Capítulo 3

O Problema Amostral – Inferências e Comparações

No Capítulo 2 foram apresentadas diversas distribuições de probabilidade que representam diferentes problemas em que variáveis aleatórias estão envolvidas. No entanto, esses modelos probabilísticos dependem de parâmetros que, na maioria absoluta das vezes, não podem ser determinados *a priori*. Por exemplo, na distribuição binomial descrita pela Equação (2.5), quem é o parâmetro p? E na distribuição normal descrita pela Equação (2.53), quem são os parâmetros μ (média) e σ^2 (variância)? Repare que uma pessoa desavisada poderia dizer que a média μ e a variância σ^2 são os valores calculados pela definição de média da Equação (1.71) e de variância da Equação (1.72). No entanto, para que a média e a variância sejam calculadas a partir das definições introduzidas pelas Equações (1.71) e (1.72), é necessário que a distribuição de probabilidades normal da Equação (2.53) esteja perfeitamente definida, o que significa que μ e σ^2 devem ser conhecidos. Essa contradição indica claramente que os parâmetros da distribuição têm que ser obtidos de outra forma, que não a partir das definições introduzidas nos Capítulos 1 e 2. Se o problema analisado tiver caráter multivariável, como aqueles abordados nas Seções 2.8 a 2.10, o número de parâmetros da distribuição pode ser muito grande. Portanto, é necessário desenvolver técnicas que permitam inferir os parâmetros que descrevem os modelos probabilísticos, para que eles de fato possam ser úteis para a análise de problemas reais.

Mas por que é tão importante que se conheça a distribuição de probabilidades que está associada a um determinado problema? A resposta fundamental dessa questão é que, se as curvas de distribuição de probabilidades que descrevem as flutuações aleatórias observadas em certos problemas são conhecidas, então é possível comparar os problemas e discriminar aqueles resultados que devem ser (e os que não devem ser) esperados. O primeiro caso constitui o conjunto de procedimentos chamados de **testes de hipóteses**. A pergunta típica que gera esse conjunto de procedimentos é: *"Será que uma certa propriedade ou conjuntos de resultados obtidos das diferentes curvas de distribuição analisadas podem ser considerados iguais (diferentes)?"*. Como será visto nos próximos capítulos, o analista é chamado todo o tempo a opinar sobre essa questão, para saber se um processo ou conjunto de resultados permanece constante ou está mudando. O segundo caso constitui o conjunto de problemas chamados de **determinação dos intervalos de confiança**. A pergunta típica que gera esse conjunto de procedimentos é: *"Qual é o conjunto de resultados mais provável?"*, ou ainda *"Que resultados podem ser descartados com certo grau de confiança?"*. Como veremos nos capítulos seguintes, respostas para essas questões permitem racionalizar sobre a qualidade dos resultados obtidos experimentalmente e sobre o conteúdo de informação

disponível para análise. Além disso, as respostas dessas perguntas quase sempre geram procedimentos de projeto e rotinas de decisão, como visto no Exemplo 2.3.

Para resolver as questões propostas acima, é necessário amostrar o sistema; isto é, tomar medidas representativas do problema estocástico considerada. O objeto fundamental desse capítulo é discutir como medidas experimentais podem ajudar o analista a definir as distribuições de probabilidade que descrevem as flutuações observadas e, dessa forma, permitir a comparação de resultados e a tomada de decisão.

3.1. Definição de Intervalo de Confiança

Para que seja possível tomar decisões, é preciso decidir que resultados podem ser considerados normais (ou seja, têm grande probabilidade de ocorrer) e que resultados devem ser considerados anormais (ou seja, que têm probabilidade tão baixa de ocorrer que podem ser descartados na grande maioria das vezes). Para tanto, define-se como o intervalo de p% de confiança ao conjunto de resultados que, segundo a curva de distribuição de probabilidades considerada, concentra p% dos resultados admissíveis. Portanto, são descartados os (100–p%) resultados menos prováveis, sendo (100–p%)/2 desses resultados localizados na extremidade inferior e (100–p%)/2 desses resultados localizados na extremidade superior. A Figura 3.1 ilustra esse conceito.

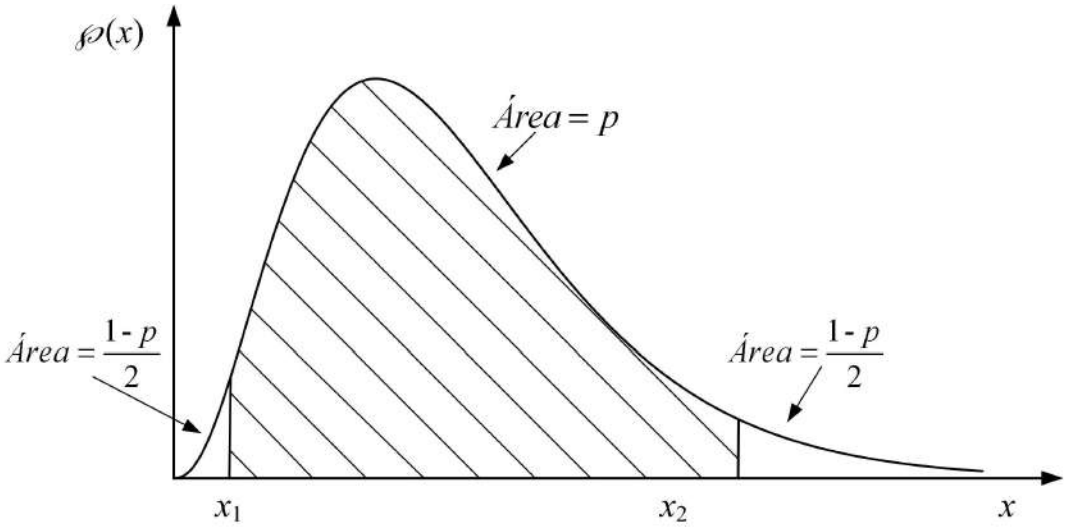


Figura 3.1 - Ilustração gráfica do conceito de intervalo de confiança.

Portanto, se (x_1, x_2) são os limites de confiança com p% de probabilidade de uma certa variável x, descrita por uma curva de densidade de probabilidades $\wp(x)$, então

$$P_{AC}(x_1) = \int_{x_{mín}}^{x_1} \wp(x)\,dx = \frac{1-p}{2} \tag{3.1}$$

$$P_{AC}(x_2) = \int_{x_{mín}}^{x_2} \wp(x)\,dx = 1 - \frac{1-p}{2} = \frac{1+p}{2} \tag{3.2}$$

Os exemplos a seguir ilustram o procedimento de análise proposto.

Exemplo 3.1 - Admita que dois catalisadores industriais distintos seguem diferentes padrões de decaimento de atividade. No primeiro caso, sabe-se que a distribuição de tempo de vida segue a curva exponencial típica, na forma

$$\wp_1(t) = \frac{\exp\left(-\frac{t}{10}\right)}{10}$$

onde t é dado em horas. No segundo caso, sabe-se que a distribuição de tempo de vida segue uma curva gama, na forma

$$\wp_2(t) = \frac{2^{20}}{\Gamma(20)} t^{19} e^{-2t}$$

Comparando-se as médias e variâncias das duas distribuições, obtém-se no primeiro caso (Equações (2.29-30))

$$\mu_{1T} = 10 \text{ e } \sigma_{1T}^2 = 100$$

e no segundo (Equações (2.50-51))

$$\mu_{2T} = 10 \text{ e } \sigma_{2T}^2 = 5$$

Portanto, vê-se que, embora os dois catalisadores apresentem tempos médios de vida iguais (10 h), o tempo de vida do segundo catalisador é muito mais uniforme que o tempo de vida do primeiro catalisador. Dessa maneira, parece muito mais fácil decidir sobre o momento de troca do catalisador no processo industrial no segundo caso que no primeiro. Para ilustrar esse efeito, no primeiro caso o intervalo de confiança de 95% ($p = 0.95$, $(1-p)/2 = 0.025$, $(1+p)/2 = 0.975$) para o tempo de vida do catalisador é

$$(0.25, 36.89)^1_{95\%}$$

enquanto para o segundo é

$$(6.1, 14.8)^2_{95\%}$$

Repare que se o nível de confiança exigido for maior e igual a 98% ($p = 0.98$, $(1-p)/2 = 0.01$, $(1+p)/2 = 0.99$), então os intervalos para cada catalisador são, respectivamente:

$$(0.10, 46.05)^1_{98\%} \text{ e } (5.54, 15.92)^2_{98\%}$$

os quais são intervalos de confiança mais largos devido ao aumento no nível de confiança exigido. A Figura 3.2 ilustra graficamente as duas distribuições de probabilidade analisadas.

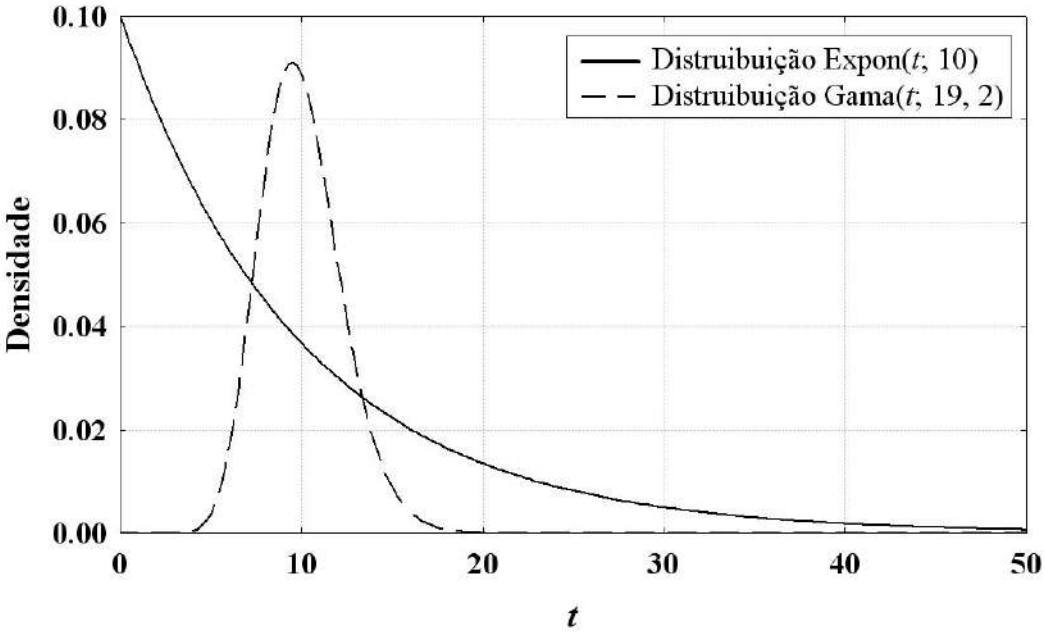


Figura 3.2 - Comparação entre as duas distribuições de tempo de vida dos catalisadores.

Exemplo 3.2 - Freqüentemente é necessário calcular integrais de curvas de densidade de probabilidade, para cômputo de médias, variâncias, intervalos de confiança, etc. Na maior parte dos problemas, no entanto, soluções analíticas não estão disponíveis. Temos portanto que calcular as integrais numericamente.

Muitas técnicas numéricas foram desenvolvidas para o cômputo de integrais e não se pretende aqui fazer uma revisão dessas técnicas. Contudo, uma técnica de integração muito simples está ilustrada na Figura (1.22) e nas Equações (1.66-69). É a chamada **técnica do retângulo para integração**, definida como

$$\overline{x}_i = \frac{x_{i+1} + x_i}{2}, \; x_i = x_{mín} + (i-1)\Delta x$$

$$I = \int_{x_1}^{x_2} F(x)\, dx \approx \sum_{i=1}^{NR} F(\overline{x}_i)\Delta x$$

$$NR = \frac{x_2 - x_1}{\Delta x}$$

que consiste fundamentalmente em aproximar a integral pela soma das áreas dos retângulos que têm base igual a Δx (precisão da integração) e altura igual ao valor da função no ponto médio do intervalo Δx considerado. Portanto, o cálculo das integrais necessárias para a análise dos dados não deve ser considerada uma dificuldade intransponível. Muito pelo contrário, essas integrais podem ser calculadas até com certa facilidade.

Por exemplo, seja a curva exponencial do Exemplo 3.1, dada por

$$\wp(t) = \frac{\exp\left(-\frac{t}{10}\right)}{10}$$

cujo valor médio é conhecido e igual a 10. Numericamente, o valor médio pode ser obtido na forma

$$\mu_T = \int_0^{\infty} t\wp(t)\,dt \approx \int_0^{100} t\wp(t)\,dt \approx \sum_{i=1}^{NR} \bar{t}_i \frac{\exp\left(-\frac{\bar{t}_i}{10}\right)}{10}\Delta t$$

$$\bar{t}_i = \frac{t_{i+1} + t_i}{2},\ t_i = 0 + (i-1)\Delta t$$

$$NR = \frac{100 - 0}{\Delta t}$$

A Tabela 3.1 ilustra a qualidade dos resultados obtidos para diferentes valores de Δt. Observe que a convergência dos resultados é bastante rápida, à medida que a precisão da integração aumenta (Δt diminui). Um resíduo final é observado porque a integral é computada até o limite máximo de 100, que serve como referência para o limite superior infinito.

Tabela 3.1 - Convergência do procedimento de integração numérica usado para o cálculo da média da curva Expon(*t*; 10).

Δt	100	10	5	1	0.5	0.1	0.05
NR	1	10	20	100	200	1000	2000
I	3.369	10.377	10.097	9.999	9.996	9.995	9.995

Para fins de tomada de decisão, todo resultado observado que não estiver contido no intervalo de confiança pode ser considerado anormal (improvável), de maneira que ele indica a mudança de comportamento do sistema estudado ou o aparecimento de um novo fato, até então desconsiderado. Deve ser enfatizado que, ao se definir o intervalo de confiança com *p*% de probabilidade, define-se implicitamente que as decisões estarão erradas (100-*p*)% das vezes. Portanto, pode-se dizer que o estabelecimento do nível de confiança é equivalente à definição da fração de vezes que um erro pode ser tolerado. Por exemplo, ao se dizer que uma variável aleatória está num certo intervalo 95% das vezes, diz-se simultaneamente que ela não está naquele intervalo 5% das vezes por razões meramente aleatórias. Portanto, ao se dizer que a observação de um valor fora do intervalo de confiança indica uma mudança, erra-se 5% das vezes.

Erroneamente costuma-se acreditar que, quanto maior o nível de confiança exigido, menor o intervalo de confiança. Preste atenção que o resultado correto é exatamente o oposto: quanto maior o nível de confiança exigido, mais largo o intervalo

de confiança. Isso ocorre porque é necessário incluir maior quantidade de resultados possíveis, à medida que aumenta o grau de confiança exigido. Isso cria um problema para o processo de tomada de decisão muito interessante:

a) Para aumentar a confiança e diminuir o risco de erro no processo de tomada de decisão, aumenta-se o nível de confiança exigido;
b) À medida que se aumenta o nível de confiança, aumenta-se simultaneamente o conjunto de resultados possíveis e diminui-se o número de resultados considerados pouco prováveis, tornando o processo de tomada de decisão sobre o que é possível e o que não é possível mais difícil.

Por exemplo, considere os resultados obtidos no Exemplo 3.1 com a distribuição gama. Suponha ainda que foi observada perda de atividade para uma pastilha de catalisador após 6 horas de operação. Será que algo mudou no processo? No limite de 95% de confiança (portanto a probabilidade de tomar uma decisão errada é de 5% ou 1 em 20) é possível dizer que algo estranho ocorreu, pois o tempo de vida de 6h é pouco provável. No entanto, no limite de 98% de confiança (portanto a probabilidade de tomar uma decisão errada é de 2% ou 1 em 50) não é possível dizer que ocorreu mudança no processo, já que 6h é um valor provável. No limite de 100% de confiança, qualquer valor seria possível! Veja que fica muito mais difícil detectar falhas quando o nível de confiança exigido sobe, embora as decisões sejam sempre tomadas com mais segurança.

Pelas razões discutidas acima, não é possível generalizar nem recomendar de forma absoluta um nível ótimo de confiança para determinação dos intervalos de confiança e tomada de decisão. Cada processo e cada analista definem o intervalo de confiança adequado para a análise executada. Se uma eventual decisão equivocada não envolve riscos nem custos muito grandes, pode-se trabalhar com níveis de confiança mais baixos e aumentar a velocidade do processo de detecção de falhas e/ou mudanças do processo (essa é uma estratégia arrojada). Se uma eventual decisão equivocada pode comprometer seriamente a segurança e/ou a economia do processo, deve-se trabalhar com níveis de confiança mais altos, sabendo-se que essa estratégia certamente provocará atrasos no processo de tomada de decisão (essa é uma estratégia conservadora). Os níveis típicos de confiança utilizados para tomadas de decisão são os níveis de 90%, 95%, 98% e 99%, com utilização muito mais freqüente dos níveis de confiança de 95% e 98%.

Exemplo 3.3 - Conforme discutido na seção anterior, a curva normal é muito utilizada para representação de erros de medida. Portanto, é muito conveniente determinar os limites típicos de confiança para variáveis que apresentam flutuações normalmente distribuídas.

A Tabela A.1 encaminhada no Apêndice apresenta as probabilidades da curva normal, parametrizada na forma

$$\text{Normal}(u;0.1),\ u=\left(\frac{x-\mu_X}{\sigma_X}\right)$$

onde u representa a variável x normalizada. A Tabela A.1 só contém as probabilidades acumuladas de valores positivos de u, uma vez que a curva normal é simétrica e

$$P_{AC}(u) = 1 - P_{AC}(-u)$$

Para ler a Tabela A.1, considere a linha 1.0 e a coluna 0.05, onde se encontra o número 0.8531. Nesse caso,

$$P_{AC}(1.05) = 0.8531$$
$$P_{AC}(-1.05) = 1 - 0.8531 = 0.1469$$

Usando a Tabela A.1, para obter o intervalo de confiança de 90%, procura-se o limite inferior onde $P_{AC}(u_1) = 0.05$ e o limite superior onde $P_{AC}(u_2) = 0.95$. Segundo a Tabela A.1, $u_2 \approx 1.65$ ($P_{AC}(1.65) = 0.9505$). Pela simetria da curva normal, conclui-se que $u_1 \approx -1.65$ ($P_{AC}(-1.65) = 1 - 09505 = 0.0495$). Logo, os limites de 90% de confiança de uma variável distribuída normalmente são

$$x_1 = \mu_X - 1.65\sigma_X < x < \mu_X + 1.65\sigma_X = x_2$$

Usando a Tabela A.1, para obter o intervalo de confiança de 95%, procura-se o limite inferior onde $P_{AC}(u_1) = 0.025$ e o limite superior onde $P_{AC}(u_2) = 0.975$. Segundo a Tabela A.1, $u_2 \approx 1.96$ ($P_{AC}(1.96) = 0.9750$). Pela simetria da curva normal, conclui-se que $u_1 \approx -1.96$ ($P_{AC}(-1.96) = 1 - 09750 = 0.0250$). Logo, os limites de 95% de confiança de uma variável distribuída normalmente são

$$x_1 = \mu_X - 1.96\sigma_X < x < \mu_X + 1.96\sigma_X = x_2$$

Usando a Tabela A.1, para obter o intervalo de confiança de 98%, procura-se o limite inferior onde $P_{AC}(u_1) = 0.01$ e o limite superior onde $P_{AC}(u_2) = 0.99$. Segundo a Tabela A.1, $u_2 \approx 2.33$ ($P_{AC}(2.33) = 0.9901$). Pela simetria da curva normal, conclui-se que $u_1 \approx -2.33$ ($P_{AC}(-2.33) = 1 - 09901 = 0.0099$). Logo, os limites de 98% de confiança de uma variável distribuída normalmente são

$$x_1 = \mu_X - 2.33\sigma_X < x < \mu_X + 2.33\sigma_X = x_2$$

Usando a Tabela A.1, para obter o intervalo de confiança de 99%, procura-se o limite inferior onde $P_{AC}(u_1) = 0.005$ e o limite superior onde $P_{AC}(u_2) = 0.995$. Segundo a Tabela A.1, $u_2 \approx 2.58$ ($P_{AC}(2.58) = 0.9951$). Pela simetria da curva normal, conclui-se que $u_1 \approx -2.33$ ($P_{AC}(-2.58) = 1 - 09951 = 0.0049$). Logo, os limites de 99% de confiança de uma variável distribuída normalmente são

$$x_1 = \mu_X - 2.58\sigma_X < x < \mu_X + 2.58\sigma_X = x_2$$

Esses limites de confiança serão muito utilizados para análise de dados ao longo das seções e capítulos posteriores.

3.2. O Problema de Amostragem

Os exemplos da seção anterior mostram que, uma vez conhecida a distribuição de probabilidades que governa um certo problema estocástico, muitas informações úteis e procedimentos de tomada de decisão podem ser construídos. No entanto, a situação real é muito distinta da situação considerada até aqui, pois quase nunca é possível saber *a priori* qual é a distribuição de probabilidades que governa um fenômeno. Pior ainda, mesmo quando a forma da função de distribuição é conhecida, ainda assim os parâmetros que caracterizam a distribuição de probabilidades em geral não são conhecidos. Para medir grandezas físicas, como a temperatura, é possível construir equipamentos de medição, como um termômetro. Infelizmente, não há equipamentos que possam ser conectados aos problemas físicos para determinar as curvas de distribuição de probabilidades dos diferentes problemas. Como proceder então? A resposta é: **EXPERIMENTANDO**!!!

A Equação (1.4), reproduzida abaixo, utilizada para definir a probabilidade de um evento em um problema discreto, mostra que é possível construir um histograma de probabilidades em um problema discreto a partir da repetição do experimento um número suficientemente grande de vezes. Mas o que é um número suficientemente grande de vezes?

$$p_i = \lim_{f_i \to \infty} \left(\frac{f_i}{\sum_{j=1}^{NR} f_j} \right) = \lim_{N_T \to \infty} \left(\frac{f_i}{N_T} \right) \tag{1.4}$$

Exemplo 3.4 - Uma moeda é jogada para o alto várias vezes e a fração de vezes em que se obtém o resultado Cara é lançada no gráfico da Figura 3.3.

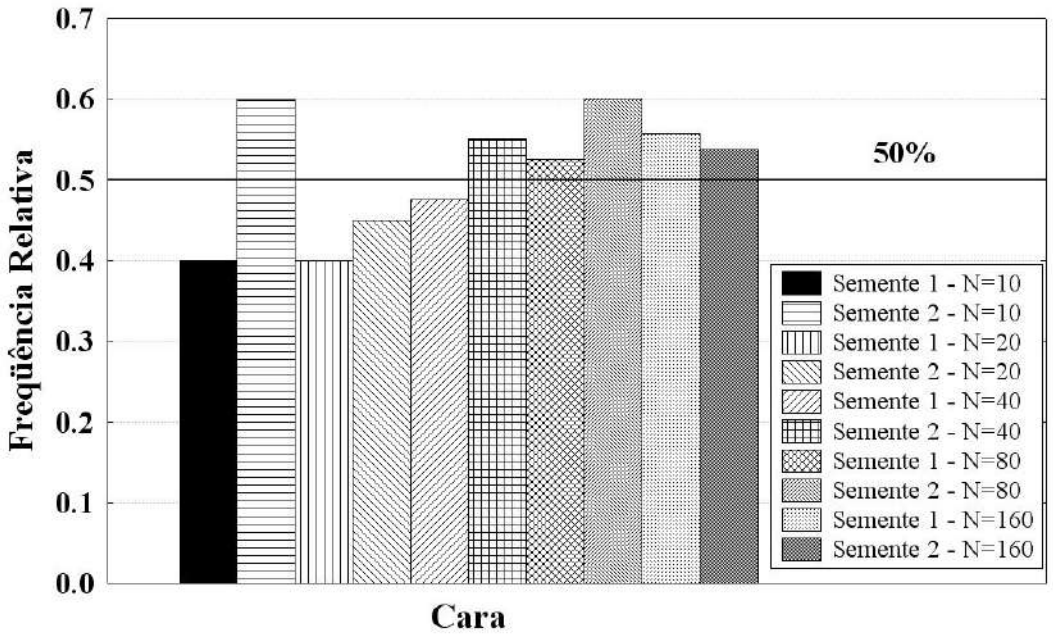


Figura 3.3 - Fração de vezes em que se obtém o resultado Cara no experimento da moeda para várias simulações diferentes

Os experimentos foram realizados no computador, usando-se a seguinte função para geração de números aleatórios com distribuição uniforme

$$X_{k+1} = 11X_k - \text{Trunc}\left(11X_k\right)$$

com sementes X_1=0.40634930 e X_2=0.75832446. A seguinte regra foi usada para decidir sobre o resultado da simulação: X_k < 0.5 é Coroa e X_k > 0.5 é Cara. Podem ser observados grandes desvios do valor nominal, mesmo quando o número de experimentos é bastante grande. Portanto, o infinito pode estar realmente longe!!!! Isso indica de forma clara uma vez mais que não é realista acreditar que as distribuições de probabilidade possam ser construídas unicamente da medida de dados experimentais, já que um número de repetições extremamente elevado pode ser necessário.

Exemplo 3.5 - Uma forma conveniente de gerar curvas de probabilidade acumulada em problemas contínuos a partir da experimentação é admitir uma vez mais a validade da regra de integração por retângulos. Nesse caso, admitindo-se que vários valores foram medidos e foram organizados de forma crescente

$$X_1 \le X_2 \le X_3 \ldots \le X_{N-1} \le X_N$$

pode-se admitir que cada um desses valores limita um intervalo de igual probabilidade, dado que foram esses os intervalos amostrados pela repetição do experimento. Repare que essa argumentação é extremamente questionável, dado que a repetição do procedimento de medida, de forma geral, não resultará na mesma seqüência de valores. No entanto, se essa argumentação é aceita, então

$$P_{AC}\left(X_i\right) = \frac{i}{N+1}$$

onde o denominador (N+1) designa o número de intervalos contínuos definidos pelos N pontos amostrados. Se a mesma função de geração de números aleatórios definida no Exemplo 3.4 e as mesmas sementes são usadas para gerar os pontos experimentais, obtêm-se os resultados apresentados na Figura 3.4. Deve ser observado como as curvas de densidade acumulada são diferentes nos diferentes procedimentos de amostragem, mesmo quando 40 pontos experimentais distintos são amostrados. Isso indica uma vez mais que não é realista acreditar que as distribuições de probabilidade possam ser construídas unicamente da medida de dados experimentais, já que um número de repetições extremamente elevado pode ser necessário.

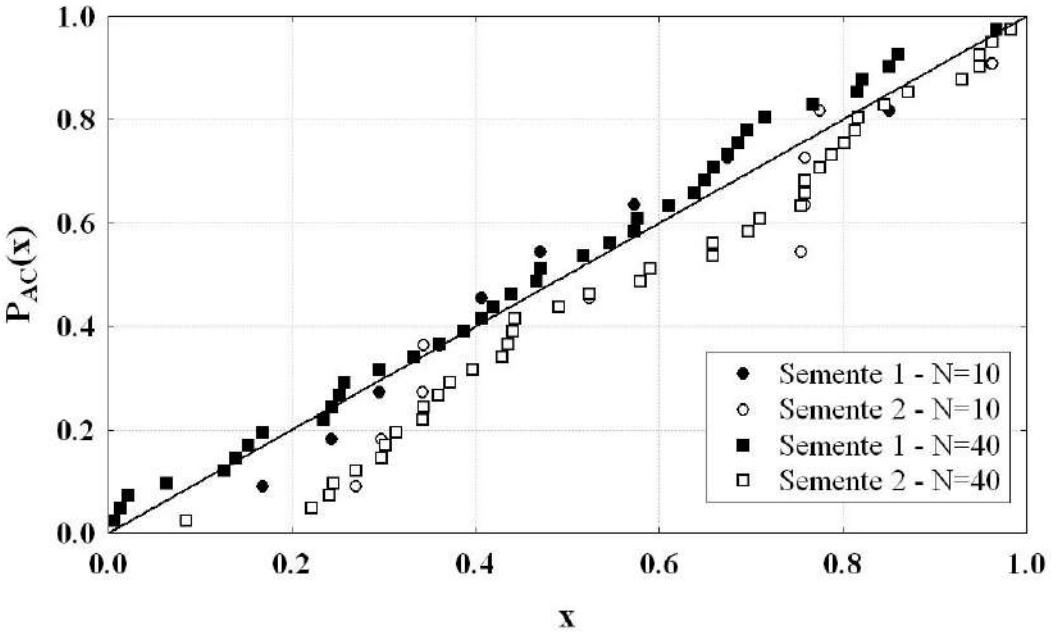


Figura 3.4 - Probabilidade acumulada de pontos gerados pelo gerador de pontos pseudo-aleatórios no Exemplo 3.3, admitindo-se que os intervalos são igualmente prováveis.

Portanto, verifica-se uma vez mais que o infinito pode estar realmente longe!!!!

Os Exemplos 3.4 e 3.5 mostram que, mesmo em problemas muito simples, o número de repetições experimentais necessárias para se construir um histograma ou uma curva de densidade de probabilidades com precisão pode ser muito grande. Na maior parte dos problemas de interesse da engenharia e das ciências básicas, não é possível realizar tantos experimentos por causa do tempo e do custo necessário para a experimentação. Dessa forma, o analista tem que conviver com muitas incertezas a respeito da distribuição real de probabilidades que pode ser associada a um problema físico. Por isso, muito freqüentemente hipóteses são formuladas a respeito de como as curvas de distribuição de probabilidade regulam a flutuação de grandezas físicas reais, como mostrado no Capítulo 2. Conseqüentemente, dificuldades adicionais podem aparecer durante o processo de tomada de decisão, já que algumas medidas flutuam aleatoriamente, já que não se conhece com suficiente precisão a curva de distribuição de probabilidades que governa o problema e uma vez que as hipóteses formuladas não são necessariamente verdadeiras.

Nesse contexto, o uso de modelos de distribuição de probabilidades, como aqueles apresentados no Capítulo 2, é bastante conveniente, pois reduz a busca da distribuição de probabilidades à busca de uns poucos parâmetros que são necessários para descrevê-los. Infelizmente, no entanto, na grande maioria das vezes os modelos são escolhidos sem grande fundamentação teórica ou experimental e muito pouca atenção tem sido dada na literatura técnica às conseqüências práticas que podem resultar de uma escolha mal feita do modelo de distribuição de probabilidades. Por isso, há que se ter cuidado na hora de escolher o modelo mais adequado para descrever as flutuações observadas. (Testes de aderência serão formulados nesse e nos próximos capítulos para

ratificar ou não o modelo de distribuição de probabilidades utilizado para descrever os fenômenos físicos. Como veremos, essa escolha é fundamental para a correta formulação dos problemas de estimação de parâmetros e planejamento experimental.)

3.1.1. Médias e Variâncias Amostrais

Como mostrado no Capítulo 2, na maior parte dos modelos analíticos de distribuições de probabilidades é possível fazer uma associação direta entre os parâmetros do modelo e os valores da média e da variância. Como esses valores são extremamente importantes para caracterizar em torno de que valores e de quanto flutuam os dados experimentais, parece claro que o problema fundamental de ajuste da maior parte dos modelos probabilísticos, e em particular da curva normal, é a determinação da média e da variância a partir dos dados experimentais amostrados. Portanto, admitamos a princípio que um certo conjunto de valores amostrais x_1, x_2, ..., x_N foi obtido a partir da repetição de um certo experimento aleatório. A questão fundamental então é: como obter μ_X e σ_X^2 a partir desse conjunto de dados amostrados?

De acordo com as Equações (1.7) e (1.71), reproduzidas abaixo, o valor médio pode ser obtido a partir do histograma ou da densidade de probabilidades como:

$$\mu_X = \sum_{i=1}^{NR} p_i x_i \tag{1.7}$$

$$\mu_X = \int_{x_{min}}^{x_{max}} x \wp(x)\, dx \tag{1.71}$$

No entanto, de acordo com a discussão dos parágrafos anteriores, não se conhecem as distribuições reais de probabilidade do problema, mas apenas um conjunto de dados amostrados. Como conciliar então a realidade e os objetivos pretendidos? Para isso, formulemos a seguinte hipótese:

Hipótese Fundamental 1.1 - A Hipótese do Experimento Bem Feito

Admita que cada valor experimental pode ser obtido de forma semelhante, seguindo procedimentos idênticos de experimentação e sem vícios na execução dos experimentos. Assim, admita que as flutuações observadas encerram a realidade da natureza experimental do problema e não são influenciadas por erros ou vícios cometidos pelo analista. Nesse caso, cada dado representa igualmente a grandeza experimental desconhecida, em torno da qual as observações experimentais flutuam. Portanto, cada observação experimental pode ser considerada igualmente provável e a cada uma das observações x_1, x_2, ..., x_N pode ser associada a mesma probabilidade $p_i = 1/N$ de que este seja o melhor valor para representar a medida física real.

Se a hipótese do experimento bem feito é aceita, então, por analogia direta com a Equação (1.7), é possível escrever:

$$\overline{X} = \sum_{i=1}^{N} p_i x_i = \sum_{i=1}^{N} \left(\frac{1}{N}\right) x_i = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.3}$$

onde $\overline{X}$ é a chamada **média amostral** do conjunto de dados. Antes que se seja tentado a confundir $\overline{X}$ com μ_X, é conveniente perceber os resultados apresentados no exemplo abaixo.

Exemplo 3.6 - Nas Tabelas 3.2 e 3.3 apresentam-se as médias amostrais calculadas para os problemas analisados nos Exemplos 3.4 e 3.5.

Tabela 3.2 - Médias amostrais obtidas no Exemplo 3.4.

N	10		20		40		80		160		∞
Semente	1	2	1	2	1	2	1	2	1	2	-
$\overline{X}$	0.500	0.400	0.600	0.450	0.500	0.425	0.538	0.438	0.513	0.506	0.500

Tabela 3.3 - Médias amostrais obtidas no Exemplo 3.5.

N	10		20		40		80		160		∞
Semente	1	2	1	2	1	2	1	2	1	2	-
$\overline{X}$	0.518	0.483	0.559	0.422	0.512	0.488	0.547	0.516	0.521	0.513	0.500

Observe que a média amostral flutua de experimento para experimento em torno da média verdadeira, igual a 0.500 em ambos os casos. A média amostral, portanto, não deve ser confundida com a média real da distribuição de probabilidades amostrada, que o analista a princípio desconhece.

O Exemplo 3.6 mostra claramente que a média amostral $\overline{X}$ flutua e, por isso, não deve ser confundida com a média verdadeira μ_X da distribuição. (Se houver dúvidas a esse respeito, lembre que o valor médio do experimento dos dados é 3.5, como mostrado no Exemplo 1.4. No entanto, parece perfeitamente normal jogar o dado três vezes e obter o número 1 três vezes seguidas, resultando na média amostral $\overline{X}$=1.) Mais ainda, se a média amostral flutua de experimento para experimento (nesse caso o experimento consiste em tomar amostras de tamanho N), ela é também uma variável aleatória, assim como os dados amostrados x_i. Portanto, a média amostral $\overline{X}$ deve ser encarada como uma variável aleatória que flutua em torno de certo valor médio e com certa variância, que devem a princípio ser caracterizados, assim como a distribuição de probabilidades que descreve as flutuações de $\overline{X}$. Mas certamente a conseqüência mais importante dessa discussão é que não devemos ter esperanças de obter o valor real da média μ_X, a não ser que tenhamos a distribuição real de probabilidades do problema, o que, segundo a discussão apresentada na seção anterior, de maneira geral não é possível. Dessa forma, se tivermos que obter informações sobre o problema a partir da experimentação (amostrando), **nunca saberemos qual é o valor verdadeiro da média μ_X.**

Embora a discussão anterior pareça um pouco frustrante, ela coloca a perspectiva verdadeira que o experimentador deve ter em relação aos dados obtidos a

partir da observação experimental. Não apenas os dados flutuam, em função dos diversos erros experimentais apresentados nas seções iniciais, como também os valores obtidos a partir da manipulação desses dados, como a média amostral, também flutuam. Dessa forma, o experimentador tem que aprender a conviver com essas incertezas e a caracterizar as flutuações com que convive. Em particular, para o procedimento de cálculo da média amostral é possível escrever as seguintes propriedades.

Propriedade 3.1 - Se os experimentos x_i, i=1...N, são todos realizados em condições idênticas e flutuam em torno da média verdadeira μ_X, a média amostral $\overline{X}$ também flutua em torno do valor médio verdadeiro μ_X.

$$\mathrm{E}\left\{\overline{X}\right\} = \mathrm{E}\left\{\frac{\sum_{i=1}^{N} x_i}{N}\right\} = \frac{1}{N}\mathrm{E}\left\{\sum_{i=1}^{N} x_i\right\} = \frac{\sum_{i=1}^{N}\mathrm{E}\{x_i\}}{N} = \frac{\sum_{i=1}^{N}\mu_X}{N} = \mu_X \tag{3.4}$$

Repare que a Propriedade 3.1 (Equação (3.4)) dá o alento de garantir que, embora o valor da média amostral não possa ser confundido com o valor da média real, na média o valor da média amostral é igual ao valor da média real. (*Observe como a propriedade de linearidade da média foi útil para escrever a Equação (3.4).*) Isso significa que, se o experimento usado para obtenção da média amostral for repetido infinitas vezes, na média o experimento resultará na obtenção da média real. No entanto, na prática o experimento será realizado UMA ÚNICA VEZ, para uma amostra de tamanho N. Por isso, a Propriedade 3.1 não garante a obtenção do valor médio verdadeiro para um conjunto finito de experimentos, mas garante a **consistência** do procedimento experimental usado. Podemos ao menos garantir que a média amostral flutua em torno do valor médio verdadeiro. No entanto, como ambos x_i e $\overline{X}$ flutuam ao redor da mesma média verdadeira μ_X, qual seria então a utilidade de se calcular a média amostral? A Propriedade 3.2 responde a essa pergunta.

Propriedade 3.2 - Se as medidas experimentais x_i, i=1,...,N, são medidas independentes ($\sigma^2_{X_i,X_j} = 0,\ i \neq j$) realizadas em condições idênticas e flutuam todas em torno da mesma média verdadeira μ_X com variância σ^2_X, então a média amostral $\overline{X}$ flutua em torno do valor médio verdadeiro μ_X com variância igual a $\sigma^2_{\overline{X}} = \sigma^2_X / N$.

$$\mathrm{Var}\left\{\overline{X}\right\} = \mathrm{E}\left\{\left(\overline{X}-\mu_X\right)^2\right\} = \mathrm{E}\left\{\left(\frac{\sum_{i=1}^{N} x_i}{N}-\mu_X\right)^2\right\} = \mathrm{E}\left\{\left(\frac{\sum_{i=1}^{N} x_i - N\mu_X}{N}\right)^2\right\} =$$

$$\frac{1}{N^2}\mathrm{E}\left\{\left(\sum_{i=1}^{N} x_i - N\mu_X\right)^2\right\} = \frac{1}{N^2}\mathrm{E}\left\{\left(\sum_{i=1}^{N}\left(x_i-\mu_X\right)\right)^2\right\} = \tag{3.5}$$

$$\frac{1}{N^2}\mathrm{E}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_j-\mu_X\right)\left(x_i-\mu_X\right)\right\} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left\{\left(x_j-\mu_X\right)\left(x_i-\mu_X\right)\right\} =$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma^2_{X_i,X_j} = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2_{X_i} = \frac{N\sigma^2_X}{N^2} = \frac{\sigma^2_X}{N}$$

A Propriedade 3.2 (Equação (3.5)) é extremamente importante porque ela mostra de forma inequívoca que a variância da média amostral é inversamente proporcional ao tamanho da amostra considerada. Logo, quanto maior o tamanho N da amostra a partir da qual foi obtido o valor da média amostral, menor o nível de incerteza desse valor. Assim, a grande utilidade do cálculo do valor amostral médio é a redução do conteúdo de incerteza sobre o valor da média real μ_X. (Observe que o Exemplo 2.13 ilustra bem esse efeito de redução da incerteza com o aumento de N.) É possível inclusive planejar o tamanho da amostra para que se tenha um nível especificado de flutuação no valor da média amostral, se uma avaliação da variância experimental de uma única medida é conhecida. No entanto, o conteúdo de incerteza só vai para zero no limite em que N vai a infinito, o que é impossível do ponto de vista prático. Dessa forma, sempre haverá algum conteúdo de incerteza sobre o valor real de μ_X.

Exemplo 3.7 - Suponha que a cada medida x_i, i=1,...,N, de uma mesma população é associado o peso w_i, i=1,...,N. Suponha ainda que

$$\overline{X} = \sum_{i=1}^{N} w_i x_i$$
$$0 < w_i < 1$$
$$\sum_{i=1}^{N} w_i = 1$$

Nesse caso, a Propriedade 3.1 pode ser escrita na forma:

$$\mathrm{E}\left\{\overline{X}\right\} = E\left\{\sum_{i=1}^{N} w_i x_i\right\} = \sum_{i=1}^{N} w_i \mathrm{E}\left\{x_i\right\} = \sum_{i=1}^{N} w_i \mu_X = \mu_X \sum_{i=1}^{N} w_i = \mu_X$$

enquanto a Propriedade 3.2 pode ser escrita como

$$\mathrm{Var}\left\{\overline{X}\right\} = \mathrm{E}\left\{\left(\overline{X}-\mu_X\right)^2\right\} = E\left\{\left(\sum_{i=1}^{N} w_i x_i - \mu_X\right)^2\right\} = \mathrm{E}\left\{\left(\sum_{i=1}^{N} w_i\left(x_i-\mu_X\right)\right)^2\right\} =$$

$$\mathrm{E}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j\left(x_j-\mu_X\right)\left(x_i-\mu_X\right)\right\} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathrm{E}\left\{\left(x_j-\mu_X\right)\left(x_i-\mu_X\right)\right\} =$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma^2_{X_i,X_j} = \sum_{i=1}^{N} w_i^2 \sigma^2_{X_i} = \sigma^2_X \sum_{i=1}^{N} w_i^2 < \sigma^2_X$$

de maneira que qualquer média ponderada dos dados amostrados flutua em torno do valor médio μ_X com variância inferior à dos dados amostrados. Isso mostra que há um certo grau de arbitrariedade na definição da média amostral da Equação (3.3), já que qualquer média ponderada dos números amostrados também satisfaz as Propriedades 3.1 e 3.2 definidas anteriormente. Por isso, retornaremos a esse problema no Capítulo 4, para aumentar um pouco mais a significação teórica da Equação (3.3).

A mesma discussão apresentada para a média amostral pode ser agora estendida para a medida amostral da variância. Nesse caso, as Equações (1.36) e (1.72), reproduzidas abaixo

$$\sigma^2_{XX} = \mathrm{Var}\left\{\mathrm{x}\right\} = \mathrm{E}\left\{\left(x_i-\mu_X\right)^2\right\} = \sum_{i=1}^{NR} p_i\left(x_i-\mu_X\right)^2 \tag{1.36}$$

$$\sigma^2_{XX} = \int_{x_{min}}^{x_{max}} \left(x-\mu_X\right)^2 \wp\left(x\right) dx \tag{1.72}$$

e a hipótese do experimento bem feito sugerem a seguinte definição para a **variância amostral**, s^2_X

$$s^2_X = \sum_{i=1}^{N} p_i\left(x_i-\overline{X}\right)^2 = \sum_{i=1}^{N}\left(\frac{1}{N}\right)\left(x_i-\overline{X}\right)^2 = \frac{\sum_{i=1}^{N}\left(x_i-\overline{X}\right)^2}{N} \tag{3.6}$$

No entanto, antes que a Equação (3.6) seja aceita como medida adequada da variância amostral (o que de fato ela não é, como será mostrado ao longo desta seção), é conveniente observar o Exemplo 3.8.

Exemplo 3.8 - Nas Tabelas 3.4 e 3.5 apresentam-se as variâncias amostrais calculadas a partir da Equação (3.6) para os problemas analisados nos Exemplos 3.4 e 3.5.

Tabela 3.4 - Variâncias amostrais obtidas no Exemplo 3.4.

N	10		20		40		80		160		∞
Semente	1	2	1	2	1	2	1	2	1	2	-
s^2_X	0.250	0.240	0.240	0.248	0.250	0.244	0.249	0.246	0.249	0.250	0.250

Tabela 3.5 - Variâncias amostrais obtidas no Exemplo 3.5.

N	10		20		40		80		160		∞
Semente	1	2	1	2	1	2	1	2	1	2	-
s_X^2	0.137	0.094	0.107	0.084	0.098	0.078	0.083	0.082	0.083	0.083	0.083

Observe que a variância amostral flutua de experimento para experimento em torno de valores próximos das variâncias verdadeiras, iguais a 0.250 no primeiro caso e 0.083 no segundo caso. A variância amostral, portanto, não deve ser confundida com a variância real da distribuição de probabilidades amostrada, que o analista a princípio desconhece.

Assim como no caso da média amostral, o Exemplo 3.8 mostra claramente que a variância amostral s_X^2 flutua e, por isso, não deve ser confundida com a variância verdadeira σ_X^2 da distribuição. Mais ainda, se a variância amostral flutua de experimento para experimento (nesse caso o experimento consiste em tomar amostras de tamanho *N*), ela é também uma variável aleatória, assim como os dados amostrados x_i. Portanto, a variância amostral também deve ser encarada como uma variável aleatória que flutua em torno de certo valor médio e com certa variância, que devem a princípio ser caracterizados, assim como a distribuição de probabilidades que descreve as flutuações de s_X^2. Como no caso da média amostral, não devemos ter esperanças de obter o valor real da variância σ_X^2, a não ser que tenhamos a distribuição real de probabilidades do problema, o que de maneira geral não é possível, como já discutido. Dessa forma, se tivermos que obter informações sobre o problema a partir da experimentação (amostrando), nunca saberemos qual é o valor verdadeiro da variância σ_X^2. No entanto, como no caso anterior e mostrado a seguir, é possível escrever um conjunto de propriedades bastante úteis para a variância amostral.

Propriedade 3.3 - Se os experimentos x_i, i=1...N, são realizados de forma independente em condições idênticas e flutuam em torno da média verdadeira μ_X com variância σ_X^2, a Equação (3.6) **NÃO** fornece uma avaliação consistente da variância amostral, sendo necessário reescrever a Equação (3.6) na forma:

$$s_X^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N-1} \tag{3.7}$$

A variância amostral definida pela Equação (3.7) flutua em torno do valor real da variância σ_X^2.

Para mostrar a Propriedade 3.3, é conveniente primeiramente abrir a Equação (3.7) em termos dos desvios em relação à média verdadeira, em geral desconhecida. Assim,

$$s_X^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N} = \frac{\sum_{i=1}^{N}\left(x_i - \sum_{j=1}^{N}\frac{x_j}{N}\right)^2}{N} = \frac{\sum_{i=1}^{N}\left(Nx_i - \sum_{j=1}^{N}x_j\right)^2}{N^3} =$$

$$\frac{\sum_{i=1}^{N}\left(N\left(x_i - \mu_X\right) - \sum_{j=1}^{N}\left(x_j - \mu_X\right)\right)^2}{N^3} =$$

$$\frac{\sum_{i=1}^{N}\left(N^2\left(x_i - \mu_X\right)^2 - 2N\left(x_i - \mu_X\right)\sum_{j=1}^{N}\left(x_j - \mu_X\right) + \left[\sum_{j=1}^{N}\left(x_j - \mu_X\right)\right]^2\right)}{N^3} = \tag{3.8}$$

$$\frac{\sum_{i=1}^{N}\left(x_i - \mu_X\right)^2}{N} - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i - \mu_X\right)\left(x_j - \mu_X\right) + \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(x_j - \mu_X\right)\left(x_k - \mu_X\right) =$$

$$\frac{\sum_{i=1}^{N}\left(x_i - \mu_X\right)^2}{N} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i - \mu_X\right)\left(x_j - \mu_X\right)$$

Agora, o valor médio da Equação (3.8) pode ser calculado como

$$\mathrm{E}\left\{s_X^2\right\} = \frac{\sum_{i=1}^{N}\mathrm{E}\left\{\left(x_i - \mu_X\right)^2\right\}}{N} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left\{\left(x_i - \mu_X\right)\left(x_j - \mu_X\right)\right\} =$$

$$\frac{\sum_{i=1}^{N}\sigma_{X_i}^2}{N} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{X_i,X_j}^2 = \frac{N\sigma_X^2}{N} - \frac{1}{N^2}\sum_{i=1}^{N}\sigma_{X_i}^2 = \frac{N\sigma_X^2}{N} - \frac{N\sigma_X^2}{N^2} = \frac{(N-1)}{N}\sigma_X^2 \tag{3.9}$$

Repare que a Equação (3.9) mostra que, na média, a Equação (3.6) leva um valor de variância amostral menor que o valor da variância real do problema. Esse é um defeito inaceitável do procedimento de inferência do valor real da variância. Para corrigir o resultado, no entanto, o procedimento a seguir é muito fácil: basta multiplicarmos o resultado obtido por N e dividirmos o resultado por (N-1), o que resulta na Equação (3.7) e na Propriedade 3.3. Diz-se, portanto, que a variância amostral definida na Equação (3.7) é uma avaliação **consistente** da variância real do problema. Deve ficar bem claro que a necessidade de apresentar o valor (N-1) no denominador da Equação (3.7) nada tem de arbitrário - muito pelo contrário. É exatamente essa correção que permite obter, na média, uma inferência consistente da variância real do problema a partir dos dados amostrados. O valor (N-1) é chamado de **número de graus de liberdade** do problema, representado usualmente por ν. Como no caso da média amostral, o fato da Equação (3.7) fornecer uma medida consistente da variância não significa que a variância amostral obtida em um problema particular é igual à variância verdadeira e desconhecida do problema. Para que isso fosse verdade, seria necessário obter a média a partir de infinitas repetições do problema físico investigado, o que não é possível. Portanto, nunca saberemos de fato qual é o valor real da variância do problema a partir de dados amostrados. No entanto, a Equação (3.9) oferece ao menos o consolo

de que o valor obtido para a variância amostral a partir da Equação (3.7) flutua ao redor do valor verdadeiro da variância.

Propriedade 3.4 - Se os experimentos x_i, i=1...N, são realizados de forma independente em condições idênticas e flutuam em torno da média verdadeira μ_X com variância σ_X^2, então a variância amostral descrita pela Equação (3.7) flutua em torno de σ_X^2 com variância igual a:

$$\mathrm{Var}\left\{s_X^2\right\} = \mathrm{E}\left\{\left(s_X^2 - \sigma_X^2\right)^2\right\} = \frac{2\sigma_X^4}{N-1}\left[1 + \frac{N-1}{2\,N}\left(k_X^4 - 3\right)\right] \tag{3.10}$$

onde k_X é a kurtose, definida na Equação (1.57).

A Equação (3.10) pode ser mostrada com facilidade substituindo-se a Equação (3.8) no lado esquerdo da Equação (3.10) e efetuando-se as operações necessárias. Essa demonstração fica deixada como exercício para o leitor interessado por causa do excessivo número de manipulações algébricas necessárias. Contudo, a Equação (3.10) é muito importante porque ela indica de forma inequívoca que o nível de flutuação da variância amostral cai continuamente, à medida que aumenta o tamanho do conjunto de dados amostrados, convergindo para zero quando N vai a infinito. Dessa maneira, quanto maior o tamanho do conjunto amostral, maior a precisão com que se obtém o valor da variância amostral. Para o caso muito específico em que os dados amostrados seguem uma distribuição normal, então $k_X^4 = 3$ (*Esse é um resultado clássico para a curva normal. Lembre-se que a curva normal é uma curva bi-paramétrica, de maneira que, fixados média e variância, todos os demais momentos da curva de distribuição ficam também automaticamente fixados.*) e a Equação (3.11) ganha a forma mais simples

$$\mathrm{Var}\left\{s_X^2\right\} = \mathrm{E}\left\{\left(s_X^2 - \sigma_X^2\right)^2\right\} = \frac{2\sigma_X^4}{N-1} \tag{3.11}$$

Observe que as Equações (3.7) e (3.10-11) mostram que é impossível fazer qualquer inferência sobre a variância real de um problema se apenas um dado é medido (N-1 = ν = 0). Esse resultado é obviamente pertinente, pois não é possível ter mesmo qualquer noção de espalhamento dos dados se apenas um dado experimental está disponível.

A Equação (3.7) pode ser então utilizada automaticamente para descrever o **desvio padrão amostral**,

$$s_X = \sqrt{s_X^2} \tag{3.12}$$

a **covariância amostral**,

$$s_{XY}^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{N-1} \tag{3.13}$$

e o **coeficiente de correlação amostral**,

$$r_{XY} = \frac{s_{XY}^2}{s_X s_Y} \tag{3.14}$$

De forma similar à mostrada nos casos anteriores, as Equações (3.12-14) definem formas consistentes de avaliar as grandezas de interesse para a análise a partir de dados amostrados. Também de forma similar, essas grandezas amostrais devem ser encaradas como variáveis estocásticas, sujeitas a flutuações que convergem para zero quando o tamanho do conjunto de dados amostrados vai para infinito.

Exemplo 3.9 - A covariância amostral, definida pela Equação (3.13), pode ser colocada na forma

$$s_{XY}^2 = \frac{\sum_{i=1}^{N}\left[N\left(x_i-\mu_X\right)-\sum_{j=1}^{N}\left(x_j-\mu_X\right)\right]\left[N\left(y_i-\mu_Y\right)-\sum_{j=1}^{N}\left(y_j-\mu_Y\right)\right]}{N^2\left(N-1\right)}$$

e

$$s_{XY}^2 = \frac{\sum_{i=1}^{N}\left[N\left(x_i-\mu_X\right)-\sum_{j=1}^{N}\left(x_j-\mu_X\right)\right]\left[N\left(y_i-\mu_Y\right)-\sum_{j=1}^{N}\left(y_j-\mu_Y\right)\right]}{N^2\left(N-1\right)} =$$

$$\frac{N^2\sum_{i=1}^{N}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)}{N^2\left(N-1\right)} - 2\frac{N\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i-\mu_X\right)\left(y_j-\mu_Y\right)}{N^2\left(N-1\right)} +$$

$$\frac{\sum_{k=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i-\mu_X\right)\left(y_j-\mu_Y\right)}{N^2\left(N-1\right)} =$$

$$\frac{N^2\sum_{i=1}^{N}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)}{N^2\left(N-1\right)} - \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i-\mu_X\right)\left(y_j-\mu_Y\right)}{N^2\left(N-1\right)}$$

Aplicando o operador de média e admitindo que as medidas x_i e y_i obtidas de um mesmo experimento podem estar correlacionadas entre si, mas não com medidas de experimentos distintos, então

$$\mathrm{E}\left\{s_{XY}^2\right\} = \frac{N^2\sum_{i=1}^{N}\mathrm{E}\left\{\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)\right\}}{N^2\left(N-1\right)} - \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{E}\left\{\left(x_i-\mu_X\right)\left(y_j-\mu_Y\right)\right\}}{N^2\left(N-1\right)} =$$

$$\frac{N^2\sum_{i=1}^{N}\sigma_{XY}^2}{N^2\left(N-1\right)} - \frac{N\sum_{i=1}^{N}\sigma_{XY}^2}{N^2\left(N-1\right)} = \sigma_{XY}^2$$

que mostra que a Equação (3.13) de fato permite uma inferência consistente da covariância entre dois conjuntos de dados.

3.3. Distribuições e Intervalos de Confiança de Grandezas Amostrais

Como as grandezas amostrais devem ser encaradas como variáveis aleatórias e sujeitas a flutuações, cuja variância depende do tamanho N do conjunto amostrado, torna-se pertinente perguntar sobre a forma da curva de distribuição que governa as flutuações das grandezas amostrais. De maneira geral, essa pergunta pode ser respondida através do procedimento ilustrado abaixo para uma função genérica dos pontos amostrais.

Seja uma função genérica dos pontos amostrais definida como $f(x_1,...,x_N)$. Suponha que é possível explicitar a dependência inversa do valor de x_N, para que o valor de $f(x_1,...,x_N)$ atinja um valor especificado f_1 na forma $x_N = g(x_1,...,x_N\text{-}1,f_1)$. Então a seguinte igualdade pode ser escrita

$$\int_{f_1}^{f_2} \wp_f(f)\,df = \int_{x_1} \wp(x_1) \ldots \int_{x_{N-1}} \wp(x_{N-1}) \left[\int_{g(x_1,\ldots,x_{N-1},f_1)}^{g(x_1,\ldots,x_{N-1},f_2)} \wp(x_N)\,dx_N \right] dx_{N-1} \ldots dx_1 \qquad (3.15)$$

onde são feitas (N-1) integrações sobre as (N-1) variáveis que podem flutuar independentemente para gerar os valores especificados da função f e uma integração sobre o valor de x_N, que especifica de fato os valores desejados de f. Se f_1 é o valor mínimo admissível para a função $f(x_1,...,x_N)$, então a Equação (3.15) pode ser rescrita como

$$P_{AC}(f_2) = \int_{x_1} \wp(x_1) \ldots \int_{x_{N-1}} \wp(x_{N-1}) \left[\int_{g(x_1,\ldots,x_{N-1},f_1)}^{g(x_1,\ldots,x_{N-1},f_2)} \wp(x_N)\,dx_N \right] dx_{N-1} \ldots dx_1 \qquad (3.16)$$

cuja derivação gera a curva de densidade de probabilidades $\wp_f$ de $f(x_1,...,x_N)$.

Para ilustrar de forma mais clara o uso das Equações (3.15-16), suponha que se deseja conhecer a função densidade de probabilidades da média entre dois pontos, obtidos segundo uma distribuição de probabilidades arbitrária $\wp(x)$. Nesse caso, deseja-se conhecer a função distribuição de probabilidades da seguinte transformação

$$f(x_1, x_2) = \overline{X} = \frac{x_1 + x_2}{2}$$

que resulta na transformação inversa

$$g\left(x_1, \overline{X}\right) = x_2 = 2\overline{X} - x_1$$

Obviamente, o valor mínimo de $\overline{X}$ é o valor mínimo de x_i, de maneira que

$$P_{AC}\left(\overline{X}\right) = \int_{x_{mín}}^{x_{máx}} \wp\left(x_1\right)\left[\int_{2x_{mín}-x_1}^{2\overline{X}-x_1} \wp\left(x_2\right)dx_2\right]dx_1$$

Procedimentos semelhantes podem ser gerados para as demais variáveis amostrais. Dessa forma, o importante é perceber que a densidade de probabilidades de uma grandeza calculada a partir de variáveis aleatórias (e, portanto, essa grandeza também é a princípio uma variável aleatória) pode ser obtida a partir de procedimentos matemáticos bem definidos. Isso não significa dizer que soluções analíticas estão sempre disponíveis, dado que as transformações matemáticas são complexas e muitas vezes intratáveis analiticamente.

Exemplo 3.10 - Para a distribuição uniforme no intervalo (0,1), mostram-se a seguir as funções de densidade de probabilidade para a média e a variância amostrais obtidas a partir de dois pontos. Para a média amostral

$$P_{AC}\left(\overline{X}\right) = \int_0^1 \left[\int_0^{2\overline{X}-x_1} dx_2\right]dx_1$$

É preciso lembrar que a distribuição uniforme é igual a zero fora do intervalo (0,1), de maneira que as seguintes relações de desigualdade precisam ser satisfeitas:

$$0 < x_1 < 1,$$
$$0 < 2\overline{X} - x_1 < 1$$

ou

$$0 < x_1 < 1,$$
$$2\overline{X} - 1 < x_1 < 2\overline{X}$$

Mas só é possível satisfazer ambas as desigualdades se

$$0 < x_1 < 2\overline{X} \quad \text{se } \overline{X} < 0.5$$
$$2\overline{X} - 1 < x_1 < 1 \quad \text{se } \overline{X} > 0.5$$

Portanto, para o caso da média amostral, resulta que

$$P_{AC}\left(\overline{X}\right) = \int_0^{2\overline{X}}\left[\int_0^{2\overline{X}-x_1} dx_2\right]dx_1 = \int_0^{2\overline{X}}\left[2\overline{X} - x_1\right]dx_1 = 2\overline{X}^2 \quad \text{se } \overline{X} < 0.5$$

$$P_{AC}\left(\overline{X}\right) = \int_0^{2\overline{X}-1}\left[\int_0^1 dx_2\right]dx_1 + \int_{2\overline{X}-1}^1\left[\int_0^{2\overline{X}-x_1} dx_2\right]dx_1 = \qquad \text{se } \overline{X} > 0.5$$
$$= \int_0^{2\overline{X}-1} dx_1 + \int_{2\overline{X}-1}^1 \left[2\overline{X} - x_1\right]dx_1 = 4\overline{X} - 2\overline{X}^2 - 1$$

e portanto

$$\wp\left(\overline{X}\right) = 4\overline{X} \qquad se\ \overline{X} < 0.5$$
$$\wp\left(\overline{X}\right) = 4 - 4\overline{X} \quad se\ \overline{X} > 0.5$$

que é a distribuição triangular do Exemplo 1.13. Logo, a distribuição triangular do Exemplo 1.13 pode ser interpretada como a distribuição da média de dois pontos obtidos a partir da distribuição uniforme. Observe que a distribuição triangular concentra os valores da média amostral ao redor de 0.5 mesmo quando as medidas isoladas estão uniformemente distribuídas no intervalo [0,1], como descrito pela Propriedade 3.2.

No caso da variância amostral, é conveniente ver primeiramente que o valor mínimo admissível para a variável é igual a zero, obtido quando os dois pontos amostrados são iguais. Além disso,

$$s_X^2 = \frac{\left[x_1 - \left(\frac{x_1+x_2}{2}\right)\right]^2 + \left[x_2 - \left(\frac{x_1+x_2}{2}\right)\right]^2}{1} = \frac{\left[\frac{x_1-x_2}{2}\right]^2 + \left[\frac{x_2-x_1}{2}\right]^2}{1} = 2\left[\frac{x_1-x_2}{2}\right]^2$$

de tal maneira que, para qualquer valor especificado de s_X^2, valores menores que esses são encontrados no intervalo

$$x_1 - \sqrt{2s_X^2} < x_2 < x_1 + \sqrt{2s_X^2}$$

Dessa forma, a Equação (3.16) pode ser escrita como

$$P_{AC}\left(s_X^2\right) = \int_0^1\left[\int_{x_1-\sqrt{2s_X^2}}^{x_1+\sqrt{2s_X^2}} dx_2\right]dx_1\,,\ s_X^2 < 0.5$$

Como no problema anterior, é necessário garantir que

$$0 < x_1 < 1\,,$$
$$x_1 - \sqrt{2s_X^2} > 0\,,$$
$$x_1 + \sqrt{2s_X^2} < 1$$

ou

$$0 < x_1 < 1,$$
$$x_1 > \sqrt{2s_X^2},$$
$$x_1 < 1 - \sqrt{2s_X^2}$$

que só podem ser satisfeitas se

$$\sqrt{2s_X^2} < x_1 < 1 - \sqrt{2s_X^2},\ s_X^2 < 0.5$$

Para que a desigualdade acima seja satisfeita, é necessário que

$$\sqrt{2s_X^2} < 1 - \sqrt{2s_X^2},\ s_X^2 < 0.125$$

Portanto

$$P_{AC}\left(s_X^2\right) = \int_0^{\sqrt{2s_X^2}} \left[\int_0^{x_1+\sqrt{2s_X^2}} dx_2 \right] dx_1 + \int_{\sqrt{2s_X^2}}^{1-\sqrt{2s_X^2}} \left[\int_{x_1-\sqrt{2s_X^2}}^{x_1+\sqrt{2s_X^2}} dx_2 \right] dx_1 + \int_{1-\sqrt{2s_X^2}}^{1} \left[\int_{x_1-\sqrt{2s_X^2}}^{1} dx_2 \right] dx_1 \quad , \ s_X^2 < 0.125$$

$$P_{AC}\left(s_X^2\right) = \int_0^{1-\sqrt{2s_X^2}} \left[\int_0^{x_1+\sqrt{2s_X^2}} dx_2 \right] dx_1 + \int_{1-\sqrt{2s_X^2}}^{\sqrt{2s_X^2}} \left[\int_0^{1} dx_2 \right] dx_1 + \int_{\sqrt{2s_X^2}}^{1} \left[\int_{x_1-\sqrt{2s_X^2}}^{1} dx_2 \right] dx_1 \quad , \ 0.125 < s_X^2 < 0.5$$

resultando em

$$P_{AC}\left(s_X^2\right) = 2\left(\sqrt{2s_X^2} - s_X^2\right),\ 0 < s_X^2 < 0.5$$

e portanto

$$\wp\left(s_X^2\right) = 2\left(\frac{1}{\sqrt{2s_X^2}} - 1\right),\ 0 < s_X^2 < 0.5$$

que mostra que as variâncias amostrais pequenas são mais provavelmente obtidas que as variâncias amostrais grandes. A curva de densidade é inclusive singular no ponto s_X^2=0.

O Exemplo 3.10 mostra que, mesmo em problemas supostamente muito simples, a obtenção formal das curvas de distribuição que descrevem as flutuações de grandezas amostrais pode ser muito complexa. Isso se deve ao fato de que múltiplas combinações de resultados podem levar aos mesmos valores amostrais. Por isso, optamos nesse texto em apresentar os resultados clássicos da literatura, sem mostrar os procedimentos que tornam possível a obtenção dessas soluções. O leitor interessado pode consultar a literatura adicional apensada ao final do capítulo para informações matemáticas mais detalhadas a esse respeito.

É interessante observar, no entanto, que o computador pode auxiliar bastante a tarefa numérica de gerar as curvas de distribuição de probabilidades, uma vez fixadas a distribuição de probabilidades da variável amostrada e o tamanho *N* do conjunto de dados, como mostrado no Exemplo 2.13. Para tanto, pode-se utilizar o procedimento numérico descrito a seguir. O procedimento, normalmente chamado de **Procedimento de Monte Carlo**, consiste em gerar muitos números aleatórios (*ND* números, com *ND* da ordem de milhares) que seguem a distribuição de probabilidades estudada e computar as grandezas amostrais a partir de conjuntos contendo *N* desses dados. Dessa forma, muitos valores são obtidos para as grandezas amostrais a partir de *N* dados que seguem a distribuição considerada. Obtém-se assim uma amostra fidedigna da distribuição das grandezas amostrais. As curvas de probabilidade acumulada podem então ser obtidas, como mostrado nos Exemplos 2.13 e 3.5. Esse tipo de procedimento numérico pode ser executado com facilidade em computadores pessoais para quaisquer distribuições de probabilidades e para qualquer tamanho amostral considerado, como ilustrado a seguir no Exemplo 3.11.

Algoritmo 3.1 - Geração de curvas de distribuição de grandezas amostrais.

Fixados *N*, tamanho da amostra, e *ND*, número de dados amostrais
1- Gerar *N* dados com distribuição uniforme (ver Seção 2.4);
2- Transformar os *N* dados para a distribuição desejada (ver Equações 2.24-25);
3- Calcular a grandeza amostral desejada (ver Seções 3.1-3.3);
4- Repetir o procedimento até que sejam gerados *ND* valores amostrais;
5- Construir o histograma (ver Exemplo 2.13) ou a curva de probabilidades acumuladas (ver Exemplo 3.5) e, a partir delas, obter as curvas de densidade de probabilidade.

Exemplo 3.11 - Para o cômputo das médias e variâncias amostrais a partir de dois pontos aleatórios distribuídos uniformemente no intervalo (0,1), como mostrado no Exemplo 3.10, é possível calcular os intervalos de confiança na forma:

Confiança de 95%:

$$P_{AC}\left(\overline{X}_1\right) = 2\overline{X}_1^2 = 0.025 \;\Rightarrow\; \overline{X}_1 = 0.1119$$

$$P_{AC}\left(\overline{X}_2\right) = 4\overline{X}_2 - 2\overline{X}_2^2 = 0.975 \;\Rightarrow\; \overline{X}_1 = 0.8881$$

$$P_{AC}\left(s_{X1}^2\right) = 2\left(\sqrt{2s_{X1}^2} - s_{X1}^2\right) = 0.025 \;\Rightarrow\; s_{X1}^2 = 7.91x10^{-5}$$

$$P_{AC}\left(s_{X2}^2\right) = 2\left(\sqrt{2s_{X2}^2} - s_{X2}^2\right) = 0.975 \;\Rightarrow\; s_{X2}^2 = 0.354$$

O Algoritmo 3.1 é usado nesse exemplo para gerar a distribuição desejada numericamente. A função de distribuição uniforme foi gerada usando-se o procedimento

$$X_{k+1} = 11\,X_k - \mathrm{Trunc}\left(11\,X_k\right)$$

com semente X_1=0.75832446 (ver Seção 2.4). Fez-se *ND* igual a 2000 e *N*=2. Os resultados obtidos e ordenados em ordem crescente são apresentados nas Figuras 3.5 e 3.6. Os limites apresentados separam os menores 2.5% (50 menores valores) e os maiores 2.5% (50 maiores valores) valores calculados, de maneira que entre eles encontram-se 95% dos valores obtidos.

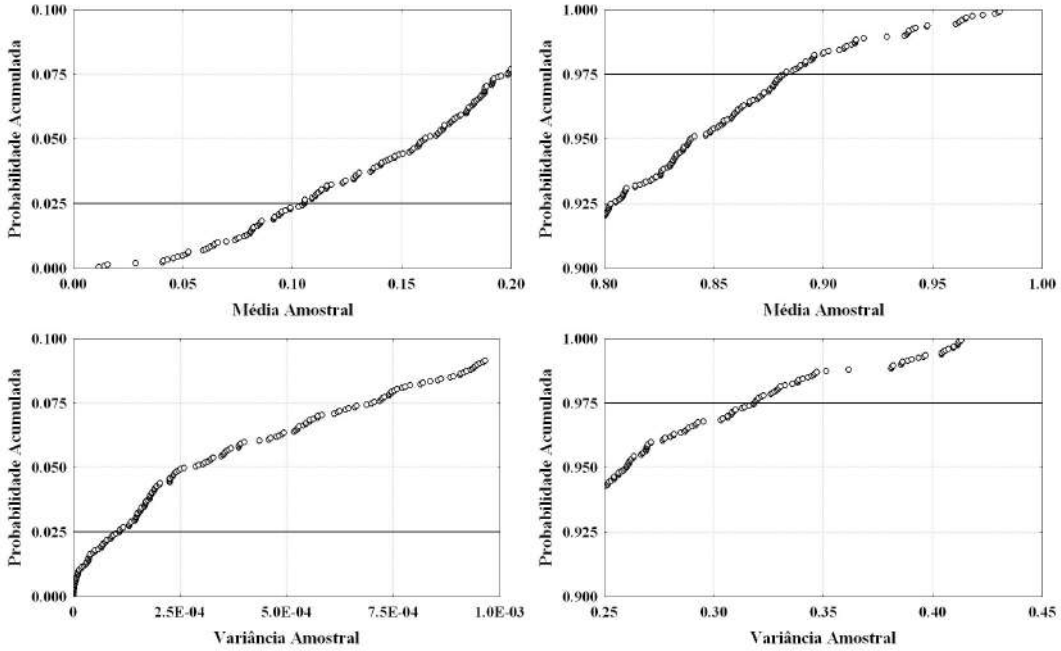


Figura 3.5 - Limites de confiança da média e variância amostrais obtidos numericamente.

Vê-se que os resultados podem ser considerados muito bons, se comparados aos valores calculados de forma exata. Os limites de confiança obtidos para a média amostral são aproximadamente iguais a 0.11 e 0.88, enquanto os limites de confiança obtidos para a variância amostral são aproximadamente iguais a $1.2x10^{-4}$ e 0.32. Vê-se, contudo, que ainda há razoável grau de incerteza nos valores dos limites de confiança, a despeito do número elevado de pontos experimentais considerados. Observa-se uma vez mais que o número de dados necessários para a adequada representação de curvas de distribuição de probabilidades pode ser muito elevado. Apesar disso, quando toda a faixa de valores admissíveis é considerada, observa-se concordância bastante boa entre as curvas geradas numérica e teoricamente.

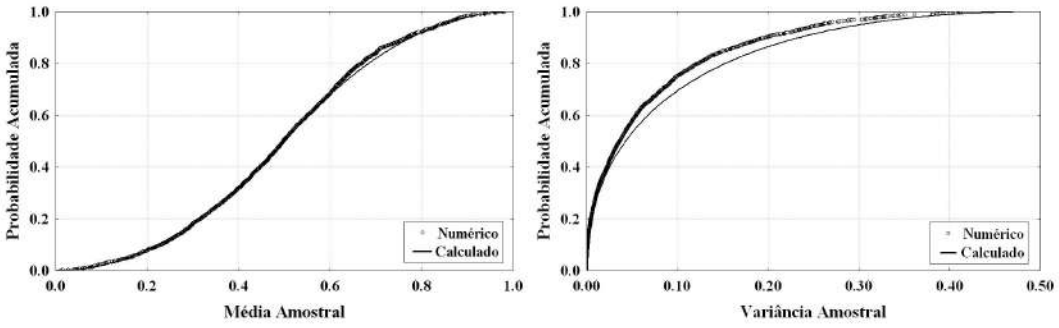


Figura 3.6 - Probabilidades acumuladas das médias e variâncias amostrais em toda a faixa de valores admissíveis.

Apesar dos resultados anteriores terem ilustrado a dificuldade de gerar teoricamente as curvas de distribuição de probabilidades de grandezas amostrais, alguns resultados clássicos são disponíveis para o caso em que as medidas experimentais estão sujeitas a flutuações normais.

3.3.1. A Distribuição t de Student

Seja x uma variável aleatória sujeita a flutuações normais, com média μ_X e variância σ_X^2. Sejam N o número de amostragens independentes de x feitas e $\overline{X}$ e s_X^2 as média e variância amostrais obtidas. Pode-se mostrar que a variável normalizada t, definida como:

$$t = \frac{\overline{X} - \mu_X}{\frac{s_X}{\sqrt{N}}} \tag{3.17}$$

está distribuída na forma

$$\wp(t) = \mathrm{Stud}(t;\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)} \tag{3.18}$$

onde ν é o número de graus de liberdade e Γ representa a função gama, definida pela Equação (2.46). A forma da distribuição t de Student (publicada originalmente por W.S. Gosset, sob o codinome de Student, donde vem o nome normalmente usado para referenciar essa importante distribuição estatística) está mostrada na Figura 3.7, enquanto valores para as probabilidades acumuladas são apresentados na Tabela A.2 do Apêndice.

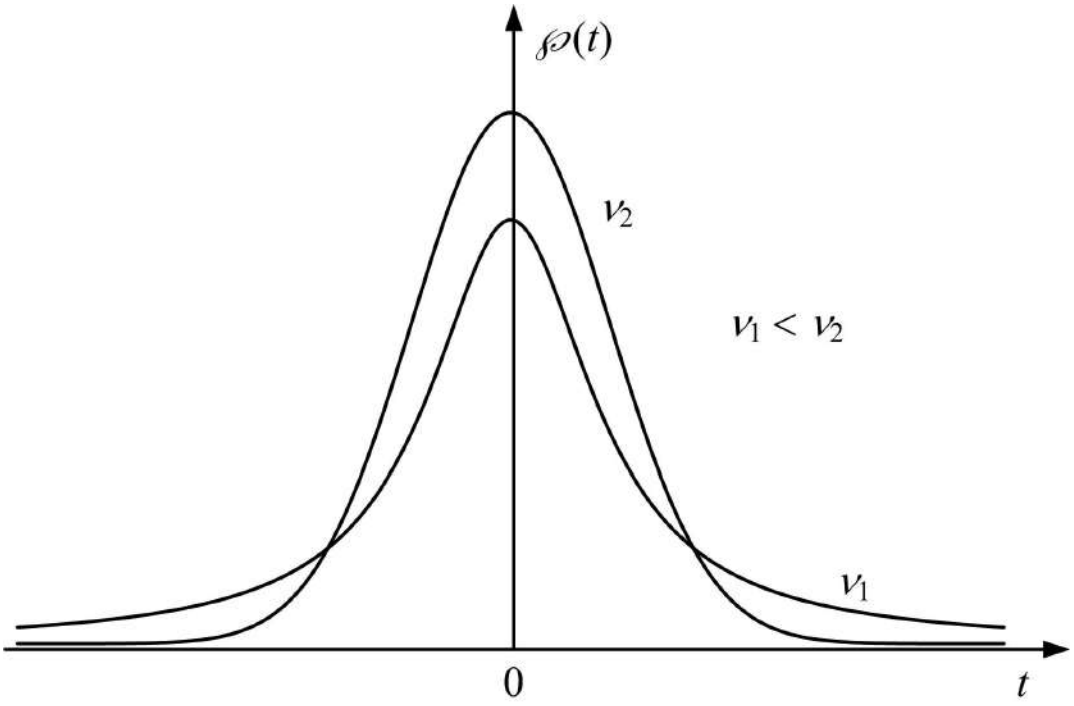


Figura 3.7 - Ilustração da distribuição t.

A Figura 3.7 mostra que a distribuição *t* é simétrica em relação ao eixo *y* de coordenadas e é definida sobre todo o domínio real (-∞,+∞). Além disso, a distribuição *t* depende de um único parâmetro, ν, que representa o tamanho do conjunto amostral. Quanto maior o valor de ν, mais estreita é a distribuição em torno do valor médio t=0, em função das menores incertezas existentes sobre o valor real da média quando *N* aumenta. A distribuição *t* tem enorme importância prática porque permite impor limites precisos sobre a região de confiança onde deve estar a média verdadeira, a partir de valores amostrados, como mostrado nos exemplos que se seguem.

Exemplo 3.12 - Admita que testes de atividade catalítica foram realizados em condições supostamente idênticas, resultando no seguinte conjunto de dados:

Tabela 3.6 - Dados de atividade catalítica obtidos experimentalmente.

i	1	2	3	4	5	6	7	8	9	10
x_i **(g/h g)**	0.450	0.467	0.431	0.440	0.452	0.458	0.438	0.462	0.447	0.452

onde *i* designa o experimento realizado e x_i designa a atividade medida, em gramas de produto por hora por grama de reagente. Nesse caso,

$$\overline{X} = \frac{\sum_{i=1}^{10} x_i}{10} = 0.450$$

$$s_X^2 = \frac{\sum_{i=1}^{10} (x_i - 0.450)^2}{9} = 93.2 \cdot 10^{-6}$$

$$s_X = \sqrt{s_X^2} = 9.65 \cdot 10^{-3}$$

Sabemos, no entanto, que não devemos confundir a média e a variância amostrais com a média e a variância verdadeiras da distribuição. Para construir o intervalo de confiança da média real a partir dos valores amostrais, podemos contar com o auxílio da distribuição *t*.

Suponha que um nível de confiança de 95% é requerido. Nesse caso, deseja-se obter os valores de t_1 e t_2 tais que:

$$\mathrm{P_{AC}}\left(t_1; 9\right) = 0.025\,,\ \mathrm{P_{AC}}\left(t_2; 9\right) = 0.975$$

Esses valores podem ser obtidos da integração da Equação (3.18) e estão mostrados na Tabela A.2. Na linha referente a 9 graus de liberdade e na coluna referente a uma probabilidade acumulada de 0.975 encontra-se o valor $t_2 = 2.262$. Como a distribuição t é simétrica em relação ao eixo *y*, conclui-se que $t_1 = -2.262$. Pode-se dizer, portanto, que com 95% de confiança

$$-2.262 < t = \frac{0.450 - \mu_X}{\dfrac{9.65 \cdot 10^{-3}}{\sqrt{10}}} < 2.262$$

ou

$$0.450 - 2.262 \frac{9.65 \cdot 10^{-3}}{\sqrt{10}} < \mu_X < 0.450 + 2.262 \frac{9.65 \cdot 10^{-3}}{\sqrt{10}}$$

e

$$0.443 < \mu_X < 0.457$$

Portanto, embora não seja possível dizer qual é o valor verdadeiro da média, é possível definir o intervalo onde ela deve ser encontrada, com um certo grau de confiança, desde que os dados medidos estejam sujeitos a flutuações normais. Para os níveis de confiança de 98% e 99%, os resultados obtidos são respectivamente iguais a:

$$\mathrm{P_{AC}}\left(t_1;\ 9\right) = 0.010\,,\ \mathrm{P_{AC}}\left(t_2;\ 9\right) = 0.990$$

$$-2.821 < t = \frac{0.450 - \mu_X}{\dfrac{9.65 \cdot 10^{-3}}{\sqrt{10}}} < 2.821$$

$$0.450 - 2.821 \frac{9.65 \cdot 10^{-3}}{\sqrt{10}} < \mu_X < 0.450 + 2.821 \frac{9.65 \cdot 10^{-3}}{\sqrt{10}}$$

$$0.441 < \mu_X < 0.459$$

e

$$P_{AC}\left(t_1;\ 9\right) = 0.005\,,\ P_{AC}\left(t_2;\ 9\right) = 0.995$$

$$-3.250 < t = \frac{0.450 - \mu_X}{\frac{9.65 \cdot 10^{-3}}{\sqrt{10}}} < 3.250$$

$$0.450 - 3.250\frac{9.65 \cdot 10^{-3}}{\sqrt{10}} < \mu_X < 0.450 + 3.250\frac{9.65 \cdot 10^{-3}}{\sqrt{10}}$$

$$0.440 < \mu_X < 0.460$$

Como já discutido em exemplos anteriores, quanto maior o grau de confiança exigido, maior o intervalo de confiança obtido, tornando mais difícil o processo de tomada de decisão.

Deve ficar bem claro que o Exemplo 3.12 acima admite implicitamente que a medida experimental está distribuída de forma normal e que todas as medidas de fato representam o mesmo fenômeno. Só assim é possível usar a distribuição *t* de Student. Caso a distribuição da medida amostrada original não seja normal ou caso o conjunto de medidas represente coisas diferentes, a utilização da distribuição *t* não faz qualquer sentido. Nesse caso, outra distribuição da média amostral deveria ser gerada ou o Algoritmo 3.1 deveria ser usado, como ilustrado no Exemplo 3.11. É verdade, no entanto, que como conseqüência do Teorema do Limite Central (ver Seção 2.6), a distribuição *t* converge para a curva normal à medida que *N* aumenta, independentemente da distribuição de probabilidades que deu origem aos dados amostrados. Portanto, para *N* suficientemente grandes (*Temos visto que isso pode representar valores inconcebíveis para a prática experimental. Portanto, cuidado com essas hipóteses!*), é possível dizer que $\overline{X}$ está distribuído normalmente em torno de μ_X, com variância igual a $\sigma_{\overline{X}}^2 = s_X^2 / N$.

Exemplo 3.13 - Suponha que tenha sido admitida distribuição normal para a média amostral. Então, segundo a Tabela A.1 da curva normal, para limite de confiança de 95%, podem ser obtidos os seguintes valores:

$$P_{AC}(u_1; 9) = 0.025\,,\; P_{AC}(t_2; 9) = 0.975$$

$$-1.960 < u = \frac{0.450 - \mu_X}{\frac{9.65 \cdot 10^{-3}}{\sqrt{10}}} < 1.960$$

$$0.450 - 1.960\frac{9.65 \cdot 10^{-3}}{\sqrt{10}} < \mu_X < 0.450 + 1.960\frac{9.65 \cdot 10^{-3}}{\sqrt{10}}$$

$$0.444 < \mu_X < 0.456$$

resultando numa visão mais otimista que a real da região onde se encontra a média verdadeira. Para valores menores de *N*, como usados na prática experimental, essas

diferenças podem vir a ser muito grandes, de forma que o uso dessa aproximação raramente pode ser justificado.

Exemplo 3.14 - Suponha que o seguinte conjunto de dados, mostrado de forma ordenada na Tabela 3.7, é gerado a partir de um gerador de números uniformemente distribuídos no intervalo (0,1), como no Exemplo 3.3.

Tabela 3.7 - Conjunto de dados gerados de acordo com uma distribuição uniforme em (0,1).

i	1	2	3	4	5
x_i	0.007	0.176	0.337	0.884	0.927

Nesse caso,

$$\overline{X} = \frac{\sum_{i=1}^{10} x_i}{5} = 0.466$$

$$s_X^2 = \frac{\sum_{i=1}^{10} (x_i - 0.466)^2}{4} = 0.175$$

$$s_X = \sqrt{s_X^2} = 0.418$$

Se a região de confiança da média é calculada como no Exemplo 3.12, para um grau de confiança de 99%

$$P_{AC}(t_1; 9) = 0.005 \quad ; \quad P_{AC}(t_2; 9) = 0.995$$

$$-4.604 < t = \frac{0.466 - \mu_X}{\frac{0.418}{\sqrt{5}}} < 4.604$$

$$0.466 - 4.604\frac{0.418}{\sqrt{5}} < \mu_X < 0.466 + 4.604\frac{0.418}{\sqrt{5}}$$

$$-0.395 < \mu_X < 1.321$$

O resultado obtido acima é absurdo, pois sabemos que a média está, com 100% de confiança, contida no intervalo (0,1). Ela jamais pode ser negativa ou maior que 1, como calculado, porque os pontos estão sendo gerados com a distribuição uniforme. Onde está o erro do procedimento usado? O erro fundamental cometido foi usar a distribuição *t*, válida para valores amostrados que seguem uma distribuição normal, e não uma distribuição uniforme. Isso mostra de maneira inequívoca como as hipóteses feitas a respeito dos dados podem ser importantes para a análise. Portanto, se a função de densidade de probabilidades que gera os pontos aleatórios não é conhecida, o uso da distribuição *t* de Student para interpretar médias amostrais pode ser temerário.

3.3.2. A Distribuição Chi-Quadrado (χ^2)

Seja x uma variável aleatória sujeita a flutuações normais, com média μ_X e variância σ_X^2. Sejam N o número de amostragens independentes de x feitas e $\overline{X}$ e s_X^2 as média e variância amostrais obtidas. Pode-se mostrar que a variável normalizada χ^2, definida como:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \overline{X}}{\sigma_X} \right)^2 \tag{3.19}$$

está distribuída na forma

$$\wp\left(\chi^2\right) = \mathrm{Chi}\left(\chi^2;\nu\right) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \left(\chi^2\right)^{\left[\frac{\nu}{2}-1\right]} e^{-\left(\frac{\chi^2}{2}\right)} \tag{3.20}$$

apresentando

$$\mathrm{E}\left\{\chi^2\right\} = \nu \tag{3.21}$$

$$\mathrm{Var}\left\{\chi^2\right\} = 2\nu \tag{3.22}$$

onde ν é o número de graus de liberdade e Γ representa a função gama, definida pela Equação (2.46). A forma da distribuição χ^2 está mostrada na Figura 3.8, enquanto valores para as probabilidades acumuladas são apresentados na Tabela A.3 do Apêndice.

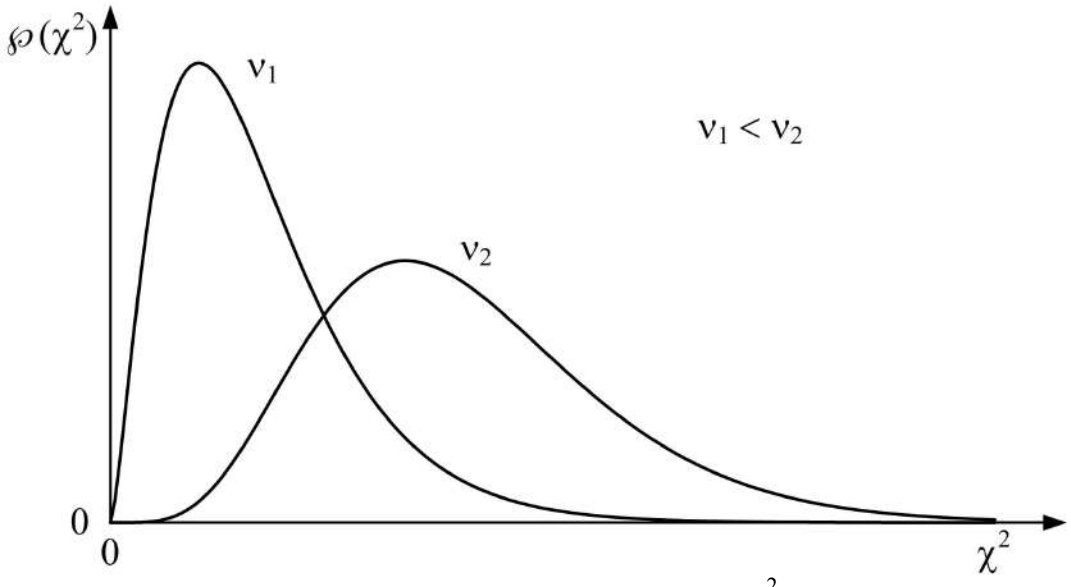


Figura 3.8 - Ilustração da distribuição χ^2.

A Figura 3.8 mostra que a distribuição χ^2 não apresenta qualquer eixo de simetria e é definida sobre o domínio real positivo [0,∞). Além disso, a distribuição χ^2 depende de um único parâmetro, ν, que representa o tamanho do conjunto amostral. Quanto maior o valor de ν, mais larga é a distribuição em torno do valor médio $\chi^2=\nu$. A distribuição χ^2 tem enorme importância prática porque, dentre muitas outras coisas, permite impor limites precisos sobre a região de confiança onde deve estar a variância verdadeira, a partir de valores amostrados, como mostrado nos exemplos a seguir. Para tanto, observe que

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i - \overline{X}}{\sigma_X}\right)^2 = \frac{(N-1)}{\sigma_X^2}\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{(N\text{-}1)} = (N-1)\frac{s_X^2}{\sigma_X^2} \tag{3.23}$$

Além disso, somas normalizadas como a apresentada na Equação (3.19) aparecem com muita freqüência em problemas práticos, como mostrado nos próximos capítulos.

Exemplo 3.15 - No Exemplo 3.12, foram analisados 10 dados de atividade de catalisador em réplicas experimentais independentes. As média e variância amostrais obtidas foram:

$$\overline{X} = \frac{\sum_{i=1}^{10} x_i}{10} = 0.450$$

$$s_X^2 = \frac{\sum_{i=1}^{10}\left(x_i - 0.450\right)^2}{9} = 93.2\cdot 10^{-6}$$

$$s_X = \sqrt{s_X^2} = 9.65\cdot 10^{-3}$$

Sabemos, no entanto, que não devemos confundir a média e a variância amostrais com a média e a variância verdadeiras da distribuição. Para construir o intervalo de confiança da variância real a partir dos valores amostrais, podemos contar com o auxílio da distribuição χ^2.

Suponha que um nível de confiança de 95% é requerido. Nesse caso, deseja-se obter os valores de χ_1^2 e χ_2^2 tais que:

$$P_{AC}\left(\chi_1^2;9\right) = 0.025,\ P_{AC}\left(\chi_2^2;9\right) = 0.975$$

Esses valores podem ser obtidos da integração da Equação (3.20) e estão mostrados na Tabela A.3. Na linha referente a 9 graus de liberdade e na coluna referente a uma probabilidade acumulada de 0.025 encontra-se o valor $\chi_1^2 = 2.700$. Na linha referente a

9 graus de liberdade e na coluna referente a uma probabilidade acumulada de 0.975 encontra-se o valor $\chi_2^2 = 19.023$. Pode-se dizer, portanto, que com 95% de confiança

$$\chi_1^2 = 2.700 < \chi^2 = (N-1)\frac{s_X^2}{\sigma_X^2} < 19.023 = \chi_2^2$$

ou

$$(N\text{-}1)\frac{s_X^2}{\chi_2^2} < \sigma_X^2 < (N\text{-}1)\frac{s_X^2}{\chi_1^2}$$

e

$$9\frac{93.2\cdot 10^{-6}}{19.023} < \sigma_X^2 < 9\frac{93.2\cdot 10^{-6}}{2.700}$$

e

$$44.1\cdot 10^{-6} < \sigma_X^2 < 311.7\cdot 10^{-6}$$

De forma similar, para graus de confiança de 98% e 99%, os resultados obtidos são respectivamente iguais a:

$$P_{AC}(\chi_1^2;9) = 0.010,\ P_{AC}(\chi_2^2;9) = 0.990$$

$$\chi_1^2 = 2.088 < \chi^2 = (N-1)\frac{s_X^2}{\sigma_X^2} < 21.666 = \chi_2^2$$

$$(N\text{-}1)\frac{s_X^2}{\chi_2^2} < \sigma_X^2 < (N\text{-}1)\frac{s_X^2}{\chi_1^2}$$

$$9\frac{93.2\cdot 10^{-6}}{21.666} < \sigma_X^2 < 9\frac{93.2\cdot 10^{-6}}{2.088}$$

$$38.7\cdot 10^{-6} < \sigma_X^2 < 401.7\cdot 10^{-6}$$

e

$$P_{AC}(\chi_1^2;9) = 0.005,\ P_{AC}(\chi_2^2;9) = 0.995$$

$$\chi_1^2 = 1.735 < \chi^2 = (N-1)\frac{s_X^2}{\sigma_X^2} < 23.589 = \chi_2^2$$

$$(N\text{-}1)\frac{s_X^2}{\chi_2^2} < \sigma_X^2 < (N\text{-}1)\frac{s_X^2}{\chi_1^2}$$

$$9\frac{93.2\cdot 10^{-6}}{23.589} < \sigma_X^2 < 9\frac{93.2\cdot 10^{-6}}{1.735}$$

$$35.6\cdot 10^{-6} < \sigma_X^2 < 483.5\cdot 10^{-6}$$

Vê-se, portanto, que as incertezas existentes durante a obtenção do valor real da variância podem ser muito grandes, quando *N* é pequeno.

Deve ficar bem claro que o Exemplo 3.15 acima admite implicitamente que a medida experimental está distribuída de forma normal e que todas as medidas de fato representam o mesmo fenômeno. Só assim é possível usar a distribuição χ^2. Caso a distribuição da medida amostrada original não seja normal ou caso o conjunto de medidas represente coisas diferentes, a utilização da distribuição χ^2 não faz qualquer sentido e resultados espúrios, como aqueles mostrado no Exemplo 3.14, podem ser obtidos.

Exemplo 3.16 - Observe no Exemplo 3.15 que o fator $(N-1)/\chi_1^2$ diz quantas vezes maior a variância real pode ser, quando comparada à variância amostral. Por isso, esse número é apresentado abaixo para alguns valores típicos.

Tabela 3.8 - Fatores que dizem quantas vezes maior que a variância amostral a variância real pode ser.

	N=1	**2**	**3**	**5**	**10**	**20**	**30**	**40**	**50**	**100**
95%	∞	1018	39.5	8.26	3.33	2.13	1.81	1.65	1.55	1.35
98%	∞	6366	99.5	13.5	4.31	2.49	2.03	1.82	1.69	1.43
99%	∞	25460	199.5	19.3	5.19	2.78	2.21	1.95	1.80	1.49

Observe na Tabela 3.8 que com cinco réplicas é possível apenas garantir a ordem de grandeza da variância verdadeira. Para garantir o primeiro algarismo significativo (incertezas inferiores a 100% do valor medido) da variância verdadeira são necessárias entre 20 e 30 réplicas! Quando o número de réplicas chega a 100, as incertezas são da ordem ainda de 35 a 50% do valor medido! Para que a incerteza seja inferior a 10% do valor medido são necessárias 900 (95%), 1250 (98%) ou 1500 (99%) réplicas, o que é inaceitável do ponto de vista do trabalho científico experimental. Por isso, teremos sempre que conviver com incertezas muito grandes em relação aos reais valores da variância experimental.

A Tabela 3.8 também mostra que as incertezas da variância real caem muito rapidamente para pequenos valores de N (por exemplo, caem cerca de duas ordens de grandeza quando N é incrementado de 2 para 3), mas depois decaem muito lentamente para valores elevados de N (por exemplo, decaem cerca de uma ordem de grandeza quando N é incrementado de 5 para 30). Por isso, raramente há justificativas para que se reproduza um dado experimental mais do que 5 vezes, uma vez que ganhos apreciáveis de certeza requereriam aumento muito grande do número de réplicas experimentais. Por isso, uma regra heurística de repetição pode ser formulada, recomendando a replicação de dados não mais do que 5 vezes, a não ser que seja muito fácil repetir o experimento.

3.3.3. A Distribuição F de Fisher

Sejam x e y variáveis aleatórias sujeitas a flutuações normais, com médias μ_X e μ_Y e variâncias σ_X^2 e σ_Y^2. Sejam N_1 e N_2 os números de amostragens independentes de x e y feitas, sendo que $\overline{X}$ e $\overline{Y}$ e s_X^2 e s_Y^2 são as médias e variâncias amostrais obtidas. Pode-se mostrar que a variável normalizada F, definida como:

$$F = \frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2} \tag{3.24}$$

está distribuída em conformidade com a seguinte função de densidade de probabilidades

$$\wp(F) = F(F;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_1}{2}\right)} \nu_1^{\left(\nu_1/2\right)} \nu_2^{\left(\nu_2/2\right)} \frac{F^{\left(\frac{\nu_1}{\nu_2}-1\right)}}{\left(\nu_1 F + \nu_2\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}} \tag{3.25}$$

com

$$\mathrm{E}\{F\} = \frac{\nu_2}{\nu_2 - 2} \tag{3.26}$$

$$\mathrm{Var}\{F\} = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2} \tag{3.27}$$

onde ν é o número de graus de liberdade e Γ representa a função gama, definida pela Equação (2.46). A forma da distribuição F está mostrada na Figura 3.9, enquanto valores para as probabilidades acumuladas são apresentados na Tabela A.4 do Apêndice.

A Figura 3.9 mostra que a distribuição F é definida sobre o domínio real positivo $[0,\infty)$. A distribuição F depende ainda de dois parâmetros, ν_1 e ν_2, que representam os tamanhos dos conjuntos amostrais analisados. Quanto maiores os valores de ν_1 e ν_2, mais estreita é a distribuição, uma vez que as variâncias amostrais tendem a se aproximar das variâncias reais. Além disso, a distribuição F apresenta a seguinte propriedade de simetria:

$$P_{AC}(F;\nu_1,\nu_2) = p\% \;\Rightarrow\; P_{AC}\left(\frac{1}{F};\nu_2,\nu_1\right) = 100 - p\% \tag{3.28}$$

que é induzida pela própria definição do valor de F. A Equação (3.28) diz que se a probabilidade de se encontrar um valor de F inferior a um certo marco é igual a p% para dois conjuntos 1 e 2, ao se inverter a definição dos conjuntos 1 e 2 os resultados devem ser qualitativamente idênticos. Como a definição dos conjuntos foi invertida, o valor do marco também tem que ser. Nesse caso, o que era maior passa a ser menor e vice-versa.

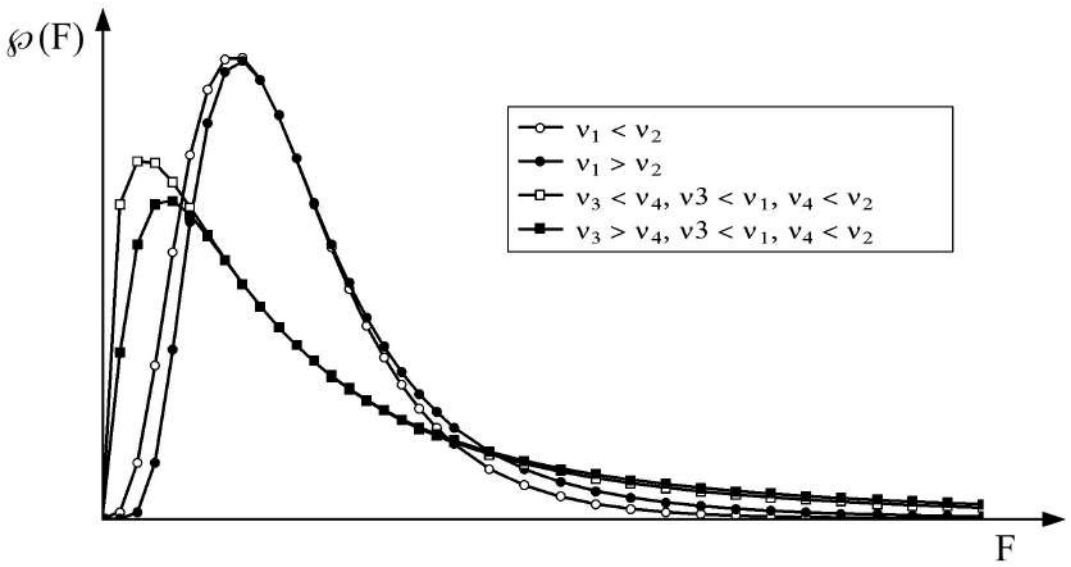


Figura 3.9 - Ilustração da distribuição F.

A distribuição F tem enorme importância prática porque permite estabelecer comparações muito mais eficientes entre diferentes variâncias amostrais que aquelas obtidas com a distribuição χ^2. Para tanto, observe que se as variâncias reais dos dois conjuntos de dados analisados são supostamente iguais, então

$$F = \frac{s_X^2}{s_Y^2} \tag{3.29}$$

que é o formato básico de F usado nos exercícios seguintes.

Exemplo 3.17 - Se dois conjuntos de dados supostamente equivalentes (variâncias reais supostamente iguais) contêm 3 e 5 dados amostrados, respectivamente, quão diferentes podem ser as variâncias obtidas?

De acordo com os resultados do Exemplo 3.16, as diferenças observadas podem ser muito grandes. Dados 2 e 4 graus de liberdade, respectivamente, e fixando o grau de confiança em 95%, procuram-se os valores de F tais que

$$P_{AC}\left(F_1; 2, 4\right) = 0.025\,,\ P_{AC}\left(F_2; 2, 4\right) = 0.975$$

Esses valores podem ser obtidos diretamente da integração da Equação (3.25) ou através da Tabela A.4. Nesse caso, como a distribuição F é biparamétrica, são apresentadas várias tabelas para valores preestabelecidos da probabilidade acumulada. Usando a Tabela montada para a probabilidade acumulada de 0.975, na coluna relativa ao grau de liberdade igual a 2 e na linha relativa ao grau de liberdade igual a 4 obtém-se o valor F_2=10.649. Não há tabela disponível para a probabilidade acumulada de 0.025. Nesse caso, usando a propriedade de simetria descrita pela Equação (3.28), na tabela de probabilidade acumulada de 0.975, na coluna relativa ao grau de liberdade igual a 4 e na

linha relativa ao grau de liberdade igual a 2 obtém-se o valor de F_1=1/39.248. Portanto, com 95% de confiança

$$\frac{1}{39.248} < F = \frac{s_X^2}{s_Y^2} < 10.649$$

quando o conjunto *x* tem três medidas amostrais e o conjunto *y* tem cinco medidas amostrais.

De forma similar, para 98% de confiança

$$P_{AC}\left(\frac{1}{F_1}; 4, 2\right) = 0.990\,,\ P_{AC}\left(F_2; 2, 4\right) = 0.990$$

$$\frac{1}{99.249} < F = \frac{s_X^2}{s_Y^2} < 18.000$$

Deve ficar bem claro que o Exemplo 3.17 acima admite implicitamente que as medidas experimentais estão distribuídas de forma normal e que todas as medidas de fato representam o mesmo fenômeno. Só assim é possível usar a distribuição F. Caso a distribuição da medida amostrada original não seja normal ou caso o conjunto de medidas represente coisas diferentes, a utilização da distribuição F pode não fazer qualquer sentido, resultando em resultados espúrios, como aquele mostrado no Exemplo 3.14.

3.4. Fazendo Comparações Entre Grandezas Amostrais

Com enorme freqüência, o analista é chamado a decidir se medidas amostrais podem ser consideradas equivalentes ou não. De forma mais específica, deseja-se saber se o valor médio real ou se a variância real do problema pode estar mudando ou pode ter mudado durante os estudos experimentais. Como veremos nos capítulos seguintes, essa questão pode exercer enorme influência sobre o tratamento dos dados e a interpretação final do conjunto de dados experimentais.

Uma forma muito simples de estabelecer essas comparações e tomar decisões está baseada na construção dos intervalos de confiança para a variável considerada. Por exemplo, sejam α e β as grandezas comparadas (por exemplo, médias ou variâncias amostrais) e sejam $\alpha_1 < \alpha < \alpha_2$ e $\beta_1 < \beta < \beta_2$ os respectivos intervalos de confiança para um grau de confiança p% especificado. Então, admitindo que $\alpha_1 < \beta_1$, as grandezas α e β são distintas com grau de confiança *p*% se $\alpha_2 < \beta_1$; ou seja, se não há interseção entre os intervalos considerados.

Exemplo 3.18 - Admita que dois estudantes diferentes obtiveram os seguintes dados de titulação no laboratório:

Tabela 3.9- Medidas de titulação obtidas por dois alunos.

	1	2	3	4	5	6	7
1- Volume (ml)	76.48	76.43	77.20	76.25	76.48	76.48	76.6
2- Volume (ml)	77.10	78.4	77.2	76.2	77.7	76.8	-

As médias e variâncias amostrais são iguais a

$$\overline{X_1} = \frac{\sum_{i=1}^{7} x_i}{7} = 76.56 \text{ e } \overline{X_2} = \frac{\sum_{i=1}^{6} x_i}{6} = 77.23$$

$$s_1^2 = \frac{\sum_{i=1}^{7} (x_i - 76.56)^2}{6} = 0.0906 \text{ e } s_2^2 = \frac{\sum_{i=1}^{6} (x_i - 76.56)^2}{5} = 0.5707$$

$$s_1 = \sqrt{s_1^2} = 0.301 \text{ e } s_2 = \sqrt{s_2^2} = 0.755$$

Os intervalos de confiança da média e variância amostrais do primeiro conjunto podem ser obtidos a partir das distribuições *t* e χ^2, como feito nas seções anteriores. Fixando o grau de confiança em 95% e levando-se em conta que ν_1=*N*-1=6, para a média

$$P_{AC}(t_1;6) = 0.025,\ P_{AC}(t_2;6) = 0.975$$

$$-2.447 < t = \frac{76.56 - \mu_1}{\frac{0.301}{\sqrt{7}}} < 2.447$$

$$76.56 - 2.447\frac{0.301}{\sqrt{7}} < \mu_1 < 76.56 + 2.447\frac{0.301}{\sqrt{7}}$$

$$76.28 < \mu_1 < 76.84$$

e para a variância

$$P_{AC}(\chi_1^2;6) = 0.025,\ P_{AC}(\chi_2^2;6) = 0.975$$

$$\chi_1^2 = 1.237 < \chi^2 = (N_1 - 1)\frac{s_1^2}{\sigma_1^2} < 14.449 = \chi_2^2$$

$$6\frac{0.0906}{14.449} < \sigma_1^2 < 6\frac{0.0906}{1.237}$$

$$0.03762 < \sigma_1^2 < 0.4394$$

Os intervalos de confiança da média e variância amostrais do segundo conjunto podem ser também obtidos a partir das distribuições *t* e χ^2. Fixando o mesmo grau de

confiança de 95% para fins de comparação e levando-se em conta que ν_1=*N*-1=5, para a média

$$P_{AC}\left(t_1;6\right)=0.025\,,\;P_{AC}\left(t_2;6\right)=0.975$$

$$-2.571<t=\frac{77.23-\mu_2}{\frac{0.755}{\sqrt{6}}}<2.571$$

$$77.23-2.571\frac{0.755}{\sqrt{6}}<\mu_2<77.23+2.571\frac{0.755}{\sqrt{6}}$$

$$76.44<\mu_2<78.03$$

e para a variância

$$P_{AC}\left(\chi_1^2;5\right)=0.025\,,\;P_{AC}\left(\chi_2^2;5\right)=0.975$$

$$\chi_1^2=0.831<\chi^2=\left(N_2-1\right)\frac{s_2^2}{\sigma_2^2}<12.833=\chi_2^2$$

$$5\frac{0.5707}{12.833}<\sigma_2^2<5\frac{0.5707}{0.831}$$

$$0.2224<\sigma_2^2<3.434$$

Comparando-se os intervalos de confiança da média, observa-se que no limite de 95% de confiança há interseção dos intervalos na faixa $76.44<\mu_1,\;\mu_2<76.84$, de maneira que não é possível dizer que as médias são diferentes. De forma similar, para as variâncias obtém-se interseção na região $0.2224<\sigma_1^2,\;\sigma_2^2<0.4394$, de maneira que não é possível dizer que as variâncias são diferentes. Logo, por esses critérios as medidas dos dois alunos poderiam ser consideradas equivalentes e, por isso, até misturadas em um único conjunto de dados.

De forma similar, aplicando o teste F para 95% de confiança

$$P_{AC}\left(\frac{1}{F_1};5,6\right)=0.975\,,\;P_{AC}\left(F_2;6,5\right)=0.975$$

$$\frac{1}{5.9876}=0.1670<F=\frac{s_1^2}{s_2^2}<6.9777$$

O valor de F obtido foi

$$F = \frac{s_1^2}{s_2^2} = \frac{0.0906}{0.5707} = 0.1587$$

que não satisfaz a desigualdade acima. Portanto, no limite de confiança de 95%, o valor de F obtido experimentalmente pode ser considerado pouco provável. Logo, é pouco provável que as variâncias reais dos dois problemas sejam iguais. Logo, com 95% de confiança, pode-se dizer que o segundo aluno lidou com mais flutuações experimentais do que o primeiro, indicando que os experimentos conduzidos pelo primeiro aluno são mais precisos.

Repare que as conclusões obtidas com os intervalos de confiança da variância e com o teste F são distintas. Isso não é incomum; muito pelo contrário. No entanto, o teste F tem capacidade muito maior de detectar diferenças de variâncias amostrais que os intervalos de confiança obtidos com a distribuição χ^2. Por isso, pode-se afirmar com 95% de certeza que os conjuntos amostrais podem ter a mesma média, mas têm variâncias distintas. Portanto, não parece haver argumentos que justifiquem a mistura dos dados, já que os dois conjuntos não parecem ter sido amostrados de uma mesma população.

Deve ficar bem claro que o Exemplo 3.18 acima admite implicitamente que as medidas experimentais estão distribuídas de forma normal e que todas as medidas de fato representam o mesmo fenômeno. Só assim seria justificável o uso das distribuições t, χ^2 e F para a análise. Caso as medidas amostradas não sejam distribuídas normalmente ou caso os conjuntos de medidas representem coisas diferentes, a utilização dessas distribuições pode não fazer qualquer sentido, resultando em resultados espúrios, como aquele mostrado no Exemplo 3.14.

As comparações feitas através dos intervalos de confiança são muito simples e podem ser executadas com facilidade. No entanto, a literatura está repleta de testes comparativos desenvolvidos para condições particulares, onde informações adicionais são conhecidas. Não é objetivo desse texto discorrer longamente sobre esse assunto e o leitor interessado pode buscar informações adicionais nas referências apensadas ao final do capítulo. No entanto, algumas dessas situações particulares são apresentadas a seguir.

3.4.1. Testes Adicionais para a Média

Condição especial 1 - Seja uma média histórica μ_X e a respectiva variância σ_X^2, obtidas com número elevado de graus de liberdade e consideradas iguais aos valores verdadeiros. Deseja-se saber se uma nova média amostral $\overline{X}$, obtida a partir de um novo conjunto de dados de tamanho *N*, é compatível com os dados passados. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Nesse caso, a variável

$$u = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{N}}} \tag{3.30}$$

é normalmente distribuída, com média zero e variância igual a 1. Logo, a curva normal pode ser usada para gerar os intervalos de confiança de $\overline{X}$ e verificar se o valor obtido é compatível com o esperado.

$$\mu_X - u_1 \frac{\sigma_X}{\sqrt{N}} < \overline{X} < \mu_X + u_2 \frac{\sigma_X}{\sqrt{N}} \tag{3.31}$$

Condição especial 2 - Seja uma média histórica μ_X, obtida com número elevado de graus de liberdade e considerada igual ao valor verdadeiro. Deseja-se saber se uma nova média amostral $\overline{X}$, obtida a partir de um novo conjunto de dados de tamanho N, é compatível com os dados passados. Desconhece-se σ_X^2, mas se conhece s_X^2. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Nesse caso, a variável

$$t = \frac{\overline{X} - \mu_X}{\frac{s_X}{\sqrt{N}}} \tag{3.32}$$

segue a distribuição t, com ν=N-1 graus de liberdade. Logo, a distribuição t pode ser usada para gerar os intervalos de confiança de $\overline{X}$ e verificar se o valor obtido é compatível com o esperado.

$$\mu_X - t_1 \frac{s_X}{\sqrt{N}} < \overline{X} < \mu_X + t_2 \frac{s_X}{\sqrt{N}} \tag{3.33}$$

Condição especial 3 - Dois conjuntos de dados com $\left(\overline{X_1}, s_1^2, N_1\right)$ e $\left(\overline{X_2}, s_2^2, N_2\right)$ estão disponíveis. Deseja-se saber se as médias podem ser consideradas diferentes. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Como os dados flutuam normalmente, as médias amostrais também flutuam normalmente com variâncias desconhecidas e iguais a σ_1^2/N_1 e σ_2^2/N_2. A diferença entre as médias amostrais, $D = \overline{X_1} - \overline{X_2}$, flutua com variância $\sigma_D^2 = \sigma_1^2/N_1 + \sigma_2^2/N_2$. Se as populações são similares, $\sigma_D^2 = \sigma^2\left[1/N_1 + 1/N_2\right]$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Admitindo-se que as médias são iguais, porque as populações são semelhantes, e que se conhece a variância verdadeira dos dados σ^2, então a variável

$$u = \frac{D}{\sigma_D} \tag{3.34}$$

tem distribuição normal, com média zero e variância igual a 1. Assim,

$$-u_1\sigma_D < D < -u_2\sigma_D \tag{3.35}$$

Se a variância real não é conhecida, admitindo-se que os conjuntos são similares e que têm a mesma variância verdadeira, então

$$s_{1+2}^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} \tag{3.36}$$

é uma estimativa melhor da variância da medida, com $\nu_1 + \nu_2$ graus de liberdade. Assim,

$$s_D^2 = s_{1+2}^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right) \tag{3.37}$$

é uma estimativa da variância de D com $\nu_1 + \nu_2$ graus de liberdade. Logo, a variável

$$t = \frac{D}{s_D} \tag{3.38}$$

segue a distribuição t, com $\nu_1 + \nu_2$ graus de liberdade, de forma que

$$-t_1 s_D < D < -t_2 s_D \tag{3.39}$$

Exemplo 3.19 - O desempenho de dois tipos de gasolina é apresentado abaixo:

Gasolina	**1**	**2**
Milhas/galão (média)	22.7	21.3
Desvio padrão amostral	0.45	0.55
Número de carros em que foram feitas as medidas	5	5

$$D = \overline{X_1} - \overline{X_2} = 1.4$$

$$s_{1+2}^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} = \frac{4 \cdot 0.45^2 + 4 \cdot 0.55^2}{4+4} = 0.2525$$

$$s_D^2 = s_{1+2}^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right) = 0.2525 \left(\frac{1}{5} + \frac{1}{5} \right) = 0.101$$

$$s_D = 0.3178$$

$$t = \frac{D}{s_D} = 4.405$$

Para 8 graus de liberdade e 95% de confiança,

$$-2.306 < t < 2.306$$

Conclui-se, portanto, que o valor observado de t é pouco provável e que as gasolinas são diferentes com 95% de confiança.

É importante observar que testes similares podem ser utilizados para verificar se uma determinada média difere significativamente de zero, por exemplo. Este teste é bastante importante para a estimação de parâmetros, como será visto nos capítulos posteriores.

3.4.2. Testes Adicionais para a Variância

Condição especial 1 - Seja uma média histórica μ_X e a respectiva variância σ_X^2, obtidas com número elevado de graus de liberdade e consideradas iguais aos valores verdadeiros. Deseja-se saber se uma nova variância amostral s_X^2, obtida a partir de um novo conjunto de dados de tamanho *N*, é compatível com os dados passados. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Nesse caso, a variável

$$\chi^2 = (N-1)\frac{s_X^2}{\sigma_X^2} \tag{3.40}$$

segue a distribuição χ^2, com ν=*N*-1 graus de liberdade. Logo

$$\chi_1^2 \frac{\sigma_X^2}{(N-1)} < s_X^2 < \chi_2^2 \frac{\sigma_X^2}{(N-1)} \tag{3.41}$$

3.4.3. Testes Adicionais de Aleatoriedade

Condição especial 1 - Seja uma média histórica μ_X e a respectiva variância σ_X^2, obtidas com número elevado de graus de liberdade e consideradas iguais aos valores verdadeiros. Deseja-se saber se as flutuações das medidas amostrais em um conjunto de tamanho *N* podem ser admitidas normais.

Nesse caso, a variável

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\right)^2 \tag{3.42}$$

segue a distribuição χ^2, com $\nu = N$ graus de liberdade. Logo,

$$\chi_1^2 < \chi^2 < \chi_2^2 \tag{3.43}$$

Condição especial 2 - Deseja-se saber se as flutuações das medidas amostrais em um conjunto de tamanho *N* podem ser admitidas normais.

Nesse caso, a variável

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \overline{X}}{s_X} \right)^2 \tag{3.44}$$

segue a distribuição χ^2, com ν=N-1 graus de liberdade. Logo,

$$\chi_1^2 < \chi^2 < \chi_2^2 \tag{3.45}$$

Condição especial 3 - Deseja-se saber se as flutuações das medidas amostrais em um conjunto de tamanho N seguem uma distribuição estatística particular.

Esse problema pode ser tratado de forma mais rigorosa usando-se as ferramentas de estimação de parâmetros apresentadas nos próximos capítulos. No entanto, uma técnica muito usada consiste em construir uma tabela na forma:

Intervalo	**Limites do Intervalo**	**Probabilidade do Intervalo**	**Número total de observações**
1	$x_0 < x < x_1$	$1/NI$	N_1
2	$x_1 < x < x_2$	$1/NI$	N_2
...	...	...	...
NI	$x_{NI\text{-}1} < x < xN_{\mathrm{I}}$	$1/NI$	N_{NI}

que divide o domínio de definição da distribuição que está sendo testada em NI intervalos igualmente prováveis. Então, o número de observações efetuadas em cada intervalo é distribuído na tabela. Para analisarmos os dados, é conveniente observar que um ponto experimental pode estar ou não no intervalo considerado (2 respostas são possíveis) e que a probabilidade de acerto ($1/NI$) é conhecida. Logo, o número provável de pontos colhidos em cada intervalo pode ser previsto com a curva binomial. Os valores observados são então comparados com aqueles obtidos pela curva binomial, para um dado grau de confiança. Se todos os valores observados estão em conformidade com a previsão efetuada com a distribuição binomial, então a curva de probabilidade originalmente proposta pode ser considerada plausível; caso contrário, a curva de probabilidade proposta deve ser descartada. Se N é o número total de pontos considerado, um procedimento heurístico consiste em fazer $NI = \sqrt{N}$. Sabe-se que se NI < 5, o poder de discriminação dessa técnica é muito baixo, o que mostra uma vez mais a necessidade de grande número de réplicas para um ajuste adequado da curva de distribuição de probabilidades.

Exemplo 3.20 - No Exemplo 3.5 foi gerada a seguinte seqüência de pontos experimentais que seguem uma distribuição uniforme:

Tabela 3.10 - Números aleatórios com distribuição uniforme no intervalo (0,1), gerados como no Exemplo 3.5.

0.0109	0.1194	0.3298	0.3970	0.4607	0.6282	0.7481	0.8654
0.0306	0.1610	0.3369	0.4055	0.4766	0.6725	0.7573	0.9101

0.0316	0.2291	0.3416	0.4423	0.5192	0.6732	0.7680	0.9237
0.0498	0.2430	0.3475	0.4476	0.5202	0.7062	0.7706	0.9493
0.0680	0.3138	0.3665	0.4518	0.5482	0.7227	0.8227	0.9702

A média e variância amostrais são iguais a

$$\overline{X} = \frac{\sum_{i=1}^{40} x_i}{40} = 0.4884$$

$$s_x^2 = \frac{\sum_{i=1}^{40} \left(x_i - 0.4884\right)^2}{39} = 0.07952$$

$$s_X = \sqrt{s_X^2} = 0.2820$$

Deseja-se saber se a curva normal pode representar de forma adequada esse conjunto de dados aleatórios. Para isso, admitindo que $\mu_X = \overline{X}$, que $\sigma_X^2 = s_X^2$ e que $NI = \sqrt{40} = 6$, monta-se a seguinte Tabela de distribuição dos dados.

Tabela 3.11 - Distribuição dos pontos da Tabela 3.10 em intervalos de igual probabilidade da curva normal.

Intervalo	**Limites do Intervalo**	**Probabilidade do Intervalo**	**Número total de observações**
1	$-\infty < x < 0.2156$	1 / 6	7
2	$0.2156 < x < 0.3669$	1 / 6	8
3	$0.3669 < x < 0.4884$	1 / 6	7
4	$0.4884 < x < 0.6099$	1 / 6	3
5	$0.6099 < x < 0.7612$	1 / 6	7
6	$0.7612 < x < \infty$	1 / 6	8

Os limites de confiança de 95% obtidos a partir da curva binomial, com m=40 e p=1/6 (ver Seção 2.1) são 2 ($P_{AC}(2;40,1/6) \cong 0.025$) e 12 ($P_{AC}(12;40,1/6) \cong 0.975$). Logo, o número de observações em cada um dos intervalos analisados deve estar entre 2 e 12, com 95% de confiança. Como essa condição é satisfeita em todos os intervalos da Tabela 3.11, não é possível dizer que os dados da Tabela 3.10, gerados segunda uma distribuição uniforme, não seguem uma distribuição normal. Vê-se uma vez mais como é difícil definir de forma inequívoca a curva de distribuição de probabilidades que rege um determinado problema físico. A Figura 3.10 confirma claramente o resultado e mostra como pode ser difícil discriminar diferentes curvas de densidade de probabilidade mesmo quando um número razoável de pontos está a disposição, como no caso.

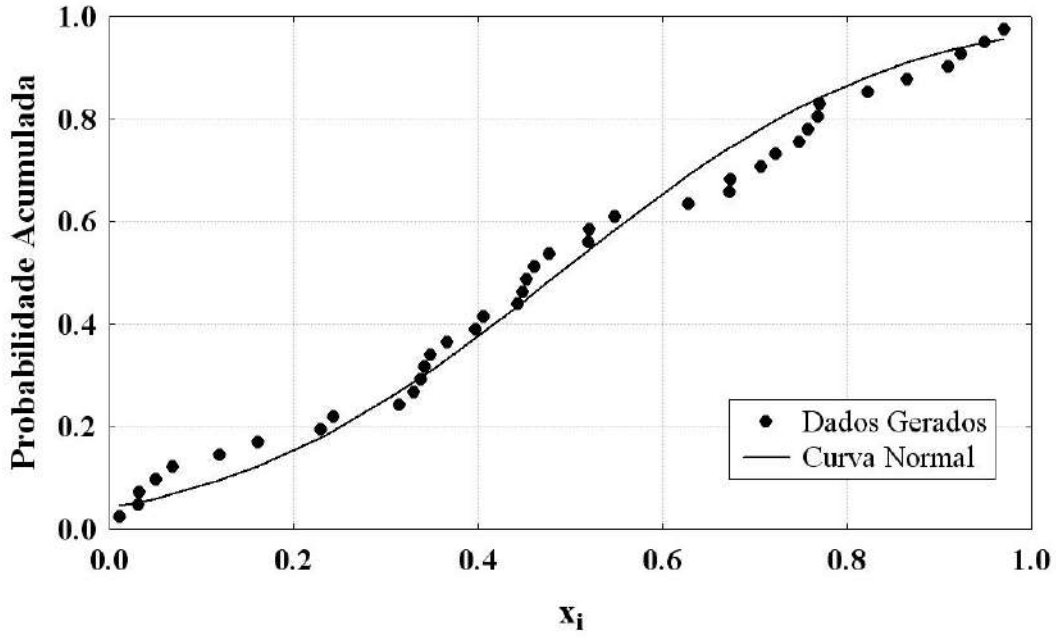


Figura 3.10 - Ajuste normal aos dados da Tabela 3.10.

3.4.4. Testes Adicionais de Independência dos Dados

Condição especial 1 - Dois conjuntos de dados com $\left(\overline{X}, s_X^2, N\right)$ e $\left(\overline{Y}, s_Y^2, N\right)$ estão disponíveis. Deseja-se saber se os dados podem estar correlacionados. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Nesse caso, a medida de dependência é dada pela covariância ou pelo fator de correlação (ver Seção 1.6). No entanto, como saber se a medida é significativa? Um teste bastante simples é baseado na Equação (1.40)

$$\text{Var}\{x+y\} = \text{Var}\{x\} + 2\text{Covar}\{x,y\} + \text{Var}\{y\} \tag{1.40}$$

Se os dados são independentes, a variância da soma (diferença) é a soma das variâncias. Se os dados não são independentes, a variância da soma (diferença) é diferente da soma das variâncias. O teste consiste em verificar com o teste F se a diferença observada é inferior ou não àquela que poderia ser causada por mera flutuação aleatória.

Exemplo 3.21 - O seguinte conjunto de dados está disponível:

x:	1	2	3	4	5
y:	1.1	1.9	3	3.9	5.1

que resultam nas grandezas amostrais

$$\overline{X} = 3 \qquad s_X^2 = 2.50 \qquad s_X = 1.5811$$
$$\overline{Y} = 3 \qquad s_Y^2 = 2.51 \qquad s_X = 1.5843$$
$$s_{XY}^2 = 2.50 \qquad \rho_{XY} = \frac{s_{XY}^2}{s_X s_Y} = 0.998$$

Para a soma (diferença) de *x* e *y*, as grandezas amostrais são

$$\overline{X+Y} = 6 \qquad s_{X+Y}^2 = 10.1 \qquad s_{X+Y} = 3.1639$$

$$\overline{X-Y} = 0 \qquad s_{X-Y}^2 = 0.01 \qquad s_{X+Y} = 0.1$$

Fixando-se o limite de confiança em 95%, para quatro graus de liberdade obtém-se

$$\frac{1}{9.6045} < F < 9.6045$$

Para os dois casos analisados

$$\mathrm{F} = \frac{\mathrm{s}_{\mathrm{X-Y}}^2}{\left(\mathrm{s}_{\mathrm{X}}^2 + \mathrm{s}_{\mathrm{Y}}^2\right)} = \frac{0.01}{5.01} = 0.002 \quad ; \quad \mathrm{F} = \frac{\mathrm{s}_{\mathrm{X+Y}}^2}{\left(\mathrm{s}_{\mathrm{X}}^2 + \mathrm{s}_{\mathrm{Y}}^2\right)} = \frac{10.01}{5.01} = 2.00$$

Vê-se, portanto, que as diferenças observadas na variação das diferenças não poderiam ser explicadas por flutuações puramente aleatórias. Assim, pode-se dizer que a covariância (e o fator de correlação) entre *x* e *y* são significativos com 95% de confiança.

O resultado obtido não deve impressionar demais o leitor, pois esse problema era fácil de resolver. Na maior parte dos casos, poucos pontos resultam quase sempre em baixa qualidade de resolução dos termos de correlação.

Condição especial 2 - Um conjunto de dados com $\left(\overline{X}, s_X^2, N\right)$ está disponível. Deseja-se saber se os dados obtidos são realmente aleatórios ou se podem estar correlacionados entre si. Admite-se que as medidas amostrais flutuam de acordo com a curva normal.

Nesse caso, é conveniente definir a **função de auto-correlação** na forma

$$C_{X_k} = \frac{\sum_{i=1}^{N-k}\left(x_i - \overline{X}_0\right)\left(x_{i+k} - \overline{X_k}\right)}{N-k-1} \tag{3.46}$$

ou na forma

$$C_{X_k} = \frac{\sum_{i=1}^{N-k}\left(x_i - \overline{X}_0\right)\left(x_{i+k} - \overline{X_k}\right)}{\left[\sum_{i=1}^{N-k}\left(x_i - \overline{X}_0\right)^2\right]^{0.5}\left[\sum_{i=1}^{N-k}\left(x_{i+k} - \overline{X}_k\right)^2\right]^{0.5}} \tag{3.47}$$

em que é calculada a covariância (Equação (3.46)) ou a correlação (Equação (3.47)) de dados amostrais deslocados de *k* unidades no tempo. Nesse caso, $\overline{X_0}$ é a média amostral dos primeiros *N*-*k* valores amostrados, enquanto $\overline{X_k}$ é a média amostral dos últimos *N*-*k* valores amostrados. A função de auto-correlação pode fornecer importantes pistas sobre a existência de dinâmica (não aleatoriedade) entre os dados amostrados e sobre a existência de efeitos experimentais indesejados. No entanto, para evitar a tomada equivocada de conclusões, a significância dos valores calculados com a Equação (3.46) deve ser sempre testada, como ilustrado no Exemplo 3.21. Como procedimento heurístico, recomenda-se que (*N*-*k*) seja sempre igual ou superior a 20 para uso eficiente das Equações (3.46-47).

Fundamentalmente, a função de auto-correlação mostra se existe uma memória entre dados que se sucedem em uma série de dados. Se existe uma relação determinística entre os dados (por exemplo, os dados representam a resposta de um processo a uma dada perturbação), as correlações são significativas e se aproximam do valor unitário. Se os dados são corrompidos por erros experimentais e/ou as perturbações do processo são muito freqüentes, as correlações tendem a diminuir à medida que o atraso *k* aumenta. Dessa forma, é possível definir um horizonte de memória do processo, que é o máximo valor de k para o qual ainda se observam correlações significativas entre os dados. Essa informação pode ser fundamental em vários problemas.

Um exemplo típico de aplicação prática dos espectros de auto-correlação é a análise do comportamento dinâmico de processos. Se um processo opera em condições estacionárias (todas as variáveis se mantêm aproximadamente constantes ao longo do tempo), as flutuações dos dados refletem apenas os erros de medida e operação do processo (ou seja, as flutuações são essencialmente aleatórias), de forma que o espectro de auto-correlação deve apresentar correlações muito próximas de zero para qualquer valor de *k* considerado. Assim, se correlações pronunciadas são observadas para valores de *k* baixos, esse é um indício claro de que o processo opera de forma dinâmica na freqüência de amostragem dos dados e que qualquer tentativa de interpretação dos dados deve ser feita à luz de um modelo dinâmico do processo. Portanto, o espectro de auto-correlação auxilia na definição da melhor estratégia de modelagem matemática dos dados disponíveis. Além disso, o máximo valor de *k* para o qual as correlações ainda podem ser consideradas significativas (k_{max}) é uma constante de tempo que caracteriza o processo e o procedimento de amostragem. Esse dado pode conter importante conteúdo de informação para a implementação de rotinas de controle de processo e simulação. Por exemplo, o uso de simuladores estacionários só deveria ser usado para descrição do processo se os dados estão amostrados com freqüência inferior àquela definida por k_{max}, para que seja possível filtrar a influência dinâmica que um dado da seqüência exerce sobre o outro. Mais ainda, esquemas de controle devem coletar informações do processo com freqüência superior àquela definida por k_{max}, para que seja possível capturar a

informação dinâmica e corrigir efeitos causados por perturbações indesejadas do processo.

Exemplo 3.22 - Para o conjunto de dados ilustrado abaixo na Figura 3.11, calcula-se o espectro de auto-correlação da Figura 3.12. Vê-se de forma clara que as correlações diminuem lentamente, à medida que a distância entre os dados aumenta, e tornam-se não significativas após um certo tempo.

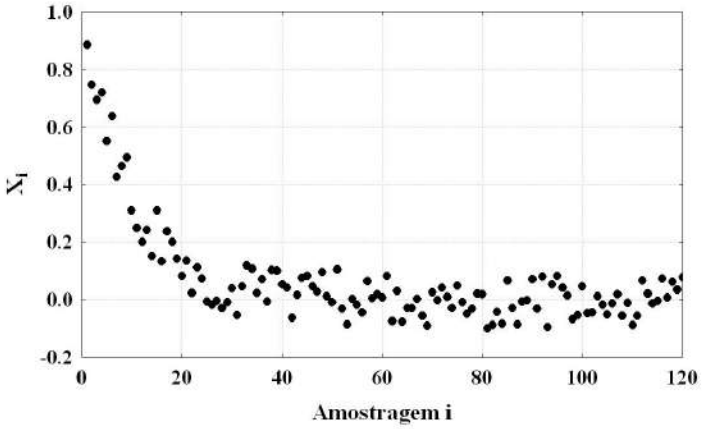


Figura 3.11 - Dados amostrados num processo de experimentação.

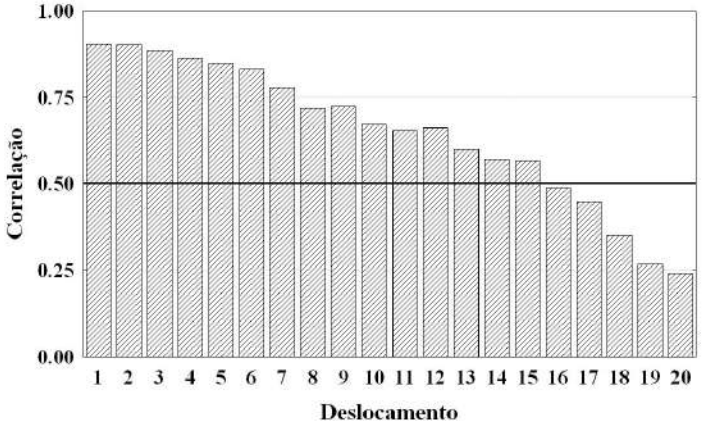


Figura 3.12 - Função de auto-correlação para os dados da Figura 3.11.

Considerando-se que correlações da ordem de 0.5 já são bastante fracas, observa-se na Figura 3.12 que o horizonte de memória característica do processo é de 16 unidades de amostragem (k_{max}). Esse deslocamento dá uma idéia da dinâmica do processo e de quão longe uma informação inserida no processo de experimentação permanece influenciando os demais resultados obtidos. Se comportamento aleatório fosse desejado, como durante a execução de réplicas experimentais, os dados deveriam ser recusados.

3.4.5. Testes Adicionais para *Outliers*

Outlier é a expressão usada genericamente para designar pontos experimentais que parecem não se adequar a uma distribuição particular de probabilidades definida pela grande maioria dos demais pontos experimentais. Quase sempre a detecção de *outliers* visa a eliminação desses pontos suspeitos de não fazerem parte do conjunto. Essa é uma questão muito controversa da prática estatística, em particular quando poucos pontos experimentais estão disponíveis, e será analisada algumas vezes nos capítulos que seguem. De uma forma cautelosa, como descrito por E.J. Gumbel (*Technometrics*, **2**, 165, 1960): *"A rejeição de outliers em bases puramente estatísticas é e continua a ser um procedimento perigoso. Sua existência pode ser a prova de que a população estudada não é, na realidade, o que se assumiu que fosse."*

Se o número de graus de liberdade é pequeno, o melhor teste para detecção de *outliers* parece ser primeiramente a repetição da medida experimental e em segundo lugar a comparação estatística dos resultados amostrais obtidos quando o candidato a *outlier* é removido ou adicionado ao conjunto de dados. Se as comparações resultarem em conclusões de equivalência, a decisão mais sensata será manter o candidato a *outlier* no conjunto de pontos experimentais, a não ser que sobre ele pairem dúvidas de erros grosseiros.

Exemplo 3.23 - Os seguintes dados foram obtidos para a concentração de uma espécie química em uma solução mineral

x (ppm): 23.2 23.4 23.5 24.1 25.5

havendo desconfiança de que o último ponto seja na realidade um *outlier*. Para analisar a questão, para um grau de confiança de 95%, o conjunto amostral que contém o *outlier*

$$\overline{X} = 23.94 \qquad s_X^2 = 0.873 \qquad s_X = 0.934 \qquad \nu = 4$$
$$22.78 < \mu_X < 25.10$$

é comparado com o conjunto amostral que não contém o *outlier*

$$\overline{X} = 23.55 \qquad s_X^2 = 0.150 \qquad s_X = 0.387 \qquad \nu = 3$$
$$22.93 < \mu_X < 24.17$$

$$\frac{1}{9.9792} < F = \frac{0.873}{0.150} = 5.82 < 15.101$$

Como as médias e variâncias obtidas com e sem o *outlier* são estatisticamente semelhantes, não parece razoável descartar o candidato a *outlier* do conjunto de pontos.

3.5. A Região de Confiança em Problemas Multidimensionais

Chama-se de região de confiança com probabilidade *p* àquela região do espaço de variáveis que concentra uma probabilidade definida e igual a *p* das possíveis flutuações observáveis no problema. Em um problema unidimensional, a definição da

região de confiança é extremamente simples, pois consiste simplesmente em descartar as extremidades inferior e superior dos valores menos prováveis que concentram probabilidades (1-*p*)/2. Em um problema multidimensional, no entanto, a definição da região de confiança pode não ser um problema bem posto, pois diferentes regiões, com diferentes formas, podem resultar numa mesma concentração de probabilidades. Essa questão está ilustrada no Exemplo 3.24 a seguir.

Exemplo 3.24 - Considere a distribuição exponencial de probabilidades definida para duas variáveis no Exemplo 2.14.

$$\wp\left(x_1; x_2\right) = 2e^{(-x_1-2x_2)}$$

Pode-se então construir regiões de confiança com forma quadrada, com lados de tamanho 2*a* e centradas ao redor do ponto médio, na forma

$$\int_{1-a}^{1+a}\int_{0.5-a}^{0.5+a} 2e^{(-x_1-2x_2)}dx_2dx_1 = 2\int_{1-a}^{1+a} e^{(-x_1)}\int_{0.5-a}^{0.5+a} e^{(-2x_2)}dx_2dx_1 =$$

$$2\left[\frac{e^{(-x_1)}}{-1}\right]_{1-a}^{1+a}\left[\frac{e^{(-2x_2)}}{-2}\right]_{0.5-a}^{0.5+a}$$

cuja confiança depende do valor de *a*. Como ambas as variáveis x_1 e x_2 são estritamente positivas, o maior valor admissível para a é 0.5 (lados iguais a 1). Portanto, o maior quadrado centrado em torno da média representa uma confiança de 33.15%.

Alternativamente, pode-se também construir regiões de confiança com forma retangular, com lados de tamanhos proporcionais a 2:1 e centradas ao redor do ponto médio, na forma

$$\int_{1-2\,a}^{1+2\,a}\int_{0.5-a}^{0.5+a} 2e^{(-x_1-2x_2)}dx_2dx_1 = 2\int_{1-2\,a}^{1+2\,a} e^{(-x_1)}\int_{0.5-a}^{0.5+a} e^{(-2x_2)}dx_2dx_1 =$$

$$2\left[\frac{e^{(-x_1)}}{-1}\right]_{1-2a}^{1+2a}\left[\frac{e^{(-2x_2)}}{-2}\right]_{0.5-a}^{0.5+a}$$

De forma análoga, o maior desses retângulos admissível tem lados iguais a 2 e a 1. Nesse caso, o retângulo máximo admissível concentra uma confiança de 74.76%. Logo, parece claro que existe um retângulo com os lados na proporção 2:1 e centrado em torno do ponto médio que concentra a mesma confiança do quadrado com lado de comprimento igual a 1. Na realidade, esse retângulo tem os lados com comprimentos iguais a 1.44 e 0.72, nas direções de x_1 e x_2 respectivamente.

Da mesma forma que feita entre o retângulo e o quadrado no caso anterior, diferentes regiões de forma retangular, circular, elipsoidal, etc., podem ser desenhadas para conter a mesma probabilidade de observação dos dados que a região quadrada

proposta inicialmente. Logo, não é possível definir a forma da região de confiança de forma inequívoca sem que restrições adicionais sejam impostas ao problema.

3.5.1. A Geometria da Região de Confiança da Curva Normal Multidimensional

Como mostrado no Exemplo 3.24, não é possível definir uma região de confiança de forma inequívoca em problemas multidimensionais sem que se imponham restrições adicionais ao problema. No caso particular da curva normal multidimensional, uma propriedade muito importante é o fato de que a curva apresenta a forma de um chapéu ou sino, convergindo para o valor zero à medida que as variáveis tendem a infinito em quaisquer direções do espaço. Portanto, é possível desenhar curvas de nível fechadas, onde a densidade de probabilidade se mantém constante. Por isso, para o caso da curva normal multidimensional, define-se a região de confiança com probabilidade *p* àquela região do espaço de variáveis que é limitada por uma superfície onde todos os pontos estão associados a um mesmo valor da densidade de probabilidade e onde a integral da função densidade de probabilidade é igual a *p*. O conceito de região de confiança aqui proposto pode ser facilmente compreendido se imaginarmos que a função densidade de probabilidade descreve um relevo no espaço e as superfícies que delimitam regiões de diferentes probabilidades são as curvas de nível, como mostrado na Figura 3.13..

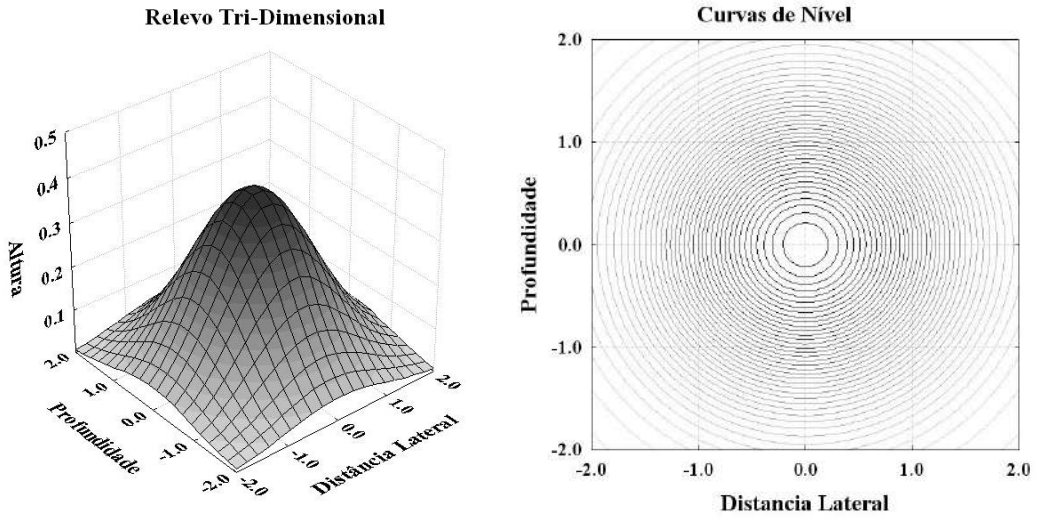


Figura 3.13 - Definição da região de confiança para a curva normal multidimensional.

No caso da curva normal, a definição da região de confiança está associada ao expoente da Equação (2.72), dado que os demais termos da equação são constantes e não dependem do ponto experimental considerado. Sendo assim, as curvas de nível que limitam as regiões de confiança satisfazem a Equação (3.48) abaixo:

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\mathbf{V}_{\mathbf{X}}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)=c \tag{3.48}$$

onde *c* é uma constante que caracteriza o nível da função densidade de probabilidade e, portanto, o grau de confiança. Quanto menor o valor de *c*, maior o grau de confiança,

uma vez que a função normal tende a zero para valores muito grandes. A região de confiança é então aquela que satisfaz a Equação (3.49)

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \mathbf{V}_\mathbf{X}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) \le c \tag{3.49}$$

constituída pelos pontos interiores em relação à curva de nível.

As Equações (3.48-49) são muito estudadas na Álgebra e caracterizam um conjunto particular de curvas chamadas de formas quadráticas. Este nome deve-se ao fato de que, depois de feitas as multiplicações vetoriais, a Equação (3.48) pode ser colocada na forma:

$$\sum_{i=1}^{NX}\sum_{j=1}^{NX}\left(v_{ij}^{-1}\right)\left(x_i-\mu_i\right)\left(x_j-\mu_j\right)=c \tag{3.50}$$

que é a generalização de uma polinomial de segundo grau para várias variáveis. v_{ij}^{-1} é o elemento *ij* da inversa da matriz de covariâncias de **x**.

Como a matriz $\mathbf{V_X}$ é positiva definida, a curva definida pela Equação (3.48) é uma forma quadrática muito especial, que recebe o nome de **hiper-elipse**; ou seja, uma elipse no espaço de dimensão *NX*. Portanto, a região de confiança obtida a partir da curva normal é sempre uma elipse no espaço de variáveis de dimensão *NX*. O problema é que o estudo da Equação (3.48) na forma proposta é bastante dificultado pelo fato da matriz $\mathbf{V_X}$ não ser diagonal, o que faz com que todos os termos quadráticos apareçam, como na Equação (3.50). Portanto, antes de estudar as características da hiper-elipsóide que define a região de confiança, é conveniente diagonalizá-la. Para tanto, lembremos do problema clássico de valores característicos, colocado como encontrar os números λ (valores característicos) e vetores **d** (vetores característicos) que satisfazem a seguinte equação:

$$\mathbf{V_X d} = \lambda \mathbf{d} \tag{3.51}$$

ou seja

$$\left(\mathbf{V_X}-\lambda\mathbf{I}\right)\mathbf{d}=\mathbf{0} \tag{3.52}$$

O sistema de equações (3.52) é um sistema linear clássico. Para que existam soluções não triviais da Equação (3.52), é necessário que a matriz ($\mathbf{V_X}$ - $\lambda\mathbf{I}$) seja singular; ou seja, que seu determinante seja igual a zero. Portanto, a equação

$$\det\left(\mathbf{V_x}-\lambda\mathbf{I}\right)=0 \tag{3.53}$$

é a equação que permite calcular os valores característicos do sistema. Uma vez obtidos os valores característicos do sistema, a Equação (3.51) pode ser utilizada para que sejam obtidos os vetores característicos. Como a matriz ($\mathbf{V_X}$ - $\lambda\mathbf{I}$) é singular, infinitos vetores característicos satisfazem a Equação (3.51). Para normalizar e definir de forma única a solução do problema, é conveniente tomar como solução, dentre as infinitas soluções

existentes, aquela cujo vetor tem tamanho unitário. Deve ser ainda enfatizado que a Equação (3.53) resulta sempre em um polinômio de grau *NX*, que portanto admite até *NX* diferentes raízes ou valores característicos. Como a matriz $\mathbf{V_X}$ é positiva definida e simétrica, é possível garantir que todos os seus valores característicos são números reais e positivos.

A Equação (3.51) pode ser re-escrita de forma compacta, englobando todas as soluções características do sistema ao mesmo tempo, na forma:

$$\mathbf{V_X}\left[\mathbf{d_1}\vdots\mathbf{d_2}\vdots\cdots\vdots\mathbf{d_{NX}}\right]=\left[\mathbf{d_1}\vdots\mathbf{d_2}\vdots\cdots\vdots\mathbf{d_{NX}}\right]\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ 0 & 0 & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_{NX}\end{bmatrix} \tag{3.54}$$

que pode então ser usada como definição da matriz diagonal dos valores característicos e da matriz de vetores característicos na forma:

$$\mathbf{V_X D = D\Lambda} \tag{3.55}$$

onde

$$\mathbf{\Lambda}=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_{NX}\end{bmatrix} \tag{3.56}$$

e

$$\mathbf{D}=\left[\mathbf{d_1}\vdots\mathbf{d_2}\vdots\cdots\vdots\mathbf{d_{NX}}\right] \tag{3.57}$$

Desta forma, é possível representar a matriz $\mathbf{V_X}$ como o produto de matrizes

$$\mathbf{V_X = D\Lambda D^{-1}} \tag{3.58}$$

onde $\mathbf{\Lambda}$ tem estrutura diagonal.

Exemplo 3.25 - Seja a matriz $\mathbf{A}=\begin{bmatrix}1 & -1\\ 0 & 2\end{bmatrix}$. Neste caso, os valores característicos são iguais a:

$$\det\left(\mathbf{A}-\lambda\mathbf{I}\right)=\det\left(\begin{bmatrix}1-\lambda & -1\\ 0 & 2-\lambda\end{bmatrix}\right)=\left(1-\lambda\right)\left(2-\lambda\right)-0\left(-1\right)=0$$

$$\lambda^2-3\lambda+2=0$$

cujas raízes são:

$$\lambda = \frac{-(-3) \pm \sqrt{(-3)^2 - 4(1)(2)}}{2(1)} = \begin{cases} 1 \\ 2 \end{cases}$$

Assim, os vetores característicos podem ser obtidos como:

$$\begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = 1\begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{bmatrix} a - b = a \\ 2b = b \end{bmatrix} \Rightarrow \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix} = \mathbf{d_1}$$

A solução com tamanho unitário é $\mathbf{d_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$$\begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = 2\begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{bmatrix} a - b = 2a \\ 2b = 2b \end{bmatrix} \Rightarrow \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -b \\ b \end{bmatrix} = \mathbf{d_2}$$

A solução com tamanho unitário é $\mathbf{d_2} = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$.

Desta forma, $\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ e $\mathbf{D} = \begin{bmatrix} 1 & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} \end{bmatrix}$.

Calculando-se a matriz inversa de **D** como

$$\mathbf{D^{-1}} = \frac{1}{\det(\mathbf{D})}\begin{bmatrix} d_{22} & -d_{12} \\ -d_{21} & d_{11} \end{bmatrix} = \frac{1}{\left(\frac{\sqrt{2}}{2}\right)}\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{bmatrix}$$

chega-se finalmente à representação diagonalizada de **A** como

$$\mathbf{A} = \begin{bmatrix} 1 & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{bmatrix}$$

Como além de positiva definida, a matriz $\mathbf{V_X}$ é simétrica, é possível mostrar que $\mathbf{D^{-1}} = \mathbf{D^T}$, de forma que nos problemas que nos interessam mais diretamente, é possível escrever:

$$\mathbf{V_x} = \mathbf{D\Lambda D^T} \tag{3.59}$$

Substituindo a Equação (3.59) na Equação (3.48), a equação que descreve a superfície que envolve a região de confiança ganha a forma:

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \mathbf{D\Lambda^{-1}D^T}\left(\mathbf{x}-\boldsymbol{\mu}\right) = c \tag{3.60}$$

Finalmente, redefinindo as variáveis do problema como

$$\mathbf{z} = \mathbf{D^T}\left(\mathbf{x}-\boldsymbol{\mu}\right) \tag{3.61}$$

a Equação (3.60) ganha a forma

$$\mathbf{z^T\Lambda^{-1}z} = c \tag{3.62}$$

que tem a forma explícita

$$\sum_{i=1}^{NX} \frac{z_i^2}{\lambda_i} = c \tag{3.63}$$

facilmente identificável como uma elipse centralizada no ponto central e com semi-eixos com comprimentos iguais a $\sqrt{c\lambda_i}$. Repare que *c*, ou o grau de confiança exigido, não exerce qualquer influência sobre o formato da região de confiança, excetuando-se obviamente o aumento proporcional de todos os semi-eixos da elipse. Por isso, quase sempre o fator *c* é desprezado durante a análise, já que ele apenas muda de forma absolutamente proporcional os eixos da elipse. Esses resultados indicam que as regiões de confiança obtidas para a curva normal para diferentes níveis de confiança formam uma estrutura semelhante à da cebola, em que as regiões com maior confiança envolvem completa e proporcionalmente as regiões de menor confiança.

O conjunto de transformações introduzidas através da Equação (3.61) representa uma translação para o zero e uma rotação da elipse, de forma a fazer com que os seus semi-eixos coincidam com os eixos ortogonais e que o centro da elipse coincida com a origem dos eixos de coordenadas. As transformações da Equação (3.61) são isométricas, no sentido de que elas preservam a forma original da figura geométrica, como ilustrado na Figura 3.14.

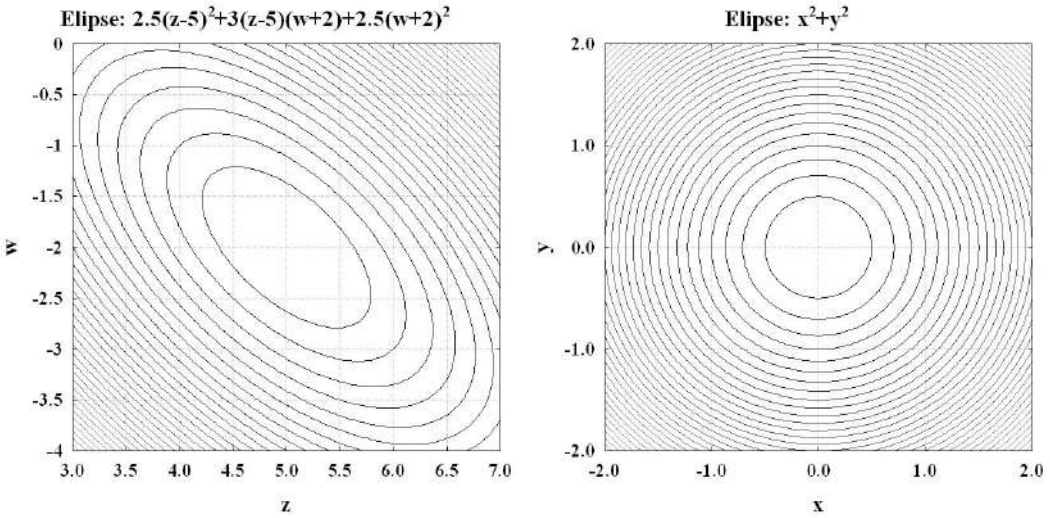


Figura 3.14 - Transformações geométricas devidas às mudanças de coordenadas.

A partir da Equação (3.63) fica relativamente fácil extrair muitas informações sobre a geometria da região de confiança de um problema descrito pela curva normal multidimensional. As informações mais importantes são:

1- A região de confiança da curva normal multidimensional é uma hiper-elipse, cujos eixos têm comprimentos proporcionais a $\sqrt{\lambda_i}$, onde λ_i, i=1,...,NX, são os valores característicos de $\mathbf{V_X}$;

2- A assimetria máxima da hiper-elipse que descreve a região de confiança, ou fator de esfericidade, definida como a razão entre os comprimentos extremos de seus eixos, pode ser dada por

$$\phi = \sqrt{\frac{\lambda_{MIN}}{\lambda_{MAX}}} \tag{3.64}$$

3- Como o traço de uma matriz (a soma dos elementos da diagonal principal) é igual à soma de seus valores característicos, ou seja,

$$\mathrm{tr}\left(V_X\right) = \sum_{i=1}^{NX} v_{ii} = \sum_{i=1}^{NX} \lambda_i \tag{3.65}$$

o traço da matriz de covariâncias é igual à soma dos comprimentos quadrados de seus eixos;

4- Como o volume de uma elipse é proporcional ao produto do comprimento de seus eixos, conclui-se que o volume da região de confiança é proporcional à raiz quadrada do produto dos valores característicos de $\mathbf{V_X}$. Como o produto dos valores característicos de uma matriz é idêntico ao valor do determinante da matriz, é possível escrever

$$\text{Volume} \approx \sqrt{\det\left(V_X\right)} = \prod_{i=1}^{NX} \sqrt{\lambda_i} \tag{3.66}$$

Portanto, os valores característicos da matriz de covariâncias $\mathbf{V_X}$ guardam muitas informações a respeito da geometria da região de confiança da distribuição normal. Repare que distribuições probabilísticas não normais podem apresentar geometria da região de confiança bastante distinto do aqui apresentado.

Exemplo 3.26 - Seja a distribuição de probabilidades exponencial apresentada abaixo:

$$\wp\left(x\right) = \left[\frac{1}{2^{NX}}\right]\left[\frac{1}{\prod_{i=1}^{NX}\tau_i}\right]\exp\left(\sum_{i=1}^{NX} -\frac{\left|x_i - \mu_i\right|}{\tau_i}\right)$$

cujo vetor de médias e matriz de covariâncias são dados por

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{NX} \end{bmatrix}, \quad \mathbf{V_X} = \begin{bmatrix} 2\tau_1^2 & 0 & \cdots & 0 \\ 0 & 2\tau_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\tau_{NX}^2 \end{bmatrix}$$

A região de confiança da distribuição exponencial pode também ser obtida explorando-se a simetria da distribuição em torno do centro $\boldsymbol{\mu}$ e o fato de que a função converge suavemente para o zero nos limites de infinitamente positivos ou negativos. Assim, como no caso da curva normal, a região de confiança pode ser dada pela equação

$$\sum_{i=1}^{NX} \frac{\left|x_i - \mu_i\right|}{\tau_i} = c$$

onde c é uma constante relacionada ao grau de confiança desejado. A equação que define a forma da região de confiança é a equação de 2^{NX} planos, a depender do sinal adotado para o termo na função módulo. Esses planos cruzam os eixos coordenados nos pontos

$$x_i = \mu_i \pm \tau_i c$$

Como os planos definidos pela equação se interceptam nos mesmos $2NX$ pontos, esses pontos constituem os vértices de um poliedro regular, cujas faces planas são os planos que conectam os vértices em cada um dos quadrantes definidos quando os eixos coordenados são centrados em $\boldsymbol{\mu}$. O poliedro é formado então por 2^{NX} faces e $2NX$ vértices. Os eixos do poliedro são paralelos aos eixos coordenados, conectam vértices opostos e têm comprimentos iguais a $2c\,\tau_i$. Assim, no espaço bidimensional a região de confiança tem a forma de um losango, com centro em $\boldsymbol{\mu}$ e eixos paralelos aos eixos coordenados. No espaço tridimensional a região de confiança tem a forma de um

octaedro regular, com faces triangulares, centro em $\boldsymbol{\mu}$ e eixos paralelos aos eixos coordenados. E assim por diante.

É muito importante perceber que a Equação (3.61) sugere uma mudança de variáveis na forma

$$z_i = \sum_{j=1}^{NX} d_{ij}\left(x_j - \mu_j\right) \tag{3.67}$$

onde d_{ij} representa o j-ésimo componente do i-ésimo vetor característico de $\mathbf{V_X}$. Se os valores característicos são ordenados de forma que

$$\lambda_1 > \lambda_2 > \ldots > \lambda_{NX} \tag{3.68}$$

então as variações observadas podem ser decompostas ao longo das direções definidas pelos vetores característicos, sendo que as variações são máximas ao longo de $\mathbf{d_1}$ (direção que define o maior eixo da hiper-elipse) e mínimas ao longo da direção $\mathbf{d_{NX}}$ (direção que define o menor eixo da hiper-elipse). Por isso, os vetores característicos são freqüentemente chamados de **direções principais** de variação, enquanto os valores característicos são usados para definir as direções do espaço ao longo das quais as variações são mais importantes. Quando um ou mais dos valores característicos apresentam ordem de magnitude muito inferior às dos demais, é possível sugerir a redução do número de variáveis do problema, já que isso indica que uma ou mais combinações de variáveis permanecem essencialmente constantes no conjunto de dados.

Exemplo 3.27 - Seja o vetor de médias $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ e a matriz de covariâncias $\mathbf{V_X} = \begin{bmatrix} 100 & 9 \\ 9 & 1 \end{bmatrix}$, cujos valores característicos são

$$\det\left(\begin{bmatrix} 100-\lambda & 9 \\ 9 & 1-\lambda \end{bmatrix}\right) = (100-\lambda)(1-\lambda) - 81 = \lambda^2 - 101\lambda + 19 = 0$$

$$\lambda = \frac{101 \pm \sqrt{101^2 - 4.19}}{2}$$

$$\lambda_1 = 100.81153\,,\ \lambda_2 = 0.18847$$

Observa-se que as flutuações ocorrem principalmente ao longo da direção 1, enquanto as flutuações observadas ao longo da direção 2 são comparativamente pouco importantes. Isso sugere que há apenas uma variável aleatória no problema, e não duas, como sugerido pela matriz de covariâncias e observações experimentais. A direção principal de variação pode ser obtida como,

$$\begin{bmatrix} 100 & 9 \\ 9 & 1 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = 100.81153\begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{matrix} 100a + 9b = 100.81153\,a \\ 9a + b = 100.81153\,b \end{matrix} \Rightarrow\ a = 11.0901b$$

Para obter o vetor unitário

$$\mathbf{d_1} = \begin{bmatrix} 11.0901 \\ 1 \end{bmatrix} \Rightarrow \|\mathbf{d_1}\| = \sqrt{11.0901^2 + 1^2} = 11.13509$$

Assim

$$\mathbf{d_1} = \frac{1}{11.13509} \begin{bmatrix} 11.0901 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9960 \\ 0.0898 \end{bmatrix}$$

que sugere a seguinte mudança de variáveis

$$z_1 = 0.9960x_1 + 0.0898x_2 - 1.1756$$

que é a verdadeira variável aleatória do problema.

A segunda direção de variação pode ser obtida como,

$$\begin{bmatrix} 100 & 9 \\ 9 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 0.18847 \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{matrix} 100a + 9b = 0.18847a \\ 9a + b = 0.18847b \end{matrix} \Rightarrow a = -0.09017b$$

Para obter o vetor unitário

$$\mathbf{d_2} = \begin{bmatrix} -0.09017 \\ 1 \end{bmatrix} \Rightarrow \|\mathbf{d_2}\| = \sqrt{0.09017^2 + 1^2} = 1.00406$$

Assim

$$\mathbf{d_2} = \frac{1}{1.00406} \begin{bmatrix} -0.09017 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.0898 \\ 0.9960 \end{bmatrix}$$

que sugere a seguinte variável se mantém essencialmente constante e igual a zero

$$z_2 = -0.0898x_1 + 0.9960x_2 - 1.9022 = 0$$

Portanto

$$x_2 = 0.09016x_1 + 1.9098 = 0$$

3.6. Conclusões

Foi mostrado nesse capítulo que, em geral, os parâmetros que caracterizam as curvas de distribuição de probabilidades em problemas estocásticos (em particular a média e a variância) não podem ser jamais obtidos por métodos empíricos. Nesses casos, é preciso definir procedimentos consistentes de inferência, a partir de dados amostrados empiricamente. Contudo, as grandezas amostradas constituem também

variáveis aleatórias, sujeitas a flutuações e incertezas. É necessário, portanto, descrever como essas grandezas flutuam e definir a forma das respectivas distribuições de probabilidade.

No caso particular de medidas sujeitas a flutuações normais, mostrou-se que a média amostral flutua de acordo com a distribuição t de Student, que pode ser utilizada para fins de determinação dos intervalos de confiança dos valores amostrados e para comparações entre valores amostrados em diferentes conjuntos de dados. De forma similar, mostrou-se que a variância amostral flutua de acordo com a distribuição χ^2, que também pode ser utilizada para fins de determinação dos intervalos de confiança dos valores amostrados e para comparações entre valores amostrados em diferentes conjuntos de dados. Contudo, comparações de variâncias obtidas em diferentes conjuntos de dados podem ser feitas de forma mais eficiente com o auxílio da distribuição F de Fisher.

Finalmente, foi mostrado que a geometria natural das regiões de confiança em problemas multidimensionais, descritos adequadamente pela distribuição normal, é a geometria das formas elípticas. Nesse caso, os valores característicos e vetores característicos que caracterizam a matriz de covariâncias do problema representam respectivamente os conteúdos de incertezas e as direções características de flutuações do problema analisado.

3.7. Leitura Adicional

Como já discutido ao final dos Capítulos 1 e 2, a literatura dedicada à apresentação e discussão do problema amostral é imensa. Não cabe aqui, portanto, uma revisão extensa dessa área. O leitor interessado encontrará centenas de livros que abordam esses assuntos em qualquer biblioteca dedicada à Matemática e à Engenharia.

Como já apresentado anteriormente, um texto clássico relacionado ao uso e aplicação dos conceitos discutidos no Capítulo 3 em problemas de Engenharia é apresentado em

"Process Analysis by Statistical Methods", D.M. Himmelblau, John Wiley & Sons, New York, **1970**.

Um outro texto clássico sobre análise e comparação de dados experimentais é apresentado por

"*Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*",G.E.P. Box, W.G. Hunter e J.S. Hunter, John Wiley & Sons, New York, **1978**.

Uma discussão mais formal sobre as propriedades matemáticas associadas ao problema de inferência estatística e aos testes de hipóteses é apresentada em

"Probability and Statistical Inference. Volume 1: Probability", J.G. Kalbfleisch, Springer-Verlag, New York, **1985**.

"Probability and Statistical Inference. Volume 2: Statistical Inference", J.G. Kalbfleisch, Springer-Verlag, New York, **1985**.

"Probability and Statistics. Theory and Applications.", G. Blom, Springer-Verlag, New York, **1989**.

Textos básicos sobre a álgebra de matrizes e formas quadráticas, em especial sobre o cálculo de valores e vetores característicos, podem ser encontrados em

"Matrix Computations", G.H. Golub e C.F. van Loan, The John Hopkins University Press, Baltimore, **1996**.

"Linear Algebra and Its Applications", G. Strang, Harcourt Brace Jovanovich College Publishers, Orlando, **1988**.

"Advanced Engineering Mathematics", C.R. Wylie e L.C. Barrett, McGraw-Hill, New York, **1985**.

3.8. Exercícios Sugeridos

1- Suponha que você está insatisfeito com a reprodutibilidade de uma certa técnica experimental e não pode comprar um novo equipamento e nem pode melhorar a técnica disponível. O que você pode fazer para melhorar a precisão das análises efetuadas? Será que você pode obter uma precisão arbitrariamente pequena para uma técnica experimental? Justifique.

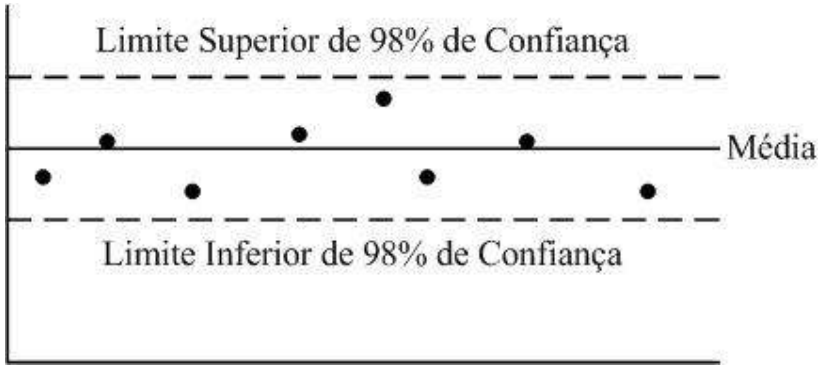
2- Suponha que a análise de dados históricos disponíveis no laboratório indiquem que a variância de uma certa medida experimental é igual a $\sigma^2 = 1$. Como você poderia propor um sistema de amostragem que reduzisse em 10 vezes a variância das medidas? Justifique.

3- Quatro turmas de operadores trabalham numa empresa química. O desempenho das quatro turmas deve ser avaliado. Você é o engenheiro recomendado para isso. Para tanto, você deve analisar os dados de conversão do reator químico onde se processa a reação. Os dados disponíveis são os seguintes:

	Turma 1	Turma 2	Turma 3	Turma 4
1	0.892	0.850	0.775	0.915
2	0.910	0.875	0.872	0.921
3	0.880	0.880	0.650	0.917
4	0.900	0.842	0.881	0.911
5	0.920	0.900	0.910	0.907
6	0.905	0.910	0.720	0.899
7	0.860	0.891	0.851	0.912
8	0.920	0.905	0.820	0.910
9	0.904	0.870	0.730	0.907

10	0.930	0.865	0.780	0.913
11	0.921	0.880	0.792	0.905
12	0.872	0.891	0.751	0.898
13	0.897	0.832	0.891	0.902
14	0.880	0.886	0.950	0.911
15	0.911	0.872	0.971	0.907
16	0.908	0.907	0.918	0.906
17	0.915	0.652	0.863	0.913
18	0.882	0.871	0.721	0.908
19	0.920	0.915	0.753	0.906
20	0.900	0.870	0.828	0.909

a) Calcule as médias e variâncias amostrais para cada conjunto de dados;
b) Calcule os intervalos de confiança da média e da variância para cada conjunto de dados. Explicite as hipóteses usadas;
c) Aplique os testes cabíveis e verifique se as turmas são ou não equivalentes;
d) Verifique se os dados de cada grupo podem estar correlacionados aos dados dos demais;
e) Construa um gráfico na seguinte forma:



Para cada turma, verifique se há *outliers*; ou seja, pontos fora da região de confiança. Podem ser observadas tendências de aumento ou decréscimo de conversão?
f) Você mandaria alguma turma para treinamento?

4- Seja o conjunto de dados relativos à variável x_i retirados do computador com a rotina RANDOM:

	00	10	20	30	40
1	0.1025	0.2217	0.3737	0.8341	0.0910
2	0.1147	0.3344	0.4521	0.4298	0.9511
3	0.9508	0.1351	0.5811	0.6315	0.1223
4	0.7212	0.6227	0.9123	0.4726	0.8711
5	0.4393	0.5111	0.7314	0.6215	0.5661

6	0.6161	0.7502	0.3122	0.5871	0.6161
7	0.0012	0.8192	0.4659	0.2012	0.9813
8	0.1200	0.9095	0.2197	0.3191	0.6715
9	0.8837	0.0195	0.7382	0.4615	0.2328
10	0.4141	0.5823	0.1180	0.9867	0.9142

a) Calcule média e variância para a lista de medidas disponíveis.
b) Faça $z_i = x_i$ e $y_i = x_{i+1}$.. Calcule o coeficiente de correlação entre z e y. Você consegue observar alguma tendência?
c) Divida os dados em 10 classes, de forma que

$$Classe_1 = 0 \le x_i \le 0.10, ..., Classe_{10} = 0.9 \le x_i \le 1.00$$

Monte o histograma de freqüência das classes.
d) A distribuição obtida é supostamente uniforme. Os dados confirmam isso? Admitindo-se que

$$\wp(x) = \begin{cases} 0, & x < 0 \\ 1, & 0 \le x \le 1 \\ 0, & x > 1 \end{cases}$$

calcule a média e a variância esperadas.
e) As médias e variâncias obtidas podem ser consideradas equivalentes às teóricas? Quais os limites de confiança dos dados obtidos?

5- Suponha que um problema estocástico envolve duas variáveis sujeitas a flutuações normais. Suponha ainda que o vetor de médias e a respectiva matriz de covariâncias são dados por:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad , \quad \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad , \quad \mathbf{V}_X = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

a) Calcule a forma da região de confiança (faça $c = 1$ na Equação (3.48));
b) Calcule as direções principais e interprete os resultados;
c) Como você descreveria a região de confiança, com um nível de confiança correspondente a $c = 1$, onde você espera encontrar valores de x_1 e x_2?

$$x_1^{\min} \le x_1 \le x_1^{\max}$$
$$x_2^{\min} \le x_2 \le x_2^{\max}$$

6- Três valores medidos estão disponíveis: 1.0, 1.5 e 8.0.
a) Caracterize estatisticamente os dados;
b) Suponha que o experimentador desconfia do último valor medido. Que conselho você daria ao experimentador?
c) Admita que um quarto valor é obtido e é igual a 1.3. A sua opinião muda? E se o quarto valor obtido for igual a 5.0? E se for igual a 9.1?