Capítulo 1

Estadística descriptiva

Con estas nota de clase iniciamos el curso. El objetivo de este primer capítulo es dar unas nociones básicas de descripción de datos. Tiene que quedar claro que hay muchas otras técnicas básicas y avanzadas de descripción de datos que no veremos. Para la resolución de problemas se tienen que utilizar tanto la resolución tradicional con papel, lápiz y calculadora, como la resolución con hojas de cálculo (Excel,OpenOffice...) o paquetes estadísticos (SPSS, R,...), o programas de propósito todavía más general (Mathematica, Octave...). Algunos de estos paquetes están disponibles en los ordenadores de las aulas de informática de la universidad y de estos algunos otros tienen licencias que os permiten bajarlos por internet de forma gratuita; como el OpenOffice (www.openoffice.org)o el paquete estadístico R (http://www.r-project.org) en versiones para cualquier sistema operativo usual.

1.1. Conceptos básicos

La estadística es aquella ciencia que tiene por objeto dar métodos tanto para la recopilación, organización y análisis de datos que provienen de un grupo de individuos, así como para la decisión de aceptar o rechazar ciertas afirmaciones o leyes.

Conceptos básicos:

- Población: Conjunto de todos los individuos que tienen en común alguna característica observable y de los que se desea estudiar un determinado fenómeno. Sus características se definen como parámetros.

Tipos de población: finita o infinita.

- Muestra: Es un subconjunto de la población del que se espera represente a la población y en el que se efectúa el estudio del fenómeno. Sus características se definen como estadísticos.

1.1.1. Estadística descriptiva

La Estadística Descriptiva se define como aquella ciencia dedicada a describir las regularidades o características de un conjunto de datos (muestra).

Tareas de la Estadística Descriptiva:

- Organización de los datos numéricos de la muestra mediante tablas y representaciones gráficas.
- Análisis de los datos obtenidos mediante la obtención de índices representativos de la muestra como son las medidas de tendencia central y de dispersión.

1.1.2. Estadística Inferencial

La Estadística Descriptiva basa su estudio sobre las muestras. Ver si éstas son representativas de la población es tarea de la estadística inferencial.

La misión principal de la Estadística Inferencial es extraer conclusiones de las características de la población mediante una muestra representativa de la misma.

1.2. Estadística descriptiva

1.2.1. Datos y series estadísticas

El análisis estadístico parte siempre de un conjunto de datos. Dado un conjunto de objetos cualesquiera (individuos, países, municipios, etc...), la observación de una determinada característica o medida de ésta (cualidad o atributo) da lugar a un dato estadístico.

Ejemplos:

- Población: Países del mundo.

Característica a estudiar: Producto Interior Bruto (P.I.B.). Los datos estadísticos serán los valores del P.I.B. de los países en cuestión.

- Población: Estudiantes de segundo curso de Informática.

  Característica a estudiar: Altura. Los datos estadísticos serán los valores de la altura en cm. para cada estudiante.

1.2.2. Clasificación de los datos

Una clasificación elemental de los datos estadísticos es la siguiente, dividida en tres criterios:

- Tipo de dato
  - Cualitativos o de atributos: cuando la comparación entre ellos sólo puede ser de igualdad o desigualdad.

    Por ejemplo: color de los ojos, afiliación política, lugar de residencia, etc,...
  - Ordinales: cuando los datos no son numéricos y la comparación entre ellos establece un orden.

    Por ejemplo: estado de ánimo (valores posibles: depresivo, normal y eufórico), estudios (valores posibles: ninguno, primarios, secundarios, superiores), etc...
  - Cuantitativas: cuando los datos son numéricos. Entre los datos cuantitativos podemos señalar dos tipos más:
    - Discretos: cuando entre dos posibles valores no hay otro. Por ejemplo: número de hijos de una familia, número de letras de una palabra en un texto, etc,...
    - Continuas: cuando entre dos posibles valores, siempre podemos encontrar otro valor posible. Por ejemplo: altura, intereses de una cuenta bancaria, etc,...
- Dimensión
  - Unidimensionales: si sólo se considera una única característica.

    Ejemplos: altura, edad, etc,...

- Multidimensionales: si se consideran conjuntamente varias características.

  Ejemplos: edad y altura, altura y peso, edad, altura y sexo, etc,...

- Tiempo

  - Atemporales: cuando los datos no están referidos, o no se considera, el momento de tiempo en el que fueron obtenidos.

    Ejemplos: color de los ojos de cierto conjunto de individuos, peso de los estudiantes que han asistido a la clase de hoy, etc,...

  - Temporales o series cronológicas: en caso contrario.

    Ejemplos: P.I.B. anual de España durante el periodo 1980 hasta 2004, número de turistas llegados al aeropuerto de Palma el mes de agosto durante los años 1970 al 2004, etc,...

1.2.3. Descripción de una serie

Una vez realizada la recogida de datos, se ha de hacer una representación numérica y descriptiva de éstos que se adecue de la mejor manera posible al estudio que se desea realizar.

Cuando los datos son atemporales y unidimensionales, es habitual presentarlos en forma de distribución de frecuencias asociando a cada modalidad o valor las veces que se repite (frecuencias absolutas).

En el caso en que los datos sean bidimensionales y atemporales se suele hacer una tabla de frecuencias denominada también como tabla de contingencia.

En el caso en que los datos sean series cronológicas, se representan como una función matemática en el tiempo, es a decir, una serie de valores (t, Y_t) donde el primer elemento es el tiempo y el segundo valor es el dato en ese tiempo.

1.2.4. Representación gráfica

Una vez descrita la serie estadística en forma de tabla, el paso siguiente es hacer una representación gráfica de la misma porque lo interesante es observar de golpe el aspecto general de los datos.

Veamos con unos cuantos ejemplos en los que esquemáticamente veremos las representaciones gráficas más habituales.

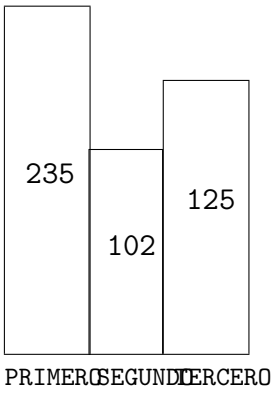

Figura 1.1: Alumnos de los distintos cursos de Informática

- Diagramas de barras. Como ejemplo, ver en la figura 1.1 la representación gráfica de la cantidad de alumnos que hay en distintos cursos de informática.
- Gráficos de sectores: Es un gráfico circular dividido en sectores donde cada sector representa el tanto por ciento de individuos que pertenecen a una determinada modalidad.
- Pictogramas: Son representaciones gráficas que guardan relación con el objeto de estudio estadístico.

Diagramas causa-efecto

Se utilizan en las empresas e industrias, para representar los factores que influyen en un fenómeno.

Por ejemplo, consideremos el diagrama de la figura 1.2:

En la figura anterior tenemos un problema en el que inciden las materias primas (MP, 2 diferentes), los métodos de elaboración (ME, 3 diferentes), la temperatura (TE, 15 grados a 20, o bien de 20 a 30) y los turnos de trabajo (TU, 2 diferentes).

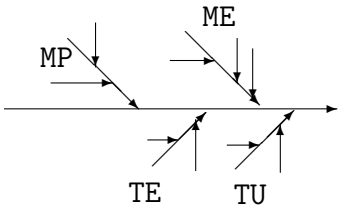

Figura 1.2: Diagrama causa-efecto

1.3. Variables unidimensionales

1.3.1. Descripción numérica

La representación ordenada de las observaciones de una muestra se hace mediante una tabla numérica, en la que aparecen los valores de la variable y sus frecuencias absolutas; número de veces que aparece cada dato en la muestra. Además la tabla se puede completar con las frecuencias absolutas acumuladas.

Mas concretamente, sean $x_1, \ldots, x_n$ las observaciones de una muestra, de tamaño n. Supongamos que los distintos valores que aparecen en las muestra son $X_1, \ldots, X_J$ y que, si es posible están ordenados de menor a mayor:

$$X_1 < X_2, < \ldots, X_J.$$

Denotaremos por n_1 las veces que aparece el valor X_1 en la muestra, n_2 las veces que aparece el valor X_2, ..., y n_J las veces que aparece el valor X_J. Las frecuencias absolutas serán, por lo tanto los valores: $n_j,\ j = 1, \ldots, J$.

Es evidente que se verifica la siguiente relación:

$$n = \sum_{j=1}^{J} n_j$$

Las frecuencias relativas se definen como el cociente entre las absolutas y el tamaño de la muestra: $f_j = \frac{n_j}{n}$. La frecuencia relativa de X_j es el tanto por uno de veces que aparece en la muestra. En ocasiones se utilizan los tantos por cien, por mil,..., pero en la práctica para el cálculo es más cómodo utilizar tantos por uno.

Cuando los datos pueden ser ordenados, se define la frecuencia absoluta acumulada N_j del valor X_j como el número de observaciones que son menores o iguales a X_j. Se verifica la siguiente relación:

$$N_j = \sum_{k=1}^{j} n_k.$$

La frecuencia relativa acumulada F_j del valor X_j es el cociente entre N_j y n, que corresponde a la suma de las frecuencias relativas de los datos anteriores a X_j. Así podemos escribir

$$F_j = \sum_{k=1}^{j} f_k = \frac{N_j}{n}.$$

Todos los resultados anteriores se pueden presentar en forma de tabla, como por ejemplo la que sigue:

X_j	n_j	N_j	f_j	F_j
X_1	n_1	N_1	f_1	F_1
X_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_J	n_J	N_J=n	f_J	$F_J = 1$
Suma $\sum$	n		1	

Cuando los datos son continuos y de una precisión elevada (por ejemplo el tiempo en segundos, con una precisión hasta milisegundos, de una transmisión) o discretos con un número elevado de posibles valores (por ejemplo tamaño en bits de ficheros en un HD), se corre el riesgo de que las frecuencias resuman escasamente la muestra, es decir que las frecuencias absolutas de cada valor sean 1 o a lo más 2. En ambos casos se suele recurrir al conteo de datos por grupos o intervalos de valores a los que se denomina clases; es lo que se llama recuento de datos agrupados.

Consideremos el caso del peso en kilogramos de una persona. Cuando decimos yo peso 60 kilos ¿qué estoy diciendo en realidad? o si consideramos la edad y digo que tengo 21 años ¿qué estoy diciendo en realidad?

En la variable continua peso en kilos tenemos que los valores se calculan hasta las unidades, si estamos haciendo una medida en forma correcta decir

que pesamos 60 Kg. debería ser equivalente a decir que pesamos 60 ± 0.5 Kg, es decir nuestro instrumento de medida debería medir así; pues el error cometido será la mitad de la precisión del instrumento de medida. Lo mismo sucede con la edad; cometemos menos error si decimos que tenemos 21 años cuando tengamos 21 ± 0.5 años [1]. Así que 60 Kg. corresponde al intervalo $(59.5, 60.5)$ este tipo de extremos recibe el nombre de límites reales y puede abarcar más de un tipo de dato, por ejemplo el intervalo $(59, 5, 70, 5)$. Por contra tenemos los a veces llamados límites aparentes así podríamos definir el agrupamiento de 60 a 70 Kg.

De forma más general tenemos que si las observaciones vienen dadas con una precisión de una cifra decimal, los extremos reales de los intervalos serán de la forma #.#5, donde el símbolo # simboliza un dígito para la parte decimal y uno o varios para la entera.

Por ejemplo si nos dan los datos:

$$1.3, 3.6, 4.7, 4.9, 1.2, 0.6,$$

unos posibles intervalos de agrupamiento con límites reales y de amplitud 1 son:

$$[0.55, 1.55),$$
$$[1.55, 2.55),$$
$$[2.55, 3.55),$$
$$[3.55, 4.55),$$
$$[4.55, 5.55).$$

Si los valores vienen dados con dos cifras decimales de precisión, los extremos de los intervalos serían de la forma #.##5.. Por ejemplo, si los datos son:

$$0.23, 1.26, 3.54, 5.76, 8.76, 3.67,$$

unos posibles intervalos con límites reales de amplitud 2 son:

$$[0.225, 2.225),$$
$$[2.225, 4.225),$$
$$[4.225, 6.225),$$
$$[6.225, 8.225),$$
$$[8.225, 10.225).$$

[1] Es evidente que las personas no hacemos esto y que decimos que tenemos 21 años hasta el día anterior a nuestro aniversario, con lo cual durante la mitad del año, de cada año de nuestra vida, estamos cometiendo un error superior a medio año.

Para escoger el primer extremo se suele calcular el mínimo de la muestra y se toma como valor mínimo el extremo inferior del límite real de ese valor.

En el primer ejemplo, el valor mínimo era 0.6; por lo tanto el primer extremo es $0.6 - 0.05 = 0.55$. En el segundo ejemplo, el valor mínimo es 0.23; por lo tanto, el primer extremo será $0.23 - 0.005 = 0.225$.

Los otros extremos se obtienen sumando una amplitud, de la misma precisión que los datos, desde el valor mínimo.

Como receta general, que no es de obligado cumplimiento, a la hora de agrupar se recomienda:

I) Decidir el número de clases a considerar. Este número no debe ser inferior a 5 y como máximo entre 15 y 20. Se pueden utilizar las siguientes heurísticas, si J es el número de clases tomar $J \geq \sqrt{n}$ (para tamaños muestrales inferiores a 150) o también $2^J \geq n$.

II) Seleccionar los límites de clase que definen los intervalos, de forma que, si es posible, todos tengan la misma amplitud, salvo quizás los extremos.

III) Intentar no dejar clases con frecuencias muy bajas, para evitar esto se pueden unir estas clases a una de sus adyacentes.

A cada clase o agrupamiento se le asigna ahora un valor representativo que recibe el nombre de marca de clase. Se suele tomar, salvo que se diga lo contrario, como marca de clase el punto medio de un intervalo; que se obtiene dividiendo por dos la amplitud del mismo.

En el primer ejemplo las marcas de clase son:

$[0.55, 1.55)$	1.05
$[1.55, 2.55)$	2.05
$[2.55, 3.55)$	3.05
$[3.55, 4.55)$	4.05
$[4.55, 5.55)$	5.05

mientras que para el segundo son estas:

$[0.225, 2.225)$	1.225
$[2.225, 4.225)$	3.225
$[4.225, 6.225)$	5.225
$[6.225, 8.225)$	7.225
$[8.225, 10.225)$	9.225

La tabla final es:

intervalos	(Marca de clase) X_j	n_j	N_j	f_j	F_j
$[L_1, L_2)$	X_1	n_1	N_1	f_1	F_1
$[L_2, L_3)$	X_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_J, L_{J+1})$	X_I	n_I	N_I	f_I	F_I
Suma $\sum$		n		1	

Ejemplo 1 *Consideremos las puntuaciones de* 50 *aspirantes a un puesto de trabajo:*

8	11	11	8	9	10	16	6	12	19
13	6	9	13	15	9	12	16	8	7
14	11	15	6	14	14	17	11	6	9
10	19	12	11	12	6	15	16	16	12
13	12	12	8	17	13	7	12	14	12

La tabla de frecuencias agrupadas con límites reales y amplitud fija de los intervalos 3 *es:*

intervalos	X_j	n_j	N_j	f_j	F_j
$[5.5, 8.5)$	*7*	*11*	*11*	*0.22*	*0.22*
$[8.5, 11.5)$	*10*	*11*	*22*	*0.22*	*0.44*
$[11.5, 14.5)$	*13*	*17*	*39*	*0.34*	*0.78*
$[14.5, 17.5)$	*16*	*9*	*48*	*0.18*	*0.96*
$[17.5, 20.5)$	*19*	*2*	*50*	*0.04*	*1.00*

1.3.2. Descripción gráfica

La representación gráfica de los datos cuantitativos discretos se hace mediante diagramas de barras.

El gráfico 1.3 nos muestra el diagrama de barras de las frecuencias absolutas y absolutas acumuladas para variables discretas. Las frecuencias absolutas n_i son las alturas de las barras con base el punto X_i. Las frecuencias absolutas acumuladas N_i son también las alturas de las barras con base el punto X_i.

La descripción gráfica de los datos continuos (agrupados) se hace mediante histogramas. En la figura 1.4 tenemos un ejemplo de histograma. En

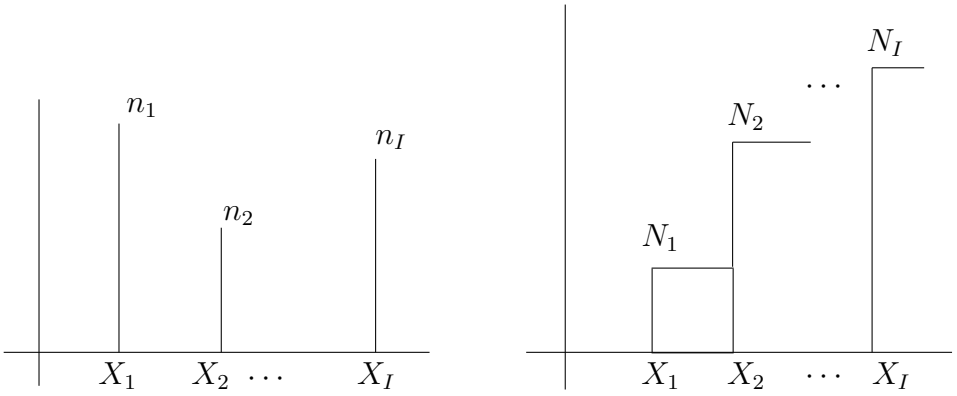

Figura 1.3: Frecuencias absolutas. Variables discretas

este caso es el histograma de las frecuencias absolutas a la izquierda y a la derecha tenemos el gráfico de las frecuencias absolutas acumuladas. Las frecuencias absolutas n_j del gráfico de la izquierda representan las àreas de los rectángulos de la base $L_{j+1} - L_j$ (amplitud del intérvalo de clase) mientras que las frecuencias absolutas acumuladas N_j del gráfico de la derecha representan las alturas de los rectángulos de base $L_{i+1} - L_i$. La curva que une los pares ordenados (X_j, h_j) se llama polígono de frecuencias absolutas (léase igual para relativas), mientras que el polígono de frecuencias absolutas acumuladas (de forma similar para relativas) es el formado por los puntos $(L_1, 0), (L_2, N_1), \ldots, (L_{J+1}, N_J)$.

Ejemplo 2 *Consideremos las puntuaciones de los* 50 *aspirantes del ejemplo 1. Tomamos intervalos de amplitud* 3. *El histograma de frecuencias absolutas con el correspondientes polígono de frecuencias acumuladas se muestra en la figura 1.5.*

Notemos que las alturas de los rectángulos se calculan teniendo en cuenta que la amplitud de los intervalos es 3*:*

$$h_1 = \frac{n_1}{3} = \frac{11}{3} = 3.6666, \qquad h_2 = \frac{n_2}{3} = \frac{11}{3} = 3.666,$$

$$h_3 = \frac{n_3}{3} = \frac{17}{3} = 5.6666, \qquad h_4 = \frac{n_4}{3} = \frac{9}{3} = 3,$$

$$h_5 = \frac{n_5}{3} = \frac{2}{3} = 0.666.$$

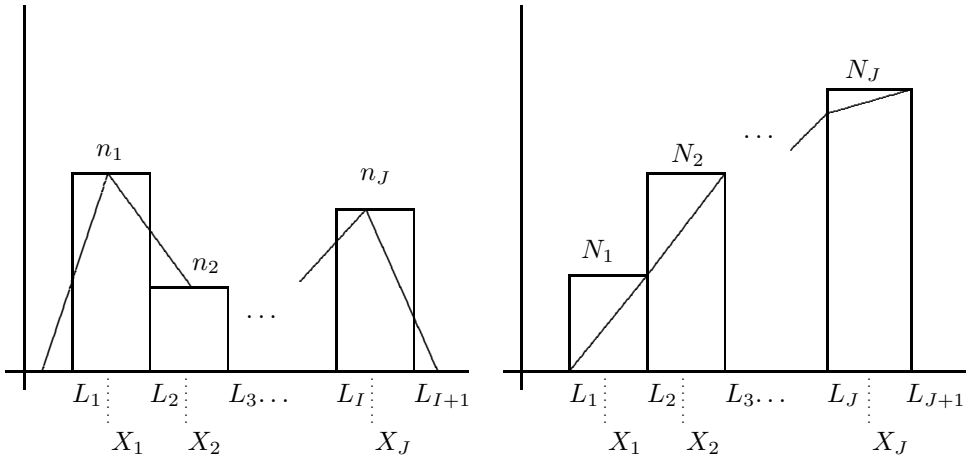

Figura 1.4: Frecuencias absolutas. Variables continuas

1.4. Análisis de las distribuciones

El análisis de las distribuciones de una variable o dato estadístico consiste en reducir los datos estadísticos a unas pocas medidas o índices, que reciben el nombre de estadísticos, que nos permitan una interpretación de las regularidades de todo el colectivo.

Tenemos los siguientes tipos de medidas:

- Medidas de posición: Intentan representar toda la distribución. Las más importantes son la media aritmética, la moda y la mediana.
- Medidas de dispersión: Intentan señalar la dispersión o separación del conjunto de datos respecto a las medidas de posición adoptadas. Las más importantes son la varianza, la desviación típica, el coeficiente de variación y los recorridos.
- Medidas de simetría y apuntamiento: Estudian si el polígono de frecuencias relativas es simétrico y lo *estirado* que está (apuntamiento). Se suele comparar este polígono con la curva de frecuencias de una distribución ideal llamada normal o campana de Gauss.
- Otras como las medidas de concentración; que no veremos. Por ejemplo el índice de Gini.

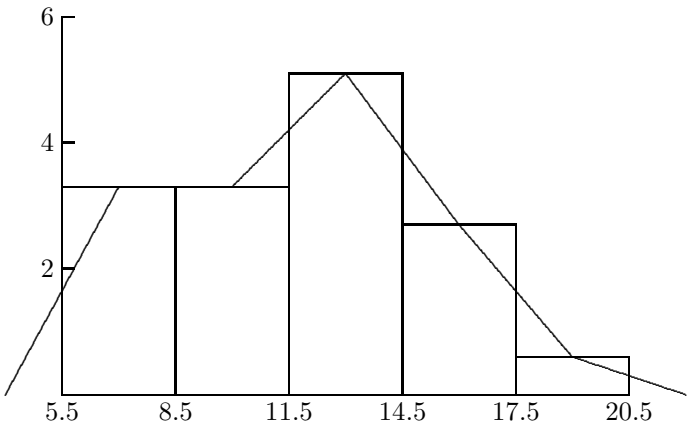

Figura 1.5: Histograma de frecuencias absolutas (ejemplo 1)

Nota importante: Algunas de estas medidas sólo se pueden calcular cuando tenga sentido operar con los datos, es decir, si estos son cantidades o al menos órdenes. Si tengo que una variable que responde al deporte que practica una persona de determinada población, aunque la variable esté codificada a valores enteros, no tiene sentido hacer la media aritmética. En lo que sigue dejaremos al lector que decida, siempre de forma razonada, que estadísticos no son aplicables a estas variables.

1.4.1. Medidas de posición

Veremos aquí las más conocidas medidas de posición.

Media aritmética

La media aritmética es la medida de tendencia central más utilizada, simboliza el valor central de toda la distribución. Su fórmula general es[2]:

$$overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{j=1}^{J} n_j X_j}{n} = \sum_{j=1}^{J} f_j X_j.$$

[2]Como se ve damos dos fórmulas una para datos no agrupados y otra para datos posiblemente agrupados, en lo que sigue no especificaremos cuales son las fórmulas para datos agrupados o no.

Para el caso de distribuciones de variables discretas, los X_j son los posibles valores de la variable mientras que en el caso continuo, son las marcas de clase de los intervalos.

Una de las propiedades fundamentales de la media es que si hacemos una transformación lineal de los datos digamos $Y = aX + b$[3] donde X son los valores de la variable, la relación entre la media aritmética de Y y la de X es:

$$\overline{y} = a\overline{x} + b.$$

Medias armónica y geométrica

Las medias armónica y geométrica no son de gran utilidad salvo en problemas concretos. Se calculan de la siguiente forma:

$$M_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\sum_{j=1}^{I} \frac{n_j}{X_j}}, \quad M_g = \sqrt[n]{\prod_{i=1}^{n} x_j} = \sqrt[n]{\prod_{j=1}^{I} X_j^{n_j}}.$$

Estas medias tienen restricciones sobre los datos, no pueden tener datos nulos, y en general se utilizan para datos positivos.

Ejemplo 3 *Consideremos los siguientes datos:*

10	5	2	7	9	5	7	6	5	9
12	2	6	6	9	12	6	6	6	4
9	7	12	11						

La media aritmética de los datos anteriores sin agrupar en intervalos es:

$$\overline{x} = \frac{10 + 5 + 2 + \cdots + 12 + 11}{24} = \frac{173}{24} = 7.20833$$

Si los agrupamos en intervalos de amplitud 3*, la media será (hacemos*

[3]Multiplicar por una constante positiva se suele denominar cambio de escala. Sumar una constante a una varible recibe el nombre de cambio de origen. Así podemos decir que la media arimética queda igual de afectada por los cambios de escala y origen en los datos.

primero la correspondiente tabla de frecuencias)

intervalos	X_j	n_j	n_jX_j
$[1.5, 4.5)$	3	3	9
$[4.5, 7.5)$	6	12	72
$[7.5, 10.5)$	9	5	45
$[10.5, 13.5)$	12	4	48
Suma		24	174

$$\overline{x} = \frac{174}{24} = 7.25$$

Notemos que los valores difieren ya que el agrupamiento provoca una pérdida de información.

Media general de orden m

Definimos la media general $M_{(m)}$ de orden m como:

$$M_{(m)} = \left(\frac{\sum_{i=1}^{n} n_i x_i^m}{n}\right)^{\frac{1}{m}} = \left(\frac{\sum_{j=1}^{J} n_j X_j^m}{n}\right)^{\frac{1}{m}}$$

Se cumple que:

$$M_{(-1)} = M_h;\ M_{(0)} = M_g;\ M_{(1)} = \overline{x}$$

Además se cumple que $M_{(m)}$ es una función creciente en m y por lo tanto :

$$M_h \leq M_g \leq \overline{x}.$$

Mediana y percentiles

La mediana es aquel valor que, cuando consideremos todos los valores de la muestra ordenados, ocupa el lugar central. Es decir, quedan la misma cantidad de valores a su izquierda que a su derecha.

Supongamos que los datos ordenados son $x_1, x_2, \ldots x_n$, la **mediana** vale:

$$\begin{cases} x_{\frac{n+1}{2}}, & \text{si } n \text{ es impar,} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{si } n \text{ es par.} \end{cases}$$

Por ejemplo, la mediana de los datos $1, 2, 5, 8, 8, 9, 11$ vale 8 ya que 8 es el que ocupa el lugar central y la mediana de $2, 3, 3, 4, 5, 6$ vale $\frac{3+4}{2} = 3.5$.

La manera anterior de calcular la mediana no es práctica en el caso en que haya muchos datos es muy costoso ordenarlos (orden n^2 o $n\log(n)$). Veamos alguna manera de cálculo aproximado de la mediana a partir de la tabla de distribución de frecuencias.

Necesitaremos las columnas de frecuencias absolutas y la de frecuencias absolutas acumuladas:

intervalos	X_j	n_j	N_j
$[L_1, L_2)$	X_1	n_1	N_1
$[L_2, L_3)$	X_2	n_2	N_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_I, L_{I+1})$	X_I	n_I	N_I
$\sum$		n	

Llamaremos intervalo crítico para la mediana al primer intervalo en el que su frecuencia absoluta acumulada supere o iguale a $\frac{n}{2}$. Denotemos por $[L_c, L_{c+1})$ el intervalo crítico. Sea N_{c-1} la frecuencia absoluta acumulada del intervalo anterior al crítico. En el caso en que el interval crítico sea el primero, $N_{c-1} = 0$. Sea n_c la frecuencia absoluta del intervalo crítico. Sea $A_c = L_{c+1} - L_c$ la amplitud del intervalo crítico. Calcularemos la **mediana** mediante:

$$M = L_c + A\frac{\left(\frac{n}{2} - N_{c-1}\right)}{n_c}.$$

La justificación de la fórmula anterior es la siguiente: si representásemos las frecuencias absolutas acumuladas entre los extremos de los intervalos, la mediana seria la antiimagen de $\frac{n}{2}$ en el intervalo crítico haciendo una interpolación por rectas (ver figura 1.6).

Los percentiles son una generalización de la mediana. La mediana es el percentil 50 ya que deja el 50 % de las observaciones a su izquierda.

En general el **percentil** P es aquel valor que deja el P % de las observaciones a su izquierda. El cálculo, dada la distribución de frecuencias es semejante al cálculo de la mediana.

Definimos el intervalo crítico en este caso como el primer intervalo del que su frecuencia absoluta acumulada supera o iguala a $\frac{n \cdot P}{100}$.

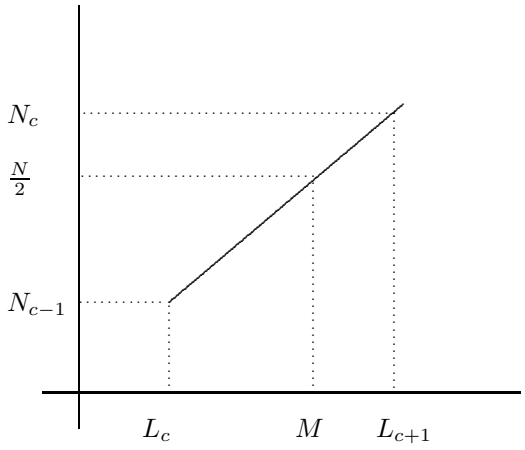

Figura 1.6: Interpretación geométrica de la Mediana

Sean entonces, $[L_c, L_{c+1})$ el intervalo crítico, N_{c-1} la frecuencia absoluta acumulada del intervalo anterior al crítico y n_c la frecuencia absoluta del intervalo crítico. Si denominamos por A_c a la amplitud del intervalo crítico, la fórmula para calcular el **percentil** P es:

$$M_p = L_c + A_c \frac{\left(\frac{n \cdot P}{100} - N_{c-1}\right)}{n_c}.$$

Ejemplo 4 *Calculemos la mediana, sin agrupar, de los siguientes datos:*

$$14, 15, 16, 18, 18, 18, 18, 19, 20, 20, 22.$$

El tamaño de la muestra es $n = 11$ *observaciones y ya están ordenadas. El lugar central, es el que ocupa el sexto puesto, el valor que ocupa este lugar es el* 18*, por lo tanto, la mediana es* 18*.*

En la siguiente muestra tenemos un número par de datos:

$$24, 25, 26, 26, 27, 27, 27, 29.$$

El tamaño muestral es $n = 8$ *observaciones que ya están ordenadas. El lugar central estará entre el cuarto y el quinto puesto. Los datos que ocupan estos lugares son el* 26 *y el* 27*. Por lo tanto la mediana vale*

$$M = \frac{26 + 27}{2} = 26.5.$$

Ejemplo 5 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	N_j
$[1.5, 4.5)$	3	3	3
$[4.5, 7.5)$	6	12	15
$[7.5, 10.5)$	9	5	20
$[10.5, 13.5)$	12	4	24

Tenemos que $n = 24$ *y que* $\frac{n}{2} = 12$. *El intervalo crítico es:* $[4.5, 7.5)$ *La mediana valdrá entonces:*

$$M = 4.5 + 3\frac{(12-3)}{12} = 6.75.$$

Percentil 25*:* $25\,\% \Rightarrow \frac{n \cdot P}{100} = 6$. *Intervalo crítico:* $[4.5, 7.5)$.

$$M_{25} = 4.5 + 3\frac{(6-3)}{12} = 5.25$$

Percentil 75*:* $75\,\% \Rightarrow \frac{n \cdot P}{100} = 18$. *Intervalo crítico:* $[7.5, 10.5)$.

$$M_{75} = 7.5 + 3\frac{(18-15)}{5} = 9.3$$

En general se habla de cuantiles para denominar a todos estos estadísticos. Los cuartiles que dividen a la población en cuartos son llamados cuartiles, así el primer cuartil Q_1 deja a su izquierda el $25\,\%$ de las observaciones, el segundo cuartil Q_2 es la mediana y el tercer cuartil Q_3 deja a su izquierda el $75\,\%$ de las observaciones. También se habla de los deciles que son los estadísticos que dividen a la población en décimas partes.

Moda

La moda de una muestra es un valor que tenga la frecuencia absoluta más grande. En consecuencia la moda no tiene por qué ser única puede haber más de un valor con frecuencia absoluta máxima. Si una distribución tiene una sola moda diremos que es unimodal, si dos bimodal, ... La presencia de dos modas puede indicar la existencia de dos poblaciones diferenciadas en la muestra (por ejemplo el peso según sexo).

En el caso en que tengamos una tabla de distribuciones, para encontrar la moda, hemos de localizar el intervalo o intervalos con frecuencia absoluta más alta.

Sean $[L_j, L_{j+1})$ los extremos del intervalo con frecuencia absoluta máxima.

Para calcular la moda podemos utilizar la siguiente fórmula, en la que suponemos que todos los intervalos tienen la misma amplitud A (en caso contrario se utilizan otras aproximaciones):

$$M_o = L_j + A\frac{n_{j+1}}{(n_{j-1} + n_{j+1})}.$$

Donde:

- A: amplitud de los intervalos
- n_{j-1}: frecuencia absoluta del intervalo anterior al de frecuencia máxima.
- n_{j+1}: frecuencia absoluta del intervalo posterior al de frecuencia máxima.

Ejemplo 6 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	N_j
$[1.5, 4.5)$	3	3	3
$[4.5, 7.5)$	6	12	15
$[7.5, 10.5)$	9	5	20
$[10.5, 13.5)$	12	4	24

El intervalo con la frecuencia absoluta mas alta es el $[4.5, 7.5)$. *Por lo tanto, la moda vale:*

$$M_0 = 4.5 + 3\frac{5}{(3+5)} = 6.375.$$

1.4.2. Medidas de dispersión

Una vez estudiadas las medidas de posición, vamos a estudiar algunos estadísticos que miden lo separadas que están las observaciones entre sí.

Algunas medidas de dispersión respecto a la media aritmética son la varianza, la desviación típica, la desviación media respecto de la media y el coeficiente de variación.

Las medidas de dispersión respecto a la a la mediana es la desviación media respecto de la mediana.

Otras medidas de dispersión son el recorrido, el rango , el recorrido intercuartílico, el rango intercuartílico.

Varianza y desviación típica o estándar

La varianza y la desviación típica nos indican si los datos están muy dispersos respecto de la media aritmética $\overline{x}$.

La fórmula del cálculo de la varianza es:

$$s^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(X_j - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{J} n_j X_j^2 - \overline{x}^2.$$

o bien para datos sin agrupar

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{n} x_i^2 - \overline{x}^2.$$

La segunda expresión es más útil que la primera de cara al cálculo de la varianza.

La propiedad fundamental de la varianza es que minimiza las desviaciones al cuadrado respecto a cualquier punto X_0. Es decir:

$$\min_{X_0} \frac{1}{n}\sum_{j=1}^{J} n_j(X_j - X_0)^2 = s^2.$$

La desviación típica o estándar es la raíz cuadrada positiva de la varianza:

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{J} n_j X_j^2 - \overline{x}^2}.$$

Por motivos que veremos en temas posteriores existe otra fórmula para el cálculo de la varianza de una muestra a la que en ocasiones se le denomina cuasivarianza o también se le llama varianza muestral [4]:

$$\tilde{s}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{j=1}^{n} n_j(X_j - \overline{X})^2.$$

Notemos que la cuasivarianza es una pequeña corrección de la varianza, en lugar de dividir por el tamaño muestral se divide por el tamaño muestral

[4] Quizá algunos de vosotros descubra aquí el motivo por el que las calculadoras llevan dos teclas s_n^2 o σ_n^2 y s_{n-1} o σ_{n-1}.

menos 1. Para muestras grandes la corrección puede resultar insignificante pero para muestras pequeñas es necesaria.

Cuando la variable X se vea afectada por un cambio lineal: $Y = aX + b$, la varianza de Y cumple la siguiente relación:

$$s_Y^2 = a^2 s_X^2.$$

De aquí deducimos que la varianza es independiente respecto a cambios de origen y que queda afectada por el cuadrado de los cambios de escala.

Para las desviaciones típicas tendremos:

$$s_Y = |a| s_X.$$

Ejemplo 7 *Consideremos la siguiente distribución de frecuencias*

intervalos	X_j	n_j	$n_j X_j$
$[9.5, 29.5)$	19.5	38	741.0
$[29.5, 49.5)$	39.5	18	711.0
$[49.5, 69.5)$	59.5	31	1844.5
$[69.5, 89.5)$	79.5	20	1590.0
Sumas		107	4886.5

Vamos a calcular la varianza

Primero calculamos la media:

$$\overline{x} = \frac{4886.5}{107} = 45.6682$$

Para calcular la varianza hemos de añadir dos columnas a la tabla anterior:

X_j	X_j^2	$n_j X_j^2$
19.5	380.25	14449.50
39.5	1560.25	28084.50
59.5	3540.25	109747.75
79.5	6320.25	126405.00
Suma		278686.75

La varianza y la desviación típica valen:

$$s_X^2 = \frac{278686.75}{107} - 45.6682^2 = 518.962$$

$$s_X = \sqrt{518.962} = 22.7807$$

Coeficiente de variación

El coeficiente de variación se define como el cociente entre la desviación típica y la media aritmética, se utiliza para variables en las que la media represente a la magnitud de los datos (por ejemplo si todos son positivos y la distribución es unimodal) :

$$CV = \frac{s}{\overline{x}}.$$

El coeficiente de variación es independiente del cambio de escala. Más concretamente, si hacemos el cambio lineal de la variable X: $Y = aX$, con $a > 0$, el coeficiente de variación de la variable Y es el mismo que el de la variable X:

$$CV_Y = CV_X.$$

El coeficiente de variación será útil para comparar la dispersión de distribuciones medidas en diferentes escalas.

Ejemplo 8 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	$n_j X_j$
$[9.5, 29.5)$	19.5	38	741.0
$[29.5, 49.5)$	39.5	18	711.0
$[49.5, 69.5)$	59.5	31	1844.5
$[69.5, 89.5)$	79.5	20	1590.0
Sumas		107	4886.5

La media y la desviación típica son:

$$\overline{x} = 45.6682, \quad s_X = 22.7807$$

Por lo tanto el coeficiente de variación es:

$$CV = \frac{s}{\overline{x}} = \frac{22.6807}{45.6682} = 0.4988$$

Desviación media

La desviación media es un índice de dispersión respecto a la mediana o a la media. Queda definido por:

$$D_M = \frac{1}{n}\sum_{j=1}^{J} n_j |X_j - M|.$$

donde M es la mediana o la media aritmética.

La propiedad fundamental de la desviación media respecto a la mediana es que minimiza las desviaciones en valor absoluto respecto de un punto cualquiera X_0. Es decir:

$$\min_{X_0} \frac{1}{n}\sum_{j=1}^{J} n_j |X_j - X_0| = D_M.$$

Ejemplo 9 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	$n_j X_j$
$[0.5, 15.5)$	8	4	32
$[15.5, 30.5)$	23	4	92
$[30.5, 45.5)$	38	2	76
Sumas		10	200

Vamos a calcular la desviación media:

Calculamos la media

$$\overline{x} = \frac{200}{10} = 20$$

Añadimos dos columnas más a la tabla de frecuencias:

X_j	$\lvert X_j - \overline{x}\rvert$	$n_j \lvert X_j - \overline{x}\rvert$
8	*12*	*48*
23	*3*	*12*
38	*18*	*36*
Sumas		*96*

La desviación es:

$$D_M = \frac{96}{10} = 9.6$$

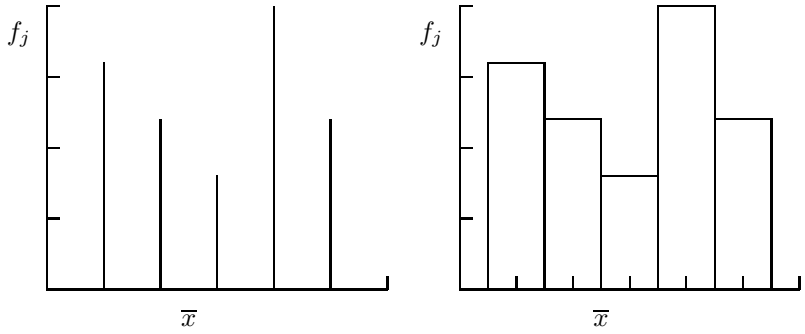

Figura 1.7: Diagrama de barras e histograma de las frecuencias relativas

Recorrido

Otra medida de dispersión es el recorrido. Se define como la diferencia entre el valor máximo y mínimo de los valores observados.

También se utiliza el recorrido intercuartílico que es $Q_3 - Q_1$; la diferencia entre el tercer y primer cuartil. También se pueden calcular recorridos con deciles, percentiles y cuantiles en general.

1.4.3. Perfil de una distribución

El perfil de una distribución viene determinado por alguno de sus polígonos de frecuencias. Es mejor utilizar las frecuencias relativas ya que no dependen del tamaño de la muestra (ver figura 1.7). La idea es encontrar la curva a donde tiende el polígono de frecuencias cuando la muestra se hace grande, que en definitiva sería la curva de frecuencias de toda la población.

Una curva continua en forma de campana llamada curva de Gauss [5] puede servir como un modelo matemático ideal para comparar el perfil de cualquier distribución. Esta curva corresponde a la gráfica de la función:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

donde μ se aproxima por $\overline{x}$ y σ por s. Su representación gráfica es la de la figura 1.8, gaussiana o campana de gauss, para el caso (estándar) en el que $\mu = 0$ y $\sigma = 1$.

[5] Es una buena broma pedir a un amigo el cálculo de la primitiva de la curva de Gauss.

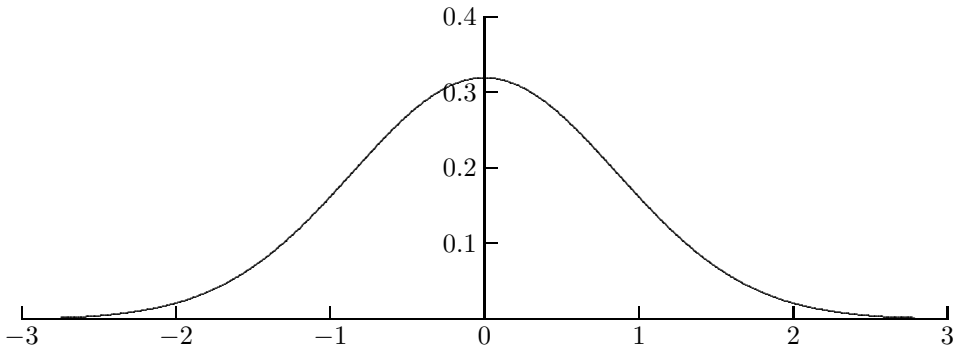

Figura 1.8: Curva normal o campana de Gauss

Las propiedades más importantes de la curva normal son:

a) Está definida para cualquier real y es siempre positiva.

b) El área comprendida entre la curva y el eje de abcisas vale siempre 1 para cualquier valor de μ y $\sigma > 0$.

c) Es simétrica respecto a la recta vertical $X = \mu$ y en este punto tiene un máximo absoluto que vale $\frac{1}{\sqrt{2\pi}\sigma}$.

d) Tiene dos puntos de inflexión en $x = \mu \pm \sigma$.

e) El eje de abcisas es una asíntota de la curva.

Las medidas de simetría y apuntamiento se suelen referir a la correspondiente distribución normal; aquella en la que los parámetros se estiman por $\mu = \overline{x}$ y $\sigma = s$. (o por la cuasivarianza).

Se entiende, entonces, que la distribución normal es simétrica y es perfecta respecto al apuntamiento. Es decir, que no es ni apuntada ni chata.

1.4.4. Medidas de simetría

Para ver si una distribución es simétrica o asimétrica por la derecha o por la izquierda se toma como índice de simetría:

$$g_1 = \frac{m_3}{s^3},$$

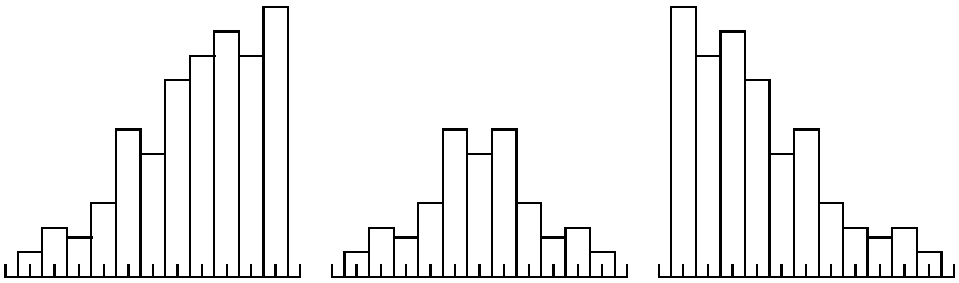

Figura 1.9: Histogramas con diferentes tipos de simetría. Izquierda: $g_1 < 0$, centro: $g_1 = 0$, derecha: $g_1 > 0$.

donde m_3 es el momento central de tercer orden y se calcula de la siguiente forma:

$$m_3 = \frac{1}{n}\sum_{j=1}^{J} n_j(X_j - \overline{x})^3,$$

y s es la desviación típica. Tenemos, entonces que:

- Si $g_1 > 0$, la distribución es asimétrica por la derecha o tiene asimetría positiva.
- Si $g_1 = 0$, la distribución es simétrica o el índice no decide.
- Si $g_1 < 0$, la distribución es asimétrica por la izquierda o tiene asimetría negativa.

Ejemplo 10 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	n_jX_j	$n_jX_j^2$
$[14.5, 19.5)$	17	4	68	1156
$[19.5, 24.5)$	22	6	132	2904
$[24.5, 29.5)$	27	8	216	5832
$[29.5, 34.5)$	32	11	352	11264
$[34.5, 39.5)$	37	35	1295	47915
$[39.5, 44.5)$	42	100	4200	176400
$[44.5, 49.5)$	47	218	10246	481562
Sumas		382	16509	727033

La media y la varianza valen:

$$\overline{x} = \frac{16509}{382} = 43.2173, \quad s_X^2 = \frac{727033}{382} - \left(\frac{16509}{382}\right)^2 = 35.49$$

Calculemos el coeficiente de asimetría g_1. Para hacerlo, hemos de añadir una columna más a la tabla anterior:

X_j	n_j	$n_j(X_j - \overline{x})^3$
17	4	-72081.33
22	6	-57308.66
27	8	-34121.16
32	11	-15525.84
37	35	-8411.41
42	100	-180.37
47	218	11799.67
Sumas	382	-175829.09

El momento de tercer orden vale:

$$m_3 = \frac{-175829.09}{382} = -460.285$$

A continuación calculamos el índice de asimetría:

$$g_1 = \frac{m_3}{s^3} = \frac{-460.285}{\left(\sqrt{35.49}\right)^3} = -2.18$$

Por lo tanto podemos decir que se trata de una distribución asimétrica por la izquierda o negativa

El índice de simetría es independiente de cambios lineales de la forma $Y = aX + b$, con $a > 0$, es decir:

$$g_1(X) = g_1(Y).$$

En otras palabras, el índice de simetría no queda afectado por cambios de origen, ni por cambios de escala positivos, mientras que para cambios de escala negativos cambia el signo de la simetría (ejercicio).

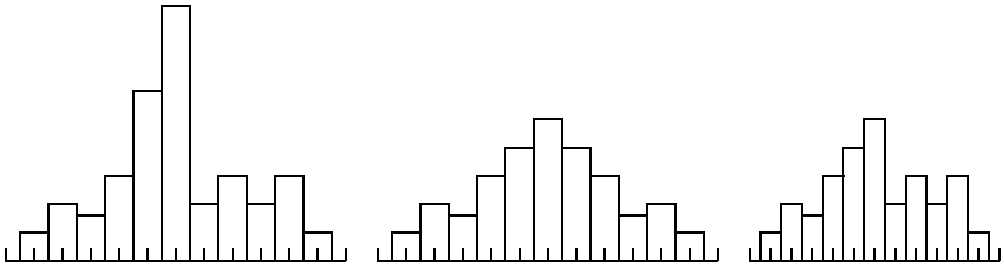

Figura 1.10: Histogramas de los tres tipos de apuntamiento

1.4.5. Medidas de apuntamiento

Las medidas de apuntamiento nos miden si el perfil de una distribución muestral está muy apuntado o no en comparación con un perfil ideal, como por ejemplo el de la campana de gauss asociada. Para estudiar el apuntamiento se utiliza un índice basado en el momento de cuarto orden, que recibe el nombre de coeficiente de apuntamiento o curtosis[6]:

$$g_2 = \frac{m_4}{s^4} - 3,$$

donde m_4 es el llamado momento central de cuarto orden y se calcula de la siguiente forma:

$$m_4 = \frac{1}{n}\sum_{j=1}^{J} n_j (X_j - \overline{x})^4,$$

y s es la desviación típica. Tenemos, pues que:

- Si $g_2 > 0$, la distribución es puntiaguda o leptocúrtica.
- Si $g_2 = 0$, la distribución es similar a la normal o mesocúrtica.
- Si $g_2 < 0$, la distribución es achatada o platicúrtica.

[6] En inglés *kurtosis*.

Ejemplo 11 *Consideremos la siguiente distribución de frecuencias:*

intervalos	X_j	n_j	n_jX_j	$n_jX_j^2$
$[14.5, 19.5)$	17	4	68	1156
$[19.5, 24.5)$	22	6	132	2904
$[24.5, 29.5)$	27	8	216	5832
$[29.5, 34.5)$	32	11	352	11264
$[34.5, 39.5)$	37	35	1295	47915
$[39.5, 44.5)$	42	100	4200	176400
$[44.5, 49.5)$	47	218	10246	481562
Sumas		382	16509	727033

La media y la varianza valen:

$$\overline{x} = \frac{16509}{382} = 43.22, \quad s_X^2 = \frac{727033}{382} - \left(\frac{16509}{382}\right)^2 = 35.49$$

El coeficiente de apuntamiento g_2. Hemos de añadir una columna a la tabla:

intervalos	X_j	n_j	$n_j(X_j - \overline{x})^4$
$[14.5, 19.5)$	17	4	1889776.11
$[19.5, 24.5)$	22	6	1215933.65
$[24.5, 29.5)$	27	8	553352.38
$[29.5, 34.5)$	32	11	174157.64
$[34.5, 39.5)$	37	35	52296.07
$[39.5, 44.5)$	42	100	219.56
$[44.5, 49.5)$	47	218	44634.88
Sumas		382	3930370.29

El momento de cuarto orden vale:

$$m_4 = \frac{3930370.29}{382} = 10288.93$$

A continuación, calculamos el índice de apuntamiento

$$g_2 = \frac{m_4}{s^4} - 3 = \frac{10288.93}{35.49^2} - 3 = 5.17$$

Por lo tanto se trata de una distribución puntiaguda o leptocúrtica.

El índice de apuntamiento es independiente respecto cambios lineales de la forma $Y = aX + b$, es decir:

$$g_2(X) = g_2(Y).$$

El índice g_2 no queda afectado por cambios de origen ni de escala.

1.5. Variables multidimensionales

Hasta ahora sólo hemos estudiado una variable, es evidente que en la realidad interesa el comportamiento conjunto de dos o más variables. En cualquier disciplina técnica o científica, economía, ciencias de la computación, bioinformática, telecomunicaciones,... son muy utilizados los conceptos de asociación, independencia y otros, entre dos o más variables. Para introducirlos estudiaremos el caso más sencillo; el de las variables estadísticas bidimensionales. En lo que respecta a esta sección cada individuo de la población tiene asociado más de un valor o cualidad observada. Per ejemplo peso y altura de un grupo de personas, peso y sexo, altura y nivel de estudios, ... Por ejemplo si estudiamos el peso (p) y la altura (h) de una población una muestra genérica de tamaño n tendría el siguiente aspecto:

$$(p_1, h_1), (p_2, h_2), \ldots, (p_n, h_n),$$

donde (p_i, h_i) es el peso y la estatura correspondientes a la observación i-ésima.

Otro ejemplo sería el estudio de la relación entre los turistas llegados a nuestra isla y el año de llegada. Los datos serían:

$$(t_1, n_1), (t_2, n_2), \ldots, (t_N, n_N),$$

donde t_i es el año i-ésimo y n_i = número de turistas llegados ese año.

1.5.1. Descripción numérica: caso bidimensional

Supongamos que tenemos (X, Y) un par de variables que se pueden medir conjuntamente en un individuo de la población que se desea estudiar.

Supondremos que las variables son discretas.

Sean $\{X_1, X_2, \ldots, X_I\}$ los valores posibles de X y $\{Y_1, Y_2, \ldots, Y_J\}$ los de Y.

El conjunto de valores que puede tomar la variable conjunta (X, Y) son:

$$\{(X_1, Y_1), \ldots, (X_1, Y_J), (X_2, Y_1) \ldots, (X_2, Y_J), \ldots, (X_I, Y_1), \ldots, (X_I, Y_J)\}.$$

Sean n_{ij} la frecuencia absoluta correspondiente al valor (X_i, Y_j), o sea, es el nombre de individuos de la muestra que tienen la variable X igual a X_i y la variable Y igual a Y_j.

Toda esta información se puede resumir en la siguiente tabla de frecuencias absolutas o tabla de contingencia:

X/Y	Y_1	Y_2	$\ldots$	Y_j	$\ldots$	Y_J	$n_{i\bullet}$
X_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1J}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_I	n_{I1}	n_{I2}	$\ldots$	n_{Ij}	$\ldots$	n_{IJ}	$n_{I\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet J}$	$N = n_{\bullet\bullet}$

En la tabla anterior, los valores de $n_{i\bullet}$ representan el número de individuos con $X = X_i$, $n_{\bullet j}$ el número de individuos con $Y = Y_j$ y n es el nombre total de individuos.

Ejemplo 12 *Consideremos la siguiente muestra de tamaño* 12 *de dos características conjuntas; la edad y peso de unas personas:*

$$\begin{array}{cccc} (20,75) & (20,75) & (30,75) & (40,85) \\ (30,65) & (20,75) & (40,85) & (30,65) \\ (20,65) & (40,75) & (30,65) & (20,75) \end{array}$$

La variable X es "edad" y toma los valores $\{20, 30, 40\}$ y la variable Y es "peso" y toma los valores $\{65, 75, 85\}$.

La tabla de frecuencias será:

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

En el caso en que las variables X y Y sean agrupadas se hace una tabla semejante al caso discreto. En este caso, los datos de las dos variables X e Y se agrupan en intervalos de clase. La tabla de valores queda como se muestra a continuación:

X/Y	intervalos	$[L'_0, L'_1)\cdots[L'_{j-1}, L'_j)\cdots[L'_{J-1}, L'_J)$					
intervalos	M. Clase	c'_1	$\cdots$	c'_j	$\cdots$	c'_J	$n_{i\bullet}$
$[L_0, L_1)$	c_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
$[L_1, L_2)$	c_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{i-1}, L_i)$	c_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{I-1}, L_I)$	c_I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
	$n_{\bullet j}$	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

En la tabla anterior las c_i son las marcas de clase correspondientes a los intervalos de la variable X y las c'_j son las marcas de clase correspondientes a los intervalos de la variable Y.

Ejemplo 13 *Consideremos la siguiente tabla que nos da el peso y la estatura de* 15 *individuos:*

Individuo	X=peso	Y=estatura
1	65	1.6
2	62	1.6
3	71	1.6
4	72	1.7
5	75	1.8
6	80	1.6
7	74	1.6
8	77	1.7
9	81	1.8
10	90	1.8
11	89	1.7
12	83	1.8
13	82	1.8
14	81	1.7
15	71	1.7

Tomamos intervalos de amplitud 10 *para la variable* X*=peso. Así los intervalos para* X *empiezan en el límite real del mínimo peso* 62*:*

$$[61.5, 71.5),\ [71.5, 81.5),\ [81.5, 91.5).$$

Tomamos intervalos de amplitud 0.1 *para la variable* Y*=talla. Los intervalos para* Y *empiezan en el límite real de la mínimo altura* 1.6*.*

$$[1.55, 1.65),\ [1.65, 1.75),\ [1.75, 1.85).$$

La tabla de frecuencias agrupadas conjunta será:

X/Y	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

1.5.2. Distribuciones marginales

Consideremos una distribución conjunta de las variables (X, Y) donde X toma valores

$$\{X_1, X_2, \ldots, X_I\},$$

mientras que Y toma los valores

$$\{Y_1, Y_2, \ldots, Y_J\},$$

con la correspondiente tabla de frecuencias conjunta n_{ij}.

A la distribución unidimensional de la variable X la llamaremos distribución marginal de X y es la que toma los valores:

$$\{X_1, X_2, \ldots, X_I\},$$

y para la que la frecuencia absoluta correspondiente a X_i es:

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij},$$

es decir, la frecuencia absoluta del valor X_i es el número total de individuos que tienen la variable $X = X_i$.

De la misma forma, la distribución marginal de Y es aquella variable unidimensional que toma los valores

$$\{Y_1, Y_2, \ldots, Y_J\},$$

y para la que la frecuencia absoluta correspondiente al valor Y_j vale:

$$n_{\bullet j} = \sum_{i=1}^{I} n_{ij},$$

o sea, el número total de individuos observados que tienen la variable $Y = Y_j$.

Las tablas de frecuencias correspondientes a las distribuciones marginales son :

Distribución marginal de la variable X

X_i	$n_{i\bullet}$
X_1	$n_{1\bullet}$
X_2	$n_{2\bullet}$
$\vdots$	$\vdots$
X_i	$n_{i\bullet}$
$\vdots$	$\vdots$
X_I	$n_{I\bullet}$
	n

Distribución marginal de la variable Y

Y_i	$n_{\bullet j}$
Y_1	$n_{\bullet 1}$
Y_2	$n_{\bullet 1}$
$\vdots$	$\vdots$
Y_i	$n_{\bullet j}$
$\vdots$	$\vdots$
Y_I	$n_{\bullet J}$
	n

Ejemplo 14 *Consideremos una distribución conjunta (X, Y) no agrupada con tabla de frecuencias:*

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Las distribuciones marginales de X e Y son:

Distribución marginal de X	
X_i	$n_{i\bullet}$
20	5
30	4
40	3
	12

Distribución marginal de Y	
Y_i	$n_{\bullet j}$
65	4
75	6
85	2
	12

Ejemplo 15 *Consideremos una distribución conjunta (X, Y) en este caso de valores agrupados con tabla de frecuencias:*

X/Y	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Les distribuciones marginales de X e Y son:

Distribución marginal de X		
Intervalo	X_i	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	4
$[71.5, 81.5)$	76.5	7
$[81.5, 91.5)$	86.5	4
		15

Distribución marginal de Y		
Intervalo	Y_i	$n_{\bullet j}$
$[1.55, 1.65)$	1.6	5
$[1.65, 1.75)$	1.7	5
$[1.75, 1.85)$	1.8	5
		15

1.5.3. Distribuciones condicionadas

Consideremos una distribución conjunta de variables (X, Y) donde X toma valores

$$\{X_1, X_2, \ldots, X_I\},$$

e Y toma valores

$$\{Y_1, Y_2, \ldots, Y_J\},$$

con la correspondiente tabla de frecuencias conjunta n_{ij}.

Consideremos un valor concreto de la variable Y, Y_j. Definimos la distribución condicionada de X respecte al valor Y_j de Y y lo denotaremos por $X/Y = Y_j$ como aquella distribución unidimensional que toma los mismo valores que X, es decir,

$$\{X_1, X_2, \ldots, X_I\},$$

y tal que la frecuencia absoluta del valor X_i (a la que denotaremos por $n_{i/j}$) se define como el número de individuos observados que tienen $X = X_i$ e $Y = Y_j$.

De la misma manera, podemos considerar un valor concreto de la variable X, X_i. Definimos distribución condicionada de Y respecto del valor X_i y la denotaremos por $Y/X = X_i$ como aquella distribución unidimensional que toma los mismo valores que Y,

$$\{Y_1, Y_2, \ldots, Y_J\},$$

y tal que la frecuencia absoluta del valor Y_j (a la que denotaremos por $n_{j/i}$) se define como el número de individuos observados que tienen la $Y = Y_j$ y la $X = X_i$.

Observemos que existen tantas distribuciones condicionadas $X/Y = Y_j$ como valores distintos toma Y y que existen tantas condicionales $Y/X = X_i$ como valores distintos toma X.

Ejemplo 16 *Consideremos una distribución conjunta (X, Y) sin agrupar, con tabla de frecuencias:*

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Fijemos $Y = 75$. La tabla de frecuencias de la distribución $X/Y = 75$ es:

$X_i/Y = 75$	$n_{i/75}$
20	4
30	1
40	1
	6

Fijemos por ejemplo $X = 30$. *La tabla de frecuencias de la distribución* $Y/X = 30$ *es:*

$Y_j/X = 30$	$n_{j/30}$
65	3
75	1
85	0
	4

Ejemplo 17 *Consideremos una distribución conjunta* (X, Y) *caso agrupado con tabla de contingencia:*

X/Y	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Fijemos por ejemplo $Y = 1.6$. *La tabla de frecuencias de* $X/Y = 1.6$ *es:*

Intervalo	X_i	$n_{i/1.6}$
$[61.5, 71.5)$	66.5	3
$[71.5, 81.5)$	76.5	2
$[81.5, 91.5)$	86.5	0
		5

Fijemos por ejemplo $X = 86.5$. *La tabla de frecuencias de* $Y/X = 86.5$ *es:*

Intervalo	Y_j	$n_{j/86.5}$
$[1.55, 1.65)$	1.6	0
$[1.65, 1.75)$	1.7	1
$[1.75, 1.85)$	1.8	3
		4

1.5.4. Momentos bidimensionales

Consideremos una distribución conjunta de las variables (X, Y) donde X toma valores

$$\{X_1, X_2, \ldots, X_I\},$$

mientras que Y toma los valores

$$\{Y_1, Y_2, \ldots, Y_J\},$$

con la correspondiente tabla de frecuencias conjunta n_{ij}.

Vamos a estudiar los momentos bidimensionales de primer y segundo orden.

Los momentos de primer orden son la media de X $(\overline{x})$ y la media de Y $(\overline{y})$. Se calculan de la siguiente forma:

$$\overline{x} = \frac{\sum_{i=1}^{I} n_{i\bullet} X_i}{n}, \quad \overline{y} = \frac{\sum_{j=1}^{J} n_{\bullet j} Y_j}{n}.$$

Los momentos de segundo orden son la varianza de X (s_X^2), la varianza de Y (s_Y^2) y la covarianza de X e Y (s_{XY}).

Las fórmulas respectivas son:

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{I} n_{i\bullet}\left(X_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{I} n_{i\bullet} X_i^2 - \overline{x}^2,$$

$$s_Y^2 = \frac{1}{n}\sum_{j=1}^{J} n_{\bullet j}\left(Y_j - \overline{y}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_{\bullet j} Y_j^2 - \overline{y}^2,$$

$$s_{XY} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(X_i - \overline{x})(Y_j - \overline{y})}{n} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} X_i Y_j}{n} - \overline{x}\cdot\overline{y}.$$

Ejemplo 18 *Consideremos las siguientes distribución de las variables* (X, Y)*:*

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Los momentos de primer orden son:

$$\overline{x} = \frac{n_{1\bullet}X_1 + n_{2\bullet}X_2 + n_{3\bullet}X_3}{n} = \frac{5\cdot 20 + 4\cdot 30 + 3\cdot 40}{12} = 28.333,$$

$$\overline{y} = \frac{n_{\bullet 1}Y_1 + n_{\bullet 2}Y_2 + n_{\bullet 3}Y_3}{n} = \frac{4\cdot 65 + 6\cdot 75 + 2\cdot 85}{12} = 73.333$$

Los momentos de segundo orden son:

$$s_X^2 = \frac{n_{1\bullet}X_1^2+n_{2\bullet}X_2^2+n_{3\bullet}X_3^2}{n} - \overline{x}^2 = \frac{5\cdot 20^2+4\cdot 20^2+3\cdot 40^2}{12} - 28.333^2 = 63.888,$$

$$s_Y^2 = \frac{n_{\bullet 1}Y_1^2+n_{\bullet 2}Y_2^2+n_{\bullet 3}Y_3^2}{n} - \overline{y}^2 = \frac{4\cdot 65^2+6\cdot 75^2+2\cdot 85^2}{12} - 73.333^2 = 47.222,$$

$$\begin{aligned} s_{XY} = & \frac{1}{n}\left(n_{11}X_1Y_1 + n_{12}X_1Y_2 + n_{13}X_1Y_3 + n_{21}X_2Y_1 + n_{22}X_2Y_2\right. \\ & \left. +n_{23}X_2Y_3 + n_{31}X_3Y_1 + n_{32}X_3Y_2 + n_{33}X_3Y_3\right) - \overline{x}\cdot\overline{y} \\ = & \frac{1}{12}\left(1\cdot 20\cdot 65 + 4\cdot 20\cdot 75 + 0\cdot 20\cdot 85 + 3\cdot 30\cdot 65 + 1\cdot 30\cdot 75 + \right. \\ & \left. 0\cdot 30\cdot 85 + 0\cdot 40\cdot 65 + 1\cdot 40\cdot 75 + 2\cdot 40\cdot 85\right) - 28.333\cdot 73.333 \\ = & \ 22.222 \end{aligned}$$

Ejemplo 19 *Consideremos una distribución conjunta (X, Y) (datos agrupados) con tabla de frecuencias:*

X/Y	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Los momentos de primer orden son:

$$\overline{x} = 76.5, \quad \overline{y} = 1.7$$

Los momentos de segundo orden son:

$$s_X^2 = 53.333, \quad s_Y^2 = 0.006, \quad s_{XY} = 0.4$$

1.5.5. Independencia e incorrelación

Vamos a introducir dos conceptos nuevos: el de independencia y el de incorrelación.

El concepto de independencia formaliza la idea conocer el valor de la variable X no aporta información alguna sobre el valor de Y y viceversa.

Dada una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij}, diremos que X e Y son independientes si:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}, \text{ para todo } i \in \{1, 2, \ldots, I\} \text{y para todo } j \in \{1, 2, \ldots, J\}.$$

En el caso en que la relación anterior falle para un i y un j diremos que las dos variable no son independientes.

Ejemplo 20 *En este ejemplo las variables X e Y no son independientes:*

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

ya que por ejemplo

$$\frac{n_{11}}{n} \neq \frac{n_{1\bullet}}{n}\frac{n_{\bullet 1}}{n}, \quad \frac{1}{12} \neq \frac{5}{12}\cdot\frac{4}{12}.$$

En cambio en este otro caso, si son independientes:

X/Y	65	75	85	
20	3	2	1	6
30	6	4	2	12
40	6	4	2	12
	15	10	5	30

Dejamos al lector la comprobación como ejercicio.

El concepto de incorrelación formaliza la idea de relación lineal en el sentido de que las variables crecen de forma lineal conjuntamente (relación directa) o bien si una crece, la otra decrece (relación inversa). Dada una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij}, diremos que X e Y son incorreladas si su covarianza $s_{XY} = 0$.

La relación que existe entre los dos conceptos introducidos, el de independencia y el de incorrelación viene dada por la siguiente propiedad:

Teorema 21 *Si las variables X e Y son independientes entonces son incorreladas.*

El recíproco del teorema anterior no es cierto en general. Podemos decir que independencia implica incorrelación pero lo contrario no.

independencia $\Rightarrow$ incorrelación

Demostración del teorema:
Si X es independiente de Y tenemos que

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}.$$

Por lo tanto:

$$\begin{aligned} s_{XY} = & \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(x_i-\overline{x}\right)\left(y_j-\overline{y}\right) \\ = & \sum_{i=1}^{I}\left(x_i-\overline{x}\right)\frac{n_{i\bullet}}{n}\sum_{j=1}^{J}\left(y_j-\overline{y}\right)\frac{n_{\bullet j}}{n} \\ = & \ 0\cdot 0 = 0, \end{aligned}$$

teniendo en cuenta que si X es una variable unidimensional con valores $\{x_1, x_2, \ldots, x_I\}$, con las correspondientes frecuencias absolutas $\{n_1, n_2, \ldots, n_I\}$, tenemos:

$$\sum_{i=1}^{I} n_i\left(x_i-\overline{x}\right) = \sum_{i=1}^{I} n_i x_i - n\overline{x} = 0.$$

1.6. Asociación, concordancia y correlación

1.6.1. Introducción

En esta sección estudiaremos si existe algún tipo de relación entre dos variables X e Y.

Hasta ahora sabemos cuando dos variables son independientes o no. En caso de que no se sean independientes, nos interesará medir el grado de dependencia que tienen, es decir, si son "muy dependientes o no".

Para medir la dependencia utilizaremos una serie de coeficientes como son el coeficiente de contingencia de Pearson y en el caso de variables con sólo dos valores el coeficiente de contingencia de Yule.

1.6.2. Coeficientes de asociación

Coeficiente de contingencia o de correlación de Pearson

Dada una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij}, definimos el coeficiente de correlación de Pearson como:

$$C_P = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

donde χ^2 es el llamado estadístico de Pearson, que se define como:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

Las n_{ij} son las conocidas como frecuencias empíricas u observadas y las expresiones $\frac{n_{i\bullet} n_{\bullet j}}{n}$ se las conocen como frecuencias teóricas; o sea, las frecuencias conjuntas que tendrían las variables (X, Y) si fueran independientes.

Por lo tanto, cuando más cerca estén las frecuencias empíricas de las teóricas, más pequeño será el estadístico de Pearson χ^2 y el coeficiente de contingencia de Pearson C_P.

Ejemplo 22 *Consideremos la siguiente distribución conjunta:*

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

La tabla anterior es la de frecuencias empíricas. A continuación construimos la tabla de frecuencias teóricas:

X/Y	65	75	85	
20	1.66	2.5	0.83	5
30	1.33	2	0.66	4
40	1	1.5	0.5	3
	4	6	2	12

Como podemos observar, las frecuencias teóricas no coinciden con las empíricas. Por lo tanto, deducimos que las dos variables X e Y no son independientes.

Vamos a calcular ahora el coeficiente de contingencia de Pearson C_P.

En primer lugar hemos de calcular el estadístico χ^2:

$$\chi^2 = \frac{(1-1.66)^2}{1.66} + \frac{(4-2.5)^2}{2.5} + \frac{(0-0.83)^2}{0.83} + \frac{(3-1.33)^2}{1.33} + \frac{(1-2)^2}{2} + \frac{(0-0.66)^2}{0.66} + \frac{(0-1)^2}{1} + \frac{(1-1.5)^2}{1.5} + \frac{(2-0.5)^2}{0.5}$$

$$= 10.916$$

Por último calculamos el coeficiente de contingencia de Pearson C_P:

$$C_P = \sqrt{\frac{10.916}{12 + 10.916}} = 0.690$$

Propiedades C_P:

1) El valor de C_P es mayor o igual que 0 y menor que 1. En el caso en que X e Y sean independientes, las frecuencias empíricas y teóricas coinciden y $C_P = 0$.

2) Cuanto más dependientes son las variables X e Y, C_P se aproxima más a 1.

Por lo tanto si C_P es pequeño podemos decir que el grado de dependencia es bajo , mientras que si C_P aumenta es alto.

Para el caso más trivial en el que tengamos una tabla 2×2, es decir $I = J = 2$, o sea cuando las variables sólo toman dos valores cada una, se utiliza otro coeficiente, el coeficiente de contingencia de Yule. Se define así:

$$C_\gamma = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

Recordamos al lector que la tabla de frecuencias tendrá el siguiente aspecto:

X/Y	Y_1	Y_2	
X_1	n_{11}	n_{12}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	n

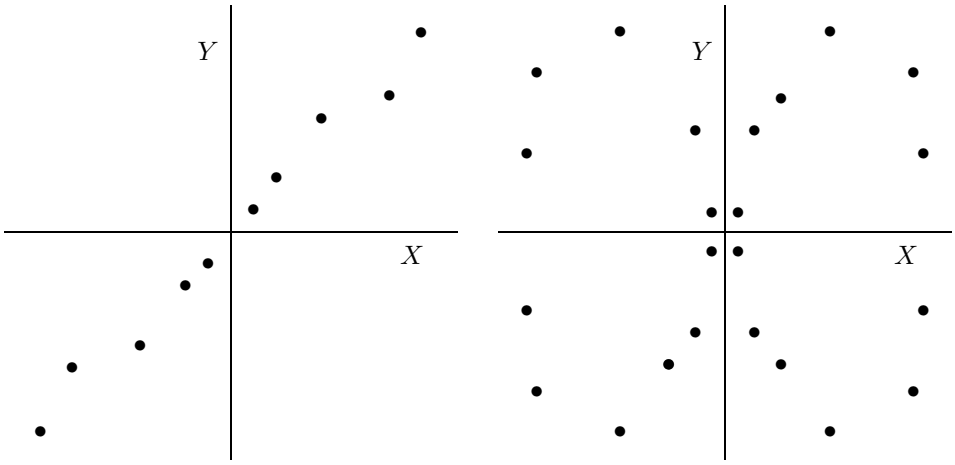

Figura 1.11: Nube de puntos con relación lineal y sin relación lineal

Ejemplo 23 *Calculemos el coeficiente de contingencia de Yule de la siguiente distribución conjunta:*

X/Y	Y_1	Y_2	
X_1	2	8	10
X_2	3	7	10
	5	15	20

$$C_\gamma = \frac{2 \cdot 7 - 3 \cdot 8}{2 \cdot 7 + 3 \cdot 8} = -0.263158$$

El coeficiente de contingencia de Yule siempre está entre -1 y 1. En caso de independencia entre las variables, se tiene que $C_\gamma = 0$.

1.6.3. Correlación lineal

El problema que nos planteamos en este punto es saber si existe una relación lineal entre X e Y, es decir, si existen dos valores numéricos a y b tales que:

$$Y \approx a + bX.$$

La relación lineal no tiene por que ser perfecta. Lo que nos interesa es medir esa relación lineal. En el gráfico de la izquierda de la figura 1.11 se

vislumbra una relación lineal mayor que en el de la derecha ya que podemos encontrar un a recta que aproxime mejor Y en función de X.

Vamos a introducir un coeficiente que mide la relación lineal entre dos variables. Este coeficiente es el coeficiente de correlación lineal de Pearson r_{XY} y se define de la manera siguiente

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y}.$$

Propiedades de r_{XY}:

1) $-1 \leq r_{XY} \leq 1$

2) $r_{XY} = r_{YX}$

3) Interpretación de r_{XY}:

- Si $r_{XY} > 0$ y a medida que se aproxima a 1, aumenta la relación lineal positiva entre las dos variables X e Y; lo que quiere decir que si X crece, la variable Y también y si X decrece, Y también, Obsérvese la parte izquierda de la figura 1.11 como ejemplo de este caso.
- Si $r_{XY} < 0$ y su valor está muy cerca de -1 quiere decir que hay una buena relación lineal negativa entre las dos variables X e Y; lo que significa que si la variable X crece, la variable Y decrece o viceversa. Como ejemplo ver el gráfico de la figura 1.12.
- Si $r_{XY} = 0$ o es pequeño, quiere decir que no hay ningún tipo de relación lineal entre las variables X e Y.
- Si $r_{XY} = \pm 1$, hay relación lineal exacta entre X e Y, o sea, existen dos números reales a y b tales que $Y = a + bX$.

Ejemplo 24 *Consideremos la siguiente distribución conjunta de la variable* (X, Y)*: (ver ejemplo 18)*

Los valores de s_X^2, s_Y^2 *y de* s_{XY} *son:*

$$s_X^2 = 63.888, \quad s_Y^2 = 47.222, \quad s_{XY} = 22.222$$

El coeficiente de correlación lineal vale:

$$r_{XY} = \frac{22.222}{\sqrt{63.888 \cdot 47.222}} = 0.405$$

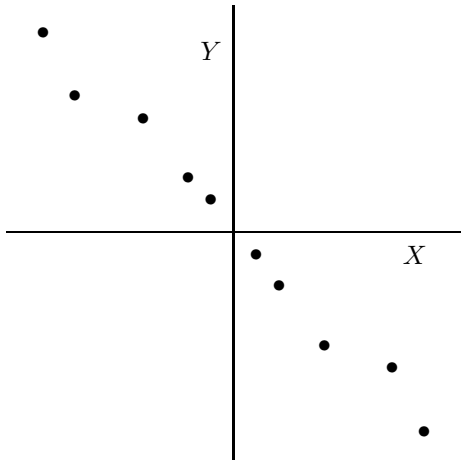

Figura 1.12: Nube de puntos con relación lineal negativa

1.6.4. Correlación ordinal

Vamos a estudiar ahora la relación que existe entre dos ordenaciones dadas por una muestra de datos bidimensionales.

Los estadísticos que miden este tipo de relaciones reciben el nombre de coeficiente de correlación ordinal y nos darán medidas de la similitud de las dos ordenaciones a lo que se suele llamar concordancia.

Más concretamente, consideremos un conjunto de individuos y los ordenamos según dos criterios. Tendremos así dos ordenaciones de los individuos. Estas ordenaciones las podemos disponer como si se tratara de una estadística bidimensional, donde la primera componente de la observación de un individuo correspondería al número de orden del primer criterio de ordenación y la segunda componente al otro.

Por ejemplo consideremos las observaciones en 5 humanos de su peso X en Kg. y estatura Y en metros:

Individuo i	(X_i, Y_i)	Orden X	Orden Y
Individuo 1	$(80, 1.75)$	3	2
Individuo 2	$(75, 1.92)$	2	4
Individuo 3	$(85, 1.67)$	4	1
Individuo 4	$(66, 1.80)$	1	3
Individuo 5	$(90, 2.00)$	5	5

Si ordenamos los individuos en orden ascendente (de menor a mayor) según el peso quedan así:

Rango Peso	1	2	3	4	5
Individuo	4	2	1	3	5

mientras que si los ordenamos en orden ascendente según su altura:

Rango	1	2	3	4	5
Individuo	3	1	4	2	5

Tenemos así dos ordenaciones de números ordinales enteros que reciben el nombre de rangos[7]. En general podemos escribir:

$$\begin{aligned} \text{para } X \rightarrow \quad & \{r_{x_1}, r_{x_2}, r_{x_3}, \ldots, r_{x_n}\}, \\ \text{para } Y \rightarrow \quad & \{r_{y_1}, r_{y_2}, r_{y_3}, \ldots, r_{y_n}\}, \end{aligned}$$

donde los valores de r_{x_i} y r_{y_i} dan el lugar que ocupa el valor x_i o el y_i en cada una de las muestras ordenadas. Estos valores están comprendidos entre 1 y n, luego son dos permutaciones de orden n.

Las diferencias entre las ordenaciones son:

$$d_i = r_{x_i} - r_{y_i}\,; i = 1, 2, \ldots, n.$$

El coeficiente de correlación ordinal o por rangos de Spearman queda definido por:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}.$$

De hecho, r_S no es más que el coeficiente de correlación lineal introducido en la sección anterior aplicado a los rangos.

Propiedades de r_S:

- Si $r_S = 1$, las dos ordenaciones coinciden; o sea, $r_{x_i} = r_{y_i}$ para cualquier i entre 1 y n.

[7]El cálculo de rangos se complica en el caso de empates, es decir cuado hay valores repetidos en las series de datos. En estos casos se puede romper el empate de varias maneras.

- Si $r_S = -1$, la ordenación de Y es exactamente la opuesta a la de X, es decir , $r_{x_i} = r_{y_{n-i+1}}$ para cualquier i entre 1 y n.

- El coeficiente r_S está siempre comprendido entre -1 y 1. Si $r_S > 0$, podemos decir que las dos ordenaciones son del mismo sentido y si $r_S < 0$, las dos ordenaciones son de sentidos opuestos.

Ejemplo 25 *Consideremos la muestra anterior de pesos y estaturas de* 5 *individuos:*

Individuo i	(X_i, Y_i)	r_{X_i}	r_{Y_i}	d_i^2
Individuo 1	$(80, 1.75)$	3	2	1
Individuo 2	$(75, 1.92)$	2	4	4
Individuo 3	$(85, 1.67)$	4	1	9
Individuo 4	$(66, 1.80)$	1	3	4
Individuo 5	$(90, 2.00)$	5	5	0
Σ				18

Luego tenemos que :

$$r_s = 1 - \frac{6 \cdot 18}{5 \cdot (25-1)} = 1 - \frac{108}{120} = 0.1$$