

2 CONCEITOS BÁSICOS DE PROBABILIDADE E ESTATÍSTICA

2.1. Introdução

As propriedades físicas e mecânicas de solos e rochas são em geral naturalmente dispersas. Os métodos probabilísticos podem ajudar na análise destas dispersões, facilitando o entendimento dos dados. Neste capítulo são fornecidos conceitos básicos de probabilidade ligados à análise geotécnica. Em especial, serão abordadas as aplicações as análises de estabilidade de taludes.

Estes conceitos são detalhados em livros clássicos de probabilidade e estatística, tais como Ang e Tang (1975), e mais recentemente, Baecher e Christian (2003).

2.2. Variáveis Aleatórias

Uma variável é aleatória quando puder assumir diferentes valores como resultado de um experimento aleatório. As variáveis aleatórias se dividem em discretas e contínuas:

- Variável Aleatória Discreta: admite apenas um número limitado de valores.
- Variável Aleatória Contínua: admite qualquer valor dentro de um intervalo dado, ou pode tomar todos os valores de um intervalo.

Na maioria das situações na engenharia geotécnica as variáveis utilizadas são as variáveis aleatórias contínuas.

2.3. Função Densidade de Probabilidade

Para uma variável aleatória contínua, a função utilizada para representar a distribuição de probabilidade é a função densidade de probabilidade ($f_x(x)$). Esta função descreve a forma da curva de distribuição da probabilidade de ocorrência de cada valor da variável aleatória contínua.

Existem vários tipos de função de densidade de probabilidade, sendo que as distribuições mais utilizadas em engenharia são: Normal, Log-normal, Exponencial, Gamma, Beta. Dentre estas distribuições, as mais usadas em engenharia geotécnica são a distribuição Normal (ou gaussiana) e a distribuição Log-normal.

Uma forma alternativa de poder visualizar os dados (análogo à função de densidade de probabilidade) é mediante uma função de distribuição cumulativa (CDF), que indica a probabilidade de uma variável ter um valor menor ou igual a um valor selecionado (figura 2.1), na realidade CDF é a integral da correspondente função de densidade de probabilidade.

A função cumulativa da densidade de probabilidade ($F_x(x)$), e definida por:

$$F_x(x_i) = P(x \leq x_i) = \int_{-\infty}^{x_i} f_x(x) \partial x \quad (2.1)$$

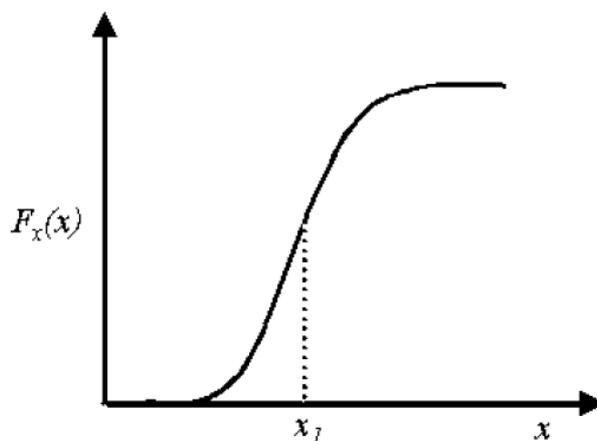


Figura 2.1 – Função de distribuição cumulativa.

2.3.1. Distribuições de Variáveis Aleatórias

Neste trabalho, são apresentadas a distribuição Normal (ou gaussiana) e Log-normal, por serem as mais usadas em engenharia geotécnica.

2.3.1.1. Distribuição Normal

A distribuição Normal é a mais familiar das distribuições de probabilidade e também uma das mais importantes em estatística. Esta distribuição tem a forma de sino (Figura 2.2).

A equação da curva Normal é especificada usando-se dois parâmetros: a média populacional μ e o desvio padrão populacional σ , (ou a variância populacional σ^2). Denota-se $N(\mu, \sigma^2)$ a curva Normal com média μ e variância σ^2 . A média refere-se ao centro da distribuição, e o desvio padrão (ou variância) ao espalhamento de curva. A distribuição normal é simétrica em relação à média.

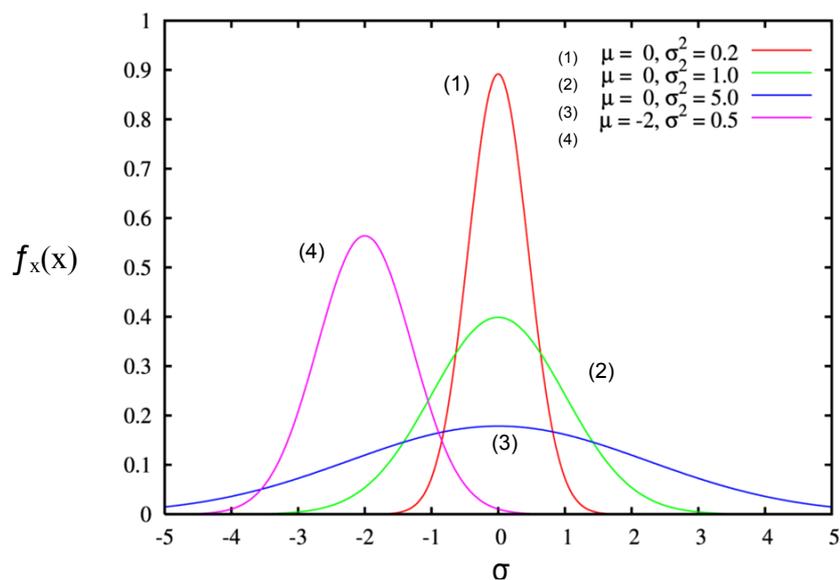


Figura 2.2 - Variação de uma densidade de probabilidade normal, para vários valores de μ, σ^2 .

A equação que descreve o comportamento da função de densidade normal é:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (2.2)$$

A área sob a curva normal é sempre unitária. Então, para quaisquer dois valores específicos pode-se determinar a proporção de área sob a curva entre esses dois valores. Para a distribuição Normal, a proporção de valores abrangida por um, dois ou três desvios padrão da média, é indicada na Tabela 2.1.

Tabela 2.1 – Proporção de área sobre a distribuição normal.

Faixa de valores	Proporção
$\mu \pm 1\sigma$	68,3%
$\mu \pm 2\sigma$	95,5%
$\mu \pm 3\sigma$	99,7%

Pode-se observar nesta tabela que o 99,7% de todos os dados estão dentro de $\pm 3\sigma$. Esta é a principal justificação da “*Three-Sigma Rule*”, proposta por Daí e Wang (1992) e usada por Duncan (1999, 2000) para obter o desvio padrão de uma amostra.

Na prática deseja-se calcular a probabilidade de uma variável (por exemplo, o fator de segurança FS) ser menor ou igual a certo valor (por exemplo, 1,0). Para isso, a variável x cuja distribuição é $N(\mu, \sigma^2)$ e transformada numa forma padronizada Z com distribuição $N(0,1)$.

Onde Z é dada por:

$$Z = \frac{x - \mu}{\sigma} = \left(\frac{FS - E(FS)}{\sigma_{FS}} \right) \quad (2.3)$$

Z é a probabilidade de que FS possa ser menor ou igual a qualquer número (por exemplo, $FS = 1$).

2.3.1.2. Distribuição Log-Normal

Uma variável aleatória x tem a distribuição Log-Normal quando seu logaritmo possuir a forma de uma distribuição Normal. Uma das características mais importante desta distribuição é não admitir dados negativos.

A função de densidade está caracterizada por ter dois parâmetros, média populacional μ , e o desvio padrão populacional σ , a equação que caracteriza esta função de densidade é:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{\ln(x) - \mu^2}{2\sigma^2}\right] \quad (2.4)$$

A Figura 2.3 ilustra algumas distribuições Log-Normal com diferentes valores de σ .

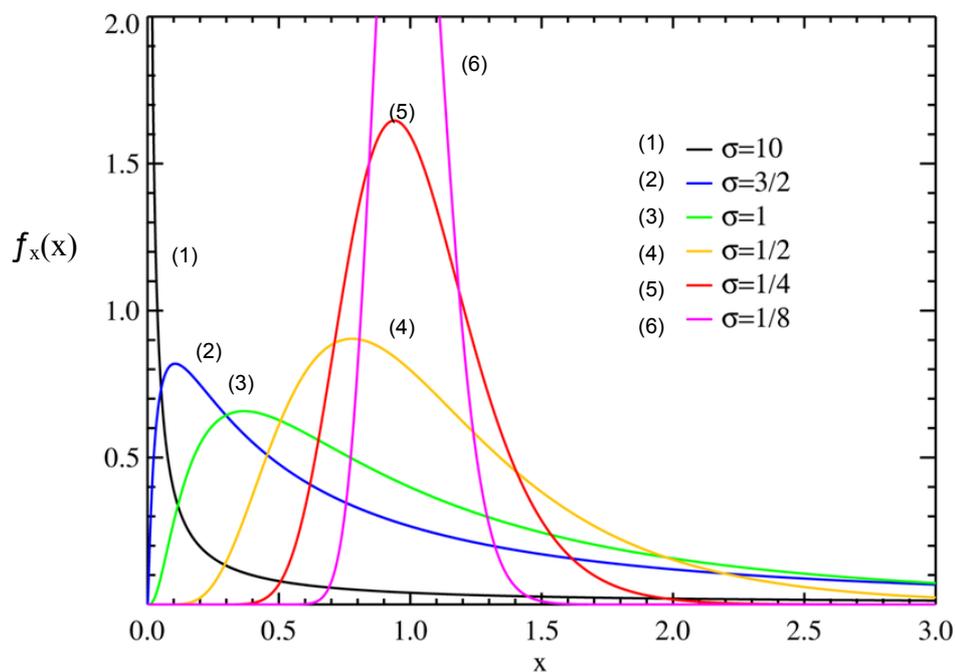


Figura 2.3 - Distribuições log-normal da variável x .

2.3.1.3. Distribuição de Parâmetros Geotécnicos

Como foi mencionado, a distribuição Log-Normal é sempre positiva, como os parâmetros usados em problemas geotécnicos, por exemplo, o valor do fator de segurança, sempre podem ter valores positivos, então esta distribuição é usualmente adequada.

Baecher e Christian (2003) demonstraram que a probabilidade do fator de segurança ser negativo quando se usa uma distribuição normal é muito baixa, pelo que a distribuição normal também é válida para problemas geotécnicos.

Na realidade, deveriam se fazer provas de adequabilidade para poder determinar o melhor ajuste dos dados. As mais usadas são:

- Método Chi Quadrado;
- Método Kolmogorov Smirnov;
- Método Andersen Darling.

2.4. Tratamento Estatístico dos Dados

2.4.1. Análise Gráfico da Amostra

A maneira mais comum de representar os dados é mediante um histograma, também denominada gráfico de barras, onde os dados são agrupados em intervalos. A altura da barra em cada intervalo, mostra a quantidade de valores dentro deste intervalo. A Figura 2.4 apresenta um gráfico de barras típico.

Pode-se observar que o histograma é dividido em intervalos de igual largura. A dificuldade que se encontra na hora de fazer um histograma é definir o valor dos intervalos quando se tem poucos dados. Sturges (1926) sugere o uso da equação 2.5 para definir o número de intervalos (k), onde n é o número de dados.

$$k = 1 + 3.3 \log_{10} n \quad (2.5)$$

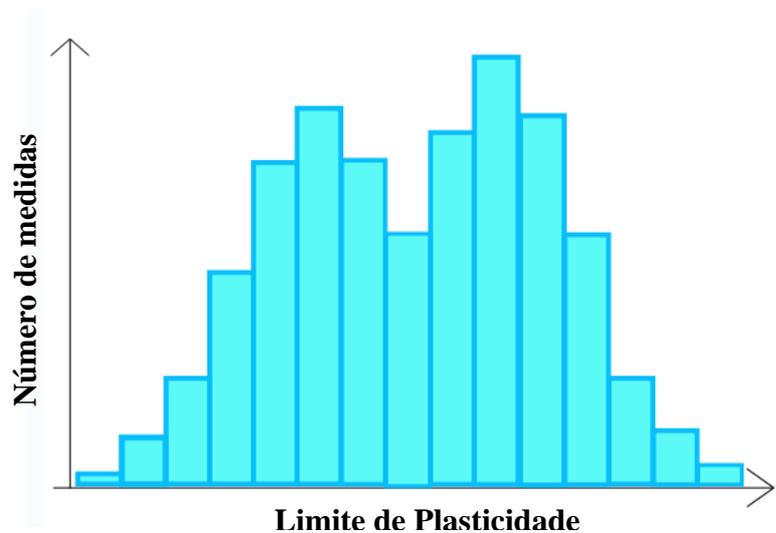


Figura 2.4 – Histograma típico.

O uso de histogramas é muito limitado quando se dispõe de um número pequeno de dados. Recomenda-se por tanto o uso de funções de distribuição de probabilidade (Normal ou Log-normal), ou ainda o uso de funções de distribuição cumulativa (CDF).

2.4.2. Análise da Amostra

Os parâmetros comumente utilizados para a análise da amostra são a média e a variância da amostra.

A média da amostra também denominada média aritmética é representado por:

$$E(x) = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.6)$$

A média da amostra indica o centro de gravidade de uma distribuição de probabilidade.

A variância da amostra relaciona-se com os quadrados dos desvios da variável x em relação à média $E(x)$, sendo definida por:

$$V(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - E(x))^2 \quad (2.7)$$

Onde o desvio padrão da amostra (σ) é definido como a raiz quadrada da variância, denotado por:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - E(x))^2} \quad (2.8)$$

O desvio padrão expressa a dispersão de uma variável aleatória com respeito ao valor esperado ou a média.

Uma maneira mais conveniente de expressar a dispersão dos dados é com o coeficiente de variação da amostra $Cov(X)$, definido pela equação 2.9.

$$Cov(X) = \frac{\sigma(x)}{E(x)} \quad (2.9)$$

Uma estimativa rápida do desvio padrão pode ser alcançada mediante o uso da amplitude da amostra (r_x), que é definida por:

$$r_x = x_{\max} - x_{\min} \quad (2.10)$$

Onde r_x é a diferença entre os valores máximo e mínimo da amostra. Se os dados são normalmente distribuídos, pode-se obter uma estimativa de σ , com base na amplitude da amostra r_x e num fator N_n , o qual é definido por:

$$\sigma = N_n (x_{\max} - x_{\min}) \quad (2.11)$$

A equação 2.11 foi proposta por Burington e May (1970) e o fator N_n pode ser obtido pela Tabela 2.2, onde n é o número de amostras.

Tabela 2.2 – Valores do fator N_n para estimar o desvio padrão, Burintong e May (1970).

n	N_n	n	N_n	n	N_n
2	0,886	11	0,315	20	0,268
3	0,510	12	0,307	30	0,244
4	0,486	13	0,300	50	0,222
5	0,430	14	0,294	75	0,208
6	0,395	15	0,288	100	0,199
7	0,370	16	0,283	150	0,19
8	0,351	17	0,279	200	0,18
9	0,337	18	0,275		
10	0,325	19	0,271		

Este método de estimativa rápida do desvio padrão somente pode ser usado quando a dispersão dos dados é simétrica (Baecher e Christian, 2003), devido a ser afetado fortemente pelos extremos. Para dispersões assimétricas, este método não é recomendado. Em geral, as propriedades geotécnicas são simétricas, no caso de propriedades hidráulicas, a distribuição é usualmente assimétrica.

Duncan (1999, 2000) descreveu o uso da “*Three Sigma Rule*” proposta por Daí e Wand (1992), para o cálculo do desvio padrão em geotecnia. É baseada no fato de que o 99.7% (Figura 2.5) dos dados normalmente distribuídos estão dentro de $\pm 3\sigma$, sendo o desvio padrão (σ) obtido pela expressão 2.12.

$$\sigma = \frac{HCV - LCV}{6} \quad (2.12)$$

Onde HCV é o maior valor concebível para o parâmetro em questão e LCV é o menor valor concebível (ou possível).

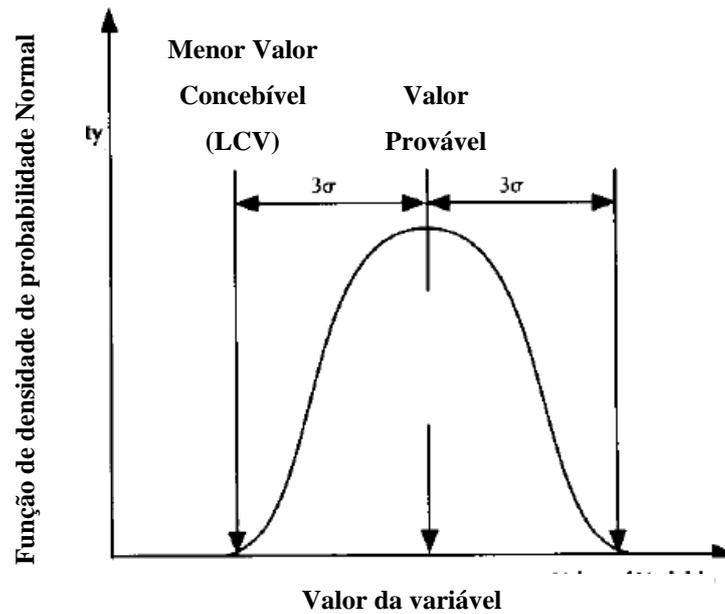


Figura 2.5 – “Three Sigma Rule”, Daí e Wand (1992).

Duncan (2000) estendeu o uso da “Three Sigma Rule” a um método gráfico, como apresentado na figura 2.7, que é um caso prático encontrado na geotecnia.

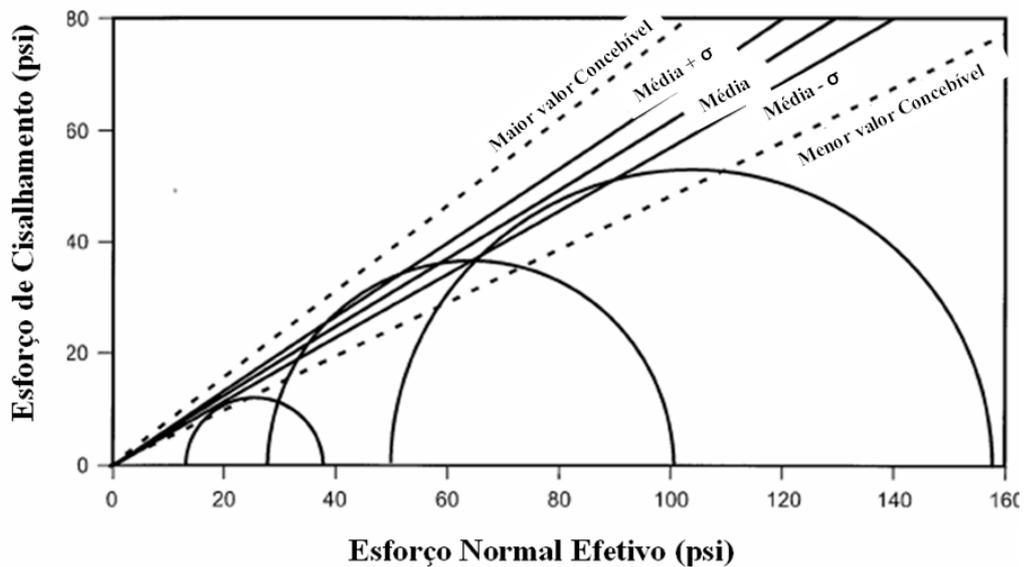


Figura 2.6 – “Three Sigma Rule” aplicado graficamente na obtenção dos parâmetros de resistência c e ϕ , Duncan (2000).

Duncan (2000) sugere o uso da “*Three Sigma Rule*” para obtenção do desvio padrão, e pode ser usada quando dados limitados são fornecidos ou quando nenhum dado é encontrado, e também para julgar razoáveis coeficientes de variação publicados na literatura.

O método de “*Three Sigma Rule*” foi criticado por Christian e Baecher (2001) devido ao fato de que a obtenção dos valores mais alto e mais baixo (HCV, LCV), dos parâmetros requer engenheiros com bom senso e experiência.

Utilizando a equação 2.11 para uma obtenção rápida do desvio padrão na forma:

$$\sigma = \frac{(x_{\max} - x_{\min})}{N_{n^*}} \quad (2.13)$$

Para 10 ensaios (n=10) o valor de N_{n^*} é 3,078 (inverso de N_n na Tabela 2.2). Ao usar a “*Three Sigma Rule*” (equação 2.12), o valor de N_{n^*} tem que ser igual a 6,0. Duncan (2001) sugere que o uso da “*Three Sigma Rule*” tem que ser acompanhado com a Tabela 2.2 e requer um engenheiro experiente, na determinação do provável desvio padrão do parâmetro.

2.4.3. Análise da Correlação entre Variáveis

Em qualquer análise geotécnica é usual se lidar com mais de uma variável aleatória (coesão, ângulo de atrito, peso específico do solo, poropressões, largura da camada, etc.). A incerteza de uma variável pode estar associada à incerteza de outra, não sendo independente, o que pode afetar significativamente o resultado da análise.

O coeficiente de correlação entre duas variáveis define como a variação em um parâmetro pode afetar o valor da outra variável.

Vieira (1999) comenta sobre o uso indiscriminado, ou o mau uso que pode ser dado à interpretação da correlação entre variáveis. Segundo o autor, a

correlação entre duas variáveis nem sempre significa uma relação de causa e efeito. Muitas vezes existe uma terceira variável, não estudada, que determina tanto os aumentos em x como os aumentos (ou diminuições) em y .

Portanto, a correlação entre variáveis não indica que uma causa a outra, mostra apenas que existe uma relação linear acidental entre elas.

Wolff e Harr (1987), Mostyn e Li (1993), e Mais, Giasi e Cherubini (2003) mostraram a grande sensibilidade que existe na probabilidade de ruptura para a consideração de algum tipo de coeficiente de correlação.

No caso dos parâmetros de resistência de Mohr Coulomb (c , θ), é atribuído algum tipo de correlação. Quanto maior é o ângulo de atrito (θ), menor pode ser a coesão (c). Assim poderia haver uma correlação negativa, e a incerteza no ângulo de atrito estaria muito influenciada na incerteza da coesão.

O coeficiente de correlação sempre deve ser obtido quando houver suficientes dados disponíveis. Na ausência de dados suficientes, não é recomendável a obtenção de algum tipo de coeficiente de correlação.

Antes de definir o coeficiente de correlação é preciso definir a covariância entre duas variáveis (x , y), dada por:

$$C(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - E(x)) \cdot (y_i - E(y)) \quad (2.14)$$

No caso em que x e y sejam independentes, $C(x, y)$ é igual a zero. Agora se pode definir o coeficiente de correlação $\rho(X, Y)$ dado por:

$$\rho(x, y) = \frac{C(x, y)}{\sigma_x \sigma_y} \quad (2.15)$$

Onde σ_x e σ_y são os desvios padrão das variáveis x , y .

É importante deixar claro que o coeficiente de correlação é uma medida só da associação linear entre variáveis. Assim duas variáveis altamente dependentes de uma forma não linear podem ter um coeficiente de correlação igual a zero.

O intervalo do coeficiente de correlação (equação 2.15) é entre +1 e -1, indicando perfeita correlação linear positiva para $\rho = +1$. Por outro lado $\rho = -1$ indica perfeita correlação linear negativa, enquanto $\rho = 0$ indica que não há associação entre as variáveis consideradas.

Holtz e Krizek (1972) mostram um coeficiente de correlação negativo entre o teor de umidade e a densidade seca (Figura 2.7). Eles também mostram um baixo coeficiente de correlação positivo entre a densidade seca e a resistência à compressão não confinada (Figura 2.8).

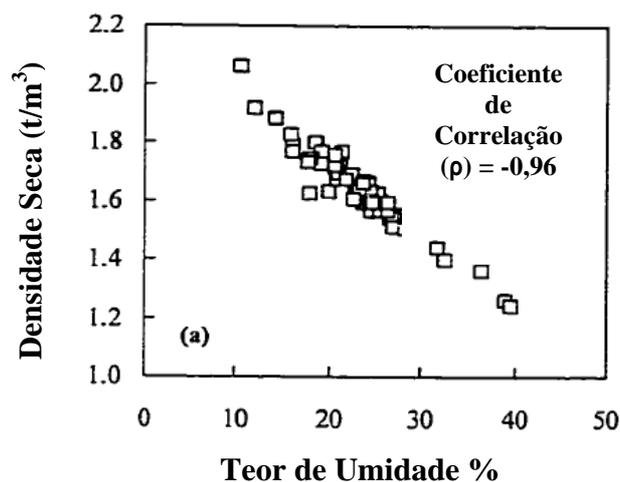


Figura 2.7 – Correlação negativa forte entre duas variáveis, Holtz e Krizek (1972).

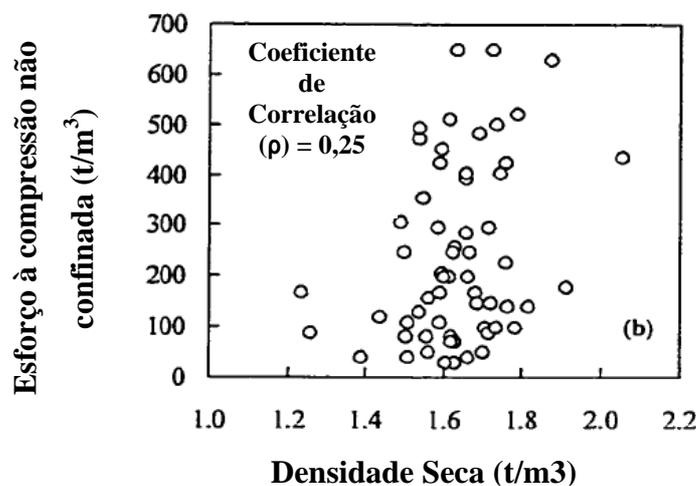


Figura 2.8 - Correlação positiva baixa entre duas variáveis, Holtz e Krizek (1972).

No caso de problemas de estabilidade de taludes, as variáveis usuais são os parâmetros de resistência, peso específico do solo e condições de poropressão, porém suas incertezas não são independentes e devem ser determinadas.

Alonso (1976) mostra que a contribuição da incerteza da densidade do solo é insignificante às demais incertezas presentes numa análise de estabilidade de taludes. Isto é devido ao fato de que a determinação do peso específico do solo, obtida em laboratório, geralmente é precisa e mostra um pequeno desvio padrão.

As análises de estabilidade de taludes são pouco sensíveis às variações do peso específico do solo, pelo que este parâmetro geralmente é tomado como um dado determinístico e não uma variável aleatória. É usual, por tanto que o coeficiente de correlação do peso específico com as propriedades de resistência do solo seja desprezado.

2.4.4. Valores de Coeficiente de Variação Usuais em Solos

Na literatura encontra-se diferentes valores de coeficiente de variação (Cov), definidos pela equação 2.13. Estes valores de Cov podem ser de grande utilidade para se validar um determinado parâmetro de solo. Valores típicos estão apresentados na Tabela 2.3.

O coeficiente de variação (Cov) descreve quanto os dados estão longe da média, sendo, portanto função da variação espacial dos dados. Segundo Phoon e Kulhawy (1999a), o maior problema em usar os coeficientes de variação encontrados comumente na literatura é que estes não removem a tendência que podem apresentar os dados de solo (por exemplo, com a profundidade).

A não remoção da tendência dos dados pode superestimar o Cov, a menos que o parâmetro não mostre uma tendência com a profundidade, ou seja, constante, como mostrado na Figura 2.9.

Tabela 2.3 – Coeficientes de variação típicos de solos.

Características do Solo	Cov(%)	Referência
Densidade	3-7	Harr (1987), Kulhawy (1992)
Ângulo de Atrito Efetivo (ϕ') (Areias)	2-13	Harr (1987), Kulhawy (1992), Duncan (2000)
	5-15	Lumb (1974), Hoeg e Murarka (1974), Singh (1971).
Ângulo de Atrito (argilas)	12-56	Lumb (1974), Singh (1971)
Resistência não Drenada (Su)	13-40	Harr (1987), Kulhawy (1992), Lacasse e Nadim (1996)
	20-50	Lumb (1974),), Singh (1971)
Resistência não drenada (Su), Ensaio de palheta.	10-20	Kulhawy (1992)
	10-40	Phoon e Kulhawy (1999a)
N de SPT	15-45	Harr (1987), Kulhawy (1992)
	10-70	Phoon e Kulhawy (1999a)
Coeficiente de adensamento	33-68	Duncan (2000)
Pressão de Pré adensamento (σ'_a)	10-35	Harr (1984), Duncan (2000), Lacasse e Nadim (1996)
Limite de Liquidez	2-48	Lumb (1974), Singh (1971), Kuhn (1971), Mitchell (1993).
Limite de Plasticidade	9-29	Lumb (1974), Singh (1971), Kuhn (1971), Mitchell (1993).
Índice de Plasticidade	7-79	Lumb (1974), Singh (1971), Kuhn (1971),

A Figura 2.9 mostra uma variância total de 45 bpf^2 considerando tendência nula com a profundidade. Quando é considerada uma tendência linear, obtida por métodos de regressão, a variância residual encontrada é de 11 bpf^2 . Isto representa só 25% da variância total, mostrando claramente que a consideração da tendência diminui em 75% a variância da variável N_{SPT} .

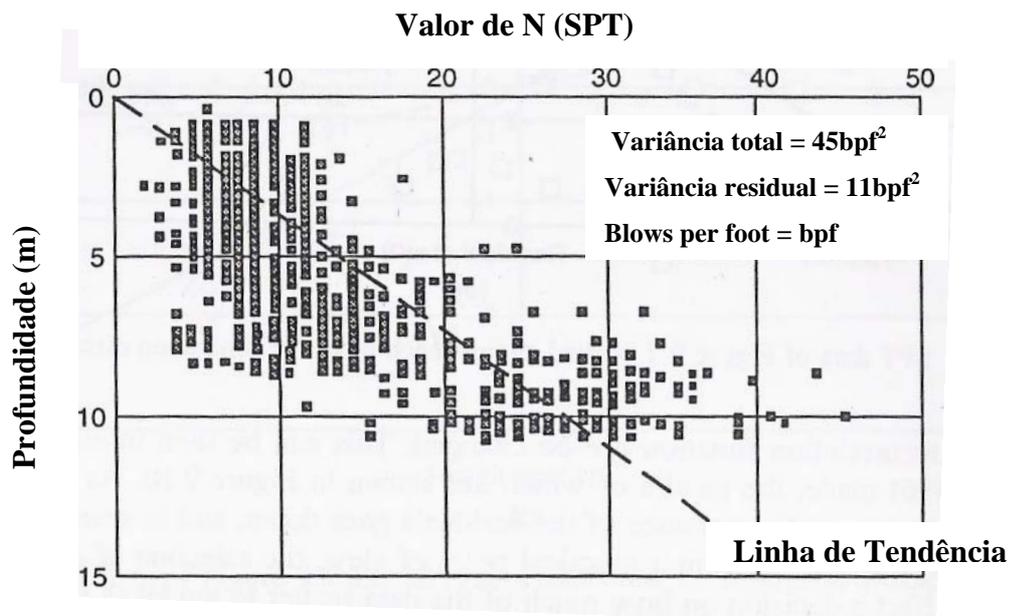


Figura 2.9 – Efeito da não remoção da tendência no cálculo da variância ou Cov, em dados de N do SPT em areias, Baecher e Christian (2003).

Todos os métodos descritos neste capítulo não consideram a existência de uma tendência para o cálculo da variância (ou desvio padrão ou covariância) pelo que ferramentas adicionais devem ser desenvolvidas para a melhor quantificação da variância. Estas ferramentas estão descritas no Capítulo 3.