NAB
Núcleo de Estudos em
Biomassa e Gerenciamento de Água

Elementos Básicos da Quimiometria

-Núcleo de Estudos em Biomassa e Gerenciamento de Água (NAB)-

Dezembro/ 2009
Versão 01

Elaborado	Aprovado	Data
Gerente da Qualidade	Coordenador do NAB	11/12/2009

Sumário

SUMÁRIO 2
1. INTRODUÇÃO 5
2. CONCEITOS BÁSICOS 6
3. SOFTWARE ESTATÍSTICO 7
3.1. MINITAB 8
3.2. SAS 8
3.3. SPSS 8
3.4. STATISTICA DA STATSOFT 8
4. MÉTODOS MULTIVARIADOS 9
4.1. ANÁLISE DE AGRUPAMENTO HIERÁRQUICO (HCA) 9
4.1.1. UNIFICAÇÃO OU AGRUPAMENTO EM ÁRVORE 11
4.1.2. AGRUPAMENTO POR K-MÉDIAS 14
4.2. ANÁLISE DE AGRUPAMENTO (CLUSTER ANALYSIS) 19
4.2.1. PADRONIZAÇÃO DOS DADOS 20
4.2.2. ESCOLHA DO COEFICIENTE DE SEMELHANÇA 20
4.2.3. COEFICIENTES DE SEMELHANÇA PARA VARIÁVEIS QUANTITATIVAS 21
4.2.4. ESCOLHA DA ESTRATÉGIA DE AGRUPAMENTO 22
4.2.4.1. LIGAÇÃO SIMPLES (SINGLE LINKAGE) 22
4.2.4.2. UNWEIGHTED PAIR- GROUP METHOD USING ARITHMETIC AVERAGES (UPGMA) 22
4.2.4.3. LIGAÇÃO COMPLETA (COMPLETE LINKAGE) 23
4.2.4.4. MÉTODO DE WARD'S 23
4.2.5. APLICAÇÃO DA METODOLOGIA 23
4.2.6. MATRIZ FENÉTICA DE SEMELHANÇA (F) 24
4.3. ANÁLISE DE COMPONENTES PRINCIPAIS (PCA) 28
4.3.1. METODOLOGIA – CÁLCULO DOS COMPONENTES PRINCIPAIS 29
4.3.2. CÁLCULO DO PRIMEIRO COMPONENTE PRINCIPAL – (EIXO X) 30
4.3.3. CÁLCULO DO SEGUNDO COMPONENTE PRINCIPAL – (EIXO Y) 31
4.3.4. VARIÂNCIA CONTIDA EM CADA COMPONENTE PRINCIPAL 32
4.3.5. CORREÇÃO DE CADA VARIÁVEL COM O COMPONENTE PRINCIPAL 33

4.3.6. CONSTRUÇÃO GRÁFICA BIDIMENSIONAL (CP1 X CP2) 34
4.3.7. RESULTADOS 37
4.4. ANÁLISE DE CORRESPONDÊNCIA 39
4.5. ANÁLISE DISCRIMINANTE 39
4.5.1. DISTÂNCIA GENERALIZADA DE MAHALANOBIS 40
4.6. ESCALONAMENTO MULTIDIMENSIONAL 41
4.6.1. DIAGRAMA DE SHEPARD 41
5. PLANEJAMENTO DE EXPERIMENTOS 42
5.1. OBJETIVOS 42
5.2. APLICAÇÕES 42
5.3. GLOSSÁRIO 43
5.4. PRINCÍPIOS BÁSICOS 43
5.5. ETAPAS PARA O DESENVOLVIMENTO DE UM PLANEJAMENTO DE EXPERIMENTOS 44
5.6. PLANEJAMENTO INICIAL 44
5.7. PROJETO E ANÁLISE DE EXPERIMENTOS 46
5.8. ESTRATÉGIAS NO PLANEJAMENTO DE EXPERIMENTOS 46
5.9. ROTEIRO PARA A REALIZAÇÃO DE UM EXPERIMENTO 46
5.9. ESTUDO EXPERIMENTAL 47
5.10. EXEMPLOS DO TIPO 2^K 47
5.11. PROJETO DE EXPERIMENTOS FATORIAL DO TIPO 2^K 47
6. TIPOS DE PLANEJAMENTO 48
6.1. PLANEJAMENTO COM UM ÚNICO FATOR 48
6.1.1. ANOVA PARA O MODELO DE EFEITOS FIXOS 50
6.1.2. ONE-WAY ANOVA 53
6.1.3. ABORDAGEM ANOVA 54
6.1.3. VERIFICAÇÃO DA ADEQUAÇÃO DO MODELO 56
6.2. PLANEJAMENTO ALEATÓRIO COM BLOCOS COMPLETOS 58
6.3. PLANEJAMENTO COM BLOCOS INCOMPLETOS 61
6.4. PLANEJAMENTO USANDO O QUADRADO LATINO 63
6.5. PLANEJAMENTO FATORIAL 66
6.5.1. FATORIAL 2^2 66
6.5.2. FATORIAL 2^3 69
6.5.3.1 FATORIAL 2^K COM 1 REPETIÇÃO 74

6.5.3.2. ADIÇÃO DE PONTOS CENTRAIS DO PLANEJAMENTO FATORIAL 2^K 78
6.6. EXPERIMENTOS (RESUMO) 80
6.6.1. PROJETO E ANÁLISE DE EXPERIMENTOS 80
6.6.2. ESTRATÉGIAS NO PLANEJAMENTO DE EXPERIMENTOS 81
6.6.3. ROTEIRO PARA A REALIZAÇÃO DE UM EXPERIMENTO 81
6.6.4. ESTUDO EXPERIMENTAL 81
6.7. EXEMPLOS 82
7. RESUMO 83
8. REFERÊNCIAS BIBLIOGRÁFICAS 87
9. SUMÁRIO DE REVISÕES 88

1. Introdução

A quimiometria é uma área que se refere à aplicação de métodos estatísticos e matemáticos, assim como aqueles baseados em lógica matemática, a problemas de origem química.

Com a sofisticação crescente das técnicas instrumentais, impulsionada pela invasão de microprocessadores e microcomputadores no laboratório químico, tornaram-se necessários tratamentos de dados mais complexos do ponto de vista matemático e estatístico, a fim de relacionar os sinais obtidos (intensidades por exemplo) com os resultados desejados (concentrações).

As análises quantitativas que eram realizadas na maioria das vezes por "via úmida" como titulação, precipitação e reações específicas, que são demoradas e muitas vezes pouco precisas, estão cada vez mais sendo substituídas por técnicas instrumentais como: Ressonância Magnética Nuclear, Espectroscopia no Infravermelho, Espectroscopia no visível/ultravioleta, Espectrometria de Massas, Cromatografia, Polarografia, Análise por Injeção em Fluxo,ICP-MS,IEOS, etc., que aliam a velocidade de análise com uma boa qualidade de resultados. Nessas técnicas instrumentais não é obtida uma informação direta do resultado, mas sim uma grande quantidade de sinais (curvas, picos) que podem ser tratados para uma possível quantificação das várias espécies presentes.

A associação de técnicas instrumentais e métodos quimiométricos permitem que os resultados analíticos sejam obtidos de forma sistemática e com confiabilidade estatística, permitindo reduzir o custo e o tempo dos experimentos, além de diminuir a emissão de resíduos. Através desta associação podem-se monitorar propriedades críticas durante o processo de fabricação, de modo a minimizar o risco sobre perdas de lotes, otimizando, assim, as operações de produção, propiciando às empresas aumentar sua competitividade e, também, acelerar o desenvolvimento e lançamento de novos produtos no mercado.

Muita ênfase tem sido dada aos sistemas multivariados, nos quais ao se analisar uma amostra qualquer, é possível medir muitas variáveis simultaneamente. Nesses sistemas, a conversão da resposta instrumental no dado químico de interesse, requer a utilização de técnicas de estatística multivariada, álgebra matricial e análise numérica. Essas técnicas se constituem no momento na melhor alternativa para a interpretação de dados e para a aquisição do máximo de informação sobre o sistema.

De todos os ramos da química clássica, talvez a química analítica tenha sido a mais afetada pelo desenvolvimento recente da instrumentação química associada a computadores. De fato, a *"Chemometrics Society"*, organização internacional dedicada ao uso e desenvolvimento de métodos em quimiometria, é composta principalmente por químicos interessados em problemas analíticos. Atualmente, é muito raro se encontrar qualquer periódico respeitável sobre pesquisas em química analítica, que não traga artigos reportando dados obtidos com o auxílio de microcomputadores, ou tratados por matemática multivariada ou métodos estatísticos, sempre com o objetivo de melhorar a qualidade dos resultados ou facilitar a sua interpretação.

Os dados analíticos são normalmente obtidos com o objetivo de caracterizar objetos (doenças,amostras,indivíduos,plantas, solos,etc.). Esta caracterização é relativamente

simples, quando o número de variáveis é pequeno. Atualmente, através da tecnologia dos computadores, a quantidade de informações que podem ser tratadas e armazenadas é muito grande, complexa e variada. Na posse dessas informações, a questão que surge naturalmente é como interpretá-las e obedecendo a natureza multivariada, como extrair informação relevante?

Nós, seres humanos, possuímos capacidades únicas de reconhecer padrões. Este processo é complexo, levando cientistas a várias tentativas de exploração do seu mecanismo e ao desenvolvimento de metodologias matemáticas como as redes neurais ou a inteligência artificial, que permitam reconhecer padrões num conjunto de dados. A capacidade humana de identificação por reconhecimento visual vai até a terceira dimensão, devendo portanto, para fazer uso da nossa capacidade de padronizar, representar os dados em três ou duas dimensões. A questão fundamental reside em transformar informação m-dimensional em tri ou bidimensional. Isto pode ser feito através de várias técnicas entre as quais se incluem Análise de Agrupamento, Análise de Componentes Principais, Escalonamento Multidimensional, Análise de Correspondência, Análise Discriminante, etc.

Este documento é de caráter meramente orientativo, para maiores esclarecimentos, consultar a bibliografia em referência.

2. Conceitos Básicos

Tratamento

Um tratamento é uma condição imposta ou objeto que se deseja medir ou avaliar em um experimento. Normalmente, em um experimento, é utilizado mais de um tratamento. Ex.: equipamentos de diferentes marcas, diferentes tamanhos de peças, doses de um nutriente em um meio de cultura, quantidade de lubrificante em uma máquina, temperatura de armazenamento de um alimento.

Unidade Experimental ou parcela

Unidade experimental ou parcela é onde é feita a aplicação do tratamento. É a unidade experimental que fornece os dados para serem avaliados.

Repetição

É o número de vezes que um tratamento aparece no experimento. O número de repetições em um experimento vai depender também dos recursos disponíveis, do tipo de experimento (delineamento), da variabilidade do experimento ou da variável resposta.

Variável resposta ou variável dependente

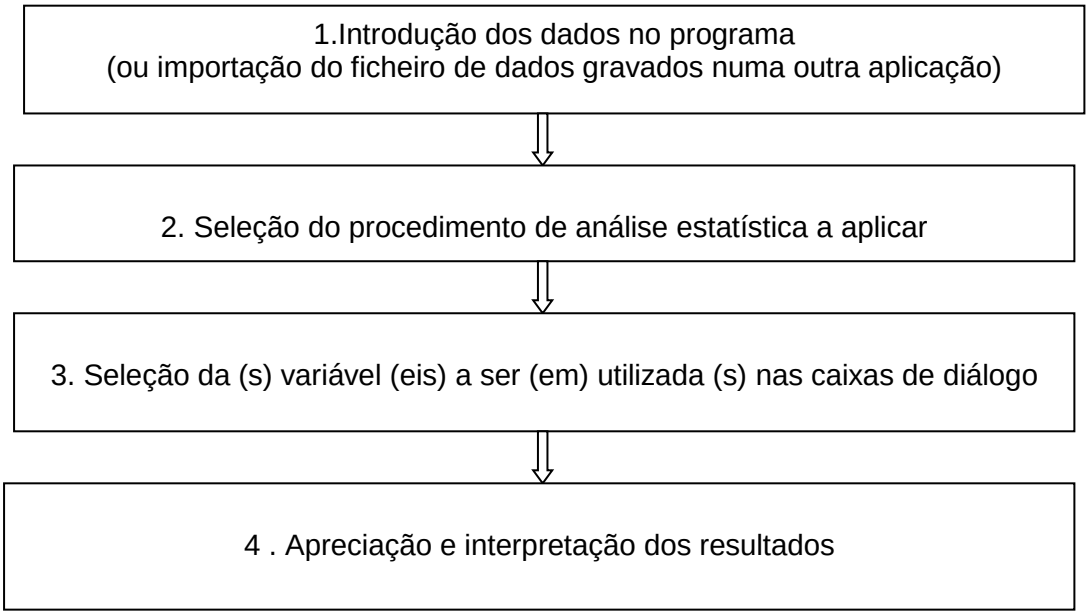
Uma variável qualquer característica que apresenta variação. Ex.: altura das pessoas, comprimento de uma peça, teor de enxofre em óleo diesel.

Delineamento Experimental (Design)

É a forma como os tratamentos ou níveis de um fator são designados as unidades ou parcelas. A análise de variância é baseada no delineamento experimental utilizado.

3. Software Estatístico

Existem vários softwares estatísticos no mercado para a análise estatística de dados. Deste modo, pretende-se realizar uma apresentação sucinta de alguns desses programas. A utilização desses softwares deve se sempre suportada por um adequado conhecimento das técnicas estatísticas envolvidas, ou orientadas por quem detenha esses conhecimentos. De uma forma genérica e simplificada, todos os aplicativos estatísticos, lidam com a análise estatística de dados estruturada em quatros etapas:



Hoje em dia, o software estatístico adquiriu uma grande importância nos meios acadêmicos, empresarial e administrativo, entre outros, quer pela sua facilidade de utilização, quer pela eficácia no tratamento de grandes conjuntos de dados.

O objetivo fundamental é a apresentação simplificada dos diferentes aplicativos, com forte incidência na interface gráfico desses programas.

3.1. Minitab

O Minitab é um software estatístico de capacidade intuitivo, permitindo se obter estatísticas descritivas, simulações e distribuições, inferência estatística elementar, análise da variância, regressão, análise de dados categóricos, métodos não paramétricos, análise de séries temporais, etc.

3.2. SAS

O SAS é um conjunto integrado de aplicativos informáticos com capacidade para o tratamento de grandes volumes de dados. A funcionalidade do sistema é construída em volta de quatro aspectos: acesso, gestão, análise e apresentação de dados.

O núcleo de todo o sistema SAS é o SAS base, o qual permite criar tabelas e proceder à manipulação de dados.

3.3. SPSS

O SPSS (Statistical Package for the Social Sciences) é um software modular, integrando todas as etapas do processo analítico desde o planejamento, acesso e gestão de dados. Sendo uma solução modular, permite-lhe adicionar novas funcionalidades e integrar outros produtos de software autônomos, apresentando sempre a mesma estrutura de utilização. A linha dos produtos SPSS pode ser dividida em três categorias:

- SPSS Base System: é o módulo principal for *Windows*, inclui um conjunto alargado de procedimentos para acesso, manipulação, análise e apresentação de dados, todos eles acessíveis a partir de uma interface simples de utilizar.

- Módulos adicionais SPSS.

- Software *stand – alone* (produtos autônomos) integrável com o *SPSS Base System*.

3.4. Statistica da Statsoft

É um aplicativo autônomo que inclui estatísticas descritivas (correlações, testes t e outros testes para as diferenças entre grupos, tabelas de freqüências e cruzamentos), métodos de regressão múltipla, métodos paramétricos, rotinas de ANOVA/ MANOVA, módulos das distribuições e um vasto conjunto de ferramentas para gráficos.

Nota: O Excel é uma poderosa folha de cálculo, que para além de múltiplas funcionalidades, nos permite ainda fazer a análise estatística de dados, através de um conjunto de funções e procedimentos avançados, os quais se encontram sob o comando “Análise de Dados”, aceitável como Suplementos no menu Ferramentas

(*tools*). Através deste comando, podemos aceder a uma vasta gama de procedimentos estatísticos, desde a análise mais simples como a estatística descritiva (tabelas de freqüência, médias, modas, desvio padrão, etc), até análises mais complexas (análise de variância- ANOVA, regressão, etc). Para estudos orientativos quanto a utilização do Excel, verificar a referência bibliográfica (Manual de Estatística utilizando o Excel).

4. Métodos Multivariados

Existem vários métodos multivariados de análise multivariada com finalidades diversas entre si. Portanto, primeiramente é necessário saber que conhecimento se pretende gerar, ou melhor, o que se pode afirmar a respeito dos dados. Para exemplificar esta diversidade, vamos propor alguns objetivos e indicar alguns métodos possíveis. Quando o interesse é verificar como as amostras se relacionam, ou seja, o quanto estas são semelhantes segundo as variáveis utilizadas no trabalho, destacam dois métodos que podem ser utilizados, a análise por grupamento hierárquico (HCA) e a análise por componentes principais (PCA). Quando a finalidade principal é fazer previsão, por exemplo, quando temos muitas variáveis independentes e queremos encontrar uma variável dependente, a regressão linear múltipla e redes neurais são métodos indicados para esta situação. Com a finalidade bem diversa, existem métodos de análise multivariada que podem ser utilizados na etapa inicial de uma pesquisa, na própria escolha das variáveis que descreverão o sistema. Isto é muito comum nos casos em que, um processo necessita de ser otimizado. Dentre os métodos que servem para otimização, citamos o simplex e o planejamento fatorial.

Os métodos estatísticos são escolhidos de acordo com os objetivos da pesquisa, por isto, mostrar, predizer ou otimizar os resultados obtidos por diferentes métodos. Portanto, a estatística multivariada, com os seus diferentes métodos, difere de uma prateleira de supermercado abarrotada de produtos com a mesma função, pois cada método tem sua fundamentação teórica e sua aplicabilidade. .

4.1. Análise de agrupamento Hierárquico (HCA)

O termo Análise de Agrupamentos, primeiramente usado por (Tyron, 1939) na realidade comporta uma variedade de algoritmos de classificação diferentes, todos voltados para uma questão importante em várias áreas da pesquisa: *Como organizar dados observados em estruturas que façam sentido, ou como desenvolver taxonomias capazes de classificar dados observados em diferentes classes*. Esta análise consiste no tratamento estatístico de cada amostra como um ponto no espaço multidimensional descrito pelas variáveis escolhidas. É possível, nesta técnica, tratar cada variável como um ponto no espaço multidimensional descrito de acordo com o interesse em cada situação. Quando uma determinada amostra é tomada como no espaço das variáveis, é possível calcular a distância deste ponto a todos os outros pontos, constituindo-se assim uma matriz que descreve a proximidade entre todas as amostras estudadas.

Biólogos, por exemplo, têm de organizar dados observados em estruturas que "façam sentido", ou seja, desenvolver taxonomias. Zoologistas confrontados com uma variedade de espécies de um determinado tipo, por exemplo, têm de conseguir

classificar os espécimes observados em grupos antes que tenha sido possível a descrição desses animais em detalhes de formas a se destacar detalhadamente as diferenças entre espécies e subespécies.

A idéia aqui é a de um processo ***data-driven***, ou seja, dirigido pelos dados observados de forma a agrupar esses dados segundo características comuns que ocorram neles.

Este processo deve levar em conta a possibilidade de se realizar inclusive uma organização hierárquica de grupos, onde a cada nível de abstração maior, são também maiores as diferenças entre elementos contidos em cada grupo, da mesma forma que espécies animais do mesmo gênero têm muito em comum entre si, mas espécies animais que possuem apenas o filo ou a ordem em comum possuem pouca similaridade.

a) Significância Estatística

Observe que as discussões até o momento não mencionaram a questão da significância estatística ou de seu teste. A Análise de Agrupamentos é na verdade uma coleção de diferentes algoritmos que ***agrupam objetos***. O ponto aqui é que, utilizamos métodos de análise de agrupamentos quando não possuímos nenhuma hipótese *a priori* sobre a estrutura ou comportamento de nossos dados e necessitamos iniciar com alguma coisa. Por que então não deixar um software descobrir quais regularidades são interessantes no conjunto dos dados? Por causa disso, testes de significância estatística ainda não são apropriados nesta altura do campeonato.

b) Áreas de Aplicação

Técnicas de agrupamento têm sido aplicadas em uma enorme gama de áreas. (Hartigan 1975) já provê uma visão geral ampla de vários estudos publicados acerca da utilização de técnicas de Análise de Agrupamentos. Na área médica, por exemplo, agrupamento de doenças por sintomas ou curas pode levar a taxonomias muito úteis. Em áreas da psiquiatria, por exemplo, considera-se que o agrupamento de sintomas como paranóia, esquizofrenia e outros é essencial para a terapia adequada. Na arqueologia, por outro lado, também se tem tentado agrupar civilizações ou épocas de civilizações com base em ferramentas de pedra, objetos funerários, etc. De forma geral, toda vez que se faz necessário que se classifique uma "montanha" de dados desconhecidos em pilhas gerenciáveis, se utiliza métodos de agrupamento.

c) Métodos de Agrupamento

Há dois algoritmos de agrupamento de dados baseados em métodos estatísticos interessantes para efeitos de classificação de padrões. Vamos analisar e discutir cada um dos dois abaixo e vamos, ao final, discutir como utilizar os resultados da aplicação destes métodos para o particionamento de um grupo de dados cujo comportamento intrínseco ainda é desconhecido, na confecção de sistemas de reconhecimento de padrões que sejam capazes de automaticamente classificar novas observações em uma das classes "detectadas" por um destes dois métodos.

Nota: Consultar o Manual de Estatística com relação a retirada de *outliers*, normalidade dos dados e homogeneidade das variâncias, caso seja necessário.

4.1.1. Unificação ou Agrupamento em Árvore

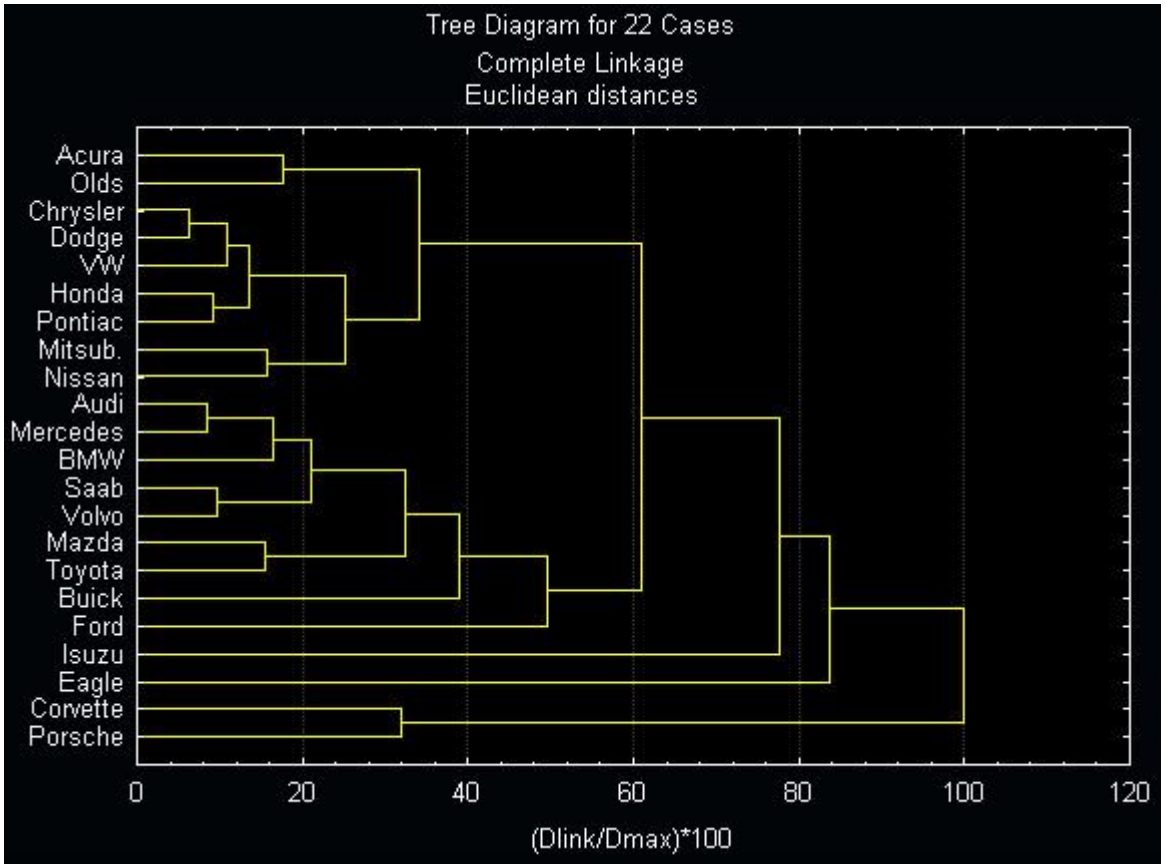
O agrupamento em árvore (*Tree Clustering*) tem por objetivo a construção de taxonomias de vários níveis. Ele é considerado um método de agrupamento aglomerativo hierárquico.

Suponha um conjunto de dados como o abaixo, proveniente do conjunto de exemplos do software Statistica, descrevendo vários carros através de um conjunto de atributos seus. Em quantas classes sensatas podemos dividir este conjunto de carros?

Data: Cars.sta* (5v by 22c)

	Performance, fuel economy, and approximate price for various automobiles				
	1 Preço	2 Aceleração	3 Frenagem	4 Manutenção	5 Kilometragem
Acura	-0,521072363	0,477252671	-0,00657103855	0,381619066	2,07875356
Audi	0,865652474	0,208033216	0,31869537	-0,0913735792	-0,677061608
BMW	0,495859184	-0,801539742	0,192202878	-0,0913735792	-0,153805564
Buick	-0,613520685	1,68874022	0,933087475	-0,20962174	-0,153805564
Corvette	1,23544576	-1,8111127	-0,494470651	0,972859872	-0,677061608
Chrysler	-0,613520685	0,0734234878	0,427117506	-0,20962174	-0,153805564
Dodge	-0,705969008	-0,195795968	0,481328574	0,145122743	-0,153805564
Eagle	-0,613520685	1,21760617	-4,19889364	-0,20962174	-0,677061608
Ford	-0,705969008	-1,54189324	0,987298543	0,145122743	-1,7235737
Honda	-0,42862404	0,409947807	-0,00657103855	0,0268745821	0,369450479
Isuzu	-0,79841733	0,409947807	-0,0607821066	-4,23005922	1,0671252
Mazda	0,126065894	0,679167263	-0,133063531	0,499867227	-1,7235737
Mercedes	1,05054912	0,00611862399	0,119921454	-0,0913735792	-0,153805564
Mitsub.	-0,613520685	-1,00345433	0,0837807415	0,381619066	0,718287842
Nissan	-0,42862404	0,0734234878	-0,00657103855	0,263370905	0,997357732
Olds	-0,613520685	-0,734234878	0,40904715	0,381619066	2,11363729
Pontiac	-0,613520685	0,679167263	0,535539642	0,145122743	0,195031798
Porsche	3,4542055	-2,21494188	-0,295696735	0,618115388	-1,02589897
Saab	0,588307506	0,679167263	0,246413946	0,263370905	0,0206131169
Toyota	-0,0588307506	1,21760617	0,22834359	0,73636355	-0,851480289
VW	-0,705969008	-0,128491104	0,101851098	0,381619066	0,195031798
Volvo	0,218514217	0,611862399	0,13799181	-0,20962174	0,369450479

O objetivo deste algoritmo é o de unificar objetos em classes ou grupos sucessivamente maiores através da utilização de alguma medida de similaridade ou de distância. Um resultado típico deste enfoque é uma árvore hierárquica, como no exemplo do **dendrograma** abaixo:

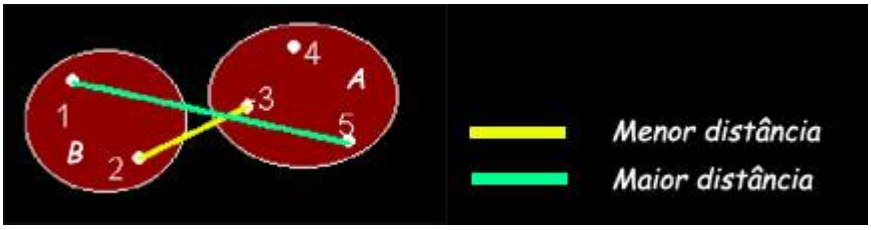


Exemplo de uma árvore hierárquica em uma classificação de tipos de carros de acordo com uma série de características que os descrevem.

Para realizar este gráfico acima, o nome de cada instância (não são classes) foi utilizado como variável na posição da variável dependente, neste caso o modelo do carro. As classes são dadas pelos ramos da árvore, que é construída de trás para frente pelo método: começamos com ramos individuais e vamos juntando ramos de acordo com a distância (nesse caso euclidiana) entre instâncias, de forma a agrupá-las em classes, até gerarmos a raiz da árvore.

Para construir a árvore utilizamos alguma medida de distância entre classes. Chamamos esta distância de **distância de conexão** ou *linkage distance*. Há três filosofias de análise da distância de conexão ao fazer-se a montagem da árvore:

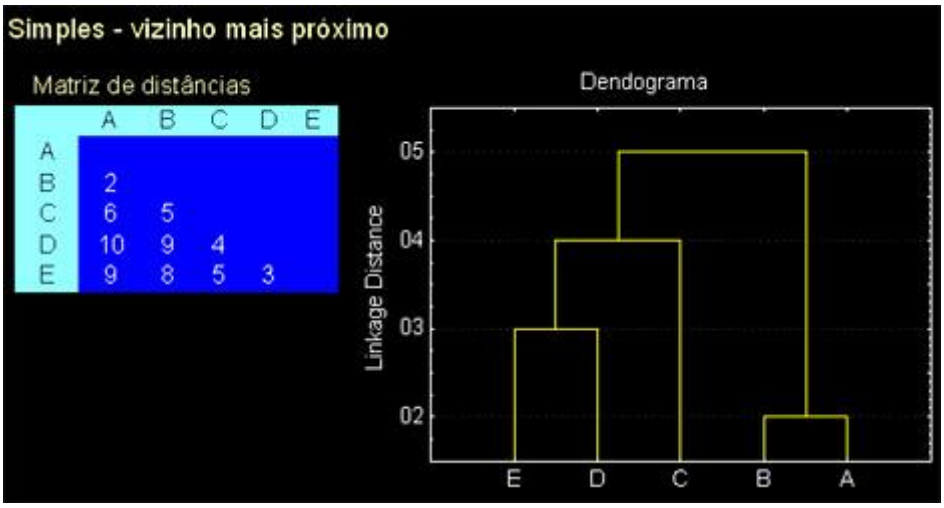
- **Simples** - consideramos a distância entre os vizinhos mais próximos como a distância entre agrupamentos. Neste caso no exemplo abaixo d(A, B) = d(2,3)
- **Completa** - consideramos a distância entre os vizinhos mais distantes como a distância entre agrupamentos. Neste caso no exemplo abaixo d(A, B) = d(1,5)
- **Média** - Consideramos a distância média segundo a fórmula abaixo como a distância entre agrupamentos.

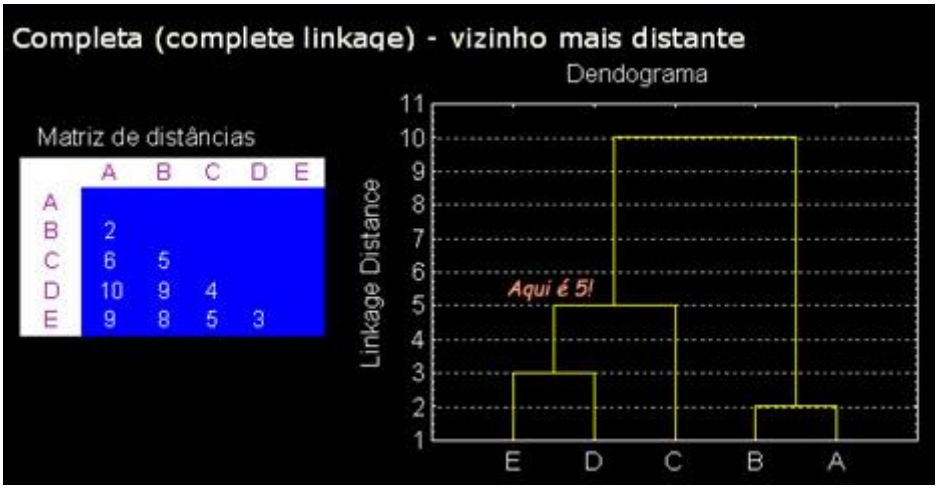


A fórmula para cálculo da distância média $d_{média}$ é dada pela média das distâncias entre todos os pares de pontos:

$$\overline{d} = \frac{d(1,3) + d(1,4) + \ldots + d(2,5)}{6}$$

Para montar a árvore de classificação ou dendograma, procedemos unindo sempre grupos apresentando a menor distância de acordo com uma das três regras acima. Veja os exemplos abaixo para esclarecer suas dúvidas:





Para utilizarmos o método para nos dar um determinado conjunto de classes, percorremos a árvore então a partir da raiz, até termos o número de classes que nos agrada mais. No caso acima, por exemplo, se percorrermos a partir da raiz (Dlink/Dmax = 1.0) em direção às folhas e pararmos em Dlink/Dmax = 0.7, onde teremos 4 classes, dadas cada qual pelo seu ramo correspondente.

Lembre-se que, à medida que você se move para a direita no diagrama de árvore, aumentando as distâncias de conexão, agrupamentos cada vez maiores são formados, com cada vez maiores diversidades intra-agrupamentos. Se este gráfico mostra um platô claro, isto significa que muitos clusters foram formados a

aproximadamente a mesma distância de conexão. Esta distância poderia ser um local de corte para a divisão dos dados em grupos ou classes.

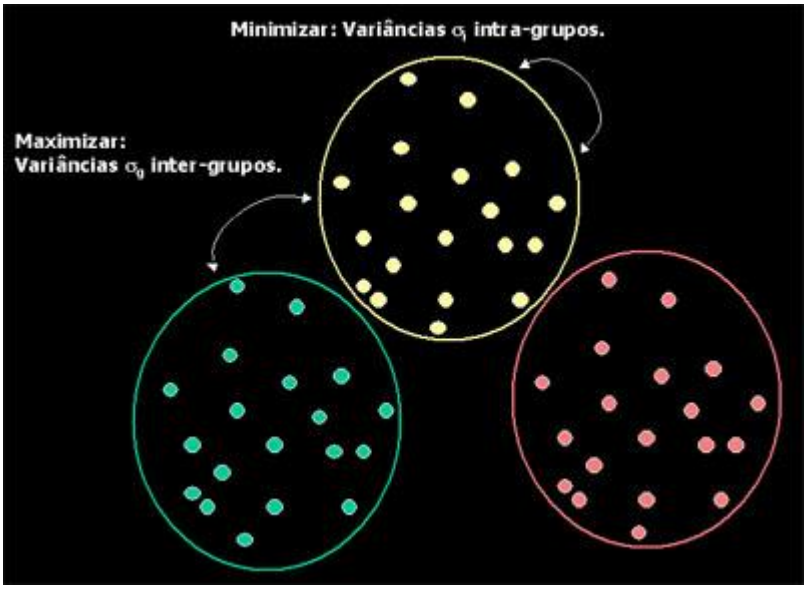
4.1.2. Agrupamento por k-Médias

Este segundo método de Análise de Agrupamentos é um método de **agrupamento não-hierárquico por repartição**. Este método de agrupamento é muito diferente do método de Agrupamento em Árvore . Suponha que você já tem as hipóteses a respeito do número de conjuntos em seus casos ou variáveis. Você quer informar ao computador para formar exatamente 3 conjuntos que devem ser tão distintos quanto o possível. Este é o tipo de pesquisa que pode ser feita pelo algoritmo de aglomeramento por *k*-Médias. O método *k*-Médias produzirá exatamente ***k*** diferentes conjuntos com a maior distinção possível entre eles.

No exemplo dos carros acima, o pesquisador consumidor pode ter um "pressentimento" da experiência de aquisição de carros anteriores ou de análises de mercado que os carros caem basicamente em três categorias diferentes no que diz respeito à relação custo-benefício. Ele pode querer saber se esta intuição pode quantificada, isto é, se a análise de agrupamento por k-Médias das medidas da relação custo-benefício dada pelas variáveis descritoras dos carros produziria certamente os três conjuntos de marcas de carros como esperados. Assim, as médias das diferentes medidas de relação custo-benefício (frenagem, manutenibilidade, etc) para cada conjunto representariam uma maneira quantitativa de expressar a hipótese ou intuição do pesquisador.

Computacionalmente, você pode pensar neste método como a Análise de Variância (ANOVA) "ao contrário" ; O programa começará com os k-conjuntos aleatórios, e moverá então os objetos entre estes conjuntos com o objetivo de: (1) minimizar a variabilidade dentro dos conjuntos e (2) maximizar a variabilidade entre conjuntos. Isto é semelhante ao "ANOVA , mas ao contrario" no sentido que o teste de significância ANOVA avalia a variabilidade entre - grupos de encontro a variabilidade intra-grupo ao calcular o teste de significância para a hipótese em que as médias dos grupos são diferentes para cada grupo. Em k-Médias, as tentativas do programa de mover objetos (por exemplo, casos) dentro e fora dos grupos (conjuntos) para ter resultados ANOVA mais significativos. (porque, entre outros resultados, os resultados ANOVA são a saída padrão da análise de agrupamento).

Para avaliar a precisão da classificação, você pode comparar a variabilidade intra-grupo (que é pequena se a classificação é boa) para a variabilidade intergrupos (que é grande se a classificação é boa). Em outras palavras, você pode fazer uma análise de variâncias padrão entre - grupos para cada dimensão (caso ou variável). Para um dado ***k*** e para cada variável, uma medida de discriminação entre grupos.

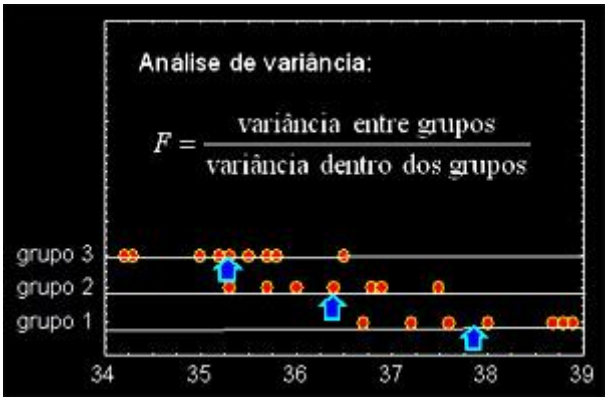


*A figura acima ilustra o processo de Análise de Agrupamentos através do método das k-Médias: O processo iterativo é realizado por combinação de elementos em **k** grupos até que se obtenha uma combinação que maximize o cálulo das variâncias entre grupos e que minimize o cálculo das variâncias intra-grupos.*

a) Avaliando a Qualidade de um Agrupamento Gerado

Um agrupamento gerado por este algoritmo é único e o melhor para um dado ***k,*** e nenhum mínimo local no processo iterativo é do conhecimento deste autor. Porém, nem todo agrupamento gerado é útil para efeitos de classificação, pois um determinado ***k***, por exemplo *k*=5, pode refletir um número de classes inadequado para a divisão de uma determinada população de observações. É necessário encontrar o valor de ***k*** que melhor reflita a "divisão natural" da população de dados observada. Para avaliar se uma classificação é apropriada, pode-se comparar a variância intracluster (que deverá ser pequena se a divisão em classes for adequada) à variância inter-clusters (que deverá ser grande se a classificação em categorias for boa). Isto significa que uma boa divisão de um conjunto de observações em grupos ou categorias é aquela onde os elementos de uma mesma categoria são os mais parecidos entre si (menor variância intra-cluster) e onde os elementos de grupos diferentes são os mais diferentes entre si possíveis (variância inter-cluster ou intergrupos). Isto é dito verificar a robustez dos grupos de objetos ou categorias geradas.

Para tanto, você pode realizar uma análise de variância padrão intergrupos. Para um dado ***k*** e para cada variável, uma medida de discriminação entre grupos pode ser dada pelo cálculo das variâncias inter- e intra-grupos e pelo coeficiente de discriminação **F**, dado abaixo:



Pode-se utilizar ***F*** como medida da qualidade de um determinado conjunto de classes como subdivisão natural de um conjunto de dados, quando não se sabe se para determinado conjunto de dados 2,3 ou 4 classes são o mais adequado.

b) Algoritmo Básico do Método das k- Médias

1. Padronize ou estandardize todos os dados, descrevendo cada variável em termos de distância de seu valor em desvios-padrão da sua média.
2. Fixa-se o número de agrupamentos desejado = ***k***;
3. Divida os casos aleatoriamente nos ***k*** grupos;
4. Calcula-se o centróide de cada grupo;
5. Com os dados padronizados, calcula-se, para cada caso, a distância euclidiana em relação ao centróide de cada grupo;
6. Transfira o caso para o grupo cuja distância ao centróide é mínima;
7. Repita (4), (5) e (6) até que nenhum caso seja mais transferido.

O exemplo do conjunto de dados das flores do gênero Iris utilizado para a Análise de Discriminantes também é um bom exemplo que pode ser utilizado com o algoritmo de k-médias para produzir um conjunto de classes naturais. Ele nos permite uma boa análise das variâncias:

	Análise de variância com 2 agrupamentos				
	SS entre	gl	SS dentro	gl	F
comp. sep.	76,7	1	72,31	148	157,0
larg. sép.	54,2	1	94,76	148	84,7
comp. pét.	126,9	1	22,13	148	848,6
larg. pét.	117,3	1	31,68	148	548,1
	Análise de variância com 3 agrupamentos				
	SS entre	gl	SS dentro	gl	F
comp. sep.	111,5	2	37,50	147	218,6
larg. sép.	77,6	2	71,39	147	79,9
comp. pét.	137,3	2	11,72	147	860,6
larg. pét.	130,7	2	18,28	147	525,7
	Análise de variância com 4 agrupamentos				
	SS entre	gl	SS dentro	gl	F
comp. sep.	115,3	3	33,67	146	166,7
larg. sép.	93,9	3	55,08	146	83,0
comp. pét.	138,6	3	10,42	146	647,1
larg. pét.	134,5	3	14,48	146	452,0

c) Exemplo de Agrupamento por k-Médias

Tome novamente o caso dos modelos de carros dado acima. A meta do algoritmo k-means é encontrar um particionamento ótimo para dividir um número de objetos em k agrupamentos. Este procedimento irá mover objetos de agrupamento para agrupamento com o objetivo de minimizar a variância intra-grupo e maximizar a variancia entre-grupos. Neste exemplo, você pode identificar 3 agrupamentos nos dados sobre carros, usando o metodo de agrupamento em árvore (Joining). Agora você verá que o tipo de solução fornecida pelo agrupamento por k-medias irá sugerir 3 grupos.

Se executarmos este exemplo no software Statistica, setando "casos" como objetos a agrupar e k=3, obtemos o seguinte relatório, após o término do processamento da análise de agrupamentos:

- Number of variables: 5
- Number of cases: 22
- K-means clustering of cases
- Missing data were casewise deleted
- Number of clusters: 3
- Solution was obtained after 3 iterations

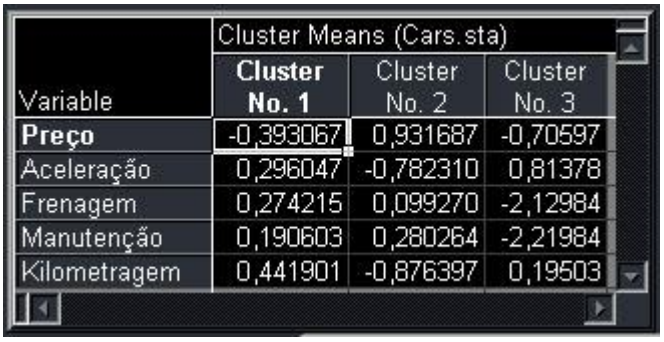
Este processamento nos provê vários conjuntos de valores que vão nos permitir utilizar o resultado para um classificador. A primeira etapa é observar a análise da variância, que para nosso caso é como abaixo:

	Analise da Variância					
Variable	**Between SS**	df	Within SS	df	F	signif. p
Preço	9,08159	2	11,91841	19	7,23881	0,004602
Aceleração	6,74790	2	14,25210	19	4,49794	0,025163
Frenagem	10,11892	2	10,88108	19	8,83457	0,001938
Manutenção	10,87750	2	10,12250	19	10,20857	0,000975
Kilometragem	7,99118	2	13,00882	19	5,83575	0,010573

Nesse caso, a julgar pela magnitude e níveis de significância dos valores de **F**, as variáveis *Manutenção, Frenagem e Preço* representam os critérios mais importantes para a atribuição de objetos a grupos.

d) Utilizando o Resultado da Análise de Agrupamentos para Construir um Classificador de Padrões

Esses valores acima somente servem para nos permitir avaliar a qualidade dos clusters gerados, são pouco interessantes para um classificador. Já as médias dos valores de cada cluster, que representam uma espécie de centróide de cada cluster:

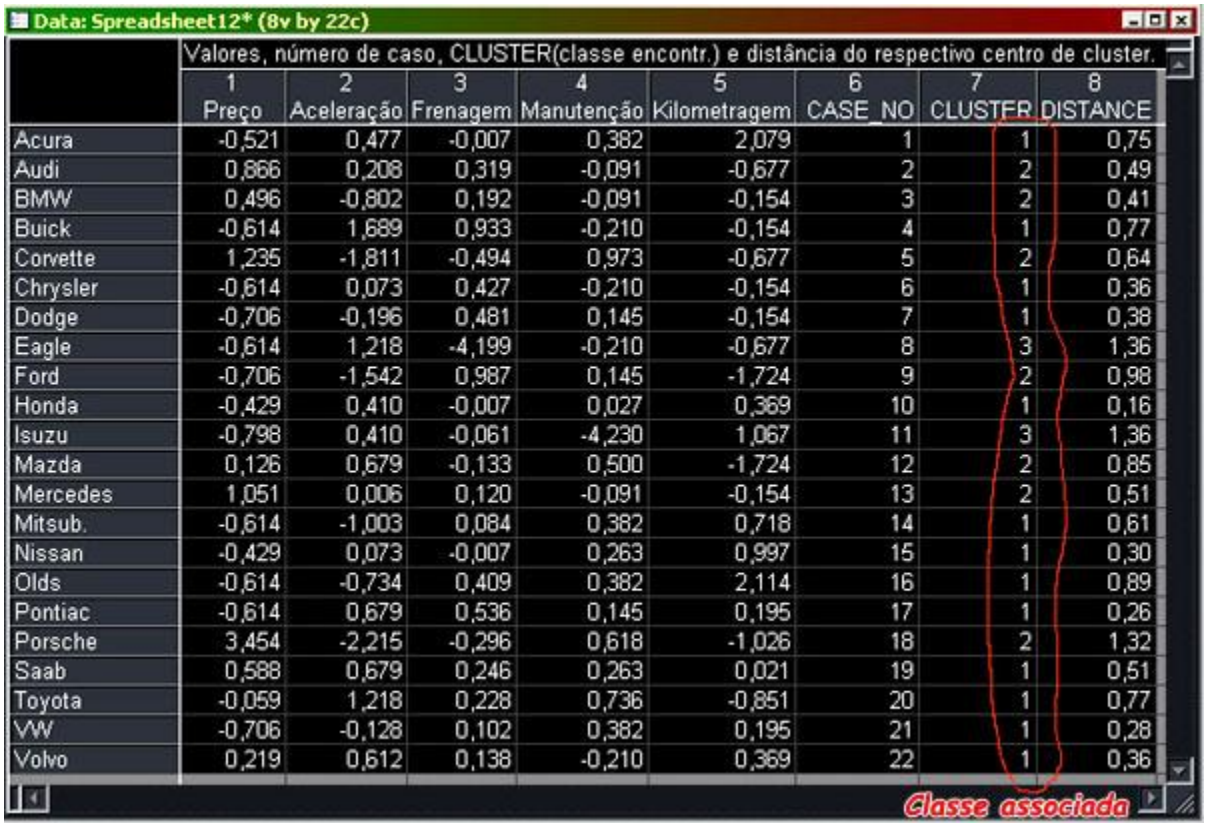


Cluster Means (Cars.sta)

Variable	Cluster No. 1	Cluster No. 2	Cluster No. 3
Preço	-0,393067	0,931687	-0,70597
Aceleração	0,296047	-0,782310	0,81378
Frenagem	0,274215	0,099270	-2,12984
Manutenção	0,190603	0,280264	-2,21984
Kilometragem	0,441901	-0,876397	0,19503

Considerando que cada uma destas 3 médias representa um ponto no R^5, que é o espaço de casos deste conjunto de dados, poderíamos considerar estes valores como uma espécie de protótipos de referência para um classificador. Mas temos também a opção de usar as próprias instâncias de casos e sua respectiva classificação.

Para saber como foram classificados os casos, pode-se observar a tabela de classificação geral de caso por grupo/classe:



Data: Spreadsheet12* (8v by 22c)

Valores, número de caso, CLUSTER(classe encontr.) e distância do respectivo centro de cluster.

	1 Preço	2 Aceleração	3 Frenagem	4 Manutenção	5 Kilometragem	6 CASE_NO	7 CLUSTER	8 DISTANCE
Acura	-0,521	0,477	-0,007	0,382	2,079	1	1	0,75
Audi	0,866	0,208	0,319	-0,091	-0,677	2	2	0,49
BMW	0,496	-0,802	0,192	-0,091	-0,154	3	2	0,41
Buick	-0,614	1,689	0,933	-0,210	-0,154	4	1	0,77
Corvette	1,235	-1,811	-0,494	0,973	-0,677	5	2	0,64
Chrysler	-0,614	0,073	0,427	-0,210	-0,154	6	1	0,36
Dodge	-0,706	-0,196	0,481	0,145	-0,154	7	1	0,38
Eagle	-0,614	1,218	-4,199	-0,210	-0,677	8	3	1,36
Ford	-0,706	-1,542	0,987	0,145	-1,724	9	2	0,98
Honda	-0,429	0,410	-0,007	0,027	0,369	10	1	0,16
Isuzu	-0,798	0,410	-0,061	-4,230	1,067	11	3	1,36
Mazda	0,126	0,679	-0,133	0,500	-1,724	12	2	0,85
Mercedes	1,051	0,006	0,120	-0,091	-0,154	13	2	0,51
Mitsub.	-0,614	-1,003	0,084	0,382	0,718	14	1	0,61
Nissan	-0,429	0,073	-0,007	0,263	0,997	15	1	0,30
Olds	-0,614	-0,734	0,409	0,382	2,114	16	1	0,89
Pontiac	-0,614	0,679	0,536	0,145	0,195	17	1	0,26
Porsche	3,454	-2,215	-0,296	0,618	-1,026	18	2	1,32
Saab	0,588	0,679	0,246	0,263	0,021	19	1	0,51
Toyota	-0,059	1,218	0,228	0,736	-0,851	20	1	0,77
VW	-0,706	-0,128	0,102	0,382	0,195	21	1	0,28
Volvo	0,219	0,612	0,138	-0,210	0,369	22	1	0,36

Classe associada

Para a construção de um classificador de novos casos, é esta tabela que é a mas interessante. Uma opção de uso desses resultados é considerar o cluster associado a cada caso como sendo a classe encontrada para o caso, ou seja, sua classificação e utilizar esta informação em um método de aprendizado supervisionado para criar um classificador para novos casos ou simplesmente utilizar esta lista de casos "classificados". Este método pode ser uma coisa complexa como uma rede neural ou uma coisa simples como k-NearestNeighbour.

e) Exemplo de Análise de Agrupamentos :Implementação de dois principais métodos de análises de Agrupamentos.

Implemente os dois métodos de Análise de Agrupamentos vistos acima em um programa de computador, de tal forma que o usuário possa ler um arquivo de dados organizado em colunas divididas por tabs e escolher entre um dos dois métodos. Os requisitos da implementação de cada um dos dois métodos são dados abaixo:

Requisitos da Implementação do Agrupamento em Árvore

- O programa deverá permitir ao usuário escolher entre Distância de Hamming e Distância Euclideana como métricas de similaridade.
- O programa deverá permitir ao usuário escolher entre as três **distâncias de conexão** ou *linkage distances*.
- O programa deverá mostrar ao usuário o dendrograma da classificação gerada.
- O programa deverá permitir que se exporte o arquivo de dados com a classe atribuída a cada objeto/observação.

Requisitos da Implementação do Agrupamento por k-Médias

- O programa deverá permitir ao usuário escolher o valor de k.
- O programa deverá permitir ao usuário gerar classificações com vários ks e guardar o resultado de cada uma, de forma que se possa compará-las.
- O programa deverá permitir que o usuário opte pelo cálculo dos valores de variância interna e externa e de F para realizar a análise de dados dos resultados produzidos.
- O programa deverá permitir que se exporte o arquivo de dados com a classe atribuída a cada objeto/observação.

A teoria e a aplicabilidade deste modelo será discutida no item 4.2.

4.2. Análise de Agrupamento (Cluster Analysis)

Esta análise tem por finalidade reunir, por algum critério de classificação, as unidades amostrais em grupos, de tal forma que exista máxima homogeneidade dentro do grupo e máxima heterogeneidade entre os grupos.

Veremos a análise de grupamentos com ênfase à classe dos métodos seqüenciais, aglomerativos, hierárquicos e não sobrepostos (SAHN). A representação dos resultados é feita num gráfico com estrutura de árvore denominado de dendograma.

Os procedimentos de análise de grupamento surgiram com a preocupação de biólogos e psicólogos, em avaliar numericamente as semelhanças ou não entre os organismos com vistas de esquemas de classificação.

Usualmente, o interesse se volta para o grupo de objetos ou indivíduos semelhantes, em termos de suas características (variáveis).

Para o processamento de Análise de grupamento após terem sido escolhidas as variáveis, devemos utilizar 3 procedimentos:

- padronização dos dados,
- escolha do coeficiente de semelhança,
- escolha da estratégia de grupamento.

4.2.1. Padronização dos Dados

Há duas razões para a padronização de matriz de dados. Primeiro, as unidades associadas aos atributos podem arbitrariamente afetar no grau de similaridade entre os objetos. Com a padronização esse efeito é eliminado. Segundo, a padronização faz com que os atributos contribuam com o mesmo peso no cálculo do coeficiente de similaridade entre objetos. Se uma variável possui um intervalo de valores superior a outro de uma variável, certamente a primeira variável contará com um peso maior na determinação do grau de similaridade entre os objetos. Este efeito pode ser compensado pela padronização.

Muitas são as funções de padronização utilizadas para a padronização da matriz de dados. Podemos admitir a seguinte matriz de dados: a matriz de dados tem **n** atributos (*j*= 1,2...,p) e têm **n** objetos (i = 1,2...,n). Na matriz de dados, o valor do *i*-ésimo objeto e *j*-ésimo atributo será denotado por **Xij**. O correspondente valor padronizado na matriz de dados será representado por **Zij**.

Dentre as funções de padronização a mais utilizada na prática é, para cada **i** fixo:

$$Z_{ij} = \frac{Xij - \overline{X}_j}{S_j}$$ que é utilizada no *software* do Statistica.

Após essa padronização, as variáveis passam a ter média igual a zero e a variância unitária.

4.2.2. Escolha do coeficiente de semelhança

Um conceito fundamental na aplicação da Análise de Agrupamento é a escolha de um coeficiente que quantifique o quanto dois objetos são parecidos. Estes coeficientes podem ser divididos em duas categorias: medidas de similaridade e medidas de dissimilaridade.

Na primeira, quanto maior o valor observado, mais similar serão os objetos e na segunda, quanto menor o valor mais similar serão os objetos.

Por exemplo, o coeficiente de correlação é uma medida de similaridade, pois quanto maior o seu valor maior a associação, enquanto que a distancia euclidiana é uma medida de dissimilaridade, pois quanto menor o valor mais próximo está um do outro.

4.2.3. Coeficientes de semelhança para variáveis quantitativas

Distância Euclidiana: sendo a mais utilizada. É a distância geométrica num espaço m – multidimensional. O cálculo da distância entre dois objetos A e B nesse espaço é feito através da fórmula matemática:

$$d_{AB} = \sqrt{(X_{1A} - X_{1B})^2 + (X_{2A} - X_{2B})^2 + ... + (X_{mA} - X_{mB})^2}$$

Distância Euclidiana Média: É derivada da distância anterior e também bastante utilizada. Possui duas considerações importantes; a primeira é que ela pode ser usada na ausência de dados para algumas coordenadas e a segunda permite acumular evidências empíricas sobre os níveis de similaridade. O cálculo da distância entre dois objetos A e B é feito através da fórmula matemática:

$$d_{AB} = \frac{1}{\sqrt{m}} \sqrt{(X_{1A} - X_{1B})^2 + (X_{2A} - X_{2B})^2 + ... + (X_{mA} - X_{mB})^2}$$

Distância de Manhattan: Esta distância e simplesmente a diferença comum no espaço m-dimensional. Na maioria dos casos, esta medida produz resultados semelhantes à distância euclidiana simples. Porém, esta medida sofre influência do efeito de grandes diferenças (ouliers). Esta distância é calculada pela seguinte fórmula matemática:

$$d_{AB} = |X_{1A} - X_{1B}| + |X_{2A} - X_{2B}| + ... + |X_{mA} - X_{mB}|$$

Distância de Chebychev: Esta medida de distância pode ser apropriada em casos quando a pessoa quer verificar se dois objetos são “diferentes”, num espaço m-multidimensional. Essa distância é calculada pela seguinte fórmula:

$$d_{AB} = máximo|X_{iA} - X_{iB}, onde, i = 1,2...,n$$

Distância Potência: Ás vezes o pesquisador pode querer aumentar ou diminuir o peso progressivo que é provocado pela dimensão nas quais os objetos respectivos são muitos diferentes. Isto pode ser realizado pela Distancia Potência. Esta distância é calculada pela seguinte equação matemática:

$$d_{AB} = \sqrt[n]{\left(x_{1A} - X_{1B}\right)^p + \left(x_{2A} - X_{2B}\right)^p + ... + \left(x_{mA} - x_{mB}\right)^p}$$

O paramêtro p controla o peso porgressivo que é dado para as diferenças em dimensões individuais e o paramêtro n controla o peso progressivo que é dado para diferenças maiores entre os objetos. Se p e n são iguais a dois, entao esta distância é igual a distância euclidiana.

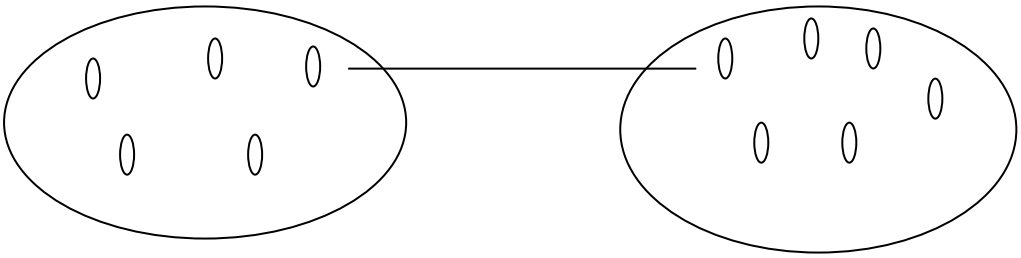
4.2.4. Escolha da estratégia de agrupamento

Dentre os métodos de agrupamento mais freqüentemente utilizados, destacam-se aqueles que caracterizam por serem Seqüenciais, Aglomerativos, Hierárquicos e Sem sobreposição (SAHN – Sequencial, Agglomerative, Hierarquical and Nonoverlaping methods).

Os métodos mais utilizados são:

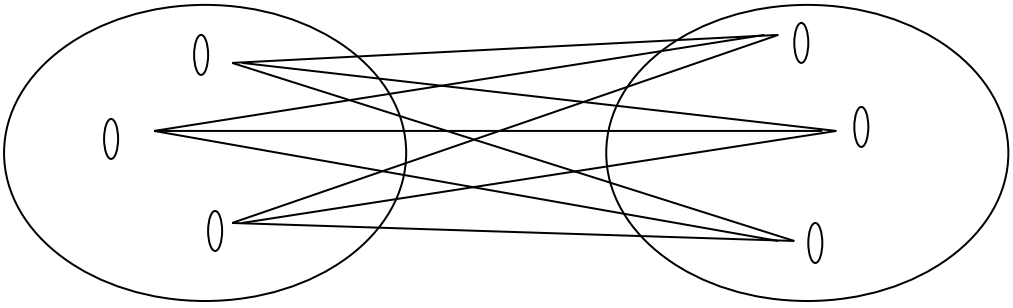
4.2.4.1. Ligação Simples (Single Linkage)

- a distância entre os grupos é definida como sendo a distância entre os elementos mais próximos (menor distância) dos dois grupos.



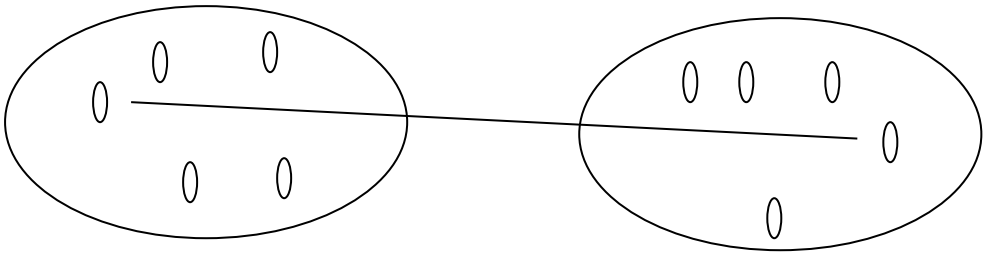
4.2.4.2. Unweighted pair- Group Method using arithmetic averages (UPGMA)

- (Não ponderado de agrupamento aos pares): neste método, a distância entre os dois grupos é definida como a média das distâncias entre todos os pares de valores de um grupo com o outro.



4.2.4.3. Ligação Completa (Complete Linkage)

– a distância entre dois grupos é definida como sendo a distância entre os indivíduos mais distantes dos dois grupos (distância máxima).



4.2.4.4. Método de Ward's

Este método forma grupos de dados buscando minimizar a soma das diferenças entre os elementos de cada grupo e o valor médio do grupo, minimizando o desvio padrão entre os dados de cada grupo.

Dendograma; É o gráfico bidimensional resultante da análise de agrupamento onde mostra os conglomerados formados.

4.2.5. Aplicação da Metodologia

O exercício desenvolvido a seguir mostra os passos dos cálculos da construção de um dendograma. A tabela abaixo contém 3 variáveis X_1,X_2 e X3 e seis unidades amostrais.

Unidade	X_1	X_2	X_3
1	2,00	4,00	5,00
2	3,00	5,00	3,00
3	6,00	4,00	3,00
4	4,00	2,00	1,00
5	3,00	1,00	1,00
6	6,00	1,00	4,00

Padronizando as colunas através da fórmula:

$$z_{ij} = \frac{xij - \bar{x}_j}{s_j}$$

Onde i = 1,...,n unidades para cada j = 1,..,p variáveis, construímos a tabela abaixo.

Unidade	X_1	X_2	X_3
1	-1,195	0,677	1,352
2	-0,598	1,258	0,104
3	1,195	0,677	0,104
4	0,000	-0,484	-1,144
5	-0,598	-1,064	-1,144
6	1,195	-1,064	0,728

A seguir foi construída a matriz fenética de semelhança (F) que será utilizada como matriz de partida na construção do dendograma. Essa matriz foi criada utilizando-se de coeficientes de similaridade ou dissimilaridades entre as unidades. Neste exemplo foi utilizada a distância euclidiana que é um coeficiente de dissimilaridade, pois quanto menor a distância mais similaridades são as unidades ou grupos.

4.2.6. Matriz fenética de semelhança (F)

A distância euclidiana entre dois pontos A ($x_1,...x_n$) e B ($y_1,...,y_n$) pertencentes ao Rn é dada por:

$$d_{AB} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

Para exemplificar os cálculos, determinaremos a distância entre as unidades 1 e 2 e a seguir entre 3 e 5:

$$F_{12} = \sqrt{(-1{,}195 + 0{,}598)^2 + (0{,}677 - 1{,}258)^2 + (1{,}352 - 0{,}014)^2} = 1{,}50$$

$$F_{35} = \sqrt{(-1{,}195 + 0{,}598)^2 + (0{,}677 - 1{,}064)^2 + (0{,}104 - 1{,}144)^2} = 2{,}79$$

Após os cálculos a matriz fenética de semelhança (F) tem estes resultados:

$$F = \begin{array}{c} \begin{array}{cccccc} (1) & (2) & (3) & (4) & (5) & (6) \end{array} \\ \left[\begin{array}{cccccc} 0{,}00 & 1{,}50 & 2{,}70 & 3{,}00 & 3{,}10 & 3{,}02 \\ & 0{,}00 & 1{,}88 & 2{,}22 & 2{,}64 & 3{,}00 \\ & & 0{,}000 & 2{,}08 & 2{,}79 & 1{,}85 \\ & & & 0{,}000 & 0{,}83 & 2{,}30 \\ & & & & 0{,}000 & 2{,}59 \\ & & & & & 0{,}00 \end{array}\right] \end{array} \begin{array}{l} (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \end{array}$$

Como estratégia de agrupamento, utilizaremos o método AVERAGE LINKAGE.

O processo se inicia, na matriz F, entre as unidades mais similares (menor distância).

Neste caso, a menor distância é 0,83 (unidades 4 e 5) que indicaremos por (4,5).

A seguir devemos reestruturar a matriz F originando a matriz F': para facilitar monte a matriz sem os elementos referentes a (4,5), conforme modelo a seguir:

F' =	(1)	(2)	(3)	(4)	(6)	
	0,00	1,50	2,70		3,02	(1)
		0,00	1,88		3,00	(2)
			0,000		1,85	(3)
				0,000		
						(4,5)
					0,00	(6)

Cálculo dos elementos de F':

$$d(45),1 = média(d_{41}; d_{51}) = média(3{,}00; 3{,}10) = 3{,}05$$

$$d(45),2 = média(d_{42}; d_{52}) = média(2{,}22; 2{,}64) = 2{,}43$$

$$d(45),3 = média(d_{43}; d_{53}) = média(2{,}08; 2{,}79) = 2{,}44$$

$$d(45),6 = média(d_{46}; d_{56}) = média(2{,}30; 2{,}59) = 2{,}45$$

Após completar a matriz F' com os valores temos:

F' =	(1)	(2)	(3)	(4,5)	(6)	
	0,00	1,50	2,70	3,05	3,02	(1)
		0,00	1,88	2,43	3,00	(2)
			0,000	2,44	1,85	(3)
				0,000	2,45	(4,5)
					0,00	(6)

A seguir, a menor distância em F' acontece entre as unidades 1 e 2 que indicaremos por (1,2). Devemos reestruturar a matriz F' originando a matriz F'': para facilitar monte a matriz sem os elementos referentes a (1,2):

F'' =

	(1,2)	(3)	(4,5)	(6)	
	0,00				(1,2)
		0,00	2,44	1,85	(3)
			0,000	2,45	(4,5)
				0,000	(6)

Cálculos dos elementos de F'':

$$d(12),3 = média(d_{13}; d_{23}) = média(2{,}70; 1{,}88) = 2{,}29$$

$$d(12),(45) = média(d_{1(45)}; d_{2(45)}) = média(3{,}05; 2{,}43) = 2{,}74$$

$$d(12),6 = média(d_{16}; d_{26}) = média(3{,}02; 3{,}00) = 3{,}01$$

Após substituir estes valores em F'' obtemos a matriz:

F'' =

	(1,2)	(3)	(4,5)	(6)	
	0,00	2,29	2,74	3,01	(1,2)
		0,00	2,44	1,85	(3)
			0,000	2,45	(4,5)
				0,000	(6)

A seguir, o menor valor em F'' é 1,86 (unidades 3 e 6) que indicaremos por (3,6).

Devemos reestruturar a matriz F'' originando a matriz F'''. Para facilitar monte a matriz sem os elementos referentes a (3,6):

F''' =

	(1,2)	(4,5)	(3,6)	
	0,00	2,74		(1,2)
		0,00		(3)
			0,000	(4,5)

$$d(3,6),(1,2) = média(d_3,(1,2); d_{6,(1,2)}) = média(2{,}29;3{,}01) = 2{,}65$$

$$d(3,6) = média(d_3,(_{4,5}); d_{6,(4,5)}) = média(2{,}44;2{,}45) = 2{,}45$$

Com esses resultados obtemos a matriz:

$$F''' = \begin{array}{c} \begin{array}{ccc} (1,2) & (4,5) & (3,6) \end{array} \\ \begin{bmatrix} 0{,}00 & 2{,}74 & 2{,}65 \\ & 0{,}00 & 2{,}45 \\ & & 0{,}000 \end{bmatrix} \end{array} \begin{array}{l} (1,2) \\ (4,5) \\ (3,6) \end{array}$$

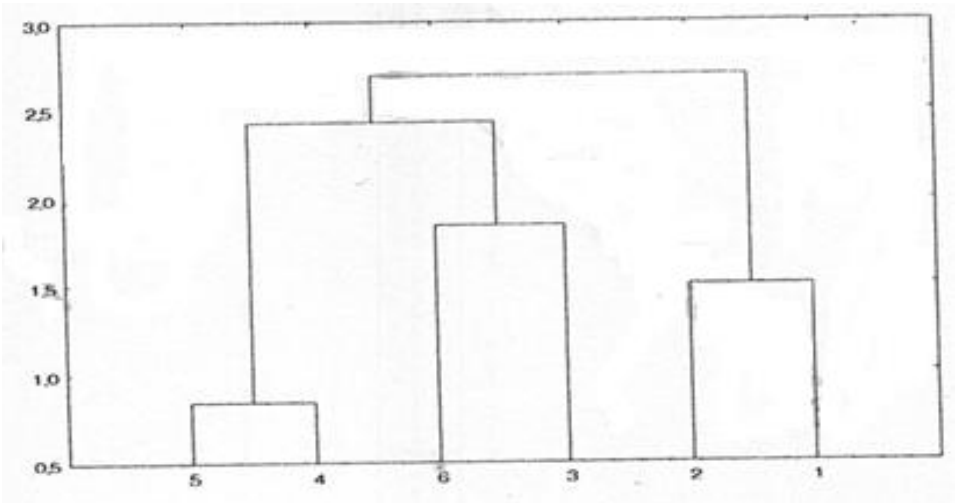
Finalmente, a menor distância em F''' é 2,45 {(3,6) com (4,5)}, que indicaremos por (3,6,4,5).

$$d_{(3,6),(4,5)} = média(d_{(3,6)}); d(_{(4,5)}) = média(2{,}44;2{,}65) = 2{,}695$$

Com estes resultados, obtivemos a matriz:

$$F'''' = \begin{array}{c} \begin{array}{cc} (1,2) & (3,6,4,5) \end{array} \\ \begin{bmatrix} 0{,}00 & 2{,}695 \\ & 0{,}00 \end{bmatrix} \end{array} \begin{array}{l} (1,2) \\ (3,6,4,5) \end{array}$$

Assim, o dendograma (ou fenograma) tem a seguinte distribuição de unidades:



4.3. Análise de Componentes Principais (PCA)

A análise de componentes principal é uma técnica estatística poderosa que pode ser utilizada para a redução do número de variáveis e para fornecer uma visão estatisticamente privilegiada de conjunto de dados, fornecendo as ferramentas adequadas para identificar as variáveis mais importantes no espaço de componentes principais.

Os fundamentos da PCA são apresentados descrevendo os passos matemáticos e estatísticos a partir das necessidades de interpretação adequada de matriz de dados. O entendimento exaustivo do assunto requer o conhecimento de operações com matrizes e por isso optamos por uma abordagem conceitual usando as noções de álgebra linear.

Esta técnica consiste em reescrever as variáveis originais em novas variáveis, denominadas componentes principais, através de uma transformação de coordenadas. A transformação matemática das coordenadas pode ser feita de diversas maneiras conforme o interesse.

Os componentes principais são as novas variáveis geradas através de uma transformação matemática especial realizada sobre as variáveis originais. Cada componente principal é uma combinação linear de todas as variáveis originais.Por exemplo, um sistema com 8 variáveis, após a transformação terá 8 componentes principais. Cada uma destas componentes, por sua vez será descrita como uma combinação linear das 8 variáveis originais. Nessas combinações, cada variável terá uma importância ou peso diferente.

As variáveis podem guardar entre si correlações que são suprimidas nas componentes principais, ou seja, as componentes principais são ortogonais entre si. Deste modo, cada componente principal traz uma informação estatística diferente das outras. A segunda característica importante é decorrente do processo matemático - estatístico de geração de cada componente que maximiza a informação estatística para cada uma das coordenadas que estão sendo criadas. As variáveis originais têm a mesma importância estatística, enquanto que as componentes principais são tão mais importantes que podemos até desprezar as demais. Destas características podemos compreender como a análise dos componentes principais:

a) Podem ser analisadas separadamente devido à ortogonalidade, servindo para interpretar o peso das variáveis originais na combinação das componentes principais mais importantes,
b) Podem servir para visualizar o conjunto da amostra apenas pelo gráfico das duas primeiras componentes principais, que detêm maior parte da informação estatística.

A análise de componentes principais é executada com o objetivo de simplificar a descrição de um conjunto de variáveis inter-relacionadas.Na PCA as variáveis não são discriminadas como independentes ou dependentes como na análise de regressão. Todas são tratadas como variáveis.

A técnica pode ser entendida como um método de transformação das variáveis originais em novas variáveis não correlacionadas. Cada componente principal é uma combinação linear das variáveis originais. Uma medida da quantidade de informação explicada por cada componente principal é a sua variância. Por esta razão os componentes principais são ordenados em ordem decrescente de sua variância, ou seja, o componente principal que contém mais informação é o primeiro, sendo o último aquela componente principal com menos informação.

Os componentes principais são não correlacionados o que é interessante pois um pesquisador estando com um problema envolvendo várias variáveis originais de complexo inter-relacionamento pode analisar um conjunto menor de variáveis não correlacionadas que são as componentes principais. Num conjunto grande de variáveis nem todas tem quantidade de informação relevante podendo através da ACP selecionar aquelas que mais possuem quantidade de informação relevante.

Análise de Componentes Principais é considerada uma técnica estatística exploratória utilizada na tentativa de compreender o inter-relacionamento entre as variáveis originais. A primeira aplicação da ACP foi no campo de testes educacionais.

4.3.1. Metodologia – Cálculo dos Componentes Principais

Seja um conjunto de equações simultâneas:

$AX = \lambda x$ ou $AX - \lambda x = 0$ ou $(A - \lambda I)X = 0$ (1):

Onde A é a matriz de correlação de coeficientes A_{ij} (matriz de correlações ou matriz de variâncias e covariâncias); X é um vetor desconhecido formado por cada Xi, λ constante (autovalor).

Adotaremos no exemplo a seguir a mesma matriz padronizada utilizada na análise de agrupamento do item 4.2.6.

A matriz C (3x3) de variâncias e covariâncias entre as 3 variáveis em estudo vale:

$$\rightarrow \mathbf{C} = \begin{bmatrix} 1 & -0{,}2775 & 0{,}0001 \\ -0{,}2775 & 1 & 0{,}4226 \\ 0{,}0001 & 0{,}4226 & 1 \end{bmatrix}$$

Assim:

$$|C - \lambda I| = 0 \Rightarrow$$

$$\left| \begin{bmatrix} 1 & -0{,}2775 & 0{,}0001 \\ -0{,}2775 & 1 & 0{,}4226 \\ 0{,}0001 & 0{,}4226 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = 0$$

$$\Rightarrow \quad \begin{bmatrix} 1-\lambda & -0{,}2775 & 0{,}0001 \\ -0{,}2775 & 1-\lambda & 0{,}4226 \\ 0{,}0001 & 0{,}4226 & 1-\lambda \end{bmatrix} = 0 \quad \Rightarrow$$

$$\lambda^3 - 3\lambda^2 + 2{,}744403\lambda - 0{,}744379 \cong 0$$

Esta equação é denominada equação característica e tem raízes (autovalores ou *eigenvalues*) aproximadamente:

$\lambda_1 \approx 1{,}5057$, $\lambda_2 \approx 1{,}0001$ e $\lambda_3 \approx 0{,}4942$

Substituindo cada λ_i no sistema de equações simultâneas (1) obtemos os autovetores (*eingenvector*) ω_i.

4.3.2. Cálculo do Primeiro Componente Principal – (eixo X)

Substituindo cada $\lambda_i = 1{,}5057$ gera o componente principal que retém quantidade maior da informação total. O sistema com esse autovalor fica:

$$\begin{cases} -0{,}5057X_1 - 0{,}2775X_2 + 0{,}0001X_3 = 0 \\ -0{,}2775X_1 - 0{,}5057X_2 + 0{,}4226X_3 = 0 \\ +0{,}0001X_1 + 0{,}4226X_2 - 0{,}5057X_3 = 0 \end{cases}$$

Este sistema deve ser indeterminado para que se tenha solução não nula. Isto quer dizer que a matriz dos coeficientes deve ser singular, ou seja, seu determinante é nulo.

É necessário normalizar os coeficientes da equação para que a solução seja única, isto é, os autovetores devem ser unitários e ortogonais entre si. Eliminemos a última equação e escolhemos $X_3=1$.

$$\begin{cases} 0{,}5057X_1 + 0{,}2775X_2 = 0{,}0001 \\ 0{,}2775X_1 + 0{,}5057X_2 = 0{,}4226 \end{cases}$$

Cujas soluções são $X_1 = -0{,}655868$ e $X_2 = 1{,}19557$

Temos para o autovalor $\lambda_1=1{,}5057$, o autovetor de $\omega_1 = (-0{,}655868;\ 1{,}195577;\ 1)$. O módulo deste vetor vale:

$$\sqrt{(-0{,}655868)^2 + (1{,}195577)^2 + (1{,}000)^2} = 1{,}691025$$

Assim, os coeficientes normalizados do primeiro componente principal valem:

$$a_{11} = \frac{-0{,}655868}{1{,}691025} = -0{,}387852$$

$$a_{21} = \frac{1{,}195577}{1{,}691025} = 0{,}707013$$

$$a_{31} = \frac{1{,}000}{1{,}691025} = 0{,}591357$$

Assim, o primeiro componente (CP1) principal vale:

$$CP1 = -0387858X_1 + 0{,}707013X_2 + 0{,}591357X_3$$

4.3.3. Cálculo do Segundo Componente Principal – (eixo Y)

O segundo maior autovalor $\lambda_2 \approx 1{,}0001$ gera o segundo componente principal que retém a maior parte da variabilidade que sobrou da quantidade armazenada no primeiro componente principal. O sistema com esse autovalor fica:

$$\begin{cases} -0{,}0001X_1 - 0{,}2775X_2 + 0{,}0001X_3 = 0 \\ -0{,}2775X_1 - 0{,}0001X_2 + 0{,}4226X_3 = 0 \\ +0{,}0001X_1 + 0{,}4226X_2 - 0{,}0001X_3 = 0 \end{cases}$$

Procedendo como anterior eliminemos a última equação e escolhemos $X_3=1$.

$$\begin{cases} -0{,}0001X_1 - 0{,}2775X_2 + 0{,}0001 = 0 \\ -0{,}2775X_1 - 0{,}0001X_2 + 0{,}4226 = 0 \end{cases}$$

cujas soluções são $X_1=1{,}522882$ e $X_2=-0{,}000188$. Assim, tem-se o autovalor $\lambda_1=1{,}0001$ e o autovetor $\omega_2 = (1{,}522882;-0{,}000188;1)$. O módulo deste vetor vale:

$$\sqrt{(1{,}522882)^2 + (-0{,}000188)^2 + (1{,}000)^2} = 1{,}821858$$

Os coeficientes normalizados do segundo componente principal valem:

$$a_{12} = \frac{1{,}522882}{1{,}821858} = 0{,}835895$$

$$a_{22} = \frac{-0{,}000188}{1{,}821858} = -0{,}000103$$

$$a_{32} = \frac{1{,}000}{1{,}821858} = 0{,}548890$$

Então, o segundo componente principal (CP2) vale:

$$CP2 = 0{,}835895X_1 - 0{,}000103X_2 + 0{,}548890X_3$$

4.3.4. Variância contida em cada componente principal

Variância contida em cada componente principal

Seja C a matriz de covariâncias dos dados originais padronizados.

Seja λ_n a h-ésima raiz característica (autovalor).

Seja $\lambda_1+\lambda_2+...+\lambda_n$= Traço (C).

A % da variância total contida em cada componente vale: $CP_h = \frac{\lambda_h}{Traço(C)}100$

% da variância total contida em $CP_1 = \frac{\lambda_1}{3}100 = \frac{1{,}5057}{3}100 = 50{,}19\%$

% da variância total contida em $CP_2 = \frac{\lambda_2}{3}100 = \frac{1{,}0001}{3}100 = 33{,}33\%$

% da variância total contida em $CP_3 = \frac{\lambda_3}{3}100 = \frac{0{,}4942}{3}100 = 16{,}47\%$

Notamos que quem retém mais informação (maior variabilidade das variáveis originais) é o componente principal 1 e os dois em conjunto (1 e 2) retém 83,52%.

Obs.: Vários autores sugerem escolher autovalores acima de 1 pois geram componentes com quantidade relevante de informação das variáveis originais. Abaixo de 1 a quantidade de informação retida no componente não é relevante.

4.3.5. Correção de cada variável com o componente principal

É obtida através da fórmula: $r_{xj}(cp_h) = \frac{a_{jh}\sqrt{\lambda_h}}{s_j}$

Onde:

s_j = desvio padrão da variável j;

a_{jh} = coeficiente da variável j no h-ésimo componente principal;

λ_h = h-ésima raiz característica (autovalor) da matriz de covariância.

No exemplo dado a correlação de cada variável no componente principal 1 (CP1) vale:

S_j = desvio padrão da variável j;
a_{jh} = coeficiente da variável j no h-ésimo componente principal,
λ_h = h-ésima raiz característica (autovalor) da matriz de covariância.

No exemplo, dado a correlação de cada variável no componente principal 1 (CP1) vale:

Temos $s_1 = 1$; $a_{11} = -0{,}387852$; $a_{21} = 0{,}707013$; $a_{31} = 0{,}591357$ e $\lambda_1 \approx 1{,}5057$

Assim:

$$r_{x_1}(cp_1) = \frac{-0{,}387852\sqrt{1{,}5057}}{1} = -0{,}475921$$

$$r_{x2}(cp_1) = \frac{0{,}707013\sqrt{1{,}5057}}{1} = 0{,}867554$$

$$r_{x3}(cp_1) = \frac{0{,}591357\sqrt{1{,}5057}}{1} = 0{,}725636$$

A importância de cada variável no componente principal 2 (CP2) vale, temos $s_1=1$; $a_{12}=0{,}835895$; $a_{22}= -0{,}000103$; $a_{32}= 0{,}548890$ e $\lambda_2 = 1{,}0001$. Assim:

$$r_{x_1}(cp_2) = \frac{0{,}835895\sqrt{1{,}0001}}{1} = 0{,}835936$$

$$r_{x2_1}(cp_2) = \frac{-0{,}000103\sqrt{1{,}0001}}{1} = -0{,}0001110$$

$$r_{x_3}(cp_{12}) = \frac{0{,}548890\sqrt{1{,}0001}}{1} = 0{,}548891$$

Resumindo:

Assim em ordem decrescente de importância no primeiro componente principal temos X_2, X_3 e X_1. Portanto as variáveis que mais discriminam no eixo horizontal são X_2 e X_3 nessa ordem e com correlações positivas indicando que unidades mais á direita possuem maior influência dessas variáveis e OTU's mais à esquerda possuem certa influência da variável X_1 (correlação negativa).

Quanto ao segundo componente principal, temos X_1 e X_3, nessa ordem, como variáveis discriminadoras e com correlações positivas, indicando que as unidades mais acima possuem maior influência da variável X_1 e também um pouco da variável X_3.

4.3.6. Construção gráfica bidimensional (CP1 x CP2)

Com os dois eixos podemos construir, um gráfico bidimensional mostrando, a nova distribuição das unidades.

$CP1 = -0{,}387852X_1 + 0{,}707013X_2 + 0{,}591357X_3$ (abscissas das OTU's)

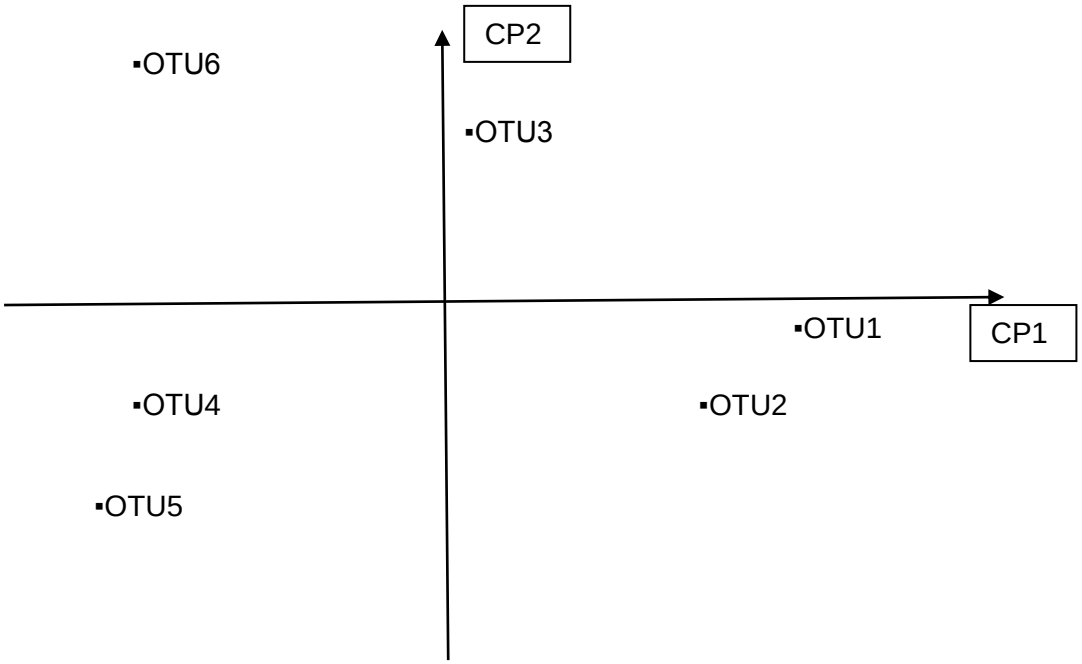
$$OTU(1) = (-0{,}387852)(-1{,}1950 + (0{,}707013)(0{,}677) + (0{,}591357)(1{,}352) = 1{,}74$$
$$OTU(2) = (-0{,}387852)(-0{,}598) + (0{,}707013)(1{,}258) + (0{,}591357)(0{,}104) = 1{,}18$$
$$OTU(3) = (-0{,}387852)(1{,}195) + (0{,}707013)(0{,}677) + (0{,}591357)(0{,}104) = 0{,}07$$
$$OTU(4) = (-0{,}387852)(0{,}000) + (0{,}707013)(-0{,}484) + (0{,}591357)(-1{,}144) = -1{,}01$$
$$OTU(5) = (-0{,}387852)(-0{,}598) + (0{,}707013)(-1{,}064) + (0{,}591357)(-1{,}114) = -1{,}19$$
$$OTU(6) = (-0{,}387852)(-1{,}195) + (0{,}707013)(-1{,}064) + (0{,}591357)(0{,}728) = -0{,}78$$

$CP2 = 0{,}835895X_1 - 0{,}000103X_2 + 0{,}548890X_3$ (ordenadas das OTU's)

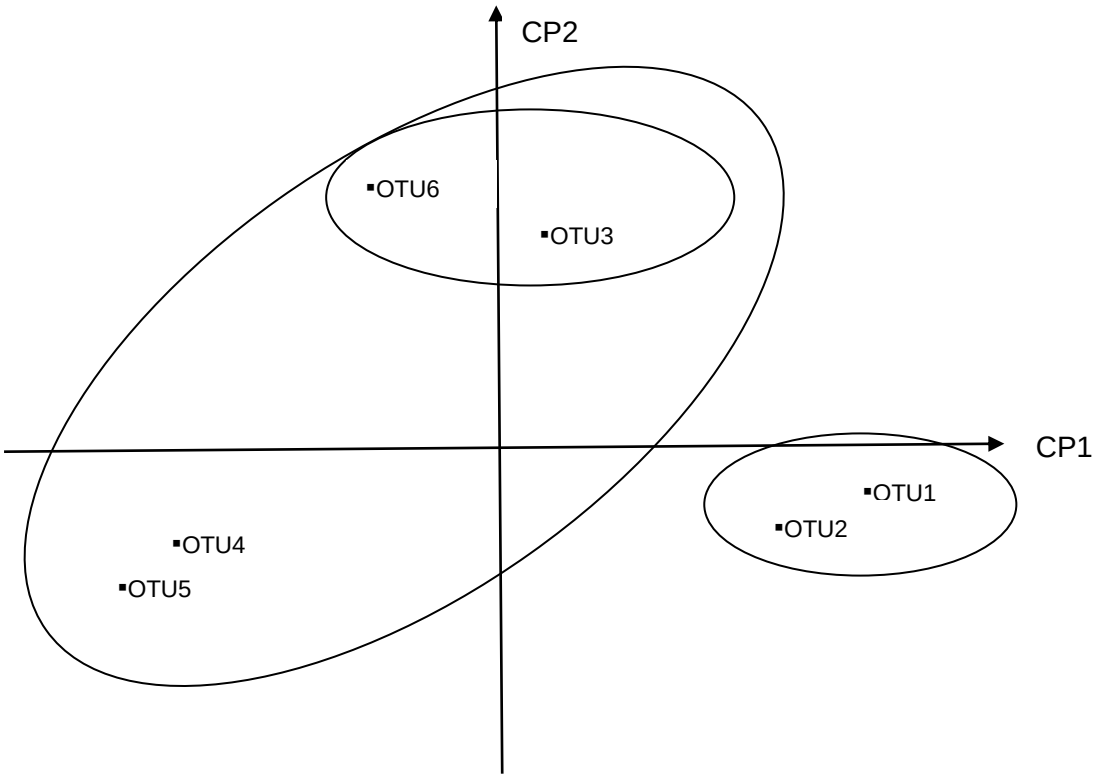
$$OTU(1) = (0{,}835895)(-1{,}195) + (-0{,}000103)(0{,}677) + (0{,}548890)(1{,}352) = -0{,}25$$
$$OTU(2) = (0{,}835895)(-0{,}598) + (-0{,}000103)(1{,}258) + (0{,}548890)(0{,}104) = -0{,}44$$
$$OTU(3) = (0{,}835895)(1{,}195) + (-0{,}000103)(0{,}677) + (0{,}548890)(0{,}104) = 1{,}05$$
$$OTU(4) = (0{,}835895)(0{,}000) + (-0{,}000103)(-0{,}484) + (0{,}548890)(-1{,}144) = -0{,}62$$
$$OTU(5) = (0{,}835895)(-0{,}598) + (-0{,}000103)(-1{,}064) + (0{,}548890)(-1{,}144) = -1{,}12$$
$$OTU(6) = (0{,}835895)(1{,}195) + (-0{,}000103)(-1{,}064) + (0{,}548890)(0{,}728) = -1{,}39$$

Resumindo:

Unid.	Variáveis			Componentes	
	X1	X2	X3	CP1	CP2
1	-1,195	0,677	1,352	1,74	-0,25
2	-0,598	1,258	0,104	1,18	-0,44
3	1,195	0,677	0,104	0,07	1,05
4	0,000	-0,484	-1,144	-1,01	-0,62
5	-0,598	-1,064	-1,144	-1,19	-1,12
6	1,195	-1,064	0,728	-0,78	1,39



Podemos reconstruir o gráfico bidimensional da análise de componentes principais usando a informação dendograma obtido anteriormente gerando um resumo mais completo dos resultados obtidos.

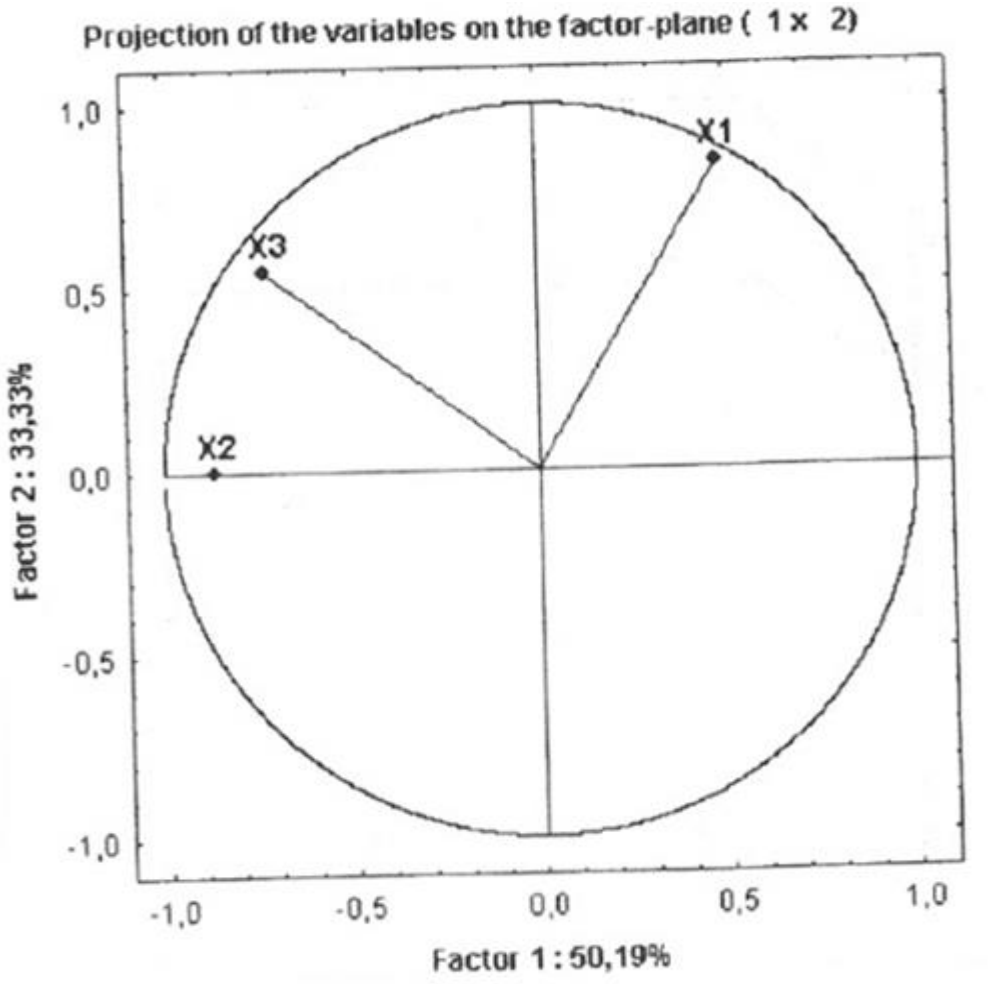


Conclusões:

Podemos dizer que temos dois ou três grupos formados dependendo do objetivo desejado. As variáveis X_2 e X_3 são responsáveis pela localização à direita do grupo formado pelas unidades 1 e 2, enquanto que a variável X1 tem alguma influência sobre os grupos formados pelas unidades 4 e 5 e pelas unidades 3 e 6 (localizadas mais a esquerda). Esta discussão se prendeu ao primeiro componente principal. Quanto ao segundo componente principal, podemos dizer que as variáveis que separaram os grupos formados pelas unidades 3 e 6 (acima) e unidades 4 e 5 (abaixo) foram X_1 fortemente e em parte X_3. Podemos admitir que as variáveis X_2 e X_3 são as mais discriminantes na formação dos grupos (maior correlação com o componente 1 – acima de 0,72 – que por sua vez é o que mais retém informação das variáveis originais – 50,19%). Não podemos desprezar a variável X^1 (correlação de 0,83 com componente principal 2) embora este componente tem menos informação que o anterior (aproximadamente 33%).

4.3.7. Resultados

- projeção das variáveis segundo os dois primeiros componentes principais:



- Correlação das variáveis com cada um dos componentes principais:

Factor-variable correlations (factor loadings), based on corre (Spreadsheet1) lations

	Factor 1	Factor 2	Factor 3
X1	0,476201	0,835944	0,272820
X2	-0,867689	0,000000	0,497108
X3	-0,725339	0,548815	-0,415554

- Autovalores e estatísticas:

Eigenvalues of correlation matrix, and related statistics (Spreadsheet1) Active variables only

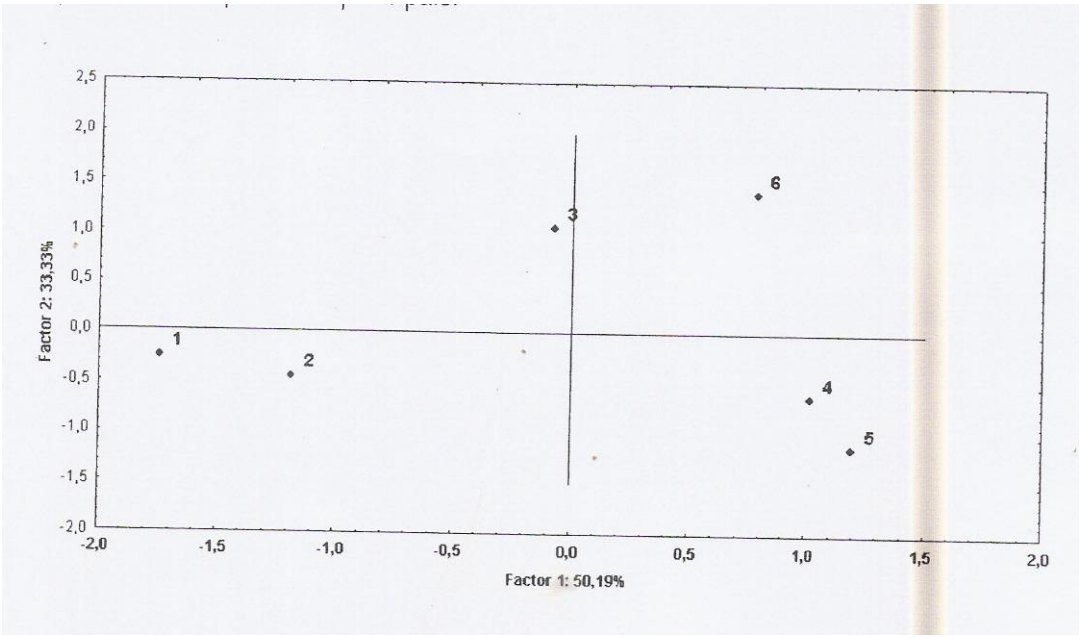
	Eigenvalue	% Total	Cumulative	Cumulative
1	1,505767	50,19224	1,505767	50,1922
2	1,000000	33,33333	2,505767	83,5256
3	0,494233	16,47443	3,000000	100,0000

- Autovetores e estatísticas:

Eigenvectors of correlation matrix (Spreadsheet1)

	Factor 1	Factor 2	Factor 3
X1	0,388071	0,835944	0,388071
X2	-0,707107	0,000000	0,707107
X3	-0,591102	0,548815	-0,591102

- Gráfico de linha mostrando os autovalores:



4.4. Análise de Correspondência

As análises de Correspondência (Simples (AC) ou Múltipla (ACM) são importantes recursos de análise multivariada aplicadas em conjunto de dados estritamente categóricos, por sua flexibilidade e facilidade de interpretação.

A Análise de Correspondência, na sua versão bivariada e multivariada, pode ser classificada no conjunto de técnicas associadas a mapas percepções /intuitivos. Esses mapas são definidos com "representação virtual das percepções dos objetos de um indivíduo em duas ou mais dimensões. Normalmente, este mapa tem níveis opostos de dimensões nos extremos dos eixos X e Y. Por exemplo, um mapa pode ser identificado nos extremos do eixo X como de "doce" a "azedo" e nos extremos do eixo Y como de "alto preço" a "baixo preço." Cada objeto tem, então, uma posição espacial no mapa percentual refletindo a relativa similaridade ou preferência em relação a outros segundo as dimensões do mapa perceptual.

Tais tipos de análises permitem que o analista visualize, como em um mapa geográfico, as proximidades (similaridades e dissimilaridades) entre os estímulos propostos no trabalho de pesquisa. O nome Análise de Correspondência deve-se ao fato de as linhas e colunas de uma tabela a ser transformado em unidades correspondentes , o que facilita sua representação conjunta.

A representação gráfica da AC é especialmente rica em informação, permitindo que o analista depreenda, rapidamente, as relações entre as variáveis.

4.5. Análise Discriminante

É a técnica de dependência mais utilizada. É aplicada quando a variável dependente é categórica (nominal ou não métrica) e as variáveis independentes são métricas. Por exemplo, queremos distinguir entre risco de crédito alto e baixo. Se tivéssemos uma medida de risco de crédito, poderíamos utilizar uma análise de regressão multivariada o que não é o caso, pois somente podemos saber se uma pessoa se encontra numa categoria de risco ou não. Esta é uma medida do tipo categórica (variável dependente) na qual se pode aplicar análise discriminante. Quando se têm duas classificações, a técnica é conhecida como Análise Discriminante de dois ou mais grupos e quando se tem 3 ou mais classificações, a técnica é conhecida como Análise Discriminante de dois ou mais grupos (MDA).

Após a definição dos grupos, são coletados dados individuais dos elementos de cada grupo. A análise discriminante procurar estimar a combinação linear das características individuais de cada elemento que melhor discrimina entre os grupos pré-estabelecidos. Outra vantagem da análise discriminante é reduzir o espaço dimensional das variáveis independentes para G-1 dimensões, onde G e o número de grupos estabelecidos a priori. No caso de dois grupos apenas, a análise é transformada em uma única dimensão. É utilizada também para classificar novos elementos dentro de um dos grupos.

A análise discriminante implica em obter um valor teórico que é uma combinação linear das variáveis independentes que discrimine melhor entre os grupos definidos a priori segundo:

$$Z_{IJ} = a + W_1 X_{1k} + W_2 X_{2k} + W_3 Y_{3k} + ... + W_n Y_{nk}$$

Onde:

Z_{IJ} = valor da função discriminante j para o objeto k

a = constante

W_1 = ponderação discriminante para a variável independente

X_{ik}= variável independente i para o objeto k.

Cada grupo gera uma função discriminante.

Supõe-se que as variáveis independentes venham de amostras de populações com distribuição normal multivariada e que se tenha, nos grupos, homogeneidade nas matrizes de variância / covariância das variáveis.

4.5.1. Distância Generalizada de Mahalanobis

A distância generalizada de Mahalanobis (D^2) é usada como uma técnica de comparação quanto à separação entre diversos grupos permitindo avaliar a extensão e a direção dos afastamentos entre os valores médios das variáveis usadas na discriminação.

O valor da distância generalizada D^2 ligando dois grupos é um número puro, com propriedades da distância comum, e mede a extensão com que diferem entre si em tamanho e forma.

Sendo $\bar{x}_i$ e $\bar{x}_j$ os vetores de médias dos grupos i e j e S a estimativa combinada da matriz de dispersão dentro dos grupos, a Distância Generalizada de Mahalanobis entre os grupos i e j é usualmente estimada por:

Sendo $D_{ij}^{\ 2} = \left[\bar{x}_i - \bar{x}_j\right]\left[S\right]^{-1}\left[\bar{x}_i - \bar{x}_j\right]$

Este método de representação de diferenças entre os grupos leva em conta qualquer correlação que exista entre as variáveis usadas e é também independente das unidades de medida com que as variáveis estão expressas.

4.6. Escalonamento Multidimensional

A técnica escalonamento multidimensional (MDS) é utilizada na construção de mapas perceptuais que melhor representem as semelhanças de objetos. Pode ser comparada com outras técnicas de interdependência, tais como, análise de fator e análise de agrupamento. A análise de fatores agrupa variáveis que se correlacionam enquanto que a análise de agrupamento agrupa objetos similares. Com MDS é possível analisar qualquer tipo de matriz, similaridade ou dissimiralidade, bem como correlação.

A medida mais comum que é usada para avaliar o bom ajuste de uma configuração particular que reproduza a configuração original é a medida do stress. Quando menor seu valor melhor será configuração encontrada. O stress de Kruskal é calculado pela fórmula:

$$Stress = \sqrt{\frac{(d_{ij} - \hat{d}_{ij})^2}{(d_{ij} - \overline{d})^2}}$$

Onde $\overline{d}$ = distância média no mapa

$\hat{d}_{ij}$ = Distância obtida através das medidas de similaridade

d_{ij} = distâncias originais

Segundo Kruskal o valor do stress tem a seguinte classificação:

STRESS	AJUSTE
20 %	POBRE
10 %	ACEITÁVEL
5 %	BOM
2,5 %	EXCELENTE
0 %	PERFEITO

4.6.1. Diagrama de Shepard

O gráfico de pontos onde o eixo y representa as distâncias referentes a mapa perceptual e o eixo x, as similaridades originais, é denominado de Diagrama de Shepard. Quando maior a coincidência dessas medidas mais perfeita é a configuração particular representada no mapa perceptual e por conseqüência, menor o stress. Desvios entre essa medidas indicam falta de ajuste.

5. Planejamento de Experimentos

Técnicas de planejamento experimental permitem alterar, de forma simultânea e sistemática, todas as variáveis relevantes envolvidas no processo ou no desenvolvimento de um produto, de tal maneira que a influência de cada variável possa ser estimada de forma precisa. A relação entre as variáveis envolvidas e a resposta ou propriedades do sistema (produto ou processo) é então descrita através de modelos matemáticos que fornecem ao investigador um completo entendimento de seu domínio experimental e com um número mínimo de experimentos, permitindo a rápida tomada de decisões e evitando gastos desnecessários com novos experimentos. Cabe ressaltar que métodos multivariados permitem estimar interações entre fatores, o que não é possível ser avaliados através de métodos univariados

O planejamento de experimentos (*Design of Experiments* – DOE) é uma técnica utilizada para se planejar experimentos, ou seja, para definir quais dados, que quantidade e em que condições os dados devem ser coletados durante um determinado experimento, buscando, basicamente satisfazer dois grandes objetivos: a maior precisão estatística possível na resposta e o menor custo. É portanto, uma técnica de extrema importância para a indústria pois seu emprego permite resultados mais confiáveis economizando dinheiro e tempo. A sua aplicação no desenvolvimento de novos produtos é muito importante, onde uma maior qualidade dos resultados dos testes pode levar a um projeto com desempenho superior seja em termos de suas características funcionais bem como sua robustez.

5.1. Objetivos

1) Determinação das variáveis relevantes em um determinado estudo;
2) Determinação da curvatura de um plano;
3) Determinação de pontos de máximo, mínimo e de inflexão das variáveis de resposta;
4) Determinação dos níveis das variáveis independentes que levam a um ótimo do experimento;
5) Determinação da composição ideal de uma mistura de componentes.

5.2. Aplicações

1) Avaliação e comparação de configurações básicas de projeto;
2) Avaliação de diferentes materiais;
3) Seleção de variáveis de projeto;
4) Determinação de variáveis de projeto que melhorem o desempenho de produtos.
5) Obtenção de produtos que sejam mais fáceis de fabricar, que sejam projetados, desenvolvidos e produzidos em menos tempo, que tenham melhor desempenho e confiabilidade que os produzidos pelos competidores.

5.3. Glossário

- **Fatores ou tratamentos:** são as variáveis de controle ou entrada,

- **Níveis:** correspondem às faixas de valores das variáveis de controle,

- **Variável resposta**: parâmetro de saída, resultante de uma variação nas variáveis de entrada,
- **Aleatorização:** é a prática de realizar a escolha das corridas ou pontos experimentais por meio de um processo aleatório. Esta prática simples em muitos casos garante as condições de identidade e independência dos dados coletados e evita erros sistemáticos.

- **Blocos:** São agrupamentos de dados para eliminar fontes de variabilidade que não de interesse do expectador.

5.4. Princípios Básicos

1) Replicação - obtenção do erro experimental,

2) Aleatoriedade - os experimentos devem ser realizados de forma aleatória para garantir não tendenciosidade nos resultados,

3) Blocagem - aumento da precisão de um experimento, pela eliminação de ruídos

.
Esses três princípios básicos de um planejamento de experimentos são *replicação*, *aleatoriedade* e *blocagem*. Fazer um experimento com réplicas é muito importante por dois motivos. O primeiro é que isto permite a obtenção do erro experimental. A estimativa desse erro é básica para verificar se as diferenças observadas nos dados são estatisticamente diferentes. O segundo motivo se refere ao fato de que, se a média de uma amostra for usada para estimar o efeito de um fator no experimento, a replicação permite a obtenção de uma estimativa mais precisa desse fator.

Os métodos estatísticos requerem que as observações, ou os erros, sejam variáveis aleatórias distribuídas independentemente. Os experimentos, com suas réplicas, devem ser realizados de forma aleatória, de modo a garantir a distribuição equânime de todos os fatores não considerados. Explicando melhor: em um estudo da influência, na dureza de um compósito, de diferentes ponteiras de uma máquina de ensaios mecânicos, podem-se utilizar corpos de prova provenientes de bateladas diferentes, que podem ter (mas teoricamente não deveriam), por exemplo, diferentes teores de fibras. Na realização dos testes, esses corpos de prova devem ser distribuídos de forma aleatória entre as ponteiras.

A blocagem é uma técnica extremamente importante, utilizada industrialmente que tem o objetivo de aumentar a precisão de um experimento. Em certos processos, pode-se controlar e avaliar, sistematicamente, a variabilidade resultante da presença de fatores conhecidos (*nuisance factors*) que perturbam o sistema, mas que não se tem interesse em estudá-los. A blocagem é usada, por exemplo, quando uma determinada medida

experimental é feita por duas diferentes pessoas, levando a uma possível não homogeneidade nos dados. Outro exemplo seria quando um determinado produto é produzido sob as mesmas condições operacionais, mas em diferentes bateladas. De modo a evitar a não homogeneidade, é melhor tratar cada pessoa e batelada como um bloco. Esta técnica será apresentada mais adiante.

As experiências devem ser realizadas seqüencialmente. A primeira delas, chamada *experimento de peneiramento (screen experiment),* é usada para determinar que variáveis são importantes (variáveis críticas). As experiências subseqüentes são utilizadas para definir os níveis das variáveis críticas identificadas anteriormente, que resultam em um melhor desempenho do processo.
Em suma, o que se quer aqui é obter um modelo matemático apropriado para descrever certo fenômeno, utilizando o mínimo possível de experimentos. O planejamento experimental permite eficiência e economia no processo experimental e o uso de métodos estatísticos na análise dos dados obtidos resulta em *objetividade científica* nas conclusões.

5.5. Etapas para o desenvolvimento de um Planejamento de Experimentos

Coleman & Montgomery propõem as seguintes etapas para o desenvolvimento de um planejamento de experimentos na indústria:

- Caracterização do problema,
- Escolha dos fatores de influência e níveis,
- Seleção de variáveis de resposta,
- Determinação de um modelo de planejamento de experimento,
- Condução de um experimento,
- Análise de dados,
- Conclusões e recomendações.

5.6. Planejamento Inicial

1) Definir as variáveis (podem ser qualitativas e quantitativas);

2) Definir os níveis importantes;

3) Os resultados devem ser analisados e modificações pertinentes devem ser feitas no planejamento experimental.

Todo planejamento experimental começa com uma série inicial de experimentos, com o objetivo de definir as variáveis e os níveis importantes. Podem-se ter variáveis qualitativas (tipo de catalisador, tipo de equipamento, operador, etc.) e quantitativas (temperatura, pressão, concentração, índice de inflação, ph do meio, etc.). Os resultados devem ser analisados e modificações pertinentes devem ser feitas no planejamento experimental. A Figura 1 apresenta um resumo desta estratégia inicial.

É importante frisar que os métodos que serão descritos aqui não substituem a imaginação e o bom senso, mas eles ajudam a economizar tempo e dinheiro, uma vez que eles conduzem à objetividade da análise de resultados.

Antes de começar a realizar os experimentos, os objetivos e os critérios devem estar bem claros, de modo a dar subsídios para a escolha:

1. Das variáveis envolvidas nos experimentos;
2. Da faixa de variação das variáveis selecionadas;
3. Dos níveis escolhidos para essas variáveis. No caso de muitos fatores, é melhor escolher inicialmente dois níveis;
4. Da variável de resposta;
5. Do planejamento experimental. Nessa etapa, há que se considerar o tamanho da amostra (número de réplicas), a seleção de uma ordem de realização dos experimentos e se há vantagem em fazer a blocagem dos experimentos; dos métodos de análise dos resultados dos experimentos.

Os métodos estatísticos são usados para guiar uma tomada objetiva de decisão.

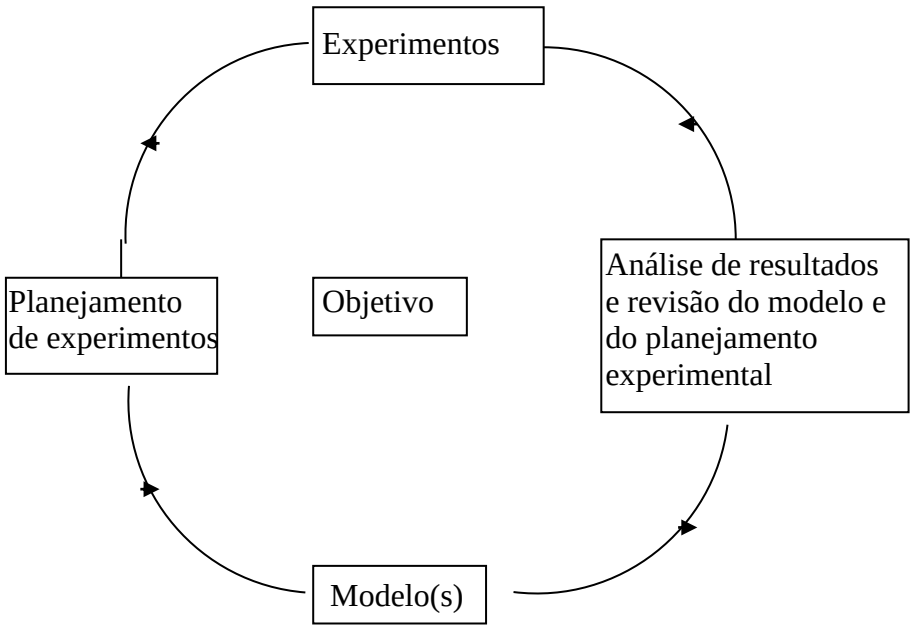


Figura 1 Estratégia Inicial.

As quatro etapas iniciais, conhecidas como planejamento pré-experimental, geralmente envolvem um grupo de pessoas. O sucesso de um planejamento de experimentos depende muito dessa etapa inicial.

5.7. Projeto e Análise de Experimentos

• Planejar o experimento para se ter informações suficientes (em termos dos objetivos da pesquisa) com o menor número possível de ensaios.

• Analisar os dados de forma compatível com o projeto experimental realizado.

5.8. Estratégias no Planejamento de Experimentos

1. Reconhecer, estabelecer e delimitar claramente o problema;

2. Identificar os possíveis fatores que podem afetar o problema em estudo;

3. Verificar quais fatores que poderão ser mantidos fixos e, portanto, não terão os seus efeitos avaliados no estudo experimental;

4. Identificar, para cada fator, o intervalo de variação e os níveis que entrarão no estudo;

5. Escolher um projeto experimental adequado, isto é, saber como combinar os níveis dos fatores de forma que se possa resolver o problema proposto com o menor custo possível;

6. Escolher a resposta adequada, ou seja, a variável *Y* que mede adequadamente o resultado (a qualidade, o desempenho, etc.) do processo ;

7. O planejamento de como será a análise dos dados do experimento.

5.9. Roteiro para a Realização de um Experimento

1. Identificar e estabelecer o problema;
2. Escolha dos fatores(k) e de seus níveis(*b*):bk;
3. Seleção da variável resposta (*Y*);
4. Escolha do projeto experimental;
5. Realização do experimento;
6. Análise estatística dos dados; significativo versus relevante.
7. Conclusões e recomendações.

5.9. Estudo Experimental

Manipula-se de forma planejada, certas *variáveis independentes* ou *fatores* (*A, B, C,...*) para verificar o efeito que esta manipulação provoca numa certa *variável dependente* ou *resposta Y.*

Processo

5.10. Exemplos do Tipo 2^k

a) Encontrar a melhor condição de operação de um processo químico:

- A resposta Y pode ser o rendimento da reação química e os fatores podem ser: o tempo de reação (A) e a temperatura de reação (B).

b) Verificar quais são os fatores que mais interferem na resistência à compressão (Y) de um concreto. Os fatores a serem estudados podem ser:

- O tempo de hidratação (A),
- A dosagem do cimento (B),
- A qualidade do cimento (C),
- O uso de aditivos (D).

5.11. Projeto de experimentos fatorial do Tipo 2^k

• Supõe-se:

1. Aleatorização;

2. Variação aleatória (erro experimental) *com distribuição normal;*

3. Para fatores quantitativos, supõem-se *efeitos lineares*;

4. Dados balanceados (*mesmo número* de observações em *cada combinação dos níveis* dos fatores).

6. Tipos de Planejamento

6.1. Planejamento com um único Fator

Em muitos experimentos diários, está-se interessado em comparar várias condições das variáveis independentes e analisar se existem diferenças entre elas. Imagine que se tenha a níveis de um fator e *n* réplicas. As corridas experimentais devem ser realizadas de forma completamente aleatória, a fim de eliminar os efeitos de variáveis que causem distúrbios (*nuisance factors*). Por exemplo, suponha que, ao ligar uma máquina, exista um tempo necessário para o seu aquecimento inicial. Se corridas forem feitas nesse período com amostras usando certo nível do fator, conclusões erradas podem ser obtidas, visto que esse tempo inicial pode afetar a variável de resposta.

Resumindo:

Problema: um único fator, com *a* níveis e *n* réplicas,
Objetivo: determinar diferença entre os níveis do fator,
Corridas: aleatórias.

Considere o exemplo em que se quer analisar a influência de cinco concentrações de algodão na resistência à tensão de um tecido sintético, conforme dados na Tabela 1.

Tabela 1. Dados da Resistência à Tensão do Algodão

% de Algodão	Réplicas				
	1	**2**	**3**	**4**	**5**
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

Neste exemplo, temos 5 réplicas (n = 5) e 5 níveis (a = 5)

Esses dados podem ser modelados através de uma equação matemática, cujo objetivo é poder prever novos valores da variável de resposta em pontos diferentes daqueles usados nos experimentos, mas dentro da faixa operacional.

Um modelo simples, conhecido como **modelo da média** (*means model*), pode ser usado para descrever o conjunto de pontos experimentais:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad (1)$$

y_{ij} = cada observação experimental
μ_i = média das observações experimentais para o nível *i*
ε_{ij} = erro aleatório associado à medição experimental da variável dependente

com i = 1, 2, ..., a, j = 1, 2, ..., n, μ_i sendo a média do i-ésimo tratamento e ε_{ij} sendo uma componente do **erro aleatório** que incorpora todas as outras fontes de variabilidade no experimento, incluindo medida, fatores incontroláveis diferenças entre unidades experimentais (como material de teste, equipamentos, etc.) e ruído em geral do processo (como variabilidade ao longo do tempo, efeitos do ambiente, etc.). É suposto que esse erro seja uma variável aleatória e tenha uma distribuição normal e independente, com média zero e variância constante, σ^2, para todos os níveis do fator; assim, $E(y_{ij}) = \mu_i$.

Outra maneira de expressar a Equação (1), conhecida como **modelo dos efeitos**, é:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad (2)$$

y_{ij} = cada observação experimental
μ = média global das observações experimentais
τ_i = efeito do i-ésimo tratamento
ε_{ij} = erro aleatório associado à medição experimental da variável dependente

em que μ é um parâmetro comum a todos os tratamentos, conhecido como **média global**, e τ_i é um parâmetro característico do i-ésimo tratamento, conhecido como **efeito do *i*-ésimo tratamento.**

$$\mu = \frac{\sum_{i=1}^{a} \mu_i}{a} \qquad (3)$$

$$\tau = \sum_{i=1}^{a} \tau_i = 0 \qquad (4),$$

pelo fato de $\mu_i = \mu + \tau_i$.

A Equação (2) é mais largamente utilizada, visto que μ é uma constante e τ_i representa desvios dessa constante quando os tratamentos específicos são aplicados. Essa equação é chamada também de **análise de variância univariável ou de um único efeito** (*one-way or single-factor analysis of variance*).

Os tratamentos podem ser variáveis fixas ou aleatórias. Eles serão considerados variáveis fixas, quando os seus níveis forem escolhidos pelo experimentalista. Nesse caso, as conclusões obtidas, utilizando testes de hipóteses relativos às médias dos tratamentos, serão aplicadas somente aos níveis considerados na análise e não a tratamentos similares que não tenham sido considerados explicitamente. Essa desvantagem pode ser superada se os a tratamentos forem amostras aleatórias retiradas de uma população maior de tratamentos. O conhecimento acerca de tratamentos particulares perde o significado, sendo mais importante agora testar hipóteses sobre a variabilidade de τ_i, tentando estimá-la.

O modelo que considera os tratamentos como variáveis fixas é chamado de **modelo de efeitos fixos** e o que considera variáveis aleatórias é chamado de **modelo de efeitos aleatórios** ou **modelo de componentes de variância**.

6.1.1. ANOVA para o Modelo de Efeitos Fixos

Está-se interessado em saber se existe alguma diferença entre as médias dos tratamentos; isto é, se $E(y_{ij}) = \mu_i = \mu + \tau_i$. As hipóteses são então:

H_0: $\mu_1 = \mu_2 = ... = \mu_a$ (5)
H_1: $\mu_i \neq \mu_j$ para pelo menos um par (i,j), ou pela Equação (5)

H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$
H_1: $\tau_i \neq \tau_j$ para no mínimo um valor de i

Para testar essas hipóteses, a melhor opção é a abordagem da análise de variância, como já visto anteriormente. O princípio básico desse método é dividir a variabilidade total em seus componentes. Essa variabilidade total é expressa em temos da **soma quadrática total**, ***SQ*$_T$:**

$$SQ_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{tt})^2 \qquad (6)$$

que se for dividida por N –1 graus de liberdade, tem-se a **variância da amostra** dos y's, que é uma medida padrão da variabilidade.

A soma quadrática total pode ser dividida nos seguintes termos:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{tt})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}[(\bar{y}_{it} - \bar{y}_{tt}) + (y_{ij} - \bar{y}_{it})]^2 \qquad (7)$$

que após manipulações algébricas e mostrando que o termo do produto cruzado é zero, fica-se com:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{tt})^2 = n\sum_{i=1}^{a}(\bar{y}_{it} - \bar{y}_{tt})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2 \qquad (8)$$

O primeiro termo do segundo membro representa a soma quadrática das diferenças **entre** (*between*) as médias dos tratamentos e a média global e a soma quadrática das diferenças das observações dentro (*within*) dos tratamentos e a média dos tratamentos. Essa última diferença é devida somente a erros aleatórios.

A Equação (8) pode ser expressa também por:

$$SQ_T = SQ_{Tratamentos} + SQ_E \qquad (9),$$

em que $SQ_{Tratamentos}$ representa a soma quadrática dos tratamentos (entre os tratamentos) e SQ_E representa a soma quadrática do erro (dentro dos tratamentos). Os graus de liberdade dessas somas são, respectivamente, a – 1 e N - a.

Se a Equação (9) for dividida por graus de liberdade apropriados, fica-se com duas estimativas para a variância da amostra, σ^2: uma baseada na variabilidade inerente dentro dos tratamentos e outra baseada na variabilidade entre os tratamentos. Claro

que se não houver diferença entre as médias dos tratamentos, essas estimativas serão iguais

.

De modo a determinar as estimativas das variâncias de cada termo do segundo membro da Equação (8), pode-se proceder conforme detalhamento a seguir.

Expressando a soma quadrática sob outra forma, tem-se:

$$SQ_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2 = \sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2\right] \qquad (10)$$

Se o termo entre colchetes for dividido por *n* – 1, ter-se-á a variância da amostra no *i*-ésimo tratamento:

$$S_i^2 = \frac{\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2}{n-1} \qquad (11)$$

Essas variâncias podem ser combinadas (ponderadas) para resultar uma estimativa simples da variância da população, tendo como objetivo o aumento no número de graus de liberdade.

$$\sum_{i=1}^{a}(n-1)S_i^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2 \qquad (12)$$

$$\frac{\sum_{i=1}^{a}(n-1)S_i^2}{\sum_{i=1}^{a}(n-1)} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{it})^2}{\sum_{i=1}^{a}(n-1)} = \frac{SQ_E}{(N-a)} \qquad (13)$$

O último termo da Equação (13) corresponde à estimativa combinada da variância comum dentro de cada um dos <u>*a*</u> tratamentos, sendo chamado de **média quadrática do erro, *MQ*$_E$**.

Se não houver diferenças entre as *a* médias dos tratamentos, σ^2 pode ser estimada através de:

$$\frac{SQ_{Tratamentos}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{it} - \bar{y}_{tt})^2}{a-1} = n\left(\frac{\sigma^2}{n}\right) \qquad (14)$$

sendo (σ^2/n) a variância das médias dos tratamentos. O termo $SQ_{Tratamentos}/(a-1)$ = $MQ_{Tratamentos}$ é chamada de **média quadrática de tratamentos**.

Como os valores esperados das médias quadráticas dos tratamentos e do erro são dados por (para detalhes, ver Montgomery, 2001):

$$E(MQ_{Tratamentos}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} \qquad (15)$$

$$E(MQ_{Erro}) = \sigma^2 \qquad (16)$$

Percebe-se então que se não houver diferença nas médias dos tratamentos (o que implica que τ_i = 0), $MQ_{Tratamentos} = \sigma^2$. Logo, pode-se detectar diferenças nas médias dos tratamentos se $MQ_{Tratamentos}$ e MQ_E forem comparados.

O teste de hipótese apresentado anteriormente utiliza à estatística *F*, dada por:

$$F_o = \frac{MQ_{Tratamentos}}{MQ_{Erro}} \quad , \qquad (17)$$

com *a* – 1 e *N* – *a* como graus de liberdade do numerador e denominador, respectivamente. Se a hipótese nula for verdadeira, as duas médias quadráticas estimam o mesmo valor para σ^2. Mas se a hipótese nula for falsa, o numerador será maior do que o denominador, implicando-se em se ter

$$F_o > F_{\alpha,a-1,N-a} \quad , \qquad (18)$$

sendo α o nível de significância.

O intervalo de confiança para a média do *i*-ésimo tratamento é dado por:

$$\bar{y}_{it} - t_{\alpha/2,a(n-1)}\sqrt{\frac{MQ_E}{n}} \le \mu_i \le \bar{y}_{it} + t_{\alpha/2,a(n-1)}\sqrt{\frac{MQ_E}{n}} \qquad (19)$$

O intervalo de confiança para a diferença entre as médias de dois tratamentos é dado por:

$$\bar{y}_{it} - \bar{y}_{jt} - t_{\alpha/2,a(n-1)}\sqrt{\frac{2MQ_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{it} - \bar{y}_{jt} + t_{\alpha/2,a(n-1)}\sqrt{\frac{2MQ_E}{n}} \qquad (20)$$

No caso de não se dispor de todas as informações experimentais (planejamento desbalanceado), as equações das somas quadráticas devem ser modificadas para:

$$SQ_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{tt}^2}{N} \qquad (21)$$

$$SQ_{Tratamentos} = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i} - \frac{y_{tt}^2}{N} \qquad (22)$$

Sempre que possível, deve-se escolher um planejamento balanceado, visto que ele é mais potente e é menos sensível a pequenos desvios da suposição de igualdade de variâncias para os *a* tratamentos.

Ao final de qualquer planejamento, precisa-se verificar a adequação do modelo matemático obtido e a validade das suposições feitas.

6.1.2. One-Way ANOVA

H_0: $\mu_1 = \mu_2 = ... = \mu_a$
H_1: $\mu_i \neq \mu_j$ para pelo menos um par (i,j)

H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$
H_1: $\tau_i \neq \tau_j$ para no mínimo um valor de i

Tratamento (nível)	Observações				Totais	Médias
1	y_{11}	y_{12}	...	y_{1n}	y_{1t}	$\bar{y}_{1t}$
2	y_{21}	y_{22}	...	y_{2n}	y_{2t}	$\bar{y}_{2t}$
.	.	.		.	.	
.	.	.		.	.	.
.	.	.		.	.	.
						.
a	y_{at}	y_{a2}	...	y_{an}	y_{at}	$\bar{y}_{at}$
					y_{tt}	$\bar{y}_{tt}$

$$y_{it} = \sum_{i-1}^{n} y_{ij} \quad (23)$$

$$\bar{y}_{it} = \frac{y_{it}}{n} \quad (24)$$

$$y_{tt} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \quad (25)$$

$$\bar{y}_{tt} = \frac{y_{tt}}{N} \quad (26)$$

6.1.3. Abordagem ANOVA

$$SQT = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{tt}\right)^2 \qquad (27)$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{tt}\right)^2 = i = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{y}_{it} - \bar{y}_{tt}) + (y_{ij} - \bar{y}_{it})\right]^2 \qquad (28)$$

$SQ_T = SQ_{Tratamentos} + SQ_{ERRO}$ *(29)*

$SQ_{Tratamentos}$ - representa a soma quadrática das diferenças **entre** (*between*) as médias dos tratamentos e a média global

SQ_{ERRO} - representa a soma quadrática das diferenças das observações dentro (*within*) dos tratamentos e a média dos tratamentos. Causada apenas pelos erros aleatórios.

$$\frac{SQ_{Tratamentos}}{a-1}$$ (30), variância baseada na variabilidade entre os tratamentos.

$$\frac{SQ_{Erro}}{N-a}$$ (31), variância baseada na variabilidade inerente dentro dos tratamentos

Se não houver diferença entre as médias dos tratamentos, essas estimativas serão iguais.

$$E(MQ_{Tratamentos}) = \sigma 2 + \frac{n\sum_{i=1}^{a}\tau_i^{\,2}}{a-1} \qquad (32)$$

$$E(MQ_{Erro}) = \sigma^2 \qquad (33)$$

Se não houver diferença nas médias dos tratamentos (τi = 0):

$MQ_{Tratamentos} = \sigma^2$ (34)

Conclusão: Pode-se detectar diferenças nas médias dos tratamentos se $MQ_{Tratamentos}$ e MQ_E forem comparados

$$F_0 = \frac{MQTratamentos}{MQErro} \quad (35)$$

com $a - 1$ e $N - a$ como graus de liberdade do numerador e denominador, respectivamente.

H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$
H_1: $\tau_i \neq \tau_j$ para no mínimo um valor de i
H_0 verdadeira: $MQ_{Tratamentos}$ e MQ_{Erro} estimam o mesmo valor de σ^2
H_0 falsa: $MQ_{Tratamentos} > MQ_{Erro}$, implica em se ter:

$F_o > F_{\alpha,a-1,N-a}$, sendo α o nível de significância. (36)

Exemplo:
Para os dados experimentais obtidos na tabela abaixo, verifica-se a concentração de algo que afeta a resistência à tensão do tecido sintético. Usa-se α = 0,01.

Testar as hipóteses que se quer testar:

H_0: $\tau_1 = \tau_2 = ... = \tau_5 = 0$
H_1: $\tau_i \neq 0$ *para no mínimo um i*

% de algodão	Réplicas					Totais	Médias
	1	2	3	4	5		
15	7	7	15	11	9	49	9,8
20	12	17	12	18	18	77	15,4
25	14	18	18	19	19	88	17,6
30	19	25	22	19	23	108	21,6
35	7	10	11	15	11	54	10,8
						y_{tt} =376	$\overline{y}_{it}$ =15,04

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade d.f.	Média Quadrática MQ	F	p
Concentração de algodão	475,76	4	118,94	14,76	0,000009
Erro	161,20	20	8,06		
Total	636,96	24			

Analisando os resultados obtidos na tabela 2, verifica-se que o valor do teste F calculado (14,76) foi maior do que o valor tabelado ($F_{0,05;4;20}$ = 2,87), levando-se então

a rejeição da hipótese. Analisando o valor de p, verifica-se também que ele foi < 0,01; o nível de significância requerido pelo problema.

Conclusão: A concentração influência a resistência do tecido à tensão.

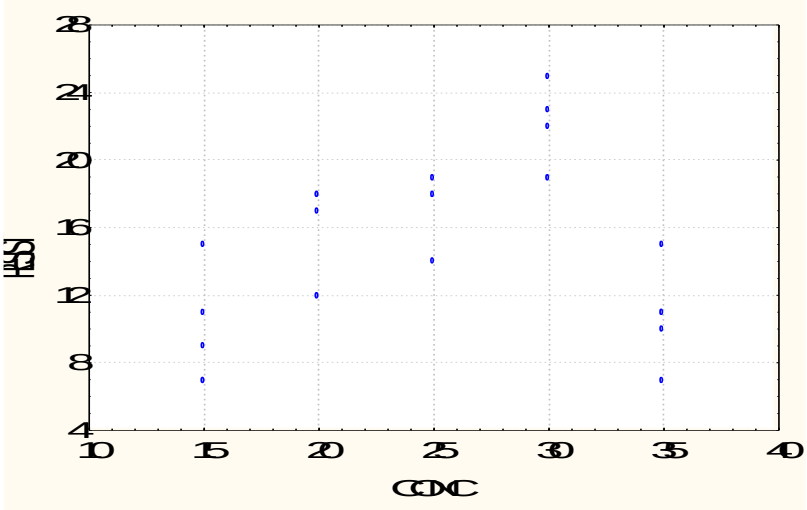
6.1.3. Verificação da Adequação do Modelo

Suposição - erros, ε_{ij}, são considerados aleatórios e distribuídos normal e independentemente, com média zero e variância constante σ^2.

Verificação - vários testes devem ser feitos nos resíduos, $e_{ij} = y_{ij} - \hat{y}_{ij}$.

a) Teste de Normalidade

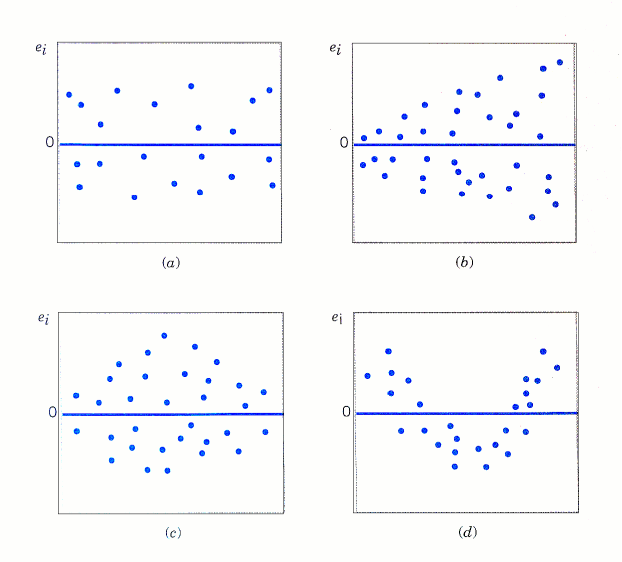
Tabela xx – Variância Constante



b) Presença de *Outliers*

$$dij = \frac{e_{ij}}{\sqrt{MQ_E}} \qquad (37)$$, não devem ser maiores do que ± 2.

C) Aspecto Esperado dos Resíduos



Caso (a): aspecto esperado.

Caso (b): aumento da variância (corrigir, transformando *y* em *ln y*, $\sqrt{y}$, 1/*y*.

Caso (c): desigualdade na variância.

Caso (d): termos de ordens mais elevadas devem ser adicionados ao modelo matemático.

A influência da concentração de algodão na resistência do tecido pode ser vista de forma qualitativa, através de gráficos, e de forma quantitativa, através de uma equação matemática.

No caso de se querer determinar diferenças entre as médias dos 5 níveis, poderia-se pensar em usar o teste *t* para todos os pares possíveis de médias. A questão é que tal procedimento causaria um aumento no erro tipo I. Explicando melhor; há 10 pares possíveis e se a probabilidade de aceitar corretamente a hipótese nula para cada teste individual for 1 - α = 0,95, a probabilidade de aceitar corretamente a hipótese nula para todos os 10 pares é $(0,95)^{10} \cong 0,60$, se os testes forem independentes. Observa-se assim, um aumento do erro tipo I de 0,05 para cerca de 0,40.

A melhor abordagem quando se quer comparar várias médias é a análise de variância. Para isso, considere *a* **tratamentos** ou **níveis diferentes** de um **único fator** e *n* réplicas, dando origem à Tabela do exemplo 1 (página xxx), em que os y_{ij} correspondem aos valores experimentais da variável de resposta, os y_{it} correspondem à soma das observações para a linha *i* (*i*-ésimo tratamento), os $\bar{y}_{it}$ correspondem às médias das observações para a linha *i*, y_{tt} correspondem à soma de todas as observações e $\bar{y}_{tt}$ corresponde à média de todas as observações.

$$y_{it} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{it} = y_{it} / n \qquad (38)$$

$$y_{tt} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{tt} = y_{tt} / N \qquad (39)$$

sendo $N = na$, o número total de observações.

6.2. Planejamento Aleatório com blocos completos

Com o objetivo de estudar a blocagem, suponha que se queira saber se quatro diferentes tipos de ponteiras (considera-se cada ponteira como um tratamento) de uma máquina de ensaios produzem diferentes leituras de dureza de um compósito. Imagine agora que sejam realizados quatro ensaios para cada ponteira. Existe assim um único fator (variável independente importante), que é o tipo de ponteira. No total, serão necessários 16 experimentos, envolvendo então 16 peças, que podem ser levemente diferentes entre si, resultando em uma possível variação nos valores de dureza da peça. Como resultado, o erro experimental refletirá tanto o erro aleatório como a variação entre as peças. Uma maneira de diminuir esse erro experimental é remover a variabilidade entre as peças, fazendo a blocagem; ou seja, considerando cada peça como sendo um bloco, sujeito ao teste com as quatro ponteiras. Tem-se assim um experimento mais uniforme que serve para comparar a influência das ponteiras na dureza do material. A esse planejamento, dá-se o nome de *planejamento aleatório com blocos completos.* Por completo entende-se que cada bloco contém todos os tratamentos (todas as ponteiras atuam neste bloco). A Tabela 14.5 apresenta o planejamento aleatório com blocos completos para o caso do teste de dureza do compósito.

Tabela 3. Planejamento Aleatório com Blocos Completos para a Dureza de um Compósito

	Peça			
Tipo de ponteira	1	2	3	4
1	9,3	9,4	9,6	10,0
2	9,4	9,3	9,8	9,9
3	9,2	9,4	9,5	9,7
4	9,7	9,6	10,0	10,2

O planejamento aleatório com blocos completos é um dos mais usados planejamentos experimentais. Geralmente, matéria-prima, pessoas e tempo são as fontes mais usuais de variabilidade nos experimentos.

O modelo matemático mais simples para representar o planejamento é dado a seguir:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \end{cases} \qquad (40)$$

em que μ é a média global, τ_i é o efeito do *i*-ésimo tratamento (ponteira, no caso), β_j é o efeito do *j*-ésimo bloco (peça no caso) e ε_{ij} é o erro aleatório que segue a distribuição normal padrão. Colocando na forma matricial, tem-se:

$$\underline{\underline{y}} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1b} \\ y_{21} & y_{22} & \cdots & y_{2b} \\ \vdots & \vdots & \cdots & \vdots \\ y_{a1} & y_{a2} & \cdots & y_{ab} \end{pmatrix} \qquad (41),$$

sendo então *a* o número total de tratamentos (ponteiras no caso) e *b* o número total de blocos (peças no caso). O produto *ab* é o número total, *n*, de experimentos. Cada coluna da matriz representa um bloco. Há apenas uma observação por tratamento em cada bloco; existe uma aleatoriedade na ordem de realização dos tratamentos dentro de cada bloco. Assim, os blocos representam uma restrição à aleatoriedade.

Está-se interessado em saber se os quatro tratamentos (ponteiras) fornecem o mesmo valor médio de dureza; ou seja:

H_o: $\mu_1 = \mu_2 = ... = \mu_a$ (42)

H_1: no mínimo um $\mu_i \neq \mu_j$

Outra maneira de expressar a Equação (42) é:

H_o: $\tau_1 = \tau_2 = ... = \tau_a = 0$ (43)

H_1: $\tau_i \neq 0$, no mínimo em um *i,* visto que a média do *i*-ésimo tratamento é:

$$\mu_i = (1/b) \sum_{j=1}^{b} (\mu + \tau_i + \beta_j) = \mu + \tau_i. \qquad (44)$$

O desvio do ponto experimental y_{ij} em relação ao valor médio de todos os pontos experimentais pode ser decomposto nos termos mostrados na Equação (41).

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}) \qquad (45),$$

em que

$$\bar{y}_i = \frac{\sum_{j=1}^{b} y_{ij}}{b} \qquad \bar{y}_j = \frac{\sum_{i=1}^{a} y_{ij}}{a} \qquad \bar{y} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}}{n} \qquad (46)$$

Como antes, se a Equação (41) for elevada ao quadrado, e aplicado o somatório sobre todos os pontos experimentais, obtém-se a Equação (43), uma vez que os termos cruzados se anulam.

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y})^2 = b\sum_{i=1}^{a}(\bar{y}_i - \bar{y})^2 + a\sum_{j=1}^{b}(\bar{y}_j - \bar{y})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2 \qquad (47)$$

$$SQ_{Total} = SQ_{Tratamentos} + SQ_{Blocos} + SQ_{Erro} \quad (48)$$

A Tabela 4 apresenta uma análise de variância para a blocagem proposta.

Tabela 4 – Análise de Variância para o Planejamento Aleatório com Blocos Completos

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade (d.f.)	Média Quadrática
Tratamentos	$\mathbf{SQ_{Tratamentos}}$	$\boldsymbol{a}$ **- 1**	$MQ_{Tratamentos} = \frac{SQ_{Tratamentos}}{a-1}$
Blocos	$\mathbf{SQ_{Blocos}}$	$\boldsymbol{b}$ **- 1**	$MQ_{Blocos} = \frac{SQ_{Blocos}}{b-1}$
Erro	$\mathbf{SQ_{Erro}}$	$\boldsymbol{(a-1)(b-1)}$	$MQ_{Erro} = \frac{SQ_{Erro}}{(a-1)(b-1)}$
Total	$\mathbf{SQ_{Total}}$	$n-1$	$MQ_T = \frac{SQ_T}{n-1}$

O coeficiente de determinação é dado pela Equação (35):

$$r^2 = \frac{SQ_{Modelo}}{SQ_{Total}} = \frac{SQ_{Tratamentos} + SQ_{Blocos}}{SQ_{Total}} \quad (49)$$

Como antes, realiza-se o teste F com o objetivo de validar a hipótese H_o.

$$F(a-1,(a-1)(b-1)) = \frac{MQ_{Tratamentos}}{MQ_{Erro}} > F_{\alpha}(a-1,(a-1)(b-1)) \quad (50),$$

com α sendo a significância do teste. Se a Equação (49) se verificar, a hipótese H_o é rejeitada, concluindo-se que diferentes tratamentos (ponteiras) influenciam a dureza do material. Para saber o efeito dos blocos, seria natural realizar o teste F, calculado da seguinte forma:

$$F(b-1,(a-1)(b-1)) = \frac{MQ_{Blocos}}{MQ_{Erro}} > F_{\alpha}(b-1,(a-1)(b-1)) \quad (51)$$

Alguns pesquisadores acham que isto não é correto, pois o teste F só pode ser realizado para experimentos completamente aleatórios, o que não é o caso, visto que a aleatoriedade só existe dentro dos blocos. Montgomery, em seu livro de Planejamento e Análise de Experimentos, não utiliza esse teste, mas diz que a razão MQ_{Blocos}/MQ_{Erro} fornece uma idéia do efeito dos blocos na variável de resposta. Se o seu valor for grande, pode-se inferir que o bloco tem uma grande influência na redução de ruído (variabilidade), sendo útil então para melhorar a precisão da comparação das médias dos tratamentos.

Para os dados na Tabela 3, os resultados podem ser visualizados na Tabela 5.

Tabela 5 – Resultados do Planejamento Aleatório com Blocos Completos

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade (d.f.)	Média Quadrática (d.f.)	F	*p-level*
Tratamentos	**38,50**	**3**	**12,83**	**14,44**	**0,0009**
Blocos	**82,50**	**3**	**27,50**		
Erro	**8,00**	**9**	**0,89**		
Total	**129,00**	**15**			

O resultado obtido aqui seria diferente, caso se tivesse usado a abordagem do planejamento completamente aleatório com um único fator. Os termos referentes aos blocos seriam somados ao termo do erro, resultando na Tabela 6.

Tabela 6 – Resultados do Planejamento Completamente Aleatório com um Único Fator

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade (d.f.)	Média Quadrática (d.f.)	F	*p-level*
Tratamentos	**38,50**	**3**	**12,83**	**1,40**	**0,219902**
Erro	**90,50**	**12**	**7,54**		
Total	**129,00**	**15**			

Como se pode notar, a conclusão seria oposta à obtida anteriormente, mostrando a importância da formação de blocos para evitar o aumento no termo do erro. Esse aumento faz com que diferenças importantes entre as médias dos tratamentos não possam ser detectadas.

6.3. Planejamento com blocos incompletos

Às vezes, por motivos alheios, tais como falta de matéria-prima, danos na aparelhagem experimental, alto custo de obtenção de muitos dados experimentais, etc., alguns dados não são disponíveis para todos os tratamentos e todos os blocos.

Assim, os tratamentos não são mais ortogonais aos blocos. De modo a contornar tal problema, existem duas alternativas. A primeira seria uma análise aproximada, que consiste em estimar a informação perdida e usá-la como se fosse uma informação real. A segunda alternativa seria fazer uma análise exata, que consiste em usar o teste geral de significância da regressão.

No caso de uma abordagem aproximada, o objetivo é estimar o valor inexistente através da minimização da soma quadrática do erro. A equação resultante é dada a seguir, cujo desenvolvimento pode ser encontrado no livro de D. C. Montgomery, 2001.

$$y_{ij}^{*} = \frac{ay_{it}^{*} + by_{tj}^{*} - y_{tt}^{*}}{(a-1)(b-1)}$$ (52), em que y_{ij}^{*} é o valor inexistente que se quer estimar, y_{it}^{*} é a soma da variável dependente em todos os blocos para o *i*-ésimo tratamento, y_{tj}^{*} é a soma da variável dependente em todos os tratamentos para o *j*-ésimo bloco e y_{tt}^{*} é a soma total da variável dependente em todos os blocos e em todos os tratamentos.

Voltando ao caso do exemplo da dureza dos compósitos, imagine agora que os dados estariam na Tabela 7.

Tabela 7. Planejamento Aleatório com Blocos Incompletos para a Dureza de um Compósito

Tipo de Ponteira	Peça			
	1	2	3	4
1	9,3	9,4	9,6	10,0
2	9,4	9,3	--	9,9
3	9,2	9,4	9,5	9,7
4	9,7	9,6	10,0	10,2

Utilizando a Equação (51), tem-se:

$$y_{ij}^{*} = \frac{4(28{,}6) + 4(29{,}1) - 144{,}20}{(3)(3)} = 9{,}6$$

Esse valor está próximo àquele da Tabela 3. A nova tabela da ANOVA é dada a seguir. Como se pode observar, a mesma conclusão seria obtida.

Tabela 8 – Resultados do Planejamento Aleatório com Blocos Incompletos (ANOVA da tabela 7)

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade (d.f.)	Média Quadrática (d.f.)	F	*p-level*
Tratamentos	**39,98**	**3**	**12,83**	**17,12**	**0,0008**
Blocos	**79,53**	**3**	**27,50**		
Erro	**6,22**	**8**	**0,89**		
Total	**125,73**	**14**			

No caso de muitas ausências, elas poderiam ser estimadas como antes, ou seja, achando o mínimo da média quadrática do erro, em relação a cada valor inexistente e resolvendo o sistema resultante de equações algébricas.

6.4. Planejamento usando o Quadrado Latino

Foi visto anteriormente a importância do conceito de blocagem de uma variável. Imagine agora que se tenha mais de uma variável que confira aleatoriedade ao sistema. Suponha que, além de 4 peças diferentes, tenham-se agora 4 operadores diferentes. Por conseguinte, têm-se dois fatores que provocam distúrbios no sistema: tipo de peça e operador. O planejamento adequado para esse tipo de problema consiste em testar cada ponteira exatamente uma vez para cada peça, sendo cada ponteira testada exatamente uma vez para cada operador. A seguinte tabela ajuda a visualizar o problema.

Tabela 9 - Planejamento Quadrado Latino

Peça	Operador			
	1	2	3	4
1	A= 9,3	B= 9,4	C= 9,6	D= 10,0
2	B= 9,4	C= 9,3	D= 9,8	A= 9,9
3	C= 9,2	D= 9,4	A= 9,5	B= 9,7
4	D= 9,7	A= 9,6	B= 10,0	C= 10,2

As letras latinas se referem aos 4 tipos de tratamentos ou ponteiras. Esse tipo de planejamento é dito quadrado, pois o número de colunas deve ser igual ao número de linhas. Ele elimina duas fontes de variabilidade, fazendo a blocagem em duas direções. Os planejamentos mais comuns são 4x4, 5x5 e 6x6.

O modelo estatístico para o quadrado Latino é:

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \varepsilon_{ijk} \quad (53),$$

sendo y_{ijk} a observação da *i*-ésima linha e *k*-ésima coluna para o *j*-ésimo tratamento, μ a média global, α_i o efeito da *i*-ésima linha, τ_j é o efeito do *j*-ésimo tratamento, β_k é o efeito da *k*-ésima coluna e ε_{ijk} é o erro aleatório. Nesse caso, as somas quadráticas são:

$$SQ_T = SQ_{Linhas} + SQ_{Colunas} + SQ_{Tratamentos} + SQ_E \quad (54)$$

As somas quadráticas são calculadas da seguinte forma:

$$SQ_{Tratamentos} = \frac{1}{a}\sum_{j=1}^{a} y_{tjt}^2 - \frac{y_{ttt}^2}{N} \quad (55)$$

$$SQ_{Linhas} = \frac{1}{a}\sum_{i=1}^{a} y_{itt}^2 - \frac{y_{ttt}^2}{N} \quad (56)$$

$$SQ_{Colunas} = \frac{1}{a}\sum_{k=1}^{a} y_{ttk}^2 - \frac{y_{ttt}^2}{N} \quad (57)$$

$$SQ_{Colunas} = \frac{1}{a}\sum_{k=1}^{a} y_{ttk}^2 - \frac{y_{ttt}^2}{N} \qquad (58)$$

$$SQ_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{a}\sum_{k=1}^{a} y_{ijk}^2 - \frac{y_{ttt}^2}{N} \qquad (59)$$

A tabela da ANOVA se torna agora:

Tabela 10 - Tabela da ANOVA para o Planejamento com Quadrado Latino

Fonte de Variação	Soma Quadrática (SQ)	Graus de Liberdade (d.f.)	Média Quadrática (MQ)	F_o
Tratamentos	$SQ_{Tratamentos}$	$a - 1$	$MQ_{Tratamentos} = \frac{SQ_{Tratamentos}}{a-1}$	$F_o = \frac{MQ_{Tratamentos}}{MQ_E}$
Linhas	SQ_{Linhas}	$a - 1$	$MQ_{Linhas} = \frac{SQ_{Linhas}}{a-1}$	
Colunas	$SQ_{Colunas}$	$a - 1$	$MQ_{Colunas} = \frac{SQ_{Colunas}}{a-1}$	
Erro	SQ_{Erro}	$(a-2)(a-1)$	$MQ_{Erro} = \frac{SQ_{Erro}}{(a-2)(a-1)}$	
Total	SQ_{Total}	$a^2 - 1$	$MQ_T = \frac{SQ_T}{a^2-1}$	

Tabela 11 - Valores das Médias Marginais para o Exemplo 9

Means and Standard Deviations (latin.sta)

EXPERIM. DESIGN: 4 by 4 Latin Square
REDUCTIN; Mean = 20,0000 Sigma = 4,44222

Effect	Level	Means	Paramet. Estimate	Std.Dev.
DRIVER	ONE	23,00000	3,00000	2,943920
	TWO	24,00000	4,00000	3,162278
	THREE	15,00000	-5,00000	1,414214
	FOUR	18,00000	-2,00000	2,449490
CAR	AUDI	19,00000	-1,00000	3,651484
	MERCEDES	20,00000	-,00000	6,976150
	TOYOTA	19,00000	-1,00000	2,000000
	CHRYSLER	22,00000	2,00000	4,966555
ADDITIVE	A_ONE	18,00000	-2,00000	2,943920
	A_TWO	22,00000	2,00000	5,597619
	A_THREE	21,00000	1,00000	5,228129
	A_FOUR	19,00000	-1,00000	4,242640

A informação contida nessa tabela é melhor visualizada através do seguinte gráfico.

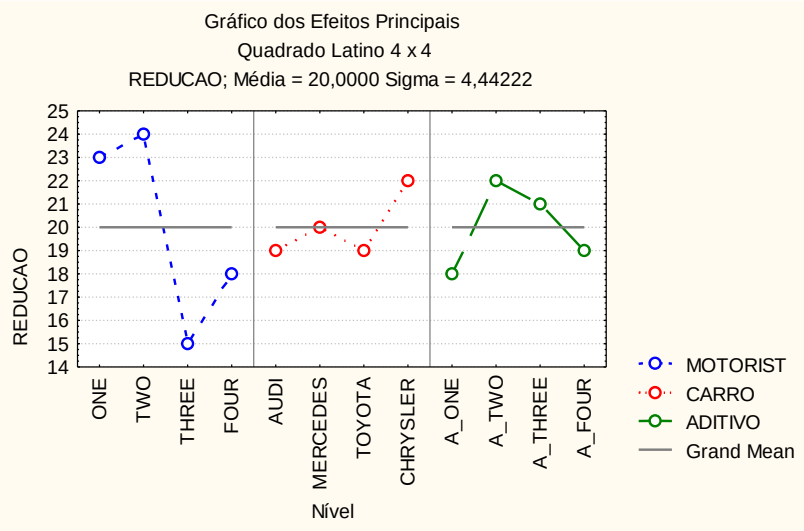


Figura 1 – Gráfico das Médias da Tabela 11.

Várias conclusões qualitativas podem ser tiradas:

a) Considerando a variável motorista, nota-se uma diferença enorme entre os motoristas 2 e 3.
b) O motorista 3 é o que pior dirige.
c) O carro da marca Chrysler é o que tem melhor redução no consumo de combustível. Pode-se dizer que não existe diferença grande entre os outros modelos.
d) O aditivo 2 é o melhor.

Esse tipo de planejamento não permite fazer uma análise cruzada; ou seja, fazer uma combinação entre todas as variáveis. Não se consegue saber se poderia se obter uma maior redução no consumo de combustíveis se outras combinações fossem feitas. Isso será feito mais adiante, quando se estudar planejamento fatorial.

A Tabela 14 apresenta a análise de variância, mostrando que o tipo de aditivo influencia a redução no consumo de combustíveis. Pode-se ver também que o tipo de motorista afeta e a sua blocagem é importante, o mesmo não acontecendo com o tipo de carro.

Tabela 14 - Tabela da ANOVA para os Dados da Tabela 11

Analysis of Variance (latin.sta)

EXPERIM. DESIGN	4 by 4 Latin Square REDUCAO; Mean = 20,0000 Sigma = 4,44222				
Effect	SS	df	MS	F	p
MOTORIST	216,0000	3	72,00000	27,00000	,000699
CARRO	24,0000	3	8,00000	3,00000	,116960
ADITIVO	40,0000	3	13,33333	5,00000	,045197
Residual	16,0000	6	2,66667		

6.5. Planejamento Fatorial

Os planejamentos fatoriais são amplamente utilizados em experimentos envolvendo vários fatores onde é necessário estudar o efeito conjunto destes fatores na resposta.

Os que serão abordados nesse trabalho serão: 2^2 , 2^3 e 2^k.

6.5.1. Fatorial 2^2

Nesse caso, tem-se 2 fatores cada um com dois níveis, produzindo 4 tratamentos ((1), a, b e ab).

A	B		TOTAL
	b_1	b_2	
a_1	(1)	b	(1) + b
a_2	a	ab	a + ab
TOTAL	(1)+a	b+ab	

A estimativa dos efeitos fatoriais (efeitos médios) é dada por:

$$A = \frac{1}{2}\left[\frac{(a-(1))}{r} + \frac{(ab-b)}{r}\right] = \frac{1}{2r}\left[(a+ab)-((1)+b)\right] \qquad (60)$$

$$B = \frac{1}{2}\left[\frac{(b-(1))}{r} + \frac{(ab-a)}{r}\right] = \frac{1}{2r}\left[(b+ab)-((1)+a)\right] \qquad (61)$$

$$AxB = \frac{1}{2}\left[\frac{(ab-b)}{r} - \frac{(a-(1))}{r}\right] = \frac{1}{2r}\left[((1)+ab)-(a+b)\right] \qquad (62)$$

O quadro de sinais (coeficientes dos contrastes) para obtenção dos Efeitos é:

Combinação de Tratamento	Efeito Fatorial			
	I	A	B	AB
(1)	+	-	-	+
a	+	+	-	-
b	+	-	+	-
ab	+	+	+	+

As somas de quadrados dos efeitos fatoriais são dadas por:

$$SQA = \frac{[(a+ab)-((1)+b)]^2}{4r} \qquad (63)$$

$$SQB = \frac{[(b+ab)-((1)+a)]^2}{4r} \qquad (64)$$

$$SQAxB = \frac{[((1)+ab)-(a+b)]^2}{4r} \qquad (65)$$

$$SQtotal = \sum_{ijk} Y_{ijk}^2 - \frac{Y_{...}^2}{4r} \qquad (66)$$

$$SQE = SQtotal - SQA - SQB - SQAB \qquad (67)$$

Exemplo de Fatorial 2^2:

Fator A: efeito de concentração do reagente: níveis de 15% (baixo) e 25% (alto).
Fator B: presença de catalisador: ausência (baixo) e presença (alto).
Resposta: tempo de reação de um processo químico.
Nº de repetições: 3.

	Repetição			
Tratamentos	1	2	3	Total
A baixo, B baixo ⇔ (1)	28	25	27	80
A alto, B baixo ⇔ a	36	32	32	100
A baixo, B alto ⇔ b	18	19	23	60
A alto, B alto ⇔ ab	31	30	29	90

Total=330

Vamos calcular a estimativa dos efeitos médios, utilizando das equações 60, 61 e 62.

$$A = \frac{1}{2(3)}[(100+90)-(80+60)] = \frac{50}{6} = 8{,}33$$

$$B = \frac{1}{2(3)}\left[(60+90)-(80+100)\right] = \frac{-30}{6} = -5{,}00$$

$$AxB = \frac{1}{2(3)}\left[(80+90)-(100+60)\right] = \frac{10}{6} = 1{,}67$$

Vamos calcular as somas dos quadrados dos efeitos fatoriais, utilizando-se as equações 63, 64, 65, 66 e 67.

$$SQA = \frac{(50)^2}{4(3)} = 208{,}33$$

$$SQB = \frac{(-30)^2}{4(3)} = 75{,}00$$

$$SQAxB = \frac{(10)^2}{4(3)} = 8{,}33$$

SQErro = SQTotal - SQA-SQB-SQAxB = 323,00 - 208,33 - 75,00 - 8,33 = 31,34

$$SQTotal = \sum_{ijk} Y_{ijk}{}^2 - \frac{Y...^2}{4(3)} = 28^2 + ...29^2 - \frac{330^2}{4(3)} =$$

$$9398 - 9075 = 323$$

Resumindo:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
A	208.33	1	208.33	53.15**
B	75.00	1	75.00	19.13**
AB	8.33	1	8.33	2.13
Erro	31.34	8	3.92	
Total	323.00	11		

**Significativo a 1%

6.5.2. Fatorial 2^3

Nesse caso tem-se 3 fatores cada um com 2 níveis, produzindo 8 tratamentos ((1), a, b, c, ab, ac, bc e abc).

A estimativa dos efeitos fatoriais (efeitos médios) é dada por:

$$A = \frac{1}{4}\left[\frac{(a-(1))}{r} + \frac{(ab-b)}{r} + \frac{(ac-c)}{r} + \frac{(abc-bc)}{r}\right] \quad (68)$$
$$= \frac{1}{4r}\left[a + ab + ac + abc - ((1) + b + +c + bc)\right]$$

$$B = \frac{1}{4r}\left[b + ab + bc + abc - (1) - a - c - ac\right] \quad (69)$$

$$C = \frac{1}{4r}\left[c + ac + bc + abc - (1) - a - b - ab\right] \quad (70)$$

$$AB = \frac{1}{2}\left\{\frac{1}{2}\left[\frac{(ab-b)}{r} - \frac{(a-(1))}{r}\right] + \frac{1}{2}\left[\frac{abc-bc}{r} - \frac{(ac-c)}{r}\right]\right\} \quad (71)$$
$$= \frac{1}{4r}\left[ab - b - a + (1) + abc - bc - ac + c\right]$$

$$AC = \frac{1}{4r}\left[(1) - a + b - ab - c + ac - bc + abc\right] \quad (72)$$

$$BC = \frac{1}{4r}\left[(1) + a - b - ab - c - ac + bc + abc\right] \quad (73)$$

$$ABC = \frac{1}{4r}\{[abc - bc] - [ac - c] - [ab - b] + [a - (1)]\} \qquad (74)$$

$$= \frac{1}{4r}[abc - bc - ac + c - ab + b + a - (1)]$$

O quadro de sinais para obtenção dos efeitos é:

Combinação de Tratamento	Efeito Fatorial							
	I	A	B	AB	C	AC	BC	ABC
(1)	+	-	-	+	-	+	+	-
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
ab	+	+	+	+	-	-	-	-
c	+	-	-	+	+	-	-	+
ac	+	+	-	-	+	+	-	-
bc	+	-	+	-	+	-	+	-
abc	+	+	+	+	+	+	+	+

A soma de quadrados dos efeitos fatoriais é dada por:

$$SQ_{efeitofatorial} = \frac{(contraste)^2}{8r} \qquad (75)$$

Exemplo de Fatorial 2^3:

Fator A: efeito da porcentagem de gaseificação: 10% e 12%
Fator B: pressão de operação no enchimento: 25 psi e 25 psi
Fator C: velocidade da esteira: 200 e 250
Resposta: volume de bebida gaseificada embalada em cada garrafa
Nº de repetições: 2

% de Gaseificação	Pressão de Operação (B)			
	25 psi		30 psi	
	Velocidade da Esteira (C)		Velocidade da Esteira (C)	
	200	250	200	250
10	-3 -1	-1 -0	-1 -0	1 1
	- 4 = (1)	- 1 = c	- 1 = b	2 = bc
12	0 1	2 1	2 3	6 5
	1 = a	3 = ac	5 = ab	11 = abc

As estimativas calculadas dos efeitos médios foram calculadas através das equações :

$$A = \frac{1}{4r}\left[a - (1) + ab - b + ac - c + abc - bc\right]$$

$$= \frac{1}{8}\left[1 - (-4) + 5 - (-1) + 3 - (-1) + 11 - 2\right] = \frac{1}{8}\left[24\right] = 3.00$$

$$B = \frac{1}{4r}\left[b + ab + bc + abc - (1) - a - c - ac\right]$$

$$= \frac{1}{8}\left[-1 + 5 + 2 + 11 - (-4) - 1 - (-1) - 3\right] = \frac{1}{8}\left[18\right] = 2.25$$

$$C = \frac{1}{4r}\left[c + ac + bc + abc - (1) - a - b - ab\right]$$

$$= \frac{1}{8}\left[-1 + 3 + 2 + 11 - (-4) - 1 - (-1) - 5\right] = \frac{1}{8}\left[14\right] = 1.75$$

$$AB = \frac{1}{4r}\left[ab - a - b + (1) + abc - bc - ac + c\right]$$

$$= \frac{1}{8}\left[5 - 1 - (-1) + (-4) + 11 - 2 - 3 + (-1)\right] = \frac{1}{8}\left[6\right] = 0.75$$

$$AC = \frac{1}{4r}\left[(1) - a + b - ab - c + ac - bc + abc\right]$$

$$= \frac{1}{8}\left[-4 - 1 + (-1) - 5 - (-1) + 3 - 2 + 11\right] = \frac{1}{8}\left[2\right] = 0.25$$

$$BC = \frac{1}{4r}\left[(1) + a - b - ab - c - ac + bc + abc\right]$$

$$= \frac{1}{8}\left[-4 + 1 - (-1) - 5 - (-1) - 3 + 2 + 11\right] = \frac{1}{8}\left[4\right] = 0.50$$

$$ABC = \frac{1}{4r}\left[abc - bc - ac + c - ab + b + a - (1)\right]$$

$$= \frac{1}{8}\left[11 - 2 - 3 + (-1) - 5 + (-1) + 1 - (-4)\right] = \frac{1}{8}\left[4\right] = 0.50$$

As somas de quadrados dos efeitos Fatoriais são:

$$SQA = \frac{(24)^2}{16} = 36.00$$

$$SQB = \frac{(18)^2}{16} = 20.25$$

$$SQC = \frac{(14)^2}{16} = 12.25$$

$$SQAB = \frac{(6)^2}{16} = 2.25$$

$$SQAC = \frac{(2)^2}{16} = 0.25$$

$$SQBC = \frac{(4)^2}{16} = 1.00$$

$$SQABC = \frac{(4)^2}{16} = 1.00$$

SQTotal = 78.50 SQErro = 5.50

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Percentagem de Gaseificação (A)	36.00	1	36.00	57.14**
Pressão (B)	20.25	1	20.25	32.14**
Velocidade da Esteira (C)	12.25	1	12.25	19.44**
AB	2.25	1	2.25	3.57
AC	0.25	1	0.25	0.40
BC	1.00	1	1.00	1.59
ABC	1.00	1	1.00	1.59
Erro	5.00	8	0.63	
Total	78.00	15		

**Significativo a 1%

6.5.3. Fatorial 2^k

Os métodos de análise podem ser generalizados para o caso do fatorial 2^k (*k* fatores com 2 níveis).

Assim, o contraste, AB...K = (a ± 1) (b ± 1)...(k ± 1).

Por exemplo, o contraste AB no fatorial 2^3 é dado por:

(a-1) (b-1) (c+1) = abc+ab+c+(1)-ac-bc-a-b

As somas de quadrados dos efeitos fatoriais são dadas por:

$$SQ_{efeitofatorial} = \frac{(contrasteAB...K)^2}{r2^K}$$

A tabela de análise de variância tem a seguinte estrutura geral; supondo o Delineamento Completamente Casualizado na aleatorização dos Tratamentos.

CAUSAS DE VARIAÇÃO	GL
K efeitos principais	
A	1
B	1
.	.
.	.
.	.
K	1
C_k^2 INTERAÇÕES SIMPLES	
AB	1
AC	1
.	.
.	.
.	.
JK	1
C_k^3 INTERAÇÕES TRÍPLICES	
ABC	1
ABD	1
.	.
.	.
.	.
1 INTERAÇÃO DE K FATORES	
ABCD...K	1
ERRO EXPERIMENTAL	$2^k(r-1)$
TOTAL	$r2^k - 1$

6.5.3.1 Fatorial 2^k com 1 repetição

O nº de tratamentos em um delineamento fatorial 2k aumenta com o número de fatores.

Nesses casos é impossível obter uma estimativa propriamente dita do erro experimental.

Para poder testar os efeitos fatoriais considera-se as interações de ordem elevada desprezíveis e assume-se que as mesmas produzem uma estimativa do erro experimental.

Exemplo de Fatorial 2^k com repetição:

Fator A: temperatura: A0; A1
Fator B: pressão: B0; B1
Fator C: concentração de reagente:C0; C1
Fator D: taxa de mistura: D0; D1
Resposta: a influência de fatores (quatro) na taxa de filtração de um produto químico
Nº de repetições: 1

Resumindo:

	A_0				A_1			
	B_0		B_1		B_0		B_1	
	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1
D_0	45	68	48	80	71	60	65	65
D_1	43	75	45	70	100	86	104	96

Vamos assumir que as interações tríplices e quádruplas são desprezíveis.

O SQ do erro é calculado por:

SQErro= SQABC+SQABD+SQACD+SQBCD+SQABCD (76)

(com 5 gl)

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
A	1870.56	1	1870.56	73.15
B	39.06	1	39.06	1.53
C	390.06	1	390.06	15.25*
D	855.56	1	855.56	33.46**
AB	0.06	1	0.06	< 1
AC	1314.06	1	1314.06	51.39**
AD	1105.56	1	1105.56	43.24**
BC	22.56	1	22.56	< 1
BD	0.56	1	0.56	< 1
CD	5.06	1	5.06	< 1
Erro	127.84	5	25.57	
Total	5730.94	15		

ABC, ABD,ACD,BCD,ABCD são as interações desprezíveis

	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD
1)	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+
a	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-
b	-	+	-	-	+	-	+	-	+	-	+	+	-	+	-
ab	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+
c	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-
ac	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+
bc	-	+	-	+	-	+	-	-	+	-	+	-	+	-	+
abc	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
d	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-
ad	+	-	-	-	-	+	+	+	+	-	-	-	-	+	+
bd	-	+	-	-	+	-	+	+	-	+	-	-	+	-	+
abd	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-
cd	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+
acd	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-
bcd	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
abcd	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Algoritmo de Yates

Tratamento	Resposta	(1)	(2)	(3)	(4)	estimativa do efeito	SQ	Efeito
(1)	45	116	229	502	1121	-	-	-
a	71	113	273	619	173	21,23	1870,5625	A
b	48	128	292	20	25	33,13	39,0625	B
ab	65	145	327	153	1	0,13	0,0625	AB
c	68	143	43	14	79	9,88	390,0625	C
ac	60	149	-23	11	-145	-18,13	1314,0622	AC
bc	80	111	116	-16	19	2,38	22,5625	BC
abc	65	166	37	17	15	1,88	14,0625	ABC
d	43	26	-3	44	117	14,63	855,5625	D
ad	100	17	17	35	133	16,63	1105,5625	AD
bd	45	- 8	6	-66	-3	-0,38	0,5625	BD
abd	104	- 15	5	-79	33	4,13	68,0625	ABD
cd	75	57	-9	20	-9	-4,13	5,0625	CD
acd	86	59	-7	-1	-13	-1,63	10,5625	ACD
bcd	70	11	2	2	-21	-2,63	27,5625	BCD
abcd	96	26	15	13	11	1,38	7,5625	ABCD

coluna (1): 1a metade soma dos adjacentes na coluna resposta
2a metade segundo-primeiro na coluna resposta
coluna (2): idem na coluna (1)
coluna (3): idem na coluna (2)
coluna (4): idem na coluna (3)

Efeito: $(4) \div r\,2^{k-1} \Rightarrow (4) \div 8$

$\uparrow\uparrow$

$1\ \ 2^3$

Comentários:

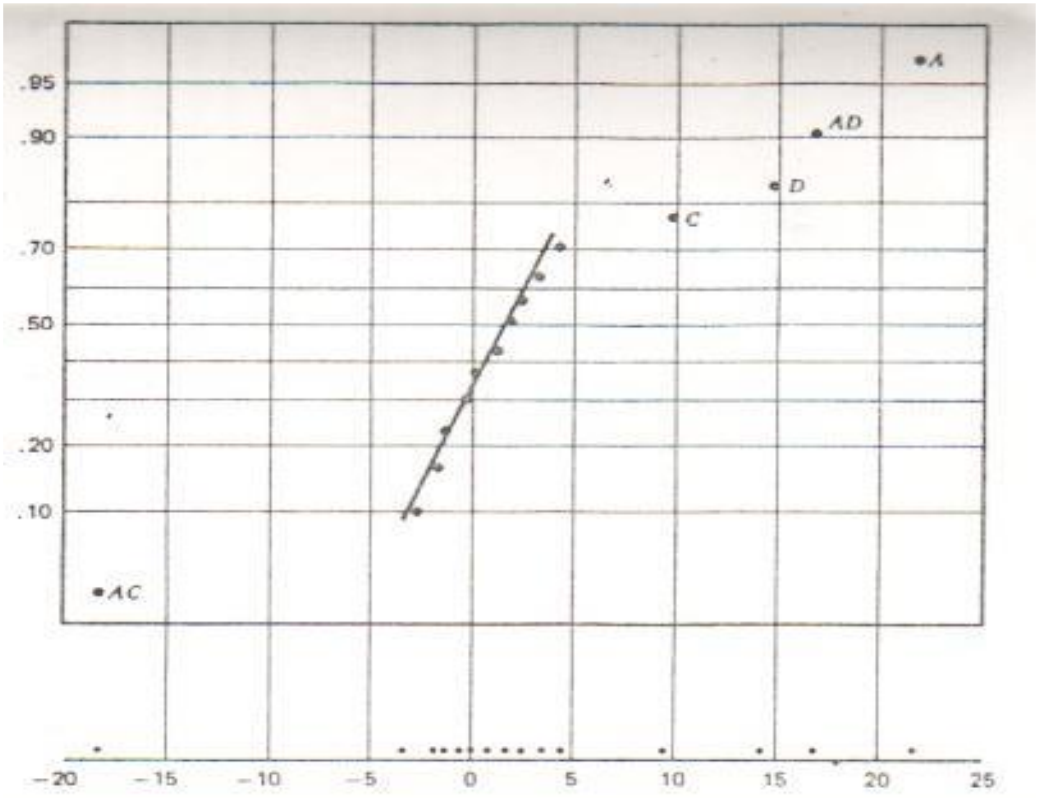
As interações de ordem elevada poderão não ser desprezíveis. Mas como saber quais são ou não são desprezíveis?

Uma maneira simples de verificar se os efeitos são desprezíveis seriam plotar as estimativas dos efeitos em papel de probabilidade normal. Os efeitos desprezíveis são normalmente distribuídos e estarão numa reta num gráfico de probabilidade normal.

Então:

Cálculos para construção do gráfico de Probabilidade Normal

Ordem (j)	Efeito	(Eixo x):Estimativa	(Eixo y):(j - .5)/15
15	A	21,23	.9667
14	AD	16,63	.9000
13	D	14,63	.8333
12	C	9,88	.7667
11	ABD	4,13	.7000
10	B	3,13	.6333
9	BC	2,38	.5667
8	ABC	1,88	.5000
7	ABCD	1,38	.4333
6	AB	0,13	.3667
5	CD	-0,38	.3000
4	BD	-1,13	.2333
3	ACD	-1,63	.1667
2	BCD	-2,63	.1000
1	AC	-18,13	.0333



Comentários:

Efeitos pequenos → sobre uma reta
Efeitos Grandes → fora da reta
Interações tríplices e quádruplas sobre a reta → desprezíveis
Desde que o efeito de B (pressão) é não sig. e todas interações que envolvem B são desprezíveis podemos descartar B do experimento e analisar como se fosse um experimento 2^3 com os fatores A, C e D com 2 repetições.

Assumindo que o fator B é desprezível

C. Variação	GL	SQ	QM	F
A	1	1870,56	1870,56	83,36**
C	1	390,06	390,06	17,35**
D	1	855,56	855,56	38,13**
AC	1	1314,06	1314,	58,56**
AD	1	1105,56	1105,56	49,27**
CD	1	5,06	5,06	<1
ACD	1	10,56	10,56	<1
Erro	8	179,52	22,44	
TOTAL	15	5730,94		

6.5.3.2. Adição de pontos centrais do planejamento fatorial 2^k

Um aspecto importante a ser observado é a suposição da linearidade em delineamentos 2k. É preciso verificar se pode sustentar que o modelo é linear (1ª ordem) ou se há possibilidade de ser quadrático (2ª ordem).

Quando rodamos um delineamento 2k assumimos antecipadamente um ajuste linear, entretanto se as variáveis explicativas forem quantitativas há possibilidade de esta relação não ser dessa ordem

Uma maneira de nos preservarmos quanto à possibilidade de ser um modelo de segunda ordem é adicionado pontos centrais no delineamento 2k.

Uma importante razão para adicionarmos pontos centrais é o fato de eles não impactarem em delineamentos 2k.

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum\sum_{i<j} \beta_{ij} x_i x_j + \varepsilon \qquad (77)$$

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum\sum_{i<j} \beta_{ij} x_i x_j + \sum_{j=1}^{k} \beta_{ij} x_j^2 + \varepsilon \qquad (78)$$

$$SS_{purequadratic} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C} \qquad (79)$$

$$H_0 : \sum_{j=1}^{k} \beta_{ij} = 0$$

$$H_1 : \sum_{j=1}^{k} \beta_{ij} \neq 0$$

Exemplo para Adição de pontos centrais ao planejamento 2^k

Engenheiro químico está estudando um processo, com 2 variáveis de interesse.
Ele não tem certeza que a suposição de linearidade está satisfeita.
Decide conduzir um experimento 2^k com uma repetição, aumentando 5 pontos centrais.

$$MSE = \frac{SS_E}{n_C - 1} = \frac{\sum (y_i - \bar{y})^2}{centerpoints} = \frac{\sum (y_i - 40{,}46)^2}{n_C - 1} = \frac{0{,}1720}{4} = 0{,}0430 \qquad (80)$$

Média dos ptos centrais=40,46
Média dos ptos do delin. Fatorial=40,425
40,425 - 40,46 = -0,035 (pequeno)

Exemplo para Adição de pontos centrais ao planejamento 2^k

$$SS_{purequadratic} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C} \qquad (81)$$

$$SS_{purequadratic} = \frac{(4)(5)(-0{,}035)^2}{4+5} = 0{,}0027 \qquad (82)$$

$$H_0 = \beta_{11} + \beta_{22} = 0$$

A hipótese nula não pode ser rejeitada
Conclusão: o modelo é de 1ª ordem (linear)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

C.V.	G.L.	Soma dos quadrados	Quadrado médio	F_0	P-Value
A(tempo)	1	2,4025	2,4025	55,87	0,0017
B(temperatura)	1	0,4225	0,4225	9,83	0,0350
AB	1	0,0025	0,0025	0,06	0,8185
Quadrático	1	0,0027	0,0027	0,06	0,8185
Erro	4	0,1720	0,0430		
Total	8	3,0022			

6.6. Experimentos (Resumo)

6.6.1. Projeto e Análise de Experimentos

Planejar o experimento para se ter informações suficientes (em termos dos objetivos da pesquisa) com o menor número possível de ensaios.

Analisar os dados de forma compatível com o projeto experimental realizado.

6.6.2. Estratégias no Planejamento de Experimentos

1. Reconhecer, estabelecer e delimitar claramente o problema;

2. Identificar os possíveis fatores que podem afetar o problema em estudo;

3. Verificar quais fatores que poderão ser mantidos fixos e, portanto, não terão os seus efeitos avaliados no estudo experimental;

4. Identificar, para cada fator, o intervalo de variação e os níveis que entrarão no estudo;

5. Escolher um projeto experimental adequado, isto é, saber como combinar os níveis dos fatores de forma que se possa resolver o problema proposto com o menor custo possível;

6. Escolher a resposta adequada, ou seja, a variável Y que mede adequadamente o resultado (a qualidade, o desempenho, etc.) do processo; o planejamento de como será a análise dos dados do experimento;

6.6.3. Roteiro para a realização de um Experimento

1. Identificar e estabelecer o problema;
2. Escolha dos fatores (k) e de seus níveis (b);
3. Seleção da variável resposta (y);
4. Escolha do projeto experimental;
5. Realização do experimento;
6. Análise estatística dos dados (significado x relevante).
7. Conclusões e recomendações.

6.6.4. Estudo Experimental

Manipula-se de forma planejada, certas *variáveis independentes* ou *fatores* (*A, B, C,...*) para verificar o efeito que esta manipulação provoca numa certa *variável dependente* ou *resposta Y.*

6.7. Exemplos

A)Exemplos do Tipo 2k:

Ex 1:

- Encontrar a melhor condição de operação de um processo químico;

- A resposta Y pode ser o rendimento da reação química e os fatores podem ser: o tempo de reação (A) e a temperatura de reação (B).

Ex 2:

- Verificar quais são os fatores que mais interferem na resistência à compressão (Y) de um concreto. Os fatores a serem estudados podem ser: tempo de hidratação; a dosagem (A); a qualidade do cimento (C) e o uso de aditivos (D).

B)Projeto de experimentos fatorial do tipo 2k:

Supõe-se aleatorização, variação aleatória (erro experimental) com distribuição normal, para fatores qualitativos, supõem-se efeitos lineares e dados balanceados (mesmo número de observações em cada combinação dos níveis dos fatores).

7. Resumo

Classificação dos Planejamentos de Experimentos	**Aplicação**	**Estrutura**	**Informações obtidas**
Completamente Aleatorizado com um único fator	Apropriado quando somente um fator experimental está sendo estudado	O efeito do fator é estudado por meio da alocação ao acaso das unidades experimentais aos tratamentos (níveis do fator). Os ensaios são realizados em ordem aleatória.	Estimativa e comparações dos efeitos dos tratamentos Estimativas da variância
Fatorial	Apropriado quando vários fatores devem ser estudados em dois ou mais níveis e as interações entre os fatores podem ser importantes	Em cada repetição completa do experimento todas as combinações possíveis dos níveis dos fatores (tratamentos) são estudadas. A alocação das unidades experimentais aos tratamentos e a ordem de realização dos ensaios são feitas de modo aleatório.	Estimativas e comparações dos efeitos dos fatores Estimativa dos possíveis efeitos de interações Estimativa da variância
Fatorial 2k em blocos	Apropriado quando o número de ensaios necessários para o planejamento em k fatores em 2 níveis é muito grande para que sejam realizados sob condições homogêneas	O conjunto completo de tratamentos é divido em subconjuntos de modo que as interações de ordem mais alta são confundidas com os blocos. São tomadas observações em todos os blocos. Os blocos surgem geralmente como conseqüência de restrições de tempo, homogeneidade de materiais, etc.	Fornece as mesmas estimativas do planejamento fatorial, exceto algumas interações de ordem mais alta que não podem ser estimadas porque estão confundidas com os blocos.

Classificação dos Planejamentos de Experimentos	Aplicação	Estrutura	Informações obtidas
Fatorial 2k fracionário	Apropriado quando existem muitos fatores (k muito grande) e não é possível coletar observações em todos os tratamentos	Vários fatores são estudados em dois níveis, mas somente um subconjunto do fatorial completo é executado. A formação dos blocos algumas vezes é possível.	Estimativas e comparações dos efeitos de vários fatores Estimativa de certos efeitos de interação (alguns efeitos podem não ser estimáveis) Certos planejamentos fatoriais fracionários quando k é pequeno) não fornecem informações suficientes para estimar a variância
Blocos aleatorizados	Apropriado quando o efeito de um fator está sendo estudado e é necessário controlar a variabilidade provocada por fatores perturbadores conhecidos. Estes fatores perturbadores (material, tempo, pessoas, etc.) são divididos em blocos ou grupos homogêneos	São tomadas observações correspondentes a todos os tratamentos (níveis do fator) em cada bloco. Usualmente os blocos são considerados em relação a um único fator perturbador.	Estimativas e comparações dos efeitos dos tratamentos livres dos efeitos do bloco Estimativas dos efeitos do bloco Estimativa da variância

Classificação dos Planejamentos de Experimentos	Aplicação	Estrutura	Informações obtidas
Blocos Incompletos Balanceados	Apropriado quando todos os tratamentos não podem ser acomodados em um bloco	Os tratamentos testados em cada bloco são selecionados de forma balanceada: dois tratamentos quaisquer aparecem juntos em um mesmo bloco o mesmo número de vezes que qualquer outro par de tratamentos	Idêntico ao planejamento em blocos aleatorizados. Os efeitos de todos os tratamentos são estimados com igual precisão
Blocos Incompletos Parcialmente Balanceados	Apropriado quando um planejamento em blocos incompletos balanceados necessita de um número de blocos excessivamente grandes	Alguns pares de tratamentos aparecem juntos n1 vezes, outros pares aparecem juntos n2 vezes,..., e os pares restantes aparecem juntos m vezes.	Idêntico ao planejamento em blocos aleatorizados, mas os efeitos dos tratamentos são estimados com diferentes precisões
Quadrados de Youden	Similares aos quadrados latinos, mas o número de linhas, colunas e tratamentos não precisam ser iguais	Cada tratamento ocorre uma vez em cada linha. O número de tratamentos deve ser igual ao número de colunas. Os blocos são formados em relação a duas variáveis perturbadoras	Idêntico ao planejamento em quadrados latinos

Classificação dos Planejamentos de Experimentos	Aplicação	Estrutura	Informações obtidas
Quadrados Latinos	Apropriado quando um fator de interesse está sendo estudado e os resultados podem ser afetados por duas outras variáveis experimentais ou por duas fontes de heterogeneidade. É suposta a ausência de interações	O quadrado latino é um arranjo para permitir dois grupos de restrições de bloco. Os tratamentos são distribuídos em correspondência à s colunas e linhas de um quadrado. Cada tratamento aparece uma vez em cada linha e uma vez em cada coluna. Os números de tratamentos devem ser iguais ao número de linhas e colunas do quadrado. Os blocos são formados em relação a duas variáveis perturbadoras, as quais correspondem à s colunas e linhas do quadrado.	Estimativas e comparações dos efeitos dos tratamentos livres dos efeitos das duas variáveis bloco. Estimativas e comparações dos efeitos das duas variáveis de bloco Estimativa da variância
Hierárquico	Experimentos com vários fatores em que os níveis de um fator (B) são similares mas não idênticos para diferentes níveis de outro fator (A). Ou seja, o j-ésimo nível de B quando A está no nível 1 é deferente do j-ésimo nível de B quando A está no nível 2 e assim por diante	Os níveis do fator B estão aninhados abaixo dos níveis do fator A	Estimativas e comparações dos efeitos dos fatores Estimativa da variância

Classificação dos Planejamentos de Experimentos	Aplicação	Estrutura	Informações obtidas
Superfície de resposta	O objetivo consiste em fornecer mapas empíricos ou gráficos de contorno. Estes mapas ilustram a forma pela qual os fatores, que podem ser controlados pelo pesquisador, influenciam a variável resposta	Os níveis dos fatores são vistos como pontos no espaço de fatores (muitas vezes multidimensional) no qual a resposta será registrada	Mapas que ilustram a natureza e a forma da superfície de resposta

8. Referências Bibliográficas

Error, Measurements and results in chemical analysis; K. Eckschlager, M. Sc., Ph.

Planejamento e otimização de experimentos - Benício de Barros Neto, Ieda Spacino Scarminio, Roy Edward Bruns,

Design and analysis of experiments - Douglas C. Montgomery,

Planejamento de Experimentos utilizando o Statistica – Verônica Calado e Douglas Montgomery.

Apostila de Técnicas de Análise Multivariada da Statsoft.

http://alea-estp.ine.pt/html/statofic/html/*dossiê*/html/*dossiê*.html

http://www.inf.ufsc.br/~patrec/agrupamentos.html

Cerqueira, E.O; Poppi, R.J.; Kubota, L.; Mello. C., *Quim. Nova, 23,* 2000, 690- 698.

Galvão, R. K. H.; Araújo, M. C. U.; Saldanha, T. C. B.; Visane, V.; Pimentel, M.F. “Estudo Comparativo Sobre Filtragem de Sinais Instrumentais usando Transformadas de Fourier e wavelet” , *Quim. Nova*, 24, 2001, 874-884

D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbeke, "Handbook of Chemometrics and Qualimetrics: Part B", Elsevier, Amsterdam, 1998.

M. Otto, "Chemometrics -Statistics and Computer Application in Analytical Chemistry", Wiley-VCH, Weinheim, 1999.

H. Martens e T. Naes, "Multivariate Calibration", Wiley, New York, 1991.

B. Barros Neto, I. S. Scarminioe R. E. Bruns, “Como Fazer Experimentos. Pesquisa e desenvolvimento na ciência e na indústria”, 2a ed., Ed. da UNICAMP, Campinas, 2003.

Hair Anderson Tatham Black, “Analise Multivariada De Dados” - Bookman Editora

Maria Célia Garcia Alvarez, “Quimiometria”.

9. Sumário de Revisões

Edição	*Item (s) Revisado (s)*	*Responsável pela Revisão*	*Data da Revisão*
01	*Emissão Inicial*	*Olivia Woyames Pinto*	*11/12/2009*