RAYMUNDO COSTA E SOUZA

SÍNTESE E ANÁLISE NA ESTATÍSTICA

BAHIA - 1959 - BRASIL

S Í N T E S E E A N Á L I S E N A E S T A T Í S T I C A

Tese apresentada à Faculdade de Ciências Econômicas da Universidade da Bahia para inscrição em concurso de docência livre da cadeira de Estatística Metodológica.

À MINHA MÃI E

EM MEMÓRIA DE

MEU PAI

PREFÁCIO

A deliberação da douta Congregação da Faculdade de Ciências Econômicas da Universidade da Bahia, confiando-nos no ano de 1954 a regência da cadeira de Estatística Metodológica do curso superior de Ciências Econômicas, cuja regência vimos exercendo ininterruptamente, impôs como conseqüência natural do desempenho da missão didática e da resultante sistematização do nosso conhecimento, a elaboração desta tese para os misteres de habilitação ao concurso de livre-docência da referida disciplina.

O presente trabalho, embora não pretenda dar uma visão completa do assunto, foi efetuado em correspondência à aspiração do autor de estabelecer no âmbito da estatística, uma conciliação entre duas diretrizes metodológicas ortodoxamente opostas em suas finalidades - a síntese e a análise. Entregamo-nos, por isso, voluntária e pacientemente, à seleção dos elementos necessários à execução da exaustiva porém agradável tarefa de expor os resultados da pesquisa empreendida no sentido de verificar a compatibilidade, complementariedade e concomitância dos procedimentos sintéticos e analíticos, manifestadas desde os primeiros vestígios de maturação do método estatístico.

Nas três primeiras partes do texto, encontra-se a discussão em têrmos gerais da natureza e aplicabilidade daquelas duas metodologias, quando então o leitor após perceber a função dasíntese, na apreciação de índole informativa, do único fator quantitativo existente na amostra, tomará contato com a análise nas considerações de teor explicativo em tôrno de dois ou mais fatôres quantitativos no campo amos

tral, assimilando já nesta segunda fase a ação complementar, de caráter simbiótico do sistema binário síntese-análise.

Ressalta em tôda esta composição, o ângulo sob o qual divisamos a maturação do método estatístico, tanto o começo do processo evolutivo com a passagem da simples descrição de fenômenos homogêneos à interpretação de fenômenos heterogêneos, quanto a intensidade do devir neste último setor em razão da complexidade amostral, ressalvando-se entretanto a relatividade da evolução aí esboçada, pois alguns outros aspectos comparativos da estatística proporcionam igualmente estruturações conceituais transformistas.

Apesar de termos nos reportado no escrito, às conexões da estatística com a indução, não é nosso objetivo examinar especìficamente a questão concernente às bases que o método estatístico estabelece, para a generalização ao universo dos valores calculados na amostra, mas sim estudar a maturação acima mencionada no campo da própria amostra. Procuramos assim demonstrar as mutações do "modus faciendi" metodológico, desde que a síntese deixando de ser operada isoladamente na fase descritiva, torna-se complemento eficiente da análise no estágio superior interpretativo.

Na descrição, limitamos a apreciação sintética aos problemas da contração e dispersão, utilizando as técnicas do promédio e do desvio padrão; na parte da interpretação, delineamos o ciclo evolutivo do método, adotando um critério fundado nas teorias da regressão e correlação linear, com o intento de mostrar a complementação sintético-analítica e a diferenciação da análise paralelamente à repetição da síntese, deixando perceber a evolução da metodologia e finalmente enunciando uma lei reguladora das mudanças operadas em sua contextura.

Precedendo a parte final conclusiva, dois capítulos foram acrescentados para propiciar a identificação dos esquemas ou modêlos abstratos aos planos concretos.

À Superintendência do Instituto de Economia e Finanças da Bahia, aos membros do seu "staff" e ao récem-falecido amigo e Secre

tário Geral dessa instituição, Dr. Daniel Quintino da Cunha, o autor sente-se reconhecido pelas facilidades que lhe foram oferecidas quanto ao livre acesso à bibliotéca para as necessárias consultas.

Somos igualmente gratos a todos os mestres, amigos e colegas que acataram a idéia de elaboração dêste trabalho, para cuja consecução concorreram decisivamente, dando-nos apoio através de constante estímulo e expressiva motivação.

Salvador, Março de 1959

RAYMUNDO COSTA E SOUZA

PRIMEIRA PARTE

CONDIÇÕES DE PERQUIRIÇÃO ESTATÍSTICA

- CAPÍTULO I -

ASPECTOS UNIDIMENSIONAIS E PLURIDIMENSIONAIS

A progressiva expansão do conhecimento científico depende da extensão do aperfeiçoamento das conceituações, à medida que os fundamentos das várias ciências são meticulosamente estabelecidos mediante critérios bem definidos e métodos cada vez mais apurados.

A substituição de modalidades qualitativas, quer das propriedades abstratas induzidas de nossas sensações (1) ou dos sistemas de valores especiais (2), por elementos quantitativos, faz emergir a noção de grandeza em proveito de maior precisão nos resultados provenientes das pesquisas.

Efetivamente, os números estão hoje presentes em quase todos os setores do conhecimento (3). Direta ou indiretamente são usados como instrumentos indispensáveis aos trabalhos científicos, mesmo entre aquêles de caráter social e econômico. A ninguem passa desperce-

---

(1) A observação do mundo exterior dá lugar à assimilação das propriedades abstratas dos objetos, isto é, aquelas que induzimos das nossas próprias sensações - comprimento, temperatura, massa, côr, luz, intervalo de tempo etc. -, cujos caracteres em se manifestando aos sentidos, traduzem os fenomenos físicos e oferecem os elementos que possibilitam a formação dos respectivos conceitos. (ALLEN, R.G.D, Análisis Matemático para Economistas. Madrid, M. Aguilar, 1946, pág. 11).

(2) Modalidades científicas há que não podem ser absorvidas em percepções sensoriais, mas subordinam-se a juízos de valores como sói acontecer na apreciação de atos coletivos no setor das ciências sociais, onde estes fatos humanos por sua natureza peculiar, representam valores. (GRANGER, GILLES GASTON. Lógica e Filosofia das Ciências, São Paulo, Edições Melhoramentos, 1955, pág, 140).

(3) YULE, G. UDNY e KENDALL, M.G. Introdução à Teoria da Estatística. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística, 1948, págs. 19-20.

bido que um atributo qualitativo se torna mais significativo quando es tá associado a uma expressão quantitativa. Os números encerram sempre precisão, e visto que as ciências em sua generalidade procuram apresentar conclusões acuradas, não se pode estranhar que elas os empreguem freqüentemente.

As características exclusivas a cada um dos grupos comumente diferenciados em ciências exatas e ciências sociais, não condicionam por si só o aparecimento de uma grandeza, mas sim a possibilidade de revestir forma quantitativa os aspectos de um fenômeno físico, meteorológico, sociológico, demográfico, econômico etc.

As crescentes exigências de acuidade na pesquisa e rigidez dignidade na apresentação do material quantitativo em diversas ciências, sucedendo ao emprêgo cada vez mais intensivo de processos e técnicas numéricas, fizeram imperativo o aparecimento de um método específico apto a permitir o estudo de certos aspectos dos agregados de números. Isto porque se faz mister exprimir tão precisamente quanto possível, algumas características e também prováveis relações, que permaneceram ocultas embora houvessem os agregados sido submetidos aos ordinários meios de perquirição científica.

Por conseguinte, tentando superar as restrições inerentes a alguns dos métodos científicos no tocante a um grau de suficiente precisão que se procura atribuir aos resultados da pesquisa, foram introduzidos no âmbito das ciências em geral, procedimentos especiais de teor matemático fundados em modos sintéticos e meios analíticos, compreensivos da importante e útil metodologia estatística.

A superposição das magnitudes de uma grandeza, abrangendo os números reais limitados ao campo de variabilidade do fenômeno, suscita o exame da sua variação na amplitude do agregado, a fim de que êste venha a exibir, por aplicação do método estatístico, caracteres não evidenciados devidamente sob o crivo dos métodos ortodoxos. Considerando-se uma variável X que toma os valores X_i (i = 1,2,3,....n) distribuídos por todo o intervalo de variação da grandeza, repetidos al-

guns dêles conforme a intensidade com que ocorrem durante a observação, o cientista dispõe de uma massa de dados quantitativos, que não obstante originalmente difusa, revelará após metódica elaboração, traços peculiares à evolução do fenômeno.

As séries estatísticas têm origem em tais superposições de magnitudes, exprimindo normalmente a dinâmica do próprio fenômeno, o qual sob forma quantitativa se apresenta mais sensível à inquirição de suas variações características, contanto que a observação seja efetuada numa amostra representativa de tôdas as ocorrências semelhantes possìvelmente existentes no universo.

Em qualquer pesquisa é necessário, atendendo aos requisitos de menor custo e economia de tempo, destacar para o estudo específico que se pretende realizar, uma parcela do conjunto que envolve todos os pormenores dos padrões de variabilidade da grandeza, ou grandezas se houver mais de uma, cujo conjunto é denominado <u>universo</u> ou <u>população</u> irrestritamente aos campos físico, biológico, social, econômico, demográfico etc., não importando a natureza do fenômeno. O emprêgo do vocábulo é extensivo aos aspectos numéricos finitos e infinitos das ciências em geral, porquanto em sua maioria, os fenômenos são suscetíveis de decomposição em variáveis e por isso dão lugar à apreciação das componentes quantitativas.

Sejam contínuas ou discretas, consoante os valores consecutivos das séries difiram por quantidades finitas ou infinitamente pequenas, ora as variáveis aparecem isoladas ora conjugadas, resultando dessa alternativa comum a tôda manifestação da vida física ou da ambientação social, a existência de universos ou populações monovariáveis e universos ou populações bivariáveis e multivariáveis.

A homogeneidade ou heterogeneidade dos universos físicos, biológicos, sociológicos etc., conforme haja influência única de uma grandeza na formação do fenômeno ou coexistam duas ou mais grandezas, dita critérios distintos embora correlatos a serem adotados na verificação do regime das variações, em face de circunstâncias simples-

mente monovariáveis e de outras mais complexas de natureza diferencia da. Obviamente a sistemática não pode ser a mesma em cada uma destas hipóteses, diferindo os processos e técnicas usados na elucidação dos caracteres de uma problemática fenomênica, de acôrdo com o grau de complexidade que esta encerra.

O caráter unitário dos universos monovariáveis, dado à homogeneidade do fenômeno, alvitra a adoção de um critério unidimensional para proceder fácil e seguramente por meio da síntese, à perquirição mais rápida das características de uma determinada variável. A natureza dos universos bivariáveis e multivariáveis, por sua vez, consubstanciando a heterogeneidade de um fenômeno, demanda a escolha de um critério pluridimensional, o qual requer orientação analítica em oposição ao espírito de síntese das investigações realizadas sob as condições anteriores, a fim de que se possam estabelecer com maior clareza possível nesta segunda etapa, as relações provàvelmente existentes entre duas ou mais variáveis.

Há pois idéia de representação sumária, de condensação, enfim de síntese, tôda vez que aplicamos o critério unidimensional ao conjunto monovariável; calculam-se os chamados valores sinóticos, reduzindo-se a totalidade dos dados a uma única constante, a qual devidamente elaborada, resume as características da série em X_i ($i = 1,2,3,\ldots,n$). Portanto, é inteiramente descritiva esta fase, onde os valores únicos têm a função de representar sintèticamente uma situação global na amplitude da série estudada.

O critério pluridimensional, todavia, ultrapassa os limites da operação descritiva regida pelos processos baseados na monovariação, pois se propõe através da dissociação, em harmonia com o processo analítico, a estabelecer relações de dependência entre duas ou mais variáveis, bem como avaliar o sentido da variação e a intensidade da relação entre elas, emergindo daí os problemas de regressão e correlação estatísticas.

Em virtude da complexidade do universo e igualmente da amostra, esta fase de elaboração pluridimensional que requer operações mais refinadas, é essencialmente interpretativa, porquanto envolvendo a necessidade de comparação entre duas ou mais componentes quantitativas de um fenômeno heterogêneo, numa amostra integrante de universo idêntico, comprova empìricamente os fundamentos de hipóteses formuladas à base de precedentes observações ou mediante premissas admitidas verdadeiras. Destarte, o assentamento de relações prováveis entre duas ou mais variáveis, fornece extraordinário subsídio ao método genérico da indução, inclusive tornando as explicações desta metodologia mais compatíveis com a realidade científica moderna.

A análise de um universo bivariável ou multivariável através da respectiva amostra e a subseqüente formulação de uma lei fundamentada nas relações quantitativas dependenciais, é tão importante para o desenvolvimento científico em vista da sua íntima analogia com a indução, que W.A. Shewart (4) discutindo a acentuada semelhança entre a estatística e o método indutivo, expõe que na investigação da existência de fatos, leis e causas a partir de dados empíricos, os cientistas têm interêsse de reunir, apresentar e interpretar elementos qualitativos e quantitativos, procedimento êste comum aos campos do método científico, da lógica e da estatística. (5)

---

(4) SHEWART, W.A. Annual Survey of Statistical Technique: Development in Sampling Theory. In Econometrica, Vol 1, n. 3, julho 1933, pag. 226.

(5) Esta semelhança é também posta em destaque por Sixto Rios, quando observa ser a indução de uma lei em tôrno de certo fenômeno experimental, uma maneira de comprová-la mediante novas experiências, atribuindo-se então uma medida da confiança que nos merecem tais experimentos como comprovação da própria lei. RIOS, SIXTO. Introduccion a los Metodos de la Estadistica. Madrid, Nuevas Gráficas S.A., 1952, pag. 4.

- CAPÍTULO II -

SÍNTESE E ANÁLISE

O critério pluridimensional, orientando a interpretação através da análise aplicada às modalidades quantitativas do fenômeno, não exclui nem dispensa o conurso da síntese, em visando o obtenção de resultados científicos. Na fase descritiva, a síntese é operação única, concordante com o critério unidimensional, não se requerendo então o complemento : de nenhum outro processo no ato de apreciação dos caracte res de um agregado numérico constitutivo da amostra monovariável. Na fase interpretativa, porém, apesar de ser a análise o principal modo de elucidação da especificidade de dois ou mais agregados numéricos compreensivos da amostra, ela coexiste com a síntese em tôda a extensão da pesquisa.

Verdadeiramente, enquanto se processam operações estatísticas sob o critério pluridimensional, ocorre a constante necessidade de calcular-se valores sinóticos sôbre as diferentes séries da amostra, cujos cálculos tanto mais exigidos serão quanto maior for o número de variáveis, subordinado no entanto todo êste modus faciendi sintético à marcha da análise, a qual circunscreve e domina completamente o campo da investigação. Crescendo a complexidade amostral em razão do maior número de variáveis incluídas no âmbito da observação, as operações sinóticas tendem a repetir-se nas suas formas primitivas, ao passo que os meios analíticos se diferenciam em combinações cada vez mais apuradas.

Poder-se-ia, portanto, induzir uma lei de evolução da metodologia estatística, enunciando que, ultrapassados os objetivos da descrição e atingida a fase interpretativa, as aplicações neste último setor demonstram a manutenção das modalidades sintéticas concomitante-

mente à sensível e sucessiva diferenciação do grau de análise.

Esta asserção não significa que deixemos de considerar a importância relativa das operações simplesmente sintéticas, de teor aproximadamente uniforme, tôda vez que a homogeneidade do fenômeno configurado em amostra monovariável exigir estritamente a sua aplicação Entretanto, o método estatístico requer como condição necessária de maturação, a tomada de contato com a incessante diversificação dos procedimentos analíticos suscitada pela heterogeneidade do fenômeno, tal qual êste se apresenta nas amostras bivariáveis e multivariáveis. Esta circunstância, conquanto mantenha a síntese operação básica e integrante das novas elaborações, introduz os fecundos recursos da análise em proveito da precisão que se intenta ordinàriamente atribuir aos resultados de uma pesquisa em qualquer seara científica onde seja verificada a influência de dois ou mais fatôres.

Conseqüentemente, concluimos que a síntese pode ser praticada isoladamente - no curso da descrição - ou conexa porém sempre subordinada à análise - no caso de interpretação -. Se houver descrição, a síntese, compreendendo o cálculo de valores sinóticos, constitui uma operação autônoma, mas na interpretação assume caráter acessório, participando como elemento estrutural da análise para determinação das relações estocásticas e dos coeficientes especiais.

Enfim, uma reflexão sôbre o contôrno evolutivo do método estatístico envolve preliminarmente a consideração de um critério unidimensional, aspecto conceitual basilar que rege a teorização e aplicação da metodologia sintética, tendo esta por objeto as amostras monovariáveis e por finalidade a descrição de alguns caracteres quantitativos dessas parcelas amostrais; utiliza-se aí um processo condensativo consubstanciado na organização de dados e na técnica calculatória dos valores sinóticos, em busca de resultados apenas informativos. A redução a que se submete uma série estatística, por motivo da aplicação sintética consoante a natureza do próprio critério unidimensional, imprime ao método um sentido descritivo, de caráter mais limitativo do que extensivo no conjunto geral do estudo estatístico. Em seguida, surge o

critério pluridimensional, aspecto conceitual transcendente ao caráter estritamente primário do anterior, por já propiciar a penetração nas fases secundária e terciária de uma pesquisa, conduzindo à adoção de uma metodologia analítica onde o objeto são as amostras de duas ou mais variáveis e a finalidade é a interpretação dos efeitos de fenômenos complexos no campo amostral. Agora é o processo dissociativo coordenado por uma técnica calculatória de relações estocásticas e coeficientes especiais, que produz resultados essencialmente explicativos com respeito às séries estatísticas conjugadas, dando lugar ao sentido interpretativo do método, por excelência extensivo no âmbito da estatística.

- C A P Í T U L O III -

ESSÊNCIA DA ESTATÍSTICA

Antes de discutirmos os pormenores da verificação do regime das variações de uma ou de várias grandezas, sempre com o fito de esboçar a evolução natural do método estatístico através das suas fases sucessivas - síntese e análise -, julgamos conveniente discernir os fundamentos desta metodologia por meio de alguns conceitos essenciais a uma razoável percepção do seu alcance.

Dentre as definições encontradas na literatura estatística nacional, destacamos a fim de evidenciar os propósitos dos que executam investigações em bases estatísticas, os enunciados dos professôres L. Nogueira de Paula e Milton da Silva Rodrigues.

O Prof. Nogueira de Paula diz ser a estatística " o processo metodológico que consiste em deduzir das massas de observações valores sinalético e leis mais ou menos prováveis". (6)

O Prof. Milton da Silva Rodrigues explica que a "estatística é o método que tem por objeto o estudo dos agregados e por fim a determinação das suas tendências características limites". (7)

Quanto às conceituações de cultores da estatística no estrangeiro, Corrado Gini considera a estatística uma "técnica apropri

---

(6) NOGUEIRA DE PAULA, LUIZ. Metodologia da Economia Política. Rio de Janeiro, Irmãos Pongetti, 1942, pág. 73.

(7) RODRIGUES, MILTON DA SILVA. Elementos de Estatística Geral. São Paulo, Companhia Editôra Nacional, 1939, pág. 30

ada para o estudo quantitativo dos fenômenos que necessitam de coleções ou massa de observações". (8)

Examinando cada um dos enunciados acima transcritos, em proporção suficiente a distinguir os seus pontos comuns, podemos perceber sem dificuldades a idéia central remanescente em todos êles, após excluídas as diferenças de forma decorrentes ùnicamente do modo de apresentá-los literalmente. Todos aquêles conceitos nos fazem encarar a estatística, fundamentalmente, como uma observação metódica por meio de processos e técnicas especiais, destinada a estudar sintética e analìticamente, tão minuciosamente quanto possível, dados quantitativos e numerosos subordinados à influência de causas múltiplas.

Quer olhemos a estatística segundo êste ou aquêle conceito, ressaltam os aspectos comuns a todos êles, identificados tanto na repetição das modalidades quantitativas do fenômeno no tempo ou no espaço, quanto na impossibilidade de descobrir-se em virtude das múltiplas e complexas causas atuantes, a verdadeira natureza do fenômeno no ato da observação. Então, perscrutada a regularidade das séries estatísticas, anotando-se por exemplo, o número n' de vêzes que aparece um determinado resultado num grupo maior de n modalidades repetidas, cuja razão $\frac{n'}{n}$ tende a aproximar-se de um valor fixo, torna-se possível especificar algum aspecto característico do fenômeno. Assim, esta condição que distingue em primeiro plano a natureza aleatória ou estocástica inerente à formação dos fenômenos complexos, (9) ergue a motivação principal do emprêgo da estatística à observação científica.

Em tôdas aquelas definições, implícitos estão os conteúdos da síntese e da análise regendo a execução de pesquisas em tôrno

---

(8) GINI, CORRADO., Os Fundamentos e o Alcance do Método Estatístico. In Revista Brasileira de Estatística. Ano IX, Julho/Setembro 1948, n. 35, pág. 301.

(9) RIOS, SIXTO. Op. cit., pág. 2.

de certa feição conexa às modalidades repetidas de um fenômeno. Sejam "valores sinaléticos e leis mais ou menos prováveis" como expõe L. Nogueira de Paula, ou "tendência características limites" consoante a terminologia de Milton da Silva Rodrigues, ou "técnica especial adequada ao estudo quantitativo dos fenômenos em massa" tal qual o pronunciamento de Gini, subentendem êstes juízos ora o procedimento condensativo da descrição - síntese - ora o procedimento dissociativo da interpretação - análise -, contanto que ligados a observações acumuladas, quantitativas e de manifesta complexidade.

Historicamente, a necessidade prática dos procedimentos estatísticos provem da inaplicabilidade do clássico método experimental ao estudo de algumas ciências cujos fenômenos são complexos, influenciados por várias causas extremamente modificáveis ao ponto de não permitir ao observador perceber a tendência dos próprios fenômenos. Estão nesta condição, dentre outras, as ciências sociais, nas quais a complexidade se faz sentir em tôda plenitude, exigindo dos pesquisadores métodos mais apropriados à apreensão das relações criadas no mecanismo evolutivo da sociedade. A propósito, na esfera da ciência econômica, Oskar Anderson salienta a íntima e espontânea associação da estatística à economia, escrevendo: -"...the relation between statistical analysis and economic theory is analogous to that between experimental and theoretical physics.... In physics conclusions deduced from hypothesis...are verified with the aid of experiments. In economics the place of physical hypotheses is taken by the several special theories, such as the quantity theory of money; while for experiments are substituted statistical investigation and analysis which "verify" theory and give it concrete substance by means of averages, index numbers, combination tables, correlation coeficients...." (10)

O tratamento especificamente estatístico, portanto, apareceu inevitàvelmente substituindo o método experimental em alguns ca-

---

(10) ANDERSON, OSKAR M. Statistics. In Encyclopedia of the Social Sciences, Vol. 14. New York, The Mac Millan Company, 1954, pag. 371.

sos e complementando-se em outros. Quando um fenômeno é insuscetível de análise experimental, não sendo por conseguinte possível a separação de algumas causas prováveis a fim de apreciar-se a extensão dos efeitos de uma única causa, ou mesmo que o seja restam ainda os chamados "êrros experimentais" motivados por alterações no contrôle da experiência, os recursos estatísticos servem à verificação e indagação nos limites da amostra, dos fatôres que possìvelmente provocaram variabilidade nos efeitos repetidos, os quais se apresentam distintos entre si nas observações sucessivas, dado à contingência aleatória.

Quanto à impraticabilidade da renovação de uma modalidade fenomênica rigorosamente nas mesmas condições anteriores, inclusive no círculo das ciências denominadas exatas, conclui Albert Waugh relativamente à física e à astronomia, que os cientistas defrontam em suas investigações êrros de observação, razão por que são muitas vêzes conduzidos a aplicarem o método estatístico além dos processos experimentais, realizando desta maneira uma apreciação mais acurada de tôdas as variações. (11) . Tais êrros provenientes das medições consecutivas da grandeza em questão, cujas mensurações conquanto efetuadas sob condições análogas dão lugar a resultados diferentes, são também reportados por Corrado Gini, que reputa a estatística um complemento indispensável à eliminação dos êrros acidentais perturbadores das observações astronômicas, físicas, meteorológicas etc. (12)

Incontestàvelmente, os eruditos mestres tanto no Brasil como no estrangeiro, são unânimes com referência às condições de aplicabilidade da estatística na estrutura quantitativamente ampliada de um fenômeno e desdobrada nos valores superpostos de uma ou mais grandezas, em que a complexidade oriunda da natureza das variações estocásticas, mostra-se inatacável pelos habituais métodos de investigação científica.

---

(11) WAUGH, ALBERT E. Elementos de Estatística. Porto Alegre, Editôra Globo, sem indicação de ano, pags. 2-3

(12) GINI, CORRADO, Curso de Estatística. Barcelona, Editorial Labor S.A., 1953, pags. 16-17.

Sejam fenômenos naturais ou sociais, assumem à medida que evoluem, aspectos de crescente complexidade, constituindo universos homogêneos ou heterogêneos, conforme haja variação única ou múltipla ao longo da sua trajetória. Não resta outra alternativa senão usar segundo o critério adaptável ao tipo de variação observada na amostra, os processos e técnicas peculiares ao método estatístico, aplicando-se a síntese ou a análise com devida justeza respectivamente para destacar características ou determinar relações no conjunto de dados quantitativos tomados em grande número e sujeitos a um paradigma aleatório.

SEGUNDA PARTE

MODOS SINTÉTICOS

- CAPÍTULO IV -

SÍNTESE POR RECOMPOSIÇÃO

A metodologia sintética, de caráter limitativo, envolve uma idéia e determina uma ação de natureza ainda primária sôbre os agregados numéricos, de acôrdo com as necessidades merarente informativas da observação unidimensional. Destarte, o processo condensativo desdobra-se em duas operações complementares, quando aplicado à caracterização dos conjuntos monovariáveis: -

a) - recomposição dos elementos brutos prèviamente coletados. É neste sentido que se fala da ordenação de dados estatísticos para composição da amostra.

b) - indiciação dos elementos já organizados mètodicamente em distribuições. Por esta forma, calculam-se valores sinóticos, espécies de indicadores que dão uma idéia em têrmos sumários, das características de um conjunto amostral.

Sobressaem, portanto, inicialmente como fase introdutória de arranjo dos dados, uma síntese por recomposição, cujo papel é dar em forma condensada uma figuração mais expressiva aos elementos quantitativos esparsos; depois, uma síntese por indiciação, constituindo a descrição pròpriamente dita, que se propõe a destacar por meio de um só valor, tendências peculiares a fenômenos simples.

A síntese por recomposição aparece para desempenhar o seu papel condensativo, no ato da organização de uma amostra, consistindo no agrupamento metódico dos elementos numéricos primitivamente difusos; estabelecem-se grupos especiais de valores, compreensivos de parcelas dos agregados, seguindo-se o cômputo de quantas vêzes ocorrem os valores da variável em cada um dêsses grupos ou classes, assim co-

mo a distribuição global destas ocorrências. Tal problema de disposição sistemática dos dados estatísticos é de transcendental importância, havendo portanto razões de ordem teórica e prática que demonstram a necessidade de fazer preceder à síntese por indiciação, um trabalho meticuloso de síntese por recomposição.

Efetivamente, a descrição estatística encontra apoio no novo aspecto que se dá à série primitiva, quer seja assinalado apenas o número de valores repetidos da variável ou seja especificado o número de ocorrências dos valores da variável em cada uma das classes especiais prèviamente estabelecidas. Daí surgem as "freqüências absolutas", número de vêzes que se repete uma determinada magnitude da variável, ou quantidade dos valores que ocorrem entre os limites daquelas classes definidas. A freqüência absoluta é simbolizada por f_i ($i = 1,2,3,....n$) onde f representa a intensidade dos valores X_i ($i = 1,2,3,....n$).

A amostra monovariável, resultante da conjugação de grupos parcelados da variável paralelamente à anotação das respectivas freqüências absolutas, compreende uma distribuição de freqüência, cuja construção atendendo aos requisitos de ordenação metódica, consentâneos com os princípios da síntese por recomposição, permite que os caracteres de monovariação sejam mais fàcilmente localizados pelas técnicas específicas à síntese por indiciação. Uma série estatística, portanto, agrupada em coleção amostral desta espécie, além de ser produto de um procedimento condensativo no sentido da recomposição de dados, supre também o material de que se servirá êste mesmo procedimento porém já em fase de indiciação, cuja função é sobremaneira descrever sumàriamente por meio de valores sinóticos, certas particularidades nos limites de um agregado monovariável.

Especialmente às variáveis discretas, ilustraremos a sua condensação em distribuições de freqüência, tomando o número de vêzes que 6 moedas em 200 tentativas, apresentam na queda as alternativas 0 cara, (6 coroas), 1 cara (5 coroas), 2 caras (4 coroas), 3 caras (3 coroas), 4 caras (2 coroas), 5 caras (1 coroa) e 6 caras (0 coroas). Reunindo todos os resultados derivados da experiência, ter-se-

á mostrado como se repetem aquelas diferentes combinações entre as 6 moedas ao cabo dos 200 lançamentos. Identificada a queda da moeda com o reverso para cima - cara - a um acontecimento favorável ou sucesso, então a coordenação da seqüência de valores $X_1 = 0$, $X_2 = 1$, $X_3 = 2$, ... , mediante o critério da síntese por recomposição, gera a seguinte distribuição de freqüência, a qual exibe modalidades repetidas da variável no âmbito amostral.

Valores da variável X_i - sucessos	Freqüência f_i
0	2
1	19
2	46
3	62
4	47
5	20
6	4
	200

Dados hipotéticos.

Nas séries estatísticas contínuas, a recomposição dos seus valores em distribuições de freqüência, além de apresentá-los em feição sintética, mostra a divisão da amostra monovariável em grupos ou classes, equivalendo esta condensação de variáveis contínuas ao problema de classificação das observações originais.(13)

Referindo, por exemplo, X_i (i = 1,2,3,.....n) aos valores de uma série contínua particularizada aos salários por hora de 100 operários, podemos submeter tais valores à técnica condensativa da síntese por recomposição e apresentá-los em distribuição de freqüência como segue: -

---

(13) CANSADO, ENRIQUE. Apuntes de Estadística General. Edição mimeografada. Centro Interamericano de Enseñanza de Estadística Economica y Financiera. Santiago, 1953, pag. 15.

Valor da variável	Freqüência
X_i - cruzeiros	f_i
40,1 - 45	5
45,1 - 50	15
50,1 - 55	20
55,1 - 60	35
60,1 - 65	13
65,1 - 70	9
	100

Dados hipotéticos

Estas duas distribuições de freqüência construídas em observância aos princípios e técnicas da síntese por recomposição, correspondendo a aspectos de amostras monovariáveis, podem ser também apresentadas em diagramas. As figuras 1 e 2 representam respectivamente o histograma e o polígono de freqüência da condensação acima expressa em números, aquela referente aos sucessos obtidos na experiência com 6 moedas lançadas 200 vezes. As figuras 3 e 4, por outro lado, traduzem da mesma forma, histograma e polígono de freqüência da condensação numérica relativa aos níveis salariais arrolados entre 200 operários.

Do exposto, depreende-se que a síntese dos agregados numéricos, conduzida por recomposição dos valores originais de uma variável, visa sobretudo a apresentação mais significativa possível dos dados estatísticos contidos na amostra monovariável.

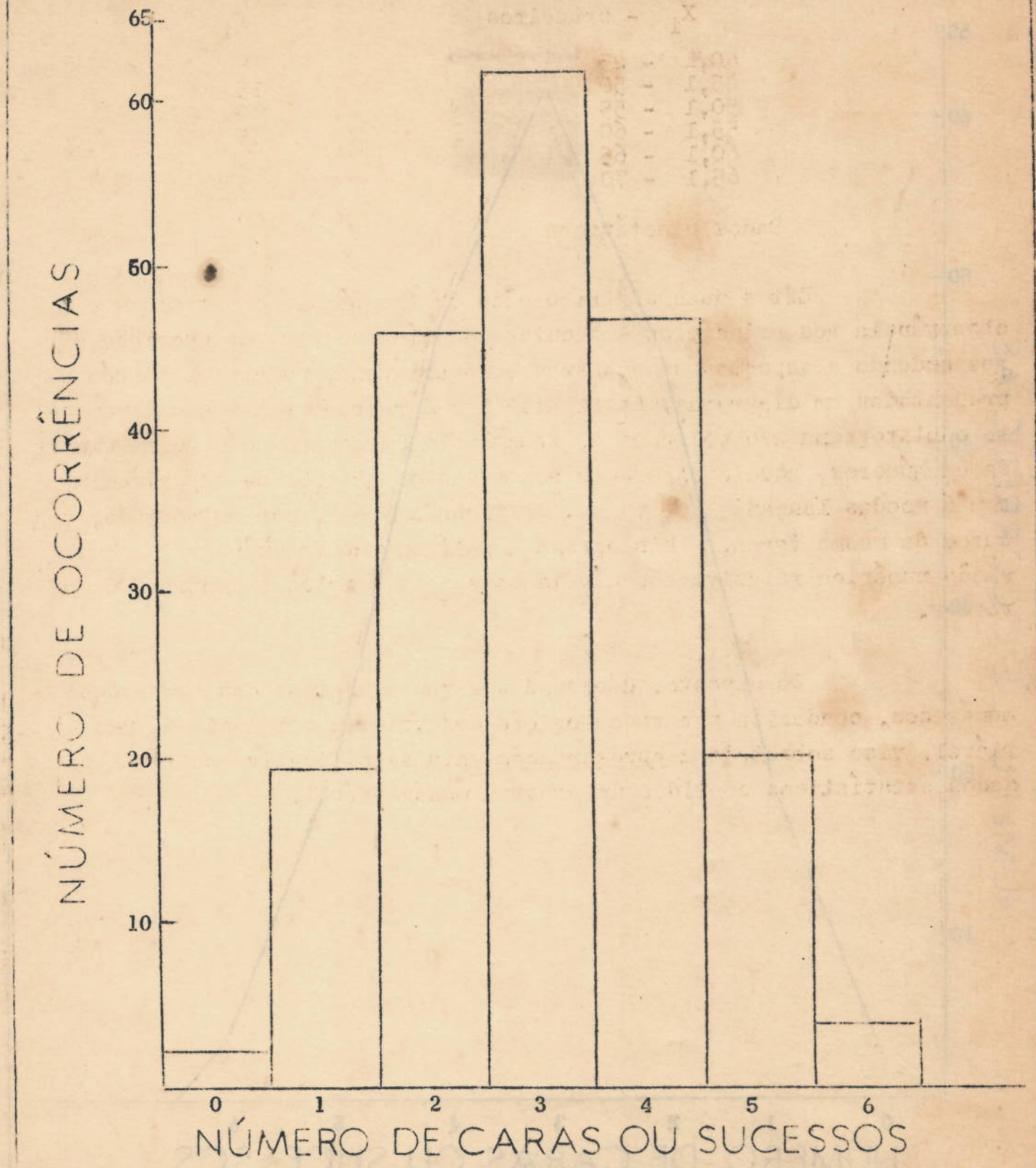


FIGURA 1- Histograma . Distribuição de 200 lançamentos de 6 moedas, classificados segundo o número de caras ou sucessos. (Dados hipotéticos)

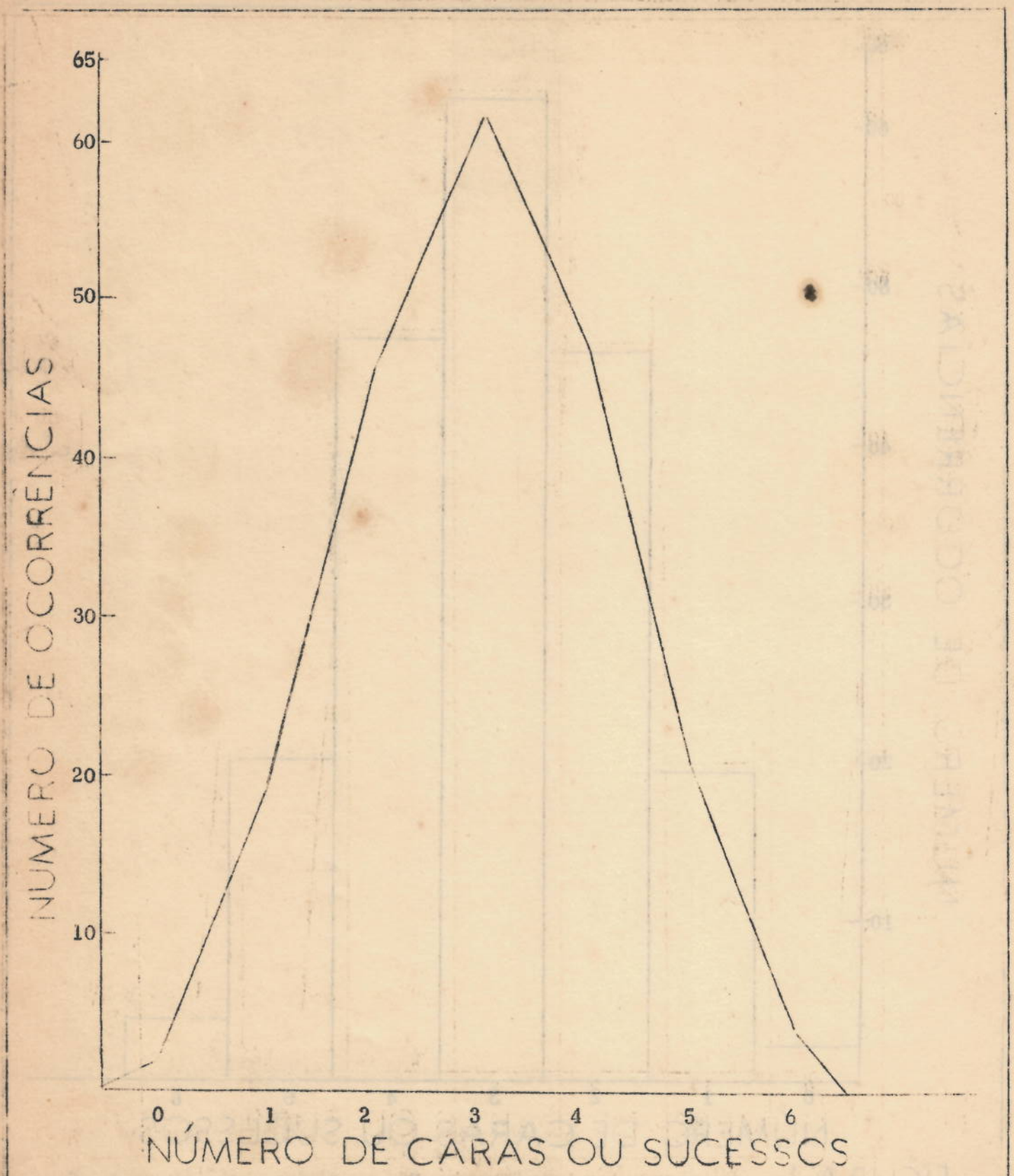


FIGURA 2- Polígono de freqüência. Distribuição de 200 lançamentos de 6 moedas, classificados segundo o número de caras ou sucessos (Dados hipotéticos).

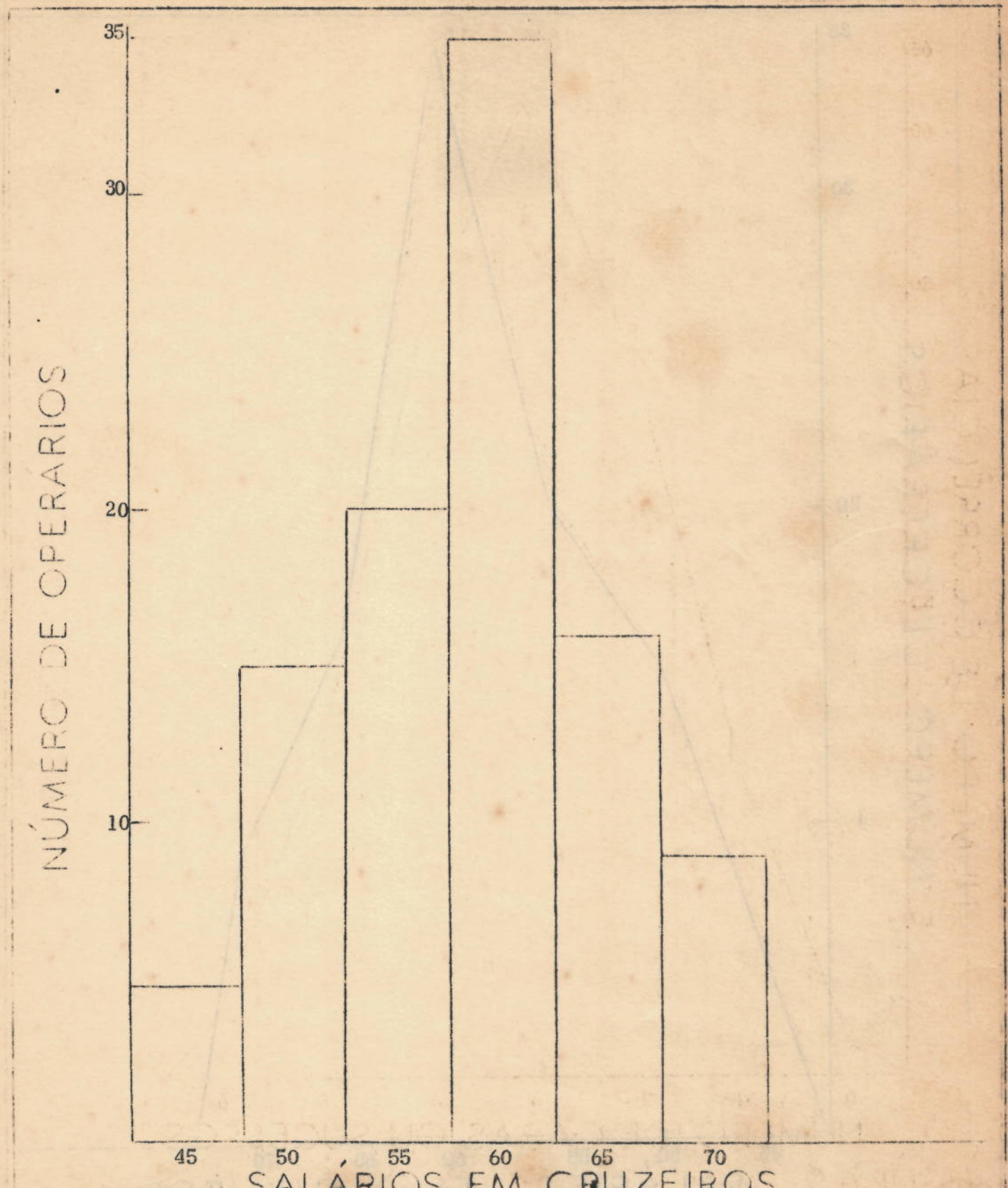


FIGURA 3- Histograma. Distribuição de 100 operários, classificados segundo os salários por hora (Dados hipotéticos).

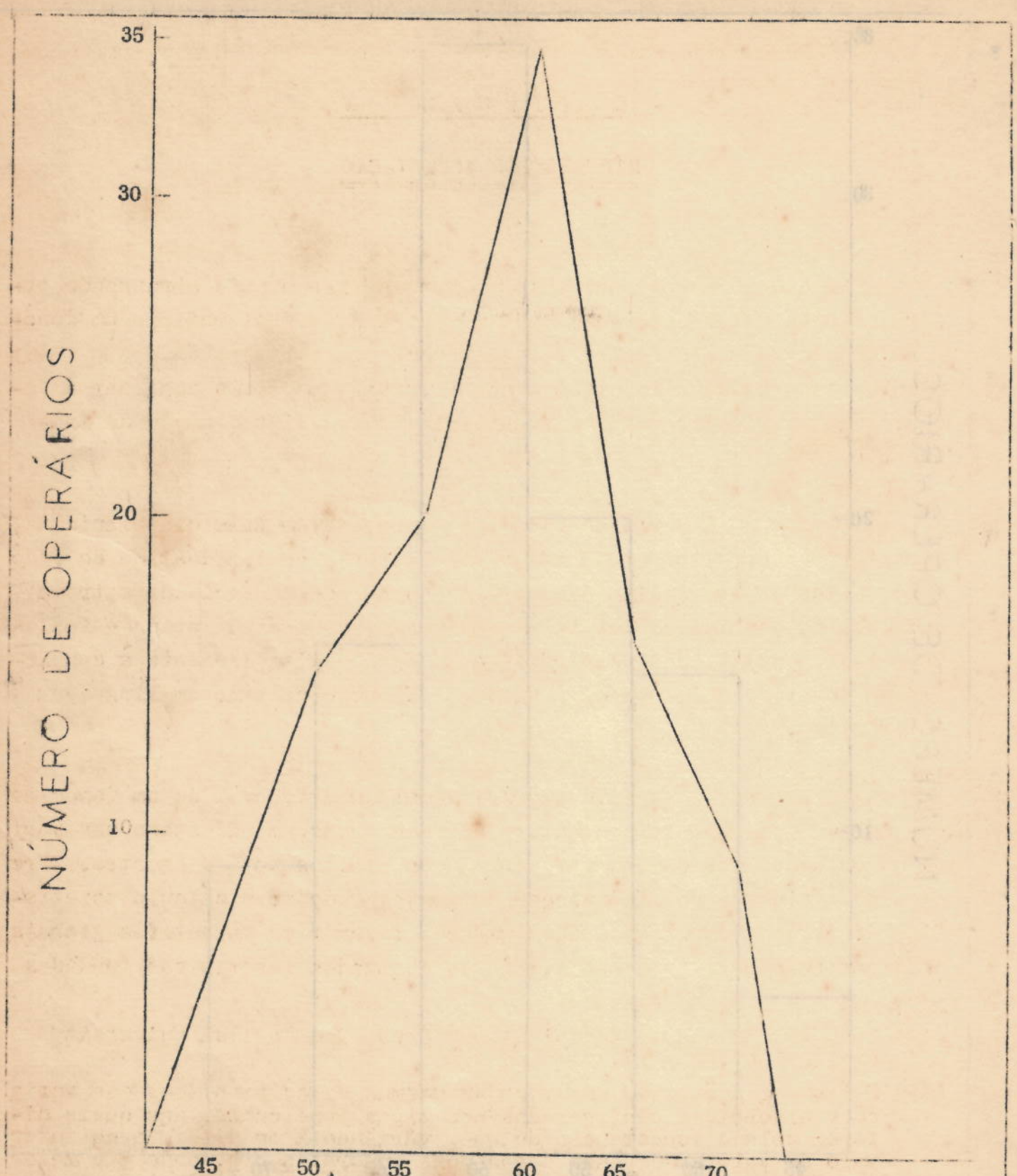


FIGURA 4- Polígono de freqüência. Distribuição de 100 operários, classificados segundo os salários por hora. (Dados hipotéticos)

- C A P Í T U L O V -

SÍNTESE POR INDICIAÇÃO

Em qualquer distribuição de freqüência abrangendo observações em grande número de uma variável X, a qual reflete um fenômeno complexo, tão singular é o padrão de sua variabilidade que a terminologia anglo-saxônia criou o vocábulo variate para designar a variável cujos valores estão circunscritos a uma distribuição de freqüência.(14)

A variação especial observada em tais distribuições, caracteriza perfeitamente a natureza aleatória ou estocástica do fenômeno, pelos efeitos muito diversificados em conseqüência da extrema sutileza na combinação dos fatôres causais, cuja peculiaridade fez estender-se o qualificativo aleatório ou estocástico inerente à complexidade do próprio fenômeno, à variável X que o exprime em linguagem estatística.(15)

O aspecto excessivamente modificável de um fenômeno por motivo da complexa contextura das concausas, manifestado nas variações estocásticas de uma variável X nos limites de uma amostra, abre o campo à síntese por indiciação, desenvolvendo-se o cálculo estatístico de algumas constantes destinadas a resumir os caracteres globais da distribuição de freqüência. São as operações descritivas fundadas

---

(14) GOODMAN, RICHARD. Statistics. Londres, The English University Press, 1957, pág 17.

(15) Estas variações não encerram as mesmas características das variações algébricas configuradas nos tipos funcionais, das quais diferem pela circunstância de que, variando X em determinado intervalo de valores e sendo dados dois números a e b, onde b é maior do que a, podemos determinar - P (a<X<b) -, introduzindo o símbolo P a noção de probabilidade na identificação dos caracteres da variável. (CLARK, CHARLES E. An Introduction to Statistics. New York, John Wiley & Sons, Inc., 1953, pag. 53)

na aplicabilidade do critério unidimensional às amostras monovariáveis, que produzirão os valores únicos aptos a evidenciar sumàriamente alguma particularidade proveniente da natureza aleatória do fenômeno.(16)

Realizando-se a descrição com a síntese por indiciação, são reduzidos os elementos quantitativos de uma amostra monovariável a valores representativos de uma situação global de X na respectiva distribuição de freqüência, onde a idéia de síntese dirige a determinação de constantes indicadoras das tendências de contração e distensão dos valores de variável.

A síntese por indiciação impõe de início o nivelamento dos valores desiguais porém homogêneos de uma variável X, calculando-se para isso o promédio como valor sinótico, quando então é substituído o conjunto total de valores da variável aleatória, por aquêle valor típico que mede a contração de X. Obtemos desta maneira um valor central ao redor do qual se acham repartidas as freqüências dos demais valores com maior ou menor uniformidade conforme a natureza da distribuição.

Na fixação de um só elemento quantitativo suscetível de caracterizar o grau de contração da variável X, é preciso que o valor estabelecido para êsse fim esteja o mais próximo possível dos outros valores da distribuição. O promédio será tanto mais preciso quanto menor for o seu afastamento em relação às demais magnitudes de X,

---

(16) Na concepção exclusivamente matemática, é a associação da probabilidade com a variável, que permite o conceito de variável aleatoria. Assim, se considerarmos a variavel X capaz de tomar os valores X_i (i=1,2,3,....n), sendo P_i (i=1,2,3,...n) as probabilidades que ela tem de assumir os respectivos valores, de tal modo que $\Sigma P_i=1$, diz-se que X é uma variável aleatória. Em nosso trabalho, entretanto, quando especificamos a natureza aleatória de uma variável, referimo-nos mais especialmente aquela classe de fenômenos cujas variações se sucedem com caracteres peculiares, dado a sua inerente complexidade. É portanto, ao aspecto empírico, estatístico, e não ao matemático propriamente dito, que ligamos a conceituação estocástica. Critério idêntico no tratamento da matéria encontra-se em GOODMAN, RICHARD, Op. cit., passim.

e apoiados neste requisito Alvarez e Orejana conceituam promédio como a "constante que mede o valor com respeito ao qual apresentam menos diferença todos os demais valores da distribuição".(17)

Ajusta-se perfeitamente êste conceito aos promédios básicos conhecidos sob a denominação de médias, as quais A. Julin define como "expressões do estado quantitativo normal de um determinado fenômeno";(18) dentre estas médias indentificaremos os tipos aritmético e geométrico aos fundamentos matemáticos dos valores típicos dessa natureza.

Sendo X_i (i= 1,2,3,...n) os valores de uma variável aleatória limitados a uma distribuição monovariável e P a constante que descreve sumàriamente a tendência de contração no agregado numérico, podemos estabelecer: $-\sum_{i=1}^{n} f_i (X_i - P)^2 = E$

Atendendo ao requisito da máxima aproximação entre P e os valores X_i, precisamos achar um valor tal para P que torne mínima a função acima. Derivando-a em relação a P, anulando a derivada e resolvendo a equação resultante: -

$$\frac{dE}{dP} = 2\sum f_i (X_i - P) \qquad (19)$$

$$\frac{dE}{dP} = 2\sum f_i X_i - 2\sum f_i P$$

$$2\sum f_i X_i - 2\sum f_i P = 0$$

$$\sum f_i X_i = \sum f_i P$$

---

(17) ALVAREZ, M. GARCIA e OREJANA, J. AYUSO. Estadistica. Madrid, Editorial S.A.E.T.A., 1946, pág. 219.

(18) Apud CARVALHO, BULHÕES. Estatística, Método e Aplicação. Rio de Janeiro, Tip. Leuzinger, 1933, pag. 208.

(19) Neste e em vários outros casos ulteriores, omitiremos o índice i e o limite inscrito acima do somatório $\sum$, com o fito de facilitar a marcha das deduções.

$$P = \frac{\sum f_i X_i}{\sum f_i} \qquad (5.1)$$

Fazendo $f_i = 1$, resulta:

$$P = \frac{\sum X_i}{N} \qquad (5.2)$$

A expressão (5.1) define especialmente a média aritmética de uma variável X em distribuição de freqüência, ou quando houver simplesmente magnitudes repetidas na amplitude de X. Por outro lado,(5.2) corresponde à média aritmética simples, aplicável a uma série cujos valores ocorrem com igual intensidade ao longo da amplitude de X.

Medindo os desvios em têrmos de razões entre cada valor da variável aleatória e a sua média, representaremos: - $x_i = \frac{X_i}{P}$

Aplicando logarítmos: -

$$\log x_i = \log X_i - \log P \qquad (5.3)$$

Exprimindo a soma dos produtos das diferenças logarítmicas ao quadrado pelas respectivas freqüências, e substituindo o primeiro membro da (5.3) por E: -

$$E = \sum f_i (\log X_i - \log P)^2$$

O valor de logP que torna mínima esta função é: -

$$\log P = \frac{\sum f_i \log X_i}{\sum f_i} \qquad (5.4)$$

Dando à expressão (5.4) um novo aspecto em têrmos de P:-

$$P = \sqrt[\sum f_i]{\text{II } X_i^{f_i}}$$

Fazendo $f_i = 1$, temos: -

$$P = \sqrt[N]{\text{II } X_i} \qquad (5.6)$$

Semelhantemente à distinção estabelecida quanto ao promédio aritmético, a expressão (5.5) define a média geométrica de uma variável, seja nos casos de distribuições de freqüência ou havendo apenas magnitudes repetidas na gama de X. O cálculo da (5.6), por sua vez, conduz à média geométrica simples, isto é, se a intensidade dos valores X_i for a mesma em tôda a escala de variação.

Êstes promédios ou valores médios especiais são calculados quando a propensão de contração da variável aleatória se verifica no centro da distribuição, refletindo simetria entre os valores de X limitados à observação anostral. São obtidos desta maneira os valores sinóticos, que revelan as características de simetria resultantes de condensação especìficamente no centro da distribuição. Aferindo-se esta tendência pelos menores desvios aritméticos de X a contar da média $(X_i - P)$, empreganos o promédio aritmético, enquanto se reserva o promédio geométrico aos casos em que tal tendência é medida pelos menores desvios geométricos de X em relação à média $\left(\frac{X_i}{P}\right)$.

Conforme depreendemos das deduções que levaram às conclusões expressas nas fórmulas (5.1) e (5.5), na hipótese de haver desvios em têrmos aritméticos - diferenças -, a média aritmética será mais estável e precisa como medida de contração; mas se o caso for de desvios em têrmos geométricos - razões -, o promédio geométrico indubitàvelmente oferecerá maior estabilidade e precisão.

Em oposição ao papel da síntese por recomposição que conduz apenas a uma apresentação mais perfeita dos dados estatísticos, os meios da síntese por indiciação, entretanto, desenvolvem-se no sentido de prover informação tão condensada quanto possível dos elementos de uma amostra monovariável.

- C A P Í T U L O VI -

DERIVAÇÕES DA SÍNTESE POR INDICIAÇÃO

Visto que duas distribuições podem ter o mesmo promédio, diferindo porém uma da outra quanto aos afastamentos dos valores das respectivas variáveis em relação àquele valor sinótico, cabe determinar a variabilidade de X em tôrno do ponto de máxima condensação. Esta medida da proporção em que os diversos valores de X se distanciam do valor central da distribuição, envolve um procedimento calculatório especial, sendo para isso indispensável o assentamento de outra categoria de valores sinóticos, desta vez com o objetivo de descrever as características de distensão da variável nos limites da amostra monovariável.

Uma das constantes mais apropriadas à avaliação desta distensão numa distribuição de freqüência, é o desvio padrão, o qual se determina em têrmos de afastamentos quadraticos de cada lado da média aritmética, pois a soma dos desvios ao quadrado é a menor possível, quando o valor atribuído ao promédio equivale à média aritmética.(20) Esta particularidade define a estabilidade daquela constante como medida de distensão, tôda vez que se desejar mensurar a variabilidade de X em têrmos de seus afastamentos com respeito ao promédio aritmético da distribuição.

Originalmente podemos exprimir o desvio padrão de uma variável aleatória X em distribuição de freqüência ou não sendo êste especialmente o caso, mas se houver repetição de algumas das magnitudes X_i, da seguinte maneira: -

---

(20) Demonstramos esta propriedade à pág 41, Cap.5, ao abordarmos a questão da máxima aproximação, ou inversamente mínimo afastamento que X_i deve manter em relação ao valor sinótico destinado a medir o grau de contração de uma variável.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_1)^2}{N}} \qquad (6.1)$$

Elevando ao quadrado ambos os seus membros, temos a variância: -

$$s_x^2 = \frac{\sum_{i=1}^{n} f_i (x_i)^2}{N} \qquad (6.2)$$

Sendo x_i por definição igual a $(X_i - P)$, isto é, a diferença entre o valor da variável e o promédio aritmético da distribuição, daremos à expressão (6.2) o seguinte aspecto: -

$$s_x^2 = \frac{\sum_{i=1}^{n} f_i (X_i - P)^2}{N} \qquad (6.3)$$

Êste quociente, que equivalendo ao momento de segunda ordem da variável X com origem na própria média, é uma forma especial da expressão geral dos momentos de ordem r em tôrno do promédio aritmético: -

$$m_r = \frac{\sum_{i=1}^{n} f_i (X_i - P)^r}{N}$$

A orientação do estudo concernente à distensão, no sentido de atribuir valores sinóticos essenciais à sua avaliação matemática, dimana fundamentalmente da teoria dos momentos estatísticos. De modo geral, a determinação do momento de certa ordem ao redor da média, em função do momento de igual ordem porém tomado a partir de uma origem arbitrária, obedece ao seguinte princípio: -

Sendo $z_i = X_i - A$, isto é, o excesso do valor da variável aleatória X sôbre o valor de A, em que $A \neq P$, o primeiro momento da variável X em relação à origem arbitrária será: -

$$\mu_1 = \frac{\sum f_i z_i}{N}$$

Substituindo z_i por $(X_i - A)$, acharemos: -

$$\mu_1 = \frac{\sum f_i (X_i - A)}{N} = \frac{\sum f_i X_i}{N} - \frac{\sum f_i A}{N}$$

Portanto: -

$$\mu_1 = P - A = d$$

Assim,

Assim, o primeiro momento em relação à origem arbitrária A, equivale à diferença entre a média e essa origem arbitrária, cuja diferença chamamos d.

Para o caso particular do momento de ordem 0, temos: -

$$\mu_0 = \frac{\sum f_i (X_i - A)^0}{N} = \frac{\sum f_i}{N} = \frac{N}{N} = 1$$

Retornando à expressão $z_i = X_i - A$, podemos transformá-la em: -

$$z_i = (X_i - A) = (X_i - P) + (P - A) = x_i + d$$

Então: -

$$x_i = z_i - d$$

Para achar a expressão geral do momento de ordem r relativamente à média, em função do momento r em tôrno de uma origem arbitrária, precisamos atender a que: -

$$\sum (f_i x_i^r) = \sum \left(f_i (z_i - d)^r \right) = \sum f_i z_i^r - C_r^1 \, d \sum \left(f_i z_i^{r-1} \right) + \ldots$$

.....

$$.... + C_r^2 d^2 \sum (f_i z_i^{r-2}) - C_r^3 d^3 \sum (f_i z_i^{r-3}) + ... \pm d^r \sum f_i$$

Dividindo por N: -

$$\pi_r = \mu_r - C_r^1 d\mu_{r-1} + C_r^2 d^2 \mu_{r-2} - C_r^3 d^3 \mu_{r-3} + ... \pm d^r$$

Sendo $n = 1$, temos: -

$$\pi_1 = \mu_1 - \mu_1 = 0$$

$$\pi_1 = 0$$

No caso de $n = 2$, resulta: -

$$\pi_2 = \mu_2 - d^2$$

$$\pi_2 = \mu_2 - \left(\mu_1^2\right) \quad (6.4)$$

Para n = 3: -

$$\pi_3 = \mu_3 - 3\mu_1\mu_2 + 2\left(\mu_1^3\right) \quad (6.5)$$

Revertendo à expressão (6.3) e desenvolvendo o segundo membro: -

$$s_x^2 = \frac{\sum f_i X_i^2 - 2P\sum f_i X_i + \sum f_i P^2}{N}$$

Ou

$$s_x^2 = \frac{\sum_{i=1}^{n} f_i X_i^2}{N} - P^2 \quad (6.6)$$

Deduzimos assim uma fórmula alternativa para determinação da variância e subseqüentemente do desvio padrão, em cuja fórmula figuram os valores da variável X e a média. Compreende-se que a expressão (6.6) corresponde ao momento de segunda ordem da média, em função do momento de igual ordem a partir da origem 0, da mesma forma que a (6.4) define o momento de segunda ordem da média em função do momento segundo relativamente a uma origem arbitrária.

Fazendo $f_i = 1$ em (6.2) e (6.6), verificamos que: -

$$s_x^2 = \frac{\sum_{i=1}^{n} x_1^2}{N} \quad \text{e} \quad s_x^2 = \frac{\sum_{i=1}^{n} X_i^2}{N} - P^2$$

Ou em têrmos de desvio padrão: -

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{N}} \quad \text{e} \quad s_x = \sqrt{\frac{\sum_{i=1}^{n} X_1^2}{N} - P^2}$$

Diferentemente das expressões (6.2) e (6.6), estas duas aplicam-se aos casos onde não existem valores repetidos no curso da observação, sendo a intensidade portanto a mesma para cada um dos elementos componentes da série.

Delimita o desvio padrão a amplitude dentro da qual se distendem os valores de X abaixo e acima do promédio aritmético, indicando o campo de variabilidade da distribuição, circunscrito aproximadamente aos marcos $P \pm s_x$.

Argúi-se que o desvio padrão, expresso na mesma unidade da variável X, sendo portanto uma medida absoluta, nem sempre avidencia uma distensão ampla ou restrita na amostra monovariável; duas distribuições podem ter igual desvio padrão, apesar de ser uma mais instável do que a outra. É necessário, por conseguinte, a fim de avaliar-se rigorosamente a maior ou menor propensão com que os valores X_i se afastam do promédio, estabelecer uma medida relativa de distensão, comparando-se

para isso o desvio padrão ao promédio aritmético P, geralmente sob forma percentual. Para preencher esta finalidade, Karl Pearson propôs a razão - $\frac{s_x 100}{P}$ - como coeficiente de variação, possibilitando dêste modo a comparação da distensão de duas ou mais distribuições, o que seria impraticável ao desvio padrão por sí só realizar.

TERCEIRA PARTE

ESTRUTURAÇÕES ANALÍTICAS

- CAPÍTULO VII -

COEXISTÊNCIA SINTÉTICO-ANALÍTICA

Consoante discorremos no capítulo 1, o pesquisador não depara ùnicamente com circunstâncias em que, face à homogeneidade da amostra, basta uma elaboração sintética através de processos e técnicas descritivos, simplesmente destinada à verificação das características de contração e distensão de uma única variável X.

Ao contrário, muitas vêzes urge inquirir sôbre uma situação consubstanciada na heterogeneidade da amostra, cuja condição exige a adoção de procedimentos analíticos e dá lugar às interpretações da evolução de uma variável X conjuntamente com outra variável Y, que tomam os valores X_i (i =1,2,3,.....n) e Y_i(i=1,2,3,.....n).(21)

Há pelo menos duas variáveis aleatórias - X e Y - a considerar na amostra: - duas grandezas diferentes, representativas da duplicidade de fatôres componentes do problema e da heterogeneidade do universo em questão. Por isso, nos casos de amostras bivariáveis e notadamente das multivariáveis, devem ser mais refinados os meios calculatórios aplicáveis ao estudo do regime da variação das grandezas integrantes do fenômeno. O emprêgo dessa instrumentação mais aperfeiçoada, indispensável ao estudo de amostras dessa natureza, faz-se em harmonia com o sentido interpretativo, por sua índole extensivo, especialmente derivado do critério pluridimensional.

Enquanto as amostras monovariáveis não requerem senão a aplicação do processo meramente sintético, baseado no critério unidimensional, assumindo portanto as pesquisas nesse setor um caráter puramente descritivo, as amostras bivariáveis e multivariáveis exigem a utili

---

(21) Baseamos nossa explanação na hipótese de dados não agrupados, em que a cada valor de X correspondente somente um valor de Y e a cada valor de Y corresponde um só valor de X.

zação de uma metodologia sensivelmente analítica de origem pluridimensional. Conseqüentemente, as pesquisas nesta última fase, de maior alcance científico, distinguem-se pela natureza profundamente explicativa dos seus resultados. Tomamos contato, em decorrência das indagações apoiadas na identificação do instrumental estatístico às modalidades de bivariação e multivariação de uma amostra, com operações extensivas, opostas àquelas que, por serem menos dilatadas e suficientes aos estudos restritos a amostras monovariáveis, revelam-se de certo modo operações limitativas.

Ao aspecto abstrato do critério pluridimensional que disciplina em geral a perquirição das mudanças em amostras de duas ou mais variáveis, junta-se o aspecto concreto do processo dissociativo e da técnica calculatória das relações estocásticas já na fase de aplicação de cunho analítico aos elementos da própria amostra.

Retomando aquí e desenvolvendo o tema anteriormente abordado, concernente ao problema da metodologia analítica e sua conexão com a sintética, aduzimos que na fase interpretativa da estatística, análise e síntese não são procedimentos polares, reciprocamente repulsivos, mas sim complementares por coexistirem no curso da pesquisa estas duas diretrizes metodológicas. Verifica-se, ao invés de exclusão, uma condição de subordinação da síntese, cujas operações de indiciação por meio de valores sinóticos tendem a repetir-se nas séries constitutivas da amostra. A tipologia analítica adaptável ao propósito de interpretação especial em dada pesquisa, determina a natureza e a extensão da síntese por indiciação aplicável aos fins ainda descritivos que se faz mister atender em relação a cada uma das distribuições ou séries do campo amostral.

O estudo que ora começamos, mostrará como o processo dissociativo da análise traduzido na determinação das relações estocásticas e dos coeficientes especiais, supõe o processo condensativo da síntese por indiciação através dos valores sinóticos dos espécimes de médias, desvios padrões etc., os quais constituindo elementos iniciais, vão integrar a estrutura das próprias relações estocásticas e dos referidos coeficientes. Nem por isso, entretanto, perde a pesqui-

sa bivariável ou multivariável o seu caráter analítico, tanto em seu objeto quanto em seus resultados, da mesma forma que, nas ciências humanas, conserva-se analítica uma investigação de fundo enonômico, quando por exemplo, ao estudar a dinâmica do fenômeno do desemprêgo, o economista procura antes de tudo reconstituir um conjunto de acontecimentos passados, tal qual etapa inicial de escolha dos elementos fundamentais, a fim de poder destacar a importância relativa do papel de certos fatôres aos quais a sua análise o conduz. O caminho sintético da reconstituição de ocorrências pretéritas de igual natureza, tão sòmente complementar, não invalida nem subestima o caráter precìpuamente analítico da pesquisa empreendida pelo economista.

No domínio da análise, percebemos o método estatístico na trajetória da sua maturação relativa,(22) quando em indagações sôbre as variações inerentes à complexidade da amostra consequente à heterogeneidade do fenômeno, identificam-se tanto o aspecto conceitual formulado em concepções pluridimensionais quanto o modo objetivo de exame das sucessões de magnitudes de diferentes grandezas,às questões específicas compreendidas na interpretação.

Convém ilustrar de que maneira se constata a manutenção das modalidades do procedimento sintético simultaneamente à crescente diferenciação do grau de elaboração analítica, discutindo em primeiro lugar um caso de duas variáveis aleatórias - X e Y -, expressões quantitativas das intensidades de dois fatôres co-atuantes na formação de um fenômeno complexo. As combinações tècnicamente mais apuradas e peculiares aos meios analíticos, surgem no instante em que são superadas as limitações da simples fase sintética, advindo dessa circunstância o desdobramento da análise em proporções acentuadas,dado à necessidade de um estudo interpretativo imposto pela crescente complexidade dos fenômenos,cujos fenômenos são revelados quantitativamente por múltiplas

---

(22) Temos o cuidado de qualificar de relativa esta maturação, porque a passagem da descrição à interpretação envolve apenas uma das componentes de um sistema gerador da evolução estatística. Outro aspecto que poderia denotar maturação, é a passagem do sentido descritivo ao inferencial, comportando êste último uma diferenciação oriunda da aplicação consistente da teoria das probabilidades.

séries estatísticas.

Consideremos nestas condições um caso específico de bivariação, onde os vários pares de magnitudes X_i e Y_i das duas variáveis, são submetidos a uma meticulosa análise bivariável, para que os agregados numéricos mostrem uma relação interdependencial aproximada ou média, a qual em observações de amostras semelhantes se verifica sem discordância appeciável do resultado original.

Em geral, o critério pluridimensional, regulador da análise estatistica, consolida a sistemática da indução científica, pois tanto mais eficiente e precisa será a verificação de uma dada hipótese pela observação dos efeitos de um fenômeno, quanto mais sedimentadas estiverem as conclusões indutivas nas relações médias estabelecidas por meios particulares da interpretação, cujas relações dão uma medida do grau de aderência entre o que conhecemos teòricamente e aquilo que assimilamos pràticamente através do cálculo específico.

A análise estatística de uma amostra bivariável compreende inicialmente a fixação de uma equação que possa indicar a natureza da relação e a taxa de variação entre as duas variáveis, seguindo-se a determinação dos coeficientes específicos que ora medem o grau de relação entre as duas variáveis ora servem para avaliar a proporção de variabilidade nos valores de Y - variável dependente - imputada à concomitante variação nos valores do X - variável independente.

Ao se estabelecerem esta equação e êstes coeficientes relativamente a duas variáveis, sem embargo do seu caráter estocástico assimilável sobretudo por uma metodologia analítica, a síntese funciona acessòriamente, na medida em que o cálculo de valores sinóticos efetuado repetidamente nas séries parciais da amostra, faz emergir os elementos primários indispensáveis à estruturação daquêles instrumentos matemático-estatísticos destinados à interpretação de fenômenos complexos. Ao invés de serem os processos sintéticos excluídos com a interposição da metodologia analítica, êles se tornam entretanto necessários à consecução definitiva desta.

A espécie de dependência que a análise estatística sob critério bidimensional procura esclarecer, difere dos tipos de dependência funcional onde a análise puramente matemática permite calcular para Y um valor rigorosamente determinado pelo valor de X. O objetivo do estatista ao examinar uma amostra bivariável de dimensão suficiente, é descobrir se os diversos pares de valores (X_i, Y_i) apresentam alguma tendência sugestiva de aproximação a uma forma matemática definida, em quanto aumenta sucessivamente o número de elementos da amostra.

Visto ser peculiar a variação observada entre as duas variáveis aleatórias nos limites da amostra bivariável, embora não seja uniforme por causa da interferência no universo de fatôres desconhecidos ou inespecíficos cuja influência se reflete na amostra, o conhecimento de X - variável independente - faculta-nos mediante análise estatística, conhecer algo acêrca da variável dependente Y. Não havendo òbviamente uma relação rígida, de tipo funcional, verifica-se contudo uma relação de natureza estocástica, aleatória ou estatística entre as duas variáveis.(23)

Gràficamente, a tendência que as duas variáveis conjuntamente encerram, pode ser revelada num diagrama de dispersão, em que os valores X_i são abcissas e os Y_i ordenadas. Na figura 5 os valores elevados de X estão associados aos altos valores de Y, ao passo que na figura 6 os grandes valores de X estão associados aos baixos valores de

---

(23) Nem sempre é observada a rigor na taxinomia estatística, a conceituação de dependência estatística distintamente da de dependência funcional. Alguns, inclusive Mordecai Ezekiel, que pontifica os estudos da teoria e método da correlação estatística, estende a denominação relação funcional aqueles dois tipos de dependência, estabelecendo a distinção apenas quanto ao modo de determinação destas relações. Segundo êle, enquanto se pode determinar matematicamente relações funcionais a partir de conclusões experimentais onde são mantidas constantes tôdas as influências exceto a que se esta analisando, muitos problemas, porém, requerem sejam determinados estatisticamente as suas relações funcionais; ha, portanto, uma relação funcional em qualquer caso, podendo esta relação no entanto ser encarada sob o ângulo matematico ou estatístico, em consonancia com o criterio que tiver sido adotado na sua determinação. (EZEKIEL, MORDECAI. Methods of Correlation Analysis. New York, John Wiley & Sons Inc., 1956, págs. 34-46.

Y. Em ambos os casos as variáveis estão linearmente relacionadas, pois os pontos relativos à junção da cada par de valores das duas variáveis, ajustam-se aproximadamente a uma linha reta, a qual na figura 5 é ascendente e na figura 6 é descendente. Diz-se também que existe correlação linear positiva no primeiro caso e correlação linear negativa no segundo.

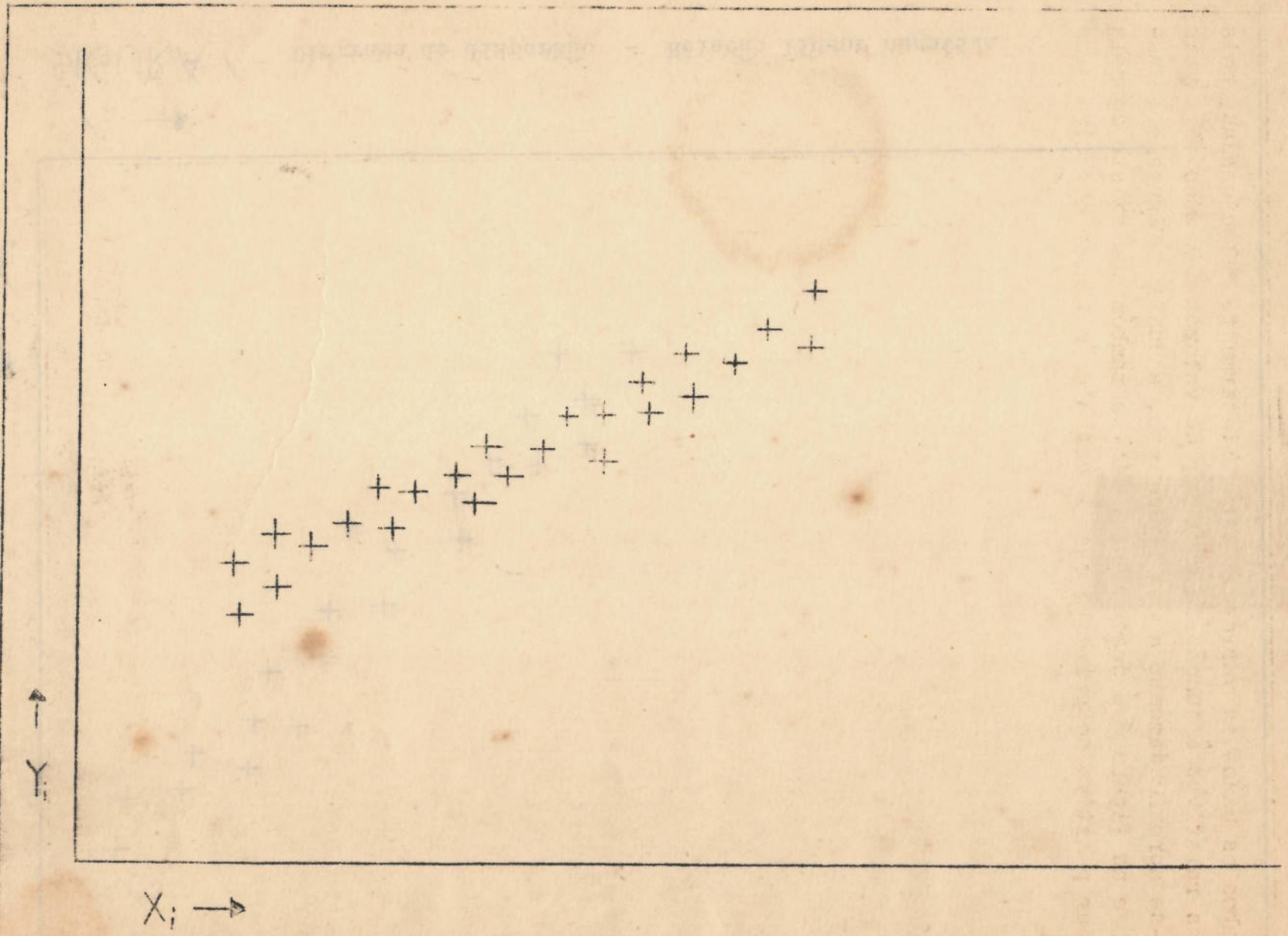


FIGURA 5- Diagrama de dispersão - Relação linear positiva.

FIGURA 6- Diagrama de dispersão - Relação linear negativa

- CAPÍTULO VIII -

REGRESSÃO SIMPLES

O problema discutido em primeiro plano na análise estatística de uma amostra bivariável, conforme assinalamos, diz respeito ao assentamento da expressão algébrica que exprime a relação média entre as duas variáveis, implicando isto no equacionamento destas variáveis de tal maneira que uma delas considerada dependente, fique bem expressa em função dos valores atribuídos à outra tomada como independente. Temos, portanto, $Y = \phi(X)$, se Y depender de X, ou inversamente $X = \phi(Y)$ se a variável X estiver influenciada por Y.

Havendo amostras bivariáveis formadas por pares (Y_i, X_i), quer os valores dos dois agregados aumentem simultaneamente, distribuídos segundo os pontos do diagrama da figura 5, ou os valores de um dos agregados aumentem - por exemplo os Y_i - enquanto os X_i diminuem, segundo uma série de pontos idênticos àqueles ilustrados no diagrama da figura 6, observa-se certa proporcionalidade na variação dos valores de cada par. Esta é a característica dos tipos de regressão linear, fazendo aparecer no campo da estatística as regressões retilíneas ou lineares, expressas matemàticamente pela equação que define a linha reta: - $Y = a + bX$.

Esta equação ressalta, na amostra bivariável, a dependência de Y para X, demonstrando a influência que um fator exerce sôbre o outro, ao qual está relacionado no âmbito do fenômeno global em universo de natureza igual à da amostra. Além de deixar bem claro qual a natureza da relação - ascendente ou descendente - entre as duas variáveis, conforme seja positivo ou negativo o parâmetro b, a equação permite igualmente aferir a mudança advinda à variável dependente Y quando a variável independente X varia unitàriamente, equivalendo êste último aspecto à taxa de variação entre os dois fatôres, cuja taxa está subor

tendida no valor numérico do mencionado parâmetro.

Concernente aos qualificativos dependente e independente atribuídos formalmente às variáveis Y e X que figuram respectivamente no primeiro e no segundo membro da equação, muitas vêzes sem nenhuma dependência causal unilateral entre as variáveis, alguns autores julgam aquêles dois vocábulos impróprios dado ao seu teor genérico. Assim Wold(24) propõe se aceite em substituição os têrmos variável-efeito (Y) e variável-causa (X) quando houver sido constatado diante de dados empíricos uma situação de dependência efetivamente causal, e variáveis a explicar (Y) e variáveis explicativas (X) no caso de considerar-se o aspecto puramente formal da posição relativa das duas variáveis na equação.(25) Ademais, a terminologia corrente abrange outras denominações para as variáveis Y e X, ao tempo em que se introduz a interpretação causal nos diferenciados campos científicos, encontrando-se em experiências biológicas os têrmos dosagem e variável-efeito, em psicologia experimental as variáveis estímulo e resposta, refletindo ambos os casos um tratamento estatístico de dados experimentais.(26) Ao lado destas existem outras designações aplicáveis à análise de elementos gerais não provenientes de qualquer observação concreta.

Esta redução linear conduz ao problema de atribuir valores numéricos aos parâmetros a e b da equação $Y = a \pm bX$, sobrelevando está transformação as primeiras contingências de repetição constante das modalidades sintéticas, não obstante a aplicação analítica que já se faz sôbre amostras bivariáveis. Deixa entrever a análise bivariável que a interpretação requer, a fim de preservar a sua proficiência, uma apreciável flexibilidade em seus meios com referência à síntese

---

(24) WOLD, HERMAN. Demand Analysis. New York, John Wiley & Sons, Inc., 1953, pag. 33

(25) Variáveis a explicar e variáveis explicativas equivalem a regressand e regressor, em cocordância com a tradução dêstes dois vocábulos saxonios, publicada pelo Centro de Estudos Econômicos anexo ao Instituto Nacional de Estatística de Portugal, após realização do Seminario de Econometria em maio de 1953 na cidade de Lisboa.

(26) Ibidem. pág 324.

por indiciação aplicada às séries da amostra, pois a recorrência das formas descritivas originais, cria modalidades analíticas mais apuradas na proporção do acréscimo de variáveis na amostra.

Começaremos com a valiosa instrumentação matemático-estatística dos momentos a partir da origem 0 e da média de ordem r e s em relação às duas variáveis X e Y respectivamente. Apesar de envolver duas variáveis, esta condição suscita em essência uma operação sintética singular, porquanto os referidos momentos dão sempre lugar à apreciação das características de contração e distensão, não hevendo por conseguinte razões de negar-se o caráter descritivo a uma determinação desta natureza, embora distinta das suas congêneres no concernente à apresentação de duas variáveis justapostas no curso da síntese por indiciação.

Dêste modo, o momento referido à origem 0 de ordem r com respeito a X e s com relação a Y, desde que não haja valores repetidos das variáveis, toma a forma de: -

$$a_{rs} = \frac{1}{N} \sum_{i=1}^{n} X_i^r Y_i^s \qquad (8.1)$$

Em particular à expressão acima, temos: - (27)

$$a_{10} = \frac{1}{N} \sum_{i=1}^{n} X_i = \bar{X}$$

$$a_{01} = \frac{1}{N} \sum_{i=1}^{n} Y_i = \bar{Y}$$

$$a_{11} = \frac{1}{N} \sum_{i=1}^{n} X_i Y_i$$

Os momentos em relação à média, admitidas as mesmas condi-

(27) Simbolizaremos daqui em diante as médias aritméticas de X e Y por $\bar{X}$ e $\bar{Y}$.

ções são: -

$$m_{rs} = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X})^r (Y_i - \bar{Y})^s \qquad (8.2)$$

Particularizando a expressão (8.2), teremos: -

$$m_{10} = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

$$m_{20} = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X})^2 = a_{20} - \bar{X}^2 = s_x^2$$

$$m_{02} = \frac{1}{N} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = a_{02} - \bar{Y}^2 = s_y^2$$

$$m_{11} = \frac{1}{N} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = a_{11} - \bar{X}\bar{Y} = s_{xy} \qquad (8.3)$$

Êste último momento - $m_{11} = s_{xy}$ - chamado especialmente "covariância", "momento-produto" ou "momento misto", derivado de síntese por indiciação e definindo uma singularidade descritiva, é muito importante e útil na análise.

Considerando outra vez um diagrama de dispersão cujos pontos estejam distribuídos aproximadamente em sentido linear, cumpre ajustar uma linha reta aos dados reais representados pelos pontos inseridos no gráfico cartesiano. Esta linha denominada tecnicamente reta de regressão de Y sôbre X se Y for a variável dependente, exige na determinação dos seus parâmetros o concurso da minimização quadrática, de modo que os valores teóricos da variável dependente calculados através da respectiva equação, discrepem o menos possível dos valores efetivos.

Êste requisito de mínimo afastamento dos valores reais da variável dependente a partir dos seus valores teóricos obtidos através da equação da linha reta, é atendido contanto que nós ajustemos a reta

com uma técnica tal que a soma dos quadrados daquêles desvios seja mínima.(28)

Ora, cada ponto correspondente a um par de valores (X_i, X_i) não ajustado exatamente, mantém uma distância d em relação à linha adaptada, que é o afastamento entre o dito ponto e a linha, tal qual vemos no esquema diagramático da figura 7.

A condição acima prevista de ser d uma distância menor possível, envolve a minimização de $\sum d^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$, ou seja, da função

$$E = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Desenvolvendo o segundo membro e anulando as derivadas parciais $\frac{dE}{da}$ e $\frac{dE}{db}$, encontraremos: -

$$\sum (Y_i - a - bX_i) = 0$$
$$\sum X_i (Y_i - a - bX_i) = 0$$

Efetuando a soma e dividindo todos os têrmos por N: -

$$\bar{Y} - a - b\bar{X} = 0$$

$$a_{11} - a\bar{X} - ba_{20} = 0$$

Estas são as equações normais compreensivas de um sistema de duas equações do 1º grau. Tirando o valor de a na primeira e substituindo-o na segunda, vem: -

---

(28) Assim como um promédio é escolhido de maneira que a soma dos quadrados dos afastamentos entre o seu valor e os dos restantes da série, seja um mínimo, também a soma dos quadrados dos desvios entre os valores reais da variável Y e os seus valores teóricos, deve ser a menor possível. Vide capítulo 5, págs. 40-43 para fins comparativos.

$$a_{11} - \bar{XY} + b\bar{X}^2 - bs_x^2 - b\bar{X}^2 = 0$$

Ou:

$$m_{11} - bs_x^2 = 0$$

Portanto: -

$$b_{yx} = \frac{m_{11}}{s_x^2} = \frac{s_{xy}}{s_x^2} \qquad (8.4)$$

Determinamos desta forma o parâmetro b em função da covariância de X e Y (s_{xy}) e da variância de X (s_x^2), cujo parâmetro neste caso especial denominado coeficiente de regressão de Y sôbre X correspondente ao coeficiente angular da reta, ordinàriamente é representado por b_{yx}. Afere êste coeficiente, em média, o acréscimo proporcional da variável dependente, decorrente do aumento unitário da variável independente, equivalendo assim à taxa de variação entre os dois fatôres correlatos. Além disso, o seu sinal indica a natureza da própria relação média, isto é, se a variável dependente aumenta ou diminui à medida que cresce a variável independente, em suma, se a regressão é positiva ou negativa. Fornecendo êste parâmetro um dos elementos da equação de regressão linear, teremos para o caso em que a equação é referida à média: -

$$y = b_{yx}\bar{x} \qquad (8.5)$$

Referente à origem O, a equação de regressão será: -

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X}) \qquad (8.6)$$

Efetuando o produto indicado no segundo membro e isolando Y no primeiro, ressulta: -

$$Y = b_{yx}X - b_{yx}\bar{X} + \bar{Y}$$

Então teremos: -

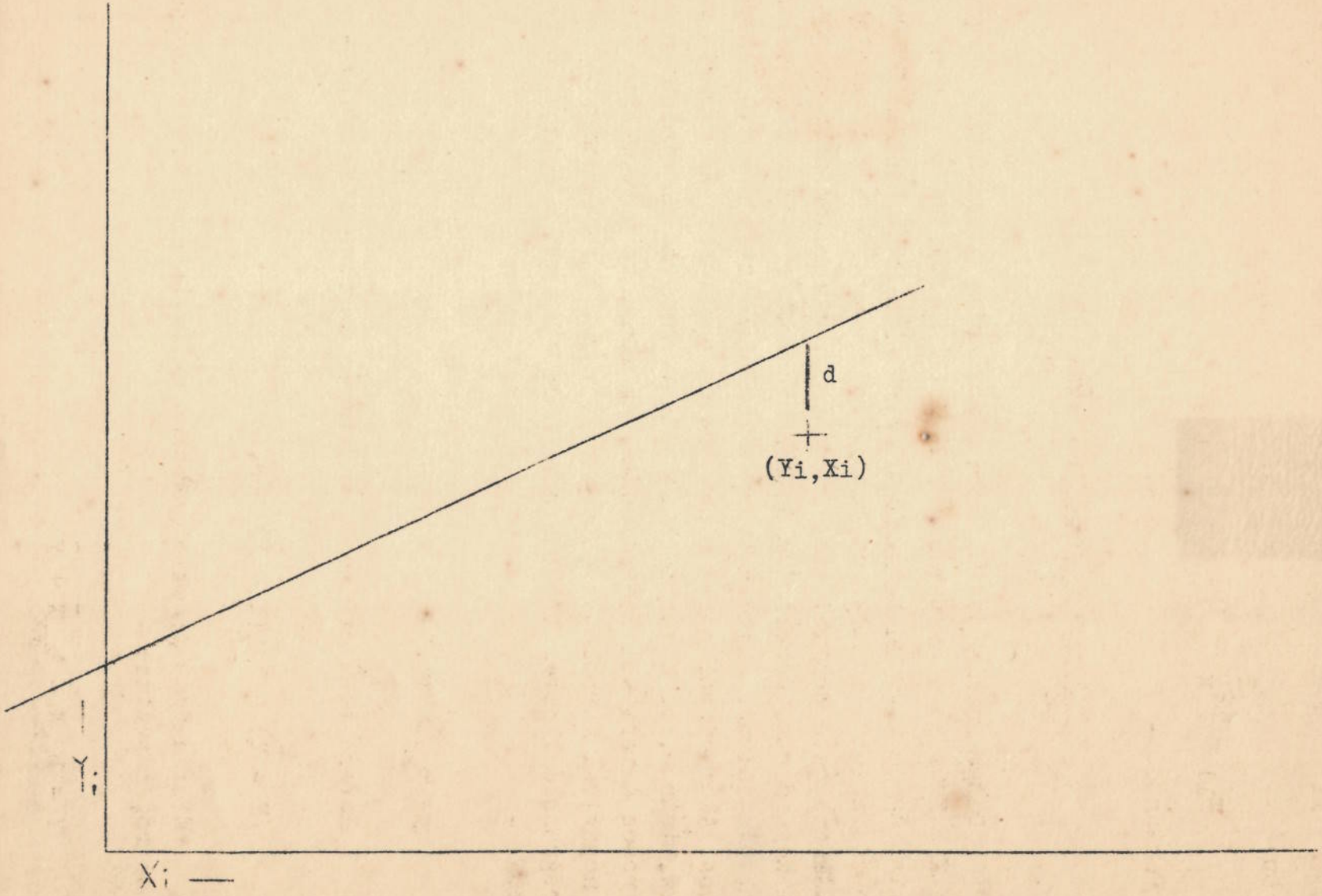
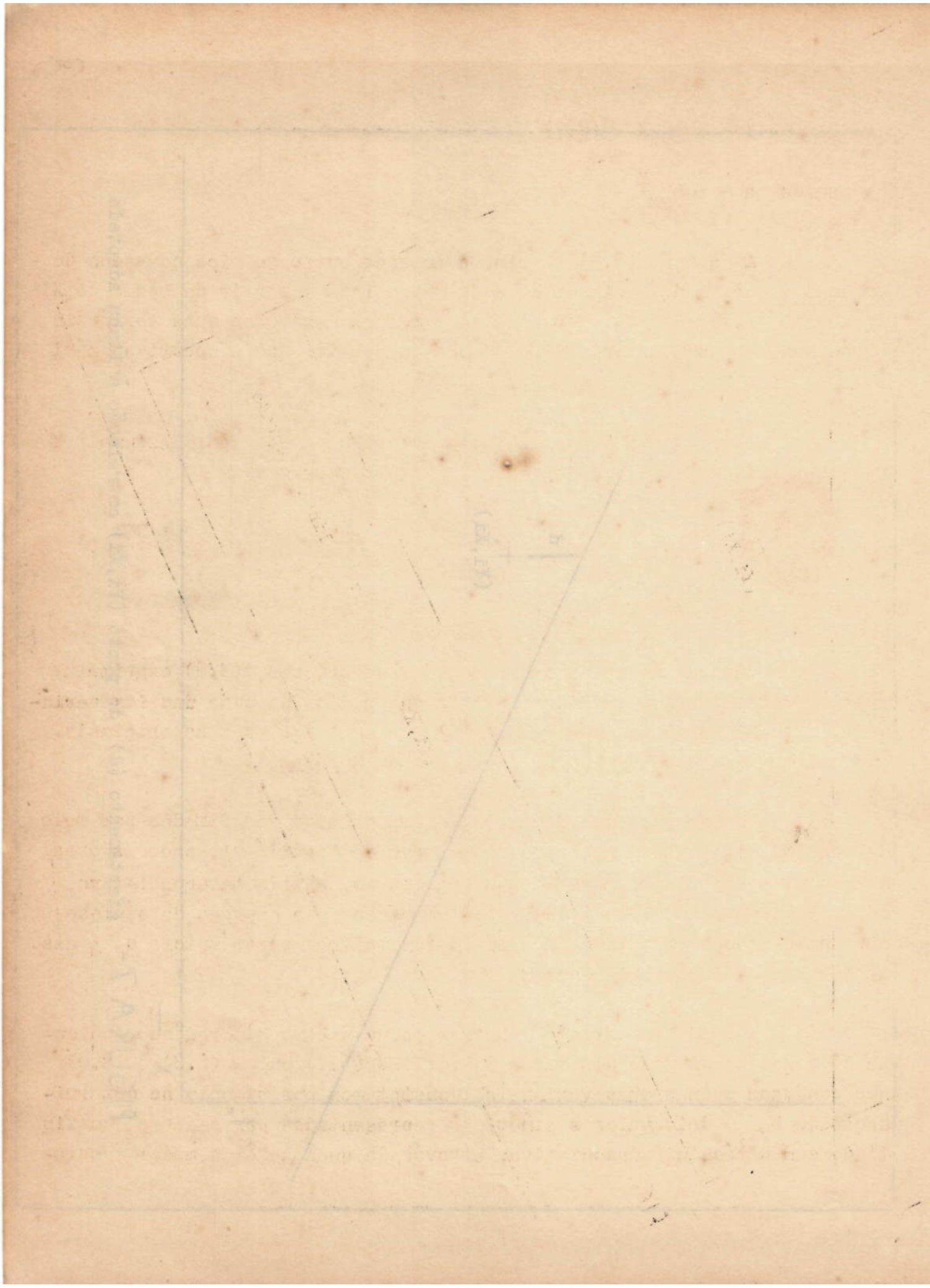


FIGURA 7- Afastamento (d) do ponto (Yi,Xi) em relação à linha adaptada



$$Y = a+b_{yx}X \quad (8.7)$$

sendo $a = \bar{Y}-b_{yx}\bar{X}$

A equação (8.5) exprime a relação entre desvios contados de um lado a partir da média de Y, e de outro lado a partir da média de X. A equação (8.7), por sua vez, indica a relação entre as duas variáveis expressas nas escalas primitivas e não em desvios das respectivas médias.

Invertendo a posição das variáveis na equação, fazendo Y variável independente e X dependente, estabeleceremos: -

$$x = b_{xy}y \quad (8.8) \quad \text{e}$$

$$X = a+b_{xy}Y \quad (8.9)$$

Anàlogamente às duas equações anteriores, (8.8) exprime a relação entre desvios tomados a partir das médias de cada uma das variáveis, e (8.9) refere-se aos valores da variável nas escalas originais. Nestas, b_{xy} é o coeficiente de regressão de X sôbre Y.

Estas relações de naturaza estocástica são fixadas por meio da análise estatística aplicada a um agregado bivariável, onde as duas componentes são dissociadas em observância ao sentido interpretativo, supondo porém esta dissociação, preliminarmente, o emprêgo da síntese por indiciação no que toca ao cálculo dos valores paramétricos de b nas mencionadas relações aleatórias.

Imediatamente ao penetrarmos no âmbito heterogêneo, calculando equações simples dos tipos de (8.5) e (8.7) ou de (8.8) e (8.9) que conjugam apenas duas variáveis, constatamos por exemplo na determinação de b_{yx}, dois valores sinóticos representados por s_{xy} e s_x^2 coexistindo com a técnica dissociativa, através da qual aquêles modêlos esto-

cásticos são assentados.

Começada assim a diferenciação do método ante uma amostra bivariável, encontram-se simultânea e necessàriamente síntese e análise, a primeira desenvolvida em modalidades quase sempre idênticas desde as elaborações iniciais da segunda, surgindo aí as mutações na contextura da metodologia. Suplantada desta forma a fase descritiva, tem o seu ponto de partida a evolução da interpretação estatística, ganhando esta em nitidez com o aumento do número de variáveis na amostra; as operações sintéticas repetidas em suas formas originais ocorrem concomitantemente à inovação dos meios analíticos.

As equações de regressão proporcionam prever aproximadamente as magnitudes prováveis da variável dependente com base nas magnitudes reais da variável independente, sendo por isso chamada também equação de estimativa, tornando-se a linha de regressão um excelente instrumento de análise teórico-científica quando através dela se procura explicar a mudança em uma das variáveis como efeito das alterações de outra variável que sôbre ela exerce influência. Nestas condições, a equação de regressão oferece no dizer de Wold, uma explicação hipotética da variável-efeito, porquanto se tenta estabelecer através dêste processo estatístico de regressão, um modêlo teórico identificável ao fenômeno analisado.(29)

De fato, conhecida a sucessão de valores que a variável independente assume no curso da observação, pode-se atribuir fàcilmente à variável dependente a escala de valores calculáveis por meio da equação de regressão, cujos valores teóricos tanto mais aproximados estão dos efetivos, quanto mais acurado houver sido o processo de ajustamento da reta e maior o número de elementos na amostra bivariável.(30)

---

(29) WOLD, HERMAN. Op. cit., pág. 30

(30) A dispersão entre os valores reais e teóricos da variável dependente, definida pela variância residual, serve de fundamento ao conceito de erro padrão de estimativa, o qual, derivado da idéia de correlação, representa-se por $s_{ya} = s_{yi}\sqrt{(1 - r)^2}$, Mede esta expressão o grau de estabilidade na distribuição dos valores de Y em tôrno da linha de regressão linear simples.

- CAPÍTULO IX -

CORRELAÇÃO BRUTA

Colocada a análise estatística em têrmos de regressão, onde o parâmetro b das respectivas equações deriva de uma instrumentação calculatória fundada em valores sinóticos, evidenciando de início que nos problemas de bivariação, a síntese por indicação persiste integrando em caráter acessório o processo analítico, prosseguiremos o estudo da interpretação com a questão da correlação, medida esta especìficamente estatística, imprescindível às análises acuradas executadas dentre a maioria das ciências. Grande vantagem do procedimento da correlação é que o seu cálculo, ao contrário do da regressão, propicia uma medida abstrata independente da unidade em que estão expressas as variáveis; faculta-se portanto ao pesquisador verificar o grau de relação entre as duas variáveis, assim como a proporção da variabilidade global de Y explicável pela equação de regressão, diante da concomitante variação nos valores da variável X.

Quanto maior porção da variabilidade total se torna explicada pela reta de regressão, refletindo diferenças de certo modo pequenas entre os valores teóricos e os correspondentes dados empíricos, tanto mais sensível será a relação entre as duas variáveis. Ajustada a reta aos vários pontos do diagrama de dispersão e estimadas as magnitudes prováveis da variável dependente, deparamos com desvios em tríplice aspecto: - 1) desvios entre os valores reais da variável dependente e a respectiva média; 2) desvios entre os valores calculados da variável dependente e a média dos valores reais da mesma variável; 3) desvios entre os valores reais e os estimados da variável dependente. Daí dimanam três diferentes espécies de variâncias: - variância total ($s^2_{y_i}$) e variância explicada ($s^2_{y_c}$) e variância residual ou não explicada ($s^2_{y_a}$) respectivamente.(31)

---

(31) CAMARA, LOURIVAL. Correlação. Apostilha mimeografada. Ano 1954, Série I-B, Rio de Janeiro, Escola Brasileira de Estatística, pag. 7.

Conseqüentemente, teremos: -

$$s_{y_i}^2 = \frac{\sum (Y_i - \bar{Y})^2}{N} \qquad (9.1)$$

$$s_{y_c}^2 = \frac{\sum (Y_c - \bar{Y})^2}{N} \qquad (9.2)$$

$$s_{y_a}^2 = \frac{\sum (Y_i - Y_c)^2}{N} \qquad (9.3)$$

Em duas séries relativas à mesma variável dependente, uma delas composta dos valores efetivamente observados (Y_i) e a outra integrada pelos valores estimados através da equação de regressão (Y_c), segue-se que $\bar{Y}_i = \bar{Y}_c = \bar{Y}$ porquanto $\sum Y_i = \sum Y_c$.

Em face da identidade: -

$$(Y_i - \bar{Y}) = (Y_c - \bar{Y}) + (Y_i - Y_c)$$

deduzimos que: -

$$\frac{\sum (Y_i - \bar{Y})^2}{N} = \frac{\sum (Y_c - \bar{Y})^2}{N} + \frac{\sum (Y_i - Y_c)^2}{N}$$

Ou

$$s_{y_i}^2 = s_{y_c}^2 + s_{y_a}^2 \qquad (9.4)$$

A variância total ($s_{y_i}^2$) tem, portanto, duas componentes: - $s_{y_c}^2$ (variância explicada) e $s_{y_a}^2$ (variância residual ou não explicada).

Visto que a razão entre a variância explicada e a variância total mede o grau em que a última se torna explicada pela equação de regressão, demonstrando forte ou fraca relação entre as variáveis

conforme seja elevada ou reduzida a proporção de variância explicada para a variância total - $\frac{s^2_{y_c}}{s^2_{y_i}}$., denomina-se coeficiente de determinação linear simples ou bruto êsse quociente muito valioso no estudo da correlação estatística. E a raiz quadrada do dito coeficiente equivale ao coeficiente de correlação linear simples ou bruto. As expressões algébricas destas constantes, diante do exposto, serão: -

$$r^2_{yx} = \frac{s^2_{y_c}}{s^2_{y_i}} \quad (9.5) \text{ - coeficiente de determinação}$$

$$r_{yx} = \sqrt{\frac{s^2_{y_c}}{s^2_{y_i}}} \quad (9.6) \text{ - coeficiente de correlação.}$$

Ezequiel[32] estabelece a distinção entre o coeficiente de determinação e o de correlação, aduzindo que o primeiro mede a proporção de todos os elementos de variância existentes em Y também presentes em X, ao passo que o segundo mede a proporção da variância em uma das variáveis (Y) quando associada à variação da outra (X), aferindo destarte a importância relativa da concomitância dos dois fatôres. Enquanto o coeficiente de correlação (r) indica simplesmente o grau de relação entre as duas variáveis, o coeficiente de determinação (r^2) mede em têrmos de variância, a proporção de variabilidade na variável dependente, devida às mudanças da variável independente.

Como da expressão (9.4) podemos deduzir que $s^2_{y_c} = s^2_{y_i} - s^2_{y_a}$, substituindo esta diferença em (9.5), acharemos: -

---

(32) EZEKIEL, MORDECAI. Op. cit., págs. 138-139.-

$$r^2_{yx} = \frac{s^2_{y_i} - s^2_{y_a}}{s^2_{y_i}} = 1 - \frac{s^2_{y_a}}{s^2_{y_i}}$$

Com a introdução que fizemos da expressão da variância não explicada, as fórmulas para o coeficiente de determinação e de correlação passarão a ser respectivamente:-

$$r^2_{yx} = 1 - \frac{s^2_{y_a}}{s^2_{y_i}} \qquad (9.7)$$

$$r_{yx} = \sqrt{1 - \frac{s^2_{y_a}}{s^2_{y_i}}} \qquad (9.8)$$

Sendo r nulo, não há correlação linear entre as duas variáveis. Êste coeficiente pode variar entre - 1 e + 1, verificando-se perfeita correlação linear inversa ou negativa quando r é igual a -1 e perfeita correlação linear direta ou positiva no caso de ser r igual a + 1. Semelhantemente, há correlação linear inversa se $r < 1$ e correlação linear direta se $r > 1$. O sinal de r é o mesmo de b_{yx} ou b_{xy} na equação de regressão. Em virtude destas propriedades, Cramer considera o coeficiente de correlação bruto como uma medida do grau de linearidade da distribuição.(33)

A vantagem de exprimir-se a correlação em função das variâncias residual e total, consiste na obtenção de um resultado tal que derivado da fórmula (9.8) cresce em razão do maior grau de relação entre as duas variáveis, as quais tanto mais associadas estão quanto maior porção da variabilidade total em Y estiver interpretada por meio da equação de regressão.

---

(33) CRAMER, HARALD. Metodos Matematicos de Estadistica. Madrid, Aguilar, S.A. de Ediciones, 1953, pág. 320.

Podemos também dar outra forma ao coeficiente de correlação a partir das expressões (9.7) e (9.3). Mudada a origem para o ponto das médias, ou seja, tomando-se as duas variáveis em têrmos de desvios ao redor das respectivas médias ($\overline{Y}$ e $\overline{X}$) conforme expresso em (8.5), transformar-se-á a (9.3) em: -

$$s^2_{y_a} = \frac{\Sigma (y_i - b_{yx}x_i)^2}{N}$$

Elevando ao quadrado o numerador e efetuando a soma indicada, teremos: -

$$s^2_{y_a} = \frac{\Sigma y_i^2 - 2b_{yx} \Sigma x_i y_i + b^2_{yx} \Sigma x_i^2}{N}$$

cuja equação se reduzirá a: -

$$s^2_{y_a} = s^2_y - 2b_{yx}s_{xy} + b^2_{yx} s^2_x \qquad (9.9)$$

Como $b_{yx} = \frac{s_{xy}}{s^2_x}$ de acôrdo com a (8.4), resulta após substituição em (9.9): -

$$s^2_{y_a} = s^2_y - \frac{s^2_{xy}}{s^2_x} \qquad (9.10)$$

Substituindo na expressão (9.7) o valor de $s^2_{y_a}$ conforme a (9.10), encontraremos: -

$$r^2_{yx} = 1 - \frac{s^2_y - \frac{s^2_{xy}}{s^2_x}}{s^2_y}$$

Por conseguinte: -

$$r^2_{yx} = \frac{s^2_{xy}}{s^2_x \; s^2_y} \quad (9.11) \text{ e}$$

$$r_{yx} = \frac{s_{xy}}{s_x \; s_y} \quad (9.12)$$

Exprimimos assim o coeficiente de correlação em função da covariância e dos desvios padrões das duas variáveis X e Y.

Multiplicando e dividindo a expressão (9.12) por s_x, podemos dar uma outra forma a r_{yx}.

$$r_{yx} = \frac{s_{xy} \; s_x}{s_x \; s_y \; s_x} = \frac{s_{xy} \; s_x}{s^2_x \; s_y}$$

Como $\frac{s_{xy}}{s^2_x} = b_{yx}$, resulta que: -

$$r_{yx} = b_{yx} \; \frac{s_x}{sy} \quad (9.13)$$

Desta forma, o coeficiente de correlação está determinado em função do coeficiente de regressão e dos desvios padrões das duas variáveis. Permite esta fórmula o cálculo de uma nova equação para b_{yx}, da qual participa r_{yx}. Operando a transformação necessária, temos:-

$$b_{yx} = \frac{r_{yx}}{\frac{s_x}{s_y}} = r_{yx} \; \frac{s_y}{s_x}$$

Portanto: -

$$b_{yx} = r_{yx} \frac{s_y}{s_x} \qquad (9.14)$$

Anàlogamente, $b_{xy} = r_{xy} \frac{s_x}{s_y}$ (9.15)

Consideremos as expressões (9.14) e (9.15) que definem os coeficientes de regressão linear em função do coeficiente de correlação. Multiplicando-as entre si, acharemos: -

$$b_{yx}\, b_{xy} = r \frac{s_y}{s_x} \cdot r \frac{s_x}{s_y}$$

Segue-se então que: -

$$b_{yx}\, b_{xy} = r^2 \qquad (9.16) \quad \text{ou}$$

$$\sqrt{b_{yx}\, b_{xy}} = r_{yx} \qquad (9.17)$$

Isto é, a média geométrica dos coeficientes de regressão é igual ao coeficiente de correlação.

As expressões algébricas fundamentais identificadas aos coeficientes de determinação e correlação, (9.7), (9. 8), (9.11) e (9.12), constituindo excelentes instrumentos de análise, encerram as idéias básicas da síntese por indiciação, pois na dedução de r^2 e subseqüentemente de r, integram-se valores sinóticos conexos a cada um dos agregados, referidos tanto à contração quanto à distensão dêstes. A descrição continua, portanto, preenchendo a sua função complementar, conquanto indispensável, no curso da interpretação, nesta segunda fase de análise bivariável distinta da primeira (regressão) por motivo da introdução do estudo concernente ao grau de relação e da variabilidade explicavel, cujas facetas estatísticas são mensuradas por aquêles coeficientes.

- C A P Í T U L O X -

REGRESSÃO MÚLTIPLA

Nos casos de relação linear múltipla, o próprio qualificativo está a indicar uma condição mais complexa do que a precedente. Trata-se agora de três ou mais variáveis aleatórias na estruturação do fenômeno, em cuja circunstância as amostras destacadas de universos ainda mais heterogêneos, exigem o critério pluridimensional no tratamento de questões dessa natureza. Persistem as operações extensivas, objetivando em primeiro lugar a dedução de uma relação estocástica entre a variável dependente e as demais, sob a forma de equação de regressão, e posteriormente o cálculo dos coeficientes de determinação e correlação.

Tomadas amostras a três variáveis, defrontamos com situações específicas ao ponto de tornar possível reafirmarmos o estreito vínculo da síntese à análise, na órbita da maturação do método estatístico; os problemas atinentes à regressão e correlação múltipla, requerem uma maior dosagem sintética não obstante o ser nas mesmas modalidades anteriores, ao lado de análises mais penetrantes.

O primeiro aspecto da problemática envolve a estimativa dos valores de uma das variáveis, com base nos valores efetivos das outras variáveis, por meio da equação de regressão múltipla. Adotaremos uma terminologia semelhante àquela utilizada anteriormente nos casos de regressão linear simples, quanto à denominação de variável dependente àquela que se julga estar sob influência das demais e variáveis independentes tôdas as outras também observadas e estudadas nos limites da amostra.

Nesta exposição designaremos sempre a variável dependente por X_1 e as outras por X_2, X_3 etc., havendo por conseguinte um fator X_1 relacionado a outros fatôres X_2, X_3.....X_n, êstes últimos independentes entre si, de forma que qualquer um dêles possa assumir indistintamente um valor para determinadas combinações de valores dos demais.(34)

---

(34) ALVAREZ, M. GARCIA e OREJANA, J. AYUSO. Op. cit., pág. 443.

Considerando o caso mais simples de relação linear múltipla, em que há apenas três variáveis, uma dependente e duas independentes, a respectiva equação de regressão é: -

$$X_1 = a_{1,23} + b_{12,3}X_2 + b_{13,2}X_3 \qquad (10.1)$$

Como contra-partida da linha de regressão nas relações lineares a duas variáveis, a equação (10.1) define a superfície de regressão de X_1 com respeito a X_2 e X_3, permitindo-se calcular o valor mais provável de X_1 para determinada posição dos valores das demais variáveis.(35)

Considerando desvios dos valores de X_1, X_2 e X_3 em relação às respectivas médias aritméticas, expressos quantitativamente por x_1, x_2 e x_3, a equação (10.1) passa à forma: -

$$x_1 = b_{12,3}x_2 + b_{13,2}x_3 \qquad (10.2)$$

Esta mudança de origem é fàcilmente compreensível diante das identidades que exprimem os desvios em função dos próprios valores da variável, como sejam: - $x_1 = X_1 - \overline{X}_1$; $x_2 = X_2 - \overline{X}_2$; $x_3 = X_3 - \overline{X}_3$. Substituindo-os na equação (10.2) resultará: -

$$X_1 - \overline{X}_1 = b_{12,3}(X_2 - \overline{X}_2) + b_{13,2}(X_3 - \overline{X}_3)$$

Efetuando as multiplicações indicadas e as necessárias transposições, teremos: -

$$X_1 = b_{12,3}X_2 + b_{13,2}X_3 + (\overline{X}_1 - b_{12,3}\overline{X}_2 - b_{13,2}\overline{X}_3)$$

Então: -

$$X_1 = a_{1,23} + b_{12,3}X_2 + b_{13,2}X_3, \text{ sendo}$$

$$a_{1,23} = \overline{X}_1 - b_{12,3}\overline{X}_2 - b_{13,2}\overline{X}_3$$

---

(35) Loc. cit.

Nas equações (10.1) e (10.2), o parâmetro $b_{12,3}$ é o coeficiente de regressão da variável dependente X_1 sôbre a variável independente X_2, mantida constante a outra variável independente X_3. Anàlogamente, o parâmetro $b_{13,2}$ equivale à regressão de X_1 sôbre X_3, sem alteração nos valores de X_2. São especìficamente denominados coeficientes parciais ou líquidos de estimativa em correspondência à condição "coeteris paribus" que rege as previsões no campo das ciências sociais.(36)

Na determinação dos valores paramétricos de $b_{12,3}$ e $b_{13,2}$, começamos a admitir que, sendo a equação de regressão múltipla um instrumento destinado a estimar os valores da variável dependente X_1 em função dos valores das variáveis independentes X_2 X_3, é indispensável estabelecer expressões tais que as magnitudes de X_1 calculadas através da equação, estejam aproximadas tanto quanto possível dos dados efetivamente observados pelo pesquisador.

Por isso, a máxima proximidade ou recìprocamente a mínima dispersão entre os valores reais e teóricos da variável dependente X_1, subordina-se à condição de ser $\sum d^2 = \sum_{i=1}^{n} (x_1 - b_{12,3}x_2 - b_{13,2}x_3)^2$ um mínimo, o que implica em minimizar a função:-

$$E = \sum_{i=1}^{n} (x_1 - b_{12,3}x_2 - b_{13,2}x_3)^2$$

Desenvolvendo o segundo membro e anulando as derivadas parciais $\frac{dE}{db_{12,3}}$ e $\frac{dE}{db_{13,2}}$, acharemos:-

$$\sum (b_{12,3}x_2^2 + b_{13,2}x_2x_3 - x_1x_2) = 0$$

$$\sum (b_{12,3}x_2x_3 + b_{13,2}x_3 - x_1x_3) = 0$$

---

(36) CROXTON, FREDERICK E. e COWDEN, DUDLEY J. Estatística Geral e Aplicada. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística. 1952. pág. 847.

Efetuando a soma e as devidas transposições, resulta o seguinte sistema de equações normais: -

$$\Sigma x_1 x_2 = b_{12,3} \Sigma x_2^2 + b_{13,2} \Sigma x_3 x_2$$

$$\Sigma x_1 x_3 = b_{12,3} \Sigma x_2 x_3 + b_{13,2} \Sigma x_3^2$$

Dividindo-se todos os seus têrmos por N, estas equações transformam-se em: -

$$s_{12} = b_{12,3} s_2^2 + b_{13,2} s_{32} \qquad (10.3)$$

$$s_{13} = b_{12,3} s_{23} + b_{13,2} s_3^2 \qquad (10.4)$$

Das equações (10.3) e (10.4) deduzimos respectivamente: -

$$b_{12,3} = \frac{s_{12} - b_{13,2} s_{32}}{s_2^2} \quad \text{ou}$$

$$b_{12,3} = b_{12} - b_{13,2} b_{32} \qquad (10.5)$$

$$b_{13,2} = \frac{s_{13} - b_{12,3} s_{23}}{s_3^2} \quad \text{ou}$$

$$b_{13,2} = b_{13} - b_{12,3} b_{23} \qquad (10.6)$$

Substituindo na expressão (10.5) o valor de $b_{13,2}$ segundo a (10.6): -

$$b_{12,3} = b_{12} - b_{32} b_{13} + b_{32} b_{12,3} b_{23}$$

Segue-se então que: -

$$b_{12,3} - b_{32}b_{12,3}b_{23} = b_{12} - b_{32}b_{13}$$

$$b_{12,3}(1 - b_{32}b_{23}) = b_{12} - b_{32}b_{13}$$

$$b_{12,3} = \frac{b_{12} - b_{32}b_{13}}{(1 - b_{32}b_{23})} \qquad (10.7)$$

De maneira análoga, obtém-se para o outro coeficiente: -

$$b_{13,2} = \frac{b_{13} - b_{23}b_{12}}{(1 - b_{23}b_{32})} \qquad (10.8)$$

As expressões (10.7) e (10.8) permitem o cálculo dos coeficientes líquidos da equação de regressão múltipla a três variáveis, mediante os coeficientes brutos de uma equação a duas variáveis. São aspectos especiais de uma expressão genérica pela qual se pode determinar qualquer coeficiente de regressão de ordem p em função das regressões de ordem p-1. No caso vertente em que há uma variável dependente e duas independentes, as fórmulas acima comportam cálculos preliminares de valores sinóticos em escala mais ampla do que no caso de apenas duas variáveis, pois a determinação de cada um dos coeficientes parciais $b_{12,3}$ e $b_{13,2}$ requer necessàriamente o cálculo de três covariâncias (s_{13}, s_{12} e s_{32}) e duas variâncias (s_3^2 e s_2^2), através dos coeficientes brutos b_{12}, b_{13}, b_{23} e b_{32}.

Estivessemos porventura analisando regressões múltiplas com mais de três variáveis, sempre uma delas dependente e as demais independentes entre si, suceder-se-iam as covariâncias e variâncias, cujos valores sinóticos, elementos integrantes que são dos coeficientes de ordem inferior, apareceriam em maior quantidade à medida que fôssemos acrescentando variáveis à amostra. Havendo n variáveis, a expressão geral do coeficiente de regressão -

$$b_{12,34\ldots n,} = \frac{b_{12,34\ldots(n-1)} - b_{1n,34\ldots(n-1)}b_{n2,34\ldots(n-1)}}{1 - b_{2n,34\ldots(n-1)}b_{n2,34\ldots(n-1)}}$$

dá uma idéia da extrema difusão da síntese por indiciação na análise de regressão, quando se considera a contextura complexa de uma amostra composta de muitas variáveis. Confirma-se assim a coexistência sintético-analítica, de modo a não deixar dúvida quanto à manutenção das formas descritivas originais entremeadas com os aspectos modificados da interpretação.

Os coeficientes $b_{12,3}$ e $b_{13,2}$ da equação de regressão múltipla, têm função análoga aos seus congêneres na equação de regressão simples: - $b_{12,3}$ mede especìficamente, em têrmos médios, a proporção de variação em X_1 associada à alteração unitária em X_2, sem mudanças em X_3, correspondendo à taxa de variação entre a variável dependente X_1 e a variável independente X_2; $b_{13,2}$ afere, em média, o "quantum" de variabilidade que ocorre em X_1 por unidade de variação em X_3, não se verificando concomitantemente modificações em X_2, equivalendo à taxa de variação entre a variável dependente X_1 e a variável independente X_3. Os sinais de $b_{12,3}$ e $b_{13,2}$ indicam a natureza da relação média entre as variáveis.

- CAPÍTULO XI -

CORRELAÇÃO MÚLTIPLA

Calculada a regressão com mais de duas variáveis, sucede o cômputo dos coeficientes de determinação e correlação. O coeficiente de determinação múltipla afere o grau de variabilidade global em X_1 na amostra multivariável, explicável pela equação de regressão, ante a mudança das demais variáveis X_2, X_3,....X_n conjugadas; o coeficiente de correlação múltipla, por seu turno, indica o grau de relação entre a variável dependente única (X_1) e certo número de variáveis independentes que se reunem (X_2, X_3,..X_n), na extensão em que as variações da primeira se entrosam com a variação agregada das demais.(37) Nestas condições, tôdas as variáveis independentes constituem uma espécie de série independente única, identificando-se então o coeficiente de correlação múltipla a uma medida da relação entre a variável dependente e a série independente, semelhantemente ao coeficiente de correlação simples que liga sòmente duas variáveis.(38)

A fórmula do coeficiente de correlação múltipla, o qual é usualmente simbolizado por R, tem origem na consideração dos desvios entre os valores reais e os teóricos X_1, cujos desvios denominaremos ε_i

---

(37) Em tais casos, o efeito de uma das variáveis independentes sôbre a dependente, é constante para todos os outros valores que as demais variáveis independentes assumem, amprestando assim um caráter aditivo aos resultados derivados da correlação múltipla. Quando a natureza e o grau da relação entre uma variável dependente e uma independente, diferem segundo a intensidade de outra variável independente, verifica-se correlação conjunta, cujo tratamento excluimos do nosso estudo.

(38) MILLS, FREDERICK CECIL. Métodos Estatísticos Aplicados à Economia e aos Negócios. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística, 1952, pág, 500.

$$\varepsilon_i = X_{1c} - X_{1i} = a_{1,23} + b_{12,3}X_{2i} + b_{13,2}X_{3i} + \ldots\ldots\ldots - X_{1i} \quad (11.1)$$

Multiplicando todos os têrmos por ε_i e somando, temos:

$$\sum \varepsilon_i^2 = a_{1,23} \sum \varepsilon_i + b_{12,3} \sum X_{2i}\varepsilon_i + b_{13,2} \sum X_{3i}\varepsilon_i + \ldots\ldots\ldots - \sum X_{1i}\varepsilon_i$$

Como $\sum \varepsilon_i = 0$, $\sum X_{2i}\varepsilon_i = 0$, $\sum X_{3i}\varepsilon_i = 0$

$$\sum \varepsilon_i^2 = - \sum X_{1i}\varepsilon_i \quad (11.2)$$

Multiplicando todos oe têrmos da (11.1) por X_{1i} e somando, encontramos: -

$$\sum X_{1i}\varepsilon_i = a_{1,23} \sum X_{1i} + b_{12,3} \sum X_{1i}X_{2i} + b_{13,2} \sum X_{1i}X_{3i} + \ldots\ldots - \sum X_{1i}^2 \quad (11.3)$$

Substituindo na (11.2) o valor de $\sum X_{1i}\varepsilon_i$ segundo a (11.3): -

$$\sum \varepsilon_i^2 = \sum X_{1i}^2 - a_{1,23} \sum X_{1i} - b_{12,3} \sum X_{1i}X_{2i} - b_{13,2} \sum X_{1i}X_{3i} + \ldots\ldots$$

Dividindo por N: -

$$\frac{\sum \varepsilon_i^2}{N} = \frac{\sum X_{1i}^2 - a_{1,23} \sum X_{1i} - b_{12,3} \sum X_{1i}X_{2i} - b_{13,2} \sum X_{1i}X_{3i} \ldots\ldots}{N}$$

(11.4)

Transladando a origem ao ponto onde estão situadas as médias ($\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$), isto é, exprimindo a (11.4) em têrmos de desvios ao redor das médias de três variáveis, a variância residual ou não explicada de X_1 (s_{1a}^2), será: -

$$s_1^2 a = \frac{\sum x_{1i}^2 - b_{12,3} \sum x_{1i}x_{2i} - b_{13,2} \sum x_{1i}x_{3i}}{N}$$

Portanto: -

$$s_1^2 a = s_1^2 - b_{12,3}\, s_{12} - b_{13,2}\, s_{13} \quad (11.5)$$

De modo geral, $R_{1,23\ldots n} = \sqrt{1 - \frac{s_{1a}^2}{s_1^2}}$, anàlogamente ao caso da correlação linear simples, mede a correlação em função das variâncias residual e total, entre o fator configurado estatisticamente variável X_1 e os fatôres conjugados subentendidos pelas variáveis X_2, $X_3 \ldots X_n$. Substituindo na referida fórmula, s_{1a}^2 por seu valor conforme a (11,5), consoante ao aspecto de três variáveis, temos: -

$$R_{1,23} = \sqrt{1 - \frac{s_1^2 - b_{12,3}s_{12} - b_{13,2}s_{13}}{s_1^2}} \quad \text{ou}$$

$$R_{1,23} = \sqrt{\frac{b_{12,3}\, s_{12} + b_{13,2}\, s_{13}}{s_1^2}} \quad (11.6)$$

R não tem sinal porque a relação pode ser positiva com algumas variáveis independentes e negativas com outras.

Neste procedimento de correlação a três variáveis, comparativamente àquele usado para igual fim em amostra bivariável, afigura-se-nos mais sensível a análise em razão da própria repetição de formas sintéticas expressas por momentos-produtos e variâncias. É certo que no problema de correlação em amostra bivariável, a expressão - $\frac{s_{xy}}{s_x \cdot s_y}$ - não exige senão o cálculo de uma covariância e de dois desvios padrões; no tocante à correlação em amostra trivariável, porém, a diferenciação da análise está evidenciada na expressão - $R_{1,23} = \sqrt{\frac{b_{12,3}s_{12} + b_{13,2}s_{13}}{s_1^2}}$ - para cuja aplicação se faz mister a determinação prévia não só de dois

momentos-produtos diretamente (s_{12} e s_{13}), mas também do cálculo dêstes mesmos momentos, da covariância s_{23} e das necessárias variâncias, através do cômputo dos parâmetros brutos que integram as fórmulas dos parâmetros parciais $b_{12,3}$ e $b_{13,2}$, os quais entram na composição de - $R_{1,23}$.

A intensificação do emprêgo das formas sintéticas primitivas, conforme mostramos, proporciona só por sí, a diferenciação da análise, fazendo-nos assim crer que a repetição dos valores descritivos conduz a interpretações cada vez mais distintas, contanto que tenhamos amostras com número crescente de variáveis. A interferência de uma grande dosagem descritiva na estruturação de sensíveis interpretações, é percebida no processo geral de avaliação do coeficiente de correlação múltipla para qualquer número de variáveis, por meio da equação abaixo, onde a manutenção das formas sintéticas anteriores em operações consecutivas, gera as próprias condições de aperfeiçoamento da análise: -

$$R_{1,23\ldots n} = \sqrt{\frac{b_{12,34\ldots n}s_{12} + b_{13,24\cdots n}s_{13} + \ldots + b_{1n,23\ldots(n-1)}s_{1n}}{s_1^2}}$$

- CAPÍTULO XII -

CORRELAÇÃO PARCIAL

A discussão da variação múltipla invoca o problema das relações parciais ou líquidas, pois enquanto se dispõem de observações acumuladas dos efeitos da ação conjugada e constante de n fatôres ..... ($X_1, X_2, X_3 \ldots X_n$), é preciso conhecer muitas vêzes a dependencia exclusivamente entre dois daquêles fatôres, sendo para tal fim anuladas as influências motivadas pelos demais.

É questão idêntica à correlação simples em amostra bivariável, exceto no tocante à neutralização que se faz da influência de fatôres outros que agem decisiva e inevitàvelmente sôbre aquêle que o pesquisador tratava no estudo da correlação bruta, como se estivesse subordinado tão sòmente a um único agente específico.

Ora, quando a observação e o cálculo não confinados a dois - fatôres - X_1 e X_2 - sem a preocupação de remover a influência de outras - componentes - X_3, $X_4 \ldots X_n$ - que necessàriamente atuam, as constantes estabelecidas para medir a taxa de variação e o grau de relação entre aquelas duas variáveis aparentes, resultam de certo modo distorcidas, por se haver omitido os meios de avaliar e anular os efeitos da influência de alguns possiveis co-fatôres.

O procedimento linear simples sofre limitações, face à inadequacidade da sua técnica em neutralizaro influxo dos co-fatôres, os quais tanto quanto os dois analisados, integram o fenômeno global. Por isso, no processo da correlação parcial, conservam-se inalteradas tôdas aquelas variáveis cujos efeitos não desejamos que pesem nas conclusões da análise destinada a aferir o grau de relação particularmente à variação de dois elementos.

Tornam-se, por conseguinte, mais eficientes, com a introdução dos conceitos e técnicas das relações líquidas, os métodos de análi

se teóricase prática, oriundos dos processos básicos da correlação bruta. Desta maneira, são sentidas as novas formas de maturação do método estatístico, porquanto os princípios que presidem à determinação dos coeficientes líquidos, representam um ponderável avanço comparativamente às limitações da análise de linearidade simples. As restrições interpostas aos resultados identificados às condições de correlação linear bruta, por não terem sido compensadas na amostra bivariável os efeitos de outros fatôres integrantes do fenômeno, recomendam a adoção da técnica pertinente à correlação líquida ou parcial, que revela as circunstâncias efetivas de relação entre duas variáveis, sem que sôbre estas se exerça a influência de um terceiro fator.

Em tais casos de relações parciais ou líquidas, é mais dilatado o campo de repetição da síntese por indiciação, em correspondência à maior penetração da análise. O sistema binário síntese-análise reunindo descrição e interpretação em verdadeira osmose, comprova diante da sua aplicação ao problema da linearidade líquida, uma contingência natural de manifesta evolução do método estatístico.

Em geral, variando agregadamente X_1, X_2, X_3,.....X_n, estamos aptos a determinar especialmente a correlação entre X_1 e X_2 devidamente resguardada das mudanças verificadas em X_3, X_4 ....X_n; pode-se igualmente avaliar a correlação entre X_2 e X_3 em particular, fora da influência a que está também a primeira subordinada, proveniente das variações de X_1,X_4.....X_n. Em qulquer uma destas condições, são eliminadas as variações devidas a outros agentes que podem distorcer a relação verdadeiramente existente entre os dois considerados na análise. Obedecendo a êste critério, o cientista se protege contra os azares de resultados pouco precisos, pois suas conclusões não serão afetadas por causas que devem permanecer controladas.

O coeficiente de correlação linear ou líquida constitui a medida da influência de cada uma das variáveis independentes - X_2, X_3, ...X_n - sôbre a variável dependente X_1, aferindo assim a importância relativa das variáveis independentes na integração do fenômeno complexo; mostra a relação existente entre o fator considerado efeito e um dos co-agentes, não se ressentindo por isso o resultado apresentado pelo

pesquisador, das influências de outros fatôres componentes do fenômeno.

Enquanto no procedimento de correlação múltipla, o objetivo é estabelecer uma medida da importância de tôdas as variáveis independentes combinadas, na análise da correlação parcial se intenta medir a importância de cada uma das variáveis separadamente, para o que se eliminam as alterações associadas às variàveis independentes remanescentes. (39)

A obtenção de uma constante apta a definir esta relação de carater líquido, ou alternativamente de um valor conexo que possa exprimir a proporção de variação na variável dependente X_1 explicável pela concomitante variação de uma das variáveis independentes - X_2, X_3 ou X_n, subordina-se ao princípio de que aquela parte de variabilidade do fator dependente não explicada pela ação dos fatores independentes, pode ser interpretada mediante a inclusão de um novo fator. (40)

Sendo, por exemplo, $(1 - r^2_{12})$, a variabilidade de X_1 não explicada por uma equação de regressão linear simples, e $(1 - R^2_{1,23})$ a variabilidade de X_1 não explicada por uma equação de regressão múltipla, em cuja circunstância houve interferência da variável X_3, deduziremos que relativamente à primitiva variabilidade não interpretada $(1 - r^2_{12})$, a parte explicada pela nova variável X_3 é dada pelo quociente:-

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

Após simplificação, transformamos esta expressão em:

$$r^2_{13,2} = 1 - \frac{1 - R^2_{1,23}}{1 - r^2_{12}} \quad (12.1)$$

indicando a porção de variabilidade que, embora não explicada pela simples conjunção dos fatôres X_1 e X_2, passa a ser interpretada depois de

---

(39) EZEKIEL, MORDECAI, Op. cit., pág. 213
(40) EZEKIEL, MORDECAI. Op. cit., pág. 214

incluído o fator X_3. Mede, portanto, a influência da variável independente X_3 sôbre a variável dependente X_1, com os efeitos de X_2 eliminados.

Complementarmente, $r_{13,2} = \sqrt{1 - \frac{1 - R^2_{1,23}}{1 - r^2_{12}}}$ (12.2) supre o coeficiente de correlação parcial entre X_1 e X_3, aferindo o grau de relação entre estas duas variáveis, mantida constante a influência de X_2.

Por outro lado, $r^2_{12,3} = 1 - \frac{1 - R^2_{1,23}}{1 - r^2_{13}}$ (12.3) e

$r_{12,3} = \sqrt{1 - \frac{1 - R^2_{1,23}}{1 - r^2_{13}}}$ (12.4) servem à avaliação dos efeitos de X_2 diretamente sôbre X_1 e o grau de relação entre estas duas variáveis, mantida constante a influência do co-fatôr X_3.

O aspecto que reveste a constante expressiva da correlação linear parcial entre X_1 e X_2, havendo três variáveis, em confronto com a feição da sua correspondente no caso de relação retilínea simples, mostra claramente ser a primeira, resultante de análises mais penetrantes através de recorrentes operações sintéticas no campo amostral. Neste particular, ainda mais incontestável se torna a diferenciação interpretativa concomitantemente à repetição descritiva, ao examinarmos a fórmula geral do coeficiente de correlação parcial entre X_1 e X_2, para N variáveis, em função dos coeficientes de correlação múltipla: -

$$r_{12,34....n} = \sqrt{1 - \frac{1 - R^2_{1,234...n}}{1 - R^2_{1,34....n}}}$$

QUARTA PARTE

IDENTIFICAÇÃO DA TEORIA AOS FATOS

- CAPÍTULO XIII -

APLICAÇÃO DA SÍNTESE POR INDICIAÇÃO

As operações descritivas da estatística encontram na construção dos números-indices de grandezas econômicas, um dos mais frutíferos e promissores campos de aplicação. A problemática é realmente de monovariação, onde existem apenas agregados numéricos da mesma natureza, constituídos da superposição dos valores de uma única grandeza, seja o preço, a produção agrícola ou industrial, a renda nacional, a quantidade ou o valor da exportação etc.

Escolhemos o campo econômico para aplicação da teoria exposta nas partes anteriores dêste trabalho, tanto neste capítulo como no próximo concernente à análise, por já haver sido há muito assentado um julgamento definitivo sôbre a essencilidade do método estatístico na elucidação de múltiplos aspectos dos fatos econômicos, cuja classe de fenômenos é traduzível sòmente em modalidade quantitativa, forma coletiva e grande complexidade. Desde 1925 Warren Persons expunha que o material de interêsse do economista na qualidade de estatístico, consiste em dados quantitativos relacionados aos fenômenos em massa conexos às atividades de produção e consumo da riqueza no seio da humanidade. (41).

A síntese por indiciação, fundada no critério unidimensional, a que nos reportamos no capítulo 5, é adequada e suficiente à verificação das características do fenômeno homogêneo desdobrado quantitativamente nas sucessivas magnitudes de uma grandeza econômica.

O protótipo de descrição estatística em tôrno de fatos econômicos, é possìvelmente aquêle que compreende a redução ou nivelamento dos valores desiguais porém homogêneos de uma grandeza econômica, a

---

(41) PERSONS, WARREN.M. Statistics and Economic Theory. In The Review of Economic Statistics, Vol VII, Julho 1925, nº 3, pag, 182.

Fim de mostrar em que ponto aproximadamente ocorre a contração dos referidos valores. Alcança-se êste propósito através dos números-índices, cuja técnica elaborativa repousa no princípio dos valores sinóticos da classe dos promédios. Semelhantemente ao que dirige o estatístico na fixação de um promédio para medir a tendência central de uma variável X em conjunto monovariável, a construção de números-índices na ciência econômica está bàsicamente impregnada da idéia de prover valores sumários, afastados o mínimo possível dos demais componentes do agregado numérico de natureza econômica.

No exemplo concreto a seguir, o campo de observação será a amostra monovariável compreensiva da produção per-capita de alguns produtos da lavoura baiana destinados ao consumo interno, expressa em quilograma, no qüinqüênio 1950-1954, cujos dados foram destacados de um quadro de período retroativo mais dilatado, organizado por Helio Sento Sé, economista do Instituto de Economia e Finanças da Bahia.(42) A descrição das características dêste fenômeno, implicando na síntese por indiciação, abrange o cálculo dos indispensáveis números-índices sôbre os valores discriminados no Quadro 1.

(42) SENTO-SÉ, HÉLIO. Estrutura e Desenvolvimento da Lavoura na Bahia. 1945-1954. Salvador, Instituto de Economia e Finanças da Bahia, 1957. Edição mimeografada. Apêndice.

- Quadro 1 -

Estimativa da produção per-capita de alguns produtos da lavoura destinados ao consumo interno na Bahia - 1950-1954

Produtos	Produção per-capita - Kg.				
	1950	1951	1952	1953	1954
Arroz	3,9	3,0	2,2	2,3	3,6
Banana	28,3	26,8	26,0	28,8	29,0
Batata doce	7,7	7,4	7,3	8,8	10,3
Batata inglêsa	0,4	0,4	0,3	0,5	0,4
Café	4,4	3,6	3,0	3,7	3,2
Cana-de-açucar	37,0	36,0	35,6	35,8	36,7
Feijão	9,5	13,4	8,0	11,2	12,2
Mandioca	135,3	114,3	112,3	115,5	142,7

FONTE: - INSTITUTO DE ECONOMIA E FINANÇAS DA BAHIA.

Convertendo-se êstes valores originais a relativos, com referência ao ano de 1950, tem lugar o quadro 2.

- Quadro 2 -

Estimativa da produção per-capita de alguns produtos da lavoura destinados ao consumo interno na Bahia
1950-1954

Produtos	Produção per-capita - em números relativos				
	1950	1951	1952	1953	1954
Arroz	100	77	56	59	92
Banana	100	95	92	102	102
Batata doce	100	96	95	114	134
Batata inglêsa . . .	100	100	75	125	100
Café	100	82	68	84	73
Cana-de-açucar . . .	100	97	96	97	99
Feijão	100	141	84	118	128
Mandioca	100	84	83	85	105
Totais	800	772	649	784	833

O promédio aritmético estudado no capítulo 5, expresso na fórmula (5.2), dá origem às médias aritméticas de preços relativos, por alguns denominadas índices aritméticos simples, cuja determinação efetuaremos em conexão aos relativos acima. Denominando R_{oi}, R_{1i}, R_{2i}, ... R_{ni} cada uma das sucessões de relativos nos anos considerados na pesquisa, a síntese por indiciação através do referido promédio, produz os seguintes índices:(43)

(43) De modo geral, $R_{ni} = \frac{B_{ni}\ 100}{B_{oi}}$ em que B_{ni} equivale à produção per-capita de um dos gêneros em determinado ano, e B_{oi} representa a produção per-capita da mesma mercadoria no ano-base. B_n é uma grandeza derivada, resultante do quociente $\frac{K_n}{P_n}$, sendo K_n a produção total e P_n a pupulação em igual período.

$$I\ (1950) = \frac{\sum R_{oi}}{N} = \frac{800}{8} = 100$$

$$I\ (1951) = \frac{\sum R_{1i}}{N} = \frac{772}{8} = 97$$

$$I\ (1952) = \frac{\sum R_{2i}}{N} = \frac{649}{8} = 81$$

$$I\ (1953) = \frac{\sum R_{3i}}{N} = \frac{784}{8} = 98$$

$$I\ (1954) = \frac{\sum R_{4i}}{N} = \frac{833}{8} = 104$$

O promédio aritmético que exprimimos sob a fórmula (5.1), propicia o cálculo de um índice aritmético mais complexo, o índice aritmético ponderado ou simplesmente a média aritmética ponderada de preços relativos, que mede a influência proporcional de cada um dos relativos na amplitude do agregado. Também os promédios geométricos expressos matemàticamente em (5.5) e (5.6) no capítulo 5, aplicam-se à determinação de números-índices através do processo geométrico.

Acompanhando a marcha calculatória dos índices aritméticos simples, notamos que para cada ano, invariàvelmente reduzimos ou nivelamos os valores desiguais porém homogêneos, consubstanciados nos relativos da produção per-capita agrícola, obedecendo êste procedimento à síntese por indiciação peculiar à descrição estatística. Os índices dêste modo determinados, representam sintèticamente o comportamento da produção per-capita do conjunto de gêneros, no ano considerado, apreciando-se complementarmente o sentido da evolução dessa grandeza econômica no período 1950-1954 por meio da comparação entre si dos índices obtidos em cada ano.

As medidas de distensão que são também instrumentos unidimensionais amplamente utilizáveis no domínio descritvo da estatística desem

penham porém no setor de aplicação econômica, um papel mais eficaz como elementos integrantes da análise de amostras bivariáveis e multivariáveis sob o cunho pluridimensional da interpretação. No capítulo seguinte, examinaremos um caso concreto onde a metodologia analítica, de caráter interpretativo, será aplicada a um problema de multivariação econômica.

- CAPÍTULO XIV -

APLICAÇÃO DA ANÁLISE

Discorremos através dos capítulos 7 a 12, da terceira parte, sôbre a maneira pela qual a síntese por indiciação se combina com o processo dissociativo da análise, enquanto são computadas as relações estocásticas e também os coeficientes de determinação e correlação. Isto porque interfere a todo instante o cálculo de desvios em tôrno de pontos característicos, incorporando-se esta técnica à dedução dos parâmetros das equações de regressão e dos mencionados coeficientes.

Comprovaremos na ciência econômica a simbiose da síntese com a análise, estudando os aspectos imanentes da primeira coexistindo com os modos cambiantes da última, na interpretação do fenômeno dos transportes terrestres em suas conexões com a renda nacional do setor agro-industrial, referente às condições observadas em algumas regiões do Brasil no ano de 1956.

O exame da dependência e correlação entre as variáveis compreendidas neste problema econômico, através da metodologia analítica de base pluridimensional, permite constatarmos em têrmos práticos a evolução do método, com a transformação do simples caráter primário da descrição nas feições secundária e terciária da interpretação.

Para compôr a amostra multivariável abrangendo três fatôres específicos, a saber, extensão ferroviária, distância rodoviária e renda nacional das atividades agrícolas e industriais, consideraremos estas variáveis identificadas às circunstâncias que prevalecem nos seguintes Estados, escolhidos em cada uma das regiões fisiográficas brasileiras: - Pará (Norte); Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco e Alagoas (Nordeste); Sergipe, Bahia, Espíri

to Santo e Rio de Janeiro (Leste); Paraná e Santa Catarina (Sul); Mato Grosso e Goiás (Centro-Oeste).(44)

A pesquisa consistirá então através da análise desta amostra trivariável consoante o critério pluridimensional, em interpretar quantitativamente a influência das longitudes das estradas de ferro e de rodagem na formação da renda nacional oriunda de fontes agro-industriais.

Constituída a amostra nas condições acima, encerrando as 16 unidades federadas os caracteres da situação ferro-rodoviária em conexão com a renda da agricultura e indústria, em vários trechos do território nacional que apresentam aspectos de certo modo uniformes sob o ponto de vista econômico,(45) procuraremos primeiramente estabelecer uma relação estocástica entre as três variáveis - renda, rêde ferroviária e rêde rodoviária -, deduzindo para isso uma equação de regressão, e depois avaliar o grau ou estreiteza da relação entre as mesmas variáveis por meio do coeficiente de correlação.

É um caso evidente de variação múltipla, porquanto há mais de duas variáveis no problema, sendo por hipótese admitida a renda nacional agro-industrial dependente do alcance de penetração dos caminhos ferroviários e rodoviários. É mister, pois, para apreciar o nexo que se julga existir entre a renda e as extensões quilométricas das ferrovias e rodovias, verificar estatìsticamente aquela hipótese através dos processos de repressão e correlação múltipla, em que sôbre

---

(44) Seleção denominada "sistemática" na tipologia pertinente à técnica da amostragem. Os diversos Estados foram tomados de cada uma das áreas fisiográficas brasileiras, segundo uma orientação prévia do analista fundada na relativa homogeneidade, e não de maneira casual.

(45) Atendendo às normas da amostragem sistemática, foram excluídos desta análise os Estados de São Paulo, Rio Grande do Sul e Minas Gerais, porquanto existem aí condições mais pronunciadas de desenvolvimento econômico. Havendo nessas áreas grande densidade de capitais privados e públicos, polarização de trabalho especializado e concentração de outros fatôres de expansão, é razoável atribuir uma parte ponderável da renda agro-industrial daquelas três unidades federadas, a influência de tais condições específicas, as quais se verificam, entretanto, em menor proporção noutros Estados, sobretudo naquêles situados nas periferias do Norte e Nordeste.

X_1 - renda -, atuam agregadamente X_2 - extensão ferroviária e - X_3 - distância rodoviária. Os diagramas de dispersão das figuras 8 e 9 orientam o analista quanto à relação aproximadamente linear de cada uma das longitudes ferroviáriase rodoviária para a renda agro-industrial.

Reconhecemos que êste estudo não tem caráter indutivo, por ser impraticável a generalização de resultados sob a forma de extrapolação, pois do universo original de pequena amplitude selecionamos a maior parte dos grupos relativamente homogêneos para uma apreciação específica; no entanto, a interpretação dêste fenômeno multivariável através das suas dependências e relações, constituindo uma operação analítica, dar-nos-á oportunidade de evidenciar, não obstante as limitações provenientes do fato de haver poucos elementos na amostra, as mudanças no agregado trivariável representativo das condições existentes nos Estados especificados, mudanças essas que são refletidas nos coeficientes de regressão, determinação e correlação.

Para iniciar a tarefa, organizemos os quadros 3-A, 3-B e 3-C, nos quais estão arroladas as várias unidades federadas ao lado dos valores que lhes correspondem relativamente às três variáveis e aos produtos derivados, assim como as necessárias operações sintéticas cujos resultados estão colocados abaixo das referidas tabelas.

A análise aplicada à amostra trivariável da renda agro-industrial sob o influxo dos transportes ferroviários e rodoviários, consiste originalmente na adaptação de uma relação estocástica aos elementos do problema. Se bem que as três componentes sejam dissociadas em observância à análise, esta dissociação requer entretanto o prévio emprêgo da síntese por indiciação no cálculo dos parâmetros da equação de regressão múltipla.

As fórmulas paramétricas apropriadas ao caso de três variáveis, referidas matemàticamente nas expressões (10.7) e (10.8) do capítulo 10, dão ensejo à repetição de modalidades sintéticas à medida que se determinam sucessivamente na sua contextura, os coeficientes brutos de regressão linear - b_{12}, b_{13}, b_{32} e b_{23} -, mediante os quais são definidos os coeficientes líquidos ou parciais. - $b_{12,3}$ e $b_{13,2}$

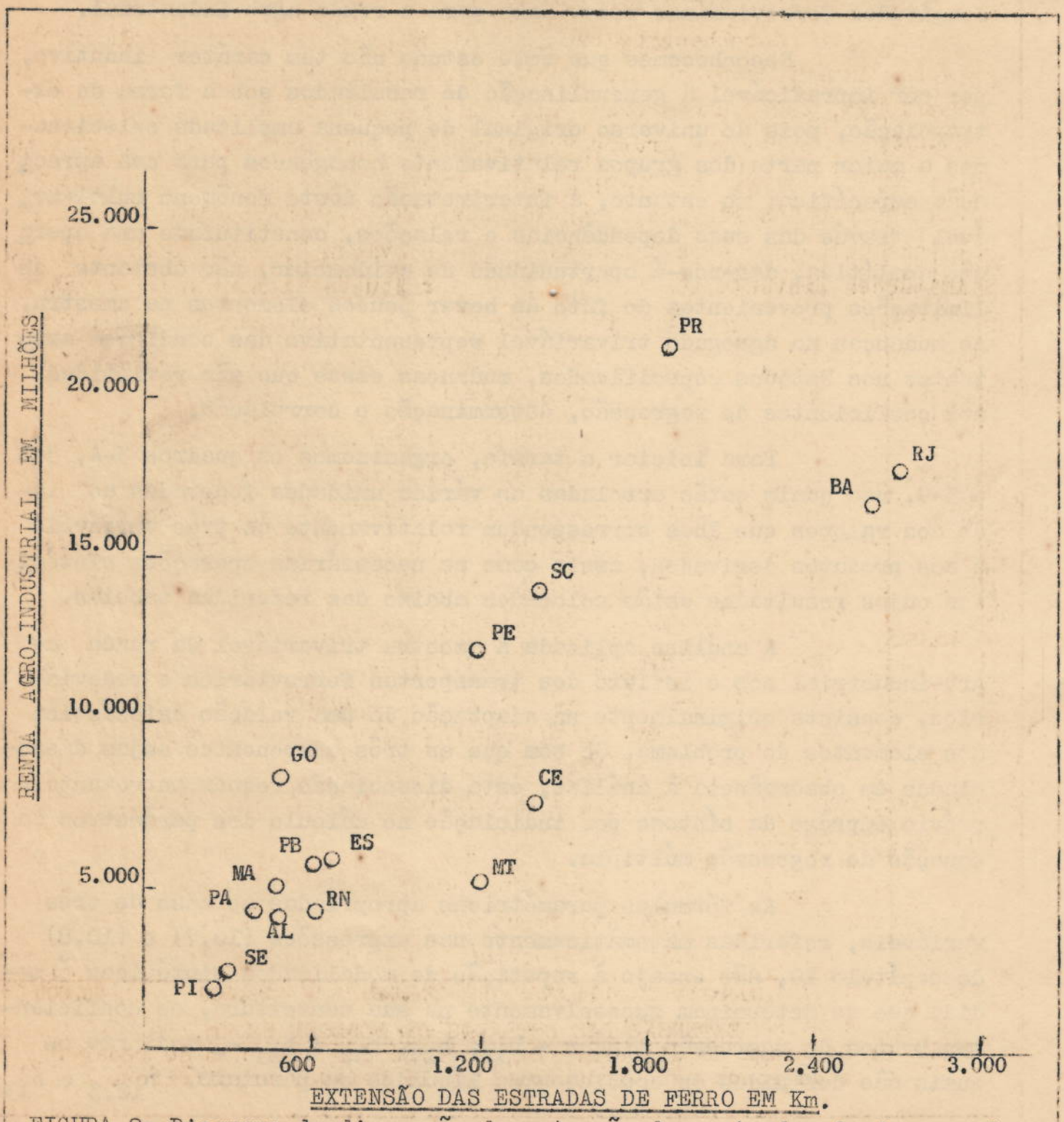


FIGURA-8: Diagrama de dispersão da extensão das estradas de ferro e da renda agro-industrial em 16 Estados brasileiros.

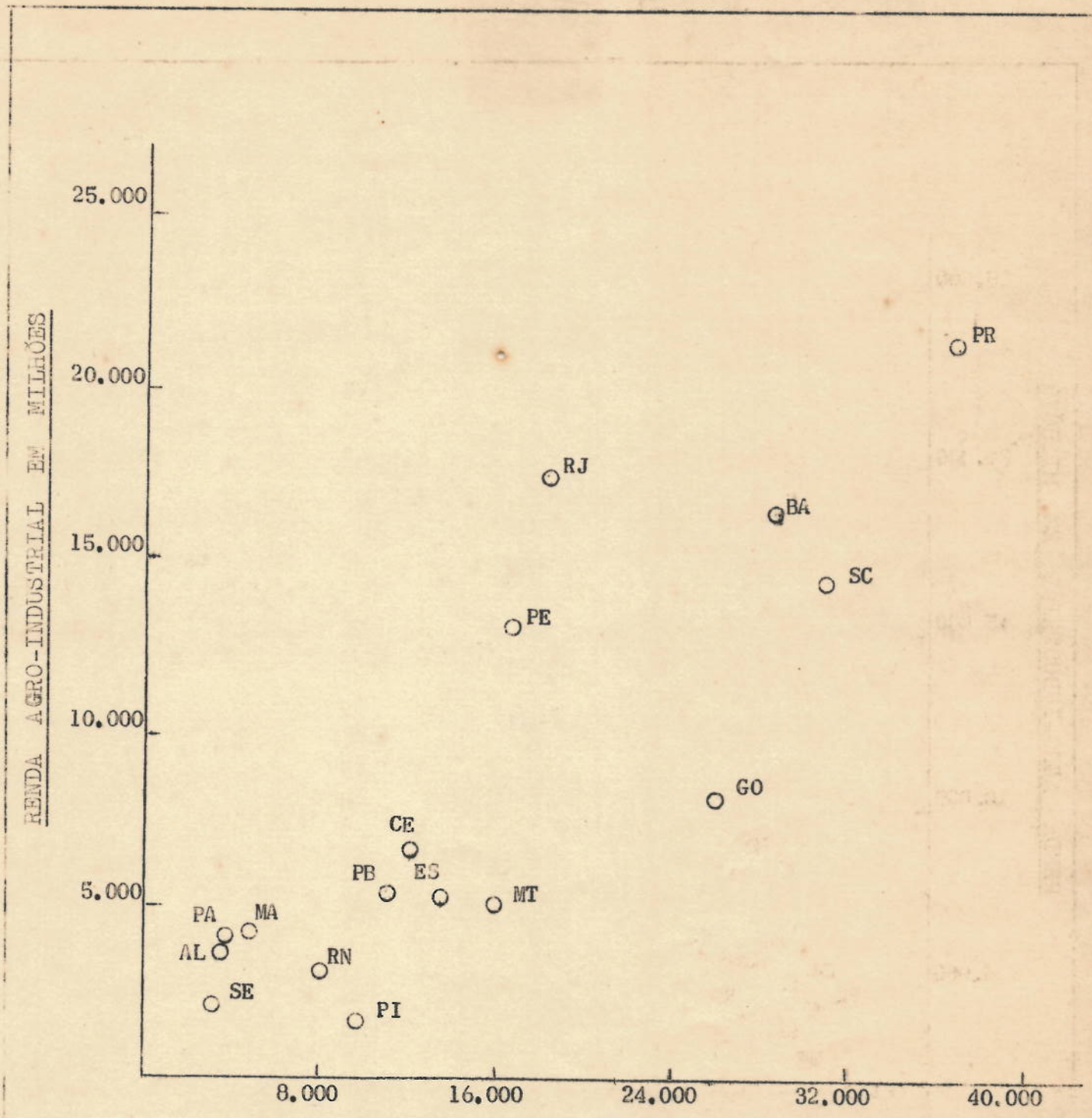
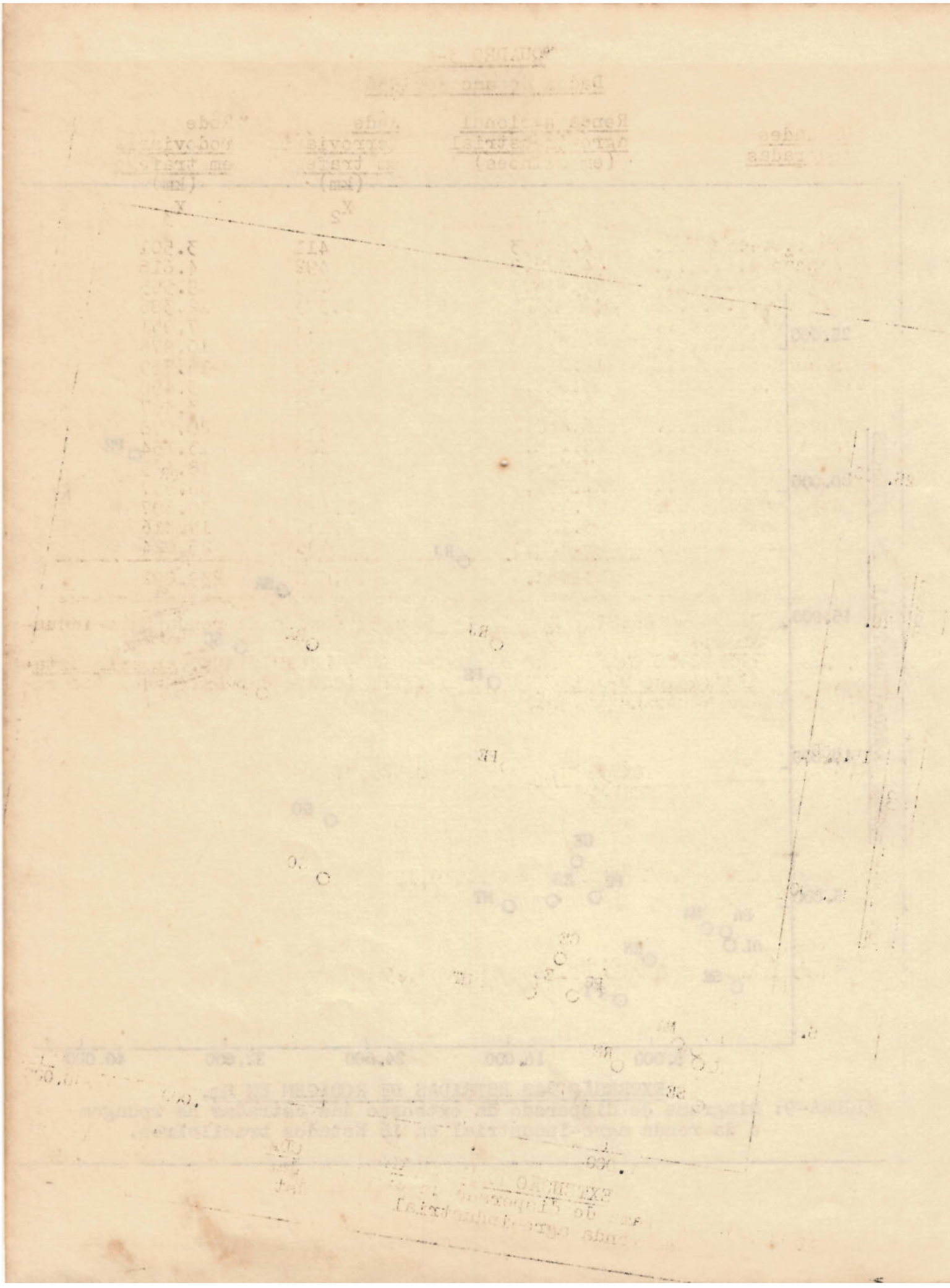


FIGURA-9: Diagrama de dispersão da extensão das estradas de rodagem e da renda agro-industrial em 16 Estados brasileiros.



QUADRO 3-A

Dados do ano de 1956

Unidades federadas	Renda nacional agro-industrial (em milhões) X_1	Rêde ferroviária em trafego (km) X_2	Rêde rodoviária em trafego (km) X_3
Pará	4.049,3	411	3.501
Maranhão	4.126,9	492	4.618
Piauí	1.768,5	243	9.555
Ceará	7.432,2	1.395	12.386
Rio G. do Norte ..	3.218,0	614	7.951
Paraíba	5.717,8	608	10.976
Pernambuco	12.288,4	1.183	16.759
Alagoas	3.813,4	474	3.490
Sergipe	2.298,1	297	3.407
Bahia	16.440,2	2.593	28.078
Espírito Santo ...	5.418,0	663	13.254
Rio de Janeiro ...	17.565,7	2.676	18.423
Paraná	21.351,5	1.875	36.557
Santa Catarina ...	14.115,3	1.412	30.597
Mato Grosso	5.166,7	1.195	15.316
Goías	8.410,3	495	25.024
	133.180,3	16.626	239.892

FONTES:- BANCO DO BRASIL, Relatório de 1957 (dados da renda agro-industrial)
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Anuário Estatístico do Brasil-1957. Ano XVIII (dados das extensões das redes de transportes).

$$\bar{X}_1 = \frac{\sum X_1}{N} = \frac{133.180,3}{16} = 8.323,77$$

$$\bar{X}_2 = \frac{\sum X_2}{N} = \frac{16.626}{16} = 1.039,13$$

$$\bar{X}_3 = \frac{\sum X_3}{N} = \frac{239.892}{16} = 14.993,25$$

QUADRO 3-B

Dados do ano de 1956

Unidades Federadas	X_1X_2	X_1X_3	X_2X_3
Pará	1.664.262,3	14.176.599,3	1.438.911
Maranhão	2.030.434,8	19.058.024,2	2.272.056
Piauí	429.745,5	16.898.017,5	2.321.865
Ceará	10.367.919,0	92.055.229,2	17.278.470
Rio G. do Norte	1.975.852,0	25.586.318,0	4.881.914
Paraíba	3.476.422,4	62.758.572,8	6.673.408
Pernambuco	14.537.177,2	205.941.295,6	19.825.897
Alagôas	1.807.551,6	13.308.766,0	1.654.260
Sergipe	682.535,7	7.829.626,7	1.011.879
Bahia	42.629.438,6	461.607,935,6	72.806.254
Espirito Santo	3.592.134,0	71.810.172,0	8.787.402
Rio de Janeiro	47.005.813,2	323.612.891,1	49.299.948
Paraná	40.034.062,5	780.546.785,5	68.544.375
Santa Catarina	19.930.803,6	431.885.834,1	43.202.964
Mato Grosso	6.174.206,5	79.133.177,2	18.302.620
Goiás	4.163.098,5	210.459.347,2	12.386.880
	200.501.457,4	2.816.668.592,0	330.689.103

$$s_{12} = \frac{\Sigma X_1X_2}{N} - \bar{X}_1\bar{X}_2 = \frac{200.501.457,4}{16} - (8.323,77)(1.039,13) =$$

$$12.531.341,09 - 8.649.479,12 = 3.881.861,97$$

$$s_{13} = \frac{\Sigma X_1X_3}{N} - \bar{X}_1\bar{X}_3 = \frac{2.816.668.592}{16} - (8.323,77)(14.993,25) =$$

$$176.041.787 - 124.800.364,55 = 51.241.422,45$$

$$s_{23} = \frac{\Sigma X_2X_3}{N} - \bar{X}_2\bar{X}_3 = \frac{330.689.103}{16} - (1.039,13)(14.993,25) =$$

$$20.668.068,94 - 15.579.935,87 = 5.088.133,07$$

QUADRO 3-C

Dados do ano de 1956

Unidades federadas	X_1^2	X_2^2	X_3^2
Pará	16.396.830,49	168.921	12.257.001
Maranhão	17.031.303,61	242.064	21.325.924
Piauí	3.127.592,25	59.049	91.298.025
Ceará	55.237.596,84	1.946.025	153.412,996
Rio G. do Norte	10.355.524,00	376.996	63.218.401
Paraíba	32.693.236,84	369.664	120.472.576
Pernambuco	151.004.774,56	1.399.489	280.864.081
Alagoas	14.542.019,56	224.676	12.180.100
Sergipe	5.281.263,61	88.209	11.607.649
Bahia	270.280.176,04	6.723.649	788.374.084
Espírito Santo	29.354.724,00	439.569	175.668.516
Rio de Janeiro	308.553.816,49	7.160.976	339.406.929
Paraná	455.886.552,25	3.515.625	1.336.414.249
Santa Catarina	199.241.694,09	1.993.744	936.176.409
Mato Grosso	26.694.788,89	1.428.025	234.579.856
Goiás	70.733.146,09	245.025	626.200.576
	1.666.415.039,61	26.381.706	5.203.457.372

$$s_1^2 = \frac{\sum X_1^2}{N} - \bar{X}_1^2 = \frac{1.666.415.039,61}{6} - (8.323,77)^2 \approx$$

$$= 104.150.939,98 - 69.285.147,01 = 34.865.792,97$$

$$s_1 = \sqrt{34.865.792,97} = 5.904,73$$

$$s_2^2 = \frac{\sum X_2^2}{N} - \bar{X}_2^2 = \frac{26.381.706}{16} - (1.039,13)^2 =$$

$$= 1.648.856,63 - 1.079.791,16 = 569.065,47$$

$$s_2 = \sqrt{569.065,47} = 754,36$$

$$s_3^2 = \frac{\sum X_3^2}{N} - \bar{X}_3^2 = \frac{5.203.457.372}{16} - (14.993,25)^2 =$$

$$= 325.216.085,75 - 224.797.545,56 = 100.418.540,19$$

$$s_3 = \sqrt{100.418.540,19} = 10.020,91$$

- $b_{12,3}$ $b_{13,2}$ -; Esboça-se concomitantemente nessa etapa preliminar de trivariabilidade, uma diferenciação analítica que tende a discernir a índole interpretativa do nosso estudo de multivariabilidade econômica.

Os valores sinóticos essenciais à determinação de cada um dos coeficientes brutos e subsequentemente dos coeficientes líquidos, estão já calculados ao pé dos quadros 3-A, 3-B e 3-C. Substituídos nas expressões abaixo - b_{12}, b_{13}, b_{32} e b_{23} -, as quais são formas especiais da (8.4) deduzida no capítulo 8, teremos efetuado uma síntese por indiciação, porém complementar e estritamente subordinada ao procedimento dissociativo da análise, dado ao caráter multivariável da amostra e às finalidades pluridimensionais daí decorrentes.

$$b_{12} = \frac{s_{12}}{s_2^2} = \frac{3.881.861,97}{569.065,47} = 6,821$$

$$b_{13} = \frac{s_{13}}{s_3^2} = \frac{51.241.422,45}{100.418.540,19} = 0,510$$

$$b_{32} = \frac{s_{32}}{s_2^2} = \frac{5.088.133,07}{569.065,47} = 8,941$$

$$b_{23} = \frac{s_{23}}{s_3^2} = \frac{5.088.133,07}{100.418.540,19} = 0,051$$

$$b_{12,3} = \frac{b_{12} - b_{32}b_{13}}{(1 - b_{32}b_{23})} = \frac{6,821 - 8,941 \times 0,510}{(1 - 8,941 \times 0,051)} = 4,156$$

$$b_{13,2} = \frac{b_{13} - b_{23}b_{12}}{(1 - b_{23}b_{32})} = \frac{0,510 - 0,051 \times 6,821}{(1 - 0,051 \times 8,941)} = 0,298$$

Diante dos dois últimos parâmetros, a equação de regressão linear múltipla adaptável ao problema concernente à influência das longitudes ferroviáriase rodoviária sôbre a renda agro-industrial, referidas as variáveis aos seus valores originais, será: -(46)

$$X_1 = -462,843 + 4,156X_2 + 0,298X_3 \quad (14.1)$$

cuja equação mede os reflexos que a influência agregada dos tráfegos ferroviário e rodoviário imprime à renda agro-industrial.

O valor 4,156 atribuído ao coeficiente de regressão parcial - $b_{12,3}$ -, exprimindo a taxa de variação entre X_1 e X_2, mede a alteração normal na magnitude da renda agro-industrial, associada à variação unitária na quilometragem ferroviária em regiões que possuem a mesma extensão rodoviária; o valor 0,298 atribuído ao coeficiente de regressãosparcial - $b_{13,2}$ - equivalendo à taxa de variação entre X_1 e X_3, afere a mudança normal ocorrida na grandeza da renda agro-industrial, vinculada à variação unitária na quilometragem rodoviária em regiões de igual extensão ferroviária. Avalia-se assim com certa precisão a dependência existente entre X_1 e X_2, ou entre X_1 e X_3, mantendo-se constante em cada um dêstes casos a outra variável.

Os valoresde X_1 derivados da equação (14.1) são estimativas da renda em função da variação agregada de X_2 e X_3, respectivamente os comprimentos da rêde ferroviária e da rodoviária nas áreas especificadas nesta pesquisa de interpretação econômica.

Prosseguindo nossa análise multivariável, é oportuno e essencial efetuar o cálculo do coeficiente de correlação múltipla - $R_{1,23}$ -, o indicador do grau de relação entre o fator dependente (X_1) e os dois outros que agem agregadamente (X_2 e X_3). Alguns dos valores sinóticos obtidos com base nos elementos dêste problema trivariável, se levados à equação (11.6) do capítulo 11, deixa perceber novamente a

---

(46) Na equação (14.1) do tipo $X_1 = a_{1,23} + b_{12,3}X_2 + b_{13,2}X_3$, o parâmetro $a_{1,23} = \bar{X}_1 - b_{12,3}\bar{X}_2 - b_{13,2}\bar{X}_3$. Fazendo as devidas substituições, resulta o valor - 462,843 para $a_{1,23}$.

diferenciação analítica no exame das condições de relação múltipla nessa problemática econômica, paralelamente à repetição sintética.

O emprêgo da referida fórmula, após substituição de valores, resulta em: -

$$R_{1,23} = \sqrt{\frac{b_{12,3}s_{12} + b_{13,2}s_{13}}{s_1^2}} =$$

$$= \sqrt{\frac{4,156 \times 3.881.861,97 + 0,298 \times 51.241.422,45}{34.865.792,97}} =$$

$$= \sqrt{0,90068114} = 0,9490$$

A renda agro-industrial, está por conseguinte, relacionada bem acentuadamente com a ação conjugada das extensões ferroviária e rodoviária, tal qual demonstra o alto coeficiente de correlação múltipla.

O quadrado de $R_{1,23}$, ou seja, o coeficiente de determinação múltipla - $R^2_{1,23} = 0,90068$ - mostra que cêrca de 90,07% da variação ocorrida na magnitude da renda pode ser interpretada ou explicada pela equação de regressão, desde que variem agregadamente as quilometragens das estradas de ferro e rodagem.

Êstes dois coeficientes, no entanto, medem especificamente o grau de relação entre as três variáveis, assim como a variabilidade explicada através da equação de regressão, apenas na medida em que se considera a variável dependente única - X_1 - sofrendo os efeitos da ação agregada das duas variáveis independentes X_2 e X_3. O conhecimento do grau de relação da variável dependente para cada uma das duas variáveis independentes, tanto quanto a proporção de variabilidade do fator dependente não explicada por uma das variáveis independentes mas que passa a ser interpretada pela outra, envolve a deter

minação dos coeficientes de correlação e determinação parcial.

Efetivamente, os coeficientes de correlação e determinação líquidos ou parciais da renda relativamente ao comprimento da rêde ferroviária, aferem o volume da renda agro-industrial associado à longitude ferroviária - X_2 -, eliminando-se os efeitos da variação de X_3 - extensão rodoviária. Por outro lado, a longitude rodoviária - X_3 - associada à grandeza da renda agro-industrial - X_1 - com os efeitos de X_2 - extensão ferroviária - eliminados, é mensurável pelos coeficientes de correlação e determinação líquidos ou parciais da renda relativamente à longitude rodoviária.

As expressões (12.2) e (12.4) estabelecidas no capítulo 12, servirão ao propósito especial de analisar parcialmente a condição de trivariabilidade econômica ora estudada, cujo processo subentendido na dissociação das três componentes duas a duas, implica necessàriamente em cada caso, na complementação calculatória dos valores sinóticos. Não há, pois, que subtrair nesta nova circunstância, o emprêgo de modos analíticos principais simultâneamente com operações sintéticas acessórias.

Efetuando as substituições nas fórmulas apropriadas, determinam-se os valores dos dois coeficientes de correlação parcial: -

$$r_{12,3} = \sqrt{1 - \frac{1 - R^2_{1,23}}{1 - r^2_{13}}} = \sqrt{1 - \frac{1 - (0,9490)^2}{1 - (0,8660)^2}} =$$

$$= \sqrt{0,60279447} = 0,7764$$

$$r_{13,2} = \sqrt{1 - \frac{1 - R^2_{1,23}}{1 - r^2_{12}}} = \sqrt{1 - \frac{1 - (0,9490)^2}{1 - (0,8715)^2}} =$$

$$= \sqrt{0,58701074} = 0,7662$$

Os resultados $r_{12,3} = 0,7764$ e $r_{13,2} = 0,7662$ significam que a renda agro-industrial está um pouco mais vinculada à extensão f ferroviária do que à rodoviária, diferindo a estreiteza da relação em 1,3%, em têrmos relativos, ao considerarmos a influência alternada das longitudes das ferrovias e rodovias.

É importante comparar a grandeza dos coeficientes de correlação líquida à magnitude dos coeficientes de correlação bruta, a fim de verificar a intensidade da verdadeira relação, quando se eliminam os efeitos de um terceiro fator atuante no campo amostral.

Os coeficientes brutos r_{12} e r_{13} que já entraram na composição das fórmulas dos coeficientes parciais acima especificadas, são calculáveis através das formas particulares da equação (9.12) deduzida no capítulo 9. Tomando-se os dados concernentes à renda agro-industrial relacionada aos transportes terrestres, teremos:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{3.881.861,97}{5.904,73 \times 754,36} = 0,8715$$

$$r_{13} = \frac{s_{13} s_{13}}{s_1 s_3} = \frac{51.241.422,45}{5.904,73 \times 10.020,91} = 0,8680$$

A respeito da influência dos caminhos ferroviários, a comparação do coeficiente de correlação líquida - $r_{12,3} = 0,7764$ - ao respectivo coeficiente de correlação bruta - $r_{12} = 0,8715$ -, demonstra que a tentativa de mensurar a variação entre a renda agro-industrial e a longitude ferroviária sem primeiro eliminar-se a influência também exercida pela extensão rodoviária, resulta de certo modo imprecisa, pois r_{12} reflete do mesmo modo os efeitos advindos ao nível da renda, provenientes da variação específica das condições rodoviárias.

Por outro lado, o confronto do coeficiente de correlação líquida - $r_{13,2} = 0,7662$ - com o seu correspondente bruto - $r_{13} = 0,8660$ - revela que não se procurando prèviamente anular os efeitos das variações quilométricas ferroviárias sôbre a renda agro-inudstrial, o valor atribuído a r_{13} encerra além dos efeitos oriundos das variações na extensão das rodovias, aquelas outras influências dimanadas das mudanças longitudinais das ferrovias.

Em ambos os casos, a aplicação do coeficiente bruto concorreu para sobreestimar a medida do grau de relação entre as duas variáveis econômicas, desvantagem esta sanada com a introdução do coeficiente líquido ou parcial, tècnicamente mais aperfeiçoado e mais eficaz quanto aos resultados apresentados.

Os coeficientes de determinação, decorrentes da elevação ao quadrado daquêles coeficientes de correlação, suprem conclusões úteis no tocante à avaliação da influência das variáveis X_2 e X_3 atuando separadamente sôbre X_1. O exame de $r^2_{12,3}$ indica ser a extensão ferroviária suscetível de explicar 0,6027 ou cêrca de 60,3% da variabilidade não interpretada pela atuação do fator rodoviário; da mesma forma, a apreciação de $r^2_{13,2}$ mostra que a longitude rodoviária explica 0,5870 ou cêrca de 58,7% da variabilidade remanescente após aferida ùnicamente a influência do fator ferroviário.

O caráter extensivo da interpretação está assim confirmado no estudo dêste fenômeno complexo de natureza econômica, em cuja amostra trivariável analisamos a influência múltipla e parcial da penetração ferro-rodoviária em certas áreas do território nacional. Na fixação da equação de regressão e dos coeficientes de determinação e correlação, consistindo na dissociação da complexidade estrutural da amostra com o objetivo de esclarecer os efeitos fenomênicos traduzidos na conjugação das três variáveis, constatamos uma fase de essencial diferenciação analítica sedimentada emo operações sintéticas complementares.

QUINTA PARTE

CONCLUSÃO

- CAPÍTULO XV -

A MATURAÇÃO METODOLÓGICA

Em face do expôsto nos capítulos incluídos nas partes precedentes dêste trabalho, a conclusão espontânea é que o método estatístico proporciona à ciência uma instrumentação perfeitamente exeqüível diante das circunstâncias confrontadas em diferentes condições fenomênicas, garantindo através dos seus processos e técnicas, resultados menos eivados dos "êrros de observação" comumente verificados no curso de uma pesquisa.

Erguido em bases quantitativas, formulam-se os seus conceitos em princípios matemáticos e constrói-se a sua sistemática em consonância com a lógica dedutiva. Contudo, sem embargo da sua gênese recional e do desenvolvimento de teorias estatísticas estribadas em postulados e axiomas matemáticos, cujas teorias visam principalmente a construção de modêlos teóricos que se admitem ajustar à realidade dos fatos, as condições de estruturação e aplicação do método estatístico sugerem depender a sua essencialidade e proficiência, mais da sua flexibilidade em interpor-se a contingências científicas diversificadas no tempo e no espaço, do que da identificação dos seus caracteres ao apriorismo inerentes às concepções matemáticas.

O ângulo sob o qual se percebe a perspectiva global do tratamento estatístico, envolve a superposição das modalidades quantitativas de um fenômeno, tão extensivamente quanto possível na amplitude de uma amostra, a fim de poder ser constatada alguma regularidade no intervalo da observação de uma ou de muitas variáveis.

De início ressaltamos a necessidade de proceder, ante a sucessão das magnitudes de uma ou mais grandezas abarcando um conjunto fenomênico, à verificação do regime de sua variação nos limites amostrais, segundo um dos critérios, unidimensional ou pluridimensional, os

quais guiam as pesquisas respectivamente nas fases primária e secundária ou terciária de variação, a primeira destas fases estruturada em amostras monovariáveis e as demais consubstanciadas em amostras bivariáveis e multivariáveis.

Se adstrita ao critério unidimensional empregado especialmente no campo monovariável, a observação das mudanças não requer senão a adoção de uma metodologia sintética, onde a recomposição e a indiciação satisfazem perfeitamente aos objetivos do sentido descritivo; neste estágio primário da estatística, em que se considera apenas uma variável, o processo reverte na descrição dos caracteres de um fenômeno homogêneo, usando-se para êste mister como parte integrante do método geral, a condensação, a qual subentende além da organização dos dados - síntese por recomposição -, a técnica calculatória dos valores sinóticos - síntese por indiciação.

Um tanto limitativas em seus propósitos sumários, as operações sintéticas provêem resultados apenas informativos em tôdas as implicações descritivas. Apesar de ser a síntese fecunda no domínio da descrição, por meio de cálculos autônomos, aplicáveis à estrutura de amostras monovariáveis, não se pode todavia notar enquanto ela estiver confinada ao setor descritivo, a evolução natural que o método estatístico tende a experimentar diante de circunstâncias outras surgidas no meio fenomênico.

Portanto, em problemas monovariáveis, os quais por sua natureza requerem especìficamente um traço informativo ou indício acêrca das características do fator quantitativo único, a condensação importa no nivelamento de valores desiguais porém homogêneos, todos pertencentes à mesma série estatística, medindo-se então no âmbito amostral, o grau de contração e de dispersão da variável. Cada uma destas condições, entretanto, conforme explanamos no decorrer dêste trabalho, impõe uma técnica especial destinada à determinação dos valores sinóticos necessários à descrição dos seus caracteres: - o grau de contração é referido pelos promédios e a dispersão é avaliada, dentre outras constantes, pelo desvio padrão.

Para adaptação dos fundamentos da metodologia sintética a uma situação real, resolvemos descrever o caso de monovariabilidade econômica configurada nas características da produção per-capita de alguns produtos agrícolas da Bahia, utilizando a síntese por indiciação especificada ao promédio aritmético simples calculado em relação aos dados brutos do problema. Compreendeu esta descrição a construção de números-índices, os quais confrontados entre sí serviram à aferição do comportamento daquela grandeza econômica durante o período considerado.

Por outro lado, nas reflexões quanto ao tratamento de amostras bivariáveis e multivariáveis, a verificação do regime das mudanças de duas ou mais variáveis sob o critério pluridimensional, reveste e forma de explicação em tôrno de um fenômeno heterogêneo. Depara da esta fase complexa, de natureza secundária e terciária, torna-se imprescindível a adoção de uma metodologia analítica, onde o processo dissociativo manifestado na técnica calculatória das regressões e dos coeficientes especiais, marca o sentido interpretativo no setor da estatística.

De cunho extensivo por seu objetivo de evidenciar dependências e correlações, as operações analíticas contràriamente às sintéticas, conduzem aos fins elucidativos que abrangem as motivações superiores de tôda indagação científica. As conclusões estatísticas no âmbito interpretativo, convém aqui reafirmar, apresentam-se profundamente diferenciadas em razão do número de variáveis da amostra, conquanto a sucessiva diferenciação do grau de análise comporte, denunciando subordinação da síntese, uma concomitância de modalidades sintéticas repetidas na formas originais.

Superadas as limitações intrínsecas da descrição, a qual visa apenas prover indícios sôbre fenômenos homogêneos em amostras monovariáveis por meio de operações sintéticas únicas e autônomas, a interpretação introduzida na pesquisa com a necessidade que se tem de explicar um fenômeno heterogêneo em amostras de duas ou mais variáveis, deixa perceber desde a apreciação dos simples agregados bivariáveis, que as aplicações essencialmente analíticas aí efetuadas,

supõem a síntese por indiciação através de valores sumários dos tipos de médias, desvios padrões, variâncias, covariâncias etc., os quais calculados a título de elementos iniciais, passam a compor a estrutura das próprias relações estocásticas e dos coeficientes, ambos de natureza analítica.

Verificamos já a coexistência sintético-analítica, embora a síntese de caráter acessório nas elaborações analíticas, mostrando então como os valores sinóticos interferem constantemente no cálculo dos parâmetros das equações de regressão e dos elementos componentes dos coeficientes de determinação e correlação.

Apresentou-se-nos a oportunidade de identificar a metodologia analítica a um caso concreto de multivariabilidade, interpretando em amostra trivariável econômica, a influência das rêdes ferroviária e rodoviária, agregada e parcialmente, na formação da renda agro-industrial em dezesseis estados brasileiros. A dissociação dos três fatôres econômicos consistiu no estabelecimento de uma equação de regressão múltipla, bem como na fixação dos coeficientes de correlação e determinação múltiplos e parciais, requerendo êste processo analítico, entretanto, o complemento da síntese por indiciação no cálculo dos valores sinóticos integrantes daquêles instrumentos de análise. Compreendeu esta interpretação, tanto a especificação das taxas de variação entre as condições de penetração daquelas duas vias de transporte e a renda, quanto a mensuração do grau de ralação e da variabilidade explicada entre os fatôres correlatos.

Portanto, na interpretação estatística, síntese e análise não são procedimentos opostos, mas sim complementares, por coexistirem êstes dois sentidos metodológicos no curso das perquisas de índole explicativa. Esta simbiose, quando constatada em amostras bivariáveis, conquanto ainda em estado embrionário, revela as primeiras manifestações de evolução do método; de fato, realizando-se a investigação nesse estágio secundário da estatística, começam a surgir as transmudações na estrutura amostral, cuja decomposição das duas séries mediante os processos de regressão e correlação, assenta os fundamentos das ilações estabelecidas nas várias ciências. A maior complexidade

decorrente da crescente heterogeneidade do meio fenomênico, faz-nos entrar em contato com a fase terciária ou as demais de maior hierarquia do método, quando se pode observar mutações mais pronunciadas na amostra; a dissociação de três ou mais agregados por meio de regressões e correlações diversificadas, produzindo diferenciação na análise, é precisamente ocasionada por cálculos sintéticos que se repetem à medida que aumenta o número de séries da amostra.

É mais significativa, por conseguinte, a maturação do método estatístico ante as interpretações de amostras trivariáveis, acentuando-se o grau de evolução metodológica com a diferenciação analítica, enquanto se acrescentam variáveis à amostra. Os cálculos sucessivamente repetidos de valores sinóticos dos mesmos tipos ou semelhantes, que ocorrem na contigência de crescimento da amostra multivariável, definem a marcha de análises concomitantes, cada vez mais acuradas, as quais circunscrevem e subordinam a síntese, desde que os amplos objetivos da interpretação suplantam os limitados propósitos da descrição.

Os processos gerais de avaliação dos coeficientes de regressão parcial e dos coeficientes de correlação múltipla e parcial, no caso de n variáveis, comparados aos meios de determinação das respectivas constantes na hipótese de dependência e relação simplesmente bivariável, isto é, transladando-se a interpretação do plano secundário para os outros e afinal suposta ela adaptável a uma fase de variação ideal como seja o caso genérico, permitem que se note a manutenção das formas sintéticas originais coexistindo com a extrema modificação dos aspectos analíticos sobrevenientes, cuja combinação simbiótica gera à medida que cresce a quantidade de variáveis, o aperfeiçoamento do próprio método estatístico.

Após considerados consecutivos casos de interpretação, que diferem entre si quanto ao maior número de fatôres relacionados na amostra multivariável, cujas divergências amostrais requerem operações analíticas sucessivamente mais penetrantes, resta induzirmos diante da sequência dos aspectos de crescente complexidade, uma lei fundamental da evolução do método estatístico. Faz-se mister para isso,

tomar as diferentes expressões quer dos coeficientes de regressão ou dos coeficientes de correlação, com número cada vez maior de variáveis até atingir o nível enésimo de variação; ganha muito em complexidade a contextura amostral através dêsses distintos graus de mutações, refletindo a fase ideal de variação que se exprime matemàticamente pelo caso geral, a interposição natural de uma instrumentação analítica extremamente sensível em face de ser imprescindível esclarecer, tão minuciosamente quanto possível, os efeitos de fenômenos cujas causas múltiplas se combinam com máxima sutileza.

A acentuada diversificação da análise, tal qual patenteada na composição das mencionadas fórmulas gerais, expressivas do coeficiente líquido da equação de regressão múltipla, do coeficiente de correlação múltipla e do coeficiente de correlação parcial, respectivamente -

$$b_{12,34\ldots n} = \frac{b_{12,34\ldots(n-1)} - b_{1n,34\ldots(n-1)}\, b_{n2,34\ldots(n-1)}}{1 - b_{2n,34\ldots(n-1)}\, b_{n2,34\ldots(n-1)}};$$

$$R_{1,23\ldots n} = \sqrt{\frac{b_{12,34\ldots n} s_{12} + b_{13,24\ldots n} s_{13} + \ldots + b_{1n,23\ldots(n-1)} s_{1n}}{s_1^2}} \quad \text{e}$$

$$r_{12,34\ldots n} = \sqrt{1 - \frac{1 - R^2_{1,234\ldots n}}{1 - R^2_{1,34\ldots n}}},$$

demonstrando a extrema difusão da síntese por indiciação incidente sob as formas originais dos conhecidos valores sinóticos, repetidas vêzes, no seguimento do processo dissociativo da análise, dá ensejo à indução de uma lei destinada a situar nos devidos têrmos, as fases de maturação do método mercê das alternativas, conquanto distintas mas complementares e simbióticas, da síntese e da análise.

Há pois razões em generalizarmos a conclusão de que, superados os restritos objetivos da descrição, realizados exclusivamente

por operações autônomas da síntese por indiciação, e alcançados os dilatados propósitos da interpretação, fundamentados na complementariedade sintético-analítica, constata-se na fase interpretativa da estatística, a manutenção das modalidades sintéticas concomitantemente à sucessiva e constante diferenciação do teor da análise.

A simples modalidade monovariável de um fenômeno homogêneo, requerendo para elucidação dos seus caracteres nada mais do que uma apreciação descritiva, determina ùnicamente a aplicação da síntese segundo o critério unidimensional de investigação científica; a complexa feição bivariável ou multivariável de um fenômeno heterogêneo, entretanto, exigindo por sua natureza um estudo interpretativo, impõe o emprêgo da análise fundamentada no critério pluridimensional, indispensável em pesquisas mais rigorosas. Se a razão de ser da descrição é constituída pelos meros intentos de apresentação e informação acêrca das características de um único fator em amostra simples, o fim da interpretação é sempre a explicação das mudanças verificadas no modo de combinação dos fatôres em amostras complexa.

Apesar de ser o estágio um tanto limitativo da descrição, atendido tão sòmente pela síntese, a fase extensiva da interpretação se processa através do sistema binário síntese-análise, cuja ação mostra como o método estatístico avança em sua trajetória evolutiva. A maturação da metodologia estatística, implicando no próprio desenvolvimento científico, acompanha a rigidez da coexistência sintético-analítica, na proporção em que a análise alcança graus sucessivamente mais diferenciados, sendo por bem dizer esta concomitância de natureza simbiótica a condição necessária daquela evolução, em harmonia com a especificidade do próprio sentido interpretativo.

BIBLIOGRAFIA

ALLEN, R.G.D. Análisis Matemático para Economistas. Tradução da edição inglêsa por Emilio Figueroa. Madrid, M. Aguilar, 1946.

ANDERSON, OSKAR. Statistics. In Encyclopedia of the Social Sciences. Vol. 14. New York, The Mac Millan Company, 1954.

ALVAREZ, M. GARCIA e OREJANA, J. AYUSO. Estadistica. Madrid, Editorial S.A.E.T.A., 1946.

BANCO DO BRASIL. Relatório de 1957. Rio de Janeiro, 1957.

CLARK, CHARLES E. An Introduction to Statistics. New York, John Wiley & Sons, Inc., 1953.

CÂMARA, LOURIVAL. Correlação. Rio de Janeiro, Apostila da série I-B da Escola Brasileira de Estatística, 1954.

CRAMER, HARALD. Metodos Matematicos de Estadistica. Tradução da edição original em inglês por Enrique Cansado. Madrid, Aguilar S. A. de Ediciones, 1953.

CROXTON, FREDERICK E. e COWDEN, DUDLEY J. Estatística Geral e Aplicada. Tradução da 8a edição norte-americana. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística, Conselho Nacional de Estatística, 1952.

CANSADO, ENRIQUE. Apuntes de Estadistica General. Edição mimeografada. Centro Interamericano de Enseñanza de Estadistica Economica y Financiera. Santiago, 1953.

CARVALHO, BULHÕES. Estatística, Método e Aplicação. Rio de Janeiro, Tip. Leuzinger, 1933.

EZEKIEL, MORDECAI. Methods of Correlation Analysis. New York, John Wiley & Sons, Inc., 1956.

GINI, CORRADO. Curso de Estadística. Tradução da edição italiano de 1946-1947 por Jorge Stecher Navarra. Madrid, Editorial Labor S/A, 1953.

GINI, CORRADO. Os Fundamentos e o Alcance do Método Estatístico. In Revista Brasileira de Estatística. Ano IX. Julho/Setembro 1948, n. 35.

GRANGER, GILLES GASTON. Lógica e Filosofia das Ciências. São Paulo, Edições Melhoramentos, 1955.

GOODMAN, RICHARD. Statistics. Londres, The English University Press, 1957

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Anuário Estatístico do Brasil - 1957. Ano XVIII. Rio de Janeiro, 1957.

INSTITUTO NACIONAL DE ESTATÍSTICA. Trabalhos do Seminário de Econometria dirigido pelo Prof. H. Wold. Lisboa, Publicações do Centro de Estudos Economicos, 1953.

MILLS, FREDERICK CECIL. Métodos Estatísticos Aplicados à Economia e aos Negócios. Tradução da 2a. edição norte-americana por H. Alvim Pessoa. Rio de Janeiro, Serviço Gráfico do Instituto Brasileiro de Geografia e Estatística, 1952.

NOGUEIRA DE PAULA, LUIZ. Metodologia da Economia Política. Rio de Janeiro, Irmãos Pongetti, 1942.

PERSONS, WARREN M. Statistics and Economic Theory. In Review of Economic Statistics. Vol. VII. Julho 1925, nº 3.

RIOS, SIXTO. Introduccion a los Metodos de la Estadistica. Madrid, Nuevas Graficas S.A., 1952.

RODRIGUES, MILTON DA SILVA. Elementos de Estatística Geral. São Paulo, Companhia Editôra Nacional, 1939.

SHEWART, W. A. Annual Survey of Statistical Technique: Developments in Sampling Theory. In Econometrica. Vol. I, nº 3, Julho 1933.

SENTO-SÉ, HELIO. Estrutura e Desenvolvimento da Lavoura na Bahia 1945-54. Edição preliminar mimeografada. Salvador, Instituto de Economia e Finanças da Bahia, 1957.

WOLD, HERMAN. Demand Analysis. New York, John Wiley & Sons, Inc., 1953.

WAUGH, ALBERT E. Elementos de Estatística. Tradução da 2a edição norte-americana por Ernesto Pellanda. Porto Alegre, Editôra Globo. Sem indicação de ano.

YULE, G. UDNY e KENDALL, M.G. Introdução à Teoria da Estatística. Tradução da 13a edição inglesa por Evandro de Oliveira Silva. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística, 1948.

ÍNDICE

QUARTA PARTE

QUINTA PARTE

E R R A T A

Página	Linha	Onde se lê	Leia-se
14	11	acuidades	acuidade
15	22	contínuas u discretas	discretas ou contínuas
53	rodapé	correspondente	corresponde
65	3	(X_i, X_i)	(Y_i, X_i)
71	4	indicação	indiciação
86	2 e 9	E_i	$\mathcal{E}_i$
90	1	teóricas	teórica
103	4	ferroviárias	ferroviária
111	3	ferroviárias	ferroviária
119	18	inerentes	inerente