MÉTODOS QUANTITATIVOS APLICADOS EM GEOGRAFIA: UMA INTRODUÇÃO

BARBARA-CHRISTINE NENTWIG SILVA (*)

O objetivo deste trabalho é o de apresentar métodos matemático-estatísticos mais comumente empregados na Geografia, discutindo importância, vantagens e desvantagens da aplicação dos mesmos em nossa disciplina. Espera-se, com isto, poder contribuir de forma sistemática para o ensino e a pesquisa geográfica no Brasil, considerando que os referidos métodos, apesar de largamente utilizados em inúmeros países, dando margem a uma extensa bibliografia (Greer-Wootten, 1972), ainda necessitam de maior divulgação e emprego entre nós. Para tanto, nota-se a existência de grande motivação por parte de muitos estudantes e professores quanto ao emprego imediato de métodos quantitativos avançados, tais como análise fatorial e análise de superfície de tendência, faltando-lhes, em geral, imprescindíveis informações quanto aos conhecimentos metodológicos que devem necessariamente anteceder a aplicação citada. Talvez, por isso, não se tenha conseguido entre nós uma difusão ainda mais rápida e eficiente dos métodos quantitativos. Assim, o objetivo deste artigo é discutir a aplicação dos métodos quantitativos introdutórios à Geografia, esperando-se analisar os métodos mais avançados em uma etapa posterior.

1. NECESSIDADE DO EMPREGO DE MÉTODOS QUANTITATIVOS

1.1. Necessidade e importância

Na pesquisa e no ensino da Geografia existe, em termos gerais, abundância de dados, sendo muito difícil, senão impossível, tratar conjuntos muito numerosos sem emprego de métodos quantitativos visando permitir a redução das informações a formas manejáveis e interpretáveis.

Existe também a possibilidade de análises mais profundas dos dados disponíveis, de solução de problemas complexos e de exploração de novos campos não passíveis de serem descobertos unicamente através da simples observação dos dados brutos. É preciso notar igualmente que, tanto para os casos de dados muito numerosos como para os pouco numerosos, os métodos quantitativos possibilitam maior objetividade e precisão na análise, podendo evitar longas e muitas vezes superficiais descrições verbais. Com o emprego destes métodos, os geógrafos desenvolvem uma lógica bem mais crítica, sendo orientados a pensar de forma rigorosa e pre-

(*) Instituto de Geociências da Universidade Federal da Bahia, Salvador.

cisa, evitando generalizações vagas baseadas sobre evidências insuficientemente analisadas. Além disso, os métodos não-quantitativos aplicados aos mesmos dados levam, em numerosas ocasiões, a resultados diferentes, permitindo variadas interpretações, enquanto os métodos quantitativos possibilitam a obtenção de resultados idênticos utilizando iguais procedimentos para os mesmos problemas e, consequentemente, uma única interpretação. Por outro lado, os métodos quantitativos permitem ao pesquisador importante economia de recursos e tempo.

A necessidade de empregar métodos quantitativos na Geografia é, também, reforçada pelo caráter de linguagem científica, interdisciplinar e universal que os mesmos apresentam. Assim, por exemplo, um índice de correlação r = 0,9 tem o mesmo sentido para qualquer pesquisador, não importando a sua disciplina, se ele conhece a técnica de correlação. Uma explicação verbal como "alta correlação" é muito menos precisa, mesmo quando afirmado por pesquisadores de uma única disciplina.

Finalmente, a importância da abordagem quantitativa deve ser ressaltada na contribuição que a mesma oferece à aplicação da Geografia na solução de problemas de diversas naturezas, através do oferecimento de eficientes modelos analíticos, preditivos e de planejamento.

1.2. Abordagem científica

A utilização dos métodos quantitativos na Geografia deve ser colocada em uma perspectiva mais ampla, qual seja, a da abordagem científica como um todo. Com isto, é destacada a sua correta posição no processo científico, evitando-se tanto a depreciação quanto a superestimação dos referidos métodos.

Entendemos por ciência um método de estudo, ou seja, um processo onde se constrói, passo a passo, um modelo da realidade, supervisionado e manejável. Esta realidade pode envolver somente fenômenos naturais ou humanos e ainda uma combinação dos dois. Com isto, afastamo-nos da idéia de ciência como o estudo de certos conjuntos de fenômenos, por exemplo, os naturais em oposição aos humanos, problema ainda relevante em nossos dias. Cole e King (1968, p.18) afirmam, a este respeito, que a possibilidade de uma ciência existir é determinada não pelo *o que* mas pelo *como*, não pelo assunto mas pelo método. Assim sendo, todos os assuntos podem ser objeto de investigações científicas. Cole e King (p. 18-19) sugerem uma sequência de etapas ou passos para uma pesquisa geográfica, que será mostrada a seguir, de forma simplificada. Antes, porém, é preciso deixar claro que esta sequência não é relevante para todas as situações na Geografia, devendo servir, sobretudo, como sugestão para perguntas que o geógrafo deve fazer a si mesmo quando trabalhando em pesquisa.

a) inicialmente é recomendável ter em vista um objetivo para o estudo, em lugar de coletar material esperando encontrá-lo no decorrer do trabalho ou só no fim deste. Este pode ser um problema relevante a resolver, uma hipótese ou um modelo teórico a testar, mesmo emprestado de uma outra disciplina e que podemos tentar aplicar em uma situação especial. Na formulação dos objetivos, bem como no próprio desenvolvimento dos trabalhos, devem estar envolvidos obrigatoriamente todos os aspectos conceituais pertinentes, apresentados de forma lógica e consistente;

b) uma vez definido um objetivo, é normalmente necessário fazer a coleta de informações que, na Geografia, podem ser conseguidas de maneira direta ou indireta, primária ou secundária. Podem também ser obtidas através de trabalho de campo, material publicado em forma verbal ou numérica ou de fontes como a fotografia aérea ou mapas existentes. As amostragens são muitas vezes necessárias, mas os geógrafos a utilizam de forma muito menos numerosa e precisa do que outros pesquisadores;

c) os dados devem ser preparados e guardados. Se são apresentados em um mapa, este torna-se um tipo de armazenagem. Dados de interesse geográfico podem também ser armazenados em tabelas e particularmente em matrizes. A forma de armazenagem não precisa ser a página impressa: métodos modernos permitem, por exemplo, armazená-los em grandes quantidades sobre cartões de computação ou fitas;

d) depois de preparar os dados, estes devem ser processados. Cada vez mais são utilizados, na Geografia, métodos matemáticos e inferências estatísticas derivadas do processo matemático. Muitas análises precisam de computação, porque sem isso o tempo gasto e, consequentemente, os custos, seriam enormes mesmo quando calculadas com máquinas convencionais de mesa.

e) com estes passos, os resultados deverão ser encontrados. A interpretação dos resultados será dada em forma compreensível, seja verbal, numérica, cartográfica ou de alguma outra forma que possa ser entendida. Os resultados confrontados com os objetivos iniciais podem indicar muitas direções: – podem ser tão insatisfatórios que o pesquisador deve recomeçar o trabalho a partir do primeiro ou segundo passo; [1] – podem indicar a necessidade de trabalho suplementar; – podem também ser apresentados como um estudo completo.

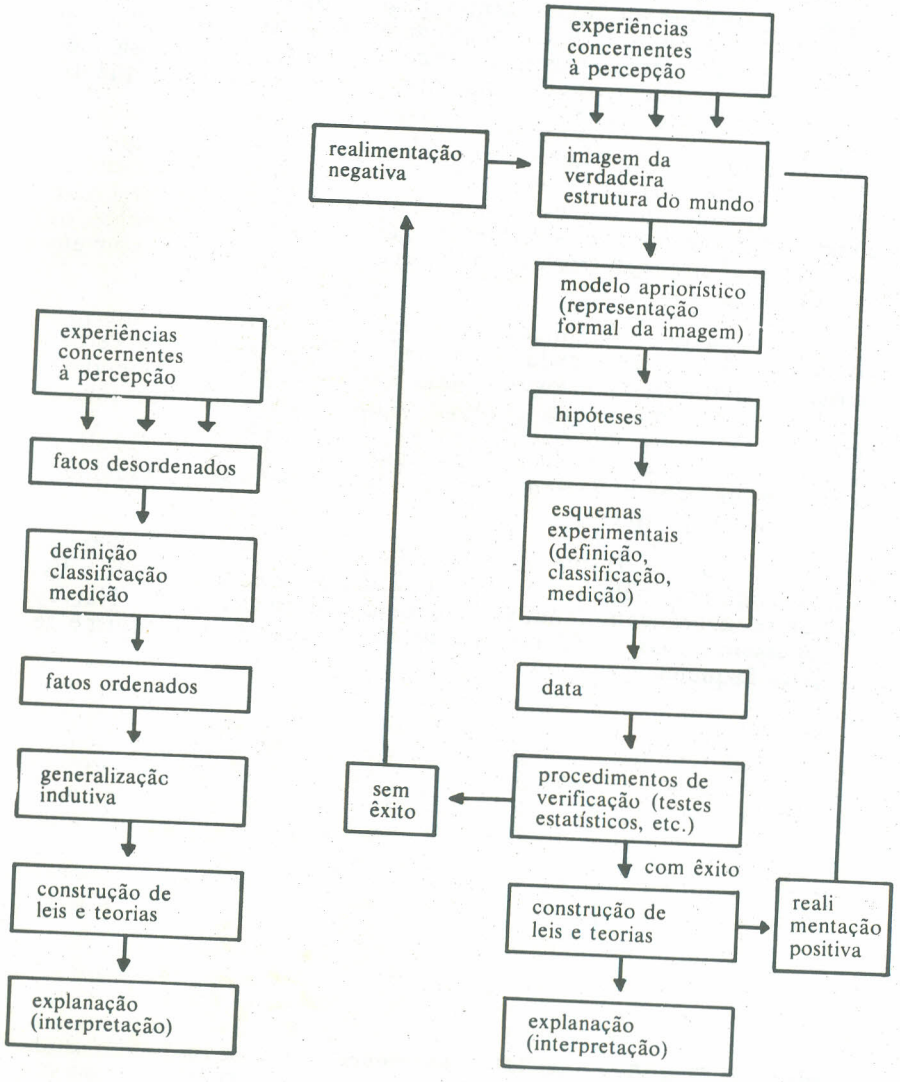
Dois outros autores, Harvey e Daugherty, apresentam graficamente as principais etapas do método científico aplicado à Geografia, fazendo distinção entre abordagem indutiva e dedutiva. Primeiramente, Harvey (1969) apresenta duas abordagens na pesquisa geográfica, conforme se observa no Esquema 1.

1. Raramente na Geografia o pesquisador admite a possibilidade de encontrar resultados negativos em seus trabalhos, indicando seja o reinício dos mesmos por um outro caminho, seja a necessidade de complementação. Entretanto, este fato é de grande importância no processo científico, sendo comumente aceito em outras disciplinas.

Esquema 1

ABORDAGENS NA PESQUISA GEOGRÁFICA

(seg. D. Harvey – *Explanation in Geography* – Nova York, 1969, p. 34)



a) abordagem indutiva

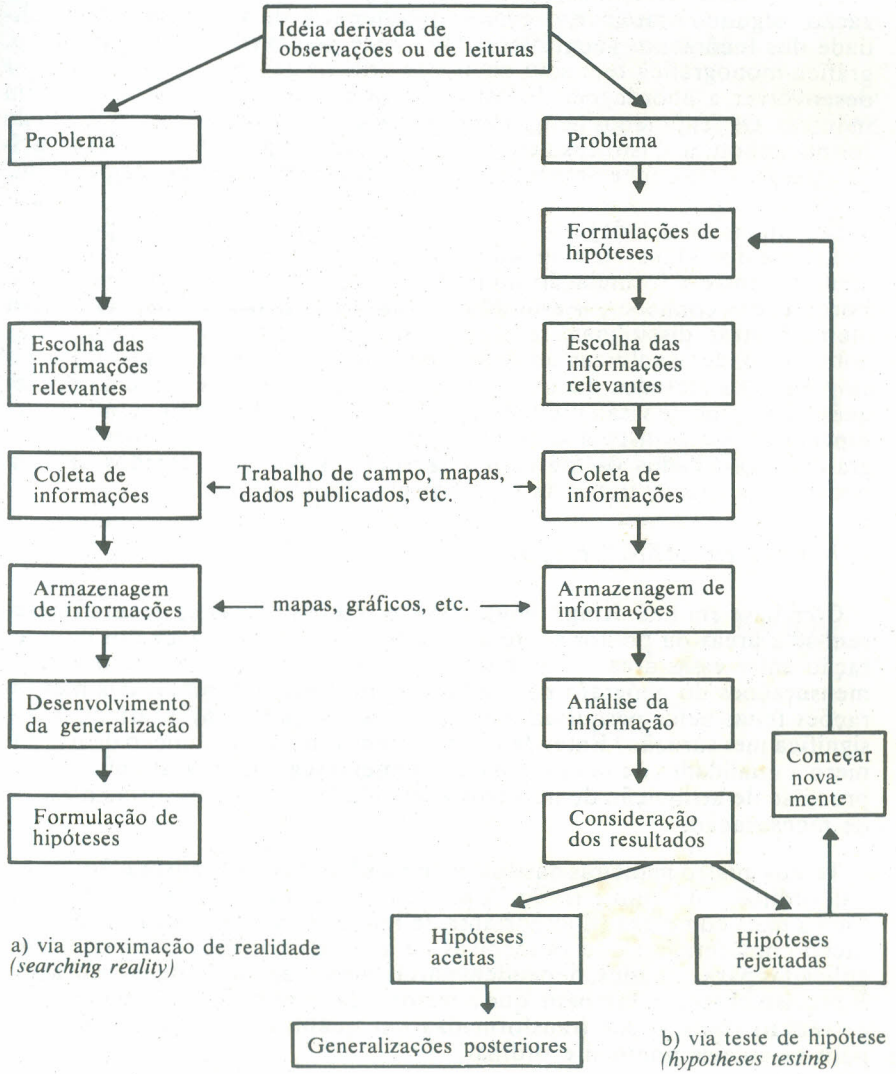
b) abordagem dedutiva

Mais recentemente, Daugherty (1974) propôs as seguintes abordagens (ver Esquema 2).

Esquema 2

DUAS ABORDAGENS NA PESQUISA GEOGRÁFICA

(seg. R. Daugherty – *Science in Geography* – Data Collection – Londres – 1974 – p. 7)



a) via aproximação de realidade *(searching reality)*

b) via teste de hipótese *(hypotheses testing)*

Alguns comentários devem ser feitos sobre estes dois esquemas:

Primeiramente, destaca-se a importância da generalização como objetivo final nas *duas* formas de abordagem para uma pesquisa geográfica. Entretanto, é preciso reconhecer que a Geografia, por aceitar durante muito tempo os seus fenômenos como sendo excepcionais, – como decorrência de sua localização individualizada sobre o espaço, não sendo por isto mesmo passíveis de sujeição a leis e princípios gerais –, utilizou muito mais e de forma limitada a abordagem indutiva. A chamada abordagem idiográfica-monográfica, tão amplamente discutida nos últimos anos no Brasil, corresponde exatamente ao acima exposto, deixando de desenvolver o processo de construção teórica, segundo Harvey, ou de generalização, segundo Daugherty, como consequência da implítica individualidade dos fenômenos geográficos. Mais recentemente, a abordagem idiográfica-monográfica tem sido objeto de críticas gerais, o que acabou por desenvolver a abordagem dedutiva e a complementação da abordagem indutiva. Os fenômenos geográficos passaram a ser também encarados de forma nomotética, isto é, sujeitos a leis (no sentido probabilístico) e princípios gerais, cujo conhecimento deve ser justamente o objetivo da pesquisa.

Em ambas as abordagens destaca-se, de qualquer forma, a importância da análise dos dados. As etapas a isto referentes correspondem à fase intermediária entre a formulação do problema e a proposição dos resultados. Forçoso é reconhecer, a esta altura, que a Geografia, confrontada com muitas outras disciplinas, teve também uma limitação quanto às suas potencialidades analíticas ao deixar de valorizar durante muito tempo o emprego dos métodos quantitativos. Talvez tenha sido ainda uma consequência do tipo de visão predominantemente idiográfica-monográfica, que supervalorizou os aspectos qualitativos e subjetivos dos problemas geográficos. Colocados os métodos quantitativos em seus devidos termos, podemos, a seguir, tratar de seus aspectos preliminares.

2. *NÍVEIS DE MENSURAÇÃO*

Com base em L. J. King (1969), as mensurações típicas do geógrafo referem-se a áreas ou pontos sobre a superfície da terra e a aspectos da interação entre essas áreas ou pontos. É preciso ressaltar, entretanto, que as mensurações do geógrafo não são diferentes, em si mesmas, das mensurações feitas pelos pesquisadores de outras disciplinas. Mas, afinal, o que significa mensuração? Entendemos por mensuração a 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.

Temos quatro maneiras básicas, ou níveis básicos, de mensuração: nominal, ordinal, intervalo e razão. A primeira maneira é a mais simples, a última a mais complexa. É importante definir as maneiras básicas de mensuração considerando que as técnicas de análise estatística que podem ser aplicadas para os dados, dependem parcialmente da escala de mensuração. É preciso observar, também, que a maioria das descrições qualitativas, isto é, verbais, podem ser transformadas em quantitativas, particularmente para as escalas nominal e ordinal.

2.1. *Escala nominal*

Esta escala é utilizada para classificar objetos ou fenômenos em termos de igualdade dos seus atributos e numerá-los. A forma mais simples de mensuração nominal é a divisão em duas classes, que são identificadas com os números zero e um. Henshall e King (1966, p. 77) apresentam um exemplo de uma escala nominal em que se pode verificar a existência ou não de atributos – produtos agrícolas e pecuária – para fazendas de Barbados, com o número *um* indicando a existência do produto agrícola e o *zero* a sua ausência. Esta mesma classificação pode ser ampliada e utilizada em inúmeras situações de pesquisa visando transformar aspectos qualitativos em quantitativos. Um exemplo de utilização na Geografia seria dado na matriz abaixo – chamada de conectividade – em que aparecem as ligações aéreas bidirecionais entre um grupo de cidades do Estado da Bahia.

Cidades	Salvador	Feira de Santana	Bom Jesus da Lapa	Barreiras	Itapetinga	Vitória da Conquista
Salvador	0	0	1	1	1	1
Feira de Santana	0	0	0	0	0	0
Bom Jesus da Lapa	1	0	0	1	0	0
Barreiras	1	0	1	0	0	0
Itapetinga	1	0	0	0	0	1
Vitória da Conquista	1	0	0	0	1	0

1 = existência de ligações aéreas
0 = inexistência
Fonte: Nordeste Linhas Aéreas. Salvador. 1977.

Esta matriz representa-se graficamente da seguinte forma (fig. 1):

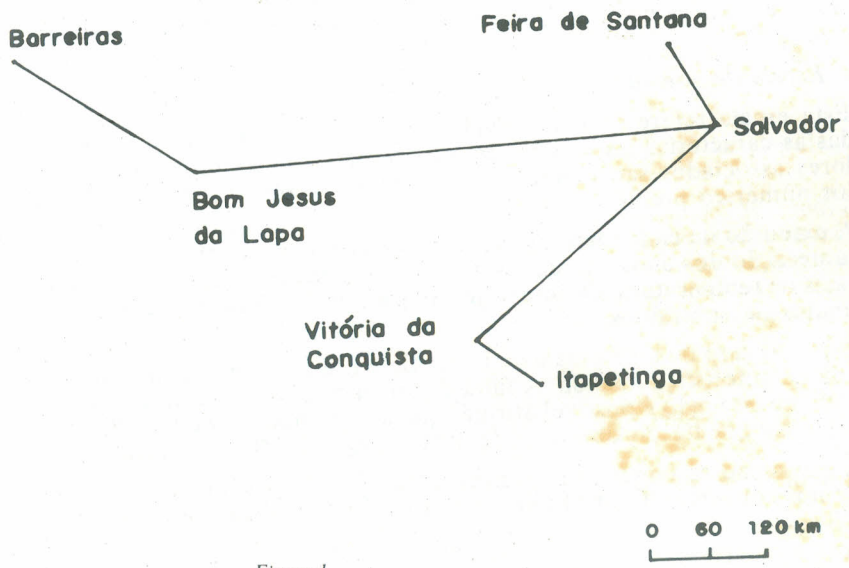


Figura 1

Cada observação na mensuração nominal pertence a uma só classe, sendo estas últimas, por consequência, mutuamente exclusivas. Também é preciso notar que as operações aritméticas comuns de adição e multiplicação não podem ser aplicadas às mensurações nominais, mas é possível extrair outras informações numéricas como a ocorrência de *zero* e *um*, ou seja, a frequência. É possível igualmente identificar a frequência de cada classe e, assim, a classe modal, ou seja, aquela onde se fixa a maioria das observações. A frequência de cada classe, por sua vez, pode ser expressa como percentagem do número total.

2.2. Escala ordinal

Esta escala é utilizada quando os fenômenos ou observações são passíveis de serem arranjados segundo uma ordem, isto é, segundo a grandeza ou preferência. Assim, as expressões qualitativas são arranjadas segundo uma ordem como, por exemplo, a classificação hierárquica dos níveis educacionais (primeiro grau, segundo grau, universitário) em uma seqüência numérica como um, dois, três. Outro exemplo seria a classificação das classes de renda de uma população também segundo uma sequência numérica um, dois, três, quarto... A escala de dureza é um exemplo clássico de escala ordinal.

A aplicação desta forma de mensuração é somente possível quando se desenvolve uma sequência qualitativa na qual é lógico colocar um fato antes do outro. Na escala ordinal, as operações aritméticas também não devem ser feitas. Assim, em uma classificação de hotéis, por exemplo, em três níveis hierárquicos – luxuosos, médios e simples – não se pode dizer que os hotéis luxuosos sejam duas vezes melhores que os médios. Sabemos, por suposição, que os hotéis luxuosos são de nível hierárquico superior aos médios quanto a uma série de características (conforto e serviços), mas não temos meios para quantificar esta diferença na escala ordinal. Como na escala anterior, é possível contar a frequência de cada clase para indicar a classe modal. A mediana pode também ser determinada nesta escala de mensuração.

2.3. Escala de intervalo

Esta escala refere-se a um nível de mensuração em que a escala tem todas as características de uma escala ordinal, mas os intervalos entre os valores associados são conhecidos e cada observação pode receber um valor numérico preciso.

A extensão de cada intervalo sucessivo é constante como, por exemplo, a numeração dos anos, variações de altitude através de curvas de níveis e escalas de temperatura. Os intervalos diferentes são passíveis de serem adicionados ou subtraídos.

O ponto zero de uma escala de intervalo é arbitrário e não indica ausência da característica medida. A falta de um zero absoluto é uma desvantagem. Com isto, não é possível afirmar que uma temperatura de 20ºC é duas vezes mais quente do que uma de 10ºC porque o 0ºC é arbitrário.

A utilização matemática é limitada a transformações lineares (1), isto é, podemos converter ºC em ºFahrenheit através da equação ºF = 32 + 1,8ºC.

(1). $x' = ax + b$ $(a > o)$

Cada transformação linear dessa forma preserva a informação no dado original. Transformações mais complexas alteram as relações. Como vimos, não tem sentido dizer que a temperatura de 30ºC seria duas vezes mais quente do que 15ºC. Caso fosse possível, isto significaria que 86ºF (= 30ºC) seria duas vezes superior a 59ºF (= 15ºC).

2.4. Escala de razão

É a mais precisa de todas, referindo-se a um nível de mensuração em que a escala tem todas as características de uma escala de intervalo, sendo que o ponto zero é uma origem verdadeira. Nesta escala, o zero indica ausência de propriedade. Como exemplo desta escala podemos citar: escala métrica, número, idades e pesos de pessoas, distância, produção renda *per capita* etc.

Com esta escala é possível comparar os valores não só observando as diferenças mas também comparando os estimativos absolutos. Assim, uma densidade de zero pessoas/km² quer indicar que nenhuma pessoa está na área e uma densidade de 30 pessoas quer indicar três vezes mais que 10 pessoas/km². Ou que a razão 100/40 = 20/8 = 5/2. Ou, ainda, pode-se dizer que um intervalo é três vezes maior que outro e também que o valor numa determinada posição é duas vezes maior que o mesmo em outra posição.

3. CONCEITOS BÁSICOS

Como vimos, o geógrafo, sendo confrontado com uma grande quantidade de dados, deve inicialmente reduzí-los a uma forma manejável para que ele possa responder às suas hipóteses ou questões preliminares e chegar a proposições explanatórias. Para tanto, necessário é o conhecimento de conceitos e métodos introdutórios que antecederão aos complexos métodos da análise quantitativa atualmente utilizados.

3.1. População, amostra e variável

O conceito de população ou universo refere-se à totalidade existente das observações individuais, podendo ser limitadas no tempo e/ou no espaço como, por exemplo, uma pesquisa que se faça sobre a distribuição dos escravos no Brasil, em 1800. Por outro lado, a população do ponto de vista estatístico é considerada finita ou infinita.

Por amostra entende-se a análise de pequena parte da população selecionada segundo regras específicas.

Yeates (1974, p.10) define uma variável como conjunto de medidas sobre, um característico específico, que não tem valores constantes porém, variações no valor. A variável que pode assumir teoricamente qualquer valor entre dois dados fixados como, por exemplo, números inteiros, é chamada de contínua. Temperaturas, distâncias, áreas e alturas são exemplos deste caso. A variável que assume um só valor fixado, não podendo ter valor intermediário, é chamada de discreta. Um exemplo seria o número de crianças numa família, a população e veículos no tráfego. É possível converter dados contínuos em discretos se fizermos o arredondamento dos valores, por exemplo, para números inteiros.

Uma constante é um símbolo ou número que tem um só valor, qualquer que seja o problema. Na fórmula para encontrar a circunferência do círculo $C = 2\pi r$ os símbolos 2 e π são constantes.

Spiegel (1971, p. 2-3) menciona que muitas vezes é conveniente estender o conceito de variável a entidades não-numéricas, o que é do especial interesse para a Geografia.

3.2. Arredondamento

Existem muitas situações onde é aconselhável, senão necessário, fazer o arredondamento dos dados, ou seja, a redução do número de dígitos. Isto pode ocorrer, por exemplo, se o pesquisador tem impressão de que certo dado levantado seja impreciso. Assim, um levantamento da população bovina de uma região indicando 60.421 bovinos pode ser arredondado para 60.000, quando se reconhece as dificuldades do próprio levantamento.

Por outro lado, muitas vezes, é aconselhável cortar um número mesmo sabendo ser este correto. Neste caso, o objetivo é o de dar uma idéia mais fácil para tratar as informações. Quando se fala, por exemplo, do valor da exportação do fumo na Bahia, não é necessário referir-se, em geral, aos valores em centavos.

Um terceiro tipo de justificativa sobre a utilização do arredondamento ocorre quando um número representa valor adquirido através de cálculos como, por exemplo, a temperatura média de um lugar. Neste caso, mesmo sendo os cálculos precisos – por exemplo, 24,621ºC – não é necessário indicar todos os números, devendo se fazer o arrendondamento de tal maneira que combine o melhor possível, com a exatidão da medida, no caso 24,6ºC.

As regras para se fazer um arredondamento são simples, mas muitas vezes não são obedecidas. Isto pode ter o efeito de uma acumulação de inexatidões ou erros desde o começo da pesquisa.

Um número que deve ser arredondado não muda se o número que o segue é menor do que cinco. Se o número que deve ser arredondado é seguido de número maior do que cinco, ou de um número cinco seguido de outros com exceção de 0, neste caso ocorre o aumento de 1. Se o número a ser arredondado é seguido de cinco só, ou de cinco seguido de zeros, ele fica sem ser mudado se é número par, mas aumentado de 1 se é ímpar. Se muitos desses números devem ser somados, isto tem como objetivo, fazer com que se obtenha, em média, o mesmo número de dados aumentados com relação aos diminuídos.

Assim, para resumir as possibilidades de arredondamento apresentamos os seguintes exemplos:

número	arredondamento para o número inteiro
64,8	65
64,4	64
83,5	84
83,50	84
84,5	84
84,50	84
84,51	85

4. *TÉCNICAS DE AGRUPAMENTO*

Se temos muitas observações é preciso, através de diversos passos, arranjar os dados numa forma manejável. Existem vários métodos que permitem a redução dos dados a uma forma supervisionável, de tal maneira que uma comparação com outros dados seja também possível. A escolha do método para o início e desenvolvimento de um trabalho científico, dependerá basicamente da natureza dos dados e do objetivo e profundidade da análise.

A tabela 1 mostra as taxas de urbanização municipais no Estado de Alagoas, em 1970. Foram indicadas as taxas de todos os noventa e quatro municípios do referido Estado. Para simplificar o nosso exemplo, fazemos o arredondamento para números inteiros das taxas de urbanização e depois arranjamos as informações em forma de um rol (tab. 2).

O rol corresponde a disposição dos dados em ordem de grandeza crescente ou descrescente. Assim, pode-se determinar a amplitude total (range), ou seja, a diferença entre o maior e o menor valor, que seria no nosso exemplo 92% – 6% = 86%. Além disso, podemos contar quantas vezes ocorre uma observação e indicar a moda que é o valor que ocorre com maior frequência. A mediana, que é o valor central, é também fácil de ser determinada.

TABELA 1 – Taxas de urbanização municipais no Estado de Alagoas, 1970. (População da sede municipal sobre a população total, em %).

Água Branca	8	Igreja Nova	17	Palmeira dos Índios	42
Anadina	24	Inhapi	9	Pão-de-Açúcar	35
Arapiraca	46	Jacaré dos Homens	35	Passo de Camaragibe	35
Atalaia	13	Jacuípe	23	Paulo Jacinto	42
Barra de Santo Antônio	38	Japaratinga	22	Penedo	41
Barra de São Miguel	54	Jaramataia	34	Piaçabuçu	50
Batalha	44	Joaquim Gomes	13	Pilar	52
Belém	20	Jundiaí	6	Pindoba	17
Belo Monte	17	Junqueiro	14	Piranha	19
Boca da Mata	14	Lagoa da Canoa	10	Poço das Trincheiras	7
Branquinha	18	Limoeiro de Anadia	6	Porto Calvo	28
Cacimbinhas	15	Maceió	92	Porto de Pedras	23
Cajueiro	30	Major Isidoro	16	Porto Real do Colégio	29
Campo Alegre	24	Maragogi	15	Quebrângulo	29
Campo Grande	20	Maravilha	23	Rio Largo	58
Canapi	8	Marechal Deodoro	37	Roteiro	77
Capela	24	Maribondo	36	Santa Luzia do Norte	72
Carneiros	18	Mar Vermelho	8	Santana do Ipanema	34
Chá Preta	9	Mata Grande	13	Santana do Mundaú	12
Coité do Nóia	10	Matriz de Camaragibe	44	São Brás	40
Colônia Leopoldina	38	Messias	17	São José da Laje	25
Coqueiro Seco	79	Minador do Negrão	9	São José da Tapera	7
Coruripe	15	Monteirópolis	30	São Luís do Quitunde	32
Delmiro Gouvêia	62	Murici	26	São Miguel dos Campos	34
Dois Riachos	23	Novo Lino	18	São Miguel dos Milagres	22
Feira Grande	13	Olho d'Água das Flores	37	São Sebastião	7
Feliz Deserto	62	Olho d'Água do Casado	43	Satuba	44
Flexeiras	18	Olho d'Água Grande	14	Tanque d'Arca	30
Girau do Ponciano	8	Olivença	9	Taquarana	9
Ibateguara	22	Ouro Branco	28	Traipu	16
Igaci	11	Palestina	41	União dos Palmares	31
				Viçosa	30

TABELA 2 – Taxas de urbanização municipais no Estado de Alagoas (em %), 1970.

ROL

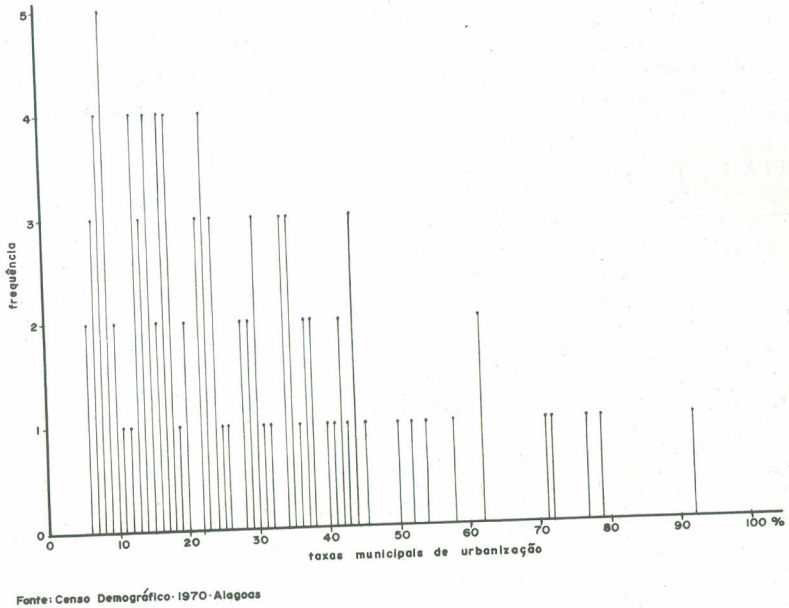
6	14	23	38
6	15	24	38
7	15	24	40
7	15	24	41
7	15	25	42
8	16	26	42
8	16	28	43
8	17	28	44
8	17	29	44
9	17	29	44
9	17	30	46
9	18	30	50
9	18	30	52
9	18	31	54
10	18	32	58
10	19	34	62
11	20	34	62
12	20	34	71
13	22	35	72
13	22	35	77
13	22	35	79
13	23	36	92
14	23	37	
14	23	37	

O diagrama de dispersão é um gráfico onde são registradas ocorrências sem classificação, permitindo a visão de como se distribuem os dados (fig. 2). Assim, observa-se a amplitude total e os dados que ocorrem com maior frequência. Através do diagrama de dispersão consegue-se, também, ver se existe agrupamento natural dentro dos dados, ou seja, a delimitação experimental (ou preliminar) de classes. Entretanto, quando os dados são muito numerosos, a construção de um rol ou de um diagrama de dispersão é trabalhosa ou mesmo impossível. Muitas vezes, o pesquisador recebe melhor impressão das observações, isto é, melhor imagem arranjando os dados em classes, em forma de tabela ou gráfico. Tem isto o efeito de eliminar os extremos.

Desta forma, a chamada tabela de frequência ou distribuição de frequência é uma disposição tabular dos dados por classes, juntamente com as frequências correspondentes. Para fazer a construção levantam-se as seguintes perguntas importantes: quantas classes precisam ser feitas para determinados dados e qual seria o intervalo de classe?

Como vimos no diagrama de dispersão, poderíamos fazer a divisão em classes, segundo o agrupamento natural, mas neste caso não teríamos, nem em progressão aritmética nem em progressão geométrica, um tamanho regular de classes. Mas a maior desvantagem do agrupamento natural é que, muitas vezes, é difícil ver onde fazer uma separaçao e o resultado é

bem subjetivo, sem base científica. Cada pesquisador determinaria provavelmente um outro número de classes para o mesmo conjunto de dados originais. Assim, é mais aconselhável aplicar um outro método para a divisão em classes.



Fonte: Censo Demográfico - 1970 - Alagoas

Figura 2

Para reduzir a perda de informações, um número pequeno de classes não pode ser feito. Por outro lado, classes muito numerosas não introduzem redução do grande número de dados originais, o que é justamente o objetivo para conseguir uma forma supervisionável. Com número grande de classes, os detalhes ficariam demonstrados. É necessário encontrar um número razoável de classes. Uma possibilidade seria através da fórmula de Sturges, que dá uma estimativa do número de classes (k) a ser utilizada. A fórmula de Sturges é: $k = 1 + 3{,}3 \log n$ onde n é o número total de observações e *log* é o logarítmo para a base 10. No nosso exemplo da taxa de urbanização em Alagoas temos $k = 1 + 3{,}3 \log 94$, considerando que se trata de 94 municípios. Assim, k seria 7,51123 e arredondando o número temos 8 classes. Devemos lembrar que a fórmula dá uma indicação do número de classes, não sendo uma fórmula inflexível.

Para achar o intervalo de classes dividimos a amplitude total através do mero de classe ou seja 86/8 = 10,75. Neste caso a amplitude dentro do intervalo de classe pode ser 10% e de uma classe à outra 11%.

É preciso salientar que o número de classes e sua amplitude dependem do número de ocorrências e da extensão total dos referidos dados. Assim, para cada conjunto de dados deve-se calcular de novo o número de classes e a amplitude.

Uma outra classificação provisória seria com base na regra geral, segundo a qual o número de classes nunca deveria ser maior do que cinco vezes o logaritmo do número total de observações, o que indicaria, no exemplo de Alagoas, não mais de 10 classes.

Depois de ter recebido uma sugestão do número de classes, segundo a fórmula de Sturges, e encontrado o intervalo do mesmo, podemos agrupar os nossos dados na tabela de freqüência.

TABELA 3 – Tabela de freqüência das taxas municipais de urbanização do Estado de Alagoas, 1970

intervalo de classe	ponto médio	freqüência (absoluta)	freqüência relativa (%)
6-16%	11	31	32,98
17-27%	22	23	24,47
28-38%	33	20	21,28
39-49%	44	9	9,57
50-60%	55	4	4,26
61-71%	66	3	3,19
72-82%	77	3	3,19
83-93%	88	1	1,06

Temos oito classes, onde cada uma tem o seu limite inferior, que é o seu valor mais baixo, e o limite superior, ou seja o valor mais alto. No nosso exemplo, a primeira classe, que vai de 6 - 16%, tem o limite inferior de 6% e o limite superior de 16%.

Os chamados limites reais de classe são obtidos adicionando-se o limite superior de uma classe ao inferior da seguinte e dividindo-se a soma por 2. No exemplo citado as classes com os limites reais teriam o seguinte intervalo: 5,5-16,54; 16,5-27,5; 27,5-38,5, etc.

A frequência absoluta se acha facilmente, contando as ocorrências para cada classe. Dividindo a frequência absoluta pelo número total das ocorrências achamos a frequência relativa, que é um valor importante para comparações. Ela pode ser indicada em forma de valores porcentuais.

O ponto médio da classe, que juntamos na tabela de frequência, é o valor representativo de cada uma. Esse valor acha-se somando o limite inferior ao limite superior de classe e dividindo-se a soma por dois.

Da tabela de frequência podemos ver que o agrupamento dos dados introduz aproximação. A maioria das observações cai na primeira classe, a minoria na última. A distribuição assimétrica é destacada.

Se tivéssemos trabalhado com dados contínuos poderíamos ter usado intervalos de classes abertos no lado direito. Isso é, intervalos onde o limite inferior pertence a classe, mas o limite superior pertence a classe superior. Por exemplo [6-17); [17-28) ... ou seja, isto quer dizer que a classe envolve valores de seis até menor do que dezessete. Cada valor maior ou igual a seis cai na primeira classe. Se o valor fosse de 16,99, este valor pertenceria também ainda à primeira classe, enquanto um valor de dezessete é da segunda.

Depois de ter estabelecido a tabela de frequência podemos utilizá-la para construir um histograma que é a representação gráfica de distribuições de freqüência, dando assim, uma boa impressão visual do conjunto de dados. Sobre a abscissa são indicados os intervalos de classe com os limites reais e com centro nos pontos médios respectivos, e sobre a ordenada, as frequências absolutas ou relativas (fig. 3).

Como vimos no gráfico, o histograma dá boa impressão visual da forma de distribuição. A assimetria se destaca claramente. A maioria dos municípios tem baixa taxa de urbanização no Estado de Alagoas.

O histograma que indica as frequências em forma relativa chama-se *histograma de frequência relativa* (fig. 3). Para os problemas geográficos é bastante útil juntar ao mesmo gráfico a frequência absoluta e relativa, como se vê na figura 3.

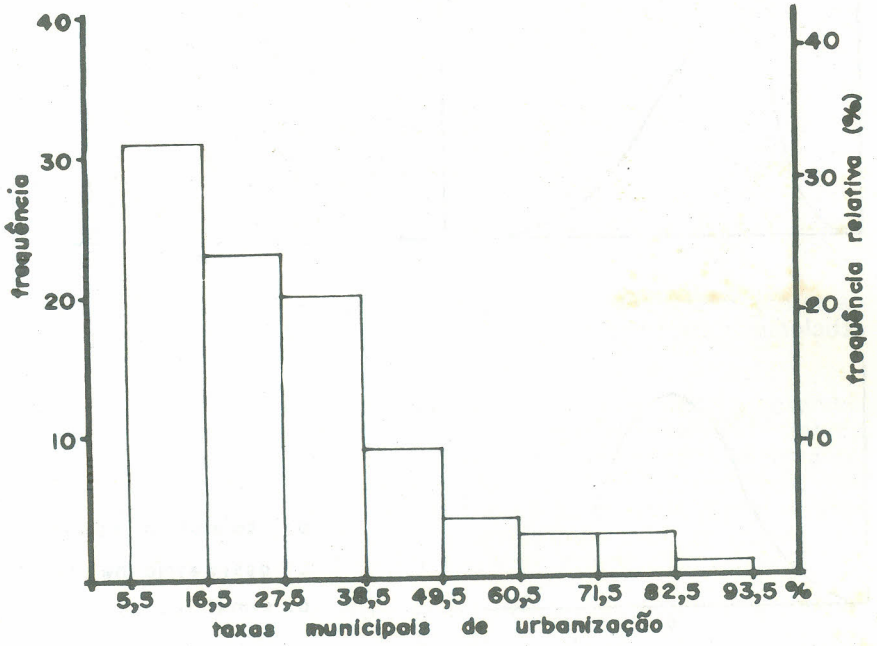


Figura 3

O polígono de frequência, outro tipo de representação gráfica, tem indicado na abscissa os pontos médios, na ordenada as frequências absolutas ou relativas. Os pontos são unidos através de uma linha reta. O polígono de frequência é utilizado particularmente para indicar a forma da distribuição de freqüência de um conjunto de dados. Na interpretação de um polígono de frequência, deve-se lembrar que a altura dos pontos que são unidos para formar o polígono, representa as frequências das classes, nas quais os dados foram agrupados e não os dados individuais. Fala-se de polígono de frequência relativa se as frequências relativas são indicadas. Aqui também podemos juntar no mesmo gráfico, sobre eixos diferentes, as frequências absolutas e relativas. Com o aumento do número de classes o polígono de frequência tende a ser uma curva, assim chamada curva de frequência.

Como mencionamos anteriormente, no exemplo da taxa de urbanização, temos uma distribuição de frequência assimétrica, ou seja, a maioria dos dados está no lado esquerdo do gráfico. Fala-se neste caso de assimetria positiva, ou que a curva de frequência é desviada para a direita (fig. 4). Este tipo de distribuição é particularmente comum na geografia humana. Um outro tipo de assimetria, a assimetria negativa, onde a maioria dos dados se encontra no lado direito, é a mais rara na geografia (fig. 4). Neste caso se diz também que a curva é desviada para a esquerda. Por outro lado, a forma simétrica da distribuição é encontrada muitas vezes na geografia física (fig. 4).

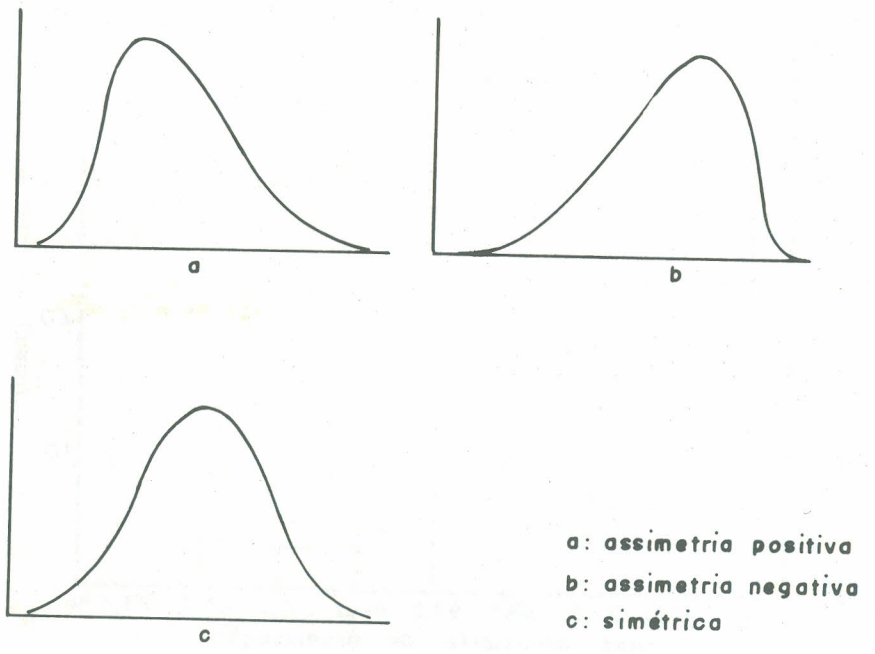


Figura 4

Além destes exemplos gráficos, os dados podem ser representados através do polígono de frequência acumulada ou ogiva (fig. 5). Na tabela de frequência acumulada, ou simplesmente distribuição de frequência acumulada, que serve de base para o gráfico, são representadas as frequências acumuladas para o nosso exemplo da taxa de urbanização no Estado de Alagoas. Podemos acumular a frequência total dos valores abaixo ou igual de qualquer limite superior de classe, ou acumular a frequência total acima ou igual ao limite inferior de classe. A tabela da frequência acumulada mostra o resultado para o nosso exemplo.

TABELA 4

		Freqüência acumulada	Freqüência acumulada (%)
abaixo ou igual a	5	0	0
	16	31	32,98
	27	54	57,45
	38	74	78,72
	49	83	88,30
	60	87	92,55
	71	90	95,74
	82	93	98,94
	93	94	100,00
acima ou igual a	6	94	100,00
	17	63	67,02
	28	40	42,55
	39	20	21,28
	50	11	11,70
	61	7	7,45
	72	4	4,26
	83	1	1,06
	94	0	0

Como nos gráficos anteriores é aconselhável, na Geografia, juntar ao mesmo gráfico uma segunda ordenada com as frequências acumuladas relativas percentuais (fig. 5).

O polígono de frequência acumulada, seja para as frequências absolutas seja para as relativas, é muito útil na intepretação geográfica, embora ainda pouco utilizado. Através das ogivas é possível deduzir rapidamente, no nosso exemplo da taxa de urbanização, o número de municípios com taxa de urbanização entre duas determinadas porcentagens. Assim, através do gráfico podemos bem resumir, à uma forma manejável, um grande conjunto de dados e, ao mesmo tempo, fornecer informações valiosas.

Um método útil para a interpretação na Geografia é a *curva de Lorenz*, que tem como objetivo indicar até que ponto uma distribuição é desigual (não uniforme) em comparação à uma distribuição totalmente uniforme. Essa curva é tradicionalmente utilizada para mostrar a distribuição da renda em relação à população, mas ela tem muitas outras possibilidades de

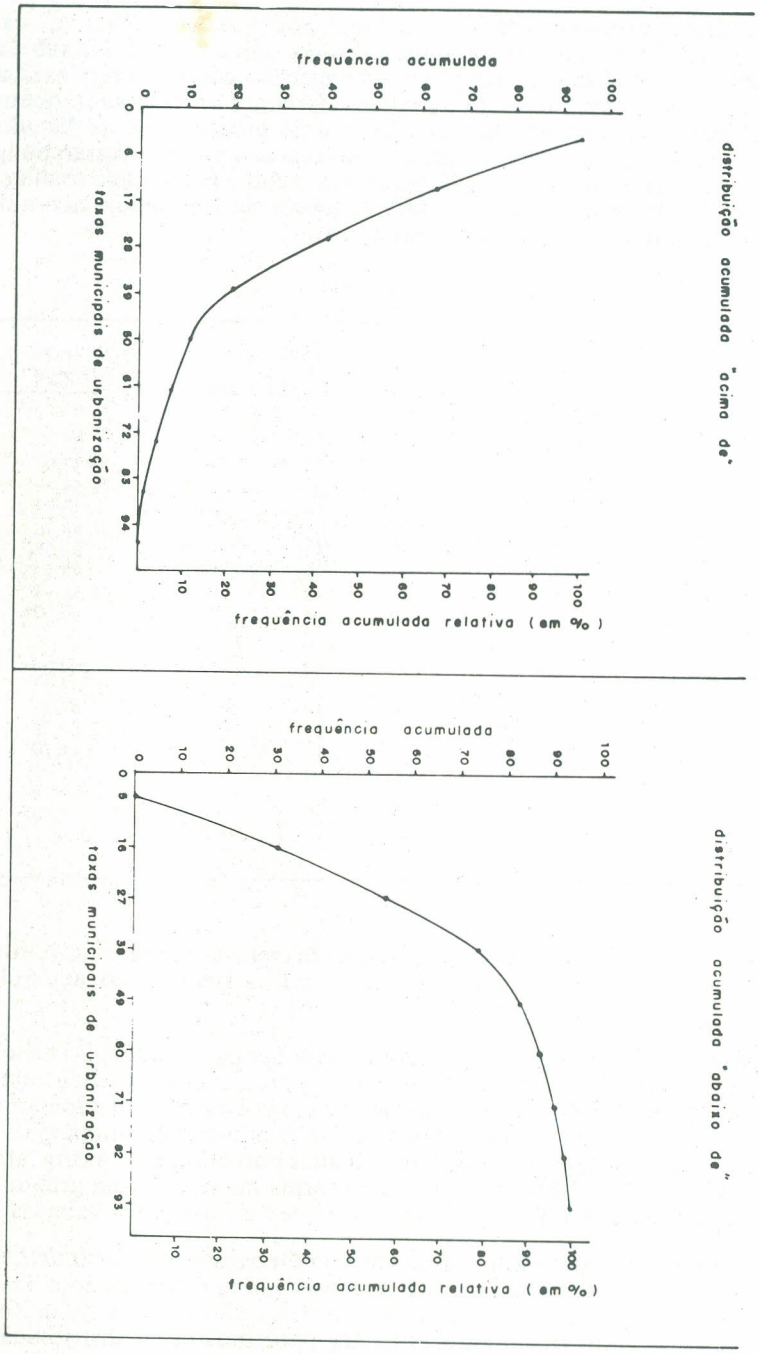


Figura 5

aplicação na Geografia, como expressar, por exemplo, a relação da distribuição da população e a área ocupada, da população e da alimentação, da renda e do número de crianças, etc. Para melhor mostrar o efeito da curva de Lorenz podemos ilustrá-la no exemplo da distribuição da renda em relação à população da área urbana em Salvador.

Foram levantadas as rendas "per capita" da população e fixados em sete classes os níveis de renda mensal. As porcentagens de pessoas referentes à cada classe e as da renda recebida, segundo as classes, podem ser vistas na tabela 5. As frequências acumuladas correspondentes às frequências relativas porcentuais das pessoas e da renda, foram determinadas e colocadas sobre um gráfico, onde a abscissa representa a população e a ordenada a renda, sendo que os dois eixos têm escalas comparáveis (fig. 6).

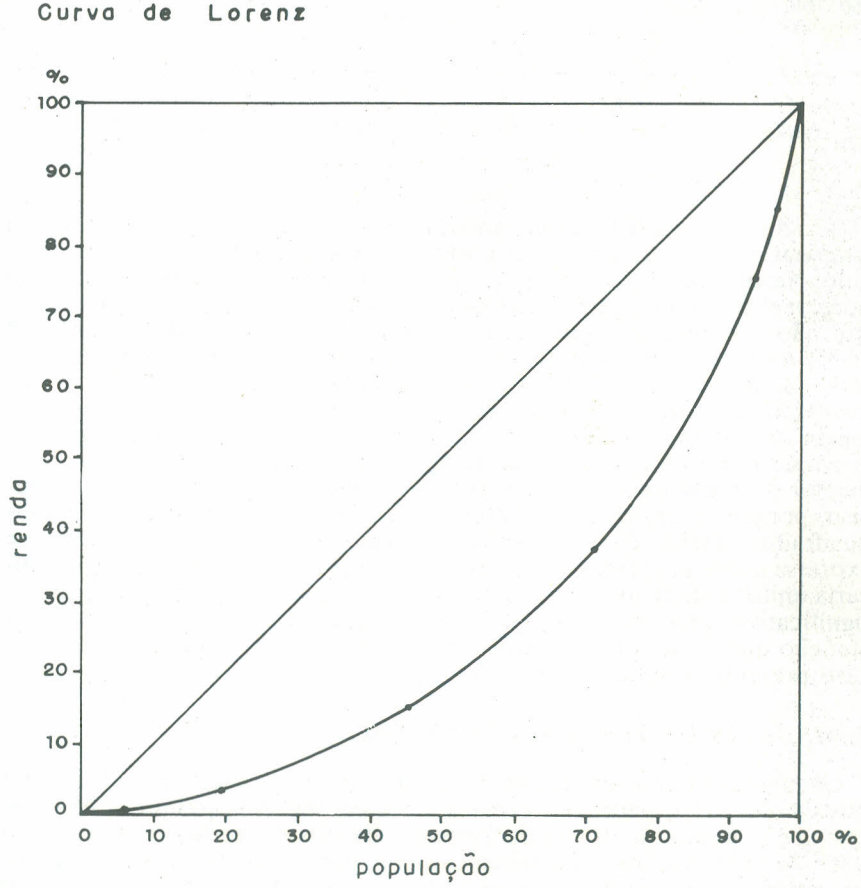


Figura 6

TABELA 5 – Distribuição da população e da renda familiar, na área urbana de Salvador, por níveis de renda *per capita* mensal – 1966

níveis de renda *per capita* mensal (Cr$ 1,00)	dados simples		dados acumulados	
	porcentagem de pessoas	porcentagem da renda total recebida	percentagem de pessoas	porcentagem da renda total recebida
menos de 10	5,7	0,6	5,7	0,6
10 - 20	14,4	3,2	20,1	3,8
20 - 40	26,0	11,7	46,1	15,5
40 - 80	25,4	22,0	71,5	37,5
80 - 160	22,1	38,0	93,6	75,5
160 - 240	3,3	9,7	96,9	85,2
240 - 400	3,1	14,8	100,0	100,0

Fonte: ETENE/BNB (1969) p. 43
Observação: O salário mínimo em Salvador era, na época, de Cr$ 82,50.

Se a distribuição fosse totalmente uniforme, a curva seria uma linha reta diagonal, ou seja, a curva estaria totalmente sobre a hipotenusa do triângulo. Neste caso, a porcentagem da renda recebida seria exatamente proporcional à porcentagem das pessoas nas diversas classes. No nosso caso isto não acontece, porque quase a metade das pessoas (46,1%) tinha só 15,5% da renda. Visualmente podemos determinar que quanto menos uniforme a distribuição mais côncava é a curva. Podemos ainda indicar a uniformidade da distribuição calculando a área abaixo da curva como porcentagem da distribuição teorética, totalmente uniforme. A desvantagem da curva de Lorenz é que a interpretação visual é só uma aproximação rudimentar e o mencionado cálculo também não fornece um índice muito preciso, porque se realiza normalmente na base de contagem de pequenos quadrados, abaixo da curva e no triângulo. A área abaixo da curva é expressa como porcentagem da área do triângulo. Um índice de 100% indicaria uniformidade total, enquanto um índice de 20-30% espelha um muito significativo grau de agrupamento. No entanto, precisa-se chamar a atenção que índices iguais podem resultar de curvas bem diferentes. No caso presente o índice seria 52,75%.

5. *MEDIDAS DA TENDÊNCIA CENTRAL*

Os métodos que resumem os dados em forma de tabelas e gráficos sem dúvida são úteis, particularmente no começo de uma pesquisa. Mas, ao resumir os dados, já através da formação de classes, perde-se uma quantidade de informações. Não sabemos mais, por exemplo, como é a distribuição dos dados dentro da classe, isto é, se os dados se agrupam perto do ponto médio, do limite inferior ou superior da mesma, ou se eles são distribuídos regularmente dentro desta. Por outro lado, esta perda de detalhes é

recompensada através da possibilidade de chamar a atenção sobre as características do conjunto de dados.

Os métodos tabulares e gráficos também não fornecem informação exata e quantitativa sobre a distribuição. Se temos, por exemplo, dois histogramas podemos ver que eles diferem entre si, mas não podemos até agora indicar, através de um valor numérico, como eles diferem. Assim, para melhor descrever uma sequência de dados ou uma distribuição de frequência, devemos ainda encontrar algumas medidas características, os chamados parâmetros. Para algumas descrições, particularmente a distribuição normal, que tem a forma regular, simétrica, em forma de um sino, dois parâmetros servem para fazer uma boa descrição. Um mede a tendência central o outro a dispersão. Para outras distribuições, precisamos ainda de duas outras medidas: a da simetria e a da curtose.

Se observarmos um conjunto de dados, percebemos que em quase todos os dados os valores tendem a se agrupar em torno de um valor central. Parece que este é típico para o conjunto de dados e que localiza o valor central da distribuição.

Para a Geografia três tipos das chamadas medidas da tendência central, ou medidas de localização, são importantes. Vamos apresentar as medidas e depois discutir as vantagens e desvantagens.

5.1. *Média aritmética*

A média aritmética, ou simplesmente a *média*, encontra-se adicionando todos os valores e dividindo-os pelo número total dos valores. Em termos matemáticos, a fórmula para a média aritmética é a seguinte:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

onde

$\bar{X}$ = a média

Σ = a letra grega "sigma" que significa "soma"

X = o valor individual

n = o número de ocorrências ou observações

subscrito i = um índice

1 e n = a amplitude do somatório, ou seja, devem-se somar todos os dados do primeiro até o último que é o n-ésimo.

Ao invés de escrever $\sum_{i=1}^{n} X_i$ podemos fazê-lo também $X_1 + X_2 + X_3 + \ldots\ldots\ldots + X_n$ o que significaria a mesma operação.

Aplicando a fórmula da média aritmética para o nosso exemplo da taxa de urbanização em Alagoas temos

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2594}{94} = 27{,}5957 = \sim 27{,}60, \text{ ou seja,}$$

a taxa de urbanização média no Estado de Alagoas é de 27,60%.

Para poder utilizar esta fórmula devemos ter os nossos dados em forma não agrupada, o que não é sempre o caso para os dados geográficos. Se os dados são já fornecidos agrupados, o que ocorre muitas vezes na Geografia, podemos encontrar a média aritmética através da fórmula seguinte:

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{K} f_i}$$

onde

K = o número de classes

f_i = a frequência de i-ésima classe

X_i = o ponto médio da i-ésima classe

$$\sum_{i=1}^{K} f_i = n$$

Se o número de observações é grande e o intervalo de classes pequeno, a aproximação para a média $\bar{X}$ dos dados não agrupados é bastante parecida. Para nosso exemplo podemos sair da seguinte tabela:

TABELA 6

Ponto médio X_i	freqüência f_i	produto $f_i X_i$
11	31	341
22	22	506
33	20	660
44	9	396
55	4	220
66	3	198
77	3	231
88	1	88
	94	2.640

Então, a média aritmética para dados agrupados seria

$$\bar{X} = \frac{2640}{94} = 28{,}0851 = \sim 28{,}09$$

Comparando este valor com o dos dados não agrupados, constatamos aproximação relativamente grande.

5.2. Mediana

Para achar a mediana, uma outra medida da tendência central, devemos arranjar os nossos dados em ordem crescente ou decrescente e em seguida encontrar o valor central. Se o número total das ocorrências é número par, neste caso a mediana está entre os dois números centrais, ou seja deve-se somar os dois valores centrais e dividir por dois.

Para determinar a mediana para a taxa de urbanização no Estado de Alagoas podemos sair do rol, já construído anteriormente. Temos n = 94 e a mediana deve-se encontrar entre o 47-iésimo e o 48-iésimo valor. Sendo que o 47-iésimo valor é 23% e o 48-iésimo também, a mediana é (23 + 23) : 2 = 23%. É o valor para qual a metade da frequência total fica situada abaixo e a outra metade acima dele. Ou, em outras palavras: 50% dos municípios tem uma taxa de urbanização abaixo de 23 e 50% dos municípios tem uma taxa de urbanização acima de 23%.

Em algumas situações da pesquisa geográfica, a mediana é uma medida importante. Por exemplo, determinar a mediana dos salários é um melhor método, para indicar a tendência central do que a média aritmética, porque as poucas muito altas rendas aumentariam o valor da média aritmética de tal maneira, que esta não seria mais uma medida típica das rendas.

A construção do rol é trabalhosa se temos muitos dados. Neste caso, é melhor construir uma tabela da distribuição de frequência para encontrar a mediana para dados agrupados. Este cálculo da mediana é baseado na interpolação linear dentro da classe que contém a mediana. Percebemos uma boa aproximação da mediana se os valores são distribuídos uniformemente dentro da classe. Saíndo assim da hipótese da equirepartição dos valores dentro da classe, aplicamos a seguinte fórmula.

$$m = L_1 + \frac{\frac{n}{2} - (\Sigma f)_1}{f_{mediana}} \cdot c$$

L_1 = limite inferior real da classe mediana, i.e. da classe que contém a mediana.

n = frequência total

$(\Sigma f)_1$ = soma de todas as frequências das classes inferiores à mediana

$f_{mediana}$ = frequência da classe mediana

c = amplitude do intervalo da classe mediana

Para melhor explicação, voltamos ao nosso exemplo da taxa de urbanização em Alagoas. A tabela seguinte resume os fatos importantes que precisamos.

TABELA 7

limites reais de classes	frequência	frequência acumulada
5,5 - 16,5	31	31
16,5 - 27,5	23	54
27,5 - 38,5	20	74
38,5 - 49,5	9	83
49,5 - 60,5	4	87
60,5 - 71,5	3	90
71,5 - 82,5	3	93
82,5 - 93,5	1	94

Considerando que a primeira classe só engloba trinta e uma frequências, a mediana deve estar dentro da segunda. Logicamente podemos reconstruir a fórmula mencionada através da regra de três. Na segunda classe no intervalo de classe de 11 temos 23 frequências. Queremos saber que amplitude de intervalo corresponderia à frequência 94/2 - 31, ou seja à 16. Assim, 11 : 23 = x : 16. O resultado é x = 7,7.

Desta maneira, a mediana para dados agrupados é 16,5 (o limite inferior real da classe mediana) mais 7,7, obtendo-se como resultado 24,2, i.e., determinamos a mediana para dados agrupados como sendo 24,2%, no exemplo da taxa de urbanização no Estado de Alagoas. Vemos que este valor difere um pouco do valor da mediana, encontrado para dados não agrupados, sendo que a distribuição não é tão uniforme dentro da classe.

Um outro meio para encontrar a mediana é através da ogiva da distribuição. A mediana corresponde à frequência acumulada relativa igual à 50%. Se desenhamos as ogivas "acima de" e "abaixo de", no mesmo gráfico, a intersecção das duas corresponde ao valor de 50%.

5.3. Moda

A moda, um terceiro método para medir a tendência central, é o valor que ocorre com maior frequência. Existem conjuntos de dados sem moda e, por outro lado, se a moda existe ela pode não ser única. Por exemplo, o pequeno conjunto de dados 4, 6, 9, 13, 16, 17 não tem uma moda. Entretanto, o conjunto de 11, 6, 6, 3, 5, 7, 7, 7, 8, 8, 8 tem duas modas: 7 e 8. Diz-se também que o conjunto é bimodal. No conjunto 6, 1, 3, 2, 2, 4, 4, 4, 4 a moda é 4 sendo assim unimodal. No nosso exemplo da taxa de urbanização em Alagoas a moda é 9. Como vimos, a moda às vezes não existe ou temos mais de uma. Assim, para dados não agrupados, a moda não é uma medida muito útil da tendência central.

Se os dados são agrupados encontramos normalmente uma classe com uma frequência máxima. Essa classe chama-se modal. Para a taxa de urbanização em Alagoas, a classe modal é a primeira onde 31 dos 94 municípios caem dentro.

Existem relações mútuas entre as três medidas da tendência central. Se temos um conjunto de dados com distribuição totalmente simétrica, normal, neste caso o valor da média aritmética, da mediana e da moda é igual. Por outro lado, se a distribuição de um conjunto de dados tem uma distribuição com assimetria positiva, os três valores médios são diferentes uns dos outros. A média tem o valor mais alto e a moda o menor. Verificamos este caso com o nosso exemplo de Alagoas, onde a distribuição indica assimetria positiva. Como vimos, a moda tem com 9% o menor valor, a mediana com 23% um valor intermediário e a média aritmética com 27,60%, o maior valor. Considerando que a simetria positiva é muito frequente nos conjuntos de dados geográficos, a relação entre os três tipos da tendência central mostra frequentemente esta tendência mencionada. Uma distribuição com assimetria negativa tem a média aritmética com o menor valor e a moda com o maior. A mediana se encontra entre o valor da média e da moda.

5.4. Algumas observações sobre as medidas de tendência central

As três medidas de tendência central são utilizadas na Geografia. A escolha de uma delas pelo pesquisador dependerá principalmente do conjunto de dados e também do objetivo da pesquisa. Ele deve estar cons-

ciente da vantagem e desvantagem de cada uma destas medidas, para cada situação.

A moda indica o que é comum e frequente. Mas de vez em quando não é fácil dizer exatamente onde está a moda. Isto ocorre particularmente quando temos uma distribuição multimodal, com mais ou menos os mesmos valores. Por outro lado, a classe modal varia também em função da seleção das classes. Assim, a moda pode tender a ser uma forma imprecisa do valor médio. Por outro lado, a moda não tem realmente qualidades matemáticas; não podemos tirar dela outras características do conjunto de dados. Entretanto, para a apresentação gráfica, a moda é um eficiente método.

Na determinação da mediana cada ocorrência tem o mesmo peso, tratando-se de um valor baixo como de um valor alto. Assim, conjuntos bem diferentes podem ter a mesma mediana, como mostra o exemplo seguinte, onde cada vez a mediana é 300.

1, 5, 300, 800, 1000

290, 295, 300, 301, 302

A mediana também não é realmente um valor matemático.

A média aritmética, por seu lado, é matematicamente fundamentada. No cálculo, cada ocorrência tem o peso segundo sua magnitude. É importante se conscientizar destas consequências: este fato é particularmente desvantajoso quando a distribuição é assimétrica. Se temos, por exemplo, uma forte assimetria positiva, o que ocorre muitas vezes na Geografia, os poucos altos ou mesmo extremamente altos valores pesam muito e têm, como consequência, que a média se desloca em direção aos extremos altos e é mais alta do que a mediana ou a moda. Nestes casos de forte assimetria, positiva ou negativa, é melhor indicar a mediana ou a moda. Maior a assimetria, menos representativa é a média aritmética.

6. MEDIDAS DE DISPERSÃO

Para bem descrever um conjunto de dados não basta só indicar a tendência central. Particularmente quando queremos fazer a comparação entre dois ou mais conjuntos de dados, poderia acontecer que os valores médios são os mesmos ou quase assim, para os diversos grupos de dados, mas a distribuição varia muito. Um conjunto pode ter os dados próximos à média, o outro bem disperso. As medidas da tendência central descrevem só um aspecto dos dados, mas elas não fornecem informação sobre um outro aspecto da mesma importância, isto é, sobre o grau de dispersão, ou seja, o grau com que os dados tendem a dispersar-se em torno de um valor central. Desse modo, além das medidas da tendência central devemos conhecer as medidas de dispersão ou de variação.

Existem muitas medidas de dispersão. As mais comuns são a amplitude total, o desvio quartílico, o desvio médio e o desvio padrão, sendo que a amplitude total é a medida mais simples e elementar e o desvio padrão a mais complexa. Como em relação às medidas da tendência central, aqui também o pesquisador deve escolher a medida mais adequada para seu trabalho.

6.1. Amplitude total

A amplitude total, como medida mais simples de dispersão, é rápida para encontrar e dá uma primeira impressão sobre a dispersão dos dados. Se temos por exemplo dois conjuntos de dados

1, 4, 7, 10, 13 e

4, 5, 7, 8, 11

os dois têm a mesma média aritmética que é 7, mas a dispersão é bem diferente. No primeiro exemplo a dispersão vai de 1 até 13, ou seja, a amplitude total é 12. No segundo exemplo a dispersão é de 4 até 11 e assim a amplitude total é 7. Em termos gerais a amplitude total pode ser definida como X(max) - X (min), onde X (max) e X(min) são os valores maiores e menores dos n-valores.

A medida da amplitude total dá uma rápida informação sobre a dispersão dos dados, mas é medida imprecisa porque o cálculo envolve só dois valores observados, não importa se trabalhamos com muitos dados ou poucos. Não se tem informação alguma sobre a distribuição dos dados dentro do intervalo ou sobre o número de dados que estão perto da média.

Por exemplo, nos conjuntos 1, 2, 6, 6, 6, 6, 6, 10, 11 a média é 6, a amplitude total 10 e os dados estão perto da média, enquanto que no conjunto 1, 1, 1, 1, 6, 11, 11, 11, 11, , a média é também 6 e a amplitude 10, mas os dados não se agrupam perto deste. Neste exemplo, vê-se que diversos conjuntos de dados podem ter a mesma média e a mesma amplitude total, mas distribuições diferentes dentro dos conjuntos de dados.

Se o número de observações aumenta, a amplitude total cresce normalmente. Entretanto, para um pequeníssimo número de valores, a amplitude total pode dar resultado satisfatório. Ela distorce a realidade quando há valores extremos nos dados em consideração. Por último, podemos dizer que não é aconselhável utilizar a amplitude total para comparar a dispersão em diversos conjuntos de dados, se estes não tem, pelo menos aproximadamente, o mesmo número de observações. Assim visto, presumimos que a amplitude total é uma medida relativamente boa de variação para um pequeno número de dados, mas se o número aumenta, esta medida é menos aconselhável.

Na Geografia, o método da amplitude total aplica-se particularmente bem em pesquisas climatológicas. Por exemplo, podemos mapear o valor e a amplitude total das temperaturas médias anuais ou de um determinado mês do ano, dos últimos 20 ou 30 anos, das estações existentes no Brasil, ou em determinados Estados, e ligar as áreas com as mesmas amplitudes através de isolinhas. O resultado cartográfico oferece boa impressão da variação dentro dos anos. Ao invés de tomar a variável temperatura, podemos também tomar, por exemplo, a da precipitação.

6.2. Desvio quartílico

Já foi determinada a mediana dividindo-se o número total de observações em duas partes iguais, assim 50% dos dados se localizam acima da

mediana e 50% abaixo. Estas duas partes podem ser subdivididas de novo, de tal maneira que encontramos quatro grupos com o mesmo número de ocorrências. Neste caso, cada grupo envolve 25% dos dados. O quartil que separa as 25% dos mais baixos valores chama-se quartil inferior (Q1); Q2, o limite de 50%, corresponde à mediana e Q3, o quartil superior é o que separa os 25% dos mais altos valores. A diferença entre o 3º e o 1º quartil (Q3 - Q1) é a amplitude interquartílica que engloba os 50% centrais dos dados.

Observa-se que quanto menor a amplitude interquartílica, mais são os dados agrupados em torno da mediana, isto é, maior é a concentração em torno do valor central. Assim, a amplitude interquartílica é índice rudimentar da dispersão. Se a curva de distribuição fosse simétrica, neste caso os quartis 1 e 3 seriam (Q3 - Q1)/2 distante da mediana. A igualdade ou a diferença entre as distâncias de cada quartil à mediana informa sobre a simetria ou assimetria da distribuição.

No exemplo da taxa de urbanização, com 94 observações, tínhamos determinada a localização da mediana (Q2) entre a 47-ésima e a 48-ésima ocorrência que corresponde à 23%. Sendo que 47 ocorrências caem abaixo deste valor e 47 acima, o valor de Q1 deve ser o 24-ésimo ou seja 14% e o valor de Q3 o 71-ésimo ou seja 37%. A amplitude interquartílica é assim 37-14 = 23. Dividindo este valor por dois temos a amplitude semi-interquartílica ou o desvio quartílico.

Este método tem as vantagens e desvantagens da mediana. Para poucos dados é um cálculo rápido, muitas vezes utilizado para dados da climatologia, particularmente da precipitação. Mas não é realmente uma medida da dispersão das ocorrências em torno do valor central, porque, como no caso da mediana, a magnitude da ocorrência não é considerada. Trabalhamos só com o número de ocorrências entre ou acima de alguns pontos, mas não consideramos a magnitude, o valor da ocorrência.

A amplitude interquartílica fornece descrição adequada dos 50% centrais da distribuição, mas não considera os extremos da distribuição. Assim tem justamente a desvantagem contrária da amplitude total. A amplitude total determina só os extremos e ignora a distribuição dos itens perto da média. Por outro lado, o desvio quartílico não é influenciado pelos valores abaixo de Q1 ou acima de Q3.

Ao invés de calcular os quartis, podem ser calculados também outros percentis.

Uma vez que a ogiva seja construída, será fácil determinar os valores de Q1, Q2 e Q3, sendo que eles correspondem à 25, 50, e 75% das ocorrências.

6.3. Desvio médio

O desvio médio é obtido calculando as distâncias das ocorrências a partir da média, ou seja, calculando os desvios da média. Estes desvios são determinados para cada ocorrência devendo-se calcular em seguida a média destas distâncias, que são tomadas em termos absolutos. O resultado é o desvio médio dos valores individuais da média. O desvio médio considera cada valor de um conjunto de dados.

Em termos matemáticos a definição é:

$$\text{Desvio médio} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

onde $\bar{X}$ = a média aritmética

$|X - X|$ = o valor absoluto do desvio X_i em relação à $\bar{X}$. O valor absoluto é indicado através de duas linhas verticais.

Quando calculamos o desvio médio, não consideramos a direção dos desvios individuais, isto é, se eles são abaixo ou acima da média, mas simplesmente os valores absolutos.

6.4. Variância e desvio padrão

Ao invés de utilizar o desvio médio dos valores da média, na prática o desvio médio quadrado, a chamada variância, é mais utilizada. Neste caso, não retiramos o sinal utilizando os valores absolutos, mas elevando ao quadrado todos os desvios. Desta maneira, o sinal torna-se sempre positivo. A soma dos desvios da média elevados ao quadrado é dividida pelo número total de observações. A variância é definida através da fórmula:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

ela é a média dos desvios da média ao quadrado. Como os desvios são elevados ao quadrado, a variância é expressa em unidades quadradas.

Mais importante do que a variância é o desvio padrão, que indica a dispersão nas mesmas unidades de medidas como os dados originais. Este é definido por:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

O desvio padrão é a raiz da média dos quadrados do desvio da média e é medida do desvio dos valores singulares do valor central do conjunto de dados. Se os valores estão próximos uns dos outros, a soma dos quadrados é pequena e, desta maneira, também o desvio padrão. Por outro lado, se os valores estão bem distantes uns dos outros a soma dos quadrados é grande. Ou seja, em outras palavras: maior a dispersão, maior o desvio padrão.

O cálculo do desvio médio, da variância e do desvio padrão é mostrado no exemplo da precipitação anual de Salvador durante 20 anos (1956-1975) (tab. 8). Neste exemplo a média aritmética é 2.076,0 mm de precipitação anual, o desvio médio é de 432,3 mm e o desvio padrão 542,0 mm. Observa-se que o desvio padrão é maior do que o desvio médio. Para distribuições normais o desvio padrão é 1,25 vezes o desvio médio.

6.5. Métodos alternativos para o cálculo da variância e do desvio padrão

Com muitos dados é fatigante calcular a variância ou o desvio padrão segundo a fórmula indicada, mas podemos encontrar facilidades no processo do cálculo, que é baseada na modificação da fórmula, de tal maneira que o número de cálculos individuais seja reduzido. Isto economiza tempo de trabalho reduzindo-se assim a possibilidade de se cometer erros.

TABELA 8 – Precipitação anual em Salvador (1956-1975). Cálculo do desvio médio, da variância e do desvio padrão

Precipitações anuais (em mm) X_i	$(X_i - \bar{X})$	$(X_i-\bar{X})^2$
946,1	- 1129,9	1276674,0
1389,8	- 686,2	470870,4
1527,8	- 548,2	300523,2
1559,9	- 516,1	266359,2
1752,6	- 323,4	104587,6
1821,9	- 254,1	64566,8
1838,0	- 238,0	56644,0
1842,7	- 233,3	54428,9
1936,6	- 139,4	19432,4
1938,7	- 137,3	18851,3
1958,3	- 117,7	13853,3
2150,5	74,5	5550,3
2208,0	132,0	17424,0
2409,7	333,7	111355,7
2427,6	351,6	123622,6
2459,7	383,7	147225,7
2479,7	403,7	162973,7
2647,1	571,1	326155,2
2807,0	731,0	534361,0
3417,7	1341,7	1800158,9
41519,4	8646,6	5875618,2

$\bar{X} = \frac{41519,4}{20} = 2075,97 = 2076,0$ mm

desvio médio $= \frac{8646,6}{20} = 432,3$ mm

variância $= \frac{5875618,2}{20} = 293780,9$

desvio padrão $= \sqrt{293780,9} = 542,0$ mm

Fonte: Ministério da Agricultura, Departamento Nacional de Meteorologia.

A fórmula da variância é, como vimos;

$$s^2 = \frac{\Sigma (X - \bar{X})^2}{n}$$. Infere-se que, podemos escrever:

$$s^2 = \frac{\Sigma (X^2 - 2\bar{X}X + \bar{X}^2)}{n}$$

$$s^2 = \frac{\Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}^2}{n}$$

$$s^2 = \frac{\Sigma X^2}{n} - 2\bar{X}^2 + \bar{X}^2$$

$$s^2 = \frac{\Sigma X^2}{n} - \bar{X}^2$$

assim, para a variância as fórmulas seguintes podem ser utilizadas:

$$s^2 = \frac{\Sigma X^2}{n} - \bar{X}^2 \qquad \text{ou} \qquad s^2 = \frac{\Sigma X^2 - (\Sigma X)^2/n}{n}$$

Para o desvio padrão aplicam-se estas fórmulas:

$$s = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} \qquad \text{ou} \qquad s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n}}$$

Nestas alternativas da fórmula precisamos só 25 ou 27 operações individuais para calcular o desvio padrão para a precipitação anual em Salvador, ao invés de 43 operações segundo a primeira versão da fórmula. Assim, de um lado economizamos tempo, mas por outro lado surge a desvantagem de termos números bem grandes causados pela elevação ao quadrado.

O cálculo segundo a fórmula $s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n}}$

é mostrado na tabela 9, no exemplo da precipitação anual em Salvador. Observa-se que o resultado é idêntico ao resultado da tabela 8.

TABELA 9 – Cálculo da variância e do desvio padrão para a precipitação anual em Salvador (1956-1975). (Cálculo simplificado)

Precipitações anuais (em mm) X_i	X_i^2
946,1	
1389,8	895105,2
1527,8	1931544,0
1559,9	2334172,8
1752,6	2433288,0
1821,9	3071606,8
1838,0	3319319,6
1842,7	3378244,0
1936,6	3395543,3
1938,7	3750419,6
1958,3	3758557,7
2150,5	3834938,9
2208,0	4624650,3
2409,7	4875264,0
2427,6	5806654,1
2459,7	5893241,8
2479,7	6050124,8
2647,1	6148912,1
2807,0	7007138,4
3417,7	7879249,0
	11680673,3
41519,4	
	92068647,0

$$s^2 = \sqrt{\frac{\Sigma X^2 - (\Sigma X)\,2/n}{n}}$$

$$s^2 \quad \frac{92068647,0 - \dfrac{1723860576,4}{20}}{20}$$

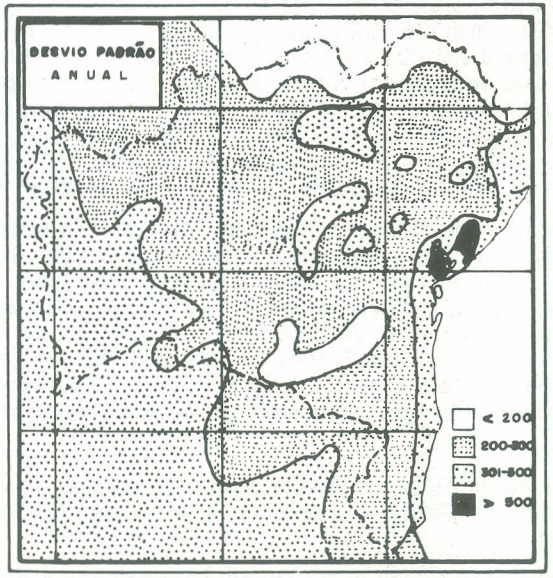
$$s^2 = 293780,9$$

$$s = 542,0 \text{ mm}$$

6.6. Aplicação das medidas de dispersão

Na Geografia, as medidas de dispersão são de importância especial em casos que as medidas da tendência central, isoladamente, não têm muito sentido. Segundo Bahrenberg e Giese (1975, p. 43), isto é particularmente relevante para vários problemas na climatologia, onde justamente a variação, ou seja, a irregularidade, muitas vezes tem mais consequência e é mais significativa do que o simples valor da média.

O desvio padrão é importante medida de dispersão, podendo ser utilizada para diversos problemas na Geografia. Ao invés de calcular o desvio padrão só de uma estação meteorológica, como fizemos para Salvador, poderíamos calculá-lo para todas as estações de um país. Estado, etc, mapear os resultados e unir os lugares do mesmo valor do desvio padrão, através de isolinhas. Isto pode ser realizado, por exemplo, para cada mês ou para os dados anuais, como mostra a fig. 7, no exemplo do Estado da Bahia. Destacam-se acentuadas diferenças na variação dentro do Estado, sendo que as maiores variabilidades de precipitações obviamente, ocorrem nas áreas mais chuvosas. Estes tipos de mapas podem servir de maneira excelente para a análise das possibilidades agro-climáticas. Poderíamos também mapear, ao inveś dos desvios padrões, as amplitudes totais para todas as estações construindo linhas da mesma variabilidade. Normalmente, neste caso, o resultado é satisfatório, porque se trata de poucos dados para cada estação.



Fonte: Bahia: SEPLANTEC (1976) p. 31, fig. 6

Figura 7

Outro exemplo da aplicação geográfica do desvio padrão será, com base em Bahrenberg e Giese (1975, p.45), o seguinte: normalmente os rendimentos dos cultivos são indicados como rendimento /ha, o que, na maioria dos casos, tem sentido. Se temos, por outro lado, cultivos em regiões onde os rendimentos variam muito de ano para outro por causas climáticas, neste caso o rendimento médio/ha, isoladamente tomado, não representa a realidade. Só junto com o desvio padrão o valor médio tem sentido. Quando se trata de uma pesquisa sobre uma região mais ou menos extensa, seria interessante mapear os resultados dos rendimentos médios/ha e os desvios padrões. Uma tal representação possibilita destacar com precisão as regiões agrícolas favoráveis das não favoráveis.

6.7. Medidas relativas

Se queremos comparar a variabilidade entre diversos conjuntos de dados, que tem médias bem diferentes ou unidades de medidas diferentes, neste caso o coeficiente de variação é uma medida melhor, indicando a variação relativa. Ele é definido em termos matemáticos pela fórmula:

$$V = \frac{s}{\bar{X}}$$ e geralmente expresso em porcentagem.

Se temos, por exemplo, uma estação registrando precipitação anual de 686,1 mm e um desvio padrão de 419,3 mm, como é o caso da estação de Bonito, no município de Utinga/Bahia, e uma outra estação como Conde, no mesmo Estado, com precipitação anual de 1392,2 mm e um desvio padrão de 414,7 mm, constatamos que a variação absoluta, ou seja, os desvios padrões são quase idênticos, mas as médias bem diferentes. Assim, é parcial dizer que a dispersão é mais ou menos a mesma. Essa disseminação pluviométrica tem consequências graves numa região com pouca precipitação anual, onde a média é baixa, mas pode ter pequena ou nenhuma consequência para a agricultura em região chuvosa. Calculando o coeficiente de variação ele indica para Bonito o valor relativo de 61,1% e para Conde 29,8%, ou seja, a instabilidade relativa de Bonito é muito maior.

Uma desvantagem deste coeficiente é que ele não é utilizável se $\bar{X}$ está próximo de zero. Este fato ocorre relativamente pouco nos dados geográficos, mas, para citar novamente exemplo de precipitação, apontamos a estação de Correntina, também no Estado da Bahia, onde durante vinte e cinco lustros aconteceu só em um determinado ano, chuva no mês de junho. Assim, a média dos vinte e cinco anos do mês de junho é 0,1 mm e o desvio padrão 0,5. Considerando que o coeficiente de variação daria por causa das razões mencionadas uma grande distorção, ele é neste exemplo, desaconselhável.

Uma outra medida de variabilidade relativa, parecida ao coeficiente de variação é achada dividindo-se o desvio médio pela média (tomada em termos absolutos).

A terceira medida indicando a variabilidade relativa é dividir o desvio quartílico pela mediana. Mas, de todas estas três medidas, o coeficiente de variação é o mais indicado. Um exemplo da aplicação do coeficiente de variação é dado no Atlas Climatológico da Bahia (1976), onde foram calcu-

lados e mapeados os coeficientes de variação de precipitação para as estações do Estado, segundo os meses e o total do ano. A interpretação, em relação aos meses, deve ser feita com cuidado. Em alguns casos, os dados das estações onde não tem precipitação, ou muito pouca em determinados meses, podem levar, segundo nossos argumentos anteriores, a uma grande distorção.

É claro que o coeficiente de variação pode ser aplicado também para problemas de outras áreas da Geografia. Fuchs (1960), por exemplo, utilizou-o com sucesso para determinar a variação intraurbana da qualidade residencial. Ele mediu essa qualidade segundo os custos e fez a diferenciação entre cidades centrais de áreas metropolitanas e cidades não centrais. O resultado da pesquisa é que as comunidades não centrais mostram, na maioria das vezes, variabilidade menor, ou seja, são mais uniformes na qualidade residencial do que as cidades centrais.

7. A CURVA NORMAL DA DISTRIBUIÇÃO DE FREQUÊNCIA

7.1. Características da curva normal

Analisamos, até agora, que existem muitos métodos para caracterizar, de forma eficiente, um conjunto de dados. Os métodos que têm mais fundamentos matemáticos são a média aritmética e o desvio padrão. Se dada distribuição de frequência for normal, neste caso a curva de distribuição terá a forma de sino e é simétrica em torno de um ponto central. Esta chamada distribuição normal é, também, denominada *distribuição de Gauss*. É uma distribuição contínua e a suposição básica para a aplicação de muitos métodos estatísticos. Considerando que vários tipos de conjuntos de dados da Geografia têm esta distribuição normal, ou quase normal, ela pode assim servir como importante modelo teórico.

Se temos diversos conjuntos de dados com o mesmo desvio padrão, mas médias diferentes, outrossim, a forma da curva é, em todos os casos, a mesma. Por outro lado, se os diversos conjuntos de dados tem a mesma média, mas desvios padrões diferentes, neste caso resultam curvas com formas diferentes. No geral, podemos dizer que maior o desvio padrão, mais alta e estreita será esta. (fig. 8).

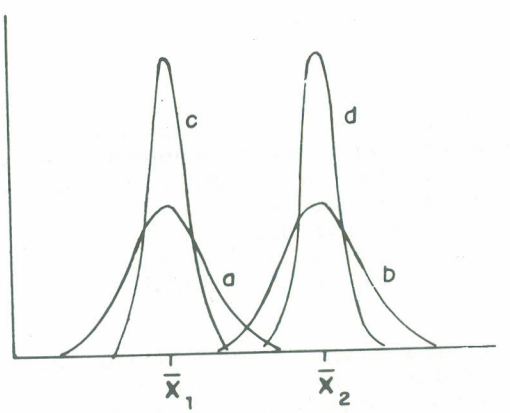


Figura 8

Dentro da área limitada pela curva e o eixo *x* temos a indicação de todas as ocorrências. Se o desvio padrão é, como vimos, responsável pela forma da curva, logicamente ele resume também o número de ocorrências. A frequência esperada de observações é representada através da área abaixo da curva, ou seja, a frequência esperada entre duas classes é representada através da área entre estes limites abaixo da curva. A área total abaixo da curva é igual a soma das frequências esperadas *1,0* ou *n*, dependendo se as frequências foram calculadas em termos relativos ou absolutos.

A curva normal tem as seguintes características: (fig. 9)

68,26% das ocorrências encontram-se entre $(\bar{X} - 1s)$ e $(\bar{X} + 1s)$

95,44% das ocorrências encontram-se entre $(\bar{X} - 2s)$ e $(\bar{X} + 2s)$

99,74% das ocorrências encontram-se entre $(\bar{X} - 3s)$ e $(\bar{X} + 3s)$

99,99% das ocorrências encontram-se entre $(\bar{X} - 4s)$ e $(\bar{X} + 4s)$

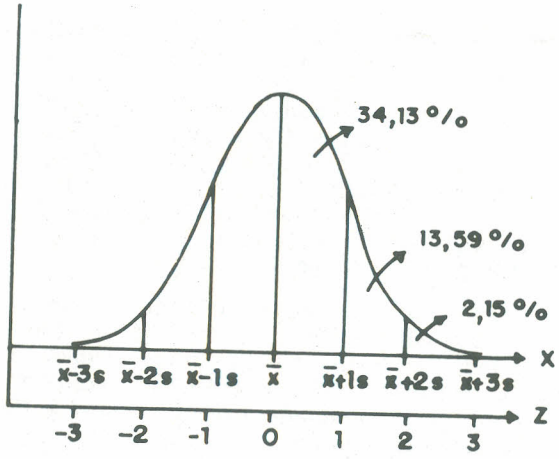


Figura 9

Isto quer dizer que um desvio maior de um desvio padrão da média ocorre mais ou menos uma vez sobre três, ou sejam em 31,74% das ocorrências. Um desvio de mais de dois desvios padrões ocorre mais ou menos uma vez sobre vinte e duas ocorrências, ou seja em 4,56% dos casos e um desvio de mais de três desvios padrões ocorre só mais ou menos em 384 ocorrências, ou seja em 0,26% dos casos. Lembramos que a curva normal é assintótica ao eixo *x* e, por consequência, o valor de 100% teoricamente nunca é atingido.

Para a distribuição normal há número infinito de tipos de curvas e não só um tipo de distribuição normal, porque a média e o desvio padrão podem assumir qualquer valor. Mas estas diversas curvas normais podem ser reduzidas

através de uma transformação para uma forma padronizada. A transformação indica para cada valor de *X* um valor $z = \frac{X - \bar{X}}{s}$ onde $\bar{X}$ é a média aritmética e *s* o desvio padrão, *Z* tem a distribuição normal, a média zero e o desvio padrão um. A tabulação da distribuição normal padronizada é mostrada em qualquer livro de estatística como, por exemplo, Spiegel (1971, p. 562). Considerando que a curva é simétrica, a tabela indica a proporção da área entre a média e um ponto acima de zero em termos de desvios padrões. Por exemplo, a área entre a média e x = 1 é 0,3413, ou seja 34,13% da área total da curva está entre estes limites, ou seja, 34,13% das ocorrências de um conjunto de dados que tem uma distribuição normal estão dentro destes limites (fig. 9).

7.2. Aplicação da distribuição normal

A distribuição normal é importante para a Geografia porque, à parte de outras possibilidades, ela pode servir para fazer predições valiosas. Sabemos que muitos conjuntos de dados, particularmente da Geografia Física, são distribuidos de forma mais ou menos normal.

Tomando o nosso pequeno exemplo da precipitação anual em Salvador, durante vinte anos, tínhamos calculado $\bar{X}$ = 2.075,97 mm e o desvio padrão 542,02 mm. Assim:

$\bar{X}$ +	1 desvio padrão	=	2.075,97 +	542,02	=	2.617,99mm
$\bar{X}$ -	1 desvio padrão	=	2.075,97 -	542,02	=	1.533,95mm
$\bar{X}$ +	2 desvios padrões	=	2.075,97 +	1.084,04	=	3.160,01mm
$\bar{X}$ -	2 desvios padrões	=	2.075,97 -	1.084,04	=	991,93mm
$\bar{X}$ +	3 desvios padrões	=	2.075,97 +	1.626,06	=	3.702,03mm
$\bar{X}$ -	3 desvios padrões	=	2.075,97 -	1.626,06	=	449,91mm

Baseando-nos sobre este exemplo podemos agora perguntar: quantos anos devemos esperar precipitação entre $\bar{X} \pm 1$ desvio padrão, ou seja, entre 1.533,99mm e 2.617,99mm? Segundo a tabela vimos que entre $\bar{X} \pm 1$ desvio padrão devem cair teoricamente 68,26% das observações. No caso presente esta percentagem corresponderia à 13,66 anos. Na realidade quatorze anos dentre vinte apresentam estes índices de precipitação. Entre $\bar{X} \pm 2$ desvios padrões, ou seja entre 449,91 mm e 3.702,03mm, deveriam cair 19,94 anos, sendo na realidade vinte anos. O nosso exemplo mostra que a predição está bem próxima da realidade. Considerando que tínhamos só um exemplo de vinte anos, não podemos esperar nenhum ano com precipitação acima ou abaixo de três desvios padrões. A realidade confirma a previsão.

Por outro lado, suponhamos que num caso semelhante existam dados indicando precipitação em vários anos acima de três desvios padrões, que segundo a probabilidade não deveriam ocorrer com tal freqüência. Neste caso, devemos duvidar dos levantamentos efetuados, sendo provável a presença de erros nos dados levantados.

Podemos ainda formular muitos outros tipos de perguntas, como por exemplo: qual a probabilidade que em Salvador caiam menos de 1.533,95mm, ou mais de 2.617,99mm de chuva, por ano? Sabemos segundo a tabela que entre a média e menos um desvio padrão, que corresponde em nosso caso à 1.533,95mm, deveriam cair 34,13% das observações. A metade da curva envolve 50% das ocorrências e assim achamos o valor de 15,87% (50,00% - 34,13%). Isto é, 15,87% dos anos deveriam ter precipitação abaixo de 1.533,95 mm ou seja 3 anos dentro dos 20, o que é exatamente o caso na realidade. Sendo a curva simétrica, mais de 2.617,99 mm de chuva (= 1 desvio padrão) deveriam ser esperados também em 15,87% dos anos.

Uma outra pergunta caberia: em quantos porcentos de anos podemos esperar que existam menos de 2.000 mm de chuva? O cálculo é feito segundo a fórmula $z = \frac{X - \bar{X}}{s}$. Nesta fórmula, $\bar{X}$ é a média aritmética, no nosso exemplo 2.076,0 mm, s é o desvio padrão de 542,0 mm. X é no caso 2.000 mm. O resultado é que z = -0,14, que corresponde à 5,57% na tabela (Spiegel, 1971, p.562). Lembramos que este valor indica a probabilidade entre a média e o valor de z. Assim, a probabilidade de ter um valor à esquerda de z = - 0,14 é de 44,43% (50,00% - 5,57%). Isto quer dizer que podemos esperar em Salvador, em 44,43% dos anos, precipitações abaixo de 2.000 mm.

Este último tipo de pergunta é muito importante para pesquisas agrícolas, particularmente em regiões com problemas de secas. Eis a pergunta: em quantos anos podemos esperar pelo menos uma determinada quantidade de chuva é respondida segundo o mesmo esquema. Por outro lado, a resposta da pergunta sobre qual seria a probabilidade de que um certo valor venha a ser ultrapassado, efetua-se da mesma maneira. Segundo tal pergunta, podemos calcular as probabilidades para cada estação existente de um país, estado ou região, e mapear os resultados através de isolinhas do mesmo valor de probabilidade.

7.3. Testes gráficos

Existe um método que permite ver rapidamente como é uma distribuição observada em relação à normalidade. Devemos sair da distribuição de freqüência acumulada, como mostra a tabela 4, onde foram calculadas as frequências acumuladas em porcentagem. Acumulamos os percentuais de frequência do menor para o maior valor. Sobre a abscissa do gráfico colocamos os limites superiores de cada classe e sobre a ordenada as porcentagens da frequência acumulada correspondentes. A diferença para os gráficos conhecidos até agora é que a ordenada tem uma escala probabilística, ou seja, o papel para fazer o gráfico é o chamado papel de probabilidade, que se encontra já pronto para ser utilizado.

Chamamos a atenção que a escala de probabilidade não tem o valor de 0% nem de 100%. Se a distribuição fosse normal, o resultado sobre este papel de probabilidade seria uma linha reta, isto é, os pontos se colocariam sobre uma linha a ser desenhada. Numa distribuição desviada para a direita as frequências superiores são desviadas para a direita da linha reta. Na distribuição desviada à esquerda as frequências superiores estão à esquerda da linha reta.

Através do gráfico podemos também estimar a média e o desvio padrão. Sabemos que numa distribuição normal a média e a mediana coincidem. A mediana encontra-se desenhando uma linha perpendicular para a abscissa da intersecção do valor de 50% da ordenada com a curva de frequência acumulada.

A estimativa do desvio padrão se consegue de maneira semelhante, colocando perpendiculares para a abscissa da intersecções de 15,87% e de 84,13% com a curva acumulada, porque 68,26% dos itens se encontram entre X + 1 e X - 1 desvio padrão. Medindo a diferença entre as duas perpendiculares e dividindo-a por 2, temos uma estimativa do desvio padrão.

7.4. Transformação de dados para a curva normal

Se constatarmos que a distribuição em questão não é normal, temos meios de tentar transformá-la matematicamente em normal, ou quase, para poder aplicar os métodos adequados para distribuições normais. Como mencionamos já várias vezes, muitas distribuições geográficas são desviadas para a direita. Quando transformamos todos os dados originais em logarítmos na base 10, o resultado é na maioria das vezes distribuição que se aproxima à normal. Neste caso fala-se que a distribuição é log-normal.

Um mesmo teste gráfico pode confirmar rapidamente se a transformação em logarítmos resulta numa distribuição normal. Colocamos os dados originais dos limites superiores de cada classe sobre a abscissa que tem, ao contrário do exemplo anterior, a escala em logarítmos. Assim, economizamos o tempo de transformar os dados originais em logarítmos. Como no exemplo anterior, a ordenada tem a escala probabilística. Este tipo de papel pode ser também obtido pronto para ser utilizado.

Ficamos conscientes que para a transformação logarítmica todas as observações devem ser maiores de zero. Se não é o caso, uma constante deve ser adicionada a todos os valores.

Entretanto, a transformação logarítmica não é a única que pode ser aplicada para distribuições com assimetria positiva. Muitas vezes a transformação em raízes quadradas dá também bons resultados. Para a assimetria negativa, a transformação para a curva normal é, em numerosas ocasiões, satisfatória calculando os quadrados dos dados originais. Assim, é preciso ressaltar a necessidade de examinar bem os dados originais antes de começar a pesquisa, para poder trabalhar com técnicas estatísticas que têm como base a curva normal. Para decidir a transformação a ser aplicada perde-se, às vezes, tempo. Em casos difíceis a melhor maneira é a de testar várias transformações possíveis e de escolher a que mais se aproxima de uma curva normal.

Berry e Horton (1970, p. 67-69) dão um exemplo da aplicação da log-normalidade, testando as distribuições das cidades segundo os tamanhos em trinta e oito países diferentes do mundo. A pesquisa conclui que treze dos trinta e oito países têm uma distribuição log-normal, indicada através do método gráfico mencionado em forma de uma linha reta. São países grandes, como a China, ou muito pequenos, como a Suiça, desenvolvidos como os Estados Unidos ou subdesenvolvidos como a Coréia. Países com longas tradições urbanas e altamente desenvolvidos tem a distribuição de

cidades tamanhos muito similares. O resultado desta pesquisa dá uma proposição de um modelo gráfico das distribuições das localidades segundo os tamanhos, variando entre os casos de primazia e log-normalidade (Berry e Horton, 1970, p. 73).

Este modelo de distribuição cidade/tamanho, de Berry, foi aplicado para as comunidades acima de 5.000 habitantes do Estado da Bahia (Xavier e Silva, 1973). Mostrou-se que a Bahia evolui de uma situação de primazia (em 1940, quase 80% da população urbana estadual morava em cidades de até 7.000 habitantes) a uma situação em que, em 1970, a tendência a log-normalidade é flagrante (Xavier e Silva, 1973, p. 111-112).

Concluindo, é preciso destacar novamente as potencialidades dos métodos quantitativos aqui apresentados, objetivando uma eficiente análise dos dados geográficos, e o caráter preliminar dos mesmos no amplo contexto metodológico atualmente disponível. Com efeito, devidamente colocados no conjunto dos procedimentos de uma pesquisa científica, o conhecimento destes métodos é básico para numerosas interpretações, e para que se possa almejar a aplicação de métodos mais avançados.

É nossa intenção apresentar, em uma etapa posterior, os métodos quantitativos avançados, aplicados à Geografia, nos moldes desta publicação.

BIBLIOGRAFIA

Bahia. SERPLANTEC. Centro de Planejamento da Bahia, CEPLAB (1976). *Atlas climatológico do Estado da Bahia; análise espacial da pluviosidade.* Doc. nº 2. Salvador.

Bahia. SERPLANTEC. Centro de Planejamento da Bahia, CEPLAB (1977). *Atlas climatológico do Estado da Bahia.* Doc. nº 4. Salvador.

Bahrenberg, G. e Giese, E. (1975). *Statistische Methoden und ihre Anwendung in der Geographie.* Stuttgart: Teubner.

Berry, B. J. L e Horton, F. E. (1970). *Geographic perspectives on urban systems.* Englewood Cliffs, N. J.: Prentice-Hall.

Cole, J. P. e King, C. A. M. (1968). *Quantitative geography.* London: John Willey and Sons.

Daugherty, R. (1984). *Science in geography: Data collection.* London: Oxford University Press.

ETENE/BNB (1969). O consumo de produtos industriais na cidade de Salvador. *Revista Econômica,* Ano 1, nº 2, p. 42-52.

Fuchs, R. J. (1960). Intraurban variation of residencial quality. *Economic Geography.* Vol. 36, p. 313-325.

Greer-Wootten, B. (1972). *A bibliography of statistical applications in Geography.* Comission on College Geography. Association of American Geographers. Technical Paper 9. Washington (D.C.).

Gregory, S. (1968). *Statistical methods and the geographer.* 2ª ed. London: Longman.

Harvey, D. (1969). *Explanation in Geography.* New York: St. Martin's.

Henshall, J. D. e King, L. J. (1966). Some structural characteristics of peasant agriculture in Barbados. *Economic Geography.* Vol. 42, p. 74-84.

King, L. J. (1969). *Statistical analysis in geography.* Englewood Cliffs, N. J.: Prentice-Hall.

Nentwing Silva, B. C., Galbraith, J. H. e Silva, S. C. Bandeira de Melo (1974). Técnica estatística para agrupamento e mapeamento de informações geográficas. *Bol. de Geografia Teorética.* Vol 4, nºs 7 e 8, p. 29-42.

Sokal, R. R. e Rohlf, F. J. (1969). *Biometry: The principles and practice of statistics in biological research.* San Francisco: Freeman.

Spiegel, M. R. (1971). *Estatística).* Rio de Janeiro: McGraw-Hill.

Toyne, P. e Newby, P. T. (1971). *Techniques in human geography.* Basingstoke and London: Macmillan.

Xavier, E. A. e Silva, S. C. Bandeira de Melo (1973). Considerações em torno de uma política de desenvolvimento regional para o Estado da Bahia. In: *Projeto de regionalização administrativa para o Estado da Bahia,* p. 107-115. Salvador: SEPLANTEC.

Yates, M. (1974). *An introduction to quantitative analysis in Human Geography.* New York: McGraw-Hill.

RESUMO

Métodos quantitativos aplicados em Geografia: uma introdução

O objetivo deste trabalho é o de analisar os métodos matemático-estatísticos mais freqüentemente utilizados na Geografia. Colocados na ampla problemática da abordagem científica, são discutidas a importância, as vantagens e as desvantagens da aplicação dos métodos quantitativos em Geografia. São analisados, posteriormente, os diferentes níveis de mensuração, os conceitos básicos, as técnicas de agrupamento, as medidas de tendência central, as medidas de dispersão e a curva normal de distribuição de freqüência. As potencialidades dos referidos métodos são ressaltadas objetivando uma eficiente análise dos dados geográficos. Em um segundo trabalho serão analisados os métodos quantitativos avançados aplicados à Geografia.

ABSTRACT

Quantitative methods applied in Geography: an introduction

The work aims to analyse the mathematical and statistical methods most frequently used in Geography. Placed in the wide perspective of the scientific analyse, this paper presents the importance, advantages and disadvantages of the application of such methods in Geography. The following aspects are then analysed: levels of measurement, basic concepts, grouping techniques, measures of central tendency, measures os dispersion and the normal distribution of frequency. The potential of the quantitative methods are showed in conclusion and in a second paper the more advanced methods will be analysed.