

LIÇÕES FUNDAMENTAIS DE
GEOESTATÍSTICA

Introdução a **Geoestatística**

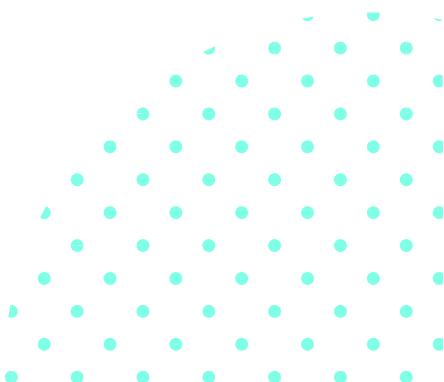
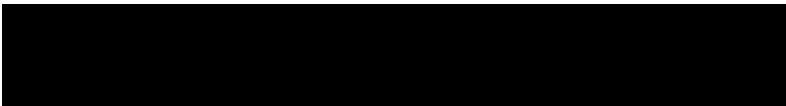
Com aplicações em R, GSLIB e SGEMS

David A. Drumond

Fernanda G. F. Niagini

João Felipe C.L. Costa

Roberto M. Rolo



Geoestatística

Introdução aos princípios e aplicações em R, GSLIB e SGeMS

David Alvarenga Drumond

Fernanda Gontijo Fernandes Niquini

Roberto Mentzingen Rolo

João Felipe Coimbra Leite Costa

David Alvarenga Drumond
Fernanda Gontijo Fernandes Niquini
Roberto Mentzingen Rolo

Geoestatística - Introdução aos princípios e aplicações em R, GSLIB e SGeMS

1 edição

Belo horizonte
22/09/2017

Aos meus pais pelo eterno carinho e apoio.

*Sempre que te perguntarem se podes fazer um trabalho,
respondas que sim e te ponhas em seguida a aprender como se faz.*

F. Roosevelt



Conteúdo

1	Prefácio	13
2	Introdução a geoestatística	17
2.1	Introdução ao capítulo	17
2.2	Afinal, o que é geoestatística?	18
2.3	Qual é o objeto de estudo da geoestatística?	21
2.4	O que podemos fazer com a geoestatística?	23
2.5	Como utilizar a geoestatística?	26
2.6	O que a geoestatística não faz?	27
2.7	Questões éticas na avaliação de depósitos minerais	29
2.8	Alguns conceitos iniciais sobre jazidas minerais	30
2.8.1	Minério	30
2.8.2	Teor de corte e teor crítico	31
2.8.3	Continuidade	31
2.8.4	Diluição	32
2.8.5	Recursos e reservas minerais	32
2.8.6	Precisão e Exatidão	34
2.9	Conclusões	36
2.10	Exercícios	36

3	Variáveis aleatórias regionalizadas	39
3.1	Introdução ao capítulo	39
3.2	Variáveis aleatórias	40
3.3	Função de distribuição acumulada - fda	42
3.4	Função de densidade de probabilidade - fdp	44
3.5	Variáveis regionalizadas	44
3.6	Funções aleatórias	47
3.7	Hipótese de estacionaridade	52
3.8	Momentos estatísticos	55
3.9	Ergocidade	59
3.10	Homocedasticidade e heterocedasticidade	59
3.11	Relação Volume Variância	60
3.12	Conclusões	64
3.13	Exercícios	64
4	Estatística univariada	67
4.1	Introdução	67
4.2	Estatísticas pontuais	70
4.2.1	Medidas de tendência central	72
4.2.2	Medidas de posição	74
4.2.3	Medidas de dispersão	75
4.2.4	Assimetria	76
4.2.5	Coeficiente de variação	77
4.2.6	Conjugando estatísticas pontuais	78
4.3	Validação do banco de dados e valores outliers	79
4.4	Descrição espacial das amostras	83
4.5	Histograma	86
4.6	Inferência Estatística	90
4.6.1	Famílias de distribuições estatísticas	90
4.7	Distribuição t-Student	96
4.8	Dimensionamento de malhas regulares	97
4.9	Exercícios	98

5	Estatística bivariada	99
5.1	Introdução	99
5.2	Probabilidade condicional e Esperança condicional	100
5.2.1	Probabilidades condicionais e conjuntas	100
5.2.2	Esperança condicional	102
5.3	Ferramentas gráficas	104
5.3.1	Gráfico Q-Q plot	104
5.3.2	Gráfico p-p plot	106
5.3.3	Gráfico de dispersão	108
5.4	Regressão linear	110
5.5	Intervalo de segurança para a regressão linear	113
5.6	Regressão linear múltipla	115
5.7	Coeficiente de correlação	116
5.8	Exercícios	119
6	Métodos clássicos e desagrupamento	121
6.1	Introdução	121
6.1.1	Princípio da mudança gradual	122
6.1.2	Princípio dos pontos mais próximos	123
6.1.3	Princípio da generalização	124
6.2	Composição	125
6.3	Composição em seções verticais	126
6.4	Determinação de volumes	128
6.5	Inverso do quadrado da distância - IQD	129
6.6	Tesselação de Delunay	130
6.7	Polígonos de Thiessen	131
6.8	Estatísticas desagrupadas	135
6.8.1	Polígonos de influência	137
6.8.2	Desagrupamento por células	137
7	Continuidade Espacial	141
7.1	Definição de continuidade espacial e variografia	141
7.2	Dependência espacial	142
7.3	Hipótese de estacionaridade	144

7.4	Funções experimentais de continuidade espacial	145
7.4.1	Efeito dos dados sobre os valores experimentais	145
7.4.2	Funções de continuidade espacial mais comuns	145
7.4.3	Outras funções experimentais	147
7.4.4	Parâmetros de busca	150
7.5	Modelagem de funções de continuidade espacial	152
7.5.1	Modelos de variogramas permissíveis	152
7.5.2	Parâmetros das funções de continuidade	152
7.5.3	Modelos de continuidade espacial mais comuns	153
7.5.4	Anisotropia	155
7.5.5	Funções de continuidade espacial cruzadas	157
7.5.6	Modelo linear de correacionalização	158
7.5.7	Modelagem automática de variogramas	158
8	Krigagem	161
8.1	Introdução	161
8.2	Krigagem Ordinária	164
8.3	Krigagem Simples	165
8.4	Krigagem de blocos	166
8.5	Influência nos pesos da krigagem	168
8.5.1	Influência do modelo de continuidade espacial nos pesos	169
8.5.2	Influência dos parâmetros do variograma	169
8.5.3	Efeito da geometria das amostras	172
8.6	Estratégia de procura	174
8.7	Validação da krigagem	176
8.7.1	Verificação do comportamento dos mapas krigado e das amostras	176
8.7.2	Comparação da média global com a média das amostras	177
8.7.3	Análise de deriva de bandas do mapa	177
8.7.4	Validação cruzada	178
8.7.5	Verificação de pesos negativos	178
9	Mudança de suporte	181
9.1	Mudança de suporte	181
9.1.1	Correção afim	182
9.1.2	Transformação lognormal indireta	183

9.2	Curva de teor e tonelagem	183
9.2.1	Curvas de teor e tonelagem derivadas de histogramas das amostras	184
9.2.2	Curvas de teor e tonelagem a partir de distribuição de probabilidades contínuas das amostras	185
9.2.3	Curvas de teor e tonelagem baseadas na dispersão dos blocos estimados	185
9.2.4	Curvas de teor e tonelagem baseadas na estimativa dos blocos	186
9.2.5	Erros associados à determinação da curva de teor-tonelagem	186
10	Estimativa x Realidade	187
10.1	Introdução	187
10.1.1	Controle de teores do minério	188
10.1.2	Uso de fatores de comparação - forma clássica	188
10.1.3	Uso de fatores de comparação - forma probabilística	189
10.1.4	Críticas à geoestatística	190
A	Geoestatística multivariada	193
A.1	Modelos multivariados	195
A.1.1	Krigagem simples com médias locais variáveis	196
A.1.2	Krigagem com deriva externa	196
A.1.3	Cokrigagem	198
A.1.4	Influência dos dados secundários	200
A.1.5	Condição não tradicional e tradicional da cokrigagem	200
A.1.6	Cokrigagem Colocada	200
B	Geoestatística utilizando o software R	201
B.1	Introdução	201
B.2	Instalação do R	203
B.3	RStudio	204
B.4	Noções preliminares	205
B.5	O R como uma calculadora	206
B.6	Utilizando funções no R	206
B.7	Operadores Relacionais	207
B.8	Operadores Lógicos no R	207
B.9	Pedindo ajuda no R	208
B.10	Pacotes do R	209

B.11	Criando vetores	209
B.12	Condicional	211
B.13	Repetições	211
B.14	Concatenação de funções	212
B.15	DataFrames	212
B.16	Mapa de localização	213
B.17	Histogramas	216
B.18	Boxplots	217
B.19	Regressão Linear	219
B.20	Vizinho mais próximo	221
B.21	Variograma	223
B.22	Validação Cruzada	229
B.23	Krigagem	231
C	Geoestatística utilizando o GSLib	235
C.1	Introdução	235
C.2	A execução do GSLIB	236
C.3	Entrada de dados	236
C.4	Exemplos de aplicação do GSLIB	237
C.4.1	Criando um histograma com o HISTPLT	237
C.4.2	Criando um gráfico de dispersão com o SCATPLT	240
C.4.3	Criando um mapa de localização com o LOCMAP	241
C.5	desagrupamento utilizando células móveis com o DECLUS	243
C.6	Convenção da orientação de eixos de anisotropia do GSLIB	245
C.7	Variograma experimental (GAMV/ VARGPLT)	246
C.8	Modelagem de variogramas (VMODEL/VARGPLT)	249
C.9	Validação Cruzada com (KT3D/LOCMAP)	252
C.10	Krigagem com (KT3D/PIXELPLT)	256
D	Geoestatística utilizando o SGeMS	261
D.1	Importando um arquivo de pontos no SGeMS	262
D.2	Visualização dos dados - Mapa de localização	265
D.3	Criação do histograma	266

Bibliografia 269



1. Prefácio

*Se seus problemas têm solução,
aprendes a solucioná-los, se não têm
solução, aprendes a não preocupar.
Os problemas só surgem quando não
se aprende a agir, no primeiro caso
aja, no segundo contemple.*

...

A geoestatística é, sem dúvida, uma das mais belas ferramentas para trabalharmos com incerteza em projetos que envolvem análise espacial e engenharia. Não é somente um ponto de partida para analisar e avaliar, mas uma proposição da humildade humana, nossa capacidade da incompreensão extensiva do universo a nossa volta.

Quando pensamos na ciência tradicional, desenvolvida nos primórdios do século XVIII com o advento do iluminismo, vemos claramente o ser humano tentando desenvolver ferramentas para explicar o universo em suas nuâncias, criando modelos físicos e matemáticos das representações de problemas reais. Estes modelos, sólidos e claros, possuem uma capacidade incrível de reproduzibilidade, podendo ser aplicados em diferentes contextos e situações. Apesar desta maleabilidade, em certos momentos problemas complexos nos foram apresentados, sendo incapazes de serem descritos pelas formas simples por estes antigos modelos.

As técnicas geoestatísticas, como as demais ferramentas modernas de estatística, marcam um ponto na história, quando assumimos a impossibilidade de entender todos os processos, mas que possamos descrevê-los de forma verossímil, criando modelos matemáticos que se aproximam da realidade, desconhecida e muitas vezes intangível. Os modelos estatísticos neste caso não são reproduutíveis, sendo impossível a um avaliador aplicar o mesmo modelo para diferentes depósitos minerais. No entanto, as técnicas são simples e eficientes, reproduzindo padrões sobre a geologia e fenômenos espaciais de forma eficiente.

O surgimento da geoestatística está diretamente relacionado aos trabalhos do professor [George Matheron](#) a partir de análises sobre estudos estatísticos de [Singel](#) em minas de ouro da África do Sul. Depósitos de ouro são conhecidos pela sua complexidade geológica, apresentando alta variabilidade em suas propriedades químicas e geológicas. Os métodos clássicos de avaliação de recursos se demonstraram inefficientes para lidar com esta complexidade do problema, pois consideravam apenas questões geométricas de disposição das amostras, sem considerar sua interdependência espacial. Surgiu então a geoestatística, que pretendia adicionar informação não apenas pelo posicionamento de amostras, mas pela sua dependência espacial.

Diferentemente dos métodos estatísticos clássicos, em que consideramos as amostras independentes entre si, na geoestatística consideramos que o posicionamento espacial, volume das amostras e orientação possuem forte relação com as avaliações que realizamos. Esta proposição torna os métodos geoestatísticos ainda eficientes por mais de 50 anos de desenvolvimento nas avaliações de recursos minerais, pois corrobora com a questão física de formação dos controles geológicos destes depósitos.

Após anos de desenvolvimento nas técnicas geoestatísticas, diferentes pesquisas e conhecimentos foram derivados das técnicas iniciais de Matheron. Mesmo assim, esta disciplina ainda parece obscura nos cursos de engenharia de minas, geologia e geografia não somente no Brasil como no mundo. Quando focamos no contexto brasileiro, o problema se acentua, pois são poucas as bibliografias escritas em português sobre este assunto. Pensando nisso, nós alunos do curso de pós-graduação Engenharia de Minas da Universidade Federal do Rio Grande do Sul, juntamente com a orientação dos profissionais da universidade, iniciamos este livro como um projeto para uma série de publicações sobre [Geoestatística](#), [Planejamento de Lavra](#), [Amostragem de jazidas](#), [Métodos de cubagem](#) entre outras disciplinas da mineração.

A abordagem adotada neste livro de [Introdução a Geoestatística](#) envolve a chamada geoestatística linear clássica, que envolve desde métodos tradicionais de geoestatística, de caracterização da continuidade espacial e da krigagem, principalmente ordinária e simples, que envolvem principalmente os primeiros trabalhos desenvol-

vidos na década de 70 pelo professor George Matheron. Ao final do livro temos seções dedicadas exclusivamente para a apresentação da geoestatística em softwares gratuitos e linguagem de programação em R. O livro se desenvolve nos seguintes capítulos:

- **Capítulo 2:** Este capítulo apresenta uma introdução a respeito da ciência da geoestatística, explicando os principais conceitos para entendermos algumas questões chaves, como *O que é a geoestatística?*, *O que podemos fazer com a geoestatística?*, *Como podemos utilizar a geoestatística?*
- **Capítulo 3:** Apresenta o objeto principal do estudo da geoestatística, a teoria das variáveis regionalizadas. Apresentamos a conceituação e o formalismo matemático deste conjunto de técnicas.
- **Capítulo 4:** Inserimos conceitos iniciais sobre estatística univariada, ou seja, as técnicas utilizadas para avaliação de apenas uma única variável. Para isso mostramos estatísticas gráficas, estatísticas pontuais, enfocando principalmente no contexto da mineração.
- **Capítulo 5:** Inserimos conceitos iniciais sobre estatística multivariada, ou seja, aquela que analisa informações conjuntas de duas ou mais variáveis. São apresentadas estatísticas pontuais, estatísticas gráficas, entre outras ferramentas específicas deste assunto.
- **Capítulo 6:** Neste capítulo apresentamos as principais formas de análise de agrupamento para amostras espaciais. O objetivo destas técnicas é criar condições para reduzir o efeito de amostragens localizadas e ponderar as estatísticas para que indiquem sua representação espacial verdadeira.
- **Capítulo 7:** Introduzimos os conceitos de continuidade espacial e dependência espacial das amostras, mostrando por exemplo, a estimativa e modelagem de funções variograma e covariograma.
- **Capítulo 8:** Apresentamos as principais técnicas de estimativa geoestatística linear, como krigagem simples e ordinária.
- **Capítulo 9:** Introduzimos os conceitos de mudança de suporte, e as principais ferramentas para analisar a mudança de volumes estimados.
- **Capítulo 10:** Apresentamos as principais técnicas utilizadas para avaliar os modelos krigados e estimativas realizadas, como por exemplo, as curvas de teor e tonelagem comumente utilizadas nos trabalhos de avaliação de recursos.

A seção final de cada capítulo apresenta uma série de exercícios com arquivos de dados apresentados no material de apoio deste livro. Ao final destes capítulos apresentamos os apêndices A, B e C, com avaliações geoestatísticas realizadas no depósito *Walker Lake*. Os dados deste depósito podem ser encontrados junto com o material de apoio.



2. Introdução a geoestatística

A estimativa de recursos é o processo de criação de um reflexo tridimensional da mineralização in situ baseado em amostras esparsas utilizando conhecimento geológico corrente e um caminhão carregado de senso comum.

*The art and the science of resource estimation
Jacqui Coombes*

2.1 Introdução ao capítulo

Este capítulo inicial pretende demonstrar os primeiros passos para entender a geoestatística. Afinal, o que é a geoestatística? Qual é o seu objeto de estudo? O que podemos fazer ou não com a geoestatística? Para que um marceneiro possa fazer uma cadeira, por exemplo, ele precisa entender de suas ferramentas e de seu funcionamento, para que possa selecionar as mais adequadas para o seu trabalho. Entendendo o que é a geoestatística e como podemos utilizá-la, principalmente no setor da mineração, estabelecemos um vínculo necessário para a aplicação correta desta poderosa ferramenta.

2.2 Afinal, o que é geoestatística?

A importância das substâncias metálicas na indústria mineral brasileira é historicamente associado aos tempos da Colônia, procurando rotas inicialmente no estado de Minas Gerais. Segundo o relatório da Agência Nacional de Mineração (2018), a produção das principais substâncias metálicas no país atingiram um valor de 41,7 bilhões de reais em exportação. A produção mineral rende impostos para as regiões produtoras, que ao mesmo tempo permitem o desenvolvimento da economia local e geração de empregos. A decisão da extração e produção mineral, no entanto, advém do conhecimento geológico e de estimativas dos corpos minerais aos quais muitas vezes não se possui acesso. Os corpos geológicos, muitas vezes, não apresentam informações superficiais de fácil acesso, como afloramentos que permitem a definição da **atitute** das camadas, por isso é necessário realizar amostragens em grande profundidade como na obtenção de **testemunhos de sondagem**. A partir de uma sonda são retirados fragmentos de rocha à profundidades de até 400m, permitindo que tenhamos informações da composição direta das rochas naquela região. A figura 2.1 exemplifica a forma cilíndrica apresentada pelo testemunho de sondagem obtido em campanhas de pesquisa mineral.

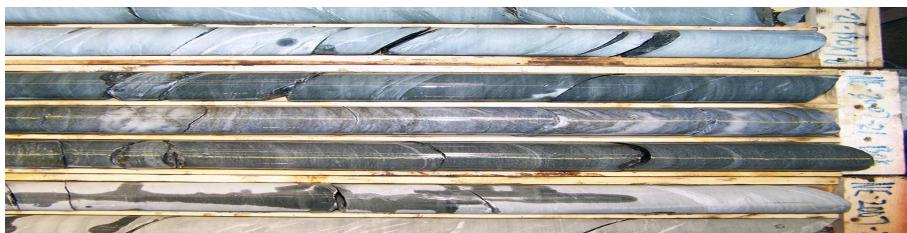


Figura 2.1: Testemunhos de sondagem de rochas. Amostras retiradas a partir da perfuração do solo, que apresentam uma informação contínua vertical das rochas e mineralogias presentes em uma região.

Desta forma, a única informação que possuímos é a informação vertical fornecida pelos testemunhos, como demonstrado na figura 2.2. As decisões da mineração não podem ser estabelecidas sem o conhecimento das informações entre os furos de sondagem, que podem representar malhas espaçadas em muitos metros. A amostragem exaustiva dos depósitos minerais também é inviável economicamente, pois em alguns casos, cada metro de amostra sondada pode corresponder a um valor de \$100 a \$400 reais.

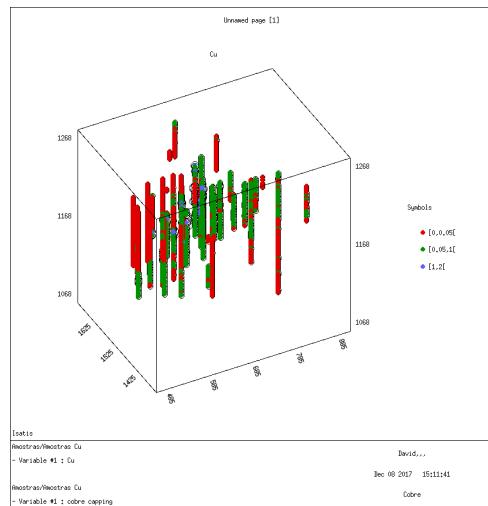


Figura 2.2: Testemunhos de sondagem representados em um software de mineração, para um depósito de cobre.

A geoestatística é a ciência que permite a espacialização das informações obtidas em um volumes menores, para um domínio maior, de forma a permitir o planejamento e tomada de decisões na mineração, e o estudo sistemático dos corpos mineralizados. A figura 2.3 demonstra a espacialização dos dados obtidos na figura 2.2 dos testemunhos de sondagem de cobre.

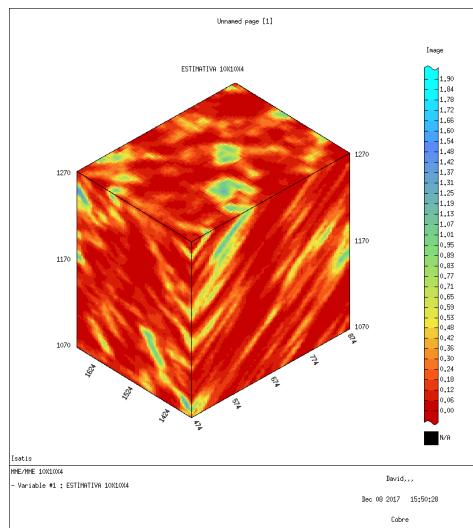


Figura 2.3: Espacialização a partir das amostras do testemunho de sondagem obtidos na figura 2.2, para um depósito de cobre.

A partir de um modelo espacializado é possível planejar a mineração e tomar a decisões na lavra, como a criação de **modelos econômicos**, a determinação das **cavas matemáticas**, o **sequenciamento das operações**. Segundo Rossi and Deutsch [2013], o objetivo principal da geoestatística consiste em 4 etapas principais:

1. Obtenção de amostras e administração da amostragem
2. Interpretação geológica e modelagem
3. Interpolação dos teores
4. Acesso às incertezas geológicas

A **obtenção de amostras e administração da amostragem** consiste no conjunto de técnicas utilizada para obter uma malha de amostragem que reduza o erro obtido pela interpolação espacial das propriedades de interesse. Por exemplo, a malha de amostragem em um depósito de ferro bandado, conhecido como (*Banded Iron Formations*) , pode ser dimensionada para que se reconheçam os minerais deletérios durante a fase de metalurgia.

A **interpretação geológica e modelagem** permite reconhecer as dimensões e formas do corpo geológico e principais estruturas. Um dos grandes desafios da mineração é conseguir reconhecer os limites dos corpos minerais e sua forma. A geoestatística permite utilizar técnicas que auxiliem no reconhecimento das formas destes corpos de maneira grosseira, a partir da *modelagem implícita*. O objetivo do uso destas técnicas é reduzir a quantidade de trabalho demandada pelo geólogo para que se possam produzir modelos condizentes com a realidade, e ao mesmo tempo, poupar trabalho excessivo pelo desenvolvimento de seções verticais.

A **Interpolação de teores** consiste no objetivo principal deste livro, em que realizamos a espacialização das propriedades de interesse das amostras para um domínio espacial maior. Esta espacialização pode ser realizada para apenas uma variável (caso univariado), ou para diversas variáveis em conjunto (caso multivariado). O objetivo principal da interpolação é garantir, com maior segurança possível, que um volume direcionado da lavra para o beneficiamento mineral possua **valor esperado**, ou **valor médio**, correto.

O **Acesso às incertezas geológicas** pode ser realizado a partir de técnicas avançadas de geoestatística como a simulação, ou geoestatística não-linear. Pretende-se desta forma tentar reconhecer as incertezas locais de uma propriedade do depósito, e avaliar quão díspares podem ser as medições em regiões do depósito que desconhecemos. A incerteza geológica é, sem dúvida, um dos fatores que mais afetam o **risco** do empreendimento mineiro. Para que os investidores possam verificar o risco de seus investimentos, foram criados os **códigos de mineração**, que criaram padrões nomear regiões do depósito mineral com maior ou menor incerteza quanto uma propriedade de interesse, geralmente aquela de retorno econômico.

Segundo [Matheron \[1963\]](#), criador da geoestatística, podemos definí-la tal como:

R "Geoestatística, na sua maior aceitação, consiste no estudo da distribuição do espaço de valores úteis para engenheiros de minas e geólogos, como teores, espessura da camada, ou acumulação, incluindo as práticas mais importantes para a avaliação de depósitos minerais- [Matheron \[1963\]](#)

Atualmente o uso da geoestatística compreende uma diversidade enorme de áreas, desde a **engenharia civil**, **engenharia agrícola**, **engenharia ambiental**, **geografia**, **engenharia hídrica** e até mesmo em áreas que não se resumem à dados geograficamente referenciados, mas espacialmente referenciados em objetos ou seres, como a **mecânica** ou **medicina**. Podemos entender a geoestatística sob uma perspectiva mais ampla, abordando o estudo das incertezas a cerca de fenômenos temporalmente ou espacialmente localizados. O professor [Goovaerts \[1997\]](#), demonstra claramente a nossa dificuldade de entender as incertezas:

R "A respeito da incerteza ... ela surge do nosso conhecimento imperfeito do fenômeno, dependente dos dados e ainda mais dependente do modelo, em que o modelo especifica nossas decisões (concepções) a priori do fenômeno. Nenhum modelo tal como a medida da incerteza, pode ser objetiva.- [Goovaerts \[1997\]](#)

Podemos entender então as limitações acerca dos modelos geoestatísticos. Estamos sempre **dependentes das amostras recolhidas para a avaliação**, como também a escolha dos modelos que melhor representam as características de um fenômeno. A geoestatística constitui atualmente a área que melhor consegue caracterizar a incerteza geológica, dada as condições de amostragem que obtemos na mineração e de muitos problemas georeferenciados. Definimos a geoestatística como:

Definição 2.2.1 — Geoestatística. *A geoestatística é a ciência capaz de transformar as informações obtidas por amostras georeferenciadas em conhecimento, a partir da caracterização da incerteza geológica, da interpretação destes dados, das inferências e estimativas, e da tomada de decisão pelo reconhecimento do fenômeno estudado.*

2.3 Qual é o objeto de estudo da geoestatística?

A geoestatística é a ciência que permite o estudo de variáveis regionalizadas. O capítulo 3 trará informações a respeito desta teoria, que compõe o objeto principal do estudo da geoestatística. Uma variável regionalizada é aquela que pode assumir um valor específico no espaço. Este valor é **determinístico**, gerado a partir de fenômenos que muitas vezes não conhecemos. Por não conseguirmos acessar as informações a respeito desta variável, optamos por utilizar uma metodologia **estocástica** para acessar a nossa **incerteza** a cerca do fenômeno que estudamos. Desta

forma pensamos na variável regionalizada com um aspecto **dicotômico**, a medida que possui valor real onde conhecemos, e valor aleatório onde desconhecemos.

O físico Erwin Schrödinger desenvolveu um problema em 1935 muito similar a esta condição das variáveis regionalizadas. O experimento foi chamado de "Gato de Schrödinger". O experimento propunha que um gato fosse preso em uma caixa, com um veneno que poderia aleatoriamente ser liberado, matando o gato em seguida. Para quem observa a experiência do lado de fora, não há como detectar se o gato está vivo ou morto, logo o estado de sobrevivência do gato é indefinido, dependendo da real observação de dentro da caixa. A figura 2.4 demonstra este experimento.

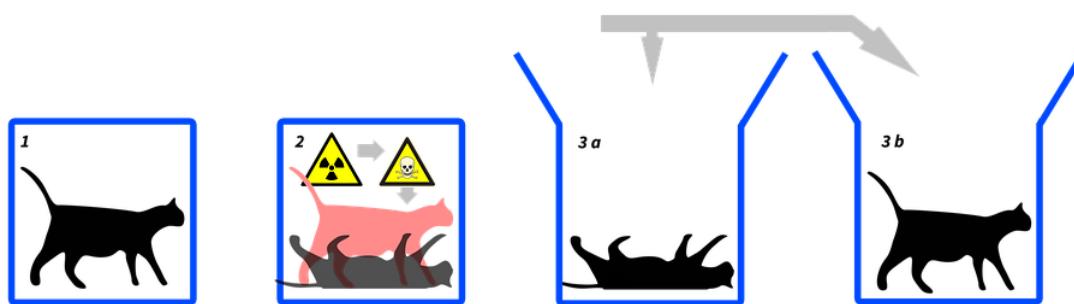


Figura 2.4: Experiência do "Gato de Schrödinger", proposta em 1935 pelo físico Austríaco Erwin Schrödinger. 1) O gato está dentro da caixa, 2) Passado um tempo o observador externo não sabe o que está dentro da caixa. 3a) Abre-se a caixa e define-se que o gato está morto. 3) Abre-se a caixa e define-se que o gato está vivo.

Quando pensamos em termos da mineração, o problema de se definir minério ou estéril é similar ao do "Gato de Schrödinger". Não sabemos de fato o que será minerado deve ser enviado para o beneficiamento mineral, a não ser que de fato retirarmos o material do local. As **variáveis regionalizadas** funcionam de forma bem semelhante, possuindo este aspecto ao mesmo tempo determinístico e aleatório. Um depósito mineral é um evento geológico realizado durante milhões de anos. Durante o tempo de existência humana é quase impossível que estes depósitos minerais se modifiquem. Desta forma o corpo mineral, ou o "gato" já está dentro da caixa há muito tempo, porém nos é impossível determinar o seu atual estado sem que ocorra a mineração.

Ao estudar as variáveis regionalizadas, a geoestatística propõe encontrar valores e relações que desconhecemos, sem obtermos informações diretas do depósito mineral. Considerando a variabilidade inerente destas variáveis podemos criar modelos estatísticos que possam inferir propriedades que desconhecemos em outras regiões do depósito mineral.

2.4 O que podemos fazer com a geoestatística?

A partir da avaliação dos depósitos minerais utilizando a geoestatística, podemos caracterizar o fenômeno espacial e quantificar as incertezas para diferentes **variáveis**. Uma variável é uma característica de interesse de estudo no depósito mineral, que se modifica segundo seu posicionamento no espaço. No estudo da mineração possuímos uma série de diferentes tipos de variáveis associadas ao depósito mineral, tais como:

1. **Químicas:** Teores de elementos químicos de interesse, ou de elementos deletérios prejudiciais no processamento mineral
2. **Físicas:** Dureza, densidade, condutibilidade térmica, condutibilidade hidráulica, saturação
3. **Geológicas:** Litologia, composição mineralógica, número de falhas, RQD (Rock quality index)
4. **Processamento:** Recuperação metalúrgica, recuperação mássica, moabilidade, consumo de reagentes
5. **Operacionais:** Resistência a penetração, consumo de explosivos, tempo de carregamento
6. **Econômicas:** Preço de mercado, valor presente líquido

O entendimento de cada uma destas variáveis permite a tomada de decisão de lavra de uma parte constituinte do depósito mineral. O uso de modelos geoestatísticos para cada uma destas variáveis deve, no entanto, ser específico para cada tipo de variável calculada. Neste livro introdutório abordamos principalmente os problemas que se relacionam com variáveis consideradas **aditivas**. Algumas variáveis que podem ser consideradas aditivas, e que representam o maior escopo de trabalho dos avaliadores de depósito mineral são **teor do elemento metálico, quantidade de metal de interesse, massa do minério e acumulação**. [Carrasco et al. \[2008\]](#) demonstra o conceito de aditividade de variáveis, expresso por:

 "Quantidades consideradas aditivas são aquelas que a quantidade média é igual a média das quantidades.- [Carrasco et al. \[2008\]](#)"

A variável teor, por exemplo, pode ser considerada uma variável aditiva, pois a média aritmética dos teores de duas regiões de mesma forma e volume é idêntico ao valor médio do teor nestas regiões. No caso da recuperação metalúrgica, por

exemplo, não há possibilidade de se considerar a média aritmética como valor médio, sendo impossível utilizar a geoestatística linear para estimar valores de recuperação metalúrgica. Carrasco et al. [2008] ainda afirma a necessidade de variáveis aditivas para se realizar o processo de estimativa diretamente por meio da geoestatística linear clássica.

R "Uma quantidade dita não aditiva, não pode modelar sua variabilidade espacial ou estimar diretamente- Carrasco et al. [2008]

Para estimar ou avaliar variáveis ditas não-aditivas recorremos aos métodos de **geoestatística não linear** ou **simulação geoestatística**. Estes métodos não serão abordados neste volume deste livro, apenas abordaremos os conceitos primários da *geoestatística linear*, que envolvem o tratamento de variáveis aditivas. No entanto, para o leitor iniciante, é importante entender que os métodos deste livro apenas se aplicam para **variáveis aditivas** e para amostra com o mesmo volume e forma, conceito denominado de **suporte amostral**. As estimativas realizadas no depósito mineral geralmente são feitas em volumes maiores, chamado de **suporte da estimativa** e compõe a chamada **unidade seletiva de lavra**. Segundo Rossi and Deutsch [2013] uma unidade seletiva de lavra pode ser caracterizada como:

R "Mínimo volume de material ao qual o minério e o estéril podem ser separados, em função do método de lavra e da seletividade- Rossi and Deutsch [2013]

Podemos então discretizar o espaço em pequenos blocos, para se realizar a estimativa nestes locais. Este é chamado de **modelo de blocos**, representado na figura 2.5.

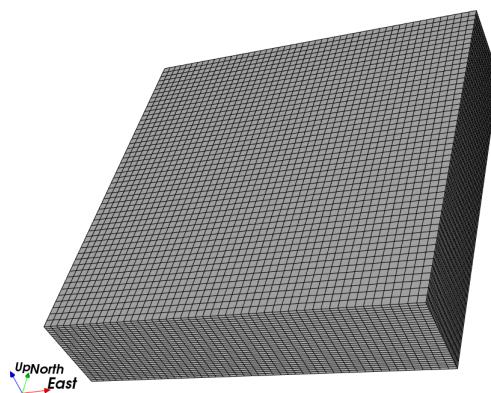


Figura 2.5: Representação de um modelo de blocos e discretização do espaço.

Assim podemos dividir o espaço a ser estimado em pequenos volumes de decisão na lavra. Para fins de planejamento mineral, quanto menor o tamanho destes blocos,

melhor a facilidade do planejamento. No entanto, para fins de avaliação de depósitos, blocos de tamanho pequeno produzem estimativas espúrias.

O equilíbrio entre estas duas vontades deve ser encontrado para constituir o tamanho adequado da unidade seletiva de lavra. Uma das regras de ouro da mineração geralmente afirma que: **o tamanho do bloco não deve ser inferior a 1/4 do tamanho da malha de amostragem**. Esta é uma afirmação atrela o tamanho do bloco geralmente a uma malha de amostragem bem definida e calculada, o que muitas vezes não condiz com as questões práticas.

Proposição 2.4.1 Segundo a regra de ouro da geoestatística um bloco estimado não deve ter tamanho inferior a 1/4 do espaçamento da malha de amostragem. Quando considerada uma malha irregular este tamanho não pode ser menor que 1/4 do valor esperado dos espaçamentos. O valor esperado pode ser calculado a partir da média aritmética dos espaçamentos.

A discretização do depósito mineral em diferentes domínios nem sempre ocorre somente em modelos de blocos. Diferentes formas de caracterização dos volumes no espaço pode ser utilizada na geoestatística e no planejamento minera. A Figura (2.6) demonstra alguns exemplos de divisão do espaço. Algumas delas como **polígonos de influência** e **triangulação de Delunay** representam antigas formas de estimativa de um depósito mineral, mas que, no entanto, ainda são usuais por outras formas de análise.

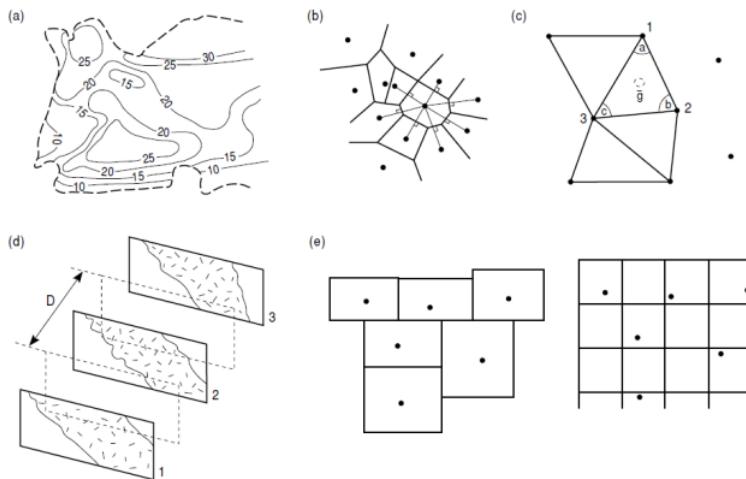


Figura 2.6: Figura demonstrando diversas apresentações de uma propriedade do depósito mineral. a) isolinhas b) Polígonos de influência c) triangulação d) seção paralelas e) blocos irregulares f) blocos regulares

A partir da definição das variáveis do depósito mineral o engenheiro é capaz de estimar a viabilidade técnica e econômica do depósito mineral. O artigo 6, do

decreto 9.406 de 12 de junho de 2018 do código mineral vigente define:

Definição 2.4.1 — Jazida Mineral. *Toda a massa de substância mineral ou fóssil, que aflore na superfície ou já exista no solo, no subsolo, no leito ou no subsolo do mar territorial, da zona econômica exclusiva ou da plataforma continental, que tenha valor econômico*

A partir da avaliação de depósitos minerais, conseguimos então, identificar regiões econômicas e capazes do aproveitamento industrial.

2.5 Como utilizar a geoestatística?

A geoestatística clássica utilizada neste livro não aborda conceitos matemáticos complexos, mas ainda sim constitui base para a resolução de muitos problemas de estimativa na mineração. Mesmo que os cálculos não sejam complexos, são de certa forma muito onerosos computacionalmente. As resoluções podem ser demoradas e alguns casos exigirem alta performance computacional. Desta forma, o uso de algoritmos refinados se torna cada vez mais importante nas análises geoestatísticas.

Durante a fase de avaliação das jazidas minerais o computador exerce função essencial como ferramenta de estudo. Uma quantidade substancial de softwares estão disponíveis em meio comercial e alguns aplicativos livres também existem. Softwares comerciais são mais custosos, mas possuem suporte técnico e manutenção de seus sistemas. Apresentam código fechado ao público externo e pertencente geralmente aos proprietários. Softwares gratuitos geralmente são disponibilizados por universidades, possuem código aberto ao público e podem ser facilmente obtidos via Internet.

Uma das bibliotecas gratuitas mais importantes é sem dúvida o GSLIB (Geostatistical Software Library) e apresenta além dos executáveis do programa seus algoritmos, escritos em Fortran 90 e disponibilizados no site. Os programas são administrados pelo doutor Clayton Deutsch e Emmanuel Schnetzler. Mais informações sobre o pacote de softwares pode ser encontrado no site www.gslib.com ou no guia de uso [DeutschCV \[1998\]](#). Neste livro abordamos nas seções A, B e C o uso dos principais softwares e linguagem R.

O alinhamento da geoestatística com o desenvolvimento de algoritmos cria dependências com as disciplinas de programação. Os engenheiros e geólogos estão cada vez mais alinhados com o desenvolvimento de algoritmos, principalmente com as linguagem R e Python, pela sua simplicidade e facilidade de implementação.

O uso dos softwares de mineração geralmente requerem que os arquivos de dados sejam organizados eficientemente em formatos pré-estabelecidos, gerados pelas

campanhas de exploração. Essa compilação dos dados é trabalhosa e necessita de uma validação primordial, tornando o trabalho de preparação dos dados às vezes muito mais demorado que as implementações dos programas.

Entre as aplicações mais comuns encontradas em softwares de mineração, temos:

- Uma grande variedade de procedimentos de avaliação dos dados (estatísticas, gráficos, etc.:)
- Determinação da qualidade dos dados e dos protocolos de amostragem
- Modelagem tridimensional e visualização de formas geológicas complexas e distribuição das amostras.
- Preparação de seções planas e verticais
- Gráficos de contorno tanto do teor como de outras variáveis
- Caracterização da continuidade espacial (Variogramas automáticos, mapas de variograma, variogramas experimentais e modelagem)
- Modelagem de blocos do depósito
- Metodologias de cálculos de recurso e reservas
- Avaliações dos efeitos de vários métodos de mineração
- Determinação da viabilidade econômica de depósitos

Alguns destes softwares podem ainda incluir ferramentas de planejamento de mina, tal como otimização de cava, sequenciamento, desenho de cava, etc. A grande quantidade de ferramentas adicionadas nestes programas geralmente os tornam pouco específicos para análises espaciais, obtendo apenas algumas rotinas específicas para trabalhos mais simples.

2.6 O que a geoestatística não faz?

Toda ferramenta possui suas limitações. A geoestatística é a melhor ferramenta para análises espaciais até então criada, mas ela possui limitações no uso de seus modelos. Primeiramente **a geoestatística não é uma caixa preta**. Isso significa que uma boa análise do depósito mineral não depende exclusivamente de apertar um botão, como muitos modelos mais simples fazem. Para criar modelos geoestatísticos adequados eles devem passar por uma intensa avaliação e reavaliação dos parâmetros de ajuste destes modelos.

Em segundo lugar **a geoestatística não é uma bola de cristal**. Quando realizamos estimativas estamos susceptíveis a erros e incertezas. Uma boa estimativa dos depósitos minerais propiciará redução dos erros, mas inevitavelmente não podemos sempre esperar resultados exatos. Além das condições relacionadas a escolha e refinamento dos modelos, também temos condições inerentes da incerteza geológica. [Maranhao \[1985\]](#) demonstra a classificação de jazidas minerais no ponto de vista da avaliação de reservas, identificando quatro grupos principais:

1. **Grupo 1.** Pertence aos depósitos estratiformes, cujos representantes típicos são as jazidas sedimentares de origem marinha, que possuem grandes dimensões, forma mais ou menos constante, e regularidade na distribuição de teores. Também incluem neste grupo as jazidas metamórficas de ferro, tais como nos depósitos do quadrilátero ferrífero. Também apresenta alguns depósitos de disposição horizontal ou a subhorizontal como jazidas de calcário, carvão, sais, gipsita e alguns depósitos para construção civil, como gnaisses e granitos.
2. **Grupo 2.** O segundo grupo apresenta corpos minerais interrompidos ou levemente interrompidos e uma distribuição mais irregular dos teores que do primeiro grupo. Estas representam as jazidas de alteração superficial como depósitos de níquel e bauxita, depósitos com pequenas intrusões alcalinas, como carbonatitos e sienitos, jazidas de rochas ultrabásicas e hidrotermais.
3. **Grupo 3** O terceiro grupo geralmente enquadra jazidas de forma variável e mineralização muito irregular, compondo os principais depósitos auríferos, platinoides e diamantes. Também aborda os depósitos de veios polimetálicos e depósitos de forma lenticular, como de cobre e níquel.
4. **Grupo 4** Representa o grupo mais irregular de todos, compondo pegmatitos de pedras preciosas, alguns veios hidrotermais com metais raros e nobres e algumas jazidas ultrabásicas de platina e diamante.

Quanto maior a irregularidade do depósito e heterogeneidade de suas propriedades, maior será a dificuldade dos modelos geoestatísticos de predizerem com exatidão os resultados. Dependendo da **continuidade espacial** da propriedade e da sua **dispersão**, a aplicação de modelos simples ou complexos simplesmente não altera o nosso conhecimento sobre a **incerteza geológica**, pois a complexidade do depósito mineral é tão grande, e as amostragens realizadas em tão pouca quantidade, que se torna mais fácil jogar uma moeda para cima para decidir se devemos ou não lavrar um depósito. Neste caso, quando os métodos geoestatísticos falham,

é necessário rever as metodologias de amostragem, e tentar encontrar soluções que simplifiquem as variáveis do problema.

Desta forma a geoestatística também tem outra limitação: **Os modelos geostatísticos requerem amostras realizadas em quantidade e qualidade adequada para gerar resultados satisfatórios.** Esta talvez seja uma das limitações mais difíceis de se conseguir abordar dentro da mineração. Para estimar de forma adequada precisa-se de amostras, e amostras são caras. Muitas empresas deixam de amostrar adequadamente seus depósitos minerais com finalidade de redução de custos, mas acabam por avaliar mal seus depósitos minerais, e consequentemente, obtém baixo lucro ou inviabilizam o uso sustentável dos recursos minerais. O termo qualidade, também é uma questão muito importante. Muitas vezes as amostragens realizadas possuem protocolos mal dimensionados. Alguns métodos de amostragem de jazidas também devem ser conduzidos de forma bem precisa para realizarem estimativas adequadas, mas apesar de ser uma das etapas mais importantes na mineração, as empresas muitas vezes colocam a tarefa nas mãos de profissionais pouco qualificados.

2.7 Questões éticas na avaliação de depósitos minerais

Avaliar depósitos minerais é uma atividade incerta, devido a natureza dos depósitos minerais, no entanto, não há justificativa para o mal uso das técnicas, nem ao mesmo para decisões arbitrárias que não envolvam decisões puramente lógicas ou racionais. Infelizmente o setor mineral acaba por ser alvo de pessoas com má conduta, por ser uma área de grandes riquezas. Esta não é, com certeza, a personalidade da grande maioria dos trabalhadores que se dedicam diariamente no setor mineral, mas pessoas acabam por utilizar a justificativa do "incerto" para vender depósitos minerais subvalorizados. Um avaliador de depósitos minerais deve realizar sua tarefa friamente, analisando a viabilidade do depósito independente se ele gerará riquezas ou não.

É importante também para os gestores e gerentes de minas entenderem a natureza do problema, e que as incertezas geológicas produzirão muitas vezes resultados diferentes dos pretendidos. A mineração trata do aproveitamento de recursos que são limitados pelo tempo geológico de sua criação. Enquanto o ser humano ainda não controlar o tempo, é indiscutível que temos de aproveitar os recursos minerais existentes da melhor forma que consigamos. A avaliação de depósitos minerais é o alicerce das decisões na mineração, por isso é impreterível que os processos sejam realizados de forma mais correta possível.

Proposição 2.7.1 *Está nas mãos do avaliador de depósitos minerais a determinação das condições necessárias para a progressão da lavra. O desenvolvimento de seus projetos deve seguir sempre com conhecimento e idoneidade, pois é dele que deriva o trabalho de pessoas, o aproveitamento correto dos recursos minerais e da sociedade que aproveita estes recursos*

2.8 Alguns conceitos iniciais sobre jazidas minerais

Apresentamos nesta seção alguns dos principais conceitos de engenharia de minas, necessários para a realização de trabalhos de geoestatística e avaliação de depósitos no setor mineral. Apesar deste livro possuir foco na geoestatística, consideramos adequado entender conceitos gerais da mineração, que influenciam nas decisões tomadas pela avaliação dos depósitos.

2.8.1 Minério

A definição de minério talvez seja uma das mais importantes na produção mineral. A sua determinação permite o aproveitamento econômico dos recursos minerais, decidindo o que deve ou não ser lavrado e aproveitado. Segundo [Hustrulid et al. \[2006\]](#), a definição de minério pode ser considerada como:

R "Um agregado mineral com um ou mais sólidos minerais aos quais podem ser minerados, ou dos quais um ou mais produtos minerais podem ser extraídos com lucro". [Hustrulid et al. \[2006\]](#)

Isto significa que nem em todas as ocasiões um minério será extraído com a finalidade de se obter lucro pela venda. Em alguns casos, as questões econômicas da extração mineral podem ser contra intuitivas neste sentido, devido a políticas externas, estados de guerra, monopolização da produção, entre diversos outros fatores. Neste caso preferimos adotar o conceito de minério a partir do seu benefício, nem sempre ele sendo econômico. Definimos minério como:

Definição 2.8.1 — Minério. *Minério é todo agregado mineral ou fóssil cabível de aproveitamento técnico, que possibilita um benefício, seja ele econômico ou social, de forma a propiciar os interesses das diferentes componentes da sociedade, sejam elas a União, as forças sociais ou mineradores.*

Um exemplo bem característico de minérios explotados contra o senso econômico são os minerais radioativos, de monopólio da União. É de interesse estratégico de um país deter estes recursos capazes de produzir energia e armas, sendo muitas vezes gastos valores acima do valor do minério para sua extração.

2.8.2 Teor de corte e teor crítico

O conceito de teor de corte (ou cutoff) é definido como aquele em que o valor do conteúdo metálico ou mineral, em um certo volume de rocha, permite sua extração econômica. Os teores de corte são usados para distinguir blocos de minério e estéril em vários estágios da evolução da estimativa da jazida mineral (exploração, desenvolvimento e produção). O teor crítico, no entanto, representa o teor ao qual se delimita o limite entre prejuízo e lucro. [Rendu \[2014\]](#) define o teor de corte como:

R "O teor de corte geralmente é definido como a mínima quantidade de um produto de valor ou metal que em uma tonelada métrica deve conter para que este material seja enviado para a planta de beneficiamento" - [Rendu \[2014\]](#)

2.8.3 Continuidade

A continuidade é um termo derivado em toda a história da matemática e da ciência desde tempos remotos. Talvez uma das primeiras concepções da continuidade seja com o paradoxo de Zenão, que conta a história da corrida de Aquiles e a tartaruga. [Srivastava and Parker \[1989\]](#) demonstram o sentido da continuidade como:

R "Uma descrição da similaridade ou da dissimilaridade entre pares de valores com uma função de sua separação do vetor h " - [Srivastava and Parker \[1989\]](#)

Em outras palavras podemos dizer que a continuidade espacial é representada pela similaridade entre medidas que se localizam em regiões diferentes no espaço. Os fenômenos geológicos, neste caso, apresentam uma importante característica derivada de suas gêneses: Na maioria dos casos, medidas de propriedades realizadas mais próximas tendem a ser mais similares entre si do que medidas realizadas em grandes distâncias. Caracterizar a similaridade dos fenômenos geológicos é a chave para garantir que as estimativas e a caracterização da incerteza geológicas possam ser realizadas.

Definição 2.8.2 — Continuidade espacial. Definimos a continuidade espacial como a regularidade com que uma propriedade é medida em amostras aproximadas no espaço. Se as diferenças entre as amostras for pequena, dizemos que o material é contínuo ou similar. Quando o material é muito diferente de amostras pouco espaçadas dizemos que ele é discreto ou dissimilar. Na geoestatística definimos a continuidade a partir de uma direção do espaço, podendo ela se apresentar diferencialmente de acordo com a direção adotada.

2.8.4 Diluição

Segundo [Susaeta et al. \[2008\]](#) a diluição se refere ao estéril que não é separado do minério durante a operação da lavra. Este estéril é misturado com o minério e enviado para a usina de beneficiamento. Enquanto aumenta a quantidade de material enviado para a usina a diluição diminui o teor que deveria ser estimado e enviado para a usina corretamente. A estimativa de depósitos minerais é realizada desconsiderando os efeitos de produção e planejamento. Isto significa que os valores realmente lavrados não correspondem aos volumes planejados e induzem diferenças naturais da estimativa. O processo de comparação entre os valores reais obtidos na usina e os valores estimados pode ser definido como **aderência do planejamento de lavra**.

Definição 2.8.3 — Aderência de lavra. *Aderência do planejamento de lavra é todo o processo de comparação entre os valores estimados do depósito mineral e os obtidos durante a operação, seja durante a mineração, ou dos valores obtidos durante o beneficiamento mineral.*

As incertezas geológicas presentes no depósito mineral, ou as diferenças do planejamento da operação podem trazer discordâncias quanto os volumes e qualidade do material estimado e realmente lavrado, causando diluição do minério. Existem diferentes tipos de diluição durante a extração mineral. A **diluição interna** ocorre quando existem partes de estéril dentro do volume estimado do minério. Algumas vezes a amostragem pode não computar veios ou lentes de estéril dentro do bloco de decisão de lavra, dado que o volume das amostra é muito inferior ao volume das amostras. A **diluição externa** ocorre quando o planejamento mineral aborda parte do material não definido como minério, o que é comum nas regiões de contato do corpo geológico. Também há a chamada **diluição operacional**, que ocorre quando o desmonte de rochas realiza a fragmentação em regiões acima do planejado, chamado de *overbreak*, ou abaixo do planejado, chamado de *underbreak*.

2.8.5 Recursos e reservas minerais

A definição de recursos e reservas minerais são alternativas para publicidade de declarações públicas relativo às incertezas geológicas do depósito mineral. A CBRR (Comissão Brasileira de Recursos e Reservas) identifica a declaração pública como:

Definição 2.8.4 — Declaração pública. *Declarações públicas são preparadas para informar investidores ou potenciais investidores e seus conselheiros sobre os resultados da exploração, recursos minerais ou reservas minerais. Elas incluem, mas não se limitam, a relatórios anuais ou trimestrais das entidades, notas à im-*

prensa, memorandos informativos, documentos técnicos, publicações em website e apresentações públicas.

A partir de declarações públicas, as empresas podem indicar os volumes de metais e de massas estimados com base no conhecimento da incerteza geológica. Esta alternativa foi criada na década de 70, principalmente após o escândalo da empresa Bre-X, após constatado salgamento das minas de ouro em Busang na Indonésia. Definindo **Recursos** e **Reservas** minerais, o minerador classifica seus potenciais de produção segundo a incerteza geológica. A CBRR também define Recurso Mineral como:

Definição 2.8.5 — Recurso Mineral. *Um Recurso Mineral é uma concentração ou ocorrência de material sólido de interesse econômico dentro ou na superfície da crosta terrestre onde forma, teor ou qualidade e quantidade que apresentem perspectivas razoáveis de extração econômica.*

A definição de Recurso está ligada diretamente ao conhecimento da incerteza geológica. Os códigos de mineração não definem as técnicas necessárias para se definir os volumes de depósito de acordo com estas incertezas, apenas indicam que deve-se usar alguma técnica pertinente para isto. A responsabilidade desta definição cai diretamente à pessoa competente responsável pela auditoria. Estes Recursos minerais podem ser divididos em ordem crescente de confiabilidade geológica de acordo com as categorias de Inferido, Indicado e Medido. A CBRR também define Reserva Mineral como

Definição 2.8.6 — Reserva Mineral. *Uma Reserva Mineral é a parte economicamente lavrável de um Recurso Mineral Medido e/ou Indicado. Isso inclui diluição e perdas que podem ocorrer quando o material é lavrado ou extraído e é definido apropriadamente pelos estudos nos níveis de Pré-Viabilidade ou de Viabilidade que incluem aplicação de Fatores modificadores.*

Ou seja, para transformar um recurso em reserva mineral é necessário que se prove a viabilidade da extração do minério, seja ela econômica, social, ambiental ou política. Isto é realizado a partir dos fatores modificadores. A CBRR também define os fatores modificadores como:

Definição 2.8.7 — Fatores Modificadores. *Fatores Modificadores são considerações usadas para converter Recursos Minerais em Reservas Minerais. Esses incluem, mas não se limitam a considerações sobre: a lavra, o processamento, a metalurgia, a infraestrutura, a economicidade, o mercado, os aspectos legais, ambientais, sociais e governamentais*

As Reservas Minerais podem se dividir em provável, quando medida a partir de

um Recurso Indicado e, em algumas circunstâncias de um Recurso medido. A reserva provada é aquela que possui alta confiabilidade, representando recursos medidos. A figura 2.7 demonstra graficamente os resultados da exploração mineral em recursos e reservas minerais, também apresentando sua forma de conversão, de acordo com o conhecimento geológico e os fatores modificadores.

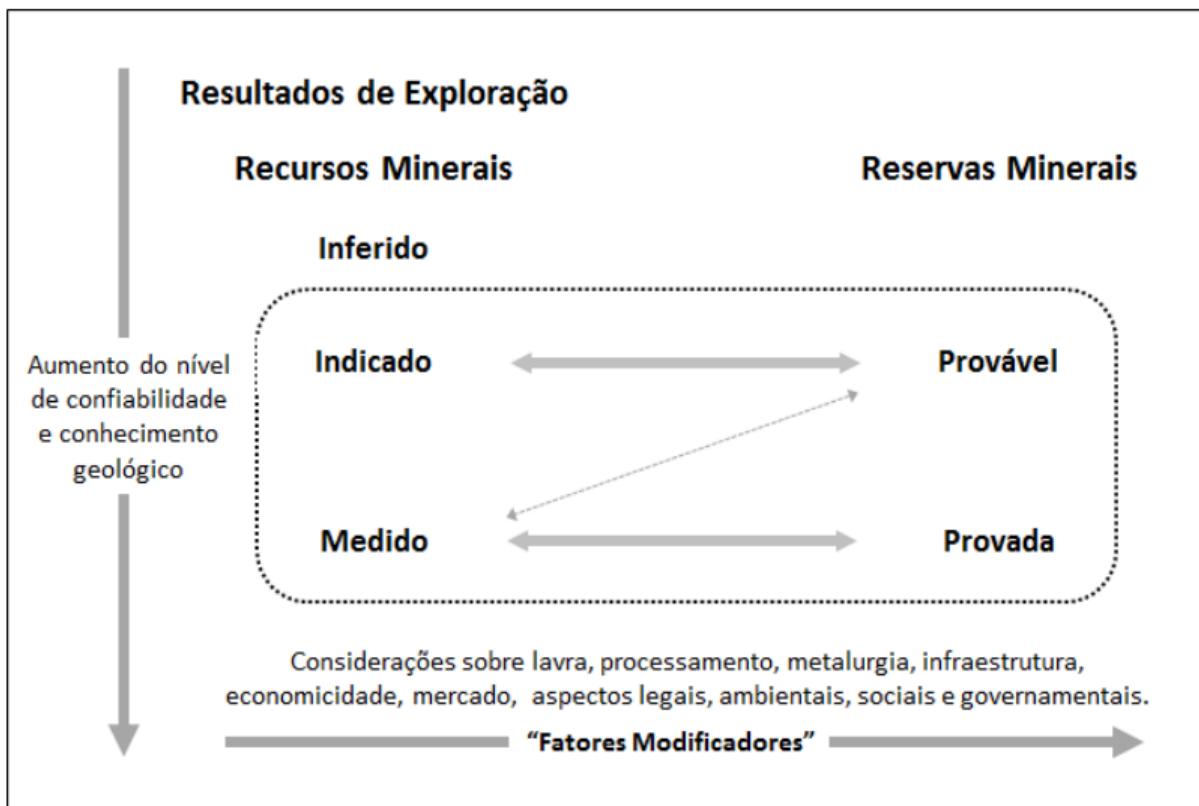


Figura 2.7: Figura demonstrando a classificação de jazidas em recursos e reservas. Linhas indicando a transição entre as classificações

2.8.6 Precisão e Exatidão

Uma das premissas utilizadas na geoestatística clássica é que os resultados das amostras obtidas é um valor fixo. Esta afirmação na maioria dos casos não é realista, pois as amostragens na mineração podem apresentar diferentes valores referentes aos erros de amostragem.

Proposição 2.8.1 *Para os processos de geoestatística clássica, os valores das amostras georeferenciados são determinísticos, a medida que apresentam volume, posicionamento e propriedades constantes. Isto não se aplica a todos os métodos geoestatísticos como o KVME (Kriging with Measurement Error Variance) Delhomme [1978]*

Esta variação das amostras quanto ao valor esperado por elas pode ser definido por duas propriedades: **Exatidão** e **Precisão**. A Exatidão pode ser exemplificado como a proximidade de uma estimativa com a realidade, enquanto precisão é a medida da dispersão entorno de uma estimativa. Analogamente a precisão e a exatidão podem ser comparadas com um jogo de dardos como na figura (2.8), em que pretendemos atingir o centro do alvo. Quanto mais próximo forem os disparos do centro, melhor será a sua exatidão, e quanto mais próximos forem os disparos entre si, significa que são mais precisos. Disparos podem ser precisos, no entanto, não exatos. Disparos podem ser exatos por se localizarem em média próximos do centro, mas podem ser imprecisos se distanciarem entre si.

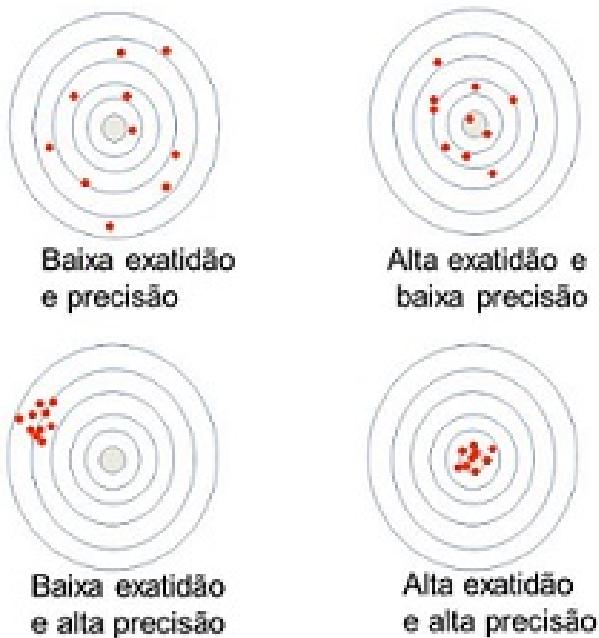


Figura 2.8: Figura demonstrando os conceitos de exatidão e precisão. O centro do alvo é o valor verdadeiro que pretende-se alcançar com os disparos. Disparos entorno do centro são considerados exatos. Disparos próximos aos outros são considerados precisos

A amostragem na mineração ainda sofre um outro problema, quanto a reproduzibilidade. Na verdade este é um problema para a maioria dos fenômenos espaciais, pois quando amostramos uma região não há como amostrar novamente, pois estas amostras geralmente são **destrutivas**. Além disso, ao amostrar em um local específico, uma amostra mesmo que próxima já se configura como uma amostra diferente. Os trabalhos do professor Gy [2012] invocam os principais conceitos e teorias a respeito da amostragem a granel, utilizada na mineração.

Eventualmente diversos fatores podem causar as variações e erros na amostra-

gem. Há vários tipos de erros potenciais na estimativa de reservas minerais incluindo:

- Erro de amostragem
- Erros de análise química.
- Erros de densidade (É comum em muitos casos considerar a densidade do material constante ao longo do depósito)
- Erros da geologia, durante as fases de determinação da continuidade espacial e geometria do depósito mineral.
- Na escolha do método de lavra adotado que pode não atender as questões de seletividade do minério e estéril de forma ótima.
- A diluição do minério com a encaixante.
- Erro humano (inserção de valores errados no banco de dados, de casas decimais, et.)
- Fraude (salgamento de amostras, substituições de amostras, dados não representativos, etc.)

2.9 Conclusões

Neste capítulo inicial apresentamos os principais conceitos relacionados à geoestatística e ao planejamento de mina. Definimos o que é esta ciência que será abordada ao longo de todo o livro, o que ela pode realizar ou não, e sua importância dentro do contexto da mineração. Entendemos que a geoestatística é uma ferramenta para auxiliar na compreensão do desconhecido, e que é inerente ao empreendimento mineral, pois raras são as alternativas ao qual possuímos informação sistemática ao longo de todo o depósito mineral.

2.10 Exercícios

Exercícios 2.1 Segundo a definição de Carrasco et al. [2008], sabemos que uma variável é aditiva se o seu valor médio é igual a média de seus valores. Discutimos ao longo do texto que as variáveis teor e conteúdo metálico são variáveis aditivas, capazes de serem utilizadas nos modelos clássicos que abordamos neste livro. Desta forma identifique variáveis na mineração que podem ser consideradas

aditivas ou não.



Exercícios 2.2 Realize um "brainstorm" e pense todas as possibilidades que podem sofrer uma mina que possam tornar um minério em um estéril. Por exemplo, a descoberta de uma outra jazida de uma empresa concorrente mais próximo do mercado consumidor pode aumentar o preço do minério e tornar parte do recurso inutilizável por um tempo. E quais seriam os fatores que fazem um estéril se tornar minério?



Exercícios 2.3 Pretende-se determinar se uma unidade seletiva de lavra é um minério ou estéril. O custo fixo de extração do material é 5 um/ton. O custo de mineração por tonelada movimentada é 2 um/ton. A relação estéril/minério é 3/2. A Recuperação metalúrgica é de 95% e o preço do minério é de 100 um/ton. O teor do elemento útil do bloco é 2%.



Exercícios 2.4 Os dados da tabela seguinte demonstram um conjunto de valores estimados e dados reais obtidos. Determine:

- O viés das estimativas. (Diferença entre a média dos valores estimados e a dos reais)
- Considere o cut-off como 2g/ton. Determine: A proporção dos valores estimados como minério que realmente são minério. A proporção dos valores estimados como estéreis que realmente são estéreis.

Estimados	Real
2.05	2.0
2.03	2.02
1.01	1.32
2.31	3.45
3.02	1.02
2.76	2.19
3.08	4.01
3.74	3.67
1.02	1.43
1.00	1.01
2.03	1.05





3. Variáveis aleatórias regionalizadas

Todas as vezes que eu leio relatórios estatísticos, eu tento imaginar meu contemporâneo infeliz, a Pessoa Média, a quem, de acordo com estes relatórios, possui 0.66 filhos, 0.032 carros e 0.046 TVs.

Kato Lomb

3.1 Introdução ao capítulo

A geoestatística é uma ciência que se iniciou nos anos de 1950, com estudos de Krige [1960] na África do Sul a respeito de valores estimados em distribuições lognormais de ouro. Em 2012 o professor Daniel Krige recebeu a Ordem de Baobab, uma condecoração do presidente da África do Sul, pelas suas excepcionais contribuições para a economia, ciência, medicina, inovações tecnológicas e serviços comunitários. Durante seus 30 anos de idade, se tornou pioneiro no uso da estatística para avaliação de depósitos de ouro para um número limitado de furos de sondagem. As ideias do pesquisador foram fortemente abraçadas pela França após a tradução de seus artigos em língua nativa em 1995, o que gerou a fundação do centro de Geoestatística em Fontainebleau, corroborando para os estudos do professor George Matheron, e a criação da **teoria das variáveis regionalizadas**.

Este primeiro capítulo introduz a geoestatística a partir do seu objeto de estudo, as variáveis regionalizadas. Explicamos os principais conceitos abordados pela teoria clássica, e como eles se relacionam no entendimento dos fenômenos espacializados. Maiores informações podem ser encontradas nas obras de Matheron [1963] ou nos livros base de Isaaks and Srivastava [1989] e Goovaerts [1997]

3.2 Variáveis aleatórias

Alguns conceitos iniciais sobre estatística são necessários antes que possamos aprofundar os conceitos de geoestatística. Um dos principais conceitos utilizados para o entendimento de fenômenos aleatórios é o de **variável aleatória**.

Definição 3.2.1 — Variável aleatória. *Uma variável aleatória é uma função de um espaço amostral S nos números reais.* Casella and Berger [2010]

Imagine que tenhamos um saco com grandes quantidades de pedras coloridas vermelhas e azuis. Nosso espaço amostral seria portanto $S = \{\text{pedras vermelhas, pedras azuis}\}$. Se quisermos determinar uma variável aleatória que seja definida pela amostragem de duas pedras poderíamos ter o seguinte resultado $Z = \{(\text{pedra vermelha, pedra azul}), (\text{pedra vermelha, pedra vermelha}), (\text{pedra azul, pedra azul})\}$. Uma variável aleatória geralmente é definida a partir de uma letra maiúscula, enquanto uma realização, ou seja, um resultado desta variável aleatória é definido por uma letra minúscula. A figura 3.1 é um exemplo de uma variável aleatória, pois para cada valor possível dentro do espaço amostral de diferentes litologias é associado um valor inteiro.

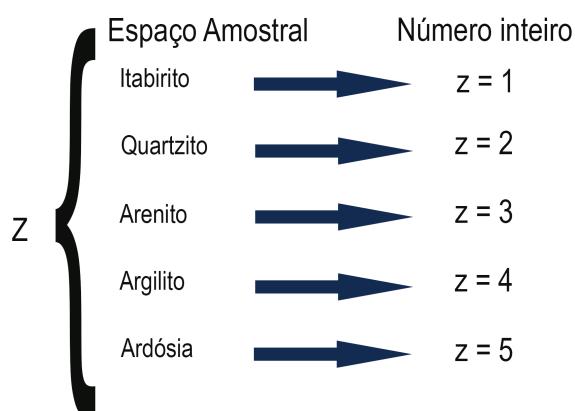


Figura 3.1: Exemplo de variável aleatória indicadora. Para cada possível valor de litologia do depósito é associado um valor inteiro.

Note, no entanto, que atribuir um valor para esta variável não significa dar uma maior importância ou uma menor importância para cada litotipo. Colocar um valor inteiro igual a 1 para o Itabirito não significa considerá-lo mais importante que as demais litologias. Neste caso dizemos que esta variável é **cardinal**, pois o valor associado de cada componente do espaço amostral a um valor inteiro não está diretamente ligado com sua importância, ao contrário de variáveis **ordinais** ao qual seu número associado é diretamente expresso pela sua importância.

As variáveis aleatórias são divididas geralmente em duas classes na geoestatística, considerando **variáveis aleatórias reais**, que podem apresentar valores dentro do conjunto de dados reais, ou **variáveis indicadoras**, quando consideramos que podem assumir valores inteiros. Exemplos de variáveis reais, por exemplo, são as de teores dos elementos metálicos, enquanto variáveis indicadoras são representadas pelas litologias presentes no depósito mineral.

Variáveis ditas **contínuas** são aquelas que possuem um espaço amostral infinito e **não contável**, geralmente representada por um conjunto de valores reais. Quando medimos teores, por exemplo, o resultado de uma amostra pode variar infinitamente dentro de um intervalo de 0% a 100%. Apesar desta limitação, o número de realizações que podem advir desta variação são infinitas, pois naturalmente o valor 5,6740 % é diferente do valor 5,6741 %, mesmo que muito próximos.

Em contrapartida, variáveis discretas são **contáveis**, mesmo que seu espaço amostral seja infinito. Se um subconjunto deste espaço amostral for considerado é possível conseguir definir para ele uma probabilidade. Variáveis discretas estão geralmente ligadas ao conjunto de números inteiros.

Para uma variável aleatória pode ser atribuído uma probabilidade (Pr) de ocorrência para cada uma de suas realizações. A ideia de probabilidade mais básica está relacionada com a **frequência de ocorrência relativa de um evento**, ou também chamada de abordagem frequentista. Nossa variável aleatória demonstrada pela figura 3.1 pode ser associada a uma probabilidade de acordo com a figura 3.2, considerando a proporção de rochas de cada tipo dentro do domínio geológico estimado.

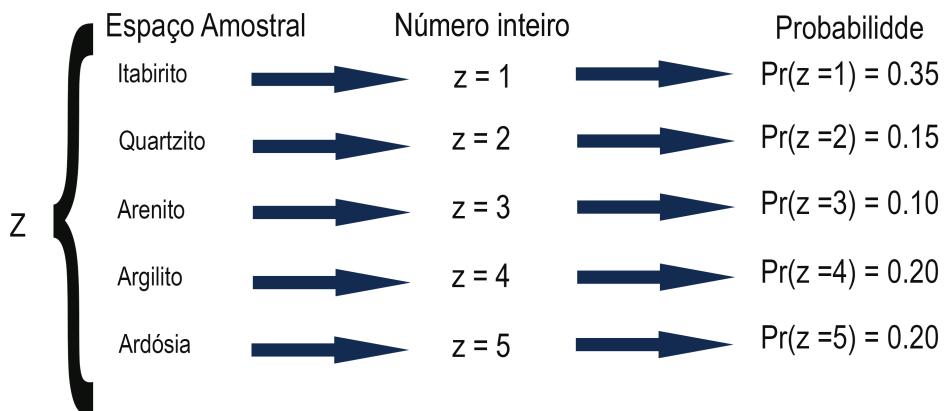


Figura 3.2: Exemplo de variável aleatória indicadora. Para cada possível valor de litologia do depósito é associado um valor inteiro. Uma probabilidade é atribuída para a frequência relativa de cada litologia.

Em outras palavras a probabilidade é semelhante a uma métrica de proporção das realizações de uma variável aleatória. Na verdade a probabilidade pode ser qualquer medida, desde que satisfaça os **Axiomas de Kolmogorov**.

1. A probabilidade de um evento é um número não negativo, dentro do intervalo $[0,1]$.
2. A probabilidade do espaço amostral é 1.
3. Se n eventos são mutuamente exclusivos, a probabilidade da união destes eventos é igual a soma das probabilidades individuais.

Os conceitos de probabilidade são estudados na matemática dentro da **teoria dos conjuntos** que é a base para a fundamentação da estatística. Para maiores informações da teoria base em probabilidade, axiomas de Kolmogorov e teoria dos conjuntos, aconselhamos ler as referências de [Alencar \[2014\]](#) e [FEITOSA et al. \[2011\]](#).

3.3 Função de distribuição acumulada - fda

Para cada elemento de uma variável indicadora podemos associar um valor de probabilidade, ou de frequência da apresentação deste elemento. Por exemplo se consideramos que um depósito mineral possui apenas dois tipos de rocha, podemos dizer que o tipo 1 representa 30% de frequência no depósito mineral, enquanto o tipo

2 apresenta 70% de frequência. Associar uma probabilidade para variáveis aleatórias indicadoras é intuitivo. No entanto, não conseguimos definir a probabilidade de um elemento para variáveis aleatórias reais contínuas, pois o espaço amostral é infinito. Não conseguimos associar, por exemplo, a probabilidade de um teor ser 5,67%. Neste caso utilizamos uma abordagem intervalar, associando a probabilidade a um intervalo de valores reais, logo é possível dizer que o depósito mineral possui probabilidade de 40% dos teores variarem de 5,67% a 9,32%. Uma função de distribuição acumulada é representada pela probabilidade de uma variável aleatória assumir um valor igual ou menor a um determinado limite. Definimos então a **Função de distribuição acumulada** $F(z)$ tal como

$$F(z) = \Pr(Z \leq z) \quad (3.1)$$

A figura 3.3 indica a função de distribuição para variáveis contínuas e discretas. Em A) possuímos uma função discreta que pode assumir apenas valores inteiros de 1 a 8. Em B) possuímos uma função contínua de valores que se alteram no intervalo [1,8]

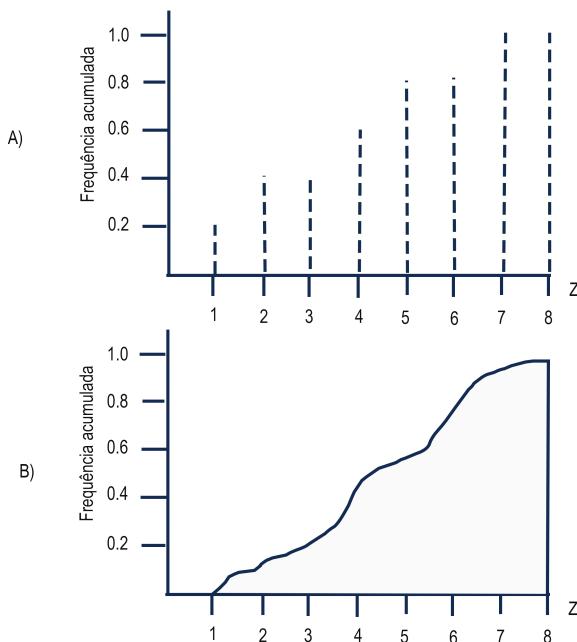


Figura 3.3: Função de distribuição acumulada - fda para variáveis discretas A) e contínuas B)

3.4 Função de densidade de probabilidade - fdp

No caso de distribuições contínuas, em que os valores podem ser determinados para qualquer valor dentro de um domínio real, é permitido utilizar princípios de cálculo para medir informações dessas distribuições. Como não podemos definir o valor da probabilidade em um ponto específico precisamos utilizar o conceito de probabilidade intervalar. A **Função de densidade de probabilidade-fdp** pode ser determinada como a probabilidade um valor assumir este valor dada uma variação infinitesimal.

$$f(z) = \lim_{\delta \rightarrow 0} [Pr(Z > z, Z < z + \delta)] \quad (3.2)$$

Logo a relação entre a função de distribuição acumulada e a função de densidade de probabilidades pode ser expressa por

$$F(z) = \int_{-\infty}^z f(z) dz \quad (3.3)$$

Proposição 3.4.1 *Pode parecer que o valor da densidade de probabilidade seja equivalente ao valor da probabilidade assumindo uma realização z de variável aleatória Z , no entanto esta visão está errada! A ideia de probabilidade está diretamente ligada na ideia de frequência relativa de um evento. Valores de variáveis contínuas possuem espaço amostral incontável, o que significa que não conseguimos medir o seu tamanho, sendo ele infinito. Quantos seriam os possíveis resultados, por exemplo, de um teor de uma amostra apresentar? Se fosse possível associar uma probabilidade a um valor de uma variável aleatória real, todos estes valores seriam iguais a zero, pois sua frequência em nada representa na imensidão de valores possíveis. A função de densidade de probabilidade é na verdade uma medida de taxas de variação da probabilidade.*

3.5 Variáveis regionalizadas

Matheron [1963] , pai fundador da geoestatística, iniciou o conceito de **variável regionalizada**, para exemplificar os fenômenos espaciais. Quando um fenômeno exibe uma certa **estruturação espacial**, dizemos ele ser regionalizado. Os fenômenos geológicos, por exemplo, exibem estruturação característica na sua formação, o que significa que os corpos geológicos apresentam geometrias características de suas gêneses. A variável regionalizada $z(x)$ denota um valor conhecido em um determinado ponto x , sendo apenas um resultado neutro puramente descritivo, sem

interpretação probabilística. Em outras palavras, a variável regionalizada é o valor real encontrado no depósito mineral para cada ponto x no espaço.

Definição 3.5.1 — Variável Regionalizada. *Uma variável regionalizada $z(x)$ representa a medida de uma propriedade qualquer, seja ela o teor do elemento metálico, quantidade de metal ou acumulação, definida em um ponto x no espaço de coordenadas definidas*

É impossível para nós conhecer o valor real de $z(x)$ para cada ponto no espaço, pois implicaria em muito mais que uma amostragem sistemática por todo o domínio do depósito mineral. Desta forma a variável regionalizada apresenta aspectos contraditórios, porém complementares para a definição do modelo geoestatístico:

- **Apresenta uma componente aleatória** onde não conseguimos amostrar ou definir a variável. Isto marca o aspecto irregular da variável. Seguindo a notação estatística, nos locais onde a variável regionalizada não é definida, denotamos $Z(x)$ para informar que nestes locais ela assume aspecto de uma variável aleatória. Sendo Ω o universo que pode ser composto a variável, definimos a variável aleatória em local desconhecido da variável regionalizada como $Z(x) : \Omega \rightarrow \mathbb{R}$.
- **Apresenta uma componente estruturada** nos locais onde é determinada, como por exemplo, pelos métodos de amostragem, representada pela própria forma $z(x)$, convencionalmente pela notação estatística como a realização da variável aleatória $Z(x)$ no suporte x .

Proposição 3.5.1 *Pode parecer um tanto estranho que algo possa assumir condições dicotômicas desta forma. Ao mesmo tempo que consideramos que algo existe e é determinístico, também consideramos que algo é aleatório e transitório. Na verdade as coisas são como sempre são, o que fazemos é assumir que em certos casos, não conseguimos definir algo, e em outro sabemos muito bem o que é. A aleatoriedade, na verdade, nunca existiu. Aleatoriedade é nosso princípio de humildade em não entendermos como os fenômenos ocorrem.*

A observação da variável regionalizada, não ocorre, no entanto em um ponto do espaço. Pontos são abstrações matemáticas de dimensão infinitesimal, uma condição geralmente para que possamos aplicar o princípio de continuidade dos modelos. As nossas observações são realizadas em amostras com volumes específicos e em grandes regiões que queremos estimar. Matheron [1963] apresenta os principais conceitos de domínio e suporte. Um domínio é uma região onde a variável regionalizada é diferente de zero. No nosso livro apresentamos o domínio das estimativas pela

notação D , enquanto os domínios de um bloco ou painel de lavra são apresentados por V .

Definição 3.5.2 — Domínio. *Domínio de uma variável regionalizada pode ser considerada qualquer região onde a variável apresenta valor diferente de zero. Por exemplo, a região da mina onde pretendemos estimar valores desconhecidos pode ser considerada como um domínio de estimativa D .*

Matheron [1963] apresenta também o conceito de suporte, sendo este relacionado com a capacidade de entendimento da variável regionalizada $z(x)$. De certa forma, é impossível conhecer o valor da variável regionalizada em um ponto x , pois o que detemos é o conhecimento da variável em um volume v , representando um testemunho de rocha, ou um fragmento de rocha.

Definição 3.5.3 — Suporte. *Suporte é o volume e forma v ao qual se detém o conhecimento da variável regionalizada $z_v(X)$*

Em alguns casos, pela dimensão do domínio estimado em relação ao suporte, este é quase observado como um ponto. Imagine um fragmento de rocha de 10cm^3 e um painel a ser estimado de 200m^3 . A diferença de ordem de grandeza entre a amostra e o painel é gigantesca.

Proposição 3.5.2 *Dizemos que do ponto de vista matemático é quase impossível definir a variável regionalizada $z(x)$, pois é quase impossível amostrar em um ponto. No entanto, esta é uma observação muito purista, que desconsidera os aspectos de engenharia. Em alguns casos uma amostra pode ser visualizada como uma realização da variável regionalizada $z(x)$, pois o volume da amostra é tão inferior ao domínio, que se torna praticamente uma dimensão pontual*

Uma das condições de aplicação da geoestatística clássica, que considera o uso de variáveis aditivas, é que o suporte das amostras utilizado nas estimativas deve ser o mesmo. Isto significa que o volume dos testemunhos utilizados para estimativa, amostras de canais, ou outros tipos de amostragens devem ter todos mesma forma, tamanho e volume. Esta é também outra questão impraticável, pois é impossível principalmente em rochas, obter regularidade nas amostras desta forma. Para contornar esta situação nos utilizamos os métodos chamados de **regularização**, que permitem criar amostras de mesmo tamanho.

Estas definições são as clássicas apresentadas pelo professor George Matheron em seus primeiros trabalhos sobre a teoria das variáveis aleatórias regionalizadas. Existe muita confusão entre diferentes autores para a representação destes conceitos de **suporte** e **domínio**, sendo muitas vezes o domínio do painel chamado de suporte do painel. De acordo com a definição de suporte, seria necessário conhecer o valor

real do painel, o que é impossível, sendo mais adequada a nomeclatura de domínio do painel. Estas divergências de conceituação não prejudicam o estudo da geoestatística como um todo, mas acabam por criar diferentes formas de notação e algumas vezes dificultam a leitura dos textos. O mais importante em se ter em mente é que este volume, seja do domínio ou do suporte, pode alterar os resultados das suas estimativas, na chamada **relação volume e variância**.

Esta ambiguidade da variável regionalizada permite tratamento de forma diferenciada segundo os objetivos de cada estudo. Podemos, ora tratar a variável regionalizada apenas como valores dispostos no espaço, ora dar um tratamento probabilístico para estes valores. Matheron [1963] aborda estes dois princípios como

- **Métodos transitivos** Considera a hipótese de estacionaridade, mas não implica em qualquer hipótese probabilística, sendo métodos apenas descritivos da variável regionalizada $z(x)$. Esta abordagem é utilizada principalmente na geoestatística clássica abordada neste livro. Faremos os cálculos geoestatísticos considerando apenas a descrição dos valores amostrados em uma determinada região, sem premissas sobre uma possível distribuição de probabilidades local.
- **Teoria intrínseca** Utiliza a interpretação probabilística da variável regionalizada $Z(x)$, também considerando hipóteses de estacionaridade. Esta metodologia é amplamente utilizada nos métodos considerados não-lineares e nas simulações geoestatísticas, em que se pretende determinar não apenas um valor esperado determinístico para um volume estimado, mas também uma distribuição de probabilidades.

3.6 Funções aleatórias

Como dissemos anteriormente a variável regionalizada possui uma componente tanto determinística, onde conhecemos os valores da variável, como uma componente aleatória, em locais onde se desconhece a propriedade de interesse. Este aspecto dicotômico é trocado por alguns autores ao usarem da **teoria intrínseca** e estabelecerem a variável aleatória sobre termos exclusivos de uma interpretação probabilística. Uma visão um pouco mais abstrata da variável aleatória é entender que sua componente determinística é apenas um resultado ou uma realização da variável aleatória naquele local, e que $Z(x)$, chamada em alguns casos de **função aleatória** é uma função que associa a qualquer ponto do espaço uma variável aleatória. A figura 3.4 demonstra o resultado de uma amostragem $z(x = x_1)$ no ponto x_1 .

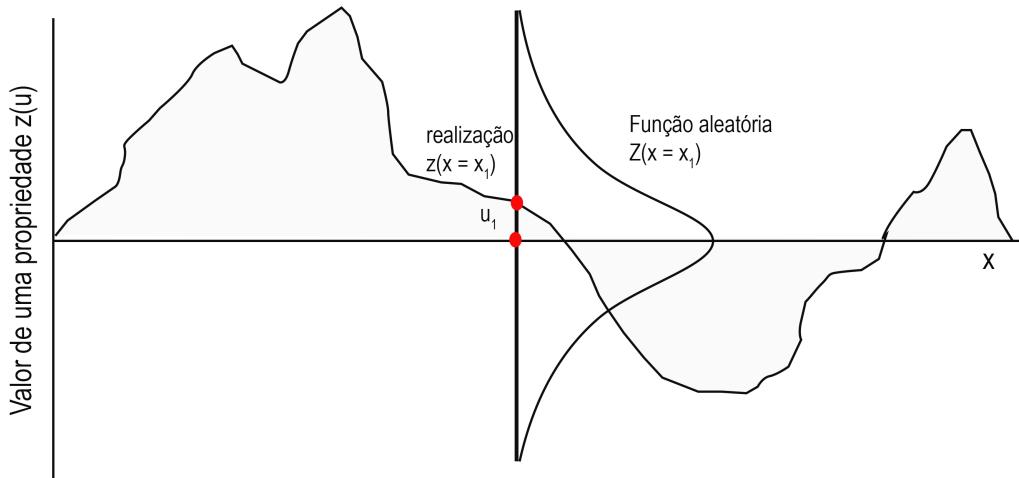


Figura 3.4: Demonstração do resultado amostrado $z(x = x_1)$ como uma realização da função aleatória $Z(x = x_1)$. No ponto x_1 o valor amostrado é apenas um resultado de uma função que desconhecemos, que associa uma distribuição de probabilidades naquele local.

Em muitos os casos não é possível conhecer esta função geradora do depósito mineral, apenas tomamos como hipótese que ela existe e é uma combinação de variáveis aleatórias em todo o espaço. Na geoestatística muitas vezes consideramos que esta função pode ser representada como uma combinação linear destas variáveis, chamada de **geoestatística linear**, ou **geoestatística clássica**. Ao tomarmos esta simplificação proposta pela teoria intrínseca, a demonstração das técnicas geoestatísticas se tornam bem mais fáceis, por isso, durante este texto, pretendemos utilizar o conceito da função aleatória em vez da forma tradicional da variável regionalizada proposta por Matheron.

Definição 3.6.1 — Função aleatória. *Uma função aleatória pode ser descrita como uma função que associa a cada ponto no espaço x uma variável aleatória $Z(x = x_1)$, sendo x_1 o ponto de coordenadas especificado.*

Esta função aleatória é composta de uma amalgama de diversas variáveis aleatórias, cada uma em um ponto do espaço. A análise geoestatística destes valores permite decompormos esta função em duas componentes principais de acordo com os valores esperados de cada uma destas variáveis. O valor esperado tende a ser o de maior probabilidade de ocorrência em um determinado local. Desta forma podemos decompor a função aleatória em duas componentes principais, o **resíduo** e o valor de **tendência**. Por definição, a função aleatória pode ser expressa por $Z(x) = R(x) + m(x)$, sendo que os resíduos possuem média igual a zero.

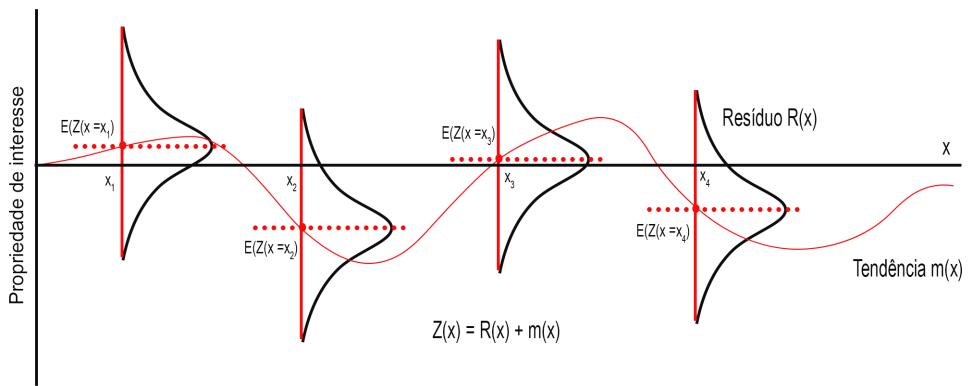


Figura 3.5: Decomposição da função aleatória a partir da determinação de sua tendência, indicada por $m(x)$, e o seu resíduo $R(x)$.

É comum na geoestatística assumirmos algumas hipóteses quanto a função aleatória. A **hipótese de estacionaridade de segunda ordem** afirma que o valor da tendência deve ser constante em todo o domínio considerado e o resíduo deve possuir variância constante para todo o domínio. Como desconhecemos a função aleatória, e nunca conseguimos determinar as variáveis aleatórias em cada ponto considerado, a hipótese de estacionaridade é sempre assumida, e nunca conseguimos comprová-la. Observe a série de números gerados na figura 3.6.

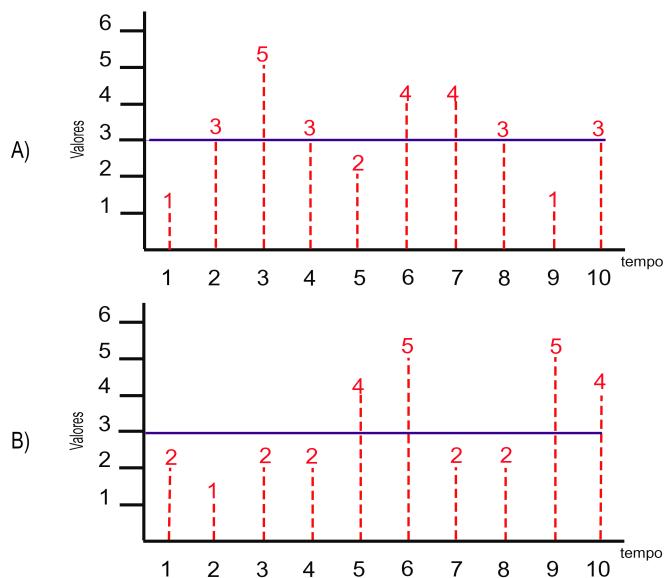


Figura 3.6: Série de números gerados em A e B. O valor médio destas séries é 2.9.

Ao observá-los, provavelmente você deve estar imaginando que foram feitos jogando-se dados na mesa. As séries A e B possuem média muito próxima do que seria de um dado de seis lados, e variam de 1 a 6. Na verdade, você está parcialmente certo,

eu gerei estes números a partir de dados. A diferença, no entanto, é que a série B foi gerada metade por um dado tetraédrico e metade por um dado cúbico, enquanto os dados da série A foram gerados apenas por um dado cúbico. Os valores médios reais que deveriam ser consideradas para este modelo são os representados na figura 3.7.

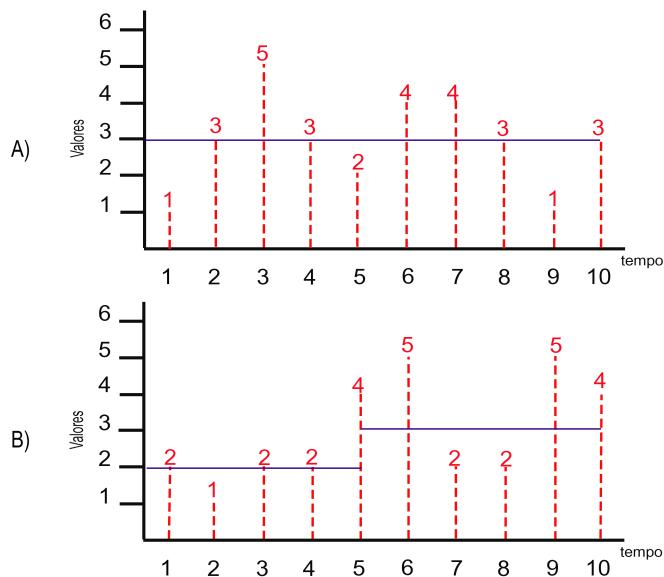


Figura 3.7: Série de números gerados. Em A) os dados foram gerados a partir de um dado tetraédrico, enquanto em B) foram gerados por um dado cúbico. Apesar dos valores médios globais serem idênticos, as médias locais são diferentes.

Apesar das médias globais serem exatamente as mesmas, as séries locais possuem distribuições distintas, com média e variância diferentes. Na verdade tanto a série B como série A poderiam ter sido geradas com o mesmo dado de seis lados. A diferença clara, quando consideramos cálculos **probabilísticos** com cálculos **estatísticos**, é que a probabilidade requer conhecimento sobre o fenômeno gerador, enquanto a estatística pretende inferir situações a partir das informações dos **dados**. Na verdade, a única informações que temos a todo momento no depósito mineral são amostras e informações indiretas como geofísica e geoquímica.



A decisão de observar uma configuração particular dos dados como estacionário como o resultado de uma função aleatória estacionária está fortemente ligada com a decisão de que estas amostras podem ser unidas juntas. Nenhuma destas decisões pode ser checada quantitativamente, não são certas ou erradas e nenhuma prova das suas validades é possível. No entanto podem ser julgadas como apropriadas ou não. Isaaks and Srivastava [1989]

Apesar de o fenômeno gerador ser completamente distinto para a metade dos

dados na série B, não é custoso unir estas diferentes distribuições sobre a mesma hipótese comum. Desta forma, assumir a estacionaridade neste caso é válido, dado que não conhecemos como estas informações foram construídas.

Em alguns casos, no entanto, não parece ser muito sábio adotar a hipótese de estacionaridade de segunda ordem. Observe a imagem da figura 3.8. A série é crescente com diferenças de valores iguais a 1 começando de um 1, 2, 3, 4, 5. Se você perguntasse para uma criança qual seria o próximo número na sequência ela diria 6. Se utilizássemos geoestatística para estimar o próximo número considerando a estacionaridade dos valores, o resultado seria 3.

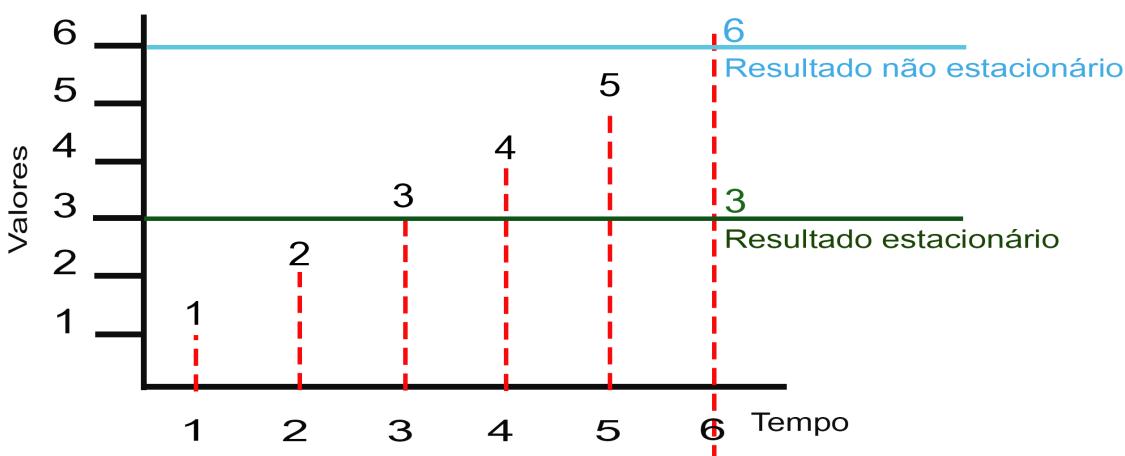


Figura 3.8: Série crescente de números. O próximo número da sequência a ser estimado considerando um modelo não estacionário seria 7, enquanto para o modelo estacionário seria de apenas 6

Não parece ser sábio adotar o número 3 neste caso. Se o fenômeno gerador desta série fosse realizado por um dado, seria tão equiprovável encontrarmos o número 3 ou o número 6 na próxima realização. No entanto a informação condicionada pela série parece nos instruir com clareza que existe este padrão a partir da observação indireta dos dados, nós desconhecemos como estes dados foram gerados. Utilizando a geoestatística nos reconhecemos que existe uma **estruturação** presente nesta sequência, que condicionalmente os dados gerados parecem seguir uma ordem, e que o próximo número gerado tente a apresentar uma variação talvez equivalente como ao dos anteriores, podendo ser 4 ou até mesmo 6.

R *Uma das ideias mais importantes na geoestatística é considerar que a função aleatória gera variáveis aleatórias condicionadas ao longo do espaço. A ideia de continuidade implica que qualquer informação próxima tende a ser mais parecida do que informações muito distantes. Isto é fisicamente plausível,*

principalmente quando pensamos na geologia. As rochas que estão próximas tendem a possuir propriedades físico-químicas muito mais semelhantes do que quando consideramos uma distância muito grande. Por isso nos é intuitivo considerar o número estimado como 6 e não 3, pois ele representa um comportamento condicionado por medidas sucessivas, logo $Pr\{Z(x_6) = 6|z(x_1) = 1, z(x_2) = 2, z(x_3) = 3, z(x_4) = 4, z(x_5) = 5\} > Pr\{Z(x_6) = 3|z(x_1) = 1, z(x_2) = 2, z(x_3) = 3, z(x_4) = 4, z(x_5) = 5\}$

Estes casos também são chamados na geoestatística de **deriva**, ou seja, que existem mudanças graduais na tendência dos dados. Em alguns casos é bastante lógico na mineração considerar a deriva. A topografia, por exemplo, quando analisada em determinadas escalas e situações pode ser continuamente ascendente ou descendente. Neste caso descartar a hipótese de estacionaridade de segunda ordem é sábio.

R *O custo de aceitar o uso de um modelo inapropriado é que as propriedades estatísticas dos valores estimados divergirão de modelos homólogos Isaaks and Srivastava [1989]*

3.7 Hipótese de estacionaridade

Como visto anteriormente podemos realizar hipóteses a respeito da função aleatória, geradora dos fenômenos geoestatísticos. Estas hipóteses são decisões que não podem ser numericamente definidas, mas que em casos convém serem julgadas, para que as estimativas não retornem valores não condizentes com a realidade. A escolha de um tipo de estacionaridade significa que adotamos um critério que considere um conjunto de dados com um comportamento **homogêneo**. Uma das hipóteses utilizada pela geoestatística mais importantes, e que não constitui critério de escolha, é a chamada de **hipótese estrita**. Diferentemente da hipótese de estacionaridade de segunda ordem, esta é adotada automaticamente quando se opta por um método geoestatístico e não é passível de decisão. A principal ideia da estacionaridade estrita é que o fenômeno é homogêneo em uma mesma direção no espaço, sendo **invariante por translação**.

R *A ideia de areia em um jarro é uma boa imagem da estacionaridade de uma função aleatória em três dimensões, pelo menos enquanto a areia estiver bem ordenada (de outra forma se esta jarra vibrar, os grãos finos se depositarão na base, criando não estacionaridade vertical) Chiles and Delfiner [2009]*

Uma forma geométrica de pensarmos na hipótese estrita é pelo uso de fractais. Fractais são figuras geométricas autosimilares, em que cada um de seus componentes

carregam características da informação como um todo. A figura 3.9 representa um fractal. Estas formas autosimilares são muito comuns na natureza, seja no padrão desenhado por cristais de gelo, pela forma das plantas e principalmente nas rochas. A geologia em pequena escala muitas vezes é uma repetição que se traduz em grande escala.

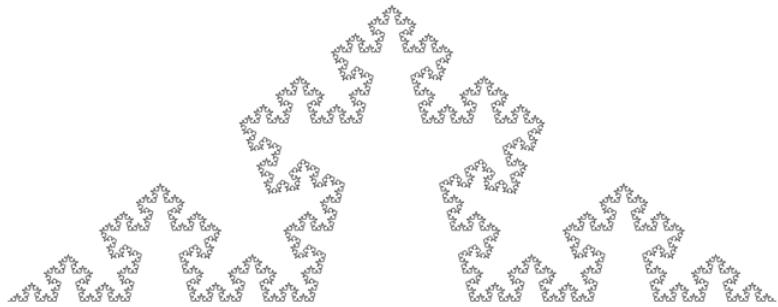


Figura 3.9: Fractal gerado a partir da repetição sistemática de estruturas cada vez menores.

Proposição 3.7.1 *Todo o conhecimento humano somente advém do entendimento de padrões. As diferentes disciplinas, sejam elas humanas, biológicas ou exatas, apenas diferenciam quanto ao objeto de estudo. Não há diferença nenhuma entre um físico que entende padrões referentes ao movimento de planetas, um linguista que estuda o padrão de idiomas, um historiador que verifica padrões no tempo, ou um matemático que verifica o padrão das formas. A natureza também age desta forma, pois esperamos acordar no dia seguinte com o sol sobre as montanhas. Até mesmo dentro de fenômenos que parecem ser puramente aleatórios, podemos encontrar motivos pelos quais podemos entender padrões. Independente de fenômenos serem estacionários ou não estacionários, a geoestatística procura simplesmente estas formas no espaço, representações que apesar de não serem físicas, são mímicas da natureza da existência destes fenômenos*

Esta repetição de comportamentos em uma direção leva a seguinte definição matemática. Um fenômeno dito estacionário estrito significa que $Pr\{Z(x_1) < z(x_1), Z(x_2) < z(x_2), \dots, Z(x_k) < z(x_k)\} = Pr\{Z(x_{1+h}) < z(x_{1+h}), Z(x_{2+h}) < z(x_{2+h}), \dots, Z(x_{k+h}) < z(x_{k+h})\}$, sendo h um vetor de direção determinada. A hipótese de estacionariedade estrita significa que existe um grau de repetição no comportamento da variável ao longo de uma direção, no entanto, o fenômeno espacial pode apresentar deriva. Outra questão a ser abordada é o fato de que a adoção da estacionariedade é dependente da escala analisada. Um fenômeno considerado não estacionário pode assumir comportamento estacionário local. Observa a série de

dados representada pela figura 3.10. Quando analisado o comportamento global da função aleatória esta apresenta nitidamente uma tendência nos dados, no entanto, quando considerada uma escala menor do vetor h , este mesmo comportamento pode ser tomado como estacionário. Este fenômeno também é chamado de **quasi estacionário**.

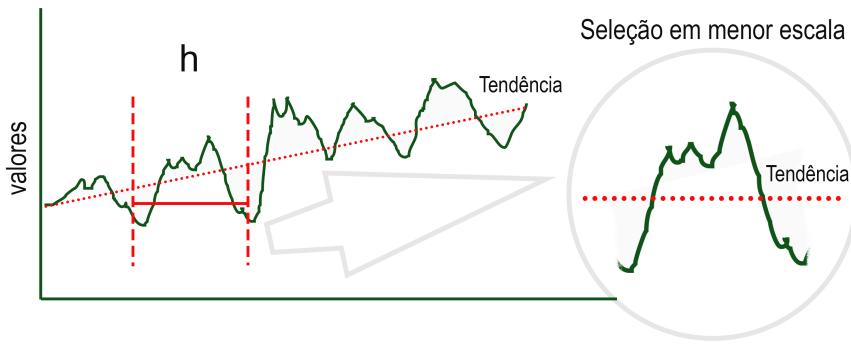


Figura 3.10: Comportamento analisado de uma série não estacionária quando analisada em um domínio global, e estacionária quando analisada em um domínio menor de comprimento h .

Proposição 3.7.2 *Uma das maiores contribuições da geoestatística para a ciência talvez tenha sido a concepção de que os fenômenos podem ser dependentes da escala analisada. Dependendo da observação nossas hipóteses a respeito do fenômeno podem mudar. Isto é fisicamente compatível com a ideia da geologia. Analisar um depósito mineral em uma grande extensão de área, com toda a certeza é diferente quando observamos variações de tamanho centimétrico. A própria observação da Terra quando vista do espaço apresenta belos tons azuis e brancos, mas quando aproximamos a escala de uma região do tamanho de um país, notamos como nossa visão é diferente e muito mais variável.*

A hipótese de estacionaridade estrita é uma hipótese realizada sobre a característica do fenômeno, não dos resultados das amostras. A hipótese de estacionaridade de segunda ordem, no entanto, é uma hipótese relacionada com os **momentos estatísticos** do fenômeno. A principal ideia dos momentos estatísticos é que eles representam de alguma forma o resumo da distância entre os dados, desta forma quando pensamos na estacionaridade intrínseca, ou na estacionaridade de segunda ordem, pensamos na possível homogeneidade da distância entre os dados.

O conceito de **estacionaridade intrínseca**, desta forma, apresenta também outra forma de conceber esta homogeneidade, quando estabelecemos que uma variação $Y_h(x) = Z(x + h) - Z(x)$ é estacionária de segunda ordem. Em outras palavras dizemos que existe homogeneidade quando consideramos a diferenças entre variáveis

aleatórias geradas pela função aleatória. Segundo [Chiles and Delfiner \[2009\]](#), se a hipótese de estacionaridade intrínseca pode ser considerada e não ocorre uma tendência, então o valor médio da função aleatória é constante, e o valor esperado de $Y_h(x)$ é zero.

3.8 Momentos estatísticos

Como dito anteriormente, momentos estatísticos são representações da distância entre dados. As medidas de distância, podem ser por exemplo, medidas da tendência central dos dados ou medidas da dispersão destes dados. A figura 3.11 representa os conceitos de **tendência central** e de **dispersão**.

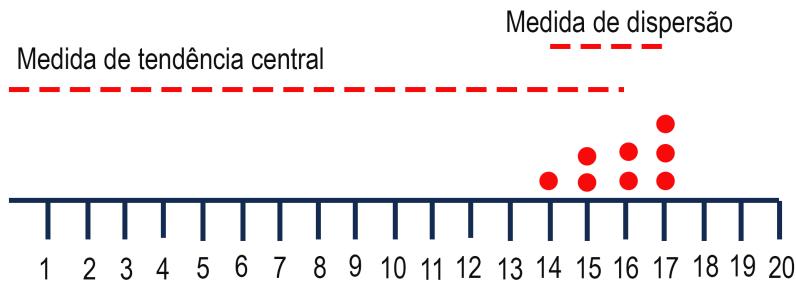


Figura 3.11: Exemplos de momentos estatísticos para um conjunto de dados. O valor médio representa o quanto distante está o centro dos dados, enquanto a dispersão apresenta quanto agregados estão estes dados.

A principal medida da distância do centro dos dados é chamada de **esperança matemática**, e pode ser representada para variáveis contínuas como.

$$E(Z) = \int_{z=-\infty}^{z=+\infty} z f(z) dz \quad (3.4)$$

No caso de variáveis discretas, podemos determinar a esperança matemática como

$$E(Z) = \sum_{i=-\infty}^{+\infty} Pr(z_i) z_i \quad (3.5)$$

Muitas vezes há confusões ao se dizer que a esperança matemática representa o valor mais provável que determinada variável pode possuir. No entanto, a esperança

matemática é simplesmente uma medida da distância do centro dos dados, sendo que este centro pode ser pouco provável ou nem mesmo existir. Observe a figura 3.12. A esperança matemática neste caso representa o centro de uma distribuição com probabilidade muito baixa de ocorrência.

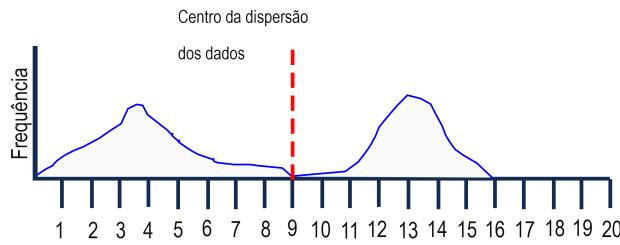


Figura 3.12: Exemplo da esperança matemática de uma distribuição multimodal. O valor da probabilidade para este centro de dispersão é praticamente nulo.

O que de fato ocorre é que os centros de dispersão dos dados da maioria dos problemas de engenharia não são multimodais, ou seja, apresentam vários picos nas distribuições de densidade de probabilidade com na figura. Neste caso a esperança matemática pode representar os valores mais prováveis de ocorrência da dispersão dos dados.

Pela definição da esperança matemática, algumas propriedades podem ser diretamente derivadas. A multiplicação da variável aleatória Z por um valor constante c , implica na seguinte condição.

$$E(cZ) = cE(Z) \quad (3.6)$$

O valor esperado de uma variável aleatória constante pode ser relacionada pela seguinte propriedade

$$E(c) = c \quad (3.7)$$

A demonstração, na verdade é muito simples, já que advém da própria definição de esperança matemática

Demonstração. Valor esperado de uma constante é igual a ela mesma

$$\begin{aligned} E(c) &= \sum_{i=-\infty}^{+\infty} Pr(c)c \\ E(c) &= c \sum_{i=-\infty}^{+\infty} Pr(c) \\ \text{Como: } \sum_{i=-\infty}^{+\infty} Pr(c) &= 1 \\ E(c) &= c \end{aligned}$$

■

A partir da definição da esperança matemática como uma medida de centralidade da distribuição dos dados, podemos derivar outros momentos representando diferentes distâncias desta distribuição. Os diferentes momentos matemáticos podem caracterizar diferentes distâncias relativas à **centralidade**, **dispersão**, **assimetria**, **forma**. A geoestatística, na verdade, foca sua análise principalmente nos momentos de primeira e segunda ordem.

R *Em aplicações da mineração, a lei de probabilidades espacial nunca é requerida, principalmente porque os dois primeiros momentos da função são suficientes para providenciar uma solução aproximada para muitos problemas encontrados Journel and Huijbregts [1978]*

Outro momento estatístico importante é a variância, definida como o momento de segunda ordem centrado. A variância pode ser considerada como uma medida de dispersão, demonstrada por

$$Var(Z) = E(Z - E(Z))^2 \quad (3.8)$$

Esta forma tradicional da variância pode ser substituída por outra representação a partir de

$$Var(Z) = E(Z^2) - E(Z)^2 \quad (3.9)$$

A prova desta relação também é facilmente demonstrada a partir das propriedades da esperança matemática e pela definição da variância.

Demonstração. Relação entre as equações 3.8 e 3.9

$$\begin{aligned}Var(Z) &= E(Z - E(Z))^2 \\Var(Z) &= E(Z^2 - 2ZE(Z) + E(Z)^2) \\Var(Z) &= E(Z^2) - E(2ZE(Z)) + E(Z)^2 \\Var(Z) &= E(Z^2) - 2E(Z)E(Z) + E(Z)^2 \\Var(Z) &= E(Z^2) - 2E(Z)^2 + E(Z)^2 \\Var(Z) &= E(Z^2) - E(Z)^2\end{aligned}$$

■

Os momentos estatísticos de primeira e segunda ordem, representados pela esperança matemática e pela variância representam medidas tomadas de uma única variável aleatória. Para relacionar diferentes variáveis aleatórias utilizamos comumente a **covariância**, esta representada pela similaridade entre duas variáveis aleatórias. Considere as variáveis aleatórias Z e Y . Podemos representar a covariância pela seguinte relação

$$Cov(Z, Y) = E((Z - E(Z))(Y - E(Y))) \quad (3.10)$$

Se as variáveis Z e Y apresentam médias idênticas iguais a m , então a covariância pode ser representada por

$$Cov(Z, Y) = E(ZY) - m^2 \quad (3.11)$$

A prova desta relação pode ser facilmente obtida

Demonstração. Relação da Covariância considerando médias idênticas iguais a m

$$\begin{aligned}Como: E(Z) &= E(Y) = m \\Cov(Z, Y) &= E((Z - m)(Y - m)) \\Cov(Z, Y) &= E(ZY - Zm - Ym + m^2) \\Cov(Z, Y) &= E(ZY) - E(Zm) - E(Ym) + E(m^2) \\Cov(Z, Y) &= E(ZY) - mE(Z) - mE(Y) + E(m^2) \\Cov(Z, Y) &= E(ZY) - m^2 - m^2 + m^2 \\Cov(Z, Y) &= E(ZY) - m^2\end{aligned}$$

Se as variáveis Y e Z forem idênticas, a covariância entre as duas variáveis aleatórias é equivalente a variância. A prova pode ser demonstrada por

Demonstração. Prova de que a covariância é idêntica a variância para $Z=Y$

$$\text{Como : } Y = Z \rightarrow Cov(Z, Y) = Cov(Z, Z)$$

$$C(Z, Z) = E((Z - E(Z))(Z - E(Z)))$$

$$C(Z, Z) = E(Z - E(Z))^2$$

$$C(Z, Z) = Var(Z) \vee Var(Y)$$

3.9 Ergocidade

A ergocidade é uma das propriedades mais importantes da função aleatória. A ideia é de que cada vez ao qual analisamos um volume maior no espaço, a tendência é que o valor médio deste volume se aproxime cada vez mais do valor médio do fenômeno. Matematicamente podemos definir a propriedade da ergocidade como

$$\lim_{V \rightarrow \infty} \frac{1}{|V|} \int_{x \in V} Z(x) dx = m \quad (3.12)$$

Em que $|V|$ é o volume considerado e m o valor da média do fenômeno.

Definição 3.9.1 — Ergocidade. A Ergocidade pode ser caracterizada como a propriedade da função aleatória de convergência dos valores médios se aproxime a um valor constante m , de acordo com um domínio V considerado.

Alguns fenômenos tendem a apresentar dispersões infinitas, crescentes de acordo com o desenvolvimento da função aleatória. Estes fenômenos podem apresentar dispersão infinita, tal como o fenômeno de movimento browniano.

3.10 Homocedasticidade e heterocedasticidade

Além do comportamento da estacionariedade dos valores médios da função aleatória, também é importante qualificar os fenômenos geoestatísticos a partir do comportamento da variância. A figura 3.13, por exemplo, demonstra um comportamento crescente da variância de acordo com o desenvolvimento da série.

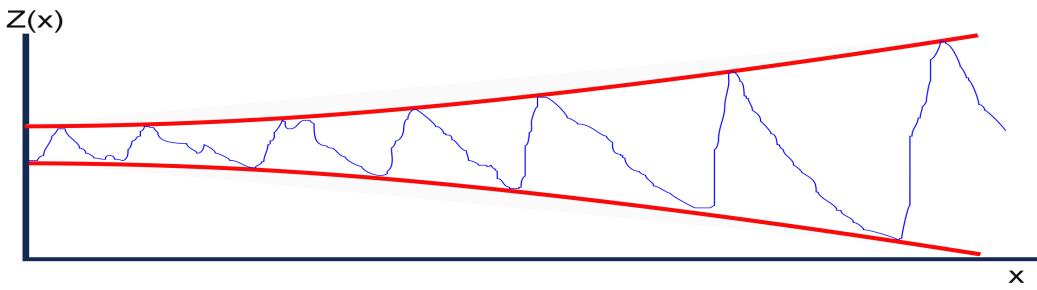


Figura 3.13: Fenômeno de heterocedasticidade representando variabilidade infinita da função aleatória

Estes fenômenos são chamados de **heterocedásticos**, e aumentam a variabilidade de acordo com o incremento da direção. Fenômenos constantes quanto a dispersão local são chamados de **homocedásticos**

Definição 3.10.1 — homocedasticidade. A hipótese de homocedasticidade, igual (*homo*) dispersão (*scedasticidade*), implica que a variância da função aleatória é constante para todo e qualquer ponto x representado no domínio D . A heterocedasticidade, no entanto, implica que a variância aumenta ao longo da função aleatória.

3.11 Relação Volume Variância

Na geoestatística, a variável aleatória se manifesta em todos os pontos no espaço. No entanto, nem sempre é possível reconhecer a variável em um suporte pontual, e para fins de engenharia precisamos entender a variável aleatória dentro de domínios específicos, sejam eles os domínios da amostra, na unidade seletiva de lavra, ou dentro de um domínio de estimativa. Na geoestatística clássica, apenas se determina os **valores esperados** destas variáveis dentro de um domínio, principalmente os de primeira e segunda ordem, não se importando com o reconhecimento das distribuições locais.

O processo de se determinar volumes esperados dentro de um domínio é chamado na geoestatística de **regularização**.

R *Muito raramente, em prática, o valor dos dados pontuais $z(x)$ estão disponíveis. Mais comumente o valor dos dados $z_v(x)$ em um certo suporte $v(x)$ estão disponíveis, como por exemplo uma amostra de testemunho, ou mais genericamente o volume de uma amostra. O valor médio $z_v(x)$ é chamado de regularização das variáveis pontuais $z(y)$ dentro do domínio $v(x)$ Journel and Huijbregts [1978]*

A regularização permite com que medidas realizadas dentro de um domínio estipulado possuam mesmo volume, permitindo com que suas propriedades sejam compatíveis para fins de estimativa na geoestatística.

Assumindo que a função aletória é contínua, e que uma combinação linear de variáveis aleatórias pode ser expressa, podemos definir um valor regularizado em um espaço amostral, definido pelo seu suporte. Considere v como o suporte amostral, logo o seu valor regularizado pode ser descrito como

$$Z_v = \frac{1}{|v|} \int_{x \in v} Z(x) dx \quad (3.13)$$

Da mesma forma o valor regularizado dentro da unidade seletiva de lavra pode ser definido por

$$Z_V = \frac{1}{|V|} \int_{x \in V} Z(x) dx \quad (3.14)$$

Em último caso podemos definir o valor médio dentro do domínio de estimativa

$$Z_D = \frac{1}{|D|} \int_{x \in D} Z(x) dx = m \quad (3.15)$$

Em que m é o valor esperado do fenômeno considerado, e constante, se considerada a propriedade da ergocidade. Considerando os diferentes domínios v , V , e D , podemos determinar três diferentes relações $(v|V)$, $(v|D)$ ou $(V|D)$, representadas pelas relações entre amostras e unidade seletiva de lavra, amostras e domínio de estimativa e unidades seletivas de lavra e domínio de estimativa.

Para indicarmos a variabilidade ao qual os valores amostras em um domínio estão dispersos quanto um valor de referência estimado, utilizamos uma estatística chamada de **variância de dispersão**, denotada pela letra D^2 . A ideia da variância está diretamente associada ao conceito de **entropia**, ou grau de desorganização.

Proposição 3.11.1 *A variância de dispersão é uma das medidas mais importantes na geoestatística e está associada ao conceito de entropia, ou de desorganização dos dados. Quando você considera, por exemplo, a variabilidade de um pixel de uma foto em relação ao seu valor médio, com toda a certeza este será mais disperso que valores médios de partes do corpo na foto, como rostos e mãos, em relação a este valor central. A ideia de que nosso conhecimento sobre um fenômeno pode ser afetado pela dispersão da informação é essencial, principalmente nas técnicas de*

mudança de suporte que serão vistas futuramente.

A **variância de dispersão** é portanto uma medida da variabilidade entre estes domínios, considerando os valores regularizados. A variância de dispersão amostra e domínio de estimativa pode ser definida por

$$D^2(v|V) = \frac{1}{N} \sum_{i \in V} [Z_{v_i} - Z_V]^2 \quad (3.16)$$

Sendo N o número de pontos amostrais regularizados de suporte v dentro do domínio estimado V . Se considerarmos o suporte $(.)$ como o suporte pontual, podemos definir a variância de dispersão ponto amostra por

$$D^2(.|v) = \frac{1}{|v|} \sum_{x \in v} [Z(x) - Z_v]^2 \quad (3.17)$$

Em que $|v|$ é o volume constituído pelo suporte amostral v e todos os seus pontos internos. E a variância de dispersão ponto e domínio estimado por

$$D^2(.|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x) - Z_V]^2 \quad (3.18)$$

Em que $|V|$ é o volume constituído pelo suporte amostral V e todos os seus pontos internos. Uma das relações importantes da variância de dispersão pode ser determinada pela diferença entre variâncias de dois suportes, tal como

Demonstração. Prova da relação da variância de dispersão entre ponto e bloco estimado como a diferença entre a variância do fenômeno e da variância do bloco

estimado.

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x) - Z_V]^2$$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - 2Z(x)Z_V + Z_V^2]$$

como $\sum_{x \in V} Z(x)Z_V = Z_V^2$, tal que $Z_V = constante$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - Z_V^2]$$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} ([Z(x)^2 - m^2] - [Z_V^2 - m^2])$$

Pela hipótese de estacionaridade de segunda ordem:

$$\frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - m^2] = Var(Z(x)) = s^2(\cdot|.) , \text{ e}$$

$$[Z_V^2 - m^2] = Var(Z(V)) = s^2(V|V) , \text{ logo}$$

$$D^2(\cdot|V) = D^2(\cdot|.) - D^2(V|V)$$

■

Analogamente as relações $s^2(\cdot|v) = s^2(\cdot|.) - s^2(v|v)$ e $s^2(v|V) = s^2(V|V) - s^2(v|v)$ podem ser derivadas. Podemos encontrar então a seguinte identidade

$$s^2(\cdot|V) = s^2(\cdot|v) + s^2(v|V) \tag{3.19}$$

Esta também é chamada de **relação de krige** ou relação da aditividade de variâncias de krige. Quando consideramos a dispersão de valores de uma variável em domínios maiores como V , esta tende a ser maior que consideramos no suporte amostral v . Este princípio também é chamado de **volume e variância**, ou seja, quanto maior for a diferença entre os suportes amostrais e o domínio de estimativa, menor será nossa acurácia nestas previsões. A ideia da variabilidade de acordo com a mudança do volume estimado ou do suporte amostral está diretamente associada à definição de uma imagem, no conceito da geoestatística. Observe a figura 3.14. Em A) possuímos os valores exatos do fenômeno estudado. Podemos notar que os resultados são uma representação fiel de uma representação física, podem ser realmente consideradas um "mapa" dos valores distribuídos no espaço. Em B) verificamos os valores estimados, que se apresentam de forma pixelada e não apresentam

uma definição adequada do problema. No entanto, cada bloco estimado no mapa B) guarda uma correlação alta com os valores médios tomados da região no mapa A). Dizemos que as estimativas geoestatísticas não são uma ferramenta boa para produzir "mapas", já que estes são reproduções fidedignas dos fenômenos espaciais, mas o valor esperado dentro de um bloco em B) tende a ser cada vez mais próximo do valor médio real na região quanto menor for a definição e maior o tamanho do bloco.

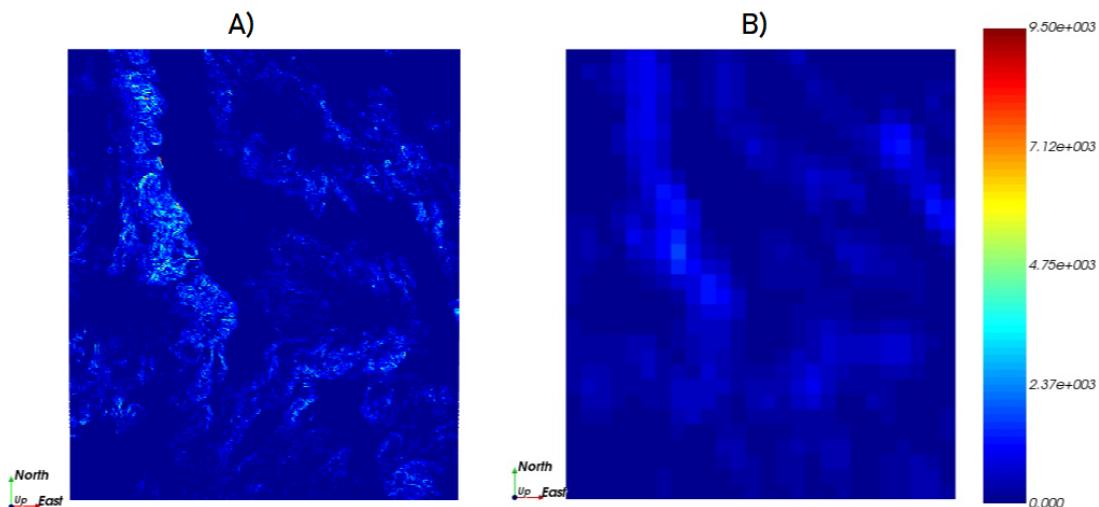


Figura 3.14: Relação do conceito de volume e variância apresentado em imagens. Em A) possuímos o valor exaustivo de um banco de dados, enquanto em B) apresentamos os valores krigados. É possível notar que as estimativas não reproduzem as feições naturais do fenômeno, no entanto, cada bloco estimado é

3.12 Conclusões

Neste capítulo aprendemos um pouco sobre a teoria das variáveis regionalizadas, um conceito determinado na década de 70 pelo professor George Matheron, e que evoluiu ao longo do tempo, facilitando o estudo de variáveis georeferenciadas. Estes conceitos iniciais são abstratos, porém poderosos, pois permitem constituir as bases de hipóteses utilizadas nos modelos geoestatísticos.

3.13 Exercícios

Exercícios 3.1 Enumere em uma lista todas as variáveis aleatórias regionalizadas

que você possui em seu objeto de estudo. Indique ao lado se elas são somáticas ou não. Ex.: Teor-> somático, Condutibilidade hidráulica -> não somático. ■

Exercícios 3.2 Cinco ações de uma mineradora possuem rentabilidade de 5, 10, 20, 4 e 5 Unidades monetárias. Se a probabilidade de renda destas ações forem iguais a 40%, 35%, 10%, 10% e 5% qual é o valor esperado para a renda de todas as ações. Resp.: 8.15 UM ■

Exercícios 3.3 Cinco amostras possuem valor de teor iguais a 2%, 2.5%, 2.3%, 2.1% e 2.7%. Se o volume das amostras é de 5, 4, 3, 5 e 7 cm^3 qual é o teor médio das amostras. Resp.: 2,34% ■

Exercícios 3.4 Prove que o valor do resíduo da função aleatória é ortogonal à sua tendência, ou seja $Cov(R, m) = 0 \forall x \in D$ sendo D o domínio do depósito. ■

Exercícios 3.5 Prove que a covariância de duas variáveis aleatórias independentes seja igual a zero. Dica.: Tome o valor de $E(XY) = E(X)E(Y)$ ■



4. Estatística univariada

Estatística: a ciência que diz que se eu comi um frango e tu não comeste nenhum, teremos comido, em média, meio frango cada um.

Pitigrilli

4.1 Introdução

As avaliações geoestatísticas geralmente se iniciam com uma avaliação global das amostras. Nesta primeira etapa, o objetivo principal é **descrever** e **inferir** informações sobre o comportamento geral das amostras. A chamada **estatística descritiva** representa o conjunto de técnicas necessárias para resumir informações da realidade observada das amostras, usando formas numéricas ou gráficas para caracterizá-las. Já a chamada estatística inferencial ocupa em tomar inferências da população de dados a partir de informações das amostras. O estudo sistemático das variáveis em termos globais não representam o fenômeno estudado, mas partem do ponto de vista necessário para o início da pesquisa, podendo avaliar inconsistências nos dados e possíveis comportamentos que possam indicar situações favoráveis ou desfavoráveis na análise espacial.

Proposição 4.1.1 *Usualmente, os sistemas aos quais estudamos não podem ser isolados em variáveis discretas e independentes. Estes fatores influenciam os primeiros*

passos da pesquisa, em como e onde coletar espécies ou observações - Borradaile [2013]

A palavra estatística, non entanto, apresenta duplo sentido. Pode representar a **teoria estatística** ou as medidas realizadas pelos dados. Alguns conceitos iniciais são de extrema importância quando consideramos o uso da estatística clássica univariada

Definição 4.1.1 — População. *conjunto de elementos que tem pelo menos uma característica em comum. No caso da geoestatística a população pode ser considerada analogamente ao conjunto possível de todas as realizações em um domínio geológico considerado*

Definição 4.1.2 — Amostra. *Amostra pode ser considerada como um subconjunto de elementos de uma população. Existem diferentes tipos de amostragens na mineração, como sondagens diamantadas, amostragens de canal, medições de nível freático, etc.*

Em muitos casos é comum representar este conjunto de dados por tabelas. Sumários que caracterizam as informações de cada subconjunto de amostras no espaço. Muitos softwares de geoestatística e planejamento mineral caracterizam os furos de sondagem a partir de dois ou três arquivos. Geralmente o primeiro arquivo consta uma tabela sobre o posicionamento da boca dos furos na superfície, caracterizando seu posicionamento espacial em um plano cartesiano $\langle x,y,z \rangle$. O segundo arquivo geralmente representa a direção dos furos e o comprimento realizado em cada manobra. E um terceiro arquivo geralmente apresenta as propriedades medidas em cada manobra realizada do testemunho.

Uma questão importante a ser considerada nas estatísticas descritivas é sua capacidade de resumo da informação. Estatísticas numéricas são uma alternativa importante para formar concepções que auxiliam na tomada de decisão, mas ao mesmo tempo reduzem a sensibilidade sobre outras questões dos dados. Imagine a figura 4.1. Temos duas fontes de temperatura equidistantes, uma com 270° e outra a -270° . Apesar da diferença abrupta de temperaturas, a temperatura média da parede entre elas é apenas 0° . Um ser humano conseguiria sobreviver facilmente se ocupasse apenas o espaço entre estas duas fontes de temperatura, mas morreria se afastasse delas.

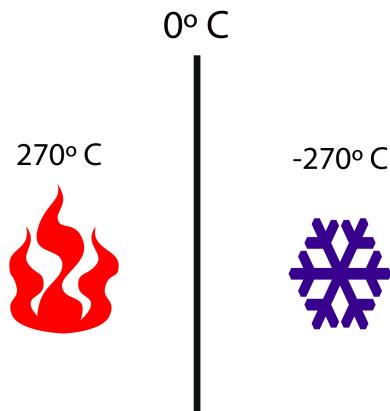


Figura 4.1: Duas fontes de temperatura equidistantes, uma quente e outra fria. A temperatura média da parede que separa estas fontes é igual a média das temperaturas, o que não representa toda a complexidade do fenômeno.

Apesar da média ser uma medida muito útil para ser utilizada na descrição dos dados, utilizada sozinha pode gerar interpretações erradas sobre o problema. Convenciona-se que o uso de estatísticas descritivas deve ser múltiplo, optando por utilizar não apenas uma, mas diferentes técnicas de avaliação.

O uso de estatísticas descritivas permite em muitos casos

1. Avaliar se as proporções globais possam estar acima do cut-off esperado
2. Identificar a facilidade da aplicação dos métodos clássicos de acordo com as distribuições de frequência
3. Auxiliar no dimensionamento de malhas de sondagem principalmente nas etapas iniciais (*greenfield*) na mineração
4. Identificar a possibilidade da divisão de domínios se apresentadas frequências multimodais

Proposição 4.1.2 *Estatísticas univariadas são a primeira alternativa para analisar dados. Quando as amostras ainda são escassas, principalmente nas fases iniciais da pesquisa mineral, estas ferramentas são extremamente úteis para avaliarem de forma genérica os resultados das campanhas. Se bons resultados podem ser gerados a partir de estatísticas univariadas, a confiança no projeto aumenta suas perspectivas, no entanto, se os dados demonstrarem condições pobre das estatísticas, ainda podemos apostar em uma melhor avaliação do depósito*

É importante salientar que informações a partir de estimativa e interpolação não podem gerar dados além dos limites estipulados pela estatística descritiva univariada. Qualquer método de inferência não extrapola os valores mínimos e máximos de um depósito mineral. Descrever é antes de tudo um passo que necessita encontrar propriedades de algo. A descrição deve conter os aspectos mais importantes de um depósito mineral, tal como mínimo e máximo encontrados, valores médios, dispersão. Da mesma forma que desenhar é uma atividade altamente explicativa para descrever um problema, as estatísticas gráficas desempenham papel fundamental na avaliação inicial.

4.2 Estatísticas pontuais

Como dito anteriormente, o conceito estatística pode ser dúvida, ao mesmo tempo que enfoca na 'teoria estatística' ou em **funções aplicadas em dados**. Quando estas funções forem aplicadas em todos os dados de um universo são chamadas de **parâmetros**. Qualquer função realizada a partir de dados pode ser considerada uma estatística, ou um **estimador**, no entanto, algumas delas são mais usuais, por conseguirem a partir de dados aproximar as estatísticas de seus respectivos parâmetros.

Outra forma de resumir e descrever os dados é através de estatísticas pontuais. Elas resumem a informação do conjunto de amostras em uma única medida descrevendo-o como um todo.

Definição 4.2.1 — Estatísticas pontuais. *Estatísticas pontuais são funções realizadas a partir dos dados para calcular valores que representam propriedades do conjunto. Dentre as categorias mais conhecidas possuímos medidas de centralidade, dispersão, assimetria, achatamento*

Se fôssemos comparar a descrição pontual com o retrato falado de um criminoso, cada estatística seria apenas uma parte do rosto, a média o nariz e a variância as orelhas, por exemplo. Uma das ferramentas utilizadas para entender estas estatísticas visualmente também é conhecida como faces de Chernoff, como demonstrado na figura 4.2

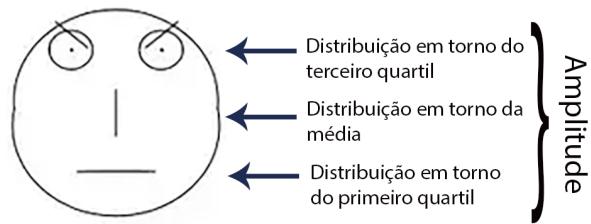


Figura 4.2: Exemplo das faces de Chernoff e as características das estatísticas com estruturas da face.

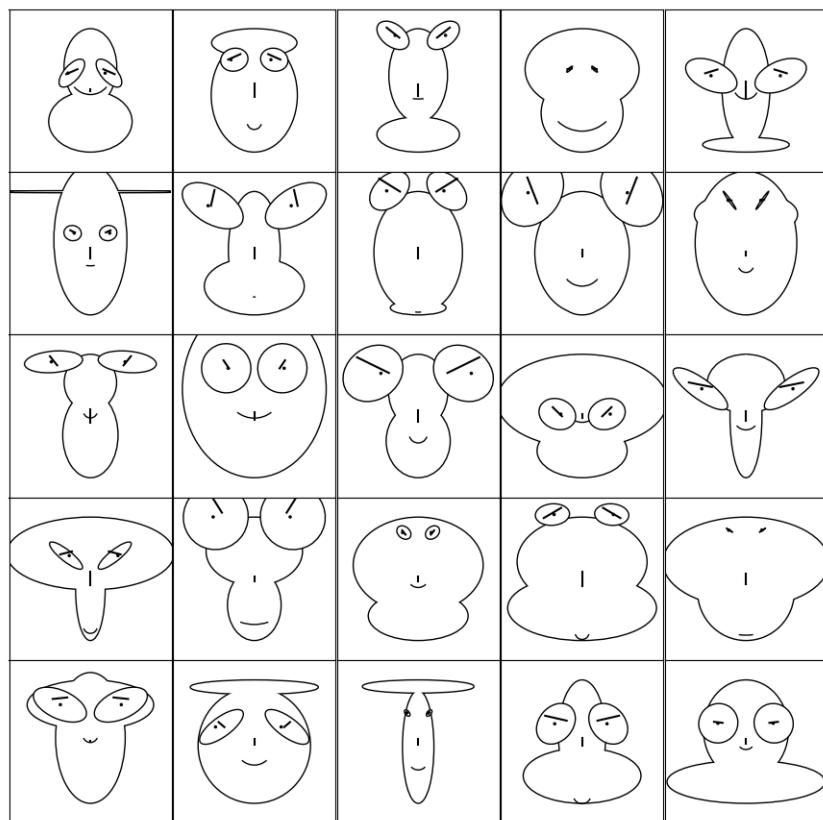


Figura 4.3: Diferentes faces de Chernoff para diferentes variáveis

Proposição 4.2.1 *Faces de Chernoff foram criadas por Herman Chernoff em 1973, como uma forma de representar dados multivariados de forma a ser discernido facilmente por um observador humano. As faces constituem em linhas desenhadas em duas dimensões que contém uma série de estruturas faciais.* - Morris et al. [2000]

É importante salientar que apenas uma estatística pontual não é uma medida que garante informação completa a respeito de um conjunto de dados. Um depósito mineral pode ter valor médio de 50g de ouro por tonelada, enquanto outro tenha 45g

de ouro por tonelada, e ainda assim o segundo depósito seja mais rico, pois a análise deve ser realizada sobre as proporções gerais dos dados. Isso acontece porque as medidas pontuais de tendência central como a média devem estar sempre associadas com uma medida de dispersão. Se o depósito de 50 g por tonelada possuir uma menor dispersão, e o depósito de 45 g/ton possuir uma maior, para um dado cut-off o depósito de 45g/ton pode ser mais rico.

A Figura (4.4) demonstra esta situação graficamente. Notamos que a distribuição A, apesar de possuir uma média menor que a distribuição B, ainda assim relata um depósito mais rico para o cut-off considerado.

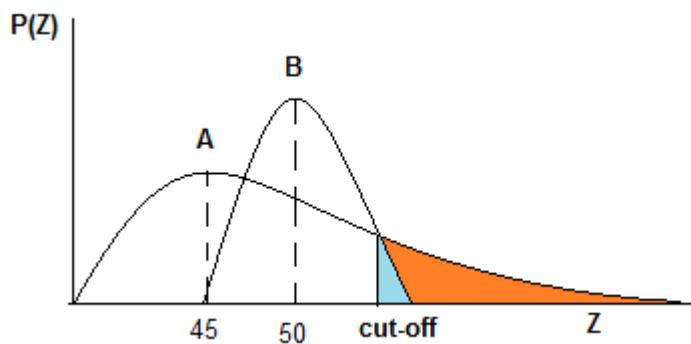


Figura 4.4: Exemplo de duas distribuições A e B relatando um depósito mais rico A com média menor que B. Área azul mostrando a contribuição da distribuição B acima do cut-off e área laranja mostrando a contribuição de A acima do cut-off

4.2.1 Medidas de tendência central

As medidas de tendência central são estatísticas calculadas a partir das amostras que representam o centro de massa do conjunto. Analogamente ao ponto de equilíbrio de uma barra, estas representam o centro de dispersão dos dados. Note que esta é uma convenção matemática. O valor médio não representa necessariamente um valor do conjunto de amostras e nem tão pouco pode representar um valor mais provável, mas apenas um centro da dispersão dos dados.

Proposição 4.2.2 *Se lançarmos um dado de seis lados centenas de vezes, e anotarmos o valor realizado em cada jogada, teremos uma tabela com cada número e sua possível frequência. É esperado que a média deste conjunto de dados seja 3.5, pois a frequência entre os números obtidos nos lançamentos será aproximadamente parecida $(6 + 1)/2$. Este valor apesar não é real, pois não podemos obter metades de uma face de um dado, mas representa o centro de dispersão destes valores.*

As medidas de tendência central mais comuns são a média aritmética, a moda,

a média ponderada e a mediana.

Média aritmética

A média aritmética pode ser descrita segundo a equação (4.2) em que x são os valores das amostras e n o número de amostras. Se a média aritmética for calculada a partir de uma população finita de todos os seus elementos a média \bar{x} é equivalente ao valor esperado da variável $\mu = E(X)$.

$$\bar{x} = \frac{1}{n} \sum_{i=0}^n x_i \quad (4.1)$$

Muitas vezes é necessário calcular a média aritmética de um agrupamento de dados a partir de um histograma, por exemplo. Neste caso podemos calcular a média aritmética como

$$\bar{x} = \frac{1}{n} (f_1 c_1 + f_2 c_2 + \dots + f_n c_n) = \frac{1}{n} \sum_{i=0}^n f_i c_i \quad (4.2)$$

Em que f_i é a frequência de cada classe c_i .

Moda

Para variáveis inteiras, a informação mais importante é a frequência de cada valor da variável. Neste caso uma das informações importantes de tendência central é a moda, como o valor com maior frequência nos dados. No caso de variáveis reais contínuas, frequências são desprovidas de significado, sendo impossível calcular seus valor nas amostras, apenas por classes.

Definição 4.2.2 — Moda. A moda M_0 de uma amostra é a observação com maior frequência nos dados

A moda nem sempre é um valor fixo, pois diferentes valores ou classes podem possuir mesma frequência. Quando um histograma apresenta dois picos, este também é chamado de **bimodal**. Quando apenas um é apresentado, chamamos o histograma de **unimodal**

Média ponderada

A média ponderada considera que cada valor pode possuir uma importância diferenciada, e a ele é associado um valor chamado **peso**. A equação (4.3) demonstra o

valor de uma média ponderada

$$\bar{x} = \frac{\sum_{i=0}^n p_i x_i}{\sum_{i=0}^n p_i} \quad (4.3)$$

Em que p_i corresponde o peso de cada um dos valores para as n variáveis possíveis. A relação de cada peso pela soma total destes pesos também é chamado de ponderador e pode ser representado pela equação (4.4)

$$\lambda_i = \frac{p_i}{\sum_{i=1}^n p_i} \forall i \quad (4.4)$$

A média ponderada pode ser reescrita em termos de seus ponderadores de acordo com a equação (4.5)

$$\bar{x} = \sum_{i=0}^n \lambda_i x_i \quad (4.5)$$

Mediana

A mediana é uma representação do valor associado a aproximadamente 50% da frequência total dos dados.

Definição 4.2.3 — Mediana. Se o número de elementos (n) for ímpar, a mediana é igual a $\frac{n+1}{2}$ elemento. Se o número de elementos for par, então a mediana é igual a média do $\frac{n}{2}$ elemento e o $\frac{n}{2} + 1$ elemento

Proposição 4.2.3 Suponha que a amostra consiste em 10 observações: 6, 3, 4, 7, 4, 6, 7, 6, 5, 3, nós teremos um número de elementos $n = 10$, sendo este valor par. Ordenando o conjunto de dados teremos 3, 3, 4, 4, 5, 6, 6, 6, 7, 7. Então a mediana é igual a média entre o 5º e o 6º elemento, correspondendo ao valor de $(5 + 6)/2 = 5,5$.

4.2.2 Medidas de posição

As medidas de posição são aquelas tomadas em relação a outras, ou seja em seu contexto geral com outros valores. Entre elas as mais comuns são os **percentis**, **quartis**, e **decis**

Percentis ou quantil

Uma das formas de se avaliar a posição dos dados é quanto a sua frequência. Um percentil ou quantil representa o valor correspondente a uma proporção total dos

dados.

Definição 4.2.4 — Percentil ou quantil. Um percentil ou quantil c_p de uma amostra corresponde ao valor imediatamente superior ou igual a $100xp\%$ e imediatamente inferior a $100x(1 - p\%)$ dos dados

Proposição 4.2.4 Suponha que a amostra a seguinte amostra: $\{6, 3, 4, 7, 4, 6, 7, 6, 5, 3, 4, 2\}$ nós teremos um número de elementos $n = 10$. Ordenando o conjunto de dados teremos $\{2, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7\}$. logo as proporções dos dados serão $\{2 : 8\%, 3 : 17\%, 4 : 25\%, 5 : 8\%, 6 : 25\%, 7 : 17\%\}$. As proporções acumuladas serão equivalentes a $\{2 : 8\%, 3 : 25\%, 4 : 50\%, 5 : 58\%, 6 : 83\%, 7 : 100\%\}$. Então o percentil de 67% será o valor imediatamente superior a 58% e inferior a 83%. Utilizando uma interpolação linear temos que $(67\% - 58\%) * (6 - 5) / (83\% - 58\%) + 5 = 5.36$.

Quartis

O **quartil** são medidas de posição que correspondem a 4 posicionamentos especiais dentro do conjunto de dados. O primeiro quartil representa o **o percentil de 25%**, o segundo quartil representa **o percentil de 50% ou a mediana**, e o terceiro quartil representa **o percentil de 75% dos dados**.

Proposição 4.2.5 . Se obtivermos um conjunto de dados iguais a 50, 34, 27, 54, 25, 43, 15, 12 contendo 8 valores então podemos ordená-los em crescente de tal forma que teremos 12, 15, 25, 27, 34, 43, 50, 54. O valor do primeiro quartil será, segundo os dados ordenados, 15. O terceiro quartil será 43. E a mediana será igual a 27.

4.2.3 Medidas de dispersão

Outras medidas importantes são as de dispersão. Entre as mais comuns podemos citar a **variância**, o **desvio padrão** e a **amplitude** dos dados.

Amplitude

A forma mais simples de se medir a dispersão dos dados é considerar sua amplitude. A maior vantagem em se definir a amplitude é sua simplicidade de cálculo, porém esta estatística é muito afetada por valores extremos

Definição 4.2.5 — Amplitude. Corresponde a diferença do valor máximo obtido nos dados x_{max} com o valor mínimo x_{min} .

Intervalo Interquartil

Uma forma de se avaliar uma medida de dispersão menos afetada pelos valores extremos é o intervalo interquartil

Definição 4.2.6 — Intervalo interquartil. Corresponde a diferença do valor do terceiro quartil (Q_{75}) com o valor do primeiro quartil (Q_{25}).

Variância

A variância pode ser descrita pela equação (4.7)

$$s^2 = \frac{\sum_{i=0}^n (x_i - \bar{x})^2}{n - 1} \quad (4.6)$$

Em que $n - 1$ é o número de graus de liberdade da amostra, tal que este pode ser definido pelo número de amostras menos o número de estatísticas utilizadas durante o cálculo. Note que para a operação da variância precisamos antes determinar o valor da média. É uma medida que não apresenta as mesmas unidades que a das amostras, para isso geralmente utilizamos o desvio padrão, que pode ser calculado como a raiz quadrada dos valores da variância ($s = \sqrt{s^2}$).

Em alguns casos também é possível calcular a variância para classes e não para valores, assim como a média aritmética. Neste caso podemos calcular a variância a partir de

$$s^2 = \frac{\sum_{i=0}^n f_i (c_i - \bar{c})^2}{n - 1} \quad (4.7)$$

Em que c corresponde ao valor da classe e f_i o valor da frequência associada aquela classe.

4.2.4 Assimetria

Outra medida pontual importante também é a assimetria. Esta se caracteriza pela diferença de proporções de uma distribuição de amostras segundo ao redor de seu valor mais frequente.

A figura (4.5) demonstra a distribuição de dados assimétrica. O item a) representa uma distribuição assimétrica positiva, enquanto o item b) representa uma distribuição assimétrica negativa. A assimetria positiva é caracterizada por um valor da mediana abaixo do valor médio, enquanto a assimetria negativa se caracteriza por uma alta proporção de valores altos.

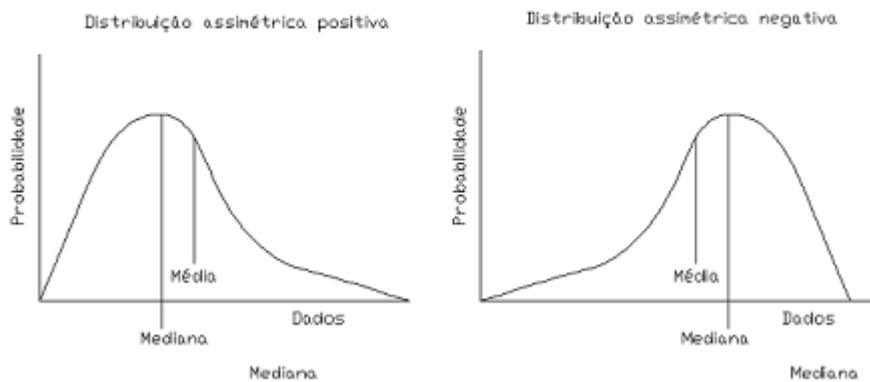


Figura 4.5: Assimetria de uma distribuição de dados a) Assimetria positiva b) assimetria negativa

Uma das medidas de assimetria mais comuns é o coeficiente de Pearson que pode ser expresso pela equação (4.8)

Coefficiente de assimetria de Pearson

$$S_p = 3(\bar{x} - M_e) / s \quad (4.8)$$

Em que M_e é a moda dos dados, \bar{x} é o valor médio das amostras e s é o desvio padrão das amostras.

Proposição 4.2.6 Imagine uma variável com valor de média $\bar{x} = 198.89$, valor de mediana $M_e = 128.15$ e valor de desvio padrão igual a $s = 180.56$. O valor do coeficiente de assimetria será igual a $3(192.89 - 128.30)/180.56 = 1.07$, demonstrando assimetria nos dados.

Distribuições com característica de assimetria positiva são muito comuns na avaliação de depósitos minerais, principalmente no tratamento de commodities erráticos tal como ouro e diamante. Nesses depósitos podem ocorrer anomalias raras e uma amostra constituir em alto valor. Esta propriedade também é chamada de efeito pepita e será melhor tratada no capítulo de Continuidade espacial.

4.2.5 Coeficiente de variação

Em certos momentos é importante comparar variáveis aleatórias de tipos diferentes. Para sabermos se uma distribuição é mais errática que outra, neste caso, não bastariamos comparar seus valores de variância. Valores que possuam médias maiores tendem a apresentar dispersões também maiores. Para isso utilizamos o coeficiente de variação, que nada mais é do que o desvio padrão de uma distribuição pelo seu valor médio. Desta forma "igualamos" diferentes distribuições em um único coeficiente

comparativo.

O coeficiente de variação pode ser dado pela equação (4.9)

$$CV = \frac{s}{\bar{x}} \quad (4.9)$$

Os coeficientes de variação são medidas importantes para a pesquisa mineral, porque são a primeira forma utilizada para classificar depósitos minerais segundo sua regularidade. O livro de [Maranhao, 1985] demonstra a classificação de depósitos minerais de acordo com o coeficiente de variação, tal como na tabela 4.1.

Tabela 4.1: Regularidade dos depósitos minerais de acordo com a classificação do coeficiente de variação

Regularidade	Coeficiente de variação	Exemplo
Regulares	$5\% < CV < 40\%$	Jazidas de ferro, manganês, níquel, cobalto
Irregulares	$40\% < CV < 100\%$	Jazidas de fluorita, barita, grafita, corídon
Muito irregulares	$100\% < CV < 150\%$	Jazidas de tungstênio em tactitos, ouro
Extremamente irregulares	$CV > 150\%$	Pegmatitos com berilo, tantalita, columbita

Valores de coeficiente de variação maiores representam geralmente um maior desafio para a aplicação de técnicas de geoestatística, pois geralmente apresentam alta variabilidade ou erraticidade dos dados.

4.2.6 Conjugando estatísticas pontuais

Como dito anteriormente, é sempre importante conjugar estatísticas pontuais diferentes de forma a garantir a melhor informação possível. Uma destas alternativas é adicionar ao valor médio um número de desvios padrões de forma a garantir que um conjunto de dados esteja situado dentro destes limites $(\bar{x}) \pm ks$. Para isso utilizaremos uma das mais renomadas relações estatísticas.

A desigualdade de Chebyshev é uma identidade que implica em um valor mínimo de probabilidade para que uma realização esteja dentro de um intervalo múltiplo do desvio padrão. Podemos definir a equação (4.10) como a desigualdade de Chebyshev.

$$P(|X - \mu| \geq k\sigma) \leq 1/k^2 \quad (4.10)$$

Em que X é o valor da variável aleatória, μ é o valor da média da população, σ é o valor do desvio padrão da população e k é uma constante proporcional. A desigualdade de Chebyshev é independente do valor da distribuição de probabilidades para

a variável aleatória. Apesar de não possuirmos os valores (μ, σ) correspondentes aos parâmetros da população, podemos estimar os valores a partir das estatísticas das amostras. Se o número de amostras for grande o suficiente e as técnicas de amostragem bem selecionadas, podemos dizer que $(\mu \sim \bar{x}, \sigma \sim s)$

Para um k igual a 2, sabemos que existe uma probabilidade de no mínimo 75 por cento de que o valor da amostra esteja em dois desvios padrões da média. Podemos caracterizar as amostras então por uma medida de posição e de dispersão conjuntamente. Ao descrever as amostras é bem claro que devemos associar no mínimo dois de seus parâmetros, como por exemplo, dizer que as amostras de teor de ouro possuem valores entre $(50 \pm 20) \text{ ppm}$ em que 20 representaria dois desvios padrões de 10 ppm e 50 ppm seu valor médio.

4.3 Validação do banco de dados e valores outliers

A primeira etapa da geoestatística é a validação das amostras. Devemos antes de tudo verificá-las para que não encontremos valores discrepantes (outliers) ou incoerências nos dados. Análises realizadas com valores muito discrepantes pode acabar gerando resultados espúrios e inconsistentes com a realidade.

Definição 4.3.1 — Outlier. *Um outlier é considerado um valor ou observação caracterizado pela sua relação entre o restante de observações que fazem parte das amostras. O seu distanciamento em relação as observações é essencial para fazer sua caracterização. Estas observações também são chamadas de 'anormais', contaminantes, estranhas, extremas ou aberrantes - Figueira [1998]*

É importante entender que os dados anômalos nem sempre são valores errados. Eles podem ser valores reais representantes de uma anomalia da natureza. Poderíamos encontrar, por exemplo, em um depósito de ouro uma pepita com um valor agregado muito alto, mas apesar de ser um dado correto ele não representa o conjunto de amostras como um todo. Machado [2012] indica que o surgimento de valores anômalos podem ocorrer por diversas formas, entre elas:

1. **Valores errôneos:** As possíveis causas são os erros de análise ou de digitação, troca de amostras, contaminações de amostras ou até mesmo fraude.
2. **Valores pertencentes a outra população:** Podem ocorrer devido à mistura de diferentes teores ou litologias, ou que possuem processo formacional em tempos geológicos distintos. A revisão dos domínios geológicos, neste caso é recomendado, de forma a tratar e estimar os dados separadamente.

3. Valores pertencentes a mesma população: Podem ocorrer eventos metagenéticos que favoreçam a concentração de uma propriedade em parte do depósito. Estes eventos estão relacionados também ao chamado **efeito pepita**, em que proporções erráticas podem aparecer apenas em locais distintos do depósito, em regiões pequenas.

A Tabela 4.2 é um exemplo de como valores anômalos podem aparecer. Nota-se claramente que as amostras 1 e 3 estão erradas. Primeiramente porque não existem valores de teor percentuais acima de 100% e também porque não existem teores descritos como letras. No entanto, a amostra 4 também está errada, porque o minério composto por limonita não pode apresentar um valor de teor de ferro de 72%, pois é incompatível com a química da mineralogia.

Tabela 4.2: Tabela de teores do minério de ferro

Índice	Minério	Teor(%)
1	Hematita compacta	120%
2	Hematita granular	53%
3	Magnetito	0.i3
4	Limonita	72%

Proposição 4.3.1 *Pode até mesmo parecer um clichê, mas a melhor forma de se analisar outliers é com bom senso. Devemos entender o problema, analisá-lo profundamente na hora de limparmos o banco de dados. Permitir que bancos de dados sejam transmitidos antes de uma boa verificação pode resultar no fracasso de uma análise destes dados.*

Diversas são as formas de identificação de valores outliers. Técnicas para valores em apenas uma variável são muito conhecidas, no entanto, deve-se entender que um valor anômalo depende de sua dimensão analisada. Uma amostra outlier considerando variáveis distintas pode não ser um valor um valor anômalo quando considerado um problema multivariado.

Uma das ferramentas mais comuns para identificação de valores anômalos é o gráfico boxplot. Ele demonstra a disposição dos dados em um eixo e limita os valores das amostras em uma caixa contendo os quartis das amostras. Os valores que se situam acima ou abaixo das retas formadas pela adição e subtração 1,5 vezes o intervalo interquartil dos valores máximo e míno dos dados representam outliers. O intervalo interquartil é também determinado como a diferença entre os valores do terceiro quartil e do primeiro quartil. A figura 4.6 demonstra o gráfico boxplot e suas dimensões.

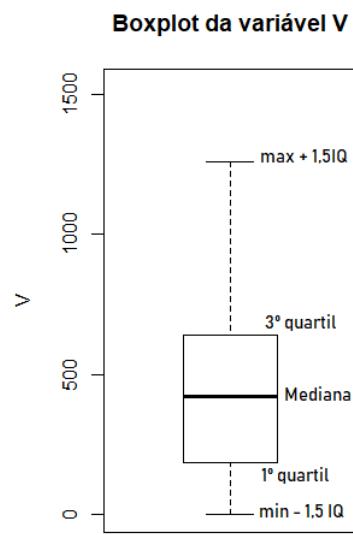


Figura 4.6: Representação de um gráfico de caixa dividida entre os intervalos das amostras

Os valores anômalos ou outliers são demonstrados na figura 4.8 como pontos circulados fora das barras que representam os limites de aceitação dos valores da amostra.

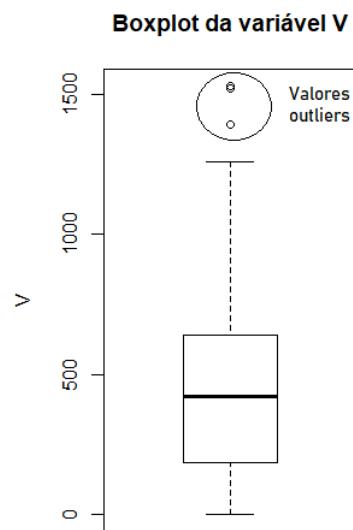


Figura 4.7: Representação dos valores outliers no gráfico boxplot - Pontos circulados em vermelho

Muito cuidado deve ser utilizado com esta ferramenta. Em alguns casos distribuições de dados assimétricas podem gerar no gráfico boxplot uma quantidade de valores anômalos absurdas. A melhor forma de lidar com valores outliers é o bom senso, ferramentas são úteis, mas não devem ser o critério determinante na maioria dos casos.

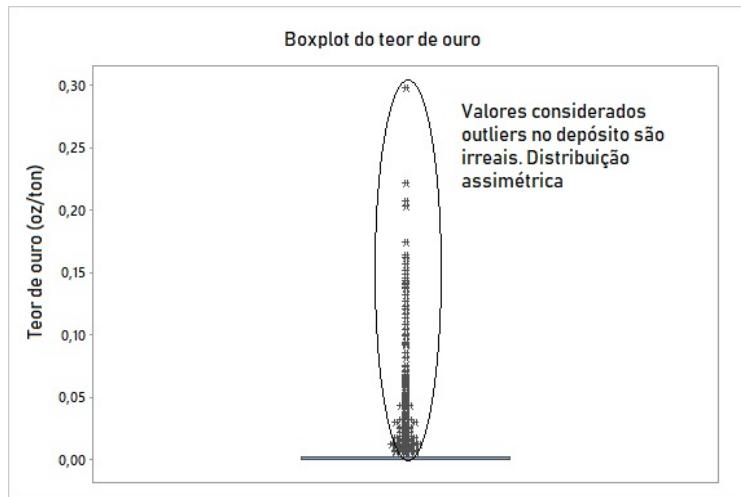


Figura 4.8: Valores outliers em uma distribuição assimétrica dos dados. Nota-se que grande parte da informação é considerada outliers. Neste caso é necessário bom senso para não se remover informações desnecessárias e prejudicar a análise de dados.

Após a identificação de valores anômalos é possível realizar o tratamento destes dados. É imprescindível entender que bancos de dados **nunca** devem ser alterados, apenas estatísticas. A alteração ou remoção de dados é considerada uma atitude imoral para analistas de dados.

- R** É importante entender que um banco de dados **nunca** deve ser alterado. Apenas as estatísticas são cabíveis a manutenção. A alteração de dados reais pode ser considerada um ato imoral, principalmente na mineração, onde o trabalho, segurança e condições de vidas de muitas pessoas estão em jogo.

Diversas alternativas podem ser utilizadas para o tratamento de valores outliers. Dentre elas podemos citar

1. **Truncamento:** Após identificar valores outliers é possível normalizar seus valores para os valores extremos (máximo ou mínimo), ao desconsiderá-los. O truncamento de dados na geoestatística deve ser feito de forma a evitar que os valores anômalos não alterem significativamente as estatísticas globais. Como regra de ouro considera-se que o truncamento deve ser feito sem que se altere mais do que 10% dos valor médio das amostras.

2. **Remoção:** Em alguns casos a remoção dos valores outliers pode ser feita. Se a proporção de dados removidos for alta, é possível alterar excessivamente as estatísticas, por isso muito cuidado deve ser feito ao considerar uma amostra como outlier.
3. **Reescalonamento:** Dependendo da distância relativa dos outliers com o contexto geral das amostras é possível realizar uma redução de suas distâncias até o valor máximo desconsiderando-os.

4.4 Descrição espacial das amostras

A geoestatística é uma ciência que prevê a utilização de informações no espaço, e para isso muitas vezes utilizamos informações de mapas. Mapas são representações visuais de uma região que são dotados de informações como **escala, legenda, título, orientação**.

Mapas de localização destas amostras são uma ferramenta gráfica muito importante para determinar o comportamento de variáveis no espaço. Mapas devem ser feitos de forma cuidadosa, representando escalas condizentes com o objeto de estudo e garantindo a melhor visualização possível das amostras.



A qualidade desejada de um mapa varia de acordo com a investigação. Tipicamente a representação de um mapa deve ser limpa, sem valores altos ou baixos ambíguos, e mostrar os dados o menos distorcidos possível com um mínimo de artefatos computacionais - Gustavsson et al. [1997]

Estas informações nos permitem identificar regiões consideradas mais ricas, regiões onde ocorrem agrupamentos característicos dos dados, e o layout das malhas de amostragem. A figura 4.9 demonstra um depósito polimetálico de Jura. O atributo é o tipo de rocha de um dado período geológico.

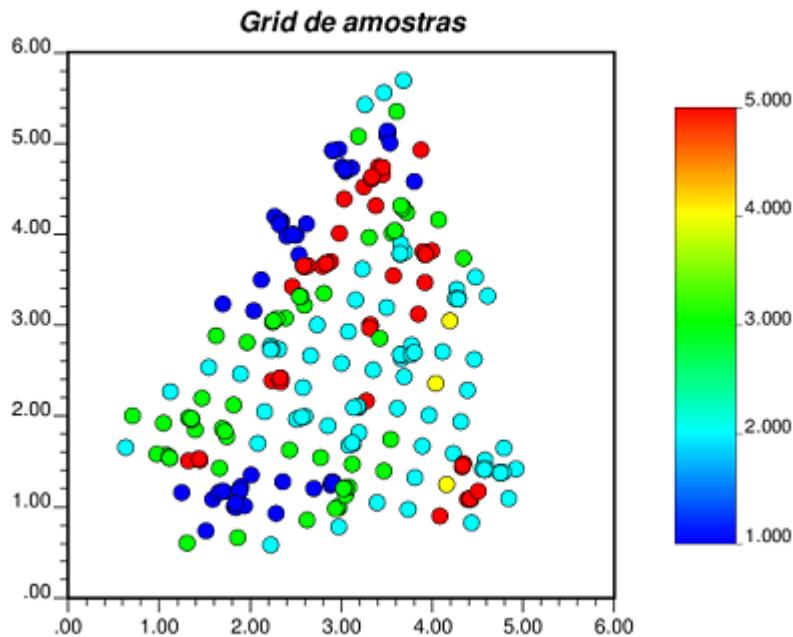


Figura 4.9: Disposição das amostras no espaço. Cores diferenciadas mostrando tipos de rocha em períodos geológicos diferentes

Podemos ver que as amostras estão dispostas de forma irregular em um formato de delta de um rio. A orientação do tipo de rocha 1 se encontra ao oeste e parte ao sul, enquanto a do tipo 5 se encontra distribuído mais ao norte. Qualquer estimativa realizada a partir desta configuração de amostras deve respeitar os valores iniciais. Se por exemplo, iniciássemos uma exploração cujo o interesse seria o litotipo 1, provavelmente começariámos a retirar o material de oeste para leste para reduzir o fluxo de caixa do empreendimento.

A 4.10 demonstra a propriedade de teor de Cádmio obtida nestas amostras no depósito. Podemos verificar sua distribuição segundo esta disposição deltaica.

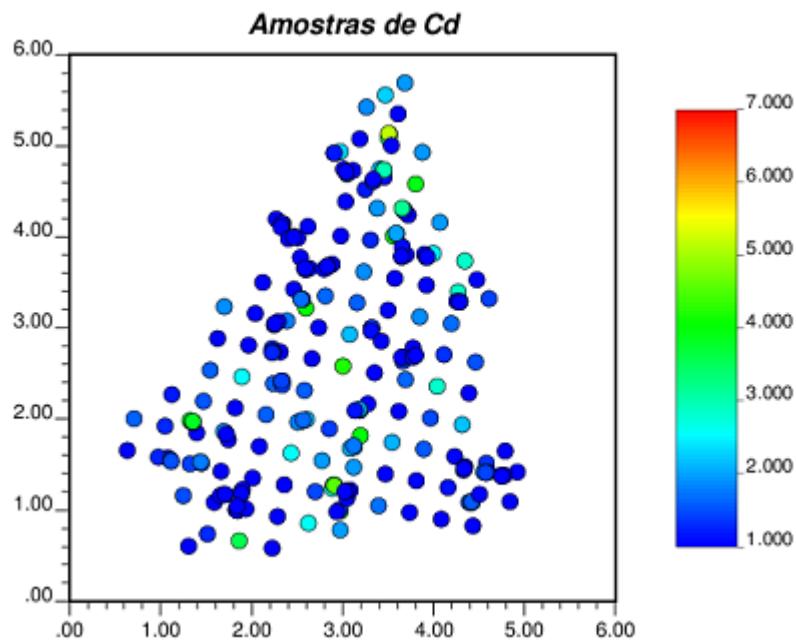


Figura 4.10: Disposição do Cd

As informações disponíveis nestes mapas nos permite associar as informações entre as variáveis do tipo litológico e o teor de Cádmio. Notamos que o litotipo 1 parece ter maior correlação com valores baixos do teor de Cádmio do que o litotipo 2, que parece ter correlação com valores um pouco mais altos. Esta análise visual nos permite entender o comportamento de certas variáveis e sua disposição no espaço, buscando explicações para os valores destas propriedades. Além das informações obtidas em mapa também podemos visualizar amostras e propriedades em um espaço tridimensional. A figura 4.11 demonstra a disposição de amostras e a visualização do comportamento de uma propriedade binária no espaço.

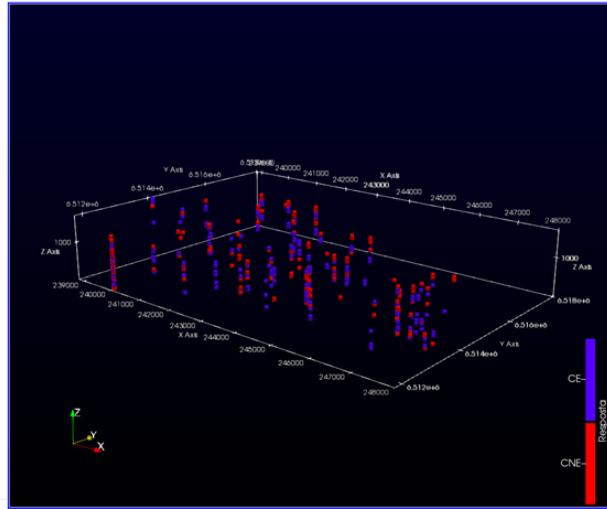


Figura 4.11: Informações de amostras obtidas em três dimensões.

4.5 Histograma

A descrição das estatísticas das amostras é uma forma inicial para aglomerar um conjunto de informações extensos. Um gráfico de grande utilidade para verificar a frequência dos dados é o histograma.

Definição 4.5.1 — histograma. *Um histograma é uma ferramenta gráfica, representada por um gráfico de barras que condiciona os valores de uma variável com suas frequências.*

Esta ferramenta é essencial principalmente em três condições:

1. **Classificação:** Quando possuímos classes distintas o histograma apresenta diretamente a proporção de cada classe considerada
2. **Contagem:** Quando a variável constitui em valores inteiros, cada valor desta variável pode ser diretamente associada a sua frequência.
3. **Contínuo:** Quando os valores são reais, podemos atribuir intervalos de classe (ou em inglês *bins*) aos quais estes valores estão inseridos. Dependendo do número de intervalos de classe e seu tamanho o histograma pode apresentar diferentes formas.

Uma das proposições utilizadas para o cálculo do número ótimo de intervalos de classes é pela fórmula de Sturges 4.11

$$\hat{h} = \frac{\text{amplitude dos dados}}{1 + \log(n)} \quad (4.11)$$

Em que a amplitude dos dados é relacionado a diferença do máximo e do mínimo das amostras e n é o número de amostras. A figura 4.12 representa um histograma da variável Cádmio do depósito de Jura. Podemos notar como a distribuição dos dados se comporta nesta variável, como aspectos de simetria, valores médios, e inclusive possíveis valores outliers, quando as barras de frequência são pequenas e distanciadas da maioria.

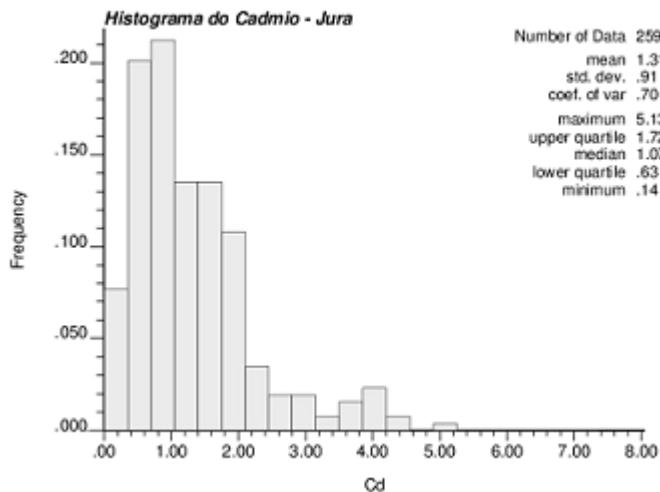


Figura 4.12: Histograma do Cd

A observação de uma frequência de uma classe é diretamente relacionada ao tamanho desta. Na figura 4.13 podemos ver que a classe de teores de 0,04 a 0,75 g/ton ocupa uma proporção de 20 % dos dados.

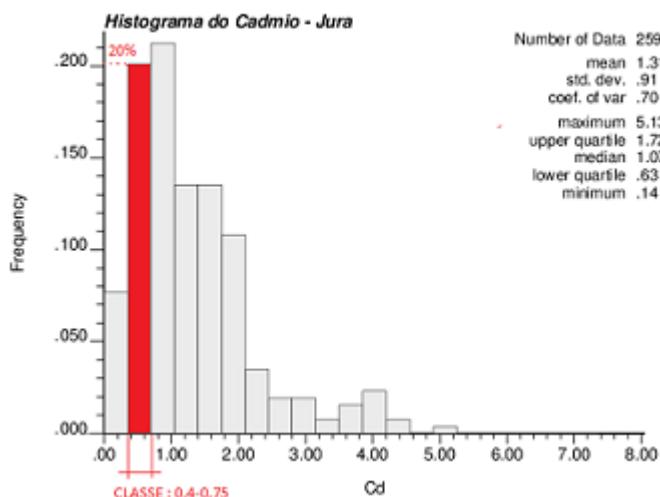


Figura 4.13: Histograma do Cd - Classe marcada

A construção de histogramas envolve sempre a criação de intervalos de classe de mesmo tamanho. Alterar apenas o tamanho de uma classe em detrimento de outras pode ser considerado um uso abusivo das estatísticas, enviesando a percepção de outras pessoas sobre as verdadeiras frequências dos dados.

R

Não é correto alterar o tamanho de apenas um intervalo de classe em detrimento dos outros. Esta prática é mal vista, e pode ser intuitivamente criada para gerar vies na percepção dos leitores quanto as frequências de determinados valores.

Assim como em outras estatísticas, a utilização dos histogramas favorece o entendimento global dos dados, mas prejudica no entendimento apurado da variável. A escolha do tamanho do intervalo é uma variável importante para a observação desta estatística gráfica. Valores de classe com tamanho muito grande apresentaram frequências maiores, mas perderão a forma natural dos dados. Valores de classe com tamanho muito pequeno apresentarão baixa frequência e se tornarão mais achados, dificultando a visualização das proporções da variável.

Proposição 4.5.1 *A escolha do tamanho do intervalo de classe é fundamental para verificar a forma do histograma e sua representação real. Valores de tamanho muito pequenos ou grandes podem gerar gráficos pouco intuitivos, escondendo a real simetria, valores médios e dispersão dos dados. É importante que um histograma caracterize visualmente os dados de forma a representar as estatísticas numéricas a serem calculadas.*

Outra forma de representar um histograma é na sua forma acumulada. Neste caso cada valor das frequências de uma variável são aumentadas em ordem crescente, do menor valor das amostras até o maior valor. A figura 4.14 é uma demonstração do gráfico acumulado.

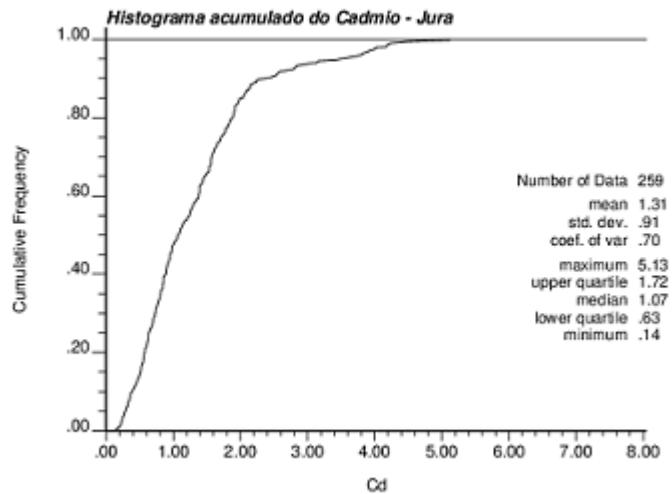


Figura 4.14: Histograma do Cd acumulado

A figura 4.15 demonstra a leitura do gráfico acumulado. Podemos notar por este gráfico que 60 por cento dos valores estão abaixo do teor de 1,5g/tonelada.

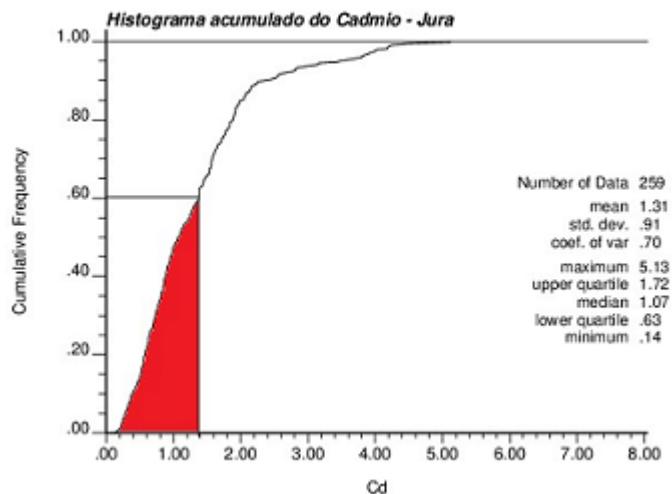


Figura 4.15: Histograma do Cd acumulado - Leitura

O formato dos histogramas pode adicionar importantes informações sobre a distribuição dos dados, como por exemplo a assimetria. Na figura 4.16 podemos observar dois histogramas de depósitos minerais diferentes, um simétrico de Ferro em A) e um de alumínio assimétrico em B). Quando consideramos as técnicas clássicas de avaliação de depósitos a assimetria dos dados pode dificultar os métodos convencionais, o que torna depósitos de alta assimetria mais difíceis de reproduzirem estimativas condizentes.

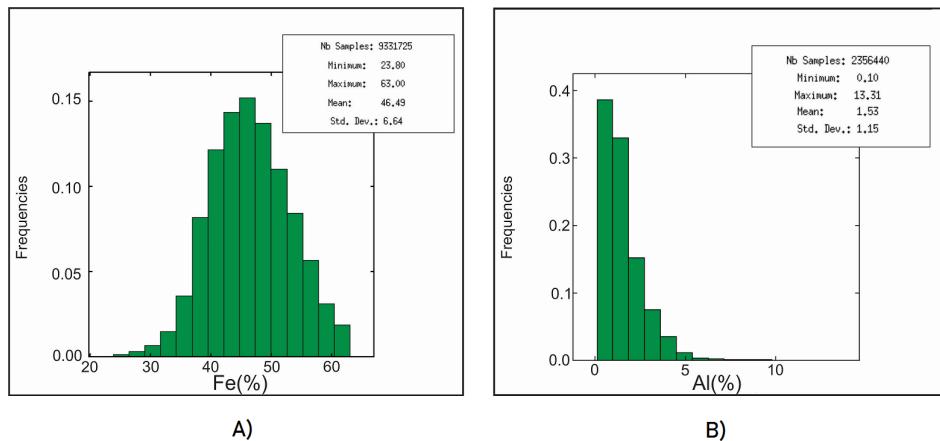


Figura 4.16: Simetria para diferentes histogramas - a) histograma simétrico, b) histograma assimétrico

O formato do histograma também é um importante parâmetro para a inferência de distribuições de probabilidade. A partir dele podemos visualizar uma possível distribuição de probabilidade e dar um "chute" para testarmos se esta se encaixa na distribuição das amostras. Distribuições de frequências centradas podem ter como candidato um modelo de ajuste gaussiano, por exemplo. Distribuições assimétricas podem se encaixar, por exemplo, em um modelo lognormal.

4.6 Inferência Estatística

Após analisados os dados amostrais podemos utilizar funções para modelar populações dos dados. Na maioria dos casos não precisamos conhecer *a priori* as distribuições da população, mas em alguns casos como na geoestatística não-linear, conhecer uma distribuição teórica de probabilidades pode facilitar estudos para entender problemas mais complexos

Definição 4.6.1 — Inferência estatística. *Inferência estatística é o método pelo qual deduzimos informações da população dos dados com base em informações das amostras*

,

4.6.1 Famílias de distribuições estatísticas

Uma função de densidade de probabilidade de uma variável aleatória nada mais é do que uma função $p(X = x)$ que correlaciona cada realização x da variável aleatória X a uma dada probabilidade. Como consequência da definição algumas condições estão associadas:

- $p(x) \leq 1 \forall x$
- $\int_{-\infty}^{\infty} p(x)dx = 1$ para distribuições contínuas
- $\sum_{x=-\infty}^{\infty} p(x) = 1$ em que a e b são limites para a distribuição discreta

Distribuição de Poisson

Esta é uma distribuição discreta amplamente utilizada para experimentos ditos de eventos "raros", ou seja, utilizada para modelar eventos que a probabilidade de ocorrência é diretamente proporcional ao tempo de espera.

Em filas de caminhões, por exemplo, é muito comum a utilização da função de distribuição de Poisson para medir a probabilidade de chegada de um equipamento, pois é de se esperar que para um pequeno intervalo de tempo após a saída de um caminhão da frente de lavra, a probabilidade da chegada de outro seja pequeno. Outro exemplo é a frequência de fraturas em uma rocha. É de se esperar que para tamanhos pequenos de rocha a quantidade de fraturas seja pequena, enquanto para tamanhos grandes de rocha essa densidade aumente.

A função de distribuição de Poisson pode ser escrita segundo a equação (4.12)

$$P(X = x) = \frac{\exp^{-\lambda} \delta^x}{x!} \quad (4.12)$$

Em que x é uma realização da variável aleatória X , $P(X = x)$ é a probabilidade associada àquele evento e $\lambda = E(X)$ sendo o parâmetro da função. Na maioria dos casos aproximamos $E(X) \sim \bar{x}$. Tal como qualquer distribuição de probabilidades sabemos que a soma de todos os eventos possíveis deve gerar um resultado igual a 1. Podemos demonstrar isso de acordo com a prova

Demonstração. Sabendo que a função exponencial pode ser aproximada por uma

série de Taylor como a seguir temos :

$$e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$$

Então:

$$\begin{aligned} \sum_{x=0}^{\infty} P(X = x) &= \sum_{x=0}^{\infty} \frac{\exp^{-\lambda} \lambda^x}{x!} \\ \sum_{x=0}^{\infty} P(X = x) &= \exp^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ \sum_{x=0}^{\infty} P(X = x) &= \exp^{-\lambda} \exp^{\lambda} = 1 \end{aligned}$$

■

Distribuição Gaussiana

Esta talvez seja uma das funções de densidade de probabilidade mais populares e representa um grande papel na geoestatística. As equações de estimativa lineares que serão apresentadas neste livro são também analogamente chamadas de **equações normais**. Isto se deve pelo fato de que os resultados obtidos em variáveis gaussianas são os mais precisos possíveis dentro de todas outras distribuições na geoestatística. Quanto mais próximo for a distribuição das amostras de uma distribuição gaussiana, melhores serão os resultados de uma estimativa geoestatística.

Proposição 4.6.1 Consideremos uma variável Z , gaussiana e estacionária (em prática a variável que pode ser aproximada de um histograma por uma gaussiana), com média m e variância σ^2_Z , a hipótese de permanência da normalidade indica que uma variável Y estimada segue uma distribuição de mesma forma e média $m = E(Z) = E(Y)$ e variância $\sigma^2_Z \neq \sigma^2_Y$ -Journel and Huijbregts [1978]

O formato de uma distribuição gaussiana é tipicamente na forma de um sino (*bell shape*), centrado em um valor médio e com uma variância característica. A figura 4.17 demonstra uma distribuição gaussiana típica com média igual a 5 e variância igual a 2.

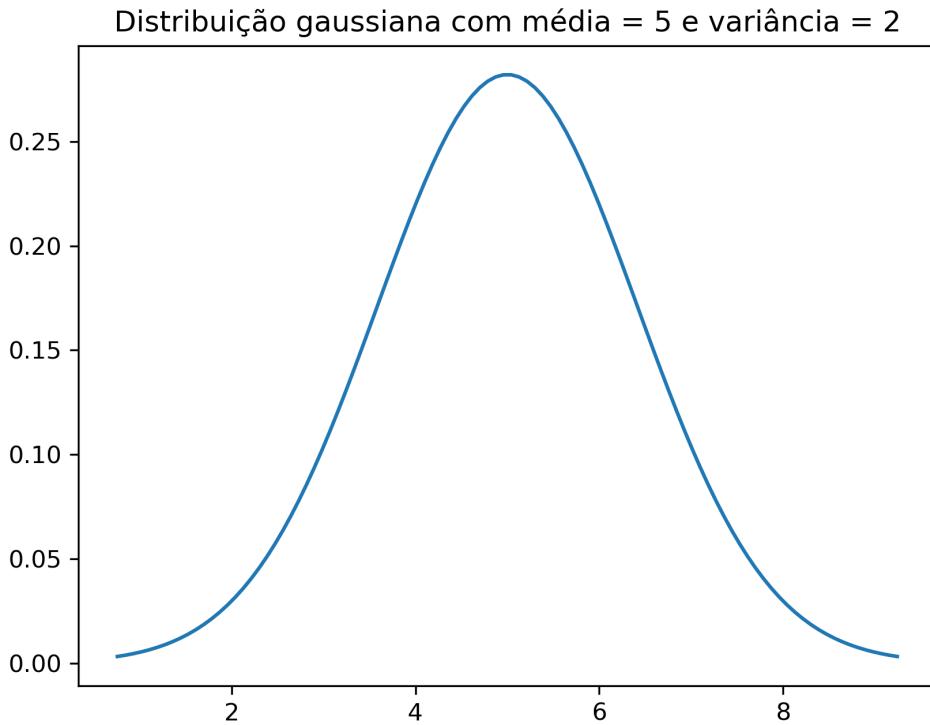


Figura 4.17: Forma de uma distribuição gaussiana com média 5 e variância 2

A distribuição é um modelo simétrico e descrito por dois parâmetros, a média da população e a variância. A função de densidade de probabilidade da distribuição pode ser desrita segundo a equação (4.13)

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4.13)$$

Em que σ^2 é a variância da distribuição aleatória e μ é a média. O caso particular da distribuição gaussiana é quando sua média é igual a zero e variância é igual a 1, neste caso temos uma distribuição padronizada segundo a equação (4.14)

$$P(X = x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} \quad (4.14)$$

Uma variável aleatória pode ser padronizada segundo a relação (4.20)

$$X_p = (X - \mu)/\sigma \quad (4.15)$$

Que nada mais é do que uma operação de deslocamento da variável aleatória pela sua média e encurtamento da distribuição pelo seu desvio padrão.

Para demonstrar que a distribuição gaussiana possui soma de todos os seus eventos igual a 1 devemos antes lembrar que ela é uma distribuição simétrica, logo a soma dos valores à esquerda do valor médio da distribuição é idêntico à soma dos valores à direita da distribuição. A integral da função gaussiana não possui uma antiderivada para utilizarmos explicitamente, por isso o truque utilizado é provar que o quadrado da integral da gaussiana é equivalente a 2π . Logo temos:

Demonstração. Prova do somatório de uma função gaussiana ser igual a 1

$$\begin{aligned} Int^2 &= \left(\int_{-\infty}^{\infty} e^{-\frac{(x)^2}{2}} dx \right)^2 = 4 \int_0^{\infty} e^{-\frac{(t)^2}{2}} dt \int_0^{\infty} e^{-\frac{(u)^2}{2}} du \\ &4 \int_0^{\infty} \int_0^{\infty} e^{-\frac{(t^2+u^2)}{2}} dt du \end{aligned}$$

Alterando para coordenadas polares

$$\begin{aligned} &4 \int_0^{\infty} \int_0^{\pi/2} r e^{-\frac{r^2}{2}} dr d\theta \\ &2\pi \int_0^{\infty} r e^{-\frac{r^2}{2}} dr \\ &2\pi \end{aligned}$$

Logo se: $Int^2 = 2\pi$

$$Int = \sqrt{2\pi}$$

portanto :

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1$$

■

Distribuição Lognormal

A distribuição lognormal é uma distribuição assimétrica e positiva, geralmente associada na mineração com depósitos de elementos raros , tais como ouro, diamante e platina. Pode ser considerada uma distribuição cujo seu logaritmo é normalmente distribuído. A equação (4.16) demonstra a função de densidade de probabilidade para a distribuição lognormal.

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp^{\frac{(-\log(x))^2}{2}} \quad (4.16)$$

O Valor esperado da distribuição pode ser demonstrado segundo a equação (4.17)

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \quad (4.17)$$

A figura 4.18 apresenta a forma assimétrica da distribuição lognormal, para um distribuição com média 5 e variância igual a 2.

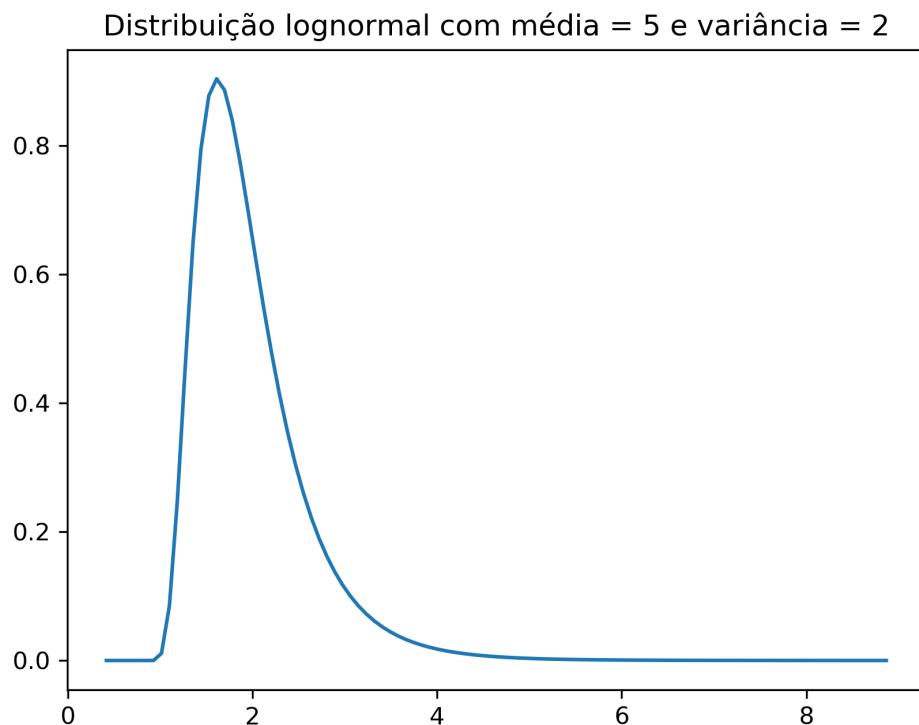


Figura 4.18: Forma de uma distribuição lognormal com média 5 e variância 2

Estimando a média da população

O processo de inferência estatística resume-se em determinar características da população a partir de dados amostrais. Podemos estimar o valor real da média da função aleatória $Z(x)$ a partir do estimador $\hat{Z}(x)$ a partir da média aritmética $\sum_{i=1}^n Z(x_i)/n$ em que n constitui um número grande de variáveis aleatórias em diferentes suportes i . A equação (4.18) apresenta este processo.

$$E(\hat{Z}(x)) = E\left(\sum_{i=1}^n Z(x_i)\right)/n = \left(\sum_{i=1}^n E(Z(x_i))\right)/n = \left(\sum_{i=1}^n m\right)/n = m \quad (4.18)$$

Sobre a hipótese de estacionaridade da média, sabemos que a média das variáveis aleatórias é igual a média da função aleatória. Ou seja, sob a hipótese de estacionaridade de segunda ordem podemos considerar que a média das amostras é um bom estimador para a média da população ou do depósito mineral.

Enquanto a variância no entanto temos segundo a equação (4.19)

$$Var(m) = Var \left(\sum_{i=1}^n Z(x_i)/n \right) = \sum_{i=1}^n 1/n^2 Var(Z(x_i)) = \sigma^2/n \quad (4.19)$$

Em outros termos, sob a hipótese de estacionaridade, a variância da média populacional tende a reduzir de acordo com o número de amostras tomadas. Isso também é chamado de efeito de suporte, pois quanto mais informações temos com a amostragem, mais o valor esperado de uma função aleatória tende a ser o correto. Quanto maior a quantidade de amostras utilizadas em uma estimativa, menores serão os erros associados a esta estimativa média local. A figura (4.19) demonstra como o valor médio tende a cada vez se aproximar mais da média das amostras com o aumento do número de amostras.

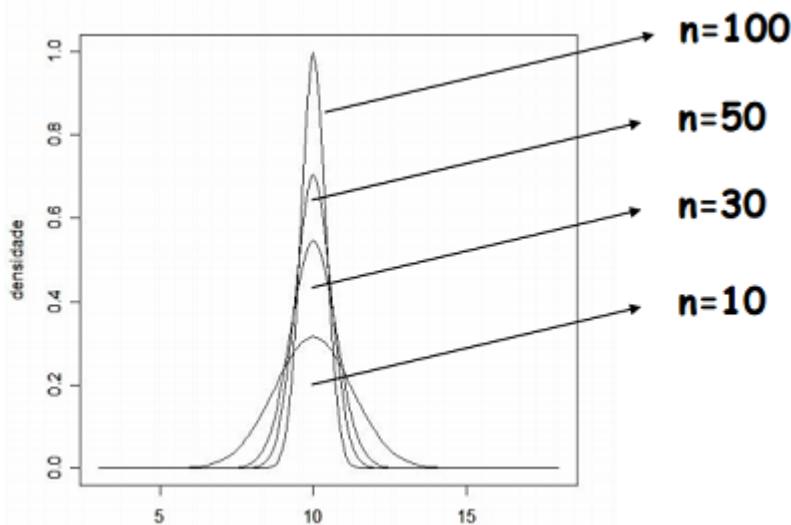


Figura 4.19: Figura demonstrando o efeito de suporte para um número crescente de amostras. O aumento do número de amostras tende a concentrar a função de densidade de probabilidade entorno do valor médio

4.7 Distribuição t-Student

Para determinarmos a distribuição gaussiana geralmente assumimos o conhecimento a respeito da variância da população. Se considerarmos a **distribuição de valores**

médios, sabemos que se $Z_{x1}, Z_{x2}, \dots, Z_{xn}$ são amostras normalmente distribuídas normalmente $\phi(m, \sigma^2)$, então a quantidade

$$Z_p = \frac{(\bar{Z(x)} - \mu)}{\sigma/\sqrt{n}} \quad (4.20)$$

É distribuída com variável aleatória $\phi(0, 1)$. A distribuição dos valores médios $(\bar{Z(x)} - \mu)/(\sigma/\sqrt{n})$ segue a distribuição chamada de t-Student, com $n - 1$ graus de liberdade. Quando o número de amostras tende a crescer, aproximadamente de 30, a distribuição t-Student converge para a distribuição normal padrão $\phi(0, 1)$. Por isso dizemos que para estudos estatísticos iniciais, precisamos de pelo menos 30 amostras para se ter uma melhor compreensão da média.

4.8 Dimensionamento de malhas regulares

Em campanhas de prospecção preliminares é rotineiro utilizar técnicas estatísticas convencionais para estimar o tamanho e posicionamento de malhas de amostragem. No estágio inicial é necessário cobrir uma certa área de forma a verificar suas potencialidades. A medida que os estudos avançam, as amostragens tendem a aumentar e se tornarem mais densas, e estudos geoestatísticos mais avançados são realizados. A área de influencia de uma perfuração pode ser calculada pela equação (4.22)

$$A_0 = \frac{A}{n} \quad (4.21)$$

Os estudos iniciais são fortemente afetados pela regularidade do depósito mineral. Depósitos erráticos como veios de ouro tendem a necessitar de malhas mais adensadas que depósitos regulares como os de carvão mineral.

R O principal fator que controla a densidade da malha de perfuração é a regularidade do depósito e, por isso, a malha tem de ser cada vez mais densa, à medida que trabalham depósitos onde a variabilidade na forma ou qualidade (teor e conteúdo) é maior - Maranhao [1985]

Para encontrarmos o número mínimo de amostras segundo o erro esperado para amostragem, utilizamos a equação

$$N = \frac{(t.CV)^2}{E^2} \quad (4.22)$$

Em que t é o valor da variável t-Student para um nível de confiabilidade, CV é o valor do coeficiente de variação do depósito mineral e E é o valor do erro aceitável para a estimativa.

Proposição 4.8.1 *Considere um depósito mineral com coeficiente de variação igual a 51,98%, um valor de confiabilidade para a média de 95% (t -student = 2.20, para 12 amostras), e um erro aceitável para uma medida de no máximo 20%. A área pesquisada é igual a $70.000m^2$, e realizaremos 12 amostras. Logo o erro que cometemos é $E = \sqrt{\frac{(t.CV)^2}{N}} = \sqrt{\frac{(2.20 \cdot 51.98)^2}{12}} = 32.7\%$*

4.9 Exercícios

Exercícios 4.1 Considere o conjunto de amostras com teores de ferro contendo unicamente hematita Fe_2O_3 e sílica SiO_3 . Os valores são (45, 69, 80, 35, 56, 78) %. Determine os valores outliers do problema considerando a massa atômica do ferro igual 56g/mol e do oxigênio igual a 16g/mol. Resp.: 80% e 78% ■

Exercícios 4.2 Considere o conjunto de amostras com teores (2.4, 5.0, 7.6, 4.3, 2.7, 8.9) g/ton todos com o mesmo suporte. Encontre o valor da média, da variância, do desvio padrão do conjunto de amostras. Resp.: $\bar{x} = 5.7$, $s^2 = 5.06$, $s = 2.25$ ■

Exercícios 4.3 Um geólogo precisa decidir entre duas metodologias de amostragem para um dado elemento de pesquisa. Entre elas temos a sonda diamantada e o pó de perfuratriz. As incertezas do custo da pesquisa estão diretamente relacionadas com a variabilidade da recuperação, desejando o método com o menor risco associado. Para isso mediu-se a recuperação dos testemunhos e do pó retirado pela máquina. A recuperação dos testemunhos fora de 90% com um desvio padrão de 30%, enquanto a do pó foi de 70% com uma variação de 20%. Deseja-se saber qual método utilizar. Resp.: Pó de perfuratriz « CV ■



5. Estatística bivariada

Como em outras artes, a ciência da dedução e análise é uma que não pode ser adquirida por um longo e paciente estudo, nem é a vida longa o suficiente para permitir qualquer mortal se ater a mais alta perfeição nela.

Sherlock Holmes em 'Um estudo em Vermelho'

5.1 Introdução

Na análise de bancos de dados geralmente se torna necessário comparar duas populações diferentes. Em um depósito mineral, por exemplo, podemos ter diversas variáveis presentes. Em alguns casos a relação entre elas pode ser um indício dos fenômenos genéticos de formação das rochas. Em outros casos apenas estamos interessados em como uma informação secundária pode estar relacionada com uma primária de interesse. Seria proveitoso para nós, por exemplo, traçar um modelo que definisse a chance de obter uma amostra com certo teor em contrapartida de outra amostra com o teor de uma variável diferente. Em um depósito vulcanogênico sulfetado podemos estar interessados em prever a quantidade de um elemento

metálico a partir do enxofre da rocha encaixante. Enfim, toda a informação que relaciona duas variáveis pode ser descrita pela estatística bivariada.

Diferentemente da estatística univariada, a comparação de histogramas de variáveis diferentes não é uma alternativa interessante sobre o ponto de vista prático. É muito difícil determinar a relação entre duas amostras simplesmente pelas suas proporções individuais. Para isso definimos algumas ferramentas que facilitam ao modelador entender a relação entre duas variáveis distintas visualmente e numericamente.

As seções que se prosseguem mostrarão algumas das ferramentas utilizadas para se caracterizar distribuições bivariadas. Inicialmente apresentamos as **ferramentas gráficas** mais utilizadas e depois algumas **estatísticas pontuais** utilizadas.

5.2 Probabilidade condicional e Esperança condicional

5.2.1 Probabilidades condicionais e conjuntas

Probabilidades não são nada além de métricas de conjuntos, proporções de acordo um espaço amostral (Ω). Estas proporções podem tomar diferentes características quando analisamos não apenas um conjunto individual, mas a interação entre eles. Muitas vezes não estamos interessados em determinar as probabilidades ou frequências individuais de uma variável aleatória. É interessante, por exemplo, determinar combinações entre variáveis e suas possíveis relações. E se desejarmos saber qual é a frequência de um minério e que seu conteúdo tenha um determinado valor de impureza? Se denotarmos X como o evento de ser minério, e Y , a variável que denota seu limite de impurezas, podemos denotar a probabilidade de $P(X, Y)$ como sendo a probabilidade de que "*Um material seja minério e apresente impurezas acima do limite desejado*". A forma mais simples de se entender probabilidades é de acordo com um diagrama de Venn, como na figura 5.1

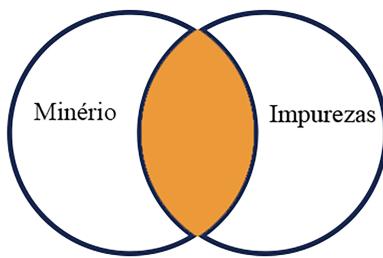


Figura 5.1: Demonstração de eventos disjuntos entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos representado pela probabilidade $P(X, Y)$

Observe a tabela 5.1. Notamos na coluna três o número de vezes que o minério

considerado possui uma impureza maior ou igual a 0,005. Neste caso sabemos que há 2 valores em cinco em que isso ocorre. Logo a probabilidade conjunta é $P(\text{Minério}) \cap P(\text{Impureza} \geq 0,005) = 2/5 = 40\%$

Tabela 5.1: Tabela da relação entre um dado minério e uma impureza

Minério	Impureza	$\text{Minério} \cap (\text{Impureza} \geq 0,005)$
Sim	0,005	1
Não	0,007	0
Não	0,008	0
Sim	0,006	1
Sim	0,003	0

Em alguns casos também é importante determinar a conjunção entre os eventos, ou a probabilidade de $P(X \vee Y)$. Neste caso queremos determinar "*Um material seja minério ou apresente impurezas acima do limite desejado*". Note que a conjunção 'ou' é um conectivo lógico muitas vezes dispare do seu uso corriqueiro no português. Ser um ou outro na verdade não é uma escolha entre um dos elementos, mas uma soma dos eventos. A representação da conjunção pode ser vista na figura 5.2.

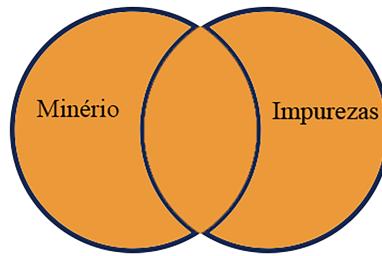


Figura 5.2: Demonstração da conjunção de eventos entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos representado pela probabilidade $P(X \vee Y)$

Estas relações lógicas envolvem o conhecimento entre os eventos independentemente. Conhecer $P(X, Y)$ é exatamente o mesmo que conhecer $P(B, A)$. Em alguns casos devemos entender o conceito de dependência na estatística, expresso pela probabilidade condicional. Neste caso queremos saber "*dado que um material apresente impurezas, qual é sua probabilidade de ser minério*". Esta é uma afirmação muito diferente da obtida nos outros casos, pois para sabermos se algo é minério, precisamos saber antes se ele contém impurezas. A probabilidade condicional $P(X|Y)$ pode ser demonstrada pela figura 5.3 como a relação da área laranja pela área hachurada.

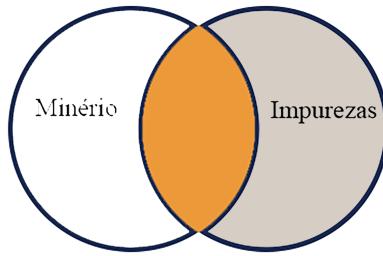


Figura 5.3: Demonstração da probabilidade condicional entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos disjuntos. A probabilidade condicional é a relação entre a área laranja pela área hachurada.

Esta relação também é chamada de teorema de Bayes, e envolve a equação (5.1)

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \quad (5.1)$$

Proposição 5.2.1 *As probabilidades condicionais expressam um importante conceito na geoestatística, a dependência entre variáveis aleatórias. Quando estudamos fenômenos espaciais, os valores obtidos em um suporte específico x_1 são muito mais dependentes de x_2 do que x_3 , se a distância de $\{x_1, x_2\}$ for menor que a distância de $\{x_1, x_3\}$*

5.2.2 Esperança condicional

A partir da definição de probabilidade condicional também é possível determinar a esperança condicional. Ela nada mais é que o valor médio obtido de uma variável Y dado que a variável X assuma um valor específico x . Por exemplo, podemos determinar qual é a probabilidade do material ser contaminado Y , dado que a presença de um litotipo X seja $x = \{\text{itabirito}\}$, de acordo com a equação (5.2).

$$E(Y|X = x) = \sum_{y \in Y} y P(Y = y|X = x) \quad (5.2)$$

Considere que Y seja uma variável binária tal que assuma o valor 0 para o elemento contaminado, e valor 1 para não contaminado. A variável X pode assumir os valores de itabirito e calcário no problema. Analisando a tabela 5.2 podemos determinar as probabilidades de acordo com os respectivos valores apresentados.

Tabela 5.2: Tabela da relação entre um minério contaminado e litotipo

Y	X	$(Y=0 X=\text{itabirito})$	$(Y=1 X=\text{itabirito})$
contaminado	itabirito	1	0
descontaminado	calcário	0	0
descontaminado	calcário	0	0
contaminado	itabirito	1	0
descontaminado	itabirito	0	1
$P(Y=y X=x)$		2/3	1/3

O valor esperado condicional pode ser obtido a partir da tabela pode ser calculado por (5.3)

$$E(Y|X = \text{itabirito}) = 2/3.0 + 1/3.1 = 1/3 \quad (5.3)$$

No caso de variáveis reais, aos quais as probabilidades não são explícitas diretamente, opta pelo uso de estatísticas intervalares. Neste caso desejamos obter $E(Y|x_1 < X < x_2)$. Podemos obter, por exemplo, o valor da recuperação metalúrgica de carvão, dado que os valores de enxofre estejam entre 5% e 6%, por exemplo. O valor da esperança condicional considerando uma distribuição contínua das variáveis X e Y pode ser expressa pela equação (5.4)

$$E(Y|X) = \int_{-\infty}^{-\infty} y f_{Y|X}(y, x) dy \quad (5.4)$$

Onde $f_{Y|X}(y, x)$ representa a função de densidade de probabilidade condicional. A figura 5.4 apresenta como calcular estatísticas condicionais considerando estatísticas intervalares. O histograma, ou a distribuição dos dados são consideradas dentro dos limites específicos $x_1 < X < x_2$, para um determinado tamanho da classe.

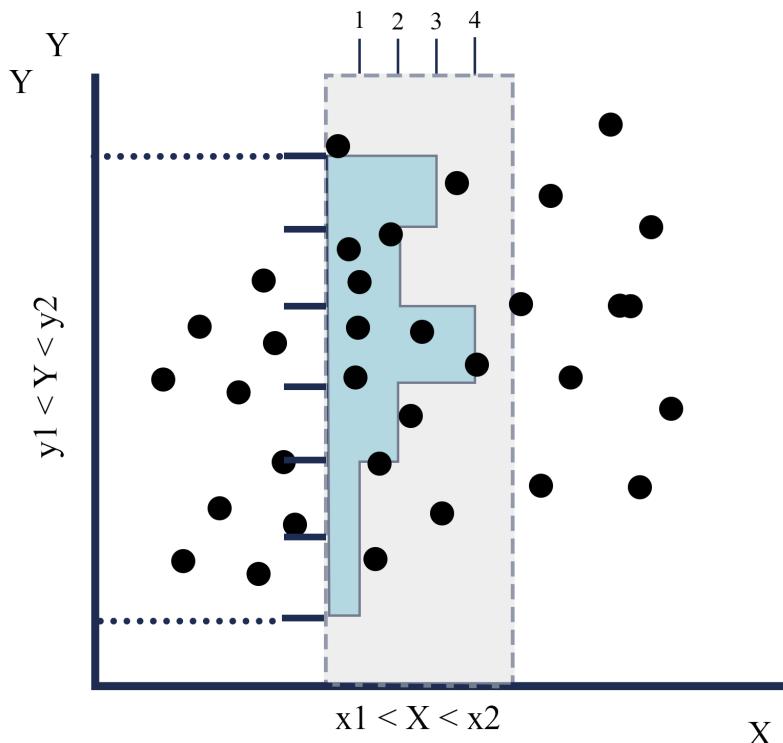


Figura 5.4: Demonstração do histograma condicional considerando um intervalo de $x_1 < X < x_2$ e $y_1 < Y < y_2$. Podemos considerar

5.3 Ferramentas gráficas

5.3.1 Gráfico Q-Q plot

O gráfico q-q plot é uma ferramenta para uma primeira análise de diferentes distribuições de variáveis aleatórias. Para cada par conjugado são plotados os quantis de uma variável juntamente com outra. Variáveis que possuam distribuição semelhante tendem a apresentar um comportamento segundo uma reta $y = x$, de inclinação 45° .

Quando as variáveis apresentam a mesma forma, mas deslocamentos diferentes, ou seja, médias diferenciadas, o gráfico q-q plot apresenta o mesmo formato de uma reta, mas um deslocamento em sua abscissa. Quando as distribuições possuem formas semelhantes, mas variâncias diferentes, a distribuição tende a ter uma inclinação diferente. No entanto, quando distribuições possuem assimetrias e formas diferentes, o gráfico q-q plot tende a produzir uma convexidade diferente. A Figura 5.5 demonstra o gráfico q-q plot das variáveis Cobalto e Cádmio.

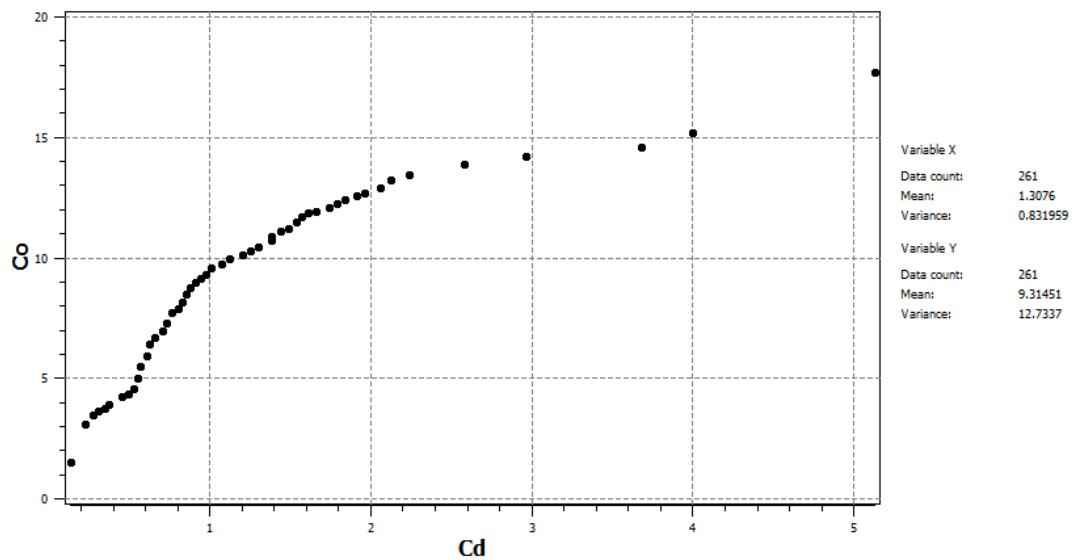


Figura 5.5: Gráfico QQ-Plot de Cobalto e Cádmio. Nota-se uma curvatura característica demonstrando pequena correspondência entre as duas populações. Cada ponto representa o mesmo quantil para cada variável

Nota-se na figura que o formato do q-q plot é convexo, demonstrando que as distribuições de dados seguem leis diferenciadas. A figura 5.6 demonstra a comparação entre os histogramas.

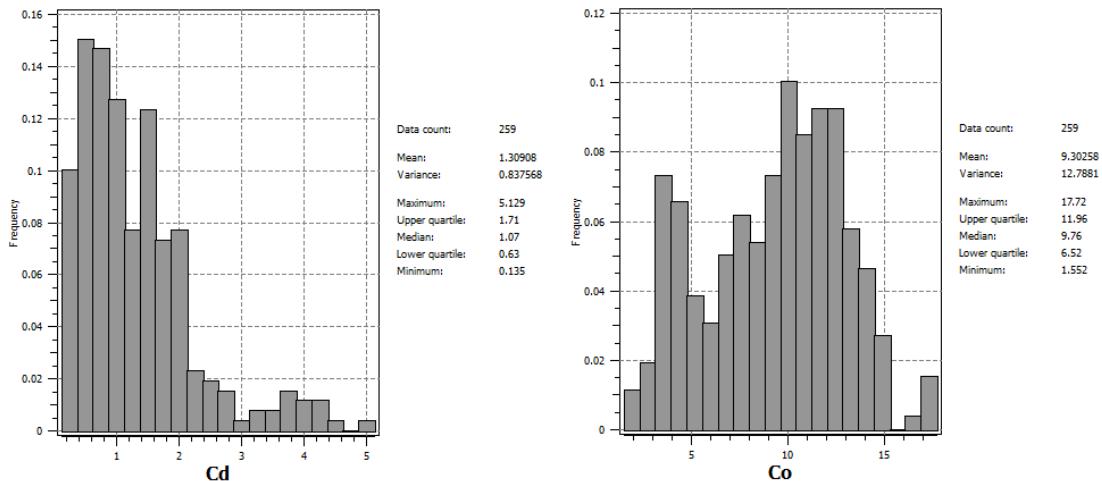


Figura 5.6: Diferenças entre os histogramas de Cádmio e Cobalto

Proposição 5.3.1 *O gráfico q-q plot é uma alternativa para comparar distribuições de variáveis diferentes. A utilização da ferramenta, deve ser no entanto, utilizada com sabedoria. Valores outliers podem prejudicar a comparação entre as distribui-*

ções, o que não significa que possam ser identificadas como possíveis distribuições semelhantes.

Gráficos q-q plot podem ser utilizados não apenas entre amostras, mas também com uma combinação de uma variável amostrada e os quantis teóricos de uma distribuição. Uma das formas de se averiguar a normalidade de uma distribuição é comparar as amostras padronizadas $Z_{pad} = (Z - \bar{x})/S$ com uma variável gaussiana padrão $\phi(0, 1)$. A figura 5.7 demonstra

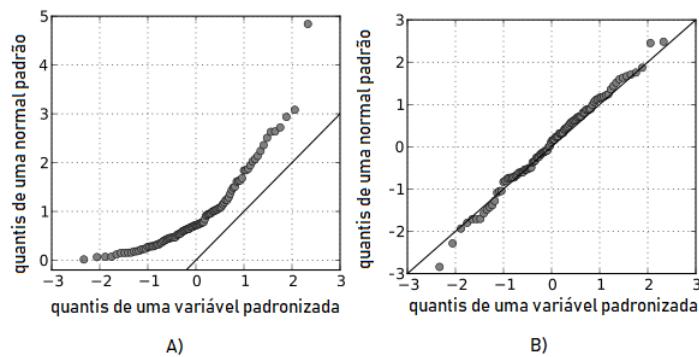


Figura 5.7: Gráfico da utilização do q-q plot para ajuste de uma distribuição. Quantis de uma amostra padronizada comparadas com quantil de uma distribuição gaussiana padrão. A) Mal ajuste. B) Bom ajuste

5.3.2 Gráfico p-p plot

Semelhante ao gráfico q-q plot temos o gráfico p-p plot. O gráfico de probabilidades implica nos pares conjugados que indicam a mesma probabilidade ($Pr(Z < z), Pr(Y < z)) \forall z \in Z, Y$). A figura 5.8 demonstra o gráfico da variável Cobre pela de Cromo.

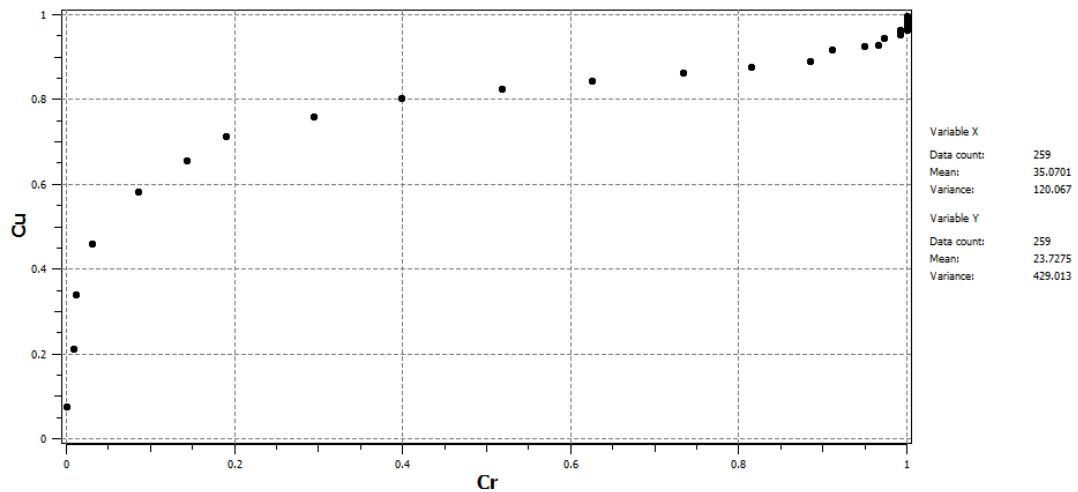


Figura 5.8: Gráfico PP-Plot de Cobre e Cromo. Nota-se uma curvatura característica demonstrando pouca correspondência entre as duas populações. Cada ponto representa o percentual acumulado para o mesmo valor da variável aleatória

Podemos notar que as diferenças demonstradas no gráfico p-p plot se reproduzem nas diferenças entre os histogramas de cobre e cromo na figura 5.9

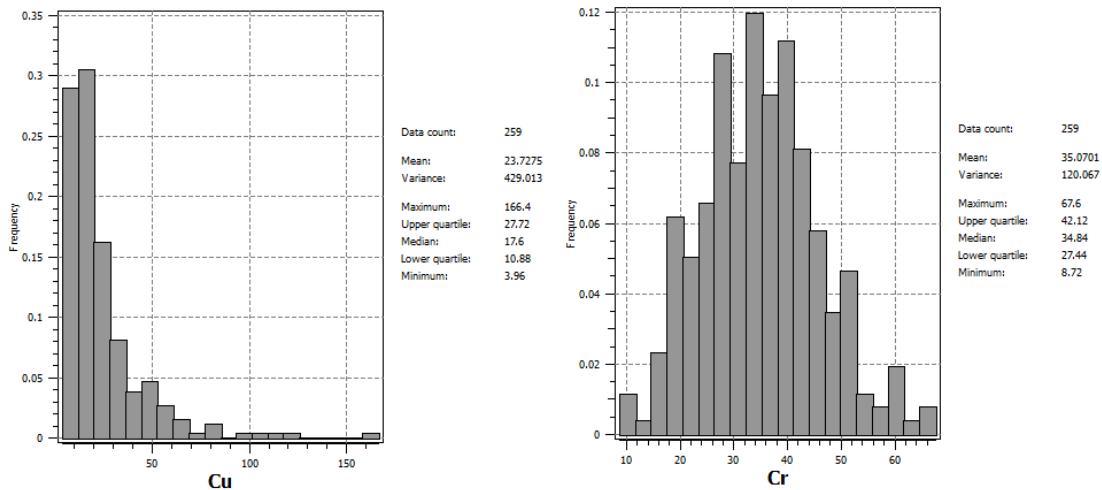


Figura 5.9: Diferenças entre os histogramas de Cobre e Cromo

A análise do gráfico é feita de forma semelhante ao QQ-plot, no entanto, este gráfico é muito mais sensível à mudança de escala das variáveis. Ele é mais vantajoso quando a ordem de grandeza das variáveis analisadas for semelhante. Neste caso estamos comparando a relação de percentuais acumulados diferentes para o mesmo valor da variável aleatória.

5.3.3 Gráfico de dispersão

O gráfico de dispersão apresenta dados de duas variáveis dispostos nos eixos cartesianos. Temos uma variável **preditora** (X) e uma variável **resposta** (Y). Os pares conjugados $(x_i, y_i) \in X, Y$ representam pontos em um plano cartesiano.

Proposição 5.3.2 *Uma das primeiras utilizações da regressão linear foi no estudo da importância de tendências entre gerações. Durante o período de 1893-1898, E. S. Pearson organizou uma coleção de $n=1375$ alturas de mães do reino Unido abaixo de 65 anos e uma de suas filhas acima de 18. O interesse era computar o tamanho das mães ($Mheight$) com o tamanho das filhas ($Dheight$) como preditor. Se todas as mães possuirem tamanho igual suas filhas, os dados estarão dispostos em uma reta de inclinação 45° . A linearidade proposta pela dispersão identifica que mães mais altas geralmente possuem filhas mais altas. - Weisberg [2005]*

Para a utilização do gráfico os dados devem estar colocados. Isso significa que a amostra 1 deve ter a mesma origem da amostra 2, ou o mesmo suporte. Logo só podemos realizar um gráfico de dispersão com vetores de amostras do mesmo tamanho.

Caso a amostragem apresente dados inválidos para uma variável devemos utilizar um filtro para separar apenas os dados colocados. Existem técnicas estatísticas que permitem o tratamento de dados perdidos ou inexistentes, mas nada substitui a amostra em termos de informação sobre o objeto de estudo. A figura (5.10) demonstra um gráfico de dispersão entre a variável cromo e cobalto.

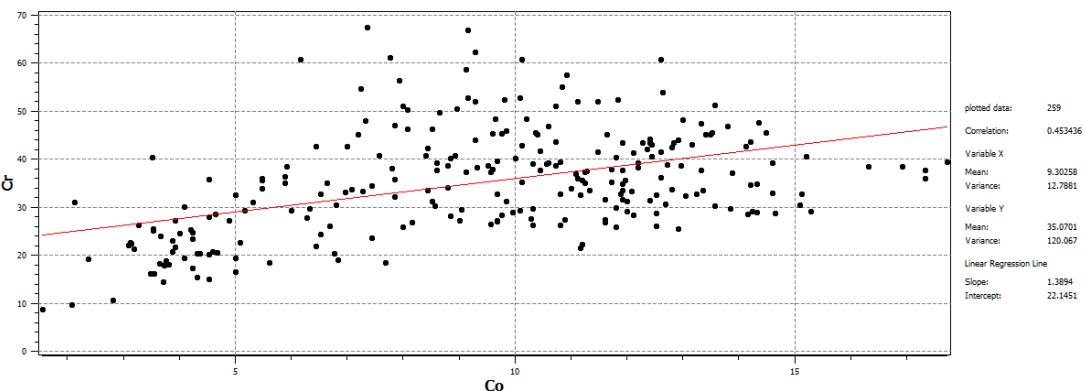


Figura 5.10: Gráfico de dispersão da variável Cromo e Cobalto. Nota-se dependência linear positiva entre as variáveis.

Nota-se pela figura que as variáveis possuem **dependência linear positiva** entre a variável Cromo e Cobalto. Isso significa que amostras com valor grande

de cromo podem apresentar valores grandes de cobalto. O contrário também pode acontecer, alguns minerais como quartzo e piroxênio são inversamente proporcionais em rochas magmáticas. À medida em que se aumenta o teor de quartzo tende-se a reduzir o teor de piroxênio na amostra de rocha. Neste caso possuímos uma **dependência linear negativa**

A Figura 5.11 demonstra os tipos de correlação lineares possíveis. Em 5.11 -a temos a correlação linear positiva em que o aumento da variável X aumenta o valor de Y, em 5.11 -b temos a correlação linear negativa em que o aumento do valor X tende a diminuir o valor de Y e em 5.11 -c temos um caso de independência entre as variáveis, tal que o aumento da variável X não altera o valor da variável Y.

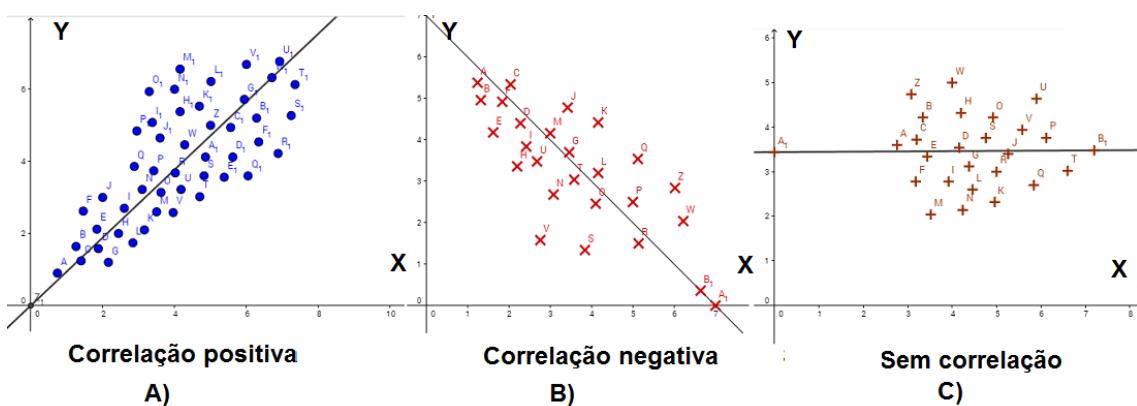


Figura 5.11: Figura demonstrando os tipos de correlação linear possíveis. A) Correlação linear positiva, B) Correlação linear negativa, C) Sem correlação

Os gráficos de dispersão também são uma boa medida para a visualização de valores outliers. A figura (5.12) demonstra a dispersão anterior mas com uma área circulada de pontos que não estão dentro do comportamento linear das variáveis. Neste caso para valores intermediários de Cobalto temos grandes valores de Cromo.

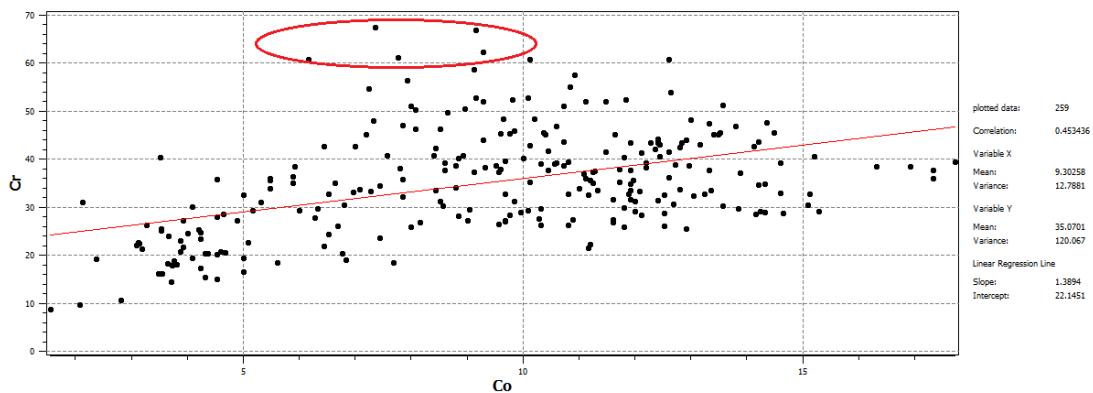


Figura 5.12: Gráfico de dispersão da variável Cromo e Cobalto demonstrando valores outliers. Círculo vermelho indica possíveis valores fora dos padrões das variáveis conjuntas

Muitas vezes um valor outlier em um gráfico bivariado não é demonstrado no tratamento individual das amostras. Muito cuidado deve ser tomado para a retirada de pares anômalos das estatísticas, pois eles podem gerar novos valores discrepantes e não demonstrarem um padrão de maior correlação entre as variáveis.

5.4 Regressão linear

O modelo de regressão linear simples é aquele em que definimos uma dependência diretamente proporcional entre a variável resposta Y e preditora X . Podemos associar o valor esperado da variável resposta dado valores da variável preditora tal que $E(Y|X = x) = \beta_0 + \beta_1 x$. Note que $E(Y|X = x)$ corresponde ao **valor médio da variável Y condicionado a um valor x da variável X** , e β_0 e β_1 , também são o **intercepto da reta no eixo das abcissas e a tangente do ângulo de inclinação da reta**. Muitas pessoas acabam por não entender que a regressão linear pode não apresentar uma representação acurada da resposta dado um valor da variável preditora, porque o valor estimado pela regressão não é o valor da variável resposta, mas sim seu valor esperado condicionado. Muitas variáveis apresentam alta dispersão em torno de seus valores médios e podem não ser estimativas plausíveis. A solução da **regressão linear ordinária** geralmente advém do método dos mínimos quadrados. Considere $\hat{y}_i = \hat{E}(Y|X = x)$ como um estimador para $E(Y|X = x)$, logo teremos que o resíduo pode ser demonstrado pela equação $\hat{y}_i - y_i$ segundo a equação

Demonstração. Regressão linear pelo método dos mínimos quadrados

$$\begin{aligned}\epsilon_i &= \hat{y}_i - y_i \\ \epsilon_i^2 &= \hat{y}_i^2 + y_i^2 - 2\hat{y}_i y_i \\ \epsilon_i^2 &= (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 + y_i^2 - 2(\hat{\beta}_0 + \hat{\beta}_1 x_i)y_i\end{aligned}$$

A soma dos erros quadráticos para cada i

$$\sum_{i=0}^n \epsilon_i^2 = \sum_{i=0}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 + \sum_{i=0}^n y_i^2 - \sum_{i=0}^n 2(\hat{\beta}_0 + \hat{\beta}_1 x_i)y_i$$

Tomando as derivadas parciais segundo os parâmetros: $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{aligned}1) \frac{\partial \sum_{i=0}^n \epsilon_i^2}{\partial \hat{\beta}_0} &= 2\hat{\beta}_0 n + 2\hat{\beta}_1 \sum_{i=0}^n x_i - \sum_{i=0}^n 2y_i = 0 \\ 2) \frac{\partial \sum_{i=0}^n \epsilon_i^2}{\partial \hat{\beta}_1} &= 2\hat{\beta}_1 \sum_{i=0}^n x_i^2 + 2 \sum_{i=0}^n \hat{\beta}_0 x_i - 2 \sum_{i=0}^n y_i x_i = 0\end{aligned}$$

■

A obtenção dos parâmetros pode ser facilmente encontrada isolando os termos das equações 1 e 2. Podemos então determinar

Demonstração. Obtenção dos parâmetros da regressão

$$\hat{\beta}_0 = \left(\sum_{i=0}^n y_i - \hat{\beta}_1 \sum_{i=0}^n x_i \right) / n$$

Substituindo em 2

$$\begin{aligned}\hat{\beta}_1 \sum_{i=0}^n x_i^2 + \sum_{i=0}^n \left(\sum_{j=0}^n y_j - \hat{\beta}_1 \sum_{j=0}^n x_j \right) x_i / n - \sum_{i=0}^n y_i x_i &= 0 \\ \hat{\beta}_1 \sum_{i=0}^n x_i^2 - \hat{\beta}_1 \bar{x} \sum_{i=0}^n x_i + \bar{x} \sum_{i=0}^n y_i - \sum_{i=0}^n y_i x_i &= 0 \\ \hat{\beta}_1 \left(\sum_{i=0}^n x_i^2 - \bar{x} \sum_{j=0}^n x_j \right) &= \sum_{i=0}^n y_i x_i - \sum_{j=0}^n y_j \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=0}^n y_i x_i - \sum_{j=0}^n y_j \bar{x}}{\sum_{i=0}^n x_i^2 - \bar{x} \sum_{j=0}^n x_j}\end{aligned}$$

■

O problema de regressão linear se torna então um problema de otimização ao encontrar o menor somatório dos desvios quadráticos. A figura (5.14) demonstra graficamente o problema da regressão linear. Neste caso os valores dos coeficientes lineares das retas podem ser encontrados a partir de derivação simples ou por meio de métodos numéricos, como a utilização do **método Simplex**. O resultado também é análogo ao **método de máxima verossimilhança** considerando a distribuição dos resíduos como gaussianos.

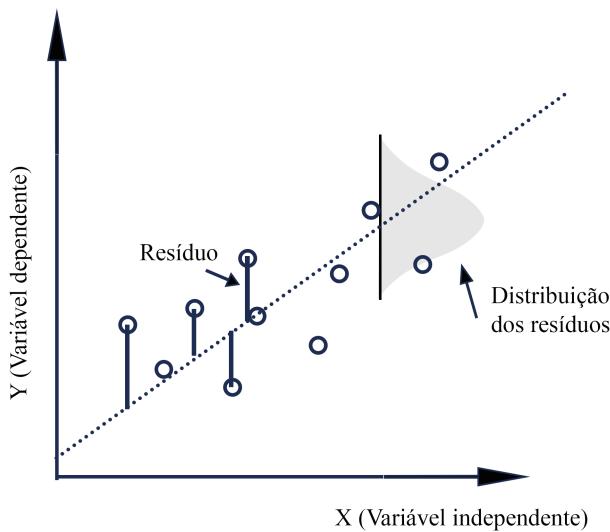


Figura 5.13: Explicação da regressão linear entre a variável independente X e a variável dependente Y. Barras verticais representando os desvios das amostras com o valor médio.

Como a regressão linear assume a minimização dos valores quadráticos, os parâmetros β_0 e β_1 podem ser fortemente afetados por valores outliers. As propostas mais modernas de regressão prevem que a regressão seja utilizada apenas em parte dos dados, e avaliada com outra quantidade dos dados. Os dados utilizados para a estimativa dos parâmetros é chamada de Assim conseguimos avaliar a qualidade da predição de acordo com a dispersão destes resíduos.

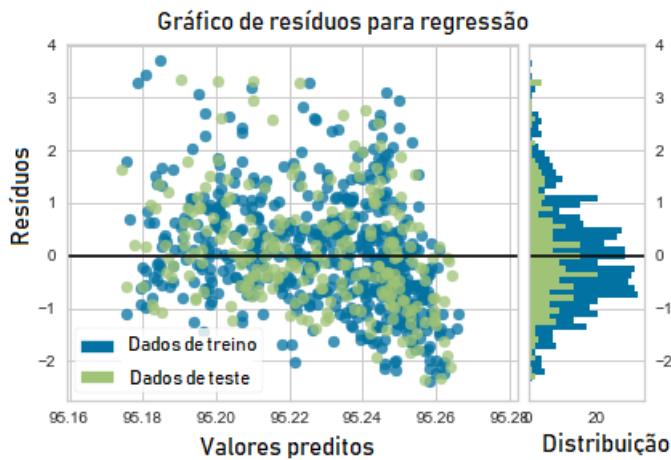


Figura 5.14: Gráfico da dispersão de resíduos nos bancos de dados de treino e teste. Dados de treino utilizados para determinar os parâmetros da regressão e dados de teste para avaliar as diferenças do modelo de predição

Na geoestatística utilizamos estimadores lineares, semelhantes ao processo de regressão linear. É observado que para aplicações na mineração que utilizem os métodos geoestatísticos clássicos, o erro do valor estimado é ligeiramente diferente de uma distribuição gaussiana, para problemas lineares e estacionários. Temos uma maior confiabilidade do valor esperado estimado utilizando krigagem ordinária do que utilizando regressão linear ordinária. Na verdade, o método de regressão linear ordinária é pouco usual nos dias atuais, considerando os diferentes modelos possíveis e robustos (menos influenciados pelos valores outliers), no entanto, ainda é um método muito popular pela sua simplicidade e facilidade de aplicação.

R *Em aplicações da mineração, a distribuição dos erros da função aleatória são geralmente simétricos com um crescimento mais pronunciado na moda e caudas alongadas do que para distribuições normais com mesma média e variância. Então em relação a uma distribuição normal, há menos erros na região próxima ao valor estimado e mais erros nas caudas. - Journel and Huijbregts [1978]*

5.5 Intervalo de segurança para a regressão linear

A determinação do modelo de regressão consiste em estimar parâmetros β_0 e β_1 para econtrarmos o valor estimado $\hat{y}_i = E(Y|X = x_i)$. No entanto, se a nuvem de pontos determinada pela regressão for esparsa, o valor \hat{y}_i não possui capacidade preditiva e pode encontrar-se dentro de limites amplos. Considerando que a distribuição dos resíduos seja normalmente distribuída, podemos encontrar os limites da regressão

para bandas superiores e inferiores, determinando assim a confiabilidade desta reta regredida. A equação (5.5) demonstra como podem ser calculadas as bandas de incerteza da regressão de acordo com um nível de significância estipulado.

$$\hat{y}_i \pm t_{(n-2,p)}^* s_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}} \quad (5.5)$$

Em que x_i é o valor da variável X, t^* é o valor da distribuição de t-student para um grau de liberdade igual a $n - 2$ e nível de significância p, enquanto s_y pode ser demonstrado segundo a equação (5.6)

$$s_y = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} \quad (5.6)$$

Em que y_i é o valor da coordenada y para um ponto amostral i. Ou seja s_y é o valor do desvio padrão entre os valores amostrais e os valores médios estimados pela regressão.

A figura (5.15) demonstra o intervalo de segurança para o valor regredido.

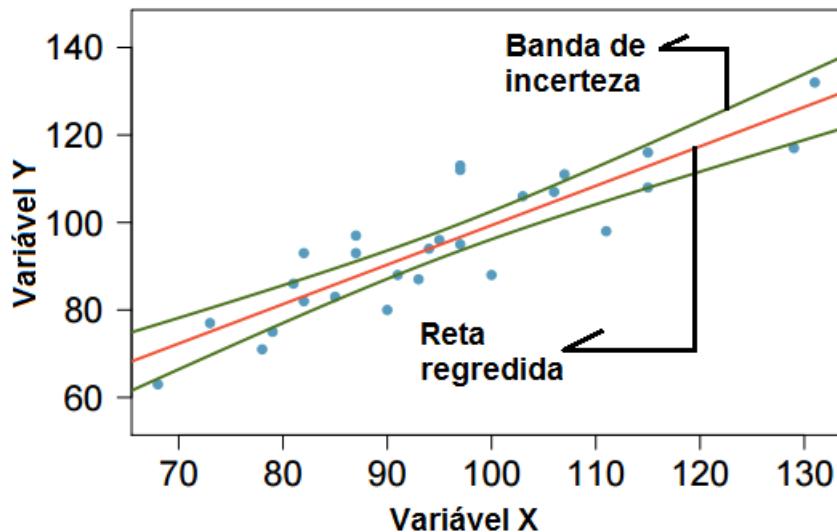


Figura 5.15: Demonstração do intervalo de confiança para a regressão linear. Banda de incerteza adicionada como limite inferior e superior dado pela equação (5.5)

Nota-se que as bandas apesar de acompanharem o valor de regressão linear não são retas, apresentando um maior estreitamento na região mediana da dispersão. A confiabilidade do centro de dispersão da reta regredida é sempre maior.

Proposição 5.5.1 Os intervalos de confiança para a regressão linear estipulam que os resíduos seguem uma distribuição gaussiana. Isto porém, pode não se apresentar na prática. A única garantia que temos para que este resíduo seja considerado gaussiano, é se por ventura, a distribuição dos dados segue uma lei de probabilidades **multigaussiana**. Este é o pressuposto de técnicas mais avançadas de geoestatística não-linear. As bandas de incerteza devem ser consideradas como uma alternativa para verificar dados discrepantes, mas não como uma métrica de decisão na detecção de valores outliers.

5.6 Regressão linear múltipla

Quando pensamos em apenas uma variável preditora, a determinação de \hat{y}_i se limita a encontrar o valor de $E(Y|X = x_i)$. Porém quando múltiplas variáveis são relacionadas o problema se torna encontrar $E(Y|X^1 = x_i^1, X^2 = x_i^2, \dots, X^n = x_i^n)$. O modelo de regressão linear múltipla é um caso extendido da regressão linear simples para múltiplas variáveis. Neste caso temos um conjunto de n variáveis preditoras e uma variável resposta

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \dots + \hat{\beta}_n x_i^n = \sum_{j=0}^p \hat{\beta}_j x_i^j \quad (5.7)$$

Onde x_i^0 é sempre igual a 1 para j variáveis de 0 a p . O problema se resume a encontrar $\hat{\beta}_p$ parâmetros que aproximem melhor a combinação dos valores das variáveis x_i^p de \hat{y}_i . Podemos definir o problema de regressão linear múltipla a partir de sua forma matricial pela equação (5.8)

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ 1 & x_2^1 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \quad (5.8)$$

Ou de forma simplificada pela equação (5.9)

$$\bar{Y} = \bar{X}\bar{\beta} \quad (5.9)$$

Onde \bar{Y} representa o vetor da variável resposta, \bar{X} a matriz das variáveis preditoras e $\bar{\beta}$ os parâmetros. A obtenção dos parâmetros a partir da regressão múltipla

pode ser facilmente encontrado através de operações matriciais.

Demonstração. Obtenção dos parâmetros da regressão

$$\begin{aligned}\bar{Y} &= \bar{X}\bar{\beta} \\ \bar{X}\bar{Y} &= \bar{X}\bar{X}\bar{\beta} \\ (\bar{X}\bar{X})^{-1}\bar{X}\bar{Y} &= \bar{\beta} \\ (\bar{X}\bar{X})^{-1}\bar{X}\bar{Y} &= \bar{\beta} \\ X^\dagger\bar{Y} &= \bar{\beta} \\ \text{tal que } X^\dagger &= (\bar{X}\bar{X})^{-1}\bar{X}\end{aligned}$$

■

Em que X^\dagger também é chamada de pseudo-inversa de X . Se no caso da regressão ordinária simples obtínhamos um valor regredido a partir da minimização do resíduo de uma variável, neste momento obtemos o resíduo a partir de uma combinação de múltiplas variáveis. O erro quadrático pode ser obtido a partir da equação (5.10).

$$\sum_i \epsilon_i^2 = \sum_i \left(y_i - \sum_{j=0}^p \hat{\beta}_j x_i^j \right)^2 \quad (5.10)$$

A obtenção dos parâmetros β_p pode ser calculado a partir das técnicas de minimização dos resíduos, formando um sistema de p derivadas parciais. Um dos grandes problemas da regressão linear múltipla é o fato de que as grandezas de variáveis diversas podem ser diferentes e impactar de forma diferenciada nos pesos da regressão. Esse problema de dimensão geralmente pode ser minimizado se padronizarmos as variáveis como visto no capítulo 4. Outra questão envolvendo a regressão múltipla é o fato de que valores outliers conseguem ser ainda mais prejudiciais que a regressão linear ordinária, afetando muito a estimativa dos pesos. Uma tentativa de excluir certos efeitos de valores discrepantes é introduzir uma constante adicional chamada de **regularização** (λ) multiplicando os valores dos parâmetros β_p .

5.7 Coeficiente de correlação

Observamos na seção anterior que se duas variáveis são dependentes, podemos assumir que existe uma probabilidade $P(Y|X = x)$. No entanto, as probabilidades condicionais parecem não fornecer um quadro geral da dependência de uma variável

aleatória Y , pois precisamos saber qual valor a variável X deve assumir. É necessário ter uma métrica para avaliarmos o quanto forte ou fraca é a dependência entre as variáveis. Imagine o caso onde temos um depósito hidrotermal de ouro, associado principalmente a rochas magmáticas sulfetadas. Se existir uma alta dependência do conteúdo de ouro com o de enxofre, podemos assumir que o conhecimento de uma variável auxiliará no conhecimento da outra. Porém valores pequenos de teor de enxofre podem ser menos dependentes do teor de ouro do que para altos valores do teor de enxofre. Esta discrepância dado alguns limites pode ser favorável para o uso de probabilidades condicionais nas caudas da distribuição de enxofre, mas não garante uma visão geral da dependência linear entre estas variáveis.

R *Tanto na natureza como em vários problemas de engenharia nos deparamos com a dependência entre diferentes variáveis. Em muitos casos estas dependências podem ser modeladas linearmente. Em outros casos, quando conhecemos propriedades físicas relacionáveis, podemos utilizar transformações lineares, capazes de transformar modelos não lineares em lineares.*

No capítulo 3 observamos a correlação como uma medida de dependência entre variáveis aleatórias. A covariância teórica pode ser estimada a partir de sua covariância experimental pela equação (5.11)

$$\hat{Cov}_{X,Y} = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (5.11)$$

Onde \bar{x} e \bar{y} são as médias aritméticas entre as variáveis X e Y para um número de amostras n. A covariância experimental pode ser muito bem comparada ao produto escalar obtido pela multiplicação de dois vetores $X * Y = \|X\| \|Y\| \cos(\theta)$, em que $\cos(\theta)$ é chamado de cosseno diretor da projeção de X em Y . Quando a projeção do vetor X corresponde ao vetor Y temos o máximo valor possível, no entanto, quando estes vetores são perpendiculares temos um valor de $\cos(90^\circ) = 0$, e portanto $\hat{Cov}_{X,Y} = 0$

A covariância é uma medida muito susceptível a valores outliers, pois como X e Y podem ter unidades discrepancytes, o produto das duas pode ser mais influenciado por aquela variável que apresentar maiores valores. Imagine que procuremos a relação entre a massa em quilos dos testemunhos e o teor de um elemento químico. Como a massa dos testemunhos poderá variar de valores acima da unidade, como por exemplo 12kg e os teores apenas com valores decimais, como 0.12, a importância dada para as variações do peso serão maiores. Isto torna a covariância pouco comparativa, apenas se considerarmos a normalização das variáveis. A alternativa para

isto é normalizarmos a covariância pelos desvios padrões das respectivas variáveis, o que gera o **coeficiente de correlação de Pearson** (5.12).

$$\rho_{X,Y}^p = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5.12)$$

Os valores de $\rho_{X,Y}$ podem variar de -1 a 1, sendo 1 quando apresenta correlação positiva perfeita e -1 quando apresenta correlação negativa perfeita. Quando $\rho_{X,Y}^p = 0$ temos um indicativo de independência entre as variáveis tal que $\hat{Cov}_{X,Y}$ é igual a 0. Podemos obter a relação entre o coeficiente de Pearson e a correlação pela equação (5.13)

$$\rho_{X,Y}^p = \frac{\hat{Cov}_{X,Y}}{\sqrt{S_x^2 S_y^2}} \quad (5.13)$$

Onde $\hat{Cov}_{X,Y}$ representa a covariância estimada, e S_x^2 e S_y^2 as variâncias experimentais das amostras x e y. Para que consigamos calcular o valor do coeficiente de variação e da correlação, precisamos que existam valores correspondentes tanto para as amostras x e y, ou seja, precisamos de dados **colocados**. Se por ventura houverem dados faltantes, não conseguiremos calcular a covariância ou o coeficiente de correlação.

Em alguns casos não desejamos observar a dependência entre os valores da variável, mas precisamos saber qual é a dependência da **ordem** dos dados. Para isto realizamos uma medida chamada de **rank ou posto**, que representa a ordem de uma amostra em seu conjunto.

Definição 5.7.1 — Rank ou posto. *Um rank ou posto é a ordem crescente de uma amostra em seu respetivo conjunto. Por exemplo, um conjunto de amostras com valores $x = 3, 41, 2, 57, 8, 9, 6$, possuirá um rank R_x tal qual $R_x = 2, 6, 1, 7, 4, 5, 3$*

O chamado **coeficiente de correlação de Spearman** nada mais é que a correlação de Pearson considerando seus respectivos postos.

$$\rho_{X,Y}^s = \frac{\hat{Cov}_{R_x, R_y}}{\sqrt{S_{R_x}^2 S_{R_y}^2}} \quad (5.14)$$

Enquanto a correlação de Pearson é uma medida linear de dependência, o coeficiente de Spearman é uma medida não linear, demonstrando a correlação entre

crescimentos e descrescimentos das variáveis independente dos valores dos dados. Uma função parabólica tal que $Y = \beta_1 X^2 + \beta_0$ apresentará um valor de coeficiente de correlação de Pearson muito baixo, porém um valor de Coeficiente de Spearman igual a 1.

5.8 Exercícios

Exercícios 5.1 Os dados da tabela abaixo representam valores de Au e cobre medidos concumitamente nos mesmos testemunhos de sondagem. Com estes dados, pede-se:

- Determine a covariância dos dados
- Determine o coeficiente de correlação.
- As amostras são dependentes positivamente ou negativamente?
- Faça um gráfico de regressão linear entre as variáveis ouro e cobre

Au	Cobre
0.012	2.0
0.015	2.02
0.013	1.32
0.070	3.45
0.012	1.02
0.067	2.19
0.090	4.01
0.08	3.67
0.012	1.43
0.011	1.01
0.011	1.05

Exercícios 5.2 Os dados da tabela abaixo representam valores de X e Y. Faça um gráfico de dispersão e determine o par de valor outlier para o gráfico.

X	Y
0.729	1.546
0.757	1.683
0.140	0.175
0.575	0.963
0.408	0.726
0.402	1.104
0.616	1.321
0.958	5.02
0.9136	1.873
0.527	0.853
0.470	0.960



6. Métodos clássicos e desagrupamento

Difficultés rencontrées dans le développement d'une Géostatistique linéaire. A cette émergence lente et difficile, nous apercevons plusieurs sortes de raison, les unes historiques, d'autres simplement psychologiques, et d'autres encore qui correspondent a des problèmes de fond, à de véritables difficultés méthodologiques, ou épistémologiques qui n'ont été vraiment élucidées qu'à la fin des années 60 ou au début des années 70

G. Matheron

6.1 Introdução

Apesar de antiga, a geoestatística se iniciou como um alternativa para tentativas de avaliação de depósitos minerais antes da década de 70. Os métodos chamados de clássicos eram relativamente eficientes para condições de depósitos minerais mais homogêneos e de classificação estatística dentro dos grupos considerados regulares. No entanto, devido a intensa atividade industrial humana, é necessário aproveitarmos cada vez mais depósitos minerais complexos. Diferentemente dos métodos

geoestatísticos, os métodos convencionais baseiam-se apenas na distribuição geométrica entre as amostras e não na correlação e dependência entre variáveis aleatórias. Apesar de ultrapassados, os métodos de avaliação clássica ainda são utilizados em depósitos de baixa variabilidade e com quantidades de amostras muito baixas, como por exemplo, depósitos estratiformes de argila ou areia. Em alguns casos bem específicos os métodos clássicos podem até mesmo apresentar resultados superiores aos métodos geoestatísticos seguindo o princípio da navalha de Occan.

Definição 6.1.1 — Navalha de Occan. *"A navalha de Occam é um princípio lógico atribuído ao filósofo medieval William de Occam. O princípio estabelece que não se deve assumir mais suposições do que necessário. Também é chamado de princípio da parsimônia. Este princípio envolve todo a modelagem científica e construção de teorias. Ele nos incentiva a escolher de um grupo de modelos equivalentes para um dado fenômeno aquele mais simples. Para todo modelo, a navalha de Occan nos ajuda a 'cortar' aqueles conceitos, variáveis ou construções que não conseguem realmente explicar o fenômeno. Ao fazer isso, o desenvolvimento do modelo se torna bem mais simples, e há menos chances de introduzir inconsistências, ambiguidades ou redundâncias"* -[Heylighen \[1997\]](#)

Veja bem que adotar os métodos clássicos em detrimento da geoestatística utilizando a navalha de Occan só possui bons resultados em dois casos. No primeiro o problema é tão simples e o depósito mineral tão homogêneo e pouco amostrado, que se torna factível o uso de um modelo extremamente simplificado. No segundo caso temos um problema tão complexo e variável que se torna impossível encontrar aparentemente um padrão qualquer no fenômeno, sendo mais fácil adotar valores médios pela distância do que realmente utilizar um método geoestatístico refinado. Segundo o professor [Yamamoto \[2001\]](#), os métodos chamados de clássicos baseiam-se no princípio da **interpretação**, aos quais determinam valores a partir de duas amostras contíguas. É possível a partir da disposição das amostras encontrar valores estimados. Estes princípios segundo o professor são:

1. Mudança gradual ou lei de função linear
2. Pontos mais próximos ou esfera de igual influência
3. Generalização ou empírico

6.1.1 Princípio da mudança gradual

O princípio da mudança gradual indica que uma mudança de uma propriedade acontece de forma contínua de uma amostra pontual P_1 até uma amostra pontual P_2 .

Pelo princípio da navalha de Occan, a função utilizada para realizar esta transição geralmente é uma variação linear. Dada uma propriedade como o teor T a ser estimada a partir das propriedades T_1 e T_2 de dois pontos amostrais no espaço, com respectivas distâncias D_1 e D_2 da origem, realizamos a interpolação deste valor a partir de sua distância D entre os pontos pela equação (6.1).

$$T = T_1 + \frac{(D - D_1)(T_2 - T_1)}{(D_2 - D_1)} \quad (6.1)$$

A figura 6.1 apresenta o resultado geométrico da interpolação linear realizada entre os pontos amostrais.

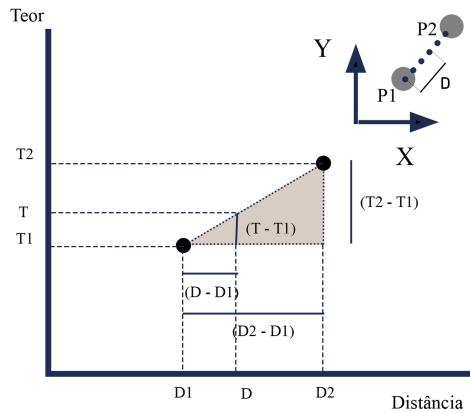


Figura 6.1: Princípio da mudança gradual de um teor para duas amostras no espaço P_1 e P_2 , variação linear dos teores para uma distância D considerada.

6.1.2 Princípio dos pontos mais próximos

O princípio dos pontos mais próximos assume que o valor interpolado é igual ao valor da amostra mais próxima dele. Este processo também é chamado de **vizinho mais próximo**. Dado um conjunto de amostras no espaço $P = \{P_1, P_2, P_3, \dots, P_n\}$ com propriedades $T = \{T_1, T_2, T_3, \dots, T_n\}$, com respectivas posições espaciais segundo um eixo cartesiando (x, y, z) , o valor de uma propriedade T_0 para uma amostra P_0 no espaço assume o valor

$$T_0 = T_i | i \rightarrow \min(\|P_0, P_i\|), \forall i \in [1, n] \quad (6.2)$$

Em que $\|P_0, P_i\|$ representa a distância euclidiana entre o par conjugado de pontos no espaço para n amostras, $\min()$ representa o mínimo valor. A figura 6.2

apresenta o princípio dos pontos mais próximos. Cinco pontos amostrais são apresentados P_1 , P_2 , P_3 , P_4 e P_5 . Como a distância mínima entre o ponto a ser amostrado P_0 até o ponto amostral mais próximo é D_2 , T_0 recebe o valor de T_2 .

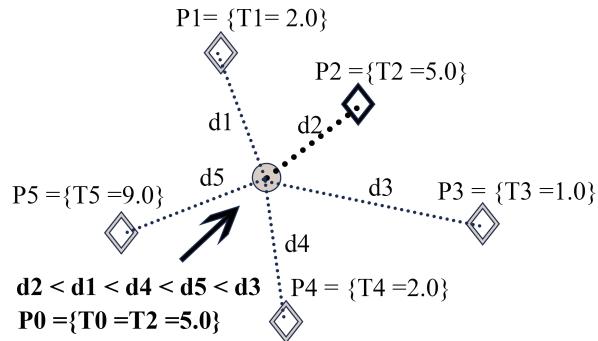


Figura 6.2: Princípio dos pontos mais próximos. Conjunto de pontos amostrais P_1 , P_2 , P_3 , P_4 , P_5 para um ponto amostral estimado P_0 . Como a menor distância euclidiana entre o ponto P_0 é o ponto P_2 , assumimos que a propriedade $T_0 = T_2$.

6.1.3 Princípio da generalização

Segundo critérios geológicos é possível realizar a extração de uma dada propriedade segundo a continuidade do depósito mineral. Este princípio é justificado nas fases inciais de pesquisa para ajustar uma dada tendência das propriedades do depósito mineral. A figura 6.6 demonstra a extração do teor a partir do conhecimento de uma falha geológica entre os pontos amostrais P_1 e P_2 . A partir da atitude da camada é estipulado uma variação segundo os teores extrapolados.

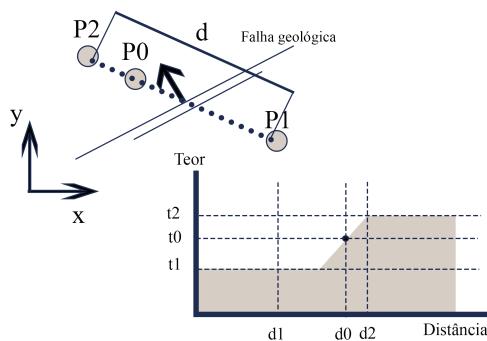


Figura 6.3: Princípio da generalização e extensão dos teores a partir de um critério geológico. Falha geológica observada entre os pontos amostrais P_1 e P_2 . Ponto estimado P_0 é determinado a partir de diferenças entre os teores e inclinação da camada.

6.2 Composição

Para realizar a análise geoestatística é necessário prover de amostras com mesmo **suporte**. Como testemunhos de sondagem podem representar fragmentos de tamanho distintos, é necessário realizar a **regularização** dos tamanhos das amostras. Isto significa que ao invés de tomarmos as propriedades referentes a apenas os fragmentos dos testemunhos de sondagem (litotipo, teores, propriedades físicas, etc.), tomamos a amostra como um valor de média ponderada entre os diversos tamanhos de fragmentos dos testemunhos. Observe a figura 6.4. Os fragmentos do testemunho de sondagem são respectivamente os valores $\{l_1, l_2, l_3, l_4\}$. Para considerarmos este testemunho de sondagem como medidas representativas para a geoestatística consideramos duas composições de tamanho $\{C_1, C_2\}$, em que $\{x_1, x_2\}$ representam as respectivas partes dos fragmentos $\{l_1, l_2\}$ em C_1 . Se t_1, t_2, t_3, t_4 são propriedades aditivas como os teores dos fragmentos, podemos encontrar o valor da propriedade da composição t_{C1} pela relação (6.3)

$$t_{C1} = \frac{(x_1/l_1)t_{l_1} + (x_2/l_2)t_{l_2}}{(x_1/l_1) + (x_2/l_2)} \quad (6.3)$$

Em que (x/l) representa a proporção de um dado fragemento dentro da composta. Podemos generalizar o caso da composição para n fragmentos de acordo com a relação (6.4)

$$t_C = \frac{\sum_{i=0}^n (x_i/l_i)t_{l_i}}{\sum_{i=0}^n (x_i/l_i)} \quad (6.4)$$

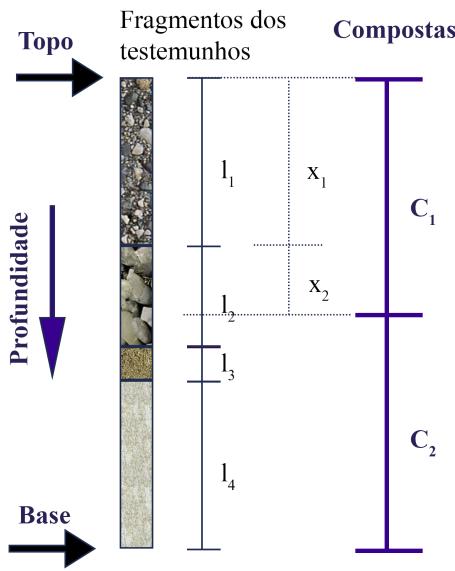


Figura 6.4: Composição realizada em testemunho com 4 fragmentos $\{l_1, l_2, l_3, l_4\}$ e tamanhos de composta igual $\{C_1, C_2\}$. $\{x_1, x_2\}$ representam respectivamente as proporções dos fragmentos $\{l_1, l_2\}$ dentro da composta C_1 .

A escolha do tamanho da composição do testemunho geralmente está associada ao comprimento da manobra. Em alguns casos quando são realizados furos com manobras de tamanho diferenciado pode se optar pela moda dos valores de manobra.

Proposição 6.2.1 *Devemos lembrar antes de realizar a composição de uma dada propriedade se ela pode ser caracterizada como uma variável aditiva, tal como teores e massas. A composição de testemunhos de sondagem considerando propriedades não aditivas deve ser realizada a partir dos estimadores adequados para as tendências centrais. Por exemplo, se considerarmos a velocidade de propagação de um pulso sísmico nos fragmentos, sabemos que a média correta de velocidade é a **harmônica** e não a média **aritmética**.*

6.3 Composição em seções verticais

É comum durante o processo de estimativa, o geólogo realizar seções interpretadas de um depósito mineral, considerando os aspectos estruturais do depósito, o controle geológico entre outras características que formam um **modelo geológico**. Diferentemente de um **modelo estatístico** que prevê o conhecimento de variáveis aleatórias do depósito mineral, o modelo geológico é uma representação dos diferentes litotipos dispostos no espaço, e nem sempre podem ser correlacionados com suas devidas propriedades. Para incorporar os valores das propriedades nas interpretações geológicas

podemos fazer um "preenchimento" das seções a partir de valores médios obtidos nelas. Este processo segue o princípio da generalização tomando o valor médio de uma seção interpretada como valor médio dos testemunhos contidos dentro daquela seção, ou daqueles bem próximos a seção interpretada. Observe a figura 6.5, temos 6 compostas realizadas em furos $\{F_1, F_2, F_3, F_4, F_5, F_6\}$ correspondendo a área cinza da seção interpretada. Os valores de teores são respectivamente $\{t_1, t_2, t_3, t_4, t_5, t_6\}$ de comprimentos $\{l_1, l_2, l_3, l_4, l_5, l_6\}$

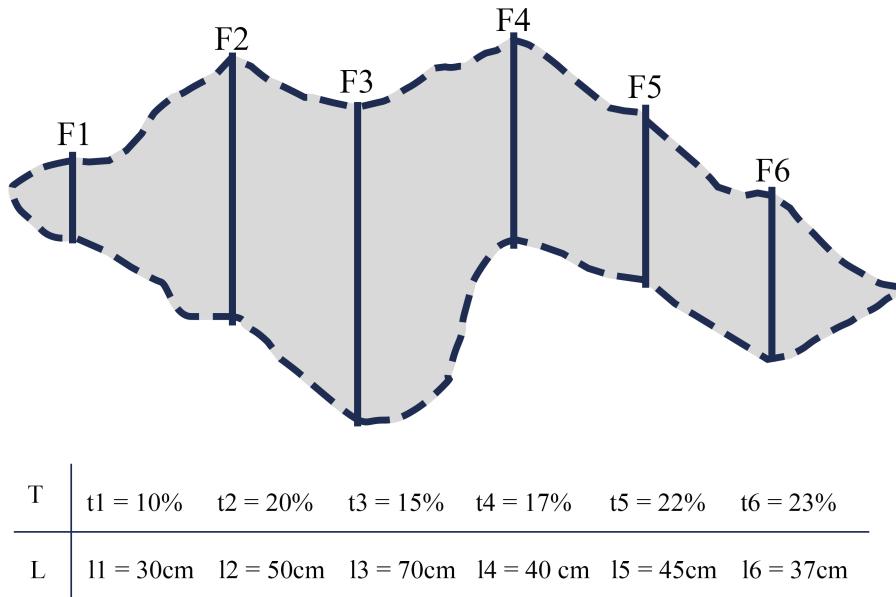


Figura 6.5: Composição realizada em testemunho com 4 fragmentos $\{l_1, l_2, l_3, l_4\}$ e tamanhos de composta igual $\{C_1, C_2\}$. $\{x_1, x_2\}$ representam respectivamente as proporções dos fragmentos $\{l_1, l_2\}$ dentro da composta C_1 .

Para obtermos o valor médio da seção podemos realizar a seguinte operação de composição (6.5)

$$t_C = \frac{\sum_{i=1}^6 t_i l_i}{\sum_{i=1}^6 l_i} = \frac{10 * 30 + 20 * 50 + 15 * 70 + 17 * 40 + 22 * 45 + 23 * 37}{30 + 50 + 70 + 40 + 45 + 37} = 17.91\% \quad (6.5)$$

6.4 Determinação de volumes

Um dos valores mais importantes obtidos na produção mineral é o volume do material a ser extraído, esse processo também é chamado de **cubagem**. A mineração consiste em movimentar um grande volume de rochas de diferentes litotipos. Infelizmente as informações obtidas das rochas são descritas apenas por amostras de pequeno volume e extensão, como em testemunhos de sondagem, durante as primeiras etapas da pesquisa mineral. Em alguns casos é possível obter trincheiras ou verificar as estruturas geológicas em painéis de mina subterrânea. Estes volumes estimados são geralmente obtidos pela interpretação geológica de um profissional engenheiro de minas ou geólogo, capaz de entender os processos de gênese destes depósitos minerais. Uma das alternativas consiste em realizar seções verticais contendo informações dos furos de sondagem e realizar a extrapolação de volumes a partir da extensão das áreas interpretadas. Dado uma série de seções S de área A , o volume entre seções pode ser obtido a partir da equação (6.6)

$$V = \sum_{i=1}^{n-1} D_i(A_i + A_{i+1})/2 \quad (6.6)$$

Onde D_i é a distância entre as seções A_i e A_{i+1} para n seções consideradas. Os volumes obtidos pelas seções das pontas é obtido a partir de extrapolação dado uma distância considerada aceitável pelo geólogo, baseando-se em critérios de continuidade. A figura 6.6 apresenta a interpolação realizada a partir da interpretação geológica das seções e extrapolação destas para obtenção do volume.

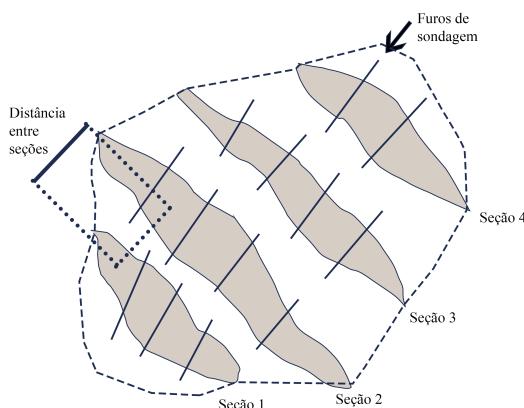


Figura 6.6: Obtenção dos volumes dos corpos geológicos a partir de extrapolação de seções verticais interpretadas. Cinco seções consideradas e os respectivos furos de sondagens utilizados para interpretar os volumes.

Proposição 6.4.1 Apesar de ser um método antigo de avaliação de volumes, as interpretações de seções ainda são utilizadas na prática. Apesar da dificuldade operacional deste método com a digitalização de seção por seção, permite a geólogos e engenheiros de minas incorporarem informações importantes para o modelo, como adição de estruturas geológicas, possíveis zonas de alteração, entre outras características típicas do depósito analisado. Métodos recentes como a **modelagem implícita** auxiliam muito na velocidade de produção dos modelos, mas geralmente precisam de um cuidado maior ao incorporar outras informações após serem gerados

6.5 Inverso do quadrado da distância - IQD

Um dos interpoladores mais antigos conhecidos é o uso do inverso do quadrado da distância. A justificativa da utilização deste ponderador é que muitos problemas físicos ocorrem a partir de leis de decaimento quadráticas. O uso de outras potências também pode ser utilizado, mas a medida que esta potência cresce, os valores mais próximos das regiões estimadas tendem a ser mais valorizados, reduzindo a suavização da interpolação. Dada uma propriedade t como o teor de uma amostra P , situada a uma distância d do centroide de uma célula estimada, a média estimada desta célula pode ser calculada pela equação (6.7):

$$t_m = \frac{\sum_{i=1}^n \frac{1}{d_i^p} t_i}{\sum_{i=1}^n \frac{1}{d_i^p}} \quad (6.7)$$

Em que p é o grau do polinômio considerado para n amostras situadas nas redondezas da célula estimada. Se o número de amostras é grande, o cálculo dos ponderadores geralmente é realizado apenas com os valores mais próximos, para isto utilizando um algoritmo que filtre segundo a proximidade das amostras da célula estimada. Observe a figura 6.7.

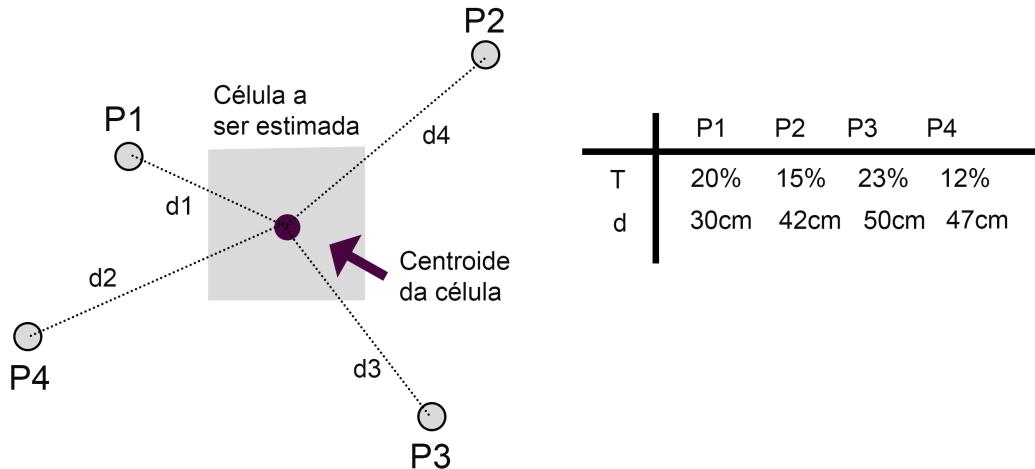


Figura 6.7: Obtenção dos volumes dos corpos geológicos a partir de extração de seções verticais interpretadas. Cinco seções consideradas e os respectivos furos de sondagens utilizados para interpretar os volumes.

Para 4 pontos mais próximos da célula estimada, temos respectivamente os valores de teor (T) e a distância ao centroide da célula. Desta forma podemos calcular o valor médio da célula a partir de (6.8)

$$t_m = \frac{1/(30^2)20 + 1/(42^2)15 + 1/(50^2)23 + 1/(47^2)12}{1/(30^2) + 1/(42^2) + 1/(50^2) + 1/(47^2)} = 17,92\% \quad (6.8)$$

6.6 Tesselação de Delunay

A interpolação espacial pode ser realizada a partir da **tesselação de Delunay**, dividindo o espaço entre as amostras em triângulos. Cada amostra é univocamente ligada a dois pontos mais próximos, sendo esta uma solução geométrica única. Os valores médios de cada triângulo é obtido a partir da média ponderada entre os tamanhos da composta e os valores das propriedades. Dado três pontos $\{P_1, P_2, P_3\}$ com respectivos valores de teor $\{t_1, t_2, t_3\}$ e comprimentos $\{l_1, l_2, l_3\}$. O valor médio de cada triângulo pode ser obtido a partir de (6.9)

$$t_m = \frac{t_1l_1 + t_2l_2 + t_3l_3}{l_1 + l_2 + l_3} \quad (6.9)$$

Observe a figura 6.8. Estão dispostas sete amostras no espaço formando os

triângulos de Delunay, e o triângulo $\{P_1, P_2, P_3\}$ está destacado.

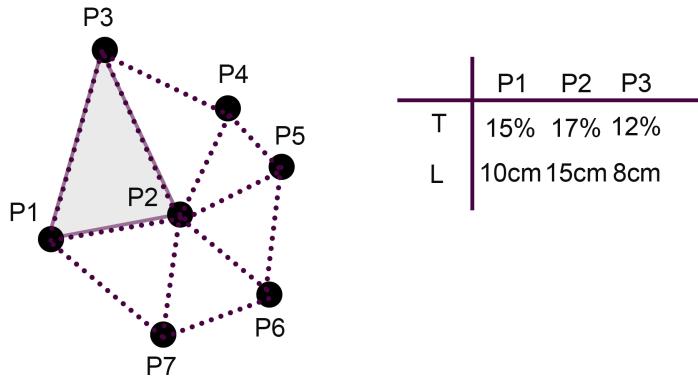


Figura 6.8: Valor médio obtido em um triângulo de Thiessen. 3 pontos amostrais $\{P_1, P_2, P_3\}$, com valores de teor $\{t_1, t_2, t_3\}$.

Considerando $\{t_1, t_2, t_3\}$ os teores relativos a cada amostra, e $\{l_1, l_2, l_3\}$ o tamanho das compostas. Podemos calcular o valor do teor médio como

$$t_m = \frac{15 * 10 + 17 * 15 + 12 * 8}{10 + 15 + 8} = 15,18\% \quad (6.10)$$

6.7 Polígonos de Thiessen

Uma das formas mais tradicionais de avaliação de propriedades georeferenciadas em duas dimensões é a utilização dos chamados polígonos de Thiessen. Cada polígono representa uma área correspondente de influência para uma determinada propriedade. Ao utilizarmos o princípio da generalização, podemos estender o valor de uma propriedade para toda a região considerada por este polígono. A disposição geométrica dos polígonos é única, todos eles são convexos e as relações espaciais destas figuras estão presentes em muitas questões envolvidas na natureza, como por exemplo, a formação de colméias ou de bolhas de sabão.

Proposição 6.7.1 "Em 1911 um climatologista A. H. Thiessen sugeriu um método para representar a precipitação de dados baseados na disposição das estações de tempo. Ele definiu regiões baseadas em uma série de pontos no plano (estações de tempo) em "regiões mais próximas pela linha média entre as estações considerando as estações mais próximas". Baseado em sua proposta o termo polígono de Thiessen tem sido comumente utilizado na geografia definindo os polígonos formados pelo critério de proximidade no plano" -Brassel and Reif [1979]

Para criar um polígono de Tiessen é necessário realizar quatro etapas principais:

1. Determinar o ponto amostral considerado centroide do polígono. Unir aos pontos mais próximos semi-retas ligando o centroide.
2. Determinar os semi-planos formados pela reta perpendicular as semi-retas que ligam o centroide pela metade da distância entre eles.
3. Determinar os vértices do polígono a partir da interseção entre as retas determinadas no item 2.
4. Ligar todos os vértices do polígono. A solução é única e gerará um polígono convexo.

A figura 6.9 exemplifica a formação dos polígonos de Tiessen a partir da configuração geométrica de 6 pontos amostrais $\{P_1, P_2, P_3, P_4, P_5, P_6\}$, sendo P_1 considerado o centroide do polígono. Qualquer propriedade tomada deste ponto amostral é estendida para toda a área do polígono considerado.

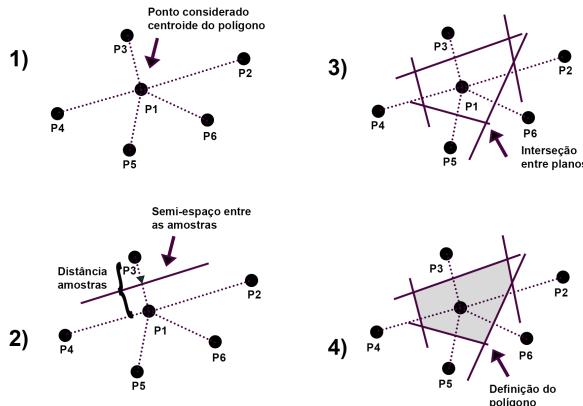


Figura 6.9: 4 etapas para a geração do polígono de Thiessen a partir de um ponto amostral considerado centroide (P_1). 1) Determinar o ponto amostral considerado centroide do polígono. Unir aos pontos mais próximos semi-retas ligando o centroide. 2) Determinar os semi-planos formados pela reta perpendicular as semi-retas que ligam o centroide pela metade da distância entre eles. 3) Determinar os vértices do polígono a partir da interseção entre as retas determinadas no item 2. 4) Ligar todos os vértices do polígono. A solução é única e gerará um polígono convexo.

Proposição 6.7.2 *A solução por polígonos de Thiessen é puramente geométrica. Nenhuma consideração é feita sobre a relação entre as propriedades de uma amostra no espaço. Apenas é realizada a extensão desta propriedade para uma área considerada de influência da amostra.*

Os polígonos de Thiessen possuem algumas propriedade geométricas. O vértice de um polígono de Thiessen é correspondente ao centro geométrico do círculo formado pelo centroide do polígono e dois pontos mais próximos. A figura 6.10 apresenta esta propriedade. Ao unir triângulos entre os pontos mais próximos é criada a chamada **tesselação de Delunay**, em que o baricentro do triângulo é também correspondente ao vértice do polígono de Thiessen.

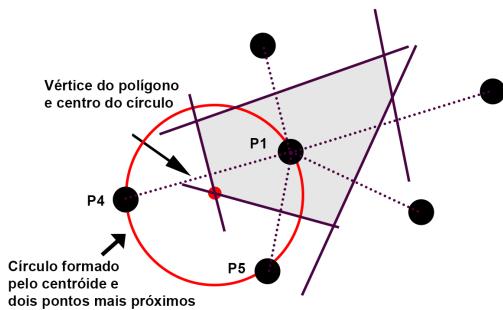


Figura 6.10: Propriedade dos polígonos de Thiessen. Centro do círculo composto pelo baricentro e dois pontos mais próximos forma um vértice do polígono de Thiessen.

É comum para os pontos que se situam nas extremidades do conjunto de dados não possuirem uma solução de polígono fechado, como ocorre nos pontos amostrais interiores. Neste caso geralmente se forma uma extração da área de influência. Esta extração pode considerar quesitos geológicos ou puramente geométricos, mas é um critério muito mais subjetivo que indicativo. A figura 6.12 apresenta esta extração.

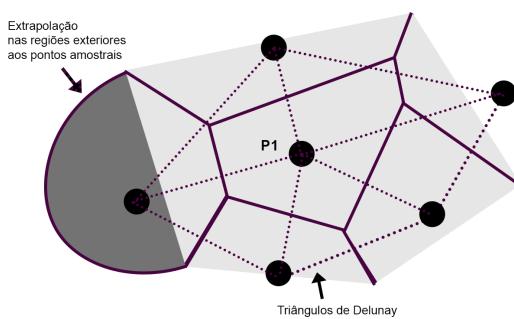


Figura 6.11: Extração realizada nos pontos amostrais situados na extremidade. Extração apresentada pelo um arco de cor cinza escuro.

Para obter o teor médio do depósito a partir dos polígonos de Thiessen basta

considerar a média ponderada entre as áreas de influência, tal como na equação

$$t_m = \frac{\sum_{i=1}^n A_i t_i}{\sum_{i=1}^n A_i} \quad (6.11)$$

Em que A representa a área do polígono de Thiessen para cada ponto amostral i e t representa o teor da área representado pelo valor da amostra em seu centroide. A solução tridimensional para os polígonos de Delunay é os chamados poliedros de Delunay. A solução tridimensional é complicada, e geralmente envolve elementos de topologia de alto nível computacional e matemático. A simplificação utilizada é utilizar o princípio do vizinho mais próximo, dividindo o espaço em uma malha de tamanho infinitesimal e atribuindo em cada célula a propriedade do ponto amostral mais próximo. Quando o tamanho da célula tende a um valor infinitesimal a solução pelo vizinho mais próximo converge para os polígonos ou poliedros de Thiessen. A figura 6.12 representa um mapa dos polígonos de Thiessen a partir do método do vizinho mais próximo e uma malha de tamanho unitário.

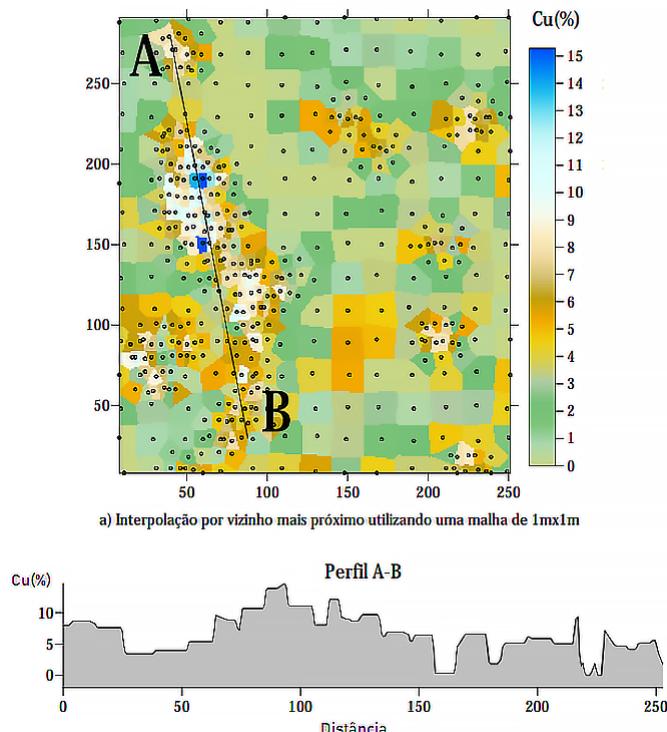


Figura 6.12: Aproximação dos polígonos de Thiessen a partir do vizinho mais próximo, utilizando uma malha de 1mx1m

6.8 Estatísticas desagrupadas

As malhas de amostragem representam uma importante fonte de informação para a realização dos métodos geoestatísticos. O posicionamento das amostras caracteriza a informação espacial a ser representada pelos métodos de estimativa. Quando pensamos em termos de representatividade a amostragem regular no espaço é a que melhor representa as características do depósito mineral. Em alguns casos é possível distribuir a malha segundo a continuidade espacial do depósito, permitindo um maior afastamento nas regiões mais contínuas do depósito e menos espaçadas nas regiões menos contínuas. A figura 6.13[A] apresenta esquematicamente uma malha regular e irregular. Em 6.13[B] é apresentado a disposição da malha segundo a continuidade do corpo geológico.

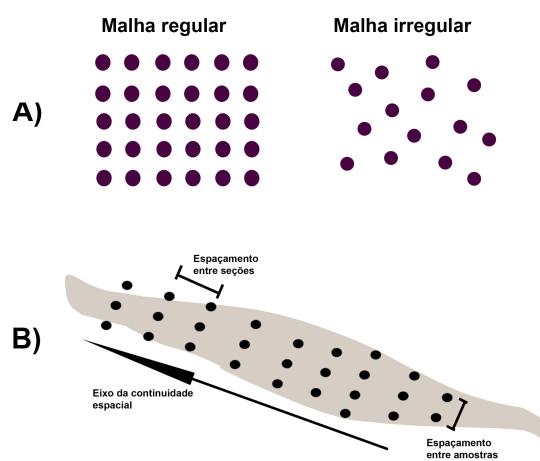


Figura 6.13: A) Representação de uma malha de amostragem regular e uma malha de amostragem irregular B) disposição de amostras segundo a continuidade espacial do corpo geológico.

Nos problemas de engenharia geológica, a disposição das malhas de sondagem dependem de diversos fatores, como por exemplo, terrenos de maior declividade que impedem a utilização de sondas, áreas de proteção ambiental, regiões de córregos ou outros fatores que podem impedir a formação de uma malha regular. Além disso, pela natureza de risco da atividade econômica, é comum adensar amostragens em lugares específicos onde possuam maior interesse para a mineração, como teores metálicos mais altos, ou litotipos de maior importância. Isto forma um agrupamento ou *cluster*. A figura 6.14 apresenta esquematicamente a formação de um agrupamento nas amostras.

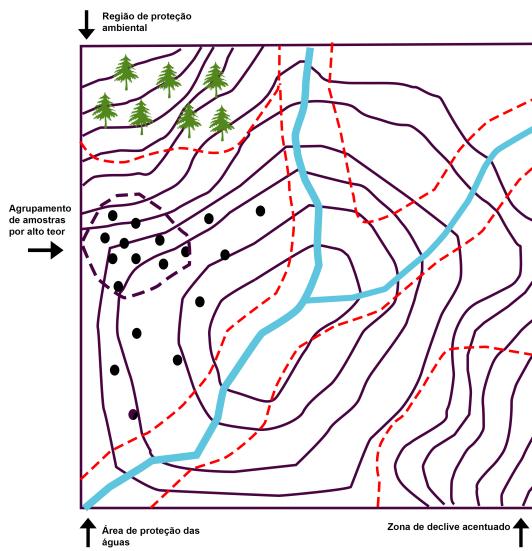


Figura 6.14: Representação de uma área amostrada. Obstáculos para a amostragem representados pela presença de áreas de preservação, terreno com maiores inclinações e área de reservas hídricas. Amostragem irregular realizada a oeste do desenho do mapa.

Calcular estatísticas considerando apenas os dados sem sua disposição espacial pode resultar em enviesamento. Se muitas análises são realizadas apenas em locais onde ocorre alto teor, os resumos estatísticos produziram também resultados com alto valor, mesmo que eles não correspondam à representação do domínio de estimativa.

R "É natural que os dados georeferenciados coletados são de uma forma não representativos. Amostragens preferenciais em áreas de interesse são intencionais e facilitadas pela intuição geológica, dados análogos e amostragens anteriores. A prática de coletar amostras agrupadas ou espacialmente enviesadas é encorajada pelas restrições técnicas e econômicas, tal como produções futuras, acessibilidade e custos de análise dos laboratórios" -[Pyrcz and Deutsch \[2003\]](#)

Um dos maiores erros cometidos por iniciantes ao considerar o desagrupamento de amostras é substituir os valores das amostras pelos valores dos pesos de desagrupamento. A alteração realizada pelo desagrupamento deve ser feita apenas sobre as estatísticas e não sobre seu valor bruto.

Definição 6.8.1 — Desagrupamento. *Dada uma estatística $\phi(Z)$ a partir de uma variável aleatória Z , uma estatística desagrupada θ é aquela que pode ser aplicada de tal forma que $\theta(\phi(Z))$ considerando as distâncias euclidianas relativas entre as amostras.*

As duas principais técnicas utilizadas para desagrupamento são os polígonos de influência, ou de Thiessen vistos anteriormente e o desagrupamento por células.

6.8.1 Polígonos de influência

O desagrupamento das amostras pode ser realizado a partir de áreas de influência como no caso dos polígonos de Thiessen. A frequência de cada valor pode ser alterada pela área do polígono respectivamente. Observe a figura 6.15. Cada ponto amostral $\{P_1, P_2, P_3, P_4, P_5, P_6\}$ possui uma área gerada pelo vizinho mais próximo em um grid de tamanho de célula conhecida. Abaixo podemos ver um histograma representando a frequência destes pontos. Se considerarmos apenas seus valores brutos, e não sua disposição espacial, cada ponto assume um valor de frequência igual a 1. No caso da utilização das áreas pelos polígonos

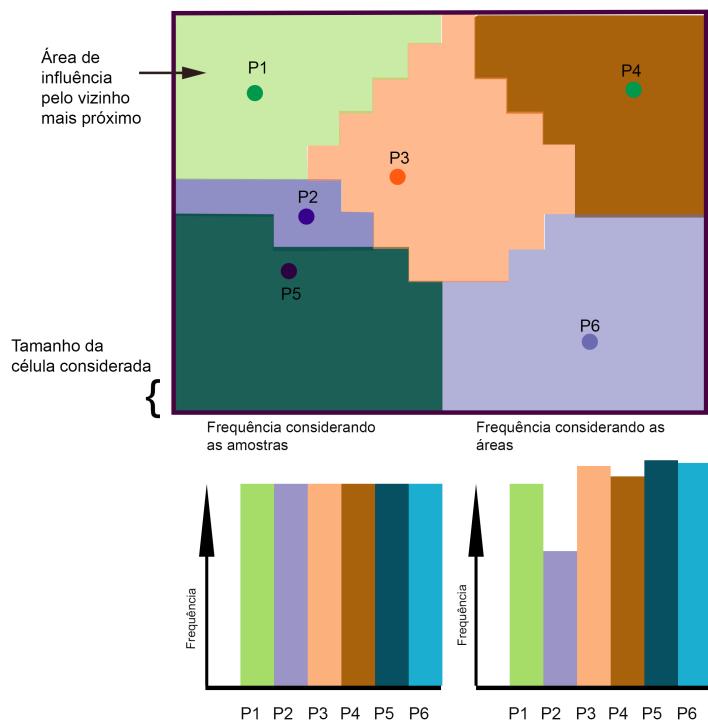


Figura 6.15: Representação de uma área amostrada. Obstáculos para a amostragem representados pela presença de áreas de preservação, terreno com maiores inclinações e área de reservas hídricas. Amostragem irregular realizada a oeste do desenho do mapa.

Definição 6.8.2 — Desagrupamento por polígonos de influência. *Dado uma amostra Z com uma realização z , $F(Z = z)$ representa a frequência de um elemento da amostra. Logo $F(Z = z) = A(z)$, sendo $A(z)$ a área de influência de um elemento z da amostra.*

6.8.2 Desagrupamento por células

O desagrupamento realizado por polígonos de influência, gera uma solução única, e não permite encontrar pesos diferentes para as amostras, no entanto, o método de desagrupamento por células é flexível, permitindo ajustar parâmetros que indicarão o melhor resultado. O método considera a divisão do espaço em 'células' de mesma dimensão, tal que o peso de cada amostra é dado pelo número de amostras contidas dentro de cada célula. Observe a figura 6.16. A célula da linha 1 e coluna 1 apresenta apenas duas amostras, o que significa que cada uma receberá um peso de $1/2$.

	Coluna 1	Coluna 2	Coluna 3	Coluna 4	Coluna 5	Coluna 6
Linha 1	P1 P2	P3	P4 P5	P6		
Linha 2			P7 P8	P9		
Linha 3			P10	P11 P12	P13	
Linha 4				P14	P15	P16 P17

Figura 6.16: Representação do desagrupamento das células em um espaço bidimensional.

Evidentemente o tamanho da célula definirá o peso do desagrupamento. Uma célula muito grande que ocupe toda a extensão territorial analisada terá peso idêntico a $1/n$, sendo n o número de amostras. Logo o ponderador das amostras será igual a equação 6.12

$$p_{t_i} = (1/n) / \left(\sum_{i=1}^n 1/n \right) = 1/n \quad (6.12)$$

Exatamente igual a média aritmética dos valores. Da mesma forma se forem escolhidos tamanhos de células tão pequenas que apenas uma amostra esteja contida, teremos um valor de peso igual a 1, também obtendo o valor da média aritmética. A escolha do tamanho da célula deve ser feita entre estes dois casos extremos, aos quais teremos o menor valor desagrupado da média. A figura 6.17 demonstra a procura do tamanho da célula quadrada mais próxima do menor valor da média desagrupada, definindo assim o resultado que pretendemos.