2010

Universidad de La Habana.
Facultad de Economía.
Departamento de Estadística e Informática.

Profesora: MSc. Mirielys Miranda Iglesias.

[ELEMENTOS DE ESTADÍSTICA PARA LA INVESTIGACIÓN.]

MATERIAL DE ESTUDIO DE LA ASIGNATURA:

Curso Propedéutico, Maestría en Gestión de la Información.

MEDIDAS DESCRIPTIVAS.

Las medidas descriptivas constituyen los indicadores que extraen, representan y resumen el comportamiento de la información.

Dentro de los indicadores se les denomina como estadísticos o estadígrafos si se calculan a partir de la información recolectada en una muestra o parámetros, si se calculan a partir de una población. Además, se pueden clasificar según su función en estadísticos de: posición, dispersión, deformación y apuntamiento.

A continuación se analizan los estadísticos más importantes dentro de cada grupo.

- **MEDIDAS DE POSICIÓN:** Describen la posición o localización de la distribución de frecuencias a lo largo del eje horizontal y por lo general, es posible elegir algún valor promedio que describa todo un conjunto de datos. Las medidas de posición son promedios y pueden ser de tendencia central o no.

Con frecuencia se utilizan, como las más importantes medidas de tendencia central: la media aritmética, la mediana, la moda y la media geométrica. Son medidas de posición pero no de tendencia central: las cuantilas entre las que se encuentran las cuartilas, las decilas y los percentiles, pero éstas no se van a estudiar, por no estar en el programa de estudio.

1. **MEDIA ARITMÉTICA.** También denominada ***media***, es el promedio ó medida de tendencia central, que se utiliza con mayor frecuencia. Se define como la suma de los valores de xi entre el total de elementos. Se representa: En la muestra por $\overline{X}$, y en la población por μ ó M(x).

   Forma de cálculo:

- Datos primarios $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

- Datos Organizados $\overline{X} = \sum_{i=1}^{n} \frac{X_i n_i}{n}$ *ó* $\overline{X} = \sum_{i=1}^{n} X_i f_i$ En el caso de las clases recordar sustituir siempre Xi por la marca de la clase.

Propiedades de la Media: Estas propiedades son de mucha importancia y de utilización práctica en el desarrollo de las Estadísticas.

1.- M(k) = k La media de una constante es igual a la propia constante

2.- M(kx) = K M(x) La media de una constante por una variable es igual a la constante por la media de la variable.

3.- M(k + x) = k + M(x) La media de una constante más una variable es igual a la constante más la media de la variable.

4.- M(x + y) = M(x) + M(y) La media de la suma de dos variables es igual a la suma de las medias de ambas variables.

5.- M(X - μ) = 0 La media de las desviaciones con respecto a la media es igual a cero.

Desviaciones es la diferencia entre el valor de la variable y un valor fijo, cuando este valor fijo es la media se le llama desviaciones con respecto a la media.

6.- M(X - μ)^2 = mínimo. La media de las desviaciones con respecto a la media al cuadrado, es un mínimo.

Desventajas de la media:

1. La media se ve afectada por la ocurrencia de valores extremos, por su propia definición, esto quiere decir que si hay algunos valores grandes que difieren del resto, la distribución se deformará a la derecha, por el contrario si hay algunos valores pequeños que difieren del resto, la distribución se deformará a la izquierda. Y el valor de la media tiende a buscar la deformación y por tanto no será representativa de esa distribución. Una distribución se dice que es deformada cuando al dibujarla no se

asemeja a una campana boca abajo, sino que un lado se alarga más que el otro.

2. Otra desventaja de la media, es que no se puede calcular cuando existen intervalos abiertos, a menos que se eliminen los mismos y esto, aumenta la pérdida de información.

3. **MEDIA GEOMÉTRICA:** Se utiliza para hallar promedios cuando la variable X mide la razón o la proporción, se denota por g_x y se calcula a través de la raíz enésima del producto de cada X_i por su n_i correspondiente, sin distinción entre datos simples u organizados.

Forma de cálculo:

- Datos primarios y Organizados: **Mg_x** = $\sqrt[n]{\prod Xi^{ni}}$

$$Mg = \sqrt[n]{x_1^{n1} x_2^{n2} \ldots x_n^{n_n}}$$

2. **LA MODA:** Se define como el valor más frecuente, en un conjunto de datos. El valor modal siempre es el de mayor frecuencia y en el caso de la organización en clases.

Se denota por Mo ó Md y puede no existir en una distribución, ó existir más de una.

Forma de cálculo:

- Datos primarios: Luego de ordenar los valores de la variable (Xi) se escogerá siempre como Moda el valor que más se repite.
- Datos Organizados pero no Agrupados: El valor modal será el valor de "X" que le corresponde al máximo relativo. (Mayor frecuencia absoluta). Así un máximo relativo será aquel valor que supera al que le antecede y al que le sigue en las frecuencias absolutas

  Máximo relativo: nj-1 < nj > nj+1.

- Datos Organizados y Agrupados: El cálculo de la Moda será a través de la siguiente fórmula:
  Mo = Lj-1 + Cj [(nj - nj-1)/(nj -nj-1) + (nj - nj+1)] donde se cumple que nj es el Máximo Relativo.

Se utiliza en ocasiones como medida de tendencia central y se obtiene fácilmente a partir de un arreglo ordenado. A diferencia de la media, la moda no se afecta ante la ocurrencia de valores extremos. Las distribuciones, pueden tener una, dos o tres modas y en dependencia se les dirá que son unimodal, bimodal o plurimodal; en el caso de no repetirse ningún valor se considerará la no existencia de la misma.

3. **LA MEDIANA (Me)**: Se define como el valor que supera a no más del 50% de las observaciones y es superado, por no más del 50% de las observaciones.

Forma de cálculo:

- Datos primarios: primero, se deben ordenar de menor a mayor. Luego, se busca la posición del valor mediano en el arreglo ordenado; ahora para determinar la Me existen 2 reglas y están en dependencia, del número de observaciones.

  REGLA 1: Si el tamaño de la muestra es un número impar, la mediana está representada por el valor numérico correspondiente a la posición del centro de las observaciones ordenadas.

  REGLA 2: Si el tamaño de la muestra es un número par, entonces la posición mediana, será la semi-suma de los dos valores centrales o la media de los dos valores centrales de las observaciones ordenadas.

- Datos están Organizados pero No Agrupados: Se busca la clase mediana, en las frecuencias absolutas acumuladas. De la manera siguiente:

1.- Se determina la fracción n/2, que ubica el centro de la distribución.

2.- Se representa por **Nj a la menor frecuencia acumulada que supere a n/2** por tanto, quedaría que: **Nj-1 ≤ n/2 < Nj.** De la expresión anterior se pueden presentar dos situaciones:

a- Que **n/2 > Nj-1**; y en ese caso, el valor mediano será el valor de la variable que le corresponde a Nj; esto es, **Me = Xj**

b.- Que **n/2 = Nj-1;** y entonces el valor mediano, será la media del valor de la variable que le corresponde a Nj y el valor de la variable que le corresponde a Nj-1, esto es **Me = (Xj + Xj-1) / 2**

- Si los datos están Organizados y Agrupados: Se utiliza la siguiente fórmula:

  1- **Me = L_{j-1} + C_j [(n/2 - N_{j-1})/ n_j] cuando se cumple que n/2 >N_{j-1}**

  2- **Me = L_{j-1} cuando se cumple que: n/2 = N_{j-1}.** Ya que el resto de la expresión se hace cero, pues dos cosas que son iguales su diferencia es cero.

Ventajas de la mediana:

La mediana no se ve afectada por datos extremos, es por ello que en esos casos es más representativa que la media, como medida de tendencia central.

Observación: Si resulta ser el primer intervalo mediano, se plantea que no existe Nj-1 y no se podría calcular la Me. No obstante se suele hacer el supuesto de que al no existir intervalo anterior, entonces Nj-1 = 0.

- **MEDIDAS DE DISPERSIÓN:** Sólo a través de ellas es que se puede determinar si la medida de posición es significativa o representativa de la distribución.

Entre las medidas de dispersión o variabilidad, se van a estudiar: la varianza, la desviación típica y el coeficiente de variación.

Dentro de los estadísticos de dispersión, el más utilizado es la varianza, precisamente por sus propiedades.

1. **LA VARIANZA:** Se define como la media aritmética del cuadrado de las desviaciones de la variable con respecto, a su media**.** Se denota por $\mathbf{S^2}$ en la muestra y por s^2 o **V(x)** en la Población.

Forma de cálculo:

- Datos primarios, viene dado por:

$$S^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 / n \quad \text{ó} \quad \text{también} \quad S^2 = \sum_{i=1}^{n} X_i^2 / n - \overline{X}^2$$

- Para datos Organizados: (no agrupados o agrupados)

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2 n_i}{n} \quad \text{ó} \quad \text{también} \quad S^2 = \sum_{i=1}^{n} \frac{X_i^2 n_i}{n} - \overline{X}^2$$

Mientras mayor sea la varianza más dispersos estarán los valores alrededor de la media, y mientras más pequeña, menos dispersión habrá, pero estas son cantidades elevadas al cuadrado y no se puede interpretar como tal y tampoco da una idea clara.

Propiedades de la Varianza:

1.- V(x) ≥ 0 La varianza es un número no negativo

2.- V(k) = 0 La varianza de una constante es igual a cero

3.- V(X ± k) = V(x) La varianza de la suma de una variable más una constante es igual a la varianza de la variable.

4.- V(kx) = k^2 V(x) La varianza del producto de una constante por una variable es igual a la constante al cuadrado por la varianza de la variable.

Como se ha dicho la varianza no se puede interpretar por estar su resultado elevado al cuadrado, por lo que es conveniente contar con otro estadístico que basado en el valor de la varianza sirva para dar una medida de la dispersión en las mismas unidades o dimensiones en que están expresados los datos y este estadístico es: La Desviación Típica.

2. **LA DESVIACION TÍPICA:** Se define como la raíz cuadrada positiva de la Varianza. Se denota por **S** en la muestra y por s en la población.

La desviación típica es una magnitud no negativa, pero no cumple las restantes propiedades de la Varianza, pues la extracción de la raíz no lo permite.

Forma de cálculo: $S = + \sqrt[2]{S^2}$

En ocasiones resulta necesario contar con un estadístico que refleje la dispersión sin depender de la magnitud de las observaciones, esto es que sea un valor relativo. Esta necesidad surge generalmente, cuando se comparan las dispersiones entre varios conjuntos expresados en unidades de tiempo diferentes. Este estadístico es el Coeficiente de Variación y que se verá a continuación:

3. **COEFICIENTE DE VARIACIÓN:** Se define como la razón entre la desviación típica y la media y su interpretación debe ser en %, por lo cual debe multiplicarse por 100 el resultado. Se denota por **C.V** tanto para datos en su forma primaria u organizada.

Forma de cálculo: C.V= $\frac{S}{\bar{X}} \times 100\%$

Este coeficiente es una medida de dispersión igual que la desviación típica, pero la diferencia radica en que es una medida relativa, mientras que la varianza y la desviación son medidas absolutas.

Ventajas del C.V:

1- Permite comparar distribuciones que estén expresadas en unidades de medidas diferentes ya que el resultado viene expresado en porciento.

2- También se debe señalar que por ser el resultado en porciento, es de más fácil comprensión definir si la variabilidad dada por el coeficiente de Variación es más alta o más baja, porque en ocasiones a través de la desviación típica esto se hace más difícil y a veces ni siquiera es posible.

- **MEDIDAS DE PORCIÓN:**

  1. **COEFICIENTE DE PROPORCIÓN:**

  Forma de cálculo: $\hat{p} = \frac{x}{n}$ Indica la proporción de elementos que cumplen con la característica evaluada, posteriormente en los próximos temas se realizará un mayor énfasis en su utilización.

- **MEDIDAS DE CORRELACIÓN.**

  1. **LA COVARIANZA:** Es una medida de la variación conjunta de cada variable respecto a su media. La covarianza se debe utilizar para medir la variación conjunta (covariación) de las variables X e Y.

  Forma de cálculo:

$$Cov(xy) = \sum_{j=1}^{m}\sum_{x=1}^{k}(XiYj)f(x_i\, y_j) - \overline{X}\,\overline{Y} \quad \text{ó} \quad \sum_{j=1}^{m}\sum_{x=1}^{k}(XiYj)\, nij/n - \overline{X}\,\overline{Y}$$

Es una medida de dispersión de los valores de la variable y una medida de la relación entre ellas.

Propiedades de la Covarianza:

1.- Si X e Y son independiente entonces Cov(xy) = 0, sin embargo el recíproco no se cumple, esto es que la Cov(xy) = 0, no es una condición suficiente para decir que las variables X e Y son independientes.

2.- V(x+y) = V(x) + V(y) + 2Cov(xy)

V(x-y) = V(x) + V(y) - 2Cov(xy)

Si X e Y son independientes, entonces V(x + y)= V(x) + V(y) ya que la Cov(xy)=0

Desventaja de la Covarianza: Esta medida no debe ser utilizada de modo exclusivo para medir la relación entre las dos variables, ya que es sensible al cambio de escala. Cambio de escala, quiere decir que si cada valor de la variable X, se multiplica por una constante "b", la media aritmética queda multiplicada por esa constante.

Dado que la covarianza se ve afectada por un cambio de escala, es necesario definir una medida de relación entre dos variables, y que no esté afectada por los cambios de unidad de medida.

2. **EL COEFICIENTE DE CORRELACIÓN.** Si se divide la Covarianza por el producto de las desviaciones típicas de cada variable, se obtiene el Coeficiente de Correlación lineal de Pearson. Se interpreta como el grado de la relación mientras que la covarianza sólo informaba el tipo de relación que existía entre las variables a partir del signo si este era positivo o negativo. Este coeficiente permite cuantificar el grado de relación o asociación que existe entre las variables, suponiendo que entre ellas pudiera existir una relación lineal.

Se representa en la muestra por **r** y en la población por r

Forma de cálculo:

$$\mathbf{r_{xy} = Cov_{(xy)} / S_x S_y}$$

Propiedades del Coeficiente de Correlación:

1. Carece de unidad de medida
2. Es invariable en transformaciones lineales (cambio de origen y escala)
3. Solo toma valores comprendidos entre –1 y 1. Cuando |r| está muy próximo a 1 existirá una relación lineal, muy fuerte entre las variables.
4. Cuando $r \approx 0$, puede afirmarse que no existe relación lineal entre las variables.
5. Si X y Y son independientes, entonces r(xy) = 0

Otro aspecto importante a comprobar en las distribuciones bidimensionales es la independencia entre las variables

Se considera que X y Y se distribuyen independientemente si y solo si:

- **f(xy) = f(xi)f(yj)**

Cuando X y Y se distribuyen independientemente se puede decir que las distribuciones de X y Y son independientes o simplemente que X y Y son independientes.

ALGUNAS CONSIDERACIONES DE INTERÉS PARA EL APOYO EN EL CÁLCULO DE LOS ESTADÍGRAFOS DE CORRELACIÓN:

DISTRIBUCIONES BIDIMENSIONALES

A lo largo de los capítulos precedentes habíamos estudiado la distribución de una variable que expresaba la medida de un carácter cuantitativo. Pero para una población dada, se pueden estudiar simultáneamente dos o más caracteres cuantitativos diferentes. Por ejemplo, se puede medir sobre un cuadro de salarios a la vez el salario percibido y la antigüedad en la empresa o, sobre una población de estudiantes, la nota obtenida en una prueba y la edad de los candidatos.

DISTRIBUCION BIDIMENSIONAL DE FRECUENCIAS

De forma general, si se estudian sobre una misma población y si se miden por las mismas unidades estadísticas un carácter X y un carácter Y(ambos cuantitativos) se obtienen dos series, de las variable X e Y.

Considerando simultáneamente las dos series, es decir, para cada unidad estadística el par de los valores (xi,yj) que le corresponde, se suele decir que estamos ante una estadística de dos dimensiones, o variable estadística bidimensional.

Se puede, evidentemente, estudiar separadamente la distribución de la población según el carácter X o el carácter Y, y resumir cada una de las distribuciones, por ejemplo, calculando $\overline{X}$ $\overline{Y}$, Sx, Sy, etc., pero puede ser interesante considerar simultáneamente los dos caracteres a fin, de estudiar las posibles relaciones entre ellos y poder responder a cuestiones como, ¿existe una relación entre los valores del carácter X y el carácter Y? Por ejemplo, ¿existe una relación entre la nota obtenida y la edad del candidato?

Aunque la formulación de estas cuestiones sea "neutra", se puede percibir lo que, al menos implícitamente, pueden significar: ¿existe una relación causal entre X e Y?, ¿la antigüedad en la empresa determina el nivel salarial?, ¿La nota obtenida depende de la

edad del candidato? Pero ningún instrumento estadístico puede permitir afirmar que existe una relación de causalidad entre dos caracteres. Sin embargo, existen instrumentos estadísticos que permiten revelar la existencia de coincidencias entre los valores de dos variables; y a partir de la constatación de esas coincidencias se puede eventualmente formular la hipótesis de una relación causal entre los dos caracteres. Este es el interés fundamental del estudio de dos variables estadísticas con dos dimensiones.

Si existen coincidencias estadísticas entre los valores de dos caracteres, si existe una relación entre las dos variables, las coincidencias pueden ser más o menos fuertes, y la intensidad de la relación puede variar entre dos extremos: ausencia total de ligazón o relación, o ligazón perfecta. Vamos a estudiar ahora los dos extremos antes de entrar en el tema, más complicado, de las situaciones intermedias.

INDEPENDENCIA Y RELACION FUNCIONAL DE DOS VARIABLES.

Cuando no existe relación entre dos variables, se dice que las variables son independientes. Inversamente, cuando la relación entre dos variables es perfecta, se dice que las variables están relacionadas funcionalmente, lo que significa que su relación puede ser expresada bajo la forma y = f (x).

Diremos que Y depende funcionalmente de X cuando podamos establecer una aplicación que nos transforme los elementos de X en elementos de Y.

Pues bien, desde el punto de vista de la estadística, lo verdaderamente importante es que a través de esa función, se pueden determinar inequívocamente los elementos de Y conocidos los de X (o viceversa).

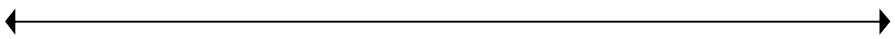
Un ejemplo de este tipo de relación podría ser la existente entre el espacio y el tiempo, para una velocidad determinada, en el movimiento uniforme, ya que sabemos que s = v.t, y esto nos permite una determinación exacta de "s" para los diferentes "t".

Pero existen otros tipos de características como estatura, peso, consumo, y renta, opiniones sobre cierto tema y nivel de educación etc., en los que no cabe duda de que existe interrelación, pero en los que es imposible definir sobre ellas una aplicación en el sentido estrictamente matemático. Es decir, no dependen funcionalmente una de otra.

Ahora bien, estadísticamente es claro que el peso depende en cierta forma de la estatura, el consumo de la renta, etc. Este tipo de relación no expresable a través de una determinada aplicación es la conocida como dependencia estadística. Y así como la dependencia funcional no admite grados, la dependencia estadística si los admite, ya que pueden existir dependencias más o menos fuertes.

La situación opuesta a la dependencia funcional es la independencia completa entre los dos fenómenos (por ejemplo, precio de varas de bambú en China y producción de acero en España).

Estos tipos de dependencia se podrían representar en un segmento de la recta real, en donde en un extremo se situaría la dependencia funcional y en el contrario la independencia. Los puntos intermedios corresponderían a los diferentes grados de dependencia estadística (Fig.6.1).



dependencia ←---------dependencia ----------→independencia

funcional estadística

Para distinguir la dependencia existente entre dos variables (consumo y renta) o entre dos atributos (opiniones y nivel de educación) hablaremos de asociación entre variables y de contingencia entre atributos. Este último aspecto, el relativo a atributos, será estudiado más adelante en el capítulo XIV con más detenimiento.

DISTRIBUCIONES BIDIMENSIONALES. TABLAS DE CORRELACION Y DE CONTINGENCIA.

Sea una población estudiada simultáneamente según dos caracteres X e Y; representaremos gen,ricamente la distribución de variables por (x_i ,y_j ;nij), donde x_i, y_j son dos valores cualesquiera y n_{ij} es la frecuencia conjunta del valor i-,simo de X con el j-,simo de Y.

Una forma de disponer los resultados es la conocida como "tabla de correlación", que es una tabla de doble entrada como la siguiente:

X \ Y	y^1	y_2	...	y_j	...	y_k	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	$n_1.$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	$n_2.$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	$n_{i.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	$n_{h.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.k}$	N

Por ejemplo, n_{11} nos dice el número de veces que se ha presentado x_1 conjuntamente con y_1; n_{12}, la frecuencia conjunta de x_1 con y_2, etc.

El número total de individuos observados es N.

Si la distribución bidimensional es de atributos, la tabla de doble entrada se llama de contingencia. Por ejemplo, supongamos que podemos aglutinar las diferentes respuestas a una cierta pregunta en cuatro categorías A, B, C, Y D, y que intentamos relacionar estas respuestas con el nivel cultural. La tabla de contingencia podría ser la siguiente:

	Estudios		
Respuestas	Est. Primarios	Bachillerato	Est. Superiores
A	n_{11}	n_{21}	n_{31}
B	n_{12}	n_{22}	n_{32}
C	n_{13}	n_{23}	n_{33}
D	n_{14}	n_{24}	n_{34}

Otra forma de disponer los resultados, a veces más cómoda es:

xi	yj	nij
x_1	y_1	n_{11}
.	.	.
.	.	.
.	.	.
xi	yj	nij
.	.	.
.	.	.
.	.	.
x_h	y_k	n_{hk}
		N

DISTRIBUCIONES MARGINALES

Puede ocurrir que nos interese a partir de una distribución bidimensional, estudiar aisladamente cada una de las variables. De esta forma tendríamos dos distribuciones unidimensionales, que serían las distribuciones de X y de Y, respectivamente.

Para obtenerlas tenemos que determinar las frecuencias marginales. En la distribución marginal de X, tenemos que hallar cuántas veces se repite cada valor de xi con independencia de que aparezca conjuntamente o no con algún valor de Y. As¡ el número de veces que se repite en total x1, con independencia de los valores de Y, según la tabla de correlación, será :

$$n_1. = n_{11} + n_{12} + ... + n_{1k} = \sum_{j=1}^{k} nij$$

que se lee "n subíndice uno punto" y que es la frecuencia marginal de x_1. Para un valor í-ésimo de X, su frecuencia marginal será:

$$n_i. = n_{i1} + n_{i2} + ...+ n_{ik} = \sum_{j=1}^{k} n_{ij}$$

Análogamente, para la distribución marginal de Y, la frecuencia marginal de un valor j-esimo cualquiera será:

$$n.j = n_{1j} + n_{2j} + ... + n_{hj} = \sum_{i=1}^{h} n_{ij}$$

Vemos pues que la última fila y la última columna del cuadro corresponden a las frecuencias marginales.

Las distribuciones marginales de frecuencias serán:

X		Y	
xi	ni.	yj	n.j
x_1	$n_1.$	y_1	$n._1$
x_2	$n_2.$	y_2	$n._2$
.	.	.	.
.	.	.	.
.	.	.	.
xi	ni.	yj	n.j
.	.	.	.
.	.	.	.
.	.	.	.
x_h	$n_h.$	y_k	$n._k$
	N		N

Es claro que: $\sum_{i=1}^{h} ni. = \sum_{j=1}^{k} n.j = \sum_{i=1}^{h}\sum_{j=1}^{k} nij = N$

DISTRIBUCIONES CONDICIONALES

Análogamente las distribuciones condicionales de Y a un valor i-ésimo de X serán:

yj/xi	nj/i
y_1	n_{i1}
y_2	n_{i2}
.	.
.	.
.	.
yj	nij
.	.
.	.
.	.
y_k	n_{ik}
	ni.

Las frecuencias relativas de las distribuciones condicionadas a algún valor de "Y", ó a algún valor de "X" serán, respectivamente

f i/j = nij/n.j f j/i = nij/ni.

Vamos a ver un ejemplo. Sea la siguiente tabla de doble entrada:

Y / X	1	2	3	4	ni.
5	1	2	1	3	7
10	2	1	3	2	8
15	3	2	1	2	8
n.j	6	5	5	7	23

Se pide calcular:(a) la distribución marginal de la Y

(b) la distribución condicionada de X/Y = 2

Estas distribuciones serán:

	(a)			(b)		
yj	n.j	n.j/N		xi/y=2	ni/2	f i/2
1	6	6/23		5	2	2/5
2	5	5/23		10	1	1/5
3	5	5/23		15	2	2/5
4	7	7/23			5	1
	23	1				

INDEPENDENCIA ESTADISTICA

Dos variables X e Y se dice que son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales, es decir:

$$nij/N = ni./n \cdot n.j/N \quad \text{para todo } i,j$$

En este caso, las frecuencias relativas condicionadas serán:

$$f\ i/j = nij/n.j = ni.(n.j/N)/n.j = ni./N$$

$$f\ j/i = nij/ni. = ni.(n.j/N)/ni. = n.j/N$$

Es decir, las frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias relativas marginales, lo que nos indica el condicionamiento, en cuanto tal, no existe: las variables son independientes, puesto que en las distribuciones marginales se estudia el comportamiento de una variable con independencia de los valores que pueda tomar la otra.

RESUMEN DE LAS HABILIDADES A EVALUAR:

- ***SABER CALCULAR TANTO PARA DATOS SIN AGRUPAR COMO AGRUPADOS LAS MEDIDAS DESCRPTIVAS ORIENTADAS.***