Capítulo 9

MEDIDAS DE VARIABILIDAD O DISPERSIÓN

umario

Objetivos específicos

Al finalizar el estudio del capítulo, el estudiante será capaz de:

1. Explicar la importancia que tienen el examen y medición de la variabilidad de un conjunto de datos.
2. Calcular e interpretar las principales medidas de variabilidad absoluta.
3. Señalar la importancia del concepto de variabilidad relativa y explicar en qué situaciones debe usarse el coeficiente de variación.
4. Seleccionar la medida de variabilidad más adecuada para cada caso.
5. Calcular e interpretar puntajes estandarizados.
6. Conocer la técnica denominada diagrama de caja.

Resumen

En este capítulo, se estudia la variabilidad o dispersión de los datos. Para cada una de las medidas se discute el concepto básico en que descansa, su definición y el procedimiento de cálculo empleado, tanto en datos sueltos como agrupados en distribución de frecuencias. Se distingue entre variabilidad absoluta y relativa. También se incluye una sección donde se hace referencia a la estandarización de notas, y otra donde se trata la técnica denominada diagrama de caja.

9.1. LA VARIABILIDAD Y SU IMPORTANCIA

Al analizar y describir un conjunto de datos, se tienen en mente, por lo general, tres objetivos: determinar la forma como se distribuyen los datos, identificar los valores centrales –el "centro"– de esa distribución y establecer la medida como los datos se concentran o dispersan alrededor de esos valores centrales.

Las medidas de tendencia central, ya sea la moda, la mediana o la media aritmética, dan información muy útil sobre la distribución pero solo abordan un aspecto de los datos, la ubicación del centro de esta. No dicen nada acerca de la medida en que los datos se agrupan o dispersan alrededor de los valores centrales; por este motivo, aunque en casos muy específicos, un promedio puede dar una descripción adecuada del conjunto de datos, en casi la totalidad de las situaciones, los valores centrales resultan insuficientes, por eso deben calcularse medidas de dispersión para completar el panorama descriptivo de la distribución.

Desde esta perspectiva, los promedios son medidas muy útiles para describir los conjuntos de datos y conocer el centro de la distribución pero, casi sin excepción, los valores incluidos en esta difieren del valor central, aunque el grado de dispersión varíe de un conjunto a otro. Además, puede suceder que conjuntos con promedios muy similares varíen significativamente en su dispersión.

Todo esto hace que sea casi tan importante conocer un promedio como la variabilidad de los datos a su alrededor, de manera que sea posible una mejor descripción del conjunto y una comparación más eficaz con otros conjuntos o distribuciones. Esto sucede porque la validez de un promedio para resumir o representar el conjunto de datos para el cual se calculó depende, en grado sumo, de si los datos individuales se dispersan o

se concentran alrededor de él. Cuanto más concentrados estén los datos alrededor del promedio aritmético, por ejemplo, mucho más confianza se tendrá en este valor para caracterizarlos. Si la moda del número ideal de hijos es 3 en una población, y una proporción muy grande de mujeres se concentra en ese valor, este se puede utilizar con seguridad para describir las preferencias reproductivas de la población, como un todo, ya que pocos se alejan en forma significativa de él.

Una situación diferente se plantea con el ingreso familiar, que es muy variable. En el caso de esta característica, conocer la mediana, por ejemplo, es muy útil; pero, al usarla para hacer descripciones del nivel de ingreso de la población, no debe perderse de vista que una fracción importante de las familias tiene un ingreso muy superior o muy inferior a la mediana, y por ello las conclusiones que se saquen deben reforzarse o matizarse usando otras medidas de posición como los cuartiles, por ejemplo.

Conviene señalar que el concepto de variabilidad juega un papel clave dentro de la estadística. Si los fenómenos no se repitieran o lo hicieran sin variaciones, la estadística casi no tendría razón de ser. Pero la realidad es que la mayoría se repiten y lo hacen mostrando variaciones de mayor o menor intensidad, de ahí la importancia que tiene en el mundo moderno, al suministrar técnicas y procedimientos válidos y confiables para analizar esos hechos que se repiten y hacer inferencias acerca de ellos no obstante la variabilidad presentada.

El interés por la variabilidad de los datos se origina, entonces, en la certeza de que varían, pero tienen una tendencia natural a agruparse alrededor de los valores centrales, entre más concentrados estén los datos alrededor de ellos, más representativos son del conjunto y más confiables las conclusiones que se logren usando esos valores centrales o promedios.

La importancia del concepto de variabilidad se hace aún más clara, si se nota como en la práctica, puede suceder que varios conjuntos de datos tengan la misma media aritmética y, sin embargo, su dispersión sea muy diferente, tal como puede apreciarse en la figura 9.1.

Note que los conjuntos tienen la misma media aritmética $(\overline{X} = 5)$, pero su dispersión o variabilidad es muy diferente: mientras que en el grupo A todos los valores son iguales a 5, es decir, no existe dispersión, en B sí hay cierto grado de variabilidad y en el grupo C, la dispersión es aún mayor. Es más, en este último grupo ni siquiera hay un valor que sea igual al promedio y esto puede dar una idea de las conclusiones erróneas a las cuales se puede llegar si no se toma en cuenta la dispersión de los datos con respecto a esta medida. Es evidente que solo el conocimiento del valor central $\overline{x}$ no es suficiente para caracterizar un conjunto de datos.

	Suma	Promedio
a) 5, 5, 5, 5, 5	25	5
b) 4, 5, 5, 5, 6	25	5
a) 1, 3, 4, 7, 10	25	5

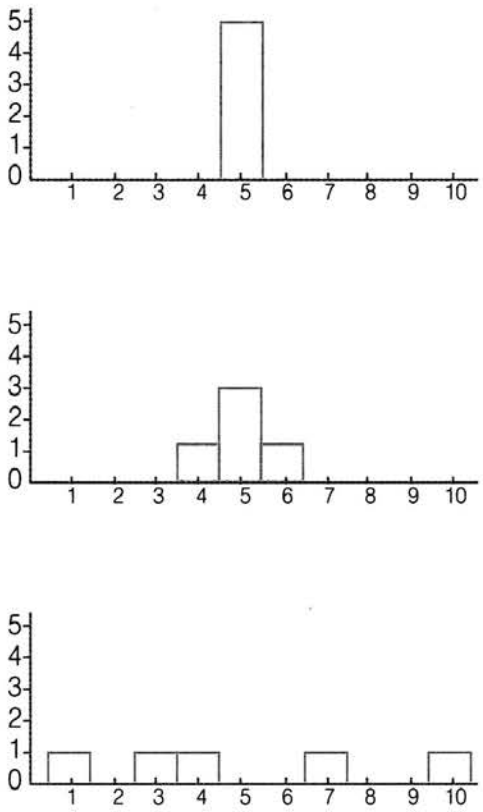

Figura 9.1. Ilustración de tres distribuciones con igual media y diferente variabilidad

Para los investigadores, la variabilidad es un fenómeno natural y corriente del cual tienen clara conciencia; un odontólogo, por ejemplo, cuando estudia las edades de erupción de una pieza dental, sabe que esta sale en los niños a diferentes edades, dependiendo de la influencia de muchos factores: herencia genética, alimentación, salud, etc., por lo tanto, procura establecer la edad más corriente de erupción de la pieza –mediana, promedio o moda– y, conocer lo que podría llamarse un intervalo de variación "normal", dentro del cual puede esperarse que les salga la pieza a un alto porcentaje de niños. Asimismo, todos los educadores saben que los estudiantes difieren en cuanto a rendimiento académico, y se manifiesta, por ejemplo, en las diferentes notas que se obtienen cuando un grupo es sometido al mismo examen o prueba. Por eso, si el educador desea hacer una evaluación del rendimiento mostrado en una cierta prueba, no solo debe calcular la nota promedio, sino también otras medidas que le permitan saber entre qué notas se encuentra la mayoría de los alumnos, cuál es la nota más alta y cuál la más baja, entre otras.

Un agrónomo, por su parte, interesado en comparar el efecto de dos fertilizantes sobre la producción de maíz, decide probar su efectividad aplicándolos a dos grupos de plantas de maíz y luego contrasta la producción promedio por planta de ambos grupos. Tiene clara conciencia de que una serie de elementos extraños, diferentes a los dos fertilizantes que desea comparar, tales como condición biológica de las plantas, tipo de

suelo, cambios climáticos, influyen sobre el experimento, introduciendo variaciones que afecten los resultados. Estas pueden agruparse bajo el nombre de "variabilidad experimental", eventualmente desvirtuarían el experimento, de tal manera que si encontrara diferencias importantes entre la producción de los dos grupos de plantas, no sería posible establecer con claridad si se debe a diferencias en la efectividad de los fertilizantes o al efecto de otros factores extraños no controlados; por otra parte, también puede ser que ese efecto perturbador actúe disimulando o anulando las diferencias de eficacia entre los fertilizantes.

Debido a lo anterior, el agrónomo procura reducir en lo posible el efecto de esos factores externos y la "variabilidad experimental" utilizando un diseño experimental adecuado, el cual incluye procedimientos muy conocidos, como plantas de una misma variedad, dividir la tierra en bloques pequeños y asignar el fertilizante aplicado a cada uno en forma aleatoria. Además, una vez realizado el experimento, procederá a medir esa variabilidad no debida a los fertilizantes que se comparan, y usarla para decidir si la diferencia entre la producción promedio para los dos tipos de productos puede ser considerada como real o si es un efecto aleatorio de la variabilidad experimental.[1]

9.2. LA MEDICIÓN DE LA VARIABILIDAD. EL RECORRIDO Y LA DESVIACIÓN MEDIA

Han sido propuestas diferentes formas de medir la variabilidad o dispersión de un conjunto de datos, cada una posee ventajas y limitaciones conceptuales y prácticas. La elección de una de ellas, en un caso concreto, dependerá de la situación particular que se considera y de si, en ese caso, las ventajas de su utilización superan sus limitaciones respecto a las otras medidas. Aquí se discutirán las medidas de variabilidad más conocidas, a saber:

a) El recorrido o amplitud

b) La desviación media

c) La desviación estándar y la variancia

d) El coeficiente de variación

Seguidamente, se discutirán el *recorrido* y la *desviación media;* las otras se tratarán en los apartados siguientes.

1. Este tipo de problemas corresponden a la prueba de hipótesis y se discuten en el capítulo 13.

9.2.1. El recorrido o amplitud

Una forma natural de apreciar la variabilidad es considerar los valores extremos del grupo, esto da origen al recorrido o amplitud, que se define como la diferencia entre el valor mayor y el menor del conjunto de datos. Para los siguientes datos: 3, 10,2, 8, 7, el recorrido es 8:

$$\text{Recorrido} = 10 - 2 = 8.$$

No obstante lo simple de su cálculo y lo fácil que resulta percibir su significado, el recorrido no es muy usado debido a ciertas limitaciones. La más importante, como puede apreciarse en su definición, es la de no considerar todas las observaciones del grupo o muestra de datos, sino únicamente el mayor valor y el menor. Esta característica hace que dependa sensiblemente del número de datos y tienda a aumentar al crecer, ya que es bastante probable que, entre las nuevas observaciones agregadas, aparezca una más pequeña o más grande a las existentes, eso incrementará en el valor del recorrido.

En la práctica, el recorrido se utiliza cuando se desea una medida simple de la variabilidad, cuando –por falta de tiempo– no se pueden emplear medidas más complejas o cuando el número de casos (la muestra) es pequeño, situación en la cual el recorrido es tan informativo como cualquiera de las otras medidas más elaboradas.

El recorrido es muy usado en el área de las aplicaciones del control estadístico de procesos. Esto se debe a que es una medida muy fácil de entender y calcular, además, en muchos de los procedimientos no se dispone ni del tiempo ni del personal necesario para calcular otras medidas menos simples.

9.2.2. La desviación media

La necesidad de definir una medida de dispersión que considere, para su cálculo, todos los datos y no esté tan estrechamente ligada al número de ellos, lleva casi automáticamente a la conclusión de que debe basarse en las desviaciones o diferencias de los datos individuales, respecto de un valor central o típico. Esta línea de razonamiento conduce, lógicamente, a considerar la suma de las desviaciones de los datos, con respecto a la media aritmética, como una posible medida de dispersión. Sin embargo, como fue indicado en la sección 8.2, propiedad 2, esa suma de desviaciones de las observaciones, con respecto a la media aritmética siempre es *igual a cero*, circunstancia que impide su uso como medida de dispersión.

Para obviar este problema, se puede emplear la suma de los *valores absolutos* de las diferencias y dividirla por el número de datos para obtener una medida de dispersión promedio o por observación. Así, se origina la llamada *desviación media.*

$$DM = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n} = \frac{\text{desviaciones}}{\text{número de datos}}.$$

Note que se consideran los valores absolutos de las diferencias, y para ello se escribe la diferencia $x_i - x$ entre dos barras verticales, forma convencional utilizada para indicar el valor absoluto de un número.

Su cálculo se ilustra, seguidamente, para los valores 3, 10, 2, 8 y 7.

x_i	$x_i - 6$	$\lvert x_i - 6 \rvert$
3	-3	3
10	+4	4
2	-4	4
8	+2	2
7	+1	1
	0	14

$$\bar{X} = \frac{\sum x_i}{n} = \frac{30}{5} = 6.$$

$$DM = \frac{14}{5} = 2,8.$$

9.3. LA DESVIACIÓN ESTÁNDAR Y LA VARIANCIA: CONCEPTO, DEFINICIÓN Y CÁLCULO EN DATOS SIMPLES Y AGRUPADOS

La *desviación media*, no obstante las ventajas conceptuales que reúne, casi no se utiliza debido a que requiere el manejo de valores absolutos por una parte, y por el hecho de existir otra medida, basada también en las desviaciones respecto a la media aritmética, que es mucho más cómoda y útil, además tiene numerosas utilidades prácticas y teóricas. Esta medida es la desviación estándar o típica que emplea, en lugar de los valores absolutos, los cuadrados de las desviaciones.

$$\text{Desviación estándar} = \sqrt{\frac{\sum(\text{desviaciones})^2}{\text{número de datos}}}.$$

La desviación estándar indica cuánto se alejan, en promedio, las observaciones de la media aritmética del conjunto. Es la medida de dispersión más usada en estadística, tanto en aspectos descriptivos como analíticos, también tiene mucha importancia su cuadrado, que recibe el nombre de variancia.

$$\text{variancia} = \frac{\sum(\text{desviaciones})^2}{\text{número de datos}}.$$

Como es más cómodo trabajar con la fórmula sin el radical, se procurará, en lo sucesivo, desarrollar los cálculos utilizando *la variancia* y extraer, al final del proceso, la raíz cuadrada para obtener la *desviación estándar*.

A continuación se discutirán con todo detalle las características e importancia de la desviación estándar y de la variancia, así como los métodos más apropiados para su cálculo.[2] Sin embargo, es conveniente hacer algunas observaciones acerca de la definición de variancia, según se considere una muestra o toda la población.

Como se explicó oportunamente, en muchas ocasiones, el estudio de una población se realiza al observar solo una muestra de sus elementos, no todos. Las medidas o valores calculados a partir de ella se utilizan luego para representar o estimar los de la población en los cuales se está interesado. Con el propósito de establecer claramente si el cálculo se realizó para toda la población o para una muestra, se acostumbra indicar con símbolos diferentes cada una de las situaciones. Comúnmente, se usan letras latinas mayúsculas o griegas para apuntar los valores de la población y latinas minúsculas para los calculados a partir de los datos de la muestra (estimadores). Además, es corriente emplear la letra N para señalar el número total de elementos en la población y la n para representar el tamaño de la muestra. A continuación, se presentan los símbolos y definiciones para el *promedio* (media aritmética) y la *variancia*, según se refieran a la población o a una muestra:

Grupo de referencia	Promedio	Variancia
Muestra (n)	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$
Población (N)	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Inmediatamente llama la atención que, al definir s^2, se utiliza $n-1$ como divisor, en vez de n. Esto obedece a que de acuerdo con la teoría estadística, al dividir por $n-1$ se obtiene una mejor estimación del valor poblacional σ^2 (variancia de la población). Sin embargo, si la muestra es grande, no tiene importancia alguna usar n o $n-1$ como

2. Al igual que las medidas de tendencia central, las de dispersión pueden calcularse usando las funciones estadísticas de los programas apropiados o las funciones de hojas de cálculo como la de Excel.

divisor, ya que el resultado numérico que se obtendrá será prácticamente el mismo; en cambio, si la muestra es pequeña, entonces sí es importante emplear la fórmula apropiada, o sea, la correspondiente a s^2.[3]

9.3.1. El cálculo de la variancia en datos sin agrupar

Seguidamente, se presenta el cálculo de la variancia s^2 cuando se tiene una muestra de n datos sin agrupar; la fórmula correspondiente aparece en el cuadro anterior. Al utilizar la fórmula y sacar luego la raíz cuadrada, puede obtenerse el valor de la *desviación estándar* (s). Se usan los valores 3, 10, 2, 8 y 7.

x_i	$x_i - x$	$(x_i - x)^2$
3	-3	9
10	4	16
2	-4	16
8	2	4
7	1	1
30		46

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30}{5} = 6.$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1} = \frac{46}{4} = 11,5.$$

$$s = \sqrt{11,5} = 3,39.$$

Una forma alternativa de calcular la variancia es con la expresión "fórmula para cálculos":

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{n-1}$$

a la cual se llega realizando ciertas transformaciones algebraicas en el numerador de s^2.

En seguida, se repite el cálculo de s y s^2, con la "fórmula para cálculos" recién comentada:

x	x^2
3	9
10	100
2	4
8	64
7	49
30	226

3. En el libro, siempre que se trabaje con muestras, se utilizará s^2.

$$\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \frac{(30)^2}{5} = \frac{900}{5} = 180.$$

$$\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_1\right)^2}{n} = 226 - 180 = 46.$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{46}{4} = 11,5.$$

$$s = \sqrt{11,5} = 3,39.$$

El detalle anterior busca, en esencia, mostrar las características de la fórmula y las opciones de cálculo de la variancia. Debe hacerse notar que, usualmente, el cálculo de este estadístico se realiza con una función específica de una hoja de cálculo –como la de Excel, por ejemplo– o de un programa estadístico.[4]

9.3.2. El cálculo de la variancia en datos agrupados en una distribución de frecuencias

Cuando los datos están agrupados en una distribución de frecuencias, la fórmula de cálculo de variancia debe ajustarse para que dependa de los puntos medios y de las frecuencias.

Si se trata de una muestra, la forma que adopta es la siguiente:

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n-1}.$$

Donde:

x_i = punto medio de la clase *i*.

f_i = frecuencia absoluta de la clase *i*.

$n = \sum_{i=1}^{k} f_i$, o sea el número de frecuencias u observaciones.

K = número de clases.

De igual manera, la "fórmula para cálculos" es

$$s^2 = \frac{\sum x^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{n}}{n-1}.$$

4. En el caso del Excel lo que se hace es digitar los datos en la hoja de cálculo y después utilizar las funciones estadísticas VAR y DESVESTA. Si esto se hace, se obtienen los valores 11,5 y 3,39.

A continuación, se ilustra el cálculo de la variancia y de la desviación estándar para la distribución de frecuencias que se ha venido tomando como ejemplo en los capítulos precedentes, referente al peso en kilogramos de 60 estudiantes hombres de dos colegios privados (capítulo 7).

Cuadro 9.1

CÁLCULO DE LA VARIANCIA Y DE LA DESVIACIÓN ESTÁNDAR PARA LA DISTRIBUCIÓN DE FRECUENCIAS

PUNTOS MEDIOS (x_i) (1)	FRECUENCIA (f_i) (2)	$x_i f_i$ (3) = (1) (2)	$x_i^2 f_i$ (4)=(1)(3)
47	1	47	2 209
52	6	312	16 224
57	8	456	25 992
62	19	1 178	73 036
67	9	603	40 401
72	6	432	31 104
77	5	385	29 645
82	3	246	20 172
87	3	261	22 707
	60	3 290	261 490

$$\frac{(\sum x_i f_i)^2}{n} = \frac{3920^2}{60} = 256\,106{,}67.$$

$$s^2 = \frac{261490 - 256106{,}67}{59} = 91{,}24.$$

$$s = \sqrt{s^2} = \sqrt{91{,}24} = 9{,}55.$$

Si s^2 se calcula con los datos originales, es decir, sin agrupar, se obtiene:

$$s = \sqrt{s^2} = \sqrt{93{,}93} = 9{,}69.$$

Si se compara la desviación estándar calculada a partir de los datos individuales, con la obtenida de la distribución de frecuencias, se advierte que no coinciden y la última es ligeramente inferior. La diferencia se debe al hecho de que la fórmula de la variancia, para el caso de las distribuciones de frecuencias, supone que todas las observaciones dentro de una clase son iguales al punto medio, lo cual, como es evidente, no se cumple en la mayoría de los casos y produce, consecuentemente, una ligera subestimación de la variancia (y de la desviación estándar). La discrepancia, sin embargo, entre la variancia real y la estimada a partir de la distribución de frecuencias es por lo general bastante pequeña como para ser ignorada para fines prácticos.

Ejemplo 1

Un consultor estudió una muestra de 40 tiendas de una gran ciudad e investigó, entre otras cosas, el monto anual gastado en publicidad por estas. La distribución de frecuencias siguiente resume la información recogida. Calcule la desviación estándar.

Cuadro 9.2
DISTRIBUCIÓN DE FRECUENCIAS

Miles de pesos	Número de tiendas	x_i	$x_i f_i$	$x_i^2 f_i$
10-19	4	15	60	900
20-29	14	25	350	8 750
30-39	8	35	280	9 800
40-49	6	45	270	12 150
50-59	3	55	165	9 075
60-79	4	70	280	19 600
80-99	1	90	90	8 100
Total	40		1495	68 375

$$s^2 = \frac{68\,371 - \frac{1495^2}{40}}{39} = \frac{68\,375 - 55\,875{,}6}{39}$$

$$= \frac{12\,499{,}4}{39} = 320{,}5.$$

$$s = \sqrt{320{,}5} = 17{,}9.$$

9.3.3. La variancia de un conjunto formado por la combinación de dos o más grupos de datos

Cuando se combinan dos o más grupos de datos, la variancia del conjunto puede obtenerse a partir de las variancias y medias de los grupos individuales. Para ilustrar este punto, considere el caso de dos poblaciones, 1 y 2, con las siguientes características:

Población	Tamaño	Media	Variancia
1	N_1	μ_1	σ_1^2
2	N_2	μ_2	σ_2^2

Como es sabido, la media del conjunto de las $N = N_1 + N_2$ observaciones es la media ponderada de las medias de cada una de las poblaciones. En símbolos:

$$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}.$$

En cuanto a la variancia del conjunto, σ^2, puede probarse matemáticamente que viene dada por la expresión:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\mu_1 - \mu)^2 + N_2(\mu_2 - \mu)^2}{N_1 + N_2}.$$

Suponga ahora que se tienen las dos poblaciones mostradas abajo:

Población	Elementos	Tamaño	Media	Variancia
1	1,3	$N_1 = 2$	$\mu_1 = 2$	$\sigma_1^2 = 1$
2	2,4,6,8	$N_2 = 4$	$\mu_2 = 5$	$\sigma_2^2 = 5$

Si se juntan las dos poblaciones en una sola y se utilizan las fórmulas antes presentadas, se tendrá, para el conjunto:

$$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2} = \frac{2 \cdot 2 + 4 \cdot 5}{6} = \frac{24}{6} = 4.$$

$$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\mu_1 - \mu)^2 + N_2(\mu_2 - \mu)}{N_1 + N_2}$$

$$= \frac{2 \cdot 1 + 4 \cdot 5 + 2 \cdot 2^2 + 4 \cdot 1^2}{6}$$

$$= \frac{2 + 20 + 8 + 4}{6} = \frac{34}{6} = \frac{17}{3}.$$

Ahora bien, la población combinada se conforma por $N = 6$ elementos: 1, 2, 3, 4, 6, 8; si se considera esta población de 6 elementos, y se usan las definiciones conocidas de media y variancia se obtiene:

$$\mu = \frac{\sum_{i=1}^{N} x_1}{N} = \frac{1 + 2 + 2 + 4 + 6 + 8}{6} = \frac{24}{6} = 4.$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 + (8-4)^2}{6}$$

$$\frac{9 + 4 + 1 + 0 + 4 + 16}{6} = \frac{34}{6} = \frac{17}{3}.$$

Como puede notarse, los resultados coinciden plenamente con los obtenidos con las fórmulas descritas para la media y la variancia de grupos combinados.

9.3.4. Variancia dentro y entre grupos

La propiedad de la variancia recién comentada se mantiene sin importar el número de grupos que se combinen. Por ejemplo, si se trata de k grupos, se tendría:

$N = N_1 + N_2 + N_3 + ... + N_k$, y la fórmula de la variancia para el total de los N datos, σ^2, sería:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$= \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \ldots + N_k\sigma_k^2 + N_1(\mu_1 - \mu)^2 + N_2(\mu_2 - \mu)^2 + \ldots + N_k(\mu_k - \mu)^2}{N}.$$

La cual podría resumirse en la expresión:

$$\sigma^2 = \underbrace{\frac{\sum_{i=1}^{k} N_i\sigma_i^2}{N}}_{\sigma_D^2} + \underbrace{\frac{\sum_{i=1}^{k} N_i(\mu_i - \mu)^2}{N}}_{\sigma_E^2}.$$

La fórmula anterior revela que, cuando se combinan grupos de datos, la variancia del conjunto está formada por dos partes: a) la media ponderada de las variancias de los grupos (σ_D^2); y b) la variancia de las medias de los grupos (σ_E^2). Los estadísticos denominan, a la primera, variancia *dentro*, ya que mide la variancia dentro de los grupos; y a la segunda, variancia *entre*, puesto que mide cómo varían las medias de los grupos respecto a la media general.

Esta propiedad es fundamental en estadística y constituye la base de la técnica denominada "análisis de variancia", que tiene importantes aplicaciones en el diseño de experimentos, en la regresión y correlación y en el muestreo.

Ejemplo 2

Considere una población formada por 20 mujeres casadas, mayores de 40 años, de quienes se conoce su educación formal y media y la variancia del número de hijos tenidos durante la vida (datos hipotéticos).

Educación primaria	Educación secundaria	Educación universitaria
$N_1 = 8$	$N_2 = 5$	$N_3 = 7$
$\mu_1 = 6$	$\mu_2 = 3$	$\mu_3 = 2$
$\sigma_1^2 = 3$	$\sigma_2^2 = 2$	$\sigma_3^2 = 2$

a) ¿Cuáles son los valores de μ y de σ^2?

$$\mu = \frac{\mu_1 N_1 + \mu_2 N_2 + \mu_3 N_3}{N_1 + N_2 + N_3} = \frac{6 \cdot 8 + 3 \cdot 5 + 2 \cdot 7}{20}$$

$$= \frac{48 + 15 + 14}{20} = \frac{77}{20} = 3{,}85.$$

$$\sigma^2 = \frac{\sum_{i=1}^{k} N_i\sigma_i^2}{N} + \frac{\sum_{i=1}^{k} N_i(\mu_i - \mu)^2}{N}$$

$$= \frac{8 \cdot 3 + 5 \cdot 2 + 7 \cdot 2}{20} + \frac{8(6 - 3{,}85)^2 + 5(3 - 3{,}85)^2 + 7(2 - 3{,}85)^2}{20}$$

$$= \frac{48}{20} + 64{,}55 = 2{,}40 + 3{,}23 = 5{,}63.$$

b) ¿Cuál de las dos variancias es más importante: la variancia *entre* o *dentro* de grupos?

$$\frac{\sigma_D^2}{\sigma^2} = \frac{2,40}{5,63} = 0,426.$$

$$\frac{\sigma_E^2}{\sigma^2} = \frac{3,23}{5,63} = 0,574.$$

Puede apreciarse que la variancia entre grupos representa un 57,4% de la variancia total. Esto señala que la variabilidad en el número de hijos tenidos durante la vida, puede explicarse, estadísticamente, en un 57%, por diferenciase en el nivel de la educación de las mujeres. En consecuencia, si todas las mujeres tuvieran el mismo nivel de educación, la variancia en el número de hijos tenidos sería, únicamente, un 43% de la observada en el grupo de datos analizado. A nivel más sustantivo, el resultado señala que la escolaridad de las mujeres es un factor determinante en el número de hijos tenidos.

La relación $R^2 = \frac{\sigma_E^2}{\sigma^2}$ es utilizada como una medida de asociación entre variables, en este caso fecundidad y educación, y recibe el nombre de "razón de determinación".

9.4. VARIABILIDAD RELATIVA. EL COEFICIENTE DE VARIACIÓN

Una situación corriente, en la investigación y en general, en el análisis de datos, es la necesidad de comparar dos o más conjuntos en cuanto a su variabilidad. Si se dan los datos en las mismas unidades y si los promedios de los conjuntos, es decir, la magnitud natural de los datos, son bastante similares, la desviación estándar es una herramienta perfectamente apropiada para realizar la comparación. Pero si alguna de las condiciones antes citadas no se cumple, la desviación estándar, y cualquier medida absoluta de dispersión, pierde casi toda su utilidad para este propósito.

Si los datos están expresados en diferentes unidades, es obvio que no puede compararse su variabilidad utilizando la desviación estándar, ya que carece de sentido contrastar, por ejemplo, una desviación estándar expresada en kilogramos con otra en centímetros o minutos. Por otra parte, aun cuando los conjuntos de datos estén en la misma unidad de medida, la diferencia entre sus promedios puede ser tan importante que haga completamente inadecuada la comparación directa de las desviaciones estándar. Tal sería el caso, por ejemplo, de una situación en la cual se estuviera examinando la variabilidad en el peso de conejos y en el de caballos, es evidente que una variación de un kilogramo tiene muy diferente importancia según el animal.

Se hace necesario, entonces, disponer de valores independientes de las unidades de medida y de la magnitud general de los datos que se consideren; con este propósito, se

utilizan las llamadas medidas de dispersión relativa, la más importante es el coeficiente de variación.

El coeficiente de variación indica el interés de la desviación estándar en relación con el promedio aritmético. Su definición es la siguiente:

$$\text{Coeficiente de variación} = \frac{\text{Desviación estándar}}{\text{Media aritmética}} \cdot 100.$$

Note que se ha multiplicado por 100. De acuerdo con la simbología presentada, se tendrían las siguientes fórmulas según se trate de una población o de una muestra:

$$C.V. = \frac{\sigma}{\mu} \cdot 100,$$

$$C.V. = \frac{s}{\bar{x}} \cdot 100.$$

Su definición obedece a las necesidades mencionadas anteriormente de contar con una medida independiente de las unidades y de la magnitud general de las observaciones. Al dividir la desviación estándar (s) –medida de dispersión absoluta– por la media aritmética $(\bar{x})$ –medida de tendencia central– se eliminan las unidades,[5] por una parte y, por otra, la inclusión del promedio en el divisor permite corregir el efecto que tiene la magnitud general de los datos sobre la desviación estándar. En otras palabras, si la desviación estándar es grande porque los datos en sí son grandes, al dividirse entre la media aritmética ese factor es eliminado o controlado. En cuanto a la multiplicación por 100, no tiene otro propósito que el de "amplificar" el número relativo y hacer más cómodo su uso.

Para el ejemplo del peso de los 60 estudiantes hombres de colegios privados, que se ha venido considerando a lo largo del texto, los cálculos realizados indicaron un $\bar{x} = 64,93$ y $s = 9,69$; por ello, el coeficiente de variación resulta ser:

$$C.V. = \frac{s}{\bar{x}} \cdot 100 = \frac{9,69}{64,93} \cdot 100 = 14,92\,\%.$$

Este resultado puede interpretarse diciendo que la desviación estándar representa un 14,62% de la media aritmética; sin embargo, su verdadero significado y utilidad aparecen cuando se realizan comparaciones.

5. Tanto la desviación estándar como la media aritmética están en unidades concretas; así, si se trata de peso, ambas vendrían dadas en kilogramos, por ello, al dividir una entre la otra, desaparece la unidad de referencia y queda solo un número que no se refiere a una unidad determinada.

Ejemplo 3

Los siguientes datos se refieren a estatura en centímetros de niñas de 2 y 16 años.

Edad en años	Estatura promedio	Desviación estándar
2	84	3
16	160	5

En términos absolutos, es evidente que hay mayor variabilidad en el grupo de niñas de 16 años, pues la desviación estándar es superior; sin embargo, al calcular los coeficientes de variación se descubre que son muy parecidos, resulta más bien ligeramente inferior el correspondiente a niñas de 16 años.

$$C.V = \frac{3}{84} \cdot 100 = 3{,}6\,\% \text{ (2 años)},$$

$$C.V = \frac{5}{160} \cdot 100 = 3{,}1\,\% \text{ (16 años)}.$$

Se concluye, entonces, que la dispersión relativa en ambos grupos de niñas es muy similar, y la mayor dispersión absoluta del grupo de 16 años se origina en el hecho de que esas niñas, por ser de más edad son, en general, más altas.

9.5. ESTANDARIZACIÓN DE NOTAS Y PUNTAJES

Otra aplicación muy útil de la desviación estándar se presenta en la estandarización de notas o puntajes. Con gran frecuencia, educadores y psicólogos necesitan comparar el nivel de rendimiento mostrado por los estudiantes en asignaturas de muy diferente naturaleza, por ejemplo entre historia y matemáticas. La comparación directa de los puntajes o calificaciones obtenidos por los alumnos en cada una de esas asignaturas, evidentemente, tiene poco valor. Las diferencias en la naturaleza del campo específico de conocimiento, en la metodología pedagógica empleada en su enseñanza y en la forma y exigencia de las evaluaciones utilizadas en cada una de ellas, hacen que, de partida, deban esperarse diferencias importantes en las calificaciones obtenidas, las cuales no pueden explicarse exclusivamente al rendimiento o capacidad de los estudiantes. Por estas circunstancias, entonces, resulta importante determinar, en cada asignatura, cuál es la posición relativa que ocupa la calificación de un dicente en particular, en relación con las obtenidas por sus compañeros, y luego hacer la comparación entre ambas materias. ¿Cómo se logra lo anterior?

Para visualizar mejor el procedimiento que debe seguirse, se ilustra con el ejemplo 4.

Ejemplo 4

Don Alonso tiene dos hijos: Cecilia y Efraín. Cecilia sigue la carrera de Derecho mientras Efraín estudia Ingeniería Industrial. Un día, a la hora de almuerzo, Cecilia informa a la familia que obtuvo un 85% como nota promedio en el curso de Derecho Constitucional, el cual considera como el más difícil de los que llevó en el semestre. Al ser interrogado por su hermana sobre la nota obtenida en el curso que más lo obligó a estudiar, Efraín responde, cabizbajo, haber conseguido un 70% de promedio en el curso de Probabilidad y Estadística. El padre, entonces, felicita a Cecilia por su buen rendimiento académico y reprende a Efraín por no seguir el ejemplo de su hermana. ¿Tiene don Alonso razones objetivas para sumir esa actitud?

En términos absolutos, obviamente Cecilia salió mejor que Efraín pero, ¿cómo se comparan los resultados de cada uno en relación con las notas obtenidas por sus compañeros en sus respectivos cursos? ¿No podría ser que en Derecho se califique con menos rigor que en Ingeniería, o en esta última las materias sean más difíciles? Efraín, pensando en esto y acordándose de algo indicado por el profesor de Probabilidad y Estadística, lo llama por teléfono y le consulta el problema. El profesor le dice que sí le puede ayudar, pero necesita conocer cuál es el promedio y la desviación estándar de las notas obtenidas por el grupo de Derecho Constitucional (las de Probabilidad y Estadística él ya las tiene calculadas, ¡naturalmente!), Efraín las consigue y se reúne dos días después con su profesor. Este, bajo la mirada curiosa de Efraín, anota y calcula lo siguiente:

Derecho Constitucional	Probabilidad y Estadística
$\mu_D = 82$	$\mu_P = 60$
$\sigma_D = 6$	$\sigma_P = 8$
Nota de Cecilia: $X_C = 85$	Nota de Efraín: $X_E = 70$
Nota de Cecilia estandarizada:	Nota de Efraín estandarizada:
$Z_C = \frac{X_C - \mu_D}{D}$	$Z_E = \frac{X_E - \mu_P}{P}$
$= \frac{85 - 82}{6}$	$= \frac{70 - 60}{8}$
$= 0,50.$	$= 1,25.$

Al observar estos resultados, el profesor comenta a Efraín que, como este lo sospechaba, don Alonso se equivocó al reprenderlo por su desempeño académico y, si bien no estuvo mal que felicitara a Cecilia, una felicitación más calurosa debía haber sido para él, pues obtuvo un mejor rendimiento, con respecto a sus compañeros, que Cecilia. En efecto, el profesor le explica que las notas estandarizadas comparan, en primer término, las desviaciones de las notas específicas con respecto a la media del grupo, o sea las $x_i - \mu$. Esto elimina la distorsión introducida en la comparación de las notas reales, por el hecho de que los promedios de ambos grupos sean tan diferentes. Para una comparación válida, sin embargo, este ajuste no basta, es necesario, además, eliminar el efecto de las diferencias en variabilidad percibidas dentro de las notas, en cada uno de los cursos.

Esto se logra dividiendo las desviaciones con respecto a la media, $x_i - \mu$, entre la desviación estándar de las notas de esa asignatura. De esta manera, se obtiene un número que dice cuánto se aleja la nota específica (la de Cecilia, por ejemplo) del promedio de su grupo, en término de unidades de desviaciones estándares. En este caso $x_C - \mu_D = 85 - 82 = 3$, Cecilia superó en 3 puntos el promedio del grupo, y esa diferencia representa solo la mitad de la desviación estándar (3/6). Por el contrario, la nota de Efraín supera el promedio del grupo en 10 puntos, y esa desviación representa 1,25 desviaciones estándar. Los valores 0,50 y 1,25 son independientes de la magnitud general de las notas y de su variabilidad, por ello permiten una comparación válida.

Como la interpretación directa de estos valores, tal como se han presentado, puede resultar incómoda, es posible hacer una transformación que permita obtener notas estandarizadas con una media y una desviación estándar "cómodas". Bajo el supuesto, por ejemplo, de que la nota de aprobación es de 70 puntos y una desviación estándar de 10 puntos es razonable, la expresión para el cálculo de dicha nota sería:

Nota estandarizada $= 70 + 10(Z)$, y las correspondientes para Cecilia y Efraín serían las siguientes:

Nota de Cecilia: $70 + 10 \cdot 0{,}50 = 70 + 5 = 75$.

Nota de Efraín: $70 + 10 \cdot 1{,}25 = 70 + 12{,}5 = 82{,}5$.

Estos valores indican que, si la nota promedio en cada asignatura fuera de 70 y la desviación estándar 10, Efraín recibiría un 82,5 en Probabilidad y Estadística y Cecilia un 75 en Derecho Constitucional.

9.6. MEDIA Y VARIANCIA DE VARIABLES DICOTÓMICAS

Usualmente, cuando se piensa en promedio o en variancias, se consideran solo variables cuantitativas (continuas o discretas). Los conceptos son aplicables, sin embargo, al caso de variables dicotómicas como el sexo, la condición de ocupado o desocupado, la tenencia de celular, y aún a variables cualitativas como el estado civil o la provincia de nacimiento, bajo ciertas condiciones.

Considere, por ejemplo, poseer o no celular. En este caso, se recurre a definir una variable dicotómica x tal que asume el valor 1 cuando la persona tiene celular y el valor 0 cuando no lo tiene. Con esta definición, la fórmula $\sum X_i = N_1$ da el número de personas con celular y $N - N_1$ representa a quienes no tienen. A su vez

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{N_1}{N} = P$$

indica la proporción de personas que poseen celular, dentro de la población de interés, y $1 - P = Q$ la de personas que no lo tienen.

En general, puede demostrarse que, si se tiene una variable x: (0,1) (o sea dicotómica) como la anterior, la variancia viene dada por:

$$\text{Variancia} = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = P(1-P) = PQ.$$

Donde P representa la proporción de elementos en la población con la característica de interés –tener celular, por ejemplo–, y $1 - P = Q$ la proporción que no lo tiene.

Si se tiene una muestra de tamaño n, las expresiones correspondientes serán:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = p.$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{n}{n-1}p = p(1-p) \approx pq.$$

Note que si la muestra es medianamente grande, el factor $\frac{n}{n-1}$ es muy cercano a 1, y por eso, en la práctica, se utiliza comúnmente $s^2 = p(1-p) = pq$.

Ejemplo 5

Suponga que se toma una muestra de 8 familias y para cada una de ellas se determina si tienen computadora en la vivienda o no.

Cuadro 9.3
FAMILIAS CON O SIN COMPUTADORA

FAMILIA	CONDICIÓN	VARIABLE X
1	Tiene computadora	1
2	Tiene computadora	1
3	No tiene computadora	0
4	Tiene computadora	1
5	No tiene computadora	0
6	No tiene computadora	0
7	Tiene computadora	1
8	Tiene computadora	1

¿Cuál es la media y la variancia de la variable?

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5}{8} = 0,625.$$

$$s^2 = \frac{n}{n-1}p(1-p) = 1,143 \cdot 0,625 \cdot 0,375 = 0,268.$$

Si la variable de interés no es dicotómica, sino categórica, como el estado civil, por ejemplo, lo que se hace es "dicotomizarla", seleccionando la categoría de interés (soltero, por ejemplo) y agrupando todas las demás (no soltero). Hecho esto, se aplica la misma técnica ya vista para las variables dicotómicas.

En resumen, debe concluirse que para las variables cualitativas o categóricas es posible calcular el promedio y la variancia. Esta propiedad tiene singular importancia, en la inferencia estadística, cuando se considera el muestreo de proporciones.

9.7. ILUSTRACIÓN INTEGRADA DE DISTRIBUCIONES DE FRECUENCIAS Y MEDIDAS DE TENDENCIA CENTRAL Y VARIABILIDAD

Un economista tiene datos sobre el consumo anual de mariscos por familia, en kilogramos, de las dos regiones en las cuales se divide un país: zona costera y central. Estos datos fueron recogidos en una encuesta por muestreo reciente. Para la zona central ya se han calculado algunas medidas de posición, obteniéndose: media aritmética = 8,4; Mo = 6,1 y Me = 7,4. Se sabe, además, que el número total de familias en la zona central es de 600 000, es decir, representan un 40% del total del país.

Cuadro 9.4

CONSUMO ANUAL DE MARISCOS POR FAMILIA

CONSUMO MENSUAL en kilogramos	NÚMERO DE FAMILIAS ENCUESTADAS	
	Zona central	Zona costera
1 a 3	25	6
4 a 6	100	75
7 a 9	88	50
10 a 12	40	30
13 a 15	22	20
16 a 21	18	10
22 a 30	7	9
TOTAL	300	200

a) Interprete el valor 7,4 para la zona central.

b) Calcule Me, Mo y $\bar{x}$ para la zona costera.

c) ¿Dónde consumen más mariscos las familias, en la zona costera o en la central? Justifíquelo.

d) ¿Cuál es el consumo promedio para todo el país?

e) ¿En cuánto estimaría el volumen total de consumo de mariscos del país, para el año de la encuesta?

SOLUCIÓN

a) La mitad de las familias de la zona central consumen menos de 7,4 kg de mariscos al año, y la otra mitad más de esa cantidad.

b)

Cuadro 9.5

CÁLCULO DE ME, MO y $\bar{x}$ PARA LA ZONA COSTERA

CLASES	f_i	F_a	x_i	$x_i f_i$
1-3	6	6	2	12
4-6	75	81	5	375
7-9	50	131	8	400
10-12	30	161	11	330
13-15	20	181	14	280
16-21	10	191	18,5	185
22-30	9	200	26	234
	200			1816

$$M_e = L_i + \left[\frac{\frac{n}{2} - F_a}{f_i}\right]c = 6,5 + \frac{100 - 81}{50} \cdot 3$$
$$= 6,5 + 1,14 = 7,64.$$

$$M_e = L_i + \left[\frac{f_1}{f_1 + f_2}\right]c = 3,5 + \frac{69}{69 + 25} \cdot 3$$
$$= 3,5 + 2,2 = 5,70.$$

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{1816}{200} = 9,08.$$

c) Es parecido en ambas zonas, ya que, aun cuando la moda en la zona central es un poco mayor: 6,1 contra 5,7, las medianas son muy similares.

d) El consumo promedio para todo el país se puede calcular haciendo una media ponderada de los consumos medios de ambas zonas. Recuerde que un 40% de las familias está en la zona central y un 60% en la costera. Entonces:

$$\bar{x} = w_1\bar{x}_1 + w_2\bar{x}_2 = 0,40 \cdot 8,4 + 0,60 \cdot 9,1 = 8,82.$$

e) El consumo total de mariscos en el país se obtiene multiplicando el número total de familias por el consumo promedio por familia. El número de ellas en la zona central es de 600 000 (40%), en la costera 900 000 (60%).

$$\text{Consumo total} = \hat{x} = N\bar{x} = 150\,000 \cdot 8,82 = 13\,230\,000 \text{ kg.}$$

9.8. MEDIDAS DE POSICIÓN O CUANTILIOS

Los cuantilios o medidas de posición tienen como propósito describir la posición que guarda un valor específico de la distribución, con respecto al resto de los valores, cuando todos han sido ordenados. Los más usados son los cuartiles y percentiles, aunque en ciertas áreas también se recurre a los deciles.

La mediana es una medida típica de posición, aunque también se utiliza como medida de tendencia central por ubicarse en el centro de la distribución,. Las otras medidas de posición no están cerca del centro, con excepción del cuartil 1, el decil 5 y el percentil 50, los cuales son iguales a la mediana.

Como ya se sabe, la mediana divide el conjunto en dos partes iguales, es decir, la mitad de los valores son inferiores a ella y la otra mitad, superiores. En forma similar, se definen los cuartiles, que son valores de la variable, los cuales dividen el conjunto en cuartas partes. Así, el primer cuartil Q_1 es un valor tal que una cuarta parte de las observaciones son menores que él y $\frac{3}{4}$ partes, mayores; el segundo cuartil Q_2 es igual a la mediana y el Q_3 sobrepasa a $\frac{3}{4}$ partes de las observaciones, solo es superado por un cuarto de ellas.

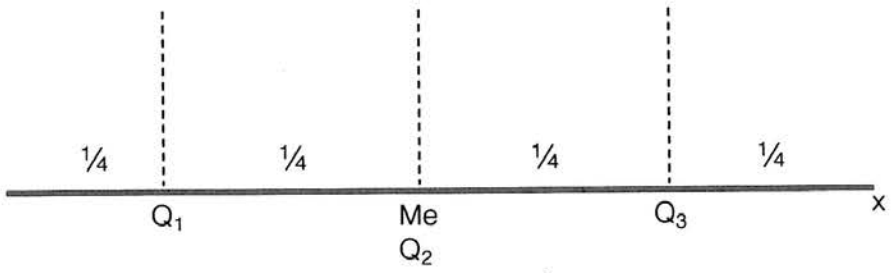

Figura 9.2. Quantilios

Además de la mediana y los cuartiles, existen también los deciles, son nueve valores que dividen el conjunto ordenado en diez partes iguales o décimas; y los noventa y nueve percentiles, los cuales lo dividen en centésimas.

Obviamente, tanto la mediana como los cuartiles y deciles constituyen casos particulares de los percentiles y pueden expresarse como tales. Por ejemplo: $M_e = P_{50}, Q_3 = P_{75}$, $D_4 = P_{40}$, etc. Resulta, entonces, que si se tiene una fórmula para el cálculo de los percentiles, puede usarse para calcular cualquier cuantilio.

Para el cálculo de los percentiles se debe, en primer lugar, ordenar los datos de menor a mayor, una vez hecho esto se aplica la fórmula siguiente:

$$P_m = \frac{m}{100}(n+1) \text{ término}$$

$$1 \leq m \leq 99.$$

Donde:

P_m = percentil m. Valor tal que un $\frac{m}{100}$ de las observaciones son menores o iguales que él y un $1 - \frac{m}{100}$ son mayores.

m = número que indica el percentil deseado. Por ejemplo, si $n = 43$, esto significa que se quiere el percentil 43, se escribe P_{43}. Evidentemente, m variará entre 1 y 99.

n = número total de observaciones que forma el grupo estudiado.

Ejemplo 6

Calcular el percentil 86 para los datos sobre los pesos de 60 estudiantes que se han venido considerando. En este caso:

$$n = 60 \text{ y } m = 86.$$

$$P_{86} = \frac{86}{100}(60+1) \text{ término} = \frac{86 \cdot 61}{100} \text{ término} = 52,46 \text{ término.}$$

El percentil 86 está dado por el 52,46 término del conjunto de los 60 datos ordenados. Sin embargo, el 52,46 término no existe y deberá entonces hacerse una interpolación lineal entre los términos 52 y 53, usando la conocida *regla de tres*. Seguidamente, se incluyen los datos ordenados y se calcula el percentil 86.

45	56	61	64	68	75
52	56	61	64	68	75
52	56	62	64	69	77
52	57	62	64	70	79
53	57	62	65	70	80
53	60	62	65	71	84
53	60	62	66	71	84
55	60	63	67	72	86
55	61	63	67	73	87
55	61	63	67	75	88

Término	Valor
52	75
52,46	P_{86}
53	77

$$\frac{77-75}{53-52} = \frac{P_{86}-75}{52,46-52} \Rightarrow 2 = \frac{P_{86}-75}{0,46}.$$

$$P_{86} = 75 + 2 \cdot 0,46 = 75 + 0,92 = 75,92.$$

El valor $P_{86} = 75,92$ significa que un 86% de los estudiantes tiene un peso inferior a 75,92 kg y un 14%, más de ese peso.

Respecto al cálculo de los otros cuantilios, como ya se indicó, es cuestión de expresarlos en percentiles. Así, por ejemplo, si se desea obtener el D_6, simplemente se procede a calcular el percentil 60. Como ejercicio, compruebe $P_{35} = 61$; $D_4 = 62$ y $Q_1 = 57,75$. Cuando se tienen datos agrupados, el cálculo de los percentiles se realiza utilizando la siguiente fórmula:

$$P_m = L_i + \frac{\frac{m}{100}(n - F_a)}{f_i} \cdot c.$$

Donde:

P_m = percentil m.

M = número que indica el percentil que se quiere.

n = número total de observaciones, o sea, la suma de las frecuencias absolutas $(\sum f_i)$.

L_i = límite inferior de la clase donde está el percentil m.

f_i = frecuencia absoluta de la clase donde está P_m.

F_a = frecuencia acumulada "menos de" de la clase anterior a aquella donde está el percentil m.

c = intervalo de la clase donde está el percentil m.

Observe que la fórmula de la mediana simplemente es un caso particular de la fórmula anterior, cuando $m = 50$.

Como ejemplo, considere el cálculo del percentil 21 para la distribución de frecuencias correspondientes a los pesos de los 60 estudiantes hombres de los dos colegios privados, que se incluye seguidamente.

Cuadro 9.6

CÁLCULO DEL PERCENTIL 21, PARA LA DISTRIBUCIÓN DE FRECUENCIAS DE LOS PESOS DE LOS 60 ESTUDIANTES HOMBRES DE LOS DOS COLEGIOS PRIVADOS

CLASES	Frecuencia absoluta (f_i)	Frecuencia acumulada (F_i)
44,5 – 49,5	1	1
49,5 – 54,5	6	7
54,5 – 59,5	8	15
59,5 – 64,5	19	34
64,5 – 69,5	9	43
69,5 – 74,5	6	49
74,5 – 79,5	5	54
79,5 – 84,5	3	57
84,5 – 89,5	3	60
TOTAL	60	

Cálculo de P_{21}:

a) $\frac{m}{100} \cdot n = \frac{21}{100} \cdot 60 = 12,6$.

b) Clase donde está P_{21}: $54,5 - 59,5$.

c) Intervalo de clase: $c = 59,5 - 54,5 = 5$.

$$P_{21} = 54,5 + \frac{12,6 - 7}{8} \cdot 5 = 54,5 + 3,5 = 58.$$

El valor obtenido $P_{21} = 58$ significa que un 72% de los escolares tienen pesos inferiores a 58 kilos y un 28%, mayor. Como ejercicio compruebe $P_{77} = 72,17$ y $P_{24} = 59,13$.

Para terminar esta parte dedicada a las medidas de posición o cuantilios, es pertinente hacer algunos comentarios acerca de su interpretación.

En realidad, los cuantilios son valores que dividen un grupo de observaciones ordenadas, de acuerdo con su magnitud, en fracciones; es decir, delimitan un segmento del grupo. Sin embargo, en la práctica es corriente hablar de ellas como si fueran categorías, y así se dice: "su nota quedó en el tercer cuartil" o "los alumnos incluidos en el primer decil", queriendo indicar que la nota del estudiante quedó dentro del 25% superior o, en el segundo caso, el grupo de alumnos que obtuvo notas inferiores al D_1 es decir, el 10% con las notas más bajas. Debe quedar claro, entonces, que los cuantilios son valores para indicar posición, pero en el lenguaje corriente es usual que se utilicen también como categorías.

Seguidamente, se incluyen dos ejemplos con el propósito de ilustrar el uso e interpretación de estas medidas.

Ejemplo 7

En una universidad europea, se aplica a los alumnos nuevos una prueba de habilidad general que tiene un máximo de 500 puntos. Un estudiante costarricense se somete a ella y obtiene 230 puntos. Al conocer el resultado, se siente desalentado porque no obtuvo ni siquiera una puntuación igual a la mitad de los puntos posibles. Pero, en realidad, ¿se puede decir que salió mal?

Si se conoce el criterio que sigue la universidad mencionada para admitir alumnos, no sería difícil decidir si salió mal o bien; es cuestión de determinar si su nota está por debajo o encima del mínimo exigido. Sin embargo, puede obtenerse una apreciación sobre el rendimiento dado por el estudiante, al considerar el de los otros participantes. Suponga (que se sabe) que la nota obtenida, 230 puntos, equivale al percentil 75. ¿Qué significa esto? Que el estudiante costarricense obtuvo una nota superior a la del 75% de los participantes. Su resultado, desde este punto de vista puede considerarse bastante satisfactorio.

9.8.1. Recorrido intercuartil (RIC)

Las medidas de posición, como los cuartiles y los deciles, informan respecto a la variabilidad del conjunto de datos; con los primeros, se define una medida de dispersión muy interesante y útil, denominada recorrido intercuartil, y es la diferencia entre el cuartil 3 y el 1, de la distribución: $RIC = Q_3 - Q_1$.

25% más bajo | | | 25% más alto

Q_1 Q_2 Q_3

| ← RIC → |

50% central

Figura 9.3. Recorrido intercuartil

Esta medida tiene la ventaja de dar una idea de la dispersión del 50% central de los datos, y excluye el efecto de los valores extremos. Además, resulta muy útil cuando se tienen clases abiertas y no es posible calcular la variancia.

9.9. DIAGRAMA DE CAJA[6]

El diagrama de caja es un procedimiento gráfico basado en los cuartiles, permite visualizar, en una forma organizada, un grupo de datos. Está compuesto por un rectángulo, la "caja" y dos brazos, los "bigotes"; para construirla se necesitan cinco medidas: el valor mínimo y máximo, los cuartiles Q_1, Q_2 y el Q_2 (es igual a la mediana). Suministra información sobre la ubicación de los valores mínimo y máximo, de los cuartiles y el espacio donde se concentra el 50% central de los datos y la simetría de la distribución. Se puede ajustar para que señale también el número y ubicación de los valores atípicos.

Para explicar la construcción del diagrama y su interpretación, se usará la siguiente información que corresponde a la duración en minutos de las entrevistas realizadas en una encuesta telefónica de $n = 449$ casos, llevadas a cabo en la última semana de la campaña política del 2010.

Valor mínimo	5 minutos
Primer cuartil Q_1	10 minutos
Mediana Q_2	11 minutos
Tercer cuartil Q_3	14 minutos
Valor máximo	30 minutos

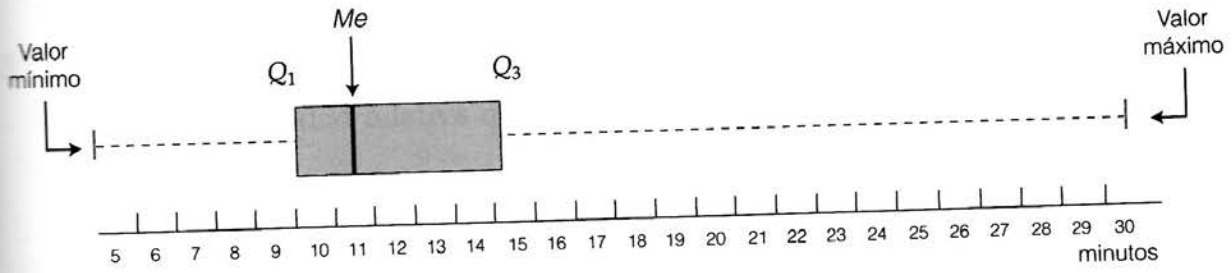

Figura 9.4. Diagrama de caja

6. Esta técnica descriptiva es conocida en inglés como *Box plot o box-and-whisker diagram*. El término whisker se debe a la apariencia de bigotes de gato de las líneas que señalan los valores máximos y mínimos. Fue propuesta en los años setenta por J. Tukey, un estadístico muy conocido, especialista en análisis de datos.

El diagrama muestra varios detalles interesantes:

a) El 50% central de las duraciones está entre 10 y 14 minutos.

b) El recorrido intercuartil, distancia entre el Q_1 y Q_3 igual a $14 - 10 = 4$.

c) La distribución es asimétrica hacia la derecha, o sea tiene una asimetría positiva. Esto es señalado por dos piezas de información: la mediana no está en el centro de la caja sino mucho más cerca del Q_1 y, por consiguiente, el área de la caja está situada a la derecha de ella es mucho mayor que el área a la izquierda; la longitud del "bigote derecho –distancia entre el Q_3 el valor máximo– es mucho mayor que la longitud del "bigote" izquierdo (16 minutos *vrs.* 5 minutos).

EJERCICIOS DE AUTOEVALUACIÓN

I. **Selección única.** A continuación se le dan 4 preguntas de selección única, marque con una equis (x) la opción que conteste en forma correcta y verdadera la proposición dada.

1. En relación con el recorrido, ¿cuál de las siguientes afirmaciones es correcta?
   a) Su cálculo toma en cuenta todas las observaciones del conjunto.
   b) Su magnitud depende solamente de los valores extremos del conjunto.
   c) Es, en realidad, una medida compleja de variabilidad.
   d) Se utiliza cuando se requiere una medida precisa de la variabilidad.

2. La utilidad de la media aritmética, para caracterizar un conjunto de datos, es afectada por su dispersión:
   a) En gran medida
   b) Escasamente
   c) De ninguna manera
   d) Siempre

3. La desviación estándar indica una de las siguientes alternativas:
   a) La variabilidad relativa que existe entre cada uno de los datos y la media del conjunto.
   b) El porcentaje que representa la suma de las desviaciones de los datos, con respecto al promedio, en relación con el total de las observaciones.
   c) En cuánto se alejan los datos en promedio de la media aritmética del conjunto.
   d) La magnitud de las diferencias entre cada observación y la mediana.

4. La variancia entre grupos tiene la siguiente definición:
   a) Media ponderada de las variancias de los grupos.
   b) Variancia ponderada de las medias de los grupos.
   c) Variancia total dividida entre la variancia dentro de grupos.
   d) Suma de las variancias de los grupos dividida entre la variancia total.

II. Completar. A continuación se le da una pregunta que debe completar en los espacios indicados para tal efecto.

1. Un valor o medida de cierta característica –promedio, variancia, proporción, etc.– calculado a partir del total de elementos en la población se denomina _____ y el que se calcula con base en una muestra y se utiliza para representarlo se llama _____.

III Asocie. Cada concepto de la izquierda asócielo con la característica que le corresponda en la columna de la derecha.

a) Desviación estándar

b) Recorrido

c) Coeficiente de variación

d) Desviación media

___ Toma en cuenta solo el valor mayor y el menor del conjunto.

___ Para su cálculo se utilizan valores absolutos.

___ Su cuadrado recibe el nombre de variancia.

___ Se utiliza muy poco en la práctica.

___ Es una medida de dispersión relativa.

___ Utiliza los cuadrados de las desviaciones con respecto al promedio.

___ Se utiliza cuando se quiere comparar la variabilidad de dos o más conjuntos de datos.

IV. Falso y verdadero. Cada una de las cuatro aseveraciones que complementan la frase inserta a continuación, puede ser verdadera o falsa. Indíquelo con (V) o con (F), según corresponda.

1. Cuando se quiere comparar, en forma válida, la variabilidad de dos o más conjuntos de datos:

___ a) Se pueden utilizar las desviaciones estándares si los datos están dados en las mismas unidades y tienen magnitudes similares.

___ b) Se usa el coeficiente de variación si los promedios de los conjuntos son muy diferentes.

___ c) Se deben utilizar puntuaciones tipificadas.

___ d) Siempre se debe usar el coeficiente de variación.

V. Desarrollo. A continuación se le dan 6 preguntas que usted debe desarrollar según lo estudiado en el capítulo.

1. ¿Por qué es importante, cuando se analiza un conjunto de datos, además de conocer los valores que lo resumen o representan, medir su dispersión o variabilidad?

2. ¿Cuáles son los motivos por los cuales la fórmula para calcular la variancia debe modificarse cuando se tienen datos agrupados en una distribución de frecuencias?

3. En una fábrica mediana hay tres departamentos: administración, producción y distribución. Un consultor hizo un estudio en la empresa y recogió información sobre diferentes características de los empleados de cada uno de los tres departamentos, entre ellas el salario anual en dólares. Los principales resultados se resumen a continuación:

	Administración	Producción	Distribución
Número de empleados	20	50	30
Salario promedio	18 000	12 000	12 500
Moda	9000	8750	9150
Mediana	10 000	10 250	9950
Salario mayor	40 000	16 000	18 000
Salario menor	5000	5000	5000
Desviación estándar	6000	2000	2500
Coeficiente de variación	33%	17%	20%

   a) ¿En qué departamento ganan más los empleados? ¿Por qué piensa así?

   b) ¿En qué departamento los salarios presentan una mayor variación absoluta? Justifique su respuesta.

   c) ¿En qué departamento los salarios presentan una mayor variación relativa? ¿Por qué?

4. En el mismo estudio, antes citado, se preguntó a los empleados la frecuencia mensual con la que comían fuera de la casa. Los resultados obtenidos fueron los siguientes:

	Administración	Producción	Distribución
Número de empleados	20	50	30
Promedio	6	1	11
Variancia	2	1	3

   a) Encuentre el número promedio de veces por mes que comen fuera, para el conjunto de empleados de la fábrica.

   b) Calcule la variancia general, es decir, para el conjunto de todos los trabajadores.

c) Calcule R^2.

d) ¿Existe asociación entre la frecuencia con que se come fuera y el departamento donde se trabaja?

5. Un estudiante obtuvo una nota de 8,40 en el examen final de Matemáticas y una de 9,00 en el final de Castellano. La media aritmética y la desviación estándar, para el total de estudiantes, fueron: 7,60 y 1,00 en matemáticas y 8,20 y 1,60 en Castellano, respectivamente. ¿En cuál de las materias dio mejor rendimiento el estudiante? Razone su respuesta. La escala de calificación es de 0 a 10, con 7,00 como nota mínima para aprobar.

6. La empresa productora de lácteos "Dos Cipreses" realizó un pequeño estudio de mercado sobre la aceptación de sus productos en un barrio residencial de la capital del país, donde se entrevistaron 22 amas de casa. Una de las preguntas se refirió a si en el hogar se consumía regularmente algún tipo de yogur. Las respuestas fueron las siguientes:

Ama de casa N°	Consumo de yogur en el hogar	Ama de casa N°	Consumo de yogur en el hogar	Ama de casa N°	Consumo de yogur en el hogar
1	Si	9	No	17	Si
2	No	10	Si	18	No
3	No	11	No	19	No
4	Si	12	Si	20	Si
5	Si	13	No	21	No
6	No	14	Si	22	No
7	Si	15	No		
8	No	16	No		

a) Calcule la media y la variancia para el atributo consumo de yogur.

b) Si se sabe que en el barrio hay 300 residencias particulares, ¿en cuánto estimaría usted el número en las cuales se consume yogur?

RESPUESTA A LOS EJERCICIOS DE AUTOEVALUACIÓN

I. Selección única

1. b) 2. a) 3. c) 4. b)

II. Completar

1. Un valor o medida calculado a partir del total de elementos en la población se denomina valor de la población; el que se calcula con base en una muestra y se utiliza para representarlos se llama valor de la muestra o estimador.

III. Asocie

b) Toma en cuenta solo el valor mayor y el menor del conjunto.

d) Para su cálculo se utilizan valores absolutos.

a) Su cuadrado recibe el nombre de variancia.

d) Se utiliza muy poco en la práctica.

c) Es una medida de dispersión relativa.

a) Utiliza los cuadrados de las desviaciones con respecto a la media.

c) Se utiliza cuando se quiere comparar la variabilidad de dos o más conjuntos de datos.

IV. Falso o verdadero

a) V b) V c) F d) F

V. Desarrollo

1. Es importante porque la capacidad de un valor típico o de resumen para representar al conjunto de datos depende, lógicamente, en gran medida, de si estos se concentran alrededor de él. Cuanto más agrupados estén los datos individuales, alrededor del promedio, por ejemplo, mucha mayor confianza se tendrá en usar este valor para representar o caracterizar el conjunto de ellos.

2. Cuando se tienen los datos agrupados en una distribución de frecuencias no se conocen los valores individuales dentro de las clases y no es posible, por lo tanto, calcular las desviaciones de cada una de las observaciones con respecto a la media aritmética. Para solucionar este problema, se supone que cada una de las frecuencias de las clases es igual al punto medio y se procede luego a hacer la diferencia

entre el punto medio y la media aritmética, elevarla al cuadrado y multiplicarla por la frecuencia de la clase. La suma de todos estos productos, uno para cada clase, constituye una estimación de la suma de desviaciones al cuadrado que requiere el cálculo de la variancia.

3. a) Para determinar en cuál departamento ganan más los empleados, en un caso como este, en el que la distribución es claramente asimétrica, es necesario usar la moda o la mediana, que son poco afectadas por valores extremos. Dada la poca variación que presentan la moda y también la mediana entre dependencias, y el hecho de que el departamento con la moda más alta (distribución) tiene la mediana más baja, y el que tiene la mediana más alta tiene la moda más baja (producción). Parece razonable concluir que los empleados ganan, en "promedio", aproximadamente igual en los tres departamentos.

   b) En el de administración, al ser la desviación estándar de los salarios mucho mayor que en los otros dos departamentos.

   c) En el de administración porque su coeficiente de variación es el más elevado (33%).

4. a) El número promedio de veces por mes que comen fuera, para el conjunto de empleados, se calcula en la siguiente forma:

$$\mu_1 = \frac{N_1\mu_1 + N_2\mu_2 + N_3\mu_3}{N_1 + N_2 + N_3} = \frac{20 \cdot 6 + 50 \cdot 1 + 30 \cdot 11}{20 + 50 + 30} = \frac{120 + 50 + 330}{100} = 5.$$

   b) Variancia para el conjunto. Se aplica la fórmula correspondiente:

$$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2}{N} + \frac{N_1(\mu_1 - \mu_2)^2 + N_2(\mu_2 - \mu)^2 + N_3(\mu_3 - \mu)^2}{N}$$

$$\sigma^2 = \frac{20 \cdot 2 + 50 \cdot 1 + 30 \cdot 3}{100} + \frac{20(6-5)^2 + 50(1-5)^2 + 30(11-5)^2}{100}$$

$$= \frac{180}{100} + \frac{1900}{100} = 20{,}80.$$

   c) Valor de R^2

$$\sigma_E^2 = \frac{N_1(\mu_1 - \mu)^2 + N_2(\mu_2 - \mu)^2 + N_3(\mu_3 - \mu)^2}{N} = \frac{1900}{100} = 19.$$

$$R^2 = \frac{\sigma_E^2}{\sigma^2} = \frac{19}{20{,}80} = 0{,}9135.$$

   d) Sí existe asociación entre la frecuencia con que se come fuera y el departamento donde se trabaja, ya que la variancia de la variable comer fuera puede explicarse en un 91,35% por el departamento donde se trabaja.

5.

$x_m = 8,40 \quad \mu_m = 7,60 \quad \sigma_m = 1,00$ (Matemática)

$x_c = 9,00 \quad \mu_c = 8,20 \quad \sigma_c = 1,60$ (Castellano)

Dado que los promedios y las desviaciones estándar difieren, para poder determinar en qué materia dio mejor rendimiento el estudiante, hay que estandarizar las notas así:

$$\text{Matemática } z_m = \frac{x_m - \mu_m}{\sigma_m} = \frac{8,40 - 7,60}{1,00} = 0,8.$$

$$\text{Castellano } z_c = \frac{x_c - \mu_c}{\sigma_c} = \frac{9,00 - 8,20}{1,60} = 0,5.$$

A pesar de que en Castellano la nota fue superior, al considerar el rendimiento observado en cada grupo de alumnos, se tiene un mejor rendimiento en matemática, es decir, la posición relativa del estudiante en Matemática es superior a la de Castellano (0,8 > 0,5).

Con el procedimiento que se siguió, se obtiene un número el cual indica cuánto se aleja la nota específica (la de Matemática, por ejemplo) del promedio de su grupo, en términos de unidades de desviaciones estándares. Los valores de 0,8 y 0,5 son independientes de la magnitud de las notas y de su variabilidad y por ello permiten una comparación válida.

Para facilitar aún más la interpretación, se puede hacer la siguiente transformación para obtener notas estandarizadas con una media de 7 puntos (nota mínima para aprobar los cursos) y una desviación estándar de 1 punto.

Nota de Matemática $7 + 1 \cdot 0,8 = 7,80$.

Nota de Castellano $7 + 1 \cdot 0,5 = 7,50$.

Estas indican que si la nota promedio en cada asignatura fuera 7 y la desviación estándar 1, el estudiante recibiría un 7,8 en Matemática y un 7,5 en Castellano.

6. a) Para calcular la variancia y la media debe definirse una variable dicotómica x, que asume el valor 0 si en la vivienda no se consume yogur y 1 si se consume. La suma de los valores de X_i es entonces una suma de ceros y unos:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1 + 0 + 0 + \ldots + 1 + 0 + 0}{22} = \frac{9}{22} = 0,409 = p.$$

La media es realmente una proporción $p = 0,409$ o 40,9% e indica que 41% de las familias del barrio residencial consumen yogur.

La fórmula para calcular la variancia de la media sería la que corresponde a una proporción:

$$s^2 = \frac{n}{n-1}p(1-p) = \frac{22}{21} \cdot 0,409 \cdot (1 - 0,409) = 0,253.$$

b) Para estimar el número de residencias particulares del barrio donde se consume yogur, se multiplica el número total de viviendas (N = 300) por la proporción que consume yogur:

$$Np = 300 \cdot 0,409 = 123.$$

El número estimado de residencias particulares que consume yogur en el barrio es 123.