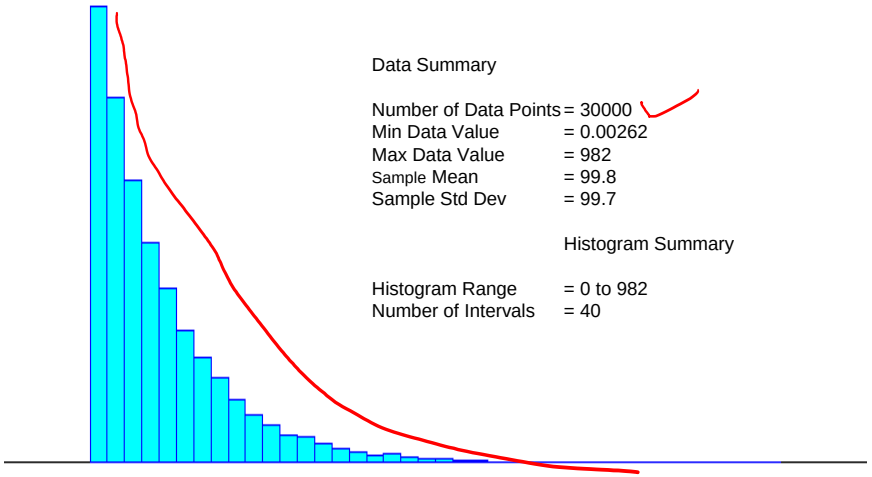
TRATAMENTO DE DADOS

Prof. Dr. Joao Ferreira Netto

Agosto de 2020

Processos estocásticos ⇒ "probabilidades"

MOTIVAÇÃO PARA A ANÁLISE DE DADOS



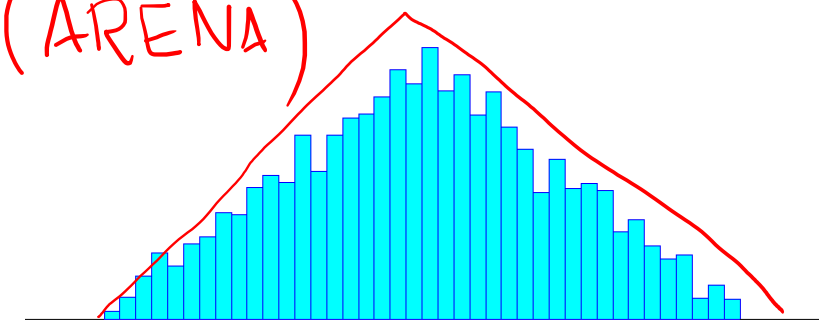
Data Summary

Number of Data Points	= 30000
Min Data Value	= 0.00262
Max Data Value	= 982
Sample Mean	= 99.8
Sample Std Dev	= 99.7

Histogram Summary

Histogram Range	= 0 to 982
Number of Intervals	= 40

Input Analyzer (ARENA)

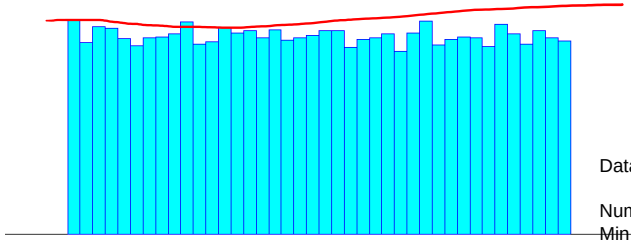


Data Summary

Number of Data Points	= 5000
Min Data Value	= 2.62
Max Data Value	= 196
Sample Mean	= 99.9
Sample Std Dev	= 41

Histogram Summary

Histogram Range	= 2 to 197
Number of Intervals	= 40



Data Summary

Number of Data Points	= 30000
Min Data Value	= 1.18e-005
Max Data Value	= 200
Sample Mean	= 99.6
Sample Std Dev	= 57.8

Histogram Summary

Histogram Range	= 0 to 200
Number of Intervals	= 40

Objetivo

intervalos entre as chegadas

PROCEDIMENTO PARA O TRATAMENTO DE DADOS

estatística descritiva

Trata-se de uma seleção de técnicas e conceitos já consagrados na Estatística, dispostos numa seqüência simples tal que auxilia o analista no tratamento de dados para simulação discreta.

Os conceitos e técnicas estatísticas são apresentados, bem como sua aplicação utilizando programa MINITAB. As referências bibliográficas para a montagem que fundamentam o procedimento proposto foram: Scheffé(1959), Peres(1986), Bussab(1988 e 1990) e Botter(1996).

Dados obtidos a partir de medição ou de bases de dados existentes

⇒ Big Data

- **Quem solicita o trabalho acredita que conhece bastante o problema real e os condicionantes gerais do mesmo (o que não é sempre verdade);** Objetivo

- **Quem elabora o modelo de simulação, em geral, solicita uma grande massa de dados, com informações sobre o problema a ser solucionado;**

- **As informações nem sempre estão disponíveis e, às vezes, nem foram coletadas ou armazenadas em meios facilmente acessíveis.**

Automação ⇒ sensores ✓

Obtenção dos Dados Necessários

- **Medição;**
- **Consulta a bancos de dados existentes;**
- **Consulta a informações externas ao ambiente onde o problema está inserido.**

Os problemas mais comuns com os Dados

•**quem montou o banco de dados não montou um manual de instruções para quem vai anotar os dados e depois inseri-los no Banco;**

•**pessoas de diferentes setores anotam esses dados e, ocorrendo dúvidas, os dados não são anotados, podendo até mesmo ocorrer a anotação dos dados por "estimativa" do valor que o processo resultar;**

falta de uniformização

•**erros no preenchimento de tabelas também são bastante comuns.**

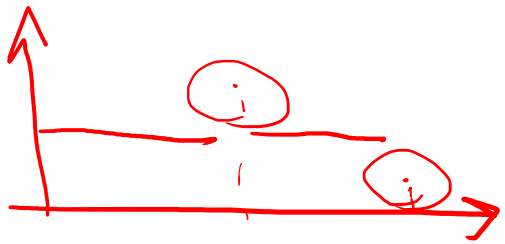
Procedimento para o Tratamento de Dados em Simulação Discreta

→ objetivo

→ conhecer o BD / medir os dados (registrar)

- **Ordenação dos Dados;** (temporal)
- **Avaliação Descritiva;**
- **Visualização dos Dados;** (gráficos)
- **Limpeza dos Dados;** (retirada de outliers)
- **Agrupamento;** (clusterização)
- **Seleção da Distribuição de Probabilidade.**

MINITAB



Resumo das Ações para a Limpeza e Análise de dados

Passo	Metodologia/Programa Utilizados	Análises
1 – Ordenação dos dados	Excel ou MINITAB	Erros ou omissões de preenchimento; campo de variação dos dados
2 – Avaliação Descritiva	Medidas de Posição e Dispersão/MINITAB	Para cada variável independentemente: ▪Comparação da média com a média aparada ▪Comparação da média, moda e mediana ▪Comparação da amplitude e intervalo entre quartis ▪Avaliação do Coeficiente de Variação
3 – Visualização	Histogramas ou Gráfico de Barras, Gráfico de Freqüência Acumulada, Diagrama de Dispersão Medidas de Associação/ MINITAB	Para cada variável independentemente: ▪Avaliar assimetria e achatamento ▪Avaliar concentrações de dados nas classes de freqüência. Para pares de variáveis: ▪Visualizar a associação e calcular correlação linear.

Resumo das Ações para a Limpeza e Análise de dados

4 – Limpeza de Dados	Gráfico de Caixas, Discrepâncias Bidimensionais/MINITAB Boxplot	Para cada variável independentemente: ▪Consultar quem conhece o processo e com base nos pontos discrepantes apontados pelo gráfico de caixas, eliminá-los ou mantê-los Para pares de variáveis: ▪Avaliar gráficos de dispersão em busca de pontos que induzem a avaliações de associações erradas
5 – Agrupamento	Gráfico de Caixas, ANOVA e Análise de Agrupamentos/MINITAB Clusterização	Para os diversos fatores existentes, associados a cada variável: ▪Verificar se os gráficos de caixa por fator são diferentes entre si ; ▪Efetuar comparação de médias pela ANOVA. Para as diversas variáveis associadas a cada elemento da amostra: ▪Proceder a uma Análise de Agrupamento.
6 – Seleção da Distribuição de Probabilidades	Gráfico de Probabilidades e Testes de Aderência/ MINITAB ou INPUT ANALYSER	Para cada variável independentemente: ▪Buscar uma distribuição teórica que seja aderente aos dados “limpos” e “agrupados” ou adotar a distribuição empírica.

Ordenação dos Dados

É muito simples;

É de fundamental importância para o analista;

Estabelece o reconhecimento dos limites das variáveis.

Ordenação dos Dados

Identifica as discrepâncias, como por exemplo:

Mistura de dados numéricos com alfanuméricos, lacunas em dados emparelhados; valores extremos incompatíveis

Avaliação Descritiva

- **Medidas de Ordem (máximo e mínimo);**
- **Separatrizes (Quartis, Descis e Pertencis);**
- **Medidas de Tendência Central (Média, Média Aparada, Mediana e Moda);**
- **Medidas de Dispersão (Amplitude, Intervalo Inter-Quartil, Variância, Desvio- Padrão e Coeficiente de Variação).**

Avaliação Descritiva
Comparações Importantes

- **Média com a Média Aparada - efeitos dos extremos;**

- **Média, Moda e Mediana - assimetrias, achatamento;**

- **Amplitude e Intervalo Inter-Quartis - efeito dos extremos;**

- **Coeficiente de Variação - tamanho do desvio-padrão em relação à média.**

Visualização dos Dados

Histograma;

Gráfico de freqüência acumulada;

Diagrama de dispersão e a correspondente medida de associação.

Visualização dos Dados
Análises Relevantes

Identificação de 2 ou mais “modas”;

Comportamentos do Gráfico de freqüência acumulada - primeira identificação da distribuição;

Possibilidade de eliminação de uma variável, se houver grande associação.

Exemplo de Acompanhamento no. 1

Tabulação de dados de duração de viagens (tempo em horas)

Tempo	Freqüência	Tempo	Freqüência
0	55	30	2
2	1	35	2
4	1	40	1
5	59	44	1
6	2	50	1
7	3	57	1
8	2	60	1
9	1	70	2
10	28	75	2
13	2	95	1
15	10	660	1
20	6	1245	1
28	1		
N =	187		

Exemplo de Acompanhamento no. 1

Com a simples tabulação dos dados detectou-se:

•em 55 viagens encontra-se registrado o tempo zero, que é um valor claramente impossível;

•percebe-se a existência de um número excessivo de viagens com tempo de duração múltiplo de 5: 5, 10, 15 horas, etc. Existe uma tendência observada na população em geral em arredondar valores numéricos para múltiplos de 5;

•dois valores apresentam-se muito superiores aos demais (660 e 1245).

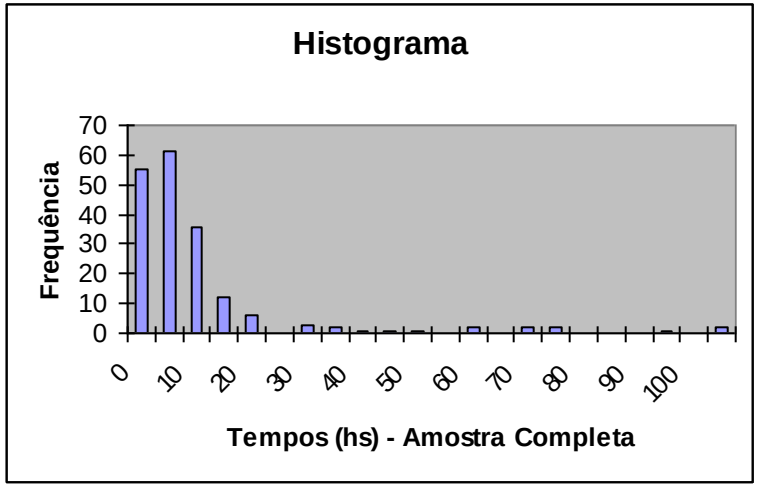
Exemplo de Acompanhamento no. 1

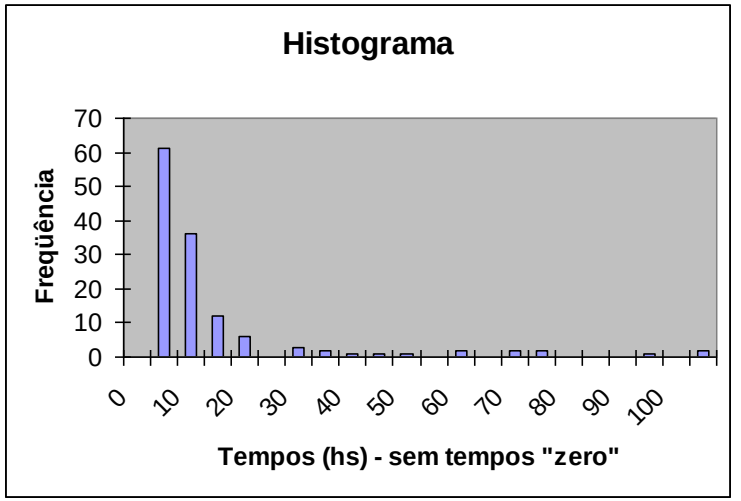
Medidas descritivas para a variável tempo

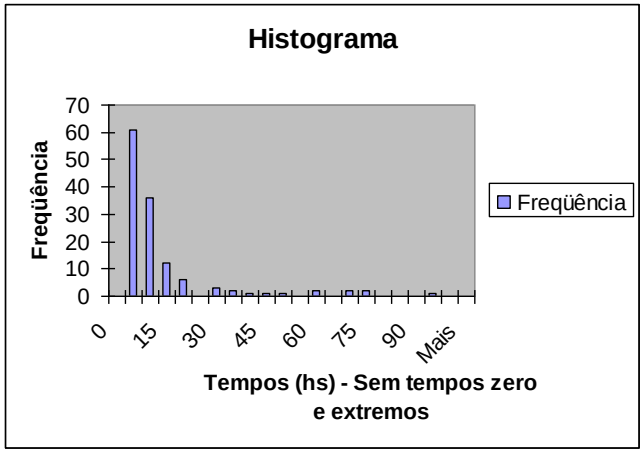
Medida Descritiva	Amostra Completa	Excluindo o valor 1245	Excluindo os valores 1245 e 660
Média	27,5	18,2	13,3
Média aparada de 10%	11,1	10,7	10,3
Mediana	7,5	7,0	7,0
Desvio-padrão	121,8	58,8	16,4
Q_1	5,0	5,0	5,0
Q_3	14,5	13,0	13,0
N	132	131	130

Exemplo de Acompanhamento no. 1

Histogramas para a variável tempo







Limpeza dos Dados

Gráfico de caixas;

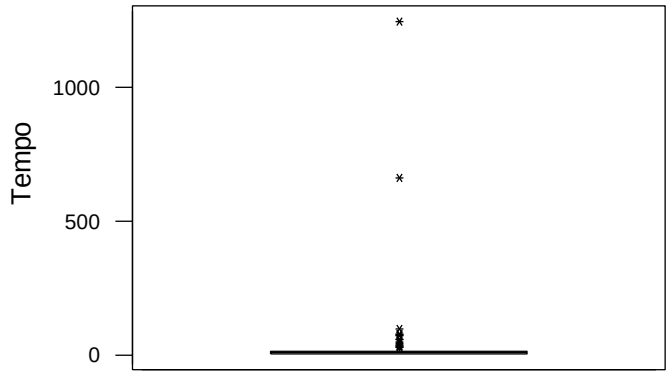
Investigar os pontos amostrais fora do intervalo e retirá-los após uma análise;

Valores discrepantes bidimensionais.

Gráfico de caixas

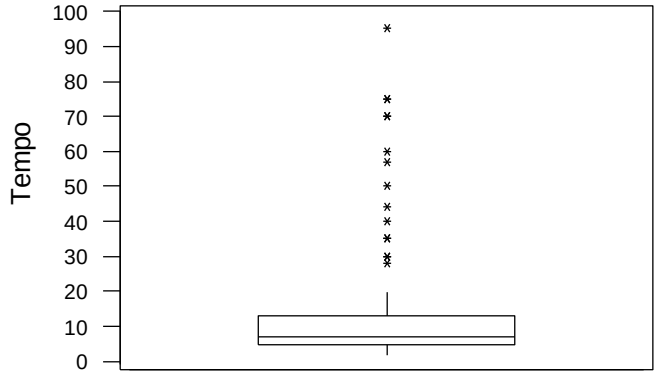
Para o exemplo de Acompanhamento no. 1

Gráfico de caixas da variável TEMPO - Amostra Completa



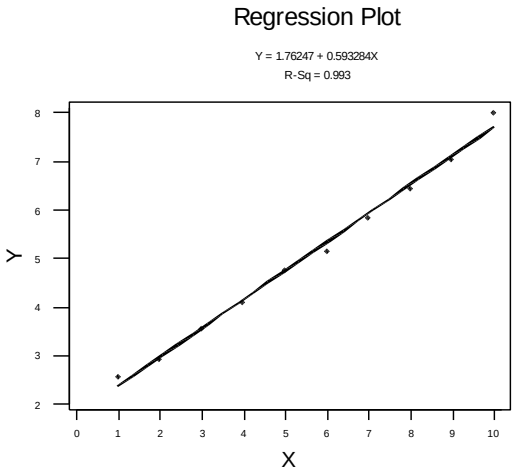
foram excluídas as observações iguais a zero

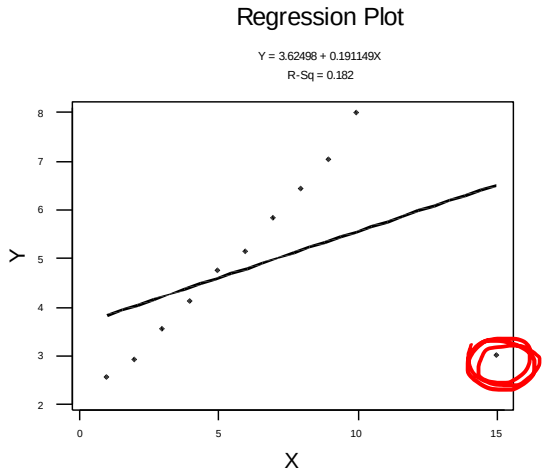
Gráfico de caixas da variável TEMPO
omitindo os valores 1245 e 660



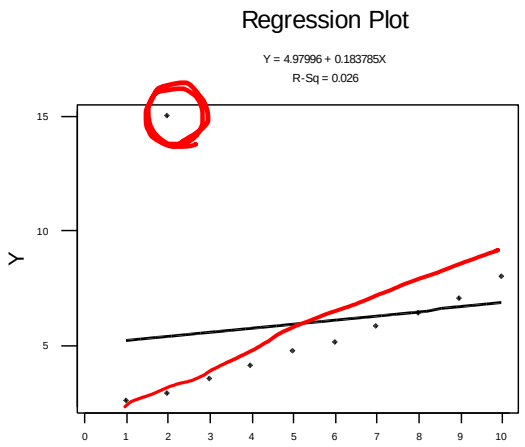
foram excluídas as observações iguais a zero

Valores discrepantes bidimensionais

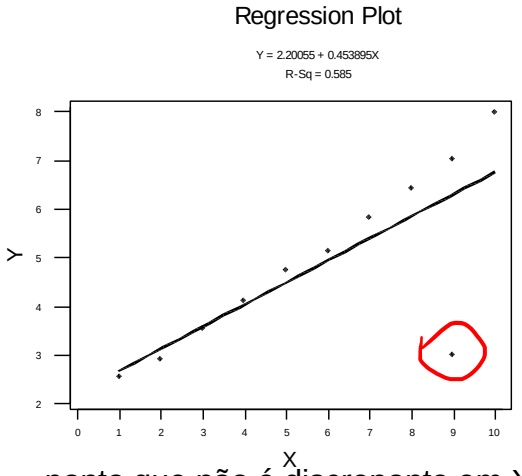




Ponto que é discrepante em X, não é em Y e é bidimensional



Ponto que é discrepante em Y, não é em X e é bidimensional



ponto que não é discrepante em X, não é em Y e é bidimensional

Agrupamento

A investigação da existência de grupos dentro dos dados amostrais coletados pode ser feita das seguintes formas:

- **Investigação por meio de gráfico de caixas;**
- **Análise de variância;** ANOVA
- **Análise de agrupamentos.** Clusterização

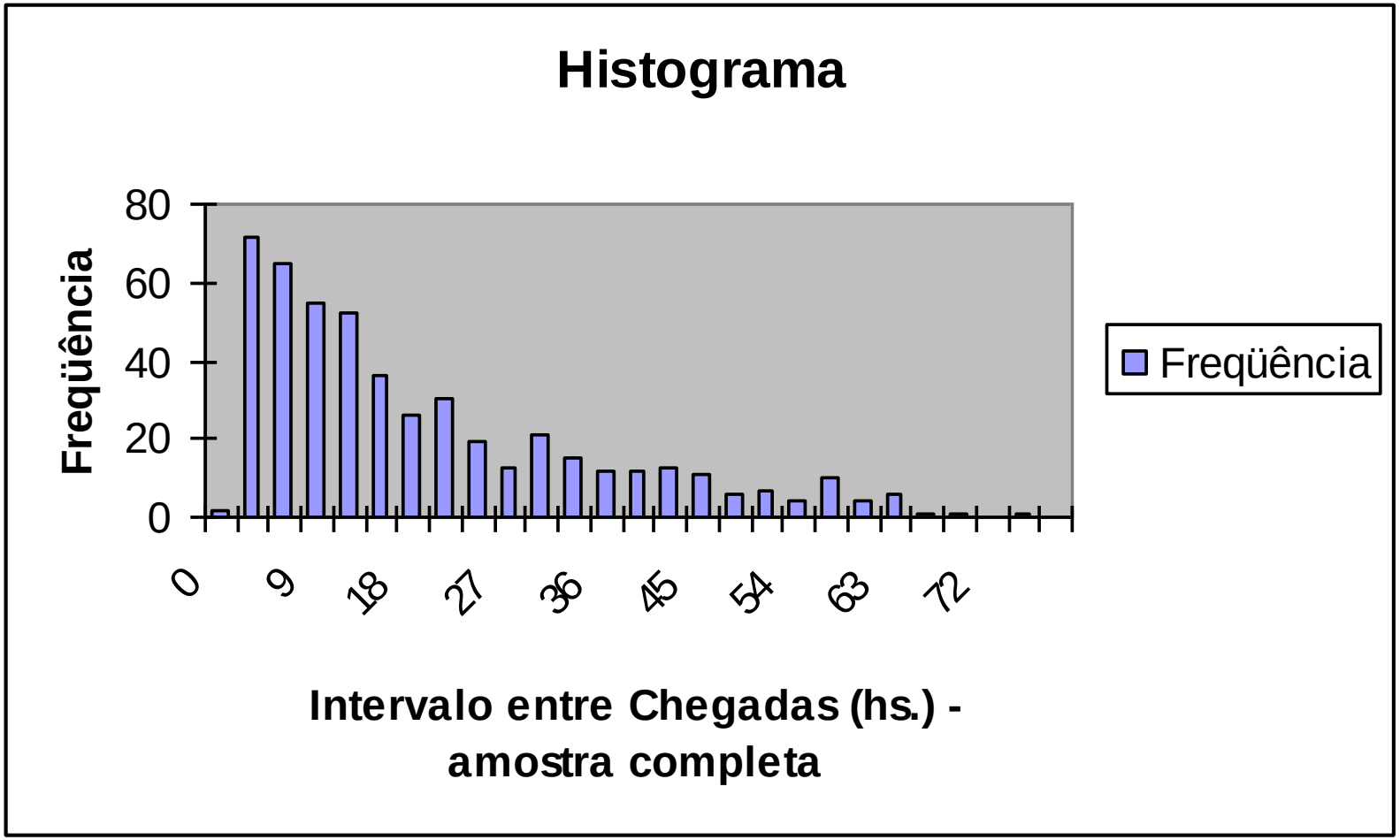
Exemplo de Acompanhamento no. 2

- **A amostra contém o instante da chegada de 495 navios**
- **A primeira coluna indica o tipo de navio: 1-Conteiner, 2-Carga Geral, 3-Refrigerado**
- **A segunda coluna indica o instante que o navio chegou (em dias)**

1	0
2	0
3	0
1	1,664373
2	2,485113
1	3,737042
1	3,953346
1	4,581805
1	4,760281
1	4,87695
2	4,97031
....	.....

Exemplo de Acompanhamento no. 2

- **Histograma para a amostra 495 navios**



Investigação por meio de gráfico de caixas – Resultados para o Exemplo de Acompanhamento no. 2

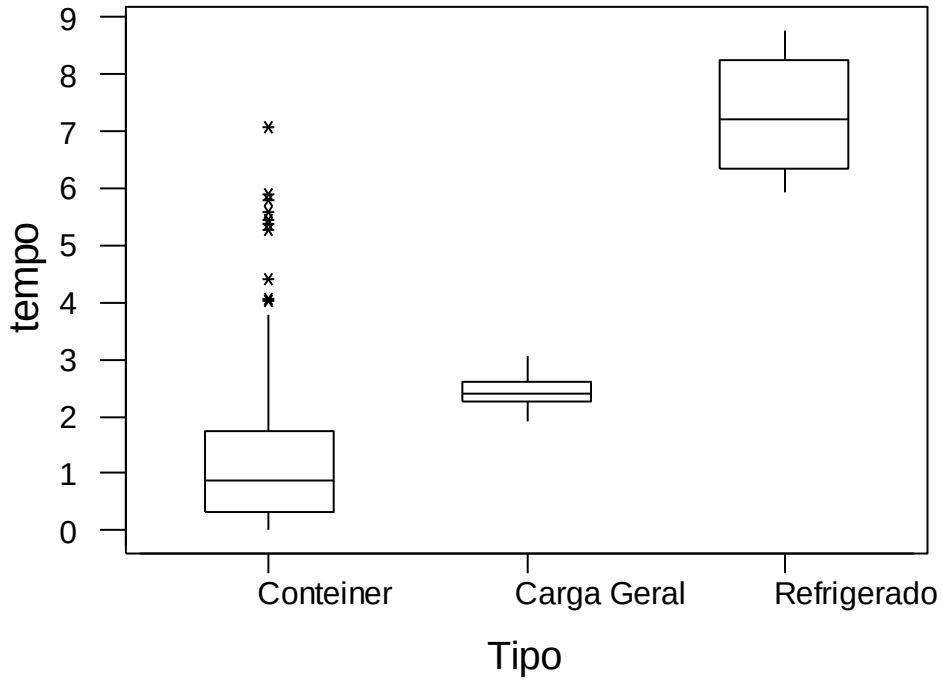


Gráfico de caixas para o intervalo entre as chegadas de três tipos de navios

Análise de Variância

Requisitos para aplicação da forma clássica:

-Distribuição Normal e igualdade de variâncias

Relaxação da igualdade de variâncias se a razão entre a maior e a menor for inferior a 5 (Scheffé – 1958);

Relaxação da Distribuição (com pouco desvio em relação a normal), mas exigência de igualdade de variâncias (Nether - 1995);

Aplicação de Teste não paramétrico (Kruskal- Wallis) para comparar igualdade das distribuições. Caso o teste seja rejeitado, o mesmo não indica se a diferença é devida às médias, ou às variâncias ou na forma das distribuições (Conover – 1980).

Análise de variância – Resultados para o Exemplo de Acompanhamento no. 2

One-way ANOVA: Intervalos versus Fator

Analysis of Variance for Interval

Source	DF	SS	MS	F	P
Fator	2	1552,330	776,165	809,14	0,000
Error	489	469,071	0,959		
Total	491	2021,401			

Level	N	Mean	StDev
1	293	1,2435	1,1939
2	150	2,4206	0,2361
3	49	7,3050	0,9632

Individual 95% CIs For Mean Based on Pooled StDev

```
-----+---------+---------+---------+---------+-
        (-*-)
             (-*-)
                          (-*-)
-----+---------+---------+---------+---------+-
                      2,0       4,0       6,0       8,0
```

Pooled StDev = 0,9794

Tukey's pairwise comparisons

Family error rate = 0,0500

Intervals for (column level mean) - (row level mean)

	1	2
2	-1,4072 -0,9469	
3	-6,4153 -5,7077	-5,2616 -4,5072

Análise de variância das médias de intervalo entre chegadas de três tipos de navios.

Análise de Agrupamentos (*Clusters Analysis*)

É um técnica descritiva simples que pode ser aplicada quando o analista quer agrupar elementos, que são representados por diversas variáveis. As etapas são:

- **Escolher as variáveis e normalizar;**
- **Montar uma matriz de distâncias euclidianas;**
- **Escolher a menor distância e reduzir a matriz, montando um agrupamento;**
- **Recalcular a matriz de distâncias e repetir o passo anterior, até que só dois grupos permaneçam ou que o valor da distância "d" aumente muito de uma redução para outra.**

Exemplo de Acompanhamento no. 3

Características de nove terminais de contêineres

Terminal	Volume de contêineres movimentado por mês (T.E.U.)	Área disponível (m^2)
A	10000	40000
B	5800	12000
C	3000	15000
D	12000	19000
E	8500	30000
F	4200	8000
G	6000	18500
H	2800	9000
I	7000	25000
Média	6588,9	19611,1
Desvio-Padrão	3135,5	10493,4

Exemplo de Acompanhamento no. 3

Dados padronizados dos nove terminais de contêineres

Terminal	Volume de Contêineres	Área
A	1,09	1,94
B	-0,25	-0,73
C	-1,14	-0,44
D	1,73	-0,06
E	0,61	0,99
F	-0,76	-1,11
G	-0,19	-0,11
H	-1,21	-1,01
I	0,13	0,51

Exemplo de Acompanhamento no. 3

Distâncias euclidianas entre os nove terminais (primeira matriz)

Terminais	A	B	C	D	E	F	G	H
B	2,99							
C	3,27	0,94						
D	2,10	2,09	2,90					
E	1,07	1,92	2,26	1,53				
F	3,57	0,64	0,77	2,70	2,51			
G	2,41	0,62	1,01	1,91	1,36	1,15		
H	3,74	1,00	0,58	3,09	2,70	**0,46**	1,36	
I	1,72	1,30	1,59	1,69	0,68	1,85	0,70	2,03

Exemplo de Acompanhamento no. 3

Distâncias euclidianas entre os oito terminais mais o grupo FH (segunda matriz e após a primeira redução)

Terminais	A	B	C	D	E	G	FH
B	2,99						
C	3,27	0,94					
D	2,10	2,09	2,90				
E	1,07	1,92	2,26	1,53			
G	2,41	**0,62**	1,01	1,91	1,36		
FH	3,65	0,82	0,67	2,89	2,60	1,26	
I	1,72	1,30	1,59	1,69	0,68	0,70	1,94

Exemplo de Acompanhamento no. 3

Agrupamento Final (Após a sétima redução)

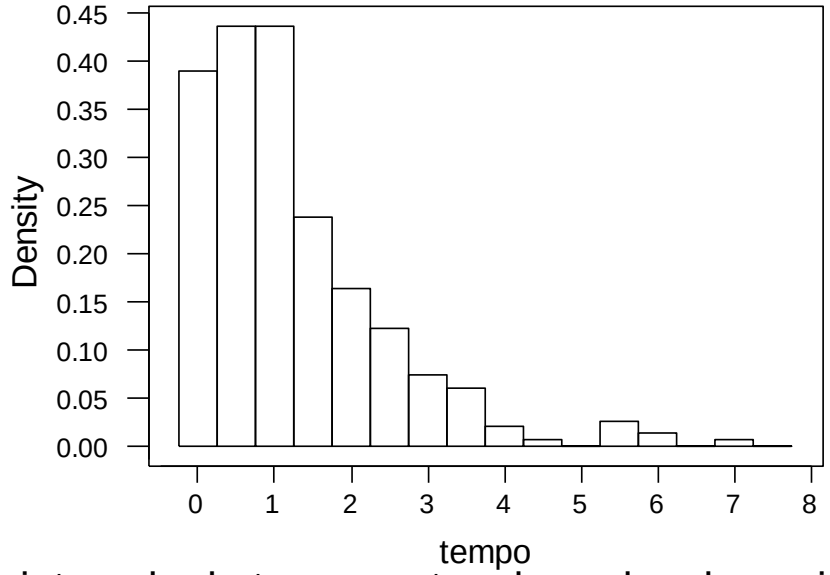
Terminais/grupos	FHCGB
IEAD	2,42

Seleção da Distribuição de Probabilidade

A seleção de probabilidades será feita utilizando-se três métodos. São eles:

- **Investigação gráfica e gráfico de probabilidades;**
- **Teste de aderência de Qui-Quadrado;**
- **Teste de aderência de Kolmogorov- Smirnov.**

Investigação gráfica e gráfico de probabilidades – Resultados para o Exemplo de Acompanhamento no. 2



Histograma do intervalo de tempo entre chegadas de navios do tipo contêiner

distribuição fortemente assimétrica (descarta-se a normal e a uniforme)
e seu histograma é semelhante ao das distribuições exponencial, Weibull ou gama;

Investigação gráfica e gráfico de probabilidades – Resultados para o Exemplo de Acompanhamento no. 2

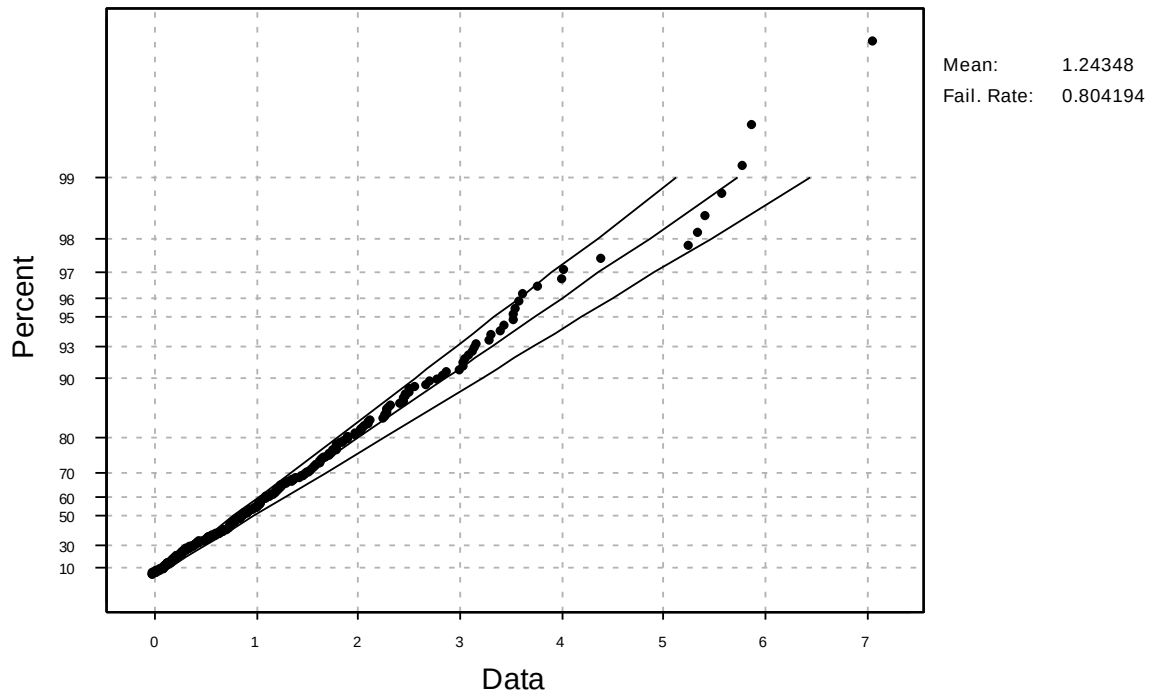


Gráfico de probabilidades exponencial para os intervalos de tempo entre chegadas de navios do tipo contêiner

Testes de aderência

Um teste de aderência é um teste de hipóteses que verifica se um determinado conjunto de dados foi gerado através de uma distribuição especificada.

Teste de aderência de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov baseia-se na comparação entre a distribuição acumulada amostral e a função distribuição acumulada teórica, que se acredita ter gerado os dados (F(x)).

A estatística do teste, que tem como hipótese nula que os dados foram gerados segundo uma distribuição específica contra uma hipótese alternativa, que diz que os dados não foram gerados segundo essa distribuição, é dada por:

$$D = \sup_x \left| F(x) - S(x) \right|$$

Teste de aderência Qui-Quadrado

A expressão abaixo mostra como calcular a estatística E para utilização no teste de aderência proposto.

$$E_k = \frac{O_k - T_k}{T_k}$$

$$E = \sum_{K=1}^{K} E_k^2$$

O valor E_k é a diferença entre o número observado de elementos (O_k) e o valor teórico (T_k) e divide-se o valor obtido pelo valor teórico da classe (T_k). A somatória dos valores desses valores E_k, para todas as K classes envolvidas determina a estatística E, cuja distribuição é uma QUI-QUADRADO com K-1-n graus de liberdade, em que n é o número de parâmetros estimados a partir da amostra coletada.

Teste de aderência Qui-Quadrado e Kolmogorov-Smirnov – Resultados para o Exemplo de Acompanhamento no. 2

Distribution Summary **(intervalo entre chegadas – dias)**

Distribution: **Exponential**
Expression: **EXPO(1.24)**
Square Error: **0.000845**

Chi Square Test
Number of intervals = 8
Degrees of freedom = 6
Test Statistic = 6.78
Corresponding p-value = 0.357

Kolmogorov-Smirnov Test
Test Statistic = 0.0447
Corresponding p-value > 0.15

Algumas observações sobre os testes de aderência

•Os pacotes computacionais fornecem os níveis descritivos para diversas distribuições teóricas testadas, classificados em ordem crescente.

•Mais de uma distribuição teórica poderá aderir àquela amostra para um mesmo nível de significância adotado.

•Sugere-se que o analista utilize a distribuição mais conhecida e cujos parâmetros são mais facilmente calculados. Por exemplo, uma exponencial ao invés de uma beta.

•Uma situação inversa também pode ocorrer, ou seja, a lista de distribuições apresentadas pelos pacotes computacionais apresenta níveis descritivos muito baixos, mesmo para a melhor distribuição por eles indicada. Nesse caso, recomenda-se distribuição denominada empírica, que nada mais é do que a representação, por intervalos, do gráfico de freqüência acumulada obtido a partir da amostra.

Conclusões

- **Identificou-se uma lacuna na literatura nacional e internacional no que se refere ao tratamento e análise de dados para elaboração de modelos de simulação discreta;**

- **Foi feita uma proposta de alteração do processo de simulação;**

- **Foi desenvolvido de um procedimento simples eficiente de análise de dados.**

STAT:FIT

Prof. Dr. Rui Carlos Botter

Teste de Aderência com o auxílio de um software

Março de 2017

STAT:FIT

Procure em “todos os programas” SIMUL e selecione STATFIT

A versão estudantil é mais restrita, mas permite obter a curva que melhor adere aos dados.

Os dados devem estar sem cabeçalho e a separação da parte decimal com ponto.

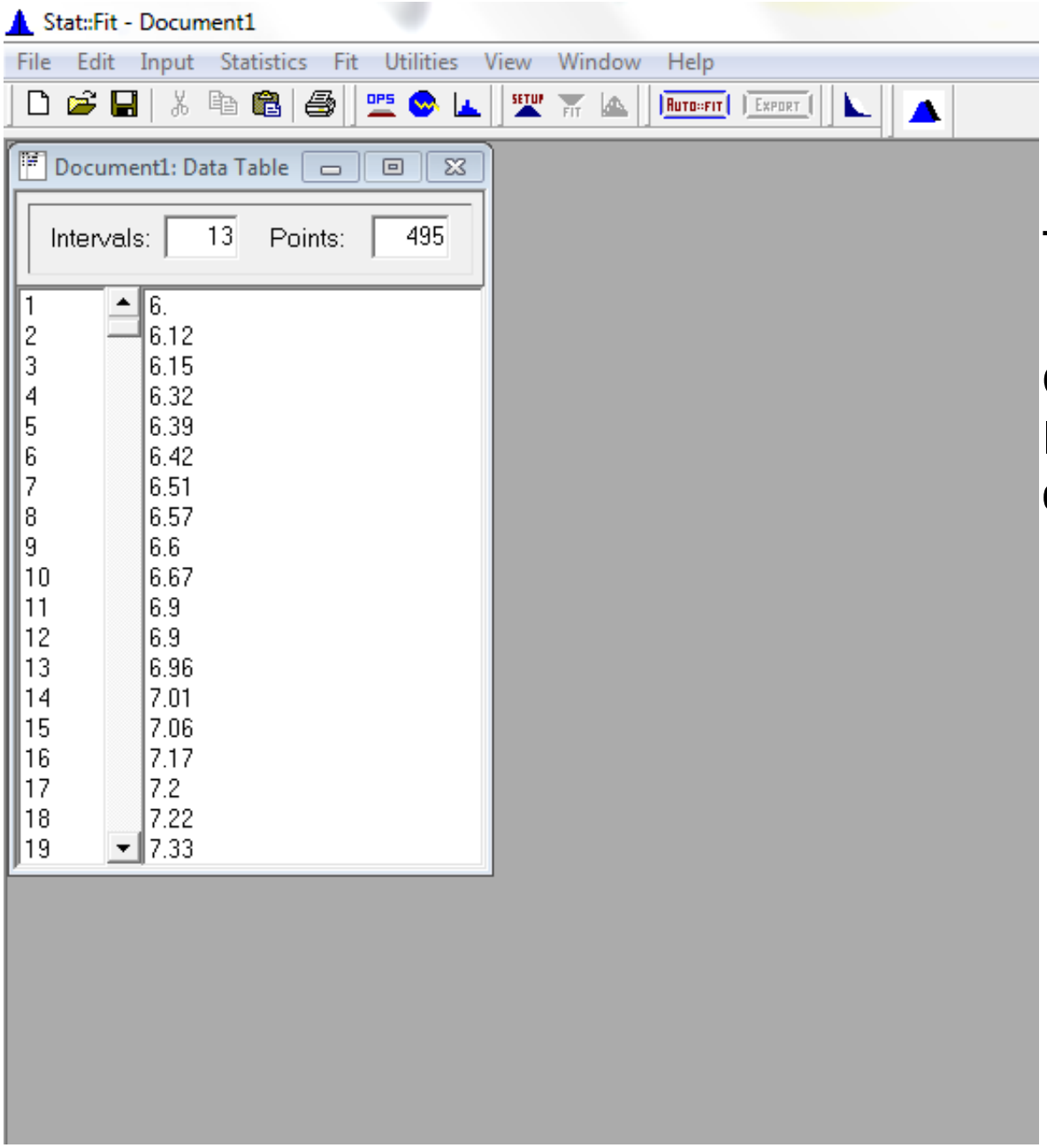
STAT:FIT

No arquivo pallet.txt aparecem os dados para navios que transportam pallets.

Foram retirados os valores zero e muito elevados, mas para outros arquivos de dados caberia uma análise detalhada prévia dos mesmos buscando-se a retirada dos outliers.

Somente com o arquivo de dados previamente analisado é que se recomenda utilizar o STATFIT

Copiar o conjunto de dados e “colar”diretamente na planilha da direita do STAT FIT

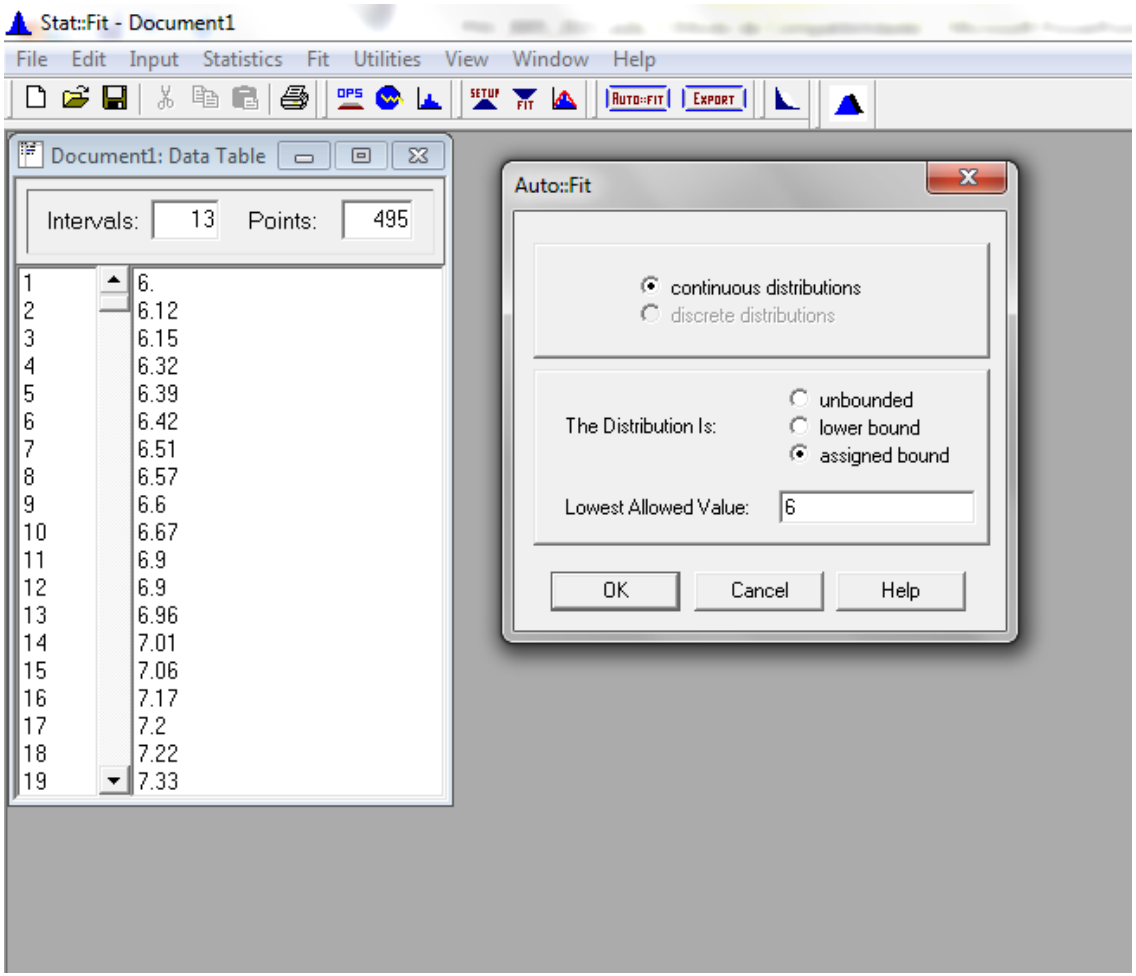


Tela do Statfit

Com os dados dos pallets Inseridos na coluna da direita

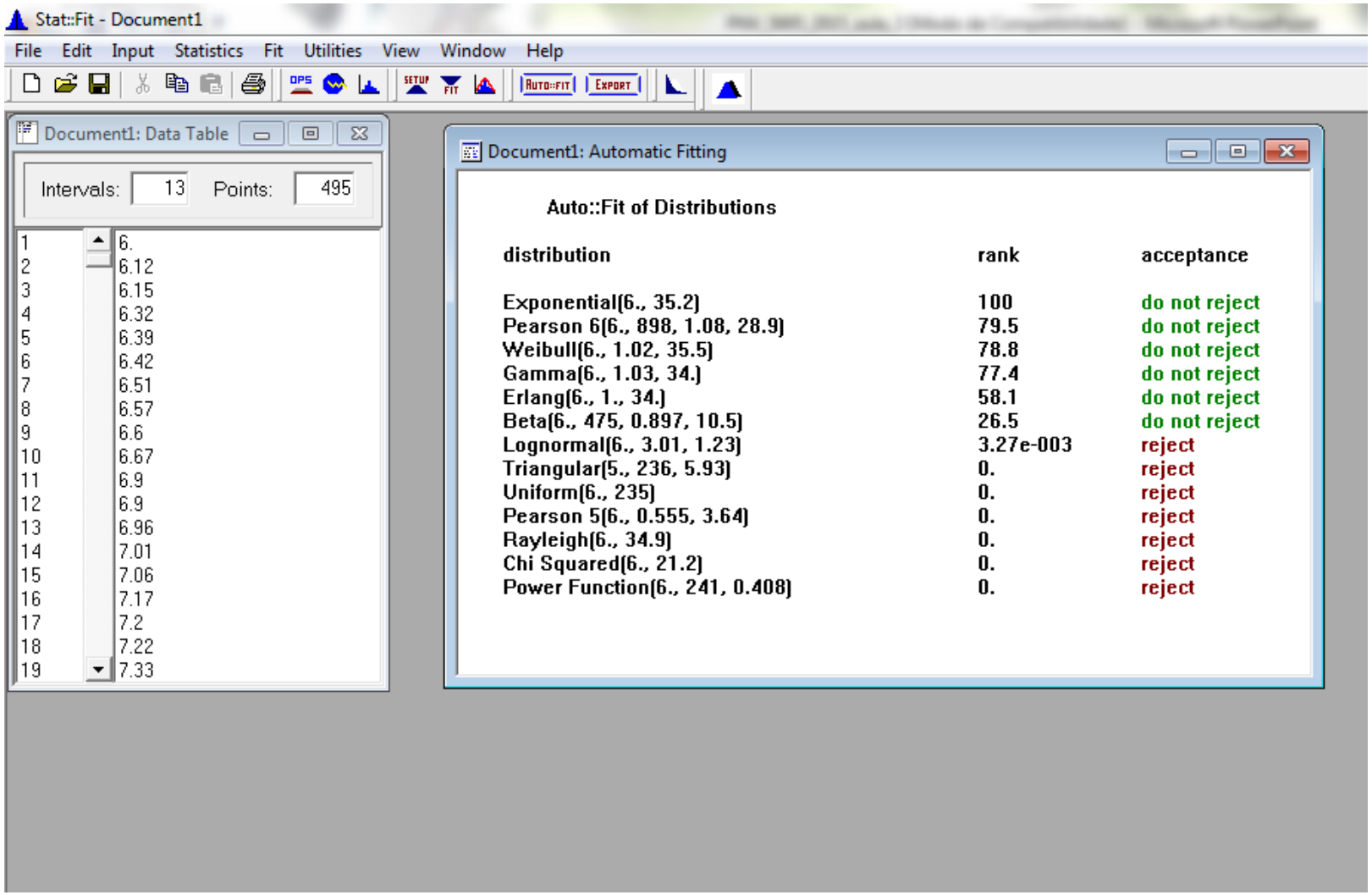
STAT:FIT

Selecione parte superior a opção: FIT
em seguida Auto:: FIT e OK



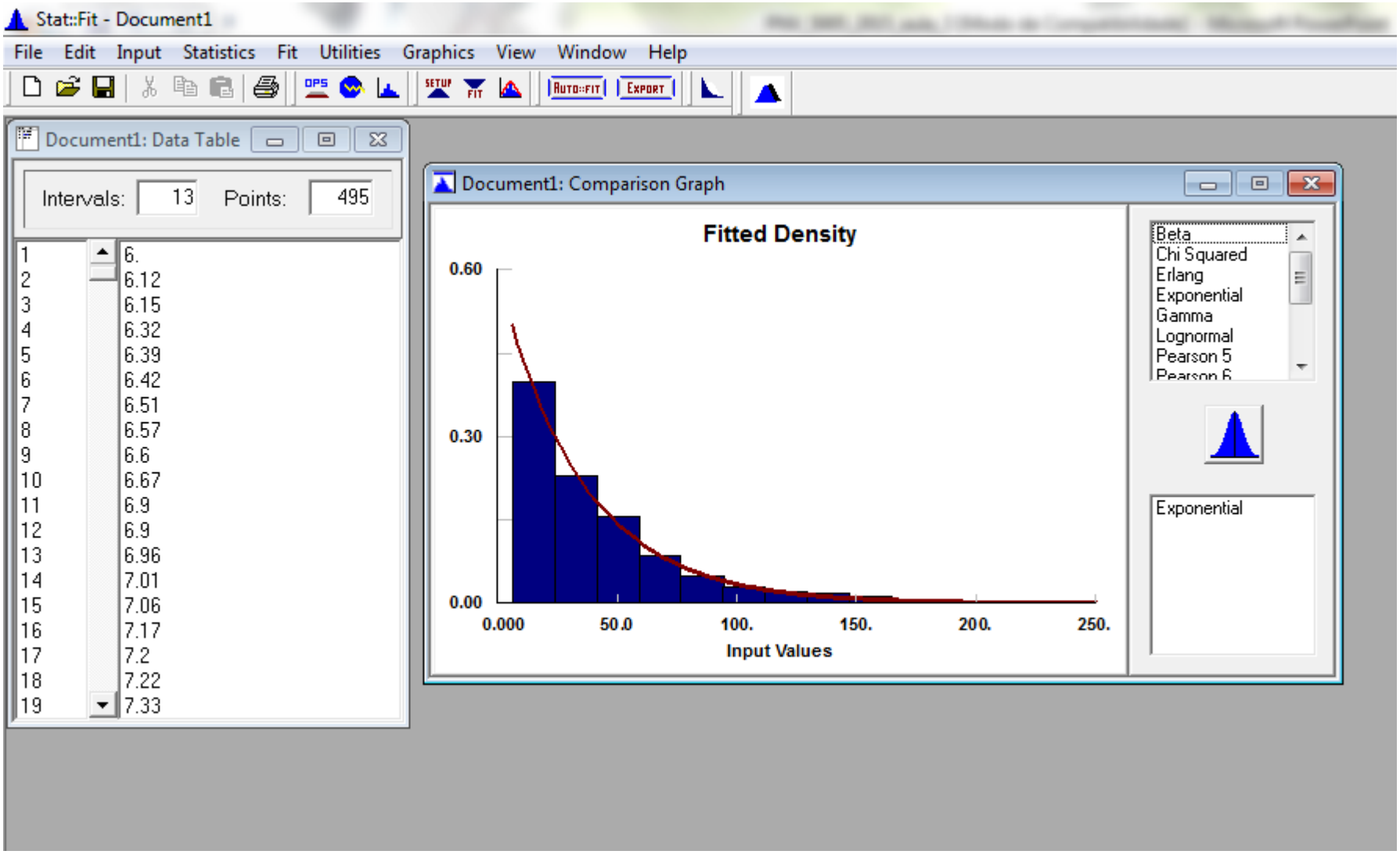
STAT:FIT

A curva que melhor aderiu, o RANK e se o teste de aderência foi ou não aceito aparecerá em um nova janela



STAT FIT

Depois na opção FIT, selecione: Result Graphs e escolha “density”



STAT FIT

CUIDADO:

O STAT FIT analisa qualquer conjunto de dados: Já previamente “limpos” ou não;

Deixar de analisar os dados previamente e inserí-los diretamente no STAT FIT pode levar a obtenção de resultados ruins a respeito dos mesmos