Universidade de São Paulo
BRASIL

Análise estatística multivariada aplicada a processos químicos

- **Escola Politécnica**
- **Departamento de Engenharia Química**

- Prof. Dr. Cláudio Augusto Oller do Nascimento
- Prof. Dr. Roberto Guardani

- 2007

Parte 1. Conceitos Básicos de Estatística

Introdução

Os sistemas de instrumentação e de controle de processos industriais viveram um grande avanço nos últimos vinte anos, como conseqüência da evolução na microeletrônica e na área computacional. Houve um aumento impressionante da quantidade de informações disponíveis aos operadores de processos químicos, com a disseminação de instrumentos medidores das variáveis de processo, assim como dos sistemas de transmissão, concentração e armazenamento de dados. Atualmente, praticamente todas as instalações industriais de processos produtivos são providas desses sistemas. Em instalações petroquímicas, por exemplo, é comum a existência de milhares de sensores de vazão, temperatura, nível e pressão, além de analisadores de composição em correntes e equipamentos, capazes de medir em tempo real uma grande quantidade de variáveis de interesse. Atualmente os engenheiros(as) de processo contam com informações valiosas, com enorme quantidade de dados de processo, o que possibilita monitorar em detalhes variáveis específicas, ou o desempenho de equipamentos e de determinadas áreas de unidades industriais.

A aplicação das técnicas de estatística multivariada, nesses casos, pode fornecer informações quanto a correlações entre grupos de variáveis, em geral não evidentes em análises baseadas em pares de variáveis, apenas. Uma vez identificados padrões de correlação, essas técnicas podem ser aplicadas para identificação de diferentes regimes de operação da unidade, assim como na detecção de situações anômalas, erros de medidas em sensores, ou falhas em unidades. Dados históricos de variáveis de processo têm sido usados também na otimização de condições de operação de unidades industriais.

Este curso apresenta técnicas de análise estatística multivariada adequadas à aplicação em processos da indústria química, visando extrair informações sobre o comportamento de unidades industriais a partir da exploração de dados de operação, e que constituem a base para a construção de modelos estatísticos e para a implementação de controles estatísticos em unidades industriais. Os conceitos apresentados, assim como o treinamento com exercícios baseados em casos industriais reais, constituem ferramentas de grande valor para a análise de processos industriais. O objetivo é oferecer um conjunto de métodos estatísticos adequados à aplicação a casos de interesse, constituíndo-se, assim, em complemento da formação de profissionais de engenharia de processos.

Lista de símbolos

$E(x)$	esperança, ou valor esperado, da variável x
e	erro aleatório; coordenadas em uma base ortogonal após rotação (Fig. 3.1)
e_j	j-ésimo componente principal
freq	freqüência de amostragem
GL	número de graus de liberdade
M_t	momento de ordem t de uma distribuição
m	número de componentes principais selecionados
N	número total de observações em uma população
n	número total de observações em uma amostra
$P(x)$	probabilidade acumulada, Eq. 1.11
p	número de variáveis em um sistema multivariado
$p(x)$	função densidade de probabilidade, Eq. 1.10
Q	variável qualquer em um processo
R	matriz (pxp) de correlação das p variáveis aleatórias
r_{kj}	coeficiente de correlação entre as variáveis k e j
SD^2_{ih}	distância estatística entre as observações i e h
SCP_{jK}	soma dos produtos cruzados das variáveis centradas na média X_j e X_k
SS_j	soma dos quadrados da variável aleatória X_j
SSCP	matriz (pxp) da soma dos quadrados e produtos cruzados das p variáveis
s^2	desvio padrão calculado para uma amostra
w_{kj}	peso da variável aleatória x_j no k-ésimo componente principal
X_j	variável aleatória j centrada na média
X	vetor vertical de variáveis aleatórias centradas na média, com p linhas, correspondente a uma dada observação
x	variável aleatória j
x	matriz de dados experimentas, com n linhas e p colunas
$x(i,k)$	i-ésima observação da k-ésima variável
x_{med}	média de uma variável calculada para uma amostra

Subscritos:

i	refere-se à i-ésima observação
k	refere-se à k-ésima variável
j	refere-se à j-ésima variável, ou ao j-ésimo componente principal
min, med, max	valores mínimo, mediano e máximo de uma variável

Símbolos gregos

ε	erro na medição de uma variável *Q*
μ	média de uma população
$\boldsymbol{\Sigma}$	matriz (*pxp*) de covariância das *p* variáveis aleatórias
σ	desvio padrão de uma população
τ	erro sistemático em uma medida
θ	ângulo de rotação dos eixos de coordenadas (Fig. 3.1)

Bibliografia

BARROS NETO, B., SCARMINIO, I.S., BRUNS, R.E. Como fazer experimentos: pesquisa e desenvolvimento na indústria. Ed. Unicamp, Campinas, 2001.

BERTHOUEX, P.M., BROWN, L.C. Statistics for environmental engineers. Lewis Publishers, 2nd. Ed., New York, 2002.

HAIR Jr, J.F., ANDERSON, R.E., TATHAM, R.L., BLACK, W.C. Multivariate data analysis. Prentice Hall, 5th. Ed, Upper Saddle River, 1998.

HIMMELBLAU, D. M. Process analysis by statistical methods. John Wiley & Sons, 1970

JOHNSON, R.A., WICHERN, D.W. Applied multivariate statistical analysis. Prentice Hall, 4th. Ed, Upper Saddle River, 1998.

JOLLIFFE, I.T. Principal component analysis. Springer-Verlag, New York, 1986.

SHARMA, S. Applied multivariate techniques. John Wiley & Sons, Inc, New York, 1996.

Estatística descritiva

Neste texto, o resultado de uma medição qualquer em um processo em um dado instante de tempo será denominado de *observação*. Uma dada observação pode ser constituída por um único dado experimental, ou, no caso de sistemas com múltiplas variáveis, por um conjunto de valores. Uma observação será denominada de *x(i,k)*, sendo *i* o índice da observação na série de observações (com *i* variando de 1 a *N* observações) e *k* o índice da variável (com *k* variando de 1 a *P* variáveis). Por exemplo, supondo que numa unidade industrial sejam consideradas medidas de vazão, temperatura e pH de uma corrente de efluente ao longo do tempo. Supondo que se deseje estudar uma série de 100 medidas ao longo de um determinado período de tempo, então o número de observações *N* é igual a 100 e *P* é igual a 3. Um dado qualquer, como, por exemplo, *x(20,3)* refere-se à vigésima observação da variável pH, da série de 100 observações. A base de dados referente a esse exemplo seria então constituída por uma matriz de dados x, com 100 linhas e 3 colunas.

Erros de medidas, precisão e exatidão

As medidas de variáveis de processo na indústria apresentam normalmente flutuações, na forma ilustrada na Figura 1.1, na qual é mostrado o gráfico de uma série temporal dos valores medidos, *x,* da vazão, *Q*, em uma corrente de processo. Observam-se tendências de longo período (ou baixa freqüência) juntamente com flutuações de amplitudes e freqüências variáveis.

Toda medição tem imperfeições que dão origem a um erro no resultado. O erro pode ser representado como:

$$x = Q + \varepsilon \,, \tag{1.1}$$

em que o erro ε pode ter um componente sistemático, τ, também chamado “bias”, e um componente aleatório, *e* Assim, tem-se:

$$x = Q + (e + \tau) \,. \tag{1.2}$$

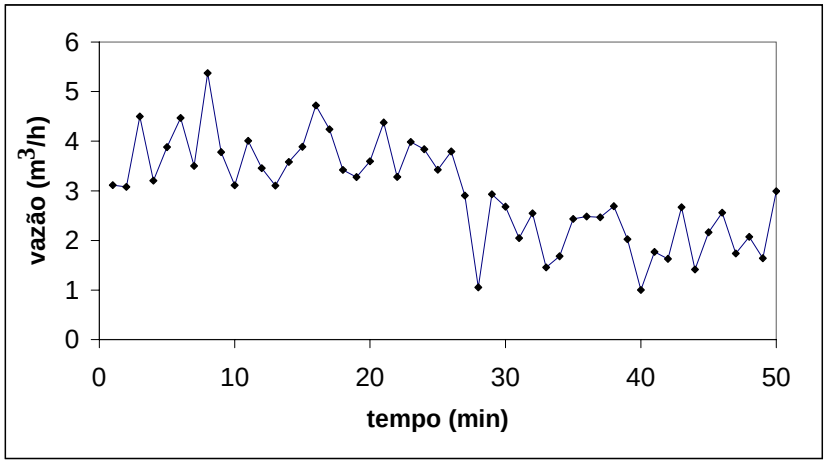


Figura 1.1. Exemplo de série temporal de variável medida em unidade industrial.

Erros sistemáticos ("bias") causam desvios persistentes entre o valor medido e o valor devido de uma variável, e normalmente são causados por defeitos em metodologia ou procedimentos de medição, ou por descalibração de instrumentos. Podem ser eliminados ou diminuidos pela adoção de correções de procedimentos ou com a calibração de instrumentos de medição. A magnitude dos erros sistemáticos não pode ser estimada, a menos que se conheça o valor real da variável medida.

Erros aleatórios têm origem em variações temporais, espaciais, estocásticas ou imprevisíveis das grandezas de influência. Não podem ser eliminados, mas seu valor médio tende a zero com o aumento do número de observações, e sua variação pode ser quantificada aplicando-se conceitos estatísticos.

Precisão é uma medida do espalhamento de medições repetidas da mesma variável. O espalhamento se deve ao erro aleatório. Medições precisas possuem erros aleatórios pequenos.

Exatidão, ou acurácia é a resultante dos dois tipos de erros. Uma medição com boa exatidão possui erro sistemático zero e erro aleatório mínimo. A Figura 1.2 ilustra esse conceito, para um caso de comparação entre 4 medidores de pH, quando o valor real da variável é 8,0. Somente o medidor D apresenta boa exatidão.

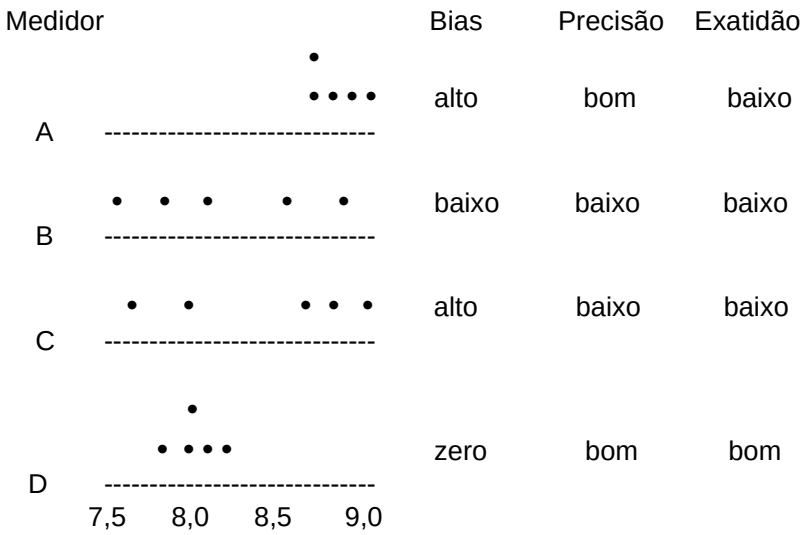


Figura 1.2. Comparação do erro sistemático (“bias”), precisão e exatidão de 4 medidores de pH (5 medidas cada) para o valor real da variável igual a 8,0.

Representação de dados de processo

A primeira etapa na visualização de dados, a partir das planilhas em que normalmente são apresentados, é a construção de gráficos com séries temporais de cada variável, como mostrado na Figura 1.1. Cada ponto no gráfico da série temporal representa um valor armazenado pelo sistema de aquisição de dados, o qual foi ajustado para aquisição de sinal proveniente do sensor com a freqüência de 1 min^{-1}. Em estudos dedicados a caracterizar oscilações de variáveis de processo, é recomendado que a freqüência de amostragem dos dados seja no mínimo igual a 2.$freq_{min}$, em que $freq_{min}$ é a freqüência de corte, ou seja, a menor freqüência que se deseja considerar no estudo. A preparação dos dados pode necessitar de procedimentos de suavização (“data smoothing”), baseados em médias móveis no tempo, com segmentos de tempo adequados. A Figura 1.3 ilustra o efeito da suavização: cada curva suavizada foi obtida com o valor médio da variável em diferentes segmentos de tempo, cada um deles envolvendo um número de pontos, *k*, diferente. Para um conjunto de *n* observações de uma variável *x*, a média no tempo para segmentos de *k* pontos cada é obtida pela Eq. 1.3.

$$\bar{x}(k) = \frac{1}{k} \sum_{j=i-k+1}^{i} x_j \qquad i = k,k+1,...,N \tag{1.3}$$

Como mostrado na Figura 1.3, o efeito da suavização é eliminar oscilações de maior freqüência, mantendo as tendências de períodos mais longos. O número de pontos a ser incluído no cálculo da média móvel depende do objetivo de cada estudo específico. O cálculo da média pode ser feito com segmentos superpostos ou seqüenciais. A superposição de segmentos diminui a possibilidade de haver descontinuidade nas séries temporais finais obtidas, como pode ser visto comparando-se as Figuras 1.4 e 1.5.

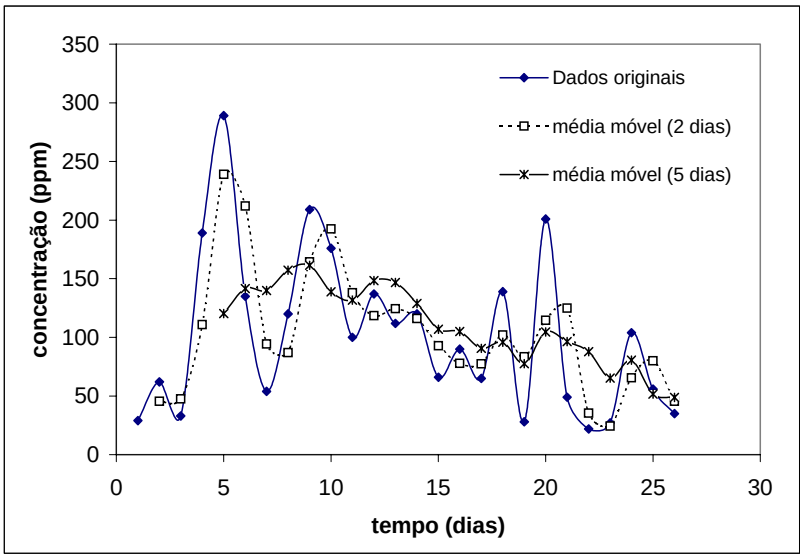


Figura 1.3. Efeito da suavização dos dados: as curvas mostram as observações originais com medidas diárias de concentração de um contaminante em uma corrente de efluente, as médias móveis de 2 dias e de 5 dias.

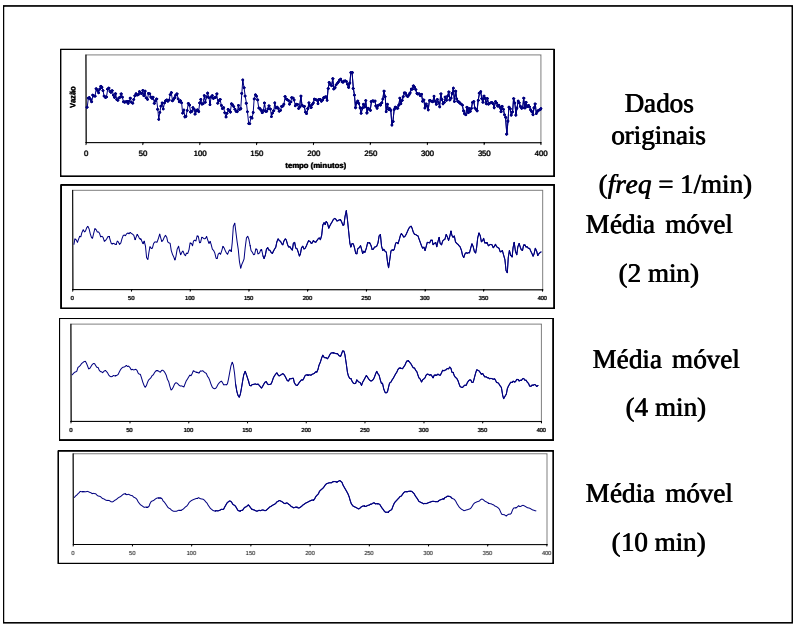


Figura 1.4. Suavização dos dados por segmentos de tempo superpostos.

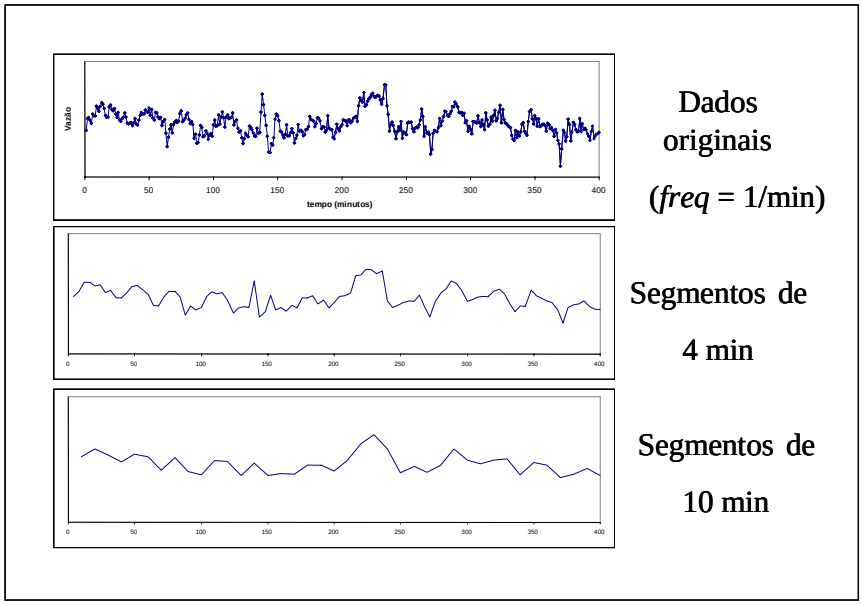


Figura 1.5. Suavização dos dados por segmentos de tempo seqüenciais.

Uma forma usual de representar séries de dados é por meio de gráficos de freqüência, denominados de histogramas. Na Figura 1.6, um conjunto de 27 medições de concentração, apresentadas na tabela, são dispostas em um gráfico e seus valores classificados por classe de concentração, gerando um histograma na forma clássica, com gráfico de barras, em que a altura de cada barra corresponde ao número de observações em cada classe (ou intervalo) de valores. A base de cada barra cobre o intervalo de valores de cada classe. O histograma pode ser expresso na forma de freqüência relativa, ou seja, o número de observações em cada classe dividido pelo total de observações (*n*). Histogramas possibilitam uma estimativa inicial do valor dominante em um conjunto de dados (aquele com maior freqüência de ocorrência), bem como visualizar o grau de dispersão dos dados em torno do valor dominante e observar se a simetria da distribuição. Além disso, podem ser identificadas observações situadas fora da distribuição, ou seja, dados anômalos (conhecidos como “outliers”).

Distribuições de dados

Dados de processo são caracterizados a partir de parâmetros das suas distribuições. Os parâmetros mais utilizados para caracterizar uma população são a média e a variância. A média, μ, também denominada valor esperado, ou esperança de *x*, *E(x)*, para uma população com *N* observações, é dada por:

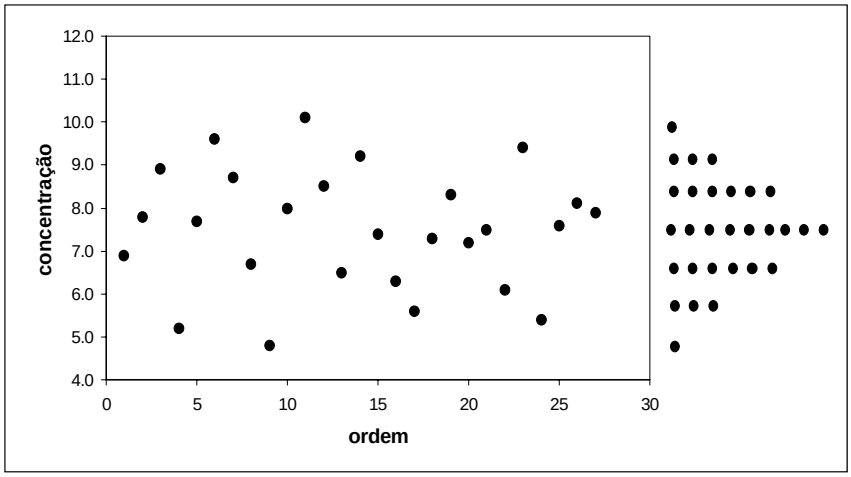
$$E(x) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1.4)$$

A variância, σ^2, expressa a dispersão dos dados em relação à média:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (1.5)$$

1	2	3	4	5	6	7	8	9	10	11	12	13	14
6,9	7,8	8,9	5,2	7,7	9,6	8,7	6,7	4,8	8,0	10,1	8,5	6,5	9,2
15	16	17	18	19	20	21	22	23	24	25	26	27	
7,4	6,3	5,6	7,3	8,3	7,2	7,5	6,1	9,4	5,4	7,6	8,1	7,9	

Dados originais (27 medidas de concentração)



Representação e classificação dos dados

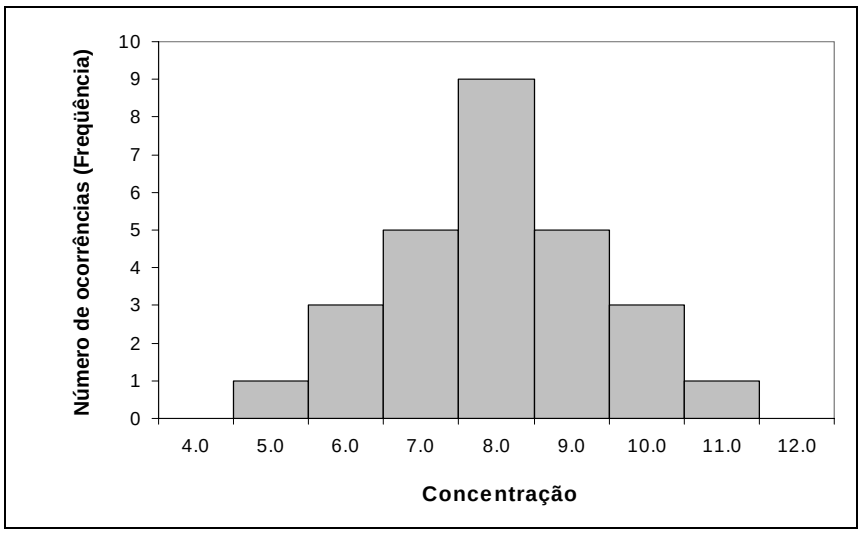


Figura 1.6. Seqüência para obtenção de histogramas de dados.

A divisão por (N – 1) deve-se ao número de graus de liberdade, GL, associados ao cálculo da variância, ou seja, o número de informações necessárias para que a Eq. 1.5 fique determinada. Para o cálculo da média, são necessárias as N observações ($GL = N$); para o cálculo da variância, são necessárias N – 1 observações e a média (pois com N – 1 observações e a média calcula-se a N-ésima observação). A variância tem valor positivo ou nulo. A dispersão é normalmente representada pela raiz quadrada da variância, ou seja, o desvio padrão, σ:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}} \tag{1.6}$$

Outras características de populações de dados utilizadas comumente são:

Moda, ou valor dominante: valor da variável correspondente ao maior valor na distribuição de freqüências de ocorrência, o que corresponde ao máximo no histograma.

Mediana: valor que divide uma população de observações, quando estas são ordenadas.

Exemplo: considerando a seguinte população com 9 dados, ordenados em ordem crescente:

dados: 22 24 24 25 27 30 31 35 40; as propriedades dessa população são:

média: μ = 28,7; desvio padrão: σ = 5,9; mediana: x_{med} = 27;

considerando que seja incluído um dado a mais, ficando a população com 10 dados:

dados: 22 24 24 25 27 30 31 35 40 42; as propriedades são:

μ = 30; σ = 7; x_{med} = 28,5;

substituindo-se o último dado, 42, por um valor bem maior, por exemplo 85, a população fica:

dados: 22 24 24 25 27 30 31 35 40 *85*; as propriedades são:

μ = 34,3; σ = 18,7; x_{med} = 28,5.

Nesse exemplo, ao ser adicionado um dado a mais à população (de 9 para 10), todos os parâmetros foram afetados. Ao ser substituído um dos dados (42) por outro de valor consideravelmente maior (85), tanto a média quanto o desvio padrão foram afetados (este muito mais, pois a dispersão da distribuição aumentou muito). No entanto, não houve alteração na mediana. Esta é uma propriedade importante da mediana: é pouco sensível a variações nos valores extremos dos dados de uma população, o que o torna menos suscetível a alterações em casos de presença de dados anômalos.

Há outros parâmetros para caracterizar uma população de dados, que são utilizados em aplicações específicas, como o coeficiente de variação, quartis, percentis etc, cuja definição

pode ser encontrada em livros de estatística. Podem-se representar graficamente distribuições de observações na forma de caixa e alongamentos (“box and whiskers”), como mostrado na Fig. 1.7. O gráfico possibilita visualizar rapidamente características importantes de uma população de observações. Começando pela caixa, que cobre 50% dos dados, com limites no quartil referente aos 25% menores valores das observações e no quartil correspondente aos 75% menores valores. Um quartil corresponde a uma quarta parte das observações, dispostas em ordem crescente de valor. Os segmentos, ou alongamentos (“whiskers”), cobrem toda a faixa de valores das observaçõe da variável (de x_{min} a x_{max}). Novos valores medidos, que se situem fora da faixa, são considerados anômalos, podendo corresponder a medidas erradas, por exemplo. No gráfico, é indicada a mediana, que divide a população em duas partes iguais e dá uma idéia inicial sobre a simetria da distribuição dos dados.

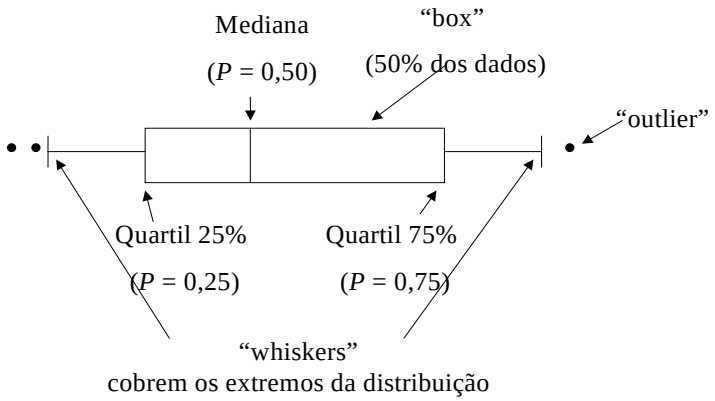


Figura 1.7. Representação de distribuição de dados na forma de “box” e “whiskers”.

Função Densidade de Probabilidade

Seja $\Delta P(x)$ a probabilidade de ocorrência de um dado valor da variável aleatória x, pertencente a uma população. Então, das propriedades de probabilidade, sabe-se que:

$$\sum_{i=1}^{N} \Delta P(x_i) = 1 \tag{1.7}$$

Se x for uma variável contínua, então a probabilidade existe para qualquer valor de x dentro do domínio da variável e pode-se expressar a Eq. 1.7 na forma integral:

$$\int_{x} dP(x) = 1 \tag{1.8}$$

Definem-se as funções probabilidade acumulada, *P(x)*, de ocorrência de um dado valor menor ou igual a *x* e a função densidade de probabilidade, *p(x)*, da seguinte forma (Figura 1.8):

$$P(x) = \int_{-\infty}^{x} dP(x), \qquad 0 \le P(x) \le 1 \tag{1.9}$$

$$p(x) = \frac{dP}{dx} \tag{1.10}$$

sendo, então:

$$P(x) = \int_{-\infty}^{x} p(x)dx \tag{1.11}$$

Portanto,

$$P(\infty) = \int_{-\infty}^{\infty} p(x)dx = 1 \tag{1.12}$$

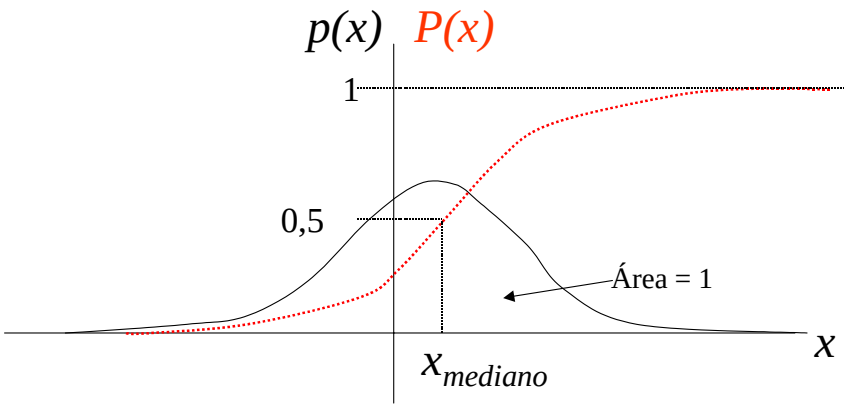


Figura 1.8. Aspecto típico e propriedades das funções *P(x)* e *p(x).*

Pode-se, então, expressar a média, μ, de uma variável *x* contínua, como:

$$\mu = \int_{-\infty}^{\infty} x.p(x)dx = \int_{-\infty}^{\infty} x.dP(x) \tag{1.13}$$

Graficamente, a média representa a área indicada na Figura 1.9.

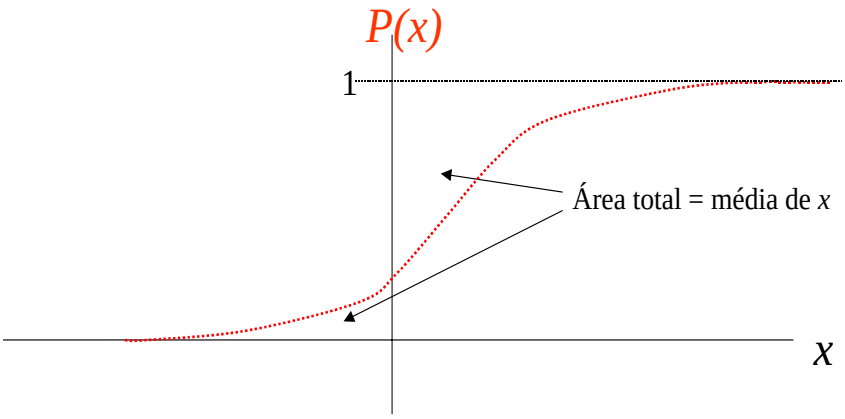


Figura 1.9. Representação gráfica da média, $\bar{x}$.

A variância é expressa como:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x) dx \tag{1.14}$$

A variância é chamada de momento de ordem 2 de *x* em torno da média, expressando assim uma medida da dispersão quadrática dos dados em torno da média. De modo geral, define-se um momento de ordem *t* de uma distribuição de valores de *x* como:

$$M_t = \int_{-\infty}^{\infty} x^t \, p(x) dx \tag{1.15}$$

Os momentos são importantes para uma série de cálculos associados a processos químicos, relacionados a propriedades de distribuições de dados, assim como na engenharia de processo que trata de distribuições em populações, como em processos industriais de polimerização, cristalização, e no estudo de escoamento de fluidos em vasos de processo.

Tipos comuns de distribuições

Quando se trabalha com variáveis aleatórias, ou seja, cujo valor é afetado por fatores não controlados, é necessário incorporar nos cálculos a probabilidade de uma variável assumir um determinado valor. As ferramentas disponíveis para estimativa de intervalos para valores de variáveis e parâmetros de distribuições baseiam-se em formas típicas para algumas situações características. Para variáveis binárias, por exemplo, como na contagem de falhas em sistemas de instrumentação em uma indústria, ou de falhas na qualidade de um produto, é comum o uso de funções de distribuição binomial, ou de distribuição geométrica. Nos casos em que eventos devem ser classificados em duas ou mais categorias, utiliza-se a distribuição multinomial. Para contagens de tempo entre eventos em um dado processo, ou de freqüências de falhas em um sistema qualquer, por exemplo, utiliza-se a distribuição de Poisson. Expressões matemáticas para essas funções possibilitam o cálculo de probabilidades de números de falhas, ou de

distribuição de períodos de tempos entre falhas, por exemplo. Descrições e aplicações dessas distribuições são apresentadas em livros sobre estatística. As variáveis com as quais engenheiros de processo normalmente se ocupam são variáveis contínuas (como vazões, composições, pressão, ou temperatura), coletadas na forma de séries temporais, e o interesse é obterem-se informações sobre a unidade industrial baseadas na distribuição de valores dessas variáveis ao longo do tempo. Os valores de variáveis aleatórias contínuas apresentam curvas de distribuições na forma de sinos, sendo a mais comum a distribuição normal.

A distribuição normal é baseada no fato de que as medidas de uma variável aleatória qualquer são afetadas por um conjunto de flutuações de muitos fatores independentes. Com essa hipótese, o teorema do limite central prevê que os valores da variável têm distribuição da função densidade de probabilidade segundo uma função gaussiana, ou normal.

Teorema do limite central

Se a flutuação total no valor de uma variável aleatória for o resultado da soma das flutuações de muitas variáveis independentes e de importância aproximadamente igual, então a distribuição de valores tenderá para a distribuição normal, não importando a natureza das distribuições das variáveis individuais. O exemplo clássico é o jogo de dados. A distribuição de probabilidades para o lançamento de um dado não viciado por um grande número de vezes é mostrada na Figura 1.10a: os valores possíveis são os inteiros de 1 a 6, com probabilidades iguais. Porém, se forem anotados os valores médios de muitos lançamentos de, por exemplo, 5 dados (ou as médias de cada 5 lançamentos de um mesmo dado), então há 5 eventos independentes, com igual probabilidade, contribuindo para o valor da média. O resultado da distribuição da média é mostrado na Figura 1.10b. À medida que aumenta o número de dados que compõem a média, a distribuição tende a uma curva em forma de sino. Por exemplo, a Figura 1.10c mostra a distribuição da média de 10 lançamentos de dados.

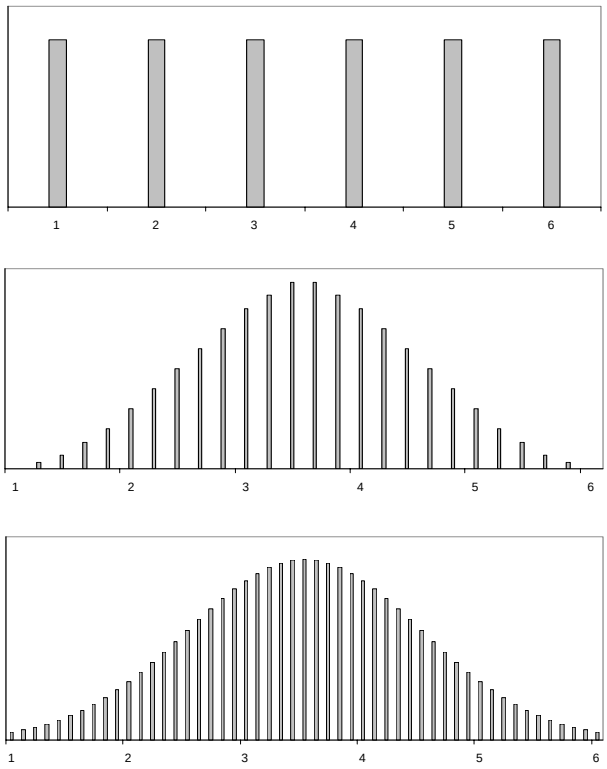


Figura 1.10. Distribuição de probabilidades da média do lançamento de 1 (a), 5 (b) e 10 (c) dados não viciados (Barros Neto *et al*, 2001).

A distribuição normal, ou distribuição gaussiana, de uma população com média μ e desvio padrão σ tem função densidade de probabilidade expressa na forma da Eq. 1.14.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.14}$$

A Figura 1.11 apresenta um gráfico dessa função, para o caso em que $\mu = 0$ e $\sigma = 1$.

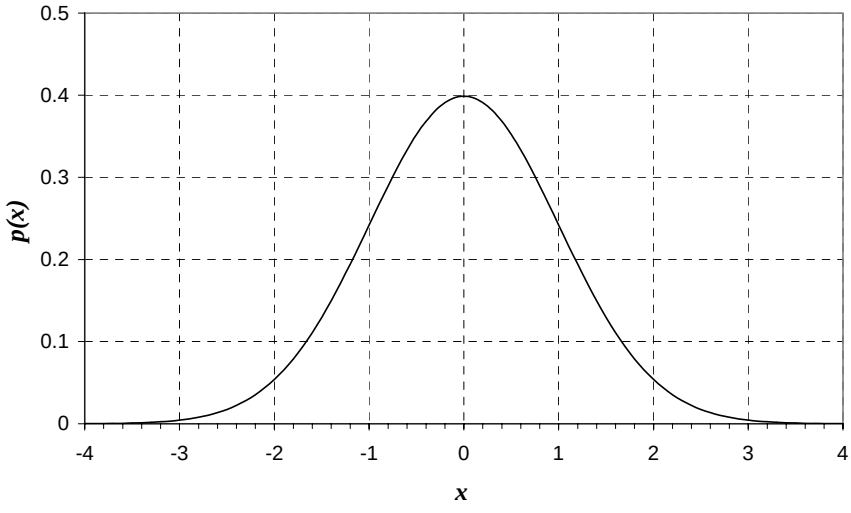


Figura 1.11. Distribuição normal da função densidade de probabilidade para uma variável aleatória *x*, com $\mu = 0$ e $\sigma = 1$.

Algumas propriedades importantes da distribuição normal:

1) A área sob a curva corresponde à probabilidade. A área total, portanto, vale 1.
2) A curva é simétrica em relação à média e em forma de sino.
3) A probabilidade do valor da variável estar no intervalo entre a média e $\pm\ \sigma$ é de 68,26%; a probabilidade da medida estar no intervalo entre a média e $\pm\ 3\sigma$ é de 99,73%.

Portanto, praticamente todos os dados estão contidos no intervalo entre a média e $\pm 3\sigma$. As probabilidades são calculadas aplicando-se a Eq. 1.11. Por exemplo, a probabilidade de que $\mu - \sigma < x < \mu + \sigma = P(\mu + \sigma) - P(\mu - \sigma)$.

Amostra e População

Uma população com *N* elementos constitui o conjunto completo de observações de uma variável. Uma amostra é um subconjunto da população, com *n* observações. Supõe-se que uma população seja um grande conjunto de *N* observações, do qual são retiradas amostras. Uma amostra representativa tem as mesmas características da população da qual a amostra foi retirada. As características de uma população podem ser estimadas a partir de características da distribuição de valores na amostra. A média, x_{med}, e a variância, s^2, de uma amostra com *n* observações são expressas como:

$$x_{med} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.15}$$

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - x_{med}\right)^2}{n-1} \tag{1.16}$$

Para amostras contendo *n* observações cada, extraídas aleatoriamente e independentemente de uma população com distribuição normal com média μ e variância σ^2, pode-se demonstrar que:

As médias amostrais x_{med} têm distribuição normal, com média μ e variância igual a:

$$s^2 = \frac{\sigma^2}{n} \tag{1.17}$$

ou seja, à medida que o número de observações na amostra aumenta, a variância da distribuição da média x_{med} tende a 0.

Normalmente as estimativas de parâmetros da distribuição de uma população são feitas utilizando-se variáveis como a variável aleatória *t*, que segue a distribuição de Student, com *n* – 1 graus de liberdade e é definida como:

$$t = \frac{x_{med} - \mu}{s/\sqrt{n}} \tag{1.18}$$

A variância da amostra, s^2, tem distribuição de valores definida pela variável aleatória χ^2, que segue a distribuição qui quadrado, com n – 1 graus de liberdade, definida como:

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2} \tag{1.19}$$

Detalhes a respeito das deduções e uso dessas variáveis, bem como os valores das curvas de distribuição para diferentes valores de *GL* podem ser vistas em livros texto sobre estatística.

Intervalos de confiança

As variáveis t e χ^2 são importantes na estimativa de intervalos de confiança para a média e a variância populacionais a partir das características de uma amostra. Com base nas definições das Eqs. 1.18 e 1.19, os intervalos de confiança são definidos, para a média e variância, como:

$$x_{med} - t_{n-1,(1-\alpha)}\frac{s}{\sqrt{n}} < \mu < x_{med} + t_{n-1,(1-\alpha)}\frac{s}{\sqrt{n}} \tag{1.20}$$

$$(n-1)\frac{s^2}{\chi^2_{n-1,(1-\alpha)}} < \sigma^2 < (n-1)\frac{s^2}{\chi^2_{n-1,(\alpha)}}\,. \tag{1.21}$$

Nas Eqs. 1.20 e 1.21, o subscrito (n – 1) refere-se ao número de graus de liberdade envolvido nos cálculos; o subscrito (1 - α) refere-se ao grau de significância adotado no cálculo do intervalo. O termo α é o grau de confiança, ou seja, a probabilidade de que a característica estimada da população (média na Eq. 1.20, variância na Eq. 1.21) tenha valor dentro do intervalo. Na maioria dos casos adota-se α = 0,95, ou seja, a probabilidade de que a característica da população tenha valor maior ou menor que o intervalo é de 2,5% (no total, portanto, a probabilidade de erro na estimativa do intervalo é de 5%, igual a 1 - α). Livros-texto de estatística (por exemplo, Barros Neto *et al*, 2001) apresentam tabelas com valores de t e χ^2 para diferentes valores de *GL* e α. Para amostras com número de observações n maior que 30, os valores de t e χ^2 têm variações muito pequenas. As variáveis aleatórias t e χ^2 são utilizadas freqüentemente em testes de hipóteses, para comparações entre médias e variâncias de diferentes amostras.

Procedimento para estimar intervalo de confiança para a média μ de uma população, a partir de uma amostra com *n* observações

Passo 1: definir o nível de confiança, α (p. ex. 95%);

Passo 2: determinar a solução *t* da equação: $F(t)=\frac{1}{2}(1+\alpha)$, usando uma tabela de distribuição de *t* de Student para *n* – 1 graus de liberdade (*n* = número de observações);

Passo 3: calcular a média x_{med} e a variância s^2 da amostra;

Passo 4: calcular $t\frac{s}{\sqrt{n}}$;

O intervalo de confiança é: $x_{med}-t_{n-1}\frac{s}{\sqrt{n}}<\mu<x_{med}+t_{n-1}\frac{s}{\sqrt{n}}$

Procedimento para estimar intervalo de confiança para a variância σ^2 de uma população, a partir de uma amostra com *n* observações

Passo 1: definir o nível de confiança, α (p. ex. 95%);

Passo 2: determinar as soluções χ_1^2 e χ_2^2 das equações: $F(\chi_1^2)=\frac{1}{2}(1-\alpha)$ e $F(\chi_2^2)=\frac{1}{2}(1+\alpha)$, usando uma tabela de distribuição de qui-quadrado para *n* – 1 graus de liberdade (*n* = número de observações);

Passo 3: calcular $(n-1)\frac{s^2}{\chi_1^2}$; $(n-1)\frac{s^2}{\chi_2^2}$, em que s^2 é a variância da amostra;

O intervalo de confiança é: $(n-1)\frac{s^2}{\chi_2^2}<\sigma^2<(n-1)\frac{s^2}{\chi_1^2}$

Tamanho mínimo de uma amostra

Podem-se utilizar os mesmos conceitos para determinar o número mínimo de observações que devem estar contidas em uma amostra, *n*, para que o desvio entre a média da amostra, x_{med}, e a média da população, μ, seja menor que uma dada tolerância, ε, com um grau de confiança α. Neste caso, é necessário conhecer o valor de s^2, o que pode ser feito a partir de dados históricos de amostras da variável *x*, com um número de observações conhecido.

O tamanho da amostra pode ser determinado, para $\varepsilon = \pm(x_{med} - \mu)$, por:

$$n \geq \left(\frac{t_{GL,\alpha} . s}{\varepsilon} \right)^2 . \qquad (1.22)$$

Na Eq. 1.22, *GL* é o número de graus de liberdade utilizado par calcular *s*, o desvio padrão de uma amostra anteriormente conhecida da variável. Pode-se utilizar, para o valor da variável *t*, tanto o valor α, quanto 1 - α, porque a distribuição de *t* é simétrica.