ESTADÍSTICA
GUIA No 5

MEDIDAS DE DISPERSIÓN

Introducción

Al estudiar características o variables de una población o muestra, siempre se manifiestan discrepancias o diferencias en los resultados individuales de las observaciones. La variabilidad es algo inherente a cada fenómeno aleatorio, y origina en ellos cierta homogeneidad o heterogeneidad, según que las discrepancias o diferencias sean pequeñas o grandes. A este grado de variabilidad, de diferencia entre observaciones es a lo que se llama dispersión.

Ocurre entonces, cuando se quiere asignar un número a cada grado de variabilidad, que surgen diferentes medidas de dispersión. Las definiciones de estas medidas se pueden establecer entre valores determinados de la muestra de datos o entre todos los valores y un valor de referencia, que suele ser una medida de tendencia central, como la media aritmética o la mediana, con el propósito de que la medición se vea poco influenciada por las propias unidades de medida de los valores cuya dispersión se desea estimar.

Se pueden reconocer al menos dos tipos generales de medidas de dispersión. Por ejemplo, Fernández y Fuentes (1995) sugieren distinguir entre dos tipos de medidas de dispersión. A las medidas de dispersión expresadas en términos de la misma unidad de medida que los datos, se las llaman medidas de dispersión absoluta, y a las que se expresan de manera adimensional, es decir, de manera independiente a las unidades de medición, las llaman medidas de dispersión relativa.

El reconocimiento de la existencia de la variabilidad como punto de partida para el estudio de la aleatoriedad y la construcción de modelos estadísticos, hace que las medidas de dispersión sean necesarias para efectuar comparaciones significativas entre grupos de observaciones. De hecho, cuando se mide la dispersión de los valores de una variable respecto a una de sus medidas de tendencia central, se está midiendo el grado de representatividad que dicha medida de tendencia central tiene respecto al conjunto de datos que pretende resumir. Así pues, a mayor dispersión se tendrá una menor representatividad de la medida de posición y viceversa. Además, la medición con este tipo de medidas debe ser no negativa y consonante con el nivel de dispersión en el sentido de que valores pequeños del estadígrafo en uso deben reflejar un nivel bajo de dispersión y viceversa.

Esta cuestión de la representatividad se puede precisar un poco más con un ejemplo. Suponga que en el estudio de dos grupos de familias A y B, de quince familias cada grupo, la distribución del número de hijos se tiene como se muestra en la siguiente tabla.

Grupo A		Grupo B	
Número de hijos x_i	Frecuencia f_i	Número de hijos x_i	Frecuencia f_i
1	2	0	7
2	11	2	1
3	2	4	7
Total	15	Total	15

Se puede observar que en ambos grupos la media aritmética del número de hijos es dos. Entonces a primera vista se puede afirmar que el comportamiento de los dos grupos es el mismo respecto al número de hijos. Sin embargo, es evidente que el grupo B, presenta los datos más dispersos que el grupo A. Por lo tanto la media aritmética es más representativa de lo que sucede en el grupo A, ya que en éste los resultados se apartan menos de la media aritmética que en el grupo B.

Como se acaba de ver en el ejemplo anterior la media aritmética caracteriza mejor al grupo A de familias que al B, respecto al número de hijos. En general, para caracterizar una distribución de frecuencias, las medidas de tendencia central se deben acompañar de una medida de dispersión que ponga de manifiesto el grado de representatividad del conjunto de datos.

Algunos ejemplos de medidas de dispersión son el recorrido, la desviación media, la desviación estándar, el rango medio, la desviación intercuartílica, la varianza y el coeficiente de variación. En lo que sigue, se hará una descripción de las medidas dispersión absoluta que son más utilizadas, las principales medidas de dispersión relativa, y finalmente se presentará una serie de ejemplos, para ilustrar el cálculo y utilización de las mismas.

5.1 Recorrido (Re).

El recorrido o rango de dispersión (Re), se define como la diferencia entre el valor máximo y el valor mínimo de los datos. Aunque se considera que es una medida imperfecta, cuando es razonable suponer que los datos se distribuyen de manera uniforme, entonces se espera que si, por ejemplo, el mínimo y el máximo están comprendidos entre 3 y 26, los datos presentarán más alejamiento mutuo que si los mismos datos están comprendidos entre 13 y 19, cuya diferencia es menor.

De todas maneras el rango tiene la ventaja de ser muy fácil de calcular y es recomendable tenerlo en cuenta cuando hay pocos datos por analizar. Sin embargo, el hecho de depender exclusivamente del máximo y el mínimo, puede ocasionar el que no refleje de manera apropiada la dispersión de una distribución de datos, cuando se tiene una buena cantidad de datos con valores intermedios. Además, no es posible su aplicación en los casos en que alguno de los valores, máximo o mínimo, como ocurre en ocasiones, quede indeterminado.

Este tipo de inconvenientes ponen de manifiesto la necesidad de considerar otras medidas de dispersión. Por ejemplo, cuando los valores próximos al máximo y el mínimo de una serie de datos están excesivamente alejados del resto, la consideración de un recorrido más corto, prescindiendo de un porcentaje determinado de los datos más alejados, puede dar una idea de la dispersión del conjunto de datos más acorde con la realidad, que si se emplea la diferencia entre los valores más extremos. Por ello, alternativas que algunas veces se contemplan son el intervalo intercuartílico ($Q_3 - Q_1$), el interdecílico ($D_9 - D_1$) o el intercentílico ($P_{99} - P_1$).

5.2 Desviaciones medias.

La suma de todas las desviaciones respecto a la media aritmética de una distribución de frecuencias, como se señaló en el capítulo anterior, vale cero. Por lo tanto, la media aritmética de dichas desviaciones no sirve para medir la dispersión de los valores de una variable. Sin embargo al considerar el valor absoluto de las desviaciones respecto a una medida de tendencia central como la media aritmética o la mediana, permite definir tres tipos de desviaciones que se comentan enseguida.

5.2.1 Desviación media absoluta.

La desviación media es la media aritmética de los valores absolutos de las diferencias de los datos respecto de la media aritmética. Con datos agrupados se puede escribir así:

$$D_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{k}\left|x_i - \bar{x}\right|f_i$$

Donde se tienen k valores diferentes de los datos o k intervalos de clase, según que la variable considerada sea discreta o continua, y N es el total de datos. Para datos sin agrupar se considera que n es el total de datos y se expresa así:

$$D_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Respecto a la desviación media es apropiado señalar que al considerar la función $D(u) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - u\right|$ asociada a los posibles promedios de los valores absolutos de la desviaciones respecto a u, se puede demostrar (ver por ejemplo, Cansado (1967)) que el punto en que se minimiza esta función es en el valor de la mediana. Por ello, si se usan desviaciones medias para cuantificar la dispersión, quizás sea preferible utilizar el promedio de los valores absolutos de las desviaciones respecto a la mediana, medida que se pasa a considerar enseguida.

5.2.2 Desviación media respecto a la mediana.

La desviación media respecto a la mediana es la media aritmética de los valores absolutos de las desviaciones de los datos respecto a la mediana y se puede expresar para datos agrupados como:

$$D_{Me} = \frac{1}{N}\sum_{i=1}^{k}\left|x_i - Me\right|f_i$$

Y para datos sin agrupar se expresa como $D_{Me} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$.

Las letras k, N, n, etcétera, tienen la misma interpretación que en el caso de la desviación media.

5.2.3 Desviación mediana.

La desviación mediana se define como la mediana de la distribución cuyos valores son las desviaciones, en valor absoluto, de los datos respecto a la mediana. Por ejemplo, si los valores de una variable son 2, 4, 8, 11, 13, 17 y 21, su mediana es Me = 11. De manera que los valores absolutos de las desviaciones respecto a la mediana son 0, 2, 3, 6, 7, 9 y 10, cuya mediana es 6, por lo tanto la desviación median es 6.

La interpretación que se le puede dar a la desviación mediana es similar a la que se le puede dar a la desviación intercuartílica ($Q_3 - Q_1$), en el sentido de recoger la variación entre el 50% de los datos intermedios. En realidad, cuando la distribución es simétrica, ambas medidas coinciden.

5.3 Varianza (S^2).

La varianza es una de las medidas de dispersión más mencionadas en la literatura estadística. En realidad de todas las medidas de dispersión la varianza y la desviación estándar (que se presenta en el siguiente apartado), son las más importantes para un desarrollo teórico de la estadística. El propósito de la varianza es medir la mayor o menor dispersión de los valores de una distribución de datos respecto a la media aritmética. Cuanto mayor sea la varianza mayor dispersión existirá y por tanto menor representatividad se podrá atribuir a la media aritmética. En términos agrupados la varianza se define como:

$$S^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$

Y para datos sin agrupar, se define así:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Quizás el principal problema con la varianza es que su valor no se exprese en las mismas unidades que la variable analizada, sino elevada al cuadrado, lo cual dificulta su interpretación. No obstante, debido a sus propiedades matemáticas la varianza goza de excelente reputación.

Algunas de las propiedades que se pueden destacar de la varianza son las siguientes:

- Si se considera la función de variable real definida como $F(u) = \frac{1}{N}\sum_{i=1}^{k}(x_i - u)^2 n_i$ se tiene que valor donde es mínima para u es la media aritmética.
- Por la manera cono está definida, una suma de cuadrados, nunca es negativa y sólo puede ser nula cuando todos los valores son iguales.
- Además, si $y_i = k.x_i + c$ entonces $S_y^2 = k^2 S_x^2$
- La siguiente igualdad también se utiliza con frecuencia

$$\frac{1}{N}\sum_{i=1}^{k}(x_i - u)^2 n_i = \frac{1}{N}\sum_{i=1}^{k}(x_i)^2 n_i - (\bar{x})^2$$

5.4 Desviación estándar (S).

Ya se ha dicho que la varianza no viene expresada en las mismas unidades de medida que las de los datos. Sin embargo, la raíz cuadrada de la varianza nos lleva a la desviación estándar también conocida como desviación típica. Se define como la raíz cuadrada con signo positivo de la varianza. En su versión para datos agrupados, se presenta así:

$$S = +\sqrt{S^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}$$

Y para datos sin agrupar así:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La desviación estándar es la más utilizada entre las medidas de dispersión y satisface las mismas propiedades que se mencionaron para la varianza. Sin embargo, otra propiedad, no mencionada antes, que es interesante y relevante mencionar, se deduce a continuación.

Suponga que $x_1, x_2, \ldots, x_n$, es una colección de valores numéricos de los datos de una distribución. Entre todas las diferencias $(x_i - \bar{x})^2$ para i =1,2, …, n seleccione todas aquellas diferencias cuyos valores x_i verifiquen la desigualdad $|x_i - \bar{x}| \geq k$, donde k designa un número positivo.

Ahora suponga que $(x_{i1} - \bar{x})^2$, $(x_{i2} - \bar{x})^2, \ldots, (x_{ip} - \bar{x})^2$ son las p cantidades que satisfacen la desigualdad. Entonces

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \frac{1}{n}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2$$

Por otra parte, como $|x_{ij} - \bar{x}| \geq k$ para j = 1,2, …, p, se tiene entonces que $|x_{ij} - \bar{x}|^2 \geq k^2$ y por lo tanto $\sum_{j=1}^{p}(x_{ij} - \bar{x})^2 \geq \sum_{j=1}^{p} k^2 = pk^2$, por lo tanto

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \frac{1}{n}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2 \geq k^2\frac{p}{n}$$

Nótese que el cociente p/n que aparece al final de la desigualdad representa la frecuencia relativa de los x_i tal que $|x_i - \bar{x}| \geq k$. Si p/n se denota más bien como $fr(|x_i - \bar{x}| \geq k)$, entonces se tiene que

$$\frac{S^2}{k^2} \geq fr(|x_i - \bar{x}| \geq k)$$

Pero dado que en una distribución de frecuencias se satisface la igualdad $fr(|x_i - \bar{x}| \geq k) + fr(|x_i - \bar{x}| < k) = 1$, entonces se llega a:

$$fr(|x_i - \bar{x}| < k) \geq 1 - \frac{S^2}{k^2}$$

Si ahora se elige el valor de k como tS^2, la desigualdad anterior se transforma en la siguiente:

$$fr(|x_i - \bar{x}| < tS^2) \geq 1 - \frac{1}{t^2}$$

La desigualdad obtenida se puede ver como la interpretación frecuencial de la llamada desigualdad de Tchevichev utilizada en estadística matemática y teoría de la probabilidad. Para este caso le da el siguiente sentido a la desviación estándar: la proporción de datos que caen en el intervalo $(\bar{x} - tS, \bar{x} + tS)$ es a lo menos $1 - (1/t^2)$. Por ejemplo, la proporción de datos incluidos en el intervalo $(\bar{x} - 2S, \bar{x} + 2S)$ es al menos

$1 - (1/2^2) = ¾$, es decir, del 75% del total; mientras que los datos que caen en el intervalo $(\overline{x} - 3S, \overline{x} + 3S)$ es como mínimo del $1 - (1/3^2) = 8/9 = 0{,}88$, que equivale al 88%. Se ve pues, que la desviación estándar es una medida bastante precisa de la dispersión de los datos en torno a la media aritmética de la distribución y por ello goza de tanta reputación.

Para finalizar, se tiene que la desviación estándar siempre dará un valor mayor o igual al de la desviación media, puesto que la media cuadrática de las observaciones $|x_i - \overline{x}|$ es mayor o igual que la media aritmética de éstas, es decir $D_{\overline{x}} \leq S$.

5.5 Coeficiente de variación media de Pearson (CV_x).

Todas las anteriores medidas de dispersión que fueran consideradas antes, son medidas de dispersión absoluta, ya que se expresan en términos de la unidad que se utiliza para hacer mediciones. Las medidas de dispersión relativa, evaden este problema al considerar cocientes entre una medida de dispersión absoluta (excepto la varianza) y una medida de tendencia central.

En este sentido el coeficiente de variación media de Pearson, indica la relación existente entre la desviación típica de una muestra y su media, ya que se define como $CVx = \frac{S}{\overline{X}}$.

Al dividir la desviación típica por la media se convierte la medición en un valor libre de la unidad de medida. Así pues, si comparamos la dispersión en varios conjuntos de observaciones, el que tenga menor dispersión será el que tenga menor coeficiente de variación.

Este coeficiente es quizás el más importante y fiable de las medidas de dispersión relativa, entre otras razones por venir expresado en términos de dos estadísticas bien reconocidas que en general son objetivas y representativas de un conjunto de datos. Además, permite comparaciones de variación de conjuntos de datos expresados en diferentes unidades de medida. El principal inconveniente del coeficiente de variación media de Pearson (y de otros coeficientes definidos de manera similar), es que al ser un coeficiente inversamente proporcional a la media aritmética, cuando ésta tome valores cercanos a cero, a menos que se lleve a cabo un cambio de origen en los datos.

5.7 Ejemplos.

Ejemplo 1. Altura de unas palmeras

Las alturas de 5 palmeras son 4 metros, 6 metros, 10 metros, 8 metros y 20 metros. Si las medidas se cambian a decímetros, ¿cómo cambiará la desviación estándar?

a) Aumentará en 10
b) Disminuirá en 10
c) Aumentará en un factor de 10
d) Disminuirá en un factor de 10
e) No cambiará

Discusión. Este ítem pretende valorar si se reconoce la manera como se afecta la desviación estándar cuando se introduce un cambio en la escala de los datos y en este caso la respuesta correcta es la opción (c). Los distractores (a) y (b), expresan que el cambio en la escala de los datos tiene un efecto aditivo., lo cual es falso. La opción (d) aunque sugiere que si hay un cambio multiplicativo no se reconoce el sentido correcto en que se da. Y por último, la elección de la opción (e) sugiere que se piensa equivocadamente, en que la desviación estándar es invariante ante cambios de escala.

Ejemplo 2. Trabajo perdido en una empresa.

Durante los últimos veinte días laborables, el número total de horas de trabajo perdidas diariamente en una empresa de cien obreros viene dada por los datos: 1, 3, 1, 1, 2, 4, 2, 2, 1, 2, 800, 6, 8, 400, 1, 5, 4, 6, 3, 1.

a) Si se supone que la jornada laboral es de ocho horas diarias, ¿qué porcentaje medio de horas se han perdido en esos días?

b) Encuentre la desviación absoluta media, y la desviación media respecto a la mediana y con base en esta información valore, entre la media y la mediana, cuál de ellas es más representativa de la tendencia central de los datos.

Discusión. En esta empresa el número de horas diarias de trabajo corresponde a 100 x 8 = 800. Si se denota con x_i el número de horas de trabajo perdidas en un día i, el cociente $x_i/800$ representa la proporción de horas de trabajo en ese día. También se puede expresar $x_i/800$ en términos porcentuales multiplicando por 100. Así, $(100 \cdot x_i)/800 = x_i/8$ %.

En la tabla que sigue se organiza la información de los datos suministrados.

Horas perdidas por días x_i	Frecuencia absoluta f_i	Porcentaje por día $x_i/8$ %	Porcentaje total %
1	6	0,125	0,750
2	4	0,250	1,000
3	2	0,375	0,750
4	2	0,500	1,000
5	1	0,625	0,625
6	2	0,750	1,500
8	1	1,000	1,000
400	1	50,000	50,000
800	1	100,000	100,000
Total	N=15		156,625

El porcentaje medio de horas perdidas a lo largo de los veinte días es la media aritmética de los porcentajes totales (última columna de la tabla). Por lo tanto el porcentaje medio de horas de trabajo perdidas en términos de la media aritmética es 156,625/20 = 7,831.

Para determinar el valor de la desviación absoluta media respecto a la media aritmética y respecto a la mediana, se organizan los cálculos intermedios en la siguiente tabla.

x_i	f_i	F_i	$x_i \cdot f_i$	$\lvert x_i - \bar{x} \rvert \cdot f_i$	$\lvert x_i - Me \rvert \cdot f_i$
1	6	6	6	369,90	9
2	4	10	8	242,60	2
3	2	12	6	119,30	1
4	2	14	8	117,30	3
5	1	15	5	57,65	2,5
6	2	17	12	113,30	7
8	1	18	8	54,65	5,5
400	1	19	400	337,35	397,5
800	1	20	800	737,35	797,5
Total	20		1.253	2.149,40	1.225,0
Medias			62,65	107,50	61,3

La media aritmética de horas diarias de trabajo perdido es $\bar{x}$ = 1253/20 = 62,65. Entonces la desviación media respecto a la media aritmética se obtiene del cociente 2149,4/20 = 107,5. La mediana de horas diarias de trabajo perdido se ubica entre el dato 10 y el 11, por lo que entonces la mediana es (2+3)/2 = 2,5. Por lo tanto, la desviación absoluta respecto a la mediana se obtiene del cociente 1255/20 = 61,3.

El tamaño de la desviación absoluta media respecto a la media aritmética sugiere poca representatividad para la media aritmética. En realidad, la desviación absoluta media respecto a la media aritmética viene más influenciada por los valores extremos 400 y 800, algo atípicos dentro de la serie de datos. La mediana, al considerar los datos extremos no por su valor sino por la posición que ocupan dentro del conjunto ordenado de los datos, refleja de forma más realista la tendencia central. De hecho el valor de la desviación absoluta media respecto a la media aritmética, casi duplica el valor de la desviación absoluta media respecto a la mediana. Las consideraciones anteriores sugieren entonces que la mediana es más representativa que la media.

Ejemplo 3. Valoración de la gestión del alcalde.

Para conocer la conformidad de los habitantes de Bogotá, acerca de la gestión realizada por el actual alcalde de la ciudad, durante el periodo en el que ha despeñando sus funciones, se practicó una encuesta de opinión a 740 personas, en donde se calificaba la gestión del alcalde en una escala de 0 a 10. Los resultados de la encuesta fueron los que se muestran en la siguiente tabla. Determine la media aritmética de las calificaciones arrojadas por la encuesta y estime la representatividad de dicha media.

Calificación de la gestión	Número de encuestados
[0, 1)	50
[1, 3)	60
[3, 4)	90
[4, 6)	100
[6, 8)	240
[8, 9)	120
[9, 10]	80

Discusión. Una disposición práctica para exhibir los cálculos que se requieren para hallar la media y la varianza de la muestra se presentan en la tabla de la página siguiente.

De los datos de la tabla se puede encontrar la media aritmética como 4420/740 = 5,97. La varianza resulta del cociente 5104, 46/ 740 = 6,9 y la desviación estándar es la raíz cuadrada de 6,9, es decir, 2,63. También es posible hallar la varianza con la expresión alternativa dada por $\frac{1}{N}\sum_{i=1}^{k}(x_i)^2 f_i - (\bar{x})^2$ de donde se obtiene $(31.505/740 - (5,97)^2 = 6,9$.

Calificación	f_i	x_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$	$x_i^2 \times f_i$
[0, 1)	50	0,5	25	1497,67	12,5
[1, 3)	60	2,0	120	947,07	240,0
[3, 4)	90	3,5	315	550,40	1102,5
[4, 6)	100	5,0	500	94,67	2500,0
[6, 8)	240	7,0	1680	253,15	11760,0
[8, 9)	120	8,5	1020	766,30	8670,0
[9, 10]	80	9,5	760	995,19	7220,0
Total	740		4420	5104,46	31505,0

Observe que el valor de la desviación estándar resulta ser menor que una vez el valor de la media aritmética. Si este hecho se considera como criterio práctico, se tiene que la media es aceptablemente representativa.

Ejemplo 4. Reacción ante una vacuna para la gripe

Como parte de una investigación para combatir la gripe común, un grupo de 500 personas se distribuyó en cincuenta grupos de de diez personas cada grupo y se les aplicó una vacuna experimental. Luego se anotó el número de personas por grupo que presentó reacción ante la vacuna. Los datos obtenidos se muestran en la siguiente tabla:

Número de personas por grupo que reaccionan a la vacuna	0	1	2	3	4	5	6	7	8	9	10
Número de grupos	9	9	8	8	5	3	3	2	1	1	1

a) Encuentre la media aritmética y la desviación estándar σ del número de personas por grupo que tuvieron reacción ante la vacuna.

b) ¿Qué porcentaje de personas reacciona ante la vacuna entre $(\bar{x} - \sigma, \bar{x} + \sigma)$ y entre $(\bar{x} - 2\sigma, \bar{x} + 2\sigma)$?. σ es la desviación estándar de la variable X.

Discusión. Una disposición práctica para exhibir los cálculos para hallar la media y la desviación estándar de la muestra se presentan en la siguiente tabla, donde x_i denota el número de personas por grupo con reacción ante la vacuna y f_i el número de grupos.

x_i	f_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$
0	9	0	73,62
1	9	9	31,14
2	8	16	5,92
3	8	24	0,16
4	5	20	6,50
5	3	15	13,74
6	3	18	29,58
7	2	14	34,28
8	1	8	26,42
9	1	9	37,70
10	1	10	50,98
Total	50	143	310,02

Con base en la información de la tabla se tiene que la media aritmética se obtiene como 143/50 = 2,86. Para la varianza se calcula 310,02/50 = 6,2, de donde la desviación estándar, al sacar la raíz cuadrada, da 2,49.

En cuanto al literal (b,) se tiene que entre $(\bar{x} - S) = 2{,}86 - 2{,}49 = 0{,}37$ y $(\bar{x} + S) = 2{,}86 + 2{,}49 = 5{,}35$, hay 1x9 + 2x8 + 3x8 + 4x5 + 5x3 = 84 personas, mientras que entre $(\bar{x} - 2S) = 2{,}86 - 2x(2{,}49) = -2.12$ y $(\bar{x} + S) = 2{,}86 + 2x(2{,}49) = 7{,}84$, hay 9 + 84 + 6x3 + 7x2 = 125 personas. En el primer caso el porcentaje de personas a una desviación de la media es de 84/143 = 58,74% y a dos desviaciones de la media hay 125/143 = 87,41%.

Observe que los resultados son consistentes, con lo que dice la versión frecuencial de la llamada desigualdad de Tchevichev.

Ejemplo 5. Temperaturas registradas en un observatorio

En un observatorio meteorológico de Canadá se llevó un registro de las temperaturas, en grados centígrados, durante los primeros 59 días del año 2008 y se anotaron en la tabla que se muestra a continuación.

Temperatura (ºC)	Número de días
[-12, -8)	2
[-8, -5)	4
[-5, -2)	8
[-2, 0)	18
[0, 4)	17
[4, 6)	6
[6, 8)	3
[8, 10]	1

a) Encuentre los coeficientes de variación cuartílica y de variación media de Pearson y evalué cuál de los dos coeficientes mide de manera más fiable la dispersión relativa de las temperaturas.

b) Si se transforma la medición de la temperaturas de la escala de grados centígrados a la escala de grados Fahrenheit (ºF = 32 + 9/5xºC) ¿Cuál coeficiente resulta más fiable?

Discusión. Dado que para calcular los coeficientes de variación cuartílica y media de Pearson se requiere determinar el valor de los cuartiles primero y tercero, la media aritmética y la desviación estándar, en la tabla de la página siguiente se disponen algunos de los cálculos requeridos. Para encontrar los cuartiles se debe empezar por determinar las posiciones de los cuartiles las cuales resultan de calcular N/4 = 59/4 = 14,75 y 3N/4 = (3x59)/4 = 44,25. Entonces, aplicando la fórmula general dada en el ejercicio 26 del capítulo anterior, para establecer el valor de un cuantil i, tomando s = 3, es decir:

$$C_i(s) = L_{i-1} + \frac{\frac{i \cdot N}{s} - F_{i-1}}{f_i} \cdot a_i \text{ para } i = 1, 2, \ldots, s\text{-}1.$$

donde L_{i-1}, f_i y a_i designa el límite inferior, la frecuencia absoluta y la amplitud del intervalo, respectivamente, de la clase a la que pertenece el cuantil y F_{i-1} la frecuencia acumulada absoluta de la clase anterior a ella. Así se obtiene

$$Q_1 = -2 + \frac{14,75 - 14}{18} \times 2 = -1,971 \text{ y } Q_3 = 0 + \frac{44,25 - 32}{17} \times 4 = 2,882$$

Temperatura (ºC)	f_i	x_i	F_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$
[-12, -8)	2	-10,0	2	-20	201,36
[-8, -5)	4	-6,5	6	-26	170,77
[-5, -2)	8	-3,5	14	-28	99,91
[-2, 0)	18	-1,0	32	-18	19,24
[0, 4)	17	2,0	49	34	65,71
[4, 6)	6	5,0	55	30	147,97
[6, 8)	3	7,0	58	21	145,58
[8, 10]	1	9,0	59	9	80,39
Total	59			2	930,93

La media aritmética de la temperatura es 2/59 = 0,034 ºC, la varianza se obtiene de 930,93/59 = 15,78, y la desviación estándar se obtiene al sacar la raíz cuadrada a este número dando 3,972. ºC. De lo anterior se llega a que el coeficiente de variación cuartílica es:

$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{2,882 - (-1,971)}{2,882 + (-1,971)} = 4,973$$

Mientras que el coeficiente de variación de Pearson da:

$$V_{\bar{x}} = \frac{S}{\bar{x}} = \frac{3,972}{0,034} \cong 117,17 = 11,71\%$$

Como se puede notar, el valor del coeficiente de variación de Pearson resulta muy distorsionado debido a la proximidad de la media aritmética al valor cero. En este caso resulta más razonable utilizar el coeficiente de variación cuartílica.

Ahora bien, cuando se cambia la escala de los datos aplicando la relación ºF = 32 + 9/5xºC, se obtiene la siguiente tabla de frecuencias.

Temperatura (ºF)	f_i	x_i	F_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$
[10,4; 17,6)	2	14,0	2	28,0	652,40
[17,6; 23,0)	4	20,3	6	81,2	553,29
[23,0; 28,4)	8	25,7	14	205,6	323,70
[28,4; 32,0)	18	30,2	32	543,6	62,34
[32,0; 39,2)	17	35,6	49	605,2	212,91
[39,2; 42,8)	6	41,0	55	246	479,43
[42,8; 46,4)	3	44,6	58	133,8	471,68
[46,4; 50,0]	1	48,2	59	48,2	260,47
Total	59			1891,6	3016,22

Ahora los cuartiles inferior y superior vienen dados por

$$Q_1 = 28,4 + \frac{14,75 - 14}{18} \times 3,6 = 28,55 \text{ y } Q_3 = 32 + \frac{44,25 - 32}{17} \times 7,2 = 37,19$$

Con estos resultados se obtiene el coeficiente de variación cuartílica y el coeficiente de variación de Pearson así:

$$V_Q = \frac{37,19 - 28,55}{37,19 + 28,55} = 0,131 = 13,1\%$$

$$V_{\bar{x}} = \frac{\sqrt{3016,22/59}}{1891,6/59} \cong 0,223 = 22,3\%$$

En este caso con ambos coeficientes se manifiesta una baja dispersión relativa, siendo el coeficiente de variación de Pearson más fiable que el de la variación cuartílica, dado que el primero tiene en cuenta toda la información de los datos, mientras que el segundo solamente la posición ordenada de los valores de los datos.

Ejemplo 6. Pesos de dos grupos de estudiantes.

El médico de un colegio registra las medias aritméticas y las varianzas de los pesos de dos grupos A y B, como se muestran en la siguiente tabla:

Grupo	Media	Varianza
A	64 kg	1 ,4 kg^2
B	68 kg	1,1 kg^2

a) Si se sabe que la media aritmética de los dos grupos es 67, ¿en qué proporción están los tamaños de los dos grupos A y B?

b) ¿Cuál es la varianza conjunta de los dos grupos?

Discusión. Suponga que N_A y N_B son los tamaños de la muestras de los grupos A y B. Como 67 corresponde a la media ponderada de las medias de los grupos A y B, se puede plantear que $67 = \frac{N_A \cdot 64 + N_B \cdot 68}{N_A + N_B}$. De donde se tiene que $67 \cdot (N_A + N_B) = 64 \cdot N_A + 68 \cdot N_B$, entonces $(67 - 64) \cdot N_A = (68 - 67) \cdot N_B$; que es lo mismo que $3 \cdot N_A = N_B$. Es decir, N_A y N_B están en proporción de uno a tres.

Para encontrar la varianza ponderada se requiere realizar un poco de álgebra. Supóngase que x_i, x_2, …, x_{N_A}, $\bar{x}$ y S_x^2 son los pesos del grupo A, su media y su varianza, respectivamente, y que y_i, y_2, …, y_{N_B}, $\bar{y}$ y s_y^2 los pesos, la media y la varianza relativas al grupo B. Si $\bar{z}$ y S_z^2 representa la media y la varianza del grupo completo se tiene que:

$$S_z^2 = \frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_A}(x_i - \bar{z})^2 + \sum_{i=1}^{N_B}(y_i - \bar{z})^2\right] = \frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_A}((x_i - \bar{x}) + (\bar{x} - \bar{z}))^2 + \sum_{i=1}^{N_B}((y_i - \bar{y}) + (\bar{y} - \bar{z}))^2\right]$$

$$= \frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_A}(x_i - \bar{x})^2 + 2(\bar{x} - \bar{z}) \cdot \sum_{i=1}^{N_A}(x_i - \bar{x}) + \sum_{i=1}^{N_A}(\bar{x} - \bar{z})^2\right] +$$

$$\frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_B}(y_i - \bar{y})^2 + 2(\bar{y} - \bar{z}) \cdot \sum_{i=1}^{N_B}(y_i - \bar{y}) + \sum_{i=1}^{N_A}(\bar{y} - \bar{z})^2\right]$$

Pero dado que $\sum_{i=1}^{N_A}(x_i - \bar{x}) = \sum_{i=1}^{N_B}(y_i - \bar{y}) = 0$, entonces se tiene que:

$$S_z^2 = \frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_A}(x_i - \bar{x})^2 + N_A(\bar{x} - \bar{z})^2\right] + \frac{1}{N_A + N_B}\left[\sum_{i=1}^{N_B}(y_i - \bar{y})^2 + N_B(\bar{y} - \bar{z})^2\right]$$

$$= \frac{N_A}{N_A + N_B}\left[\frac{1}{N_A}\sum_{i=1}^{N_A}(x_i - \bar{x})^2 + (\bar{x} - \bar{z})^2\right] + \frac{N_B}{N_A + N_B}\left[\frac{1}{N_B}\sum_{i=1}^{N_B}(y_i - \bar{y})^2 + (\bar{y} - \bar{z})^2\right]$$

$$= \frac{N_A}{N_A + N_B}\left[S_x^2 + (\bar{x} - \bar{z})^2\right] + \frac{N_B}{N_A + N_B}\left[S_y^2 + (\bar{y} - \bar{z})^2\right]$$

Por lo tanto $S_z^2 = \frac{N_A\left[S_x^2 + (\bar{x} - \bar{z})^2\right] + N_B\left[S_y^2 + (\bar{y} - \bar{z})^2\right]}{N_A + N_B}$

Reemplazando los datos de medias y varianzas dados en el enunciado y expresando N_B en términos de N_A se obtiene:

$$S_z^2 = \frac{N_A\left[1,4 + (64 - 67)^2\right] + 3N_A\left[1,1 + (68 - 67)^2\right]}{N_A + 3N_A} = 4,175$$

Observe que aunque las varianzas de cada grupo son relativamente pequeñas, la del grupo en conjunto es casi cuatro veces más grande. Esto pone de manifiesto una diferencia significativa entre los valores de las medias de cada grupo.

Ejemplo 7. Tiempo de atención en un hospital.

En un hospital se ha llevado el registro, sobre el tiempo de espera para ser atendidos, de los últimos 320 pacientes que han acudido a la unidad de atención de urgencias. Los datos se presentan en la siguiente tabla:

Determine la media aritmética y la mediana de esta distribución de datos y mida la dispersión de los datos en torno a estas estimaciones de tendencia central.

Tiempo de espera	f_i
[0; 5)	3
[5;10)	31
[10; 15)	102
[15; 20)	63
[20; 25)	54
[25; 30)	43
[30; 35)	12
[35; 40)	6
[40; 45)	5
[45; 50]	1
Total	320

Discusión. Para empezar vale la pena recordar que la representatividad de la media se debe evaluar con la desviación estándar, mientras que la de la mediana es preferible evaluarla con base en la desviación media respecto a la mediana.

En la tabla que sigue se presentan los primeros cálculos para hallar los valores de las estimaciones requeridas.

Tiempo de espera	f_i	x_i	F_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$	$\lvert x_i - \bar{x} \rvert \cdot f_i$	$\lvert x_i - Me \rvert \cdot f_i$
[0; 5)	3	2,5	3	7,5	759,03	14,4	43,2
[5;10)	31	7,5	34	232,5	3687,33	9,4	291,4
[10; 15)	102	12,5	136	1275,0	3558,15	4,4	448,8
[15; 20)	63	17,5	199	1102,5	51,74	0,6	37,8
[20; 25)	54	22,5	253	1215,0	904,97	5,6	302,4
[25; 30)	43	27,5	296	1182,5	3555,94	10,6	455,8
[30; 35)	12	32,5	308	390,0	2383,61	15,6	187,2
[35; 40)	6	37,5	314	225,0	2187,43	20,6	123,6
[40; 45)	5	42,5	319	212,5	2902,54	25,6	128,0
[45; 50]	1	47,5	320	47,5	846,45	30,6	30,6
Total	320			5890	20837,19		2048,8

La media aritmética se obtiene como 5890/320 = 18,41 minutos. Para obtener la mediana, primero ubicamos la posición la calcular N/2 = 320/2 = 160; entonces la mediana es Me =15 + [(160-136)/63]x5 = 16,9 minutos. Para la obtención de la desviación estándar, se le saca la raíz cuadrada a la varianza dada por 20.837, 19/2 = 65,11, para obtener 8,06 minutos.

En cuanto a la obtención de la desviación media respecto a la mediana, resulta de 2.048,8/320 = 6,4 minutos.

El valor de la desviación estándar en relación con el de la media aritmética es 2,28 veces menor que la media aritmética, mientras que en el caso de la desviación media respecto a la mediana es de 2,64 veces menor que la mediana. Como hay una diferencia de (2,64-2,28) = 0,36, bajo el criterio mencionado antes, es preferible utilizar la mediana. Sin embargo, el valor un poco más alto de la media aritmética advierte que hay algunos pocos pacientes que tienen que esperar tiempos muy grandes.

Ejemplo 8. Variaciones en la crianza de animales.

Se tienen dos zoocriaderos de iguanas, cada uno con 200 iguanas. En el zoocriadero A los animales son alimentados con una mezcla de sorgo-yerbas-harina de plátano, mientras que los animales del zoocriadero B son alimentados con una mezcla de maíz-yerbas-harina de yuca. Estas diferencias en la alimentación han acarreado desarrollos desordenados en las iguanas. Dos empleados, Anatoly y Boris, son encargados de observar, medir y clasificar los animales. Anatoly se encargó del zoocriadero A y Boris del zoocriadero B. Los empleados entregaron las siguientes tablas:

¿En cuál de los dos zoocriaderos se presenta mayor desorden en el desarrollo de los animales?

Peso (lb)	Cantidad
[1.5-2.0)	15
[2.0-2.5)	20
[2.5-3.0)	25
[3.0-3.5)	30
[3.5-4.0)	30
[4.0-4.5)	35
[4.5-5.0)	45

Longitud (cm)	Cantidad
[35.5-44.0)	45
[44.0-52.5)	35
[52.5-61.0)	30
[61.0-69.5)	30
[69.5-78.0)	25
[78.0-86.5)	20
[86.5-95.0)	15

Discusión. Vale aclarar que para comparar la dispersión de dos conjuntos de datos en donde se manejen diferentes unidades de medidas, se debe usar el coeficiente de dispersión de Pearson.

Como los datos están agrupados y las variables son continuas se requieren las marcas de clases, las cuales aparecen en las siguientes tablas para los diferentes intervalos de clases.

Peso (lb)	Cantidad	Marcas
[1.5-2.0)	15	1.75
[2.0-2.5)	20	2.25
[2.5-3.0)	25	2.75
[3.0-3.5)	30	3.25
[3.5-4.0)	30	3.75
[4.0-4.5)	35	4.25
[4.5-5.0)	45	4.75

Longitud (cm)	Cantidad	Marcas
[35.5-44.0)	45	39.75
[44.0-52.5)	35	48.25
[52.5-61.0)	30	56.75
[61.0-69.5)	30	65.25
[69.5-78.0)	25	73.75
[78.0-86.5)	20	82.25
[86.5-95.0)	15	90.75

Sea P la variable Peso (en libras) y L la variable longitud (en centímetros). Si f_k y M_k son las frecuencias absolutas y las marcas de clases, respectivamente, entonces $\bar{P} = \frac{\sum_{k=1}^{k=7} f_k M_k}{200} = 3.5625$ y $\bar{L} = \frac{\sum_{k=1}^{k=7} f_k M_k}{200} = 59.9375$ son las medias aritméticas de las variables P y L.

Las varianzas de las variables P y L son $S_P^2 = \frac{\sum_{k=1}^{k=7} f_k \left(M_k - \bar{P}\right)^2}{200} = 0.92109375$ y $S_L^2 = \frac{\sum_{k=1}^{k=7} f_k \left(M_k - \bar{L}\right)^2}{200} = 266.196094$, y las desviaciones estándar son S_P=0.9597 y S_L=16.3155. Por lo tanto, los coeficientes de variación son $CV_P = \frac{S_P}{\bar{P}} = 0.26939$ y $CV_L = \frac{S_L}{\bar{L}} = 0.27221$, para las variables P y L respectivamente.

Como se puede apreciar, los desarrollos han sido muy similares en los dos conjuntos de datos, presentándose ligeramente mayor variación en el zoocriadero de Boris.

Ejemplo 9. La recta que mejor se ajusta a los puntos.

Se tienen diez puntos y dos rectas. Los puntos son A(1,4), B(2,2), C(3,5), D(4,3), E(5,6), F(6,4), G(7,6), H(8,4), J(9,8) y K(10,4). Las ecuaciones de las rectas son -4x+15y=45 y -4x+15y=47. ¿Cuál de las dos rectas se ajusta mejor al conjunto de puntos?

Discusión. Una manera de determinar cuál de las dos rectas se ajusta mejor al conjunto de puntos es considerar las distancias verticales entre los puntos y cada una de las rectas, y luego calcular la varianza o desviación estándar de estas distancias para cada recta. Por ejemplo, la distancia vertical entre un punto $P(x_1,y_1)$ y una recta con ecuación y=mx+b es $|mx_1 + b - y_1|$. Al conjunto de distancias se le calcula la media aritmética y finalmente se calcula la desviación estándar. El conjunto de distancias con menor dispersión corresponde a las distancias de la recta que mejor se ajusta, la cuál es la recta "más cercana" al conjunto de puntos.

A continuación se presentan estos cálculos para las rectas $y1 = \frac{45 + 4x}{15}$ y $y2 = \frac{47 + 4x}{15}$.

X_k	Y_k	$Y1_k = \frac{45 + 4X_k}{15}$	$\lvert Y1_k - Y_k \rvert$	$\lvert \overline{Y1} - Y_k \rvert^2$
1	4	3,267	0,733	0,360
2	2	3,533	1,533	0,040
3	5	3,800	1,200	0,018
4	3	4,067	1,067	0,071
5	6	4,333	1,667	0,111
6	4	4,600	0,600	0,538
7	6	4,867	1,133	0,040
8	4	5,133	1,133	0,040
9	8	5,400	2,600	1,604
10	4	5,667	1,667	0,111

Tabla: Distancias con respecto -4x+15y=45.

La media aritmética de las distancias $|Y1_k - Y_k|$ es $\overline{Y1} = \frac{\sum_{k=1}^{k=7} |Y1_k - Y_k|}{10} = 1.333$;

La varianza de las distancias $|Y1_k - Y_k|$ es $S^2(Y1) = \frac{\sum_{k=1}^{k=7} |\overline{Y1} - Y_k|^2}{10} = 0.293$ y la desviación estándar es S(Y1)=0.542.

X_k	Y_k	$Y2_k = \frac{47 + 4X_k}{15}$	$\lvert Y2_k - Y_k \rvert$	$\lvert \overline{Y2} - Y_k \rvert^2$
1	4	3,400	0,600	0,538
2	2	3,667	1,667	0,111
3	5	3,933	1,067	0,071
4	3	4,200	1,200	0,018
5	6	4,467	1,533	0,040

6	4	4,733	0,733	0,360
7	6	5,000	1,000	0,111
8	4	5,267	1,267	0,004
9	8	5,533	2,467	1,284
10	4	5,800	1,800	0,218

Tabla: Distancias con respecto -4x+15y=47.

La media aritmética de las distancias $|Y2_k - Y_k|$ es $\overline{Y2} = \frac{\sum_{k=1}^{k=7} |Y2_k - Y_k|}{10} = 1.333$;

La varianza de las distancias $|Y2_k - Y_k|$ es $S^2(Y2) = \frac{\sum_{k=1}^{k=7} |\overline{Y2} - Y_k|^2}{10} = 0.276$ y la desviación estándar es S(Y2)=0.525.

Al comparar las dispersiones se concluye que la recta que mejor se ajusta al conjunto de puntos es -4x+5y=47.

5.8 Ejercicios.

1. A continuación se presenta la información dada por diez estudiantes con respecto a la distancia, medida en cuadras, del lugar en donde ellos viven, al colegio en donde estudian.

40	50	30	45	45	48	35	60	36	10

a) ¿Con qué medidas estadísticas se puede resumir la distancia que tiene que recorrer un estudiante para ir de su hogar al colegio? ¿Alguna de esas medidas es más apropiada? Explique.

b) ¿Con base en qué medida estadística se puede resumir la variabilidad de las distancias recorridas por los estudiantes? ¿Alguna de esas medidas es más apropiada? Explique.

c) ¿Qué representaciones gráficas se podrían utilizar para ilustrar la situación? ¿Alguna de esas representaciones gráficas es más apropiada? Explique.

2. La siguiente información presenta los datos en miles de pesos de los salarios de secretarias que trabajan en cuatro empresas diferentes:

Empresa 1:	350	400	350	2100		
Empresa 2:	350	400	350	400	550	
Empresa 3:	350	350	350	350	1300	
Empresa 4:	300	400	500	600	700	800

¿Con qué medidas estadísticas de tendencia central y de dispersión sería apropiado resumir el comportamiento de los salarios de las secretarias de cada una de las empresas anteriores?

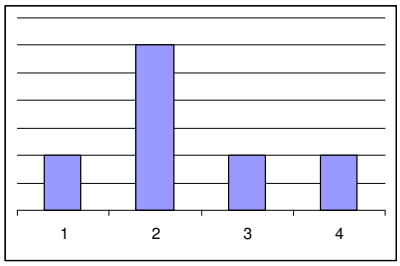
3. Construya un conjunto de diez datos que tenga un promedio de 39.9 y una desviación estándar de 0.

4. Proponga tres conjuntos, cada uno de 10 datos que satisfagan las siguientes condiciones: promedio 6 y desviación estándar 1; promedio 10 y desviación estándar 1; promedio 7 y desviación estándar 2.

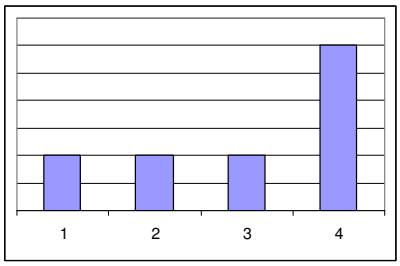
5. Construya un conjunto de diez datos con las siguientes características: promedio 39.9; que todos los datos sean diferentes; y que la distancia entre cualquier par de datos contiguos, una vez ordenados de manera ascendente o descendente, sea la misma. Con respecto al valor de la desviación estándar que se obtuvo con los datos iniciales del ejercicio 1, ¿qué relación de orden espera encontrar entre las desviaciones estándar correspondientes a la distribución del ejercicio 1 y a la que acaba de construir? ¿qué efecto puede tener sobre la desviación el aumentar o disminuir la distancia entre los datos?

6. Construya un conjunto de diez datos con las siguientes tres características: promedio de 39.9; que los datos contengan sólo dos valores diferentes, y tal que los dos valores diferentes ocurran con distinta frecuencia. Bajo las condiciones anteriores, intente establecer una relación entre los dos valores de frecuencias de los datos y las dos distancias de los datos al promedio.

7. A continuación se presenta la representación gráfica de un par de distribuciones:



Distribución 1.



Distribución 2.

¿Cuál de las dos distribuciones le parece que es más dispersa? ¿Qué efecto puede tener sobre los valores de las medidas de dispersión, el que las frecuencias de los valores de las distribuciones anteriores se cambien pero manteniéndose la misma relación de 3 a 1 que se insinúa en las gráficas?

8. Construya un conjunto de diez datos con las siguientes tres características: promedio de 39.9; que los datos contengan sólo dos valores diferentes, y tal que los dos valores diferentes ocurran con igual frecuencia. Luego calcule el rango y la desviación estándar. Luego, proponga otros conjuntos que satisfagan las mismas condiciones anteriores y trate de identificar un patrón de relación entre la desviación estándar y el rango.

9. Construya dos nuevos conjuntos de datos U y V, que satisfagan simultáneamente la siguientes condiciones: la desviación estándar de los elementos de U debe ser mayor que la desviación estándar de los elementos de V, el rango de los elementos de U debe ser menor que el rango de los elementos de V.

10. En un zoocriadero destinado a la cría de chigüiros se ha descuidado la alimentación de estos animales y se ha presentado un desarrollo inesperado. Se han clasificado los animales en 10 grupos, teniendo en cuenta sus pesos en kilogramos. La siguiente tabla muestra la cantidad de animales en cada categoría de pesos:

Pesos	Cantidad de animales
35.00 - 40.00	20
40.10 - 45.00	25
45.10 - 50.00	30
50.10 - 55.00	10
55.10 - 60.00	15
60.10 - 65.00	20
65.10 - 70.00	25
70.10 - 75.00	35
75.10 - 80.00	10
80.10 - 85.00	10

a) Calcule la media y la desviación estándar para estos datos y evalúe la representatividad de la media como medida de tendencia central, ¿Será preferible la mediana?

b) Verifique la versión frecuencial de la desigualdad de Tchevichev para los casos de dos desviaciones respecto a la media y tres desviaciones respecto a la media

11. En un colegio, los estudiantes de grado 10 se reparten en cuatro grupos {A, B, C, D} de igual cantidad de estudiantes para las asignaturas no deportivas. Se practica el examen final de física. La siguiente tabla muestra las calificaciones obtenidas por los estudiantes en cada grupo:

A	12	56	36	52	52	57	43	35	50	31	38
	72	67	31	51	66	53	52	61	60	38	63
	45	77	24	52	51	35	49	43	90	54	46
B	52	77	45	49	57	66	67	61	50	68	49
	64	66	46	68	57	52	63	50	59	47	52
	64	46	12	66	79	62	29	50	45	39	73
C	33	34	49	36	55	60	57	54	45	47	69
	84	56	39	52	88	36	60	61	54	65	47
	52	42	56	25	37	46	57	65	65	63	52
D	56	70	38	69	57	60	82	66	25	58	58
	61	53	44	74	73	60	23	50	33	51	55
	33	61	62	71	56	77	77	46	57	39	49

a) ¿Qué porcentaje x de las notas de los estudiantes satisface las desigualdades?

(I) $\bar{x} - \sigma < x < \bar{x} + \sigma$ (II) $\bar{x} - 2\sigma < x < \bar{x} + 2\sigma$ (III) $\bar{x} - 3\sigma < x < \bar{x} + 3\sigma$

b) ¿En cuál de las asignaturas se presenta mayor dispersión?

12. Como parte de un programa de control de calidad en la producción de baterías para usar en diferentes aparatos eléctricos, se someten a una prueba de duración 64 baterías de tipo A y 105 baterías de tipo B, provenientes de dos fabricantes diferentes. Los resultados obtenidos se organizan en la siguiente tabla:

Tiempo de duración (en días)	Tipo A (frecuencia)	Tipo B (frecuencia)
[90; 120)	6	7
[120; 150)	9	12
[150; 180)	18	31
[180; 210)	21	29
[210; 240)	7	22
[240; 270)	3	4

a) Compare la variabilidad de ambas distribuciones de datos en términos de coeficientes de dispersión relativa.

b) Comente acerca de la fiabilidad de los coeficientes que fueron considerados en el literal anterior.

13. Un examen de Cálculo se aplicó a los cuatro grupos de grado 11 de una institución. En la siguiente tabla se presentan las frecuencias absolutas.

Notas	Grupo A	Grupo B	Grupo C	Grupo D
[0.0 – 0.5)	8	14	2	12
[0.5 – 1.0)	6	6	4	14
[1.0 – 1.5)	8	4	6	8
[1.5 – 2.0)	10	8	10	6
[2.0 – 2.5)	12	4	12	5
[2.5 – 3.0)	8	8	10	5
[3.0 – 3.5)	8	8	6	6
[3.5 – 4.0)	6	6	4	8
[4.0 – 4.5)	14	10	2	14
[4.5 – 5.0]	8	6	14	12

Calcular para cada grupo el Coeficiente de Variación de Pearson y ordénelos de menor a mayor grado de heterogeneidad.

14. En un zoocriadero destinado a la cría de chigüiros para exportación se ha descuidado la alimentación de los animales y se ha presentado un desarrollo inesperado en estos. Se han clasificado los animales en 10 grupos, teniendo en cuenta sus pesos en kilogramos. La siguiente tabla muestra la cantidad de animales en cada categoría de pesos:

Pesos	Cantidad de animales
35.00 - 40.00	20
40.10 - 45.00	25
45.10 - 50.00	30
50.10 - 55.00	10
55.10 - 60.00	15
60.10 - 65.00	20
65.10 - 70.00	25
70.10 - 75.00	35
75.10 - 80.00	10
80.10 - 85.00	10

a) Calcule la media y la desviación estándar para estos datos y evalúe la representatividad de la media como medida de tendencia central, ¿Será preferible la mediana?

b) Verifique la versión frecuencial de la desigualdad de Tchevichev para los casos de una desviación respecto a la media y dos desviaciones respecto a la media

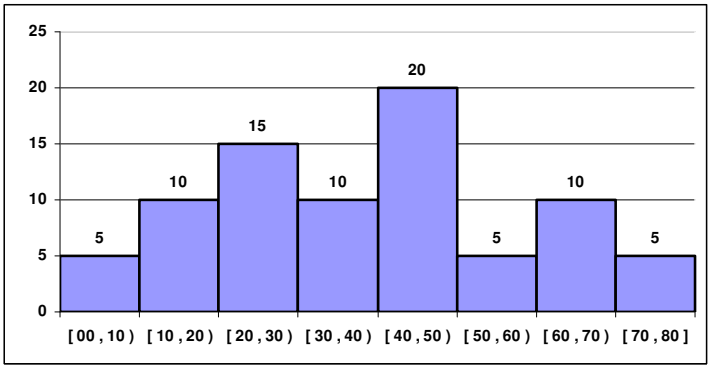
15. Se tienen dos zoocriaderos (A y B) de iguanas, cada uno con 200 iguanas. En el zoocriadero A los animales son alimentados con una mezcla de sorgo-yerbas-harina de plátano, mientras que los animales del zoocriadero B son alimentados con una mezcla de maíz-yerbas-harina de yuca. Estas diferencias en la alimentación han producido desarrollos desordenados en las iguanas. Dos empleados, Anatoly y Boris, son encargados de observar, medir y clasificar los animales. Anatoly se encargó del zoocriadero A y Boris del zoocriadero B. Desafortunadamente Anatoly tomó los pesos y Boris tomó la longitud nariz-cola, y con eso entregaron las siguientes tablas:

Pesos (kg)	Cantidad
[4,00-6,50)	20
[6,50-9,00)	35
[9,00-11,5)	55
[11,5-14,0)	40
[14,0-16,5)	25
[16,5-18,0)	15
[18,0-20,5]	10

Longitudes (cm)	Cantidad
[50-57)	10
[57-64)	20
[64-71)	20
[71-78)	25
[78-85)	45
[85-92)	55
[92-99]	25

De acuerdo al coeficiente de variación, ¿en cuál de los dos zoocriaderos se presenta mayor desorden en el desarrollo de los animales?

16. El siguiente diagrama representa la distribución de frecuencias de los valores de una variable continua X. Calcule el coeficiente de variación de Pearson.



17. La serie final de un campeonato de béisbol fue disputada por los equipos A y B, durante la temporada en cada equipo participaron 40 jugadores. Al final de la serie se contabilizaron los batazos de hit

conectados por los dos equipos y se construyeron las distribuciones que se muestra en las siguientes tablas:

Equipo A	
Hits	Jugadores
100-125	2
125-150	3
151-175	5
176-200	1
201-225	9
226-250	8
251-275	7
275-300	5

Equipo B	
Hits	Jugadores
125-160	8
161-195	7
196-230	5
231-265	6
266-270	4
271-305	5
306-340	3
341-375	2

¿En cuál de los dos equipos el ritmo de bateo fue más homogéneo durante la temporada?

18. Un embarque de 20 computadoras similares que se envía a un distribuidor contiene 8 aparatos defectuosos. Una escuela escoge aleatoriamente 10 de estas computadoras y las compra. Se define la variable aleatoria X como el número de computadoras defectuosas entre las computadoras compradas. ¿Cuál es la varianza de la variable X?

19. Un juego consiste en lanzar cinco dados, apostar $1000 y ganar $1000 por cada "cinco" que aparezca, es decir, si le salen n cincos se gana 1000n pesos. Otro juego consiste en lanzar seis dados, apostar $1250 y ganar $1250 por cada "cinco" que aparezca, es decir, si le salen n cincos se gana 1250n pesos. En ambos juegos, si al jugador no le sale el número apostado, entonces pierde el doble del dinero apostado. ¿En cual de los dos juegos varía en mayor grado la ganancia?

20. En un salón de juegos se encuentran dos objetos (A y B) de tiro al blanco, los cuales están formados por 5 círculos concéntricos de radios 10 cm, 20 cm, 30 cm, 40 cm y 50 cm. Un hombre que dispara al blanco en el objeto A recibe 50 puntos, 40 puntos, 30 puntos, 20 puntos o 10 puntos, según pegue en la zona 1 (círculo pequeño), zona 2, zona 3, zona 4 o zona 5 (anillos circulares). Un hombre que dispara al blanco en el objeto B recibe 45 puntos, 40 puntos, 35 puntos, 30 puntos o 20 puntos, según pegue en la zona 1 (círculo pequeño), zona 2, zona 3, zona 4 o zona 5 (anillos circulares). La probabilidad de que el disparo haga contacto con cualquiera de las 5 zonas del blanco es 1/3, y la probabilidad de no dar en el blanco es 2/3. Si X se define como el puntaje ganado por jugador que dispara en el objeto A, y Y se define como el puntaje ganado por un jugador que dispara en el objeto B, ¿En cuál de los dos objetos hay mayor variabilidad en las ganancias obtenidas?