Modelagem dos Dados de Entrada

Simulação Discreta

Filipe Saraiva

Conteúdo

Conteúdo

Introdução

Nessa aula serão discutidas as maneiras pelas quais dados de entrada em um projeto de simulação discreta são obtidos e pré-tratados.

Também serão apresentadas aplicações muito diretas dos conceitos de Estatística Descritiva vistos anteriormente.

Introdução

No geral, o procedimento de modelagem dos dados compreende três fases:

- Coleta de dados;
- Tratamento dos dados;
- Inferência

Essas 3 etapas serão discutidas nessa apresentação.

Conteúdo

Coleta de Dados

A coleta de dados para simulação discreta se inicia com a escolha das variáveis de entrada do sistema a ser simulado.

É importante ter muito clara a diferença entre **dados de entrada** e **dados de saída** – o que se pretende obter – da simulação.

Coleta de Dados

A maioria dos sistemas que procuramos modelar possui algum fenônemo aleatório que o governa. Exemplos:

- Tempo de operação em uma máquina - repetindo a mesma operação, mas levando tempos diferentes para processá-la;
- Filas de banco - clientes chegam em horários e quantidades diferentes;
- Reposição de refil de bebidas em bares - tempo que o consumidor leva para consumir uma bebida varia de maneira aleatória.

Apesar da natureza aleatória dessas variáveis, é possível prever seu comportamento probabilístico a partir da observação em tempos anteriores.

Coleta de Dados

Exemplo

Um gerente de um supermercado quer estudar o tamanho da fila nos caixas de supermercado. Quais variáveis deverão ser analisadas?

- Número de prateleiras no supermercado
- Os tempos de atendimento nos caixas
- O número de clientes em fila
- O tempo de permanência dos clientes no supermercado
- O tempo de chegada sucessiva dos clientes nos caixas

Coleta de Dados

Exemplo

Um gerente de um supermercado quer estudar o tamanho da fila nos caixas de supermercado. Quais variáveis deverão ser analisadas?

- Número de prateleiras no supermercado
- **Os tempos de atendimento nos caixas**
- O número de clientes em fila – **Essa medida é resultado!**
- O tempo de permanência dos clientes no supermercado
- **O tempo de chegada sucessiva dos clientes nos caixas**

Coleta de Dados

Definidas as variáveis de entrada, cabe agora ao pesquisador mensurar um número considerável desses valores com a finalidade de criar uma amostra representativa do sistema.

Em que pese essas variáveis serem aleatórias, uma medição correta e precisa de uma amostra será suficiente para encontrarmos um modelo probabilístico que rege o fenômeno.

Coleta de Dados

O trabalho de medição dos dados é um trabalho de campo:

- Ir ao local e, utilizando um cronômetro, medir o tempo em que o fenômeno acontece, anotar o valor obtido quando o fenômeno finaliza, zerar o cronômetro, fazer nova medição do fenômeno, etc.
- O tamanho das observações deve ser entre 100 e 200 amostras. Menor que 100 pode comprometer a observação do fenômeno probabilístico, enquanto maior que 200 não traz ganho significativo.
- Coletar e anotar as observações na ordem, para permitir análise de correlação.
- Deve-se ter clareza se o fenômeno varia conforme o dia, horário ou outra variável relacionada com o momento em que foi colhido. Estudos devem levar essa característica em conta.

Coleta de Dados

Para o estudo da entrada de dados, supomos a medição dos clientes que chegam ao supermercado durante um determinado horário. A tabela abaixo apresenta as medições, em segundos, dessa entrada, compondo uma amostra de 200 observações.

11	5	2	0	9	9	1	5	1	3
3	3	7	4	12	8	5	2	6	1
11	1	2	4	2	1	3	9	0	10
3	3	1	5	18	4	22	8	3	0
8	9	2	3	12	1	3	1	7	5
14	7	7	28	1	3	2	11	13	2
0	1	6	12	15	0	6	7	19	1
1	9	1	5	3	17	10	15	43	2
6	1	13	13	19	10	9	20	19	2
27	5	20	5	10	8	2	3	1	1
4	3	6	13	10	9	1	1	3	9
9	4	0	3	6	3	27	3	18	4
6	0	2	2	8	4	5	1	4	18
1	0	16	20	2	2	2	12	28	0
7	3	18	12	3	2	8	3	19	12
5	4	6	0	5	0	3	7	0	8
8	12	3	7	1	3	1	3	2	5
4	9	4	12	4	11	9	2	0	5
8	24	1	5	12	9	17	728	12	6
4	3	5	7	4	4	4	11	3	8

Coleta de Dados

Com os dados medidos, finaliza-se a etapa de coleta dos mesmos e passa-se então para o tratamento desses dados.

Conteúdo

Tratamento dos Dados

Nessa etapa iniciam-se um conjunto de análises para explorar o conjunto de dados e então compreender o fenômeno probabilístico expresso neles.

Tratamento dos Dados

Utilizando Estatística Descritiva, obtem-se as seguintes medições para o conjunto de dados:

Medidas de Centralidade	
Média	10,43
Mediana	5
Moda	3
Mínimo	0
Máximo	728
Medidas de Dispersão	
Amplitude	728
Desvio Padrão	51,41
Variância	2.643,81
Coeficiente de Variação	492,74%
Coeficiente de Assimetria	0,144

Tratamento dos Dados

Na análise, descobrimos que o mínimo é 0 (diferentes clientes entrando juntos) e o máximo é 728 (ou seja, mais de 12 minutos!). A média é 10,44 segundos.

Essa primeira avaliação permite verificar o quão discrepante alguns dados se encontram. No caso, aquele 728 é uma medida que deve permanecer no conjunto de dados ou ser removida?

Os dados discrepantes são chamados de *outliers* e merecem uma avaliação detida sobre o que fazer com eles.

Tratamento dos Dados

Antes de discutirmos *outliers*, vale a pena entender o impacto dele nas medições do conjunto de dados.

Abaixo temos algumas medições com e sem o valor 728 do conjunto:

	Com *outlier*	**Sem *outlier***
Média	10,43	6,829
Mediana	5	5
Máximo	728	43
Amplitude	728	43
Desvio Padrão	51,41	6,60
Variância	2.643,81	43,59
Coeficiente de Variação	492,74%	96,68%

Conteúdo

Outlier

Dados não usuais em conjuntos de dados são chamados *outliers*.

As razões mais comuns para o seu surgimento são erro na coleta de dados ou a ocorrência de um evento raro e inesperado.

Outlier

Razões para o surgimento de *outliers*:

- Erro na coleta de dados – falha em sensor que realiza coleta, mas também ocorre quando feita de maneira manual. Erro na atualização da tabela de dados, e outras. No geral, quando esse é o motivo do *outlier*, devemos remover a observação.
- Eventos raros – um *outlier* difícil de se lidar pois situações atípicas podem ocorrer durante as medições.

Em todo caso, a avaliação se o *outlier* deve ou não ser removido da base de dados é uma decisão importante que deve levar em conta o bom senso e honestidade sobre o fenômeno em estudo.

Outlier

Existem algumas técnicas que permitem uma avaliação se determinado dado é um *outlier* a ser removido ou não.

Fazendo Q_1 e Q_3 os respectivos quartis 1 e 3, temos o cálculo da **Amplitude Interquartil** (A) dada por:

$$A = Q_3 - Q_1$$

Outlier

A partir da Amplitude Interquartil, temos os seguintes conjuntos de valores discrepantes:

Outlier moderado	valor $< Q_1 - 1{,}5A$	valor $> Q_3 + 1{,}5A$
Outlier extremo	valor $< Q_1 - 3A$	valor $> Q_3 + 3A$

Outlier

Para o exemplo, tem-se:

- $Q_1 = 2$
- $Q_3 = 9$
- $A = Q_3 - Q_1 = 7$ (Amplitude Interquartil)

Assim, os valores discrepantes serão:

Outlier moderado	valor < -8,5	valor > 19,5
Outlier extremo	valor < -19	valor > 30

Outlier

Para o exemplo, teremos 11 valores como *outliers* moderados e 2 como *outliers* extremos (43 e 728).

Em nossa avaliação, apenas o valor 728 deveria ser removido, visto que ele difere em grande medida dos demais.

Entretanto, cabe o comentário: é importante avaliar bem se o *outlier* deve ser removido ou não.

Conteúdo

Análise de Correlação

Removido os *outliers*, é necessário ainda fazer uma avaliação se o conjunto de dados contém valores independentes ou não.

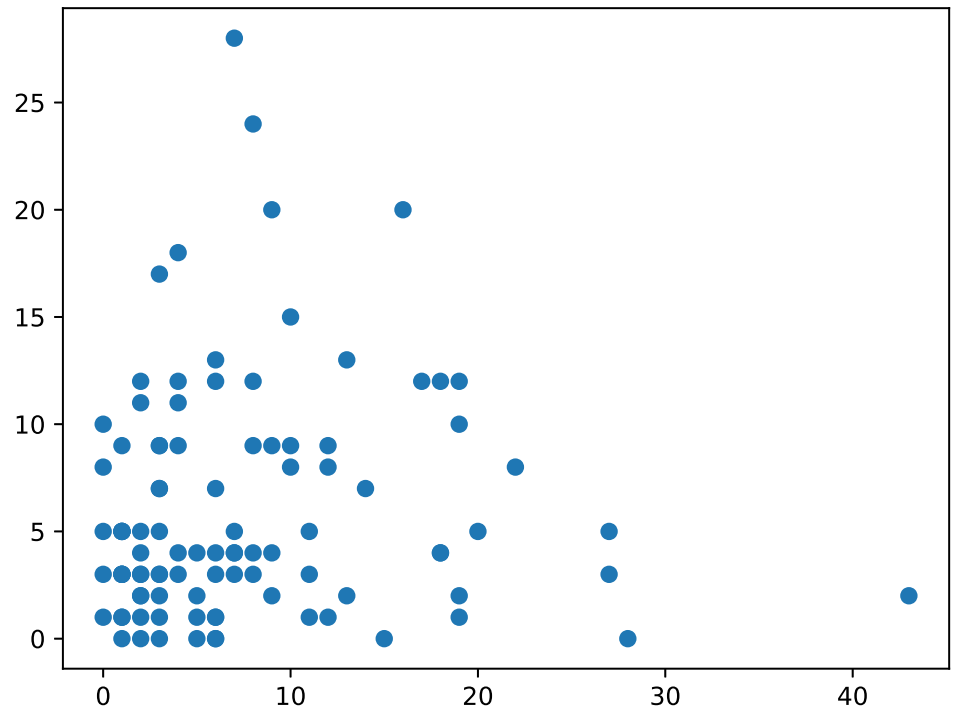
Essa situação em geral não é válida quando os dados apresentam uma “curva de aprendizado”. Nesses casos, significa que a obtenção dos dados foi realizada em alguma situação onde o operador foi “aprendendo” a tarefa ao longo do tempo, não demonstrando o tempo real de execução da tarefa.

Análise de Correlação

Para fazer essa análise, basta criar um gráfico de dispersão com o conjunto de dados, na ordem em que foram auferidos, e verificar se ele está com os pontos bem dispersos ou se há alguma aparência de "evolução" neles.

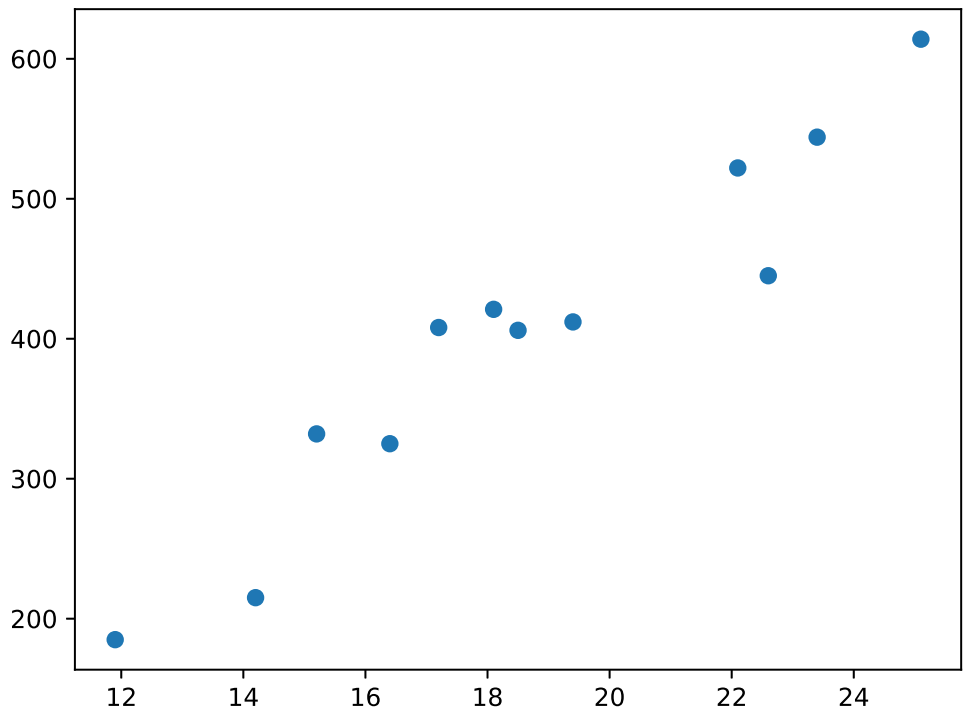
Sendo as observações os valores expressos por x_1, x_2, x_3, x_4, ..., $x_{(n-1)}$, x_n, o gráfico a ser criado deve ser feita pelos pares (x_1, x_2), (x_3, x_4), ..., $(x_{(n-1)}, x_n)$.

Análise de Correlação



Acima temos o gráfico de dispersão para o conjunto de dados que trabalhamos no exemplo. Ele não aparenta ter alguma evolução.

Análise de Correlação



No exemplo acima temos um gráfico que apresenta um tipo de “evolução”, que seria a curva de aprendizado.

Análise de Correlação

Em Python, podemos utilizar o pacote **matplotlib.pyplot** para gerar esse gráfico:

```
import matplotlib.pyplot as plt
# sendo x e y as listas com as abscissas e ordenadas do gráfico
plt.plot(x, y, 'o')
plt.show()
```

Conteúdo

Inferência

Parte-se agora para a última etapa do pré-processamento dos dados, a Inferência.

Nessa etapa, tenta-se encontrar o modelo probabilístico que mais se adequa aos dados mensurados, de forma que o comportamento das variáveis de entrada seja descoberto.

Conteúdo

Histograma

O primeiro passo da análise para inferência é criar um histograma do conjunto de dados.

Um histograma é um gráfico de frequência que mostra como uma população de dados está distribuída – ou seja, ele mostra quantas vezes temos repetido um determinado valor ou intervalo de valores no conjunto.

Histograma

Para iniciar a análise, precisamos calcular o número de classes a serem utilizadas no histograma. A fórmula para esse cálculo é dada por:

$$K = 1 + 3{,}3\log_{10} n$$

Onde:

- K – número de classes (portanto, deve ser arredondada para um inteiro)
- n – número de observações na amostra

Histograma

Para nosso exemplo, teremos:

$K = 1 + 3{,}3\log_{10} n$
$K = 1 + 3{,}3\log_{10} 199$
$K = 8{,}59 \approx 9$

Lembrando que como tiramos o *outlier*, agora temos 199 observações.

Histograma

Em Python utilizaremos a biblioteca **math** e a função padrão **round** para realizar esse cálculo da seguinte forma:

```
import math
round(1 + 3.3 * math.log10(199))
```

Histograma

Com o número de classes sendo igual a 9, procedemos para calcular o tamanho do passo para o intervalo de valores em classe. Para tanto, dividimos a amplitude pelo número de classes:

$$h = \frac{A}{K}$$

Onde:

- h – tamanho de cada classe
- A – amplitude
- K – número de classes

Histograma

Para o exemplo fica:

$$h = \frac{A}{K}$$
$$h = \frac{43}{9}$$
$$h = 4{,}8$$

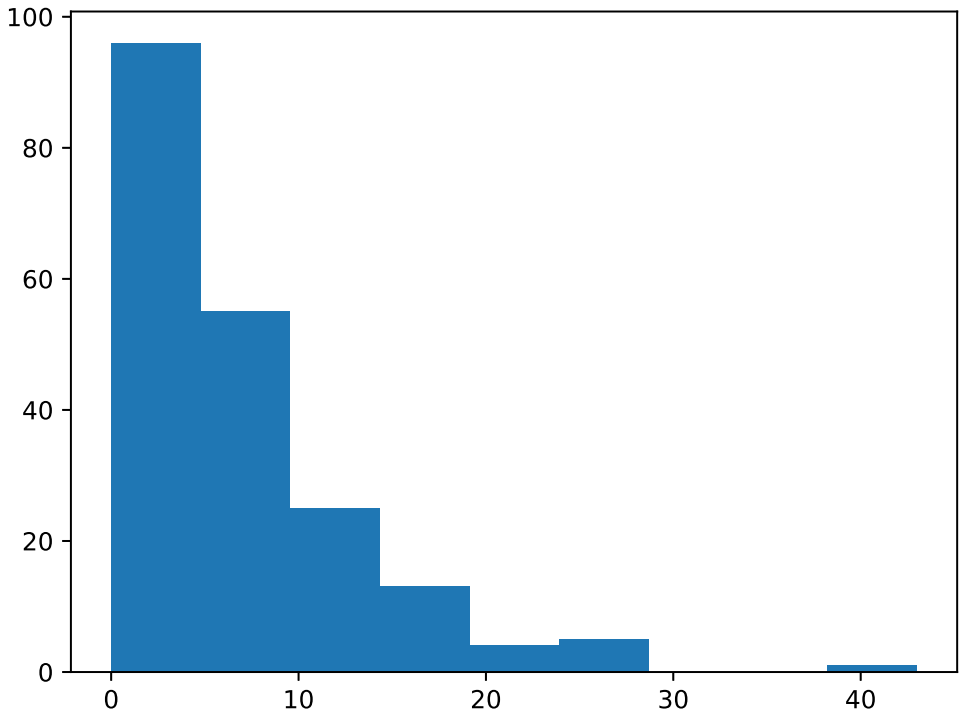
Ou seja, teremos portanto 9 classes com intervalos de valores com passos de 4,8 para cada.

Histograma

A tabela abaixo apresenta as classes, seus respectivos intervalos de valores, e a frequência de cada classe no conjunto de dados.

Classes	Intervalos	Frequência
1	valor $\leq$ 4,8	96
2	4,8 < valor $\leq$ 9,6	55
3	9,6 < valor $\leq$ 14,3	25
4	14,3 < valor $\leq$ 19,1	13
5	19,1 < valor $\leq$ 23,9	4
6	23,9 < valor $\leq$ 28,7	5
7	28,7 < valor $\leq$ 33,4	0
8	33,4 < valor $\leq$ 38,2	0
9	38,2 < valor	1

Histograma



A figura acima apresenta o histograma para o conjunto de dados do exemplo trabalhado.

Histograma

A exemplo dos gráficos anteriores, também utilizaremos o **matplotlib.pyplot** para gerar o gráfico de histograma. Para tanto, precisamos além dos dados, calcular o número de classes e passá-lo como argumento na função.

```
import matplotlib.pyplot as plt
# sendo x o conjunto de dados e k o número de classes
plt.hist(x, bins=k)
plt.show()
```

Conteúdo

Teste de Aderência

Com o histograma criado, a dúvida que fica é: “será que os dados podem ser modelados a partir de alguma distribuição de probabilidades específica?”

Se isso for possível, o fenômeno modelado pode ser modelado utilizando essa função de probabilidade, e ela pode ser utilizada para gerar dados de entrada.

Para verificar se o fenômeno medido é compatível com alguma distribuição, fazemos os **Testes de Aderências**.

Teste de Aderência

Os testes de aderência nada mais são do que verificar se o histograma é similar a alguma função de probabilidade conhecida.

Há 2 maneiras de realizar essa verificação: a partir de uma análise “visual” do histograma e dos gráficos de distribuições, e a partir de métodos matemáticos para esse teste.

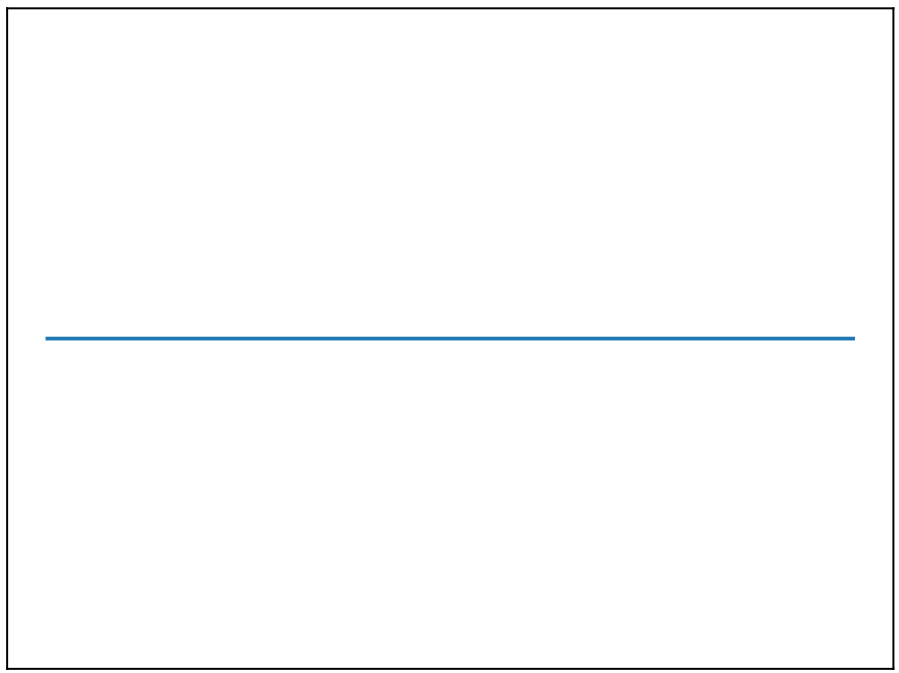
Teste de Aderência

Sobre o método gráfico, precisamos relembrar as principais funções de distribuição probabilidade utilizadas. Veremos as seguintes:

- Uniforme
- Exponencial
- Normal
- Lognormal
- Triangular

Nos slides a seguir, a **Função** significa a Função Densidade de Probabilidade de cada distribuição.

Teste de Aderência – Distribuição Uniforme



Parâmetros:

a - menor valor;

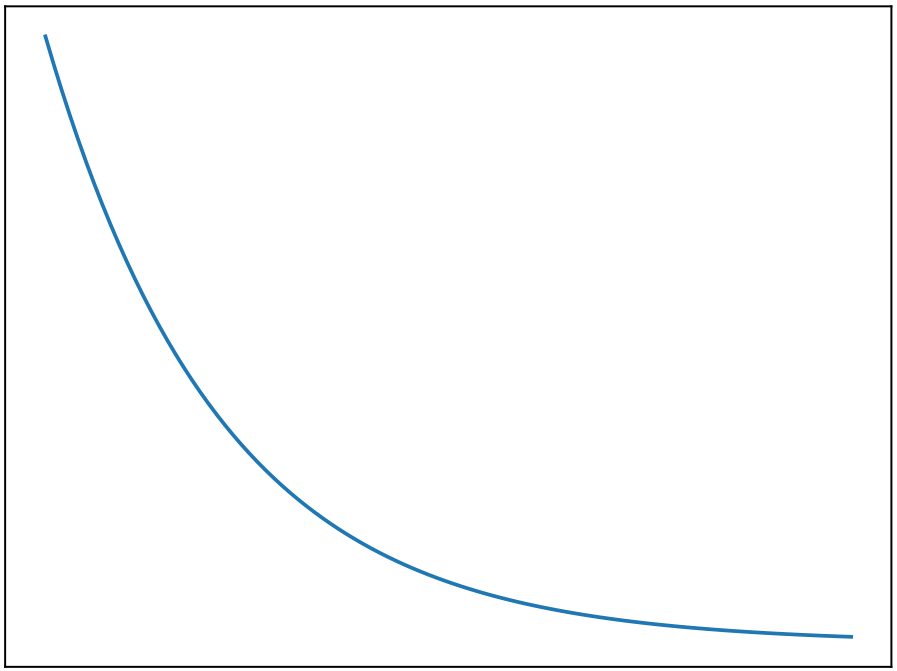
b - maior valor.

Função: $\frac{1}{b-a}$

Média: $\frac{a+b}{2}$

Variância: $\frac{(b-a)^2}{12}$

Teste de Aderência – Distribuição Exponencial



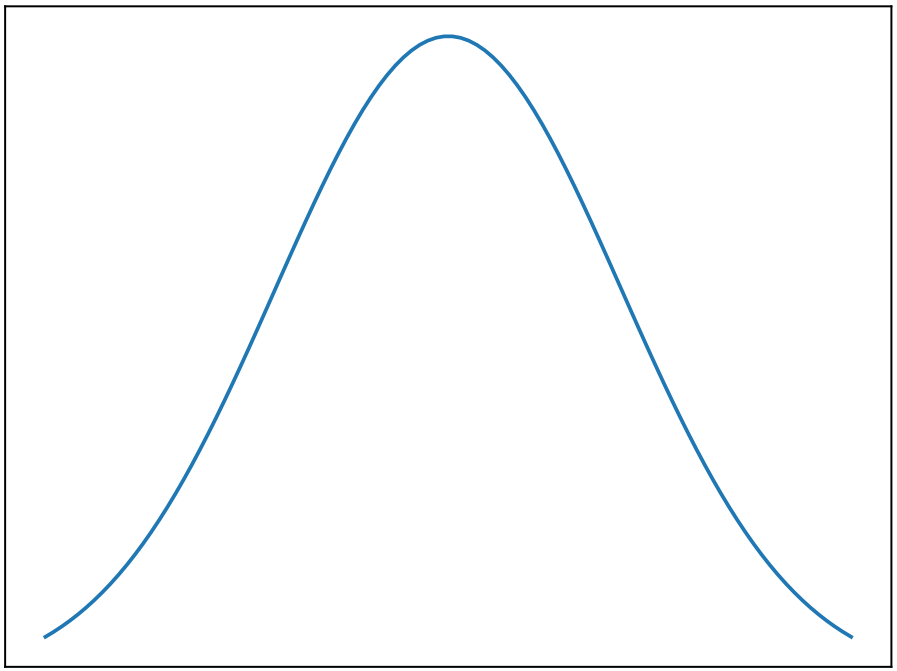
Parâmetros:
λ - taxa de ocorrências

Função: $\lambda e^{-\lambda x}$

Média: $\frac{1}{\lambda}$

Variância: $\frac{1}{\lambda^2}$

Teste de Aderência – Distribuição Normal



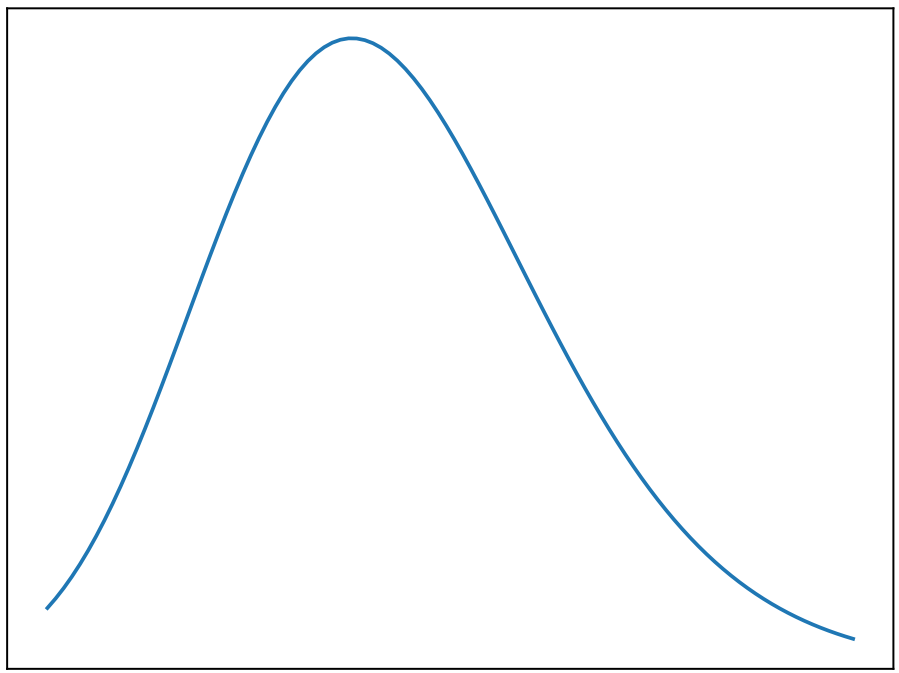
Parâmetros:
σ^2 - variância
μ - média

Função: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Média: μ

Variância: σ^2

Teste de Aderência – Distribuição Lognormal



Parâmetros:

σ - forma ou dispersão

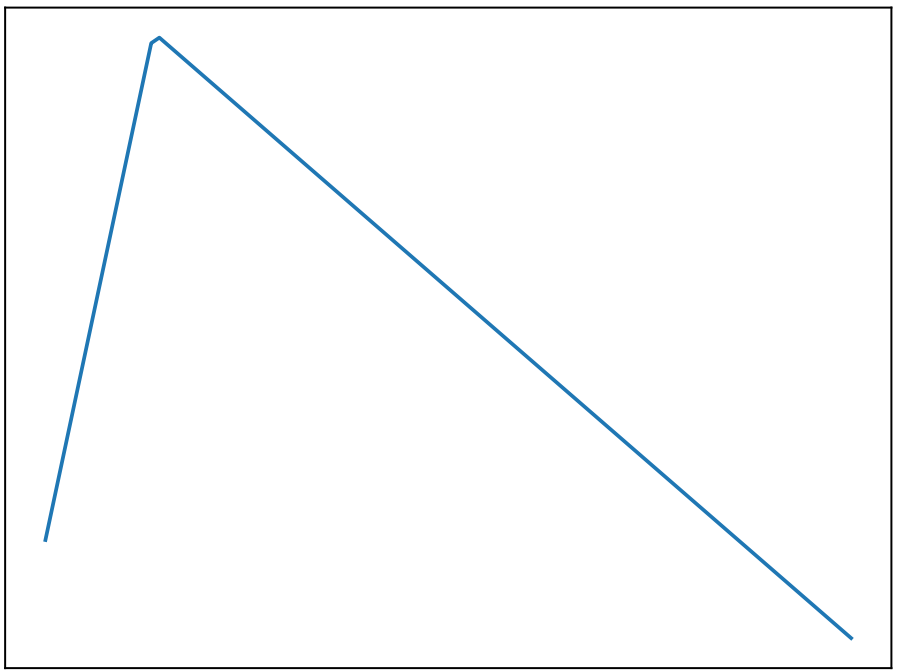
μ - escala ou posição

Função: $\frac{1}{\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$

Média: $e^{\mu+\frac{\sigma^2}{2}}$

Variância: $e^{2\mu+\sigma^2}(e^{\sigma^2})$

Teste de Aderência – Distribuição Triangular



Parâmetros:

a - menor valor

b - maior valor

m - moda

Onde: $a < m < b$

Função:

$\frac{2(x-a)}{(m-a)(b-a)}$, para $a \leq x \leq m$

$\frac{2(b-x)}{(b-a)(b-a)}$, para $m \leq x \leq b$

Média:

$\frac{a+m+b}{3}$

Variância:

$\frac{a^2+m^2+b^2-ma-ab-mb}{18}$

Teste de Aderência

Conhecida as principais distribuições, o teste de aderência visual consistirá portanto de verificar visualmente qual delas consegue melhor representar o histograma obtido do fenômeno em medição.

Entretanto, é importante notar que esse tipo de abordagem é sujeito a falhas e nem sempre preciso.

Teste de Aderência

Para uma maior precisão no teste de aderência é recomendado aplicar algum método matemático/estatístico para a verificação.

A literatura traz alguns, e nessa aplicaremos um deles: o **Teste de Kolmogorov-Smirnov** (KS).

Esse teste consiste em observar a distância máxima entre a função acumulada das observações e a função acumulada de algum modelo teórico de uma distribuição. Caso essa distância esteja acima de um limiar, pode-se considerar que a distribuição observada é aderente à distribuição teórica comparada.

Teste de Aderência

Para realizar o teste KS é necessário criar uma tabela com as seguintes colunas:

1. **Valor observado** – identificar quais são os valores presentes na base de dados e ordená-los em ordem crescente;
2. **Frequência observada** – contabilizar quantas vezes cada valor aparece na base de dados;
3. **Frequência acumulada observada** – somar as frequências observadas a cada linha, tanto do valor na linha quanto dos valores menores;
4. **Frequência acumulada observada normalizada** – divide-se a frequência acumulada observada pelo número total de dados;

Ao final dessa primeira parte teremos calculado a frequência de elementos observados na base e também a frequência normalizada.

Teste de Aderência

Para realizar o teste KS é necessário criar uma tabela com as seguintes colunas (continua):

5. **Frequência teórica esquerda** – a partir de uma função de distribuição teórica, calcula-se qual o valor dessa função para a entrada do **Valor observado** presente na linha.
6. **Frequência teórica direita** – o valor da função para a entrada do **Valor observado** da linha seguinte.
7. **Distância para a esquerda** – o valor absoluto da diferença entre a **frequência teórica esquerda** e a **frequência acumulada observada normalizada**.
8. **Distância para a direita** – o valor absoluto da diferença entre a **frequência teórica direita** e a **frequência acumulada observada normalizada**.
9. **Maior distância** – o maior valor entre as distâncias esquerda e direita para cada linha.

Teste de Aderência

Utilizando o exemplo da aula, teríamos as seguintes primeiras linhas na tabela:

V.O.	**F.O.**	**F.A.O.**	**F.A.O.N.**	$\mathbf{F}_{esq}$	$\mathbf{F}_{dir}$	$\mathbf{D}_{esq}$	$\mathbf{D}_{dir}$	**D**
0	13	13	0.07	0	0.14	0.07	0.07	0.07
1	23	36	0.18	0.14	0.25	0.04	0.08	0.08
2	18	54	0.27	0.25	0.35	0.02	0.08	0.08
3	26	80	0.4	0.35	0.44	0.05	0.04	0.05
...	...	...	...	...	...	...	...	...

Teste de Aderência

Sobre alguns dos dados que são obtidos através de fórmulas matemáticas, temos:

Frequência Acumulada Observada Normalizada (F.A.O.N.) – divide-se F.A.O. pelo número total de elementos da base (no exemplo, 199).

Frequência Teórica Esquerda e Direita (F_{esq} e F_{dir}) – primeiro é necessário escolher qual distribuição será comparada com os dados observados.

Para esse exemplo na tabela, decidiu-se comparar com a distribuição exponencial.

Para fazer o cálculo das frequências teóricas, é necessário conhecer a função acumulada da distribuição. Para a distribuição exponencial, essa função é dada por $1 - e^{-\lambda x}$, onde λ é o inverso da média da distribuição.

Teste de Aderência

Frequência Teórica Esquerda e Direita (F_{esq} e F_{dir}) (continua)

Portanto, calculando a média da distribuição (6,83) teremos $\lambda = 0,146$. Assim, a fórmula será $1 - e^{-0,146x}$.

F_{esq} será a aplicação do valor observado (primeira coluna) nessa função, enquanto F_{dir} será a aplicação do próximo valor observado.

D_{esq} e D_{dir} – são respectivamente |F_{esq} - F.A.O.N.| e |F_{dir} - F.A.O.N|

D – é $max(D_{esq}, D_{dir})$.

Teste de Aderência

Finalmente, para verificar o teste de aderência, utilizamos o maior valor de **D** e verificamos a tabela do KS para saber como calcular o valor D_{crtico} dado um nível de significância α.

Na tabela do KS o valor a ser buscado é relacionado com o α e o número de observações. Fazendo $\alpha = 0.05$ e lembrando que temos ($n =$) 199 observações, a tabela nos indicará que o valor D_{crtico} deve ser calculado a partir da fórmula $\frac{1{,}36}{\sqrt{n}}$.

Assim: $D_{crtico} = \frac{1{,}36}{\sqrt{199}} = 0,0964$.

Como esse valor é maior que o maior valor de **D**, então a distribuição é aderente ao conjunto de dados.

Teste de Aderência

Atenção

A Tabela do KS comentada no slide anterior está disponível como material complementar da aula.

Conteúdo

Conclusões

- Tratamento de dados em simulação discreta é essencial para realização das simulações;
- Os dados precisam ser medidos em campo ou coletados a partir de tabelas, e em seguida é necessário encontrar uma distribuição probabilística que modele de que maneira o fenômeno se comporta;
- Com os histogramas e o teste de aderência via KS, é possível verificar se determinados dados é condizente com alguma distribuição probabilística.

Modelagem dos Dados de Entrada

Simulação Discreta

Filipe Saraiva