ELEMENTOS DE AMOSTRAGEM PROBABILÍSTICA

APLICADA AO CENSO*

*Curso ministrado, sob os auspícios da Administração do Ponto IV no Brasil, pelo sr. THOMAS B. JABINE, técnico do "Bureau of the Census" dos Estados Unidos da América. Anexos compilados sob a orientação do sr. THOMAS B. JABINE. Tradução do inglês de ANNA STERENBERG, revista por VINICIUS FONSECA.

PREÂMBULO

A série de nove palestras que se segue foi apresentada ao pessoal do Núcleo de Planejamento Censitário, entre julho e novembro de 1956. As palestras referem-se à amostragem probabilística e suas aplicações nos trabalhos censitários, com um emprêgo mínimo de matemática em sua apresentação. Não pretendem dar ao leitor tôda a teoria necessária ao planejamento de uma amostra e ao cálculo de estimativas e de variâncias. Sua finalidade é a de informar sôbre o que é a amostragem probabilística, como pode ser usada nos censos e inquéritos e, acima de tudo, como interpretar os resultados de uma amostra em vista dos erros de amostragem a que estão sujeitos.

Foram incluídos nas palestras vários anexos, contendo gráficos, exemplos, experimentos e outros materiais utilizados com o fim de ilustrar os princípios teóricos estudados. Há vários apêndices especiais. O primeiro, uma tradução do artigo " A Amostragem no Censo de População e Habitação de 1950", de Joseph Steinberg e Joseph Waksberg, aos quais devemos a permissão para traduzir e usar o trabalho. O segundo, o relatório de uma amostra experimental selecionada dos questionários do Censo Demográfico do Brasil de 1950; êste trabalho foi hábilmente dirigido pelo Sr. Heitor da Câmara Vellozo, com a assistência técnica do autor. Segue-se um relatório sôbre duas amostras experimentais, selecionadas dos questionários do Censo Comercial do Brasil de 1950, relativos ao Distrito Federal. Êste projeto foi realizado, e o relatório preparado, pelo Sr. Rudolf W. F. Wuensche, sob a orientação técnica do autor. O quarto apêndice consiste da tradução de excertos da mais recente publicação oficial dos Estados Unidos sôbre o assunto, que descrevem os processos de amostragem utilizados no Inquérito Periódico de População do "Bureau do Censo". Finalmente, há um vocabulário de têrmos e expressões utilizados em amostragem, às instruções e resultados de dois experimentos realizados durante o curso, e uma bibliografia.

O autor deseja expressar sua gratidão às seguintes pessoas, cuja contribuição ou cooperação lhe foi de grande valia no preparo dêste trabalho. Em primeiro lugar, o Dr. Armando Rabello, Diretor do Núcleo de Planejamento Censitário, cuja antevisão da utilidade potencial das técnicas de amostragem no trabalho censitário o levou a solicitar a realização desta série de palestras aos seus funcionários; Dr. Henrique Maia Penido, do Serviço Especial de Saúde Pública, e Dr. E. Ross Jenney, da Divisão de Saúde e Saneamento do Instituto de Assuntos Interamericanos, por terem possibilitado ao autor dedicar seu tempo a êste empreendimento; Sr. Vinícius Fonseca, do N.P.C. por ter feito a coordenação necessária às reuniões do grupo e também por suas críticas hábeis e muitas sugestões úteis; Sr. Moysés Kessel, do I. B.G.E., e Dr. Jacques Noel Manceau, do Serviço Especial de Saúde Pública, por seu excelente trabalho como intérpretes durante as palestras; Srs. Martiniano Barbosa

Moreira e Herbert Wilkes Júnior por seu inestimável auxílio na reunião e preparação do material publicado; Sr. William Gelabert e seus colegas dos Serviços Audio-Visuais do Instituto de Assuntos Interamericanos, que construiram um aparelho muito engenhoso a fim de demonstrar o comportamento de amostras repetidas; Sra. Anna Sterenberg, do Núcleo, por um ótimo trabalho de tradução em setor altamente técnico; Sra. Lea Rotberg Costa, do SESP; Sr. Mário de Andrade Medeiros e Sra. Maria Alice Martins Secco do N.P.C., que datilografaram os esboços preliminares e finais das palestras.

Finalmente, merecem agradecimentos os membros do Núcleo por sua amável recepção, sua paciência quanto às dificuldades de tradução, e pelo interêsse demonstrado pela aplicação destas técnicas ao seu trabalho. Desta forma recompensaram grandemente os esforços para a preparação dêste material.

Thomas B. Jabine

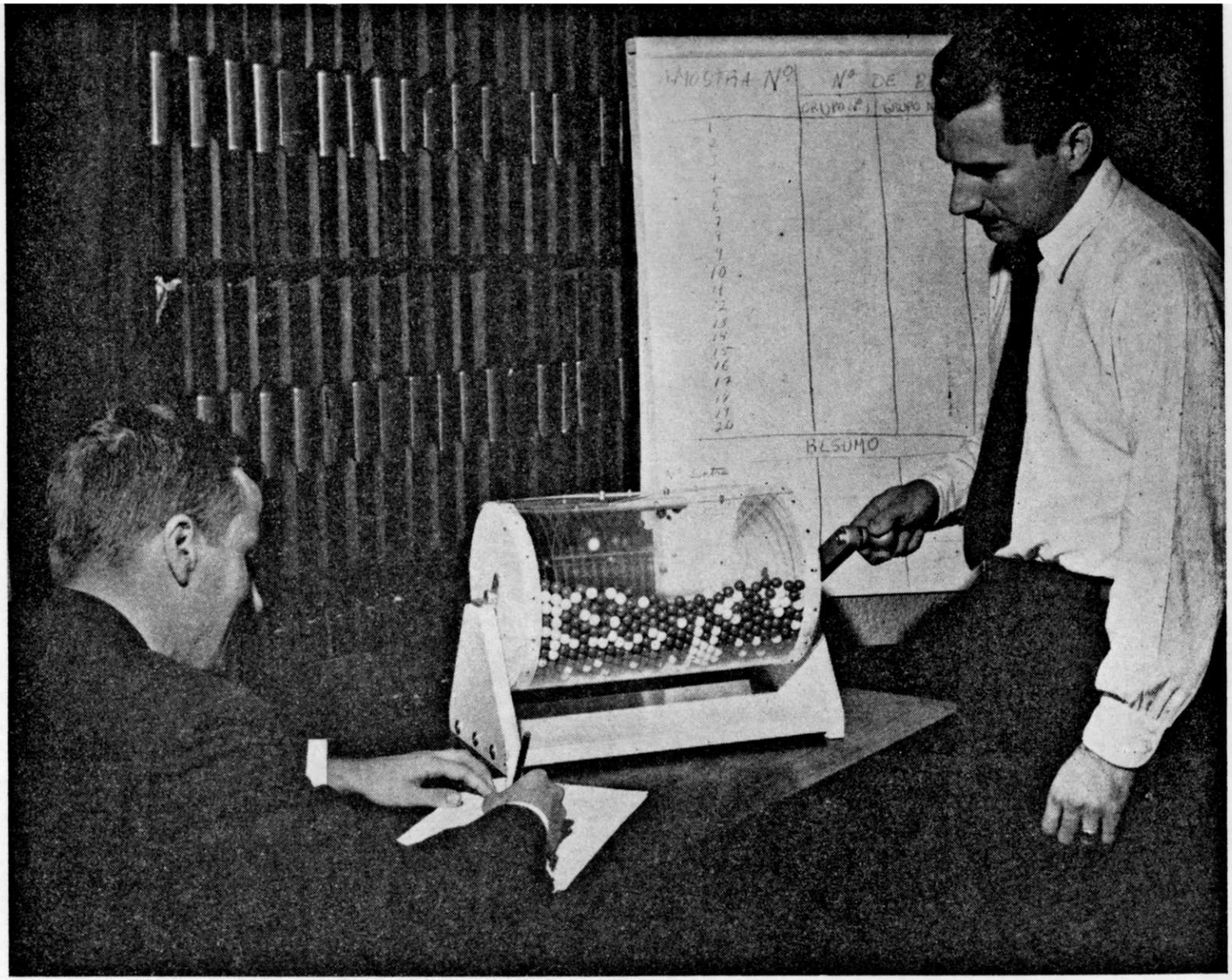


Figura n.º 1 — O "amostrador" usado para os experimentos.

ERRATA

Pág.	Linha	Onde se lê	Leia-se
6	32	a) implica repetição da experiência	a) implica repetição do experimento
2ª Palestra Anexo 1 1ª pág.	17	Permutações, combinações e arranjos	2. Permutações, combinações e arranjos
2ª Palestra Anexo 1 1ª pág.	26	(n ≤ n)	(m ≤ n)
2ª Palestra Anexo 1 2ª pág.	27	Probabilidades compostas	4. Probabilidades compostas
2ª Palestra Anexo 1 3ª pág.	32	P (A + B)= 1/36 + 3/36 = =3/36 = 1/12	P (A + B)= 1/36 + 2/36= =3/36 = 1/12
2ª Palestra Anexo 1 4ª pág.	8	retirada não é retirada não é resposta	retirada não é resposta
11	21	ou, então, quando o universo é limitado	ou, então, quando o universo é ilimitado
14	8	2ª - Qual	2 - Qual
17	21	(V. Anexo B)	(V. Anexo 2)
20	22	Também é possível computar	Também é possível calcular
21	2	duas famílias:	duas famílias, sem reposição:
4ª Palestra Anexo 2	-	-	No final do 4º problema: "Êste é um bom processo?"
24	-		Na indicadora do primeiro quadro: 0,00 a 0,09 0,10 a 0,19 0,20 a 0,29 0,30 a 0,39 0,40 a 0,49 0,50 a 0,59 0,60 a 0,69 0,70 e mais

Pág.	Linha	Onde se lê	Leia-se
24	Cabeçalho do segundo quadro	ESCALA DE VALORES	INTERVALOS DE VALORES
24	Rodapé do segundo quadro	... estimativas que caem em escalas especificadas ...	... estimativas que caem em intervalos especificados ...
24	13	... de que a estimativa caia nas escalas mais reduzidas.	... de que a estimativa caia nos intervalos mais reduzidos.
25	24	(ou cêrca de (2/3 σ)	(ou cêrca de 2/3 σ)
25	36	(ou 510 ± (10)	(ou 510 ± 10)
5ª Palestra Anexo 2 2ª pág.	18	de declaração ou emissão de declaração	de declaração ou omissão de declaração.
5ª Palestra Experimento nº 1	16	$\sigma \hat{p} = \frac{\sqrt{0,30 \times 0,70}}{100} = 0,0458$	$\sigma\hat{p} = \sqrt{\frac{0,30 \times 0,70}{100}} = 0,0458$
28	1	(Anexo 1)	(Anexo 1, à página 33)
28	3	$\sigma^2 = \frac{1}{6} \times 10,650 = 1775$; $\sigma = 42$	$\sigma^2 = \frac{1}{6} \times 20,650 = 3442$; $\sigma = 58,6$
28	9	(Anexo 1)	(Anexo 1, à página 33)
28	15	70% (de 42 para 13)	80% (de 59 para 13)
32	14	(estabelecimentos A, B e C do exemplo)	(estabelecimentos A, B, E e H do exemplo)
32	33	de representar o Brasil,	de representar a população do Brasil,
33	Quadro 1 Linha Total Col. (6)	10 650	20 650
40	17	= Pessoas da amostra com a característica A / Pessoas recenseadas na amostra - x	= Pessoas da amostra com a característica A / Pessoas recenseadas na amostra x
41	4	- Total de pessoas da amostra	± Total de pessoas da amostra
44	17	da amostra por 20,4).	da amostra por 19,6).

Pág.	Linha	Onde se lê	Leia-se
Bibliografia 1ª pag.	Item I, C, 2	Mostelles	Mosteller
Bibliografia 2ª pag.	Item II, 2	Mostelles	Mosteller

ÍNDICE

Anexos

PRIMEIRA PALESTRA

Espero que êste curso atenda às vossas necessidades. Embora tenha conferenciado sôbre vários aspectos da amostragem no Bureau do Censo dos Estados Unidos, a muitos grupos de estagiários e visitantes de todo o mundo, nunca tentei antes abranger todo o campo da amostragem. Felizmente existem agora diversos livros e outros documentos que facilitam a tarefa. Êste progresso é recente - o primeiro livro relativamente completo sôbre amostragem foi publicado em 1949 (Yates). Hoje, existem pelo menos cinco: Yates; Deming; Hansen, Hurwitz e Madow; Cochrane; Sukhatme. Além disso, há uma série crescente de trabalhos em revistas estatísticas e outras publicações.

Baseado no meu conhecimento dêsse material e em seis anos de experiência no Bureau do Censo, trabalhando com alguns dos pioneiros no campo da amostragem, tentarei esboçar uma introdução à amostragem que sirva a vossas necessidades o máximo possível. A fim de ser bem sucedido, necessito de vossas críticas e sugestões sôbre o conteúdo do curso e o método de exposição. Acolherei assim as perguntas, em qualquer ocasião. Isto é particularmente importante por serem as palestras pronunciadas em inglês, exigindo tradução. Não há nada mais desanimador do que falar a um grupo que apenas ouve, pois isto geralmente significa dificuldade de entendimento ou falta de interêsse da parte do ouvinte. Com a vossa cooperação poderemos evitar que se chegue a esta situação. Fiquei impressionado com o que o Dr. Rabello me contou sôbre a maneira pela qual o Núcleo funciona como equipe. Se nos dedicarmos ao curso, com êste espírito, certamente realizaremos algo de valioso.

Espero que, além de proveitoso, êste curso também seja agradável. Para mim, desde que aprendi algo a respeito, a amostragem tornou-se um assunto fascinante. É empolgante o conhecimento que se pode ter das características de um grupo, com o grau de precisão desejado, estudando uma amostra dos indivíduos que o constituem. Dessa maneira torna-se possível obter maior número de dados, com economia dos recursos utilizados. Nenhum país desejaria onerar-se mais de uma vez, em 5 ou 10 anos, com a realização de um censo completo. Com o emprêgo da amostragem, no entanto, torna-se possível obter os mesmos dados (pelo menos para o país em conjunto, e grandes áreas geográficas) anualmente, trimestralmente, ou mesmo mensalmente. Além disso, êsses dados podem ser obtidos muito mais ràpidamente do que em um censo completo, devido ao menor volume das apurações.

Há também algo de fascinante no fato de fenômenos que aparentemente não se enquadram a um padrão perceptível, quando examinados individualmente, tornarem-se previsíveis em seu comportamento, quando encarados em um grupo. Desde que todos trabalhais com estatística, estais inteirados dêsse fato. Creio que, trabalhando com amostras, podereis apreciar alguns exemplos marcantes.

Aquêles, de vós, que já se manifestaram sôbre o nosso "seminá - rio" pediram um mínimo de teoria e um máximo de aplicação prática. Estou satisfeito por terdes êste desejo, e gostaria de dizer porque.

A amostragem é um instrumento a ser usado no trabalho censitá - rio, e poderia compará-la, neste sentido, ao equipamento de tabulação IBM.

O técnico censitário não precisa saber como armar um painel, como manejar uma classificadora, como consertar uma tabuladora que não funciona adequadamente. Para êsses trabalhos, dispõe de um técnico.

De maneira semelhante , o técnico censitário que usa a amostragem não é obrigado a planejar com detalhes a amostra mais eficiente, dada uma determinada situação; a deduzir fórmulas para os erros de amostragem; ou a aplicar processos complicados de estimativas. Para isso, precisa de um técnico, um perito em amostragem, com antecedentes em estatística matemática.

O técnico de IBM, entretanto, não é quem decide sôbre as tabulações que devem ser feitas, as máquinas a utilizar, a ordem das tabulações, etc. Cabem ao técnico censitário estas questões importantes, levando em conta seu amplo conhecimento e experiência, com relação a todos os fatôres envolvidos.

De maneira semelhante , como técnicos censitários, mais cedo ou mais tarde enfrentareis problemas que envolvam a magnitude e a maneira pela qual as técnicas de amostragem devam ser utilizadas nas operações que dirigis. Não tenho dúvida a respeito, porque as vantagens da amostragem são tão grandes que o seu emprêgo está ganhando rápida aceitação em todo o mundo, particularmente em repartições estatísticas governamentais.

Podeis ser solicitados a resolver questões como as seguintes:

a) Distribuindo uma amostra de certa maneira entre várias áreas, obtereis resultados com erros de amostragem determinados . Êstes erros são bastante pequenos, ou seria preferível gastar uma certa quantia de dinheiro a fim de reduzí-los de 10%? Se dispuzerdes de um orçamento fixo, preferirieis redistribuir a amostra, a fim de reduzir o êrro em certas áreas, aumentando-o em outras?

b) Propõe-se que a codificação dos questionários seja verificada a base de amostra, e apenas aquêles lotes em que o índice de erros ultrapasse determinados limites sejam verificadas por completo. Sugere-se um plano de amostragem que rejeite lotes como os seguintes:

Taxa de êrro no lote inteiro (%)	Probabilidade de rejeição
0 - 1	0,05
1 - 2	0,50
2 ou mais	0,99

Será êste plano aceitável?

c) Selecioncu-se uma amostra que fornece dados com precisão sa-

tisfatória para Estados e Regiões. Também podem deduzir-se dados por Município, mas terão erros de amostragem relativamente grandes: devem ser tabulados e publicados êsses dados?

A fim de responder a perguntas como estas, tereis de compreender o conceito de êrro de amostragem. Que significa dizer que uma certa estimativa apresenta um coeficiente de variação de 5%? Quanto vale reduzir êste êrro para 4%, ou 2%?

Genèricamente, desejareis saber por outro lado que passos devem ser dados quando se usa a amostragem, que precauções especiais devem ser tomadas na análise e publicação de dados nela baseados, quais os diferentes métodos de amostragem e as condições em que são usados, e, acima de tudo, quais as aplicações da amostragem no trabalho censitário.

É meu objetivo apresentar a matéria com um mínimo de teoria e um máximo de aplicação. Não obstante, é impossível estudar as aplicações práticas inteligentemente sem compreender alguma coisa das teorias de que dependem.

Neste curso, falaremos principalmente sôbre a amostragem probabilística. Para a maioria dos objetivos, êste é o único tipo de amostragem aceitável, fato que tentarei demonstrar posteriormente.

Amostra probabilística é aquela cujos resultados podem ser interpretados de acôrdo com a teoria da probabilidade. Por esta razão, dedicarei as duas seguintes palestras a uma descrição daqueles pontos da teoria da probabilidade importantes para a amostragem probabilística. Tentarei fazê-lo sem utilizar complexos dados matemáticos, e incluirei uma ou duas experiências a fim de demonstrar o funcionamento das leis da probabilidade.

Em seguida, apreciaremos os conceitos básicos da amostragem - a população, a amostra, a estimativa baseada na amostra e o êrro de amostragem. De novo, faremos algumas experiências para verificar como êstes conceitos se conduzem na prática.

Estudaremos depois, de modo geral, os diferentes tipos de população, as diversas modalidades de planejamento possíveis para uma amostra dessas populações - amostra simples ao acaso, amostra de conglomerados, amostra sistemática, e assim por diante. Veremos paralelamente os diferentes tipos de estimativas que podem ser usados, e alguns dos problemas encontrados na estimativa dos erros de amostragem.

Tudo que foi mencionado requererá cêrca de quatro palestras. Prosseguiremos, então, descrevendo e discutindo aplicações específicas da amostragem no trabalho censitário. Visto como minha experiência foi adquirida no Bureau do Censo dos EE.UU., serei forçado a depender principalmente do trabalho que ali se tem realizado. Contudo, como sabeis, dispomos de algumas experiências com os resultados dos Censos brasileiros de 1950, os quais

fornecerão material interessante para discussão.

Gostaria finalmente de dedicar algum tempo à discussão do Inquérito Permanente de População do Bureau do Censo dos EE.UU. Trata-se de uma aplicação da amostragem que pertence mais ao campo da estatística permanente do que ao censitário; todavia, será útil falar-vos a respeito, primeiro porque representa uma das técnicas de inquérito por amostra mais cuidadosamente administradas e atualizadas do mundo, e segundo, porque apreciaria que considerásseis a praticabilidade e conveniência de possuir o Brasil um inquérito nacional dêsse tipo.

Até hoje, tem havido muito pouco uso das técnicas de amostragem no Brasil, mas há provas de que os estatísticos brasileiros reconhecem sua importância, e estão ansiosos por aprendê-las e aplicá-las. Acredito que as técnicas de amostragem apresentarão um desenvolvimento muito rápido nos anos vindouros. Não há veículo melhor para êste desenvolvimento do que os Censos de 1960. Não só se oferece a possibilidade de divulgar maior número de resultados em tempo mais curto e a menor custo, como também o uso bem sucedido da amostragem nesses censos mostrará o caminho a muitos outros grupos, tanto no govêrno como fora do mesmo, para a sua aplicação nas atividades estatísticas.

Há uma expressão nos EE. UU. -"getting in on the ground floor" - que significa estar entre os primeiros a tomar parte em algum novo empreendimento. Penso que sois favorecidos em ter a oportunidade e o desejo de vos tornardes pioneiros no que concerne ao uso da amostragem. Creio e espero que vosso conhecimento e experiência neste campo ser-vos-á de valor crescente nos anos que virão.

Após estas promissoras perspectivas sôbre o uso da amostragem, permiti-me acrescentar algumas palavras de cautela. A amostragem não é a solução mágica para todos os vossos problemas. A um tempo, pode alcançar ótimos resultados e causar graves embaraços, se não fôr usada corretamente. Deveis saber o que estais fazendo. Não quero dizer que, como técnicos censitários, deveis ter tôdas as fórmulas na ponta da língua. É preciso porém que tenhais à vossa disposição, sempre que aplicardes a amostragem, alguém que tenha competência para idealizar um plano de amostra eficiente, para evitar tendenciosidades ocultas, e dar-vos bons conselhos sôbre todos os problemas técnicos de amostragem. Mesmo os peritos cometerão erros de vez em quando. Mas uma das coisas boas da amostragem é que, geralmente, realizando-se pequena amostra adicional, ou aplicando-se ponderações diferentes ao realizar a estimativa, pode-se corrigir tais erros. Ao passo que, um êrro no planejamento de um censo, terá de esperar dez anos a fim de ser emendado.

Deveis insistir, por outro lado, para que o perito em amostragem torne razoàvelmente claro, sem recorrer a complexas fórmulas matemáticas, o que está propondo, e por que. Eu não confiaria em um perito que pedisse pa

ra aceitardes suas sugestões, sem discussão, por ignorardes os mistérios em que se baseassem. Espero sinceramente que, após estas reuniões e como resultado de vossos ulteriores estudos e experiências, compreendereis os "mistérios" nos quais se baseiam as boas técnicas de amostragem.

SEGUNDA PALESTRA

PROBABILIDADE

Porque é importante conhecer algo sôbre probabilidade?

Discutiremos êste assunto com detalhes mais adiante; no momento, tentarei apresentar-lhes uma síntese da matéria.

Há muitas maneiras de selecionar amostras, mas a finalidade é sempre a mesma: em cada caso, desejamos saber alguma coisa acêrca de um

universo

de que a amostra é uma parte.

Naturalmente não podemos obter um resultado exato, visto como estamos usando uma

amostra

Como, então avaliar o risco que corremos?

Para alguns tipos de amostra, não temos meios de fazê-lo. Se assim acontecesse para com todos os tipos da amostragem, relutaríamos em realizá-la. Felizmente, há um método de amostragem que resolve esta dificuldade. Trata-se da

amostragem probabilística

Com isto, queremos referir-nos a uma amostragem feita de forma que os resultados sigam as leis da probabilidade. Descobrimos que, dessa maneira, podemos usar a própria amostra não sòmente para estimar características de um universo, como também para conhecer os tipos de erros a que tais estimativas estão sujeitas.

Falaremos pois sôbre amostragem probabilística. Isto não quer dizer que não haja situações em que poderiam ser empregados outros métodos; tais situações, porém, são raras, especialmente no trabalho censitário.

Eis porque devemos discutir em primeiro lugar a probabilidade.

QUE É PROBABILIDADE?

O conceito de probabilidade envolve questão profunda, que tem sido discutida com interêsse por filósofos e cientistas. Podemos entretanto dizer, com segurança, que a probabilidade tem as seguintes características:

a) implica repetição da experiência, com resultados que podem ser classificados;

b) não pode ser conhecida com exatidão;

c) implica ignorância do que vai acontecer, dependendo nesse sentido do conhecimento prévio do problema em foco.

COMO MEDIMOS A PROBABILIDADE?

A expressão matemática da probabilidade deve ser um número entre 0 e 1. A soma das probabilidades de todos os resultados possíveis de uma expe-

riência deve igualar a 1, desde que êsses resultados sejam exaustivos - isto é, abranjam todos os casos possíveis, mùtuamente exclusivos, da experiência.

Convém esclarecer, até o perfeito entendimento, a significância de uma probabilidade igual a 1, bem como a significância de uma probabilidade de 0.

Discutir o caso de um ponto numa linha.

COMO CONHECER A PROBABILIDADE DE UM DETERMINADO ACONTECIMENTO?

Algumas probabilidades constituem apenas estimativas aproximadas. Exemplos:

Qual a probabilidade de que êste prédio esteja aqui amanhã?

Qual a probabilidade de que um homem morra, em resultado de uma operação cirúrgica?

Qual a probabilidade de que um cavalo ganhe determinada corrida?

Outras são bastante exatas. Exemplos:

Qual a probabilidade, no jogo de cara ou coroa, de ser cara o resultado de determinada jogada?

Qual a probabilidade, no jogo de dados, de se obterem 4 pontos?

Considerando-se um certo volume de um gás, submetido a determinada temperatura e pressão, qual a probabilidade de um certo número de moléculas chocar-se, durante determinado período de tempo?

Em qualquer caso, entretanto, nunca saberemos a probabilidade exata (exceto em casos de menor importância, em que a probabilidade é 0 ou 1) de um acontecimento, porque a probabilidade é uma expressão do que nós imaginamos que vai acontecer. Nossa estimativa da probabilidade pode basear-se em resultados de numerosas experimentações semelhantes, ou pode ser uma simples conjetura.

A utilidade da probabilidade depende de como essas estimativas concordem com a realidade. Podemos jogar dados ou cartas e saber o que esperar, porque as probabilidades estão próximas da realidade.

De maneira semelhante se comporta a ciência da mecânica: apresenta leis que predizem muito exatamente os resultados da aplicação de fôrças sôbre os objetos. Entretanto, estas leis são, realmente, expressões da probabilidade, porque os objetos considerados compõem-se de moléculas que se movem mais ou menos ao acaso. Diz-se que, teòricamente, é possível um tijolo levantar-se do chão, de repente, se por acaso a maioria, ou tôdas as suas moléculas, agirem dessa forma. Para todos os fins práticos pode-se asseverar, no entanto, que isto não acontecerá, porque a probabilidade é muito pequena.

Gostaria de salientar, outrossim, que a amostragem probabilística é um instrumento útil, porque permite selecionar uma amostra aleatória de tal maneira que torne conhecida a probabilidade de qualquer parte do universo parti-

cipar da amostra. Explicarei mais tarde, com detalhes, a maneira de fazê-lo.

PROBABILIDADES DE ACONTECIMENTOS COMPLEXOS

Falamos sôbre as probabilidades de acontecimentos simples. A teoria da probabilidade ensina como agir com as probabilidades, a fim de predizer o que acontecerá em situações que envolvam combinações complexas dêstes acontecimentos simples.

Exemplos:

1. Considerando, no jogo de cara ou corôa, que $p = 1/2$, para obter o resultado de cara, qual a probabilidade de:

   a. se uma moeda fôr jogada 10 vêzes, 7 resultados ou mais serem cara?
   b. exatamente 5 serem cara?
   c. haver pelo menos uma série de 3 caras consecutivas?

2. Se $p = 1/6$, no jogo de dados, para obter um número de 1 a 6, jogando-se 4 dados de uma vez, qual a probabilidade de:

   a. todos os resultados serem diferentes?
   b. haver exatamente dois 6 ?
   c. a soma ser 15 ou mais?

Empregando a teoria da probabilidade, podemos calcular as probabilidades de cada um dêstes acontecimentos. Ao fazê-lo não levamos em conta o fato de que as probabilidades reais não são exatas. Manipulamos quantidades teóricas. Entretanto, se as probabilidades verdadeiras são próximas do que esperamos, então temos alguns resultados úteis.

Não pretendo tentar ensinar-lhes como realizar estas operações em geral. Entretanto, gostaria de dar algumas ilustrações do que quero dizer.

Tomemos um dos exemplos anteriores: jogando-se 4 dados de uma vez, qual a probabilidade de que todos os resultados sejam diferentes?

Primeiro, calculamos o número total de resultados possíveis.

Exemplo:

1111, 1112, 1113, etc.

isto é, $6 \times 6 \times 6 \times 6 = 1296$

Dêstes, quantos não têm duplicação ?

Primeiro enumeramos as diferentes combinações de 4 números, sem levar em consideração a ordem:

1234	1345	2356
1235	1346	2456
1236	1356	3456
1245	1456	
1246	2345	
1256	2346	

Obtemos 15 combinações. Cada uma delas tem diferentes ordens possíveis, ex.:

1234	2134	3124	4123
1243	2143	3142	4132
1324	2314	3214	4213
1342	2341	3241	4231
1423	2413	3412	4312
1432	2431	3421	4321

Um total de 24, para cada combinação.

Existem, portanto, 15 x 24 = 360 modos de obter todos os resultados diferentes.

A probabilidade de 4 resultados diferentes é

$$\frac{360}{1296} = 0{,}278 \text{ ou } 1 \text{ em } 3{,}6 \text{ tentativas}$$

Suponhamos que não sabemos como calcular esta probabilidade. Poderíamos então estimá-la, realizando vários jogos com os quatro dados. Essas experiências funcionariam como uma amostra do conjunto de todos os jogos possíveis de realizar com quatro dados, não tendenciosos. Naturalmente, quanto maior fôr a amostra - o número de jogos realizados - tanto maior a fidedignidade da estimativa obtida.

Deí uma ilustração de como a teoria da probabilidade nos ajuda a resolver problemas complexos. Gostaria de apresentar - mas não resolver - alguns outros problemas, só para lhes dar uma idéia dos tipos de questões a que a teoria da probabilidade se pode aplicar.

1. Um amigo lhes diz que tem três filhos. Qual a probabilidade de que sejam todos do mesmo sexo? Todos meninos? Duas meninas e um menino?

2. Um médico preparou 100 contas de pagamento para enviar a seus clientes. Sua enfermeira colocou-as nos envelopes ao acaso, sem se preocupar em confrontar os nomes nas contas com os dos envelopes. Qual a probabilidade de que pelo menos uma chegue à pessoa adequada?

3. Tome-se uma superfície plana, com linhas paralelas distanciadas 10 cm uma da outra. Atire-se um palito de 5 cm de comprimento sôbre a superfície, ao acaso. Qual a probabilidade de que caia cruzando com uma das linhas?

4. Tomem 50 pessoas ao acaso, e perguntem as datas de seus aniversários. Qual a probabilidade de que pelo menos duas tenham o mesmo aniversário (mês e dia do mês)?

Talvez a nossa atenção aos jogos de azar e a outros assuntos ainda menos práticos tenha-lhes dado a impressão de que a teoria da probabilidade é de utilidade limitada. Esta teoria resultou, de fato, do interêsse pelos jogos de azar, e nos primeiros tempos talvez fôsse considerada de valor limitado. Conta-se de Benjamin Franklin que, inquirido sôbre o valor de suas experiências com a eletricidade, perguntou - "Que valor tem um bebê recém-nascido?"

A teoria da probabilidade "desenvolveu-se" desde então, ao ponto

de apresentar uma larga variedade de usos. Por exemplo:

- Na genética - a fim de predizer os modos pelos quais caracte - rísticas de pessoas, animais ou plantas serão transmitidas às gerações seguin - tes.

- Nos problemas de "tráfego", principalmente nas operações telefônicas: que quantidade de equipamento é necessário a fim de controlar um certo número de telefones?

- Na física nuclear - a fim de predizer o comportamento estatístico de massas de pequenas partículas, que se comportam individualmente ao acaso.

E por último, mas não menos importante, a amostragem probabilística.

Antes de deixar a probabilidade, gostaria de voltar mais uma vez ao jôgo e apresentar mais um caso:

Há um jôgo de dados (conhecido como "craps" em inglês) que consiste em:

O jogador joga dois dados

a. se o total fôr dois, três ou doze, êle perde imediatamente;

b. se o total for sete ou onze, êle ganha imediatamente;

c. se obtem 4, 5, 6, 8 ou 10, continua jogando até que, ou repita o número original, ou jogue um sete:

1. Se repetir o número original em primeiro lugar, vence;

2. Se jogar um sete em primeiro lugar, perde.

Qual é a sua probabilidade de ganhar?

Gostaria também de mencionar um sofisma comum: a chamada "lei das médias" - se um certo resultado se repete várias vêzes, torna-se menos e menos provável que se repita em tentativas sucessivas. Por exemplo:

Se jogarmos cara 5 vêzes seguidas, que poderemos dizer sôbre a jogada seguinte?

Ou ainda: Se jogarmos cara 999 vêzes seguidas, o que podemos saber sôbre a jogada seguinte?

Note-se que a moeda não se lembra do que aconteceu ...

Vamos agora resumir o que estudamos sôbre a probabilidade.

Que é probabilidade? É um número ligado a um de vários possíveis resultados de uma experiência, que exprime uma expectativa daquele resultado.

Quando ela é útil? Quando se verificam ocorrências consistentes com as probabilidades estabelecidas.

Qual sua utilidade na amostragem probabilística? Permitir a seleção da amostra de tal maneira que a probabilidade de seleção de cada unidade do universo seja conhecida quase exatamente. Neste caso, poderemos, empregando a teoria da probabilidade, predizer o comportamento da amostra e fazer afirmações precisas sôbre como a mesma diferirá do universo.

Segunda Palestra

CÁLCULO DE PROBABILIDADES (Resumo)

1. Definição clássica de probabilidade - Se um experimento puder dar N resultados mutuamente exclusivos e igualmente prováveis, e se n dêstes resultados possuirem um atributo A, então a probabilidade de A é a fração n/N, expressar por

$$P(A) = n/N$$

Exemplos:

a. Jogando-se um dado, qual a probabilidade de resultar um número par?

Resposta: A representa o aparecimento de um número par. O dado possui 6 faces, das quais 3 são números pares

$$P(A) = 1/2$$

b. Jogando-se uma moeda três vêzes sucessivas, qual a probabilidade de aparecer cara tôdas as três vêzes?

Resposta: A representa o aparecimento de 3 caras. Há 8 resultados possíveis (2x2x2), dos quais apenas um é três caras sucessivas

$$P(A) = 1/8$$

Permutações, combinações e arranjos

Ao coligir seja o número de resultados mutuamente exclusivos e igualmente prováveis de um acontecimento seja o número de resultados com um atributo particular, é necessário, com freqüencia, usar análise combinatória. Seguem-se algumas definições e fórmulas para os três principais tipos de contagem:

Permutação - A permutação de n itens significa um arranjo daqueles itens em alguma ordem especificada. Se os n itens forem todos diferentes, há $n! = n(n-1)\ldots 2.1$ permutações diferentes. Por exemplo, as permutações das três letras a b c são: a bc, ac b, ba c, ca b, bc a, c ba.

Combinação - uma combinação de m entre n itens ($n \leq n$) significa uma série de m dos n itens cuja composição é especificada, mas cuja ordem não é especificada. Se os n itens forem todos diferentes, há

$$\frac{n!}{(n-m)!\,m!} = \frac{n\,(n-1)\ \ldots\ (n-m+2)\,(n-m+1)}{1.2\ \ldots\ldots\ m}$$

combinações possíveis de m itens. As letras a, b, c combinadas tomadas duas de cada vez são ab, ac, bc. As combinações geralmente são denotadas pelo símbolo .. C_m^n ou $\binom{n}{m}$.

Arranjo - um arranjo de m entre n itens ($m \leq n$) significa uma série de m entre n itens com composição e ordem especificadas. Se os n itens forem todos diferentes, há

$$\frac{n!}{(n-m)!} = n\,(n-1)\ \ldots\ (n-m+2)\,(n-m+1)$$

arranjos possíveis dos m itens. Os arranjos das letras a, b, c consideradas duas de cada vez são ab, ba, ac, ca, bc, cb.

Nota: - Não há equivalente para a palavra arranjo em inglês. O têrmo permutação é usado tanto para permutações como para arranjos.

3. Alguns exemplos utilizando análise combinatória

a. Uma urna contém 4 bolas, numeradas 1, 2, 3 e 4, respectivamente. Uma pessoa de olhos vendados retira as bolas uma a uma. Qual a probabilidade de que as retirará em ordem numérica?

Resposta: n=1, visto com há apenas uma ordem numérica pròpriamente.

$$N = 4! = 4x3x2x1=24$$

$$P(A) = \frac{1}{24}$$

b. De um baralho de 52 cartas, tiram-se 4 ao acaso. Qual a probabilidade de que tôdas as quatro cartas serão de naipes diferentes?

Resposta: Neste problema, a ordem das 4 cartas não tem nenhuma importância. Visto como cada um dos 4 naipes tem 13 cartas, há

$$13^4 = 13x13x13x13$$

possíveis combinações de 4 cartas, nas quais tôdas são de diferentes naipes. O número total de combinações possíveis é $C_4^{52} = \frac{52x51x50x49}{1x2x3x4}$

$$\text{Portanto, } P(A) = \frac{13^4}{C_4^{52}} = 0,105$$

c. Em uma liga de futebol de 8 times, possuindo cada qual o seu próprio estádio, quantos jogos são necessários de maneira que cada time jogue com cada um dos outros times uma vez no próprio estádio e uma vez no estádio do outro time.

Resposta: A solução é dada pelo número de arranjos de 2 em 8 times. A ordem pode ser considerada como determinando o time doméstico. O número de jogos a serem realizados é 8x7 = 56

Probabilidades compostas

a. Sejam A e B dois atributos do resultado de um experimento e P (A) e P (B) as probabilidades de A e B. Então, a probabilidade de que o resultado terá pelo menos um dos dois atributos A e B é igual à soma das suas probabilidades individuais menos a probabilidade da sua ocorrência conjunta, isto é:

$$P(A+B) = P(A) + P(B) - P(AB)$$

Êste procedimento pode ser estendido a qualquer número finito de atributos.

b. Se A e B forem atributos mútuamente exclusivos, então a fórmula precedente torna-se

$$P(A+B) = P(A) + P(B)$$

c. Diz-se que um par de atributos é <u>independente</u> se a ocorrência de um dêles não afeta a ocorrência do outro.

A probabilidade de que um experimento produza um resultado com ambos os atributos independentes A e B é o produto das suas probabilidades separadas, isto é,

$$P(AB) = P(A)\,P(B)$$

d. O símbolo P (A/B) representa a probabilidade condicional de B, dado A. Em outras palavras, P(A/B) é a probabilidade de que o resultado de um experimento tenha o atributo A, dado o fato de que possui o atributo B.

A probabilidade da ocorrência conjunta de dois atributos quaisquer, A e B, é igual à probabilidade de um dos dois multiplicada pela probabilidade condicional do outro, isto é,

$$P(AB) = P(A)\,P(B/A) = P(B)\,P(A/B)$$

Se os atributos A e B fôrem <u>independentes</u>, isto se reduzirá à regra dada na parte C.

5. <u>Exemplos de probabilidades compostas</u>

a. (Ilustrando 4 a) Lança-se um dado. Qual a probabilidade de que o resultado seja ímpar ou maior que três?

Resposta:

$P(A) = 3/6$ visto como há três números ímpares.

$P(B) = 3/6$ visto como há três números maiores que 3.

$P(AB) = 1/6$ visto como há um número (5) que é ímpar e maior do que 3.

$P(A+B) = 3/6 + 3/6 - 1/6 = 5/6$

b. (Ilustrando 4 b) Lançam-se dois dados. Qual a probabilidade de que a soma das duas faces seja 2 ou 3?

Resposta: Êstes experimento tem $6x6 = 36$ resultados igualmente prováveis.

$P(A) = 1/36$, visto como há apenas um resultado com soma 2 (1,1)

$P(B) = 2/36$, visto como há duas maneiras de obter a soma de 3 (2,1 e 1,2)

$P(A+B) = 1/36 + 3/36 = 3/36 = 1/12$, visto como A e B são claramente mútuamente exclusivos.

c. (Ilustrando 4 c) Uma urna contém 10 bolas vermelhas e 10 brancas. Extrai-se uma bola ao acaso e repõe-se na urna, depois extrai-se uma segundo bola ao acaso. Qual a probabilidade de que ambas as bolas sejam vermelhas?

Resposta: Visto como a primeira bola é reposta, as duas extrações são independentes.

Portanto,

$$P(AB) = P(A)\,P(B) = \frac{10}{20} \times \frac{10}{20} = \frac{1}{4}$$

d. (Ilustrando 4 d) Da urna do exemplo anterior, retiram-se duas bolas sucessivamente. A primeira bola retirada não é reposta. Qual a probabilidade de que ambas as bolas sejam vermelhas?

Reposta: Visto como a primeira bola retirada não é retirada não é resposta, o resultado da segunda seleção depende do resultado da primeira. Portanto,

$$P(AB) = P(A)\,P(B/A) = \frac{10}{20} \times \frac{9}{19} = \frac{9}{38}$$

Note-se que a probabilidade de duas bolas vermelhas é menor quando a primeira bola não é reposta.

Êstes dois exemplos ilustram os processos de amostragem com reposição e amostragem sem reposição.

Segunda Palestra

Derivação da probabilidade de ganhar no jôgo de "Craps"

(vêr pág. 10)

O jôgo de "craps" joga-se da maneira seguinte: Um jogador joga dois dados:

a. Se o total dos dois dados fôr 2, 3 ou 12, êle perde imediatamente.

b. Se o total dos dois dados fôr 7 ou 11, êle ganha imediatamente.

c. Se o total fôr 4, 5, 6, 8, 9 ou 10, êle continua jogando até que repita o mesmo número ou jogue um sete.

1. Se repetir o número original em primeiro lugar, êle ganha.
2. Se jogar um sete em primeiro lugar, perde.

O quadro que se segue mostra como se deriva a probabilidade de ganhar:

Total dos dois dados (1)	Probabilidade dêste total na primeira jogada (2)	Probabilidade condicional de ganhar com êste total na 1ª jogada (3)	Produto (2) x (3) (4)
2	1/36	0	0
3	2/36	0	0
4	3/36	3/9	0,02777
5	4/36	4/10	0,04444
6	5/36	5/11	0,06313
7	6/36	1	0,16667
8	5/36	5/11	0,06313
9	4/36	4/10	0,04444
10	3/36	3/9	0,02777
11	2/36	1	0,05555
12	1/36	0	0
Total	36/36 = 1	—	0,49290

Probabilidade de ganhar = 0.49290

Explicação: Quando se jogam dois dados, há 36 resultados igualmente prováveis, isto é, (1,1) (1,2) (1,3) .... (5,6)(6,6). O numerador de cada probabi-

lidade na coluna (2) é o número de resultados que dão o total na (1). Por exemplo, um total de seis pode ser obtido de 5 maneiras - (1,5) (2,4) (3,3) (4,2) (5,1). O denominador de cada probabilidade é 36, o número total de resultados possíveis.

As regras do jôgo estabelecem que a probabilidade condicional de ganhar, para um determinado total na primeira jogada é zero se o total fôr 2,3 ou 12 e é um se o total fôr 7 ou 11. Para os outros totais - 4, 5, 6, 8, 9 e 10, temos que derivar a probabilidade de repetir êste total antes de jogar um total de 7.

Tirando um exemplo específico, derivemos a probabilidade de jogar um 4 antes de jogar um 7. Chamemos P_4 e P_7 a probabilidade de jogar 4 e 7 respectivamente em uma única jogada. Então, a probabilidade de jogar um quatro em primeiro lugar, que indicaremos por P, é igual à probabilidade de jogar um 4 na primeira jogada, mais a probabilidade de não jogar 4 ou 7 na primeira mas jogar 4 na segunda, mais a probabilidade de não jogar 4 ou 7 em nenhuma das duas primeiras jogadas, mas jogar 4 na terceira, etc. Isto pode ser escrito

$$P = P_4 + (1-P_4-P_7)\,P_4 + (1-P_4-P_7)^2\,P_4 + \ldots$$

$$= P_4 \sum_{n=0}^{\infty} (1-P_4-P_7)^n$$

A série infinita é uma série geométrica de forma $1+x+x^2+\ldots$ cuja soma é dada por $\frac{1}{1-x}$, de maneira que temos

$$P = P_4 \frac{1}{1-(1-P_4-P_7)} = \frac{3}{36} \times \frac{1}{1-\left(1-\frac{3}{36}-\frac{6}{36}\right)}$$

$$P = \frac{1}{3}$$

Esta probabilidade foi expressa por 3/9 na coluna (3) a fim de mostrar que é igual ao número de maneiras de produzir um 4 dividido pelo número de maneiras de produzir um 4 ou um 7. As probabilidades para 5, 6, 8, 9 e 10 são tôdas derivadas da mesma forma.

Na coluna (4) temos o produto da probabilidade de jogar um certo número na primeira jogada e a probabilidade condicional de ganhar com aquêle número na primeira jogada, isto é, a probabilidade conjunta de jogar aquêle número na primeira jogada e ganhar. A soma destas probabilidades para todos os resultados possíveis da primeira jogada dá a probabilidade global de ganhar.

TERCEIRA PALESTRA

INTRODUÇÃO À AMOSTRAGEM

Com as noções sôbre probabilidade que estudamos, estamos preparados para examinar a amostragem, e como se lhe aplica a teoria da probabilidade.

Comecemos por enunciar algumas definições:

Amostragem - é o ato de selecionar unidades elementares de um universo definido, a fim de conhecer algo sôbre êsse universo, ou mais exatamente, a fim de estimar certas características que lhe são peculiares.

Unidades elementares - são as unidades cujas características devem ser medidas.

Universo - é o conjunto de unidades elementares que desejamos medir.

Amostra - constitui-se de um grupo de unidades elementares retiradas de um universo.

Estimativa - representa um valor de uma característica do universo, deduzido da amostra.

Precisão - é a diferença entre o valor real e a estimativa.

Por que extraímos amostras? Pelos seguintes motivos:

1. Em algumas situações, é impraticável a realização de um levantamento completo. Exemplo: testes de destruição, como munição; exame de sangue; ou, então, quando o universo é limitado, como nas mensurações de comprimento, pêso, etc.
2. Obtençao de maior número de informações pelo mesmo custo. Pelo preço de um inquérito decenal que abranja todos os estabelecimentos industriais do Distrito Federal, por exemplo, é possível realizar quatro inquéritos por amostra, em intervalos menores de tempo, com resultados suficientemente fidedignos.
3. Obtenção da mesma informação, a menor custo. A economia, em algumas situações, é verdadeiramente notável, porque, em geral, é o tamanho da amostra, e não a percentagem, que controla os erros de amostragem.
4. A fim de economizar tempo. Por exemplo, na obtenção de tabulações preliminares de resultados censitários.

Qual o preço que pagamos por tais vantagens?

É sempre arriscado basear decisões em resultados de uma amostra, porque êsses resultados estão sujeitos a erros de amostragem. Isto entretanto não nos deve desanimar, porque:

1. Podemos selecionar amostras de modo a conhecer com exatidão o risco que corremos, nesse sentido;

2. Podemos tornar êsse risco tão pequeno quanto o desejarmos, aumentando o tamanho da amostra;

3. Na realidade, podemos até ganhar em precisão, porque, como sabem todos os que lidam com censos, uma contagem completa não quer dizer, necessàriamente, uma contagem exata.

Que se deve fornecer a um técnico em amostragem, para que possa planejar uma amostra? Duas informações fundamentais:

1. Qual o universo a ser investigado pela amostra.

A fim de responder a esta pergunta, devemos definir as unidades elementares. Exemplos:

- População (habitantes de uma determinada área geográfica)
- Estabelecimentos agropecuários
- Estabelecimentos industriais

Podemos, aliás, modificar nosso universo a fim de economizar despesas desnecessárias e obter um plano de amostragem que seja prático. Por exemplo, em uma amostra da população, no Brasil, podemos omitir certas tribus indígenas.

2. Qual a informação desejada.

É realmente espantoso, mas há pessoas que pretendem realizar inquéritos sem que possam especificar, exatamente, o tipo de informação que desejam obter. Suponhamos que, em um inquérito de famílias (por hipótese, já definimos família) devemos obter informações sôbre a renda média familiar. Devemos então saber:

a) Para que período (um determinado ano, mês ou semana?)

b) Que deve ser incluído na renda (como tratar as heranças, prejuízos comerciais, juros sôbre bonificações, etc.)?

c) Quais os diferentes tipos de renda que desejamos distinguir?

d) Quais as classificações cruzadas desejadas (por tamanho da família, raça, área geográfica, etc.)?

É necessário, portanto, definir exatamente o que se deseja estimar. Deve-se reconhecer que há um limite para a quantidade de detalhes que uma amostra pode fornecer. Quando se está informado sôbre as despesas necessárias à sua realização, pode-se modificar-lhe o plano.

Êstes dois pontos - definição do universo e das informações desejadas - podem parecer insignificantes a muitas pessoas, mas estou certo que todos vocês compreendem a sua importância. Como ilustração, gostaria de examinar os inquéritos de opinião pública (Gallup) que se aplicam à "predição" de eleições, nos Estados Unidos. Muitas pessoas têm-me dito - "Como é que a amostragem pode ser útil? Veja o que acontece com os inquéritos Gallup".

Como responder a isso? Ora, queremos predizer uma eleição median-

te a amostragem. Qual o universo de que devemos retirar a amostra para obter a previsão? - Evidentemente, êsse universo abrangerá tôdas as pessoas que deverão votar na eleição em causa. Podemos fazer uma amostragem dessa população, antes do dia da eleição? - Evidentemente não, assim como ninguém pode dizer ao certo se vai votar.

Admitindo todavia que pudéssemos selecionar uma amostra representativa, que informação desejaríamos? Desejaríamos saber como cada pessoa votaria na eleição. Podemos sabê-lo com certeza? Não, em virtude de

a) pessoas que ainda não se decidiram;
b) pessoas que mudam de idéia;
c) votos anulados.

Creio que ficou claro o motivo pelo qual não se pode predizer eleições com um grau de êrro conhecido, pelo Inquérito Gallup ou outro qualquer (note-se, outrossim, que nem me referi ao plano de amostragem utilizado). Exatamente os mesmos problemas surgiriam, aliás, se quizéssemos levantar um censo completo dos votantes potenciais, antes da eleição.

Existe uma terceira informação, igualmente importante, de que o técnico em amostragem necessitará. Pode ser formulado por uma, ou por ambas as questões seguintes:

a) qual o grau de precisão que se requer?
b) quais os recursos disponíveis?

Respondendo-se à pergunta (a), o técnico de amostragem poderá fornecer o custo da amostra. Respondendo-se à pergunta (b), poderá estimar o grau de precisão, dentro do orçamento previsto.

Na prática, raramente podem-se obter respostas diretas e imediatas para estas perguntas. É mais fácil começar por determinar as especificações quanto à informação desejada, e os recursos disponíveis; em seguida, após calcular a precisão dos resultados que podem ser obtidos por êsse preço, pode-se decidir:

- gastar mais dinheiro
- gastar menos dinheiro
- obter dados mais precisos
- obter dados menos precisos
- pedir informações mais detalhadas
- pedir informações menos detalhadas
- modificar o universo a ser investigado
- levantar um censo completo
- desistir do plano.

A adoção de um plano de amostragem pode ser comparado ao problema de recém-casados que procuram casa para morar. Têm uma idéia do tipo de moradia

que desejam, e começam a procurar o que há disponível, pelo preço que podem ou querem pagar. Finalmente, por uma série de ajustes entre vontade e necessidade, chegam a uma decisão - que pode ser, até, a de morar com a família de um dêles.

Determinamos, de modo geral, o que é a amostragem, e salientamos que, a fim de planejar uma amostra para um determinado projeto, precisamos responder a três perguntas:

1 - Qual o universo a ser compreendido na amostra?
2ª - Qual a informação necessária?
3ª - Qual a precisão requerida?

ou

3b - Quais os recursos disponíveis?

Examinemos agora a questão: Que características deve ter um planejamento de amostragem, para ser aceitável?

Prevenì-los-ei de que, para quase todos os fins dos trabalhos censitários, a resposta indica a amostra probabilística. Com isto em mente, vamos estudar algumas amostras não probabilísticas, e examinar os problemas peculiares.

Exemplo 1 - Suponhamos que, em eleições passadas, os resultados para Presidente da República em certo Município do Estado de São Paulo tenham sempre coincidido com os totais nacionais, no que se refere à ordem dos candidatos mais votados. Fazemos um levantamento dos votantes dêsse Município, antes das eleições, com o intuito de prever as suas preferências eleitorais. Será válida a previsão? (Comentar o processo adotado).

Exemplo 2 - Compramos uma cesta de maçãs. Examinamos duas, de cima da cesta; como parecem boas e gostosas, compramos tôda a cesta. Teremos agido acertadamente? (Comentar).

Exemplo 3 - Queremos estimar o consumo de café no Distrito Federal. Tiramos do catálogo 100 números de telefones, ao acaso; ligamos e perguntamos às pessoas que atenderem quantos cafêzinhos tomam por dia. É bom êste processo? Podemos usá-lo a fim de estimar o consumo total? O consumo per capita? (Comentar).

Exemplo 4 - (continuação do 3). Suponha-se que, depois de determinar a distribuição da população do Distrito Federal segundo a idade e o sexo, selecionamos uma amostra de 100 pessoas, com a mesma distribuição por idade e sexo. Enviamos entrevistadores que devem obter um certo número de entrevistas para cada grupo (idade e sexo) de pessoas. Comentar êste processo.

Que têm em comum todos êsses exemplos?

Segundo ensinam Hansen, Hurwitz e Madow ("Sample Survey Methods

and Theory" Vol. I, pág. 9):

> "Quando a determinação dos indivíduos (unidades elementares) a serem incluídos em uma amostra implica julgamento pessoal, não se pode obter uma medida objetiva de fidedignidade dos resultados da amostra, porque os vários indivíduos podem ter oportunidades diversas, e desconhecidas, de serem sorteados".

As amostras dêsse tipo, são conhecidas sob várias denominações: amostras de conveniência, amostras intencionais, amostras propositais, amostras por quota, etc.

Definiremos a amostra probabilística com palavras de Hansen, Hurwitz e Madow, (op. cit., pág. 11): "... o planejamento de uma amostra fornece uma amostra probabilística quando a probabilidade de inclusão na amostra é conhecida, e não é zero, para cada um dos indivíduos ou unidades elementares da população".

Quando usamos uma amostra probabilística, a precisão dos resultados de amostra pode ser medida da própria amostra. Em outras palavras, podemos estabelecer um limite superior na diferença entre o valor "real" e a estimativa da amostra, de forma que a probabilidade da diferença que excede êste limite seja tão pequena quanto desejarmos.

Além disso, quando se segue corretamente o processo da amostra probabilística, remove-se a possibilidade de tendenciosidade do planejamento da amostra. Estas vantagens - possibilidade de estimar o grau de precisão, através da amostra, e eliminação de tendenciosidade na seleção - são grandemente importantes. Tão importantes que, em princípio, se deve preferir uma amostra probabilística a uma amostra intencional, mesmo que a última contenha muito mais indivíduos, a igual custo. Isto não quer dizer que a amostragem intencional nunca deva ser usada.

Exemplos de amostras intencionais adequadas:

1. Exame de sangue.
2. Experiência preliminar com um questionário.

De um modo geral, entretanto, quando se trabalha com um universo finito (o que se dá na maior parte do trabalho censitário e de pesquisas), e se deseja estimar características dêsse universo (em contraposição ao exame de alguns casos, destinado a ganhar experiência), planejando-se incluir na amostra mais que um número muito pequeno de elementos, deve-se empregar a amostragem probabilística.

QUARTA PALESTRA

AMOSTRAGEM ALEATÓRIA SIMPLES

Definimos amostra probabilística, mas até agora nada dissemos sôbre como selecioná-la, ou o que fazer com a mesma, depois de selecioná-la. Podemos dizer que um inquérito por amostra, ou outro emprêgo de amostragem, consiste das seguintes etapas:

1. Planejamento da amostra
2. Seleção da amostra
3. Estimação
4. Estimação dos erros de amostragem

O planejamento de uma amostra implica a escolha, entre os muitos tipos possíveis de planos de amostragem, daquele que se acredita ser o mais eficiente em uma determinada situação. Discutiremos esta questão posteriormente. Em primeiro lugar quero discorrer sôbre os princípios básicos compreendidos nas etapas 2-4, examinando como se relacionam ao tipo mais simples possível de amostra - a amostra aleatória simples.

Para ser mais específico, examinarei uma amostra aleatória simples, retirada sem reposição de um universo limitado.

Suponhamos que, em um universo de N elementos, decidimos selecionar uma amostra de n elementos. O universo não está agrupado, estratificado ou aglomerado sob nenhuma forma.

Desejamos que, neste caso, cada elemento tenha uma probabilidade conhecida e igual de seleção. Quais são as maneiras pelas quais poderíamos fazê-lo? Há várias, por exemplo, poderíamos usar cartões, fichas, moedas, etc. Porém o processo mais conveniente e seguro será o uso de uma

Tabela de números aleatórios

A palavra aleatório é difícil de definir com precisão. De um modo geral, significa "sem um padrão ou modêlo discernível". A construção de uma tabela de números aleatórios é bastante complexa, e requer a aplicação de tôda uma série de testes estatísticos a fim de identificar quaisquer desvios da aleatoriedade. Como exemplo de uma tabela de números aleatórios, veja-se o Anexo 1.

O uso da tabela, entretanto, é razoàvelmente simples e oferece grandes vantagens sôbre os métodos anteriormente examinados. Examinemos-lhe a:

a) conveniência; e

b) fidedignidade.

Vejamos como aplicar a tabela de números aleatórios a uma amostra aleatória simples de 10 objetos em 100. (Note-se que o têrmo amostra aleatória simples tem um sentido muito preciso - não só cada elemento, mas cada grupo de

n elementos, deve ter a mesma probabilidade de seleção. Isto distingue a amostra aleatória simples das amostras de conglomerados, sistemáticas ou outras).

Exemplo prático: ilustrar a seleção de 10 números aleatórios entre 1 e 100. (01=1, 02=2 etc. e 00=100; abandonar as repetições).

O emprêgo de uma tabela de números aleatórios nem sempre é muito simples. Suponhamos, por exemplo, que quizéssemos 10 números aleatórios entre 1 e 300. Como selecioná-los-íamos da tabela? (Ilustrar).

As mais conhecidas fontes de tabelas de números aleatórios, nos Estados Unidos, são:

a) The Rand Corporation, A Million Random Digits, The Free Press, Glencoe, III.

b) H. Burke Horton (Título desconhecido - tabela de 105 000 números aleatórios), Interstate Commerce Comission, Washington D.C., 1949.

c) Kendall e Smith, Tracts for Computers nº 24 (contém 100 000 números) Cambridge, 1939.

d) Fisher e Yates, Statistical Tables for Biological, Medical and Agricultural Research, Oliver & Boyd, 1938. (contém 100 000)

Muitas obras sôbre estatística, principalmente as que se referem à amostragem, contêm também páginas de números aleatórios.

Examinaremos a seguir alguns problemas em que o uso dos números aleatórios é de proveito (V. Anexo B.)

Em têrmos de tempo e esfôrço, a seleção da amostra com o emprêgo da tabela de números aleatórios representa uma parte muito pequena do processo de amostragem. É entretanto, a etapa principal, por que mediante esta seleção é que se obtém uma amostra probabilística, possibilitando, pois interpretar os resultados de acôrdo com as leis da probabilidade.

Retornemos ao exemplo da amostra aleatória simples de 10 objetos em 100. Demonstramos como é selecionada. A providência seguinte seria obter a informação desejada com referência a cada um dos 10 elementos selecionados.

Exemplos de características: a) atributos; b) variáveis

De posse dessa informação, procedemos à estimação das características do universo.

Processos de estimação

Vejamos primeiramente quais as características que podem ser estimadas, e como o são.

a) Atributos (ex. côr das fichas, em um recipiente).

1. Número (de 100 fichas da população, quantas são vermelhas?)
2. Proporção (de 100 fichas, qual a proporção das vermelhas?)

b) Variáveis (ex.: renda familiar)

1. Total (renda total de 100 pessoas)
2. Média (total dividido pelo número de pessoas)
3. Mediana (renda individual ultrapassada por metade da população)
4. Amplitude total (diferença entre a mais alta e a mais baixa)

Há outras, mas estas são as mais comuns.

Suponhamos que nossa amostra aleatória simples consiste de 10 fichas de uma população de 100 fichas vermelhas e brancas. Como poderíamos usá-la a fim de estimar:

a) proporção de fichas vermelhas na população?
b) número de fichas vermelhas na população?

Suponhamos que nossa amostra consiste de 10 estabelecimentos agropecuários, selecionados aleatòriamente de uma população de 100. Como estimaríamos:

a) área total?
b) área média?

Tôdas estas estimativas têm a propriedade de não serem tendenciosas.

Uma estimativa sem tendenciosidade é aquela cujo valor esperado é o valor do universo. O valor esperado corresponde ao valor de todas as amostras possíveis obtendo-se a estimativa de cada uma e calculando a média para tôdas. (Nota: esta definição só se aplica no caso de serem tôdas as amostras igualmente prováveis, como em uma amostragem aleatória simples. Se não forem tôdas igualmente prováveis, então deve-se usar uma média ponderada).

Ilustração - Um universo de 5 estabelecimentos agropecuários com áreas (hectares) de 5, 20, 25, 30, 100. A área total é 180 ha, a área média é 36 ha. Considerando-se amostras de 2 unidades, há 10 possíveis:

Amostra	Elementos	Estimativa da área média	Estimativa da área total
1	5,20	12,5	62,5
2	5,25	15,0	75,0
3	5,30	17,5	87,5
4	5,100	52,5	262,5
5	20,25	22,5	112,5
6	20,30	25,0	125,0
7	20,100	60,0	300,0
8	25,30	27,5	137,5
9	25,100	62,5	312,5
10	30,100	65,0	325,0
Total	-	360,0	1.800,0
Média (valor esperado)	-	36,0	180,0

Não se deve chegar à conclusão de que a estimativa não tendenciosa é o único tipo aceitável de estimativa. A estimativa tendenciosa é simplesmente aquela cujo valor esperado não é igual ao valor real da característica estimada. A diferença entre o valor esperado da estimativa e o valor real denomina-se tendenciosidade.

O fato de que uma estimativa é tendenciosa não significa necessàriamente que há algo "errado" com a amostra ou o método de estimação empregado. A conveniência de um determinado método de estimação é julgada pelo efeito combinado do seu êrro de amostragem mais a sua tendenciosidade. Uma estimativa tendenciosa com um êrro de amostragem pequeno e uma pequena tendenciosidade será preferível a uma estimativa não tendenciosa com um êrro de amostragem grande. Há, de fato, um tipo particular de estimativa, conhecido como estimativa de razão, que, embora tendencioso, resulta freqüentemente em um êrro global menor do que o da estimativa não tendenciosa correspondente. Isto será examinado com maiores detalhes na Palestra nº 8.

A fim de compreender êstes conceitos mais perfeitamente, devemos examinar o conceito de precisão - que se mede pelo êrro de amostragem de uma estimativa.

Êrro de amostragem

Acredito que tenha ficado claro para todos que, de modo geral, quanto maior uma amostra, tanto mais precisas as estimativas das características de universo.

Também deve ficar claro, se pensarmos um pouco a respeito, que a precisão de uma estimativa baseada em uma amostra de um determinado tamanho depende da variabilidade, no universo, do item que estimamos.

A medida desta variabilidade chama-se desvio padrão. É computada da forma que se segue (ilustrar com número de pessoas por família):

1. Computar o número médio de pessoas por família.

2. Para cada família, computar o desvio da média, elevar ao qua-

drado e somar para tôdas as famílias.

3. Dividir pelo número total de famílias. Isto é conhecido como variância.

4. Tirar a raiz quadrada. Êste é o desvio padrão.

Ilustração numérica

Família	Nº de pessoas	Diferença da média	Quadrado da diferença
1	2	- 2	4
2	5	+ 1	1
3	4	0	0
4	6	+ 2	4
5	1	- 3	9
6	4	0	0
7	4	0	0
8	8	+ 4	16
9	3	- 1	1
10	3	- 1	1
TOTAL	40		36

$$\text{Média} = \frac{40}{10} = 4$$

$$\text{Variância} = \frac{1}{10} \times 36 = 3{,}6$$

$$\text{Desvio padrão} = \sqrt{3{,}6} = 1{,}9$$

A ilustração acima dá-nos o desvio padrão de uma característica do universo. Também é possível computar o desvio padrão de uma estimativa baseada em uma amostra, exatamente da mesma maneira. De fato, podemos interpretar o desvio padrão acima como o desvio padrão de uma estimativa da média baseada em uma amostra de uma unidade do universo de 10 famílias.

Que acontece, se aumentamos o tamanho da amostra? Computemos o desvio padrão de uma estimativa da média baseada em uma amostra de duas famílias:

ESTIMATIVA DA MÉDIA (1)	Nº DE AMOSTRAS COM ESTA ESTIMATIVA (2)	DIFERENÇA DA MÉDIA DA POPULAÇÃO (3)	QUADRADO DA DIFERENÇA (4)	(2) x (4) (5)
1,5	1	- 2,5	6,25	6,25
2,0	2	- 2,0	4,00	8,00
2,5	5	- 1,5	2,25	11,25
3,0	5	- 1,0	1,00	5,00
3,5	8	- 0,5	0,25	2,00
4,0	6	0	0	-
4,5	6	0,5	0,25	1,50
5,0	4	1,0	1,00	4,00
5,5	3	1,5	2,25	6,75
6,0	3	2,0	4,00	12,00
6,5	1	2,5	6,25	6,25
7,0	1	3,0	9,00	9,00
TOTAL	45	-	-	72,00

$$\text{Variância} = \frac{72}{45} = 1,6$$

$$\text{Desvio padrão} = \sqrt{1,6} = 1,3$$

Poderíamos fazer a mesma coisa para amostras de 3,4, 5 etc., até 10, e obteríamos os seguintes resultados:

TAMANHO DA AMOSTRA	VARIÂNCIA
1	3,6
2	1,6
3	0,9
4	0,6
5	0,4
6	0,3
7	0,17
8	0,10
9	0,05
10	0,00

Visto como estamos falando sôbre amostragem sem reposição, uma amostra de tamanho 10 terá sempre todos os elementos do universo e, portanto, não estará sujeita a nenhum desvio do valor dêsse universo.

Suponhamos que, em lugar de 10 famílias, temos dois outros universos, um de 20 famílias, e outro de 1 000 000 de famílias. Em cada caso, a proporção de famílias de cada tamanho corresponderá à do universo de 10 famílias, isto é, 10% têm uma pessoa, 10% têm 2, 20% têm 3, etc. Neste caso, a variância e o desvio padrão de tôdas as três populações serão os mesmos. O quadro seguinte mostra-nos qual será a variância da média estimada para vários tamanhos de amostras.

TAMANHO DA AMOSTRA	VARIÂNCIA DA MÉDIA ESTIMADA PARA UMA POPULAÇÃO DE		
	10	20	1 000 000
1	3,6	3,6	3,6
2	1,6	1,7	1,8
3	0,9	1,1	1,2
4	0,6	0,8	0,9
5	0,4	0,6	0,7
6	0,3	0,4	0,6
7	0,17	0,35	0,51
8	0,10	0,28	0,45
9	0,05	0,23	0,40
10	0	0,19	0,36

Isto ilustra o importante princípio de que, exceto quando a amostra equivale a uma grande proporção do total, o fator decisivo na determinação da precisão das estimativas é o número absoluto da amostra, e não a sua proporção em relação ao valor real do universo.

Em outras palavras, uma amostra de 1 000 de um universo de 10 000, e uma amostra de 1 000 de um universo de 10 000 000 fornecerão estimativas de médias ou proporções com quase a mesma precisão, contanto que os dois universos tenham composição semelhante. Fixando êste fator em mente, pode-se evitar o desperdício de usar amostras desnecessàriamente grandes, devido à crença errônea de que há algo de mágico ou desejável em uma amostra de 10%, independentemente do tamanho da população.

TABELA DE NÚMEROS ALEATÓRIOS

Exemplo de uma página

4652	3819	8431	2150	2352	2472	0043	3488
9031	7617	1220	4129	7148	1943	4890	1749
2030	2327	7353	6007	9410	9179	2722	8445
0641	1489	0828	0385	8488	0422	7209	4950
8479	6062	5593	6322	9439	4996	1322	4918
9917	3490	5533	2577	4348	0971	2580	1943
6376	9899	9259	5117	1336	0146	0680	4052
7287	0983	3236	3252	0277	8001	6058	4501
0592	4912	3457	8773	5146	2519	3931	6794
6499	9118	3711	8838	0691	1425	7768	9544
0769	1109	7909	4528	8772	1876	2113	4781
8678	4873	2061	1835	0954	5026	2967	6560
0178	7794	6488	7364	4094	1649	2284	7753
3392	0963	6364	5762	0322	2592	3452	9002
0264	6009	1311	5873	5926	8597	9051	8995
4089	7732	8163	2798	1984	1292	0041	2500
9376	7365	7987	1937	2251	3411	6737	0367
3039	3780	2137	7641	4030	1604	2517	9211
8971	8653	1855	5285	5631	2649	6696	5475
0373	4153	5199	5765	2067	6627	3100	5716
9092	4773	0002	7000	7800	2292	2933	6125
2464	1038	3163	3569	7155	2029	2538	7080
3027	6215	3125	5856	9543	3660	0255	5544
5754	9247	1164	3283	1865	5274	5471	1346
4358	3716	6949	8502	1573	5763	5046	7135
7178	8324	8379	7365	4577	4864	0629	5100
5035	5939	3665	2160	6700	7249	1738	2721
3318	0220	3611	9887	4608	8664	2185	7290
9058	1735	7435	6822	6622	8286	8901	5534
7886	5182	7595	0305	4903	3306	8088	3899
3354	8454	7386	1333	5345	6565	3159	3991
3415	7671	0846	7100	1790	9449	6285	2525
3918	5872	7898	6125	2268	1898	0755	6034
6138	9045	6950	8843	6533	0917	6673	5721
3825	1704	2835	4677	4637	7329	3156	3291
1349	0417	9311	9787	1284	0769	8422	1077
4234	0248	7760	6504	2754	4044	0842	9080
6880	3201	7044	3657	5263	0374	7563	6599
0714	5008	5076	1134	5342	1608	5179	0967
3448	6421	3304	0583	1260	0662	7257	0766
5711	7343	7539	3684	9397	5335	4031	1486
2588	3301	0553	2427	3598	2580	7017	9176
8581	4253	7404	5264	5411	3431	3092	8573
8475	6322	3949	9675	6533	1133	8776	2216
0272	5624	8549	5552	7469	2799	2822	9620
7383	7795	7939	2652	4456	6993	2950	8573
5126	2089	7729	0945	3901	4445	7117	8186
2064	3760	0939	7319	5939	3432	2030	4752
9315	8185	7805	6294	7072	6491	4012	1016
6814	8752	3462	6001	3302	3895	7371	3432

Quarta Palestra

ANEXO 2

NÚMEROS ALEATÓRIOS

Problemas concernentes à sua utilização

1. Usando a tabela de números aleatórios, como escolher 10 números aleatórios entre 1 e 5 ?

2. Usando a tabela de números aleatórios, como escolher 10 números aleatórios entre 1 e 7 ?

3. Como resolver o problema 2 com uma moeda?

4. Deseja-se selecionar números aleatórios entre 1 e 10. Propõe-se o processo seguinte:

   Jogue dois dados:

   a) Se o total que aparecer estiver entre 2 e 10, use êste total como número aleatório;

   b) Se o total fôr 11, considere-se 1 como o número aleatório;

   c) Se o total fôr 12, não atribua nenhum número desta jogada.

QUINTA PALESTRA

Significância do Desvio-Padrão

Vimos como se calcula o desvio-padrão de uma estimativa baseada numa amostra. Verificamos outrossim que, de um modo geral, depende fundamentalmente do tamanho da amostra, e apenas secundàriamente da sua proporção em relação ao universo.

Estudemos agora outra questão de real importância.

Que significa dizer que uma estimativa tem um certo desvio-padrão?

Para responder a isto, devemos examinar primeiramente o conceito de distribuição de freqüência. Todos sabem, em têrmos práticos, o que seja uma distribuição de freqüência. Costuma-se representá-la gràficamente, como ilustram as figuras 1 e 2 do Anexo 1. No primeiro caso, denomina-se uma distribuição de freqüência descontínua; no segundo, distribuição de freqüência contínua.

Em ambos, a escala vertical pode representar: a) o número de casos, com os valores indicados na escala horizontal; ou b) a proporção de todos os casos com aquêle valor (distribuição de freqüência relativa). A escala horizontal pode representar hectares, cruzeiros, idade ou qualquer outra característica. Algumas vêzes, trata-se apenas com valores positivos; em outras, podem haver valores positivos e negativos.

Um tipo particular de distribuição que consideraremos calcula-se da seguinte maneira: de determinado universo, digamos um grupo de 10 000 famílias, retiram-se repetidas amostras de 100 famílias, e para cada uma dessas amostras determina-se o tamanho médio da família. As estimativas obtidas podem ser classificadas em uma distribuição de freqüência representada gràficamente na forma da figura 3 (Anexo 1).

Se expressarmos esta distribuição de freqüência em números relativos, obteremos um método adequado para verificar a variabilidade das estimativas do tamanho médio de família, nas repetidas amostras de 100 famílias, mediante a distribuição proporcional das estimativas segundo as diferenças para com o valor real. Por exemplo, se o universo de 10 000 famílias apresentasse a distribuição do caso anterior (V. 4ª palestra, pág. 20) poderíamos obter os se-

guintes resultados, em uma série grande de amostras:

DIFERENÇA DO VALOR REAL (4,0)	PROPORÇÃO DE ESTIMATIVAS COM ESTA DIFERENÇA
0,00 - 0,09	0,404
0,10 - 0,19	0,302
0,20 - 0,29	0,180
0,30 - 0,39	0,079
0,40 - 0,49	0,026
0,50 - 0,59	0,007
0,60 - 0,69	0,001
0,70 +	menos de 0,0005

Distribuição do tamanho médio de família estimado de amostras de 100 famílias.

Qual o valor prático de tudo isto?

No exemplo, se retirássemos uma amostra de 100 famílias e estimássemos o número médio de pessoas por família, poderíamos determinar de antemão que a estimativa teria 7 "chances", em 10, de apresentar diferença de até 0,20 (ou 5%), em relação ao valor real. Da mesma forma, a probabilidade de que apresente até 10% de diferença seria de 97 em 100; ou ainda, de 999 em 1 000, tratando-se de uma diferença de até 15%.

Por outras palavras, variando a escala de casos favoráveis, podemos tornar a probabilidade tão grande quanto o desejarmos de que caia nesta escala.

ESCALA DE VALORES	PROBABILIDADE DA ESTIMATIVA PARA CADA CLASSE DE VALORES
3,90 - 4,10	0,404
3,80 - 4,20	0,706
3,70 - 4,30	0,886
3,60 - 4,40	0,965
3,50 - 4,50	0,991
3,40 - 4,60	0,998
3,30 - 4,70	0,999
3,20 - 4,80	quase certeza

Proporção de estimativas que caem em escalas especificadas (amostras de 100 famílias)

Aumentando o tamanho da amostra, aumentaremos as probabilidades de que a estimativa caia nas escalas mais reduzidas.

Contudo, ainda não alcançamos o estágio da utilidade prática, pois, em nosso exemplo, começamos com um universo completamente conhecido, ao passo que, geralmente, desejamos retirar amostras de universos desconhecidos.

Felizmente, verificou-se que pràticamente tôdas as estimativas baseadas em amostras razoàvelmente grandes seguem o que se chama a distribuição normal (ilustrada gràficamente pela figura 4 do Anexo 1).

A experiência demonstrou que essa distribuição é determinada por um número extremamente grande de processos naturais. A definição da distribuição normal é uma fórmula matemática, e pode-se demonstrar matemàticamente que as distribuições da maioria das estimativas baseadas em amostras aproximam-se da distribuição normal, à medida que aumenta o tamanho da amostra.

A figura 5 (Anexo 1) mostra a representação gráfica da distribuição das médias baseadas em amostras de um universo assimétrico. Vê-se por ela que a distribuição da média da amostra é normal, mesmo que a distribuição do universo seja não-normal e assimétrica.

A distribuição normal dá-nos uma resposta simples à questão do grau em que as estimativas por amostra se desviam do valor real. Se uma estimativa por amostra segue a distribuição normal, então sabemos que

68,3% de tôdas as estimativas diferirão do valor real em menos de 1 desvio-padrão (1σ)

95,4% de tôdas as estimativas diferirão do valor real em menos de dois desvios-padrão (2σ)

99,7% de tôdas as estimativas diferirão do valor real em menos de três desvios-padrão (3σ)

Pode-se fazer afirmação semelhante para qualquer múltiplo do desvio-padrão. Em especial, podemos dizer que 50% de tôdas as estimativas diferirão do valor real em menos de 0,6745 (ou cêrca de $(2/3\sigma)$.

Assim, tendo-se o valor real e o desvio-padrão da estimativa da amostra que apresente distribuição normal, podemos fazer afirmações precisas sôbre o desvio provável da estimativa do valor real.

Na prática, entretanto, quando efetuamos a amostra de um universo, não sabemos nem o valor real da estimativa (se o soubéssemos, não haveria razão para a amostra), nem o seu desvio-padrão. Portanto, devemos:

1. obter a estimativa pela amostra
2. estimar o desvio-padrão também através da amostra.

Suponhamos que a estimativa da média seja 510, e seu desvio-padrão estimado seja 20. Admitindo-se uma distribuição normal, poderíamos afirmar que o valor real acha-se provàvelmente na escala

490 - 530 (ou $510 \pm (1\sigma)$.

Não podemos atribuir uma probabilidade a esta afirmação, porque o valor real, ou está, ou não está na escala. O que podemos dizer é que, se repetíssemos êste mesmo processo de amostragem muitas vêzes, e de cada vez determinássemos um intervalo semelhante de um desvio-padrão de cada lado da estimativa, em 68,3% dos casos o valor real estaria nesse intervalo.

Gràficamente, isto apareceria conforme a figura 6 (Anexo 1), que se refere a um caso em que a média real é igual a 500.

O que acabo de descrever-lhes (com pequenas modificações) é conhecido como intervalo de confiança. É um intervalo baseado inteiramente em valores de amostras, e não requer nenhum conhecimento dos valores do universo. A hipótese de que a distribuição da amostra seja normal basta para tornar válida esta interpretação.

Para nossa conveniência, gostaria de resumir aqui os têrmos usados a fim de descrever os erros de amostragem. Já definimos o desvio-padrão e a variância, que é o quadrado do desvio-padrão.

O coeficiente de variação de uma estimativa é igual ao desvio-padrão dividido pela estimativa. Em realidade, é um desvio-padrão relativo. O coeficiente de variação ao quadrado é conhecido às vezes como variância relativa. Seguem-se alguns exemplos:

ESPECIFICAÇÃO	TIPO DE ESTIMATIVA		
	Número	Proporção	Percentagem
Estimativa	200	0,30	10%
Desvio-padrão	20	0,03	1%
Variância	400	0,0009	-
Coeficiente de variação	0,10 ou 10%	0,10 ou 10%	0,10 ou 10%
Variância relativa	0,01	0,01	0,01
Intervalo de confiança de 95%.....	160 - 240	0,24 - 0,36	8% - 12%

Medidas de variabilidade.

Omitiu-se a variância da percentagem porque é difícil dar significação ao quadrado da percentagem. Note-se que o coeficiente de variação pode ser expresso como proporção, de acôrdo com a definição acima, e também como percentagem, o que é praxe rotineira.

Para melhor elucidação do exposto, consulte-se o Anexo 2 a esta Palestra (exposição sôbre erros de amostragem, usada para tabulações preliminares do Censo de População de 1950, dos Estados Unidos).

G R Á F I C O S

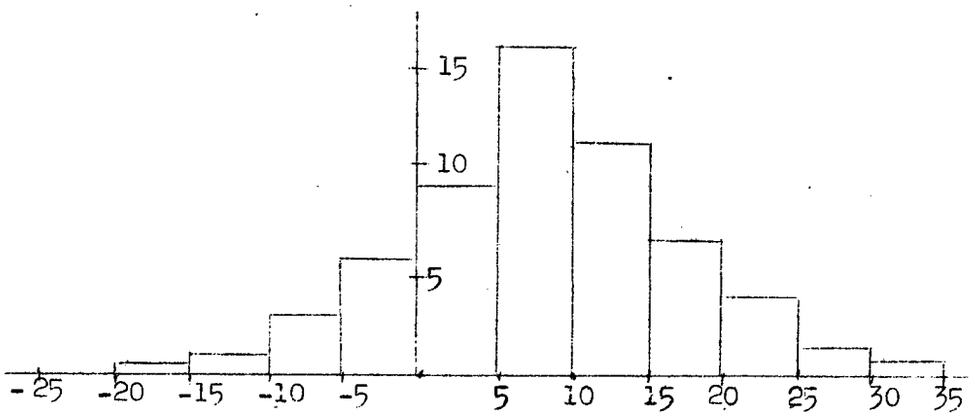


Fig. 1. Distribuição de freqüência descontínua

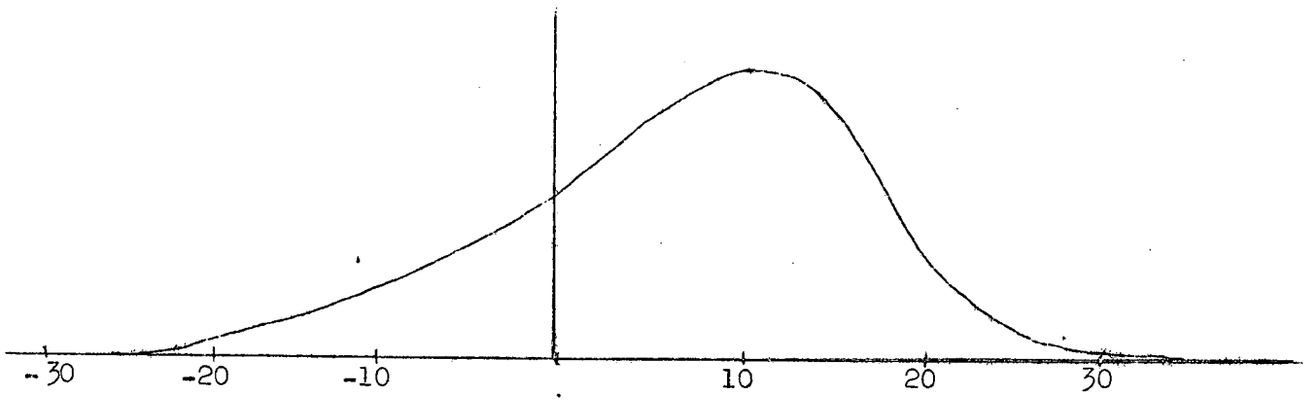


Fig. 2. Distribuição de freqüência contínua

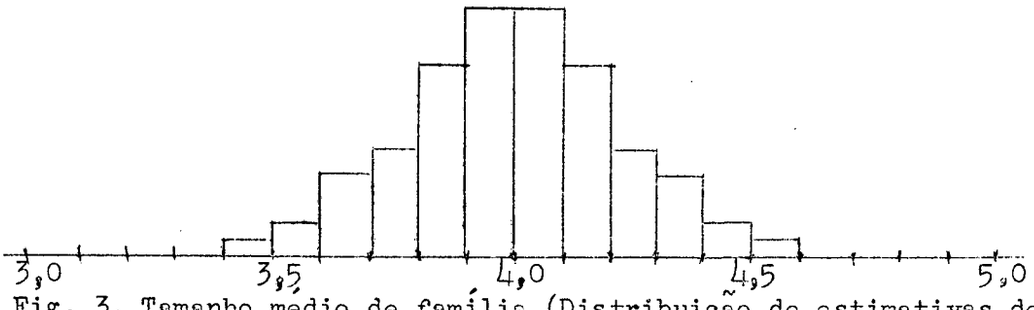


Fig. 3. Tamanho medio de família (Distribuição de estimativas de amostras de 100 famílias)

G R Á F I C O S
(conclusão)

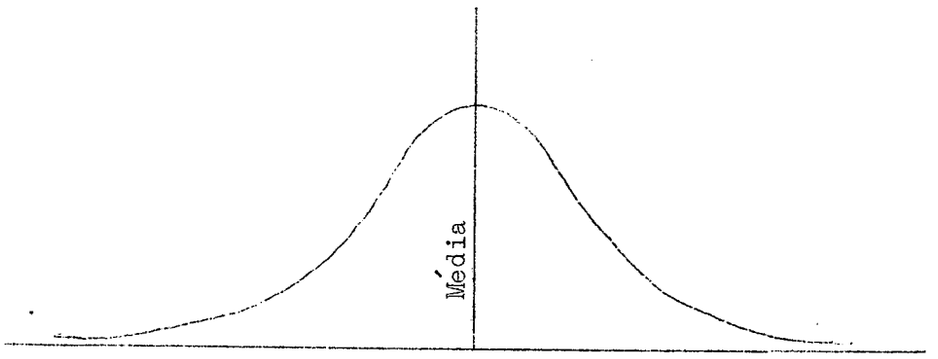


Fig. 4. Distribuição normal

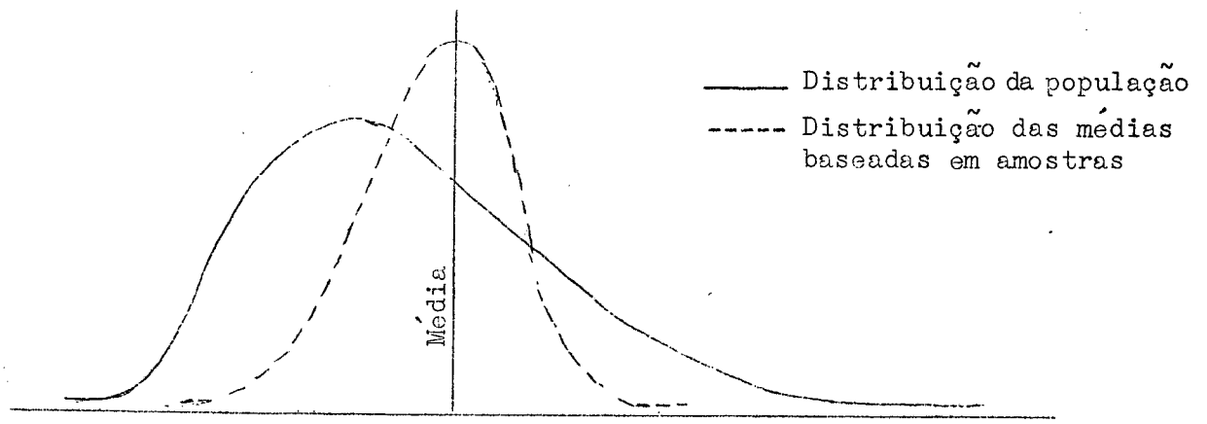


Fig. 5. Distribuição de médias baseadas em amostras de um universo assimétrico

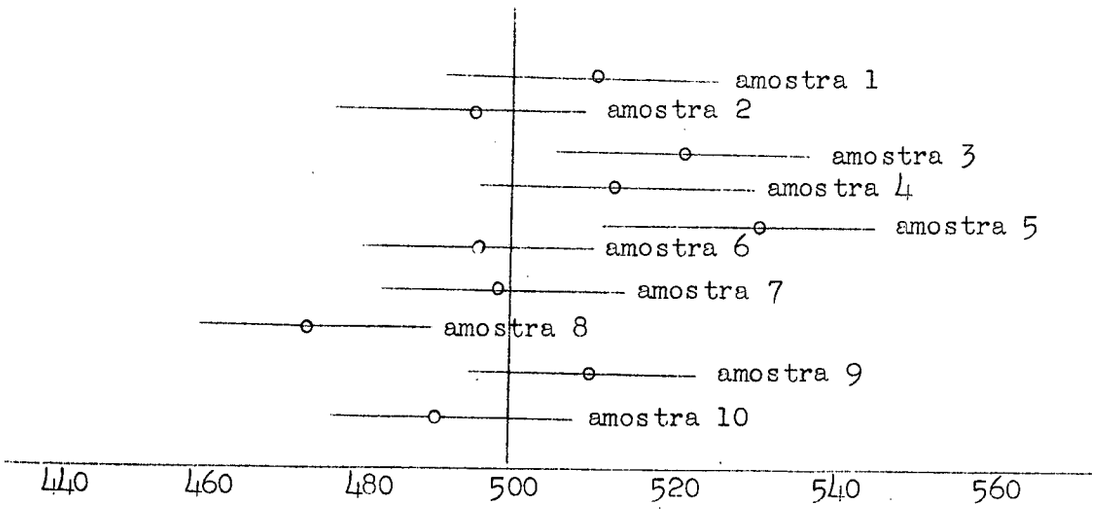


Fig. 6. Ilustração do intervalo de confiança

FIDEDIGNIDADE DAS ESTIMATIVAS DE 1950*

O quadro a seguir apresenta o desvio-padrão aproximado das estimativas de tamanhos selecionados (exceto para a população não branca). O desvio-padrão é uma medida de variabilidade da amostragem. Há 2, em 3 "chances" de que a diferença devida à variabilidade da amostragem entre uma estimativa e a cifra que se obteria pela contagem completa do universo seja menor do que o desvio-padrão. O número pelo qual o desvio-padrão deve ser modificado, a fim de se obterem outros ímpares considerados apropriados, pode ser encontrado na maioria dos textos sôbre estatística. Por exemplo, há 19 "chances", em 20, de que a diferença seja menor que o dôbro do desvio-padrão, e 99 "chances" em 100, menor que 2 vêzes e meia o desvio-padrão. As diferenças citadas neste relatório são, de um modo geral, pelo menos duas vêzes tão grandes quanto o seu desvio-padrão.

QUADRO A - DESVIO-PADRÃO DE UM NÚMERO ESTIMADO
(Escala de 2 "chances" em 3)

TAMANHO DA ESTIMATIVA	DESVIO-PADRÃO	TAMANHO DA ESTIMATIVA	DESVIO-PADRÃO
1 000 ...	2 000	5 000 000..	92 000
5 000 ...	3 000	7 500 000..	111 000
10 000 ...	4 000	10 000 000..	127 000
25 000 ...	7 000	20 000 000..	174 000
50 000 ...	10 000	30 000 000..	204 000
100 000 ...	13 000	40 000 000..	226 000
250 000 ...	21 000	50 000 000..	240 000
500 000 ...	30 000	60 000 000..	250 000
1 000 000 ...	42 000	70 000 000..	255 000
2 500 000 ...	66 000	100 000 000..	242 000

As cifras acima indicam o desvio-padrão das características da população total e da população branca. Para a população não branca, o desvio-padrão é maior, aumentando para cêrca do dôbro da cifra relativa às estimativas de 10 000 000.

As características da população rural também apresentam variabilidade ligeiramente maior do que a indicada no Quadro A.

*U.S. Department of Commerce, Bureau of the Census, Washington 25, D.C., 1950 Census of Population - Preliminary Reports, Series PC - 7, Nº 6, pág. 5.

QUADRO B - DESVIO-PADRÃO DE UMA PERCENTAGEM ESTIMADA
(Escala de 2 "chances" em 3)

BASE DA PERCENTAGEM	DESVIO-PADRÃO DA PERCENTAGEM ESTIMADA SEGUNDO O TAMANHO DA PERCENTAGEM				
	2 ou 98	5 ou 95	10 ou 90	25 ou 75	50
250 000	1,2	1,8	2,5	3,6	4,2
500 000	0,8	1,3	1,8	2,6	3,0
1 000 000	0,6	0,9	1,2	1,8	2,1
5 000 000	0,2	0,4	0,6	0,8	1,0
10 000 000	0,2	0,3	0,4	0,6	0,6
20 000 000	0,2	0,2	0,3	0,4	0,4
40 000 000	0,1	0,2	0,2	0,3	0,4
60 000 000	0,1	0,1	0,2	0,2	0,2
100 000 000	0,1	0,1	0,1	0,2	0,2

Além da variação da amostragem e de outras limitações mencionadas acima, as estimativas estão sujeitas a tendenciosidades devidas a erros de declaração ou emissão de declarações. A possível conseqüência de tais tendenciosidades não foi considerada nas medidas de fidedignidade: os dados obtidos mediante uma contagem completa de tôdas as pessoas também estariam sujeitos a essas tendenciosidades.

INTERVALOS DE CONFIANÇA (RESUMO)

1. Para qualquer característica que segue uma distribuição normal, seja uma estimativa ou um valor do universo, podemos dizer exatamente que proporção do universo cai em um determinado intervalo, contanto que conheçamos o desvio-padrão da característica. Especìficamente, podemos dizer que cêrca de 2/3 de todos os valores estão dentro do intervalo de um desvio-padrão em ambos os lados da média.

2. A maioria das estimativas de amostras razoàvelmente grandes (digamos 100 casos ou mais) seguem uma distribuição que é aproximadamente normal, mesmo quando as amostras são retiradas de universos ligeiramente assimétricos.

3. Ao efetuar amostra de um universo conhecido, podemos dar a probabilidade de que a nossa estimativa caia dentro de qualquer intervalo especificado. Em particular, podemos dizer que há uma probabilidade de cêrca de 2/3 de que a estimativa esteja dentro do intervalo de valores que vão do valor do universo menos um desvio-padrão (da estimativa) ao valor do universo mais um desvio-padrão. Se bem que, na prática, apenas retiramos uma amostra, esta probabilidade deve ser interpretada em têrmos de amostras repetidas; mas o intervalo mencionado é o mesmo para tôdas as amostras, porque é baseado em valores do universo.

4. Ao efetuar a amostragem de uma população desconhecida, podemos dar a probabilidade de que um intervalo baseado em valores de amostra contenha o valor real. Especìficamente, podemos dizer que há uma probabilidade de cêrca de 2/3 de que o intervalo que vai da estimativa da amostra menos um desvio-padrão (estimado da amostra) à estimativa da amostra mais um desvio-padrão contenha o valor do universo. Esta probabilidade também deve ser interpretada em têrmos de amostras repetidas; e, contràriamente ao intervalo para o universo conhecido, êste intervalo variará de uma amostra para outra, porque se baseia em valores de amostra (a estimativa e seu desvio-padrão estimado).

SEXTA PALESTRA

Tipos de amostragem

Já descrevemos com detalhes a amostra aleatória simples. De um universo constituído de N unidades elementares, selecionamos uma amostra de n unidades, de forma que haja a mesma probabilidade de seleção para cada unidade e cada combinação possível de n unidades. Tratando-se de amostragem aleatória simples, e todos os tipos de amostragem de universos finitos, devemos especificar inicialmente se a amostragem é

com reposição ou

sem reposição

Na amostragem com reposição, cada unidade elementar tem uma "chance" de seleção em cada uma das n seleções, tivesse ou não sido selecionada uma ou mais vêzes anteriormente. Na amostragem sem reposição, em cada seleção sucessiva apenas aquelas unidades que não foram selecionadas anteriormente têm "chance" de seleção.

No trabalho consitário, usa-se normalmente a amostragem sem reposição, porque

1. os erros de amostragem resultam ligeiramente menores, dado um determinado tamanho de amostra.
2. evita a inclusão de algumas unidades elementares (questionários, cartões, etc.) nas tabulações, mais de uma vez.

A única vantagem da amostragem com reposição consiste em ligeira simplificação nas fórmulas necessárias a fim de estimar erros de amostragem. Isto habitualmente não compensa a perda em eficiência.

A amostra aleatória simples é raramente usada na prática. Geralmente usam-se combinações de amostra de conglomerado, amostra estratificada, amostra sistemática e outros métodos que descreveremos. Entretanto, todos êstes métodos são, de um certo modo, extensões do método de amostragem aleatória simples; portanto, esta última é a base de tôda a nossa teoria e prática de amostragem.

Amostragem estratificada - Suponhamos que se deseja retirar uma amostra de 2, de um universo de 4 estabelecimentos industriais, a fim de estimar o número médio de empregados; os estabelecimentos têm o seguinte número de empregados.

Estabelecimento	Nº de empregados	
A	200	
B	250	Total de empregados = 500
C	20	Nº médio por estabelecimento = 125
D	30	

Se retirarmos uma amostra aleatória simples de 2 estabelecimen -

tos, há 6 amostras possíveis, conforme indica o quadro 1 (Anexo 1).

No caso em foco, a variância e o desvio-padrão equivalem a:

$$\sigma^2 = \frac{1}{6} \times 10{,}650 = 1775 \; ; \qquad \sigma = 42$$

Note-se que as amostras 1 e 6 dão estimativas muito deficientes, resultando em um grande êrro de amostragem.

Suponhamos agora que dividimos o universo em duas partes, ou estratos, consistindo o primeiro estrato dos estabelecimentos A e B, e o segundo dos estabelecimentos C e D Selecionamos uma amostra de 1 estabelecimento de cada um dos dois estratos. Há então 4 amostras possíveis, conforme o quadro 2 (Anexo 1).

Nesse caso, a variância e o desvio-padrão ficam reduzidos:

$$\sigma^2 = \frac{1}{4} \times 650 = 162{,}5 \; ; \qquad \sigma = 13$$

Neste exemplo, elimina-se a possibilidade de selecionar amostras consistentes dos dois maiores, ou dos dois menores estabelecimentos (amostras 1 e 6 no primeiro grupo), e, como resultado, reduz-se o desvio-padrão da estimativa de quase 70% (de 42 para 13).

Na prática, trabalharemos certamente com universos maiores, e não saberemos os valores das características que tentaremos estimar. Entretanto, comumente possuímos sôbre o universo informações intimamente relacionadas com os itens que desejamos investigar. Por exemplo, se desejássemos retirar uma amostra de estabelecimentos industriais no Distrito Federal, a fim de obter informações atuais sôbre emprêgo, fôlhas de pagamento, produção, etc., poderíamos considerar o universo de estabelecimentos recenseados no Censo de 1950 e dividí-lo em diversos estratos, de acôrdo com o número de empregados em 1950, ou alguma outra característica. Sem dúvida, muitos dos estabelecimentos terão números diferentes de empregados por ocasião do inquérito, mas quanto mais a distribuição de tamanho no inquérito atual fôr semelhante à distribuição de 1950, tanto mais a amostra estratificada será mais eficiente do que a aleatória.

No nosso exemplo, usamos a mesma taxa de amostragem - 1 em 2, ou 50% - em ambos os estratos. Em muitas situações, será mais eficiente variar a taxa de amostragem nos diferentes estratos. Em uma amostra de estabelecimentos industriais ou agropecuários, verificamos com freqüência que há um pequeno grupo de unidades extremamente grandes, por exemplo, estabelecimentos agropecuários de mais de 100 000 hectares, que responde por uma grande parte da área total ou outra característica. Estas unidades devem formar um estrato separado para serem "selecionadas" segundo a taxa de 100%; por outras palavras, devem ser recenseadas uma a uma.

Há uma teoria matemática, denominada teoria da repartição ótima,

que possibilita determinar o melhor meio possível de distribuir uma amostra de determinado tamanho, em diversos estratos. Em geral, esta distribuição dependerá de duas coisas:

1. variabilidade de unidades dentro do estrato (expressa pelo desvio-padrão); e

2. custo da obtenção das unidades de amostra dentro de cada estrato (note-se, por exemplo, que o custo unitário do recenseamento de famílias censitárias seria maior nas áreas rurais do que nas áreas urbanas; estando os outros fatôres equiparados, usaríamos uma taxa de amostragem mais elevada nas áreas urbanas).

Há muitos meios pelos quais podemos estratificar, e muitas razões para a estratificação. Ao fazer uma amostragem da população de uma cidade, por exemplo, poderíamos classificar instituições tais como hospitais, escolas, prisões, conventos, etc., em um estrato separado, seja porque as pessoas que vivem nestas instituições possuem características diferentes, seja porque se deva selecionar a amostra de uma forma diferente. Se fizéssemos uma amostragem afim de obter dados sôbre renda, poderíamos, mediante breve inquérito, dividir a cidade em áreas de condição econômica elevada, baixa e intermediária, e distribuir a amostra segundo êsses estratos, conforme indica a teoria da repartição ótima.

Veremos muitas outras aplicações da amostragem estratificada à medida que prosseguirmos.

Amostragem de conglomerado - Na amostragem estratificada, dividimos o universo em dois ou mais grupos e selecionamos uma amostra de unidades elementares de cada grupo. Na amostragem de conglomerado, dividimo-lo em um grande número de conglomerados, ou grupos de unidades elementares, e tiramos amostras dêsses conglomerados.

Por exemplo, na amostra experimental do Censo de 1950, para o Município de Vitória, a unidade elementar de amostragem foi o indivíduo, mas a seleção da amostra foi feita considerando-se cada boletim com número terminado em 3. Um boletim podia conter uma família grande, ou apenas um indivíduo.

Ao retirar uma amostra de respostas censitárias para uma grande área (um Estado por exemplo), poderíamos usar unidades maiores, tal como o município, o distrito, o setor ou a pasta, como conglomerados. Em tais casos, provàvelmente selecionaríamos uma amostra dêsses conglomerados e depois selecionaríamos sub-amostras dentro de cada um dos conglomerados de amostra. O conglomerado também poderia ser uma área apresentada em um mapa, e, tendo selecionado uma amostra de áreas, poderíamos então prosseguir recenseando tôdas as pessoas que vivessem na área, ou todos os estabelecimentos agropecuários que nela se encontrassem.

Tópico para discussão - Que devemos fazer, se um estabelecimento agropecuário tem partes em mais de uma área?

Em quase todos os casos encontrados na prática, uma amostra de conglomerados apresentará erros de amostragem maiores do que os de uma amostra aleatória simples com igual número de unidades elementares. A relação depende da composição dos conglomerados, conforme se apresenta nos dois exemplos seguintes:

Exemplo 1: Um universo tem 5 famílias, com 5 pessoas em cada família. Quatro das famílias são totalmente católicas romanas, e a 5ª família é toda ela protestante. Queremos estimar a distribuição de pessoas, por religião, no universo. Se retirarmos uma amostra de conglomerado consistindo de 1 das 5 famílias (20%), é claro que a nossa distribuição da amostra apresentará todos como Católicos Romanos ou todos Protestantes, resultado fraco, em qualquer caso. Por outro lado, se retirarmos amostras aleatórias simples de 5 pessoas (20%), haverá alguns exemplos com um Protestante e 4 Católicos Romanos, dando uma estimativa igual ao valor do universo, e, em geral, estaremos muito mais perto do valor real. É evidente que a amostragem aleatória simples é mais eficiente neste caso.

Exemplo 2 : Temos um universo de 5 famílias. Cada família tem 4 pessoas - 2 homens e 2 mulheres. Queremos estimar a distribuição de indivíduos, por sexo, no universo.

Neste caso, se retirarmos amostras de conglomerado consistindo cada uma de uma das 5 famílias, teremos sempre exatamente o resultado correto - 50% de homens e 50% de mulheres. Por outro lado, se retirarmos amostras aleatórias simples do mesmo tamanho - 4 pessoas - haverá algumas amostras com todos homens e outras com todos mulheres. Evidentemente, a amostragem de conglomerados é mais eficiente neste caso.

Visto como a amostragem aleatória simples é geralmente mais eficiente do que a amostragem de conglomerado, raramente usaríamos a última, quando mais não fôsse pelo fato de que devemos considerar fatôres de custo e tempo na determinação do melhor plano de amostragem a ser usado em uma dada situação. Suponhamos, por exemplo, que desejamos retirar uma amostra de 1 000 pessoas do Brasil. Considere-se o custo relativo para:

A. selecionar uma amostra de 1 000 pessoas, ao acaso, do país inteiro, e entrevistá-las; e

B. selecionar uma amostra de 100 municípios e uma sub-amostra de 10 pessoas de cada um dos 100 municípios, e entrevistá-las.

Evidentemente, o custo da alternativa B é menor, caso seja mesmo possível usar a alternativa A. Considerando as alternativas de outra maneira, caso tenhamos uma quantia certa de dinheiro para gastar, o uso da amostragem de conglomerado permitir-nos-á empregar uma amostra maior, e o aumento em tamanho freqüentemente compensará a perda em eficiência.

Como no caso da amostragem estratificada, há uma teoria matemática que torna possível determinar o tamanho ótimo de conglomerado em uma determinada situação. Os cálculos tornam-se às vêzes bastante complexos: por exemplo, quando os conglomerados são desiguais quanto ao tamanho, e há mais de um estágio de amostragem. Felizmente, têm sido feitos muitos estudos cuidadosos cujos resultados podem ser aplicados a muitas das situações encontradas na prática.

Amostragem com probabilidade proporcional ao tamanho - Até o momento, temos falado apenas sôbre planos de amostragem nos quais tôdas as unidades elementares, ou conglomerados, em um determinado universo ou estrato, têm a mesma probabilidade de seleção. Se atentarmos para a definição de amostragem probabilística, entretanto, vemos que ela não requer probabilidades iguais, mas apenas que a probabilidade de seleção para cada unidade seja conhecida e diferente de zero.

Vimos que nos universos que apresentam variação considerável em tamanho, tais como os estabelecimentos agropecuários ou industriais, podemos reduzir o êrro de amostragem para um determinado tamanho de amostra grupando as unidades elementares em estratos, de acôrdo com o tamanho, e efetuando a amostragem com coeficientes progressivamente mais altos nos estratos que contenham as maiores unidades.

Um meio alternado e mais preciso de conseguir o mesmo resultado consiste em retirar amostra de tôdas as unidades de um único grupo, dando porém a cada unidade uma probabilidade de seleção exatamente proporcional ao seu tamanho. O exemplo a seguir ilustrará como se consegue isto.

Exemplo - Temos um universo de estabelecimentos agropecuários para os quais se conhece a área em 1950.

ESTABELECIMENTO	ÁREA (ha)	ESTABELECIMENTO	ÁREA (ha)
A	5	E	5
B	5	F	25
C	10	G	15
D	30	H	5

Prosseguimos acumulando as áreas, conforme se segue:

Est. agropec.	Área (ha)	Área acumulada	Nºs. aleatórios
A	5	5	01-05
B	5	10	06-10
C	10	20	11-20
D	30	50	21-50
E	5	55	51-55
F	25	80	56-80
G	15	95	81-95
H	5	100	91-100

A cada estabelecimento corresponderia tantos números aleatórios quantas unidades de área tivesse. Dêsse modo, um estabelecimento de 30 hectares (estabelecimento D do exemplo) teria 30 "chances", em 100, de ser selecionado, ao passo que um estabelecimento de 5 hectares (estabelecimentos A, B e C do exemplo) só teriam 5 "chances", em 100.

Para selecionar um estabelecimento, proceder-se-ia à seleção de um dos números aleatórios entre 01 e 100 (ou 00 e 99). A cada estabelecimento está associado um grupo específico de números aleatórios, de tamanho variável conforme o tamanho do estabelecimento, e ordenado de acôrdo com a área acumulada. De modo que a um número entre 51 e 55, por exemplo, só pode corresponder na seleção o estabelecimento E; ou ainda, a um número entre 21 e 50, só corresponde o estabelecimento D.

Se quisêssemos que a amostra incluísse mais de um estabelecimento agropecuário, selecionaríamos números aleatórios adicionais e incluiríamos os estabelecimentos agropecuários correspondentes na amostra. Deve-se tomar cuidado com êste processo a fim de distinguir claramente entre a amostragem com e sem reposição, e aplicar as ponderações apropriadas, ao fazer as estimativas baseadas na amostra. O técnico em amostragem, nesse caso, deve, geralmente, especificar a norma seguida.

A amostragem com probabilidade proporcional ao tamanho é frequentemente utilizada quando se retiram amostras de grandes conglomerados, tais como municípios. Por exemplo, ao retirar uma amostra de 100 municípios, a fim de representar o Brasil, é evidente que aos municípios como o Distrito Federal e São Paulo deve ser dada maior probabilidade de seleção do que aos pequenos municípios do interior.

Note-se que o "tamanho" das unidades não será, em geral, igual ao da característica que queremos estimar. Por exemplo, em uma amostra de estabelecimentos agropecuários, o "tamanho" poderia ser o número de bovinos em 1952, embora desejássemos estimar o número de bovinos em 1956. Um estabeleci-

mento agropecuário que possuísse várias centenas de bovinos em 1952, e portanto, uma grande probabilidade de seleção, poderia não possuir bovinos em 1956. Isto não introduziria tendenciosidade nos nossos resultados; apenas aumentaria os erros de amostragem a que estão sujeitos. Habitualmente, será possível encontrar uma medida de tamanho que esteja bastante aproximada do assunto em consideração, a fim de aumentar a eficiência da amostra, quando se efetua amostragem com probabilidade proporcional ao tamanho.

Anexo 1

QUADROS

Quadro 1

NÚMERO DA AMOSTRA (1)	ESTABELECIMENTOS (2)	TOTAL DE EMPREGADOS (3)	ESTIMATIVA DA MÉDIA (4)	DIFERENÇA DA MÉDIA REAL (5)	QUADRADO DA DIFERENÇA (6)
1	AB	450	225	100	10 000
2	AC	220	110	15	225
3	AD	230	115	10	100
4	BC	270	135	10	100
5	BD	280	140	15	225
6	CD	50	25	100	10 000
Total	-	-	750	-	20 650

Quadro 2

NÚMERO DA AMOSTRA (1)	ESTABELECIMENTOS		TOTAL DE EMPREGADOS (4)	ESTIMATIVA DA MÉDIA (5)	DIFERENÇA DA MÉDIA REAL (6)	QUADRADO DA DIFERENÇA (7)
	I Estrato (2)	II Estrato (3)				
1	A	C	220	110	15	225
2	A	D	230	115	10	100
3	B	C	270	135	10	100
4	B	D	280	140	15	225
Total	-	-	-	500	-	650

SÉTIMA PALESTRA

Amostragem sistemática

O uso da amostragem sistemática no trabalho consitário é muito comum. Há dois tipos freqüentemente usados.

O primeiro tipo requer a seleção de cada K ésima unidade de uma lista de tôdas as unidades do universo.

O segundo tipo exige que cada unidade tenha algum número de identificação ligado à mesma, de modo a permitir a seleção de tôdas as unidades com números terminados em um certo dígito, ou dígitos determinados.

Êsses dois tipos de amostragem sistemática se equivalem quando tôdas as unidades têm números de uma única série consecutiva. Em muitos casos práticos, entretanto, há omissões ou duplicações no sistema de numeração, que devem ser levadas em conta na escolha do processo a empregar.

De certo modo, a amostragem sistemática pode ser considerada como uma forma da amostragem de conglomerado. Isto pode ser ilustrado por um exemplo:

Suponhamos que se deseje selecionar uma amostra de 1% dos telefones do Distrito Federal. Poderíamos conseguí-lo selecionando um número aleatório entre 00 e 99, de modo a incluir na amostra todos os telefones com números terminados nesse dígito. O que fizemos, de fato, foi dividir o universo em 100 conglomerados, consistindo cada um de todos os telefones com números terminados em dois dígitos especificados; e depois, selecionar um dêsses conglomerados ao acaso. Tais conglomerados, contudo, são de um tipo muito especial, porque cada um está espalhado por todo o Distrito Federal, pelo que a amostra de um conglomerado torna-se razoàvelmente eficiente. Se, porém, os nossos conglomerados consistissem de todos os telefones de um quarteirão, ou de uma certa "estação" telefônica, certamente não poderíamos confiar em uma amostra de um dêsses conglomerados.

Em muitas situações, a amostragem sistemática será mais eficiente do que a amostragem aleatória simples, porque introduz um certo grau de estratificação. Isto acontece particularmente quando a ordem de listagem ou numeração tem alguma significância geográfica. Suponhamos que queremos selecionar uma amostra de 10% das casas numeradas de uma determinada rua. Se as listarmos em ordem numérica e selecionarmos cada décima casa, estaremos automàticamente seguros da representação proporcional em cada secção da rua. Isto é importante, porque as casas e famílias com as mesmas características demográficas, sociais e econômicas tendem a se aglomerar. Considerando-se cada décima casa, estamos na realidade formando estratos de 10 casas cada, e selecionando uma casa de cada estrato. A única diferença é que as seleções de unidades dentro dos estratos não são independentes - a primeira determina a seleção da amostra inteira.

Pode-se usar o mesmo exemplo a fim de ilustrar um dos perigos da amostragem sistemática. Suponhamos que, ao invés de a décima casa numerada, decidíssemos selecionar tôdas as casas com números terminados em um certo dígito, tomado ao acaso. Há pelo menos duas maneiras pelas quais êste processo pode causar dificuldade.

Em primeiro lugar, a maioria dos sistemas de numeração de ruas têm todos os números ímpares de um lado da rua e os números pares do outro. Portanto, a amostra proposta teria casas de um lado apenas da rua. Tècnicamente, não seria uma amostra tendenciosa, visto como tôdas as casas têm uma probabilidade conhecida de seleção; entretanto, seria ineficiente (resultaria em erros grandes de amostragem) se houvesse quaisquer diferenças substanciais entre as casas dos dois lados da rua.

Em segundo lugar, muitos sistemas de numeração de ruas atribuem um grupo de números (geralmente 100) a cada quarteirão, tenha ou não número correspondente de casas; de maneira que a primeira casa em um quarteirão, de um lado, poderia ter sempre um número terminado em 00, por exemplo, e a primeira do outro lado, um número terminado em 01. Assim, duas das nossas dez amostras possíveis conteriam uma proporção extraordinàriamente alta de casas de esquina, e tôdas as outras uma proporção baixíssima. Também isso seria ineficiente pois, como freqüentemente se verifica, as casas de esquina tendem a possuir características diferentes das outras casas.

Mesmo o método de tomar cada décima casa numerada é traiçoeiro. Se não houve lacunas nos números para grandes secções da rua, verificaremos novamente que a amostra não está igualmente distribuída por ambos os lados da rua. Neste exemplo, seria conveniente usar um intervalo de amostragem ímpar em lugar de par, ou fazer outra modificação adequada.

Outra ilustração dos perigos da amostragem sistemática: suponhamos que se numeram os quarteirões de uma cidade, e depois seleciona-se uma amostra sistemática de 1 em 5. Devemos tomar cuidado a fim de evitar a seguinte situação:

1	2	(3)	4	5
10	9	(8)	7	6
11	12	(13)	14	15
20	19	(18)	17	16
21	22	(23)	24	25

Neste exemplo, selecionamos cada quinto quarteirão, começando com 3, e verificamos que todos os quarteirões selecionados ficaram enfileirados. Provàvelmente esta amostra não seria eficiente.

Sempre que usarmos uma amostra sistemática, portanto, é extremamente importante considerar as modificações periódicas das características do universo. Determinadas essas modificações, devemos proceder do modo que o intervalo de amostragem não coincida com os períodos do universo.

Apesar dêstes problemas, a amostragem sistemática é extremamente comum no trabalho censitário devido à sua conveniência e facilidade de aplicação.

É conveniente porque elimina a necessidade de aplicar um número aleatório isolado para cada unidade da amostra. É fácil de empregar em um Censo ou inquérito, porque permite darem-se instruções simples ao entrevistador para selecionar a respectiva amostra. As operações de amostragem em grande escala, tais como o uso de perguntas de amostra no censo, seriam pràticamente impossíveis sem a aplicação dos princípios da amostragem sistemática.

Há uma certa dificuldade na estimação dos erros de amostragem de estimativas baseadas em uma amostra sistemática. Isto se dá porque as estimativas dos erros de amostragem são obtidas examinando-se os desvios de tôdas as unidades de amostragem em relação ao seu valor médio; mas em uma amostra sistemática, realmente só temos uma unidade, ou conglomerado, de amostragem, determinada ao acaso, e portanto, não há desvios a calcular. Entretanto, isto não constitui um problema sério visto como existem muitas aproximações que podem ser usadas a fim de obter estimativas razoàvelmente boas dos erros de amostragem.

Amostragem dupla e seqüencial

É difícil definir estas expressões com precisão, visto como têm sido usadas para descrever vários tipos diferentes de planos de amostragem. Contudo, o princípio básico da amostragem seqüencial consiste em selecionar uma amostra em partes sucessivas, verificando os resultados de cada parte antes de selecionar a seguinte, a fim de decidir se a amostragem adicional é necessária, e, neste caso, até que ponto o é, e qual o melhor planejamento de amostra a ser usado.

A amostragem dupla pode referir-se a uma amostra seqüencial em duas partes; entretanto, também é aplicada a certos tipos de processos de sub-amostragem nos quais a informação da amostra inicial é usada a fim de servir como base para estratificação, ou para melhorar a estimativa final.

Algumas ilustrações talvez esclareçam melhor o assunto:

Exemplo 1 - Amostra seqüencial, a fim de aceitar ou rejeitar um lote - Temos um grande grupo de cartões perfurados, e desejamos saber se a proporção de erros de perfuração é bastante baixa para que possamos aceitar o lote, isto é, admitir que sejam tabulados sem verificá-los todos. Desejamos aceitar o lote se estivermos razoàvelmente certos de que o coeficiente de êrro (percentagem de todos os cartões que têm um ou mais erros) é de 5% ou menos.

Podemos abordar êste problema da forma seguinte:

1. Tomamos uma amostra aleatória de n cartões, verificamo-los e calculamos o coeficiente de êrro na amostra.

2. Conhecido o coeficiente de êrro na amostra, nossa medida seguinte será

   a. Se o coeficiente de êrro da amostra exceder certo limite, digamos 10%, rejeitamos o lote, isto é, verificamos todos os cartões.

   b. Se o coeficiente de êrro da amostra é inferior a um certo limite, digamos 1%, aceitamos o lote sem verificação.

   c. Se o coeficiente de êrro da primeira amostra acha-se entre êstes dois limites, isto é, na escala de 1 a 10%, selecionamos uma segunda amostra de tamanho n.

3. Se os resultados da primeira amostra requeressem a seleção de uma segunda amostra (caso c), selecionaríamos esta amostra adicional, verificaríamos os cartões e calcularíamos o coeficiente de erros das duas amostras combinadas.

4. Teríamos novamente três alternativas, da mesma forma que após a primeira amostra, mas a escala de coeficientes de erros indicativa de amostragens posteriores seria mais estreita, por exemplo:

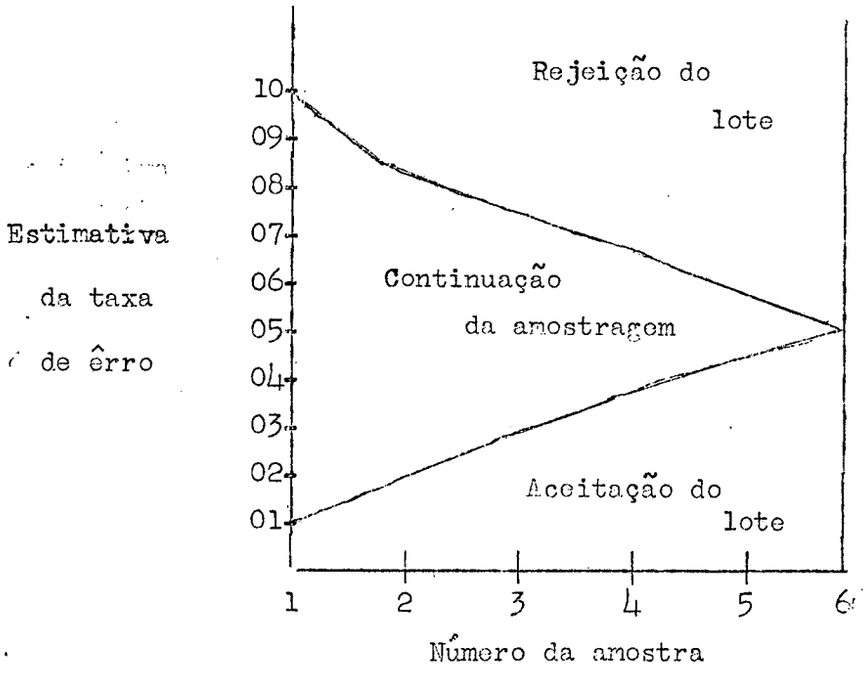
   Menos de 2,5% - aceitamos

   Mais de 8,0% - rejeitamos

   Entre 2,5% e 8,0% - continuamos efetuando a amostragem

5. Em amostras sucessivas, se necessário, seguimos o mesmo processo, de forma que os limites para aceitação e rejeição aproximam-se cada vez mais (convergindo para a casa de 5%) até que, adotássemos uma decisão final, com base na estimativa do coeficiente de êrro das amostras combinadas.

Êste processo pode ser ilustrado gràficamente.



PLANO DE AMOSTRAGEM SUCESSIVA

Perguntas:

1. Por que a região para continuação da amostragem se torna cada vez menor, à me-

dida que cresce o número de amostras?

2. Por que não selecionamos imediatamente uma única amostra de tamanho equivalente às seis amostras menores?

Exemplo 2 - Temos uma lista de estabelecimentos industriais e queremos selecionar uma amostra que inclua 1 em 10 dos estabelecimentos com 100 empregados ou mais, e 1 em 50 daqueles com menos de 100 empregados.

Os estabelecimentos da lista não se encontram em ordem de tamanho, mas admitamos que estejam numerados em uma única série contínua, começando por 1.

Uma forma muito simples de selecionar a amostra seria procurar cada décimo estabelecimento da lista (digamos, aqueles com números terminados em 2) e verificar o número de empregados. Quando contassem 100 ou mais empregados, seriam incluídos na amostra. Por outro lado, cada 50º estabelecimento (digamos, os de números terminados em 12 e 62) seria também incluído na amostra, independentemente do número de empregados.

Esta amostra teria uma variância ligeiramente mais alta do que a obtida mediante a seleção de exatamerte 1/10 dos grandes e 1/50 dos pequenos estabelecimentos, com base em listas elaboradas por ordem de tamanho. Mas a diferença não é muito grande, e em alguns casos pode tornar-se caro ou impraticável redistribuir as unidades da lista.

Perguntas:

1. Por que a amostra dupla teria uma variância maior do que a estratificada?

2. Suponhamos que os estabelecimentos não fôssem numerados. Poderíamos ainda usar esta técnica?

Exemplo 3 - Estamos retirando uma amostra de unidades domiciliares no Distrito Federal, a fim de estimar a proporção das desocupadas. Somos solicitados a estimar esta proporção com um coeficiente de variação de 10% ou menos. Sabemos apenas que o coeficiente de vacância se acha provàvelmente entre 1 e 10 por cento, mas não conhecemos quais os bairros que apresentam coeficientes altos, ou baixos.

Uma forma de proceder seria retirar uma única amostra de cada bairro bastante grande que assegurasse a fidedignidade do resultado independentemente do coeficiente de vacância e de sua variabilidade entre os quarteirões, ou outra unidade de amostragem. Êste processo resultaria quase que certamente em um coeficiente de variação muito inferior a 10%, a um custo extra considerável.

A outra alternativa seria efetuar a amostra em dois estágios. No

primeiro, retiraríamos uma amostra de cada bairro bastante grande a fim de assegurar a fidedignidade desejada, mas só no caso de serem mínimos os coeficientes de vacância e a variabilidade entre as unidades de amostragem. Como resultado dessa amostra, teríamos estimativas de

a. coeficiente de vacância em cada bairro; e
b. grau de variabilidade dentro dos bairros.

Obtida esta informação, poderíamos planejar eficientemente a amostra do segundo estágio, a fim de dar o coeficiente de variação requerido, distribuindo a maior parte da amostra restante pelos bairros com coeficientes altos de vacância e alta variabilidade. Desta maneira, reduziríamos ao mínimo o custo total, usando a distribuição ótima da amostra combinada entre os bairros.

Estes três exemplos ilustram os empregos principais da amostragem dupla e seqüencial, métodos que são particularmente úteis quando pouco se sabe de antemão a respeito do universo de que se deve retirar a amostra, ou quando a estratificação é importante devido à variação no tamanho das unidades elementares, mas o custo da estratificação de todo o universo é alto. Por outro lado, deveríamos ter em mente pelo menos duas desvantagens da amostragem seqüencial:

a. o custo de uma amostra em dois ou mais estágios tende a ser um pouco mais alto que o de uma amostra do mesmo tamanho retirada em um único estágio.
b. a necessidade de tabular os resultados de cada estágio antes de proceder ao seguinte atrasará a conclusão do inquérito, ou outra operação por amostragem.

RESUMO - Cada um dos tipos de amostragem que estudamos - amostragem aleatória simples, amostragem estratificada, amostragem de conglomerado, amostragem sistemática, amostragem com probabilidade proporcional ao tamanho e amostragem seqüencial - tem suas vantagens e desvantagens peculiares, cuja importância deve ser avaliada separadamente cada vez que planejamos uma amostra para uma situação particular. Não existe fórmula matemática que possa ser usada a fim de determinar, em cada situação, uma combinação dêsses métodos que forneça a precisão desejada, ao menor custo. Portanto, o planejamento de uma amostra é apenas em parte uma ciência exata; em parte, é tambem uma arte que requer a habilidade de aplicar a experiência passada a situações novas, e fazer o uso máximo de todos os materiais e dados disponíveis. Devemos lembrar, igualmente, que existem diversos modos de fazer estimativas de uma amostra, de maneira que a escolha do melhor processo de estimação também faz parte do problema do planejamento de uma amostra. Na secção a seguir descreveremos as características dos principais tipos de estimativas usadas com relação às amostras probabilísticas.

OITAVA PALESTRA

Processo de estimação

Há muitas maneiras de fazer estimativas por meio de uma amostra. Geralmente também se lança mão de alguma informação não proveniente da amostra. A melhor estimativa, neste caso, seria a que tirasse máximo proveito da informação disponível. Por exemplo:

Após levantar um censo, decidimos fazer tabulações preliminares baseadas em uma amostra de 5% de todos os setores. Para estimativas de totais, como número de pessoas presentes, número de homens no grupo de 14 a 19 anos de idade, número de mulheres solteiras, etc., poderíamos obter uma estimativa sem tendenciosidade (valor esperado igual ao valor real) simplesmente multiplicando os totais da amostra por 20. Contudo, não é êste necessàriamente o melhor método.

Visto como haverá, sem dúvida, algum tipo de contagem do total de pessoas recenseadas em cada setor (para fins de pagamento e contrôle, por exemplo) pode-se utilizar essa contagem como base para melhores estimativas de totais. Por exemplo, poderíamos estimar totais da forma seguinte:

$$\text{Estimativa de pessoas com a característica A} = \frac{\text{Pessoas da amostra com a característica A}}{\text{Pessoas recenseadas na amostra}} \times \text{Contagem de pessoas recenseadas em todos os setores}$$

Esta espécie de estimativa é conhecida como estimativa de razão, assim chamada porque a amostra é usada a fim de estimar a razão de duas variáveis aleatórias. Neste exemplo, emprega-se a amostra para estimar a proporção de pessoas com a característica considerada; pode-se multiplicar essa proporção por um total baseado na contagem completa das pessoas recenseadas a fim de obter a estimativa do número de pessoas com a característica.

Nem tôda estimativa de razão é necessàriamente uma estimativa não tendenciosa, isto é, nem sempre o valor esperado é igual ao valor real do universo. Isto pode ser verificado pelo seguinte exemplo: em um universo de 4 famílias, desejamos estimar a proporção de pessoas de 40 anos e mais, retirando uma amostra de 2 famílias:

Família	Total de pessoas	Pessoas de 40 anos e mais
A	4	1
B	6	3
C	10	2
D	5	2
Total	25	8

$$P = \frac{8}{25} = 0{,}320$$

Retiramos uma amostra de 2 famílias e estimamos a proporção de pessoas de 40 anos e mais:

Estimativa da proporção de pessoas de 40 anos e mais = Pessoas de 40 anos e mais na amostra ÷ Total de pessoas da amostra

Se tirarmos tôdas as amostras possíveis e computarmos o valor esperado desta estimativa, verificaremos que o mesmo não é igual ao valor real.

Amostra	Valor da estimativa
AB	4/10 = 0,400
AC	3/14 = 0,214
AD	3/ 9 = 0,333
BC	5/16 = 0,313
BD	5/11 = 0,455
CD	4/15 = 0,267
Soma	1,982

Valor esperado de $P = \frac{1}{6} \times 1{,}982 = 0{,}330$; $P = 0{,}320$

Há uma diferença, ou tendenciosidade, de 0,010.

Para as grandes amostras, êste tipo de tendenciosidade é normalmente muito pequeno e pode ser seguramente desprezado. Além do mais, a estimativa de razão é consistente, o que significa que à medida que cresce o tamanho da amostra, a tendenciosidade da estimativa aproxima-se de zero, ou, por outras palavras, o valor esperado da estimativa de razão aproxima-se do valor real.

Nos casos em que temos de escolher entre uma estimativa não tendenciosa e uma estimativa de razão, devemos considerar a correlação entre os dois itens envolvidos na razão. Se fôr elevada, a estimativa de razão será melhor. Se fôr baixa, ou negativa, será preferível a estimativa não tendenciosa. A questão pode ser posta de outra forma: que é menos variável - a característica que pretendemos estimar, ou sua razão em relação a outra característica cujo valor real conhecemos?

Embora se afirme que, para as grandes amostras, a tendenciosidade da estimativa de razão é pequena, devemos tomar cuidado, a fim de distinguir entre os casos em que aplicamos uma proporção estimada a um total, ou contagem, conhecidos, e os casos em que aplicamos essa proporção a um valor "estimado" de outra maneira.

Por exemplo, se levantássemos um inquérito à base de áreas da população do Brasil, usando uma amostragem probabilística, poderíamos considerar as duas seguintes maneiras de estimar o número total de pessoas com uma determinada característica.

1. Faz-se uma estimativa simples não tendenciosa baseada na amostra.

2. Usa-se a amostra a fim de estimar a proporção de pessoas com a característica considerada e multiplica-se por uma estimativa independente da população total do Brasil, baseada por exemplo, na taxa de crescimento entre os censos de 1940 e 1950, ou na projeção corrigida com apôio nos dados sôbre nascimentos, óbitos e migrações a partir do censo de 1950.

Se a estimativa independente da população total fôsse razoàvelmente aproximada do valor real, então, quase certamente lucrariamos para a maioria das características usando uma estimativa de razão; por outro lado, se a estimativa independente diferisse substancialmente do valor real, a diferença se refletiria nas estimativas de tôdas as características, contrabalançando qualquer vantagem decorrente do fato de que as razões são mais estáveis do que os números absolutos, para a maioria das características da população.

A fim de escolher o tipo de estimativa a usar neste caso, teríamos que especular sôbre o possível êrro de nossa estimativa independente da população total. Se o inquérito estivesse sendo levantado dentro de um ou dois anos após o censo mais recente, poderíamos concluir que seria conveniente uma estimativa de razão, visto como, não levando em conta como foi feita a estimativa independente, o seu êrro não poderia ser muito grande se a contagem censitária fôsse exata. Mas se o último censo tivesse sido levantado vários anos atrás, ou se fôsse recente mas o considerássemos não fidedigno, estaríamos mais inclinados a escolher a estimativa não tendenciosa, com o seu êrro de amostragem conhecido, de preferência à estimativa de razão com o seu êrro de amostragem um pouco menor mas com uma tendenciosidade de amplitude desconhecida. A escolha é difícil, requerendo não só o conhecimento dos erros de amostragem de diferentes tipos de estimativas, mas também um conhecimento considerável dos erros alheios à amostragem relacionados com os censos e com vários tipos de estimativas periódicas da população.

Além da estimativa simples não tendenciosa e das estimativas de razão, há um terceiro tipo comumente usado - a estimativa diferencial. Como a estimativa de razão, faz uso de uma característica "auxiliar" (para que se dispõe de uma contagem completa ou de uma estimativa independente de qualquer espécie) a qual é correlacionada com a característica que estamos estimando. De fato, é possível, interpretar a estimativa de razão como um caso especial da estimativa diferencial. Como exemplo da estimativa diferencial, considere-se o seguinte:

Temos um universo de estabelecimentos comerciais para os quais conhecemos o valor das vendas do ano passado (baseado em uma contagem completa), e retiramos uma amostra a fim de estimar as vendas dêste ano. Para os estabelecimentos da amostra, teremos informações tanto sôbre as vendas dêste ano como sôbre as do ano passado, de maneira que podemos fazer estimativas não tendenciosas das vendas de ambos os anos.

Uma estimativa diferencial das vendas dêste ano seria:

Total conhecido das vendas do ano passado	+	Estimativa não tendenciosa das vendas dêste ano	-	Estimativa não tendenciosa das vendas do ano passado

Em outras palavras, estimamos as vendas dêste ano através da diferença entre as vendas dêste ano e as do ano passado para os estabelecimentos da amostra, expandindo êste valor a fim de representar tôdas as lojas, e acrescentando-o ao total conhecido das vendas do ano passado para tôdas elas.

A estimativa diferencial desta forma não é tendenciosa, isto é, seu valor esperado é igual ao valor real. O exemplo acima ilustra o uso mais comum da estimativa diferencial: estimar o valor corrente de alguma característica para a qual temos um valor anterior baseado em uma contagem completa. Poderíamos também ter usado uma estimativa de razão para êste fim, conforme se segue:

$$\text{Estimativa de razão das vendas dêste ano} = \frac{\text{Estimativa não tendenciosa das vendas dêste ano}}{\text{Estimativa não tendenciosa das vendas do ano passado}} \times \text{Total conhecido das vendas do ano passado}$$

Em geral, pode-se dizer que a amostragem em ocasiões sucessivas a fim de estimar tanto os níveis correntes como as mudanças de mês a mês ou de ano a ano cria alguns dos problemas mais complicados do planejamento e estimação de amostras. Por exemplo, em uma amostra repetida deve ficar claro que obteremos melhores estimativas das mudanças mensais incluindo as mesmas unidades em nossa amostra, cada mês, do que retirando amostras diferentes cada mês. Por outro lado, verificamos que a cooperação dos informantes decresce após permanecerem na amostra diversos meses, de modo que nos arriscamos a obter uma tendenciosidade demasiado grande devido à omissão de declarações. Podemos ajustar esta situação mudando apenas parte da amostra cada mês, de modo que cada informante seja mantido na amostra por um período de vários meses, mas seja eventualmente substituído por outro. São necessários alguns tipos muito complexos de estimativas a fim de obter os melhores resultados possíveis de uma amostra dêste tipo.

Pràticamente, em qualquer aplicação da amostragem ao trabalho censitário, existirá a possibilidade de reduzir os erros de amostragem das estimativas usando alguma forma de estimativa de razão, estimativa diferencial ou outra, contanto que não seja simples e não tendenciosa. Não obstante, nem sempre tiramos plena vantagem destas possibilidades devido ao trabalho adicional envolvido nas tabulações. A fim de fazer uma estimativa simples não tendenciosa geralmente necessita-se apenas tabular o número de casos da amostra com uma determinada característica e multiplicar êsse total por um número inteiro. Em planejamentos mais complicados de amostras talvez tenhamos que tabular separadamente vários grupos nos quais se tenham usado diferentes taxas de amostragem, expandir cada total de grupo pelo inverso da taxa de amostragem, e somar os números resultantes. Êste ainda é um

processo relativamente simples.

Para a estimativa de razão, por outro lado, de um modo geral, devemos examinar detidamente todos os processos necessários à estimativa não tendenciosa, e depois, ainda, realizar outras multiplicações usando geralmente razões fracionárias. Se forem usadas razões diferentes para vários subgrupos, talvez seja necessário fazer várias destas multiplicações para cada item. Em uma amostra que esteja sendo usada a fim de estimar um grande número de itens, isto significa um trabalho adicional considerável.

Às vezes é possível simplificar os cálculos de estimativas de razões duplicando ou omitindo aleatòriamente um pequeno número de casos. Por exemplo, se retirarmos uma amostra de 5% de setores e verificarmos que contém 5,1% do total conhecido de pessoas recenseadas, poderíamos omitir aleatòriamente 2% ..... $\left(\frac{5,1 - 5,0}{5,0} \times 100\right)$ dos casos de amostra e depois usar um fator de 20 a fim de expandir os totais da amostra. Pode-se demonstrar que dêsse modo se obtém pràticamente um êrro de amostragem equivalente ao de uma verdadeira estimativa de razão (obtida pela multiplicação dos totais originais da amostra por 20,4).

Mesmo com estas simplificações, restam muitos casos em que o tipo mais eficiente de estimativa requer uma grande quantidade de cálculos. Em tais casos, será necessário usar um tipo de estimativa mais simples, embora pouco menos eficiente, a menos que se disponha de um equipamento moderno de cálculo, de alta velocidade.

Estimação e apresentação de erros de amostragem

A estimação de erros de amostragem, embora requeira uma quantidade relativamente pequena de trabalho em comparação com a coleta e tabulação de dados, apresenta problemas técnicos mais difíceis do que qualquer outra parte do processo de amostragem. Por esta razão é ela às vêzes desprezada. Com freqüência, verificam-se relatórios de inquéritos por amostra nos quais se usaram bons métodos de amostragem probabilística, mas não se deram quaisquer informações sôbre os erros de amostragem dos resultados. Isto é lamentável, visto como a principal vantagem de uma amostra probabilística consiste em se poder fazer afirmações precisas sôbre os erros prováveis dos resultados. Em qualquer relatório de resultados por amostra devemos sempre tentar dar uma indicação sôbre a magnitude dos erros de amostragem.

Por que é difícil estimar e publicar erros de amostragem? Há várias razões:

1. Complexidade dos cálculos - A fórmula para o êrro de amostragem de uma estimativa simples não tendenciosa, obtida de uma amostra aleatória simples, é bastante simples, e os valores numéricos são fáceis de calcular. Para os casos correntes, entretanto, em que o planejamento de amostra pode envolver conglomeração, estratificação e outras variações; em que se usa algum tipo de estimativa de

razão; e em que se utilizam sub-amostras a fim de estimar variâncias, etc. - nesses casos, as fórmulas para os erros de amostragem são bastante complicadas. Para um planejamento de amostra moderadamente complicado, o cálculo do êrro de amostragem para uma única característica pode requerer umas 20 ou 30 diferentes adições, subtrações, multiplicações, divisões e extrações de raízes quadradas. Tudo isto, e ainda o cálculo da soma dos quadrados dos valores da característica para as unidades individuais de amostragem, que se faria em quase todos os casos. Um planejamento realmente complexo pode requerer centenas de cálculos separados para cada item.

2. Grande número de estimativas - Em uma aplicação típica da amostragem ao trabalho censitário, tal como a obtenção de tabulações preliminares, é preciso fazer estimativas de centenas ou mesmo milhares de itens. O cálculo de erros de amostragem para todos êsses itens, embora possa ser relativamente simples para cada um, é um grande empreendimento.

3. Planejamentos de amostra para os quais não se dispõe de estimativa não tendenciosa da variância - Qualquer estimativa de erros de amostragem deve ser baseada na variação observada em relação à média de unidades individuais de amostragem incluídas na amostra. Há muitos planejamentos, entretanto, nos quais a amostra contém apenas uma unidade, no sentido em que é usada aqui, ou pelo menos sòmente uma unidade em cada estrato. Por exemplo, já assinalamos que tirar uma amostra sistemática consistente de cada $K^{ésima}$ unidade equivale a dividir a população em K conglomerados, consistindo cada um de elementos com intervalos de K unidades, e selecionar um dêstes conglomerados aleatòriamente. Uma estimativa não tendenciosa da variância teria que ser baseada na variação do valor médio de 2 ou mais dêstes conglomerados. Visto como a amostra inclui apenas um conglomerado, não podemos fazer uma estimativa não tendenciosa. Em outro planejamento de amostra comumente usado, um país é dividido em um grande número de estratos, contendo cada qual um grupo de municípios (ou condados), e apenas um município (ou condado) é selecionado de cada estrato. O problema neste caso é idêntico.

4. Dificuldade de explicar erros de amostragem em têrmos simples - Os conceitos empregados a fim de interpretar os erros de amostragem tal como o conceito de intervalo de confiança, não são simples. Às vezes as exposições sôbre erros de amostragem e as explicações de como usá-los não são dadas porque se julga que o consumidor dos dados não as entenderá ou interpretá-las-á incorretamente. Como resultado, o consumidor comum de dados interpretará as estimativas como se fôssem contagens completas, e o consumidor melhor informado, que conheça algo sôbre erros de amostragem, ficará relutante em usar os dados, visto que não sabe quais os riscos envolvidos.

Apresentamos uma série de problemas relacionados com o cálculo e a apresentação de erros de amostragem. Como fazer face a êstes problemas? Se-

guem-se algumas das resoluções elaboradas.

1. Uso de aproximação nos cálculos - Quando é impossível uma estimativa não tendenciosa da variância, devido ao planejamento da amostra, ou quando a fórmula exata é muito complexa, é possível, comumente, adotar uma aproximação, isto é, uma estimativa tendenciosa da variância, que esteja razoàvelmente próxima do valor real. Quando assim se procede, é comum escolher uma aproximação que dê uma super-estimativa da variância, de modo que nossas decisões quanto aos resultados significantes da amostra deverão ser tomadas com cautela.

2. Uso de amostras interpenetrantes - Um sistema que tem ganho popularidade nos últimos anos, principalmente nas áreas em que os recursos não permitem um grande volume de cálculos, é o uso das amostras interpenetrantes. Por isto entendemos que, em lugar de usar uma única amostra de tamanho n, usamos K amostras de tamanho n/K (K é geralmente 5 ou 10), selecionando cada uma das K amostras independentemente, mas usando o mesmo processo de seleção para cada uma. Começamos então a fazer estimativas de cada uma das K amostras, usando o mesmo processo de estimação para cada uma. O resultado líquido é que podemos estimar erros de amostragem de maneira muito simples, independentemente da complexidade do planejamento e dos processos de estimação da amostra, computando a variação das K estimativas sôbre seu valor médio, que é a estimativa baseada na soma das K amostras.

   Embora êste sistema pareça atraente à primeira vista, verificamos que há um preço a ser pago por esta simplificação de estimação da variância. Êste preço consiste em um aumento nos erros de amostragem, resultante do fato de que as K amostras interpenetrantes devem ser selecionadas independentemente. Assim, se K é 5 e o planejamento original pedia a seleção de uma unidade de cada um dos vários estratos, o planejamento usando amostras interpenetrantes requeria tornar os estratos 5 vêzes tão grandes e selecionar 5 unidades de cada um dos novos estratos. É quase certo que esta amostra será menos eficiente.

   O problema que enfrentamos ao decidir se devemos ou não usar êste sistema pode ser pôsto da seguinte maneira: que preferimos, um planejamento de amostra que reduza ao mínimo nossos erros de amostragem, mesmo que não possamos estimar êstes erros de amostragem exatamente? Ou um planejamento de amostra que dê maiores erros de amostragem, mas para o qual possamos obter estimativas muito precisas dêstes erros de amostragem?

3. Estimação de erros de amostragem para um número limitado de itens - Quando se usa uma amostra a fim de estimar um grande número de itens, é possível às vêzes estimar erros de amostragem para um grupo representativo dêstes itens, e com base neste grupo tirar algumas conclusões sôbre a magnitude ou o nível geral dos erros de amostragem para todos os itens. O processo usado, aproximadamente, é o que se segue:

a. Selecionar em um inquérito demográfico um grupo de itens, certificando-se de que tôdas as classes de itens no inquérito estão representadas, bem como de que êsses itens estão bem distribuídos de acôrdo com o tamanho, variando dos menores grupos à população total incluída no inquérito.

b. Calcular a variância para cada um dos itens selecionados.

c. Traçar as estimativas de variância em um gráfico, usando o tamanho da estimativa como eixo horizontal.

d. Traçar uma curva nos pontos marcados. Se o gráfico revelar que certos grupos de itens, que tendem a agrupar-se (como côr, religião, etc.), não parecem seguir o padrão geral, não incluir êstes pontos quando traçar a curva.

e. A curva traçada pode então ser usada para dar um valor aproximado do êrro de amostragem para uma estimativa de qualquer tamanho. (Como exemplo de erros de amostragem deduzidos dêste processo, ver os dois quadros do Anexo 2 da 5ª Palestra).

4. Crítica das exposições analíticas baseadas em resultados de amostras - Não há um único método geralmente aceito de apresentar e explicar os erros de amostragem em um relatório que contém dados baseados em uma amostra. Às vêzes os erros de amostragem são apresentados nos mesmos quadros, ao lado das estimativas. Com maior freqüência, são apresentados em outro quadro, como parte de uma introdução ou apêndice. Às vezes não há explicação alguma sôbre o significado dos erros de amostragem; em outros casos, dá-se uma explicação muito precisa, tècnicamente correta. O método usado dependerá de fatôres tais como a finalidade do relatório, e o público a que se destina.

Há, entretanto, um processo que pode e deve ser sempre seguido. Se o relatório contém uma análise dos dados, tôdas as exposições incluídas nesta análise devem ser testadas quanto à significância estatística, e modificadas as exposições conforme requeiram os resultados dêstes testes. Assim, se os dados indicam que o número médio de pessoas por família é maior nas áreas rurais do que nas urbanas, devemos computar o desvio-padrão da diferença entre as duas estimativas. Se o desvio-padrão fôr, digamos, o dôbro da diferença estimada, somos forçados a concluir que a diferença podia ter ocorrido simplesmente devido a erros de amostragem, e, portanto, ou não mencionaríamos isto em nossa análise, ou afirmaríamos que não era estatísticamente significante. Se entretanto a diferença fôr grande em relação a seu desvio-padrão (por exemplo, igual ao dôbro) deveremos mencionar esta diferença em nossa análise, e usá-la da maneira que parecer apropriada. Seguindo êste processo os elaboradores dos dados pelo menos não serão culpados de apresentar ao consumidor conclusões que não sejam justificadas pelos dados.

A fim de resumir o assunto dos erros de amostragem, podemos considerar os dois principais aspectos. O primeiro - cálculo dos erros de amostra-

gem - deve ser incumbência do técnico em amostragem, que pode deduzir as fórmulas corretas e recomendar o uso de métodos apropriados de aproximação, quando necessário. O segundo aspecto - apresentação - é algo que deve ser decidido conjuntamente por tôdas as pessoas interessadas no assunto. O técnico censitário deve decidir, com base no seu conhecimento de como e por quem os relatórios serão usados, qual será o método mais conveniente de apresentar erros de amostragem. O técnico em amostragem deve ser consultado a fim de garantir que as explicações dadas sejam tècnicamente corretas.

NONA PALESTRA

Observações gerais sôbre a aplicação da amostragem nos censos e inquéritos

1. Fontes para amostras - Já discutimos a importância de dar uma definição precisa ao universo de que se tira a amostra. De importância, também, é a escolha da "fonte" de que a amostra será selecionada. Por fonte queremos dizer: grupo de questionários, coleção de mapas, arquivo de cartões, ou outros materiais de que se pode retirar a amostra diretamente. Cada fonte conterá alguns, ou todos os elementos do universo; além disso, poderá conter outros elementos que não estejam no universo.

Considere-se o problema de planejar uma amostra geral de população do Distrito Federal, digamos, dois anos após o levantamento de um censo. Uma das fontes que poderíamos usar seriam os questionários do censo. Outra poderia ser uma lista, ou um mapa, com a divisão completa do Distrito Federal em quarteirões ou unidades semelhantes.

A segunda fonte - lista de quarteirões - "conteria" todos os elementos do universo, visto como cada morador do Distrito Federal poderia estar associado ùnicamente a um daquêles quarteirões. Entretanto, a simples seleção de uma amostra de quarteirões não seria suficiente, no caso, visto como se deseja selecionar uma amostra de famílias ou indivíduos, e coligir as informações necessárias. Além do que, e como a população dos quarteirões varia consideràvelmente de tamanho uma amostra com igual probabilidade de seleção para cada um dêles seria ineficiente.

A primeira fonte - questionários censitários - poderia ser usada de modo satisfatório afim de retirar uma amostra eficiente de indivíduos. Esta fonte, evidentemente, não incluiria todos os indivíduos do universo a ser selecionado, por outro lado, contaria alguns já não existentes no mesmo universo. Poderia de qualquer forma ser usado para selecionar uma amostra de domicílios (particulares e coletivos), ao invés de indivíduos. Ainda assim, porém, haveria um grupo de domicílios não existentes na época do censo, que ficariam fora da amostra. Em uma área de rápido crescimento, como o Distrito Federal, o número de novos domicílios acrescidos em um período curto poderá ser substancial.

Felizmente, podemos chegar a uma solução que utiliza ambas as fontes mencionadas e em grande parte se beneficia das vantagens respectivas.

Procede-se da maneira seguinte:

a. Usam-se os questionários para selecionar uma amostra eficiente de domicílios (amostra aleatória simples, ou constituída de pequenos conglomerados). Recenseiam-se tôdas as pessoas que vivam naqueles domicílios.
b. Seleciona-se uma amostra de quarteirões. Nestes quarteirões faz-se uma listagem de todos os domicílios e determinam-se quais os que não foram incluí

dos no censo. Recenseiam-se todos, ou uma sub-amostra dêstes domicílios.

A fim de tornar êste planejamento mais eficiente, seria conveniente fazer um ligeiro inquérito preliminar para identificar os quarteirões com grande quantidade de novas construções e dar a êstes uma chance maior de seleção na amostra de quarteirões. O grosso da amostra viria, contudo, da amostra mais eficiente de domicílios censitários, visto como o grosso do universo estaria contido nestes domicílios.

O exemplo acima ilustra dois princípios básicos. Primeiro: convém geralmente, usar uma lista ou seu equivalente como fonte, quando a lista contiver uma grande proporção dos elementos do universo. Segundo: visto como as listas raramente são completas ou atualizadas, com freqüência será necessário suplementar a amostra por listas com uma amostra à base de áreas, a fim de assegurar aos elementos que não figuraram na lista representação na amostra global.

No exemplo acima, consideraram-se sòmente duas fontes possíveis. Em muitos estudos, haverá um grande número delas. Considere-se, por exemplo, o problema de retirar uma amostra do universo constituído de todos os médicos residentes em certo Estado. Algumas fontes possíveis são:

1. Listas de graduados de escolas brasileiras de Medicina
2. Listas de membros de sociedades médicas estaduais ou nacionais
3. Registros de diplomas na Diretoria do Ensino Superior, do M.E.C.
4. Arquivos do registro do Serviço Nacional de Fiscalização da Medicina, do Ministério da Saúde
5. Questionários da Campanha Nacional de Estatística do IBGE, dando os nomes de médicos de cada município
6. Respostas do censo mais recente, apresentando a ocupação de cada pessoa recenseada
7. Lista dos municípios e distritos do Estado

A seleção da melhor fonte ou combinação de fontes, em tais casos, requer o estudo cuidadoso das propriedades de cada uma. Embora a amostragem através de listas ou arquivos de cartões pareça simples e conveniente à primeira vista, logo se verifica que, na maioria dos casos, contêm, em maior ou menor grau, omissões, duplicações, itens obsoletos, declarações incorretas e outros defeitos. Ao invés de aceitarmos essas fontes sem restrições, devemos fazer testes objetivos a fim de determinar sua relação com o universo em estudo; e ter em mente que uma lista ou arquivo de cartões incompleto pode ser usado freqüentemente com vantagem se fôr suplementado de maneira adequada com uma amostra à base de áreas.

* * *

2. Listas de grandes estabelecimentos - É de grande interêsse para a eficiência dos planejamentos de amostra, em quase todos os setores da atividade censitária,

o conhecimento prévio dos elementos, ou estabelecimentos, excepcionalmente grandes em relação à média. Nas amostras de população, devemos considerar o domicílio coletivo, que poderá conter várias centenas de pessoas. No setor agrícola, verifica-se, com freqüência, que alguns estabelecimentos agropecuários representam uma proporção substancial da área total dos estabelecimentos. No caso da indústria, um pequeno número de estabelecimentos manufatureiros pode concorrer com a maior parte da produção total.

Êsses grandes estabelecimentos, principalmente na agricultura e indústria, constituem tanto uma vantagem como um perigo para quem realiza a amostra. É uma vantagem possuir essas unidades, porque, mediante um levantamento completo de um número muito pequeno de unidades, pode-se conhecer uma proporção substancial das características investigadas sem erros de amostragem. Nesses casos, pode-se tolerar um êrro de amostragem relativamente grande nas estimativas de totais para as restantes unidades menores e ainda obter erros de amostragem aceitáveis para os totais globais. O que, por sua vez, permite usar uma amostra muito menor do que seria possível em outras condições.

O perigo neste caso é que, ao estratificar o universo de modo que as grandes unidades possam ser tratadas em separado, pode-se acidentalmente deixar de incluir algumas das grandes unidades no grupo apropriado. Uma ou duas grandes unidades que entrem errôneamente em grupos de amostra podem elevar o êrro de amostragem muito acima do que seria obtido se tôdas as grandes unidades fôssem identificadas adequadamente antes da amostragem. Portanto, é conveniente, e, em última análise, econômico, qualquer esfôrço para a identificação dos grandes estabelecimentos. Não basta, por exemplo, em um inquérito de estabelecimentos manufatureiros, simplesmente listar os grandes recenseados no censo mais recente (esta listagem serve como primeira medida e deve ser cuidadosamente executada e verificada). Além disso, devem-se consultar outras fontes, como as associações comerciais, os registros de repartições do govêrno, dados periódicos coligidos por agentes municipais de estatística e indivíduos bem familiarizados com a indústria nas localidades abrangidas pelo inquérito. Mesmo que esta investigação adicional forneça apenas um número pequeno de estabelecimentos grandes adicionais, o resultado significará uma redução valiosa no êrro de amostragem.

* * *

3. O Censo de População como base para inquéritos por amostragem - Muitas pessoas têm a crença errônea de que é impossível realizar um inquérito da população por amostra, a menos que já tenha sido levantado em época próxima um censo completo da área em questão. Como já descrevemos, no entanto, a técnica de amostragem à base de áreas (amostragem de conglomerado) torna possível usar as técnicas da amostragem probabilística sem possuir uma lista dos elementos do universo. Não obstante, a existência de um censo recente da mesma população torna possível o planejamento muito mais eficiente de amostra, combinando-se amostragem à base de áreas e de listas, conforme foi discutido na Secção I desta palestra. Por-

tanto, uma questão a ser considerada no planejamento de um censo de população é: como organizá-lo, a fim de aumentar seu valor como base para futuras amostras, consistentes com outros objetivos?

A contribuição mais importante que o censo pode emprestar nesse sentido é fornecer informações sôbre número e características das unidades domiciliares localizadas em pequenas áreas. As informações por setores já representam um auxílio útil à amostragem, mas, para o máximo aproveitamento, o censo deve fornecer as seguintes informações:

a. Cada área urbana deve ser dividida em quarteirões numerados, e o recenseador deve registrar o número do quarteirão de cada unidade domiciliar recenseada (além de auxiliar a amostragem, êste processo torna possível fazer tabulações censitárias por quarteirões de cidade). Essas tabulações têm sido consideradas muito úteis pelos administradores municipais, autoridades escolares, funcionários da saúde pública, etc. nos países que as adotam.

b. Em outras áreas, deve-se fornecer ao recenseador o mapa mais detalhado disponível do seu setor, e instruí-lo no sentido de nele registrar a localização de cada unidade domiciliar recenseada, mediante um símbolo e número apropriado. Êste processo também representa mais que um auxílio à amostragem: ajuda o recenseador a verificar se abrangeu todo o setor, e torna muito mais fácil para outra pessoa verificar o seu trabalho.

É desnecessário dizer que, sendo útil êsse material, deve-se exigir de cada recenseador que siga êstes processos, e devolva seu mapa com os questionários completados. Embora isto possa parecer uma carga adicional desnecessária na época do censo, seu valor será verificado quando chegar a ocasião de usar os materiais censitários como base para amostragem.

* * *

4. Relações entre os erros de amostragem e os erros alheios à amostragem - Como todos sabem, mesmo uma contagem completa está sujeita a muitas espécies de erros que a fazem diferir do "valor real". Êstes erros originam-se de respostas inexatas, não observância das instruções pelos recenseadores, codificação incorreta, erros de perfuração e tabulação mecânica, etc.

Êstes mesmos fatôres, e outros adicionais, prevalecem para uma investigação baseada em amostra. Portanto, devemos lembrar-nos que a diferença entre nossa estimativa de amostra e o valor real deve-se apenas em parte a erros de amostragem, e que também pode haver alguma diferença originada dêstes outros tipos de erros, a que nos referimos como erros alheios à amostragem.

Quais são as implicações práticas disso? Em primeiro lugar, ao interpretar os resultados de um inquérito por amostra, não devemos permitir que, pelo fato de se ter usado uma grande amostra (e portanto, obtido estimativas com erros de amostragem muito pequenos) se esqueça que as estimativas podem diferir bastante do valor real, devido aos erros alheios à amostragem. Tomando

um exemplo extremo, suponha-se que tentamos estimar a renda média familiar no Brasil através de uma amostra probabilística de 100 000 famílias, em que se fêz a cada chefe de família uma única pergunta, "Qual foi sua renda familiar no mês passado?" Com uma amostra de 100 000 famílias, o êrro de amostragem das estimativas seria expresso por um coeficiente de variação de menos de um por cento. Entretanto, o êrro de amostragem apenas dá uma idéia da diferença provável entre a estimativa e o valor que seria obtido entrevistando tôdas as famílias do Brasil mediante processos idênticos (isto é, fazendo apenas uma pergunta sôbre renda). Entretanto, se prèviamente estabelecer-se um conceito preciso de renda familiar, essa contagem completa forneceria um valor que também poderia diferir bastante do "valor real".

Em segundo lugar, a consideração de erros alheios à amostragem leva-nos ao seguinte importante princípio: Em qualquer investigação estatística, o método ótimo será um que resulte, em média, nas menores diferenças entre os valores reais e as estimativas, levando em conta tanto os erros de amostragem como os erros alheios à amostragem. Êste princípio freqüentemente é ignorado. Um êrro comum é gastar quase todos os recursos de um inquérito a fim de reduzir os erros de amostragem tirando uma amostra grande e gastar pouco ou nada em uma tentativa de controlar a qualidade dos dados.

Vejamos algumas das maneiras pelas quais podemos usar nossos recursos com o intuito de reduzir os erros alheios à amostragem. Em primeiro lugar podemos fazer inquéritos-pilôto a fim de experimentar diferentes processos de campo e questionários, e escolher o mais eficiente. Antes e durante o levantamento real, podemos cuidar de fornecer treinamento e supervisão adequada aos recenseadores. Em um inquérito permanente, ou em um censo, podemos fazer verificações posteriores para determinar a adequabilidade da cobertura e do conteúdo, e assim, saber, quais os itens que causam dificuldades, e porquê. Ao fazer tabulações, podemos verificar o trabalho de codificação e perfuração. Esta última é uma fonte de erros alheios à amostragem que tem sido ressaltada freqüentemente. Em muitos casos, pode-se substituir a verificação completa das operações de processamento pela inspeção por amostragem.

É difícil, sem dúvida, medir precisamente o efeito de algumas destas atividades na melhoria da exatidão global. Não obstante, a experiência tem demonstrado, e sem dúvida continuará a demonstrar, que uma proporção substancial de recursos dedicados à coleta de dados estatísticos deve ser dirigida no sentido da redução dos erros alheios à amostragem.

* * *

5. Contrôle da execução de processos de amostragem - Estivemos falando sôbre erros alheios à amostragem, em geral. Examinemos agora um grupo particular de erros alheios à amostragem - aquêles diretamente relacionados ao uso da amostragem. Neste grupo incluem-se erros na seleção da amostra e na aplicação dos processos de estimação apropriados ao planejamento de amostra utilizado.

Estando outros fatôres equiparados, é melhor usar um plano de amostragem no qual o contrôle completo da seleção seja realizado na repartição central, de maneira que os recenseadores possam ser instruídos precisamente sôbre quais as pessoas ou estabelecimentos que devem ser incluídos na amostra. Entretanto, isto nem sempre é praticável. Suponha-se, por exemplo, que temos uma lista de famílias e desejamos entrevistar um adulto, selecionado ao acaso, de cada família. Não podemos fazer esta seleção até que saibamos o número de adultos de cada família. Deveremos mandar o recenseador ao local a fim de determinar êste número e trazê-lo à repartição central, de maneira que aí se possa fazer a seleção aleatória, e depois mandá-lo de volta, a fim de entrevistar a pessoa selecionada? Ou dar instruções e material ao recenseador para que selecione um dos adultos aleatòriamente de cada casa, de modo que possa determinar o número de adultos, fazer a seleção aleatória e completar a entrevista, tudo em uma só visita? Evidentemente, o segundo método é mais econômico, contanto que se possa planejar instruções simples, que o recenseador possa seguir ao fazer a seleção aleatória. Isto não é tão fácil quanto parece. Se deixarmos que o recenseador faça a seleção, devemos ter meios de verificar o seu trabalho. Seria possível, por exemplo, que o recenseador manipulasse o processo de seleção da amostra de tal maneira que a pessoa selecionada sempre fôsse um dos adultos presentes por ocasião de sua primeira entrevista, de modo a evitar a necessidade de outras visitas. Se êle fôsse pago à base de entrevista, teria um forte incentivo para fazer exatamente isso. Ou poderiam ocorrer erros simplesmente devido à interpretação errônea das instruções.

Seja feita inteira ou parcialmente na repartição central, importa para a seleção que haja instruções detalhadas por escrito, a respeito de todo o mecanismo do processo. Estas instruções devem prever tôdas as situações que possam ocorrer na seleção da amostra. Por exemplo: a) ao selecionar uma amostra de questionários de família, que se deve fazer com os questionários suplementares usados para grandes famílias? b) supondo que se hajam preenchido questionários referentes a domicílios vazios, como deverão ser tratados?

Um êrro comum na seleção de amostra é o de fazer substituições para as unidades que estão incluídas na fonte mas não pertencem à população de que se faz a amostra.

Em um inquérito de população, admita-se que foram selecionadas 5% das casas na área do inquérito, verificando-se que um pequeno número delas se acham desocupadas, ou foram demolidas. Devemos substituir outras casas ocupadas por estas? A resposta é clara, se percebermos que a amostra de 5% pode ser efetuada, em média, de maneira a conter não só 5% das casas ocupadas mas também 5% das casas vazias, demolidas ou qualquer outro tipo incluído na fonte da qual a amostra foi selecionada.

Isto não significa que não devamos nunca selecionar uma amostra suplementar, se a original revelar-se demasiado pequena, No exemplo acima, caso verifiquemos que a fonte contém grande proporção de casas desocupadas (de modo

que o número de famílias da amostra é muito menor que se esperava), podemos selecionar uma amostra adicional, usando qualquer método que pareça apropriado, contanto que se mudem os processos de estimação a fim de adequadamente levar em conta a amostra suplementar.

As pessoas familiarizadas com censos compreendem a importância do contrôle, isto é, de certificar-se que em cada estágio da coleta e processamento, cada questionário, setor, pasta ou outra unidade é apropriadamente considerada. Quando se trabalha com amostras, o contrôle adquire maior importância, porque cada unidade de amostra representa várias unidades na população. Felizmente, a redução no volume do material, quando se emprega amostra, torna mais fácil manter um contrôle exato. Sem um bom contrôle, mesmo uma grande amostra não pode produzir dados fidedignos.

* * *

6. Registros e relatórios de inquéritos por amostra - Sempre que se seleciona uma amostra, para qualquer fim, deve-se manter um registro completo de cada estágio dos processos usados na seleção. Devem-se rotular cuidadosamente fôlhas de trabalho e listagens, de maneira que qualquer pessoa interessada possa determinar exatamente quais os passos levados a efeito, e em que ordem. Os mesmos comentários aplicam-se aos materiais usados na confecção de estimativas e no cálculo de variâncias.

Não se fazem estas coisas apenas para satisfazer uma curiosidade vã. Freqüentemente ocorrem questões sôbre a validade de determinadas estimativas e às vêzes estas questões só podem ser determinadas por um exame cuidadoso dos registros básicos de seleção e estimação da amostra. Além disso, êstes registros serão inestimáveis a fim de planejar novos inquéritos no mesmo setor. Os elementos apropriados a êste fim incluem não só os dados sôbre variâncias e componentes de variâncias, como também os dados sôbre os custos ligados ao planejamento, coleta e tabulação do inquérito.

A finalidade de uma investigação estatística, seja um censo ou um inquérito por amostra, é obter dados estatísticos. A publicação dos resultados dará geralmente ênfase primária à apresentação e análise dos resultados. Entretanto, também é importante que os dados e a análise sejam acompanhados por uma descrição do planejamento da amostra ou do inquérito, e uma exposição de erros de amostragem e outros detalhes de metodologia necessários a uma interpretação correta dos resultados. Esta informação pode aparecer como parte do relatório básico, como apêndice ou publicação separada, dependendo do tipo de investigação e do emprêgo que será feito dos dados.

A fim de ter certeza de que todos os dados pertinentes à metodologia estão considerados, não se pode fazer mais do que seguir o esbôço fornecido na publicação das Nações Unidas intitulado "The Preparation of Sampling Survey Reports" (Item 3 da Secção IV da Bibliografia).

Se uma investigação estatística fornece a informação especificada nesse esbôço, qualquer interpretação errônea dos resultados será de responsabilidade do consumidor, e não da repartição que fêz a investigação.

Êste documento das Nações Unidas, embora destinado primàriamente a servir como ajuda na preparação de relatórios, pode ser lido com bastante proveito por qualquer pessoa que se preocupe com o planejamento, execução ou interpretação de inquéritos por amostragem.

A AMOSTRAGEM NO CENSO DE POPULAÇÃO E HABITAÇÃO DE 1950*

Joseph Steinberg e Joseph Waksberg

Bureau do Censo

Um importante passo à frente no uso da amostragem, nos trabalhos consitários, teve lugar no Censo de População e Habitação de 1950. Como parte do Censo de 1940, o Bureau do Censo contou pela primeira vez com métodos modernos de amostragem a fim de aumentar a eficiência de suas operações (1, 2). Em 1950, fêz-se uso mais extenso da amostragem de características da população do que em 1940, e pela primeira vez se coligiram muitos itens sôbre habitação à base de amostra. Fizeram-se muitas outras aplicações da amostragem para a obtenção de tabulações rápidas e no processamento e contrôle das operações. As finalidades do emprêgo de cada processo eram ou reduzir o custo do Censo diminuindo as despesas de coleta ou tabulação, ou reduzir a duração do tempo necessário para divulgar publicações. Em muitos casos, ambos os objetivos foram servidos simultâneamente. Como resultado da redução das despesas, o Bureau do Censo estava habilitado a obter informações sôbre maior número de itens e fornecer tabulações mais detalhadas do que seria possível de outra forma. Ao mesmo tempo, foi possível diminuir o custo total e melhorar a supervisão e contrôle do trabalho de campo.

A metodologia usada na amostragem, para vários fins, de um universo de mais de 150 milhões de pessoas e 45 milhões de unidades domiciliares, e os vários tipos de problemas que foram resolvidos a fim de satisfazer os requisitos de fidedignidade e custo podem ser de interêsse aos técnicos no campo da amostragem demográfica e a outros consumidores de estatísticas consitárias.. Segue-se uma breve descrição das aplicações mais importantes da amostragem que fizemos e os ensinamentos que colhemos.

As amostras básicas de população e habitação - Para muitos tipos

---

* Êste documento foi preparado para ser apresentado na III.ª Reunião Anual da Associação Americana de Estatística, Boston, 28 de dezembro de 1951.

(1) Hansen, M.H., e Hurwitz, W.N.: "Sampling Methods Applied to Census Work", History, Operations and Organization of the Bureau of the Census, Vol. I, págs. 29-45, Bureau do Censo, novembro de 1946.

(2) Stephan, F.F., Deming, W.E. e Hansen, M.H.: "The Sampling Procedure of the 1940 Population Census", Journal of the American Statistical Association, Vol. 34, págs. 615-630, dezembro de 1940.

de características da população e das unidades domiciliares, as necessidades de informações são bem supridas por dados obtidos através de uma amostra suficientemente grande que forneça a necessária fidedignidade. Para certos dados, dos quais a lei não exige contagens completas, muitas vêzes é inútil a despesa de coligí-los de todo o universo. Uma proporção bastante grande das informações coligidas no Censo de 1950 foi obtida ou por uma amostra de 20% da população ou por uma amostra de 20% das unidades domiciliares, levantadas simultâneamente com os itens obtidos na base de 100%. Obtiveram-se características adicionais da população por meio de uma amostra de 3 1/3%. Visto como a precisão dos resultados da amostra depende principalmente do seu tamanho absoluto, e não da proporção da população incluída na mesma, a precisão relativa dos dados da amostra de 20% será menor para as áreas pequenas do que para as áreas grandes. Os ítens incluídos na amostra de 20% foram aquêles cujos dados-resumo deviam ser tabulados para áreas pequenas mas não com grandes detalhes, e cuja análise não requeria contagens completas. As tabulações de dados obtidos por amostra deviam ser feitas geralmente para áreas de 2 500 habitantes ou mais, mas as tabulações cruzadas só para áreas muito maiores. A amostra de 3 1/3% foi usada para características da população cuja publicação foi planejada apenas para áreas muito grandes, tais como o total do país, regiões, estados ou cidades com 100 000 habitantes ou mais, com graus variados de tabulações cruzadas.

Determinar se um item devia estar na amostra de 20% ou na contagem completa dependeu da precisão e da quantidade de detalhes que a informação requeria. Além disso, foram levadas em conta as despesas envolvidas e a exatidão esperada nas respostas a algumas das perguntas. As contagens da população por idade, sexo, raça e residência e algumas características básicas das unidades domiciliares, por exemplo, foram consideradas fundamentais e necessárias com alta precisão para tôdas as áreas, tanto pequenas como grandes e jamais houve dúvida de que estas deviam ser obtidas na base de censo completo. Outras perguntas foram mantidas na base de 100% ou porque os dados eram relativamente econômicos em um censo completo, ou porque, em muitos casos, eram necessários com um grau muito elevado de exatidão (por exemplo, nacionalidade), ou ainda porque se desejavam, com freqüência, tabulações detalhadas para áreas relativamente pequenas. Além disso, parecia imprudente ir longe demais em um único censo no sentido da dependência da amostragem, sem experiências adicionais quanto à eficiência da amostra de 20% em fornecer resultados satisfatórios em relação a áreas relativamente pequenas ou tabulações detalhadas para áreas maiores. A transferência de outros itens para a amostra será considerada, sem dúvida, no censo seguinte. Ao mesmo tempo, a experiência do Censo de 1950 pode sugerir que alguns itens sejam coligidos em uma amostra maior ou em censo completo (por exemplo, ano em que o prédio foi construído).

As características obtidas à base de amostra podem ser encontradas examinando-se o questionário básico de população e habitação (Figura 1). Os itens da amostra de 20% da população incluem entre outros - renda, instrução e migração. A amostra de 3 1/3% da população incluiu principalmente perguntas detalhadas sôbre estado conjugal e fecundidade. As perguntas sôbre habitação obtidas à base de amostra abrangem itens como ano de construção, equipamento de aquecimento, etc.

As amostras de 20% e 3 1/3% da população também são a base de parte do programa de tabulação de alguns tipos de dados coligidos de tôda a população. Alguns dos dados básicos de família, por exemplo "número de casais", estão sendo tabulados da amostra de 20%. As tabulações analíticas mais elaboradas dos dados de família, que só serão publicadas para áreas relativamente grandes, foram planejadas numa amostra que é basicamente de 3 1/3%, principalmente devido ao alto custo da transcrição.

Cada questionário de população continha 30 linhas, uma para cada pessoa recenseada. De cinco em cinco linhas figurava uma linha de amostra, com quesitos adicionais para a amostra de 20%. As informações da amostra de 3 1/3% foram obtidas para as pessoas recenseadas na última linha da amostra de 20% de cada questionário. A amostra da habitação foi organizada de maneira diferente. O processo utilizado foi grupar as perguntas de amostra em 5 séries e obter informações sôbre uma série para cada unidade domiciliar. Cada questionário previa 12 unidades domiciliares.

As amostras de população e habitação são ambas amostras basicamente sistemáticas, sendo a unidade de amostragem uma pessoa e uma unidade domiciliar, respectivamente. A experiência de 1940 (1) e os estudos realizados antes do Censo de 1950, utilizando material do Censo de 1940 (Figura 2), indicaram que se poderiam esperar alguns problemas de uma amostra sistemática se não se introduzissem variações no padrão sistemático. Isto se deve principalmente às praxes censitárias que resultam em padrões cíclicos de relacionar as pessoas, recenseando-se os chefes da família na primeira linha de um questionário, pelo menos uma vez, e freqüentemente, várias vêzes, em cada um dos 250 000 distritos censitários. Esperava-se que alguns desvios de um padrão de amostragem sistemática fôssem causados pela própria presença, nos questionários preenchidos, de linhas que não continham registros e linhas com observações dos recenseadores. A fim de reduzir ainda mais o efeito do padrão sistemático, usaram-se cinco versões diferentes de questionário. Como resultado, cada linha do questionário de população foi parte da amostra em uma e sòmente uma das cinco versões. De maneira semelhante, cada grupo de amostra de habitação foi relacionado em primei-

(1) Stephan, F.F., Deming, W.E. e Hansen, M.H.: "The Sampling Procedure of the 1940 Population Census", Journal of the American Statistical Association, Vol. 34, págs. 615-630, dezembro de 1940.

ro lugar em apenas um dos cinco questionários. Dentro de cada distrito censitário (que possui em média 200 unidades domiciliares), os questionários foram divididos de maneira aproximadamente igual entre as cinco versões.

O planejamento básico do questionário usado nestas amostras tem efeitos muito importantes sôbre o tipo de dados que podem ser publicados. Por exemplo, na amostra de 20% da população, a unidade de amostragem foi a pessoa. Por conseguinte, para a maioria das características da população obtidas na base de amostra, não se dispõe de informações para tôda a família. Isto significa que não se podem fazer algumas classificações cruzadas de dados de amostra que poderiam ser convenientes, do ponto de vista analítico. De maneira semelhante, também não se podem fornecer classificações cruzadas da maioria dos itens de habitação obtidos por amostra e as classificações cruzadas dos itens de população e habitação por amostra só são possíveis na base de uma amostra de 4%. Entretanto, o plano adotado foi simples e compreendeu instruções simples. Além disso, a facilidade com que os processos podiam ser seguidos de forma mecânica pelos recenseadores e o fato de que a experiência censitária de 1940 indicava que seria bem sucedido um tipo de amostra de população por linha, levou a empregar o planejamento.

Experimentou-se a praticabilidade e eficiência do emprêgo da família censitária, em lugar do indivíduo, como unidade de amostra. Isto se fêz em pequena escala em provas experimentais anteriores ao Censo e em maior escala em uma série de áreas no Censo. Os resultados das análises destas provas orientarão o Bureau na determinação da natureza da amostra em Censos subseqüentes. A impressão até hoje é que o uso da família censitária como unidade de amostra tem várias vantagens e desvantagens. Um dos principais problemas é a dificuldade de fornecer instruções simples aos recenseadores sôbre a forma de tratar as instituições e outros grandes domicílios coletivos que não podem ser incluídos na amostra baseada na família censitária. Contudo, uma amostra de família censitária permite a apuração de estatísticas de família na base de amostra.

Na publicação dos dados baseados na amostra de 20%, as estimativas do número de pessoas e unidades domiciliares com características especificadas estão sendo obtidas, em todos os casos, multiplicando-se por cinco o número que figura na amostra com estas características. As estimativas de percentagens estão sendo obtidas, em cada caso, usando valores de amostra tanto para o numerador como para o denominador. Seria possível, em muitos casos, computar estimativas mais perfeitas de totais usando-se processos de estimação de razões (Isto está sendo esperado para a amostra de 3 1/3% da população e a amostra de 4% da habitação). Estas estimativas de razões são particularmente úteis quando o total em questão representa uma alta percentagem de outra classe, para a qual tanto se pode obter uma estimativa por amostra como uma contagem completa. As

estimativas de razão não estão sendo usadas para estas estimativas de amostra devido ao maior custo necessário à sua preparação do que para as estimativas deduzidas multiplicando-se os mesmos resultados por cinco. Entretanto, nos dados publicados, os consumidores estão sendo supridos de informações suficientes, de maneira que podem aplicar geralmente estimativas de razão quando o desejarem.

Deve-se observar que podem resultar vantagens notáveis nas estimativas de características da população através das estimativas de razão, em certas condições. Por exemplo, se em um condado* hipotético as estimativas do número e percentagem de pessoas de 7 a 13 anos de idade que estão matriculadas na escola são 900 e 90%, respectivamente, o resultado do emprêgo de uma estimativa de razão para a população total de 7 a 13 anos de idade é reduzir o êrro padrão de cêrca de 55 para cêrca de 20. Se a percentagem fôsse inferior a 90%, as vantagens seriam menores. A Figura 3 mostra a redução no êrro-padrão que deve ser esperada sob várias condições. Conforme se pode ver, a menos que as proporções sejam bastante altas, as vantagens resultantes não são bastante grandes para justificar as computações requeridas. Para a amostra de 20% das características da habitação, visto como não são possíveis tanto contagens completas como estimativas por amostra para detalhes maiores que o número total de unidades domiciliares, a redução do êrro-padrão geralmente é insignificante, não se recomendando estimativas de razões.

Uma secção posterior dêste documento examinará os contrôles estabelecidos para evitar a tendenciosidade nas amostras, e os métodos usados na estimação do grau de variabilidade da amostragem a ser esperado.

As amostras para as tabulações preliminares** - Um dos problemas enfrentados pelo Bureau na época do Censo Decenal foi se devia divulgar dados preliminares apresentando as características obtidas no Censo ou publicar apenas os resultados definitivos. A tarefa de crítica, codificação e tabulação dos dados antes de publicar os resultados finais era tremenda. Os dados para o país em conjunto e grandes sub-áreas tinham interêsse imediato para muitas decisões políticas. Durante a década passada, muitas estatísticas, para o país em conjunto, eram obtidas através do Inquérito Corrente de População (1, 2). Entretanto, para as principais sub-áreas, só se dispunha de dados fragmentários. A fim de atender a estas exigências de informações preliminares, decidiu-se pu-

---

*N.T. - Equivale a município.

**Preparado com Albert Mindlin

(1) Hansen, M.H. e Hurwitz, W.N.: "A New Sample of the Population", Estadistica, Vol. II, págs. 483-497, dezembro de 1944.

(2) Goldfield, E.D.; Steinberg, J. e Welch, E.H.: "The Monthly Report on the Labor Force", Estadistica, Vol. VI, págs. 66-67, março, 1948.

blicar informações do Censo de 1950 à base de amostra, sòmente para a nação, quatro regiões, os dez maiores estados e as 57 maiores áreas metropolitanas padronizadas do país. As publicações das informações sôbre população e habitação (1) foram divulgadas durante a primeira metade de 1951. Os dados-resumo dos Estados Unidos foram publicados em fevereiro de 1951, mais de um ano e meio antes da publicação antecipada dos resultados finais para os Estados Unidos.

Estas tabulações preliminares por amostra também supriram o Bureau de um instrumento conveniente para verificações e análises internas. Por exemplo, fêz-se uso extenso da amostra preliminar a afim de determinar taxas de omissão de declarações, no Censo, em relação a várias características. Outro emprêgo importante visava oferecer orientação na determinação de especificações da tabulação final.

Foram atendidas, sem dúvida, as necessidades públicas através dêste programa antecipado por amostra. Por outro lado, foi possível enfrentar muitos problemas com muito mais antecedência. A amostra preliminar revelou a extensão das diferenças existentes entre os dados coligidos no Censo e o Inquérito Corrente de População. Além disso, fizeram-se tabulações especiais a fim de satisfazer necessidades importantes não atendidas pelo programa normal de tabulação.

Empregaram-se dois métodos a fim de obter êstes dados preliminares - um para a nação, regiões e os estados maiores, e outro para as áreas metropolitanas padronizadas. A primeira amostra foi planejada com o fim de fornecer fidedignidade aproximadamente igual de características apresentadas por uma proporção fixa da população total em cada um dos estados, e de modo semelhante, para cada uma das regiões. Isto resultou em uma amostra de tamanho aproximadamente igual em cada um dos dez estados e também nas quatro regiões. A amostra foi estratificada por distritos censitários (ED) urbanos, rurais e institucionais, selecionando-se em cada estrato uma amostra de ED e uma sub-amostra de pessoas e unidades domiciliares. As estimativas das variâncias de unidades dentro dos ED e entre os ED foram feitas na base de outros estudos correlatos que, considerados com as despesas envolvidas, nos possibilitaram determinar um planejamento ótimo. A amostra final teve uma taxa dentro do ED de cêrca de 1 em 60 pessoas e 1 em 60 unidades domiciliares. A amostra global conteve cêrca de 150 000 pessoas e 46 000 unidades domiciliares em 14 000 ED. A informação foi então transcrita dos questionários censitários, em relação a pessoas especificadas nas linhas de amostra de 20% e em relação às unidades domiciliares dos chefes da família que figuravam na amostra. Esta foi uma das primeiras operações executadas. Das informações transcritas, prepararam-se cartões de perfuração separados.

---

(1) - Bureau do Censo, 1950 Census of Population and Housing Preliminary Reports, Série PC-5, PC-6, PC-7, HC-3, HC-4 e HC-5, 1951.

Para as mais importantes cidades e áreas metropolitanas do país, empregou-se uma técnica diferente a fim de interferir o menos possível nas operações de processamento das respostas censitárias. Em cada um dêstes lugares selecionou-se uma amostra de ED à qual se deu alta prioridade em tôdas as operações censitárias. Então, logo que se tornaram disponíveis todos os cartões perfurados daqueles ED selecionados, fizeram-se as tabulações necessárias de acôrdo com os cartões referentes às pessoas da amostra de 20% da população e, geralmente, a tôdas as unidades domiciliares. Os cálculos preliminares levaram ao emprêgo de cêrca de 100 ED em cada área para a desejada fidedignidade das características gerais (1). Nas áreas metropolitanas que continham as maiores cidades, para as quais se publicaram estimativas separadas para a cidade central e para tôda a área, o número de ED na amostra aumentou. Nestes lugares, o número de ED necessários no resíduo, fora da cidade central, foi determinado de maneira a fornecer fidedignidade igual para proporções idênticas, na cidade central e na área tôda. Os dados publicados basearam-se em estimativas de razões.

Forneceu-se em 1950 uma série mais extensa de dados preliminares do que em 1940, mas para uma série mais limitada de áreas. Entretanto, a amostra preliminar em 1940 implicou a transcrição de dados para cêrca de 6 1/2 milhões de pessoas - trabalho consideràvelmente maior do que empreendemos em 1950. O tipo de amostragem em áreas metropolitanas padronizadas criou alguns problemas especiais. Classes da população que tendem a se concentrar, tais como a população institucional, os não brancos, estudantes universitários, pessoal militar, ocupantes de navios no mar, etc., criam problemas especiais quando se usam grandes conglomerados como elementos de amostra (1). As grandes instituições, instalações militares, e áreas portuárias constituiram ED separados e claramente identificados, e êstes eram estabelecidos como estrato separado. Foi pouco o que se pôde fazer a respeito de outros conglomerados. A Figura 4 apresenta a alta variabilidade de amostragem dêstes grupos. O ponto 34 representa a população institucional de 14 anos e mais. Os pontos 5, 14 e 36 são características da população não branca. O efeito sôbre os dados para os Estados Unidos, regiões e estados não foi grande demais, mas em algumas áreas metropolitanas foi sério. Resultou na impossibilidade de apresentarmos dados separados para algumas destas classes da população. Um problema especial foi o das pessoas de 18-24 anos que freqüentavam a escola. Para os não brancos, os dados só puderam ser apresentados para algumas das áreas metropolitanas padronizadas sulistas.

---

(1) Hansen, MH. e Hurwitz, W.N.: "Relative Efficiencies of Various Sampling Units in Population Inquiries", Journal of the American Statistical Association, Vol. 37, págs. 89-94, março de 1942.

É proveitoso comparar a eficiência dos planejamentos de amostra com algumas alternativas como as seguintes (1, 2):

1. Uma amostra simples ao acaso, do mesmo tamanho: Em Massachussetts, em que a taxa de amostragem dentro do ED foi 1 em 60, resultando em um tamanho médio de amostra de cêrca de 12 pessoas por ED, a variância de um planejamento simples ao acaso é cêrca de 88% daquela do planejamento realmente utilizado. As vantagens resultantes da estratificação quase compensaram completamente o decréscimo na eficiência devido ao uso de uma amostra de conglomerado. Em Cleveland, em que o coeficiente de amostragem dentro do ED foi 1 em 5, resultando em um tamanho médio de amostra de cêrca de 150 pessoas por ED, a variância de um planejamento simples ao acaso é sòmente cêrca de 39% daquela do planejamento real. Isto revela os resultados que se obtêm ao considerar apenas alguns conglomerados e uma taxa grande dentro dos conglomerados.

2. Um planejamento de conglomerado do mesmo tamanho que não prevê a estratificação urbana, rural e institucional: Em Massachussetts, a estratificação oferece uma redução de cêrca de 29% na variância.

3. Um planejamento de conglomerado do mesmo tamanho no qual os ED são selecionados ao acaso, ao invés de sistemàticamente: Em Cleveland, há uma redução de cêrca de 31% na variância mediante uma seleção de ED sistemática invés de ao acaso.

O Inquérito da Propriedade Residencial*- Como parte do Censo de Habitação, planejou-se uma série de perguntas sôbre os aspectos financiais das propriedades residenciais não agrícolas hipotecadas. No Censo de Habitação de 1940, fêz-se um número limitado de tais perguntas em relação às propriedades não comerciais ocupadas pelo proprietário, constituídas de uma a quatro unidades domiciliares. Para 1950, o âmbito das perguntas expandiu-se e obtiveram-se dados sôbre propriedades ocupadas por proprietários e locatários.

Em lugar de obter êstes dados para tôdas as propriedades residenciais, achamos preferível obtê-los na base de amostra subseqüente ao recenseamento normal. Além do custo relativamente alto de coleta dêstes dados, a experiência de 1940 sugeriu que os erros de declaração eram bastante elevados quando se coligiam os dados financiais simultâneamente com outros itens no Censo. Presume-se que isto se devia ao fato de que o informante habitual no Censo é a dona de casa, que freqüentemente não está bem informada sôbre o financiamento da

---

(1) Madow, W.G. e Madow L.H.: "On the Theory of Systematic Sampling" (ver nota (1) da pág. 1). The Annals of Mathematical Statistics, Vol. XV, págs. 1-24, março de 1944.

(2) Cochran, W.G.: "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Population". The Annals of Mathematical Statistics, Vol. XVIII, págs. 164-177, junho de 1946.

* Preparado com Nathan Lieder.

casa. Julgou-se que seria um processo mais conveniente obter êstes dados na base de amostra e através do proprietário. Achou-se que para as áreas de tabulação consideradas os erros de amostragem introduzidos seriam provàvelmente mais do que compensados pela redução nos erros de declaração. Os planos de publicação formulados após entendimento com um grupo consultivo exigiram tabulações para cada uma das quatro regiões e as grandes áreas metropolitanas padronizadas.

A amostra foi planejada a fim de fornecer estimativas das características financiais das propriedades residenciais com uma fidedignidade fixa para cada uma das quatro principais regiões do país e para 25 das maiores áreas metropolitanas padronizadas. Por exemplo, visávamos um coeficiente de variação de cêrca de 5% (no nível de um desvio-padrão) para um item que é 30% do total das propriedades hipotecadas ocupadas pelo proprietário. Uma condição adicional imposta sôbre a amostra era que devia fornecer boas estimativas do total da dívida hipotecária não saldada.

Uma propriedade residencial hipotecada é, em geral, um único prédio ou vários prédios principalmente dedicados ao uso residencial que tenham sido empenhado pelo proprietário como seguro por um empréstimo. Não existe uma fonte de que se possa obter uma relação de tôdas essas propriedades hipotecadas ou mesmo de tôdas as propriedades residenciais. Contudo, pôde-se selecionar uma amostra de prédios baseada numa amostra das unidades domiciliares não agrícolas recenseadas no Censo de Habitação de 1950. Êste foi escolhido, portanto, como método para selecionar a amostra de propriedades - a amostra a ser compreendida daquelas propriedades que incluíam os prédios selecionados. Depois que o questionário básico do Censo de 1950 foi preenchido e devolvido pelo recenseador, selecionou-se a amostra para o Inquérito da Propriedade Residencial.

Empregou-se uma amostra de duas fases. Com base nas estimativas do custo de várias fases do inquérito e em estimativas das variâncias dentro da unidade primária de amostragem e entre as unidades primárias de amostragem, determinou-se o número ótimo de unidades primárias de amostragem a ser selecionado e o tamanho esperado da amostra dentro de cada uma. É interessante notar que as estimativas pesquisadas com antecedência, tanto para o custo direto da repartição local como para o custo total da coleta, de viagens e de processamento por questionário, foram aproximadamente o dôbro do custo real. Visto como a distribuição ótima é uma função da proporção dos dois custos, as elevadas estimativas antecipadas do custo ainda levam a distribuições da amostra próximas da ótima.

As unidades primárias de amostragem formaram-se, consistindo, em geral, de um condado ou vários condados contíguos e foram grupadas em estratos de tamanho quase igual dentro de uma região, na base de homogeneidade de características de habitação. A Figura 5 apresenta os critérios usados na estratifi

cação. Os dados especificamente relacionados às propriedades ocupadas por inquilinos não estavam disponíveis e não podiam, assim, ser incluídos entre os critérios. Selecionou-se uma unidade primária de amostragem de cada estrato com probabilidade proporcional ao tamanho da amostra (1). A medida de tamanho foi a população não agrícola estimada em 1950.

Dentro de cada uma das unidades primárias de amostragem selecionadas, selecionou-se uma amostra de prédios não agrícolas baseada em dados coligidos no Censo de Habitação de 1950. Antes da seleção final, os prédios foram estratificados no Censo de 1950 como hipotecados ou não hipotecados, ou se eram ocupados por inquilino. Os prédios ocupados por inquilinos foram ainda estratificados por número de unidades domiciliares. Estabeleceu-se uma amostra auto-ponderada, para cada classe de prédio, ajustando o coeficiente dentro da unidade primária de amostragem de maneira que as probabilidades de seleção resultantes eram as mesmas para todos os prédios de uma determinada propriedade dentro de cada região. O processo de amostragem é apresentado com pormenores na Figura 5. Os questionários censitários foram planejados com o fim de coligir dados sôbre situação quanto a hipoteca, sòmente para as unidades domiciliares ocupadas por proprietário. Um estudo pilôto anterior ao censo revelou que aproximadamente 8% daquelas unidades recenseadas como não hipotecadas estavam realmente hipotecadas. Por conseguinte, a amostra final foi selecionada das que foram registradas como não hipotecadas, bem como das que foram registradas como hipotecadas.

O mesmo estudo pilôto e estimativas de outras fontes indicaram que aproximadamente um têrço de todos os prédios ocupados por inquilino estavam hipotecados. Esta estimativa foi utilizada a fim de determinar o tamanho da amostra dos inquilinos para fornecer a fidedignidade requerida.

A fim de fornecer estimativas adequadas da dívida total não saldada, as grandes propriedades de aluguel foram consideradas como um estrato separado. Estas propriedades têm um impacto muito grande sôbre a variância de estimativas de agregados, embora o número destas propriedades de aluguel seja pequeno comparado ao total de tôdas as outras propriedades de aluguel hipotecadas. Selecionou-se uma amostra sistemática de uma lista de tais propriedades (ver Figura 5). A fim de evitar tendenciosidade na amostra, estabeleceu-se um processo para assegurar que uma propriedade nesta lista não teve oportunidade de seleção na parte da amostra selecionada do Censo de Habitação. As grandes propriedades foram selecionadas por coeficientes de uma em quatro, uma em duas, ou com certeza, dependendo do tamanho.

---

(1) Hansen, M.H. e Hurwitz, W.N.: "On the Theory of Sampling From Finite Populations", The Annals of Mathematical Statistics, Vol. XIV, pags. 333-362, dezembro de 1943.

A coleta por via postal, seguido por uma coleta de campo, de uma amostra de não declarantes após vários inquéritos adicionais por via postal, foi adotada como a técnica de inquérito mais eficiente. Algumas considerações levaram a esta decisão. Achou-se que usando a coleta por via postal, a qualidade das respostas seria provàvelmente bastante alta, visto como o dono da propriedade poderia consultar seus registros de acôrdo com suas conveniências. Outra consideração foram as despesas de viagem relativamente elevadas envolvidas na coleta direta de campo. Em particular, soube-se que uma proporção substancial das propriedades da amostra seria não hipotecada. Os processos por via postal eliminariam a necessidade de viagens dispendiosas a muitas destas propriedades. De modo semelhante, as informações sôbre as características financiais de propriedades de aluguel deviam ser obtidas dos proprietários das mesmas. O Censo não fornecia informação sôbre a localização do proprietário e um processo por via postal era um meio econômico de obter seu nome e endereço. Além disso, uma prova experimental revelou que se podia esperar uma taxa bastante alta de respostas a um "canvass" por via postal.

O número de questionários a serem enviados pelo correio e o tamanho da sub-amostra foram então calculados de modo a fornecer a fidedignidade predeterminada ao menor custo (1). Após o inquérito foi possível determinar que as estimativas, anteriores ao inquérito, do custo da preparação burocrática, estavam apreciàvelmente abaixo do custo real. Além disso, os custos de coleta e viagem por questionário para o recenseamento de campo não eram tão altos quanto se imaginava. Se os custos reais fôssem conhecidos, ou mais estreitamente aproximados, teria sido possível começar com um tamanho menor de amostra e depois tentado obter questionários pela coleta de campo de todos os não declarantes à coleta por via postal. Felizmente, entretanto, havia esta amostra maior, visto como a amostra completa para as propriedades de aluguel era necessária nas amostras das áreas metropolitanas padronizadas, como resultado de super-estimação antecipada do número total de propriedades de aluguel hipotecadas nestas áreas.

Finalmente, obtiveram-se questionários de cada credor hipotecário bem como do devedor hipotecário. O questionário do credor foi usado para obter alguns tipos de informação para os quais não se podiam obter dados do devedor (tipo de possuidor da hipoteca, por exemplo). Foi possível, ainda, verificar as respostas de um devedor a algumas perguntas feitas ao credor sôbre as características da hipoteca. Não teria sido suficiente, entretanto, obter a informação do credor apenas, porque alguns dos dados básicos da propriedade ne-

---

(1) Hansen, M.H. e Hurwitz, W.N.: "The Problem of Non-Response in Sample Surveys", Journal of the American Statistical Association, Vol. 41, págs. 517-529, dezembro de 1946.

cessários só são possíveis através do devedor.

Os dados do inquérito ainda não estão publicados. Em geral, as estimativas da proporção de hipotecas, anteriores à amostra, parecem justificadas. Contudo, em algumas das grandes áreas metropolitanas padronizadas, não se obteve o número desejado de propriedades de aluguel hipotecadas. Isto se explica, provàvelmente, primeiro por uma estimativa antecipada inexata do número total de prédios de aluguel nestas áreas, e segundo, por uma proporção significativa de propriedades que revelaram ser ocupadas pelo proprietário, embora figurassem na amostra como ocupadas por inquilino. A diferença entre a classificação de uma propriedade usando os dados do Censo e os resultados de um inquérito pode ser explicada, em pequena parte, por uma mudança na condição de propriedade entre o período do Censo e a época em que os questionários do inquérito foram coligidos. Entretanto, os erros do recenseamento são responsáveis, provàvelmente, pela maior parte da diferença.

A fim de estimar a extensão dos erros de resposta semelhantes, no próprio inquérito, selecionou-se uma sub-amostra de 2% de tôdas as propriedades do inquérito cujos proprietários declararam não estar hipotecadas. Os registros legais destas propriedades foram verificados em escritórios locais de condados, tribunais, etc. Como resultado, estima-se que menos de 1% das propriedades registradas neste inquérito como não hipotecadas eram hipotecadas, em realidade.

De um modo geral, obteve-se uma taxa muito elevada de respostas aos questionários por via postal. A Figura 6 dá uma análise de taxas de respostas ao recenseamento por via postal. A grande proporção de propriedades de aluguel com proprietários não identificados durante a fase postal do inquérito - 30% de tôdas as propriedades de aluguel da amostra - foi causada pela falta de tempo suficiente para obter êstes endereços, por endereços falhos e insuficientes obtidos dos questionários censitários, propriedades vazias para as quais não se puderam obter informações pelo correio, e por informações incorretas fornecidas pelos inquilinos. Ocorreram endereços insuficientes de propriedades, no censo, especialmente nas zonas rurais. Os recenseadores não relacionaram os endereços para correspondência, em tais áreas. Descreviam, tanto quanto podiam, a localização dos prédios. O nome do ocupante e o distrito era bastante, muitas vêzes, para localizar o prédio no inquérito subseqüente. Freqüentemente, porém, isto não adiantava. O problema era particularmente sério em relação aos prédios vazios, especialmente em áreas de veraneio, etc., em que uma descrição idêntica podia-se aplicar à maioria dos prédios em um conglomerado. O processo por via postal não funcionou eficientemente em relação a tais propriedades.

Além dos problemas postais da distribuição de questionários pa-

ra as pessoas apropriadas, os recenseadores tinham dificuldade em localizar alguns dêstes prédios. Nestes casos, o recenseador devia fazer um levantamento dos prédios residenciais na área geral e selecionar um prédio em uma forma predeterminada, livre de tendenciosidade. Pode ser conveniente investigar totalmente a praticabilidade de se obterem endereços postais nos questionários censitários, no futuro, caso se considerem inquéritos subseqüentes por amostra baseados nos endereços censitários.

Um dos aspectos singulares dêste inquérito foi a verificação das respostas feita como parte do próprio inquérito. Quando se compararam os questionários dos devedores e credores, para a mesma propriedade, os erros de declaração se revelaram. Tìpicamente, quando existiam diferenças, os dois questionários pareciam de diferentes propriedades - não concordavam sôbre os têrmos do empréstimo, nem sôbre o número de imóveis abrangidos pela hipoteca. Êstes problemas foram resolvidos em parte por via postal, posteriormente.

As estimativas da variabilidade da amostragem ainda não foram computadas no inquérito total. Entretanto, as variâncias calculadas empregando questionários para as propriedades adquiridas em 1949 e na primeira metade de 1950, indicam que foi atingida provàvelmente, de um modo geral, uma fidedignidade ligeiramente melhor do que se visava.

Tendenciosidade e variâncias * - Conforme se indicou anteriormente, tomaram-se medidas para determinar e controlar tendenciosidades. As tendenciosidades nos dados baseados em uma amostra podem surgir de três fontes pelo menos (1) : a) o planejamento da amostra pode estar tendencioso; b) a situação de entrevista pode ser uma fonte de tendenciosidade; e c) as técnicas de processamento utilizadas para tabular os dados pode resultar em dados tendenciosos. A segunda e a terceira fontes de tendenciosidade não são problemas apenas dos inquéritos por amostra, visto como ocorrem também nos censos completos. Considerando o censo em conjunto, a qualidade das operações de processamento foi controlada por meio dos programas de verificação de amostra descritos na secção seguinte. Estabeleceram-se alguns contrôles sôbre as tendenciosidades surgidas durante a coleta, através de uma supervisão de campo vigorosa. Isto não eliminou erros de declaração e organizou-se um Inquérito Pós-Censitário para estudar, na base de amostra, a natureza dos erros de declaração. Êste inquérito será examinado ligeiramente, mais adiante, neste documento.

A coleta simultânea à base de amostra, em conjunto com o censo, criou a possibilidade de tendenciosidades na seleção da amostra, particularmente quando a maior parte do trabalho foi conduzida por pessoal descentralizado .

---

* Preparado com Herman Hèss

(1) Deming, W.E.: "On Errors in Surveys", American Sociological Review, Vol. IX, pags. 359-369, agôsto de 1944.

A principal preocupação foi causada pelas amostras de 20% da população e da habitação. Isto surgiu parcialmente como resultado da importância destas amostras nos planos de tabulação. Preocupou, além disso, a possibilidade de que aparecesse uma tendenciosidade, porque estas amostras eram geralmente as bases de outros programas de amostra, alguns dos quais foram anteriormente examinados. Como garantia, deu-se grande ênfase ao exame da possibilidade de tendenciosidades na seleção destas amostras.

As instruções ao recenseador previam uma ordem específica de recenseamento de pessoas dentro da família censitária. As famílias censitárias deviam ser recenseadas sucessivamente sem deixar linhas em branco, exceto sob condições aceitáveis. Quando se seguiram êstes processos a amostra não sofreu tendenciosidade. Exceto por uma manipulação muito hábil da ordem de contagem das famílias, os processos permitiram a verificação da seleção da amostra. Estabeleceram-se três séries de contrôle. A primeira abrangeu a inspeção do trabalho do recenseador pelo seu supervisor à medida que o trabalho prosseguia, e a sua correção sempre que se encontravam problemas sérios. Além disso, à medida que o trabalho de coleta era completado em cada condado, o Supervisor de Distrito da área devia apresentar um relatório contendo a população total do condado e o número de pessoas para as quais se obtiveram informação por amostra. Um exame destas duas cifras serviu para indicar se algum recenseador deixou de recensear um número significativo de pessoas nas linhas de amostra. Descobriu-se que apenas um punhado dos 130 000 recenseadores cometeram erros sérios o bastante para afetar, em grau significativo, o tamanho da amostra de 20% da população. Uma fonte maior de dificuldade que esta verificação revelou foi o fato de que alguns recenseadores deixaram de completar tôdas as informações suplementares que deviam ser indagadas às pessoas da amostra, particularmente no caso da população institucional. Cêrca de 7 000 cartas foram enviadas de Washington a fim de permitir preencher as informações dêste tipo quando um grupo de pessoas para as quais faltavam informações da amostra se concentrava em uma área. Receberam-se respostas a mais da metade destas cartas.

Fêz-se uma verificação mais detalhada na época da tabulação, a fim de assegurar que não houvessem tendenciosidades significativas para algumas características selecionadas. Na amostra da população êste contrôle foi estabelecido no nível da área de publicação. Pràticamente não se encontraram erros bastante grandes que merecessem correção. É interessante notar, entretanto, que, como conseqüência desta verificação, descobriram-se problemas metodológicos no programa de tabulação, que foram então resolvidos. Informações adicionais sôbre a possibilidade de tendenciosidades foram fornecidas por um estudo em pequena escala baseado em uma amostra de distritos censitários.

Visto como nos questionários censitários, mesmo quando as instru

ções foram seguidas, havia linhas que não continham declarações ou observações do recenseador, o processo empregado não assegurou automàticamente uma amostra exata de 20% de pessoas ou unidades domiciliares em cada localidade. Outros desvios podiam ocorrer quando as instruções não eram seguidas. Entretanto, a análise do tamanho da amostra da população revelou que para a população total, os desvios de 20% são insignificantes, em muitas áreas. Espera-se que uma situação semelhante se apresente na amostra da habitação. Em relação a determinadas características, os desvios são um pouco maiores, mas não o bastante para causar qualquer preocupação. As Figuras 7 e 8 apresentam alguns dados sôbre o tamanho da amostra de 20% da população. Pode-se verificar nos níveis de condados e estados que há evidência de algumas tendenciosidades muito pequenas não descobertas no nível do ED na amostra de 20% da população, devido ao fato de os recenseadores terem deixado de seguir as instruções exatamente. A amostra da população total é muito ligeiramente, mas significativamente inferior a 20% e a amostra dos chefes de família, um pouco mais inferior. Na amostra de 20% da habitação, não há evidência de qualquer tendenciosidade, no momento. Entretanto, a investigação da amostra de habitação ainda não foi completada.

Na amostra de 3 1/3%, além dos problemas inerentes aos métodos de coleta e processamento, houve uma complicação adicional causada por um método de seleção inicial de amostra ligeiramente tendencioso. A amostra de 3 1/3% consistiu de pessoas que apareciam na última linha de amostra de cada questionário de população. Os processos de coleta determinavam que sempre que uma série de 12 linhas de habitação num questionário eram totalmente completadas, os recenseadores não deviam completar a parte de população do questionário, mas sim começar uma nova fôlha. O efeito disto foi uma diminuição na representação das pessoas que viviam em pequenas famílias censitárias, na amostra de 3 1/3%. Entretanto, a necessidade de prover um esquema simples de amostragem que os recenseadores pudessem compreender fàcilmente e seguir mecânicamente, parecia indicar que êste era o melhor plano a seguir.

A análise de uma amostra dos questionários completos indicou que na amostra, conforme foi selecionada inicialmente, por causa do planejamento do questionário, a proporção das pessoas não incluídas, a fim de obter uma amostra de 3 1/3% livre de tendenciosidade, atingiu aproximadamente 4%. A fim de corrigir a tendenciosidade no método de seleção inicial, como parte do processo de crítica, tôdas as fôlhas em que as declarações sôbre habitação foram completadas antes de se chegar à última linha de população foram consideradas como constituindo um estrato separado e uma amostra de 3 1/3% de pessoas foi selecionada de maneira não tendenciosa dêste estrato. Contudo, as respostas são "ignorado" ou preenchidas na fase da crítica para os poucos itens de 3 1/3% para tais pessoas. A amostra de 3 1/3% de pessoas compreende os dois grupos.

Nenhuma análise da qualidade do trabalho no censo seria completa sem uma menção do Inquérito Pós-Censitário e do Confronto do Inquérito Corrente de População (CPS) com o Censo. A fim de investigar a exatidão do recenseamento, de acôrdo com a definição usada, fêz-se um inquérito pós-censitário como suplemento do censo. Um grupo de recenseadores altamente treinados reexaminavam, na base de amostra, informações obtidas pelos recenseadores usuais. Houve duas grandes categorias de informação que êste inquérito procurou verificar: contagens diminuídas ou aumentadas das pessoas e unidades domiciliares, denominadas "verificação do campo abrangido"; e exatidão das informações obtidas, denominadas "verificação do conteúdo".

O inquérito foi uma combinação de amostra à base de áreas e amostra por listagem - a primeira, para verificar o campo abrangido, a segunda para verificar o conteúdo. Ambas estas amostras foram realizadas em uma amostra de cêrca de 280 unidades primárias de amostragem.

Os questionários empregados no PES (Inquérito Pós-Censitário) foram bastante detalhados. Destinavam-se a descobrir as pessoas e as unidades domiciliares duplicadas e omitidas, caracterizar as famílias cujos dados se confundiam, e fornecer comparações de respostas às perguntas censitárias, não só com os resultados do recenseamento de 1950 mas também com outras fontes independentes, tais como registros de nascimentos, imigração, previdência social e matrícula escolar. Quando estas tabulações fôrem completadas e publicadas fornecerão ao consumidor de material censitário uma base para avaliar a exatidão dos resultados, e auxiliarão o Bureau do Censo no planejamento de futuros questionários.

O Bureau do Censo realiza o Inquérito Corrente de População (CPS), que fornece a base para os relatórios mensais sôbre a mão-de-obra, assim como vários outros relatórios suplementares publicados em intervalos periódicos. Previu-se que surgiriam diferenças entre os resultados dêste inquérito e do Censo, à parte daquelas devidas à variabilidade da amostragem, surgindo tais diferenças de técnicas diferentes de recenseamento, programas de treinamento e planejamento de questionários. O Bureau achou que, como realizador tanto do Censo como da estatística corrente, tinha a responsabilidade de explicar as razões das diferenças. As respostas das famílias no CPS em abril de 1950 estão sendo comparadas no CPS e no Censo. Também está sendo feita uma comparação do campo abrangido. Os resultados, além de explicar as diferenças, também devem ajudar a esclarecer os conceitos dos vários itens, conforme foram realmente aplicados pelos recenseadores, considerados diferentes dos projetados pelo planejamento.

É praxe do Bureau do Censo fornecer aos consumidores de dados censitários informações sôbre êrros de amostragem de dados publicados com base em amostras. No caso de amostragem simples ao acaso, isto não apresenta problema. Dada a população de uma área e o coeficiente de amostragem, o êrro-padrão de qualquer estimativa é uma simples função do tamanho da estimativa e os erros-pa-

drão podem ser expressados na forma de um simples quadro ou podem ser computados fàcilmente pelos consumidores dos dados. Entretanto, quando se usam amostras por etapas múltiplas ou sistemáticas (como no caso de tôdas as amostras censitárias), tais expressões simples não podem ser usadas e o êrro-padrão só pode ser formulado em relação ao planejamento da amostra específica empregada. O êrro-padrão de cada estimativa depende da distribuição na população da característica cujo tamanho está sendo estimado e uma expressão exata da precisão requereria um cômputo separado do êrro-padrão para cada característica estimada. Evidentemente, isto não é prático. Não só o custo é proibitivo, mas a expressão dos erros-padrão requereria tantos volumes quanto os de resultados.

O processo seguido geralmente é o de computar as variâncias de um número limitado de características típicas. Constrói-se então um diagrama de dispersão traçando-se estas variâncias contra o tamanho das estimativas. (Na prática, usam-se os coeficientes de variação ao quadrado ao invés de variâncias). Quando êstes diagramas de dispersão são examinados verifica-se geralmente que os pontos tendem a cair em curvas mais ou menos bem definidas da forma $V_x^2 = a/\frac{b}{x}$. As curvas são então ajustadas aos pontos por métodos de regressão reduzindo ao mínimo os resíduos relativos ao quadrado do V_x^2, supondo-se então que as curvas representarão a precisão da classe inteira de estimativas baseadas na amostra. Podem-se então construir quadros simples apresentando o êrro-padrão como uma função do tamanho da estimativa, independentemente das características particulares daquele tamanho. Quando se usam curvas diferentes, elas geralmente representam grupos diferentes de características, por exemplo, estimativas de brancos versus não brancos, características de indivíduos ou famílias censitárias, etc. Uma série típica de pontos e uma curva são apresentadas na Figura 4.

Tanto as amostras de população como de habitação de 20% empregadas são bàsicamente amostras sistemáticas, embora as diferentes versões dos questionários e a existência de linhas em branco tivessem o efeito de introduzir variações periódicas no molde sistemático. Para estas amostras, calcularam-se variâncias das características apresentadas na Figura 9. As variâncias das amostras de 20% das características obtidas na base de 100% foram estimadas selecionando uma amostra de distritos censitários e calculando a variância de população das cinco amostras possíveis de 20% em cada um dos distritos censitários selecionados (Figura 10). As diferenças seriais das observações relativas às pessoas nas linhas da amostra de 20% foram usadas para investigar as variâncias de algumas das mesmas características da base de 100% e de uma série de características por amostra no mesmo grupo de distritos censitários. Além do mais, a amostra básica de distritos censitários foi subdividida em quatro grupos independentes e os cálculos foram realizados em separado para cada um dêstes grupos a fim de investigar a estabilidade das estimativas da variância.

Estas técnicas resultaram virtualmente nas mesmas estimativas da

variabilidade da amostragem, para a maioria dos itens, e as variâncias assim obtidas, comparadas intimamente com aquelas que teriam resultado de uma amostra simples ao acaso de unidades individuais (as binomiais), descontando-se as linhas em branco. Os dados para a amostra de população são apresentados nas Figuras 9 e 11. O fato de que as estimativas da variância não dependem dos distritos censitários específicos empregados para estimar as variâncias, pode ser verificado na Figura 11, que revela que as quatro amostras independentes resultaram virtualmente nas mesmas estimativas de variância. A Figura 9 mostra como as variâncias estimadas pelos dois diferentes métodos são próximas das variâncias que surgiriam da suposição de uma amostra simples ao acaso.

A despeito do fato de que ambas as amostras de 20% da população e da habitação se aproximassem bastante do binômio, quando se combinavam as duas a fim de fornecer a amostra de 4% da habitação (as perguntas de habitação da amostra para as unidades domiciliares cujos chefes caiam nas linhas de amostra) resultavam erros-padrão relativamente elevados em comparação com a amostra de 20%. Isto é causado por um modêlo fortemente cíclico que resulta da combinação das duas amostras. As estimativas de razões serão usadas para diminuir as variâncias dos dados publicados.

São necessárias algumas modificações nos métodos de estimação de variância, no caso das amostras usadas para as tabulações preliminares por amostra, o Inquérito Pós-Censitário, o Inquérito da Propriedade Residencial, e outros planejamentos por etapas múltiplas. No caso da amostra preliminar, as variâncias para as estimativas estaduais e regionais foram obtidas na hipótese de que os distritos censitários foram selecionados ao acaso dentro de cada estrato. Quando a taxa de amostragem é muito baixa, uma seleção sistemática para as características tratadas tem pràticamente o mesmo efeito que a seleção ao acaso. Na estimação das variâncias na amostra para as grandes SMA(áreas metropolitanas padronizadas) e cidades, os coeficientes de amostragem dos distritos censitários, em algumas áreas, eram um em seis, e o efeito da seleção sistemática não podia ser ignorado. Foram então usados os quadrados das diferenças sucessivas dos totais dos distritos censitários. Deve-se notar que êste método exagera ligeiramente a variância. Planejam-se processos semelhantes para o Inquérito da Propriedade Residencial e o Inquérito Pós-Censitário utilizando a técnica de grupos de estratos agregados com algumas modificações trazidas pelas variações nos planejamentos de amostra.

Para as variâncias das estimativas de razões e proporções, usou-se a expressão $V_x^2 - V_y^2$, sendo V_x^2 o quadrado do coeficiente de variação do numerador e V_y^2 do denominador. Esta é uma boa aproximação, se (como é freqüentemente o caso de proporções de tamanho pequeno ou moderado) $\frac{x}{y}$ e y não fôrem altamente correlacionados.

Fizeram-se algumas aproximações adicionais a fim de simplificar os

cálculos. As estimativas reais publicadas para uma única região representaram uma soma de estimativas de razões para estados e o resíduo da região. Contudo, a hipótese levantada no cálculo da variância, foi que a estimativa de razão foi introduzida como passo final e baseou-se na soma de estratos de estimativas sem tendenciosidades. A investigação empírica revelou que o êrro introduzido por esta aproximação era insignificante. Por exemplo, na região Centro-Norte, a proporção entre a variância do método final e a variância do outro para a característica "assalariados e pessoas que recebem ordenado em trabalho privado" foi 1,03.

Inspeção do processamento por amostra* - Os métodos de amostragem, além de representarem um papel importante na coleta de informações suplementares e na publicação preliminar dos dados censitários, têm sido extremamente úteis na avaliação e contrôle da qualidade das operações de coleta e processamento. O contrôle da qualidade foi introduzido pela primeira vez nas operações censitárias de processamento em 1940 (1); no Censo de 1950, os planos de inspeção por amostragem foram usados em quase tôdas as operações básicas e suplementares de processamento.

A coleta e processamento dos dados censitários oferecem oportunidades únicas para aplicação de técnicas de contrôle de qualidade a operações não industriais. Dois fatôres correlacionados justificam o uso dos contrôles por amostra. O primeiro é a impossibilidade prática de conseguir exatidão de 100% nas operações de coleta, crítica, codificação e perfuração, de tal magnitude. Mesmo que se suponha que várias revisões completas de cada operação possam eliminar todos os erros bem nítidos, sempre há características tão difíceis de classificar que mesmo dois técnicos especializados não concordarão sôbre o código ou anotação apropriada do recenseador. O segundo fator é o custo. Cada verificação de 100% de uma operação aumenta o custo original de 50 a 100%. As necessidades dos consumidores de dados censitários não são tão rigorosas que um pequeno número de êrros não possam ser tolerados (2). A exatidão aumentada resultante de mesmo uma única inspeção de 100% de cada operação comparada à inspeção por amostra não aumentaria o valor dos dados para os consumidores o bastante para justificar o custo aumentado. O objeto de inspeção e verificação não é, pois, exatidão de 100% - é medir a qualidade do trabalho e conservá-lo em um nível aceitável. Não é necessário verificação de 100% para conseguir êstes objetivos.

Cada uma das principais operações de processamento estava sujeita

---

*Preparado com Thomas Jabine.

(1) Deming, W.E. e Geoffrey, L.: "On Sample Inspection in the Processing of Census Returns", Journal of the American Statistical Association, Vol. 36, págs. 351-360, setembro de 1941.

(2) Truesdell, L.E.: "The Problem of Quality in Census Data, "Estadistica", Vol. IX, págs. 163-171, junho de 1951.

a inspeção por amostra em algum momento. Na Figura 12 resumem-se os principais planos de inspeção utilizados. Esta figura ilustra a variedade de situações em que foram aplicadas as técnicas de inspeção por amostra e revela como estas técnicas variaram de acôrdo com as características particulares de cada operação. As figuras 13 e 14 descrevem com mais detalhes dois dos planos empregados. A Figura 13 descreve a verificação por amostra da codificação dos questionários de população. Êste plano foi projetado principalmente para controlar a exatidão de cada codificador, embora a amostragem para aceitação no nível do ED (distrito censitário) também estivesse compreendida. Serviu como modêlo para todos os planos que foram adotados subseqüentemente para outras operações de codificação e transcrição. A Figura 14 descreve a inspeção por amostra de contagens da população fora da sede. Estas contagens foram feitas nos escritórios de distrito e foram apurações por ED do número de pessoas recenseadas. Foram planejadas principalmente para fins de fôlha de pagamento e resultados preliminares. Com base em experiências passadas, achou-se que estas contagens eram suficientemente exatas para servir como base para contagens finais da população por estados, eliminando assim a necessidade de uma operação adicional de contagem. A fim de determinar se era êste o caso, foi planejado um plano de amostragem para aceitação, para determinar se as contagens fora do órgão central eram exatas, dentro do limite desejado nos níveis estaduais e nacionais.

Nestas operações foram seguidos princípios de particular importância no contrôle de processos dêste tipo. Alguns dêstes princípios podem ser óbvios a qualquer um que fêz trabalho de contrôle de qualidade; outros são talvez peculiares às operações de escritório dêste tipo.

1. Os contrôles primários devem ser sôbre o trabalho de indivíduos, visto como a diferença entre indivíduos é um componente principal da variação na qualidade. Êste princípio tinha que ser violado ocasionalmente, quando o coeficiente de produção em uma operação era tão baixo que o trabalho de um indivíduo por um período de uma ou duas semanas, selecionado por amostra em um nível razoável era insuficiente para fornecer uma estimativa fidedigna de qualidade.

2. A qualidade do trabalho de um indivíduo pode variar significativamente com o tempo. Portanto, é desejável computar coeficientes individuais de êrro pelo menos de duas em duas semanas e preferìvelmente tôda semana. Além disso, cada pessoa deve submeter-se a um período de qualificação (habilitação), no início da operação, durante o qual seu trabalho é verificado em 100%.

3. Devem-se fazer todos os esforços para conservar as características operacionais tão simples quanto possível. O trabalho a ser verificado deve ser designado por uma pessoa em cada unidade de trabalho em lugar de tornar cada verificador responsável pela aplicação das instruções de amostragem. Em geral, um modêlo sistemático de seleção com inícios ao acaso é preferível a um aleatório. As medidas de qualidade devem ser adaptáveis a sistemas de manutenção de registros.

Na codificação, por exemplo, o coeficiente de êrro foi obtido dividindo-se os erros totais de todos os tipos por população codificada. Embora altamente correlacionado com os efeitos de erros de codificação nas tabulações finais, um coeficiente de êrro dêste tipo não fornece uma medida exata dos desvios que ocorrerão em qualquer lote específico. Uma cifra exata requereria contrôles separados não só sôbre cada característica codificada, mas também sôbre todos os códigos possíveis para cada característica. Desde que isto, evidentemente, estava fora de consideração, era necessário restringir-se a estudos ocasionais para determinar a relação entre a qualidade resultante da codificação e a qualidade dos resultados finais.

4. A amostragem de conglomerados foi usada, até certo ponto, em todos os planos. A vantegem, em têrmos de produção, especialmente quando o conglomerado era um ED, tornou possível conseguir a fidedignidade desejada, mais econômicamente, mesmo que fôsse requerida uma amostra um pouco maior. É difícil determinar de antemão exatamente até que ponto a acumulação de erros afetará as características de operação de um determinado plano. Portanto, é conveniente examinar os dados operacionais cuidadosamente para determinar as verdadeiras características de operação de cada plano.

5. É essencial fazer algum tipo de verificação do trabalho dos verificadores. Nunca se operou sob a ilusão de que verificação de 100% significava exatidão de 100%, mas foi um tanto chocante descobrir, instituindo uma verificação sistemática do trabalho do verificador de codificação, que os verificadores estavam comunicando, em média, apenas cêrca de metade dos erros introduzidos. Achou-se que o único método eficaz de avaliar a qualidade do trabalho do verificador nas operações de codificação e crítica era pré-verificar, isto é, relacionar mas não corrigir os erros do codificador antes da verificação normal e então, depois que a verificação tivesse tido lugar, comparar esta lista com as comunicações de erros do verificador. Se a unidade de trabalho a ser verificada não contivesse erros de codificação, "implantávam-se" alguns a fim de reduzir o número de unidades de trabalho manuseadas. A experiência ensinou que a reverificação, isto é, a verificação após o trabalho normal do verificador, parece levar a enganos, visto como infelizmente não é incomum os verificadores corrigirem os erros sem comunicá-los. Nas operações tais como a perfuração, em que a verificação é uma operação independente, não é necessária a pré-verificação.

6. A maioria dos planos de contrôle foram operados a um custo de 20 a 40% do custo da inspeção de 100%. Um nível de inspeção de x% não assegura automàticamente uma economia de custo de 100-x%. As estimativas de custo devem levar em conta inspeção de 100% durante o período de treinamento, custos fixos do tratamento de lotes de inspeção, inspeção de 100% de lotes rejeitados e manutenção extra de registros.

Os resultados dêstes planos de inspeção por amostra em têrmos de qualidade e custo foram animadores. Entretanto, permanecem muitas oportunidades

para o aperfeiçoamento dos planos existentes e a instituição de contrôles adicionais por amostra. Em particular, é necessário um programa eficaz para controlar a qualidade da coleta. Muito se tem feito no sentido de avaliar a qualidade de um censo depois que o mesmo teve lugar, mas até agora não se fêz bastante para controlar a qualidade enquanto se realiza a coleta. Êste e outros problemas oferecerão um desafio contínuo àqueles que estão interessados em aplicar as técnicas de contrôle de qualidade ao trabalho de inquérito.

Agradecimento

Os autores sentem-se particularmente penhorados a Morris H. Hansen, Diretor Assistente para as Normas Estatísticas do Bureau do Censo e William N. Hurwitz, Chefe da Secção de Pesquisas Estatísticas, sob cuja direção técnica se realizou todo o trabalho descrito neste documento. Estas aplicações da amostragem no trabalho censitário foram incentivadas por A. Ross Eckler, Diretor Interino do Bureau do Censo, Howard G. Brunsman, Chefe, e Henry S. Shryock, Wayne F. Daugherty e Robert B. Voight, Chefes Assistentes da Divisão de População e Habitação.

A maioria dosquadros e gráficos mencionados neste artigo, embora de considerável interêsse, não são essenciais para a compreensão dos processos básicos descritos. Portanto, a fim de simplificar o trabalho e economizar tempo, as figuras não consideradas essenciais deixaram de ser traduzidas e reproduzidas. Sòmente as figuras 1, 6 e 13, que se julgou de especial importância, são incluídas no final do artigo. Para os que estejam interessados, a série completa de figuras poderá ser encontrada no texto original, em inglês.

FIGURA 1

Porções do Questionário de População do Censo de 1950 e tipos de informações obtidas através das amostras de 20% e 3 1/3%

(questionário de 30 linhas, havendo de 5 em 5 linhas uma linha de amostra)

	Número da linha	Nome da rua, avenida ou estrada	Número da casa (e apartamento)	NOME Qual o nome do chefe desta família? Quais os nomes de tôdas as outras pessoas que moram aqui?	CONDIÇÃO EM RELAÇÃO AO CHEFE DA FAMÍLIA Registre a condição da pessoa em relação ao chefe da família, como: Chefe Cônjuge Filha Neto Sogra Hóspede Cônjuge do hóspede Empregada, etc.
		1	2	7	8
	1				
LINHA DE AMOSTRA (2)					
	3				
	4				
	5				
	6				
LINHA DE AMOSTRA (7)					

Informação solicitada pela amostra de 20%

1. Residência em 1949
2. País de nascimento dos pais
3. Matrícula escolar e anos escolares completados
4. Renda individual e familiar em 1949
5. Serviço nas fôrças armadas

Informação solicitada pela amostra de 3 1/3%

1. Última ocupação e ramo da atividade econômica das pessoas que deixaram de trabalhar.
2. Informações sôbre casamento e fecundidade para as pessoas não solteiras.

FIGURA 1 (cont.)

PORÇÃO DO QUESTIONÁRIO DE HABITAÇÃO DO CENSO DE 1950

(Questionário de 12 linhas; modelos de perguntas de amostra repetidas em ciclos de 5 linhas)

12	13	14	15	16
BANHEIRO - USO EXCLUSIVO 1 ☐ Para uso exclusivo desta unidade 2 ☐ Compartilhado por outra unidade 3 ☐ Nenhum banheiro para uso desta unidade	BANHEIRA OU CHUVEIRO INSTALADO 1 ☐ Para uso exclusivo desta unidade 2 ☐ Compartilhado por outra unidade 3 ☐ Nenhuma banheira ou chuveiro para uso desta unidade	a. EQUIPAMENTO (1) DE AQUECIMENTO 1 ☐ Vapor encanado ou água quente 2 ☐ Fôrno de ar aquecido 3 ☐ Outros meios com chaminé 4 ☐ Outros meios sem chaminé 5 ☐ Não aquecida V ☐ Vazia	a. COMBUSTÍVEL DE AQUECIMENTO MAIS USADO 1 ☐ Carvão 2 ☐ Madeira 3 ☐ Gás de iluminação 4 ☐ Gás engarrafado 5 ☐ Combustível líquido 6 ☐ Eletricidade 7 ☐ Outros combustíveis 8 ☐ Não aquecida V ☐ Vazia	FORMA DE OCUPAÇÃO Ocupada 1 ☐ Por proprietário 2 ☐ Por locatário 3 ☐ Mediante cessão Vazia 4 ☐ Para alugar 5 ☐ Para venda, sòmente 6 ☐ Não está para alugar ou vender
BANHEIRO - USO EXCLUSIVO 1 ☐ Para uso exclusivo desta unidade 2 ☐ Compartilhado por outra unidade 3 ☐ Nenhum banheiro para uso desta unidade	BANHEIRA OU CHUVEIRO INSTALADO 1 ☐ Para uso exclusivo desta unidade 2 ☐ Compartilhado por outra unidade 3 ☐ Nenhuma banheira ou chuveiro para uso desta unidade	b. Esta unidade possui iluminação elétrica? (2) 1 ☐ Sim 2 ☐ Não V ☐ Vazia	b. Que tipo de refrigerador possui esta unidade? 1 ☐ Elétrico, a gás ou outro refrigerador mecânico 2 ☐ Geladeira com resfriamento a gêlo 3 ☐ Outro tipo de refrigeração 4 ☐ Nenhum V ☐ Vazia	FORMA DE OCUPAÇÃO Ocupada 1 ☐ Por proprietário 2 ☐ Por locatário 3 ☐ Mediante cessão Vazia 4 ☐ Para alugar 5 ☐ Para venda, sòmente 6 ☐ Não está para alugar ou vender

FIGURA 6

Análise das respostas por via postal no Inquérito da Propriedade Residencial

CLASSE DE QUESTIONÁRIOS	TÍTULO DE POSSE E SITUAÇÃO QUANTO A HIPOTECA, DECLARADOS NO CENSO			
	Proprietário - Propriedade hipotecada	Proprietário - Propriedade não hipotecada	Locatário (1)	Total
TOTAL	62 384	34 547	157 792(2)	254 723
Número enviado pelo menos uma vez	62 129	34 475	104 807(3)	201 411
Número recebido durante a fase postal	51 086	28 831	79 983	159 900
Questionários aceitáveis de unidades não hipotecadas	4 105	24 811	52 988	81 904
Questionários aceitáveis de unidades hipotecadas	36 409	2 201	14 842	53 452
Questionários incompletos ou inconsistentes	10 752	1 819	12 153	24 544
Número dos que não responderam	6 332	2 751	20 929	30 012
Número dos que voltaram sem chegar ao destinatário, devido a endereços falhos	4 966	2 965	5 996	13 927
Porcentagem recebida (do total enviado)	82,2	83,6	76,3	79,4
Porcentagem recebida (do número total)	81,9	83,5	50,7	62,8

(1) Inclui prédios selecionados dos questionários censitários e propriedades seleciona das da lista de grandes propriedades.

(2) Exclui cêrca de 5 000 propriedades eliminadas do inquérito porque eram de propriedade do govêrno, principalmente não residenciais, etc.

(3) Compreende aquêles registrados como proprietários de propriedades alugadas, em resposta a inquérito postal e telefônico dos locatários; não se pode enviar questionários para as propriedades cujos proprietários não estavam identificados na data do envio postal.

FIGURA 13

PROCESSOS DE OPERAÇÃO PARA VERIFICAÇÃO DE CODIFICAÇÃO POR AMOSTRA

A. Qualificação para verificação por amostra

OPERAÇÃO E SEMANA DE PRODUÇÃO	NO FIM DESTA SEMANA:	
	Qualifica-se para verificação por amostra se a taxa de êrro* fôr menos de	Dispensado da operação se a taxa de êrro fôr mais de
Codificação geral		
1	3%	**
2	5%	15%
3	5%	11%
4	5%	9%
5	5%	7%
6	5%	5%
Codificação de ocupação		
1	1%	**
2	2%	6%
3	2%	5%
4	2%	4%
5	2%	3%
6	2%	2%

* A taxa de êrro é obtida dividindo os erros totais pela população trabalhada.

** Nenhum codificador será dispensado antes do fim da segunda semana de produção.

B. Contrôle dos Operadores após a qualificação:

Um codificador será considerado "fora de contrôle" nas seguintes condições:

1. Codificação geral - Se a sua taxa de êrro (estimada da amostra) fôr mais de 7% em uma semana, ou mais de 6% para um período de duas semanas.

2. Codificação de ocupação - Se a sua taxa de êrro fôr mais de 3% em uma semana, ou mais de 2,5% para um período de duas semanas.

A "medida" a ser tomada quando um operador está fora de contrôle será geralmente verificação 100% para um pequeno período e retreinamento, seguido por requalificação ou dispensa da operação.

FIGURA 13 (cont.)

PROCESSOS DE OPERAÇÃO PARA VERIFICAÇÃO DE CODIFICAÇÃO POR AMOSTRA (cont.)

C. Padrões para aceitação dos E.D.

CODIFICAÇÃO GERAL		CODIFICAÇÃO DE OCUPAÇÃO	
Número de linhas verificadas no E.D.	Rejeitada se houver mais erros do que	Número de linhas verificadas no E.D.	Rejeitada se houver mais erros do que
1 - 9	1	1 - 19	1
10 - 19	2	20 - 39	2
20 - 29	3	40 - 59	3
.	.	.	.
.	.	.	.
.	.	.	.
190 - 199	20	180 - 199	10
.	.	.	.
.	.	.	.
.	.	.	.

D. Método de selecionar as linhas a serem verificadas

Codificação geral - Um conglomerado de 5 linhas de dois em dois questionários, 5 em 60 ou 8 1/3%.

Codificação de ocupação - Um questionário em cada 10, 30 em 300 ou 10%.

E. Limites superiores da qualidade média da produção esperada*

Codificação geral - 2,5%.

Codificação de ocupação - 1,0%.

Em têrmos de erros que seriam encontrados se a operação fôsse sujeita a verificação 100%.

F Custos do plano em têrmos do custo da codificação (Ver nota na pág. seguinte).

COMPONENTE	VERIFICAÇÃO DA CODIFICAÇÃO GERAL	VERIFICAÇÃO DA CODIFICAÇÃO DA OCUPAÇÃO
Verificação 100% durante o período de qualificação	6,7%	5,9%
Custo direto da verificação por amostra	7,5%	6,0%
Custo máximo para a verificação 100% de E.D. rejeitados e de operadores fora de contrôle	3,0%	3,0%
Registros, supervisão, retreinamento, etc.	2,8%	5,1%
Reserva	5,0%	5,0%
Custo total como porcentagem do custo da codificação	25,0%	25,0%

FIGURA 13 (concl.)

PROCESSOS DE OPERAÇÃO PARA VERIFICAÇÃO DE CODIFICAÇÃO POR AMOSTRA (concl.)

Nota - Os custos acima foram estimativas feitas antes que o plano fôsse pôsto em operação.

Os custos finais demonstraram ser um pouco mais altos - cêrca de 31% para a verificação da codificação geral e 34% para a verificação da codificação de ocupação. Os dois principais fatôres responsáveis pelos custos mais altos foram:

1. A mobilidade da mão-de-obra era muito maior do que se esperava originalmente, principalmente devido ao mercado de trabalho mais escasso que se seguiu ao irrompimento das hostilidades na Coréia. Isto aumentou grandemente a quantidade de verificação 100% necessária para qualificar novos codificadores.

2. O custo direto da verificação por amostra foi consideràvelmente mais alto do que havia sido previsto originalmente.

I.B.G.E. - Conselho Nacional de Estatística
NÚCLEO DE PLANEJAMENTO CENSITÁRIO

RELATÓRIO SÔBRE UMA EXPERIÊNCIA DESTINADA A ESTUDAR ALGUNS PROCESSOS DE AMOSTRAGEM, DOS QUESTIONÁRIOS USADOS NO CENSO DEMOGRÁFICO DE 1950

T. B. Jabine

a) Município de Vitória

- Julho 1956 -

CONTEÚDO DÊSTE RELATÓRIO

Em março de 1956, o Núcleo de Planejamento Censitário foi criado, para planejar os Censos Econômicos e Demográficos do Brasil em 1960. Uma das atividades do Núcleo consiste em determinar as técnicas de amostragem que podem ser usadas para reduzir o custo, melhorar a qualidade, e permitir a publicação dos resultados do Censo de 1960, com mais presteza.

O Núcleo acredita que essa é justamente uma das mais simples e, ao mesmo tempo, uma das mais valiosas aplicações da técnica de amostragem. Portanto a primeira etapa consistiria em tirar amostras experimentais dos questionários do Censo Demográfico de 1950, para determinadas áreas, reproduzindo, tanto quanto possível, as condições que teriam sido encontradas em 1950.

Os objetivos dessa experiência foram:

1 - Fazer tabulações baseadas nos dados das amostras as quais poderão ser comparadas com os resultados do Censo já publicados, ilustrando assim o grau de precisão a ser esperado dos vários tipos de amostras que poderiam ser usadas.

2 - Dar ao "Staff" do Núcleo uma idéia das modificações que poderiam ser introduzidas no processamento, tabulação e técnicas de publicação, necessárias quando usamos métodos modernos de amostragem.

3 - Obter informação sôbre custo e variabilidade, necessários para determinar o mais oficiente plano de amostragem a ser usado no futuro.

A primeira amostra foi selecionada dos questionários do Censo Demográfico de 1950 do município de Vitória, Estado do Espírito Santo. Posteriormente uma amostra separada foi selecionada do resto do Estado, e os resultados combinados, a fim de fazer estimativa para todo o Estado.

Os processos usados e os resultados obtidos para o município de Vitória, são apresentados neste relatório. Os resultados para o Estado do Espírito Santo serão apresentados em outro relatório.

PLANO DE AMOSTRAGEM

A população da qual foi tirada a amostra consiste de:

Todos os Boletins de Família (C.D. 1.01), Listas de Domicílio Coletivo (C.D. 1.02) e Boletins Individuais (C.D. 1.03) referentes ao município de Vitória. Êsses boletins e listas estavam contidos em 44 pastas e foram numerados dentro de cada pasta e independentemente do tipo, obedecendo a série natural dos números: 1, 2, 3, 4 ...

A Amostra consiste de:

Todos os Boletins e Listas com números de série terminados em 3 (inclusive os suplementos de qualquer C.D. 1.01 com número terminado em 3). A composição exata da amostra, comparada com a da população, foi a seguinte:

ITEM (1)	TOTAL DO MUNICÍPIO (2)	AMOSTRA	
		Nº (3)	% baseada na coluna (2) (4)
C.D. 1.01 (Família)	9 239	935	10.12
C.D. 1.02 (Coletivo)	115	10	8.70
C.D. 1.03 (Individual)	2 831	284	10.03
Pessoas Recenseadas	52 532	5 280	10.05

Estimativas:

Foram feitas estimativas dos principais itens encontrados na publicação IBGE intitulada: Seleção dos Principais Dados do Estado do Espírito Santo . Duas espécies de estimativas foram feitas para cada item.

Estimativa sem tendenciosidade:

Obtida, simplesmente, multiplicando o total da amostra, para cada característica, por 10.

Estimativa de razões:

Obtida, para cada característica, multiplicando o total de pessoas recenseadas (presume-se no caso que êsse dado seria conhecido antes de concluída a tabulação preliminar) pela proporção de pessoas na amostra, com essa característica.

Erros de Amostragem:

O cálculo dos erros referentes a ambas as estimativas, foi feita para uma amostra representativa das características. Êsses erros de amostragem foram estimados a partir dos resultados da amostra, e o cálculo foi levado a efeito independente do conhecimento das diferenças que existiam entre os totais do censo e as estimativas baseadas nos dados da Amostra.

* * *

ROTEIRO PARA A SELEÇÃO E PROCESSAMENTO DA AMOSTRA

1 - Selecione todos os boletins e listas com números terminados em 3.

2 - Para cada pessoa registrada nesses boletins perfure um cartão do tipo padrão.

3 - A partir dos cartões perfurados, tabule os dados para obter os totais de cada característica desejada.

4 - Multiplique os totais da amostra por 10 para obter a estimativa sem tendenciosidade.

5 - Selecione uma subamostra correspondente a um quinto dos boletins da amostra primitiva.

6 - Use os totais de famílias e os totais de indivíduos da subamostra para calcular os erros de amostragem das estimativas sem tendenciosidade referentes as características selecionadas.

7 - Calcule o número de cartões que devem ser duplicados ou subtraídos para que o número total de pessoas seja exatamente igual a um décimo do total conhecido de pessoas recenseadas.

8 - Selecione cartões ao acaso. Duplique ou subtraia o número de cartões indicado no item anterior.

9 - Tabule os dados dêsse grupo de fichas a fim de obter os totais da amostra referentes as características desejadas.

10 - Multiplique os totais da amostra por 10 para obter as estimativas das razões.

(Nota - Os itens 7 - 10, são virtualmente equivalentes ao processo de estimar a razão descrita acima e elimina a necessidade de multiplicar os resultados da amostra por um número decimal, neste caso 9,9492).

11 - Usando a mesma subamostra selecionada no item 5, calcule os erros de amostragem para as estimativas das razões referentes as características selecionadas.

* * *

RESULTADOS

Nas tabelas que seguem, comparamos os totais do Censo com as estimativas da amostra para um grupo representativo de 40 itens. Êsses 40 itens foram escolhidos independente das diferenças existentes entre os dados do censo e as estimativas baseadas nos dados da amostra.

São também feitas comparações de 3 distribuições percentuais típicas, também selecionadas independentemente das diferenças existentes.

Finalmente, damos uma tabela comparando as distribuições da idade-sexo baseadas no censo e na amostra.

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	Das razões	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor Absoluto	%	Valor Absoluto	%
Pessoas presentes							
5 a 9 anos							
Homens	2 782	2 620	2 580	- 165	- 5.9	- 202	- 7.3
Mulheres	2 724	2 840	2 770	116	4.3	46	1.7

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor Absoluto	%	Valor Absoluto	%
50 a 59 anos							
Homens	1 539	1 420	1 440	- 119	- 7.7	- 99	- 6.4
Mulheres	1 480	1 420	1 440	- 60	- 4.1	- 40	- 2.7
Situação do domicílio							
Quadro urbano							
Homens	19 379	18 620	19 050	- 759	- 3.9	- 329	- 1.7
Mulheres	22 431	22 060	22 410	- 371	- 1.7	- 21	- 0.1
15 a 19 anos							
Homens	2 120	1 960	2 010	- 160	- 7.5	- 110	- 5.2
Mulheres	3 254	3 150	3 150	- 104	- 3.2	- 104	- 3.2
Quadro suburbano e rural							
0 a 4 anos							
Homens	716	920	810	204	28.5	94	13.1
Mulheres	703	830	740	127	18.1	37	5.3
Côr							
Brancos							
5 a 9 anos							
Homens	1 425	1 270	1 290	- 155	- 10.0	- 135	- 9.5
Mulheres	1 432	1 550	1 510	118	8.2	78	5.4
20 a 24 anos							
Homens	1 355	1 230	1 230	- 125	- 9.2	- 125	- 9.2
Mulheres	1 610	1 420	1 410	- 190	- 11.8	- 200	-12.4
Pretos							
Homens	2 347	2 760	2 550	413	17.6	203	8.6
Mulheres	3 022	3 520	3 360	498	16.5	338	11.2
Pardos							
Total	17 968	17 680	17 700	- 288	- 1.6	- 268	- 1.5
10 a 14 anos							
Homens	1 089	990	1 010	- 99	- 9.1	- 79	- 7.3
Mulheres	1 156	1 150	1 160	- 6	- 0.5	4	0.3
Estado Conjugal							
Solteiro							
20 a 24 anos							
Homens	2 079	1 810	1 790	- 269	- 12.9	- 289	- 13.9
Mulheres	1 939	1 850	1 840	- 89	- 4.6	- 99	- 5.1
Cádados							
Homens	7 528	7 600	7 620	72	1.0	92	1.2
Mulheres	7 884	8 000	8 020	116	1.5	136	1.7

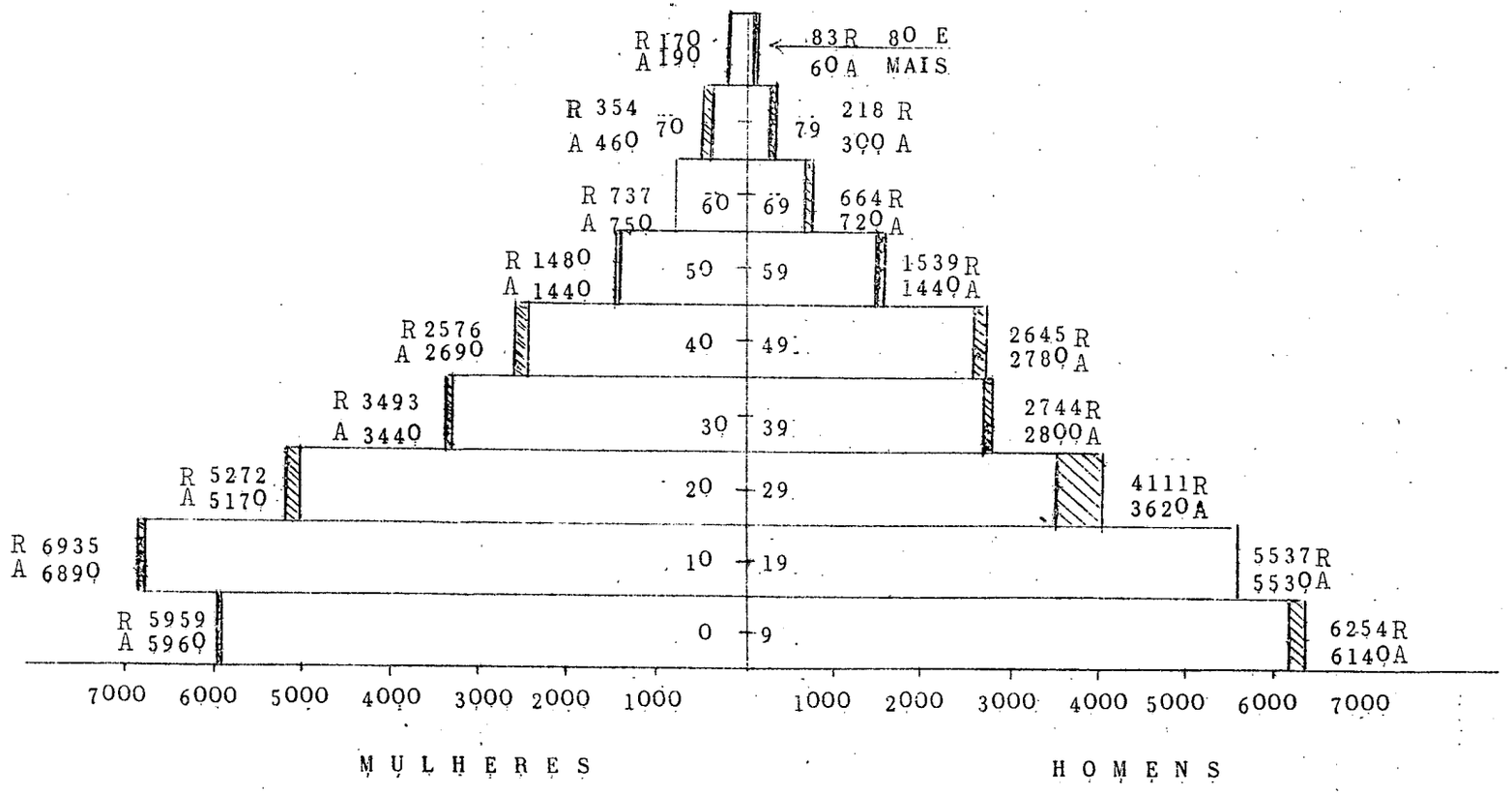
ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor Absoluto	%	Valor Absoluto	%
Religião							
Católicos Romanos							
25 a 29 anos							
Homens	1 521	1 320	1 360	- 201	- 13.2	- 161	- 10.6
Mulheres	1 993	2 040	2 040	47	2.4	47	2.4
Protestantes	2 180	2 750	2 690	570	26.1	510	23.4
Brasileiros natos							
Homens	23 244	23 020	22 900	- 224	- 1.0	- 344	- 1.5
Mulheres							
30 a 39 anos	3 458	3 380	3 380	- 78	- 2.3	- 78	- 2.3
Bahianos	805	770	750	- 35	- 4.3	- 55	- 6.8
Mineiros							
Mulheres	907	910	880	3	0.3	- 27	- 3.0
Homens	1 134	1 250	1 260	116	10.2	126	11.1
Instrução							
Sabem ler							
10 a 14 anos	4 927	5 010	4 930	83	1.7	3	0.1
25 a 29 anos							
Mulheres	1 621	1 710	1 710	89	5.5	89	5.5
Não sabem ler e escrever							
Homens	5 034	5 020	4 840	- 14	- 0.3	- 194	- 3.9
Mulheres	7 687	8 000	7 930	313	4.1	243	3.2
Atividade							
Industria de transformação							
Homens	1 944	1 760	1 770	- 184	- 9.5	- 174	- 9.0
Prestação de serviços							
Homens	1 772	1 720	1 740	- 52	- 2.9	- 32	- 1.8
Mulheres	2 932	3 020	3 020	88	3.0	88	3.0
Atividades sociais							
Condições inativas							
20 a 29 anos	329	340	340	11	3.3	11	3.3
DISTRIBUIÇÕES PERCENTUAIS							
Côr	100.0	100.0	100.0				
Brancos	54.0	52.7	53.2	- 1.3	- 2.4	- 0.8	- 1.4
Pretos	10.5	12.4	11.7	1.9	18.1	1.2	11.4
Pardos	35.3	34.8	35.0	- 0.5	- 1.4	- 0.3	- 0.8
Amarelos e ignorados	0.2	0.1	0.1	- 0.1	- 50.0	- 0.1	- 50.0

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor Absoluto	%	Valor Absoluto	%
Estado conjugal-homens	100.0	100.0	100.0				
Solteiros	45.4	44.1	44.0	- 1.3	- 2.9	- 1.4	- 3.1
Casados	51.0	52.7	52.9	1.7	3.3	1.9	3.7
Viuvos	3.1	2.9	2.9	- 0.2	- 6.5	- 0.2	- 6.5
Desquitado, divorciado e sem especificação	0.5	0.3	0.2	- 0.2	- 40.0	- 0.3	- 60.0
Brancos	100.0	100.0	100.0				
0 a 4 anos	12.8	12.9	12.8	0.1	0.8	0.0	0.0
5 a 9 anos	10.4	10.6	10.4	0.2	1.9	0.0	0.0
10 a 14 anos	11.4	12.0	11.8	0.6	5.3	0.4	3.5
15 a 19 anos	12.5	12.2	12.0	- 0.3	- 2.4	- 0.5	- 4.0
20 a 29 anos	18.5	17.2	17.2	- 1.3	- 7.0	- 1.3	- 7.0
30 a 39 anos	12.8	12.5	12.7	- 0.3	- 2.3	- 0.1	- 0.8
40 a 49 anos	10.5	11.2	11.4	0.7	6.7	0.9	8.6
50 a 59 anos	6.2	5.8	6.0	- 0.4	- 6.5	- 0.2	- 3.2
60 a 69 anos	3.0	3.6	3.6	0.6	20.0	0.6	20.0
70 a 79 anos	1.3	1.4	1.5	0.1	7.7	0.2	15.4
80 anos e mais	0.5	0.5	0.5	0.0	0.0	0.0	0.0
Idade ignorada	0.1	0.1	0.1	0.0	0.0	0.0	0.0

POPULAÇÃO PRESENTE POR SEXO E GRUPOS DE IDADES

MUNICÍPIO DE VITÓRIA - ESTADO DO ESPÍRITO SANTO

Confronto entre os resultados do censo e da amostra



R - Valor real, obtido pelo Recenseamento

A - Valor da estimativa sem tendenciosidade, obtida da amostra

▨ - Diferença (+ ou -) entre R e A

ERROS DE AMOSTRAGEM

Estimativas baseadas em amostras estão sujeitas a erros de amostragem. Êstes erros de amostragem dependem primeiramente do tamanho da amostra e do tamanho da estimativa. A tabela abaixo dá aproximadamente os erros de amostragem, para estimativas de vários tamanhos, tanto para as estimativas sem tendenciosidade como para as estimativas de razão. Para as sem tendenciosidade, as estimativas maiores têm erros absolutos de amostragem maiores mas erros relativos menores. Para a estimativa de razão quando o tamanho da estimativa aproxima-se do número total de pessoas recenseadas, ambos os erros absolutos e relativos de amostragem diminuem.

TAMANHO DA ESTIMATIVA	ÊRROS DE AMOSTRAGEM (um desvio padrão) DE			
	Estimativa sem tendenciosidade		Estimativa de razão	
	Absoluto	Relativo (%)	Absoluto	Relativo (%)
100	38	37.6	38	37.6
250	59	23.8	59	23.7
500	84	16.8	84	16.7
1 000	119	11.9	118	11.8
2 500	188	7.5	184	7.3
5 000	267	5.3	253	5.1
10 000	378	3.8	338	3.4
25 000	603	2.4	430	1.7
50 000	860	1.7	185	0.4
52 532 *	883	1.7	0	0.0

*Pessoas recenseadas.

Êstes erros de amostragem podem ser interpretados da maneira seguinte:

Para um item censitário de tamanho dado, existe probabilidade de 2 em 3 de que uma estimativa baseada numa amostra do tipo usado nesta experiência não diferirá do valor verdadeiro (censo) de mais que o êrro de amostragem que se vê na tabela para aquêle tipo de estimativa. Há probabilidade de 19 em 20 de que a estimativa não diferirá do valor verdadeiro de mais que o dôbro do êrro de amostragem apresentado na tabela.

Exemplo:

Para uma característica atribuída a 10 000 pessoas no Censo do Município de Vitória, as probabilidades são de 2 em 3 que uma estimativa sem tendenciosidade baseada na amostra usada não diferirá do valor censitária mais de 378, isto é, estará entre os limites de 9.622 para 10.378. Existem probabilidades de 19 em 20 que a estimativa ficará em 9.244 o 10.756.

Há certas características que tendem a agrupar-se por família, isto é, ou todos os membros da família tem a característica ou nenhum dêles a tem. Isto é particularmente verdadeiro no que concerne religião e côr. Quando se usa uma amostra de famílias, estimativas dêstes tipos de características estão sujeitas a erros de amostragem maiores do que as estimativas de outras características como idade, sexo e muitas outras não sujeitas a êste tipo de agrupamento. Portanto, os erros de amostragem apresentados na tabela acima são subestimativas para estimativas relacionadas a côr e religião, particularmente para os grupos menores de côr e religião.

ESTIMATIVAS SEM TENDENCIOSIDADE VS ESTIMATIVAS DE RAZÃO

A principal vantagem da estimativa de razão é que ela reduz os erros de amostragem para itens grandes como é demonstrado na tabela do capítulo anterior. Para itens pequenos pouca diferença há entre a estimativa de razão e as estimativas sem tendenciosidade. Uma análise dos 40 itens para os quais as 2 estimativas foram comparadas com os censos totais mostra:

Para 21 itens a estimativa de razão estava mais próxima do valor verdadeiro

Para 11 itens a estimativa sem tendenciosidade estava mais próxima do valor verdadeiro

Para 8 itens a diferença foi a mesma para ambas as estimativas.

Ao decidir sôbre a estimativa a usar é necessário balancear o aumento da fidedignidade obtido com a estimativa de razão contra o seu custo maior, resultante da necessidade de realizar operações processuais adicionais antes de fazer as tabulações (Itens 7 e 8 no roteiro).

Perda de eficiência resultante do uso de uma amostra BOLETINS em vez de uma amostra de INDIVÍDUOS

A amostra para o Município de Vitória consistiu de aproximadamente 10% de todos os boletins. Na base da bem conhecida teoria de amostragem, é possível computar os erros de amostragem de estimativas baseadas numa amostra simples ao acaso contendo o mesmo número de indivíduos. Isto foi feito, e na tabela que se segue, os resultados são comparados com os erros de amostragem prèviamente calculados para estimativas baseadas em nossa amostra de boletins.

Tomando por base o êrro de amostragem da amostra simples ao acaso de indivíduos podemos ver que o aumento médio no êrro de amostragem devido à amostragem de boletins usando a estimativa de razão, é de 20%. Uma vez que o êrro de amostragem é grosseiramente proporcional à raiz quadrada do tamanho da amostra, podemos

TAMANHO DA ESTIMATIVA	ÊRRO DE AMOSTRAGEM (1 desvio padrão) PARA		
	AMOSTRA SISTEMÁTICA DE BOLETINS		Amostra simples ao acaso de indivíduos
	Usando estimativa sem tendenciosidade	Usando estimativa de razão	
100	38	38	32
250	59	59	50
500	84	84	70
1 000	119	118	99
2 500	188	184	154
5 000	267	253	212
10 000	378	338	284
25 000	603	430	361
50 000	860	185	155
52 532*	883	0	0

* Pessoas recenseadas

ver que seria necessário aumentar a amostra dos boletins de cêrca de 42%, isto é, selecionar 14% de todos os boletins em vez de 10% para reduzir os erros de amostragem ao nível daqueles que resultaria de uma amostra simples ao acaso de indivíduos da ordem de 10%:

CONCLUSÕES

Baseadas nesta experiência podemos concluir que:

1º - O plano de amostragem usando para o Município de Vitória foi conduzido corretamente e produziu os resultados fidedignos previstos;

2º - Um plano de amostragem dêste tipo poderia ser usado para fazer as tabulações preliminares do censo de 1960 tornando possível a publicação de resultados com muito maior antecedência que nos censos anteriores;

3º - Os manipuladores dos dados censitários deveriam ser consultados para determinar se os resultados desta amostra são suficientemente fidedignos para os seus propósitos (lembrando que os resultados baseados no censo completo seriam publicados numa data posterior) ou se deveria ser usada uma amostra maior de forma a reduzir os erros de amostragem;

4º - O pessoal do Núcleo não teve dificuldade em seguir as instruções para a seleção e tabulação da amostra e para o cálculo dos erros de amostragem.

5º - Se forem aplicados métodos de amostragem nos censos de 1960 é essencial que o Núcleo obtenha os serviços de tempo integral de uma ou mais pessoas com conhecimento de estatística matemática e com treinamento e experiência da teoria e aplicações da amostragem.

I.B.G.E. - Conselho Nacional de Estatística
NÚCLEO DE PLANEJAMENTO CENSITÁRIO

RELATÓRIO SÔBRE UMA EXPERIÊNCIA DESTINADA A ESTUDAR ALGUNS PROCESSOS DE AMOSTRAGEM, DOS QUESTIONÁRIOS USADOS NO CENSO DEMOGRÁFICO DE 1950

T. B. Jabine

b) Estado do Espírito Santo

- Setembro 1956 -

INTRODUÇÃO

Êste relatório é a. parte b do Relatório Sôbre Uma Experiência Destinada a Estudar Alguns Processos de Amostragem, dos Questionários Usados no Censo Demográfico de 1950, e deve ser considerado em conjunto com a parte a.

A parte a do relatório expõe os objetivos gerais da experiência de obtenção de amostras do censo de 1950 e descreveu o esbôço da amostra e os resultados obtidos para o Município de Vitória. Subseqüentemente, obteve-se uma segunda amostra para representar o resto do Estado do Espírito Santo e os resultados das duas amostras foram combinadas para dar estimativas para o Estado.

O propósito dêste relatório é descrever o desenho da amostra e os processos da estimação usados para o interior do Estado e apresentar os resultados inclusive os erros de amostragem, para o Estado, como um todo.

PLANO DE AMOSTRAGEM

O universo do qual foi retirada a amostra, consistiu de

Todos os Boletins de Família (Modêlo C.D. 1.01), Listas de Domicílio Coletivo (Modêlo C.D. 1.02) e Boletins Individuais (Modêlo C.D. 1.03) coletados no Estado do Espírito Santo, com exceção do Município de Vitória. Êstes boletins e listas estavam contidos em 628 pastas; dentro de cada pasta estavam numerados em séries consecutivas a começar por 1, independentemente do "modêlo" de boletins ou lista.

A amostra consistiu de

Aproximadamente 1% dos boletins e listas do Universo, sendo selecionada em duas fases.

Sua composição exata foi a seguinte:

ESPECIFICAÇÃO	TOTAL DO INTERIOR DO ESTADO	TAMANHO DA AMOSTRA	
		Números absolutos	Números relativos (%)
C.D. 1.01 (Família)	146 703	1 490	1,02
C.D. 1.02 (Coletivo)	499	2	0,40
C.D. 1.03 (Individual)	5 721	25	0,44
Total de Boletins e Listas	152 923	1 517	0,99
Pessoas recenseadas	*819 535	8 148	0,99

*Total definitivo

1. Na primeira fase, a população foi dividida em unidades primárias

de amostragem, assim consideradas as pastas que continham questionários referentes a uma única "situação" (urbana, suburbana ou rural). Quando contivessem questionários de mais de uma "situação", cada grupo correspondente a determinada "situação" foi considerado uma unidade primária separada. Uma amostra de tais unidades primárias (pastas ou parte das pastas) foi selecionada com probabilidade proporcional ao tamanho, e com estratificação por "situação". Foram incluídos nessa amostra:

SITUAÇÃO	UNIDADES PRIMÁRIAS DE AMOSTRAGEM
Urbana	8
Suburbana	5
Rural	49
Total	62

ou aproximadamente 10% de tôdas as pastas.

2. Na segunda fase, uma sub-amostra sistemática de boletins e listas foi selecionada de cada uma das unidades primárias, adotando-se determinada freqüência em cada caso, de modo que, em combinação com a probabilidade de seleção para as unidades primárias, se obteve uma probabilidade global de seleção de 1 em 100 para cada boletim no universo. Em média, foram selecionados 25 boletins de cada unidade primária de amostragem.

Estimativas para o Interior:

Obtiveram-se estimativas sem tendenciosidade para cada característica multiplicando o total da amostra por 100. Estimativas de razão foram obtidas multiplicando o total de pessoas recenseadas em cada situação pela proporção de pessoas que, na amostra correspondente a essa situação possuiam a característica em causa. Por presunção, êsses totais seriam disponíveis para utilização com os resultados da amostra preliminar. Devido a certas aproximações usadas na elaboração dessas estimativas, o valor da amostra para o total de pessoas recenseadas difere ligeiramente do valor obtido pelo Censo.

Estimativas para o Estado em conjunto

Para o total do Estado, as estimativas sem tendenciosidade e as estimativas de razão foram obtidas pela simples adição das estimativas para o interior às correspondentes a Vitória.

Erros de amostragem:

Os erros de amostragem das estimativas para o Estado inteiro foram computados sòmente para as estimativas de razão, uma vez que geralmente gozam de preferência sôbre as estimativas sem tendenciosidade. A amostra para o interior do Estado foi selecionada na forma de 5 sub-amostras interpenetrantes, de modo a facilitar a estimativa das variâncias.

Nas tabelas seguintes, comparam-se os totais do Censo com as estimativas da amostra, para vários itens. A primeira tabela dá os valores para os itens principais.

Na segunda tabela comparam-se os totais do Censo com estimativas, para um grupo representativo de 40 itens. Êsses 40 itens foram escolhidos independentemente das diferenças existentes entre os dados do Censo e as estimativas baseadas nos dados da amostra.

São também feitas, na tabela 3, comparações de 3 distribuições percentuais típicas, também selecionadas independentemente das diferenças existentes.

Finalmente, dá-se na Tabela 4 a compração das distribuições de idade e sexo, baseadas no Censo e na amostra.

Tabela nº 1 - Estimativas dos Itens Principais

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor absoluto	%	Valor absoluto	%
População recenseada.....	872 067	867 600	870 390*	- 4 467	- 0,5	- 1 677	- 0,2
Homens	443 565	440 730	442 970	- 2 835	- 0,6	- 595	- 0,1
Mulheres	428 502	426 870	427 420	- 1 632	- 0,4	- 1 082	- 0,3
População presente	861 562	856 390	860 240	- 5 172	- 0,6	- 1 322	- 0,2
Homens	436 939	433 680	436 850	- 3 259	- 0,7	- 89	0,0
Mulheres	424 623	422 710	423 390	- 1 913	- 0,5	- 1 233	- 0,3
de 15 anos e mais	477 794	480 720	485 550	+ 2 926	+ 0,6	+ 7 756	+ 1,6
Homens	241 654	240 710	244 000	- 944	- 0,4	+ 2 346	+ 1,0
Mulheres	236 140	240 010	241 550	+ 3 870	+ 1,6	+ 5 410	+ 2,3
Quadro urbano	136 106	132 380	134 660	- 3 726	- 2,7	- 1 446	- 1,1
Homens	64 343	63 120	63 850	- 1 223	- 1,9	- 493	- 0,8
Mulheres	71 765	69 260	70 810	- 2 503	- 3,5	- 953	- 1,3
Quadros suburbano e rural ..	725 456	724 010	725 580	- 1 446	- 0,2	+ 124	0,0
Homens	372 596	370 560	373 000	- 2 036	- 0,5	+ 404	+ 0,1
Mulheres	352 860	353 450	352 580	+ 590	+ 0,2	- 280	- 0,1

*Devido a certas aproximações usadas na elaboração das estimativas de razão, o valor da amostra para o total de pessoas recenseadas difere ligeiramente do valor censitário.

Tabela nº 2 - 40 Itens Representativos

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor absoluto	%	Valor absoluto	%
Pessoas presentes							
5 a 9 anos							
Homens	61 849	60 720	59 680	- 1 129	- 1,8	- 2 169	- 2,5
Mulheres	59 434	59 440	58 570	6	0,0	- 864	- 1,5
50 a 59 anos							
Homens	21 947	22 420	22 740	473	2,2	793	3,6
Mulheres	18 097	19 320	19 240	1 223	6,8	1 143	6,3
Situação do domicílio							
Quadro urbano							
Homens	64 343	63 120	63 850	- 1 223	- 1,9	- 493	0,8
Mulheres	71 763	69 260	70 810	- 2 503	- 3,5	- 953	- 1,3
15 a 19 anos							
Homens	7 018	5 860	6 010	- 1 158	-16,5	- 1 008	-14,4
Mulheres	9 757	10 050	9 750	293	3,0	- 7	0,1
Quadro suburbano e rural							
0 a 4 anos							
Homens	67 153	66 320	66 810	- 833	- 1,2	- 343	- 0,5
Mulheres	64 459	61 930	61 240	- 2 529	- 3,9	- 3 219	- 5,0
Côr							
Brancos							
5 a 9 anos							
Homens	36 110	37 570	36 690	1 460	4,0	580	1,6
Mulheres	34 992	36 450	36 610	1 458	4,2	1 618	4,6
20 a 24 anos							
Homens	23 241	26 230	26 630	2 989	12,9	3 389	14,6
Mulheres	24 896	27 320	26 710	2 424	9,7	1 814	7,3
Pretos							
Homens	51 726	45 860	46 850	- 5 866	-11,3	- 4 876	- 9,4
Mulheres	50 719	45 920	45 960	- 4 799	- 9,5	- 4 759	- 9,4
Pardos							
Total	253 423	231 080	231 200	- 22 343	- 8,8	- 22 223	- 8,8
10 a 14 anos							
Homens	17 323	15 890	16 110	- 1 433	- 8,3	- 1 213	- 7,0
Mulheres	16 542	13 750	14 060	- 2 792	-16,9	- 2 482	-15,0

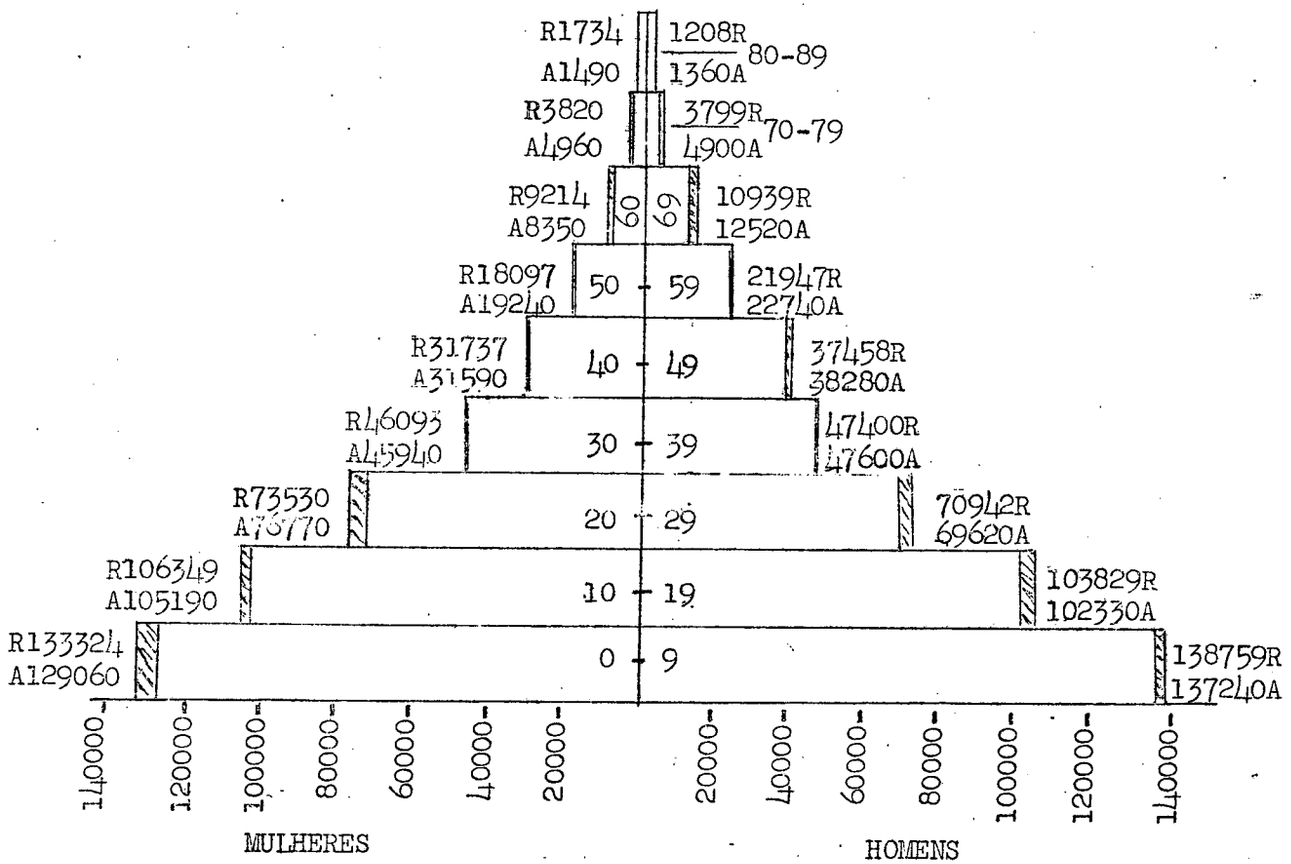
ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão.	
				Valor absoluto	%	Valor absoluto	%
Estado conjugal							
Solteiros							
20 a 24 anos							
Homens	29 483	30 010	30 090	527	1,8	607	2,1
Mulheres	18 333	19 150	19 440	817	4,5	1 107	6,0
Casados							
Homens	132 277	133 900	135 220	1 623	1,2	2 943	2,2
Mulheres	134 443	136 200	136 920	1 757	1,3	2 477	1,8
Religião							
Católicos romanos							
25 a 29 anos							
Homens	27 295	26 320	26 360	- 975	- 3,6	- 935	- 3,4
Mulheres	27 701	29 040	30 240	1 339	4,8	2 539	9,2
Protestantes	75 076	70 950	74 390	- 4 126	- 5,5	- 686	- 0,9
Brasileiros natos							
Homens	433 185	429 720	432 900	- 3 465	- 0,8	- 285	- 0,1
Mulheres							
30 a 39 anos	45 954	46 080	45 780	126	0,3	- 174	- 0,4
Bahianos	5 490	2 670	2 450	- 2 820	-51,4	- 3 040	-55,4
Mineiros							
Homens	27 430	24 310	25 480	- 3 120	-11,4	- 1 950	- 7,1
Mulheres	23 481	21 250	22 060	- 2 231	- 9,5	- 1 421	- 6,1
Instrução							
Sabem ler							
10 a 14 anos	48 741	48 710	47 630	- 31	- 0,1	- 1 111	- 2,3
25 a 29 anos							
Mulheres	13 665	14 510	15 510	845	6,2	1 845	13,5
Não Sabem ler e ignorado							
Homens	196 813	199 420	201 940	2 607	1,3	5 127	2,6
Mulheres	222 787	223 400	223 430	613	0,3	643	0,3
Ramo da atividade principal							
Indústrias de transformação							
Homens	14 587	10 760	10 970	- 3 827	-26,3	- 3 617	-24,8
Prestação de serviços							
Homens	8 023	7 420	7 640	- 603	- 7,3	- 383	- 4,7
Mulheres	10 656	10 020	10 420	- 636	- 6,0	- 236	- 2,2
Atividades sociais	5 728	5 000	5 000	- 727	-12,7	- 728	-12,7
Condições inativas							
20 a 29 anos	2 579	2 540	2 340	- 39	- 1,5	- 239	- 9,3

ITEM	CENSO TOTAL	ESTIMATIVAS A PARTIR DA AMOSTRA		DIFERENÇA ENTRE O TOTAL DO CENSO E AS ESTIMATIVAS			
		Sem tendenciosidade	De razão	Estimativas sem tendenciosidade		Estimativas de razão	
				Valor absoluto	%	Valor absoluto	%
Côr	100,0	100,0	100,0				
Brancos	58,6	62,1	62,1	3,5	6,0	3,5	6,0
Pretos	11,9	10,7	10,8	- 1,2	-10,1	- 1,1	- 9,2
Pardos	29,4	27,0	26,9	- 2,4	- 8,2	- 2,5	- 8,5
Amarelos e ignorados	0,1	0,2	0,2	0,1	50,0	0,1	50,0
Estado conjugal - homens ...	100,0	100,0	100,0				
Solteiros	42,0	40,6	40,9	- 1,4	- 3,3	- 1,1	- 2,6
Casados	54,7	55,6	55,5	0,9	1,6	0,8	1,5
Viúvos	3,1	3,7	3,5	0,6	19,4	0,4	12,9
Desquitados, divorciados e sem especificação	0,2	0,1	0,1	- 0,1	-50,0	- 0,1	-50,0
Brancos	100,0	100,0	100,0				
0 a 4 anos	17,9	17,6	17,6	- 0,3	- 1,7	- 0,3	- 1,7
5 a 9 anos	14,1	13,9	13,7	- 0,2	- 1,1	- 0,4	- 2,2
10 a 14 anos	12,8	12,5	12,3	- 0,3	- 2,3	- 0,5	- 3,9
15 a 19 anos	11,2	11,2	11,3	0,0	0,0	0,1	0,9
20 a 29 anos	16,9	17,2	17,4	0,3	1,8	0,5	3,0
30 a 39 anos	11,0	10,8	11,0	- 0,2	- 1,8	0,0	0,0
40 a 49 anos	7,8	8,1	7,9	0,3	3,8	0,1	1,3
50 a 59 anos	4,6	4,7	4,7	1,1	2,2	0,1	2,2
60 a 69 anos	2,4	2,5	2,5	0,1	4,2	0,1	4,2
70 a 79 anos	0,9	1,0	1,1	0,1	11,1	0,2	22,2
80 anos e mais	0,3	0,4	0,4	0,1	33,3	0,1	33,3
Idade ignorada	0,1	0,1	0,1	0,0	0,0	0,0	0,0

POPULAÇÃO PRESENTE POR SEXO E GRUPOS DE IDADES

ESTADO DO ESPÍRITO SANTO

Confronto entre os resultados do Censo e da amostra



R = Valor real, obtido pelo Recenseamento

A = Valor da estimativa sem tendenciosidade, obtido da amostra

= Diferença (+ ou -) entre R e A

ERROS DE AMOSTRAGEM

Estimativas baseadas em amostras estão sujeitas a erros de amostragem. Êstes erros de amostragem dependem primeiramente do tamanho e desenho da amostra e do tamanho da estimativa. A tabela abaixo dá aproximadamente os erros de amostragem para estimativas da razão de vários tamanhos, para o Estado do Espírito Santo.

TAMANHO DA ESTIMATIVA	ÊRROS DE AMOSTRAGEM (um desvio padrão)	
	Absoluto	Relativo (%)
1 000	342	34.2
2 500	540	21.6
5 000	763	15.3
10 000	1 076	10.6
25 000	1 685	6.7
50 000	2 350	4.7
100 000	3 220	3.2
250 000	4 575	1.8
500 000	5 000	1.0
750 000	3 490	0.5

Êstes erros de amostragem podem ser interpretados da maneira seguinte:

Para um item censitário do tamanho dado, existe probabilidade de 2 em 3, de que a estimativa baseada numa amostra do tipo usado nesta experiência não diferirá do valor verdadeiro (censo) de mais que o êrro de amostragem correspondente na tabela acima. Há probabilidade de 19 em 20, de que a estimativa não diferirá do valor verdadeiro mais que o dôbro de êrro de amostragem correspondente.

Exemplo:

Se o censo mostra que 100,000 pessoas do Estado do Espírito Santo têm uma característica particular, as probabilidades são de 2 em 3 de que uma estimativa de razão baseada no tipo de amostra usado não diferirá do valor censitário de mais de 3,220, isto é, estará entre os limites de 96,780 e 103,220. Existem probabilidades de 19 em 20 que a estimativa ficará entre 93,560 e 106,440.

Certas características tendem a agrupar-se por família, isto é, ou todos os membros da família têm a característica, ou nenhum dêles a tem. Isto é particularmente válido no que concerne à religião e côr. Quando se usa uma amostra de famílias, as estimativas destas características estão sujeitas a erros de amostragem maiores do que as estimativas de outras características, como idade, sexo e muitas outras não sujeitas a tal tipo de agrupamento. Portanto, os erros de amostragem apresentados na tabela não se aplicam às estimativas relacionadas a côr e religião, particularmente nos grupos de menor tamanho.

Além disso, certas características tendem a agrupar-se em unidades maiores, tais como Município ou Distrito. Na amostra para o interior, onde sòmente 10% das pastas foram incluídas, nem todos os municípios e distritos foram representados. Por conseguinte, estimativas de características que tendam a agrupar-se em certas áreas do Estado terão erros de amostragem maiores do que aquêles apresentados na tabela acima. No presente caso, verifica-se que isso acontece especialmente quanto às estimativas do número de pessoas nascidas em Estados vizinhos, como poderá ser observado nos resultados apresentados para pessoas nascidas no Estado da Bahia, onde a diferença observada é bem maior do que o êrro de amostragem dado na tabela.

CONCLUSÕES

1. A amostra para o interior do Estado foi planejada de modo a fornecer, em combinação com a amostra para Vitória, estimativas para o Estado em conjunto, as quais deveriam apresentar, erros relativos de amostragem equivalentes aos de Vitória. Na realidade, as estimativas de razão para o Estado têm erros de amostragem um pouco menores do que os de Vitória.

2. Se fôr usada uma amostra de duas fases, como a efetuada para o Interior do Estado, será interessante, provàvelmente, aumentar o número de pastas incluídas na amostra, reduzindo o tamanho da amostra dentro de cada pasta. Isto é particularmente importante se desejarmos reduzir os erros de amostragem para certos itens, como a anturalidade. Está sendo feito um estudo para medir o efeito sôbre o êrro de amostragem decorrente da utilização de pequenos agrupamentos.

3. Para a maioria dos ítens, não é muito grande o aumento de erros de amostragem decorrente da utilização de uma amostra de famílias, em vez de uma simples amostra ao acaso de indivíduos. Entretanto, há certas classes de estimativas, tais como as de côr e religião relativas a grupos pequenos, para as quais a utilização de uma amostra de família dá um êrro de amostragem consideràvelmente maior.

4. Devido à possibilidade de redução dos erros de amostragem, tanto para tabulações preliminares como para tabulações de perguntas de amostra, o Núcleo deveria considerar cuidadosamente a possibilidade de usar em 1960 em questionário de um tipo que facilitasse a seleção de uma amostra de indivíduos, em vez de uma amostra de famílias. Isto não elimina necessàriamente a possibilidade da utilização de um questionário de tipo familiar.

APLICAÇÃO DA AMOSTRAGEM NO CENSO COMERCIAL *

Experimentos realizados para o comércio varejista, com base nos dados do Distrito Federal

Refere-se êste relatório a dois experimentos de amostragem levados a efeito, para o Distrito Federal, com base nos questionários do Censo Comercial de 1950 (comércio a varejo). Os planos elaborados para a seleção das amostras tinham por objetivo apresentar estimativas concernentes a: Valor das vendas, pessoal ocupado, salários, despesas diversas, estoques e número de estabelecimentos. Admitiu-se um êrro de amostragem de até 5% para o valor das vendas, no primeiro experimento; e para o número de pessoas ocupadas, no segundo.

Pretendeu-se ainda demonstrar que a eficiência de uma amostra estratificada não requer necessàriamente a separação física de todos os questionários. Nos casos em foco, os questionários tanto de comércio varejista como do comércio atacadista, encontravam-se acondicionados indiscriminadamente em 64 pastas, por ordem dos setores de coleta. Cada pasta possuía em média 300 questionários, numerados de 1 até o final.

Primeiro experimento - A fim de tornar eficiente a amostra no referente ao total do valor das vendas, usou-se uma estratificação dos estabelecimentos em três grupos(segundo o valor de vendas). Como os demais itens (pessoal ocupado, salários, etc.)estão relacionados com o valor das vendas, tal estratificação também reduziria o êrro de amostragem quanto às estimativas correspondentes.

Os resultados reais do Censo fornecem uma distribuição dos estabelecimentos varejistas segundo diferentes faixas do valor de vendas, o que permitiu determinar: a) que êsse valor, para os estabelecimentos compreendidos na faixa de Cr$ 1 milhão a menos de Cr$ 5 milhões, representa cêrca de 1/3 do total das vendas no Distrito Federal, somando 1 496 estabelecimentos; b) que cêrca de 1/4 do referido valor corresponde aos estabelecimentos da faixa do Cr$ 5 milhões e mais (295 estabelecimentos); e c) que os restantes 5/12 do total das vendas eram realizados pelos estabelecimentos (13.374) da faixa de menos de Cr$ 1 milhão. Essa verificação indicou que o grupo de Cr$ 5 milhões e mais devia ser computado integralmente na amostra, e também que o grupo de 1 milhões de cruzeiros a menos de 5 milhões devia ter representação percentual bem mais elevada do que o grupo de menos de 1 milhão, embora êste representasse cêrca de 88% do total dos estabelecimentos.

Com apôio nesses elementos, procedeu-se à seleção de uma amostra sistemática na base de 10% sôbre o total dos questionários, empregando-se números alea

---

* Êste relatório, apresentado pelo sr. Rudolf Wuensche, Assistente-Técnico do Núcleo de Planejamento Censitário, refere-se a experimentos realizados com base no material de coleta do Censo Comercial de 1950, visando a testar a aplicabilidade da amostragem para a obtenção de resultados preliminares em um levantamento de tal natureza. A experiência foi levada a efeito pelo Assistente-Técnico Rudolf Wuensche, sob a orientação geral do sr. T. Jabine.

tórios para evitar qualquer tendenciosidade. Dentro dessa amostra, foi ainda examinada uma sub-amostra de 1%, sendo que, da primeira, retiraram-se sòmente questionários de estabelecimentos varejistas com valor de vendas entre Cr$ 1 milhão e Cr$ 5 milhões; e da segunda, questionários de estabelecimentos varejistas com qualquer valor de vendas. Os questionários de estabelecimentos varejistas com vendas de Cr$ 5 milhões e mais foram retirados na totalidade. Os resultados da seleção foram os seguintes:

1º Exp. - Seleção da amostra segundo grupos de vendas

GRUPOS DE VENDAS (Cr$ 1 000)	TOTAL DE ESTABELECIMENTOS	TAMANHO DA AMOSTRA	
		Números absolutos	Números relativos
0 a 999	13 374	132	0,98
1 000 a 4 999	1 496	160	10,69
5 000 e mais	295	295	100,00
TOTAIS	15 165	587	3,87

Como se verifica do quadro acima, a amostra selecionada limitou-se a 587 questionários, ou sejam, 3,87% do número total. Procedeu-se à apuração do valor de vendas e demais itens partindo dos questionários obtidos na amostra de 1%, prosseguindo-se com os da amostra de 10%, e finalmente, os do grupo totalmente computado. Obtiveram-se os resultados que se seguem:

1º Exp. - Apuração dos questionários selecionados

ESPECIFICAÇÃO	ESTABELECIMENTOS	PESSOAL OCUPADO	(Cr$ 1 000)			
			Salários	Despesas Diversas	Vendas	Estoques
Amostra de 1%	132	311	1 440	4 830	32 685	10 579
Amostra de 10%	160	1 395	19 130	35 503	328 366	95 931
Grupo de 5 000 000,00 e mais ...	295	13 478	272 952	420 103	3 845 303	869 545

As estimativas simples sem tendenciosidade foram obtidas multiplicando-se os dados da amostra de 1% por 100, e os da amostra de 10%, por 10. Procedeu-se em seguida à soma dos produtos obtidos, acrescentando-se-lhe os dados referentes ao grupo de Cr$ 5 milhões e mais. Os resultados são apresentados no quadro abaixo, que também mostra o êrro de amostragem relativo a cada item pesquisado.

1º Exp. - Resultados da amostra

ITEM PESQUISADO	VALOR REAL (Censo) -a-	ESTIMATIVA (amostra) -b-	DIFERENÇA (b-a)		ÊRRO DE AMOSTRAGEM RELATIVO (um desvio padrão)
			Absoluta	%	
Estabelecimentos (nº)	15 165	15 095	- 70	0,46	4,3
Pessoal ocupado	56 910	58 528	+ 1 618	2,84	3,5
Salários (Cr$ 1 000)	573 809	608 252	+ 34 443	6,00	4,8
Despesas diversas (Cr$ 1 000) .	1 251 605	1 258 133	+ 6 528	0,52	4,2
Vendas (Cr$ 1 000)	9 745 937	10 397 463	+ 651 526	6,68	3,8
Estoques (Cr$ 1 000)	2 566 610	2 886 755	+ 320 145	12,47	7,2

Segundo Experimento - Sua finalidade foi estimar os mesmos itens do primeiro, estratificando-se os questionários da amostra segundo o números de pessoas ocupadas, em vez de o valor de vendas. Dada a relação existente entre o pessoal ocupado e os outros itens, tal estratificação também diminuiria o êrro de amostragem para todos êles.

A distribuição dos estabelecimentos segundo grupos de pessoal ocupado (resultados do Censo de 1950) indicou que: a) cêrca de 1/2 do total do pessoal ocupado referia-se aos estabelecimentos que contavam de 0 a 5 pessoas (13 325 estabelecimentos); b) 1/4 do total referia-se ao grupo de 6 a 20 pessoas, somando 1 600 estabelecimentos; e c) o quarto restante, ao grupo de 21 pessoas e mais, somando 240 estabelecimentos. Isto mostrou que a amostra a selecionar deveria conter todos os questionários com 21 e mais pessoas, e que o grupo de 6 a 20 pessoas precisaria ter representação percentual bem maior do que a do grupo de 0 a 5 pessoas, apesar dêste último representar cêrca de 88% do total de estabelecimentos.

Examinaram-se duas amostras sistemáticas independentes, uma na base de 10% sôbre o total de questionários, sendo retirados sòmente os questionários de estabelecimentos varejistas que apresentassem 6 a 20 pessoas ocupadas, e outra na base de 2% sôbre o mesmo total, na qual se retiraram os questionários de estabelecimentos varejistas que estivessem compreendidos no grupo de 0 a 5 pessoas. Os questionários de estabelecimentos varejistas com 21 e mais pessoas foram retirados na totalidade. Na seleção das amostras de 2 e 10% empregaram-se números aleatórios, para evitar qualquer tendenciosidade. Eis os resultados da seleção:

2º Exp. - Seleção da amostra segundo pessoal ocupado

GRUPOS DE PESSOAL OCUPADO	TOTAL DE ESTABELECIMENTOS	TAMANHO DA AMOSTRA	
		Números absolutos	Números relativos
0 a 5 pessoas	13 325	255	1,91
6 a 20 pessoas	1 600	168	10,50
21 e mais pessoas	240	240	100,00
TOTAIS	15 165	663	4,37

Pelo quadro acima verifica-se que a amostra selecionada somou 663 questionários, havendo um acréscimo de 76 questionários em relação ao primeiro experimento. Totalizaram-se o pessoal ocupado e demais dados dos questionários obtidos na amostra de 2%, passando-se aos da amostra de 10%, e finalmente aos dos questionários de 21 pessoas e mais. Os resultados constam do quadro seguinte:

2º Exp. - Apuração dos questionários selecionados

ESPECIFICAÇÃO	ESTABELECIMENTOS	PESSOAL OCUPADO	Cr$ 1 000			
			Salários	Despesas Diversas	Vendas	Estoques
Amostra de 2%	255	545	2 499	9 016	59 195	15 917
Amostra de 10%	168	1 566	18 574	40 506	343 180	90 031
Grupo de 21 pessoas e mais ..	240	13 817	269 500	383 839	3 029 305	735 091

Obtiveram-se as estimativas simples sem tendenciosidade multiplicando os dados da amostra de 2% por 50, e os da amostra de 10% por 10. Totalizaram-se os produtos, por item pesquisado, incluindo os totais dos dados do grupo de 21 pessoas e mais. O quadro abaixo mostra as estimativas finais, comparadas com o valor real (Censo), como também o êrro de amostragem relativo, por item.

2º Exp. - Resultados da amostra

ITEM PESQUISADO	VALOR REAL (Censo) -a-	ESTIMATIVA (amostra) -b-	DIFERENÇA (b-a)		ÊRRO DE AMOSTRAGEM RELATIVO (em desvio padrão)
			Absoluta	%	
Estabelecimentos	15 165	14 670	- 495	3,3	3,1
Pessoal ocupado	56 910	56 727	- 183	0,3	3,1

2º Exp. - Resultados da Amostra (conclusão)

ITEM PESQUISADO	VALOR REAL (Censo) -a-	ESTIMATIVA (amostra) -b-	DIFERENÇA (b-a)		ÊRRO DE AMOSTRAGEM RELATIVO (um desvio padrão)
			Absoluta	%	
Salários (Cr$ 1 000)	573 809	580 190	+ 6 381	1,1	3,9
Despesas diversas (Cr$ 1 000)	1 251 605	1 239 699	- 11 906	1,0	4,6
Vendas (Cr$ 1 000)	9 745 937	9 420 855	- 325 082	3,3	4,8
Estoques (Cr$ 1 000)	2 566 610	2 431 251	- 135 359	5,3	5,8

Conclusões

Os resultados obtidos pelo primeiro experimento estão coerentes com a teoria que prevê a probabilidade, de 2 em 3, de que a estimativa baseada numa amostra do tipo usado não apresente uma variação de mais de um desvio-padrão do valor verdadeiro, e a probabilidade de 19 em 20, de que a estimativa não varie mais do que o dôbro do êrro de amostragem correspondente. Com efeito, neste experimento verificamos que os itens estabelecimentos (número), pessoal ocupado e despesas diversas apresentaram diferença não superior a um desvio-padrão, e os itens salários, vendas e estoques, diferença inferior a dois desvios-padrão.

No segundo experimento, para cinco dos seis itens investigados as estimativas estão dentro de um desvio-padrão do valor real, e o sexto item (número de estabelecimentos) apresenta uma variação de menos de dois desvios-padrão.

* * * * *

Convém esclarecer que estas amostras forma selecionadas com fins exclusivamente experimentais. Não se deve concluir, portanto, que a aplicação do processo visando à divulgação de dados preliminares de futuros censos econômicos leve a resultados semelhantes, no que respeita aos erros de amostragem. Na verdade, êsses erros podem ser reduzidos ao nível desejado, dentro de cada item investigado. Para isto, seria viável, por exemplo, aumentar o tamanho da amostra. Nesse caso, porém, havia de levar-se em conta a maior duração das apurações (que necessàriamente ocorreria), bem como o acréscimo de despesas.

Os elementos obtidos pelos dois experimentos descritos indicam, todavia, ser possível obter estimativas com maior grau de exatidão, e portanto diminuir os erros de amostragem, modificando as bases da estratificação, de modo a torná- la mais eficientes. Os acréscimos de tempo e de despesa resultantes seriam desprezí-

veis, ao passo que a redução dos erros de amostragem poderia ser bastante satisfatória.

Estudos preparatórios adicionais estão sendo levados a efeito, tendo em vista o seguinte:

1 - Determinar o número máximo de itens específicos que poderão ser apresentados por Unidade da Federação, ou Região;

2 - Estudar a aplicação de estimativas de razão, para concorrer na diminuição dos erros de amostragem;

3 - Fixar o item básico de estratificação mais conveniente: - valor das vendas ou pessoal ocupado -, e novo escalonamento para os estratos;

4 - Estudar planos de amostragem que se beneficiem de um entrosamento pré-estabelecido com a coleta, podendo-se obter com isso redução de tempo e de despesa na operação de amostragem.

ESTATÍSTICAS DE MÃO-DE-OBRA

Conceitos e métodos dos levantamentos periódicos realizados pelo "Bureau of the Census" dos EE. UU.

Introdução - Na primeira palestra desta série prometeu-se que seria incluído (ver Página 4) um estudo do Inquérito Periódico de População do Bureau do Censo dos Estados Unidos. Embora o Inquérito Periódico de População não constitua uma aplicação da amostragem ao trabalho censitário, vale a pena estudá-lo como exemplo de algumas das mais recentes e mais refinadas aplicações da teoria da amostragem ao trabalho prático de inquéritos. Além disso, oferece uma excelente ilustração de como o material e os resultados do censo podem fornecer uma sólida base para inquéritos por amostra eficientes.

Para a descrição dêste inquérito, escolhemos a tradução de alguns excertos da publicação oficial do Bureau do Censo sôbre os conceitos e métodos empregados(1). Êste relatório foi publicado em julho de 1954. A única alteração importante no planejamento e processos do inquérito desde aquela época foi a expansão da amostra para cêrca de 35.000 famílias em 330 áreas, em 1956.

Esta expansão reduziu os erros de amostragem e permite a publicação regular de dados com base regional.

Os principais aspectos do planejamento da amostra e dos métodos de inquérito que devem ser observados são:

1. A formação e estratificação das unidades primárias de amostragem, consistindo de condados isolados ou grupos de condados.

2. A seleção, com probabilidade proporcional ao tamanho, de uma unidade primária de amostragem de cada estrato.

3. O uso extensivo da amostragem de área a fim de selecionar setores dentro das unidades primárias de amostragem escolhidas.

4. A maneira pela qual os dados do censo de 1950 e as estimativas periódicas independentes da população são usados a fim de melhorar as estimativas baseadas na amostra.

5. O sistema de substituição gradual das famílias de amostra a fim de evitar sobrecarregar desnecessàriamente os informantes, reduzindo ao mesmo tempo os erros de amostragem das estimativas de mudança.

6. A rapidez com que os dados de um determinado período são coligidos, tabulados e liberados ao público.

---

(1) Bureau do Censo dos E.E.U.U., "Concepts and Methods Used in the Current Labor Force Statistics Prepared by the Bureau of the Census", Current Population Reports, Séries P-23, Nº 2, July 30, 1954.

"CONCEITOS E MÉTODOS EMPREGADOS NAS ESTATÍSTICAS PERIÓDICAS DA MÃO-DE-OBRA PREPARADAS PELO BUREAU DO CENSO"

As informações periódicas sôbre emprêgo, desemprêgo e dados correlatos são apuradas todos os meses através do Inquérito Periódico de População do Bureau do Censo. Êste inquérito é realizado todos os meses com uma amostra cientificamente selecionada representando a população civil (não-institucional).

Todo mês publicam-se estimativas da condição de emprêgo da população de 14 anos e mais em "The Monthly Report on the Labor Force", Current Population Reports, Series P-57. As estimativas da mão-de-obra, emprêgo agrícola e não-agrícola, desemprêgo e pessoas fora da mão-de-obra são apresentadas por idade e sexo. Para as pessoas empregadas publicam-se em separado as estimativas daquelas que estão trabalhando efetivamente nos seus empregos durante a semana do inquérito e as que possuem empregos mas não estão trabalhando. Apresenta-se uma distribuição de horas trabalhadas durante a semana do inquérito para aquelas que estão trabalhando. Aquelas que possuem empregos mas não estão trabalhando são classificadas de acôrdo com o motivo por que não estão trabalhando. Para aquelas que estão procurando emprêgo apresenta-se a duração do desemprêgo. A publicação regular mensal é geralmente divulgada cêrca de três semanas após a coleta dos dados. O inquérito regular sôbre a mão-de-obra é suplementado com um programa de inquéritos adicionais, coordenados com os levantamentos mensais, que se destinam a fornecer estatísticas mais detalhadas sôbre problemas especiais a fim de ajudar a qualificar e interpretar os totais gerais publicados todos os mêses. Os resultados dêstes estudos geralmente são publicados em Current Population Reports, Series P-50. Por exemplo, devido à utilidade de tais dados em diagnosticar as tendências econômicas, obtêm-se atualmente informações trimestrais sôbre as características dos trabalhadores em horário parcial. Dêstes estudos, publicam-se estimativas sôbre o número e as características dos que trabalham em horário parcial devido à diminuição de trabalho, mobilidade do trabalhador, insuficiência de empregos de horário integral e outros fatores econômicos, bem como os que trabalham em horário parcial por vontade própria ou por várias razões pessoais. Outro item importante fornecido por êstes estudos trimestrais é o número de pessoas desempregadas que procuram trabalho em horário parcial, ao invés de em horário integral.

Usam-se outros inquéritos suplementares para medir o número de trabalhadores que têm mais de um emprêgo e em mais de um ramo de atividade; a renda anual de pessoas e famílias; o tamanho do grupo que trabalha em qualquer momento durante o período de um ano; os ganhos e horas trabalhadas de trabalhadores agrícolas assalariados durante o ano; e fenômenos econômicos semelhantes. Através dêstes inquéritos suplementares, bem como de detalhes fornecidos nas estatísticas mensais, é possivel reagrupar várias categorias de pessoas de maneiras diferentes a fim de fornecer, por exemplo, medidas de desemprêgo ou emprêgo de acôrdo com definições alternativas.

O planejamento do inquérito

A amostra do Inquérito Periódico de População estende-se por 230 áreas

de amostra abrangendo 453 condados e cidades independentes. Um total de 24.000 a 26.000 unidades domiciliares e outros alojamentos de moradia são selecionados para a amostra em qualquer momento, obtendo-se também entrevistas completas, mensalmente, acêrca de 20.000 a 22.000 famílias. Do restante, cêrca de 500 a 1.000 são famílias das quais se devem coligir informações, porém isto não é feito porque os ocupantes não são encontrados em casa após visitas repetidas, estão temporàriamente ausentes ou não são encontradas por outras razões. As outras 2.500 a 3.500 unidades selecionadas representam as encontradas vazias, ocupadas por pessoas com residências em outra parte ou que de alguma outra forma não devem ser recenseadas. O tamanho da amostra global varia através do tempo, em parte devido ao acaso mas também devido ao crescimento da população e à criação de novas famílias. Em cada dois a três anos, à medida que a amostra se expande com o crescimento da população, é necessário diminuir ligeiramente a razão de amostragem a fim de manter a carga de trabalho na média de aproximadamente 25.000 unidades selecionadas prescritas no orçamento do inquérito.

Seleção das áreas de amostra - Tôda a área dos Estados Unidos, que consiste aproximadamente de 3.000 condados, foi dividida em cêrca de 2.000 unidades primárias de amostragem. Com algumas pequenas exceções, uma unidade primária de amostragem (PSU) consiste de um condado ou alguns condados contíguos. Cada área metropolitana padronizada (SMA) constituiu um PSU separado. Ao combinar condados a fim de formar PSUs, cada PSU foi definido de maneira a ser tão heterogêneo quanto possível. Poder-se-ia conseguir maior heterogeneidade incluindo mais condados. Entretanto, outra consideração importante foi tornar o PSU suficientemente compacto em área, de maneira que uma pequena amostra abrangendo todo êle pudesse ser eficientemente recenseada sem despesas indevidas de viagem. Uma unidade primária típica, por exemplo, incluiu tanto os residentes urbanos como os rurais, ou tanto os níveis econômicos elevados como os baixos, e forneceu, tanto quanto possível, diferentes ocupações e ramos de atividade.

Os PSU foram então grupados em 230 estratos. As 44 maiores áreas metropolitanas padronizadas e certas outras SMA eram estratos por si mesmas. Contudo, de um modo geral, um estrato consistiu de uma série de PSUs tão parecidos quanto possível em várias características tais como região geográfica, densidade de população, taxa de crescimento na década 1940-1950, percentagem de não brancos, principal ramo de atividade, tipo de agricultura, etc. Êstes são os mesmos tipos de critérios usados na estratificação do antigo planejamento de 68 áreas, instituído em 1943 mas interrompido recentemente, embora de certo se usassem informações mais recentes. Exceto as 44 maiores SMA e as 16 outras áreas, cada uma das quais é um estrato completo, os estratos foram estabelecidos de maneira que seus tamanhos em têrmos da população de 1950 fôssem aproximadamente iguais. Quando um PSU era um estrato por si mesmo, caía automàticamente na amostra. De cada um dos outros estratos, selecionou-se um PSU aleatòriamente para inclusão na amostra, tendo-se feito a seleção de tal maneira que a probabilidade da seleção de qualquer unidade fôsse proporcional à sua população de 1950.

Por exemplo, dentro de um estrato a probabilidade de que um PSU com uma população de 50.000 habitantes fôsse selecionado era duas vêzes a de uma unidade com população de 25.000.

As 230 áreas resultantes são aquelas nas quais se realiza o inquérito.

Seleção de famílias de amostra - Para cada estrato usa-se no presente momento (1954) uma razão de amostragem global de cêrca de 1 em 2.250. A razão de amostragem empregada em cada área particular de amostra (PSU de amostra) depende da proporção que a população da área de amostra (na data do Censo de 1950) representava da população do estrato. Assim, em uma área de amostra que era um décimo do estrato, a razão de amostragem dentro do PSU que resulta é 1 em 225, atingindo a razão desejada de 1 em 2.250.

Dentro de cada um dos 230 PSUs, usam-se métodos de amostragem de áreas na seleção de famílias específicas. Em cada PSU, o número de famílias a serem recenseadas mensalmente é determinado pela aplicação da razão de amostragem dentro do PSU em lugar da atribuição de uma quota fixa. Êste processo torna possível à amostra refletir quaisquer mudanças na população. Por exemplo, se com base no Censo de 1950 se usar uma razão de amostra de 1 em cada 225 em uma área de amostra, o número de famílias esperado na amostra será maior do que o obtido por uma quota fixa em áreas em que o número de famílias aumentou desde o censo. Em áreas em que o número de famílias diminuiu, o número esperado de famílias de amostra será menor. Desta forma, a amostra pròpriamente dita reflete a mudança da distribuição da população e evita a distorsão que resultaria da aplicação de quotas fixas de famílias, ou pessoas, baseadas na população em data mais recente.

Na aplicação dos métodos de amostragem de áreas, usaram-se vários estágios de amostragem dentro de cada PSU selecionado. Em primeiro lugar, uma amostra de unidades administrativas usada para os Censos de População e Habitação de 1950 (distritos censitários) foi selecionada, com a probabilidade de seleção de qualquer uma destas proporcional à sua população de 1950. Êsses distritos censitários selecionados foram então subdivididos em setores, isto é, pequenas áreas de terra com limites bem definidos tendo em geral um "tamanho" esperado de cêrca de seis unidades domiciliares ou outros alojamentos de moradia. Quando fôssem insuficientes as estradas, rios, e outras características de terreno que poderiam ser usadas para subdividir um distrito censitário, alguns dos setores resultantes eram várias vêzes o "tamanho" médio desejado de seis famílias. Para cada distrito censitário subdividido, designou-se um setor para a amostra, com a probabilidade de seleção proporcional ao "tamanho" estimado do setor. Quando as informações preliminares disponíveis indicavam que o setor continha cêrca de seis famílias, tôdas as unidades dentro dos limites do setor deviam ser incluídas na amostra. Nos casos em que as informações preliminares indicavam um "tamanho" de setor de várias vêzes seis unidades, devia ser feita uma listagem de campo de todos os alojamentos de moradia do setor e selecionada uma sub-amostra sistemática de maneira a conseguir o equivalente de um setor que é recenseado completamente.

Ao subdividir os distritos censitários em setores e ao determinar com antecedência o "tamanho" aproximado de cada setor, fêz-se uso de vários materiais. Nas maiores localidades urbanas, as informações concernentes ao número de unidades de cada quarteirão foram obtidas dos boletins das Estatísticas de Quarteirões publicados dos resultados dos Censos de População e Habitação de 1950 para 209 das cidades de 50.000 habitantes ou mais. Em conjunto com êstes boletins, fêz-se uso considerável de mapas Sanborn de grande escala, que existem à venda no comércio e que são relativamente atualizados para a maioria dos centros urbanos de tamanho médio e grande, e mostram o esbôço geral de cada estrutura dentro dos quarteirões. Quando êstes dados não estavam disponíveis, a localização e o número de unidades domiciliares de pequenas áreas geográficas ligadas por estradas, rios, etc., foram obtidos dos mapas usados pelos recenseadores nos Censos de População e Habitação de 1950 ou de visitas especiais de campo. Os distritos censitários dos centros urbanos - cujos materiais cartográficos geralmente eram mais precisos - foram mais fàcilmente subdivididos em setores compactos (com uma média de seis unidades) do que os das áreas rurais, mas uma proporção considerável dos setores rurais resultantes eram dêsse tamanho também. Surgiram algumas variações no tamanho efetivo dos setores quando os materiais cartográficos, embora suficientemente detalhados, eram desatualizados devido a um número considerável de novas construções ou porque continham erros.

Rotação da amostra - Parte da amostra é modificada todos os meses. Uma razão primordial para alternar a amostra é evitar os problemas de falta de cooperação que ocorrem quando se entrevista uma lista constante de pessoas, indefinidamente. A fim de realizar esta rotação da amostra em base gradual, os materiais cartográficos e outros para várias amostras são preparados simultâneamente. Para cada amostra, identificam-se oito sub-amostras sistemáticas (grupos de rotação) de setores. Um determinado grupo de rotação é entrevistado por um total de oito meses, divididos em dois períodos iguais. Permanece na amostra por quatro meses consecutivos em um ano, deixa a amostra durante os oito mêses seguintes, e depois retorna para os mesmos quatro meses civis do ano seguinte. Em qualquer mês, um oitavo dos setores de amostra estão no seu primeiro mês de coleta, outro oitavo em seu segundo mês, etc., com o último oitavo pela oitava vez na amostra (o quarto mês do segundo período de coleta). Por êste sistema, 75 % dos setores de amostra são comuns de mês para mês e 50 % de ano para ano. Êste processo fornece um grau considerável de superposição de mês para mês, e de ano para ano na lista de pessoas entrevistadas (reduzindo assim as descontinuidades na série de dados) sem sobrecarregar nenhum grupo específico de famílias com um período indevidamente longo de inquérito.

Técnicas de inquérito - O quadro de supervisores imediatos de campo consiste de aproximadamente 60 Supervisores e Assistentes de Supervisores de Distrito, localizados em 34 centros. Durante o período de coleta do Inquérito Periódico de População (CPS), todos os meses, dedicam êles a maior parte de seu tempo ao con

trôle e supervisão do inquérito, embora possuam várias outras tarefas durante o resto do mês. Supervisionam, ao todo, um quadro de cêrca de 350 entrevistadores em horário parcial.

Cada mês, durante a semana civil que contém o dia 15, êstes entrevistadores entram em contato com uma pessoa de responsabilidade nas famílias de amostra no Inquérito Periódico de População. Na data da primeira visita a uma família, o entrevistador prepara um cadastro dos membros da família, inclusive suas características pessoais (data de nascimento, sexo, raça, estado civil e condição como veterano) e sua relação com o chefe da família. Êste cadastro é atualizado em cada entrevista subseqüente a fim de tomar conhecimento de novos residentes ou residentes que se retiraram, mudanças no estado civil e itens semelhantes. As informações sôbre características pessoais são então obtidas todos os meses para fins de identificação e para classificações cruzadas com as características econômicas da população de amostra.

Em cada visita mensal, o entrevistador faz uma série de perguntas padronizadas sôbre a atividade econômica durante a semana precedente (contendo a semana civil o oitavo dia do mês, chamada a "semana do inquérito") de cada membro da família de 14 anos e mais. A finalidade principal destas perguntas é classificar a população de amostra em três grupos econômicos básicos - os empregados, os desempregados e aquêles que não estão na mão-de-obra.

Fazem-se perguntas adicionais todos os meses para suplementar os dados básicos. Para os empregados, obtêm-se informações sôbre horas trabalhadas durante a semana de inquérito, juntamente com uma descrição do emprêgo atual, e, para aquêles temporàriamente ausentes do emprêgo, o motivo por que não trabalharam na semana do inquérito. Para os desempregados, obtêm-se informações sôbre a duração da fase em que estavam procurando emprêgo e uma descrição do seu último emprêgo. Para aquêles que estão fora da mão-de-obra, registra-se sua atividade principal durante a semana de inquérito - dona de casa, estudante ou outra atividade.

Os questionários usados no inquérito são de uma forma especial conhecida como questionários de "document-sensing". Em lugar de anotar as informações, o entrevistador, para a maioria dos itens, marca um sinal em um círculo que representa a resposta correta, usando um tipo especial de lapis. Os formulários preparados desta forma podem ser convertidos em cartões de perfuração por uma máquina especial de leitura, evitando assim a preparação manual para o cartão de perfuração. O processo também reduz a codificação das respostas a um mínimo, visto como a posição de cada círculo no formulário representa por sí só um código.

PROCESSO DE ESTIMAÇÃO

Os questionários de leitura mecânica (formulários) contendo as informações obtidas para cada pessoa da amostra são recebidos na repartição de Washington na semana seguinte à coleta. Os dados em bruto são convertidos em cartões de perfu-

ração por meio de uma reprodutora mecânica. Podem-se preparar estimativas tabulando êstes cartões com um pêso fixo (o inverso da razão de amostragem - aproximadamente 2.250 no presente) após descontar as famílias que não foram entrevistadas. Entretanto, a fim de aumentar a fidedignidade das estatísticas da mão-de-obra derivadas da amostra, usam-se dois estágios de estimativas de razão e uma "estimativa composita". É possível conseguir êste processo um tanto complicado rápida e automàticamente devido à disponibilidade da UNIVAC, uma computadora eletrônica de dígitos, de alta velocidade. Os principais passos a tomar são os seguintes:

Ajustamento para as famílias não entrevistadas - Os pesos para tôdas as famílias entrevistadas são ajustados até o grau necessário a fim de descontar os domicílios ocupados que não se pôde entrevistar devido à ausência dos moradores, estradas intransitáveis, recusas ou por outras razões. Êsse ajustamento é feito em separado para certos grupos de PSUs e, dentro dêstes, segundo a côr (brancos, não-brancos) - grupo de domicílios por residência (urbana, rural não agrícola, rural agrícola). A proporção de domicílios de amostra não entrevistados geralmente é de cêrca de 3 a 5%.

Estimativas de razão - A distribuição da população selecionada para a amostra pode diferir um pouco da nacional no que respeita a características básicas como idade, côr, sexo e residência agrícola ou não agrícola, entre outras coisas. Estas características particulares da população estão intimamente relacionadas com a participação da mão-de-obra e outras medidas principais feitas através da amostra. Portanto, algumas das estimativas de amostra podem ser consideràvelmente melhoradas quando, pela ponderação apropriada dos resultados originais, a população de amostra é levada a concordar o mais perfeitamente possível com a população total, no que respeita à distribuição, das referidas características. Tal ponderação é realizada através de dois estágios de estimativas de razão, da forma seguinte:

1. Primeiro estágio - O primeiro estágio das estimativas de razão leva em conta as diferenças na data do último censo na distribuição por côr e residência da população estimada dos PSUs da amostra e a da população total de cada uma das quatro principais regiões do país. As distribuições independentes da população total por residência, classificadas em combinação com a côr, não são disponíveis em base periódica. Em seu lugar, usando os dados censitários de 1950, criaram-se razões da população total estimada por côr e residência em uma determinada região, baseadas em PSUs de amostra em relação à população total correspondente da região. Tal estimativa de razão não implica que a razão existente em 1950 fique inalterada em uma data corrente. As estimativas dos PSUs de amostra basearam-se nas contagens censitárias totais, não em contagens de inquéritos por amostra. Ao deduzir estas razões, os PSUs autorepresentados foram excluídos dos cômputos, visto como representam sòmente a si mesmos na amostra do Inquérito Periódico de População. Nas tabulações dos resultados

mensais do Inquérito Periódico de População, os pesos para tôdas as famílias de amostra dos PSUs que não são autorepresentadas em uma determinada região são multiplicados pela razão da população para aquela região em relação à classe apropriada de côr e residência.

2. Segundo estágio - O segundo estágio das estimativas de razão leva em conta as diferenças periódicas entre a distribuição observada na amostra e a da população nacional, por idade, côr e sexo. Todos os meses são preparadas estimativas independentes de tôda a população, segundo estas características, as quais são calculadas atualizando-se os dados censitários mais recentes (1950), levando em conta o envelhecimento da população, a mortalidade e a migração entre os Estados Unidos e outros países*. Os resultados da amostra do Inquérito Periódico de População (considerando os pesos determinados após o primeiro estágio das estimativas de razão) são, com efeito, usados para determinar apenas a distribuição percentual dentro de um determinado grupo de idade, côr e sexo, por condição quanto a emprêgo e várias outras características. Ao desenvolver as estatísticas em números absolutos, estas distribuições percentuais são multiplicadas pela estimativa independente da população para o grupo adequado de idade, côr e sexo.

Estimativa compósita - O último estágio na preparação de estimativas faz uso de uma estimativa compósita. Por êsse processo, obtém-se uma média ponderada de duas estimativas do mês corrente para qualquer item específico. A primeira estimativa é o resultado dos dois estágios de estimativas de razão mencionados anteriormente. A segundo consiste da estimativa compósita para o mês precedente, à qual foi adicionada uma estimativa da diferença verificada em cada item, entre o mês precedente e o mês presente, baseando-se na parte da amostra comum aos dois meses (75%). Embora os pesos dos dois componentes dessa estimativa compósita não sejam necessàriamente iguais, no caso presente foram 1/2 cada um. Pesos iguais, no caso descrito, aumentam a fidedignidade para quase todos os itens, no processo de estimação após os dois primeiros estágios de estimativas de razão.

Para a maioria das estatísticas importantes do inquérito, esta estimativa compósita conduz a uma redução adicional no êrro de amostragem, quando superposta aos dois estágios das estimativas de razão descritas anteriormente; sendo que para alguns itens a redução é substancial. As vantagens obtidas quanto à fidedignidade são maiores nas estimativas de variação mensal, embora também se obtenham vantagens nas estimativas de determinado mês ou nas de variação anual ou de outros intervalos de tempo.

Fontes de erros nas estimativas do inquérito - As estimativas do inquérito são sujeitas a erros de amostragem, isto é, os decorrentes do fato de as estimativas mensais basearem-se em informações de uma amos-

---

*Veja U.S. Bureau of the Census, Current Population Reports, Series P-25, Nº 93, april 26, 1954 para uma descrição do método usado na preparação destas estimativas independentes de população.

tra relativamente pequena e não de tôdas as pessoas do universo. Além disso, como em qualquer trabalho de pesquisa, os resultados estão sujeitos a erros de coleta e de apuração.

Os erros de classificação nos inquéritos da mão-de-obra podem ser particularmente grandes no caso de pessoas com participação ocasional na mão-de-obra. Êstes erros podem ser causados pelos entrevistadores, informantes, ou por ambos, ou ainda, por falhas no planejamento do questionário. Os entrevistadores do Inquérito Periódico de População trabalham principalmente em horário parcial. São mais bem treinados do que a maioria do pessoal de campo, com experiências repetidas neste inquérito e treinamento direto ou por via postal. Além disso, através da crítica dos questionários, preenchidos, uma observação constante durante a coleta e uma reverificação sistemática de parte de suas tarefas pelo pessoal de supervisão de campo, o trabalho dos entrevistadores é mantido sob contrôle satisfatório e seus erros ou deficiências trazidos diretamente à sua atenção.

Apesar dêsse contrôle, os entrevistadores podem às vêzes deixar de fazer as perguntas na forma prescrita. Como a variação na formulação das perguntas resulta em diferenças de declaração, êste fator pode determinar alguns erros ou falta de uniformidade nas estatísticas.

De maneira semelhante, os dados ficam limitados pelo conhecimento adequado da informação por parte do informante e pela sua boa vontade em responder com exatidão. Habitualmente uma única pessoa, em geral a dona de casa, informa pela família tôda. O informante pode não conhecer todos os fatos sôbre os demais membros da família ou pode estar possibilitado de informar satisfatòriamente a respeito de suas atitudes ou intenções. Por exemplo, a dona de casa provàvelmente saberá que seu marido está trabalhando, mas nem sempre saberá exatamente quantas horas êle trabalhou ou a natureza precisa do seu trabalho.

As estimativas do inquérito estão sujeitas a vários outros tipos de erros além dos já mencionados. Alguns dêles são:

1. Omissão de declarações - Para cêrca de 3 a 5% das unidades da amostra não são obtidas informações em determinados mêses devido a ausência temporária dos ocupantes ou por várias outras razões. Embora se faça um ajustamento nos pesos para as famílias entrevistadas, a fim de descontar as que não o foram, estas ainda representam uma possível fonte de tendenciosidade. De maneira semelhante, são omitidas algumas informações, para um número relativamente pequeno de famílias entrevistadas, porque o declarante as ignora ou porque o entrevistador deixa de fazer certas perguntas ou registrar respostas. Suprem-se geralmente essas omissões, durante o processamento dos questionários, com base nas distribuições obtidas para as pessoas de características semelhantes.

2. Estimativas independentes de população - As estimativas independentes de população usadas no processo de estimação (ver "estimativas de razão", pág. 6) também podem constituir uma fonte de êrro, embora o resultado final de seu uso consti

tua uma melhoria substancial da fidedignidade estatística de muitas das cifras importantes. Podem ocorrer erros nas estimativas independentes de população devido à omissão de certos grupos da população ou a erros na declaração da idade no último censo (que serve de base para as estimativas) ou a problemas semelhantes nos fatôres de variação da população (mortalidade, imigração, etc.), desde aquela data.

3. Erros de processamento - Embora haja um programa de contrôle de qualidade da codificação e um contrôle rigoroso de tôdas as outras fases de processamento e tabulação dos resultados, alguns erros são quase inevitáveis em uma operação estatística considerável dêste tipo. É provável, entretanto, que o êrro líquido que ocorre de processamento seja bastante desprezível.

Medida da exatidão dos resultados - A moderna teoria da amostragem fornece métodos para medir a extensão dos erros devidos à amostragem, em que a probabilidade de seleção de cada membro da população é conhecida, como no caso da amostra do Inquérito Periódico de População. Existem também métodos para medir o efeito da variabilidade das declarações no Inquérito Periódico de População. Uma medida de variabilidade da amostragem indica a extensão da diferença que pode ser esperada quando só é pesquisada uma amostra da população. Uma medida de variabilidade das declarações, por outro lado, indica a extensão da diferença que pode ser esperada devido a tipos compensadores de erros provenientes do procedimento de diferentes entrevistadores e das respostas dos informantes; êsses erros tenderiam a anular-se no levantamento de uma população bastante grande. Na prática, estas duas fontes de êrro-amostragem e variabilidade das declarações, conforme foram definidas acima - são estimadas conjuntamente através dos resultados do inquérito. Os cômputos, entretanto, não levam em conta o efeito da tendenciosidade das declarações, isto é, quaisquer erros sistemáticos de declaração, tais como os que ocorreriam se, em sua maior parte, os informantes tendessem a exagerar o número de horas trabalhadas. As tendenciosidades das declarações ocorrem da mesma maneira, tanto em um censo completo como em uma amostra e, de fato, podem ser menores em um inquérito por amostra bem realizado porque aí é possível pagar o prêço necessário a fim de coligir as informações com maior precisão.

Estimativas da variabilidade total devido à amostragem e a variação nas respostas são fornecidas em "The Monthly Report on the Labor Force" e em outras publicações baseadas no Inquérito Periódico de População, e a interpretação dos dados no texto destas publicações é feita à luz da possível variabilidade das cifras. Em geral, cifras pequenas e pequenas diferenças entre cifras estão sujeitas a uma variação relativamente grande e devem ser interpretadas com cuidado. A existência da computadora eletrônica de alta velocidade permitirá fornecer muito mais detalhes sôbre êste assunto do que foi possível até agora.

A medida da tendenciosidade das declarações é um dos aspectos mais difíceis nos inquéritos e censos. Os estudos sistemáticos sôbre êste assunto são agora parte integrante do Inquérito Periódico de População, mas em muitos casos as técnicas

disponíveis não são suficientemente precisas para fornecer estimativas satisfatórias dos erros provenientes de tendenciosidades nas declarações. Acham-se em progresso muitas experimentações com a finalidade de desenvolver mensurações mais precisas e de usar as informações de maneira a melhorar a exatidão geral do inquérito.

VOCABULÁRIO
DE
AMOSTRAGEM

(Limitado aos assuntos estudados neste Documento)

VOCABULÁRIO DE AMOSTRAGEM

Introdução - É lamentável que na ocasião em que estas palestras foram pronunciadas, o excelente trabalho intitulado Vocabulário Brasileiro de Estatística, por Milton da Silva Rodrigues, ainda não estivesse publicado. Se estivesse disponível, ter-se-ia feito uma tentativa conscienciosa a fim de adaptar-se às definições fornecidas pelo Professor Rodrigues. Naturalmente, o glossário de têrmos que se segue deve dar as definições dos têrmos efetivamente usados nas palestras. Entretanto, a fim de tornar possível uma "tradução" em têrmos dados no Vocabulário, fêz-se o seguinte:

1. As definições do Vocabulário do Professor Rodrigues são usadas diretamente, sempre que possível. Em tais casos, o têrmo é precedido por um asterisco(*).

2. Nos casos em que um dos têrmos usados nas palestras é explicado pela definição do Prof. Rodrigues para algum outro têrmo, o têrmo original é precedido por um asterisco e o têrmo do Prof. Rodrigues é dado entre parênteses, precedido pelas letras VBE, por exemplo:

* Números aleatórios (VBE Números Equiprováveis)

Quando não se usar a definição do Vocabulário do Prof. Rodrigues, será geralmente porque o têrmo não foi definido pelo Prof. Rodrigues ou porque o autor preferiu usar uma definição menos técnica ou não matemática, ou adicionar algumas observações especiais explanatórias.

Uma secção especial no fim dêste vocabulário é dedicada ao exame de têrmos usados a fim de indicar relações gerais entre estimativas de amostra e valores do universo; têrmos como precisão, êrro, fidedignidade, etc.

O Vocabulário do Prof. Rodrigues emprega o têrmo probalística em lugar de probabilística. Êste último têrmo é usado neste vocabulário simplesmente porque foi usado nas palestras, e não por antipatia ao primeiro.

Secção A - Vocabulário

* ACASO - A noção de acaso, bem como as que dela se derivam (aleatório, ao acaso, etc.) é primitiva. Uma fidelidade estrita ao sentido clássico do determinismo científico levar-nos-á a defini-lo como sendo um complexo de numerosíssimas causas cujas atuações individuais desconhecemos.

* ALEATÓRIO - devido ao acaso (q.v.). Também se diz casual e acidental.

* AMOSTRA - É todo conjunto cujas propriedades se estudam com o fim de generalizá-las a outro conjunto de que os elementos daquele são considerados provenientes.

AMOSTRA DE CONVENIÊNCIA - É um tipo de amostra não probabilística (q.v.)

AMOSTRA DUPLA - Amostra seqüencial (q.v.) consistindo de apenas duas sub-amostras.

* AMOSTRA ESTRATIFICADA - Reunião, Sn, das m amostras independentes Sn_1, Sn_2, ..., Sn_m (onde n_i denota tamanho e $\sum_i n_i = N$) tais que S_{n_i} (i= 1, 2, 3, ...,m) é oriunda do i-ésimo estrato de uma dada população.

AMOSTRA INTENCIONAL - É um tipo de amostra não probabilística (q.v.)

AMOSTRA NÃO PROBABILÍSTICA - É qualquer amostra selecionada de tal maneira que um ou mais membros do universo tenha probabilidade nula ou desconhecida de seleção. V. AMOSTRAGEM A ESMO

AMOSTRA PROBABILÍSTICA - É uma amostra selecionada de tal maneira que cada elemento do universo tenha uma probabilidade conhecida, diferente de zero, de ser selecionada.

AMOSTRA PROPOSITAL - É um tipo de amostra não probabilística (q.v.)

AMOSTRA SEQÜENCIAL - É uma série de sub-amostras independentes, selecionadas de tal maneira que os resultados de dada sub-amostra são conhecidos antes que a sub-amostra seguinte seja selecionada. As informações das primeiras m sub-amostras podem ser usadas a fim de determinar a maneira pela qual a $(m + i)^{\text{ésima}}$ amostra é selecionada.

* AMOSTRA, TAMANHO DA - é o número de elementos que a compõem.

* AMOSTRAGEM A ESMO - é o processo de seleção de amostra que, não sendo sujeito a uma disciplina estrita, fica entregue ao critério do pesquisador; êste lança mão de diversos recursos de bom-senso, nem sempre isentos de equação pessoal, a fim de obter uma amostra cuja definição se aproxime da de amostra acidental I. (Embora esta expressão não tenha sido usada nas palestras, sua definição descreve muito exatamente a maneira pela qual a maioria das amostras não probabilísticas são selecionadas).

* AMOSTRAGEM ALEATÓRIA SIMPLES (VBE Amostragem equiprobabilística) - É aquela em que, sendo N_{ij} o número de elementos de um conjunto C_j no momento da i-ésima estração (i = 1,2, ..., n_j, sendo $N_{ij} = N_j$ o tamanho inicial de C_j), a probabilidade com que um dêles é escolhido para a amostra, sendo a mesma para todos os N_{ij} elementos, é igual a $1/N_{ij}$. É o mesmo que, para certos autôres, amostragem acidental; para outros, amostragem simples. O conjunto C_j tanto pode ser apenas uma das partes de uma população, como pode ser idêntico a ela tôda.

* AMOSTRAGEM COM PROBABILIDADE PROPORCIONAL AO TAMANHO (VBE Amostragem proporcional ao tamanho) - É o processo de seleção probabilística em que as probabilidades associadas às diversas unidades amostrais da população são proporcionais aos tamanhos dessas unidades. Como tamanho das unidades amostrais podem-se adotar as intensidades do próprio atributo em causa, conhecidas por meio de um censo ou pesquisa anterior, ou, ainda, as intensidades de qualquer outro atributo altamente correlacionado com o mesmo.

* AMOSTRAGEM COM REPOSIÇÃO - processo de seleção de amostra em que cada

elemento da população que é escolhido para a amostra é, após sua observação, devolvido à população originária, antes de se fazer qualquer nova extração.

* AMOSTRAGEM DE ÁREAS - É aquela em que as unidades amostrais são áreas. Também se diz amostragem na base de áreas.

* AMOSTRAGEM POR CONGLOMERADOS - É aquela em que as unidades amostrais são conglomerados (q.v.).

* AMOSTRAGEM POR QUOTAS - Processo de amostragem no qual os trabalhadores de campo recebem tarefas específicas quanto ao número de unidades amostrais a serem escolhidas de cada estrato, mas a seleção, ela própria, é feita a esmo por êsses trabalhadores.

AMOSTRAGEM PROBABILÍSTICA - É um processo de seleção pelo qual se obtém uma amostra probabilística. Na maior parte das vêzes, êste processo envolve o uso de uma tabela de números aleatórios.

* AMOSTRAGEM SEM REPOSIÇÃO - Processo de seleção de amostra em que uma mesma unidade amostral não pode figurar mais de uma vez na amostra.

AMOSTRAGEM SISTEMÁTICA - É a amostra de um universo em que os elementos foram ordenados antes da seleção e os elementos da amostra selecionados de acôrdo com intervalos iguais, com início aleatório. Há também uma forma de amostra sistemática em que todos os elementos do universo são numerados e a amostra consiste daqueles com certos dígitos finais.

* AMOSTRAS INTERPENETRANTES - São amostras independentes, oriundas da mesma população e obtidas pelo mesmo processo de amostragem.

* AMPLITUDE TOTAL - De um conjunto de valores é o módulo da diferença entre o maior e o menor dêles.

ARRANJO - Um arranjo de m entre n itens ($m \leq n$) significa uma série de m entre n itens com composição e ordem especificadas. Se os n itens fôrem todos diferentes, há

$$\frac{n!}{(n-m)!} = n\,(n-1)\,\ldots\,(n-m+2)\,(n-m+1)$$

arranjos possíveis dos m itens. Os arranjos das letras a, b, c consideradas duas de cada vez são ab, ba, ac, ca, bc, cb.

Nota: não há equivalente para a palavra arranjo em inglês. O têrmo "permutation" é usado tanto para permutações como para arranjos.

ATRIBUTO - É uma característica cujas alternativas não são passíveis de sujeição a uma ordem antural. Exemplos: a nacionalidade, a religião, etc. (O Prof. Rodrigues define êstes como atributos não ordenáveis; o que nós chamamos de variáveis (q.v.) êle define como atributos ordenáveis).

* CARACTERÍSTICA (VBF Atributo) - Tudo aquilo que se diz ou é, próprio de um ser, podendo ser constante ou variável, qualitativo ou quantitativo.

CARACTERÍSTICA AUXILIAR - É uma característica cujos valores são conhecidos para todos os elementos do universo, usada a fim de aumentar a eficiência do processo de seleção ou estimação, como na amostragem estratificada, amostragem com probabilidade proporcional ao tamanho, estimativas de razão e de diferenças.

COEFICIENTE DE VARIAÇÃO (de uma estimativa) - É o desvio-padrão de uma estimativa dividido pelo seu valor esperado.

COEFICIENTE DE VARIAÇÃO (de uma característica do universo) - É o desvio-padrão da característica dividido por seu valor médio.

COMBINAÇÃO - Combinação de m entre n itens (m ≤ n) significa uma série de m dos n itens cuja composição é especificada, mas cuja ordem não é especificada. Se os n itens forem todos diferentes, há

$$\frac{n!}{(n-m)!\,m!} = \frac{n\,(n-1)\,\ldots\,(n-m+2)\,(n-m+1)}{1.2\,\ldots\ldots\ldots\,m}$$

combinações possíveis de m itens. As letras a, b, c combinadas tomadas duas de cada vez são ab, ac, bc. As combinações geralmente são denotadas pelo símbolo C_m^n ou $\binom{n}{m}$.

* CONGLOMERADO - É uma unidade de amostragem formada por unidades menores, em geral contíguas. Exemplo: a unidade "domicílio" é um conglomerado de unidades de "pessoas físicas".

CONSISTÊNCIA - Diz-se que uma estimativa tem a propriedade da consistência se a proporção de estimativas de amostra que diferem do valor que está sendo estimado em menos que uma pequena quantidade especificada se aproxima de 100% à medida que o tamanho da amostra aumenta.

CONTAGEM COMPLETA - Contagem na qual o valor da característica que está sendo estudada é determinado para todos os elementos do universo.

DESVIO-PADRÃO (de uma estimativa) - É a raiz quadrada do desvio quadrático médio da estimativa de seu valor esperado, isto é, a raiz quadrada da variância. (O Prof. Rodrigues refere-se ao desvio-padrão de uma estimativa como o êrro padrão).

DESVIO-PADRÃO (de uma característica do universo) - É a raiz quadrada do desvio quadrático médio da característica do seu valor médio, isto é, a raiz quadrada da variância. (O Prof. Rodrigues refere-se ao desvio-padrão de uma característica do universo como o afastamento padrão).

* DISTRIBUIÇÃO DE FREQÜÊNCIA - É a série estatística que se obtém, distribuindo os N individuos que compõem uma dada coletividade pelos diversos valores ou classes de valores, de um mesmo atributo que fornecerá a ordem de classificação. Exemplo:

N indivíduos distribuídos segundo os números daqueles que apresentam estaturas compreendidas entre tais e tais valores.

* DISTRIBUIÇÃO NORMAL (VBE Distribuição normal [unidimensional]) - É a da variável aleatória , cuja densidade de freqüência no ponto = é dada pela função de x

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(\xi-\alpha)^2}{2\sigma^2}} \qquad (-\infty < x < \infty)$$

onde $\alpha = E(\xi)$ e $\sigma^2 = E(\xi - \alpha)^2$.

ÊRRO DE AMOSTRAGEM - Em geral, é a diferença, resultante do uso da amostragem, entre uma estimativa e o valor que teria sido obtido por uma contagem completa, usando os mesmos processos. O têrmo é usado às vêzes a fim de referir-se especificamente ao desvio-padrão de uma estimativa. (O Prof. Rodrigues usa êrro amostral ou êrro de amostragem para referir-se à diferença entre uma determinada estimativa e o valor do universo).

* ESTIMATIVA (VBE Estimador) - De um parâmetro de dada população, é tôda função de elementos de amostra oriunda dessa população que mantém para com o parâmetro uma certa relação probabilística.

* ESTIMATIVA DE RAZÃO (VBE Estimativa-razão) - Resultado de um processo de estimação de um parâmetro Θ_y da população dos y que se baseia na observação dos y em uma amostra, na observação dos valores, que êsses mesmos indivíduos apresentam, de um outro atributo x e no conhecimento do valor populacional do correspondente parâmetro Θ_x da população dos x. A estimativa-razão $\hat{\alpha}_{yR}$, da média aritmética da população dos y, por exemplo, será

$$\hat{\alpha}_{yR} = \frac{\bar{y}}{\bar{x}}\,\alpha_x x$$

onde $\bar{x}$ denota a média aritmética dos x, em uma amostra; $\bar{y}$ a m.a dos y, na mesma amostra e α_x a m.a. populacional dos x.

* ESTIMATIVA DE REGRESSÃO (VBE Estimativa-regressão) - Resultado de um processo de estimação de um parâmetro Θ_y da população dos y que se baseia na equação de regressão de y sôbre outra variável x, numa amostra de portadores dessas duas variáveis, e no conhecimento do valor do correspondente parâmetro Θ_x na população dos x. Por exemplo, a estimativa-regressão, $\hat{\alpha}_{y\,L\,r}$, da média aritmética da população dos y, suposta linear a regressão de y sôbre x, será

$$\hat{\alpha}_{y\,L\,r} = \bar{y} + b\,(\alpha_x - \bar{x})$$

onde $\bar{y}$ denota a média aritmética dos y numa amostra; $\bar{x}$ a m.a. dos x, na mesma amostra; b o coeficiente de regressão de y sôbre x, nessa mesma amostra e α_x a m.a da população dos x.

ESTIMATIVA DIFERENCIAL - É uma estimativa de regressão com b = 1

ESTIMATIVA INDEPENDENTE - Com respeito a uma determinada investigação por amostragem, é uma estimativa de uma característica que não é baseada nos resultados de amostra, mas que é usada para verificar os resultados da amostra ou para melhorá-los fazendo estimativas de razão.

* ESTIMATIVA SEM TENDENCIOSIDADE (VBE Estimador não viesado) - É o estimador $\hat{\theta}$ do parâmetro θ tal que, sendo n o tamanho das amostras em que se baseia seu cálculo, a esperança matemática de $\hat{\theta}$ é igual a θ independentemente de n. Também se diz estimador não-viciado e estimador imparcial.

* ESTRATIFICAÇÃO - Processo, ou resultado, da decomposição de um conjunto de unidades amostrais em estratos, segundo as alternativas de um ou de mais atributos chamados contrôles.

* ESTRATO - Qualquer um dos subconjuntos de unidades amostrais em que a população é decomposta, antes de se fazer a seleção de uma amostra.

EXPANSÃO DE TOTAIS DA AMOSTRA - É o ato de fazer uma estimativa, usando os resultados de uma amostra. Para a estimativa simples sem tendenciosidade de um total, isto consiste em multiplicar os totais da amostra pelo inverso da fração de amostragem.

* EXPERIÊNCIA - É qualquer ato ou acontecimento cujos resultados podem ser observados ou medidos de alguma forma.

INTERVALO DE CONFIANÇA - Às vêzes, em lugar de usar uma amostra a fim de calcular uma única estimativa de uma característica do universo, preferimos construir um intervalo dentro do qual pode encontrar-se o valor real. Intervalo de confiança é um intervalo cujos limites são calculados dos valores da amostra. Em amostras repetidas, os limites do intervalo variarão, mas o método de calcular os limites dos valores da amostra permanece constante. Quando dizemos que há uma probabilidade de p porcento que êste intervalo contenha o valor do universo, queremos dizer em realidade que os intervalos calculados desta forma conterão o valor do universo em aproximadamente p porcento de uma série de amostras repetidas. (Para uma definição matemática precisa, ver o VBE)

* MÉDIA (VBE Média aritmética) - De uma coleção de n valores x_i é o quociente

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

da divisão de sua soma pelo seu número.

* MEDIANA - I. De um conjunto de 2n + 1 valores ordenados, ou rol, é o va

lor que é precedido e seguido pelo mesmo número, n, de valores.

II - De um conjunto de 2n valores ordenados, ou rol, é a média aritmética entre os valores de ordem n e de ordem n + 1.

III - De uma distribuição de freqüência de valores de x é o valor da abscissa x cuja ordenada divide ao meio a área do histograma dessa distribuição.

IV - De uma distribuição teórica de freqüência é o valor da abscissa x cuja ordenada divide ao meio a área total da respectiva curva de freqüência.

* NÚMEROS ALEATÓRIOS (VBE Números equiprováveis) - São números, dispostos em tabela, usados para a seleção acidental de amostras; admite-se que tôda seqüência dêsses números, tais como aparecem tabulados, é acidental, isto é, a probabilidade de escolha de um é a mesma que a de qualquer outro. Também chamados de números aleatórios (de Tippet) e de números de Tippett.

PERMUTAÇÃO - A permutação de n itens significa um arranjo daqueles itens em alguma ordem especificada. Se os n itens fôrem todos diferentes, há ..... $n! = n(n-1) \ldots 2.1$ permutações diferentes. Por exemplo, as permutações das três letras abc são: abc, acb, bac, cab, bca, cba.

PONDERAÇÕES - Em geral, são os fatôres pelos quais os valores de amostra são multiplicados a fim de calcularem-se estimativas de totais do universo.

PROCESSO DE SELEÇÃO ALEATÓRIA - É qualquer processo de seleção que resulta em uma amostra probabilística, isto é, uma em que cada elemento do universo tenha uma probabilidade, conhecida e diferente de zero, de seleção.

SELEÇÃO DA AMOSTRA - É a designação dos elementos específicos do universo que devem ser incluídos na amostra.

SIGNIFICÂNCIA ESTATÍSTICA - Diz-se que a diferença entre dois valores, um ou ambos os quais são estimados de uma amostra, tem significância estatística se a probabilidade fôr pequena (geralmente 5% ou menos) de que a diferença pudesse dever-se ùnicamente a erros de amostragem.

* SUB-AMOSTRA (VBE Amostra, sub) - É amostra de amostra.

* TAXA DE AMOSTRAGEM (VBE Amostragem, fração de) - De uma amostra, é o quociente da divisão do tamanho da amostra pelo tamanho da população.

TENDENCIOSIDADE - É a diferença entre o valor esperado de uma estimativa e o valor real do universo. (O Prof. Rodrigues emprega o têrmo "viés").

TENDENCIOSIDADE DO PLANEJAMENTO DA AMOSTRA (OU NA SELEÇÃO) - É todo aspecto no planejamento ou seleção da amostra que resulta na impossibilidade de obter uma amostra probabilística verdadeira.

* UNIDADE ELEMENTAR - Suporte do atributo cuja observação constitui o fim de um levantamento por amostra. Também se diz unidade de análise.

* UNIVERSO (VBE População) - No sentido da inferência ou indução estatística, é todo conjunto de indivíduos para o qual se pretendem generalizar as propriedades encontradas nos conjuntos de elementos extraídos daquele. (Com algumas exceções, o têrmo universo tem sido usado de preferência a população a fim de evitar confusão com o emprêgo demográfico de população).

VALOR ESPERADO (DA ESTIMATIVA) - É o valor médio de uma estimativa para tôdas as amostras possíveis. Se as amostras não fôrem tôdas igualmente prováveis, deve-se usar uma média ponderada.

VALOR REAL (DO UNIVERSO) - É o valor "verdadeiro" de uma característica, isto é, o valor que seria obtido se se fizesse uma contagem completa e tôdas as observações, medidas e cálculos estivessem completamente livres de êrro.

VARIÂNCIA (de uma estimativa) - É o desvio quadrático médio da estimativa do seu valor esperado.

VARIÂNCIA (de uma característica do universo) - É o desvio quadrático médio da característica do seu valor médio.

VARIÂNCIA RELATIVA (de uma estimativa) - É a variância de uma estimativa dividida pelo quadrado do valor esperado da estimativa.

VARIÂNCIA RELATIVA (de uma característica do universo) - É a variância da característica dividida pelo quadrado do valor médio da característica.

VARIÁVEL ALEATÓRIA - É tôda variável cujos valores são determinados por um processo aleatório.

SECÇÃO B - Exame de expressões relacionadas com os erros de amostragem e os erros alheios à amostragem

Expressões tais como "precisão" e "êrro de amostragem" foram usadas com muita liberalidade nesta série de palestras. Enquanto que "precisão" foi definida na terceira palestra como "a diferença entre o valor real e a estimativa", a mesma foi, em realidade, usada alternadamente para referir-se a:

1. A diferença definida
2. A diferença entre a estimativa e seu valor esperado.

A expressão "desvio-padrão" foi aplicada indiscriminadamente a características de população e a estimativas de amostras. Outros exemplos semelhantes poderiam ser citados.

Nesta secção tentar-se-á oferecer uma definição um pouco mais precisa a êste grupo de expressões, seguindo a terminologia do Prof. Rodrigues, sempre que possível. Ao estudar estas definições, há três distinções importantes que se devem ter em mente:

1. A distinção entre os erros de amostragem e os erros alheios à amostragem. Os erros de amostragem surgem unicamente do fato de que baseamos nossas estimativas em uma amostra de indivíduos em lugar de fazer uma contagem completa. Êles representam as diferenças entre uma estimativa e seu valor esperado, e não levam em conta nenhuma diferença que possa existir entre o valor esperado de uma estimativa e o valor "real" do universo que está sendo estimado.

2. A distinção entre uma diferença observada em uma determinada amostra e uma função que represente, de alguma forma o valor médio ou esperado desta diferença em uma série de amostras repetidas.

3. A distinção entre o desvio-padrão ou variância de uma característica do universo, que expressa a maneira pela qual os valores da característica para elementos individuais variam da média do universo; e o desvio-padrão ou variância de uma estimativa, que expressa a maneira pela qual as estimativas baseadas em amostras repetidas variam do valor esperado ou média em tôdas as amostras possíveis. (Esta distinção já foi observada na secção anterior nas definições de coeficiente de variação, desvio-padrão, variância e variância relativa).

Tendo estas definições em mente, definiremos agora algumas das expressões mais específicas: Êrro real é a diferença, para uma determinada amostra, entre uma estimativa e o valor do universo que está sendo estimado. É representado simbòlicamente por $\hat{\theta} - \theta$.

Êrro aparente é a diferença, para uma determinada amostra, entre uma estimativa e seu valor esperado. É representado simbòlicamente por $\hat{\theta} - E(\hat{\theta})$.

Em inglês, faz-se geralmente referência a estas quantidades simplesmente como "differences" com a qualificação apropriada. O Prof. Rodrigues parece considerar o "êrro amostral" ou "êrro de amostragem" como equivalente ao êrro real. Prefiro reservar estas expressões para o valor generalizado do "êrro aparente".

Afastamento padrão é uma medida do desvio de valores individuais de uma característica do universo da sua média global. Especìficamente, é a raiz quadrada da média dos desvios individuais da média ao quadrado, isto é

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Êrro padrão é a medida análoga do desvio de uma estimativa de amostra do seu valor esperado, isto é

$$\sqrt{E\left[\hat{\theta} - E(\hat{\theta})\right]^2}$$

Êrro quadrático médio é uma medida do desvio de uma estimativa do valor real do universo. É o mesmo que o êrro padrão com o valor esperado da estimativa substituído pelo valor real do universo, isto é

$$\sqrt{E\,(\hat{\theta} - \theta)^2}$$

Em inglês, tanto o afastamento padrão como o êrro padrão são conhecidos como "standard deviation". O êrro quadrático médio é conhecido como "root mean square error". O professor Rodrigues considera o desvio-padrão como equivalente ao afastamento padrão; nas palestras, entretanto, serviu tanto para afastamento padrão como para êrro-padrão.

Os quadrados das expressões afastamento padrão e êrro padrão são aparentemente ambos considerados como variância no Vocabulário do Prof. Rodrigues. Não encontro em seu Vocabulário um têrmo para o quadrado do êrro quadrático médio.

Consideremos agora algumas das expressões mais gerais. Êrro de amostragem, na minha opinião, deve sempre referir-se a diferenças ao acaso devidas à amostragem, isto é, a diferença entre uma estimativa e seu valor esperado. Pode-se referir à diferença observada para uma determinada amostra ou a uma medida generalizada tal como o êrro padrão, a variância, a variancia relativa ou o coeficiente de variação.

Três têrmos que são ainda mais gerais são precisão, exatidão e fidedignidade. É talvez inútil tentar fazer distinção entre êles conforme são aplicados à amostragem. Devem referir-se a um método de estimação ou a uma determinada estimativa e podem ser usados a fim de referir-se a diferenças desta estimativa do valor esperado, do valor real ou de ambos, conforme o caso. Não devem ser aplicados a uma amostra ou método de amostragem, visto como não se pode avaliar precisão ou seu equivalente sem ter em mente um método específico de estimação. Tècnicamente, não existe uma amostra fidedigna.

A N E X O S

CURSO DE AMOSTRAGEM

Experimento nº 1

Objetivo do experimento: Selecionar repetidas amostras de 100 bolinhas contidas no cilindro, usando-as para estimar a proporção de bolas brancas, e verificar até que ponto essas estimativas seguem uma distribuição normal. Neste experimento utilizamos amostras de um universo conhecido, e assim podemos prever a proporção de estimativas que cairão em um determinado intervalo.

O cilindro contém

300 bolas brancas
700 bolas vermelhas
1 000 bolas ao todo

A proporção de bolas brancas no universo é: $p = \frac{300}{1\,000} = 0,30$

Nossa estimativa dessa proporção é:

$$\hat{p} = \frac{\text{Nº de bolas brancas na amostra}}{100}$$

O desvio padrão dessa estimativa é:

$$\sigma_{\hat{p}} = \sqrt{\frac{0,30 \times 0,70}{100}} = 0,0458$$

Admitindo que a distribuição da estimativa seja normal, podemos esperar que:

- 68% aproximadamente das estimativas estarão entre 0,2542 e 0,3458, isto é, $0,3000 \pm 0,0458$ (Cf. 5ª Palestra, pg. 25);
- 95% aproximadamente das estimativas estarão entre 0,2084 e 0,3916, isto é, $0,3000 \pm 2 \times 0,0458$ (idem).

Para testar essa hipótese, cada participante dêste experimento deverá usar o "amostrador" para selecionar 5 amostras de 100 bolas cada, observando o processo indicado a seguir, e registrando os resultados na tabela abaixo.

1. O recipiente metálico do interior do cilindro está destinado a receber 10 bolas. Rode o cilindro duas ou três vêzes, e em seguida ponha o cilindro em posição que permita que o recipiente fique para cima. Se nele ficarem mais de 10 bolas, mova o cilindro ligeiramente, até que as bolas-extra caiam; se ficarem menos de 10 bolas, rode novamente o cilindro até obter o número requerido.

2. Quando obtiver as 10 bolas no cilindro, registre na tabela o número de bolas brancas, correspondente à "jogada", e repita a experiência. Ao todo, deverão ser feitas 50 "jogadas" (10 para cada amostra de 100 bolas).

3. Depois de 3 ou 4 "jogadas", sacuda o cilindro no sentido lateral, a fim de que as bolas fiquem bem misturadas.

Nota: Os resultados que obtenha cada participante serão posteriormente computados, para a verificação que se intenta realizar.

RESULTADOS DO EXPERIMENTO

BOLAS BRANCAS (Número obtido)	AMOSTRAS DE 100				
	1	2	3	4	5
1ª jogada					
2ª "					
3ª "					
4ª "					
5ª "					
6ª "					
7ª "					
8ª "					
9ª "					
10ª "					
TOTAL					

Resultados do Experimento nº 1

Doze pessoas realizaram êste experimento, tendo cada qual registrado o número de bolas brancas em 5 amostras de 100 bolas, dando um total de 60 amostras de 100. Em cada amostra de 100, o número esperado de bolas brancas foi 30.. Os resultados reais foram os seguintes:

Nº de bolas brancas	Nº de amostras	Nº de bolas brancas	Nº de amostras
21	1	32	6
22	0	33	4
23	0	34	3
24	5	35	5
25	5	36	0
26	5	37	0
27	2	38	0
28	6	39	0
29	8	40	0
30	5	41	1
31	4	TOTAL	60

A relação desta distribuição com os valores previstos é apresentada pelo seguinte quadro:

DIFERENÇA ENTRE O NÚMERO OBSERVADO E O NÚMERO ESPERADO DE BOLAS BRANCAS		AMOSTRAS COM RESULTADOS NESTE INTERVALO		
Valor absoluto	Desvios-padrões	Nº	%	% esperada
0 - 4,6	Menos de 1 σ	43	71,6	68,3
4,9 - 9,2	Entre 1 e 2 σ	16	26,7	27,1
9,2 e mais	Mais de 2 σ	1	1,7	4,6
Total		60	100,0	100,0

A concordância entre as distribuições observadas e as teóricas apresentadas nas 2 últimas colunas é satisfatória para êsse número de amostras. Com outro teste, podemos computar a estimativa de p baseada no grupo todo de 60 amostras.

$$\text{Temos } \hat{p} = \frac{1\,766}{6\,000} = 0{,}294$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0{,}30 \times 0{,}70}{6\,000}} = 0{,}006$$

De maneira que a diferença entre $\hat{p}$ e o valor real (0,300) é aproximadamente igual a um desvio-padrão, diferença que seria excedida pelo acaso em cêrca de 1 em cada 3 amostras de 6 000.

CURSO DE AMOSTRAGEM

Experimento nº 2

Objetivo: Demonstrar o emprêgo dos intervalos de confiança, quando se procede a uma amostragem de universo desconhecido.

No cilindro será colocado um determinado número de bolas (não inferior a 500, ao todo), obedecendo a qualquer proporção entre a quantidade de bolas vermelhas e brancas. Tanto o número total de bolas quanto a proporção entre as duas parcelas serão ignorados pelos participantes do experimento.

A êles caberá, então, repetir as operações prescritas no Experimento nº 1, de modo a obter 50 amostras de dez bolas cada, registrando o resultado obtido (número de bolas brancas na amostra) na Tabela A (anexa). Não esquecer de misturar as bolas lateralmente, de vez em quando.

Para cada grupo de 10 amostras deve-se completar os dados solicitados no quadro B (Resumo), consoante as seguintes instruções:

. Col. (1) - Número total de bolas brancas dividido por 100.

Cols. (2) e(3) - Registre os limites superior e inferior do intervalo de confiança correspondente ao valor registrado na coluna (1), limites êsses dados na Tabela C (anexa).

Col. (4) - Esta coluna só será preenchida quando a proporção real de bolas brancas fôr revelada (após a realização do experimento). Marque "sim" se a proporção real estiver dentro dos limites indicados nas cols. (2) e (3) (inclusive quando fôr igual a um dêles). Marque "não" se o valor real estiver fora dêsses limites.

Os intervalos de confiança estão planejados de tal forma que podemos esperar que a resposta na col. (4) seja "sim" em cêrca de 95% das amostras de 100.

* * * *

Instruções do experimento nº 1

1. O recipiente metálico do interior do cilindro está destinado a receber 10 bolas. Rode o cilindro duas ou três vêzes, e em seguida ponha-o em posição que permita que o recipiente fique para cima, com o número de bolas requerido.

2. Quando obtiver as 10 bolas no recipiente, registre na tabela o número de bolas brancas, correspondentes à "jogada", e repita a experiência. Ao todo, deverão ser feitas 50 "jogadas" (10 para cada amostra de 100 bolas).

3. Depois de 3 ou 4 "jogadas", sacuda o cilindro no sentido lateral, a fim de que as bolas fiquem bem misturadas.

QUADRO A - VALORES OBSERVADOS

(Bolas brancas)

JOGADAS	NÚMERO DE BOLAS BRANCAS, POR AMOSTRA				
	Amostra 1	Amostra 2	Amostra 3	Amostra 4	Amostra 5
1ª jogada					
2ª "					
3ª "					
4ª "					
5ª "					
6ª "					
7ª "					
8ª "					
9ª "					
10ª "					
TOTAIS					

QUADRO B - RESUMO DAS AMOSTRAS DE 100

AMOSTRA	PROPORÇÃO ESTIMADA DE BOLAS BRANCAS (1)	LIMITES DO INTERVALO DE CONFIANÇA		VALOR REAL NO INTERVALO	
		Inferior (2)	Superior (3)	Sim (4)	Não
Amostra 1					
Amostra 2					
Amostra 3					
Amostra 4					
Amostra 5					

TABELA C

INTERVALOS DE CONFIANÇA DE UMA ESTIMATIVA DE PROPORÇÃO

(Amostra de 100)

ESTIMATIVA DA PROPORÇÃO DE BRANCAS	LIMITE INFERIOR	LIMITE SUPERIOR	ESTIMATIVA DA PROPORÇÃO DE BRANCAS	LIMITE INFERIOR	LIMITE SUPERIOR
0,00	0,00	0,03	0,51	0,41	0,61
0,01	0,00	0,05	0,52	0,42	0,62
0,02	0,00	0,06	0,53	0,43	0,63
0,03	0,01	0,08	0,54	0,44	0,64
0,04	0,02	0,10	0,55	0,45	0,65
0,05	0,02	0,11	0,56	0,46	0,66
0,06	0,03	0,12	0,57	0,47	0,66
0,07	0,03	0,13	0,58	0,48	0,67
0,08	0,04	0,15	0,59	0,49	0,68
0,09	0,05	0,16	0,60	0,50	0,69
0,10	0,05	0,17	0,61	0,51	0,70
0,11	0,06	0,18	0,62	0,52	0,71
0,12	0,07	0,20	0,63	0,53	0,72
0,13	0,08	0,21	0,64	0,54	0,73
0,14	0,08	0,22	0,65	0,55	0,74
0,15	0,09	0,23	0,66	0,56	0,75
0,16	0,10	0,24	0,67	0,57	0,76
0,17	0,11	0,25	0,68	0,58	0,77
0,18	0,11	0,26	0,69	0,59	0,77
0,19	0,12	0,27	0,70	0,60	0,78
0,20	0,13	0,28	0,71	0,62	0,79
0,21	0,14	0,30	0,72	0,63	0,80
0,22	0,14	0,31	0,73	0,64	0,81
0,23	0,15	0,32	0,74	0,65	0,82
0,24	0,16	0,33	0,75	0,66	0,83
0,25	0,17	0,34	0,76	0,67	0,84
0,26	0,18	0,35	0,77	0,68	0,85
0,27	0,19	0,36	0,78	0,69	0,86
0,28	0,20	0,37	0,79	0,70	0,86
0,29	0,21	0,38	0,80	0,72	0,87
0,30	0,22	0,40	0,81	0,73	0,88
0,31	0,23	0,41	0,82	0,74	0,89
0,32	0,23	0,42	0,83	0,75	0,89
0,33	0,24	0,43	0,84	0,76	0,90
0,34	0,25	0,44	0,85	0,77	0,91
0,35	0,26	0,45	0,86	0,78	0,92
0,36	0,27	0,46	0,87	0,79	0,92
0,37	0,28	0,47	0,88	0,80	0,93
0,38	0,29	0,48	0,89	0,82	0,94
0,39	0,30	0,49	0,90	0,83	0,95
0,40	0,31	0,50	0,91	0,84	0,95
0,41	0,32	0,51	0,92	0,85	0,96
0,42	0,33	0,52	0,93	0,87	0,97
0,43	0,34	0,53	0,94	0,88	0,97
0,44	0,34	0,54	0,95	0,89	0,98
0,45	0,35	0,55	0,96	0,90	0,98
0,46	0,36	0,56	0,97	0,92	0,99
0,47	0,37	0,57	0,98	0,94	1,00
0,48	0,38	0,58	0,99	0,95	1,00
0,49	0,39	0,59	1,00	0,97	1,00
0,50	0,40	0,60			

Resultados do Experimento nº 2

Também 12 pessoas realizaram êste experimento, cada qual registrando 5 amostras de 100 bolas cada, e determinando os intervalos de confiança adequados a cada amostra, através das tabelas apresentadas nas instruções.

Em seguida à seleção destas amostras, verificou-se que a proporção real de bolas brancas do cilindro foi 0,290. Determinou-se, então, que 54, ou 90%, dos 60 intervalos continham o valor real. A proporção esperada era 95%.

A estimativa de p baseada em tôdas as 60 amostras foi

$$\hat{p} = \frac{1\ 687}{6\ 000} = 0{,}281$$

O êrro-padrão de $\hat{p}$ é 0,006, como no Experimento nº 1, de maneira que a diferença entre os valores observado e esperado é

$$0{,}290 - 0{,}281 = 0{,}009 = 1{,}5\ \sigma$$

A P Ê N D I C E (A)

BIBLIOGRAFIA ABREVIADA SÔBRE AMOSTRAGEM E SUAS APLICACÕES AO TRABALHO CENSITÁRIO

Nota: Destina-se esta bibliografia a fornecer uma lista de material de consulta às pessoas interessadas em aprender algo sôbre amostragem, e suas aplicações ao trabalho censitário, sem que necessitem estudar detalhadamente a teoria matemática ligada à mesma. Não se trata, assim, de uma bibliografia completa sôbre a amostragem.

Nas publicacões relacionadas, os capítulos enumerados são os que requerem menor conhecimento de matemática.

As obras assinaladas com asterisco (*) acham-se à disposição dos interessados no Grupo Técnico de Coordenação.

I. LIVROS

A. Sôbre amostragem

1. Cochran, W.G., Sampling Techniques, John Wiley & Sons, New York, 1953.
* 2. Deming, W. Edwards, Some Theory of Sampling, John Wiley & Sons, New York, 1950 (Capítulos 1, 2, Secção A do capítulo 4, e capítulo 12). (Existe também em espanhol).
* 3. Hansen, M.H.; Hurwitz, W.N. e Madow, W.G., Sample Survey Methods and Theory (2 volumes: Vol. I - Methods and Applications, Vol. II - Theory), John Wiley & Sons, New York, 1953 (Vol. I, capítulos 1-3 e 12).
4. Madow, W.G., Teoria dos Levantamentos por Amostragem, Tipografia da Empresa Nacional de Publicidade, Lisboa, 1950 (Em português).
5. Sukhatme, P.V., Sampling Theory of Surveys with Applications, Iowa State College and Indian Society of Agricultural Statistics, 1954. (Capítulo 1).
* 6. Yates, F., Sampling Methods for Censuses and Surveys, Charles Griffin & Co., London, 1949. (Capítulos 1-5).

B. Sôbre probabilidade

* 1. Feller, W., Probability Theory and its Applications, John Wiley & Sons, New York, 1950.
* 2. Fry, T.C., Probability and its Engineering Uses, Van Nostrand, New York, 1928.

C. Contrôle de qualidade

* 1. Grant, Eugene, Statistical Quality Control, McGraw-Hill, New York, 1946.
* 2. Statistical Research Group (Freeman, Friedman, Mosteller e Wallis), Sampling Inspection, McGraw-Hill, 1948.

II. ARTIGOS EM REVISTAS CIENTÍFICAS

Nota: JASA - Journal of the American Statistical Association

* 1. Chevry, M.G., "Contrôle de um Recenseamento através da amostragem à base de áreas", Revista Brasileira de Estatística, Ano XIV, Nº 53, jan./março de 1953, págs. 13-18. (Português)

* 2. Cochran, W., Mostelles, F. e Tukey, J. W., "Principles of Sampling", JASA, Vol. 49 (1954), págs. 13-35.

* 3. Deming, W.E., "On Sampling by a System of Replicated Drawings with Equal Probabilities and Without Stages, to Gain Efficiency and Simplicity, Especially in the Estimates of the Standard Error", Reference document, III Inter American Statistical Conference, 1955.

4. Deming, W.E. e Geoffrey, L., "On Sampling Inspection in the Processing of Census Returns", JASA, Vol. 36 (1941), págs. 351-360.

5. Deming, W.E. e Hansen, M. "On Some Census Aids to Sampling", JASA, Vol. 38 (1943), págs. 353-357.

* 6. Deming, W.E.; Hansen, M.H. e Stephan, F.F., "The Sampling Procedure of the 1940 Population Census", JASA, Vol. 35 (1940), págs. 615-630.

* 7. Eckler, A.R., "Extent and Character of Errors in the 1950 Censuses of Population and Housing", conferência pronunciada perante o New York Chapter of American Statistical Association, 1953.

* 8. Eckler, A.R. e Pritzker, L., "Measuring the Accuracy of Enumerative Surveys", Proceedings of the International Statistical Institute, Twenty-seventh Session, 1951.

* 9. Goldfield, E.D.; Steinberg, J. e Welch, E.H., "The Monthly Report on the Labor Force", Estadística, Vol. 6, março de 1948.

10. Gurney, M.; Hansen, M.H. e Hurwitz, W.N., "Problems and Methods of a Sample Survey of Business", JASA, Vol. 41 (1946), págs. 173-189.

*11. Hansen, M.H. e Hauser, P.M., "On Sampling in Market Surveys", Journal of Marketing, Vol. 9 (1944), págs. 26-31.

*12. Hansen, M.H., "Sampling of Human Populations", ISI, Twenty-fifth Session, 1947.

*13. Hansen, M.H. e Hauser, P.M., "Area Sampling - Some Principles of Sample Design", Public Opinion Quarterly, Vol. 9, (1945), págs. 183-193.

*14. Hansen, M.H. e Hurwitz, W.N., "A New Sample of the Population", Estadística, Vol. 2 (1944), págs. 483-497.

*15. Hansen, M.H. e Hurwitz, W.N., "The Problem of Non-Response in Sample Surveys", JASA, Vol. 41 (1946), págs. 517-529.

*16. Hansen, M.H.; Hurwitz, W.N.; Nisselson, H. e Steinberg, J., "The Redesign of the Census Current Population Survey", JASA, Vol. 50 (1955), págs. 701-719.

*17. Hansen, M.H.; Hurwitz, W.N. e Pritzker, L., "The Accuracy of Census Results", American Sociological Review, Vol. 18, agôsto de 1953. (Existe

também em espanhol, em Estadística, Vol. 13, nº 46, págs. 74-85).

18. Jessen, R.J. e King, A.J., "The Master Sample of Agriculture", JASA, Vol. 40 (1945), págs. 38-56.

* 19. Kriesberg, M. e Voight, R., "Some Principles of Processing Census and Survey Data", JASA, Vol. 47 (1952), págs. 222-231.

* 20. Marks, E.; Mauldin, W.P., e Nisselson, H., "The Post-Enumeration Survey of the 1950 Census: A Case History in Survey Design", JASA, Vol. 48 (1953), págs. 220-243.

* 21. Meter, J., "Some Applications of Statistics for Auditing", JASA, Vol. 47 (1952), págs. 6-24.

* 22. Steinberg, J. e Waksberg, J., "Sampling in the 1950 Census of Population and Housing", Proceedings of the American Statistical Association, The 111th Annual Meeting, 1951. (Existe uma tradução em português).

23. Steines, P.O., "A Source of Bias in One of the Samples of the 1950 Census", JASA, Vol. 46 (1951), págs. 110-114.

* 24. Stevens, W.L., "Estimation of the Brazilian Coffee Harvest by Sampling Survey", JASA, Vol. 50 (1955), págs. 775-787.

* 25. Sukhatme, P.V., "A Amostragem nas Estatísticas Agrícolas", Revista Brasileira de Estatística, Ano XVI, Nº 61, jan./março de 1955, págs. 15-18. (Português)

* 26. Yates, Frank, "Métodos de Amostragem em Censos e Levantamentos", Revista Brasileira de Estatística, Ano XII, Nº 47, julho-setembro de 1951, págs. 279-290 (Português)

* 27. Žarković, S.S., "Sampling Methods in the Yugoslav 1953 Census of Population", JASA, Vol. 50 (1955), págs. 720-737.

III. PUBLICAÇÕES DO BUREAU DO CENSO DOS ESTADOS UNIDOS.

* 1. Deming, W.E. e Stephan, F.F., "On the Sampling Methods in the 1940 Population Census", março de 1941.

2. Sampling Staff, A Chapter in Population Sampling, 1947 (Existe também em espanhol).

* 3. Papers on Labor Force Statistics in the United States, 1952 (Ver especialmente os documentos 8-15 e 22)

* 4. "The 1950 Censuses - How They Were Taken", Procedural Studies of the 1950 Censuses, Nº 2. (Ver especialmente: Capítulo 1, págs. 5-8; Capítulo 4, pág. 24; Capítulo 5, pág. 26).

5. Além do mencionado, são de interêsse algumas exposições introdutórias, em que se descrevem processos de amostragem utilizados, nas publicações censitárias norte-americanas (censos de população, agricultura, indús-

tria, etc.) e nas publicações das estatísticas permanentes dos Estados Unidos.

IV. DIVERSOS

* 1. Inter American Statistical Institute, Statistical Vocabulary (Dicionário de têrmos estatísticos em português, inglês, espanhol e francês).

2. Nisselson, H., "Aplicacion de los Metodos de Muestreo en la Elaboracion de Censos", Bogotá, Contraloria General de la Republica de Colombia, 1950.

* 3. United Nations, Statistical Commission, Subcommission on Statistical Sampling, "The Preparation of Sampling Survey Reports", Statistical Papers, Series C, Nº 1 (Revised), 1950 (Existe também em espanhol).

V. ITENS ADICIONAIS

(Os itens que se seguem, todos em português, foram adicionados após ter sido preparada a lista original)

1. Dereymaeker, R. "Utilização dos métodos de amostragem nas estatísticas oficiais dos Estados Unidos", Revista Brasileira de Estatística, Ano XVII, Nº 65, janeiro/março de 1956, págs. 21-31.

2. Frias, Roque Garcia, "Os censos de 1950 e a aplicação da amostragem," Revista Brasileira de Estatística, Ano XI, Nº 43, julho/setembro de 1950, págs. 379-390

3. Kingston, Jorge, "Dimensionamento de amostras", Revista Brasileira de Estatística, Ano V, Nº 19, julho/setembro de 1944, pags. 299-304.

4. Madow William G. "Porque usamos amostras", Revista Brasileira de Estatística, Ano VII, Nº 27, julho/setembro de 1946, págs. 489-502.

5. Montenegro, Tulo Hostílio, "O comprimento do período como característica estatística do estilo, "Revista Brasileira de Estatística, Ano XVI, Nº 63, julho/setembro de 1955, págs. 193-274.

6. De Morais, O. Alexander, "Posição, fundamento e aplicação da amostragem por seleção ao acaso no campo da estatística administrativa", Revista Brasileira de Estatística, Ano XI, Nº 36, outubro/dezembro de 1948, págs. 627-636.

7. Rodrigues, Milton da Silva, "Vocabulário Brasileiro de Estatística", Revista Brasileira de Estatística, Ano V, Nº 18, abril/junho de 1944, págs. 187-297.

8. Rodrigues, Milton da Silva, Vocabulário Brasileiro de Estatística,

Conselho Nacional de Estatística, I.B.G.E., Rio de Janeiro, 1956.

NOTA: O item 3 da Secção IV, "The Preparation of Sampling Survey Reports", foi traduzido para o português aparecendo sob o título "Recomendações básicas sôbre amostragem", na Revista Brasileira de Estatística, Ano XII, Nº 45, janeiro/março de 1951, págs. 85-90.