

ANEXO III

CONCEITOS BÁSICOS DE ESTATÍSTICA

Existem funções cujo comportamento é perfeitamente previsível. Estas funções são denominadas *determinísticas*. A função $f(x) = 2x - 4$ é uma função determinística uma vez que seu valor está perfeitamente caracterizado quando x é definido. Funções determinísticas são muito empregadas em modelos matemáticos idealizados.

O mundo real não é composto apenas por funções determinísticas. Certas propriedades, como por exemplo a resistência mecânica de um material, a vida de uma lâmpada, a soma de dois dados honestos jogados ao acaso ou a temperatura máxima em Curitiba no mês de janeiro, variam de amostra para amostra. Um valor médio é obtido, porém é impossível prever exatamente qual o valor a ser encontrado na própria amostra a ser testada.

Funções que apresentam imprevisibilidade são denominadas de *aleatórias*. Como são imprevisíveis, não podem ser equacionadas através dos recursos usuais da matemática determinística. Ferramentas estatísticas são necessárias para tal.

III.1 Distribuição de Probabilidade

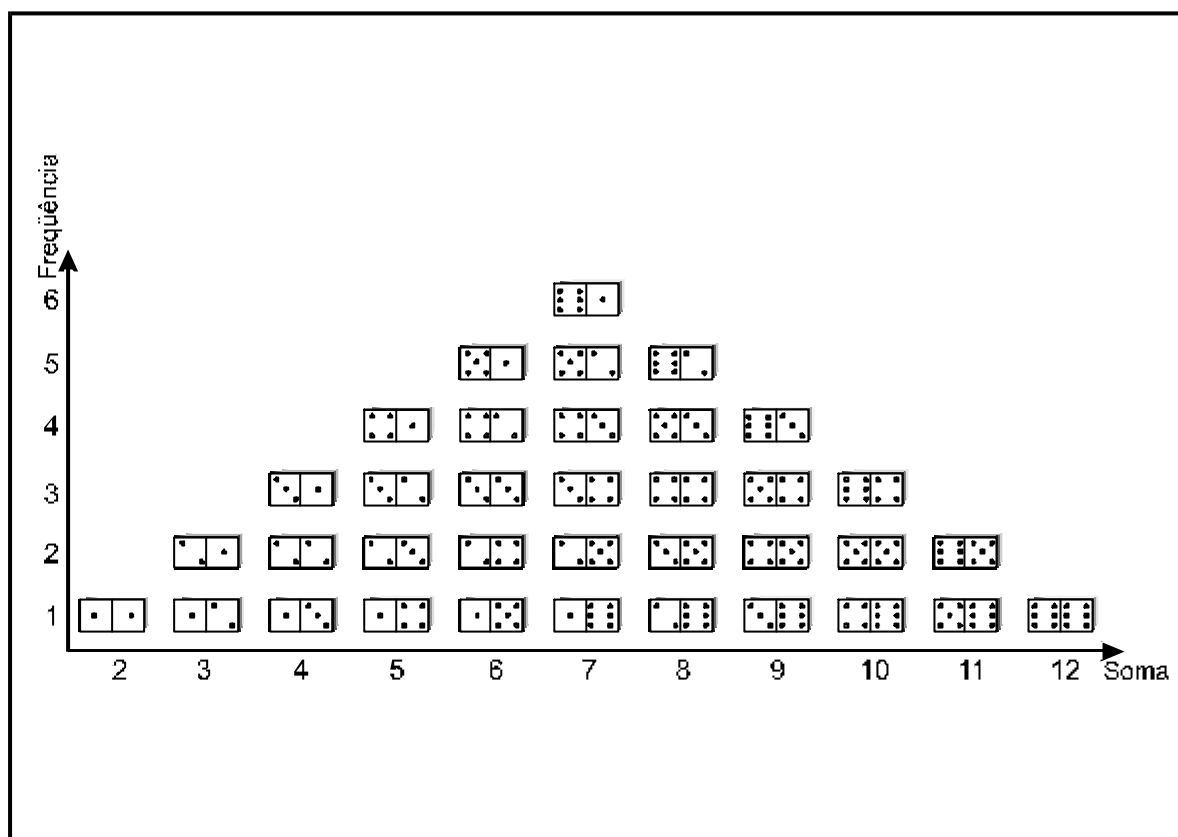
A soma de dois dados honestos pode resultar em qualquer número entre 2 e 12. Embora exista apenas uma única combinação de dados que resulte em 2 (1+1), nota-se que existem seis diferentes combinações de dados cuja soma resulta em 7 (1+6, 2+5, 3+4, 4+3, 5+2, 6+1). As chances de que a soma de dois dados lançados ao acaso resulte em 7 são maiores do que resultem em 2. Em outras palavras, a probabilidade de 7 ser obtido é maior do que 2.

A figura III.1 melhor caracteriza o universo das possíveis combinações dos dados que levam a cada soma. No eixo horizontal estão representados os valores possíveis para a soma, enquanto que no eixo vertical representa-se o número de combinações que resultam naquela soma, ou seja, a frequência com que aquele evento se manifesta. No total são 36 combinações possíveis.

Para determinar a probabilidade de que uma determinada soma seja obtida, é suficiente dividir o número de combinações que resultam naquela soma pelo número de combinações totais possíveis. A probabilidade de que 7 seja obtido como soma é de $6/36$ ou $1/6$. As chances de obter 8 são de $5/36$.

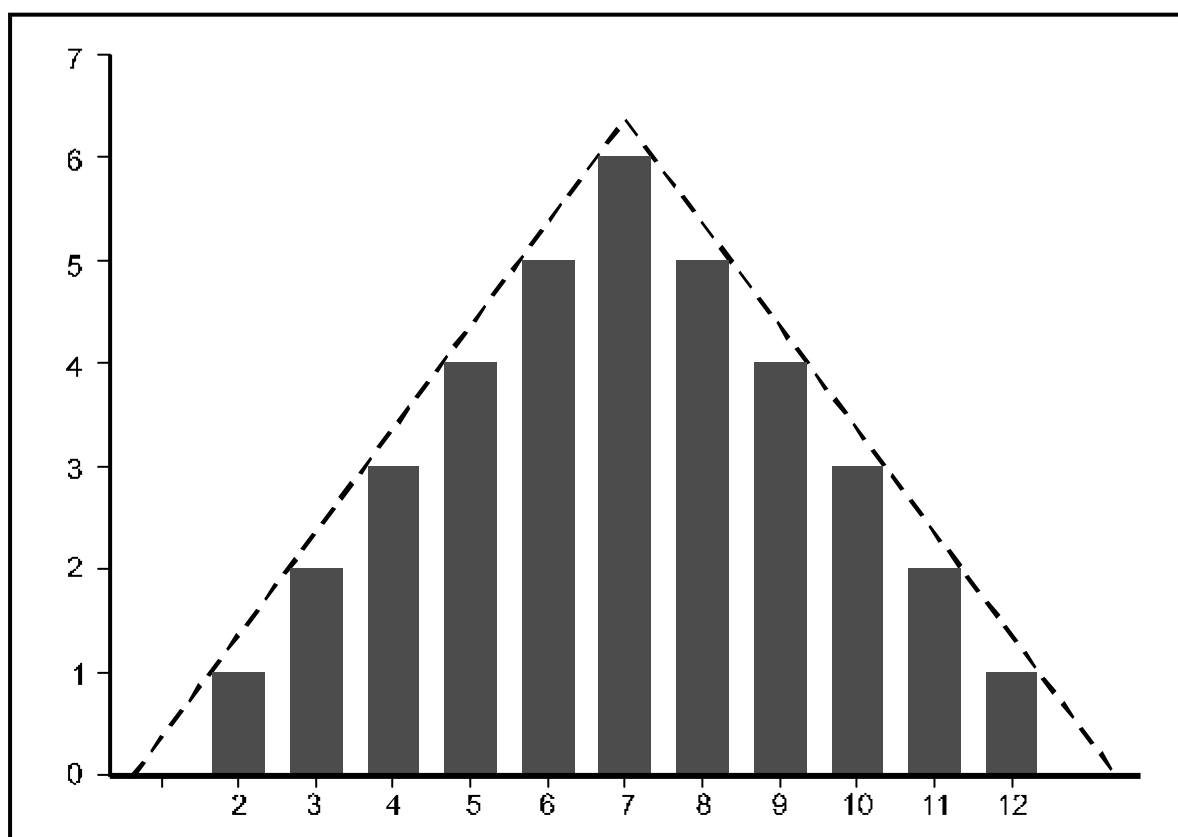
A probabilidade de que um valor situado dentro de uma faixa de valores seja obtido pode ser calculado pela soma das probabilidades individuais. Assim, as chances de que a soma esteja dentro da faixa 7 ± 1 é calculado por $5/36 + 6/36 + 5/36$, que são as probabilidades de se obter 6, 7 e 8 respectivamente, o que resulta em $16/36$ ou $4/9$. Verifica-se que as chances de que qualquer valor entre 2 e 12 seja obtido são de 1 (100%).

O gráfico da figura III.1 pode ter a frequência expressa em termos relativos. Para tal, divide-se a frequência de cada evento pelo número total de eventos do universo possível. No caso, divide-se cada frequência por 36. A figura III.2 mostra o gráfico resultante. Este gráfico das frequências relativas recebe o nome de *função densidade de probabilidade*, representada por $p(x)$, onde x



AAG - 11/97 - MCG 023

Figura III.1 - Combinações Possíveis para a Soma de Dois Dados Honestos



AAG - 11/97 - MCG 024

Figura III.2 - Densidade de probabilidade para a Soma de dois Dados Honestos

O representa cada evento envolvido e $p(x)$ a probabilidade deste evento ocorrer. No caso da soma de dois dados honestos, $p(7) = 1/6$, $p(6 \leq x \leq 8) = 4/9$, $p(-\infty < x < +\infty) = 1$.

A soma de dois dados é uma variável discreta, isto é, pode assumir apenas alguns valores inteiros e bem definidos. Porém, freqüentemente, encontra-se na natureza funções aleatórias contínuas, isto é, podem assumir qualquer valor real. Ao se analisar estatisticamente o comportamento de uma máquina ensacadeira que, idealmente, deveria empacotar 1,00 kg do produto por saco, verifica-se, na prática, que isto não ocorre sempre. Por imperfeições no seu mecanismo, sacos com massas, por exemplo entre 0,98 kg e 1,02 kg podem resultar. Embora seja muito difícil calcular teoricamente a função densidade de probabilidade desta ensacadeira, é possível determiná-la aproximadamente através de um grande número de observações experimentais.

O aspecto da função densidade de probabilidade de uma função aleatória contínua é uma curva contínua. A figura III.3.a ilustra $p(x)$ para uma ensacadeira com distribuição de probabilidade normal ou gaussiana. Nota-se que $p(x)$ é também uma função contínua. Neste caso, não há sentido em se determinar a probabilidade de que um determinado valor real venha a ocorrer, mas apenas de que faixas venham a ocorrer. Por exemplo, para determinar as chances de que sacos $1,00 \pm 0,02$ kg sejam obtidos determina-se a área abaixo da curva $p(x)$, representada por $P(0,98 \leq x \leq 1,02)$ ¹, entre estes limites, isto é:

$$P(0,98 \leq x \leq 1,02) = \int_{x=0,98}^{x=1,02} p(x) dx \quad (\text{III.1})$$

Deve-se notar uma importante propriedade de $p(x)$: $P(-\infty < x < +\infty) = 1$, isto é, a integral de $p(x)$ entre os limites $-\infty$ e $+\infty$, que corresponde à probabilidade de x estar dentro destes limites, sempre resulta em 1.

A figura III.b apresenta a função densidade de probabilidade de outra ensacadeira com características diferentes. Nota-se que, embora a área total sob $p_b(x)$ seja também unitária, esta é uma curva mais fechada que $p_a(x)$. A máquina que possui $p_b(x)$ apresenta maior probabilidade de resultar sacos com valores mais próximos do ideal que a primeira, portanto é uma máquina melhor. Já a máquina que possui $p_c(x)$ é a pior de todas por apresentar probabilidade relativamente altas de que valores que se afastam bastante do ideal venham a ocorrer. A característica que diferencia estas três ensacadeiras é a chamada *dispersão* que é maior quanto maior for o "espalhamento" da curva $p(x)$, isto é, a dispersão de $p_c(x)$ é maior que a dispersão de $p_b(x)$.

O desvio padrão (σ) é um parâmetro estatístico empregado para medir a dispersão de uma função aleatória. É tanto maior quanto maior for a dispersão. No caso da figura III.3 é evidente que $\sigma_c > \sigma_a > \sigma_b$. σ é calculado por:

$$s = \lim_{n \rightarrow \infty} \sqrt{\frac{\sum_{i=1}^{\infty} (x_i - m)^2}{n}} \quad (\text{III.2})$$

onde:

x_i é o valor do evento "i"

μ é o valor médio de todos os eventos

¹ Aqui o símbolo $p(x)$ é empregado para a função densidade de probabilidade enquanto $P(y)$ representa a probabilidade do evento y ocorrer.

Outro parâmetro importante que caracteriza uma função aleatória é o seu valor central, isto é, seu valor médio (μ). μ é calculado por:

$$m = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{\infty} x_i \quad (\text{III.3})$$

III.2 Distribuição Normal

Uma das distribuições estatísticas mais comumente encontradas na prática é a distribuição *normal* ou *gaussiana*. O *teorema do limite central* demonstra que a combinação de um grande número de fatores de natureza aleatória, com qualquer distribuição, aproxima-se da distribuição normal à medida que aumenta o número dos fatores envolvidos. A forma da função densidade de probabilidade $p(x)$ da distribuição normal assemelha-se a de um sino, como mostrada na figura III.4. Apresenta simetria em torno do valor central (médio). O desvio padrão desta distribuição corresponde à distância entre o valor central e o ponto de inflexão de $p(x)$, isto é, o ponto onde a segunda derivada de $p(x)$ é zero. Sua função densidade de probabilidade é:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

onde:

$$z = \frac{x - m}{s}$$

e

μ é o valor médio
 σ é o desvio padrão

A distribuição das dimensões de um lote de peças fabricadas por uma máquina, a distribuição em um alvo de tiros dados por um atirador, os erros de medição e a temperatura média do dia 21 de abril de cada ano são exemplos de distribuições normais.

O cálculo da probabilidade de que uma dada função aleatória com distribuição normal esteja dentro de uma faixa de valores é também calculada pela equação (III.1), isto é, pela integral definida de $p(x)$ entre os limites considerados. No caso da distribuição normal não se pode exprimir a integral de $p(x)$ como uma função simples. É comum encontrar esta integral na forma de tabelas normalizadas. Entretanto, existem alguns valores particulares que, por serem muito empregados na prática, devem ser citados.

Se tratando de uma função aleatória com distribuição normal, valor médio μ e desvio padrão σ , é possível calcular as seguintes probabilidades:

$$\begin{aligned} P(\mu - \sigma < x < \mu + \sigma) &= 0.6826 \\ P(\mu - 3\sigma < x < \mu + 3\sigma) &= 0.9973 \\ P(\mu - 1.96\sigma < x < \mu + 1.96\sigma) &= 0.95 \\ P(\mu - 2.58\sigma < x < \mu + 2.58\sigma) &= 0.99 \\ P(\mu - 3.30\sigma < x < \mu + 3.30\sigma) &= 0.999 \end{aligned}$$

III.3 A Natureza Aleatória do Erro de Medição

Sabe-se que é impossível efetuar uma medição absolutamente isenta de erros. Seja em função do sistema de medição ou em função do mensurando ou do operador, o erro de medição está

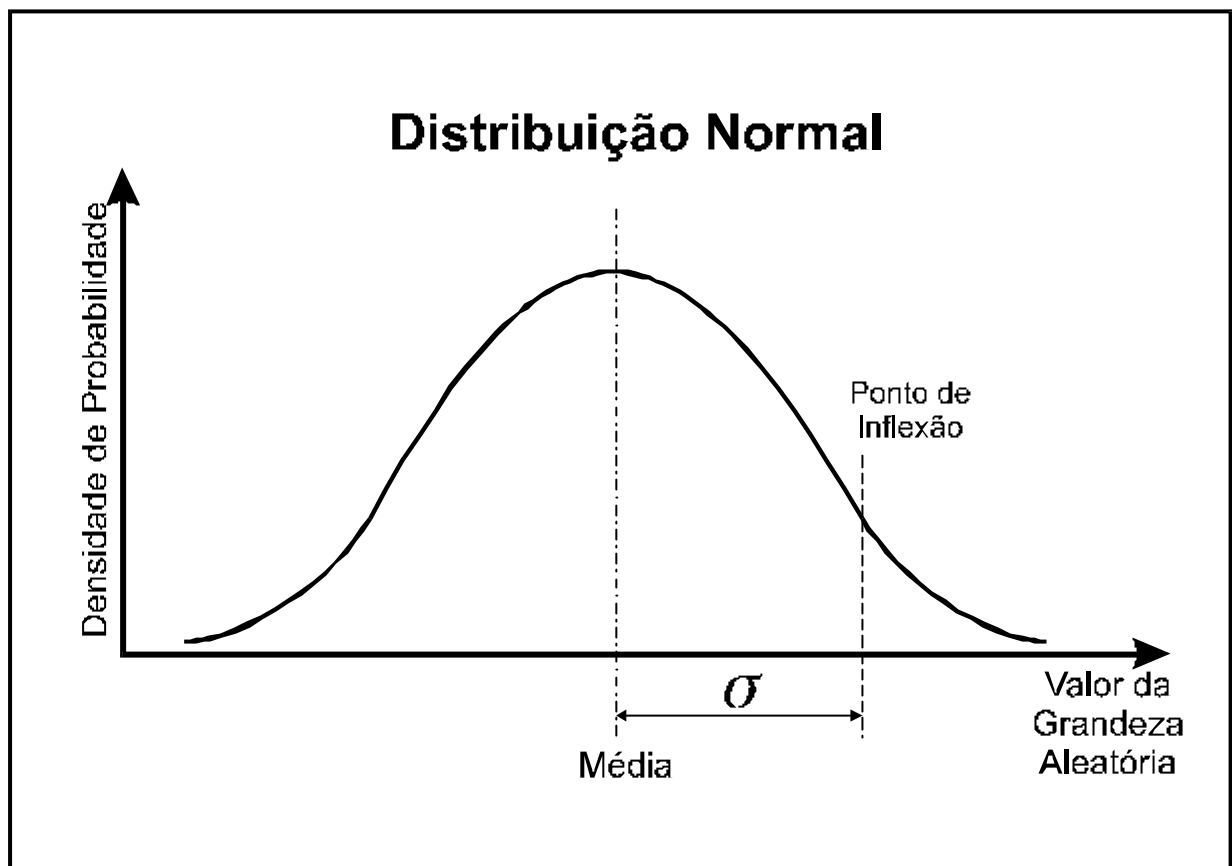


Figura III.3 - Distribuição Normal ou Gaussiana

AAG - 1º/97 - MCG 025

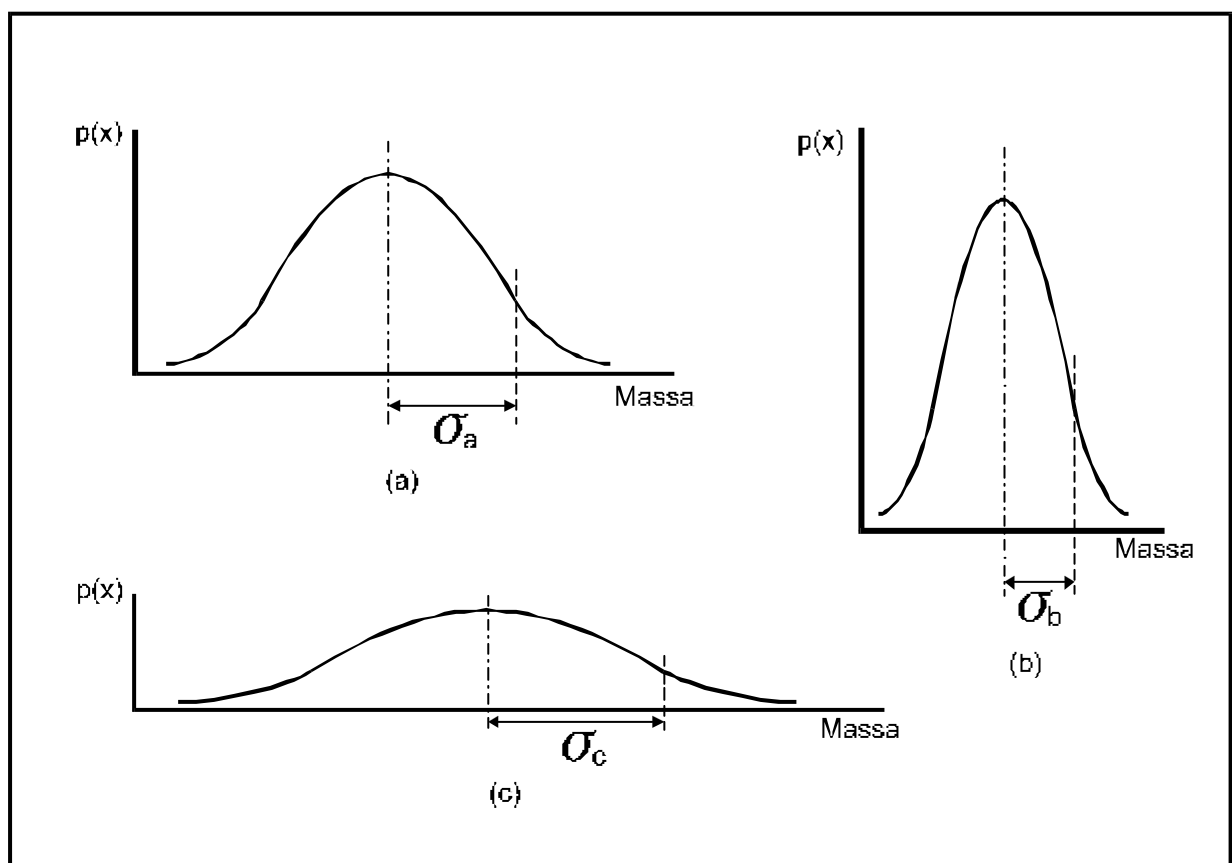


Figura III.4 - Densidade de Probabilidade para Três Ensacadeiras

a) Ensacadeira normal; b) Ensacadeira com pequena dispersão; c) Ensacadeira com grande dispersão

AAG - 1º/97 - MCG 025

sempre presente. Ao se repetir a medição de um mensurando invariável, com o mesmo sistema de medição e nas mesmas condições, como por exemplo a medição repetitiva da massa de uma peça com a mesma balança, verifica-se, com frequência, que o valor obtido não se repete.

O erro de medição presente em cada indicação pode ser determinado pela diferença entre a indicação e o valor verdadeiro convencional, isto é, $E = I - VVC$. Em um SM ideal, este erro deveria ser sempre nulo. Porém, nota-se que este erro é na verdade uma função aleatória com distribuição aproximadamente normal.

O valor médio do erro de medição é o *erro sistemático* (E_s), que só poderia ser determinado baseada em um número infinito de observações, por:

$$E_s = MI - VVC$$

onde:

$$MI = \frac{1}{n} \sum_{i=1}^n I_i$$

(III.5)

e

MI é a média de infinitas indicações
VVC é o valor verdadeiro convencional

Se um número finito de observações é envolvido, a equação (III.5) pode ainda ser usada para estimar o erro sistemático. Neste caso, esta estimativa recebe o nome específico de *tendência* (T_d).

A parcela aleatória do erro de medição é simplesmente chamada de *erro aleatório*. Tratando-se de uma função aleatória, cada valor medido possui um erro aleatório diferente, e dado por: $E_{aj} = I_j - MI$. A sua caracterização é realizada através da medida da dispersão da distribuição normal associada, isto é, do desvio padrão (σ).

Define-se a *repetitividade* (Re), como sendo a faixa que, com uma probabilidade estatística definida, conterá o erro aleatório. É comum adotar a probabilidade de 95% como aceitável para a Re ($\hat{\epsilon}$). Assim, 95% dos erros aleatórios estarão dentro desta faixa. A Re é estimada por:

$$Re(95\%) = \pm 1,96 \, s$$

(III.6)

Porém, como será visto no próximo item, a estimação de σ não é tão direta.

III.4 Amostra Versus População

Os conceitos de média (μ) e desvio padrão (σ) são válidos para uma função aleatória. Para caracterizá-los perfeitamente pelas equações III.2 e III.3 é necessário envolver um número infinito de valores observados desta função, isto é, toda a *população*.

Na prática, não se tem tempo para coletar um número infinito de valores. É comum considerar apenas uma *amostra* de n valores desta população. A média e o desvio padrão da população são *estimados* a partir da média, do desvio padrão e do tamanho da amostra. A média e o desvio padrão da amostra são calculados por:

² Alguns autores adotam 99,7%, o que corresponde a $\pm 3\sigma$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{III.7})$$

e

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (\text{III.8})$$

Esta estimativa só é confiável para valores grandes de n . Se amostras pequenas são envolvidas ($n < 200$), é necessário aplicar um coeficiente de correção (t) conhecido como *coeficiente t-Student*. O coeficiente t -Student é função da probabilidade de enquadramento desejada (P) e do tamanho da amostra (n). A figura III.5 apresenta valores tabelados para “ t ” como função de n e de P . Assim, a repetitividade associada ao erro aleatório pode ser estimada por:

$$Re = \pm t \cdot s \quad (\text{III.9})$$

A média verdadeira da população (μ), calculada a partir dos parâmetros da amostra, não pode ser determinada exatamente. Alguma incerteza ainda resultará. Pode-se mostrar que a média da população estará situada dentro da seguinte faixa, determinada de *intervalo de confiança da média*:

$$\bar{x} \pm \frac{t}{\sqrt{n}} s = \bar{x} \pm \frac{Re}{\sqrt{n}} \quad (\text{III.9})$$

onde:

$\bar{x}$ é a média da amostra
 s é o desvio padrão da amostra
 t é o coeficiente t -Student
 n é o tamanho da amostra

III.5 Outras Distribuições Estatísticas

Existem situações na prática onde é conveniente modelar certos efeitos ou fenômenos por meio de outras distribuições distintas da normal. Neste texto, não será discutida a aplicabilidade das diversas distribuições em problemas de metrologia.

a) Distribuição retangular

É caracterizada por apresentar a mesma densidade de probabilidade para todos os valores dentro dos limites dados por “ $m - a$ ” e “ $m + a$ ”, e zero fora destes (figura III.6). Seu desvio padrão é dado por:

$$s = \frac{a}{\sqrt{3}}$$

b) Distribuição triangular

É caracterizada por apresentar máxima probabilidade para o valor médio e decrescer linearmente até zero nos limites dados por “ $m - a$ ” e “ $m + a$ ”, e zero fora destes (figura III.7). Seu desvio padrão é dado

por: $s = \frac{a}{\sqrt{6}}$

Número de valores Individuais (n)	Valores do coeficiente t de Student			
	P = 68,3% 1,00 σ	P = 95,0% 1,96 σ	P = 99,0% 2,58 σ	P = 99,7% 3,00 σ
2	1.839	12.706	63.656	235.774
3	1.322	4.303	9.925	19.206
4	1.198	3.182	5.841	9.219
5	1.142	2.776	4.604	6.620
6	1.111	2.571	4.032	5.507
7	1.091	2.447	3.707	4.904
8	1.077	2.365	3.499	4.530
9	1.067	2.306	3.355	4.277
10	1.059	2.262	3.250	4.094
11	1.053	2.228	3.169	3.957
12	1.048	2.201	3.106	3.850
13	1.044	2.179	3.055	3.764
14	1.041	2.160	3.012	3.694
15	1.038	2.145	2.977	3.636
16	1.035	2.131	2.947	3.586
17	1.033	2.120	2.921	3.544
18	1.031	2.110	2.898	3.507
19	1.029	2.101	2.878	3.475
20	1.028	2.093	2.861	3.447
21	1.026	2.086	2.845	3.422
22	1.025	2.080	2.831	3.400
23	1.024	2.074	2.819	3.380
24	1.023	2.069	2.807	3.361
25	1.022	2.064	2.797	3.345
26	1.021	2.060	2.787	3.330
27	1.020	2.056	2.779	3.316
28	1.020	2.052	2.771	3.303
29	1.019	2.048	2.763	3.291
30	1.018	2.045	2.756	3.280
40	1.014	2.023	2.708	3.204
60	1.009	2.001	2.662	3.132
120	1.005	1.980	2.618	3.064
∞	1.000	1.960	2.580	3.000

**Figura III.5 - Tabela de Coeficientes de Student
para vários níveis de probabilidade**

AAG - 05/99 MCG 041

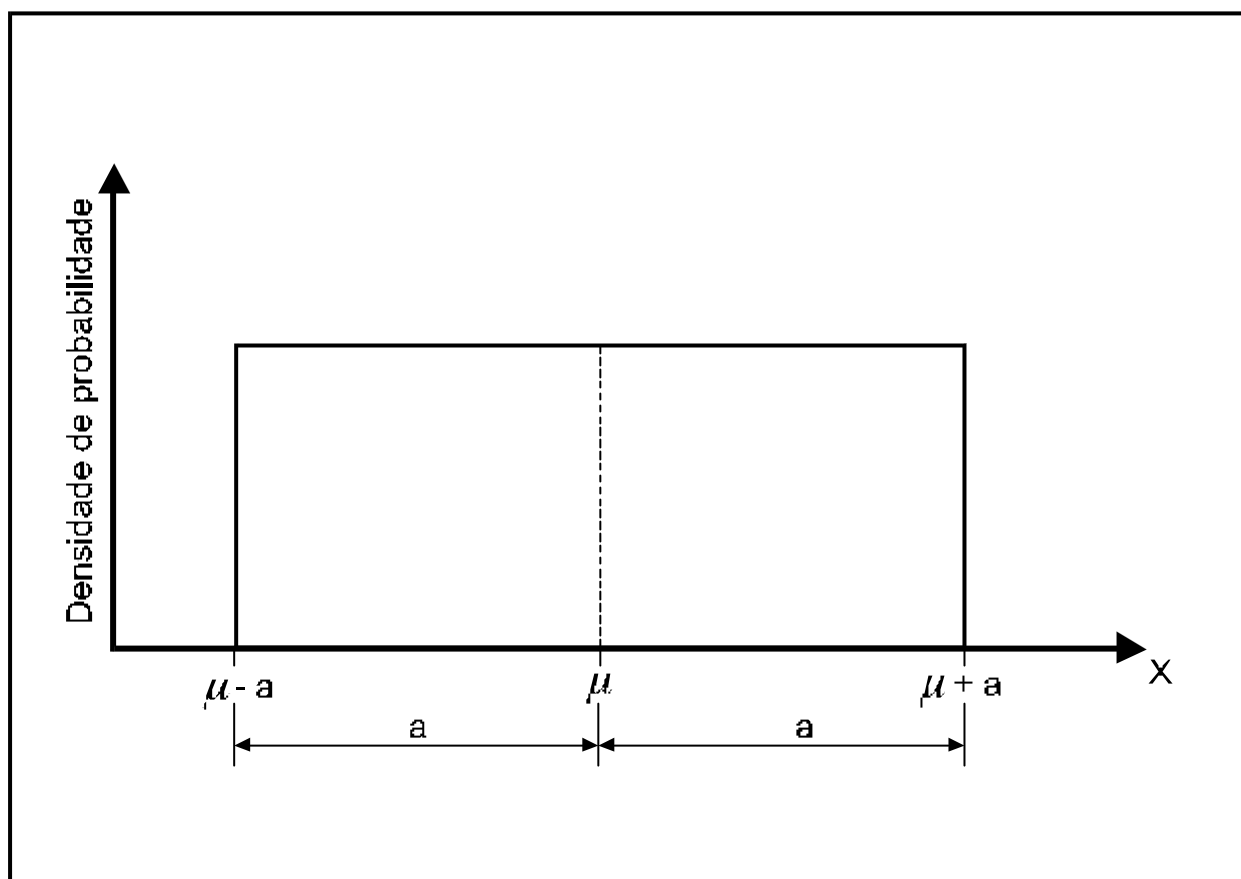


Figura III.6 - Distribuição de Probabilidade Retangular

AAG - 11/97 - MCC 015

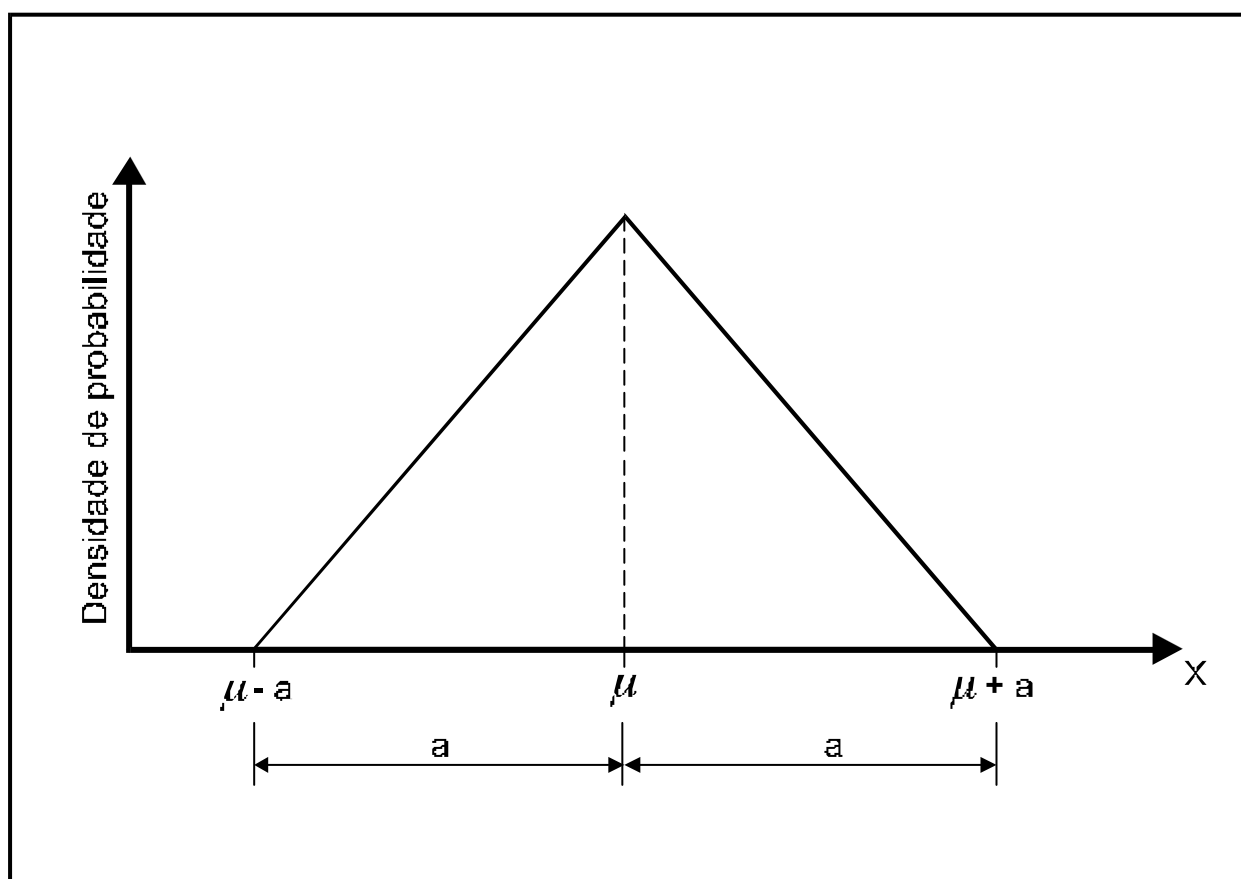


Figura III.7 - Distribuição de Probabilidade Triangular

AAG - 10/97 - MCC 027

Exercícios para o Anexo III

III.1. Classifique as variáveis abaixo como determinísticas ou aleatórias:

- A) A distância indicada no odômetro de automóveis que percorrem o trecho Florianópolis-Curitiba pela mesma estrada
- B) O horário do pôr-do-sol de uma mesma cidade ao longo do ano
- C) A massa de “uma pitada de sal” que uma cozinheira acrescenta todo dia no feijão
- D) A vida de uma lâmpada de 60W de um mesmo lote de fabricação
- E) O seno do terço do quadrado de um número

III.2. Qual a probabilidade da soma de três dados honestos estar entre 5 e 7 inclusive?

III.3. Classifique as seguintes variáveis aleatórias como discretas ou contínuas:

- A) A massa de uma “pitada de sal”
- B) A medida da massa de uma “pitadas de sal” obtida de uma balança com resolução de 0,1g
- C) A vida de lâmpadas de um mesmo lote de fabricação
- D) O tempo expresso em horas correspondente à vida de uma pessoa do sexo masculino residente em uma dada cidade
- E) As várias medidas efetuadas da massa de uma mesma peça, efetuadas pela mesma balança

III.4. Sendo $p(x)$ a função densidade de probabilidade de uma variável aleatória contínua, determine expressões para o cálculo da probabilidade desta variável:

- A) Ser maior que o valor x_z
- B) Ser menor que o valor x_b
- C) Sendo $x_a > x_b$, estar entre estes dois valores
- D) Sendo $x_a > x_b$, ser maior que x_a ou menor que x_b

III.5. Qual a probabilidade de uma variável aleatória com distribuição normal com média 18,00g e desvio padrão 0,12g situar-se dentro da faixa $18,00 \pm 0,36g$? e na faixa $18,24 \pm 0,12g$?

III.6. Calcule a média amostral, o desvio padrão amostral e o intervalo de confiança do erro aleatório dos dados abaixo. Calcule também o intervalo de confiança dentro do qual estará a média verdadeira população completa (equação III.9)

12,8	12,5	13,0	13,1
12,6	12,9	13,1	12,8
12,3	12,8	12,6	12,7

III.7. Supondo que os dados da questão anterior referem-se à calibração de uma balança onde a mesma massa padrão de $12,500 \pm 0,002g$ foi medida diversas vezes, o que possível afirmar sobre o erro sistemático e sua incerteza?