido da seguinte forma:

Caso Discreto: $$E[X] = \sum_{i}^{n} x_i P(X = x_i)$$

Caso Contínuo: $$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

> **Variância:** A variância é o valor que mede a variabilidade, ou seja o quão dispersão estão as probabilidades em relação ao valor esperado. A variância é definida da seguinte forma:

$$V[X] = \sigma^2 = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Onde, $E[X^2] = \sum_{i}^{n} x_i^2 P(X = x_i)$ e analogamente para o caso contínuo $E[X^2] = \int_{-\infty}^{\infty} x^2 p(x)dx$.

Exemplo de aplicação - Caso Discreto: Seja a variável aleatória X com distribuição abaixo, calcule $E[X]$ e $V[X]$:

> P(X = 1) = 0.3
> P(X = 2) = 0.4
> P(X = 3) = 0.2
> P(X = 4) = 0.1

No trecho de código abaixo exemplifica uma forma de ser implementado a resolução do exercício acima usando o *Python*:

```
# Vetor de eventos
X = [1, 2, 3, 4]

# Vetor de probabilidades
P = [0.3, 0.4, 0.2, 0.1]

# Cálculo do Valor esperado
esp = np.dot(X, P)
```

```
# Mostra o valor esperado
print("Valor esperado: ", np.round(esp, 2))

# Cálculo da variância
var = np.round(np.dot(np.power(X, 2), P) - np.power(esp, 2), 2)

# Mostra a variância
print("Variância:      ", var)
```

Exemplo de Aplicação - Caso Contínuo: A variável X tem função de densidade de probabilidade dada por:

$f(x) = \frac{x^2}{3}$, se $-1 \leq x \leq 2$, caso contrário seria 0.

Para o caso contínuo, precisa-se realizar o cálculo de uma integral, onde será utilizado uma função própria da biblioteca *SciPy*:

```
# Carrega a função quad para aproximar o valor da integral
from scipy.integrate import quad
```

A função *quad* irá aproximar o valor da integral ao valor calculado teórico, com uma margem de erro bem pequena. A implementação da resolução em código *Python* encontra-se a seguir:

```
# Função para a equação do valor esperado
def funcao_vlr_esperado(x):
    return x*(x*x)/3

# Cálculo da integral e o erro a partir da função anterior
esp, erro1 = quad(funcao_vlr_esperado, -1, 2)

# Print do valor esperado
print("Valor Esperado: ", esp)
print("Erro da Integral: ", erro1)

# Função para a equação do valor esperado x^2 a partir da função anterior
def funcao_variancia(x):
    return (x*x)*(x*x)/3

# Integral de x^2
esp_x2, erro2 = quad(funcao_variancia, -1, 2)

# Cálculo da variância
var = esp_x2 - esp*esp

# Print da variância
print("Variância:      ", var)
print("Erro da Integral: ", erro2)
```

4. Principais Distribuições

Existem algumas distribuições largamente utilizadas para o levantamento de probabilidade de eventos, sendo as principais delas descritas nos tópicos a seguir.

4.1 Distribuição de Bernoulli

A distribuição de Bernoulli é uma distribuição discreta para um espaço amostral $k \in \{0, 1\}$, baseando a probabilidade em **sucessos** e **falhas**, onde a probabilidade de sucesso de um evento ($k = 1$) é igual a p e a probabilidade de falha ($k = 0$) seria o valor complementar $1 - p$. A função que descreve a distribuição de Bernoulli pode ser definida como:

$$P(X = k) = p^k(1 - p)^{(1-k)}$$

Os valores para o valor esperado e a variância para a distribuição de Bernoulli são respectivamente:

Valor Esperado: $E[X] = p$;

Variância: $V[X] = p(1 - p)$.

Mas no caso da distribuição de Bernoulli, trata-se apenas para um evento isolado, como por exemplo o lançamento de uma moeda. Quando o problema envolve eventos **com repetições**, utiliza-se o caso geral da distribuição de Bernoulli que seria uma **Distribuição Binomial.**

4.2 Distribuição Binomial

Seja a variável aleatória baseado em n repetições de Bernoulli, temos que a definição da distribuição binomial é dada por:

$$P(X = k) = \binom{n}{k} p^k(1 - p)^{(n-k)} = \frac{n!}{k!(n - k)!} p^k(1 - p)^{(n-k)}$$

Os valores para o valor esperado e a variância para a distribuição binomial são respectivamente:

Valor Esperado: $E[X] = np$;

Variância: $V[X] = np(1 - p)$.

Exemplo de Aplicação: Em uma caixa há 8 bolas brancas e 4 pretas. Retira-se 5 bolas com reposição. Calcule a probabilidade de que:

A) saiam duas bolas brancas:

Uma forma de resolver este problema seria justamente desenvolver uma função em Python que realize o cálculo da probabilidade para a distribuição binomial:

```
# Carrega a função que calcula o fatorial
from math import factorial

# Cria uma função para o cálculo da probabilidade binomial
def binomial(k, n, p):
    C = (factorial(n)/(factorial(n - k)*factorial(k))) # Cálculo da combinação
    return C*np.power(p, k)*np.power(1 - p, n - k)     # retorna o valor da probabilidade
```

Definida a função, aplica-se aos conceitos propostos pelo exercício:

```
# Número de retiradas
n = 5

# Número de brancas
k = 2

# Probabilidade de uma bola branca
p = 8/12

# Mostra o resultado teórico
print("A probabilidade para este evento será: ", np.round(binomial(k, n, p), 3))
```

A biblioteca *SciPy* também tem uma aplicação direta para o cálculo pontual da probabilidade binomial utilizando uma função chamada *probability mass function*. Resolvendo o mesmo exercício com a função do *SciPy*:

```
# Carrega as funções para probabilidade binomial
from scipy.stats import binom
```

```
# Número de retiradas
n = 5

# Número de brancas
k = 2

# Probabilidade de uma bola branca
p = 8/12

# Mostra o resultado teórico
print("A probabilidade para este evento será: ", np.round(binom.pmf(k, n, p), 3))
```

B) saiam pelo menos 3 pretas:

Para este exemplo, como é buscado a probabilidade de pelo menos 3 bolas pretas para 5 lançamentos, pode ocorrer também os eventos 4 e 5 bolas pretas. Dessa forma, precisa ser trabalho com a probabilidade acumulada dos eventos, uma forma de implementar isso em *Python* foi desenvolvida no código abaixo:

```
# Número de retiradas
n = 5

# Probabilidade de uma bola preta
p = 4/12

# Calculando a probabilidade acumulada
prop = 0                          # Define como 0 o valor inicial da probabilidade
for k in range(3, 6):             # Desenvolve um laço entre 3 e 5 lançamentos
       prop += binomial(k, n, p)  # Calcula-se o valor da probabilidade e soma ao valor anterior

# Mostra o resultado da probabilidade
print("A probabilidade para este evento será: ", np.round(prop, 3))
```

Analogamente, o *SciPy* tem uma função chamada *cumulative distribution function* para estes casos de probabilidade acumulada. No caso do exemplo, seria mais conveniente calcular a probabilidade acumulada até duas retiradas de bolas e utilizar a **probabilidade complementar**, pois neste exemplo limita-se a 5 lançamentos mas poderiam ser milhares de lançamento. Então é uma aplicação mais simples trabalhar com a probabilidade complementar:

```
# Número de retiradas
n = 5

# Probabilidade de uma bola preta
p = 4/12

# Limite das retiradas
k = 2

# Calculando a probabilidade acumulada até 2 retiradas
prop = binom.cdf(k, n, p)

# Calcula a probabilidade complementar
prop_comp = 1 - prop

# Mostra o resultado da probabilidade complementar
print("A probabilidade para este evento será: ", np.round(prop_comp, 3))
```

4.3 Distribuição Poisson

Uma variável aleatória tem distribuição de Poisson quando podemos descrever um evento em relação a uma taxa/contagem de ocorrência, normalmente chamada de μ, sendo $\mu > 0$. Dessa forma a equação para a distribuição de Poisson será definida como:

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

Os valores para o valor esperado e a variância para a distribuição de Poisson serão respectivamente:

Valor Esperado: $E[X] = \mu$;

Variância: $V[X] = \mu$.

Exemplo de Aplicação: Em uma central telefônica chegam 300 ligações por hora. Sabendo que segue uma distribuição de Poisson, qual é a probabilidade de que:

A) Em um minuto não ocorra ligações?

Resolvendo o exemplo implementando uma função em *Python* para o cálculo da distribuição de *Poisson*:

```
# Definindo uma função para a distribuição de Poisson
def Poisson(k, mu):
    return np.exp(-mu)*(mu**k)/factorial(k)

# Definindo a taxa de ocorrência
mu = 5 # 300 chamadas/ 60 minutos = 5 chamadas por minuto

# Frequência procurada
k = 0   # Não ocorrer ligações

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(Poisson(k, mu), 3))
```

Analogamente, resolvendo o mesmo exemplo com biblioteca *SciPy* utilizando a função *pmf*:

```
# Carrega a função para distribuição de Poisson
from scipy.stats import poisson

# Definindo a taxa de ocorrência
mu = 5 # 300 chamadas/ 60 minutos = 5 chamadas por minuto

# Frequência procurada
k = 0   # Não ocorrer ligações

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(poisson.pmf(k, mu), 3))
```

B) Ocorra pelo menos 4 ligações?

Implementando novamente uma função em *Python* para o cálculo da distribuição de *Poisson*, neste caso utilizando de probabilidade acumulada:

```
# Definindo a taxa de ocorrência
mu = 5 # 300 chamadas/ 60 minutos = 5 chamadas por minuto

# Frequência procurada
k = 3  # Até 2 ligações para usar a probabilidade complementar (no range o 3 é excludente)

# Laço para o cálculo da probabilidade
prop = 0                     # Inicializa a probabilidade com zero
for i in range (0, k):       # Laço para o cálculo de probabilidade
    prop += Poisson(i, mu)   # Calcula o Poisson e soma com a probabilidade anterior

# Cálculo da probabilidade complementar
```

```
prop_comp = 1 - prop

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(prop_comp, 3))
```

Analogamente, resolvendo o mesmo exemplo com biblioteca *SciPy* utilizando a função *cdf*:

```
# Carrega a função para distribuição de Poisson
from scipy.stats import poisson

# Definindo a taxa de ocorrência
mu = 5 # 300 chamadas/ 60 minutos = 5 chamadas por minuto

# Frequência procurada
k = 2  # Até 2 ligações para usar a probabilidade complementar

# Cálculo da probabilidade acumulada até 2 ligações
prop = poisson.cdf(k)

# Cálculo da probabilidade complementar
prop_comp = 1 - prop

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(prop_comp, 3))
```

4.4 Distribuição Exponencial

Uma variável aleatória contínua tem uma distribuição exponencial quando queremos avaliar o tempo decorrido entre dois eventos consecutivos, diferente do Poisson que avalia de acordo com uma contagem de ocorrências em um espaço de tempo. A função densidade de probabilidade que descreve a distribuição exponencial pode ser descrita como:

$f(x) = \alpha e^{-\alpha x}, x \geq 0$

No caso para $x < 0$, a probabilidade de $f(x) = 0$

Os valores para o valor esperado e a variância para a distribuição exponencial serão respectivamente:

Valor Esperado: $E[X] = \frac{1}{\alpha}$;

Variância: $V[X] = \frac{1}{\alpha^2}$.

Exemplo de Aplicação: O intervalo de tempo, em minutos, entre emissões consecutivas de uma fonte radioativa é uma variável aleatória contínua que segue uma distribuição exponencial com parâmetro $\alpha = 0.2$. Qual a probabilidade de que ocorra uma emissão em um intervalo inferior a 2 minutos?

Resolvendo o exemplo com uma implementação em *Python*:

```
# Define a função densidade de probabilidade para a exponencial
def funcao_exp(x):
    alpha = 0.2                    # Define o parâmetro alpha
    return alpha*np.exp(-alpha*x)  # Devolve o valor da função densidade

# Integral da função de densidade
prop, erro = quad(funcao_exp, 0, 2)

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(prop, 3))
```

Para o caso de utilizar a biblioteca *SciPy*, como a integral pega o intervalo contínuo entre 0 e 2, basta utilizar da função *cdf* para resolver o exercício:

```
# Carrega as funções para a distribuição exponencial
from scipy.stats import expon

# Define o parâmetro alpha
alpha = 0.2

# Mostra o resultado
print("A probabilidade para este evento será: ", np.round(expon.cdf(x = 2,              # x é o limite superior
                                                         scale = 1/alpha), 3))  # scale representa o parâmetro alpha
```

4.5 Distribuição Uniforme

A distribuição uniforme é uma distribuição bem simples e não possui parâmetros, a única diferença é que só vai haver probabilidade para um determinado evento x, se $x \in [a, b]$. Dessa forma a equação de densidade de probabilidade para a distribuição uniforme é dado por:

$$f(x) = \frac{1}{b-a}, a \leq x \leq b$$

E para o caso de $x \notin [a, b]$, a função de densidade será $f(x) = 0$

Os valores para o valor esperado e a variância para a distribuição uniforme serão respectivamente:

Valor Esperado: $E[X] = \frac{a+b}{2}$;

Variância: $V[X] = \frac{(b-a)^2}{12}$.

Materiais Complementares

Documentação do SciPy;

Artigo *5 Probability distribution you should know as a data scientist* escrito por Harsh Maheshwari;

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.