11. ESTADíSTICA DESCRIPTIVA

11.1 Introducción

El objeto de cualquier investigación estadística es la toma de información acerca de los individuos de cierto colectivo llamado *población estadística*. Cuando nos preguntamos por las características de un grupo homogéneo de individuos (entendido "individuo" en sentido amplio: una persona, un coche, una empresa), estamos fijando una *población* de estudio.

Cada elemento de una población es un *individuo,* pudiendo clasificarse estas poblaciones estadísticas en poblaciones *finitas e infinitas*, de acuerdo con el número de individuos incluidos en las mismas. Así, son ejemplos de poblaciones finitas: los estudiantes de una universidad, los habitantes de un país, etc; como ejemplos de poblaciones infinitas: el conjunto de las "infinitas" tiradas de un dado, el conjunto de los "infinitos" tornillos producidos por una fábrica, etc.

Cuando se estudia el total de la población se dice que se ha realizado un censo. Debido a que el proceso de observación de una característica para todos los individuos de una población puede resultar costoso (económica o humanamente), o sencillamente irrealizable es frecuente la elección de un subconjunto de observaciones de la población inicial. Este subconjunto de observaciones constituye lo que llamamos *muestra*.

La *Estadística Inductiva* se ocupa de extraer información sobre distintas características de interés de la población a partir de la información contenida en una muestra. La parte de la Estadística que trata solamente de describir y analizar un grupo dado de datos sin sacar conclusiones o inferencias acerca de la población que los ha generado se llama *Estadística Descriptiva o Deductiva*.

11.2 Variables estadísticas unidimensionales

A cada una de las características de la población que nos interese estudiar le llamaremos variable estadística. Las variables estadísticas consideradas para la descripción de los individuos de la muestra puede presentar distintas *modalidades o estados*: la variable sexo presenta la modalidades hombre o mujer, la variable categoría laboral puede presentar las modalidades administrativo, técnico y directivo. Según la naturaleza de las modalidades, las variables estadísticas se dividen en variables *cualitativas* (por ejemplo la variable sexo), las cuales toman valores no medibles, y variables *cuantitativas* (estatura, peso,...),

que toman valores medibles (valores numéricos). Éstas últimas se clasifican, a su vez, en variables *discretas* y *continuas*.

Las variables discretas se caracterizan porque toman, a lo sumo, tantos valores distintos como números naturales hay. Para las variables continuas, sin embargo, sólo se puede señalar un intervalo de valores (pensemos en la variable estatura), dentro del cual la variable pueda tomar cualquier posible valor.

11.3 Tablas de frecuencias

Si se supone que $x_1, x_2, \ldots, x_n$ son los valores medidos en una muestra de tamaño n de una variable X que presenta k modalidades $c_1, c_2, \ldots, c_k$, y si n_i es el número de individuos de la muestra que presentan la modalidad c_i $(i = 1, 2, \ldots, k)$.

Al número n_i se le denomina *frecuencia absoluta de la modalidad* c_i, definiéndose *la frecuencia relativa de* c_i como el número $f_i = n_i/n$ $(i = 1, 2, \ldots, k)$, que representa la proporción de individuos de la muestra que presentan dicha modalidad.

También se pueden definir las frecuencias acumuladas o acumulativas, siempre que los valores de la variable sean susceptibles de ser ordenados. Así, la *frecuencia absoluta acumulada* como:

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{j=1}^{i} n_i \quad (i = 1, 2, \ldots, k)$$

que representa el número de individuos de la muestra que presentan alguna de las i primeras modalidades, y la *frecuencia relativa acumulada*

$$F_i = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_i = \frac{N_i}{n} \quad (i = 1, 2, \ldots, k)$$

que representa la proporción de individuos de la muestra que presentan alguna de las i primeras modalidades.

Se puede representar la distribución de frecuencias de la variable unidimensional X en una tabla con las modalidades y sus respectivas frecuencias. Si la variable es cualitativa, se tomarán como modalidades las distintas respuestas observadas en la muestra; si la variable es discreta (que tome pocos valores distintos), las modalidades coincidirán con los distintos valores (dispuestos en orden creciente) medidos en la muestra; si la variable es continua (o bien discreta pero que toma muchos valores distintos), se tomarán como modalidades los intervalos de clase.

11.3.1 Intervalos de clase

Cuando se trata de describir una variable continua (o discreta que toma muchos valores distintos) se utiliza la técnica de construir intervalos de valores, llamados *intervalos de clase*, en los cuales se agruparán los datos observados en la muestra. Así cada intervalo es considerado una modalidad, tomándose como frecuencia absoluta de cada modalidad el número de observaciones agrupadas en el intervalo correspondiente.

Una vez construidos los intervalos de clase se elige un representante en cada uno de ellos, llamado *marca de clase*, que usualmente es tomado como el punto medio del intervalo pudiendo utilizarse posteriormente estos representantes para el cálculo (con datos agrupados) de distintas medidas descriptivas.

El número de intervalos debe ser mayor al aumentar el tamaño muestral, por lo que se recomienda construir tantos intervalos como el número entero (entre 5 y 20) más próximo a $\sqrt{n}$.

En general, salvo que conozcamos alguna información que nos incline a construirlos de amplitudes distintas, construiremos intervalos de la misma amplitud. Para evitar posibles coincidencias de los extremos de clase con alguna observación muestral, es habitual precisar estos extremos con una cifra decimal más que las utilizadas en las anotaciones muestrales.

Ejemplo 1: La empresa XYZ tiene un total de 512 trabajadores clasificados en 4 categorías laborales según la tabla de frecuencias siguiente:

PUESTO

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Administrativo	178	34.8	34.8	34.8
Técnico	208	40.6	40.6	75.4
Especialista	99	19.3	19.3	94.7
Directivo	27	5.3	5.3	100.0
Total	512	100.0	100.0	

Figura 11-1 Tabla de frecuencias de la variable puesto de trabajo de la empresa XYZ.

11.4 Medidas características de una variable

Las medidas descriptivas se dividen, de acuerdo con su cometido, en: medidas *de tendencia* (también se dicen *de posición* o *de localización*), medidas *de dispersión* (o *de variabilidad*) y medidas de *forma.*

Las medidas de localización nos indican el ámbito de fluctuación de los datos; las de dispersión, el grado de concentración de las observaciones en torno

a un punto central; y las de forma, las distorsiones que sufre la distribución de frecuencias en la muestra con respecto a una situación estándar que sirve de referencia.

11.4.1 Medidas de tendencia

Las principales medidas de tendencia son las llamadas medidas de *tendencia central*, diseñadas para indicar el punto en torno al cual se sitúan los datos. Por esta razón, tales medidas se utilizan, a menudo, como representantes del conjunto muestral (o de su *tamaño*). Las medidas de tendencia central usuales son la *media*, la *mediana* y la *moda*. La media (aritmética) se define como

$$\overline{x} = \frac{x_1 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

$$\overline{x} = \frac{n_1 c_1 + ... + n_k c_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{n} f_i c_i.$$

La *mediana* es el punto antes del cual se encuentran el 50% de los casos en la muestra. Es una medida de tendencia central que se prefiere en algunos casos a la media, al estar menos influenciada por datos extremos.

Para variables de tipo discreto, la *moda* se define como el valor más repetido. Para variables continuas hay que redefinir el concepto de moda, ya que no se apreciarán repeticiones sistemáticas en los datos. Se define el *intervalo modal* como el intervalo de clase con la mayor frecuencia asociada.

11.4.2 Medidas de dispersión

Estas medidas tienen por cometido señalar la cantidad de variabilidad en la muestra. Para medir la variabilidad en términos absolutos, utilizaremos fundamentalmente la *desviación típica* o el *rango*, cantidades que dependen de las unidades de medida. Existen índices de dispersión que no tienen tal dependencia, dando una idea relativa de la cantidad de variabilidad muestral. El más usado es el *coeficiente de variación*.

Empezando por las medidas absolutas de variabilidad, la desviación típica de la muestra se define como

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Cuando los datos se presentan en una tabla de frecuencias, se tiene más bien

$$s = \sqrt{\frac{n_1(c_1 - \overline{x})^2 + ... + n_k(c_1 - \overline{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \overline{x})^2}$$

Al cuadrado de la desviación típica se le llama *varianza.*

El coeficiente de variación se define, supuesto $\overline{x} > 0$, como el cociente entre la desviación típica y la media de la muestra. Carece de unidades, por lo que es adecuado cuando se comparan variabilidades en muestras de muy distinto tamaño o tomadas en unidades distintas.

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

Medidas de forma

La forma en que se distribuyen los datos puede ser más o menos fiel a un estándar. El modelo de referencia por antonomasia es el *modelo normal o gaussiano.* Este modelo es simétrico y tiene una única moda (que coincide con los valores medio y mediano).

Como medida de la falta de simetría, consideramos el *coeficiente de asimetría de Fisher,* que se define como

$$\gamma_1 = \frac{1}{s^3}\frac{(x_1 - \overline{x})^3 + ... + (x_n - \overline{x})^3}{n} = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \overline{x})^3$$

Al igual que el coeficiente de variación, el coeficiente de asimetría carece de unidades. En una situación perfectamente simétrica (como la gaussiana), encontraríamos el coeficiente γ_1 es nulo. Si $\gamma_1 > 0$, decimos que los datos presentan una asimetría positiva (o hacia la derecha). Si $\gamma_1 < 0$, decimos que la situación es asimétrica negativa (o hacia la izquierda).

Una medida del exceso (o carencia) de apuntamiento en la distribución de los datos está dada por el *coeficiente de curtosis de Fisher*

$$\gamma_2 = \frac{1}{s^4}\frac{(x_1 - \overline{x})^4 + ... + (x_n - \overline{x})^4}{n} - 3 = \frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3.$$

Este coeficiente (también carente de unidades) nos dice si la aglomeración de los datos en torno a la media se produce de forma gaussiana también llamada distribución mesocúrtica ($\gamma_2 = 0$) o si, por contra, la distribución es más aplastada que la normal que denominaremos planicúrtica ($\gamma_2 < 0$) o más apuntada en cuyo caso diremos que es leptocúrtica ($\gamma_2 > 0$).

Ejemplo 2: Dados los siguientes datos: 3,2,5,7,6,4,1,9,5,7,6,4,2 correspondientes a las notas de matemáticas de 13 alumnos, calcular su media, desviación típica y coeficiente de variación.

La media resulta:

$$\overline{x} = \frac{n_1 c_1 + ... + n_k c_k}{n} = \frac{61}{13} = 4.6923.$$

La desviación típica es:

$$s = \sqrt{\frac{n_1(c_1 - \overline{x})^2 + ... + n_k(c_1 - \overline{x})^2}{n}} = 2.2321.$$

El coeficiente de variación resultaría:

$$CV = \frac{s}{\overline{x}} = \frac{2.2321}{4.6923} = 0.4757.$$

11.5 Representaciones gráficas

Uno de los procedimientos habitualmente utilizados para la descripción de datos consiste en el uso de gráficos, que, en general, se construyen con el propósito de presentar de modo gráfico la información recogida en las tablas de frecuencias. Existe un amplio catálogo de posibles construcciones gráficas, a continuación se describen algunas de las más frecuentemente utilizadas:

11.5.1 Variables cualitativas

Diagrama de barras Para la construcción de este gráfico se parte de un sistema de ejes de coordenadas, representándose en el eje de abscisas las modalidades de la variable y en el eje de ordenadas las frecuencias (n_i ó f_i). A continuación, sobre cada modalidad se levanta un rectángulo o barra de altura igual a la frecuencia (absoluta o relativa) representada, siendo todos los rectángulos de igual base.

Ejemplo 3: Los datos del ejemplo 1 del puesto de trabajo de la empresa XYZ pueden verse representados en el diagrama de barras de la figura 1.2.

Diagrama de sectores La idea de este gráfico es semejante a la del diagrama de barras, sólo que en vez de utilizar un sistema de ejes de coordenadas para representar las frecuencias, se trata de repartir un círculo de radio arbitrario en sectores proporcionales a la frecuencia de cada modalidad.

En el caso del ejemplo de la empresa XYZ el diagrama de sectores puede apreciarse en la figura 1.3.

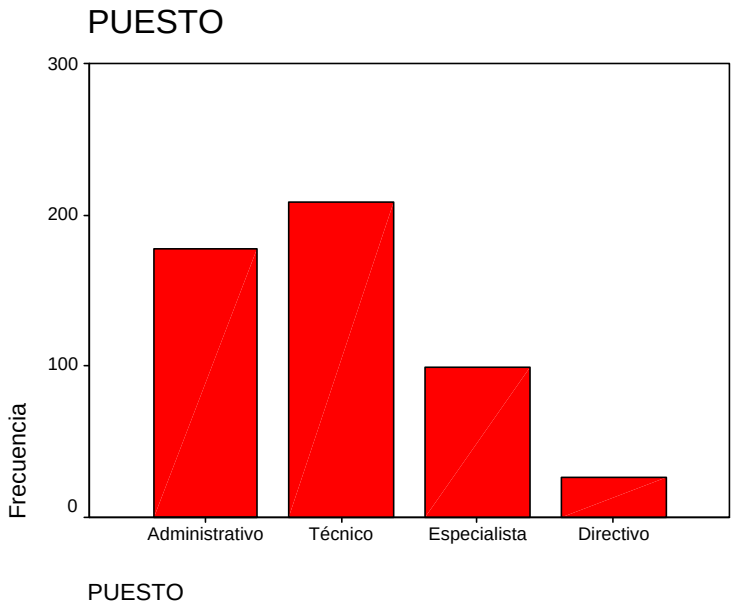

Figura 11-2 Diagrama de barras de la variable puesto.

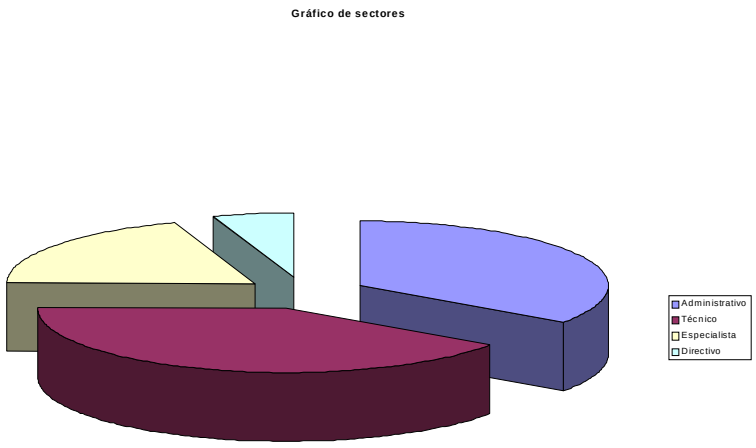

Figura 11-3 Gráfico de sectores de la empresa XYZ.

11.5.2 Variables cuantitativas discretas

Diagrama de barras Se construye de la forma anteriormente descrita para variables cualitativas, pudiendo sustituirse las barras por segmentos de recta.

Diagrama acumulativo de frecuencias Su construcción se realiza representando en un sistema de ejes de coordenadas los puntos (c_i, N_i) ó (c_i, F_i), según se representen frecuencias absolutas o relativas acumuladas, uniéndose a continuación de forma escalonada los puntos representados.

11.5.3 Variables cuantitativas continuas

Histograma Esta representación gráfica es el equivalente continuo del diagrama de barras. Partiendo de un sistema de ejes de coordenadas, se representan en el eje de abscisas los extremos e_i de los intervalos de clase y a continuación, tomando como base de cada rectángulo la amplitud c_i del intervalo de clase correspondiente, se levanta sobre cada uno de los intervalos un rectángulo de altura $h_i = n_i/c_i$ ó $h'_i = f_i/c_i$, según se representen frecuencias absolutas o relativas.

De esta manera se consigue representar la idea de densidad de frecuencia por unidad de amplitud asociada al concepto de variable continua. En el caso más general en que todos los intervalos sean de la misma amplitud, podrán tomarse las alturas de los rectángulos iguales a las frecuencias de los correspondientes intervalos.

Uniendo los punto medios de las bases superiores de cada rectángulo del histograma se obtine la representación gráfica llamada *polígono de frecuencias.* También se puede representar los correspondientes polígonos acumulativos sin más que representar las frecuencias acumuladas.

Ejemplo 4: El salario correspondiente a los técnicos de la empresa XYZ puede verse en la figura 1.4:

Las medidas características para el ejemplo de la empresa XYZ, para el salario de la variable puesto en el caso de los técnicos están recogidas en la Tabla de la figura 1.5.

Ejemplo 5: En la figura 1.6 puede apreciarse el histograma correspondiente a unos datos con distribución simétrica y mesocúrtica, cuyo coeficiente de asimetría es -0.01 y con coeficiente de curtosis o apuntamiento -0.087.

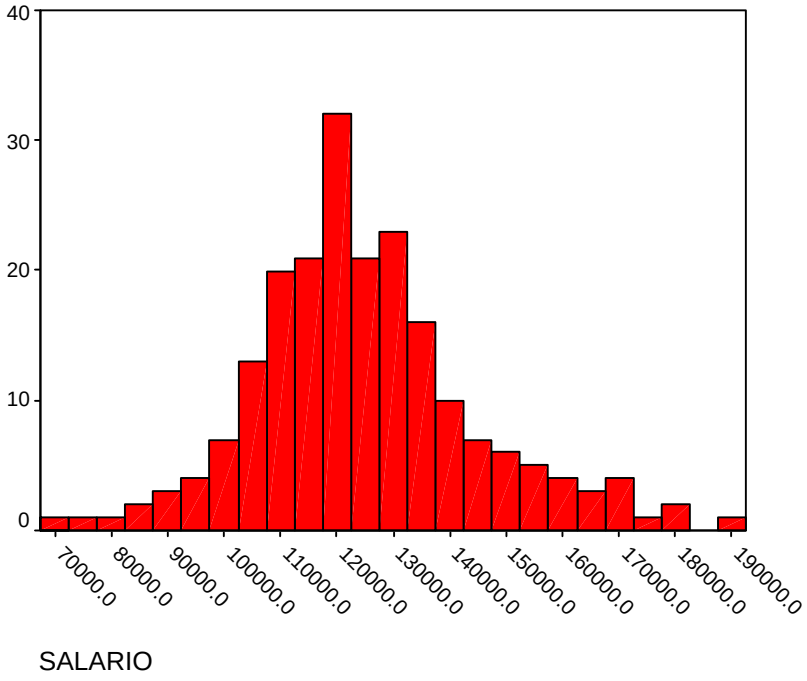

Figura 11-4 Histograma de la variable salario.

Estadísticos descriptivos

	N	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
SALARIO	208	124991.39	19664.984	3.87E+08	.504	.169	.825	.336
N válido (según lista)	208							

Figura 11-5 Medidas características para la variable salario de los técnicos de XYZ.

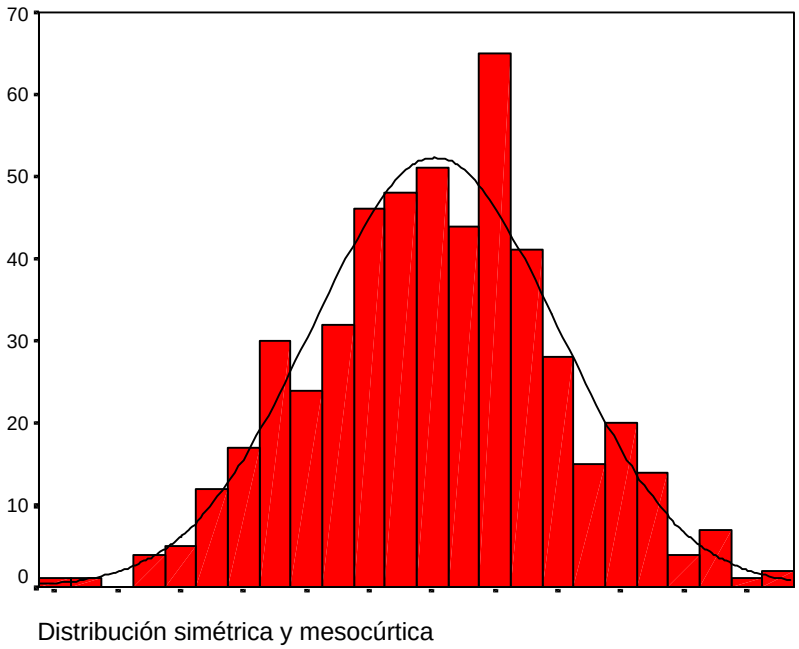

Figura 11-6 Histogram con curva Normal o gausiana.

11.6 Variables estadísticas bidimensionales

En muchas ocasiones interesa el estudio de más de una variable estadística, y es frecuente que se agrupen las variables de dos en dos para analizar posibles relaciones. En este tema analizaremos las técnicas estadísticas descriptivas para dos variables. De acuerdo con la forma en que viene dada la información distinguimos dos situaciones: el análisis de *tablas de contingencia* (asociadas a variables de tipo cualitativo) y los *estudios de correlación* (adecuados para variables numéricas).

11.7 Tablas de contingencia

Los datos en la tabla siguiente muestran el tipo de puesto de trabajo y la variable sexo de la empresa XYZ. Las cantidades que incluye la tabla son frecuencias absolutas: por ejemplo, se encontraron 144 individuos administrativos y mujeres, y 172 técnicos y hombres. Las sumas por filas y columnas que figuran en los márgenes de la Tabla se dicen *frecuencias marginales*; estas cantidades representan el comportamiento de las dos variables en la muestra cuando cada una se considera sin tener en cuenta la otra. Las sumas de ambos márgenes

Tabla de contingencia PUESTO * SEXO

Recuento

		SEXO		
		Hombre	Mujer	Total
PUESTO	Administrativo	34	144	178
	Técnico	172	36	208
	Especialista	20	79	99
	Directivo	17	10	27
Total		243	269	512

Figura 11-7 Tabla de doble entrada de la empresa XYZ.

dan una cantidad común: el tamaño muestral (512 sujetos observados). Por lo que respecta al número de hombres, tenemos un 47.46% ($100 \times \frac{243}{512}$) del total, y 269 mujeres (*frecuencias marginales relativas* de la variable "sexo"). En cuanto al número de administrativos tenemos un 34.77% ($100 \times \frac{178}{512}$), 208 técnicos, 99 especialistas y 27 directivos (*frecuencias marginales relativas* de la variable "puesto de trabajo").

Una tabla como la anterior se llama *tabla de contingencia* o *tabla de doble entrada*. Para analizar la relación entre las variables que definen la tabla de contingencia recurrimos primeramente a las llamadas *frecuencias condicionadas.* Estas frecuencias indican cómo se distribuyen los subgrupos de individuos definidos por una variable (*variable condicionante*) entre las distintas modalidades de la otra *(variable condicionada*). Así, para los datos de la Tabla anterior, las frecuencias de la variable "puesto" condicionada a "ser mujer" (es decir, en tal subgrupo de 269 individuos) son:

administrativa	$144/269 = 0.535$
técnica	$36/269 = 0.1334$
especialista	$79/269 = 0.294$
directiva	$10/269 = 0.0372$

En particular, observamos que el 53.5% de los individuos mujeres son administrativas.

11.8 Regresión y correlación

El propósito del estudio de los modelos de regresión es la construcción de modelos matemáticos que permitan explicar la relación de dependencia existente entre una *variable respuesta o dependiente* Y, y una o más *variables explicativas, regresoras o independientes* X, pudiendo, por ejemplo, utilizarse

estos modelos como herramienta para la predicción de nuevos valores de la variable respuesta a partir del conocimiento de valores particulares que han tomado la variable o variables explicativas.

Se trata de utilizar la información contenida en una muestra de la variable bidimensional (X, Y) para construir modelos matemáticos de la forma:

$$Y = m(X) + \varepsilon,$$

donde ε representa el error cometido al predecir el valor de Y mediante el valor que resulta de evaluar la función m en un valor de X. La *función de regresión* m determina, por tanto, la relación de dependencia entre X e Y.

Si se supone que m tiene una forma lineal $m(x) = a + bx$ nuestro interés es determinar los parámetros a y b del modelo. Para este fin puede observarse la forma que adopta el diagrama de dispersión construido con las observaciones (x_i, y_i) para encontrar la recta que "mejor se ajusta" al diagrama de dispersión o nube de puntos.

11.8.1 El método de mínimos cuadrados

Si se observan los valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, n pares de observaciones de una variable bidimensional (X, Y), y supongamos que se desea ajustar una recta de la forma $y = a + bx$ al diagrama de dispersión de (X, Y), el método de mínimos cuadrados proporciona un criterio para la determinación de los valores desconocidos de los parámetros, construido en base a la medida de ajuste consistente en minimizar la función $\Phi(a, b)$ siguiente:

$$\Phi(a, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2,$$

obteniéndose que los valores que minimizan $\Phi(a, b)$ son los números

$$\hat{a} = \overline{y} - \hat{b}\overline{x}, \quad \hat{b} = \frac{s_{xy}}{s_x^2}$$

Por tanto la expresión de la recta de regresión es:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

La recta de regresión anterior suele conocerse como *recta de regresión de Y sobre X*, ya que ha sido obtenida considerando que Y es la variable respuesta y que X es la variable explicativa. En esta aparece la *covarianza* s_{xy} de la variable bidimensional (X, Y), definida como:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Si intercambiamos los papeles de X e Y, obtendremos (en general) una recta de regresión distinta, llamada *recta de regresión de X sobre Y*, de ecuación

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$

Asimismo, llamaremos *coeficientes de regresión* a las pendientes de las rectas de regresión de Y sobre X, $\beta_{yx} = s_{xy}/s_x^2$, y de X sobre Y, $\beta_{xy} = s_{xy}/s_y^2$. Si las rectas de regresión coinciden (dependencia lineal exacta), se verifica que $\beta_{yx} = \beta_{xy}^{-1}$ $(\beta_{yx}\beta_{xy} = 1)$.

11.8.2 Correlación

Una forma de medir la adecuación de un modelo lineal es mediante la definición del *coeficiente de correlación de Pearson*:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{1}{n s_X s_Y}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Este coeficiente toma valores entre -1 y 1; valores próximos a 1 o -1 indican una alta adecuación del modelo lineal, o una *alta correlación*.

Por otra parte, si el modelo lineal es inadecuado (es decir, si las variables Y y X no guardan una asociación o dependencia lineal) se encontrarán valores de r_{XY} en torno a cero.

Otro índice de interesante interpretación es el *coeficiente de determinación* R^2, que se define como el cuadrado de r_{XY}. El valor R^2 está comprendido entre 0 y 1, y representa la cantidad de variabilidad de Y explicada por el modelo lineal de predicción.

En el supuesto de que el modelo ajustado corresponda a la recta de regresión de Y sobre X, la *predicción* del valor de Y, conocido un valor particular x_0 de X, se realizará sustituyendo en la ecuación del modelo ajustado x por el valor numérico x_0, de manera que la predicción de Y con el modelo de regresión lineal simple es $\hat{y}_0 = \hat{a} + \hat{b}x_0$.

Ejemplo 6: Se han recogido las notas de 14 alumnos en las materias de cálculo y estadística quedando reflejadas en la tabla de la figura 11.8.

Podríamos calcular y representar la recta de regresión correspondiente a los datos de la tabla 11.8 en la gráfica 11.9.

Resúmenes de casos

	NOTAS_ES	NOTAS_CA
1	4.50	5.50
2	6.00	5.00
3	7.00	5.50
4	3.00	2.50
5	2.50	2.00
6	6.00	7.50
7	8.00	8.50
8	9.00	9.50
9	7.00	8.50
10	2.00	3.00
11	4.00	5.50
12	1.00	.50
13	2.00	2.50
14	6.00	4.00
Total N	14	14

Figura 11-8 Notas de 14 alumnos en estadística y cálculo.

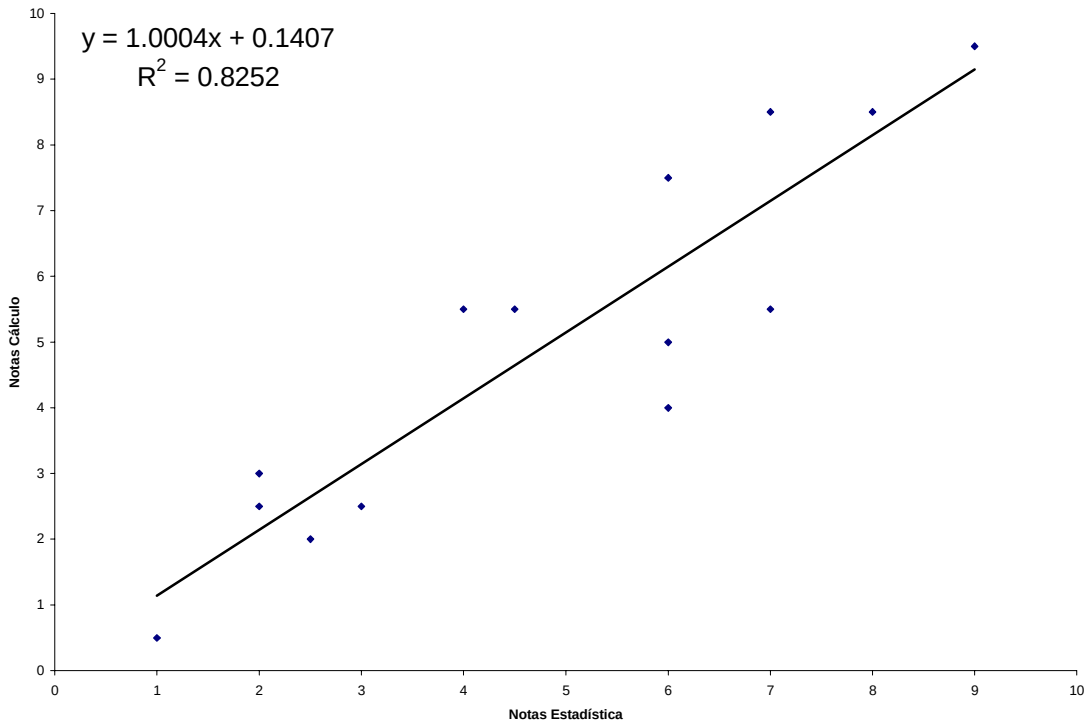

Figura 11-9 Ajuste lineal de la variable notas de estadística frente a notas de cálculo.