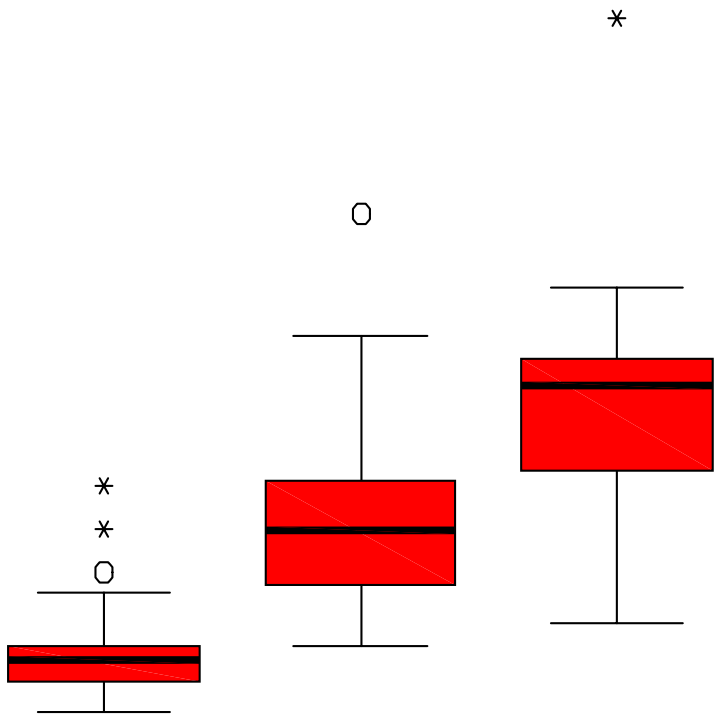
ESTADÍSTICA APLICADA EN LAS CIENCIAS SOCIALES Y HUMANAS

ESTADÍSTICA I



César N. AGUIRRE M. Fernanda NIÑO Eduardo F. SIMONETTI

EDITORIAL UNIVERSITARIA DE MISIONES

San Luis 1870
Posadas - Misiones – Tel-Fax: *(03752)* 428601
Correos electrónicos:
edunam-admini@arnet.com.ar
edunam-direccion@arnet.com.ar
edunam-produccion@arnet.com.ar
edunam-ventas@arnet.com.ar

Colección: Cuadernos de Cátedra
Coordinación de la edición: Nicolás Capaccio
Tapa: Francisco Sánchez
Compaginación y armado de interiores: Amelia E. Morgenstern
Corrección: Hedda Giraudo - Amelia E. Morgenstern

ISBN 987-9121-98-8
1º reimpresión
Impreso en Argentina

Universidad Nacional de Misiones
Posadas, 2005

Aguirre, César
Estadística aplicada en las ciencias sociales y humanas / César Aguirre; M. Fernanda Niño y Eduardo F. Simonetti; coordinado por Rodolfo Nicolás Capaccio - 1a ed. 1a reimp. - Posadas: Univ. Nacional de Misiones-Editorial Universitaria, 2005.
240 p.; 30x21 cm. (Cuadernos de Cátedra)
ISBN 987-9121-98-8
1. Sociología I. Niño, M. Fernanda, II. Simonetti, Eduardo F., III. Rodolfo Nicolás Capaccio, coord. IV. Título
CDD 301.

PRESENTACIÓN

¿POR QUÉ SABER DE ESTADÍSTICA?

La Estadística ha ganado reconocimiento como disciplina importante en la formación profesional universitaria de diferentes campos del conocimiento. Sus aportes a esta formación ocurren en dos niveles: el *primero* como disciplina contributiva a la preparación científica de los estudiantes, proporcionando los conocimientos indispensables e irremplazables en la producción, tratamiento y análisis de datos cuantitativos. El *segundo* nivel se manifiesta en el plano de la formación general de los individuos que deben desenvolverse en el mundo actual, intensamente conectado por las modernas telecomunicaciones, y cada vez más interdependiente en lo económico y social. "*En este nuevo mundo es importante ser capaz de orientarse en la red de información disponible, mucha de la cual es cuantitativa. El ciudadano debe moverse entre masas de datos cuantitativos que muchas veces son contradictorios y requieren de un mínimo de conciencia sobre la forma en que tales datos se recogen, organizan, analizan e interpretan. Como resultado se precisan nuevas habilidades*"[1].

El ciudadano medio de hoy necesita reunir conocimientos que le sirvan para decodificar la cuantiosa información que recibe por diferentes medios, que le permitan juzgar la calidad de esa información, que le sean útiles para comprender ideas expresadas por otros y formar las propias, que le sirvan para construir y comunicar su propia información.

La Estadística es la disciplina que aporta los recursos culturales y prácticos que el ciudadano necesita para desenvolverse en la nueva sociedad de la información, y la enseñanza de la disciplina tiene el fin de generar y difundir una "cultura estadística" dirigida a dotar al ciudadano común de tales recursos y habilidades.

OBJETIVOS GENERALES

El curso de estadística aplicada tiene el propósito de *promover la formación* de *usuarios competentes* de esta disciplina y sus herramientas. Usuarios con habilidades que le permitan *vincular* los *conocimientos estadísticos* con la *solución de problemas* de su campo disciplinar o profesional específico.

Es decir, se trata de desarrollar en los alumnos la capacidad de **abordar y resolver, desde la Estadística, problemas de producción de información** con fines científicos o de tomar decisiones.

Los ***objetivos generales*** derivados de este propósito de formación, son:

1. Promover en los alumnos el desarrollo de habilidades intelectuales del *pensamiento-razonamiento estadístico*. Ello supone abordar la solución de un problema de trabajo desde los siguientes elementos básicos:

- la *necesidad de datos* para alcanzar una correcta comprensión del fenómeno o problema en estudio (*impulso estadístico)*;
- la idea de la *"transnumeración*", entendida como la habilidad de construir y producir *datos como representaciones de aspectos de un sistema real* para lograr una *mejor comprensión* de dicho sistema (datos que capturan información significativa sobre elementos del sistema);
- la necesidad de la *observación masiva* (cantidad numerosa de datos: *datos estadísticos*) como fundamento del *análisis estadístico* de los fenómenos;
- la presencia de la *variación* en los datos (y en los fenómenos que ellos representan) y el *error e incertidumbre* como consecuencia de ella;

[1] OTTAVIANI, M. G. (1999): *Notas sobre los Desarrollos y Perspectivas en Educación Estadística*. International Association for Statistical Education -IASE-.

- la necesidad del *resumen, descripción y modelización* de la variación.

2. Capacidad para *formalizar un problema de trabajo* o investigación en términos (preguntas) estadísticos.

3. Capacitar para *la construcción y obtención* de los datos que requiere la solución de un problema.

4. *Conocer herramientas* estadísticas para el tratamiento y análisis de datos y *comprender sus fundamentos* lógicos, limitaciones, propiedades, etc.

5. Manejar la *tecnología auxiliar disponible* para la aplicación de las *herramientas estadísticas*.

6. *Integrar* apropiadamente diferentes *herramientas estadísticas* en el análisis de un problema (estrategia de abordaje estadístico del problema).

7. *Interpretar los resultados* estadísticos en el contexto del problema de estudio y *comunicar los hallazgos* o respuestas a las preguntas iniciales (información-comunicación).

ENFOQUE DEL CURSO

Tratándose de un Curso de Estadística para "*no estadísticos"* y considerando los objetivos generales citados anteriormente, la estrategia pedagógica a seguir **enfatizará la *comprensión conceptual* de los contenidos** a desarrollar. Es decir, se privilegiará la *conceptualización de la Estadística*, por sobre la demostración y el tratamiento matemático de sus conocimientos.

También se pondrá mucho énfasis en **vincular los conceptos y herramientas con la solución de problemas reales de investigación o de decisiones**. De ahí que **el *cálculo estadístico no constituye una actividad central*** del curso y se realizará –únicamente- con propósitos pedagógicos.

En este primer nivel del Curso de Estadística se presentarán con la mayor profundización posible, temas relativos a un análisis descriptivo de los datos (Estadística Descriptiva). En esta pretensión, resulta básico e insoslayable el abordaje de los siguientes tópicos:

- ✓ La Investigación Estadística (Unidad 1)
- ✓ Organización y Descripción Inicial de los Datos (Unidad 2)
- ✓ Los valores que Caracterizan al Conjunto de Datos (Unidad 3)
- ✓ Análisis de la Variación y Asimetría (Unidad 4)
- ✓ Estudio de la Relación entre Variables (Unidad 5)
- ✓ Los Números Índices (Unidad 6)

Este Curso de Estadística ha sido pensado como una **propuesta no presencial** de formación. Por ello, los contenidos y las actividades han sido organizados y producidos de tal manera que el trabajo pueda auto-administrarse sin grandes dificultades.

El presente material incluye, para cada unidad, lo que se dio en llamar "***Notas de Cátedra***", en las que se ***desarrollan los conceptos teóricos centrales***, orientando el aprendizaje [2]. Además, las "Notas de Cátedra" remiten periódicamente a las "***Guías de actividades****",* en las que se proponen consignas de trabajo (teóricas y/o prácticas) que facilitan la comprensión de los conceptos tratados y le plantean situaciones concretas de análisis de datos que favorecen el desarrollo de las habilidades necesarias para el tratamiento de datos y la producción de información.

Al final de las "***Guías de Actividades"*** se presenta un propuesta de trabajo denominada ***"Evaluación Parcial de la Unidad…."***. Se trata de una actividad de síntesis de los conocimientos teóricos y prácticos desarrollados en la unidad, mediante la cual se podrán evaluar los avances en sus conocimientos.

[2] En cada unidad, se recomienda bibliografía complementaria.

ORIENTACIONES PARA EL USO DEL MATERIAL

El material incluye señales que intentan orientar la lectura y subrayar aquellas cuestiones en las que se debe poner especial atención. También ofrece esquemas y gráficos que sintetizan un conjunto de conceptos y las relaciones que se pueden establecer entre ellos.

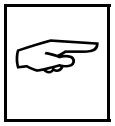
Además de los recursos utilizados tradicionalmente para destacar algún aspecto parcial de la presentación (uso de negritas y/o cursivas), en las Notas de Cátedra se han utilizado un conjunto de íconos que señalan partes diferentes -en cuanto a su naturaleza- del desarrollo teórico. Así, aparecerán:

En general, los temas se presentan a partir de situaciones que ponen en evidencia la necesidad de nuevas herramientas de análisis. Estas situaciones se traducen en preguntas de investigación y estadísticas, las que requieren el uso de herramientas específicas para encontrar una respuesta. El icono señala el carácter de **planteo general** del texto.

Destaca en el texto los **conceptos y definiciones.**

Indica el desarrollo de un **ejemplo** donde se utilizan los conceptos presentados.

Enfatiza lo **importante** de algunas cuestiones consideradas en el desarrollo del tema.



Señala la parte del texto donde se realiza la **interpretación** de los resultados estadísticos obtenidos.

Advierte sobre la necesidad de **hacer un alto en la lectura** y realizar la actividad que se indica.

Señala las **consignas de trabajo** a realizar, en las Guías de Actividades de cada unidad.

En todas las unidades se han incluido ***esquemas*** de los contenidos tratados, una síntesis de "***lo visto***", una enumeración de los ***conceptos centrales*** y las ***habilidades*** que se pretendieron transmitir. El propósito de incluir estos recursos es **brindarle una mirada global** sobre la unidad y simultáneamente **destacar las ideas centrales que la estructuran**, así como las habilidades que se asocian a esas ideas en la práctica. Consecuentemente, le recomendamos especial atención a estas diferentes formas de síntesis de cada unidad, ya que constituyen otra forma de aproximación a los conceptos desarrollados y le permitirán reorientar una segunda lectura del material.

Los Autores:

AGUIRRE, César Norberto
Estadístico (Universidad Nacional de Rosario -Argentina-), postgrado en Estadística y Cuentas Nacionales (Instituto de Estudios Sociales de La Haya -Holanda-), Especialización en Administración Estratégica Universitaria (Universidad de Quebec -Canadá-), Especialización en Administración Estratégica de Negocios (Universidad Nacional de Misiones).
Profesor Regular Titular de Estadística (Facultad de Humanidades y Ciencias Sociales -UNaM-), Profesor Titular de Estadística (Maestría de Gestión Pública -UNaM-), ex Docente de Cursos de Postgrado en Análisis Exploratorio de Datos (Programa PRESTA, Universidad Libre de Bruselas-Unión Europea).

NIÑO, María Fernanda
Profesora de Matemática, Física y Cosmografía, Inst. Sup. del Profesorado "Pbro. Dr. Antonio Saenz". Maestría en Docencia Universitaria, Fac. de Ingeniería-Univ. Nac. Misiones (etapa elaboración de tesis).
Ayudante de Primera (Regular) de Estadística (Fac. de Humanidades y Ciencias Sociales de la UNaM). Ex docente de Cursos de Posgrado de Métodos Estadísticos Multivariados Aplicados a las Ciencias Humanas Y Sociales, del Programa PRESTA (Univ. Libre de Bruselas), auspiciado por la Unión Europea. Docente en el Curso de Estadística Aplicada, (Maestría en Gestión Pública de la Facultad de Ciencias Económicas-UNaM). Ex docente tutor del Seminario "Metodología y Técnicas de la Investigación Social" (Maestría en Gerencia y Administración de Proyectos Sociales -UNaM-)

SIMONETTI, Eduardo Francisco
Estadístico (Universidad Nacional de Rosario -Argentina-), Master en Desarrollo Económico para América Latina, Universidad Internacional de Andalucía – Sede Iberoamericana de La Rábida (España).
Profesor Titular Regular de "Indicadores Socioeconómicos", (Facultad de Humanidades y Ciencias Sociales de la UNaM). Docente en el Curso de Estadística Aplicada, (Maestría en Gestión Pública de la Facultad de Ciencias Económicas-UNaM). Docente del Seminario "Sistemas de Información y Herramientas Informáticas para la Gestión de Programas Sociales" (Maestría en Gerencia y Administración de Proyectos Sociales -UNaM-).

ÍNDICE

Unidad 1: La Investigación Estadística

Página

1. Introducción 11
2. Intentando Definir la Estadística 11
3. Problema de Trabajo e Investigación Estadística 12
3.1. Las Preguntas de Investigación 13
3.2. Las Preguntas estadísticas 13
4. Los Datos 13
5. Las Variables 15
6. Conjunto de Datos: Datos Estadísticos 17
7. Fuentes de Datos 18
8. Investigación por Censo y por Muestra 19
9. ¿Qué Hemos Visto? 20
Esquema-La Estadística en el Proceso de Investigación 21
Esquema – Estructura del Curso – Estadística Descriptiva 22
Bibliografía 23

Unidad 2: Organización y Descripción Inicial de los Datos

1. Los Datos y la Información 25
2. La Primera Organización de los Datos: la Matriz de Datos 25
3. El Análisis de la Matriz de Datos 28
4. Las Distribuciones de Frecuencias en el Análisis Univariado 29
4.1. Variables categóricas 30
- el recurso numérico 30
- el recurso gráfico 31
4.2. Variables numéricas 32
4.2.1. Variables numéricas con pocos valores diferentes 32
- el recurso numérico 32
- el recurso gráfico 33
4.2.2. Variables numéricas con muchos valores diferentes 34
- el recurso numérico 34
- el recurso gráfico 39
4.3. Transformaciones de las frecuencias absolutas 42
4.3.1. Las frecuencias relativas 42
4.3.2. Las frecuencias acumuladas 43
4.3.3. La curva de Lorenz y el índice de Gini 45
4.4. Otras consideraciones sobre los recursos gráficos 52
4.5. Esquema – Tipos de gráficos univariados 55
5. ¿Qué Hemos Visto? 56
Esquema – El Análisis de Datos: Distribuciones de Frecuencias 57
Bibliografía 58

Unidad 3: Los Valores que Caracterizan al Conjunto de Datos

Página

1. ¿Por qué son Necesarios? ... 59
2. ¿Cuáles Son? ... 60
3. Media Aritmética ... 60
3.1. Principales Propiedades de $\bar{x}$ ... 61
3.2. Cálculo de la Media ... 64
3.2.1. Datos sin resumir ... 64
3.2.2. Datos agrupados en arreglo de frecuencias ... 64
3.2.3. Datos agrupados en una distribución con intervalos ... 65
4. La Mediana ... 66
4.1. Principales propiedades de Ma ... 67
4.2. Determinación de la Ma ... 68
4.2.1. Datos numéricos sin resumir ... 68
4.2.2. Datos numéricos en arreglo de frecuencias ... 69
4.2.3. Datos numéricos en una distribución con intervalos ... 70
4.2.4. Datos categóricos ordinales ... 71
5. El Modo ... 72
5.1. Principales Propiedades del Mo ... 73
5.2. Determinación del Mo ... 73
5.2.1. Para arreglos de frecuencias y datos categóricos ... 73
5.2.2. Para una distribución con intervalos ... 74
6. Cuartiles, Deciles, Centiles ... 75
6.1. Los Cuartiles ... 76
6.2. Los Deciles ... 77
6.3. Los Centiles ... 78
6.4. Curva de Lorenz asociada a las medidas de posición ... 78
7. ¿Cómo Integrar estas Medidas de Resumen? ... 80
7.1. El resumen de los cinco números ... 80
7.2. El diagrama de Caja (*Box-plot*) ... 81
8. ¿Qué Hemos Visto? ... 82
Esquema – Valores que Caracterizan un Conjunto de Datos ... 83
Bibliografía ... 84

Unidad 4: Análisis de la Variación y Asimetría

1. ¿Por qué Evaluar la Variabilidad y la Asimetría? ... 85
2. ¿Cómo Medir la Variabilidad? ... 86
2.1. Para variables numéricas ... 86
2.1.1. Las medidas absolutas ... 87
A) El Rango, Amplitud o Recorrido ... 87
B) El Rango Intercuartil ... 88
C) Desviación Media ... 88
D) Desviación Mediana ... 90
E) Variancia y Desviación estándar ... 91
2.1.2. Las medidas relativas ... 92
F) Coeficiente de variación ... 93
G) Coeficiente de Desviación Media ... 94

Página

H) Coeficiente de Desviación Mediana 94
2.2. Dispersión para variables categóricas 94
3. ¿Cómo Medir la Asimetría? 97
3.1. Coeficiente de Asimetría de Pearson 98
3.2. Coeficiente intercuartílico de Bowley 99
4. ¿Qué Hemos Visto? 101
Esquema – Medidas de Dispersión y Asimetría 102
Bibliografía 103

Unidad 5: Estudio de la Relación entre Variables

1. ¿Por qué Estudiar la Relación entre Variables? 105
2. La Relación entre Variables Categóricas 108
2.1. El recurso numérico 108
2.2. El recurso gráfico 114
2.2.1. Gráficos compuestos 114
2.2.2. Gráficos de partes componentes 115
3. La Relación entre Variables Categóricas y Numéricas 116
3.1. El recurso numérico 116
3.2. El recurso gráfico 119
4. La Relación entre Variables Numéricas 120
4.1. El recurso gráfico 120
4.2. El recurso numérico 124
4.2.1. El análisis de regresión lineal simple 124
4.2.2. El coeficiente de correlación lineal de Pearson 127
5. ¿Qué Hemos Visto? 129
Esquema – Estudio de la Relación entre Variables 130
Bibliografía 131

Unidad 6: Los Números Índices

1. ¿Qué son y cuál es su utilidad? 133
2. Los Números Índices Simples 134
2.1. El Relativo Simple de Base Fija 134
2.2. El Relativo Simples de Eslabón 135
2.3. El Relativo Simple en Cadena 136
3. Los Números Índices Compuestos 137
3.1. El Índice de Agregados no Ponderados 138
3.2. El Índice de Promedio de Relativos no Ponderados 139
3.3. Los Índices de Agregados Ponderados 141
3.3.1. El índice de Laspeyres 141
3.3.2. El índice de Paasche 143
3.4. Los Índices de Promedios Ponderados de Relativos 144
3.4.1. El índice promedio ponderado de relativos de Laspeyres 144
3.4.2. El índice promedio ponderado de relativos de Paasche 145
4. Algunas Consideraciones Especiales Temas Especiales 146
4.1. El Índice de Valor 146

Página

4.2. El Cambio de Base de un Número Índice 147
4.3. El Empalme de Dos Números Índices Solapados.....148
4.4. Procedimiento de Números Índices en Cadena.....149
4.5. La Deflación de una Serie.....149
5. Problemas en la Construcción de los Números Índices 150
5.1. La Selección de la Muestra 150
5.2. La Elección del Período Base 151
5.3. La Ponderación Adecuada 151
5.4. La Selección del Promedio.....151
5.5. Los Cambios de Producto.....151
6. ¿Qué Hemos Visto? 152
Bibliografía 153

Anexo: GUÍA DE ACTIVIDADES

Unidad 1: La Investigación Estadística.....157
Unidad 2: Organización y Descripción Inicial de los Datos 167
Unidad 3: Los Valores que Caracterizan al Conjunto de Datos.....177
Unidad 4: Análisis de la Variación y Asimetría.....185
Unidad 5: El Estudio de la Relación entre Variables.....189
Unidad 6: Los Números Índices 199

Bibliografía General.....157

UNIDAD 1: *LA INVESTIGACIÓN ESTADÍSTICA*

4. Introducción

Al iniciar el aprendizaje de Estadística elemental*, "aplicada a la solución de problemas"*, probablemente a Ud. se le plantean interrogantes como los siguientes:

- ✓ *¿A qué tipo de problemas nos referimos?*
- ✓ *¿Cómo abordar la búsqueda de respuestas a un tema/problema desde la Estadística?*

A lo largo de este primer bloque de contenidos, Ud. encontrará la información para explicar estas preguntas.

2. Intentando Definir la Estadística

Hemos seleccionado algunos autores, quienes se refieren a la disciplina estadística del siguiente modo:

Daniel, W. W. pp 1

"La palabra estadística tiene relación con aquellos conceptos y técnicas que se emplean en la recopilación, organización, resumen, análisis, interpretación y comunicación de información numérica".

Anderson, Sweeney y Williams; pp 3

"En un sentido amplio, la estadística es el arte y la ciencia de reunir, analizar, presentar e interpretar datos*. Especialmente en los negocios y en la economía,* una razón básica *para esa recopilación, presentación e interpretación de datos,* es proporcionar a los administradores *y a quienes toman decisiones,* una mejor comprensión del entorno *para permitirles* tomar mejores decisiones*".*

Moore, D. S.; pp XXI y XXII

"La estadística es la ciencia que trata sobre la obtención de información a partir de datos numéricos*[...] Para la mayoría de las personas que utilizan la estadística, e incluso para muchos estadísticos profesionales, la estadística* es la disciplina que proporciona instrumentos e ideas *que permiten* utilizar datos numéricos *para* profundizar en la comprensión *de distintos temas [...] A pesar de que la estadística se fundamenta en una sólida base matemática, nuestro interés se centra en la* estadística aplicada*, que se puede dividir en tres campos de estudio: el* análisis de datos*, la* obtención de datos *y la* inferencia estadística".

Mood, A. M. pp 3

"La concepción profana de estadística suele incluir la recogida de grandes masas de datos y la presentación de éstos en tablas y gráficos; puede incluir también el cálculo de totales, promedios, porcentajes, etc. En todo caso, esta concepción tiene unos treinta años de retraso; estas operaciones, más o menos rutinarias, constituyen solamente parte incidental de la estadística de hoy." Estadística es también el diseño de experimentos, el diseño de sobrevisiones muestrales, la reducción y el proceso de datos y otras muchas cuestiones. (...) Describiremos la estadística como la tecnología del método científico que proporciona instrumentos para la toma de decisiones cuando prevalecen condiciones de incertidumbre.

Más allá de los matices que diferencian a estas ideas entre sí, todas ellas coinciden en ciertos elementos que conforman un mismo concepto básico de *estadística aplicada*, el que bien podríamos sintetizar del siguiente modo:

Es una disciplina que aporta los conocimientos y herramientas insustituibles para:

- **Diseñar y aplicar procedimientos de recolección de datos** (experimentos, muestras, censos, registros administrativos y fuentes secundarias), referidos a un conjunto numeroso de personas, animales, objeto, etc.; necesarios para el estudio de un fenómeno de nuestra esfera de interés científico, o de toma de decisiones.
- **Organizar y resumir** los datos masivos recogidos.
- **Describir y analizar** a las personas, animales u objetos observados, mediante los datos organizados y resumidos.
- **Realizar inferencias** sobre la población de la que provienen los datos recogidos, cuando estos se originan en procedimientos muestrales.
- **Obtener conocimientos e información** sobre el fenómeno en estudio, a partir de **interpretar** los resultados del análisis estadístico**.**

3. Problema de Trabajo e Investigación Estadística

A menudo y cada vez con mayor frecuencia, sea como profesionales, como investigadores, como administradores, como personas de negocios, como docentes o como simples ciudadanos; deseamos *conocer en la forma mas completa y convincente* posible, el estado o el comportamiento de algún aspecto de la realidad que nos rodea. Por ejemplo:

- Como administradores públicos necesitamos describir la situación del sistema de salud de la provincia, con el fin de diseñar políticas (tomar decisiones) para mejorarlo.
- Como investigadores de la economía regional, deseamos explicar la evolución que han tenido la producción y los precios del tabaco en los últimos años, y pronosticar sus comportamientos hacia el futuro.
- Como empresarios de la actividad turística, necesitamos conocer el perfil de los grupos turísticos que visitan el Parque Nacional Iguazú para elaborar estrategias de marketing a aplicar en los centros emisores mas importantes de la Argentina.
- Como docentes o directivos del sistema educativo oficial, deseamos dimensionar el fenómeno de la violencia estudiantil en el nivel medio.
- Como ciudadanos deseamos calificar a nuestros gobernantes y su gestión de gobierno.
- Como científicos sociales nos proponemos conocer la situación laboral de la mujer en nuestro país y, de este modo, contrastar ciertas proposiciones (hipótesis) que nos formulamos sobre el tema.
- Como gerentes de una empresa pública, necesitamos explicar en todas sus dimensiones el fenómeno del ausentismo de los funcionarios, con el fin de tomar decisiones al respecto.

En fin, los planteos pueden ser muy variados y estar relacionados con las más diversas esferas de las ciencias y de la vida cotidiana del hombre de nuestros días.

A este tipo de cuestiones las consideramos un ***problema de trabajo*** (***problema del entorno real*** o simplemente *problema*) porque **se originan en *preguntas*** (explícitas o implícitas) que nos formulamos **sobre el tema**. Preguntas que llevarán a la **búsqueda de evidencias *consistentes*** *y precisas* que permitan encontrar las mejores respuestas. Esto es, que motivarán la necesidad de *investigar* sobre el tema.

La investigación basada en métodos estadísticos debe ser previamente diseñada por el investigador; ajustándose a principios, conceptos y procedimientos plenamente reconocidos y aceptados para tal fin: *la* ***metodología de investigación cuantitativa****.*

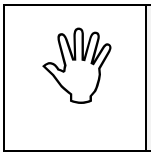
El ***diseño metodológico*** de una investigación particular podrá ser más o menos complejo, dependiendo ello de la *complejidad del fenó*meno en estudio, del carácter de *los resultados buscados* y de las *condiciones prácticas* bajo las que se llevará a cabo, entre otras razones.

Lo cierto es que todo trabajo de estadística aplicada debe, necesariamente, responder a cierto *diseño previo* (aunque más no fuere, simple y elemental), el que deberá ser convenientemente formalizado y explicitado.

Un buen **diseño metodológico** de la investigación (y del consecuente plan de acción para llevarla a cabo) **es de extrema importancia para**:

- ***orientar*** correctamente la ***construcción y obtención de los datos*** apropiados al problema y a la solución buscada,
- ***asignar validez*** a los resultados que se obtengan de los datos recogidos,
- ***optimizar los esfuerzos*** de todo tipo que se dediquen al trabajo,
- ***valorar las conclusiones*** de una investigación.

Los temas metodológicos de una investigación cuantitativa escapan a los alcances del curso [3]. Sin embargo, presentaremos en los apartados siguientes algunos conceptos que son necesarios para facilitar la comprensión de la estadística, desde el enfoque que proponemos.



Actividad Nº 1

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

3.1. Las preguntas de investigación

Toda investigación surge y es guiada por una o varias *preguntas generales* (explícitas o implícitas) o supuestos que el investigador formula sobre las cuestiones centrales de su problema de trabajo. El valor de estas preguntas (*preguntas de investigación*) es el de acotar el tema de trabajo, esbozar el objeto de estudio y orientar la estrategia de abordaje del tema.

Consideremos ahora como ejemplo el artículo que Ud. acaba de leer. Aunque en el texto no se expresan explícitamente los propósitos (interrogantes o hipótesis) que guiaron el trabajo, podemos imaginar algunas preguntas e hipótesis que formularon los investigadores. Por ejemplo las siguientes:

- ✓ ¿Qué dimensión tiene el mercado de usuarios de Internet en Argentina?
- ✓ ¿Se ha expandido este mercado en los últimos años o se ha mantenido relativamente estable?
- ✓ Internet es una herramienta mayormente utilizada por adolescentes y jóvenes, con fines recreativos y educativos.

3.2. Las preguntas estadísticas

Cada una de estas preguntas generales, a su vez, derivará en otras preguntas más específicas que tenderán a expresar el problema en términos numéricos concretos. Por ejemplo, algunas podrían ser:

- ✓ ¿Cuántos son los usuarios efectivos de Internet en nuestro país?
- ✓ ¿En qué medida ha crecido el número de usuarios en los últimos años?
- ✓ ¿Qué edad tienen en general, los usuarios de Internet?, ¿cuál es la edad más frecuente?
- ✓ ¿Son las mujeres y los adolescentes los que más utilizan el servicio?
- ✓ ¿cuáles son los fines más difundidos entre los usuarios de la red?
- ✓ ¿Predominan los usuarios de un determinado nivel socioeconómico?
- ✓ ¿Con qué intensidad (cantidad de horas diarias) se utiliza el servicio?

4. Los Datos

Es evidente que la secuencia ***Problema*** → ***Investigación Estadística*** → ***Respuesta*** supone la presencia de un elemento sustancial: *"los datos"*.

[3] Para aquellos que deseen profundizar las cuestiones metodológicas, les sugerimos la lectura de: BARANGER, D.: "Construcción y Análisis de Datos", Editorial Universitaria UNaM, Posadas 2000.

El *problema* nos coloca ante la *necesidad* de reunir *indicios/evidencias* (datos) suficientemente *capaces de informar* sobre los aspectos del contexto que nos preocupan.

Mediante el diseño y la práctica de la *investigación* se resolverá a *qué datos recurrir, cómo obtenerlos y cómo utilizarlos* apropiadamente.

Finalmente, los resultados y **conclusiones del *análisis* que se realice sobre los datos**, **aportarán la *información y respuestas* al problema** planteado. Entonces:

¿Qué es un dato?

Es el **registro** (numérico o no) que se obtiene como resultado **de observar** cierta **característica** de interés en **un "individuo"** (persona, animal, cosa o entidad de naturaleza abstracta) que constituye el objeto de estudio.

En este concepto se resumen las siguientes ideas centrales:

IMPORTANTE

el dato supone*:*

- ➔ **un "individuo"** que ha sido **observado/medido** en cierta **característica** de interés;
- ➔ esta ***observación/medición*** → se realiza mediante criterios e instrumentos previamente determinados;
- ➔ el **dato** se materializa → en el **registro de la medición** realizada.

Un dato cobra *significado* por el *"**individuo**"* al que se remite, por la ***característica*** de ese "individuo" que representa y por *la **forma*** en que esa característica ha sido medida. Es decir, **un dato reproduce información si y solo si se expresa en relación con su contexto**.

Por ejemplo: el número *"36"* y la palabra "*media*", por sí solas no aportan información relevante. En cambio, si las relacionamos con el contexto en el que se inscriben, aclarando que se trata de la *"edad en años cumplidos"* de un *"usuario argentino de Internet"*, quien pertenece a la *"clase socioeconómica" "media";* pasan a representar una buena información sobre el "individuo" observado.

Unidad de Análisis / Unidad de observación / Elemento / Individuo

Es la persona, animal, cosa o entidad de naturaleza abstracta, sometido a la observación/medición y a la cual harán referencia los datos.

En cada tema de estudio particular, la unidad de análisis tendrá una entidad específica. Por ejemplo:

- "**persona** residente en la República Argentina que en el año 2001 que es **usuaria** del servicio de Internet",
- "**establecimiento hotelero** de la ciudad de Puerto Iguazú".
- "**mercado** misionero del tabaco".

Utilizaremos indistintamente los términos ***"individuo"*** o ***"elemento"*** para referirnos en forma genérica a las ***unidades de análisis*** de la investigación, cualquiera sea su naturaleza. Así entonces, un árbol de la ciudad de Bs. As. es un "individuo", como también lo es un usuario de Internet entrevistado en el G. Bs. As, un establecimiento carcelario de la Patagonia o un turista encuestado en Puerto Iguazú.

POBLACIÓN EN ESTUDIO

Es el conjunto de **todas las unidades de análisis** que serán consideradas en la investigación.

Por ejemplo:

- "**todos** los **usuarios** del servicio de Internet en la Argentina, en el año 2001",
- "**todos** los **hoteles, residenciales, hosterías, etc.** existentes en la ciudad de Puerto Iguazú (Misiones), en el mes de julio de 2001".

La **población** en estudio **se define por la *naturaleza* de los elementos** que la forman, por el ***espacio geográfico*** en el que se ubican los elementos y el **período de *tiempo*** que se toma como referencia.

IMPORTANTE

En todo trabajo estadístico es de extrema importancia una precisa definición de la unidad de análisis y la población en estudio, dado que los datos y conclusiones que de ellos se obtengan, remitirán a esos individuos, en el espacio y tiempo definidos.

5. Las Variables

Toda característica de los individuos que es relevante en una investigación, sin dudas variará a lo largo de la población en estudio. La *edad* de los usuarios de Internet varía de uno a otro, lo mismo que la situación ocupacional de cada uno de ellos o la cantidad de horas diarias que cada usuario dedica a estar conectado en la red.

Variable

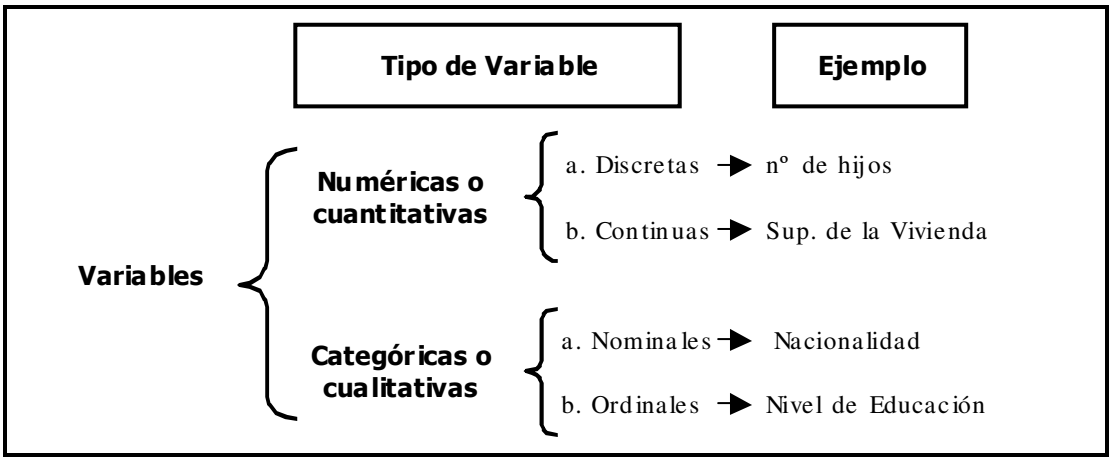
Denominaremos ***variable en estudio*** o simplemente *variable,* a **toda característica que será observada/medida en los individuos de la población en estudio.**

Vemos en nuestro ejemplo que fueron varias y muy diferentes las variables que se utilizaron para describir a los usuarios de Internet. Cada uno de las personas entrevistadas fue observada en características como las siguientes:

- ✓ edad (en años cumplidos),
- ✓ sexo,
- ✓ lugar de residencia,
- ✓ situación ocupacional,
- ✓ nivel socioeconómico,
- ✓ frecuencia semanal de conexión a la red,
- ✓ cantidad de horas de uso de la red,
- ✓ lugar de donde se conecta a la red,
- ✓ etc.

Algunas de estas variables se expresan como una *cantidad numérica* atribuible a cada individuo observado: la *edad*, la *frecuencia semanal de conexión*, la *cantidad de horas de uso*. **Otras en cambio**, **expresan cierto *atributo*** del individuo observado: el *sexo* de la persona, la *situación ocupacional* del individuo, el *lugar desde donde se conecta a la red*, etc., etc., etc.

En el esquema siguiente presentamos la forma en que se clasifican las variables según como se expresen sus datos (cantidades numéricas o atributos) y a su vez la sub-clasificación que se puede hacer para cada tipo de variables:



Variables Numéricas o Cuantitativas

Denotan una **cantidad** del individuo observado y sus datos se expresan en números (diámetro del tallo del árbol, antigüedad como usuario de la red, ingreso del grupo familiar, etc.).

Discretas	Continuas
• sus datos solamente pueden expresarse mediante **números enteros;** • generalmente son el resultado de la enumeración o el conteo de ciertos elementos en la unidad de observación.	• sus datos se expresan en números no enteros (números reales); • generalmente son el resultado de mediciones con unidades de medida preestablecidas como: kilowatios hora, centímetros, kilogramos, dólares, minutos, etc.
Por ejemplo: número de personas que son miembros del hogar, número de sucursales que integran la cadena de una firma, cantidad de árboles implantados en una manzana, etc.	**Por ejemplo**: la estatura (que puede ser de 1,874 m), el tiempo de conexión a Internet (que puede ser 1,25 horas), etc.

Variables Categóricas o Cualitativas

Denotan una cualidad del individuo, y sus datos se expresan como una **categoría** predefinida del atributo observado (la cualidad "sexo" admite las categorías varón-mujer, la variable "lugar de residencia" puede expresarse mediante las categorías Posadas-interior de la provincia-Otras provincias-Otros países, la variable "nivel socioeconómico de los usuarios de Internet" se expresa mediante las categorías alta-media alta-media-media baja-baja).

Nominales	Ordinales
• sus datos se expresan con categorías que únicamente permiten clasificar a los individuos, sin establecer ningún tipo de orden o jerarquía entre ellos.	• sus datos se expresan con categorías que además de clasificar a los individuos, permiten establecer un orden entre ellos, aunque sin establecer "distancias" exactas entre las diferentes categorías.
Por ejemplo: las categorías varón-mujer de la variable sexo, las categorías católico-protestante-luterano-evangelista-etc. de la variable religión, las categorías Oberá-Eldorado- Apóstoles-El Soberbio- etc, de la variable lugar de residencia. En ninguno de estos casos se puede establecer una jerarquía entre ellos, por la categoría que detenta cada uno de ellos.	**Por ejemplo**: la variable nivel socioeconómico de los usuarios de Internet, cuyas categorías son "alta", "media alta", "media", "media baja" y "baja". También puede ser la variable estado de salud de un paciente si se lo clasifica en "muy bueno", "bueno", "regular", "grave", "muy grave". **NOTA**: En estos ejemplos los individuos pueden ser ordenados (en forma ascendente o descendente) según la categoría de la variable en que se ubica cada uno de ellos, pero no sabemos, exactamente, cuánto peor es el estado "grave" con respecto al "regular".

Obsérvese que la **variable** denota una **característica** observable del "individuo" en estudio (nivel socioeconómico, estado de salud, ingreso del grupo familiar mensual, estatura). Y **cada variable admite diferentes "valores"** (números o categorías) posibles de ser observados en las unidades de análisis. Por ejemplo: para la variable nivel socioeconómico se han definido como **posibles valores** a las categorías "alta", "media alta", "media", "media baja" y "baja". En cambio, la variable ingreso familiar tendrá como valores posibles a **números** comprendidos en el rango que va desde el ingreso más bajo posible al más alto de la población.

IMPORTANTE

En consecuencia, **la distinción de los datos (y las variables)** según su tipo (cuantitativos, categóricos nominales u ordinales) es extremadamente importante **para el uso correcto de las herramientas estadísticas.** Como veremos más adelante, algunas herramientas solamente son aplicables a ciertos tipos de datos y a otros no.

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.

6. Conjunto De Datos: Datos Estadísticos

Los métodos y técnicas de la estadística no son aplicables a observaciones individuales. Requieren de conjuntos "suficientemente grandes" de datos, recogidos mediante la observación sistemática de un número "suficientemente grande" de individuos.

En la masividad de los datos, la estadística se ocupa de estudiar las variaciones entre ellos para encontrar, describir, explicar e inducir; tendencias y regularidades de los individuos.

En resumen, el buen uso de las herramientas estadísticas supone un conjunto numeroso de datos (numéricos o categóricos): **"datos estadísticos"**

Los datos estadísticos de una variable en estudio se pueden obtener:

Observando al **mismo tiempo** y **bajo las mismas condiciones**, a un **gran número de individuos diferentes**

→ **observación transversal** → conjunto o serie de datos transversales

Observando a un **mismo individuo**, bajo las mismas condiciones, a lo largo de **numerosos momentos diferentes de tiempo:**

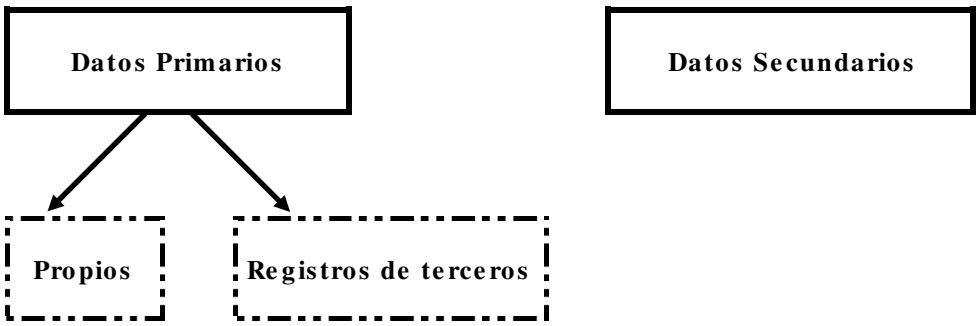
→ **observación longitudinal** → datos longitudinales o serie temporal/ cronológica

En el estudio de los usuarios de Internet se utilizaron ambos tipos de datos:
Por un lado se observó la variable "cantidad de usuarios" en la unidad de análisis "mercado de Internet en Argentina", a lo largo de diferentes períodos anuales consecutivos, dando lugar a una serie cronológica de cinco datos (Gráfico: "Usuarios de Internet en la Argentina").

Por el otro, se observaron transversalmente a 1.400 usuarios, en diferentes variables de interés (edad, sexo, nivel socioeconómico, situación ocupacional, lugar de residencia, etc.), dando lugar a un conjunto de 1.400 datos transversales por cada una de ellas. Es decir, que esta observación resultó en tantos conjuntos de 1.400 datos cada uno, como variables diferentes fueron observadas de esta manera.

7. Fuentes de Datos

Los datos a emplear en una investigación pueden provenir de diferentes fuentes u orígenes y encontrarse en diferentes estados de elaboración. Podemos considerar entonces:



Datos Primarios

Son aquellos que se encuentran en la forma original en que fueron registrados ("datos brutos"), sin haber sufrido ningún tipo de tratamiento o elaboración posterior.

Este tipo de datos, según su **fuente,** pueden ser:

Propios	Registros de terceros
Cuando fueron diseñados con el fin específico de la investigación y expresamente recolectados por quien los utilizará.	Son datos primarios que se recopilan con fines ajenos a los de la investigación, pero que por su definición y procedimientos de captación se ajustan a nuestras necesidades. Generalmente se trata de datos que se registran con fines administrativos.
Por ejemplo: los datos recogidos mediante la encuesta realizada a los usuarios de Internet.	**Por ejemplo**: los datos que sobre sus clientes llevan los diferentes servidores de la red. Otro ejemplo: los datos que se registran en el legajo de cada cliente de una empresa o de cada alumno de la UNaM.

Datos Secundarios

Son aquellos que fueron producidos (diseñados y recopilados) por terceros, con un fin ajeno al de la investigación y que ya han sido sometidos a alguna forma de elaboración posterior. En consecuencia, estos datos siempre se originan en terceras fuentes.

Por ejemplo: los datos que publican las oficinas de estadística de instituciones públicas, de las empresas, etc.

8. Investigación por Censo y por Muestra

La población en estudio puede ser observada (transversalmente) de dos maneras:

Enumeración completa	Por muestra
Consiste en observar las variables de estudio en ***todos los individuos*** **que forman la población**. Usualmente se denomina *"censo"* a esta forma de recopilación de datos.	Consiste en ***seleccionar una parte*** de la población (*la* ***muestra***), ***observar a los individuos elegidos*** en las variables en estudio*, elaborar conclusiones* a partir de los datos *de la muestra* y, cuando esto es posible, *generalizar* estas conclusiones al conjunto de *toda la población* de origen (***inferir*** conclusiones sobre la población a partir de los resultados muestrales).

El estudio de las *"2.000.000 de personas conectadas a Internet"* se basó en una **muestra** de solo *"1.400 casos"* efectivamente observados. Sin embargo, **las conclusiones** extraídas del análisis de estos casos **se atribuyen a toda la población**. Por ejemplo:

- ✓ "el 50 por ciento *de los usuarios de la Red* tiene más de 35 años",
- ✓ "4 de cada 10 *usuarios* son mujeres",
- ✓ "sólo el 3 por ciento *de los navegantes* está desocupado".

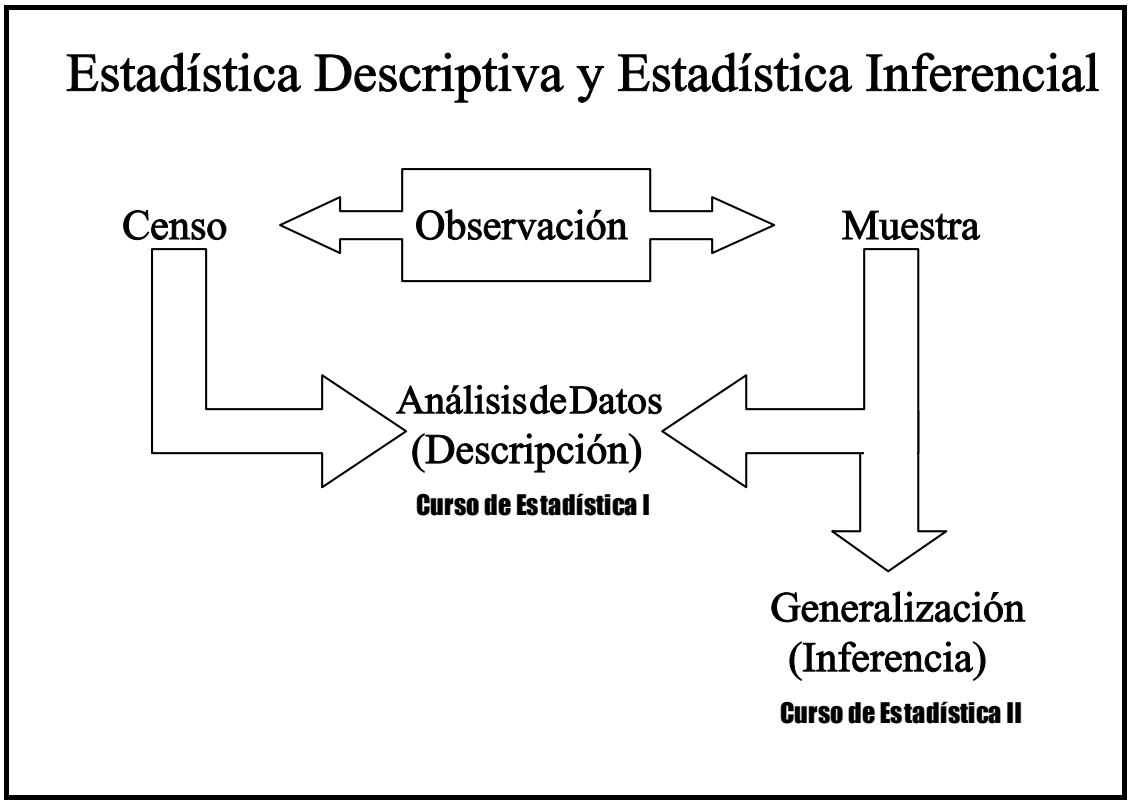
Los procedimientos de observación por muestras y de generalización *(inferencia)* de los resultados, nos llevan a ciertas preguntas clave como las siguientes:

- ✓ ¿Cuáles son los argumentos para realizar un estudio por enumeración completa o por muestra?
- ✓ ¿Cómo elegir una muestra de manera que reproduzca (sea "representativa") lo mejor posible a la población en estudio?
- ✓ ¿Qué mecanismos o procedimientos se deben aplicar para generalizar correctamente las conclusiones de la muestra?
- ✓ ¿Qué exactitud o confiabilidad pueden tener estas generalizaciones?

Es decir, la **investigación basada en muestras** nos coloca frente a **dos temas centrales** de la Estadística:

Muestreo	Estadística inductiva o inferencial
Que trata sobre los procedimientos y técnicas para seleccionar muestras de una población.	Que aporta los conocimientos para realizar generalizaciones (inferencias) confiables de los resultados muestrales.

Ambos temas serán tratados en el curso más avanzado de Estadística II. Hasta tanto, Ud. debe tener presente que, a pesar de lo extremadamente relevante que significa distinguir una investigación basada en "censos" de aquellas basadas en "muestras", ***las técnicas y herramientas para la descripción inicial de los datos (Estadística Descriptiva) que presentaremos en este curso, son comunes a ambas situaciones de trabajo*.**



Actividad Nº 3

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad.

9. ¿Qué Hemos Visto?

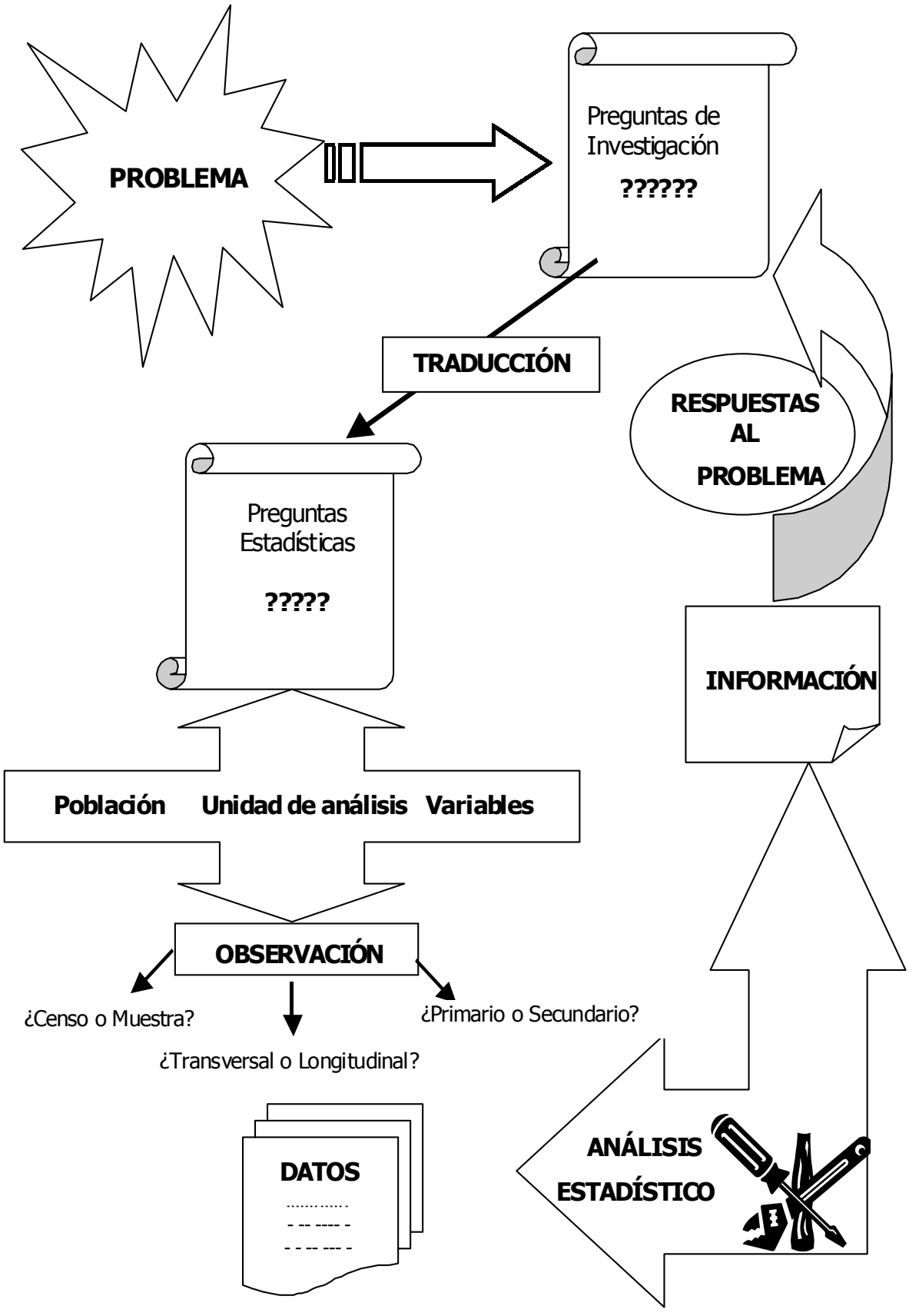
El propósito de esta unidad es introducir al lector interesado en temas de investigación estadística, en los conceptos básicos que resulta imprescindible manejar cuando se utiliza esta disciplina.

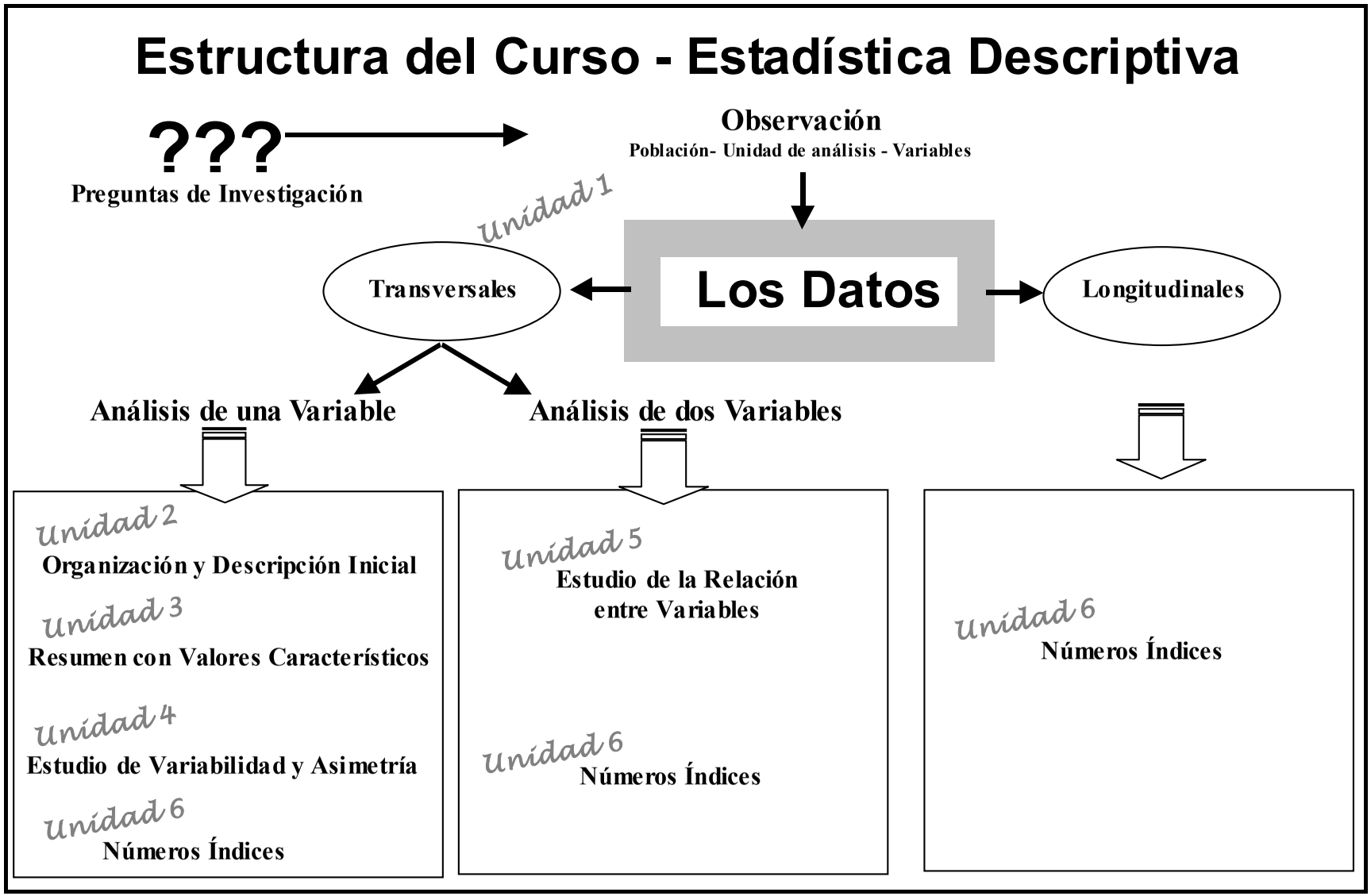
Así, inscribiendo el uso de la estadística en un proceso de investigación o toma de decisiones, se presentaron -en el marco de la producción de información- aquellos elementos teóricos recurrentes en cualquier situación de trabajo que implique el análisis estadístico. De esta manera, se formalizan en la presentación los conceptos de: dato, unidad de análisis, población y variable.

Dado que la posibilidad de utilizar cualquiera de las técnicas estadísticas, está condicionada por el tipo de variables que se quieren analizar, se puso especial atención en la clasificación de variables que resultan de las diferentes formas en que se registran los datos. Así hemos distinguido variable cualitativas y cuantitativas (con sus respectivas sub-clasificaciones), diferenciando además, las observaciones transversales y longitudinales.

Finalmente, se realizó una distinción de los datos según la fuente de la cual se obtienen (Primarios y Secundarios) y el tipo de investigación que realizamos según se observan todas las unidades de análisis de la población (censo) o una parte de ella (muestra).

LA ESTADÍSTICA EN EL PROCESO DE INVESTIGACIÓN




Estructura del Curso - Estadística Descriptiva
???
Preguntas de Investigación
Observación
Población- Unidad de análisis - Variables
Unidad 1
Los Datos
Transversales
Longitudinales
Análisis de una Variable
Análisis de dos Variables
Unidad 2
Organización y Descripción Inicial
Unidad 3
Resumen con Valores Característicos
Unidad 4
Estudio de Variabilidad y Asimetría
Unidad 6
Números Índices
Unidad 5
Estudio de la Relación entre Variables
Unidad 6
Números Índices
Unidad 6
Números Índices

Bibliografía

ANDERSON, D; SWEENEY, D.; WILLIAMS, T (1999): *Estadística para Administración y Economía*. International Thomson Editores, México. Páginas 1 a 21.

DANIEL, WAYNE (1985): *Estadística con aplicación a las ciencias sociales y a la educación,* McGraw-Hill, México.

MOORE, DAVID (1995): *Estadística Aplicada Básica*. Antoni Bosch Editor, Barcelona. Páginas: XXI a XXIV, 1 a 5 y 6 a 7 (punto 1.2)

MOOD, A. M. (1965): *Introducción a la Teoría de la Estadística*. Aguilar, Madrid (3ra. Edición).

Conceptos Centrales

- Preguntas de investigación y preguntas estadísticas
- Dato
- Unidad de análisis o "individuo"
- Población en estudio
- Variable
- Tipos de variables
- Datos transversales y longitudinales
- Datos primarios (propios y de registros) y secundarios
- Investigación por enumeración completa y por muestra
- Muestreo e inferencia estadística

Habilidades

- *Identificar* en un trabajo de investigación estadística las preguntas que lo orientaron (de investigación y estadísticas).
- *Distinguir* en una situación concreta: la población en estudio, la unidad de análisis, las variables de interés y el tipo de variables a que corresponde cada una.
- Reconocer para cualquier situación de trabajo que se presenta: si se trata de datos longitudinales o transversales, el tipo de fuente utilizada, y si corresponde a un relevamiento por muestra o por enumeración completa.

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 descriptas 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

Variable nº 2 — **Variable nº "p"**

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

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, ...x_i, ...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_1, z_2, z_3, ... z_{25}.....z_{139}), representará al conjunto (3,3,1,.....,2.....3) de datos de la matriz correspondiente a las carreras cursadas.

Actividad Nº 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***.

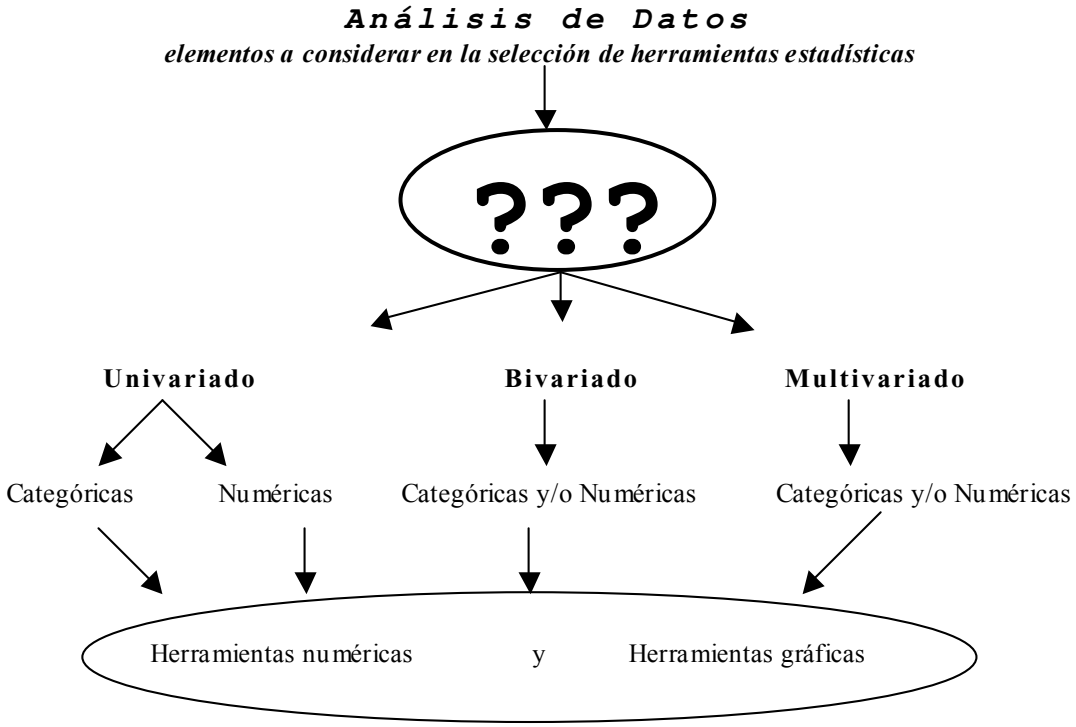
Otro aspecto a tener en cuenta al considerar la herramienta apropiada para el análisis [1] es ***el tipo de variable*** con el que se está trabajando: cuantitativas, o cualitativas (ordinales o nominales).

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 análisis [2], 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.

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

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 tabla [4] como la siguiente:

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

Nombre de la variable

Valores de la variable

SEXO	***nº de estudiantes***
Varón	30
Mujer	109
Total	**139**

Cantidad de varones observados

Frecuencias absolutas

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.

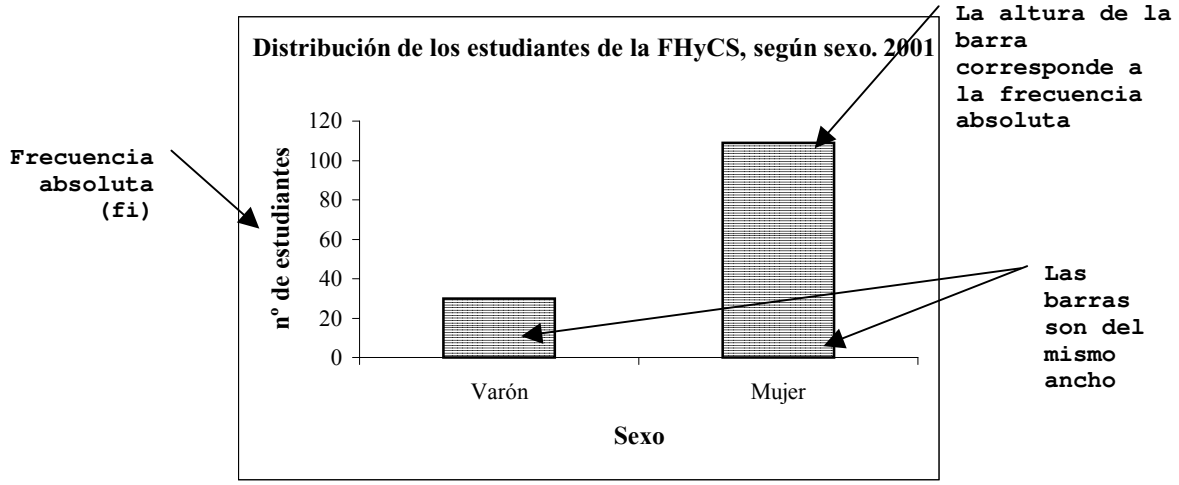
[4] Es importante destacar que toda tabla se puede identificar:

- 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 **fuente** 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.

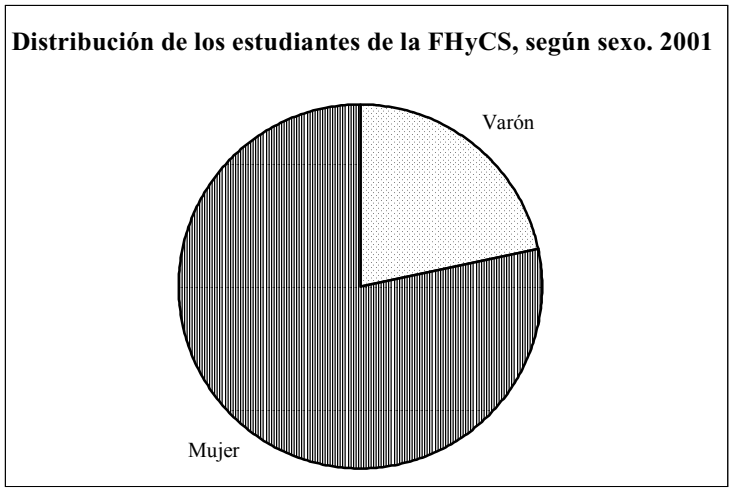
- el recurso gráfico

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 casos [5].



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 360º, 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*.*

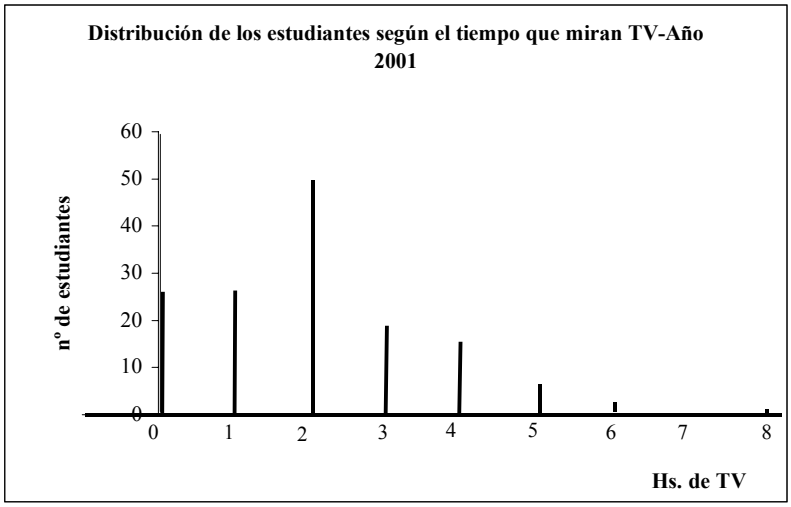
IMPORTANTE

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 Nº 3

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 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 arreglo*[6], 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

Ocho Intervalos de clase

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**

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 (Li y Ls).*
- ***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 [(Li+Ls)/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, las 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:

$$k \cong 1+3{,}3.\log n$$

donde: ***k*** es el número de clases que se quiere determinar,
n es el número total de datos y
log es el logaritmo [7] 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 $a \cong \frac{47\text{-}17}{8} = 3{,}75$ 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 partes [8]:

- *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.

Tallo → **1 4 | 7** ← hoja

[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).

El tallo puede tener tantos dígitos como sean necesarios en tanto que la hoja solo un dígito.

Diagrama de tallo-hoja para la Edad

```
 Frecuencia     Tallo -  Hoja

    (57)      1 | 77777788888888888888888888899999999999999999999999999999
    (58)      2 | 00000000111111111122222222223344455556666667777788999999999
    (17)      3 | 00111233344557788
     (4)      4 | 0147                                  Representa a
Ancho Tallo:          10                              los 8 estudiantes de 29 años
Cada hoja:        1 caso o "individuo"
```

En el diagrama anterior vemos que los tallos (1,2,3,4) están ordenados de manera vertical. A la derecha de cada uno, se disponen en forma creciente las hojas, cada una de las cuales corresponde al último dígito de cada una de las edades observadas para ese tallo.

A la izquierda de cada tallo se consignan las frecuencias correspondientes a ese tramo de edad. Así, en el tramo más largo del diagrama (el segundo) se indica que hay 58 estudiantes que registran edades entre 20 y 29 años. Podemos ver entonces que **este diagrama combina una distribución de frecuencias y un gráfico de barras horizontales**.

En términos generales, el **diagrama de tallo-hoja** presenta:

- ✓ los tallos **ordenados** en forma vertical y creciente;
- ✓ a la derecha de cada tallo, las **hojas** donde cada una **representa un dato** (salvo que se indique lo contrario al pie del diagrama);
- ✓ las **hojas ordenadas en forma creciente**, dentro de cada tallo, donde el valor que asumen corresponde al dígito siguiente al tallo.

El diagrama nos ***permite analizar la distribución*** de los individuos, y decidir límites para construir intervalos de clase que expresen de manera apropiada el comportamiento de los datos: la forma de la distribución (concentraciones, valores poco frecuentes, extremos, etc.).

A partir de la forma de la distribución que nos presenta este diagrama, podemos tomar decisiones para construir los intervalos de clase. Una distribución posible de las edades, resultaría al dividirlas en tantas clases como décadas se identifican en el diagrama tallo-hoja. Esta decisión comportaría una gran pérdida de información, así por ejemplo: se perdería aquella que supone conocer el número importante de estudiantes con 19 años (edad más frecuente que surge de una simple observación el gráfico). Entonces, se hace necesario proponer intervalos que resuman los datos sin tanta pérdida de información.

Una alternativa en este sentido sería analizar el interior de cada tramo de edades, considerando el contexto de los datos (estudiantes del primer y segundo año de la universidad) y las recomendaciones generales de la construcción de una distribución de frecuencias por intervalos de clase[9].

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

Edades más frecuentes

Intervalos de clase de diferente amplitud con una clase abierta

Edad	nº de estudiantes
17-18	28
19-20	37
21-22	20
23-26	15
27-30	17
31-35	11
36 y más	8
Total	**136**

Frecuencia máxima

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

[9] Queremos hacer notar que los cortes son subjetivos y consecuentemente la alternativa propuesta es solo una de las posibles que se pueden construir.

- ***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

Intervalos de clase de diferente amplitud

Intervalo de clase abierta

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

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

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad***.**

- el recurso gráfico

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

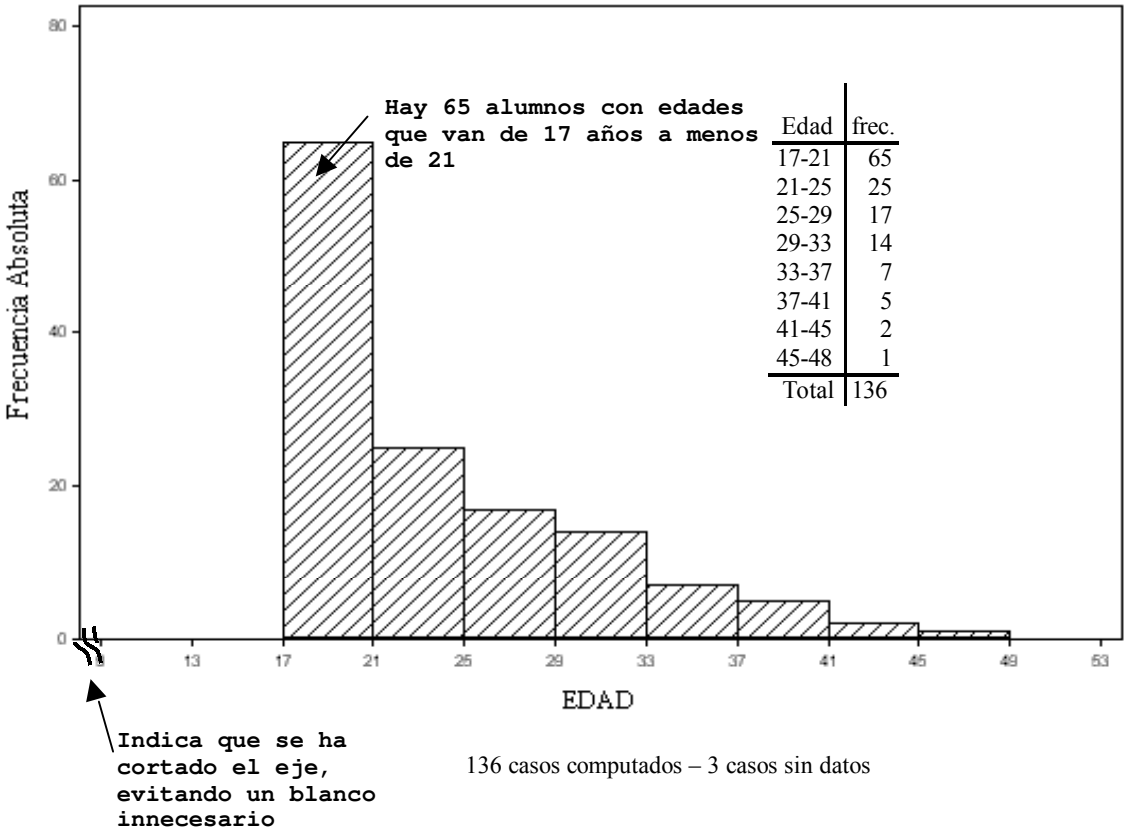
Histograma

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



Edad	frec.
17-21	65
21-25	25
25-29	17
29-33	14
33-37	7
37-41	5
41-45	2
45-48	1
Total	136

136 casos computados – 3 casos sin datos

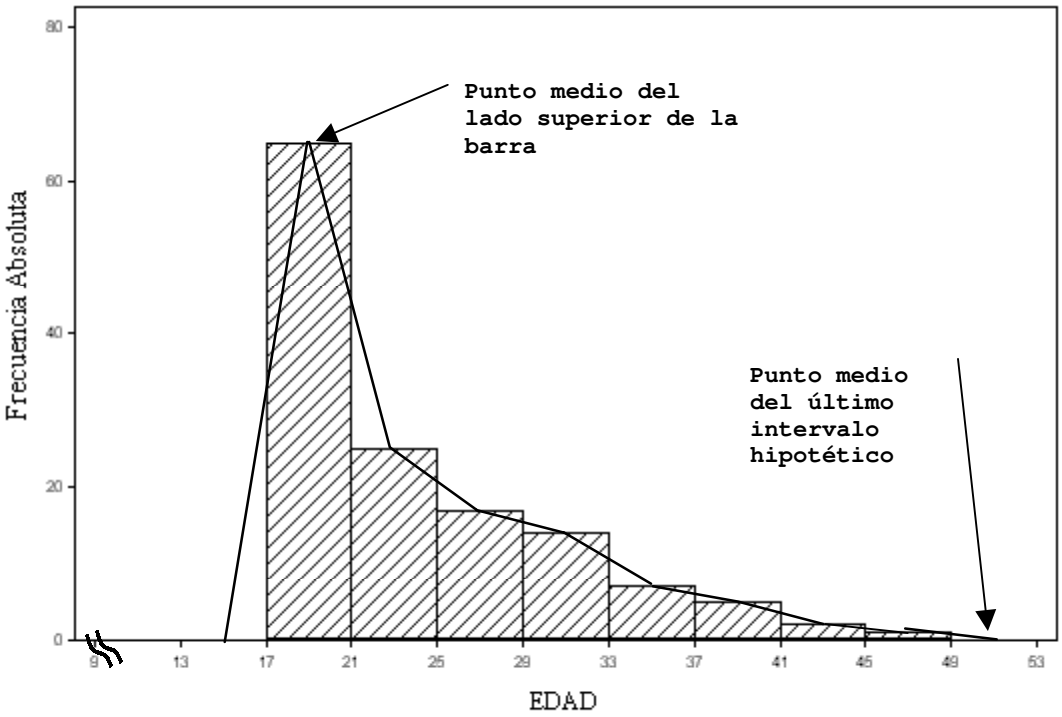
IMPORTANTE

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 finalidad [10]. 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.

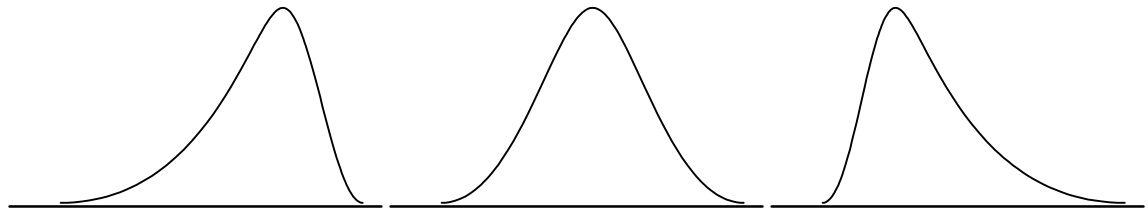
El **diagrama de tallo-hoja** que presentáramos anteriormente, también funciona como un recurso exploratorio que nos permite ***captar la forma*** de la distribución, sin perder los valores individuales que se agrupan en los distintos intervalos. De hecho, este es uno de los usos más frecuentes del diagrama 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.

Formas típicas de una distribución de frecuencias



Asimétrica a izquierda **Simétrica** **Asimétrica a derecha**

Es la "cola" de la distribución, la que califica el tipo de asimetría

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

	***Actividad Nº 6*** *Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 6 de la Guía de Actividades correspondiente a esta unidad***.**

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.

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

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**

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

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.

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

IMPORTANTE

Estas frecuencias tienen sentido únicamente para datos ***numéricos*** o datos ***categóricos 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**				

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

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

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 7 de la Guía de Actividades correspondiente a esta unidad***.**

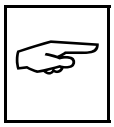
Cuando se trata de una ***variable ordinal***, el razonamiento es análogo al desarrollado para las variables numéricas. Por ejemplo en el caso de la variable Nivel de estudios del Padre la información se podría organizar en una tabla como la siguiente:

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:

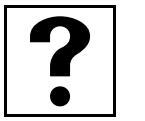


- Los 17 estudiantes cuyos padres tienen secundario incompleto, representan el 12,8%.
- Son 103 los estudiantes cuyos padres no superaron el secundario incompleto (tienen un nivel de estudios de secundario incompleto o menos). Estos representan el 77,4% del total de los estudiantes.
- Los que tienen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

Actividad Nº 8

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 8 de la Guía de Actividades correspondiente a esta unidad **.**

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:

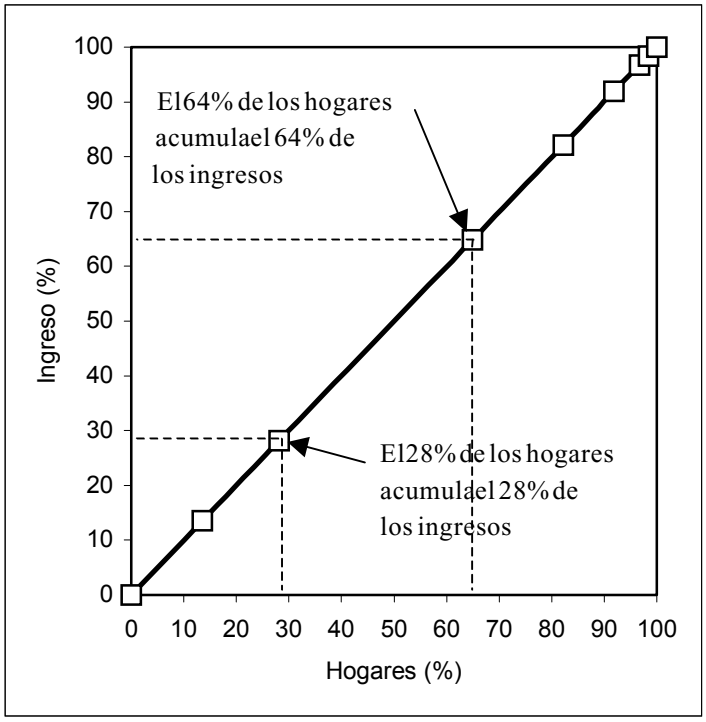
- ✓ ¿La superficie de tierra productiva de la provincia, aparece concentrada entre pocos propietarios?
- ✓ ¿Cómo se distribuye el ingreso entre los hogares de la ciudad de Posadas?
- ✓ ¿Cuál es la distribución de los 37 millones de argentinos según el tamaño de las localidades?
- ✓ etc.

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.

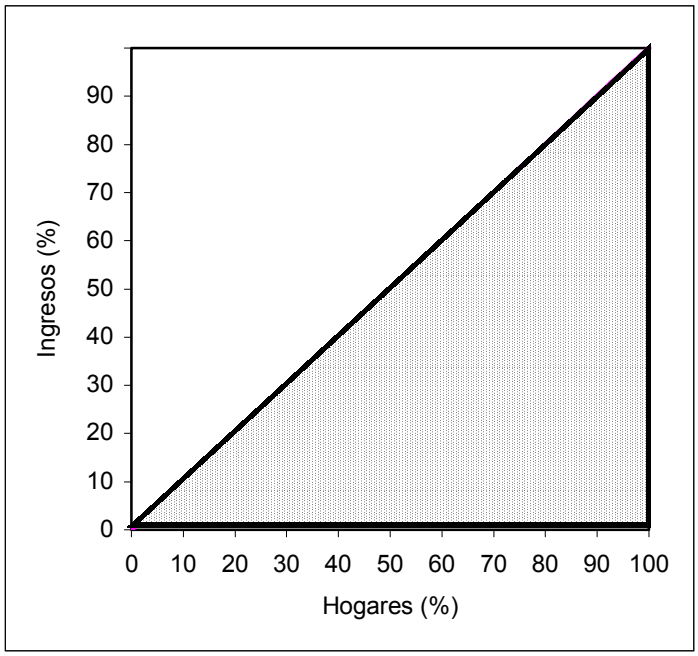
Una situación de estas características se puede representar gráficamente, utilizando un sistema de ejes cartesianos, en el que cada punto queda definido por el *porcentaje de hogares* y su correspondiente *porcentaje de ingresos*, obteniendo una gráfica como la siguiente.

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).

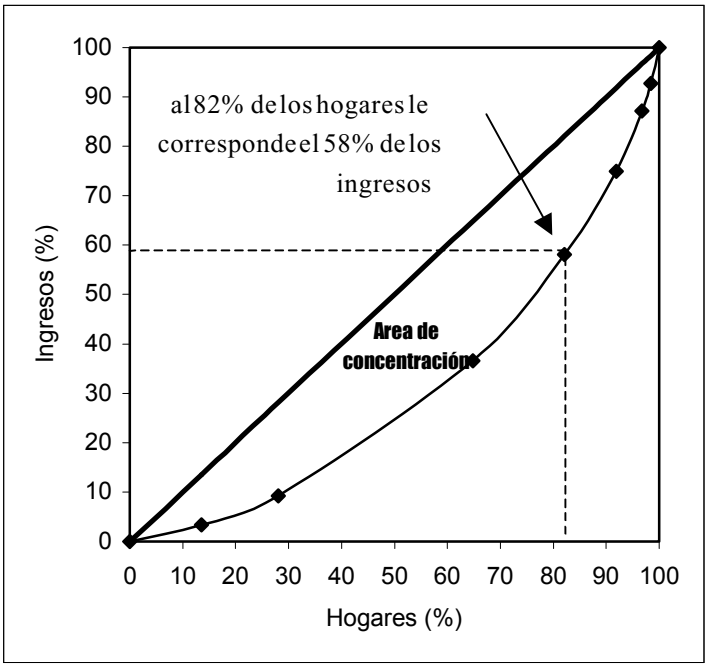
Curva de Lorenz para una situación de máxima concentración



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 co 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 x x_i) expresa el monto o volumen total de ingresos percibido por los hogares de esa clase. Así por ejemplo: 450 x 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 x 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.

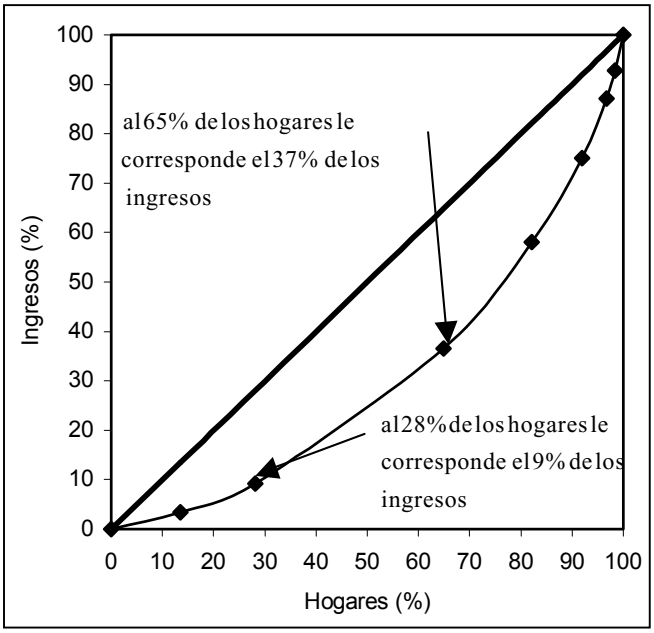
Así entonces, a manera de ejemplo, podemos observar en la fila sombreada que, los 2.736 hogares con ingresos menores que $1.243, acumulan $1.611.711; esto significa **que el 82% del total de hogares que menos ganan, participan con sólo el 58% del monto total de ingresos percibido por el conjunto de familias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



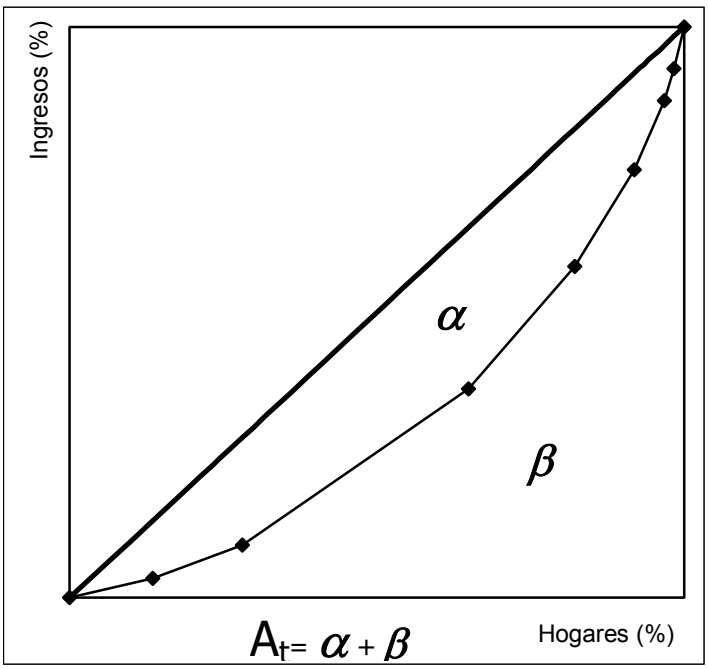
Esta gráfica tiene la ventaja de permitirnos apreciar de manera sencilla el nivel de concentración de la variable en estudio. En nuestro ejemplo, vemos que la curva define un área que está más cercana a la situación de equidistribución que a la de máxima concentración, y podríamos entonces calificarla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

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

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

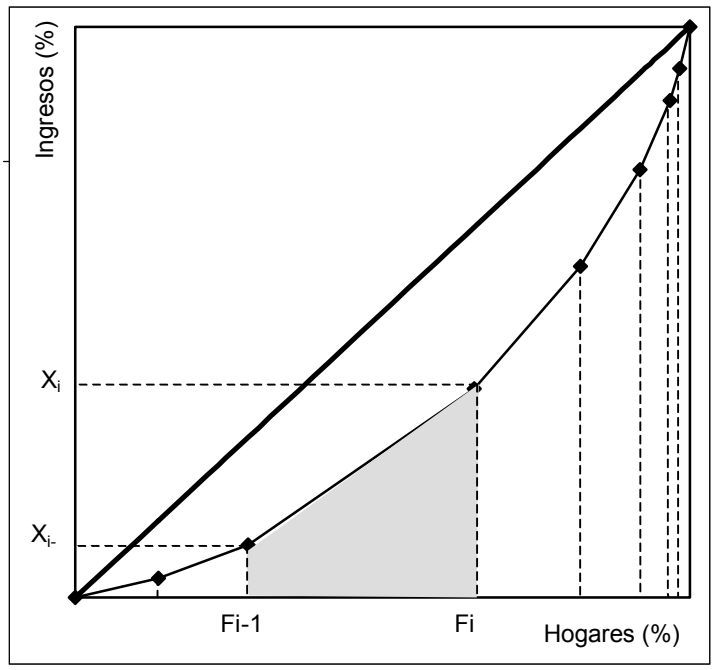
$I_G=1$ cuando se trata de una situación de **máxima concentración** ($\alpha=A_t$)

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 - β); de lo que resulta:

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

El área total A_t se determina como la mitad del área del cuadrado de lado 100; esto es 5.000. El problema se reduce a 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.

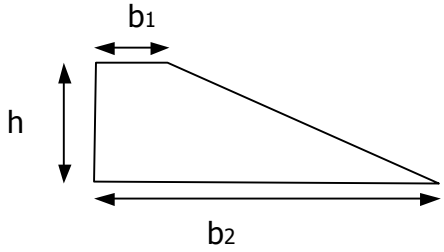
Gráfica de Lorenz: elementos para la determinación del área residual β



Recordemos que el área de un trapecio se obtiene como:

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

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:

$$\frac{(X_{i-1}+X_i)\cdot (F_i-F_{i-1})}{2}$$

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

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 esta rá dada por:

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

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

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

Luego:

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

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

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

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)	$(X_{i-1}+X_i).(F_i-F_{i-1})$ (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:

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

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 Nº 9

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 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: $I_G = 1 - \sum_{i=1}^{k} (X_{i-1}+X_i).(F_i - F_{i-1})$

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.

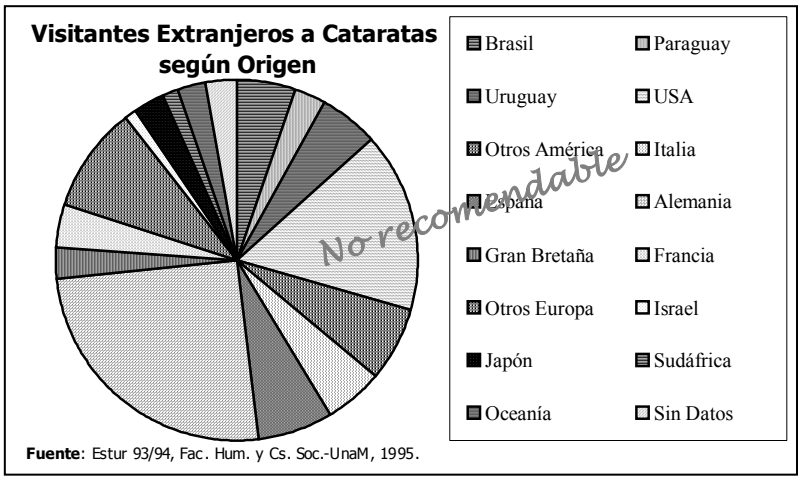
IMPORTANTE

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 construidos 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.).

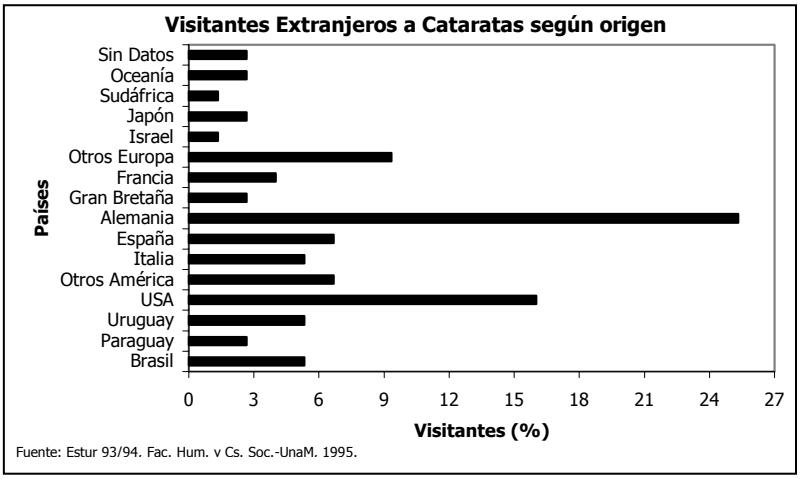
Algunos gráficos que ilustran los aspectos señalados precedentemente:



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.



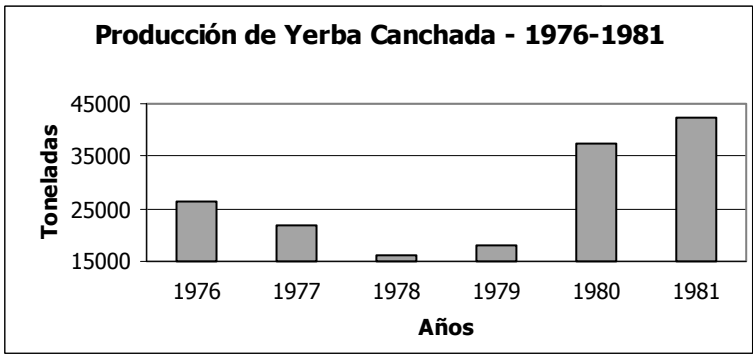
Para presentar esta misma información una alternativa es utilizar un gráfico de barras horizontales [13] como el siguiente.

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.

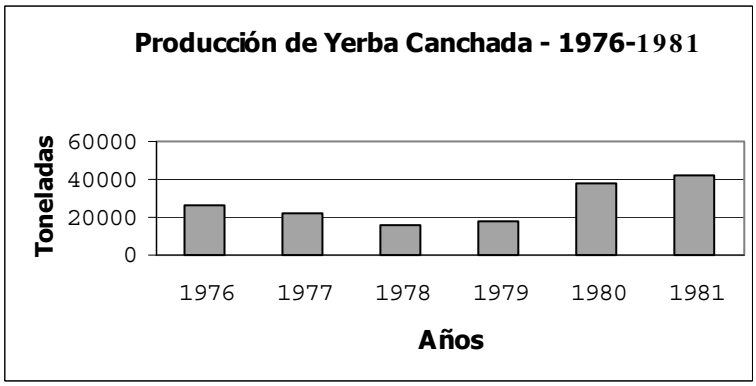
b) Modificando las escalas se pueden producir, para un mismo conjunto de datos, distorsiones en los gráficos que generan en un observador desprevenido impresiones totalmente diferentes respecto al comportamiento de los mismos. Esto obliga a ser muy cuidadoso tanto en la construcción (en el caso de quien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



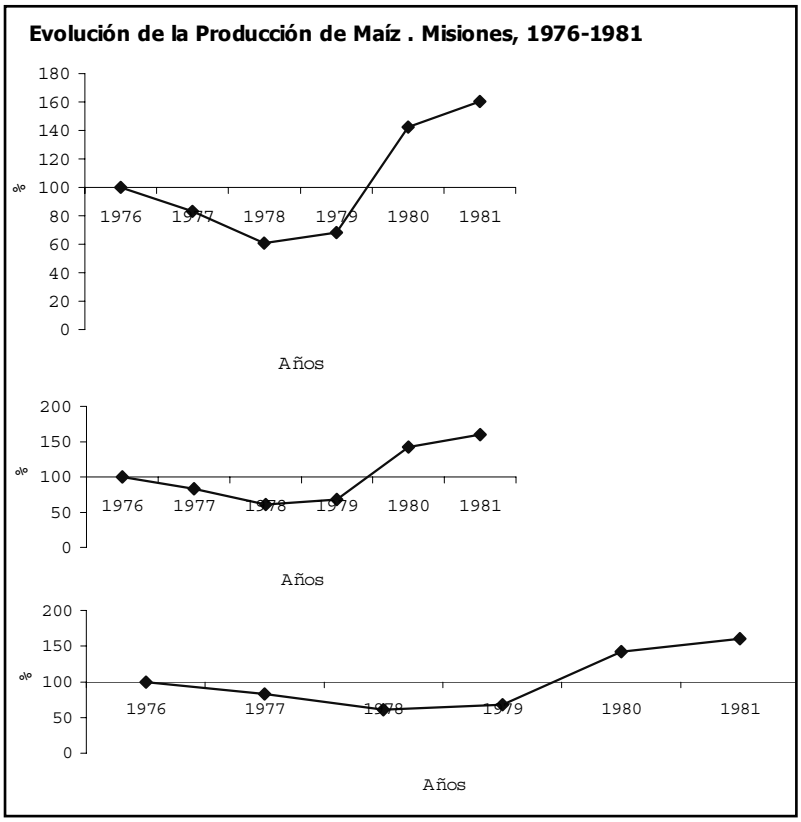
En este segundo Gráfico se muestra la escala vertical desde cero y, en consecuencia, la altura de las barras es proporcional a la producción en toneladas.

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**.

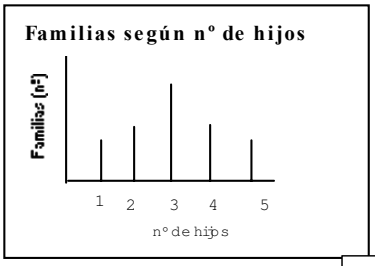
Con estos ejemplos no pretendemos agotar los casos de distorsiones que se pueden producir a la hora de utilizar el recurso gráfico, sino más bien alentar una actitud crítica cuando se construyen gráficos, y también cuando se interpretan gráficos ya construidos.

4.5. Esquema de gráficos univariados

Variable

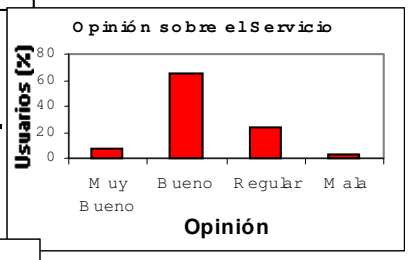
Con Ejes Cartesianos

Bastones



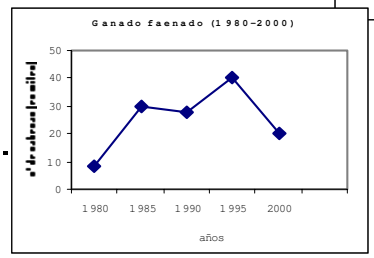
Numéricas

Barras



Categóricas

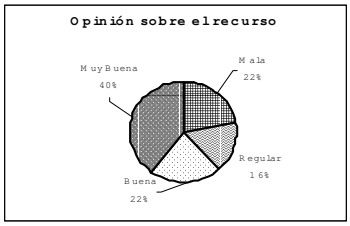
Líneas



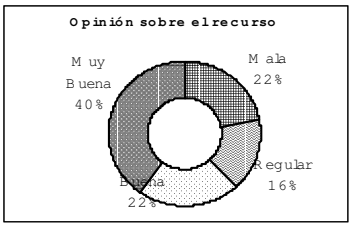
Numéricas

Sectores

Circular



Anillo

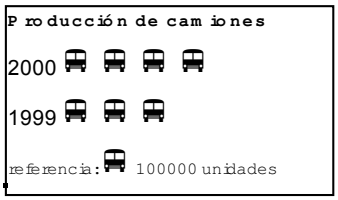


Categóricas

Otros

Mapas

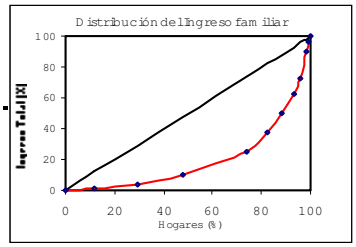
Pictograma



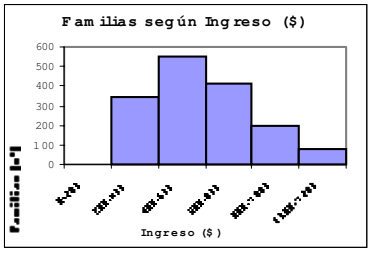
Numéricas y Categóricas

Especiales

Lorenz



Histograma



Numéricas

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.*

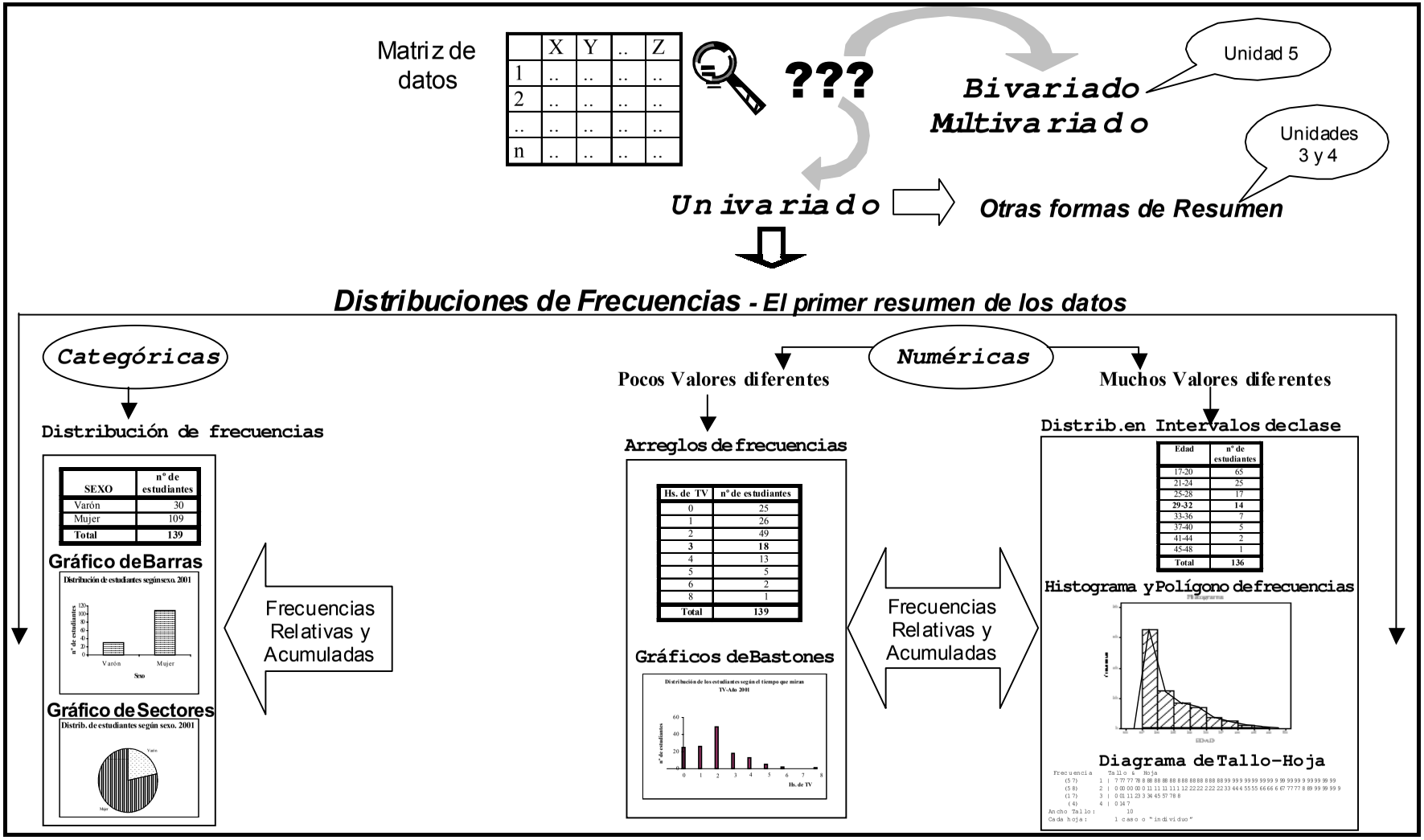
En el análisis univariado, el primer resumen de los datos son las **distribuciones de frecuencias***, para cuya construcción debemos considerar inicialmente el tipo de variable a trabajar (***numérica o categórica***).*

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.


Matriz de datos
X Y .. Z
1 .. ..
2 .. ..
.. ..
n .. ..
???
Bivariado
Multivariado
Unidad 5
Univariado
Otras formas de Resumen
Unidades 3 y 4
Distribuciones de Frecuencias - El primer resumen de los datos
Categóricas
Distribución de frecuencias
SEXO | n° de estudiantes
Varón | 30
Mujer | 109
Total | 139
Gráfico de Barras
Gráfico de Sectores
Frecuencias Relativas y Acumuladas
Numéricas
Pocos Valores diferentes
Arreglos de frecuencias
Hs. de TV | n° de estudiantes
0 | 25
1 | 26
2 | 49
3 | 18
4 | 13
5 | 5
6 | 2
8 | 1
Total | 139
Gráficos de Bastones
Frecuencias Relativas y Acumuladas
Muchos Valores diferentes
Distrib.en Intervalos de clase
Edad | n° 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
Histograma y Polígono de frecuencias
Diagrama de Tallo-Hoja

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.

UNIDAD 3: LOS VALORES QUE CARACTERIZAN AL CONJUNTO DE DATOS

1. ¿Por qué son Necesarios?

En el Capítulo anterior hemos analizado herramientas estadísticas elementales que permiten resumir grandes masas (conjuntos) de datos primarios (categóricos o numéricos), convirtiéndolos en expresiones comprensibles y operables como lo son las tablas y los gráficos de las distribuciones de frecuencias. Además, hemos introducido algunas medidas simples que ayudan a la interpretación de tales resúmenes: frecuencias relativas y acumuladas.

La correcta utilización de esas herramientas descriptivas nos permitirá elaborar ciertas conclusiones sobre los "individuos" observados. Por ejemplo, analizando las tablas y gráficos del capítulo anterior, en las que se resumen diferentes grupos de datos relativos a los estudiantes del Curso de Estadística, podríamos afirmar entre otras cosas que[1]:

- ✓ *el 13% de los alumnos dedica 3 horas diarias a mirar TV*,
- ✓ *109 alumnos del curso son mujeres*,
- ✓ *90 estudiantes tienen 24 años o menos*.

A menudo, el análisis y descripción que deseamos realizar requiere de medidas capaces de ***resumir*** aún más al conjunto de datos, expresándolo en ***un solo "valor"*** (número o categoría de la variable en estudio) que lo ***represente***. Expresiones de síntesis como las siguientes facilitarán la comprensión global del fenómeno que expresan los datos que se analizan y, además, harían más sencilla la comparación entre distintas series de datos:

- ✓ *"Los grupos turísticos registran una estadía **promedio de 3 noches** en Puerto Iguazú"*.
- ✓ *"Es llamativo que el 50 por ciento de los usuarios de la red tiene **más de 50 años**"*.
- ✓ *"El **fresno** es el árbol que **más abunda** en la ciudad de Buenos Aires, con más del 40% del total de ejemplares*".

En los tres ejemplos, cada uno de los conjuntos de datos analizados (pernoctes en Puerto Iguazú, edad de los usuarios de Internet y variedad de los árboles de la CBA), queda ***resumido y expresado*** por un único valor de la variable en estudio: *"3 noches"*, *"50 años"* y *"fresno"*. Estas son las medidas estadísticas denominadas *"de tendencia central"*.

IMPORTANTE

Es oportuno reiterar que las medidas presentadas en el Capítulo anterior (frecuencias absolutas, relativas, etc.) y las que veremos en esta unidad, se **emplean de igual modo y con idénticos fines** de **resumen y descripción**, ya sea cuando se trata de **datos muestrales** como de **datos poblacionales** ("censales"). Es decir que, tanto los **conceptos** como la **forma de calcularlas** y la **interpretació**n de los resultados, son los mismos en ambas situaciones de trabajo.

En Capítulos posteriores distinguiremos el significado que adquieren estas medidas (estadístico muestral/estimador o parámetro) según provengan de datos muestrales o poblacionales.

[1] Sugerimos que el lector identifique las medidas estadísticas utilizadas en cada una de estas afirmaciones y que, aplicándolas a los datos de los ejemplos citados, verifique que todas ellas sean correctas.

2. ¿Cuáles Son?

Las medidas de tendencia central de un conjunto de datos son valores que ***tienden a ubicarse en el centro de la distribución*** (de ahí su nombre), cuando esta reúne ciertas condiciones: es unimodal [2] y la mayor concentración de los datos (mayores frecuencias) ocurre alrededor de los valores centrales de la variable observada.

Son varias las medidas de resumen llamadas de tendencia central: las que se construyen mediante ***alguna forma*** (aritmética, geométrica, cuadrática o armónica) de ***promediar todos los datos*** del conjunto y las que se ***basan en un solo dato*** de la serie (mediana y modo). En este curso analizaremos solo las tres de uso más común:

- el promedio aritmético o *"media aritmética",*
- la *moda o modo*, y
- la *mediana.*

IMPORTANTE

A lo largo del texto iremos introduciendo la notación matemática ("fórmulas") de las herramientas estadísticas que analizaremos y, en ciertos casos, de algunas demostraciones relacionadas con ellas.

Como regla general, estas expresiones estarán a continuación del concepto estadístico que representan. Por ello, ***recomendamos firmemente*** *centrar la atención y asegurarse de* ***comprender primero el concepto,*** *luego su formalización matemática, y por último el procedimiento de cálculo.*

3. Media Aritmética

Concepto

La media aritmética $\bar{x}$ de un conjunto de datos de una **variable numérica** "X", es el resultado de **sumar todos** los valores del conjunto y **dividir esa suma** por el total **n** de observaciones que componen el conjunto [3].

Simbología: La notación usual para representar a la media aritmética es: $\bar{x}, \bar{y}, \bar{z}$, etc., dependiendo de la letra (X, Y ó Z) adoptada para simbolizar a la variable en estudio. La distinción entre letras mayúsculas ($\overline{X}$) y minúsculas ($\bar{x}$) generalmente se reserva para diferenciar una media poblacional (mayúscula) de una muestral (minúscula). En este curso ***utilizaremos única e indistintamente*** la notación $\bar{x}$, debiendo el lector tener presente la advertencia anterior.

De igual modo, las letras *n* y *N* son usualmente reconocidas para distinguir en forma simbólica al total de observaciones de una muestra (*n*) y al total de datos de una población (*N*). Utilizaremos el símbolo *n* indistintamente.

Así entonces, si tomáramos los $n = 136$ datos [4] de la variable "Y" (columna) "*edad*", registrados en la matriz *"Estudio de los Alumnos de Estadística I"* del Capítulo anterior, el ***promedio*** o ***media aritmética*** o simplemente ***"media"*** de ese conjunto de observaciones, será:

$$\bar{y} = \frac{19 + 27 + 26 + 28 + \ldots\ldots\ldots + 30}{136} = \frac{3180}{136} \cong 23,4 \text{ años}$$

`valor promedio o "media aritmética" del conjunto`

`total de datos`

[2] El concepto de distribución "unimodal" quedará debidamente aclarado en puntos posteriores de esta unidad.

[3] Nótese que por tratarse de una medida "calculada" con los datos, ***solo es aplicable*** a datos de ***variables numéricas***.

[4] No declaran su edad 3 estudiantes.

Vemos en el ejemplo cómo la media aritmética resume en un solo número toda la información del conjunto de individuos observados: "*se trata de un grupo de 136 estudiantes cuya edad promedio es de, aproximadamente, 23 años"*.

Actividad Nº 1

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

En Fórmula

Sea $\{x_1, x_2, x_3, x_4, x_5, \ldots .. x_i, \ldots\ldots x_n\}$; un conjunto de *n* observaciones de la ***variable numérica*** "X". Según la definición anterior, el valor $\bar{x}$, promedio o media aritmética del conjunto, será:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_i + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Notaciones Equivalentes

Otras formas matemáticas equivalentes para expresar al promedio, son las siguientes:

$$\bar{x} = \sum \frac{xi}{n} \qquad\qquad \bar{x} = \frac{1}{n} \sum xi$$

3.1. Principales Propiedades de $\bar{x}$

La media aritmética reúne ciertas propiedades que es importante conocer para utilizarla correctamente como resumen de un conjunto de datos, o bien para resolver algunos problemas que pueden surgir en su aplicación práctica.

- ***Primera Propiedad***

  Si dos de los términos de la expresión $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ son conocidos, se puede determinar el tercero de ellos mediante un simple pasaje de términos. Cuando se conocen $\bar{x}$ y *n,* la suma $\sum_{i=1}^{n} x_i$ se podrá determinar haciendo el producto de $\bar{x}$ por *n*. En símbolos:

  $$\sum_{i=1}^{n} x_i = \bar{x} \,.\, n$$

  Esta propiedad matemática nos permitiría saber, por ejemplo, que las $n = 32$ cárceles federales [5] de todo el país alojan un total de 60.416 internos, ya que cada una de ellas tiene una media de 1.888 presos. Esto es así porque:

  $$\sum_{1}^{32} x_i = 32 \,.\, 1888 = 60416$$

- ***Segunda Propiedad***

  El promedio es una ***medida calculada*** a partir de todos y cada uno de los datos de una serie, en consecuencia resume apropiadamente la información del conjunto. Sin embargo, por esta propiedad, en ciertas situaciones de trabajo puede perder eficacia como medida *"representativa"* del conjunto de datos.

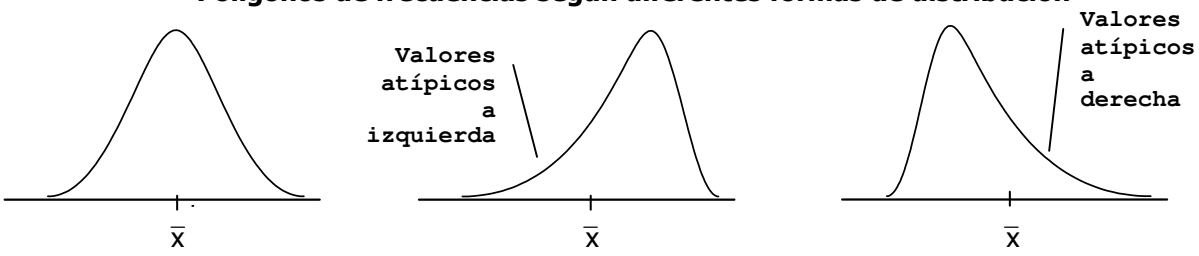
[5] Revisar el ejemplo del Párrafo Nº 2 de la Actividad Nº 1.

Cuando en la serie de observaciones existen valores extremos o *"atípicos"*, estos influirán en el valor de $\bar{x}$, pudiendo llegar a distorsionarlo de tal modo que no represente al *"común"* de los datos del conjunto (es una medida *"no resistente"*). Veamos el siguiente ejemplo:

$\bar{x} = 11,6$ es el promedio de los siguientes datos: $\{12, 10, 9, 16, 11\}$. En cambio, si el conjunto fuera $\{12, 10, 9, \mathbf{160}, 11\}$; el promedio resultaría: $\bar{x} = 40,4$. El valor atípico (160) afecta a $\bar{x}$ ***alejándola de la tendencia central*** del conjunto, resultando esta en un ***valor muy diferente*** al de los datos *normales* de la serie (12, 10, 9 y 11).

Entonces, ¿el promedio de 40,4 representa apropiadamente al *"común"* de los datos del conjunto? No, porque *"**no resiste**"* el efecto del valor extremo [6] y se ***desplaza de la tendencia central hacia el lado*** del valor atípico.

Polígonos de frecuencias según diferentes formas de distribución



Resumiendo: en un conjunto de datos en el cual los valores atípicos tienen un peso significativo (difieren mucho de los valores *"regulares"*), el ***promedio aritmético***, por ser una medida *"no resistente"*, ***debe ser analizado con cuidado.*** Esto es así porque -como en el ejemplo anterior- puede resultar fuertemente desplazado de la tendencia central e inducir a interpretaciones erróneas acerca del conjunto de datos que resume.

IMPORTANTE

La presencia de valores extremos en una distribución se manifiesta por formas (histogramas y polígonos de frecuencias) marcadamente asimétricas. De ahí la importancia de realizar una cuidadosa exploración previa (gráfica y numérica) de los datos.

- ***Tercera Propiedad***

Se denomina ***residuo o desvío individual*** de un dato cualquiera de la serie, con respecto ***a la media aritmética*** de todo el conjunto, a la ***diferencia entre el valor de ese dato y el valor*** de $\bar{x}$.

Retomando el ejemplo de las edades de los alumnos del curso de Estadística, el residuo o desvío con respecto a la edad promedio de 23 años, de cada uno de los datos del conjunto será:

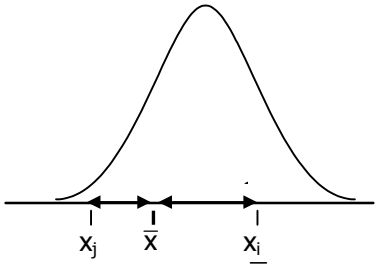
Dato (x_i)	**Desvío** ($d_i=x_i-23$)
19	-4
27	4
26	-3
28	5
....	...
x_i	x_i-23
...	...
30	7
	$\Sigma d_i=0$

[6] Los valores extremos pueden serlo por defecto o por exceso como en este ejemplo.

Cada desvío con respecto al valor de la media de todo el conjunto podrá ser negativo, nulo o positivo, según el valor del dato sea menor, igual o mayor al del promedio. Así, el desvío del primer dato x_1=19 años es: d_1=19-23=-4 años. El desvío del segundo dato x_2=27 años es: d_2=27-23=+4 años y así sucesivamente hasta el último dato x_{139}=30 años, cuyo desvío es: d_{139}=30-23=+7 años.

En forma simbólica, el desvío de un dato genérico x_i se expresa: $d_i=x_i - \bar{x}$ y para un conjunto $\{x_1, x_2, x_3, x_4, x_5, \ldots .. x_i, \ldots\ldots x_n\}$ de observaciones, habrá n residuos individuales $\{d_1, d_2, d_3, d_4, d_5, \ldots .. d_i, \ldots\ldots d_n\}$.

Es de notar que los desvíos (desprovistos del signo positivo o negativo) miden la ***"distancia"*** que separa a ***cada individuo*** observado del ***promedio general*** del grupo. Por ejemplo: el segundo individuo de la serie se diferencia en 4 años del promedio general de 23 años, mientras que la distancia al promedio del individuo 139, es de 7 años.



Los residuos de un conjunto de datos, con respecto a $\bar{x}$, tienen la propiedad de que la suma de todos ellos (cada uno con su signo negativo, nulo o positivo) es siempre igual a cero:

$$\sum_{i=1}^{n} \left(x_i - \bar{x}\right) = \sum_{i=1}^{n} di = 0$$

Es decir que, por esta propiedad, ***la suma*** (*-4+4-3+5.....+7*) de los 139 residuos individuales de las edades de los estudiantes de Estadística, **será igual a cero**[7].

- ***Cuarta Propiedad***

  En ciertas ocasiones de trabajo disponemos de dos o más promedios aritméticos, que resumen a diferentes conjuntos de datos de ***una misma variable***.

  Por ejemplo: por datos recogidos se sabe que el salario mensual promedio de n_1= 107 agentes públicos provinciales ***varones*** es $\bar{y} = \$1133,25$, mientras que el salario medio de n_2=73 empleadas ***mujeres*** es $\bar{z} = \$862,07$.

  En estas condiciones podría resultar útil conocer el promedio que resume a los salarios de ***todos los agentes públicos***, considerados como un solo conjunto de observaciones (n = 180 en total). La *"**media de medias**"* es el promedio que resuelve cuestiones como la planteada. Esta *"media de medias"* a la que simbolizaremos con la notación $\bar{\bar{x}}$ (ó $\bar{\bar{z}}$ ó $\bar{\bar{y}}$), se define del siguiente modo:

  Sea $\bar{y}$ la media aritmética de n_1 observaciones de cierta variable en estudio, y $\bar{z}$ la media de otro conjunto de n_2 datos de la misma variable; el promedio aritmético $\bar{\bar{x}}$ ***de ambas medias*** (*"media de medias")* será[8]:

$$\bar{\bar{x}} = \frac{n_1 . \bar{y} + n_2 . \bar{z}}{n_1 + n_2}$$

[7] Esta propiedad puede ser verificada en forma completa, utilizando el conjunto de 5 datos $\{12, 10, 9, 16, 11\}$ del ejemplo anterior.

[8] Es muy importante tener presente que los datos z_i e y_i deben ser *conceptualmente "promediables"* entre sí, de tal modo que $\bar{\bar{x}}$ represente un concepto válido y comprensible.

En consecuencia, *el salario promedio general de* **todos los agentes públicos** *del ejemplo será de $1023, 27 porque:*

$$\bar{x} = \frac{107 \,.\, 1133,25 + 73 \,.\, 862,07}{180} = \$1023,27$$

3.2. Cálculo de la Media

El procedimiento a seguir para el cálculo de $\bar{x}$ dependerá del estado en el que se encuentran los datos a trabajar. Esto es:

- ✓ ¿se trata de datos en el estado *"bruto"* de la matriz de datos (sin ninguna forma de resumen)?,
- ✓ ¿se trata de datos resumidos en un arreglo de frecuencias?,
- ✓ ¿se trata de datos resumidos en una distribución de frecuencias con intervalos?

IMPORTANTE

Recomendamos especialmente a los estudiantes del curso, familiarizarse con el manejo de algún *software* que les permita resolver los cálculos estadísticos mediante el uso de computadoras.

Seguidamente presentamos los procedimientos para el cálculo *manual* de $\bar{x}$ (con la ayuda de una calculadora común) con dos propósitos:

- que puedan revisar los conocimientos teóricos desde el cálculo aplicado a ejercicios concretos,
- que puedan resolver problemas de trabajo aun cuando no disponen del auxilio informático.

3.2.1. Datos sin resumir

El procedimiento de cálculo consiste en aplicar estrictamente y paso a paso, el concepto de la media aritmética. O sea: *"sumar todos los datos del conjunto y luego, dividir esa suma por el total n de observaciones de la serie"*.

3.2.2. Datos agrupados en arreglo de frecuencias

El resumen en arreglo de frecuencias permite identificar a cada dato por su valor individual y, por ello, el cálculo se realiza de igual modo que en la situación anterior*: sumando todas las observaciones individuales y dividiendo la suma por n*.

Retomemos el arreglo de frecuencias que resume la distribución de los alumnos del curso de Estadística, según las horas diarias que dedican a ver televisión.

Alumnos de Estadística según el tiempo diario que miran TV

Horas TV (x_i)	**Estudiantes (f_i)**
0	25
1	26
2	49
3	18
4	13
5	5
6	2
8	1
Total	**139**

El promedio de este grupo de datos será:

$$\bar{x} = \frac{\overbrace{0+0+....+0}^{25\ veces}+\overbrace{1+1+....+1}^{26\ veces}+\overbrace{2+2+....+2}^{49\ veces}+\overbrace{3+3+....+3}^{18\ veces}+.....+6+6+8}{139}$$

o sea:
$$\bar{x} = \frac{0.25+1.26+2.49+...+6.2+8.1}{139} = \frac{275}{139} = 2 \text{ horas diarias}$$

Es decir que, estando los datos resumidos en un arreglo de frecuencias, el procedimiento de cálculo de la media consiste en: *"multiplicar cada dato de la serie por su correspondiente frecuencia absoluta, sumar entre sí todos los productos y, finalmente, dividir la suma resultante por el total n de datos"*.

A esta forma de promediar los datos se la llama **"media ponderada por las frecuencias"** y simbólicamente se expresa como:

$$\bar{x} = \frac{\sum x_i . f_i}{n}$$

IMPORTANTE

Nótese que la media ponderada calculada a partir de un arreglo de frecuencias, reproduce ***estrictamente*** al ***concepto original*** del promedio, en tanto se trata de: *"la suma de todas las observaciones dividida por el total de datos*".

3.2.3. Datos agrupados en una distribución con intervalos

Cuando los datos se encuentran agrupados en una distribución con intervalos, es necesario basar el cálculo de $\bar{x}$ en un procedimiento que no considere a los valores individuales, ya que estos no son conocidos en esta situación de trabajo.

En el ejemplo siguiente se presenta la distribución de n = 72 "grupos turísticos"[9] observados en Puerto Iguazú, resumidos en intervalos del "*gasto total*[10] *del grupo en un día completo de estadía*" en el lugar.

Turistas Según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)	Pto. Medio (x_i)
00 - 55	19	27,5
55 - 110	20	82,5
110 - 165	18	137,5
165 - 220	7	192,5
220 - 275	4	247,5
275 - 330	3	302,5
330 - 385	1	357,5
Total	**72**	

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

La tabla permite saber, por ejemplo, que 20 grupos gastaron en un día entre $55 y $110, pero no es posible conocer el gasto exacto de cada uno de ellos individualmente.

[9] Conjunto de personas (familiares o no) que comparten el mismo presupuesto de viaje.

[10] Comprende el gasto por todo concepto (alojamiento, alimentación, transporte, esparcimiento, servicios varios, compras, etc.) por *"grupo turístico"*, en 24 horas corridas de permanencia en Pto. Iguazú.

El cálculo de la media en esta situación de trabajo, se basa en asumir a cada dato individual (desconocido) como **equivalente al valor del punto medio** o "marca" de la clase en que se ubica. Por ejemplo, se asumirá que el gasto individual de cada uno de los 18 grupos comprendidos entre $110 y $165, fue equivalente a $137,5. De igual modo, asumiremos que el gasto individual de cada grupo comprendido entre $275 y $330 fue equivalente a $302,5 y así sucesivamente para todos los datos de la distribución.

Al reemplazar los datos individuales por el valor del punto medio de clase que los representa, el promedio resultará de un cálculo similar al anterior. Es decir:

$$\bar{x} = \frac{27{,}5 \,.\, 19 + 82{,}5 \,.\, 20 + 137{,}5 \,.\, 18 + 192{,}5 \,.\, 7 + 247{,}5 \,.\, 4 + 302{,}05 \,.\, 3 + 357{,}5 \,.\, 1}{72}$$

O sea:

$$\bar{x} = \frac{8085}{72} = \$112{,}30 \text{ de gasto promedio diario por grupo}$$

Nuevamente, la media se obtiene por un procedimiento "ponderado por las frecuencias" del tipo $\bar{x} = \frac{\sum x_i . f_i}{n}$, en el cual los valores "x_i" ahora son las **marcas de cada clase** y los valores "f_i" son las correspondientes **frecuencias absolutas de clase**.

IMPORTANTE

Nótese que el valor de la media que resulta por esta forma de cálculo no es exacto, en tanto se basa en los puntos medios de clase y no en los datos originales. Se obtiene entonces, un valor "aproximado" al "verdadero valor" del promedio.

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. La Mediana

A diferencia de los promedios (la media aritmética en nuestro caso) que resultan de una operación **basada en todos los datos** de la serie, la mediana marca la tendencia central del conjunto tomando en consideración a **uno solo de ellos**.

Concepto

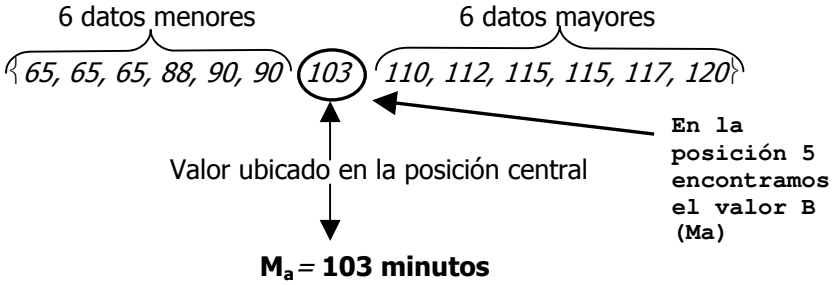
La mediana ($\mathbf{M_a}$) de una distribución es **el dato** que ocupa **la posición central** del conjunto de observaciones, debiendo estar los datos **previamente ordenados** en forma ascendente (o descendente) de magnitud.

Símbología: son diversos los símbolos aceptados para representar a esta medida: M_{dn}, M, M_{ed}, M_d, M_e, X_5, X_{me}; entre otros. Nuevamente, las letras mayúsculas y minúsculas se reservan para distinguir lo "*poblacional*" de lo "*muestral*". En este curso emplearemos indistintamente la notación $\mathbf{M_a}$.

Consideremos como ejemplo la siguiente serie de datos numéricos, referidos al "*tiempo en minutos*" que le requirió realizar un examen de Estadística a un grupo de n = 13 alumnos:

Minutos: { 120, 65, 110, 117, 65, 115, 88, 90, 103, 112, 90, 65, 115 }

El conjunto **ordenado** *en forma ascendente*[11] *resulta:*



M_a = 103 minutos

Es decir que la mediana es el **valor que se ubica en el centro del conjunto de datos ordenados** y, como tal, divide a la serie en **dos grupos con igual cantidad de observaciones** (aproximadamente la mitad): uno de ellos contiene a todos los **casos que son inferiores o iguales** al valor mediana, y el otro a todos los **casos iguales o superiores** *a él.*

Por ello, la M_a representa al **"individuo medio"** de la muestra o población en estudio: (en esta característica observada) el alumno que utilizó **"103 minutos"** para resolver el examen, es el alumno medio del grupo, ya que por debajo de él se ubican la mitad de sus compañeros y por encima la otra mitad.

4.1. Principales Propiedades de M_a

- **Primera Propiedad**

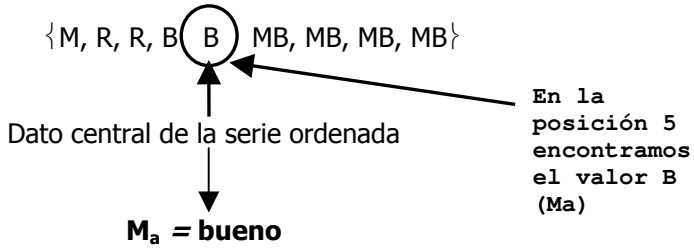
  Es una medida basada en un concepto fácilmente comprensible, que requiere de operaciones simples para aplicarla (ordenar y ubicar la posición central).

- **Segunda Propiedad**

  Siendo M_a el dato que ocupa el ***lugar central de la distribución ordenada***, el concepto tiene significado y, en consecuencia, ***es aplicable a datos categóricos ordinales***. Veamos el ejemplo siguiente en el que se analizan las respuestas sobre la *"calificación a la Fiesta Provincial de La Flor*[12]*"* (Montecarlo, Misiones, año 2001), obtenidas en un relevamiento efectuado a $n = 9$ personas mayores de 16 años que asistieron al evento.

  Calificaciones: {R, MB, MB, B, M, MB, R, MB, B}

  El conjunto **ordenado** resultará:



M_a = bueno

A ambos lados de la categoría mediana se ubica la misma cantidad de observaciones, unas de **categoría igual o inferior** a M_a y las otras, de **categoría igual o superior** a ella.

Es decir, aproximadamente *el 50% de los visitantes del ejemplo, asignó a la Fiesta una calificación "buena" o inferior y la otra mitad la calificó como "buena" o superior.*

- **Tercera Propiedad**

  La mediana de ***datos numéricos*** tiene la propiedad de ser ***"resistente"*** a la presencia de valores extremos en el conjunto de observaciones. Retomando el ejemplo de los minutos que les

[11] Idéntico resultado se obtendría si el orden en los datos fuera descendente.

[12] Las categorías posibles de respuesta fueron: muy bueno ***(MB)***, bueno ***(B)***, regular ***(R)***, malo ***(M)*** y muy malo ***(MM)***.

llevó a los 13 alumnos de Estadística realizar el examen, si ***reemplazáramos*** el dato del primer alumno (65) por el valor 5 minutos; la ***mediana del conjunto permanecería inalterada*** en:

$$M_a = 103 \text{ minutos}$$

Lo mismo ocurriría si se **reemplazara** el dato más alto de la serie (120) por cualquier valor atípico para ese conjunto de observaciones (por ejemplo 720 ó 7200).

Nótese que en estos ejemplos, la cantidad de n = 13 observaciones de la serie se mantiene inalterada, ya que suponemos la sustitución de un valor original por otro atípico. Es decir, la M_a **es resistente a valores extremos si no se modifica el tamaño n** del conjunto de datos.

- **Cuarta Propiedad**

En cambio, si al conjunto original se agregaran 2 nuevos alumnos (ahora $n = 15$) con 109 y 118 minutos respectivamente, la serie ordenada resultaría:

$$\{65, 65, 65, 88, 90, 90, 103, \underset{\uparrow}{109}, 110, 112, 115, 115, 117, 118, 120\}$$

$$\mathbf{M_a = 109 \text{ minutos}}$$

Es decir que la M_a es una medida que puede alterarse si se ***modifica la cantidad de datos*** de la serie.

- **Quinta Propiedad**

Por ser una medida que representa a todo el conjunto de datos mediante uno solo de sus valores, ***cuando se trabaja con datos numéricos*** la M_a no aporta elementos sobre la conformación general del grupo de observaciones (e individuos en consecuencia): *¿hay datos atípicos en la distribución?*, *¿cuán diferentes son los valores extremos en relación con los datos "comunes"?*

Retomando el ejemplo de Actividad Nº 2, si dijéramos que: *"la mitad de los 97 funcionarios (incluidos los 7 cargos gerenciales) de la empresa perciben haberes netos mensuales superiores a $753"*[13]; sin conocer los datos originales, no sabríamos que en el conjunto en estudio se incluyen valores tan extremos como $4927,....., $5124,...$6701 y $6890.

4.2. Determinación de la M_a

El procedimiento a seguir para determinar[14] el valor mediana de una distribución en estudio, dependerá del tipo de datos que se trate (numéricos u ordinales) y del estado de elaboración en que se encuentran (datos brutos, arreglos de frecuencias, distribución con intervalos).

4.2.1. Datos numéricos sin resumir

- Si el número de observaciones es impar

Cuando los datos en análisis son **numéricos** y el **número n** de observaciones que forman el conjunto **es impar**, habrá un **único valor** que ocupará la **posición central** del conjunto ordenado (ejemplos anteriores de n = 13 ó n = 15 estudiantes en el examen de Estadística). En esta situación el procedimiento consistirá en **ordenar rigurosamente** los datos por su magnitud (sentido ascendente o descendente) y luego, **identificar el valor que se ubica en el lugar central del conjunto ordenado** (que deja igual cantidad de datos a ambos lados). Ese valor es la mediana del conjunto.

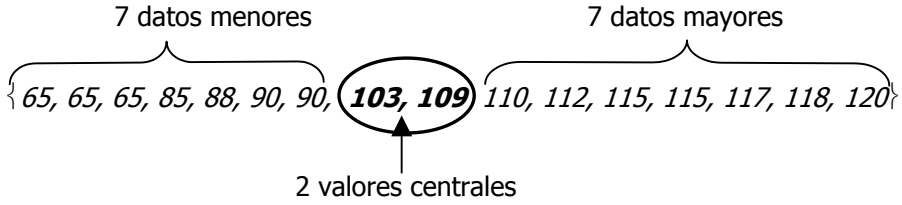
- Si el número de observaciones es par

Cuando el **número n** de observaciones de la serie **es par**, serán dos los valores centrales del conjunto ordenado, que separarán la misma cantidad de datos hacia

[13] Recomendamos realizar el ejercicio de verificar la exactitud de esta afirmación.

[14] Nótese que hablamos de *"determinar"* y no de *"calcular"* M_a, porque se trata de una medida "no calculada". Si bien analizaremos procedimientos basados en fórmulas y cálculos numéricos con los datos, en todos los casos se trata de razonamientos para **identificar el valor central** de la serie ordenada, tal como se define esta medida.

ambos lados. Por ejemplo, supongamos que fueron n = 16 los alumnos que rindieron el examen de Estadística:



En este caso la M_a se determina **por convención**, **promediando ambos datos centrales**. Es decir:

$$Ma = \frac{103+109}{2} = 106 \text{ minutos}$$ [15]

4.2.2. Datos numéricos en arreglo de frecuencias

En esta situación de trabajo el razonamiento debe seguir los mismos pasos anteriores, considerando que en el arreglo de frecuencias los **datos ya se encuentran ordenados por magnitud**. El problema entonces consiste en:

a- ubicar el lugar central del conjunto ordenado (posición del valor M_a),

b- identificar el valor (o los valores si n es par) que ocupa esa posición (o esas posiciones).

Retomemos como ejemplo la distribución de los alumnos del curso de Estadística, según las horas diarias que dedican a la TV:

Alumnos de Estadística según el tiempo diario que miran TV

Horas TV (x_i)	**Estudiantes (f_i)**	F_a
0	25	25
1	26	51
Ma → **2**	**49**	**100**
3	18	118
4	13	131
5	5	136
6	2	138
8	1	139
Total	**139**	

a- Ubicación del lugar central de la distribución ordenada

- Si el número de observaciones es impar (ej.: n = 139), el conjunto ordenado de menor a mayor ocupará 139 posiciones [16] y una sola de ellas será la central: $\frac{139+1}{2} = 70$, de tal modo que a su izquierda quedarán 69 datos menores o iguales y a su derecha otros 69 datos mayores o iguales.

Tenemos así que, tratándose de un **número impar** de observaciones, la **posición o lugar central** de la distribución se determina mediante:

[15] Notar que en este caso, M_a no es exactamente un dato de la serie. La medida toma el valor "teórico" que resulta de promediar los dos datos centrales y, en consecuencia, ocupa un lugar también "teórico", ubicado entre ambos valores.

[16] Imagine a los 139 valores individuales ordenados uno al lado del otro sobre una recta horizontal. El primero será "0" (se repite por 25 veces) y el último será 8 (una sola vez).

$$\text{Posición Ma} = \frac{n+1}{2}$$

- Si el número de observaciones es par (ej.: $n = 160$ alumnos), serán dos las posiciones centrales ($\text{Posición } 80 = \frac{160}{2}$ y $\text{Posición } 81 = \frac{160}{2} + 1$) las que dejan igual cantidad de observaciones hacia ambos lados (79 en este caso).

Tratándose de un **número par de datos**, las **dos posiciones centrales se determinan** mediante:

$$\text{Posición}_1 = \frac{n}{2} \quad y \quad \text{Posición}_2 = \frac{n}{2} + 1$$

b- Determinación del valor Ma

Habiendo identificado la posición central (o las dos posiciones cuando n es par) del conjunto ordenado, el problema ahora es identificar el dato (o los datos) que se ubica(n) en ese lugar. Para ello nos valemos de las frecuencias acumuladas (en el sentido "menor que"), razonando en el ejemplo anterior del siguiente modo:

- ✓ Hasta el valor 1 de la distribución **se acumulan** 51 datos ordenados y, en consecuencia, ninguno de ellos (valores 0 y 1 del arreglo) alcanzan la **posición 70**.
- ✓ Al pasar al valor 2 ya son 100 las observaciones acumuladas, lo que significa que uno de los 49 datos iguales a 2 es el que ocupa la posición central 70.
- ✓ Es decir: la $M_a = 2$ horas diarias.

Este valor de la mediana nos indica que "aproximadamente la mitad de los alumnos entrevistados dedica 2 horas diarias o menos a ver TV" (obviamente la otra mitad, dedica 2 horas o más por día).

El razonamiento es idéntico cuando el **número n de casos** del conjunto **es par**, teniendo en cuenta que ahora el problema consiste en identificar los valores que ocupan las **dos posiciones centrales** y luego, determinar M_a como el promedio entre ambos datos.

4.2.3. Datos *numéricos* en una distribución con intervalos

En esta situación de trabajo la mediana no puede ser determinada **exactamente** porque, al ser **desconocidos** los **datos individuales** que forman el conjunto en estudio, no hay manera de reconocer el valor que ocupa la posición central de la serie ordenada [17]. Por ello, el procedimiento consiste en **estimar** la M_a mediante el siguiente razonamiento:

a. determinar el punto medio "teórico" (o centro geométrico) de la serie haciendo:

$$\text{Posición Ma} = \frac{n}{2}$$

b. analizando las frecuencias acumuladas ("menor que"), identificar la clase o intervalo ("clase mediana") de la distribución en la que se ubica dicha posición;

c. **estimar** el valor mediana aplicando la siguiente **fórmula de interpolación**:

$$Ma = L_i + \frac{\frac{n}{2} - Fa_{(i-1)}}{fi} \cdot a$$

siendo:

M_a : valor estimado de la mediana,
L_i : límite inferior de la "clase mediana",
$\frac{n}{2}$: punto medio de la serie de datos,
$Fa_{(i-1)}$: frecuencia acumulada anterior a la "clase mediana",
f_i : frecuencia absoluta de la "clase mediana",
a : amplitud de la "clase mediana".

Retomemos el ejemplo del gasto diario de los turistas en Pto. Iguazú

[17] Es de notar que los datos se encuentran ***ordenados por la magnitud de sus intervalos.***

Turistas según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)	F_a
00 - 55	19	19
55 - 110	**20**	**39**
110 - 165	18	57
165 - 220	7	64
220 - 275	4	68
275 - 330	3	71
330 - 385	1	72
Total	**72**	

clase M_a →

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

- Punto medio de la distribución:

$$\frac{n}{2} = \frac{72}{2} = 36$$

- Analizando las frecuencias acumuladas se observa que la primera clase reúne a los **primeros 19 datos** ordenados de la distribución y, en consecuencia, ninguno de ellos alcanza al **punto medio 36**.

  Al pasar a la segunda clase ya **son 39 los datos acumulados** en sentido ascendente de magnitud, razón por la cual entre los 20 datos de esta clase se encuentran los dos valores centrales de la distribución. Es decir, ésta es la "clase mediana"[18].

- Localizada la clase donde se ubica M_a, su **valor estimado** resultará de hacer:

$$Ma = 55 + \frac{36 - 19}{20}.55 = \$101,75$$

lo que permite decir: "la mitad de los grupos turísticos tienen un gasto diario de aproximadamente $101,75 o menos".

4.2.4. Datos categóricos ordinales

Cuando los datos en análisis son **ordinales** y se encuentran resumidos en una tabla de frecuencias, el procedimiento sigue un razonamiento similar al de la situación "datos numéricos en arreglo de frecuencias". O sea:

a- ubicar el lugar central (o los lugares si n es par) del conjunto ordenado (posición de la categoría M_a),

b- identificar el valor (o los valores si n es par) que ocupa esa posición (o esas posiciones).

Consideremos el ejemplo sobre los usuarios de la empresa misionera de servicios eléctricos:

Opinión de los Usuarios sobre el Servicio Eléctrico de Mnes. (EMSA)

Opinión	Usuarios	F_a
Muy Malo	3	3
Malo	20	23
Regular	151	174
Bueno	**469**	**511**
M. Bueno	42	685
TOTAL	**685**	

M_a →

Fuente: Departamento TISE-FHyCS. 1994

[18] La clase de la mediana siempre es aquella cuya frecuencia acumulada "menor que", resulta ***igual o inmediatamente mayor*** a: $\frac{n}{2}$ ó $\frac{n+1}{2}$, según corresponda.

- Posición central (en este caso *n es impar*): $\frac{n+1}{2} = \frac{686}{2} = 343$
- **Localizada la posición central** del conjunto ordenado, nos valemos de las frecuencias acumuladas para **identificar al dato que se ubica en ese lugar**. La categoría "muy malo" acumula 3 observaciones, la categoría "malo", 23 observaciones y 174 son las opiniones "regular" o menos. Al pasar a la categoría siguiente ya son 511 los datos acumulados, razón por la que uno de los 469 datos "bueno" es el que ocupa el lugar central 343. En consecuencia M_a = "bueno".

Así: "*aproximadamente la mitad de los usuarios entrevistados, tienen una opinión* ***"buena" o superior*** *sobre el servicio eléctrico que reciben*".

Si el número *n* de datos de la serie fuera **par** (por ejemplo n = 734 usuarios), existirían dos posiciones centrales: $\text{Posición}_1 = \frac{n}{2}$ y $\text{Posición}_2 = \frac{n}{2} + 1$ (lugares 367 y 368 en nuestro ejemplo). Con la ayuda de las frecuencias acumuladas, se podrá localizar la $\boldsymbol{M_a}$ identificando los datos (categoría) que se ubican en estos lugares.

Actividad Nº 3

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad***.**

5. El Modo

Concepto

El modo ($\mathbf{M_o}$) de un conjunto de observaciones (numéricas o categóricas nominales u ordinales) es el **valor individual que más veces se repite** en la serie. $\mathbf{M_o}$ será el valor más típico, más recurrente o bien, el que reúne la **mayor frecuencia absoluta** entre todos los valores (categorías) individuales observados en el conjunto de datos que se analiza.

Símbología: algunos símbolos utilizados para representar a esta medida son: M_{do}, X_{mo}; entre otros. Nuevamente, las letras mayúsculas y minúsculas se reservan para distinguir lo "*poblacional*" de lo "*muestral*". En este curso emplearemos indistintamente la notación $\mathbf{M_o}$.

En este caso tenemos también una medida que toma en consideración a **una sola de las observaciones**, aunque no siempre se ubica en los valores centrales de la serie de datos.

Tomando como ejemplo la serie de datos referidos al "tiempo en minutos" que le requirió realizar un examen a los alumnos de Estadística y a las "calificaciones a la Fiesta Provincial de La Flor", realizadas por 9 visitantes al evento, tendremos:

dato más frecuente

Minutos: { *65, 65, 65, 88, 90, 90, 103, 110, 112, 115, 115, 117, 120* }

M_o= 65 minutos

dato más frecuente

Calificaciones: { M, R, R, B, B, MB, MB, MB, MB }

M_o = muy bueno

5.1. Principales Propiedades del M_o

- **Primera Propiedad**

  Es una medida conceptualmente simple, fácil de interpretar y de comunicar, que requiere únicamente del conteo para ser determinada.

- **Segunda Propiedad**

  Por no requerir de ninguna forma de orden en los datos, tiene significado y ***es aplicable a datos categóricos nominales*** (es la única de las tres medidas de tendencia central que hemos tratado, posible de ser utilizada con este tipo de datos).

- **Tercera Propiedad**

  Cuando la diferencia entre la frecuencia máxima observada (frecuencia modal) con alguna de las restantes no es muy grande, el M_o como medida característica de la distribución pierde relevancia.

IMPORTANTE

Puede ocurrir que en un conjunto de datos se encuentren dos o más valores que reúnen la misma frecuencia absoluta máxima[19] (en nuestros ejemplos si tuviéramos ***dos alumnos más*** con 90 y 115 minutos respectivamente o bien, ***dos visitantes más*** que califiquen la *Fiesta de la Flor* como *Regular*). En tales casos las distribuciones resultarían ***bimodal*** (dos valores con la misma frecuencia máxima) o ***multimodal*** (tres o más valores con esta propiedad) y no es posible determinar un único valor/categoría M_o para toda la serie.

5.2. Determinación del M_o

5.2.1. Para arreglos de frecuencias y datos categóricos

Si los datos individuales se encuentran sin agrupar, lo recomendable es resumirlos previamente en un arreglo de frecuencias (o en una tabla de frecuencias para datos categóricos). Encontrándose los datos presentados de esta manera, la determinación del M_o simplemente se remite a ubicar en la distribución, el valor o categoría al que corresponde la mayor frecuencia absoluta.

Consideremos el siguiente ejemplo:

Estudiantes del Curso de Estadística. FHyCS-Año 2001

según Sexo

Sexo	Estudiantes
Varón	30
Mujer ← M_o	**109**
Total	**139**

Fuente: elaboración propia.

según el Tiempo Diario que Miran TV

Horas TV (x_i)	Estudiantes (f_i)
0	25
1	26
2 ← M_o	**49**
3	18
4	13
5	5
6	2
8	1
Total	**139**

[19] Esta situación es muy raro que ocurra si el número (n) de observaciones es "suficientemente grande".

Usuarios del Servicio Eléctrico de Misiones (EMSA), Según Opiniones sobre la Calidad del Servicio

Opinión	Usuarios
M. Bueno	42
Bueno	**469**
Regular	151
Malo	20
Muy Malo	3
TOTAL	**685**

M_o

Fuente: Departamento TISE-FHyCS. 1994

Así entonces:

"*las mujeres predominan en el grupo de estudiantes de Estadística y lo más común o frecuente son los alumnos que dedican 2 horas diarias a ver TV*", *y*
"*la opinión de que el servicio eléctrico es bueno, es la más típica entre los usuarios de la Empresa de Electricidad de Misiones*".

5.2.2. Para una distribución con intervalos

En la situación de trabajo en la que los datos son numéricos y se encuentran resumidos en una distribución con intervalos (como el ejemplo de los gastos turísticos que se presentan a continuación), el M_o debe determinarse mediante el siguiente **procedimiento de estimación**, aceptado por convención:

Turistas según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)
00 - 55	19
55 - 110	**20**
110 - 165	18
165 - 220	7
220 - 275	4
275 - 330	3
330 - 385	1
Total	**72**

clase modal

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

Asumiendo que la clase que presenta la **mayor frecuencia absoluta** de la distribución (**"clase modal"**) es la que **contiene entre sus datos** al valor modal, una vez identificada el valor del M_o se puede **estimar** mediante el siguie*nte* **procedimiento de interpolación**:

$$M_o = L_i + \frac{d_1}{d_1 + d_2} . a$$

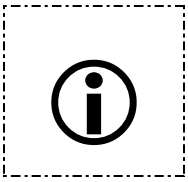
siendo:
L_i: límite inferior de la clase modal,
d_1: la diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase inmediata anterior a la modal,
d_2: la diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase inmediatamente posterior a la modal,
a: amplitud de la clase modal.

En nuestro ejemplo resultará:

L_i = 55 **d_1 =** 20-19 = 1 **d_2 =** 20-18 = 2 **a =** 55

$$Mo = 55 + \frac{1}{1+2} . 55 = \$73,3 \text{ diarios}$$

O sea: "*estimamos que el gasto más frecuente entre los 72 casos observados, es de $73,3 diarios*".



IMPORTANTE

Este procedimiento para estimar el modo de datos numéricos agrupados en clases es ***altamente sensible a la forma*** en que se define la distribución. Esto es: al número de intervalos y a la amplitud de cada uno de ellos.

El siguiente ejemplo ilustra sobre este problema. El mismo grupo de n = 9 datos se organiza de 3 maneras distintas:

Situación A

Datos	fi
65 ← M_o	**2**
70	1
72	1
73	1
81	1
82	1
86	1
87	1
Total	**9**

Situación B

Datos	fi
65 - 69	2
70 – 74 ← *Clase M_o*	**3**
75 – 79	0
80 – 84	2
85 - 89	2
Total	**9**

Situación C

Datos	fi
65 - 69	2
70 – 79 ← *Clase M_o*	3
80 - 89	**4**
Total	**9**

El **modo verdadero** de la serie es $M_o = 65$ ya que se trata del valor del conjunto con mayor frecuencia (Situación A).

En la Situación B la **clase modal** es la segunda de la distribución (**70–74**) y aplicando el procedimiento de estimación por interpolación resulta: $M_o = 70{,}75$.

En la Situación C el M_o se ubicará en la tercera clase (**80–89**), resultando su estimación: $M_o = 81{,}5$ [20].

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.

6. Cuartiles, Deciles, Centiles

Utilizando medidas de tendencia central podemos describir a los grupos turísticos observados en Puerto Iguazú señalando, por ejemplo, que:

"*se trata de grupos que observan un promedio de $112,30 diarios de gasto por todo concepto; siendo $73,30 la suma que diariamente gastan con mayor frecuencia y la mitad de los grupos analizados destina $101,75 o más por día a satisfacer sus necesidades*".

Esta descripción permite una buena **comprensión global** de los datos elaborados y, por ende, de los individuos analizados; pero muy poco o nada nos informa sobre aspectos más específicos del fenómeno en estudio. Por ejemplo:

- ✓ ¿por encima de qué valor se ubican los turistas que más gastan? o en términos más concretos, ¿qué nivel del gasto corresponde al 10% de los turistas que más gastan?,
- ✓ ¿por debajo de qué monto se ubican los grupos que menos gastan diariamente?,
- ✓ ¿entre qué valores están los niveles de gastos centrales?,
- ✓ etc.

[20] Sugerimos verificar los resultados de las situaciones A y B.

Es decir, en la descripción de un conjunto de datos, **las medidas de tendencia central no dan cuenta de la diversidad de situaciones (variabilidad o dispersión) que se presentan**. Es preciso entonces, agregar a esta información otros elementos que permitan una descripción más completa, haciendo referencia a otras características de la distribución [21].

En todo conjunto de datos (numéricos u ordinales) se pueden determinar **ciertos valores característicos que amplían la información** proporcionada por las medidas sintéticas de tendencia central sobre los individuos que se analizan. Estos datos, ubicados en posiciones estratégicas del conjunto, permiten conocer aspectos de su composición y estructura, que aportan nuevos elementos para el análisis. Es decir, las preguntas señaladas precedentemente, pueden responderse a partir de ciertos **datos ubicados estratégicamente** en una **distribución ordenada**.

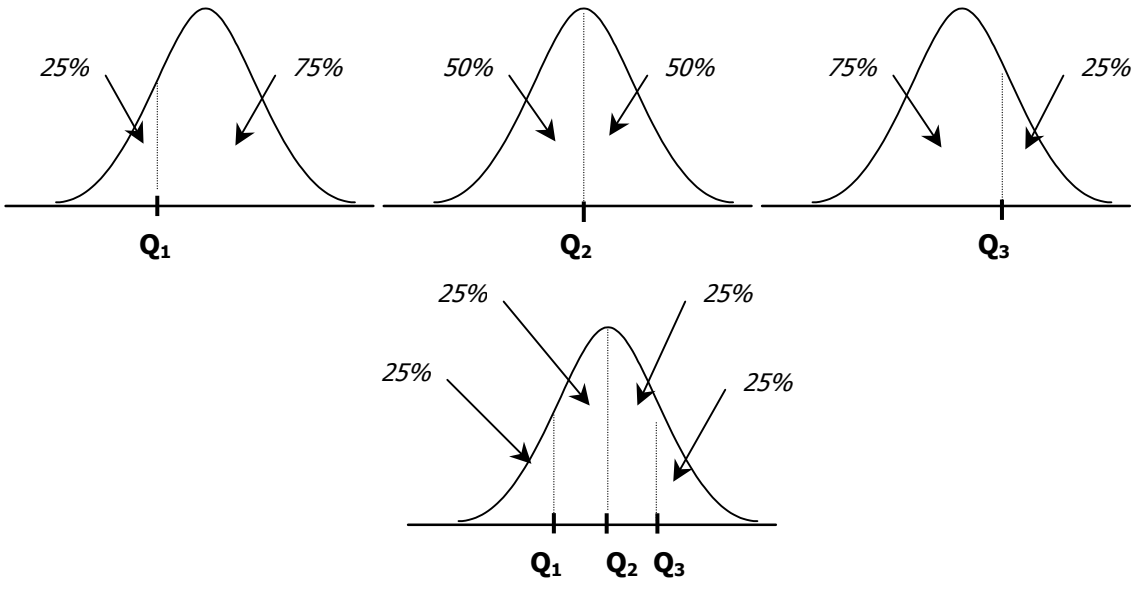
Las *medidas de posición*: **cuartiles, deciles y centiles**, son las que permiten individualizar a los datos que reúnen las condiciones señaladas.

6.1. Los Cuartiles

En toda distribución de datos **numéricos o categóricos ordinales** es posible hallar **tres observaciones individuales** que dividen al **conjunto, previamente ordenado en forma ascendente**, en cuatro partes iguales, cada una de ellas con el 25% de los datos.

- **Cuartil 1 – Primer Cuartil (Q_1):** es aquel valor del conjunto de observaciones que se ubica en una posición tal que a uno de sus lados deja una cuarta parte (25%) de los datos que son menores o iguales a él, y hacia el otro lado las tres cuartas partes (75%) de los datos que son mayores o iguales que él (es el valor que se ubica en la posición ¼ del conjunto ordenado).
- **Cuartil 2 – Segundo Cuartil (Q_2):** coincide con la mediana ya que divide al conjunto en dos partes, cada una con la mitad de los datos: $Q_2 = M_a$.
- **Cuartil 3 – Tercer Cuartil (Q_3):** es el dato situado en la posición que deja ¾ de las observaciones menores o iguales que él hacia un lado y ¼ de las observaciones mayores o iguales que él hacia el otro lado (el dato que se ubica en la posición ¾ de la serie ordenada).

Gráficamente



[21] Una vez más: no se trata de reducir la descripción de un conjunto de datos en un único valor, por más expresivo que el mismo pueda resultar, sino de comunicar *la forma* de la distribución en la que se expresa la disparidad y repetición de los valores de la variable.

Ejemplo:

Para la distribución de los grupos turísticos según el nivel de gasto diario en Iguazú, los cuartiles resultan:

Q_1 = \$52,11 Q_2 = Ma = \$101,75 Q_3 = \$155,83

Es decir que:

"*Una cuarta parte de los grupos (los 18 grupos que menos gastan) registra un nivel de gasto diario igual o inferior a \$52,11, mientras que el 25% de los que más gastan se ubican en \$155,83 ó más por día. Es decir que el 50% (36) de los grupos centrales registra un nivel de gasto comprendido entre \$52,11 y \$155,83 diarios*".

"*Considerando que el gasto mediana es de \$101,75, una cuarta parte de los turistas registra gastos diarios entre \$52,11 y \$101,75, y otra cuarta parte gasta entre \$101,75 y \$155,83*".

Determinación de los Cuartiles

El procedimiento para determinar Q_1 y Q_3 de una distribución sigue un razonamiento análogo al de la mediana, pero considerando que ahora se trata de identificar a los datos localizados en las posiciones ¼ y ¾ del conjunto ordenado. Para ello procedemos de la siguiente manera:

- **Localizamos las posiciones de los cuartiles**; la manera más sencilla de obtenerlas es:

$$\text{Posición } Q_1 = \frac{n}{4} \quad \text{y} \quad \text{Posición } Q_3 = 3.\frac{n}{4}$$

En nuestro ejemplo de los gastos turísticos, la posición del cuartil 1 será:

$$\text{Posición } Q_1 = \frac{72}{4} = 18$$

- Posteriormente, **inspeccionando las frecuencias acumuladas**, individualizamos los datos que ocupan las posiciones cuartílicas deseadas.
- **Cuando los datos son numéricos y se encuentran resumidos en una distribución con *intervalos***, primero debemos ***ubicar la clase del cuartil***, y luego ***estim*ar** su valor mediante el siguiente cálculo:

$$Q_1 = L_i + \frac{\frac{n}{4} - Fa_{(i-1)}}{f_i}.a \qquad \text{y} \qquad Q_1 = L_i + \frac{\frac{3.n}{4} - Fa_{(i-1)}}{f_i}.a$$

Donde los datos a considerar en cada una de estas expresiones (L_i, $Fa_{(i-1)}$, f_i, a) toman como referencia a las clases de Q_1 y Q_3 respectivamente, con significado idéntico al explicado para determinar la M_a en esta situación de trabajo.

En el ejemplo de los gastos turísticos:

La clase del cuartil 1 es la primera (0-55), por consiguiente podemos estimar el Q_1 de la siguiente manera:

$$Q_1 = L_i + \frac{\frac{n}{4} - Fa_{(i-1)}}{f_i}.a = 0 + \frac{18-0}{19}.55 = 52,11$$

Siguiendo el procedimiento indicado, verifique el valor correspondiente al tercer cuartil.

6.2. Los Deciles

Son los nueve valores de la distribución ordenada en forma ascendente que la dividen en diez partes iguales, cada una de ellas con el 10% de los datos.

- **Decil 1 – Primer Decil (D_1):** es aquel *valor del conjunto de observaciones* que se ubica en una posición tal que, a uno de sus lados *deja al 10% de los datos que son menores o iguales a él* y,

hacia el otro lado, el 90% de los datos restantes que son mayores o iguales que él (es el valor que separa el primer décimo del conjunto ordenado en forma ascendente).

- **Deciles 2, 3, 4, 5, 6, 7, 8 y 9 (D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9):** se definen trasladando el concepto de D_1 al segundo décimo, tercer décimo....., noveno décimo de la serie ordenada en forma ascendente ($D_5 = M_a$).

  En este caso, la forma sencilla de **ubicar la posición** de un decil genérico "i" (para i = 1, 2, 3, 4, 5, 6, 7, 8 ó 9) será mediante el cociente:

$$\frac{i \,.\, n}{10}$$

  Luego, la determinación seguirá los pasos ya explicados y la **estimación por interpolación** se basará en:

$$D_i = L_i + \frac{\frac{i.n}{10} - Fa_{(i-1)}}{f_i} \,.\, a$$

6.3. Los Centiles (C_1, C_2, ..., C_{98}, C_{99})

Son noventa y nueve valores de la distribución ordenada en forma ascendente, que la dividen en cien partes iguales, cada una de ellas con el 1% de los datos.

La posición del "i"-ésimo centil (siendo i = 1, 2, 3,....., 98 ó 99) se determina por:

$$\frac{i.n}{100}$$

La estimación por interpolación resulta de aplicar la siguiente operación **a la clase del centil** genérico "i":

$$C_i = L_i + \frac{\frac{i.n}{100} - Fa_{(i-1)}}{f_i} \,.\, a$$

Actividad Nº 5

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad***.**

6.4. La curva de Lorenz asociada a medidas de posición

Como vimos en la **unidad anterior**, es posible asociar a cierto tipo de variables (ingreso, propiedad de la tierra, etc.) la gráfica de Lorenz, que nos permitirá analizar el grado de concentración/distribución de estos recursos en la población en estudio. En aquel momento, se presentó la construcción de esta gráfica a partir de una **tabla de frecuencias construida en base a intervalos de igual amplitud**; sin embargo, es posible hacerlo construyendo intervalos de distinta amplitud, cada uno de los cuales incluya la misma cantidad de individuos, de tal forma que la frecuencia relativa porcentual en cada uno de ellos sea del 25%, o del 10%, etc. Esto significa construir intervalos cuyo límite superior coincide con los cuartiles (tendríamos cuatro intervalos), o con los deciles (diez intervalos), etc.

Consideremos por ejemplo la distribución de los hogares según el ingreso familiar en la ciudad de Formosa. Se puede ver en el Cuadro siguiente que los hogares aparecen distribuidos en intervalos de clase de diferente amplitud, de manera que cada uno de los mismos agrupa aproximadamente un 10% del total de los hogares (4329 hogares). De esta manera estamos presentando los datos en una **distribución según deciles de ingreso**.

Distribución de los Hogares según ingreso familiar – Formosa, octubre 1997

Decil	Escala Ingresos	Hogares (%)	Ingreso total Por Decil (miles)	Porcentaje de Ingreso	Ingreso medio por decil
1	20-200	10	549	1,9	127
2	200-250	10	976	3,3	225
3	250-330	10	1281	4,3	296
4	330-400	10	1603	5,4	371
5	400-500	10	1901	6,4	439
6	500-600	10	2316	7,8	533
7	600-710	10	2796	9,4	652
8	720-980	10	3584	12,1	830
9	980-1330	10	4935	16,7	1134
10	1330-10449	10	9668	32,7	2219
Total		**100 (43288)**	**29609**	**100,0**	**684**

Fuente: INDEC – EPH. 1998

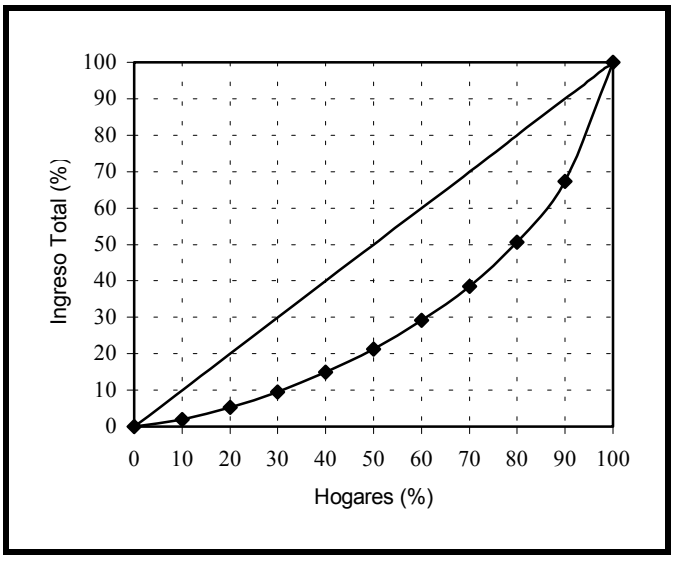
En la tabla se aprecia que, entre los hogares de la ciudad de Formosa, existe una concentración de los ingresos: el 10% de los hogares que más ganan concentran el 32,7% del total de los ingresos, mientras que el 10% de los hogares más pobres acumulan sólo el 1,9%. Esta situación produce una brecha entre "ricos" y "pobres", en la que el **ingreso promedio del último decil ($2219) es 17,5 veces mayor que el ingreso promedio del primer decil**. Esta comparación se podría extender a otros grupos, por ejemplo comparar el primer 20% de los hogares (primer quintil) que acumula sólo el 5,2% frente al último 20% que acumula el 49,4% del total de los ingresos; y así sucesivamente.

La curva de Lorenz tiene la ventaja de expresar las situaciones de equidad/inequidad de manera más general, permitiendo apreciar el comportamiento de la variable en forma inmediata.

Según hemos visto en la unidad anterior, para construir la curva de Lorenz tenemos que realizar las siguientes transformaciones: acumular los porcentajes de hogares y acumular los porcentajes de ingresos totales por decil.

Distribución de los Hogares según deciles de ingreso - Formosa, octubre 1997

Decil	Escala Ingresos	Hogares Acum. (%)	Ingresos Acum. (%)
1	20-200	10	1,9
2	200-250	20	5,2
3	250-330	30	9,5
4	330-400	40	14,9
5	400-500	50	21,3
6	500-600	60	29,1
7	600-710	70	38,5
8	720-980	80	50,6
9	980-1330	90	67,3
10	1330-10449	100	100,0



La curva así construida expresa de manera elocuente la **concentración del ingreso** que existe en los hogares de Formosa, y el hecho de haber utilizado los deciles facilita la lectura comparativa de los datos.

Actividad Nº 6

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 6 de la Guía de Actividades correspondiente a esta unidad.

7. ¿Cómo Integrar estas Medidas de Resumen?

Hemos presentado hasta aquí una serie de valores característicos de una distribución que nos permite señalar diferentes aspectos del conjunto de datos que se analiza. Cada una de estas medidas dirige nuestra mirada hacia algún rasgo de interés de ese conjunto, las que serán más ilustrativas en tanto sean integradas en una descripción que totalice todos los aspectos destacables, generando así una "buena imagen" de esa distribución.

7.1. El resumen de los cinco números

Una forma aceptada y eficaz de integrar diferentes medidas descriptivas es la que se conoce como "**el resumen de los cinco números**", en la que se consideran:

- $\mathbf{X_{mín}}$: el mínimo
- $\mathbf{Q_1}$: el cuartil 1
- $\mathbf{M_a}$: la mediana
- $\mathbf{Q_3}$: el cuartil 3
- $\mathbf{X_{máx}}$: el máximo

Con estos valores, estamos describiendo la distribución identificando un valor de tendencia central (la mediana), dos valores entre los cuales se concentran el 50% de los datos centrales (Q_1 y *Q_3) y otros dos valores entre los cuales se dispersa el conjunto total de los datos ($X_{mín}$ y $X_{máx}$).*

Si consideramos los gastos diarios de los grupos turísticos, podemos describir mediante este criterio al conjunto de las observaciones utilizando los siguientes valores:

$\mathbf{X_{mín}}$= \$ 0 $\mathbf{Q_1}$= \$52,11 $\mathbf{M_a}$= \$101,75 $\mathbf{Q_3}$ = \$155,83 $\mathbf{X_{máx}}$= \$385

"*La mitad de los grupos turísticos no superan los $101,75 de gasto diario, aunque los gastos observados varían $0 y $385. Por otro lado, el 50% de los gastos centrales se ubican entre $52,11 y $155,83".*

Así como el resumen de los cinco números resulta un recurso apropiado para hacer una descripción de la distribución, también se pueden incorporar otros valores característicos que expresen nuevas especificidades del conjunto de datos. En este sentido, es posible agregar al análisis, otras medidas que nos permitan dar una mejor idea de la forma de la distribución. Por ejemplo, utilizando además de los cinco números vistos, los deciles 1 y 9 en un resumen que podríamos llamar "de los siete números".

$\mathbf{X_{mín}}$= \$ 0 $\mathbf{D_1}$= \$20,8 $\mathbf{Q_1}$= \$52,11 $\mathbf{M_a}$= \$101,75 $\mathbf{Q_3}$ = \$155,83 $\mathbf{D_9}$= \$231 $\mathbf{X_{máx}}$= \$385

Al comentario anterior basado en los cinco números, se podría agregar que:

"El 10% de los que menos gastan no superan los $20,8 diarios, mientras que un 10% de los grupos turísticos, gastan diariamente $231 o más".

IMPORTANTE

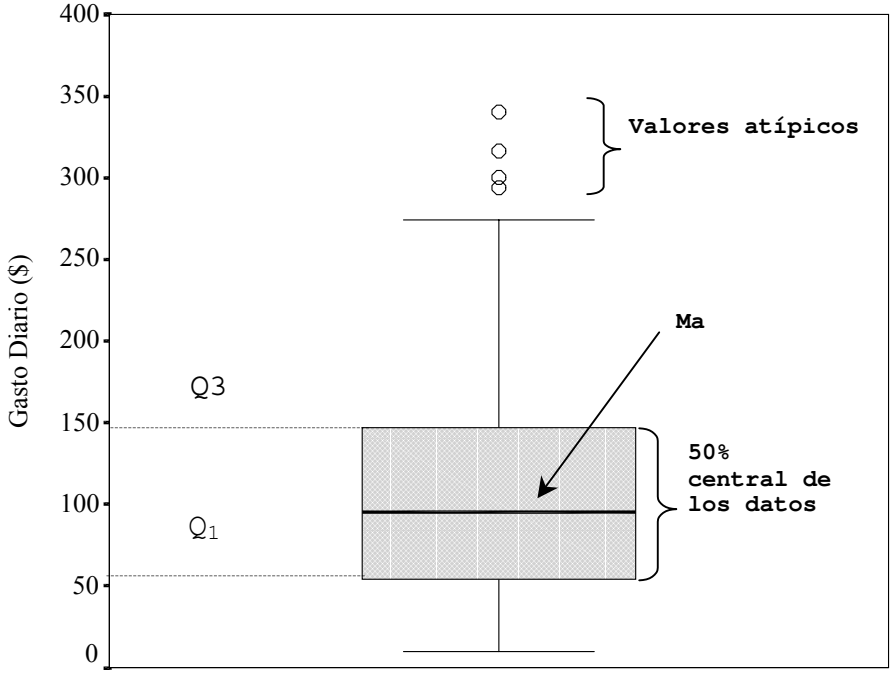
La decisión del número de valores característicos a utilizar para la descripción, e incluso qué deciles incorporar, depende de las particularidades de la distribución: número de casos, forma, número de valores diferentes que tome la variable y propósitos del análisis.

7.2. El diagrama de Caja (*Box-plot*)

El **recurso gráfico asociado al resumen de los cinco números** es lo que se conoce como Diagrama de Caja [22]. En este diagrama se utiliza un rectángulo (caja) que limitado por los cuartiles uno y tres, incluye en su interior el 50% de los datos centrales; dentro de la caja se señala la mediana con un segmento. A partir de esos límites del rectángulo, se grafican líneas -llamadas "bigotes"- con una longitud igual a 1,5 veces la distancia entre el cuartil 1 y el 3 [23]. Posteriormente –fuera de los "bigotes"- el gráfico identifica aquellos valores atípicos (*outliers*), que están a más de 1,5 veces la distancia Intercuartil (1,5 . R Q) de los extremos de la caja.

A continuación presentamos el diagrama de Caja construido a partir de los datos individuales de los gastos realizados diariamente por los 72 grupos turísticos.

Diagrama de Caja: Distribución de los gastos diarios.
Pto Iguazú, Feb. '94



En este gráfico podemos ver que los gastos diarios de los turistas tienen un comportamiento bastante simétrico en el 50% de los datos centrales (la mediana se ubica en el centro de la caja, a igual distancia de los cuartiles uno y tres). El conjunto total de los datos muestra una asimetría a la derecha, (el bigote superior es más largo que el inferior e incluso se aprecia la presencia de cuatro grupos turísticos con gastos atípicos). Por otro lado el "bigote" inferior está indicando una mayor concentración de los gastos menores, no hay valores atípicos pequeños e incluso no se identifica ningún grupo que no haya realizado gastos (el "bigote" no alcanza al valor $0).

Este tipo de recurso gráfico resulta muy ilustrativo y en consecuencia recomendable cuando queremos comparar dos o más distribuciones [24].

Vemos entonces que el diagrama de caja permite visualizar una serie de aspectos interesantes de la forma del conjunto de los datos:

- *Presencia de* ***valores atípicos***

[22] También denominado "Diagrama de Caja con bigotes" o en inglés "*Box-Plot*".

[23] En la unidad siguiente, se podrá ver que esta distancia entre el cuartil 1 y el 3 es una medida de variabilidad que se conoce como *Rango intercuartil* (RQ).

[24] El uso del *box-plot* para la comparación de conjuntos de datos, será tratado posteriormente en la Unidad 5.

- ***Simetría del conjunto central*** *de los datos (equidistancia o no de la mediana a los cuartiles).*
- ***Simetría del conjunto total*** *de datos (forma de la caja y longitud de los bigotes).*
- ***Dispersión en cada una de las zonas*** *en las que queda dividida la distribución (la longitud de cada parte, expresa la mayor o menor proximidad de los datos entre sí).*
- *El* ***rango*** *de la distribución (distancia entre el valor máximo y mínimo).*

Estas características del diagrama hacen que el mismo resulte **útil** (junto con el de *tallo-hoja*) en la **etapa inicial exploratoria** de los datos, previo a la construcción de una distribución de frecuencias y cálculo de las medidas resumen, ya que -como hemos visto- **la forma de la distribución condiciona el posterior tratamiento y resumen de los datos**.

Actividad Nº 7

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 7 de la Guía de Actividades correspondiente a esta unidad***.**

8. ¿Qué Hemos Visto?

En esta unidad hemos avanzado un paso más en el camino del tratamiento y análisis estadístico elemental de los datos.

Efectuados los primeros resúmenes numéricos y gráficos, para una primera lectura del fenómeno que representan los datos (unidad 2), el análisis a menudo requiere de instrumentos que permitan un **mayor resumen de la información**.

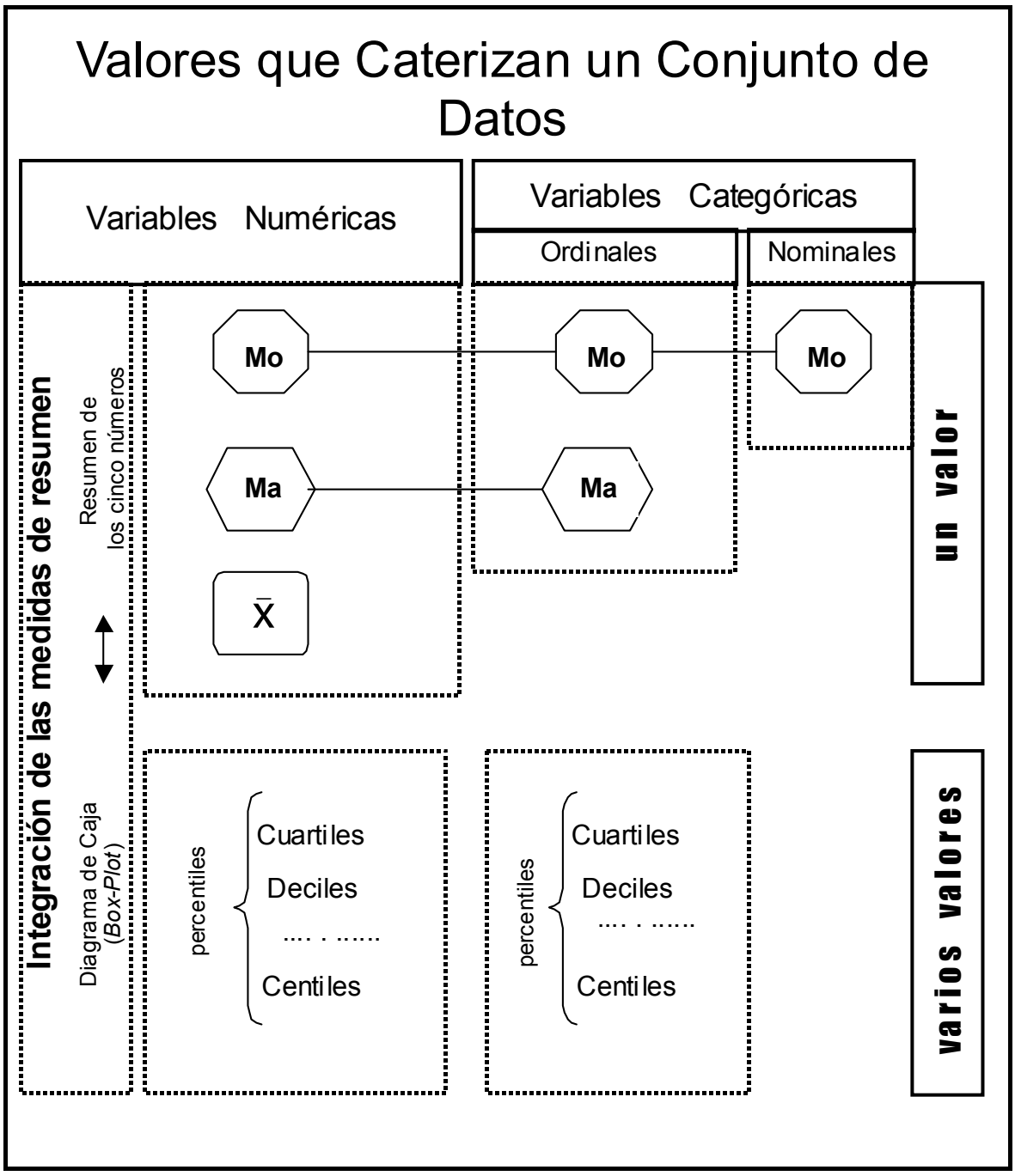
Las medidas de **tendencia central** tienen este propósito, y su aplicación en un problema particular **dependerá básicamente de las necesidades de información que motivan el análisis**, del **tipo de datos** con los que se trabaja y de las **propiedades del conjunto** como un todo.

El buen dominio del concepto, propiedades y limitaciones de cada una de ellas es el requisito para utilizarlas correctamente.

Además, hemos presentado las **diferentes medidas de posición** que permiten complementar la comprensión de un conjunto de datos, informando -con distintos niveles de detalle- sobre su estructura.

En todos los casos, el énfasis está puesto en facilitar la comprensión conceptual de cada herramienta, para luego pasar al plano de la formalización matemática elemental y del cálculo aplicado a ejemplos de fácil comprensión.

En relación con esto último, reiteramos la recomendación a quienes puedan hacerlo, de utilizar la informática como auxiliar del trabajo estadístico.


Valores que Caterizan un Conjunto de Datos
Variables Numéricas
Variables Categóricas
Ordinales
Nominales
Integración de las medidas de resumen
Resumen de los cinco números
Diagrama de Caja (Box-Plot)
Mo
Mo
Mo
Ma
Ma
X̄
un valor
percentiles
Cuartiles
Deciles
....
Centiles
percentiles
Cuartiles
Deciles
....
Centiles
varios valores

Bibliografía

BARBANCHO, A. (1978): *Estadística Elemental Moderna*. Ed. Ariel, Barcelona, España. pág. 117-123, 127-132, 134-138.

BLALOCK, H. M.(1978): *Estadística Social,* FCE, México. pág. 67-72, 81-83.

UNIVERSIDAD NACIONAL DE CÓRDOBA (1993): *Estadística aplicada a la Investigación. Curso a distancia*. Fac. de Cs. Económicas, Córdoba, Módulo III pág. 1-42.

Conceptos Centrales

- Media aritmética: concepto y propiedades.
- Mediana: concepto y propiedades.
- Modo: concepto y propiedades.
- Cuartiles, deciles, centiles: concepto y aplicación.

Habilidades

- Reconocer la utilidad, alcances y limitaciones de cada una de las medidas resumen presentadas.
- Identificar para una situación de trabajo, las medidas de Tendencia Central y Posición que podrían utilizarse para una buena descripción de los datos.
- Conocer los fundamentos que guían los procedimientos para la obtención de estas medidas.
- Interpretar en términos de un problema, las medidas y gráficos asociados a una distribución (*Box-plot* y Curva de Lorenz).
- Saber comunicar en un informe las características de un conjunto de datos, integrando los distintos recursos estadísticos aprendidos hasta el momento.

UNIDAD 4: ANÁLISIS DE LA VARIACIÓN Y ASIMETRÍA

1. ¿Por qué Evaluar la Variabilidad y la Asimetría?

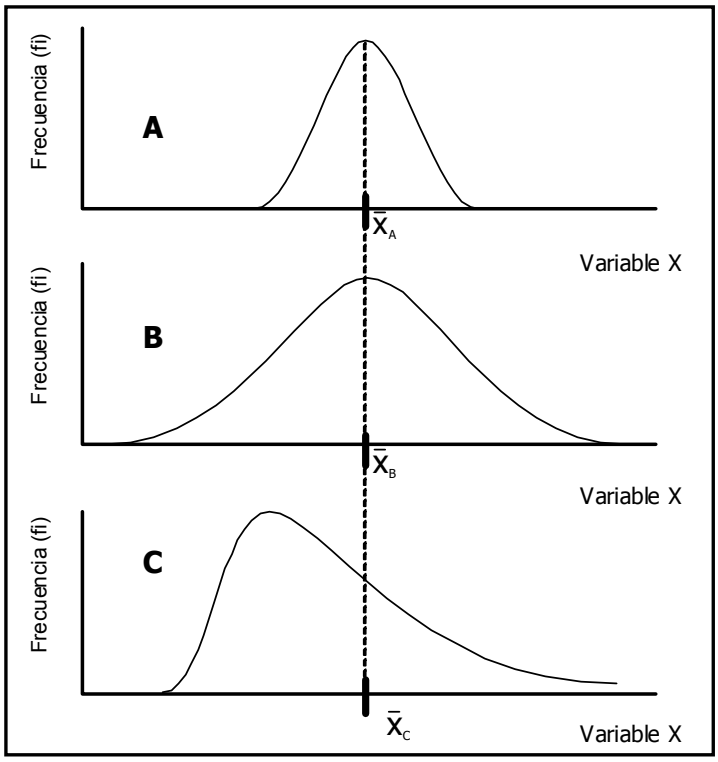
No se investiga lo obvio, aquello que encuentra una respuesta simple y evidente. Las preguntas que nos formulamos generalmente aluden a situaciones complejas, comprenden fenómenos en los que las características de interés presentan valores diversos, no son uniformes.

Dicho en términos estadísticos, los datos que obtenemos en relación con alguna pregunta de investigación, **varían a través del conjunto de unidades observadas**, y "controlar" esa variabilidad es el fin último en la tarea de describir los datos y producir información.

Hasta aquí todas las medidas o herramientas presentadas intentaban, de diferentes maneras, resumir los datos para lograr una mejor **descripción de esa diversidad**. Así, las **distribuciones de frecuencias** (en su forma numérica o gráfica) nos permiten ***presentar y describir los diferentes valores observados***. En tanto que las ***medidas resumen*** desarrolladas en la unidad anterior, nos facilitan la ***descripción de los individuos a través de un conjunto de valores característicos*** que intentan dar cuenta de la variabilidad.

Asimismo, debemos destacar que **la representatividad de las medidas de tendencia central se vincula estrechamente con la dispersión de los datos** y (concretamente en el caso de la media) con la ***simetría*** de la distribución[1]**.** Consideremos los siguientes gráficos donde se representan tres distribuciones de frecuencias (polígonos A, B y C) que registran un mismo valor para la media.

Distribuciones con igual media aritmética y diferente variabilidad y/o simetría



Evaluando los gráficos, es posible concluir que la media aritmética resulta mucho más representativa del conjunto de datos en la distribución **A** (simétrica y con menor variabilidad) que en las situaciones **B** (simétrica pero con valores más dispersos en torno a la media) y **C** (también más dispersa y asimétrica).

[1] Esto pone de manifiesto que tanto la variabilidad como la asimetría de la distribución son aspectos a considerar a la hora de evaluar estas medidas. Recordar que: cuando se observa la presencia de valores atípicos, el ***promedio aritmético debe ser analizado con cuidado,*** porque puede resultar fuertemente desplazado de la tendencia central e inducir a interpretaciones erróneas acerca del conjunto de datos que resume (Ver Unidad 3).

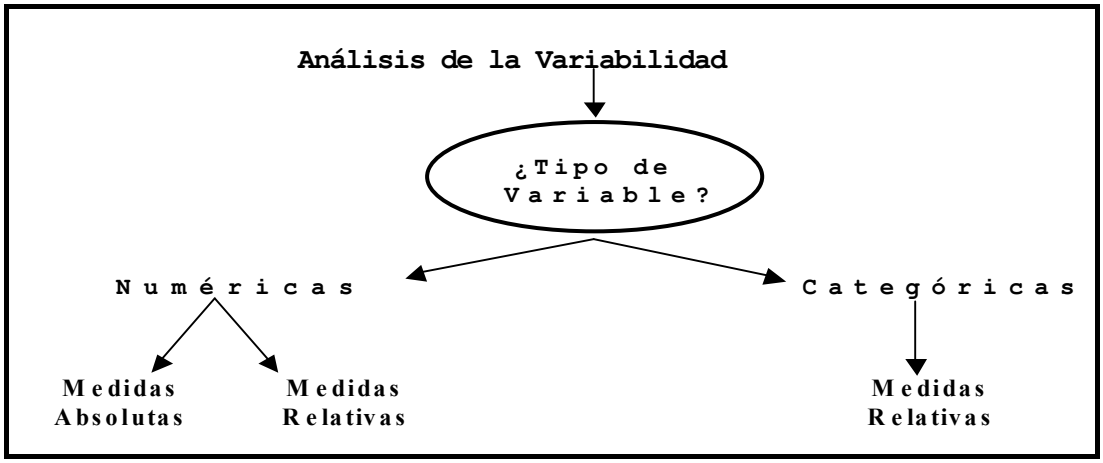
A estas características que hacen a *la forma* (variabilidad y simetría[2]) de la distribución, le podemos asociar *medidas que resuman en números la "cantidad de variación" y el "grado de asimetría"*, valores que nos permitirán comparar distintos conjuntos de individuos.

En esta unidad abordaremos -en primera instancia- cómo medir la variabilidad, para posteriormente presentar aquellas medidas del grado de asimetría de una distribución.

2. ¿Cómo Medir la Variabilidad?

¿Qué significa medir la variabilidad? Obtener un único número que exprese qué tan ***dispersos o diferentes*** son entre sí el conjunto de valores observados o -lo que es lo mismo- que indique cuán homogéneos son los individuos en términos de la característica en cuestión.

Si bien el concepto de variabilidad es único, las medidas son distintas según se trate de variables numéricas o categóricas. Además, para las variables numéricas podemos identificar medidas absolutas y relativas.



2.1. Para variables numéricas

Por tratarse de variables *medidas en una escala de intervalo*, la dispersión de los valores observados se puede expresar directamente por la diferencia aritmética entre esos valores. En consecuencia, cuanto mayor sea la diferencia entre dos valores, podemos aseverar que mayor será la variación que existe entre esos dos datos.

Veamos en un sencillo ejemplo, las ideas anteriores: tenemos seis individuos para los cuales se han registrado sus notas en Historia y Matemática.

Simboliza al segundo individuo

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Historia	7	8	7	6	7	7	**7**
Nota Matemática	4	8	5	9	10	6	**7**

Es la nota de Matemática del cuarto individuo

Se puede observar que los promedios de las notas en estas materias son coincidentes. Sin embargo, la variabilidad en las notas de Historia es claramente menor que en las de Matemática; así la mayor variación que se registra entre las notas de Historia es de 2 puntos (entre $\mathbf{i_2}$ y $\mathbf{i_4}$, que son los individuos más diferentes entre sí), mientras que en Matemática, la mayor diferencia es de 6 puntos (entre $\mathbf{i_5}$ y $\mathbf{i_1}$). Estamos en condiciones de afirmar para este pequeño conjunto de observaciones que, a pesar de que la medida resumen es la misma, los conjuntos son diferentes: las notas de Matemática

[2] Aunque no lo desarrollaremos en este curso, otro aspecto a considerar en el análisis de la forma es lo que se conoce como *curtosis*.

son más heterogéneas (están más dispersas) que las de Historia. El promedio en Historia ***"representa mucho mejor"*** al rendimiento de los estudiantes en esa asignatura, que la nota promedio de Matemática al correspondiente conjunto de datos.

IMPORTANTE

Las medidas de tendencia central ocultan la variabilidad del conjunto de datos. Por ello, cuantificar la variabilidad constituye un complemento imprescindible en la descripción de una distribución.

Conocer (medir) la variación de los datos permite:

- ➔ describir esta característica inherente a todo conjunto de observaciones,
- ➔ evaluar la *"calidad"* de las medidas de tendencia central, y
- ➔ comparar mejor diferentes grupos de datos mediante sus promedios.

En general, las situaciones no serán tan evidentes, ni el número de datos tan pequeños como en el ejemplo anterior; lo que obliga a construir medidas que nos permitan resumir y evaluar esa variabilidad.

2.1.1. Las medidas absolutas

Para la construcción de medidas absolutas de variación se pueden adoptar dos perspectivas:

➔ **Considerar el campo de variación de las variables**: las medidas obtenidas expresan la extensión o amplitud de variación de los datos que se están considerando. Se identifican en este grupo: el *Rango* y el *Rango Intercuartil*.

➔ **Considerar las variaciones de los datos individuales:** estas medidas resumen en un valor la totalidad de las variaciones de los datos individuales. Entre estas medidas se destacan: la *Desviación Media*, la *Desviación Mediana*, la *Variancia* y el *Desvío Estándar*.

Considerando el campo de variación de las variables, tenemos:

A) El Rango, Amplitud o Recorrido: indica la extensión en la que varían la totalidad de los datos; es la mayor diferencia que se puede registrar entre dos valores de la variable.

Esta medida se calcula como la diferencia entre el máximo valor y el mínimo valor observado de la variable.

$$R = x_{máx} - x_{mín}$$

En el ejemplo de las notas el rango para la variable "nota de Matemática" es de 6 (R = 10 - 4), lo que indica que la totalidad de las notas observadas se registran en un campo o extensión de variación de 6 puntos. En el caso de las "notas de Historia" esta amplitud de variación es de 2 puntos.

Cuando los datos están agrupados en intervalos de clase, dado que no conocemos exactamente el máximo y el mínimo, el rango se obtiene [3] haciendo la diferencia entre el límite superior de la última clase y el límite inferior de la primera:

$$R = L_{sk} - L_{í1} \quad \text{(donde k es el número de clases)}$$

Comentarios*:*

Es una medida de muy fácil cálculo, que permite una aproximación rápida a la variabilidad de los datos.

Al tomar sólo los valores máximo y mínimo, si se observan **valores muy atípicos**, puede brindar una **idea distorsionada** sobre la variabilidad como característica del conjunto.

Dos distribuciones con el **mismo rango** pueden tener dispersión "interna" de los **datos muy diferentes** (el conjunto de los valores pueden estar más o menos concentrados).

[3] Estrictamente se trata de una estimación ya que desconocemos los verdaderos valores máximos y mínimos.

B) Rango intercuartil: indica la extensión en la que varían el 50% de los datos centrales de la distribución.

Se calcula como la diferencia entre el tercer y el primer cuartil.

$$RQ = Q_3 - Q_1$$

Comentarios:

Muchas veces es preferible medir la variabilidad del 50% de los datos centrales, descartando el 25% de los valores más bajos y el 25% de los más altos, para evitar así la distorsión que puede provocar la presencia de valores atípicos.

Simultáneamente, estamos prescindiendo en este caso de la mitad de las observaciones.

Para describir la distribución de las edades de los alumnos del curso de Estadística podemos utilizar algunas de las medidas de resumen presentadas en la unidad anterior.

Mediana	21 años
Mínimo	17 años
Máximo	47 años
Cuartil 1	19 años
Cuartil 3	27 años

A estas medidas las podemos complementar con medidas de variación. Así tenemos:

Rango: R = 47 - 17 = 30 años

Rango intercuartil: RQ= 27 - 19 = 8 años

A partir de este conjunto de medidas se puede decir que: *la mitad de los alumnos de Estadística tienen 21 años o menos, y los más jóvenes tienen 17 años. Las edades de los estudiantes varían en una amplitud de 30 años, lo que implica una diferencia de 30 años entre el/(los) alumno/s más jóven/es y el/(los) de más edad. El 50% de los estudiantes con las edades centrales difieren a lo sumo en 8 años.*

Recordar que el Diagrama de Caja (*Box-Plot*) es un recurso gráfico apropiado para el análisis de la distribución en general y de la variabilidad y asimetría en particular. (Ver unidad 3).

Actividad Nº 1

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

Considerando las variaciones de los datos individuales tenemos:

Una alternativa que facilita construir estas medidas es tomar los ***desvíos de cada uno de los valores individuales con respecto a un punto elegido como referencia***. Generalmente este valor de referencia es una medida de tendencia central.

C) Desviación media: esta medida se construye tomando ***todos* los *desvíos individuales*** con respecto a la media aritmética.

Como hemos definido, *un **desvío individual** es la diferencia entre un valor de la variable y la media aritmética*: $d_i = (x_i - \bar{x})$. Es decir que tendremos tantos desvíos individuales como individuos hayamos observado.

En el ejemplo de las notas de Matemática tendríamos los seis desvíos siguientes:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Matemática	4	8	5	9	10	6	**7**
Desvíos individuales a la $\bar{x}$	-3	1	-2	2	3	-1	

Se puede ver que, mientras el individuo 1 está 3 puntos por debajo de la media, el individuo 5 está en esa misma cantidad por encima de la media.

Para resumir en un único número la variabilidad de las seis observaciones, podemos recurrir al promedio pero, como ya hemos señalado en la unidad anterior, *la suma de los desvíos a la media es cero* [4]. Para resolver este problema vamos a sumar los desvíos absolutos, es decir el valor de los desvíos prescindiendo de su signo.

En términos del problema tenemos que la *Desviación Media* se obtiene como:

$$DM = \frac{3+1+2+2+3+1}{6} = \frac{12}{6} = 2 \text{ puntos}$$

Se interpreta que, *en promedio, las notas de matemática se desvían de la media en 2 puntos*.

Desviación Media (DM):

Es el promedio de los desvíos individuales (en valor absoluto) con respecto a la media aritmética.

$$DM = \frac{\sum |x_i - \bar{x}|}{n} = \frac{\sum |d_i|}{n}$$

Las barras simbolizan "valor absoluto"

Comentario:

Cuando estamos en presencia de distribuciones en las que se observan **valores atípicos** (marcadamente asimétricas) la media como medida resumen *no es recomendable*, y en consecuencia *tampoco lo es la desviación media* como medida de variabilidad.

Para el caso de las edades de los alumnos del curso de Estadística, la *Desviación Media* calculada a partir de los valores individuales, es: DM = 5,14 años (Ud. podría controlar este resultado, calculando la DM a partir de los datos que figuran en la Unidad 2).

Para datos organizados en una distribución de frecuencias:

- Si se trata de un **arreglo** de frecuencias y se va a obtener la desviación media en forma manual, la expresión de cálculo es:

$$DM = \frac{\sum |x_i - \bar{x}| . f_i}{n} = \frac{\sum |d_i| . f_i}{n}$$ donde f_i es la frecuencia del valor x_i

- Cuando los datos están *agrupados en* ***intervalos de clase***, *y no se dispone de los valores individuales*, se podrá estimar la Desviación Media, considerando que el x_i de la fórmula se corresponde con el punto medio de la clase.

En el caso de la edad de los estudiantes, si desconociéramos los valores individuales de esta variable y contáramos únicamente con los datos organizados en una distribución de frecuencias en intervalos de clase, podríamos estimar la Desviación Media realizando las operaciones que se indican en la Tabla.

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

Punto Medio de clase — **Desvíos individuales**

Edad	nº de estud. (f_i)	PM	d_i = (PM– 23,6)	$\|d_i\| . f_i$
17-20	65	18,5	-5,1	331,5
21-24	25	22,5	-1,1	27,5
25-28	17	26,5	2,9	49,3
29-32	14	30,5	6,9	96,6
33-36	7	34,5	10,9	76,3
37-40	5	38,5	14,9	74,5
41-44	2	42,5	18,9	37,8
45-48	1	46,5	22,9	22,9
Total	**136**			**716,4**

Fuente: elab. propia en base a datos del "Estudio de los Alumnos de Estadística"

[4] Recordar que por una propiedad de la media la suma de los desvíos individuales a la media siempre es cero. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Dividiendo la suma de los desvíos en valores absolutos (716,4) por el número de casos (136), tenemos una Desviación Media estimada en 5,27 años.

"Las edades de los alumnos de estadística se dispersan –en promedio- con respecto a la media en 5,27 años".

D) Desviación mediana: si evaluamos que la media no es una buena medida resumen de los datos y optamos por la mediana como medida de tendencia central, sería apropiado utilizar una medida de dispersión relacionada a la mediana. Así entonces, de manera análoga a la desviación media, tenemos que:

Desviación Mediana (DMa):

Es el promedio de los desvíos individuales (en valor absoluto) con respecto a la mediana.

$$DMa = \frac{\sum |x_i - Ma|}{n}$$

Comentarios:

- Para datos organizados en distribuciones de frecuencias, valen los mismos comentarios que para el cálculo de la Desviación Media.

$$DMa = \frac{\sum |x_i - Ma| . f_i}{n}$$

donde:

$\mathbf{f_i}$ es la frecuencia del valor x_i

$\mathbf{x_i}$ son los valores observados de la variable en el caso de un arreglo de frecuencias, o el punto medio de la clase en el caso de una distribución en intervalos de clase.

Calculamos la Desviación Mediana para las Notas de Matemática:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	Ma
Nota Matemática	4	8	5	9	10	6	7
Desvíos a la Ma	-3	1	-2	2	3	-1	

Promedio de los valores centrales 6 y 8

$$DMa = \frac{\sum |x_i - Ma|}{n} = \frac{3+1+2+2+3+1}{6} = 2 \text{ puntos}$$

En consecuencia, *las notas de Matemática se desvían, en un promedio de 2 puntos, de la mediana*.

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

La edad de los estudiantes es una distribución marcadamente asimétrica a la cha y la mediana (Ma = 21,5) será la mejor medida resumen de los datos. Así, lo más apropiado es utilizar la desviación mediana, que se obtiene mediante las operaciones que se presentan en la Tabla:

Edad	n° de estud. (f_i)	PM	d_i = (PM – 21,5)	$\|d_i\| . f_i$
17-20	65	18,5	-3	195
21-24	25	22,5	1	25
25-28	17	26,5	5	85
29-32	14	30,5	9	126
33-36	7	34,5	13	91
37-40	5	38,5	17	85
41-44	2	42,5	21	42
45-48	1	46,5	25	25
Total	136			674

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

Desvíos individuales a la mediana

Suma del producto de los **desvíos absolutos** individuales a la mediana por la frecuencia

Luego: $DMa=\frac{674,0}{136}=4,96$ años

"Esta medida indica que en promedio las edades de los estudiantes se desvían de la mediana en 4,96 años".

E) Variancia y Desviación estándar: en el cálculo de la desviación media se tomaron los valores absolutos de los desvíos evitando así que la suma nos dé cero. Otro criterio para solucionar este mismo problema sería elevar esos desvíos al cuadrado, obteniendo de esta manera una nueva medida de variabilidad que se conoce como Variancia.

Esta medida se simboliza utilizando la letra griega "sigma" elevada al cuadrado (σ^2).

El cálculo de la variancia para las notas de Matemática es:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Matemática	4	8	5	9	10	6	**7**
Desvíos individuales a la $\bar{x}$	-3	1	-2	2	3	-1	

¿. . . ?

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(-3)^2 + (1)^2 + (-2)^2 + (2)^2 + (3)^2 + (-1)^2}{6} = \frac{28}{6} = 4,7 \text{ (puntos)}^2$$

Variancia **(σ^2)**:

Es el promedio de los cuadrados de los desvíos a la media aritmética.

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Comentarios:

- La *variancia* y el *desvío estándar* son, fundamentalmente por razones de orden teórico, las medidas *más utilizadas* para cuantificar la variabilidad de un conjunto de datos.
- Dado que los desvíos a la media están elevados al cuadrado, la variancia ***se expresa en una unidad de medida que es el cuadrado de la unidad de medida de la variable original***. Esto dificulta la interpretación del resultado en términos del problema.

La unidad de medida en la que queda expresada la variancia no es interpretable en términos de la variable que se analiza. Hasta aquí sólo la podemos considerar como una cuantificación de la variabilidad existente en los datos.

Para resolver este problema, se calcula la *raíz cuadrada de la variancia*, que resulta en una nueva medida llamada ***Desvío Estándar*** (σ), la que queda expresada en la unidad original.

$$\sigma = \sqrt{\sigma^2}$$

En el ejemplo de las notas de Matemática el desvío estándar será:

$$\sigma = \sqrt{4,7} = 2,2 \text{ puntos}$$

"*Las notas de matemática de los alumnos se dispersan en promedio en torno a la media en 2,2 puntos*".

Si no contáramos con los datos originales, el cálculo de la variancia y el desvío estándar para las edades de los estudiantes de estadística, a partir de la tabla, sería:

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

Desvíos individuales a la media

Desvíos al cuadrado

Edad	nº de estud. (f_i)	PM	d_i = (PM– 24,1)	d_i^2	$d_i^2 . f_i$
17-20	65	18,5	-5,1	26,0	1690,0
21-24	25	22,5	-1,1	1,2	30,0
25-28	17	26,5	2,9	8,4	142,8
29-32	14	30,5	6,9	47,6	666,4
33-36	7	34,5	10,9	118,8	831,6
37-40	5	38,5	14,9	222,0	1110,0
41-44	2	42,5	18,9	357,2	714,4
45-48	1	46,5	22,9	524,4	524,4
Total	**136**				**5709,6**

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

La variancia es: $\sigma^2 = \frac{\sum d_i^2 . f_i}{n} = \frac{5709,6}{136} = 42,0$

El desvío estándar es: $\sigma = \sqrt{42} = 6,48$ años.

Entonces, *los estudiantes del curso tienen una media de 24,1 años y sus edades -en promedio- se dispersan con respecto a ese valor 6,48 años*.

Para datos agrupados en distribuciones de frecuencias:

La expresión de cálculo es:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 . f_i}{n}$$

donde: $\mathbf{f_i}$ es la frecuencia del valor x_i

$\mathbf{x_i}$ son los valores observados de la variable en el caso de un arreglo de frecuencias, o el punto medio de la clase en el caso de una distribución en intervalos de clase.

Actividad Nº 2

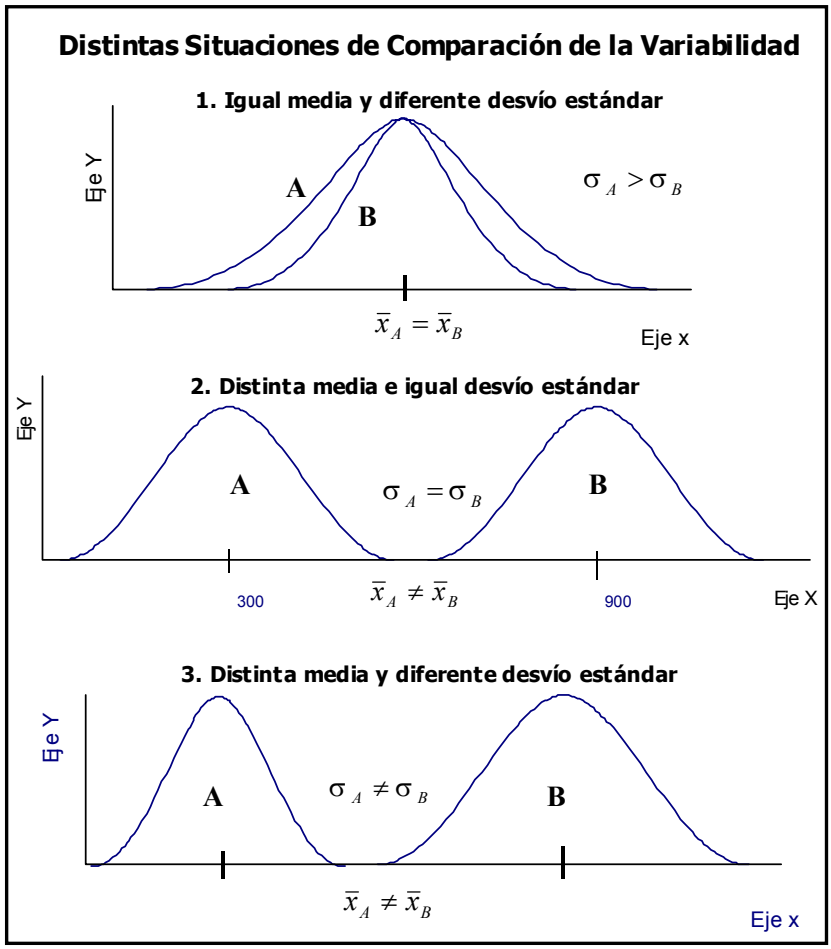
*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 2 de la Guía de Actividades correspondiente a esta unidad***.**

2.1.2. Las medidas relativas

Con frecuencia nos vemos en situaciones de tener que ***comparar la variabilidad de diferentes conjuntos de datos***. Así por ejemplo, comparar los ingresos de grupos pertenecientes a distintos estratos sociales, las edades de grupos en diferentes etapas de la vida, las temperaturas en distintas regiones del planeta, etc.

Existen **diferentes situaciones** que se pueden presentar **al comparar distribuciones**. En el esquema siguiente se presentan, en términos generales, esas situaciones de comparación.

El **primer Gráfico** está expresando una situación en la cual debemos comparar la variabilidad de dos grupos que -medidos en la misma variable- tienen **medias iguales** y **dispersiones diferentes**. Es fácil de concluir que en la distribución B los individuos son más homogéneos que en la otra.



La dificultad de comparar no se presenta tan clara en las otras dos situaciones (2 y 3).

Cuando las variables están medidas en la misma escala (situación 2), no es difícil de ver que:

- una variación de 2 años entre escolares, no implica la misma heterogeneidad de los individuos (en cuanto a: intereses, preferencias y habilidades) que esa misma variación entre universitarios, o
- una dispersión de $50 pesos en el ingreso mensual de gerentes de empresa, no los diferencia (en cuanto al nivel de vida o consumo), de la misma manera que esa misma variación lo hace entre sus obreros, etc.
- Es aún más evidente la dificultad de comparar la hogeneineidad de los individuos cuando las distribuciones tienen valores distintos de promedio y dispersión absoluta (situación 3). Por ejemplo esto ocurriría si queremos comparar:
    - la variación en el consumo de energía eléctrica de los hogares y de las industrias. Si conociéramos que el desvío estándar en el consumo de los hogares es de 100 Kw y entre las industrias es de 1500 Kw; no tenemos información suficiente para concluir sobre la mayor o menor homogeneidad en alguno de las poblaciones, dado que -como podemos suponer- sus promedios son sustancialmente diferentes.

En consecuencia, **para valorar la dispersión de un grupo y poder compararlo con otro**, se hace ***necesario evaluar la dispersión en términos relativos a las magnitudes de esas variables en cada uno de los grupos***. Esto significa que, comparar la cantidad de dispersión de dos grupos, exige construir ***medidas relativas de variabilidad***.

Esta necesidad de ***relativizar la variabilidad***, se evidencia también cuando se busca comparar la homogeneidad de dos conjuntos de observaciones en términos de dos ***variables expresadas en unidades de medida distintas***. Por ejemplo, queremos ver si nuestros estudiantes se parecen más entre sí (son más homogéneos) en cuanto al tiempo que miran televisión (en horas), que en relación a su edad (en años); los turistas que visitan Puerto Iguazú se parecen más entre sí en términos de sus años de estudio que de sus gastos, etc. Así, los interrogantes nos conducirían a comparar la dispersión de la edad de los alumnos con la dispersión en el tiempo que miran TV; y la variabilidad de gastos de los turistas, con la variabilidad en los años de estudio. Ambas situaciones son incomparables en términos de variabilidad absoluta.

F) Coeficiente de variación

Es la medida relativa de dispersión más utilizada dado que se construye a partir de la desviación estándar que, como hemos dicho, es la medida de dispersión más difundida.

Coeficiente de Variación (CV):

Definido como:

$$CV = \frac{\sigma}{\bar{x}}.100$$

indica la cantidad de variación expresada como un porcentaje de la media aritmética.

Comentarios:

- Si las medias aritméticas de dos conjuntos son iguales (o aproximadamente) las medidas absolutas serán suficientes para la comparación.

	Edad	**Hs. TV**
N	136	139
$\bar{X}$	23,4 años	2,0 hs.
σ	6,4 años	1,5 hs.
CV	27,3 %	75,8 %

En el ejemplo de los estudiantes, podemos ver que las *edades se dispersan en promedio un 27,3% del valor de la media aritmética*, mientras que el *tiempo que miran TV tiene una dispersión del 75,8% del promedio general*. En conclusión, "*el grupo es mucho más homogéneo en términos de sus edades que en relación con sus hábitos como televidentes".*

Existen **otras medidas relativas** de variación que se construyen de manera análoga al coeficiente de variación, según sea la medida absoluta de dispersión que se considere. Así tenemos:

G) Coeficiente de Desviación Media

$$CDM = \frac{DM}{\bar{x}}.100$$

H) Coeficiente de Desviación Mediana

$$CDMa = \frac{DMa}{Ma}.100$$

donde: *DM* es la desviación media y *DMa* es la desviación mediana.

IMPORTANTE

En la práctica *no se construyen sucesivamente* todas las medidas que hemos presentado sino que, a partir de la medida de resumen seleccionada como más representativa de la ***tendencia central, se seleccionará una medida de dispersión que la complemente***, y consecuentemente se construirá la medida relativa correspondiente a esa medida absoluta.

Una vez más: la **utilización de determinadas medidas** es el resultado de una **decisión del investigador** y surge de considerar las características de ese particular conjunto de datos que se está analizando.

Actividad Nº 3

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad **.**

2.2. Dispersión para variables categóricas

Como es de suponer, la construcción de una ***medida de dispersión*** para variables categóricas (nominales u ordinales) no se basa en el desvío de los datos individuales a una medida de tendencia central; su lógica es totalmente diferente. En estos casos, ***¿cómo entenderíamos y valoraríamos diferentes situaciones de dispersión?***

Supongamos que se observan seis "individuos" en una variable con dos categorías: Cat1 y Cat2 de una escala nominal u ordinal. Tendríamos así situaciones de:

- **Dispersión Nula (máxima concentración**): cuando todas las observaciones corresponden a ***una sola de las categorías*** posibles. Es decir alguna de las siguientes dos situaciones.

Situación A

Variable	nº individuos
Cat1	6
Cat2	0
Total	6

Todos los individuos presentan la característica Cat1

Situación B

Variable	nº individuos
Cat1	0
Cat2	6
Total	6

Todos los individuos presentan la característica Cat2

- **Máxima Dispersión (Mínima Concentración**) las observaciones se distribuyen entre las diferentes categorías de manera tal que, en todas, haya la misma cantidad de casos.

Variable	nº individuos
Cat1	3
Cat2	3
Total	6

- **Dispersión intermedia**: Cuando las observaciones se distribuyen entre las categorías de modo desigual pero sin llegar al extremo de concentrarse todas en una sola de ellas. Por ejemplo; situaciones como las siguientes:

Variable	nº individuos
Cat1	4
Cat2	2
Total	6

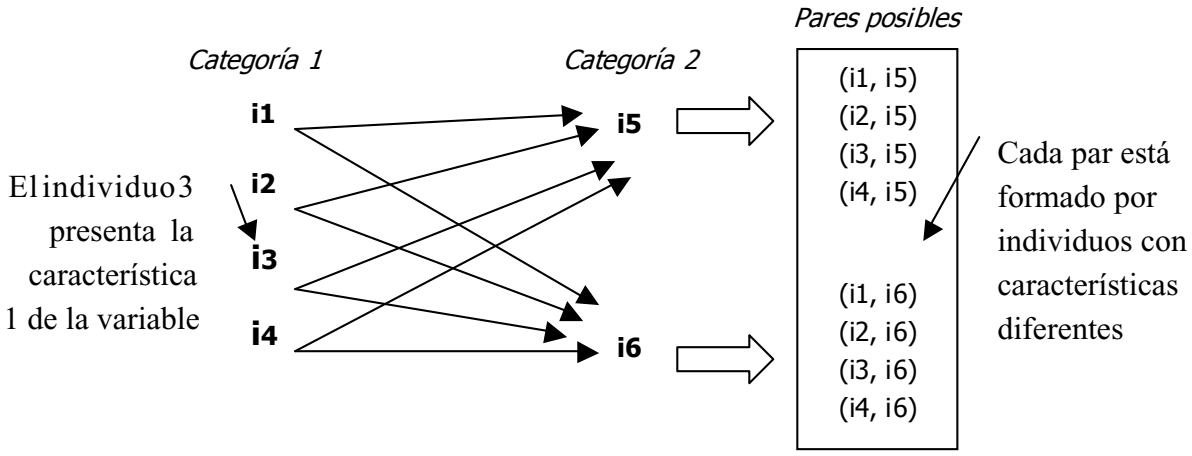
Algunas de las categorías tiene más casos que las otras

Variable	nº individuos
Cat1	1
Cat2	5
Total	6

A partir del concepto de dispersión para datos categóricos, podemos ver la ***lógica que sirve de base para la construcción del Índice de Dispersión***.

El índice de dispersión para una variable de *dos categorías* se obtiene a partir del número de pares de individuos [5] que se pueden construir combinando los elementos de una categoría con todos los de otra. Hay que tener en cuenta que, en este caso, **cada par es una combinación de individuos diferentes** en términos de la variable que se está analizando. Por ejemplo, si se tratara de la variable sexo, cada par estaría integrado por un hombre y una mujer. Así, para una variable cuya distribución presenta cuatro individuos en una categoría y dos en la otra, los pares que se pueden formar serían:

[5] Si la variable tiene tres categorías serán ternas, si tiene cuatro serán grupos de cuatro individuos y así siguiendo.



En la tabla siguiente resumimos, para el ejemplo de seis observaciones en una variable de dos categorías, el número de ***pares posibles de elementos con atributos diferentes que se pueden construir para cada nivel de dispersión***.

Nivel Dispersión	nº individuos en Cat1	nº individuos en Cat2	Nº pares posibles
Nula	6	0	0
Intermedia 1	5	1	5
Intermedia 2	4	2	8
Máxima	3	3	9

En la tabla anterior se puede ver que, *a medida que **crece el nivel de dispersión** de la variable, **aumenta el número de pares posibles** a construir.*

Se observa que la situación de máxima dispersión se corresponde con el mayor número de pares posibles y que la dispersión nula no permite construir ningún par. En consecuencia, el número de pares de diferentes elementos podría constituir una medida absoluta de la heterogeneidad de los individuos en términos de la variable en estudio.

Es posible entonces usar esta relación para construir una *medida relativa* de *dispersión*, de tal manera que sea útil para comparar distintas distribuciones.

Indice de Dispersión (ID)

Se define como el cociente entre el número de pares que corresponde a la distribución observada, sobre el número de pares posibles que corresponde a la situación de máxima dispersión (igual distribución de casos entre las categorías). Por lo tanto; el índice varía entre 0 y 1.

$$0 \leq \mathbf{ID} \leq 1$$

Donde:

ID = 1 en la situación de máxima dispersión (o mínima concentración),

ID = 0 en la situación de dispersión nula (o total concentración).

Si consideramos como distribución observada una de las que en el ejemplo hemos llamado *situación intermedia* (intermedia 2), el índice resulta:

$$\mathbf{ID} = \frac{\text{nº pares observados}}{\text{nº pares posibles en situación de Máx. Dispersión}} = \frac{8}{9} = 0,89 \quad \text{u} \quad 89\%$$

Cuando el número de categorías y/u observaciones es relativamente grande, la determinación del número de pares posibles y de pares observados se vuelve dificultoso. En estos casos el **ID** se determina mediante la siguiente fórmula :

$$ID=\frac{k(n^2-\sum f_i^2)}{n^2(k-1)}$$

donde:

k : número de categorías de la variable

n : total de casos

fi : cantidad de observaciones o frec. Abs. en la categoría i-ésima.

Veamos la utilidad de este índice para comparar la heterogeneidad del motivo de la búsqueda de trabajo entre los hombres y las mujeres.

Motivo de la búsqueda de trabajo por sexo - Posadas-1986.

Motivo de Búsqueda	**Varones**	**Mujeres**
Completar Ingreso Familiar Básico	1140	262
Ampliar Ingreso Familiar Básico	452	490
Otros Motivos	578	702
Total	**2.170**	**1.454**

Fuente: EPH, mayo 1986.

Para la desviación de los varones el índice resulta:

$$ID=\frac{3[(2170)^2-(452^2+1140^2+578^2)}{2170^2(3-1)}=0,91$$

En el caso de las mujeres será:

$$ID=\frac{3[(1454)^2-(262^2+490^2+702^2)}{1454^2(3-1)}=0,93$$

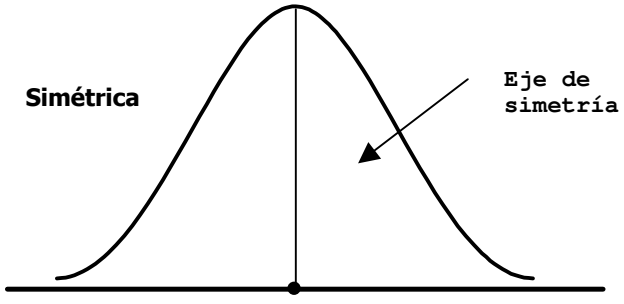
Ambos grupos presentan una alta dispersión (ID cercano a 1). Dado que el ID de las mujeres es mayor, *"las mujeres son ligeramente más heterogéneas que los hombres en cuanto al motivo por el que buscan trabajo".*

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.

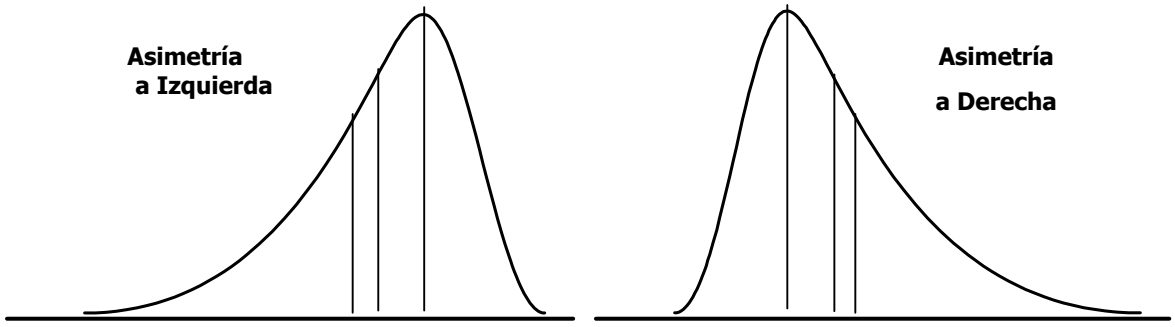
3. ¿Cómo Medir la Asimetría?

Como señaláramos oportunamente la "silueta" de la forma de la distribución (polígono de frecuencias) nos da una idea acerca de la simetría del conjunto de datos. Así teníamos que, en la situación de simetría, cada mitad de la curva es una imagen espejada de la otra mitad y la recta que hace de "espejo" (eje de simetría) es la que pasa por las medidas de tendencia central (media, mediana y modo, que coinciden en el mismo valor).



$\overline{x}$=Ma=Mo

A medida que la distribución se hace más asimétrica hacia uno u otro lado (derecha e izquierda), las medidas de tendencia central tienden a alejarse unas de otras, siendo la media -por estar afectada por los valores extremos- la que más se desplaza hacia la cola de la distribución (ver gráficos siguientes).



$\bar{x} < Ma < Mo$ $\qquad$ $Mo < Ma < \bar{x}$

Vemos en los Gráficos que, en el caso de una asimetría a la izquierda, la media es menor que la mediana y esta a su vez, menor que el modo. Inversamente, en la asimetría a derecha será el modo asume el menor valor y la media la mayor de las tres medidas. Se puede ver además que la mediana, siempre toma un valor intermedio entre las otras dos medidas, ubicándose más próxima a la media [6].

A medida que la **asimetría crece** en una u otra dirección, también las **distancias entre la media y el modo, y la media y la mediana, crecen**. En consecuencia, podemos utilizar estas diferencias ($\bar{x}$- Mo , o $\bar{x}$- Ma) como **medidas absolutas de la asimetría de una distribución**. Además se puede ver que si la asimetría es a la izquierda, $\bar{x}$- Mo dará un valor negativo, en tanto que si la asimetría es a la derecha esta diferencia será positiva.

En síntesis:

$$\bar{x}\text{-Mo} = 0 \Rightarrow \textbf{Simetría}$$

$\bar{x}$-Mo < 0 $\Rightarrow$ **Asimetría negativa** $\qquad$ $\bar{x}$-Mo > 0 $\Rightarrow$ **Asimetría positiva**

Además, cuanto mayor sea el valor absoluto de la diferencia, mayor será el grado de asimetría de la distribución

$$\text{A mayor } |\bar{x}\text{-Mo}| \quad \Rightarrow \quad \text{mayor asimetría}$$

Para poder **comparar la asimetría** de distribuciones de variables medidas en distintas escalas o presentadas para valores con distinta magnitud, la solución es **construir medidas relativas** de asimetría.

3.1. Coeficiente de asimetría de Pearson

Una de las medidas de asimetría más difundidas, es el **Coeficiente de Asimetría de Pearson** que calcula esa diferencia en cantidad de desvíos estándar.

Coeficiente de Asimetría de Pearson **(CAP)**

Se define como:

$$CAP = \frac{\bar{x} - Mo}{\sigma}$$

[6] En casos de asimetría moderada, la mediana se ubica -próxima a la media- a un tercio de la distancia entre la media y el modo.

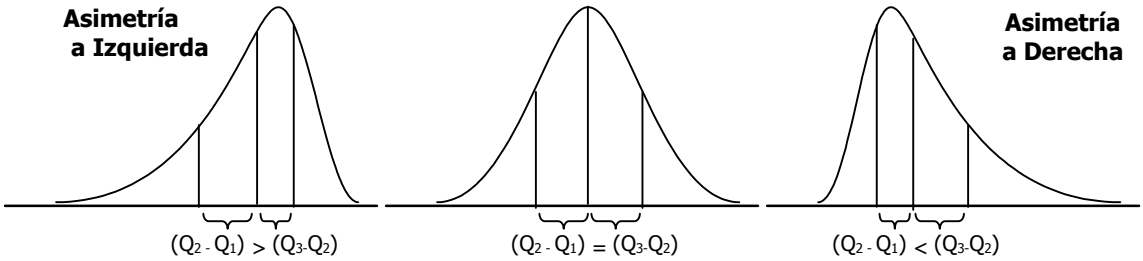
Comentarios

- La **magnitud absoluta** del coeficiente indica la **"cantidad de desvíos estándar"** a los que se encuentra la media del modo.
- Se lo puede expresar en porcentaje, multiplicando por 100 el resultado de la expresión anterior.
- Si el coeficiente es ***igual a cero***, estamos en una situación de ***simetría perfecta***.
- En situaciones de ***asimetría***, el coeficiente puede tomar valores positivos o negativos:
  - Los valores ***positivos*** están indicando una **asimetría a la *derecha***.
  - Los valores ***negativos*** indican una **asimetría a la *izquierda***.
- En términos teóricos, este coeficiente puede tomar valores que **varían entre –3 y +3**.

3.2. Coeficiente intercuartílico de Bowley

Una medida alternativa del grado de asimetría se puede plantear a partir de las distancias que se observan entre los cuartiles. En una situación de simetría los cuartiles 1 y 3 estarán equidistantes de la mediana. Es decir: $Q_3-Q_2 = Q_2-Q_1$

Ahora bien, si la ***distribución es asimétrica, estas distancias no serán iguales*** y variarán con el grado de asimetría; en consecuencia, las diferencias entre estas distancias pueden usarse como base para medir la asimetría de una distribución.



Tomando en cuenta esta característica de las distancias intercuartílicas, Bowley propone una medida relativa que expresa estas diferencias en términos del recorrido intercuartílico.

Coeficiente intercuartílico de Bowley (CAB)

Se define como:

$$CAB=\frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}$$

De esta expresión, se deduce otra más sencilla para el cálculo manual.

$$CAB=\frac{Q_3+Q_1-2.Q_2}{Q_3-Q_1}$$

Comentarios:

- En situaciones de *asimetría*, el coeficiente puede tomar valores positivos o negativos:
  - Los valores *positivos* están indicando una asimetría a la *derecha*.
  - Los valores *negativos* indican una asimetría a la *izquierda*.
- En términos teóricos este coeficiente puede tomar valores que ***varían entre –1 y +1***.
- Según Bowley:
  - un valor de 0,1 (o –0,1) puede considerarse una ***asimetría moderada;***
  - un valor de 0,3 (o –0,3) puede considerarse como una ***marcada asimetría***.
- El coeficiente es *igual a cero*, en una situación de *simetría perfecta*.
- El coeficiente será 1 (o –1) cuando el Q1 (o Q3) coincida con la mediana.

Como parte de un estudio de medición de audiencia radial, se llevó a cabo una encuesta a 150 hogares de la ciudad para medir el tiempo de escucha de dos radios locales, entre las 16 y las 19 horas. Los resultados de esta observación se presentan en las tablas siguientes:

FM Guaraní

Tiempo de escucha (minutos)	Hogares (nº)
0 – 15	14
15 – 30	18
30- 45	20
45 – 60	25
60 – 75	45
75 – 90	18
90- 105	7
105 – 120	3
TOTAL	**150**

FM Acuario

Tiempo de escucha (minutos)	Hogares (nº)
0 – 15	3
15 – 30	45
30- 45	25
45 – 60	20
60 – 75	18
75 – 90	18
90- 105	14
105 – 120	7
TOTAL	**150**

MEDIDA	FM Guaraní	FM Acuario
$\overline{x}$	54,1 min	52,5 min
Ma	59,1 min	46,9 min
Mo	66,3 min	28,4 min
Q1	34,1 min	26,5 min
Q3	71,8 min	76,3 min
σ	25,8 min	28,9 min

El promedio de escucha en ambas radios es similar, aunque es de destacar que la mitad de los oyentes de radio Guaraní escuchan aproximadamente una hora o menos en esa franja horaria, mientras que la mitad de la audiencia de FM Acuario no excede los 47 minutos. Se destaca la diferencia en los tiempos más frecuentes de escucha (66 min. en Guaraní, y 28 min. en Acuario).

La heterogeneidad de los tiempos de audiencia es levemente mayor en FM Acuario (CVg = 0,48 y CVa = 0,55). A su vez, la distribución de los tiempos de escucha en FM Guaraní tienden a concentrarse en los valores más altos, mientras que los de FM Acuario en los valores más bajos; esto se manifiesta en los coeficientes de asimetría (negativo para el primer caso y positivo en el segundo). Además, es mayor el grado de asimetría en FM Acuario (0,83 veces el desvío estándar).

$$CAPg=\frac{54,1-66,3}{25,8}=-0,47 \qquad CAPa=\frac{52,5-28,4}{28,9}=0,83$$

Si ***analizamos la asimetría en el 50% central*** de los tiempos de escucha de ambas radios, se aprecia que en el caso de FM Guaraní es marcada la asimetría a izquierda en el grupo central, en tanto que en FM Acuario es moderada y a derecha.

$$CABg=\frac{(71,8-59,1)-(59,1-34,1)}{71,8-34,1}=-0,33 \qquad CABa=\frac{(76,3-46,9)-(46,9-26,5)}{76,3-26,5}=0,18$$

IMPORTANTE

Las diferencias entre el **coeficiente de Pearson y el de Bowley** están expresando con claridad que, aun cuando ambos miden asimetría, lo hacen sobre la base de criterios diferentes: el primero mide la asimetría de toda la distribución, mientras el segundo se refiere únicamente a los datos centrales. En consecuencia **aportan información complementaria** sobre esta característica de la distribución.

Actividad Nº 5

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad***.**

4. ¿Qué Hemos Visto? (*)

En esta unidad hemos avanzado en la descripción de la forma de una distribución, presentando herramientas que nos permiten medir dos características centrales: ***variabilidad y asimetría***. Estas medidas **complementan las medidas resumen** presentadas en el Capítulo anterior.

Así entonces, hemos presentado medidas de dispersión para variables numéricas que se construyen sobre la base de diferentes **criterios: rango o campo de variación de los datos, y distancia de las observaciones a una medida de tendencia central** que se toma como referencia. Surgen entonces una serie de **medidas que expresan la cantidad de variabilidad en términos absolutos***.*

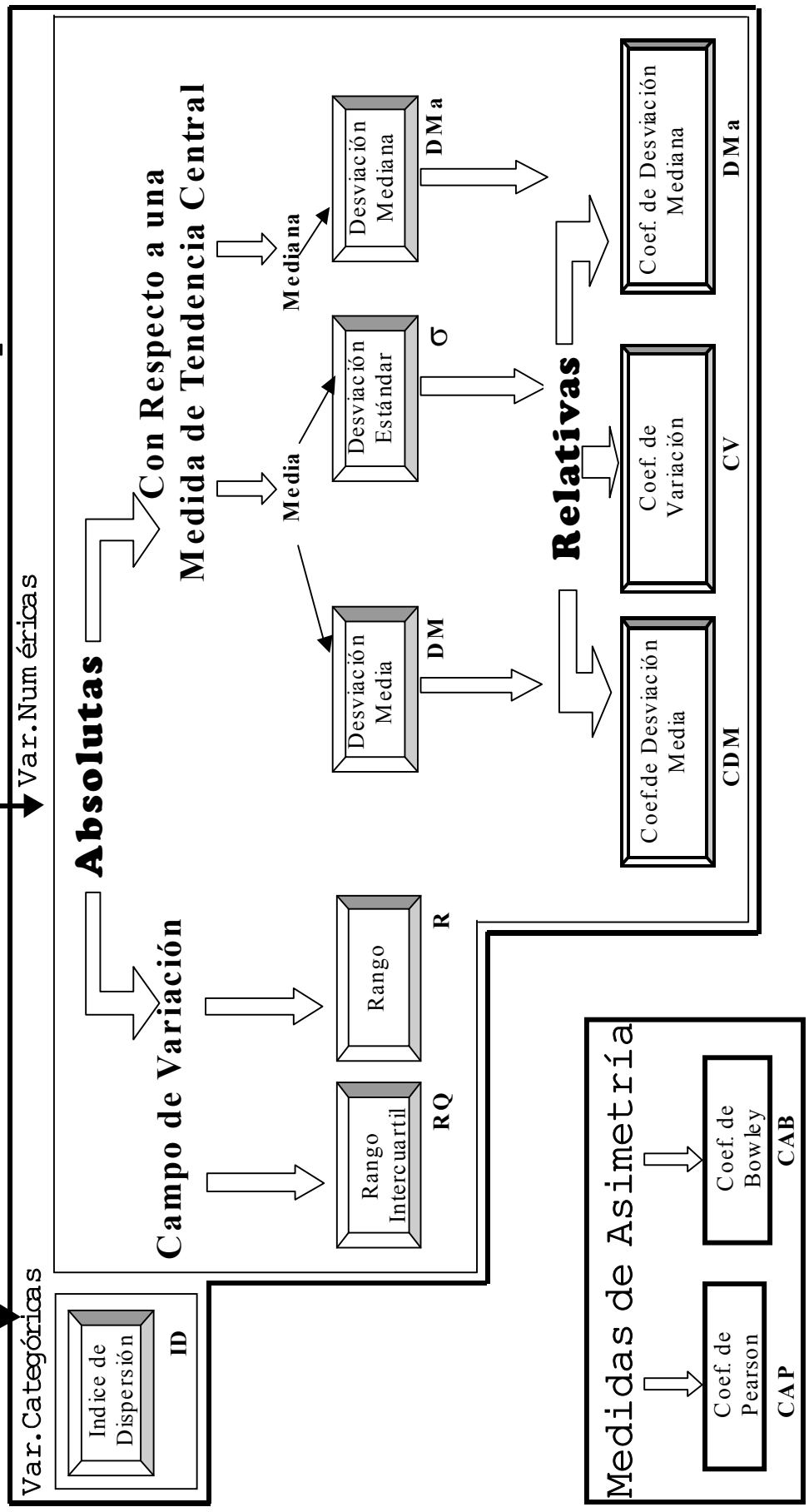
Para resolver cuestiones de **comparabilidad** de diferentes distribuciones, presentamos además **medidas relativas de dispersión**, transformando las principales medidas absolutas.

También, para medir la dispersión de **variables categóricas**, propusimos un **Índice de dispersión**.

Finalmente, presentamos medidas que valoran y permiten comparar el **grado de asimetría** de distintas distribuciones.

En todos los casos, se analizó la variabilidad o la asimetría, en ejemplos que ayuden a la interpretación y comunicación de estas herramientas de análisis, destacando su complementariedad con otras herramientas de análisis.

(*) Ver esquema en la página siguiente


Medidas de Variación o Dispersión
Var.Numéricas
Var.Categóricas
Indice de Dispersión
ID
Absolutas
Con Respecto a una Medida de Tendencia Central
Media
Mediana
Desviación Media
DM
Desviación Estándar
σ
Desviación Mediana
DMa
Campo de Variación
Rango
R
Rango Intercuartil
RQ
Relativas
Coef.de Desviación Media
CDM
Coef. de Variación
CV
Coef. de Desviación Mediana
DMa
Medidas de Asimetría
Coef. de Pearson
CAP
Coef. de Bowley
CAB

Bibliografía

BARBANCHO, A. (1978): *Estadística Elemental Moderna*. Ed. Ariel, Barcelona, España. Páginas: 145-146.

BLALOCK, H. M. (1986): *Estadística Social*, México, FCE. Páginas: 90 a 102.

SHAO, S. (1967): *Estadística Para Economistas y Administradores de Empresas*. Herrero Hermanos S.A., México. Páginas: 218 a 237.

UNIVERSIDAD NACIONAL DE CÓRDOBA (1993): *Estadística aplicada a la Investigación. Curso a distancia*. Fac. de Cs. Económicas, Córdoba, 1993. Módulo IV. Páginas: 3-16.

Conceptos Centrales

- Variabilidad / Dispersión.
- Necesidad de medir la variabilidad.
- Criterios para construir medidas absolutas de dispersión para variables numéricas
- Necesidad de utilizar medidas relativas de dispersión o variabilidad.
- El concepto de dispersión para variables categóricas y la medición asociada.
- Concepto de Asimetría y criterios para su medición.

Habilidades

- Seleccionar y obtener las medidas de variabilidad más apropiadas a una situación de trabajo.
- Interpretar las diferentes medidas en términos del problema.
- Comparar la variabilidad de diferentes distribuciones.
- Seleccionar y obtener medidas de asimetría.
- Interpretar las diferentes medidas de asimetría.
- Describir la forma de una distribución integrando las diferentes medidas de resumen conocidas.
- Comunicar los resultados del análisis.

UNIDAD 5: EL ESTUDIO DE LA RELACIÓN ENTRE VARIABLES

1. ¿Por qué Estudiar la Relación entre Variables?

Como habíamos señalado oportunamente [1] cuando se inicia una investigación se formulan interrogantes que nos remiten al análisis de una, dos o más variables. En las unidades anteriores hemos desarrollado las herramientas necesarias para el estudio univariado, que resulta una etapa insoslayable en el análisis de los datos, y que nos permitió una primera aproximación a la comprensión del fenómeno en estudio, respondiendo así algunas preguntas iniciales.

En el análisis de los estudiantes de Estadística, a partir de esa primera exploración es posible responder: *¿es heterogéneo el grupo en cuanto a la edad?; ¿hay predominio de mujeres?; ¿sus padres han alcanzado el nivel universitario?; ¿se trata de estudiantes provenientes de hogares de bajos ingresos?*, etc.

Estamos ahora en situación de poder avanzar en nuestro análisis y abordar cuestiones que ofrecen un mayor interés de investigación, en tanto permiten encontrar alguna explicación –al menos parcial- de ciertos hechos, poder predecir el comportamiento de algunas características a partir del conocimiento de otras, contrastar algunas hipótesis de investigación que vinculan dos variables, etc. En definitiva, lo que nos proponemos en esta etapa de la investigación es ***analizar para un mismo conjunto de individuos la relación que existe entre las variables***.

En términos concretos y en relación con los estudiantes de Estadística, resulta de interés en esta

- ✓ ¿Difiere el nivel de ingresos según sea el lugar de residencia de los padres?
- ✓ A mayor ingreso del hogar de los estudiantes mayor nivel de estudios del padre.
- ✓ Entre los hombres, ¿es más frecuente encontrar estudiantes con estudios superiores previos a la carrera que cursan actualmente?
- ✓ Las mujeres, ¿dedican más tiempo a mirar televisión?
- ✓ A mayor edad es menor la cantidad de horas dedicadas a mirar TV.
- ✓ A medida que decrece la edad, decrece también el tiempo que se dedica al estudio.
- ✓ etc.

Todas estas preguntas encontrarán respuesta a partir de un análisis bivariado.

> Según una encuesta de Gallup realizada en julio de 2000, el 41% de los argentinos manifestaba temor al desempleo. Este temor "*aumenta a medida que disminuyen el poder adquisitivo (clase baja, 51%, contra 17% de las clases alta y media alta) y el nivel de educación de los encuestados (46% entre aquellos con educación primaria y 33% en aquellos con estudios secundarios), entre los más jóvenes (48% entre los menores de 35 años) y los residentes en el interior y el conurbano (43%, en promedio, contra 29% de la Capital Federal)*". (Diario La Nación, 06/08/2000).
>
> A partir de una encuesta dirigida por la Sociedad de Estudios Laborales (SEL), se pudo saber que "*el promedio de los egresados universitarios y terciarios gana 1.158 pesos. Y aquí un dato llamativo: al discriminar las cifras por sexo, los hombres perciben una media de 1.648 pesos, mientras que las mujeres apenas alcanzan a 878 pesos*". (Diario La Nación, 8/8/2000).

[1] Ver en la Unidad 2 el apartado: "3. El Análisis de la Matriz de Datos".

Conclusiones como las presentadas precedentemente son el resultado de haber realizado un análisis bivariado. Intentando responder en el primer caso preguntas como: ¿varía el temor al desempleo según sea el nivel de educación de los encuestados?; ¿y entre los diferentes grupos de edad? ¿y según sea el lugar de residencia? En el segundo caso, además de querer conocer el nivel de ingresos de los universitarios en general, la pregunta a responder era: ¿hombres y mujeres, perciben ingresos diferentes?

Al ***analizar la relación*** entre variables hay tres aspectos a considerar:

- ✓ la **existencia** de relación (¿*hay relación*?)
- ✓ la **forma** en que se produce esa relación (¿*cómo se da*?)
- ✓ la **fuerza** de la relación (¿*cuán intensa es*?)

Lo que se puede observar en los ejemplos anteriores, es que **existe** una relación entre las variables:

En el primero, se observa que **existe relación** porque al variar el nivel económico de los individuos también varía la incidencia del temor a la desocupación; la **forma** queda expresada al decir que, "el temor aumenta cuando disminuye el nivel económico". En el texto no aparece una valoración de la intensidad.

En el segundo estudio, se aprecia que **hay una relación** entre el sexo y el nivel de ingresos, dado que según sea el sexo varía el nivel de ingreso; para caracterizar la **forma** de esa relación se puede decir que "en promedio, los ingresos resultan menores para las mujeres". Tampoco aquí se valora explícitamente la intensidad de esa relación.

Relación entre variables

En términos generales podemos hablar de una relación entre variables, cuando en un mismo conjunto de individuos se observa un comportamiento sincrónico o coordinado en el comportamiento de las mismas (al cambiar los valores de una variable cambian al mismo tiempo y de manera determinada, los valores de la otra).

En el estudio de la relación entre dos variables, podemos explorar la existencia o no de una relación, o bien si tuviera sentido, determinar si una de las variables explica o causa los cambios registrados en la otra. En el último caso existiría una variable "explicada" o "respuesta" y una variable "explicativa". (Moore, 1998).

A las variables **explicativas** se las reconoce también como **independientes**, en tanto que a las **variables respuesta** como **dependientes**.

Var. ***respuesta o dependiente***: mide el resultado de un estudio.

Var. ***explicativa o independiente***: intenta explicar los resultados observados.

En el estudio de Gallup citado anteriormente, la edad, el nivel de educación y el nivel económico serían variables que "explican" los niveles registrados de la variable en estudio: el temor al desempleo (variable respuesta o dependiente). Los conceptos de variables explicativas o explicadas suponen el control de algunas variables a través de experimentos.

IMPORTANTE

En las Ciencias Sociales, no se realizan experimentos como en otras ciencias en las cuales se puede efectuar un control estricto de las variables explicativas. Los valores de las distintas variables simplemente son observados y -en estos casos- puede existir o no una relación de causa-efecto entre las variables cuya relación se estudia.

Para iniciar un análisis bivariado, es necesario **considerar dos aspectos centrales** que hacen a cuestiones de diferente orden:

- ✓ la **naturaleza de la relación** entre las variables;
- ✓ el **tipo de variables** que se están analizando.

En cuanto a su naturaleza, según Barbancho [2] se pueden identificar los siguientes tipos de relaciones entre variables:

a) **Dependencia causal unilateral**: en este caso, una variable influye a la otra pero no al contrario. Ej: la cantidad de lluvia influye en el rendimiento del trigo; el nivel de educación en la preferencia del tipo de lectura; el nivel de ingresos en la selección del lugar de alojamiento; etc.

b) **Interdependencia**: la influencia es recíproca, y se produce por lo tanto en las dos direcciones; hay dependencia causal bilateral. Ej.: el precio de un producto en el mercado y la cantidad demandada de ese producto; la posición frente al aborto y la afiliación política; la elección de un lugar de vacaciones y el medio de transporte utilizado; etc.

c) **Dependencia indirecta**: dos variables pueden estar relacionadas por la intervención de una tercer variable que influye en ambas. Ej.: la tasa de natalidad y el consumo de proteínas de origen animal (la tercera variable sería el nivel de vida); el número de accidentes de tránsito y la cantidad de semáforos (esta relación se explica por la concentración urbana); etc.

d) **Covariación casual**: es el caso de dos variables que presentan un comportamiento sincronizado aun cuando esta relación puede ser totalmente casual o accidental. A esta conclusión se llega naturalmente cuando se sabe que entre ambas no existe ningún vínculo directo o indirecto que justifique tal relación observada.

IMPORTANTE

La **decisión sobre la naturaleza de la relación entre las variables es ajena a la Estadística**. Solo es posible determinarla a partir del conocimiento del tema que se está estudiando. Sin embargo, esta definición es fundamental para la interpretación de los resultados.

A su vez, **el *tipo de variables*** [3] que se están analizando ***determinará las herramientas estadísticas*** disponibles. Así tenemos que:

Si se trata de...		**Recurrimos a ....**
Dos variables categóricas	→	Tablas de contingencia
Una variable numérica y una categórica	→	Comparación de medias entre grupos
Dos variables numéricas	→	Análisis de Correlación

En todos estos casos podremos recurrir a alternativas gráficas o numéricas como herramientas de análisis.

Actividad Nº 1

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

[2] BARBANCHO, Alfonso: *Estadística elemental moderna*. Ed. Ariel Barcelona, España, 1978.

[3] Antes de iniciar el desarrollo de cada una de estas herramientas de análisis, creemos conveniente señalar una cuestión de terminología que puede conducir a confusión a un lector desprevenido. Mientras algunos autores utilizan el término ***asociación*** como sinónimo de **relación**, otros reservan el término ***asociación*** cuando se trata de la relación entre variables categóricas y hablan de ***correlación*** para referirse a la relación entre variables numéricas. En la presentación de esta unidad adoptaremos este último criterio.

2. La Relación entre Variables Categóricas

Cualquier análisis estadístico supone la organización y/o resumen de los datos. En el análisis univariado organizábamos los datos en tablas de frecuencias simples, indicando la cantidad (o porcentaje) de individuos que presentaban un determinado valor de la variable.

Ahora bien, si pretendemos responder preguntas del tipo:

- ✓ ¿Cuántas personas de *nivel socioeconómico* alto *opinan* que el servicio eléctrico es bueno?
- ✓ ¿Cuántos *hombres leen frecuentemente el periódico*? Y, ¿cuántas *mujeres*?
- ✓ Entre los que *nunca leen revistas*, ¿cuántos son *hombres*?
- ✓ Entre nuestros estudiantes del curso de Estadística, de los que vienen de colegios privados ¿cuántos son varones y cuántas mujeres?
- ✓ etc.

Tendremos que **describir a los individuos mediante el tratamiento simultáneo de dos variables categóricas**. Ante esta necesidad, nos debemos preguntar:

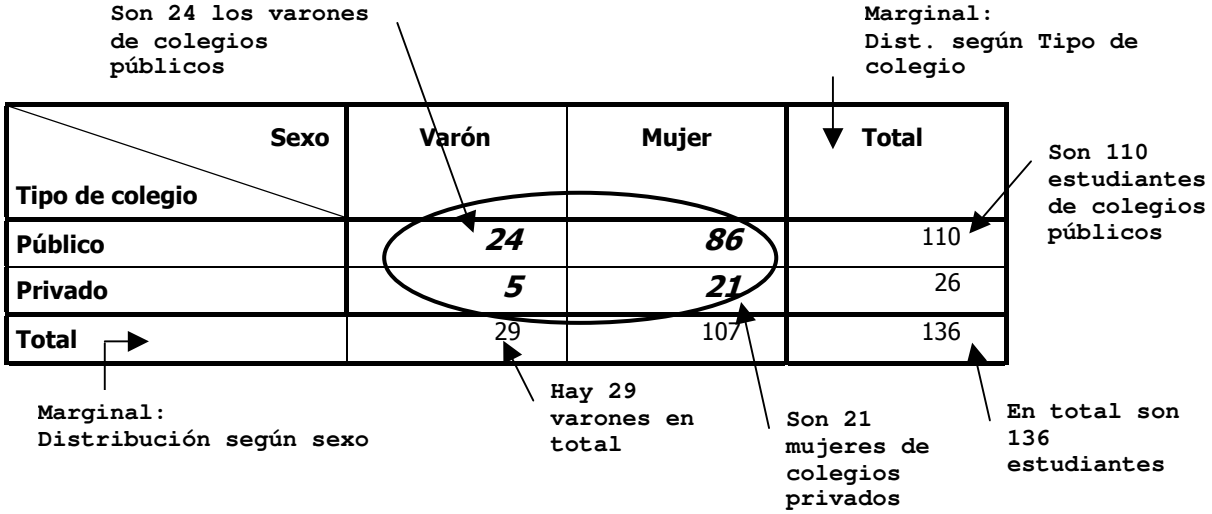
¿Cómo presentar los datos para describir a los individuos a partir de dos variables categóricas simultáneamente?

2.1. El recurso numérico

Si intentáramos responder a la pregunta sobre cantidad de hombres y mujeres que vienen de colegios privados y públicos, podríamos **contar** en la matriz de datos **cuántos individuos cumplen simultáneamente la doble condición de**:

- ser **mujer** y haber asistido a un colegio **público,**
- ser **mujer** y haber asistido a un colegio **privado,**
- ser **varón** y haber asistido a un colegio **público,** y
- ser **varón** y haber asistido a un colegio **privado.**

Si realizado el conteo en la matriz de datos, observamos que fueron 86 las mujeres que asistieron a un colegio público, y 24 los varones; y a un colegio privado asistieron 21 de las mujeres y 5 de los varones, podríamos organizar estos datos en una tabla como la siguiente:



Sexo / Tipo de colegio	Varón	Mujer	Total
Público	***24***	***86***	110
Privado	***5***	***21***	26
Total	29	107	136

Esta forma de organizar los datos se conoce como ***tabla de contingencia***. En el ***cuerpo de la tabla*** (zona resaltada) se presenta la **distribución conjunta** que da cuenta del número de individuos que presentan cada una de las combinaciones posibles de las categorías de las variables. Se distribuyen así los 136 estudiantes según la doble clasificación: "tipo de colegio" y "sexo".

En toda tabla de contingencia podemos distinguir:

- Los ***Marginales***: corresponden a la última fila y la última columna de la tabla que, encabezados por la palabra "total", presentan la distribución univariada según "sexo" (última fila) y según "tipo de colegio" (última columna). Se puede leer entonces que de nuestros 136 entrevistados, 29 son hombres y 107 mujeres; a la vez que 110 estudiantes asistieron a establecimientos públicos y 26 lo hicieron a privados.
- Las ***Filas***: presentan la distribución de los individuos que vienen de establecimientos públicos o privados según el sexo. En la primera fila, tenemos la distribución según el sexo de los 110 individuos que asistieron a establecimientos públicos.
- Las ***Columnas***: presentan la distribución de varones y mujeres por tipo de colegio. En la primera columna, tenemos la distribución de los 29 varones según el tipo de colegio al que asistieron.
- Las ***Celdas***: consignan las frecuencias correspondientes a la combinación de pares de categorías de las variables. Así, en la segunda celda de la primera fila se puede leer que hay 86 estudiantes que asistieron a establecimientos públicos y son mujeres.

> ***Tabla de contingencia:***
>
> Es una tabla que presenta la distribución de los individuos clasificados según dos variables categóricas simultáneamente.

Hasta aquí sólo hemos presentado la **tabla de contingencia como una forma de organización de los datos cuando se consideran simultáneamente dos variables**. A partir de esta tabla, podemos responder a la pregunta que nos formuláramos inicialmente: ¿cuántos varones y cuántas mujeres vienen de colegios privados?

A los efectos de avanzar en el estudio de las relaciones entre variables nos podemos plantear una situación que permita ilustrar ese proceso de análisis.

En un estudio sobre hábitos alimenticios, una de las cuestiones de interés era conocer sobre el consumo de productos dietéticos. En particular, la investigación se planteaba como hipótesis que existía una mayor preferencia por este tipo de productos entre las mujeres. Se observaron 850 individuos de los cuales reproducimos en forma parcial la matriz de datos con las variables *Sexo* y *Consumo de Productos Dietéticos*.

Matriz (parcial) sobre el consumo de productos dietéticos

Individuos	Sexo	Consumo de Productos Dietéticos
1	Hombre	Consume
2	Hombre	No consume
3	Mujer	Consume
4	Mujer	Consume
5	Hombre	No consume
6	Mujer	Consume
7	Hombre	Consume
8	Mujer	No consume
9	Hombre	No consume
10	Mujer	No Consume
11	Mujer	Consume
12	Hombre	No consume
...	**...**	**...**
850	Mujer	Consume

A partir del conteo de los datos de la matriz, construimos la siguiente tabla de contingencia.

Distribución de los Individuos según Sexo y Consumo de Productos Dietéticos

	Consumo de Productos Dietéticos		
Sexo	**Consumen**	**No Consumen**	**Total**
Hombres	150	300	**450**
Mujeres	350	50	**400**
Total	**500**	**350**	**850**

En los ***marginales*** de la tabla se observa que *los 850 entrevistados se distribuyen en 500 que declaran consumir productos dietéticos y 350 que no lo hacen. A su vez, considerando el sexo, esos mismos 850 individuos se clasifican en 450 hombres y 400 mujeres.*

En el ***cuerpo*** de la tabla (que contiene la *distribución conjunta*) podemos ver que, *del total de individuos observados son: 150 los hombres que consumen productos dietéticos y 300 los que no consumen, 350 mujeres que declaran consumir estos productos y 50 que no lo hacen*.

Ahora bien:

¿cómo valorar si es "importante" la cantidad de hombres no consumidores o de mujeres consumidoras, etc.?

Una alternativa es **apreciar esta información en relación con el total de individuos observados**, lo que conduce a una tabla como la siguiente.

Distribución de los Individuos según Consumo de Productos Dietéticos y Sexo (%)

	Consumo de Productos Dietéticos		
Sexo	**Consumen**	**No Consumen**	**Total**
Hombres	18	35	**53**
Mujeres	41	6	**47**
Total	**59**	**41**	**100 (850)**

Cada uno de los números de la tabla corresponde a un **porcentaje calculado sobre el total de casos** observados (850). Así por ejemplo:

- ✓ El 53% de los entrevistados son hombres.
- ✓ El 59% de los individuos consumen productos dietéticos.
- ✓ El 18% de los casos, son hombres que consumen productos dietéticos.
- ✓ El 6% de los individuos son mujeres que no consumen
- ✓ etc.

Así entonces, esta tabla sirve para describir el porcentaje de individuos que registra cada par de características. En este tipo de tablas es importante consignar:

- ✓ que los **valores corresponden a porcentajes** (se lo puede hacer en el título).
- ✓ el **total de casos** sobre el cual están calculados los porcentajes; generalmente se lo incluye entre paréntesis al lado del 100%.

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***.**

Ahora bien, resuelta la organización de los datos y realizada una primera lectura de los mismos, estamos en condiciones de **estudiar la relación** entre estas dos variables. Estudiar **la existencia** de relación entre las variables nos remite a preguntas como:

- ✓ ¿Es diferente el comportamiento de hombres y mujeres en cuanto al consumo de productos dietéticos?
- ✓ ¿Varía la composición por sexo de los consumidores y no consumidores?

Responder a estas preguntas nos conduce a **dos lecturas diferentes de la tabla**. Así compararíamos:

- la distribución del **consumo entre los hombres** *vs*. el **consumo entre las mujeres** para responder la primera pregunta, y
- la **distribución según sexo entre los consumidores** *vs.* la **distribución según sexo entre los no consumidores** *para la segunda.*

Si observáramos que la **distribución** del consumo **es igual** en hombres y mujeres, concluiríamos que **no existe** relación entre las variables (o las variables son independientes). También ocurriría lo mismo si la distribución por sexo es igual entre consumidores y no consumidores.

La necesidad de comparar nos lleva al **cálculo de porcentajes** (principalmente cuando las subpoblaciones presentan un número de individuos muy diferentes).

Ahora bien:

¿Cómo calcular los porcentajes?, ¿sobre qué total los calculamos?

Para comparar el consumo de hombres y mujeres, tomamos los porcentajes dentro de cada fila. Así, tendremos tres totales de referencia (ó 100%) para cada una de las filas: el total de hombres (450), el total de mujeres (400) y el total de individuos observados (850).

Distribución del Consumo de Productos Dietéticos según Sexo (%)

	Consumo de Productos Dietéticos		
Sexo	***Consumen***	***No Consumen***	**Total**
Hombres	33	67	**100** (450)
Mujeres	87	13	**100** (400)
Total	**58**	**41**	**100** (850)

300.100/450= 67% de los hombres no consumen

Los hombres son en total 450

41% del total de casos son no consumidores

50.100/400= 13% de las mujeres no consumen

Comparando en la Tabla la distribución de los hombres y las mujeres según el consumo, *se hace evidente que el comportamiento varía con el sexo. Puede decirse entonces que* ***existe una relación entre ambas variables*** *o que el* ***sexo y el consumo de productos dietéticos no son independientes****.*

En cuanto a la **forma** en que se da la relación, deberíamos poder responder **cómo es** esa relación:

- ✓ ¿son las mujeres más consumidoras que los hombres?, o ¿son los hombres los que tienden a un mayor consumo de los mismos?

En la tabla, se puede ver que:

"Mientras el 33% de los hombres consume productos dietéticos, en el caso de las mujeres ese porcentaje alcanza el 87%".

Otra manera de expresar la **misma** información que en el párrafo anterior, sería decir:

"Entre los hombres hay un 67% de no consumidores, mientras entre las mujeres este porcentaje es del 13%".

Las expresiones anteriores están indicando de manera implícita que son las mujeres las que presentan una mayor inclinación hacia el consumo de los productos dietéticos (la forma en que se produce la relación).

Actividad Nº 3

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad***.**

Para comparar la composición por sexo de consumidores y no consumidores, tomamos los porcentajes "dentro" de cada columna. Así tenemos tres totales de referencia (ó 100%): el total de consumidores (500), el total de no consumidores (350) y el total de individuos observados (850).

Distribución de los Individuos por Sexo Según Consumo (%)

Sexo	Consumo de Productos Dietéticos: Consumen	No Consumen	Total
Hombres	**30**	**86**	**52** ← El 52% de los individuos son hombres
Mujeres	**70**	**14**	**48**
Total	**100** (500)	**100** (350)	**100** (850)

$\frac{150}{500}.100$ **= 30% de los consumidores son hombres**

$\frac{300}{350}.100$ **= 86% de los no consumidores son hombres**

Dado que:

"*Mientras entre los consumidores, las mujeres representan el 70%, entre los no consumidores de productos dietéticos estas constituyen solo el 14%*" [4].

Nuevamente aquí podemos concluir que **existe relación** entre ambas variables (la composición por sexo de los consumidores es diferente a la composición de los no consumidores), y la **forma** en que se produce esa relación es que **los consumidores son mayoritariamente mujeres**.

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***.**

Pero…

¿cuál es la mejor manera de calcular los porcentajes?

Cualquiera de las dos últimas tablas permiten **apreciar si existe relación** entre las variables. Así, conociendo el sexo de un individuo podemos predecir con buenas posibilidades de acertar si será consumidor de productos dietéticos (ej. si se trata de un hombre puedo predecir que será un no

[4] El resultado de la comparación también puede expresarse como *"El 30% de los consumidores son hombres, mientras entre los no consumidores los hombres constituyen el 86%".*

consumidor y acertaré con esta predicción en 67 de cada 100 casos); a su vez, conociendo que no es consumidor podemos arriesgar, con bastante chance de acertar, cuál será el sexo del individuo.

Si consideramos la necesidad de explicar el comportamiento de una de las variables, tiene sentido pensar que el sexo "explica" el consumo de estos productos, y no que el consumo "explica" el sexo; entonces resulta más apropiada para este caso la tabla en la que se compara el consumo según el sexo (tabla con porcentajes calculados en el sentido de las filas).

En este punto del análisis podríamos plantearnos encontrar una medida o un único valor que resuma la ***fuerza o intensidad*** de la relación entre las variables en estudio, y es indudable que una medida de estas características tiene -entre otras ventajas- la posibilidad de comparar la fuerza de la relación que se observa en distintas tablas.

Una aproximación intuitiva a la **evaluación de la fuerza** de la relación entre las variables en una tabla de contingencia, puede lograrse calculando lo que se conoce como una **diferencia de proporciones o porcentajes**. Para ello, y tomando el ejemplo del consumo de productos dietéticos, se procedería de la siguiente manera: considerando al sexo como variable explicativa debemos comparar el comportamiento de hombres y mujeres, en cuanto al consumo de productos dietéticos. En otras palabras, queremos responder a la pregunta: ¿quiénes presentan mayor tendencia a consumir productos dietéticos: los hombres o las mujeres? Para encontrar respuesta a esta pregunta, habíamos visto que debíamos calcular los porcentajes de consumo sobre el total de hombres y sobre el total de mujeres (en la tabla construida corresponde a "porcentaje en el sentido de las filas").

Así, nos encontrábamos con que "*mientras el 33% de los hombres consume productos dietéticos, en el caso de las mujeres ese porcentaje alcanza el 87%*". En consecuencia, "*entre los hombres se registra un 54% (33%-87%) menos de consumidores que entre las mujeres*". Este último cálculo, que expresa numéricamente la diferencia del consumo entre los hombres y las mujeres, se conoce como diferencia de proporciones.

Distribución del Consumo de Productos Dietéticos según Sexo y Diferencia de proporciones (d)

	Consumo de Productos Dietéticos	
Sexo	**Consumen**	**No Consumen**
Hombres	33	67
Mujeres	87	13
d	**-54**	**54**

La diferencia de proporciones nos indica la fuerza de la relación entre las variables y en términos teóricos puede tomar valores entre 0 y 1 (0 y 100 si se trata de porcentajes). Se puede comprender que, si todas las mujeres son consumidoras y todos los hombres no consumidores (o viceversa), la variable sexo explica totalmente el consumo y la relación es perfecta; en este caso la diferencia de proporciones alcanzaría el valor 1 (100%). Si el comportamiento de hombres y mujeres fuera idéntico (igual proporción de mujeres que de hombres que consumen) estaríamos en una situación de "no-relación" y la diferencia de proporciones sería igual a 0. En síntesis, cuanto mayor es la diferencia de proporciones más fuerte es la relación entre las variables.

$\mathbf{0 \leq d \leq 1}$

Si d=0 $\rightarrow$ se trata de una situación de **independencia** o **no relación** entre las variables.

Si d=1 $\rightarrow$ se trata de una situación de **perfecta relación** entre las variables.

De alguna manera, con la diferencia de proporciones estamos formalizando un proceso que realizamos "naturalmente" al analizar una tabla de contingencia cuando comparamos los porcentajes.

IMPORTANTE

Debe observarse que, **según sea la forma** en que se calculan **los porcentajes** ("consumo según sexo" o "sexo según consumo") las **diferencias obtenidas pueden ser distintas** ya que los marginales no serán necesariamente iguales: no son simétricos. Es decir, **no hay un único valor que resuma la relación** presente en la Tabla. (Determine Ud. la diferencia de proporciones del "sexo según consumo").

Cuando se trate de tablas de **una o ambas variables con más de dos categorías**, hay **más de una diferencia de proporciones** y, en consecuencia, no se obtiene un único número que sintetice la fuerza de la relación.

La Estadística ofrece diversos coeficientes construidos según criterios también diferentes que responden a esta intención, los que no serán tratados en esta presentación dado que escapan a los alcances propuestos para este curso [5].

Actividad Nº 5

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad***.**

2.2. El recurso gráfico

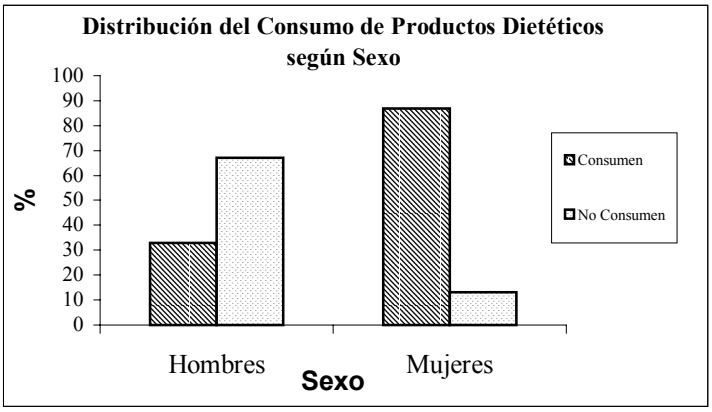
Dado que se trata de variables categóricas, se utilizan gráficos de barras, en el que solo uno de los ejes es numérico. Básicamente pueden distinguirse dos tipos de gráficos:

- ✓ los ***gráficos compuestos***, y
- ✓ los de ***partes componentes***.

En estos gráficos las barras pueden ser horizontales o verticales, y las frecuencias pueden ser absolutas o relativas.

2.2.1. Gráficos compuestos

En este tipo de gráficos, para cada categoría de una de las variables se presenta la distribución de frecuencias según la segunda variable. Cada barra tiene una altura que se corresponde con la frecuencia (absoluta o relativa).



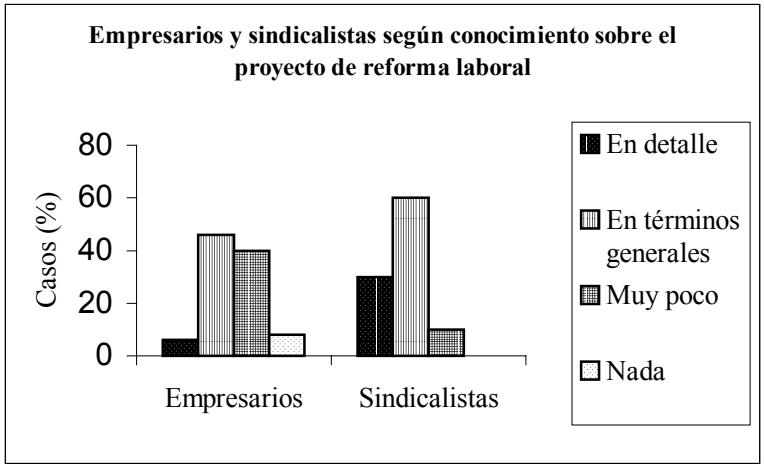
Este gráfico corresponde a la tabla en la que para cada sexo se presenta la distribución (relativa) según el consumo. En consecuencia, el gráfico nos permite comparar la presencia de consumidores y no consumidores en cada sexo.

Se aprecia claramente que *la presencia de consumidores de productos dietéticos es predominante entre las mujeres, mientras entre los hombres son minoría*.

Aún sin contar con la tabla de contingencia, este tipo de gráficos facilita las comparaciones. Así por ejemplo, en el gráfico siguiente se presenta la distribución entre empresarios y sindicalistas, del nivel de conocimiento que tenían sobre el proyecto de reforma laboral; rápidamente se puede ver que entre los sindicalistas el nivel de conocimientos era mayor ("en detalle" y "en términos generales" son

[5] Al lector interesado le sugerimos remitirse a textos que le dedican especial atención a este tema, tal el caso de BARANGER, D.: *Construcción y Análisis de datos*, Editorial Universitaria de la Univ. Nac. de Misiones, Posadas, 2000.

las categorías predominantes), mientras que entre los empresarios alcanza relevancia la categoría "muy poco" e incluso algunos "nada" sabían sobre el proyecto.

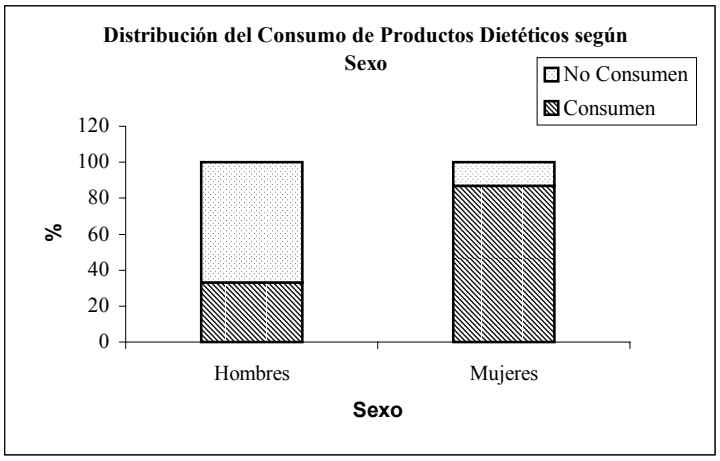


Fuente: elab. propia basándose en datos publicados en el diario Perfil 31/5/98

2.2.2. Gráficos de partes componentes

Es similar al anterior, en el sentido de presentar la distribución de una de las variables dentro de cada categoría de la segunda. Se los puede representar en términos absolutos o relativos y la altura de cada barra se corresponde con la frecuencia absoluta o el 100%.

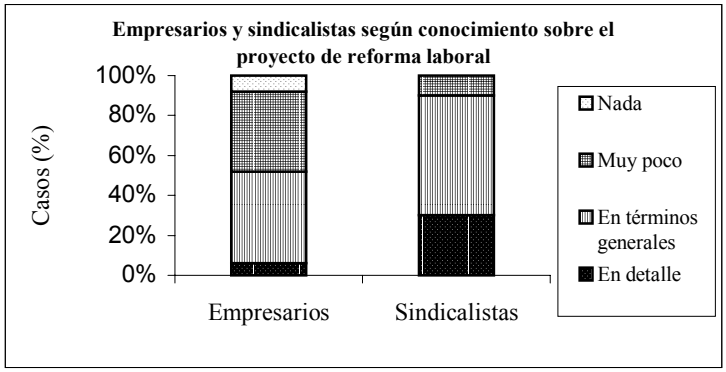
Cada barra es subdividida en tantas categorías como tiene la otra variable. La altura de cada subdivisión se corresponde con la frecuencia absoluta (o relativa) de la categoría correspondiente.



Una vez más, el gráfico muestra claramente la *importancia que tiene entre los hombres la categoría "no consumidores de productos dietéticos", mientras que entre las mujeres esa categoría es de poca importancia*.

Para el ejemplo del conocimiento de empresarios y sindicalistas sobre el Proyecto de Reforma Laboral, el gráfico compuesto sería el que se presenta.

Este tipo de gráficos **pierde su capacidad de favorecer las comparaciones cuando crece el número de categorías** de una o ambas variables.



Fuente: elab. propia basada en datos publicados en el diario Perfil 31/5/98

Actividad Nº 6

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 6 de la Guía de Actividades correspondiente a esta unidad***.**

3. La Relación entre Variables Categóricas y Numéricas

Es muy frecuente que nos formulemos preguntas del tipo:

- ✓ ¿Los salarios que perciben las mujeres difieren del que perciben los hombres?
- ✓ ¿El rendimiento escolar de los estudiantes en el examen de Lengua varía según se trate de escuelas rurales o urbanas?
- ✓ ¿El gasto en regalos y *souvenir* difiere según la forma de organización del viaje de los turistas (cuenta propia o tours)?
- ✓ ¿El número de hijos por familia es distinto según sea el nivel socioeconómico?

Buscar respuestas a estos interrogantes nos conduce al análisis de la relación entre una variable cualitativa y una cuantitativa. Ahora bien,

¿Cómo se manifestaría la existencia de una relación entre una variable categórica y una variable numérica?

Por ejemplo, podemos decir que, si encontramos que un gasto alto en *souvenir* y regalos se corresponde con una cierta forma de organización del viaje, y viceversa, para una cierta forma de organización del viaje es probable observar un gasto elevado en regalos y *souvenir*, entonces diríamos que las variables "gasto en regalos y *souvenir*" y "forma de organización del viaje" están relacionadas. En fin, se busca en este caso, identificar si la forma de organización del viaje de los turistas, explica –en alguna medida- el gasto en regalos y *souvenir* que los turistas hacen.

En términos generales, **en este tipo de análisis intentaríamos ver si los valores de la variable numérica al ser reagrupados según las categorías de la segunda variable, constituyen clases diferentes entre sí**.

Por ejemplo, un mayor número de hijos en las familias de Nivel Socioeconómico Bajo que en las de nivel Medio y Alto; un rendimiento escolar más alto en las escuelas urbanas que en las rurales; un ingreso más alto entre los hombres, etc.

Desde esta perspectiva, el problema nos remite a **resumir la información de manera de poner en evidencia la existencia o no de este comportamiento en las variables** en estudio.

3.1. El recurso numérico

La idea entonces es ***comparar la distribución de la variable numérica entre tantas clases o grupos como categorías tenga la variable cualitativa***. En este sentido valen todas las herramientas presentadas para el análisis univariado.

Análisis de la relación

Para analizar la relación entre una **variable cuantitativa y una cualitativa**, se comparan las **distribuciones de la variable numérica entre las clases definidas por las categorías de la variable cualitativa**. *Para ello se utilizarán* **las medidas de tendencia central más representativas**.

En general, en la literatura estadística clásica se propone a la media aritmética como medida de comparación.

A los efectos de ejemplificar el razonamiento propio de este análisis, nos proponemos estudiar la relación entre el "Nivel de Estudios del Padre"[6] de nuestros estudiantes de Estadística, y el "Ingreso Familiar". A continuación presentamos la distribución del ingreso familiar para cada una de las subpoblaciones que quedan determinadas por las categorías de la variable "estudios del padre".

Nivel de Estudios Padre	n	Mín.	Máx.	Media	Mediana	Desv. Estándar	CV	Asimetría
No terminaron Primario	23	145	1300	475,4	400,0	286,5	60,3	0,79
Completaron Primario y no Secundario	57	80	2000	621,6	500,0	428,1	68,9	0,85
Completaron Secundario o más	22	200	2000	956,8	800,0	647,2	67,6	0,73

Tallo – hoja: Ingreso familiar según Nivel de Estudios del Padre

No terminaron Prim. (1)

```
Frec.  Tallo -  Hoja

 1      1 .  4
 4      2 .  0005
 4      3 .  0004
 6      4 .  000005
 3      5 .  005
 0      6 .
 0      7 .
 3      8 .  000
 2 Extremos    (>=1000)

Ancho del Tallo:  100
Cada Hoja:  1 caso(s)
```

Complet. Prim y no Secund. (2)

```
Frec.  Tallo -  Hoja

 5      0 .  01111
14      0 .  22333333333333
17      0 .  44445555555555555
 6      0 .  667777
 4      0 .  8889
 5      1 .  00001
 1      1 .  3
 2      1 .  55
 3 Extremos    (>=1600)

Ancho del Tallo:      1000
Cada Hoja:       1 caso(s)
```

Complet. Secund. o más (3)

```
Frec.  Tallo -  Hoja

 9    0 .  233344444
 3    0 .  888
 4    1 .  0000
 2    1 .  68
 4    2 .  0000

Ancho del Tallo:      1000
Cada Hoja:       1 caso(s)
```

(1) Incluye a quienes nunca asistieron o tienen Primario incompleto
(2) Incluye a quienes completaron el primario o tienen secundario incompleto
(3) Incluye a quienes completaron el secundario, o iniciaron o completaron un nivel de superior de educación.

De la comparación de las medidas de tendencia central presentadas en la tabla anterior, podríamos concluir que **existe una relación** entre el nivel de estudios del padre y el ingreso de la familia ya que es importante la diferencia tanto entre las medias como entre las medianas de los tres grupos. Además, esa relación se da ***de la forma:*** "*a un mayor nivel de estudios le corresponde, en promedio, un mayor nivel de ingresos*"[7].

De la observación del diagrama tallo-hoja surge que las tres clases o grupos presentan concentraciones de los ingresos en los primeros tramos y, también en todos los casos, algunos pocos valores atípicos de ingresos altos. En consecuencia, las tres distribuciones tienen algún grado de asimetría a la derecha. En todas ellas la media aritmética aparece "alejada de la tendencia central" en un mismo sentido (hacia la derecha).

Esta apreciación se expresa numéricamente en el cuadro anterior, donde el coeficiente de Asimetría de Pearson indica una asimetría bastante similar entre ellas, con el mayor valor para el grupo con estudios intermedios. También a este grupo le corresponde la mayor dispersión en términos relativos.

Existen medidas que permiten cuantificar la **fuerza** de la relación entre las variables, entre las que merece destacarse la denominada "**razón de correlación**". La lógica que subyace a la construcción de esta medida se basa en la idea de que **cuanto mayor sea**

[6] A los efectos de facilitar el análisis, la variable original fue recodificada en tres categorías.

[7] Esta manera de expresar la forma de la relación es posible en este caso, porque la variable categórica es ordinal. Si tuviéramos por ejemplo Nacionalidad, la descripción sería del tipo "*a los de la nacionalidad A les corresponde mayores ingresos que a los de la nacionalidad B, etc.*".

la relación entre ambas variables **más homogéneo será** el comportamiento de **la variable numérica en cada uno de los grupos definidos por la variable cualitativa**. Esto se traduce en que la variable cualitativa define clases de individuos con valores en la variable numérica muy similares entre sí y diferentes a los valores de los individuos de las otras clases.

En otras palabras, si la relación es fuerte, estaremos en condiciones de predecir con bastante certeza el valor que toma la variable numérica conociendo la categoría a la que pertenece el individuo observado; en nuestro ejemplo: si existe una relación fuerte, conociendo el nivel de estudio del padre podríamos predecir, con poco margen de error, el ingreso de la familia.

En consecuencia, en este análisis de la relación no solo debemos centrar nuestra atención en la comparación de medidas de tendencia central, sino que **debemos prestar especial atención a la variabilidad que se observa en cada grupo**.

Para la construcción de la *razón de correlación*, se hace necesario introducir un concepto asociado a la variabilidad que expresa lo siguiente:

Descomposición de la variabilidad total

La variabilidad total de la variable numérica se puede descomponer en la suma de la **variabilidad dentro** de los grupos o clases definidos por la variable categórica, **más** la **variabilidad entre** los distintos grupos (Teorema de Huygens).

Es decir:

Suma de Cuadrados total = Suma de Cuadrados intra-clase + Suma de cuadrados entre-clase

En símbolos: **SCT** = **SCintra** + **SCentre** **(8)**

Donde:

SCT = suma de los cuadrados de los desvíos individuales con respecto a la media general.

SCintra = suma de los cuadrados de los desvíos de cada individuo con respecto a la media del grupo al que pertenece.

SCentre = suma de los cuadrados de los desvíos de las medias de cada grupo con respecto a la media general.

De acuerdo con la lógica planteada para construir la ***razón de correlación***, esperamos que cuanto más fuerte sea la relación entre las variables menor será el *SCintra* y mayor el *SCentre*; o sea, si la relación es perfecta la variabilidad total se debe a la variabilidad *entre* los grupos, en tanto que será igual a cero la variación *dentro* grupos (todos los valores del grupo son iguales entre sí). Podemos expresar la *razón de correlación* (simbolizada con la letra griega "*eta*" al cuadrado: η^2) como:

$$\eta^2 = \frac{SCentre}{SCT} \qquad \text{donde: } 0 \leq \eta^2 \leq 1$$

Si calculamos las sumas de cuadrados correspondientes al ejemplo de los ingresos familiares y el nivel de estudios del padre, tenemos[9]:

[8] Formalmente el teorema se expresa: $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_i - \overline{y}_j)^2 + \sum_{j=1}^{h} n_j.(\overline{y}_j - \overline{y})^2$; donde h es la cantidad de categorías de la variable cualitativa, n_j el número de individuos de cada categoría, $\overline{y}_j$ es la media aritmética de cada una de las subpoblaciones; $\overline{y}$ es la media general de la variable numérica Y.

[9] Los resultados de la suma de cuadrados, así como el valor de η^2, se obtienen fácilmente a través de cualquier programa estadístico. De ahí el énfasis puesto en transmitir la lógica de la construcción y funcionamiento de este índice y no en los cálculos que el mismo demanda.

	Suma de Cuadrados	
Entre grupos	$\sum_{j=1}^{h} n_j.(\bar{y}_j - \bar{y})^2$	3061288
Intra grupos	$\sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_i - \bar{y}_j)^2$	20863881
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	23925169

$$\eta^2 = \frac{\text{SCentre}}{\text{SCT}} = \frac{3061288}{23925169} = 0{,}128$$

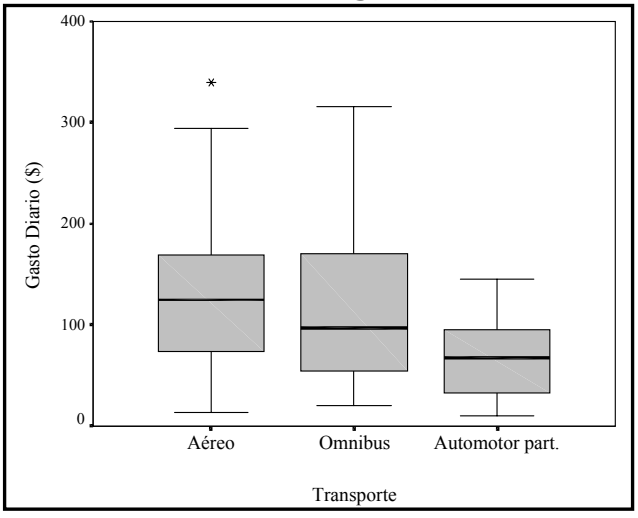
Podemos advertir que si bien, la diferencia entre las medidas de tendencia central eran importantes, la "razón de correlación" está *indicando una relación débil entre las variables.* Esto se debe a que el reagrupamiento generado a partir del nivel de estudio del padre, no produce grupos suficientemente homogéneos dentro de ellos y muy diferentes entre sí. Así, en los diagramas de tallo-hoja construidos inicialmente, se puede ver que -sobre todo en las dos primeras clases- existe un "solapamiento" de los ingresos, producto de la dispersión de esta variable dentro de cada grupo; incluso se puede destacar que el menor ingreso observado de todo el conjunto de datos se da en el nivel intermedio de educación y no en el más bajo. En consecuencia, podemos señalar que el nivel de educación del padre no discrimina bien el ingreso familiar.

3.2. El recurso gráfico

Dado que se trata de la comparación de distribuciones univariadas de una variable numérica, valen para este caso los recursos gráficos que se presentaron oportunamente y, para un análisis completo, es interesante incluir en los gráficos la ubicación de la media y la mediana.

Por ejemplo, construir tantos ***histogramas o polígonos*** como clases o grupos queden determinados por la variable categórica. El ***diagrama de tallo-hoja*** presentado en el ejemplo constituye simultáneamente -como ya hemos dicho- un recurso gráfico y numérico pertinente para este tipo de análisis. Otro recurso muy útil y expresivo para la comparación es el diagrama de Caja (*Box-Plot*), tal como se presenta en el siguiente ejemplo.

Distribución de grupos turísticos según gasto diario y medio de Arribo. Puerto Iguazú, Febrero '94



La comparación de los tres diagramas nos indica que aquellos que viajan por automotor presentan en general gastos de menor nivel y más concentrados (menos dispersos) que los que arribaron a Iguazú en otro medio de transporte. A su vez, entre los que viajan en ómnibus se observa una mayor variabilidad de los gastos (tanto en el 50% central como en el total de datos), con una asimetría hacia la derecha, expresada por una mayor dispersión en la mitad de los que más gastan (tanto la parte superior de la caja como el bigote superior son más extensos que sus correspondientes inferiores). Además, los que viajan en transporte aéreo tienen una mediana de gastos, superior a los otros dos grupos, con una ligera simetría a la izquierda en los valores centrales, una asimetría general a la derecha y un grupo con un gasto atípico.

Actividad Nº 7

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

4. La Relación entre Variables Numéricas

Muchas veces nos encontramos en situación de querer responder preguntas que refieren a la relación de dos variables numéricas. Así por ejemplo, podemos plantearnos preguntas expresadas de la forma...

- ✓ ¿al aumentar el número de años de estudio, aumenta el ingreso?,
- ✓ ¿al aumentar el número de automóviles por habitantes, aumenta el número de accidentes de tránsito?,
- ✓ ¿al disminuir el gasto en publicidad, disminuye la demanda de un producto?,
- ✓ ¿cuánto más tiempo se invierta en el estudio es mayor la calificación?,
- ✓ ¿cuanto mayor es el número de médicos por habitantes en un país, cómo varía la tasa de mortalidad infantil?,
- ✓ ¿al aumentar la antigüedad de un automóvil, aumenta el costo de mantenimiento?,
- ✓ etc.

En todas estas cuestiones el objetivo es indagar si, al cambiar el valor de una de las variables, varía en forma coordinada el valor de la otra variable. En definitiva, nos estamos preguntando por la variación conjunta o ***covariación*** de dos variables numéricas.

Dos variables X e Y (ambas numéricas) están ***correlacionadas***, si al aumentar o disminuir los valores en una de ellas (los de X por ejemplo) se observa una modificación definida (aumento o disminución) en los valores observados en la otra variable (Y).

En esta intención de analizar la correlación, el recurso gráfico aparece como un instrumento inmediato, simple y de fácil interpretación para poner en evidencia la existencia o no de la relación entre las dos variables numéricas.

4.1. El recurso gráfico

Grupos Turísticos según Número de Componentes y Gasto Total de un Día

GRUPO	COMPONENTES	GASTO ($)
1	1	92
2	5	235
3	1	70
4	6	505
5	2	149
6	6	460
7	2	149
8	6	343
9	2	220
10	3	155
11	5	275
12	3	180
13	4	146
14	4	280
15	5	240
16	3	160

Cuando se trata de dos variables que se miden en una escala numérica, es posible utilizar un sistema de coordenadas cartesianas ortogonales para la representación gráfica.

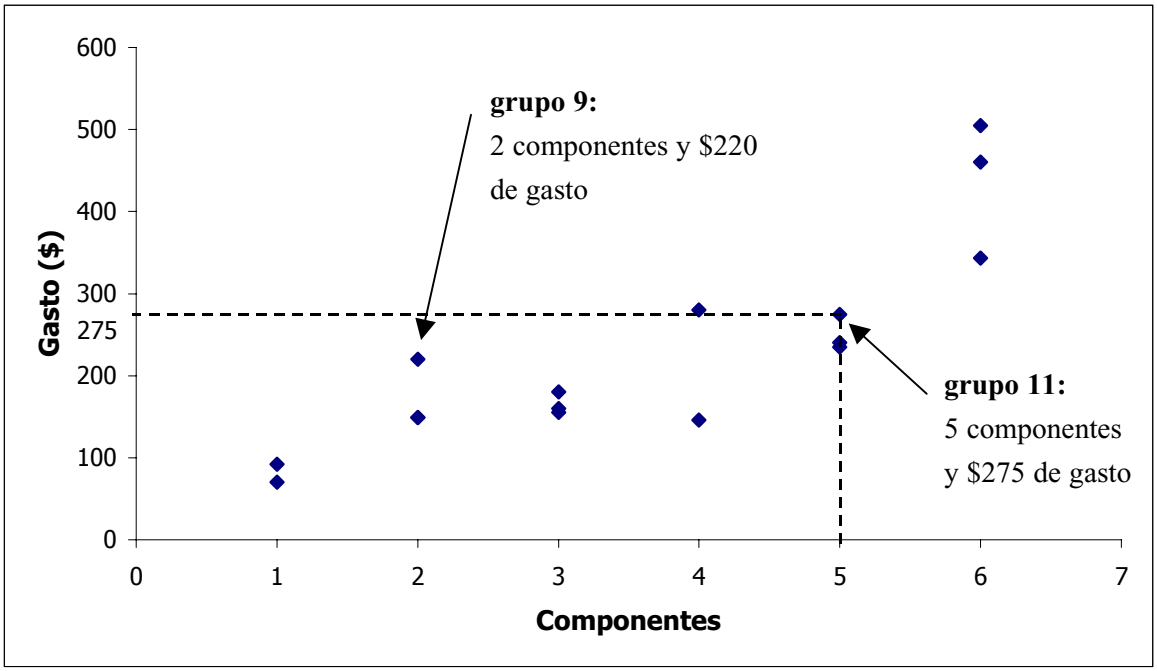
Analicemos a manera de ejemplo, la covariación entre el ***número de componentes*** de los 16 grupos turísticos que visitaron el Parque Nacional Iguazú en febrero de 1994 y el ***gasto diario*** que estos mismos grupos realizaron. Según la definición de correlación, la existencia de una relación entre estas dos variables significaría que al aumentar el número de componentes el gasto diario debería variar de una manera definida.

Observando la matriz de datos, al comparar los valores registrados por los grupos turísticos en ambas variables se puede apreciar -aún con dificultad- que en general **a los más numerosos** les corresponden **mayores niveles de gastos**, lo que nos **permite suponer la existencia de una relación** entre las dos variables. En este caso, además, podemos suponer que la

naturaleza de la relación es "causal", siendo el *número de componentes* la variable "que explica" el *gasto* de los grupos.

Esa comparación de los grupos turísticos (que en este caso son las unidades de análisis) se facilita considerablemente si **se representa gráficamente cada grupo según los valores registrados en ambas variables.**

Distribución de los grupos turísticos según el número de componentes y gasto diario



Así, en este tipo de gráficos se ubica en el eje de las X aquella variable que actúa como "independiente", mientras que, en el eje de las Y, la variable considerada "dependiente"[10]. En el plano de representación aparecerán **tantos puntos como unidades de análisis** o individuos se hayan observado, correspondiéndole como coordenadas a cada uno de ellos los valores registrados en cada variable. A cada punto se lo ubica por un par ordenado (x; y).

Así, en nuestro ejemplo, el grupo identificado con el número 11, aparece ubicado en el plano con una coordenada en el eje X igual a 5 y una coordenada en el eje Y de 275.

El grupo 11 ➔ es el punto con coordenadas (5, 275)

Representados todos los individuos de esta manera, se obtiene lo que se conoce como ***Diagrama de Dispersión***.

Actividad Nº 8

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 8 de la Guía de Actividades correspondiente a esta unidad***.**

En el diagrama de dispersión anterior se aprecia inmediatamente que **los grupos turísticos con un mayor número de componentes presentan -*en términos generales*- un gasto más alto**. Se comprueba –en este caso- un comportamiento sincrónico de las variables donde, al crecer los valores de X, también crecen los valores de Y.

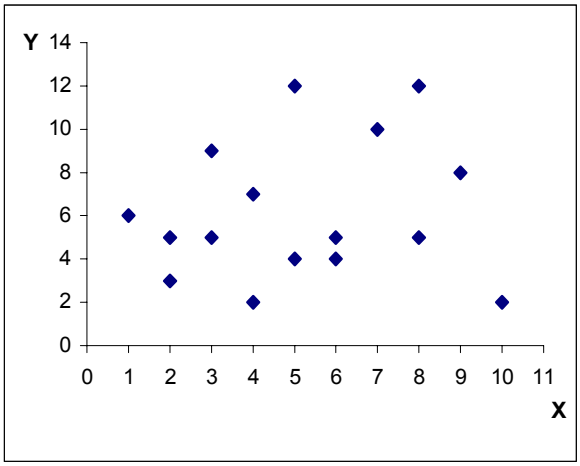
[10] Cuando se trata de una relación causal, la X corresponde a la variable *explicativa*, en tanto que la Y a la variable *explicada*. Además recordemos que la designación de una variable como dependiente o independiente no es una cuestión estadística, sino una decisión que corresponde al conocimiento del investigador sobre el fenómeno que está estudiando.

A través de los **diagramas de dispersión** *podemos estudiar:*

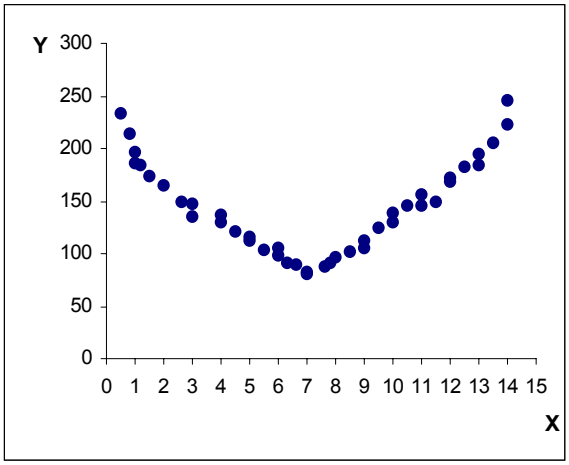
- ✓ ***si existe relación*** *entre las variables,*
- ✓ *caracterizar* ***la forma*** *de la relación, y*
- ✓ *apreciar* ***la intensidad*** *de esa relación.*

¿Cómo se manifestaría gráficamente la relación entre dos variables numéricas?

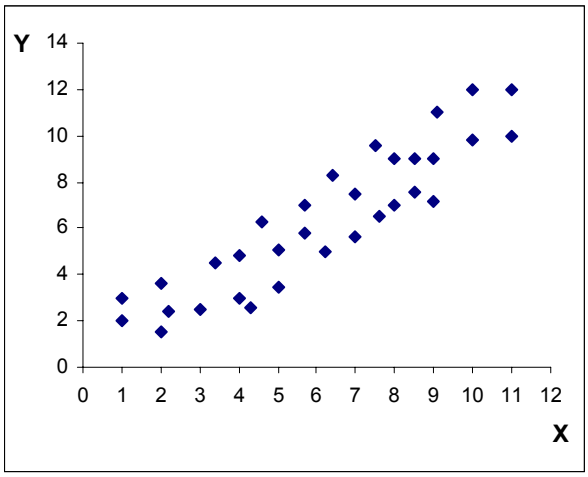
(a) No hay relación



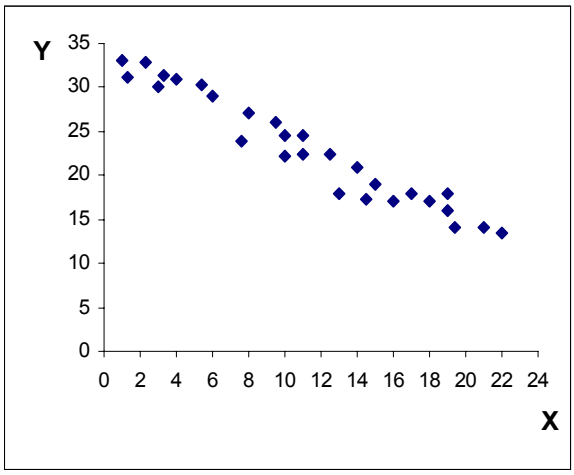
(b) Relación Curvilínea / No lineal



(c) Relación lineal positiva



(d) Relación lineal Negativa



Como hemos dicho, para que exista relación entre las variables, a las variaciones en los valores de una de ellas le corresponderán variaciones definidas en la otra. Este comportamiento no se observa en el gráfico (a), mientras que sí ocurre en los tres restantes.

En el **gráfico (a):**

- ✓ Vemos que a las variaciones en X, le corresponden variaciones "imprevisibles" en Y. A valores crecientes de X, se suceden tanto valores decrecientes como crecientes de Y; no se aprecia una forma definida en el diagrama de dispersión. En consecuencia no hay relación entre ambas variables.

En el **gráfico (b):**

- ✓ Se puede ver que los cambios en X se corresponden con variaciones definidas en Y. En consecuencia, ***existe relación*** entre ambas variables.

- ✓ Esos cambios son tales que, para valores crecientes de X, los valores de Y decrecen hasta un cierto punto para posteriormente comenzar a aumentar, describiendo los puntos una figura que se asemeja a una parábola. Así entonces puede decirse que su ***forma es curvilínea***.
- ✓ Además, dado que los puntos se ajustan casi perfectamente a esa parábola, podemos decir que la ***relación es fuerte*** (para un valor dado de X es posible predecir con bastante precisión el valor esperado de Y).

En el **gráfico (c):**

- ✓ Se puede ver que los cambios en X se corresponden con variaciones definidas en Y. En consecuencia, ***existe relación*** entre ambas variables.
- ✓ Esos cambios son tales que, a valores crecientes de X, le corresponden valores crecientes de Y, describiendo los puntos una figura que se asemeja a una recta. Así entonces puede decirse que su ***forma es lineal y creciente (también llamada lineal positiva).***
- ✓ Respecto a esa recta imaginaria, los puntos presentan un nivel de dispersión tal que nos permite calificar como ***moderada la intensidad*** de esa relación (para un valor de X podemos predecir un valor de Y, pero con cierto margen de error).

En el **gráfico (d):**

- ✓ Se puede ver que los cambios en X se corresponden con variaciones definidas en Y. En consecuencia, ***existe relación*** entre ambas variables.
- ✓ Esos cambios son tales que, a valores crecientes de X, le corresponden valores decrecientes de Y, describiendo los puntos una figura que se asemeja a una recta. Así entonces puede decirse que su ***forma es lineal y decreciente (también llamada lineal negativa).***
- ✓ Respecto a esa recta imaginaria, los puntos presentan un bajo nivel de dispersión, de manera que nos permite calificar como ***fuerte la intensidad*** de esa relación (para un valor de X podemos predecir con poco margen de error el valor correspondiente de Y).

En este curso, nos abocaremos exclusivamente al estudio de las ***relaciones lineales***.

En nuestro ejemplo sobre el estudio de la relación entre el número de componentes de los grupos turísticos y el gasto diario que realizan, observando el diagrama de dispersión podemos concluir que: **existe una relación entre las variables**, que esa relación es **de forma lineal y positiva** (al aumentar el número de componentes se registra un aumento "en promedio" del gasto diario) y que la intensidad se podría calificar provisionalmente como moderada.

Sobre este último aspecto avanzaremos en el apartado siguiente, presentando una forma de cuantificar la fuerza de la relación de dos variables cuantitativas.

IMPORTANTE

Debemos destacar que ***el análisis de la correlación comienza siempre por un estudio del diagrama de dispersión***, a partir del cual evaluamos si tiene sentido o no ***pensar en la existencia de una relación*** entre las variables consideradas y, en el caso que sea ***lineal, pasar a calcular una medida que exprese la intensidad de la relación***.

Actividad Nº 9

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 9 de la Guía de Actividades correspondiente a esta unidad***.**

4.2. El recurso numérico

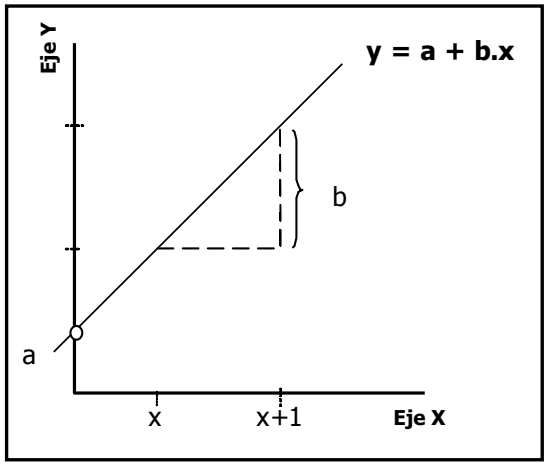
Para el caso de relaciones lineales entre las variables, desarrollaremos:

- **a.** ***el análisis de regresión***: un método que nos permite obtener la "mejor" recta que describe la relación observada, y
- **b.** ***el coeficiente de correlación***: una medida para cuantificar la fuerza de la relación.

4.2.1. El análisis de regresión lineal simple

El objetivo es **describir la relación** observada en el diagrama de dispersión, **con un modelo matemático** (una ecuación) que nos permita predecir los valores de Y correspondientes a valores dados de X. Dado que se trata de una relación lineal, ese modelo matemático a obtener corresponde a la *ecuación de una recta*.

Ecuación de la recta : $y = a + bx$

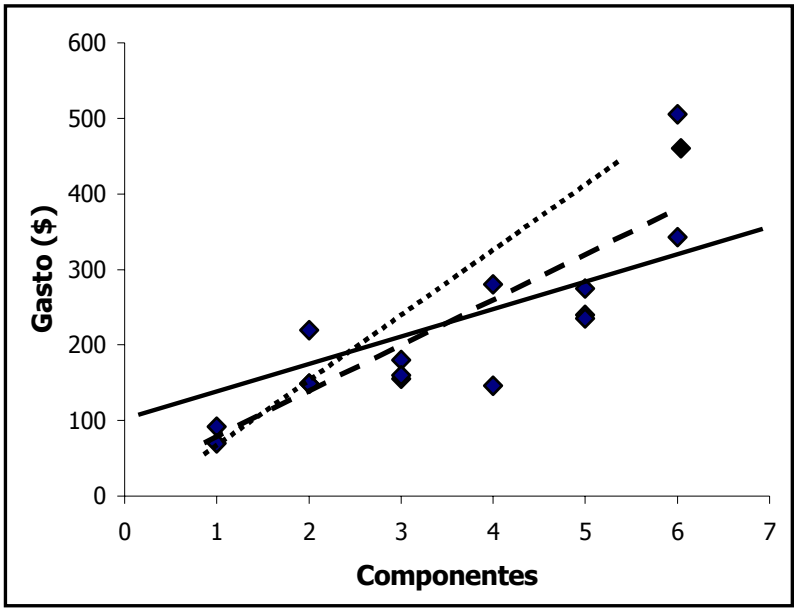


Donde:

a : es la ordenada al origen (valor de *y* cuando *x* = 0; punto en que la recta corta al eje Y).

b : es la pendiente de la recta (es lo que varía y por cada unidad de variación en x)[11]. Tomará valores positivos si al aumentar X aumenta Y (relación lineal positiva), y negativo si al aumentar X disminuye Y (relación lineal negativa).

Debemos buscar una recta que exprese o "ajuste", de la mejor manera posible, los datos observados. Intuitivamente podríamos pensar que será aquella recta que pase "lo más cerca posible" de todos los puntos que representan a los datos.



A mano alzada se pueden trazar varias rectas que "en apariencia" responden a ese propósito general, tal como las que se presentan en el gráfico. Ejemplo: puedo trazar rectas que pasen por pares de puntos que resulten usuales (no atípicos) dentro del conjunto, identificando así tantas rectas como pares de puntos no atípicos se encuentren.

Pero...

¿cuál es la recta que mejor ajusta a la nube de puntos?

Antes de definir un método para encontrar esta recta, es necesario precisar que el modelo matemático encontrado nos permitirá determinar para cada valor x_i de X, un valor estimado $\hat{y}_i$

[11] La pendiente se define como la tangente del ángulo que forma la recta con el semieje positivo de las X.

de Y. Ese par de valores $(x_i;\hat{y}_i)$ define un punto que "cae" sobre la recta. En nuestro ejemplo, utilizando el modelo, tendremos para cada número de componentes la estimación de un gasto diario.

Las diferencias que se registran entre cada valor observado (y_i) y el correspondiente valor estimado por el modelo ($\hat{y}_i$), constituye lo que se define como ***error de estimación***: $e_i = y_i - \hat{y}_i$

Debe destacarse que el modelo va a estimar un valor "promedio" de Y para cada valor de X (observe que, para cada valor de X: tamaño de grupo, pueden existir distintos valores de Y: gasto diario [12]). En consecuencia, la estimación no es exacta en términos de lo que puede efectivamente observarse para cada grupo, de ahí la presencia de los errores individuales.

Encontrar **la recta que mejor ajusta a la nube de puntos significa minimizar estos errores**. A partir de esta condición se define el siguiente **criterio para estimar la recta** que mejor ajusta las observaciones:

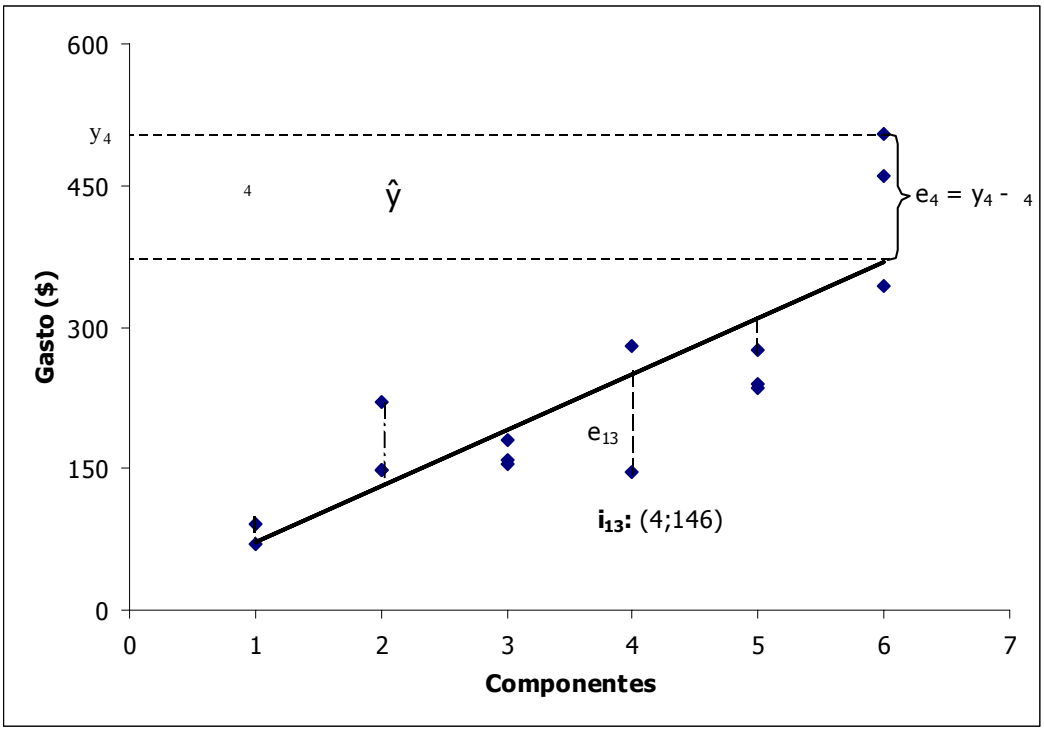
Criterio de *mínimos cuadrados*

Es aquel mediante el cual obtenemos la **recta que hace mínima la suma de los errores al cuadrado.** En símbolos quedaría expresado como:

$$\sum \left(y_i - \hat{y}_i \right)^2 = \sum \left(y_i - a - b.x_i \right)^2 = \text{mínimo}$$

Donde: a y b son las incógnitas a determinar

Determinación de la recta y errores $\hat{y}_4$ $\hat{y}_4$ de estimación en el ajuste de mínimos cuadrados



El criterio de mínimos cuadrados presentado, permitirá estimar los parámetros *a* y *b* del modelo (ecuación de la recta) que mejor ajusta nuestra nube de puntos [13]. Soslayando los procedimientos matemáticos requeridos para su determinación, encontramos que estos parámetros o coeficientes de regresión se pueden calcular mediante las siguientes expresiones.

[12] Es fácil de comprender que -en nuestro ejemplo- grupos de igual número de componentes pueden realizar distintos niveles de gasto diario. Ej: grupos 13 y 14, o los grupos 2, 11 y 15, etc.

[13] Los valores de los coeficientes a y b se obtienen fácilmente a través de cualquier programa estadístico. Nuevamente aquí resulta importante comprender la lógica para determinar la recta que mejor ajusta la nube de puntos y la utilidad de contar con este modelo, más que los cálculos que requieren la determinación de estos coeficientes.

Coeficientes de regresión

Pendiente: $b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$ **Ordenada al origen:** $a = \bar{y} - b\bar{x}$

A modo de ejemplo, estimamos la ecuación de la recta que describe la relación entre *gasto diario* y el *número de componentes* de los grupos turísticos que visitan Puerto Iguazú. Para el cálculo de los coeficientes de regresión *a* y *b* operamos de la siguiente manera.

Cálculos para determinar los valores de *a* y *b*

GRUPO	COMPONENTES	GASTO	x.y	x^2
1	1	92	92	1
2	5	235	1175	25
3	1	70	70	1
4	6	505	3030	36
5	2	149	298	4
6	6	460	2760	36
7	2	149	298	4
8	6	343	2058	36
9	2	220	440	4
10	3	155	465	9
11	5	275	1375	25
12	3	180	540	9
13	4	146	584	16
14	4	280	1120	16
15	5	240	1200	25
16	3	160	480	9
Suma	***58***	***3659***	***15985***	***256***

Cálculo de la Pendiente: $b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$

$$b = \frac{16.15985 - 58.3659}{16.256 - (58)^2} = \frac{255760 - 212222}{4096 - 3364} = 59,5$$

b=59,5

A partir del valor de b podemos concluir que el aumento de un integrante en el grupo turístico incrementará el gasto diario, en promedio, en $59,5.

Cálculo de la Ordenada al origen: $a = \bar{y} - b\bar{x}$

$$\bar{x} = \frac{58}{16} = 3,6 \qquad \bar{y} = \frac{3659}{16} = 228,7$$

Entonces, $a = 228,7 - 59,5.3,6 = 14,5$

a=14,5

Reemplazando estos coeficientes en la ecuación de la recta $y = a + bx$, tenemos:

y=14,5+59,5.x

La ventaja de contar con un modelo matemático que expresa la relación entre estas variables es que ***nos permite hacer pronósticos***. Así, si quisiéramos estimar el gasto diario de un grupo de 8

personas, le damos a x el valor 8 y obtenemos una estimación del gasto promedio para un grupo turístico de 8 integrantes.

$$y = 14,5 + 59,5.8 = 490,5$$

Entonces*, si un grupo turístico tiene 8 componentes esperaríamos que realice un gasto diario de $490,5*.

IMPORTANTE

- ✓ Cuando realizamos un análisis de regresión ***estamos suponiendo que existe una relación causal que va de X a Y*** (X es la variable explicativa e Y la variable explicada). Como consecuencia, antes de realizar este análisis estadístico, será preciso que el investigador decida -basándose en su conocimiento del tema- cuál es el sentido de la causalidad.
- ✓ Cuando el pronóstico se realiza para valores de la variable independiente que están fuera del recorrido observado (en nuestro caso grupos de 7 o más integrantes), se habla de una ***extrapolación***. Cuando el pronóstico se refiere a un valor que está dentro del recorrido observado (1 a 7 integrantes en el ejemplo) hacemos una ***intrapolación*** y en estos casos es cuando podemos calcular el error cometido con nuestra estimación media en relación con el valor de y efectivamente observado (el gasto diario medio de los grupos con ese número de componentes).
- ✓ La ***extrapolación*** -en términos generales- irá perdiendo precisión a medida que nos alejamos del campo de variación observado. Ahora bien, *¿cuál es el límite para hacer una extrapolación*? Esto ***dependerá del fenómeno en estudio*** y, en consecuencia, solo puede ser respondido a partir del conocimiento sobre el tema.
- ✓ La ***intrapolación*** será tanto ***más eficiente cuanto menor sea la dispersión*** de los puntos en torno a la recta [14].
- ✓ En términos generales, la predicción será tanto más eficiente cuanto mayor sea la fuerza de la correlación entre las variables.

Actividad Nº 10

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 10 de la Guía de Actividades correspondiente a esta unidad***.**

4.2.2. El coeficiente de correlación lineal de Pearson (*r*)

Este coeficiente que se propone como medida de la fuerza y sentido de la relación entre dos variables numéricas, cuantifica la dispersión de las observaciones (puntos del diagrama) en torno a la recta de regresión estimada. Por esta razón a este coeficiente se lo denomina también ***Coeficiente de correlación lineal***.

Así, si tenemos dos variables X e Y con medias $\bar{x}$ e $\bar{y}$; y desviación estándar σ_x y σ_y, el coeficiente de correlación se define como [15]: $r = \frac{1}{n} \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{\sigma_x . \sigma_y}$

[14] Sobre este aspecto del análisis de regresión y particularmente el uso del modelo de regresión lineal para efectuar predicciones, ver Bibliografía propuesta para esta unidad.

[15] En algunos textos en el coeficiente r se utiliza (n-1) en lugar de n. Esta distinción, que será tratada en la Estadística Inferencial, se justifica en aquellos casos en los que se trabaja con una muestra y no con la población total.

Esta expresión, que tiene en su numerador la variación conjunta o covariancia de X e Y, y en el denominador los desvíos estándar de cada una de las variables, rara vez es utilizada en la práctica. Esto es así en primer lugar porque los paquetes de análisis estadísticos (incluido Excel) lo calculan a partir de la matriz de datos original, y en el caso de tener que obtenerlo manualmente es más operativo recurrir a **la fórmula de trabajo** que se presenta a continuación:

$$r = \frac{n\sum x.y - \sum x.\sum y}{\sqrt{\left[n.\sum x^2 - \left(\sum x\right)^2\right].\left[n.\sum y^2 - \left(\sum y\right)^2\right]}}$$

Valores posibles de *r*

El coeficiente r puede tomar todos los valores comprendidos entre –1 y 1.

$$-1 \leq r \leq 1$$

Un valor de *r* positivo indica una relación lineal directa o positiva, mientras que si *r* es negativo la correlación entre las variables es indirecta o negativa.

A su vez, los valores de r **"cercanos" a 1 o –1** están ***señalando un correlación fuerte*** entre las variables, mientras que los ***"cercanos" a 0 indican una relación débil o inexistente***.

- **r = 0** No existe relación lineal entre *x* e *y,* pero puede existir una relación no lineal [16].
- **r = 1** Relación lineal ***perfecta positiva*** (directa)
- **r = -1** Relación lineal ***perfecta negativa*** (inversa)

IMPORTANTE

- ✓ El análisis de la correlación se debe ***iniciar con un estudio del diagrama de dispersión***, a partir del cual decidiremos si es pertinente pensar en la existencia de una relación lineal.
- ✓ En el análisis de correlación, ***no se supone una relación de causalidad entre X e Y*** (a diferencia de la regresión); en consecuencia es indistinta la designación de qué variable funciona como X y cuál como Y.
- ✓ Cuando es posible suponer una relación causal entre las variables es informativo calcular **el coeficiente de determinación** ($\mathbf{R^2}$) que se obtiene elevando el coeficiente de correlación (r) al cuadrado. Así $\mathbf{R^2 = r^2}$.
- ✓ El ***coeficiente de determinación*** se interpreta como: ***la proporción de la variabilidad de Y que está explicada por la variabilidad de X***. Es usual expresar este coeficiente en porcentaje.

En el ejemplo de la relación entre número de componentes de los grupos turísticos y gastos diarios que estos realizan, pudimos observar en el diagrama de dispersión que existía una relación lineal positiva, y además de la observación del gráfico dedujimos una relación de intensidad moderada. Estamos ahora en condiciones de poder cuantificar la fuerza de la relación. Así, realizados los cálculos con la fórmula de trabajo y utilizando los datos de la matriz presentada en páginas anteriores, surge que el coeficiente de correlación es [17]:

$$\mathbf{r} = 0{,}85$$

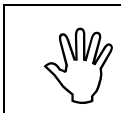
El valor de ***r*** obtenido corrige nuestra impresión visual indicando que "*la **relación** entre las variables es **fuerte y positiva** (o directa)*". Como podemos suponer una relación causal entre X e Y, tiene sentido en este caso obtener el coeficiente de determinación $\mathbf{R^2}$.

[16] Si existe otro tipo de relación, se manifestará en el diagrama de dispersión.

[17] Invitamos al lector a que controle el cálculo realizado.

$R^2 = 72{,}3\%$

Lo que indica que "*un 72% de la variación en los gastos diarios está explicada por las variaciones en el número de componentes del grupo"*.



Actividad Nº 11

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 11 de la Guía de Actividades correspondiente a esta unidad***.**

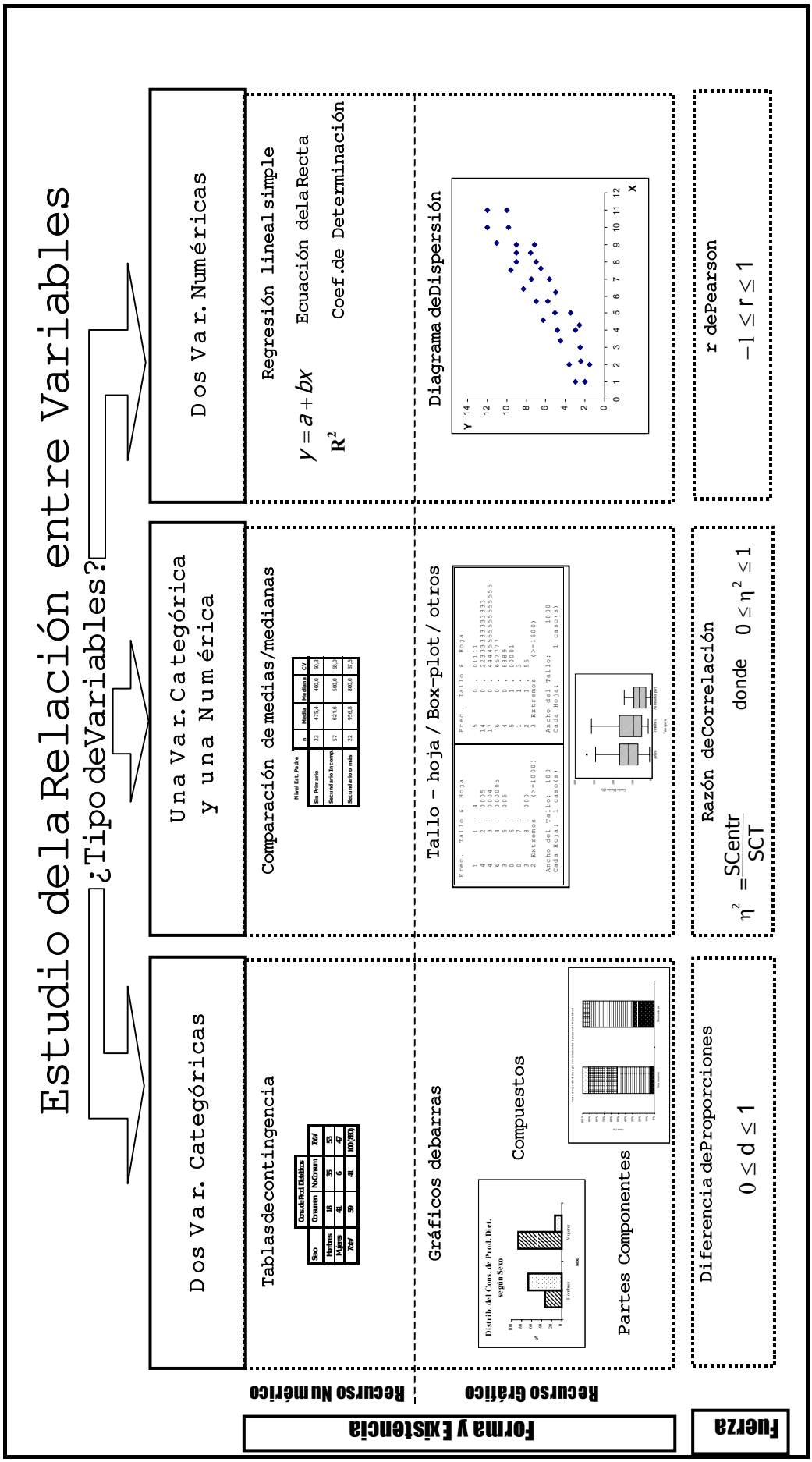
5. ¿Qué Hemos Visto?

En esta presentación, una vez precisado el ***tipo de cuestiones*** que estamos tratando de responder con el ***análisis bivariado*** de los datos, comenzamos por señalar la necesidad de preguntarnos sobre el ***tipo de variables*** que están involucradas en el estudio, como así también por la ***naturaleza de la relación*** que se puede establecer entre ellas, dado que estos dos aspectos condicionan tanto las posibilidades de análisis (las herramientas a las que podemos recurrir) como el alcance de los resultados de nuestro estudio (la posibilidad de hacer pronósticos, explicar o simplemente describir la relación).

Para el análisis de las relaciones, distinguimos estrategias diferentes según el tipo de variable: 1) ***Análisis de Tablas de Contingencia***, para dos variables cualitativas, 2) la ***comparación de medias***, en el caso de una variable cualitativa y una cuantitativa, y 3) el ***análisis de regresión y correlación lineal*** cuando se trata de dos variables cuantitativas.

Hemos destacado, además, que en este tipo de análisis existen ***tres aspectos*** que deben ser considerados cualquiera sea el tipo de variables: a) la determinación de la ***existencia*** de la relación entre las variables, b) la ***forma*** en que se da esa relación, y c) la ***fuerza*** de esa relación.

En todos los casos hemos presentado *herramientas* que nos permitían establecer la ***existencia o no de la relación***, ***describir la forma*** en que se producía esta relación, como así también una medida (**diferencia de proporciones, razón de correlación y *coeficiente de correlación***) para valorar la intensidad de la relación entre esas variables. Cuando se trata del análisis de dos ***variables numéricas,*** presentamos además la determinación de un ***modelo matemático*** **que permite hacer predicciones** cuando la relación existente es lineal y de naturaleza causal (***análisis de regresión lineal***).


Estudio de la Relación entre Variables
¿Tipo de Variables?
Dos Var. Categóricas
Una Var. Categórica y una Numérica
Dos Var. Numéricas
Tablas de contingencia
Gráficos de barras
Compuestos
Partes Componentes
Comparación de medias/medianas
Tallo – hoja / Box-plot / otros
Regresión lineal simple
Ecuación de la Recta
Coef. de Determinación
Diagrama de Dispersión
Razón de Correlación
donde
r de Pearson
Diferencia de Proporciones
Forma y Existencia
Fuerza
Recurso Numérico
Recurso Gráfico

Bibliografía

BARBANCHO, Alfonso: *Estadística elemental moderna*. Ed. Ariel Barcelona, España, 1978, pág. 211 a 221 y 237 a 245.

COLL, Sebastián; GUIJARRO, Marta: *Estadística aplicada a la historia y a las Ciencias Sociales*. Edic. Pirámide, Madrid, 1998, pág. 235 a 241 y 259 a 263.

DANIEL, Wayne: *Estadística con aplicación a las ciencias sociales y a la educación.* McGraw-Hill, México, 1985, pág. 315- 331.

MOORE, David: *Estadística aplicada básica*, Antonio Bosch ed., Barcelona, 1998 (1ra. Ed. 1995). Pág. 90 a 157.

Conceptos Centrales de esta Unidad

- Distribuciones bivariadas.
- Relación entre variables.
- Naturaleza de la relación entre las variables.
- Los tres aspectos del estudio de relación entre variables: existencia, forma y fuerza.
- Tablas de contingencia y estudio de relación entre variables cualitativas.
- Estudio de la relación entre una variable cualitativa y cuantitativa.
- Relación entre variables cuantitativas: Diagrama de dispersión.
- Análisis de regresión: modelo matemático y predicción.
- Análisis de correlación: coeficiente de Pearson.

Habilidades

- *Identificar* las herramientas numéricas y gráficas apropiadas para el estudio de la relación entre dos variables, cualquiera sea su tipo.
- *Construir* el resumen gráfico o numérico apropiado para analizar la relación entre las variables en estudio.
- *Interpretar* esos resúmenes gráficos o numéricos.
- *Evaluar* la existencia, la forma y la fuerza de la relación entre variables, cualquiera sea su tipo.
- *Realizar pronósticos* basándose en modelos de regresión lineal simple.
- *Comunicar* los resultados del análisis.

UNIDAD 6: LOS NÚMEROS ÍNDICES

1. ¿Qué son y cuál es su utilidad?

Cuando analizamos las condiciones socioeconómicas de una región, de una provincia, de un país, reiteradamente nos encontramos ante la situación de tener que valorar la evolución en el tiempo o en el espacio de variables numéricas, referidas a aspectos diversos de la realidad. Es habitual que debamos encontrar respuestas a preguntas del tipo:

- ✓ ¿en cuánto se incrementó el costo de vida durante el último año?
- ✓ ¿cuál fue el aumento del precio de la harina en el último mes?
- ✓ ¿es mayor o menor la producción de té en Misiones en relación con la de Corrientes?
- ✓ ¿ascendió el número de visitantes al Parque Nacional Iguazú respecto al año anterior?
- ✓ ¿crecieron las ventas de la empresa durante el último trimestre?
- ✓ etc.

Así, las variaciones de los precios de diversos artículos, del costo de una canasta de bienes, de la cantidad de visitantes a un centro turístico, del volumen producido mensualmente por una fábrica, etc., pueden ser datos estratégicos a la hora de planificar una actividad o tomar decisiones.

La **comparación relativa** de los cambios de los valores de una variable, ya sea a través del tiempo o del espacio, generalmente brinda al analista una idea más precisa de la magnitud de tales cambios que la simple comparación en términos absolutos. En efecto, la comprensión del cambio experimentado es más clara si la explicamos diciendo que "*la superficie cultivada con yerba mate aumentó un 9,4% entre 1991 y 1998*", que si señaláramos "*la superficie cultivada creció en 15 mil ha*" en ese período de tiempo.

En otros problemas es necesario **cuantificar mediante un único valor** la magnitud de los cambios relativos de un conjunto de variables heterogéneas, como, por ejemplo, las variaciones conjuntas de los precios de venta de distintos artículos, de la cantidad consumida de diferentes productos, etc.

Los **números índices** son las técnicas estadísticas que nos permitirán resolver este tipo de problemas.

Los números índices

Son **medidas estadísticas** que sirven para **comparar magnitudes** de una o más variables en un período (o lugar) dado, con la magnitud de esa misma o mismas variables en otro período (o lugar) de referencia llamado *base*.

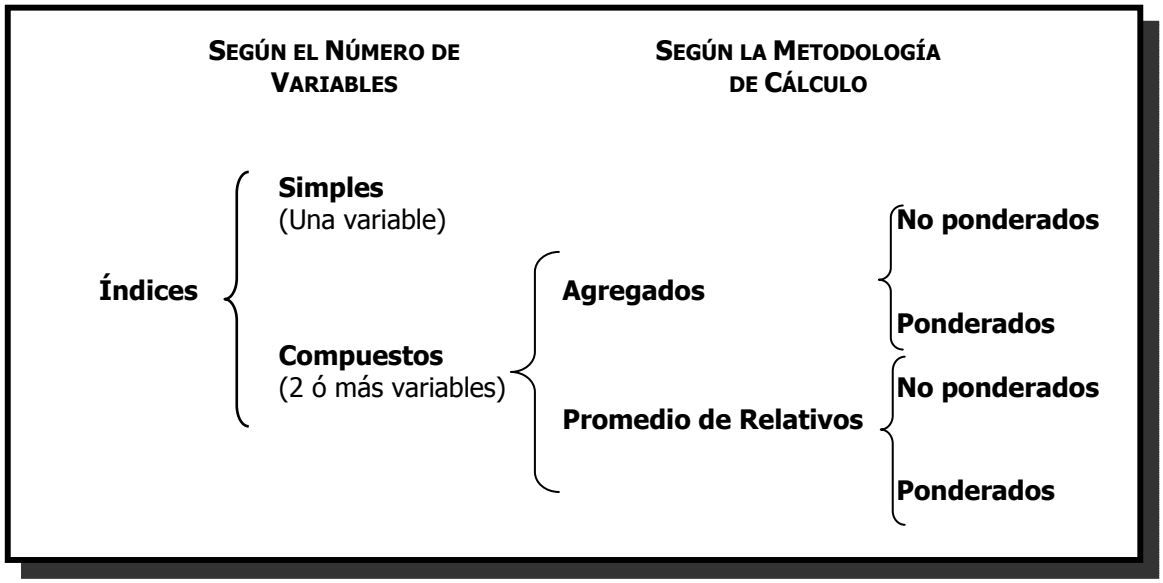
Según el número de variables con las que se trabaja en la construcción de un número índice, se los puede agrupar en dos grandes capítulos:

- **Números Índices Simples**: se construyen para medir los cambios o variaciones (a través del tiempo o del espacio) de una sola variable.
- **Números Índices Compuestos**: miden los cambios conjuntos de dos o más variables.

Tomando en cuenta la metodología utilizada para su construcción y cálculo, los índices compuestos se diferencian en **índices de agregados** y del **promedio de relativos**, pudiendo a su vez clasificarse cada uno de ellos en **no ponderados** y **ponderados**.

El esquema siguiente resume la clasificación de los números índices según sea el número de variables que intervienen en su construcción y el método de cálculo específico de cada uno de ellos. En este curso presentaremos solo las fórmulas de los índices cuyo uso es más generalizado en la práctica: índices relativos simples, índice compuesto de agregado no ponderado, índice del promedio de relativos no ponderado, índices de Laspeyres e índices de Paasche. En la bibliografía recomendada para este Capítulo el lector podrá ampliar estos conocimientos básicos con otros métodos.

Diferentes Tipos de Números Índices



2. Los Números Índices Simples

Como ya señaláramos, estos índices tienen la finalidad de medir los cambios o variaciones de los datos x1, x2, x3, ... xi,... xt de una única variable X. Estos valores pueden resultar de observaciones realizadas a una única unidad de análisis a través de diferentes momentos de tiempo (datos longitudinales), como por ejemplo son *los precios mensuales de la yerba mate durante los últimos doce meses en la ciudad de Posadas*; u observaciones realizadas transversalmente como por ejemplo *los precios de la yerba mate en el último mes en las capitales provinciales de la Argentina*.

Considerando que el tratamiento metodológico es similar para una u otra situación, los ejemplos que presentaremos a lo largo de la unidad harán únicamente referencia a datos recogidos en forma cronológica (series de tiempo). Por lo tanto los valores de una variable genérica X, observados en "t" períodos consecutivos de tiempo (quinquenios, años, meses, semanas, días, etc.), se simbolizarán del siguiente modo:

"i-ésimo" período

Períodos	1	2	3	. . . i . . .	t
Valores de X	x_1	x_2	x_3	. . . x_i . . .	x_t

Valor de X correspondiente al i-ésimo período

Si la variable en estudio fuera el precio de un producto o servicio registrado en diferentes períodos, el símbolo genérico a utilizar será "**pi**" (en lugar de x_i) que denota: *el precio del artículo en cuestión, registrado en el i-ésimo período de la serie*.

Según el tipo de interrogante que nos planteemos sobre el comportamiento de la variable que estamos analizando, se pueden realizar diversas operaciones que dan lugar a diferentes números índices.

2.1. Índice Relativo Simple de Base Fija (Rs)

Este índice se construye para mostrar las variaciones relativas (porcentuales) **en los valores de una sola variable, referidos** todos estos cambios a **un único valor** de la serie llamado **valor del período base**.

El índice relativo simple de base fija mide la variación de la variable en estudio entre un período "i" dado de la serie y otro período fijo llamado "*base*" (al que simbolizamos con "**o**"). Se lo obtiene haciendo:

$$R_S{}^{i/o} = \frac{x_i}{x_o} \bullet 100$$

Donde:

x_i : es el valor de X en el período i de interés (o "*período dado*").

x_o: es el valor de X en el período elegido como base.

A manera de ejemplo, consideremos la serie de precios de la *yerba mate canchada* durante el período comprendido entre los años 1992 y 2000. En este caso deseamos medir la variación relativa de los precios de cada período de la serie, con respecto al valor del año 1992 (año base elegido arbitrariamente en este ejemplo) [1].

Precios Corrientes de la Yerba Mate Canchada y Variaciones de los Precios
Período: 1992-2000.

Años	Precios ($/Kg)	IPRs (1992=100)
1992	**0,67**	**100,0**
1993	0,65	97,0
1994	0,66	98,5
1995	0,67	100,0
1996	0,54	80,6
1997	0,43	64,2
1998	0,38	56,7
1999	0,35	52,2
2000	0,34	50,7

Año base →

$R_s\ 97/92 = \frac{0,43}{0,67} \bullet 100 = 64,2\%$

El precio decreció un 49,3% (100-50,7)

Fuente: Dir. de Economía Agraria. Min. de Asuntos Agrarios. Posadas, Misiones. 2002.

El índice relativo simple de base fija del año '93 con base en el año '92, resulta de:

$$R_s\ {}^{93}\!/_{92} = \frac{0,65}{0,67} \bullet 100 = 97,0\%$$

Es decir que en 1993 el precio de la yerba mate canchada **decreció un 3%** (100-97) **con respecto al valor registrado en el año base**.

Analizando los índices relativos simples para la serie completa, notamos que el precio de la yerba mate canchada muestra un comportamiento decreciente a lo largo del período considerado ya que, a partir de 1995, año en el que se produce una ligera recuperación y alcanza un precio igual al registrado en el año base, decrece sostenidamente hasta alcanzar el menor valor en el año 2000, en el cual registra una caída del 49,3% con relación al precio de 1992.

2.2. El Relativo Simple de Eslabón (Re)

Este índice mide los cambios relativos de una sola variable entre dos **períodos sucesivos** (años, meses, semanas, días, etc.) de una misma serie. Es decir, permite expresar en porcentajes la variación ocurrida en los datos entre un período "i" cualquiera y el período inmediato anterior ("i-1"). Cuando nos informan que *"según los datos que difundió ayer el INDEC, el valor de la canasta básica para una familia tipo subió en setiembre un 2,05%..."* (Clarín del martes 8/10/02), la operación realizada para obtener esta información es un índice de estas características.

[1] A los fines de este índice **c**ualquier período de la serie puede ser adoptado como "base". En cada problema particular de trabajo el investigador deberá decidir el período base más conveniente, según las recomendaciones que se explican más adelante.

El índice relativo simple de eslabón mide las **variaciones relativas** de una variable en estudio **entre períodos consecutivos**, por lo que se conocen también como relativos simples con base móvil. Se lo obtiene haciendo:

$$R_e i/(i-1) = \frac{x_i}{x_{i-1}} \bullet 100$$

Donde:

x_i : es el valor de la variable en un período cualquiera de la serie.

x_{i-1}: es el valor correspondiente al período anterior.

Consideremos nuevamente el ejemplo anterior de la serie de precios corrientes de la *yerba mate canchada*, pero en este caso queremos conocer la evolución de los precios entre cada período (año en nuestro caso) y el inmediato anterior.

El índice de precios relativo simple en eslabón del año '95 (con respecto al año 1994) se obtiene de:

$$R_s\, 95/94 = \frac{0,67}{0,66} \bullet 100 = 101,5\%$$

Es decir que el precio de la yerba mate canchada del año 1995 **aumentó el 1,5% con respecto al precio anterior**. La evolución de los índices en eslabón para la serie completa se presenta en la tabla siguiente [2]:

Precios Corrientes de la Yerba Mate Canchada y Variaciones Anuales. Período: 1992-2000.

Años	Precios ($/Kg)	Re (%)
1992	0,67	-
1993	0,65	97,0
1994	0,66	101,5
1995	0,67	101,5
1996	0,54	80,6
1997	0,43	79,6
1998	0,38	88.4
1999	0,35	92,1
2000	0,34	97,1

$$R_e\, 97/96 = \frac{0,43}{0,54} \bullet 100 = 79,6\%$$

Fuente: Dir. de Economía Agraria. Min. de Asuntos Agrarios. Pdas., Mnes. 2002.

Con excepción de los años 1994 y 1995 en los que el índice registra una ligera recuperación del 1,5% con respecto al año anterior, a lo largo del período analizado los precios corrientes de la yerba mate canchada muestran un comportamiento progresivamente decreciente, ya que los valores disminuyen sostenidamente de un año a otro desde 1995 en adelante, observando la mayor caída en 1997 con un descenso del 20,4% en relación con el precio de 1996.

2.3. El Relativo Simple en Cadena (Rc)

Es frecuente que a partir de los índices en eslabón se necesite obtener los cambios relativos de una variable con referencia a un único período base. En este caso precisamos determinar, por ejemplo, **cuánto se incrementó el costo de la canasta básica de una familia tipo a lo largo del año**, conociendo los aumentos –proporcionados por el INDEC– que se produjeron mensualmente. En este tipo de situaciones recurrimos a los índices relativos en cadena.

[2] Se debe tener en cuenta que el 100% para cada valor de la serie corresponde al período inmediato anterior.

Los relativos simples en cadena se obtienen como el producto del relativo en eslabón correspondiente al período en estudio ("**i**") por los sucesivos relativos en eslabón entre ese período y la base, sin incluir al de esta. Es decir:

$$R_c\, i/o = R_e\, i/(i-1) \bullet R_e\, (i-1)/(i-2) \bullet \ldots \bullet R_e\, 1/o \bullet 100$$

Por lo tanto si, conociendo los índices relativos en eslabón, quisiéramos saber cuál fue la variación que registraron los precios corrientes de la yerba mate canchada del año 2000, con referencia al período base 1995 [3], la operación que debemos realizar es:

$$R_c\, {}^{2000}\!/_{1995} = 0,971 \bullet 0,921 \bullet 0,884 \bullet 0,796 \bullet 0,806 \bullet 100 = 50,7\%$$

Actividad Nº 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. Los Números Índices Compuestos

Se construyen para mostrar los **cambios colectivos** de un conjunto de variables (ya no más de una sola variable), las que generalmente se refieren a conceptos económicos tales como precios, cantidades (producidas, vendidas, compradas, etc.) o valores [4] de grupos de artículos que interesan por alguna razón especial. Así, por ejemplo, recurrimos a estos índices cuando estamos interesados en conocer:

- ✓ la evolución de los precios de los cultivos agrícolas de Misiones,
- ✓ el aumento en el volumen de las exportaciones de cereales de la Argentina durante cierto período,
- ✓ cuánto aumentó la canasta de productos alimenticios durante el último mes,
- ✓ etc.

Es decir que con los índices compuestos estaremos interesados en medir las fluctuaciones relativas conjuntas de "**n**" variables distintas, para cada una de las cuales se registran datos a lo largo de "**t**" períodos de tiempo (años, meses, semanas, días, etc). Así, las magnitudes correspondientes a las "**n**" variables en los "**t**" períodos, se simbolizan como:

Tiempo / Variable	1	2	...	J	...	t
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1t}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2t}
...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{it}
...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nt}

Dato de la "i-ésima" variable (X_i) registrado en el "j-ésimo" período de la serie

Siendo:

x_{21}: valor observado (dato) de la segunda variable (X_2) registrado en el primer período de la serie.
x_{i2}: dato de la "i-ésima" variable (X_i) registrado en el segundo período de la serie.
x_{nj}: dato de la "n-ésima" variable (X_n) registrado en el "j-ésimo" período de la serie.
x_{nt}: dato de la "n-ésima" variable (X_n) registrado en el "t-ésimo" (último) período de la serie.

[3] Contando con los precios corrientes, es evidente que resulta más sencillo obtener la misma información calculando un índice relativo simple.

[4] El "valor" (v) de un artículo se define como el producto del precio por la cantidad; es decir: $\mathbf{v = p_{ij} \cdot q_{ij}}$.

Si las variables en análisis fueran los precios de n artículos diferentes, el símbolo genérico que se adopta (en lugar de x_{ij}) es "$\mathbf{P_{ij}}$" que denota: *precio del i-ésimo artículo considerado, registrado en el j-ésimo período de la serie*.

Como ya fuera señalado, de acuerdo con la forma de obtener este tipo de índices se pueden distinguir los **índices de agregados** de los **índices promedios de relativos**, los que a su vez pueden ser "**no ponderados**" o "**ponderados**".

3.1. Índice de Agregado no Ponderado

Con este índice se miden las variaciones producidas en magnitudes que surgen de agregar cantidades simples (ej.: precios de los cereales, cantidades exportadas de productos agrícolas, etc.).

Al índice de agregado no ponderado se lo define como la suma de las magnitudes de todas las variables consideradas, para un mismo período dado "j" de la serie; dividida por la suma de todas las magnitudes correspondientes a esas mismas variables en el período elegido como base. El valor del índice expresado en porcentaje se obtiene haciendo:

$$I_{j/o} = \frac{\sum_{i=1}^{n} x_{ij}}{\sum_{i=1}^{n} x_{io}} \cdot 100$$

Donde:
x_{ij}: es la magnitud correspondiente a la "i-ésima" variable/artículo en el período j.
x_{io}: es la magnitud de esa misma variable/artículo en el período base.

Si las variables en estudio fueran los precios de una canasta de n artículos diferentes, el índice de agregado no ponderado (para cierto período "j" con base en otro período "o" de la misma serie) resultará:

$$IP_{j/o} = \frac{\sum_{i=1}^{n} p_{ij}}{\sum_{i=1}^{n} p_{io}} \cdot 100$$

En el cálculo de este índice se **considera una unidad de cada bien**, y **expresa el precio total** (de ventas, compras, etc.) **de los "n" artículos en cada período**, **como un porcentaje del precio de esos mismos artículos en el período base**.

Consideremos la serie de precios anuales de la *yerba mate canchada* y del *brote de té verde*, registrados en el período comprendido entre los años 1992 y 2000 (Note que analizamos n=2 artículos-variables diferentes, cada una de ellas observada a lo largo de 9 años -períodos- consecutivos). Ahora el problema es medir la evolución conjunta de los precios de ambos productos, tomando como base los precios observados en el año 1992.

Precios Corrientes de la Yerba Mate Canchada y el Brote de Té.
Variaciones de los Precios. Período: 1992-2000.

Años	Yerba Mate ($/Kg)	Té ($/kg)	$\sum_{i=1}^{2} p_{ij}$	IP (1992=100)
1992	0,67	0,060	**0,730**	100,0
1993	0,65	0,058	0,708	97,0
1994	0,66	0,070	0,730	100,0
1995	0,67	0,057	0,727	99.6
1996	0,54	0,055	0,595	81,5
1997	0,43	0,055	0,485	66.4
1998	0,38	0,075	0.455	62,3
1999	0,35	0,050	0.400	54,8
2000	0,34	0,050	0.390	53,4

Es el resultado de sumar el precio de 1 kg de yerba y 1 kg. de té en 1992

$$IP_{97/92} = \frac{0,485}{0,730} \cdot 100 = 66,4\%$$

Fuente: Dir. de Economía Agraria. Min. de Asuntos Agrarios. Posadas, Misiones. 2002.

Así entonces, el índice de precios de agregado no ponderado para el año '93, tomando como período de comparación al año '92, resulta:

$$IP\,{}^{93}\!/_{92} = \frac{0,65 + 0,058}{0,67 + 0,060} \cdot 100 = \frac{0,708}{0,730} \cdot 100 = 97,0\%$$

En consecuencia, los precios de la yerba mate canchada y del brote de té verde en 1993 **decrecieron, en conjunto, un 3%** (100-97) **con relación a los precios que registraron ambos productos en el año base 1992.**

Por lo tanto, a lo largo del período analizado los precios de estos cultivos muestran, en conjunto, un comportamiento en general decreciente con respecto a los precios de 1992. Solamente en el año 1994 los precios logran una ligera recuperación alcanzando el mismo nivel del año base y luego decrecen sostenidamente hasta alcanzar su menor valor en el año 2000, cuando el índice mide una caída del 46,6% respecto de los precios de 1992.

IMPORTANTE

- Al ser "no ponderado", este índice **le asigna igual importancia al cambio absoluto de cada variable**. Así, aquellas variables con magnitudes altas impactarán más en el resultado final del índice.
- En el caso de los precios, **la unidad de medida de cada artículo introduce una ponderación no deseada**. Es de esperar que artículos fraccionados en unidades mayores tengan precios relativamente mayores.
- **No se puede calcular el agregado simple de** cantidad cuando las variables que intervienen en su construcción **están expresadas en unidades diferentes**.

3.2. Índice de Promedio de Relativos no Ponderado

Como su nombre lo indica, consiste en promediar magnitudes relativas referidas a las variaciones individuales de series de precios, cantidades o valores.

Se lo define como el promedio no ponderado de los relativos simples (cada uno de ellos calculado para un mismo período "j" dado y un mismo período base predeterminado), para las "n" variables consideradas en el análisis. El valor del índice es expresado en porcentaje y se lo obtiene haciendo:

$$I_{j/o} = \frac{\sum_{i=1}^{n} \frac{x_{ij}}{x_{io}}}{n} \cdot 100$$

Donde:
x_{ij}: es la magnitud correspondiente al "i-ésimo" artículo en el período j.
x_{io}: es la magnitud correspondiente al "i-ésimo" artículo en el período base.

Para calcular el índice del promedio de relativos se deben realizar los siguientes pasos:

a. obtener las variaciones relativas (relativos simples) de cada variable para el mismo período "j" y con la misma base,

b. obtener la suma de los relativos para el período "j" considerado,

c. dividir la suma obtenida por el número total "n" de variables incluidas en el índice.

Si se tratara de un índice de precios, se lo obtiene mediante la siguiente expresión:

$$IP_{j/o} = \frac{\sum_{i=1}^{n} \frac{p_{ij}}{p_{io}}}{n} \cdot 100$$

Consideremos nuevamente la serie de precios de la yerba mate canchada y el brote de té para el período 1992 - 2000.

Precios Corrientes y Variaciones de los Precios de la Yerba Mate Canchada y Brote de Té. Período: 1992-2000.

Años	Yerba Mate		Té		$\sum_{i=1}^{2} \frac{p_{ij}}{p_{io}} \cdot 100$	IP (1992=100)
	($/Kg)	Rs (´92=100)	($/kg)	Rs (´92=100)		
1992	0,67	100,0	0,060	100,0	200,0	100,0
1993	0,65	97,0	0,058	96,7	193,7	96,9
1994	0,66	98,5	0,070	116,7	215,2	107,6
1995	0,67	100,0	0,057	95,0	195,0	97,5
1996	0,54	80,6	0,055	91,7	172,3	86,2
1997	0,43	64,2	0,055	91,7	155,9	78,0
1998	0,38	56,7	0,075	125,0	181,7	90,9
1999	0,35	52,2	0,050	83,3	135,5	67,8
2000	0,34	50,7	0,050	83,3	134,0	67,0

Fuente: Dirección de Economía Agraria. Ministerio de Asuntos Agrarios. Posadas, Misiones. 2002.

El índice de precios del promedio de relativos no ponderado del año '95, tomando como referencia el año '92, resulta de:

$$IP^{95}/_{92} = \frac{\frac{0,67}{0,67} + \frac{0,057}{0,069}}{2} \cdot 100 = \frac{195,0}{2} = 97,5\%$$

En 1995 el precio de la yerba mate canchada y el brote de té **decrecieron -en promedio- un 2,5%** (100-97,5) **con relación a los precios registrados en el año base.**

Nuevamente, notamos que este índice también nos muestra la persistente caída de los precios de los dos artículos en conjunto ya que, considerados aisladamente, el comportamiento de los precios del té (relativos simples) muestra variaciones muy diferentes a las de la yerba mate (relativos simples). En conjunto, los precios de ambos cultivos son, año a año, inferiores a los de 1992. La excepción es el año 1994 en el cual los precios, en promedio, superan a los de la base en un 7,6%. Los menores precios de la serie analizada se registran en el año 2000 para el cual el índice muestra una caída conjunta de ambos productos del orden del 33,0% con respecto a 1992.

IMPORTANTE

- Por ser los relativos **números abstractos, desprovistos de toda unidad** de medida, **este índice supera las principales limitaciones** asignadas al índice de agregados no ponderados.
- El que se utilice para su cómputo **un promedio aritmético simple, puede ser metodológicamente inapropiado** en el caso de magnitudes relativas.
- Este índice **le asigna igual representatividad en el promedio a cada relativo**; esto hace que variaciones absolutas pequeñas pero relativamente grandes impacten más en el valor final del índice que variaciones grandes en términos absolutos pero pequeñas en términos relativos.

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.

En general, los índices compuestos presentados hasta aquí adolecen del mismo defecto: **la falta de ponderación** de las variables que lo constituyen. Estos índices compuestos serán más eficientes en la medida en que cada una de las variables esté convenientemente "ponderada" por un factor que exprese su importancia relativa en el conjunto.

Se puede apreciar que los números índices se aplican principalmente a dos tipos de variables económicas: **precios** y **cantidades**. Tratándose de precios, las ponderaciones utilizadas más frecuentemente son las respectivas cantidades (de venta, de compra, de producción, etc) y, si se trata de cantidades, lo usual es ponderar por los precios respectivos.

3.3. Los Índices de Agregados Ponderados

Al construir un índice de precios (o de cantidades) podemos tomar la decisión de ponderar por las cantidades (o precios) del año base, del año que se está analizando o por un valor que promedia ambas magnitudes. Según sea la ponderación que adjudiquemos a cada variable al construir el índice, vamos a estar en presencia de un tipo particular de índice de agregados ponderados.

3.3.1. El Índice de Laspeyres

Para la construcción de este índice se utilizan como ponderaciones magnitudes (cantidades o precios) del año base. Si se trata de un índice de precios (IP^L), este se obtendrá haciendo:

$$IP^L\ j/o = \frac{\sum_{i=1}^{n} p_{ij} \cdot q_{io}}{\sum_{i=1}^{n} p_{io} \cdot q_{io}} \cdot 100$$

Donde:

p_{ij}: es el precio correspondiente al "i-ésimo" artículo en el período j.

p_{io}: es el precio correspondiente al "i-ésimo" artículo en el período base.

q_{io}: es la cantidad correspondiente al "i-ésimo" artículo en el período base.

La aplicación de la fórmula de Laspeyres para un período "j" dado (tomando como base otro período **"o"** predeterminado), supone realizar los siguientes pasos:

a. multiplicar el precio de cada artículo en el período "j" dado por la cantidad de ese mismo artículo registrada en el período base,

b. realizar la suma de los productos así calculados, a través de los n artículos que intervienen en el índice,

c. multiplicar el precio de cada artículo en el año base por la correspondiente cantidad en el mismo período base y sumar estos productos a lo largo de todos los artículos,

d. dividir la suma realizada en b por la suma realizada en c y, luego, al resultado multiplicar por cien.

Es obvio que para el cálculo de este índice se necesita más información (datos) que para el cálculo de los índices no ponderados que hemos visto. En efecto, el índice de precios de Laspeyres requiere de datos de cantidades (compradas, vendidas, producidas, etc.) de cada uno de los artículos que lo integran para, al menos, el período seleccionado como base.

Vamos a analizar la evolución de los precios de la yerba mate canchada y el brote de la hoja verde de té mediante un índice de Laspeyres.

Producción, Precios Corrientes y Variaciones de los Precios de la Yerba Mate Canchada y Brote de Té. Período: 1992-2000.

Años	Yerba Mate		Té		$\sum_{i=1}^{2} p_{ij}q_{io}$	IP^L ('92 = 100)
	Producción (kg.)	$/kg.	Producción (kg.)	$/kg		
1992	198.000.000	0,67	191.800.000	0,060	144.168.000	100,0
1993	230.000.000	0,65	226.300.000	0,058	139.824.400	97,0
1994	280.000.000	0,66	209.954.000	0,070	144.106.000	100,0
1995	270.000.000	0,67	211.000.000	0,057	143.592.600	99,6
1996	270.000.000	0,54	203.400.000	0,055	117.469.000	81,5
1997	280.000.000	0,43	220.000.000	0,055	95.689.000	66,4
1998	245.000.000	0,38	265.000.000	0,075	89.625.000	62,2
1999	231.000.000	0,35	266.300.000	0,050	78.890.000	54,7
2000	280.000.000	0,34	228.000.000	0,050	76.910.000	53,3

Valor de la producción del año ´92 a los precios corrientes de cada año

Fuente: Dirección de Economía Agraria. Ministerio de Asuntos Agrarios. Posadas, Misiones. 2002.

El índice de precios del año '96 tomando como referencia el año '92, se obtiene haciendo:

$$IP^{L}{}_{96/92} = \frac{0,54 \cdot 198.000.000 + 0,055 \cdot 191.800.000}{0,67 \cdot 198.000.000 + 0,060 \cdot 191.800.000} \cdot 100 = 81,5\%$$

En 1996 el precio de la yerba mate canchada y el brote de té **decrecieron -en promedio- el 18,5% con relación a los precios registrados en el año base.**

El precio de estos cultivos expone –en promedio– un comportamiento decreciente. A partir de 1994 los valores decrecen sostenidamente a lo largo del período analizado registrando el menor valor de la serie en el año 2000, en el que se produce una caída conjunta del 46,7% respecto a los precios de 1992.

El **índice de cantidad de Laspeyres** (IQ^L) es la contrapartida del índice de precios análogo, donde **las ponderaciones a ser usadas serán los precios del año base**. Así, el mismo se obtiene mediante la aplicación de la siguiente fórmula:

$$IQ^{L}{}_{j/o} = \frac{\sum_{i=1}^{n} q_{ij} \cdot p_{io}}{\sum_{i=1}^{n} q_{io} \cdot p_{io}} \cdot 100$$

Este índice de cantidades agregadas ponderadas responde a la siguiente pregunta:

- ✓ ¿cuánto se gastará (o recibirá) en el período dado con relación al período base si compramos (o vendemos), **a los precios del año base**, cantidades variables de los mismos artículos?

El índice de Laspeyres de cantidad para el año '96 tomando como referencia el año '92, se obtiene haciendo:

$$IQ^{L}{}_{96/92} = \frac{270.000.000 \cdot 0,67 + 203.400.000 \cdot 0,060}{198.000.000 \cdot 0,67 + 191.800.000 \cdot 0,060} \cdot 100 = 133,9\%$$

En 1996 la producción de yerba mate canchada y brote de té crecieron -en promedio- el 33,9% con relación a la producción obtenida en 1992.

3.3.2. El Índice de Paasche

Es un índice en el cual se utilizan como ponderaciones, magnitudes (cantidades o precios) del año en estudio. Si se trata de un índice de precios (IP^P), este se obtendrá de la siguiente manera:

$$IP^P\ j/o = \frac{\sum_{i=1}^{n} p_{ij} \cdot q_{ij}}{\sum_{i=1}^{n} p_{io} \cdot q_{ij}} \cdot 100$$

Donde:

p_{ij}: es el precio correspondiente al "i-ésimo" artículo en el período j (período en estudio).

p_{io}: es el precio correspondiente al "i-ésimo" artículo en el período base.

q_{ij}: es la cantidad correspondiente al "i-ésimo" artículo en el período dado o en estudio.

El valor de este índice debe interpretarse como: **"las cantidades producidas en el período en estudio tienen un % más (o menos) de valor de lo que esa misma lista hubiera tenido en el año base"**.

Si se tratara de un **índice de precios al consumidor**, estaríamos comparando el costo efectivo en el período dado con el costo teórico en el año base, para mantener el estándar de vida del período dado.

El índice de Paasche de los precios corrientes de la yerba mate canchada y el brote de té del año '96, tomando como referencia el año '92, se obtiene haciendo en este caso:

$$IP^{P\ 96/92} = \frac{0{,}54 \cdot 270.000.000 + 0{,}055 \cdot 203.400.000}{0{,}67 \cdot 270.000.000 + 0{,}060 \cdot 203.400.000} \cdot 100 = 81{,}3\%$$

En 1996, el precio de la yerba mate canchada y el brote de té decrecieron -en promedio- el 18,7% con relación a los precios que obtuvieron en 1992.

El **índice de cantidad de Paasche** (IQ^P) es la contrapartida del índice de precios, donde **las ponderaciones serán los precios del período dado**. Así, este índice se obtiene mediante la aplicación de la siguiente fórmula:

$$IQ^P\ j/o = \frac{\sum_{i=1}^{n} q_{ij} \cdot p_{ij}}{\sum_{i=1}^{n} q_{io} \cdot p_{ij}} \cdot 100$$

El índice de cantidad de Paasche responde a la siguiente pregunta:

- ✓ ¿cuánto se gastará (o recibirá) en el período dado con relación al período base si compramos (o vendemos), **a los precios del año en estudio**, cantidades variables de los mismos artículos?

El índice de Paasche de cantidad para el año '96 tomando como referencia el año '92, se obtiene haciendo:

$$IQ^{P\ 96/92} = \frac{270.000.000 \cdot 0{,}54 + 203.400.000 \cdot 0{,}055}{198.000.000 \cdot 0{,}54 + 191.800.000 \cdot 0{,}055} \cdot 100 = 133{,}6\%$$

En 1996 la producción de yerba mate canchada y brote de té crecieron -en promedio- el 33,6% con relación a la producción registrada en el período base.

IMPORTANTE

En general -y aún cuando miden lo mismo- los índices de Laspeyres y Paasche darán resultados diferentes por utilizar diferentes ponderaciones, lo que no indica que uno sea mejor que el otro.

- El índice de **Laspeyres tiene a favor la sencillez de su cálculo** pues requiere de menos información que el de Paasche. Conocidos los precios y cantidades del período base, solo requiere actualizar los precios o cantidades del período en cuestión.
- La fórmula de **Laspeyres,** al utilizar como ponderación los precios o cantidades del período base, **es rígida** y no permite eliminar aquellos artículos del conjunto que en el transcurso del tiempo han ido perdiendo importancia en relación con los restantes, ya sea porque han dejado de producirse, adquirirse o venderse, o porque otros bienes sustitutos los han desplazado. Por ello, cada determinado número de años exige una actualización de las ponderaciones.
- La fórmula de **Paasche es más flexible**, pues al utilizar ponderaciones móviles permite la eliminación, incorporación o sustitución de artículos sin afectar al índice y sin necesidad de modificar la base.

Actividad Nº 3

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad.

3.4. Índice de Promedio Ponderado de Relativos

En estos índices, las **ponderaciones utilizadas son los valores de los ítems utilizados en la construcción del índice**, donde –como se reseñara oportunamente– el valor del artículo se define como el producto del precio por la cantidad ($v_{it} = p_{it} \cdot q_{it}$). Al igual que todos los promedios ponderados vistos hasta aquí, estos promedios ponderados de relativos se calculan multiplicando cada relativo por su ponderación y dividiendo la suma de los productos por la suma de las ponderaciones.

Para el cálculo de estos índices se pueden utilizar los **valores del año base** ($p_{io} \cdot q_{io}$), **del año dado** ($p_{ij} \cdot q_{ij}$) o **ponderaciones teóricas** ($p_{ij} \cdot q_{io}$ ó $p_{io} \cdot q_{ij}$). Según sea la ponderación adoptada, se obtendrán índices equivalentes a los de Laspeyres y Paasche presentados anteriormente.

3.4.1. El Índice Promedio Ponderado de Relativos de Laspeyres

En este índice se utilizan como ponderaciones los **valores correspondientes al año base**. Si se trata de un índice de precios, este se obtendrá de la siguiente manera:

$$IP^{L}\ j/o = \frac{\sum_{i=1}^{n} \frac{p_{ij}}{p_{io}} p_{io} \cdot q_{io}}{\sum_{i=1}^{n} p_{io} \cdot q_{io}} \cdot 100$$

Donde:

p_{ij}: es el precio correspondiente al "i-ésimo" artículo en el período j.

p_{io}: es el precio correspondiente al "i-ésimo" artículo en el período base.

q_{io}: es la cantidad correspondiente al "i-ésimo" artículo en el período base.

El **índice de cantidad** se va a obtener utilizando las mismas ponderaciones, pero en este caso considerando como **variables los relativos de cantidad** de cada uno de los n artículos contemplados. Así, este índice se obtiene mediante la aplicación de la siguiente expresión:

$$IQ^{L}\ j/o = \frac{\sum_{i=1}^{n} \frac{q_{ij}}{q_{io}} p_{io} \cdot q_{io}}{\sum_{i=1}^{n} p_{io} \cdot q_{io}} \cdot 100$$

3.4.2. Índice Promedio Ponderado de Relativos de Paasche

En este índice se utilizan **valores teóricos como ponderaciones**. Si se trata de un índice de precios, se obtendrá de la siguiente manera:

$$IP^{P}\ j/o = \frac{\sum_{i=1}^{n} \frac{p_{ij}}{p_{io}} p_{io} \cdot q_{ij}}{\sum_{i=1}^{n} p_{io} \cdot q_{ij}} \cdot 100$$

Donde:
p_{ij}: es el precio correspondiente al "i-ésimo" artículo en el período j.
p_{io}: es el precio correspondiente al "i-ésimo" artículo en el período base.
q_{it}: es la cantidad correspondiente al "i-ésimo" artículo en el período en estudio.

Mientras en el índice de precios se utilizan como ponderaciones los valores de la producción en el período en estudio a los precios del año base, en el **índice de cantidad** se van a utilizar los valores de la producción del año base a los precios del período en estudio. Así, este índice se obtiene utilizando la siguiente fórmula:

$$IQ^{P}\ j/o = \frac{\sum_{i=1}^{n} \frac{q_{ij}}{q_{io}} p_{ij} \cdot q_{io}}{\sum_{i=1}^{n} p_{ij} \cdot q_{io}} \cdot 100$$

IMPORTANTE

Algunas de las ventajas que presentan los índices promedios de relativos son:

- Los precios o las cantidades relativas para **cada ítem en los agregados constituyen un índice simple**, que a menudo da información valiosa para el análisis.
- Cuando se introduce un nuevo bien para reemplazar a otro usado anteriormente, **los relativos para un nuevo ítem pueden empalmarse a los relativos para el antiguo, utilizando las ponderaciones de valores anteriores**.
- Cuando un índice se calcula seleccionando un ítem de cada uno de los numerosos grupos de artículos, se pueden utilizar los valores de cada grupo como ponderaciones.
- Cuando se construyen **diferentes índices promedios de relativos**, todos ellos de la misma base, **se pueden combinar para formar un nuevo índice**.

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.

4. Algunas Consideraciones Especiales

4.1. El Índice de Valor

Como ya fuera señalado, el valor de un bien se define como el producto de su precio por su cantidad ($v = p \cdot q$). A su vez, el valor de un agregado de bienes es la suma de los valores individuales de los bienes que integran ese agregado ($\sum v_{ij} = \sum p_{ij} \cdot q_{ij}$).

El cambio en el valor de un agregado de valores se mide mediante un **índice de valor (IV)**, que se define como:

$$IV\ j/o = \frac{\sum_{i=1}^{n} p_{ij} \cdot q_{ij}}{\sum_{i=1}^{n} p_{io} \cdot q_{io}} \cdot 100$$

Donde:

p_{ij}: es el precio correspondiente al "i-ésimo" artículo en el período j (período dado o en estudio).

p_{io}: es el precio correspondiente al "i-ésimo" artículo en el período base.

q_{ij}: es la cantidad correspondiente al "i-ésimo" artículo en el período j.

q_{io}: es la cantidad correspondiente al "i-ésimo" artículo en el período base.

En este caso **no es necesario introducir ponderación alguna**, porque esta es inherente a los valores mismos.

Se puede apreciar que **los precios y cantidades del numerador del índice de valor son variables respecto al denominador** y -en consecuencia- su resultado no puede responder a las preguntas que responden los índices de precio y cantidad. Tenemos entonces que, cuando –con el paso del tiempo– los precios crecen (ej.: un período inflacionario), resulta difícil poder apreciar si las modificaciones que se produjeron en el índice **se deben a variaciones en las cantidades, a variaciones en los precios o variaciones que se produjeron en ambas variables al mismo tiempo [5]**.

Vamos a presentar la evolución del valor de la *yerba mate canchada y el brote de la hoja verde de té.*

Producción, Precios Corrientes y Evolución del Valor de la Producción de la Yerba Mate Canchada y Brote de Té. Misiones, 1992-2000.

Años	Yerba Mate		Té		$\sum_{i=1}^{2} p_{ij}q_{ij}$ ← Valor de la producción	IV ('92 = 100)
	kg.	$/kg.	kg.	$/kg		
1992	198.000.000	0,67	191.800.000	0,060	144.168.000	100,0
1993	230.000.000	0,65	226.300.000	0,058	162.625.400	112,8
1994	280.000.000	0,66	209.954.000	0,070	199.496.780	138,4
1995	270.000.000	0,67	211.000.000	0,057	192.927.000	133,8
1996	270.000.000	0,54	203.400.000	0,055	156.987.000	108,9
1997	280.000.000	0,43	220.000.000	0,055	132.500.000	91,9
1998	245.000.000	0,38	265.000.000	0,075	112.975.000	78,4
1999	231.000.000	0,35	266.300.000	0,050	94.165.000	65,3
2000	280.000.000	0,34	228.000.000	0,050	106.600.000	73,9

Fuente: Dirección de Economía Agraria. Ministerio de Asuntos Agrarios. Posadas, Misiones. 2002.

[5] Se debe tener en cuenta que, en el caso de que ambas variables experimenten cambios, estos se pueden producir en forma tal que: los **precios y cantidades crecen o decrecen simultáneamente** (provocando un cambio conjunto en el mismo sentido), **o una de estas variables crece mientras la otra decrece**, dependiendo la variación del índice de valor, de cómo se compensan las magnitudes de variación producida en los precios y las cantidades.

El valor de la yerba mate canchada y el brote de té creció hasta 1994 (con un 38,4% presenta el mayor incremento de la serie), para luego comenzar a disminuir sostenidamente hasta 1999, en el que se registra una caída en el valor de estos productos que lo ubican un 34,7% por debajo del que se registrara en 1992. En el año 2000 se observa una ligera recuperación respecto al valor que se registrara en el año anterior.

El análisis de esta serie **no nos permite discriminar cuánto de la variación observada se debe a modificaciones en los precios y cuánto a cambios en las cantidades producidas**, a menos que elaboremos los índices de precios y cantidades correspondientes [6].

El índice de valor puede considerarse como el producto del índice de precios por el índice de cantidad, pero esta división del agregado de valores en sus factores de precio y cantidad se cumple siempre que el índice utilizado para el cómputo de los dos factores sea **consistente**. Es decir, **un número índice es consistente si el producto del índice de precios por el índice de cantidad coincide con el índice de valor**.

$$IV = IP \,.\, IQ$$

Se puede comprobar que ni el índice de Laspeyres ni el de Paasche cumplen con esta propiedad, pero el producto de un índice de precios de Laspeyres por uno de cantidad de Paasche (y viceversa) dan el índice de valor [7], lo que permite recomendar –obviando otras consideraciones que se deben tener en cuenta al construir un índice– que "*si al construir un índice de precios ponderamos por las cantidades del año base, al elaborar el correspondiente índice de cantidades resulta conveniente ponderar por los precios del año dado (y viceversa), para que los niveles de precio y cantidad sean consistentes*".

4.2. El Cambio de Base de un Número Índice

Si se desea cambiar la base de un índice para hacerla más reciente o para comparar dos índices con bases diferentes, el procedimiento es muy sencillo: se debe **dividir cada número índice de la serie por el valor del índice correspondiente al período que se quiere adoptar como base**.

Consideremos el siguiente ejemplo, en el que tenemos al Índice Mayorista Nivel Industrial y deseamos transformarlo para cambiar la base al año 1992.

Evolución del Índice Mayorista Nivel Industrial. 1992 - 2000

Años	Índice Mayorista Nivel Industrial
1992	96,0
1993	97,5
1994	98,2
1995	105,6
1996	109,6
1997	109,7
1998	106,2
1999	108,3 (*)
2000	111,4 (*)

(*) Valores estimados

Fuente: Misiones, Instituto Provincial de Estadística y Censos (IPEC)

Para transformar esta serie de Índices Mayoristas en nueva serie con base en el año 1992, debemos dividir todos los valores de la serie por el valor del índice correspondiente a ese año (96,0%). Así, al año 1999 le va a corresponder el valor que se obtiene al hacer:

[6] Por los cálculos realizados anteriormente, podemos saber que, por ejemplo, el crecimiento del 8,9% que tuvo el valor de estos productos en 1996 se debió al efecto conjunto de una caída de los precios del 17,5% (índice de precios de Laspeyres), y un aumento de la producción del 33,6% (índice de cantidad de Paasche).

[7] Por ejemplo: $IP^L \cdot IQ^P = \dfrac{\sum p_{ij} q_{io}}{\sum p_{io} q_{io}} \bullet \dfrac{\sum p_{ij} q_{ij}}{\sum p_{ij} q_{io}} = \dfrac{\sum p_{ij} q_{ij}}{\sum p_{io} q_{io}} = IV$

$$IMA_{'92=100}(1999) = \frac{108,3}{96,0} \cdot 100 = 112,8\%$$

La serie reconvertida con este criterio resultaría en:

Evolución del Índice Mayorista Nivel Industrial. 1992 - 2000

Años	Índice Mayorista Nivel Industrial (1992=100)
1992	100,0
1993	101,6
1994	102,3
1995	110,0
1996	114,2
1997	114,3
1998	110,6
1999	112,8 (*)
2000	116,0 (*)

$\frac{96,0}{96,0} \cdot 100$

$\frac{98,2}{96,0} \cdot 100$

(*) Valores estimados

Fuente: Elab. propia basándose en datos del IPEC.

4.3. El Empalme de Dos Números Índices Solapados

Las ponderaciones de un índice pueden estar desactualizadas (algo muy común cuando utilizamos un índice de Laspeyres) y debemos entonces construir un nuevo índice, renovando los factores de ponderación. Así, tendremos una nueva serie que deberá dar continuidad histórica a la serie anterior y consecuentemente exige lo que se conoce como "empalmar ambas series".

En el ejemplo siguiente tenemos dos series que fueron empalmadas en 1996:

Año	1er Índice (1)	2do Índice (1)
1993	100,0	87,0
1994	95,0	82,6
1995	101,0	87,8
1996	115,0	100,0
1997	126,5	110,0

$\frac{95,0}{115,0} \cdot 100$

(1) *Los valores grisados se obtuvieron mediante los dos diferentes métodos de empalme que se pueden utilizar.*

El empalme de las series se puede realizar de dos maneras:

- **Haciéndolos continuos con el índice antiguo**

En este caso se empalma en el período que es base del nuevo índice; la relación del antiguo al nuevo índice que se produce en este período prevalece para los períodos que siguen. Así, en el ejemplo, para todo período posterior, por regla de tres simple se establece que:

$115,0 / 100,0 = x / 110,0 \Rightarrow x = (115,0 / 100,0) \cdot 110,0 = 126,5$

Es decir, **para cambiar la base del nuevo índice con el antiguo**, **se deben multiplicar los valores del nuevo índice por un factor constante** equivalente a la razón entre el nuevo y el viejo índice en el período de empalme (en el ejemplo este valor es 1,15).

- **Haciéndolos continuos con el nuevo índice**

Para hacer continuo el antiguo índice con el nuevo, hay que realizar un cambio de base dividiendo –tal como fuera desarrollado precedentemente– todos los valores anteriores a la nueva base por el valor correspondiente a este período.

4.4. Procedimiento de Números Índices en Cadena

Nuevos artículos son introducidos casi continuamente a los mercados, lo que obliga a revisar periódicamente la lista de artículos y los factores de ponderación correspondientes.

Con este fin se utiliza el procedimiento de eslabones, en el cual se construye un índice tomando como base el período inmediato anterior; estos índices –como hemos visto– pueden ser encadenados de nuevo a un período base común por un proceso de multiplicación.

IMPORTANTE

El procedimiento de números índices en cadena **es útil porque permite efectuar cambios en la composición del índice de un período a otro**, pero se debe tener en cuenta que **la comparabilidad estricta se reduce a los números índices que siguen inmediatamente a la base fijada**.

Cuando los artículos son continuamente sustituidos por nuevos, el significado del índice de encadenamiento se vuelve cada vez más dudoso en el tiempo y, tal vez, pasado cierto tiempo no se pueda describir qué mide el índice.

4.5. La Deflación de una Serie

Las series de datos sobre el valor de alguna magnitud económica (consumo, producción, ventas, inventario, etc.) habitualmente se expresan valuadas según los **precios corrientes** (el precio efectivo) de cada período. En otras palabras, en los períodos en que las variaciones de precio son importantes los cambios en el valor de los bienes no son indicativos de cambios de cantidad, a menos que podamos eliminar de la serie el efecto de las variaciones en los precios. Al procedimiento de **quitar en las series el efecto de los aumentos de precios,** se lo denomina **"*deflactar*"** la serie o **expresarla a precios constantes de un período base**. El índice de precios utilizado en esta función recibe el nombre de "***deflactor***" o "***deflactador***" de la serie.

Para deflactar una serie de valores expresados a precios corrientes, se debe dividir a cada uno de ellos por un índice de precios adecuado correspondiente al mismo período considerado y luego multiplicar el resultado por cien[8]**.** Debe observarse que ambas series (de valor y de precios) tengan la misma base. La nueva serie de valor que así resulta ("*deflactada*" o "*a precios constantes*") refleja las variaciones debidas, únicamente, a la fluctuación de las cantidades (volumen de ventas, de producción, etc.), quedando anulado el efecto de los precios en los cambios del valor.

Evolución del Valor de la Yerba Mate Canchada y el Brote de Té a Precios Corriente y Precios Constantes. Misiones, 1992-2000.

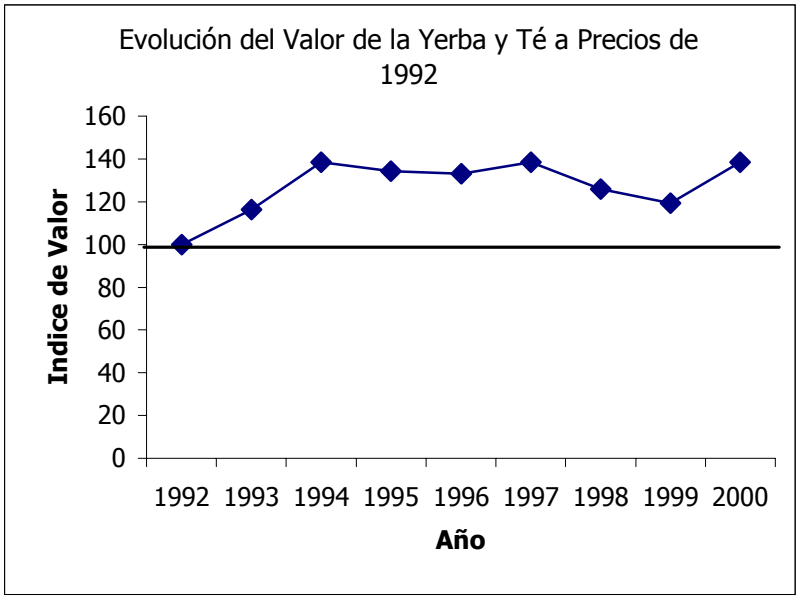
Años	IV ('92 = 100)	IP^L ('92 = 100)	IV (precios de 1992)
1992	100,0	100,0	100,0
1993	112,8	97,0	116,3
1994	138,4	100,0	138,4
1995	133,8	99,6	134,3
1996	108,9	81,5	133,0
1997	91,9	66,4	138,4
1998	78,4	62,2	126,0
1999	65,3	54,7	119,4
2000	73,9	53,3	138,6

Coincide con un Índice de Cantidad

$$\frac{91,9}{66,4} \cdot 100$$

Fuente: Dir. de Economía Agraria. Min. de Asuntos Agrarios. Posadas, Misiones. 2002.

[8] La función de deflactar es una de las aplicaciones más frecuentes de los índices de precios.



En el Cuadro y Gráfico anterior se presenta la evolución de la serie deflactada a los precios de 1992 del valor conjunto de la yerba mate canchada y del brote de té. **En este caso, se utilizó como deflactador el Índice de Precios de Laspeyres** correspondiente a estos productos; **es de esperar entonces que la serie a precios constantes así obtenida coincida con el Índice de Cantidades de Paasche**[9], atendiendo lo que se planteara precedentemente al tratar la propiedad de consistencia de los índices.

Actividad Nº 5

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

5. Problemas en la Construcción de los Números Índices

5.1. La Selección de la Muestra

Lo más importante que se debe señalar en este aspecto es que **el muestreo aleatorio es raramente utilizado** en la construcción de números índices. Los índices se construyen a partir de muestras seleccionadas deliberadamente, dependiendo la representatividad del índice del hecho de que todos o la mayoría de los precios de los bienes que se juzgan importantes en la población sean incluidos en su construcción. Hacemos referencia tanto a los bienes que serán incluidos en la construcción del índice como a las unidades de observación en la que se van a observar precios y cantidades.

Es evidente que el juicio de quien construye el índice y el conocimiento de los datos que se investigan tiene importancia primordial. En el caso de un índice de precios, **el que construye el índice** es quien **debe decidir cuáles son los bienes a ser incluidos**, **cómo se deben definir los precios**, **dónde y cuándo se deben reunir las cotizaciones** de los precios, etc.

Cuando pretendemos observar la evolución de los precios en la economía del país, la decisión sobre los productos que se van consideran para su construcción, y los referentes para obtener los precios (y cantidades) se toma en función del objetivo planteado para el índice. Por ejemplo, si el propósito es describir el comportamiento de las actividades económicas **en general**, se buscará construir un índice ampliamente representativo, tanto en el tipo de productos que se incluyan como en las entidades que realizan transacciones con esos productos; indudablemente se trata en este caso de un índice que se modificará lentamente porque refleja la evolución media de una gran variedad de productos (ej.: el índice de precios mayoristas). En cambio, si el propósito es resaltar algunos

[9] Si $IP^L \cdot IQ^P = IV \Rightarrow IQ^P = IV/IP^L$.

aspectos sintomáticos de la economía, se seleccionan algunas series que reflejan el comportamiento de sectores particulares (ej.: el índice de producción industrial). Este tipo de índices, al promediar productos más homogéneos en términos de su comportamiento, reflejan de manera inmediata los efectos de la economía sobre el sector de actividad al que se refieren.

5.2. La Elección del Período Base

La base debe ser seleccionada en forma cuidadosa de modo que no surjan resultados e interpretaciones erróneas. Existen dos reglas básicas a seguir en la selección de la base:

1. Que el valor de la base sea **"típico o normal"** en relación con el conjunto de valores de la serie. Es decir, ni demasiado alto ni demasiado bajo en relación con los valores de los demás períodos ya que si esto ocurriera el índice aparecerá crónicamente depreciado o sobreestimado según el caso. El valor de la base puede considerarse típico **si coincide con la tendencia general de la serie**.
2. El valor de la base debe ser **relativamente reciente**. Un período base muy alejado en el pasado hace a los números índices recientes menos representativos porque los valores individuales contenidos en el índice tienden a dispersarse con el tiempo. Además, las ponderaciones deben ser actualizadas ya que interesa comparar las fluctuaciones con algún cuadro de referencia similar al actual.

5.3. La Ponderación Adecuada

Puede observarse que **solo se requiere una exactitud aproximada en las ponderaciones** para que un índice sea útil en la práctica. Cada procedimiento de ponderación tiene sus méritos teóricos y prácticos, como así también sus inconvenientes, siendo importante observar que:

- al cambiar la ponderación también cambia el significado del índice; por lo tanto la ponderación **depende del tipo de pregunta que deseamos responder;**
- cuando dos tipos de ponderaciones pueden rendir información similar, se podrá **recurrir a la que requiere menos esfuerzo de cálculo o permite una interpretación más precisa o proporcione una mayor consistencia teórica**.

5.4. La Selección del Promedio

Desde un punto de vista estrictamente matemático al promediar relativos, la media geométrica o armónica resultarían más eficientes que **la media aritmética**. Sin embargo esta última **es la más utilizada por su facilidad de cálculo y**, fundamentalmente, porque **su significado es más fácil de interpretar**.

La representatividad de los promedios obtenidos depende de la forma de distribución de los relativos; si los valores están ampliamente dispersos el índice puede perder significado. Al respecto **se ha demostrado que los relativos calculados a partir de una base reciente tienen una pronunciada tendencia central y la proporción de relativos bajo la clase modal es grande.** Cuando más remota es la base, la distribución se hace más dispersa y negativamente asimétrica, con una proporción menor de relativos en la clase modal. Esto sugiere que el índice es más representativo cuando la base es más reciente.

También se observa una tendencia central más marcada en grupos de ítems que son más homogéneos (Ej.: productos agrícolas, bienes durables a los consumidores, etc.). Por lo tanto, en forma ideal, un índice -como cualquier otro promedio- debería ir acompañado de una medida de dispersión.

5.5. Los Cambios de Producto

En una economía dinámica, los bienes son reemplazados permanentemente por productos nuevos. Puesto que la significación de un índice depende de la constancia de significado del surtido de bienes que lo conforman, la comparación de los niveles de precios o cantidades a partir de puntos distantes en el tiempo puede ser de difícil interpretación o carente de sentido. Para atender este tipo de problemas se utilizan los índices en cadena, con todas las dificultades que ello acarrea según se viera precedentemente.

Por otra parte, mediante estos procedimientos **no es posible** presentar **evidencia cuantitativa que permita observar los cambios en la calidad de los productos**.

6. ¿Qué Hemos Visto?

Hemos desarrollado en esta unidad distintas maneras de obtener **números índices**; estos índices, que en rigor constituyen **maneras particulares de promediar magnitudes**, son una forma clásica y difundida de **analizar y presentar la evolución de diferentes series**, particularmente aquellas que se refieren a precios, cantidades y valores. En la presentación quedó expresado que este recurso es válido para analizar series de tiempo como así también para realizar el análisis de otro tipo de series numéricas.

Así, tomando como ejemplos series de tiempo, fueron presentados teórica y prácticamente diferentes tipos de números **índices simples** (para una sola variable) y **compuestos** (dos o más variables) **ponderados y no ponderados**, realizando en cada caso la interpretación de los valores obtenidos y expresando los alcances y limitaciones de las fórmulas utilizadas.

Se consideraron además algunas **cuestiones vinculadas a la utilización** de los números índices y otras que se refieren a problemas que se deben atender en **la construcción** de los mismos.

Bibliografía

ANDERSON, D. R, SWEENEY, D. J., WILLIAMS, T. A.: *Estadística para Administración y Economía***.** 7ª Edición. Cap. 17. Internacional Thomson Editores. México, 1999.

FERRUCCI, Ricardo J.: *Instrumental para el Estudio de la Economía Argentina*. Cap. 3. EUDEBA, Buenos Aires. 1990.

FREDIANI, Ramón O.: *Medición del Desarrollo Económico y Social de las Provincias Argentinas*, CIPESP. 1979.

YA-LUN CHOU: *Análisis Estadístico*. Edit. Interamericana. México. 1972.

YAMANE, Taro: *Estadística*, Edit. Harla S.A. México. 1974.

Conceptos Centrales de esta Unidad

- Números índices: concepto y utilidad.
- Números índices simples: diferentes tipos, concepto y propiedades.
- Números índices compuestos ponderados y no ponderados: diferentes tipos, concepto y propiedades.
- Valor y deflactación de una serie (precios constantes).

Habilidades

- Saber construir los diferentes tipos de números índices.
- Conocer los alcances y limitaciones de las fórmulas utilizadas.
- Poder analizar, interpretar e informar sobre los datos obtenidos.

Anexo:
Guías de Actividades

UNIDAD 1: *LA INVESTIGACIÓN ESTADÍSTICA*

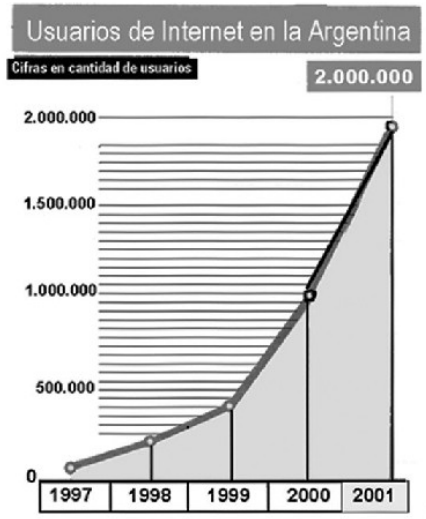
Actividad Nº 1

Hasta la década del '80, el uso de Internet estuvo reservado a especialistas con conocimientos específicos de computación, que utilizaban la red también para fines específicos. La masificación de Internet se da en la década de los '90, con la creación de un entorno "amigable" que facilitaba la comunicación *máquina-usuario.* Desde entonces, su uso en el mundo ha *variado* tanto en número de usuarios, como en las características de las personas usuarias. La *variación* en el tipo de usuarios puede deberse tanto a la disponibilidad de computadoras como al bajo costo del servicio de acceso y la diversidad de usos que ofrece la red de redes, ampliando entonces la gama de personas que se interesan en este nuevo recurso. Pero, ¿qué es lo que ocurre en la Argentina?

A continuación encontrará la trascripción del *artículo* "***Ya hay dos millones de personas conectadas a Internet****",* publicado por el diario Clarín el 24/06/2001 mediante el cual se dan a conocer los resultados de una investigación realizada por el estudio D'Alessio/Irol/Harris. El texto constituye un buen ejemplo de una investigación basada en métodos estadísticos y contiene una serie de elementos que nos permitirán abordar y ejemplificar conceptos centrales de esta parte del curso.

Ud. deberá leer atentamente todo el artículo, registrando los aspectos centrales del informe: ***para qué y cómo se realizó el estudio***, así como ***cuáles son las principales conclusiones a las que arribaron los investigadores***.

Es importante que realice cuidadosamente esta actividad porque la iremos utilizando en la presentación teórica de los temas siguientes.



"Ya hay dos millones de personas conectadas a Internet"

Diario Clarín, 24/06/01

La seducción del ciberespacio se vuelve cada día más irresistible en todo el mundo. Y los argentinos no son la excepción. En el país ya hay 2.000.000 de personas conectadas a Internet. Si bien el número es bajo respecto de las previsiones de algunos especialistas -aseguraban que en 2001 serían 3.000.000 los usuarios de la Red-, la cifra es más que significativa si se compara con 2000. Es que hace sólo un año había 1.000.000 de personas conectadas a Internet. Y en 1997 no llegaban a 70.000. Los datos surgen de una encuesta realizada por el estudio D'Alessio/Irol /Harris, en base a una investigación sobre 1.400 casos. En www.clarin.com se publica una extensa nota con todos los datos de la encuesta, que está acompañada por numerosos gráficos interactivos.

Más de 35 años

Del trabajo se desprenden conclusiones más que interesantes. Por ejemplo, que más del 50 por ciento de los usuarios de la Red tienen más de 35 años. Esto echa por tierra la teoría que sostiene que los "grandes" tienen muchas dificultades para manejar las nuevas tecnologías. Las conclusiones dicen que 4 de cada 10 usuarios son mujeres y que solo el 3 por ciento de los navegantes está desocupado.

Que haya 2.000.000 de argentinos conectados a Internet significa que el 6 por ciento del país tiene acceso a la Red. La cifra es alta comparada con el resto de América latina, donde el promedio ronda el 3 por ciento. Pero es baja en relación a Europa, donde el índice es del 20%. Y no tiene punto de comparación con la realidad de Estados Unidos y Canadá, países en los que el 60 por ciento de la población está conectada a Internet.

Los datos dicen que en la Argentina hay 1.265.000 hogares en los que algún miembro de la familia tiene acceso a la Red. Esto es, en el 14 por ciento de los 9.000.500 hogares del país vive alguien que se conecta habitualmente a Internet, independientemente del lugar. Y esto es así porque, según la encuesta, el 9 por ciento tiene conexión desde

su vivienda y el restante 5 por ciento lo hace desde el trabajo, un cibercafé o un locutorio.

La distribución de los 1.265.000 hogares dentro del país tampoco es demasiado equitativa. Mientras que el 27 por ciento de las viviendas (907.000) de la Ciudad de Buenos Aires y sus alrededores están conectada, en el resto del país solo tiene acceso a Internet el 6 por ciento de las familias, es decir, 357.000. Esta es una constante en países como Brasil o México, donde las grandes ciudades (Río de Janeiro, San Pablo o el Distrito Federal) concentran la mayor cantidad de usuarios.

Focalizando el estudio en la región metropolitana de Buenos Aires, se deduce que el uso exclusivo en el hogar está estrechamente relacionado con el nivel socioeconómico. En los sectores más altos (ABC1 y C2) prevalece la conexión desde la vivienda propia, mientras que en los segmentos medios y bajos sucede a la inversa: la mayoría se conecta desde el trabajo o un cibercafé. Y esto es así ya que no todos tienen una computadora en el hogar. También se da, por supuesto, el hecho de que hay muchos usuarios que se conectan en más de un lugar. Durante el día, por ejemplo, lo hacen desde el trabajo y por la noche desde su casa. Según la encuesta, la frecuencia de conexión de los usuarios argentinos es mayor del promedio mundial. El 68% de los entrevistados aseguró ingresar a Internet todos los días, mientras que el uno % lo hace menos de una vez por semana.

Quienes más horas navegan por la Red son los usuarios que cuentan con más antigüedad en el ciberespacio: el 85 por ciento de los que se conectan todos los días lleva más de 5 años navegando. En cuanto a la cantidad de horas de conexión a la red, el estudio habla de tres tipos de usuarios: el "heavy" (más de 4 horas todos los días), el "medium" (de 2 a 3 horas y entre 4 y 6 días) y el "light" (menos de una hora y media entre 1 y 3 días). Los usuarios del primero y segundo tipo suman el 88 por ciento del total.

Usuarios 2001

La investigación da por tierra con el mito de que la Web es para adolescentes. Los números muestran que la franja que va de 25 a 34 años concentra la mayoría de conectados. Son cerca de 640.000, es decir, el 32 por ciento. También es llamativo que el 50 por ciento de los usuarios tiene más de 35 años. Estos datos relativizan los prejuicios tecnológicos que hay con respecto a Internet, como que los mayores se sienten "trabados" para ingresar a la Red. Si bien en los comienzos del ciberespacio la gran mayoría de los navegantes eran hombres, hoy en día el 40 por ciento de los usuarios argentinos son mujeres. En cuanto al perfil del navegante, el 97 por ciento de los usuarios trabaja y el 59 por ciento es el principal sostén económico del hogar. El 53 por ciento está en pareja y muchos de ellos también son padres.

Lentamente, y a pesar de las trabas económicas, la clase media también está ingresando al ciberespacio. Se estima que cuenta con 3.000.000 de usuarios, un 16 por ciento del total. Pero de la integración de las franjas media y media - baja depende la expansión de Internet en la Argentina.

INFORME: HORACIO BILBAO
De la Redacción de clarin.com

Actividad Nº 2

Cuando se planifica o se intenta comprender una investigación desarrollada con métodos estadísticos es necesario, por un lado, identificar claramente la situación problemática abordada por el estudio, y que puede precisarse en alguna o varias preguntas de investigación. Simultáneamente, es necesario definir con precisión (o reconocer) algunas características del trabajo (*población, unidad de análisis*, etc.) para determinar el alcance que podemos dar a la interpretación de los resultados.

El objetivo de esta actividad es que analice el artículo siguiente e identifique tanto los **aspectos relativos al problema de investigación** como aquellos **conceptos estadísticos** necesarios para evaluar o comprender los resultados.

"En promedio, hay entre ocho y nueve árboles por cuadra en Buenos Aires"

Diario Clarín, 08/07/01

Sin contar los que están en plazas y parques, suman más de 400.000 ejemplares. Hay unas 500 especies distintas. El más abundante es el fresno, con el 40% del total. El 13% del arbolado urbano sufre alguna enfermedad.

Mayoría de fresnos, cientos de plantas exóticas. Abundante presencia de palmeras y muchos árboles afectados por cables y zanjas que dificultan su crecimiento. Estos son, a grandes rasgos, los primeros resultados del censo de árboles que hace un año puso en marcha la Secretaría de Medio Ambiente del Gobierno porteño junto con las Facultades de Agronomía y Ciencias Exactas de la UBA y la empresa Sistemas Catastrales S.A. Esta es la primera vez, desde que se comenzó con el arbolado urbano a fines del siglo XIX, que se realiza un censo global sobre la cantidad

de árboles de la ciudad y el estado en que se encuentran.

El trabajo es minucioso y se realizó recorriendo las 12 mil manzanas de la ciudad para contabilizar, uno por uno, cada árbol plantado en la veredas, verificar a qué especie pertenece y conocer su estado sanitario. Todos estos datos permitieron obtener un diagnóstico exacto de la situación del arbolado urbano.

Con un promedio de 8 a 9 árboles por cuadra, Buenos Aires cuenta con más de 400 mil ejemplares fuera de los árboles que se encuentran en los espacios verdes. La mayoría son árboles pero también hay gran cantidad de arbustos y un número considerable de palmeras.

En 1885 la ciudad contaba con menos de 100 mil ejemplares. Y la década del 40 fue la de mayor plantación y reposición de ejemplares. Después, distintos factores hicieron disminuir el número de árboles. El censo permitirá mantener datos actualizados para organizar futuras plantaciones.

El crecimiento desmesurado de la ciudad en la últimas décadas impidió un mayor desarrollo de los árboles en las veredas. Entradas de garajes y paradas de colectivos fueron algunos de los obstáculos para ubicar mayor cantidad de árboles en los frentes. "Si bien Buenos Aires tiene un déficit de espacios verdes, la cantidad de árboles que hay en las veredas es razonable", dijo Norberto Laporta secretario de Medio Ambiente de la Ciudad. Sin embargo, las autoridades consideran que haría falta plantar, por lo menos, 25 mil árboles más. "Estos resultados nos van a permitir actuar con mayor certeza sobre los árboles existentes y sobre qué políticas aplicar en el futuro", destacó.

Según los resultados arrojados por el censo, el 13 por ciento del arbolado urbano padece alguna plaga o enfermedad, que se manifiesta principalmente por distintos tipos de cavidades en sus troncos. Sin embargo, el porcentaje de árboles secos es muy bajo: menos del 3 % se encuentra en esa condición.

Uno de los datos más interesantes que surgen del censo es la cantidad de especies que se encuentran en las calles de Buenos Aires: más de 500 distintas, muchas de las cuales exóticas.

El **fresno** es el árbol que más abunda en la ciudad con más del 40% del total de ejemplares. Después le siguen el plátano (9%), el paraíso (8,5%), el ligustro (4%), el tilo (4%) y el ficus benjamina (3%). Pero las especies autóctonas tienen escasa presencia en la ciudad: apenas un 2,2% de tipas y un 2% de jacarandaes.

El predominio del fresno, originario de América del Norte, tiene que ver con su resistencia a las plagas y enfermedades y con su crecimiento rápido. Por eso, a partir e la década del ´80, se decidió plantarlo en reemplazo de sauces, álamos y gomeros, que habían provocado problemas con las cañerías subterráneas y en las veredas por sus raíces invasivas.

El censo mostró algunos datos, por lo menos, llamativos. Uno es la importante presencia de **palmeras** en la ciudad: más de 1000 ejemplares, la mayoría del tipo pindó, una especie autóctona del norte del país. Las palmeras se encontraron principalmente en zonas como Villa Devoto y alrededor de varios centros comerciales.

El otro dato llamativo es la gran cantidad de **plantas y árboles exóticos** que hay en las veredas porteñas. El censo descubrió especies tropicales como el mango y la guayaba, algunos ginko bilobas, originarios de la China, aloe vera y otras más raras, sobre todo para desarrollarse en el reducido espacio de un cantero, como las secuoias y las araucarias.

"Este fenómeno está relacionado con la **intervención directa de los vecinos** que muchas veces plantan ejemplares sin conocer cuáles son los más adecuados para cada lugar. Los árboles se adaptan, pero terminan sacrificados por las condiciones en que deben crecer", explicó Gabriela Campari, coordinadora general del censo.

Además, algunos factores como el **cambio climático** que sufrió la ciudad en los últimos años permitió el desarrollo de ciertas especies – como las tropicales- que, décadas atrás, no hubieran crecido.

De acuerdo con el censo, el 42% de los árboles porteños tiene entre **20 y 30 años.** Y un 12% son añejos, de más de 60 años.

Por otra parte, un 18 % de los árboles que viven en la ciudad padece algún tipo de interferencia que afecta su normal crecimiento. Cables aéreos, zanjas subterráneas, veredas rotas por las empresas de servicios, entre otras causas, interfieren en el desarrollo de los árboles. Y otro 13% sufre algún tipo de **maltrato** por los carteles o cestos de residuos clavados en sus troncos, pintadas con aerosol, hilos de pasacalles y troncos destrozados o quemados. La mayoría de los ejemplares suele adaptarse pero termina teniendo una vida útil mucho más corta.

Los barrios de casa más bajas mostraron una mayor presencia de árboles, mientras que el micro y macro centro todavía son áreas donde el gris predomina sobre el verde.

Los ficus rompen las veredas y obstruyen los desagües

"Este censo permitirá aplicar políticas más pedagógicas entre los vecinos que suelen realizar podas clandestinas o plantar especies poco recomendables", dijo Fabio Márquez, coordinador del área de nuevos paisajes verdes de la Dirección de Espacios verdes porteña.

En este sentido, los mayores problemas los causa el ficus, de la familia del gomero. El ficus puede crecer hasta el tamaño de un ombú y sus raíces suelen romper las veredas. Además, sus hojas, que contiene látex, se pudren mucho más tarde que las de otras especies y permanecen obstruyendo los desagües por más tiempo.

Por otra parte, los árboles constituyen el patrimonio natural de la ciudad y, en muchos casos, tardaron varias

décadas en alcanzar su plenitud. "Sin embargo, hay gente que no lo entiende. Es el caso de algunos comerciantes que sacan árboles de su frente porque les tapan las marquesinas – contó Márquez -. Por eso, a partir de ahora, se va a tener en cuenta un criterio más paisajístico, con cada plantación. No es lo mismo una calle con edificios torre que una con casas bajas; una avenida que una callecita más angosta".

Por ejemplo, a partir de la remodelación prevista para la avenida Corrientes, donde hoy no hay casi árboles, se van a plantar decenas de ejemplares de ibirapitá, un árbol autóctono que crece relativamente rápido y resiste bastante la contaminación.

Hoy, dos de las especies más frecuentes en la ciudad, el paraíso y el plátano, o bien son muy poco resistentes a la polución (los paraísos) o, aunque resisten más que otros la contaminación (los plátanos), provocan alergias en las personas.

Resultados de una encuesta entre 1.500 vecinos

Mientras se realizaba el primer censo de árboles de la ciudad, la Secretaría de Medio Ambiente porteña organizó una encuesta entre vecinos (realizada por los Centros de Gestión y Participación y los auxiliares vecinales en las calles). El objetivo era conocer la opinión de los ciudadanos sobre los árboles que tiene en los frentes de su casas. Sobre mas de 1.500 encuestados, éstos fueron los resultados.

- La mayoría cree que el mayor beneficio que le aportan los árboles es la sombra. También la oxigenación, la belleza y la amortiguación de los ruidos.
- El 90% de los consultados prefiere las especies que dan sombra y las que tienen flores llamativas.
- Entre los principales problemas enumerados por los encuestados, en primer lugar está la alergia (casi siempre coincide con la presencia de plátanos), y en segundo lugar, la obstrucción de los desagües por las hojas.
- Otros encuestados se quejan de las ramas que tapan las luces y algunos temen la caída de los ejemplares. No faltan tampoco los que se quejan porque las hojas le ensucian el auto.
- Muchos admiten que ellos mismos plantaron el ejemplar frente a su casa. Y la mayoría eligió el ficus.
- Algunos chicos admiten que les gustan los árboles porque pueden treparse. Y otros confiesan que les asustan las sombras de sus copas.

Diario Clarín, 08/07/01

Leyendo atentamente el artículo…

1. ¿Cuáles fueron los *motivos* que condujeron a realizar este estudio? En otras palabras, ¿cuál es la importancia de los resultados de este trabajo?
2. El artículo podría *titularse con la pregunta general* que orientó la investigación. ¿Qué pregunta elegiría Ud. como título para esta nota?
3. Para poner un subtítulo podría *desagregarse esa pregunta general en varias preguntas* que ilustren sobre aspectos más específicos de este trabajo de investigación. Proponga algunas sub- preguntas.
4. Defina con la mayor precisión posible, ¿a *"qué"* o *"quiénes"* (objetos o sujetos) se está describiendo en este estudio? (Unidad de análisis).
5. Defina el *conjunto total de esos elementos* a los que se refiere la investigación (Población bajo estudio).
6. ¿Cuáles son las características o *variables* de esos elementos que se consideraron relevantes para responder los objetivos propuestos?
7. ¿A qué *tipo de variable* (numérica o categórica) corresponde cada una de las identificadas en el punto anterior?
8. Basándose en la lectura de los resultados del estudio, identifique algunas de las *preguntas estadísticas*, en que se tradujeron las preguntas de investigación.

Actividad Nº 3

Para continuar el análisis del artículo de la actividad anterior: "***En promedio, hay entre ocho y nueve árboles por cuadra en Buenos Aires",*** deberá responder a las siguientes preguntas:

Leyendo atentamente el artículo...

1. ¿Cómo fueron obtenidos los datos? (observación transversal o longitudinal; censo o muestra).
2. A modo de síntesis del análisis anterior complete el siguiente cuadro (que en lo sucesivo denominaremos "*Ficha técnica*"). Esta *ficha* indica algunas características de la investigación, que es fundamental tomar en cuenta para evaluar el alcance de las conclusiones de cualquier investigación estadística. Realice esta tarea con la información disponible en el texto. Es posible que no tenga toda la información necesaria, cuando esto sea así, déjelo indicado.

Fuente (Organización/es que realizó el estudio):
Reseña de los objetivos:
Población:
Unidad de análisis:
Fuentes de datos utilizadas (primarias o secundarias):
Tipo de observación realizada (transversal o longitudinal):
Tipo de estudio (enumeración completa o muestra):
Tamaño de la población:
Tamaño de la muestra *:
Fecha de realización:
si corresponde

Actividad Nº 4

En esta actividad encontrará distintos ejemplos que le permitirán revisar sus conocimientos sobre los principales temas tratados en la Unidad Nº 1.

A. Una aerolínea distribuye entre los pasajeros que embarcan a uno de sus vuelos (Vuelo BA 178), el siguiente cuestionario:

OPINIÓN DE LOS PASAJEROS DEL VUELO BA 178

Formulario nº: .....

Sr. Pasajero: como nuestra intención es seguir mejorando nuestros servicios, le rogamos complete este formulario y lo entregue a nuestro personal.

1) Tiempo de espera para el Check-in (en minutos): .....

2) Califique como *Muy Bueno, Bueno, Regular, Malo, Muy Malo* a los siguientes aspectos del servicio:

	MB	B	R	M	MM
a. **Cordialidad** del personal en el Check-in	☐	☐	☐	☐	☐
b. **Información** recibida en el Check-in	☐	☐	☐	☐	☐
c. **Anuncios** para el embarque	☐	☐	☐	☐	☐
d. **Cordialidad** del personal de embarque	☐	☐	☐	☐	☐

Muchas gracias

1. ¿Cuál es la unidad de análisis a la que se referirán los datos recogidos mediante el cuestionario?
2. ¿Cuántas variables fueron observadas?
3. ¿A qué tipo de variables (cuantitativa o cualitativa) corresponde cada una?
4. ¿Cuáles son los valores posibles de cada una de las variables en observación?

B. Cada una de las expresiones siguientes corresponden a resultados de alguna investigación basada en herramientas estadísticas.

Identifique en cada caso: la *unidad de análisis* a la que se refieren, la *variable* en estudio, y el *tipo de variable* (numérica o categórica).

- Un 50% de los jubilados cobra menos de $500 mensuales.
- El 74% de los egresados universitarios ve poco probable la pérdida de su empleo.
- El 27,8% del total de los hogares del país tiene como jefe de hogar a una mujer.
- En la Argentina, hay más mujeres que hombres.
- En la pcia. de Mnes. hay 14 municipios de primera categoría (más de 10.000 habitantes).
- En el Gran Bs. As. el 40% de los mayores de 60 años vive en hogares de tres o más personas.
- En el 85% de los municipios de la provincia de Misiones los varones son mayoría.
- En el año 1996, siete universidades del país registraron más de 10.000 nuevos inscriptos.
- En general, las mujeres trabajan menos horas que los hombres.
- La mayoría de los turistas que visitaron Bs. As. en Semana Santa, llegaron en auto o micro y una cuarta parte de esos turistas eran extranjeros.

C. Si realizáramos una encuesta ***a todos*** los estudiantes que están cursando en el presente año, Estadística I en la Fac. de Hum. y Cs. Soc. de la UNaM, con el propósito de conocer sus características personales y ocupacionales. Le preguntamos entre otras cuestiones:

- *¿Qué edad tiene usted?* (en años cumplidos).
- ¿Trabaja?
- ¿Cuántas horas semanales trabaja? (si *no trabaja* escriba 0).
- ¿En qué ciudad reside habitualmente?
- ¿Dispone de conexión a Internet en su casa?

a. ¿Cuáles son las **variables** que estudiaríamos en este ejemplo?
b. Para cada una de esas variables indique su **tipo** (numérica: continua - discreta o categórica: nominal-ordinal).
c. Describa o indique los valores posibles de cada las variables anteriores.
d. ¿Cuál es la **población** en estudio? ¿Cuál la unidad de análisis?
e. Proponga **otras variables** que considere relevantes al propósito del trabajo: una nominal, una ordinal, una numérica; indicando para cada una de ellas sus valores posibles.
f. ¿Cómo calificaría a esta forma de observación: **transversal o longitudinal**?
g. Tal como está propuesto el trabajo, ¿se trata de una observación por **muestra o enumeración completa**?
h. En las condiciones mencionadas en el punto anterior, ¿será necesario realizar "***inferencias estadísticas***"? Justifique.

D. El gerente de una importante agencia de viajes se propone diseñar una estrategia de ventas para la próxima temporada alta de invierno. Para ello, y con el propósito de conocer mejor las características y preferencias de sus clientes, realizará un estudio a partir de los datos que registra la agencia en la *"Base de Clientes".* Ha decidido trabajar solamente con aquellos que registran alguna operación (compra de pasajes, de excursiones, reservas hoteleras, etc.), realizada en la temporada Diciembre-Febrero de 2002.

Nuestro gerente se ha planteado algunas *preguntas generales* que guiarán su trabajo, y las ha concretado en otras *preguntas estadísticas* para orientar la búsqueda y el análisis de los datos.

En el listado siguiente, aparecen *mezcladas* las preguntas de uno y otro tipo.

- ¿Qué tipo de productos compran?
- ¿Son mayoritariamente grupos familiares?
- ¿Cuántas personas viajan solas?
- ¿Quiénes son nuestros clientes?; ¿Cuál es el perfil sociodemográfico de mis clientes?
- ¿Qué porcentaje de clientes compran únicamente billetes de avión?
- ¿Qué edad tienen nuestros clientes?, ¿predominan los jóvenes o los adultos?
- ¿Son los clientes jóvenes los que eligen más frecuentemente los viajes al exterior?
- ¿Qué forma de pago eligen?
- ¿Pagan mayoritariamente en efectivo o con tarjeta?
- ¿Prefieren pagar en cuotas?
- ¿Cuál es el rango de gasto más frecuente?
- ¿Qué proporción de clientes gastan más de $2.000?

a. Reconozca en cada una si se trata de una pregunta de investigación o una pregunta estadística.

b. Para cada pregunta estadística, defina la o las variables para las que Ud. recogería datos de los registros de la empresa.

c. Proponga otras preguntas estadísticas para cada pregunta general de investigación.

d. ¿De qué tipo es la fuente de datos que se utilizaría en este trabajo?

e. ¿Cómo definiría usted a la población en estudio?

EVALUACIÓN PARCIAL -Unidad I-

Dos son los propósitos de esta actividad. El primero es ofrecerle a Ud. un problema de trabajo que le permitirá **revisar y ejercitar sus conocimientos** sobre los temas de esta primera unidad. El segundo propósito es permitirnos a los docentes **evaluar sus avances** en el aprendizaje.

El texto que se transcribe a continuación fue extraído del informe final de la investigación *"Satisfacción del Cliente"*, realizada por la Licenciada en Turismo CRISTINA TETZLAFF (diciembre de 1999), como Monografía de Grado para alcanzar el diploma de licenciatura.

El estudio tiene por tema general el análisis y evaluación de la "calidad del servicio" que presta a sus pasajeros un importante hotel de la ciudad de Posadas[1]. En la presentación del estudio, la autora señala que *"el presente trabajo, por consiguiente, puede ser de gran utilidad para el hotel, por cuanto se tratará de determinar el grado de satisfacción de los clientes para la posterior elaboración de un Plan de Marketing, proponiendo estrategias tendientes a satisfacer las expectativas y necesidades del cliente y consecuentemente cumplir con las metas de la empresa".*

Los objetivos generales propuestos a tales fines son los siguientes:

[1] Al que identificaremos como "NHTL" (nuestro hotel).

"*1. Detectar y analizar el grado de satisfacción que generan en los huéspedes del hotel "NHTL", los servicios prestados por el mismo.*

2. Proporcionar información que permita elaborar y desarrollar estrategias para el mejoramiento de la calidad y, consecuentemente, de marketing."

La investigación se refiere a los huéspedes del hotel, registrados durante los meses de marzo, abril y julio de 1998. Ante las limitaciones de tiempo y recursos para realizar el trabajo, se observó una muestra fortuita de 150 pasajeros en el período señalado.

Además de los correspondientes a la encuesta, y con el propósito de analizar la evolución de la demanda del hotel en el período 1994-1998, se utilizaron los datos sobre ocupación mensual de NHTL que se presentan en la tabla siguiente.

Porcentaje mensual de ocupación de NHTL
Enero de 1994/Octubre de 1998[2]

Mes	1994	1995	1996	1997	1998
Enero	34,52	27,58	22,77	22,55	26,80
Febrero	37,11	24,25	22,26	21,00	32,80
Marzo	35,55	33,52	26,68	26,71	43,53
Abril	45,57	23,05	25,80	37,83	40,99
Mayo	39,06	20,74	27,13	37,13	29,59
Junio	44,23	22,34	31,43	32,87	23,05
Julio	55,45	33,33	38,97	48,87	47,58
Agosto	49,29	26,59	39,42	36,74	44,09
Septiembre	45,40	27,50	34,03	40,17	51,71
Octubre	43,62	37,06	35,68	29,84	37,57
Noviembre	41,15	32,00	31,43	34,83	
Diciembre	39,48	26,74	29,69	32,39	
TOTAL	42,57	27,90	30,44	33,41	

En relación con este aspecto el informe expresa: "... *el año de mayor ocupación ha sido 1994, esto es dos años después de haber iniciado sus actividades el hotel*.

En el año 1995 hubo una fuerte capacidad ociosa durante todo el período. Las causas de este fenómeno pudieron haber sido varias, una de ellas el cierre del hotel durante una semana en el mes de abril para la regularización de ciertos aspectos internos de la empresa. Otra pudo haber sido la caída de la bolsa mexicana, cuya repercusión, conocida como "efecto tequila" afectó la economía de muchos países del mundo entre ellos la de la Argentina. ... [*en*] *los años 1996, 1997 y parcialmente a 1998, se observa como fue recuperándose lentamente el hotel después de su caída de 1995."*

En cuanto al **"Perfil de los huéspedes",** y basado en el análisis de los datos de las encuestas, en el informe se señalan las siguientes conclusiones:

Tipo de Huéspedes de NHTL

Tipo	Pasajeros	%
Habitual	53	35,3
No Habitual	97	64,7
TOTAL	**150**	**100,0**

"De acuerdo con el criterio tomado para segmentar a los huéspedes, el 35% de ellos son habituales. Esto está directamente relacionado con el motivo de visita, ya que en su mayoría son hombres de negocios, que vienen a la ciudad de Posadas por razones de trabajo.

Debido a este motivo laboral de visita, la mayoría de los huéspedes se hospeda solo en el hotel y un considerable porcentaje lo hace con colegas.

[2] Los porcentajes de ocupación están basados en las **habitaciones** ocupadas, independientemente del número de camas con que cuente cada una de ellas.

Si tenemos en cuenta el lugar de origen de los encuestados, vemos que el 91% reside en la Argentina, de los cuales el 50% proviene de Bs. As. (Capital y Gran Bs. As.), el 11% de la provincia de Corrientes, y el porcentaje restante de Santa Fe y Chaco.

Al margen del porcentaje de encuestados que reside habitualmente en nuestro país, el 3% proviene del Paraguay y los demás son oriundos de Brasil, Chile, EEUU y España.

Finalmente, en cuanto al perfil de la demanda, podemos decir que la mayoría de los huéspedes del hotel NHTL tiene entre 30 y 50 años de edad, y que el 86% es de sexo masculino, lo cual está relacionado con el motivo de visita, negocios.

En cuanto a la ocupación o profesión de la demanda las encuestas dieron como resultado que el 52% son profesionales, el 25% son empleados y el porcentaje restante lo conforman comerciantes, empresarios, gerentes y otros. Este alto porcentaje de profesionales y personas calificadas en general, nos da la pauta de la importancia que tienen sus opiniones en cuanto a la calidad de los servicios, por el hecho de que generalmente ya conocen otros hoteles, ya sea a nivel nacional o Internacional, y por lo tanto son conocedores de los servicios que debe brindar un hotel de categoría cuatro estrellas".

Con respecto al ***Grado de satisfacción*** de los huéspedes del hotel NHTL, el informe señala:

"El grado de satisfacción de un cliente depende de la relación entre las expectativas que tenía respecto a lo que pensaba que debía recibir y las percepciones sobre lo que recibió".

¿Logró Satisfacer sus expectativas?

Grado de Satisfacción	Pasajeros	%
Superó sus expectativas	6	4,0
Logró satisfacer	131	87,0
No logró satisfacer	13	9,0
TOTAL	**150**	**100,0**

Cuando se consultó a los huéspedes respecto a si lograron satisfacer sus expectativas, el 87% respondió que sí logró satisfacerlas, el 9% que no logró satisfacción, y el 4% sostuvo que sus expectativas hacia el hotel fueron superadas.

Al solicitar que califiquen su experiencia en el hotel, los huéspedes sostuvieron que esta fue buena o muy buena, en porcentajes similares (aproximadamente el 45% para cada categoría de respuesta). Y respecto a si volverían a alojarse en el hotel, el 99,3% sostuvo que sí lo haría".

1. El informe anterior se basa en diversos conjuntos de datos, algunos de ellos originados en la observación transversal y otros en la longitudinal. ***Deberá identificar*** cuáles son los ***datos transversales*** y cuáles los ***datos longitudinales*** utilizados.
2. En el caso de los datos longitudinales, ***deberá identificar***: unidad de análisis observada, variable en estudio y período de la serie de datos utilizada.
3. Elaborar una ***síntesis metodológica*** de la encuesta realizada a los pasajeros de NHTL, ***indicando***:
   - *unidad de análisis observada*,
   - *población en estudio*,
   - *alcance del relevamiento* (enumeración completa, por muestra), *tipo y tamaño de la muestra* (si corresponde).
4. ***Identificar*** en el texto y ***listar todas las variables*** utilizadas para describir el *"Perfil de los huéspedes"* y su *"Grado de satisfacción"*.
5. Para ***cada una de las variables listadas*** en el punto anterior agregar con todo el detalle posible:
   - definición de la variable en cuestión y su tipo (numérica: discreta, continua, categórica: nominal, ordinal),
   - identificación de los *"valores"* que son mencionados en el texto.

UNIDAD 2: *ORGANIZACIÓN Y DESCRIPCIÓN INICIAL DE LOS DATOS*

Actividad Nº 1

En la Guía de Actividades de la Unidad 1 hemos trabajado sobre una encuesta realizada a los pasajeros de una aerolínea (Actividad 4.A). A continuación presentamos los formularios completados por algunos de los pasajeros:

OPINIÓN DE LOS PASAJEROS DEL VUELO BA 178

Formulario nº: ...**1**....

Sr. Pasajero: como nuestra intención es seguir mejorando nuestros servicios, le rogamos complete este formulario y lo entregue a nuestro personal.

1) Tiempo de espera para el Check-in (en minutos): ..**60**.....

2) Califique como *Muy Bueno, Bueno, Regular, Malo, Muy Malo* a los siguientes aspectos del servicio:

	MB	B	R	M	MM
a. **Cordialidad** del personal en el Check-in	X				
b. **Información** recibida en el Check-in		X			
c. **Anuncios** para el embarque		X			
d. **Cordialidad** del personal de embarque			X		

Muchas gracias

OPINIÓN DE LOS PASAJEROS DEL VUELO BA 178

Formulario nº: ...**2**....

Sr. Pasajero: como nuestra intención es seguir mejorando nuestros servicios, le rogamos complete este formulario y lo entregue a nuestro personal.

1) Tiempo de espera para el Check-in (en minutos): ..**80**.....

2) Califique como *Muy Bueno, Bueno, Regular, Malo, Muy Malo* a los siguientes aspectos del servicio:

	MB	B	R	M	MM
a. **Cordialidad** del personal en el Check-in		X			
b. **Información** recibida en el Check-in	X				
c. **Anuncios** para el embarque		X			
d. **Cordialidad** del personal de embarque			X		

Muchas gracias

OPINIÓN DE LOS PASAJEROS DEL VUELO BA 178

Formulario nº: ...**3**....

Sr. Pasajero: como nuestra intención es seguir mejorando nuestros servicios, le rogamos complete este formulario y lo entregue a nuestro personal.

1) Tiempo de espera para el Check-in (en minutos): ..**45**.....

2) Califique como *Muy Bueno, Bueno, Regular, Malo, Muy Malo* a los siguientes aspectos del servicio:

	MB	B	R	M	MM
a. **Cordialidad** del personal en el Check-in	X				
b. **Información** recibida en el Check-in		X			
c. **Anuncios** para el embarque		X			
d. **Cordialidad** del personal de embarque		X			

Muchas gracias

.....

.....

OPINIÓN DE LOS PASAJEROS DEL VUELO BA 178

Formulario nº: ...**136**....

Sr. Pasajero: como nuestra intención es seguir mejorando nuestros servicios, le rogamos complete este formulario y lo entregue a nuestro personal.

1) Tiempo de espera para el Check-in (en minutos): ..**120**.....

2) Califique como *Muy Bueno, Bueno, Regular, Malo, Muy Malo* a los siguientes aspectos del servicio:

	MB	B	R	M	MM
a. **Cordialidad** del personal en el Check-in		X			
b. **Información** recibida en el Check-in			X		
c. **Anuncios** para el embarque		X			
d. **Cordialidad** del personal de embarque	X				

Muchas gracias

Basándose en los formularios, construya una matriz de datos para organizar esta información y complete con los datos de los formularios anteriores

Actividad Nº 2

Durante el primer semestre de 2000, el movimiento internacional de pasajeros extranjeros que llegaron a la ciudad de Buenos Aires por los Aeropuertos Jorge Newbery, y Ezeiza, y el Puerto de Buenos Aires totalizó 1.934.854 personas. Estos extranjeros que ingresaron al país estaban conformados por 959.688 turistas procedentes de países del Mercosur, 205.095 chilenos, 162.528 provenientes del resto de América Latina, 274.749 de Estados Unidos y Canadá, 286.358 de Europa, y el resto de los pasajeros provienen de "otros países".

(**Fuente**: Sec. de Desarrollo Económico del Gob. de la ciudad de Bs. As., basándose en datos del INDEC).

Basándose en la información del texto anterior, construir una tabla de distribuciones de frecuencias que resuma esos datos y el gráfico que considere más apropiado para presentar el aporte turístico de los diferentes países o regiones.

Actividad Nº 3

Durante el mes de mayo/94 se desarrolló una encuesta por muestreo en el Parque Nacional de Iguazú, con el objeto de recabar información sobre los hábitos de los turistas que visitaban este recurso. Entre otras cuestiones, se les indagaba sobre la cantidad de noches (pernoctes) que pensaban permanecer en la región. Seguidamente se detallan los datos obtenidos sobre esta variable, correspondientes a cada una de las 156 encuestas realizadas en esa oportunidad.

Sobre la base de los datos presentados a continuación, construir la distribución de frecuencias en sus formas numérica y gráfica, y describir el comportamiento de los turistas de la muestra en relación con esta característica.

encuesta	noches	encuesta	noches	encuesta	noches	encuesta	noches	encuesta	noches
1	1	33	2	65	1	97	1	129	3
2	1	34	1	66	2	98	2	130	1
3	1	35	3	67	1	99	1	131	3
4	1	36	1	68	3	100	1	132	2
5	1	37	2	69	2	101	1	133	2
6	1	38	2	70	2	102	1	134	1
7	1	39	1	71	2	103	2	135	2
8	3	40	1	72	1	104	2	136	2
9	2	41	1	73	2	105	1	137	1
10	2	42	3	74	2	106	2	138	1
11	2	43	1	75	3	107	3	139	1
12	3	44	1	76	1	108	3	140	1
13	3	45	2	77	1	109	2	141	1
14	1	46	2	78	1	110	3	142	1
15	1	47	2	79	1	111	3	143	1
16	3	48	1	80	1	112	2	144	3
17	1	49	1	81	1	113	3	145	2
18	2	50	3	82	2	114	1	146	2
19	1	51	1	83	2	115	1	147	2
20	2	52	3	84	3	116	2	148	3
21	1	53	2	85	2	117	2	149	3
22	2	54	2	86	1	118	3	150	2
23	1	55	4	87	1	119	2	151	1
24	2	56	3	88	1	120	1	152	3
25	3	57	3	89	1	121	1	153	2
26	1	58	2	90	2	122	3	154	1
27	2	59	3	91	2	123	1	155	2
28	3	60	1	92	2	124	3	156	2
29	3	61	1	93	2	125	1		
30	1	62	2	94	2	126	1		
31	1	63	2	95	1	127	1		
32	2	64	2	96	2	128	2		

Actividad Nº 4

Tomando en consideración la tabla siguiente:

Distribución de la población por edades – Pcia. de Misiones. 1991

Edad	Población
0-9	219.474
10-19	175.189
20-29	118.516
30-39	101.689
40-49	70.091
50-59	49.739
60-69	32.611
70-79	15.704
80-89	5.001
90 y más	901
TOTAL	**788.915**

Fuente: INDEC- Censo Nac. de Pob. y Vivienda – 1991

Haga un comentario sobre las características más destacables de la distribución de la población de Misiones según edades.

Actividad Nº 5

En el estudio sobre los alumnos de Estadística se preguntó entre otras variables el ingreso mensual del hogar, con el propósito de disponer de un indicador del nivel económico de los estudiantes del curso. Los datos obtenidos se presentan en un arreglo de frecuencias y en un diagrama de tallo-hoja.

Ingreso	Frecuencia (*)
80	1
130	1
145	1
150	2
180	1
200	6
250	1
300	14
340	1
350	6
400	11
450	3
480	1
500	14
550	1
560	1
600	1
650	1
700	3
750	1
800	8
850	1
900	1
1000	9
1100	1
1300	2
1500	2
1600	2
1800	2
2000	5
Total	**104**

(*) 35 estudiantes no declaran el ingreso del hogar.

Ingreso: Diagrama de Tallo-Hoja

```
Frecuencia  Tallo - Hoja

    1        0 .  8
    5        1 .  34558
    7        2 .  0000005
   21        3 .  000000000000004555555
   15        4 .  000000000005558
   16        5 .  0000000000000056
    2        6 .  05
    4        7 .  0005
    9        8 .  000000005
    1        9 .  0
    9       10 .  000000000
    1       11 .  0
    0       12 .
    2       13 .  00
    0       14 .
    2       15 .  00
    9 Extremos    (>=1600)

Ancho del tallo:  100
Cada hoja:       1 caso
```

A partir de ellos:

- Presente los datos en una tabla que resuma mejor los ingresos del hogar de los estudiantes, utilizando el o los criterios que considere más apropiado(s) para definir esos intervalos de clase. Comente las decisiones adoptadas para obtener la tabla anterior.
- Justifique la necesidad de utilizar intervalos de clases para esta distribución.

Actividad Nº 6

- Para la distribución en intervalos de clase de la actividad anterior, construya el histograma y polígono de frecuencias correspondiente.
- Tomando en consideración la tabla y gráficos, comente el comportamiento de la variable ingresos del hogar. Le sugerimos que para esta descripción tome en cuenta tanto la *forma* de la distribución, como los detalles numéricos que pueda aportar desde la lectura de la tabla.

Actividad Nº 7

- Para la tabla construida en la actividad 5, calcule las frecuencias relativas y acumuladas (absolutas y relativas).
- Con estas transformaciones de las frecuencias absolutas, Ud. dispone de otras herramientas de análisis que le permiten enriquecer su lectura anterior de los datos. Escriba nuevamente su comentario sobre los ingresos.

Actividad Nº 8

En el trabajo sobre el curso de Estadística se requirió también el *Nivel de Estudios de la Madre* del estudiante. En la tabla siguiente se presenta la distribución de frecuencias correspondientes.

Estudiantes del curso de Estadística según
Nivel de estudios de la Madre - FHyCS-Año 2001

Nivel Estudios Madre	Frecuencia
Ninguno	2
Primario Incompleto	33
Primario Completo	42
Secundario Incompleto	23
Secundario Completo	14
Terc./Univ. Incompleto	7
Terc./Univ. Completo	15
TOTAL	**136**

(*) Hay 3 estudiantes que no declaran el nivel de estudios de su madre.

- Elabore para esta tabla las transformaciones que considere necesarias, y compare esta distribución con la del nivel de estudios de los padres que se presentó en las notas de cátedra.
- Destaque a manera de conclusión aquellos aspectos que considere más relevantes para describir las semejanzas y diferencias en la educación formal de los padres de los estudiantes.
- Proponga gráficos que expresen las semejanzas y diferencias observadas.

Actividad Nº 9

La distribución que sigue corresponde a jefes de hogares de la ciudad de Posadas, distribuidos según escala de ingreso. Los datos fueron obtenidos por la Encuesta Permanente de Hogares en 1993.

Ingresos monetarios de jefes de hogares – Posadas, 1993

Ingreso ($)	Jefes de Hogares	Jef. de Hog. Acum.	Ingreso Total Acum. ($)
235-280	184	184	47380
280-325	342	526	150835
325-385	2354	2880	987505
385-465	864	3744	1354705
465-545	738	4482	1727395
545-630	546	5028	2048170
630-725	379	5406	2190865
725-835	486	5892	2569945
TOTAL	**5892**		

Fuente: Encuesta Permanente de Hogares

- Construir la Curva de Lorenz.
- Obtener el coeficiente de Gini.
- Interpretar los resultados obtenidos.

Actividad Nº 10
Práctico complementario

En esta actividad encontrará distintos ejemplos que le permitirán revisar sus conocimientos sobre los principales temas tratados en la Unidad Nº 2.

1. ¿Cuál es el propósito de construir tablas de distribuciones de frecuencias?
2. En relación con la matriz de datos ¿cuál es la información que se pierde al construir una tabla de frecuencias? Evalúe la situación para el caso de distribuciones de variables categóricas, arreglos y distribuciones en intervalos de clase.
3. Explique en qué situación se plantearía la necesidad de construir:
   a. una distribución con clases abiertas,
   b. una distribución con clases vacías o intervalos de distinta amplitud,
   c. un gráfico en el cual se "corte" el eje de abscisas.
4. Si la representación del polígono de frecuencias de los ingresos de los empleados de comercio de la provincia de Misiones resulta en una gráfica marcadamente asimétrica a la derecha y la correspondiente a los gerentes de esas mismas empresas da fuertemente asimétrica a la izquierda ¿cuál sería su conclusión sobre los ingresos de empleados y gerentes?
5. En el artículo del mercado de Internet, Ud. puede leer:

> *"...más del 50 por ciento de los usuarios de la Red tienen más de 35 años".* A lo que más adelante se agrega: *"Los números muestran que la franja que va de 25 a 34 años concentra la mayoría de conectados. Son cerca de 640.000, es decir, el 32 por ciento. También es llamativo que el 50 por ciento de los usuarios tiene más de 50 años. Estos datos relativizan los prejuicios tecnológicos que hay con respecto a Internet, como que los mayores se sienten "trabados" para ingresar a la Red*'.

a. ¿Qué variable se analiza en el párrafo?
b. ¿Qué transformaciones de las frecuencias absolutas se necesitaron construir para escribir esas conclusiones?
c. ¿Está de acuerdo con que?:
   - ➔ "... la franja que va de 25 a 34 años concentra la **mayoría** de conectados".
   - ➔ "... el 50 por ciento de los usuarios tiene más de 50 años".

   Justifique.

Según el artículo, hoy en día 4 de cada 10 usuarios son mujeres. Además, "*En cuanto al perfil del navegante, el 97 por ciento de los usuarios trabaja y el 59 por ciento es el principal sostén económico del hogar. El 53 por ciento está en pareja...*".

a. ¿Qué variables se analizan en este párrafo?
b. Con esa información, reconstruya las tablas de frecuencias correspondientes a esas variables.

6. En las tablas siguientes se presentan las distribuciones del plantel de empleados de una empresa de servicios, discriminados por sexo según sus salarios mensuales en mayo de 1996. El propósito de este resumen es analizar si la empresa tiene una política salarial diferencial por sexo.

HOMBRES

Ingreso ($)	Empleados
300 – 450	2
450 – 600	39
600 – 750	126
750 – 900	26
900 – 1050	8
1050 – 1200	20
TOTAL	**221**

MUJERES

Ingreso ($)	Empleados
300 – 450	55
450 – 600	116
600 – 750	32
750 – 900	11
900 – 1050	1
1050 - 1200	1
TOTAL	**216**

a- ¿Cuál es la proporción de hombres con ingresos inferiores a $600? ¿Y la proporción de mujeres?

b- ¿Cuántos hombres y cuántas mujeres ganan por lo menos $900?

c- ¿Entre qué ingresos se ubica la mayoría de los hombres? ¿Y entre cuáles la mayoría de las mujeres?

d- Describa la situación salarial de ambos grupos y señale si a su criterio existe una política de la empresa que establece diferencia salarial entre los sexos.

EVALUACIÓN PARCIAL -Unidad 2-

Dos son los propósitos de esta actividad. El primero es ofrecerle a Ud. un problema de trabajo que le permitirá **revisar y ejercitar sus conocimientos** sobre los temas de esta segunda unidad. El segundo propósito es permitirnos a los docentes **evaluar sus avances** en el aprendizaje.

Esta propuesta está basada en el estudio ESTUR 93/94 realizado por la Fac. de Hum. y Cs. Soc. a solicitud de la Secretaría de Turismo de la Pcia. de Misiones y financiado por el Consejo Federal de Inversiones (CFI). A los efectos de simplificar esta práctica del curso, hemos seleccionado sólo algunos aspectos de la encuesta realizada en los *lugares de alojamiento* a turistas que viajaron por *cuenta propia* en el mes de *febrero*.

A continuación se describen algunas definiciones metodológicas del estudio.

OBJETIVO

Esta encuesta está dirigida a los turistas alojados en la ciudad de Puerto Iguazú, con el fin de conocer sus hábitos y preferencias turísticas, su evaluación sobre diferentes componentes (infraestructura, equipamiento, etc) de la oferta y la estructura y nivel del gasto turístico.

UNIDAD DE ANÁLISIS
Grupos turísticos primarios que se encuentran hospedados en hoteles y establecimientos similares (residenciales, cabañas, etc) y campings. El informante será uno de sus miembros mayor de 16 años y preferentemente el que ejerce el liderazgo del grupo.

VARIABLES SELECCIONADAS:

1) Nivel de alojamiento

1. Nivel I: 4 y 5 estrellas 2. Nivel II: 3 estrellas 3. Nivel III: 2 estrellas
4. Nivel IV: Residenciales 5. Nivel V: Alojamiento en carpas, casa rodante / *motor home*, etc

2) Lugar de residencia: discriminando los residentes en Misiones, en otras provincias argentinas (registradas individualmente) y en otros países, también distinguidos individualmente.

Nota: Los números que aparecen en la matriz de datos son los códigos asignados a cada lugar de residencia. (ejemplo: 54 corresponde a la pcia. de Misiones). Ud. no necesitará el detalle de estos códigos para la tarea que deberá realizar.

3) Medio de transporte para el arribo a la Región:

1. automotor privado 2. ómnibus de línea regular 3. ómnibus servicio especial
4. aéreo en vuelo regular 5. aéreo especial (*charter*) 6. otros 7. Sin Datos

4) Total de componentes del grupo primario entrevistado**,** incluyendo al informante**.**

5) Opinión del informante sobre la arquitectura y urbanización de Pto. Iguazú: discriminando por niveles de satisfacción (1. Buena, 2. Regular, 3. Mala, 4. Sin Opinión).

6) Gasto total efectivamente realizado por el grupo primario, durante el día de permanencia en el Área inmediato anterior a la entrevista

Para las variables seleccionadas presentamos a continuación la matriz de datos y tablas y gráficos para algunas de ellas.

MATRIZ DE DATOS

ENCU	NIVEL	RESI	COMPO	TRANS	ARQ	GTOT
1	2	6	2	4	4	125
2	2	6	2	1	2	75
3	2	6	3	1	1	181
4	3	82	1	2	1	109
5	3	6	4	1	2	202
6	3	18	2	2	2	79
7	3	6	1	2	1	33
8	4	6	3	1	1	71
9	5	6	2	1	1	30
10	1	42	4	1	1	75
11	1	2	2	4	1	34
12	1	2	4	1	2	20
13	1	2	1	4	2	40
14	2	6	2	2	2	113
15	2	102	3	1	3	210
16	3	26	4	4	1	100
17	4	6	2	2	1	76
18	4	6	3	2	2	110
19	1	6	5	1	2	316
20	2	6	6	4	1	125
21	2	6	1	2	3	30
22	3	6	7	1	1	100
23	3	6	2	2	1	62
24	5	6	4	1	1	43
25	5	18	3	1	1	60
26	5	6	2	2	1	29
27	1	2	5	1	2	572

ENCU	NIVEL	RESI	COMPO	TRANS	ARQ	GTOT
28	4	6	2	2	1	140
29	5	6	5	1	1	18
30	5	6	4	1	2	45
31	1	107	2	4	1	205
32	3	6	3	1	2	135
33	4	82	2	1	3	75
34	5	22	4	1	1	00
35	1	54	3	1	1	190
36	1	6	6	1	1	270
37	2	34	5	1	1	140
38	2	6	2	2	1	319
39	2	2	3	2	1	151
40	5	6	1	2	1	24
41	5	6	6	1	1	300
42	2	6	2	1	1	79
43	2	6	5	1	2	150
44	2	14	4	1	1	119
45	2	26	4	4	1	104
46	5	6	3	1	1	59
47	5	6	4	1	4	72
48	1	6	3	1	1	110
49	1	6	4	1	3	20
50	1	2	5	1	3	75
51	1	6	5	1	2	70
52	1	6	5	1	1	75
53	4	6	2	2	1	66
54	2	6	4	1	1	632

(Continuación)

ENCU	NIVEL	RESI	COMPO	TRANS	ARQ	GTOT
55	3	6	1	4	1	129
56	3	6	1	2	1	22
57	4	54	1	2	1	44
58	2	6	3	1	1	131
59	2	14	3	1	1	99
60	2	22	5	1	3	113
61	5	18	2	1	2	41
62	5	6	6	1	1	102
63	1	82	6	1	1	95
64	1	18	7	1	1	95
65	1	2	2	1	1	378
66	1	105	1	4	2	160
67	1	2	4	4	4	18
68	1	2	4	1	1	158
69	2	6	5	1	1	217
70	4	6	4	1	1	100
71	4	110	2	4	1	50
72	4	105	2	2	2	51
73	1	6	2	4	1	143
74	2	14	4	1	2	273
75	2	105	2	4	2	165
76	2	18	6	1	2	165
77	2	54	5	1	2	379
78	4	14	2	2	3	76
79	4	2	1	2	2	37
80	1	30	4	1	1	130
81	3	2	3	2	1	149
82	4	133	2	2	1	72
83	4	128	3	2	1	57
84	4	6	2	2	1	77
85	4	105	1	4	4	42
86	5	82	3	1	1	56

ENCU	NIVEL	RESI	COMPO	TRANS	ARQ	GTOT
87	5	6	4	1	2	67
88	4	26	2	1	1	06
89	4	104	2	2	2	62
90	4	110	2	4	1	79
91	5	6	2	1	1	55
92	1	2	3	2	1	260
93	1	6	2	1	1	380
94	4	82	3	1	1	68
95	4	6	5	1	2	275
96	4	14	4	1	1	148
97	2	50	4	1	1	287
98	2	50	3	1	1	163
99	4	82	3	1	1	197
100	4	6	2	2	1	36
101	4	6	4	1	1	95
102	1	6	2	1	2	222
103	2	6	2	1	1	340
104	3	6	3	2	1	158
105	3	14	5	1	1	180
106	1	6	3	1	1	194
107	1	2	4	1	2	512
108	1	26	7	1	1	120
109	1	6	8	1	1	680
110	2	82	4	1	2	240
111	2	6	7	1	1	1520
112	2	6	3	2	2	162
113	2	2	5	1	1	387
114	2	6	6	1	1	210
115	1	6	2	1	1	460
116	2	2	3	1	2	645
117	4	2	2	4	1	100
118	4	14	4	1	1	117

Algunas tablas y gráficos:

Resid. Habitual	**Frec. Abs.**	**Fr (%)**
Cap. Fed.	16	13,6
Bs. As.	58	49,2
Córdoba	7	5,9
Corrientes	5	4,2
Chaco	2	1,7
Chubut	4	3,4
Entre Ríos	1	,8
Formosa	1	,8
La Pampa	1	,8
Mendoza	2	1,7
Misiones	3	2,5
Santa Fe	7	5,9
Italia	1	,8
Inglaterra	1	,8
Alemania	4	3,4
España	1	,8
Australia	2	1,7
Perú	1	,8
Canadá	1	,8
Total	**118**	**100,0**

Medio de Arribo	**Frec. Abs.**	**Fr (%)**
Automotor Privado	74	62,7
Ómnibus Regular	28	23,7
Aéreo Regular	16	13,6
Total	**118**	**100,0**

Opinión sobre Arq. y Urb.	**Frec. Abs.**	**Fr (%)**
Buena	78	66,1
Regular	29	24,6
Mala	7	5,9
Sin Opinión	4	3,4
Total	**118**	**100,0**

GTOT: Diagrama de Tallo - Hoja

Frecuencia	Tallo	-	Hoja
4	0	.	0011
11	0	.	22222333333
12	0	.	444444555555
21	0	.	666666777777777777777
4	0	.	9999
13	1	.	0000000011111
7	1	.	2222333
9	1	.	444445555
5	1	.	66666
5	1	.	88999
5	2	.	00111
1	2	.	2
1	2	.	4
4	2	.	6777
1	2	.	8
3	3	.	011
0	3	.	
1	3	.	4
2	3	.	77
1	3	.	8
8 Extremos			(>=387)

Ancho del tallo: 100
Cada hoja: 1 caso

Utilizando toda la información disponible escriba un informe destinado a comunicar los resultados de la encuesta a las autoridades turísticas de la provincia, en el cual se describan las características de los turistas alojados en Puerto Iguazú. Incluya en el mismo, las Tablas y Gráficos que considere pertinentes.

UNIDAD 3: Los Valores que Caracterizan al Conjunto de Datos

Actividad Nº 1

A continuación se presentan cuatro párrafos que aluden a diferentes temas de trabajo y reproducen conclusiones basadas en medias aritméticas calculadas a partir de conjuntos de datos también diferentes.

En cada párrafo encontrará la información necesaria para contextualizar esos resultados (unidad de análisis, población, variable observada, datos transversales/longitudinales, relevamiento muestral o censal, etc).

Párrafo 1:

"Una encuesta realizada en el año 2001 a 1.297 alumnos de escuelas primarias de la ciudad de Buenos Aires, reveló que los escolares (a esa fecha la población era de 150.000 estudiantes en todas las escuelas primarias porteñas) dedican en promedio 13 horas semanales a ver televisión y (también en promedio por alumno) 6 horas semanales a la lectura de libros, diarios y revistas".

Párrafo 2:

"Según los resultados de un censo realizado en el año 2000, en las 32 prisiones del Servicio Penitenciario Federal distribuidas en todo el país, había una cantidad media de 1.888 presos alojados en cada una de ellas".

Párrafo 3:

"De acuerdo con datos oficiales, en el período de 8 años comprendido entre 1991/98, se registraron exportaciones misioneras de yerba mate por un monto anual promedio de U$S 22.852.325".

Párrafo 4:

"Un estudio realizado en Capital Federal y el Gran Buenos Aires en el mes de abril de 2001, en el cual fueron encuestados 1.200 comercios del total de 4.200 establecimientos que forman el sector "autoservicios"[3], permitió conocer que estos negocios en promedio, facturan $3.000 por día y funcionan en locales cuya superficie media es de 550 metros cuadrados".

Leyendo detenidamente los ejemplos, Ud. deberá:

1. Proponer algunas preguntas estadísticas que encuentren su respuesta en los promedios mencionados.
2. Describir detalladamente al conjunto de datos que resume cada una de las medias aritméticas empleadas en el análisis (Identificar la unidad de análisis y la variable observada en cada caso, cantidad de datos de la serie, datos longitudinales o transversales, muestrales o por enumeración completa).
3. Explicar paso a paso el procedimiento que seguiría para obtener/calcular estos promedios, si dispusiera de los datos originales utilizados en cada uno de ellos (*puede resultarle útil, primero, reconstruir simbólicamente cada una de las series o conjunto de datos*).

[3] Pequeños supermercados de barrio que comercializan productos de almacén, de limpieza, verdulería, carnicería, bazar etc.

Actividad Nº 2

Primera Parte

El trabajo que dio origen a los datos que analizaremos inmediatamente tenía el propósito general de aportar información sobre diferentes características de los ***obreros y empleados calificados*** de una empresa industrial de Misiones. (Entre otras: estudios alcanzados, antigüedad en la empresa, sexo, edad, estado civil, área de trabajo, cantidad de días y de horas trabajadas en el mes anterior, etc.).

La población bajo análisis se componía de los *n*=90 obreros y empleados (excluidos los funcionarios de nivel gerencial o superior) que conformaban la plantilla de personal permanente de la empresa, al mes de marzo de 1998.

El relevamiento alcanzó a todos los "individuos" de la población y los datos se recopilaron de los legajos personales y otras fuentes administrativas disponibles.

Una de las variables observadas fue:

Z: *"Haber mensual neto percibido por el empleado en el mes inmediatamente anterior".*

A continuación encontrará la serie de datos originales de esta variable (expresados en $), en el estado en el que fueron registrados en la matriz de datos:

Empl/ obrero	Haber ($)
1	571
2	545
3	846
4	632
5	558
6	880
7	567
8	623
9	753
10	511
11	633
12	719
13	641
14	824
15	887
16	588
17	740
18	846
19	729
20	523
21	476
22	613
23	883
24	899
25	852
26	932
27	845
28	904
29	743
30	723

Empl/ obrero	Haber ($)
31	778
32	603
33	681
34	456
35	479
36	808
37	741
38	631
39	587
40	567
41	846
42	782
43	667
44	891
45	914
46	460
47	833
48	927
49	582
50	701
51	740
52	661
53	578
54	857
55	841
56	771
57	756
58	543
59	845
60	738

Empl/ obrero	Haber ($)
61	967
62	775
63	589
64	478
65	490
66	932
67	778
68	772
69	803
70	545
71	927
72	945
73	780
74	867
75	982
76	716
77	809
78	541
79	537
80	890
81	717
82	756
83	690
84	765
85	822
86	645
87	743
88	560
89	656
90	784

Luego de explorar detenidamente el conjunto de datos (por ejemplo, empleando un diagrama de tallo-hoja), su tarea consistirá en comprobar si son correctas las tres frases siguientes, y si alguna de ellas no lo fuera, tendrá que elaborar la expresión que a su juicio es verdadera.

En todos los casos sus conclusiones deben ser acompañadas de los fundamentos teóricos y/o de cálculos en los que se basan.

1. *"Los empleados y obreros de la empresa perciben un haber neto promedio de $\bar{z} = \$684,63$ mensuales; siendo el salario más bajo de \$456 y el más alto de \$982".*

2. *"Si a los datos de los 90 empleados y obreros **se agregan** los haberes que perciben los 5 subgerentes y 2 gerentes de la empresa, el salario medio de **todos los funcionarios** que componen la planta permanente asciende a $\bar{z} = \$1.057,70$. Este promedio refleja el buen nivel de los salarios que abona la firma a sus funcionarios".*

Haberes netos de gerentes y subgerentes:

$$\{4.927, 4.523, 4.852, 5.124, 4.970, 6.701, 6.890\}$$

3. *"El empleado que figura en el orden 76 difiere en menos de \$2 (d_{76}=\$-1,89) del haber promedio general de los 90 asalariados observados, mientras que los empleados del orden 21 y 66 se diferencian de dicho promedio en \$-241,89 y \$214,11, respectivamente. La suma de los residuos de todos los empleados y obreros es nula".*

4. ¿A cuánto asciende la "suma de los haberes netos" de los 90 empleados?

Segunda Parte

En la tabla siguiente se presenta la distribución que resume los datos sobre *"camas disponibles"* en 190 hosterías y residenciales relevados en una encuesta.

Hosterías y residenciales según el número de camas disponibles

Cantidad camas	Host/Resid. (f_i)
0-19	15
20-39	32
40-59	60
60-79	47
80-99	23
100-119	10
120-139	3
TOTAL	190

¿Cuál es el número promedio de camas disponibles por establecimiento?

Actividad Nº 3

1. Las 3 frases siguientes expresan algunas de las conclusiones que se pueden obtener al describir los datos sobre el *"haber mensual neto"* percibido por los obreros calificados del ejemplo anterior.

Su actividad consistirá en comprobar la veracidad de cada una de estas afirmaciones y, en caso de encontrar que alguna de ellas es errónea, tendrá que elaborar la conclusión correcta.

Nuevamente, las respuestas deben ser fundamentadas con argumentos teóricos y/o de cálculos.

a. *"La mitad de los 90 obreros y empleados calificados de la empresa, percibe haberes netos mensuales superiores o iguales a $740. La otra mitad de los salarios se ubica por debajo de dicho valor".*

b. *"Al incorporar en el análisis a los gerentes y subgerentes de la firma, el haber neto mediana de los funcionarios se eleva a $987,60 por lo que, la mitad de todo el personal percibe haberes iguales o inferiores a esa suma".*

c. *"Tal incremento en el valor medio de los haberes se debe a los haberes extremadamente atípicos de los subgerentes y gerentes de la empresa".*

2. Determinar

El número de *"camas disponibles"*, por debajo del cual se ubican los 95 establecimientos hoteleros (hosterías y residenciales) más pequeños, analizados en la actividad anterior.

Actividad Nº 4

Continuando con los datos de los dos ejemplos anteriores, su actividad consistirá en:

1. Determinar el haber mensual típico de los 90 obreros y empleados de la firma y la cantidad más frecuente de camas disponibles en las hosterías y residenciales observados.

2. Explicar detalladamente (paso a paso) el procedimiento seguido para obtener ambos resultados.

3. Analizar críticamente estos resultados y comentar sus conclusiones.

4. Comprobar que si se incorporan al análisis los haberes de los 7 gerentes y subgerentes, el valor típico de la distribución no se modifica.

Actividad Nº 5

Ampliar el análisis de las series de datos anteriores (en el caso de los haberes netos, trabajar con el conjunto original de 90 datos), utilizando las medidas de posición que considere pertinentes para completar la descripción de los individuos observados en cada ejemplo.

Actividad Nº 6

Basándose en las tablas que presentan -para dos departamentos de la pcia. de Misiones- los datos sobre las explotaciones agropecuarias distribuidas según deciles de superficie (tamaño), realizar:

a. Compare la superficie total acumulada por el 30% de las explotaciones más pequeñas de ambos departamentos.

b. Establezca "la brecha" entre el 10% de las explotaciones más grandes y más pequeñas en ambos departamentos.

c. Construya las gráficas de Lorenz y determine los coeficientes de Gini.

d. Concluya sobre la situación de la distribución de la tierra en estos departamentos.

Distribución de las explotaciones agropecuarias según por superficie. Departamentos de San Pedro y Oberá -Misiones- 1981

Decil	SAN PEDRO (1)			OBERÁ (2)		
	Superf. Total (has.)	**Explot. Acum. (%)**	**Sup. Total Acum. (%)**	**Superf. Total (has.)**	**Explot. Acum. (%)**	**Sup. Total Acum. (%)**
1	229	10,0	0,080	1181	10,0	0,75
2	744	20,0	0,338	8325	20,0	6,06
3	744	30,0	0,596	12479	30,0	14,00
4	744	40,0	0,855	12480	40,0	21,95
5	1354	50,0	1,325	12480	50,0	29,90
6	1659	60,0	1,901	12480	60,0	37,86
7	1659	70,0	2,477	12480	70,0	45,81
8	2628	80,0	3,399	12480	80,0	53,76
9	7879	90,0	6,126	18971	90,0	65,85
10	270314	100,0	100,000	53604	100,0	100,00
Total	**287954**			**156960**		

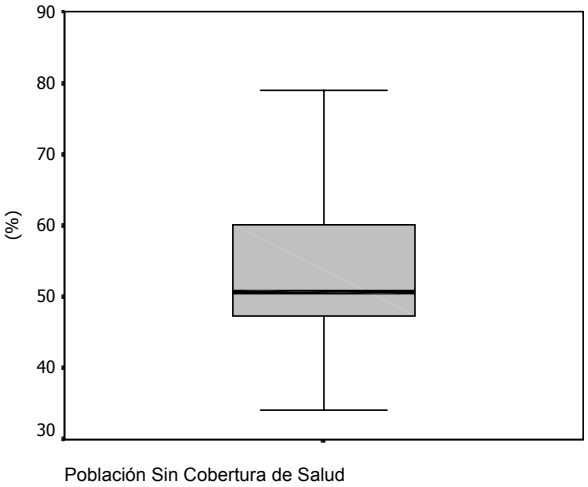
(1) El total de explotaciones censadas en San Pedro fue 443.
(2) El total de explotaciones censadas en Oberá fue 6.522.

Fuente: INDEC. Censo Nacional Agropecuario. 1981.

Actividad Nº 7

*A partir de los datos correspondientes al "**porcentaje de Población sin Cobertura de Salud**" en los 75 municipios de la pcia. de Misiones, según el Censo Nacional de Población y Vivienda – 1991, se pudieron obtener los siguientes valores y el Diagrama de Caja (Box-Plot) que permiten caracterizar esa distribución.*

- $X_{mín}$: *34,20 %*
- Q_1: *47,28 %*
- M_a: *50,79 %*
- Q_3: *60,60 %*
- ***Xmáx:*** *78,99 %*



Basándose en el diagrama y los valores característicos obtenidos, describir la situación sanitaria en los municipios de Misiones, según el "Porcentaje de Población sin Cobertura de Salud".

Actividad Nº 8
Práctico Complementario

1. En la Actividad Nº 3 de la Unidad anterior, Ud. resumió en forma numérica y gráfica los datos de una muestra de 156 turistas observados en el Parque Nacional Iguazú, referidos a la cantidad de noches (pernoctes) que planeaban permanecer en la región. También comenzó la descripción de los turistas desde esta característica en estudio. Trabajando con los mismos datos, su actividad consistirá en:

a. Determinar la cantidad promedio de noches de estadía, la estadía más frecuente y la cantidad mediana de pernoctes, y dar su opinión crítica de los resultados que obtuvo.

b. Explicar el procedimiento empleado para obtener cada una de estas medidas.

c. Elaborar un pequeño informe descriptivo de los turistas analizados.

2. El párrafo siguiente resume algunas conclusiones sobre la distribución por edades de los habitantes de Misiones, censados en 1991 (Unidad 2 - Actividad Nº 4).

- ✓ *"El primer cuarto de la población más joven de Misiones no superaba los 9 años de edad; y*
- ✓ *la cuarta parte de las personas de mayor edad, registraba 38 años o más.*
- ✓ *Un 10% de los habitantes (aproximadamente 78.900 personas) se encontraban con 55 años cumplidos o más edad a la fecha del Censo"*.

Su tarea consistirá en confirmar la exactitud de estas afirmaciones, justificando su respuesta.

3. Trabajando sobre el *"nivel de estudios"* alcanzado por las madres de los alumnos del curso de Estadística (Actividad Nº 8–Unidad 2):

- Completar la descripción determinando las siguientes medidas: M_a, M_o, D_3 y D_7.
- Redactar sus conclusiones al comparar ambas distribuciones.

4. Trabajando sobre los salarios de hombres y mujeres analizados en ejercicio 6 – Actividad Nº 10, Unidad 2:

Complete sus conclusiones acerca de la política salarial que lleva a cabo la empresa, utilizando las medidas que considere pertinente incluir en el análisis.

EVALUACIÓN PARCIAL -Unidad 3-

Como actividad de evaluación de la Unidad anterior, Ud. comenzó a analizar y describir -en términos de seis variables relevantes- a una muestra de 118 turistas que viajaron por cuenta propia a las Cataratas del Iguazú, en el mes de febrero de 1994.

Su trabajo consistirá en integrar al análisis ya iniciado, las medidas de tendencia central y de posición que considere pertinentes y apropiadas a los datos de cada una de las variables en estudio.

Para cada una de las medidas que decida utilizar en el análisis, deberá:

a. justificar su aplicación considerando cuestiones como: el propósito descriptivo que persigue, el tipo de datos con el que trabaja, las propiedades de la medida utilizada, las características más destacables del conjunto de datos, etc.,
b. explicar detalladamente el procedimiento seguido para calcular/determinar cada una de ellas.

La actividad se completará con la redacción del informe mediante el cual comunicará sus conclusiones (ahora ampliadas) a las autoridades turísticas de la provincia, agregando las tablas y gráficos que considere pertinentes.

UNIDAD 4: Análisis de la Variación y Asimetría

Actividad Nº 1

A partir de los datos sobre el "*Haber mensual neto percibido por el empleado en el mes inmediatamente anterior"* del personal de una empresa industrial de Misiones (Unidad 3, Actividad 2), observar:

- ¿Cuál es la amplitud de variación de los ingresos del personal?
- ¿Cuál es el campo de variación de los ingresos de los gerentes y subgerentes?
- ¿En qué extensión varían los ingresos del 50% central de los obreros y empleados calificados de la empresa?

Actividad Nº 2

1. Para los **haberes de empleados y obreros** de la empresa industrial de la actividad anterior, elabore un breve informe que describa esa distribución, utilizando para ello las medidas de tendencia central, posición y variación que considere pertinentes.
2. De acuerdo con lo analizado en el punto 2 de la Actividad 2-Unidad nº 3, sobre el salario promedio de todo el personal de la empresa, ***indique y justifique*** cuáles serían las medidas de tendencia central y dispersión que utilizaría para describir la variabilidad de los datos y complementar el análisis ya realizado.

Actividad Nº 3

1. El intendente de la ciudad de Leones-Cba., analizando la cantidad de metros mensuales de bacheo realizados (arreglo de pozos en el asfalto) y comparándola con la producción de la ciudad de Bs. As., observó con satisfacción que, si bien el promedio mensual era de 80 metros contra 1500 metros en Bs. As., la producción de esta tarea en su municipio mostraba una mayor regularidad (σ_{Leones} = 30 metros y σ_{BA} = 200 metros); ¿es realmente más regular la tarea en el municipio de Leones? Justifique su respuesta.
2. A partir de la encuesta permanente de hogares se pudo determinar que en 1998 el ingreso de los hogares de las ciudades de Santiago del Estero y Jujuy sorprendentemente presentan la misma desviación mediana. ¿Bajo qué condiciones se podría asegurar que los ingresos de los hogares en ambas ciudades son igualmente heterogéneos?

Actividad Nº 4

Describa brevemente el nivel de estudios de padres y madres de los estudiantes de Estadística, considerando en la descripción la heterogeneidad que presentan estos conjuntos de datos.

Estudiantes del curso de Estadística según Nivel de estudios del Padre y de la Madre-FHyCS-Año 2001

Nivel de Estudios del Padre	nº de estudiantes (*)
Ninguno	3
Prim. Incompleto	27
Prim. Completo	56
Sec. Incompleto	17
Sec. Completo	17
Terc./Univ. Incomp.	7
Terc./ Univ. Comp.	6
Total	**133**

Nivel de Estudios de la Madre	nº de estudiantes (**)
Ninguno	2
Primario Incompleto	33
Primario Completo	42
Secundario Incompleto	23
Secundario Completo	14
Terc./Univ. Incomp.	7
Terc./Univ. comp.	15
Total	**136**

(*) Hay 6 estudiantes que no declaran el nivel de estudios de su padre.
(**) Hay 3 estudiantes que no declaran el nivel de estudios de su madre.
Fuente: elaboración propia basándose en datos del "Estudio de los Alumnos de Estadística".

Actividad Nº 5

Evalúe el nivel de asimetría de la distribución que sigue (ya trabajada en la Unidad 3), utilizando los coeficientes de asimetría que conoce.

Describa esta característica de los datos y evalúe a partir de esta información si el promedio obtenido en la unidad anterior es una medida representativa del conjunto.

Hosterías y residenciales según el número de camas disponibles

Cantidad camas	Host/Resid (f_i)
0-19	15
20-39	32
40-59	60
60-79	47
80-99	23
100-119	10
120-139	3
TOTAL	**190**

Actividad Nº 6
Práctico Complementario

1. Se cuenta con información sobre el **gasto *per cápita* diario** efectuado el día inmediato anterior a la entrevista y la **cantidad de componentes** de los 118 grupos turísticos entrevistados entre

quienes visitaron Parque Nacional Iguazú en febrero de 1994. Calculadas algunas medidas de resumen sobre esta información, se obtuvo:

MEDIDA	**Gasto *per cápita***	**Componentes**
Mínimo	$ 0,00	1 pers.
Máximo	$ 230,00	8 pers.
$\overline{x}$	$ 51,45	3,32 pers.
Ma	$ 37,00	3 pers.
Mo	$ 39,50	2 pers.
Q1	$ 21,71	2 pers.
Q3	$ 62,68	4 pers.
σ	$ 47,62	1,60 pers.

Sobre la base de la información anterior, evaluar la veracidad de las siguientes afirmaciones, haciendo un comentario en cada caso.

1. La distribución del *gasto per cápita* es más dispersa que la distribución de componentes del grupo.
2. La distribución del *número de componentes* es más simétrica.
3. En la distribución del *gasto per cápita* la media es menos representativa del conjunto.
4. El 50% central de la distribución del *gasto per cápita* es más asimétrica que esa misma proporción de datos centrales en el número de componentes.

2. En una encuesta realizada por FIEL y la Fac. de Hum. y Cs. Soc. en el año 1991 se consultó a los habitantes de la ciudad de Posadas sobre su opinión en relación con medidas que se debían tomar y calidad del servicio de las empresas del Estado Nacional y Provincial. En relación con las empresas provinciales, se pudo observar que las opiniones sobre las medidas a tomar se distribuían en los distintos niveles socioeconómicos, de la siguiente manera.

Opinión sobre medidas a tomar con empresas provinciales en diferentes niveles socioeconómicos – Pdas. 1991

Opinión sobre medidas a tomar	**NES BAJO**	**NES MEDIO**	**NES ALTO**
Vender totalmente	11	21	15
Vender parcialmente	15	40	30
Mejorarlas	100	190	43
No vender	7	9	4
Otra medida	0	6	5
Sin opinión	12	13	3
Total	**145**	**279**	**100**

Fuente: Encuesta FIEL-FHyCS – Junio 1991

¿En qué estrato socioeconómico se observa un mayor consenso en relación con las medidas que deberían adoptarse para las empresas del estado provincial?

EVALUACIÓN PARCIAL -Unidad 4-

En la Unidad 2, hemos trabajado las distribuciones del plantel de empleados de una empresa de servicios, discriminados por sexo según sus salarios mensuales en mayo de 1996. El propósito era analizar si la empresa tiene una política salarial diferencial por sexo.

Distribución de los salarios mensuales de empleados de una empresa de servicios, discriminados por sexo. Mayo de 1996

HOMBRES

Ingreso ($)	Empleados
300 – 450	2
450 – 600	39
600 – 750	126
750 – 900	26
900 – 1050	8
1050 – 1200	20
TOTAL	**221**

MUJERES

Ingreso ($)	Empleados
300 – 450	55
450 – 600	116
600 – 750	32
750 – 900	11
900 – 1050	1
1050 – 1200	1
TOTAL	**216**

Basándose en la información resumida en las Tablas anteriores:

1. Tomando en cuenta las características de asimetría y variabilidad de esas distribuciones, revise críticamente la pertinencia de las medidas resumen calculadas para estas distribuciones en la unidad anterior (Actividad 8-punto 4). Escriba sus conclusiones justificándolas.
2. Sobre la base de todos los elementos de análisis con los que cuenta en este momento, redacte un informe sobre la política salarial de la empresa en relación al sexo, incluyendo en el mismo los gráficos y medidas que considere apropiados.

UNIDAD 5: El Estudio de la Relación entre Variables

Actividad Nº 1

Cada una de las preguntas siguientes plantea la necesidad de un análisis bivariado. Para cada una de ellas, identifique:

a) las variables que intervienen y su tipo;

b) la naturaleza de la relación que puede suponerse entre esas variables.

- ¿Difiere el nivel de ingresos según sea el lugar de residencia de los padres?
- Los mujeres, ¿dedican más tiempo a mirar televisión?
- ¿Cuando decrece la edad, decrece también el tiempo que se dedica al estudio?
- ¿Según sea el *nivel socioeconómico* varía la *opinión* sobre los servicios públicos?
- ¿Los *hombres leen más frecuentemente el periódico*?
- ¿Los salarios que perciben las mujeres difieren del que perciben los hombres?
- ¿El rendimiento escolar de los estudiantes en el examen de Lengua varía según se trate de escuelas rurales o urbanas?
- ¿El número de hijos por familia es distinto según sea el nivel socioeconómico?
- ¿Al aumentar el número de automóviles por habitantes, aumenta el número de accidentes de tránsito?
- ¿Al disminuir el gasto en publicidad, disminuye la demanda de un producto?,
- Cuanto mayor es el número de médicos por habitantes en un país, ¿varía la tasa de mortalidad infantil?
- ¿Al aumentar la antigüedad de un automóvil, aumenta el costo de mantenimiento?

Actividad Nº 2

Al finalizar un curso de especialización para abogados, se pide a los participantes su opinión sobre la calidad del mismo. El propósito es conocer si hay alguna relación entre la opinión y la especialidad del participante. Se presenta la matriz de datos y la especialidad de cada participante.

a) Construya una tabla que presente la clasificación bivariada de los participantes del curso según su especialidad y opinión.

b) ¿Cuántos abogados son especialistas en lo laboral?

c) ¿Cuántos participantes calificaron el curso como bueno?

d) ¿Cuántos abogados con especialidad en lo laboral han calificado el curso como bueno?

e) ¿Cuántos participantes con especialidad en lo civil y comercial lo calificaron como Malo?

f) ¿Qué porcentaje de abogados asistentes se especializan en Laboral?

g) ¿Qué porcentaje de los asistentes califican el curso como Regular y son especialistas en lo Civil y Comercial?

h) ¿Qué porcentaje de los asistentes calificó el curso como Bueno?

Matriz sobre el Curso de especialización

Individuos	Opinión	Especialidad
1	Bueno	Laboral
2	Malo	Civil y Comercial
3	Bueno	Civil y Comercial
4	Bueno	Laboral
5	Malo	Civil y Comercial
6	Bueno	Laboral
7	Bueno	Civil y Comercial
8	Bueno	Laboral
9	Regular	Civil y Comercial
10	Regular	Civil y Comercial
11	Bueno	Laboral
12	Regular	Laboral
13	Bueno	Laboral
14	Bueno	Laboral
15	Regular	Civil y Comercial
16	Regular	Laboral
17	Bueno	Laboral
18	Malo	Civil y Comercial
19	Bueno	Civil y Comercial
20	Bueno	Laboral
21	Bueno	Laboral
22	Malo	Civil y Comercial
23	Regular	Civil y Comercial
24	Malo	Laboral
25	Malo	Civil y Comercial
26	Bueno	Laboral
27	Bueno	Civil y Comercial

Actividad Nº 3

En la tabla siguiente se presentan los datos de la población urbana y rural de la Argentina en 1914, por grandes regiones geográficas.

Población urbana y rural de la Argentina por regiones. Año 1914 (en miles)

REGIONES	Urbana	Rural	Total
Pampeana	3604	2200	**5804**
Cuyana	145	368	**513**
Nordeste	135	331	**466**
Noroeste	260	735	**995**
Patagónica	12	94	**106**
Total	**4156**	**3728**	**7884**

Fuente: Recchini de Lattes, Zulma: *Aspectos demográficos de la urbanización en la Argentina, 1869-1960*. Centro de Inv. Soc.-Inst. Torcuato Di Tella. CELADE.

En 1914, **¿variaba la composición (urbana y rural) de la población entre las diferentes regiones?**

Calcule los porcentajes en fila y compare las regiones.

Actividad Nº 4

Basándose en la Tabla anterior, calcule los porcentajes en columna y compare la distribución por Regiones de cada tipo de asentamiento (urbano y rural). Describa esa comparación.

Actividad Nº 5

A. La siguiente tabla muestra la clasificación de todos los empleados de una empresa de transportes según edad y categoría de empleo. Los datos fueron relevados por la empresa en 1998.

Distribución de los empleados según edad y categoría de empleo - 1998

	Categoría de Empleo			
Grupos de Edad	**Pers. Ejecutivo**	**Administrativos**	**Obreros**	**Total**
Menos de 40 años	28	5	160	**193**
40 años o más	14	40	67	**121**
Total	**42**	**45**	**227**	**314**

- ¿Varía la edad según sea la categoría de empleo? ¿Varía la categoría de empleo según las edades?
- Analice si existe o no relación entre estas variables y, de observar una relación, describa su forma.

B. La siguiente Tabla muestra la distribución de la población según zonas y condición de pobreza, Argentina 2001.

Zonas	Condición de Pobreza		Total
	Pobre	**No Pobre**	
Cap. Fed. y Reg. Pampeana	9.058.454	14.906.587	**23.965.041**
Resto País	5.581.978	6.480.022	**12.062.000**
Total	**14.640.432**	**21.386.609**	**36.027.041**

Fuente: SIEMPRO elaboración basándose en EPH-INDEC - Octubre 2001

- La incidencia de la pobreza, ¿es diferente según las zonas del país?
- Calcule la medida que exprese la fuerza de la relación.
- Escriba su conclusión sobre la relación entre las variables Zona y Condición de Pobreza de la población.

Actividad Nº 6

A. Con relación a la actividad anterior-punto A:

- Construya un gráfico que muestre la composición por edades de cada categoría de empleo.
- Construya un gráfico que muestre la distribución de las categorías de empleo en cada grupo de edad.

B. Con relación a la actividad anterior-punto B:

- Construya un gráfico que muestre la distribución por zonas según condición de pobreza.
- Construya un gráfico que muestre la incidencia de la pobreza en cada una de las zonas.

Actividad Nº 7

A. Según la Encuesta de Desarrollo Social realizada por la Secretaría de Desarrollo de la Nación, los *Ingresos medios* y el *número de personas promedio del hogar* en 1997, registraban por regiones los siguientes valores.

	Total País	Cuyo	Gran Bs. As.	NEA	NOA	Pampeana	Patagónica
Ingreso medio del hogar	1136,7	992,2	1377,1	815,6	915,6	949,9	1190,9
Promedio de pers./hogar	3,7	4,1	3,5	4,3	4,6	3,5	3,9

Sobre la base de los datos aportados escriba sus conclusiones en relación con el ingreso y las personas por hogar en las distintas regiones del país.

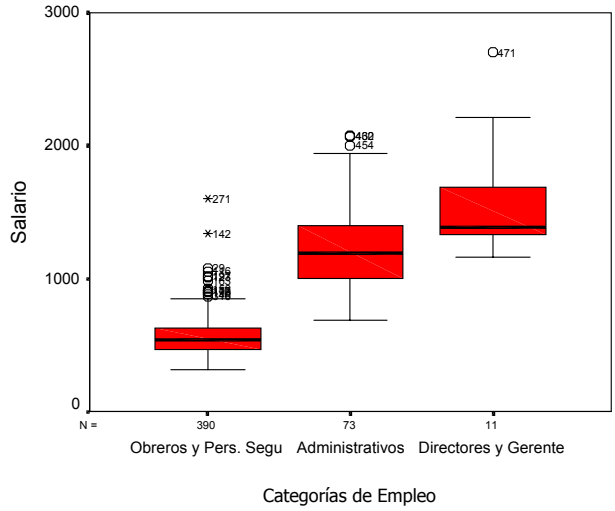
B. El gerente de personal de una empresa del sector alimentación debe exponer, ante el nuevo directorio, la política salarial que la empresa ha llevado hasta el momento en materia de remuneraciones. Para fundamentar su exposición cuenta con los datos que se presentan a continuación.

Caracterización de la distribución del salario mensual según categorías de empleo–2002

Categoría de Empleo	n	Mín.(*)	Máx. (*)	Media(*)	Mediana (*)	Desv. Estándar (*)	CV (%)
Obreros y Pers. Seguridad	390	315	1600	561,1	540,0	147,3	26,3
Administrativos	73	688	2075	1232,2	1187,5	324,4	26,3
Directores y Gerentes	11	1163	2700	1593,9	1385,0	471,7	29,6

(*) En ($)

Diagrama de Caja (*Box-Plot*)
Distribución del Salario mensual según categoría de empleo.



Suma de Cuadrados para el cálculo de η^2

Entre grupos	36.929.198
Intra grupos	18.237.400
Total	55.166.598

A partir de la información que surge del análisis de los datos, elabore un informe que describa la situación salarial de los empleados y fundamente la exposición del gerente de personal.

Actividad Nº 8

Según datos del Censo Nac. de Población y Vivienda 1991, en las provincias del país se registraban las siguientes tasas de analfabetismo y mortalidad infantil.

PROVINCIA	ANALFAB.	MORT. INF.
Cdad. Bs. As.	0,69	12,20
Buenos Aires	2,35	18,80
Catamarca	4,52	25,60
Córdoba	3,18	16,30
Corrientes	9,34	22,80
Chaco	11,31	28,20
Chubut	4,47	19,10
Entre Ríos	4,92	19,60
Formosa	8,18	29,80
Jujuy	6,68	24,00
La Pampa	4,03	12,10
La Rioja	4,01	19,30

PROVINCIA	ANALFAB	MORT. INF.
Mendoza	4,56	16,70
Misiones	8,30	21,30
Neuquén	5,33	13,80
Río Negro	5,60	15,50
Salta	6,72	20,50
San Juan	4,25	21,50
San Luis	4,31	19,90
Santa Cruz	2,19	16,00
Santa Fe	3,66	16,30
Stgo del Estero	8,64	16,60
Tierra del Fuego	1,10	11,20
Tucumán	4,96	19,60

Construya el diagrama de dispersión (*aquellos que conozcan el Excel pueden encontrar en el menú de Gráficos una opción para este tipo de diagrama*).

Actividad Nº 9

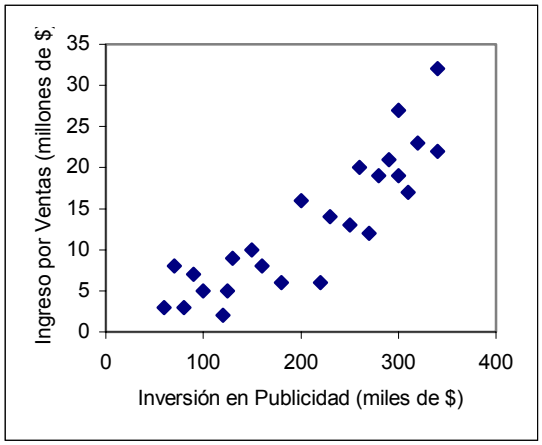
Analizando el diagrama de dispersión anterior describa el tipo de relación observada.

Actividad Nº 10

Debido a la sobreoferta de yerba mate, la Cámara de Molineros está interesada en expandir el consumo de este producto en países en que es poco conocido. A tal fin realiza un estudio para el año 1998 sobre empresas que exportan yerba mate, observando las variables ***inversión en publicidad*** (en miles de $) e ***ingresos por ventas*** (expresadas en millones de pesos), con el objetivo de analizar la relación entre las mismas. Los datos sobre las 25 empresas observadas son:

Empresa	Inv. en Pub. (miles de $)	Ing. por Ventas (millones de $)
1	250	13
2	310	17
3	320	23
4	125	5
5	80	3
6	150	10
7	270	12
8	90	7
9	260	20
10	300	27
11	130	9
12	200	16
13	60	3
14	100	5
15	230	14
16	340	22
17	300	19
18	290	21
19	70	8
20	220	6
21	280	19
22	340	32
23	160	8
24	180	6
25	120	2

Diagrama de Dispersión



1. Explicar si en este caso es aceptable la aplicación del análisis de regresión.
2. Fundamentar la elección de la variable dependiente, y explicar la naturaleza de la relación entre las variables.
3. Utilizando los valores de a =-3,48 y b = 0,08 calculados a partir de los datos de la tabla; indicar la ecuación de la recta de regresión estimada y representarla gráficamente.
4. Calcular el *ingreso promedio por ventas* suponiendo una inversión en publicidad de miles $190.

Actividad Nº 11

En el análisis de los gastos en publicidad e ingresos por ventas, los cálculos de los coeficientes de correlación y de determinación, arrojaron los siguientes resultados:

r= 0,88 **R^2= 0,774**

Interpretar los valores de **r** y **R^2**

Actividad Nº 12
Práctico Complementario

1. Para cada una de las preguntas de la Actividad nº 1, indique cuál es la herramienta de análisis bivariado que utilizaría (Análisis de Tablas de Contingencia, Diferencia de medias, Análisis de correlación).

2. Sobre una muestra de 180 agentes de la administración pública provincial, se analizó la distribución por sexo y niveles de ingreso, obteniéndose los datos que se presentan a continuación.

Distribución de agentes públicos según sexo y nivel de ingreso (*)

Sexo	Nivel de Ingreso			Total
	Bajo	Medio	Alto	
Varón	26	60	21	**107**
Mujer	25	36	12	**73**
Total	**51**	**96**	**33**	**180**

(*) Los salarios fueron categorizados según el siguiente criterio:
Bajo: menor a una canasta básica (canasta que cubre necesidades mínimas para la subsistencia).
Medio: hasta 2 veces la canasta básica.
Alto: más de 2 veces la canasta básica.

1. ¿El nivel de ingresos es diferente según se trate de hombres o mujeres? Describa
2. ¿Quiénes son (en cuanto al sexo) los que tienen diferentes niveles de ingreso? Describa.

3.

Indique si las siguientes afirmaciones son verdaderas, justificando su respuesta

a. Si el coeficiente de correlación *r de Pearson* entre dos variables es *r=0* se puede asegurar que no existe relación entre ellas.

b. El coeficiente *r de Pearson* permite determinar la existencia de relación entre cualquier par de variables.

c. Cuando el coeficiente **b** de la ecuación de regresión es positivo, el coeficiente ***r*** de correlación también es positivo.

d. Si el coeficiente de correlación es negativo el coeficiente de determinación también lo es.

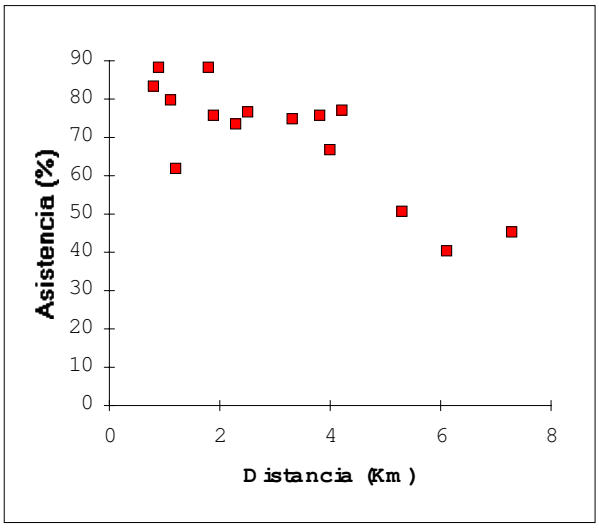
4. Según estudios realizados por la Secretaría de Trabajo de la Nación, sobre oferta de puestos de trabajo y salarios, dio como resultado que en el sector de la construcción un aumento en la demanda de trabajadores de 1000 puestos de trabajo, eleva el precio del jornal en $4,50.

a- ¿Qué tipo de análisis se realizó para llegar a esta conclusión?
b- ¿Cuál es el signo del coeficiente de correlación entre estas variables? Justifique.

5. En un estudio sobre presentismo escolar realizado en una escuela rural del interior de la provincia, se recogieron los siguientes datos de una muestra de 15 alumnos*:*

Alumno	Distancia (1) (Km)	Asistencia (2) (%)
1	0,8	83,4
2	5,3	50,8
3	2,3	73,4
4	3,8	75,6
5	4,0	67,0
6	4,2	76,9
7	3,3	75,1
8	6,1	40,3
9	0,9	88,2
10	1,2	61,8
11	2,5	76,5
12	1,1	79,6
13	7,3	45,3
14	1,8	88,3
15	1,9	75,7

Diagrama de dispersión



(1) Distancia (Km) que recorre diariamente desde su hogar a la escuela.
(2) Porcentaje anual de asistencia a clases.

a- Analizar el diagrama de dispersión y justificar la aplicación del modelo de regresión lineal.

b- Justificar el uso de la variable distancia como variable independiente del modelo.

c- Utilizando los siguientes resultados, calculados a partir de los datos de la tabla anterior:

a= 89,2 **b**= - 6,0 **r**= - 0,81 $\mathbf{R^2}$= 66,3%

Estimar el porcentaje promedio de asistencia probable, de un alumno que debe recorrer diariamente 10 km.

d- Interpretar los coeficientes ***r*** de correlación y $\mathbf{R^2}$ de determinación.

EVALUACIÓN PARCIAL -Unidad 5-

1. En la matriz que se adjunta, se presentan los datos de una muestra de 28 empleados calificados, con el fin de analizar la posible asociación entre los ***años de educación formal aprobados*** y el ***salario mensual que perciben***. A partir de esos datos:

a) Construir el diagrama de dispersión y analizar si en este caso es aceptable la aplicación del análisis de regresión lineal.

b) Utilizando los resultados que fueron calculados a partir de los datos de la tabla:

a= 119,4 **b= 35,2** **r= 0,88** $\mathbf{R^2}$**= 0,77**

b.1. Indicar la ecuación de la recta de regresión estimada y representarla gráficamente.

b.2. Calcular el salario inicial promedio de un trabajador con 9 años de educación formal aprobados.

b.3. Interpretar los coeficientes **r** de correlación y $\mathbf{R^2}$ de determinación.

Matriz de datos

Empleado	Años de Educación Formal	Sueldo Inicial ($)
1	8	408
2	8	408
3	8	420
4	8	440
5	9	488
6	10	517
7	10	480
8	11	520
9	11	450
10	11	510
11	12	508
12	12	480
13	12	540
14	12	580
15	13	450
16	13	580
17	13	564
18	13	570
19	14	510
20	14	570
21	14	600
22	15	720
23	15	680
24	16	630
25	17	810
26	17	770
27	18	680
28	18	850

2. En un estudio dirigido a los ingresantes de la UNaM, se preguntó a los estudiantes sobre ***el lugar donde recibió charlas de orientación vocacional*** y el ***tipo de colegio*** del cual egresaron. A continuación se presentan los resultados de clasificar a los alumnos en forma bivariada según sus respuestas en ambas preguntas. Los datos corresponden al año 1995.

Distribución de los Estudiantes según Lugar donde recibió orientación vocacional y Tipo de colegio- UNaM - 1995

	Lugar donde recibió orientación vocacional				
Tipo de Colegio	**No recibió**	**Colegio**	**Familia**	**Otro lugar**	**Total**
Pública	907	2082	219	465	**3673**
Privada	158	927	30	74	**1189**
Total	**1065**	**3009**	**249**	**539**	**4862**

a. Conteste las siguientes preguntas realizando los cálculos que considere convenientes si es necesario.

i. ¿Cuántos alumnos recibieron *orientación vocacional*?

ii. ¿Cuántos alumnos proceden *de un colegio público*? ¿Qué porcentaje representan *en el total de ingresantes*?

iii. Entre los ingresantes que *recibieron orientación en "Otra forma"*, ¿qué porcentaje representan *aquellos provenientes de escuelas públicas*?

iv. ¿Qué porcentaje representan *en el total de ingresantes* los que *provienen de un colegio privado y recibieron orientación en el colegio*?

b. Construya el gráfico que considere conveniente para comparar la *orientación recibida* según sea el *tipo de colegio* del cual egresaron los estudiantes**.**

UNIDAD 6: Los Números Índices

Actividad Nº 1

Para esta actividad Ud. tendrá que trabajar con los datos de la tabla siguiente. Por lo tanto le sugerimos realizar una lectura previa cuidadosa de todos sus elementos para lograr una comprensión correcta del significado y del comportamiento de los datos.

Cantidades y Precios de Exportación de Yerba Mate, Té y Tung. Misiones, 1990/2000.

Año	Exp. de Yerba Mate		Exportaciones de Té		Exportaciones de Tung	
	Tn.	U$S/kg	Tn.	U$S/kg	Tn. de Aceite	U$S/kg
1990	4.266	0,835	42.584	0,77	8.550,00	0,743
1991	9.022	1,073	34.658	0,77	8.522,00	1,019
1992	13.491	1,025	34.809	0,78	5.883,00	1,719
1993	15.689	1,065	41.872	0,88	2.497,00	1,904
1994	15.667	0,943	41.188	0,87	2.415,00	1,013
1995	37.488	0,802	40.466	0,77	3.519,00	0,918
1996	39.499	0,714	39.069	0,75	2.427,00	1,073
1997	33.277	0,677	41.465	0,77	3.978,90	1,681
1998	34.916	0,663	57.738	0,93	2.204,00	1,340
1999	30.269	0,640	51.090	0,75	1.424,00	0,944
2000	36.528	0,561	49.240	0,76	1.840,99	0,800

Fuente: Dirección General de Economía Agraria. Ministerio de Asuntos Agrarios. Provincia de Misiones. 2002.

Primera Parte

a- Tomando como período base al año 1994 (1994=100), calcular el índice relativo simple (Rs) de la cantidad y el precio de exportación de la yerba mate, para todos los períodos de la serie.

b- Interpretando los resultados que obtuvo en el punto anterior deberá decidir si cada una de las siguientes afirmaciones es ***verdadera*** o ***falsa***. En cada caso tendrá que justificar teóricamente su respuesta y, si la calificó como falsa, tendrá que redactar la interpretación correcta.

1. "*El año 1990 registra la menor cantidad exportada de la serie en estudio siendo el índice relativo simple* $Rs_{90/94}$ = 27,2%, *lo que significa una merma para ese año del* 72,8% *con respecto al volumen exportado en 1994*".
2. "*A su vez, 1999 es el año en el cual Misiones exportó la mayor cantidad de yerba mate en todo el período bajo análisis, siendo* $Rs_{99/94}$ = 252,1%, *lo que indica un aumento del* 252,1% *con respecto a la cantidad exportada en 1994*".
3. "*A partir de 1995 el precio de exportación de la yerba mate decrece sistemáticamente con respecto al precio de 1994, ya que el índice relativo simple para cada uno de esos períodos resulta*:" $Rs_{95/94}$ = 85,0% $Rs_{96/94}$ = 75,7% $Rs_{97/94}$ = 71,8% $Rs_{98/94}$ = 70,3% $Rs_{99/94}$ = 67,9% $Rs_{00/94}$ = 59,5%

c- Realizar el cálculo de los índices Rs para la cantidad y el precio de exportación de la yerba mate, del té y del tung, tomando al año 1990 como período base (1990=100) para todos ellos. Interpretar los resultados obtenidos y redactar un breve informe con sus conclusiones más relevantes.

Segunda Parte

a- Calcular el índice relativo simple en eslabón (Re) de la cantidad y el precio de exportación de la yerba mate, para todos los períodos de la serie en estudio.

b- Confirmar que:

1. "*Entre 1992 y 1993 se registra la mayor caída en la cantidad exportada de tung ya que* $Re_{93/92}$ = 42,4%, *lo que significa una disminución del* 57,6% *de un año a otro*".
2. "*El mayor incremento interanual del precio de exportación del tung se registra en 1992, con un crecimiento relativo del* 68,7% *con respecto al precio anterior,* siendo:

$$Re_{92/91} = 42,4\%$$

c- Utilizando los índices en eslabón calculados al comienzo, calcular el índice relativo simple en cadena (Rc) para los años 1999 y 2000, tomando como base a 1996. Interpretar los resultados que obtenga.

d- Realizar el cálculo de los índices relativos en eslabón (Re) para los datos de exportaciones de té (cantidad y precio) y elaborar un breve informe con sus conclusiones.

Actividad Nº 2

Supongamos por un momento que nos hemos propuesto estudiar la evolución de los precios en los primeros seis meses del año 2002 (enero-junio), de cinco artículos de consumo inevitable y permanente en nuestro hogar (los simbolizaremos con A, B, C, D y E)[4]. A los fines del trabajo hemos recurrido a nuestros registros contables hogareños, de los cuales pudimos extraer los siguientes datos de las cantidades mensuales adquiridas y de los precios mensuales promedio pagados, para cada uno de los bienes y servicios que nos ocupan y en cada uno de los meses del período en cuestión.

Cantidades Mensuales Compradas y Precios Mensuales Promedio Pagados por Cinco Artículos de Consumo Familiar . Período Enero/junio de 2002.

	A		B		C		D		E	
Mes	**Precio** ($/Kg.)	**Cant.** (Kg.)	**Precio** ($/unid.)	**Cant.** (unidad)	**Precio** ($/Kg.)	**Cant.** (Kg.)	**Precio** ($/litro.)	**Cant.** (litros)	**Precio** ($/unid.)	**Cant.** (unidad)
Enero	4,80	27	0,50	118	0,85	13	1,10	33	63,00	2
Febrero	5,70	23	0,55	118	1,50	12	1,80	33	79,20	2
Marzo	7,20	20	0,72	156	1,80	14	2,20	33	91,60	2
Abril	7,80	20	0,72	155	2,80	13	2,40	32	108,45	2
Mayo	8,10	19	0.81	156	2,20	14	2,20	34	108,45	2
Junio	6,40	23	0,81	154	2,00	14	1,85	35	97,10	2

Fuente: Datos elaborados basándose en registros propios.

Segunda Parte

a- ¿Es correcta la siguiente expresión para calcular el índice de precios de agregado no ponderado (IP) para el mes de abril con base en el mes de enero?:

$$IP_{Ab/En} = \frac{\sum_{1}^{5} p_{i4}}{\sum_{1}^{5} p_{i0}} \cdot 100 = \frac{7,80 + 0,72 + 2,8 + 2,40 + 108,45}{4,80 + 0,50 + 0,85 + 1,10 + 63,00} \cdot 100 = \frac{122,17}{70,25} \cdot 100 = 173,9\%$$

En consecuencia, *¿es correcto señalar que comprar en abril una unidad (Kg., Lt.,*

[4] Imagine bienes y servicios de consumo indispensable en los hogares, como ser: leche en envase de un litro, boleto del transporte colectivo, Kwh de luz, kilogramos de determinado corte de carne, cantidad de unidades de cierto elemento que los niños utilizan en la escuela, etc, etc.

unidad, etc.) de cada uno de los 5 artículos de la canasta costaba $122,17, *mientras que el valor de la misma compra en enero era de* $70,25? Y que, por lo tanto, ¿*los precios de abril registraron un aumento conjunto del 73,*9% *con respecto a los precios de enero*?

b- Completar el cálculo del índice de precios de agregado no ponderado para todos los períodos de la serie y confirmar los siguientes resultados:

$$IP_{Feb/En} = 126,3\% \quad IP_{Mar/En} = 147,4\% \quad IP_{May/En} = 173,3\% \quad IP_{Jun/En} = 153,9\%$$

c- Dar su interpretación y conclusiones acerca de las variaciones conjuntas de los precios en los seis meses que nos ocupan.

Segunda Parte

a- ¿Es correcta la siguiente expresión para calcular el índice de precios del promedio de relativos no ponderado (IP), para el mes de marzo con base en el mes de enero (enero=100)?.

$$IP_{Mar/En} = \frac{\sum_{1}^{5} \frac{p_{i4}}{p_{i0}}}{5} \cdot 100 = \frac{\frac{7,20}{4,80} + \frac{0,72}{0,50} + \frac{1,80}{0,85} + \frac{2,20}{1,10} + \frac{91,60}{63,00}}{5} \cdot 100 =$$

$$= \frac{1,5 + 1,44 + 2,12 + 2,00 + 1,45}{5} \cdot 100 = \frac{8,51}{5} \cdot 100 = 170,2\%$$

¿Cómo interpreta este valor del índice?

b- Completar el cálculo de los índices de precios y de cantidad por el método del promedio de relativos no ponderado, para todos los períodos de la serie que estamos analizando y dar su interpretación y conclusiones de los resultados que obtiene.

Actividad Nº 3

Imaginemos ahora a una gran empresa mayorista de viajes y turismo (EVYT) que comercializa diferentes productos turísticos ("*paquetes*") de diversos destinos del País y del exterior. El problema consiste en analizar la evolución de las ventas (cantidad comercializada y precios pagados por los clientes) de los cuatro productos de mayor demanda en la temporada alta de verano (enero y febrero), en el quinquenio 1999-2003. Para realizar este cometido contamos con los datos de la Tabla siguiente:

Evolución de las Ventas *(cantidades comercializadas y precios promedio pagados)* **de Cuatro Productos Turísticos Líderes de la Temporada Alta de Verano. Período 1999-2003.**

	A		B		C		D	
Año	**Precio** ($/unid.)	**Cant.** (unidad)	**Precio** ($/unid.)	**Cant.** (unidad)	**Precio** ($/unid.)	**Cant.** (unidad)	**Precio** ($/unid.)	**Cant.** (unidad)
1999	450	1.610	681	521	166,00	1.168	3.602	350
2000	433	1.177	748	1.011	188,30	1.073	3.579	386
2001	460	1.222	725	1.230	167,70	1.158	2.958	460
2002	583	854	1.328	583	187,80	725	6.140	233
2003	505	1.056	1.362	474	291,10	1.443	11.771	271

Fuente: Datos elaborados de registros contables de la AYVT.

Primera Parte

a- Tomando al año 2001 como base de comparación (2001=100), *¿son correctas las siguientes expresiones de cálculo para determinar el índice de precios de agregado de Laspeyres, para los años 1999 y 2003*?

$$IP^{L}_{99/01} = \frac{\sum_{1}^{4} p_{i1} q_{i0}}{\sum_{1}^{4} p_{i0} q_{i0}} \cdot 100 = \frac{450 \cdot 1.222 + 681 \cdot 1.230 + 166 \cdot 1.158 + 3.602 \cdot 460}{460 \cdot 1.222 + 725 \cdot 1.230 + 167,70 \cdot 1.158 + 2.958 \cdot 460} \cdot 100 =$$

$$= \frac{3.236.678}{3.008.746,6} \cdot 100 = 1,0757 \cdot 100 = 107,6\%$$

$$IP^{L}_{03/01} = \frac{\sum_{1}^{4} p_{i5} \cdot q_{i0}}{\sum_{1}^{4} p_{i0} \cdot q_{i0}} \cdot 100 = \frac{505 \cdot 1.222 + 1.362 \cdot 1.230 + 291,10 \cdot 1.158 + 11.771 \cdot 460}{460 \cdot 1.222 + 725 \cdot 1.230 + 167,70 \cdot 1.158 + 2.958 \cdot 460} \cdot 100 =$$

$$= \frac{8.044.128,80}{3.008.746,60} \cdot 100 = 2,673 \cdot 100 = 267,3\%$$

b- A la luz de los resultados anteriores, *¿es correcto afirmar que*:

1. *los precios de los cuatro artículos en el año 1999 fueron, en conjunto o en promedio, un 7,6% superiores a los del año 2001 ya que* $IP^{L}_{99/01}$=107,6% ?,
2. *y que, por su parte, la variación conjunta de los precios del 2003, comparada con la misma base (2001=100), es del orden del 267,3% ya que* $IP^{L}_{03/01}$=267,3% ?

c- Le sugerimos completar el cálculo de los índices de precios de Laspeyres para los años de la serie que aún no se han hecho; como así también calcular los relativos simples (Rs) para cada uno de los cuatros productos turísticos que nos ocupan. Tener el cuidado de realizar todos estos cálculos con la misma base 2001=100.
Nota: le recomendamos ordenar los resultados de sus cálculos anteriores en una Tabla como la siguiente:

	Rs (2001=100)				**IP^{L}** (2001=100)
Año	**A**	**B**	**C**	**D**	
99	97,8	93,9	98,9	121,8	107,6
00					
01	100,0	100,0	100,0	100,0	100,0
02					
03	109,8	187,9	173,6	397,9	167,3

d- Analizar minuciosamente los resultados presentados en la Tabla, intentando extraer conclusiones sobre cuestiones como las siguientes:

- ¿cómo fue la evolución o comportamiento general de los precios en su conjunto a lo largo del quinquenio que analizamos?,
- ¿cómo ha sido el comportamiento individual de los precios de cada artículo por separado, tratando de identificar los precios que variaron de manera muy diferente (aumentando o disminuyendo) a la variación conjunta que expresa el IP^{L}?

Resumir sus conclusiones en un breve informe.

Segunda Parte

a- Recordar que el índice de cantidad de Laspeyres IQ^L (**de agregado**), para el año 2003 con base 2001=100, se obtiene de:

$$IQ^L_{03/01} = \frac{\sum_1^4 q_{i5}p_{i0}}{\sum_1^4 q_{i0}p_{i0}} \cdot 100 = \frac{1.056 \cdot 460 + 474 \cdot 725 + 1.443 \cdot 167,7 + 271 \cdot 2.958}{1.222 \cdot 460 + 1.230 \cdot 725 + 1.158 \cdot 167,7 + 460 \cdot 2.958} \cdot 100 =$$

$$= \frac{1.872.939,1}{3.008.746,6} \cdot 100 = 0,622 \cdot 100 = 62,2\%$$

Que, en consecuencia, $IQ^L_{03/01} = 62,2\%$ expresa que las cantidades de los cuatro productos líderes de nuestra AVYT, comercializadas en el verano del 2003, fueron (en conjunto o promedio) un 37,8% menores que en el 2001.

b- Completar el cálculo de IQ^L de agregado para los restantes períodos de la serie y analizar la evolución conjunta del volumen de ventas en todo el quinquenio.

Le sugerimos que en su análisis contraste las variaciones del IQ^L con las variaciones del IP^L calculado en el problema anterior.

Tercera Parte

a- Corroborar que el índice de precios de agregado de Paasche -IP^P-, para el año 2000 con base 2001=100, se obtiene haciendo:

$$IP^P_{00/01} = \frac{\sum_1^4 p_{i2}q_{i2}}{\sum_1^4 p_{i0}q_{i2}} \cdot 100 = \frac{433 \cdot 1.177 + 748 \cdot 1.011 + 188,3 \cdot 1.073 + 3.579 \cdot 386}{460 \cdot 1.77 + 725 \cdot 1.011 + 167,7 \cdot 1.073 + 2.958 \cdot 386} \cdot 100 =$$

$$= \frac{2.849.408,9}{2.596.125,1} \cdot 100 = 1,098 \cdot 100 = 109,8\%$$

¿Es correcto afirmar que, siendo $IP^P_{00/01} = 109,8\%$:

- *los precios del año 2000 fueron (en conjunto o promedio) un* 9,8% *superiores que los del 2001*?,
- *¿esta variación en los precios se explica porque las cantidades comercializadas en el 2000 valorizadas a los precios vigentes en ese año, suman un valor de* $2.849.408,9; *mientras que esas mismas cantidades pero a los precios del 2001, alcanzan el valor total de* $2.596.125.1?

b- Completar el cálculo del IP^P con base 2001=100 para todos los períodos de la serie.

c- Comparar los resultados obtenidos por este índice con los valores del IP^L calculados en el problema anterior.

d- Considerando que ambos índices son diferentes métodos para medir el mismo fenómeno: la variación conjunta o promedio de los precios de los cuatro artículos líderes de nuestra AVYT, con referencia a un mismo período base:

1. *¿A qué razón atribuye Ud. el hecho de que los valores de IP^L e IP^P en general son diferentes para períodos idénticos*?
2. *¿Qué argumentos teóricos y prácticos consideraría Ud. para decidirse a utilizar uno u otro método en un problema como el que nos ocupa*?

Cuarta Parte

a- Los índices de cantidad de Laspeyres y Paasche (de agregado) para el año 2003 con base 2001=100, son los siguientes:

$$IQ^{L}_{03/01} = \frac{\sum_{1}^{4} q_{i5} p_{i0}}{\sum_{1}^{4} q_{i0} p_{i0}} \cdot 100 = \frac{1.056 \cdot 460 + 474 \cdot 725 + 1.443 \cdot 167,7 + 271 \cdot 2.958}{1.222 \cdot 460 + 1.230 \cdot 725 + 1.158 \cdot 167,7 + 460 \cdot 2.958} \cdot 100 =$$

$$= \frac{2.849.408,9}{3.008746,6} \cdot 100 = 0,623 \cdot 100 = 62,3\%$$

$$IQ^{P}_{03/01} = \frac{\sum_{1}^{4} q_{i5} p_{i5}}{\sum_{1}^{4} q_{i0} p_{i5}} \cdot 100 = \frac{1.056 \cdot 505 + 474 \cdot 1.362 + 1.443 \cdot 291,1 + 271 \cdot 11.771}{1.222 \cdot 505 + 1.230 \cdot 1.362 + 1.158 \cdot 291,1 + 460 \cdot 11.771} \cdot 100 =$$

$$= \frac{4.788.866,3}{8.044.123,8} \cdot 100 = 0,595 \cdot 100 = 59,5\%$$

Completar el cálculo de ambos índices para toda la serie.

b- Con estos resultados analizar la evolución de las cantidades comercializadas por nuestra AVYT durante el quinquenio que nos ocupa; y relacionar el comportamiento de estos índices con los respectivos índices de precios (IP^{L} e IP^{P}) calculados en los dos problemas anteriores.

Actividad Nº 4

Con esta actividad retomaremos el análisis de los datos sobre las exportaciones misioneras de yerba mate, té y tung con los que hemos trabajado en la Actividad Nº 1. Ahora, con los índices Rs de precio y de cantidad que Ud. calculó con base 1990=100 (consigna c), ejercitaremos el cálculo e interpretación de los índices ponderados del *promedio de relativos* de Laspeyres y Paasche.

a- Corroborar que los índices de precio y de cantidad del promedio de relativos de Laspeyres, para el año 1994 con base 1990=100, se determinan mediante el siguiente cálculo:

$$IP^{L}_{94/90} = \frac{\sum_{1}^{3} \frac{p_{i5}}{p_{i0}} p_{i0} q_{i0}}{\sum_{1}^{3} p_{i0} q_{i0}} \cdot 100 =$$

$$= \frac{1,129(0,835 \cdot 4.266) + 1,130(0,77 \cdot 42.584) + 1,363(0,743 \cdot 8.550)}{0,835 \cdot 4.266 + 0,77 \cdot 42.584 + 0,743 \cdot 8.550} = 116,5\%$$

$$IQ^{L}_{94/90} = \frac{\sum_{1}^{3} \frac{q_{i5}}{q_{i0}} q_{i0} p_{i0}}{\sum_{1}^{3} q_{i0} p_{i0}} \cdot 100 =$$

$$= \frac{3,673(0,835 \cdot 4.266) + 0,967(0,77 \cdot 42.584) + 0,282(0,743 \cdot 8.550)}{0,835 \cdot 4.266 + 0,77 \cdot 42.584 + 0,743 \cdot 8.550} = 109,2\%$$

b- Corroborar que los índices de precio y de cantidad del promedio de relativos de Paasche, para el año 1994 con base 1990=100, se determinan mediante el siguiente cálculo:

$$IP^{P}_{94/90} = \frac{\sum_{1}^{3} \frac{p_{i5}}{p_{i0}} p_{i0} q_{i5}}{\sum_{1}^{3} p_{i0} q_{i5}} \cdot 100$$

$$= \frac{1,129(0,835 \cdot 15.667) + 1,130(0,77 \cdot 41.188) + 1,363(0,743 \cdot 2.415)}{0,835 \cdot 15.667 + 0,77 \cdot 41.188 + 0,743 \cdot 2.415} = 113,9\%$$

$$IQ^{P}_{94/90} = \frac{\sum_{1}^{3} \frac{q_{i5}}{q_{i0}} p_{i0} q_{i5}}{\sum_{1}^{3} p_{i0} q_{i5}} \cdot 100$$

$$= \frac{3,673(0,943 \cdot 4.266) + 0,967(0,87 \cdot 42.584) + 0,282(1,013 \cdot 8.550)}{0,943 \cdot 4.266 + 0,87 \cdot 42.584 + 1,013 \cdot 8.550} = 106,7\%$$

c- Completar el cálculo de estos mismos índices (IP^L, IQ^L, IP^P e IQ^P; todos con base 1990=100) para los años 1992, 1997 y 2000. Dar su interpretación de los resultados.

Actividad Nº 5

Índice de Precios Internos al por Mayor (IPIM). Misiones, 1990/2000.

Año	IPIM (Nivel Gral.)
1990	43,04
1991	90,59
1992	96,02
1993	100,00
1994	99,81
1995	106,27
1996	109,63
1997	109,75
1998	106,22
1999	102,19
2000	106,27

Fuente: Boletín Informativo Techint 305. Enero-Marzo 2001.

Utilizando la serie correspondiente al Índice de Precios Internos al por Mayor (IPIM) nivel general, incluida en el Cuadro siguiente:

a- Actualizar los precios de los tres productos de exportación al año 2000, y

b- Deflactar los precios al año 1992

c- Interpretar los valores obtenidos

EVALUACIÓN PARCIAL -Unidad 6-

Volúmenes y Precios de Cítricos comercializados en el Mercado Central. Ciudad de Bs. As., 1994/2000

Años	Limón		Mandarina		Naranja		Pomelo	
	Toneladas	$/Tn	Toneladas	$/Tn	Toneladas	$/Tn	Toneladas	$/Tn
1994	28587,5	480,8	81677,5	468,3	119579,2	383,3	20570,7	470,0
1995	30360,1	375,8	88887,7	433,3	123022,5	396,7	23701,7	465,0
1996	31374,6	361,7	80093,3	352,5	110780,0	354,2	24824,9	441,7
1997	29057,8	356,7	85615,6	513,3	106777,6	362,5	23151,6	400,8
1998	34455,5	311,7	86752,5	373,3	119056,4	333,3	26286,2	411,7
1999	36513,0	283,3	82684,1	407,5	105257,8	372,5	24904,1	392,5
2000	34676,2	295,8	74643,3	334,2	102556,2	442,5	22758,8	444,2

Basándose en los datos del Cuadro anterior:

1. Describir el comportamiento de los volúmenes comercializados de cada uno de los productos, tomando como base el año 1994.
2. Describir la evolución de los precios de cada uno de los cuatro productos (base 1994).
3. Analizar, con base en 1994, la evolución de precios y cantidades del conjunto de los cítricos. Tomar en consideración que los índices de precios y cantidad utilizados en este caso sean consistentes.
4. Redacte un breve informe sobre el comportamiento de volúmenes y precios de comercialización de los cítricos en el MCBA durante el período considerado. En la presentación de este informe incluya aquellos gráficos que considere pertinentes.

Bibliografía General

ALAMINOS, A. (1993): Gráficos. Colección "Cuadernos Metodológicos", nº 7. Centro de Investigaciones Sociológicas, Madrid.

ANDERSON, D.; SWEENEY, J. D.; WILLIAMS, T. (1999): *Estadística para Administración y Economía*. International Thomson ed., México.

BARBANCHO, A. (1978): *Estadística Elemental Moderna*. Ed. Ariel, Barcelona, España.

BLALOCK, H. M (1986): *Estadística Social*, México, FCE.

BLANCH, N.; JOEKES, S. (1993): *Estadística aplicada a la Investigación*. Curso a distancia. Fac. de Cs. Económicas, Universidad Nacional de Córdoba, Argentina.

CHOU, Ya-Lun (1977): *Análisis Estadístico*. Ed. Interamericana, México.

COLL, S.; GUIJARRO, M. (1998): *Estadística aplicada a la historia y a las Ciencias Sociales*. Edic. Pirámide, Madrid.

CRIVISQUI, E. (1993): *Análisis Factorial de Correspondencias: un instrumento de investigación en ciencias sociales*. Laboratorio de Informática Social, Universidad Católica de Asunción, Paraguay.

DANIEL, W. (1985): *Estadística con aplicación a las ciencias sociales y a la educación*. McGraw-Hill, México.

GÓMEZ de AZEVEDO, A.; BORGES de CAMPOS, P. H. (1981): *Estadística Básica: Cursos de Ciências Humanas e de Educaçao*. Livros Técnicos e Científicos Editora S.A., Rio de Janeiro.

MOOD, A. M. (1965): *Introducción a la Teoría de la Estadística*. Aguilar, Madrid (3ra. Edición).

MOORE, D. (1998): *Estadística aplicada básica.* Antonio Bosch ed., Barcelona (1ra. Ed. 1995).

PILCHER, Donald M. (1990): *Data Analysis for the Helping Professions: A Practical Guide*, Sage Publications, California, USA.

SHAO, S. (1967): *Estadística Para Economistas y Administradores de Empresas*. Herrero Hermanos S.A., México.