Capítulo 1

DISTRIBUCIONES DE FRECUENCIAS

Un concepto básico para el manejo de los cultivos es el de *plasticidad fenotípica*, la capacidad que tienen los seres vivos para desarrollar características diferentes según cómo sea el ambiente que los rodea. Conocer los detalles de la plasticidad fenotípica de las plantas cultivadas es necesario para diseñar sistemas de cultivo en los que éstas desarrollen características deseadas. En el marco de una investigación sobre los factores que controlan el rendimiento del cultivo de girasol (*Helianthus annuus* L.), una investigadora del Departamento de Producción Vegetal de la FA-UBA cultivó parcelas experimentales de girasol con 5 plantas por m² (densidad baja) y con 10 plantas por m² (densidad alta).

Un grupo de estudiantes aprovechará dos de estas parcelas para evaluar objetivamente características de las plantas que podrían presentar plasticidad frente a las diferencias en la densidad de cultivo. En cada parcela examinarán 40 plantas y en cada una registrarán las siguientes características o variables de interés: altura del tallo (cm), el número de hojas y el sentido de la inclinación del tallo (este, oeste, norte, sur o ninguno). Con estos datos describirán la expresión de estas características en los conjuntos de plantas examinados en cada parcela. Para ello,

aplicarán una serie de herramientas estadísticas para resumir los datos de cada variable de modo de contestar las siguientes preguntas:

¿Qué valores tiene la variable en las plantas observadas en cada parcela? ¿Cuáles valores son frecuentes y cuáles son raros? ¿Cuál es el promedio de la variable entre las plantas observadas en cada parcela? ¿Cuán dispersos están los valores de la variable registrados en las plantas observadas en cada parcela?

Para cada parcela y para cada variable de interés, los estudiantes evaluarán la **frecuencia** de cada valor registrado y luego reunirán dichas frecuencias en la **distribución de frecuencias** de la variable en cada parcela. Con esa información básica construirán **tablas** y **gráficos** y calcularán **medidas de posición** y **medidas de dispersión** para describir y comparar las distribuciones de frecuencias entre las dos parcelas. Cuando las distribuciones de frecuencias de una variable de interés difieran notablemente entre las dos parcelas, los estudiantes interpretarán que han encontrado un indicio de que las plantas de girasol presentan plasticidad fenotípica en la característica en cuestión frente a la diferencia en la densidad del cultivo.

Frecuencia y distribución de frecuencias

La frecuencia es el ladrillo básico de la inferencia estadística.

Una vez que los estudiantes hayan realizado sus registros, se dispondrán a describir el conjunto de 40 plantas observadas en cada parcela en relación con cada variable de interés. El primer paso para ello será identificar qué registros diferentes realizaron (p.ej. diferentes valores de altura, diferentes sentidos de inclinación, etc.) y determinar cuántas veces se repitió cada uno, su **frecuencia**. La lista de los valores o categorías de una variable acompañados por sus correspondientes frecuencias es la **distribución de frecuencias** de dicha variable.

La distribución de frecuencias organiza la información disponible para describir cómo era el conjunto de las plantas observadas respecto de una variable de interés. Por ejemplo, la distribución de frecuencias de la variable *altura* permite establecer: (a) si a grandes rasgos las plantas eran altas o bajas y (b) si formaban un conjunto de altura homogénea o heterogénea. La primera caracterización (plantas altas o bajas) se relaciona con el **promedio** de las alturas y la segunda (altura homogénea o heterogénea) con su **variabilidad**[1].

Frecuencia absoluta y frecuencia relativa

La **frecuencia absoluta** es el número de veces que se repite algo y la **frecuencia relativa** es la proporción que representa la frecuencia absoluta en relación con el total. Por ejemplo, en la parcela de girasol con densidad baja los estudiantes observaron y registraron los sentidos de inclinación de los tallos de 40 plantas. Los números de plantas con tallos inclinados en cada sentido encontrado (números de veces en que se repitió cada sentido) son las frecuencias absolutas observadas y los cocientes entre esos números y el total de plantas observadas (40) son las correspondientes frecuencias relativas (Cuadro 1.1). La suma de todas las frecuencias relativas es igual a 1.

[1] Las nociones de frecuencia, promedio y variabilidad no sólo se aplican a la descripción de conjuntos de objetos sino también a la de conjuntos de hechos o episodios. Por ejemplo, se pueden evaluar las frecuencias, promedio y variabilidad de las intensidades de las lluvias (mm/h) que ocurren en Buenos Aires.

Cuadro 1.1. Distribución de frecuencias de los sentidos de inclinación de los tallos de 40 plantas de girasol de una parcela con densidad baja (5 plantas por m²). Las plantas estaban dispuestas en hileras con dirección norte-sur. (*fa*, frecuencia absoluta, *fr*, frecuencia relativa).

Inclinación	*fa*	*fr*
Este	4	0,100
Oeste	5	0,125
Ninguno (vertical)	31	0,775
Total	40	**1,000**

Distribución de frecuencias

La distribución de frecuencias de una variable es la especificación de las frecuencias correspondientes a cada uno de sus valores o categorías.

La tabla del Cuadro 1.1 presenta las distribuciones de frecuencias absolutas y de frecuencias relativas de la variable *inclinación de los tallos* registrada en las 40 plantas de girasol de la parcela experimental con densidad baja. En este caso sencillo, la tabla nos alcanza para notar que: (a) las plantas estaban en su mayoría en posición vertical, (b) las pocas plantas inclinadas se repartían en números similares entre aquellas inclinadas hacia cada costado de la hilera (sentidos este y oeste) y (c) ninguna planta estaba inclinada en la dirección de la hilera (sentidos norte o sur).

La comparación de la descripción precedente con la distribución de frecuencias de los sentidos de inclinación de los tallos entre las plantas de la parcela con densidad alta permite notar diferencias y similitudes (Cuadro 1.2). En esta segunda parcela: (a) la mayoría de las plantas no estaban en posición vertical sino que estaban inclinadas, (b) como en la primera parcela, también en ésta las plantas inclinadas se repartían en números similares entre aquellas inclinadas hacia cada costado de la hilera (sentidos este y oeste) y (c) en esta parcela tampoco se encontró ninguna planta inclinada en la dirección de la hilera (sentidos norte o sur).

Cuadro 1.2. Distribución de frecuencias de los sentidos de inclinación de los tallos de 40 plantas de girasol de una parcela con densidad alta (10 plantas por m²). Las plantas estaban dispuestas en hileras con dirección norte-sur. (*fa*, frecuencia absoluta, *fr*, frecuencia relativa).

Inclinación	*fa*	*fr*
Este	14	0,350
Oeste	16	0,400
Ninguno (vertical)	10	0,250
Total	40	**1,000**

Al describir y comparar estas distribuciones de frecuencias, encontramos un indicio de plasticidad fenotípica en la inclinación de los tallos de las plantas de girasol. En este caso sencillo logramos hacerlo con un mínimo resumen de los datos.

En otros casos, para describir los rasgos principales de una distribución de frecuencias, se hace necesario resumir los datos más intensamente. A tal fin, se pueden construir tablas y gráficos y calcular medidas numéricas que resumen las magnitudes de la variable (medidas de posición) o que resumen su variabilidad (medidas de dispersión). Las alternativas disponibles difieren según la variable de interés sea cuantitativa (se registre en una escala numérica) o cualitativa (se registre en un conjunto de clases o categorías). En el resto de este capítulo presentaremos estas alternativas.

Diferentes tipos de variables

Una variable es una característica o propiedad que difiere entre los elementos de un conjunto. Las variables se dividen en cuantitativas y cualitativas o categóricas. Las variables cuantitativas son las que se registran como números cuyas diferencias adquieren igual significado en cualquier parte de la escala. A su vez, las variables cuantitativas pueden ser continuas o discretas. Las continuas son aquellas que pueden tomar incontables valores en cualquier intervalo entre dos de sus valores posibles. Las discretas son las que se registran contando y por eso toman valores que pertenecen a una secuencia que se corresponde con alguna serie de números naturales. Las variables cualitativas o categóricas son las que se registran como clases de un repertorio definido. Estas variables pueden ser nominales, si las clases en que se registran no tienen un orden natural, u ordinales, si las clases tienen un orden natural. Las clases de una variable ordinal pueden identificarse con números pero las diferencias entre ellos no miden las diferencias entre las clases que representan.

Variables		Ejemplos
Cuantitativas:	**continuas**	Altura, peso, velocidad, temperatura, pH.
	discretas	Número de hojas, número de semillas.
Categóricas:	**nominales**	Sentido de inclinación (este, oeste, norte, sur, ninguno).
	ordinales	Calidad (1=pobre, 2=aceptable, 3=buena).

Distribuciones de frecuencias de variables cuantitativas

Tablas de frecuencias

La distribución de frecuencias de una variable cuantitativa se comunica con mínimo resumen en una tabla con la lista completa y ordenada de los valores de la variable acompañados por sus correspondientes frecuencias (Cuadro 1.3).

Cuadro 1.3. Distribución de frecuencias de las alturas de 40 plantas de girasol de una parcela con densidad de 5 plantas por m² (*fa*, frecuencia absoluta, *fr*, frecuencia relativa, *faa*, frecuencia absoluta acumulada, *fra*, frecuencia relativa acumulada).

Altura (cm)	*fa*	*fr*	*faa*	*fra*	Altura (cm)	*fa*	*fr*	*faa*	*fra*
193	1	0,025	1	0,025	213	3	0,075	21	0,525
197	1	0,025	2	0,05	214	1	0,025	22	0,55
198	1	0,025	3	0,075	215	5	0,125	27	0,675
199	1	0,025	4	0,1	216	1	0,025	28	0,7
201	1	0,025	5	0,125	217	2	0,05	30	0,75
202	1	0,025	6	0,15	219	2	0,05	32	0,8
204	1	0,025	7	0,175	220	1	0,025	33	0,825
205	2	0,05	9	0,225	221	3	0,075	36	0,9
207	1	0,025	10	0,25	222	1	0,025	37	0,925
208	3	0,075	13	0,325	224	1	0,025	38	0,95
209	1	0,025	14	0,35	225	1	0,025	39	0,975
210	2	0,05	16	0,40	226	1	0,025	40	1,00
212	2	0,05	18	0,45					

Además de las frecuencias absolutas y relativas de cada valor, en la tabla del Cuadro 1.3 se introducen las **frecuencias acumuladas** (absolutas y relativas). En este caso, se trata de las frecuencias de plantas con altura menor o igual que cada valor. En la tabla leemos, por ejemplo, que la **frecuencia absoluta acumulada** hasta la altura 220 cm es 33. Esto significa que entre las 40 plantas medidas, 33 tenían alturas ≤ 220 cm. Coincidentemente, se lee que la correspondiente **frecuencia relativa acumulada** es 0,825, el cociente entre 33 y 40.

En muchos casos, las tablas como la del Cuadro 1.3 resultan demasiado largas para presentar eficientemente los principales rasgos de una distribución de frecuencias. Por eso, es habitual resumir las tablas de frecuencias para variables cuantitativas mediante el recurso de dividir su escala en un número limitado de intervalos o clases (Cuadro 1.4).

Cuadro 1.4. Distribución de frecuencias de clases de altura de 40 plantas de girasol de una parcela con densidad de 5 plantas por m² (*fa*, frecuencia absoluta, *fr*, frecuencia relativa, *faa*, frecuencia absoluta acumulada, *fra*, frecuencia relativa acumulada).

Clase de altura (cm)	Marca de clase (cm)	*fa*	*fr*	*faa*	*fra*
(190, 195]	192,5	1	0,025	1	0,025
(195, 200]	197,5	3	0,075	4	0,1
(200, 205]	202,5	5	0,125	9	0,225
(205, 210]	207,5	7	0,175	16	0,4
(210, 215]	212,5	11	0,275	27	0,675
(215, 220]	217,5	6	0,15	33	0,825
(220, 225]	222,5	6	0,15	39	0,975
(225, 230]	227,5	1	0,025	40	1,00

Para construir la tabla de frecuencias del Cuadro 1.4, dividimos la escala de alturas de las plantas en **clases** o intervalos de 5 cm, abiertos a la izquierda y cerrados a la derecha. Por ejemplo, el intervalo (195, 200] incluye todos los valores de altura > 195 y ≤ 200 cm. Los extremos de cada intervalo se denominan **límites de clase** (inferior y superior) y sus puntos medios se denominan **marcas de clase**. La tabla presenta las frecuencias absolutas y relativas y las frecuencias acumuladas correspondientes a cada clase. Notemos que estas frecuencias concuerdan con las que aparecen en la tabla del Cuadro 1.3.

Gráficos de frecuencias

Las representaciones gráficas ayudan a visualizar rápidamente los principales rasgos de una distribución de frecuencias y compararlos con los de otras. Aquí describimos los tipos más comunes de gráficos de frecuencias (histogramas, polígonos, gráficos de líneas verticales, gráficos de caja y bigotes).

Histogramas

Un histograma representa la distribución de frecuencias (relativas o absolutas) de una variable continua cuya escala es dividida en clases (ver Figura 1.1). Para construirlo se dibuja un eje horizontal con la escala de la variable en cuestión y se marcan los límites de clase. Luego, tomando como base el segmento entre los límites de cada clase, se dibujan rectángulos de altura proporcional a la frecuencia de la clase correspondiente. Notemos que el histograma de la Figura 1.1 presenta una transcripción directa de la distribución de frecuencias relativas de clases que presenta la tabla del Cuadro 1.4.

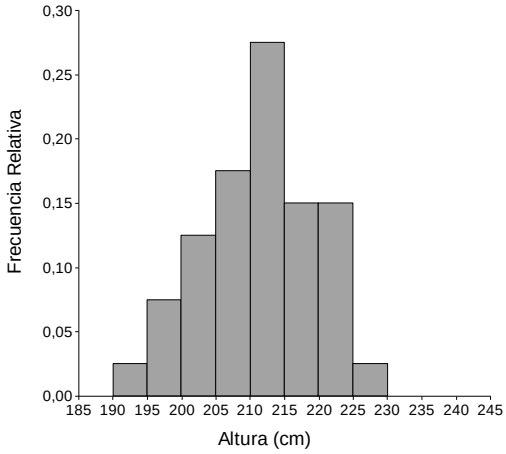


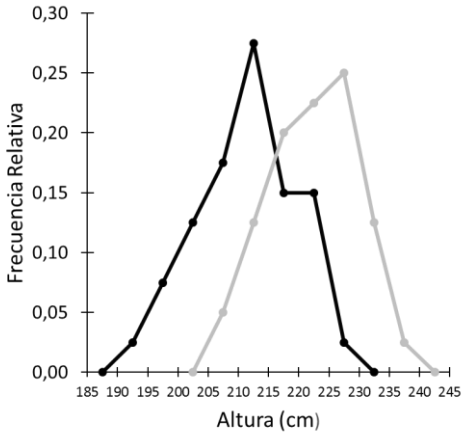
Figura 1.1. Histograma de frecuencias relativas de las alturas de 40 plantas de girasol medidas en una parcela con densidad baja (5 plantas por m²).

Este gráfico permite visualizar rápidamente la posición y la dispersión de las alturas de las plantas así como la distribución de las frecuencias entre las diferentes clases de altura.

Polígonos

Como los histogramas, los polígonos permiten representar la distribución de frecuencias (relativas o absolutas) de variables continuas con escalas divididas en clases (Figura 1.2). Para construir estos gráficos se dibuja un eje horizontal como el de un histograma y se señalan las marcas de clase. Luego se dibuja un punto *sobre cada marca de clase* a altura proporcional a la correspondiente frecuencia de clase. Por último, se unen los puntos consecutivos con líneas rectas. Notar que las ordenadas de los puntos negros de la Figura 1.2 coinciden con las alturas de los rectángulos del histograma de la Figura 1.1.

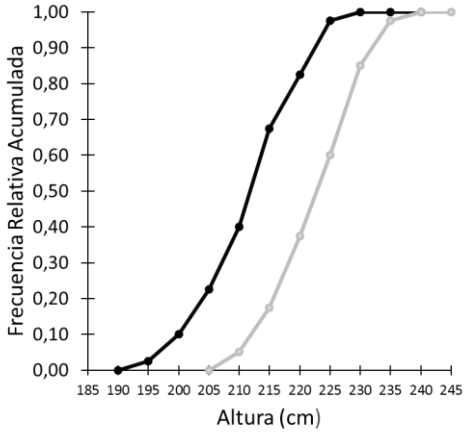
Figura 1.2. Polígonos de distribución de frecuencias relativas de las alturas de 40 plantas de girasol. Línea negra, parcela con densidad baja (5 plantas por m²). Línea gris, parcela con densidad alta (10 plantas por m²).



Estos gráficos son muy útiles para comparar dos o más distribuciones de frecuencias relativas. En la Figura 1.2 la distribución de frecuencias de las alturas de las plantas de la parcela con densidad alta aparece desplazada hacia la derecha respecto de la de las plantas de la parcela con densidad baja. Es decir que en la parcela con 10 plantas por m² las plantas más bajas eran menos frecuentes y las más altas más frecuentes que en la parcela con 5 plantas por m². Además, el gráfico muestra que las alturas mínima y máxima fueron menores en la parcela con densidad baja que en la parcela con densidad alta aunque la variabilidad de las alturas fue similar en ambas parcelas. Las diferencias que muestra la Figura 1.2 constituyen un indicio de que la altura de las plantas es una característica plástica frente a la diferencia en la densidad del cultivo.

Los polígonos son también apropiados para visualizar y comparar frecuencias acumuladas. Una alternativa es construirlos a partir de las frecuencias acumuladas por clases (Figura 1.3). En ese caso, se dibuja un punto *sobre el límite superior* de cada clase a altura proporcional a la frecuencia acumulada hasta dicha clase inclusive. Luego se unen los puntos consecutivos con líneas. La Figura 1.3 muestra que la frecuencia relativa de plantas con alturas menores al límite superior de cualquier clase fue mayor en la parcela con 5 que en la parcela con 10 plantas por m².

Figura 1.3. Polígonos de distribución de frecuencias relativas acumuladas de las alturas de 40 plantas de girasol construidos a partir de las frecuencias de clases de 5 cm. Línea negra, parcela con densidad baja (5 plantas por m²). Línea gris, parcela con densidad alta (10 plantas por m²).



Alternativamente, un polígono de frecuencias acumuladas se puede construir con mayor detalle a partir de las frecuencias de los valores individuales de la variable. En este caso, dibujamos el polígono como una función escalonada que se mantiene constante entre valores registrados consecutivos y aumenta verticalmente para cada uno en la medida de su frecuencia (Figura 1.4). Este gráfico es especialmente potente para identificar y comparar los valores de una variable que corresponden a cada frecuencia relativa acumulada (veremos más adelante que esos valores se denominan cuantiles o percentiles).

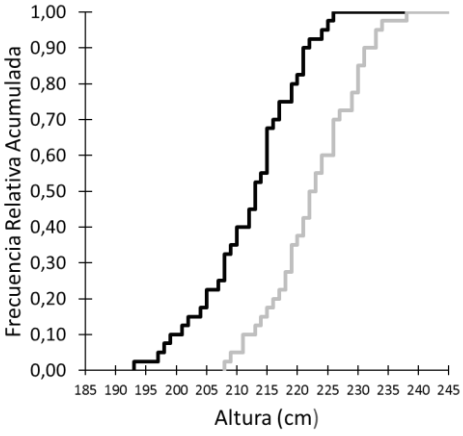


Figura 1.4. Polígonos de distribución de frecuencias relativas acumuladas de las alturas de 40 plantas de girasol construidos a partir de las mediciones individuales. Línea negra, parcela con densidad baja (5 plantas por m²). Línea gris, parcela con densidad alta (10 plantas por m²).

Gráficos de líneas verticales

Este tipo de gráfico se utiliza para representar la distribución de frecuencias de una variable cuantitativa discreta (Figura 1.5). Como los valores de estas variables son números naturales (o se pueden hacer corresponder con números naturales) su distribución de frecuencias se representa mediante líneas verticales dibujadas sobre cada valor con longitud proporcional a su frecuencia.

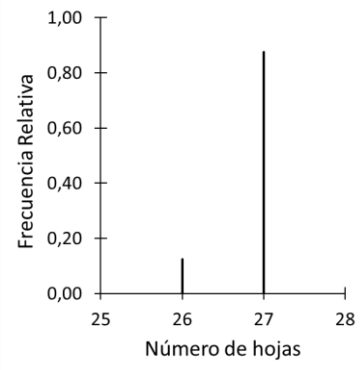


Figura 1.5. Frecuencias relativas de los números de hojas registrados en 40 plantas de girasol de la parcela con baja densidad (5 plantas por m²).

La Figura 1.5 refleja muy escasa variabilidad (gran homogeneidad) en los registros de números de hojas. Una consulta con la investigadora responsable de estas parcelas permitió concluir que esa variabilidad en los registros se debió exclusivamente a errores en los recuentos, porque todas las plantas de girasol de un cultivo producen sus hojas simultáneamente en respuesta a factores como la temperatura y la duración del día.

Posición y dispersión de una distribución de frecuencias

La posición de la distribución de frecuencias de una variable cuantitativa refleja las magnitudes de sus valores mientras la dispersión refleja su variabilidad.

Los polígonos de la Figura 1.2 muestran que las distribuciones de frecuencias de alturas de plantas de las dos parcelas tienen diferente **posición**, una está desplazada hacia la derecha de la otra. Tal diferencia de posición refleja que, tomadas en conjunto, las plantas fueron más altas en una parcela que en la otra. Además, los polígonos muestran que las distribuciones tienen similar **dispersión**. Esto indica que en una parcela la altura no fue mucho más variable o heterogénea entre plantas que en la otra parcela.

La posición y la dispersión de una distribución de frecuencias a lo largo del eje de una variable cuantitativa se evalúan numéricamente mediante **medidas de posición** y **medidas de dispersión**. Estas medidas son resúmenes muy condensados cuya importancia radica en su utilidad para realizar comparaciones cuantitativas entre dos o más distribuciones de frecuencias.

Medidas de posición para variables cuantitativas

Moda

La moda es el valor más frecuente.

La **moda** de una variable se define como el valor que tiene la máxima frecuencia. Por ejemplo, en la tabla del Cuadro 1.3 encontramos que entre las alturas de las 40 plantas de girasol que los estudiantes midieron en la parcela con densidad baja el valor más frecuente fue 215 cm. Es decir que la moda de las alturas de dichas plantas es 215 cm. Como no necesariamente un valor es más frecuente que todos los demás, algunas distribuciones de frecuencias son bimodales o polimodales o, cuando todos los valores tienen igual frecuencia, carecen de moda.

Cuando la escala de una variable se divide en clases, podemos definir la **clase modal** como aquella que reúne la máxima frecuencia. Por ejemplo, en el histograma de la Figura 1.1 se lee que la clase modal de las alturas de las plantas de girasol medidas en la parcela con densidad baja fue el intervalo (210, 215 cm].

Cuantiles, percentiles, cuartiles y mediana

Un cuantil es un valor de una variable al cual corresponde una determinada frecuencia relativa acumulada.

El **cuantil** α de una distribución de frecuencias es un valor de la variable al cual corresponde la frecuencia relativa acumulada $fra = \alpha$. Por ejemplo, el cuantil 0,15 es un valor de la variable al cual corresponde la frecuencia relativa acumulada $fra = 0{,}15$. Es común referirse a **percentiles** que no son otra cosa que los cuantiles identificados por el valor de α expresado en porcentaje. Por ejemplo, en lugar de cuantil 0,15 podemos decir percentil 15.

Los cuantiles 0,25, 0,50 y 0,75 se denominan respectivamente primer **cuartil**, segundo cuartil o **mediana** y tercer cuartil. En la Figura 1.6, los tres cuartiles de la distribución de frecuencias de las alturas de las 40 plantas de la parcela con densidad baja están señalados sobre el gráfico de frecuencias acumuladas. Podemos leer que un cuarto de las mediciones fueron $\leq 207{,}5$ cm, la mitad de fueron ≤ 213 cm y tres cuartos fueron ≤ 218 cm. Notemos que los cuartiles (como cualquier cuantil) se representan en esta figura sobre el eje horizontal y *se miden en la unidad de la variable de interés*. La **mediana** es el valor que corresponde a la mitad de la distribución de frecuencias. Por eso decimos que la mediana es una **medida de posición central**.

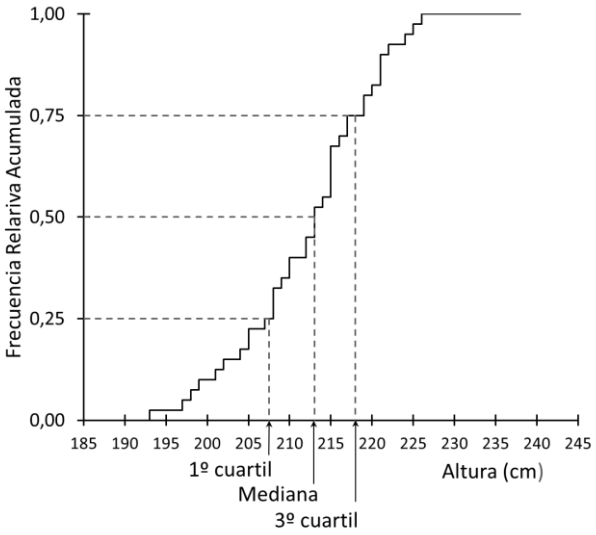


Figura 1.6. Polígono de distribución de frecuencias relativas acumuladas de las alturas de 40 plantas de girasol de la parcela con densidad baja (5 plantas por m²). Se señala la ubicación del primer cuartil (207,5 cm), la mediana (213 cm) y el tercer cuartil (218 cm).

Notemos que definimos un cuantil como ***un*** valor (no *el* valor) de la variable en cuestión al cual corresponde una cierta *fra*. La razón para ello es que puede haber más de un valor que cumpla con tal condición. Por ejemplo, la Figura 1.6 muestra que, la frecuencia relativa acumulada hasta cualquier valor del intervalo [217 cm, 219 cm) fue $fra = 0{,}75$. Por convención, el valor que asignamos al tercer cuartil es el punto medio de dicho intervalo. Análogamente, asignamos al primer cuartil el punto medio del intervalo donde $fra = 0{,}25$. En este caso, no fue necesario aplicar una convención para evaluar la mediana porque la $fra = 0{,}5$ correspondió únicamente a la altura 213 cm.

Los así llamados **gráficos de caja y bigotes** (*box plots*) resumen toda la distribución de frecuencias a partir de unos pocos cuantiles (Figura 1.7). En estos gráficos, los bordes de la caja indican el primer y tercer cuartil, la línea horizontal que corta la caja indica la mediana mientras los extremos de los bigotes indican el mínimo y el máximo de la variable (o dos percentiles como p.ej. el 5 y el 95). Los valores de todas estas medidas de posición se leen sobre el eje vertical del gráfico.

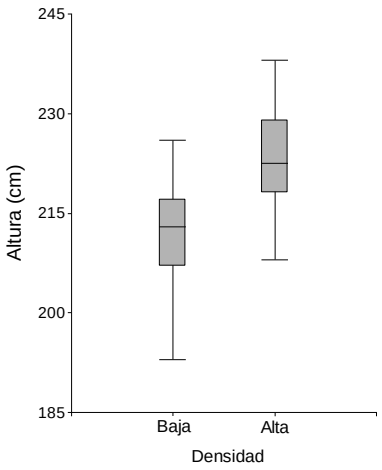


Figura 1.7. Gráfico de caja y bigotes (*box plot*) de las distribuciones de frecuencias de alturas de 40 plantas de girasol de la parcela con densidad baja (5 plantas por m²) y de la parcela con densidad alta (10 plantas por m²). Las cajas muestran los intervalos entre los valores del primer y tercer cuartil, las líneas que las atraviesan señalan el correspondiente valor de la mediana y en este caso los extremos de los bigotes señalan los valores máximos y mínimos.

Notación sumatoria

Muchos cálculos estadísticos se basan en sumas que pueden ser muy largas. Por eso, conviene utilizar una notación que permita escribirlas de manera sintética. Esta es la notación **sumatoria** que presentamos aquí mediante un ejemplo sencillo.

Consideremos una lista de números organizada como sigue:

i	1	2	3	4	5
x_i	1,9	3,0	1,5	3,0	2,5

La fila inferior contiene valores de una variable de interés denominada x. En la fila superior están los valores de una variable indicadora i que identifica la posición en la lista de cada valor de x consignado en la fila inferior. La variable indicadora aparece como **subíndice** en la expresión x_i que encabeza la fila inferior de modo que,

$$x_1 = 1{,}9, \;\; x_2 = 3{,}0, \;\; x_3 = 1{,}5, \ldots \text{ etc.}$$

La suma de los 5 valores de x es,

$$x_1 + x_2 + x_3 + x_4 + x_5 = 11{,}9$$

Mediante la notación sumatoria, escribimos esta suma como,

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

El término de la izquierda se lee como la suma de los valores de una variable llamada x cuyas posiciones en una lista son las que corresponden a los valores del subíndice i que van desde el *límite inferior* consignado debajo de la letra $\sum$ hasta el *límite superior* consignado encima de ella. Esta notación permite escribir en poco espacio sumas de un número arbitrariamente largo de sumandos. Por ejemplo,

$$\sum_{j=1}^{n} v_j$$

denota la suma de los primeros n términos de una lista de valores de una variable denominada v. Notemos que podemos usar cualquier letra para identificar el subíndice.

Media aritmética

La media aritmética es lo que vulgarmente llamamos *promedio*; se trata de la medida de posición central más importante en la inferencia estadística.

La **media aritmética** de un conjunto de valores de una variable se calcula como el cociente entre la suma y el número de dichos valores. Denotamos la media aritmética con una línea horizontal sobre la letra que identifica a la variable.

Definición:
La **media aritmética** de una variable *x* que toma *n* valores $x_1, x_2, \ldots, x_n$ es,

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad (1.1)$$

Consideremos por ejemplo los siguientes cinco valores de la variable x:

$$x_1 = 1{,}9, \quad x_2 = 3{,}0, \quad x_3 = 1{,}5, \quad x_4 = 3{,}0, \quad x_5 = 2{,}5$$

Aplicando la fórmula 1.1, calculamos su media aritmética como,

$$\begin{aligned} \bar{x} &= \frac{1}{5} \cdot \sum_{i=1}^{5} x_i \\ &= \frac{1}{5} \cdot (1{,}9 + 3{,}0 + 1{,}5 + 3{,}0 + 2{,}5) \\ &= \frac{1}{5} \cdot 11{,}9 \\ &= 2{,}38 \end{aligned}$$

La media aritmética informa *únicamente* sobre la posición central de la distribución de frecuencias sin dar idea alguna sobre su dispersión. Esto es así porque la misma suma se puede alcanzar tanto con valores que sean muy cercanos entre sí como con valores que sean muy distantes entre sí. Comparemos, por ejemplo, la media aritmética $\bar{x} = 2{,}38$ recién calculada con la media aritmética de los siguientes valores de la variable *v*,

$$v_1 = 2{,}2, \quad v_2 = 2{,}5, \quad v_3 = 2{,}5, \quad v_4 = 2{,}2, \quad v_5 = 2{,}5$$

Aplicando la fórmula 1.1 ¡encontramos que $\bar{v} = 2{,}38$! La medias aritméticas $\bar{x}$ y $\bar{v}$ son indistinguibles. Si bien los valores de *x* están más dispersos que los de *v*, ambos conjuntos están distribuidos alrededor del mismo valor de la media aritmética (Figura 1.8).

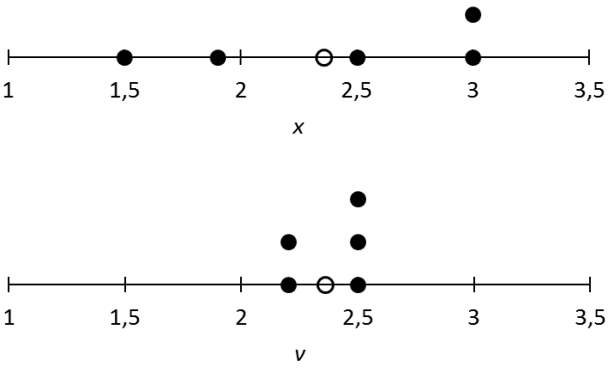


Figura 1.8. Distribuciones de cinco valores de dos variables *x* y *v*. Los puntos negros indican los valores individuales y los círculos vacíos el valor común de sus medias aritméticas.

La media aritmética se puede expresar en términos de los valores diferentes de la variable y sus correspondientes frecuencias relativas.

Cuando el conjunto de valores a promediar incluye valores repetidos, es cómodo calcular la media aritmética teniendo en cuenta sus frecuencias. Por ejemplo, para obtener $\bar{v}$ con la fórmula 1.1 hemos debido calcular,

$$\bar{v} = \frac{1}{5} \cdot (2{,}2 + 2{,}5 + 2{,}5 + 2{,}2 + 2{,}5)$$

Si reordenamos los sumandos, este mismo cálculo se puede hacer como,

$$\begin{aligned} \bar{v} &= \frac{1}{5} \cdot (2{,}2 + 2{,}2 + 2{,}5 + 2{,}5 + 2{,}5) \\ &= \frac{1}{5} \cdot (2 \cdot 2{,}2 + 3 \cdot 2{,}5) \\ &= 0{,}4 \cdot 2{,}2 + 0{,}6 \cdot 2{,}5 \end{aligned}$$

Como 0,4 y 0,6 son las *fr* de 2,2 y 2,5 respectivamente, vemos que la media aritmética se puede calcular como la suma de los valores *diferentes* de la variable multiplicados por sus correspondientes frecuencias relativas,

Definición alternativa:

Dada una variable *v* que toma *k* valores *diferentes* v_1, v_2, …, v_k cada uno con su correspondiente frecuencia relativa fr_1, fr_2, …, fr_k, la **media aritmética** de *v* es,

$$\bar{v} = \sum_{j=1}^{k} v_j \cdot fr_j \qquad (1.2)$$

Es muy importante notar que la sumatoria de la fórmula 1.2 sólo incluye *k* sumandos correspondientes a los *k* números diferentes. En cambio, la sumatoria de la fórmula 1.1 incluye *n* sumandos correspondientes a los *n* registros de la variable, aunque sean números repetidos.

La media aritmética de una variable cuya escala está dividida en clases se puede calcular aproximadamente a partir de las frecuencias relativas de clase.

La tabla de frecuencias de una variable continua resumida en clases o intervalos no contiene toda la información necesaria para calcular el valor exacto de la media aritmética (Cuadro 1.4). Sin embargo, a partir de esta tabla se puede aproximar el valor de la media aritmética como la suma de los productos de las marcas de clase por las frecuencias relativas de clase. Por ejemplo, con la información de la tabla del Cuadro 1.4 se obtiene el siguiente valor aproximado de la media aritmética de las alturas de las 40 plantas de girasol de la parcela con densidad baja.

$$\begin{aligned} \bar{h} &\approx \sum_{l=1}^{8} mc_l \cdot fr_l \\ &\approx 192{,}5\ cm \cdot 0{,}025 + 197{,}5\ cm \cdot 0{,}075 + \cdots + 227{,}5\ cm \cdot 0{,}025 \\ &\approx 211{,}375\ cm \end{aligned}$$

Para evaluar la calidad de esta aproximación la comparamos con el valor exacto de la media aritmética calculado con la fórmula 1.2 a partir de la información completa contenida en la tabla del Cuadro 1.3. Este valor es $\bar{h} = 212{,}1\ cm$.

Propiedades de la media aritmética

La media aritmética tiene dos propiedades importantes para su aplicación en numerosos procedimientos de inferencia estadística.

Propiedad I

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad (1.3)$$

Las cantidades $(x_i - \bar{x})$ se denominan **desvíos**. El valor de la media aritmética es tal que la suma de los desvíos es nula. Los desvíos positivos compensan a los desvíos negativos y la media queda, justamente, *en el medio* de los diferentes valores[2].

Propiedad II

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \sum_{i=1}^{n}(x_i - c)^2\,, \quad \text{para todo } c \qquad (1.4)$$

La suma de los cuadrados de los desvíos es la mínima suma de cuadrados de las diferencias entre los valores x_i y cualquier número real. Para interpretar esta propiedad notemos que el cuadrado del desvío es una medida de la distancia entre x_i y $\bar{x}$ (llamaremos a esta medida *distancia cuadrática*). Es decir que la suma de las distancias cuadráticas desde los valores x_i hasta la media aritmética es menor que hasta cualquier otro número. En este sentido, la media aritmética es el número que en promedio queda *más cercano* a todos los valores de la variable[3].

Medidas de dispersión de variables cuantitativas

Amplitud total y amplitud entre cuartiles

La **amplitud total** (o rango total) de una distribución de frecuencias es la diferencia entre los valores máximo y mínimo de la variable. Por ejemplo, para la distribución de frecuencias de las alturas de 40 plantas de girasol de la parcela con densidad baja,

Amplitud total = 226 cm – 193 cm = 33 cm (ver Cuadro 1.3)

Una medida alternativa es la **amplitud entre cuartiles** (o rango intercuartilar) que se calcula como la diferencia entre los valores del tercer y del primer cuartil.

Amplitud entre cuartiles = 218 cm – 207,5 cm = 10,5 cm (ver Figura 1.6)

Tanto la amplitud total como la amplitud entre cuartiles son rápidamente apreciables en los gráficos de caja y bigotes (Figura 1.7).

---

[2] Esta propiedad se demuestra aplicando primero las propiedades asociativa y conmutativa de la suma (p.ej. $a + k + b + k + c + k = [a + b + c] + 3 \cdot k$) y luego la definición 1.1 de media aritmética.

[3] Si calculamos las derivadas primera y segunda de la suma de cuadrados de $(x_i - c)$ con respecto a c, podemos comprobar que $c = \bar{x}$ corresponde a un mínimo de dicha suma porque la primera derivada es igual a 0 y la segunda derivada es mayor que cero. Para demostrar que ese mínimo es el único, mostramos que dado cualquier valor c, la suma de cuadrados de $(x_i - c)$ es $\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n \cdot (\bar{x} - c)^2$.

Varianza

La varianza es el promedio de los cuadrados de los desvíos.

La **varianza** (o variancia) de un conjunto de valores de una variable se define como la media aritmética de los cuadrados de los desvíos.

Definición:

La **varianza** del conjunto de n valores $x_1, x_2, \ldots, x_n$ de la variable x es,

$$var(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.5)$$

donde $\bar{x}$ es la media aritmética de los n valores. Notar que la varianza es una magnitud ≥ 0.

La varianza es el promedio de las distancias cuadráticas entre los valores y su media aritmética.

Dado que los cuadrados de los desvíos miden las distancias cuadráticas entre los valores de la variable y su media aritmética, la varianza es el promedio de dichas distancias. Cuanto más distantes de la media quedan los valores de una variable mayor es su varianza. Por ejemplo, la Figura 1.8, muestra que los valores la variable x, que están más dispersos que los de la variable v, aparecen más distantes de la media aritmética. Si aplicamos la fórmula 1.5 para calcular la varianza de x y la varianza de v obtenemos,

$$\begin{aligned} var(x) &= \frac{1}{5} \cdot \sum_{i=1}^{5} (x_i - \bar{x})^2 \\ &= \frac{1}{5} \cdot [\,(1{,}9 - 2{,}38)^2 + (3{,}0 - 2{,}38)^2 + \cdots + (2{,}5 - 2{,}38)^2] \\ &= 0{,}3576 \end{aligned}$$

y,

$$var(v) = \frac{1}{5} \cdot \sum_{i=1}^{5} (v_i - \bar{v})^2 = 0{,}0216$$

Estos diferentes valores de varianza reflejan la diferencia entre la dispersión de los valores de x y de los valores de v que se puede apreciar en la Figura 1.8.

La varianza se puede definir en términos de los desvíos de la variable y las frecuencias relativas de los correspondientes valores.

Dado que la varianza es la media aritmética de los cuadrados de los desvíos, se puede calcular como la suma de los cuadrados de los desvíos correspondientes a cada valor *diferente* de la variable multiplicados por las correspondientes frecuencias relativas

Definición alternativa:

Dada una variable v que tiene k valores *diferentes* $v_1, v_2, \ldots, v_k$ cada uno con su correspondiente frecuencia relativa $fr_1, fr_2, \ldots, fr_k$, la **varianza** de v es,

$$var(v) = \sum_{j=1}^{k} \left(v_j - \bar{v}\right)^2 \cdot fr_j \qquad (1.6)$$

A partir de la tabla de frecuencias del Cuadro 1.3 aplicamos la fórmula 1.6 para calcular el valor exacto de la varianza de las alturas de las 40 plantas que los estudiantes midieron en la parcela con densidad baja (5 plantas por m²).

$$\begin{aligned} var(h) &= \sum_{j=1}^{25} \left(h_j - \bar{h}\right)^2 \cdot fr_j \\ &= (193\ cm - 212{,}1\ cm)^2 \cdot 0{,}025 + \cdots + \\ &\qquad + (226\ cm - 212{,}1\ cm)^2 \cdot 0{,}025 \\ &= 63{,}89\ cm^2 \end{aligned}$$

Alternativamente, obtenemos una aproximación al valor de la varianza a partir de la tabla de frecuencias resumida en clases de altura del Cuadro 1.4.

$$\begin{aligned} var(h) &\approx \sum_{l=1}^{8} \left(mc_l - \bar{h}\right)^2 \cdot fr_l \\ &\approx (192{,}5\ cm - 211{,}375\ cm)^2 \cdot 0{,}025 + \cdots + \\ &\qquad + (227{,}5\ cm - 211{,}375\ cm)^2 \cdot 0{,}025 \\ &\approx 66{,}86\ cm^2 \end{aligned}$$

La varianza se expresa en la unidad de la variable elevada al cuadrado.

Mientras las alturas de las plantas se expresan en una escala graduada en cm, la varianza mide su dispersión a lo largo de dicha escala en cm². Es decir que las unidades de la varianza no son las mismas que las de la variable cuya dispersión mide (!). Esto es así porque la varianza es el promedio de los desvíos *elevados al cuadrado*. Para obtener una medida de dispersión más fácilmente interpretable se calcula la raíz cuadrada de la varianza o desvío estándar.

Desvío estándar

El desvío estándar es la raíz cuadrada de la varianza.

El **desvío estándar** se define como la raíz cuadrada de la varianza. De esta forma se obtiene una medida de dispersión emparentada con la varianza pero cuya unidad es la misma que la de la variable.

Definición:

El **desvío estándar** de una variable x que toma n valores $x_1, x_2, \ldots, x_n$ con varianza $var(x)$ es,

$$de(x) = \sqrt{var(x)} \qquad (1.7)$$

Por ejemplo, el desvío estándar de las alturas de las 40 plantas de girasol medidas en la parcela con densidad baja fue,

$$\begin{aligned} de(h) &= \sqrt{var(h)} \\ &= \sqrt{63{,}89\ cm^2} \\ &= 7{,}99\ cm \end{aligned}$$

Coeficiente de variación

El coeficiente de variación es el cociente entre el desvío estándar y la media aritmética.

En ocasiones interesa evaluar la dispersión de una variable en comparación con su media. Por ejemplo, el desvío estándar de las alturas de las 40 plantas de girasol de la parcela con densidad baja fue 7,99 cm. Este valor de desvío estándar parece más importante en este caso en que la media aritmética fue 212,1 cm que si correspondiera, por ejemplo, a 40 árboles de *Eucalyptus* sp. con altura media de 2514,6 cm. Para este tipo de ocasiones, el **coeficiente de variación** provee una medida de dispersión relativa al valor de la media aritmética.

Definición:

El **coeficiente de variación** de una variable x que toma n valores $x_1, x_2, \ldots, x_n$ con media aritmética $\bar{x}$ y desvío estándar $de(x)$ es,

$$cv(x) = \frac{de(x)}{\bar{x}} \qquad (1.8)$$

Aplicando la fórmula 1.8, calculamos el valor del coeficiente de variación de las alturas de las 40 plantas de girasol medidas en la parcela con densidad baja.

$$cv(h) = \frac{7{,}99\ cm}{212{,}1\ cm} = 0{,}0377 \quad \text{(o 3,77 \%)}$$

Media y varianza de funciones lineales

La media aritmética y la varianza tienen una serie de propiedades importantes referidas a funciones lineales de las variables.

Si definimos la variable $u = x + a$, donde x es una variable cuantitativa y a es una constante,

$$\bar{u} = \bar{x} + a \qquad (1.9)$$

- La media de la suma de una variable más una constante es igual a la suma de la media de la variable más la constante.

$$var(u) = var(x) \qquad (1.10)$$

- La varianza de la suma de una variable más una constante es igual a la varianza de la variable.

Si definimos la variable $v = b \cdot x$, donde b es una constante y x es una variable cuantitativa,

$$\bar{v} = b \cdot \bar{x} \qquad (1.11)$$

- La media del producto de una constante por una variable es igual al producto de la constante por la media de la variable.

$$var(v) = b^2 \cdot var(x) \qquad (1.12)$$

- La varianza del producto de una constante por una variable es igual al producto del *cuadrado* de la constante por la varianza de la variable.

Si definimos la variable $w = x + y$, donde x e y son dos variables cuantitativas,

$$\bar{w} = \bar{x} + \bar{y} \tag{1.13}$$

- La media de la suma de dos variables es igual a la suma de sus medias.

$$var(w) = var(x) + var(y) + 2 \cdot \left[\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})\right] \tag{1.14}$$

- La varianza de la suma de dos variables es igual a la suma de sus varianzas más un valor igual al doble de su **covarianza** (que puede ser negativa).

Covarianza

La covarianza mide la asociación lineal entre dos variables.

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de dos variables cuantitativas x e y, definimos la covarianza entre dichas variables como,

$$cova(x, y) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{1.15}$$

La covarianza puede ser mayor, igual o menor que cero. Esto depende de cómo sea la asociación lineal entre las dos variables. La covarianza es positiva si y generalmente aumenta cuando aumenta x, es negativa si y generalmente disminuye cuando aumenta x y es nula si no hay asociación de este tipo (lineal).

Distribuciones de frecuencias de variables categóricas

La distribución de frecuencias de una variable categórica se caracteriza por la identidad de las clases, por las frecuencias con la que cae en cada una y por la diversidad de las clases en las que cae. Para describir estas distribuciones se utilizan tablas, gráficos, la clase modal y los índices de diversidad.

Tabla de frecuencias

Las tablas de frecuencias para variables categóricas especifican las frecuencias absolutas o las frecuencias relativas de cada clase o categoría de la variable. Al comienzo de este capítulo presentamos y comparamos las tablas de frecuencias de los sentidos de inclinación de los tallos de las plantas de girasol de dos parcelas experimentales una con baja y otra con alta densidad de plantas (Cuadros 1.1 y 1.2).

Gráfico de barras

Los **gráficos de barras** son muy usados para comunicar diversos resultados estadísticos. Cuando los construimos para representar distribuciones de frecuencias de variables categóricas, disponemos las diferentes clases o categorías sobre el eje horizontal y sobre cada una dibujamos una barra o rectángulo cuya altura indica la frecuencia de la clase medida sobre el eje vertical (Figura 1.9).

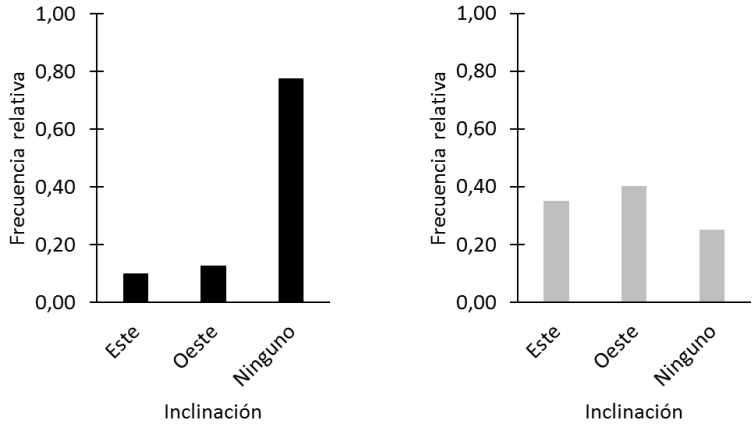


Figura 1.9. Distribuciones de frecuencias de los sentidos de inclinación de 40 plantas de girasol de dos parcelas experimentales cultivadas en hileras con dirección norte-sur. Barras negras, parcela con densidad baja (5 plantas por m²). Barras grises, parcela con densidad alta (10 plantas por m²)

Estos gráficos permiten apreciar y comparar visualmente las distribuciones de frecuencias de variables categóricas. La Figura 1.9 muestra que la distribución de frecuencias de los sentidos de inclinación en la parcela con densidad baja estuvo fuertemente concentrada en la clase *Ninguno* (plantas en posición vertical). En cambio, en la parcela con densidad alta esta clase es la que tuvo la mínima frecuencia relativa y la distribución estuvo muy poco concentrada.

Clase modal

El indicador de la posición de la distribución de frecuencias de una variable categórica es la **clase modal**. Se trata de la clase que tiene la máxima frecuencia. Como vimos, la clase modal se define también para distribuciones de variables cuantitativas cuya escala fue dividida en clases. Los gráficos de barras de la Figura 1.9 muestran que la clase modal de los sentidos de inclinación de las plantas de girasol fue *Ninguno* en la parcela con densidad baja y *Oeste* en la parcela con densidad alta. En este caso, las diferentes clases modales indican que las dos distribuciones tienen diferente posición.

Índices de diversidad

Los índices de diversidad miden la heterogeneidad entre los registros de la variable.

En las parcelas que examinaron por los estudiantes, la inclinación de las plantas era heterogénea porque algunas estaban en posición vertical, otras estaban inclinadas hacia el este y otras hacia el oeste (Figura 1.9). Esta heterogeneidad era menor en la parcela con densidad baja porque allí una gran mayoría de las plantas estaban en la misma clase de inclinación (Ninguna). Los índices de diversidad miden este tipo de heterogeneidad.

Número de clases (N0)

El índice más sencillo de la diversidad de una variable categórica es el **número de clases** con frecuencia > 0. Por ejemplo, en las dos parcelas de girasol que examinaron los estudiantes, el número registrado de sentidos de inclinación de los tallos fue 3 (Figura 1.9). Sin embargo, en la Figura 1.9 vimos que la inclinación de los tallos fue más homogénea en la parcela con baja densidad. Evidentemente, el número de clases resultó

una medida demasiado grosera para distinguir la diferencia de heterogeneidad que refleja la Figura 1.9. Para ello, es necesario utilizar medidas de la diversidad que dependen de las frecuencias relativas.

Exponencial del índice de Shannon (N1)

Una medida de diversidad de clases sensible a las diferencias entre las frecuencias relativas es el **exponencial del índice de Shannon (*N*1)**.

Definición:

Dada una variable categórica que cae en *m* clases diferentes con frecuencias relativas $fr_1, fr_2, \ldots, fr_m$, definimos el índice de diversidad ***N*1** como

$$N1 = exp\left[-\sum_{j=1}^{m} fr_j \cdot ln(fr_j)\right] \tag{1.16}$$

A partir de los valores de frecuencias relativas que aparecen en las tablas de los Cuadros 1.1 y 1.2, encontramos que los valores del índice de diversidad *N*1 fueron 1,989 para la parcela con densidad baja y 2,946 para la parcela con densidad alta. Estos valores reflejan correctamente que la diversidad de sentidos de inclinación de los tallos fue menor en la parcela con densidad baja que en la parcela con densidad alta (Figura 1.9). El índice *N*1 se puede interpretar como un *número equivalente* de clases, el número de clases de una variable con la misma heterogeneidad que la evaluada pero con igual frecuencia relativa en todas las clases.

Inversa del índice de Simpson (N2)

Una medida alternativa de diversidad de clases más sensible aún que el exponencial del índice de Shannon a las diferencias entre las frecuencias relativas es la **inversa del índice de Simpson (*N*2)**.

Definición:

Dada una variable categórica que cae en *m* clases diferentes con frecuencias relativas $fr_1, fr_2, \ldots, fr_m$, definimos el índice de diversidad ***N*2** como

$$N2 = \left[\sum_{j=1}^{m} {fr_j}^2\right]^{-1} \tag{1.17}$$

Según las frecuencias relativas que aparecen en las tablas de los Cuadros 1.1 y 1.2, los valores del índice de diversidad *N*2 de los sentidos de inclinación de las plantas examinadas fueron 1,597 para la parcela con densidad baja y 2,899 para la parcela con densidad alta. Debido a que *N*2 es más sensible que *N*1 a las diferencias en frecuencia relativa, los valores de *N*2 son menores que los de *N*1. Como el índice *N*1, el índice *N*2 se interpreta como un número equivalente de clases.

Cuestionario

1. ¿Qué significa *frecuencia*? ¿Qué es una *distribución de frecuencias*?
2. ¿A qué llamamos *frecuencia absoluta* y a qué *frecuencia relativa*? ¿Cuánto vale la suma de las frecuencias relativas?
3. ¿Cuáles son los diferentes tipos de variables? ¿Cómo se distingue cada uno?

4. ¿Qué opciones hay para representar la distribución de frecuencias de una variable cuantitativa en una *tabla*?
5. ¿A qué llamamos *frecuencia acumulada* de una variable cuantitativa?
6. ¿Cómo se representa la distribución de frecuencias de una variable cuantitativa en un *histograma*, en un *polígono*, en un *gráfico de líneas verticales* y en un *gráfico de caja y bigotes*?
7. ¿Cómo se interpretan la *posición* y la *dispersión* de la distribución de frecuencias de una variable cuantitativa?
8. ¿A qué se denomina *moda*?
9. ¿Cómo se definen un *cuantil* y un *percentil* de una distribución de frecuencias? ¿Qué unidad tienen?
10. ¿Qué son los *cuartiles* y la *mediana* de una distribución de frecuencias? ¿Qué unidad tienen?
11. ¿Qué significa la expresión $\sum_{i=1}^{n} x_i$?
12. ¿Cómo se define la *media aritmética* de un conjunto de valores de una variable cuantitativa? ¿Cómo se la calcula a partir de todos los valores y cómo se calcula a partir de los valores diferentes y sus correspondientes frecuencias relativas?
13. ¿A qué se llama *desvío*? ¿Cuánto vale la suma de los desvíos respecto de la media aritmética? ¿Qué propiedad tiene la suma de cuadrados de los desvíos?
14. ¿A qué se denomina *amplitud total* y a qué se denomina *amplitud entre cuartiles* de una distribución de frecuencias?
15. ¿Cómo se define la *varianza* de un conjunto de valores de una variable? ¿Qué valores puede tomar? ¿Cómo se interpreta?
16. ¿Cómo se calcula la varianza a partir de todos los valores de la variable y cómo se calcula a partir de los valores diferentes y sus frecuencias relativas?
17. ¿Cómo se definen el *desvío estándar* y el *coeficiente de variación*?
18. ¿Cuánto valen la media aritmética y la varianza de la suma de una variable más una constante?
19. ¿Cuánto valen la media aritmética y la varianza del producto de una variable por una constante?
20. ¿Cuánto valen la media aritmética y la varianza de la suma de dos variables?
21. ¿Cómo se define la *covarianza* entre dos variables? ¿En qué casos la covarianza es positiva, nula o negativa?
22. ¿Cómo se representa la distribución de frecuencias de una variable categórica en una *tabla* y en un *gráfico de barras*?
23. ¿Qué es la *clase modal* de una variable categórica?
24. ¿Cómo se definen el número de clases *N*0, el exponencial del índice de Shannon *N*1, y la inversa del índice de Simpson *N*2? ¿Qué miden estos índices? ¿En qué se diferencian?

Ejercicios

1.1 Las mediciones de altura de 40 plantas de girasol (*Helianthus annuus* L.) registradas por los estudiantes en la parcela con densidad alta (10 plantas por m²) se transcriben a continuación.

Planta	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Altura (cm)	229	234	218	216	219	211	223	238	233	219	227	224	226	222	209

Planta	*16*	*17*	*18*	*18*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
Altura (cm)	222	219	221	230	224	230	208	231	218	229	231	222	214	215	226

Planta	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*
Altura (cm)	211	226	223	226	217	221	220	230	213	233

a. *Organizar esta información en una tabla con las frecuencias absolutas, las frecuencias relativas y las frecuencias acumuladas (absolutas y relativas) correspondientes a cada valor registrado.*

b. *A partir de la tabla elaborada en a, identificar los valores de la moda, del mínimo, el máximo y los tres cuartiles de la distribución de frecuencias de las alturas de estas plantas y compararlos con los correspondientes a las plantas de la parcela con densidad baja (5 plantas por m²) presentados en el texto.*

c. *Calcular la media aritmética de las alturas de estas plantas y compararla con la de las 40 plantas medidas en la parcela con densidad baja.*

d. *Calcular la varianza, el desvío estándar y el coeficiente de variación de las alturas de estas plantas y comparar sus valores con los correspondientes a las 40 plantas medidas en la parcela con densidad baja.*

e. *¿Cómo se interpretan las comparaciones realizadas en b, c y d en relación con la posible plasticidad de la altura de las plantas frente a la diferencia en la densidad del cultivo entre las dos parcelas?*

1.2 Según registros publicados en meteored.com.ar, los valores de la temperatura mínima del mes de julio en el Aeropuerto de Concordia en el último cuarto del siglo veinte son los que se transcriben a continuación.

Año	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Temp. mínima julio (ºC)	0	− 4	2	1	0	− 7	− 2	2	1	− 3

Año	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Temp. mínima julio (ºC)	1	2	− 1	− 5	− 3	− 4	− 3	− 3	− 2	− 6

Año	1995	1996	1997	1998	1999
Temp. mínima julio (ºC)	− 2	− 1	− 3	2	− 1

a. *Con la información provista, elaborar una tabla de frecuencias (absolutas, relativas y acumuladas) de clases de temperatura mínima de julio. Dividir la escala de la variable en un número de clases apropiado para visualizar los principales rasgos de la distribución de frecuencias.*

b. *Confeccionar un histograma de frecuencias.*

c. *A partir de la tabla elaborada, calcular valores aproximados de la media, la varianza y el desvío estándar.*

d. *A partir de los datos originales, calcular los valores exactos de las medidas aproximadas en el punto c y evaluar los errores de aproximación.*

1.3 La Productividad Primaria Neta es la producción anual de biomasa de la vegetación. En sistemas ganaderos pastoriles se evalúa la Productividad Primaria Neta Aérea (PPNA) como una medida de la máxima producción de forraje aprovechable por los herbívoros. El histograma de la Figura 1.10 representa gráficamente la distribución de frecuencias de la PPNA anual ($g/m^2 \cdot$año) medida en 50 sitios del sur de la Pampa Deprimida donde la vegetación es una Pradera de Mesófitas.

a. *A partir de la información contenida en el histograma, construir un polígono de frecuencias y un polígono de frecuencias acumuladas por clases (recordar que las frecuencias acumuladas por clases se grafican sobre el límite superior de cada clase).*

b. *¿Qué valores aproximados tienen el primer cuartil, la mediana y el tercer cuartil de esta distribución de frecuencias?*

c. *¿Cuántos de los sitios representados en esta distribución de frecuencias tuvieron PPNA > 550 $g/m^2 \cdot$año y cuántos tuvieron PPNA ≤ 350 $g/m^2 \cdot$año?*

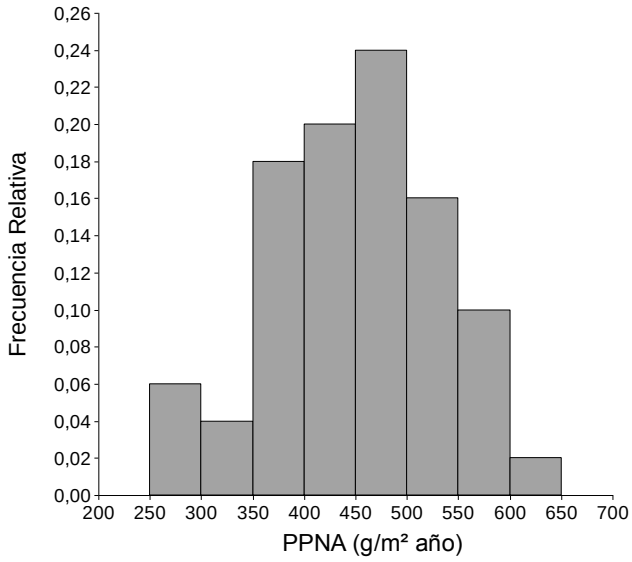


Figura 1.10. Distribución de frecuencias de los valores de productividad primaria neta aérea en 50 sitios de Pradera de Mesófitas en el sur de la Pampa Deprimida

d. Calcular valores aproximados de la media aritmética y de la varianza de la PPNA de estas praderas. Especificar las unidades correspondientes.

e. Un conjunto de 50 mediciones de PPNA en sitios con vegetación de Pradera de Hidrofitas tuvo media aritmética de 626 g/m²·año y desvío estándar de 151 g/m²·año, ¿qué diferencias habría entre el histograma aquí presentado para la Pradera de Mesofitas y el histograma de frecuencias relativas basado en dichas mediciones?

1.4 La vegetación natural de la Pampa Deprimida es un extenso pastizal con aspecto monótono. Sin embargo, cuando se lo observa en detalle, se encuentra que este pastizal es un mosaico de varias comunidades vegetales con diferente composición florística distribuidas en el paisaje en correspondencia con diferentes características de los suelos. En la Figura 1.11 se representan las distribuciones de frecuencias de los contenidos de carbono orgánico del horizonte superficial de suelos asociados con cuatro comunidades vegetales del sur de la Pampa Deprimida.

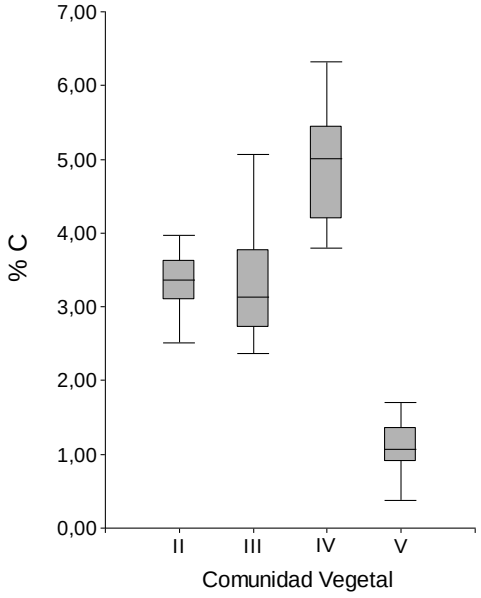


Figura 1.11. Distribuciones de frecuencias de los contenidos de Carbono orgánico (%) en el horizonte superficial de suelos asociados con diferentes comunidades vegetales del pastizal del sur de la Pampa Deprimida. Comunidad II, Pradera de Mesófitas, Comunidad III, Pradera Húmeda de Mesófitas, Comunidad IV, Pradera de Hidrófitas, Comunidad V, Estepa de Halófitas.

a. *¿Qué tipo de gráficos se presentan en la Figura 1.11?*

b. *Aproximadamente ¿qué valores tienen los máximos, mínimos y cuartiles de los contenidos de carbono del horizonte superficial de los suelos asociados con las comunidades IV (Pradera de Hidrófitas) y V (Estepa de Halófitas)?*

c. *Aproximadamente, ¿qué valores tienen las amplitudes totales y las amplitudes entre cuartiles de los contenidos de carbono del horizonte superficial de los suelos asociados con las comunidades IV (Pradera de Hidrófitas) y V (Estepa de Halófitas)?*

d. *Comparando las comunidades II (Pradera de Mesófitas) y III (Pradera Húmeda de Mesófitas), ¿cuál está asociada con mayor frecuencia con suelos cuyo contenido de carbono orgánico no supera 3 %? ¿cuál está asociada con mayor frecuencia con suelos cuyo contenido de carbono orgánico supera 4 %?*

e. *¿Cuál de las cuatro comunidades está asociada con suelos cuyos contenidos de carbono orgánico del suelo superficial son menores?*

f. *¿Cuál de las cuatro comunidades está asociada con un conjunto de suelos más heterogéneo en relación con el contenido de carbono orgánico en el horizonte superficial?*

1.5 Los herbicidas que se aplican a los cultivos eliminan malezas que compiten con las plantas cultivadas por luz, agua y nutrientes pero que también sirven como fuentes de alimento para organismos ubicados en el eslabón siguiente de la cadena trófica. Un grupo de estudiantes que investiga los impactos de las prácticas agrícolas sobre la biota de los campos de cultivo registró los pesos de *Calomys laucha* (laucha manchada) capturadas en dos lotes agrícolas de la Pampa Interior cultivados con soja, uno que había sido tratado con el herbicida Atrazina y otro que no había sido tratado con ningún herbicida. Los pesos registrados son los que se presentan a continuación:

Peso corporal de *Calomys laucha* (g)								
Lote tratado con Atrazina:	18,9	17,8	15,5	16,8	18,3	17,4	16,2	18,3
Lote no tratado:	18,7	19,5	19,2	18,9	18,2	19,9	19,5	

a. *En un mismo gráfico construir los polígonos de frecuencias relativas acumuladas de los pesos de* Calomys laucha *de cada lote como funciones escalonadas basadas en los registros individuales.*

b. *A partir de la observación del gráfico comparar la posición y la dispersión de las dos distribuciones de frecuencias.*

c. *Representar la distribución de frecuencias de los pesos de* Calomys laucha *de cada lote con un diagrama de caja y bigotes y comparar las distribuciones sobre la base de los diagramas.*

d. *Calcular la media aritmética, la varianza y el desvío estándar de los pesos de* Calomys laucha *de cada lote.*

e. *Sobre la base de la descripción de las distribuciones de frecuencia realizada en los puntos a – d, discutir la posible influencia de la aplicación de Atrazina sobre los pesos de* Calomys laucha *de estos lotes de soja.*

1.6 La Figura 1.12 representa las distribuciones de frecuencias de los totales de lluvia invernal (trimestre julio-septiembre) y estival (trimestre enero-marzo) registradas en el aeropuerto de Concordia, Entre Ríos, en el período 1981-2000.

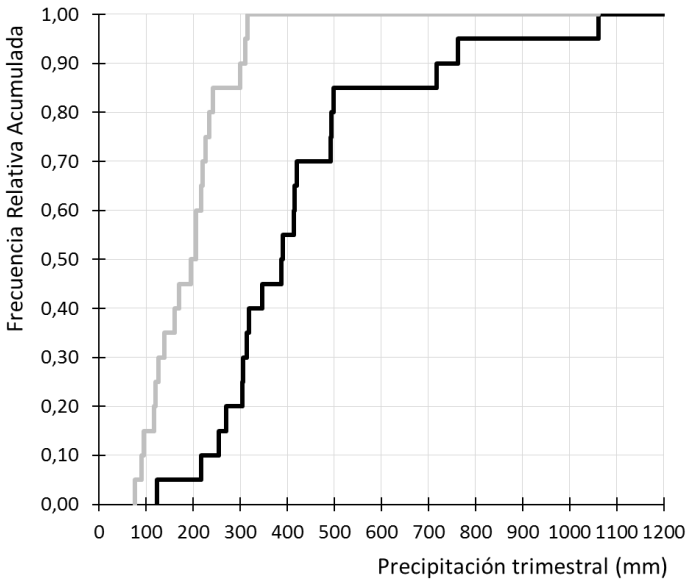


Figura 1.12. Polígonos de distribución de frecuencias relativas acumuladas de registros de precipitación trimestral en el aeropuerto de Concordia, Entre Ríos en el período 1981-2000. Línea negra, trimestre enero-marzo. Línea gris, trimestre julio-septiembre.

a. *¿En cuál de las dos estaciones llovió más? Justificar la respuesta con una medida de posición central extraída del gráfico.*

b. *¿En cuál de las dos estaciones la lluvia total fue más variable entre años? Justificar la respuesta con medidas de dispersión extraídas del gráfico.*

c. *En el período 1981-2000, los registros del aeropuerto de Concordia de lluvia otoñal (trimestre abril-junio) tuvieron mínimo de 81 mm, primer cuartil de 266 mm, mediana de 319 mm, tercer cuartil de 442 mm y máximo de 555 mm, mientras los registros de lluvia primaveral (trimestre octubre-diciembre) tuvieron mínimo de 113 mm, primer cuartil de 276 mm, mediana de 336 mm, tercer cuartil de 450 mm y máximo de 848 mm. Construir gráficos de caja y bigotes para comparar las distribuciones de frecuencias de las lluvias de las cuatro estaciones y describir la estacionalidad de las lluvias registradas. ¿En cuál estación el total de lluvia fue menor en promedio y en cuál fue más variable? ¿En qué estaciones se registraron los valores extremos más altos y más bajos?*

1.7 Con el propósito de evaluar la incidencia de la enfermedad conocida como pústula bacteriana (*Xanthomonas axonopodis* pv. *glycinea*) en un cultivo experimental de soja, una fitopatóloga observó 200 hojas de plantas de dicho cultivo y registró el número de lesiones que tenía cada una. La siguiente tabla presenta las frecuencias absolutas de hojas con cada número de lesiones registrado por esta profesional.

Lesiones por hoja	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Frecuencia absoluta	64	4	10	12	22	32	26	18	8	3	1

a. *¿De qué tipo es la variable registrada por la fitopatóloga?*

b. *Construir un gráfico de líneas para representar la distribución de frecuencias de los números de lesiones bacterianas por hoja.*

c. *Calcular y comparar los valores de la moda, la mediana y la media aritmética (notar que el valor de la media aritmética no necesariamente coincide con uno de los valores posibles de la variable registrada).*

d. *Construir un gráfico de caja y bigotes para representar esta distribución de frecuencias.*

e. *Calcular y comparar la amplitud total, la amplitud entre cuartiles y el desvío estándar de los números de lesiones por hoja. Especificar las unidades correspondientes.*

f. Escribir un texto explicativo de los principales rasgos de la distribución de frecuencias de los números de lesiones por hoja.

1.8 Demostrar las siguientes propiedades de la media y de la varianza de funciones lineales.

a. La suma de los desvíos $(x_i - \bar{x})$ *es nula (ecuación 1.3).*

b. La suma de los cuadrados de los desvíos $(x_i - \bar{x})$ *es menor que la suma de los cuadrados de las diferencias* $(x_i - c)$ *para todo* $c \neq \bar{x}$ *(ecuación 1.4).*

c. Si x *es una variable* y a *es una constante y* definimos $u = x + a$*, entonces:* $\bar{u} = \bar{x} + a$ *y* $var(u) = var(x)$ *(ecuaciones 1.9 y 1.10).*

d. Si x *es una variable* y b *una constante y* definimos $v = b \cdot x$*, entonces:* $\bar{v} = b \cdot \bar{x}$ *y* $var(v) = b^2 \cdot var(x)$ *(ecuaciones 1.11 y 1.12).*

e. Si x *e* y *son dos variables y* definimos $w = x + y$*, entonces:* $\bar{w} = \bar{x} + \bar{y}$ *y* $var(w) = var(x) + var(y) + 2 \cdot cova(x, y)$ *(ecuaciones 1.13 y 1.14).*

1.9 En la siguiente tabla se presentan los registros de los números de tormentas eléctricas y de los totales de lluvia primaveral (trimestre octubre-diciembre) en el aeropuerto de Concordia, Entre Ríos, durante el período 1981−2000.

Año	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*
Tormentas eléctricas	15	22	22	14	24	21	19	18	23	27
Lluvia (mm)	295,3	455,8	199,1	311,5	282,6	414,6	236,8	226,6	509,2	475,0

Año	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Tormentas eléctricas	25	15	25	16	17	18	25	12	11	22
Lluvia (mm)	361,3	294,8	633,4	443,2	331,9	339,8	847,7	270,1	112,6	423,8

a. Calcular la varianza de los números de tormentas eléctricas primaverales en el aeropuerto de Concordia durante el período 1981−2000. Indicar las unidades.

b. Calcular la varianza de los totales de lluvia primaveral registrados en el aeropuerto de Concordia durante el período 1981−2000. Indicar las unidades.

c. Calcular la covarianza entre los números de tormentas eléctricas y los totales de lluvia primaveral en el aeropuerto de Concordia durante el período 1981−2000*. Indicar las unidades.*

d. ¿Qué indica el signo de la covarianza calculada acerca de la asociación entre el número de tormentas eléctricas y el total de lluvias primaverales en el aeropuerto de Concordia? ¿Cómo sería la asociación si la covarianza tuviera el signo opuesto?

1.10 Las arañas son animales depredadores que controlan en buena medida los insectos perjudiciales para los cultivos. Este servicio natural depende de que el paisaje agrícola contenga refugios que aseguren la persistencia de las poblaciones de arañas de un año al siguiente. En el marco de una investigación sobre la importancia de los bordes de lote agrícola con vegetación espontánea para mantener este servicio natural, se capturaron arañas en dos tipos de sitio, lote de cultivo de soja y borde de lote con vegetación espontánea, en dos períodos del año, verano (cultivo en crecimiento) e invierno (período de rastrojo). Cada individuo capturado fue clasificado en uno de cinco gremios diferentes (grupos de arañas con hábitos similares). Las frecuencias absolutas registradas son las que figuran en la siguiente tabla.

Números de arañas capturadas por gremio en dos sitios de un paisaje agrícola en dos períodos del año. **I**. Deambuladoras pequeñas (< 10 mm) diurnas que cazan sobre las plantas. **II.** Sedentarias medianas (10 a 15 mm) crepusculares o nocturnas que cazan con telas orbiculares tejidas sobre las plantas. **III.** Deambuladoras pequeñas (< 10 mm) diurnas o nocturnas que cazan sobre el suelo. **IV**. Deambuladoras grandes (> 15 mm) nocturnas que cazan sobre el suelo. **V.** Sedentarias pequeñas (< 12 mm) que cazan con telas irregulares en el suelo o en las bases de las plantas. Datos adaptados de Liljesthröm et al, Neotropical Entomology 31:197-210 (2002)

Gremio	Lote (cultivo)		Borde	
	Verano	Invierno	Verano	Invierno
I	389	10	132	113
II	113	16	41	80
III	45	1	23	24
IV	59	14	4	4
V	21	231	19	31

a. *Construir gráficos de barras para representar las distribuciones de frecuencias relativas de gremios de arañas en cada tipo de sitio en cada período del año.*
b. *Observar los gráficos, ¿en cuál de los dos tipos de sitio la distribución de frecuencias de los gremios de arañas fluctúa más fuertemente entre verano e invierno?*
c. *¿Cuál es la clase (gremio) modal en cada tipo de sitio en cada período del año?*
d. *Calcular los índices N1 y N2 de diversidad de gremios de arañas correspondientes a cada tipo de sitio en cada período del año.*
e. *¿En cuál de los dos tipos de sitio hay mayor diversidad de gremios de arañas durante el verano y durante el invierno?*
f. *¿Cómo fluctúa la diversidad de gremios de arañas entre verano e invierno en cada tipo de sitio?*
g. *Según la descripción realizada ¿qué importancia aparente tienen los bordes de lote con vegetación espontánea para la persistencia de las poblaciones de arañas que controlan insectos perjudiciales en el cultivo de soja?*