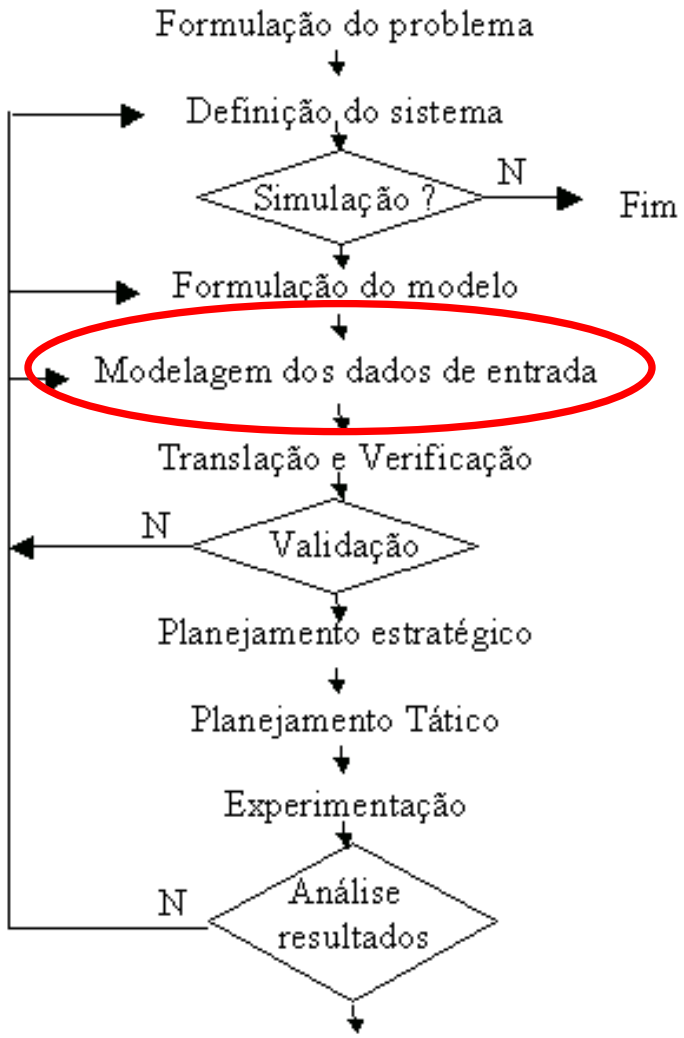
Análise e Tratamento de Dados para a Simulação

AULA 08

FACULDADE NOVA ROMA

CERTIFICAÇÃO DE QUALIDADE FGV

Introdução


Formulação do problema
Definição do sistema
Simulação ?
N
Fim
Formulação do modelo
Modelagem dos dados de entrada
Translação e Verificação
N
Validação
Planejamento estratégico
Planejamento Tático
Experimentação
N
Análise resultados

Situações relevantes na coleta de dados para um dado modelo de simulação

- Dados Existentes
  - Necessário localizá-los
  - Conseguir ter acesso a eles
- Dados Inexistentes mas possíveis de obter
  - Inquéritos
  - Estudos de amostragem
  - Experimentação
- Dados impossíveis de obter
  - Opiniões de peritos

Coletando Dados

- Geralmente difícil, caro e chato
  - Sistema pode não existir;
  - Os dados disponíveis podem não ser os desejados;
  - Podem haver mudanças no modelo em função do que se dispõe;
  - Dados podem ser incompletos;
  - Existência de muitos dados.

Coletando Dados...

- Sensibilidade dos resultados às incertezas nos dados;
- Modele o nível de detalhes de acordo com a qualidade dos dados;
- Capture a variabilidade nos dados
- Custos devem ser orçados no projeto.

Alternativas de Dados

Usando dados *"diretamente"*

- Dados são lidos de arquivos e usados no modelo (chegadas, serviços, tipos de entidades, tempos, temperaturas, etc.);
- Todos os valores são "**reais**";
- Não há elementos diferentes dos já observados;
- Pode haver falta de dados para muitas ou longas simulações;
- Perda de desempenho computacional (leitura de arquivos).
- Falta de flexibilidade operacional

Alternativas de Dados

Uso de **distribuições de probabilidades**:

- Dados gerados de acordo com a distribuição adotada;
- Grande biblioteca de distribuições
- Flexibilidade operacional
- **Outros valores,** além dos observados, poderão ser empregados (bom ou ruim ?);
- A probabilidade de ocorrência de qualquer valor no intervalo é determinada pelo perfil da distribuição
- O processo de aderência pode não ser perfeito ou adequado

Outras Alternativas

- Construir distribuição específica – sob medida: difícil,caro e demorado.

Modelagem de Dados de Entrada

Introdução

- Quando se usa distribuições de probabilidade para representar o comportamento de variáveis aleatórias, é preciso considerar que:
  - Os valores possíveis que a variável poderá assumir estarão dentro da amplitude coberta pela distribuição
  - A probabilidade de ocorrência de qualquer valor no intervalo é determinada pelo perfil da distribuição

Introdução...

- Portanto, é possível antecipar quais valores a variável poderá assumir, sem no entanto ser possível determinar quais, precisamente serão estes valores.
- A garantia do perfeito casamento entre uma distribuição teórica de probabilidades e o comportamento aleatório de uma variável de sistema passa por várias etapas.

Etapas para garantir casamento entre: distribuição de probabilidade e comportamento da variável aleatória

1. Processo de Amostragem e Coleta de Dados representativos
2. Tratamento dos Dados
3. Identificação de distribuições de probabilidade
4. Estimar parâmetros
5. Teste de Aderência (Goodness-of-Fit Tests)

1. Amostragem e Coleta de Dados

- Fontes de Dados
  - Arquivos históricos, observações do sistema, oriundos de sistemas similares, determinados com base em estimativas , baseados em afirmações, considerações teóricas, etc.

1. Amostragem e Coleta de Dados

- Amostragem

  Planejamento e observação preliminar

  - imaginar formas de coleta de dados, etc.
  - Verificar utilidade de dados coletados
  - verificar se dados são adequados para o fornecimento das distribuições que serão tomadas como entrada na simulação
  - identificar dados supérfluos

1. Amostragem e Coleta de Dados

- Amostragem: Conjunto homogêneo de dados
- Combinar dados que obedeçam ao mesmo tipo de distribuição, ao longo de um determinado período ou intervalo de tempo, em conjuntos homogêneos
- Isto permite simplificar o modelo de simulação e reduzir os custos da recolha
- Ex: verificar se os dados referentes à entrega da matéria prima em um determinado setor são homogêneos, coletar dados no período da manhã e outros no período da tarde. Verificar o comportamento nos diversos dias da semana

Amostra –Homogeneidade dos Dados

- O exemplo considera os clientes que se dirigem aos caixas.
- Períodos críticos (mais congestionados).
- Dias considerados normais (terças, quartas e quintas-feiras), com 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 1.1.

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

- 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 1.2.

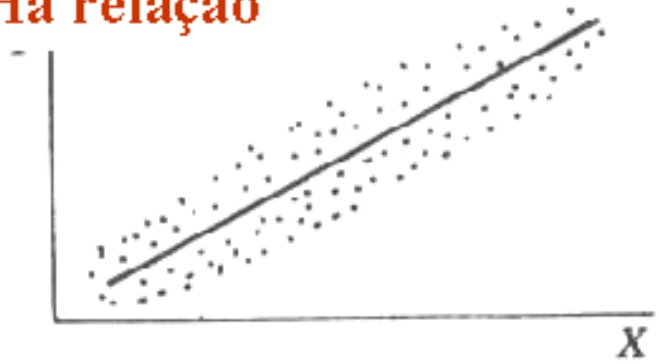
Período	Tipo de Demanda
10:00 às 11:00	A* 1,3
11:00 às 13:30	B
13:30 às 14:30	A
14:30 às 15:30	B
15:30 às 16:00	A* 1,2

- Além disso, qualquer dia de meio de semana que seja o último do mês tem demanda semelhante a da tabela 1.2. 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 1.2, acrescida de 20%.

1. Amostragem e Coleta de Dados

- Amostragem ...
- Determinar relacionamento entre variáveis (quer elas ocorram entre variáveis distintas,quer para a mesma variável) traçando o Diagrama de dispersão.
- Diagrama de Dispersão
  - uma representação gráfica dos valores correspondentes de duas variáveis, como pontos no plano, usando, como abscissas, os valores duma e, como coordenadas, os da outra.

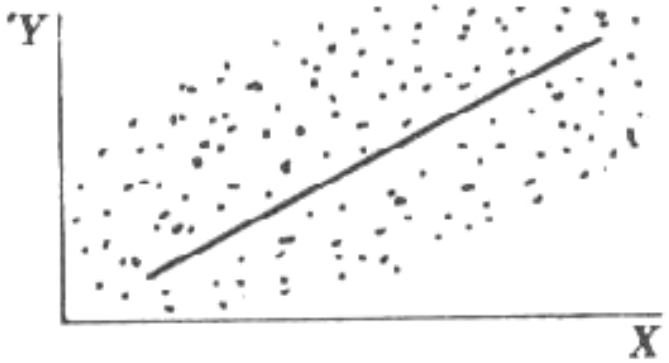
Há relação



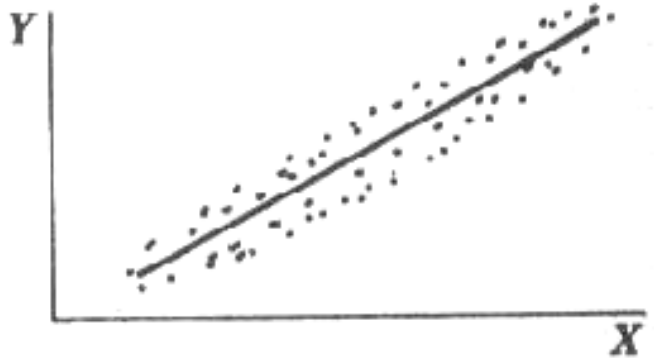
(a) Relação linear direta a

Não há relação

(b) Não há relação



(c) Relação linear direta com menor grau de relação que em (*a*)



(d) Relação linear direta com maior grau de relação que em (*a*)

Menor grau de relação do que em (a)

Maior grau de relação do que em (a)

Precauções na Recolha de Dados

- A presença de correlação significativas obriga, normalmente, à utilização de modelos mais sofisticados de análise.
- Dificilmente o modelo produz resultados realísticos se existir relação de dependência e esta for ignorada.

Precauções na Recolha de Dados ...

Independência das Observações

- considerar que uma sequência de observações, que aparentemente parecem independentes, possuam algum relacionamento (autocorrelação).

- A autocorrelação pode existir entre períodos sucessivos de tempo ou para clientes sucessivos. Ex: o tempo de serviço para o cliente i, pode estar relacionado com o tempo para servir o clente i+1

Precauções na Recolha de Dados ...

- Detecção de sazonalidade (ou outro tipo de não-estacionaridade) nos valores observados
  - A repetição periódica ao longo do tempo de padrões de variabilidade semelhantes, causam mais dificuldade na análise, em especial se não for possível removê-las.

2. Tratamento dos Dados

- Colocar dados numa forma compacta e compreensível de acordo com a pessoa que vai usá-los, de forma a que se possa extrair a informações desejadas.
- Ex: grafos/histogramas, estatísticas

Tratamento de Dados

- **Representação Gráfica -->** *Histogramas*
- **Dados brutos -** *limites* (6, 114)

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

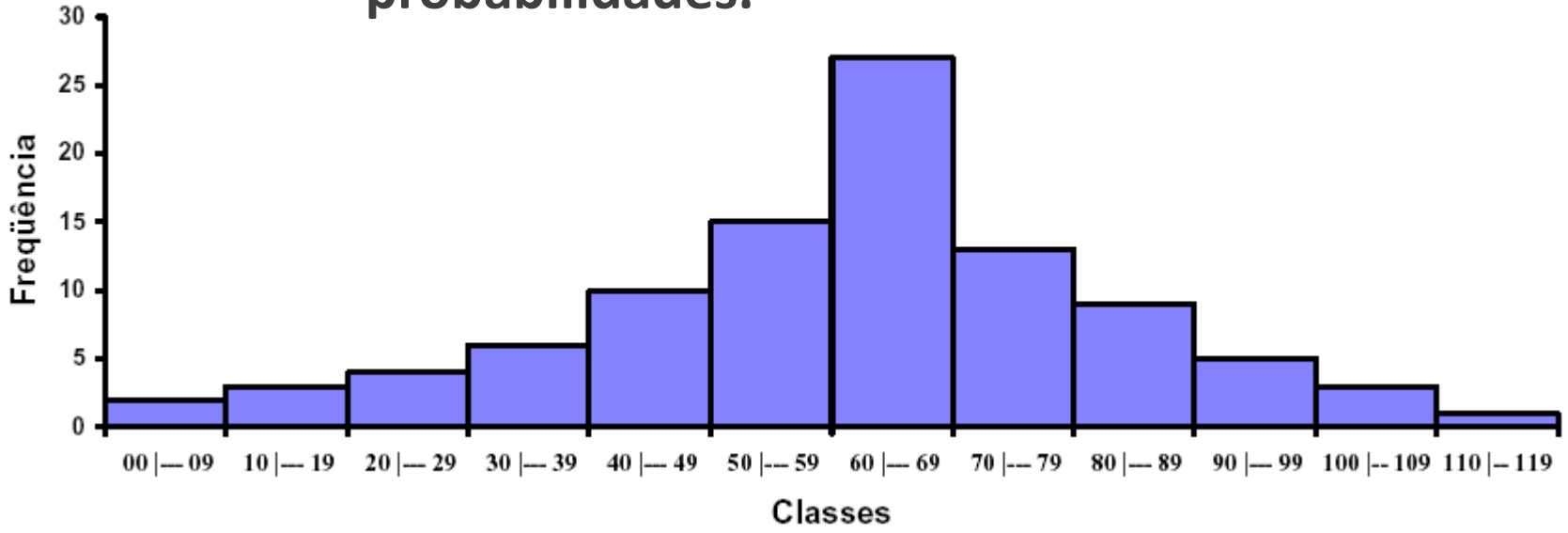
Representação Gráfica

- Tabela de distribuição de frequências

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

Representação Gráfica

- **Histograma**
  - **A utilização de gráficos (ex:histograma) são muito úteis para o delineamento da distribuição teórica de probabilidades.**



3. Identificação de distribuições de probabilidade

- Atribui-se a uma determinada distribuição teórica a responsabilidade pela geração de dados de um processo - comportamento estocástico da variáveis
- Durante o processo de ajuste das curvas, pode-se ter dificuldades em demonstrar a aderência entre os dados empíricos e aqueles da curva teórica

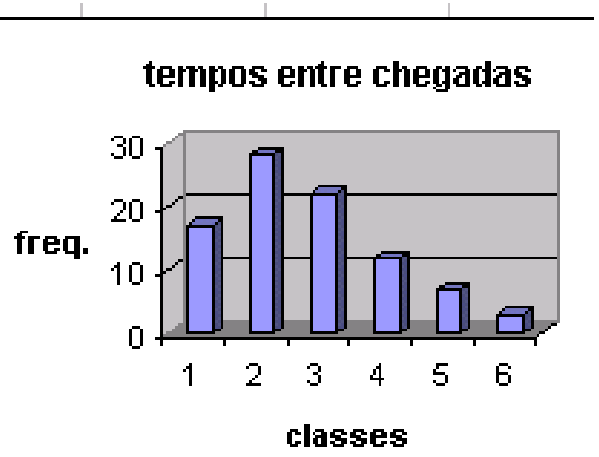
Geralmente ocorre um diagnóstico preliminar

3. Identificação de distribuições de probabilidade

- Assim, defini-se as distribuições de frequências seleciona-se uma distribuição de probabilidade

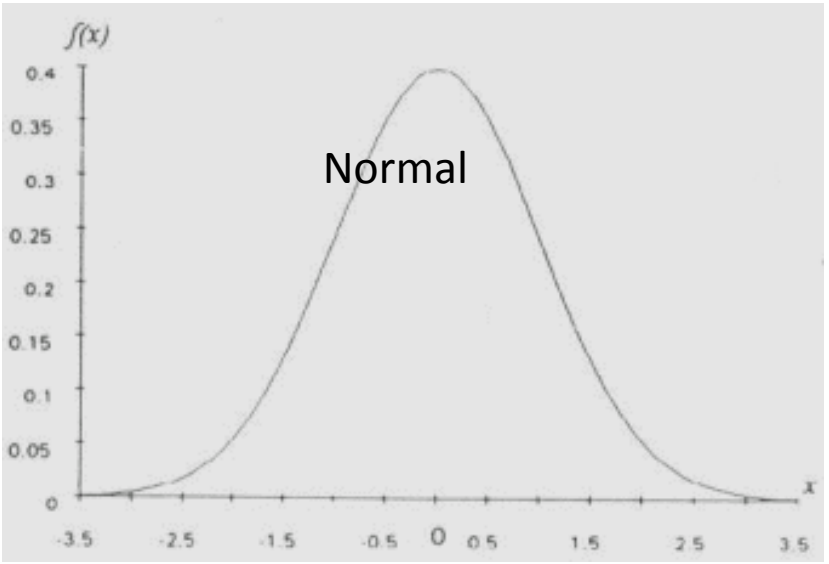
Ex.: tempos entre chegadas de clientes em um caixa de banco

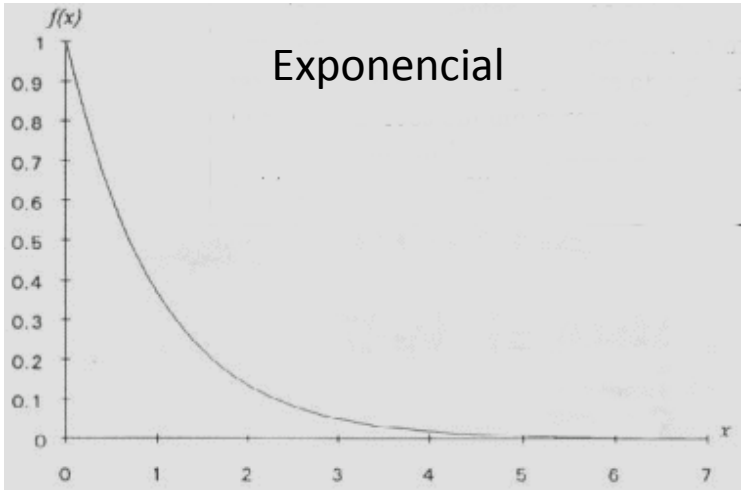
classes	freqüência
1	17
2	28
3	22
4	12
5	7
6	3
n =	89

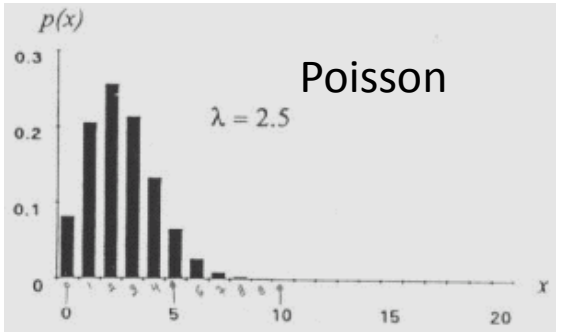


3. Identificação de distribuições probabilidade...

- Critérios para seleção da distribuição:
  - Natureza do processo sendo modelado
  - Comparação visual das curvas para achar perfil semelhante
- Exemplos

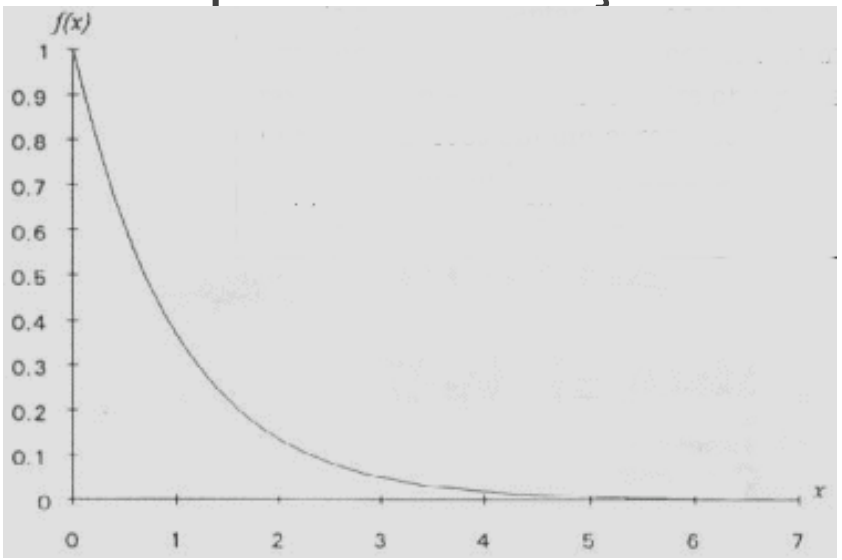






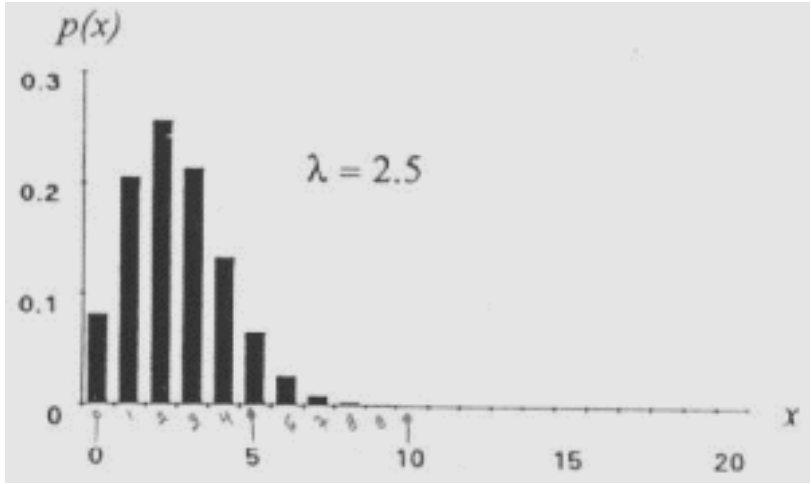
3. Identificação de distribuições probabilidade...

- **Exponencial**
  - Eventos independentes (*falta de memória*)
    - todo o fenômeno aleatório descrito por esta distribuição se caracteriza pela total imprevisibilidade, mesmo que se conheça o seu passado
  - Supõe grande variância
  - Ex:Tempos decorridos entre eventos (entre chegadas de entidades)
  - Muito usada em sistemas de filas



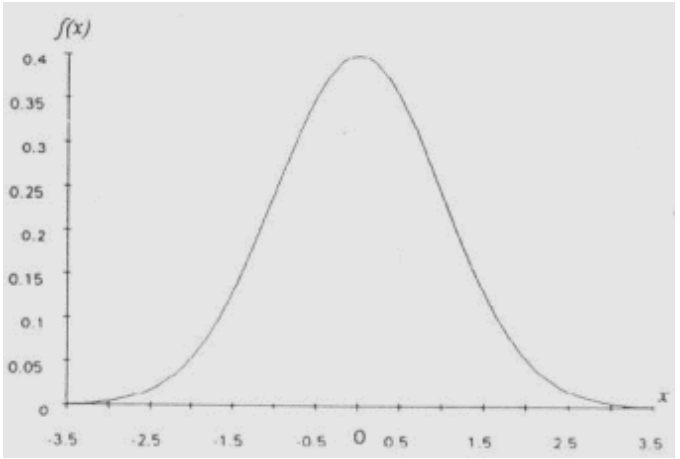
3. Identificação de distribuições de probabilidade

- **Poisson**
  - Distribuição discreta usada para modelar o número de ocorrências (valores discretos) que uma variável pode assumir ao longo de um intervalo contínuo
  - Modela o número de eventos independentes que ocorrem em um intervalo de tempo
  - Ex:
    - número de componentes que falham num intervalo de tempo



3. Identificação de distribuições de probabilidade...

- **Normal**
  - Duração de tarefas onde a variabilidade é baixa; curva simétrica. É muito usada.
  - Descreve fenômenos simétricos em torno da média
  - Usada sempre que a aleatoriedade por causa das várias fontes independentes agindo de forma aditiva



3. Identificação de distribuições de probabilidade...

- **Uniforme:**
  - Especifica que os valores compreendidos entre o mínimo e o máximo são equiprováveis.
  - O seu uso geralmente significa um completo desconhecimento da variável aleatória, conhecendo-se apenas seus limites.

4. Estimar parâmetros da distribuição escolhida

- Primeiro determina-se medidas:
  - descritivas dos dados (ex: média, moda e mediana) e
  - de dispersão (variância dos valores amostrados, desvio padrão amostral).
- Ex: se foi selecionada uma distribuição Normal, calcula-se:
  - Média
  - Desvio-padrão

Média

Média $\mu = E(x)$

$$= \sum_{i=1}^{n} p_i x_i$$ **Para variáveis discretas**

$$= \int_{-\infty}^{+\infty} x f(x) dx$$ **Para variáveis contínuas**

Soma de todos os valores possíveis, ponderada pela probabilidade de ocorrência de cada um dos valores.

Variância

A quantidade $(x-\mu)^2$ representa a distância quadrática entre x e a sua média.

A variância de x é o valor esperado desta quantidade:

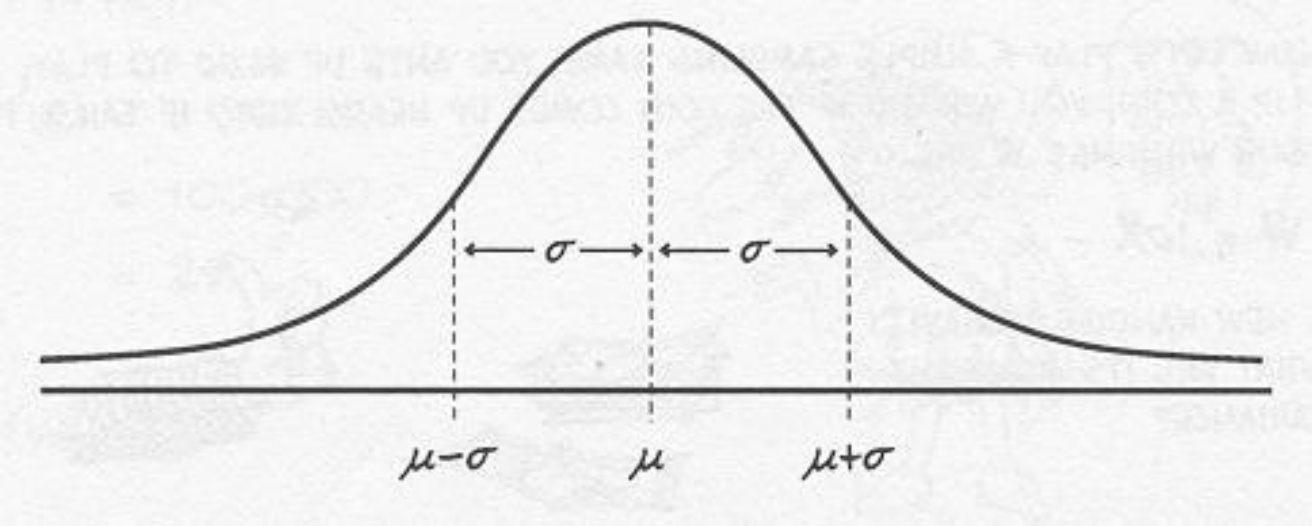
$$\mathrm{Var}(x) = E[(x-\mu)^2] = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

$$= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx$$

Variância

A variância é normalmente denotada por σ^2.

A raiz quadrada da variância é chamada de desvio padrão e é denotado por σ.



4. Estimar parâmetros da distribuição escolhida

- **Dados não Agrupados**

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Média

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

Variância

- **Dados Agrupados**

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

Média

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

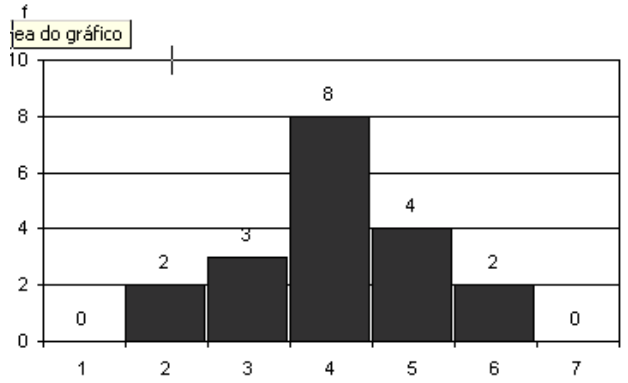
Variância

Deve-se considerar se os dados são agrupados ou não em uma distribuição de frequência.

4. Estimar parâmetros da distribuição escolhida

$$\overline{X} = \left(\sum_{i=1}^{k} f_i X_i \right) / n$$

$$S = \left(\left(\sum f_i X_i^2 - n\overline{X}^2 \right) / (n-1) \right)^{1/2}$$



# classe	inf	sup	f	X	f.X	X-x	f.x2
1	-100	2	0	-49	0	-56,1	0
2	2	4	2	3	6	-4,11	33,706
3	4	6	3	5	15	-2,11	13,296
4	6	8	8	7	56	-0,11	0,0886
5	8	10	4	9	36	1,89	14,36
6	10	12	2	11	22	3,89	30,338
7	12	100	0	56	0	48,9	0
			Σ ↓		Σ ↓		Σ ↓
			19		135		91,789
média=	7,105		n				
desvio=	2,258						

Após estas etapas

- Após se:
  1. Levantar uma hipótese sobre qual ou quais distribuições teóricas são candidatas a apresentar os dados coletados
  2. Calcular a média e desvio padrão da amostra
  3. Realizar estimativas sobre os parâmetros de distribuição
  4. **Aplicam-se testes de aderência aos dados**

5. Teste de Aderência

Geralmente os testes empregam métodos:

- Gráficos: onde a qualidade é medida de forma visual (proximidade entre o desenho da distribuição teórica e aquele referente aos dados coletados)

Verificação da qualidade da escolha da distribuição que melhor representa os dados.

Teste de Aderência

- Teóricos: verificam se o conjunto de dados amostrados não difere significativamente daquele esperado de uma distribuição teórica específica;
  - Exemplo:
    - Chi-quadrado e
    - Kolmogorov-Smirnov (K-S)

5. Teste de Aderência ...

- **Testes Paramétricos**
  - Baseiam-se no Teorema do Limite Central (TLC)
    - soma ou média de amostras de um grande número de observações aleatórias e independentes é aproximadamente normal, independente da distribuição dos valores.

Teste de Aderência

- *Ex:* ***Qui-Quadrado*** (χ^2), usado quando:
  - amostras são grandes e
  - natureza da distribuição é contínua ou discreta
  - Exige pelo menos 100 amostras
- O procedimento tem início com o arranjo das *n* observações em conjuntos de *k* classes de intervalos.

5. Teste de Aderência ...

- Segue-se o cálculo da estatística

pela seguinte fórmula:

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

$$x^2 = \frac{\sum_k (f_0 - f_e)^2}{f_e}$$

Se $x^2 = 0$, então as duas distribuições estão casando perfeitamente, quanto maior x^2 maior discrepância entre as duas distribuições

5. Teste de Aderência

- **Testes não-paramétricos**
  - não supõe nada em relação à forma da distribuição sendo testada;
  - podem ser usados quando o tamanho da amostra é pequeno ($n \leq 30$), o que exclui a aplicação do TLC

5. Teste de Aderência ...

- *Ex:* ***Kolmogorov-Smirnov*** **(K-S)**
  - Baseia-se na comparação de probabilidades acumuladas das duas distribuições (observada e teórica).
  - Permite avaliar a hipótese de que uma amostra foi retirada de uma determinada distribuição contínua especificada.
  - Não é usado em distribuições discretas.

5. Teste de Aderência...

- Na tabela a seguir tem-se dados brutos divididos em 10 classes associadas às suas frequências
- O valor da estatística K-S será obtido a partir das diferenças entre os valores acumulados das colunas:
  - Frequência Acumulada Observada
  - Frequência Acumulada Teórica
- As maiores diferenças serão observadas nas classes que iniciam em 14,00 e vão até 20,00

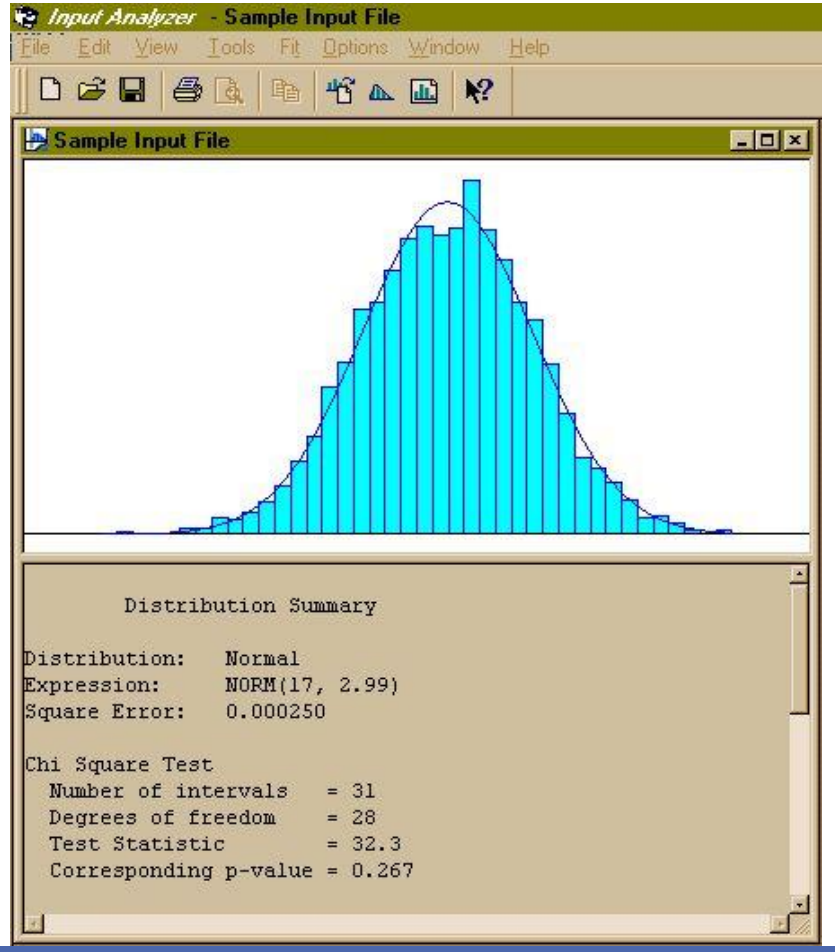
Tabela de Distribuição de Frequências

Limites das classes	Frequência relativa observada	Frequência Acumulada Observada	Frequência Acumulada Teórica	Diferenças Frequência Acumulada
10,00-12,00	0.2167	0.2167	0.1	**0.1167**
12,00-14,00	0.1167	0.3334	0.2	**0.1334**
14,00-16,00	0.1167	0.4501	0.3	**0.1501***
16,00-18,00	0.1000	0.5501	0.4	**0.1501***
18,00-20 ,00	0.1000	0.6501	0.5	**0.1501***
20,00-22,00	0.0333	0.6834	0.6	**0.0834**
22,00-24,00	0.1167	0.8001	0.7	**0.1001**
24,00-26,00	0.0333	0.8334	0.8	**0.0334**
26,00-28,00	0.0666	0.9000	0.9	**0.0000**
28,00-30,00	0.1000	1.000	1.0	**0.0000**

Modelagem de Dados de Entrada

Permite identificar distribuição teórica de probabilidades por meio de testes de aderência

- Usuário deve possuir uma amostra de dados coletada do sistema real.
- O aplicativo fornece uma expressão válida para ser empregue na simulação.



Análise de Resultados

PRÓXIMA AULA