INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - ENCE

RELATÓRIOS TÉCNICOS

Nº 001/88

CRÍTICA DE RAZÕES NO CENSO ECONÔMICO

Renato Martins Assunção (ENCE / IBGE)
Rosana de Freitas Castro (DEIND/IBGE)
José Carlos R.C.Pinheiro (ENCE / IBGE)

RIO DE JANEIRO
Junho/1988

A P R E S E N T A Ç Ã O

Este **RELATÓRIO TÉCNICO** inicia uma coleção destinada a apresentar os trabalhos de pesquisa elaborados por professores da ENCE, com eventual colaboração externa.

Executado no Laboratório de Estatística da Escola, que se destina, basicamente, à prestação de serviços técnicos a terceiros, o presente Relatório resultou de solicitação do DEIND/IBGE e estimulou a demanda por parte de vários outros órgãos do IBGE.

Os autores desta **CRÍTICA DE RAZÕES NO CENSO ECONÔMICO** contaram com a valiosa colaboração dos seguintes monitores do Curso de Graduação da ENCE: Denise da Cunha Otteiro, Renée Xavier de Menezes e Silvia Regina Nunes Constância.

DJALMA G. C. PESSOA
Diretor da ENCE

R E S U M O

O IBGE faz um trabalho de crítica dos dados fornecidos pelo respondente nos questionários do Censo Econômico. Uma parte deste trabalho procura determinar intervalos de aceitação para um conjunto de razões envolvendo diferentes campos do questionário para indústrias CE-301. A inovação mais importante na metodologia de crítica foi o emprego de curvas de rejeição ao invés de valores fixos, o que permitiu um tratamento diferenciado dos questionários de acordo com o porte dos estabelecimentos.

A B S T R A C T

IBGE Brazilian National Courses Bureau carries out data checking provided by respondents on the Economic Census questionnaires. Part of this work is setting up acceptance intervals to ratios set involving different items of the industry questionnaire CE-301. The most important methodological innovation was the use of rejection curves instead of fixed values, allowing a differencial treatment of questionnaires according to establishment's size.

CRÍTICA DE RAZÕES NO CENSO INDUSTRIAL DE 1985

SUMÁRIO

I - INTRODUÇÃO

Os Censos Econômicos de 1985, realizados pelo IBGE, têm o objetivo de investigar empresas e atividades industriais, comerciais, de serviços, de construção e de transportes, no tocante à sua atividade econômica.

As unidades básicas de levantamento são empresas e estabelecimentos, cuja definição é dada a seguir:

Empresa é a unidade jurídica caracterizada por uma firma ou razão social que engloba o conjunto de atividades econômicas exercidas em uma ou mais unidade local (UL) e responde pelo capital investido nestas atividades.

Estabelecimento é uma partição da UL para fins de levantamento estatístico, podendo, em certos casos, ser a própria unidade local. Caracteriza-se por ser uma unidade de operação, localizada em área contínua, pertencente a uma única empresa onde se desenvolve basicamente um tipo de atividade econômica.

Na fase de coleta de dados foram utilizados diferentes modelos de questionário, de acordo com as características da atividade sob investigação e da empresa como um todo.

Os questionários podem ser classificados em dois grandes grupos:

. questionários de empresa

- Modelo CE 001: aplicado na sede das empresas de médio e grande porte, contendo apenas dados sobre a empresa como um todo. É respondido pelas empresas que possuem faturamento igual ou superior a Cr$ 245.000.000,00 durante 1985, ou tenham mais de uma UL, e que estejam registradas no CGC.
- Modelo CE 002: versão simplificada, respondida pelas empresas de pequeno porte. Além de coletar dados da empresa, levanta informações das atividades nela desenvolvidas. É o único questionário respondido por estas empresas.

. questionários de atividades.

- Modelos CE 301, CE 401 e CE 501: aplicados aos estabelecimentos industriais, comerciais e de serviços, respectivamente, pertencentes às empresas que respondem ao CE 001.
- Modelos CE 601 e CE 701: aplicados aos departamentos de construção e transportes, respectivamente, das empresas cuja atividade principal é de natureza industrial, comercial ou de serviços.

Antes de serem tabulados, os dados levantados pela rede de coleta devem ser criticados, de modo a identificar e, quando possível, eliminar incorreções que possam vir a comprometer a validade dos resultados finais. Face às diferenças existentes entre os questionários, sistemas de crítica específicos foram elaborados para cada um deles. O modelo CE 301, que é o único do qual nos ocuparemos neste trabalho, passa por quatro fases de crítica, onde são identificados desde erros introduzidos na fase de transcrição, até inconsistências existentes entre os dados dos estabelecimentos e os da sede da empresa.

Este trabalho insere-se na terceira fase de crítica do modelo CE 301 e objetiva identificar questionários com valores aberrantes em uma ou mais de um conjunto de razões pré-definidas. O estabelecimento de intervalos de aceitação para cada uma das razões, levando-se em conta a disponibilidade existente de recursos e a importância relativa das razões, consiste no propósito central do estudo.

II - DEFINIÇÃO DAS RAZÕES

As variáveis de interesse neste estudo dividem-se em dois grupos: razões de batimento e razões-índices.

Existem três razões de batimento, a primeira referente ao consumo de matéria-prima, a segunda ao consumo de combustível e a terceira ao consumo de peças e acessórios. Aquela referente à matéria-prima é definida como

$$RMP = \frac{A}{B+C+D-E-F}$$

onde A= Consumo anual de M.P. em 85
B= Compras anuais de M.P. em 85
C= Recebimentos[1] de M.P. durante 1985
D= Estoque de M.P. em 31.12.84
E= Estoque de M.P. em 31.12.85
F= Transferências[1] efetuadas durante o ano de 1985

Os valores (em milhares de cruzeiros) das variáveis A,.....,F constam do questionário de cada estabelecimento industrial.

O nome "razão de batimento" baseia-se no fato de que a equação abaixo deveria se verificar:

A= B+C+D-E-F

isto é, os dois lados da equação deveriam "bater".

As definições das razões de batimento referente à combustível, RC, e peças, RP, são análogas a esta onde agora as variáveis A,...,F dizem respeito ao consumo, compras, etc... de combustível e peças.

Por uma série de motivos, as razões de batimento nem sempre são exatamente iguais à unidade. Entre estes motivos encontram-se os erros de preenchimento, inconsistência dos dados fornecidos pelo estabelecimento, erros de digitação não detectados na fase de crítica I, inflação, etc ...

---

(1) Recebimentos e transferências referem-se a movimentação de material entre estabelecimentos de uma mesma empresa.

Com relação a estas razões de batimento, o objetivo é identificar e rejeitar os questionários que têm pelo menos uma delas "muito grande" ou "muito pequena" em relação a 1. Para isto, precisamos definir o que significa ser "grande" ou "pequeno" e com isto estabelecer uma faixa de variação tolerável em torno de 1 para estas razões.

Observe por exemplo que, como

$$\frac{|A-(B+C+D-E-F)|}{B+C+D-E-F} = |RMP-1| \text{ e } \frac{|A-(B+C+D-E-F)|}{A} = \left|1 - \frac{1}{RMP}\right|$$

então RMP é função da diferença (relativa) entre A e B+C+D-E-F. Ao adotarmos a faixa de variação tolerável para RMP estaremos assegurando que os dados dos questionários não rejeitados tenham uma boa qualidade no sentido de apresentarem os erros relativos $|A-(B+C+D-E-F)|/A$ e $|A-(B+C+D-E-F)|/(B+C+D-E-F)$ pequenos. Entretanto, um erro relativo $\frac{|A-(B+C+D-E-F)|}{A}$ digamos, de 0.1 quando temos valores pequenos de A tem conseqüências muito diversas de quando temos valores grandes de A. Isto significa que a faixa de tolerância para RMP deve variar de acordo com o tamanho das variáveis A e B+C+D-E-F. Estas mesmas observações valem para as outras razões de batimento RC e RP. Resolver o problema exposto é o objetivo da metodologia apresentada nas seções III e IV .

Quanto às razões-índices, elas se referem a quatro características de cada estabelecimento:

- tempo (em meses) que o estabelecimento gasta para renovar o estoque de matéria-prima (giro de matéria-prima denotada GMP);

- tempo (em meses) que o estabelecimento gasta para renovar o estoque de produtos acabados, produtos de fabricação própria que estão estocados (giro de produtos acabados denotada GPA);

- tempo (em meses) que o estabelecimento gasta para renovar o estoque de produtos em curso, produtos de fabricação própria que ainda não estão totalmente manufaturados (giro de produtos em curso denotada GPC);

- margem de comércio definida como a proporção da receita com vendas de mercadorias em relação ao valor das compras destas mercadorias denotada MC (Obs.: estas mercadorias não são fabricadas pelo estabelecimento).

A variável giro de matéria-prima é definida como

$$GMP = \frac{(A + B)\ /2}{C/12}$$

onde A= estoque de M.P. em 31.12.84
B= estoque de M.P. em 31.12.85
C= valor do consumo (anual) de M.P.

Assim, $\frac{A + B}{2}$ é um índice do valor do estoque de M.P., $\frac{C}{12}$ é um índice do valor do consumo mensal de M.P. e portanto GMP é um índice do tempo (em meses) gasto para renovar o estoque de M.P.

As variáveis A,B,C constam do questionário do Censo Econômico. (CE 301)

De maneira análoga definem-se duas outras razões-índices. A variável giro de produtos acabados é

$$GPA = \frac{(A + B)/2}{C/12}$$

onde A= estoque de produtos acabados em 31.12.84
B= estoque de produtos acabados em 31.12.85
C= valor total da produção no ano.

As variáveis A,B,C constam do questionário do Censo Econômico.

A variável giro de produtos em curso é definida como:

$$GPC = \frac{(A + B)/2}{C/12}$$

onde A= estoque de produtos em curso em 31.12.84
B= estoque de produtos em curso em 31.12.85
C= valor total de produção no ano de 1985

As variáveis A,B,C constam do questionário do Censo Econômico.

A variável margem de comércio é definida como

$$MC = \frac{A}{B+C+D-E-F}$$

onde A= vendas de mercadorias para revenda durante 1985
B= compras de mercadorias para revenda durante 1985
C= recebimento de mercadorias para revenda durante 1985
D= estoque de mercadorias para revenda em 31.12.84
E= " " " " " " 31.12.85
F= transferências efetuadas de mercadorias para revenda durante 1985

Diferente das razões de batimento, as razões-índices não possuem um valor típico em torno do qual deveriam oscilar. Entretanto espera-se que cada uma destas razões-índices tenha uma distribuição de probabilidade bem definida e portanto, de posse desta distribuição, poderíamos definir um limite para os valores toleráveis das razões-índices. Assim, o objetivo seria determinar este limite de maneira que sejam rejeitados todos os questionários com razões-índices acima deste limite de tolerância.

CORREÇÃO DO EFEITO DA INFLAÇÃO

As variáveis que entram na definição das razões são afetadas pela inflação ao longo do ano e a correção desta influência deveria ser feita expressando todas as variáveis envolvidas em unidades monetárias de um mesmo mês.

Assim, escolhendo-se o mês de janeiro como o mês de referência, todas as variáveis referentes à estoque em 31.12.85 seriam deflacionadas utilizando-se um índice de inflação anual. Para as outras variáveis, referentes a valores agregados (tais como consumo anual de matéria-prima, compras anuais de matéria-prima, valor total da produção no ano, etc.) seria adotado o seguinte procedimento: o valor agregado total V corresponde a uma quantidade total Q de bens onde $Q=\sum_{i=1}^{12} q_i$ e q_i = quantidade de bens correspondente ao mês i. Então o valor agregado total V satisfaz $V=\sum_{i=1}^{12} V_i$ onde V_i=valor agregado correspondente aos q_i produtos do mês i. Se $q_i=q_1$ $\forall i$ então $V_i=I_i V_1$ onde I_i = índice da inflação acumulada no período de janeiro ao mês i e portanto

$$V = \sum_{i=1}^{12} V_1 I_i \qquad V_1 = \frac{V}{\sum_{i=1}^{12} I_i}$$

Denote $\sum_{i=1}^{12} I_i$ por I

O valor agregado anual corrigido pelo efeito de inflação seria então

$$V^* = 12V_1 = \frac{12\ V}{I}$$

Deste modo, as razões de batimento corrigindo-se os efeitos de inflação seriam da forma

$$R = \frac{\frac{12A}{I}}{\frac{12B}{I} + \frac{12C}{I} + D - \frac{12F}{I} - \frac{E}{I_{12}}}$$

e as razões-índices seriam da forma

$$R = \frac{\frac{1}{2}\left(A + \frac{B}{I_{12}}\right)}{\frac{C}{I}} \qquad \text{(para os giros de estoque)}$$

$$R = \frac{\frac{12A}{I}}{\frac{12B}{I} + \frac{12C}{I} + D - \frac{12F}{I} - \frac{E}{I_{12}}} \qquad \text{(para a margem de comércio)}$$

onde A,B,...,F são definidas da mesma maneira que antes em cada razão.

Esta correção da influência da inflação não foi implementada no programa de crítica, entre outros motivos, porque o efeito desta correção realizada nos dados do Censo de 1980 se mostrou pequeno.

Isto não é de se estranhar: para o ano de 1980, $I_{12} = 1.9532$ e $\sum_{j=1}^{12} I_j = 17.4004$ (usando o INPC) e portanto as razões índices referentes a giro de estoque corrigidas são da forma

$$GMP = \frac{\frac{1}{2}\left(A + \frac{B}{1.9532}\right)}{\frac{C}{17.4004}}$$

Como a razão não corrigida é da forma $\frac{\frac{1}{2}(A+B)}{C}$ então o quociente destas razões (corrigida e não-corrigida) é

$$q = \frac{\frac{\frac{1}{2}\left(A + \frac{B}{1.9532}\right)}{\frac{C}{17.4004}}}{\frac{\frac{1}{2}(A+B)}{\frac{C}{12}}} = \frac{\left(A + \frac{B}{1.9532}\right)(17.4004)}{(A+B)\,12}$$

Como $B \simeq (1.9532)A$ [1] então $q \simeq 0.98$. O mesmo cálculo para a inflação de 1985 mostrou que $q \simeq 0.90$. A superestimação que a razão não corrigida implica foi considerada pequena e acrescentando-se outros motivos operacionais foi decidido não corrigir o efeito da inflação.

TRANSFORMAÇÃO DAS RAZÕES DE BATIMENTO

As razões de batimento definidas anteriormente foram transformadas obtendo-se um novo conjunto de variáveis. Como será explicado abaixo, esta transformação teve basicamente o objetivo de eliminar a assimetria que teriam os limites de aceitação caso fossem utilizadas as variáveis iniciais.

Para estabelecer uma regra de crítica para as razões de batimento $R = \frac{A}{B+C+D-E-F} = \frac{A}{V}$ devemos obter uma faixa de aceitação tolerável para estas razões em torno do número um. Temos que se $A < V$ então $R \in (0,1)$ e se $A > V$ então $R \in (1,\infty)$ (excluímos os casos $A=0$ e $V=0$). Deste modo, a menos que $P(A < V) \gg P(V < A)$, a faixa de aceitação com limites L_1 e L_2 (com $L_1 < 1 < L_2$) deve ter $1 - L_1 < L_2 - 1$.

---

(1) ver "Correção da informação referente à estoque" à frente.

Para eliminar esta assimetria tomamos os valores absolutos dos logaritmos das razões de batimento isto é, definimos

$$R^* = |\log R| = |\log \frac{A}{V}| = |\log A - \log V|$$

e esta nova variável é simétrica em relação a A e V isto é, R^* é invariante para as duas possíveis definições de R, $R = \frac{A}{V}$ ou $R = \frac{V}{A}$. Além disso, ainda reflete uma variação relativa entre os valores A e V.

A nova variável R^* pertence ao intervalo $[0, \infty]$ e o valor "correto" desta variável é zero. Nosso objetivo passa a ser buscar um limite crítico c para esta variável de modo que se um questionário tem $R^* > c$ então ele será rejeitado.

CORREÇÃO DA INFORMAÇÃO REFERENTE A ESTOQUE

É muito natural esperar que as variáveis A e B sejam, fortemente relacionadas onde

A = estoque de matéria-prima em 31.12.84

e

B = estoque de matéria-prima em 31.12.85

Pode-se mesmo supor, para um estabelecimento fixo, que $B \simeq I.A$ onde I=índice de inflação durante 1985 e deste modo, esperava-se que log E deveria ser relacionado a log D através do modelo linear.

$$\log B = \log I + \log A + \varepsilon$$

Deve ser notado que a inclinação da reta de regressão é 1 e que o intercepto é o logaritmo do índice de inflação no ano.

O gráfico 1 do apêndice representa os valores dos pares (log B, log A) obtidos dos dados de 8.500 estabelecimentos industriais pesquisados no Censo Industrial de 1980.

O índice de inflação para o ano de 1980 foi 1.9532 e portanto log 1.9532 = 0.67 . Isto e mais a inclinação da "nuvem" de pontos mostram que nossa hipótese é bem plausível, a menos dos pontos que formam um "L" a partir da origem.

Verificou-se que este comportamento anormal detectado no gráfico explicava-se pelo início ou encerramento das atividades do estabelecimento durante o ano. São estes estabelecimentos que formam o "L" no gráfico 3. Para tomar os logarítimos de B e A, somou-se 1 em cada variável de modo a evitar o caso log (0).

Para estes estabelecimentos, o índice de giro de estoque de matéria-prima ($GMP = \frac{1}{2} \frac{A+B/I_{12}}{\frac{e}{\Sigma I_i}}$) não daria um indicador correto do tempo de renovação de estoques.

Deste modo, uma correção da informação referente ao estoque de matéria-prima se fez necessária. Então, para estes estabelecimentos que tenham A=0 ou B=0 esta correção consistiu em redefinir GMP como

$$GMP = \frac{\frac{1}{2} \text{ máx } (A , \frac{B}{I_{12}})}{\frac{C}{I_i + \ldots + I_j}}$$

onde A,B,C e I_{12}, tem o mesmo significado que antes, i=número do mês de início das atividades e j=número do mês de encerramento das atividades. Estas últimas informações adicionais também constam do questionário do Censo Econômico e como o número de estabelecimentos que necessitam ser corrigidos é relativamente pequeno, a correção não terá custo muito elevado.

Uma correção análoga deve ser feita nas razões-índices GPA e GPC. Deve ser notado que as razões de batimento e margem de comércio não necessitam desta correção.

CRITÉRIO DE REJEIÇÃO

Para a determinação dos limites críticos das regiões de variação tolerável para as razões devemos considerar dois aspectos que atuam em sentidos contrários. Se quisermos obter, ao final do processo de crítica, um alto nível de qualidade para os dados, seremos obrigados a tomar regiões de variação muito estreitas o que levará a uma grande proporção de questionários rejeitados. Entretanto, o

tempo e alto custo envolvidos na revisão destes questionários rejeitados atuam no sentido de restringir o nível de qualidade final desejado. Assim, é preciso decidir sobre o balanceamento entre o nível de qualidade final e os recursos disponíveis. Esta decisão é de natureza gerencial e a direção do Censo Econômico adotou, em face das limitações existentes, o nível de 5% do total de questionários como um limite máximo para a proporção de questionários a serem rejeitados no processo de crítica.

Como um questionário será rejeitado se pelo menos uma das sete razões consideradas for rejeitada então a limitação acima pode ser escrita como

$$0.05 = P \text{ (rejeitar um questionário)} = P \left(\bigcup_{i=1}^{n} A_i\right)$$

onde A_i = [rejeitar a i-ésima razão]

Isto é,

$$0.05 = 1 - P \left(\bigcap_{i=1}^{n} A_i^c\right) = 1 - P(\text{todas as razões "corretas"}) \quad (*)$$

Num estudo realizado com os dados do C.E. de 1980 verificou-se que a hipótese de independência entre as razões era bastante aceitável. Adotando-se esta hipótese, (*) torna-se:

$$0.05 = 1 - \prod_{i=1}^{n} (1-P(A_i)) \qquad (**)$$

Assim, os limites críticos para cada uma das sete razões (que determinam $P(A_i)$) devem satisfazer a restrição (**).

A determinação das probabilidades $P(A_i)$ foi baseada na importância relativa entre as razões. Isto significa atribuir pesos aos níveis de rejeição $P(A_i)$: quanto maior o peso, maior $P(A_i)$ o que implica num nível de qualidade maior para a razão i. A direção do C.E. decidiu que as razões referentes à matéria-prima, combustível, giro de produtos acabados e giro de produtos em curso teriam um percentual de rejeição cinco vezes maior que as razões referentes à giro de matéria-prima, lucro de revenda e peças. Deste modo, (**) torna-se

$$0.05 = 1-(1 - 5\theta)^4 \ (1 - \theta)^3$$

onde θ é o nível percentual de rejeição para GMP, MC e RP

Resolveu-se esta equação utilizando o método de Newton, encontrando $\theta = 0.0022 \Rightarrow 5\theta = 0.011$.

III - CURVAS DE REJEIÇÃO

Quando são empregadas razões entre variáveis para fins de crítica em um questionário, um aspecto fundamental da questão permanece usualmente encoberto: o da magnitude dos valores envolvidos nos cálculos. Tome-se, por exemplo, uma razão de batimento (ou seja, a resultante da divisão entre variáveis que deveriam apresentar o mesmo valor no questionário). Obviamente o valor correto deveria igualar a unidade, sendo necessário, para a crítica, estabelecer uma faixa de variação tolerável em torno deste número. Ocorre que, se para pequenos valores um erro relativo de até 10% pode ser considerado aceitável (em termos de sua influência sobre o agregado das informações, por exemplo), para valores elevados este mesmo percentual pode revelar-se desastroso.

Uma maneira de contornar esta dificuldade é através da combinação da informação associada ao erro relativo (representada, no exemplo anterior, pela razão), com alguma medida correlacionada com a magnitude dos valores utilizados no cálculo. Nesta seção, descreve-se uma abordagem metodológica de crítica que preenche estes requisitos (combinação dos erros relativos e absolutos), desenvolvida para a fase de crítica II do questionário 301 do CE-85. Saliente-se que a metodologia em questão não tem sua utilização limitada a razões de batimento, tendo sido empregada também para razões de outros tipos, como giros de estoque.

MOTIVAÇÃO DO ESTUDO

Dois aspectos necessitam ser considerados quando da elaboração de um plano de crítica: o nível final de qualidade desejado para os dados e a disponibilidade de recursos. A atuação dos dois dá-se em sentidos contrários, sendo de natureza mais gerencial do que técnica a decisão sobre o balanceamento entre ambos.

A motivação por trás do desenvolvimento da metodologia aqui descrita foi a de atingir o maior nível de qualidade possível para os dados, dentro da limitação de recursos existentes. A idéia básica é que as informações possuem importância diferenciada, em termos de influência sobre a qualidade final dos cálculos. Assim os recursos devem ser prioritariamente destinados à identificação de erros associados a valores de maior peso na composição dos agregados.

No caso particular da crítica de razões, essa linha de raciocínio conduz naturalmente ao estabelecimento de critérios diferenciados de rejeição, que levem em conta a grandeza dos valores utilizados nos cálculos. Essencialmente o que se pretende é maior flexibilidade com razões associadas a valores pouco significativos e maior rigidez com aquelas que envolvam valores expressivos.

METODOLOGIA

À princípio, considerou-se a alternativa de estratificação dos questionários segundo alguma medida correlacionada com as variáveis envolvidas nos cálculos das razões de interesse, seguida da especificação de valores críticos diferenciados por estrato. Estes últimos seriam estabelecidos de tal forma que a probabilidade de rejeição crescesse com a média dos valores assumidos pela medida de estratificação no estrato.

Considere-se, por exemplo, a razão de batimento associada a consumo e gasto com matéria-prima. Assuma-se que os questionários tenham sido divididos em três categorias: pequenos (60%), médios (30%) e grandes (10%) consumidores de matéria-prima e que seja tolerada a rejeição de até 1% dos questionários devido a esta razão. Uma questão fundamental para a aplicação desta abordagem é a definição da importância relativa entre os estratos, para fins de crítica. Uma maneira bastante natural de resolver este problema é através da atribuição de pesos relativos aos percentuais de rejeição por estrato. Suponha que, no exemplo em questão, tenham sido atribuídos pesos um, três e cinco, respectivamente, aos estratos dos pequenos, médios e grandes consumidores de matéria-prima. Isso significa que seriam rejeitados, em termos percentuais, três vezes mais questionários no segundo e cinco vezes mais no último estrato, que no primeiro. Neste exemplo particular, os percentuais de rejeição seriam: 0.5% no primeiro, 1.5% no segundo e 2.5% no último estrato. A regra de crítica ficaria estabelecida por completo com a determinação dos percentis das distribuições das razões nos estratos, associados aos percentuais de rejeição previamente mencionados.

A abordagem de crítica por estratos encerra, contudo, alguns problemas operacionais que podem dificultar sua utilização prática. O principal deles talvez seja o da construção dos estratos de questionários, consistindo-se em tarefa especialmente delicada a especificação da quantidade e limites dos mesmos. A metodologia desenvolvida neste estudo procura aproveitar a idéia central da abordagem por estratos, contornando, dentro do possível, as desvantagens a ela associadas.

DETERMINAÇÃO TEÓRICA DA CURVA DE REJEIÇÃO

Em certo sentido a metodologia de crítica por curvas de rejeição pode ser entendida como uma extensão da abordagem por estratos, com a importante diferença de que os questionários não necessitam ser estratificados, sendo considerados individualmente para fins do estabelecimento da regra de crítica.

O método pode ser resumidamente descrito da seguinte forma: a partir de uma medida de tamanho correlacionada com as variáveis que entram no cálculo das razões de interesse, determina-se uma função, denominada curva de rejeição, que fornece para cada questionário, um valor crítico. (Note-se que, a cada razão associa-se uma curva de rejeição). Traçando-se um paralelo com a abordagem anterior, os questionários poderiam ser considerados estratos unitários, adicionando-se, agora, a hipótese de existência de uma relação funcional entre os valores críticos dos estratos.

O procedimento de crítica para uma determinada razão R consiste então em:

(a) através da curva de rejeição e do valor da medida de tamanho observada no questionário obter o valor crítico;

(b) comparar o valor da razão no questionário com o valor crítico anteriormente obtido.

Duas funções, cuja notação é introduzida abaixo, são de fundamental importância para o desenvolvimento da metodologia de crítica utilizada:

$C_R(T)$ - a curva de rejeição associada à razão R, baseada na medida de tamanho T;

$P_R(T)$ - A probabilidade condicional de rejeição da razão R, dado o valor T para a medida de tamanho.

Existe uma certa dualidade entre as duas no sentido de que, conhecidas as distribuições de probabilidade de R e T, uma pode ser obtida a partir da outra. Na abordagem a seguir descrita, faz-se, a princípio, a determinação da função $P_R(T)$, a partir da especificação de um modelo e de um conjunto de restrições. A seguir, deriva-se a curva de rejeição associada, $C_R(T)$.

Pela discussão apresentada na seção anterior, a função $P_R(T)$ deve ser crescente em T, de modo a direcionar os recursos prioritariamente para os questionários de maior importância na composição dos cálculos. Afora esta, duas outras restrições devem ser obedecidas pela função $P_R(T)$.

$$0 < P_R(T) < 1 \quad (1) \quad e$$

$$E(P_R(T)) = \alpha \quad (2),$$

onde $E(.)$ representa a esperança matemática e α o percentual de rejeição associado à R.

A partir destas condições, arbitrou-se o seguinte modelo para a função de probabilidade condicional:

$$P_R(T) = Ae^{BT} + C \quad (3)$$

onde A, B e C são constantes a serem determinadas.

Note-se que pelas restrições mencionadas, as seguintes desigualdades devem ser satisfeitas:

$$A,B < 0 \quad (4);$$

$$1 > C > 0 \quad (5);$$

$$|A| < C \quad (6).$$

A obtenção das constantes A,B e C é feita a partir de um conjunto de equações derivadas das restrições citadas anteriormente:

$$P_R(t_0) = Ae^{Bt_0} + C = \delta_0\alpha \quad (7)$$

$$P_R(t_1) = Ae^{Bt_1} + C = \delta_1\alpha \quad (8)$$

$$E(P_R(T)) = AEe^{BT} + C = A\psi_T(B) + C = \alpha \quad (9).$$

onde, como antes, α representa o percentual de rejeição associado à R; t_0 e t_1 são dois valores pré-estabelecidos da medida T; δ_0 e δ_1 os pesos relativos de rejeição arbitrados para questionários de tamanhos t_0 e t_1 e ψ_T representa a função geradora de momentos de T.

Uma solução trivial para o sistema de equações acima, quando $\delta_0 = \delta_1 = 1$, é dada por B=0 e A + C = α. Como ficará claro no decorrer do texto, isso significa que a curva de rejeição assume um valor constante, igual ao percentil de ordem 1 - α da distribuição de R[1]. No presente estudo, ficar-se-á restrito aos casos em que $\delta_0 \neq \delta_1$ e B< 0, significando que $P_R(T)$ é estritamente crescente em T.

As equações (7), (8) e (9) podem ser reescritas na forma:

$$A = \frac{(\delta_1 - \delta_0)\ \alpha}{e^{Bt_1} - e^{Bt_0}} \quad (7');$$

---

1 - Está implícito, nesta afirmação, que a regra de crítica visa identificar apenas razões que excedam um determinado valor, não sendo fixado um limite inferior de variação.

$$C = [\delta_1 - \frac{(\delta_1 - \delta_0)\, e^{Bt_1}}{e^{Bt_1} - e^{Bt_0}}]\alpha \quad (8');$$

$$\frac{\psi_T(B) - e^{Bt_1}}{e^{Bt_1} - e^{Bt_0}} = \frac{1 - \delta_1}{\delta_1 - \delta_0} \quad (9')$$

Note-se que a constante B independe de α.

Uma simplificação importante é obtida tomando-se $\delta_1 = 1$ ou seja, assumindo-se que um questionário com tamanho t_1 terá a razão R rejeitada com probabilidade α. Neste caso, as equações se simplificam para:

$$A = \frac{(1 - \delta_0)\alpha}{e^{Bt_1} - e^{Bt_0}} \quad (7'')$$

$$C = \frac{(\delta_0\, e^{Bt_1} - e^{Bt_0})\alpha}{e^{Bt_1} - e^{Bt_0}} \quad (8'')$$

$$\psi_T(B) = e^{Bt_1} \quad (9'')$$

Observe-se, agora, que B depende apenas de t_1 e da distribuição de T.

Obtidos valores para A,B e C, o problema passa a ser a determinação da curva de rejeição $C_R(T)$.

Como visto nas seções precedentes, todas regras de crítica utilizadas no trabalho visam identificar razões que assumem valores por demais elevados, ou seja, são da forma:

"rejeitar o questionário quando $R > C_R$,"

onde C_R é o valor crítico associado à razão R. Na abordagem aqui apresentada, o valor crítico é dado pela função $C_R(T)$. Tem-se, então:

$$P_R(T) = P(R > C_R(T) \mid T) = 1 - F_R(C_R(T) \mid T) \quad (10)$$

onde $F_R(C_R(T)|T)$ é a função de distribuição de R condicionada em T.

Assumiu-se, no desenvolvimento do projeto, a hipótese de independência entre as razões de crítica e a medida de tamanho T.[1] A análise posterior dos dados do C.E. 80 indicou ser esta hipótese bastante aceitável. A equação (10) pode ser então simplificada:

$$P_R(T) = 1 - F_R(C_R(T)) \qquad (10')$$

Seque daí a relação fundamental para obtenção da curva de rejeição:

$$C_R(T) = F_R^{-1}(1 - P_R(T)) \qquad (11)$$

ou seja, o valor crítico $C_R(T)$ é dado pelo percentil de ordem $1-P_R(T)$ da distribuição de R.

ESTIMAÇÃO

O desconhecimento das distribuições de probabilidade das variáveis T e R torna necessário o uso de técnicas de estimação para derivação das curvas de rejeição. A distribuição de T é necessária para determinar as constantes A,B e C, que definem a função $P_R(T)$, enquanto que a de R é utilizada na obtenção da curva $C_R(T)$. A metodologia de estimação empregada é a seguir apresentada

- estimação de $P_R(T)$

A distribuição da variável T é fundamental para resolução da equação $(9'')$[2], que fornece o valor da constante B, da função $P_R(T)$.

$$\psi_T(B) = e^{Bt_1} \qquad (9'')$$

Dois procedimentos alternativos para estimação de B são apresentados. Um primeiro, de características não-paramétricas, utiliza a expansão da função geradora de momentos em série de Taylor e outro, baseia-se na especificação da distribuição de T.

. abordagem não-paramétrica

A hipótese básica aqui adotada é de que a variável T possui todos momentos finitos, podendo-se escrever:

$$\psi_T(B) = 1 + \sum_{k=1}^{+\infty} \frac{m(k)B^k}{k!} \qquad (12)$$

onde m(k) representa o k-ésimo momento da variável T.

---

[1] - Esclarecemos que esta hipótese não contradiz a idéia original da estratificação apresentada no início da seção Metodologia já que a medida de tamanho pode ser correlacionada com as variáveis que entram no cálculo das razões mas não com as próprias razões.

[2] - será assumida a simplificação mencionada, considerando-se $\delta_1=1$

O processo de estimação consiste em escolher um p suficientemente grande tal que:

$$\psi_T(B) \doteq 1 + \sum_{k=1}^{p} \frac{m(k)\, B^k}{k!} \qquad (13)$$

e substituir os parâmetros m(k) pelos momentos amostrais correspondentes, obtendo-se:

$$\hat{\psi}_T(B) = 1 + \sum_{k=1}^{p} \frac{\hat{m}(k)\, B^k}{k!} \qquad (14)$$

O estimador de B seria, então, dado pela solução da equação:

$$\hat{\psi}_T(B) = e^{Bt_1} \qquad (15)$$

que pode ser obtida numericamente através de métodos como o de Newton-Raphson, por exemplo.

A determinação do p adequado é feita iterativamente, em conjunto com a estimação de B. Define-se, inicialmente, um valor p_0, obtendo-se a estimativa associada $\hat{B}_0$. Toma-se, então, $p_1 = p_0 + 1$, com a respectiva estimativa $\hat{B}_1$. Prossegue-se até o passo n, onde se verifique a convergência das estimativas a um valor $\hat{B}$.

. abordagem paramétrica

Neste caso, é assumida uma distribuição $P_{\underset{\sim}{\theta}}$ para a variável T, com vetor de parâmetros $\underset{\sim}{\theta}$. Obtidas estimativas para $\underset{\sim}{\theta}$, encontra-se a estimativa paramétrica de B a partir da equação:

$$\psi_{T,\hat{\underset{\sim}{\theta}}}(B) = e^{Bt_1} \qquad (16)$$

Por exemplo, assumindo-se $T \sim N(\mu, \sigma^2)$ e representando-se por $\hat{\mu}$ e $\hat{\sigma}^2$ os estimadores da esperança e da variância, obtém-se o estimador $\hat{B}$ como solução de:

$$e^{\hat{\mu}B + \frac{\hat{\sigma}^2 B}{2}} = e^{Bt_1}$$

que é dada por:

$$\hat{B} = 2(t_1 - \hat{\mu}) / \hat{\sigma}^2$$

Nem sempre será possível derivar-se a solução fechada para a equação, tornando-se necessário o emprego de métodos numéricos de resolução. Considere-se, por exemplo, $T \sim \Gamma(\alpha,\beta)$ e sejam $\hat{\alpha}$ e $\hat{\beta}$ os estimadores associados. A equação (17) neste caso é dada por:

$$\left(\frac{\hat{\beta}}{\hat{\beta} - B}\right)^{\hat{\alpha}} = e^{Bt_1}$$

que não possui solução fechada.

Encontrada uma estimativa $\hat{B}$ para B, segundo uma das abordagens mencionadas anteriormente, obtêm-se estimativas para as constantes A e C por substituição de B por $\hat{B}$ nas equações (7'') e (8''). A função $P_R(T)$ será, então, estimada por:

$$\hat{P}_R(T) = \hat{A}e^{\hat{B}T} + \hat{C} \qquad (17)$$

- estimação de $C_R(T)$

A metodologia utilizada para estimação da curva $C_R(T)$ é de natureza essencialmente não-paramétrica, baseando-se em $\hat{P}_R(T)$ e na função de distribuição empírica de R, que será representada por $\hat{F}_R$.

A curva de rejeição foi definida em (11), pela relação:

$$C_R(T) = F_R^{-1}\,(1 - P_R(T))$$

Como já salientado anteriormente, os valores críticos $C_R(T)$ são dados pelos percentis de ordem $1 - P_R(T)$ da distribuição de R. A relação (11) pode ser alternativamente representada por:

$$F_R(C_R(T)) = 1 - P_R(T) \qquad (11')$$

ou ainda,

$$\begin{aligned} F_R(C_R(T)) &= \theta \\ 1 - P_R(T) &= \theta \end{aligned} \qquad (11'')$$

Em todo processo de estimação dispôs-se de uma amostra de 8500 questionários referentes ao Censo Industrial de 1980. Na amostra, a medida de tamanho T utilizada (no caso, o logarítmo do total de pessoal ocupado no estabelecimento) apresentava valores entre um mínimo (t_{min}) e um máximo (t_{max}). Assumindo-se a mesma variação para T no Censo de 1985, observa-se que só há necessidade de definir-se a curva $C_R(T)$ no intervalo $[t_{min}, t_{max}]$.

A partir da função empírica $\hat{F}_R$ e de $\hat{P}_R(T)$, podemos obter um conjunto de pontos $(T_p, \hat{C}_R(T_p))$ tais que para $p \in \mathbb{N}$,

$$[n.(1 - \hat{P}_R(T_{max}))] \leq p \leq [n.(1-\hat{P}_R(T_{min}))]$$

$$\hat{F}_R(\hat{C}_R(T_p)) = p/n \qquad (18)$$

$$1 - \hat{P}_R(T_p) = p/n$$

Note-se que, pela discussão anterior, só necessitamos determinar a curva $C_R(T)$ em $[t_{min}, t_{max}]$, o que equivale a obter as soluções das equações (11'') para θ no intervalo $[1 - P_R(t_{max}), 1-P_R(t_{min})]$. A limitação imposta sobre a variação de p faz com que p/n tome valores no intervalo $[1 - \hat{P}_R(t_{max}), 1 - \hat{P}_R(t_{min})]$. Fazendo $R_{(p)}$ representar a p-ésima estatística de ordem da variável R na amostra, obtêm-se as soluções:

$$T_p = \ln\left((1 - p/n - \hat{C}) / \hat{A}\right) / \hat{B} \qquad (19)$$

$$\hat{C}_R(T_p) = R_{(p)}$$

Assumiu-se, então, a hipótese de que os pontos $(T_p, \hat{C}_R(T_p))$ seguiam o mesmo comportamento de $(T, C_R(T))$, a menos de perturbações aleatórias. Mais especificamente, adotou-se o modelo:

$$\hat{C}_R(T) = \beta_0 e^{\beta_1 T} . \xi \qquad (20)$$

ou, equivalentemente,

$$\ln(\hat{C}_R(T)) = \ln(\beta_0) + \beta_1 T + \xi^* \qquad (21)$$

a curva de rejeição teórica, neste modelo, é dada por:

$$C_R(T) = \beta_0 e^{\beta_1 T} \qquad (22)$$

Os parâmetros β_0 e β_1 foram estimados por mínimos quadrados ordinários, no modelo linearizado, assumindo-se as hipóteses usuais sobre os resíduos ξ^*.

A curva de rejeição estimada, utilizada no processo de crítica é dada por:

$$\hat{C}_R^* (T) = \hat{\beta}_0 \, e^{\hat{\beta}_1 T} \qquad (23)$$

IV - RESTRIÇÕES DO TRABALHO

Durante a realização deste trabalho, um importante problema surgiu pela impossibilidade de acesso aos dados do Censo Econômico de 1985, que encontravam-se ainda em fase de digitação.

A alternativa adotada foi a de utilizarem-se os dados do questionário CE 301 referentes ao Censo de 1980.

Por serem as razões utilizadas na crítica variáveis adimensionais (razões de batimento) ou não-monetárias (razões-índice), a conversão dos valores para 1985 não apresentou maiores problemas. No entanto, foi necessário assumir-se a hipótese de que as distribuições de probabilidade das razões permaneceram inalteradas de um Censo para outro. Exceção foi feita à razão de margem de comércio, que mereceu tratamento diferenciado das demais.

Efetuou-se uma análise comparativa das distribuições das razões por gênero industrial e região do Brasil, não se constatando diferenças que justificassem o uso de regras diferenciadas de rejeição, por gênero ou região.

A impossibilidade de acesso aos dados de 1985 implicou também na escolha de uma medida de tamanho (T) de características não-monetárias, com vistas à estimação das curvas de rejeição definidas na seção anterior. A escolha recaiu sobre a variável "pessoal ocupado em 31/12/85", que será referenciada no restante do texto por PO. Novamente foi assumida a hipótese de igualdade entre as distribuições de probabilidade da variável para os Censos de 80 e 85.

Um derradeiro aspecto da pesquisa a ser ressaltado, diz respeito à mudança do âmbito de aplicação do questionário CE-301 do Censo de 1980 para o de 1985. No primeiro, o questionário era respondido por todos estabelecimentos de indústria, enquanto que no segundo o universo de informantes restringiu-se às empresas com faturamento igual ou superior a Cr$ 245.000.000,00, no ano de 1985, ou que possuíssem mais de uma unidade local. A compatibilização das informações foi efetuada por meio da seleção, no arquivo dos dados de 80, das empresas que teriam respondido o questionário CE-301 caso fossem adotadas as mesmas restrições utilizadas no Censo de 1985.

Nesta seção são apresentados os principais resultados obtidos a partir da aplicação da metodologia, apresentada nas seções anteriores, aos dados do Censo de 1980. Todos os cálculos foram efetuados com base em uma amostra sistemática de um terço do total de questionários disponíveis. O uso da amostra deveu-se não só à necessidade de diminuir-se o volume de informações, de modo a acelerar as rotinas de cálculo, como também ao interesse de validar-se os procedimentos de crítica em um conjunto independente de questionários, no caso os dois terços restantes.

. Medida de Tamanho (T)

Conforme mencionado na seção anterior, fez-se uso da variável PO (pessoal ocupado em 31/12/85) como medida de tamanho. Na verdade, a fim de reduzir-se a acentuada assimetria que caracteriza a distribuição dessa variável, tomou-se a transformação logarítmica de PO, doravante representada por LPO.

A análise do histograma da variável LPO, indicou que a distribuição gama ajustava-se bem aos dados, (ver gráfico 2). Os parâmetros α e β da distribuição foram estimados pelo método de momentos, tendo-se obtido:

$$\hat{\alpha} = 3.215$$

$$\hat{\beta} = 1.234$$

Desta forma, foi assumido que LPO seguia uma distribuição $\Gamma(3.625, 1.234)$.

. Probabilidade condicional de rejeição ($P_R(T)$)

A determinação de $P_R(T)$ é feita através da estimação dos três parâmetros que a definem (ver seção III). Inicialmente, procedeu-se à estimação do parâmetro B, que depende apenas da especificação da distribuição da medida de tamanho e do ponto t_1 para o qual deseja-se rejeitar a razão com probabilidade α. No caso, definiu-se este ponto como sendo a mediana de T, obtendo-se então:

$$t_1 = 2.83321$$

A constante B foi estimada a partir da equação:

$$\left(\frac{1.234}{1.234-B}\right)^{3.615} = e^{B.(2.83321)}$$

tendo-se, portanto, optado pela abordagem paramétrica de estimação, apresentada na seção III. A expressão acima é a equação (17) daquela seção, aplicada à distribuição $\Gamma(3.613, 1.234)$ e $t_1 = 2.83321$. Resolvendo-se por Newton-Raphson, obteve-se:

$$\hat{B} = -0.0846^{1}$$

Para estimação dos demais parâmetros, torna-se necessário definir o valor t_0, seu peso relativo de rejeição δ_0 e a probabilidade de rejeição não-condicional α. No caso, tomou-se:

$$t_0 = \min(T) = 0;$$

$$\delta_0 = 0,2.$$

após discussão com a gerência do Censo.

A probabilidade não-condicional de rejeição varia de razão para razão, fazendo com que os parâmetros A e C de $P_R(T)$ também variem com ela. Dois fatores respondem por esta variação, os pesos diferenciados de rejeição e os modelos assumidos para as distribuições das razões. Esta questão será abordada em detalhe no decorrer desta seção. Por ora, observe-se apenas que os parâmetros A e C podem ser expressos na forma abaixo:

$$A = A' \cdot \alpha_R$$

$$C = C' \cdot \alpha_R$$

onde A' e C' são os mesmos para todas as razões[2]. Substituindo-se os valores, definidos ou estimados anteriormente nas expressões de A' e C' obtém-se:

$$\hat{A}' = -3.754$$

$$\hat{C}' = 3.954$$

Assim sendo, a curva de rejeição estimada para uma razão R qualquer será dada por:

$$\hat{P}_R(T) = (-3.754\ e^{-0.0846T} + 3.954) \cdot \alpha_R$$

onde α_R representa o percentual de rejeição não-condicional para a razão R.

---

1 - A solução não-paramétrica indicou $\hat{B}=-0.0819$, com P=9, convergindo para uma tolerância de 10^{-7}.

2 - $A' = (1-\delta_0) \ / \ (e^{Bt_1} - e^{Bt_0})$

$C' = (\delta_0 . e^{Bt_1} - e^{Bt_0}) \ / \ (e^{Bt_1} - e^{Bt_0})$

. Percentuais de rejeição por razão (α_R).

A metodologia de determinação dos percentuais de rejeição por razão foi apresentada ao final da seção II. Faz-se menção, ali, ao uso de pesos diferenciados para os percentuais de rejeição, levando ao estabelecimento de dois grupos de razões: as muito e as menos importantes, sendo que as razões do primeiro grupo são rejeitadas com um percentual cinco vezes maior que opercentual das do segundo.

Afora este, um outro ponto contribui para a diferenciação dos percentuais de rejeição entre as razões: o próprio modelo assumido para a distribuição da razão. Foram utilizados dois tipos de modelo, um para as razões de batimento:

$$R_B = (X) + (1-X).R'_B \ ;$$

e outro para as razões-índice:

$$R_I = (1-X)R'_I$$

Em ambos os modelos X representa uma variável com distribuição de Bernoulli, com probabilidade p de sucesso igual ao percentual de questionários com valor igual a um (no caso das razões de batimento) ou zero (no caso das razões-índice). R'_B e R'_I representam respectivamente as componentes aleatórias das razões de batimento e índice. As regras de crítica são formuladas para as componentes aleatórias das razões, devendo os percentuais de rejeição dados inicialmente serem corrigidos para levar em conta a parte determinística do modelo. Assim, para uma razão qualquer R cujo percentual de rejeição α_R tenha sido obtido segundo a metodologia descrita na seção II, obtém-se o percentual de rejeição real α^*_R da seguinte forma:

$$\alpha^*_R = \frac{\alpha_R}{1-p}$$

onde, como antes, p representa o percentual de questionários com razões iguais a um (batimento) ou zero (índice).

Os percentuais α_R^* são aqueles utilizados nos procedimentos de crítica descritos nesta seção.

Os valores de α_R^* para as diferentes razões de crítica são dados abaixo:

razão	α_R	p	α_R^*
matéria prima	0.0110	0.6411	0.03065
combustível	0.0110	0.8663	0.0823
GPC	0.0110	0.8442	0.0706
GPA	0.0110	0.5514	0.0245
GMP	0.0022	0.3994	0.00366
peças	0.0022	0.6411	0.0061
margem de comércio	0.0022	0.8562	0.0153

. Curvas de rejeição ($C_R(T)$)

O procedimento de estimação das curvas de rejeição $C_R(T)$ a partir das funções $\hat{P}_R(T)$ e $\hat{F}_R$ (distribuição empírica da razão R) é descrito em detalhe ao final da seção III. Mencionou-se, então, a necessidade de estimar-se $C_R(T)$ apenas em um intervalo $[t_{min}, t_{max}]$ onde a medida de tamanho T toma valores com probabilidade um. No caso da variável LPO, este intervalo de estimação foi dado por [0,8.53621]. A hipótese por trás do uso deste intervalo é a de invariância da distribuição de LPO de um Censo para outro.

São apresentadas, a seguir, as curvas de rejeição estimadas para as diferentes razões de crítica. (ver também gráfico 3):

razão	$\hat{C}_R(T)$
matéria prima	$0.3193e^{-0.0729T}$
combustíveis	$0.2629e^{-0.028T}$
GPA	$7.3214e^{-0.097T}$
GPC	$2.5655e^{-0.155T}$
GMP	$30.468\ e^{-0.0348T}$
peças[1]	$0.3139e^{-0.0729T}$

1 - Face a inexistência de dados sobre compra e consumo de peças no Censo de 1980, utilizou-se a função estimada para a razão de matéria-prima também para esta razão.

A razão de margem de comércio recebeu um tratamento diferenciado das demais pela impossibilidade de assumir-se a hipótese de preservação de sua distribuição de probabilidade de um Censo para outro, já que esta é fortemente influenciadas pelos níveis de inflação vigentes, bastante diferentes em um período e outro. A metodologia de crítica utilizada para esta razão é descrita em detalhe ao final desta seção.

Após a análise das curvas de rejeição por parte dos técnicos do DEIND, procedeu-se à alteração do percentual de rejeição da razão associada à giro de matéria prima (GMP), já que os pontos críticos de rejeição foram considerados muito flexíveis (valores muito absurdos seriam aceitos). O percentual então utilizado foi o mesmo das razões consideradas importantes, ou seja, 0.011. O percentual global de rejeição foi elevado para 0,058 por esta modificação. A curva de rejeição estimada para GMP nestas condições foi $23.19e^{-0,102T}$, que levou a valores mais coerentes com as expectativas dos analistas do DEIND sendo, portanto, adotada.

. Validação dos Procedimentos

De modo a checar-se a validade da metodologia sugerida, submeteu-se o restante dos questionários não utilizados na etapa de estimação, aos procedimentos de crítica obtidos anteriormente. Determinou-se, então, o percentual de rejeição por razão e no total, tendo-se obtido os resultados abaixo:

razão	percentual de rejeição
matéria prima	0.015
combustível	0.009
GPA	0.012
GPC	0.014
GMP	0.009
TOTAL	0.056

Observe-se que o percentual de rejeição total não é soma dos percentuais individuais de cada razão, já que um percentual considerável dos questionários rejeitados o é em mais de uma razão. Os resultados referentes à razão de consumo de peças não puderam ser calculados já que não havia informações sobre a variável no Censo de 1980.

A análise dos percentuais obtidos indicou que a metodologia de crítica sugerida atingiu os objetivos aos quais se propunha de forma satisfatória. Como era esperado, observaram-se oscilações do percentual de rejeição real em torno do valor teórico, mas em magnitude tal que não compromete os objetivos fixados para os procedimentos de crítica. Note-se que o percentual total de rejeição ficou um pouco acima do esperado (0.058), já que o valor acima (0.056) deve ser aumentado pelas rejeições devido à margem de comércio e ao consumo de peças.

. **Margem de Comércio**

A construção de limites críticos para a variável margem de comércio (MC) foi feita de maneira diferente das demais.

Esta variável, medida em cada estabelecimento industrial, refere-se a uma atividade secundária: venda de produtos não-industrializados no estabelecimento em questão,isto é, a variável refletia o lucro obtido com revenda de produtos. Há um pequeno número de estabelecimentos que exercem esta atividade marginal a qual era muito pouco conhecida na época deste trabalho. Um dos motivos de incluí-la no procedimento de crítica era garantir dados de boa qualidade para estudos exploratórios deste aspecto do funcionamento dos estabelecimentos.

Por motivos variados não era de interesse uma crítica desta variável com base em curvas de rejeição. Buscava-se portanto apenas um intervalo de aceitação simples para os valores de MC em 1985.

O escasso conhecimento sobre a variável não incluía a influência da inflação ao longo do tempo. A característica de lucro numa atividade marginal pouco conhecida dava bases pouco sólidas ao pressuposto da invariância da variável entre períodos de tempo com níveis de inflação muito diferenciados. Assim, utilizar intervalos de aceitação calculados com dados de 1980 para a crítica dos dados de 1985 parecia um pouco arriscado.

A alternativa adotada foi obter uma pequena amostra desta variável em questionários CE-301 do Censo de 1985. Devido a dificuldades de acesso aos dados de 1985, a amostra deveria ser obtida com uma restrição: só estavam disponíveis questionários de

uma população de estabelecimentos pertencentes à Coleta Especial[1]. Grosseiramente, esta Coleta Especial (CE) referia-se a um grupo de empresas, em geral de grande porte, que recebiam tratamento diferenciado em todo o processo de realização do Censo.

Para descrever o procedimento vamos definir:

POPA = população A = Conjunto de valores da variável MC no Censo do ano A em todos os estabelecimentos que exercem a atividade de revenda.

SUBPOPA = Conjunto de valores da variável MC no Censo do Ano A em todos os estabelecimentos da CE que exercem a atividade de revenda.

Usando-se os dados de 1980 estimou-se a relação entre os valores de POP80 e SUBPOP80:

.seja POP80= $\{Y_1, \ldots Y_N\}$

e SUBPOP80 = $\{z_1 \ldots, z_n\} \subset$ POP80

. tomou-se os n percentis $Y_{(k)}$ de ordem $(K.\frac{N}{n})$ 100% de POP80

onde K= 1, ..., n. Fez-se uma regressão das estatísticas de ordem $z_{(1)}, \ldots z_{(n)}$ de SUBPOP80 em $Y_{(1)}, \ldots, Y_{(n)}$

obtendo-se $\hat{z}_{(k)} = A + B\, Y_{(k)}$ K=1,...,n

e outra regressão de $Y_{(k)}$ em $y_{(k)}$ obtendo-se

$$\hat{Y}_{(k)} = A' + B' z_{(k)} \quad K= 1,\ldots,n$$

---

(1) Estes questionários também não estavam digitados. Apesar de toda a subpopulação estar disponível foi preciso retirar uma pequena amostra pois os dados tinham de ser obtidos manualmente diretamente do questionário a um custo altíssimo.

. Estas relações usam 2 grupos de dados de um mesmo ano, 1980, e portanto entre os grupos não há diferenças devido à inflação.

. Foi feita a hipótese de que as relações lineares acima não mudaram em 1985.

Para lidar com a mudanca da distribuição de 80 para 85 utilizou-se SUBPOP80 e a amostra de SUBPOP85:

. Sejam $X_1, \ldots, x_m$ os valores ordenados obtidos da amostra de SUBPOP85.

. Sejam $z_1' < \ldots < z_m'$ os m percentis empíricos de SUBPOP80 $(n > m)$.

. Fez-se uma regressão de $z'_1, \ldots, z'_m$ em $x_1, \ldots, x_m$ obtendo-se

$$\hat{z}'_i = a+bx_i \quad i=1,\ldots, m$$

Foi obtida uma relação entre POP80 e POP85:

.Sejam C_1 e C_2 os percentis de ordem $\frac{\alpha}{2}$ 100% e $(1-\frac{\alpha}{2})100\%$ respectivamente de POP80.

Estimou-se por C_1' e C_2' os percentis de ordem $\frac{\alpha}{2}$ 100% e $(1-\frac{\alpha}{2})100\%$ respectivamente de POP 85 como

$C_1' = A' + B' (a+b (A + BC_1))$

$C_2' = A' + B' (a+b (A + BC_2))$

O intervalo $[C_1', C_2']$ foi então utilizado como intervalo de aceitação para a variável margem de comércio no processo de crítica do Censo 85. Espera-se que este intervalo rejeite aproximadamente α 100% dos valores da variável.

-- x --

NOTA POSTERIOR: Até 26/09/88 foram criticados 41% dos questionários do Censo Industrial de 85 e os percentuais de rejeição de questionários em cada razão está de acordo com o esperado.

APÊNDICE

Os gráficos aqui incluídos ajudam a esclarecer determinados tópicos apresentados durante a exposição da metodologia e dos resultados do trabalho

GRAFICO 2 - Histograma da medida de tamanho - LPO

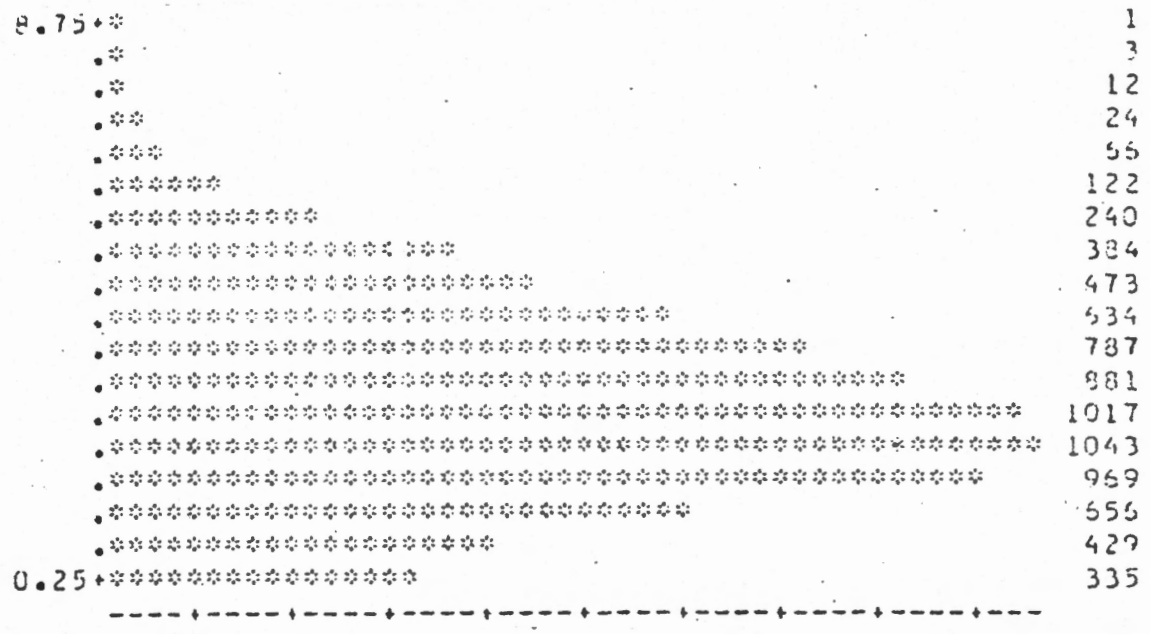


GRÁFICO 3 - Exemplo de curva de rejeição (no caso referente ao giro de matéria prima - GMP).

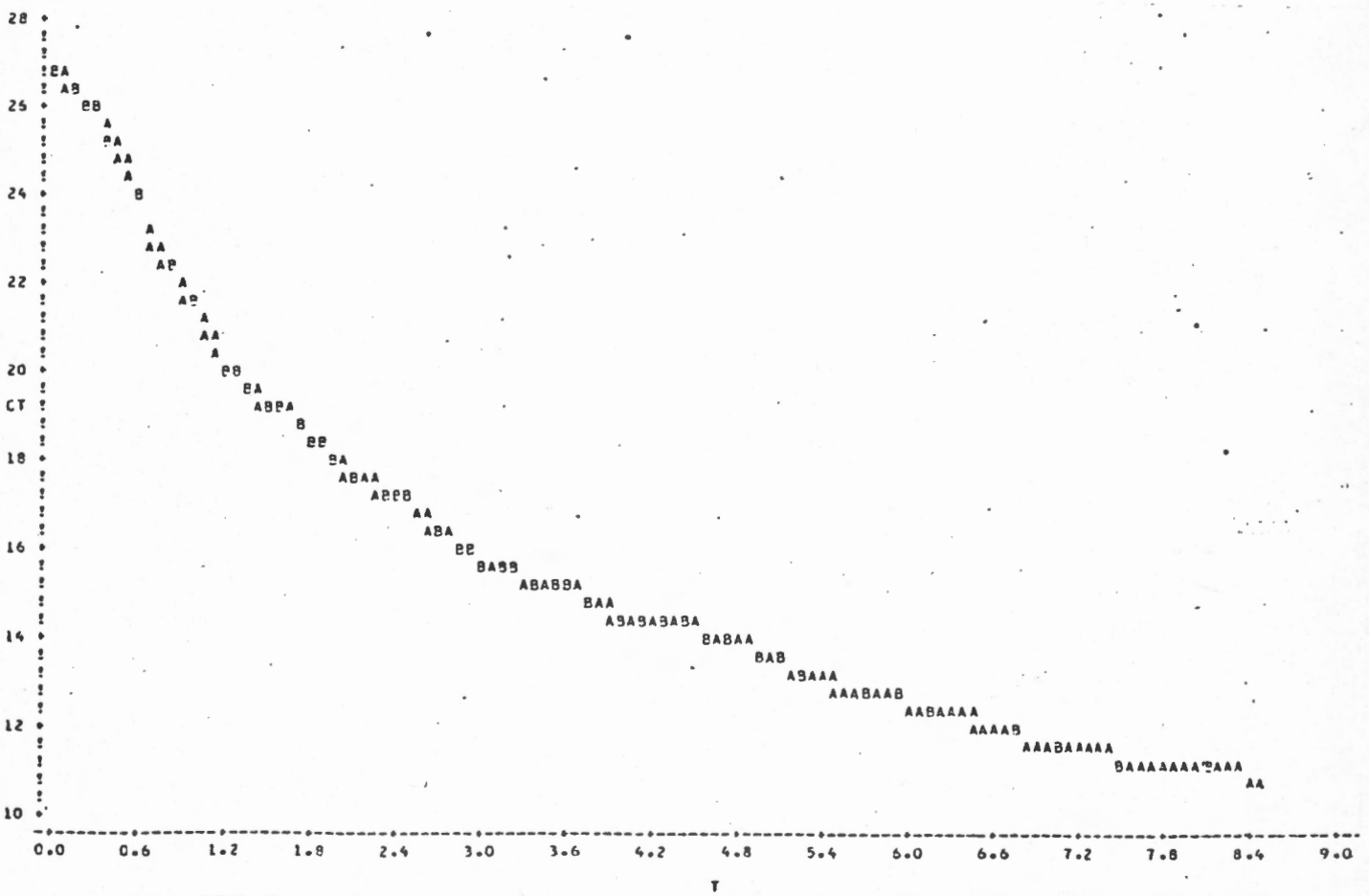
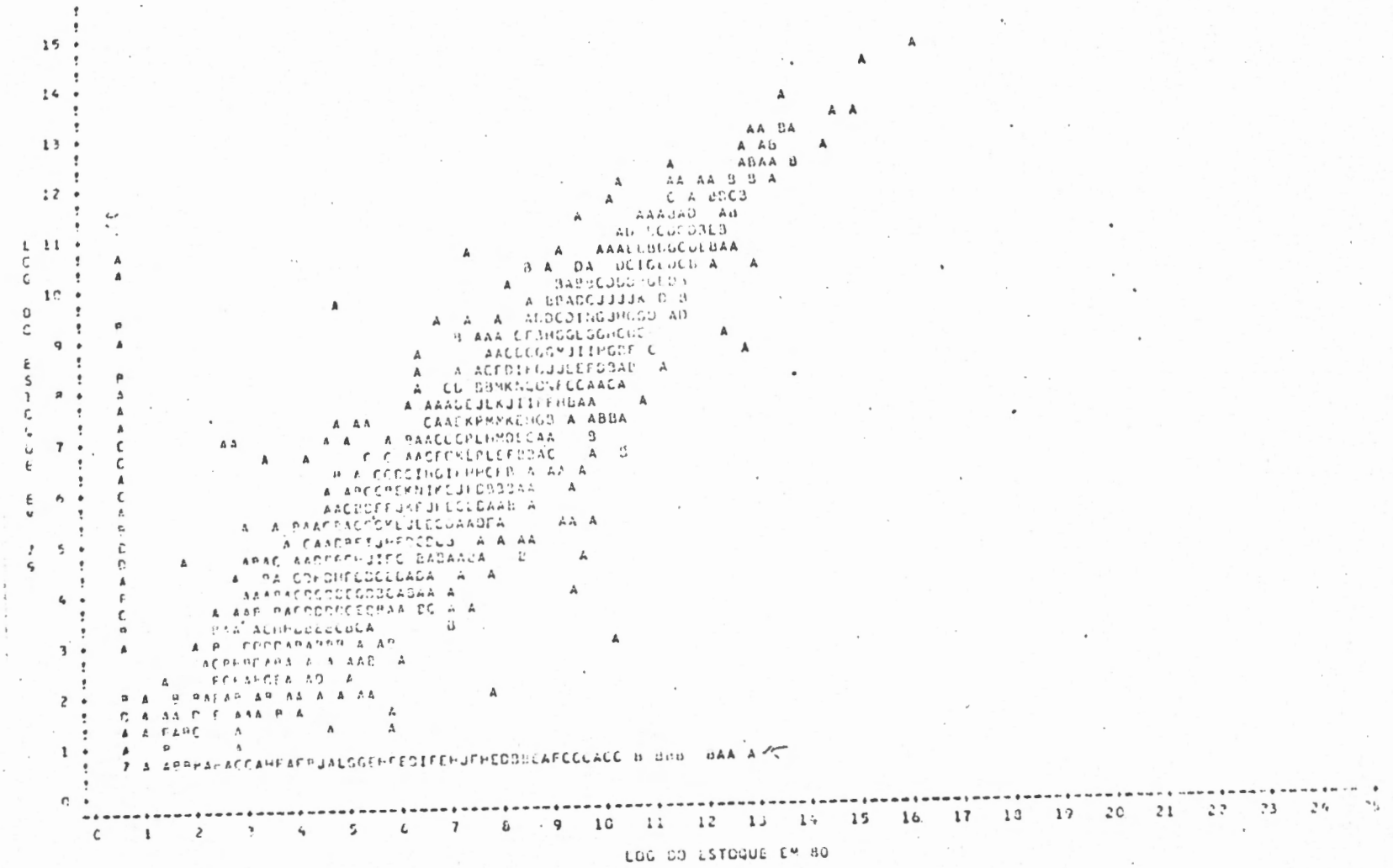


GRAFICO 1 - Relação entre os logaritmos dos valores nos estoques em 31/12/79 e 31/12/80.



REFERÊNCIAS

- ABRAMOWITZ, M. and Segun, I.A.(editors) (1968) Handbook of mathematical functions, Dover Publication, Inc., New York.

- JOHNSON, N.L. and Kotz, S. (1970) Continuous univariate distributions vol.1 e 2 Houghton Mifflin Company, Boston

- SAS User's guide: Basics, Version 5 edition (1985) SAS Institute Inc., Cary, NC