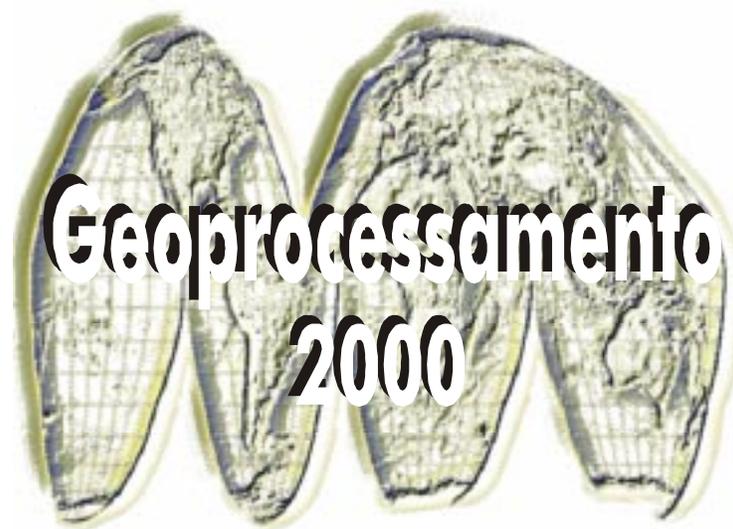


Apostila I

Estatística Básica



**Geoprocessamento
2000**

UFMG

Alexandre Diniz

1.0 INTRODUÇÃO À ESTATÍSTICA

1.1 Conceitos básicos:

- **Método**

Na Grécia antiga, *methodos*, significava caminho para se chegar a um fim.

Método – é o conjunto de etapas, ordenadamente dispostas, a serem vencidas:

- . na investigação da verdade;
- . no estudo de uma ciência;
- . ou para alcançar um determinado fim.

- **Técnica**

Modo de fazer de forma mais hábil, mais segura e mais perfeita algum tipo de atividade, arte ou ofício.

- **Conhecimento**

Conhecer é estabelecer uma relação entre a pessoa que conhece e o objeto que passa a ser conhecido.

No processo do conhecimento, o sujeito se apropria do objeto – processando-o mentalmente

Conhecer = transformar o objeto em conceito, reconstituindo-lhe em sua mente (semiótica).

Tipos de conhecimento:

- . vulgar ou empírico;
- . filosófico;
- . teológico/dogmático;
- . científico.

Dois métodos de raciocínio científico:

indução e dedução.

- **Indução**

- . Vai do particular para o geral;
- . vai dos fatos para as idéias;
- . vai das observações para as generalizações.

- **Dedução**

- . O raciocínio dedutivo parte do geral para chegar ao particular;
- . do universal para chegar ao singular;
- . das idéias para os fatos;
- . das generalizações para a observação.

- **Estatística**

Originalmente – coleção de informações de interesse para o estado sobre a população e economia.

As palavras estatística e estado têm a mesma origem latina: status.

Desenvolveu para tornar-se um método de análise muito utilizado nas ciências sociais e naturais.

- **População**

Coleção de todas as observações potenciais sobre um determinado fenômeno.

- **Amostra**

Conjunto de dados efetivamente observados ou extraídos de uma população.

Sobre os dados da amostra se desenvolvem os estudos, visando a fazer inferências sobre a população.

- **Amostragem**

. Processo de escolha da amostra;
. parte inicial de qualquer estudo estatístico;
. consiste na escolha criteriosa de elementos a serem submetidos ao estudo, para que os resultados sejam representativos, toma-se o cuidado de entrevistar um conjunto de pessoas com características sócio-econômicas, culturais, religiosas etc. tão próximas quanto possível da população.

A escolha da amostra, construção dos instrumentos, entrevistas, codificação dos dados e apuração dos resultados são etapas deste tipo de pesquisa.

1.2 Grandes áreas da estatística:

. Amostragem e planejamento de experimentos = coleta de dados.
. Estatística descritiva = organização, apresentação e sintetização de dados.
. Estatística inferencial = o conjunto de métodos para tomada de decisões, nas situações onde existem incerteza e variação.

- **Inferência**

. A tomada de decisões sobre a população, com base em estudos feitos sobre os dados da amostra, constitui o problema central da inferência estatística.
. Tais decisões sempre envolvem um grau de incerteza (probabilidade de erro).
. A inferência é feita com base em um modelo estatístico.

- **Probabilidade**

. Impossível fazer inferências estatísticas sem utilizar alguns resultados da teoria de probabilidades.
. Embora intimamente associada à estatística, tem suas características próprias.

. Busca quantificar a incerteza existente em determinada situação.

1.3 Escalas de mensuração:

- **Mensuração**

Atribuição de um número a qualidades de um objeto ou fenômeno segundo regras definidas.
O processo de atribuição de números a qualidades de objetos, forma a escala de mensuração ou escala de medida.

- **Variáveis**

Características das unidades de análise.

- **Unidades de análise**

Base da análise.
Elementos nos quais se tem interesse.

- **Tipos de variáveis**

Quatro maneiras básicas, ou níveis básicos, de mensuração (quatro tipos de variáveis):

1. nominal
2. ordinal
3. intervalar
4. razão

Importante definir os níveis de mensuração para as variáveis, porque as técnicas de análise estatística que podem ser utilizadas dependem da escala de mensuração.

- **Escala nominal**

O nível mais simples das escalas de medida;
sistema simples de classificação;
utilizada para classificar objetos ou fenômenos em termos de igualdade dos seus atributos e numerá-los;
Recurso para se classificar e rotular ou dar nomes a objetos.

O caso mais simples é formado pela divisão em duas classes que são identificadas com os números zero ou um - variável binária (0,1).
Cada observação na mensuração nominal pertence a uma só classe muito freqüente na análise geográfica;
Indica a presença ou não de determinada característica.

Ex: Municípios dentro e fora da área de atuação da SUDENE.

Características

- . classes são mutuamente excludentes;
- . operações aritméticas não podem ser aplicadas (adição e multiplicação);
- . contagem simples é possível;
- . pode-se levantar a classe modal (mais freqüente);
- . a freqüência de cada classe pode ser expressa como porcentagem do número total.

- **Escala ordinal**

Utilizada quando os fenômenos ou observações podem ser arranjados segundo uma ordenação (grandeza, preferência, importância, distância, etc..).

Ex: expressões qualitativas arranjadas segundo uma ordem:

- . hierarquia dos níveis educacionais: primeiro, segundo e terceiro graus;
- . níveis de renda: renda baixa, média e alta;
- . hierarquia urbana;
- . padrão de habitação;
- . preferência locacional;
- . escala de dureza dos minerais.

Possível quando se desenvolve uma seqüência qualitativa na qual é lógico colocar um fato antes do outro.

- . Não deve fazer operações aritméticas

Ex: classificação de hotéis em níveis hierárquicos.

Não se pode dizer que um hotel quatro estrelas é duas vezes melhor do que um hotel duas estrelas.

Sabe-se que os quatro estrelas são melhores, mas não existe meios de se quantificar esta diferença na escala ordinal.

- **Características:**

- . É possível calcular a frequência de cada classe, para indicar a classe modal;
- . Classes são mutuamente excludente;
- . Pode-se calcular coeficientes de correlação - Spearman e Kendall (estatística não paramétrica).

- **Escala intervalar**

Características:

. Tem todas as características de uma escala ordinal, porém os intervalos entre os valores são conhecidos exatamente e assim cada observação pode receber um valor numérico preciso.

. A extensão de cada intervalo sucessivo é constante:

i.e. numeração dos anos, variações de altitude através de curvas de nível e escalas de temperatura;

. O ponto zero de uma escala de intervalo é arbitrário e não indica ausência da característica medida.

. A falta de zero absoluto é uma desvantagem, pois não é possível afirmar que uma temperatura de 20 °C é duas vezes mais quente do que uma de 10 °C.

. Adapta-se a todas as operações aritméticas usuais, desde que seja mantida a ordem dos objetos e as diferenças relativas entre elas.

. A média e o desvio padrão podem ser calculados.

- **Escala de razão**

Características:

. Mais precisa de todas

. Tem todas as características de uma escala de intervalo, com a vantagem de que o ponto zero representa uma origem verdadeira (zero indica ausência de fenômeno).

Ex: escala métrica, idades e pesos de pessoas, distância, produção, renda per capita, área cultivada, capacidade, etc.

. Todas as operações são possíveis;

. Pode-se calcular qualquer razão entre duas medidas ou dois valores.

Ex:: densidade demográfica de zero pessoas por km² = nenhuma pessoa está na área.

Ex: densidade de 30 pessoas por km² = indica que existem três vezes mais do que 10/km².

. Qualquer teste estatístico paramétrico ou não paramétrico pode ser utilizado.

• **Observações**

. Conhecimento das escalas de mensuração é importante no momento de preparação de questionários.

. Perguntas devem ser elaboradas de tal maneira que as respostas sejam dadas na escala desejada.

. Pode-se formular uma pergunta de duas ou três maneiras, segundo a mensuração escolhida:

Ex: informação sobre o nível de escolaridade do chefe de família:

Escala nominal: O chefe de família é alfabetizado?

Sim Não

Escala ordinal: Qual o nível escolar do chefe de família?

1º grau 2º grau 3º grau

Escala de razão: Quantos anos frequentou a escola?

5 anos

1.4 Estatística descritiva:

. Ocupa-se da organização, apresentação e sintetização de dados.

. Parte mais conhecida

. TV ou jornais – médias, índices, gráficos.

1.4.1 Medidas de tendência central:

Busca identificar valores típicos de uma determinada distribuição.

• **Média aritmética**

. Medida de tendência central mais utilizada;

. familiar para a maioria das pessoas;

. é encontrada adicionando-se todos os valores e dividindo-se o resultado pelo número total de ocorrências:

$$\text{Média} = \sum x_i / n$$

- **Mediana**

. Valor que divide uma distribuição exatamente em duas metades.

Cálculo

. Primeiramente, arranja-se os dados em ordem crescente ou decrescente e em seguida encontra-se o valor central.

. Para os conjuntos com número ímpar de observações, a mediana é encontrada através da fórmula $n + 1/2$, onde n é o número de observações.

. O valor encontrado através da fórmula indica a ordem do termo da distribuição que representa a mediana.

. Para os conjuntos com números pares, a mediana está entre os dois números centrais $n/2$ e $n+2/2$. Após identificar esses números centrais, deve-se somá-los e dividir por dois.

. Às vezes é uma medida melhor do que a média, pois esta é influenciada por valores extremos.

- **Moda**

. Valor que ocorre com maior frequência;

. utilizada mais frequentemente quando dados estão registrados na escala nominal;

. existem conjuntos de dados sem moda;

. existem conjuntos de dados com modas múltiplas (bi-modal x unimodal).

. A exceção dos dados agrupados, a moda não é uma medida muito útil;

. neste caso a classe modal é aquela cuja frequência supera as demais.

- **Distribuição dos dados**

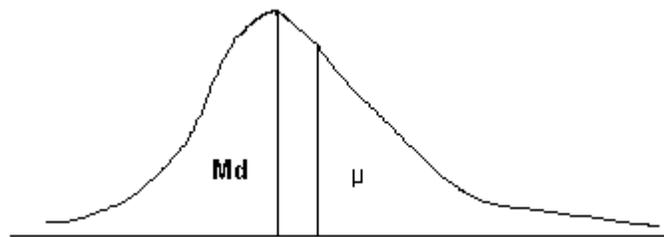
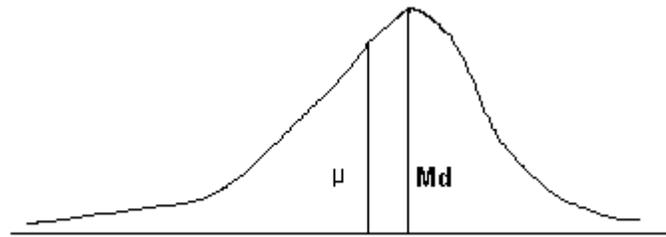
. Existem relações mútuas entre as três medidas de tendência central.

. Se temos um conjunto de dados com distribuição totalmente simétrica-normal, média, mediana e moda são idênticas.

. Se um conjunto de dados tem uma distribuição assimétrica positiva, os três valores médios são diferentes uns dos outros, sendo o valor da média superior ao da mediana.

. A simetria positiva é muito frequente nos conjuntos de dados geográficos.

. Se um conjunto de dados apresenta uma distribuição assimétrica negativa, o valor da média é menor do que o da mediana.



1.4.2 Medidas de variabilidade ou dispersão

- . Para se descrever um conjunto de dados não basta só indicar a tendência central, especialmente quando se compara dois ou mais conjuntos de dados.
- . Um conjunto pode ter todos os valores próximos à média, enquanto outro pode ter os dados mais dispersos
- . Portanto, o grau de dispersão em torno dos valores centrais é uma informação importante.

- **Amplitude total**

- . Medida mais simples de dispersão.
- . É rapidamente encontrada e dá uma primeira impressão sobre a dispersão dos dados para os conjuntos de dados:

1, 4, 7, 10, 13 e 4, 5, 7, 8, 11

- . os dois têm a média 7, mas a dispersão é bem diferente;
- . a dispersão do primeiro caso vai de 1 a 13 – amplitude total de 12;
- . a dispersão do segundo caso vai de 4 a 11 - amplitude total de 7.

. Porém, é uma medida imprecisa, pois o cálculo envolve só dois valores observados, não importa se o conjunto de dados tenha 1000 observações;

. não se tem informação alguma sobre a distribuição dos dados dentro do intervalo ou sobre o número de valores que estão perto da média.

Por exemplo nos conjuntos:

1,2,6,6,6,6,6,10,11 a média é 6 e a amplitude total é 10;

1,1,1,1,6,11,11,11,11 a média é também 6 e a amplitude também 10,

mas os dados se agrupam de modo distinto.

. a amplitude é uma boa medida de dispersão para conjuntos de dados pequenos, porém para conjuntos maiores a medida é desaconselhável.

• **Variância e Desvio Padrão**

. Na prática o desvio médio quadrado em torno da média de um conjunto de dados (variância) é mais utilizado;

. desta maneira, o sinal torna-se sempre positivo.

A soma dos desvios da média elevados ao quadrado é dividida pelo número total de observações.

$$S_x = \sqrt{\sum (x_i - \text{média})^2}$$

Ela é a média dos quadrados dos desvios em relação à média do conjunto.

Como os desvios são elevados ao quadrado, a variância é expressa em unidades quadradas e assim muito difícil de ser interpretada.

Mais importante ainda do que a variância, é o desvio padrão, que indica a dispersão nas mesmas unidades de medidas dos dados originais.

O desvio padrão é a raiz da média dos quadrados dos desvios em relação à média do conjunto e é uma medida do desvio dos valores individuais em relação ao valor central do conjunto de dados ou a raiz quadrada da variância.

Se os valores estão próximos uns dos outros, a soma dos quadrados é pequena.

Se os valores estão distantes uns dos outros, a soma dos quadrados é grande.

Nos casos em que os dados são tirados de uma amostra e se queremos estimar o desvio padrão da população da qual a amostra foi tirada, é aconselhável substituir o denominador por n-1. Com mais de 30 dados o resultado é quase idêntico.

• **Medidas de dispersão relativa**

Para comparar a variabilidade entre diversos conjuntos de dados que têm médias bem diferentes, o coeficiente de variação é uma medida melhor, indicando a variação relativa.

Facilmente obtido dividindo-se o desvio padrão pela média da distribuição.

$$V=s/x$$

Como tanto desvio padrão, quanto média são dados na mesma unidade, V é um número independente de unidades de medida.

Uma desvantagem = não é utilizável se a média está próxima de zero;
. fato que ocorre raramente nos dados geográficos, exceto em relação à temperatura e precipitação.

1.5 Probabilidade:

- . Impossível fazer inferências estatísticas sem utilizar alguns resultados da teoria de probabilidades.
- . Embora intimamente associada à estatística, tem suas características próprias.
- . Busca quantificar a incerteza existente em determinada situação,

- **Experimento aleatório** – processo de coleta de dados relativos a um fenômeno que acusa variabilidade em seus resultados

- **Espaço amostral** – conjunto de todos os resultados possíveis de um experimento (E)

Ex: Dado $E=\{1,2,3,4,5,6\}$

Gênero $E=\{\text{Homem, mulher}\}$

Quando o espaço amostral consiste em um número **finito ou infinito contável** de eventos – espaço amostral discreto;

Quando espaço amostral consiste em **todos os números reais de determinado intervalo** – espaço amostral contínuo.

- **Evento** – Subconjunto de um espaço amostral

- **Probabilidade** – possibilidade de um dado evento ocorrer

Dado – Probabilidade de 1 = $1/6$

Sexo – Probabilidade de feminino = $1/2$

As de copas – Probabilidade $1/52$

- **Distribuição de probabilidades** – distribuição de probabilidades associadas a um conjunto de eventos (espaço amostral).

- **Distribuição finita ou discreta de probabilidades** – baseada em um número contável de eventos

Ex: Experimento com dois dados – soma da combinação dos resultados

Dado 1

Dado 2

1	1	
2	2	
3	3	
4	4	
5	5	
6	6	$E = (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)$

36 combinações possíveis, logo

1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36

Associar a cada valor a sua probabilidade – distribuição de probabilidade (variável aleatória).

- **Distribuição infinita ou contínua de probabilidades** – número infinito de eventos – a curva se homogeneiza a partir de um número infinito de casos

Ex: altura, temperatura, precipitação, tempo de viagem

A distribuição pode ser encarada como um refinamento de uma distribuição bem grosseira. À medida que aumenta a precisão das medidas, um número maior de classes até que no limite temos uma curva contínua.

Exs: Número de crimes em Belo Horizonte – discreta (valores inteiros)

Tempo de percurso – contínuo

Quantidade de leite produzida – contínua

Número de perueiros – discreta

Peso do trigo – contínua

Quantidade de grãos de areia – discreta

Altura – discreta.

Vários tipos de distribuições contínuas – binomial, poisson e pascal (etc.)

Mais útil e mais utilizada é a normal.

- **Distribuição normal**

. Distribuição de probabilidade.

. A mais importante das distribuições contínuas de probabilidade.

. A curva em forma de sino.

. Tem sua origem associada aos erros de mensuração.

. Quando se efetuam repetidas mensurações de determinada grandeza com um aparelho equilibrado, não se chega ao mesmo resultado todas as vezes.

. Obtém-se um conjunto de valores que oscilam, de modo aproximadamente simétrico, em torno do valor verdadeiro.

. Ao construir um histograma desses valores e o correspondente polígono de freqüência, obtém-se uma poligonal aproximadamente simétrica.

. Supunha-se anteriormente que todos os fenômenos devessem ajustar-se a uma curva em forma de sino. Caso contrário, suspeitava-se de alguma anormalidade no processo de coleta de dados.

- . Daí o nome “curva norma”.
- . Descobriu-se depois que vários fenômenos não possuem distribuições normais
- . a distribuição normal tem papel preponderante na estatística, sendo utilizada largamente nos processos de inferência.

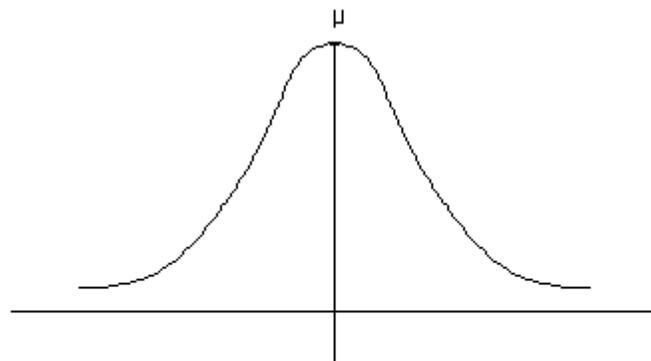
Principais características:

1. média da distribuição é μ
2. desvio padrão é σ
3. a moda ocorre em $x = \mu$
4. a curva é simétrica em relação a um eixo vertical passando por $x = \mu$
5. a curva normal é assintótica ao eixo horizontal em ambas as direções.
6. A área total sob a curva normal e acima do eixo horizontal é 1 (o eixo horizontal é o eixo dos valores de v.a. X, normal)

Propriedades:

- 68,26% das ocorrências encontram-se entre $\pm 1\sigma$
- 95,44% das ocorrências encontram-se entre $\pm 2\sigma$
- 99,74% das ocorrências encontram-se entre $\pm 3\sigma$
- 99,99% das ocorrências encontram-se entre $\pm 4\sigma$

A probabilidade de v.a . normal x estar entre a e b é igual a área sob a curva e acima do segmento horizontal



68,26% das ocorrências encontram-se entre $\pm 1\sigma$
95,44% das ocorrências encontram-se entre $\pm 2\sigma$
99,74% das ocorrências encontram-se entre $\pm 3\sigma$

1.6 Teste de hipótese:

Nos testes de hipóteses, fazemos suposições acerca dos parâmetros desconhecidos e perguntamos o quão prováveis as nossas estatísticas amostrais seriam caso essas suposições fossem de fato verdadeiras.

O objetivo: decidir se uma conjectura/suposição (hipótese) sobre determinada característica de uma ou mais populações é, ou não, apoiada pela evidência obtida a partir de dados amostrais

- **Parâmetro x Estatística**

O objetivo da estatística inferencial é fazer generalizações sobre a população com base em uma amostra retirada da própria população.

Portanto, faz-se necessário diferenciar as características da população e da amostra

- **Parâmetros**

População – parâmetros – letras gregas

Os parâmetros são valores fixos associados a população e são geralmente desconhecidos.

Ex: a média de pontos entre os estudantes de geografia pode ser desconhecida, mas o mesmo valor seria encontrado por todos os pesquisadores.

- **Estatísticas**

Amostra - estatísticas – letras romanas

As estatísticas, por outro lado, variam a cada amostra.

Caso 10 amostras de estudantes fossem selecionadas, nós raramente obteríamos os mesmos resultados.

Porém, ao contrário dos parâmetros, pode-se calcular facilmente as estatísticas para as amostras.

- **Observações**

Entretanto, é a população que nos interessa e não a amostra.

As amostras são trabalhadas por conveniência e o objetivo é fazer inferências acerca dos parâmetros da população, com base nas amostras, que são conhecidas. Amostra é um mero caminho, um passo.

Nos testes de hipóteses, fazemos especulações acerca dos parâmetros desconhecidos e então perguntamos quão provável as estatísticas seriam caso as nossas especulações fossem de fato verdadeiras.

Ao fazê-lo tentamos tomar uma decisão racional se os valores especulados para os parâmetros são razoáveis à luz das evidências.

Teste de hipótese é portanto um processo de decisão. Como a lógica no processo é complexa, segue uma discussão do procedimento

Hipótese estatística/real/alternativa (H1): qualquer afirmação sobre os parâmetros da população em estudo.

Hipótese Nula (Ho) – antítese da hipótese real.

A designação nula - Ho é a hipótese de igualdade ou nulidade – não diferença/não relação.

Erros tipo I e tipo II

Conclusão do teste	Ho verdadeira	Ho falsa
Não rejeitar Ho	Correto	Erro tipo II (β)
Rejeitar Ho	Erro tipo I (α)	Correto

• Etapas para testar uma hipótese estatística:

1. Checar os pré-requisitos dos testes.
2. Formulação das hipóteses Ho e H1.
3. Escolher uma distribuição adequada aos objetivos e a natureza dos dados .
4. Escolher o nível de significância (alfa) e estabelecer a região crítica.
5. Calcular o valor da estatística de teste com base em uma amostra de tamanho n extraída da população.
6. Tomada de decisão.

Bibliografia:

Blalock, Hubert. 1973. Social Statistics. New York, McGraw-Hill.

Gravetter, Frederick e Wallnav, Larry 1992. Statistics for the Behavioral Sciences. New York, West Publishing Company.

Gregory, S. 1973. Statistical Methods and the Geographer. London, Longman.

Hammond, Robert e McCullagh, Patrick. 1974. Quantitative Techniques in Geography – An Introduction. Oxford, Clarendon Press.

Hoel, Paul. 1981. Estatística Elementar. São Paulo, Atlas.

Martins, Gilberto e Donaire, Denis. 1979. Princípios de Estatística. São Paulo, Atlas.

Siegel, Sidney. 1975 – Estatística Não Paramétrica – Rio de Janeiro – McGraw-Hill do Brasil

Soares, José; Farias, Alfredo; César, Cibele. 1991. Introdução à Estatística. Rio de Janeiro, Guanabara Koogan.

2.0 REGRESSÃO LINEAR

2.1 Análise bivariada

Problema típico de correlação e análise de regressão:

. Existem relações entre fenômenos distintos em um conjunto de áreas?

Análises envolvem:

1. variável independente – (representada por x) – causa.
2. variável dependente – (representado por y) – efeito.

Ex: relação entre:

Taxa de fecundidade (número médio de filhos durante idade reprodutiva);

População urbana (%);

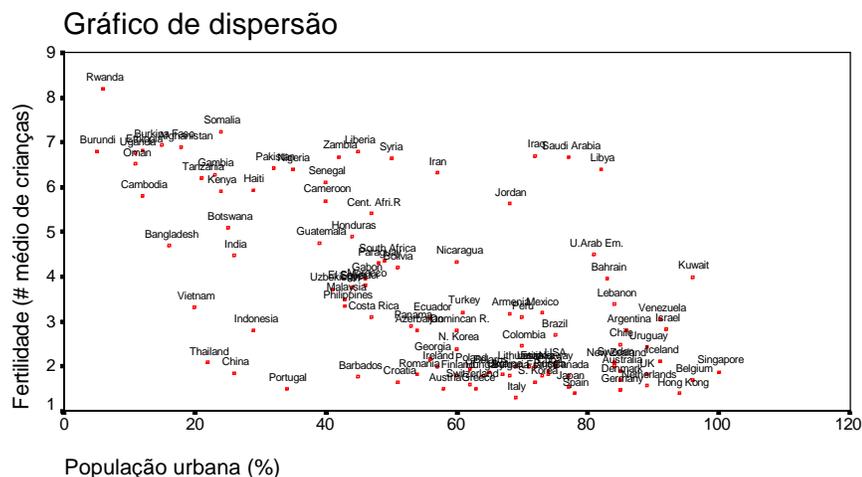
Para uma amostra de países do mundo.

Hipótese:

Quanto maior a proporção de habitantes urbanos, menor será a taxa de fecundidade

Ho: Não existe relação entre as duas variáveis

Exame do diagrama de dispersão indica que a tendência geral foi confirmada, porém para uma medição precisa, é necessário que se conheça a relação funcional entre X e Y.



Em outras palavras:

- . é importante conhecer o impacto que um aumento em X terá em Y (coeficiente de regressão);
- . é também necessário mensurar a representatividade da relação, ou o quão bem a linha de regressão define a distribuição de pontos do diagrama de dispersão (coeficiente de correlação).

2.1.1 Regressão linear simples:

Pergunta inicial:

É possível prever uma variável (Y) a partir de uma outra (X)?

A quantidade de mudança em uma variável dependente (Y), fomentada pela mudança em uma variável independente (X) é indicada pelos parâmetros da equação da regressão, indicada pela fórmula:

$$\hat{Y}_i = a_{yx} + b_{yx}X_i$$

Onde,

\hat{Y}_i é o valor estimado de Y para a *i*ésima observação;

X_i é o valor de X para a *i*ésima observação;

a_{yx} é o termo interceptor (ponto da linha de regressão que cruza o eixo dos Y)

b_{yx} é a inclinação da reta

mudança em Y a cada incremento em unidades de X

Objetivo = prever os valores de Y

O que faz a regressão linear?

. Traça através dos pontos marcados no diagrama de dispersão das variáveis X e Y, uma linha que minimiza as distâncias entre os pontos plotados.

. Minimiza a soma dos quadrados de todos os desvios verticais dos valores reais em relação à linha.

A linha de regressão é, portanto, a melhor descrição, a nível de uma reta, de uma tendência inerente a um conjunto de pontos.

Como é colocada para produzir os valores de a_{yx} e b_{yx} ?

O propósito estatístico por trás da construção da linha de regressão é colocá-la o mais próximo possível de todas as observações, de maneira que minimize os desvios quadrados entre ela e o eixo dos Y. O objetivo é minimizar:

$$\sum (Y_i - \hat{Y}_i)^2 \text{ (Variação)}$$

O objetivo é atingido ao utilizar o conceito estatístico de:

. variância;

. covariância;

. método dos quadrados mínimos;

Variância

$$S_y^2 = \frac{\sum (Y_i - \text{média de } Y)^2}{n}$$

Desvio padrão

$$S_y = \sqrt{S_y^2}$$

Covariância

$$COV_{yx} = \frac{\sum ((x_i - \text{média de } X) (Y_i - \text{média de } Y))}{N}$$

Coeficiente de regressão (b)

O coeficiente de regressão é a razão entre a covariância entre as duas variáveis e a variância na variável independente X.

$$b = \frac{\sum (x_i - \text{média de } x) (y_i - \text{média de } y)}{\sum (x_i - \text{média de } x)^2}$$

ou

$$b = \frac{COV_{yx}}{S_y^2}$$

A covariação indica o tamanho conjunto dos desvios de Y e X de suas respectivas médias, enquanto a variação indica o tamanho dos desvios em Xi. Portanto, quanto maior a covariância, maior será o impacto de X sobre Y.

O cálculo de covariâncias e variâncias envolve os valores individuais de Yi e Xi, em termos de suas distâncias das suas respectivas médias. É uma característica do método dos quadrados mínimos que a reta de regressão passe pelos ponto de interseção da média de x e de y.

Isto ajuda na determinação de a:

$$a = \text{média de } Y - b(\text{média de } X)$$

Obs:

- . A covariância é uma medida absoluta e pode ser positiva ou negativa
- . A variância só pode ser positiva

Coeficiente de correlação (r)

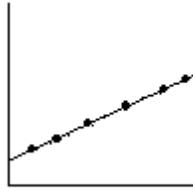
Os dois parâmetros da equação de regressão indicam a forma da relação entre Y e X, mas diz pouco sobre o grau de acuidade das estimativas de Y. Para tal, utiliza-se um parâmetro associado: coeficiente de correlação.

Existem muitos coeficientes de correlação estatística, mas trabalhar-se-á com o coeficiente de correlação de Pearson.

$$r = \frac{\sum (x_i - \text{média de } x) (y_i - \text{média de } y)}{\sqrt{[\sum (x_i - \text{média de } x)^2][\sum (y_i - \text{média de } y)^2]}} \quad \text{Covariância em X e Y} \quad \text{Raiz quadrada do produto da variação total em X e Y}$$

Duas funções:

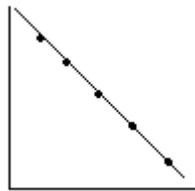
1. Examina o grau de associação de duas variáveis.
Mede até que ponto são interdependentes ou covariantes.
2. Determina a direção da correlação.
Varia de -1 a +1.



$$r = 1$$

Correlação positiva perfeita

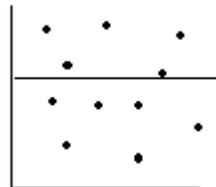
Quanto maiores os valores de x, maiores serão os valores de y



$$r = -1$$

Correlação negativa perfeita

Quanto maiores os valores de x, menores serão os valores de y



$$r = 0$$

Ausência de relação linear

Coeficiente de determinação (r^2)

O coeficiente linear de correlação r_{yx} , compara a variância na variável dependente Y com a redução na variância daquela variável, quando uma variável independente X é utilizada para estimar os valores de Y.

A proporção da variação total em Y explicada por X varia de 0 a 1.

$$r^2 = \frac{\sum (\hat{Y}_i - \text{média de } Y)^2}{\sum (Y_i - \text{média de } Y)^2} \quad \begin{array}{l} \text{variação explicada} \\ \text{variação total} \end{array}$$

Considerações

- . Difícilmente se encontra associações perfeitas ($r = +1$ ou -1)
- . Alto valor de r não significa necessariamente uma relação causal (sorvete e criminalidade)
- . Pode ser utilizada para verificação quantitativa de prováveis relações
- . Revela o grau de relação estatística, mas não explica o porque da relação
- . Coeficiente de correlação nulo ($r=0$), não indica ausência de relação - indica ausência de relação linear
- . Presença de um ou dois valores extremos podem influenciar fortemente os valores de r

Significância

Vários trabalhos que se utilizaram de regressão e/ou correlação utilizam a frase “com 5% de significância”.

Testes de significância estatística são utilizados para inferir características de uma população, com base em uma amostra. Os testes são válidos apenas se:

- . a amostra é aleatória;
- . a população foi completamente especificada.

“A correlação de -0.89 é estatisticamente significativa a 5%”

Isto indica que existe a chance de 95% de que a relação observada na amostra seja verdadeira para a população.

Testes de significância estão ligados a probabilidade de que os resultados observados na amostras não sejam relacionados à população.

Em regressão, existe um modelo para a população

$$Y = \alpha + \beta X \quad \rho = \text{correlação}$$

Que é estimado a partir de uma amostra

$$Y = a + bX \quad r_{yx} = \text{correlação}$$

Teste de significância para o coeficiente de correlação

A maneira de se testar a significância de um coeficiente de correlação é através da razão de F de Snedecor.

Lembrem-se que a variância total na variável dependente é:

$$S^2_y = \sum (Y_i - \text{média de } Y)^2 / n$$

A parcela desta variância que é explicada pela regressão é:

$$(r^2_{yx}) (S_y^2)$$

A parcela não explicada da variância é:

$$(1 - r^2_{yx}) (S_y^2)$$

Para construir o teste de F de Snedecor corrige-se esses valores, chamados de estimativas de variância, pelos seus respectivos graus de liberdade.

Existem:

(N-1) graus de liberdade na variância total

k graus de liberdade na variância explicada, sendo k o número de variáveis independentes

n-k-1 graus de liberdade na variância não explicada

$$F = \frac{(r_{yx}^2) (S_y^2)/k}{(1-r_{yx}^2) (S_y^2)/(n-k-1)} \quad \frac{\text{variância explicada/graus de liberdade}}{\text{variância não explicada/graus de liberdade}}$$

Programa informa automaticamente o nível de significância associado aos valores de F.

Teste de significância para o coeficiente de regressão

Através do teste T de Student.

$$T = \frac{b_{yx}}{SEb}$$

Seb – erro padrão da distribuição dos coeficientes de regressão

$$Seb = \frac{SE_y}{S_x \sqrt{n-2}}$$

Onde,

Se_y – erro padrão residual da regressão

S_x – desvio padrão de X

N – número de observações

Programa informa automaticamente o nível de significância associado aos valores de T.

Resíduos

Quando as observações deixam de cair na linha de regressão, o coeficiente de correlação indica o grau de ajustamento da linha de regressão no conjunto de pontos. Isto não indica, nem o sucesso da equação, ao estimar uma observação em particular, nem a variação existente em torno dos valores estimados de Y. Para tal, verifica-se os resíduos da regressão, definidos por:

$$\text{Res } Y_i = Y_i - \hat{Y}_i$$

O valor residual

Utilizados para identificar observações que estão mais distantes da linha de maior ajustamento. Pode indicar casos discrepantes, ou sugerir o uso de outras variáveis independentes que podem ser levadas em consideração na melhoria do modelo.

Resíduos positivos – valor estimado é menor do que o valor real – valor subestimado

Resíduos negativos – valor estimado é maior do que o valor real - valor superestimado

Pré-requisito da correlação e regressão

. Variáveis intervalares ou de razão

. Linearidade

Análise de regressão constrói uma linha que melhor define a distribuição de pontos;

Correlação testa a robustez desta linha, em relação a distribuição de pontos;

Caso não sejam lineares – curvilinhas – transformações.

- . Normalidade
 - variáveis normalmente distribuídas;
 - resíduos normalmente distribuídos ($Y_i - \hat{Y}_i$);
- . Variâncias iguais
- . Autocorrelação
 - valores de X são independentes entre si;
- . Variáveis independentes, sejam de fato independentes.

Aplicações das análises de regressão e de correlação simples

- . Verificação de relações entre variáveis.
- . Teste de hipóteses.
- . Predição e planejamento.

Não se deve constituir num fim, mas levar o pesquisador, especialmente através da análise e do mapeamento de resíduos, a formular ciclicamente novas hipóteses a serem testadas com o objetivo de tentar explicar a totalidade do fenômeno.

2.2 Análise Multivariada

Explora o poder de explicação que um conjunto de variáveis independentes têm quando tomadas em conjunto.

Pergunta inicial:

É possível prever uma variável (Y) a partir de um conjunto de outras (X_n)?

2.2.1 Regressão múltipla

A quantidade de mudança em uma variável dependente (Y), fomentada pelas mudanças em variáveis independentes (X_n) é indicada pelos parâmetros da equação da regressão, indicada pela fórmula:

$$\hat{Y}_{0,12} = a_{0,12} + b_{01,2}X_{1+} + b_{02,1}X_{2+/-}$$

Onde,

$\hat{Y}_{0,12}$ é o valor estimado de Y a partir das variáveis independentes X₁ e X₂;

$a_{0,12}$ é o valor intercepto (ponto do plano de regressão que cruza o eixo dos Y, onde X₁=X₂=0);

$b_{01,2}$, $b_{02,1}$ são os coeficientes de regressão parciais, indicando a inclinação das relações entre Y₀ e X₁ e X₂, respectivamente, enquanto a(s) outra(s) variável (is) é/são mantida(s) constante(s);

ϵ , erro.

O que faz a regressão linear múltipla?

- . Traça através dos pontos marcados no diagrama de dispersão das variáveis X e Y, um plano que minimiza as distâncias entre os pontos plotados.
- . Minimiza a soma dos quadrados de todos os desvios verticais dos valores reais em relação ao plano.

Correlação parcial

Trabalha os dados de tal maneira, que se pode verificar o efeito de uma variável, como se as outras não estivessem presentes na análise.

$r_{01.23-n}$ indica a correlação parcial entre a variável dependente (Y_0) e uma variável independente X_1 , mantendo o efeito das outras variáveis independentes (X_2, X_3, X_n) constantes.

Um número infinito de variáveis pode ser controlado.

Os números antes do ponto indicam as variáveis ativas, ao passo que as colocadas à direita do ponto indicam as variáveis que estão sendo controladas.

$r_{01.2}$ indica a correlação entre Y_0 e X_1 , tendo removido o efeito das relações $Y_0 = f(x_2)$ e a relação $X_1 = f(x_2)$. Essas remoções são produzidas ao regressarmos:

Y_0 em X_2 e

X_1 em X_2 e

Então, fazendo a regressão dos resíduos dessas regressões:

$$r_{01.2} = \frac{r_{01} - (r_{02})(r_{12})}{\sqrt{1 - r_{02}^2} \sqrt{1 - r_{12}^2}}$$

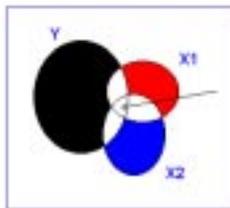
Coefficientes de regressão parciais padronizados

$b_{01.2}$ indica o aumento absoluto em Y associado a um aumento em uma unidade em X_1 , mantendo-se o efeito de X_2 constante.

Comparações entre os coeficientes b são impossíveis, uma vez que as variáveis independentes geralmente apresentam unidades de medida distintas.

A solução é padronizar os valores do coeficiente de b , transformando-os em coeficientes beta (B). Os valores de beta são dados em unidades de desvios padrão e podem ser prontamente comparáveis.

$$B_{01.2} = b_{01.2} \frac{S_{X_1}}{S_{X_2}}$$



Coefficiente de correlação múltipla

Coefficientes de correlação parciais indicam a magnitude da relação entre duas variáveis, mantendo o efeito das demais variáveis presentes na análise constantes.

O quadrado dos coeficientes de correlação parciais indicam a proporção da variância residual na variável dependente, que é associada com a variância residual na variável independente.

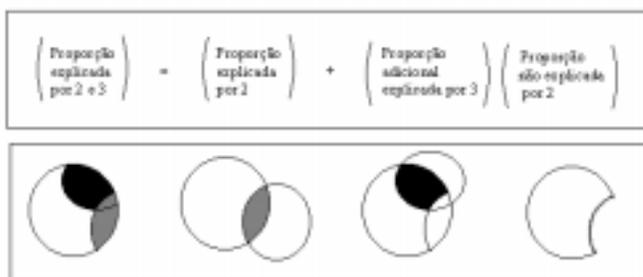
Mede a magnitude da relação entre uma variável dependente e uma série de variáveis independentes

Procede-se da seguinte maneira:

1. Primeiramente, permite-se que uma das variáveis independentes explique toda a variação possível;
2. Depois, permite-se que uma segunda variável independente explique a porção da variação deixada inexplicada pela primeira. Porém, para evitar duplicação, deve-se controlar o efeito conjunto que as duas variáveis independentes têm.
3. Então, permite-se que a terceira variável seja introduzida, controlando o efeito das outras duas variáveis independentes no modelo.

O processo segue indefinidamente, ao sabor do número de variáveis independentes no modelo.

$$R^2_{1,23} = r^2_{12} + r^2_{13,2} (1-r^2_{12})$$



Bibliografia:

Blalock, Hubert. 1973. Social Statistics. New York, Mcgraw-Hill.

Gregory, S. 1973. Statistical Methods and the Geographer. London, Longman.

Hammond, Robert e McCullagh, Patrick. 1974. Quantitative Techniques in Geography – An Introduction. Oxford, Clarendon Press.

Hoel, Paul. 1981. Estatística Elementar. São Paulo, Atlas.

Johnston, R. 1992. Multivariate Statistical Analysis Geography. New York. Longman Scientific & Technical.

King, Leslie. 1969. Statistical Analysis in Geography. Englewood Cliffs, Prentice-Hall Inc.

Martins, Gilberto e Donaire, Denis. 1979. Princípios de Estatística. São Paulo, Atlas.

Montgomery, Douglas e Peck, Elizabeth 1992. Introduction to Linear Regression Analysis. New York, John Wiley & Sons, INC.

Rummel, R. J. 1970. Applied Factor Analysis. Evanston, Northwestern University Press.

Soares, José; Farias, Alfredo; César, Cibele. 1991. Introdução à Estatística. Rio de Janeiro, Guanabara Koogan.

3.0 Componentes Principais/Análise Fatorial

Em regressão linear múltipla, busca-se compreender a relação entre um conjunto de variáveis independentes (X_n) e uma variável dependente (Y_i). O processo se dá de tal maneira, que além de todas as variáveis serem definidas/conhecidas previamente, especula-se acerca da direção da relação entre elas (positiva ou negativa).

Em componentes principais/análise fatorial, estuda-se a relação entre um conjunto de variáveis, explorando-se as inter-relações entre todas as variáveis simultaneamente. Desta maneira, todas as variáveis são ao mesmo tempo independentes e dependentes entre si. O resultado desta análise é um novo conjunto de variáveis, e a relação entre o primeiro e o segundo conjunto de variáveis é o foco da análise.

Mas por que trocar um conjunto de variáveis por outro? Três motivos distintos:

1. Para identificar grupos de variáveis inter-correlacionadas, ou a estrutura subjacente na base de dados. Neste caso, supõe-se que a lista de variáveis originais é a manifestação de um número menor de variáveis (fatores ou dimensões);
2. Simplificar os resultados pela redução do número de variáveis (dimensões);
3. Classificação de áreas/regionalização com base nas dimensões/vetores produzidos.

Início da análise

Matriz de dados:

Uma lista de “p” variáveis e “n” valores, obtidos em uma amostra.

	Variáveis			
Amostra	X1	X2	X3	Xp
1	X_{11}	X_{21}	X_{31}	X_{p1}
2	X_{12}	X_{22}	X_{32}	X_{p2}
N	X_{1n}	X_{2n}	X_{3n}	X_{pn}

A partir desta matriz de dados multivariados, obtém-se a covariância, ou correlação entre as variáveis. Trabalhar-se-á a partir da correlação de Pearson¹ entre as variáveis, como entrada para a Análise dos Componentes Principais/Fatorial.

A matriz de correlação é composta de coeficientes de correlação (r) entre todos os pares possíveis de variáveis.

O modelo

As variáveis ou atributos (X_1, X_2, \dots, X_p) são definidas como combinações lineares de k componentes/fatores não observáveis (S_1, S_2, \dots, S_k), comuns a todas as variáveis, e um fator específico (E_i) para cada variável:

$$X_1 = f(S_1 + S_2 \dots S_k)_{+/-} E_1$$

$$X_2 = f(S_1 + S_2 \dots S_k)_{+/-} E_2$$

$$X_p = f(S_1 + S_2 \dots S_k)_{+/-} E_p$$

¹ Consultar item 2.0 da apostila para discussão sobre coeficiente de correlação de Pearson.

Onde, X_1 é uma das variáveis originais;

S_1, S_2, S_k são os componentes/fatores, que por sua vez são compostos pelas variáveis originais.

Os componentes/fatores são estimados a partir das variáveis originais X_1, X_2, \dots, X_p , segundo o modelo:

$$S_j = w_{j1}X_1 + w_{j2}X_2 + \dots + w_{jp}X_p$$

Componentes principais e análise fatorial diferem na maneira como o erro é tratado. Na análise de componentes principais, os erros são tratados como componentes, de tal modo que as todas as variáveis estão relacionadas a uma série de componentes, um dos quais pode muito bem ser o seu próprio erro. Por isso, a análise de componentes principais é chamada de modelo fechado, uma vez que toda a variância associada às variáveis originais é investigada. O resultado é um conjunto de componentes que iguala o número de variáveis originais.

Já a análise fatorial exclui o erro das equações, de tal forma que, cada variável é dividida em duas partes: a variância comum (associada a outras variáveis) e a sua variância única, que é residual. A variância comum é então dividida entre o conjunto de fatores, da mesma maneira como a variância é dividida na análise de componentes principais.

Extração dos componentes/fatores

1. Estimação inicial dos fatores por meio da análise de componentes principais

Os componentes principais/fatores (S_1, S_2, \dots, S_p) são combinações lineares das p variáveis aleatórias X_1, X_2, \dots, X_p . Geometricamente, estas combinações lineares representam a seleção de um novo sistema de coordenadas, obtido pela rotação do sistema original de eixos X_1, X_2, \dots, X_p . Os novos eixos representam as direções com máxima variabilidade e fornecem uma descrição mais simples e parcimoniosa da estrutura de correlação.



Estes componentes/fatores são obtidas matematicamente de forma que a primeira (S_1) contenha a maior quantidade possível de informação total presente nas “ p ” variáveis originais. Já a segunda (S_2), que é independente da primeira, contém o máximo possível da informação restante, e assim sucessivamente. Quanto maior for a correlação entre as variáveis originais, maior é a informação contida nas primeiras componentes.

Seleção do número de componentes/fatores

A análise de Componentes Principais produz um fator para cada variável original. Na busca de simplificação dos dados, deve-se selecionar um número pequeno de fatores ($k < p$), retendo apenas aqueles que trazem grande parte da informação relevante contida nos dados originais.

A determinação do número de fatores que deve ser utilizado para representar os dados leva em consideração os autovalores, também denominados valores característicos ou *eigenvalues*, correspondentes a cada fator. Os critérios mais utilizados são os seguintes:

1. Selecionar o número de fatores que explique grande parte da variação total contida nos dados. A porcentagem da variância total contida no i -ésimo fator é dada por:

$$\% \text{ da variância total} = \frac{\text{Autovalor} \times 1000}{\text{Soma dos } p \text{ autovalor}}$$

2. Analisar a representação visual dos autovalores no gráfico Scree, observando a contribuição de cada fator.

3. Utilizar apenas os fatores cujos autovalores sejam maiores do que 1.

Interpretação dos componentes/fatores

. Matriz dos *loadings* dos componentes/fatores

Uma vez selecionados os componentes/fatores que representam satisfatoriamente a informação contida nas variáveis originais, deve-se interpretar cada componente/fator pela análise de como as variáveis originais estão relacionadas a cada componente/fator. Para isto são utilizados os valores dos coeficientes que relacionam as variáveis originais padronizadas com os fatores. Estes coeficientes são denominados *loadings* dos fatores, pois indicam o peso de cada variável no componente/fator e são equivalentes aos coeficientes de correlação (r) entre os componentes/fatores e cada variável original.

É interessante notar que a soma dos quadrados dos *loadings* de cada fator produz a variância explicada por cada um, que é uma medida da quantidade de informação existente nos dados originais que foi captada pelo fator.

Uma vez decidido o número de fatores que será considerado, deve-se dar um nome para cada fator extraído. Porém, em geral, todas as variáveis estão relacionadas como primeiro fator, dificultando a interpretação.

A técnica utilizada para melhorar a interpretação dos resultados consiste em modificar os valores dos *loadings*, de tal modo que os novos valores produzam uma matriz de *loadings* dos fatores com um estrutura simples. Isto é obtido por meio da rotação dos fatores iniciais.

Rotação

A rotação mantém a informação total presente nos componentes/fatores originais, mas faz nova atribuição das variáveis originais aos fatores;

Os principais critérios para a rotação são:

- . cada fator deve ter a maioria dos loadings o mais próximo de zero possível;
- . cada variável original deve ter poucos loadings próximos de 1 nos diversos fatores;
- . quaisquer dois fatores devem exibir padrões diferenciados de loadings baixos e altos.

Existem vários métodos de rotação. Em geral, os softwares estatísticos tem algoritmos disponíveis para a rotação ortogonal. Este tipo de rotação preserva a orientação original entre os fatores, de modo que permaneçam perpendiculares após a rotação. Os métodos de rotação ortogonal mais populares são:

1. Varimax: busca uma rotação dos fatores de forma a maximizar a variação dos quadrados dos loadings. Obtém-se, para cada fator, loadings grandes, médios e pequenos.
2. Quartimax: procura atribuir a cada variável apenas um loading elevado. Este critério tem a tendência indesejável de geral um fator global, onde todas as variáveis têm loadings elevados.
3. Equamax: busca obter uma estrutura simples com relação às linhas e colunas da matriz formada pelos loadings dos fatores. É uma combinação dos métodos varimax e quartimax.

Cálculo dos escores dos componentes/fatores

Após a extração dos componentes/fatores que resumem as variáveis originais (dimensões primárias), há interesse, na maioria das vezes, em obter os valores dos fatores correspondentes aos dados da amostra. Estes valores, nos novos eixos coordenados, são denominados escores. Os escores podem ser utilizados para construir gráficos, ou são utilizados como entrada de dados para outras técnicas estatísticas. Pode-se ainda utilizá-los no processo de classificação/regionalização.

Bibliografia:

Dillon, William R. 1984. Multivariate Analysis, Methods and Applications. New York, John Wiley & Sons, Inc.

Drumond, Fátima. Análise Dimensional. Departamento de Estatística. Ices/UFMG

Faissol, Speridião 1972. Análise Fatorial: problemas e aplicações na geografia, especialmente nos estudos urbanos. Revista Brasileira de Geografia. 34 (4): 77-100.

_____. 1972. A Estrutura Urbana Brasileira: uma visão ampliada no contexto do processo brasileiro de desenvolvimento econômico. Revista Brasileira de Geografia. 34 (3):19-123.

Johnston, R. 1992. Multivariate Statistical Analysis in Geography. New York. Longman Scientific & Technical.

Rummel, R. J. 1970. Applied Factor Analysis. Evanston, Northwestern University Press.