

UNIDAD 2: ORGANIZACIÓN Y DESCRIPCIÓN INICIAL DE LOS DATOS

1. Los Datos y la Información

Una vez obtenidos los datos primarios, recogidos mediante alguna de las estrategias de observación transversal descritas en el capítulo anterior, el investigador debe encontrar el mejor camino para convertirlos en información sobre los individuos observados; información que deberá acercar respuestas a las preguntas que dieron inicio a la investigación. En consecuencia, en la producción de esa información son los objetivos de la investigación los que definirán el curso a seguir en el tratamiento y análisis de los datos.

Cualesquiera sean los objetivos a alcanzar con el trabajo estadístico, el tratamiento inicial de los datos registrados debe comenzar por organizarlos en forma tal que se facilite su tratamiento. La manera de organización que se utiliza es la conocida como "Matriz de datos" que ordena los datos en una planilla rectangular, posibilitando su tratamiento en los programas informáticos.

2. La Primera Organización de los Datos: la matriz de datos



En el sentido práctico, es una forma de organizar los registros originales (de los cuestionarios, entrevistas, archivos, etc.), por la cual se ponen en relación los individuos con sus datos y permite visualizar estas relaciones. Consiste en un arreglo matricial de filas y columnas (elaborado manualmente o por medios electrónicos) como el siguiente:

Matriz de datos de "n" individuos y "p" variables

Individuo	Variable X	Variable Y	Variable J	Variable Z
1	x_1	y_1	j_1	z_1
2	x_2	y_2	j_2	z_2
.
.
.
I	x_i	y_i		j_i		z_i
.
.
.
n	x_n	y_n	j_n	z_n

Variable nº 2

Variable nº "p"

Fila que describe al individuo "i"

Cada fila de la matriz representa a un individuo de la muestra o población en estudio y cada columna identifica a una de las variables observadas. En las celdas se ubican los valores correspondientes a los individuos en cada una de estas variables (numéricas o categóricas).

Así entonces, la *i*-ésima fila de la matriz presentará al individuo genérico "i" de la muestra (o población) y sus datos en las "p" variables en estudio. A su vez, la *j*-ésima columna contendrá los valores de la variable "j", registrados a través de los "n" individuos observados.

- Notación básica

Emplearemos una notación sencilla para simbolizar a las variables y sus datos. Esto es, las letras mayúsculas **X** o **Y** o **Z** o **T** o **J** o **V** se utilizarán para designar a una **variable** en estudio (el concepto que enuncia la característica observada en los individuos). Por ejemplo:

- **X**: "Edad del Usuario de Internet" (expresada en "años cumplidos"),

- **Y:** "Intensidad de Uso del Servicio de Internet" (expresada en "horas diarias de conexión"),
- **Z:** "Sexo" (varón-mujer).

La letras minúsculas x , y , z , t , j , v , simbolizarán los valores de las variables observadas, y el subíndice que las acompaña (1, 2, 3, ..., "i", "n"), representa a los individuos con los que se corresponden cada uno de ellos. Así, continuando con el ejemplo anterior, tenemos que:

- **x_1 :** denotaría la edad observada en el usuario de Internet, registrado como "individuo 1" de la matriz,
- **y_i :** simbolizaría la intensidad de uso del servicio de Internet, registrada en el "individuo genérico i" de la muestra o población,
- **z_n :** representaría el sexo, observado en el "n-ésimo individuo genérico" de la muestra o población.

De ello resulta que:

- la expresión ($x_1, x_2, x_3, \dots, x_i, \dots, x_n$) denotará al **conjunto de los "n" valores** que la variable simbolizada con "X", registra a lo largo de los n individuos observados;
- los subíndices **no guardan relación con la magnitud** o valor de los datos que representan, simplemente **indican el orden en que fueron incorporados** a la matriz cada uno de los individuos;
- dos o más datos simbólicos cualesquiera (t_3 y t_n , por ejemplo) pueden registrar **valores diferentes** de la variable, **o bien a un mismo valor** de "T" que, por corresponder a distintos individuos, se representan con símbolos diferentes;
- en el caso de datos categóricos " u_i " **representa ahora a una de las categorías** de respuesta **o "valor"** de la **variable cualitativa** simbolizada con "U", categoría que fue observada en el "i-ésimo" individuo de la muestra o población.

- Un ejemplo de la matriz de datos



Los datos se originan en un relevamiento dirigido a los alumnos de diferentes carreras universitarias de grado de la Facultad de Humanidades y Ciencias Sociales (Licenciaturas en Trabajo Social, Antropología Social y Turismo; Profesorado en Ciencias Económicas y Técnico en Investigación Socioeconómica), que iniciaron en forma regular el curso del primer nivel de Estadística (Estadística I – Primer Cuatrimestre del 2001).

El propósito de este estudio era delinear un perfil socioeconómico y conocer algunos hábitos vinculados al estudio de los alumnos que cursan esta asignatura en la FHyCS. La observación se realizó como actividad inicial de la primera clase y abarcó a todos los alumnos inscriptos en la nómina (enumeración completa). El instrumento de recolección consistió en un cuestionario semi-estructurado de dieciséis preguntas, cuya aplicación fue auto-administrada por los alumnos.

En la matriz del ejemplo se ordenan los datos de sólo diez de esas variables, a saber:

- (EDAD) *Edad del alumno en años cumplidos.*
- (SEXO) *Sexo: 1: masculino, 2: femenino.*
- (CARRERA) *Carrera que cursa en la FHyCS, por la cual asiste al curso de Estadística:*

1: Profesorado en Cs. Económicas	2: Licenciatura en Turismo
3: Licenciatura en Trabajo Social	4: Licenciatura en Antropología Social
5: Técnico en Investigación Socioeconómica	
- (INGRESO) *año de ingreso a la Carrera de referencia.*
- (ESTPADRE) *nivel más alto de la educación formal, alcanzado por el padre del alumno:*

1: Ningún estudio	2: Primario incompleto
3: Primario completo	4: Secundario incompleto
5: Secundario completo	6: Superior/universitario incompleto
7: Superior/universitario completo	8: no sabe
- (ESTMADRE) *nivel mas alto de la educación formal, alcanzado por la madre del alumno: mismas categorías anteriores.*

- (RESIDEN) *lugar de residencia permanente del alumno -el que comparte con su grupo familiar-:*
 1: Posadas 2: Localidad del interior de Misiones
 3: Otro lugar del país o del extranjero
- (INGRET) *nivel del ingreso mensual total por todo concepto (salarios, rentas, etc.), del grupo familiar directo completo (incluyendo al alumno si corresponde), medido en pesos.*
- (HSESTUDI) *número aproximado de horas semanales que dedica al estudio de todas las asignaturas de su carrera, sin contar las horas de clases u otras actividades obligatorias.*
- (HSTV) *número de horas diarias que mira Televisión.*

Matriz del "Estudio de los Alumnos de Estadística I"

Alumno	Edad	Sexo	Carrera	Ingreso	Estpadre	Estmadre	Residen	IngreTOT	Hsestudi	Hstv
1	19	2	3	2000	3	3	2	180	-	3
2	27	2	3	2001	3	3	1	300	4	2
3	26	2	1	1999	4	4	1	700	4	2
4	28	2	2	1999	3	3	2	350	8	2
5	37	2	3	2001	3	3	1	1500	10	1
6	25	2	3	2000	3	3	1	500	3	0
7	20	2	2	2000	3	5	2	1500	6	3
8	29	2	3	-	3	2	1	560	3	1
9	25	1	1	1999	8	6	1	-	-	2
10	19	1	2	1999	8	7	1	-	4	3
11	18	1	3	2001	7	3	2	1000	14	5
12	18	2	3	2001	2	2	2	250	3	2
13	19	2	2	2000	4	7	1	-	3	1
14	19	1	2	2000	5	5	3	-	1	2
15	19	2	3	-	2	2	1	200	3	1
16	29	2	3	2001	3	3	1	300	8	2
17	19	2	2	2000	3	7	2	-	3	1
18	22	1	2	1999	5	4	1	-	6	2
19	19	2	2	2000	5	7	1	-	3	2
20	20	2	2	2000	8	7	1	-	10	2
21	22	2	2	2000	3	3	1	-	7	2
22	20	1	2	-	7	6	1	2000	4	1
23	22	2	3	1997	4	3	2	450	-	1
24	19	1	2	2000	7	7	2	1600	10	2
25	21	1	2	2000	4	5	2	1000	8	0
.
.
.
.
.
139	30	2	3	2001	3	3	1	400	7	3

- El ejemplo en símbolos

Estas variables y sus datos se expresarían simbólicamente del siguiente modo:

Si representáramos con **T** a la variable *estudios de la madre*, t_{13} simbolizaría el nivel de estudios alcanzado por la madre del alumno 13 $\rightarrow t_{13} = 7$ (universitario completo).

Simbolizando con **X** a la variable *ingreso total mensual del alumno y su grupo familiar*, x_{139} representará el ingreso total mensual del grupo familiar declarado por el alumno 139 $\rightarrow x_{139} = 400$.

Si fuera **Z** la variable *carrera que cursa el alumno*, el conjunto simbólico $(z_{1r}, z_{2r}, z_{3r}, \dots, z_{25}, \dots, z_{139})$, representará al conjunto $(3, 3, 1, \dots, 2, \dots, 3)$ de datos de la matriz correspondiente a las carreras cursadas.



Actividad N° 1

Antes de continuar con la lectura, es necesario realizar aquí la Actividad No 1 de la Guía de Actividades correspondiente a esta unidad.

3. El Análisis de la Matriz de Datos



Aun cuando la matriz de datos constituye una organización que facilita el acceso a los registros, es indudable que nuestra capacidad cognitiva no nos permite aprehender el comportamiento de los datos y obtener información a partir de ellos. Ante 139 registros como en el ejemplo, quizás con una mirada a la matriz podríamos saber el sexo mayoritario entre los estudiantes, pero difícilmente podremos concluir sobre el nivel educativo predominante entre los padres, y sería imposible poder establecer si existe una relación entre esta variable y el ingreso familiar.

Esta limitación de procesar mentalmente tal cantidad de información, nos obliga a recurrir a nuevas herramientas que permitan **resumir los datos** haciendo visibles aspectos que de otra forma permanecerían ocultos. Ahora bien, decidir sobre **cuáles son las herramientas más apropiadas depende en primer lugar de las preguntas** que intentemos responder y que, como ya dijimos, son las que guían todo el proceso de análisis.

En términos del estudio de los alumnos de Estadística y las necesidades de delinear un perfil socio-económico de los mismos, nos planteamos algunas preguntas como las siguientes:

1. ¿es heterogéneo el grupo en cuanto a la edad?
2. ¿hay predominio de mujeres?
3. ¿la composición por sexo varía según sea la carrera?
4. ¿en su mayoría se trata de alumnos ingresantes?
5. ¿sus padres han alcanzado el nivel universitario?
6. ¿se trata de estudiantes provenientes de hogares de bajos ingresos?
7. ¿está relacionado el ingreso de los hogares con el lugar de Residencia?
8. ¿el perfil determinado por el sexo del estudiante y su carrera, se relaciona con las horas dedicadas al estudio?

En este sintético listado de preguntas podemos distinguir aquellas que involucran a una sola variable (preguntas 1,2,4,5,6), a dos variables (preguntas 3 y 7) y a tres o más variables (pregunta 8). Para la búsqueda de respuestas a esas preguntas será necesario utilizar herramientas estadísticas diferentes **según sea el número de variables consideradas**.



- Cuando el análisis de los individuos se realiza a partir de una única variable sin tomar en cuenta el resto de la matriz, hablamos de un **análisis univariado**.
- Si el tratamiento de los datos involucra dos variables simultáneamente se trata de un **análisis bivariado**.
- Cuando trabajamos con tres o más variables simultáneamente recurrimos al **análisis multivariado**.

Otro aspecto a tener en cuenta al considerar la herramienta apropiada para el análisis¹ es **el tipo de variable** con el que se está trabajando: cuantitativas, o cualitativas (ordinales o nominales).

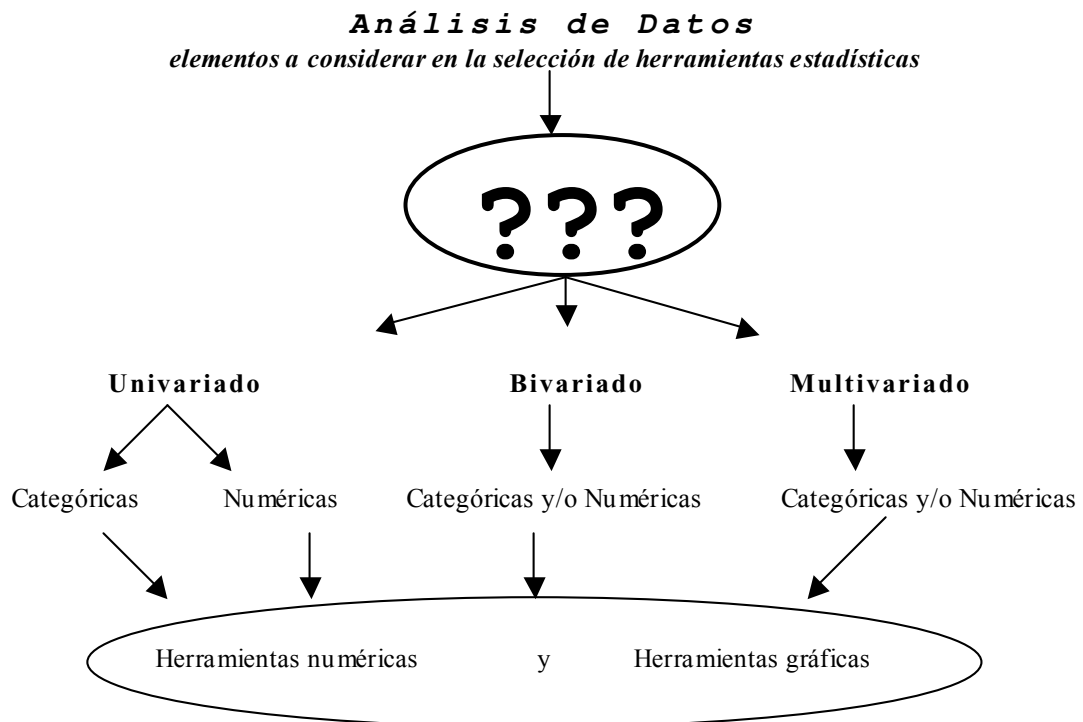
Además, las herramientas estadísticas para el análisis de datos se pueden clasificar en dos grandes familias: **numéricas y gráficas**, ambas concurrentes para hacer visible el comportamiento de los datos y complementarias en la intención de producir información.



IMPORTANTE

Priorizar las herramientas numéricas o las gráficas en el trabajo de exploración, es una decisión del investigador.

¹ Las distintas herramientas de tratamiento y análisis de datos se irán presentando según el tipo de variables involucradas.



Las herramientas que se presentarán en este curso corresponden fundamentalmente al análisis univariado y se tratan algunas de las más utilizadas del análisis bivariado.

4. Las Distribuciones de Frecuencias en el Análisis Univariado

Independientemente de la necesidad de responder aquellas preguntas que suponen el tratamiento de una única variable, cualquier análisis bi o multivariado requiere de la exploración de cada una de las variables de la matriz de datos. Las **distribuciones de frecuencias** constituyen un **primer resumen de los datos**, que nos permitirán formarnos una primera idea de cada una de las características consideradas en la investigación, construir nuevas clasificaciones, evaluar la posibilidad de aplicar otras herramientas de análisis², reformularnos algunas de las preguntas iniciales, plantear otras, etc.



La construcción de una distribución de frecuencias es un procedimiento sencillo e intuitivo que consiste en contar el número de veces que se repite cada valor de la variable en estudio (sea esta cualitativa o numérica), en el conjunto de todas las observaciones. Por ejemplo, si consideramos la variable sexo de los estudiantes de Estadística, contamos el número de veces que se presenta el valor "varón" y el valor "mujer" en el conjunto de los 139 individuos. Así, resulta que 30 es el número de veces que se repite la categoría varón y 109 la categoría mujer. Este número de repeticiones que corresponde a cada valor de la variable recibe el nombre de frecuencia absoluta.

Frecuencia absoluta:

Es el número de veces que se repite un mismo valor de la variable (una misma categoría si se trata de una variable categórica, un mismo número si la variable es numérica) en el conjunto de los "n" individuos observados.

Se simboliza con f_i (i representa en este caso el orden en que se presentan los valores de la variable).

² En unidades posteriores se presentarán otras herramientas para resumen de los datos las cuales exigen condiciones de la distribución que habrá que evaluar en esta etapa.

Distribución de frecuencias:

Consiste en un arreglo en el cual se presentan los valores de la variable y las frecuencias absolutas computadas para cada uno de ellos.

Una condición que debe cumplir la distribución de frecuencias absolutas es que la suma de todas ellas es igual al total (n) de individuos observados.

$$f_1 + f_2 + \dots + f_k = \sum_{i=1}^k f_i = n \quad (3)$$

En nuestro ejemplo, $f_1 = 30$ y $f_2 = 109$ y la suma de ambas frecuencias es igual al total de individuos observados ($n = 139$).

Si bien el concepto de distribuciones de frecuencias siempre es el mismo, la construcción cambia según se trate de variables numéricas o categóricas, y esto es así tanto para los recursos de análisis numéricos (*tablas de frecuencias*) como para los gráficos (*gráficos de distribuciones de frecuencias*). Distinguiendo estas situaciones, se presentarán las distintas herramientas estadísticas referentes a las distribuciones de frecuencias.

4.1. Variables categóricas

- el recurso numérico



Como hemos señalado, la variable sexo del ejemplo de los estudiantes de Estadística tiene dos valores posibles (varones y mujeres), y para computar las **frecuencias absolutas** que le corresponden a cada una de estas categorías realizamos un conteo del número de mujeres (109) y el número de varones (30) que aparecen entre los 139 casos registrados. Así, estaríamos distribuyendo a los 139 individuos observados en las dos categorías definidas por el sexo.

Esta clasificación se podría organizar en una tabla⁴ como la siguiente:

Distribución de estudiantes del curso de Estadística según sexo. FHyCS-Año 2001.

Nombre de la variable	SEXO	n° de estudiantes	Cantidad de varones observados
Varón	30	Frecuencias absolutas	
Mujer	109		
Valores de la variable	Total	139	Total de individuos observados

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

Es de notar que la tabla anterior resume la columna "sexo" de la matriz de datos originales, sin perder información, ganando al mismo tiempo en claridad para comprender los datos. Esta organización resumida de los datos se conoce como "Tabla de Distribuciones de Frecuencias".

³ El símbolo \sum se denomina "sumatoria" y es una forma abreviada de señalar la suma de una serie de términos; en este caso la suma de todas las frecuencias absolutas desde la primera ($i = 1$) hasta la número k .

⁴ Es importante destacar que toda tabla se puede identificar:

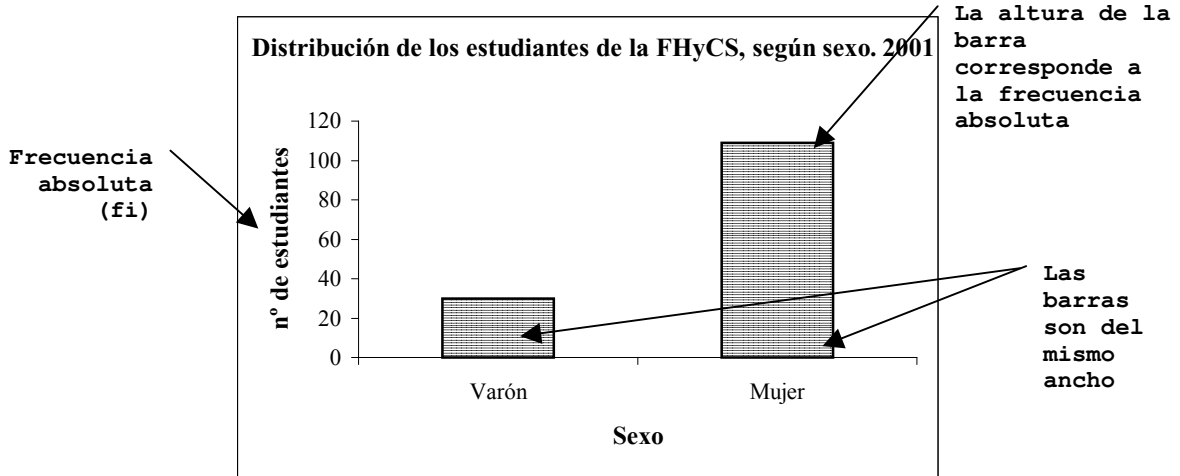
- un **título** que responda a **qué** se está describiendo, **cómo** se lo describe (en base a qué característica), **cuándo** fueron obtenidos los datos, **dónde** fueron obtenidos (lugar al que refieren);
- una **columna principal** donde se consigna el nombre de la variable y sus valores posibles y **encabezados** descriptivos del contenido de la o las columnas;
- un **cuerpo** donde están los datos;
- una **fuentes** que indica la institución, investigación, texto, etc. del cual provienen los datos;
- las **notas aclaratorias o de calce**: que sirven para clarificar alguna parte de la tabla y tienen la misma finalidad que las notas al pie en un texto. No siempre son necesarias.

- el recurso gráfico

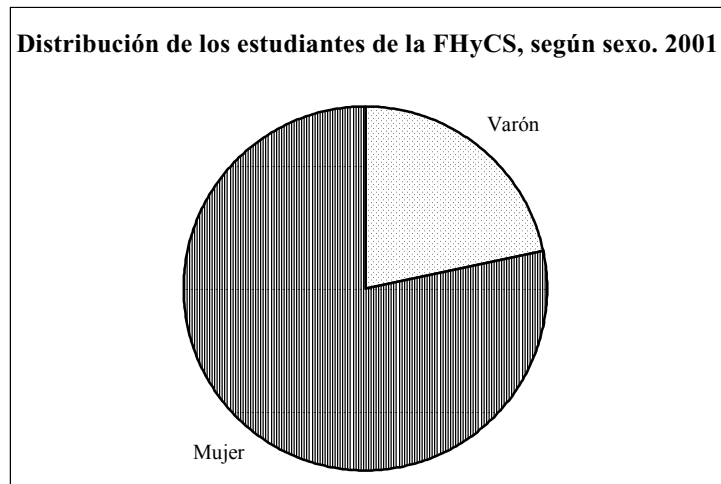


Las dos formas gráficas más utilizadas para presentar distribuciones de frecuencias de variables categóricas son: el **gráfico de barras** y el **gráfico de sectores**.

El denominado **gráfico de barra** recoge en el eje horizontal (en este caso el eje no es numérico) las categorías correspondientes a la variable (en nuestro ejemplo varón y mujer). El eje vertical (de las Y) es un eje numérico, con una escala en la que se pueden representar los valores de frecuencias observados. Las alturas de las barras de cada categoría expresan la frecuencia absoluta correspondiente.



El **gráfico de sectores o de torta**, divide una circunferencia en porciones donde cada una de ellas representa una categoría de la variable; su "tamaño" es proporcional a la frecuencia absoluta de esa categoría y el círculo representa al total de casos⁵.



A simple vista, los gráficos construidos nos permiten captar rápidamente la desigual distribución por sexo de los estudiantes del curso Estadística. Esta característica de las herramientas gráficas hacen que las mismas sean apropiadas como:

- *un recurso de análisis de los datos, y*
- *una forma efectiva de presentar y comunicar los resultados.*

⁵ La determinación del número de grados del sector correspondiente a cada categoría se obtiene razonando mediante regla de tres simple. Al total de casos (en el ejemplo 139) le corresponden 360°, consecuentemente a la categoría mujeres se le asignará un sector igual a $\frac{109}{139} \cdot 360 = 282,3^\circ$



Actividad Nº 2

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 2 de la Guía de Actividades correspondiente a esta unidad.

4.2. Variables numéricas

Cuando se construyen distribuciones de frecuencias para variables cuantitativas, **los recursos numéricos y gráficos difieren según las mismas presenten pocos o muchos valores diferentes.** Esta distinción entre las variables numéricas es al único efecto de poder destacar las particularidades de las técnicas que se utilizan en uno y otro caso.

4.2.1. Variables numéricas con pocos valores diferentes

- el recurso numérico



En el caso de una variable numérica, el criterio para resumir los datos en una tabla de frecuencias es esencialmente el mismo: a cada valor diferente que toma la variable, se le asigna el número de individuos que presentan ese valor (frecuencia absoluta).

Arreglo de Frecuencias:

Tabla en la que se presentan ordenados por magnitud (creciente o decreciente) los valores individuales observados de la variable en estudio y sus correspondientes frecuencias.

• Restricciones:

- * sólo tiene sentido en el caso de variables discretas, y
- * cuando la variable presenta pocos valores diferentes.

• Comentario: al igual que para variables categóricas se logra un resumen de los datos originales sin perder información.

La doble restricción para construir un **arreglo de frecuencias**, se cumple para pocas variables, por ejemplo "hº de hijos", "cantidad de televisores en el hogar", "hº de tarjetas de crédito disponibles en el hogar", etc.



En nuestro ejemplo, la variable "cantidad de horas diarias que mira TV" asume pocos valores diferentes y el tiempo frente al televisor está medido en horas enteras, de manera que es posible construir un arreglo de frecuencias.

Distribución de los alumnos según el tiempo que miran TV

Hs. de TV	nº de estudiantes
0	25
1	26
2	49
3	18
4	13
5	5
6	2
7	0
8	1
Total	139

Los diferentes valores de la variable

18 alumnos miran TV 3hs. diarias

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"



A partir de la lectura de la tabla, se puede señalar que mayoritariamente los alumnos miran TV 2 horas o menos por día, y son pocos los que le dedican 5 horas o más.



IMPORTANTE

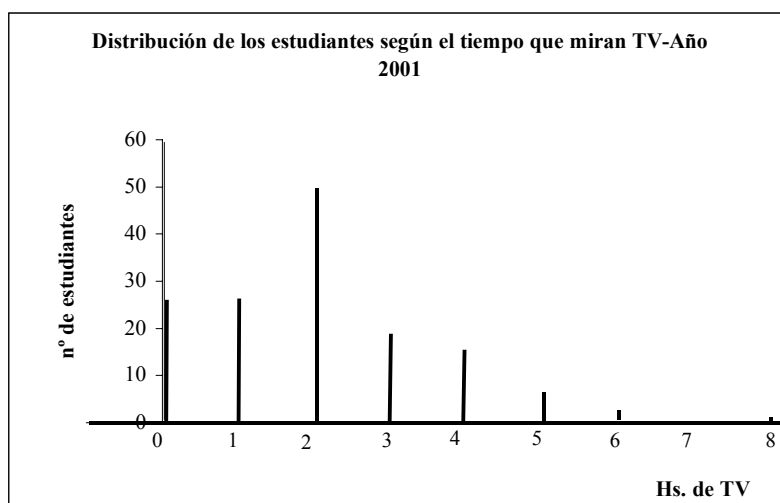
Siempre que intentamos **dar cuenta de la variabilidad** de los datos, la descripción de la distribución de frecuencias **no se agota con señalar cuál es el o los valores más frecuentes**.

Se logra comunicar esta diversidad señalando tanto los valores que más se repiten, como las singularidades, los máximos y mínimos, etc., de tal manera que la descripción genere una **buena "imagen" de la distribución** de los datos.

- el recurso gráfico



Para la representación de un arreglo de frecuencias, se recurre a un gráfico denominado **de bastones** que utiliza un sistema de ejes cartesianos, en cuyo eje de abscisas (eje X) se representan los valores de la variable y en las ordenadas (eje Y) las frecuencias absolutas. Para cada valor de la variable se levanta una línea (o bastón) cuya altura es la frecuencia absoluta correspondiente a ese valor. Debe destacarse que en este tipo de gráficos se traza una línea y no una barra, debido a que a cada valor de la variable le corresponde un punto en el eje de las abscisas.



Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"



El gráfico permite observar inmediatamente que, como se describiera a partir de los datos de la tabla, *los valores 0, 1 y 2 horas de mirar TV concentran el mayor número de alumnos y que es poco frecuente que los estudiantes miren más de 5 horas de TV.*



Actividad N° 3

Antes de continuar con la lectura, es necesario realizar aquí la Actividad N° 3 de la Guía de Actividades correspondiente a esta unidad



Consideremos ahora la edad de los estudiantes. Es importante señalar que -como en la tabla que se presenta a continuación- *si no se cumplen los requisitos señalados precedentemente para la construcción de un arreglo⁶*, la tabla de frecuencias **no constituye un buen resumen de la información**, que permita una mayor comprensión del comportamiento de los datos.

⁶ Recordemos que este tipo de distribución se utiliza en el caso de variables discretas con pocos valores diferentes.

Entre los alumnos se registran 25 edades diferentes, lo que resulta en una tabla extensa que dificulta aprehender la tendencia general de la edad de los estudiantes. En consecuencia, esta tabla no resulta un buen recurso para el análisis de la variable.

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

Edad (*)	nº de estudiantes
17	6
18	22
19	29
20	8
21	10
22	10
23	2
24	3
25	4
26	6
27	5
28	2
29	8
30	2
31	3
32	1
33	3
34	2
35	2
37	2
38	2
40	1
41	1
44	1
47	1
Total	136

(*) Hay tres estudiantes que no declaran la edad

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística".

La construcción de cualquier tabla debe lograr un equilibrio entre la mayor claridad y la menor pérdida de información; *en este caso, si bien no perdimos información tampoco hemos ganado en un resumen que permita visualizar rápidamente las principales características de la variable en estudio.*

4.2.2. Variables numéricas con muchos valores diferentes

- el recurso numérico



Una solución al problema de construir distribuciones de frecuencias para variables con muchos valores diferentes evitando las tablas extensas, es construirlas de tal manera que, en lugar de listar los valores individuales de la variable, se los presenta en **grupos de valores** para los cuales se computa su frecuencia. A esta forma de presentar los datos se la conoce como **distribución en intervalos de clase**.

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

Edad	nº de estudiantes
17-20	65
21-24	25
25-28	17
29-32	14
33-36	7
37-40	5
41-44	2
45-48	1
Total	136

Ocho Intervalos de clase

Hay 14 estudiantes que tienen entre 29 y 32 años

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística".



Leyendo la tabla, vemos que (en cuanto a su edad) el grupo es bastante heterogéneo, con edades que van desde los 17 a los 48 años; sin embargo, hay 90 estudiantes que no exceden los 24 años, y entre ellos el mayor número se concentra entre los 17 y 20 años de edad. Solamente 3 superan los 40 años. Una vez más, la **descripción de la edad de los estudiantes** no se puede reducir a la mención de lo hegemónico que resulta el grupo de edades entre 17 y 20 años. Por ello, se intenta expresar la diversidad de edades en este grupo.

Se puede ver que, de esta manera, **hemos ganado en claridad al lograr una mayor síntesis**. Debemos destacar a su vez que, mediante este procedimiento también **hemos perdido información**, dado que no podemos recuperar desde esta tabla los valores individuales de los datos. Por ejemplo: sabemos que hay 5 estudiantes que tienen entre 37 y 40 años, pero desconocemos cuáles son sus edades exactas; esto mismo vale para cada una de las clases restantes.

Esta pérdida de información hace evidente el cuidado que debemos poner al agrupar los datos en clases, es decir, al determinar la cantidad de intervalos que utilizaremos y la amplitud que daremos a los mismos.

IMPORTANTE



En las distribuciones en intervalos de clase:

- ✓ Hemos ganado en resumen y mayor claridad sobre el comportamiento de los datos.
- ✓ Conocemos la frecuencia absoluta de cada clase, pero perdemos o desconocemos la frecuencia que le corresponde a cada valor individual.
- ✓ La pérdida de información exige cuidados en la construcción de los intervalos.
- ✓ Construir una distribución en intervalos supone decidir el número de estos y su amplitud.



Actividad Nº 4

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 4 de la Guía de Actividades correspondiente a esta unidad.



Distribución en intervalos de clase:

Tabla en la que se presentan los datos agrupados en ciertas clases o intervalos de valores de la variable en estudio y las frecuencias computadas para cada clase o intervalo.

Conceptos básicos

- **Número de clases de la distribución (K):** cantidad de intervalos de clase en los que se redistribuyen los valores de la variable.
- **Límites de la clase:** los valores que delimitan cada intervalo de clase. Existe un límite inferior y uno superior para cada clase (L_i y L_s).
- **Amplitud de una clase (a):** es la diferencia entre el límite superior de esa clase y el límite superior de la clase anterior.
- **Punto medio de clase (PM):** o "marca de clase", es un valor "representativo" del intervalo que se obtiene como el promedio de los límites de la clase $[(L_i+L_s)/2]$.
- **Rango del conjunto de datos (R):** es un valor que expresa de manera global el campo de variación de los datos. Cuando se cuenta con los datos individuales se lo obtiene como: $x_{máx} - x_{mín}$; en el caso de distribuciones en intervalos de caso es la diferencia entre el límite superior de la última clase y el límite inferior de la primera.



En la distribución por edades de los alumnos, los datos se ordenaron en 8 clases de igual amplitud ($a = 4$); para la primer clase el límite inferior es 17 y el límite superior es 20, y su punto medio de clase es 18,5. Es importante destacar que por tratarse en este caso de una variable que asume valores enteros (se toma la edad en años cumplidos), fue posible construir intervalos **discontinuos**, esto es que el límite superior de una clase no coincide con el límite inferior de la siguiente, de manera que hay una pérdida de continuidad entre un intervalo y otro, lo que no supone un problema en el caso de variables discretas.

En el caso de variables continuas se construirán intervalos donde el límite superior de una clase coincide con el límite inferior de la siguiente (**continuos**). Por ejemplo en el caso de las edades se construirían intervalos de 17 a 21, 21 a 25, 25 a 29, etc. En estos casos, para que no existan problemas de decidir a qué intervalo asignar el valor que coincide con uno de los límites, se acepta la convención de que los intervalos comprenden las edades que van de 17 **a menos** de 21, de 21 **a menos** de 25, etc. De manera que, un individuo con 21 años se computa en el segundo de los intervalos definidos.

Si tomamos otro ejemplo como el *ingreso mensual total* del hogar de los estudiantes, se pueden construir intervalos de 0-250, de 250-500, 500-750, etc. Un estudiante que pertenece a un hogar con un ingreso total mensual de \$500 será asignado al tercer intervalo (de 500 a 750 pesos), porque el intervalo de 250 a 500 incluirá todos los ingresos desde 250 incluido, hasta \$499,99.

¿Qué criterios utilizar para construir los intervalos?

Esta pregunta no tiene una única respuesta. La construcción de la distribución por intervalos se puede guiar por distintos criterios, como el propuesto por **Sturges, la exploración previa de los valores individuales** y **los propósitos del análisis**. Sin embargo, pueden señalarse algunas recomendaciones.



Recomendaciones generales para la construcción:

- El número de clases no debería ser inferior a 4 ni superior a 15.
- Las clases deberán ser -en lo posible- de igual amplitud y con límites enteros.
- Evitar la presencia de clases abiertas (sin límite superior en la última clase o inferior en la primera).
- Evitar la presencia de clases vacías (intervalos de clase con frecuencia cero).
- Con la redistribución en clases, se buscará manifestar la tendencia de los datos a concentrarse en determinados valores.
- Los intervalos deben comprender todo el rango de variación de la variable.

□ **El modelo de Sturges**



Una primer respuesta sería la que propone Sturges quien, a partir del número de datos que se quieren ordenar, "recomienda" como el número de clases apropiada el resultado de la siguiente expresión:

$$k \cong 1 + 3,3 \cdot \log n$$

donde: **k** es el número de clases que se quiere determinar,
n es el número total de datos y
log es el logaritmo⁷ en base 10.

Obtenido el número de clases (**k**) la amplitud de las mismas (**a**), surge inmediatamente de hacer: $a \cong \frac{R}{k}$, donde R es el rango. Se expresa que la amplitud es *aproximadamente igual* (\cong) al resultado del cociente, porque este puede dar un valor no entero.



En nuestro ejemplo, para determinar el número de intervalos de clase, dado que *n* es igual a 136, la fórmula de Sturges sugiere: $k \cong 1 + 3,3 \log 136 \cong 8$ intervalos
 Entonces, dado que la edad máxima es de 47 años y la mínima de 17 años, la amplitud es $a \cong \frac{47-17}{8} = 3,75$ aproximadamente igual a 4.

Determinado el número de intervalos y la amplitud, construimos las clases tomando como límite inferior el menor valor observado (en nuestro ejemplo 17 años), o un valor que comprenda a ese mínimo. Así, el primer intervalo en este caso podría ser de 17 a 21 (si hacemos intervalos continuos) y, a partir de allí, los intervalos se sucederán de la siguiente manera: 21 a 25, 25 a 29 y así siguiendo hasta el intervalo que cubra el valor máximo.

El modelo de Sturges presenta como ventaja su simplicidad, y como limitación el hecho de que se basa únicamente en el número de datos observados sin tener en cuenta la distribución de los mismos. En consecuencia, debe ser tomado como una primera aproximación al número y amplitud de intervalos, la que será ajustada en función del comportamiento de los datos y tomando en cuenta las recomendaciones que hemos señalado precedentemente.

□ **La exploración previa de los valores individuales**



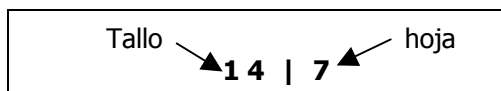
Una forma de explorar la distribución de los valores de la variable es mediante la construcción y observación del **arreglo de frecuencias**. A partir del arreglo, y tomando en cuenta las reglas generales de organización de los datos en intervalos, se puede ir construyendo un mejor resumen, mediante un proceso de prueba y error, hasta cumplir con el requisito final de expresar *de la mejor manera* posible la tendencia o comportamiento de los datos. Un procedimiento de estas características es el que hemos utilizado para la construcción de la distribución en intervalos presentada anteriormente. Sería un buen ejercicio para el lector, ensayar distintas organizaciones de los datos a partir del análisis del arreglo de las edades de los estudiantes.

Otra forma de exploración inicial es recurrir a un tipo de gráfico especial que se conoce como **diagrama de tallo-hoja** (en inglés denominado Stem & Leaf). El diagrama presenta los datos de la variable ordenados de una manera particular, en el que **se descomponen los valores** en dos partes⁸:

- *el tallo*, que toma los primeros dígitos, y
- *la hoja* que toma el dígito siguiente.



Por ejemplo, el valor 147 se puede dividir en un tallo de 14 (los dos primeros dígitos) y una hoja de 7.



⁷ Esta fórmula es muy sencilla de utilizar en el caso de contar con una calculadora que disponga de la función logarítmica.

⁸ La construcción de este Gráfico tiene muchas variantes, aquí desarrollaremos la más simple; sin embargo, para profundizar el conocimiento sobre este recurso analítico, recomendamos la lectura de Moore (1995: 19-21) y Alaminos (1993: 32-33).

□ **Los propósitos del análisis**

Los propósitos del análisis pueden guiar la construcción de intervalos de clase diferentes a los que surgen de un *modelo como el de Sturges* o del análisis de la distribución a partir del *diagrama de tallo-hoja*. Así por ejemplo, en la construcción de intervalos de clase para la variable edad, puede ser de interés del investigador reconocer la distribución según **grupos de edades que tienen sentido** en términos de que cada tramo de edad permite suponer características particulares de quienes lo integran (experiencia de vida, intereses, hábitos, trabajo, rol en el hogar, etc.). Así, podríamos imaginar intervalos de clase definidos como:

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

Edad	nº de estudiantes
17-19	57
20-29	58
30 y más	21
Total	136


Intervalos de clase de diferente amplitud →

Intervalo de clase abierta →

Hay 21 estudiantes de 30 años y más →

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

Tenemos en este caso una distribución u organización de los datos que resulta válida, aun cuando se trata de tres intervalos con distinta amplitud y uno de ellos es abierto (sin un límite superior). Lo que queremos destacar con el ejemplo, es que, al momento de construir una distribución, **por encima de cualquier criterio estadístico que se pueda tomar en cuenta, está el propósito del análisis.**

	Actividad Nº 5 <i>Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad.</i>
--	--

- el recurso gráfico

El recurso gráfico que se asocia a las distribuciones de frecuencias organizadas en intervalos de clase es el **histograma**.



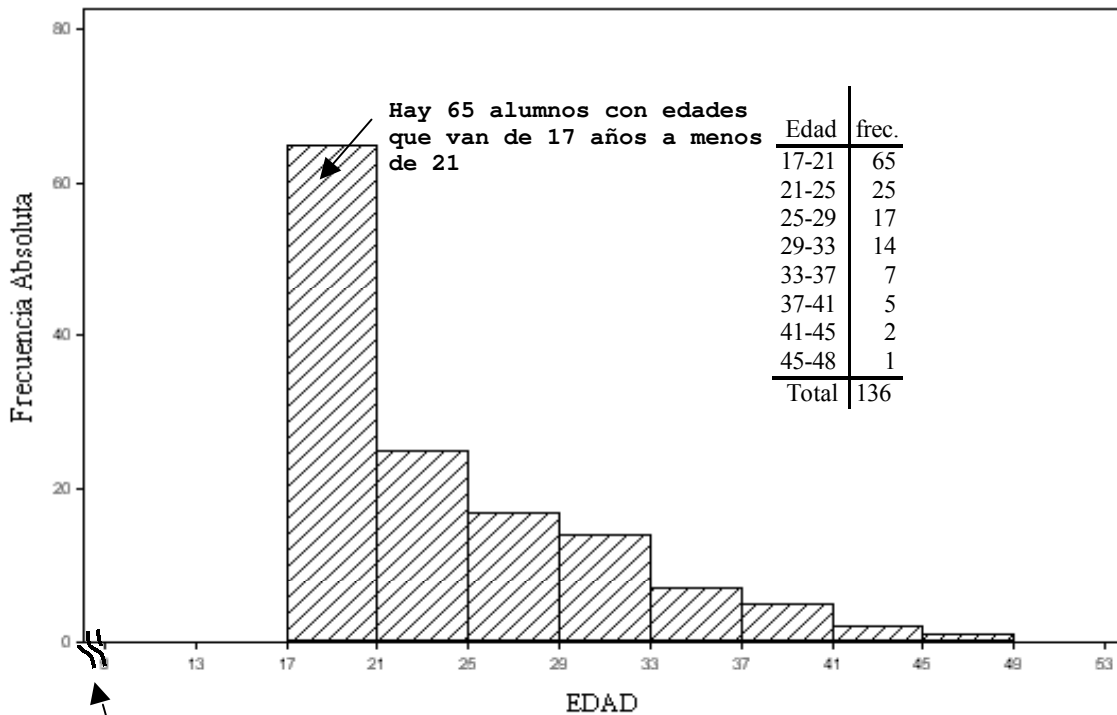
Histograma

Se trata de un **gráfico de barras en un sistema de ejes cartesianos**, en cuyo eje de las **X** se representa la **variable** en estudio, y en el eje de las **Y** las **frecuencias**. En él, se hace corresponder a cada intervalo de clase una barra cuya altura coincide con la frecuencia de esa clase.

Comentarios

1. Las barras deben cubrir todo el recorrido de la variable, lo que exige darle continuidad a los intervalos que se construyen.
2. La presencia de clases de diferente amplitud y de clases abiertas exigen soluciones particulares para graficar y es este uno de los motivos por los cuales se busca evitar este tipo de situaciones.
3. La principal utilidad de este recurso analítico es facilitar la descripción general del conjunto de datos, analizando la "forma" que toma la distribución; esto es para qué valores existen mayores concentraciones, como así también identificar aquellos muy diferentes (valores atípicos) al común de los datos del conjunto.

Histograma de Edad de los Estudiantes



Indica que se ha cortado el eje, evitando un blanco innecesario

136 casos computados – 3 casos sin datos

IMPORTANTE

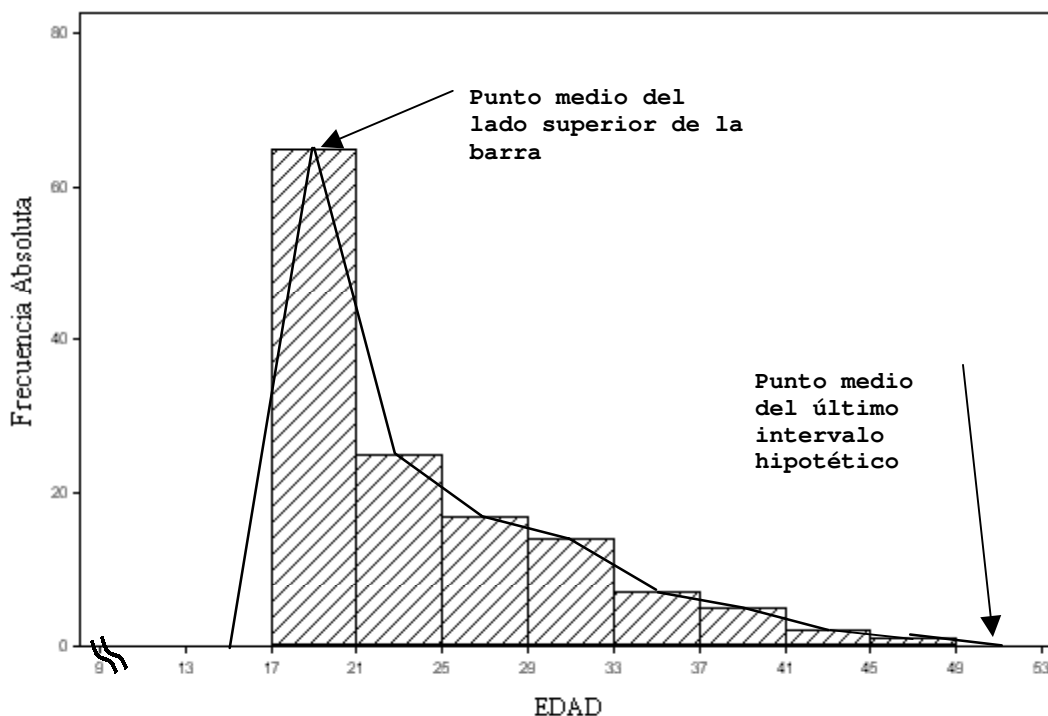
El histograma se construye con *intervalos de clase continuos* y de *igual amplitud*, que es la manera más sencilla de hacerlos, y permite **estudiar la forma de la distribución**, finalidad fundamental de este recurso. **La forma** está dada -como fuera señalado anteriormente- por los aspectos más generales (concentraciones) y singularidades (valores atípicos) que presentan los datos.



En este caso *la forma* del histograma nos indica la fuerte concentración de estudiantes entre 17 y 21 años con una sostenida disminución del número de ellos a partir de esa edad. Otra manera de expresar la *forma* de esta distribución sería señalando que en este conjunto existe una concentración de los datos en los primeros grupos de edades (es muy frecuente la presencia de estudiantes “jóvenes”) y pocos casos de estudiantes en las edades más altas.



El **polígono de frecuencias** constituye otra manera de presentar una distribución de frecuencias, que se obtiene uniendo mediante segmentos los puntos medios del lado superior de cada una de las barras de frecuencia. En los extremos, el polígono se “cierra” uniendo los extremos del primero y último rectángulo con el punto medio de un primer y último intervalo hipotético construido a este fin (en nuestro ejemplo los intervalos de 13-17 y 49-53 años de edad).

Histograma y Polígono de Frecuencias de la Edad de los Estudiantes

136 casos computados – 3 casos sin datos

El polígono se representa normalmente en forma separada al histograma ya que ambos tienen la misma finalidad¹⁰. De esta manera con el polígono obtenemos un gráfico simple, que constituye una "silueta" de la *forma de la distribución*, y en consecuencia nos permite al igual que el histograma, describir el comportamiento general del conjunto de datos.

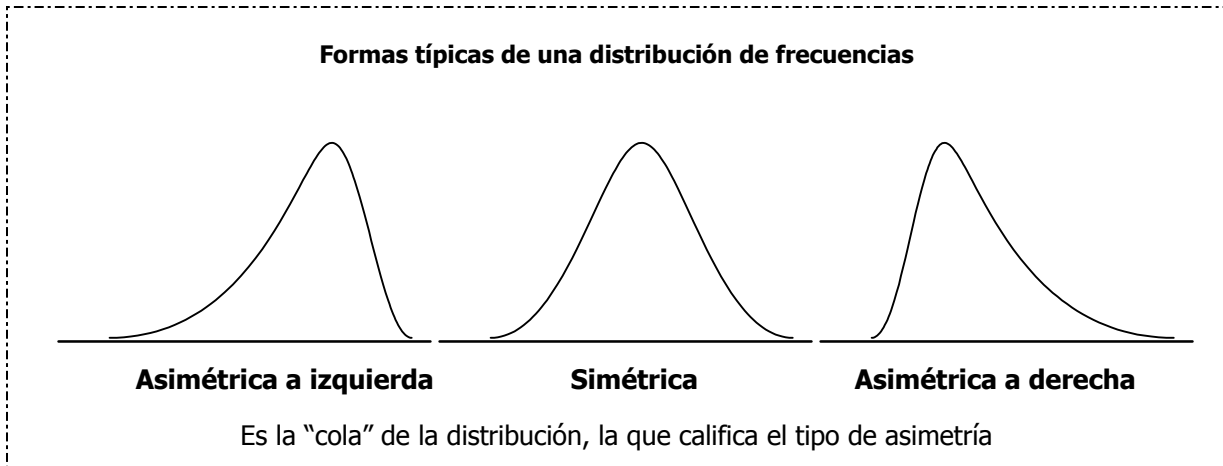
Tanto el histograma como el polígono de frecuencias son recursos fundamentales para explorar y presentar un conjunto de datos numéricos en los que tenga sentido realizar agrupamientos en intervalos de clase.

El **diagrama de tallo-hoja** que presentáramos anteriormente, también funciona como un recurso exploratorio que nos permite **captar la forma** de la distribución, sin perder los valores individuales que se agrupan en los distintos intervalos. De hecho, este es uno de los usos más frecuentes del diagrama y varios autores lo presentan como un recurso que conserva las bondades de una tabla de frecuencias y las de un histograma.

Las **distribuciones en cuanto a su forma** pueden ser de tres tipos (ver gráfico):

- **Simétricas:** cuando los datos se concentran en los valores centrales de la distribución, y las frecuencias decrecen hacia ambos extremos de manera simétrica.
- **Asimétricas a la derecha:** cuando los datos se **concentran a la izquierda** y disminuyen las frecuencias a medida que aumentan los valores de la variable.
- **Asimétricas a la izquierda:** cuando los datos se **concentran a la derecha** de la distribución y las frecuencias disminuyen gradualmente a medida que los valores de la variable decrecen.

¹⁰ Se puede demostrar además, que la superficie de todas las barras del histograma y el área comprendida bajo el polígono son equivalentes.



En el ejemplo del histograma o polígono de las edades se observa una distribución marcadamente asimétrica a la derecha.

	<p>Actividad N° 6</p> <p><i>Antes de continuar con la lectura, es necesario realizar aquí la Actividad N° 6 de la Guía de Actividades correspondiente a esta unidad.</i></p>
--	---

4.3. Transformaciones de las frecuencias absolutas



Muchas veces la necesidad de interpretar la información producida en una tabla de frecuencias absolutas y/o responder preguntas que nos formulamos en relación al comportamiento de los datos nos obligan a re-expresar o transformar la información contenida en la tabla. Por ejemplo,

- decir que "65 estudiantes tienen entre 17 y 21 años", no brinda información respecto a la importancia de este grupo en el conjunto de estudiantes observados, es más ilustrativo señalar que "el 48% de los estudiantes tienen entre 17 y 21 años".
- de la misma manera, responder a la pregunta *¿cuántos estudiantes tienen menos de 29 años?*, obligaría a recalcular la frecuencia absoluta reagrupando a los estudiantes que tienen menos de 29 años.

Con el fin de dar respuesta a este tipo de interrogantes, se re-expresan las frecuencias en otras que facilitan la lectura e interpretación: **frecuencias relativas y acumuladas**.

4.3.1. Las frecuencias relativas

Hay diversas situaciones en las que se requiere expresar la distribución de frecuencias en términos relativos al total de datos; por ejemplo:

- cuando queremos conocer la **importancia relativa de ciertos valores** o características en el conjunto de datos observados. Ejemplo: "El 40% de los árboles de Bs. As. son fresnos", para señalar la abundancia de esta variedad en la ciudad;
- cuando queremos **comparar esa importancia relativa** entre dos conjuntos de datos de diferente tamaño. Ejemplo: "El 37,6% de la población de Formosa es pobre mientras que en Misiones esa población alcanza al 24,9%", para comparar la incidencia de la pobreza en dos poblaciones de diferente tamaño;
- cuando **a partir de una muestra** queremos **sacar conclusiones** sobre la presencia de cierta característica en la población. Ejemplo: para concluir sobre el comportamiento de la población de Internet a partir de la observación de una muestra, no brinda una información pertinente decir "560 de los usuarios de Internet observados son mujeres" sino: "cuatro de cada diez usuarios de Internet son mujeres".

Frecuencia relativa (fr):



Mide la proporción de datos del conjunto que presentan un determinado valor de la variable, generalmente expresado en porcentaje.

Cálculo

Se la obtiene como el cociente entre la frecuencia absoluta de una clase (valor individual o categoría de respuesta) y el total "n" de datos.

$$fr = \frac{f_i}{n}$$

Generalmente se la expresa en porcentaje, multiplicando por 100 la expresión anterior.

$$fr(\%) = \frac{f_i}{n} \cdot 100$$

La suma de todas las frecuencias relativas porcentuales es 100.

$$\sum fr = 100$$

Estudiantes del curso de Estadística según edad- FHycS-Año 2001

Edad	nº de estudiantes	Frecuencia relativa (%)
17-18	28	20,6
19-20	37	27,2
21-22	20	14,7
23-26	15	11,0
27-30	17	12,5
31-35	11	8,1
36 y más	8	5,9
Total	136	100,0

$$\frac{28}{136} \cdot 100$$

El 11% de los estudiantes tienen entre 23 y 26 años

La suma de las frecuencias relativas siempre da 100

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"



En la tabla se puede leer, por ejemplo, que *los 15 estudiantes de entre 23 y 26 años, representan el 11% del total.*

4.3.2. Las frecuencias acumuladas



Muchas veces interesa conocer el número total (o el porcentaje) de individuos que tienen *menos que* (a lo sumo) un determinado valor de la variable o *más que* (al menos) un cierto valor. Por ejemplo: *¿cuántos estudiantes tienen hasta 22 años?* o *¿cuántos estudiantes tienen más de 26 años?*

Intentemos responder intuitivamente estos dos interrogantes. En el primer caso, deberíamos considerar a los estudiantes que tienen 17, 18, 19, 20, 21 y 22 años. El número de estudiantes con a lo sumo 22 años surgirá de sumar el total de estudiantes que tienen entre 17 y 18 años, más los que tienen entre 19 y 20, y los que tienen 21 y 22 años. Es decir que *acumulamos las frecuencias absolutas* de todos los intervalos de edades que no excedan los 22 años. En consecuencia tenemos (28+37+20 = 85) 85 estudiantes de 22 años o menos.

De manera análoga se puede razonar para encontrar la cantidad de estudiantes que tienen *más de* 26 años.

Para responder a este tipo de interrogantes resulta conveniente construir una **distribución de frecuencias acumuladas**.

- Frecuencias acumuladas "menos que" (Fa-)



Indican el número de observaciones en la distribución que son menores al límite superior de cada una de las clases (valor individual o categoría de respuesta) en que fueron organizados los datos.

Cálculo:

Para una clase genérica "i" de la distribución (o valor individual si se trata de un arreglo de frecuencias o categoría si se trata de una variable ordinal), la frecuencia acumulada menos que se obtiene sumando la frecuencia absoluta de esa clase más las frecuencias absolutas de todas las clases anteriores a ella.

$$Fa- = \sum_1^i f_i$$

- Frecuencias acumuladas "más que" (Fa+)

Indican el número de observaciones en la distribución que son mayores al límite inferior de cada una de las clases (valor individual o categoría de respuesta) en que fueron organizados los datos.

- Frecuencias acumuladas relativas (Far)

Indican la proporción o porcentaje de observaciones acumuladas respecto al total de datos.

Cálculo

Se obtiene como proporción o porcentaje de las frecuencias acumuladas absolutas ("menos que" o "más que") al total "n" de datos.

$$Far = \frac{Fa}{n} \quad \text{ó} \quad Far(\%) = \frac{Fa}{n} \cdot 100$$



IMPORTANTE

Estas frecuencias tienen sentido únicamente para datos *numéricos* o datos *categoricos en escala ordinal*.

Estudiantes del curso de Estadística según edad- FHycS-Año 2001

Edad	nº de estudiantes	Frec. relativa (%)	Frec. Acumulada Fa-	Frec. Acumulada Far- (%)	Frec. Acumulada Fa+	Frec. Acumulada Far+ (%)
17-18	28	20,6	28	20,6	136	100,0
19-20	37	27,2	65	47,8	108	79,4
21-22	20	14,7	85	62,5	71	52,2
23-26	15	11,0	100	73,5	51	37,5
27-30	17	12,5	117	86,0	36	26,5
31-35	11	8,1	128	94,1	19	14,0
36 y más	8	5,9	136	100,0	8	5,9
Total	136	100,0				

$$\frac{71}{136} \cdot 100$$

$$8+11+17$$

La acumulada relativa porcentual de la última clase es 100%


La acumulada absoluta de la última clase es "n"

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

En este cuadro se incluyen todas las formas de expresar las frecuencias y en él podemos leer en la línea grisada y a modo de ejemplo que:



- 20 estudiantes tienen entre 21 y 22 años, y constituyen el 14,7% del total del curso.
- 85 estudiantes tienen 22 años o menos y representan el 62,5% del total.
- 71 tienen 21 años o más y este grupo representa el 52,2% del total.

	Actividad Nº 7 <i>Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 7 de la Guía de Actividades correspondiente a esta unidad.</i>
---	--

Quando se trata de una **variable ordinal**, el razonamiento es análogo al desarrollado para las variables numéricas. Por ejemplo en el caso de la variable Nivel de estudios del Padre la información se podría organizar en una tabla como la siguiente:

Estudiantes de Estadística según Nivel de estudios del Padre- FHyCS-Año 2001

Nivel de Estudios del Padre	nº de estudiantes (*)	estudiantes (%)	Frecuencias Acumuladas (Fa-)	Frecuencias Acumuladas Far- (%)	Frecuencias Acumuladas (Fa+)	Frecuencias Acumuladas Far+ (%)
Ninguno	3	2,2	3	2,2	133	100,0
Prim. Incompleto	27	20,3	30	22,5	130	97,8
Prim. Completo	56	42,1	86	64,6	103	77,5
Sec. Incompleto	17	12,8	103	77,4	47	35,4
Sec. Completo	17	12,8	120	90,2	30	22,6
Terc./Univ. Incomp.	7	5,3	127	95,5	13	9,8
Terc./ Univ. Comp.	6	4,5	133	100,0	6	4,5
Total	133	100,0				


(*) Hay 6 estudiantes que no declaran el nivel de estudios de su padre.

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

En la línea grisada se lee:



- Los 17 estudiantes cuyos padres tienen secundario incompleto, representan el 12,8%.
- Son 103 los estudiantes cuyos padres no superaron el secundario incompleto (tienen un nivel de estudios de secundario incompleto o menos). Estos representan el 77,4% del total de los estudiantes.
- Los que tienen padres con secundario incompleto o más, son 47 y representan el 35,4% del total.

	Actividad Nº 8 <i>Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 8 de la Guía de Actividades correspondiente a esta unidad.</i>
---	--

4.3.3. La curva de Lorenz y el índice de Gini



La **Curva de Lorenz** es un recurso gráfico que permite **analizar el grado de concentración/desconcentración** de ciertas variables particulares. Así, para el "ingreso", la "renta", la "tenencia de la tierra", etc. tiene sentido y resulta de interés conocer la mayor o menor concentración de esos "recursos" en una cierta población en estudio. Este gráfico será útil cuando intentemos responder preguntas como:

- ✓ ¿La superficie de tierra productiva de la provincia, aparece concentrada entre pocos propietarios?
- ✓ ¿Cómo se distribuye el ingreso entre los hogares de la ciudad de Posadas?
- ✓ ¿Cuál es la distribución de los 37 millones de argentinos según el tamaño de las localidades?
- ✓ etc.

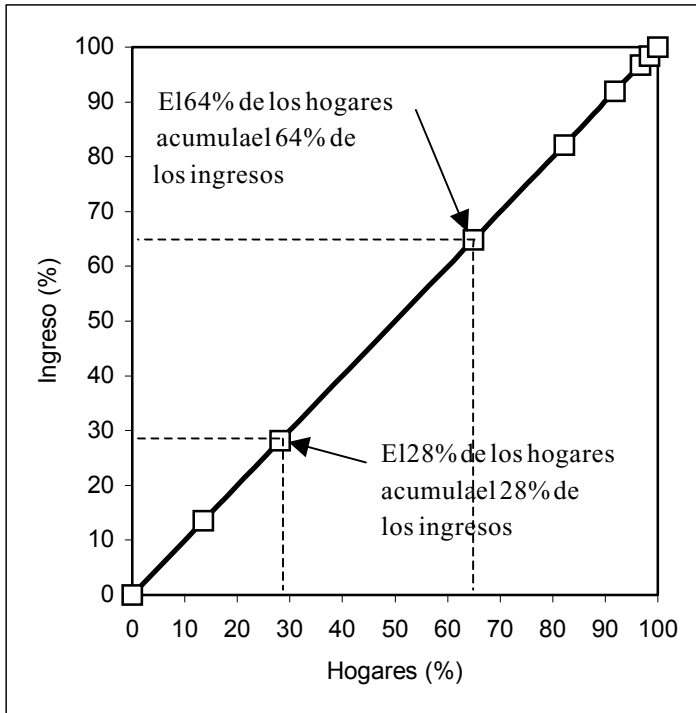
A manera de ejemplo consideremos la distribución del ingreso entre los hogares de Posadas. Analizar la distribución de estos ingresos entre los hogares, nos lleva a observar si el monto total de los ingresos registrados se reparte equitativamente (o no), entre el total de hogares; así, en una situación de equidistribución, a cada hogar le correspondería el mismo ingreso. Intuitivamente, podemos entender que, en este caso, el ingreso del 5% de los hogares representa un 5% del ingreso

total; a un 28% de los hogares le corresponderá el 28% del total de los ingresos, al 64% el 64% y así sucesivamente.



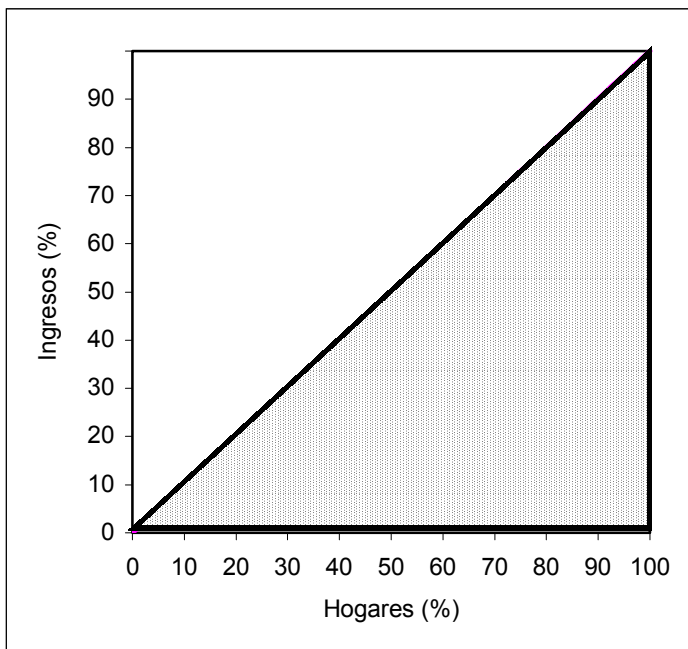
Una situación de estas características se puede representar gráficamente, utilizando un sistema de ejes cartesianos, en el que cada punto queda definido por el *porcentaje de hogares* y su correspondiente *porcentaje de ingresos*, obteniendo una gráfica como la siguiente.

Curva de Lorenz para una situación de equidistribución (o mínima concentración)



La situación de **equidistribución** queda representada entonces por la recta que divide al cuadrante en dos partes iguales (bisectriz, diagonal del cuadrado); expresando así el caso de **mínima concentración** (estrictamente nula).

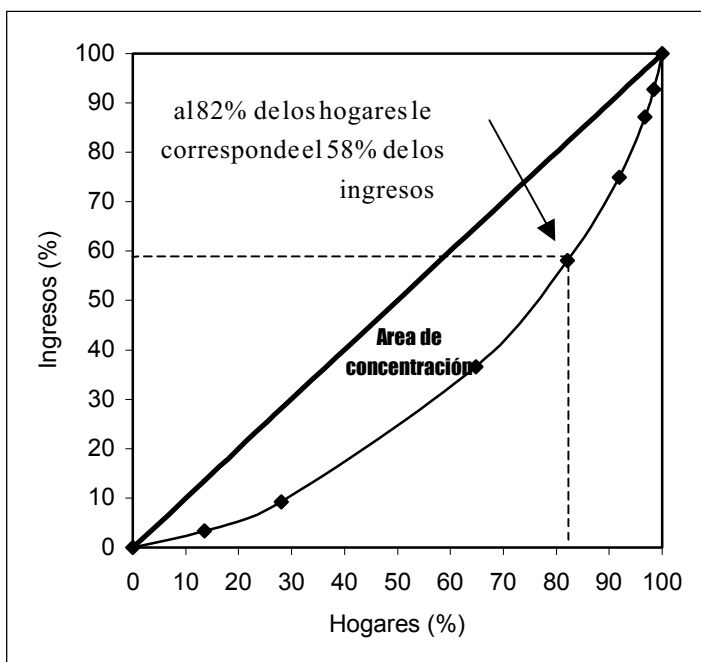
Curva de Lorenz para una situación de máxima concentración



La situación opuesta (de **máxima concentración**) estaría dada por aquel caso en que el total de los ingresos se concentra en un solo hogar. Entonces, al 10% de los hogares les corresponde el 0% de los ingresos, al 30% también el 0%, y así sucesivamente, hasta llegar al último hogar (que completa el 100%) al que le corresponde el 100% de los ingresos.

De esta manera el gráfico define un área que se corresponde con el triángulo inferior del cuadrado (área sombreada): área de **máxima concentración**. Estamos aquí nuevamente ante una situación teórica.

Curva de Lorenz para una situación de concentración intermedia



Entre estos dos extremos, de máxima y nula concentración, en la realidad encontraremos una infinidad de **situaciones intermedias**, que definirán curvas que **a medida que se alejan de la bisectriz nos hablan de situaciones cada vez menos equitativas o de mayor concentración** de la variable que se está analizando. El área definida entre la bisectriz y la curva se conoce como **área de concentración**.

En el gráfico siguiente presentamos una curva de Lorenz que representa una situación intermedia a los extremos planteados.

La construcción de la curva de Lorenz



La construcción de la curva es sencilla, debiéndose contar para ello con la distribución de frecuencia de la variable en estudio; en este caso la distribución de la variable ingreso en los 3.300 hogares de Posadas. A continuación desarrollaremos las transformaciones necesarias para disponer de los datos que se representan en la curva de Lorenz (porcentaje de ingresos que acumulan diferentes porcentajes acumulados de hogares).

Ingresos familiares mensuales- Posadas 1994

Ingresos familiares	Número de hogares (f_i)	Ingreso medio de clase (x_i)
165-249	450	207,0
249-414	486	331,5
414-829	1224	621,5
829-1243	576	1036,0
1243-1658	324	1450,5
1658-2487	162	2072,5
2487-3316	54	2901,5
3316-4146	54	3731,0
TOTAL	3330	

La Tabla anterior presenta la distribución de los ingresos monetarios mensuales percibidos por 3.330 familias de Posadas, agrupados en intervalos. Aceptando que los puntos medios representan a los datos incluidos en cada clase, el producto de cada punto medio por su correspondiente frecuencia absoluta ($f_i \times x_i$) expresa el monto o volumen total de ingresos percibido por los hogares de esa clase. Así por ejemplo: $450 \times 207,0 = \$93.150.-$ Esto significa que los 450 hogares con niveles de ingresos mensuales entre \$165 y \$249 perciben en conjunto un monto total de \$93.150.-

De igual modo los 486 hogares con ingresos entre \$249 y \$414 perciben todos juntos un monto total de ingresos de \$161.109 ($486 \times 331,5$). Es decir que utilizando los puntos medios de clase (ingreso medio de ese grupo de hogares) y las frecuencias absolutas (cantidad de hogares de la clase)

es posible obtener el ingreso total de las familias que componen esa clase, tal como se muestra en la columna (4) de la tabla siguiente.

Ingresos familiares mensuales - Posadas 1994

Ingresos familiares ⁽¹⁾	Número de hogares(f_i) ⁽²⁾	Ingreso medio de clase (x_i) ⁽³⁾	Monto total de ingresos en \$ ⁽⁴⁾
165-249	450	207,0	93150
249-414	486	331,5	161109
414-829	1224	621,5	760716
829-1243	576	1036,0	596736
1243-1658	324	1450,5	469962
1658-2487	162	2072,5	335745
2487-3316	54	2901,5	156681
3316-4146	54	3731,0	201447
TOTAL	3330		2775546

Sumando los ingresos correspondientes a cada clase, obtenemos el monto total de los ingresos percibido por el conjunto de los 3.330 hogares observados (\$2.775.546). Podemos ver además que, los 450 hogares de menores ingresos (entre \$165 y \$249) acumulan un total de \$93.150; a su vez son \$161.109 los percibidos por hogares con ingresos mensuales entre \$249 y \$414, y así sucesivamente.

El número de hogares y el monto total de los ingresos que les corresponden, pueden ser acumulados tal como se presenta en las columnas (5), (6), (7) y (8), de la Tabla siguiente.

Ingresos familiares mensuales – Posadas, 1994

Ingresos familiares ⁽¹⁾	Número de hogares (f_i) ⁽²⁾	Monto total de ingresos en \$ ⁽⁴⁾	Hogares Acum. (Fa) ⁽⁵⁾	Ing. Acum. (\$) ⁽⁶⁾	Hogares Acum.(%) ⁽⁷⁾	Ing. Acum. (%) ⁽⁸⁾
165-249	450	93150	450	93150	14	3
249-414	486	161109	936	254259	28	9
414-829	1224	760716	2160	1014975	65	37
829-1243	576	596736	2736	1611711	82	58
1243-1658	324	469962	3060	2081673	92	75
1658-2487	162	335745	3222	2417418	97	87
2487-3316	54	156681	3276	2574099	98	93
3316-4146	54	201447	3330	2775546	100	100
TOTAL	3330	2775546				

Las columnas (5) y (6) expresan en valores absolutos, el número de hogares y monto total de ingresos acumulados. Las columnas (7) y (8) presentan esos mismos valores expresados en porcentajes.

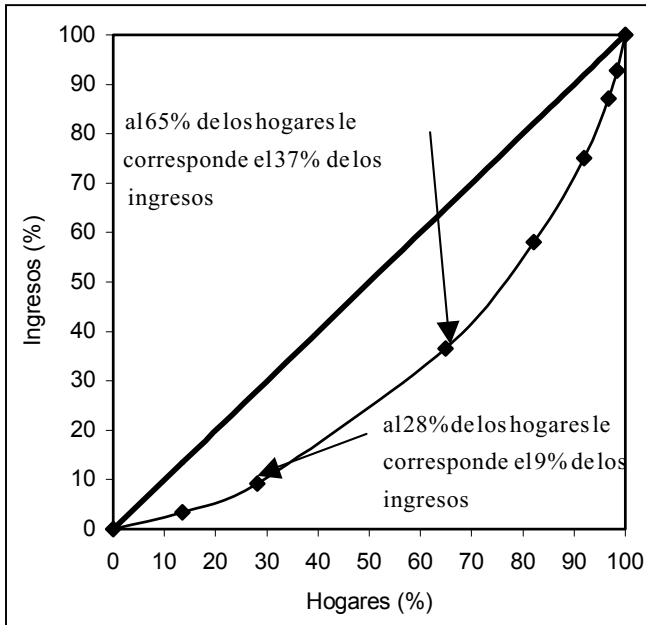
Así entonces, a manera de ejemplo, podemos observar en la fila sombreada que, los 2.736 hogares con ingresos menores que \$1.243, acumulan \$1.611.711; esto significa **que el 82% del total de hogares que menos ganan, participan con sólo el 58% del monto total de ingresos percibido por el conjunto de familias observadas.**

Con igual criterio se interpretan los valores acumulados (absolutos y relativos) para todas las clases de la distribución. Las cifras relativas presentadas en (7) y (8) permiten construir la **curva de Lorenz**. El porcentaje acumulado de los hogares (7), estará representado en el eje de abscisas y el porcentaje acumulado de los ingresos (8) en el eje de ordenadas.

De esta manera, la curva queda determinada por los puntos que tienen por abscisa el porcentaje acumulado de hogares y por ordenadas el porcentaje de ingresos acumulados correspondientes. Así por ejemplo, el primer punto que representamos estará definido por las coordenadas (14;3), el

segundo punto perteneciente a la curva tendrá coordenadas (28;9) y así sucesivamente con los diferentes pares de porcentajes que tenemos en la tabla, hasta el punto (100;100).

Curva de Lorenz. Distribución de los ingresos de 3.330 hogares de la ciudad de Posadas- 1994



Esta gráfica tiene la ventaja de permitirnos apreciar de manera sencilla el nivel de concentración de la variable en estudio. En nuestro ejemplo, vemos que la curva define un área que está más cercana a la situación de equidistribución que a la de máxima concentración, y podríamos entonces calificarla como "moderada".

Como ocurre con la mayoría de **los gráficos**, tiene como limitación el que **no nos ofrece ningún nivel de precisión y la valoración es subjetiva**. A su vez, en el caso de tener que realizar una comparación entre dos conjuntos de datos, a no ser que se trate de situaciones extremas o muy diferentes, puede resultar aventurado concluir a partir de la apreciación visual de la gráfica. Para estos casos se hace necesario **definir un recurso numérico asociado a esta gráfica** que exprese el

nivel de concentración de la variable.

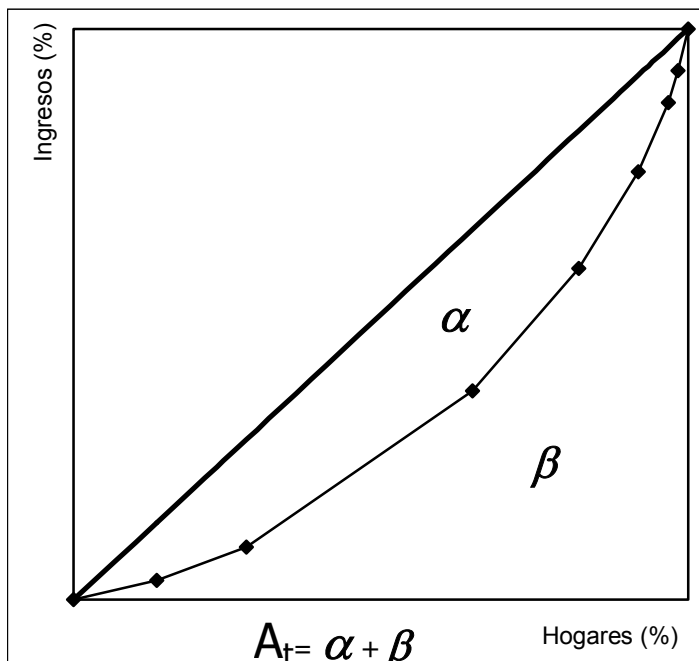
El índice de Gini



Como hemos visto, la curva define un **área de concentración** (que denominaremos α), delimitada por la recta de equidistribución y la curva obtenida; cuanto mayor sea el nivel de concentración de la variable en estudio, mayor será el área de concentración α .

También vimos que el área que se corresponde con la situación de **máxima concentración** coincide con el **triángulo inferior** determinado por la recta de equidistribución (**área total A_t**).

Gráfica de Lorenz: Área de concentración, área residual y área total



En las **situaciones intermedias**, vamos a poder identificar un área de concentración α , y un área residual β (diferencia entre el área total y el área de concentración), cumpliéndose en cualquier caso, que: $A_t = \alpha + \beta$.

El índice de Gini

Se lo define como el cociente entre el área de concentración α y el área total A_t . En símbolos:

$$I_G = \frac{\alpha}{A_t} \quad \text{siendo: } 0 \leq I_G \leq 1$$

$I_G=0$ cuando se trata de una situación de **equidistribución** ($\alpha=0$)

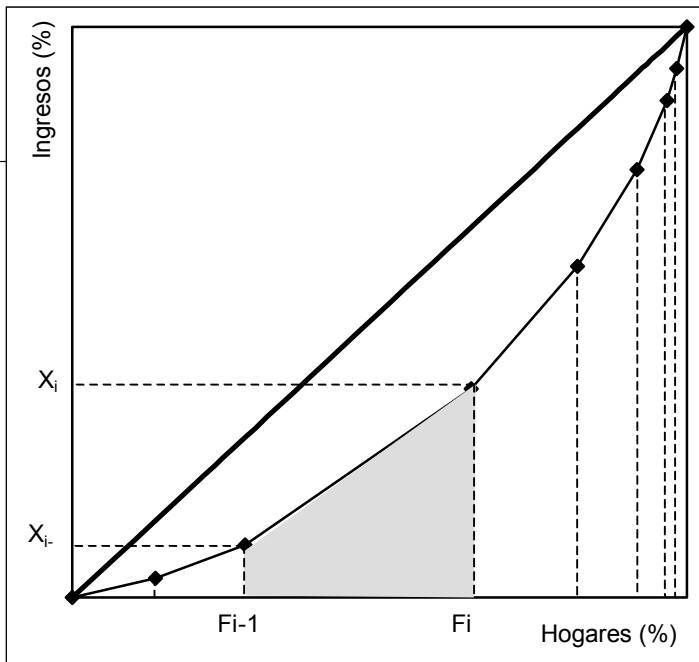
$I_G=1$ cuando se trata de una situación de **máxima concentración** ($\alpha=A_t$)

Como el cálculo del área β resulta más sencillo que el de α , al índice se lo plantea en términos de β , reemplazando α por $(A_t - \beta)$; de lo que resulta:

$$I_G = 1 - \frac{\beta}{A_t} \quad (11)$$

El área total A_t se determina como la mitad del área del cuadrado de lado 100; esto es 5.000. El problema es, por lo tanto, determinar el área β , que puede ser pensada como la sumatoria de las áreas de cada uno de los trapecios que componen el área total β . Se puede ver en el gráfico que tendremos tantos trapecios como intervalos de clase se hayan definido.

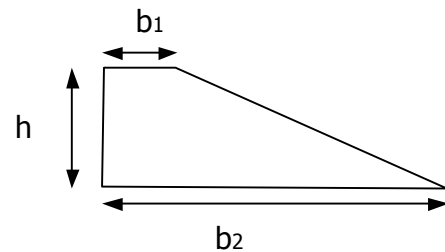
Gráfica de Lorenz: elementos para la determinación del área residual β



Recordemos que el área de un trapecio se obtiene como:

$$\frac{(b_1 + b_2) \cdot h}{2}$$

Donde:
 b_1 : base menor
 b_2 : base mayor
 h : altura



En la curva de Lorenz, y para el trapecio genérico planteado en la gráfica, tendremos:

$$b_1 = X_{i-1} \quad b_2 = X_i \quad h = F_i - F_{i-1}$$

donde: X_i es la variable acumulada en porcentaje hasta el intervalo genérico i

X_{i-1} es la variable acumulada en porcentaje hasta el intervalo anterior a i .

F_i es la frecuencia acumulada porcentual hasta el intervalo i .

F_{i-1} es la frecuencia acumulada porcentual hasta el intervalo anterior a i .

¹¹ $I_G = \frac{\alpha}{A_t} = \frac{A_t - \beta}{A_t} = 1 - \frac{\beta}{A_t}$

Entonces, el área. β está dada por:

$$\beta = \sum_{i=1}^k \frac{(X_{i-1} + X_i) \cdot (F_i - F_{i-1})}{2} \quad \text{donde } k \text{ es el número de intervalos de clase.}$$

Siendo el índice de Gini: $I_G = 1 - \frac{\beta}{A_t}$

Y el área β es: $\beta = \sum_{i=1}^k \frac{(X_{i-1} + X_i) \cdot (F_i - F_{i-1})}{2}$

Luego:

$$I_G = 1 - \frac{\beta}{A_t} = 1 - \frac{\sum_{i=1}^k \frac{(X_{i-1} + X_i) \cdot (F_i - F_{i-1})}{2}}{5000} = 1 - \frac{1}{10000} \sum_{i=1}^k (X_{i-1} + X_i) \cdot (F_i - F_{i-1})$$

En síntesis, se utiliza como **fórmula de trabajo**, la siguiente expresión:

$$I_G = 1 - \frac{1}{10000} \sum_{i=1}^k (X_{i-1} + X_i) \cdot (F_i - F_{i-1}) \quad (12)$$



Para los datos de los 3.330 hogares de Posadas, el Coeficiente de Gini, se obtendría como:

Ingresos familiares mensuales – Posadas, 1994.

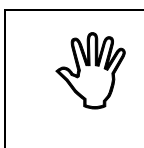
Ingresos familiares ⁽¹⁾	Hog. Acum. (%) ⁽⁷⁾	Ing. Acum. (%) ⁽⁸⁾	$X_{i-1} + X_i$ ⁽⁹⁾	$F_i - F_{i-1}$ ⁽¹⁰⁾	$(X_{i-1} + X_i) \cdot (F_i - F_{i-1})$ ⁽¹¹⁾
165-249	14	3	3	14	42
249-414	28	9	12	14	168
414-829	65	37	46	37	1702
829-1243	82	58	95	17	1615
1243-1658	92	75	133	10	1330
1658-2487	97	87	162	5	810
2487-3316	98	93	180	1	180
3316-4146	100	100	193	2	386
TOTAL					6233

Reemplazando en la fórmula:

$$I_G = 1 - \frac{1}{10000} \sum_{i=1}^k (X_{i-1} + X_i) \cdot (F_i - F_{i-1}) = 1 - \frac{1}{10000} 6233 = 1 - 0,6233 = 0,377$$



Se puede ver que el *área de concentración* representa un 37,7% del *área total*, valor que expresa una **concentración moderada de los ingresos**.




Actividad N° 9

Antes de continuar con la lectura, es necesario realizar aquí la Actividad N° 9 de la Guía de Actividades correspondiente a esta unidad.

¹² Si los valores se expresaran en términos relativos no porcentuales, la expresión del índice es: $I_G = 1 - \sum_{i=1}^k (X_{i-1} + X_i) \cdot (F_i - F_{i-1})$

4.4. Otras consideraciones sobre los recursos gráficos

Hasta aquí hemos presentado la construcción y utilidad analítica de los recursos numéricos o tabulares, así como las alternativas gráficas con las que se corresponden y complementan. El recurso gráfico ofrece una amplia gama de posibilidades que no pretendemos agotar en esta presentación, sino señalar sus principales alcances y limitaciones, a partir de las cuales el investigador, basándose en su creatividad, podrá generar nuevas alternativas. Dado que existen programas informáticos -como Excel, que permiten construir fácilmente una gran variedad de gráficos- esta presentación se dirige principalmente a precisar los criterios que se deben tomar en cuenta a la hora de seleccionar e interpretar un gráfico.



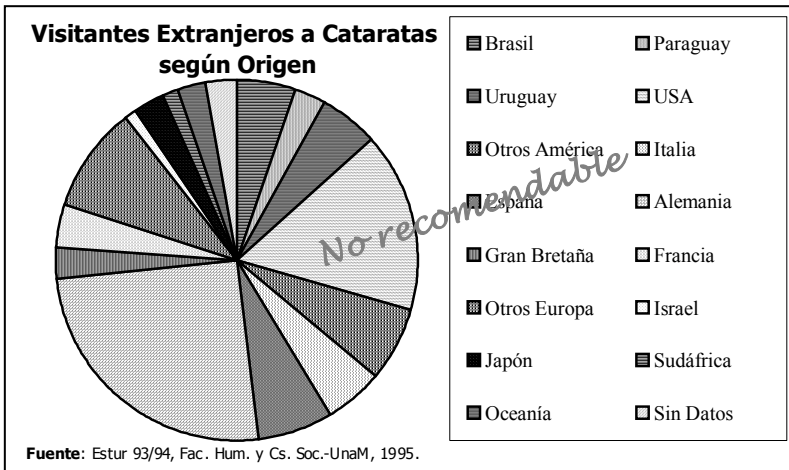
IMPORTANTE

Hemos presentado hasta aquí recursos gráficos asociados a las distribuciones de frecuencias absolutas (de sectores, de barras, de bastones, histogramas y polígonos); es necesario destacar que **esos mismos gráficos pueden ser contruidos para las distribuciones de frecuencias relativas**. Estos gráficos conservan la forma de la distribución y según sea el interés del investigador se decidirá por una u otra alternativa de representación.

Sobre este recurso queremos destacar algunos aspectos, que entendemos fundamentales:

- Los gráficos **no tienen un papel secundario** en el análisis y la presentación de datos. No son un "adorno" en los informes.
- Su capacidad de expresar de manera sencilla una gran cantidad de información los convierte en un **recurso poderoso** no solo para la presentación de resultados, sino para la **exploración y análisis** de los datos.
- Esta capacidad de transmitir mucha información en forma inmediata exige que se deban observar cuidadosamente **algunos principios**. Ellos tienen que ver con:
 - Evitar el exceso de información en un mismo gráfico.
 - Evitar la inclusión de gráficos que no aporten información relevante (son inexpresivos y se sobrecarga inútilmente el informe).
 - Seleccionar gráficos que tomen en cuenta el destinatario (científicos, de divulgación, etc.). Hay gráficos que normalmente sólo podrán ser decodificados por especialistas.
 - Respetar las reglas técnicas, fundamentalmente relativas a la construcción de las escalas, la consideración del tipo de variables, etc.; para evitar el riesgo de generar una impresión equivocada sobre los datos.
 - De los gráficos posibles para la presentación o análisis de un determinado tipo de datos, seleccionar aquellos que mejor destacan las características que interesa mostrar (estructura, evolución, participación, etc.).

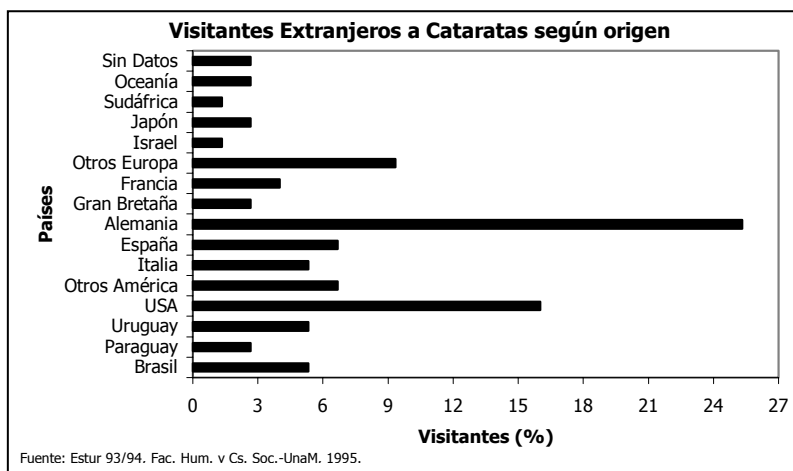
Algunos gráficos que ilustran los aspectos señalados precedentemente:



a) Queremos mostrar en un gráfico la distribución de los visitantes extranjeros a Cataratas del Iguazú según su origen. Dado que se trata de la distribución de una variable categórica un gráfico de sectores o de torta aparece como una alternativa válida de presentación para mostrar el diferente peso relativo que tienen los distintos emisores identificados.

La gran cantidad de categorías identificadas para

la variable origen, hace que este Gráfico de sectores -técnicamente correcto- resulte inapropiado dado el gran número de comparaciones que obliga a realizar para su lectura. Esto es incongruente con el propósito de la construcción de un gráfico: simplicidad e inmediatez para captar la información resumida.



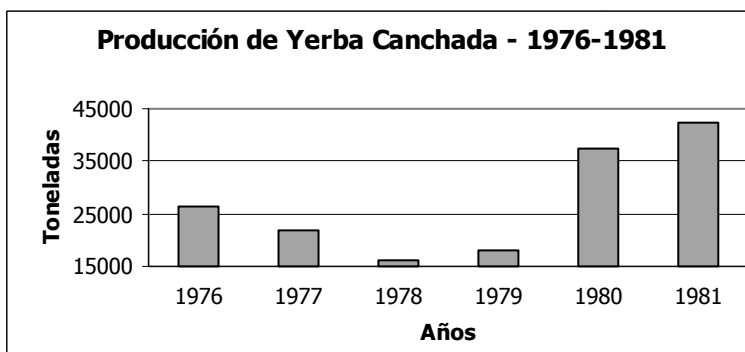
Para presentar esta misma información una alternativa es utilizar un gráfico de barras horizontales¹³ como el siguiente.

En el Gráfico se destaca inmediatamente la importante participación de visitantes de la Unión Europea, estadounidenses y otros países de Europa, como así también brasileños y uruguayos.

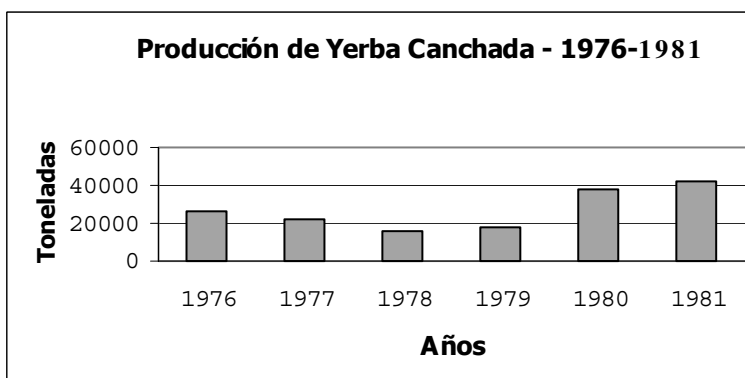
b) Modificando las escalas se pueden producir, para un mismo conjunto de datos, distorsiones en los gráficos que generan en un observador desprevenido impresiones totalmente diferentes respecto al comportamiento de los mismos. Esto obliga a ser muy cuidadoso tanto en la construcción (en el caso de quien los produce) como en la lectura de los mismos (por parte de quien los quiere interpretar).

Presentamos a continuación dos conjuntos de datos longitudinales que ejemplifican diferentes situaciones relativas a la modificación de las escalas.

b.1) Son dos gráficos sobre la producción de yerba canchada en la provincia de Misiones durante el período 1976-1981.



Aquí se presentan los datos con la producción por encima de las 15.000 toneladas. En términos gráficos significa que el eje horizontal no corta al vertical en el origen (cero), sino a la altura de los 15.000.



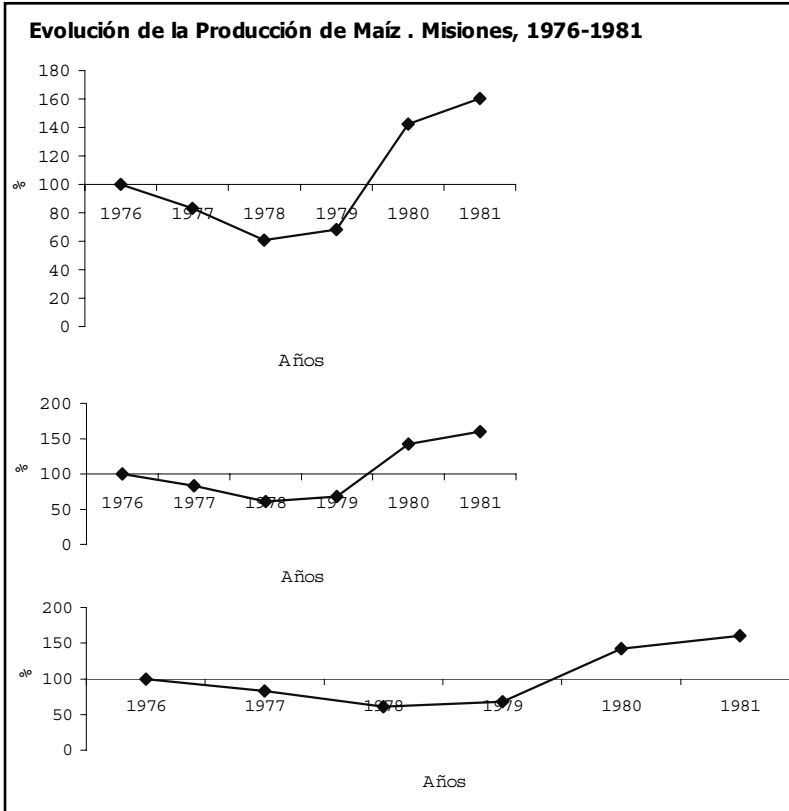
En este segundo Gráfico se muestra la escala vertical desde cero y, en consecuencia, la altura de las barras es proporcional a la producción en toneladas.

La comparación de estos Gráficos pone de manifiesto que, con la primera alternativa de representación, "exageramos" las variaciones que se producen a lo largo del

¹³ Para evitar la superposición de los nombres de las categorías (además extensos en este caso) que ocurre cuando se usa un gráfico de barras verticales.

período analizado. Ejemplo: en el primer Gráfico, la producción del año '78 pareciera representar menos de la tercera parte de la registrada en el 77. Esta impresión se corrige cuando observamos el segundo Gráfico.

b.2) Son tres Gráficos en los que se representa la evolución de la producción de maíz en Misiones entre 1976 y 1981, tomando 1976 como base (=100).



En cada uno de ellos se modifican las escalas de los ejes x e y provocando en el comportamiento de la serie impresiones visuales muy diferentes.

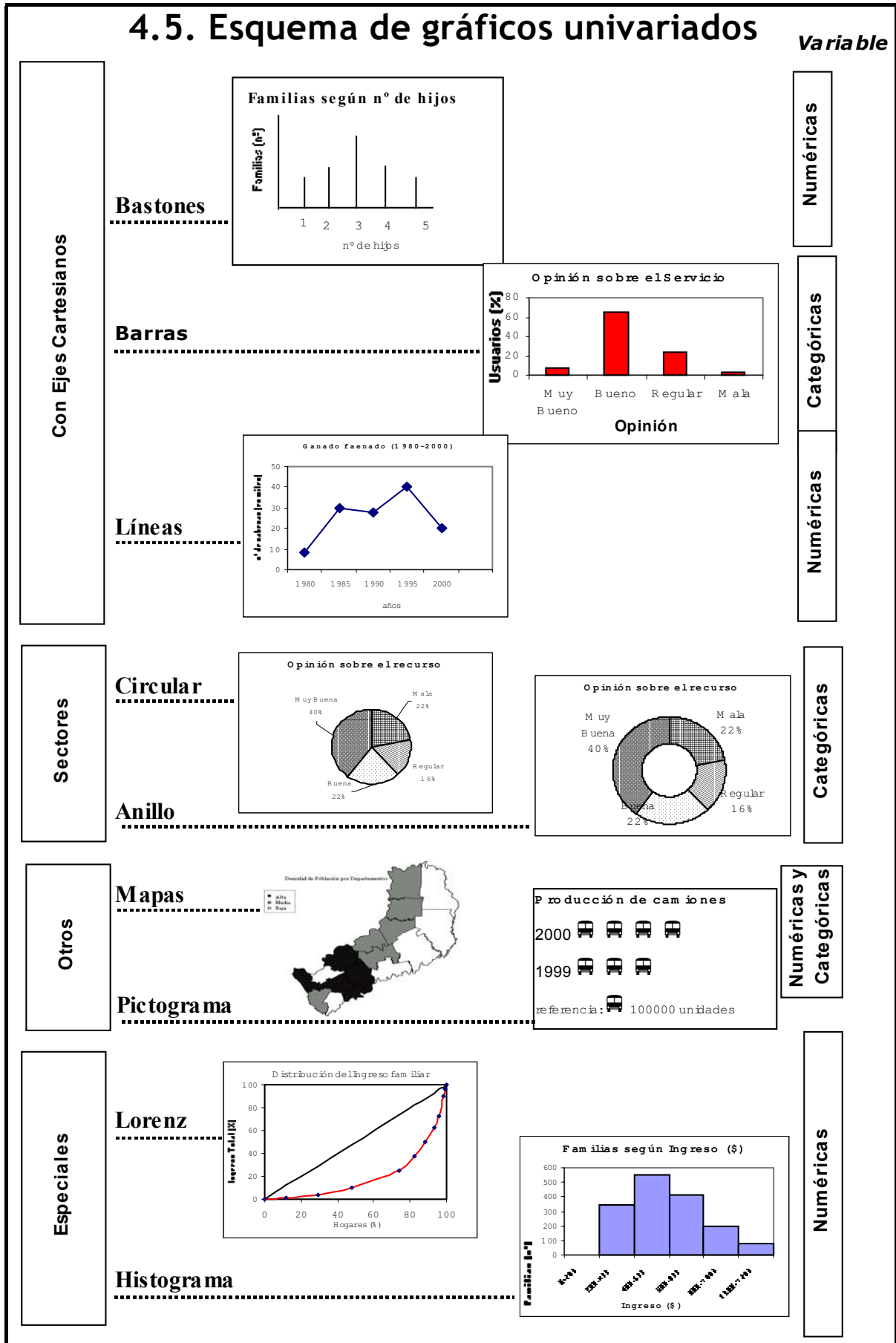
Con relación al primer Gráfico:

- ✓ en el segundo, las variaciones aparecen exageradas por haber modificado la escala del eje y,
- ✓ en tanto que en el tercero, la Gráfica suaviza la serie (los "saltos" de un año a otro parecen más pequeños) al haber modificado la escala del eje x.

La recomendación que intentamos ejemplificar en este caso, es que se debe **mantener la misma escala cuando se desean comparar distintas series.**

Con estos ejemplos no pretendemos agotar los casos de distorsiones que se pueden producir a la hora de utilizar el recurso gráfico, sino más bien alentar una actitud crítica cuando se construyen gráficos, y también cuando se interpretan gráficos ya construidos.

4.5. Esquema de gráficos univariados



5. ¿Qué Hemos Visto? (*)

En esta unidad hemos iniciado el camino del tratamiento y análisis de los datos.

*Superada la primer instancia de organizar las observaciones en una **matriz de datos** que facilita su tratamiento estadístico, comenzamos el **proceso de análisis** guiados por las **preguntas iniciales** de investigación. Estas preguntas pueden determinar la necesidad de trabajar con **una, dos o más variables** simultáneamente; sin embargo, la exploración de cada una de las variables (**análisis univariado**) es un proceso **necesario** en varios sentidos: porque nos permitirá empezar a comprender el fenómeno en estudio, reformular algunas clasificaciones, evaluar la posibilidad de aplicar otras herramientas de análisis, dar respuestas a las preguntas más simples y formularnos nuevas preguntas.*

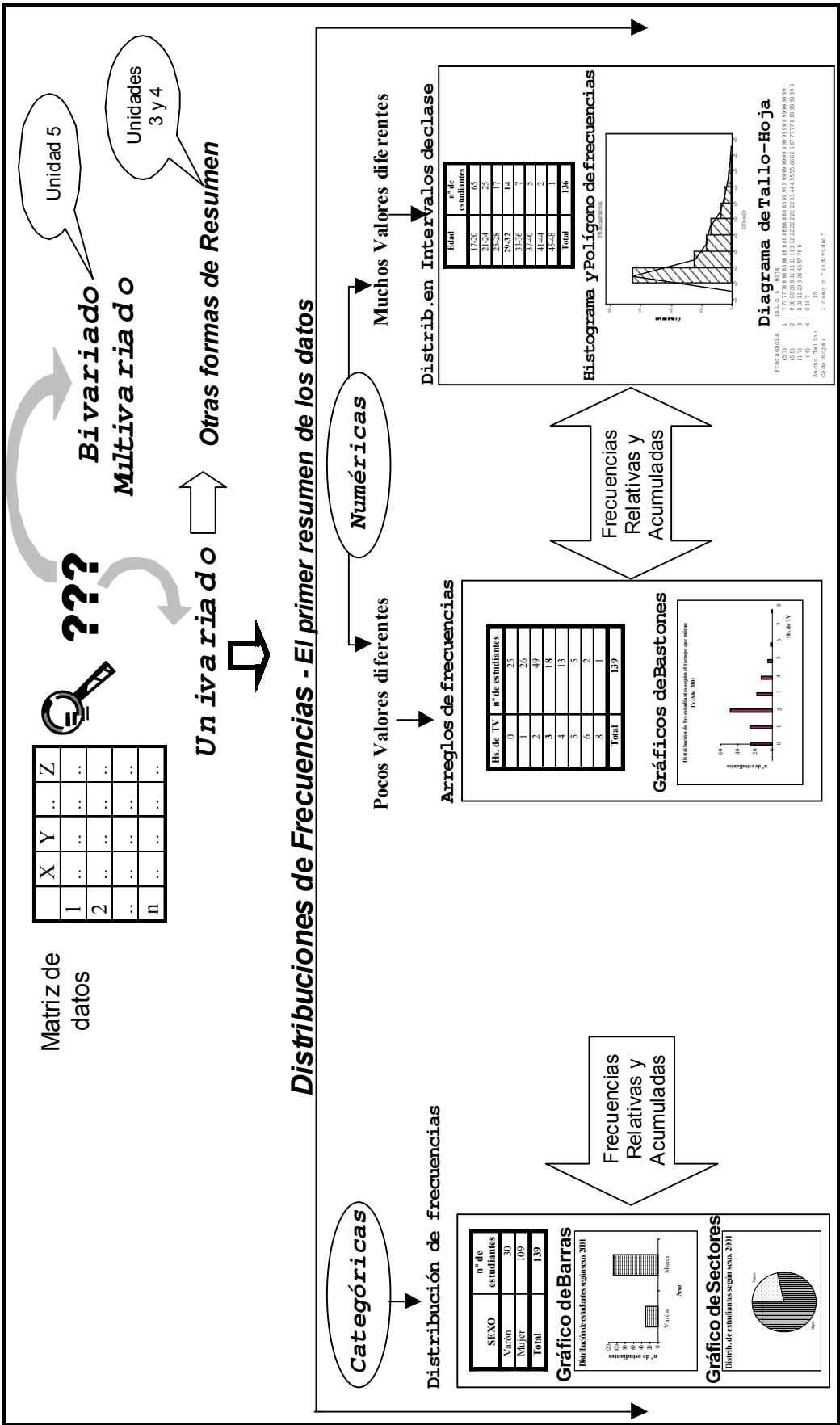
*En el análisis univariado, el primer resumen de los datos son las **distribuciones de frecuencias**, para cuya construcción debemos considerar inicialmente el tipo de variable a trabajar (**numérica o categórica**).*

*Hecha esa distinción se pueden adoptar distintas estrategias en el abordaje de los datos; así, aparecen los **recursos numéricos y gráficos** como dos herramientas poderosas y complementarias en esta tarea de comprender el comportamiento de los datos y comunicar la información producida. Priorizar una u otra herramienta en el trabajo de exploración es una decisión del investigador.*

*Además, hemos presentado transformaciones de las frecuencias absolutas (**frecuencias relativas y acumuladas**) que facilitan y enriquecen las posibilidades de análisis e interpretación de las distribuciones de frecuencias. Asociado a las transformaciones de las frecuencias se presentaron un recurso gráfico (curva de Lorenz) y un recurso numérico (Índice de Gini) que resultan de suma utilidad en el análisis de la distribución/concentración de algunas variables económicas (renta, tierra, ingreso, etc.).*

En todos los casos, hemos intentado presentar para cada herramienta el tipo de preguntas a las que pueden responder, el cuándo utilizarlas y cómo hacerlo, destacando a su vez sus alcances y limitaciones como recurso analítico y de comunicación.

(*) ver esquema en la página siguiente.



Bibliografía

MOORE, D. (1995): *Estadística Aplicada Básica*. Antoni Bosch Editor, Barcelona. Páginas: 6 a 21.

ALAMINOS, A. (1993): *Gráficos*. Colección "Cuadernos Metodológicos" nº 7. Centro de Investigaciones Sociológicas, Madrid. Páginas: 7 a 14 y 23 a 27.

BLALOCK, H. M (1986): *Estadística Social*, México, FCE. Páginas: 43 a 64.

Conceptos Centrales

- Matriz de datos.
- Distribuciones de frecuencias.
- Arreglos y distribución en intervalos de clase: tablas y gráficos
- Frecuencias relativas y frecuencias acumuladas (absolutas y relativas).

Habilidades

- Organizar un conjunto de datos en distribuciones de frecuencias.
- Construir gráficos de distribuciones de frecuencias.
- Describir la *forma* de una distribución.
- Reconocer y obtener las transformaciones necesarias de las frecuencias absolutas para responder preguntas específicas.
- Interpretar la información resumida en una distribución de frecuencias
- Comunicar los resultados del análisis.