PD ABC-389

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

John Schneider

January 31, 1990

TRIP REPORT
CARACAS, VENEZUELA
January 7-13, 1990

MARJORIE DAUPHIN
International Mathematical Statistical Staff
Statistical Methods Division

Trip Purpose and Summary

At the request of the Government of Venezuela, Marjorie Dauphin traveled to Caracas to assist the Oficina Central de Estadística e Informática (OCEI) in developing sampling procedures for the 1990 Census of Population and Housing. During her visit, Dauphin collaborated with counterparts at OCEI on refinements of the sample plan and the development of estimation procedures. Her recommendations are documented in the report titled: "Consideraciones para el Proceso de Estimación en la Muestral Censal," which is provided as Attachment B.

General Aspects of Visit

Dauphin met with the OCEI Director General and General Coordinator of the 1990 census, Mr. Francisco Ortega, who gave her an overview of the census organization and activities. Dauphin later met with the census working group chiefs, "jefes de unidades censales", who discussed with her current and upcoming activities in each of their different areas. The census day is scheduled for either the second or the third Sunday in October 1990; the exact date will be decided later.

At the end of the visit, Dauphin met again with Ortega and the census working group chiefs, and with the Chief of OCEI, Dr. Félix Seijas, to present the draft technical report and recommendations. The report points to strengths and weaknesses of the sample design and recommends specific estimation procedures. Overall, Dauphin was impressed with the degree of organization of the census operations, the great effort being exercised by the OCEI personnel, and their concern for the success of the census. Dauphin also met and briefed representatives of the U. S. Embassy and representatives of the United Nations in Caracas on the work carried out.

Best Available Document

Technical Aspects

Dauphin worked closely with Harold Martín-Caro, Chief of the Unidad Censal de Muestreo (the sampling working group), Irene Gurrea, and Luis Montero of the same group. She reviewed their proposed sample plan for the census and provided comments. Given that the sample plan was considered final at the time of her visit, Dauphin concentrated on the development of estimation procedures, including variance estimation procedures. Based on this work, Dauphin drafted a technical report titled: "Consideraciones para el Proceso de Estimación en la Muestral Censal", which she finalized upon returning to Washington. The final report is provided as Attachment B.

Dauphin also reviewed segment maps developed during the cartographic operation and found them to be adequate and detailed enough for the creation of primary sampling units (PSUs). In the Venezuelan census sample, the PSU will be a subdivision of the segment and will be created on the segment maps. Dauphin recommended some general control checks to ensure the accuracy of coverage.

Follow-Up Recommendations

It is recommended that the OCEI sampling group:

- develop specific quality control measures for the selection of the sample;
- carry out a test of the sample selection procedures including its quality control aspects;
- following the test, make revisions to the sample plan and the quality control procedures, as needed;
- refine the procedures recommended for estimation;
- develop the computer programs necessary for estimation;
- conduct a test of the estimation procedures; and
- following the test, make necessary revisions to the estimation procedures and programs.

In addition, it is recommended that OCEI consider the implementation of a Post-Enumeration Survey (PES) to measure the accuracy of the census in terms of coverage and content, and to evaluate the adequacy of the area frame for household surveys during the intercensal decade. The development of PES procedures should begin as soon as possible.

Technical Assistance Recommendations

If requested, the U. S. Bureau of the Census can provide additional technical assistance to OCEI in the following sample design areas:

- follow-up and revision of estimation procedures for census sample, following test (one 1-week visit);

- planning and development of PES sample design, analytical plans, and matching forms and procedures (one 3-week visit and two 2-week visits) - one visit may be combined with the follow-up visit for census sample estimation;

- analysis of PES (one 2-week visit);

- following the census, development of master sampling frame of enumeration areas (one 3-week visit and two 2-week visits).

Other recommendations for technical assistance were provided in Glenn Ferri's trip report of October 1989. It is necessary that visits be funded to include preparation and follow-up time in Washington. The budget provided in Ferri's report includes such preparation and follow-up time. The technical reports referred to in this report are on file in the International Mathematical Statistics Staff of the International Statistical Programs Center.

Attachments

CC:

J. Preston	AmEmbassy/Caracas (3)
A. Jefferson	UNDP/Caracas
F. Seijas	OCEI (4)
M. Broady	AID/LAC/SAM
C. Buck	AID/LAC/SAM
G. Britan	AID/PPC/CDIE

ATTACHMENT A

LIST OF PERSONS CONTACTED

Oficina Central de Estadística e Informática (OCEI)

Dr. Félix Seijas, Chief, Oficina Central de Estadística e Informática
Francisco Ortega E., Director General of OCEI and Census Coordinator
María de Aguilera, Chief, Cooperación Técnica
Harold Martín-Caro, Chief, Unidad Censal de Muestreo
Irene Gurrea, Unidad Censal de Muestreo
Luis Montero, Unidad Censal de Muestreo
Rafael Armada, Chief, Unidad de Operaciones Censales
Oscar García, Chief, Unidad Técnica Censal
Iraida Rodríguez, Chief, Unidad Censal de Reclutamiento, Selección, y Adiestramiento
Vaughan Salas, Chief, Unidad de Promoción y Publicación Censal
Joe Hernández, Chief, Unidad Censal de Procesamiento
Zoraida López, Chief, Unidad de Control y Evaluación de Actividades Censales
Iván Reina, Chief, Unidad Censal de Administración

American Embassy, Caracas

Alfred J. White, Economic Adviser
Helen Hudson

United Nations Development Program

Alfredo Jefferson, Representante Residente Adjunto

ATTACHMENT B

TECHNICAL REPORT

XD-ABC-389-A
BA 10129

CONSIDERACIONES PARA EL PROCESO DE ESTIMACION EN LA MUESTRA CENSAL

XII CENSO DE POBLACION Y VIVIENDA DE VENEZUELA (1990)

Marjorie Dauphin
U. S. Bureau of the Census
Enero de 1990

CONTENIDO

I. Selección de Secciones y Viviendas Muestrales 1

II. Proceso de Estimación: Expansión de la Muestra al Universo 2

Probabilidades de Selección 2
Selección del Tipo de Estimador 2
Post-Estratificación. 3
Estimación de Personas. 4
Estimación de Viviendas y de Hogares. 6
Asignación de la Ponderación. 6

III. Proceso de Estimación: Cálculo y Presentación de Errores Muestrales 9

Presentación 9
Varianza de un Total Basado en un Estimador de Razón 10
Estimación por el Método de Conglomerados Ultimos. 12
Estimación por el Método de Muestreo Unietápico de Conglomerados . 13
Varianza de Estimaciones de Parámetros Relativos 16

IV. Encuesta Evaluativa. 18

Anexo 1: Error Estándar para Estimador de Total Cualitativo en el Muestreo Aleatorio Simple

Anexo 2: Comentarios Generales Sobre el Plan de Muestreo

Anexo 3: Referencias Sobre el Procedimiento de Ponderación Utilizado en el Censo de Estados Unidos de 1980

El plan muestral diseñado por la Unidad Censal de Muestreo para la aplicación del cuestionario ampliado en el XII Censo de Población y Vivienda de Venezuela consiste en una muestra estratificada unietápica de conglomerados de 20%. El plan propone garantizar estimaciones separadas y fiables para cada municipio. Dentro del municipio, se crearon estratos llamados "segmentos" que son áreas geográficas de 200 viviendas aproximadamente. Las unidades primarias de muestreo son conglomerados pequeños de 20 viviendas aproximadamente llamados "secciones".

El plan muestral está documentado en el informe titulado "Diseño de la Muestra del XII Censo General de Población y Vivienda (Lineamientos Generales)". El presente informe se concentra más bien en el proceso de estimación aunque aborda algunos puntos referentes al plan de muestreo.

I. SELECCION DE SECCIONES Y VIVIENDAS

Según lo discutido en la documentación del diseño muestral preparado por la Unidad Censal de Muestreo, los segmentos son considerados como estratos para los fines de selección y estimación. La unidad primaria de muestreo (UPM) es la sección. Se crearán 10 secciones dentro de cada segmento mediante una operación de cartografía y listado. Luego, se tomará una muestra de dos secciones exactamente dentro de cada uno de los segmentos conformados. Conviene que el número de secciones muestrales sea constante para facilitar los controles operacionales. Para mayor eficiencia muestral, la selección de secciones dentro del segmento se debe hacer de forma sistemática con arranque aleatorio. Una vez seleccionadas las secciones, se incluirán en la muestra todas las viviendas que en ellas se encuentren.

Conforme a las reglas de muestreo estratificado, la selección de secciones se deberá hacer dentro de cada segmento independientemente. Es decir, la selección de ciertas secciones en un segmento no deberá influir sobre cuáles secciones se elijan en otro. Puesto que las secciones serán de igual tamaño aproximadamente, la selección podrá hacerse con probabilidades iguales. Cabe recalcar que deberá existir un control operacional sumamente estricto para garantizar que, en efecto, las secciones formadas no varíen mucho en tamaño. De lo contrario, habrá una reducción en la fiabilidad de las estimaciones, puesto que uno de los factores que más afectan el error muestral en el muestreo por conglomerados es la variabilidad en el tamaño de las UPMs.

El éxito de la operación muestral dependerá en gran parte en la calidad del listado de viviendas y de los croquis. Además de los cuidados que se deberán tomar para que tanto los segmentos como las secciones que se formen sean de igual tamaño, es muy importante que los límites de cada sección y de cada segmento sean reconocibles en el campo. Cualquier confusión en cuanto a los límites en el momento del empadronamiento podrá significar errores de cobertura tanto para la muestra como para el empadronamiento exhaustivo.

Estos controles de calidad también deben involucrar a los supervisores y a los empadronadores. Se les debe dar instrucciones específicas en sus respectivos manuales y también durante la capacitación para tratar los problemas relacionados a la cobertura. La propuesta de enviar a los supervisores al campo 15 días antes del empadronamiento para hacer un reconocimiento general de los segmentos, de las secciones y sus respectivos límites es excelente y debe tratar de cumplirse. También es aconsejable que se hagan otras pruebas en el campo, después que se formen las secciones, para probar la metodología de selección de la muestra antes de ponerla en práctica.

II. PROCESO DE ESTIMACION: EXPANSION DE LA MUESTRA AL UNIVERSO

Probabilidades de Selección

Dentro de cada dominio de estimación (municipio, estado, etc.) y dentro de cada estrato (segmento) h, la probabilidad de selección de la j-ésima vivienda es la misma que la de seleccionar la i-ésima UPM (sección) y es igual a:

$$P(UPM) = \frac{n_h}{N_h}$$

donde:

N_h = número total de secciones en el segmento h; normalmente será 10;

n_h = no. de secciones en la muestra dentro del segmento h; será siempre 2.

Normalmente esta probabilidad será 2/10 ó 20 porciento y será constante para todas las viviendas no solamente dentro del segmento sino en toda la muestra. Sin embargo, si es necesario considerar la formación de más de 10 secciones dentro de un segmento por razón de su tamaño muy grande, la probabilidad de selección sería $2/N_h$. Seguiría siendo constante dentro de un segmento pero variaría algo de segmento a segmento.

Selección del Tipo de Estimador

La técnica recomendada para expandir los resultados de la muestra a la población completa es la de estimación por relativo ó razón. La Unidad Censal de Muestreo ha contemplado esta técnica y también la de estimación por regresión pues el estimador de regresión tiende a ser más eficiente que el de razón. Sin embargo, considerando que la ganancia en precisión a menudo es muy pequeña como para compensar por el aumento en dificultad y costo de aplicación y que, además, el estimador de regresión puede tener un rendimiento pobre en muestras pequeñas (como las que seguramente habrá a nivel de celda de estimación), es aconsejable optar por el estimador de razón.

Las ventajas asociadas con el uso del estimador de razón son que:

- o aumenta la precisión (fiabilidad) de las estimaciones muestrales sobre la del estimador convencional de total;
- o garantiza que los totales basados en estimaciones de la muestra coincidan con los provenientes del recuento a 100 porciento en cada municipio;
- o automáticamente ajusta para cualquier no-respuesta que eventualmente pudiera ocurrir dentro de la muestra;
- o corrige sesgos sistemáticos que eventualmente pudieran ocurrir durante el proceso de selección de la muestra.

Post-Estratificación

El método de razón implica el cálculo de un "factor de ponderación" ó "factor de expansión" el cual se define más adelante. El área geográfica donde se calcularán las ponderaciones, área de ponderación, es el municipio. Se recomienda combinar los segmentos dentro del municipio al definir el área de ponderación porque los segmentos son relativamente pequeños (200 viviendas) y porque hay que limitar el número de celdas. (En este caso, coincide el dominio de estimación, que es también el municipio, con el área de ponderación, aunque esto no es un requisito). Dentro del municipio, es aconsejable definir ciertas celdas homogéneas, "celdas de ponderación" ó "post-estratos", en base a las características de la población. Los dominios de estimación consistirán en una agregación de esas celdas de ponderación. Puesto que existen tres unidades de análisis diferentes: viviendas, hogares, y personas, deberá existir un factor de ponderación diferente para cada una y, por lo tanto, las celdas de ponderación serán diferentes para cada una.

Además de tener la calidad de homogéneos, los post-estratos definidos deben coincidir con las categorías definidas para la tabulación. Pero también es importante limitar el número de celdas formadas para evitar que haya algunas con muy pocas observaciones ya que en éstas la varianza sería muy alta. Además, hay que tener en cuenta que el estimador de razón es sesgado y que su sesgo aumenta a medida que disminuye el número de unidades muestrales en el que está basado. En otras palabras, es importante controlar el error cuadrático medio (ECM) de las estimaciones que es una combinación de varianza y sesgo.

Por lo tanto, en el caso de personas se recomienda usar a lo sumo tres variables de clasificación dentro del municipio. Según lo propuesto por la Unidad Censal de Muestreo, la población se podrá estratificar a posteriori en grupos de edad, sexo, y condición de jefe o no jefe de hogar. El uso de la variable "jefe de hogar" permitiría que coincidieran el número de jefes de hogar y el total de hogares basado en el recuento a 100 por ciento con las estimaciones de hogares basadas en la muestra.

Los grupos de edad propuestos son 10:

0 - 4	25 - 34
5 - 9	35 - 44
10 - 14	45 - 54
15 - 19	55 - 64
20 - 24	65 ó más

Sin embargo, los grupos de edad bajos no se cruzarán por la variable "jefe de hogar". Así habrá menos de 40 celdas dentro de cada municipio para las ponderaciones de personas. La inclusión de variables de clasificación adicionales en la post-estratificación no es aconsejable, no sólo porque aumentaría el ECM en las celdas, sino también porque complicaría el proceso de computación.

En el caso de viviendas, se usaría como variable de post-estratificación el tipo de vivienda y su condición de ocupación. Para el "tipo de vivienda", se recomienda combinar los tipos menos frecuentes con el fin de limitar el número de categorías. En el caso de "condición de ocupación", importaría sólo la categoría "ocupada" puesto que no se tienen planeados cuadros elaborados con las viviendas desocupadas. Por lo tanto, es recomendable calcular la ponderación de razón independiente para las viviendas ocupadas solamente. Las desocupadas compartirían el mismo factor.

En el caso de hogares, las variables aconsejables para la post-estratificación son el tamaño del hogar (digamos tres categorías: 1, 2-5, 6+ personas), y la tenencia de la vivienda (dos categorías: propia ó no).

Estimación de Personas

Dentro de un dominio de estimación, $\hat{Y}_R$, el estimador del total poblacional por el método de razón, (separado para celdas de ponderación, combinado para segmentos) es igual a:

$$(1) \qquad \hat{Y}_R = \sum_{A}^{DOM} X_A \frac{\sum_{h}^{A} w_h y_{Ah}}{\sum_{h}^{A} w_h x_{Ah}}$$

$$= \sum_{A}^{DOM} \sum_{h}^{A} t_{Ah} y_{Ah}$$

donde:

h - segmento

DOM - dominio de estimación, incluye uno ó más municipios

A - post-estrato ó celda de ponderación definido por las variables de clasificación

X_A - número total de personas en la población dentro de la celda de ponderación A, obtenido del recuento a 100%

w_h - $\frac{N_h}{n_h}$ - recíproco de la probabilidad de selección de la sección en el segmento h

y_{Ah} - total muestral (cualitativo ó cuantitativo) para la característica de interés "y", sumado para todas las secciones dentro del segmento h, dentro de la celda de ponderación A

x_{Ah} - total de personas en la muestra, sumado para todas las secciones dentro del segmento h, dentro de la celda de ponderación A

t_{Ah} - ponderación (factor de expansión) de todos los elementos dentro del segmento h; igual a:

$$(2) \qquad t_{Ah} = w_h \frac{X_A}{\sum_h^A w_h x_{Ah}}$$

El primer factor dentro de la ponderación para cada unidad muestral, w_h, es el recíproco de la probabilidad de selección de la sección. En la medida que se mantenga constante el número total de secciones en todos los segmentos y el número de secciones seleccionadas en los segmentos, w_h será constante (igual a 5) de segmento a segmento. En este caso se cancelaría con el w_h que aparece en el denominador, y el estimador quedaría simplemente como:

$$(3) \qquad \hat{Y}_R = \sum_A^{DOM} X_A \frac{\sum_h^A y_{Ah}}{\sum_h^A x_{Ah}}$$

$$= \sum_{A}^{DCM} \sum_{h}^{A} t_A \, y_{Ah}$$

y el factor de ponderación t_A sería:

$$(4) \qquad t_A = \frac{X_A}{\sum_h x_{Ah}}, \quad \text{constante para toda h.}$$

Estimación de Viviendas y de Hogares

El procedimiento para la estimación de viviendas y de hogares, para aquellas variables provenientes de la muestra, es esencialmente el mismo que para la estimación de personas. Como se indicó previamente, deberá existir un factor de ponderación propio para la vivienda y otro para el hogar.

Es decir, se aplicarían también las fórmulas de (1) a (4). La diferencia está en la definición de las variables "x" y "y". La "x" representaría el número de viviendas ocupadas de la muestra, ó el número de hogares de la muestra, según el caso. Paralelamente, X sería el respectivo total poblacional. Asimismo, la variable "y" representaría el número de viviendas, ó de hogares, que poseyeran la característica de interés. La variable "y" también podría ser cuantitativa como en el caso de "ingreso".

Asignación de la Ponderación

El cálculo de la ponderación t_A ó t_{Ah} se debe realizar dentro de cada post-estrato "A" definido según las variables de clasificación y debe estar sujeto a ciertas restricciones. Conlleva los pasos siguientes:

1. Determinación de municipios "autosuficientes" (stand-alone areas):

   El área geográfica donde se efectuará el cálculo de ponderaciones, el área de ponderación, es el municipio. Es necesario especificar un mínimo de observaciones muestrales para calcular la ponderación dentro de un municipio con el fin de controlar el ECM. No existe una fórmula exacta, pero este mínimo se puede determinar de forma arbitraria según la experiencia empírica en otros censos. Por ejemplo, en el censo de Estados Unidos de 1980 (ver Anexo 3) se utilizó 400 como número mínimo de observaciones dentro de un área de ponderación, el "census tract"; en Venezuela en 1971, fue 2,000, y en otros países, ha sido hasta 14,000, dependiendo del tamaño del área.

Tal vez, un valor razonable para el Censo de Venezuela sería un mínimo de 500 observaciones muestrales (personas, viviendas, ú hogares, según el caso) en un municipio. Los municipios que no tuvieran por lo menos 500 observaciones tendrían que consolidarse con otros adyacentes durante el proceso de ponderación. La Unidad Censal de Muestreo podría experimentar con otros valores para tratar de equilibrar dos criterios conflictivos: la necesidad de controlar el ECM y la necesidad de mantener los municipios como entidades separadas para la estimación.

2. Determinación de celdas "autosuficientes" (stand-alone cells) dentro del municipio:

Dentro del municipio, también es necesario establecer requisitos de consolidación de celdas para controlar el ECM. Hay <u>dos</u> controles importantes que efectuar.

a. En primer lugar, en cada celda debe haber un mínimo de observaciones muestrales. Para esto no existe una "ciencia cierta". Por ejemplo, en el censo de Estados Unidos de 1980 el mínimo requerido en una celda fue 10 (ver Anexo 3); en el colombiano de 1989 fue 50. Tal vez, sería lógico usar el valor 30 como mínimo ya que es conocido en la teoría como demarcador mínimo de "muestras pequeñas" vs. "muestras grandes". Así, las celdas de menos de 30 observaciones se consolidarían con otras anexas antes de calcular la ponderación.

   En este caso también, la Unidad Censal de Muestreo podría experimentar con otros valores, analizando cuál sería la cantidad esperada de observaciones en cada celda tomando en cuenta la fracción muestral y el total poblacional en la celda. También tomaría en cuenta la necesidad de conservar la desagregación de ciertos rubros que deben coincidir en los cuadros de publicación. Este último objetivo obviamente entra en conflicto con el de controlar el ECM y se tendrán que equilibrar los dos.

b. En segundo lugar, el factor que aparece a continuación, el cual se relaciona con t_{Ah} ó t_A, debe limitarse para controlar la varianza y además detectar errores operacionales.

   El factor, $\dfrac{X_A}{\sum_h w_h \, x_{Ah}}$, es la razón de dos estimaciones separadas

   de un mismo total poblacional, el numerador por un recuento exhaustivo y el denominador en base a una muestra. Por lo tanto, el factor se debe acercar a 1 (salvo por un pequeño error muestral) en todos los casos. La Unidad Censal de Muestreo sin duda querrá establecer límites de tolerancia alrededor de esta cifra. (Por ejemplo, en el censo de Estados Unidos de 1980, el límite de tolerancia fue 4 para este factor. Ver Anexo 3.)

Si w_h es igual a 5 para toda h, el requisito es que t_A (fórmula 4) deberá acercarse a 5 y que el total poblacional X_A deberá ser más ó menos 5 veces el total muestral en la celda. Con un límite de tolerancia de 4 para el factor mencionado, t_A no podría sobrepasar 20 y el total poblacional no podría exceder 20 veces el valor muestral. Toda vez que en una celda se sobrepasaran estos límites, habría que investigar las causas con el fin de detectar eventuales errores operacionales (tales como la omisión de registros) cuales habría que remediar antes de proseguir con las estimaciones.

Las celdas que cumplan con los dos requisitos establecidos (mínimo de observaciones muestrales y el factor antedescrito manteniéndose dentro de los límites de tolerancia) serán las celdas "autosuficientes" y en ellas, se aplicará el factor de ponderación t_A ó t_{Ah} independientemente.

Aquellas celdas que no cumplan con estos criterios deberán consolidarse (collapse) con otras adyacentes antes de aplicar la ponderación. El patrón a seguir para determinar el orden en que deben c nsolidarse las celdas deberá ser determinado por la Unidad Censal de Muestreo en base a la prioridad que tendrá cada categoría de estimación en las publicaciones. Una buena referencia es el documento preparado por Benjamin Tepping para el Censo Venezolano de 1971, titulado "Proposed Procedure for Estimation Based on the Sample Data for the 1971 Census of Venezuela". Después de efectuar las consolidaciones necesarias, se podrá calcular el factor t_A ó t_{Ah} dentro de la celda consolidada.

3. Cálculo del factor t_{Ah} ó t_A para cada celda A, según la fórmula (2) ó la (4) dentro de las celdas de ponderación definidas por la post-estratificación y la consolidación (collapsing) de áreas y celdas.

4. Eliminación de fracciones para que queden ponderaciones enteras:

   El procedimiento descrito en el informe titulado "Diseño de la Muestra del XII Censo General de Población y Vivienda (Lineamientos Generales)", preparado por la Unidad Censal de Muestreo, es adecuado.

5. Anexación de la ponderación individual t_A ó t_{Ah} en los registros de persona, de vivienda, y de hogar, según corresponda.

IV. PROCESO DE ESTIMACION: CÁLCULO Y PRESENTACION DE ERRORES MUESTRALES

Presentación

Es muy importante presentar los errores de muestreo en la misma publicación donde se presenten las estimaciones de la muestra para dar a conocer su nivel de fiabilidad ó precisión. Con este fin, se pueden escoger de los cuadros a publicarse un grupo representativo de diferentes tipos de estimaciones y para ellas calcular la varianza (ver fórmulas más adelante).

Para las demás estimaciones se puede presentar un cuadro generalizado de errores muestrales tal como el que se ilustra en el Anexo 1, basado en un estimador de proporción en el muestreo aleatorio simple. La ventaja de basar los errores estándar generalizados en este estimador es que se puede calcular fácilmente (por ejemplo, usando una hoja electrónica Lotus) para una gran variedad de valores estimados. Desde luego, este cuadro debe ir acompañado de otro cuadro generalizado de los efectos de diseño (deff) aplicables. Esto es para que se refleje la varianza del estimador complejo que realmente se usó, es decir, que refleje la estratificación, la conglomeración, y el uso del estimador de razón. Siendo que la varianza es una cantidad absoluta, es preferible generalizar el deff, cantidad relativa, que la varianza propiamente.

El error estándar correcto para una estimación muestral de total basada en el diseño complejo sería:

$$\sqrt{\text{deff}} \quad x \quad \sqrt{v_{mas}(\hat{A})} = F_1 \; x \; F_2$$

El factor F_1, la raíz cuadrada del deff, sería el factor a incluir en el segundo cuadro generalizado. El deff se puede calcular en un programa casero usando, en el numerador, la fórmula de varianza definida más adelante en el rubro (6) para el total basado en un estimador de razón y, en el denominador, la fórmula indicada para el total basado en un estimador de proporción. Sería:

$$\text{deff} = \frac{v_{comp}(\hat{Y}_R)}{v_{mas}(\hat{A})}$$

donde: "comp" representa el diseño complejo que se usó y "mas" representa el muestreo aleatorio simple.

El factor F_2 se obtendría en base a la fórmula ilustrada en el Anexo 1 y es la raíz cuadrada de:

$$v_{mas}(\hat{A}) = (1-\frac{n}{N}) \frac{\hat{A}(N-\hat{A})}{n}$$

donde:

$\hat{A} = Np$

p = simple proporción muestral

$n = .20N$

<u>Fórmula para la Varianza de un Total Basado en un Estimador de Razón</u>

Dentro de cualquier dominio de estimación, el estimador de total en una subpoblación C es calculado por:

$$\hat{Y}_{RC} = \sum_{A}^{DOM} \sum_{h}^{A} t_{Ah}\, y_{CAh} \; ; \quad \text{según la fórmula (1)}$$

$$= \sum_{A}^{DOM} X_A\, \hat{R}_{CA}$$

separado para celdas de ponderación y combinado para segmentos.

donde:

$$\hat{R}_{CA} = \frac{\sum_{h}^{A} w_h\, y_{CAh}}{\sum_{h}^{A} w_h\, x_{Ah}}$$

17

Si $N_h = 10$ y $n_h = 2$ para toda h, la razón se simplifica a:

$$\hat{R}_{CA} = \frac{\sum_{h}^{A} y_{CAh}}{\sum_{h}^{A} x_{Ah}}$$

C - un subconjunto de registros (personas, viviendas, ú hogares, según el análisis) que pertenezcan a una subpoblación de interés para el análisis

h - el segmento

A - la celda de ponderación (el post-estrato)

$w_h = \frac{N_h}{n_h}$; recíproco de la probabilidad de selección

y_{CAh} - total muestral de "y" para registros pertenecientes a la subpoblación C dentro de la celda Ah

x_{Ah} - número de unidades muestrales (personas, viviendas, ú hogares) en la celda Ah

otros términos tal como definidos previamente.

La expresión para la verdadera varianza de este estimador es:

$$(5) \quad VAR(\hat{Y}_{RC}) = \sum_{A}^{DOM} [\ VAR(\hat{Y}_{CA}) + R_{CA}^2\ VAR(\hat{X}_A) - 2\ R_{CA}\ COV\ (\hat{X}_A, \hat{Y}_{CA})\ ]$$

donde:

$$\hat{Y}_{CA} = \sum_{h}^{A} w_h\ y_{CAh}$$

$$\hat{X}_A = \sum_h^A w_h \, y_{Ah}$$

Esta varianza se estima por la copia muestral:

$$(6) \quad v(\hat{Y}_{RC}) = \sum_A^{DOM} [\, v(\hat{Y}_{CA}) + \hat{R}_{CA}^2 \; v(\hat{X}_A) - 2\, \hat{R}_{CA} \; \mathrm{cov}\,(\hat{X}_A, \hat{Y}_{CA}) \,]$$

Una opción para el cálculo de la varianza es usar el paquete PC CARP ó el SUPER CARP cuales utilizan el método de los "conglomerados últimos" y producen la varianza de un estimador de razón. Para conseguir la varianza del total, se tendrían que manipular los resultados multiplicando por X_A^2 (dentro del post-estrato A), puesto que:

$$\hat{Y}_{RC} = \sum_A^{DOM} X_A \, \hat{R}_{CA}$$

También habría que manipular los resultados de CARP para calcular el deff ya que aquél producido por PC CARP refleja la eficiencia relativa del <u>mismo estimador de razón</u> en los dos diseños diferentes, no la del estimador de razón relativa a la del estimador de proporción como se desea. Otra opción es elaborar un programa casero desde un principio para calcular todas las fórmulas. En las secciones siguientes, se presentan las fórmulas que corresponden a cada método.

<u>Método de Conglomerados Ultimos</u>

Bajo el método de conglomerados últimos, que es el utilizado por PC CARP y SUPER CARP, los componentes del estimador de varianza en la fórmula (6) se calculan por medio de las siguientes fórmulas.

$$(7) \quad v(\hat{Y}_{CA}) = \sum_h^A \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{CAhi} - \frac{\hat{Y}_{CAh}}{n_h}\right)^2$$

14

donde:

$\hat{Y}_{CAhi} = \sum_{j \in C} \frac{N_h \, y_{hij}}{n_h}$; estimador convencional del total para la subpoblación C en el i-ésimo conglomerado del segmento h.

y_{hij} = valor muestral de la variable "y" en la j-ésima unidad de observación (persona, hogar, ó vivienda) dentro de la i-ésima sección muestral del segmento h. Se supone que la j-ésima unidad pertenece a la subpoblación de interés C; de no ser asi, el valor de y_{hij} se considera como cero.

$\hat{Y}_{CAh} = \sum_{i} \hat{Y}_{CAhi}$; estimador convencional del total para la subpoblación C en el segmento h.

El componente $v(\hat{X}_A)$ se estima también según la fórmula (7), pero omitiendo el sub-indice "C".

$$(8) \quad cov(\hat{X}_A, \hat{Y}_{CA}) = \sum_{h}^{A} \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{X}_{Ahi} - \frac{\hat{X}_{Ah}}{n_h}\right) \left(\hat{Y}_{CAhi} - \frac{\hat{Y}_{CAh}}{n_h}\right)$$

<u>Método Basado en Muestreo Unietápico de Conglomerados</u>

Si no se usa uno de los paquetes CARP y hay que elaborar un programa casero, se pueden simplificar las fórmulas individuales (7) y (8) ya que la muestra dentro del estrato es una muestra unietápica de conglomerados con probabilidades iguales. Las fórmulas apropiadas serían las de muestreo aleatorio simple de secciones usando y_{CAhi}, el total muestral en la sección, como "unidad de observación". Estas nuevas fórmulas se sustituirían dentro de la fórmula (6) para los respectivos componentes.

$$(9) \quad v(\hat{Y}_{CA}) = \sum_{h}^{A} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_{CAh}^2}{n_h}$$

donde:

$$s_{CAh}^2 = \frac{\sum_i^{n_h} y_{CAhi}^2 - \frac{(\sum_i^{n_h} y_{CAhi})^2}{n_h}}{n_h - 1}$$

Si $N_h = 10$ y $n_h = 2$ para toda h, entonces la fórmula (9) se reduce a:

$$(9.1) \qquad v(\hat{Y}_{CA}) = 20 \sum_h^A (y_{CAh1} - y_{CAh2})^2$$

donde "y_{CAh1}" y "y_{CAh2}" son los totales muestrales para la primera y la segunda sección muestral, respectivamente, dentro del segmento h.

El componente $v(\hat{X}_A)$ se estimaría también según la fórmula (9) pero omitiendo el sub-índice "C".

$$(10) \quad cov(\hat{X}_A, \hat{Y}_{CA}) = \sum_h^A \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_{CAh}(X,Y)}{n_h}$$

donde:

$$s_{CAh}(X,Y) = \frac{\sum_i^{n_h} x_{Ahi}\, y_{CAhi} - \frac{(\sum_i^{n_h} x_{Ahi})(\sum_i^{n_h} y_{CAhi})}{n_h}}{n_h - 1}$$

21

Si $N_h = 10$ y $n_h = 2$ para toda h, entonces la fórmula (10) se reduce a:

$$(10.1) \qquad \mathrm{cov}(\hat{X}_A, \hat{Y}_{CA}) = 20 \sum_h^A (x_{Ah1} - x_{Ah2}) (y_{CAh1} - y_{CAh2})$$

En resumen, bajo el supuesto de $N_h = 10$ y $n_h = 2$ para toda h, el proceso de estimación usando las fórmulas de muestreo unietápico de conglomerados para un estimador de razón se reduce a:

$$\hat{Y}_{RC} = \sum_A^{DOM} \sum_h^A t_A \, y_{CAh} \qquad (3)$$

$$\text{donde:} \quad t_A = \frac{X_A}{\sum_h x_{Ah}}, \quad \text{constante para toda h, y anexada ya a cada registro.} \qquad (4)$$

$$v(\hat{Y}_{RC}) = \sum_A^{DOM} [v(\hat{Y}_{CA}) + \hat{R}_{CA}^2 \; v(\hat{X}_A) - 2 \hat{R}_{CA} \; \mathrm{cov} (\hat{X}_A, \hat{Y}_{CA})] \qquad (6)$$

donde:

$$\hat{R}_{CA} = \frac{\sum_h^A y_{CAh}}{\sum_h^A x_{Ah}}$$

$$v(\hat{Y}_{CA}) = 20 \sum_h^A (y_{CAh1} - y_{CAh2})^2 \qquad (9.1)$$

$$v(\hat{X}_A) = 20 \sum_{h}^{A} (x_{Ah1} - x_{Ah2})^2 \qquad (9.1)$$

$$\mathrm{cov}(\hat{X}_A, \hat{Y}_{CA}) = 20 \sum_{h}^{A} (x_{Ah1} - x_{Ah2}) (y_{CAh1} - y_{CAh2}) \qquad (10.1)$$

"y_{CAh1}" y "y_{CAh2}" son los totales muestrales para la primera y la segunda sección muestral, respectivamente, dentro del segmento h.

Fórmula para la Varianza de Estimaciones de Parámetros Relativos

Será tal vez necesario en algunos casos, calcular la varianza para medidas relativas como lo son la media, la proporción, ó la razón. Estos son parámetros que se definen como la razón de dos totales digamos, Y_1 y Y_2. Hay tres posibles casos con diferentes formas de proceder.

Si sólo el total en el numerador es derivado de la muestra mientras que el del denominador proviene del recuento a 100 por ciento, el caso es simple. Un ejemplo sería el caso de la tasa de desempleo, donde el numerador, el número de personas desempleadas, se estimará a partir de la muestra y el denominador, el número de personas económicamente activas, provendrá del recuento exhaustivo. La varianza de una estimación de este tipo es simplemente la varianza del numerador (determinada por las fórmulas indicadas previamente en el rubro (6)) multiplicada por la constante:

$$\left[\frac{1}{\text{denominador}^2} \right]$$

Otra posibilidad es que sólo el total en el denominador sea derivado de la muestra mientras que el del numerador proviene del recuento exhaustivo. También este es un caso simple. La varianza de la estimación será la varianza del denominador (determinada por las fórmulas indicadas previamente en el rubro (6)) multiplicada por la constante:

$$\left[\frac{\text{numerador}^2}{\text{denominador}^4} \right]$$

Sin embargo, cuando los totales tanto en el numerador como en el denominador son estimados a partir de la muestra, la fórmula se vuelve muy compleja pues involucra la razón de dos razones (una doble razón). Puesto que sin duda serán pocos estos casos, se recomienda simplificar el cálculo de la varianza usando la fórmula de varianza estándar para una razón sencilla pero manteniendo la ponderación t_A ó t_{Ah}.

Sea $\hat{R}_R = \dfrac{\hat{Y}_2}{\hat{Y}_1}$; donde $\hat{Y}$ se calcula por la fórmula (1) ó (3) e incluye el factor de ponderación t_A

Entonces:

$$v(\hat{R}_R) = \frac{1}{\hat{Y}_1^2} \left[v(\hat{Y}_2) + \hat{R}_R^2 \, v(\hat{Y}_1) - 2 \hat{R}_R \, \mathrm{cov}(\hat{Y}_1, \hat{Y}_2) \right]$$

Claro está, el resultado será una subestimación no-cuantificada de la verdadera varianza, lo cual habría que mencionar en el informe metodológico.

IV. ENCUESTA EVALUATIVA

La utilidad de los datos censales es limitada si su exactitud es desconocida. Por esta razón, muchos países conducen una encuesta evaluativa ("post-enumeration survey") después del censo para evaluar la calidad de los datos censales.

- Una encuesta evaluativa proporciona medidas cuantitativas de la exactitud de los datos censales en términos de cobertura y contenido.

- También provee una base para ajustes que permitan corregir sesgos sistemáticos que de otra forma afectarían las proyecciones de población y otros análisis.

- Permite la evaluación de la calidad del marco muestral desarrollado en base al censo y que ha de usarse para las encuestas de hogares intercensales.

En realidad, la encuesta evaluativa debe diseñarse en la fase inicial de preparación del censo y debe probarse durante la fase de censo experimental. Esto permite, además de la identificación de mejoras necesarias en la metodología de la encuesta evaluativa en sí, una evaluación más efectiva de la metodología propuesta para el mismo censo permitiendo así revisiones en ella.

Se recomienda que para el Censo de Población y Vivienda de Venezuela la planeación detallada de la encuesta evaluativa comience cuanto antes y que se haga una prueba cabal de su metodología en el campo antes de ejecutarla.

Es primordial que el personal asignado a la planeación y a la ejecución de la encuesta evaluativa sea ajeno a las operaciones del empadronamiento censal. Esto es cierto tanto para el personal técnico como para el personal de campo. Esta independencia operacional es vital para la validez de los resultados de la encuesta evaluativa, cuales se deben obtener con lo más mínimo posible de influencia (sesgo) de los resultados del empadronamiento censal.

ANEXO 1

ERROR ESTANDAR PARA ESTIMACION DE TOTAL CUALITATIVO EN EL MUESTREO ALEATORIO SIMPLE

ERROR ESTANDAR PARA ESTIMACIONES DE TOTALES CUALITATIVOS
EN UNA MUESTRA ALEATORIA SIMPLE DE 20%

TOTAL ESTIMADO $\hat{A}$	TAMAÑO DE LA POBLACION, N					
	50,000	100,000	250,000	500,000	750,000	1,000,000
50	14	14	14	14	14	14
100	20	20	20	20	20	20
250	32	32	32	32	32	32
500	44	45	45	45	45	45
1,000	63	63	63	63	63	63
2,500	97	99	99	100	100	100
5,000	134	138	140	141	141	141
10,000	179	190	196	198	199	199
15,000	205	226	237	241	242	243
25,000	224	274	300	308	311	312
75,000	-	274	458	505	520	527
100,000	-	-	490	566	589	600
250,000	-	-	-	707	816	866
500,000	-	-	-	-	816	1000
750,000	-	-	-	-	-	866
1,000,000	-	-	-	-	-	-

Este cuadro puede ampliarse para incluir otros valores de N, menores ó mayores.

El error estándar es la raíz cuadrada de:

$$v(\hat{A}) = (1-\frac{n}{N})\frac{\hat{A}(N-\hat{A})}{n}$$

donde: $n = .20N$

COEFICIENTES DE VARIACION PARA ESTIMACIONES DE TOTALES CUALITATIVOS
EN UNA MUESTRA ALEATORIA SIMPLE DE 20%

TOTAL ESTIMADO $\hat{A}$	TAMAÑO DE LA POBLACION, N					
	50,000	100,000	250,000	500,000	750,000	1,000,000
50	28.3%	28.3%	28.3%	28.3%	28.3%	28.3%
100	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%
250	12.6%	12.6%	12.6%	12.6%	12.6%	12.6%
500	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%
1,000	6.3%	6.3%	6.3%	6.3%	6.3%	6.3%
2,500	3.9%	3.9%	4.0%	4.0%	4.0%	4.0%
5,000	2.7%	2.8%	2.8%	2.8%	2.8%	2.8%
10,000	1.8%	1.9%	2.0%	2.0%	2.0%	2.0%
15,000	1.4%	1.5%	1.6%	1.6%	1.6%	1.6%
25,000	0.9%	1.1%	1.2%	1.2%	1.2%	1.2%
75,000	-	0.4%	0.6%	0.7%	0.7%	0.7%
100,000	-	0.0%	0.5%	0.6%	0.6%	0.6%
250,000	-	-	0.0%	0.3%	0.3%	0.3%
500,000	-	-	-	0.0%	0.2%	0.2%
750,000	-	-	-	-	0.0%	0.1%
1,000,000	-	-	-	-	-	0.0%

Este cuadro puede ampliarse para incluir otros valores de N, menores ó mayores.

ANEXO 2

COMENTARIOS GENERALES SOBRE EL PLAN DE MUESTREO PARA LA MUESTRA CENSAL

COMENTARIOS GENERALES SOBRE EL PLAN DE MUESTREO PARA LA MUESTRA CENSAL

Una de las cualidades importantes de un plan muestral es la medida en que pueda producir estimaciones que se acerquen al verdadero valor, es decir, estimaciones exactas. La exactitud requiere un error total bajo. Esto incluye un error muestral bajo y un error no-muestral también bajo. La Unidad Censal de Muestreo realizó una labor extraordinaria y cuidadosa en preparar este plan y tomó en cuenta las varias consideraciones teóricas y prácticas necesarias para lograr un buen diseño equilibrando ambos tipos de errores. Pero ciertos requisitos y ciertas dificultades que tuvieron que enfrentar limitaron las opciones posibles. Cabe hacer unos breves comentarios para referencia futura.

El diseño más aconsejable para una muestra censal, porque es simple y a la vez produce un error total bajo, es el de muestreo sistemático dentro de cada área de empadronamiento (equivalente a un segmento). El plan muestral escogido es más complejo y menos eficiente que esta alternativa. Dentro del contexto de un censo, que es una operación de gran escala, es aconsejable simplificar la formación de áreas de empadronamiento y también el proceso de estimación. Las tres dificultades principales que no permitieron esta opción y los comentarios al respecto son discutidas a continuación.

- Primera dificultad: la necesidad de tener empadronadores especializados en el cuestionario ampliado lo cual es difícil cuando son voluntarios.

  Comentario: el nivel de preparación de los empadronadores se puede mejorar si se tienen los recursos para intensificar la capacitación para todos.

- Segunda dificultad: la preocupación por el sesgo de predominancia de hogares pequeños que se notó en el censo de 1971, el cual se atribuyó al proceso de selección de la muestra.

  Comentario: es posible controlar el sesgo de selección mediante procedimientos apropiados de control de calidad en el campo. El sesgo restante no necesariamente impide que las estimaciones tengan un error total bajo si se toma en cuenta que el error muestral que produce el muestreo sistemático es menor que el que produce el muestreo por conglomerados. Por otro lado, es importante señalar que, en otros países, por ejemplo, en Filipinas, se ha podido comprobar mediante análisis posteriores que el sesgo de predominancia de hogares pequeños no estuvo en realidad en la selección de la muestra, aunque así se supuso al principio. Más bien, fue resultado de una sistemática omisión de personas dentro de hogares grandes, tal vez para reducir la carga de trabajo, que fue mayor en las viviendas de la muestra quizás porque el cuestionario era más largo. Este tipo de sesgo no es intrínseco al plan muestral sistemático de viviendas y, por lo tanto, puede repetirse bajo cualquier plan muestral y puede suceder en el empadronamiento exhaustivo.

Tercera dificultad: un requisito administrativo que el empadronamiento debe efectuarse en un solo día lo que significa usar áreas de trabajo pequeñas de 20 viviendas para un cuestionario corto y 10 viviendas para un cuestionario ampliado.

Comentario: operacionalmente, es imposible usar el segmento como área de empadronamiento si el trabajo se tiene que completar en un día. Así es que bajo este requisito no se puede usar el plan de muestreo sistemático dentro del segmento y hay que optar por el de conglomerados pequeños. Pero es bueno señalar también que la realización de las operaciones de empadronamiento en un solo día hará más difícil los controles de calidad y los controles operacionales tanto para la muestra como para el empadronamiento a 100 por ciento. Además, este requisito requirió una operación cartográfica mucho más laboriosa porque conlleva la formación de áreas censales a tres niveles en lugar de uno: primero, la conformación de segmentos, luego la división de segmentos en secciones correspondientes a áreas de empadronamiento, y finalmente, la división de secciones en subsecciones para la aplicación del cuestionario ampliado.

ANEXO 3

REFERENCIAS SOBRE EL PROCEDIMIENTO DE PONDERACIÓN UTILIZADO EN EL CENSO DE ESTADOS UNIDOS DE 1980

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

1980 DECENNIAL CENSUS EVALUATION AND RESEARCH MEMORANDA
SERIES SMD Number 37-18

2

April 21, 1982

MEMORANDUM FOR Peter A. Bounpane, Acting Chief
Decennial Census Division

From: Charles D. Jones, Chief
Statistical Methods Division

Subject: "Final" Collapsing Criteria for the 1980 Census Full Sample Weighting Matrices

The criteria for collapsing the 1980 census full sample weighting matrices have undergone several modifications. This memorandum specifies the collapsing criteria which are considered "final" at this time. These specifications supersede those 1/ concerning the collapsing criteria which were contained in the earlier memoranda.

As was specified in the previous memoranda on this issue, the full sample weighting matrices must be collapsed prior to the iterative adjustment steps. The row, column or householder/non-householder marginal categories must meet the following conditions:

1. The uninflated marginal sample count for the rows, columns or householder/non-householder categories must be at least 10 for the persons weighting matrix and 5 for the housing unit weighting matrices;

2. The 100-percent census marginal count for the race and Spanish origin categories must be at least 150 for the persons weighting matrix and 50 for the occupied housing unit matrix;

3. The 100-percent census marginal count must be larger than the corresponding unweighted sample count; and

1/ See Section II.B.3 of 1980 Census Memorandum Series-SMD, Memorandum for Gerald J. Post from Charles D. Jones, Subject: "Memorandum No. 37-1 Final Specifications for the 1980 Census Ratio Estimation and Sample Weighting for Population and Housing," May 19, 1980. Also see "Full-Sample Collapsing Criteria," on page 3 of Memorandum for Distribution List from Henry F. Woltman, Subject: "Decisions Concerning 1980 Census Sample Weighting Procedures," December 29, 1981.

4. The ratio of the 100-percent marginal count to the initial inflated marginal sample count must be greater than or equal to the weighting area observed sampling rate 2/ and less than, "initially", 4. However, this maximum ratio of 4 will be iteratively lowered to limit the size of the weight to at most 48 3/ (or possibly 49 due to the random rounding).

The iterative collapsing procedure is carried out as follows:

(1) After the initial stage of collapsing the weighting matrix, the iterative ratio adjustment (raking) is performed for a particular weighting area;

(2) The maximum ratio of the final adjusted to the initial inflated count is obtained over all the interior cells of the weighting matrix for the weighting area;

3) If the product of this ratio and the largest initial weight for the weighting area is larger than 48, then the maximum ratio collapsing criterion is lowered from 4 to 3;

(4) The weighting matrix undergoes additional collapsing under the new criterion of 3, and the iterative ratio adjustment is then performed again;

(5) The maximum ratio described in (2) is then obtained. If the product of this new maximum interior cell ratio and the largest initial weight is still larger then 48, the above process is repeated this time with a maximum ratio collapsing criterion of 2.6.

(6) This process is repeated, as necessary, until the product of the maximum interior cell ratio and the largest initial weight is less than 48. Each additional repetition is performed by lowering the maximum ratio collapsing criterion by 0.4 (e.g., 2.6, 2.2, 1.8...). Collapsing to a single cell will be performed if necessary. (It should be noted that weighting areas are formed so that it is possible to achieve a maximum weight of 48.)

2/ The observed sampling rate is simply the ratio of the uninflated sample count to the 100-percent census count.

3/ The maximum weight has been changed from 63. For the latter, see 1980 Census Memorandum Series-SMD, Memorandum for Gerald J. Post from Charles D. Jones, Subject: "Memorandum No. 37-2: Supplement to Memorandum No. 37-1 Maximum Weight for the 1980 Census Sample Record, "October 29, 1980.

Collapsing proceeds as specified in the previous documentation 4/ until all of the above conditions are satisfied.

4/ This documentation is described in footnotes 1 and 3.

cc: Standard Distribution List
J. Thompson (SMD)
T. Harahush "
M. Fan "
J. Kim "
H. Yamauchi "

SMD:JKim:llp

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

(3)

December 29, 1981

MEMORANDUM FOR Distribution List

From: Henry F. Woltman, Chief H.W.
Census Operations Branch
Statistical Methods Division

Subject: Decisions Concerning 1980 Census Sample Weighting Procedures

This memorandum documents several decisions concerning the 1980 census sample weighting procedures which were reached during a joint meeting between SMD, POP, and DCD staff members held on December 15, 1981. The purpose of this meeting was to discuss the results of implementing the weighting procedures on the sample files available from the three test states - Montana, Rhode Island, and Delaware. The decisions reached concerned the revised method of collapsing the race/origin controls to be used for the full sample weighting, and a modified procedure for weighting the POW/MIG half-sample.

Background

As background, the basic findings of the test state weighting will be briefly discussed. A more detailed discussion of the results will appear in a forthcoming memorandum.

For the full sample weighting, eight methods were tested by varying the definition of race/origin control categories, the criteria used to test the categories to determine if collapsing was necessary and the scheme by which the categories that did not meet these criteria were combined. The eight variations tested were as follows:

(1) The race/origin controls consist of the five major races - "White," "Black," "Asian," "Indian (American)," and "Other," with Spanish and non-Spanish by race. The collapsing test criteria were a minimum of 10 sample persons (unweighted) and a ratio test involving the ratio of the 100-percent count to the initial weighted sample count for each category. Within each weighting area, the collapsing scheme always combined the "Other" race category with "White." The remaining failed races were first grouped together, and if the grouped races still failed, they were combined with the smallest race category that passed in the weighting area. Origin was then tested using the same criteria, within the remaining race categories.

(2) The same race/origin controls and test criteria as for (1). Within each weighting area, the collapsing procedure was to collapse "Other" into the largest specific race category. The remaining failed races were handled as in (1).

(3) The same race/origin controls and test criteria as for (1). However, "Other" went to White as did all other failed race categories.

(4) Black and non-Black cross classified by Spanish and non-Spanish were the only race/origin categories tested. The test criteria were the same as for (1).

(5) Black and non-Black were the only race/origin controls used (no origin). The test criteria were the same as for (1).

(6) Spanish and non-Spanish were the only race/origin controls used (no race). The test criteria were the same as for (1).

(7) No race or origin categories were used as controls.

(8) The same race/origin categories and collapsing procedures as (3). However the collapsing test criteria were changed to require at least 10 sample persons (unweighted) and 150 100-percent persons (in addition to the ratio criteria) in order for a race/origin category to pass, i.e., to allow the race/origin category to stand as a control.

Methods (1), (2) and (3) were each used to weight the sample edited detail files from Montana and Rhode Island. After examining the weighted sample counts that resulted from these methods, it became apparent that Asians and Spanish were being underestimated consistently, and that American Indians were being overestimated. Methods (4), (5), (6) and (7) were then used to weight the sample from Montana and Rhode Island. Upon examining the results of these procedures, it was found that using only the observed sample count as a test criteria for collapsing race/origin could lead to a downward bias in estimates for those race/origin categories which were sparsely scattered throughout a state or region. This observation lead to the development of method (8). Method (8) was then tested on Montana, Rhode Island and Delaware. The results of this weighting procedure were satisfactory to all concerned.

An analysis of the POW/MIG subsample weighting was only possible in Delaware, since Montana and Rhode Island had almost entirely passed through the POW/MIG coding operation before the subsampling operations were instituted. Six methods of performing this weighting were tested on the Delaware POW/MIG subsample as follows:

(1) The specified procedure where the 1980 census sample weights for persons in the POW/MIG subsample were initially inflated by the inverse of the POW/MIG coding rate for the ED's to which they belonged. The initially inflated weights then underwent a "cell-by-cell" ratio adjustment for race/origin, age and sex. The race/origin

37

categories, collapsing procedures, and collapsing test criteria were the same as for method (8) for the full sample weighting (a more conservative ratio criterion was used as a collapsing test but this criterion was rarely invoked).

(2) The collapsing test criteria for method (1) were modified to require a minimum of 5 sample persons (unweighted) and at least 150 persons in the weighted full sample count before a race/origin category was permitted to stand by itself. Furthermore, the initial weight for each POW/MIG subsample person was limited to be no larger than 48.

(3) Procedure (2) was applied with an unlimited initial weight.

(4) No race/origin controls were used for the cell-by-cell adjustment.

(5) The initial weights as described for (1) only underwent a single cell adjustment to the 100 percent count of total persons in each weighting area.

(6) The strict unbiased weight (i.e., either six or two) was used for each person as an initial weight. These "unbiased" weights also underwent a single cell adjustment to the 100 percent count of total persons in each weighting area.

Method (1) above, was implemented first for the POW/MIG weighting for Delaware. In reviewing the counts that resulted for the weighted POW/MIG subsample, it was found that Spanish were extremely low, and that American Indians were rather high. Methods (2) - (6) were then each applied to weight the POW/MIG sample for Delaware. Method (5) was found to provide the "best" results.

Decisions Concerning the 1980 Census Sample Weighting

Based on the discussion of these results, the following decisions were reached concerning the procedures to be used for production weighting of the full sample and the POW/MIG half-sample.

The three test states as well as the states of Connecticut, North Carolina and Arizona will be run in a production mode using the following collapsing criteria for the full sample and the revised weighting procedure for the POW/MIG subsample.

1. Full sample collapsing criteria

   a. For population data, the collapsing method using a minimum sample count of 10 persons and a minimum 100 percent count of 150 persons for a particular race, spanish/nonspanish, age, or sex cell will be used. Further the "Other" race category will be collapsed with the "White" race category as will any other specific race category that fails either or both of the stated criteria. For occupied housing unit data the criteria will be 5 sample housing units and 50 100 percent housing units for the race, age, sex, spanish/nonspanish householder cells of the weighting matrix.

76

b. After examining the results for the three additional states, two additional collapsing methods will be investigated: (1) collapsing the "Other" race group to the largest specific race group (rather than always collapsing with white) if there are sufficient weighting areas in a state to permit this evaluation; and (2) using no race or spanish origin controls. The weights assigned to the sample for these states from the production runs will only be changed, however, if one of the above procedures results in significantly better agreement between the weighted sample and complete count figures for race and/or spanish origin.

2. POW/MIG subsample weighting procedure

The subsample for POW/MIG will be weighted using the product of the full sample weight, the inverse of the ED POW/MIG coding rate and a single cell ratio adjustment to the weighting area total population count. Again, testing of alternative weighting procedures will be undertaken only in the event unanticipated differences between the weighted sample and complete count figures for race and/or spanish origin exist.

cc:

P. Bounpane	(DIR)		P. Berman	(POP)
B. Bailar	"		A. Young	(HOUS)
K. Wolter	"		W. Downs	"
R. Lepage	(DCD)		M. Garland	(DUSD)
S. Goldman	"		P. Zeisset	"
D. Dalzell	"		C. Jones	(SMD)
D. Dwyer	"		S. Miskura	"
D. Sexton	"		T. Harahush	"
R. Herriot	(POP)		J. Thompson	"
P. Schneider	"		J. Kim	"
N. McKinney	"		M. Fan	"
E. Fernandez	"		H. Yamauchi	"
A. Cresce	"			

SMD:JThompson:llp

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

May 11, 1989

MEMORANDUM FOR Henry Woltman
Assistant Division Chief for Census Design
Statistical Support Division

From: Rick Griffin R.G.
Chief, Census Design Branch
Statistical Support Division

Subject: Maximum Weights - 1990 Census

BACKGROUND

The iterative proportional fitting (raking) weighting procedure to be used for the 1990 Census will result in differential weights. Raking is, of course, beneficial as it improves the mean square error of estimates and adjusts for sampling biases. However, it is also well known that differential weights can result in an increase in variance. The benefits of raking are clear but the ratio of weights should be controlled. For the 1980 Census a maximum weight of 48 was used. This was four times the maximum initial weight that was allowed. The purpose of this memorandum is to provide a statistical analysis that will result in the recommendation of a procedure for determining maximum weights for the 1990 Census. An application of the recommendation for the Dress Rehearsal will be given.

ANALYSIS FOR 1990

Consider a weighting area for which a subweighting area estimate for a characteristic across three interior cells or margins of the weighting matrix is desired. Assuming the sample in the weighting area is a simple random sample of sampling fraction f, the sampling fraction for any cell is a random variable of expectation f. Suppose two of the cells have a sampling fraction of f and these two cells have the same 100 percent count. The third cell has a sampling fraction of r ($r < f$). Assuming no collapsing, the two equal size cells will have a weight of f^{-1} and the other cell will have a weight of r^{-1}. If r^{-1} is higher than the maximum weight, collapsing is performed. Assume the cell with sampling fraction r is collapsed with one of the two cells with sampling fraction f. The resulting new weight is a function of r, f and w, where w is the proportion of the total 100 percent count in the original three cells that is in each of the two equal size cells with sampling fraction f. In the Appendix, a formula for the new weight due to collapsing and a formula for the ratio of the variance for the estimate with no collapsing to the variance

40

for the estimate with collapsing is given (Note that in the Appendix $w = w_1 = w_3$). Equal variances in the three cells is assumed. Collapsing indicates that r^{-1} is larger than the maximum allowed weight.

Tables 1-3 give values of the new weight if collapsing is performed and the ratio of the variances (no collapsing to collapsing) for f values of .5, .167 and .125 respectively. Values of w range between .3 and .48. Note that the cell with a sampling fraction of r has a proportion of the total 100 percent count of the three cells equal to 1-2w. For each line of each table the r value is also given. If the r value is chosen as the maximum weight, then the estimate without collapsing will be used and the variance ratio shown in the table indicates the loss in reliability compared with collapsing. Remember that we do not want to always collapse since more cells has mean square error advantages and adjusts for sampling biases. For example, consider Table 1 for a weighting area with a sampling fraction of 1-in-2. For a w value of .45 (the two cells with sampling fraction f have 90% of the total 100 percent count in the original three cells), a maximum weight of 12 will allow a variance ratio of about 1.68 and a increase in standard error of about 30 percent. A similar analysis using Table 2 for a weighting area with a sampling fraction of 1-in-6 shows that a maximum weight of 48 will allow a increase in standard error of about 28 percent. Table 3 shows that for a weighting area with a sampling fraction of 1-in-8, a maximum weight of 64 will allow a increase in standard error of about 27 percent. Note that for lower w values the maximum weight for the same increase in standard error is lower. However, the larger the size of the cell with the sampling fraction r the less likely that the differential between r and f is significant. Thus it seems reasonable to base our decision on the case where the cell with sampling fraction r has 10 percent of the total 100 percent count of the three cells (w = .45).

Also provided are three graphs, one for each of the three 1990 Census designated sampling fractions of 1-in-2, 1-in-6 and 1-in-8. Each of these graphs shows the data in the associated table (based on the weighting area sampling fraction). The horizontal axis is the allowed maximum weight and the vertical axis is the resulting variance ratio. One line for each of the w values is shown.

Initial weighting areas have one designated sampling rate (1-in-2, 1-in-6 or 1-in-8), however if any initial weighting area has less than 400 sample persons it is combined with another initial weighting area so that no final weighting areas have less than 400 sample persons. As a result, final weighting areas can have a mixture of the three designated sampling rates. In addition, due to various nonsampling errors and the

fact that the census sample is a sample of housing units not persons the "observed" sampling fraction can be different than the designated sampling fraction. Thus, using the assumptions and methodology of this memorandum the recommendation for 1990 is to base the maximum weight for each weighting area on the overall observed sampling fraction of the weighting area according to the following:

Observed Weighting Area Sampling Fraction (f)	Maximum Weight
f less than or equal to 14.5%	64
14.5% < f < 34%	48
f greater than or equal to 34%	12

Note that since weights in the weighting area can be less than f^{-1}, the differential weight factor can be greater than (Max Weight)/f. However, further decrease in the maximum weight would limit the benefits of raking.

Application for the Dress Rehearsal

The above analysis was done for the 1990 Census designated sampling rates of 1/2, 1/6 and 1/8. The Dress rehearsal designated sampling rates were 1/2, 1/6, 1/10 and 1/20. Due to the smaller sampling fractions in some areas, larger maximum weights should be allowed. The recommendation is to use the same procedure as shown above for 1990 except that for observed sampling fractions less than 7.5% use a maximum weight of 160.

STSD Planning Group

cc: J. Thompson (STSD)
A. Navarro "
C. Swan "
J. Golmont "
S. Pak "

STSD:RGriffin

APPENDIX

Suppose a population is divided into 3 strata with common variance for the characteristic of interest. The proportions of the population in each of the strata are w_1, w_2 and w_3. Also let $w_1 = w_3$. The sampling fractions in the three strata are f, r and f respectively.

Consider two estimates of the population total. Assume simple random sampling in each strata or collapsed strata.

(1) Estimate 1, X_1, is the usual stratified estimator of total with the weight in each of the three strata equal to the inverse of the sampling fraction.

(2) Estimate 2, X_2, is formed by first collapsing strata 2 and 3 into one stratum and assuming the combined sample is a random sample from the collapsed stratum. The resulting sampling fraction in the collapsed stratum is denoted z. The estimator is the stratified estimator across stratum 1 and the collapsed stratum. The weights are the inverse of f and the inverse of z respectively.

Remembering that $w_1 = w_3$ we have

$$z = \frac{rw_2 + fw_1}{w_2 + w_1} = \frac{r(1-2w_1) + fw_1}{1-w_1}$$

$$\frac{\text{Var } (X_1)}{\text{Var } (X_2)} = \frac{2w_1(1/f-1) + (1-2w_1)(1/r-1)}{w_1(1/f-1) + (1-w_1)(1/z-1)}$$

TABLE 1

f = 0.500

W	Max Weight (r)	New Weight	Variance Ratio
0.48	10	2.13	1.24
0.48	16	2.14	1.45
0.48	20	2.15	1.60
0.48	25	2.15	1.78
0.48	30	2.15	1.96
0.45	10	2.34	1.52
0.45	16	2.38	1.99
0.45	20	2.39	2.30
0.45	25	2.40	2.70
0.45	30	2.41	3.10
0.43	10	2.49	1.66
0.43	16	2.55	2.26
0.43	20	2.57	2.66
0.43	25	2.58	3.17
0.43	30	2.59	3.67
0.4	10	2.73	1.81
0.4	16	2.82	2.54
0.4	20	2.86	3.04
0.4	25	2.88	3.66
0.4	30	2.90	4.28
0.38	10	2.90	1.88
0.38	16	3.02	2.67
0.38	20	3.07	3.20
0.38	25	3.11	3.87
0.38	30	3.13	4.54
0.3	10	3.68	1.93
0.3	16	4.00	2.75
0.3	20	4.12	3.30
0.3	25	4.22	4.00
0.3	30	4.29	4.69

TABLE 2

f = 0.167

W	Max Weight (r)	New Weight	Variance Ratio
0.48	10	6.19	1.01
0.48	30	6.39	1.15
0.48	48	6.43	1.28
0.48	64	6.45	1.40
0.48	90	6.46	1.60
0.45	10	6.47	1.03
0.45	30	7.02	1.33
0.45	48	7.14	1.64
0.45	64	7.18	1.91
0.45	90	7.23	2.36
0.43	10	6.65	1.03
0.43	30	7.47	1.43
0.43	48	7.64	1.83
0.43	64	7.72	2.19
0.43	90	7.78	2.79
0.4	10	6.92	1.04
0.4	30	8.18	1.55
0.4	48	8.47	2.07
0.4	64	8.60	2.53
0.4	90	8.71	3.29
0.38	10	7.10	1.05
0.38	30	8.69	1.61
0.38	48	9.07	2.18
0.38	64	9.24	2.70
0.38	90	9.39	3.54
0.3	10	7.78	1.06
0.3	30	11.05	1.71
0.3	48	12.00	2.37
0.3	64	12.44	2.96
0.3	90	12.86	3.94

TABLE 3

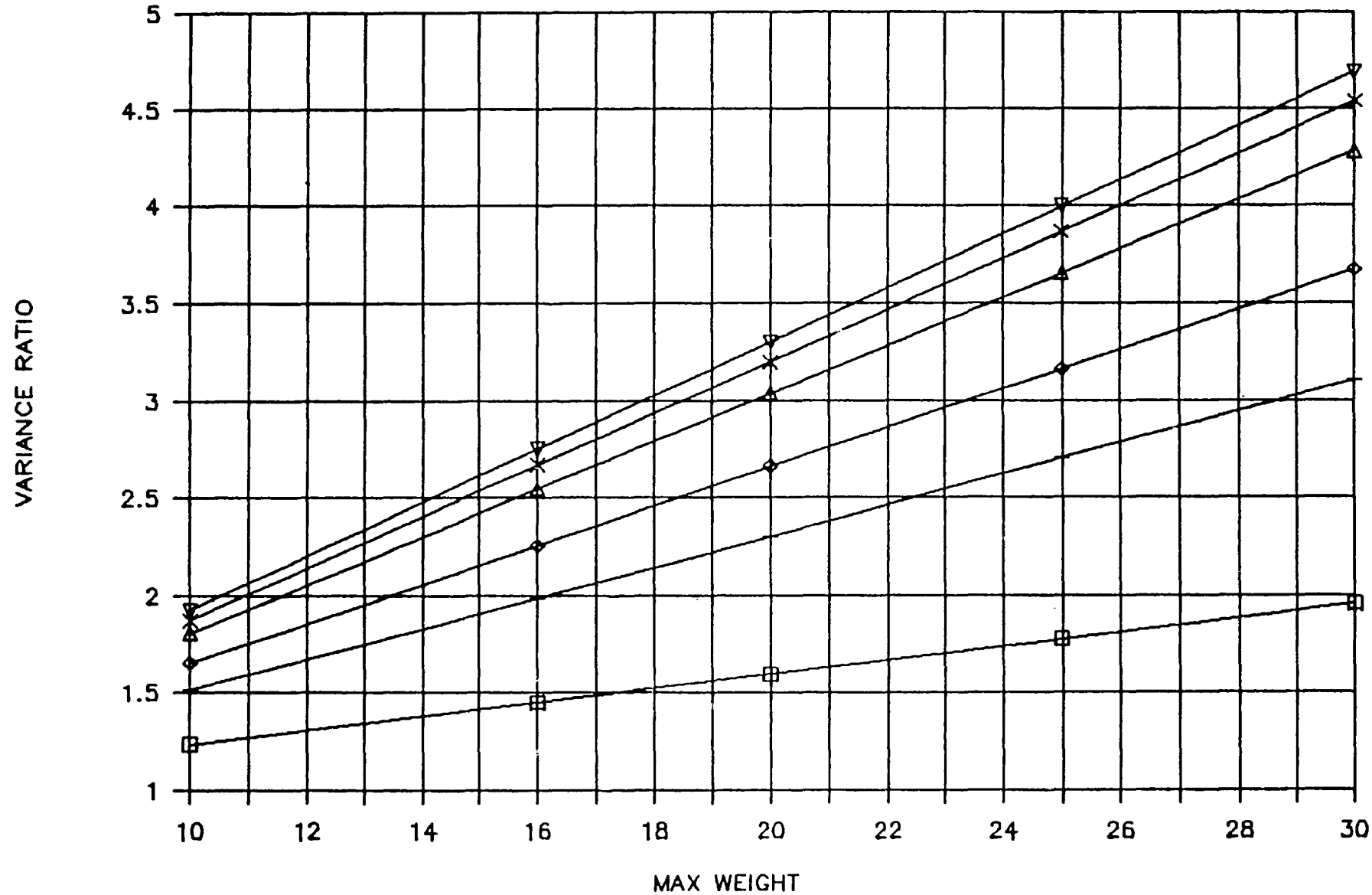
f = 0.125

M	Max Weight (r)	New Weight	Variance Ratio
0.48	10	8.13	1.00
0.48	30	8.48	1.09
0.48	48	8.55	1.18
0.48	64	8.58	1.27
0.48	90	8.60	1.41
0.45	10	8.30	1.00
0.45	30	9.23	1.20
0.45	48	9.43	1.41
0.45	64	9.51	1.61
0.45	90	9.59	1.93
0.43	10	8.41	1.01
0.43	30	9.76	1.26
0.43	48	10.06	1.54
0.43	64	10.19	1.80
0.43	90	10.31	2.22
0.4	10	8.57	1.01
0.4	30	10.59	1.33
0.4	48	11.08	1.70
0.4	64	11.29	2.03
0.4	90	11.49	2.57
0.38	10	8.67	1.01
0.38	30	11.17	1.37
0.38	48	11.81	1.77
0.38	64	12.10	2.14
0.38	90	12.36	2.75
0.3	10	9.03	1.01
0.3	30	13.77	1.43
0.3	48	15.27	1.90
0.3	64	16.00	2.33
0.3	90	16.69	3.04

VARIANCE RATIO

$f = 1/2$

W is % in each large cell



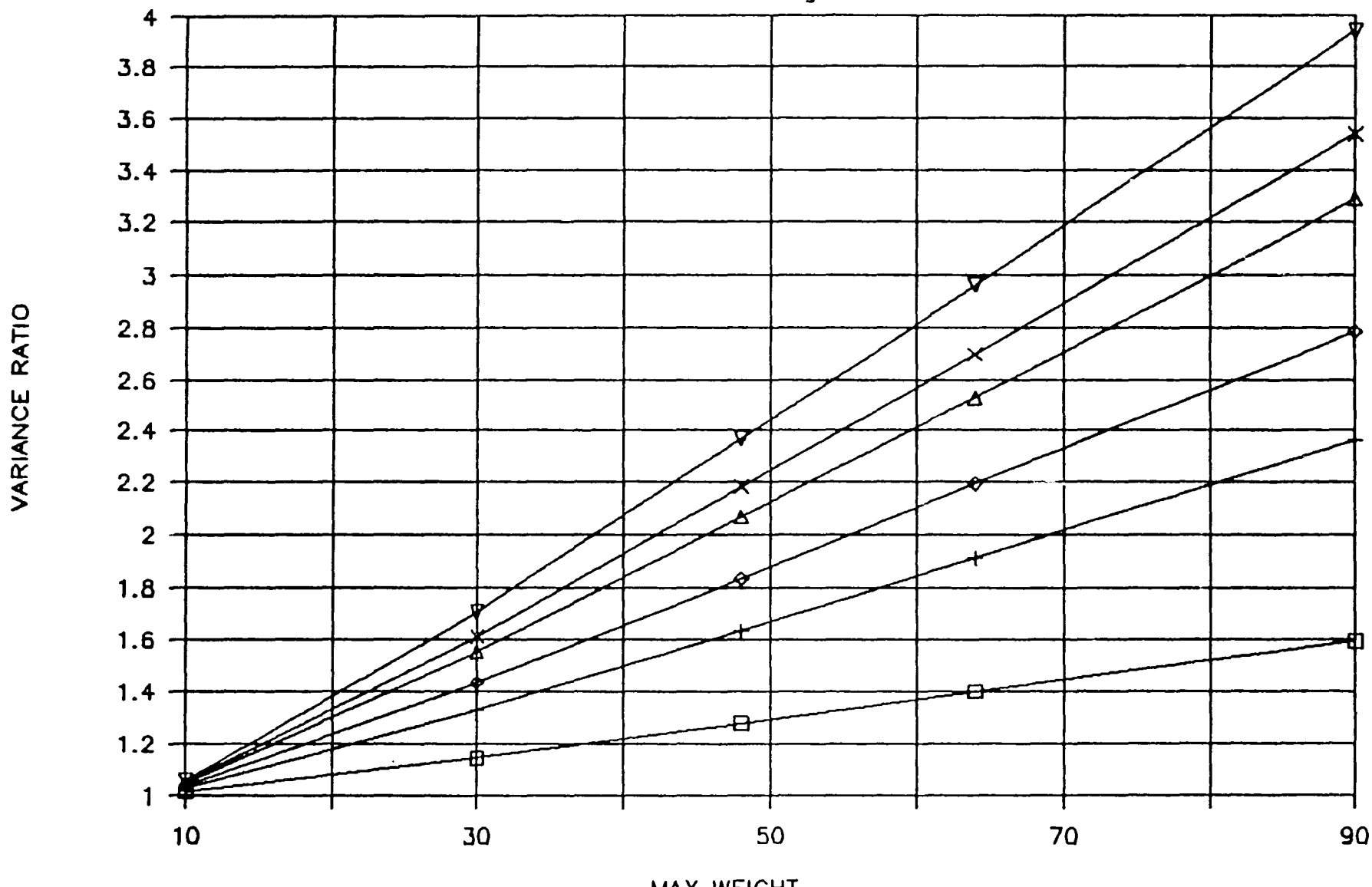
+ W = 45% ◇ W = 43% △ W = 40% × W = 38%

□ W = 48% ▽ W = 30%

VARIANCE RATIO

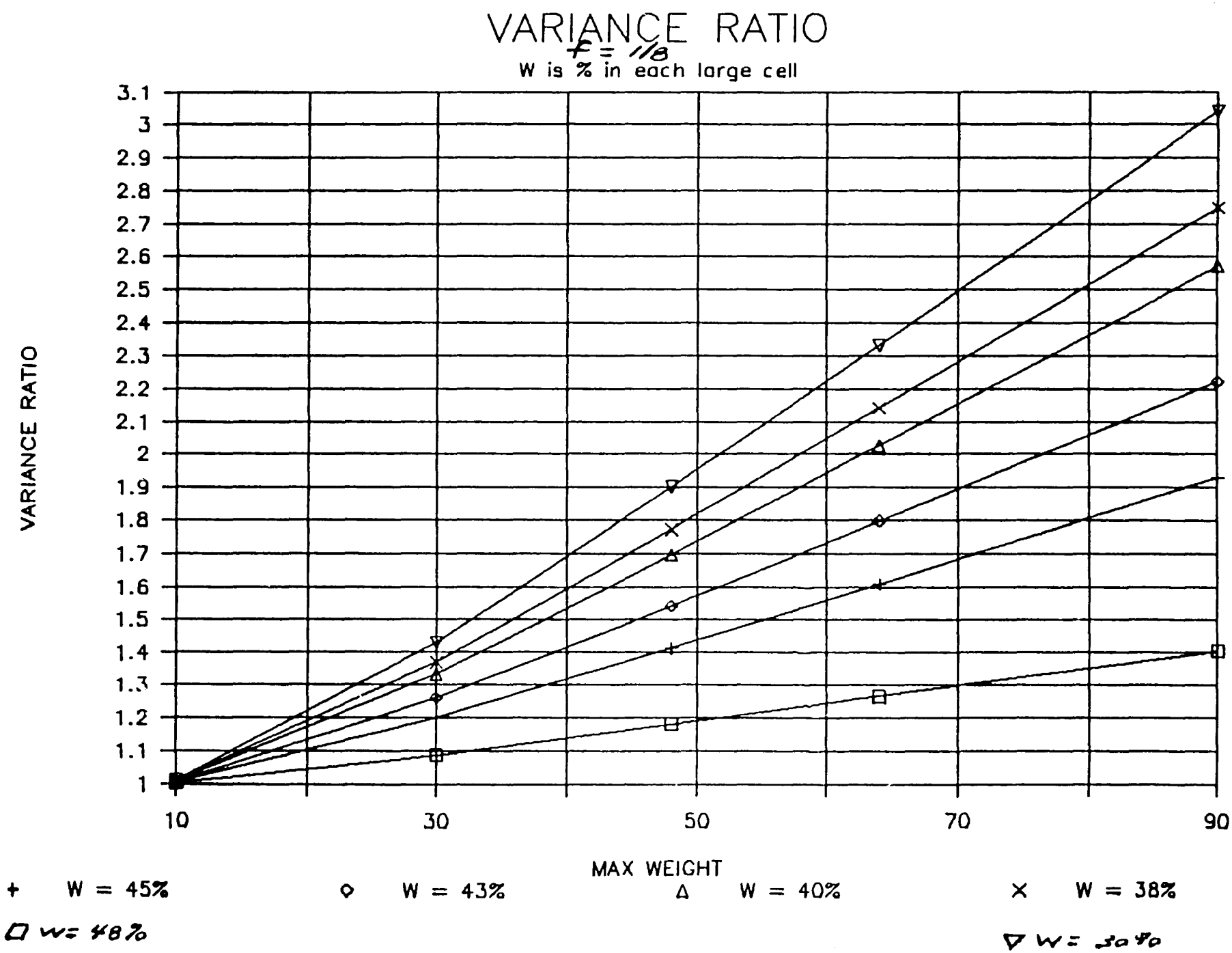
f = 1/6

W is % in each large cell



+ W = 45% ◇ W = 43% △ W = 40% × W = 38%

□ W = 48% ▽ W = 30%


VARIANCE RATIO
f = 1/8
W is % in each large cell
VARIANCE RATIO
3.1
3
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
10
30
50
70
90
MAX WEIGHT
+ W = 45%
◇ W = 43%
Δ W = 40%
× W = 38%
□ W = 48%
∇ W = 30%

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

May 11, 1989

MEMORANDUM FOR Henry Woltman
Assistant Division Chief for Census Design
Statistical Support Division

From: Rick Griffin
Chief, Census Design Branch
Statistical Support Division

Subject: Weighting Issues - 1990 Census

We are in the process of preparing draft weighting specifications for the Dress Rehearsal. These specifications will serve as a draft for 1990. Our goal is to examine all issues documented for 1980 weighting and to find solutions for these issues that are appropriate for the 1990 situation. In some cases, due to differences between the Dress Rehearsal and 1990 Census sample designs, the implementation of the solution may be different for 1990 than the Dress Rehearsal. In such cases the solution is determined with the 1990 sample design in mind and implemented as appropriate for the Dress Rehearsal. The purpose of this memorandum is to document the issues investigated so far as well as the proposed solutions.

Initial Weights

In 1980, initial weights were formed as the inverse of the observed emumeration district (ED) sampling rate (persons or housing units). Since EDs did not cross tabulation boundaries, it was important for sub-weighting area estimation that ED level estimates be accurate and reliable. In 1990, we do not have EDs. A suitable proxy is tabulation block groups. A tabulation block group is all blocks in the same tract or block numbering area (BNA) that have the same first digit of the block code and have the same place and Minor Civil Division (MCD) code. Sample data will be published for tabulation block groups. Thus concern for tabulation block group level estimates is warranted and reliable tabulation block group level estimates should result in reliable estimates for most sub-weighting area tabulation levels. The recommendation is to use tabulation block groups as the areas from which to form the initial weights for the 1990 Census. Note that the blocks in a tabulation block group all have the same designated sampling rate as did all the blocks in an ED in 1980. Statistical aspects of the initial weights issue have been examined and documented in an April 17, 1989 memorandum, Subject: Initial Weights for the 1990 Census.

Formation of Weighting Areas

In 1980, initial weighting areas were all EDs with the same

tract-MCD-Place-County codes. In 1990, initial weighting areas will be all tabulation block groups with the same tract-MCD-Place-County codes. Once initial weighting areas are formed, forming final weighting areas is basically the same as in 1980. The procedure is basically as follows. Within each place,any initial weighting area with less than 400 sample persons is combined with the initial weighting area in the place with the "closest" tract code. If the resulting weighting area has less than 400 sample persons, the procedure continues within the place. If the place collapses to one weighting area with less than 400 sample persons, the weighting area is combined with the initial weighting area outside the place with the "closest" tract code. In no case is a final weighting area allowed to cross a county boundary.

Upward Bias in 1-in-2 Sample Estimates

For the 1980 Census, an upward bias in 1-in-2 sampling rate area estimates was observed and documented in [1]. It was noted that for mixed sample (more than one designated sampling rate) weighting areas, whenever the sample for the 1-in-6 portion of a marginal weighting array category was extremely low in relation to the remainder of the ED (i.e. the reminder of the ED was sampled at about 1-in-6 but the portion of the ED in the margin was sampled at a rate of much less than 1-in-6) that the iterative ratio adjustment procedure tended to give larger weights than desired to records from the correctly sampled 1-in-2 portion of the marginal producing an upward bias for this portion. Also, the records from the 1-in-6 portion of the marginal received weights smaller than desired to estimate the 1-in-6 portion of the marginal. Although apparently not observed in 1980, if the sample in the 1-in-6 portion of a marginal in a mixed sample weighting area was extremely high in relation to the remainder of the ED, the result was an underestimate for the 1-in-2 portion. The sugested solution is to add a dimension to the weighting array which corresponds to sample type. We have checked with DOD and this will not present a problem. Thus, a dimension for 1-in-2, 1-in-6 and 1-in-8 designated sampling rates is proposed for 1990 Census weighting array. For the Dress Rehearsal this dimension has 4 sampling rates. The sampling rate dimension will be checked for collapsing last and it will be the third dimension in the iteration order after family type/household size and head/nonhead but prior to race/origin/sex/age. We want the race/origin dimension last in the iteration order since the last dimension will have exact agreement between sample estimates and 100 percent data (after collapsing).

Collapsing

The final collapsing criteria for the 1980 Census weighting are

given in [2]. Other than the maximum weights (see next section) , we plan on using these criteria for 1990.

The criteria requiring a minimum 100 percent marginal count for the race and Hispanic origin categories came about as a result of testing varying definitions of race/origin control categories and collapsing rules for these categories and is documented in [3]. It was found that "using only the observed sample count as a test criteria for collapsing race/origin could lead to a downward bias in estimates for those race/origin categories which were sparsely scattered throughout a state or region." It seems that requiring only a minimum sample count resulted in collapsing in undersampled weighting areas (for a given sparsely scattered race/origin category) but not collapsing in oversampled weighting areas. Thus, large weights were decreased amd small weights were allowed to stand resulting in underestimates. Requiring a minimum 100 percent count of size large enough to ensure that collapsing due to a small sample count is a rare event tends to take care of this problem. In theory, this same problem can occur for other dimensions of the weighting array. We have considered adding the same collapsing criteria for other margins. However, the other margins are not as "politically" sensitive or as likely to be sparsely scattered throughout a state or region. Our current proposal is not to add this additional collapsing criteria except, as in 1980, for the race/origin margins.

The above leads into a general discussion of whether, in some instances, it is better to collapse or not collapse. Apparently collapsing reduces the mean square error. However, collapsing to much, whatever that means, erodes some of the benefits of raking such as correcting for sampling biases and taking advantage of correlations between 100 percent and sample characteristics. Thus, there appears to be a trade off. Consider an example of a variation of what I believe has been called the "Philadelphia Problem". Suppose there are two races one of which needs to be collapsed due to too small a sample caused by undersampling and the other which was sampled correctly and has a sample large enough to stand alone. These two races are collapsed together. The undersampled race will be underestimated and the correctly sampled race will be overestimated. The combined races are of course correctly estimated. In spite of this, we do collapse because the mean square error of estimates is improved.

<u>Maximum Weights</u>

The raking weighting procedure to be used in the 1990 Census will result in differential weights. Raking is, of course beneficial as it improves the mean square error of estimates

and adjusts for sampling biases. However, it is also well known that differential weights can result in an increase in variance. The benefits of raking are clear but the ratio of weights should be controlled. For the 1980 Census, a maximum weight of 48 was used. This was four times the maximum initial weight that was allowed. A memorandum providing a statistical analysis and a recommended procedure for determining maximum weights for the 1990 Census and the Dress Rehearsal has been prepared (May 11,1989, Subject: Maximum Weights - 1990 Census). An independent maximum weight is calculated for each weighting area based on the observed sampling fraction. For the 1990 census, if the observed sampling fraction, f, of the weighting area is less than or equal to 14.5 percent the maximum weight is 64. If f is between 14.5 percent and 34 percent the maximum weight is 48 and if f is greater than or equal to 34 percent the maximum weight is 12. For the Dress Rehearsal the recommendation is the same except that for f less than 7.5 percent a maximum weight of 160 is used.

Race Coding

For the 1980 Census, the write-in entries to the race item were not coded during 100 percent processing. The race item was coded during sample data processing. Thus, the count of the "Other" race sample persons was much lower at the sample processing than at the 100 percent processing mainly due to the "Other" race write-in changes made during the sample coding operations. The race collapsing procedures were reviewed, [4] and [5], and it was decided to always collapse the "Other" race category with the White category [6]. Even with this change, there were possible problems. Sample counts of Indians, Asians and possibly Blacks could be low and sample counts for Whites could be high. This follows in weighting areas where persons of "Other" race were changed to any race except White. The observed sampling rate for the joint White/Other race category was effectively lowered so that sample persons in this category were assigned proportionately larger weights. Analagously, persons in the remaining race categories would have higher observed sampling rates and would receive proportionately lower weights. It was found that these problems could be corrected by placing sample persons in cells of the weighting matrix based on unedited responses (100 percent processing) and tabulating based on edited responses (sample processing).

For 1990, there will be two write-ins on the race item, one for American Indian and one for Asian Pacific Islander and Other. All these write-ins will be coded during 100 percent processing except for a Other race write-in with Q7 positive (Hispanic person). These latter cases will be coded only during sample processing but the race code on the Sample Edited Detail File

(SEDF) will not be changed. A new code will be entered on the SEDF to indicate Other race changed to a particular race. This new code, if applicable, is used for tabulation from the SEDF. Thus, there is no need to always collapse White and Other and sample persons can be placed in the weighting matrix based on the race code on the SEDF (not the new code). Thus, assuming no collapsing, for example, the sample estimate for Whites will be the number of Whites after 100 percent processing (Nw) plus the number of Other race from 100 percent processing (No) times the proportion of sample Other race persons who are Hispanic Origin and have a write-in race which was given a new code of White during sample data processing (R), (Nw + RNo). (Note that according to DOD (Don Dalzell) these other race write-in cases with Q7 positive may not be coded during sample data processing. If this is the case, then there is no difference between sample and 100 percent for race and there is no problem. If the coding is done we will make sure the new code is used.)

Miscellaneous

The following additional topics documented in [6] have been incorporated as appropriate in the draft weighting specifications.

- Additional Combinations of Initial Weighting Areas

  This guards against discrepancies between the 100 percent and the weighted sample count for total persons, total occupied housing units and total vacant housing units.

- Combining Tracts in Weighting Area Formation

  This is to deal with a non-zero suffix problem discovered for "Water Tracts". The problem probably does not exist for the Dress Rehearsal. We will check to see if it is a problem for 1990 and take appropriate action, as in 1980, if necessary.

- Differential Initial Weights

  Sample persons in housing units receive different initial weights than sample persons living in group quarters. Sample occupied and vacant housing units also receive different initial weights. This is done to reflect the fact that observed sampling rates differ for these categories (some of this is of course due to sampling bias).

  In 1980, there was an additional modification for non-military EDs sampled at 1-in-2 which contained a small military base. This was necessary because for the 1980 Census

Sample Design persons in group quarters (other than military) were sampled at the same rate as the rest of the ED. Military group quarters were sampled at a rate of 1-in-6. For 1990, all group quarters persons are sampled at a rate of 1-in-6. Thus, no modification for the initial weight of military group quarters persons is necessary.

- AUGMENTATION

  In 1980, if an ED had an observed sampling rate that was too small (less than 1/12 I believe) short form cases were converted to long (via imputation) so as to raise the observed sampling rate (to 1/12 ?). We need to look into this for 1990. Note that having a maximum initial weight takes care of some of this problem.

OTHER TOPICS EXAMINED

- A geographic adjustment to correct for differences between sample estimates and 100 percent counts for publication areas which are not composed exactly of one or more weighting areas was originally part of the 1980 Census weighting specifications [7]. However, the planned geographic adjustment was found not to eliminate the discrepancy and to result in some increase in the variance of the sample estimates [8]. Thus, this geographic adjustment was dropped for 1980 and will not be used for 1990. Note that weighting areas are formed so as to minimize the number of publication areas which are not completely defined as the union of final weighting areas. Weighting area size constraints, of course, make it necessary that this does occur for small publication areas such as block groups, and small places, tracts, and block numbering areas.

STSD Planning Group
cc: J. Thompson (STSD)
A. Navarro "
C. Swan "
J. Golmont "
S. Pak "

STSD:RGriffin

References

[1] 8/24/82, From: J. Thompson, For: The Record, Subject: Observation of an Upward Bias in 1-in-2 Sample Estimates

[2] 4/21/82, From; C. Jones, For: P. Bounpane, Subject: "Final" Collapsing Criteria for the 1980 Census Full Sample Weighting Matrices

[3] 12/29/81, From: H. Woltman, For: Distribution List, Subject: Decisions Concerning 1980 Census Sample Weighting Procedures

[4] 8/18/81, From: C. Jones, For: P. Bounpane, Subject: Collapsing the "Other" Race Category in Weighting the 1980 Census Sample

[5] 10/9/81, From: C. Jones, For: P. Bounpane, Subject: Concerns with Always Collapsing the Other Race Category into the White Race Category during the 1980 Census Sample Weighting

[6] 4/29/82, From: C. Jones, For: P. Bounpane, Subject: Documentation of Modifications to the 1980 Census Full Sample Weighting Specifications

[7] 5/10/80, From: C. Jones, For: G. Post, Subject: Memorandum No 37-1 Final Specifications for the 1980 Census Ratio Estimation and Sample Weighting for Population and Housing

[8] 5/26/81, From: C. Jones, For: P. Bounpane, Subject: Geographic Adjustment as Part of the 1980 Census Weighting Procedures: Is it Necessary?