ANEXO 1:

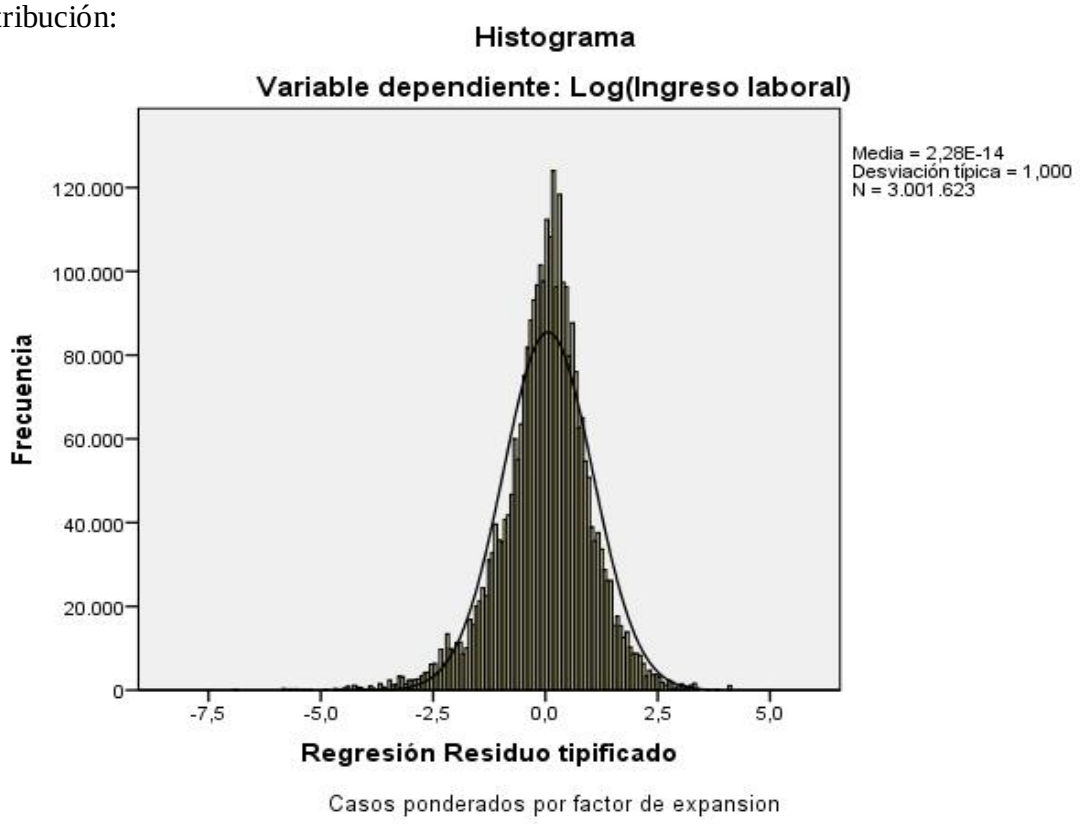
CUMPLIMIENTO DE SUPUESTOS DEL MODELO CLÁSICO DE REGRESIÓN

1. Una aclaración importante

En este punto se desarrollan las pruebas formales para verificar si nuestro modelo cumple con los supuestos del método de MCO haciendo uso del software estadístico *EViews 8*. En las secciones anteriores se utilizó el paquete estadístico SPSS 21, que permite la incorporación del "factor de expansión" calculado por el INE (ver detalles en la sección 1.5) en todos los cálculos a través de un proceso de "réplica simulada" (SPSS, Archivo de Ayuda del programa). El EViews no cuenta con esa función, por lo que los coeficientes y estadísticos estimados en este punto son ligeramente diferentes de los presentados en las secciones anteriores.

2. El supuesto de normalidad en los residuales

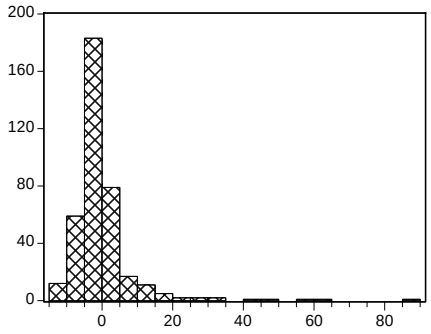
Uno de los motivos para especificar el ingreso laboral en forma logarítmica es para cumplir con el supuesto de normalidad en los residuos (y este supuesto es una condición necesaria para que sean válidas nuestras pruebas de hipótesis). Para el modelo modificado de ganancias del capital humano estimado para el 2012, los residuos presentan la siguiente distribución:



Para evaluar el cumplimiento del supuesto de normalidad en las perturbaciones vamos a utilizar la prueba de Jarque-Bera. En el panel A del Gráfico Nº 1, se muestra el resultado de esta prueba cuando se utiliza como variable dependiente al ingreso laboral; como el nivel de probabilidad asociado al estadístico de Jarque-Bera es virtualmente de cero, puede afirmarse que los residuos estimados no están distribuidos normalmente. El histograma para los residuales estimados en este modelo muestra una clara asimetría positiva (de 4.44), lo que confirma nuestra afirmación de que existe una mayor dispersión entre los ingresos laborales más altos.

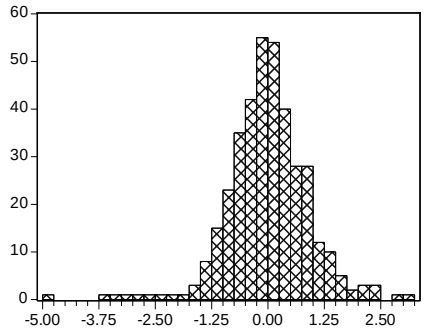
Gráfico Nº 1. Prueba de Jarque-Bera para los residuales del modelo de ganancias del capital humano, con variable dependiente diferente

A) variables dependiente: *Ingreso laboral*



Media	5.90e-16
Mediana	-1.498611
Máximo	85.93959
Mínimo	-14.77495
Dev. St.	9.302477
Skewness	4.444524
Kurtosis	32.24844
Jarque-Bera	14679.22
Probability	0.000000

B) variables dependiente: log(*Ingreso laboral*)



Media	1.89e-16
Mediana	-0.004957
Máximo	3.039783
Mínimo	-4.761263
Dev. St.	0.897521
Skewness	-0.522049
Kurtosis	6.480125
Jarque-Bera	207.3721
Probability	0.000000

FUENTE:Encuesta MECOVI 2002, INE. Elaboración Propia.

El panel B del Gráfico Nº 1 muestra el resultado de esta prueba cuando la variable dependiente es el Log (ingreso laboral). El histograma para los residuales estimados con este modelo es más parecido a la normal, pero presenta una pequeña asimetría negativa (de

-0.5) y una Kurtosis ó apuntalamiento mayor que el de la normal[42]. El nivel de probabilidad asociado al estadístico de Jarque-Bera es nuevamente de cero, por tanto, pese a la transformación logarítmica no se cumple con el supuesto de normalidad en las perturbaciones, pero es claro que con esta transformación nos acercamos más al cumplimiento del supuesto.

3. Verificación del cumplimiento de los supuestos del modelo clásico de regresión lineal

Vimos en la sección 3.6.2 que las propiedades estadísticas de insesgamiento y varianza mínima de los estimadores del modelo clásico de regresión lineal dependían del cumplimiento de 10 supuestos. Los 4 primeros eran necesarios para poder utilizar el método —es decir, aplicar sus fórmulas matemáticas—, y los cumplimos todos. Los restantes 6 supuestos hacían que nuestros coeficientes estimados sean los "mejores estimadores lineales insesgados" (Gujarati, 1997, Pg. 70); cuando estos supuestos no se cumplen es posible lograr mejores ajustes utilizando métodos diferentes al de MCO. Vamos ahora a probar si nuestro modelo cumple con esos supuestos.

En la Tabla 1. se muestran los resultados de las pruebas aplicadas para verificar el cumplimiento del supuesto 7: "*Las varianzas de los residuos condicionales a X son iguales entre sí*"[43]. Se aplicó la prueba de White (en sus dos formas) a los residuales estimados por nuestro modelo, en ninguno de los casos de detecta evidencia estadísticamente significativa de heteroscedasticidad en los datos.

Por el método gráfico puede verse que la dispersión de los residuales parece seguir un patrón aproximadamente triangular —una mayor dispersión en los valores centrales—, lo que implica heteroscedasticidad en los datos, pero no tiene el patrón que esperábamos (ver comentario en la columna 3 de la Tabla 1.).

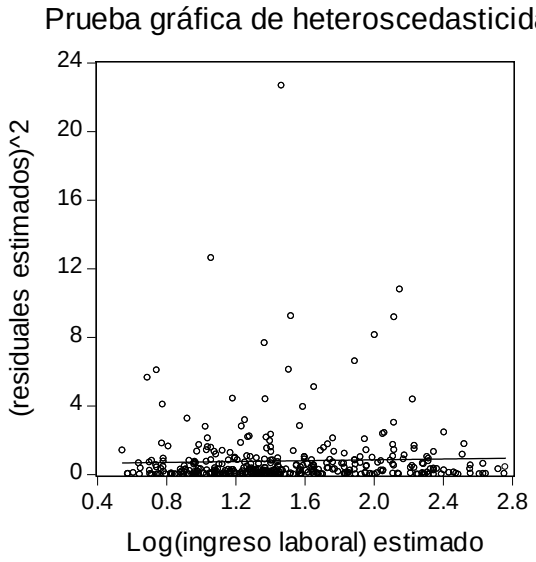
Dado que el problema de heteroscedasticidad de nuestros residuos no parece grave (la prueba general de heteroscedasticidad de White no la detecta) y que el patrón de

[42] La curva normal tiene un coeficiente de kurtosis de 3, nuestro modelo con variable dependiente Log(ingreso laboral) tiene una Kurtosis de 6.5, y el modelo con variable dependiente Ingreso Laboral lo tiene de 32.2. Por tanto, aunque la transformación logarítmica mejora la situación de la Kurtosis, no soluciona el problema de no cumplimiento del supuesto.

[43] Este supuesto es conocido técnicamente como el supuesto de **homoscedasticidad** (de *homo*=igual, y *cedasticidad*=dispersión), y cuando *no se cumple* se dice que hay heteroscedasticidad en los datos.

heteroscedasticidad es diferente al esperado, no se aplicará ninguna medida correctiva para lograr el cumplimiento de este supuesto.

Tabla 1. Verificación del cumplimiento del supuesto de "homoscedasticidad"

<table>
<tr><th>PRUEBAS APLICADAS</th><th colspan="2">HIPÓTESIS PARA LA PRUEBA Y RESULTADOS</th><th>INTERPRETACIÓN (Nivel de confianza = 95%)</th></tr>
<tr><td rowspan="3">Prueba General de heteroscedasticidad de White (sin términos cruzados)</td><td colspan="2">Hipótesis nula: Se cumple el supuesto de homoscedasticidad</td><td rowspan="3">Con un nivel de confianza de 95% y en base a la prueba de White, se acepta la hipótesis nula de que se cumple el supuesto de homoscedasticidad.</td></tr>
<tr><td>F=1.021</td><td>Prob. F(8,368)= 0.42</td></tr>
<tr><td>Obs*R^2=8.188</td><td>Prob. Chi-Square(8)= 0.42</td></tr>
<tr><td rowspan="3">Prueba General de heteroscedasticidad de White (con términos cruzados)</td><td colspan="2">Hipótesis nula: Se cumple el supuesto de homoscedasticidad</td><td rowspan="3">Con un nivel de confianza de 95% y en base a la prueba de White, se acepta la hipótesis nula de que se cumple el supuesto de homoscedasticidad.</td></tr>
<tr><td>F=1.121</td><td>Prob. F(17,359)= 0.33</td></tr>
<tr><td>Obs*R^2=18.997</td><td>Prob. Chi-Square(17)= 0.33</td></tr>
<tr><td>Método gráfico</td><td colspan="2"> 
</td><td>El grado de dispersión de los residuales respecto a los valores estimados por el modelo —los valores estimados del Log(ingreso laboral)— aumentan en los valores centrales y tienden a disminuir en los extremos. Este es un patrón de heteroscedasticidad pero no tiene la forma que esperábamos —nuestra expectativa era que el grado de dispersión en los residuales aumentara al aumentar el ingreso laboral (como lo indicaban nuestros análisis anteriores)— debido al efecto de la transformación logarítmica del ingreso (ver sección 0)</td></tr>
</table>

FUENTE:Encuesta MECOVI 2002, INE. Elaboración Propia.

Analicemos ahora el supuesto 4 del modelo clásico de regresión lineal indica que para poderse realizar la estimación de los parámetros por el método de MCO "*ninguna de las variables independientes debe ser una combinación lineal exacta de las otras*". Cómo mencionamos antes, este supuesto se denomina técnicamente: "no **multicolinealidad** perfecta entre las variables independientes". Por ejemplo, si el coeficiente de correlación entre dos variables independientes es 1 —lo que implica que cualquiera de las variables puede obtenerse a partir de la otra, usando una ecuación matemática—, las fórmulas matemáticas de MCO no pueden aplicarse.

La multicolinealidad perfecta imposibilita la utilización de MCO. La multicolinealidad alta[44] genera problemas de estimación, puede demostrarse que a medida que aumenta la multicolinealidad, los errores estándar de los coeficientes tienden a aumentar (ver Gujarati, 1997, pg. 323-328).

Dos de las variables independientes introducidas en nuestro modelo están fuertemente relacionadas: la Experiencia Laboral y la (Experiencia laboral)^2. Por lo que puede esperarse que nuestro modelo sufra de multicolinealidad alta. En la Tabla 2. puede verse que los resultados son acordes con nuestra expectativa a priori: *se detecta la presencia de multicolinealidad moderada vinculada con las variables Experiencia Laboral y la (Experiencia laboral)^2*. Así que tendremos en cuenta este hecho a la hora de probar nuestras hipótesis.

Tabla 2. Diagnóstico del Problema de Multicolinealidad

PRUEBAS APLICADAS	**INDICADORES**		**REGLAS DE EVALUACIÓN**	**INTERPRETACIÓN**
Índice de condición (IC)	Máximo valor propio=4,268	IC=18,79	Si el IC está entre 10 y 30, existe multicolinealidad entre moderada y fuerte. Por encima de 30 existe multicolinealidad severa	Nuestro modelo presenta un grado de multicolinealidad entre moderada y fuerte, pero no severa.
	Mínimo valor propio=0,012			
Factor de Inflación de varianza (FIV)	**Variable**	**FIV**	Si el FIV de una variable es mayor a 10, dicha variable presenta un alto grado de asociación lineal con alguna (o todas) las demás	Las variables Experiencia y (Experiencia)^2 presentan un alto grado de colinealidad, y son las que generan el problema detectado con el IC.
	años de estudio	6,024		
	Experiencia	13,572		
	(Experiencia)^2	13,342		
	Sexo	1,052		
	D1·(años de estudio)	5,245		
Factor de Tolerancia (TOL)	**Variable**	**TOL**	Mientras más cercano está el TOL de una variable de cero, mayor es el grado de colinealidad de esa variable con las otras	El TOL confirma que el problema de colinealidad está vinculado a las variables Experiencia y (Experiencia)^2, pero muestra que no es un problema muy grave.
	años de estudio	0,166		
	Experiencia	0,074		
	(Experiencia)^2	0,075		
	Sexo	0,950		
	D1·(años de estudio)	0,191		

FUENTE:Encuesta MECOVI 2002, INE. Elaboración Propia.

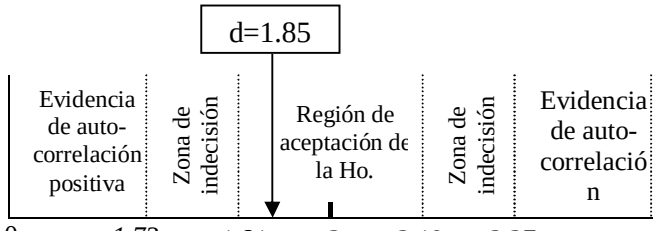
Consideremos ahora el supuesto 8: *No existe autocorrelación entre los residuos.* Si los residuos estimados a partir de nuestro modelo muestran algún patrón no aleatorio, se dice que hay autocorrelación. Un patrón no aleatorio implica que pueden usarse los valores de

[44] La multicolinealidad alta indica situaciones en las que el grado de asociación lineal entre de alguna de las variables independientes respecto a las demás es alto (mayor a 0.8).

unos residuos para estimar otros, y con esa información pueden desarrollarse modelos de estimaciones más precisas y con menor varianza que los estimadores MCO.

El grado de correlación entre los residuos depende de la forma como están ordenados. Cuando se analiza información de series de tiempo, los residuos se ordenan en función del tiempo y suelen presentarse problemas de autocorrelación. Pero en el estudio de información de corte transversal (como en nuestro caso) no suelen existir criterios definidos sobre la forma como deben ordenarse los residuos, por lo que es posible ordenarlo de forma que no se viole el supuesto de no autocorrelación.

Tabla 3. Verificación del cumplimiento del supuesto de "No Autocorrelación"

<table>
<tr><th>PRUEBAS APLICADAS</th><th colspan="3">HIPÓTESIS PARA LA PRUEBA Y RESULTADOS</th><th>INTERPRETACIÓN (Nivel de conf. = 95%)</th></tr>
<tr><td rowspan="2">Estadístico "d" de Durbin-Watson</td><td colspan="3">Hipótesis nula: No existe correlación serial de primer orden en los residuos</td><td rowspan="2">Con un nivel de confianza de 95% y en base a la prueba de Durbin-Watson, se acepta la hipótesis nula, es decir, no existe evidencia estadísticamente significativa de correlación serial de primer orden en los residuos.</td></tr>
<tr><td colspan="3"> 

Nota.- la tabla disponible para esta prueba tenia valores críticos sólo hasta tamaños muestrales de 200 o menos, debido a esto el contraste de la prueba se hizo usando los valores críticos para un tamaño muestral de 200.</td></tr>
<tr><td rowspan="6">Prueba de Breusch-Godfrey para autocorrelaciones de orden superior</td><td>Nº de rezagos incluidos</td><td colspan="2">Resultado</td><td rowspan="6">Con un nivel de confianza de 95% y en base a la prueba de Breusch-Godfrey, no se detecta evidencia estadísticamente significativa de autocorrelación.</td></tr>
<tr><td>2</td><td>Obs*R^2=2.36</td><td>Prob. Chi-Square(2)=0.31</td></tr>
<tr><td>3</td><td>Obs*R^2=2.20</td><td>Prob. Chi-Square(3)=0.53</td></tr>
<tr><td>4</td><td>Obs*R^2=2.30</td><td>Prob. Chi-Square(4)=0.68</td></tr>
<tr><td>5</td><td>Obs*R^2=3.26</td><td>Prob. Chi-Square(5)=0.66</td></tr>
<tr><td>6</td><td>Obs*R^2=3.74</td><td>Prob. Chi-Square(6)=0.71</td></tr>
</table>

FUENTE:Encuesta MECOVI 2002, INE. Elaboración Propia.

En la Tabla 3. se muestran los resultados de las pruebas que se han aplicado para verificar el cumplimiento del supuesto de autocorrelación. Puede verse claramente que la autocorrelación en nuestro caso no constituye un problema.

Examinemos ahora el supuesto 10 que dice: *el modelo de regresión está correctamente especificado, o el modelo de regresión no tiene errores de especificación*. Los errores de especificación (ver sección 3.6.2, Supuesto 10) son:

- El omitir variables relevantes
- El agregar variables irrelevantes.
- El plantear incorrectamente la forma funcional del modelo
- La existencia de una influencia mutua entre la variable dependiente y las independientes, que no está explícitamente considerada en el modelo.

Esperamos que todas las variables introducidas en el modelo sean relevantes, pero debido a la gran cantidad de factores que determinan el ingreso laboral (y que no están incluidas explícitamente en el modelo) es posible que se detecten errores de especificación por omisión de variables relevantes. En la Tabla 4. se resumen los resultados de las pruebas aplicadas para detectar posibles errores de especificación.

Tabla 4. Detección de Errores de Especificación

<table>
<tr><th>ERROR DE ESPECIFICACIÓN</th><th>PRUEBAS APLICADAS</th><th colspan="4">INDICADORES</th><th>INTERPRETACIÓN (Nivel de conf. = 95%)</th></tr>
<tr><td rowspan="6">Inclusión de variables irrelevantes</td><td rowspan="6">Significación individual de los coeficientes estimados</td><td colspan="2">Variable</td><td>Errores estándar</td><td>Prob.</td><td rowspan="6">Con un nivel de confianza de 95%, puede verse claramente que todas las variables son individualmente significativas</td></tr>
<tr><td colspan="2">años de estudio</td><td>0.0229</td><td>0.0039</td></tr>
<tr><td colspan="2">Experiencia</td><td>0.0135</td><td>0.0007</td></tr>
<tr><td colspan="2">(Experiencia)^2</td><td>0.0003</td><td>0.0265</td></tr>
<tr><td colspan="2">Género</td><td>0.0969</td><td>0.0009</td></tr>
<tr><td colspan="2">ds·(Esc-12)</td><td>0.0150</td><td>0.0443</td></tr>
<tr><td rowspan="6">Omisión de variables relevantes y forma funcional incorrecta</td><td rowspan="6">Prueba RESET de Ramsey</td><td>Regresores adicionales</td><td colspan="3">Resultado</td><td rowspan="6">Con la introducción de 4 ó 5 regresores adicionales, la prueba RESET detecta de nuestro modelo no está correctamente especificado</td></tr>
<tr><td>1</td><td>F=0.069</td><td colspan="2">Prob. F(1,370)=0.792</td></tr>
<tr><td>2</td><td>F=2.446</td><td colspan="2">Prob. F(2,369)=0.088</td></tr>
<tr><td>3</td><td>F=1.960</td><td colspan="2">Prob. F(3,368)=0.120</td></tr>
<tr><td>4</td><td>F=3.021</td><td colspan="2">Prob. F(4,367)=0.018</td></tr>
<tr><td>5</td><td>F=2.439</td><td colspan="2">Prob. F(5,366)=0.034</td></tr>
</table>

FUENTE:Encuesta MECOVI 2002, INE. Elaboración Propia.

Como se esperaba, todas las variables introducidas en el modelo son relevantes, y se detectan posibles errores de especificación por omisión de variables. Una forma de

solucionar este tipo de errores es la introducción de nuevas variables al modelo, pero en nuestro caso no se dispone de información adicional sobre otras variables importantes[45].

4. Balance del cumplimiento de supuestos

Cuando se cumplen estos supuestos puede demostrarse que:

- cada uno de los coeficientes de regresión estimados con el método de MCO a partir de una muestra siguen una distribución normal, y puede utilizarse la una distribución t para realizar prueba de hipótesis[46].
- Para la prueba de significancia global puede utilizarse la técnica del análisis de varianza (ANOVA) y la distribución F (Ver Gujarati, 1997, pg. 241-244).

Nuestro modelo presenta problemas en cuanto al cumplimiento de los supuestos del modelo clásico de regresión lineal:

a) Los residuales no están normalmente distribuidos, pero su distribución es aproximadamente normal.

b) Presenta una multicolinealidad entre moderada y fuerte, este hecho perjudica la eficiencia de las pruebas de hipótesis, pero no las invalida.

c) El patrón de distribución de los residuales estimados indica la presencia de heteroscedastidad, pero las pruebas formales no detectan evidencia estadísticamente significativa de ella. Por lo que nuestras pruebas de hipótesis se realizaran utilizando los errores estándar normales.

d) El supuesto de autocorrelación no es un problema en nuestro caso.

e) El modelo presenta posibles errores de especificación por omisión de variables relevantes.

---

[45] Por ejemplo, como señala Goleman (1996), uno de los factores determinantes más importantes del éxito laboral son las habilidades sociales. Este factor, que debería formar parte del modelo, no fue introducido porque no se cuenta con la información necesaria para cuantificarlo (de hecho, es un factor de difícil medición, hasta ahora los estudios sobre el tema han cuantificado su influencia por métodos indirectos).

[46] Pese a que los coeficientes están normalmente distribuidos no puede usarse directamente la *distribución normal* para la prueba de hipótesis porque no se conoce la varianza de regresión poblacional (σ^2). La varianza σ^2 se debe estimar a partir de los datos muestrales, esto hace que deba usarse una distribución t (con n-k grados de libertad, donde n=tamaño muestral, y k=Nº de variables introducidas en el modelo) para realizar las pruebas de hipótesis (Ver Gujarati, 1997, pg. 115-117).

Por tanto, debido a que no cumplimos con todas las condiciones ideales establecidas por los supuestos, los resultados de las pruebas de hipótesis presentadas en esta investigación *no son estrictamente válidas*, pero las pruebas de verificación presentadas en este anexo permiten ver que estamos muy cerca de cumplir adecuadamente con todos los supuestos, por lo que puede afirmarse que las inferencias realizadas en la presentación de resultados son una buena "aproximación" de los valores reales.

ANEXO 2:

DISEÑO METODOLÓGICO DE LAS ENCUESTAS DE HOGARES

1. Introducción

Bolivia ingresó como país miembro del "Programa para el Mejoramiento de las Encuestas y Medición de Condiciones de Vida en América Latina y el Caribe (MECOVI)"[47] en mayo de 1999. En noviembre del mismo año el Instituto Nacional de Estadística (INE) llevó a cabo la primera recolección de datos en el país en el marco del Programa MECOVI. Desde esa fecha, casi todos los años se realizaron encuestas a muestras aleatorias de hogares tanto en el área urbana como rural de todos los departamentos[48]. En estas encuestas se busca medir "las condiciones de vida de la población boliviana, a través de la aplicación de un cuestionario multitemático que permite investigar: las características generales sociodemográficas, salud, educación, empleo, ingresos y gastos de los miembros del hogar, y las características de la vivienda y servicios básicos de los hogares"[49], con esta información se calculan indicadores sociodemográficos y económicos, y buscan "en última instancia mejorar las condiciones de bienestar de los hogares y reducir la pobreza en el país."[50]

La información básica para esta tesis proviene del procesamiento especial de las Encuestas de Hogares realizadas por el INE entre el año 2002 al 2012. Por ese motivo, en este capítulo se expone el diseño metodológico utilizado en la realización de dichas encuestas, se describe el universo de estudio, el alcance temático y su diseño muestral, se detalla la estructura y el diseño de los instrumentos de recolección de datos (cuestionaros). También

[47] El Programa MECOVI ha sido ejecutado desde 1996 por el Banco Interamericano de Desarrollo (BID), el Banco Mundial y la CEPAL, conjuntamente con las instituciones y agencias especializadas de los países participantes. Su objetivo central "es apoyar a los países en la tarea de generar información adecuada y de alta calidad acerca de las condiciones de vida de los habitantes de la región, en cuanto a su contenido, alcance, confiabilidad, actualidad y relevancia para el diseño y evaluación de políticas" (Ver http://www.cepal.org/deype/mecovi/).

[48] El INE ensayó varias metodologías diferentes para las encuestas de hogares: entre el 1999 y el 2002 se realizaron encuestas puntuales de muestras aleatorias de hogares. Entre los años 2003 y 2004 se ejecutó la Encuesta Continua de Hogares, que hacía gran énfasis en las características de los gastos e ingresos de los hogares. En los años 2005 y 2007 se retomó la modalidad de encuestas puntuales de hogares. La Encuesta de Hogares 2008, retomó en la sección de Salud. Desde ese año, el cuestionario de encuesta no presenta grandes variaciones.

[49] Documento metodológico EH 2009, pg. 5. INE (http://www.ine.gob.bo)

[50] Ídem.

se presentan los criterios y fórmulas utilizadas por el INE para el cálculo de los factores de expansión de datos.

2. Objetivo de las Encuestas de Hogares

El objetivo general de las Encuestas de Hogares es "obtener información sobre las condiciones de vida de los hogares, a partir de la recopilación de información de variables socioeconómicas y demográficas de la población boliviana, necesarias para la formulación, evaluación, seguimiento de políticas y diseño de programas de acción en el área social."[51]

3. Marco conceptual

Región: Área geográfica utilizada para agrupar los departamentos de acuerdo a su tipo ecológico predominante. La región clasifica los departamentos en:

a) Altiplano, que comprende los departamentos de La Paz, Oruro, y Potosí.
b) Valle, que comprende los departamentos Cochabamba, Chuquisaca y Tarija.
c) Llano, que comprende los departamentos de Santa Cruz, Beni y Pando.

Área Urbana: Poblaciones con 2.000 o más habitantes

Área Rural: Poblaciones con menos de 2.000 habitantes.

Área Amanzanada: ubicadas generalmente en área urbana, presentan viviendas en un orden determinado, en espacios delimitados por calles, avenidas, etc.

Unidad Primaria de Muestreo (UPM): es un área geográfica sujeta a selección con fines de muestreo, contiene un conjunto de aproximadamente 80 a 150 viviendas en área amanzanada correspondiente a uno o varios Sectores Censales, y de 150 a 350 viviendas en área dispersa.

Vivienda Particular: se consideraron viviendas particulares a aquellas que están habitadas por hasta tres hogares (con más de tres hogares es considerada vivienda colectiva). Puede estar habitada o deshabitada al momento de realizar la visita.

Vivienda colectiva: es aquella vivienda usada como lugar de alojamiento por un conjunto de personas entre las cuales no existen vínculos familiares, y que hacen vida en común por razones de enseñanza, religión, trabajo u otros motivos. Son consideradas como tales:

[51] Documento metodológico EH 2008, pg. 6. INE. (http://www.ine.gob.bo)

hoteles, alojamientos, cuarteles, hospitales, etc. Por razones prácticas, también se considera como viviendas colectivas a aquellas que alberguen a más de tres hogares particulares. Este tipo de viviendas no es objeto de estudio de las Encuestas de Hogares.

Hogar: Unidad conformada por una o más personas, con relación de parentesco o sin él, que habitan una misma vivienda y que al menos para su alimentación dependen de un fondo común al que las personas aportan en dinero y/o especie. Una persona sola también constituye un hogar.

4. Alcance temático de las Encuestas de Hogares

Para lograr sus objetivos, las Encuestas de Hogares buscan información sobre:

- ✓ Características Sociodemográficas
- ✓ Migración
- ✓ Salud
- ✓ Educación
- ✓ Condición de actividad y características ocupacionales
- ✓ Ingresos del hogar
- ✓ Gastos del hogar
- ✓ Características de la vivienda

a) Características Sociodemográficas: busca obtener información sociodemográfica básica de cada uno de los miembros del hogar, se indaga sobre:

- Sexo
- edad
- parentesco
- idioma materno, idiomas que habla
- estado civil
- pertenencia étnica y autoidentificación

b) Migración: la encuesta identifica el cambio de residencia (al interior o exterior del país) de los miembros del hogar los últimos cinco años y el de toda la vida, así como las razones de migración.

- Migración en los últimos 5 años:

    - Interna
    - Externa
    - Razones de Migración
    - Tiempo de Residencia
- Lugar de nacimiento y migración

c) Salud: Se investiga el acceso a los servicios de salud. Se consulta sobre enfermedades respiratorias, diarreicas y de otro tipo, vacunas, gastos en salud, afiliación a seguros de salud, etc.

d) Educación: se indaga sobre:

- Alfabetismo
- Nivel de instrucción
    - Educación Básica
    - Educación Media
    - Educación Superior
- Matriculación, asistencia a un centro educativo
- Tipo de establecimiento
    - Particular o Privada
    - Fiscal, Pública, de Convenio
- Repitencia
- razones de inasistencia
- Uso individual de TICs: teléfono, celular móvil, computadora, Internet.

e) Condición de actividad y características ocupacionales: busca obtener la caracterización de la población en edad de trabajar[52] (PEA), profundizar sobre el perfil de ocupados, desocupados e inactivos, y obtener la información suficiente para las estimaciones del ingreso laboral (monetario, en especie u otras).

f) Ingresos del hogar: se indaga sobre las fuentes de ingreso no laboral (transferencias, ingresos por renta de propiedad, remesas) de los hogares. Esta información combinada con

[52] El marco conceptual usado en esta parte de las Encuestas de hogares se basan en el enfoque de la fuerza de trabajo, presentado en la sección 3.2 de este trabajo.

la obtenida sobre los ingresos laborales de los miembros (observada en el inciso anterior) permite estimar los ingresos de los hogares.

g) Gastos del hogar: recolecta información sobre el gasto corriente (como en alimentos, alquileres, educación, servicios, etc.) monetario y no monetario, y las erogaciones financieras y de capital (pago de préstamos, transferencias, ampliación de vivienda, etc.).

h) Características de la vivienda: se observa y consulta sobre:

- Tipo de vivienda (casa, departamento, choza, etc.)
- Tenencia (alquilada, propia y totalmente pagada, propia y la están pagando, etc.)
- Calidad de la construcción
- Disponibilidad de servicios
- Uso de habitaciones
- Acceso a TICs en los hogares

5. Diseño muestral

5.1. Universo de Estudio

Las Encuestas de Hogares están dirigidas "al conjunto de los hogares establecidos en viviendas particulares ocupadas de las ciudades capitales, resto urbano y área rural de Bolivia"[53], se excluye a personas que residen en viviendas colectivas (como hospitales, cuarteles, hoteles, conventos, etc.) y se incluye a personas que residen en viviendas particulares dentro de las viviendas colectivas (como serenos, porteros y cuidadores).

5.2. Cobertura geográfica

Las Encuestas a Hogares cubren las ciudades capitales, resto urbano y área rural de Bolivia.

5.3. Unidades de Observación, Análisis y de Muestreo

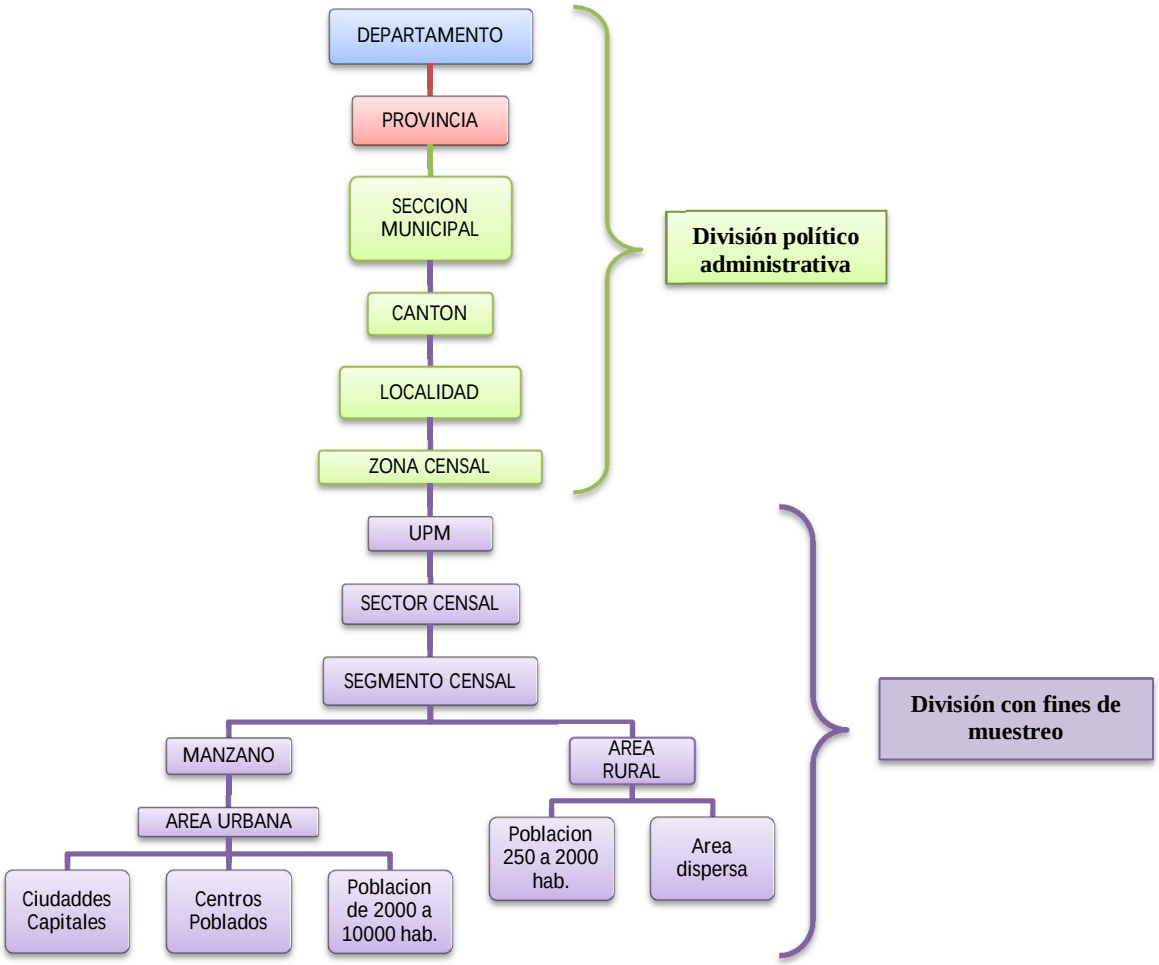
Las unidades de análisis son los **hogares** y las **personas** que forman parte de esos hogares. La unidad de muestreo en su última etapa es la **vivienda particular**, la cual "tiene permanencia fija en el tiempo y espacio, característica que la habilita para ser utilizada como unidad de selección en el diseño muestral"[54].

[53] Documento metodológico EH 2009, pg. 48. INE. (http://www.ine.gob.bo)
[54] Idem. pg. 49.

5.4. Marco Muestral

El marco muestral que se utiliza en las Encuestas de Hogares proviene del Censo Nacional de Población y Viviendas de 2001 (CNPV-2001), tiene variables que permiten la estratificación y unidades de muestreo compuesto por Unidades Primarias de Muestreo (UPMs), sectores censales, segmentos censales, y otras variables auxiliares que permiten la aplicación del muestreo complejo. Una característica importante de este marco es que es un "Marco de áreas", en el sentido de que toda Bolivia está representada en el marco[55].

Estructura del Marco Muestral



FUENTE: Documento metodológico EH 2009, pg. 49. INE. Elaboración propia

[55] Ver Documento metodológico EH 2009, pg. 49; y el Documento metodológico EH 2008, pg. 43. INE. (http://www.ine.gob.bo)

5.5. Tipo de Muestreo

El tipo de muestreo empleado en las Encuestas a Hogares es probabilístico, estratificado, por conglomerado, bietápico en el área urbana (y trietápico en el área rural)[56].

a) **Probabilístico:** porque cada una de las unidades de muestreo (las viviendas) tiene una probabilidad conocida y distinta de cero de ser seleccionada.
b) **Estratificado:** las Unidades Primarias de Muestreo (UPM) con características geográficas y poblacionales similares son agrupadas en estratos.
c) **Por conglomerados:** debido a que las UPM son conjuntos de otras unidades muestrales (las viviendas).
d) **Bietápico:** La unidad de muestreo última (la vivienda) es seleccionada en dos etapas para el área urbana:
   - Primera etapa: Selección de las Unidades Primarias de Muestreo.
   - Segunda etapa: Al interior de cada UPM se selecciona un número determinado de viviendas (unidad secundaria de muestreo).

5.6. Estratificación

El modelo de estratificación empleado combina la estratificación geográfica (urbana, rural), con una subestratificación estadística propia que emplea variables sobre características de vivienda. En este sentido se considera la unión de dos variables que forman grupos que entran dentro la condición de estratos, la primera identifica el área urbana o área rural de la unidad primaria de muestreo, y la segunda ―denominada estrato estadístico― fue construida en base a los niveles del Índice de Necesidades Básicas Insatisfechas (NBI)[57].

Los subestratos estadísticos son cuatro, y reciben las siguientes denominaciones:

1) **Estrato Alto**, son aquellas unidades muestrales que se encuentran con las necesidades básicas satisfechas.
2) **Estrato Medio Alto**, son unidades muestrales que están en el umbral de pobreza.
3) **Estrato Medio Bajo**, son unidades muestrales que están en el nivel de pobreza moderada.

[56] Documento metodológico EH 2009, pg. 50. INE. (http://www.ine.gob.bo)
[57] Estos son: Necesidades básicas satisfechas, Umbral de pobreza, Pobreza Moderada, Indigencia y Marginalidad. Ver del Documento metodológico EH 2009, pg. 51

4) **Estrato Bajo**, son unidades muestrales que están entre la indigencia y marginalidad de pobreza.

5.7. Tamaño de la Muestra

Los tamaños muestrales de cada una de las Encuestas de Hogares utilizadas se muestran en el Cuadro Nº 1 de la sección 1.5 del presente trabajo.

5.8. Proceso de Selección de la muestra

Las Unidades Primarias de Muestreo (UPM) se seleccionan de manera independiente en cada uno de los estratos explícitos. La probabilidad de selección de una UPM determinada es proporcional a su tamaño (el cual está definido por el número de viviendas que contiene).

5.9. Probabilidades de Selección y Factores de Expansión

5.9.1. Probabilidad de Selección

La probabilidad de selección de una vivienda combina las probabilidades de selección de las Unidades Muestrales en cada etapa (considerando la información disponible en el marco muestral) con los listados de actualización y selección de viviendas. En áreas urbanas, estas probabilidades vienen dadas por la siguiente expresión[58]:

$$P\left(Viv_{ijh}\right)=\left(\frac{A_h N_{jh}}{N_h}\right)\cdot\left(\frac{c}{VL_{jh}}\right)$$

donde:

$P\left(Viv_{ijh}\right)$: Probabilidad de seleccionar la i-ésima vivienda de la j-ésima UPM, del h-ésimo estrato.

A_h : Número de UPMs seleccionadas del estrato h.

N_h : Número de viviendas del estrato h en el Marco Muestral.

N_{jh} : Número de viviendas en la j-ésima UPM del estrato h.

VL_{jh} : Número de viviendas listadas en la j-ésima UPM, del estrato h.

c : Número fijo de viviendas seleccionadas en la última etapa.

[58] Ídem. pg. 54-55.

5.9.2. Factores de expansión

El inverso de la probabilidad de selección de la vivienda es el factor de expansión base. El factor final lleva los ajustes de no-respuesta en los resultados de incidencias de campo y el ajuste de la población proyectada a ese año.

Los factores de expansión base se calculan como:

$$F'_{ih} = \frac{1}{P(Viv_{ijh})} = \left(\frac{N_h}{A_h N_{jh}}\right) \cdot \left(\frac{VL_{jh}}{c}\right)$$

Una vez obtenidos los factores de expansión base, se realiza un ajuste por no respuesta.

$$F''_{ih} = F''_{ih} \cdot \left(\frac{c}{VE_{jh}}\right) = \left(\frac{N_h}{A_h N_{jh}}\right) \cdot \left(\frac{VL_{jh}}{c}\right) \cdot \left(\frac{c}{VE_{jh}}\right)$$

donde:

VE_{jh} : Número de viviendas encuestadas en la j-ésima UPM, del estrato h.

Los factores ajustados por no respuesta se corrigen, a fin de asegurar que en cada ciudad capital se obtenga la población total determinada por la proyección de población generada por el INE referida al punto medio del levantamiento, mediante la siguiente expresión:

$$F_{ih} = F''_{ih} \cdot \frac{P_h}{\hat{P}_h}$$

donde:

P_h : Población en el h-ésimo estrato, según la proyección.

$\hat{P}_h$: Población en el h-ésimo estrato, a la que expande la encuesta.

6. Estructura de los instrumentos de medición

Los cuestionarios de las Encuestas de Hogares suelen estar organizados en ocho grandes secciones, cada una de ellas dividida en “partes”, en función al alcance temático de la encuesta:

Sección 1. Parte A. Características Sociodemográficas (para todos los miembros del hogar) Permite determinar la ubicación, la localización y las características generales de las viviendas. Identificar las características de los miembros del hogar y determinar la estructura de la población por sexo, edad, núcleos familiares y relación de parentesco, etc.

Sección 2. Parte A. Migración (para todos los miembros del hogar) Investiga los desplazamientos de la población en los últimos 5 años, e indaga sobre las razones por las que se produjeron los mismos.

Sección 3. Salud - Parte A (menores de cinco años) Estas preguntas están dirigidas a conocer el estado de salud de las personas menores de cinco años: se evalúa la cobertura, estructura y gastos de los servicios de salud.

- **Parte B. (menores de tres años)** se busca conocer el grado de acceso que tienen los menores de tres años al esquema de vacunación del Programa Ampliado de Inmunización (PAI).
- **Parte C. (solo para mujeres entre trece y cincuenta años)** Permite indagar sobre las características de fecundidad de la mujer a partir de la cantidad de hijos que tubo, fechas, a quién acudió en el momento del parto, gastos en la atención pre-natal, etc.
- **Parte D. (para todos los miembros del hogar)** Investiga sobre el estado de salud de la población boliviana en el tiempo de referencia de las últimas cuatro semanas; enfermedades o accidentes, quien participó en el tratamiento, calidad del servicio, gastos, y afiliación a un seguro de salud.

Sección 4. Educación – Parte A. (personas de cinco años y más) Se indaga acerca de las características educativas de la población, principalmente aquellas referidas al alfabetismo y analfabetismo, nivel y curso de instrucción máximo alcanzado, matriculación, asistencia e inasistencia, razones de inasistencia, deserción y cobertura del sistema educativo.

Además permite establecer la cantidad de personas que asisten a establecimientos fiscales públicos, público de convenio o particulares/privados.

- **Parte B.** Identifica la frecuencia de la repitencia, sus posibles causas y las causas de la inasistencia.

- **Parte C. (personas de cinco años y más)** Este grupo de preguntas permite conocer si el informante de cinco y más años utilizó celular/móvil para comunicarse durante los últimos 12 meses. Identificar a la población que usa Internet para diferentes actividades (lugar, frecuencia y tiempo de uso), etc.

Sección 5. Empleo – Parte A (solo para personas de siete años y más) El objetivo de esta sección es clasificar a la población según su condición de actividad, permite indagar por actividades económicas y las condiciones de trabajo. Además, investigar las características del último empleo del desocupado cesante, conocer el número de personas que trabajaban en la última empresa, institución, o lugar donde el informante desempeñó sus labores.

- **Parte B. (solo para personas de siete años y más)** Esta sección nos proporciona información sobre los sectores de la actividad económica donde trabajan las personas ocupadas en su principal ocupación, la relación del trabajador con su empleo, la organización jurídica del lugar de trabajo, etc.
- **Parte C. (solo para personas de siete años y más)** indaga sobre el salario líquido que reciben los trabajadores (después de cumplir con las obligaciones de ley y los aportes a la seguridad social) y su frecuencia de recepción. Es importante mencionar que no se toman en cuenta descuentos por atrasos o anticipos.
- **Parte D. (solo para personas de siete años y más)** se busca saber cuánto gana el trabajador/a independiente en su ocupación principal. Este ingreso incluye aún los gastos que implica el tener una actividad independiente.
- **Parte E. (solo para personas de siete años y más)** Investiga la existencia de una segunda ocupación, que funciones desempeña, la actividad de la institución, si trabaja en forma dependiente o independiente, características de la administración de la empresa o institución donde trabaja, etc.
- **Parte F. (solo para personas de siete años y más)** Investiga sobre el ingreso total de la ocupación secundaria.

- **Parte G. (solo para personas de siete años y más)** busca identificar el subempleo visible y las razones por las cuales las personas en edad de trabajar se encuentran actualmente desempleadas.

Sección 6. Ingresos no laborales del Hogar – Parte A. (solo para personas de siete años y más de edad) Se busca obtener información sobre los ingresos que perciben los miembros del hogar y que no proceden de una actividad económica.

- **Parte B. (solo para personas de siete años y más de edad**) Permite identificar los Ingresos monetarios o en especie que los miembros del hogar perciben por concepto de transferencias procedentes de otros hogares.
- **Parte C. - Ingresos no laborales del Hogar – (solo para personas de siete años y más de edad**) busca medir el impacto socioeconómico de las remesas en los hogares bolivianos, a partir de la caracterización del envío, la frecuencia, recepción y destino de las remesas.

Sección 7. Gastos – Parte A. Permite estudiar las características de los gastos que realiza el hogar en la adquisición de bienes y servicios de consumo final.

- **Parte B.** Indaga sobre los gastos en educación de todos los miembros del hogar en el último mes y gastos en el último año.
- **Parte C.** Indaga sobre los gastos en alimentación dentro del hogar.
- **Parte D.** Permite conocer los gastos no alimentarios como vestimenta, transporte, comunicaciones, servicios a la vivienda, esparcimiento, servicios de cultura, etc.
- **Parte E.** Se refiere a la tenencia de bienes duraderos (equipamiento) del hogar.

Sección 8. Vivienda – Parte A. Medir la pobreza en relación a las necesidades básicas insatisfechas de acceso a servicios básicos y condiciones de la vivienda.

- **Parte B.** permite conocer la tenencia, disposición y/o acceso a la Tecnología de Información y Comunicación.

ANEXO 3:

NUESTRA HERRAMIENTA FUNDAMENTAL: LA ESTADÍSTICA

La estadística es una "caja de herramientas" para recolectar y analizar datos. Para la recolección propone protocolos de observación y de experimentación (Villaroel, 2005). Para el análisis, propone un conjunto de técnicas (fundamentadas matemáticamente) que deben emplearse en función de los objetivos y resultados esperados de la investigación:

- Si se requiere resumir los datos, se utilizan técnicas de la estadística descriptiva como la distribución de frecuencias, los promedios y las desviaciones estándar.
- Si se busca generalizar los resultados obtenidos a partir del análisis de una muestra, se utiliza las técnicas de la inferencia estadística. Estas técnicas permiten evaluar los errores que pueden cometerse al generalizar los resultados de la muestra, y probar hipótesis sobre los parámetros poblacionales usando esos valores generalizados.

a) Técnicas para el análisis descriptivo de datos

1. Medidas de posición: la media, la mediana y los percentiles.

La **media** es la medida de tendencia central para variables cuantitativas más utilizada, y es un concepto familiar para casi todas las personas. Se obtiene sumando los elementos del conjunto observado y dividiendo ese total entre el número de elementos. Su principal desventaja es que es muy sensible a la presencia de valores extremos ó atípicos, los valores atípicos tienden a alejar a la media del punto cerca del cuál se ubican la mayoría de las observaciones haciendo que esta medida pierda representatividad.

La **mediana** es otra medida de tendencia central para variables cuantitativas[59]. A diferencia de la media, la mediana no se ve afectada por la presencia de valores atípicos en los datos, lo que la vuelve una medida de tendencia central en ocasiones más robusta que la media. Para calcularla se ordenan los datos en orden ascendente o descendente y se ubica el valor que ocupa la *posición central*, el objetivo es encontrar un número que marque el límite y divida al conjunto ordenado en dos partes de igual cantidad de elementos. Si el número de elementos del conjunto es:

- Impar, la mediana es el valor del elemento que ocupa la posición central

[59] La mediana también puede calcularse para datos cualitativos ordinales (Moya, 1996, pg. 208).

- Par, la mediana es el promedio de los dos elemento que ocupan la posición central

La mediana divide el conjunto de datos en dos partes iguales (normalmente se dice que 50% de los datos de observados están por debajo de la mediana y el restante 50% está por encima).

Los **percentiles** son conceptualmente similares a la mediana, hay percentiles de 1 a 99, correspondientes a los porcentajes de 1 a 99. Así, el percentil 25 es el valor que supera a no más del 25% de los datos y es superado por el restante 75%. El percentil 1 es el valor que supera a no más del 1% de los datos y es superado por el 99% restante (Moya, 1996, Pg. 223-224).

2. Medidas de dispersión: La desviación media y la desviación estándar.

Las medidas de dispersión —como la desviación media absoluta y la desviación estándar— indican el grado de agrupamiento de los datos en torno a una medida de tendencia central. Un valor alto de la desviación media (o de la estándar) indica una gran dispersión; y un valor bajo refleja un gran agrupamiento (valores muy parecidos entre si).

Para calcular la **desviación media absoluta** se obtiene la diferencia (en valor absoluto) entre cada punto observado y la media, y se calcula el promedio de esas diferencias. Indica, por tanto, que tan lejos en promedio se encuentra cada observación respecto a la media.

La **desviación estándar** se calcula también a partir de las diferencias (también llamadas desvíos) de cada observación respecto a la media. Se suman los cuadrados de cada desvío, y ese total se divide entre el número de observaciones menos 1. De ese cociente se extrae su raíz cuadrada positiva, ese número es la desviación estándar. Una característica importante de este estadístico es que las observaciones más alejadas de la media tienen una gran influencia en su valor (mientras más alejado esté un dato de la media, mayor será su desvío correspondiente y mucho mayor el cuadrado de ese desvío).

La desviación estándar es la medida de dispersión más utilizada en la práctica, desafortunadamente no tiene una interpretación intuitivamente obvia[60]. En esta tesis la

[60] Para obtener una "idea" de la dispersión de los datos a partir de de la desviación estándar puede usarse el teorema de Tchebyshev (que predice el número mínimo de observaciones que hay en el intervalo formado por la media y un número especificado de desviaciones estándar, independientemente de la distribución de

usaremos para comparar la dispersión entre subgrupos de una misma variable, por ejemplo, para comparar la dispersión entre los ingresos de las personas que son bachilleres y los ingresos de quienes sólo cursaron primaria, aquel grupo que tenga la desviación estándar más alta será el que presente mayor dispersión en sus datos.

El uso de desviación estándar en la comparación del grado de dispersión entre grupos tiene un problema, esta medida de dispersión "es significativa solamente en relación con la media respecto a la cual se calcula" (Murillo, 1990, pg. 62). Si las medias de los grupos son muy diferentes, la desviación estándar puede conducirnos a conclusiones equivocadas. Una alternativa posible es usar el **coeficiente de variación** (que es la desviación estándar dividida entre la media) como medida de dispersión relativa.

b) Técnicas de análisis gráfico

1. Diagrama de cajas

El diagrama de cajas es una técnica gráfica univariante que permite comparar distribuciones de datos entre diferentes grupos (de la misma variable) y detectar valores extremos o típicos. Se construye a partir de 5 estadísticos de resumen:

- El percentil 75 ó cuartil superior[61] (Q_3)
- El percentil 25 ó cuartil inferior (Q_1)
- La mediana (Me)
- El extremo superior
- El extremo inferior

En la siguiente figura se muestra la estructura de un diagrama de caja. Dentro de la caja se encuentra el 50% de los datos observados. El espacio comprendido entre el cuartil superior y el inferior es más grande mientras mayor sea la dispersión de los datos (esta es una medida de dispersión denominada **recorrido intercuartílico**, $Ri=Q_3 - Q_1$).

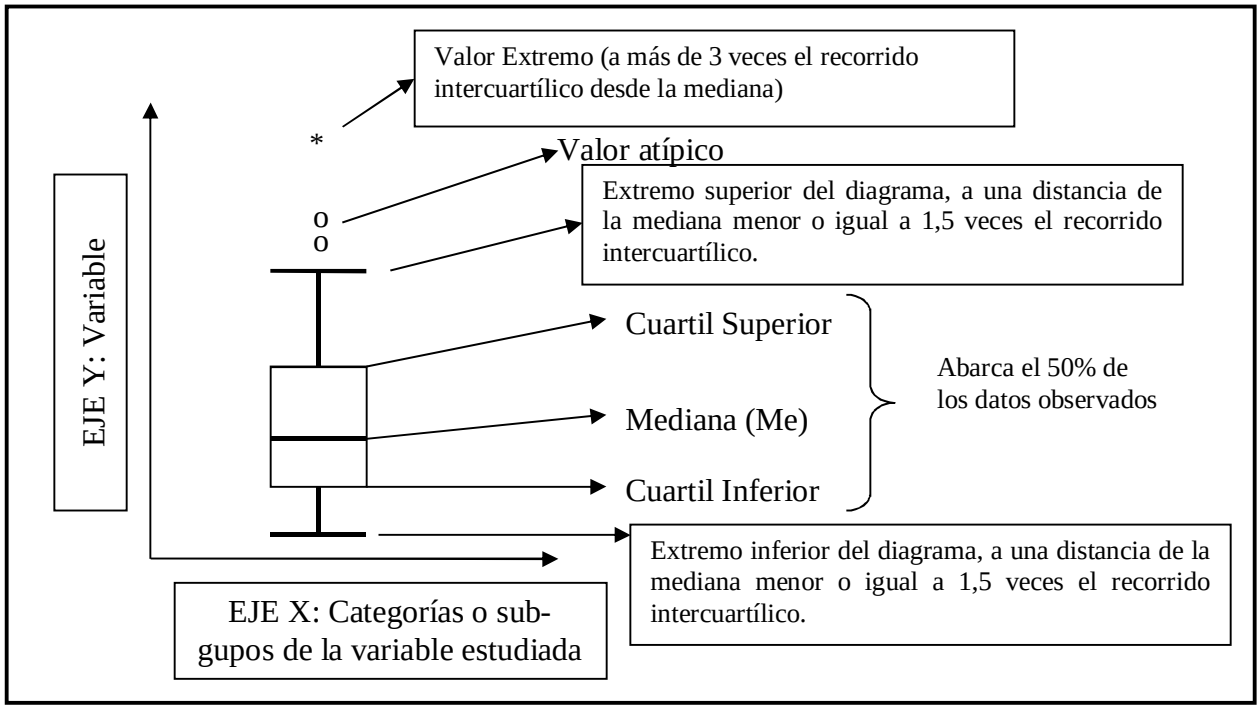
---

frecuencias del conjunto observado) o las reglas empíricas (que se basan en las características de la distribución de los datos). Para detalles puede consultarse Moya (1996, Pg. 286-289)

[61] Los cuartiles son conceptualmente similares a la mediana, "son valores que dividen un conjunto de datos ordenados en forma ascendente o descendente en 4 partes iguales" (Moya, 1996, pg.210); así el primer cuartil (ó cuartil inferior) es equivalente al percentil 25, el segundo cuartil es equivalente a la mediana, y el tercer cuartil (ó cuartil superior) es igual al percentil 75.

A partir de la caja se trazan líneas hacia arriba hasta el mayor valor observado que diste 1,5 veces (o menos) el recorrido intercuartílico, ese valor será el extremo superior del diagrama. Los valores situados por encima de dicho valor límite son considerados atípicos. Los valores observados situados a una distancia mayor a 3 veces el recorrido intercuartílico desde la mediana —es decir, valores muy alejados del resto— se denominan valores extremos[62]. Para el trazado de la parte inferior del diagrama se sigue un procedimiento similar.

Estructura de un Diagrama de Caja



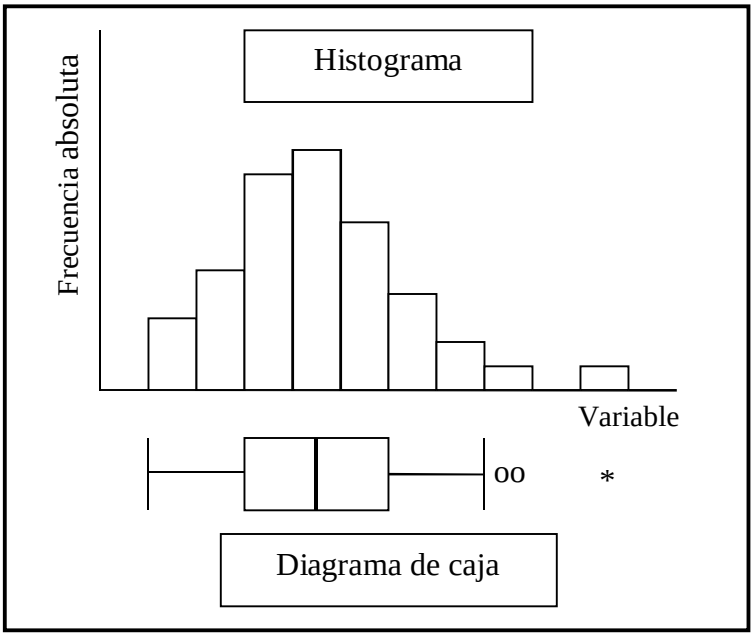
La forma del diagrama de caja está muy relacionada con la distribución de datos. El histograma es muy usado para representar la distribución de datos, en la siguiente figura se ilustra la relación entre dicha técnica gráfica y el diagrama de caja.

Puede verse que la caja coincide contiene los valores que se presenta con mayor frecuencia, y que dentro de los límites externos del diagrama se encuentran la gran mayoría de los valores. Así, un diagrama de caja es una representación simplificada de la distribución real

[62] Es frecuente, durante el proceso de depuración de una base de datos, verificar si los valores extremos corresponden a observaciones reales o a errores durante la medición o transcripción.

de frecuencias, y su simplicidad lo vuelve muy útil para la comparación (como se verá durante la presentación y el análisis de resultados).

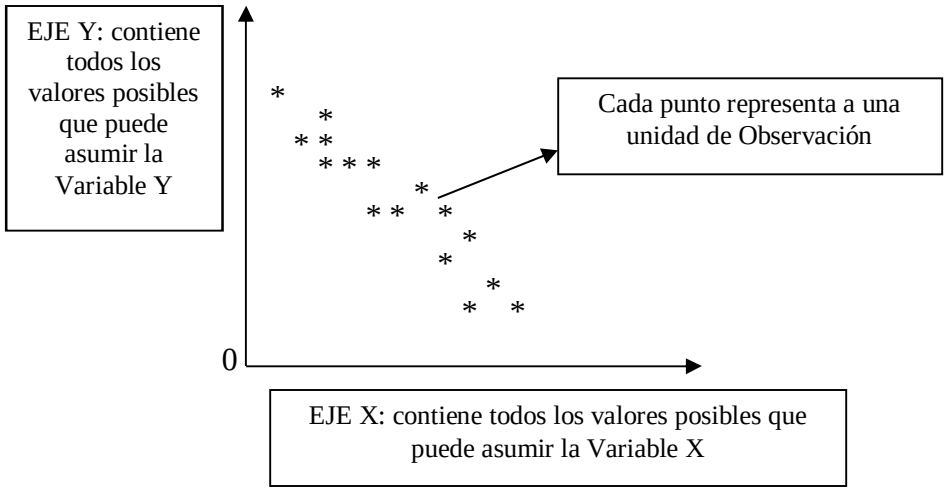
Histograma y Diagrama de Caja



2. Diagramas de dispersión

El diagrama de dispersión es una técnica gráfica bivariante ó trivariante que nos ayuda a entender las relaciones entre las variables de un conjunto de datos. En un eje de coordenadas, donde cada eje corresponde a una variable (y representa los valores que esta puede tomar), se grafican las unidades observadas. Cada punto representa una unidad observada, y sus coordenadas indican los valores de las variables correspondientes a esa unidad observada.

Diagrama de Dispersión



Si los puntos del diagrama muestran una tendencia decreciente significa que los valores altos de una variable están asociados (ó corresponden) a valores bajos de la otra variable, es decir, existe una relación inversa entre estas variables. Si, por el contrario, los puntos tienen una tendencia ascendente, a los valores altos de una variable le corresponden valores altos de la otra (relación directa).

3. Líneas de tendencia

Una línea de tendencia representa una tendencia en una serie de datos obtenidos a través de un largo período. Este tipo de líneas puede decirnos si un conjunto de datos en particular (como por ejemplo, el PIB, el precio del petróleo o el valor de las acciones) han aumentado o decrementado en un determinado período. Se puede dibujar una línea de tendencia a simple vista fácilmente a partir de un grupo de puntos, pero su posición y pendiente se calcula de manera más precisa utilizando técnicas estadísticas como las regresiones lineales. Las líneas de tendencia son generalmente líneas rectas, aunque algunas variaciones utilizan polinomios de mayor grado dependiendo de la curvatura deseada en la línea.