ENCUESTA DE MÁRGENES DE COMERCIO Y TRANSPORTE
Diseño muestral

CONTENIDO

2.1. INTRODUCCIÓN

En el presente capítulo se presenta una apreciación global de aspectos relacionados con el diseño muestral para la "**Encuesta de Márgenes de Comercio y Transporte**". El diseño muestral para la encuesta de referencia está basada en el contexto de diseños muestrales probabilística, muestreo aleatorio estratificado y muestreo sistemático.

La documentación del diseño muestral e implementación constituye una guía metodológica para los analistas y usuarios, incluyendo la selección de la muestra, asignación de la muestra, recolección de datos, preparación de archivos de los datos, construcción de factores de expansión incluyendo ajustes o conceptos de calibración para compensar las imperfecciones de la muestra, especificaciones para la estimación de errores muestrales. La documentación de la encuesta también es esencial para la unión con otras fuentes de los datos y para los varios tipos de evaluación y el análisis suplementarios.

Este capítulo está organizado como sigue. Las secciones 2.1, 2.2, 2.3, 2.4 proporcionan una introducción general de la Encuesta de Márgenes de Comercio y Transporte. La sección 2.5 considera el diseño muestral probabilística, estratificado y sistemático. Describiendo la construcción del marco muestral y los problemas asociados, definición de las unidades muestrales de selección con probabilidad proporcional al tamaño y muestreo sistemático con intervalo constante. Discute el problema de determinación del tamaño de muestra óptima exigida para satisfacer los niveles de precisión preestablecidos en términos de error estándar y el coeficiente de variación de las estimaciones y el esquema de asignación de la muestra que puede considerarse que de satisfacen las demandas para producir estimaciones a nivel nacional, ciudades capitales y sus respectivas probabilidades de selección. Sección 2.6 discute el análisis e interpretación de los datos de la encuesta y, en particular, da énfasis al hecho de que el análisis apropiado de datos de la encuesta debe tener en cuenta los rasgos del diseño muestral y estadística inferencial de la información denominada como la tabla de errores muestrales, el concepto del efecto de diseño es introducido en el contexto de muestreo aleatorio estratificado, factores a tener en cuenta en esta discusión se incluyen los requisitos de precisión preestablecidos para las estimaciones de la encuesta y las consideraciones prácticas que derivan de la organización de archivos de trabajo.

2.2. ANTECEDENTES

El antecedente inmediato del anterior Cambio de año Base de las Cuentas Nacionales de Bolivia 1990 donde se llevó adelante la primera "Encuesta de Márgenes y Canales de Comercio y Transporte" con estudios en el eje central del país (La Paz, Cochabamba y Santa Cruz de la Sierra), cuyos resultados se aplicaron al "Año Base de las Cuentas Nacionales del sector comercio.

Actualmente el cálculo del Valor Bruto de la Producción para el comercio de distribución y el transporte de carga, está bajo la responsabilidad de la Dirección de Cuentas Nacionales, en base a tasas de márgenes obtenidas para el año 1990. Estas estimaciones permiten encontrar los valores correspondientes a la producción del sector comercio para cada año.

2.3. ENCUESTAS A EMPRESAS VERSUS ENCUESTAS A HOGARES

La gestión eficaz de registros de unidades económicas no es posible tomando exclusivamente operaciones estadísticas o censos económicos. Las encuestas no permiten detectar ni las modificaciones estructurales de las unidades de producción ni su evolución temporal, con tasas significativas de renovación. Los censos económicos implican elevados costos en recursos y tiempos de proceso, lo cual supone que la información recogida quede rápidamente desactualizada.

Mientras, para las Encuestas a Hogares que realiza la Dirección de Estadísticas e Indicadores Sociales del INE el marco muestral indispensable es la información proveniente del Censo Nacional de Población y Vivienda de 5 de septiembre de 2001. donde los indicadores que captura el mismo es de tipo estructural, debido a que las variables que lo conforman varían lentamente a lo largo del tiempo; por ejemplo, no captura situaciones de pobreza reciente o coyuntural puesto que no incorpora variables como el ingreso o el empleo que pueden ser muy volátiles.

Dentro de este contexto, las ventajas que ofrece el acceso a los datos administrativos son evidentes, fundamentalmente cuando están involucradas un gran número de unidades que deben satisfacer estrictos requerimientos de mantenimiento, cobertura y calidad, establecidos bajo normativas legales muy precisas.

Esta es la práctica generalizada en el ámbito de la Unión Europea, Organización para la Cooperación y el Desarrollo Económico (OCDE) y de la Comunidad Andina de las Naciones (CAN) que se encuentra en plena armonización. La presencia de fuentes de origen tributario o social es relevante dados los buenos grados de compatibilidad con los objetivos estadísticos. Por ejemplo, fuera de la Unión, los países con sistemas más desarrollados utilizan modelos similares, variando la tipología de fuentes y su prioridad en función de sus condiciones particulares. En este sentido, a modo de resumen a continuación se presenta las diferencias que existe entre Encuesta a Hogares y Encuesta a Empresas:

Encuestas a Hogares	Encuestas a Empresas
Muestreo multietápico	Muestreo monoetápico
Marco de áreas	Marco de lista
Recogida de datos por entrevista personal	Recogida de datos por correo y apoyo telefónico
Costo elevado	Menor costo
Variables cualitativas	Variables cuantitativas

2.4. OBJETIVOS

La Encuesta de Márgenes de Comercio y Transporte permite disponer de una información básica para el conocimiento de la realidad industrial y cuantificación de la producción, valor agregado y otras variables del sector comercio

Proporcionar información precisa, fiable y en el menor tiempo posible, de las principales características producción, y valor agregado de la actividad de los diversos sectores industriales, de forma que puedan satisfacer las necesidades de información tanto nacional como internacionales en la materia de Cuentas Nacionales.

2.5. DISEÑO MUESTRAL

Es una guía metodológica diseñada para los analistas y otros usuarios para describir paso a paso las etapas de realización de una encuesta por muestreo con el objetivo de garantizar información de buena calidad que se traducen en estimaciones precisas o robustas.

2.5.1. Unidad de investigación

La unidad básica o de investigación de la encuesta es la EMPRESA COMERCIAL. Se denomina EMPRESA a toda unidad jurídica que constituye una unidad organizativa de transacciones orientadas a la compra – venta de bienes y que disfruta de una cierta autonomía de decisión, principalmente a la hora de emplear los recursos corrientes de que dispone. La empresa puede ejercer una ó más actividades en uno ó varios lugares. Se considera que una empresa, a efectos de la encuesta, si su actividad principal está incluida dentro de la Clasificación Internacional Industrial Uniforme (CIIU).

Clasificación Internacional Industrial Uniforme (CIIU): La CIIU desempeña un papel importante al proporcionar el tipo de desglose por actividad necesario para la compilación de las cuentas nacionales desde el punto de vista de la producción. La CIIU tiene por finalidad establecer una clasificación uniforme de las actividades económicas productivas. Su propósito principal es ofrecer un conjunto de categorías de actividades que se pueda utilizar cuando se diferencian las estadísticas de acuerdo con esas actividades. El propósito secundario de la CIIU es presentar ese conjunto de categorías de actividad de modo tal que las entidades se puedan clasificar según la actividad económica que realizan.

Considerando que es necesario facilitar la comparación de la información en el tiempo y también con respecto a otros países y de acuerdo a normas internacionales, se utilizará la Clasificación Industrial Internacional Uniforme de la actividad económica "**Categoría de tabulación: G - Comercio al por mayor y al por menor; reparación de vehículos automotores, motocicletas, efectos personales y enseres domésticos**", CIIU Rev. 3.

2.5.2. Ámbito

2.5.2.1 Ámbito Poblacional

Conjunto de empresas con una o más personas ocupadas remuneradas y cuya actividad principal está incluido en la CIIU. La encuesta cubre a las empresas comerciales *(*mayoristas y minoristas*)* que realicen reventa de alimentos (víveres en general), bebidas, cigarrillos y cigarros; materias primas agropecuarias y animales vivos; materias primas no agropecuarias; productos textiles, prendas de vestir y sus accesorios; calzado, artículos de cuero y sucedáneos del cuero; productos farmacéuticos, perfumería y cosméticos. Droga veterinaria; muebles, electrodomésticos, artículos y equipo de uso doméstico; muebles máquinas y equipo de oficina; computadores, programas y suministros; papel, cartón, libros, papelería, periódicos y revistas; artículos de ferretería y cerrajería, herramientas, vidrios y pinturas; materiales para la construcción; vehículos automotores y motocicletas; partes, piezas y accesorios (lujos) para automotores y motocicletas; maquinaría para la agricultura, minería, construcción, industria; gasolina, lubricantes y aditivos para automotores, combustibles sólidos, líquidos y gaseosos; otras mercancías no especificadas anteriormente*.*

ACTIVIDAD PRINCIPAL de la empresa es aquélla que genera el **mayor valor añadido**. Si no se dispone de esta información, se considera aquélla que proporcione el **mayor valor de producción**, o en su defecto, la que emplee un **mayor número de personas ocupadas.**

2.5.2.2 Ámbito Geográfico

Todo el territorio nacional correspondiente a las ciudades capitales de departamento más la ciudad de El Alto. La encuesta está diseñada para poder obtener resultados a nivel nacional y de ciudades capitales, dependiendo de la precisión de la información recolectada, mediante el coeficiente de variación.

2.5.2.3 Ámbito Temporal

La encuesta se llevó a cabo con carácter **transversal, independiente y esporádico**. Los datos solicitados se refieren al **año natural,** definido como el periodo de tiempo que coincide con el calendario solar, comenzando el 1 de octubre y terminando el 31 de diciembre de 2008. El período de referencia de la recolección de la información generalmente de 365 días, es el año anterior a la aplicación de la encuesta, en este caso 2007.

2.5.3 Marco Muestral

Un requisito imprescindible para la realización de encuestas por muestreo es la existencia de un marco muestral que permita clasificar y estratificar la población objeto de estudio, desarrollar los procesos de

selección de muestras representativas y contacto con las unidades informantes, obtener los factores de elevación necesarios para la generación final de agregados e inferir los datos al conjunto poblacional. Dentro de este contexto, se presentan las líneas básicas de gestión del registro de unidades económicas DIRCEMBOL que trata de construir, actualizar, y mantener el INE a partir de fuentes administrativas. Asimismo, se presentan los principales retos para el futuro en este campo.

Directorio Central de Empresas de Bolivia (DIRCEMBOL): Constituye el marco de referencia para la mayoría de las encuestas económicas. Las variables del marco son código de actividad y el tamaño de la empresa. También figura un código que representa el intervalo de su cifra de negocios, conforme al siguiente cuadro estadístico.

Cuadro N° 1

Bolivia: Marco muestral de las empresas comerciales por ciudad capital de departamento, según forma legal de la empresa

Forma legal	Sucre	La Paz(*)	Cochabamba	Oruro	Potosí	Tarija	Santa Cruz	Trinidad	Cobija	Total
Empresa Unipersonal	235	1790	1281	258	171	324	1518	111	138	5826
Fundación/Asociación sin fines de lucro	2	4	2	1	0	0	2	0	0	11
Otra	3	1	1	0	0	1	0	0	0	6
Otro Carácter de Entidad	0	36	11	3	4	4	21	1	4	84
Sociedad Anónima	1	114	27	1	4	1	130	0	1	279
Sociedad Anónima Mixta	0	0	1	0	0	0	0	0	0	1
Sociedad Colectiva	0	1	0	0	0	0	0	0	0	1
Sociedad Comandita	0	0	0	0	0	0	2	0	0	2
Sociedad de Responsabilidad Limitada	35	1234	453	99	32	91	1244	13	30	3231
Total	276	3180	1776	362	211	421	2917	125	173	9441

Fuente: DIRCEMBOL 2008 (*): Incluye Ciudad de El Alto

Los sectores se corresponden, en su mayoría, con el nivel de 2 dígitos (división) de la Nomenclatura de Cuentas Nacionales (NCN). En algunos casos se ha optado por una mayor desagregación a nivel de producto o servicio (8 dígitos). Cada división constituye una población independiente a efectos de muestreo como se muestra en el siguiente cuadro estadístico.

Cuadro N° 2

Bolivia: Marco muestral de las empresas comerciales por ciudad capital, según NCN

Código NCN	Sucre	La Paz(*)	Cochabamba	Oruro	Potosí	Tarija	Santa Cruz	Trinidad	Cobija	**Total**
14	19	215	118	23	13	9	148	5	11	**561**
15	4	21	6	2	1	1	10	2	4	**51**
16	11	57	22	9	5	7	20	2	3	**136**
17	26	142	102	28	16	24	143	14	3	**498**
18	62	379	311	43	24	117	351	34	8	**1329**
19	14	68	67	9	7	32	53	5	3	**258**
20	1	58	17	26	17	6	13	0	0	**138**
21	38	543	270	59	36	88	485	14	25	**1558**
22	3	67	27	4	4	12	65	3	6	**191**
73	10	113	69	18	7	16	88	13	6	**340**
99	88	1517	767	141	81	109	1541	33	104	**4381**
Total	**276**	**3180**	**1776**	**362**	**211**	**421**	**2917**	**125**	**173**	**9441**

Fuente: DIRCEMBOL 2008 (*): Incluye Ciudad de El Alto

2.5.4 Tipo de muestreo

En la encuesta se ha utilizado un **muestreo aleatorio estratificado** y la selección de unidades muestrales mediante el muestreo sistemático con intervalo constante. Puesto que, una muestra aleatoria estratificada es obtenida mediante la separación de los elementos de la población en grupos que no presentan traslapes, llamados estratos y la selección posterior de una muestra PPT (Probabilidad Proporcional al Tamaño) con arranque aleatorio de cada estrato.

2.5.5 Estratificación de las unidades de muestreo

La estratificación divide las unidades de la población mutuamente excluyentes y grupos o estratos colectivamente exhaustivos, de donde se seleccionan las muestras aleatorias por separadas de cada estrato. Un propósito primario de estratificación es mejorar la precisión de las estimaciones de la encuesta o reducir la varianza. En este caso, la formación de los estratos debe ser tal que las unidades en el mismo estrato son tan homogéneas como sea posible y las unidades entre estratos son tan heterogéneas como sea posible con respecto a las características o variables de interés de la encuesta. Otros beneficios de estratificación incluyen (i) la conveniencia administrativa y flexibilidad y (ii) garantización de representación de dominios importantes y las subpoblaciones especiales.

2.5.5.1 Variables de estratificación

El muestreo aleatorio estratificado explícito en la Encuesta hace referencia a la variable geográfica departamento y grado de urbanización denominado ciudades capitales de departamento más El Alto, y los criterios de asignación fueron los siguientes:

- Nacional
- Ciudades Capitales de Departamento
  - Sucre (muestreo forzoso)
  - La Paz + El Alto (muestreo aleatorio estratificado y muestreo sistemático)
  - Cochabamba (muestreo aleatorio estratificado y muestreo sistemático)
  - Oruro (muestreo forzoso)
  - Potosí (muestreo forzoso)
  - Tarija (muestreo forzoso)
  - Santa Cruz de la Sierra (muestreo aleatorio estratificado y muestreo sistemático)
  - Trinidad (muestreo forzoso)
  - Cobija (muestreo forzoso)
- Clasificación de la actividad comercial
  - Comercio al por mayor
  - Comercio al por menor
- Categoría jurídica de la empresa
  - Empresa Jurídica (Sociedad Anónima - SA)
  - Empresa Jurídica (Sociedad de Responsabilidad Limitada - SRL)
  - Empresa Unipersonal
- Nomenclatura de las Cuentas Nacionales (NCN)

NCN	DESCRIPCIÓN
7 a 13	Alimentos y Bebidas
14	Textiles y Productos Textiles
15	Madera Aserrada y Productos de Madera Excepto Muebles
16	Pasta de Papel, Papel y Productos De Papel, Edicion e Impresión
17	Productos de Refinación del Petroleo y Otros Combustibles
18	Sustancias y Productos Quimicos
19	Sustancias y Productos Quimicos, Productos de Minerales no Metalicos
20	Fabricacion de Metales Comunes
21	Maquinaria y Equipo
22	Productos Manufacturados Diversos
99	Determinar su **NCN (**Categoría Otros**)**

En consecuencia, cada estrato dentro de una población a muestrear (Sector de actividad) viene determinado por el cruce de las variables ciudad capital de departamento, clasificación de la actividad comercial y Categoría jurídica de la empresa. Además, aplicando los criterios definidos con respecto a muestras exhaustivas y no exhaustivas se tiene la siguiente tabla estadística:

Cuadro N° 3

Bolivia: Marco muestral de las empresas comerciales por ciudad capital, según NCN

Código NCN	Sucre	La Paz(*)	Cochabamba	Oruro	Potosí	Tarija	Santa Cruz	Trinidad	Cobija	**Total**
14	5	215	118	3	6	2	148	3	8	**508**
15	4	21	6	2	1	1	10	1	0	**46**
16	8	57	22	6	5	4	20	1	3	**126**
17	20	142	102	12	7	9	143	5	0	**440**
18	19	379	311	7	9	35	351	4	4	**1119**
19	10	68	67	2	3	13	53	4	3	**223**
20	1	58	17	24	17	6	13	0	0	**136**
21	17	543	270	16	14	34	485	8	21	**1408**
22	2	67	27	4	4	5	65	3	5	**182**
73	5	113	69	8	4	3	88	6	5	**301**
99	0	1517	767	0	0	0	1541	0	0	**3825**
Total	**91**	**3180**	**1776**	**84**	**70**	**112**	**2917**	**35**	**49**	**8314**

Fuente: DIRCEMBOL 2008 (*): Incluye Ciudad de El Alto

2.5.6 Determinación del tamaño de muestra.

El conocimiento del tamaño apropiado de una muestra nos permite estar seguros de si los resultados publicados en diversos informes de la literatura estadística tuvieron como base, además de un diseño cuidadoso, una conclusión apropiada, en función de la significancia de la diferencia observada. Para calcular el tamaño de la muestra se debe tener en cuenta los errores tipo I y II, la varianza, la magnitud del efecto de diseño, el nivel de significancia y el poder de la prueba. Para decidir qué tipo de fórmula matemática se utilizará, lo primero que se debe observar es si nuestro estudio es de valores promedio, porcentajes, razón o comparativo.

La muestra diseñada para esta investigación es de tipo proporcional, estratificada por tamaños (empresas sin ventas y empresas con ventas), sector comercio (al por mayor y al por menor) y ciudades (ciudad capital de departamento), nomenclatura de las cuentas nacionales como objeto de estudio. Para la distribución de las encuestas realizadas dentro de cada estrato (sector, tamaño y ciudad) se repartieron las encuestas a realizar de forma proporcional.

Categoría	Universo	Muestra	Error muestral máximo (+/-)
Región objetivo 1	106497	400	5.0
Resto de regiones españolas	101342	400	5.0
Industrias y construcción	45996	399	5.0
Servicios	161843	401	5.0
Empresas sin asalariados	**124050**	**293**	5.9
Empresas con asalariados	**83788**	**507**	4.4
Total	**207838**	**800**	3.5

Por ejemplo, en el ámbito geográfico es el territorio económico español, y el tamaño de la muestra es de 800 cuestionarios válidos. El reparto de las encuestas por sectores, tamaños y zonas geográficas se resume en la cuadro anterior. En ésta también figuran el tamaño del universo y el error muestral máximo en condiciones habituales de muestreo (nivel de confianza del 95,5% y probabilidad de p=q=0,5 y 2 sigma).

Otro caso referido a la Encuesta a las Pequeñas y Medianas Empresas de la república de Chile, se verifica el tamaño de la muestra, marco muestral y el coeficiente de variación como medida de la precisión de las estimaciones de interés:

Actividades		Tamaño	Marco Muestral	Muestra Efectiva			
Categoría CIUU Rev. 3	Descripción			Total	Inclusión Forzosa	Inclusión Aleatoria	CV (%)
D	Industrias Manufactureras	Mediana - Grandes	353	65	10	55	0.92
		Medianas	598	62		62	1.27
		Medianas - Pequeñas	1438	106		106	1.27
		Pequeñas - Grandes	3201	139		139	2.14
		Pequeñas - Pequeñas	8968	155		155	3.27
		Total	**14558**	**527**	**10**	**517**	**0.98**

2.5.6.1 Estratos exhaustivos

Se denomina exhaustivo cuando abarca a todas las unidades estadísticas que componen el colectivo, universo, población o conjunto estudiado. Para la presente encuesta todas las empresas que pertenecen a los estratos geográficos departamento (Sucre, Oruro, Potosí, Tarija, Trinidad y Cobija).

2.5.6.2 Estratos no exhaustivos

Cuando una encuesta no es exhaustiva debido que se seleccionan algunas, una proporción, una fracción, un subconjunto de unidades estadísticas aleatoriamente.

Para el presente estudio se aplica la no exhaustiva en las ciudades capitales de departamento del eje central (La Paz + El Alto, Cochabamba y Santa Cruz de la Sierra).

En los estratos no exhaustivos se ha utilizado una **Afijación Proporcional** mediante la siguiente expresión matemática:

$n_h = n\frac{N_h}{N}$, donde

n_h : Número de empresas seleccionadas en la muestra del estrato h.
N_h: Número total de empresas en el estrato h.
N: Número total de empresas en la ciudad capital de departamento (La Paz, Cochabamba o Santa de la Sierra)

2.5.6.3 Expresiones matemáticas

A continuación se expone y describe cada una de las expresiones matemáticas de donde se deduce la ecuación para el cálculo del tamaño de muestra:

a) **Cálculo del tamaño de muestra de la media ($\bar{y}_{st}$)**. Conforme a la siguiente notación matemática:

L: Número de estratos.
1 – α: Nivel de confianza en una determinada investigación por muestreo.
$Z_{1-\frac{\alpha}{2}}$: Valor de la distribución normal estándar del nivel de confianza.
S^2_h: Varianza muestral o cuasivarianza del estrato h.
N_h: Total poblacional en el estrato h.
ε: Límite para error de estimación o error permisible.
W_h: Factor de expansión en el estrato h.
n_h : Muestra seleccionada mediante el muestreo aleatorio simple en el estrato h.

N: Población Total
n: Tamaño de muestra

$$Z_{1-\frac{\alpha}{2}}\sigma_{\bar{y}_{st}} = Z_{1-\frac{\alpha}{2}}\sqrt{V(\bar{y}_{st})} = Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}}{N^2}}$$, entonces

$$\varepsilon = Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}}{N^2}} \quad \Rightarrow \quad \varepsilon^2 = \frac{Z^2{}_{1-\frac{\alpha}{2}}}{N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right) \quad \Rightarrow si \quad n_h = nw_h$$

$$\varepsilon^2 = \frac{Z^2{}_{1-\frac{\alpha}{2}}}{N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{nw_h}\left(1-\frac{nw_h}{N_h}\right) \quad \Rightarrow \quad n = \frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h} - \frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}\frac{nw_h}{N_h}$$

$$n + \frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} nN_h s^2{}_h = \frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h} \quad \Rightarrow \quad n\left(1+\frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N_h s^2{}_h\right) = \frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}$$

$$n = \frac{\frac{Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2 N^2}}{\frac{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N^2}} \quad \Rightarrow \quad n = \frac{Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}$$

b) **Cálculo del tamaño de muestra de una proporción o porcentaje ($\hat{p}_{st}$).** Conforme a la siguiente notación matemática:

L: Número de estratos.
1 – α: Nivel de confianza en una determinada investigación por muestreo.
$Z_{1-\frac{\alpha}{2}}$: Valor de la distribución normal estándar del nivel de confianza.
$\hat{p}_h$: Proporción o porcentaje de una determinada clase
$\hat{q}_h$: Proporción o porcentaje del complemento de una determinada clase
$S^2{}_h$: Varianza muestral o cuasivarianza del estrato h.
N_h: Total poblacional en el estrato h.
ε: Límite para error de estimación o error permisible.
W_h: Factor de expansión en el estrato h.
n_h : Muestra seleccionada mediante el muestreo aleatorio simple en el estrato h.
n: Tamaño de muestra
N: Población Total

$$n = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h s^2{}_h} = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h \hat{q}_h}{w_h}}{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h \hat{p}_h \hat{q}_h}, \quad donde \quad s^2{}_h = \hat{p}_h \hat{q}_h$$

c) **Cálculo del tamaño de muestra del total poblacional ($\hat{Y}_{st}$)**. Conforme a la siguiente notación matemática:

L: Número de estratos.
1 – α: Nivel de confianza en una determinada investigación por muestreo.
$Z_{1-\frac{\alpha}{2}}$: Valor de la distribución normal estándar del nivel de confianza.
$S^2{}_h$: Varianza muestral o cuasivarianza del estrato h.
N_h: Total poblacional en el estrato h.
ε: Límite para error de estimación o error permisible.
W_h: Factor de expansión en el estrato h.
n_h : Muestra seleccionada mediante el muestreo aleatorio simple en el estrato h.
n : Tamaño de muestra

$$Z_{1-\frac{\alpha}{2}} \sigma_{\hat{Y}_{st}} = Z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{Y}_{st})} = Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}$$

$$\varepsilon = Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}} \quad \Rightarrow \quad \varepsilon^2 = Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h} \quad \Rightarrow \quad \varepsilon^2 = Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

$$\varepsilon^2 = Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n w_h} \left(1 - \frac{n w_h}{N_h}\right) \quad \Rightarrow \quad n = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2} - \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h} \frac{n w_h}{N_h}}{\varepsilon^2}$$

$$n + \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h} \frac{n w_h}{N_h}}{\varepsilon^2} = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2} \quad \Rightarrow \quad n \left(\frac{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2} \right) = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2}$$

$$n = \frac{\frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2}}{\frac{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2}} \quad \Rightarrow \quad n = \frac{Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{w_h}}{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h s^2{}_h}$$

d) **Cálculo del tamaño de muestra del total poblacional de una clase ($\hat{A}_{st}$).** Conforme a la siguiente notación matemática:

L: Número de estratos.
1 – α: Nivel de confianza en una determinada investigación por muestreo.

$Z_{1-\frac{\alpha}{2}}$: Valor de la distribución normal estándar del nivel de confianza.

$\hat{p}_h$: Proporción o porcentaje de una determinada clase

$\hat{q}_h$: Proporción o porcentaje del complemento de una determinada clase

S^2_h: Varianza muestral o cuasivarianza del estrato h.

N_h: Total poblacional en el estrato h.

ε: Límite para error de estimación o error permisible.

W_h: Factor de expansión en el estrato h.

n_h : Muestra seleccionada mediante el muestreo aleatorio simple en el estrato h.

n : Tamaño de muestra

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2_h \frac{s^2_h}{w_h}}{\varepsilon^2 + Z^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h s^2_h} = \frac{Z^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N^2_h \frac{\hat{p}_h \hat{q}_h}{w_h}}{\varepsilon^2 + Z^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} N_h \hat{p}_h \hat{q}_h}, \quad donde \quad s^2_h = \hat{p}_h \hat{q}_h$$

(OJO SEGÚN EL CONTENIDO FALTAN LOS PUNTOS 5.7., 5.8 Y 5.8.1 AL 5.8.4.)

EL PUNTO SIGUIENTE 1.2.9. PARECE QUE PERTENECE AL PUNTO 5.8.5 DEL CONTENIDO

2.5.6.3 Asignación o afijación del tamaño de muestra

Cuando se utiliza muestreo aleatorio estratificado se debe seleccionar al menos un elemento de muestreo de cada estrato. Así que primero se determina el tamaño de la muestra y después se determina cuántos elementos se deben seleccionar de cada estrato. Dependiendo de los criterios que se tengan en cuenta para distribuir la muestra entre los estratos se tienen diferentes tipos de asignación o afijación y ellos son: proporcional, de Neyman y óptima. También considerada como una técnica estadística que se utiliza con la finalidad de asignar o distribuir la muestra en cada uno de los estratos, que a continuación se describe cada una de ellas:

a) Asignación proporcional al tamaño

Es un sistema de asignación, afijación o distribución de una muestra aleatoria estratificada en el cual se usa la misma probabilidad de selección en cada estrato. Sea N el número de unidades de la población total que forma parte de alguna muestra:

$$n = n_1 + n_2 + n_3 + ... + n_L$$

Cuando la asignación es proporcional el tamaño de cada estrato es proporcional a la población del estrato:

$n_h = n\frac{N_h}{N}$ o $\frac{n_h}{n} = \frac{N_h}{N}$, reemplazando en la expresión matemática del límite para error de estimación y despejando n que viene a ser el tamaño de muestra, mediante una probabilidad proporcional al tamaño (PPT):

$$\varepsilon = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{N^2} \sum_{h=1}^{L} N^2_h \frac{s^2_h}{n_h} \frac{N_h - n_h}{N_h}} \quad \Rightarrow \quad \varepsilon^2 = Z^2_{1-\frac{\alpha}{2}} \frac{1}{N^2 n_h} \sum_{h=1}^{L} N^2_h s^2_h \left(1 - \frac{n_h}{N_h}\right)$$

$$\varepsilon^2 = Z^2_{1-\frac{\alpha}{2}} \frac{1}{N^2 n \frac{N_h}{N}} \sum_{h=1}^{L} N^2_h s^2_h \left(1 - \frac{1}{N_h} n \frac{N_h}{N}\right) \quad \Rightarrow \quad n = \frac{Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2 N} \frac{1}{N_h} \sum_{h=1}^{L} N^2_h s^2_h \left(1 - \frac{n}{N}\right)$$

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2 N}\frac{1}{N_h}\sum_{h=1}^{L} N^2{}_h s^2{}_h - \frac{Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2 N}\frac{1}{N_h}\sum_{h=1}^{L} N^2{}_h s^2{}_h \frac{n}{N} \quad \Rightarrow \quad n = \frac{Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2 N}\sum_{h=1}^{L} N_h s^2{}_h - \frac{nZ^2_{1-\frac{\alpha}{2}}}{\varepsilon^2 N^2}\sum_{h=1}^{L} N_h s^2{}_h$$

$$n\left(1+\frac{Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N^2}\right) = \frac{Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N} \quad \Rightarrow \quad n = \frac{\frac{Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N}}{\frac{\varepsilon^2 N^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N^2}}$$

$$n = \frac{NZ^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h} \text{ o para caso del total } \quad n = \frac{NZ^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}$$

También se puede determinar la expresión matemática del tamaño de muestra para asignación proporcional en el caso de una proporción o porcentaje:

Reemplazando $S^2_i = \hat{p}_h\hat{q}_h$ en la expresión $n = \dfrac{NZ^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}{\varepsilon^2 N^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h s^2{}_h}$, se tiene:

$$n = \frac{NZ^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h \hat{p}_h\hat{q}_h}{\varepsilon^2 N^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h \hat{p}_h\hat{q}_h} \text{ o para caso del total } \quad n = \frac{NZ^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h \hat{p}_h\hat{q}_h}{\varepsilon^2 + Z^2_{1-\frac{\alpha}{2}}\sum_{h=1}^{L} N_h \hat{p}_h\hat{q}_h}$$

En la asignación de Probabilidad Proporcional al Tamaño (PPT) la muestra se reparte entre los estratos proporcionalmente a los tamaños de éstos. Este tipo de asignación se utiliza cuando los costos y las varianzas de los estratos no son muy diferentes.

b) Asignación de Neyman

Cuando se realiza un muestreo aleatorio estratificado, los tamaños muestrales en cada uno de los estratos es n_h, para ello se puede basarse en alguno de los siguientes criterios:

- Elegir los n_h de modo que minimice la varianza del estimador, para un costo específico, o bien;
- Habiendo fijado la varianza que se puede admitir para el estimador, minimizar el costo en la obtención de las muestras.

Teorema: (Asignación de Neyman). Sea E una población con N elementos, dividida en L estratos, con N_h elementos en cada estrato, donde h = 1, 2, 3, ..., L:

$N = N_1 + N_2 + N_3 + \ldots + N_L$

Sea el problema de programación no lineal que consta de la función objetivo y la correspondiente restricción, como sigue:

Min Z = $V(\bar{y}_{st})$

Sujeto a $\sum_{h=1}^{L} n_h = n$

Para la determinación de una solución óptima de un problema de programación no lineal, se aplica los multiplicadores de Lagrange, como sigue:

$$L(n_h, \lambda) = V(\bar{y}_{st}) + \lambda\left(\sum_{h=1}^{L} n_h - n\right) \qquad \text{ó} \qquad L(n_h, \lambda) = \frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}}{N^2} + \lambda\left(\sum_{h=1}^{L} n_h - n\right) \qquad \text{equivalente}$$

$$L(n_h, \lambda) = \frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\left(1 - \frac{n_h}{N_h}\right)}{N^2} + \lambda\left(\sum_{h=1}^{L} n_h - n\right) = \frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}}{N^2} - \frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{n_h}{N_h}}{N^2} + \lambda\left(\sum_{h=1}^{L} n_h - n\right)$$

Minimizando la función Lagrange $L(n_h, \lambda)$ o aplicando derivadas parciales con respecto n_h y λ respectivamente, se tiene

$$\frac{\partial L(n_h, \lambda)}{\partial n_h} = -\frac{1}{N^2 n^2{}_h}\sum_{i=1}^{L} N^2{}_h s^2{}_h + \lambda = 0 \qquad \text{y} \qquad \frac{\partial L(n_h, \lambda)}{\partial \lambda} = \left(\sum_{h=1}^{L} n_h - n\right) = 0$$

$$\lambda = \frac{1}{N^2 n^2{}_h}\sum_{h=1}^{L} N^2{}_h s^2{}_h \;\Rightarrow\; \sqrt{\lambda} = \frac{1}{N n_h}\sum_{h=1}^{L} N_h s_h \;\Rightarrow\; \sqrt{\lambda} = \frac{1}{N n \frac{N_h}{N}}\sum_{h=1}^{L} N_h s_h \;\Rightarrow\; \sqrt{\lambda} = \frac{1}{n N_h}\sum_{h=1}^{L} N_h s_h$$

$$\Rightarrow\; \frac{n}{\sum_{h=1}^{L} N_h s_h} = \frac{1}{N_h\sqrt{\lambda}} \;\Rightarrow\; \frac{n}{\sum_{h=1}^{L} N_h s_h} = \frac{1}{\frac{N_h S_h}{n_h}} \;\Rightarrow\; \frac{n}{\sum_{h=1}^{L} N_h s_h} = \frac{n_h}{N_h S_h} \;\Rightarrow\; n_h = n\left(\frac{N_h S_h}{\sum_{h=1}^{L} N_h s_h}\right)$$

Donde nh = $n\frac{N_h}{N}$: asignación proporcional

Sea n el tamaño de muestra determinado y se divide en cada estrato como:

$$n = n_1 + n_2 + n_3 + ... + n_L$$

Sea Y la variable aleatoria que representa la característica que se intenta estudiar, sobre cada estrato puede definirse como: $\bar{y}_h$ es la media aritmética obtenido en una muestra de tamaño n_i en el estrato N_h y sea $V(\bar{y}_h)$ la varianza de dicha variable aleatoria, entonces: $\sum_{h=1}^{L} V(\bar{y}_h)$, se minimiza cuando $n_h = n\frac{N_h \hat{S}_h}{\sum_{h=1}^{L} N_h \hat{S}_h}$, donde $\hat{S}^2{}_h = \frac{1}{N-1}\sum_{i=1}^{N}(Y_{hi} - \bar{Y}_h)^2$ $\begin{cases} Y_{hi} = i - ésimo\ elemento\ de\ N_h \\ Y_h = media\ poblacional\ de\ N_h \end{cases}$, se denomina cuasi-varianza del estrato N_h.

La idea de la distribución óptima, trata de jugar no sólo con el tamaño del estrato, sino que también pretende jugar con la variabilidad del mismo, de forma que parece lógico que los estratos de mayor variabilidad le correspondan muestras mayores. Si $\sigma_1, \sigma_2, \sigma_3, ..., \sigma_L$ son las desviaciones típicas de los L-estratos podemos explicar tanto los tamaños de los estratos, así como su variabilidad.

$$\frac{n_1}{N_1\sigma_1}=\frac{n_2}{N_2\sigma_2}=\frac{n_3}{N_3\sigma_3}=\cdots=\frac{n_L}{N_L\sigma_L}$$

De donde se obtienen los tamaños muestrales de la distribución óptima o distribución de Neyman (su inventor) que se obtienen por la fórmula:

$$n_h=\frac{n*N_h*\sigma_h}{N_1\sigma_1+N_2\sigma_2+N_3\sigma_3+\cdots N_L\sigma_L}$$ para h = 1,2,...., L y donde $n = n_1 + n_2 + \ldots + n_L$

$$\varepsilon=Z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}_{st}}=Z_{1-\frac{\alpha}{2}}\sqrt{V(\bar{y}_{st})}=Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{N^2}\sum_{h=1}^{L}N^2{}_h\,\frac{S^2{}_h}{n_h}\left(\frac{N_h-n_h}{N_h}\right)}=Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{N^2}\sum_{h=1}^{L}N^2{}_h\,\frac{S^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right)}$$

Reemplazando n_i por $n\left(\frac{N_hS_h}{\sum_{h=1}^{L}N_hs_h}\right)$

$$\varepsilon^2=Z^2{}_{1-\frac{\alpha}{2}}\,\frac{1}{N^2}\sum_{h=1}^{L}N^2{}_h\,\frac{S^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right)\quad\Rightarrow\quad\varepsilon^2=Z^2{}_{1-\frac{\alpha}{2}}\,\frac{1}{N^2n\left(\frac{N_hS_h}{\sum_{h=1}^{L}N_hS_h}\right)}\sum_{h=1}^{L}N^2{}_h\,\frac{S^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right)$$

$$\Rightarrow\quad n=\frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\,\frac{\sum_{h=1}^{L}N_hS_h\sum_{h=1}^{L}N^2{}_hS^2{}_h}{N_hS_h}-\frac{nZ^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\,\frac{\sum_{h=1}^{L}N_hS_h}{N_hS_h}\,\frac{N_hS_h}{\sum_{h=1}^{L}N_hS_h}\,\frac{\sum_{h=1}^{L}N^2{}_hS^2{}_h}{N_h}$$

$$\Rightarrow\quad n=\frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\,\frac{\sum_{h=1}^{L}N_hS_h\sum_{ih1}^{L}N^2{}_hS^2{}_h}{N_hS_h}-\frac{nZ^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\sum_{h=1}^{L}N_hS^2{}_h\quad\Rightarrow\quad n\left(1+\frac{nZ^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\sum_{h=1}^{L}N_hS^2{}_h\right)=\frac{Z^2{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\left(\sum_{h=1}^{L}N_hS_h\right)^2$$

$$n=\frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}N_hS_h\right)^2}{\varepsilon^2N^2+Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_hS^2{}_h}$$ o para el caso del total $$n=\frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}N_hS_h\right)^2}{\varepsilon^2+Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_hS^2{}_h}$$

También se puede determinar la expresión matemática del tamaño de muestra para asignación óptima en el caso de una proporción o porcentaje:

Reemplazando $S^2{}_h = \hat{p}_h\hat{q}_h$ en la expresión $n=\frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}N_hS_h\right)^2}{\varepsilon^2N^2+Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_hS^2{}_h}$, se tiene:

$$n=\frac{Z^{2}{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}N_h\sqrt{\hat{p}_h\hat{q}_h}\right)^2}{\varepsilon^2N^2+Z^{2}{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h\hat{p}_h\hat{q}_h}$$ o para el caso del total $$n=\frac{Z^{2}{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}N_h\sqrt{\hat{p}_h\hat{q}_h}\right)^2}{\varepsilon^2+Z^{2}{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h\hat{p}_h\hat{q}_h}$$

Una afijación de Neyman se aplica cuando existen marcadas diferencias en la variabilidad de las observaciones dentro de los estratos, es recomendable utilizar este método de distribución de la muestra, ya que además de tener en cuenta el tamaño de los estratos se tiene en cuenta la dispersión de los datos dentro de cada estrato. De ésta manera se obtendrá una muestra más grande de aquellos estratos que sean más heterogéneos.

c) Asignación Óptima

Esta técnica de distribución de la muestra se aplica cuando además de tener marcadas diferencias en la dispersión o variabilidad dentro de los estratos, el costo para obtener la información de un estrato a otro varía, se recomienda utilizar la asignación óptima. Con ésta asignación se tiene en cuenta el tamaño de los estratos, la dispersión o variabilidad dentro de ellos y el costo para recopilar la información.

Reemplazando n_i por $n\left(\frac{\frac{N_hS_h}{\sqrt{c_h}}}{\sum_{h=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}}\right)$

$$\varepsilon^2=Z^{2}{}_{1-\frac{\alpha}{2}}\frac{1}{N^2}\sum_{h=1}^{L}N^2{}_h\frac{S^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right)\quad\Rightarrow\quad\varepsilon^2=Z^{2}{}_{1-\frac{\alpha}{2}}\frac{1}{N^2n\left(\frac{\frac{N_hS_h}{\sqrt{c_h}}}{\sum_{h=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}}\right)}\sum_{h=1}^{L}N^2{}_h\frac{S^2{}_h}{n_h}\left(1-\frac{n_h}{N_h}\right)$$

Despejando n se obtiene la fórmula matemática del tamaño de muestra para $\bar{y}_{st}$:

$$n=\frac{Z^{2}{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\frac{\sum_{h=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}}{\frac{N_hS_h}{\sqrt{c_h}}}\sum_{h=1}^{L}N^2{}_hS^2{}_h-\frac{nZ^{2}{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\frac{\sum_{h=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}}{\frac{N_hS_h}{\sqrt{c_h}}}\frac{\frac{N_hS_h}{\sqrt{c_h}}}{\sum_{h=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}}\sum_{h=1}^{L}N_hS^2{}_h$$

$$n\left(1+\frac{Z^{2}{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\sum_{h=1}^{L}N_hS^2{}_h\right)=\frac{Z^{2}{}_{1-\frac{\alpha}{2}}}{\varepsilon^2N^2}\left(\sum_{i=1}^{L}\frac{N_hS_h}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_hS_h\sqrt{c_h}\right)$$

$$n=\frac{Z^{2}{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_hS_{hi}}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_hS_h\sqrt{c_h}\right)}{\varepsilon^2N^2+Z^{2}{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_{hi}S^2{}_h}$$

También se puede determinar la expresión matemática del tamaño de muestra para asignación óptima en el caso de una proporción o porcentaje:

Reemplazando $S^2_{ih} = \hat{p}_h\hat{q}_h$ en la expresión $n = \dfrac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_h S_h}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_h S_h\sqrt{c_h}\right)}{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h S^2{}_h}$, se tiene:

$$n = \frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_h\sqrt{\hat{p}_h\hat{q}_h}}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_h\sqrt{\hat{p}_h\hat{q}_h}\sqrt{c_h}\right)}{\varepsilon^2 N^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h\hat{p}_h\hat{q}_h}$$

Asimismo, se puede determinar la expresión matemática del tamaño de muestra para asignación óptima en el caso del total poblacional de una clase (una proporción o porcentaje):

Reemplazando $S^2_{ih} = \hat{p}_h\hat{q}_h$ en la expresión $n = \dfrac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_h S_h}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_h S_h\sqrt{c_h}\right)}{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h S^2{}_h}$, se tiene:

$$n = \frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_h\sqrt{\hat{p}_h\hat{q}_h}}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_h\sqrt{\hat{p}_h\hat{q}_h}\sqrt{c_h}\right)}{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_h\hat{p}_h\hat{q}_h}$$

Finalmente, se puede determinar la expresión matemática del tamaño de muestra para asignación óptima en el caso del total poblacional:

$$n = \frac{Z^2{}_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L}\frac{N_h S_{hi}}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L}N_h S_h\sqrt{c_h}\right)}{\varepsilon^2 + Z^2{}_{1-\frac{\alpha}{2}}\sum_{h=1}^{L}N_{hi}S^2{}_h}$$

A la hora de determinar el tamaño de muestra hay que tomar en cuenta varios factores: el tipo de muestreo, el parámetro a estimar, el error muestral admisible, la varianza poblacional y el nivel de confianza. Por ello antes de presentar algunos casos sencillos de cálculo del tamaño muestral delimitemos estos factores.

Parámetro. Son las medidas o datos que se obtienen sobre la población objeto de estudio (Número total de visitantes receptores o emisores).

Estadístico. Los datos o medidas que se obtienen en base a una muestra y para realizar inferencias o estimaciones para obtener los parámetros.

Error Muestral, de estimación o standard. Es la diferencia entre un estadístico y su parámetro correspondiente. Es una medida de la variabilidad de las estimaciones de muestras repetidas en torno al valor de la población, nos da una noción clara de hasta dónde y con qué probabilidad una estimación basada en una muestra se aleja del valor que se hubiera obtenido por medio de un censo completo. Siempre se comete un error, pero la naturaleza de la investigación nos indicará hasta qué medida podemos cometerlo (los resultados se someten a error muestral e intervalos de confianza que varían muestra a muestra). Varía según se calcule al principio o al final. Un estadístico será más preciso en cuanto y tanto su error es más pequeño. Podríamos decir que es la desviación de la distribución muestral de un estadístico y su fiabilidad.

Nivel de Confianza. Probabilidad de que la estimación efectuada se ajuste a la realidad. Cualquier información que queremos recoger está distribuida según una ley de probabilidad (Gauss o Student), así llamamos nivel de confianza a la probabilidad de que el intervalo construido en torno a un estadístico capte el verdadero valor del parámetro.

Varianza Poblacional. Cuando una población es más homogénea la varianza es menor y el número de entrevistas necesarias para construir un modelo reducido del universo, o de la población, será más pequeño. Generalmente es un valor desconocido y hay que estimarlo a partir de datos de estudios previos. En ese entendido, se ha realizado el procesamiento en paquete estadístico SPSS para el cálculo de número de

observaciones muestrales, gasto total, promedio de gasto diario, varianza, desviación estándar o típica y coeficiente de variación por estrato, análisis e interpretación de la información obtenida mediante la "Encuesta de Gasto Turismo Receptor y Emisor 2007" . Los mismos se presentan en las siguientes tablas estadísticas:

Cuadro N° 4

Bolivia: Distribución de la muestra de empresas comerciales por ciudad capital, según NCN

Código NCN	Sucre	La Paz(*)	Cochabamba	Oruro	Potosí	Tarija	Santa Cruz	Trinidad	Cobija	**Total**
14	5	32	20	3	6	2	26	3	8	**105**
15	4	2	4	2	1	1	1	1	0	**16**
16	8	9	10	6	5	4	10	1	3	**56**
17	20	25	27	12	7	9	21	5	0	**126**
18	19	49	56	7	9	35	53	4	4	**236**
19	10	9	28	2	3	13	14	4	3	**86**
20	1	19	17	24	17	6	9	0	0	**93**
21	17	114	72	16	14	34	100	8	21	**396**
22	2	10	4	4	4	5	15	3	5	**52**
73	5	16	17	8	4	3	12	6	5	**76**
99	0	1	0	0	0	0	1	0	0	**2**
Total	**91**	**286**	**255**	**84**	**70**	**112**	**262**	**35**	**49**	**1244**

(*): Incluye Ciudad de El Alto

2.6. Errores de muestreo

Se expresa en términos relativos

$$CV\left(\hat{X}\right)=\frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}}.100$$

donde $\hat{V}\left(\hat{X}\right)=\sum_{h}\hat{V}\left(\hat{X}_h\right)$

El valor de $\hat{V}\left(\hat{X}_h\right)$ tiene tres componentes (sumandos):

$\hat{N}_h^*\left(\hat{N}_h^*-n_h^*\right)\frac{\sum_{i=1}^{n_h^*}\left(x_i-\overline{X}_h^*\right)^2}{n_h^*\left(n_h^*-1\right)}$ debida a la variación de la variable

$\overline{X}_h^{*2}.\hat{N}_h^*\left(N_h-\hat{N}_h^*\right)\frac{N_n-n_h}{N_h\left(n_h-1\right)}$ debida a la variación de

$\sum_{k\neq h}N_k\left(N_k-n_k\right)\frac{S_k^{h^2}}{n_k}$ debida a los cambios de estrato

Siendo:

$\overline{X}_h^*=\frac{\sum_{i=1}^{n_h^*}x_i}{n_h^*}$ Y $S_k^h=\frac{\sum_{i=1}^{n_k^h}x_i^2}{n_k-1}-\frac{\left(\sum_{i=1}^{n_k^h}x_i\right)^2}{n_k\left(n_k-1\right)}$

la cuasivarianza muestral de las empresas que pasan de un estrato k cualquiera, al estrato h.

2.7. Construcción de los factores de expansión

El factor de factor de expansión, peso, factor de ponderación o factor de elevación se define como el inverso de la fracción de muestreo o la probabilidad de selección de una empresa conforme a la siguiente expresión matemática en el muestreo aleatorio estratificado: $W_h = \frac{N_h}{n_h}$

Cuadro Nº 5

Bolivia: Distribución de empresas por marco, muestra y factores de expansión, según NCN

NOMENCLATURA DE CUENTAS NACIONALES		Total de Empresas (Nh)	Empresas prorateados (Nh)	muestra de empresas (nh)	Factor de Expansión (Wh)
14	TEXTILES Y PRODUCTOS TEXTILES	561	1047	105	10
15	MADERA ASERRADA Y PRODUCTOS DE MADERA EXCEPTO MUEBLES	51	95	16	6
16	PASTA DE PAPEL, PAPEL Y PRODUCTOS DE PAPEL, EDICION E IMPRESIÓN	136	254	56	5
17	PRODUCTOS DE REFINACIÓN DEL PETROLEO Y OTROS COMBUSTIBLES	498	929	126	7
18	SUSTANCIAS Y PRODUCTOS QUIMICOS	1329	2480	236	11
19	PRODUCTOS DE MINERALES NO METALICOS	258	481	86	6
20	FABRICACION DE METALES COMUNES	138	257	93	3
21	MAQUINARIA Y EQUIPO	1558	2907	396	7
22	PRODUCTOS MANUFACTURADOS DIVERSOS	191	356	52	7
73	Alimentos y bebidas	340	634	76	8
99	No especificado	4381	1	2	0
Total		**9441**	**9441**	**1244**	**8**

Por ejemplo, el factor de expansión de la Nomenclatura de Cuentas Nacionales de "Textiles y productos de textiles (14)" es de 10, que puede ser interpretado que una empresa de la muestra representa a 10 empresas de esa actividad incluida más la empresa que ha proporcionado información.

2.8. Estimaciones

Un estimador se define como una cantidad calculada en base a las observaciones muestrales de una o más variables de interés, con la finalidad de realizar algunas inferencias de la población objetivo.

2.8.1 Determinación de los estimadores de la media

a) Estimador de la media poblacional (μ_{st}):

$$\bar{y}_{st} = \frac{1}{N}\left(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3 + \cdots + N_L\bar{y}_L\right) = \frac{\sum_{h=1}^{L} N_h\bar{y}_h}{N}$$

b) Varianza estimada de la media $\bar{y}_{st}$

$$V(\bar{y}_{st}) = V\left[\frac{1}{N}\left(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3 + \cdots + N_L\bar{y}_L\right)\right] = \frac{1}{N^2}\left[N^2{}_1V(\bar{y}_1) + N^2{}_2V(\bar{y}_2) + N^2{}_3V(\bar{y}_3) + \cdots + N^2{}_LV(\bar{y}_L)\right]$$

$$= \frac{1}{N^2}\left[N^2{}_1\frac{s^2{}_1}{n_1}\frac{N_1 - n_1}{N_1} + N^2{}_2\frac{s^2{}_2}{n_2}\frac{N_2 - n_2}{N_2} + N^2{}_3\frac{s^2{}_3}{n_3}\frac{N_3 - n_3}{N_3} + \cdots + N^2{}_L\frac{s^2{}_L}{n_L}\frac{N_L - n_L}{N_L}\right]$$

$$= \frac{\sum_{h=1}^{L} N^2{}_h\frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}}{N^2}$$

c) Desviación estándar estimada de $\bar{y}_{st}$

$$\sigma_{\bar{y}_{st}} = \sqrt{V(\bar{y}_{st})} = \sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}{N^2}}$$

d) Límite para error de estimación de $\bar{y}_{st}$

$$Z_{1-\frac{\alpha}{2}} \sigma_{\bar{y}_{st}} = Z_{1-\frac{\alpha}{2}} \sqrt{V(\bar{y}_{st})} = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_i - n_h}{N_h}}{N^2}}$$

e) Intervalo de confianza de μ_{st}

$$P\left(\bar{y}_{st} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}{N^2}} \le \mu_{st} \le \bar{y}_{st} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}{N^2}} \right) = 1-\alpha$$

Donde $S^2{}_h = \frac{\sum_{i=1}^{n_i} (y_{hi} - \bar{y}_h)^2}{n_h - 1} = \frac{\sum_{i=1}^{n_i} y^2{}_{hi} - n_h \bar{y}^2{}_h}{n_h - 1} = \frac{\sum_{i=1}^{n_{i1}} y^2{}_{hi}}{n_h - 1} - \frac{n_h \bar{y}^2{}_h}{n_h - 1}$

2.8.2 Determinación de los estimadores de un total poblacional

a) Estimador del total poblacional ($\hat{Y}_{st}$)

$$\hat{Y}_{st} = N\bar{y}_{st} = N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3 + \cdots + N_L\bar{y}_L = \sum_{h=1}^{L} N_h\bar{y}_h$$

b) Varianza estimada de $\hat{Y}_{st}$

$$V(\hat{Y}_{st}) = V(N\bar{y}_{st}) = N^2 V(\bar{y}_{st}) = \frac{N^2 \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}{N^2} = \sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}$$

c) Desviación estándar estimada de $\hat{Y}_{st}$

$$\sigma_{\hat{Y}_{st}} = \sqrt{V(\hat{Y}_{st})} = \sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}$$

d) Límite para error de estimación de $\hat{Y}_{st}$

$$Z_{1-\frac{\alpha}{2}} \sigma_{\hat{Y}_{st}} = Z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{Y}_{st})} = Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h} \frac{N_h - n_h}{N_h}}$$

e) Intervalo de confianza de Y_{st}

$$P\left(\hat{Y}_{st} - Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}} \le Y_{st} \le \hat{Y}_{st} + Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{L} N^2{}_h \frac{s^2{}_h}{n_h}\frac{N_h - n_h}{N_h}}\right) = 1-\alpha$$

2.8.3 Determinación de los estimadores de una proporción o porcentaje

a) Estimador de la proporción poblacional (p)

$$\hat{p}_{st} = \frac{1}{N}\left(N_1\hat{p}_1 + N_2\hat{p}_2 + N_3\hat{p}_3 + \cdots + N_L\hat{p}_L\right) = \frac{\sum_{h=1}^{L} N_h\hat{p}_h}{N}$$

b) Varianza estimada de $\hat{p}_{st}$

$$V(\hat{p}_{st}) = V\left[\frac{1}{N}\left(N_1\hat{p}_1 + N_2\hat{p}_2 + N_3\hat{p}_3 + \cdots + N_L\hat{p}_L\right)\right] = \frac{1}{N^2}\left[N^2{}_1V(\hat{p}_1) + N^2{}_2V(\hat{p}_2) + N^2{}_3V(\hat{p}_3) + \cdots + N^2{}_LV(\hat{p}_L)\right]$$

$$= \frac{1}{N^2}\left[N^2{}_1\frac{\hat{p}_1\hat{q}_1}{n_1 - 1}\frac{N_1 - n_1}{N_1} + N^2{}_2\frac{\hat{p}_2\hat{q}_2}{n_2 - 1}\frac{N_2 - n_2}{N_2} + N^2{}_3\frac{\hat{p}_3\hat{q}_3}{n_3 - 1}\frac{N_3 - n_3}{N_3} + \cdots + N^2{}_LN^2{}_L\frac{\hat{p}_L\hat{q}_L}{n_L - 1}\frac{N_L - n_L}{N_L}\right]$$

$$= \frac{\sum_{h=1}^{L} N^2{}_h\frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2}$$

c) Desviación estándar estimada de $\hat{p}_{st}$

$$\sigma_{\hat{p}_{st}} = \sqrt{V(\bar{x}_{st})} = \sqrt{\frac{\sum_{h=1}^{L} N^2{}_h\frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2}}$$

d) Límite para error de estimación de $\hat{p}_{st}$

$$Z_{1-\frac{\alpha}{2}}\sigma_{\hat{p}_{st}} = Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{p}_{st})} = Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sum_{h=1}^{L} N^2{}_h\frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2}}$$

e) Intervalo de confianza de P_{st}

$$P\left(\hat{p}_{st} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sum_{i=1}^{L} N^2{}_h\frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2}} \le \mu \le \hat{p}_{st} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sum_{h=1}^{L} N^2{}_i\frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2}}\right) = 1-\alpha$$

Donde

$$S^2{}_h = \frac{\sum_{i=1}^{n_h}\left(a_{hi} - \hat{p}_h\right)^2}{n_h - 1} = \frac{\sum_{i=1}^{n_h} a^2{}_{hi} - n_h\hat{p}^2{}_h}{n_h - 1} = \frac{\frac{n_h}{n_h}\sum_{i=1}^{n_h} a^2{}_{hi}}{n_h - 1} - \frac{n_h\hat{p}^2{}_h}{n_h - 1} = \frac{n_h\frac{n_{hi}}{n_h} - n_h\hat{p}^2{}_h}{n_h - 1} = \frac{n_h\hat{p}_h - n_h\hat{p}^2{}_h}{n_h - 1}$$

$$= \frac{n_h\hat{p}_h(1-\hat{p}_h)}{n_h - 1} = \frac{n_h\hat{p}_h\hat{q}_h}{n_h - 1}$$

2.8.4 Determinación de los estimadores de un total poblacional de una clase (proporción o porcentaje)

a) Estimador del total poblacional ($\hat{A}_{st}$)

$$\hat{A}_{st} = N\hat{p}_{st} = N_1\hat{p}_1 + N_2\hat{p}_2 + N_3\hat{p}_3 + \cdots + N_L\hat{p}_L = \sum_{h=1}^{L} N_h\hat{p}_h$$

b) Varianza estimada de $\hat{A}_{st}$

$$V(\hat{A}_{st}) = V(N\hat{p}_{st}) = N^2V(\hat{p}_{st}) = \frac{N^2\sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}{N^2} = \sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}$$

c) Desviación estándar estimada de $\hat{A}_{st}$

$$\sigma_{\hat{A}_{st}} = \sqrt{V(\hat{A}_{st})} = \sqrt{\sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}$$

d) Límite para error de estimación de $\hat{A}_{st}$

$$Z_{1-\frac{\alpha}{2}}\sigma_{\hat{A}_{st}} = Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{A}_{st})} = Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{L} N^2{}_h \frac{\hat{p}\ \hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}$$

e) Intervalo de confianza de A_{st}

$$P\left(\hat{A}_{st} - Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}} \le A_{st} \le \hat{A}_{st} + Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{L} N^2{}_h \frac{\hat{p}_h\hat{q}_h}{n_h - 1}\frac{N_h - n_h}{N_h}}\right) = 1 - \alpha$$

2.9. Actualización

Desde 2008, el DIRCEMBOL se actualiza con periodicidad anual. El alcance de estos trabajos afecta al 100% poblacional y permite la detección de los cambios más importantes relativos tanto a la existencia como a las principales características de las unidades registradas.

2.10. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN DE LA ENCUESTA

La mayor parte de la técnica de imputación actualmente utilizada involucra la sustitución de un valor "Imperfecto" por un pronosticado. Algunas de las técnica principales de esta clase son la imputación de la media, imputación hot – deck, imputación cold – deck, imputación por regresión, y imputación múltiple. Revisaremos a continuación las principales características de estas técnicas.

2.10.1 Imputación de media global

Éste es un método simple que, para el ítem j, asigna el mismo valor, concretamente, la media del ítem del informante, $\bar{y}_{rj}$, a cada valor faltante yjk de en el conjunto ru - rj. El método puede producir un estimador puntual razonable de la población total $t_j = \sum_U y_{jk}$, pero es menos atractivo si deseamos calcular un intervalo de confianza usando un estimador de la varianza estándar. Como intuitivamente está claro, reemplazar todos

valores de faltantes para un ítem dado por la media del informante para el ítem dará un conjunto de valores con menos variabilidad que en una muestra del tamaño igual consistente en realidad de valores observados completamente. A menos que la no – respuesta es insignificante, o a menos que un estimador de la varianza modificado es utilizado, el método podría resultar fácilmente minimizado en la estimación de la varianza os cálculos aproximados de discrepancia seriamente discretos e invalidad los intervalos de confianza.

2.10.2 Imputación media de la clase

Este método funciona dividiendo el conjunto de unidad de respuesta ru en clases de imputación de forma que los elementos en la misma clase son considerados similares. Las variables auxiliares variables son usadas para la clasificación. Para un ítem j en particular, y para todos k elementos en una clase de imputación dado, los valores perdidos son reemplazados por la media del informante en esa clase. Habrá un poco de distorsión de la distribución "Natural" de los valores, pero la distorsión es menos grave que con la imputación de media global.

2.10.3 Imputación Hot – Deck y Imputación Cold - Deck

La mejora sobre los métodos de imputación de medias es pedida creando una variabilidad más auténtica en los valores imputados. En procedimientos de imputación hot – deck las respuestas perdidas son reemplazados por valores seleccionados entre los encuestados en la encuesta actual. El procedimiento Cold - deck, por otro lado, use imputaciones sobre la base de otras fuentes que la encuesta actual, por ejemplo, las encuestas más recientes o datos históricos.

Varios procedimientos de hot – deck han sido propuestos, incluyendo la imputación aleatoria global, imputación aleatoria dentro de las clases, imputación hot – deck secuencial, imputación hot – deck hierachical, y la función de distancia combinando.

2.10.4 Imputación aleatoria global

Este método funciona como sigue: para el ítem j, un valor perdido es reemplazado por un valor observado actualmente yjk, tomado de un encuestado, un donante, extraído aleatoriamente para el ítem j de un conjunto de respuestas, rj. Aunque el método da un conjunto de datos para el ítem j con una la variación natural final, no sigue que las técnica usuales pueden ser usados directamente, por ejemplo, para calcular los estimadores de la varianza y intervalos de confianza.

2.10.5 Imputación aleatoria dentro de las clases

Esto es una alternativa a las técnicas anteriores, en la cual las clases apropiadas son formadas, de forma semejante como la imputación de media de la clase. Para un elemento en una clase dado, un valor imputado es obtenido de un donante elegido al azar en la misma clase.

2.10.6 Imputación hot – deck secuencial

Por ejemplo, los Estados Unidos en la Encuesta Actual de la Población usa este método. Tiene la ventaja que un simple paso completo del archivo de datos es suficiente completar el procedimiento de imputación. Cuando un valor perdido es descubierto sobre cierto ítem, un donante es identificado retrocediendo a través del archivo de datos al elemento más cercano que muestra un valor de respuesta para ese ítem y es en la misma clase de imputación como el receptor. El procedimiento empieza con un valor cold – deck en cada clase de imputación. Un empate de este procedimiento no aleatorio es que resulta en los usos múltiples de donantes a menudo. Una mejora, llamado imputación de hot – deck jerárquico, es usada en el suplemento de ingresos de marzo de los Estados Unidos en la Encuesta Actual de la Población. En este procedimiento de imputación aleatorio, un conjunto muy grande de clases de imputación es usado así que a menudo ningún donante puede ser encontrado para un elemento de imputación requerida. Para remediar este, la imputación de clases son derrumbado en un modo jerárquico hasta que un donante es encontrado.

2.10.7 La función de distancia combinado

Este es otro procedimiento de hot – deck. Para el item j, un valor perdido yjk es reemplazado por el valor del informante clasificado por el "Más cercano", como medida por una función de distancia definida en relación con valores conocidos de las variables auxiliares.

2.10.8 Imputación por regresión

A diferencia de la técnica hot - deck, la imputación por regresión usa relaciones estimadas entre variables. Una simple aplicación de esta idea se atribuible a Buck (1960) use los datos de los encuestados para ajustar a una regresión de una variable para el cual una o más imputaciones son necesarias sobre otras variables disponibles, asumiendo que se tiene un alto valor predictivo. Los predoctores pueden ser variables de estudio (otros ítems sobre el cuestionario) o variables auxiliares. La ecuación de la regresión ajustada es usada para proceder las imputaciones. Por ejemplo, para j = 5 ítems con las y variables y1, y2, y3, y4, y5, sea yjk el valor de yj para el k – èsimo elemento. Para un cierto elemento k ε ru – rc, suponiendo que y1k y y2k son valores perdidos así que la información grabada para ese elemento se lee (-, -, y3k, y4k, y5k). Las imputaciones para los dos espacios en blanco son obtenidas de la siguiente manera: Sea $\tilde{y}_1 = \tilde{f}_1$ (y3, y4, y5) la ecuación de regresión de y1 sobre y3, y4 y y5, ajustado usando los datos para el elemento k ε rc. La correspondiente regresión estimado de y2 sobre y3, y4 y y5 está denotado $\tilde{y}_2 = \tilde{f}_2$ (y3, y4, y5). Estas dos ecuaciones, y los tres valores grabados para el elemento k producen las imputaciones $\tilde{y}$ 1k = $\tilde{f}$ 1(y3k, y4k, y5k) y y2k = $\tilde{f}$ 2 (y3k, y4k, y5k). Las imputaciones son calculadas análogamente para los espacios en blanco correspondiente a otros elementos k ε ru.

A veces uno producir residual al azar es añadido en el futuro para reflejar la incertidumbre en el valor imputado. El uso de la regresión multivariante también ha sido propugnado para la imputación cuando un elemento tiene dos o más espacios en blanco.

2.10.9 Imputación múltiple

El método de imputación mencionado produce un simple valor imputado para cada valor perdido yjk. En general, Esto distorsiona más o menos la distribución natural de valores para ese ítem. Rubin (1987) recomienda el uso de las imputaciones múltiples, es decir "Para cada dato perdido impute algunos valores, m dice.... Éstos m valores son ordenados puesto que el primer conjunto de valores imputados para los valores faltantes son usados para formar el primer conjunto de datos completados, y así que no.... Cada conjunto dato completados es analizado usando el procedimiento estándar completo de los datos". De las m estimaciones, una simple combinación de estimaciones es calculada junto con una estimación de la varianza pooled ser usado para el intervalo de estimación. Esta estimaciòn de la varianza contiene un componente que refleja la variabilidad entre los data sets terminar; puede ser visto como una expresión de la incertidumbre del estadístico sobre qué valor hacerlo/serlo atribuir. Una desventaja con la imputación múltiple es que requiere que mayor cantidad trabaje en manejo de datos y cálculo de los cálculos aproximados.

2.11 Bibliografía

1. CARL-ERIK SÄRNDAL, BENGT SWENSSON y JAN WRETMAN. 1992. “**Model assisted survey sampling**”. Editorial Springer-Verlag.
2. SHARON L. LOHR. 2000. “**Muestreo: Diseño y análisis**”. International Thompson Editores.
3. WOLTER, Kirk M. 2007. “**Introduction to variance estimation**”. Editorial Springer-Verlag.