UNIVERSIDAD BÍBLICA LATINOAMERICANA
PENSAR • CREAR • ACTUAR

BACHILLERATO EN CIENCIAS TEOLÓGICAS
BACHILLERATO EN CIENCIAS BÍBLICAS

LECTURA SESIÓN 13

CTX 105 METODOLOGÍA DE LA INVESTIGACIÓN

Pereira M., Rodney. "Análisis de los datos" En *Pautas metodológicas para investigaciones cualitativas y cuantitativas en ciencias sociales y humanas*, coordinado por Mario Yapu, 264-290. La Paz: Programa de Investigación Estratégica en Bolivia, 2010.

Reproducido con fines educativos únicamente, según el Decreto 37417-JP del 2008 con fecha del 1 de noviembre del 2012 y publicado en La Gaceta el 4 de febrero del 2013, en el que se agrega el Art 35-Bis a la Ley de Derechos de Autor y Derechos Conexos, No. 6683.

La asignación de códigos y sus procedimientos deben documentarse en un informe detallado que se constituya en una guía para el proceso de codificación y para localizar las variables e interpretar los datos durante el análisis. En este documento debe incorporarse la codificación de las no respuestas y su tratamiento en la clasificación (Hernández/Fernández/Baptista 1999: 320).

Efectuada la codificación que permite la clasificación de la información contenida en los cuestionarios, ésta se procede a guardarla en algún medio magnético.

4. ANÁLISIS DE LOS DATOS

Una vez realizada la recolección de la información de tipo cuantitativo, comienza la fase del análisis e interpretación. Este proceso puede resultar complejo dependiendo del tipo de análisis que se pretenda realizar: descriptivo, inferencial, causal univariado o multivariado, pruebas paramétricas, entre los más importantes. El alcance de este apartado se orienta a presentar una introducción a los aspectos básicos sobre la presentación de la información y el análisis descriptivo.

4.1. Presentación de la información cuantitativa

Ya sea que se maneje información de una muestra o de una población, existe una regla práctica: cuando el conjunto de datos contenga veinte o más observaciones, la mejor manera de observar los datos es presentarlos de forma resumida elaborando tablas y gráficas apropiadas que faciliten aproximar sus principales características (Berenson/Levine 1992: 97).

4.1.1. Distribución de frecuencias

Es una tabla resumen en la que se ordenan los datos por clases o categorías. En cada clase o categoría existirá más de un dato denominado frecuencia que cumple con el atributo fijado en la clase o categoría. Debe tenerse en cuenta que al agrupar los datos se pierde alguna información de las observaciones individuales.

En la construcción de una tabla de frecuencias se debe considerar tres aspectos: La selección del número adecuado de clases o categorías, la definición del intervalo de clase, y los límites de cada clase a fin de evitar traslapes.

Selección del número de clases:

El número de clases o categorías para variables cuantitativas depende de la cantidad de datos; cuanto mayor es el número de observaciones, mayor el número de clases. Si existen pocas clases, la información que se obtiene es reducida y si hay muchas se tendrá una baja concentración de datos no permitiendo una buena descripción de éstos. Si las variables son cualitativas, el número de categorías está en función de los agrupamientos que previamente definió el investigador. Por ejemplo, para el caso de estado civil, las categorías pueden ser: soltero, casado, conviviente, separado, divorciado, viudo.

Intervalo de clase:

Es el tamaño o anchura de cada clase, sólo se aplica a variables cuantitativas. Para su cálculo, un camino es que todas las clases tengan el mismo tamaño, en este caso se busca la diferencia entre el dato mayor y el menor (denominado rango) y se divide entre el número de clases que se desean.

$$\textit{Tamaño de clase} = \frac{\textit{Rango}}{\textit{Número de clases}}$$

Si el resultado fuera un número decimal, se debe redondear al entero superior.

Una segunda opción se presenta cuando existen elevadas diferencias entre los datos y se desea analizar algunos grupos con especial atención; en esta situación los intervalos de clase no requieren ser del mismo tamaño.

Límite de clase:

Son las acotaciones para cada intervalo de clase que permiten agrupar los datos sin que éstos se sobrelapen, es decir que sólo pertenezcan a una clase. Existen intervalos cerrados, que son los que tienen un límite inferior y superior, y abiertos, en los que no se acota el límite superior o inferior ("más de" o "menos de").

Marca de clase:

Es el punto o valor que se encuentra en la mitad de los límites de cada clase y es representativo de los datos de esa clase. Se calcula de la siguiente manera:

$$Marca\ de\ clase = \frac{Límite\ inferior + Límite\ superior}{2}$$

4.1.2. Distribución de frecuencia relativa y acumulada

A partir de la distribución de frecuencias, con el propósito de mejorar el análisis, se puede elaborar las distribución de frecuencias relativas que se obtiene dividiendo la frecuencia de cada una de las clases o categorías entre el número total de observaciones, el resultado será un número menor a uno y mayor a cero denominado proporción. La expresión de la frecuencia relativa se puede expresar en porcentajes multiplicando cada frecuencia por cien. La suma de las frecuencias relativas da lugar a las frecuencias acumuladas

EJEMPLO No. 26

Tomando como referencia el ejemplo donde se propone analizar los factores que inciden en el nivel de ingreso de los ocupados de la ciudad de La Paz, la Encuesta de Hogares-Calidad de Vida del programa MECOVI (INE) ha establecido una muestra para la ciudad de La Paz de 420 viviendas donde viven 1800 personas; de éstas, 506 son ocupadas y constituyen el objeto de estudio.

Distribución de frecuencias por grupos de edad:

Para elaborar esta tabla se observó en la muestra de ocupados que las edades están comprendidas entre los 10 y 80 años, estableciéndose un rango de 80 – 10 = 70. Se definió agrupar la población en cinco intervalos de clase (con propósito didáctico). Con esta información se calculó el intervalo o tamaño de clase.

$$Tamaño\ de\ clase = \frac{Rango}{Número\ de\ clases} = \frac{70}{5} = 14$$

Los límites de clase se establecieron en base al tamaño de clase, el límite inferior de la primera clase fue de 10 años y el superior hasta menos de 24 años (23.999) (10+14). Con este mismo procedimiento se establecieron los límites para la segunda clase (24 hasta menos de 38 años) y así sucesivamente. Se evitó que existan traslapes entre los intervalos.

Definidos los aspectos anteriores se estructuró la siguiente tabla de frecuencias absolutas, relativas y acumuladas de la muestra de la población ocupada por grupos de edad.

Continúa en la página siguiente

Viene de la página anterior

Tabla de frecuencias por grupo de edad:			
Grupo de edad	**Frecuencia absoluta**	**Frecuencia relativa (%)**	**Frecuencia relativa acumulada (%)**
10 – 23 años	113	22,3	22,3
24 – 37 años	170	33,6	55,9
38 – 51 años	134	26,5	82,4
52 – 65 años	69	13,6	96,0
66 – 80 años	20	4,0	100,0
Total	**506**	**100,0**	

Distribución de frecuencias por grupos de ingresos:

En la elaboración de esta tabla se observó que el rango tenía una amplitud de Bs. 35.675 y que existía una elevada dispersión de los datos, los que se concentraban en ingresos inferiores a Bs. 2.000. Por esta razón, se optó en delimitar diferentes tamaños de clase que permitieran cubrir todo el universo y que el número de clases no fuera elevado; se optó por cinco grupos. Los resultados se muestran a continuación.

Tabla de frecuencias por grupo de ingreso:			
Grupo de ingreso	**Frecuencia absoluta**	**Frecuencia relativa (%)**	**Frecuencia relativa acumulada (%)**
00 - 500	185	36,6	36,6
501 - 1000	156	30,8	30,8
1001 - 1500	59	11,7	11,7
1501 - 2000	28	5,5	5,5
2001 - 4000	42	8,3	8,3
4001 - 36000	36	7,1	7,1
Total	**506**	**100,0**	**100,0**

4.1.3. Tablas de contingencia

Estas tablas relacionan o cruzan dos variables que pueden ser cuantitativas o cualitativas, lo cual permite, dependiendo del propósito de la investigación, tener una visión de los datos en forma mas desagregada e inferir de manera preliminar un primer nivel de relación entre las variables.

4.1.4. Supertablas

Son tablas de contingencia donde se muestran tres o más variables, las que pueden ser una combinación de variables cuantitativas (previamente ordenas por clases o categorías) con variables cualitativas.

EJEMPLO No. 27

Con base en los datos provenientes de la Encuesta de Hogares, se puede armar una tabla de contingencia que cruce los grupos de edad por sexo, los resultados se pueden mostrar en valores absolutos y relativos (porcentajes).

Tabla de contingencia Edad por Sexo

Grupos de Edad	Frec. Absolutas			Frec. Relativas Vertical			Frec. Relativas Horizontal		
	sexo		Total	sexo		Total (%)	sexo		Total (%)
	Hombre	Mujer		Hombre (%)	Mujer (%)		Hombre (%)	Mujer (%)	
10 - 23	52	61	113	19,5	25,4	22,3	46,0	54,0	100,0
24 - 37	87	83	170	32,7	34,6	33,6	51,2	48,8	100,0
38 - 51	73	61	134	27,4	25,4	26,5	54,5	45,5	100,0
52 - 65	39	30	69	14,7	12,5	13,6	56,5	43,5	100,0
66 - 80	15	5	20	5,6	2,1	4,0	75,0	25,0	100,0
Total	266	240	506	100,0	100,0	100,0	52,6	47,4	100,0

En la tabla se puede observar (frecuencias relativas horizontal) que de la muestra de la población ocupada de la ciudad de La Paz, el 52.6% son hombres y el 47.4% son mujeres. Las frecuencias relativas verticales muestran, para el caso de los hombres, que las mayores participaciones (porcentajes) se presentan en los grupos de edad comprendidos entre los 24 a 37 años (32.7%) y entre los 38 a 51 años (27.4%), en el caso de las mujeres las mayores participaciones se presentan en los tres primeros grupos de edad. Las frecuencias relativas horizontales revelan que en el grupo de edad comprendido entre los 10 a 23 años la participación (porcentaje) de la mujer (54%) es mayor a la de los hombres (46%) mientras que en los dos últimos grupos de edad, la participación de los hombres es mayor.

Una tabla de contingencia más amplia (Supertabla) podrá ser aquella que cruce los grupos de ingreso con grupos de edad clasificados por sexo.

Continúa en la página siguiente

Viene de la página anterior

Tabla de contingencia Grupos de Ingreso por Grupos de edad por sexo

Grupos de ingreso	Grupos de edad por sexo										Totales
	10 - 23		24 - 37		38 - 51		52 - 65		66 - 80		
	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	
00 - 500	27	36	19	42	9	22	11	12	5	2	185
501 - 1000	18	19	33	21	21	19	9	11	4	1	156
1001 - 1500	5	4	16	5	11	9	6	1	1	1	59
1501 - 2000	2	2	6	3	7	3	2	2	1		28
2001 - 4000			5	9	13	1	7	3	3	1	42
4001 - 36000			8	3	12	7	4	1	1		36
Totales	**52**	**61**	**87**	**83**	**73**	**61**	**39**	**30**	**15**	**5**	**506**

4.2. Representaciones gráficas

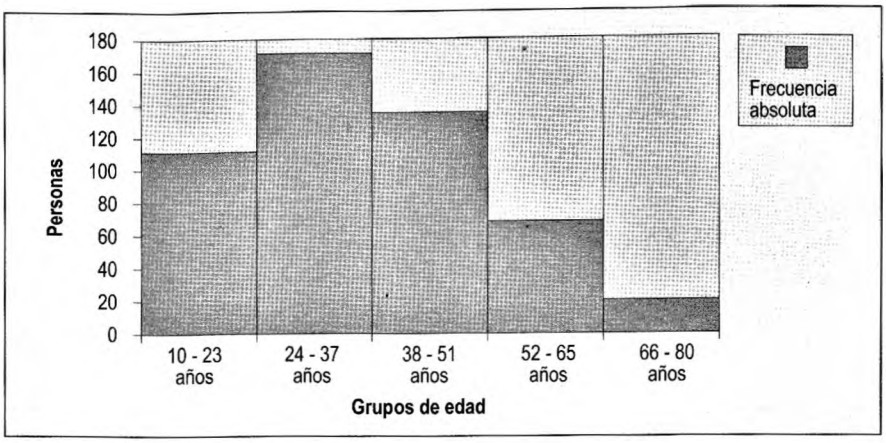
La información ordenada en tablas de frecuencia o de contingencia o cuando se trata de un número reducido de datos se puede representar en forma de gráficas, lo cual permite destacar aspectos significativos de un conjunto de datos, bien se dice "una imagen vale mas que mil palabras". Las formas más comunes de representar los datos son los histogramas, los polígonos, diagramas de barra, de pastel, radiales.

4.2.1. Los histogramas

Son una forma de representación gráfica en dos ejes mediante barras verticales, cuyo ancho está definido por los límites de clase. En el eje horizontal se representa las categorías (intervalos de clase), por ejemplo, los intervalos de edad o de ingreso, y en el eje vertical se representan el número, proporción o porcentaje de observaciones para cada intervalo de clase. La altura de cada barra representa la frecuencia de cada clase o categoría. Si en el eje vertical se representa los valores absolutos, el histograma se denomina de frecuencias absolutas; si son porcentajes, se denomina de frecuencias relativas o de proporciones.

De la tabla de frecuencias del EJEMPLO No. 26 de la población ocupada por grupos de edad se puede graficar un histograma de frecuencias absolutas que tendrá la siguiente forma:

HISTOGRAMA
Frecuencias Absolutas

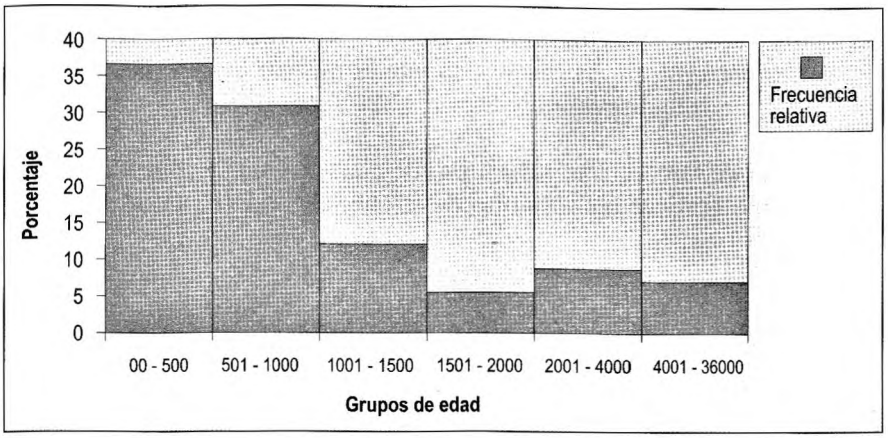


En el histograma se aprecia una aproximación de la distribución de la población ocupada por grupos de edad. Se observa que el mayor número de personas ocupadas se encuentra en el grupo de edad comprendido entre 24 a 37 años (170 personas), le sigue el grupo de 38 a 51 años (134 personas) y el de 10 a 23 años (113 personas), mientras que en los grupos comprendidos entre 52 y 65 años y 66 a 80 años se encuentran en cada uno menos de 70 personas.

Esta distribución permite inferir que más de 80% de lo ocupados se encuentran en los tres primeros grupos de edad (entre 10 y 51 años), que al aumentar la edad hasta los 37 años crece la ocupación para luego disminuir paulatinamente conforme la población se hace más vieja. El mayor número de ocupados entre el segundo y tercer grupo puede responder a varios factores, como un mayor grado de educación y experiencia, la necesidad de generar ingresos para la familia dado que la mayor parte de esta población no son solteros.

También se puede graficar un histograma de frecuencias relativas por grupos de ingreso contenida en el EJEMPLO No. 26:

HISTOGRAMA
Frecuencias Relativas



En este histograma se observa, en base a la muestra considerada en el ejemplo, que los ingresos de la mayor parte de la población ocupada (68%) no sobrepasan Bs. 1.000, en efecto 37% de la población ocupada tiene un ingreso hasta de Bs. 500 y 31% entre Bs. 501 y Bs. 1.000. En los siguientes grupos, que corresponden a ocupados con cada vez mayores de niveles de ingreso, su representatividad en el total de la muestra se reduce conforme se incrementa el ingreso. Así se observa que en el grupo de mayores ingresos se encuentra 7% de los ocupados. Este histograma permite una primera aproximación a la distribución del ingreso entre los ocupados, destacándose que la mayor parte de los ocupados tienen bajos ingresos y sólo 15% tiene ingresos por encima de Bs. 2.000.

4.2.2. Polígonos de frecuencia

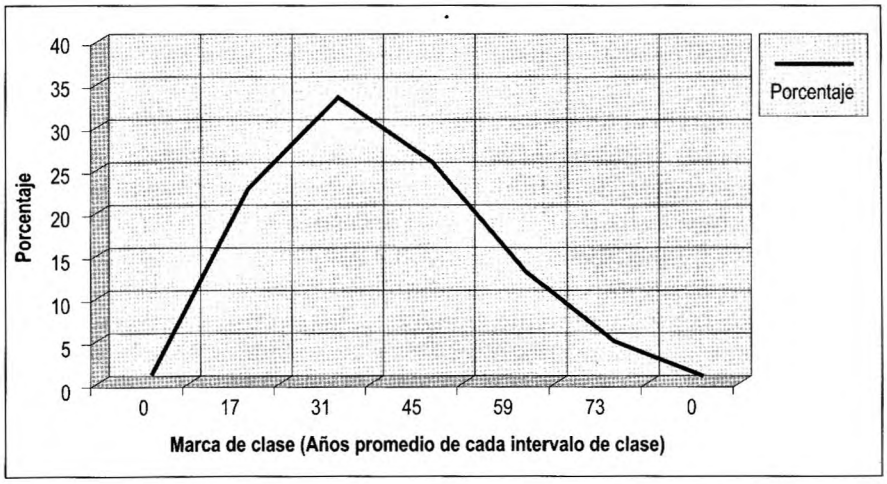
Son otra forma de representación gráfica de línea continua que se construye de modo análogo al histograma, representado en el eje horizontal las marcas de clase y en el vertical las frecuencias absolutas o relativas. Cuando se trata de propósitos comparativos entre dos poblaciones o muestras, se recomienda utilizar las frecuencias relativas.

En la construcción de los polígonos se debe tener presente que es la representación de la forma de una distribución en particular. El área debajo del polígono incluye la totalidad de las observaciones sea que se expresen en valores absolutos o relativos, por lo que es necesario conectar los puntos medios primero y último con el eje horizontal. Asimismo, el eje vertical debe mostrar el cero para no distorsionar o representar en forma equivocada los datos.

Los polígonos permiten una primera aproximación a la forma de distribución de los datos. Cuando su forma es achatada implica la existencia de dispersión de los datos, en tanto que cuando su forma es menos achatada y apuntalada, los datos se concentran alrededor de una medida de tendencia central. Asimismo, se puede ilustrar las propiedades de tendencia central, dispersión y forma.

De la tabla de frecuencias por grupos de edad sobre la población ocupada se puede graficar un polígono de frecuencias relativas que tendrá la siguiente forma:

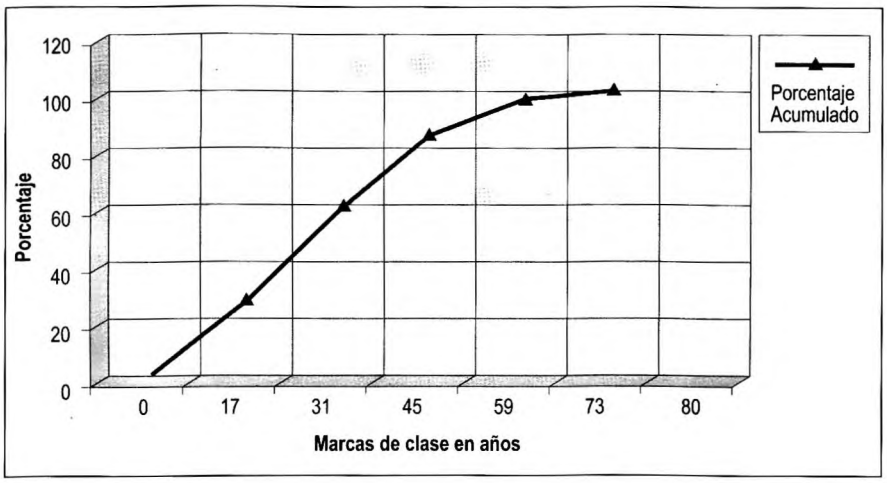
POLÍGONO
Frecuencias Relativas



En este polígono se aprecia que mas de 30% de la población ocupada tiene una edad de alrededor de 31 años y que no existe una elevada dispersión de las edades respecto a este valor. Asimismo, se observa que la distribución porcentual de las edades se aproxima a una curva normal con cierto sesgo hacia la derecha.

También se puede graficar un polígono de frecuencias acumuladas denominado ojiva. Considerando el ejemplo anterior, en el eje horizontal se representan las marcas de clase y en el vertical los porcentajes acumulados. Esta información proviene de la tabla de frecuencias acumuladas del ejemplo No. 26.

POLÍGONO
Frecuencias Acumuladas (Ojiva)



En la gráfica se observa que más de 95% de la muestra de ocupados de la ciudad de La Paz tiene hasta 59 años, que los ocupados con menos de 40 años representan más de 60% y los de menos de 17 años un poco más de 20% .

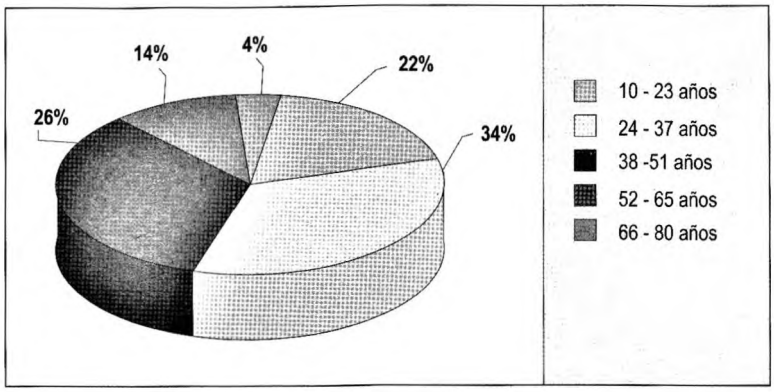
4.2.3. Gráficas de pastel

Este tipo de gráficas son utilizadas para mostrar proporciones o porcentajes de participación de una categoría o intervalo de clase en la población o muestra total que es objeto de estudio. Se expresa a través de un círculo que se divide en proporciones que corresponden al porcentaje con que cada categoría participa en el universo seleccionado (población o muestra).

La gráfica de pastel, denominada también "torta", permite visualmente detectar la representatividad de cada categoría en el universo sujeto de estudio.

Mediante una gráfica de pastel o torta y en base a la tabla de frecuencias relativas del ejemplo No. 26 se puede representar la proporción (porcentaje) con que cada grupo de edad de la población ocupada de la muestra de la ciudad de La Paz participa en la misma.

GRÁFICA DE PASTEL O TORTA
(% de participación de cada grupo de edad en la población ocupada)



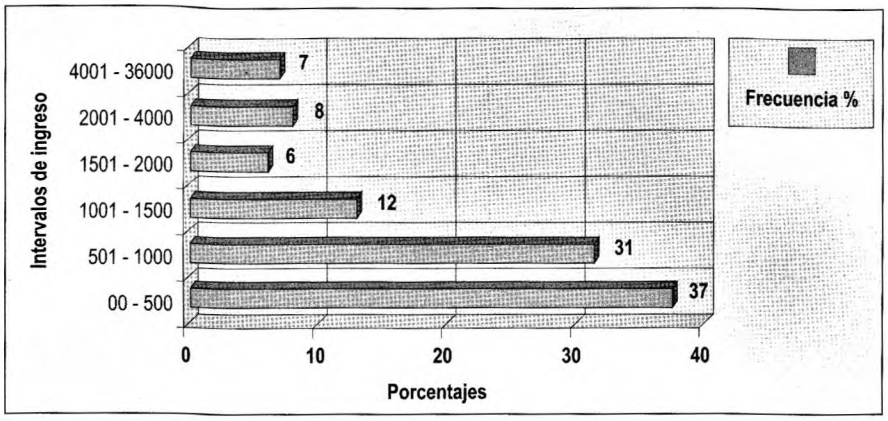
La gráfica permite una aproximación al perfil etáreo de la población ocupada, destacando que los mayores porcentajes corresponden a los tres primeros grupos de edad, comprendidos entre los 10 a 51 años (22%, 34% y 26%, respectivamente), en tanto que los grupos con más de 51 años muestran porcentajes de participación reducidos. Llama la atención que en el grupo de 66 a 80 años (edad de jubilación) se encuentren personas ocupadas, situación que puede deberse a la falta de rentas o de redes de seguridad familiar.

4.2.4. Gráfica de barras

Una forma alternativa de mostrar la representatividad de una categoría en la población total o muestra es la gráfica de barras, que puede ser de barras horizontales o verticales. Esta gráfica se construye en dos ejes, cada barra representa una categoría (intervalo de clase) que debe tener el mismo ancho para no confundir al lector. Si se opta por la forma horizontal, en el eje vertical se colocan las categorías y en el horizontal las frecuencias absolutas o relativas (porcentajes); si se prefiere la forma vertical, se debe invertir el significado de los ejes.

Para mostrar las frecuencias relativas de la muestra de la población ocupada por intervalos de ingreso la gráfica de barras es la siguiente:

GRÁFICA DE BARRAS
(% de ocupación en cada grupo de ingreso)



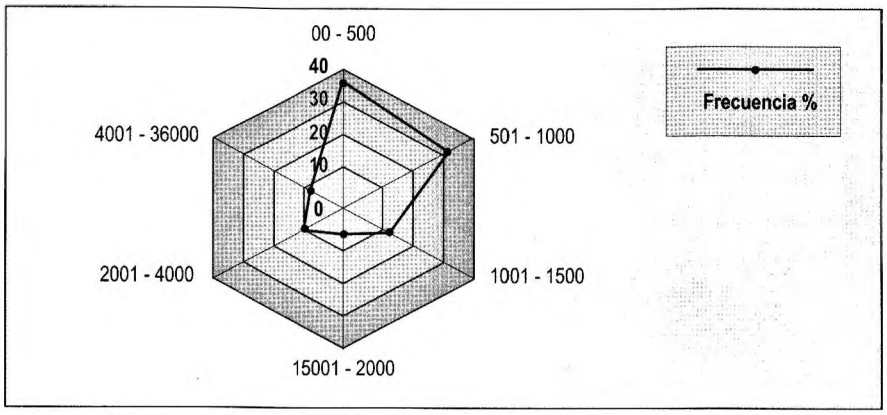
Esta grafica ratifica los comentarios efectuados en el histograma de distribución de frecuencias relativas: elevada concentración de la población ocupada con ingresos menores a Bs. 1.001 y reducidos porcentajes de participación de ocupados con ingresos elevados. Si se considera que los ingresos de los ocupados se distribuyen al interior de la familia y dada la situación que nos muestra la gráfica se puede explicar en parte la pobreza que se verifica en la ciudad de La Paz.

4.2.5. Gráfica radial

Este tipo de gráficos se utilizan generalmente para comparar dos distribuciones de frecuencias relativas cuando se utiliza ya sea dos poblaciones o muestras o se quiere representar los cambios de una distribución de frecuencias en dos a más momentos en el tiempo. En esta gráfica los ejes son polígonos concéntricos, donde cada polígono representa las frecuencias relativas, y los vértices externos muestran las categorías o intervalos de clase.

La distribución de frecuencias relativas de la población ocupada por intervalo de ingreso se representa en una gráfica radial de la siguiente forma:

GRÁFICA RADIAL
(Participación (%) de la población ocupada en cada grupo de ingreso)



Todas las representaciones gráficas contenidas en este texto y otras no incluidas, pueden elaborarse utilizando paquetes computacionales que contienen una amplia gama de posibilidades para elaborar gráficos. Dado el amplio uso de hojas de trabajo como Excel o Lotus, éstas incorporan variadas opciones para graficar a partir de ordenamientos de datos. Los paquetes estadísticos como el SPSS o STATA también tienen una amplia biblioteca de gráficas.

4.3. Análisis descriptivo de datos

Tanto la construcción de tablas de frecuencias, de contingencia y supertablas, así como la presentación gráfica de los datos permite una primera visión general de algunos aspectos destacables de la información; sin embargo, es necesario pasar a otro nivel que permita mejorar la descripción del conjunto de datos cuantitativos que son objeto de una investigación.

La descripción de los datos se puede realizar en consideración a su tendencia central, dispersión y forma, para lo cual se utilizan distintas medidas de carácter descriptivo. Si estas medidas se calculan a partir de una muestra de datos se denominan "estadísticos", y si se calculan a partir de una población se denominan "parámetros".

4.3.1. Medidas de tendencia central

La mayor parte de los conjuntos de datos muestran una tendencia a agruparse alrededor de un punto central y, por lo general, es posible elegir

algún valor promedio que describa todo un conjunto de datos. Un valor típico descriptivo como ése es una medida de tendencia central o posición. Las medidas más utilizadas son: la media aritmética, mediana y la moda.

La media aritmética:

Es el promedio o medida de tendencia central que se calcula sumando todas las observaciones de un conjunto de datos, dividiendo después ese total entre el número total de elementos de la población o muestra. La fórmula que se utiliza para el cálculo de la media aritmética para una muestra es la siguiente:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Donde; $\sum_{i=n}^{n} X_i = X_1 + X_2 + \ldots\ldots + X_n$

El símbolo $\sum$ se lee como sumatoria de todos los valores
n = Número de observaciones en la muestra
X_i = i–ésima observación de la variable X.

Esta fórmula es aplicable cuando se cuenta con pocos datos o se dispone de una computadora. Cuando el número de datos es grande, se los agrupa en tablas de frecuencia y el promedio se denomina "media aritmética ponderada", cuya fórmula es:

$$\bar{X} = \frac{\sum_{j=1}^{g} m_j f_j}{n}$$

Donde: g = número de clases o grupos
m_j = punto medio o marca de clase de la j-ésima clase
f_j = número de observaciones en la j-ésima clase o frecuencia de la j-ésima clase

El cálculo de la media se basa en todas las observaciones del conjunto de datos; ninguna otra medida de tendencia central tiene esta característica. Sin embargo, como su cálculo se basa en todas las observaciones, resulta muy afectada por valores extremos. Cuando se presenta esta situación, la media aritmética puede representar una imagen distorsionada y no es el mejor indicador para describir un conjunto de datos.

Debe aclararse que el cálculo de la media a partir de datos no agrupados (datos en su forma original) ofrece resultados reales y su cálculo con datos agrupados (tablas de frecuencias) son resultados muy aproximados a la media real.

EJEMPLO No. 28

De un grupo de seis personas de 18, 22, 25, 28, 31, 33 años, se quiere conocer el promedio de estas edades.

La media será:

$$\bar{X} = \frac{18 + 22 + 25 + 28 + 31 + 33}{6} = 26.16 \text{ años} \approx 26 \text{ años}$$

Si los datos se encuentran en una tabla de frecuencia (como la del EJEMPLO No. 26 sobre las edades de la población ocupada), el cálculo de la media implica:

- Determinar el punto medio (marca de clase) de cada intervalo. En este cálculo se incluye el límite inferior y el límite superior considerando que éste es menor al límite inferior de la siguiente clase (10+23.999)/2= 16.999, valor que puede redondearse a 17.
- Cada marca de clase se multiplica por su frecuencia absoluta.
- La suma de esta multiplicación se divide entre la suma de las frecuencias absolutas que es el tamaño de la muestra o de la población. Este resultado es la media ponderada.

Grupo de edad	Marca de clase	Frecuencia absoltuta	m * f
10 - 23 años	17	113	1.921
24 - 37 años	31	170	5.270
38 - 51 años	45	134	6.030
52 - 65 años	59	69	4.071
66 - 80 años	73	20	1.460
Total		**506**	**18.752**

$$= \frac{\sum_{j=1}^{g} m_j f_j}{n}$$

$\bar{X} = \frac{18.752}{506} = 37$ *años,* que es la medida ponderada de la muestra de la población ocupada de la ciudad de La Paz.

Continúa en la página siguiente

Viene de la página anterior

El cálculo de la media ponderada de los ingresos de la muestra de los ocupados de la ciudad de La Paz, siguiendo el mismo procedimiento, para el año 2001 fue de Bs. 2.237 al mes. Este promedio está fuertemente influenciado por pocos casos (6) de ingresos que sobrepasan los Bs.12.000. Sin considerar estos casos la media fue de Bs. 1.383.

La mediana:

Es el valor que se encuentra en el centro de un conjunto ordenado de datos y divide en dos el conjunto. Es decir que la mitad de los datos están por debajo de la mediana y la otra mitad por encima. La mediana no se ve afectada por observaciones extremas, por ello, cuando existe alguna observación extrema, resulta apropiado utilizar la mediana.

Para el cálculo de la mediana lo primero que se debe hacer es ordenar los datos en forma ascendente o descendente. Si el número de datos es impar, se aplica la siguiente fórmula:

$$Mediana = \frac{n+1}{2}$$

Si el número de datos es par también se aplica la fórmula anterior, pero entonces la posición de la mediana estará entre dos observaciones intermedias, en tal caso se calculará el promedio correspondiente a estas dos observaciones centrales. Si en la muestra se repiten valores, en el cálculo de la mediana se ignora esta situación y se calcula en base a las referencias anteriores.

El cálculo de la mediana se afecta por la cantidad de observaciones y no por la magnitud de ningún valor extremo.

EJEMPLO No. 29

En base al EJEMPLO No. 28 referido a la edad de 6 personas, se tiene el siguiente arreglo:

Persona	1	2	3	4	5	6
Edad	18	22	25	28	31	33

$Mediana = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$ valor que se encuentra entre 25 y 28 años; aplicando el promedio (25+28)/2 = 26.5 años se encuentra la mediana, que representa que el 50% de las personas son menores a esta edad y el otro 50% son mayores. Debe observarse en este caso que los valores de la media (26 años) y la mediana son muy próximos.

La mediana de ingresos de la muestra de los ocupados de la ciudad de La Paz es de Bs. 703; en este caso, se pudo apreciar una importante diferencia con la media aritmética, diferencia que se debe a la presencia de valores extremos elevados. Por esta característica se recomienda utilizar para este caso la mediana.

La mediana para datos agrupados:

Para datos agrupados en frecuencias de clase, el cálculo de la mediana se realiza con la siguiente fórmula:

$$Mediana = L + \left(\frac{n/2 - F}{f_{med}} \right) c$$

Donde:
- L = Es el límite inferior del intervalo de clase donde se encuentra la mediana
- n = Tamaño de la muestra
- F = Es la suma de las frecuencias acumuladas hasta la clase de la mediana pero sin incluirla
- f_{med} = Es la frecuencia de la clase donde se encuentra la mediana
- c = Es el ancho del intervalo de clase donde se encuentra la mediana.

EJEMPLO No. 30

Con base en la tabla de distribución de frecuencias del ingreso de los ocupados (EJEMPLO No. 26),

Grupos de ingreso	Frecuencia absoluta	Frecuencia absoluta acumulada
00 - 500	185	185
501 - 1000	156	341
1001 - 1500	59	400
1501 - 2000	28	428
2001 - 4000	42	470
4001 - 36000	36	506
Total	**506**	

El cálculo de la mediana será:

$$Mediana = L + \left(\frac{n/2 - F}{f_{med}}\right) c = 501 + \left(\frac{\frac{561}{2} - 185}{341}\right) 500 = 599.7 \cong 600 Bs$$

En el procedimiento primero se debe calcular n/2 (506/2=253). Este valor permite ubicar en la columna de frecuencias acumuladas el grupo de ingreso donde se encuentra la mediana. Dado que 253 es mayor a 185 y menor a 341, la categoría de ingreso corresponde al intervalo comprendido entre 501-1000. Luego, siguiendo la fórmula, se calculan los valores para F, c y f_{med}.

Debido a que los intervalos de ingreso no son del mismo tamaño, la mediana estimada a partir de datos agrupados difiere en más de 10% de la mediana calculada con los datos sin agrupar. Por esta razón, cuando los intervalos de clase no tienen el mismo ancho, se recomienda calcular la mediana a partir de los datos originales. Los paquetes estadísticos permiten esta opción en un tiempo muy reducido.

El cálculo de la mediana de edades de la población ocupada en la ciudad de La Paz, con el anterior procedimiento, es de 35 años y debido a que los intervalos de clase son iguales este resultado esta más próximo a la mediana calculada con datos sin agrupar.

La moda:

Es el valor de un conjunto de datos que aparece con mayor frecuencia y se la obtiene a partir de un conjunto ordenado de datos. Para datos agrupados, la clase modal es el intervalo de clase con la frecuencia más alta y el valor modal por lo general se lo aproxima como el punto medio de esa clase modal. Esta medida no está afectada por valores extremos. Un conjunto de datos puede tener más de una moda o no tener moda cuando cada uno de los datos tiene la misma frecuencia.

EJEMPLO No. 31

Considerando la tabla de distribución de frecuencias de la población ocupada por grupos de edad, el intervalo de clase que tiene el mayor número de frecuencias es el comprendido entre 24 y 37.99 años, que constituye la clase modal, y su valor modal es 31 años.

En la tabla de frecuencias de la muestra de la población ocupada por grupo de ingresos la moda se encuentra en el intervalo comprendido entre 0 y 500 Bs.

4.3.2. Medidas de dispersión

En la mayoría del conjunto de datos no todos los valores son iguales; el grado en que varían es de suma importancia para la investigación y la estadística en particular. Si no hubiera variación o dispersión de datos, no habría necesidad de la mayoría de las medidas de tendencia central y de otras que se utilizan en la estadística descriptiva. Las medidas de dispersión se orientan a describir la variabilidad de un conjunto de datos (Freund/Simon 1994: 70). Las más importantes son el rango, la varianza y la desviación estándar.

El rango:

Es la diferencia entre el valor más alto y el más bajo de un conjunto de datos. Se calcula mediante la siguiente fórmula:

$$Rango = X_{mayor} - X_{menor}$$

El rango mide la dispersión total del conjunto de datos. Su debilidad es que no toma en consideración la forma en que se distribuyen los datos

entre los valores más pequeños y los más grandes. El cálculo del rango permite una primera aproximación a la dispersión de datos.

La varianza:

Es una medida de dispersión respecto a la media; mide qué tan cerca o tan lejos están los valores de su propia media aritmética y posibilita establecer la forma en que los valores fluctúan respecto al promedio.

Para calcular la varianza de una muestra de datos no agrupados se utiliza la siguiente fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

Donde: $\overline{X}$ = media aritmética de la muestra
n = tamaño de la muestra
X_i = i-ésimo valor de la variable X

La varianza para datos agrupados (en tablas de frecuencia) se calcula como:

$$S^2 = \frac{\sum_{j=1}^{g} (m_j - \overline{X})^2 f_i}{n - 1}$$

Donde: g = número de clases o grupos
m_j = punto medio o marca de clase de la j-ésima clase
f_i = número de observaciones en la j-ésima clase o frecuencia de la j-ésima clase

Los resultados que se obtienen con la varianza son unidades al cuadrado, por ejemplo, la edad al cuadrado o ingresos al cuadrado; a fin de expresar en unidades simples se utiliza la desviación estándar.

Desviación estándar:

Es una medida de dispersión respecto a la media aritmética que se obtiene aplicando la raíz cuadrada a la varianza.

Para datos no agrupados se calcula de la siguiente manera:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}}$$

Para datos agrupados se emplea la siguiente fórmula:

$$S = \sqrt{\frac{\sum_{j=1}^{g} (m_j - \overline{X})^2 f_j}{n - 1}}$$

Los resultados que se obtienen para la varianza y la desviación típica permiten detectar que cuanto mayores son sus valores los datos están más dispersos respecto a su media. Si estos valores son pequeños, los datos estarán más concentrados alrededor de la media; si la varianza y la desviación típica son iguales a cero, significa que todas las observaciones son iguales.

Esta característica ayuda a verificar la homogeneidad o heterogeneidad de los grupos. Casi siempre, cuando se analiza aspectos relacionados con el ingreso, se observa elevada dispersión, producto de las diferencias de ingreso entre un población.

En el marco del teorema de Chebyshev, cuando se conoce la desviación estándar se puede establecer que 75% de la población se encuentra comprendida entre más dos y menos dos desviaciones estándar respecto a la media ($2\sigma + \mu - 2\sigma$), y 88.9% entre más tres y menos tres desviaciones estándar respecto a la media.

EJEMPLO No. 32

Con base en la tabla de frecuencias de los grupos de edad de la población ocupada, se puede obtener la varianza y desviación estándar para datos agrupados. Para este cálculo se busca la marca de clase, luego las diferencias entre la marca de clase y la media; estas diferencias se elevan al cuadrado y este resultado se multiplica por las frecuencias. Esta secuencia se muestra en el siguiente cuadro:

De acuerdo al EJEMPLO No. 28, la media de edad $(\overline{X})$ es de 37 años.

Grupo de edad	Marca de clase	$m_j - \overline{X}$	$(m_j - \overline{X})^2$	Frecuencia f_i	$(m_j - \overline{X})^2 f_i$
10 - 23 años	17	17-37= -20	400	113	45.200
24 - 37 años	31	31-37= - 6	36	170	6.120
38 - 51 años	45	45-37= 8	64	134	8.576
52 - 65 años	59	59-37= 22	484	69	33.396
66 - 80 años	73	73-37= 36	1296	20	25.920
Total				**506**	**119.212**

Continúa en la página siguiente

Viene de la página anterior

$\sum_{j=1}^{5}(m_j - \overline{X})^2 f_i = 11.9212$, luego la varianza será $S^2 = \frac{11.9212}{506-1} = 236$

y la desviación estándar $S = \sqrt{236} = 15.36$ *años*

En el ejemplo del ingreso de los ocupados, la desviación estándar es de Bs. 2.668.

Coeficiente de variación:

Es una medida relativa de dispersión que mide el porcentaje de variación respecto a la media. Es útil cuando se compara la variabilidad de dos o más conjuntos expresados en diferentes unidades de medición. La fórmula de cálculo es:

$$CV = \left(\frac{S}{\overline{X}}\right) 100\%$$

Donde: S = Es la desviación estándar de un conjunto de datos
$\overline{X}$ = Es la media de un conjunto de datos

EJEMPLO No. 33

El coeficiente de variación para el caso de las edades de la población ocupada será

$$CV = \left(\frac{S}{\overline{X}}\right) 100\% = \left(\frac{15.36}{37}\right) = 41.51\%$$

El coeficiente de variación para los ingresos de los ocupados será

$$CV = \left(\frac{S}{\overline{X}}\right) 100\% = \left(\frac{2.668}{1.350}\right) = 100 = 197.6\%$$

(la media y la desviación estándar provienen de datos no agrupados)

Comparando estos resultados se aprecia que la dispersión de la información respecto a la media de las edades es mucho menor a la que se observa con el ingreso; en este caso, esta medida permite inferir la existencia de fuertes diferencias del ingreso respecto a la media.

4.3.3. La forma

Es la manera en que se distribuyen los datos. Una distribución de datos puede ser simétrica cuando tiene la forma de una campana cuyo centro

divide en dos partes iguales los datos que se encuentran a su derecha e izquierda. Existe la distribución normal y la asimétrica o sesgada cuando no tiene la característica anterior.

Para describir la forma se compara la media y la mediana. Si estas dos medidas son iguales, se considera que los datos son simétricos (o con sesgo cero). Si la media es superior a la mediana, los datos tienen un sesgo positivo a la derecha. Si la media es menor a la mediana, los datos tienen un sesgo negativo o hacia la izquierda, luego:

Media >mediana =>sesgo positivo o hacia la derecha
Media <mediana =>sesgo negativo o hacia la izquierda
Media = mediana =>simetría o sesgo cero

Una medida de asimetría es el coeficiente de Pearson, que se calcula con la siguiente fórmula:

$$SK = \frac{3\ (media - mediana)}{\text{desviación estándar}}$$

Si el valor es negativo, existe sesgo hacia la izquierda; si es positivo, el sesgo es hacia la derecha; y si es cero, la distribución es simétrica.

El sesgo positivo se presenta cuando la media se ve afectada por valores muy grandes, por ejemplo, cuando se presentan ingresos muy elevados. El sesgo negativo ocurre cuando la media se reduce por valores muy pequeños, es el caso de la edad cuando predomina una población de niños y jóvenes. Los datos son simétricos cuando no hay valores extremos en dirección alguna, de manera qué los valores bajos y altos se compensan entre sí.

4.3.4. Resumen

Tanto las medidas de tendencia central (la media, mediana y moda) como las de dispersión (rango, varianza, desviación estándar) y las de simetría permiten resumir la naturaleza general de un conjunto de datos. Con sólo pocas medidas se puede tener una idea razonablemente clara de una muestra o población.

De las medidas de tendencia central se obtiene una visión sobre dónde se localiza el conjunto de datos y si es simétrico o sesgado. De las medidas de dispersión se obtiene una idea de la variación de los datos respecto a la media, que también puede interpretase como la homogeneidad o

heterogeneidad que puede presentar un determinado grupo respecto a algunas categorías que se desean investigar.

EJEMPLO No. 34

A partir de las medidas descriptivas, se puede tener una visión sobre la posible relación entre el nivel de ingresos de los ocupados de la ciudad de La Paz, con la edad (proxy de la experiencia) y nivel de escolaridad.

Grupo de ingreso	Frecuencia	Promedio de edad	Coeficiente de variación edad %	Años de Escolaridad promedio	Coeficiente variación Escolaridad %
00 - 500	185	33,6	48,9	8,4	53,4
501 - 1000	156	36,5	38,3	8,9	53,9
1001 - 1500	59	39,1	33,9	10,8	40,8
1501 - 2000	28	40,6	31,8	11,5	30,7
2001 - 4000	42	46,5	30,0	14,1	27,8
4001 - 36000	36	42,9	22,7	16,3	8,6
Total	**506**				

En el cuadro se observa que conforme aumenta el ingreso, el promedio de edad de las personas ocupadas es cada vez mayor, sin embargo, se observa la excepción en el grupo de los más altos ingresos. Esta situación invita a revisar las características de este grupo, comparar con una muestra similar y analizar caso por caso en la muestra a fin de validar o no este resultado.

Respecto a los coeficientes de variación para la edad, se aprecia que en el grupo de más bajos ingresos la dispersión es la más elevada, lo cual permite detectar que su composición etárea es heterogénea. En el grupo de los más altos ingresos se observa una menor dispersión respecto a su media de edad.

También se destaca una clara relación entre el nivel de ingreso y la escolaridad, mientras que en el grupo de ingresos bajos ésta es en promedio de 8.4 años, en el de más altos es de 16.3 años. Los coeficientes de variación de la escolaridad disminuyen a medida que aumenta el ingreso, que puede explicarse por el elevado número de ocupados por cuenta propia con distintos niveles de escolaridad que se encuentran en los tres primeros grupos, mientras que en los grupos de ingreso más elevado los aspectos de la formalidad en el trabajo exigen requisitos de escolaridad.

4.3.5. Cuantiles

Además de las medidas de tendencia central, dispersión y forma, existen otras medidas que se utilizan para resumir o descubrir propiedades de grandes conjuntos de datos cuantitativos. Los cuantiles dividen la población en proporciones iguales; así, los deciles dividen la población en diez partes, los cuartiles en cuatro, los quintiles en cinco. En este último caso, cada parte representa 20% de la población. El número de cuantiles dependerá del detalle que se quiera conocer de la población o muestra.

La construcción de los cuantiles implica los siguientes pasos:

a) Ordenar los datos de manera ascendente o descendente en función de la variable de interés, por ejemplo, para el caso de análisis de los ingresos se ordena a la población de los ingresos más bajos a los más altos.

b) Definir el número de cuantiles que se utilizarán para el análisis.

c) Crear los cuantiles de la población dividiendo el arreglo ordenado en las partes iguales que se han definido. Por ejemplo, si son quintiles de ingreso, se divide la población ordenada en cinco partes iguales, tal que cada una represente 20% de la población.

d) Luego, para cada cuantil se pueden aplicar medidas de tendencia central, de dispersión o forma.

EJEMPLO No. 35

Se ha ordenado la muestra de los ocupados de la ciudad de La Paz por ingresos de manera ascendente y se ha definido agruparla por quintiles. Realizada esta agrupación, se han calculado algunas medidas descriptivas para el ingreso, la edad y la escolaridad.

Quintiles de ingreso	Ingreso promedio mensual Bs	Promedio de edad	Años de escolaridad promedio
1o	55	30,7	8,3
2o	394	36,9	8,1
3o	703	36,6	8,9
4o	1.116	38,1	10,3
5o	4.492	43,9	14,4

Esta presentación es muy común para análisis de ingresos y de pobreza y constituye una alternativa al cuadro de frecuencias por intervalos de ingreso que se ha mostrado en los ejemplos anteriores.

En el cuadro se destacan tendencias entre los ingresos, la edad y la escolaridad. Así, a medida que nos ubicamos en quintiles más elevados de ingreso, tanto la edad como la escolaridad van en aumento. Además, se observa que la diferencia entre el promedio de edad del quintil más bajo y el más alto es de 13 años y la diferencia de escolaridad entre estos quintiles es de 6.1 años.

Estos resultados exploratorios permiten adelantar que la hipótesis formulada sobre "Los mayores niveles de ingresos de los ocupados se explican por su nivel de escolaridad y por su experiencia" se verifica de manera preliminar.

Los siguientes pasos para ratificar o no estos resultados, obtenidos a partir de una muestra, requieren el proceso de verificar su validez externa (inferencia estadística) y aplicar las denominadas pruebas de hipótesis.

4.3.6. Inferencia estadística

El alcance del presente texto no aborda los temas de la inferencia estadística, que implica un conocimiento básico de la teoría de probabilidades, las distribuciones aleatorias y las pruebas de hipótesis, entre otros aspectos.

Sin embargo, debe tenerse presente la importancia de la inferencia estadística, que es el proceso de hacer uso de los resultados de una muestra para obtener conclusiones sobre las características de una población. Este proceso ayuda a determinar la representatividad de una muestra respecto a una población dada y los márgenes de error que se puedan tener al estimar los parámetros poblacionales.