

Resumiendo Datos Numéricamente

CAPÍTULO 4

4.1. INTRODUCCIÓN

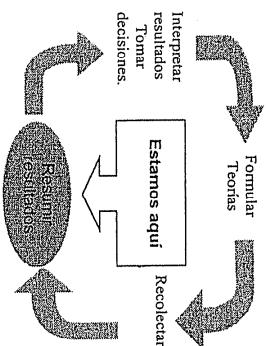
En el Capítulo 3 comenzamos a resumir y organizar nuestros datos para conseguir información de ellos. Nos concentraremos en cómo presentar nuestros datos gráficamente. Los gráficos nos permiten responder preguntas tales como:

- ¿Están la mayoría de los valores concentrados en el medio con pocos valores hacia ambos lados?
- ¿Qué valor(es) se presenta(n) (ocurre/n) con mayor frecuencia?
- ¿Hay algún valor "raro" o "extremo" comparado con el resto de los datos?

En este capítulo aprenderemos cómo tomar nuestros datos, nuestro conjunto de mediciones y resumirlos de una manera útil. Mejoraremos nuestras exhibiciones gráficas presentando varias medidas de resumen de los datos. Comenzaremos con la idea de "centro" de los datos. ¿Cuál es un valor promedio o típico? Analizaremos dos medidas del centro: la media aritmética y la mediana y presentaremos otra medida llamada moda. Veremos que estas diferentes medidas de posición nos conducen a distintas interpretaciones sobre el mismo conjunto de datos. Como vimos en el Capítulo 3, los datos varían de unidad a unidad. Otro aspecto del conjunto de datos es la dispersión de los valores.

¿Cuán dispersos están los valores? ¿Varían sobre un gran rango o están bastante cerca unos de otros? Así como los distintos tipos de datos nos conducen a distintos tipos de gráficos, ellos nos llevan a diferentes tipos de medidas de resumen numéricas. Nos inclinaremos hacia las medidas resumenes numéricas de variables cuantitativas y analizaremos algunas alternativas si los datos son cualitativos.

El propósito de este capítulo es demostrar la utilidad de unos números bien elegidos para resumir datos que han sido seleccionados.



4.2. MEDIDAS CENTRALES

En el capítulo 3 analizamos la característica: edad que influye en la presión sanguínea. El conjunto de edades del Conjunto Data 3, de los 20 individuos:



Sujeto número	Género	Edad	Dosis: de tabletas ingeridas	Presión sanguínea diastólica	
				Al comienzo del estudio	Al final del estudio
1001	M	45	2	100.2	100.1
1002	M	41	1	98.5	100.0
1003	F	51	2	100.8	101.1
1004	F	46	2	101.1	100.9
1005	F	47	3	100.0	99.8
1006	M	42	2	100.7	100.2
1007	M	43	4	100.7	100.7
1008	F	50	2	100.3	100.9
1009	M	39	1	100.5	101.0
1010	M	32	1	99.9	98.5
1011	M	41	2	101.0	101.4
1012	M	44	2	100.9	100.8
1013	F	47	2	97.4	96.2
1014	F	49	3	98.8	99.6
1015	M	45	3	100.9	100.0
1016	F	42	1	101.1	100.1
1017	M	41	2	100.7	100.3
1018	F	40	1	97.8	98.1
1019	M	45	2	100.0	100.4
1020	M	37	3	101.5	100.8

Supongamos que tiene que quedar un único número que representa la edad más típica (característica) de los 20 sujetos.

¿Qué número elegiría? Probablemente elegiría un número central del centro de la distribución de la edad. Las medidas centrales (o medidas de posición central) son: **media aritmética**, **mediana**, **media**, **fracciones**, etc. Por ahora trataremos la mediana y la media aritmética.

Si los datos provienen de una muestra, la media y la mediana son estadísticas, en cambio si se calculan con todos los datos de la población, son parámetros.

MEDIA ARITMÉTICA •••

Definición

La media aritmética de un conjunto de observaciones, se obtiene sumando las observaciones y dividiendo por el número de observaciones (n).

Para hallar la media aritmética, simplemente sumamos todos los valores observados de la variable y dividimos por el número total de observaciones.

Ejemplo: para calcular la edad promedio de los 20 sujetos en el estudio médico:

Esta es la medida más común del centro y tiene una notación con la cual nos familiarizaremos:

$$\frac{45+41+51+46+47+\dots+45+37}{20} = 43,35 \text{ años}$$

Si $x_1, x_2, x_3, \dots, x_n$ denota una muestra de n observaciones, luego la media de la muestra es llamada **x̄** y se denota:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Si en lugar de una muestra, tenemos todos los valores observados de la variable de la población, podríamos computar la media de la población de la misma manera: sumando todos los valores y dividiendo por el número total.

La media poblacional se simboliza con la letra griega μ .

EJEMPLO 4.1. Número Medio de hijos por familia

Los siguientes datos corresponden al número de hijos por familias provenientes de una muestra aleatoria simple de 10 hogares de un mismo barrio: 2, 3, 0, 2, 1, 0, 3, 0, 1, 4. La media aritmética (o comúnmente llamada promedio) de estas 10 observaciones es:

$$\bar{x} = \frac{2+3+0+2+1+0+3+0+1+4}{10} = \frac{16}{10} = 1,6$$

Obtuvimos un promedio de 1,6 hijos, si bien no es posible tener 1,6 hijos en cualquier familia, no se redondea a 2. Estamos hablando de promedio.

Suponga ahora que la observación de la última familia en un listado previo fue registrado incorrectamente, como 40 en lugar de 4. ¿Qué le pasaría a la media?

$$\bar{x} = \frac{2+3+0+2+1+0+3+0+1+40}{10} = \frac{52}{10} = 5,2$$

Advierta que 9 de las 10 observaciones son menores que la media.

La media es sensible a las observaciones extremas.

La mayoría de las demostraciones gráficas habrían detectado esta observación extrema.



¿La Media Aritmética es Siempre el Centro?

Suponga una muestra de tamaño 10.

- ¿Puede la media aritmética ser mayor que el máximo valor o menor que el mínimo valor? Si contesta sí, dé un ejemplo.

- ¿Puede la media aritmética ser el punto medio entre el mínimo y el máximo valor (cuando el mínimo no es igual al máximo)? Si contesta sí, dé un ejemplo.

- ¿Puede la media aritmética ser exactamente el 2º valor más pequeño, cuando ellos son ordenados de menor a mayor (no todas las observaciones iguales)? Si contesta sí, dé un ejemplo.

Estamos descubriendo con este ejemplo que la media puede ser cualquier percentil (los percentiles se presentan en la próxima sección).

PARA RESOLVER!!!

4.1. Una media aritmética no siempre es representativa

Los resultados de los parciales de Gastón son: 7, 98, 25, 19, y 26. Calcule la nota promedio. Explique por qué la media no es un buen trabajo para resumir las notas de Gastón.

PARA RESOLVER!!!

4.2. Medianas Combinadas

La nota promedio de 3 estudiantes es 54 y la nota promedio de otros 4 estudiantes es 76. ¿Cuál es la nota promedio de los 7 estudiantes?

EJEMPLO 4.2: Salario Promedio

1. Se obtiene una muestra de tamaño 10 de salarios mensuales. Suponga que 9 de los valores son iguales a \$1.000.- y un valor es igual a \$10.000.-.

¿Es la media una buena medida promedio?

La media es:

$$\bar{x} = \frac{9(1.000)+10.000}{10} = \frac{19.000}{10} = 1.900,$$

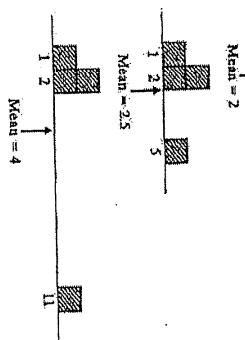
la cual no mide el centro, dado que el 90% de los salarios son inferiores a la media.

2. Se selecciona una muestra de tamaño 10. Suponga que 9 valores son \$1.000.- y el otro es \$100.000.-. ¿Cómo es esta media aritmética con respecto a la media calculada en el punto 1.?

$$\bar{x} = \frac{9(1.000)+100.000}{10} = \frac{109.000}{10} = 10.900,$$

la cual es más grande que la media de la parte 1.
La media se desvía, no hacia los valores que son iguales, sino más bien **hacia el valor extremo**. Esto puede verse a través de la representación física de la media.

La media es también definida como el punto de equilibrio... el punto donde la distribución balancearía. Si la distribución es simétrica como en la primera figura, la media estará exactamente en el centro de la distribución.



Como la observación más grande está ubicada a la derecha, la media se traslada hacia la observación extrema.

Si la distribución es asimétrica, podemos pensar en una **medida más resistente del centro**.

Una medida de tendencia central que es resistente a los valores extremos es la **mediana**.

Definición

La mediana de un conjunto de n observaciones ordenadas de menor a mayor, es un valor tal que la mitad de las observaciones es menor o igual a ese valor, y la otra mitad de las observaciones es mayor o igual a ese valor.

Para encontrar la mediana de 5 números: 4, 7, 3, 9, 5; primero necesitamos ordenar los valores observados de la variable de menor a mayor: 3, 4, 5, 7, 9. Dado que el número de observaciones es impar, la mediana que está en el centro de la secuencia ordenada de los datos es 5. Aclaro que dos observaciones son menores que 5 y dos observaciones son mayores que 5.

Si el número de observaciones es par, la mediana es cualquier número comprendido entre las dos observaciones del centro. Por ello, la mediana es el promedio de los valores numéricos correspondientes a estas dos observaciones centrales.

Así, la mediana de: 3, 4, 5, 7, 9, 11 es el promedio de 5 y 7, o sea, 6.

En general, la **posición** de la mediana puede ser encontrada haciendo $(n+1)/2$, donde n es el número de observaciones. Si $(n+1)/2$ es un número entero, contamos desde el menor valor tanto lugares como lo indica ese número y esa será la posición que ocupa la mediana. Si $(n+1)/2$ no da un número exacto, entonces contamos $(n/2)$ observaciones desde el más pequeño y promediamos esa observación con la siguiente más alta.

Las observaciones correspondientes a la variable edad de los 20 sujetos sometidos al estudio médico ordenadas:

32	37	39	40	41	41	41	42	42	43	44	45	45	45	46	47	47	49	50	51
									↗	↗									

Mediana = 43,5

10^a obs. 11^a obs.

PARA RESOLVER!!

4.3.

Encuentre la mediana del número de hijos correspondiente a una muestra de 10 familias.

**Nº de observación: 1 2 3 4 5 6 7 8 9 10
Nº de chicos: 2 3 0 1 4 0 3 0 1 2**

- a) Ordene las observaciones de menor a mayor.
- b) Localice el lugar que ocupa la mediana..
- c) Mediana = _____
- d) ¿Qué le pasa a la mediana si la observación # 5 en el primer listado fue incorrectamente listada como 40 en lugar de 4?
- e) ¿Qué ocurre si la observación # 3 del primer listado fue incorrectamente registrada como -20 en lugar de 0?

Observación

La **mediana es resistente**, es decir, ella no cambia, o cambia muy poco, con las observaciones extremas.

Otra medida... LA MODA

Definición

La moda de un conjunto de observaciones es el valor de la variable que ocurre con mayor frecuencia. Es el valor que tiene la frecuencia más alta entre todas las observaciones.

Para hallar la moda, será útil ordenar las observaciones para ver cuán a menudo ocurre, cada valor.

La moda de los valores: $\{0, 0, 0, 1, 1, 2, 2, 3, 4\}$ es 0 porque es el valor que ocurre más a menudo -un total de tres veces-.

Para las siguientes observaciones: $\{0, 0, 0, 1, 1, 2, 2, 3, 4\}$ hay dos modas. El 0 y el 2 porque ellos son los más frecuentes. Este conjunto es bimodal.

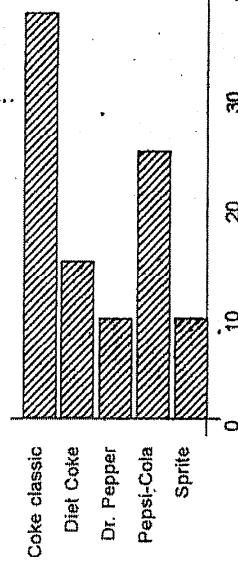
- ¿Cuál sería la moda del siguiente conjunto de datos? $\{0, 10, 2, 4, 5, 8\}$ Aquí cada uno ocurre solamente una vez. En lugar de referirnos a todas las observaciones como una moda, es práctico decir que los datos no presentan moda.

- Para el conjunto de valores: $\{0, 0, 0, 0, 1, 2, 3, 4, 4, 4, 5\}$, podríamos decir que hay dos modas 0 y 4 porque éstos son los más frecuentes entre valores vecinos.

A veces la moda no es usada como una medida de centro dado que el valor más frecuente podría estar lejos del centro de la distribución; sin embargo, es una medida que puede ser tenida en cuenta para datos cualitativos.

Consideré el gráfico de barras, el cual muestra el porcentaje de bebidas gaseosas compradas, clasificadas según marcas. [Ejemplo 3, capítulo 3]. Podemos observar que la categoría modal es la que corresponde a "Coke Classic".

Distribución de las bebidas carbonatadas clasificadas según marcas



Fuente: Datos obtenidos en base a una muestra aleatoria de 50 compras Beverage Digest. 1998

Este gráfico muestra que el mayor porcentaje de compras es el correspondiente a Coke Classic (representada por la barra más alta).

Si las distintas categorías fueran codificadas como:

- 1: Coke Classic
- 2: Diet Coke
- 3: Dr. Pepper
- 4: Pepsi-Cola
- 5: Sprite

No es razonable calcular la media ni la mediana para la variable tipo de gaseosas, dado que esa codificación fue arbitraria.

Los valores numéricos sólo sirven para distinguir las categorías y pueden ser ordenados ni promediados. A veces, las categorías suelen ser presentadas en orden: del menor al mayor porcentaje.

EJEMPLO 4.3: Diferentes medidas pueden dar diferentes impresiones

El famoso trío: media; mediana y moda, representa tres métodos diferentes para encontrar las llamadas medidas de tendencia central o de posición.

Cuando ellas resultan diferentes, nos pueden conducir a diferentes interpretaciones de los datos que estamos resumiendo.

Consideremos los ingresos anuales de 5 familias de una misma zona:

\$12.000.- \$12.000.- \$30.000.- \$90.000.- \$100.000.-

- ¿Cuál es el ingreso típico o característico para este grupo?

El **ingreso medio** es: \$48.800,-

$$\bar{x} = \frac{100.000 + 90.000 + 30.000 + 12.000 + 12.000}{5} = 48.800$$

La **mediana** es: \$30.000.

La **moda** es: \$12.000.-

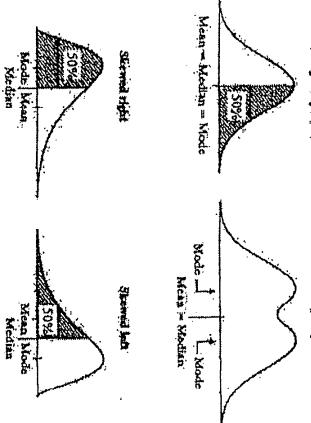
Si estamos tratando de promocionar a la zona como opulenta, tributaria, podríamos preferir el ingreso medio. Si tratamos de argumentar en contra de un incremento en los impuestos podríamos decir que los ingresos son demasiado bajos como para implementar un incremento en los impuestos y por lo tanto elegiríamos la moda. Si deseamos representar estos valores con el ingreso que está en el medio, registráramos la mediana. Tres medidas diferentes, cada una válida e informativa en las distintas situaciones.

¿QUÉ MEDIDA DE POSICIÓN CENTRAL USAR?

La moda no es muy usada, si el valor más frecuente está lejos del centro de la distribución.

La mediana, principalmente usa la información ordenada. La media usa los valores numéricos reales.

Los siguientes gráficos muestran la relación entre la media, mediana y moda.



La moda es el valor de la variable donde la distribución presenta un máximo. Dado que las áreas de un histograma representan frecuencias, la mediana es el valor tal que la mitad del área bajo la distribución está a la izquierda y la otra mitad yace a la derecha de ella.

La media es el centro de gravedad de la distribución, el punto en el cual la distribución balancearía.

Como puede observarse, si la distribución es aproximadamente simétrica, la media y la mediana estarán bastante cerca una de otra. Si tenemos una distribución que es asimétrica, la mediana será preferida como medida central. La mediana es una medida resistente del centro, resistente a los efectos de las observaciones extremas.

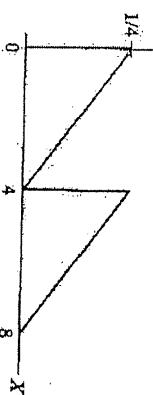
PARA PENSAR!!

Suponga que calcula la media, mediana y moda para un listado de números.

- ¿Cuál será el número de la lista que siempre aparecerá?
- Si la distribución es simétrica ¿Cuál medida de tendencia central calcularía? ¿La media o la mediana? Por qué?

PARA RESOLVER!! 4.4. Una distribución diferente

La distribución para una variable X continua está dada por la siguiente gráfica:



- a) ¿Es simétrica la distribución? **SÍ** **NO**
 b) ¿Cuál es el valor numérico de la mediana de esta distribución?

Mayor que 4 menor que 4 igual a 4

¿Por qué?

PARA RESOLVER!

4.5. La utilidad de la aleatorización

Recordemos que aprendimos la importancia de diseñar experimentos cuidadosamente. La asignación aleatoria de los sujetos a los distintos tratamientos fue un método usado para reducir el impacto del sesgo.

La aleatorización tiende a balancear los grupos con respecto a otros factores que pueden ser desconocidos por el investigador, pero podrían afectar las conclusiones.

Considere un estudio para comparar dos tipos de antibióticos para el tratamiento de anginas (estreptococos) en los chicos: Amoxicilina y Cefadroxil. En un centro sanitario, 23 chicos fueron aleatoriamente asignados para uno de los dos tratamientos. Se pensó que la edad de los chicos podría influir sobre la efectividad de los antibióticos. Las edades de los chicos para cada tratamiento se presentan a continuación.

Calcule la media, mediana y moda para cada uno de los grupos.

Compare los grupos con respecto a la edad.

T1: Amoxicilina (n=11)

Edad: 14 17 11 10 11 14 9 12 8 10 9

Media:

Mediana:

Moda:

T2: Cefadroxil (n=12)

Edad: 9 14 8 10 13 7 9 11 16 10 12 9

Media:

Mediana:

Moda:

• ¿Cómo son los grupos comparados con respecto a la edad?

Notas

- Cuando vea o escuche sobre el registro de un promedio, pregunte cuál fue computado: la media o la mediana.

- Piense o examine la distribución de los valores para afirmar si la medida de centro usada es la apropiada.



MEDIA, MEDIANA Y MODA

La medida más común de tendencia central es la media. Se calcula sumando todas las observaciones de un conjunto de datos, dividiendo ese total por el número total de elementos involucrados. La media está afectada por valores extremos (outliers y valores alejados que se encuentran en la cola de una distribución asimétrica). De allí que la media sea una buena elección como medida de centro de una distribución unimodal y aproximadamente simétrica, sin outliers.

La mediana es la medida más robusta del centro, es decir, que no está influenciada por los valores extremos. La mediana está en la mitad de las observaciones, cuando los datos están ordenados del más pequeño al más grande.

Si se tiene un número impar de valores, la mediana está en el centro de la secuencia ordenada; si en cambio se tiene un número par de valores la mediana es el promedio de los dos valores centrales y cae exactamente en la mitad de ellos. Si se tiene n observaciones luego $(n+1)/2$ nos da la posición de la mediana.

Para distribuciones asimétricas o distribuciones con outliers, la mediana tiende a ser la mejor elección como medida de tendencia central.

La moda es (son) el (los) valor(es) que ocurre(n) con mayor frecuencia. Para una distribución el modo es valor asociado con el "pico" más alto.

El valor más frecuente puede estar lejos del centro de la distribución, en ese caso la moda no es una medida real del centro. Sin embargo, la moda es la única de las tres que puede ser usada para datos cualitativos.

Ejercicios

4.1. Un artículo titulado "Un nuevo hogar sobre la Costanera" (La Capital, 7/5/99) reportó la mediana de los alquileres para los meses diciembre 98, enero, febrero y marzo 99. Explique por qué se registró la mediana en lugar de la media (o promedio) de los alquileres de las casas ubicadas sobre el río Paraná.

4.2. La columna del "Ask Marilyn" (Parade Magazine, Junio 11, 96) analiza las tres medidas de tendencia central en un intento de aprender cuál valor corresponde a un valor "típico" registrado:

Comentario del lector: "Leí un ensayo sobre sexo. El mismo decía que un hombre típico tiene seis parejas sexuales en su vida y que una mujer típica tiene 2. Suponiendo que el hombre es heterosexual y dado que el número de hombres y mujeres es aproximadamente igual, ¿Cómo puede ello ser verdad?"

Respuesta de Marilyn Ask: "Ud. ha supuesto que "típico" se refiere al promedio aritmético de números. Pero "típico o corriente" también significa "en el medio" y "más común o frecuente" (los estadísticos llaman a estas tres clases de promedio: media, mediana y moda, respectivamente).

Veamos cómo son usadas las tres medidas: Suponga que recibirá 5 invitados a su fiesta. Sus edades son 100, 99, 17, 2 y 2. Le comenta al mayordomo que en términos medios la edad es de 44 años:

$$\frac{100+99+17+2+2}{5} = \frac{220}{5} = 44$$

Sólo para estar a salvo, le comenta que la edad promedio es 17 (la edad posicionada justamente en el medio) y para asegurarse de que todo está correcto, le cuenta al cochero que la edad corriente es 2 (la edad más común).

Cada uno de los acompañantes fue tratado con pure de papa y zapallo, acompañados por el último CD de Michael Jackson y seguido con un buen cognac. ¡En fin!

Para el caso sobre el ensayo sobre el sexo "típico" pudo haberse referido a "más común", el cual se ajustaría a todos los estereotipos (Eso, si Ud. cree en los ensayos sobre sexo).

- a)** ¿Está de acuerdo con la conclusión del Marilyn Ask "que el ensayo registró a la moda como típico"?
- b)** Las edades sugeridas por Marilyn Ask proporcionan distintas medidas de los que llamamos "típicos". Piense un ejemplo de conjunto de 5 observaciones de las cuales el 80% de ellas cae por debajo de la media y para las cuales la mediana es igual al mínimo.

4.3. La edad media (o promedio) de 10 adultos en un salón es 35 años. Un adulto de 32 años ingresa al salón. ¿Puede encontrar la edad promedio de los 11 adultos? Si su respuesta es sí, hágelo. Si su respuesta es no, explique por qué no.

La edad mediana para los 10 adultos del salón es 35 años. Un adulto de 32 años entra al salón. ¿Puede hallar la nueva mediana de los 11 adultos? Si su respuesta es sí, háguela. Si su respuesta es no, explique por qué no.

4.4. Los salarios de los atletas super estrellas profesionales reciben mucha atención en la media. El contrato anual de un millón de dólares se toma como algo común entre este grupo élite. Sin embargo raramente pasa un año sin que uno o más de las asociaciones de los jugadores negocien con cada uno de los dueños de los equipos, por salarios adicionales y consideraciones benéficas para todos los jugadores de un deporte particular.

- a)** Si la Asociación de jugadores desea sustentar un argumento para conseguir salarios "promedio" más altos ¿Cuál medida de tendencia central usaría? ¿Por qué?
- b)** Para refutar tal argumento, ¿qué medida de centro los propietarios aplicarían a los salarios de los jugadores? ¿Por qué?

4.5. ¿Dónde están los cereales para niños ubicados sobre los estantes del exhibidor?

En la parte de abajo (estante 1), en la parte de arriba (estante 3) o en la parte media (estante 2) donde los chicos pueden verlos?

Los siguientes datos proveen el contenido de marcas de 76 cereales dispuestos en diferentes estantes.

ESTANTE 1 (más bajo):
10, 6, 1, 3, 2, 11, 15, 10, 11, 6, 2, 3, 0, 0, 0, 3, 3, 3, 8

Si los datos provienen de una muestra, las medidas de variación se denominan estadísticas. Si los datos constituyen la población entera, luego, las medidas de variación serán parámetros. La notación para representar la desviación estándar de una muestra diferirá de la de la desviación estándar de la población.

- ESTANTE 2 (medio):**
14, 2, 9, 13, 12, 13, 0, 13, 7, 12, 9, 11, 3, 6, 12, 3, 9, 12, 15, 5, 12
ESTANTE 3 (más alto):
6, 8, 5, 0, 8, 8, 5, 7, 7, 3, 10, 5, 3, 4, 6, 9, 11, 11, 13, 7, 2, 10, 14,
3, 0, 0, 6, 8, 6, 3, 14, 3, 3, 12

- a) Para cada estante, halle la media y la mediana, previa identificación de la variable en estudio.
b) Para cada estante, realice un histograma o diagrama de frecuencias de la variable en estudio. Describa la forma de cada distribución y comente. Relacione la mediana y media sobre la forma de la distribución.

4.3. MEDIDAS DE VARIACIÓN O DE DISPERSIÓN

Las medidas de posición central son útiles pero a menudo dan una interpretación incompleta de los datos.

Consideremos las siguientes listas de números y sus correspondientes gráficos de frecuencias. Ambos conjuntos de datos tienen la misma media, mediana y modo, pero los valores obviamente difieren en lo que a dispersión o dispersión respecta.

Los valores del listado 1 están concentrados en torno a 60, en cambio los valores del listado 2 están muy dispersos.

List 1: 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65										List 2: 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85									
Media = mediana = moda = 60										Media = mediana = moda = 60									
desv. = 3.87										desv. = 14.17									

RANGO; AMPLITUD O ALCANCE

Es la medida de la variación más simple. Es la diferencia entre la mayor y la menor de las observaciones de un conjunto de datos.

En el estudio médico, la edad de los 20 sujetos fue de 32 a 51 años, por lo que el rango resulta $51 - 32 = 19$ años.

Con frecuencia, el rango es utilizado en las informaciones del establecimiento de tiempo, las cuales proveen por ejemplo la temperatura máxima y mínima del día. Dado que el rango es calculado a partir de los dos valores más extremos, el mismo puede dar una distorsión del modelo real de variación.

O sea, si bien el rango se calcula con facilidad, su debilidad preponderante es que no toma en consideración la forma en que se distribuyen los datos entre los valores más pequeños y los más grandes. Las dos distribuciones representadas abajo tienen la misma amplitud o rango, pero el modelo de variación es muy diferente.

La primera presenta la mayoría de sus valores lejos del centro, mientras que la segunda tiene la mayoría de sus valores cerca del centro.

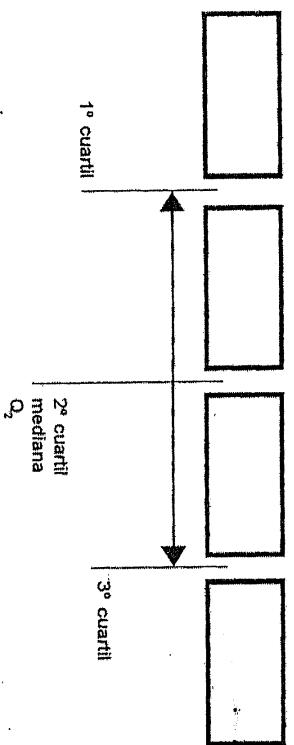


RANGO INTERCUARTÍLICO

Una medida de dispersión que aún mantiene la idea de un rango pero no está influenciada por los valores extremos es el rango intercuartílico, también llamado dispersión media. Considera la dispersión del 50% medio de los datos. La idea es dividir a los datos ordenados en cuatro partes iguales y ver la distancia de las dos partes extremas. Para dividir a los datos, primero hallamos la mediana (representada por Q2 o el valor que divide a los datos en dos mitades) y luego hallar la mediana para cada mitad.

Algunas de las medidas de variación: rango, rango intercuartílico, variancia y desvió estándar (luego se completan con otras medidas sugeridas por el Análisis Exploratorio de Datos -EDA-). Estas medidas numéricas de resumen describen la dispersión que se encuentra en los datos. Un valor grande de ella indica mayor variación.

los tres valores que dividen a los datos en cuatro partes se denominan cuartilos (cuartiles) y se representan con Q_1 , Q_2 y Q_3 .



A la diferencia entre el tercer y primer cuartilo la denominaremos **rango intercuartílico** y lo simbolizaremos **IQR** o bien **IQD** y resulta igual a $Q_3 - Q_1$ [$IQR = Q_3 - Q_1$]

HALLANDO LOS CUARTILLOS

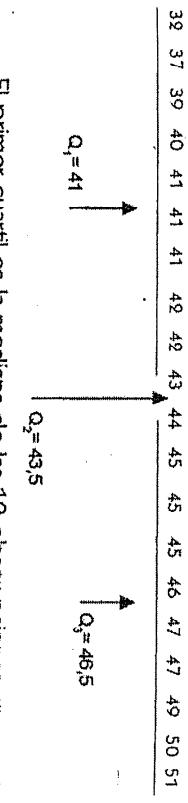
1. Hallar la mediana de todas las observaciones.
2. Primer cuartilo: Q_1 = la mediana de las observaciones que caen debajo de la mediana.
3. Tercer cuartilo: Q_3 = la mediana de las observaciones que caen por encima de la mediana.

Notas

- Cuando el número de observaciones es impar, el medio de las observaciones es la mediana. Esta observación no se incluye en las dos mitades para calcular Q_1 y Q_3 .
- Los distintos libros, calculadoras o computadoras pueden usar distintas maneras aproximadas de calcular los cuartilos, pero todos están basados sobre la misma idea.
- En la distribución asimétrica por izquierda, el primer cuartilo estará más lejos de la mediana que lo que estará el tercer cuartilo.
- Si la distribución es simétrica, los cuartilos estarán a la misma distancia de la mediana.

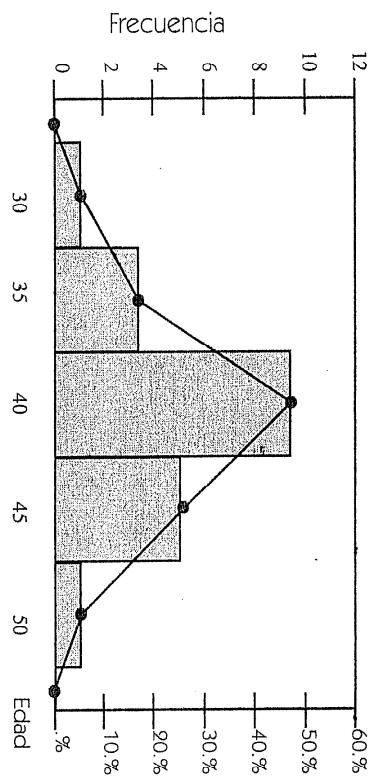
EJEMPLO 4.4. Cuartilos para la edad

En el estudio médico las edades de los 20 sujetos están listadas abajo en orden. Hallamos la mediana: $Q_2 = \tilde{x} = 43.5$ años.



El primer cuartil es la mediana de las 10 observaciones que caen por debajo de 43.5. El promedio de las 5^a y 6^a observaciones es 41 y es el Q_1 . El 3er cuartilo es la mediana de las 10 observaciones que caen por arriba de 43.5. El promedio de la 5^a y 6^a observación que caen encima de la posición de la mediana (de 43.5) es el 3er cuartilo. $Q_3 = 46.5$.

El rango para este conjunto de datos es 19 años. El rango intercuartílico es $[46.5 - 41] = 5.5$ años. Vemos que la distribución de la edad es aproximadamente simétrica y los cuartilos están casi a la misma distancia de la mediana.



El primer cuartil, Q_1 , es el valor de la variable tal que el 25% de las observaciones caen o se encuentran por debajo de él. El tercer cuartil, Q_3 , es el valor de la variable tal que el 75% de las observaciones caen por debajo de él. Los cuartilos son realmente los 25°, 50° y 75° percentiles. En general, el pésimo percentil es el valor de la variable tal que el p% de las observaciones caen por debajo de él.

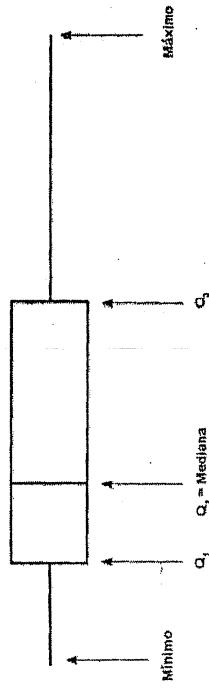
CINCO NÚMEROS DE RESUMEN

Cuando la mediana es usada como una medida de tendencia central, el IQR o RI es generalmente usado como una medida de dispersión. Si consideramos los cuartilos y agregamos, los valores mínimo y máximo, tendremos **5 medidas resumenes** de los datos.

Podemos mostrar estas 5 medidas en un gráfico llamado **Diagrama de Caja** (Box Plot).

Esas 5 medidas proveen una simplificación del conjunto entero de datos. Provee una medida del centro a través de la mediana y medidas de dispersión a través de RI o IQR y el rango. La distancia de los cuartilos a la mediana puede indicarnos asimetría, la cual es chequeada mejor mediante el examen del histograma o diagrama de tallo y hoja.

Cinco números resumen: Mínimo, Q_1 , Mediana, Q_3 , Máximo



Un **diagrama de caja** es construido como sigue:

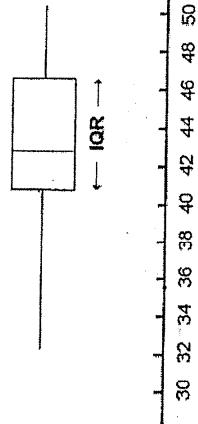
- Los cuartilos 1º y 3º son los extremos de la caja
- La línea dibujada dentro de la caja es la mediana.
- Las líneas, llamadas también "pattillas", son extendidas desde los extremos de la caja hacia cada uno de los valores $x_{(1)}$ y $x_{(n)}$.

EJEMPLO 4.5. Cinco números de resumen y Diagrama de Caja para la edad

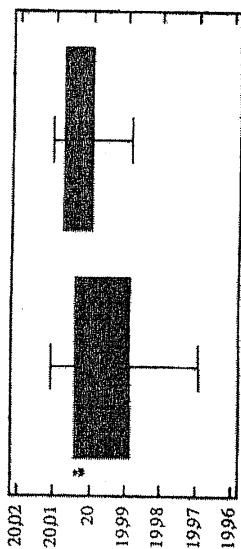
Los cinco números de resumen para el conjunto de datos referido a la edad de los sujetos son: $x(1)=32$; $Q_1=41$; $Q_2=43.5$; $Q_3=46.5$; $x(n)=51$

El Diagrama de Caja

Q_1 , Q_2 , Q_3



La distancia entre la mediana y los cuartilos es aproximadamente la misma, sustentando una distribución casi simétrica tal como se vio previamente en el histograma. Los **Box-Plots lado a lado** (múltiples) son buenos para comparar dos o más distribuciones respecto a los 5 números de resumen. La figura siguiente muestra los diagrama de caja para los dos procesos que producen partes de 20,000 cm. Podemos comparar estos diagramas con los diagramas lado a lado de Tallos y hojas.



Process 2	Process 1
1996,9	
1997,5	
1998,445	
1999,2478	
2000,012444788	
2001,1	

Nota: 911998 representa
19,989 cm.

Nota: 199814 representa
19,984 cm.

Nota: 911998 representa
19,988 cm.

Nota: 199814 representa
19,984 cm.

Aunque la mediana del primer proceso está más cerca del valor estipulado, el segundo proceso presenta menos variabilidad. Muchos paquetes estadísticos incorporan algunas modificaciones al Diagrama de Caja. Algunos ubican un * en la caja para indicar la media. De esa manera media y mediana se pueden comparar. Otros incorporan la "Regla del pulgar" para identificar potenciales outliers o valores extremos y grafican estos potenciales outliers separadamente del gráfico.

REGLA DEL PULGAR PARA IDENTIFICAR POTENCIALES OUTLIERS

- Hallar $1.5 \times \text{IQR}$
- Trazar las barreras internas:
 $Q_1 - 1.5 \times \text{IQR}$ $Q_3 + 1.5 \times \text{IQR}$
- Las observaciones que caen fuera de las barreras internas son consideradas como potenciales outliers.

Diagrama de Caja Modificado: el mismo es modificado, cuando se dibujan separadamente algunos potenciales outliers y luego extienden los patillos hasta que no haya más outliers (hasta el valor "adyacente" al outliers).

El Box Plot siguiente fue producido usando un paquete estadístico de computación. Los cinco números de resumen para estos datos son:

$x(1) = 1; Q_1 = 21; Q_3 = 32; Q_{\text{mediana}} = 32.5; x(n) = 325$. El RIO o IQR = 45; el primer paso previo al trazado de las barreras internas: $1.5 \times 45 = 67.5$, las barreras internas se trazan en (-46,5) y 133,5.



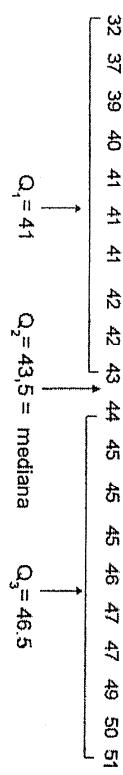
*

Observe el (*) y el círculo (o) a la derecha de las barreras internas. Ellos representan dos potenciales valores extremos o "outliers" cuyos valores son 165 y 325 aproximadamente.

El 165 es un valor "extremo" (está entre las barreras internas y externas). El 325 es un valor "lejano" situado fuera de las barreras externas. Este tema se completará en clase y se entregará material adicional.

EJEMPLO 4.5. ¿Alguna edad extrema?

Aplicamos nuestra regla para determinar si hay valores extremos o outliers en nuestro conjunto de datos referidos a la edad de los sujetos.

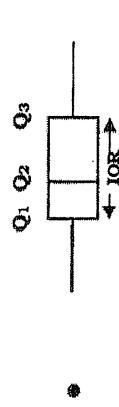


El RIO o IQR = $(46.5 - 41) = 5.5$. El primer paso $1.5 \times 5.5 = 8.25$. Las barreras internas estarán situadas en $(41 - 8.25) = 32.75$ y $(46.5 + 8.25) = 54.75$. Hay exactamente una observación que cae fuera de las barreras internas (el 32 que es el menor valor observado). Así, vemos que hay un potencial outlier basado en esta regla. Ésta será una información importante cuando analicemos las respuestas de interés, o sea, la presión sanguínea.

El Box Plot modificado para la edad se observa abajo. Advierta que la patilla izquierda llega hasta pasado 37, la observación pequeña no es considerada un outlier.

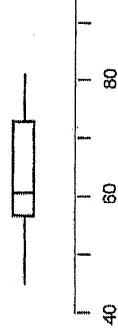
El Diagrama de Caja

a) Halle las 5 medidas de resumen para los grupos tratados. Comente sus resultados.



PARA RESOLVER!! 4.6. Costo de las zapatillas

Los precios de 12 pares de zapatillas de competición se muestran en el siguiente Box plot:



Precio

a) ¿Cuál es el rango aproximado de los precios de tales zapatillas?
Rango _____

b) ¿El 25% de las zapatillas cuestan más de qué cantidad aproximada?
\$ _____

PARA RESOLVER III 4.7. Comparando edades...estudio de los antibióticos

Recuerda

Variable: edad de los 23 niños a los que le fueron asignados aleatoriamente uno de los dos tratamientos.

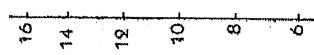
Se proveen los datos ordenados:

a) Halle las 5 medidas de resumen para los grupos tratados. Comente sus resultados.

Grupo Amoxicilina (n=11) 8 9 9 10 10 11 11 12 14 14 17
5 medidas resumen:

Grupo Cefadroxil (n=12) 7 8 9 9 10 10 11 12 13 14 16
5 medidas resumen:

b) Realice los **Box Plot lado a lado** para el estudio de los antibióticos de la parte a)



c) Usando la "regla del pulgar" ¿Hay algún(os) outlier(s) para el grupo Amoxicilina? Si los hay, modifique su Box-plot aparte de b)

d) Idem c) para el grupo Cefadroxil.

Resumitulando

Las 5 medidas de resumen y el correspondiente Box Plot nos proveen un buen análisis exploratorio de los datos. La mediana nos da la medida del centro de la distribución. La longitud de la caja, el RI o IQR, nos proporciona una medida de la dispersión y la distancia de los cuartílos a la mediana nos dan una idea de la simetría. Sin embargo, cuando examinamos los Box Plots, ellos pueden esconder baches o gaps y múltiples picos o máximos, es decir, no nos muestran en forma completa la forma de la distribución. La simetría de una distribución no implica que sea unimodal.

Suponga que el tamaño total de muestra es el mismo para 3 diseños:

a) ¿Para qué población (A, B o C), los diseños i) y ii) serán comparativamente efectivos? Explique.

b) ¿Para qué población (A, B, o C), el diseño iii) será el mejor diseño? Explique.

c) ¿Para qué población (A, B, o C), el diseño iii) será el mejor diseño? Explique.

DESVIACIÓN ESTÁNDAR

Cuando la media aritmética es usada como medida de centro, la medida de dispersión más común es la desviación estándar.

Al igual que en la media aritmética, la desviación estándar hace uso de todas las observaciones para su cálculo. Establece la forma en que los valores fluctúan con respecto a la media. Es la raíz cuadrada del promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media.

Puede pensarse como la distancia promedio entre las observaciones y el promedio. Para hallar la desviación estándar, primero hallamos la variancia, que es el promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media aritmética. La desviación estándar es la raíz cuadrada positiva de la variancia. No es fácil calcular la desviación estándar pero las calculadoras hacen esta tarea. En los dos próximos ejemplos veremos como calcular a mano la desviación estándar luego de comprender e interpretar como actúa la desviación estándar como medida de dispersión es mejor utilizar una calculadora.

EJEMPLO 4.8. Desviación Estándar: ¿Qué es?

Supongamos tener sólo tres valores en nuestra muestra:

0, 5 y 7. Y deseamos calcular la desviación estándar de la muestra.

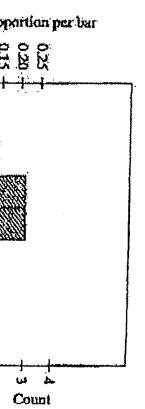
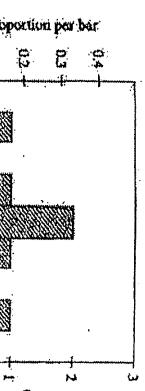
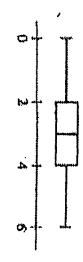
Observemos que la media aritmética es igual a 4 (compruébalo).

El cuadro siguiente muestra los desvíos de los valores observados con respecto a la media.

Repetida en la
250.

248

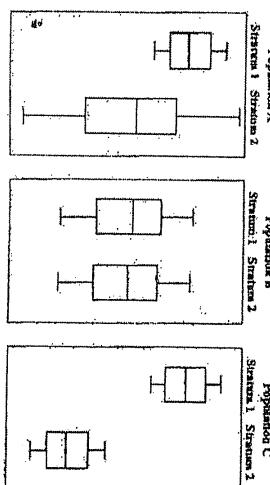
EJEMPLO 4.8. Un box plot simétrico no implica una distribución simétrica



PARA RESOLVER !!! 4.8. Diseño muestral efectivo

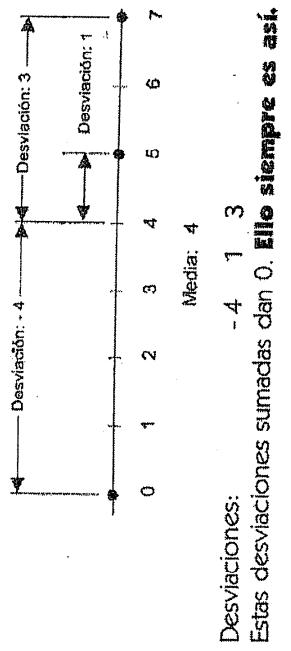
4.8. Diseño muestral efectivo

Los gráficos muestran los Boxplots lado a lado de algunas variables para dos estratos en 3 poblaciones hipotéticas: A, B y C. En cada población, las unidades están igualmente divididas en 2 estratos:



Consideré 3 diseños muestrales para estimar la verdadera, media poblacional:

- i) Muestreo simple al azar
- ii) Muestreo aleatorio estratificado tomando igual tamaños muestrales de los dos estratos.
- iii) Muestreo estratificado tomando más unidades de un estrato y pocas de otro estrato.



Supongamos que el tamaño total de muestra es el mismo para 3 diseños:
 a) ¿Para qué población (A, B o C), los diseños i) y ii) serán comparativamente efectivos? Explique.
 b) ¿Para qué población (A, B o C), el diseño ii) será el mejor diseño?
 Explique.
 c) ¿Para qué población (A, B o C), el diseño iii) será el mejor diseño?
 Explique.

DESVIACIÓN ESTÁNDAR

Cuando la media aritmética es usada como medida de centro, la medida de dispersión más común es la desviación estándar. Al igual que en la media aritmética, la desviación estándar hace uso de todas las observaciones para su cálculo. Establece la forma en que los valores fluctúan con respecto a la media. Es la raíz cuadrada del promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media. Puede pensarse como la distancia promedio entre las observaciones y el promedio.

Para hallar la desviación estándar, primero hallamos la variancia, que es el promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media aritmética. La desviación estándar es la raíz cuadrada positiva de la variancia. No es fácil calcular la desviación estándar pero las calculadoras hacen esta tarea. En los dos próximos ejemplos veremos cómo calcular a mano la desviación estándar. Luego de comprender e interpretar cómo actúa la desviación estándar como medida de dispersión es mejor utilizar una calculadora.

EJEMPLO 4.8. Desviación Estándar ¿Qué es?

Supongamos tener sólo tres valores en nuestra muestra:

0, 5 y 7 y deseamos calcular la desviación estándar de la muestra. Observemos que la media aritmética es igual a 4 (compruébelo). El cuadro siguiente muestra los desvíos de los valores observados con respecto a la media.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad [\text{Propiedad de la media aritmética}]$$

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

Dado que no podemos usar la suma de los desvíos como medida de dispersión, usaremos los desvíos al cuadrado.

Desviaciones al cuadrado: 16 1 9

La siguiente tabla resume los cálculos para computar la variancia y luego la desviación estándar.

X_i	desviaciones: (x_i - x̄)	desvíos al cuadrado: (x_i - x̄)²
0	0 - 4 = -4	16
5	5 - 4 = 1	1
7	7 - 4 = 3	9

$$\text{Media aritmética: } (\bar{x}) = 4 \quad \sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 26$$

$$\text{Observación: } \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

$$\bullet \quad \sum_{i=1}^n (x_i - \bar{x})^2 = \text{mínimo}$$

La variancia muestral (S^2) se la define como la suma de estos desvíos al cuadrado dividido por el tamaño de la muestra menos 1 ($n-1$). Luego analizaremos el por qué de dividir por $(n-1)$ y no por n .

$$\text{Variancia muestral: } S^2 = \frac{(-4)^2 + 1^2 + 3^2}{3-1} = \frac{16+1+9}{2} = 13$$

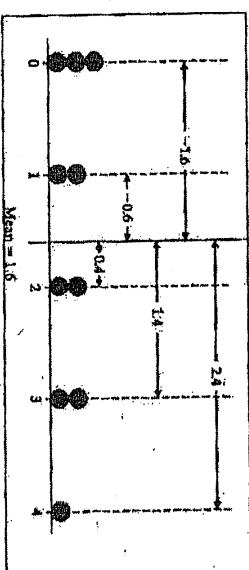
Este valor 13 es "casi" un promedio de los desvíos al cuadrado. Las desviación estándar (S) es: $S = \sqrt{13} \equiv 3,6$

EJEMPLO 4.10.

Desviación estándar para el n° de hijos por hogares

Recordemos el conjunto de datos referidos al número de chicos por hogares de una cierta vecindad:
2, 3, 0, 2, 1, 0, 3, 0, 1, 4

1. Primero calculamos la media aritmética = 1,6



2. Calculamos los desvíos de las observaciones con respecto a la media y luego los elevamos al cuadrado:

Observación	Diferencia	Squared Deviation
2	0,4	0,16
3	1,4	1,96
0	-1,6	2,56
2	0,4	0,16
1	-0,5	0,25
0	-1,6	2,56
3	1,4	1,96
0	-1,6	2,56
1	-0,6	0,36
4	2,4	5,76
intens	1,6	sum = 18,40
		sum = 0
		sum = 18,40

3. La variancia muestral es:
 $S^2 = 18,4/9 = 2,044$

4. La desviación estándar es la raíz cuadrada de la variancia:
 $\text{Desviación estándar} = S = \sqrt{2,044} = 1,43$

El número medio de chicos por hogar fue 1,6 y la desviación estándar fue 1,43. De allí que podríamos resumir estos datos diciendo: "Los hogares tienen -en promedio- 1,6 hijos con una desviación de 1,43 hijos".

La desviación estándar es aproximadamente la distancia promedio entre las observaciones y la media. Ello no significa que el número de chicos esté a 1,43 de 1,6, es decir, entre 0,17 y 3,03 chicos para todos los hogares, ya que de ser así habría 4 hogares para los cuales el número de chicos está fuera de este rango. Explicar.

Dado que la desviación estándar es la medida de dispersión más común, ella también tiene alguna simbología especial, con la cual nos familiarizaremos:

Si $X_1, X_2, X_3, \dots, X_n$, constituye una muestra de n observaciones, la variancia muestral será:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \dots = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$$

La desviación estándar = $S = \sqrt{S^2}$

Si en lugar de trabajar con datos provenientes de una muestra, lo hacemos con todas las observaciones de una población, la variancia poblacional (σ^2) es el promedio de los cuadrados de los desvíos de las observaciones con respecto a la media poblacional.

La desviación estándar poblacional, simbolizada con σ (sigma) es la raíz cuadrada de la variancia poblacional.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}; N: \text{número total de observaciones}$$

Notes

- La variancia está medida en unidades al cuadrado. Si la edad está medida en años y la variancia es 1,44, citaremos a la variancia como 1,44 años². Sacando la raíz cuadrada de la variancia obtenemos la desviación estándar expresada en unidades originales. La desviación estándar para la edad es 1,2 años, o podríamos decir que en promedio las observaciones estaban 1,2 años de la edad promedio.
- Hay argumentos estadísticos para sustentar por qué dividimos por $n-1$ en lugar de n en el denominador de la desviación muestral. Si deseamos usar la variancia muestral como un estimador de la variancia poblacional, usando el $n-1$ nos dará un "mejor" estimador. En el capítulo 9 examinaremos esta idea, y veremos que la cantidad $n-1$ será considerada como **nº de grados de libertad**.
- Dado que la suma de los desvíos de las observaciones con respecto a la media es siempre cero, sólo necesitamos conocer $n-1$ variables para determinar la restante ($(n-1)$) de las desviaciones que están libres).
- Como la media no es una medida resistente del centro y la desviación estándar usó la media en su definición, ella no es una medida de dispersión resistente. Está fuertemente influenciada por los valores extremos.
- Pensemos a la desviación estándar como (aproximadamente) la distancia promedio entre las observaciones y la media. Si todas las observaciones tienen exactamente el mismo valor, luego la desviación estándar será igual a 0 (cero). (No hay dispersión). Caso contrario la desviación estándar siempre será positiva. Cuanto más "separadas" o dispersas estén las observaciones de su media, mayor será el valor de la desviación estándar.

- a) ¿Cuál de los conjuntos de datos tiene la menor desviación estándar? A B C.
- b) ¿Cuál de los conjuntos de datos presenta la mayor desviación estándar? II I III
- c) Halle la desviación estándar para cada uno de los conjuntos y chequee sus respuestas a) y b)



PARA PENSAR !!!

Dado que dos (o más) conjuntos de n observaciones presentan la misma desviación estándar, ¿muestran ellos la misma variabilidad? De todos modos ¿qué es la variabilidad?

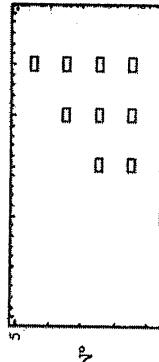
EJEMPLO 4.11.

¿Qué es la variabilidad?

Consideremos los siguientes conjuntos de datos [observe sus gráficas]

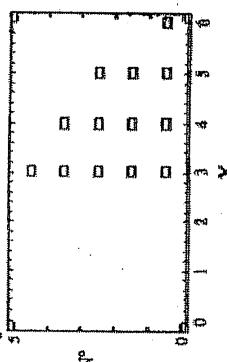
Conjunto Datos I: ($n = 13$): 2 3 3 3 4 4 4 5 5 5 5

DISTRIBUCIÓN I



Conjunto Datos II: ($n = 13$): 3 3 3 3 3 4 4 4 5 5 5 6

DISTRIBUCIÓN II



PARA RESOLVER !!!

4.9. Aumentando la dispersión

Considere los tres conjuntos de datos siguientes

- I) 20 20 20
- II) 18 20 22
- III) 17 20 23

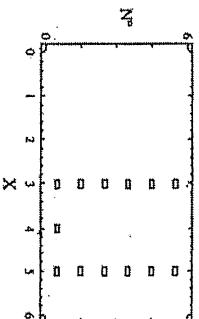
Conjunto Datos III: ($n = 13$): 2 3 3 4 4.4 4 4 4 4 5 5 6

DISTRIBUCIÓN III



Conjunto Datos IV: ($n = 13$): 2 3 3 3 3 3 4 5 5 5 5 5 5

DISTRIBUCIÓN IV



La media aritmética para cada uno de los cuatro conjuntos de datos es

$$\bar{x} = 4$$

La tabla presenta tres medidas de variabilidad para cada una de las cuatro distribuciones.

Medidas de variabilidad	I	II	III	IV
Rango 3	3	3	4	2
Rlo IQR	2	2	0	2
Desvío estndar	1	1	1	1

Referencia: A. J Nitko (1983) Educational Tests: An Introduction

Si observamos el rango: la Distribución III es más variable; si miramos el RI o IQR: la distribución III es menos variable, mientras que todas las distribuciones presentan la misma desviación estndar!!! Algunas personas asocian variabilidad con rango, otras asocian variabilidad con cuántos valores difieren de la media. Hay muchas medidas de variabilidad, generalmente la desviación estndar

es la más ampliamente usada. Pero lrecuerde! Que una distribución con la desviación estndar más pequeña no es necesariamente la distribución que presenta la mínima variabilidad con respecto a otras definiciones o a su propia definición de variabilidad.

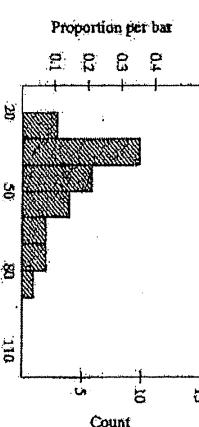
PARA PENSAR !!!

¿Qué pasaría si el último valor observado en cada uno de los cuatro conjuntos de datos fuera cambiado por 16? Recalcule las tres medidas de variabilidad para cada uno de esos conjuntos y realice algún comentario.



PARA RESOLVER III

Consideré la distribución de los precios de los calzados de competición dada por:

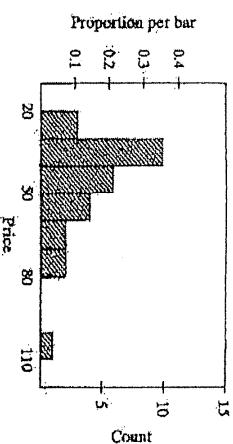


a) ¿Qué valores representarían mejor a la media y a la mediana?

Valores: 40 45
Media: 45
Mediana: 40

Explique:

b) Suponga que una observación más grande fue incorrectamente ingresada, según lo demuestra la siguiente figura:



Rodee la mejor respuesta:

La media: [incrementará]
 [decrecerá] será la misma

La mediana: incrementará
 decrecerá [será la misma]

La desviación estándar: [incrementará]
 [decrecerá] será la misma

PARA RESOLVER !!!
A.11. Desviación estándar para las edades

Las edades de los 20 sujetos del estudio médico fueron listadas en orden. Se halló la media aritmética y fue de 43,35 años.

32 37 39 40 41 41 41 41 42 43 44 45 45 46 47 47 49 50 51

a) Halle la desviación estándar para estos datos.

b) Complete la oración: En promedio, las edades de estos sujetos dis-
 tan 1,5,6 años de su media de 43,35 años.

c) ¿Cuántos de los 20 sujetos tienen edades dentro de una desviación
 estándar de la media?

d) ¿Cuántos de los 20 sujetos tienen edades dentro de dos desvia-
 ciones estándar de la media?

RANGO INTERCUARTÍLICO Y DESVIACIÓN ESTÁNDAR

El RI o IQR es la distancia entre el primer y tercer cuartilo ($Q_3 - Q_1$) y mide la dispersión del 50% central de los datos. Cuando la mediana es usada como medida de posición, generalmente el IQR se usa como medida de variabilidad. Para distribuciones asimétricas o distribuciones con outliers, el RI tiende a ser la mejor medida de dispersión si nuestro objetivo es resumir esa distribución. Agregando los valores mínimo y máximo a la mediana y cuartilos tenemos las 5 medidas o números resúmenes. Una representación gráfica de estos 5 números resúmenes es el Box Plot y la longitud de la caja corresponde al RI o IQR.

La desviación estándar es aproximadamente el promedio de las distan- cias de los valores observados y su media. La media y la desviación estándar son las medidas más usadas para distribuciones simétricas sin outliers.

En el próximo capítulo analizaremos una importante familia de dis- tribuciones simétricas llamadas distribuciones normales; para las cuales la desviación estándar es una medida resumen muy útil.

Consejo

Las medidas resúmenes numéricas presentadas en este capítulo pro- veen información acerca del centro y dispersión de una distribución, pero un gráfico tal como un histograma o un diagrama de tallo y hoja proveen la mejor "foto" de la forma general de la distribución. **¡Gráfico primero sus datos!!**

Observación

Se entregará material adicional para completar el tema de estudio con otras medidas sugeridas por Tukey en el análisis exploratorio de datos.



- 7. Suponga que está interesado en la compra de un auto y en particular desea elegir uno de los dos modelos que le interesa: modelo A y modelo B. La mayoría de las características son casi iguales; tales como precio y opciones, costo promedio de mantenimiento anual. Pero leyendo una revista automovilística encuentra que la desviación estándar

del costo de mantenimiento es más pequeña para el modelo B. Basado en esta información ¿cuál de las siguientes citas es la apropiada?

- El modelo A con una desviación estándar más grande es preferible porque un valor grande indica una menor variación en los datos.
- El modelo B, con una desviación estándar más pequeña es preferible porque un valor pequeño indica que la media es la medida más confiable de los costos de mantenimiento.
- Ellos son igualmente aceptables porque la desviación estándar no es útil para la comparación de los conjuntos de datos.

4.8. Los cinco números de resumen para la distribución de los resultados de las pruebas son: 340 460 580 780 950

- Gráfique un box plot simple para la distribución de los resultados de las pruebas.
- Suponga que Ud. obtuvo 470 en su prueba. ¿Qué puede decir sobre el porcentaje de estudiantes que obtuvieron más puntaje que Ud.?
- Si el 25% de los estudiantes obtuvo "Sobresaliente" en la prueba, ¿cuál fue el mínimo resultado que se necesitó para conseguir un "Sobresaliente"?

4.9. ¿Dónde están los cereales para niños ubicados sobre los estantes del almacén?

¿En la parte de abajo (estante 1), en la parte de arriba (estante 3) o en la parte media (estante 2) donde los chicos pueden verlos? Los siguientes datos proveen el contenido de azúcar de 76 tipos de cereales dispuestos según estantes.

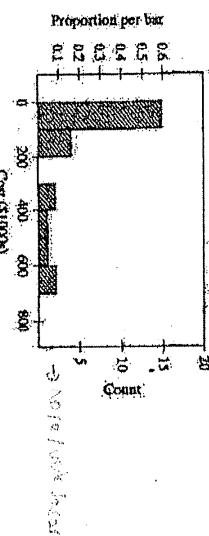
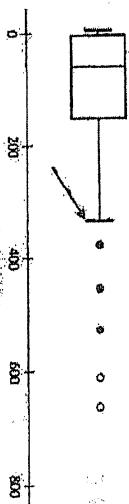
Estante 1 (más bajo): 10, 6, 1, 3, 2, 11, 15, 10, 11, 6, 2, 3, 0, 0, 0, 3, 3, 3, 8

Estante 2 (medio): 14, 2, 9, 13, 12, 13, 0, 13, 7, 12, 9, 11, 3, 6, 12, 3, 9, 12, 15, 5, 12

Estante 3 (más alto): 6, 8, 5, 0, 8, 8, 0, 5, 7, 7, 3, 10, 5, 10, 5, 3, 4, 6, 9, 11, 11, 13, 7, 2, 10, 14, 3, 0, 0, 6, 8, 6, 3, 14, 3, 3, 12

- Para cada estante, halle el RI o IQR y la desviación estándar para el contenido de azúcar.
- ¿Cuáles medidas resúmenes preferiría para describir el contenido de azúcar la media aritmética y la desviación estándar, o los 5 números resumen? Explique. (Gráfique previamente).

4.10. Suponga que Ud. es un empresario importante. Los costos iniciales, en miles de dólares, para las 25 franquicias más actuales, según "Entrepreneur magazine" están resumidos abajo:



a) La cadena de restaurantes "Hardes" tuvo el más alto costo de iniciación. ¿Cuál es el costo inicial aproximado para un restaurante "Hardes"?

- b) Describa la forma de la distribución.
c) Se dijo que para estas 25 franquicias, el costo inicial característico o típico fue de \$154.700. ¿Qué medida de tendencia central representa este valor? Explique.

- d) ¿A qué valor corresponde el marcado con una flecha sobre el box plot?

4.11. Dos docentes están comparando los resultados de los finales. Cada docente tiene 21 estudiantes.

En la clase A: un estudiante recibió 30 puntos, otro obtuvo 70 y el resto 50 puntos. En la clase B: Un estudiante obtuvo 30 puntos, otro 32, otro 34 y así sucesivamente hasta alcanzar 70 puntos.

En la clase C: 10 estudiantes obtuvieron 30 puntos, uno obtuvo 50 puntos y 10 obtuvieron 70 puntos. Las distribuciones son:

PARA RESOLVER !!!
4.12. UNA TRANSFORMACIÓN

Recuerde los datos sobre el número de chicos por hogares en 10 hogares de una cierta vecindad:

2 3 0 2 1 0 3 0 1 4

Encontramos la media =1,6 y la desviación estándar =1,43. Supongá que deseamos resumir el número de personas en un hogar. Cada hogar tiene dos adultos, por lo tanto, simplemente podemos sumar dos a cada valor observado de la variable: 4, 5, 2, 4, 3, 2, 5, 2, 3, 6.

- a) Midiendo la variación con el rango ¿Qué clase tiene mayor variación? (rodeé una respuesta)

Clase A Clase B Clase C Las 3 iguales

- b) Midiendo la variación a través del desvío estándar ¿Cuál clase presenta mayor variación? (rodeé una respuesta sin efectuar el cálculo)

Clase A Clase B Clase C Las 3 iguales

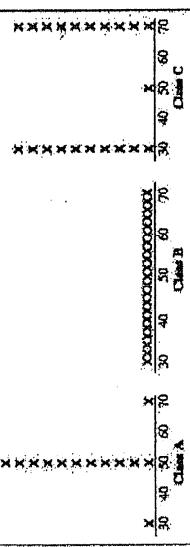
- (C) Registre los 5 números de resumen para cada clase.

Clase A:
 Clase B:
 Clase C:

4.4. TRANSFORMACIONES LINEALES Y ESTANDARIZACIÓN

Un trabajador tiene la tarea de medir las longitudes de las partes que salen de un cierto proceso de fabricación. Las 3 partes seleccionadas fueron medidas y registrado como: 2 cm, 2,5 cm y 3 cm. ¿Cuáles serán las longitudes de estas partes si la unidad de medida fuesen mm en lugar de cm?

Dado que hay 10 mm en 1 cm, tendremos 20 mm, 25 mm y 30 mm. Una conversión entre estos dos sistemas de mediciones puede ser expresada como una transformación lineal. En este caso de conversión de cm a mm, multiplicamos por 10: valor en mm = $10 \times$ (valor en cm). En esta sección exploraremos cómo tales transformaciones afectan a las distintas medidas resúmenes.



c1) Halle la media y la desviación estándar de este nuevo conjunto de observaciones y compárelas con las obtenidas para las observaciones originales. ¿Cómo cambió la media? ¿Cómo cambió la desviación estándar?

c2) Resuma cómo queda afectada la media y la desviación estándar de las observaciones si a cada valor observado de la variable se le suma la misma constante.

c3) Supongá que un chico recibe una asignación de \$3 semanales. El total de asignaciones gastadas en un hogar puede ser obtenida multiplicando cada valor de la variable original por 3:

6, 9, 0, 6, 3, 0, 9, 0, 3, 12

c4) Halle la media y la desviación estándar de este nuevo conjunto de observaciones y compárelas con los de los datos originales. ¿Cómo cambió la media? ¿Cómo cambió la desviación estándar?

c5) Resuma cómo queda afectada la media y la desviación estándar cuando cada observación del conjunto de datos original se lo multiplica por la misma constante distinta de cero.

c6) El costo de la entrada por chico para el ingreso de un parque de diversiones es de \$5. Los adultos entran gratis. Cada hogar tiene, también, un cupón para abonar \$2. Sin hacer los cálculos aritméticos, ¿puede decir cuál sería la media y la desviación estándar si multiplicamos cada observación original por 5 y luego le restamos 2 (nueva variable: $y = 5x - 2$)?

En la parte (c) del ejemplo anterior ($y = 5x - 2$) es llamada **Transformación lineal de la variable X** y en general es denotada por "**a X + b**".

Una transformación lineal es un tipo de transformación, conversión o recodificación. Ello será muy útil cuando trabajemos con una familia de distribuciones llamadas distribuciones normales.

Las reglas desarrolladas en el **PARA RESOLVER III 4.12** pueden resumirse como:

X representa los valores de la variable original; \bar{x} es la media y S_x la desviación estándar de los valores de la variable original.

Los nuevos valores, representados por Y son una transformación lineal de X : $Y = aX + b$, luego:

La media para la variable Y está dada por: $\bar{Y} = a\bar{X} + b$,

La Desviación estándar de la variable Y está dada por: $S_y = |a|S_x$

EXAMPLE 4.12. Transformaciones

En una carta reciente, uno de sus primos de Europa, comentó que el último verano había sido caluroso. En particular, la temperatura de cada día para una semana fue:

X : temperatura (grados centígrados):
lunes, martes, miércoles, jueves, viernes, sábado, domingo
40, 41, 39, 41, 41, 40, 38

Basado en estos datos, la media y la desviación estándar son:

$$\begin{aligned}\bar{x} &= 40^{\circ}\text{C} \\ S_x &= 1.15\end{aligned}$$

Supongamos que Ud. vive en N. York y no está familiarizado con la escala Celsius (o Centígrado), sino con la escala Fahrenheit: $y =$ temperatura Fahrenheit está relacionada con la escala Celsius por:

$F = \frac{9}{5}C + 32$; en términos de Y y de X :

$$Y = \frac{9}{5}X + 32$$

Por lo tanto, la $\bar{Y} = \frac{9}{5}(40) + 30 = 104$ grados Fahrenheit
 $\text{y } S_y = \frac{9}{5}, 1.15$ grados Fahrenheit.

Ahora sí puede entender cuánto calor hace en Europa. Observe que no necesita transformar cada valor. ¡Use lo que aprendió en Estadística!!

PARA RESOLVER III 4.13. Una transformación especial

Haremos una transformación especial sobre los datos originales de número de chicos por hogar:

$$2, 3, 0, 2, 1, 0, 3, 0, 1, 4$$

- El primer paso será restar la media a cada valor de la variable, o sea, calcular los desvíos:
- El segundo paso será dividir dicha diferencia por la desviación estándar S_x
- Calcule la media y la desviación estándar de esos valores transformados:

Media:
Desviación estándar:

Una variable X es estandarizada, si la variable tiene media cero y desviación estándar uno.

Observe que la variable Estandarizada: $\frac{X - \bar{X}}{S_x}$

puede ser expresada en la forma de una transformación lineal:

$$\frac{X - \bar{X}}{S_x} = \left(\frac{1}{S_x} \right) X - \left(\frac{\bar{X}}{S_x} \right) \text{ con } a = \frac{1}{S_x} \text{ y } b = \left[\frac{\bar{X}}{S_x} \right]$$

Ejercicios

- a) Resuma estos datos gráficamente y describa la distribución de los resultados de los parciales. Recuerde definir la variable que estudia.
- b) Basado en la parte a) ¿preferiría usar los 5 números de resumen o la media y la desviación para describir numéricamente estos datos? Calcule la(s) medida(s) que eligió.
- c) Retomando a los registros, se observó que el puntaje 0 corresponde a un estudiante que no realizó el parcial por haberse hospitalizado. ¿Será apropiado sacar dicho valor? Explique cómo cambian las medidas descriptivas.

- 4.12.** Un docente distribuyó los exámenes parciales a sus alumnos y anuncia que el resultado promedio fue de 76 (sobre 100). Usted recibe su examen y observa que obtuvo un porcentaje de 88. ¿Cómo se sentiría Usted? Tal vez está contento por estar encima del promedio.
- a) ¿Podría su puntaje ser el resultado más alto?
- b) ¿Podría el 50% de los estudiantes haber sacado un puntaje más alto que usted?
- c) Si el docente también comenta la desviación estándar de esos puntuajes, ¿con cuál de las dos siguientes desviaciones estándares se sentiría feliz con su resultado?
- D_S = 4 puntos o D_S = 16 puntos

- 4.13.** En un ensayo clínico, los sujetos que entraron en el experimento, son asignados para un tratamiento o a un grupo control. La colección de datos continúa hasta la finalización del ensayo. Las variables tales como peso de los sujetos, pueden resultar de importancia, por lo tanto se las registrará. Considere la siguiente información sobre los sujetos que participan en este ensayo para probar una droga particular.

HOSPITAL I		HOSPITAL II	
	Tamaño muestral	Desevio promedio	Tamaño muestral
Tratamiento	123	162 lbs	23 lbs
Control	131	164 lbs	24 lbs

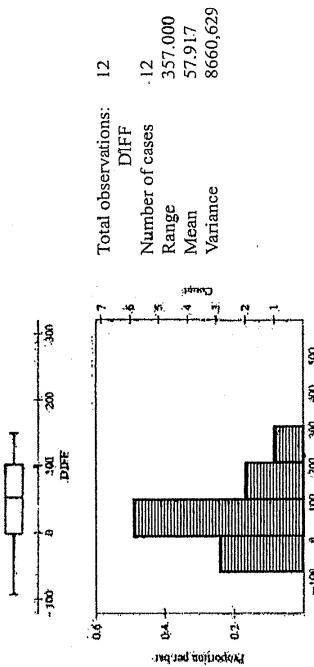
	Peso promedio	Desevio estandar	Peso promedio	Desevio estandar
Tratamiento	123	162 lbs	162	143 lbs
Control	131	164 lbs	167	178 lbs

¿Qué hospital falló en el uso de la aleatorización? Explique.

- 4.14.** ¿Qué hay acerca de un "Cero"? Se presentan las notas de los parciales (sobre 25 puntos) de los estudiantes de computación ($n = 40$ estudiantes).

19 19 20 20 22 20 17 19 20 19 14 21 21 12 17 23 17 23 16
 22 22 22 17 18 13 14 20 19 23 20 23 21 0 24 23 21 19 22
 20 21

- 4.15.** Los siguientes datos resumen la diferencia del número de aviones comprados por las 12 más grandes aerolíneas de O.S. para los años 1995 y 1998. Por ejemplo: Delta Airlines tuvo 444 en el año 1995 y se espera tener al final del 98, 583 aviones, con lo que habría una diferencia de 139 aviones.



- Use la información proporcionada para responder:
- a) ¿Aproximadamente qué porcentaje de aerolíneas se espera que tengan tan pocos aviones al final de 1998 como en 1995?
- b) American Airlines está proyectada para tener el mayor incremento en el número de aviones de 1995 a 1998. ¿Cuál es ese incremento proyectado?
- c) Basado en esos datos, el 75% de las aerolíneas tuvieron una diferencia que cayeron por debajo de aviones.

- 4.16.** Se llevó a cabo un estudio para comprobar si la cantidad de calcio en la dieta puede bajar la presión sanguínea.

A un grupo de 10 hombres se les proporcionó un suplemento de calcio diario, mientras que otro grupo control de 11 hombres recibió un placebo. Se midió la presión sistólica sanguínea de todos los hombres antes de comenzar el tratamiento y luego de 12 semanas de tratamiento.

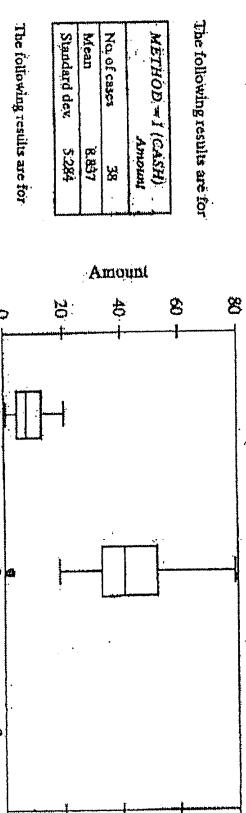
Las diferencias de presión (antes-después) para el grupo de los sujetos tratados con calcio son:

$$7 - 4 \quad 18 \quad 17 \quad -3 \quad -5 \quad 1 \quad 10 \quad 11 \quad -2$$

- a) ¿Qué indica un valor negativo?
 b) Halle la media y la desviación estándar para la diferencia de presión sanguínea.

4.17. Una cadena de supermercados opera en el Gran Rosario. Recientemente dicha cadena ofreció a sus clientes operar con tarjetas de crédito tales como Visa y Mastercard, además de los usuales opciones: efectivo o cheque.

Esta nueva opción se implementó como una prueba para ver si los clientes se incentivan para hacer compras más grandes. Luego de las primeras dos semanas, se seleccionó una muestra aleatoria de 100 clientes durante la tercera semana. Para cada cliente muestreado se registró su monto gastado (\$) y forma de pago (efectivo, cheque, tarjeta de crédito). Los datos de pagos en efectivo y cheques de los clientes están resumidos:



- a) Dé los 5 números resumen para los datos que corresponden a tarjetas de crédito y luego realice el Diagrama de caja básico en el espacio previsto arriba.
 b) Calcule la media y la desviación estándar para la forma de pago con tarjeta de crédito.

Media:
 Desviación estándar:

Algunos de los clientes que antes pagaban en efectivo o cheque, ahora pueden estar pagando con tarjetas. Antes de esa nueva opción, aproximadamente el 50% de los clientes pagaban en efectivo y aproximadamente el otro 50% lo hacían con cheques

- c) ¿Qué proporción de los clientes muestreados pagan:
 en efectivo?
 con cheques?
 con tarjeta de crédito?
 Efectúe algún comentario sobre el cambio en la distribución de la forma de pago.
 d) ¿Cuál fue el monto de la compra más pequeña que hizo el cliente que pagó en cheque?
 e) ¿Piensa Ud. que la opción de pago con tarjeta de crédito está incentivando a los clientes a realizar más compras?

TÉRMINOS CLAVES



Asegúrese que puede describir con sus propias palabras y dé un ejemplo para cada uno de los términos siguientes:

Medidas de Tendencia Central

Media

Mediana

Moda

Medidas de Variación o Dispersión

Range

Rango Intercuartílico (IQR o RI)

Cuartiles

Percentiles

Cinco números de Resumen

Diagrama de Caja

Potenciales Outliers

Diagrama de Caja modificado

Desviación estándar

Variancia

Coeficiente de Variación

Transformación Lineal

Estandarización