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:

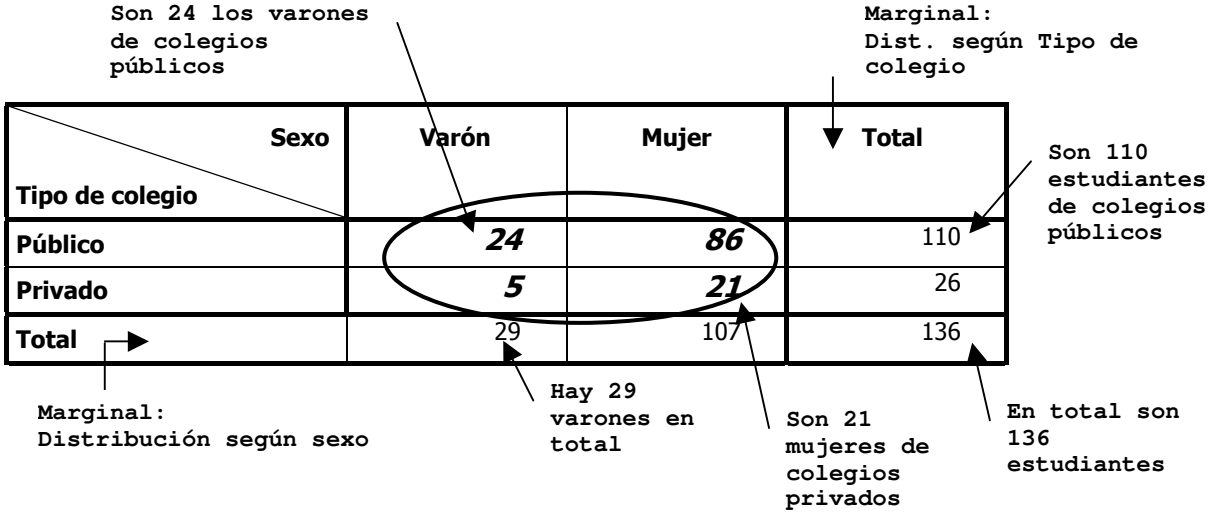
¿Cómo presentar los datos para describir a los individuos a partir de dos variables categóricas simultáneamente?

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

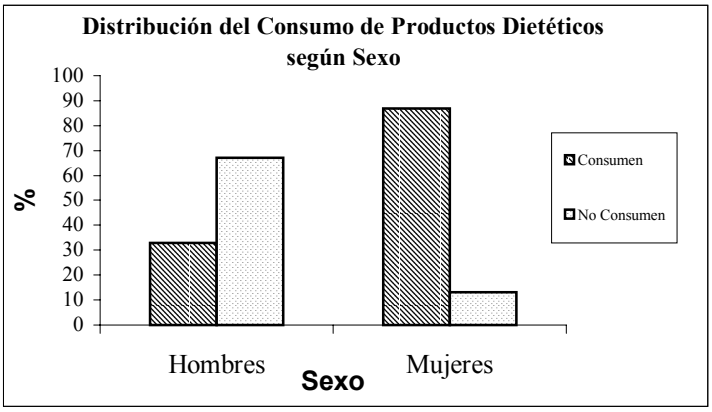
Dado que se trata de variables categóricas, se utilizan gráficos de barras, en el que solo uno de los ejes es numérico. Básicamente pueden distinguirse dos tipos de gráficos:

- ✓ 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).



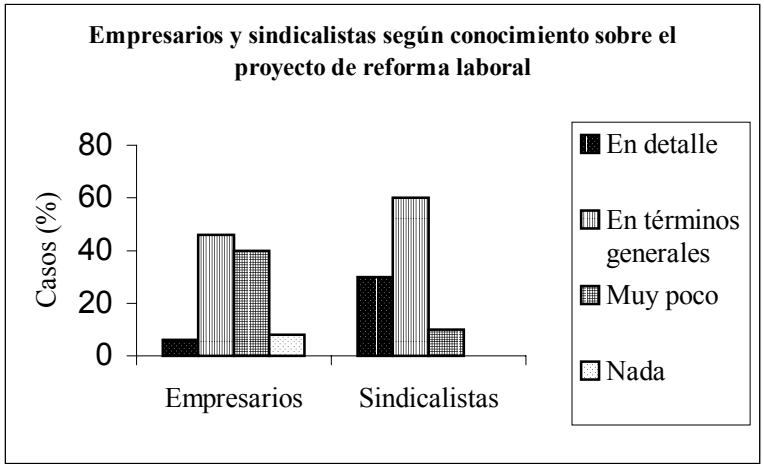
Este gráfico corresponde a la tabla en la que para cada sexo se presenta la distribución (relativa) según el consumo. En consecuencia, el gráfico nos permite comparar la presencia de consumidores y no consumidores en cada sexo.

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.

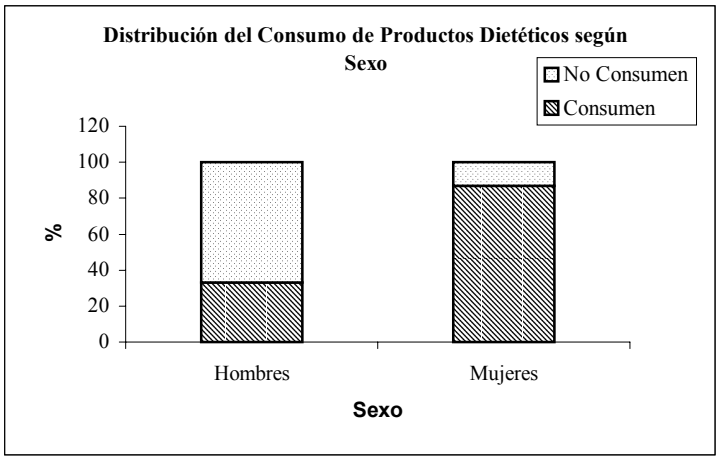


Fuente: elab. propia basándose en datos publicados en el diario Perfil 31/5/98

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%.

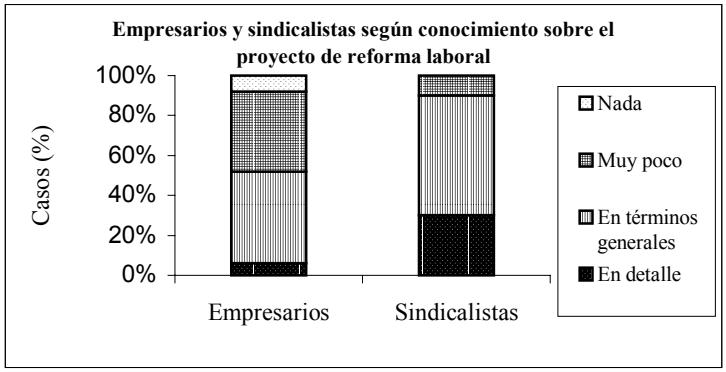
Cada barra es subdividida en tantas categorías como tiene la otra variable. La altura de cada subdivisión se corresponde con la frecuencia absoluta (o relativa) de la categoría correspondiente.



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

$$\eta^2 = \frac{\text{SCentre}}{\text{SCT}} = \frac{3061288}{23925169} = 0{,}128$$

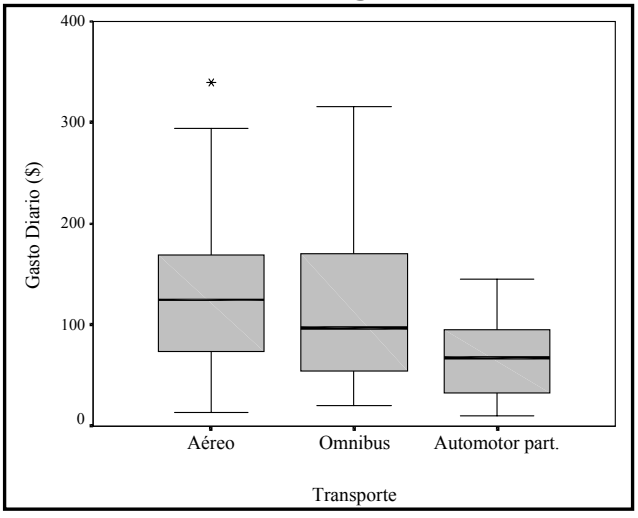
Podemos advertir que si bien, la diferencia entre las medidas de tendencia central eran importantes, la "razón de correlación" está *indicando una relación débil entre las variables.* Esto se debe a que el reagrupamiento generado a partir del nivel de estudio del padre, no produce grupos suficientemente 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.

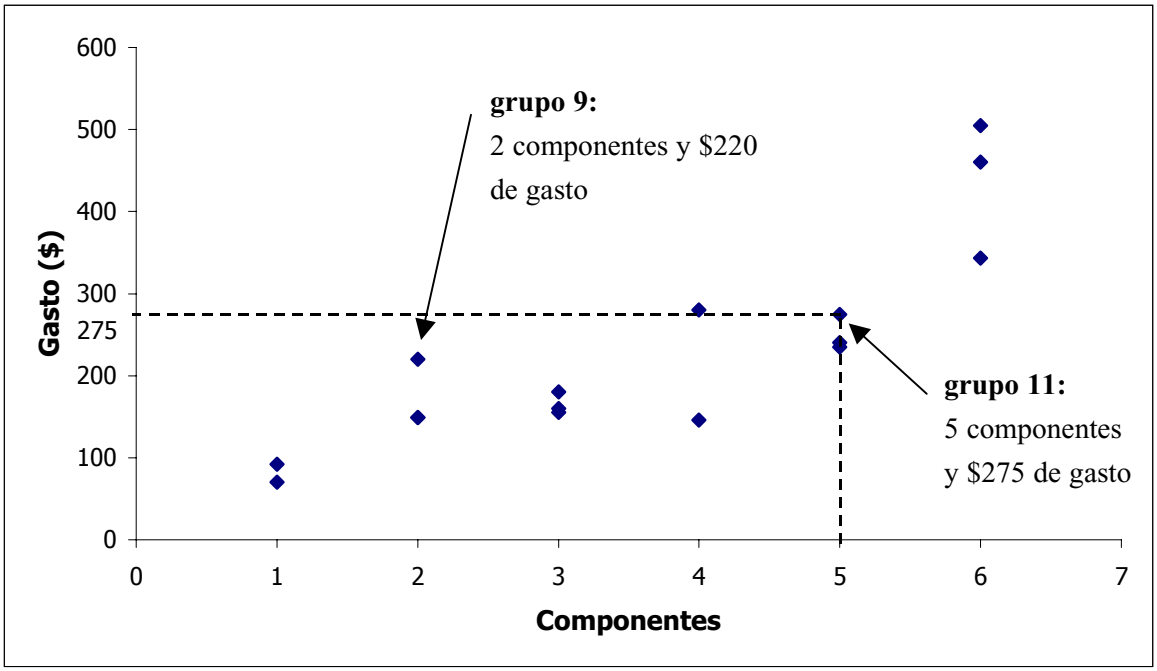
Analicemos a manera de ejemplo, la covariación entre el ***número de componentes*** de los 16 grupos turísticos que visitaron el Parque Nacional Iguazú en febrero de 1994 y el ***gasto diario*** que estos mismos grupos realizaron. Según la definición de correlación, la existencia de una relación entre 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.

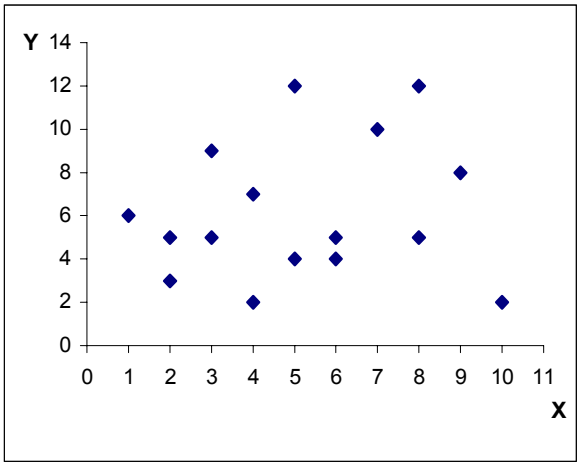
[10] Cuando se trata de una relación causal, la X corresponde a la variable *explicativa*, en tanto que la Y a la variable *explicada*. Además recordemos que la designación de una variable como dependiente o independiente no es una cuestión estadística, sino una decisión que corresponde al conocimiento del investigador sobre el fenómeno que está estudiando.

A través de los **diagramas de dispersión** *podemos estudiar:*

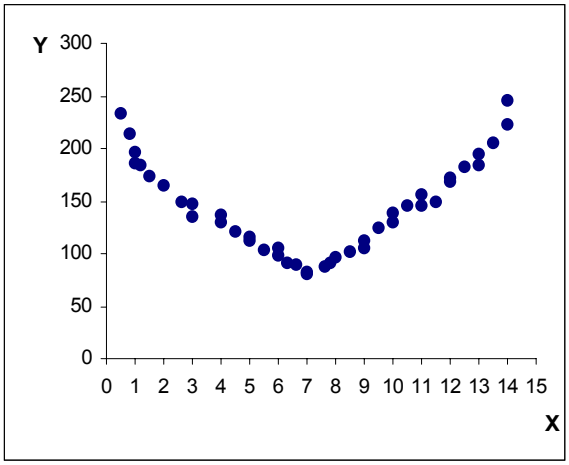
- ✓ ***si existe relación*** *entre las variables,*
- ✓ *caracterizar* ***la forma*** *de la relación, y*
- ✓ *apreciar* ***la intensidad*** *de esa relación.*

¿Cómo se manifestaría gráficamente la relación entre dos variables numéricas?

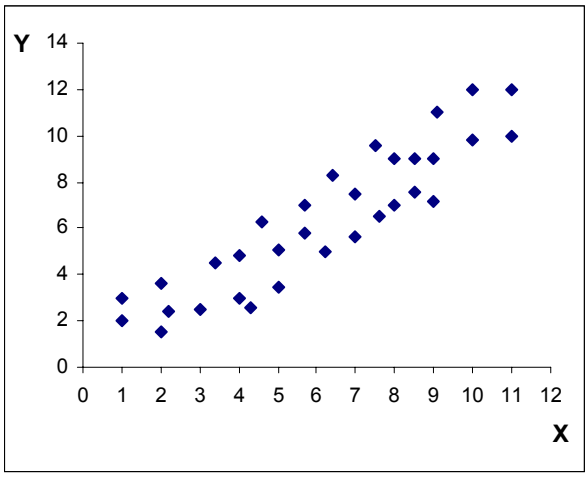
(a) No hay relación



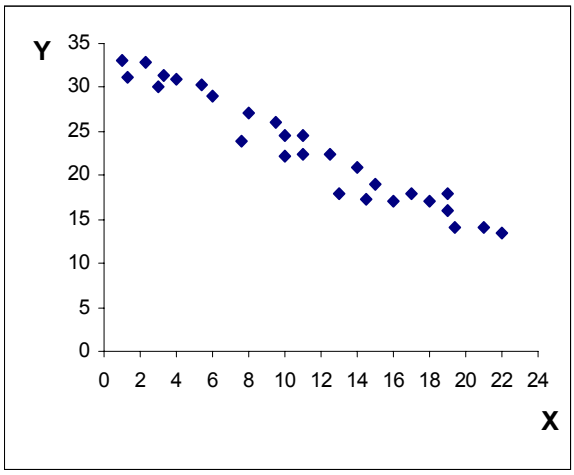
(b) Relación Curvilínea / No lineal



(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

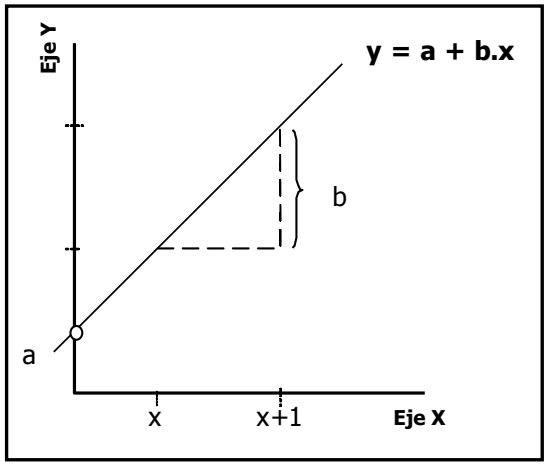
Para el caso de relaciones lineales entre las variables, desarrollaremos:

- **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 : $y = a + bx$

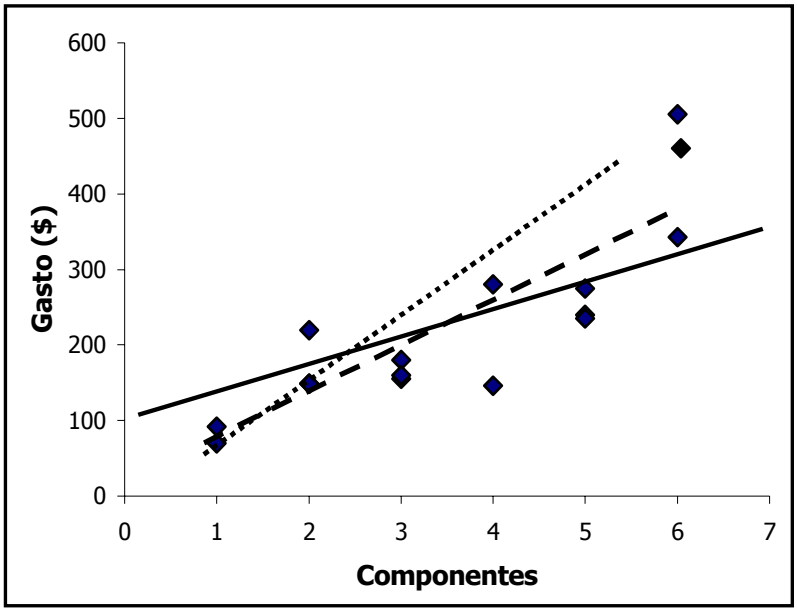


Donde:

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:

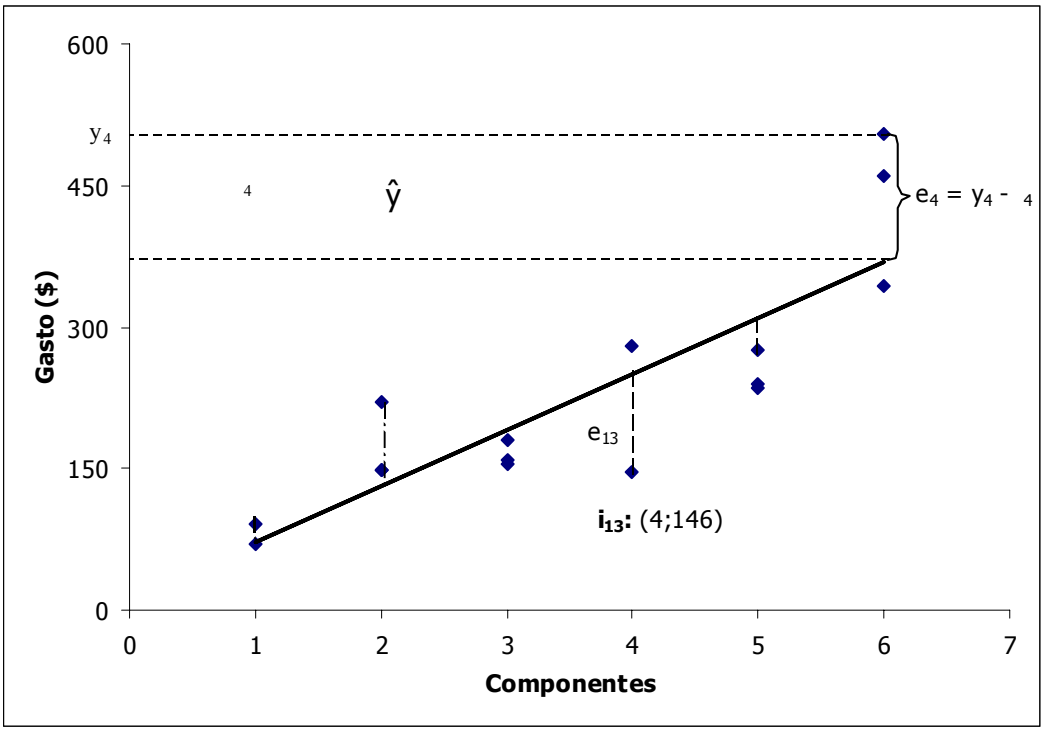
Criterio de *mínimos cuadrados*

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]:

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***.**

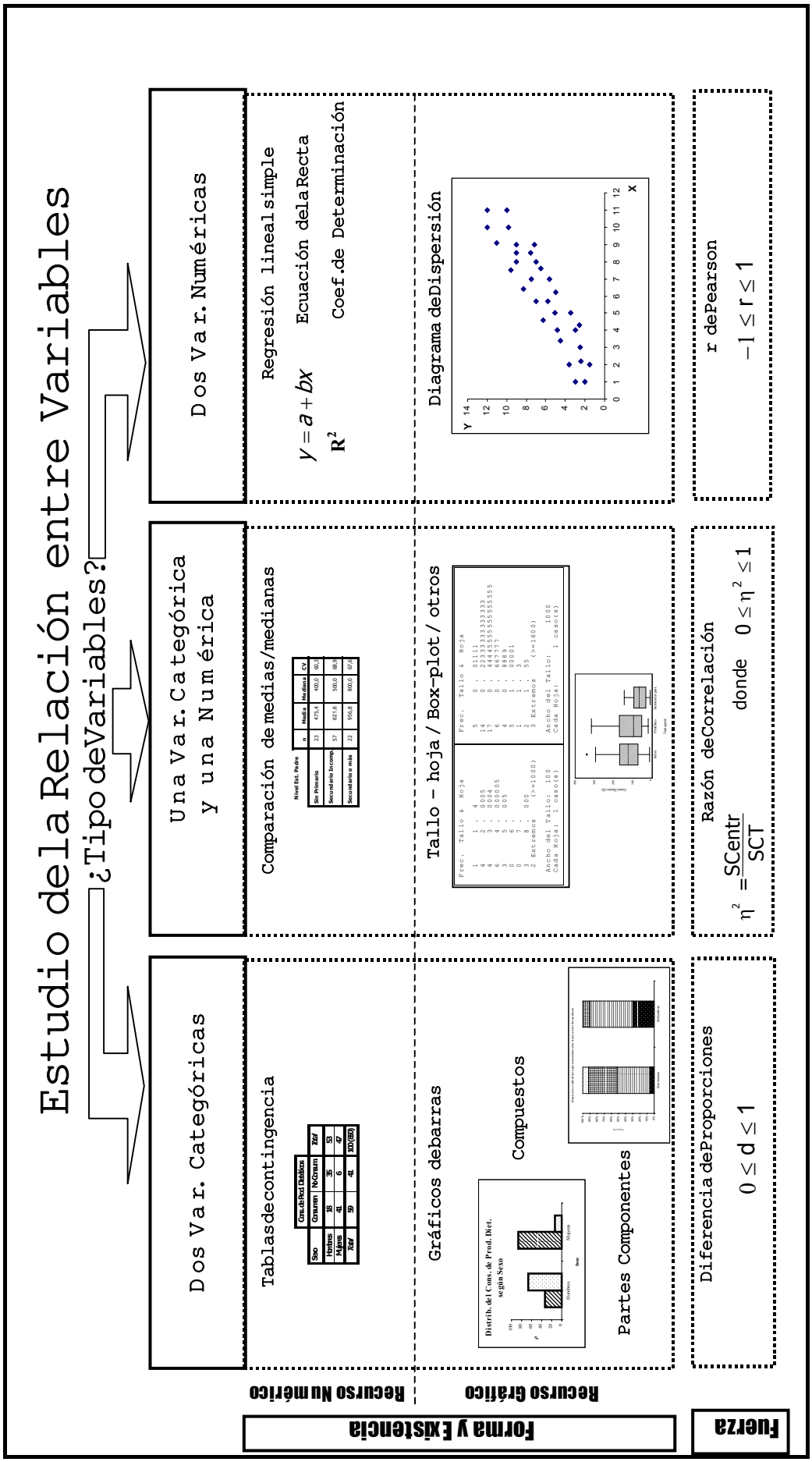
5. ¿Qué Hemos Visto?

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
Razón de Correlación
$\eta^2 = \frac{SCentr}{SCT}$ donde $0 \le \eta^2 \le 1$
Diferencia de Proporciones
$0 \le d \le 1$
Regresión lineal simple
$y = a + bx$ Ecuación de la Recta
R^2 Coef. de Determinación
Diagrama de Dispersión
r de Pearson
$-1 \le r \le 1$
Recurso Numérico
Recurso Gráfico
Forma y Existencia
Fuerza

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.