5

ANÁLISE E TRATAMENTO DE DADOS PARA A SIMULAÇÃO

Neste capítulo serão abordados tópicos relativos à análise e ao tratamento dos dados voltados a alimentação de modelos de simulação. Juntamente com as etapas anteriores: definição do problema, descrição correta do sistema e hipóteses preliminares sobre seu comportamento, esta etapa constitui um importante passo antes da fase de modelagem.

Tópicos

5.1 Introdução
5.2 Processo de Amostragem e Coleta dos dados
5.3 Tratamento dos Dados
5.4 Identificação da Distribuição Teórica de Probabilidades
5.5 Estimação de Parâmetros
5.6 Testes de Aderência
5.7 Ajuste de Distribuições com o Arena *Input Analyzer*
Sumário
Exercícios

5.1 Introdução

Modelar computacionalmente um sistema do mundo real sugere a criação de uma espécie de analogia digital deste sistema. Esta deve possuir a capacidade de se comportar de maneira semelhante ao sistema original de tal forma que, ao interagir com o usuário, permita a este a realização de experimentos com a intenção final de um maior entendimento e compreensão do sistema real.

Para que um modelo possa criar uma *história artificial* do sistema real, é fundamental que este traga consigo a possibilidade de apresentar até mesmo um comportamento estocástico, à semelhança da grande maioria dos sistemas. Em modelos voltados à simulação, este objetivo é alcançado pela utilização de distribuições de probabilidades (empíricas ou teóricas) como forma de representar a multiplicidade de ocorrências de eventos aleatórios. Alguns exemplos clássicos deste emprego na área dos sistemas de manufatura são: os tempos decorridos entre as falhas de equipamentos ou ainda o tempo necessário para repará-las. Pode-se citar ainda o intervalo de tempo entre chegadas de clientes num sistema, o tempo de passagem de entidades por um sistema, etc.

Quando se faz uso de distribuições de probabilidades para representar o comportamento de variáveis aleatórias presentes nos sistemas a serem modelados, é preciso considerar os seguintes pontos:

1. Os possíveis valores que a variável poderá assumir estarão dentro da amplitude coberta pela distribuição;
2. A probabilidade de ocorrência de qualquer valor no intervalo é determinada pelo perfil da distribuição.

Portanto, é possível antecipar que valores a variável poderá assumir sem, no entanto, ser possível determinar qual, precisamente será este valor. A garantia de realizar um perfeito casamento entre uma distribuição teórica de probabilidades e o comportamento aleatório de uma variável do sistema passa por várias etapas. Vários autores [Banks, 1984; Law, 1991, Pegden, 1994; Pidd, 1992, Jain, 1991] apontam alguns passos básicos com vista a realização desta identificação:

1. Processo de amostragem e coleta dos dados;
2. Tratamento dos dados
3. Identificação da distribuição estatística;
4. Estimação dos parâmetros da distribuição identificada;
5. Testes de aderência.

Cada um destes passos será apresentado e discutido neste capítulo.

5.2 Processo de Amostragem e Coleta dos dados

Todo o processo de identificação da distribuição de probabilidade mais adequada a expressar o comportamento da variável aleatória sob estudo começa, obrigatoriamente, com a coleta de dados. Este costuma ser, também, o marco inicial dos problemas que se enfrenta na modelagem de sistemas.

Os dados estão disponíveis? De que maneiras estão disponíveis? Como coletá-los? Como analisá-los? Estas e outras questões básicas, a respeito de dados necessários a um estudo de simulação, passam a ocupar a mente do analista uma vez iniciado os trabalhos de modelagem. O fato é que se necessita de dados para simular corretamente um sistema. Nunca se deve esquecer que ao final do processo, isto é, ao se coletar os resultados provenientes das simulações, estes dependerão, fundamentalmente, do tratamento fornecido aos dados de entrada. A simulação é uma ferramenta de avaliação logo, o *output* depende da qualidade do *input*. Desta maneira, mesmo que toda a estrutura do modelo esteja correta, se os dados de entrada forem inadequadamente coletados e analisados ou, não forem representativos do ambiente modelado, os resultados da simulação podem não ter a validade desejada, levando a análises sem fundamentos e a decisões impróprias.

5.2.1 Fontes de Dados

Na maioria dos problemas envolvendo a simulação de sistemas reais, a determinação e a coleta de dados é uma tarefa difícil e demorada. Em geral, enfrenta-se duas situações básicas: a existência ou possibilidade de obtenção de dados e a não existência ou impossibilidade de obtenção de dados. Na grande maioria das vezes enquadram-se, no primeiro caso, modelos baseados em sistemas reais existentes e acessíveis. No segundo caso, encontram-se aqueles modelos de sistemas que ainda não existem. No entanto, algumas vezes, mesmo com a existência física do sistema a ser modelado, os dados desejados não existem ou não é possível obtê-los. Muitas vezes pertencem ao passado, tornando impossível uma nova coleta. Por exemplo, coletar dados sobre a fabricação de determinados tipos de produtos que não são regularmente produzidos (problema de sazonalidade). Segundo um levantamento apresentado por Pegden (1990), na maioria dos casos e, dependendo das circunstâncias, as fontes de dados podem ser:

1. Arquivos históricos (expondo o comportamento, resultados, etc.) do sistema;
2. Provenientes de observações do sistema sob estudo;
3. Oriundos de sistemas similares;
4. Determinados com base em estimativas de operadores;
5. Obtidos com base em afirmações de vendedores de máquinas, equipamentos, etc.;
6. Estimativas de projetistas de sistemas, ou mesmo;
7. Considerações teóricas sobre o sistema.

5.2.2 Amostragem

A questão da coleta de dados com o objetivo de modelagem e simulação de sistemas não é tema comum em livros e artigos de simulação. A maior parte das discussões sobre o tema ocorre em livros de estatística [Miller et alli,, 1991 ou Mendenhall, W., 1973], ou mesmo aqueles exclusivamente dedicados ao assunto [Cochran, W.G., 1977]. Uma boa referência sobre o tema é o livro de Banks [1984], voltado à modelagem e simulação de sistemas. Nele, o autor aponta alguns pontos essenciais, no sentido de melhorar e facilitar a condução do exercício de coletar dados. Apresenta-se abaixo um resumo dos elementos apontados em Banks [1984].

1. *Planejamento e Observação Preliminar*: Iniciar o trabalho com um planejamento. Este pode começar por uma pré-observação da situação. Tentar coletar dados enquanto se observa. Durante este processo de observação preliminar, é importante que se procure imaginar algumas formas de realizar a coleta. É provável que estas formas de coleta tenham que ser modificadas antes que a versão final seja alcançada. Uma atenção para circunstâncias não usuais no período observado e de como estas circunstâncias podem ser tratadas.
2. *Utilidade dos Dados Coletados*: Tentar analisar os dados na medida de sua coleta. Determinar se estes são adequados ao fornecimento das distribuições as quais serão tomadas como entrada de dados na simulação. Verificar se algum dado que está sendo coletado não é útil para a simulação. Não existe necessidade de se coletar dados supérfluos.
3. *Conjuntos Homogêneos de Dados:* Tentar combinar dados que formem conjuntos homogêneos, isto é, que obedeçam ao mesmo tipo de distribuição ao longo de um determinado período ou intervalo de tempo. Por exemplo, para verificar se os dados referentes à entrega de matéria-prima em um determinado setor de uma fábrica são homogêneos, coletar dados em um intervalo no período da manhã e em outro no período da tarde. Verificar também o comportamento em diferentes dias da semana ou do mês. Testes estatísticos, como o teste *t*, para comparação das médias, podem ser empregados para este propósito.

4. *Relacionamento entre Variáveis:* Para determinar o relacionamento entre duas variáveis, um diagrama de dispersão pode ser construído. Muitas vezes, apenas uma simples visualização do diagrama pode indicar a existência de uma relação entre duas variáveis de interesse. Uma análise de regressão (ou técnicas estatísticas mais sofisticadas) pode ser empregada para examinar a relação e sua significância.
5. *Independência das Observações:* Considerar a possibilidade de que uma seqüência de observações, que aparentemente parecem independentes, possua algum relacionamento. A autocorrelação pode existir entre períodos sucessivos de tempo ou para clientes (entidades) sucessivos. Por exemplo, o tempo de serviço para o cliente *i*, pode estar relacionado com o tempo para servir o cliente *i*+1.

É importante deixar claro que as medidas sugeridas por Banks servem apenas como um guia ou sugestões sobre os procedimentos adequados que devem ser tomados quando da coleta e análise dos dados. Procedimentos estes que devem ser abordados com o maior cuidado possível.

Estudo de Caso: Amostragem para um Modelo de uma Agência Bancária

Para exemplificar alguns procedimentos básicos do processo de amostragem, foi desenvolvido um estudo de caso que utiliza como cenário uma agência bancária. O modelo que descreve o comportamento deste sistema é exibido na figura 5.1. O modelo original foi desenvolvido pela *Systems Modeling Corp.*, fabricante do Arena, e pode ser encontrado no diretório de exemplos que acompanha o software. O modelo que desenvolvido é baseado naquele, com pequenas modificações. Pode ser encontrado sob o título "*Transações Bancárias.doe*".

Na seqüência, expõem-se os procedimentos que se acredita sejam adequados ao processo de amostragem que deve ser realizado neste sistema com vista à coleta de dados.

Como pode ser observado na figura 5.1, a agência possui três pontos básicos de atendimento ao público:

1. Caixas internos (caixa 1 e caixa 2), funcionando no período das 10:00 às 16:00 horas;
2. Sistema *drive-thru* (atendimento ao cliente no seu próprio veículo), num caixa externo dedicado a este serviço, também funcionando no período de 10:00 às 16:00;
3. Sistema de auto-atendimento em caixas automáticos (ATM), durante as 24 horas do dia.

Durante o processo de observação preliminar do sistema, foram identificados inúmeros tipos de clientes (entidades). Dentre estes, três tipos básicos farão parte do modelo, como forma de simplificação do processo de modelagem. São eles:

1. Cliente dos caixas (atendimento por funcionários do banco);
2. Cliente das máquinas automáticas;
3. Cliente do *drive-thru*.

Para a perfeita modelagem de cada um deles, dois parâmetros são importantes e necessitam de amostragem:

1. O período de tempo decorrido entre duas chegadas de cada tipo de cliente;
2. O tempo necessário para que estes procedam com suas transações.

No modelo estes valores serão armazenados nas variáveis TECc, TECm e TECd. Respectivamente: Tempo Entre Chegadas nos Caixas, Tempo Entre Chegadas nas Máquinas e Tempo Entre Chegadas no *Drive-thru*. Para os tempos de serviços, os valores são armazenados nos atributos TSc, TSm e TSd.

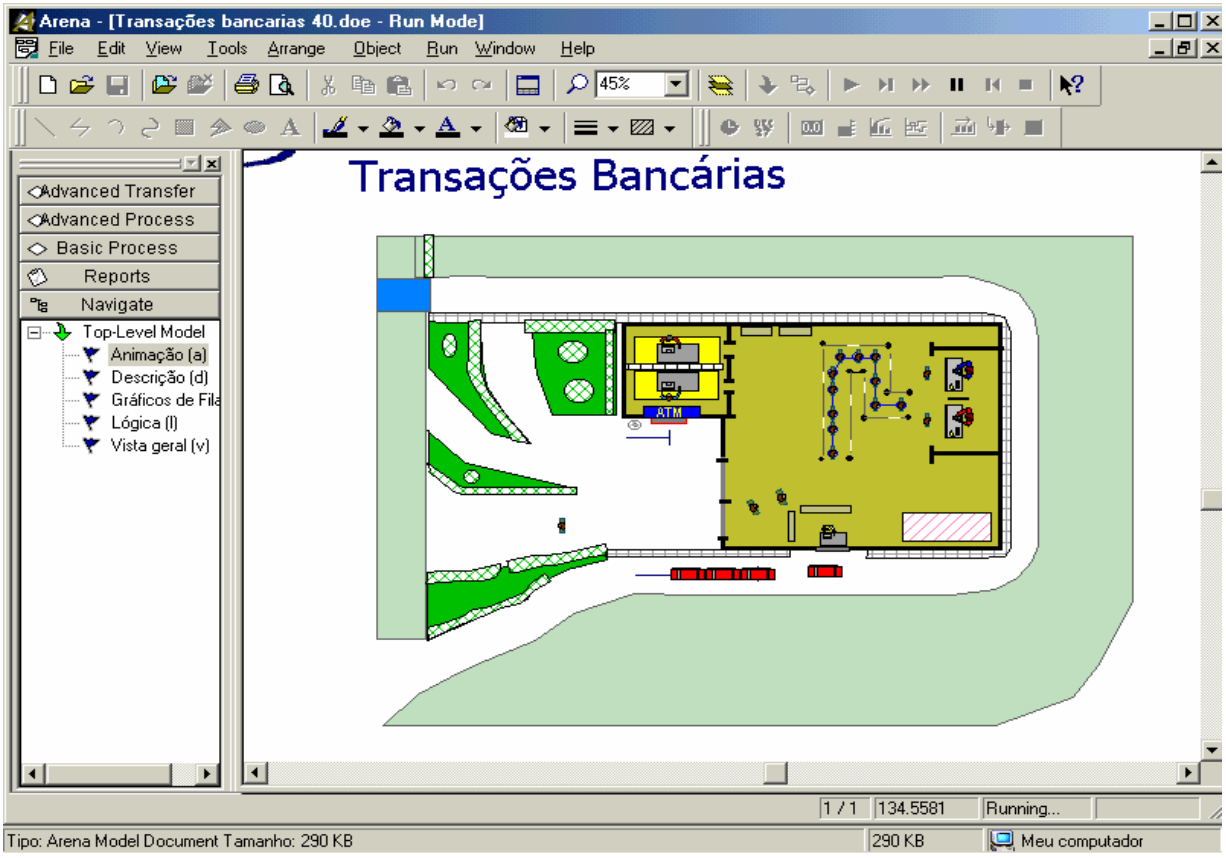

Figura 5.1: Vista do modelo Transações Bancárias criado no ARENA

Antes de iniciar os procedimentos de amostragem é preciso estar ciente dos objetivos da modelagem e simulação deste sistema. Neste caso, segundo a gerência, o importante é avaliar o desempenho do sistema considerando a satisfação dos clientes medida em função do tempo de espera pelo atendimento (tempo de fila). Além disso, também é importante medir o número de clientes nas filas, uma vez que esta medida permite definir o espaço físico necessário a este fim, bem como diminuir a taxa dos clientes que desistem de usar o sistema, tendo em vista a "expectativa", a partir da avaliação visual, de um longo tempo de espera nas filas.

Neste estudo, consideram-se, apenas, os clientes que se dirigem aos guichês dos caixas. Numa observação preliminar, visando o planejamento do processo de amostragem, o analista é informado pela gerência que, dependendo do horário e do dia da semana ou do mês, os

tempos TSc e TECc, associados a cada tipo de cliente, variam bastante. Desta maneira, será preciso montar uma estratégia para que tais variações sejam capturadas pelo processo de amostragem. Esta troca de informações com o gerente do sistema permite a obtenção de indicações sobre os horários do dia da semana e do mês em que existe maior demanda por serviços, por posto de atendimento.

Devido à escassez de recursos e de prazos exíguos para o cumprimento de etapas do projeto, a realização da amostragem ao longo de períodos muito extensos, que cubram todas ou a maioria das situações operacionais pode não ser viável. Neste caso, amostras significativas devem ser levantadas em períodos considerados mais críticos, apontados pela gerência.

No caso de sistemas em que existe o registro eletrônico de transações (comuns em sistemas prestadores de serviços), uma excelente alternativa ao procedimento acima é o exame dos registros dos caixas, uma vez que nestes, os horários de realização de cada transação ficam armazenados. No entanto, muitas vezes, a observação destes registros não é possível (por motivos de normas de segurança interna, por exemplo). Desta forma, a partir dos indicativos da gerência, deve ser montada a estratégia de amostragem.

Processo de Amostragem

De acordo com informações da gerência, considerando somente clientes que se dirigem aos caixas, os períodos críticos (mais congestionados) em dias considerados normais (terças, quartas e quintas-feiras), podem ser divididos em três níveis de demanda: A, acima da média; B, na média e C, abaixo da média. As distribuições destas demandas durante o horário comercial, das 10:00 às 16:00 horas, ocorrem de acordo com a tabela 5.1.

Período	Tipo de Demanda
10:00 às 11:00	A
11:00 às 13:30	C
13:30 às 14:30	B
14:30 às 15:30	C
15:30 às 16:00	A

Tabela 5.1: Perfil da demanda em meio da semana

Nas segundas-feiras e sextas-feiras, o perfil da demanda é semelhante, mas os níveis de demanda se modificam, conforme pode ser observado na tabela 5.2.

Período	Tipo de Demanda
10:00 às 11:00	A* 1,3
11:00 às 13:30	B
13:30 às 14:30	A
14:30 às 15:30	B
15:30 às 16:00	A* 1,2

Tabela 5.2: Perfil da demanda no início e fim da semana

Além disso, qualquer dia de meio de semana que seja o último do mês tem demanda semelhante a da tabela 5.2. Se o último dia do mês for uma sexta-feira ou o primeiro dia do mês for uma segunda-feira, o perfil da demanda segue a tabela 5.2, acrescida de 20%.

Observe-se que as informações passadas pela gerência facilitam, sobremaneira, o processo de coleta de dados, pois, uma vez definidos os valores dos parâmetros A, B e C, o perfil da demanda para os diversos períodos relativos ao cliente tradicional (caixas internos), estará determinada. No entanto, nem sempre existe tal possibilidade, exigindo que se realizem coletas de amostras sobre os inúmeros períodos de diversidade da demanda.

Outro fato importante a ser ressaltado, é que os dados acima informados são válidos apenas para o tipo de cliente em questão. Para os demais clientes, informações semelhantes devem ser colhidas junto a gerencia ou coletadas se estas não estiverem disponíveis. Uma vez que os parâmetros A, B e C, resumem o perfil de demanda em todos os períodos, estes poderão ser coletados em qualquer dos dias da semana.

Definidos os períodos em que a coleta será realizada, o próximo passo no planejamento é a determinação do tamanho das amostras. Este ponto costuma ser bastante “nebuloso” para aqueles que estão realizando este procedimento pela primeira vez. De fato, a grande maioria dos textos sobre simulação, não costuma tratar deste tema de forma clara. É preciso, muitas vezes, buscar a informação necessária junto a textos de estatística, tarefa nem sempre agradável aos menos iniciados. Neste texto, pretende-se tratar deste assunto da forma mais prática possível, e com o enfoque voltado a sua utilização em modelagem e simulação discreta de sistemas. A teoria envolvida no procedimento de coletar dados é exatamente a mesma, independente da técnica que será utilizada para a realização dos experimentos.

A palavra chave nas questões de amostragem é: “representatividade”. Isto implica que a amostra deve ser a mais representativa possível do seu próprio universo. Considere, por exemplo, as informações passadas pela gerência do banco que se está analisando. Não fora tais informações, o número de observações a serem realizadas deveria ser muito maior. Com a informação dada, o planejamento indica que se pode buscar, especificamente em determinados horários, a informação necessária, com uma boa expectativa no que se refere ao quesito representatividade, mesmo considerando que o número de observações a ser realizado será muito menor do que se houvesse a necessidade de cobrir todos os horários de funcionamento.

Mas afinal, qual deve ser o tamanho das amostras a serem coletadas durante os períodos já definidos? Neste ponto, o problema da representatividade passa a depender de uma outra questão: a variabilidade da grandeza que se está tratando. Imagine-se, por exemplo, que os tempos decorridos entre as chegadas dos clientes nos caixas do banco sejam determinísticos, todos com valores iguais a 2 min.. Neste caso, sabendo que todos os clientes chegam em intervalos de 2 min., a amostra precisaria ter tamanho 1. Na medida que cresce a variabilidade envolvendo a variável que se está avaliando, crescerá, também, o tamanho do conjunto de dados, de forma a torná-lo representativo. Isso não significa dizer, no entanto, que para variáveis com

alta variabilidade o número necessário de observações deva crescer indefinidamente, até quase alcançar o tamanho do próprio universo (população) que se deseja representar.

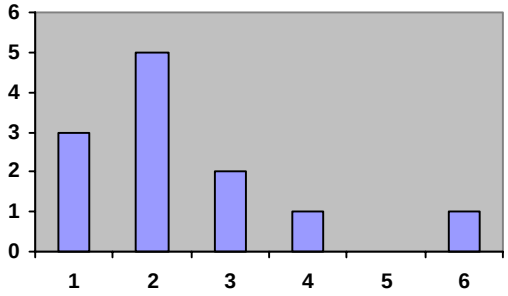

Gráfico 5.1a: 12 observações

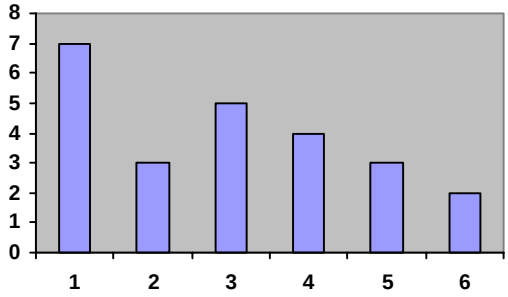

Gráfico 5.1b: 24 observações

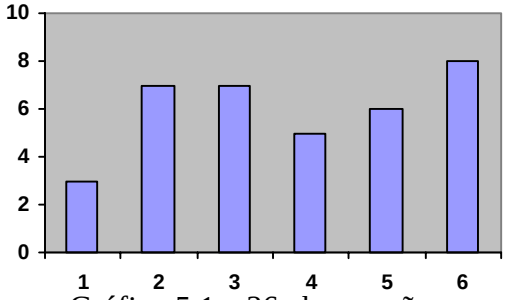

Gráfico 5.1c: 36 observações

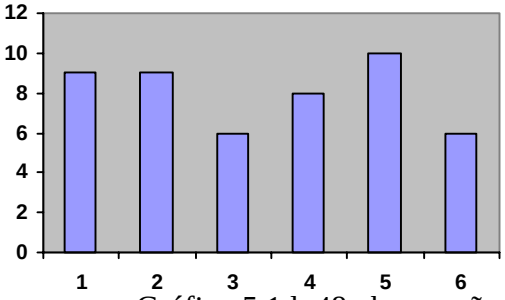

Gráfico 5.1d: 48 observações

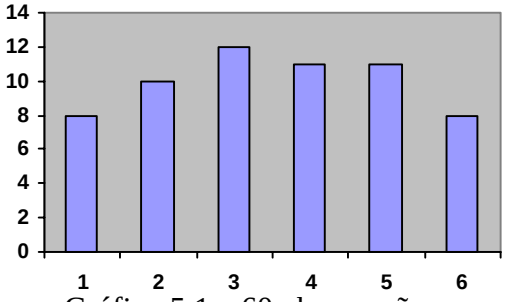

Gráfico 5.1e: 60 observações

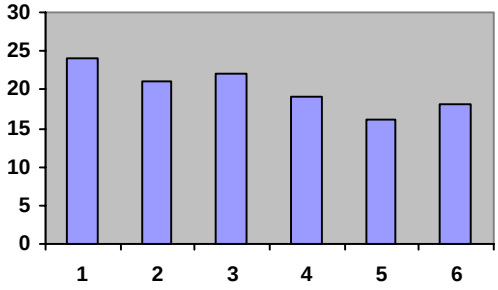

Gráfico 5.1f: 120 observações

Esta questão da representação da amostra pode ser exemplificada através do experimento de lançar um dado. Quantas vezes um dado deve ser lançado para que se possa afirmar que os seus possíveis resultados {1, 2, 3, 4, 5 e 6}, tem todos a mesma probabilidade de ocorrerem? Se você realizar tal experiência verá que, apesar da grande variabilidade desta variável, a partir de um conjunto com cerca de 50 observações é possível reconhecer esta tendência.

Observe nos Gráficos 5.1a até 5.1f, a comparação de diversos resultados desta experiência, quando simulado no Arena o lançamento de um dado, desde 12 até 120 vezes (você também pode realizar esta experiência usando o modelo *Amostra do dado.doe*).

O que se pode observar nos gráficos acima, é uma tendência a uma maior representatividade na medida em que cresce o número de observações. Embora esta variável possua grande variabilidade, a partir do gráfico 5.1d, com 48 observações a tendência a uma distribuição uniforme das observações, considerando os possíveis valores que o dado pode assumir, fica mais evidente.

Com relação as variáveis que se precisa amostrar, TECc e TSc, se verificará, mais adiante neste capítulo, tratarem-se de variáveis bastante conhecidas. Na grande maioria dos casos, é comum que o tempo decorrido entre as chegadas de clientes em um sistema como um banco, seja bem descrito por uma distribuição Exponencial. Da mesma forma, o tempo que um caixa despende atendendo um cliente pode, na maioria das vezes, ser tratado por uma variável com distribuição normal. Considerando, pois, estas duas variáveis, verifiquem o comportamento das mesmas quando se coletam vários tamanhos de amostras.

Coleta de Dados dos Tempos Entre Chegadas (TECc) dos Clientes dos Caixas

Observando as recomendações do planejamento, foram realizadas observações nos períodos indicados pela tabela 5.1. Uma vez que se tenha a prévia noção do tipo de distribuição teórica cujos parâmetros se está tentando obter, o trabalho de coleta pode ser atenuado. No caso, por experiências anteriores, sabe-se que a tendência é que a variável TECc comporte-se de acordo com uma distribuição exponencial. Por ser esta uma distribuição que apresenta grande variabilidade, quanto maior for o tamanho da amostra mais certeza se terá quanto a sua representatividade. É comum se ouvir falar em números “mágicos”, 25 ou 30, para tamanho de amostras, com sendo valores aceitáveis. Tais números não tratam exatamente do que se está aqui revisando. Quando, mais adiante neste capítulo, se discutir estimação de parâmetros, estes números voltarão a aparecer, aí sim, no contexto correto. Como se disse anteriormente, o número de observações depende da variabilidade.

Imagine-se, que o verdadeiro valor de variável TECc no período das 10:00 às 11:00 horas seja perfeitamente descrito por uma distribuição Exponencial de média 2. Veja o que acontece quando se coleta amostras com tamanhos que variam de 10 a 100 elementos e se procura estimar o parâmetro correto para a variável.

Para medir a precisão do parâmetro adotado, busca-se ajuda na ferramenta *Input Analyser*, presente no ambiente de simulação ARENA. Ao final deste capítulo faz-se uso intenso desta ferramenta. Na medida em que as amostras são geradas, solicita-se à ferramenta que analise os dados, indicando (estimando) qual seria o valor do parâmetro *média*, apresentado pelas observações. Quanto mais próximo de 2, maior a confiança.

Experimento	Tam. da Amostra	Estimativa do Parâmetro
1	10	EXPO(2,45)
2	20	EXPO(2,78)
3	30	EXPO(2,26)
4	40	EXPO(2,13)

5	50	EXPO(1,98)
6	100	EXPO(2,01)

Tabela 5.3: Tamanho da amostra *versus* precisão do parâmetro

Observando-se a tabela 5.3, verifica-se que, na medida do aumento do número de observações, mais o parâmetro avaliado se aproxima do verdadeiro valor apresentado pela população, isto é Exponencial (2). Observam-se também, pequenas diferenças acima ou abaixo do esperado decorrentes da aleatoriedade do processo de amostragem. Por exemplo, no segundo experimento, embora com mais elementos, o valor da média (2,78) é maior do que seu antecessor (2,45).

Mas qual é a importância prática em se tomar valores mais ou menos próximos do verdadeiro valor do parâmetro da distribuição? A resposta depende novamente do tipo de distribuição que se esteja tentando inferir. No caso desta distribuição (Exponencial), que determina a taxa de chegadas dos clientes no banco, isto pode ser significante. Observe-se os resultados de uma nova bateria de experimentos apresentados na tabela 5.4. Gera-se, novamente com o auxílio do *Input Analyzer* do ARENA, amostras de distribuições Exponenciais com 500 observações. Para cada uma delas aumenta-se em 0,2 o valor do parâmetro. Inicia-se com média igual a 2,0 e encerra-se com valor 2,8. Estes limites foram decorrentes dos valores apresentados na experiência anterior em que se variou o tamanho da amostra (tabela 5.3).

Experimento	Parâmetro Utilizado	Parâmetro Inferido	Valor Máx. na Amostra
1	2,0	2,03	10,5
2	2,2	1,97	14,4
3	2,4	2,44	14,8
4	2,6	2,56	23,9
5	2,8	2,81	24,6

Tabela 5.4: Parâmetro empregado *versus* Máximo valor.

Observa-se que as colunas Parâmetro Utilizado e Parâmetro Inferido possuem valores muito próximos, indicando que o número de observações (500) é suficiente e determina muita confiança nos dados. O resultado mais importante da tabela encontra-se na coluna Valor Máximo Amostrado. Observe que, na medida do aumento da "média adotada", cresce, consideravelmente, o valor máximo obtido. Isto demonstra o cuidado que deve ser adotado para a determinação dos parâmetros da distribuição teórica que se queira utilizar. Com amostras pouco significativas, como as dos primeiros experimentos apresentados na tabela 5.3, vê-se que seria possível a adoção do valor 2,78, como parâmetro da distribuição. A conseqüência está apresentada no quinto experimento da tabela 5.4, com a possibilidade de ocorrência de valores grandes, durante a simulação do modelo, como os quase 25 min. entre duas chegadas de clientes no banco.

Os números "mágicos", 25 ou 30 observações, anteriormente citados são, na verdade, decorrentes de outra realidade. Pode-se resumir a questão sob dois pontos de vista: o teórico e o

prático. Quando se realizam testes estatísticos para a estimação de parâmetros, é comum se encontrar, nos textos de estatística sobre o assunto, formulações voltadas para "pequenas ou grandes amostras". As "pequenas", dizem os textos, são aquelas com menos de 25 ou 30 observações. Na verdade, existem realmente estes tipos de testes com estas indicações sobre o número de observações. No entanto, é preciso saber que com testes estatísticos adequados e com o número mínimo de observações bem amostradas (os mágicos 25 ou 30), raramente se nega a hipótese de que o parâmetro observado na amostra seja uma boa estimativa do parâmetro populacional. Provavelmente, os dados do terceiro ou mesmo do segundo experimento da tabela 5.3, sejam perfeitamente aceitáveis num teste estatístico. Assim, sob o ponto de vista teórico, 25 ou 30 observações servem, na maioria das vezes, para comprovar, estatisticamente, determinada hipótese. É bom lembrar, no entanto que, quando se está simulando um modelo existe a possibilidade de se lidar com valores extremos, cujas conseqüências devem ser previamente entendidas, sob pena de se obter resultados que, sob o ponto de vista prático, podem ser até mesmo desastrosos.

Em resumo, as experiências que se está relatando neste estudo de caso procuram bem caracterizar os cuidados a serem tomados no processo de amostragem, especialmente no que se refere a coleta de dados. Entender as conseqüências nos resultados do projeto de simulação, por conta da adoção deste ou daquele parâmetro é, e continua sendo, a principal tarefa do analista ou modelador. Tarefa esta que, no momento, nenhuma ferramenta computacional é capaz de substituir.

Faz-se agora um pequeno parêntese no estudo de caso que se está tratando. Nos próximos tópicos, aborda-se os principais elementos envolvidos no tratamento de dados. Volta-se com o estudo de caso do sistema de transações bancárias ao final deste capítulo, quando se verá como utilizar as ferramentas que o Arena possui para o auxílio do analista nas principais tarefas envolvendo análise de dados para a modelagem e simulação discreta de sistemas.

5.3 Tratamento dos Dados

Na medida em que se têm os dados necessários disponíveis, é preciso que estes recebam um tratamento adequado de forma que se possa extrair as informações desejadas. Em outras palavras, é preciso que toda a informação contida nos dados coletados torne-se acessível e, principalmente, compreensível. Uma informação numérica pode ser comunicada de várias maneiras, dependendo do propósito e do usuário a que se destina. Por exemplo, se as informações numéricas se destinam a um gerente de controle de qualidade, é razoável apresentá-los na forma de estatísticas. Se o usuário for um operador de uma máquina, talvez o desejável seja a forma gráfica. O propósito deste tópico é ilustrar duas maneiras pelas quais torna-se possível reduzir, a formas mais compactas e compreensíveis, o grande volume de dados brutos obtidos a partir da amostragem.

5.3.1 Representação Gráfica

A representação gráfica dos dados depende do tipo de variável ao qual o dado se refere. Desta maneira, pode-se considerar representações que se referem a variáveis discretas (agrupadas e não agrupadas) ou contínuas (agrupadas).

(a) Histogramas para dados discretos e agrupados

Quando a variável que se está tratando pode assumir um número muito grande de possíveis valores (mesmo que discretos), é comum representá-la na forma de grupos ou classes de valores. Para representar graficamente os resultados obtidos o indicado é a construção de um gráfico conhecido por *histograma*.

Os procedimentos para a montagem da tabela e do histograma serão ilustrados a partir de um exemplo cujos dados estão apresentados na tabela 5.5. Os dados exibem o número de defeitos encontrados em um sofisticado programa de computador. O programa, na sua versão Beta foi distribuído gratuitamente, via Internet, para ser testado. Os usuários reportam os defeitos a uma base de dados central. Os valores da tabela indicam o número de defeitos reportados por dia, durante os 100 primeiros dias.

46	52	39	43	69	31	53	52	68	17
6	64	25	88	67	85	57	60	76	60
58	96	67	94	60	73	68	66	41	60
11	38	70	82	40	94	8	86	105	65
79	65	88	54	51	114	59	93	64	31
66	68	37	109	67	59	60	62	41	50
78	97	78	55	74	67	22	40	100	27
20	44	62	72	49	82	54	73	68	38
74	75	57	86	31	82	69	51	53	63
49	70	62	46	26	36	65	83	78	19

Tabela 5.5: Numero de defeitos/dia reportados durante 100 dias

Os procedimentos para a determinação da distribuição de freqüências e do histograma dos dados coletados são apresentados passo a passo a seguir.

Passo 1:

Identificar os limites dos valores observados. Neste caso: [6; 114];

Passo 2:

Determinação das classes. Para se montar uma tabela de distribuição de freqüências deve-se, primeiramente, escolher o número de classes e seus limites. Este número depende, fundamentalmente, do número de observações a da dispersão entre os dados obtidos. Uma boa aproximação ao número ideal de classes foi proposta por Hines e Montegomery [1980]. Segundo eles tomar a raiz quadrada do número de elementos amostrados é uma boa aproximação.

Obviamente que um julgamento do resultado obtido deve ser feito para fins de obtenção de um histograma aceitável. Se os intervalos são muito grandes o histograma será grosseiro com pouca informação sobre o comportamento da variável. No outro extremo, com muitos intervalos,

a freqüência absoluta observada em cada classe se reduz muito. Com isso, a figura torna-se falha, podendo, até mesmo, algumas das classes não apresentarem freqüências, o que implica, novamente em pouca informação e clareza. A regra geral é que não existe uma maneira rápida e simples para se determinar corretamente o número de classes ou o ponto inicial onde a primeira classe deva iniciar. Neste exemplo, adota-se 12 classes com uma amplitude de 10 pontos, isto é: a primeira classe iniciando em 0 e terminando em 9; a segunda iniciando em 10 e terminando em 19; e assim por diante.

Passo 3:

Determinar a freqüência absoluta de cada uma das classes. Para realizar a tarefa monta-se uma *tabela de distribuição de freqüências* para dados agrupados. A tabela 5.6 apresenta os resultados.

Classes (defeitos reportados)	Ponto Médio x_i	Freqüência Absoluta
0 - 9	4,5	2
10 - 19	14,5	3
20 - 29	24,5	4
30 - 39	34,5	6
40 - 49	44,5	10
50 - 59	54,5	15
60 - 69	64,5	27
70 - 79	74,5	13
80 - 89	84,5	9
90 - 99	94,5	5
100 - 109	104,5	3
110 - 119	114,5	1
		Total = 100

Tabela 5.6: Distribuição de freqüências do numero de defeitos

Passo 4:

Construção do histograma das freqüências absolutas. Na construção de um histograma, os retângulos (elemento básico deste tipo de gráfico) devem ser desenhados com suas bases centradas nos pontos médios de cada uma das classes. Estes pontos servem como elementos de representação de todos os valores incluídos em cada uma das classes. Tais pontos são, também, utilizados para fins de calculo de medidas de tendência central para dados agrupados, como a média aritmética, por exemplo. O gráfico 5.2 apresenta o resultado final do histograma.

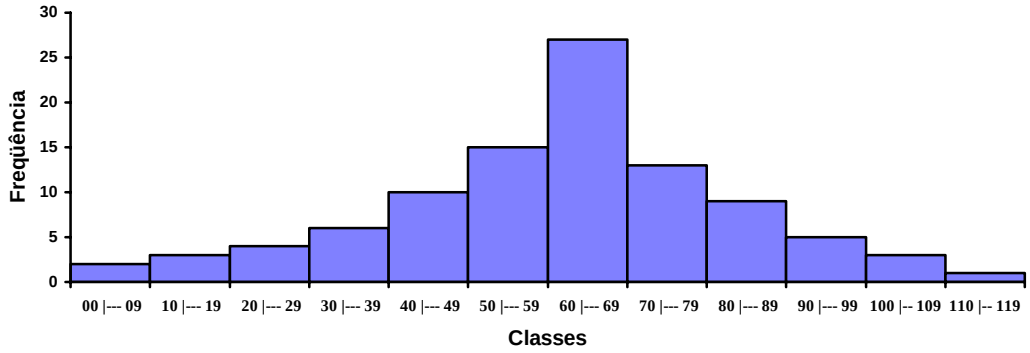

Gráfico 5.2: Histograma do número de defeitos

(b) Histogramas para dados contínuos

Embora os histogramas sejam plenamente aplicáveis a variáveis discretas (como no exemplo anterior), seu maior emprego é na representação gráfica das distribuições de freqüências de variáveis contínuas. Os procedimentos são semelhantes ao caso anterior, apenas diferindo um pouco nos métodos de determinação da amplitude e do número de classes. Um exemplo de sua construção será descrito com base nos dados apresentados na tabela 5.6. Esta tabela apresenta valores relativos a tempos de serviços em um posto de trabalho. Durante os turnos diários de 8 horas de duração, os valores foram coletados de hora em hora. A operação de aquisição dos dados foi realizada durante duas semanas não consecutivas de 5 dias de trabalho cada uma.

15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0	13.2
22.7	9.8	**6.2**	14.7	17.5	26.1	12.8	28.6	17.6	23.7
26.8	22.7	18.0	20.5	11.0	20.9	15.5	19.4	16.7	10.7
19.1	15.2	22.9	26.6	20.4	21.4	19.2	21.6	16.9	19.0
18.5	23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5	14.5
14.4	29.6	19.4	17.0	20.8	24.3	22.5	24.6	18.4	18.1
8.3	21.9	12.3	22.3	13.3	11.8	19.3	20.0	25.7	**31.8**
25.9	10.5	15.9	27.5	18.1	17.9	9.4	24.1	20.1	28.5

Tabela 5.6: Resultados da coleta de 80 valores associados aos tempos de serviço

Na tabela 5.6, pode-se observar que o menor valor, o maior valor e a diferença entre eles, obtidos no levantamento foram 6.20, 31.80 e 25.60, respectivamente. Aplicando-se a regra sugerida por Hines e Montegomery [1980], na tabela 5.6, verifica-se que o número de classes deve ser igual a 9, com a aproximação do resultado da raiz quadrada do número de dados ($\sqrt{80} = 8.94$) para o primeiro inteiro. A amplitude das classes é dada pela relação entre a diferença dos valores mínimo e máximo (6.2 e 31.8) e o número de classes. No caso das 9 classes, tem-se: 25.6 / 9 = 2.84 ≅ 3. A aproximação é sempre para o inteiro superior. A tabela de freqüências e o respectivo histograma dos dados são exibidos na tabela 5.7 e no gráfico 5.3:

Classes	Freqüências (f_j)
5.0 |--- 8.0	2
8.0 |--- 11.0	7
11.0 |--- 14.0	8
14.0 |--- 17.0	11
17.0 |--- 20.0	18
20.0 |--- 23.0	15
23.0 |--- 26.0	10
26.0 |--- 29.0	7
29.0 |--- 32.0	2
Total	**80**

Tabela 5.7: Distribuição das freqüências absolutas dos tempos de serviço

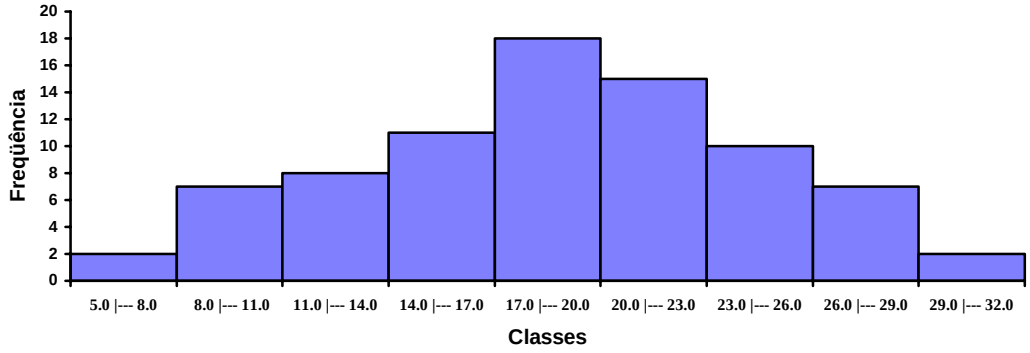

Gráfico 5.3: Histograma das freqüências dos tempos de serviços distribuídas em 9 classes

Os gráficos 5.4 e 5.5, abaixo, apresentam os mesmos dados (tabela 5.5) distribuídos sobre 5 e 26 classes, respectivamente. Embora seja possível observar um breve delineamento do comportamento da variável, tanto o primeiro como o último gráfico, não transmitem ao analista, com a mesma intensidade, a informação visual obtida junto ao histograma com 9 classes.

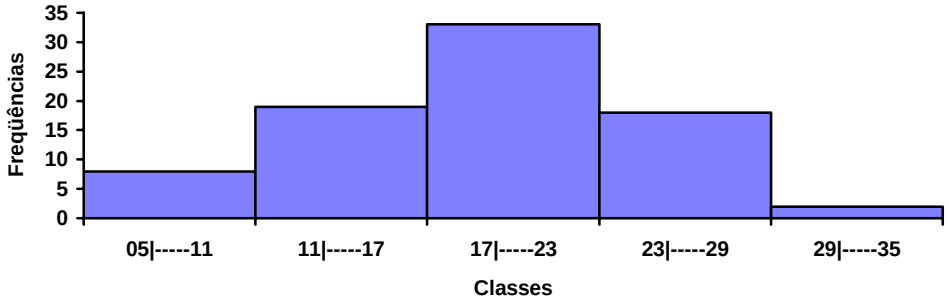

Gráfico 5.4: Histograma das freqüências dos tempos de serviços distribuídas em 5 classes

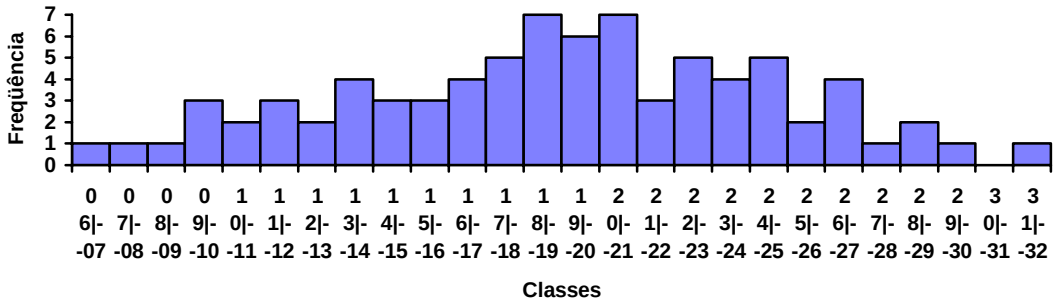

Gráfico 5.5: Histograma das freqüências dos tempos de serviços distribuídas em 26 classes

5.3.2 Medidas Descritivas e Medidas de Dispersão

Determinar os valores da média e da variância dos valores amostrados é o primeiro passo para a estimação dos parâmetros da distribuição teórica, pois, em muitas ocasiões, estes elementos servirão como estimativas diretas daqueles ou como um elemento fundamental na busca destes objetivos.

Quando os dados não estão agrupados em uma distribuição de freqüências, as seguintes fórmulas podem ser adotadas para a obtenção da média amostral ($\overline{X}$) e da variância amostral (S^2), considerando que as observações são provenientes de um conjunto de tamanho *n* e são denotadas por $x_1, x_2, \ldots, x_n$:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (5.1)$$

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} \qquad (5.2)$$

Se os dados estiverem agrupados em uma distribuição de freqüências, as equações acima podem ser modificadas como se abaixo, onde as freqüências de cada grupo são dadas por f_j e *k* representa os possíveis valores de x_j.

$$\overline{X} = \frac{\sum_{j=1}^{k} f_j x_j}{n} \qquad (5.3)$$

$$S^2 = \frac{\sum_{j=1}^{k} f_j x_j^2 - n\overline{X}^2}{n-1} \qquad (5.4)$$

Para o caso de dados contínuos, os mesmos poderão apresentar-se agrupados na forma de classes ou, não agrupados, sob a forma de dados brutos. Para o último caso, as fórmulas 5.1 e 5.2

se aplicam diretamente. Para o caso de dados agrupados, onde não mais exista a possibilidade de se lidar com os dados brutos, as fórmulas 5.3 e 5.4 podem ser aplicadas, substituindo-se os valores de x_j por m_j, onde m_j. é o ponto médio de cada classe. Neste caso, k fornece o número de classes em que a amostra foi subdividida. Por exemplo, no caso da tabela 5.7, os dados coletados foram divididos em $k = 9$ classes enquanto que na tabela de freqüências 5.8, abaixo, $k = 7$.

Exemplo

Aplicar as fórmulas 5.3 e 5.4 para calcular a média ($\overline{X}$ **)** e o desvio-padrão (s) para os valores da tabela 5.3.

Classes	m_j	f_j	$.f_j\, m_j$	$.f_i\, m_j^2$
5.0 |--- 9.0	7	3	21	147
9.0 |--- 13.0	11	10	110	1210
13.0 |--- 17.0	15	14	210	3150
17.0 |--- 21.0	19	25	475	9025
21.0 |--- 25.0	23	17	391	8993
25.0 |--- 29.0	27	9	243	6561
29.0 |--- 33.0	31	2	62	1922
Total		80	1512	31008

Tabela 5.8: Dados agrupados em sete classes ($k = 7$)

$$\overline{X} = 1512 / 80 = 18.90$$

$$S^2 = 31008 - 80.(18.90)^2 / 79 = 30.77$$

O desvio-padrão é obtido extraindo-se a raiz quadrada da variância:

$$s = (30.77)^{1/2} = 5.55$$

5.4 Identificação da Distribuição Teórica de Probabilidades

O terceiro passo no processo de análise dos dados coletados é a identificação de uma distribuição teórica de probabilidades que possa representar, da melhor maneira possível, o comportamento estocástico da variável sob análise. Neste caso, a construção de uma distribuição de freqüências e a utilização de gráficos, tais como um histograma, são muito úteis para a identificação ou delineamento da distribuição teórica de probabilidades.

A construção de um histograma permite dar inicio ao processo de inferência sobre uma distribuição teórica de probabilidades. As hipóteses sobre qual distribuição adotar devem estar baseadas no contexto do assunto investigado e no perfil do histograma obtido. Por exemplo, se os dados tratam de tempos entre chegadas e o histograma possui um perfil semelhante ao de uma distribuição exponencial, a hipótese de que os dados são gerados de acordo com aquela distribuição é quase uma garantia, uma vez que a distribuição exponencial costuma estar

associada a tempos decorridos entre chegadas de entidades em um sistema. Da mesma forma, se a variável coletada tratar, por exemplo, de pesos de pallets e o histograma apresentar-se simétrico em torno da média, assemelhando-se a uma distribuição normal, não se estará longe da verdade quando se adota esta distribuição.

No tópico que segue apresenta-se as principais distribuições teóricas de probabilidades, considerando suas características gerais, tais como, aplicações mais comuns, parâmetros, etc..

5.4.1 Principais Distribuições Teóricas de Probabilidades

As distribuições: Normal, Exponencial e Poisson são usualmente encontradas e não apresentam dificuldades em serem reconhecidas e analisadas. Apesar de apresentarem maiores dificuldades para a análise, outras distribuições como a Gama e Weibull fornecem uma grande variedade de formas e, portanto, não podem ser negligenciadas no processo de identificação.

No processo de identificação, é comum a ocorrência de um diagnóstico preliminar. Nesta fase, atribui-se a uma determinada distribuição teórica (geralmente as mais conhecidas) a responsabilidade pela geração de dados de um processo. Durante o processo de ajuste das curvas, no entanto, pode-se ter algumas dificuldades em demonstrar a aderência entre os dados empíricos e aqueles da curva teórica. É preciso examinar com a máxima cautela tais diferenças. Se as estas se encontram em um dos extremos da curva, talvez distribuições como a Gama ou Weibull possam aderir mais adequadamente aos dados.

Principais Distribuições Contínuas

Normal

Dentre as muitas distribuições contínuas citadas ao longo deste texto sem dúvida a *distribuição normal* é mais importante delas. Foi primeiramente estudada no início do século XVIII, quando alguns pesquisadores verificaram o incrível grau de regularidade associados com erros de medição. Eles descobriram que os padrões (distribuições) observados eram aproximados por uma distribuição contínua a qual se referiam como *curva normal de erros* e atribuída as leis do acaso [O'Keefe, 1989].

A conhecida curva na forma de sino descreve fenômenos simétricos em torno da média (Figuras 5.2 e 5.3). É utilizada sempre que a aleatoriedade for causada por várias fontes independentes agindo de forma aditiva. Exemplos: erros de medição ou ainda a soma ou média de amostras de um grande número de observações independentes de uma distribuição qualquer (base do *Teorema Central do Limite*). Talvez este último exemplo reflita a grande importância desta distribuição na simulação.

O *Teorema Central do Limite* estabelece que a soma ou a média resultante de um grande número de valores aleatórios e independentes é aproximadamente normal, independente da distribuição dos valores individuais. Com base neste teorema, pode-se, por exemplo, agregar os tempos de inúmeros subprocessos independentes, somando-os e substituindo-os por um único valor (soma ou média) cujo resultado tende a normalidade na medida que cresce o número de

sub-processos. Segundo Pegden (1991), a menos que estes possuam distribuições extremamente assimétricas, tal aproximação continua válida mesmo para pequenas amostras. Como será visto mais adiante no capítulo seis, este teorema ajudará, também, nos processos de análise dos resultados de simulações, uma vez que a média de uma distribuição amostral tende a normalidade, independentemente da distribuição que rege o comportamento da população da qual foi retirada.

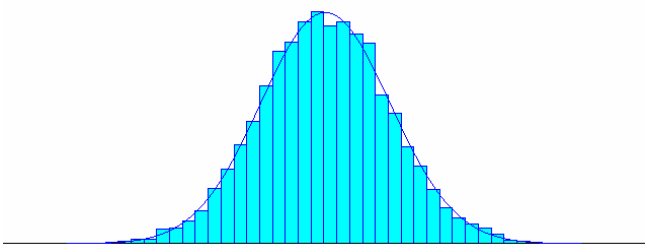

Figura 5.2: Exemplo de uma Normal com $\mu = 4$ e $\sigma = 0,5$

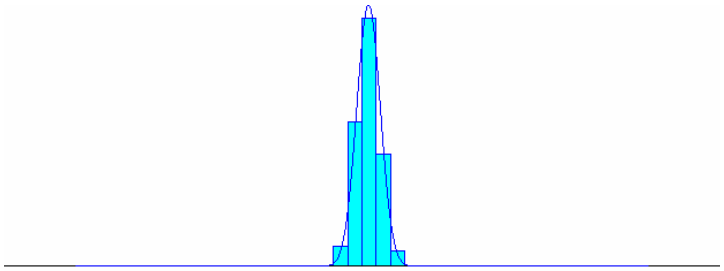

Figura 5.3: Exemplo de uma Normal com $\mu = 4$ e $\sigma = 0,01$

Uniforme

A distribuição uniforme é talvez àquela com o maior número de adjetivos. Seu emprego costuma estar associado a expressões como: a mais simples ou a que ilustra o maior desconhecimento do fenômeno aleatório sob análise. É tradicionalmente empregada quando a única informação disponível sobre a variável aleatória é que esta ocorre entre dois limites (mínimo e máximo).

Como exemplos de seu uso, pode-se citar: distância entre fonte e destino de uma entidade através de um sistema (uma mensagem em uma rede de comunicações ou uma matéria-prima em um sistema de manufatura). Sua grande importância está em seu emprego como fonte geradora de valores aleatórios no intervalo de 0 a 1, função indispensável para a aplicação dos inúmeros métodos numéricos destinados a geração de variáveis aleatórias. Na Figura 5.4 observa-se um histograma relativo a uma amostra com 10.000 pontos de uma variável aleatória com comportamento descrito por uma distribuição Uniforme no intervalo [0; 1].

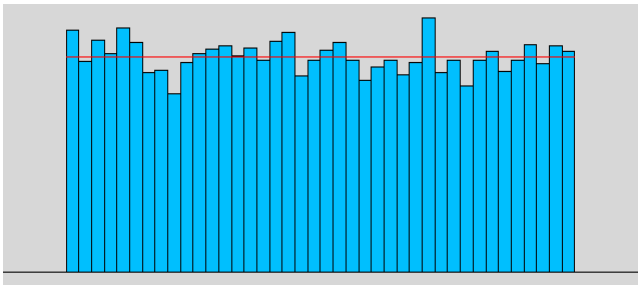

Figura 5.4: Exemplo de uma Uniforme com *a*=0 e *b*=1

Triangular

A semelhança da distribuição Uniforme, o emprego da distribuição Triangular ocorre, principalmente, quando se desconhece a curva associada a uma variável aleatória mas tem-se boas estimativas dos seus limites inferior (*a*) e superior (*b*) bem como, de seu valor mais provável (*m*). Desta forma, estimativas de comportamento baseadas na distribuição triangular oferecem perspectivas de resultados mais aderentes à realidade do que aqueles com base na distribuição uniforme. Na Figura 5.5 tem-se o histograma de uma distribuição Triangular com parâmetros *a, b* e *m* iguais a 1, 6 e 3, respectivamente.

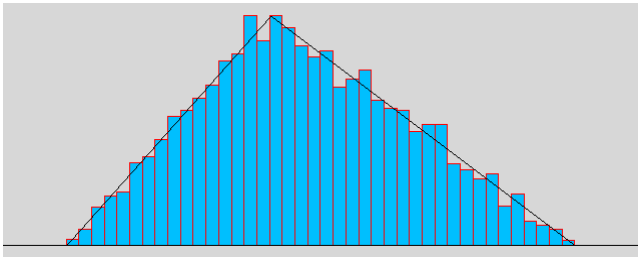

Figura 5.5: Triangular com parâmetros *a*=1, *b*=6 e *m*=3

Exponencial

A principal característica da distribuição exponencial, e razão da sua grande aplicabilidade em sistemas de filas, é sua falta de *memória*. Todo fenômeno aleatório descrito por esta distribuição se caracteriza pela total imprevisibilidade, mesmo que se conheça seu passado.

Pode-se imaginar, por exemplo, uma variável cujo comportamento é delineado por uma distribuição exponencial com média 1/λ. Suponha agora, que ao se observar uma ocorrência desta variável, se passe a contar o tempo decorrido até uma próxima ocorrência. O tempo médio decorrido entre as duas observações será 1/λ. Suponha agora que não se observe ocorrências por um longo tempo *t=x*, ainda assim, a situação será idêntica, isto é, o tempo médio esperado até a próxima ocorrência continuará sendo 1/λ. Em outras palavras, isto significa dizer que o conhecimento prévio do tempo da ocorrência do último evento não ajuda na previsão do tempo de ocorrência do próximo evento.

A exponencial é muito utilizada na modelagem de tempos decorridos entre dois eventos, particularmente se estes forem causados por um grande número de fatores independentes. Por exemplo: tempo entre duas chegadas consecutivas de entidades em um sistema ou entre duas

falhas de um equipamento. Embora não seja recomendável, devido a sua alta variabilidade, algumas vezes é empregada para caracterizar tempos de serviços.

Nas Figuras 5.6 e 5.7 pode-se observar histogramas relativos a duas amostras com 10.000 pontos cada uma delas. O primeiro refere-se a uma variável exponencialmente distribuída com média $1/\lambda = 1$. Os valores mínimo e máximo da amostra foram 0 e 9,5; respectivamente. No segundo a variável possui média $1/\lambda = 10$. Neste caso, os valores mínimo e máximo foram 0 e 95,5 respectivamente. Tais valores atestam a alta variabilidade associada a esta distribuição e os cuidados que devem ser tomados no seu emprego.

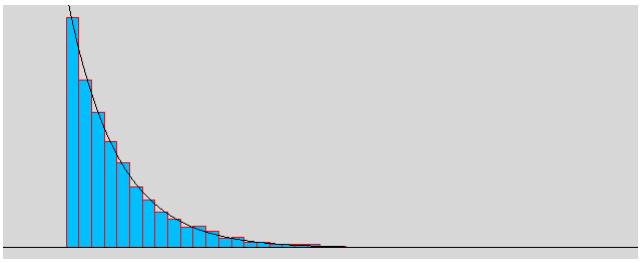

Figura 5.6: Exponencial com média $1/\lambda = 1$

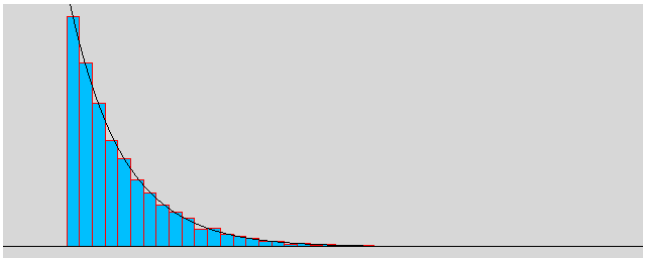

Figura 5.7: Exponencial com média $1/\lambda = 10$

Lognormal

O logaritmo natural de uma variável que segue uma distribuição normal possui uma distribuição lognormal. Quando a variável sob análise é resultante do produto de um grande número de variáveis aleatórias positivas é comum que esta variável tenha uma tendência a uma distribuição Lognormal (Figuras 5.8 a 5.10). Uma aplicação típica desta distribuição é na representação de tempos de serviços em sistemas de filas, principalmente quando esta variável possui boa aderência a distribuições não simétricas.

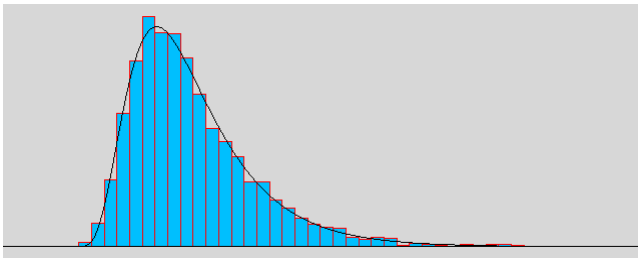

Figura 5.8: Lognormal com média =1 e desvio =0.5

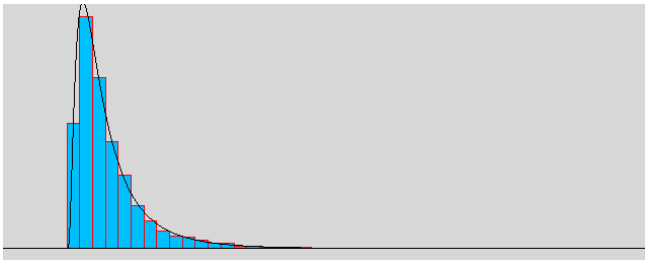

Figura 5.9: Lognormal com média =1 e desvio =1

Figura 5.10: Lognormal com media =1 e desvio =1.5

Erlang

A distribuição Erlang é utilizada como uma extensão da exponencial, especialmente quando o fenômeno aleatório é observado ao longo de diversas fases as quais podem ser descritas, de forma independente, com distribuições exponenciais. Desta forma, a soma destas *m* distribuições exponenciais de média 1/λ é uma distribuição Erlang com parâmetros 1/λ e *m.*

Pode ser empregada, por exemplo, para modelar um servidor que represente uma série de outros *m* servidores cujos tempos de processo ou serviço possam ser descritos por distribuições exponenciais com média 1/λ. Outro emprego desta distribuição é na modelagem de tempos de reparos (manutenção de sistemas) ou tempos decorridos entre falhas.

Nas Figuras 5.11 a 5.13, apresenta-se distribuições Erlang com mesma média (1/λ) e com valores crescentes para *m* (1, 2 e 3). O primeiro gráfico confunde-se com uma distribuição exponencial de mesma média.

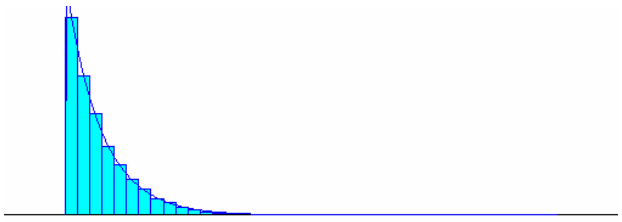

Figura 5.11: Erlang com 1/λ = 1 e *m*=1

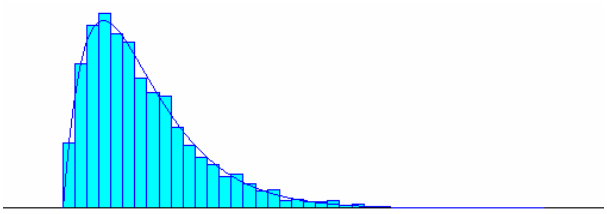

Figura 5.12: Erlang com 1/λ = 1 e *m*=2

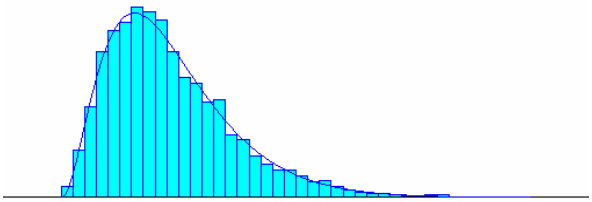

Figura 5.13: Erlang com 1/λ = 1 e *m*=3

Gama

A distribuição Gama (Figura 5.14) é uma generalização da distribuição Erlang. A diferença é que esta permite que *m* assuma valores não inteiros. Quando *m* é inteiro, as duas distribuições se confundem (Figura 5.15). É empregada em condições de modelagem semelhantes a da distribuição Erlang. A curva da distribuição Gama com média 1/λ = 1 e parâmetro *m* = 1, terá a mesma forma da Erlang, com mesmos parâmetros e da exponencial com média *a*.

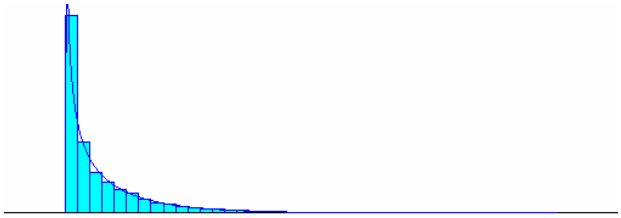

Figura 5.14: Gama com 1/λ =1 e *m* =½

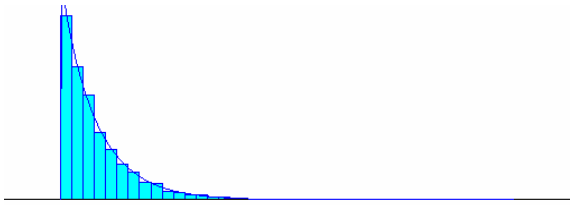

Figura 5.15: Gama com 1/λ =1 e *m* =1

Beta

A distribuição Beta (Figuras 5.16 a 5.20) é utilizada para caracterizar variáveis aleatórias cujos valores encontrem-se dentro do intervalo [0;1]. Desta maneira, uma de suas principais aplicações está na representação proporções ou frações. Como exemplos pode-se citar: proporção de defeituosos em lotes de produtos, fração de pacotes que devem ser retransmitidos, partes de trechos de estradas que anualmente necessitam reparos, etc. Outra característica desta

distribuição é o grande número de formas que ela pode assumir, dependendo de seus dois parâmetros de forma *a* e escala *b*.

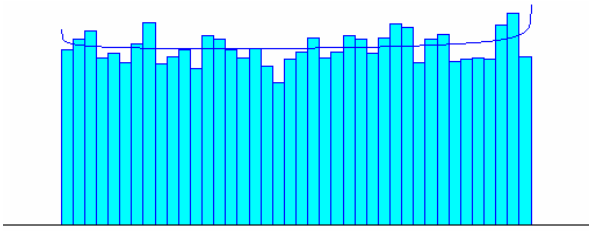

Figura 5.16: Beta com *a*=1 e *b*=1

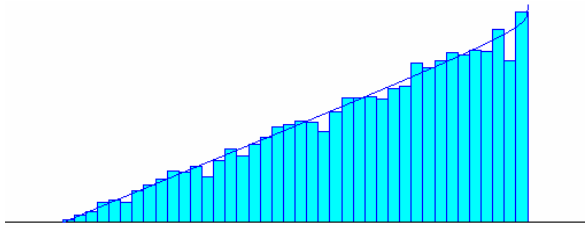

Figura 5.17: Beta com *a*=2 e *b*=1

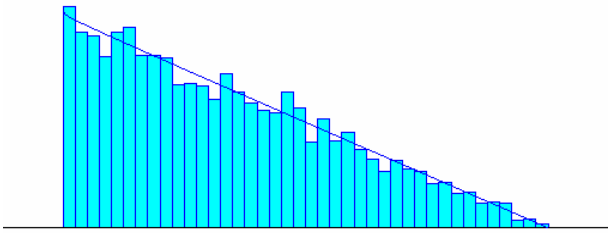

Figura 5.18: Beta com *a*=1 e *b*=2

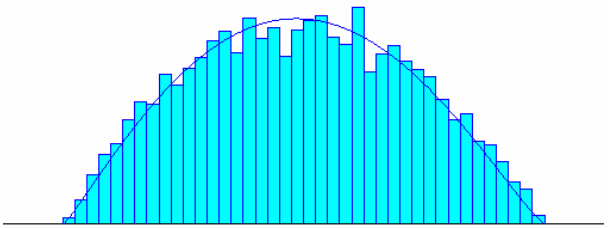

Figura 5.19: Beta com *a*=2 e *b*=2

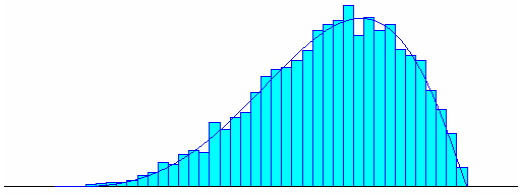

Figura 5.20: Beta com *a*=4 e *b*=2

Weibull

A principal utilização da distribuição Weibull é na representação de variáveis aleatórias que descrevam características de confiabilidade de sistemas ou equipamentos. Uma aplicação típica é na modelagem de falhas de componentes ou sistemas. Por exemplo, se um sistema é formado por inúmeros componentes que falham de forma independente e, se uma das partes falhar o sistema inteiro falha, então, o tempo decorrido entre falhas do sistema pode ser modelado utilizando-se uma distribuição Weibull (Figura 5.21 a 5.24). Assim como a Beta, esta é uma distribuição que pode assumir vários perfis dependendo de seus parâmetros, especialmente do *b* (parâmetro de forma). Por exemplo, com *b* = 3,602 a curva se assemelha a uma distribuição normal (Figura 5.24). Um perfil da distribuição Weibull poderá apresentar longas caudas à direita ou esquerda, formas de sino ou mesmo acentuados picos no valor modal.

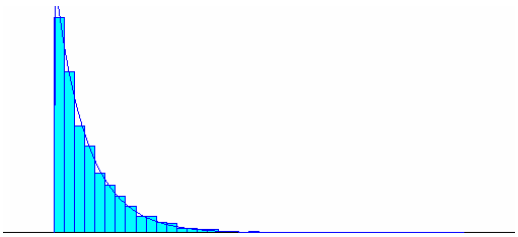

Figura 5.21: Weibull com *a*=1 e *b*=0,5

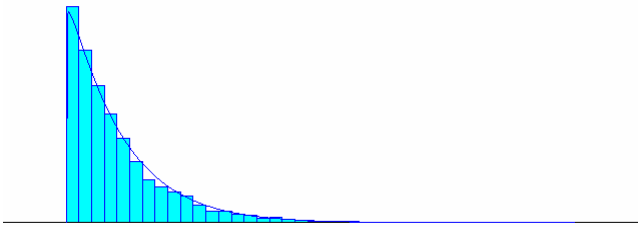

Figura 5.22: Weibull com a=1 e b=1

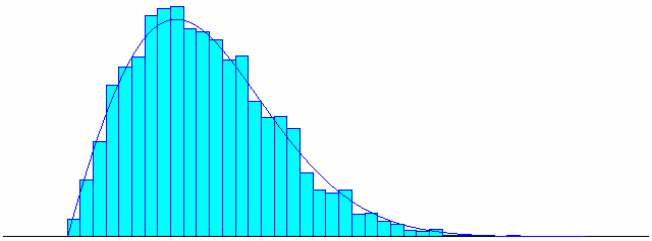

Figura 5.23: Weibull com a=1 e b=2

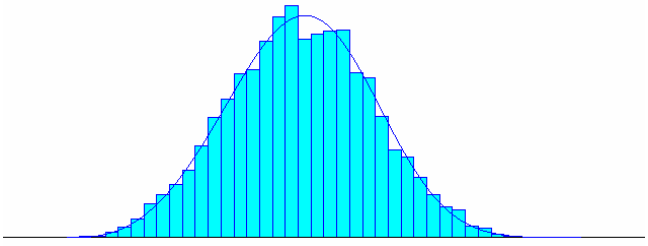

Figura 5.24: Weibull com a=1 e b=3,602

Principais Distribuições Discretas

Poisson

Na modelagem de sistemas com vistas a um tratamento analítico (por meio da Teoria das Filas, por exemplo) ou simulado, a distribuição de Poisson é, sem dúvida, uma das mais empregadas. Esta distribuição discreta é utilizada para modelar o número de ocorrências (valores discretos) que uma variável possa assumir, ao longo de um intervalo contínuo. Os exemplos de seu emprego são diversos: número de requisições feitas a um servidor em determinado intervalo de tempo, número de componentes de um sistema que falham num intervalo de tempo, número de chegadas de entidades a um sistema em determinado intervalo de tempo, etc.. Nem sempre o intervalo contínuo refere-se a intervalos de tempo. Por exemplo: número de falhas observadas a cada cem metros lineares na produção de bobinas de papel ou ainda, número de erros encontrados por lauda digitada, também são caracterizados como variáveis que se distribuem de acordo com um Poisson (Figura 5.25).

A distribuição de Poisson é particularmente apropriada quando se está modelando um sistema com inúmeras fontes independentes de chegadas. A este tipo particular de processo de chegadas, dá-se o nome de *processo de Poisson*. As distribuições Poisson e exponencial possuem um relação importante. Se os tempos decorridos entre dois eventos podem ser descritos por uma distribuição exponencial com média *a*, então diz-se que o número de ocorrências deste mesmo evento num intervalo *a* de tempo segue uma distribuição de Poisson com média $1/a = \lambda$.

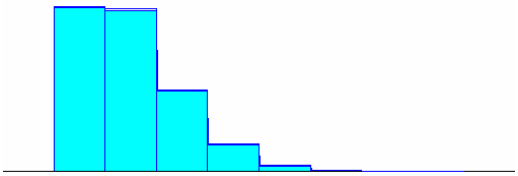

Figura 5.25: Poisson com média $1/a = 1$

Uniforme Discreta

A distribuição uniforme discreta é empregada quando a variável aleatória que está sendo modelada pode assumir apenas valores inteiros, todos com igual probabilidade, limitados a um intervalo [mínimo, máximo]. Alguns exemplos de seu emprego são: número de destinos possíveis para uma mensagem em uma rede local de computadores, número de peças que se encontram no buffer limitado em um centro de usinagem, etc..

Além destas duas, pode-se citar outras distribuições discretas tais como a geométrica, a binomial e ainda a distribuição de Bernoulli, como importantes na modelagem de sistemas.

5.5 Estimação de Parâmetros

Após o delineamento preliminar da distribuição de probabilidades, feito por meio do histograma dos dados coletados, o passo seguinte é a estimação dos parâmetros da distribuição. O

processo tem início com a determinação das medidas descritivas dos dados, tais como a média, a moda e a mediana, e das medidas de dispersão tais como a variância e o desvio-padrão amostral.

No processo de estimação dos parâmetros das distribuições sob hipóteses algumas abordagens numéricas são sugeridas pela literatura. Uma vez que se possa realizar a coleta das medidas acima citadas, as estimativas numéricas dos parâmetros são realizadas a partir da média e da variância obtidas. A tabela 5.9 apresenta algumas sugestões citadas na literatura.

Distribuição	Parâmetros	Estimadores Sugeridos
Poisson	α	$\alpha = \overline{X}$
Uniforme *(0, b)*	b	$b = \frac{n+1}{n}(max(x))$
Exponencial	λ	$\lambda = 1/\overline{X}$
Normal	μ, σ^2	$\mu = \overline{X}$; $\sigma^2 = S^2$

Tabela 5.9: Sugestões de Estimadores

No caso das distribuições Gama, Erlang e Beta, que necessitam dos parâmetros de forma (α) e de escala (β), Pegden (1990) e Law e Kelton (1991), sugerem que é possível realizar uma aproximação a estes elementos com utilização da média e da variância amostral (fórmula 5.5).

$$\alpha = (\mu / \sigma)^2 \qquad \beta = \sigma^2 / \mu \tag{5.5}$$

No caso da distribuição uniforme e da distribuição triangular, pode-se assumir que os valores mínimo e máximo são obtidos diretamente dos valores extremos dos dados amostrados, com o valor modal da distribuição triangular sendo aproximado por:

$$Mo = 3\overline{x} - (x_{min} + x_{max}) \tag{5.6}$$

Junto com a identificação das distribuições teóricas de probabilidade que regem o comportamento estocástico do sistema sob investigação, a correta estimação de seus parâmetros é um ponto de fundamental importância a fim de que o modelo construído e os resultados dele obtidos venham a possuir a maior identidade possível com este sistema. Hoje, a disponibilidade de programas computacionais que permitem a realização destas e de outras tarefas a partir dos dados amostrais, as tem tornado cada vez menos desgastante considerando apenas a necessidade de realização de cálculos matemáticos. Note, no entanto o leitor que, embora a grande maioria das tarefas relacionadas a análise de dados para a simulação possam hoje ser realizadas com o auxílio de softwares sofisticados, a compreensão dos resultados obtidos é essencial para o perfeito entendimento dos processos que estão ou venham a ocorrer no sistema real e a forma de como serão representados no modelo.

5.6 Testes de Aderência

O objetivo dos testes de aderência é a verificação da qualidade na escolha da distribuição que se acredita melhor represente os dados da população. Uma vez que se tenha levantado uma

hipótese sobre qual, ou quais distribuições teóricas são candidatas a representar os dados coletados; se tenha calculado a média e o desvio-padrão da amostra e realizado as estimativas sobre os parâmetros da(s) distribuição(ões), passa-se à última das etapas referentes à análise e ao tratamento dos dados com vistas a modelagem e a simulação de sistemas.

Assim como grande parte das etapas da análise de dados, os testes de aderência também podem ser realizados com auxílio computacional. Convém, no entanto, enfatizar uma vez mais que, mesmo adotando tal procedimento (plenamente recomendável), é fundamental que o analista entenda o que significa a aplicação do teste e seus resultados.

Usualmente os testes de aderência empregam métodos gráficos e/ou teóricos (estatísticos). Graficamente, a qualidade é medida de forma visual, isto é, de acordo com a proximidade ou "aderência" entre o desenho da distribuição teórica e aquele referente aos dados coletados. Quanto menor a diferença entre eles melhor a aderência entre os dados e a determinada distribuição. Teoricamente procura-se provar a hipótese de que o conjunto de dados amostrados não difere, de maneira significativa, daquele esperado de uma distribuição teórica especificada.

Dois dos principais métodos teóricos serão aqui revisados: Chi-quadrado e Kolmogorov-Smirnov (K-S). Ambos estão presentes na ferramenta de análise de dados do Arena, o *Input Analyser*, a qual se fará referencia e uso ao final deste capítulo. Estes testes procuram medir e avaliar os desvios entre a distribuição amostral e a teórica. A decisão de quando aplicar um ou outro teste baseia-se no tamanho da amostra disponível e na natureza da distribuição. O teste K-S é valido apenas para distribuições contínuas enquanto que o Chi-quadrado pode ser aplicado a ambos os tipos, contínuos e discretos. Em função da necessidade de pelo menos cinco observações por classe e um número razoável de graus de liberdade, não é recomendável a aplicação do teste Chi-quadrado a pequenas amostras. Geralmente, a aplicação deste teste exige conjuntos com pelo menos 100 valores, segundo alguns autores (Pegden, 1990, Law, 1991). Já o teste K-S, é aplicável a pequenas amostras.

5.6.1 Teste Chi-quadrado

Os procedimentos do teste Chi-quadrado tem início pelo arranjo das *n* observações em um conjunto de *k* classes de intervalos. Segue-se o cálculo do teste estatístico dado pela seguinte fórmula:

$$\chi^2 = \frac{\sum_k (f_0 - f_e)^2}{f_e} \qquad (5.7)$$

onde

k = número de classes ou intervalos

f_0 = frequência observada nas classes

f_e = frequência esperada nas classes

$\sum_k$ = somatório de todas as classes

Se $\chi^2 = 0$, então as duas distribuições estão "casando" perfeitamente, isto é, não existem diferenças entre a distribuição de teórica e a observada. Quanto maior o valor de χ^2, maior a discrepância entre as duas distribuições.

Pode-se demonstrar que χ^2 segue, aproximadamente, a distribuição Chi-quadrado com $\upsilon = k\text{-}1\text{-}p$ graus de liberdade, onde p é o número de parâmetros da distribuição sob hipótese, as seguintes hipóteses devem ser testadas:

1. H_0: a variável aleatória X, segue a distribuição sob hipótese com o(s) parâmetro(s) estimado(s);
2. H_1: a variável aleatória X, não segue a distribuição sob hipótese com o(s) parâmetro(s) estimado(s).

Para a decisão compara-se o valor calculado de χ^2 com os valores críticos de $\chi^2_{\alpha, k\text{-}1\text{-}p}$. Os valores críticos são fornecidos pela tabela da distribuição Chi-quadrado. A hipótese nula H_0 é rejeitada se $\chi^2 > \chi^2_{\alpha, k\text{-}1\text{-}p}$.

Na aplicação do teste é recomendável que o valor mínimo de freqüência esperada seja cinco. Se este valor for muito pequeno, ele pode ser combinado com valores esperados de classes adjacentes. Os correspondentes valores observados devem ser, da mesma forma, combinados com os de outros intervalos. Recomenda-se que para a aplicação do teste Chi-quadrado, a amostra possua pelo menos 25 elementos.

Exemplo

Com a intenção de monitorar o tráfego chamadas telefônicas sobre uma central o seguinte experimento foi realizado. A cada intervalo de cinco minutos foi registrado o número de chamadas ocorridas. Os valores esperados para o possível número de chamadas em cada intervalo são: 0, 1, 2, . . . , 13. Um total de 400 intervalos é registrado. As freqüências relativas aos valores observados foram: 3, 15, 47, 76, 68, 74, 46, 39, 15, 9, 5, 2, 0 e 1, respectivamente.

A hipótese relativa ao experimento é verificar a aderência dos dados com relação a uma distribuição de Poisson, com $\lambda = 4{,}6$. Portanto, o que se quer é testar a hipótese de poder construir um modelo deste sistema, empregando a distribuição de Poisson, com a média apontada, na geração das chamadas para a central telefônica.

Observando-se as probabilidades em uma tabela da distribuição de Poisson para a média acima e para os valores de x = 0, 1, 2, ...e 13, constrói-se a tabela 5.10.

Número de Chamadas	Freqüências Observadas	Probabilidades de Poisson	Freqüências Esperadas
0	3	0,010	4,0
1	15	0,046	18,4
2	47	0,107	42,8
3	76	0,163	65,2
4	68	0,187	74,8
5	74	0,173	69,2
6	46	0,132	52,8
7	39	0,087	34,8
8	15	0,050	20,0
9	9	0,025	10,0
10	5	0,012	4,8
11	2	0,005	2,0
12	0	0,002	0,8
13	1	0,001	0,4
	400		400,0

Tabela 5.10: Distribuições das freqüências observadas e esperadas

Os valores da coluna sob o título freqüências esperadas, são obtidos pela multiplicação dos valores da probabilidade da tabela de Poisson (3ª coluna) por 400. Portanto, de acordo com uma distribuição de Poisson com λ=4,6, a expectativa é de que em 400 períodos observados de cinco minutos, haveria 1% ou quatro destes períodos com um número de chamadas iguais a zero. O que se verifica na tabela, é que a freqüência observada foi de três chamadas.

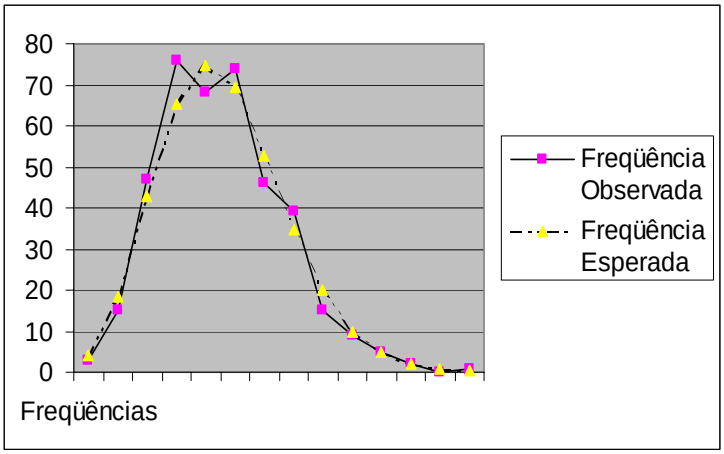

Figura 5.10: Teste de aderência visual

A partir dos valores tabelados, pode-se comparar graficamente as freqüências observadas e esperadas. A figura 5.2 apresenta as diferenças visuais do comportamento das freqüências. É possível verificar que as duas curvas apresentam um comportamento semelhante, indicando uma

tendência a aceitação da hipótese de que o comportamento da variável pode ser atribuído à distribuição de Poisson.

Para validar as conclusões obtidas de forma visual, testa-se as seguintes hipóteses sob um nível de significância α de 5%.

1. Hipótese H_0: A variável aleatória possui distribuição de Poisson com média λ=4,6;
2. Hipótese H_1: A variável aleatória não possui distribuição de Poisson com média λ=4,6;

Conforme o estabelecido, para a decisão compara-se o valor calculado de χ^2 com os valores críticos de $\chi^2_{\alpha,\, k\text{-}1\text{-}p}$. O valor crítico fornecido pela tabela da distribuição Chi-quadrado para α = 5% e $\upsilon = 10-1-1 = 8$ é igual 15,5.

É importante notar que os limites extremos da tabela possuem valor esperado e freqüência observada menor do 5, o que faz com que se agregue os valores das duas primeiras e das quatro últimas classes para se calcular as estatísticas.

Utilizando-se a fórmula do Chi-quadrado obtém-se:

$$\chi^2 = \frac{(18-22{,}4)^2}{22{,}4} + \frac{(47-42{,}8)^2}{42{,}8} + \ldots + \frac{(8-8{,}0)^2}{8{,}0} = 6{,}749$$

Como a hipótese nula H_0 é rejeitada se $\chi^2 > \chi^2_{\alpha,\, k\text{-}1\text{-}p}$ e uma vez que 6,749 < 15,5, não se pode rejeitar a hipótese de que a variável aleatória possui distribuição de Poisson com média λ=4,6. Desta maneira, pode-se afirmar, com 95% de certeza, que estatisticamente os resultados da amostra estão coerentes com a hipótese H_0.

5.6.2 Teste Kolmogorov-Smirnov (K-S)

Este teste é usado com a mesma intenção que o Chi-quadrado, isto é, testar se uma distribuição amostral segue uma determinada distribuição teórica contínua.

O teste baseia-se na comparação das probabilidades *acumuladas* das duas distribuições (observada e teórica). Para a consulta em uma tabela de valores críticos, toma-se a o maior valor K-S observado, isto é, o que corresponde ao maior desvio entre as duas distribuições.

***Exemplo*:**

Avaliar o conjunto de dados apresentados na tabela 5.11 e verificar sua aderência a uma distribuição Uniforme com α= 1%

17,38	18,09	22,47	15,29	10,33	28,98	14,70	11,26	27,49	15,90	13,47	14,43
23,73	18,09	19,09	29,29	22,12	11,86	28,31	15,79	17,48	27,78	10,27	11,94
11,77	11,72	10,72	22,20	12,05	24,28	17,33	10.42	28,78	10,16	13,63	17,31
21,56	12,61	11,76	18,37	27,00	11,86	19,90	23,92	18,61	17,38	12,66	28,29
23,17	22,28	25,24	17,58	14,66	14,41	28,59	21,72	10,56	12,48	13,02	27,84

Tabela 5.11: Dados brutos do exemplo

Na tabela 5.12 têm-se os dados brutos divididos em 10 classes associados as suas freqüências relativas e acumuladas. O valor da estatística K-S será obtido a partir das diferenças entre os valores das colunas freqüência acumulada observada e freqüência acumulada teórica.

As maiores diferenças são observadas nas classes que iniciam em 14,00 e vão até 20,00. O valor da diferença é de 0.1501. Compara-se este valor com o obtido da tabela de valores críticos do teste K-S, com α =5% e υ=60 (60 valores coletados), isto é, 0,1756. O mesmo critério de rejeição deve ser então aplicado. Como o valor crítico tabelado é maior que o valor calculado a partir dos dados da amostra, não se pode rejeitar a hipótese H_0 de que os dados levantados seguem uma distribuição Uniforme.

A escolha de qual tipo de teste utilizar quando da realização de testes de aderência baseia-se em dois fatores principais: a quantidade de elementos observados e o tipo de distribuição à qual deseja-se fazer referência. Se a distribuição sobre a qual deseja-se testar os dados for uma distribuição discreta, descarta-se a possibilidade de uso do teste K-S. Sendo aquela uma distribuição contínua, ambos os testes se aplicam.

Limites Das Classes Inf. Sup.	Freqüência Absoluta Observada	Freqüência Relativa Observada	Freqüência Acumulada Observada	Freqüência Acumulada Teórica	Diferenças Freqüência Acumulada
10,00 |–– 12,00	13	0.2167	0.2167	0.1	0.1167
12,00 |–– 14,00	7	0.1167	0.3334	0.2	0.1334
14,00 |–– 16,00	7	0.1167	0.4501	0.3	0.1501*
16,00 |–– 18,00	6	0.1000	0.5501	0.4	0.1501*
18,00 |–– 20,00	6	0.1000	0.6501	0.5	0.1501*
20,00 |–– 22,00	2	0.0333	0.6834	0.6	0.0834
22,00 |–– 24,00	7	0.1167	0.8001	0.7	0.1001
24,00 |–– 26,00	2	0.0333	0.8334	0.8	0.0334
26,00 |–– 28,00	4	0.0666	0.9000	0.9	0.0000
28,00 |–– 30,00	6	0.1000	1.0000	1.0	0.0000

Tabela 5.12: Tabela de distribuição de freqüências

O outro fator é o tamanho da amostra. Se esta for relativamente pequena, isto é, contiver menos do que 25 elementos, por exemplo, o teste Chi-quadrado não é recomendado. Lembrando,

como regra geral, é interessante que se tenha um mínimo de cinco classes cada uma com pelo menos cinco elementos para aplicação deste teste. Por outro lado, o teste K-S pode ser aplicado para conjuntos de dados de qualquer tamanho.

5.7 Ajuste de Distribuições com o Arena *Input Analyzer*

Nos itens 5.3 a 5.6, foi revisada boa parte da teoria envolvida no processo de seleção das distribuições de probabilidades, que serão empregadas na geração de dados para o modelo de simulação. Neste item, faz -se uso de uma ferramenta computacional, o Arena *Input Analyzer*, cujo propósito é auxiliar o analista nas inúmeras tarefas anteriormente abordadas: o tratamento dos dados brutos, a identificação da distribuição por meio de testes de aderência e a estimação de seus parâmetros – que visam a identificar e avaliar as melhores opções para as distribuições de probabilidades que serão empregadas.

O Arena *Input Analyzer* é uma ferramenta independente da ferramenta de modelagem. Pode ser acessada a partir da barra de ferramentas do Arena, via menu *Tools* (figura 5.10) ou acionada diretamente, a partir da janela do Arena no Windows.

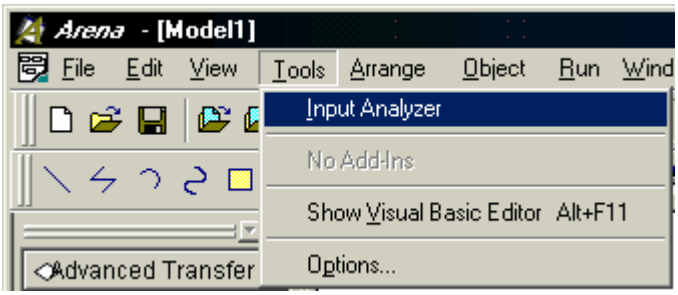

Figura 5.10: Ativando o *Input Analyzer* a partir do *Menu/Tools* do Arena

O principal propósito do *Input Analyzer* é a identificação da distribuição teórica de probabilidades por meio de testes de aderência. Para tanto, o usuário deve possuir uma amostra de dados (IID - Independente e Identicamente Distribuída) coletados no sistema real. Ao final dos procedimentos o aplicativo fornece ao uma expressão válida para ser empregada em modelos desenvolvidos no Arena. Para melhor entender o emprego desta ferramenta considera-se, novamente, o estudo de caso visto no item 5.2 quando se abordou os tópicos: Processo de Amostragem e Coleta de Dados.

Naquele estudo de caso tratado, entitulado – "Amostragem para um Modelo de uma Agência Bancária", dois importantes parâmetros exigiam um tratamento adequado para que se pudesse empregar uma distribuição de probabilidades no modelo. São eles: o período de tempo decorrido entre duas chegadas de clientes no banco e o tempo necessário para estes procederem com suas transações. No modelo, estes valores foram armazenados na variável TECc (Tempo Entre Chegadas) e no atributo TSc (Tempo Serviço).

Inicia-se analisando os dados levantados para a variável TECc. Para que se possa fazer uso do *Input Analyzer* é preciso criar um arquivo contendo os dados que serão tratados. Este arquivo poderá ser oriundo de editores de texto, planilhas eletrônicas ou banco de dados. O formato deve ser ASCII. Os dados devem estar separados por brancos (espaços, tabulações ou novas linhas).

2,26	1,86	2,65	0,10	5,52
3,01	1,81	0,07	4,74	2,47
4,65	4,91	2,17	2,88	1,15
0,98	0,67	6,63	2,01	4,06
1,74	4,74	0,72	0,17	0,97
1,93	3,02	1,06	2,75	0,17
1,34	0,51	0,26	2,05	1,64
0,46	3,64	0,30	1,41	4,69
0,45	2,59	0,10	1,61	1,02
0,49	0,64	1,38	2,08	0,50

Tabela 5.13: Amostra dos TECc no período entre 10 e 11 horas.

A tabela 5.13 apresenta os dados coletados no período entre 10 e 11 horas da manhã, que estabelece a demanda tipo A (ver tabela 5.1). Os dados pertencem a um conjunto de tamanho 50, inicialmente julgada adequada para o tipo de variável que se esta tratando. Os valores representam tempos, em minutos (formato decimal). Os dados foram salvos em um arquivo tipo texto denominado *chegadas1011.txt.*

Para que se possa tratar os dados, deve-se criar uma área de trabalho no *Input Analyzer*. Isto é feito ativando o menu *File/New* e o menu *File/DataFile/Use Existing …*, para carregar o arquivo com os dados amostrados. A figura 5.11 ilustra o procedimento.

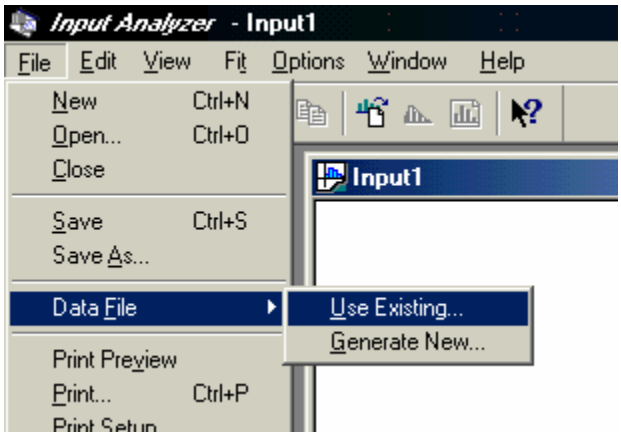

Figura 5.11: Carregando um arquivo de dados amostrados

Ao ativar a opção *Use Existing...*, uma janela padrão do Windows para busca de arquivos será aberta. O analista deve identificar o arquivo com os dados e carregá-lo. Note que será preciso alterar o tipo de extensão apresentada. Por *default*, o *Input Analyzer* sempre procura arquivos com sua extensão padrão (*.dst). Quando uma seção de trabalhos for salva a partir do

Input Analyzer, esta será a extensão utilizada pelo software. Como o arquivo exemplo possui extensão *txt*, altere o campo **.dst* para **.txt*, e clique em Abrir. Dentre os arquivos listados, aparece o arquivo *chegadas1011* (figura 5.12).

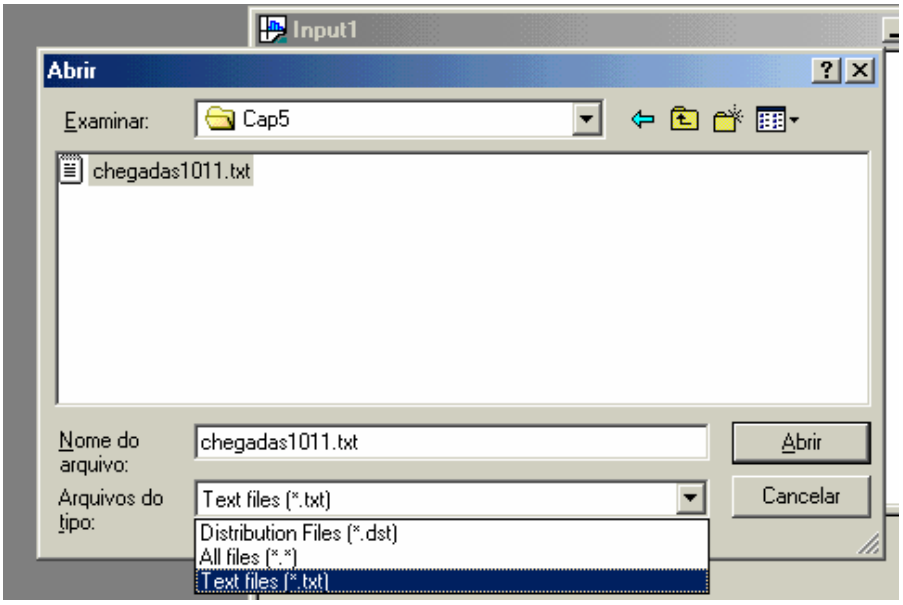

Figura 5.12. Busca do arquivo *chegadas1011.txt*

Uma vez carregado o arquivo, o *Input Analyzer* inicia seu trabalho (Figura 5.13). A primeira tarefa realizada pela ferramenta é tratar os dados brutos do arquivo, organizando-o em uma tabela de distribuição de freqüências para obtenção de um histograma (ver item 5.3.1).

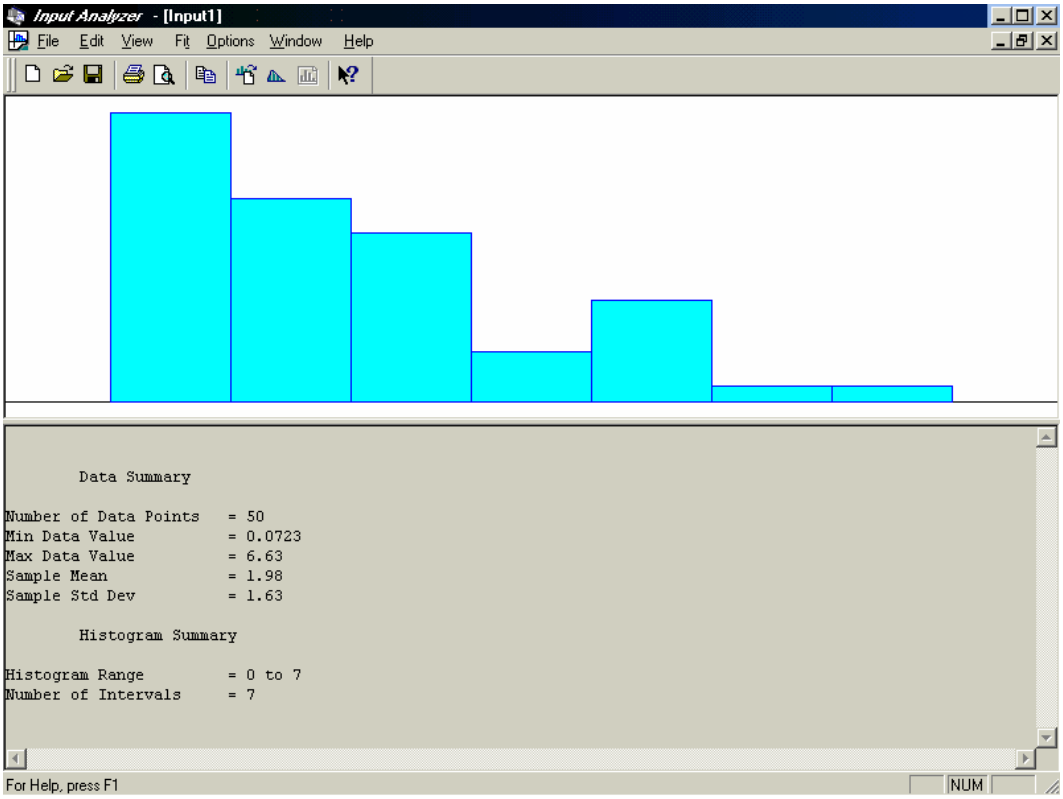

Figura 5.13: Histograma para os dados do arquivo *chegadas1011.txt*

A passo seguinte será iniciar os procedimentos para ajuste de uma distribuição de probabilidades adequada com seus respectivos parâmetros.

5.7.1 Procedimentos para o Ajuste a uma Distribuição Específica

A simples observação do histograma apresentado já permite que o analista inicie o processo de busca de uma distribuição de probabilidades que possa representar, adequadamente, o comportamento da variável aleatória sob estudo.

A partir dos conhecimentos adquiridos na breve revisão sobre as principais propriedades e aplicações de algumas distribuições de probabilidades (ver item 5.4.1 – *Principais Distribuições Teóricas de Probabilidade*) e do tipo de variável aleatória que se está lidando (a variável TECc, que expressa o tempo decorrido entre dois eventos), tem-se a expectativa de que se trata de uma variável, cujo comportamento pode ser estabelecido por uma “distribuição exponencial”.

Entenda o leitor, que este tipo de “intuição” do analista não é determinada por nenhum tipo processo esotérico ou de adivinhação. Com o tempo, e a devida experiência acumulada, torna-se comum à possibilidade de pré-determinar a distribuição de probabilidades mais adequada, mesmo sem observar o histograma dos dados. De fato, como foi visto, cada uma das distribuições revisadas pode bem caracterizar uma ou mais situações de fenômenos naturais ou experimentos que lidam com variáveis aleatórias, como aqui neste problema.

No entanto, é preciso deixar claro, que o uso do *Input Analyzer* não impõe ao usuário uma prévia experiência. Como se verá, a seguir, esta ferramenta pode ser muito útil tanto àqueles com pouco ou nenhum conhecimento em ajuste de curvas, como aos mais veteranos no assunto.

Voltando ao problema, imagine como agiria um usuário mais experimentado que, observando o histograma da figura 5.13, acreditasse se tratar de uma variável aleatória exponencialmente distribuída. A partir da situação apresentada na tela do *Input Analyzer*, o usuário tem a opção de realizar a estimação dos parâmetros e os testes de aderência, como os que foram vistos nas seções 5.5 e 5.6.

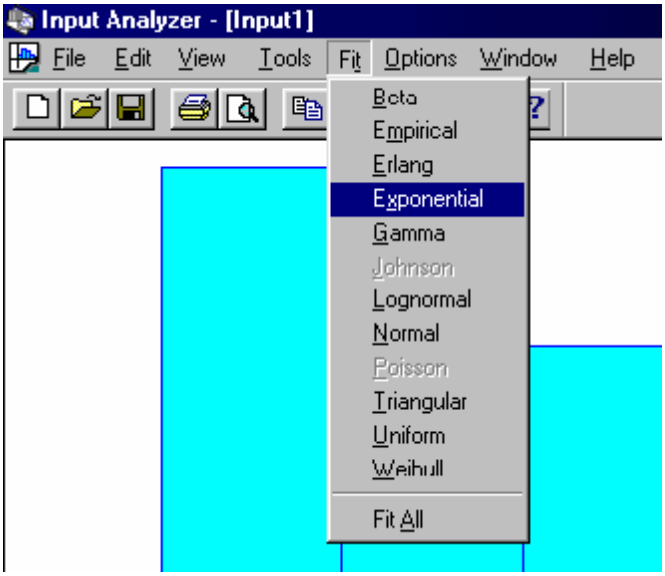

Figura 5.14: Menu Fit para a escolha da distribuição a ser testada.

O procedimento é simples. A partir do menu *Fit*, o usuário escolhe qual distribuição deseja testar. Neste caso, busca-se a distribuição *Exponencial* (Figura 5.14). Observe que a última das opções vista na figura 5.14 (*Fit All*) permite ao analista testar, de uma só vez, todas as distribuições disponíveis no *Input Analyzer*. O uso desta opção será visto no item 5.7.2 (Processo de Ajuste a todas as Distribuições). Uma vez escolhida a distribuição, a ferramenta processa e apresenta os resultados dos testes realizados.

A figura 5.15 apresenta o resultado gráfico que o *Input Analyzer* apresenta. Sobreposto ao histograma, a linha cheia exibe a função densidade da distribuição exponencial, cuja média foi estimada a partir da média dos dados observados. Pode-se verificar que, em algumas classes, a diferença entre o valor verdadeiro e os valores amostrados é bastante acentuada. Além destas informações gráficas apresentadas pelo *Input Analyzer*, a qualidade do ajuste pode ser testada, também, por meio de três medidas realizadas e expressas em valores numéricos.

Logo abaixo da janela do gráfico, uma outra janela (texto) expõe os resultados dos testes realizados. É possível observar a expressão numérica da distribuição ajustada: EXPO (1.98). Esta expressão indica que o processamento ajustou os dados, seguindo a orientação do usuário, a uma distribuição exponencial com média de 1,98 minutos. No entanto, devido a diferenças entre os

valores teóricos esperados e os valores amostrados (diferenças já observadas visualmente), o erro apresentado neste ajuste é indicado na expressão *Square Error: 0.011617*.

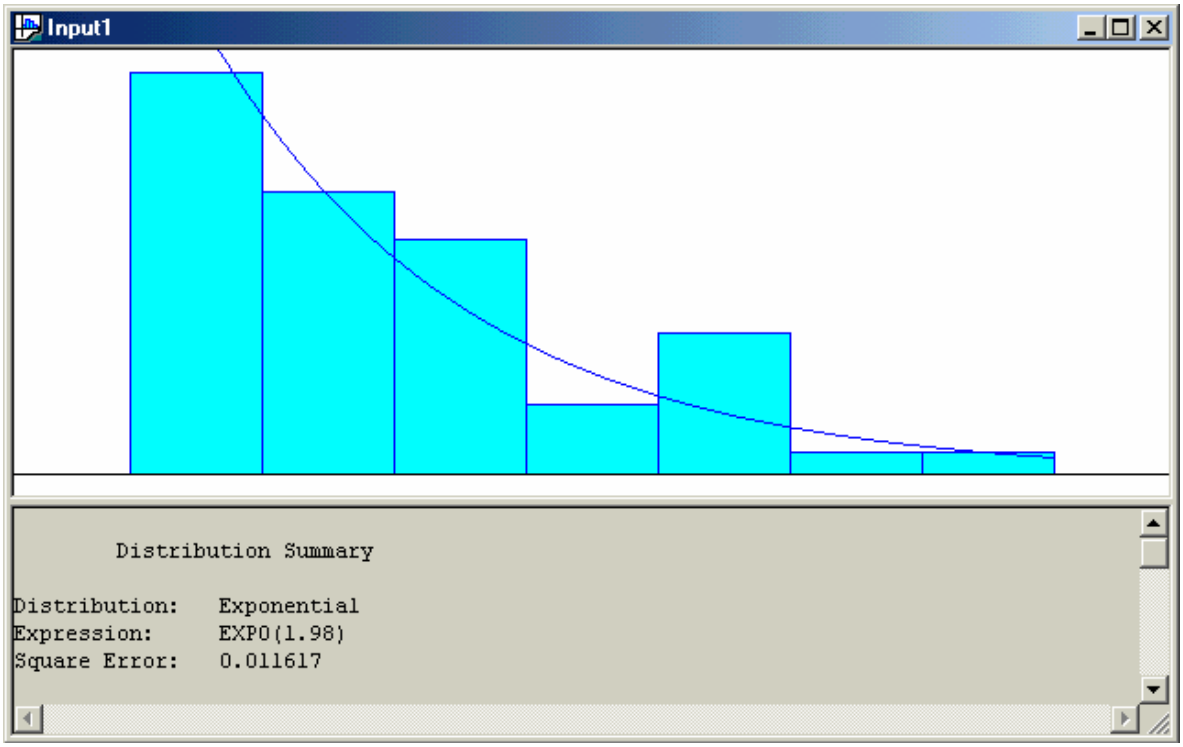

Figura 5.15: Resultado do processo de ajuste da distribuição exponencial

Este número, é o primeiro indicador de qualidade. Ele representa o valor médio das diferenças (tomadas ao quadrado), entre os valores das freqüências observadas nos dados amostrais e os valores das freqüências relativas da distribuição ajustada (no caso uma Exponencial com média 1,98). Quanto menor este valor, melhor é o ajuste.

Mais adiante será visto como é possível realizar este mesmo tipo de ajuste a todas as distribuições que se apliquem aos dados. Pode-se construir um lista ordenada das distribuições ajustadas, desde as mais adequadas aos dados (menor erro quadrado) até àquelas com os piores índices de erro.

Além das expressões relativas a distribuição e seus parâmetros, bem como ao valor do erro no procedimento de ajuste, outras informações sobre os testes realizados são apresentadas na janela sob o gráfico. São estas informações que ajudam ao analista a tomar a melhor decisão quanto à distribuição a ser adotada.

Dois testes padrões são realizados: o teste *Chi-quadrado* e o teste *Kolmogorov-Smirnov* ou *K-S*. Estes testes (ambos já tratados no item 5.6 – Testes de Aderência), são empregados para testar a hipótese de que a distribuição teórica escolhida se ajusta aos dados amostrados.

A figura 5.16 que apresenta as informações exibidas no *Input Analyzer* após a realização do ajuste (*Fit*) e dos testes de aderência. O primeiro conjunto de informações reporta o resultado do teste *Chi-quadrado* e o segundo do teste *K-S*.

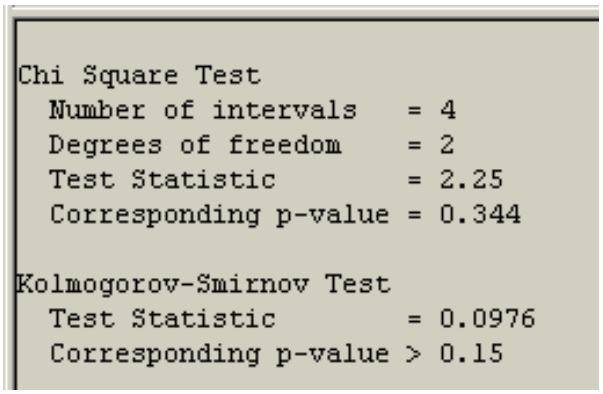

Figura 5.16: Resultados dos testes *Chi Square e K-S*.

Para melhor entender o significado dos números apresentados, lembre-se do que foi visto sobre os testes de aderência no item 5.6. Em cada teste, duas hipóteses são levantadas:

H_0: a variável aleatória *X*, segue a distribuição sob hipótese com o(s) parâmetro(s) estimado(s);
H_1 a variável aleatória *X*, não segue a distribuição sob hipótese com o(s) parâmetro(s) estimado(s).

Para a decisão, tanto num teste como no outro, compara-se o um valor calculado, obtido a partir dos dados amostrais, com valores críticos que são fornecidos por tabelas das distribuições *Chi-quadrado* e *K-S*. Quando se fala em tabelas, se está referindo ao procedimento de busca dos valores críticos em tabelas das funções, como foi realizado nos exemplos de testes de aderência vistos na seção 5.6. Obviamente que, quando tal comparação é realizada por uma ferramenta computacional como o *Input Analyzer*, esta é feita entre dois valores calculados.

A hipótese nula H_0 é rejeitada se o valor calculado for maior que o valor crítico tabelado. Dentre as informações apresentadas na Figura 5.16, observa-se os valores calculados de *Chi-quadrado* e *K-S*. Para o caso dos dados amostrados no problema, os valores críticos tabelados (Anexo 3), para um nível de significância α de 5% são: 5,99 para o *Chi-quadrado* e 0,19233 para o *K-S*, respectivamente. Considerando pois, os valores da Figura 5.16, vê-se que a hipótese H_0 não foi rejeitada em nenhum dos testes, uma vez que $2,25 < 5,99$ e $0,0976 < 0,19233$.

Adicionalmente aos valores das estatísticas, outro importante índice de qualidade do processo de ajuste é fornecido pelo *Input Analyzer*. Trata-se do valor de *p* ou como visto na Figura 5.16, *Corresponding p-value*. Segundo Kelton (1998), "o valor de *p* está associado à probabilidade de se obter um novo conjunto de dados que seja mais inconsistente com a distribuição ajustada, do que o conjunto de dados atualmente utilizado. Considerando que a distribuição ajustada seja a verdadeira distribuição da variável aleatória que se esta tratando". Em outras palavras, se o valor de *p* for "grande", existe uma grande possibilidade de o atual conjunto

de dados ser apropriado ao ajuste que se está realizando. Por outro lado, se *p* for "pequeno", é provável que se possa obter melhores resultados para o processo de aderência, se um novo conjunto de dados for utilizado. Lembre-se o processo de aderência depende, fundamentalmente, dos dados levantados. Quanto ao significado de *p* "pequeno ou grande", a literatura indica que para valores menores do que 0,05 é aconselhável não confiar nos resultados do ajuste realizado. Com valores de *p* maiores do que 0,10 pode-se dizer que a distribuição teórica obtida é candidata.

Para encerrar os procedimentos, basta copiar a expressão obtida empregando-a no modelo. No caso do modelo das transações bancárias, a expressão EXPO (1.98) é utilizada para determinar os valores a serem atribuídos a variável "Tempos entre Chegadas nos Caixas" (TECc), a qual determina o intervalo de tempo decorrido entre cada um dos clientes no banco, no período entre 10:00 e 11:00 horas.

5.7.2 Processo de Ajuste a todas as Distribuições

Considere agora, a possibilidade de que o analista não possua nenhuma experiência prévia em processos de ajustamento de curvas. Neste caso, a partir da observação do histograma apresentado na figura 5.13, seu procedimento deve ser um pouco diferente daquele tomado pelo analista experiente, relatado ao longo do item 5.7.1.

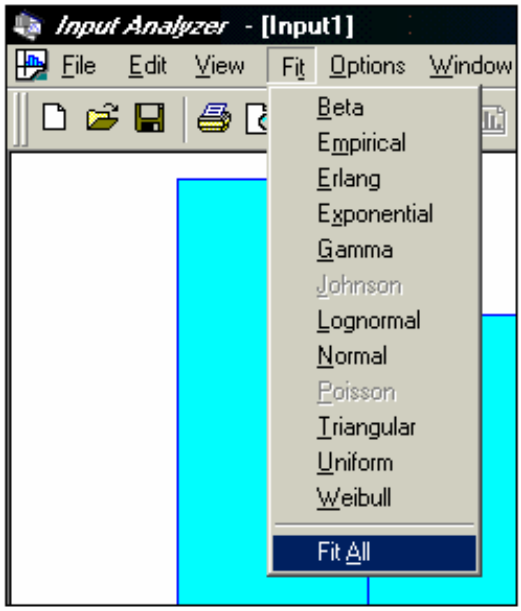

Figura 5.17: A opção Fit All do menu Fit

Conforme visto anteriormente, uma vez que os dados brutos contidos no arquivo já tenham sido tratados e o histograma montado, ao buscar a opção *Fit* do menu principal, é possível requisitar que o programa realize testes de aderência com todas as possíveis distribuições, às quais os dados se apliquem. Para tanto, basta que o usuário selecione a opção *Fit All.* (veja Figura 5.17 ao lado).

Neste ponto, o *Input Analyzer* diferencia conjuntos de dados somente números inteiros, de conjuntos contendo números reais. Para conjuntos de dados reais (basta que um elemento do conjunto contenha um ponto [.]) o processo de ajuste busca realizar os testes com as distribuições contínuas de probabilidades. Caso contrário, testa com as distribuições discretas.

Uma vez escolhida a opção Fit All, o *Input Analyzer* processa as informações e apresenta os resultados. A Figura 5.18 exibe o resultado dos testes conforme a interpretação da ferramenta.

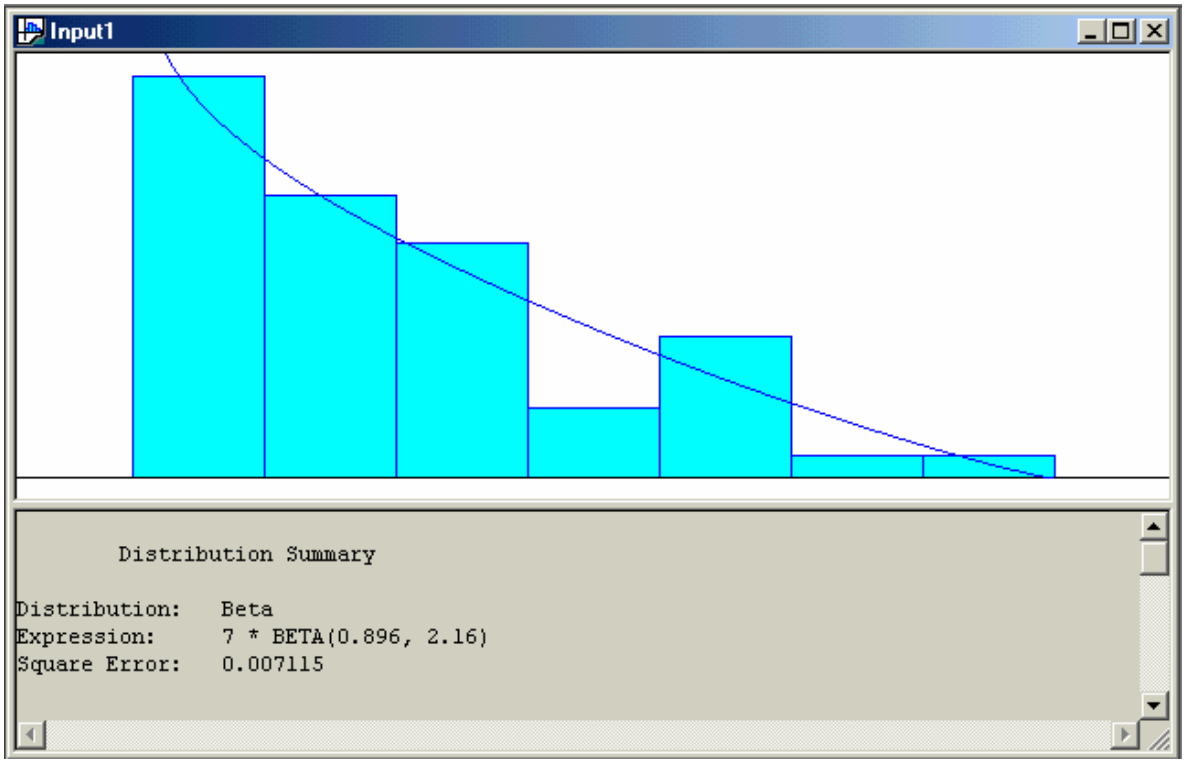

Figura 5.18: Resultado do processo de ajuste à melhor distribuição.

Como pode ser visto, a distribuição de probabilidades escolhida foi uma Beta. Sua expressão, conforme os dados amostrados é 7 + BETA (0.896, 2.16). As primeiras dúvidas que surgem na mente do usuário inexperiente são: que tipo de distribuição é essa? Posso simplesmente copiar a expressão e empregá-la no modelo? Por que o programa escolheu esta função? E a distribuição exponencial escolhida anteriormente pelo analista "experiente" é melhor ou pior que esta? Estas dúvidas são tratadas por partes.

Primeiro examina-se o processo de escolha realizado pelo *Input Analyzer*. Quando a solução anterior foi apresentada, EXPO (1.98), viu-se que o erro admitido naquela escolha foi de 0,011617. Agora, o erro apontado na Figura 5.18, é de 0,007115. Nos testes de aderência realizados pelo programa, para cada curva testada é calculado um valor para o erro. Após realizar este cálculo para todas as possíveis distribuições de probabilidade, o *Input Analyzer* ordena as soluções conforme o valor do erro. Para verificar os resultados dos cálculos efetuados pela ferramenta, basta buscar no menu *Window*, a opção *Fit All Sumary*, que contém a lista ordenada e os valores dos erros. Veja a figura 5.19

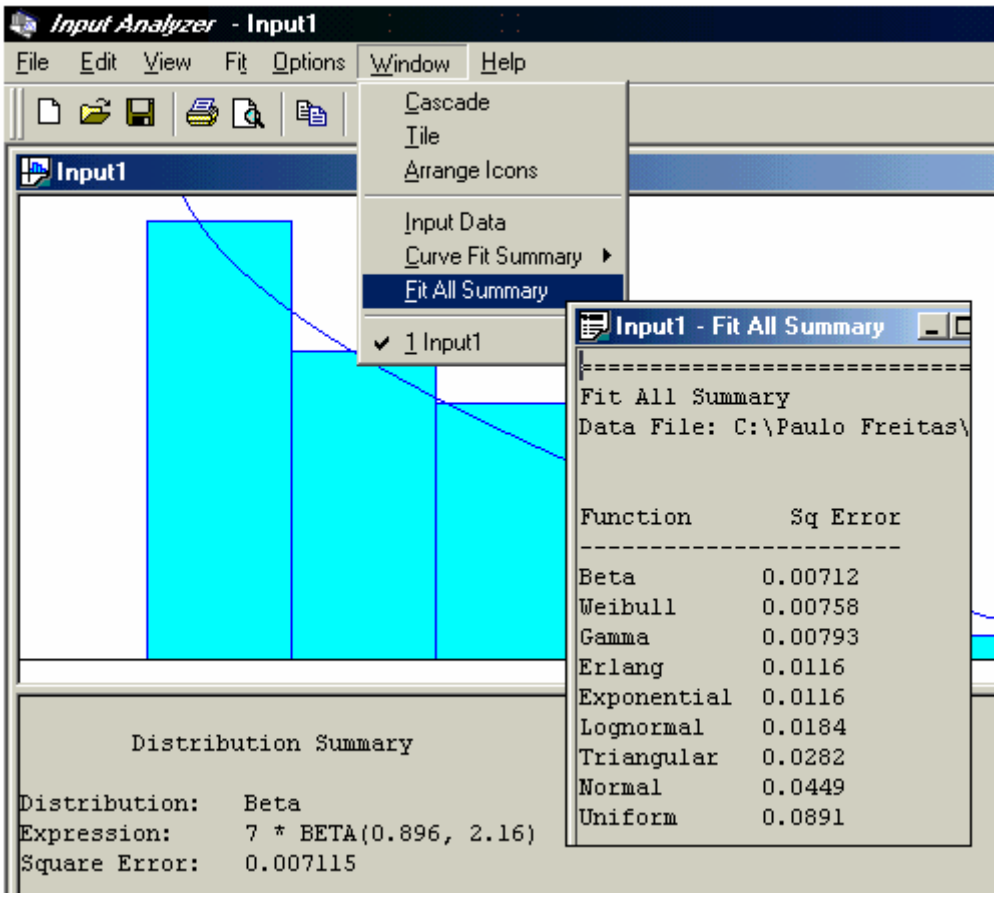

Figura 5.19: Resumo dos cálculos do erro para cada distribuição.

A lista apresenta o valor do erro ao lado de cada função testada. Agora é possível entender porque o programa escolheu a função BETA. Simplesmente porque esta apresenta o menor erro. É possível que na lista apresentada o leitor reconheça algumas distribuições. As mais comuns aos problemas introdutórios de simulação, tais como: a Exponencial, a Triangular e a Normal. Neste ponto talvez seja interessante voltar ao item 5.4 e revisar o texto sobre distribuições. Por hora, é importante fixar o processo de escolha e os pontos importantes a serem observados pelo analista neste processo.

O primeiro ponto responde uma das questões formuladas anteriormente. Posso simplesmente copiar a expressão e empregá-la no modelo? Isso dependerá de quão representativa é a amostra que foi utilizada. Se esta é uma boa representação da população à qual pertence a variável de interesse (neste caso TEC), não existem maiores impedimentos para que se adote a expressão apontada. Afinal este é o propósito de usar o *Input Analyzer*.

Observe-se agora a grandeza dos valores apresentados. Examinando a lista ordenada com as distribuições, verifica-se que as três primeiras possuem valores para o erro muito próximos uns dos outros. Isto pode significar que a adoção de qualquer uma delas, com seus respectivos parâmetros, produzirá efeitos semelhantes sobre o modelo. Lembre-se, considera-se que a amostra é correta. Relembrando o que foi visto no item 5.4 sobre as distribuições, sabe-se que as

distribuições *Beta*, *Gama e Weibull*, dependendo de seus parâmetros de forma e escala, podem ter inúmeros formatos e com relativa freqüência aderir ao conjunto de dados.

Este tipo de dúvida não é exclusiva dos menos experientes. A diferença, é que na medida que se conhece o comportamento de determinadas variáveis, pode-se ignorar a escolha principal das ferramentas de ajuste de curvas, e usando as informações fornecidas, isto é, erro quadrado, valor da probabilidade *p* e das estatísticas *Chi-quadrado* e *K-S*, assumir outra distribuição que não aquela apresentada como a melhor pela ferramenta. Em outras palavras, usar as informações fornecidas pela ferramenta é o objetivo. Quanto à decisão, procure tomá-la com base nas informações colhidas.

Considere, por exemplo, o uso da distribuição Exponencial (4ª colocada no ranking da *Figura* 5.19), em detrimento das três primeiras funções. Examine os números fornecidos na Figura 5.20, quando se pode comparar todas as funções em conjunto.

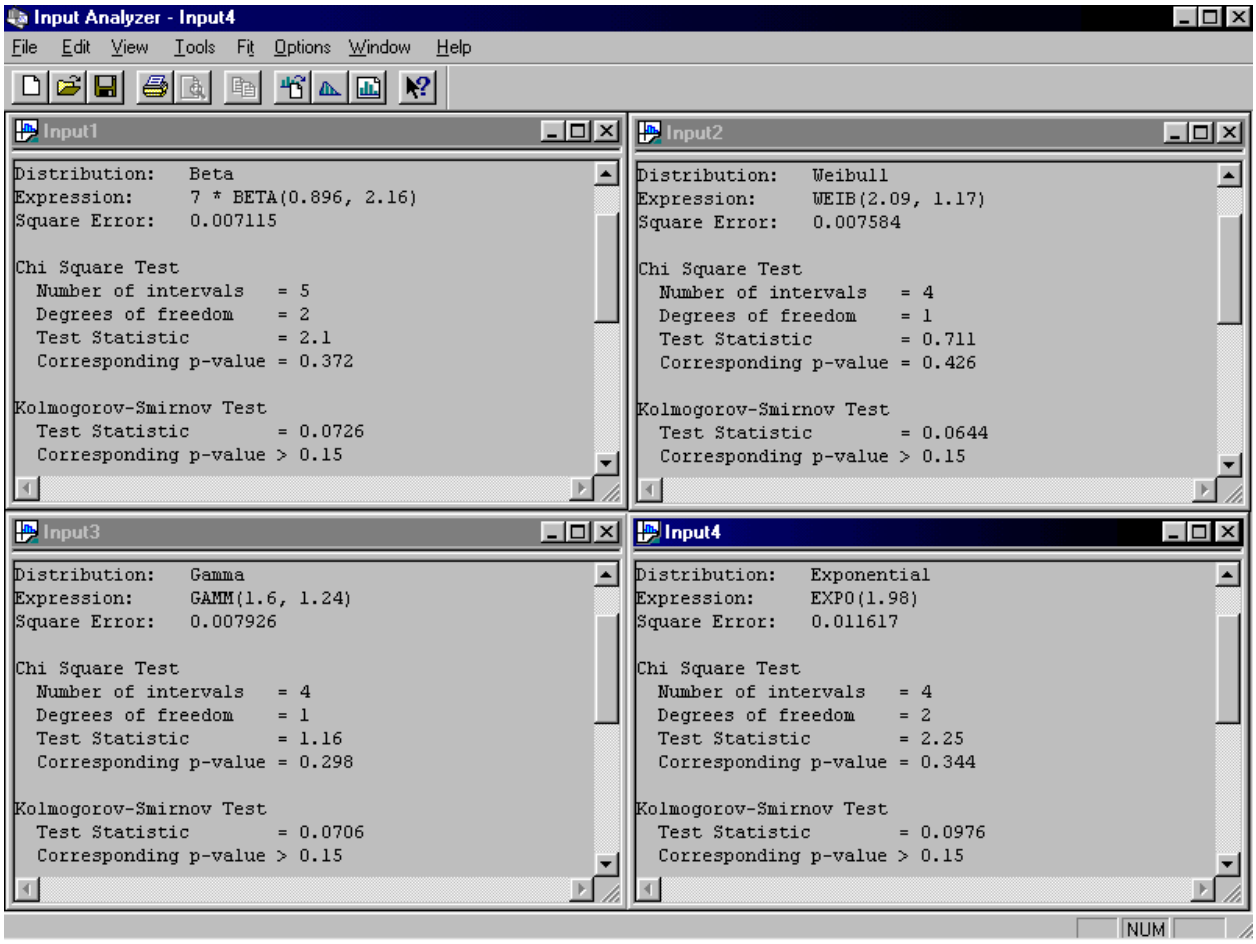

Figura 5.20: Comparação das informações fornecidas pelo *Input Analyzer* para as quatro funções.

Observando-se os principais elementos de decisão (erro, *p* e estatísticas), vê-se que nenhuma das quatro funções apresenta, nos testes realizados, valores de *p* que permita desconfiar dos ajustes realizados. Da mesma forma, todas as quatro distribuições apresentam valores para os testes estatísticos (tanto para Chi-quadrado quanto para o K-S) que implicam na aceitação da

hipótese H_0, a qual supõe que a variável aleatória X segue a distribuição sob hipótese com o(s) parâmetro(s) estimado(s). Sendo assim, a conclusão a que se chega é que, a menos do valor do erro quadrado, todas as outras informações nos permitem fazer uso de qualquer das funções, inclusive da Exponencial.

Quanto ao risco de usar uma distribuição exponencial com erro associado ao ajuste um pouco maior do que o das primeiras do ranking, este pode ser assumido, na medida que se conheça onde pode ser aplicada cada uma das distribuições de probabilidades listadas. Neste caso, a escolha seria, sem sombra de dúvidas, a distribuição Exponencial (típica distribuição associada a processos de chegadas).

Uma forma de garantir que a escolha foi correta é aumentar o tamanho da amostra e realizar novamente os testes de aderência. Lembre que a distribuição exponencial trata de variáveis aleatórias com alto grau de variabilidade e, portanto, quanto maior o conjunto de dados, melhor os resultados obtidos. Só para exemplificar, imagine que o conjunto de dados que se está tratando, recolhido nas chegadas de clientes no banco entre 10 e 11 horas da manhã, tenha agora 100 elementos.

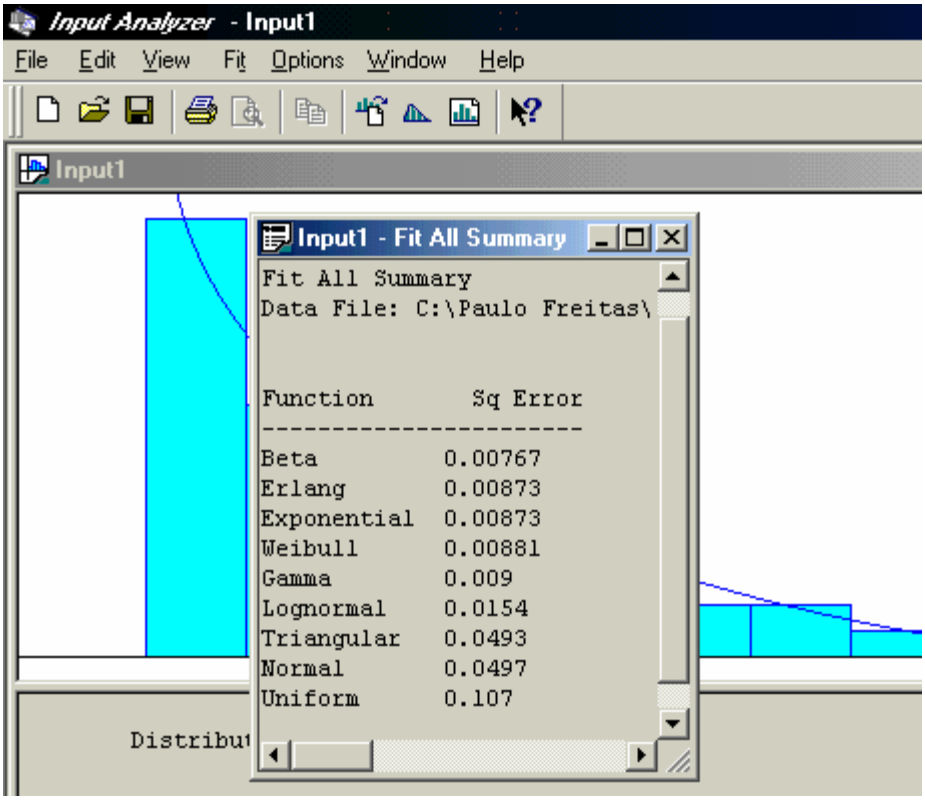

Figura 5.21: Resultados do *Fit All Sumary* para a nova amostra com 100 elementos

O novo arquivo testado, chamado "*chegadas1011-100.txt*", tem seus primeiros 50 elementos iguais aos do arquivo *chegadas1011.txt*, anteriormente utilizado. Realizando

novamente o procedimento de ajuste a todas as distribuições, *Input Analyzer* apresenta o ranking exibido na figura 5.21.

Vê-se agora, que o único argumento que ainda deixava alguma dúvida sobre adotar ou não a distribuição Exponencial foi agora derrubado. Observe que as diferenças nos valores dos erros para as primeiras quatro distribuições estão muito próximas. Não pode haver mais dúvidas quanto à possibilidade de se adotar a distribuição exponencial como àquela que determina o comportamento da variável TEC.

5.7.3 Considerações sobre o Emprego de Distribuições Teóricas de Probabilidades

Uma vez terminado o processo de ajuste da distribuição de probabilidades aos dados disponíveis na amostra, é conveniente que se deixe claro alguns pontos importantes sobre esta decisão. O emprego de distribuição de probabilidades, por mais ajustada que esta seja, pode implicar em alguns riscos quando aplicada ao modelo de simulação.

A primeira consideração a ser realizada trata dos limites extremos (mínimos e máximos) que a distribuição pode assumir. Em algumas situações, o emprego de distribuições não limitadas (como a Normal), pode implicar em circunstâncias incompatíveis com a realidade. A distribuição Normal pode assumir tanto valores positivos como negativos e, dependendo de seus parâmetros (média próxima de zero) e do propósito de seu emprego (tempo de processo é um emprego comum), pode acontecer que, durante a simulação, sejam gerados “tempos negativos”. Neste caso, o emprego de uma distribuição Triangular (limitada), talvez possa resolver o problema.

Além dos problemas causados por limites teóricos impossíveis de ocorrerem na prática, a adoção de distribuições assimétricas pode fazer com que durante a simulação valores muito altos ou muito baixos (mesmo que positivos) possam causar situações incompatíveis com a realidade. Neste caso, uma possível consideração seria adotar uma distribuição empírica, baseada nos dados amostrais. O *Input Analyzer* também pode realizar este trabalho (menu *Fit/Empirical)*. A Figura 5.22 exibe o resultado desta opção. O gráfico apresenta tanto o histograma das freqüências originais como o da freqüência acumulada. Os dados, a que o gráfico se refere, encontram-se na tabela abaixo deste.

A contrapartida, no caso de se adotar uma distribuição empírica, é que durante a simulação o modelo trabalhará, sempre, dentre dos limites estabelecidos pelos dados colhidos. Se esta, de fato, replicando situações anteriormente ocorridas e capturadas pela amostra. Nada de errado neste procedimento. Mas caberá ao analista a decisão sobre empregar uma distribuição teórica de probabilidades, que pode gerar valores pouco compatíveis com a realidade (é verdade que com baixa probabilidade) ou, aplicar uma distribuição empírica que não apresentará riscos de ocorrência de valores extremos, mas também não permitirá verificar o comportamento do sistema em situações outras, que não o já ocorrido no passado. Executar o modelo considerando as duas alternativas pode ser uma boa estratégia para auxiliar a decisão.

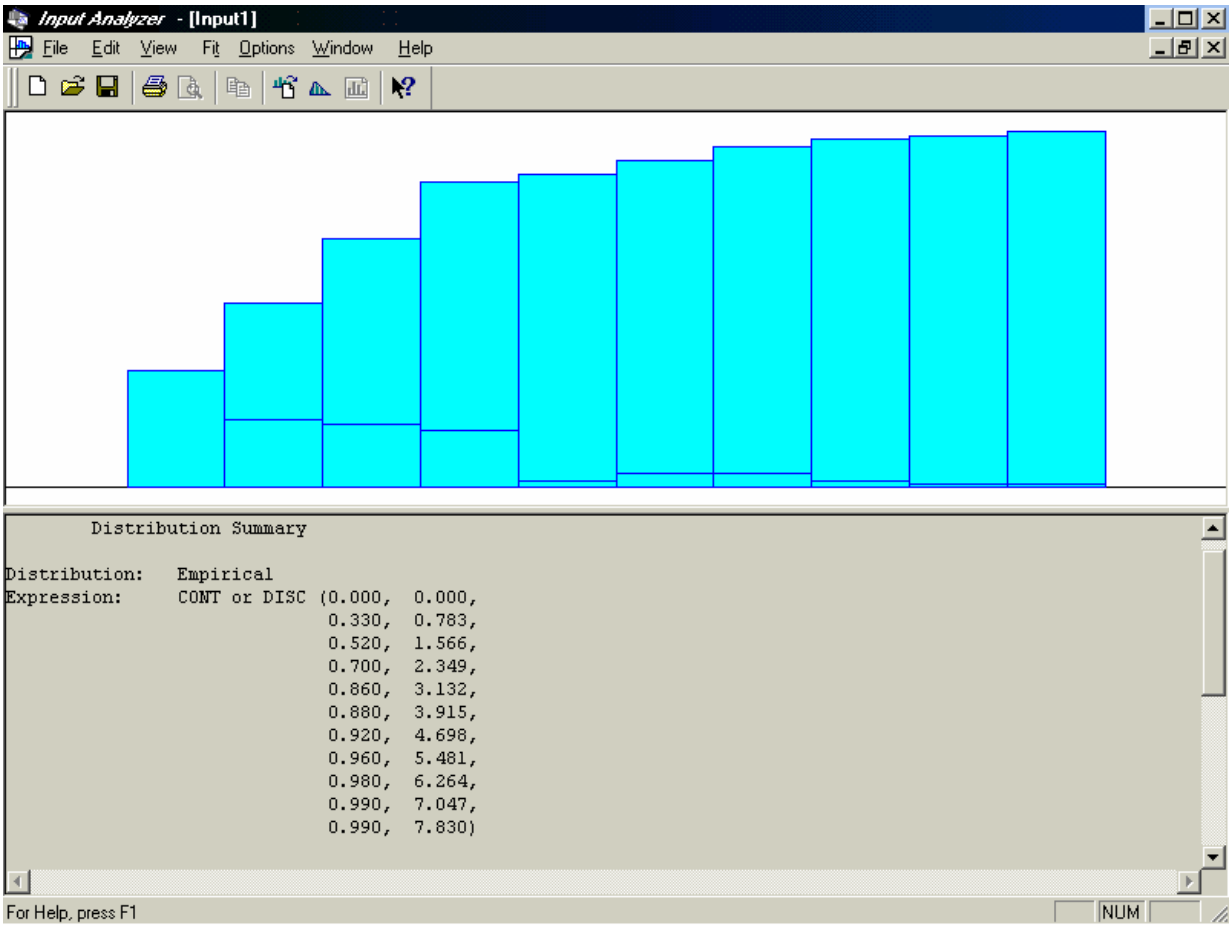

Figura 5.22: Resultado do processo de aderência a uma distribuição empírica.

Antes de se encerrar este capítulo, convém ainda alertar o usuário para outros tipos de situações que podem confundir os menos experientes. Estas incluem a presença de amostras multimodais (quando dois ou mais valores são mais freqüentes que os demais) ou conjuntos de dados que apresentam valores extremos (mas corretos).

Em situações como estas, é conveniente ao usuário subdividir a amostra antes de tentar realizar o processo de ajuste. Imagine por exemplo, a situação em que se tem dados que pretendem retratar o perfil de consumidores de um supermercado. Este conjunto, que pode ser visualizada na Figura 5.23, contém aproximadamente 1500 valores que informam o número de itens (mercadorias) com que clientes chegam ao caixa.

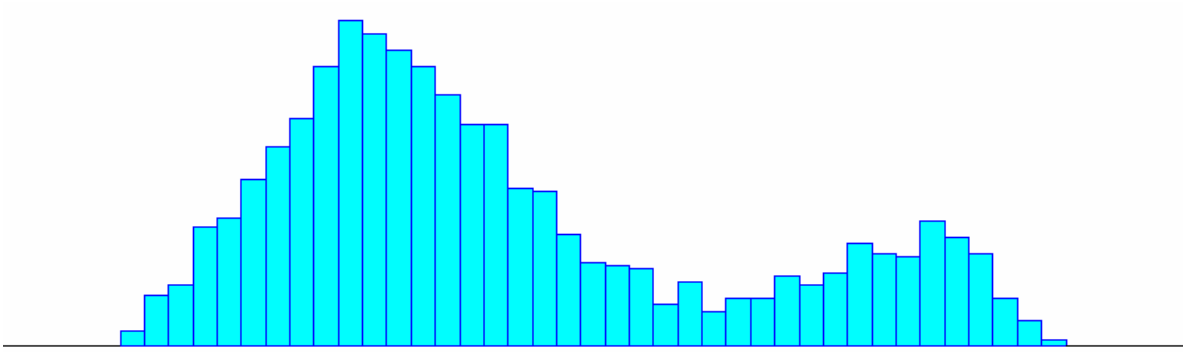

Figura 5.23: Amostra perfil de consumidores

O menor valor para o número de mercadorias foi nove itens. O maior foi 134 itens. Observa-se claramente dois picos principais nos dados do gráfico da figura 5.23. Uma análise um pouco mais criteriosa revela os principais tipos de clientes: os que compram na faixa de 20 a 30 itens e os que compram algo em torno dos 85 itens. Um tratamento adequado a este conjunto de dados requer uma divisão prévia do arquivo em dois conjuntos. O primeiro para os dados agrupados em torno do pico à esquerda da figura e o segundo para o grupo de dados em torno do pico à direita.

O *Input Analyzer* permite a realização deste tratamento sem um intervenção externa no arquivo de dados original. Pode-se estabelecer no menu *Options/Parameters/Histogram* sobre quais dados do histograma se deseja determinar o ajuste. Desta forma, tome-se, por exemplo, o limite da esquerda como sendo o ponto inferior e algum ponto entre os dois picos como sendo o limite superior. Os dados que serão utilizados podem ser facilmente deduzidos das informações que constam no texto logo abaixo do histograma (valores mínimo e máximo, média, número de classes, etc..). Examine o arquivo *mercadorias.txt*. Com ele é possível reproduzir o arquivo que aparece na Figura 5.23.

Sumário

Dentre todas as fases de um projeto que envolve a modelagem e a simulação de sistemas, a etapa inicial de definição do sistema a ser modelado em conjunto com a etapa da formulação do modelo, tanto no nível conceitual como computacional (programação), são consideradas as que apresentam as maiores dificuldades. Já coletar as informações sobre o funcionamento do sistema e saber adequadamente utilizá-las são consideradas tarefas menos nobres. No entanto, além de consumirem uma boa parte do tempo, o alcance do pleno sucesso só virá se estas tarefas forem cumpridas com o mesmo empenho e determinação empregada nas demais fases. Segundo Pegden [PEGDEN, 90], toda a base para que se alcance os objetivos em um projeto envolvendo simulação de sistemas depende, fundamentalmente, de quanto esforço for empregado em:

- Definir o problema;
- Estabelecer os objetivos do estudo ou projeto;
- Definir os limites (fronteiras) do sistema;
- Determinar os componentes e variáveis mais relevantes;
- Abstrair e levantar hipóteses sobre as relações entre componentes e variáveis;

- Estimar os valores dos parâmetros pertinentes.

Como em qualquer projeto, quanto mais erros se cometem nas fases iniciais, maiores serão as conseqüências e os custos de repará-los em estágios mais avançados. Um número que geralmente aparece na apuração dos tempos despendidos nas várias etapas aponta para um valor entre 30% e 40% do tempo total de um projeto sendo consumido nestas fases preliminares.

Exercícios

1. Desenvolva um modelo de um sistema serial simples com apenas dois processos. As entidades chegam a este sistema, com uma média de 10 min. entre cada chegada. Na medida em que chegam, as entidades são enviadas ao Processo 1. Neste posto a fila é ilimitada, e o único recurso é usado com o tempo de serviço durando, em média, 9 min.. Uma vez completado este serviço, a entidade é transferida ao Processo 2, o qual e idêntico ao Processo 1. As entidades deixam o sistema após os dois processos. As medidas de desempenho adotadas são o tamanho médio das filas de cada um dos processos e o tempo médio de permanência no sistema. Simulando o sistema por 10.000 min., realize os seguintes experimentos comparando os resultados:

Experimento	Intervalos entre Chegadas	Tempos de Serviços
1	Exponencial	Exponencial
2	Constante	Exponencial
3	Exponencial	Constante
4	Constante	Constante

2. Usando o mesmo modelo do exercício anterior, defina a distribuição Exponencial para os tempos decorridos entre as chegadas e uma distribuição Normal com média 9 para os tempos dos serviços nos dois processos. Realize experimentos atribuindo valores iguais a 1, 2 e 3 ao desvio-padrão. Cada simulação deverá ser executada durante 10.000 min.. Compare os resultados de cada rodada de simulação.

3. Ainda usando o mesmo modelo do exercício anterior, assuma que os tempos de processo possuem uma *média* de 9 min. e uma *variância* de 4 min.. Calcule os parâmetros para as distribuições *Gamma*, *Uniforme* e *Normal* que determinem tais valores. Realize simulações do sistema com estas distribuições e compare os resultados.

4. Usando o *Input Analyzer* crie um novo documento (use a opção *File/DataFile/Generate New*) contendo 50 pontos para a distribuição *Erlang* com os seguintes parâmetros: *ExpMean* igual a 12, *k* igual a 3 e *Offset* igual a 5. Uma vez obtido o conjunto de dados, realize um *Fit All*. Repita o procedimento para 500, 5.000 e 25.000 dados, usando os mesmos parâmetros para uma distribuição *Erlang*. Compare os resultados do melhor ajuste para os quatro diferentes conjuntos de dados.

5. Usando o *Input Analyzer* faça o processo de ajuste ao arquivo: "*mercadorias.txt*".